#  <br> <br> CogSci 2017 <br> <br> CogSci 2017 Proceedings of the 39th Annual Meeting of the Cognitive Science Society London, UK 26-29 July 2017 

## Computational Foundations of Cognition



Richard Granger, Dartmouth
Ulrike Hahn, Birkbeck, University of London Richard Sutton, University of Alberta
http://www.cognitivesciencesociety.org/conference2017/index.html
Program Chairs: Glenn Gunzelmann, Andrew Howes, Thora Tenbrink, Eddy Davelaar

## Table of Contents

Introduction ..... lvii
Committees ..... lx
Conference Awards ..... lxiv
Invited Presentations ..... lxvii
Tutorials
Tutorial: Recent Advances in Deep Learning ..... 1
Matthew Botvinick, Peter Battaglia
Methods for Reconstructing Causal Networks from Observed Time-Series: Granger-Causality, Transfer Entropy, and Convergent Cross-Mapping ..... 3
Fermin Moscoso del Prado Martin
Dynamic Field Theory: Conceptual Foundations and Applications in Cognitive and DevelopmentalScience5
John Spencer, Vanessa Simmering, Sebastian Schneegans
Workshops
The Fine Art of Conversation ..... 7
Saul Albert, Claude Heath, Patrick Healey
Workshop: Bridging the Gap: Is Logic and Automated Reasoning a Foundation for Human Reasoning? ..... 9
Ulrich Furbach, Steffen Hölldobler, Marco Ragni, Claudia Schon
Citizen Science, Gamification, and Virtual Reality for Cognitive Research ..... 11
Jana Jarecki, Jacob Sherson, Juho Hamari, Julia A. Bopp, Julian Jarecki, Libby Heaney, Sharon Steinemann, Pinja Haikka, Carsten Bergenholtz
Cooperative Social Intelligence: Understanding and Acting with Other ..... 13
Max Kleiman-Weiner, Yibiao Zhao, Josh Tenenbaum
Building Bridges from (Ivory) Towers: Combining Academia and Industry for Cognitive Research ..... 15
Katherine Livins, Jay Martin
The Computational Foundations of Religious Cognition: A Workshop Hosted by the International Association for the Cognitive Science of Religion (IACSR) ..... 17
Ann Taves, Dimitris Xygalatas, John Shaver
Workshop proposal: Deep Learning in Computational Cognitive Science ..... 19
Ilker Yildirim, Josh Tenenbaum
Symposia
Anthropological Contributions to Cognitive Science ..... 21
Andrea Bender, Sieghard Beller, Rita Astuti, Olivier Le Guen, Karenleigh Overmann, Peter Rácz, Fiona Jordan, Mads Solberg
Static and dynamic visual narratives, by brain and by eye ..... 23
Neil Cohn, Tom Foulsham, Tim Smith, Jeff Zacks
Game-XP: Action Games as Cognitive Science Paradigms ..... 25
Wayne Gray, Ray Perez, Martin Butz, Stuart Reeves, Matthew Sangster, Tom Stafford, Fernand Gobet
Educating Spatial Thinking for STEM Success ..... 27Mary Hegarty, David Uttal, Mike Stieff, Tom Lowrie, Stella Vosniadou
Symposium on Problem Solving and Goal-Directed Sequential Activity ..... 29
Pat Langley, Richard Cooper
LUCID science: Advancing learning through human-machine cooperation ..... 31
Timothy Rogers, Charles Kalish
Intuitive Biology and Global Challenges: Applying Theoretical Insights for Public Good ..... 33
Lauren Swiney, Deborah Kelemen, Andrew Shtulman, John Coley, Nicole Betz, Stefaan Blancke
Time in the mind of a child: Perspectives on the development of temporal cognition ..... 35
Katharine Tillman, Florie Monier, Meng Zhang, Jonathan Redshaw, Teresa McCormack
Big Data and Little Learners ..... 37
John Trueswell, Linda Smith, Josh Tenenbaum, Charles Yang
Talks: Papers
Burstiness across multimodal human interaction reveals differences between verbal and non-verbal communication ..... 39
Drew Abney, Rick Dale, Chris Kello, Max Louwerse
Enactive Mechanistic Explanation of Social Cognition ..... 45Ekaterina Abramova, Marc Slors, Iris van Rooij
Human Visual Search as a Deep Reinforcement Learning Solution to a POMDP ..... 51
Aditya Acharya, Xiuli Chen, Christopher Myers, Richard Lewis, Andrew Howes
Executive function and attention predict low-income preschoolers' active category learning ..... 57
Katherine A. Adams, George Kachergis
Simulating behavioural interventions for developmental deficits: When improving strengths produces better outcomes than remediating weaknesses ..... 63
Hala Alireza, Anna Fedor, Michael Thomas
A Bayesian Model of Memory for Text ..... 69
Mark Andrews
Numbers Uniquely Bias Spatial Attention: A Novel Paradigm for Understanding Spatial-Numerical Associations ..... 75
Lauren Aulet, Sami Yousif, Stella Lourenco
Highly Proficient Bilinguals Maintain the Language-Specific Pragmatic Constraints on Pronouns:Evidence from Speech and Gesture81
Zeynep Azar, Ad Backus, Aslı Özyürek
Rise and fall of conflicting intuitions during reasoning ..... 87
Bence Bago, Wim De Neys
The Refugees' Dilemma: not all deontological moral choices are of the same kind ..... 93
Albert Barque-Duran, Emmanuel Pothos
I know what you need to know: Children's developing theory of mind and pedagogical evidence selection ..... 99
Ilona Bass, Elizabeth Bonawitz, Patrick Shafto, Dhaya Ramarajan, Alison Gopnik, Henry Wellman
Didn't know, or didn't show? Preschoolers consider epistemic state and degree of omission when evaluating teachers ..... 105
Ilona Bass, Elizabeth Bonawitz, Hyowon Gweon
Modeling human categorization of natural images using deep feature representations ..... 111
Ruairidh Battleday, Joshua Peterson, Tom Griffiths
Semantic Typology and Parallel Corpora: Something about Indefinite Pronouns ..... 112
Barend Beekhuizen, Julia Watson, Suzanne Stevenson
How Relative is the Relative Frame of Reference? Front and back in Norwegian, Farsi, German, and Japanese ..... 118
Sieghard Beller, Andrea Bender
Quantifying infants' statistical word segmentation: a meta-analysis ..... 124
Alexis Black, Christina Bergmann
Narrowing of the Cone-of-Direct Gaze Through Reinforcement Learning ..... 130
Sam Blakeman, Denis Mareschal
Functionally localized representations contain distributed information: insight from simulations of deep convolutional neural networks ..... 136
Nicholas Blauch, Elissa Aminoff, Michael Tarr
Inferential Role Semantics for Natural Language ..... 142
Peter Blouw, Chris Eliasmith
Split-Second Detection of Cooperativeness from Faces in the Trust Game ..... 148Jean-Francois Bonnefon, Astrid Hopfensitz, Wim De Neys
Diversity in a Contrast Set Increases Generalization from a Single-Item Target ..... 149
David Bosch, David Kalkstein, Tali Kleiman
Causal learning from interventions and dynamics in continuous time ..... 150Neil Bramley, Ralf Mayrhofer, Tobias Gerstenberg, David Lagnado
Harmony in a non-harmonic language: word order learning in French children ..... 156
Guillaume Braquet, Jennifer Culbertson
How can I help? 24- to 48-month-olds provide help specific to the cause of others' failed actions ..... 162
Sophie Bridgers, Sara Altman, Hyowon Gweon
Watching Non-Corresponding Gestures Helps Learners with High Visuospatial Ability to Learn about Movements with Dynamic Visualizations: An fNIRS Study ..... 168Birgit Brucker, Björn de Koning, Ann-Christine Ehlis, David Rosenbaum, Peter Gerjets
Learning on Multi-Touch Devices: The influence of the distance between information in pop-ups and the hands of the users ..... 174
Birgit Brucker, Lara Scatturin, Peter Gerjets
A Flexible Mapping Scheme for Discrete and Dimensional Emotion Representations ..... 180
Sven Buechel, Udo Hahn
Maintenance of Perceptual Information in Speech Perception ..... 186
Wednesday Bushong, T. Florian Jaeger
Task-oriented Bayesian inference in interval timing: People use their prior reproduction experience to calibrate time reproduction ..... 192
Zhenguang Cai, Ruiming Wang, Huihui Liu, Maarten Speekenbrink
The most efficient sequence of study depends on the type of test ..... 198
Paulo Carvalho, Rob Goldstone
Is there an explicit learning bias? Students beliefs, behaviors and learning outcomes. ..... 204
Paulo Carvalho, Elizabeth McLaughlin, Ken Koedinger
A Model of Cognitive Control in the Wisconsin Card Sorting Test: Integrating Schema Theory and Basal Ganglia Function ..... 210
Andrea Caso, Richard P. Cooper
Distributional learning and lexical category acquisition: What makes words easy to categorize? ..... 216Giovanni Cassani, Robert Grimm, Steven Gillis, Walter Daelemans
Knowledge transfer in a probabilistic Language Of Thought ..... 222
Samuel Cheyette, Steve Piantadosi
Cross-situational learning of novel anaphors ..... 228Karen Clothier, Satoshi Nambu, Hajime Ono, Akira Omaki
Expected Utility in Romantic Relationships: Satisfaction as a Cue for Romantic Partnership Dissolution ..... 234
Samantha Cohen, Peter Todd
Analogical gestures foster understanding of causal systems ..... 240Kensy Cooperrider, Dedre Gentner, Susan Goldin-Meadow
Characterizing spatial construction processes: Toward computational tools to understand cognition ..... 246Cathryn S. Cortesa, Jonathan D. Jones, Gregory D. Hager, Sanjeev Khudanpur, Amy L.Shelton, Barbara Landau
Mathematical invariants in people's probabilistic reasoning ..... 252Fintan Costello, Paul Watts
Experts are better than novices when imagining wines, but not odors in general ..... 258
Ilja Croijmans, Laura Speed, Artin Arshamian, Asifa Majid
The elusive odness of or-introduction ..... 259
Nicole Cruz, David Over, Mike Oaksford
A Dynamic Tradeoff Model of Intertemporal Choice ..... 265
Junyi Dai, Timothy Pleskac, Thorsten Pachur
What is Learning? A Definition for Cognitive Science ..... 271
Jim Davies
Re-representation in comparison and similarity ..... 277
Samuel Day, Jennifer Asmuth
Comprehenders Rationally Adapt Semantic Predictions to the Statistics of the Local Environment: a Bayesian Model of Trial-by-Trial N400 Amplitudes ..... 283
Nathaniel Delaney-Busch, Emily Morgan, Ellen Lau, Gina Kuperberg, Gina Kuperberg
The Role of Imagination in Exemplar Generation: The Effects of Conflict and Explanation ..... 289 Tammy Dennis, Daniel Little
A hierarchical Bayesian model of "memory for when" based on experience sampling data ..... 295
Simon Dennis, Hyungwook Yim, Vishnu Sreekumar, Nathan Evans, Paul Garrett, Per Sederberg
Refixations gather new visual information rationally ..... 301
Yunyan Duan, Klinton Bicknell
A rational analysis of curiosity ..... 307
Rachit Dubey, Tom Griffiths
Refining the distributional hypothesis: A role for time and context in semantic representation ..... 313
Melody Dye, Michael Jones, Daniel Yarlett, Michael Ramscar
Cute Little Puppies and Nice Cold Beers: An Information Theoretic Analysis of Prenominal Adjectives ..... 319
Melody Dye, Petar Milin, Richard Futrell, Michael Ramscar
Vanishing the mirror effect: The influence of prior history $\mathcal{E}$ list composition on recognition memory ..... 325
Melody Dye, Michael Jones, Richard Shiffrin
Creating words from iterated imitation ..... 331
Pierce Edmiston, Marcus Perlman, Gary Lupyan
A Model of Event Knowledge ..... 337
Jeffrey Elman, Ken McRae
The role of intentionality in causal attribution is culturally mediated: evidence from Chinese, Mayan, and Spanish populations ..... 343
Stephanie Evers, Erika Bellingham, Katharine Donelson, Jing Du, José Antonio Jódar Sánchez, Fuyin Li, Jürgen Bohnemeyer
Explaining Enculturated Cognition ..... 349
Regina E. Fabry
A Cognitive-Pharmacokinetic Computational Model of the Effect of Toluene on Performance ..... 355
Christopher Fisher, Christopher Myers, Reem Hassan, Christopher Stevens, Charles Hack, Jeffery Gearhart, Glenn Gunzelmann
Thinking about the future: The role of spatial metaphors for time ..... 361
Stephen Flusberg, Teenie Matlock, Paul Thibodeau
Early Colour Word Learning in British Infants ..... 367Samuel Forbes, Kim Plunkett
Word Identification Under Multimodal Uncertainty ..... 373Abdellah Fourtassi, Michael Frank
Could both be right? Children's and adults' sensitivity to subjectivity in language ..... 379
Ruthe Foushee, Mahesh Srinivasan
Word Embedding Distance Does not Predict Word Reading Time ..... 385
Stefan Frank
Is Conflict Detection in Reasoning Domain General? ..... 391
Darren Frey, Wim De Neys
Sequential Effects in Prediction ..... 397
Dinis Gökaydin, Anastasia Ejova
Empirical tests of large-scale collaborative recall ..... 403
Monica Gates, Jordan Suchow, Thomas Griffiths
Faulty Towers: A hypothetical simulation model of physical support ..... 409
Tobias Gerstenberg, Liang Zhou, Kevin Smith, Joshua Tenenbaum
Gaze Shifts between Text and Illustrations are Negatively Related to Reading Fluency in Beginning Readers ..... 415
Karrie E. Godwin, Cassondra M. Eng, Anna V. Fisher
A Cognitive Model of Strategic Deliberation and Decision Making ..... 421
Russell Golman, Sudeep Bhatia
How Can Memory-Augmented Neural Networks Pass a False-Belief Task? ..... 427
Erin Grant, Aida Nematzadeh, Thomas L. Griffiths
Evidence for a facilitatory effect of multi-word units on child word learning ..... 433
Robert Grimm, Giovanni Cassani, Steven Gillis, Walter Daelemans
The nature of quantities influences the representation of arithmetic problems: evidence from drawings and solving procedures in children and adults ..... 439
Hippolyte Gros, Jean-Pierre Thibaut, Emmanuel Sander
Does Mandarin Spatial Metaphor for Time Influence Chinese Deaf Signers' Spatio-Temporal Reasoning? ..... 445Yan Gu, Yeqiu Zheng, Marc Swerts
Language and Spatial Memory in Japanese and English ..... 451Harmen Gudde, Kenny Coventry
A Computational Model for the Dynamical Learning of Event Taxonomies ..... 452
Christian Gumbsch, Sebastian Otte, Martin Butz
Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of noveltasks458Hyowon Gweon, Mika Asaba, Grace Bennett-Pierre
Categorization, Information Selection and Stimulus Uncertainty ..... 464
David Halpern, Todd Gureckis
Perceptual contrast and response assimilation in sequential categorization without feedback ..... 470 James Hampton
Radical Embodied Cognition, Affordances, and the (Hard) Problem of Consciousness ..... 476
Valerie Hardcastle
Convention-formation in iterated reference games ..... 482
Robert X.d. Hawkins, Mike Frank, Noah Goodman
Sampling frames, Bayesian inference and inductive reasoning ..... 488
Brett Hayes, Stephanie Banner, Daniel Navarro
The adaptive evolution of early human symbolic behavior ..... 494
Katrin Heimann, Riccardo Fusaroli, Sergio Gonzalez de La Higuera Rojo, Niels Nørkjær Johannsen, Felix Riede, Nicolas Fay, Marlize Lombard, Kristian Tylén
Children's social referencing reflects sensitivity to graded uncertainty ..... 495
Emily Hembacher, Benjamin demayo, Michael Frank
Investigating the Explore/Exploit Trade-off in Adult Causal Inferences ..... 501
Erik Herbst, Christopher Lucas, Daphna Buchsbaum
Modeling transfer of high-order uncertain information ..... 507
Michele Herbstritt, Michael Franke
A Priming Model of Category-based Feature Inference ..... 513
Laura Hiatt
Quantifying the impact of active choice in word learning ..... 519
Shohei Hidaka, Takuma Torii, George Kachergis
Reference Systems in Spatial Memory for Vertical Locations ..... 525
Thomas Hinterecker, Caroline Leroy, Maximilian E. Kirschhock, Mintao Zhao, Martin Butz, Heinrich H. Bülthoff, Tobias Meilinger
Teaching by Intervention: Working Backwards, Undoing Mistakes, or Correcting Mistakes? ..... 526
Mark Ho, Michael Littman, Joseph Austerweil
Computational and behavioral investigations of the SOB-CS removal mechanism in working memory ..... 532Violette Hoareau, Sophie Portrat, Klaus Oberauer, Benoit Lemaire, Gaën Plancher,Stephan Lewandowsky
Spatial Cognition and Science: The Role of Intrinsic and Extrinsic Spatial Skills from Seven to Eleven Years ..... 538
Alex Hodgkiss, Katie Gilligan, Michael Thomas, Andrew Tolmie, Emily Farran
Moving together: in the body or the mind? ..... 544Lilla Hodossy, Jorina von Zimmermann, Daniel Richardson
Modeling Sources of Uncertainty in Spoken Word Learning ..... 550
Matthias Hofer, Roger Levy
Knowledge of Cross-Linguistic Semantic Diversity Reduces Essentialist Beliefs about Categories ..... 556Kevin Holmes, Erin Luby, Sarah Husney
Decisions based on verbal probabilities: Decision bias or decision by belief sampling? ..... 557
Hidehito Honda, Toshihiko Matsuka, Kazuhiro Ueda
Semantic diversity, frequency and learning to read: A mini-mega study with children ..... 563
Yaling Hsiao, Kate Nation
Testing Statistical Learning Implicitly: A Novel Chunk-based Measure of Statistical Learning ..... 564
Erin S. Isbilen, Stewart M. McCauley, Evan Kidd, Morten H. Christiansen
Minimal covariation data support future one-shot inferences about unobservable properties of novel agents ..... 570
Julian Jara-Ettinger, Hyowon Gweon
The Semantics and Pragmatics of Logical Connectives: Adults' and Children's Interpretations of And and Or in a Guessing Game ..... 576
Masoud Jasbi, Michael Frank
Maintaining Credibility When Communicating Uncertainty: The Role of Communication Format ..... 582
Sarah Jenkins, Adam Harris, Murray Lark
A modeling link between cognitive and biological homeostasis ..... 588
Gavin Jenkins, Jordan Barnes, Paul Tupper, Mark Blair
Viewers' Sensitivity to Abstract Event Structure ..... 594
Yue Ji, Anna Papafragou
Promoting Children's Relational Understanding of Equivalence ..... 600
Kristen Johannes, Jodi Davenport, Yvonne Kao, Caroline Byrd Hornburg, Nicole McNeil
Opponent Uses of Simplicity and Complexity in Causal Explanation ..... 606
Samuel Johnson, J. J. Valenti, Frank Keil
Principles Used to Evaluate Mathematical Explanations ..... 612Samuel Johnson, Angie Johnston, Marissa Koven, Frank Keil
Statistical and Mechanistic Information in Evaluating Causal Claims ..... 618
Samuel Johnson, Frank Keil
Touch Screen Text Entry as Cognitively Bounded Rationality ..... 624
Jussi Jokinen
Using Measurement Models to Understand Eyewitness Identification ..... 625
Matthew Kaesler, John Dunn, Carolyn Semmler
A Biologically Constrained Model of Semantic Memory Search ..... 631
Ivana Kajic, Jan Gosmann, Brent Komer, Ryan W. Orr, Terrence Stewart, Chris Eliasmith
Mechanisms of overharvesting in patch foraging ..... 637
Gary Kane, Aaron Bornstein, Robert Wilson, Amitai Shenhav, Nathaniel Daw, Jonathan Cohen
Language-users choose short words in predictive contexts in an artificial language task ..... 643
Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, Simon Kirby
Preparatory Effects of Problem Posing on Learning from Instruction ..... 649
Manu Kapur
Comparing Individual and Collaborative Problem Solving in Environmental Search ..... 650
Franziska Keilmann, Stephan de La Rosa, Stephan Schwan, Ulrike Cress, Betty J Mohler, Heinrich H Bülthoff, Tobias Meilinger
Iterated Teaching Can Optimise Language Functionality ..... 651Vera Kempe, Kamil Cichon, Nicolas Gauvrit, Monica Tamariz
Dynamic Effects of Conceptual Combination on Semantic Network Structure ..... 657
Yoed Kenett, Sharon Thompson-Schill
The provenance of modal inference ..... 663
Sangeet Khemlani, Thomas Hinterecker, Philip Johnson-Laird
Numerical and Non-numerical Magnitude Estimation ..... 669
Dan Kim, John Opfer
Variables Involved in Selective Sustained Attention Development: Advances in Measurement ..... 670
Jaeah Kim, Anna Vande Velde, Erik Thiessen, Anna Fisher
Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why? ..... 676
Max Kleiman-Weiner, Alex Shaw, Josh Tenenbaum
Explaining Guides Learners Towards Perfect Patterns, Not Perfect Prediction ..... 682
Elizabeth Kon, Tania Lombrozo
Selective Information Sampling and the In-Group Heterogeneity Effect ..... 688
Elizaveta Konovalova, Gael Le Mens
Exploring the decision dynamics of risky intertemporal choice ..... 694
Emmanouil Konstantinidis, Don van Ravenzwaaij, Ben R. Newell
Consistent Probabilistic Simulation Underlying Human Judgment in Substance Dynamics ..... 700
James Kubricht, Yixin Zhu, Chenfanfu Jiang, Demetri Terzopoulos, Song-Chun Zhu, Hongjing Lu
Synchronization Assessment for Collective Behavior ..... 706Maurici A. López-Felip, Tehran J. Davis, Till D. Frank
More Siblings Means Lower Input Quality in Early Language Development ..... 711
Catherine Laing, Elika Bergelson
Uncovering visual priors in spatial memory using serial reproduction ..... 712
Thomas Langlois, Nori Jacoby, Jordan Suchow, Thomas Griffiths
A core-affect model of decision making in simple and complex tasks ..... 718Othalia Larue, Alexander Hough, Ion Juvina
Generic and Universal Generalisations: Contextualising the 'Generic Overgeneralisation' Effect ..... 724
Dimitra Lazaridou-Chatzigoga, Napoleon Katsos, Linnaea Stockall
Children's semantic and world knowledge overrides fictional information during anticipatory linguistic processing ..... 730
Ruth Lee, Craig Chambers, Falk Huettig, Patricia Ganea
The impact of descriptions and incentives on the simultaneous underweighting and overestimation of rare events ..... 736
Garston Liang, Emmanouil Konstantinidis, Aba Szollosi, Christopher Donkin, Ben R. Newell
An automatic method for discovering rational heuristics for risky choice ..... 742
Falk Lieder, Paul Krueger, Tom Griffiths
Sequential Effects in the Garner Tasks ..... 748
Deborah Lin, Daniel Little
Error-based learning: A mechanism for linking verbs to syntax ..... 754Yi Lin, Cynthia Fisher
Repeated Interactions Can Lead to More Iconic Signals ..... 760
Hannah Little, Marcus Perlman, Kerem Eryilmaz
What's worth the effort: Ten-month-old infants infer the value of goals from the costs of actions ..... 766
Shari Liu, Tomer Ullman, Josh Tenenbaum, Elizabeth Spelke
Why Does Higher Working Memory Capacity Help You Learn? ..... 767
Kevin Lloyd, Adam Sanborn, David Leslie, Stephan Lewandowsky
Social Network Limits Language Complexity ..... 773
Matt Lou-Magnuson, Luca Onnis
Learning induced illusions: Statistical learning creates false memories ..... 774
Yu Luo, Jiaying Zhao
An information-seeking account of eye movements during spoken and signed language comprehension ..... 780Kyle MacDonald, Aviva Blonder, Virginia Marchman, Anne Fernald, Michael Frank
Preschoolers appropriately allocate roles based on relative ability in a cooperative interaction ..... 786
Rachel Magid, Mary Depascale, Laura Schulz
Intuitive psychophysics: Children's exploratory play quantitatively tracks the discriminability of alternative hypotheses ..... 792
Rachel Magid, Max Siegel, Josh Tenenbaum, Laura Schulz
Leveraging Response Consistency within Individuals to Improve Group Accuracy for Rank-Ordering Problems ..... 793
Brent Miller, Mark Steyvers
A Rational Constructivist Account of the Characteristic-to-Defining Shift ..... 799
Francis Mollica, Shirlene Wade, Steve Piantadosi
An incremental information-theoretic buffer supports sentence processing ..... 805
Francis Mollica, Steve Piantadosi
Different processes for reading words learned before and after onset of literacy ..... 811
Padraic Monaghan, Ya-Ning Chang, Stephen Welbourne
Multiple variable cues in the environment promote accurate and robust word learning ..... 817
Padraic Monaghan, James Brand, Rebecca Frost, Gemma Taylor
The ecological rationality of children's option generation and decision making ..... 823Lisa Musculus, Azzurra Ruggeri, Markus Raab, Babett Lobinger
Multitasking Capability Versus Learning Efficiency in Neural Network Architectures ..... 829
Sebastian Musslick, Andrew Saxe, Kayhan Ozcimder, Biswadip Dey, Greg Henselman,Jonathan Cohen
Analogical Inferences in Causal Systems ..... 835Matthew Myers, Dedre Gentner
Individual Differences in Transfer Mediated by Conceptual Priming ..... 841
Jairo A. Navarrete
When extremists win: On the behavior of iterated learning chains when priors are heterogeneous ..... 847
Daniel Navarro, Amy Perfors, Arthur Kary, Scott Brown, Chris Donkin
Calculating Probabilities Simplifies Word Learning ..... 853Aida Nematzadeh, Barend Beekhuizen, Shanshan Huang, Suzanne Stevenson
Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words ..... 859
Aida Nematzadeh, Stephan C. Meylan, Thomas L. Griffiths
Listeners integrate speech, gesture, and discourse structure to interpret the temporal structure of complex events ..... 865
Andrea Nishimi, Esther Walker, Benjamin Bergen, Tyler Marghetis
Dynamics of Affordance Actualization ..... 871Patric Nordbeck, Laura Soter, Rachel Kallen, Anthony Chemero, Michael Richardson
Broadening the Scope of Recognition Memory ..... 877
Justin Olds, Julian Marewski
A Computational Logic Approach to Human Syllogistic Reasoning ..... 883
Ana Oliveira da Costa, Emmanuelle-Anna Dietz Saldanha, Steffen Hölldobler, Marco Ragni
Speakers' gestures predict the meaning and perception of iconicity in signs ..... 889
Gerardo Ortega, Annika Schiefner, Asli Ozyurek
A Formal Approach to Modeling the Cost of Cognitive Control ..... 895
Kayhan Ozcimder, Biswadip Dey, Sebastian Musslick, Giovanni Petri, Nesreen K. Ahmed, Theodore L. Willke, Jonathan D. Cohen
Optimization of American English, Spanish, and Mandarin Chinese over time for efficient communication ..... 901
John Pate
Relational Concept Learning via Guided Interactive Discovery ..... 907
John Patterson, David Landy, Kenneth Kurtz
The Use of Iconic Words in Early Child-Parent Interactions ..... 913
Marcus Perlman, Riccardo Fusaroli, Deborah Fein, Letitia Naigles
Evidence for the size principle in semantic and perceptual domains ..... 919
Joshua Peterson, Thomas Griffiths
Modeling the Ellsberg Paradox by Argument Strength ..... 925
Niki Pfeifer, Hanna Pankka
Causation and norms of proper functioning: Counterfactuals are (still) relevant ..... 931
Jonathan Phillips, Jonathan Kominsky
Assessing the Linguistic Productivity of Unsupervised Deep Neural Networks ..... 937
Lawrence Phillips, Nathan Hodas
Opinion Cascades and Echo-Chambers in Online Networks: A Proof of Concept Agent-Based Model ..... 943
Toby D. Pilditch
Make-or-break: chasing risky goals or settling for safe rewards? ..... 949
Pantelis Pipergias Analytis, Charley Wu, Alexandros Gelastopoulos
A computational decomposition of task-irrelevant perceptual learning ..... 955Angelo Pirrone, Tom Stafford
Inferential Pitfalls in Decoding Neural Representations ..... 961
Vencislav Popov, Markus Ostarek, Caitlin Tenison
The Relational Luring Effect: False Recognition via Relational Similarity ..... 967
Vencislav Popov, Penka Hristova
Target-to-distractor similarity can help visual search performance ..... 968
Vencislav Popov, Lynne Reder
Timing Time: Why Early Vision is Cognitively Impenetrable ..... 974
Athanassios Raftopoulos
The Wason Selection task: A Meta-Analysis ..... 980
Marco Ragni, Ilir Kola, Phil Johnson-Laird
Mental Algorithms in the Historical Emergence of Word Meanings ..... 986
Christian Ramiro, Barbara Malt, Mahesh Srinivasan, Yang Xu
A cognitive analysis of deception without lying ..... 992
Keith Ransom, Wouter Voorspoels, Amy Perfors, Daniel Navarro
Connecting stimulus-driven attention to the properties of infant-directed speech - Is exaggerated intonation also more surprising? ..... 998
Okko Rasanen, Sofoklis Kakouros, Melanie Soderstrom
When does a 'visual proof by induction' serve a proof-like function in mathematics? ..... 1004
Josephine Relaford-Doyle, Rafael Nunez
Interpreting Asymmetric Perception in Speech Perception with Bayesian Inference ..... 1010
Jie Ren, Joseph Austerweil
Quantitative Models of Human-Human Conversational Grounding Processes ..... 1016
Clayton D. Rothwell, Valerie L. Shalin, Griffin D. Romigh
Comprehenders Model the Nature of Noise in the Environment ..... 1022Rachel Ryskin, Richard Futrell, Edward Gibson
Is the strength of regularisation behaviour uniform across linguistic levels? ..... 1023
Carmen Saldana, Kenny Smith, Simon Kirby, Jennifer Culbertson
Converting Cascade-Correlation Neural Nets into Probabilistic Generative Models ..... 1029
Ardavan Salehi Nobandegani, Thomas R. Shultz
Mental Representations and Computational Modeling of Context-Specific Human Norm Systems ..... 1035
Vasanth Sarathy, Matthias Scheutz, Yoed Kenett, Mowafak Allaham, Joseph Austerweil, Bertram Malle
Attractor Dynamics in Delay Discounting: A Call for Complexity ..... 1041
Martin Schoemann, Stefan Scherbaum
Strategic exploration in human adaptive control ..... 1047
Eric Schulz, Edgar Klenske, Neil Bramley, Maarten Speekenbrink
How Does Instance-Based Inference About Event Frequencies Develop? An Analysis with a Computational Process Model ..... 1053
Christin Schulze, Thorsten Pachur, Ralph Hertwig
Prior Expectations in Linguistic Learning: A Stochastic Model of Individual Differences ..... 1059
R. Alexander Schumacher, Janet Pierrehumbert
Effects of Grammatical Gender on Object Description ..... 1060Arturs Semenuks, Webb Phillips, Ioana Dalca, Cora Kim, Lera Boroditsky
Inferring Human Interaction from Motion Trajectories in Aerial Videos ..... 1066
Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, Song-Chun Zhu
Children's spontaneous comparisons from 26 to 58 months predict performance in verbal and non-verbal analogy tests in 6th grade ..... 1072
Catriona Silvey, Dedre Gentner, Lindsey Richland, Susan Goldin-Meadow
Is Play Better Than Direct Instruction? Learning About Causal Systems through Play ..... 1078
Zi L. Sim, Kuldeep K. Mahal, Fei Xu
Conditionals, Individual Variation, and the Scorekeeping Task ..... 1084Niels Skovgaard-Olsen, David Kellen, Ulrike Hahn, Karl Christoph Klauer
A case for systematic sound symbolism in pragmatics: The role of the first phoneme in question prediction in context ..... 1090
Anita Slonimska, Sean Roberts
Learning to See People like People: Predicting Social Perceptions of Faces ..... 1096
Amanda Song, Li Linjie, Chad Atalla, Gary Cottrell
A Rational Approach to Stereotype Change ..... 1102
Jake Spicer, Adam Sanborn
Population size, learning, and innovation determine linguistic complexity ..... 1108
Matthew Spike
A rational analysis of marketing strategies ..... 1114
Nisheeth Srivastava, Ed Vul
Spatial language promotes cross-domain associations in early childhood ..... 1120Ariel Starr, Mahesh Srinivasan
Preemption in Singular Causation Judgments: A Computational Model ..... 1126Simon Stephan, Michael Waldmann
Marbles in Inaction: Counterfactual Simulation and Causation by Omission ..... 1132Simon Stephan, Pascale Willemsen, Tobias Gerstenberg
A Two-Step Signal Detection Model of Belief Bias ..... 1138
Rachel Stephens, John Dunn, Brett Hayes
Rational use of prosody predicts projection in manner adverb utterances ..... 1144Jon Stevens, Marie-Catherine de Marneffe, Shari Speer, Judith Tonhauser
A Common Neural Component for Finger Gnosis and Magnitude Comparison ..... 1150
Terrence Stewart, Marcie Penner-Wilger, Rylan Waring, Michael Anderson
Flexible integration of a navigable, clustered environment ..... 1156Marianne Strickrodt, Heinrich H. Bülthoff, Tobias Meilinger
The Semantic Spaces of Child-Directed Speech, Child Speech and Adult-directed Speech: A Manifold Perspective ..... 1157
Hao Sun, John Pate
¿From Abstract to Concrete? Evidence for designing learning platforms that adapt to user's proficiencies. ..... 1163
Michael I. Swart, Sorachai Kornkasem, Nirmaliz Colon-Acosta, Amy Hachigian, Jonathan M. Vitale, John B. Black
Supporting Low-Performing Students by Manipulating Self-efficacy in Digital Tutees ..... 1169
Betty Tärning, Magnus Haake, Agneta Gulz
The impact of practice frequency on learning and retention ..... 1175
Caitlin Tenison, John Anderson
Warm (for winter): Comparison class understanding in vague language ..... 1181
Michael Henry Tessler, Michael Lopez-Brau, Noah Goodman
Neural and computational arguments for memory as a compressed supported timeline ..... 1187
Zoran Tiganj, Karthik Shankar, Marc Howard
Picturing time: Children's preferences for visual representations of events ..... 1193
Katharine Tillman, Eren Fukuda, David Barner
Computational Exploration of Lexical Development in Down Syndrome ..... 1199Angel Tovar, Gert Westermann
The temporal dynamics of base rate neglect: People may not be intuitive statisticians after all ... ..... 1205Eoin Travers, Jonathan J. Rolison, Aidan Feeney
Promoting Spontaneous Analogical Transfer by Idealizing Target Representations ..... 1206Máximo Trench, Lucía Micaela Tavernini, Robert L. Goldstone
It's all in your head: Effects of expertise on real-time access to knowledge during written sentence processing ..... 1212
Melissa Troyer, Marta Kutas
Inhibitory Control Supports Referential Context Use in Language Production and Comprehension ..... 1218
Alison Trude, Nazbanou Nozari
A Dynamic Process Model for Predicting Workload in an Air Traffic Controller Task ..... 1224
Martina Truschzinski, Maria Wirzberger
The Dynamics of Selective Integration during Rapid Experiential Decisions ..... 1230
Konstantinos Tsetsos
Cake or Hat? Words Change How Young Children Process Visual Objects ..... 1236
Catarina Vales, Linda Smith
Looking for the Cat and Seeing the Dog: Using Visual Search to Study Semantic Knowledge in Children ..... 1242
Catarina Vales, Layla Unger, Anna Fisher
Interactivity and Ego Depletion in Insight Problem Solving ..... 1248
Frederic Vallee-Tourangeau
A computational model for decision tree search ..... 1254Bas van Opheusden, Gianni Galbiati, Zahy Bnaya, Yunqi Li, Wei Ji Ma
Approximations of Predictive Entropy Correlate with Reading Times ..... 1260
Marten van Schijndel, William Schuler
The Development of Structural Thinking about Social Categories ..... 1266
Nadya Vasilyeva, Alison Gopnik, Tania Lombrozo
The statistical significance filter leads to overconfident expectations of replicability ..... 1272
Shravan Vasishth, Andrew Gelman
Modelling dependency completion in sentence comprehension as a Bayesian hierarchical mixture
process: A case study involving Chinese relative clauses ..... 1278
Shravan Vasishth, Nicolas Chopin, Robin Ryder, Bruno Nicenboim
Interpreting actions by attributing compositional desires ..... 1284
Joey Velez-Ginorio, Max H. Siegel, Joshua B. Tenenbaum, Julian Jara-Ettinger
Non-Symbolic Exact Quantity Representation in a Language Impaired Population ..... 1290
John Verbos, Sarah Wallace, Alexander Kranjec
Audiovisual integration is affected by performing a task jointly ..... 1296
Basil Wahn, Ashima Keshava, Scott Sinnett, Alan Kingstone, Peter König
More than meets the eye: Early relational reasoning cannot be reduced to perceptual heuristics ..... 1302
Caren Walker, Alison Gopnik
Simultaneous acquisition of vocabulary and grammar in an artificial language learning task ..... 1307
Neil Walker, Christine Schoetensack, Padraic Monaghan, Patrick Rebuschat
Please Explain: Radical Enactivism's Explanatory Debt ..... 1313
Lachlan Walmsley
Learning to reinforcement learn ..... 1319
Jane Wang, Zeb Kurth-Nelson, Hubert Soyer, Joel Leibo, Dhruva Tirumala, Remi Munos, Charles Blundell, Dharshan Kumaran, Matt Botvinick
The Learning of Subordinate Word Meanings ..... 1320Hao Wang, Lila Gleitman, John Trueswell
Anticipatory Synchronization in Artificial Agents ..... 1321Auriel Washburn, Rachel Kallen, Maurice Lamb, Nigel Stepp, Kevin Shockley, MichaelRichardson
Structure Learning in Motor Control: A Deep Reinforcement Learning Model ..... 1327
Ari Weinstein, Matthew Botvinick
Children's intuitions about the structure of mental life ..... 1333Kara Weisman, Carol Dweck, Ellen Markman
The Effect of expertise and biscriptalism on letter perception: The complexity benefit ..... 1339
Robert Wiley, Brenda Rapp
Examining Multiscale Movement Coordination in Collaborative Problem Solving ..... 1345
Travis J. Wiltshire, Sune Vork Steffensen
A Computational Model for Constructing Preferences for Multiple Choice Options ..... 1351
Lena M. Wollschlaeger, Adele Diederich
Mapping the unknown: The spatially correlated multi-armed bandit ..... 1357
Charley M. Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D. Nelson, Björn Meder
What do you really think? Children's ability to infer others' desires when emotional expressions change between social and nonsocial contexts ..... 1363
Yang Wu, Laura Schulz
Discovering Multicausality in the Development of Coordinated Behavior ..... 1369
Tian Linger Xu, Drew H. Abney, Chen Yu
Unifying recommendation and active learning for human-algorithm interactions ..... 1375
Scott Cheng-Hsin Yang, Jake Alden Whritner, Olfa Nasraoui, Patrick Shafto
A non-parametric Bayesian prior for causal inference of auditory streaming ..... 1381
Tim Yates, Nathanael Larigaldie, Ulrik Beierholm
Perceived similarity mediates violations of independence in probabilistic judgments ..... 1387
James Yearsley, Daier Yuan, Jennifer Trueblood
Causal and compositional generative models in online perception ..... 1393Ilker Yildirim, Michael Janner, Mario Belledonne, Christian Wallraven, Winrich Freiwald,Josh Tenenbaum
Individual Differences in Gaze Dynamics in Risky Decision-making ..... 1394Siyuan Yin, Jennifer Trueblood
Novel Evidence for the Bilingual Advantage: Effects of Language Control on Executive Function in Balanced and Unbalanced Dual-Language Users ..... 1400W. Quin Yow, Adam C. Oei, Xiaoqian Li
Inconvenient samples: Modeling the effects of non-consent by coupling observational andexperimental results1406Yue Yu, Elizabeth Bonawitz, Patrick Shafto
Seeing Is Not Enough for Sustained Visual Attention ..... 1412Lei Yuan, Tian Linger Xu, Chen Yu, Linda Smith
Mindshaping the world can make mindreading tractable: Bridging the gap between philosophy and computational complexity analysis ..... 1418
Andrea Zeppi, Mark Blokpoel
Using mouse-tracking data to visualise decision landscapes ..... 1424
Arkady Zgonnikov, Andrea Aleni, Petri Piiroinen, Denis O’Hora
Insomniacs Misidentify Angry Faces as Fearful Faces Because of Missing the Eyes: an Eye-Tracking Study ..... 1430
Jinxiao Zhang, Antoni B. Chan, Esther E.y. Lau, Janet H. Hsiao
Low Dimensional Representations in Multi-Cue Judgement ..... 1436
Joyce Wenjia Zhao, Sudeep Bhatia, Clintin Davis-Stober
Talks: Publication-Based
Using single unit recordings in PDP and localist models to better understand how knowledge is coded in the cortex ..... 1442
Jeffrey Bowers
Interoception: The Forgotten Modality in Perceptual Grounding of Concepts ..... 1444
Louise Connell, Dermot Lynott, James Carney
Tracking Meaning Change Over Time: A Dynamic Field Theory Model ..... 1445
Kenny Coventry, Harmen Gudde, Holger Schultheis
Converging Evidence for Abstract Phonological Knowledge in Speech Processing ..... 1447
Anne Cutler
Analogy and Episodic Memory to Support Domain Learning in a Cognitive Architecture: An Exploration ..... 1449
Kenneth Forbus
Non-syntactic Processing Explains Cortical Entrainment During Speech Perception ..... 1451
Stefan Frank, Jin-Biao Yang
Comparative analysis of visual category learning ..... 1453
John Freeman, Matthew Broschard, Jangjin Kim, Leyre Castro, Edward Wasserman, Vladimir Sloutsky
Cognitive mechanisms for imitation and the detection of imitation in human dyadic interactions ..... 1455
Antonia Hamilton
Deep Networks as Models of Human and Animal Categorization ..... 1457Bradley Love, Olivia Guest, Piotr Slomka, Victor Navarro, Edward Wasserman
A Unified Model of Entropy and the Value of Information ..... 1459
Jonathan D. Nelson, Vincenzo Crupi, Björn Meder, Gustavo Cevolani, Katya Tentori
What the Shape and Material Biases Can Tell Us About Object Recognition ..... 1461
Lynn Perry
The role of learning mechanisms in understanding spoken words ..... 1463
Jennifer Rodd, Rebecca Gilbert, Hannah Betts
Far Transfer: Does it Exist? ..... 1465
Giovanni Sala, Fernand Gobet
Scientific Sensemaking: A Critical Resource for Science Learning in School ..... 1467
Christian Schunn
Unsupervised Learning in an Animal Model ..... 1469
Edward Wasserman, Leyre Castro, Vladimir Sloutsky
Posters: Papers
The effects of gesture restriction on spatial language in young and elderly adults ..... 1471
Demet Özer, Merve Tansan, Ege Ekin Özer, Katsiaryna Malykhina, Anjan Chatterjee, Tilbe Göksun
The Effects of Duration Words and Spatial-Temporal Metaphors on Perceived Duration ..... 1477
Demet Özer, Alexander Kranjec, Fuat Balcı, Tilbe Göksun
Action and actor gaze mismatch effects during spoken sentence processing ..... 1483
Dato Abashidze, Pia Knoeferle
It's Time: Quantifying the Relevant Timescales for Joint Attention ..... 1489
Drew Abney, Linda Smith, Chen Yu
The Role of Letter Frequency on Eye Movements in Sentential Pseudoword Reading ..... 1495
Cengiz Acartürk, Özkan Kılıç, Bilal Kırkıcı, Burcu Can, Ayșegül Özkan
Children's Familiarity Preference in Self-directed Study Improves Recognition Memory ..... 1501
Katherine Adams, George Kachergis, Doug Markant
From Words to Sentences $\underbrace{}_{\text {Back: Characterizing Context-dependent Meaning Representations }}$ in the Brain ..... 1507
Nora E Aguirre-Celis, Manuel Valenzuela-Rendon, Risto Miikkulainen
The Influence of Pop-Culture on Misattribution of Memory ..... 1513
Daljit Ahluwalia, Pernille Hemmer, Kimele Persaud
A Computational Model for Reasoning About the Paper Folding Task Using Visual Mental Images ..... 1519
James Ainooson, Maithilee Kunda
Legal HARKing: theoretical grounding in interaction research ..... 1525
Saul Albert, J.p. De Ruiter
Segmentation as Retention and Recognition: the RBR model ..... 1531
Raquel G. Alhama, Willem Zuidema
Phonological features in the bilingual lexicon: Insights from tonal accent in Swedish ..... 1537Nadja Althaus, Allison Wetterlin, Aditi Lahiri
Semantic Networks Generated from Early Linguistic Input ..... 1538Andrei Amatuni, Elika Bergelson
Analogical Abstraction in Three-Month-Olds ..... 1544
Erin Anderson, Yin-Juei Chang, Susan Hespos, Dedre Gentner
A Preliminary P-Curve Meta-Analysis of Learned Categorical Perception Research ..... 1550
Janet Andrews, Joshua de Leeuw, Calais Larson, Xiaoqing Xu
Reading Skill Test to Diagnose Basic Language Skills in Comparison to Machines ..... 1556
Noriko Arai, Naoya Todo, Teiko Arai, Kyosuke Bunji, Shingo Sugawara, Miwa Inuzuka, Takuya Matsuzaki, Koken Ozaki
Perception Meets Examination: Studying Deceptive Behaviors in VR ..... 1562
Carla Aravena, Mark Vo, Tao Gao, Takaaki Shiratori, Lap-Fai Yu
Against the group actor assumption in joint action research ..... 1568
Ed Baggs, Ed Baggs
Towards Automated Classification of Emotional Facial Expressions ..... 1574
Lewis Baker, Vanessa Lobue, Elizabeth Bonawitz, Patrick Shafto
Reasons and the "Motivated Reasoning Effect" ..... 1580
Cristina Ballarini, Steven Sloman
Belief Updating and Argument Evaluation ..... 1586
Megan D. Bardolph, Seana Coulson
The impact of the Digital Age in Moral Judgments ..... 1592Albert Barque-Duran, Emmanuel Pothos, James Yearsley, James Hampton
Exploring Functions of Working Memory Related to Fluid Intelligence: Coordination and Relational Integration ..... 1598Joel Edward Bateman, Damian Patrick Birney, Vanessa Loh
How Order of Label Presentation Impacts Semantic Processing: an ERP Study ..... 1604
Jelena Batinic, Andrej Savic, Vanja Kovic
The Relationship Between Executive Functions and Science Achievement ..... 1610
Drake Bauer, Sashank Varma, Keisha Varma, Martin Van Boekel, Alyssa Worley, Jean-Baptiste Quillien, Tayler Loiselle, Purav Patel
Contrasts in reasoning about omissions ..... 1611Paul Bello, Christina Wasylyshyn, Gordon Briggs, Sangeet Khemlani
Representing time in terms of space: Directions of mental timelines in Norwegian ..... 1617Andrea Bender, Kristin Sjåfjell, Annelie Rothe-Wulf, Sieghard Beller
A Bayesian model of knowledge and metacognitive control: Applications to opt-in tasks ..... 1623Stephen Bennett, Aaron Benjamin, Mark Steyvers
Modality Switch Effects Emerge Early and Increase throughout Conceptual Processing: Evidence from ERPs ..... 1629
Pablo Bernabeu, Roel Willems, Max Louwerse
Object Representation in Multiattribute Choice ..... 1635Sudeep Bhatia, Neil Stewart
The Interactive Shaping of Social Learning in Transmission Chains ..... 1641Lucas Bietti, Adrian Bangerter, Eric Mayor
Language Modality Affects Responses in Left IFG during Processing of Semantically Ambiguous Sentences ..... 1647
Lena M. Blott, Jennifer M. Rodd, Jane E. Warren
Generalized Representation of Syntactic Structures ..... 1648
Reihane Boghrati, Kate Johnson, Morteza Dehghani
Context reduces coercion costs - Evidence from eyetracking during reading ..... 1654
Oliver Bott
The Structure of Goal Systems Predicts Human Performance ..... 1660
David Bourgin, Falk Lieder, Daniel Reichman, Nimrod Talmon, Tom Griffiths
The Effects of Autonomy on Emotions and Learning in Game-Based Learning Environments ..... 1666
Amanda Bradbury, Michelle Taub, Roger Azevedo
A Computational Model of the Role of Attention in Subitizing and Enumeration ..... 1672
Gordon Briggs, Will Bridewell, Paul Bello
Effects of transmission perturbation in the cultural evolution of language ..... 1678
Thomas Brochhagen, Michael Franke
Does banana spontaneously activate yellow color? Color-related concepts help with color discrimination ..... 1684
Lionel Brunel, Philippe Servajean, Loic Heurley, Nicolas Vermeulen
Recursion in Children's Comprehension and Formulation of Algorithms ..... 1690
Monica Bucciarelli, Robert Mackiewicz, Sangeet Khemlani, Philip Johnson-Laird
Is Structural Priming in Children Facilitated by Interactions between Animacy and Syntax? ..... 1691
Leone Buckle, Anna Theakston, Elena Lieven
Equiprobability principle or "no change" principle? Examining reasoning in the Monty Hall Dilemma using unequal probabilities ..... 1697
Bruce Burns
Discovering simple heuristics from mental simulation ..... 1703
Frederick Callaway, Jessica Hamrick, Tom Griffiths
Fast and Easy: Approximating Uniform Information Density in Language Production ..... 1709
Jesus Calvillo
Learning to Learn Visual Object Categories by Integrating Deep Learning with Hierarchical Bayes ..... 1715Andres Campero, Andrew Francl, Joshua B. Tenenbaum
The paradox of relational development is not universal: Abstract reasoning develops differently across cultures ..... 1721
Alexandra Carstensen, Caren Walker
Grammar-Based and Lexicon-Based Techniques to Extract Personality Traits from Text ..... 1727
Maira B. Carvalho, Max M. Louwerse
Decomposability and Frequency in the Hindi/Urdu Number System ..... 1733
Chundra Cathcart
Is the relative distribution of verbs and nouns modulated by socio-cultural influences? Evidence from bilingual infants and toddlers in Malaysia. ..... 1739
Jun Ho Chai, Hui Min Low, Tze Peng Wong, Julien Mayor
Exploring the relations between oral language and reading instruction in a computational model of reading ..... 1740
Ya-Ning Chang, J. S. H. Taylor, Kathleen Rastle, Padraic Monaghan
Evaluating vector-space models of analogy ..... 1746Dawn Chen, Joshua Peterson, Tom Griffiths
Anticipation Effect after Implicit Distributional Learning ..... 1752
Danlei Chen, Carol Jew, Benjamin Zinszer, Rajeev Raizada
Analytic knowledge for constructing useable empirical causal knowledge: Two experiments on preschoolers ..... 1758
Patricia Cheng, Mimi Liljeholm, Catherine Sandhofer
The relationship between fairness, cognitive control, and numerical encoding ..... 1764
Nadia Chernyak, Peter Blake
Risky Decision Making for Medications: Age and Social Influence Effects ..... 1770
Amy Wing-Lam Chong, Michael Bixter, Wendy Rogers
Finding Creative New Ideas: Human-Centric Mindset Overshadows Mind-Wandering ..... 1776
Yung-Yi Juliet Chou, Barbara Tversky
Transcranial Direct Current Stimulation (tDCS) and the Face Inversion Effect: Anodal stimulation at Fp3 reduces recognition for upright faces ..... 1782
Ciro Civile, Sukhvinder Obhi, Ipl McLaren
Beliefs about sparsity affect causal experimentation ..... 1788
Anna Coenen, Neil Bramley, Azzurra Ruggeri, Todd Gureckis
Path salience in motion events from verbal and visual languages ..... 1794
Neil Cohn, Vivian Wong, Kaitlin Pederson, Ryan Taylor
Using Prior Data to Inform Initial Performance Predictions of Individual Students ..... 1800
Michael Collins, Kevin Gluck, Matthew Walsh, Michael Krusmark
Determinants of judgments of explanatory power: Credibility, Generalizability, and Causal Framing ..... 1806
Matteo Colombo, Leandra Bucher, Jan Sprenger
PACKER: An Exemplar Model of Category Generation ..... 1812
Nolan Conaway, Joseph Austerweil
Opening Up and Closing Down Discussion: Experimenting with Epistemic Status in Conversation ..... 1818
Shauna Concannon, Patrick Healey, Matthew Purver
But where's the evidence? The effect of explanatory corrections on inferences about false information ..... 1824
Saoirse Connor Desai, Stian Reimers
Learning Relational Concepts through Unitary versus Compositional Representations ..... 1830
Daniel Corral, Matt Jones
They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners ..... 1836
Benjamin R. Cowan, Holly Branigan, Habiba Begum, Lucy McKenna, Eva Szekely
Folk Attributions of Control and Intentionality Over Mental States ..... 1842
Corey Cusimano, Geoffrey Goodwin
Judgment Before Emotion: People Access Moral Evaluations Faster than Affective States ..... 1848
Corey Cusimano, Stuti Thapa Magar, Bertram Malle
The Lego hands: changing the affording location of graspable objects ..... 1854
Fabrice Da Silva, Lionel Brunel
People toss coins with more vigor when the stakes are higher ..... 1860
Gregory Dam
Perspective-Taking in Referential Communication: Does Stimulated Attention to Addressee's Perspective Influence Speakers' Reference Production? ..... 1866
Debby Damen, Per van der Wijst, Marije van Amelsvoort, Emiel Krahmer
Sex-Dependent Effects of Emotional Subliminal Visual Stimuli on a Decision-Making Task ..... 1872
Christopher Dancy, Frank Ritter, Frank Hillary, Kevin Voller
Novice to Expert continuum may affect System Response Time ..... 1878
Arindam Das, Olivia Das
Modeling Comprehension Processes via Automated Analyses of Dialogism ..... 1884
Mihai Dascalu, Laura Allen, Danielle McNamara, Stefan Trausan-Matu, Scott Crossley
Amortized Hypothesis Generation1890
Ishita Dasgupta, Eric Schulz, Noah D. Goodman, Samuel J. Gershman
The Causal Sampler: A Sampling Approach to Causal Representation, Reasoning, and Learning ..... 1896
Zachary Davis, Bob Rehder
Understanding The Role of Perception in the Evolution of Human Language ..... 1902
Isaac Davis
Leaping across the mental canyon: Analogical retrieval across disparate task domains ..... 1908
Shir Dekel, Bruce Burns, Micah Goldwater
Conversational topic connectedness predicted by Simplicity Theory ..... 1914
Jean-Louis Dessalles
The Cognitive Architecture of Recursion: Behavioral and fMRI Evidence from the Visual, Musical and Motor Domains ..... 1920
Mauricio de Jesus Dias Martins
The Temporal Cheerleader Effect: Attractiveness Judgments Depend on Surrounding Faces Through Time ..... 1925Alexia Toskos Dils, Scott Niedopytalski, Jeffrey Arroyo, Stephen Flusberg
Architectural process models of decision making: Towards a model database ..... 1931Cvetomir Dimov, Julian Marewski, Lael Schooler
Talking to Ourselves to Engage Control? Testing Developmental Relations Between Self-directed Speech, Cognitive Control and Talkativeness ..... 1937
Sabine Doebel, Yuko Munakata
The Interplay Between Self-evaluation, Goal Orientation, and Self-efficacy on Performance and Learning ..... 1943
Kit Double, Damian Birney
It's not just what we say, it's how we move: An examination of postural activity during a disclosure event ..... 1949Hannah M. Douglas, Stacie Furst-Holloway, Michael J. Richardson, Rachel W. Kallen
A theory of the detection and learning of structured representations of similarity and relative magnitude ..... 1955
Alex Doumas, Aaron Hamer, Guillermo Puebla-Ramirez, Andrea Martin
Co-ordinating Non-mutual Realities: The Asymmetric Impact of Delay on Video-Mediated Music Lessons ..... 1961
Sam Duffy, Patrick Healey
When Do Vehicles of Similes Become Figurative? Gaze Patterns Show that Similes and Metaphors are Initially Processed Differently ..... 1967
Frank Durgin, Rebekah Gelpi
Representing the Richness of Linguistic Structure in Models of Episodic Memory ..... 1973
Melody Dye, Michael Ramscar, Michael Jones
Due process in dual process: A model-recovery analysis of Smith et al. (2014) ..... 1979
Charlotte Edmunds, Andy Wills, Fraser Milton
Experimental Investigation into the Continuous Pattern of the Relationship Between Color Focality and Short-Term Memory Performance for Colors ..... 1985
Siyuan Fang, Tatsunori Matsui
Deconstructing Social Interaction: The Complimentary Roles of Behaviour Alignment and Partner Feedback to the Creation of Shared Symbols ..... 1991Nicolas Fay, Bradley Walker, Nik Swoboda
A Neural Network Model for Taxonomic Responding with Realistic Visual Inputs ..... 1997
Giorgia Fenoglio, Roberto Esposito, Valentina Gliozzi
How Many People Know? Representing the Distribution of Knowledge ..... 2003
Stanka Fitneva
The Effects of Shared Storybook Reading on Word Learning: A Meta-Analysis ..... 2006
Zoe Flack, Andy Field, Jessica Horst
Children Learn Words Better From One Storybook Illustration at a Time ..... 2007
Zoe Flack, Jessica Horst
A Unified Model of Speech and Tool Use Early Development ..... 2013
Sébastien Forestier, Pierre-Yves Oudeyer
What counts as math? Relating conceptions of math with anxiety about math ..... 2019
Ruthe Foushee, Rachel Jansen, Mahesh Srinivasan
After braking comes hasting: reversed effects of indirect associations in 2nd and 4 th graders ..... 2025
Nicole Franke, André Rölke, Ralph R. Radach, Markus J. Hofmann
TRACX2: a RAAM-like autoencoder modeling graded chunking in infant visual-sequence learning ..... 2031
Robert M. French, Denis Mareschal
Risk, Cognitive Control, and Adolescence. Challenging the Dual Systems Model ..... 2037Joanna Fryt, Tomasz Smolen, Karolina Czernecka, Amelia La Torre, Monika Szczygieł
On an effective and efficient method for exploiting "wisdom of crowds in one mind" ..... 2043
Itsuki Fujisaki, Hidehito Honda, Kazuhiro Ueda
Experimental Investigation on Top-down and Bottom-up Processing in Graph Comprehension and Decision ..... 2049Misa Fukuoka, Kazuhisa Miwa, Akihiro Maehigashi
Measures and mechanisms of common ground: backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions ..... 2055
Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten Christiansen, Mark Dingemanse
How do speakers coordinate planning and articulation? Evidence from gaze-speech lags. ..... 2061
Chiara Gambi, Matthew Crocker
Automated Generation of Cognitive Ontology via Web Text-Mining ..... 2067
Richard Gao, Thomas Donoghue, Bradley Voytek
"If It Matters, I Can Explain It": Social Desirability of Knowledge Increases the Illusion of Explanatory Depth ..... 2073
Christian Gaviria, Javier Alejandro Corredor, Zadkiel Zuluaga-Rendón
Reasoning with fundamental rights ..... 2079
Lupita Estefania Gazzo Castaneda, Antje Stemmler, Markus Knauff
The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents ..... 2085
Felix Gervits, Gordon Briggs, Matthias Scheutz
A Two-Stage Model of Solving Arithmetic Problems ..... 2091
Vladimir Glebkin, Alexander Kovtunenko, Ekaterina Krysova
Discourse Acquisition in 'Pear Stories' of Preschool-aged Children ..... 2096Vladimir Glebkin, Nikita Safronov, Varvara Sonina
Inductive reasoning influences perception of interspecies disease transmission risk ..... 2102Micah Goldwater, Molly Ireland, Nick Gaylord, Jason Van Allan, Tyler Davis
High-Performing Readers Underestimate Their Text Comprehension: Artifact or Psychological Reality? ..... 2108
Stefanie Golke, Jörg Wittwer
Computational modeling of auditory spatial attention ..... 2114Edward Golob, K. Brent Venable, Jaelle Scheuerman, Maxwell Anderson
Recruitment of the motor system in the perception of handwritten and typed characters ..... 2120
Chelsea Gordon, Ramesh Balasubramaniam
A Spiking Independent Accumulator Model for Winner-Take-All Computation ..... 2125
Jan Gosmann, Aaron Voelker, Chris Eliasmith
Folk intuitions about consciousness ..... 2131
Katharina Graben, Kai Hamburger, Markus Knauff
Pragmatic aspects of spatial language acquisition and use across languages ..... 2132
Myrto Grigoroglou, Megan Johanson, Anna Papafragou
Cognitive style predicts magical beliefs ..... 2138
Giorgio Gronchi, Jeffrey Zemla, Marco Brondi
Stopping Rules in Information Acquisition at Varying Probabilities and Consequences: an EEG Study ..... 2144
Roberto Guedes de Nonohay, Gustavo Gauer, Richard Gonzalez, Guilherme Lanning
Solving additive word problems: Intuitive strategies make the difference ..... 2150
Katarina Gvozdic, Emmanuel Sander
Age differences in language comprehension during driving: Recovery from prediction errors is more effortful for older adults ..... 2156
Katja Häuser, Vera Demberg, Jutta Kray
Asymmetric detection of changes in volatility: Implications for risk perception ..... 2162
Nigel Harvey, Matt Twyman, Maarten Speekenbrink
Compound effects of expectations and actual behaviors in human-agent interaction: Experimental investigation using the Ultimatum Game ..... 2168
Yugo Hayashi, Ryo Okada
The role of presentation order and orientation on information search and evaluations: An eye-tracking study ..... 2174
Claire Heard, Tim Rakow, Tom Foulsham
Risk and Rationality in Decisions to Commit Crime ..... 2180Rebecca Helm, Valerie Reyna, Renee Williamson, Allison Franz
Planning in Action: Interactivity Improves Planning Performance ..... 2181
Emma Henderson, Gaelle Vallee-Tourangeau, Frederic Vallee-Tourangeau
Head and Heart Metaphors for Moral Decision Making: Conceptual or Communicative? ..... 2187
Rose K. Hendricks, Paul H. Thibodeau
When metaphors in the mind become metaphors in the mouth: Documenting the emergence of a new system of linguistic metaphors for time ..... 2193
Rose Hendricks, Tyler Marghetis, Benjamin Bergen
Boosting Knowledge-Building with Cognitive Dialog Games ..... 2199Jonathan Herberg, Ilker Yengin, Praveena Satkunarajah, Margaret Tan
Leveraging mutual exclusivity for faster cross-situational word learning: A theoretical analysis ..... 2205
Shohei Hidaka, Takuma Torii, George Kachergis
Perceptions of Psychological Momentum in Basketball ..... 2211Thomas Hilbig, Matthew Welsh, Paul Delfabbro
Translating a Reinforcement Learning Task into a Computational Psychiatry Assay: Challenges and Strategies ..... 2217
Peter Hitchcock, Yael Niv, Angela Radulescu, Chris Sims
The Meanings of Morality: Investigating the psychometric properties of distributed representations of latent moral concepts ..... 2223
Joe Hoover, Katie Horton, Morteza Dehghani
Probability judgement from samples: accurate estimates and the conjunction fallacy ..... 2224
Rita Howe, Fintan Costello
Disfluencies in dialogues with patients with schizophrenia ..... 2230Christine Howes, Mary Lavelle, Patrick Healey, Julian Hough, Rose McCabe
Moral Action Changes Mind Perception for Human and Artificial Moral Agents ..... 2236
Evgeniya Hristova, Maurice Grinberg
Reconsideration on Linking Eye-movement Data with Argument Realization ..... 2242
Yujing Huang, Laine Stranahan, Jesse Snedeker
A New Model of Statistical Learning: Trajectories Through Perceptual Similarity Space ..... 2243
Elizabeth A. Hutton, Felix Hao Wang, Jason D. Zevin
The role of prior knowledge and expertise on choice of referring expression ..... 2249
Alyssa Ibarra, Jeffery Runner, Michael Tanenhaus
Beyond Almost-Sure Termination ..... 2255
Thomas Icard
The Influence of Speaker's Gaze on Sentence Comprehension: An ERP Investigation ..... 2261Torsten Jachmann, Heiner Drenhaus, Maria Staudte, Matthew Crocker
In search for the relevant space of implicit memory deficit in dyslexia ..... 2267
Sagi Jaffe-Dax, Merav Ahissar
Cognition Influencing Auditory Perception in SLD Children: Revisiting the Models of Auditory Processing ..... 2273
Saransh Jain, N.p. Nataraja
Failure to use probability of success in deciding whether to pursue one goal or two. ..... 2279
Warren James, Alasdair Clarke, Amelia Hunt
Algebra is not like trivia: Evaluating self-assessment in an online math tutor ..... 2284
Rachel Jansen, Anna Rafferty, Tom Griffiths
Reasoning ability predicts irrational worldview but not conspiracy belief ..... 2290Jan Jastrzebski, Adam Chuderski
The Motor System Does Not Use a Curvilinear Impetus Belief: Folk Physics and Embodied Cognition ..... 2296Jay Jennings, Jim Davies
Network Analysis of a Large Sample of Typical and Late Talkers ..... 2302
Eva Jimenez, Thomas Hills
Spatial language: Meaning, use, and lexical choice ..... 2308
Kristen Johannes, Barbara Landau
Belief Digitization in Economic Prediction ..... 2314
Samuel Johnson, Faith Hill
Measuring Abstract Mindsets through Syntax: Improvements in Automating the Linguistic Category Model ..... 2320
Kate M. Johnson, Reihane Boghrati, Cheryl Wakslak, Morteza Dehghani
Iconicity in Word Learning: What Can We Learn from Cross-Situational Learning Experiments?2321
Matthew Jones, Gabriella Vigliocco
Mindfulness and Fear Conditioning ..... 2327
Fergal Jones, Rossy McLaren, Ipl McLaren
A Study on the Impact of Chess Training on Creativity of Indian School Children ..... 2333
Ebenezer Joseph, Sundar Manoharan, Veena Easvaradoss, David Chandran. E
Document Similarity Misjudgment by LSA: Misses vs. False Positives ..... 2338
Kyung Hun Jung, Eric Ruthruff, Timothy Goldsmith
"The Polar Express" is Bipolar: Critical Film Reviews Influence Uncanny Valley Phenomenon in Animation Films ..... 2344
Jari Kätsyri
Self-other distinction in the motor system during social interaction: A computational model based on predictive processing ..... 2350
Sebastian Kahl, Stefan Kopp
A transfer advantage of learning diagrammatic representations of mathematics ..... 2356
Jennifer Kaminski
Gricean epistemic reasoning in 4-year-olds ..... 2362
Alyssa Kampa, Anna Papafragou
A Hebbian account of entrenchment and (over)-extension in language learning ..... 2366
Vsevolod Kapatsinski, Zara Harmon
Effects of Delayed Language Exposure on Spatial Language Acquisition by Signing Children and Adults ..... 2372Dilay Z. Karadöller, Beyza Sumer, Asli Ozyurek
A Meta-Analysis of the Joint Simon Effect ..... 2377
April Karlinsky, Keith Lohse, Melanie Lam
A toolbox of methods for probabilistic inference ..... 2383
Charles Kemp, Caleb Eddy
The Narrow Conception of Computational Psychology ..... 2389
Luke Kersten
Resolving Two Tensions in $4 E$ Cognition Using Wide Computationalism ..... 2395
Luke Kersten, Joe Dewhurst, George Deane
Processing Spatial Relations: A Meta-Analysis ..... 2401Ann-Kathrin Kießner, Marco Ragni
Decoding Virtual Agent's Emotion and Strategy from Brain Patterns ..... 2407
Eunkyung Kim, Sarah Gimbel, Aleksandra Litvinova, Jonas Kaplan, Morteza Dehghani
Decoding Partner Type in Human-Agent Negotiation using functional MRI ..... 2413
Eunkyung Kim, Jared Gilbert, Charlotte Horowitz, Jonathan Gratch, Jonas Kaplan, Morteza Dehghani
What Do We Learn from Dyslexia and Second Language Learners on the Difference Between Long-term Frequency and Short-term Sequence Repetition Effects? ..... 2419Eva Kimel, Itay Lieder, Luba Daikhin, Hilla Jacoby, Atalia Hai Weiss, Merav Ahissar
"Oops, I did it again." The impact of frequent behaviour on causal judgement. ..... 2420
Lara Kirfel, David Lagnado
Variability in advice taking in decision making ..... 2426
Miho Kitamura, Katsumi Watanabe
Inner speech in post-stroke motor aphasia ..... 2432
Vanja Kljajevic, Estibaliz Ugarte Gómez, Cristina López, Yolanda Balboa Bandeira, Agustin Vicente
The Impact of Population Structure on Models of Language Change ..... 2438
Jordan Kodner, Christopher Cerezo Falco
Eye movements during reference production: Testing the effects of perceptual grouping on referential overspecification ..... 2439
Ruud Koolen, Yorick Fliervoet
Modelling conceptual change as foraging for explanations on an epistemic landscape ..... 2445
Ismo Koponen, Tommi Kokkonen
Action Understanding in High-Functioning Autism: The Faux Pas Task Revisited ..... 2451Joanna Korman, Tiziana Zalla, Bertram Malle
Dependent Choices in Employee Selection: Modeling Choice Compensation and Consistency ..... 2457Antonia Krefeld-Schwalb, Benjamin Scheibehenne, Jörg Rieskamp, Nicolas Berkowitsch
The Influence of Prosody and Case Marking on Thematic Role Assignment in Ambiguous Action Scenes: Adults versus Children ..... 2463
Julia Marina Kroeger, Katja Muenster, Pia Knoeferle
Enhancing metacognitive reinforcement learning using reward structures and feedback ..... 2469
Paul Krueger, Falk Lieder, Tom Griffiths
Thinking and Guessing: Bayesian and Empirical Models of How Humans Search ..... 2475
Marta Kryven, Tomer Ullman, William Cowan, Josh Tenenbaum
The Puzzle of Conditionals with True Clauses: Against the Gricean Account ..... 2476
Karolina Krzyzanowska, Peter Collins, Ulrike Hahn
Cascading effect of context and competition on novel word learning ..... 2482
Sarah Kucker, Aaron Bagley
Geometric Concept Acquisition in a Dueling Deep $Q$-Network ..... 2488
Alex Kuefler, Mykel J. Kochenderfer, James L. McClelland
Perceived Difficulty of Moral Dilemmas Depends on Their Causal Structure: A Formal Model and Preliminary Results ..... 2494
Barbara Kuhnert, Felix Lindner, Martin Bentzen, Marco Ragni
Communicative efficiency in language production and learning: Optional plural marking ..... 2500
Chigusa Kurumada, Scott Grimm
Behavioral Dynamics and Action Selection in a Joint Action Pick-and-Place Task ..... 2506
Maurice Lamb, Tamara Lorenz, Stephen Harrison, Rachel Kallen, Ali Minai, MichaelRichardson
Analogies Emerge from Learning Dyamics in Neural Networks ..... 2512Andrew Lampinen, Shaw Hsu, James L. McClelland
A Hierarchical Bayesian Model of Individual Differences in Memory for Emotional Expressions ..... 2518
David Landy, L. Elizabeth Crawford, Jonathan Corbin
Simulating performance in unconscious plagiarism ..... 2524
Nicholas Lange, Timothy Hollins
When do learned transformations influence similarity and categorization? ..... 2530Steven Langsford, Andrew Hendrickson, Amy Perfors, Daniel Navarro
Preschoolers and Infants Calibrate Persistence from Adult Models ..... 2536
Julia Leonard, Max Kleiman-Weiner, Yuna Lee, Josh Tenenbaum, Laura Schulz
You can take a noun out of syntax...: Syntactic similarity effects in lexical priming ..... 2537
Nicholas Lester, Laurie Feldman, Fermin Moscoso del Prado Martin
How the Mind Exploits Risk-Reward Structures in Decisions under Risk ..... 2543
Christina Leuker, Timothy J. Pleskac, Thorsten Pachur, Ralph Hertwig
Semantic vector evaluation and human performance on a new vocabulary MCQ test ..... 2549
Joseph Levy, John Bullinaria, Samantha McCormick
"I'm Better than You at Labeling!": Preschoolers Use Past Reliability when Accepting Unexpected Labels ..... 2555
Xiaoqian Li, W. Quin Yow
How does Music Reading Expertise Modulate Visual Processing of English Words? An ERP study ..... 2561
Tze Kwan Li, Hei Yan Veronica Chan, Luhe Li, Janet H. Hsiao
Metaphor congruent image schemas shape evaluative judgment: a cross-linguistic study of metaphors for economic change ..... 2567
Patricia Lichtenstein, Ekaterina Shutova
Keystroke Dynamics Predict Essay Quality ..... 2573Aaron Likens, Laura Allen, Danielle McNamara
Visuomotor Adaptation and Sensory Recalibration in Reversed Hand Movement Task ..... 2579Jenny Lin, Yixin Zhu, James Kubricht, Song-Chun Zhu, Hongjing Lu
The development of turn-taking: Pre-schoolers may predict what you will say, but they don't use those predictions to plan a reply. ..... 2585Laura Lindsay, Chiara Gambi, Hugh Rabagliati
Mouse Tracking Shows Attraction to Alternative Targets While Grounding Spatial Relations ..... 2586
Jonas Lins, Gregor Schoener
Prediction and uncertainty in an artificial language ..... 2592
Tal Linzen, Noam Siegelman, Louisa Bogaerts
Explain, Explore, Exploit: Effects of Explanation on Information Search ..... 2598
Emily Liquin, Tania Lombrozo
Can Illness be Bright? Metaphor comprehension depends on linguistic and embodied factors ..... 2604Pei Q Liu, Louise Connell, Dermot Lynott
Want to prime exercise? Calorie labels work better than activity ones! ..... 2610
Dawn Liu, Marie Juanchihch
Failure to replicate talker-specific syntactic adaptation ..... 2616
Linda Liu, Zachary Burchill, Michael Tanenhaus, T. Florian Jaeger
Instrumental Representations of Sensorimotor Control: Representations at Immediate Level ..... 2622
Hsi-wen Liu
Disentangling perceptual and linguistic factors in parsing: Tone monitoring via ERPs ..... 2628
David J. Lobina, Josep Demestre, Jose E. Garcia-Albea
Grasping Multisensory Integration: Proprioceptive Capture after Virtual Object Interactions ..... 2634
Johannes Lohmann, Jakob Gütschow, Martin V. Butz
Goal-Directed Deployment of Attention in a Computational Model: A Study in Multiple-Object Tracking ..... 2640
Andrew Lovett, Will Bridewell, Paul Bello
'It's More Fun With My Phone': A Replication Study of Cell Phone Presence and Task Performance, ..... 2646
Ulrik Lyngs, Ryan McKay
Strategic advantages of micro-targeted campaigns: Implementing savvy persuaders in a Bayesian Agent-Based Model ..... 2651
Jens Koed Madsen, Toby D. Pilditch
Growing a Bayesian Conspiracy Theorist: An Agent-Based Model ..... 2657Jens Koed Madsen, Richard Bailey, Toby D. Pilditch
The dilution effect: Conversational basis and witness reliability ..... 2663
Jens Koed Madsen, Ulrike Hahn, Marion Vorms
Influence of using 3D images and 3D-printed objects on spatial reasoning of experts and novices ..... 2669
Akihiro Maehigashi, Kazuhisa Miwa, Masahiro Oda, Yoshihiko Nakamura, Kensaku Mori,Tsuyoshi Igami
Even when people are manipulating algebraic equations, they still associate numerical magnitude with space ..... 2675
Tyler Marghetis, Robert Goldstone, David Landy
Chunking Ability Shapes Sentence Processing at Multiple Levels of Abstraction ..... 2681
Stewart M. McCauley, Erin S. Isbilen, Morten H. Christiansen
The Interaction of Worked-Examples/ Self-Explanation Prompts and Time on Algebra Conceptual Knowledge ..... 2687
Kelly McGinn, Julie Booth, Laura Young
Semantic ambiguity effects: A matter of time? ..... 2693
Joyse Medeiros, Blair Armstrong
Inferring Intentional Agents From Violation of Randomness ..... 2699
Yuan Meng, Tom Griffiths, Fei Xu
Presentation Format Modulates Adults' Automatic Processing of Proportions ..... 2705
Rui Meng, Percival Matthews
Utilizing simple cues to informational dependency ..... 2710
Hugo Mercier, Helena Miton
Children's EEG Indices of Directed Attention during Somatosensory Anticipation: Relations with Executive Function ..... 2711
Staci Meredith Weiss, Peter Marshall
Modelling Unsupervised Event Segmentation: Learning Event Boundaries from Prediction Errors ..... 2717
Katherine Metcalf, David Leake
Objections to Computationalism. A Short Survey ..... 2723
Marcin Miłkowski
A comparison between human micro-affordances and computational classification ..... 2729
Arthur-Henri Michalland, Arthur-Henri Michalland, Denis Brouillet, Philippe Fraisse
The Role of Schema-Governed Relational Categories in Analogical Inference ..... 2735
Ricardo A. Minervino, Adrián Margni, Máximo Trench
An Investigation of Factors that Influence Resource Allocation Decisions ..... 2741
Percy Mistry, Jennifer Trueblood
Predicting Individual Differences in Working Memory Training Gain: A Machine Learning Approach ..... 2747
Shafee Mohammed, Benjamin Katz, Chelsea Parlett, Martin Buschkuehl, Susanne Jaeggi
A Longitudinal Study of Differences between Predicted, Actual, and Remembered Personal Change ..... 2748Sarah Molouki, Daniel Bartels, Oleg Urminsky
Visual Data Exploration: How Expert Astronomers Use Flipbook-Style Visual Approaches to Understand New Data ..... 2754Fernanda Monteiro Eliott, Keivan Stassun, Maithilee Kunda
The cultural evolution of complex linguistic constructions in artificial sign languages ..... 2760
Yasamin Motamedi, Marieke Schouwstra, Jennifer Culbertson, Kenny Smith, Simon Kirby
Do Accurate Metacognitive Judgments Predict Successful Multimedia Learning? ..... 2766Nicholas V. Mudrick, Michelle Taub, Roger Azevedo
Interactive Communicative Inference ..... 2772
Larry Muhlstein
Cognitive and Attentional Process in Insight Problem Solving of the puzzle game "Tangram" ..... 2778
Yoshiki Nakano
First step is to group them: Task-dynamic model validation for human multiagent herding in a less constrained task ..... 2784
Patrick Nalepka, Maurice Lamb, Rachel W. Kallen, Elliot Saltzman, Anthony Chemero, Michael J. Richardson
Conflicts Processing among Multiple Frames of Reference: An ERP Study ..... 2790
Weizhi Nan, Yanlong Sun, Hongbin Wang, Qi Li, Xun Liu
Beyond Distributed Cognition: Towards a Taxonomy of Nonreductive Social Cognition ..... 2796
Zachariah A. Neemeh, Luis H. Favela
Modeling categorical perception with auditory neurons ..... 2802
Chris Neufeld
Brief Mindfulness Meditation Improves Attention in Novices ..... 2803
Catherine Norris, Daniel Creem
Who makes use of prior knowledge in a curriculum on proportional reasoning? ..... 2809
Daniela Nussbaumer, Christian Thurn, Schumacher Ralph, Stern Elsbeth
Shared mechanisms for controlling egocentric bias during perspective taking and intertemporal choices ..... 2815Garret O'Connell, Bhismadev Chakrabarti, Chun-Ting Hsu, Anastasia Christakou, IsabelDziobek
The dot perspective task revisited: Evidence for directional effects ..... 2821Cathleen O'Grady, Thom Scott-Phillips, Suilin Lavelle, Kenny Smith
Creation of Spatial Mental Models with Figural Stimuli: Validation of the Emoji-based Spatial Integration Task ..... 2827
Polly O'Rourke, Nick Pandza, Susan Campbell
A Domain-Independent Approach of Cognitive Appraisal Augmented by Higher Cognitive Layer of Ethical Reasoning ..... 2833
Suman Ojha, Jonathan Vitale, Mary-Anne Williams
Tracking the temporal course of counterfactual understanding ..... 2839
Isabel Orenes, Juan Antonio Garcia Madruga, Isabel Gómez-Veiga, Orlando Espino, RuthInvestigating Sensitivity to Shared Information and Personal Experience in Children's Use ofMajority Information2840
Kay Otsubo, Andrew Whalen, Daphna Buchsbaum
Do Speaker's Emotions influence their Language Production? Studying the Influence of Disgust and Amusement on Alignment in Interactive Reference ..... 2846
Charlotte Out, Martijn Goudbeek, Emiel Krahmer
Nudging Problematic Smartphone Use to a Lower Level ..... 2852
Lauri Palokangas, Jyrki Suomala
Auditory and Visual Contributions to Multisensory Integration ..... 2858
Jessica Parker, Chris Robinson
The effect of acute physical activity on children's memory for language ..... 2864
Carla Pastorino Campos, John Williams
Replacing Language: Children Use Non-Linguistic Cues and Comparison in Category Formation2870
Margarita Pavlova
When is Likely Unlikely: Investigating the Variability of Vagueness ..... 2876
Kimele Persaud, Brian McMahan, Malihe Alikhani, Kevin Pei, Pernille Hemmer, Matthew Stone
Counterfactuals, indicative conditionals, and negation under uncertainty: Are there cross-cultural differences? ..... 2882
Niki Pfeifer, Hiroshi Yama
Abductive, Causal, and Counterfactual Conditionals Under Incomplete Probabilistic Knowledge ..... 2888
Niki Pfeifer, Leena Tulkki
Pragmatics Influence Children's Use of Majority Information ..... 2894Theresa Pham, Jane C Hu, Daphna Buchsbaum
A categorical (fixed point) foundation for cognition: (adjoint) corecursion ..... 2900
Steven Phillips
Dual-routes and the cost of computing least-costs ..... 2901
Steven Phillips, Yuji Takeda, Fumie Sugimoto
The Effect of Economic Scarcity Priming on Perception of Race ..... 2902Dariusz Lukasz Pilucik, Jens Koed Madsen
Perception is in the Details: A Predictive Coding Account of the Psychedelic Phenomenon ..... 2907
Sarit Pink-Hashkes, Iris van Rooij, Johan Kwisthout
Repetition improves memory by strengthening existing traces: Evidence from paired-associate learning under midazolam ..... 2913Vencislav Popov, Lynne Reder
When Less Isn't More: A Real-World Fraction Intervention Study ..... 2919
Tasha Posid, Vladimir Sloutsky
Is ambiguity detection in haptic imagery possible? Evidence for Enactive imaginings ..... 2925Wim Pouw, Asimina Aslanidou, Kevin Kamermans, Fred Paas
Categorical vs Coordinate Relationships do not reduce to spatial frequency differences ..... 2931Vishaal Prasad, Ben Cipollini, Garrison Cottrell
Refuting Overconfidence: Refutation Texts Prevent Detrimental Effects of Misconceptions on
Text Comprehension and Metacomprehension Accuracy in the Domain of Statistics ..... 2937
Anja Prinz, Stefanie Golke, Jörg Wittwer
Varieties of Numerical Estimation: A Unified Framework ..... 2943
Jike Qin, Dan Kim, John Opfer
Towards an Analogical Theory of Mind ..... 2949Irina Rabkina, Clifton McFate, Ken Forbus, Christian Hoyos
A Theory of Resonance: Towards an Ecological Cognitive Architecture ..... 2955
Vicente Raja
A model of structure learning, inference, and generation for scene understanding ..... 2956
David Raposo, Peter Dayan, Demis Hassabis, Peter Battaglia
Challenging the superficial similarities superiority account for analogical retrieval ..... 2957
Lucas Raynal, Evelyne Clément, Emmanuel Sander
Improving a Fundamental Measure of Lexical Association ..... 2963Gabriel Recchia, Paul Nulty
Priming the production of implications ..... 2969
Alice Rees, Lewis Bott
How could a rational analysis model explain? ..... 2975
Samuli Reijula
Randomness in binary sequences: Conceptualizing and connecting two recent developments ..... 2981
Stian Reimers
Using punctuation as a marker of sincerity and affective convergence during texting ..... 2986
Katherine Reynolds, Blair Casarotto, Sarah Noviski, Jennifer Roche
Influencing Network Graph Perception and Judgment: Effects of Direct Connections, Base Rates, and Visual Layout Proximity on Social Network Analysis ..... 2992
Rebecca Rhodes, Isaiah Harbison, Nathan Bos, Celeste Paul, Clay Fink, Anthony Johnson
Explicit Predictions for Illness Statistics ..... 2998Talia Robbins, Pernille Hemmer
Information theoretic factors in marking linguistic focus: A laboratory-language approach ..... 3004
Gareth Roberts, Jon Stevens
Manual Response Dynamics Reflect Rapid Integration of Intonational Information during Reference Resolution ..... 3010
Timo B. Roettger, Mathias Stoeber
Beyond candidate inferences: People treat analogies as evidence ..... 3016
Brad Rogers, David Landy
A Model-based Approach for Assessing Attentional Biases in People with Depressive Symptoms ..... 3022
Isa Rutten, Wouter Voorspoels, Ernst H. W. Koster, Wolf Vanpaemel
Is it fair? Textual effects on the salience of moral foundations ..... 3028
Estimating Causal Power between Binary Cause and Continuous Outcome ..... 3033
Motoyuki Saito, Patricia Cheng
Spatial Training and Mathematics: The Moderating Effect of Handedness ..... 3039
Giovanni Sala, Martina Bolognese, Fernand Gobet
Dissolving the Grounding Problem: How the Pen is Mightier than the Sword ..... 3045 Nancy Salay
The Causal Frame Problem: An Algorithmic Perspective ..... 3046
Ardavan Salehi Nobandegani, Ioannis N. Psaromiligkos
Experimental and Computational Investigation of the Effect of Caffeine on Human Time Perception ..... 3052
Remya Sankar, Anuj Shukla, Raju Bapi
Peculiarity doesn't trump ordinarity: On recognition memory for exceptions to the category rule ..... 3058Olivera Savic, Vladimir Sloutsky
Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior ..... 3064
K.j. Savinelli, Gregory Scontras, Lisa Pearl
Surprisingly: Marker of Surprise Readings or Intensifier? ..... 3070
Anthea Schöller, Michael Franke
Developing cognitive flexibility in solving arithmetic word problems ..... 3076
Calliste Scheibling-Sève
Aging of the Exploring Mind: Older Adults Deviate more from Optimality in Complex Choice Environments ..... 3082
Job Schepens, Ralph Hertwig, Wouter van den Bos
The impact of sleep on the formation and consolidation of spatial survey knowledge ..... 3088
Wiebke Schick, Julia Holzmann, Hanspeter Mallot
Wiebke Schick, Julia Holzmann, Hanspeter Mallot
The roles of item repetition and position in infant sequence learning ..... 3089
Christina Schonberg, Gary Marcus, Scott Johnson
Silent gesture and noun phrase universals ..... 3095
Marieke Schouwstra, Simon Kirby, Jennifer Culbertson
Discourse continuity promotes children's learning of new objects labels ..... 3101
Jessica F. Schwab, Casey Lew-Williams
The Facilitatory Effect of Referent Gaze on Cognitive Load in Language Processing ..... 3107
Mirjana Sekicki, Maria Staudte
Gestural Hesitation Reveals Children's Competence on Multimodal Communication: Emergence of Disguised Adaptor ..... 3113
Kazuki Sekine
Adaptability and Neural Reuse in Minimally Cognitive Agents ..... 3119
Matthew Setzler, Eduardo J. Izquierdo
The role of talker similarity in the perceptual learning of L2 tone categories ..... 3125
Jing Shao, Joanna Chor Yan Mak, Caicai Zhang
A Spiking Neural Bayesian Model of Life Span Inference ..... 3131
Sugandha Sharma, Aaron Voelker, Chris Eliasmith
The Role of Causality in Temporal Binding: Evidence for an Intentional Boost ..... 3137
Daniel Shiloh, Peter White, Marc Buehner
Familiarity-matching in decision making: Experimental studies on cognitive processes and analyses of its ecological rationality ..... 3143
Masaru Shirasuna, Hidehito Honda, Toshihiko Matsuka, Kazuhiro Ueda
Bridging a Conceptual Divide: How Peer Collaboration Facilitates Science Learning ..... 3149
Andrew Shtulman, Andrew Young
A Data-Driven Approach for Making Analogies ..... 3155
Mei Si, Craig Carlson
Word-object associations are non-selective in infants and young children ..... 3161
Ming Yean Sia, Julien Mayor
The tortoise wins only when the race is long: How the task environment changes the behavior of Tetris models ..... 3167
Catherine Sibert, Wayne Gray
Catherine Sibert, Wayne Gray
Information density of encodings: The role of syntactic variation in comprehension ..... 3168
Les Sikos, Clayton Greenberg, Heiner Drenhaus, Matthew Crocker
Bridging Visual Working Memory Research from Infancy through Adulthood with Dynamic Neural Field Modeling ..... 3174
Vanessa Simmering, Sammy Perone
Pupil Dilation and Cognitive Reflection as Predictors of Performance on the Iowa Gambling Task ..... 3180
Boban Simonovic, Edward Stupple, Maggie Gale, David Sheffield
Talking Through Your Arse: Sensing Conversation with Seat Covers ..... 3186
Sophie Skach, Patrick G. T. Healey, Rebecca Stewart
Individual Differences in Spontaneous Analogical Problem-Solving: The Reflective Mind Account 3191 Slavi Slavov, Penka Hristova
Gender or Community: What Drives STEM Interest Among Middle School Students? ..... 3197
Evans Smalley, Jennifer Kaminski
20-month-olds Use Social-Group Membership to Make Inductive Inferences ..... 3203
Megan Smith, Rose Scott
Thinking inside the box: Motion prediction in contained spaces uses simulation ..... 3209
Kevin Smith, Filipe Peres, Edward Vul, Joshua Tenebaum
Promoting Spontaneous Analogical Transfer: The Role of Category Status ..... 3215
Sean Snoddy, Kenneth Kurtz
Geometry-based Affordances ..... 3221
Sterling Somers
Acquiring pitch associations across modalities: the role of experience ..... 3227
Laura Speed, Ilja Croijmans, Sarah Dolscheid, Asifa Majid
Theory of Mind and Valuation during Cooperation ..... 3228Michael Spezio, Tessa Rusch, Yuqing Lei, Vanessa Hayes, Jan Glaescher
Fake News and False Corroboration: Interactivity in Rumor Networks ..... 3229
Michael Spivey
Adapting to a listener with incomplete lexical semantics ..... 3235
Sadhwi Srinivas, Barbara Landau, Colin Wilson
Children's use of lexical flexibility to structure new noun categories ..... 3241Mahesh Srinivasan, Catherine Berner, Hugh Rabagliati
Rationalizing subjective probability distortions ..... 3247
Nisheeth Srivastava, Ed Vul
Memory of relative magnitude judgments informs absolute identification ..... 3248
Nisheeth Srivastava
Exploration and Skill Acquisition in a Major Online Game ..... 3249
Tom Stafford, Sam Devlin, Rafet Sifa, Anders Drachen
Biases and labeling in iterative pragmatic reasoning ..... 3255
Jon Stevens
Actions that modify schedules of reinforcement ..... 3261
Mac Strelioff, Mimi Liljeholm
Shaping the Dynamics of Category Learning in Infants and Adults by Varying Learning Context ..... 3267 Jelena Sucevic, Nadja Althaus, Kim Plunkett
Early produced signs are iconic: Evidence from Turkish Sign Language ..... 3273
Beyza Sumer, Clara Grabitz, Aylin Kuntay
A Minimal Neural Network Model of The Gambler's Fallacy ..... 3279
Yanlong Sun, Hongbin Wang
Temporal variability in moral value judgement ..... 3285
Alexandra Surdina, Adam Sanborn
Enforced pointing gesture can indicate invisible objects behind a wall ..... 3291
Hajime Takahashi, Tetsuya Yasuda, Harumi Kobayashi
Similarities Between Objects In Analogies Framed By Schema-Governed Categories ..... 3296
Lucía Micaela Tavernini, Máximo Trench, Valeria Olguín, Ricardo A. MinervinoWhat can Hand Movements Tell us about Audience Engagement?3302Lida Theodorou, Patrick HealeyGeneralizing relations during analogical problem solving in preschool children: does blocked orinterleaved training improve performance?3308Jean-Pierre Thibaut, Micah Goldwater
Generalizing novel names in comparison settings: role of conceptual distance during learning and at test ..... 3314
Jean-Pierre Thibaut, Arnaud Witt
Solving the Puzzle to Reach the Summit: Using Metaphor to Gauge Public Perceptions of Science ..... 3320Paul Thibodeau, Stephen Flusberg, Kevin Holmes
Metaphors, Roles, and Controls in Framing Studies ..... 3326
Paul Thibodeau, Stephen Flusberg
Part-whole categorization is culture-specific ..... 3332Timothy Tilbe, Juergen Bohnemeyer, Gabriela Perez Baez, Dedre Gentner
The Stroop Effect From a Mixture of Reading Processes: A Fixed-Point Analysis ..... 3338
Gabriel Tillman, Zachary Howard, Paul Garret, Ami Eidels
Interleaving area problems in the 4 th grade classroom: What is the role of context and practice? ..... 3344
Rachael Todaro, Bradley Morris
Scarcity impairs online detection and prospective memory ..... 3350
Brandon Tomm, Jiaying Zhao
Specificity and entropy reduction in situated referential processing ..... 3356
Elli Tourtouri, Francesca Delogu, Matthew Crocker
A Cognitive Model of Social Influence ..... 3362
Greg Trafton, Malcolm McCurry, Kevin Zish, Laura Hiatt, Sunny Khemlani
Why Teach How Things Work? Tracking the Evolution of Children's Intuitions About Complexity ..... 3368Emmanuel Trouche, Aaron Chuey, Kristi Lockhart, Frank Keil
The Impact of Presentation Order on the Attraction Effect in Decision-making ..... 3374
Jennifer Trueblood, Aneesha Dasari
Interpretation and Processing Time of Generalized Quantifiers: Why your Mental Space Matters3380
Alice Ping Ping Tse, Marco Ragni
Right hemisphere lateralization and holistic processing do not always go together: An ERP investigation of a training study ..... 3386Ricky Van-yip Tso, Hangyu Chen, Yui Andrew Yeung, Terry Kit-fong Au, Janet Hsiao
"He's pregnant": simulating the confusing case of gender pronoun errors in L2 ..... 3392
Chara Tsoukala, Stefan Frank, Mirjam Broersma
Deconstructing transitional probabilities: Bigram frequency and diversity in lexical decision ..... 3398
Russell Turk, Gary Jones, Duncan Guest, Angela Young, Mark Andrews
Extraneous visual noise facilitates word learning ..... 3404
Katherine Twomey, Lizhi Ma, Gert Westermann
Comparison strategies in the change detection task are influenced by task demands. ..... 3410Rob Udale, Simon Farrell, Chris Kent
Weight matters: The role of physical weight in non-physical language across age and culture ..... 3411
Tomer Ullman, Santiago Alonso-Diaz, Stephen Ferrigno, Sarina Zahid, Celeste Kidd
Learning in the Wild: Real-World Experiences Shape Children's Knowledge Organization ..... 3417
Layla Unger, Anna Fisher
The Effects of Social Task Setting on Time Perception ..... 3423
Kerem Alp Usal, Annette Hohenberger
Instruction type and believability influence on metareasoning in a base rate task ..... 3429
Pavle Valerjev, Marin Dujmović
Text, images and diagrams as information providers ..... 3435
Sara Verbrugge, Aurélie Stas, Koen Jaspaert
The Role of Linguistic Information in Learning Abstract Words: Evidence from Children with Specific Language Impairment (SLI) ..... 3441
Gabriella Vigliocco, Marta Ponari, Courtenay Norbury
Facial Motor Information is Sufficient for Identity Recognition ..... 3447Jonathan Vitale, Benjamin Johnston, Mary-Anne Williams
Forgetting My Memories by Listening to Yours: The Impact of Perspective-Taking on Socially-Triggered Context-Based Prediction Error ..... 3453
Madalina Vlasceanu, Rae Drach, Alin Coman
Rank Aggregation and Belief Revision Dynamics ..... 3454
Igor Volzhanin, Ulrike Hahn, Dell Zhang, Stephan Hartmann
Overcoming the Tragedy of Personnel Evaluation? ..... 3460
Momme von Sydow, Niels Braus, Ulrike Hahn
Altruist vs. Egoist Detection and Individual vs. Group Selection in Personnel Management ..... 3466
Momme von Sydow, Niels Braus
Rational and Semi-Rational Explanations of the Conjunction Fallacy: A Polycausal Approach ..... 3472
Momme von Sydow
Empathic Humans Punishing an Emotional Virtual Agent ..... 3478
Laura Wächter, Barbara Kuhnert, Marco Ragni
Exploitative and Exploratory Attention in a Four-Armed Bandit Task ..... 3484
Adrian Walker, Mike Le Pelley, Tom Beesley
Actively Detecting Patterns in an Artificial Language to Learn Non-Adjacent Dependencies ..... 3490
Hao Wang, Jason Zevin, Calvin Leather, Toby Mintz
A Plausible Micro Neural Circuit for Decision-Making ..... 3491
Hui Wei, Dawei Dai, Yijie Bu
The Cognitive Reflection Test: familiarity and predictive power in professionals ..... 3497
Matthew Welsh, Steve BeggDesires influence 4- to 6-year-old children's probabilistic judgments3503Adrienne Wente, Mariel Goddu, Elyanah Posner, Teresa Garcia, María Fernández Flecha,Alison Gopnik
The Interaction of Bayesian Pragmatics and Lexical Semantics in Linguistic Interpretation:
Using Event-related Potentials to Investigate Hearers' Probabilistic Predictions ..... 3504
Markus Werning, Erica Cosentino
The Neural Mechanisms of Relational Reasoning: Dissociating Representational Types ..... 3510
Julia Wertheim, Marco Ragni
How the truth can make a great lie: An empirical investigation of the folk concept of lying by falsely implicating ..... 3516
Alex Wiegmann, Pascale Willemsen
Eliciting Middle School Students' Ideas About Graphs Supports Their Learning from a Computer Model ..... 3522
Eliane Wiese, Anna Rafferty, Marcia Linn
Counterfactual Conditionals and Normative Rules ..... 3528Meredith Wilkinson, Shira Elqayam, Valerie Thompson, David Over
Modality Differences in Timing: Testing the Pacemaker Speed Explanation ..... 3534
Emily Williams, Andrew Stewart, Luke Jones
Modeling cognitive load effects in an interrupted learning task: An ACT-R approach ..... 3540
Maria Wirzberger, Günter Daniel Rey, Josef F. Krems
Does Associative Memory Play a Role in Solving Physics Problems? ..... 3546
Katie Wood, Rossy McLaren, Ipl McLaren
Sensorimotor Learning Modulates Automatic Imitation in Visual Speech ..... 3552
Yuchunzi Wu, Bronwen Evans, Patti Adank
Whoa! Aww ... Ohh ... Hee! and Mmm: Infants' nuanced distinctions about the probable causes of emotional expressions ..... 3553
Yang Wu, Paul Muentener, Laura Schulz
Visual and Audio Aware Bi-Modal Video Emotion Recognition ..... 3554Siqi Xiang, Wenge Rong, Zhang Xiong, Min Gao, Qingyu Xiong
Teaching Versus Active Learning: A Computational Analysis of Conditions that Affect Learning ..... 3560
Scott Cheng-Hsin Yang, Patrick Shafto
Training Graph Literacy: Developing the RiskLiteracy.org Outreach Platform ..... 3566
Vincent Ybarra, Edward Cokely, Cody Adams, Margo Woller-Carter, Jinan Allan, Adam Feltz, Rocio Garcia-Retamero
Empirical constraints on computational level models of interference effects in human probabilistic judgements. ..... 3572
James Yearsley, Emmanuel Pothos, Albert Barque-Duran
Evidence for overt visual attention to hand gestures as a function of redundancy and speech disfluency ..... 3578
Amelia Yeo, Martha AlibaliPhysical problem solving: Joint planning with symbolic, geometric, and dynamic constraints3584Ilker Yildirim, Tobias Gerstenberg, Basil Saeed, Marc Toussaint, Josh Tenenbaum
Three-Way Bindings in Associative Recognition ..... 3590
Hyungwook Yim, Adam Osth, Vladimir Sloutsky, Simon Dennis
Emotional and Cognitive Interest: How Creating Situational Interest Affects Learning With Multimedia ..... 3596
Angela Yoo, Richard Catrambone
"I won't lie, it wasn't amazing": Modeling polite indirect speech ..... 3602
Erica Yoon, Michael Henry Tessler, Noah Goodman, Michael Frank
The Structure of Young Children's Numerical and Spatial Abilities ..... 3608
Christopher Young, Stephen Raudenbush, Brittney Fraumeni, Susan Levine
Alternation blindness in the perception of binary sequences ..... 3609
Ru Qi Yu, Daniel Osherson, Jiaying Zhao
Back to ABCs: Clustering Alphabetically, Rather than Semantically, Enhances Vocabulary Learning ..... 3615
Jingqi Yu, Veronica Yan, Elizabeth Bjork, Robert Bjork
Domain-General Learning of Neural Network Models to Solve Analogy Task - A Large-Scale Simulation ..... 3621
Arianna Yuan
Convincing Conversations: Using a Computer-Based Dialogue System to Promote a Plant-Based Diet ..... 3627Emma Zaal, Gregory Mills, Afke Hagen, Carlijn Huisman, John Hoeks
Pseudoneglect and development: Age-related spatial bias in bisection and drawing ..... 3633
Yordanka Zafirova, Asenia Giagtzidou, Dara Vassileva, Elena Andonova
Is Red Fire Warmer than Blue Fire? Colored Thermal Words in a Stroop Task ..... 3639
Yordanka Zafirova, Armina Janyan
The Redundancy Effect in Human Causal Learning: Evidence Against a Comparator Theory Explanation ..... 3640
Tara Zaksaite, Peter M. Jones
Modeling Semantic Fluency Data as Search on a Semantic Network ..... 3646Jeffrey C. Zemla, Joseph L. Austerweil
The Impact of Decision Agency $\mathcal{\xi}$ Granularity on Aptitude Treatment Interaction in Tutoring ..... 3652
Guojing Zhou, Min Chi
Information Seeking as Chasing Anticipated Prediction Errors ..... 3658
Jian-Qiao Zhu, Wendi Xiang, Elliot Ludvig
Interruptions Reduce Confidence Judgments: Predictions of Three Sequential Sampling Models ..... 3664
Kevin Zish, Malcolm McCurry, Nathan Aguiar, J. Gregory Trafton
Posters: Member Abstracts
Simulation and heuristics in flexible tool use ..... 3670
Kelsey Allen, Kevin Smith, Josh Tenenbaum
Predicting Preschool-Aged Children's Behavior Regulation from Attention Tasks in the Lab ..... 3671
Chelsea Andrews, Emily Coates, Kristine Kovack-Lesh, Vanessa Simmering
Why do we punish negligent behaviors? ..... 3672
Sarin Arunima, Cushman Fiery
When I say 'Black Lives Matter,' why do some people hear 'Others Lives Matter Less'? ..... 3673
E. Malemma Azumah, Jason Shepard
The Use of Ambiguous Messages as a Strategy to Appeal to Multiple Decision Makers ..... 3674
Devika Banerji, Torsten Reimer, Chris Roland
Social wayfinding in complex environments ..... 3675
Iva Barisic, Tyler Thrash, Victor R. Schinazi, Christoph Hoelscher
Projecting space into the future: peripersonal space remaps in anticipation of an object manipulation ..... 3676Anna Belardinelli, Johannes Lohmann, Martin Butz
Characterizing Human-Machine Teams with Process Algebras ..... 3677
Leslie Blaha, Robert Jasper
Bottom-up attentional cueing in category learning in children ..... 3678
Nathaniel Blanco, Vladimir Sloutsky
Infusing Cognitive Science Content in Teacher Preparation ..... 3679
Julie Booth
Using Analogical Processing to Categorize Musical Patterns ..... 3680
Janet Bourne, Elliot Chun
How infants map nonce phrases to scenes with objects and predicates. ..... 3681
Angelica Buerkin-Pontrelli, Daniel Swingley
No Tranfer of Training in Simple Addition ..... 3682
Jamie Campbell, Yalin Chen
Vowel Harmony as a Distributional Learning Problem ..... 3683
Spencer Caplan, Jordan Kodner
A comparative assessment of embodied and computational topic extraction ..... 3684
James Carney
Mapping hand to world; Development of iconic representation in gesture and homesign ..... 3685
Erica Cartmill, Lilia Rissman, Miriam Novack, Susan Goldin-Meadow
One-shot word learning under high and low sentential constraints in adult L2 learners of Chinese ..... 3686
Jenn-Yeu Chen, Yu-Shu Huang
Comparing Human Use of Fast $\mathcal{F}$ Frugal Tree with Machine-Learning Tree ..... 3687
Yee Siang Chng
Incorrect responses salience affects strategy use in a figural analogy task ..... 3688
Iwona Ciechanowska, Bartlomiej Kroczek, Adam Chuderski
Learning Object Names from Visual Pervasiveness: the Visual Statistics Predict ..... 3689Elizabeth Clerkin, Chen Yu, Linda Smith
Relations Between Intuitive Biological Thought and Scientific Misconceptions ..... 3690
John Coley, Nicole Betz, Jessica Leffers, Yian Xu, Michal Fux, Kristin de Nesnera, Kimberly Tanner
A picture falls under many categories: How ancient mathematical marks became extinct ..... 3691
Peter Coppin, Daemon Retren, Ambrose Li
ANCHORING is amodal: evidence from a signed language. ..... 3692
Qatherine Dana, Diane Brentari, Outi Bat-El, Iris Berent
Don't forget to bind: Memory binding and interference in development ..... 3693
Kevin Darby, Vladimir Sloutsky
Developmental Changes in Visual Scene Statistics ..... 3694
Christina Deserio, Jason Gold, Swapnaa Jayaraman, Rowan Candy, Linda Smith
Speed and accuracy trade-off of semantic composition involving highlighting and adjustment ..... 3695
Diana Dimitrova, Brian McElree, Petra Schumacher
A Neurodynamical Model of How Prior Knowledge Influences Visual Perception ..... 3696
Dražen Domijan, Mateja Marić
Couples Emotion Dynamics During Conversations Involving Stress and Enjoyment ..... 3697
Nicholas Duran, Ashley Randall
Information Signatures in Children's Language Environment ..... 3698Steven L. Elmlinger, Drew H. Abney, David W. Vinson, Linda B. Smith, Chen Yu
The acquisition of verb morphology in Polish and Finnish: Model and experiment ..... 3699
Felix Engelmann, Joanna Kolak, Sonia Granlund, Ben Ambridge, Julian Pine, Anna Theakston, Elena Lieven
What makes a joke funny: Analysing joke humor through single-word ratings. ..... 3700
Tomas Engelthaler, Thomas Hills
What's on your wandering mind? The content of mind wandering during text- and film comprehension ..... 3701
Myrthe Faber, Sidney Dmello
Interaction with a robot changes human motor behavior ..... 3702
Laura Fademrecht, Tobias Meilinger, Stephan Streuber, Aurelie Saulton, Heinrich Bülthoff, Rouwen Cañal-Bruland, Stephan de La Rosa
The differential effects of transmission and interaction on linguistic variation ..... 3703
Olga Feher, Kenny Smith
Production of morphologically complex words as revealed by a typing task: Morphological influences on keystroke dynamics ..... 3704
Laurie Feldman, Rick Dale, Jacolien van Rij, David Vinson
Probability matching as a cognitive basis of cultural drift ..... 3705
Vanessa Ferdinand
Is gender-fair langauge needed? How grammatical gender influences representations of discourse referents ..... 3706
Evelyn Ferstl, Lena Dietsche
Introducing a New JavaScript Framework for Professional Online Studies. ..... 3707
Holger Finger, Dorena Diekamp, Caspar Goeke
How to communicate uncertainty in severe weather forecasts? ..... 3708
Nadine Fleischhut, Stefan Herzog, Ralph Hertwig
Developing Visual Closure in Infancy ..... 3709
Samuel Forbes, Kim Plunkett
Executive Functions and Academic Achievement in a High-Poverty Sample ..... 3710
Gill A. Francis (1), Zewelanji Serpell (2), Teresa Parr (3), Michelle R. Ellefson (1)
Neural Phase Synchrony on Understanding Meanings of Symbols ..... 3711
Masayuki Fujiwara, Takashi Hashimoto, Guanhong Li, Jiro Okuda, Takeshi Konno, Kazuyuki Samejima, Junya Morita
Jumping in Japanese: Converting linguistic instructions into physical performances ..... 3712
Chie Fukada, Noriyuki Kida, Hiromichi Hagihara, Takatsugu Kojima
The Price of Fear: Developing a behavioural assessment of fear-related avoidance incorporating dynamic response measures. ..... 3713
Santiago Garcia-Guerrero, Denis O'Hora, Arkady Zgonnikov
On the road to .. somewhere? Change-blindness in event description tasks is informative about the interrelation between visual perception and language planning ..... 3714
Johannes Gerwien, Ines Marberg
Mental computations underlying morphosyntax acquisition ..... 3715
Heidi Getz, Elissa Newport
Computational Modelling of Embodied Semantic Cognition: A Deep Learning Approach ..... 3716
Ajitesh Ghose, Rick Cooper
The Perceived Duration of Vast Spaces ..... 3717
Devin Gill, Jeanine Stefanucci
The role of spatial skills in mathematics cognition: Evidence from children aged 5-10 years ..... 3718
Katie Gilligan, Alex Hodgkiss, Michael Thomas, Emily Farran
Investigating the predictions of a memory-based account of statistical learning ..... 3719
Sandrine Girard, Erik Thiessen
Motor Fluency Effects on Causal Judgment: The Role of Grip-Strength Asymmetries and Spatial-Numeric Associations ..... 3720
Kelly Goedert, Daniel Czarnowski
Comparison of directed gaze during vocalizations in bonobo and human infants ..... 3721
Ulrike Griebel, Josep Call, Eugene Buder, D. Kimbrough Oller
How does of initial inaccuracy benefit cross-situational word learning? ..... 3722
Chris Grimmick, George Kachergis, Todd Gureckis
On the Detection of "Alternative Facts" in Environmental Messages: The Effects of a Sequential versus a Simultaneous Presentation Format ..... 3723
Mona Guath, Peter Juslin
Exploring inductive bias of visual scenes ..... 3724Jessica Hamrick, David Bourgin, Thomas Langlois, Tom Griffiths
Choosing while Losing: The Effects of Valence and Relative Magnitude on Decision Dynamics. ..... 3725
Avril Hand, Denis O'Hora, Rick Dale, Petri Piiroinen
The Relationship between Anxiety, Mind Wandering and Task-switching: A Diffusion Model Analysis ..... 3726
Andree Hartanto, Hwajin Yang
Do people behave dishonestly easily? ..... 3727
Hajimu Hayashi
Hierarchical Processing of Response Production and Categorisation ..... 3728
Liusha He, Rick Cooper
Where are you? The Effect of Uncertainty and its Visual Representation on Location Judgments in GPS-like Displays ..... 3729
Mary Hegarty, Alinda Friedman, Alexander P. Boone, Trevor J. Barrett
Semantic Bootstrapping in Frames: A Computational Model of Syntactic Category Acquisition ..... 3730
John Hewitt, Charles Yang
Tangible rhythm: Sensorimotor representations of metrical structure and learning musical rhythm with gesture ..... 3731
Courtney Hilton, Micah Goldwater, Michael Jacobson
Eye movement-based probabilistic models for physical scene understanding ..... 3732
Eghbal Hosseini, Eli Pollock, Tobias Gerstenberg
Data Driven Eye Gaze Path Segmentation ..... 3733
Joseph Houpt, Mary Frame, Leslie Blaha
Case Markers Facilitate Abstraction of Syntax among Mandarin-speaking preschoolers ..... 3734
Dong-Bo Hsu
An Exploratory Study on Remote Associates Problem Solving: Evidence of Eye Movement Indicators ..... 3735
Po-Sheng Huang, Shu-Ling Peng, Jon-Fan Hu
Processing of Filler-gap Dependency in Island Constraints and its Relation to Working Memory for Non-native Speakers of English ..... 3736Kuan-Jung Huang, Jon-Fan Hu, Anastazija Rajkovic
Linguistic processes in translation: Eye-tracking reveals differential effects of phrase order and lexical choice ..... 3737
Xin Huang, Wei (sophia) Deng, Defeng Li
How reactivation strength affects memory updating ..... 3738
Scully Iiona, Hupbach Almut
An fNIRS Hyperscanning Study on Brain-Brain Interactions of a Dyad during a Joint Sentence Reading Task ..... 3739
Erdinc Isbilir, Murat Perit Cakir, Fred Cummins, Hasan Ayaz
Pre-term infants exhibit impaired prediction and learning in Audio-Visual association paradigm ..... 3740
Sagi Jaffe-Dax, Alex M Boldin, Nathaniel D Daw, Lauren L Emberson
Endpoints and Midpoints in Event Perception ..... 3741
Yue Ji, Anna Papafragou
Dynamic and multiplexed networks for working memory ..... 3742
Elizabeth Johnson, Robert Knight
Improving Perceptual Reasoning in School Children through Chess Training ..... 3743
Ebenezer Joseph, Sumathi Chandrasekaran, Uma T.r., Veena Easvaradoss
Measuring Demand Avoidance with the Demand Selection Task: Challenges and Opportunities ..... 3744
Ion Juvina, Jeffrey Nador, Othalia Larue, Randall Green, Brandon Minnery, Assaf Harel
Optimizing Mathematic Learning: Effects of Continuous and Nominal Practice Format on Transfer of Arithmetic Skills ..... 3745
Charles Kalish, Rui Meng, Ayon Seng, Percival Matthews
Priors, informative cues and ambiguity aversion ..... 3746
Lauren Kennedy, Amy Perfors, Daniel Navarro
Moral Judgments in Trolley Like Dilemmas: An Eye-Tracking Study ..... 3747
Vjeran Keric, Evgeniya Hristova
Word retrieval decline in midlife: a voxel-based morphometry study ..... 3748
Vanja Kljajevic
The relationship between verbal route descriptions and personal characteristics of empathy ..... 3749
Takatsugu Kojima
Explanatory Completeness: Evidence from Causal Chains ..... 3750
Joanna Korman, Sangeet Khemlani
Different alternative explanations can render different information relevant to explaining an event ..... 3751
Barbara Koslowski, Francoise Vermeylen
One-shot Learning and Classification in Children ..... 3752
Eliza kosoy, Brenden Lake, Josh Tenenbaum
Mentioning atypical properties of objects is communicatively efficient ..... 3753
Elisa Kreiss, Robert X.d. Hawkins, Judith Degen, Noah Goodman
Why Would 'Same' Go With 'Same'? Exploring New Factors Required For Relational Reasoning ..... 3754
Ivan Kroupin, Susan Carey
When reading is harder than a mother kucker: Top-down effects of the taboo-ness on novel wordpronunciation3755
Sarah Kucker, Lynn Perry
Effects of Variable Response-Stimulus Interval (RSI) On Sequence Learning ..... 3756
Sneha Kummetha, Anuj Shukla, Raju Bapi
Object Understanding: Exploring the Path from Percept to Meaning ..... 3757
Kenneth Kurtz, Daniel Silliman
Effect of Touch-produced Sounds on Surface Texture Perception ..... 3758
Jinhwan Kwon, Suguru Hata, Natsumi Komoto, Maki Sakamoto
Identifying Causal Direction in the Two-Variable Case ..... 3759Pablo León-Villagrá, Christopher Lucas
Intolerance to uncertainty is associated with diminished exploration ..... 3760
Jennifer Lenow, Nathaniel Daw, Elizabeth Phelps
Interplay between semantic and emotional information in visual scene processing ..... 3761
Anaïs Leroy, Sylvane Faure, Sara Spotorno
The effect of binaural beats on inhibition ..... 3762
Kimery Levering, Molly Poinan, Kristin Jay
Perception of others: Representation of immigrant groups in newspaper articles ..... 3763
Ying Li, Thomas Hills
Modelling the dynamics of integrating context into perception: in good and in poor readers ..... 3764
Itay Lieder, Vincent Adam, Maneesh Sahani, Merav Ahissar
The Effects of Familiarity and Typicality on Naming Objects and Faces ..... 3765
Yu-Ching Lin, Jon-Fan Hu, Shu-Ling Peng, Shulan Hsieh
The Effect of CSAL AutoTutor on Deep Comprehension of Text in Low-Literacy Adult Readers ..... 3766 Genghu Shi, Arthur Graesser
Adaptive response priors in context-dependent decision-making ..... 3767
Olga Lositsky, Michael Shvartsman, Robert C. Wilson, Jonathan D. Cohen
Comprehension of Chinese Classifiers in Preschool Children ..... 3768
Yu-Han Luo, Jon-Fan Hu
Cumulative response probabilities: Estimating time course of lexical activation from single-point response times ..... 3769
Sahil Luthra, James Magnuson
Guardian and Daily Mail Readers' Implicit Attitudes to Immigration ..... 3770
Dermot Lynott, Michael Walsh, Tony McEnery, Louise Connell, Liam Cross
How and when does the syllable become a reading unit? Developmental evidence in French children ..... 3771Norbert Maïonchi-Pino, Virginie Loiseau
Epistemically Suspect Beliefs can be partly explained by individual's propensity towards contradiction ..... 3772
Yoshimasa Majima
Practicing an auditory working memory task recruits lower-level auditory areas in a task-specific manner ..... 3773Tamar Malinovitch, Philippe Albouy, Merav Ahissar, Robert Zatorre
Do Relationality and Aptness Influence Conventionalization? ..... 3774
Francisco Maravilla, Dedre Gentner
Mutual Exclusivity Revisited - When Pragmatics overrides Novelty ..... 3775Hanna Marno, Dan Sperber
Search Your Feelings, Luke: Emotional Fluency Predicts Well-being and Emotional Intelligence ..... 3776
Masitah Masitah, Ying Li, Thomas Hills
Children's reasoning about data sets ..... 3777Amy Masnick, Bradley Morris, Christopher Was
Attention Modulation Effects on Visual Feature-selectivity of Neurons in Brain-inspired Categorization Models ..... 3778Saeed Masoudnia, Abdol-Hossein Vahabie, Majid Nili Ahmadabadi, Babak Nadjar Araabi
Simple and Complex Working Memory Tasks Allow Similar Benefits of Information
Compression3779fabien mathy, mustapha chekaf, Nelson Cowan
Causal asymmetry and the intuitive physics of collision events ..... 3780
Ralf Mayrhofer, Michael Waldmann
The development of interpersonal regret and its relation to prosocial choice ..... 3781
Teresa McCormack, Brian Uprichard
Interactive and embodied repair: Displaying, recognizing, and negotiating misalignment in an emerging language context ..... 3782Ashley Micklos
Misalignment increases abstraction of referring expressions ..... 3783
Gregory Mills, Gisela Redeker
The spontaneous creation of systems of conventions ..... 3784
Jennifer Misyak, Nick Chater
Refining the cognitive semantic web: The tensor method to represent the topographic emplacement of different word categories ..... 3785Eduardo Mizraji, Andrés Pomi
Picture book reading in the lives of 18-30 month old children: A diary study ..... 3786Jessica Montag, Linda Smith
Resemblance among similarity measures in semantic representation ..... 3787Maria Montefinese, David Vinson
Effects of Question Format on Test-Taker Cognition ..... 3788Jung Aa Moon, Irvin Katz, Madeleine Keehner
Magnitude of metaphor and its effect on reasoning about immigration ..... 3789
Karie Moorman, Teenie Matlock
Investigating the Impact of Sleep on Eyewitness Memory ..... 3790
David Morgan, Jakke Tamminen, Laura Mickes
Impact of testimony and prior knowledge on children's beliefs about category homogeneity ..... 3791
Kelsey Moty, Amanda Brandone
Learning to Consider Alternative Causes: Can Practice Make Us More Aware of Our Imperfection? ..... 3792
Edward Munnich, Ana Maria Hoffmann, Nancy Ortega, Emma Weinberger, Dana-Lis
Bittner, Jacqueline Tiongco, Jasmine West
Bidirectional effect of emotional contagion for pain during face-to-face interaction ..... 3793
Aiko Murata, Tatsuya Kameda, Katsumi Watanabe
Agent's symmetry elicits egocentric transformations for spatial perspective-taking ..... 3794
Hiroyuki Muto, Soyogu Matsushita, Kazunori Morikawa
Robot as Moral Agent: A Philosophical and Empirical Approach ..... 3795Shoji Nagataki, Masayoshi Shibata, Tatsuya Kashiwabata, Takashi Hashimoto, TakeshiKonno, Hideki Ohira, Toshihiko Miura, Shinichi Kubota
Iconicity vs. Systematicity in Artificial Language Learning ..... 3796
Alan Ks Nielsen, Julia Simner, Simon Kirby, Kenny Smith
The space and time of contamination: Complete, continual, spreading effects ..... 3797
Laura Niemi, Liane Young
Construction of design activity index based on the value of artifact ..... 3798
Kunio Nikata, Kentaro Inomata, Toru Sato, Keigo Kawasaki, Noriko Nagata
Indirection Explains Flexible Tuning of Neurons in Prefrontal Cortex ..... 3799
David Noelle
Five-Year-Old Children Transfer a Metacognitive Strategy to a Novel Task ..... 3800
Allison O'Leary, Vladimir Sloutsky
Do forgiving God primes strengthen support for state sanctioned punishment? ..... 3801
Katherine O'Lone, Ryan McKay
Effects of motives of search and prior experiences on online browsing performance:
Considerations from searchers cognitive load ..... 3802
Kayoko Ohtsu, Takako Sakawaki
Anticipatory Active Inference from Learned Recurrent Neural Forward Models ..... 3803Sebastian Otte, Theresa Schmitt, Martin V. Butz
Application of fuzzy logic in dyslexia user modelling to design customizing assistive technology ..... 3804
Tereza Pařilová, Eva Hladká
An Exploratory Study of the Influence of Pretend Play on Children's Self-Regulation and Language Skills ..... 3805
Tanya Paes, Michelle Ellefson
Interpreting nonsignificant findings in psychological research ..... 3806
Bence Palfi, Aba Szollosi, Barnabas Szaszi, Marton Kovacs, Mark Zrubka, Peter Szecsi, Balazs Aczel, Eric-Jan Wagenmakers
Perceived control in bounded-rational decision-making ..... 3807
Saee Paliwal, Frederike Petzschner, Ekaterina Lomakina, Klaas Enno Stephan
Applications of Cognitive Science to Enhancing Scholarly Communication ..... 3808
Purav Patel
Novel metacognitive problem solving task ..... 3809
Jwalin Patel, Amanda Aldercotte, Teresa Parr, Zewelanji Serpell, Michelle Ellefson
Comparing comparison indices: Assessing the validity of different magnitude comparison measures across presentation formats and age groups ..... 3810
Marcie Penner-Wilger, Aaron Cecala, Melissa Elfers
Language input and development during a year in an early intervention classroom ..... 3811
Lynn Perry, Emily Prince, Adriana Valtierra, Camila Rivero-Fernández, Mary Anne
Ullery, Lynne Katz, Daniel Messinger
Early Visual Evoked Potentials (VEPs) in Infant Siblings of Children with ASD, ADHD and Age-Matched Controls ..... 3812
Elena Serena Piccardi, Mark H. Johnson, Teodora Gliga
How does social touch modulate arousal states? An investigation in early development. ..... 3813
Laura Pirazzoli, Emily Jones, Teodora Gliga, Mark H. Johnson
Hand, spoon or toothbrush? Towards the understanding of the neural underpinnings of affective touch in 5 months-old infants. ..... 3814
Laura Pirazzoli, Sarah Lloyd-Fox, Mark H. Johnson, Teodora Gliga
Do we see things better when we know grammar? ..... 3815
Fenna Poletiek, Maartje Van de Velde
Patterns of Cortical Activation Correlate With Speech Understanding After Cochlear Implantation ..... 3816
Luca Pollonini, Heather Bortfeld, Michael Beauchamp, John Oghalai
Encouraging Fruit and Vegetable Consumption Through Intuitive Theory Building ..... 3817
Derek Powell, Ellen Markman
Scientific Reasoning Ability in Middle Schoolers related to MasterMind Discovery Strategy ..... 3818
Jean-Baptiste Quillien, Keisha Varma, Purav Patel, Martin Van Boekel
Neural responses decrease while performance increases with practice: A neural network model ..... 3819
Milena Rabovsky, Steven S. Hansen, James L. McClelland
Age-related top-down and bottom-up guidance on eye movements when searching in real-world scenes ..... 3820
Hanane Ramzaoui, Sylvane Faure, Sara Spotorno
The Long and Short of It: The Role of Verb Stem Vowel Duration in Sentence Processing ..... 3821Gwendolyn Rehrig, Sten Knutsen, Nicolaus Schrum, Paul de Lacy, Karin StromswoldRecycling or Trash Bin? Modeling Consumers'Recycling Behavior in a Field Study3822
Torsten Reimer, Juan Pablo Ramirez, Chris Roland, Devika Banerji
Is Neurocomputational Self-Organization a Core Mechanism of AGI Systems? ..... 3823Spyridon Revithis
Does a present bias influence exploratory choice? ..... 3824
Alexander S Rich, Todd M Gureckis
Everyday object affordance enhances automatic inhibitory control: an ERP study ..... 3825
Stefania Righi, Giorgio Gronchi, Lapo Pierguidi, Maria Pia Viggiano
Metacognitive Monitoring of Internal and External Storage and Retrieval ..... 3826
Evan Risko, Connor Gaspar, Dave McLean, Tim Dunn, Derek Koehler
From Concrete Examples to Abstract Relations: A model-based neuroscience approach to how people learn new categories ..... 3827
Gareth Roberts, Micah Goldwater, Evan Livesey, Josue Giron, Tyler Davis
Progress in building a machine that can ask interesting and informative questions ..... 3828
Anselm Rothe, Brenden Lake, Todd Gureckis
Shifting backward to say what's front? Spatial referencing of dorsal object arrangements ..... 3829
Annelie Rothe-Wulf, Andrea Bender, Julia Fischer, Nadine Lobmüller, Sieghard Beller
The time course of Intentional Binding3830
Miriam Ruess, Roland Thomaschke, Andrea Kiesel
A positive attitude increases subjective life expectancy ..... 3831
Susana Ruiz Fernández, Martin Lachmair, Juan José Rahona, Lotte Sophia Roessler, Peter Gerjets
Probability matching in choice behavior influenced by virtual rewards ..... 3832
Yuichi Saito, Miyuki G. Kamachi
ASR Systems as Models of Phonetic Category Perception in Adults ..... 3833
Thomas Schatz, Francis Bach, Emmanuel Dupoux
Which test to perform? Modeling utility of medical tests: information gain, patient risk and financial costs ..... 3834
Clara Schirren, Stefanie C. Hautz, Wolf E. Hautz, Juliane E. Kämmer, Björn Meder,Jonathan D. Nelson
Intuitive system control: Challenging the standard model of dynamic decision making ..... 3835
Wolfgang Schoppek
Learning Temporal Generative Neural Codes for Biological Motion Perception and Inference ..... 3836
Fabian Schrodt, Martin V. Butz
The Sufficiency Principle: Predicting when children will regularize inconsistent language variation ..... 3837
Kathryn D. Schuler, Jaclyn E. Horowitz, Charles Yang, Elissa L. Newport
Impaired phonological processing of lexical tones in Cantonese speakers with congenital amusia . 3838Jing Shao, Xunan Huang, Caicai Zhang
Discovering kinds of future-oriented thought using automated machine-learning techniques ..... 3839
Jason Shepard, Aiming Nie, Bridget Copley, Phillip Wolff
The influence of word-order harmony on structural priming in artificial languages ..... 3840
Kenny Smith, Olga Feher, Jennifer Culbertson
Recently rewarded task-irrelevant stimuli do not distract 2-year-olds during visual search ..... 3841
Hayley K Smith, Erik Blaser, Zsuzsa Kaldy
Disambiguating Disfluencies: What Do Speech Disfluencies Tell Us About Speech Production? . 3842 Mikhail Sokolov, John Logan
Phonological Competition during Spoken-Word Recognition in Infants and Adults ..... 3843
Marlene Spangenberg, Kim Plunkett
Word order rules in business name binomials ..... 3844
Samuel Spevack, Patricia Lichtenstein, Stephanie Shih
The time course of colour guidance in realistic scene search ..... 3845
Sara Spotorno, Ben Tatler
Perceptual decision making from correlated samples ..... 3846
Oana Stanciu, Mate Lengyel, Daniel Wolpert, Jozsef Fiser
Optimality of visual search under ambiguous stimuli ..... 3847Elina Stengård, Ronald van den Berg
It is new, but will it be good? Context-driven exploration of novel options ..... 3848
Hrvoje Stojic, Eric Schulz, Maarten Speekenbrink
Mathematical Symbol Recognition in Children ..... 3849
Sandra Street
Speaking in English, sorting in Chinese: interaction in L2 can reinforce existing categories in L13850
Ellise Suffill, Holly Branigan, Martin Pickering
The effect of overt language use on category induction ..... 3851Justin Sulik, Haily Merritt, Gary Lupyan
Respecting UP and Despising DOWN: Emotional and Body-based Image in Japanese Verbs ..... 3852
Tomohiro Taira
A time-series eye-fixation analysis of the similarity-compromise effect in multi-alternative choice ..... 3853
Tsuzuki Takashi, Chiba Itsuki
Relationship between four measures reflecting representations of fraction magnitude in adults:number line estimation, comparison, calculation of fractions, and immediate serial recall offractions3854
Yuki Tanida, Yu Koshima, Masahiko Okamoto
Slow Change: The Visual Context for Real World Learning ..... 3855
Charlene Tay, Linda Smith, Chen Yu
Effects of attention to emergent phenomena on rule discovery ..... 3856Hitoshi Terai, Kazuhisa Miwa, Sho Yokoyama, Souta Fujimura, Gotaro Nakayama
Scheduling system delays for optimal user performance: Don't predict time; let time predict! ..... 3857Roland Thomaschke, Lennart Koch, Miriam Ruess, Andrea Kiesel
Computational Foundations of Cultural Evolution: Modeling the Emergence of Systems from Higher-order Probabilistic Inference ..... 3858
Bill Thompson
Counterfactual thoughts and judgments about morally good actions ..... 3859
Shane Timmons, Ruth Byrne
Gradually ascending sound with accelerating automatic driving vehicle might change passengers, tension or anxiety: analysis of biometrical index. ..... 3860
Akitoshi Tomita, Etsuko Harada, Kozue Miyashiro, Satoshi Ando, Maito Ohmori, HiroakiYano
Does sonority influence the syllable segmentation in visual identification? Evidence in French skilled readers. ..... 3861
Méghane Tossonian, Norbert Maïonchi-Pino
Poverty of materials makes recursive combination operation evolvable ..... 3862
Genta Toya, Takashi Hashimoto
Fitting a Stochastic Model to Eye Movement Time Series in a Categorization Task ..... 3863
Paul Tupper, Thuan Pham Nguyen, Yunlong Nie, Jiguo Cao, Liangliang Wang
Gaze during utterances and silence in L1 and L2 Conversations ..... 3864
Ichiro Umata, Koki Ijuin, Tsuneo Kato, Seiichi Yamamoto
Statistical Learning Contributions to Semantic Knowledge Development ..... 3865
Layla Unger, Anna Fisher, Robert Powers
Improving Number Foundations in Preschoolers: ANS versus Symbolic Knowledge ..... 3866
Jo Van Herwegen, Hiwet Costa, Bethany Nicholson, Chris Donlan
Biological and Artificial Perspectives on Metacognition ..... 3867
Katherine Wagner
Predicting Future Performance in an ITS system via Gradient Boosting Classification ..... 3868
Breya Walker, Anne lippert, Raven Davis, Zhiqiang Cai, Cheng Qinyu, Genghu Shi,Arthur Graesser
Interactivity, Stereotype threat, and Working memory ..... 3869Anna-Stiina Wallinheimo, Harriet Tenenbaum, Adrian Banks
Judging Magnitude: Is there a Common Cognitive System for Different Types of Magnitude Judgments? ..... 3870
Rylan J. Waring, Lana M. Trick
Effects of Auditory-Feedback Delays and Musical Roles on Coordinated Timing Asymmetries in Piano Duet Performance ..... 3871
Auriel Washburn, Matthew Wright, Takako Fujioka
What influences the impact of warning labels in decisions from description-plus-experience ..... 3872
Leonardo Weiss-Cohen, Emmanouil Konstantinidis, Maarten Speekenbrink, Nigel Harvey
Later lexical development in bilinguals ..... 3873Anne White, Barbara Malt, Steven Verheyen, Gert Storms
An immersive binaural horizon for sonic data analytics ..... 3874Richard Windeyer, Dan MacDonald, David Steinman, Ambrose Li, Peter Coppin
A model of cultural co-evolution of language and mindreading ..... 3875
Marieke Woensdregt, Simon Kirby, Chris Cummins, Kenny Smith
Language as a process: An exploration among pre-adolescent Chinese EFLs ..... 3876
Hing Yi Wong, Susanna Yeung, Wen Chung Wang, Chi Wing Yip
Influences of the Matching Effects of Cognitive and Emotional Factors on Attitude Change ..... 3877
Tung-Cheng Wu, Jon-Fan Hu, Shu-Ling Peng
Is infants' mutual exclusivity response based on preference to novelty or non-name of an object? ..... 3878
Qinmei Xu, Ye Tao, Yuyan Wang, Jon-Fan Hu
Children's Attention to Semantic Content versus Emotional Tone: Differences between Two Cultural Groups ..... 3879
Yang Yang, Li Wang, Qi Wang
Knowledge partitioning in forecasting ..... 3880
Lee-Xieng Yang, Tzu-Hsi Lee
Working Memory and lexical ambiguity resolution in Chinese ..... 3881
Michael Yip
Impact of Polarity, Rationality, and Math Ability on Numerical Magnitude Knowledge ..... 3882
Laura Young, Julie Booth, Kelly McGinn
Global consequences of local complexity: evidence from recall of visually presented nonwords ..... 3883
Mackenzie Young, Colin Wilson
Children's Reasoning about Geometric Footprints ..... 3884
Christopher Young, Alana Foley, Susan Levine
Relevance Theory, Pragmatic Inference and Cognitive Architecture ..... 3885
Wen Yuan, Richard Cooper
Do you forgive past mistakes of virtual assistant? A study on changing impressions of virtual character when using its assistance multiple times ..... 3886
Masahide Yuasa
Knowledge acquiring on event chronology in Russian-language texts ..... 3887
Vera Zabotkina, Mikhail Epifanov, Elena Pozdnyakova
The Impact of Presentation Order on Category Learning Strategies: Behavioral Data and Self-Reports ..... 3888
Christina Zeller, Ute Schmid
Walking dynamics of intertemporal choice ..... 3889
Arkady Zgonnikov, Iñaki Rañó, Denis O'Hora, Kongfatt Wong-Lin
A Spatial-Temporal Analysis of a Visual Working Memory Task with EEG and ECoG ..... 3890
Qiong Zhang, Marieke van Vugt, Jelmer Borst, John Anderson
To organize or not to organize? Examining biases in search strategies using Lego building blocks ..... 3891Mona Zhu, Kimberley Yuen, Evan Risko
Emotion in Deceptive Language ..... 3892
Iraide Zipitria, Basilio Sierra, Imanol Sopena-Garaikoetxea
Author Index ..... 3893
Reviewers List ..... 3920

## Dear Cognitive Scientists,

Welcome to London for the 39th Annual Conference of the Cognitive Science Society! Our meeting brings together some of the most innovative and exciting research in cognitive science. The program features exceptional plenary talks by three leading international experts: Richard Granger (Dartmouth), Ulrike Hahn (Birkbeck, University of London), and Richard Sutton (University of Alberta). It also includes an invited symposium aimed at showcasing the theme for this year: Computational Foundations of Cognition.

CogSci 2017 received 1185 submissions, including 873 full papers, 259 member abstracts, 20 publication-based short papers, as well as 19 proposals for a symposium, 9 for a workshop, and 5 for a tutorial. After a rigorous review process, we selected 255 papers for oral presentation ( $29 \%$ ), 418 papers for poster presentation (48\%), 233 member abstracts for poster presentation, 15 publication-based talks, 9 symposia, 7 workshops, and 3 tutorials.

We hope that you enjoy the program this year and the global city of London, certainly the most vibrant and historically significant city in the United Kingdom. There are thousands of shops and restaurants near the conference venue and throughout the city. We encourage you to set aside the time to enjoy some of the many tourist attractions while you are here.

## Your Program Co-Chairs,

Glenn Gunzelmann (U.S. Air Force Research Laboratory), Andrew Howes (University of Birmingham, UK), Thora Tenbrink (Bangor University, Wales, UK), Eddy Davelaar (Birkbeck, University of London, UK)

## Acknowledgements

We are very grateful to everyone who contributed to the planning and organization of this year's meeting and to all reviewers who generously donated time to evaluate submissions. We thank the 193 Program Committee members and the 1,587 reviewers who were essential to the review process. We thank the awards committee who helped assess the prize-winning papers. All Program Committee members are listed below.

We are especially grateful for the assistance of a number of individuals and groups for many organizational aspects of the meeting. We thank Jessica Wong, the Cognitive Science Society's Conference Officer, for managing pretty much everything; Deborah Gruber, the Cognitive Science Society's Business Manager, for handling the business details; Susan Gelman, Nora Newcombe, and the Cognitive Science Society's Executive Committee and Governing Board for their constant support and advice; the Scarritt Group, Inc., for logistics and support; James Stewart of Precision Conference Solutions for maintaining the conference reviewing system; Anna Papafragou, Dan Grodner, Dan Mirman, and John Trueswell and the other members of last year's Cogsci conference committee for their helpful advice and insights; our many Committee Chairs, listed below, who enthusiastically took care of their various responsibilities; the student volunteers, listed below, who keep the conference running smoothly on site; and importantly our conference sponsors, listed below, whose generous support enabled us to focus more on the program and less on the finances.

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

- Glenn Gunzelmann, Andrew Howes, Thora Tenbrink, and Eddy Davelaar Program Co-Chairs, Cognitive Science 2017


## Sponsors

The Robert J. Glushko and Pamela Samuelson Foundation
Centre for Cognition, Computation, and Modelling
DeepMind Technologies
Frontiers In Psychology Cognitive Science
Nature Human Behaviour
Wiley
Cognitive Science Society

Thank you again for your support!

## How to Cite Your Paper

## APA formatted citation for the entire proceedings:

Gunzelmann, G., Howes, A., Tenbrink, T., \& Davelaar, E. J. (2017). Proceedings of the 39th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

## APA formatted citation for a 6-page paper:

Author A. \& Author B. (2017). This is the title of the paper. In G. Gunzelmann, A. Howes, T. Tenbrink, \& E. J. Davelaar (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society (pp. PAGES). Austin, TX: Cognitive Science Society.

## APA formatted citation for a published abstract:

Author A. \& Author B. (2017). This is the title of the abstract [Abstract]. In G. Gunzelmann, A. Howes, T. Tenbrink, \& E. J. Davelaar (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society (p. NUMBER). Austin TX: Cognitive Science Society.

## APA formatted citation for a talk/poster presentation:

Author A. \& Author B. (2017, July). This is the title of the talk or poster. Paper (or Poster) presented at the 39th Annual Conference of the Cognitive Science Society. London, England, UK.

## Organizing Committee

## Program Chairs

Glenn Gunzelmann (U.S. Air Force Research Laboratory)
Andrew Howes (University of Birmingham)
Thora Tenbrink (Bangor University)
Eddy Davelaar (Birkbeck, University of London)

## Committee Chairs

## Workshop \& Tutorials Committee

Stella Christie (Swarthmore College)
Pete Khooshabeh (U.S. Army Research Laboratory)

## Symposium \& Publication-Based Talks Committee

Chuck Kalish (University of Wisconsin, Madison)
Member Abstract Committee
Sudha Arunachalam (Boston University) Jared Novick (University of Maryland - CASL)

## Award Committee

Ion Juvina (Wright State University)

## Student Volunteer Committee

Anna Belardinelli (University of Tübingen)
Maarten Speekenbrink (University College London)
Social Media Committee
Eliane Stampfer-Wiese (University of California, Berkeley) Caitlin Tenison (Carnegie Mellon University)

Local Information
Eddy Davelaar (Birkbeck, University of London) Sabine Topf (University College London)

Sponsorship
Eddy Davelaar (Birkbeck, University of London, UK)
Andrew Howes (University of Birmingham, UK)

## Website

Aditya Acharya (University of Birmingham, UK)

## Conference Officer

Jessica Wong (University of Chicago)

## Program Committee

Abrahamson, Dor<br>Acarturk, Cengiz<br>Altmann, Erik<br>Andonova, Elena<br>Apperly, Ian<br>Arik, Engin<br>Arunachalam, Sudha<br>Ashley, Richard<br>Barley, Mike<br>Bartels, Daniel<br>Beierholm, Ulrik<br>Belardinelli, Anna<br>Beller, Sieghard<br>Bender, Andrea<br>Bergmann, Kirsten<br>Best, Brad<br>Blaha, Leslie<br>Blessing, Stephen<br>Bonnefon, Jean-Francois<br>Borovsky, Arielle<br>Bosse, Solveig<br>Bott, Lewis<br>Bowman, Howard<br>Brighton, Henry<br>Brunye, Tad<br>Bryant, David<br>Buehner, Marc<br>Burigo, Michele<br>Burns, Bruce<br>Butz, Martin<br>Cangelosi, Angelo<br>Catrambone, Richard<br>Chang, Franklin<br>Chater, Nick<br>Cheng, Peter<br>Christie, Stella<br>Claramunt, Christophe<br>Clausner, Timothy<br>Clifton, Charles<br>Cottrell, Gary

Couclelis, Helen
Craig, Scotty
Culbertson, Jennifer
Cummins, Fred
Dale, Rick
Danks, David
Danovitch, Judith
Davies, Jim
De Brigard, Felipe
de Ruiter, J.P.
de Sa, Virginia
Deák, Gedeon
Dehghani, Morteza
Echols, Catharine
Epstein, Susan
Erdfelder, Edgar
Fabrikant, Sara Irina
Farmer, George
Farmer, Thomas
Fausey, Caitlin
Feist, Michele
Fiser, Jozsef
Fisher, Anna
Foltz, Anouschka
Forbus, Ken
Frank, Mike
Frank, Stefan
Freksa, Christian
French, Bob
Fu, Wai-Tat
Galati, Alexia
Gelman, Susan
Gentner, Dedre
Gershman, Samuel
Giersch, Anne
Gluck, Kevin
Goel, Ashok
Gonnerman, Laura
Gonzalez, Cleotilde
Goodman, Noah

Griffiths, Tom
Grinberg, Maurice
Gureckis, Todd
Gweon, Hyowon
Halbrügge, Marc
Hale, John
Hartshorne, Joshua
Hayhoe, Mary
Healey, Patrick
Hegarty, Mary
Hills, Thomas
Hochmann, Jean-Remy
Houpt, Joseph
Hsiao, Janet
Hubbard, Edward
Husband, E. Matthew
Jaeger, T. Florian
Juvina, Ion
Kalish, Charles
Kan, Irene
Katherine, White
Katz, Irvin
Kello, Christopher
Khooshabeh, Peter
Kidd, Celeste
Klippel, Alexander
Kloos, Heidi
Knoeferle, Pia
Koedinger, Ken
Kopp, Stefan
Kwisthout, Johan
Levin, Beth
Levy, Roger
Lovett, Andrew
Ludvig, Ellio†
Luhmann, Christian
Lupyan, Gary
Lynott, Dermot
MacDonald, Maryellen
Maes, Alfons
Maglio, Paul
Magnani, Lorenzo
Majerus, Steve
Majid, Asifa

Mallot, Hanspeter
Malmberg, Kenneth
Mareschal, Denis
Markman, Art
Masnick, Amy
McNamara, Danielle
Medina, Jared
Meilinger, Tobias
Mello, Catherine
Mickes, Laura
Miles, Gareth
Mirman, Daniel
Montello, Daniel
Morris, Bradley
Munro, Paul
Myers, Christopher
Newcombe, Nora
Noppeney, Uta
Noveck, Ira
Novick, Jared
O'Seaghdha, Padraig
Olivetti, Marta
Pani, John
Papafragou, Anna
Patrick, Sturt
Payne, Stephen
Peebles, David
Perfors, Amy
Pustejovsky, James
Rabagliati, Hugh
Ragni, Marco
Redeker, Gisela
Reitter, David
Rips, Lance
Ritter, Frank
Rushton, Simon
Schmid, Ute
Schunn, Christian
Schwering, Angela
Sims, Chris
Sloutsky, Vladimir
Smith, Kenny
Spenader, Jennifer
Spranger, Michael

Steedman, Mark
Stell, John
Stenning, Keith
Stewart, Andrew
Stout, Dietrich
Taatgen, Niels
Tappe, Heike
Tenenbaum, Josh
Thomas, Rick
Timpf, Sabine
Trafton, Greg
Trueblood, Jennifer
Trueswell, John

Tsigaridi, Dido
Tversky, Barbara
Uttal, David
VanLehn, Kurt
Walsh, Matthew
Warren, Paul
Was, Christopher
Weidemann, Christoph
Werker, Jane $\dagger$
Wicker, Bruno
Wills, Andy
Wood, Zena

## Student Volunteers

Lena Maria Blott (University College London)
Pablo Cesar de Juan Bernabeau (Radboud University, Tillburg University)
April Karlinsky (University of British Columbia)
Alex Kuefler (Stanford University)
Anais Leroy (University of Nice Cote d'Azur)
Tamar Malinovitch (Hebrew University of Jerusalem)
Rui Meng (University of Wisconsin, Madison)
Purav Patel (University of Minnesota)
John Patterson (Binghamton University)
Francisco Quiroga-Vergara (University College London)
Hanane Ramzaoui (Cote d'Azur University)
Gwendolyn Rehrig (Rutgers University)
Megan Smith (University of California, Merced)
Sean Snoddy (Binghamton University)
Patrick Trettenbrein (University of Graz)
Siqi Xiang (Beihang University)
Amelia Yeo (University of Wisconsin, Madison)
Qiong Zhang (Carnegie Mellon University)
Mark Zrubka (Eotvos Lorand University)

## Conference Awards

## Robert J. Glushko Dissertation Prizes

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

The 2017 recipients of the Robert J. Glushko Prizes for Outstanding Doctoral Dissertations / Theses in Cognitive Science are listed below.

Alexandra Carstensen: "Universals and variation in language and thought: Concepts, communication, and semantic structure", 2016 (University of California, Berkeley)

Judith Ellen Fan: "Role of cognitive actions in learning", 2016 (Princeton University)

Julian Jara-Ettinger: "The inner life of goals: Costs, rewards, and commonsense psychology", 2016 (MIT)

Samuel Johnson: "Cognition as sense-making", 2016 (Yale University)
Dave Kleinschmidt: "Perception in a variable but structured world: The case of speech perception", 2016 (University of Rochester)

The Glushko Dissertation Prize Symposium showcases the award winning PhD research projects, moderated by Adele Goldberg (Princeton).

Friday, 3:00PM-4:40PM (King Suite)

## Marr Prize

The Marr Prize, named in honor of the late David Marr, is awarded to the best student paper at the conference. All student first authors were eligible for the Marr Prize for the best student paper. The Marr Prize includes an honorarium of $\$ 1000$ and is sponsored by The Cognitive Science Society. The winner of the 2017 Marr Prize for the Best Student Paper is:

Melody Dye, Petar Milin, Richard Futrell, \& Michael Ramscar:
Cute little puppies and nice cold beers: An information theoretic analysis of prenominal adjectives
Thursday, 10:50-11:10, Linguistic Conventions (Blenheim)

## Computational Modeling Prizes

Four prizes worth $\$ 1000$ each are awarded for the best full paper submissions to CogSci 2017 that involve computational cognitive modeling. The four prizes represent the best modeling work in the areas of perception/action, language, higher-level cognition, and applied cognition. These prizes are sponsored by The Cognitive Science Society. The winners of the 2017 Computational Modeling Prizes are listed below.

## Perception \& Action:

Tianmin Shu, Yujia Peng, Lifeng Fan, Song-Chun Zhu, \& Hongjing Lu: Inferring Human Interaction from Motion Trajectories in Aerial Videos Friday, 16:00-16:20, Interaction (Blenheim)

## Language:

## Peter Blouw \& Chris Eliasmith:

Inferential Role Semantics for Natural Language
Thursday, 10:50-11:10, Semantics \& Concepts (Buckingham)

## Higher-Level Cognition:

## Simon Stephan \& Michael Waldman:

Preemption in Singular Causation Judgments: A Computational Model Saturday, 15:00-15:20, Judgement (Windsor)

## Applied Cognition:

Jane Wang, Zeb Kurth-Nelson, Hubert Soyer, Joel Leibo, Dhruva Tirumala, Remi Munos, Charles Blundell, Dharshan Kumaran, \& Matt Botvinick: Learning to Reinforcement Learn Saturday, 10:30-10:50, Learning Mechanisms (Blenheim)

## Student Travel Awards

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

Keith Ransom
Ardavan Salehi Nobandegani
Maurici López-Felip
Christian Ramiro
Katherine Adams
Anita Slonimska
Vencislav Popov
Shari Liu
Vasanth Sarathy
Rachel Magid
Matt Lou-Magnuson
Elizaveta Konovalova
Martin Schoemann
Darren Frey
Jasmeen Kanwal
Yue Ji
Tian Xu
Francis Mollica
Michael Tessler
Dan Kim

# Invited Presentations 

## Rumelhart Prize Lecture

Lila Gleitman, University of Pennsylvania
Takes Two to Tango: The Linguistic Representation of Symmetry
Friday, July 28, 17:10-18:10, King Suite

## Rumelhart Prize Symposium

Symposium in honor of Lila Gleitman:
Finding structure and meaning in language acquisition
Panelists:
Elissa Newport, Georgetown University
Cindy Fisher, University of Illinois
John Trueswell, University of Pennsylvania
Barbara Landau, Johns Hopkins University
Friday, July 28, 10:30-12:10, King Suite

## Heineken Prize Lecture

Elizabeth Spelke, Harvard University
Core Knowledge and Composition
Thursday, July 27, 17:10-18:10, King Suite

## Thursday's Keynote

Richard Sutton, University of Alberta
Reinforcement Theories of Learning and Thinking
Thursday, July 27, 9:00-10:00, King Suite

## Friday's Keynote

Richard Granger, Dartmouth
Principles of Brain Circuit Architectures, from Percept to Concept
Friday, July 28, 9:00-10:00, King Suite

## Saturday's Keynote

Ulrike Hahn, Birkbeck, University of London
Rationality and the Role of Limited Experience
Saturday, July 29, 9:00-10:00, King Suite

## Invited Symposium

## Computational Foundations

Panelists:
Peter Dayan, University College London
Karl Friston, University College London
Matt Botvinick, DeepMind, University College London
Howard Bowman, University of Kent
Ulrike Hahn, Birkbeck, University of London
Thursday, July 27, 10:30-12:10, King Suite

# Tutorial: Recent Advances in Deep Learning 

Matthew Botvinick (botvinick@google.com)<br>DeepMind, London, U.K. Gatsby Computational Neuroscience Unit, University College London

## Peter Battaglia (peterbattaglia@google.com)

DeepMind, London U.K.

Keywords: Deep learning, reinforcement learning, artificial intelligence

## Overview

The past several years have seen a dramatic acceleration in artificial intelligence (AI) research, driven in large part by innovations in deep learning and reinforcement learning (RL) methods. The relevant developments, as showcased in a series of recent high-profile publications in Nature and elsewhere (e.g., Graves et al., 2016; Mnih et al., 2015; Silver et al., 2016), have generated intense interest in cognitive science, partially because they appear to have potentially far-reaching implications for understanding human intelligence. Unfortunately, the pace of innovation in AI has been so rapid that it is difficult for non-experts and sometimes even for experts - to stay abreast of the latest developments.
The present tutorial brings together five front-line researchers in AI, each with dual credentials in neuroscience and/or cognitive science, to provide an accessible overview and update on the most important recent developments in deep learning and deep RL. The tutorial will be aimed at a broad audience, ranging from graduate students to senior investigators, and spanning specialties from cognitive and developmental psychology to psychiatry, human factors research, and systems neuroscience. The focus will be on fundamental concepts and principles, and a central goal will be to maximize accessibility, in line with the tutorial format.

## Significance

Neural network modeling has played a pivotal role in cognitive science since at least the 1980's. Over the past decade or so, neural networks have been overshadowed to some extent by other techniques. Beginning around 2012, interest in neural network methods (often rebranded as 'deep learning') began to take off machine learning research, and have since then become the dominant approach in AI. In combination with RL methods, deep learning has enabled a series of breakthroughs in tasks ranging from image classification to game play (see Marblestone et al., 2016 for a review).
The implications of this tectonic shift for cognitive science are currently under intensive debate (Marblestone, 2016; Lake et al., 2016). It seems clear that AI innovations, including memory architectures, generative models, and deep RL techniques are likely to stimulate new hypotheses
about human cognition. At the same time, it seems likely that AI research would benefit from richer input from cognitive science.

Ironically, the potential for exchange between the two fields has been hindered by the very pace of innovation in AI. (Emblematic is the fact that the developments we will review in the present tutorial have almost all emerged since the last time the Cognitive Science meeting featured a tutorial on neural networks, just two years ago.) Our aim in the present tutorial is to mitigate this problem by providing an accessible update on the most recent key developments in deep learning and deep RL.

## Tutorial structure and activities

The tutorial will assume a half-day format, consisting of five tutorial lectures, each covering an area in which some of the most important recent innovations have arisen. As detailed below, all five lecturers are members of the research team at DeepMind in London (deepmind.com), with dual citizenship in AI and cognitive science. The material covered in each lecture will include recent work at DeepMind, but also related work from other groups.

## Participant credentials

Matthew Botvinick (Organizer) is DeepMind's Director of Neuroscience Research and Honorary Professor in the Gatsby Computational Neuroscience Unit. He holds a Ph.D in Cognitive Neuroscience from CMU's Center for the Neural Basis of Cognition, and has done extensive research in the computational neuroscience of reinforcement learning and decision making. His research at DeepMind focuses in part on meta-learning.
Peter Battaglia (Organizer) is a senior Research Scientist at DeepMind. He holds a Ph.D. in Brain and Cognitive Sciences from MIT and has done extensive research in scene representation, intuitive physics, and probabilistic inference. His research at DeepMind focuses on novel architectures for structured inference, with an emphasis understanding physical systems.
Tim Lillicrap is a senior Research Scientist at DeepMind. He holds a Ph.D in Neuroscience from Queen's University, and has done high profile research on deep reinforcement learning and biologically plausible neural network learning algorithms. His research at DeepMind focuses on the interface between reinforcement learning and memory.

Greg Wayne, a senior Research Scientist at DeepMind, holds a Ph.D. in Neurobiology from Columbia University. He has conducted high profile work in hierarchical planning and deep RL. His research at DeepMind focuses on integrative architectures for artificial intelligence.
Daan Wierstra is a senior Research Scientist and research team leader at DeepMind. He holds a Ph.D. in Artificial Intelligence from IDSIA, and has conducted high-profile research in RL and neural networks. His research at DeepMind focuses in part on computational models of imagination and planning.

## Presentations

As noted above, the tutorial will be comprised of five lectures, each covering a key area.

## Deep Reinforcement Learning (Tim Lillicrap)

Advances in deep RL have driven some of the highestprofile recent work in AI, including differentiable neural computers (Graves et al., 2016) and superhuman play in the game of go (Silver et al., 2016). This talk will provide a tutorial review of the cutting edge in deep RL research.

## Memory Architectures (Greg Wayne)

A major development over the past couple of years has been the incorporation of special modules for memory storage into deep learning AI systems (e.g., Graves et al. 2016; Santoro et al., 2016). This talk will review the state of the art, and consider the relationship with human episodic and working memory.

## Structured Models for Structured Domains (Peter Battaglia)

Recent work in AI has introduced structure into deep learning architectures, which biases such systems toward particular forms of representation. Such measures have allowed dramatic advances in modeling physical systems and other structured domains (e.g., Battaglia et al., 2016). The present talk will review this approach and discuss its relation to the notion of compositional representation in cognitive science.

## Deep Learning to Learn (Matthew Botvinick)

Recent work has explored the capacity of deep learning systems to 'learn how to learn,' leveraging previous experience to adapt more quickly to new challenges (e.g., Wang et al., 2016). This lecture will review recent progress toward endowing deep learning systems with this important capacity.

## Deep Generative Models (Daan Wierstra)

One of the most exciting developments in recent deep learning research has been the rapid progress in building rich and flexible generative models, models that support operations like imagination and forecast-based planning (e.g., Gregor et al., 2015). This lecture will review the most
recent techniques for building generative models, considering their many implications for cognitive science.

## Related events

Yildirim and colleagues will present a half-day workshop focusing on the interface between recent AI advances and cognitive science. This can be considered a companion to the present tutorial, in that it will also cover techniques other than deep learning/RL, and will more deeply explore applications to cognitive and neuroscience.

## References

Battaglia, P., Pascanu, R., Lai, M., \& Rezende, D. J. (2016). Interaction networks for learning about objects, relations and physics. In Advances in Neural Information Processing Systems (pp. 4502-4510).
Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... \& Badia, A. P. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626), 471-476.
Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., \& Wierstra, D. (2015). DRAW: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., \& Gershman, S. J. (2016). Building machines that learn and think like people. arXiv preprint arXiv:1604.00289.
Marblestone, A. H., Wayne, G., \& Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. Frontiers in Computational Neuroscience, 10.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... \& Petersen, S. (2015). Humanlevel control through deep reinforcement learning. Nature, 518(7540), 529-533.
Rezende, D. J., Mohamed, S., Danihelka, I., Gregor, K., \& Wierstra, D. (2016). One-shot generalization in deep generative models. arXiv preprint arXiv:1603.05106.
Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., \& Lillicrap, T. (2016). One-shot learning with memoryaugmented neural networks. arXiv preprint arXiv:1605.06065.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... \& Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.
Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... \& Botvinick, M. (2016). Learning to reinforcement learn. arXiv preprint arXiv:1611.05763.

# Methods for Reconstructing Causal Networks from Observed Time-Series: Granger-Causality, Transfer Entropy, and Convergent Cross-Mapping 

Fermín Moscoso del Prado Martín (fmoscoso@linguistics.ucsb.edu)<br>Department of Linguistics, South Hall 3521, UCSB<br>Santa Barbara, CA 93106 USA


#### Abstract

Keywords: convergent cross-mapping; development, dynamical systems; Granger-causality; information theory; phasespace reconstruction; time series


## Objectives and Scope

A major question that arises in many areas of Cognitive Science is the need to distinguish true causal connections between variables from mere correlations. The most common way of addressing this distinction is the design of wellcontrolled experiments. However, in many situations, it is extremely difficult -or even outright impossible- to perform such experiments. Researchers are then forced to rely on correlational data in order to make causal inferences. This situation is especially common when one needs to analyze longitudinal data corresponding to historical time-series, symbolic sequences, or developmental data. These inferences are often very problematic. From the correlations alone it is difficult to determine the direction of the causal arrow linking two variables. Worse even, the lack of controls of observational data entail that correlations found between two variables need not reflect any causal connection between them. The possibility always remains that some third variable which the researchers were not able to measure, or were actually unaware of, is the actually driver for both measured variables, giving rise to the mirage of a direct relationship between them.

In recent years, it has been shown that, under particular circumstances, one can use correlational information for making sound causal inferences (cf., Pearl, 2000). In this tutorial I will provide a hands-on introduction to the use of modern causality techniques for the analysis of observational time series. I will cover causality analyses for three types of time-series that are often encountered in Cognitive Science research:

- For numerical time-series of a predominantly stochastic nature I will discuss how to perform Granger-Causality (Granger, 1981) analyses used by econometricians, using the methodology introduced by Toda and Yamamoto (1995).
- For symbolic stochastic time-series, I will introduce the Transfer Entropy measure developed in Physics (Schreiber, 2000).
- Finally, for numerical series that can be shown to have a predominantly deterministic (even if possibly chaotic) nature, I will discuss Convergent Cross-Mapping (Clark et al., 2015; Sugihara et al., 2012), a very powerful technique recently developed in the field of ecology, that relies on the Theory of Dynamical Systems to make causal inferences.

Finally, I will demonstrate how to use each of these techniques for reconstructing networks of causal relations between large sets of variables.

## Overview of Causality Methods for Time Series

 Granger-CausalityGranger-Causality (cf. Granger, 1981) is a powerful technique developed in Econometrics for assessing whether one time sequence can be said to be the cause of another one (or viceversa). If $x$ and $y$ are stationary time sequences on discrete time $(\tau)$, in order to test whether $x$ Granger-causes $y$, one tests whether the past of $x$ is able to predict the future of $y$, over and above the predictive power that can be obtained from y's own past. Between just two variables, this is assessed using Autoregressive Models. When more than two variables are involved this is naturally extended by using Vector Autoregressive Models.

This technique is useful for the analysis of numerical time series data that are generated by a process whose nature is predominantly stochastic, which is typical of data resulting from the aggregation of multiple sources. One important requirement of the Granger-causality method is that it is limited to stationary time series. This property is sometimes difficult to guarantee in the types of series that one typically encounters in Cognitive Science, which tend to exhibit a certain degree of co-integration. This limitation can, however, be addressed by using the methodology introduced by Toda and Yamamoto (1995), which I will introduce in the tutorial.

## Transfer Entropy

Often, in Cognitive Science, researchers need to analyze sequences of discrete symbols, as could for instance be the sounds uttered by a developing child. Schreiber (2000) extended the main idea of Granger-Causality to symbolic stochastic processes. Instead of analyzing correlations between variables, one moves into using mutual informations, in the sense of Shannon (1948). Even when one's data are actually of a numerical nature, it can be actually beneficial to analyze them in symbolic terms, as the mutual informations are capable of capturing non-linear relations that could be missed by linear correlation-based methods (see, HlaváčcovaSchindler, Paluš, Vejmelka, \& Bhattacharya, 2007, for a review). This later usage will be demonstrated in the tutorial using a dataset from human speech.

## Convergent Cross-Mapping

The Granger-Causality and Transfer-Entropy approaches outlined above are suitable only for stochastic systems. In some

Cognitive Science domains, especially those dealing with longitudinal developmental data, one also encounters numerical data that can be argued to originate from a predominantly deterministic Dynamical System. Such cases can be modelled explicitly using systems of coupled differential equations. However, in many cases, only a few of the variables that relevant for the system are available to the researchers (the others being unmeasurable or plainly unknown). However, Takens' Theorem (Takens, 1981) states that the crucial properties of a dynamical system's attractor can be succesfully recovered using a single one of its variables, in what is known as Phase-State Reconstruction. Using this fact, Sugihara et al. (2012) developed the Convergent-Cross Mapping (CCM) technique, which enables recovering the direction of causality between any two time sequences generated by the same dynamical system. Importantly, and in contrast with the methods discussed above, CCM is also capable of distinguishing the case when two correlated variables are not actually causally related, but rather they are both driven by a third unstudied variable. A limitation of CCM is that it requires relatively long time series, which are often unavailable in many actual research problems. Clark et al. (2015) extended CCM to allow combining multiple short time series originating from similar processes (i.e., as if considering random effects in a regression model), introducing the "multispatial" variant of the CCM method. I will demonstrate how the multispatial CCM analyses can be performed.

## Format and Organization

This tutorial is designed to cover half a day (three hours) broken into two sections of 1.5 hours each. The first session in the tutorial will discuss the theoretical basis, conditions of applicability, advantages, drawbacks of each of the three causal analysis methods. The second session will be hands-on, guiding attendants on how to perform each of these analyses, together with the necessary diagnostics, using the R statistical software. For this, I will make use of previously published datasets, covering three different timescales: historical (Moscoso del Prado Martín, 2014), developmental (Irvin, Spokoyny, \& Moscoso del Prado Martín, 2016), and the time-scale of a typical behavioral experiment (Moscoso del Prado Martín, 2011).

## Target Audience

The tutorial is aimed at advanced graduate students, postdocs, and senior researchers wishing to use this type of causal analyses in this research. A familiarity with basic statistics and with programming (preferably using R ) will be necessary to be able to follow the theoretical arguments, and to be able to perform the analyses.

## Tutor Information

Fermín Moscoso del Prado Martín is assistant professor of Linguistics at the University of California, Santa Barbara. Previously he has held positions at the Max Planck Institute
for Psycholinguistics, the Medical Research Council - Cognition and Brain Sciences Unit, and at the Cognitive Psychology Laboratory of the French National Research Center. He holds an MEng in Computer Science by the Technical University of Madrid, and a PhD in Linguistics by the University of Nijmegen, where he was a student of Prof. R. H. Baayen. Over the last decade he has published multiple papers combining information-theoretical methods, computational modeling, corpus analyses, and psycholinguistic experiments.

## References

Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., \& Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. Ecolology, 96, 1174-1181.
Granger, C. W. J. (1981). Testing for causality. Journal of Economic Dynamics and Control, 2, 329-352.
Hlaváčcova-Schindler, K., Paluš, M., Vejmelka, M., \& Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. Physics Reports, 441, 1-46.
Irvin, J., Spokoyny, D., \& Moscoso del Prado Martín, F. (2016). Dynamical systems modeling of the child-mother dyad: Causality between child-directed language complexity and language development. In A. Papafragou, J. Trueswell, D. Grodner, \& D. Mirman (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Moscoso del Prado Martín, F. (2014). Grammatical change begins within the word: Causal modeling of the coevolution of Icelandic morphology and syntax. In P. Bello, M. Guarini, M. McShane, \& B. Scasselatti (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 2657-2662). Austin, TX: Cognitive Science Society.
Moscoso del Prado Martín, F. (2011). Causality, criticality, and reading words: distinct sources of fractal scaling in behavioral sequences. Cognitive Science, 35, 785-837.
Pearl, J. (2000). Causality: models, Reasoning and Inference. Cambridge, England: Cambridge University Press.
Schreiber, T. (2000). Measuring information transfer. Physical Review Letters, 85, 461-464.
Shannon, C. E. (1948). A mathematical theory of communication. Bell Systems Technical Journal, 27, 379-423, 623656.

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E. R., Fogarty, M., \& Munch, S. (2012). Detecting causality in complex ecosystems. Science, 338, 496-500.
Takens, F. (1981). Detecting strange attractors in turbulence. In D. A. Rand \& L.-S. Young (Eds.), Dynamical Systems and Turbulence (pp. 366-381). Berlin, Germany: Springer Verlag.
Toda, H., \& Yamamoto. (1995). Statistical inference in Vector Autoregressions with possible integrated sources. Journal of Econometrics, 66, 225-250.

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

John P. Spencer (j.spencer@uea.ac.uk)<br>School of Psychology, University of East Anglia, Norwich NR4 7TJ, United Kingdom<br>Vanessa R. Simmering (simmering @ wisc.edu)<br>Department of Psychology, University of Wisconsin - Madison, 1202 W. Johnson Street, Madison, WI 53706 USA<br>Sebastian Schneegans (ss2361@cam.ac.uk)<br>Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom

Keywords: computational model; dynamic systems; neural field model

## Objectives and Scope

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

One obstacle for researchers wishing to use DFT has been that the mathematical and technical skills required to make these concepts operational are not part of the standard repertoire of cognitive scientists. The goal of this tutorial is to provide the training and tools to overcome this obstacle. We will provide a systematic introduction to the central concepts of DFT and their grounding in both Dynamical Systems concepts and neurophysiology. We will discuss the concrete mathematical implementation of these concepts in Dynamic Neural Field models, giving all needed background and providing participants with some hands-on experience using interactive simulators in MATLAB. Finally, we will take participants through a number of selected, exemplary case studies in which the concepts and associated models have been used to ask questions about elementary forms of embodied cognition and their development.

A newly published book on Dynamic Neural Field modeling, Dynamic Thinking: A Primer on Dynamic Field Theory, covers these topics and more, with interactive simulators available to give hands-on experience to readers. We will take participants through the process of building and simulating models to illustrate key concepts in the case studies we describe in the tutorial.

## Suggested Readings

(available online, see Online Resources below)

1. Schöner, G. \& Spencer, J.P. (2015). Introduction. [Chapter 1] In G. Schöner, J. P. Spencer, \& the DFT Research Group (Eds.), Dynamic Thinking: A Primer on Dynamic Field Theory. New York, NY: Oxford University Press.
2. Schöner, G., Reimann, H., \& Lins, J. (2015). Neural Dynamics. [Chapter 2] In G. Schöner, J. P. Spencer, \& the DFT Research Group (Eds.), Dynamic Thinking: A Primer on Dynamic Field Theory. New York, NY: Oxford University Press.
3. Schöner, G. \& Schutte, A.R. (2015). Dynamic Field Theory: Foundations. [Chapter 3] In G. Schöner, J. P. Spencer, \& the DFT Research Group (Eds.), Dynamic Thinking: A Primer on Dynamic Field Theory. New York, NY: Oxford University Press.
4. Johnson, J.S. \& Simmering, V.R. (2015). Integrating Perception and Working Memory in a Three-Layer Dynamic Field Architecture. [Chapter 5] In G. Schöner, J. P. Spencer, \& the DFT Research Group (Eds.), Dynamic Thinking: A Primer on Dynamic Field Theory. New York, NY: Oxford University Press.
5. Simmering, V.R. \& Schutte, A.R. (2015). Developmental Dynamics: The Spatial Precision Hypothesis. [Chapter 10] In G. Schöner, J. P. Spencer, \& the DFT Research Group (Eds.), Dynamic Thinking: A Primer on Dynamic Field Theory. New York, NY: Oxford University Press.
6. Sandamirskaya, Y., Zibner, S., Schneegans, S., \& Schöner, G. (2013). Using Dynamic Field Theory to Extend the Embodiment Stance toward Higher Cognition. New Ideas in Psychology.

## Target Audience

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

## Schedule of Material Covered in the Tutorial

1. Conceptual foundations of Dynamical Systems Thinking and Dynamical Field Theory (DFT) - 30 minutes: embodied and situated cognition; stability as a necessary property of embodied cognitive processes; distributions of population representation as the basis of spatially and temporally continuous neural representations.
2. Dynamical Systems and Dynamic Field Theory Tutorial 90 minutes: concept of dynamical system; attractors and stability; input tracking; detection, selection, and memory instabilities in discrete neuronal dynamics; Dynamical Fields and the basic instabilities: detection, selection, memory, boost-driven detection; learning dynamics; categorical vs. graded mode of operation; practical implementation of DFT in simulators; interactive simulation; illustration of the ideas through robotic implementations;
3. Case study using DFT to understand brain-behavior relations in humans with functional neuroimaging - 60 minutes: mapping of neural activation patterns in dynamic neural fields to the hemodynamic response measured with fMRI and fNIRS; case study on the neural processes that underlie visual working memory in children and adults.
4. Case study using DFT to understand embodied cognition and its development - 90 minutes: visual and spatial working memory in infants, children, and adults; spatial precision hypothesis as a developmental mechanism visuospatial cognition
5. Case study using DFT to understand higher cognition 90 minutes: integrating location and feature information to form working memory representations of visual scenes; linking spatial language to visual perception

## Lecturers

John P. Spencer is a Professor of Psychology at the University of East Anglia in Norwich, UK. Prior to arriving in the UK, he was a Professor of Psychology at the University of Iowa and served as the founding Director of the Delta Center (Development and Learning from Theory to Application). He received a Sc.B. with Honors from Brown University in 1991 and a Ph.D. in Experimental Psychology from Indiana University in 1998. He is the recipient of the Irving J. Saltzman and the J.R. Kantor Graduate Awards from Indiana University. In 2003, he received the Early Research Contributions Award from the Society for Research in Child Development, and in 2006, he received the Robert L. Fantz Memorial Award from the American Psychological Foundation. His research examines
the development of visuo-spatial cognition, spatial language, working memory, and attention, with an emphasis on dynamical systems and neural network models of cognition and action. He has had continuous funding from the National Institutes of Health and the National Science Foundation since 2001 and has been a fellow of the American Psychological Association since 2007. He will lecture on the topics 1-3 above.
Vanessa R. Simmering is an Assistant Professor of Psychology at the University of Wisconsin - Madison. She obtained her B.S. with Honors in Psychology from the University of Iowa in 2001 and a Ph.D. in Psychology from the University of Iowa in 2008. Her research takes a dynamic systems approach to understanding cognition and development, with particular emphasis on how visuospatial cognition relates to other developing skills during early childhood. She will lecture on topic 4 above.
Sebastian Schneegans is a postdoctoral researcher in the Department of Psychology at the University of Cambridge. He obtained his PhD (Dr.-Ing.) at the Institut für Neuroinformatik, Ruhr-Universität Bochum, for his work on visual working memory, spatial cognition and spatial language within the framework of DFT. He is now developing population code models of visual working memory and testing them in psychophysical experiments. His work has been published in seven journal articles, six book chapters, and numerous conference contributions. Dr. Schneegans will lecture on mechanisms for feature binding and spatial transformations in DFT. He will lecture on topic 5 above.

## Computer Use

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

## Online Resources

We will use simulators from the free Matlab toolbox Cosivina for demonstrations. Installation instructions and documentation for the toolbox can be found on the website. Related publications, lecture material, and interactive simulators can all be found at our website, http://www.dynamicfieldtheory.org/.

# The Fine Art of Conversation 

Saul Albert (saul.albert @tufts.edu)<br>Department of Psychology, Tufts University<br>Medford, MA 02155 USA

Claude Heath (claude.heath@rhul.ac.uk)
Royal Holloway, University of London,
Egham, Surrey, UK
Patrick Healey (p.healey@qmul.ac.uk)
School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK


#### Abstract

This workshop is aimed at giving human interaction researchers the conceptual and practical apparatus to balance their representations of data (mixes of drawings and photographs in the most part), so as to "maximally incite, but also constrain" their representations, just as artists sometimes succeed in doing (Streeck, Grothues, \& Villanueva, 2009, p.28). Why-as Streeck points out-are the drawings and visualisations of interaction researchers so halting and timid, compared to the ways artists have responded to the same kinds of representational problems? Are these heavily segmented and sparsely constructed representations of interaction the result of a prevailing positivistic outlook with regard to representing shared space, where interaction is presented as staggered and discrete physical events with apparently little to connect them. The workshop seeks to redress this situation by examining the solutions that artists have arrived at when representing human interaction, and asking participants to engage in a series of activities and discussions which will re-frame their approaches to this issue.


Keywords: cognitive science; arts; interaction; drawing; embodiment; creativity; representation; comics; art; film; photography

## Depicting Human Interaction

Detailed representations focusing on social interaction in fine art are surprisingly rare. Where they do occur, they reveal something about the artist's conception of communication and their (possibly implicit) theories about how these representations are percieved and processed. Similar issues attend the representation of interaction in the Cognitive Sciences (C. P. R. Heath, 2014). Researchers have developed a diverse range of specialized methods for describing interaction; from graphic transcripts including photographs mixed with line drawings showing joint action and embodiment (Laurier, 2014), to coupled representations of the patterns of neural activation during social interaction in the brains of participants (Dumas, Nadel, Soussignan, Martinerie, \& Garnero, 2010). For thousands of years drawing, mapping, diagramming, and other forms of visual notation have been key methods for transmitting human knowledge and culture, and line drawing has been a particularly salient and widespread form of visual communication (Craig-Martin \& Martin, 1995).

None of these representations are innocent. Drawing itself is a kind of transcription, encoding our own analytical assumptions about what we see as relevant to a reader or
viewer (Ochs, 1979). A wide range of commonalities between drawing systems have evolved, such as the practice of perspective drawing, and each system has developed its own rationale, method, and objects of enquiry (Dubery \& Willats, 1972). Importantly, these approaches to representation are rarely scrutinised as empirical methods.

It is not surprising, then, that artists have become living repositories of expertise in the practice of drawing systems of all kinds, and as such have become a valued resource (Kozbelt, 2001). Visual reasoning can be examined in the ways that drawings are commonly constructed, (Van Sommers, 1984), and the ways in which novel drawing situations are spoken of and acted upon (C. P. Heath, Cameron, \& Cain, 2008). Added to this, drawing and diagramming have become standard tools in the repertoire of participatory action research (Chambers et al., 1997) applied in diversely situated engagements, (Theron, Mitchell, \& Smith, 2011). The process of interpreting visual depictions in itself is an intersubjective phenomenon, where particular methods and practices of reading determines the consequential social meanings and practical uses of the inscription (Goodwin, 2000).

This workshop will bring together a range of perspectives from the cognitive sciences and the arts, asking whether our long legacies of drawing systems in artistic and scientific representations is telling us something about our varied approaches to mind, intersubjectivity and social interaction. Researchers explore very different phenomena of interest, often using highly specialized research methods particular to their subdomain within the cognitive sciences. Many of these are represented in the list of prospective participants in the workshop which we are submitting with this outline. The proposed workshop aims to encourage new opportunities for dissemination and collaboration within and beyond the cognitive sciences, scrutinising the received and conventional methods of depicting human interaction as a starting point for conversation and exchange.

## Goals and workshop plan

Background The organizers have participated in and run workshops, specialist conferences and presentations bridging cognitive science, human interaction and the arts, including
at the 2014 meeting of the Society ${ }^{1}$. This workshop aims to: a). build upon and extend the networks of researchers established at prior meetings, and to use the clear focus on methods of depicting human interaction as a binding theme to draw together and engage the widest possible range of fields and approaches; b). to provide an entry point into the themes and discussions for a broader audience within Cognitive Science; c). to encourage cross-disciplinary researchers and especially those just starting out in their research to collaborate with others, seeking out connection points, looking at how human interaction and communication is depicted in their fields.

Format In order to create and sustain a broadened interest in the workshop, a blog page will be set up in advance so that invitees, workshop participants and others may browse examples of artworks that have already stimulated discussion amongst the organisers and can contribute their own. The blog will also provide a centralised focus for the workshop and is intended to test and support be the impetus for an illustrated publication. The central function of the blog will be gather and examine cases in which fine art has succeeded (or failed in interesting ways) to create credible depictions of interaction. Discussions on the blog both before and after the workshop will enable participants to contribute to threaded discussions alongside each artwork (link to follow on notification of acceptance of proposal).

Practical workshop activities will centre on previously worked-up examples of interactional depictions, and will use questions and criteria suggested in the workshop presentations. Activities will involve drawing out points and patterns of interest from artworks using projections onto whiteboards and screens, the results of which can later be digitally overlaid onto source artworks and can be documented on the blog. A 3-step line drawing protocol for drawing out shared interactional spaces using field inscriptions will also be used and adapted to the available 'data' -in this case the artworks under discussion.

An overview at the end of the day will compare the outputs of different groups, and will conclude with a group visit to a London gallery to look at and discuss prime examples of the kinds of artworks and drawn phenomena encountered during the day.

## Planning committee

The planning committee consists of researchers who have been working together on related projects in human interaction research, cognitive science, psychology and the arts. Each will give a short overview presentation in order to frame the subsequent activities and discussions.

- Claude Heath, "Drawing out interaction"
- Patrick Healey, "What's so difficult about drawing interaction?"
- Saul Albert, "Representing unformulated action"

[^0]The following list of participants (partial here, since numbers who have expressed interest are growing) comprises cognitive science researchers who have dealt with depictions of human interaction in their research. It also includes scientists from other fields, and artists who can bring to their highly relevant interests and methodological approaches to the crossdisciplinary objectives of the workshop.

Toby Harris (QMUL), Daniel G. Jay (Tufts), Sophie Skach (QMUL), Rosella Paulina Galindo Esparza (QMUL), Shauna Concannon (QMUL), Lida Theodorou (QMUL), Leshao Zhang (QMUL), Melissa Bliss (QMUL), Nicola Jane Plant (QMUL), Soomi Park (QMUL), Saul Albert (Tufts), Christian Heath (KCL), Yal Kreplak (EHESS), Dirk vom Lehn (King's), Jrgen Streeck (U. Texas), Eric Laurier (Edinburgh), J.P. De Ruiter (Tufts), Michael Sean Smith (UCLA), Elizabeth Stokoe \& A Dozen Eggs (Loughborough),

## References

Chambers, R., et al. (1997). Whose reality counts?: Putting the first last. Intermediate Technology Publications (ITP).
Craig-Martin, M., \& Martin, M. C. (1995). Drawing the line: Reappraising drawing past and present. South Bank Centre, London.
Dubery, F., \& Willats, J. (1972). Drawing systems. Studio Vista, London.
Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., \& Garnero, L. (2010, \#jan\#). Inter-brain synchronization during social interaction. PloS one, 5(8), e12166. doi: 10.1371/journal.pone. 0012166

Goodwin, C. (2000). Practices of seeing visual analysis: An ethnomethodological approach. In T. van Leeuwen (Ed.), Handbook of visual analysis (pp. 157-182). London: Sage Publications.
Heath, C. P., Cameron, L., \& Cain, P. (2008). The practice of three-dimensional drawing.
Heath, C. P. R. (2014). Drawing out conversation: lines around shared space (Unpublished doctoral dissertation). Queen Mary, University of London.
Kozbelt, A. (2001). Artists as experts in visual cognition. Visual Cognition, 8(6), 705-723.
Laurier, E. (2014, \#apr\#). The graphic transcript: Poaching comic book grammar for inscribing the visual, spatial and temporal aspects of action. Geography Compass, 8(4), 235-248. doi: 10.1111/gec3.12123
Ochs, E. (1979). Transcription as theory. In E. Ochs \& B. B. Schieffelin (Eds.), Developmental pragmatics (pp. 43-72). New York: Academic Press.
Streeck, J., Grothues, J., \& Villanueva, J. (2009). Gesturecraft: The manu-facture of meaning (Vol. 2). John Benjamins Publishing Company.
Theron, L., Mitchell, C., \& Smith, A. (2011). Picturing research: Drawing as visual methodology. Springer.
Van Sommers, P. (1984). Drawing and cognition: Descriptive and experimental studies of graphic production processes. Cambridge University Press.

# Bridging the Gap: Is Logic and Automated Reasoning a Foundation for Human Reasoning? 

Ulrich Furbach (uli@uni-koblenz.de)<br>Department of Informatik, University of Koblenz-Landau, Koblenz, Germany<br>Steffen Hölldobler (sh@iccl.tu-dresden.de)<br>International Center for Computational Logic, TU Dresden, Dresden, Germany<br>Marco Ragni (ragni@cs.uni-freiburg.de)<br>Cognitive Computation Lab, University of Freiburg, Freiburg, Germany<br>Claudia Schon (schon@uni-koblenz.de)<br>Department for Web Science and Technologies (WeST), University of Koblenz-Landau, Koblenz, Germany

Keywords: reasoning processes; logic; mental representation; cognitive reasoning theories; automated deduction

## Introduction

Reasoning is a core ability in human cognition. Its power lies in the ability to theorize about the environment, to make implicit knowledge explicit, to generalize given knowledge and to gain new insights. It is a well researched topic in cognitive psychology and cognitive science and over the past decade impressive results have been achieved. Early researchers starting with Störing (1908) often used propositional logic as a normative framework. Any deviation from it has been considered an error. Central results like findings from the Wason selection task (Wason, 1968) or the suppression task (Byrne, 1989) inspired a shift from propositional logic and the assumption of monotonicity in human reasoning towards other reasoning approaches. This includes but is not limited to models using probabilistic approaches (Oaksford \& Chater, 2007), mental models (Johnson-Laird, 2006), or nonmonotonic logics (Stenning \& Lambalgen, 2008). Considering cognitive theories for syllogistic reasoning show that none of the existing theories is close to the existing data (Khemlani \& Johnson-Laird, 2012). But some formally inspired cognitive complexity measures can predict human reasoning difficulty for instance in spatial relational reasoning (Ragni \& Knauff, 2013).

Automated deduction, on the other hand, is mainly focusing on the automated proof search in logical calculi. And indeed there is tremendous success during the last decades. Recently a coupling of the areas of cognitive science and automated reasoning is addressed in several approaches. For example there is increasing interest in modeling human reasoning within automated reasoning systems including modeling with answer set programming, deontic logic or abductive logic programming (Dietz \& Hölldobler, 2015; Dietz, Hölldobler, \& Wernhard, 2014). There are also various approaches within AI research for common sense reasoning (Furbach \& Schon, 2014, 2016).

Despite a common research interest - reasoning - there are still several milestones necessary to foster a better inter-
disciplinary research. First, to develop a better understanding of methods, techniques, and approaches applied in both research fields. Second, to have a synopsis of the relevant state-of-the-art in both research directions. Third, to combine methods and techniques from both fields and find synergies. E.g., techniques and methods from computational logic have never been directly applied to model adequately human reasoning. They have always been adapted and changed. Fourth, we need more and better experimental data that can be used as a benchmark system. Fifth, cognitive theories can benefit from a computational modeling. Hence, both fields - human and automated reasoning - can both contribute to these milestones and are in fact a conditio sine qua non. Achievements in both fields can inform the others. Deviations between fields can inspire to seek a new and profound understanding of the nature of reasoning.

This is the third workshop in a series of successful Bridging the Gap workshop ${ }^{1}$ located at previous conferences: 2015 at the International Conference on Automated Deduction in Berlin (CADE-25) focused on the automated reasoning aspects. 2016 at the International Conference on Artificial Intelligence in New York (IJCAI 2016) included an AI perspective. The Annual Meeting of the Cognitive Science Society is the central place for bringing a strong human centric perspective into discussion.

## Goal and Scope

The goal of this workshop is to bring together leading researchers from cognitive science, computational logics, and psychology interested in computational foundations of human reasoning - both as speakers and as audience members. Its ultimate goal is to share knowledge, discuss open research questions, and inspire new paths. Like its preceding event, it is intended to get an overview of existing approaches and make a step towards a cooperation between computational logic and cognitive science. Topics of interest include, but are not limited to the following:

- Benchmark problems relevant in both fields

[^1]- limits and differences between automated and human reasoning
- psychology of deduction and common sense reasoning,
- logics modeling human reasoning
- non-monotonic, defeasible, and classical reasoning

The workshop is planned as a half-day event. There will be an invited speaker, sponsored by IFIP TC 12 (this is why this Workshop would have the "(supported by IFIP TC 12)" in its announcement.

## Workshop Organization

Ulrich Furbach is a Senior Research Professor of Artificial Intelligence at the University of Koblenz. His research interests include knowledge management, automated reasoning, multi-agent systems, and e-learning. He is co-founder of the spin-off company wizAI (www.wizai.com), which develops knowledge management systems and information extraction tools. Steffen Hölldobler is professor for Knowledge Representation and Reasoning at the Technical University Dresden. He is currently Director of the International Center for Computational Logic and co-ordinator of the European Master's Program in Computational Logic. He is particularly interested in combining methods and techniques from computational logic and cognitive science to adequately model human reasoning and to develop connectionist systems for human reasoning. Marco Ragni is a DFG-Heisenberg fellow and associate professor at the technical faculty and the Center for Cognitive Science of the Albert-Ludwigs-University Freiburg and associated with Department. His research interests include qualitative spatio-temporal reasoning, knowledge representation and reasoning, cognitive modelling, and complex cognition with a special focus on analyzing why and how human reasoning often deviates from classical logical approaches. Claudia Schon is a postdoctoral researcher at the Institute for Web Science and Technologies at the University of Koblenz-Landau. During the last years, she was working in various projects in the area of artificial intelligence. One of these projects was the RATIOLOG project were she focused her research on commonsense reasoning and modeling human deduction.

## Target Audience

Our specific focus dovetails this years overall conference theme: "Computational Foundations of Cognitive Science". Hence, the target audience for this workshop overlaps significantly with the target audience of Cognitive Science conference. The workshops central topics (psychology of deduction, common sense reasoning, logic, non-monotonic reasoning, formal systems) are core topics of Cognitive Science with the multidisciplinary nature of the workshop being particularly appropriate for the multidisciplinary Cognitive Science conference.

## Confirmed Speakers

- E-A. Dietz Saldanha, TU Dresden, Germany
- S. Hölldobler, TU Dresden, Germany
- S. Khemlani, Naval Research Lab, USA
- B. Kowalski Imperial College London, GB
- A. Kakas, University Cyprus, Cyprus
- L. Pereira, Universidade Nova Lisboa, Portugal
- M. Ragni, University of Freiburg, Germany


## References

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. Cognition, 31, 61-83.
Dietz, E.-A., \& Hölldobler, S. (2015). A new computational logic approach to reason with conditionals. In F. Calimeri, G. Ianni, \& M. Truszczynski (Eds.), Logic programming and nonmonotonic reasoning, 13th international conference, lpnmr (Vol. 9345, p. 265-278). Springer.
Dietz, E.-A., Hölldobler, S., \& Wernhard, C. (2014). Modelling the suppression task under weak completion and well-founded semantics. Journal of Applied NonClassical Logics, 24, 61-85.
Furbach, U., \& Schon, C. (2014). Deontic logic for human reasoning. In T. Eiter, H. Strass, M. Truszczynski, \& S. Woltran (Eds.), Advances in knowledge representation, logic programming, and abstract argumentation essays dedicated to gerhard brewka on the occasion of his 60th birthday (Vol. 9060, pp. 63-80). Springer.
Furbach, U., \& Schon, C. (2016). Commonsense reasoning meets theorem proving. In M. Klusch, R. Unland, O. Shehory, A. Pokahr, \& S. Ahrndt (Eds.), Multiagent System Technologies - 14th German Conference, MATES 2016, Klagenfurt, Österreich, September 2730, 2016. Proceedings (Vol. 9872, pp. 3-17). Springer.
Johnson-Laird, P. N. (2006). How we reason. New York: Oxford University Press.
Khemlani, S., \& Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. Psychological Bulletin, 427-457.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. Oxford: Oxford University Press.
Ragni, M., \& Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. Psychological Review, 120(3), 561-588.
Stenning, K., \& Lambalgen, M. (2008). Human reasoning and cognitive science. Cambridge, MA: MIT Press.
Störing, G. (1908). Experimentelle Untersuchungen über einfache Schlussprozesse. W. Engelmann.
Wason, P. C. (1968). Reasoning about a rule. Quarterly Journal of Experimental Psychology, 20, 273-281.

# Citizen Science, Gamification, and Virtual Reality for Cognitive Research 

Jana B. Jarecki (jana.jarecki@unibas.ch, jj@janajarecki.com)<br>Department of Economic Psychology, University of Basel, Missionsstrasse 62A, 4055 Basel, Switzerland

Keywords: citizen science; volunteer science; experimental design, gamification; online experimentation, virtual reality; problem solving

## Introduction

This workshop discusses three distinct but related topics. The first topic is citizen science which involves volunteers all around the world generating data to address scientific problems and recently led to breakthroughs in the natural sciences. Citizen science typically involves volunteers playing online games while unknowingly solving real scientific problems. This approach can benefit cognitive research either indirecty - if volunteers implicitly solve computationally hard research problems disguised behind a game - or it benefits researchers directly - if players do experimental tasks, with the key benefit that online games are easily accessible for a diverse international subject pool. Also, the data from existing citizen science projects provide insights for cognitive scientists (e.g., one citizen science project involved classification of galaxies). Although citizen science data is limited with respect to precise experimental control, its benefits are worth discussing.
Our second topic is gamification, which citizen science often relies on. Gamification, in general, involves adding gamelike features to a task and could involve, for example, adding levels, points, or virtual characters to an experimental task. Gamified tasks are typically motivating for participants. While citizen science often uses gamification, also traditional laboratory studies can be supplemented with game-like elements and use technology from the gaming industry.
Our third topic is virtual reality (VR). This technology developed by the gaming industry enables players, equipped with a headset, to experience a controlled 3-dimensional environment that emulates being in the middle of a 3D scenario. Virtual reality typically causes subjective immersion by simulating a naturalistic experience of interacting with the world, while simultaneously offering full experimental control about the environmental structure and interactions. The link between (large-scale online) citizen science and (smaller-scale lab-based) VR research lies in that usually both approaches use gamification
How can cognitive scientists use gamification and virtual reality environments for their research?

## Goals and Scope

This workshop brings together experts from the natural and social sciences who have successfully launched citizen science platforms as well as gaming researchers and game developers from different disciplines. The aim is to introduce
citizen science to cognitive psychologists, stimulate a discussion about similarities and differences between laboratory and citizen science data, and explore the potential of existing citizen science platforms and data for cognitive research. A further goal is to gain insights into how best to conduct citizen science projects using gamification. Moreover, the workshop will explore the potential of virtual reality environments, fleshing out the challenges and opportunities of this novel technological opportunity for research. By the end of the workshop, all participants and speakers, will have been introduced to citizen science platforms, including how and where to run citizen science projects. They will understand the advantages and drawbacks of games for cognitive research, including an accessible way to design online applications. Lastly, the participants will have the knowledge to understand when and how virtual reality environments can be used to answer cognitive research questions in novel ways.

## Target Audience

The workshop targets researchers at all levels with an interest in investigating domains like problem solving, learning, attention, or decision making in settings suitable for interactive, gamified, or web studies. Note, that citizen science cannot offer maximum experimental control. Research interested in gamification in any form-offline or online-with a special focus on gamified experimental design and virtual reality are welcome. In particular researchers interested in cross-cultural data may be interested, as citizen science offers unique opportunities for international data collection. We also invite researchers interested in using virtual reality devices in the lab, and discussing its potential for cognitive research.

## Format

The full-day workshop involves short talks, hands-on experience with several of the games developed by the invited speakers, and the possibility to try out a task in virtual reality (using a HTC Vive). There will be two round table discussions. The first discussion focuses on the potentials and disadvantages of gamification and online data compared to laboratory data; the second discussion asks which of cognitive sciences' research problems can be fruitfully advanced by implementing the study within a virtual reality environment.

## Contributions

## Introduction of Concepts

Jana B. Jarecki, cognitive scientist at Basel University, will introduce the concepts citizen science and gamification.

Jacob Sherson, who has pioneered citizen science in quantum physics, will discuss how the citizen science platform ScienceAtHome, aims to exploit the clear mathematical framework of quantum physics and other natural science research challenges. This enables the construction of a suite of games bridging low-dimensional model challenges with complex relevant problem solving. He will also discuss the aim to turn ScienceAtHome into a large-scale social and cognitive science research platform offering insights into the individual minds and collective interactions.

Pinja Haikka, researcher at Aarhus University and head of outreach at ScienceAtHome, will talk about how to set up citizen science projects, with a special focus on QuantumMinds a citizen science project bridging quantum physics challenges and individual learning of volunteers participating in citizen science games.

## Where Citizen Science meets Cognitive Science

Ed Manley, who studies navigation skills with citizen science, discusses the Sea Hero Quest project, one of the most successful citizen science games in recent times. Originally developed to study navigation skills for dementia research, this project offers data on human navigation abilities from worldwide players from all age groups.

Carsten Bergenholtz from Aarhus University will introduce the Alice Challenge, a remote access experiment where volunteer players could remotely access and modify the settings of an actual instrument in the physics lab at Aarhus University. He discusses the challenges of remote experimental setups and discuss the advantages of running social science experiments on high-dimensional, real-life problems.

Oana Vuculescu, from the University of Aarhus, will introduce a game-based research project, the AlienGame, a sequential problem solving task to study the heuristics that individuals use in problem solving

## Designing and Building Games

Juho Hamari, is a Professor of Gamification (Associate \& tenure-track) and leads the Gamification Group spread across Tampere University of Technology, University of Turku and University of Tampere in Finland. He will give an overview about the academic literature on gamified crowdsourcing.

Nathaniel D. Phillips, cognitive scientist at the University Basel introduces the R Shiny platform as an easy yet powerful tool to build online experiments and games directly from R code. Through shiny, web application can directly interface with R which enables the researcher to conduct dynamic experiments in which the user interface is determined by cognitive modeling running behind the scene.

Julia A. Bopp, PhD candidate and player experience researcher at the Human-Computer Interaction Lab at the University of Basel, will introduce what game aspects may evoke
emotions and in turn how these emotions may influence good player experience.

Sharon T. Steinemann, PhD candidate at the Human-Computer Interaction Lab at the University of Basel, discusses meaningful game experiences. Her work investigates how ingame interactions shape experiences into being moving, thought-provoking, and personally meaningful. Findings and implications will be discussed using examples from current games with a focus on the relationship between game experiences and behavior change.

Julian Jarecki, virtual reality gaming developer at the University of Freiburg, introduces virtual reality with GraphVR and Ultimate Automizer. GraphVR is a virtual reality environment in which people can dynamically create and interact with near-real three-dimensional visualization of graph structures. The presentation will explore how VR creates an exciting opportunity to experience abstract concepts and structures.

Libby Heaney, virtual reality artist and research tutor at the Royal College of Art, will present a different way to use virtual reality, namely in the form of an exhibition that explains complex scientific matters to laypeople and makes these matters graspable.

## Hands-on Experience and Discussions

Experience Virtual Reality. We will additionally offer participants of the workshop the opportunity to directly experience and try out how virtual reality environments feel. We will have a live demo of Graph3D and other virtual reality environments.

## Invited Speakers

Jacob Sherson | Professor of Physics and Astronomy, Aarhus University | Founder of the citizen science platform ScienceAtHome
Juho Hamari | Professor of Gamification, Tampere University of Technology and University of Turku
Carsten Bergenholtz | Associate Professor of Management, Aarhus University
Ed Manley | PostDoc in spatial cognition, UCL London | Partner at the citizen science project SeaHeroQuest ( $t b c$ )
Julia A. Bopp $\mid \mathrm{PhD}$ candidate, gaming researcher, University of Basel
Julian Jarecki $\mid$ app and gaming developer with focus on development in VR, University of Freiburg
Libby Heaney | Digital and virtual reality artist and tutor at the Royal College of Art London
Nathaniel D. Phillips | PostDoc in cognitive science, University of Basel (tbc)
Pinja Haikka | PostDoc in Physics, Aarhus University | Head of Outreach at ScienceAtHome (tbc)
Sharon T. Steinemann | Phd Candidate | University of Basel

# Cooperative Social Intelligence: Understanding and Acting with Others 

Max Kleiman-Weiner (maxkw@mit.edu), Yibiao Zhao (ybz@mit.edu) \& Joshua Tenenbaum (jbt@mit.edu) Brain and Cognitive Science, Cambridge, MA 02139 USA

Keywords: multi-agent, cooperation, communication, coordination, theory-of-mind, social learning

## Theme

This workshop will focus on new developments and approaches to studying social intelligence with a specific focus on cooperation, theory-of-mind and social learning. With a diverse set of speakers and panelists, we anticipate these three themes will allow for connections to be made between developmental psychologists, cognitive scientists and artificial intelligence and robotics researchers.

1. Cooperation: How do we coordinate our limited individual capacities and perspectives to accomplish goals that noone could have completed on their own? How do we share the spoils of a cooperative activity fairly and equitably and how does this capacity develop in early childhood? How do we learn who is cooperative and who is not to be trusted? What special purpose representations have we evolved that make cooperation so robust?
2. Theory-of-Mind: How do we go from sparse, noisy, underdetermined observations of behavior to acquire abstract knowledge of latent mental states which generalize to novel situations and people? How does this capacity develop and what representations support these capacities in infancy and early childhood? How do we understand the actions and intentions of groups or even collectives of agents?
3. Social Learning: When and how do we realize that other intelligent agents are often the richest source of world knowledge in an environment? How do we actively learn from others? How do we efficiently share important cultural knowledge through teaching? How do norms and conventions originally form and how do we learn existing norms and conventions so quickly?

One motivation for understanding social intelligence is to re-engineer socially intelligent artificial agents that treat people like people and can be treated like people by people. While a world where artificial agents roam the sidewalks still feels far away, automated agents are already roaming the streets in self-driving cars. The above challenges to understanding our own social intelligence become challenges for engineering cooperative AI:

1. Cooperation: How can we build agents that can work with us on mixed teams? Will they need to be taught cooperative values or must these values be baked in from the start?
2. Theory-of-Mind: How can we build agents that can understand our intentions, how they unfold over time, and the ways in which they may change dynamically? Can agents without theory-of-mind robustly cooperate with humans? How can agents reveal their intentions in ways that are natural to us?
3. Social Learning: How can machine learning from teachers go beyond imitation and reinforcement? Can artificial agents take advantage of the human ability to teach and learn like children do?

## Speakers

We have already invited and received confirmations from six speakers (one tentative) that will form the core of our workshop. These speakers come from communities ranging from computer science and artificial intelligence to cognitive science and developmental psychology. Each speaker brings a unique perspective that will be of interest to the entire CogSci community. We believe these interdisciplinary interactions are a unique and positive element of the CogSci community and we hope to build on that foundation. We may invite 12 more speakers and are committed to organizing a gender balanced workshop.

Nick Chater (Professor of Behavioral Science, Warwick) Possible topics: Virtual bargaining, instantaneous conventions, one-shot communication, joint intentions

Joel Leibo and Thore Graepel (DeepMind) Possible topics: Deep learning for cooperation and competition

Stuart Russell (Professor of Computer Science, Berkeley) Possible topics: Cooperative inverse reinforcement learning, value alignment, cooperative robotics

Hyo Gweon (Professor of Psychology, Stanford) Possible topics: Social learning, social development, theory-of-mind, cooperation

Igor Mordatch (Professor of Robotics, CMU \& OpenAI) Possible topics: Emergent communication, deep reinforcement learning

Victoria Southgate (Professor of Psychology, Copenhagen) [tentative] Possible topics: Infant social cognition, action processing, Imitation and mimicry, Theory of mind, Motivation and learning

## Workshop Program

We plan to host a full day workshop consisting of talks given by $6-8$ invited speakers (depending on final commitment). Our intention is to explore having an extended lunch and allow (space permitted) for the presentations of posters in this topic area. This will allow for more informal interactions between interesting parties and will allow for researchers to have an additional opportunity to present their work. At the end of the workshop we will have a discussion panel with all of the speakers to synthesize the topics discussed throughout the workshop.

## Potential Financial Support

As this topic will be of interest to some of the industrial research labs we will also see if they are interested in providing some support to defer some of the registration costs. However as we already have confirmations from most of our speakers this funding will not be necessary and we believe we will have a very successful within the constraints of the funding provided by the conference.

# Building Bridges from (Ivory) Towers: Combining Academia and Industry for Cognitive Research 

Katherine A. Livins (klivins@netflix.com) \& Jay B. Martin (jaym@netflix.com)<br>Originals Product Innovation, Netflix<br>121 Albright Way, Los Gatos, CA 95032, US

Keywords: methods; basic and applied research; industry

## Objectives

This half-day workshop will discuss how to enrich research by marrying academic and industry-based work. Attendees will learn the theoretical, practical, and logistical complexities involved in advancing cognitive science across these distinct research sites.

This topic is relevant to the cognitive community because academia and industry often have common goals but distinct capabilities. For example, academics have the freedom to study almost any quantifiable question, and typically run small-scale studies performed in highly controlled settings with limited sets of participants. This approach results in high internal validity, but low statistical confidence, external validity, and limited replicability. Industry researchers are also often interested in human behavior (that of their users or clients) but typically need to further a company's business objectives with their work. However, these researchers have access to large-scale data sets and resources unmatched by the academic sector (Griffiths, 2014). This workshop will help attendees identify cases where cross-site collaborations might be useful, along with the methods necessary for carrying out such research.

## Workshop Outline

The workshop will begin with a series of talks highlighting a) the use of cognitive theory in industry contexts, b) the methodological considerations necessary for undertaking industry-centered research, and c) how to build holistic networks that pass information gleaned from industry activity back into academia-centered research. It will provide examples from the following industries in order to deliver a broad perspective.

Education Thinking, reasoning, and learning are central to cognitive science, however little attention is given to applied education. That said, a number of researchers have realized that watching how people process information in real learning contexts can not only help to shape better educational programs (e.g., Weitnauer et al., 2016), but also provide insights into cognition (e.g., Goldstone et al., 2008). This workshop will provide examples of how research in real learning settings can provide access to diverse research participants, and an ability to 'cache-out' learning theories.

Data Science and Machine Learning Cognitive science is broadly interested in observing human behavior to make claims about mental mechanisms. While data science and
machine learning can leverage any data, many companies are specifically interested in describing, understanding, and predicting their human users. This data can be important to basic research since, as Jones (2015) points out, social media, web-tracking, search-logs, and consumer reviews are huge sources of behavioral data that are typically inaccessible to academic researchers. This workshop will use data science and ML as examples of how industry is using cognitive science to model user behavior, as well as how those models can provide insights into cognition.

User Experience (UX) Cognitive science often specifies how cognitive mechanisms interact with external stimuli. While these interactions are typically described in terms of high-level principles, they govern real-world interactions such as those between a user and a product. UX focuses on applying these principles to optimize usability (Kujala et al., 2011). As a result, it relies on both cognitive research and methods. For example, it can consider work about perception and attention in order to lessen cognitive load on a human worker (e.g., Zheng et al., 2011), and it can be evaluated using measures such as eye-tracking, EEG, and reaction times (Oviatt, S.L., 2006). This workshop will use UX as an example of how academic learnings can be directly translated into concrete products, and how the development of those products can provide cognitive data.

The workshop will then provide participants with an opportunity to develop their own cross-site methodologies, with guidance from the facilitators. Finally, there will be a round-table discussion with both the facilitators and speakers where participants can ask further questions, get group feedback on their ideas, and further develop their understanding of cross-site research.

## Audience

This workshop will appeal to the broad cognitive community. As Jones (2015) states, 'Cognitive research is increasingly coming out of the laboratory', and increased interest has been demonstrated in past CogSci workshops, such as Mason \& Suri (2011). Thus, this workshop will be useful for experimentalists, who will learn about the methodological and operational considerations necessary for completing studies in industry contexts, as well as computational modelers who will learn how their models may be applied or developed through industry applications.

It will also benefit researchers at various stages in their careers. Graduate students may be interested in a window into research-oriented job trajectories outside of academia.

Likewise, younger professors interested in building academic/professional research networks, and professors attending to the flow of students into industry will value a venue for considering the positive benefits of increased connections across research sites.

## Presenters

This workshop will include speakers who have experience in both academic and industry research.

## Facilitators

This workshop will be run by Katherine Livins and Jay Martin. Livins's academic research focused on reasoning and perception, while Martin's focused on models of causal reasoning and categorization. They both now work as Data Scientists, leveraging cognitive science for the purpose of optimizing human decision-making in online working environments.

## Additional Speakers

David Landy - Indiana University, Department of Psychological and Brain Sciences. Landy's research focuses on computational and theoretical approaches to formal reasoning, mathematical cognition, and perception. He used this research to create an application called "Graspable Math", which teaches mathematical concepts by exploiting perceptual and gestural processes. He will speak on how real-world situations challenge academic research by providing richer contexts to develop and test the robustness of theories, with examples grounded in Graspable Math.

Noah Goodman - Stanford University, Advisor to Geometric Intelligence (now Uber AI Labs), Co-founder and advisor to Gamalon Labs and Ought Inc. Goodman's academic work spans cognitive modeling, probabilistic programming languages, categorization, and social reasoning. He also recently began working with Uber on applied AI and machine learning projects. He will speak about how these projects are interacting with his preexisting academic research.

Robert Rauschenberger - Exponent. Rauscehnberger's academic work focused visual attention in displays. He now works as a Managing Scientist on human factors and industrial engineering, applying his knowledge about the visual system to product design and human-product interface interaction. He will speak on the role of cognitive psychology in user research and experience design.

Nick Gaylord - CrowdFlower. Gaylord's academic research focused on the application of experimental design principles to the collection of training data for NLP models, and the role of domain-general decision making processes in human language comprehension. He now works as a Senior Data Scientist on how to curate human-generated data sets for algorithm consumption. He will speak on how skill sets and methods can translate between academia and industry.

## Workshop Structure

Livins will open by introducing and framing the workshop. Each speaker will then present a 25 -minute talk (including questions), before Martin closes by identifying consistent themes. Livins and Martin will then engage participants to identify cross-site opportunities in their own work, and help them develop a list of necessary methods and resources. The session will end with a 20-minute round-table discussion.

Table 1: Schedule.

| Event | Time |
| :--- | :--- |
| Opening Remarks: Livins | 20 minutes |
| Speaker 1: Landy | 25 minutes |
| Speaker 2: Goodman | 25 minutes |
| Speaker 3: Rauschenberger | 25 minutes |
| Speaker 4: Gaylord | 25 minutes |
| Closing Remarks: Martin | 20 minutes |
| Guided research development | 20 minutes |
| Round-table discussion | 20 minutes |
| Total time | 180 minutes |

## References

Jones, M.N. (2015). Big Data in Cognitive Science (Frontiers of Cognitive Psychology). NY: Routledge.
Goldstone, R.L., Landy, D.H., Son, J.Y. (2008). A well grounded education: The role of perception in science and mathematics. In A. Glenberg, M. DeVega, \& A. Grasser (Eds.), Proceedings of the Garachico Workshop on Symbols, Embodiment and Meaning. Universidad de La Laguna, Tenerife. (pp. 327-355).
Griffiths, T. (2014) Manifesto for a new (computational) cognitive revolution. Cognition, 135, 21-23.
Kujala, S., Vaananen-Vainio-Mattila, K., Krapanos, E., \& Sinnela, A. (2011). UX curve: A method for evaluating long-term user experience. Interacting with Computers, 23(5), 473-483.
Mason, W.A., Suri. (2011). How to use mechanical turk for cognitive science research. In L. Carlson, C. Hlscher, \& T. Shipley (Eds.), Proceedings of the 33rd annual conference of the cognitive science society (pp. 66-67). Austin, TX: Cognitive Science Society.
Oviatt, S.L., (2006). Human-centered design meetings cognitive load theory: Designing interfaces that help people think. In Proceedings of the ACM Conference on Multimedia, Special Session on "Human-Centered Multimedia System", 871-880. New York: ACM
Weitnauer, E., Landy, D., Ottmar, E. O. (2016). Graspable Math: Towards Dynamic Algebra Notations that Support Learners Better than Paper. Future Technologies Conference, San Francisco, California.
Zheng XS, Kiekebosch J, Rauschenberger R. Attentionaware human-machine interface to support video surveillance task. Proceedings, Human Factors and Ergonomics Society 55th Annual Meeting, NV, 2011.

# The Computational Foundations of Religious Cognition: A Workshop Hosted by the International Association for the Cognitive Science of Religion (IACSR) 

Ann Taves (taves@religion.ucsb.edu)<br>Department of Religious Studies, Mailcode 3130, UCSB, Santa Barbara, CA 93106, USA<br>University of California, Santa Barbara<br>Dimitris Xygalatas (xygalatas@uconn.edu)<br>Department of Anthropology, 354 Mansfield Road, Storrs CT 06269, USA<br>University of Connecticut

John Shaver (john.shaver@otago.ac.nz)
Religion Programme, P.O. Box 56, Dunedin 9012, New Zealand
University of Otago

Keywords: religious cognition; fMRI; placebo studies; natural language processing; text data mining; belief reversal; dual processing; IACSR

## Workshop Description

Religion is of global significance, and its study requires explanations from cognitive science. Currently, the cognitive science of religion consists of researchers working in an array of disciplines, employing diverse methods, including, among others: experimental research and modelling in psychology and neuroscience, and historical, archaeological, and comparative studies of religious cognition in anthropology and religious studies. The International Association for the Cognitive Science of Religion (IACSR) seeks to advance the naturalistic and cognitive study of religion by providing settings for productive dialogue across disciplinary boundaries and methodological approaches. This half-day workshop, organized by the IACSR, has three complementary goals: 1) to expose attendees to diverse methodologies for studying the computational foundations of religious cognition, 2) to provide a forum for researchers to present recent empirical findings that bear on our understanding of religious cognition, and 3) to foster new research collaborations.

To achieve these goals, the IACSR has invited three speakers whose recent work represents cutting-edge, yet diverse, methodological approaches to the study of the foundations of religious cognition. The IACSR executive board also solicited poster submissions from members and interested researchers. The first half of this workshop will consist of the three invited lectures, and the second half will involve the presentation and critical discussion of eight posters that were the highest rated by blind peer review.

## Invited talks

## The Power of Suggestibility: Using Placebo Brain Stimulation Devices to Manipulate Subjective Experience.

Michiel Van Elk, Department of Psychology, University of Amsterdam

Effects of expectancy have been studied widely in both clinical as well as experimental settings and show the powerful effects of expectations on treatment outcomes. However, most studies on expectancy effects have focused selectively on the use of inert treatments for alleviating pain or illness. Less is known about the potential of enhancing cognitive performance or human experience through expectancy manipulations. In this talk I will present a series of studies aimed at investigating the psychological and neurocognitive basis of expectancy effects on human performance and experience. First, in three studies we used a placebo 'God-helmet' to manipulate mystical experiences and to investigate the effects of self-transcendence on selfperception. In a second research line we used a placebo cognitive enhancement device and we found that belief in cognitive enhancement was associated with a stronger externalization of agency and a change in neural responses to errors. Third, in an fMRI study we found that the tendency to get absorbed in external stimuli was associated with a decreased activity of the default mode network (DMN) and the anterior cingulate cortex (ACC), thereby indicating that a process of de-selfing and reduced cognitive control could underlie suggestibility effects. These findings are integrated in the computational framework of predictive processing, which provides a unifying theory to account for religious and spiritual beliefs and experiences.
"I once was blind..." Choice Blindness and Religious Attitude Reversals

Ryan McKay, Department of Psychology, Royal Holloway London
"Choice blindness" refers to the fact that research participants often fail to notice mismatches between an outcome they choose and an outcome they receive, while nevertheless being prepared to offer justifications for choosing the outcome they did not in fact choose. Recently this phenomenon has been demonstrated in the domain of peoples' political and moral attitudes. For example, one study 'magically' exposed participants to a reversal of their previously stated attitudes, and found that many participants not only failed to detect these reversals, but constructed coherent and unequivocal arguments supporting the opposite of their original position. In this talk I will describe some recent research adapting this paradigm to the domain of religion. Our findings reveal a dramatic potential for flexibility in our religious attitudes and beliefs. I will situate this research in the broader context of attempts to manipulate religiosity.

## Mind the Text - Retracing Mental States and Cognitive Trajectories in Historical and Text-Heavy Data

Kristoffer Nielbo, Digital Text Laboratory \& Interacting Minds Centre, Department of Culture \& Society, Aarhus University

Humans exhibit a species-unique capacity for long-term planning and future-oriented cognition. This 'deep temporality' is so fundamental to human behavior, that it can be considered the hallmark of our symbolically mediated environmental interactions. Systems of cultural norms and behavior (e.g., religious groups and traditions) have a long history and develop at a time scale, which can present a challenge to the canonical methods in cognitive and experimental anthropology. How, for instance, can we approach the historical and cognitive trajectories of adherents to a religion codified several thousand years ago? The proliferation of digitized historical and text-heavy data we are currently witnessing holds part of the solution. To illustrate this within the domain of cultural cognition, we present three studies that combine techniques from Natural Language Processing (NLP) and text data mining in order to study cognitive and affective trends at multiple time scales. Study 1 compares cognitive trajectories for historically significant religious experts; study 2 explores semantic change and concepts of mind in classical Chinese literature; and study 3 presents evidence for an evolutionary-based motivational model of religious fundamentalism. We argue that when combined with domain knowledge in language and culture, NLP and text data mining are promising approaches to cultural cognition at long time scales.

## Peer-reviewed Posters

Supernatural Agents in Predictive Minds. Marc Anderson, Aarhus University. A description of what is currently regarded as one of the most promising models of perception in cognitive neuroscience, predictive coding, and the results of three experimental studies in which this framework is operationalized.

AVM: Data Structures for the Cognitive Science of Religion. Tamás Biró, Eötvös Loránd University. Describes a method for understanding religion as a complex network of attribute-value matrices (AVMs).

Flag Identity Theory (FIT): A Cognitive Explanation for Large Scale Group Cooperation and Conflict. Michael Gantley, University of Oxford, Justin Lane, Boston University. Describes Flag Identity Theory, which relies upon cognitive mechanisms for social identification and biases of loss avoidance to explain patterns of large scale cooperation.

Culturally and Developmentally Robust Intuitions about Purpose and Intentional Design in Nature: Dual Processing Evidence from China. Deb Kelemen, Boston University, Elisa Järnefelt, University of Helsinki, and Liqi Zhu, Chinese Academy of Sciences. Describes the results of two studies conducted in China: the first replicated earlier Western-based findings of a default teleological bias. In the second, participants revealed marked tendencies to view natural phenomena as created, particularly under speeded conditions and in a non-human-made rather than humanmade condition.

Computer Simulation of Large-scale Religious Systems. Justin Lane, Boston University. Describes the use of multiagent artificial intelligence to simulate individual and group level dynamics of religious cognition.

Do forgiving God Primes Strengthen Support for State Sanctioned Punishment? Katherine O'Lone, Royal Holloway London. Describes the results of a study that investigated whether the manner in which God is believed to intervene affects people's endorsement of state-sanctioned punishment.

Cognitive Foundations of Theodicy. Karolina Prochownik, Jagiellonian University. Describes a dual-process cognitive approach to the study of theodicy.
"I can't believe she's dead": The Effects of Corpse Viewing and Corpse Condition on Vigilance for Deceased Loved Ones. Claire White, UC Northridge, Daniel MT Fessler, UCLA, Pablo Gomez-Forero, UC Northidge. Describes the results of a study that examines the effect of exposure to cues of death on vigilance for agents.

# Workshop proposal: Deep Learning in Computational Cognitive Science 

Ilker Yildirim (ilkery@mit.edu) and Joshua B. Tenenbaum (jbt@mit.edu)<br>Department of Brain \& Cognitive Sciences, MIT<br>Cambridge, MA 02130 USA

Keywords: Computational models of cognition; deep learning; Bayesian models; cognitive neuroscience; computational neuroscience; computational psycholinguistics.

## Overview and significance

A new generation of deep neural network architectures has driven rapid advances in AI over the last ten years. These architectures include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and many variants and extensions. Computational cognitive scientists and neuroscientists have now begun to explore these techniques, and how they might combine with other computational tools such as Bayesian models, symbolic grammars and rule-systems, probabilistic programs, and reinforcement learning. The goal of this workshop is to bring together some of the leading researchers working at this interface, for short talks and an integrative discussion of open questions and promising directions.

Talks will cover many areas of cognition including perception, problem-solving and planning, decision-making, language and social cognition. The focus will be on models of human behavior, but the potential bridge to neural studies in humans (via fMRI) and animals (via physiology) will also be explored. Most talks will assume only a basic familiarity with neural networks, and so should be accessible to all CogSci attendees. We hope to be scheduled for an afternoon slot, and have coordinated our plans with the DeepMind's Deep Learning tutorial proposed for the morning which could serve as an introduction to more advanced methods that several talks will build on.

Workshop structure. We plan a half-day workshop comprising seven talks, each 20-25 minutes, followed by a 30 -minute panel discussion with all speakers on open questions. We will also encourage student participants to present posters on relevant work during the coffee break.

## Organizers and Presenters

Ilker Yildirim (Organizer) is a research scientist at MIT. His research spans visual and multisensory perception, computational neuroscience, and artificial intelligence
Joshua Tenenbaum (Organizer) is Professor of Computational Cognitive Science at MIT. He studies learning, perception, common-sense reasoning, and has been active in both cognitive science and artificial intelligence.
Matt Botvinick is DeepMind's Director of Neuroscience Research, and was formerly Professor of Psychology and Neuroscience at Princeton. He is a leader in computational cognitive neuroscience and reinforcement learning.

Noah Goodman is a Associate Professor of Psychology, Linguistics and Computer Science at Stanford University. His research centers on computational modeling of higherlevel cognition and probabilistic programming languages.

Thomas Griffiths is a Professor of Psychology and Cognitive Science at UC Berkeley. His group develops computational models of higher-level cognition, drawing on probabilistic, neural network, and evolutionary paradigms.

Jessica Hamrick is a PhD candidate at UC Berkeley and former intern at DeepMind. Her research focuses on mental simulation, planning and metacognition.
Tal Linzen is a postdoctoral researcher at ENS and will be Assistant Professor in the Department of Cognitive Science at Johns Hopkins starting in the fall. His research interests involve computational modeling and psycholinguistics.

Daniel Yamins is an Assistant Professor of Computational Neuroscience in the Department of Psychology at Stanford. His research lies at the intersection of neuroscience, artificial intelligence, and psychology.

## Presentations

Prefrontal cortex as a meta-reinforcement learning system (Botvinick). Two decades of neuroscience research on reward-based learning has converged on a canonical model, under which the neurotransmitter dopamine 'stamps in' associations between situations, actions and rewards by modulating the strength of synaptic connections between neurons. However, a growing number of recent findings have placed this standard model under strain. This talk draws on recent advances in AI to introduce a new theory of reward-based learning. Here, dopamine trains another part of the brain, the prefrontal cortex, to operate as its own freestanding learning system. This perspective accommodates the findings that motivated the standard model, but also deals gracefully with a wider range of observations, laying a fresh foundation for next-generation research.

Less-supervised loss functions for training models of the visual system (Yamins). Recent advances in computer vision and AI have made it possible to build deep neural networks that mimic aspects of computation in the primate and human visual system. The core idea behind these results is task-driven modeling, e.g. optimize a neural network for a complex ecologically relevant behavior, and then compare the learned model to neural responses in the brain areas thought to underlie that behavior. While this approach has produced powerful predictive models of neural representations in adult animals, its main successes so far have unfortunately relied on using heavy semantic
supervision, using large labeled datasets - data streams to which real animals do not have access. I will discuss recent work my lab has been doing to move beyond this limitation, developing less heavily supervised approaches to train deep neural network models of vision that may be more plausible models of real neural learning. These ideas rely on loss functions defined in interactive worlds that attempt to better capture the true complexities of the environment in which early juvenile development takes place.

Modeling "analysis by synthesis" in perception by combining generative models and deep inverse networks (Yildirim and Tenenbaum). In its most general form, perception can be defined as the solution to an inverse problem: identifying the world scene that gave rise to the observed retinal (or auditory, or haptic) data. It remains a mystery how the brain constructs such rich representations of the geometry of objects in a scene and their physical properties extremely quickly, in at most a few hundred milliseconds. Traditional models of how the brain solves such inference problems use iterative methods that are hard to map onto neural circuits and much too slow to explain online perception. Here we present a new approach that combines a deep neural network with a probabilistic generative model. This approach is as fast as pure feedforward models, but generates a rich description of 3D object shapes and physical properties. Applied to faces, our model explains both the tuning properties of cells in the macaque face patch system and human behavior in recognizing familiar and unfamiliar individuals.

Deep networks for amortized inference in structured probabilistic models (Goodman). I will discuss deep amortized inference for probabilistic programming languages (PPLs). This is an approach that uses a PPL to describe complex probabilistic conceptual knowledge, and captures knowledge about how to use these concepts for inference via a deep 'inference network'. I will discuss several different objectives and training methods, including variational inference and 'dream learning' (a modern variant of wake/sleep). After describing the technical setup and showing the results for a few model-learning tasks, I will speculate about the relation of deep inference networks to human procedural knowledge.

Leveraging deep learning to study representations underlying human cognition (Griffiths). Recent neural network models have resulted in significant progress in computer vision, speech recognition, and natural language processing, by learning representations of the statistical structure of complex visual, auditory, and linguistic stimuli. Understanding how the resulting models work - and how well they correspond with human perception - is an interesting scientific challenge. However, the representations that these models discover, when treated just as representations of complex stimuli, also offer the opportunity to extend the scope of psychological research.

Psychologists studying problems such as categorization or memory have tended to focus on very simple stimuli that can be carefully controlled and parameterized. This makes it possible to formulate precise theories, but at the cost of potentially losing sight of the original phenomena: Do the same models that predict how people categorize sinusoidal gratings explain how they differentiate cats and dogs? I will talk about recent work that tries to leverage representations produced by neural networks - for both images and language - to study human cognition, highlighting the promise of this approach as well as some of the challenges.

Metacontrol for Adaptive Imagination Based Optimization (Hamrick). Many machine learning systems are built to solve the hardest examples of a particular task, which often makes them large and expensive to run--especially with respect to the easier examples, which might require much less computation. For an agent with a limited computational budget, this "one-size-fits-all" approach may result in the agent wasting valuable computation on easy examples, while not spending enough on hard examples. Rather than learning a single, fixed policy for solving all instances of a task, we introduce a "metacontroller" inspired by human cognition which learns to optimize a sequence of imagined internal simulations over predictive models of the world (called "experts") in order to construct a more informed, and more economical, solution. Our approach learns to adapt the amount of computation it performs to the difficulty of the task, as well as which experts to consult by factoring in both their reliability and individual computational resource costs. The metacontroller achieves a lower overall cost (task loss plus computational cost) than more traditional fixed policy approaches, demonstrating that our approach is a powerful framework for using rich forward models for efficient model-based reinforcement learning.

Understanding neural models of human language (Linzen). Large-scale artificial neural networks have shown great promise in natural language processing, reportedly reaching human-level performance in some tasks. Yet our understanding of the capabilities of these methods is typically limited to general statistics averaged across a random sample of texts. Such coarse-grained evaluation metrics stand in marked contrast to the rich array of highly specific patterns identified by linguists and cognitive scientists. I will argue that this detailed characterization of human-level knowledge of language provides a yardstick for the desired behavior of an artificial intelligence system. Applying it to neural networks can help us understand the strengths and weaknesses of existing architectures and analyze models that are otherwise difficult to interpret. Finally, neural networks that combine powerful statistical learning with different degrees of representational assumptions can serve as useful baselines for psychological modeling. I'll discuss case studies illustrating this approach in both syntax and semantics.

# Anthropological Contributions to Cognitive Science 

Organizers<br>Andrea Bender (Andrea.Bender@uib.no) \& Sieghard Beller (Sieghard.Beller@uib.no)<br>Department of Psychosocial Science, University of Bergen, Norway<br>Presenters<br>Rita Astuti (R.Astuti@lse.ac.uk)<br>Department of Anthropology, London School of Economics (LSE), London, UK<br>Olivier Le Guen (ompleguen @ gmail.com)<br>Centro de Investigaciones y Estudios Superiores en Antropología Social, CIESAS, Mexico Ciudad, Mexico<br>\title{ Karenleigh A. Overmann (koverman@uccs.edu)<br><br>Center for Cognitive Archaeology, University of Colorado, Colorado Springs, US }<br>Peter Rácz (peter.racz@bristol.ac.uk) \& Fiona Jordan (Fiona.Jordan@bristol.ac.uk)<br>Department of Archaeology and Anthropology, University of Bristol, UK<br>Mads Solberg (Mads.Solberg@uib.no)<br>Department of Social Anthropology, University of Bergen, Norway

Keywords: anthropology; cognitive science; philosophy of science; cognition; culture; language; materiality.

Anthropology was a founding member of cognitive science (Bender et al., 2010; Gardner, 1985), sharing with other cognitive disciplines a deep interest in thinking and behavior. With its unique expertise in the cultural content, context, and constitution of cognition, it would still be essential to any comprehensive endeavor to explore the human mind (Bloch, 2012), but rather has turned into cognitive science's "missing discipline" (Boden, 2006), thus leaving important questions unanswered or even unasked. Given that substantial shares of knowledge are implicit and that cognition is situated, distributed, embodied, and grounded in various other ways, anthropological approaches provide privileged access to investigation: for arriving at reasonable hypotheses, ensuring ecological validity, and even for coming up with new research questions and paradigms (Astuti \& Bloch, 2012; Hutchins, 2010; Nersessian, 2006).
In line with recent calls for rapprochement in Topics in Cognitive Science (Bender et al., 2012; Beller \& Bender, 2015), our symposium brings together scholars that represent different branches of contemporary anthropology with distinct perspectives-including 'traditional' social anthropology, cognitive anthropology and ethno-linguistics, cognitive ecology, evolutionary anthropology, and archaeologyto present what they consider to be indispensable contributions to cognitive science.

With our selection of authors, we hope to demonstrate the value of anthropological approaches for cognitive science as well as the potential benefits of cross-disciplinary collaboration. Cognitive archaeologist Overmann discusses a theoretical perspective on how mind, behavior, and material artifacts interact to shape human cognition. Combining their expertise in linguistics and evolutionary anthropology, Rácz and Jordan investigate the design principles of kinship sys-
tems as near-universal conceptual tools. With his background in (ethno-)linguistics and cognitive anthropology, Le Guen uses Yucatec Maya sign languages to illustrate the importance of cultural practices for shaping cognitive behavior. Based on Hutchins' cognitive ecology approach, Solberg speaks to questions at the intersection of anthropology and philosophy of science by illuminating the cultural framework of science production in a biology lab. And social anthropologist Astuti concludes by taking a bird's eye view on how efforts to understand the human mind crucially benefit from acknowledging its historical origins and from taking the specific sociocultural contexts into consideration.

Based on work some of which is published in high-quality journals (such as Science, Nature, PNAS, BBS, TiCS, Current Anthropology, or Cognition), these participants will offer invaluable contributions to a more diverse, more inclusive, and hence more comprehensive cognitive science.

## Archaeology and Cognitive Science

## Karenleigh A. Overmann

Archaeology contributes to cognitive science in two key areas. First, in understanding human cognitive evolution, archaeology furnishes critical data on the timing and context of developments (Wynn, 2002). This approach assumes minds make tools: increasing complexity in material forms is an effect of, and thus signals, cognitive change related to neurological developments like encephalization. Second, archaeology provides unique insight into the ways materiality functions within the extended, enacted mind. This inverted approach-tools make minds (Malafouris, 2013)examines how material forms interact with body and brain to create meaning and experience and potentialize behavioral and psychological change. In both contributions, archaeology negotiates temporalities, centuries to millennia and longer, that can be challenging for psychological theories and methods to assimilate (e.g., Overmann, 2016).

# Cognitive and Adaptive-Historical Explanations for Kinship Diversity 

Peter Rácz \& Fiona Jordan

Kinship systems are semantic systems whose forms can be explained in terms of domain-general cognitive principles; kinship categories are optimised to be maximally distinct and as simple as possible. Kinship, then, is similar to other universal semantic categories such as colour terms. However, whereas colour terms broadly fit into one typological hierarchy, kinship systems comprise a diverse typology. Alternatively, adaptive-historical explanations emphasise how cultural traditions and social practices (particularly marriage and transfer of resources) place functional pressures on the shape of kinship systems (Jordan \& Dunn, 2010). Using a global ethnographic database of over a thousand societies we show that marriage rules and ancestry have a significant influence on the type of kinship system found in a society. This remains true if we control for the effect of lateral transmission and phylogeny. This, in turn, means that kinship is best approached by combining cognitive and historic-anthropological explanations. These results have broader implications for the understanding of lexical systems in particular and the mechanisms of human cognition in general.

## How Cultural Settings Frame Spatial Cognition: The Example of Yucatec Maya and Yucatec Maya Sign Language

On the Yucatec peninsula, the main native language is spoken Yucatec Maya (YM). However, in villages where deaf people are born, a local sign language (YMSL) was created both by deaf and their hearing kin. Although both languages are in intense contact, they are genetically different, and YMSL is not a signed version of YM. In Le Guen (2011), I showed how gestures-in addition to linguistic structures (Levinson, 2003)-can support a geocentric frame of reference. In this paper, I want to elaborate on how deaf signers using YMSL still 'inherit' the same conception of space through cultural practices.

## Exploratory Experimentation in Experimental Systems: Novel Directions for the Cognitive Anthropology of Science

Mads Solberg

It is now widely recognised that progress in many scientific disciplines, like molecular biology, are not adequately described by the hypothetic-deductive model of epistemic action through experimental falsification. Instead, cumulative progress is achieved through description and modelling of mechanisms (interacting parts that produce regularities). One view claims that mechanistic discovery proceeds through exploratory experimentation; a practice where experimentation takes on many other cognitive functions than just hypothesis-testing. Experimental systems (material,
conceptual, social, and cultural infrastructures of laboratories) set up divisions of cognitive labour and distribute cognition through time and space in ways that are critical to this process. This talk looks at how the alliance between anthropology, cognitive science, and adjacent fields like philosophy and history of science, can contribute to further developing this research area. Such collaborations are necessary for adequately explaining cultural transmission and cultural evolution in scientific knowledge, and for describing interactions between mental representation, epistemic action, and material culture in scientific experimentation. I draw on examples from a long-term cognitive-ethnographic casestudy in a community of molecular life-scientists.

## Anthropology as a Critical Friend

Rita Astuti
Anthropology is commonly listed as one of the disciplines that make up cognitive science. But what exactly is the contribution that anthropology can make to the interdisciplinary study of human cognition? The paper will argue that anthropology must take on the role of critical friend, constantly reminding other disciplines of the historical origins of all human phenomena and of the theoretical and methodological challenges that come from recognising that all aspects of human cognition develop in specific social and cultural contexts.

## References

Astuti, R., \& Bloch, M. (2012). Anthropologists as cognitive scientists. Topics in Cognitive Science, 4, 453-461.
Beller, S., \& Bender, A. (Eds.) (2015). Exploring cognitive diversity: Anthropological perspectives on cognition. Topics in Cognitive Science, 7(4).
Bender, A., Beller, S., \& Medin, D. (Eds.) (2012). Does cognitive science need anthropology? Topics in Cognitive Science, 4(3).
Bender, A., Hutchins, E., \& Medin, D. L. (2010). Anthropology in cognitive science. Topics in Cognitive Science, 2, 374-385.
Bloch, M. (2012). Anthropology and the cognitive challenge. Cambridge: Cambridge University Press.
Boden, M. A. (2006). Mind as Machine. Oxford: Clarendon Press. Gardner, H. (1985). The mind's new science. NY: Basic Books.
Hutchins, E. (2010). Cognitive ecology. Topics in Cognitive Science, 2, 705-715.
Jordan, F. M., \& Dunn, M. (2010). Kin term diversity is the result of multilevel, historical processes. Behavioral and Brain Sciences, 33, 388.
Le Guen, O. (2011). Speech and gesture in spatial language and cognition among the Yucatec Mayas. Cognitive Science, 35, 905-938.
Levinson, S. C. (2003). Space in language and cognition. Cambridge: Cambridge University Press.
Malafouris, L. (2013). How things shape the mind. Cambridge: MIT Press.
Nersessian, N. J. (2006). The cognitive-cultural systems of the research laboratory. Organization Studies, 27, 125-145.
Overmann, K. A. (2016). Beyond writing: The development of literacy in the Ancient Near East. Cambridge Archaeological Journal, 26, 285-303.
Wynn, T. (2002). Archaeology and cognitive evolution. Behavioral and Brain Sciences, 25, 389-402.

# Static and dynamic visual narratives, by brain and by eye 

Neil Cohn (neilcohn@visuallanguagelab.com)<br>Tilburg center for Cognition and Communication (TiCC)<br>P.O. Box 90153, 5000 LE Tilburg, The Netherlands<br>Tom Foulsham (foulsham@essex.ac.uk)<br>Department of Psychology, University of Essex<br>Tim J. Smith (tj.smith@bbk.ac.uk)<br>Department of Psychological Sciences, Birkbeck, University of London.

Jeffrey M. Zacks (jzacks@wustl.edu)
Department of Psychology, Washington University in St. Louis

Keywords: visual narratives; film; comics; event segmentation; discourse; eye-tracking

## Introduction

Narrative has been studied for millennia, though recent attention in the cognitive sciences has turned towards visual narratives like those found in comics (Cohn, 2013a) and films (Zacks, 2014). Most agree that the basic principles guiding comprehension involve principles that extend across the verbal and visual domains (Cohn, 2013b; Gernsbacher, 1990; Magliano \& Zacks, 2011). However, visual units of narrative-both drawn and moving-demand different affordances to retrieve and integrate information.

Unlike verbal information, the sequential units of visual narratives use an analog spatial representation, from which a comprehender must extract the relevant information, ignore or suppress the irrelevant information, and work to connect such information across a sequence of units. This involves the integration of complex event information and its interaction with narrative structures.

Such a process is further varied in the difference between static, drawn visual narratives (as in comics) and dynamic, moving ones (as in films). The introduction of movement to a sequence provides important cues and an additional layer of constraints on the effective communication of visual sequential information.

This symposium highlights this growing field within the cognitive sciences. First, the presentations focus on visual narratives of both types: static, drawn narratives, and dynamic, moving ones. Second, they split their focus between eye-tracking and cognitive neuroscience. Together, these presentations will highlight the relevance of visual narratives for studying many facets of cognition, including attention, events, narrative, and discourse.

## Do you see what I see? The curious absence of endogenous effects on gaze during cinematic narratives

Our first talk by Tim J. Smith, along with John P. Hutson (Kansas State University), Joseph P. Magliano (Northern Illinois University), and Lester C. Loschky (Kansas State University), explores the dynamic nature of film narratives. Cinematic narratives are ubiquitous but unlike textual narratives or static images, how we process edited audiovisual sequences is barely understood. From reading and scene processing we know that exogenous (i.e. stimulus demands) and endogenous factors (i.e. higher-cognitive factors such as individual differences and comprehension) compete over our overt attention, biasing where we fixate and how we process the information. However, eye-tracking studies of film viewing have demonstrated a surprising similarity in where multiple viewers direct their gaze; a phenomenon we call attentional synchrony (Smith \& Mital, 2013). Task instruction, individual differences such as expertise and age, and even differences in how the edited scenes are comprehended often fail to show gaze differences. This fragility of endogenous influence is at odds with emerging theories of active vision (Henderson, 2017). In this talk we will review several studies from our labs investigating the causes of attentional synchrony and show how filmmakers have intuited techniques to guide viewer attention in complex dynamic scenes. These findings will be used to extend the Attentional Theory of Cinematic Continuity (Smith, 2012) to include an appreciation of the dynamic interplay between exogenous and endogenous factors during cinematic narratives and the apparent dissociation between gaze and comprehension.

## Eye-tracking sequential context in scenes, comics and movies

The scenes that confront us in our everyday lives are highly structured in time and space. However, most of what we know about how people look at such scenes is based on experiments with isolated images presented in a random
order. This talk by Tom Foulsham will describe results from scenes, comics and movies which show how even minimal sequential context changes the way that visual attention is deployed. In natural scene viewing, the way that people look at isolated photographs can be compared to how they view dynamic video or to the gaze behaviour shown when people walk through a real environment. In these cases the differences observed reveal how expectations govern our attention. Building up expectations is also a key part of how visual narratives function in comics and movies. We have begun to examine how eye movement patterns reflect information processing of comic strips (Foulsham, Wybrow, \& Cohn, 2016). As expected, participants' viewing patterns change when a coherent narrative is available. The eye-tracking data can also be used to generate new experimental manipulations (e.g., mimicking fixations by zooming into particular content). These manipulations reveal how attention to particular features or moments can affect comprehension of the narrative. This technique is being pursued in both comics and video sequences, providing new insights into top-down control of attention and the exploitation of this in visual media.

## Event Comprehension and Memory in Healthy Aging and Early Alzheimer's Disease

Research on film has also shared methods with the study of visual events. In this presentation, Jeffrey M. Zacks explores these relations along with Heather R. Bailey (Kansas State University, and Christopher A. Kurby (Grand Valley State University). Events unfold in time, and viewers track the temporal dynamics of activity as part of event understanding. Adaptively tracking event dynamics is important for guiding action online and for forming durable episodic memories. Event perception and event memory both can be affected by healthy aging and by neurological disorders. Here, we describe a line of research aimed at characterizing how the visual comprehension of events is impacted by healthy aging and by early Alzheimer's disease (AD). One characteristic of aging is that older adults segment ongoing activity into events less well than do younger adults. However, this general pattern is moderated by individual differences, and is amplified by AD. Impaired event segmentation is associated with reduced subsequent memory and impaired action performance. Superior event perception is associated with greater neural synchrony in the right posterior temporal sulcus and left dorsolateral prefrontal cortex. These results suggest that interventions to improve event segmentation or online event memory representations may help visual comprehension and memory in aging and AD .

## Towards a processing model of visual narratives

The past decade has seen a rapid growth of studies on visual narrative in the cognitive and brain sciences, in static form often focusing on the sequential images in comics. Neil Cohn will summarize and integrate a growing literature of both behavioral and neurocognitive research into a model of sequential image processing. Complex visual narratives
involve an interaction between two processing streams. An ongoing semantic understanding builds meaning into a growing mental model of a visual discourse. Discontinuity across dimensions of spatial, referential, and event information then incur costs when discontinuous with the growing context. In parallel to these processes, a structural system organizes semantic information into coherent sequences using a narrative grammar that maps semantic information to categorical roles, which are then embedded within a hierarchic constituent structure. This system allows for specific predictions of structural sequencing on the basis of constructional schemas, independent of semantics. Together, these interacting streams engage an iterative process of retrieval of semantic and narrative information, prediction of upcoming information based on those assessments, and subsequent updating based on discontinuity. These core mechanisms are argued to be domain-general, as suggested by similar electrophysiological brain responses generated in response to sequential images, music, and language.

## References

Cohn, N. (2013a). The visual language of comics: Introduction to the structure and cognition of sequential images. London, UK: Bloomsbury.
Cohn, N. (2013b). Visual narrative structure. Cognitive Science, 37(3), 413-452. doi:10.1111/cogs. 12016
Foulsham, T., Wybrow, D., \& Cohn, N. (2016). Reading without words: Eye movements in the comprehension of comic strips. Applied Cognitive Psychology, 30, 566-579. doi:10.1002/acp. 3229
Gernsbacher, M. A. (1990). Language Comprehension as Structure Building. Hillsdale, NJ: Lawrence Earlbaum.
Henderson, J. M. (2017). Gaze control as prediction. Trends in Cognitive Sciences, 21(1), 15-23.
Magliano, J. P., \& Zacks, J. M. (2011). The Impact of Continuity Editing in Narrative Film on Event Segmentation. Cognitive Science, 35(8), 1489-1517. doi:10.1111/j.1551-6709.2011.01202.x
Smith, Tim J. (2012) The attentional theory of cinematic continuity. Projections 6 (1), pp. 1-27. ISSN 1934-9688.
Smith, Tim J. and Mital, P.K. (2013) Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. Journal of Vision 13 (8), ISSN 1534-7362.
Zacks, J. M. (2014). Flicker: Your Brain on Movies. Oxford, UK: Oxford University Press.

# Game-XP: Action Games as Cognitive Science Paradigms 

Organizer E Moderator: Wayne D. Gray<br>(wayne.gray.cogsci@gmail.com)<br>Rensselaer Polytechnic Institute

Keynote: Ray Perez
(ray.perez@navy.mil)
Office of Naval Research

Martin V. Butz<br>(martin.butz@uni-tuebingen.de)<br>Universität Tübingen

Stuart Reeves<br>(stuart.reeves@ nottingham.ac.uk)<br>University of Nottingham

Matthew D. Sangster Tom Stafford<br>(sangsm@rpi.edu)<br>Rensselaer Polytechnic Institute<br>(t.stafford@sheffield.ac.uk)<br>University of Sheffield

Discussant: Fernand Gobet<br>(Fernand.Gobet@liverpool.ac.uk)<br>University of Liverpool

Why games? How could anyone consider action games as experimental paradigms for Cognitive Science? In 1973, as one of three strategies he proposed for advancing Cognitive Science, Allen Newell exhorted us to "accept a single complex task and do all of it." More specifically, he told us that rather than taking an "experimental psychology as usual approach" that, we should "focus on a series of experimental and theoretical studies around a single complex task" so as to demonstrate that our theories of human cognition were powerful enough to explain, "a genuine slab of human behavior" with the studies fitting into a detailed theoretical picture. Action games represent the type of experimental paradigms that Newell was advocating and the current state of programming expertise and laboratory equipment, along with the emergence of Big Data (Griffiths, 2015) and Naturally Occurring Data Sets (NODS, Goldstone \& Lupyan, 2016), provide the technologies and data needed to realize his vision. Action Games enable us to escape from our field's regrettable focus on novice performance to develop theories that account for the full range of expertise through a twin focus on expertise sampling (across individuals) and longitudinal studies (within individuals) of simple and complex tasks.

This Symposium is inspired by the recent Action Games as Experimental Paradigms for Cognitive Science (GameXP), issue of Topics in Cognitive Science (topiCS), April 2017. It includes late-breaking work from some of the researchers represented in that topic as well as new work by new researchers.

## Symposium Presentations - in Brief

- Ray Perez provides our keynote and focuses on the long history and current promise of action games as

[^2]a research tool for understanding theory and as a delivery vehicle for training.

- Martin Butz introduces a new, hybrid cognitive architecture which takes as its domain the world of SuperMario Brothers ${ }^{\mathrm{TM}}$.
- Stuart Reeves introduces our field to Ethnomethodology and Conversation Analysis (EMCA) research and to the questions this community asks of games.
- Matt Sangster uses 1.9 million records from 539 thousand matches of teams of 5 vs 5 people who play the world's most popular game, League of Legends, to study the distinction between individual and team expertise.
- Tom Stafford uses data from a cellphone game to investigate the efficacy of sleep consolidation versus wake-time distributed practice.
- Fernand Gobet provides calm and perspective by discussing this recent flurry of Game-XP from the 50+ yr perspective of cognitive science research on Chess.


## Ray Perez - Prequel to Game-XP: Time Spent, Player Age, Effects of Game Play, and Understanding Expertise

Within two months of its release, the video game Call of Duty: Black Ops (Activsion Publishing Inc., 2010) was played more than 600 million hours worldwide, with sales of over one billion dollars (Albnanesius, 2010). U.S children spent on average 1 hr and 13 min playing video games every day in 2010, a $300 \%$ increase from 1999 (Rideout, Foehr, \& Roberts, 2010). The average age of game players is 31, with $29 \%$ under 18 years old, $32 \%$ in the $18-35$ year range, and $39 \%$ being 36 year or older. Video game playing is pervasive in our society; there is growing body of evidence that video gaming can have beneficial effects on the organization and
function of the brain and learning. However, very little is known about the long term effects of playing video games or what is learned from these games or for that matter what expert game playing looks like and how it develops. This symposium explores the nature of extreme expertise in game playing with the focus that understanding expert game playing should inform theories of cognition.

## Martin V. Butz - Mario becomes Cognitive

The SEMLINCS cognitive architecture uses the world of SuperMario ${ }^{\mathrm{TM}}$ as its task environment and endows Mario, himself, with agency. Mario learns a conceptual, generative model of its environment in the form of probabilistic production rule-like structures from its own, autonomously gathered, continuous sensorimotor experiences. As a result, SEMLINCS enables Mario not only to plan and control environmental interactions in a versatile, goal-directed, selfmotivated manner - focusing, for example, on rescuing the Princess or on gathering coins - but also to verbalize this knowledge and to receive additional knowledge linguistically (Schrodt, Kneissler, Ehrenfeld, \& Butz, 2017).

## Stuart Reeves - The Ethnomethodogy of Games

Like Cognitive Science, Ethnomethodology and Conversation Analysis (EMCA) focuses on human activities that occur during game play. Yet EMCA sets cognitive explanations of human action to one side and instead describes how such action is practically, witnessably achieved. The presentation will demonstrate the utility of the EMCA approach and emphasize the contrast between the questions asked by CogSci and EMCA (Reeves, Greiffenhagen, \& Laurier, 2017).

## Matthew Sangster - Finding the " $I$ " in Team

Can we study the contribution of individual performance in a team setting? Can we find the " I " in team? Using Big Data ( 1.9 million records from 539 thousand matches) from League of Legends ${ }^{\text {TM }}$ Sangster says "yes" and presents steps towards establishing a measure that can evaluate individuals performing in a team context.

## Tom Stafford - Sleep Consolidation: A Field Study

Few researchers would expect that data collected from a cellphone action game could be used to address questions of sleep consolidation in skill learning. Realizing that their data set of $\mathrm{Axon}^{\mathrm{TM}}$ games included long breaks between some games, they pulled out instances in which successive games occurred across sleep or non-sleep hours, with the former (but not the latter) being candidates for sleep consolidation
effects. The approach and results demonstrate the promise of Big Data to raise questions that have little or nothing to do with the paradigm used to collect the data (Stafford \& Haasnoot, 2017).

## Fernand Gobet - Discussant

Newell (1973) argued that progress in psychology was slow because its empirical research focused on answering binary questions rather than building theories that were powerful enough to explain "a genuine slab of human behavior". As measured by the work presented in this symposium and in its associated issue of Topics in Cognitive Science (Gray, 2017), Gobet (2017) attempts to assess the age-old question of any young field; namely, "are we there yet?"

## References

Gobet, F. (2017). Allen Newell's programme of research: the video-game test. Topics in Cognitive Science, 9(2).
Goldstone, R. L. \& Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. Topics in Cognitive Science, 8(3), 548-568. doi:10.1111/ tops. 12212
Gray, W. D. (2017). Games-XP: Action Games as Experimental Paradigms for Cognitive Science. Topics in Cognitive Science, 9(2), 1-19. doi:10.1111/tops. 12260
Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. Cognition, 135(SI), 21-23. doi: $\{10$. 1016/j.cognition.2014.11.026\}
Newell, A. (1973). You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In W. G. Chase (Ed.), Visual information processing (pp. 283-308). New York: Academic Press.
Reeves, S., Greiffenhagen, C., \& Laurier, E. (2017). Video gaming as practical accomplishment: ethnomethodology, conversation analysis and play. Topics in Cognitive Science, 9 (2), 1-35. doi:10.1111/tops. 12234
Rideout, V. J., Foehr, U. G., \& Roberts, D. F. (2010). Generation $m 2$ media in the lives of 8- to 18- year-olds. MenloPark, CA: Henry J. Kaiser Family Foundation.
Schrodt, F., Kneissler, J., Ehrenfeld, S., \& Butz, M. V. (2017). Mario becomes cognitive. Topics in Cognitive Science, 9(2), 1-31. doi:10.1111/tops. 12252
Stafford, T. \& Haasnoot, E. (2017). Testing sleep consolidation in skill learning: a field study using an online game. Topics in Cognitive Science, 9(2), 1-12. doi:10.1111/ tops. 12232

# Educating Spatial Thinking for STEM Success 

Mary Hegarty (mary.hegarty@psych.ucsb.edu)<br>Department of Psychological and Brain Sciences, University of California, Santa Barbara<br>David Uttal (duttal@northwestern.edu)<br>Department of Psychology, Northwestern University<br>Mike Stieff (mstieff@uic.edu)<br>Department of Chemistry and Learning Sciences Research Institute, University of Illinois, Chicago<br>Tom Lowrie (thomas.lowrie@canberra.edu.au)<br>Faculty of Education, Science, Technology and Mathematics, University of Canberra<br>Stella Vosniadou (stella.vosniadou@flinders.edu.au)<br>School of Education, Flinders University

Keywords: STEM education; spatial ability; visualization, representation, strategy, transfer.

## Overview

In recent years there has been new recognition of the importance of spatial thinking in Science, Technology, Engineering and Mathematics (STEM) disciplines, in part because of evidence that spatial ability predicts success and persistence in STEM (Wai, Lubinski \& Benbow, 2009), but is not fostered in our educational systems (National Research Council, 2006). Based on this evidence, current approaches aim to increase science achievement by training the types of general spatial skills measured by spatial ability tests. However, although there is considerable evidence that these spatial skills can be trained (Uttal, et al., 2013), there has been little evidence to date that training of general spatial skills transfers to success in STEM disciplines (Stieff \& Uttal, 2015).

In this symposium, we will take a critical approach to issues of how to educate spatial thinking, both by raising some theoretical questions about the nature of spatial thinking in STEM, and by considering a range of different approaches to enhance the development of spatial thinking at different educational levels (elementary, secondary, and college) and in different STEM disciplines. The participants will discuss a broad range of spatial challenges faced by students in STEM learning, including mastering discipline-specific spatial language, novel visuospatial representations, and the interplay between visualization and analytic reasoning strategies.

The four talks will be by researchers that differ in disciplinary expertise, methodologies, and theoretical frameworks. David Uttal, an expert in cognitive and developmental psychology will describe a program that develops $12^{\text {th }}$ grade students' spatial skills through the use of Geographic Information Systems (GIS). Mike Stieff, an expert in chemistry and learning sciences will describe how he has used theories of representational competence to design laboratory studies and classroom interventions that
improved spatial thinking in college-level chemistry by targeting students' understanding of domain-specific visuospatial representations. Tom Lowrie, an expert in mathematics education and assessment will describe an intervention conducted by elementary school teachers in the Australian school system, which improved students' spatial reasoning and transferred to mathematics achievement. Stella Vosniadou will describe laboratory studies that provide evidence for a shift from visual-spatial to analytic thinking with expertise in Geometry and Chemistry. She will interpret these results as an instance of conceptual change that raises questions about the relationship of spatial reasoning to STEM problem solving as learning progresses. Mary Hegarty will introduce the topic, moderate the symposium and lead a discussion on lessons learned about the nature of spatial thinking and how it can be best fostered in our educational systems.

## Training Spatial Problem-Solving (David Uttal)

Many studies have demonstrated that spatial skills strongly predict STEM achievement and attainment, and that spatial skills can be improved through training and experience (e.g., Uttal et al., 2013). However, most of these studies have focused only on psychometrically-assessed spatial skills, such as mental rotation. Although important, it seems likely that STEM skills will involve more than these core skills. For example, learning STEM will involve higher-order spatial skills, such as reasoning about patterns and distributions, and determining how best to represent and make decisions about spatial data. Therefore, we have created a program that emphasizes the usual of spatial data and extensive mapping challenges with Geographic Information Systems (GIS) to facilitate the development of higher-order spatial skills (Jant et al, 2013). $12^{\text {th }}$-grade students in several high schools completed the curriculum. They demonstrated improvement in spatial reasoning and more general scientific problem solving, indicating that higher-order spatial skills can be enhanced.

## Improving Spatial Thinking in STEM through Representational Competence (Mike Stieff)

Using the domain of chemistry as a context, I will explore the design of interventions that enhance spatial thinking by improving students' representational competence (i.e., skills related to interpreting, transforming, and creating visuospatial representations). Spatial thinking with visuospatial representations is central to learning and problem solving in all STEM fields. Students with low spatial ability have more difficulty interpreting visual representations in STEM courses, fueling deficit models of who can succeed in STEM fields and motivating educational interventions that aim to train general spatial ability independent of disciplinary content. Interventions developed in my laboratory include alternative strategy training (Stieff, et al. 2014), modeling activities (Stieff, et al., 2016a) and gesture (Stieff, et al., 2016). Each intervention has yielded significant improvements in representational competence and student achievement on spatially-demanding assessments. In three experiments, I will show that representational competence is highly responsive to instruction and demonstrate that students who might otherwise be excluded from STEM degree programs based on their spatial ability can attain successful learning outcomes with appropriate support.

## Developing Spatial Reasoning Programs for STEM learning: Empowering Classroom Teachers to Embed Intervention into Practice (Tom Lowrie)

Although there has been considerable research on how to improve spatial ability, few studies have considered the effect of spatial training on STEM learning (Stieff \& Uttal, 2015); despite evidence that improving spatial thinking can improve skills necessary to succeed in STEM disciplines (Uttal, et al., 2013). In fact, even very limited spatial training seems to improve student's mathematics skills (Cheng \& Mix, 2014). However, current spatial intervention programs are not likely to have much impact on school curricula, since the training is not embedded within daily classroom practices. Recently, a classroom-based spatial intervention study demonstrated improvements in students' spatial and mathematics performance (Lowrie, Logan \& Ramful, in press). The intervention was implemented by students' own classroom teachers. This presentation will focus on the need for spatial intervention programs to be framed around meaningful pedagogical frameworks, informed by cognitive science, aligned to school curricula, and implemented by classroom teachers.

## The Paradoxical Relation between Spatial Reasoning and Success in STEM (Stella Vosniadou)

I will present results from two studies which used a visual/analytic strategy task to investigate changing relations in the adoption of visual/spatial and analytic strategies in geometry and chemistry. The results showed that a) there is increasing reliance on the adoption of
analytic strategies with the development of domain expertise (see also Stieff et al., 2014), and b) that this reliance seems to depend on domain knowledge rather than on individual differences in spatial reasoning (Kospentaris et al., 2016; Vlcacholia et al., 2015). Given the convincing evidence that spatial reasoning abilities can predict success in STEM disciplines (Wai et al., 2009), the finding that problem solving in expert scientists increasingly relies on specialized, domain-specific analytic approaches raises important questions about the exact relationship between spatial reasoning and scientific problem solving.

## References

Cheng, Y.-L., \& Mix, K. S. (2014). Spatial training improves children's mathematics ability. Journal of Cognition and Development, 15(1), 2-11.
Jant, E. W., Meadow, N., Uttal, D. H., Hund, A., \& Kolvoord, R. (2013, April). Using GIS in project based curriculum: Influence on students' approach to problem solving. Poster presented at the biennial meeting of the Society for Research in Child Development, Seattle.
Kospentaris, G., Vosniadou, S., Kazi, S., \& Thanou, E. (2016) Visual and analytic strategies in geometry, Frontline Learning Research, 4(1), 40-57
Lowrie, T., Logan, T., \& Ramful, A. (in press). Visuospatial training improves elementary students' mathematics performance. British Journal of Educational Psychology.
National Research Council. (2006). Learning to think spatially. Washington, D.C.: National Academies Press.
Stieff, M., Dixon, B. L., Ryu, M., Kumi, B., \& Hegarty, M. (2014). Strategy training eliminates sex differences in STEM spatial problem solving. Journal of Educational Psychology, 106(2), 390-402.
Stieff, M., Lira, M., \& Scopelitis, S. A. (2016). Gesture as a strategic resource for spatial thinking in STEM problem solving. Cognition \& Instruction, 34(2), 80-99.
Stieff, M., Scopelitis, S. A., Lira, M., \& DeSutter, D.
(2016). Improving representational competence in organic chemistry with concrete models. Science Education, 100(2), 344-363.
Stieff, M., \& Uttal, D. (2015). How much can spatial training improve STEM Achievement? Educational Psychology Review, 27(4), 607-615.
Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., \& Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. Psychological Bulletin, 139(2), 352-402.
Vlacholia, M., Vosniadou, S., Salta, K., Roussos, P., Kazi, S., Sigalas, M., \& Tsougraki, C (2015, July). Investigating the Visual/Analytic shift in students, knowledge of chemistry, Poster presentation, at the $37^{\text {th }}$ Annual Meeting of the Cognitive Science Society, Pasadena, California, USA.
Wai, J., Lubinski, D., \& Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. Journal of Educational Psychology, 101 (4), 817-835.

# Symposium on Problem Solving and Goal-Directed Sequential Activity 

Pat Langley<br>Institute for the Study of Learning and Expertise<br>2164 Staunton Court, Palo Alto, CA 94306 USA<br>Richard P. Cooper<br>Department of Psychological Sciences, Birkbeck, University of London Malet Street, London, London WC1E 7HX UK

## Background and Motivation

Problem solving is one of the hallmarks of human cognition. The term covers a wide range of behaviors, including abilities for solving unfamiliar puzzles, designing new artifacts, generating extended plans, and pursuing complex routine activities. These each require people to carry out sequences of mental or physical steps to achieve their objectives. They can involve reasoning, subgoaling, recognizing alternatives, evaluating them, and guiding search through large spaces.

The study of problem solving played a crucial role in the early development of cognitive science as a field. Research on this topic revealed basic insights about the representations and processes that underlie high-level cognition. Empirical studies of human problem solving provided some of the first evidence for the computational nature of human thinking, and related computational models led to major theoretical advances concerning heuristic search, goal processing, expert performance, and production systems. There is little question that, without its early emphasis on problem solving, cognitive science would be a very different discipline.

In recent years work on this topic has been poorly represented at the annual Cognitive Science meeting. Some might draw the mistaken conclusion that research has stalled or that there remain no open issues. In fact, research has continued and has produced clear advances. Thus, problem solving or, more generally, goal-directed sequential activity is now typically understood within the context of the wider cognitive architecture, including how it uses domain-specific knowledge and heuristics in the service of goals. This symposium will draw together some of the recent work in this area, with the aims of highlighting progress and clarifying outstanding issues and contemporary research questions.

## Scope and Organization

The five talks in this symposium will report research that covers a wide range of issues within contemporary problem-solving research, from incubation processes on insight tasks to the use of heuristics by experts in goaldirected design. What the research has in common is a concern with activity over time that is goal directed but also situation aware.

Thomas Ormerod will examine the development and testing of computational models of insight, with a focus on capturing differences between problem-solving tasks with unitary or multiple architectures, the difficulty of modelling apparently non-monotonic processes, and whether insight is governed by special or general cognitive processes. A meta-analysis by Sio and Ormerod (2009) of incubation effects found differences on linguistic puzzles such as Remote Associates tasks and on visual puzzles like the nine-dot problem, and similar task-based differences occur with sleep and analogy. His presentation will examine the extent to which different architectures and mechanisms, such as activation of associative networks (Monaghan et al., 2013) or goal-directed search for problem representations (Ormerod et al., 2013), are needed for different puzzle types, and will report models developed for both types of problem.

Colleen Seifert will discuss creative problem solving in design, focusing on how designers intentionally introduce variation. Consideration of multiple candidate concepts early in the design process is linked to better solution outcomes, but creating divergent pathways within the sequential activity of problem solving requires additional processes oriented to this goal. As in many areas of expertise, use of analogies with past solutions or precedents can be usefully applied in creative problem solving. Her Design Heuristics approach (Yilmaz et al., 2016) distills knowledge of design precedents to serve as generative constraints to guide divergent thinking. The heuristics are captured from studies of successful design outcomes within a wide variety of problem settings, including award-winning products, a longitudinal case study of an industrial designer, and protocol studies of industrial and engineering designers. Compilation of 3450 design outcomes revealed 77 design heuristics that introduce intentional variation into the generation process. These 'cognitive shortcuts' guide processing towards more, and more varied, design solutions.

Dario Salvucci will examine people's ability to perform multiple tasks at the same time. Often, multitasking is viewed as involving two separate and distinct activities, and indeed such multitasking appears often in the everyday world (the literal and metaphorical "walking
and chewing gum"). However, multitasking often occurs in service of a single goal, with multiple 'threads' of processing performing different actions that eventually come together to complete a single purpose. Threaded cognition (Salvucci \& Taatgen, 2008, 2011) is a computational theory, embedded within the ACT-R cognitive architecture, that aims to explain the power and limitations of human multitasking. In his presentation, he will discuss the theory and implications of threaded cognition for problem solving and other goal-oriented sequential activities, especially in the context of concurrent multitasking and task interruptions.

Richard Cooper will consider problem solving in terms of a core distinction between routine and nonroutine behaviour. His architecture is based on Norman and Shallice's (1986) dual-systems theory of the control of thought and action. In this account, routine or overlearned behaviour, while goal oriented, is schema driven and controlled by an activation-based 'automatic' system. In contrast, nonroutine behaviour involves higherlevel cognitive processes that bias the routine system's activation in a deliberative, goal-directed fashion (Cooper et al., 2014). He will argue that human cognition requires: (a) explicit representation of subroutines, including their goals or effects, (b) hierarchically structured task knowledge, to support flexible and creative combination of subroutines in novel ways, and (c) control mechanisms that monitor progress towards goals, suppress prepotent response schemas, and recall relevant episodic memories to support analogical planning. He will contrast these features with those that underlie recent machine learning accounts of sequential activity.

Pat Langley will present a new architectural theory that addresses four issues typically neglected in accounts of problem solving. One is an embodied agents' need to represent and reason about both qualitative relations and quantitative attributes when describing states. A second is the relation between symbolic goals and numeric evaluation functions, which address different aspects of purpose-driven behavior. A third issue concerns the introduction of agents' top-level goals and their change over time. A final topic is the great variability observed in human problem solving, both across people and task settings. He will present a new cognitive architecture that incorporates ideas from earlier work but introduces new structures and processes that address these challenges (Langley et al., 2016).

In order to ensure coherence, presenters will each consider the problem-solving phenomena of interest, representational issues, and relevant architectural processes, such as retrieval, attention, and goal handling. Cooper and Langley will jointly moderate the session, summarizing the symposium aims, introducing each of the presenters, and ensuring the question-answer session remains timely and on topic.

## Concluding Remarks

Taken together, these presentations will offer a broad sample of current research on problem solving and sequential activity. Each speaker has contributed to this area for many years and is well known for his or her accomplishments. Their topics range from creativity and insight to routine behavior on complex tasks. Their research builds on empirical studies of cognition but also contributes to architectural accounts of the mind.

We believe that this diverse set of presentations will convince conferences attendees that problem solving remains a critical area of enquiry within Cognitive Science, with both continuing theoretical progress and outstanding challenges. We further hope that the symposium will motivate audience members to join the quest to understand this fundamental aspect of human cognition.

## References

Cooper, R. P., Ruh, N., \& Mareschal, D. (2014). The goal circuit model: A hierarchical multi-route model of the acquisition and control of routine sequential action in humans. Cognitive Science, 38, 244-274.
Langley, P., Barley, M., Meadows, B., Choi, D., \& Katz, E. P. (2016). Goals, utilities, and mental simulation in continuous planning. Proceedings of the Fourth Annual Conference on Cognitive Systems. Evanston, IL.
MacGregor, J. N., Ormerod, T. C., \& Chronicle, E. P. (2001). Information-processing and insight: A process model of performance on the nine-dot and related problems. Journal of Experimental Psychology: Learning, Memory and Cognition, 27, 176-201.
Monaghan, P., Ormerod, T., \& Sio, U. N. (2013). Interactive activation networks for modelling problem solving. Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop. San Sebastian: World Scientific.
Norman, D. A., \& Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. Davidson, G. Schwarz \& D. Shapiro (Eds.) Consciousness and self regulation (Vol. 4). New York: Plenum.
Ormerod, T. C., MacGregor, J. N., Chronicle, E. P., Dewald, A. D., \& Chu, Y. (2013). Act first, think later: Effects of maximization, minimization and lookahead on inferential planning in problem-solving. Memory \& Cognition, 41, 1096-1108.
Salvucci, D. D., \& Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. Psychological Review, 115, 101-130.
Salvucci, D. D., \& Taatgen, N. A. (2011). The multitasking mind. New York: Oxford University Press.
Sio, U.T. \& Ormerod, T.C. (2009). Does incubation enhance problem-solving? A meta-analytic review. Psychological Bulletin, 135, 94-120.
Yilmaz, S., Daly, S. R., Seifert, C. M., \& Gonzalez, R. (2016). Evidence-based design heuristics for idea generation. Design Studies, 46, 95-124.

# LUCID science: Advancing learning through human-machine cooperation. 

Timothy T. Rogers ${ }^{1}$ (ttrogers@wisc.edu) and Charles W. Kalish ${ }^{2}$ (cwkalish@wisc.edu)<br>${ }^{1}$ Department of Psychology, 1202 W. Johnson Street<br>${ }^{2}$ Department of Educational Psychology, 1025 W. Johnson Street<br>Madison, WI 53706 USA

Keywords: computational methods, modeling, education, collaboration

## Introduction

This symposium presents four collaborative research projects conducted as part of LUCID, a unique crossdisciplinary graduate training program funded by the NSF's National Research Traineeship mechanism. LUCID trains scientists from computational and behavioral disciplines to advance basic and applied research in domains where machines are used to instruct, predict, understand, respond to or learn from human users. Such human-machine interactions have a remarkably broad range of applicationin public and private education across the lifespan, industry and information technology, public and private health management, social networking and communication, robotics and human-computer interaction, national security, public policy, and, of course, basic research into the nature of learning, cognition, and intelligent behavior.

The current talks all consider how machine learning, cognitive modeling, and data-science might be integrated to address core questions in human learning and education. How can computational learning models best be leveraged to speed knowledge acquisition and breadth of transfer in educational contexts? How can we efficiently measure perceptual and cognitive structures online or in the lab? How do such structures change with increasing knowledge or expertise? How can cognitive models developed to explain behavior in simple lab-based tasks be extended to aid learning in educational contexts? And, if human beings are rational learners as most models assume, how do false beliefs arise and why are they so widespread?

The speakers consider answers to these questions that arise at the intersection of computer science, engineering, psychology, and education sciences. Sen, Meng, Matthews, Alibali, and Zhu consider how state of the art search techniques in machine learning, combined with cognitive models of human learning, can yield prescriptions for the optimal "diet" of practice in any given learning task. Mason, Nowak, and Rau describe research using a novel adaptive-sampling tool to measure the perceptual similarities discerned by undergraduates amongst diagrams of molecules, with the aim of understanding which perceptual features support or undermine a good understanding of the underlying chemical structure. Binzak, Sievert, Murphy and Hubbard apply contemporary multidimensional scaling algorithms to show that single digit number concepts differ qualitatively in experts and
novices, and consider the implications for our developing understanding of numerical representation in the brain. Finally, Frigo and Rogers describe behavioral and simulation work suggesting a new hypothesis about how and why learning can go so wrong when information propagates in social networks.

Following these talks we will briefly lay out the challenges we have encountered in pursuing crossdisciplinary training of this kind with the goal of spurring a brief discussion session in which the audience can ask the program PIs and trainees about both the science and the training approach.

## Optimizing Human Learning with Machine Teaching

A long-standing but elusive goal in machine-aided education has been to exploit cognitive models of human learning to select teaching or practice experiences for students that will efficiently lead them toward the desired knowledge state. We show how contemporary optimization methods allow theorists to discover, for any implemented learning model and desired outcome, an optimal teaching set-that is, a model training set that most efficiently produces the desired outcome given the model. We then report experiments assessing whether thus approach can be used to speed human learning, taking arithmetic as an example domain. Prior work has shown that people employ different learning strategies depending upon the structure of their practice experiences. When practice is purely symbolic (e.g. flash-card learning) people acquire item-specific knowledge that does not generalize, whereas when practice highlights underlying quantitative relationships, people learn functions that transfer well to unpracticed problems. This suggests that the optimal teaching set-the practice experiences that most rapidly produce knowledge that transfers broadly-will differ qualitatively depending on whether practice is symbolic or quantitative. We describe a series of experiments testing these predictions with participants learning new arithmetic relations through a computer-mediated teaching system that controls how practice problems are sampled. The results highlight the potential for machine teaching and cognitive modeling to boost learning in important educational domains.

## Discovering perceived relations among molecular representations

To succeed in science courses, students must learn to rapidly and effortlessly translate among different visual
representations of key representational structures with a high degree of fluency. This is a difficult task because students must learn to interpret individual representations on their own while simultaneously learning the relations among different representations. To better understand these processes, we used a novel adaptive embedding algorithm to identify which molecular representations beginning undergraduate students find similar and why (e.g., Lewis structure, ball-and-stick). Each trial of the embedding task asks participants to decide which of two candidate diagrams is most similar to a third. The algorithm adaptively selects triplets for comparison in a manner that allows for efficient estimation of perceived dissimilarities amongst all diagrams. From these dissimilarities we generated models of how different molecules are embedded in a perceptual similarity space, in the eyes of the typical undergraduate student. The result revealed an otherwise inaccessible set of visual features that jointly predict the novice similarity judgments, allowing us to identify the features salient to novice students without relying on verbal mediation. The same tool can likewise be used to identify features that govern the perceptual decisions of chemistry experts, with the ultimate aim of developing interventions that guide novice perceptual attention toward the features discerned by experts.

## Beyond Magnitude: Psychological and Neural Representations of Number Properties

In classic work Roger Shepard and colleagues (1975) employed multidimensional scaling to show that, among the graduate students and colleagues who were his subjects, single-digit number concepts encode rich structure including primeness, parity, trinity, and exponentiation. This conclusion is hard to reconcile with much contemporary work suggesting that number concepts are grounded in an innate and widely-conserved approximate magnitude estimation system. In a series of studies, we used behavioral and brain imaging methods to investigate the psychological and neural mechanisms supporting adults' sensitivity to properties of number beyond magnitude, with the aim of reconciling this discrepancy. We first replicated Shepard's result in a cohort of students and colleagues, using a triadic judgment task to estimate conceptual similarities discerned amongst single-digit numbers. We then compared these representations among expert (math and CS grad students) and non-expert (Psychology undergraduates) groups, and found that rich structure was only observed in the experts. In a third study we examined whether explicit instruction can tune number concepts, with results revealing that magnitude information strongly dominates conceptual structure in non-experts but not experts. Finally, we have begun to assess what these behavioral differences suggest about the neural representation of number concepts. Participants viewed single-digit numbers while their brains were scanned in a slow event-related fMRI design. After a delay, they were instructed to think about a specific property of that number, and then were asked to judge whether that
number matched a target number on that specific property. Using multivariate pattern classification, we assessed whether magnitude, primeness, and parity could be decoded from the neural responses measured, both before and after the important property was cued on each trial. The comparison of behavioral and brain imaging results carries important implications for an understanding of numerical cognition beyond magnitude, and for the role of expertise in reorganizing conceptual representations of numbers.

## Why do false beliefs persist in crowds?

If human learning is rational as most cognitive models propose, what explains the emergence and widespread persistence of demonstrably false beliefs? We consider a new hypothesis that stems from an important difference between learning studies in the lab versus the real world. In the lab learners typically receive a single source of correct feedback, whereas in real life learners encounter many different sources of information that vary in their knowledge, motivation, and trustworthiness. How then do learners combine information from disagreeing sources? We examined how learners weight different sources when updating their beliefs, as a function of the degree to which the sources cohere with the learner's prior beliefs. The results reveal a previously undescribed learning bias that, counterintuitively, can lead groups of learners to disagree despite overwhelmingly similar learning experiences. To understand how this learning bias might lead to the emergence and persistence of false beliefs, we report simulation experiments in which many learners provide teaching labels to one another through a social network. Each simulated learner updates its beliefs in accordance with the empirically-observed learning bias, with the consequence that the cohort fractionates into mutually distrustful subgroups that adhere to different beliefs and ignore feedback from out-group members. The work thus provides a candidate mechanism for understanding how incorrect beliefs can arise and why they persist, even if individual learners behave in accordance with rational models in lab-based studies.

## UW-Madison Participants

Computer Science: X. Zhu, A. Sen*<br>Educational Psychology: J. Binzak*, E. Hubbard, C. Kalish, P. Matthews, R. Meng*, M. Rau.<br>Electrical \& Computer Engineering: B. Mason*, R. Nowak, S. Sievert*<br>Psychology: M. Alibali, V. Frigo*, A. Murphy*, T. Rogers<br>*LUCID Trainee

# Intuitive Biology and Global Challenges: Applying Theoretical Insights for Public Good 

Lauren Swiney (l.swiney@warwick.ac.uk)<br>School of Life Sciences, University of Warwick, Coventry, CV4 7AL UK<br>Deborah Kelemen (dkelemen@bu.edu)<br>Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215 USA<br>Andrew Shtulman (shtulman@oxy.edu)<br>Department of Cognitive Science, Occidental College, Los Angeles, CA 90041 USA<br>John Coley (j.coley@northeastern.edu) \& Nicole Betz (betz.n@husky.neu.edu)<br>College of Science, Northeastern University, Boston, MA 02115 USA<br>Stefaan Blancke (st.blancke@gmail.com)<br>Department of Philosophy and Moral Sciences, Ghent University, Ghent, B-9000 Belgium

Keywords: intuitive biology; climate change; biotechnology; evolution; science literacy; applied cognitive science

## Motivation \& Overview

Questions about science literacy and the rejection of scientific consensus are once again in the spotlight, with freshly-ignited international debate over the facts of climate change, and continued controversy around the teaching of evolution. In recent years, cognitive scientists have made valuable contributions to these debates: a now substantial and diverse research field has implicated a range of cognitive, motivational and emotional factors that contribute to science acceptance, and researchers are increasingly concerned with the application of these insights to improve the quality of public debate and science-relevant policy.

In this symposium we focus on a specific strand of this research field - that related to the concepts and intuitions deployed in reasoning about the biological world. A defining feature of our symposium is the inclusion of nascent research programs exploring the role of biological reasoning in newly-emerging domains of public debate (e.g. synthetic biology), alongside more established research areas (e.g. climate change $\&$ evolutionary theory). Our core aims are two-fold: to advance key theoretical debates relating to reasoning in the biological domain via the presentation of new empirical data, and to highlight emerging best-practice in translating this basic research into applied tools in both formal learning and informal communication contexts. We are confident that this focus will be of interest not only to researchers in the broader areas of science literacy, reasoning and conceptual change, but also to those interested in the challenge of applying cognitive science research for the public good.

Within this theoretical and applied framework we bring together researchers from a variety of disciplinary backgrounds, including anthropology, philosophy, and psychology, to explore the themes of the symposium from the perspectives of human development, education, cognitive processing, moral reasoning, cultural variation, risk perception, and conceptual knowledge structure.

The symposium will consist of four talks and a panel discussion. Kelemen will present cross-cultural evidence for the developmental persistence of teleological biases, and describe the translation of these findings into earlyeducation tools. Shtulman will present data on the conceptual prerequisites for understanding evolution, and discuss implications for increasing support for evolutionrelevant policies. Coley \& Betz will present new work on intuitive reasoning about climate change. Swiney will present data on the interplay of intuitive biology and moral reasoning in shaping risk perceptions of synthetic biology and discuss related communication challenges. Blancke will lead the panel discussion, drawing on his own research bridging cognitive science and public understanding of biotechnology (Blancke et al. 2015). Together the participants have published several dozen papers in the area, including in PNAS, Psychological Science, Cognitive Science, Cognition, and Child Development.

## Kelemen: Purposefully Designing Materials for Teaching Children About Natural Selection

In a world where economies are increasingly fueled by biotechnological responses to rapidly adapting disease pathogens, pesticide-resistant insects, and climate change, understanding evolutionary processes is prerequisite for informed decision-making about bioethical issues. Despite this, the fundamental evolutionary mechanism of natural selection is one of the most misunderstood concepts in science. The roots of these misconceptions can be traced to intuitive cognitive biases emergent in early childhood. In this talk, I will overview evidence from Eastern (e.g. China) and Western (e.g. U.S.) cultures that suggests the universality and developmental persistence of biases to construe nature in terms of purpose and intentional design (e.g. Rottman et al., 2017; Schachner et al., 2017). I will describe the application of these child developmental and adult dual-processing findings to the design of explanationrich storybooks for teaching elementary school children about adaptation by natural selection. Findings reveal that after analogical discussion of two storybooks, young children accurately and enduringly generalize the theory of
natural selection. Implications for theories of conceptual change and early science education will be discussed.

Deb Kelemen is Professor of Psychological and Brain Sciences and Director of the Child Cognition Lab at Boston University.

## Shtulman: Why People Fail to Understand Evolution and Why it Matters

Evolutionary theory underlies several issues of global importance-biodiversity, conservation, antibiotics, chemotherapy, cybersecurity-but studies have shown that the general public misunderstands what evolution is and how evolution works (Shtulman \& Schulz, 2008; Shtulman \& Calabi, 2013). In this talk, I will explore three conceptual prerequisites for understanding evolution: geologic time, intraspecies variation, and intraspecies competition. All three concepts have been implicated in the discovery of natural selection in the history of science, and I will show that all three concepts explain a significant amount of variance in who understands evolution and who does not. Nevertheless, one concept in particular-intraspecies competition-explains nearly three times as much variance as that explained by the other two concepts combined. I will discuss the implications of these data for improving evolution education, as well as increasing public acceptance of evolution and public support for evolution-relevant policies.

Andrew Shtulman is Associate Professor of Psychology and Chair of the Department of Cognitive Science at Occidental College.

## Coley \& Betz: Intuitive Thinking Impacts Understanding of Global Climate Change

Although most US citizens believe that climate change is a serious issue, fewer engage in mitigative behaviors. One psychological barrier is lack of understanding of causes and effects (Bord, O’Connor \& Fisher, 2000). We examined the extent to which intuitive "cognitive construals" (essentialist, teleological, and anthropocentric thinking, Coley \& Tanner 2015) influence understanding of climate change. University students rated agreement with facts and misconceptions (consistent with cognitive construals) about climate change. We found that teleological thinking about the climate was negatively related to understanding the causes of climate change while anthropocentric thinking was positively related. Further, we found that essentialist and teleological thinking were negatively related to understanding the effects of climate change, while anthropocentric thinking was positively related. We discuss these findings in the context of broader debates about biological reasoning, and consider options for leveraging or mitigating intuitive beliefs to increase sustainable behavior.

Nicole Betz is a doctoral candidate and John Coley is Associate Professor and Director of the Conceptual Organization, Reasoning and Education Lab at Northeastern University.

## Swiney: Essentialism, Moral Reasoning, and Evaluations of Synthetic Biology

The field of Synthetic Biology (SB) is already realising its promise to re-engineer living things from the bottom-up, creating new life forms, drastically changing existing organisms, and heralding a level of human intervention in biology that challenges entrenched distinctions between the evolved and the designed. The cognitive sciences have much to offer the now-urgent public debates about the risks and benefits of such technologies, but cognitive research in this area remains in its infancy (Blancke et al., 2015). I introduce a research program drawing on theories from distinct areas of the cognitive sciences, including intuitive biology, risk perception, and moral psychology, highlighting the rich test-ground that SB provides for investigating the interplay of cognitive processes across these domains. I present data from a series of experiments in which participants evaluate specific SB technologies varying across dimensions such as the source of genetic material and the extent of genetic change. I show that both psychological essentialism and moral purity concerns shape moral judgments and risk assessments of SB, and I explore the unique challenges of applying these insights to public debate about biotechnology.

Lauren Swiney is a cognitive anthropologist and Research Career Development Fellow in the Warwick Integrative Synthetic Biology center.

## References

Blancke, S., Van Breusegem, F., De Jaeger, G., Braeckman, J., \& Van Montagu, M. (2015). Fatal attraction: the intuitive appeal of GMO opposition. Trends in Plant Science, 20(7), 414-418.
Bord, R. J., O'Connor, R. E., \& Fisher, A. (2000). In what sense does the public need to understand global climate change?. Public Understanding of Science, 9(3), 205-218.
Coley, J. D., \& Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. CBELife Sciences Education, 14(1), ar8.
Rottman, J., Zhu, L., Wang, W., Seston Schillaci, R., Clark, K. J., \& Kelemen, D. (2017). Cultural influences on the teleological stance: evidence from China. Religion, Brain \& Behavior, 7(1):17-26.
Schachner, A., Zhu, L., Li, J., \& Kelemen, D. (2017). Is the bias for function-based explanations culturally universal? Children from China endorse teleological explanations of natural phenomena. Journal of Experimental Child Psychology, 157, 29-48.
Shtulman, A., \& Schulz, L. (2008). The relation between essentialist beliefs and evolutionary reasoning. Cognitive Science, 32(6), 1049-1062.
Shtulman, A., \& Calabi, P. (2013). Tuition vs. intuition: Effects of instruction on naive theories of evolution. Merrill-Palmer Quarterly, 59(2), 141-167.

# Time in the mind of a child: Perspectives on the development of temporal cognition 

Katharine A. Tillman (katillman@ucsd.edu)<br>Department of Psychology<br>University of California, San Diego, USA

Meng Zhang (meng.zhang0904@rutgers.edu)<br>Department of Psychology<br>Rutgers University, USA

Florie Monier (florie.monier@gmail.com)
Laboratoire de Psychologie Sociale et Cognitive
Université Clermont Auvergne, France

Jonathan Redshaw (j.redshaw@uq.edu.au)<br>School of Psychology<br>University of Queensland, Australia

Teresa McCormack (t.mccormack@qub.ac.uk)<br>Department of Psychology<br>Queens University Belfast, UK

Keywords: time, cognitive development, concepts, event representation, future

## Introduction

The ways in which we experience and reason about time are fundamental aspects of human cognition. In the industrialized world, keeping track of time is vital to successful functioning in society. Ideas about the nature of time underlie many aspects of adult life: how we communicate with others, how we schedule our days, how we plan for the future, how we interpret and react to autobiographical events, and how we reason about cause and effect. Both philosophers and cognitive scientists have struggled to explain the nature and origins of this rich, multifaceted, and highly abstract concept of time. One means of exploring the nature of time in the adult mind is by asking how the ability to mentally represent and reason about time develops in children. Although some aspects of temporal cognition, like low-level duration perception, are present at birth, others, like using a clock, take children up to a decade to learn. By tracking how different time-related cognitive phenomena emerge and change across development, we may gain a fuller picture of how the many facets of time interrelate, including the biological and cultural factors that underlie them. To this end, this symposium brings together researchers from around the world to discuss five different aspects of children's temporal cognition, each of which change dramatically during the preschool years.

Each case study presented in the symposium investigates time in the context of a different cognitive system, including motor planning, spatial cognition, language, emotion, event representation, and prospective reasoning. First, Monier will discuss children's developing capacity to synchronize their movements to external temporal rhythm. Next, Tillman will examine the development of culture-specific spatial representations of time, such as the left-to-right "mental timeline." Zhang will explore how children learn time-related language, including the words "yesterday" and "tomorrow." Redshaw will investigate how children become able to hold two alternative possible futures in mind. Finally, McCormack will discuss her work on the development of emotional and value judgments about past
and future events. A leader in the field of temporal cognitive development, McCormack will also serve as symposium moderator.

## The role of motor and cognitive capacities in developmental differences in rhythmic <br> synchronization <br> Florie Monier \& Sylvie Droit-Volet

Rhythmic synchronization is the ability to synchronize a movement to an external rhythm. Prior studies have shown that, relative to older children and adults, younger children have a faster and more variable spontaneous motor tempo, and they have more difficulty slowing down their motor tempo to synchronize it with a slow external tempo. Rhythmic synchronization involves both motor and cognitive capacities. Here, we ask which of these factors drives developmental change. In a series of studies, 3- to 8-year-old children were given a spontaneous motor tempo (finger tapping) task, a synchronization task involving differing inter-stimulus-intervals (ISI), and a continuation task, in which they were asked to maintain the tempo initiated in synchronization. Neuropsychological tests were used to assess their motor and cognitive capacities. Our results showed that the variability of children's inter-tapintervals (ITI) decreased with increasing age, and that 8 -year-olds were better able to slow down their motor tempo with an external tempo than 3- or 5-year-olds. Both motor and cognitive abilities predicted individual differences in the length of ITI on the synchronization and continuation tasks. However, only cognitive capacities (i.e., short-term memory, attention-concentration) accounted for children's variability on the continuation task. These results reveal that improvement in synchronization capacities is related to the development of both motor and cognitive capacities, but that the continuation task is more cognitively demanding.

## The development of the mental timeline Katharine Tillman, Nestor Tulagan, Eren Fukuda, \& David Barner

When reasoning about time, English-speaking adults often invoke a "mental timeline" stretching from left to right. Although the direction of the timeline varies across cultures, linear representations of time have been argued to be
ubiquitous and primitive. On this hypothesis, we might predict that children also spontaneously invoke a spatial timeline when reasoning about time. However, little is known about how and when the mental timeline develops, or to what extent it is variable and malleable in childhood. We used a sticker placement task to test whether preschoolers spontaneously produce linear representations of temporal events (breakfast, lunch, and dinner) and deictic time words (yesterday, today, tomorrow), and to what degree those representations are adult-like. At age 4, children were able to make linear mappings between time and space with minimal spatial priming. However, unlike older children and adults, most preschoolers did not adopt linear representations spontaneously. Lines produced by children were also more variable in orientation and children could be easily primed to adopt an unconventional vertical timeline. Our findings suggest that preschoolers can readily form linear mappings between time and space to represent temporal sequences and past/future relationships when prompted to do so, but most do not yet do so automatically. These representations are initially flexible, and become increasingly automatic and conventionalized in the early school years.

## Children's understanding of 'tomorrow' and 'yesterday' <br> Meng Zhang and Judith Hudson

Children's use of temporal language is often taken as an indication of their understanding of time. This study used a picture-sentence matching paradigm to test children's understanding of the temporal adverbs yesterday and tomorrow. Children viewed two pictures of an object, e.g., a carved pumpkin and an intact pumpkin, while listening to a sentence, e.g., "I carved the pumpkin yesterday" or "I'm gonna carve the pumpkin tomorrow". They were asked to select one picture to match the sentence. Experiment 1 showed that 3-, 4-, and 5-year-olds all performed better when sentences were in the past tense than in the future tense. In Experiments 2 and 3, the sentences contained conflicting cues from tense and temporal adverbs, e.g., "I carved the pumpkin tomorrow". While adults selected pictures based on the temporal adverbs they heard, 4- and 5-year-olds tended to select pictures showing the outcome of actions, regardless of both tense and temporal adverb. In Experiment 4, children completed two additional tasks involving temporal reasoning. The Before \& After Task served as a baseline measure of temporal sequencing and the Yesterday \& Tomorrow Task tested children's understanding of yesterday in backward reasoning and tomorrow in forward reasoning. Results indicated that forward temporal reasoning is easier for children than backward temporal reasoning, and linguistically, they understand the term yesterday better than tomorrow.

## Young children's capacity to envision and prepare for mutually exclusive future possibilities Jonathan Redshaw, Talia Leamy, Phoebe Pincus, Jessica Crimston, \& Thomas Suddendorf

Because future events can be difficult to predict, adults often envision and prepare for multiple, even mutually exclusive alternatives. To investigate the emergence of this capacity in children, we developed a minimalist paradigm in which participants were given the opportunity to catch a ball dropped into a forked tube with two possible exits. The initial study showed that 2- and 3-year-olds often covered only one exit when preparing to catch the ball, whereas most 4 -year-olds spontaneously covered both exits from the first trial onwards. A follow-up study revealed a similar developmental pattern when the mechanism controlling the uncertain outcome was visible, rather than hidden within the tube. Additional follow-up studies, however, showed that 2- and 3-year-olds were much more likely to cover two exits when two balls were certain to drop from separate locations. These findings suggest that young children are not generally limited in reasoning about multiple future events, but rather they are specifically limited in reasoning about mutually exclusive possibilities. One potential explanation is that older children, unlike younger children, possess a metarepresentational understanding that their representations of future events can be incorrect, and so they take the opportunity to prepare for alternative versions.

## Temporal asymmetries in children's past and future thinking

Teresa McCormack, Agnieszka Jaroslawska, Patrick Burns, Aine Fitzpatrick, Jemma McGourty, \& Eugene Caruso
There are striking asymmetries in the way adults think about the past versus the future: adults typically (i) report stronger emotion when thinking about the future than the past (ii) place greater value on a future event than a past event and (iii) judge that a future event feels closer than a past event at an equivalent distance. These future biases suggest that adults are more oriented toward the future than the past. In this talk, I will describe the first developmental studies to examine whether these biases exist in children. We have found that the tendency to judge future events as feeling closer in time than equivalent past events is very robust, and can be demonstrated in children as young as 4-5 years. Children from around 6-7 years, like adolescents and adults, also report stronger emotions when thinking about the future versus the past. We also found that children from 9-10 years place greater value on future than past events. We will discuss what these findings suggest about children's thinking about the past and future, and their implications for theoretical accounts of temporal asymmetries.

# Big Data and Little Learners 

John C. Trueswell (trueswel@psych.upenn.edu)<br>(Symposium Organizer) Department of Psychology University of Pennsylvania, USA

Joshua B. Tenenbaum (jbt@mit.edu)<br>Department of Brain and Cognitive Sciences<br>Massachusetts Institute of Technology, USA

Linda B. Smith (smith4@indiana.edu)<br>Department of Psychological and Brain Sciences Indiana University, USA

Charles Yang (charles.yang @ling.upenn.edu)<br>Department of Linguistics and Computer Science<br>University of Pennsylvania, USA

Keywords: data science; deep learning; poverty of the stimulus; indirect negative evidence; language development

## Introduction

Recent advances in the data sciences, particularly within the area of language technology, have been impressive and nonincremental. For example, within the domain of language translation, the application of deep Long Short Term Memory (LSTM) neural networks to large bodies of text have resulted in a $60 \%$ reduction in translation errors from traditional methods, significantly closing the gap between machine and human performance (Wu et al., 2016). Similarly impressive advances have been observed in, e.g., speech recognition (Hinton et al., 2012), syntactic parsing (Dyer et al., 2015) and automatic content extraction (Berant et al., 2015).

Clearly, excitement is justified as a new era of linguistic technology is emerging. But should this excitement lead to a fundamental rethinking of our theories of child language and cognition? Doesn't the "poverty of the stimulus" still pose a problem for human language learners? What role do hierarchical linguistic formalisms play within statistical theories of language learning and use? This symposium brings together leading figures in cognitive science who offer different informed perspectives on these matters.

## The data and the learner from a developmental perspective

## Linda Smith (Indiana University)

The world offers data to learning systems that is massive in total scale and that comes in many forms. However, the relevant data for any learning system are only those that actually engage the learning mechanisms of that system. For living and breathing learners, this engagement begins with their sensory systems. Sensory systems are on bodies that move through the world - constrained by the physics of space and time - and thus the sampled data are constrained and ordered by space in time. Human infants learn their first words during a period in which their bodies (and brains) change dramatically and systematically and do so in ways that put those sensory systems in different parts of the data space at different points in development. The data for learning - and the learning tasks to be solved - are systematically ordered by development itself. This talk will present evidence from a large corpus of head-camera data recorded in infants' homes (over 500 million frames
extracted at 1 Hz ) that illustrate how human development (and the reality of bodies learning in space and time) fundamentally changes the questions to be asked and the computational answers to how language is learned.

## Existence proofs and computational mechanisms

## Charles Yang (University of Pennsylvania)

Mathematicians have always drawn a useful distinction between existence and constructive results. An analogy can be made in the study of language acquisition, especially in the age of Big Data and Big Machines. While distributional regularities can be captured by idealized statistical models, it is a different matter whether, and how, such regularities are exploited by computational mechanisms available to human children.

Consider the use of indirect negative evidence in language acquisition. A specific example concerns the predicative and attributive use of the so-called a-adjectives in English: "the cat is asleep/away" vs. "*the asleep/away cat". Indirect negative evidence can be formulated in certain probabilistic models of inference: the conspicuous absence of forms such as "the asleep cat" reduces the learner's confidence in the hypothesis that permits such expressions. While these models are presented as existence proofs without commitments to psychological mechanisms, they are still unlikely to succeed when evaluated against realistic statistical distribution of adjectives in a large corpus of child-directed English speech.

The alternative approach is to avoid the use of indirect negative evidence, and to develop a transparently mechanistic models that can be readily tested. I review the Tolerance Principle, a parameter-free model of inductive generalization, and its application to morphological and syntactic acquisition, including artificial language learning (joint work with Kathryn Schuler and Elissa Newport). Furthermore, the Tolerance Principle suggests that language acquisition may succeed only with small data, the kind similar to the small vocabularies of young children. This supports the view that cognitive and maturational constraints support rather than hinder language development.

## On what you can't learn from (merely) all the data in the world, and what else is needed

## Josh Tenenbaum (University of Edinburgh)

Recent successes with recurrent neural networks and other big-data techniques in AI applications raise the question of whether similar approaches might explain human language acquisition. How far can the data of language take us alone, with little other structure? I will first describe some experiments testing RNN models developed by Google that can perform some truly impressive feats in language technology, yet at the same time fail a number of basic tests of understanding syntax and semantics that cognitive scientists have long been interested in, as well as some new benchmarks that we have come up with. They often fail for interesting reasons, based on the differences between their linear (sequential) processing architecture and the hierarchical structure of thought, their emphasis on characterlevel modeling as opposed to words and phrases, and their lack of interfaces to core cognition outside language. Their successes and failures illustrate how both advocates and critics of early statistical language learning were correct Chomsky and Gleitman and Pinker were right after all, but Elman and Hinton were also right. They were just right about different things, and we can learn much by re-interpreting early debates.

As a way forward, I argue for combining smart statistics with more structured, hierarchical representations, interfacing to a cognitively grounded semantics. I report some promising results, although we are far from being able to implement this at the scale Google requires. I will also sketch ideas for how RNNs can make these more structured approaches work better, with the hope of integrating these often-opposing traditions to best make progress.

## References

Berant, J., Alon, N., Dagan, I, \& Goldberger, J. (2015). Efficient global learning and entailment. Computational Linguistics, 42, 221-263.
Dyer, C., Ballesteros, M., Ling, W., Matthews, A., \& Smith, N. (2015).Transition-based dependency parsing with stack long short-term memory. Proceedings of the $53^{\text {rd }}$ Annual Meeting of the Association for Computational Linguistics, 334-343.
Hinton, G. Deng, L., Yu, D., Dahl, G et al., (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, 29, 82-97.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W. et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv E-prints, 2arXiv:1609.08144.

# Burstiness across multimodal human interaction reveals differences between verbal and non-verbal communication 

Drew H. Abney (dhabney@indiana.edu)<br>Department of Psychological and Brain Sciences<br>Indiana University, Bloomington, IN 47405 USA

Rick Dale (rdale@ucmerced.edu)
Chris T. Kello (ckello@ucmerced.edu)
Cognitive and Information Sciences
University of California, Merced, Merced, CA 95340 USA

Max M. Louwerse (m.m.louwerse@uvt.nl)<br>Center for Cognition and Communication<br>Tilburg University, Tilburg, Netherlands


#### Abstract

Recent studies of naturalistic face-to-face communication have demonstrated temporal coordination patterns such as the synchronization of verbal and non-verbal behavior, which provides evidence for the proposal that verbal and non-verbal communicative control derives from one system. In this study, we argue that the observed relationship between verbal and non-verbal behaviors depends on the level of analysis. In a re-analysis of a corpus of naturalistic multimodal communication (Louwerse et al., 2012), we focus on measuring the temporal patterns of specific communicative behaviors in terms of their burstiness. We examined burstiness estimates across different roles of the speaker and different communicative channels. We observed more burstiness for verbal versus non-verbal channels, and for more versus less informative language sub-channels. These findings demonstrate a new method for analyzing temporal patterns in communicative behaviors, and they suggest a more complex relationship between verbal and non-verbal channels than suggested by prior studies.


Keywords: burstiness, multimodal communication, verbal and non-verbal communication

## Introduction

In cognitive science, a considerable number of studies have investigated the role of non-verbal communication in relation to verbal communication. The majority of these studies suggest an intrinsic relationship between verbal and non-verbal communication. For instance, a strong link has been shown between lexical access and gesturing, such that when people gesture, lexical access is facilitated (Rime \& Schiaratura, 1991). Also, the time gap between gesture and a familiar word is considerably shorter than the gap between gesture and an unfamiliar word (Morrel-Samuels \& Krauss, 1992), and when speech is disrupted, gestures are halted (Mayberry \& Jaques, 2000). Gesture is thought to be intrinsically related to language processing (Butterworth \& Morrissette, 1996) because most gestures occur when
people speak (McNeill, 1992), and because of evidence linking gesture with language development (Butcher \& Goldin-Meadow, 2000). In fact, non-verbal and verbal communication are sometimes argued to be so interwoven that gesture and speech are co-expressive manifestations of one integrated system, forming complementary components of one underlying process that helps organize thought (Goldin-Meadow, 2005; McNeill, 1992).

Louwerse, Dale, Bard, and Jeuniaux (2012) investigated the temporal relationship between matching behaviors in dialog partners, such as manual gesture in one speaker vs. the same manual gesture in the other speaker. By applying a cross-recurrence analysis, Louwerse et al. showed synchronized matching behavior in all categories (language, facial, gestural) that were investigated at temporal lags short enough to suggest imitation of one speaker by the other. Louwerse et al. concluded that the similarities between the different channels - verbal and non-verbal - demonstrated that the temporal structure of matching behaviors provided low-level and low-cost resources for human interaction.

So far, all studies focusing on the similarities between verbal and non-verbal communication, including Louwerse et al. (2012), focused on the temporal matching of verbal and non-verbal behavior. They tend not to investigate the temporal distribution of independent behavioral event dynamics. Complex behaviors such as human interaction tend not to show the strictest forms of synchrony, but instead are more loosely, functionally coupled (e.g., Fusaroli et al., 2014). Instead, the overall pattern of behavior, expressed in the distribution of events, may reflect particular local patterns of interaction - when one interlocutor gestures, it may sustain itself for a given period of time before waning; when another person speaks, this burst of behavior may look quite different, sustaining itself for longer, more regular periods of time. These event dynamics might paint a different picture of the relationship between verbal and non-verbal channels.

## The Property of Burstiness

Most work studying human communication is based on dyadic analyses that focus on temporal patterns across partners rather than the temporal patterns of specific behaviors produced by each partner. In the current study, the large multimodal corpus of human communication collected and reported in Louwerse et al. was re-analyzed to focus on the quantification of a particular property of behavior, burstiness.

Using the framework developed by Goh and Barabasi (2008) and extended by others (e.g., Jo, Karsai, Kertész, \& Kaski, 2012), we estimated the burstiness of verbal and nonverbal behaviors. The burstiness parameter, $B$, provides an estimate of a system's activity patterns spanning from periodic $(B=-1)$, to random ( $B=0$ ), to theoretically maximal burstiness $(B=1)$ (see Figure 1). Goh and Barabasi (2008) observed that human phenomena like human texts and email patterns have positive burstiness estimates, $B>0$, whereas human cardiac rhythms were found to have periodic burstiness estimates, $B<0$.


Figure 1: Overview of system's activity patterns spanning from periodic, to random, to theoretically optimally bursty.

## The Current Study

The goal of the current study was to investigate the temporal dynamics of behavioral events across verbal and non-verbal communicative modalities during face-to-face human interaction. We focus on the measure of burstiness, now widely used in statistical physics to capture the temporal patterns of point processes in complex network interactions.

In the first analysis section, we investigated whether or not there were differences in the burstiness of behaviors that are categorized into verbal and non-verbal channels. It is possible that verbal and non-verbal channels have similar degrees of burstiness, which would be consistent with previous work suggesting a strong intrinsic relationship. However, if the channels exhibit different degrees of burstiness, such results would suggest a more complex relationship between verbal and non-verbal communication. To further explore and understand the burstiness measure,
we also investigated the burstiness of sub-channels that constitute the language communicative channel. Our results indicate that burstiness is different for verbal versus nonverbal behaviors, and also for different aspects of language behaviors.

## Methods

## Multimodal Communication Corpus

The original task developed to collect these multimodal data is described by Louwerse et al. (2008) and Louwerse et al. (2012), who were interested in collecting multimodal structure of human interaction in order to inform avatar design for intelligent tutoring systems and other technologies. In the task, $N=24$ pairs of participants helped each other navigate a map. Each pair of participants completed 8 rounds of navigation. For each round, one participant was chosen as the "Information Giver", and other the "Information Follower." The Information Giver had a complete map, and the Information Follower had a noisy and partial map. This mismatch between maps was intended to elicit communication and predict the points at which misunderstandings were likely to occur. The participants had to use language and gesture via webcam so that the Information Follower could reconstruct a map route with the help of the Information Giver. The corpus was developed by taking these 192 recordings of interactions and coding a wide variety of behaviors. These codings were based on well-known or adapted coding schemes in discourse, along with some other semi-automated procedures (see Louwerse et al., 2008 for details). All behaviors were coded in 250 ms to encompass relatively fast behaviors such as nodding, acknowledgements, and smiling. The output from this coding procedure was a multicolumnar data format of binary point series that represented the occurrence of different behaviors at a 250 ms interval. These 250 ms intervals were the subject of our burstiness analyses.

We chose 39 behaviors that fit into four specific behaviors channels (as did Louwerse et al., 2012). Behavioral channels were categorized into two factors, Channel and Role. For the Channel factor, channels were identified as either "Face \& Head," "Manual Gesture," "Face Touch," or "Language." For the Role factor, channels were identified as either Giver or Follower. For the levels of the Role factor, all channels were included for the Giver and the Follower. See Table 1 for the behaviors that were included into each channel. The language sub-channels were annotated at the utterance-level.

Table 1: List of Channels, Sub-channels, and Behaviors

| channel | sub-channels | behaviors |
| :---: | :---: | :---: |
|  <br> head | mouth | laughing, lip tightening |
|  | eyes | blink, rolling eyes |
|  | eyebrows | asymmetrical, down- <br> frowning, out brow raiser |
| manual | head | nodding, shaking |
| gesture |  | beat, deictic, iconic, |
| meach |  | touching cheek, chinrest |
| face |  |  |

\(\left.$$
\begin{array}{ccc}\hline \text { language } & \text { dialogue acts } & \begin{array}{c}\text { acknowledgements, align, } \\
\text { check, clarify, explain, } \\
\text { instruct, query-what, } \\
\text { query-yes/no, ready, } \\
\text { reply-no, reply-what, } \\
\text { reply-yes }\end{array}
$$ <br>
\hline \& discourse \& alright, no, ok, um, well, <br>

connectives \& yes\end{array}\right]\)| descriptions | color, compass direction, <br> digit, relative direction, <br> spatial preposition |
| :---: | :---: |

## Construction of Multivariate Spike Trains and Inter-event Intervals

We are interested in estimating the burstiness of multimodal communicative behavior and are therefore working with a multivariate class of spike trains. To our knowledge, the current study provides the first steps towards dealing with burstiness in multivariate spike train corpora. The protocol converts multivariate spike trains into inter-event interval (IEI) distributions. These interval distributions help quantify the temporal clustering of communicative events across channels.

First, for each behavior, we created a spike train of onset events which excludes successive ' 1 's for prolonged events. Second, for each communicative channel (Face \& Head, Manual Gesture, Face Touch, Language), we summed the spike trains from each behavior, yielding a multimodal event series where a ' 0 ' represents a sample when no event occurred, a ' 1 ' represents a sample when one event occurred, and any number greater than 1 represents a sample when two or more events occurred. For example, a sample with a "Laughing" event and a "Nodding" event would have a " 2 " in the event series. Any sample with two or more events is considered a sample of simultaneous communicative behavior which we discuss below. Finally, IEI's were computed from the multimodal event series to construct an IEI distribution for each channel for each map task role (Giver or Follower).

An IEI is computed by considering two consecutive events, $t_{j}$ and $t_{j+1}$, and finding the temporal difference
between them, $\tau=t_{j+1}-t_{j}$. For an IEI that contains simultaneous communicative behavior (2 or more events in the same sample), an IEI, $\tau$, was computed and added to the distribution in addition to a zero for each additional event. For example, when an IEI with the second sample has 3 events, we would add to the IEI distribution (1) the corresponding $\tau$ and (2) two zeros $(0,0)$. We chose to add this component to the protocol because we wanted to treat simultaneous communicative behavior as quantitatively 'more bursty'. Adding zeros to an IEI distribution will amplify a burstiness estimate. IEI distributions for each communicative channel and each map task role were submitted to estimates of burstiness.

## Estimation of Burstiness

The burstiness parameter, $B$, is defined as,

$$
B=\frac{\sigma_{\tau}-m_{\tau}}{\sigma_{\tau}+m_{\tau}}
$$

where $\sigma_{\tau}$ is the standard deviation of the IEI distribution and $m_{\tau}$ is the mean of the IEI distribution (Goh \& Barabási, 2008; Jo, Karsai, Kertész, \& Kaski, 2012). Alternative measures of burstiness have been employed in previous studies in computational linguistics (Altmann, Pierrehumbert, \& Motter, 2009; Pierrehumbert, 2012) utilizing parameter fitting from a stretched exponential distribution (Weibull distribution). These alternative measures have provided unique insights into the dynamics of linguistic levels of description. Our decision to utilize the burstiness parameter, $B$, is twofold. First, parameter estimation from a distribution requires a minimum number of data points or IEIs. Therefore, with the properties of our corpus, parameter estimation from distribution fitting requires the implementation of confidence intervals, which can be avoided with the utilization of the burstiness parameter, $B$. Second, one goal of this study is to account for simultaneous communicative behavior as a higher degree of burstiness. The burstiness parameter, $B$, is amplified when zeros are added to the IEI distribution and therefore an ideal option for the current study. $B$ is bounded from $[-1,1]$, where $B=1$ for a theoretical maximum bursty behavior, $B=-1$ for completely regular behavior (e.g., metronome), and $B=0$ for a homogeneous Poisson process, i.e., independent events. We omitted trials that did not include reliable burstiness estimates for any of the four channels across the MapTask roles in the first analysis section ( $1.24 \%$ of trials) and for any of the three channels across the MapTask roles in the second analysis section ( $1.00 \%$ of trials).


Figure 2a and 2b: Burstiness across channels with a) Information Giver (G) and Follower (F) combined, and b) the roles separated. Error bars reflect 95\% CIs.

## Investigating Differences in Burstiness across Verbal and Non-verbal Channels

Mixed effects models (Bates et al., 2014; Team R., 2013) were utilized to determine if burstiness differed across different channels. The first set of analyses was conducted to compare burstiness estimates across role structure and communicative channels. Linear models were utilized to predict burstiness estimates. Fixed effects for these models included map task role (leader or follower), communicative channels (Face \& Head, Manual Gesture, Face Touch, and Language), and event count for each communicative channel. Event count was added into the model as a covariate to control for the potential relationship between burstiness estimates and the number of behavioral events going into the analysis. Dyad and map type were included as random effects.

If there are differences across communicative channels, we can observe such differences in a variety of ways: are there differences in the temporal structure across communicative modalities (1) collapsing burstiness estimates across MapTask roles? (2) within MapTask roles (e.g., Follower:Manual Gesture vs. Follower:Language)? and/or (3) across MapTask roles (e.g., Follower:Manual Gesture vs. Giver: Manual Gesture)?

Collapsing burstiness estimates across MapTask role, we observed a significant main effect of communicative channel, $F(3,1030)=162.55, p<.0001$ (Figure 2a). See Table 2 for results from multiple comparison tests. Overall, the language channel ( $M=.16, S E=.003$ ) was observed to be more bursty relative to the manual gesture channel ( $M=.14$, $S E=.01), b=.08, p=.009$.

Table 2: Multiple Comparisons from the random mixed effects model: ${ }^{*} p<.05,{ }^{* *} p<.01,{ }^{* * *} p<.001$.

|  | Multiple Comparisons | Beta | Z-score |
| :---: | :---: | :---: | :---: |
| Channel | Man. Gest. v. Face/Head | .08 | $7.9^{* * *}$ |
|  | Touch Face v. Face/Head | .11 | $9.6^{* * *}$ |
|  | Language v. Face/Head | .17 | $16.7^{* * *}$ |
|  | Touch Face v. Man. Gest. | .02 | $2.2^{* *}$ |
|  | Language v. Man. Gest. | .08 | $7.5^{* * *}$ |
|  | Language v. Touch Face | .05 | $4.6^{* *}$ |
| Role | Leader v. Follower | .01 | .7 |
| Int. | F:Man. Gest v. F:Lang | -.07 | $-4.79^{* * *}$ |
|  | G:Man. Gest. v. G:Lang | .05 | $3.17^{*}$ |
|  | F:Man. Gest v. G:Lang | .06 | $3.08^{*}$ |
|  | G:Man. Gest v. F:Lang | .04 | 2.79 |

The communicative channel x map task role interaction was significant, $F(3,1030)=20.97, p<.0001$, therefore, we tested for multiple comparisons using Tukey Honestly Significant Difference tests to investigate differences within and across MapTask roles (Figure 2b). At this level of the analysis, we were specifically interested in the differences between language and manual gestures, so we limit our report to those subsets of the analysis. We observed withinrole differences between language and manual gesture burstiness estimates for the Follower role ( $b=-.07, p<$ $.001)$ and for the Giver role ( $b=.05, p=.03$ ). We also observed a between-role difference for Follower: Manual Gesture v. Giver: Language ( $b=.06, p=.04$ ). The results from this analysis suggest that, across map task role, the verbal channel (i.e., language channel) had higher burstiness estimates relative to the non-verbal channels, and specifically the manual gesture channel.

## Investigating the Relative Magnitude of <br> Burstiness in the Language Channel

In the last section, we established that communicative channels exhibit temporal patterns of behavior that (1) vary across verbal and non-verbal channels and (2) are all bursty relative to exhibiting random or periodic temporal patterns. But what does it mean to be more bursty? It is important to note that these channels are made up from specific subchannels that are further made up from individual behaviors. In an effort to better understand the relative magnitude of burstiness, in this section, we focused on the language channel because this channel exhibited the highest estimates of burstiness. Specifically, we zoomed into the language channel and investigated the temporal patterns of the subchannels.

The language channel is made up of three specific subchannels: dialogue acts, discourse connectives, and descriptions. We expected to observe higher burstiness estimates for the 'descriptions' sub-channel relative to the other two channels. This hypothesis is motivated by previous research that focused on the burstiness of various linguistic levels in texts (Altmann, Cristadoro, \& Esposti,

2012; Altmann, Pierrehumbert, \& Motter, 2009). Altmann et al. (2009) observed that burstiness increased across semantic classes where 'entities' like proper nouns had higher burstiness estimates relative to predicates like in, which in turn had higher estimates than higher level operators like the. If the results observed in texts are consistent with human dialogue, we should expect to observe that descriptions like providing a relative direction will have higher burstiness estimates relative to dialogue acts like saying no or discourse connectives like saying um.


Figure 3a and 3b:
Burstiness across language channels with a) Information Giver (G) and Follower (F) combined, and b) the roles separated. Error bars reflect 95\% CIs.

Linear models were utilized to predict burstiness estimates. Fixed effects for these models included map task role (Giver or Follower), language sub-channels (Dialogue Acts, Discourse Connectives, Descriptions), and event count for each communicative channel. Similar to the previous analysis section, event count was added into the model to act as a covariate to control for the potential relationship between burstiness estimates and the number of events going into the analysis. Dyad and map type were again included as random effects. We observed that descriptions ( $M=.08, S E=.005$ ) had higher burstiness estimates relative to discourse connectives $(M=-.06, S E=.004, b=.06, p<$ .001 ) and dialogue acts ( $M=-.11, S E=.004 ; b=.17, p<$ .001) (Figure 3a). Discourse connectives and dialogue acts were both more periodic than bursty, and dialogue acts were more periodic (closer to -1) relative to discourse connectives ( $b=.11, p<.001$ ). These results suggest that various levels of verbal dialogue have different temporal patterns and such patterns have interesting parallels to previous research studying the burstiness of text corpora. We discuss these parallels in addition to the insights gained from the analysis section to better understand the pattern of results in the previous analysis section.

## Discussion

The primary goal of the current paper was to better understand the temporal patterns of verbal and non-verbal behaviors during face-to-face multimodal human communication. We submitted the multimodal corpus to an analysis of burstiness. In the first analysis section, we observed that communicative channels differed in the degree of burstiness, with the verbal channel having higher burstiness estimates relative to non-verbal channels like manual gestures, face \& head, and face touch. To add nuance to this result, in the second analysis section, we focused on better understanding the magnitude of burstiness, and zoomed into the language channel. In this analysis, we observed that a more informative sub-channel, 'descriptions', had higher burstiness estimates relative to sub-channels that focused on operators and modifiers.

Much work in the cognitive sciences has argued that verbal and non-verbal behaviors are intrinsically related via the same communicative system (Golden-Meadow, 2005; McNeill, 1992). Recent work (Louwerse et al., 2012) has made this argument by focusing on evidence of synchronization across verbal and non-verbal channels. In the current paper, we observed that, verbal and non-verbal channels differ in terms of estimates of their temporal burstiness. An important question is what these differences reflect. To begin to find an answer to this question, we examined certain language sub-channels and found higher degrees of burstiness for descriptive productions compared to pragmatic productions like dialog acts or connectives.

Considering the latter results, there are a few possible explanations for the observation that verbal and non-verbal channels exhibit different types of temporal patterns, with the verbal channel exhibiting higher burstiness estimates. The first possible explanation is that increased estimates of burstiness for the verbal channel means that more information is contained within this communicative channel relative to the non-verbal channels. This suggestion is influenced by the observations of higher degrees of burstiness in higher-level semantic classes in texts (Altmann, et al., 2009) and higher degrees of burstiness in descriptive sub-channel in dialogue (the current paper's second analysis section). If this is the case, our results point to the proposal that verbal channels during human communication are more informative relative to non-verbal channels. However, this possibility seems unlikely because our own results show that the direction of burstiness estimate differences for the language and manual gesture channels are not consistent: higher estimates for language relative to manual gesture for the information giver and higher estimates for manual gesture relative to language for the information follower.

The second possible explanation is that an important property of multimodal communication is having a collection of different types of temporal patterns across communicative channels. This proposal, what we call the
'temporal heterogeneity' hypothesis, suggests that successful communication emerges from a diverse suite of information channels that vary in temporal properties. An important adaptive property of a complex system, such as a dyadic communicative system (Dale, Fusaroli, Duran, \& Richardson, 2013; Fusaroli, Raczaszek-Leonardi, \& Tylén, 2013), is the ability for multiple components with specific intrinsic properties to self-organize to form higher-level structures (Kello \& Van Orden, 2009; Kugler \& Turvey, 1987). This proposal is amenable to the hypothesis that verbal and non-verbal channels are part of the same integrated system (Golden-Meadow, 2003; McNeill, 1992) and that gesture and speech are complementary communicative channels important for the resolution of referential expressions (Louwerse \& Bangerter, 2010; Seyfeddinipur \& Kita, 2001). The current paper contributes to this line of argument by showing, at a specific level of analysis, that verbal and non-verbal channels have different types of temporal patterns and that the heterogeneity of these temporal patterns might be important for successful communication. Another important contribution is the introduction to a simple analysis of the temporal structure of behavioral event dynamics, the burstiness analysis. Future work is required to better understand the connection between varying degrees of burstiness across diverse types of human behavioral patterns.

## Acknowledgments

This research was supported by NIH, T32 Grant \# HD07475 awarded to the first author, and NSF-IIS-0416128 awarded to the last author. The usual exculpations apply.

## References

Altmann, E. G., Pierrehumbert, J. B., \& Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. PLoS One, 4(11), e7678.
Altmann, E. G., Cristadoro, G., \& Degli Esposti, M. (2012). On the origin of long-range correlations in texts. Proceedings of the National Academy of Sciences, 109(29), 11582-11587.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. $R$ package version, 1(7).
Butcher, C., \& Goldin-Meadow, S. (2000). Language and gesture. Gesture and the transition from one-to two-word speech: When hand and mouth come together, 235-257.

Butterworth, G., \& Morissette, P. (1996). Onset of pointing and the acquisition of language in infancy. Journal of Reproductive and Infant Psychology, 14(3), 219-231.
Dale, R., Fusaroli, R., Duran, N., \& Richardson, D. C. (2013). The self-organization of human interaction. Psychology of learning and motivation, 59, 43-95.
Fusaroli, R., Rączaszek-Leonardi, J., \& Tylén, K. (2014). Dialog as interpersonal synergy. New Ideas in Psychology, 32, 147-157.
Goh, K. I., \& Barabási, A. L. (2008). Burstiness and memory in complex systems. EPL (Europhysics Letters), 81(4), 48002.
Goldin-Meadow, S. (2005). Hearing gesture: How our hands help us think. Harvard University Press.
Jo, H. H., Karsai, M., Kertész, J., \& Kaski, K. (2012). Circadian pattern and burstiness in mobile phone communication. New Journal of Physics, 14(1), 013055.
Kello, C. T., \& Van Orden, G. C. (2009). Soft-assembly of sensorimotor function. Nonlinear dynamics, psychology, and life sciences, 13(1), 57.
Kugler, P. N., \& Turvey, M. T. (1987). Information, natural law, and the self-assembly of rhythmic movement. Routledge.
Louwerse, M. M., Dale, R., Bard, E. G., \& Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. Cognitive science, 36(8), 1404-1426.
Louwerse, M. M., \& Bangerter, A. (2010). Effects of ambiguous gestures and language on the time course of reference resolution. Cognitive Science, 34(8), 15171529.

Mayberry, R., \& Jaques, J. (2000). Gesture production during stuttered speech: insights into the nature of speechgesture integration. Language and Gesture, 199-215.
McNeill, D. (1992). Hand and mind: What gestures reveal about thought. University of Chicago press.
Morrel-Samuels, P., \& Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(3), 615.
Pierrehumbert, J. B. (2012). Burstiness of verbs and derived nouns. In Shall We Play the Festschrift Game? (pp. 99115). Springer Berlin Heidelberg.

Rime, B., \& Schiaratura, L. (1991). Gesture and Speech. Chicago.
Seyfeddinipur, M., \& Kita, S. (2001, June). Gestures and self-monitoring in speech production. In Annual Meeting of the Berkeley Linguistics Society (Vol. 27, No. 1, pp. 457-464).

# Enactive Mechanistic Explanation of Social Cognition 

Ekaterina Abramova (e.abramova@ftr.ru.nl)<br>Radboud University Nijmegen; Donders Institute for Brain, Cognition, and Behaviour<br>Marc Slors (m.slors@ftr.ru.nl)<br>Radboud University Nijmegen<br>Iris van Rooij (i.vanrooij@ donders.ru.nl)<br>Radboud University Nijmegen; Donders Institute for Brain, Cognition, and Behaviour


#### Abstract

In this paper we examine an enactive approach to social cognition, a species of radical embodied cognition typically proposed as an alternative to traditional cognitive science. According to enactivists, social cognition is best explained by reference to the social unit rather than the individuals that participate in it. We identify a methodological problem in this approach, namely a lack of clarity with respect to the model of explanation it adopts. We review two complaints about a mechanistic explanatory framework, popular in traditional cognitive science, that prevent enactivists from embracing it. We argue that these complaints are unfounded and propose a conceptual model of enactive mechanistic explanation of social cognition.


Keywords: enactivism; social cognition; mechanistic explanation

## Introduction

Embodied Cognition (EC) is most generally a plea to acknowledge that the states of the body and the environment can influence cognition and that lower sensorimotor knowledge plays a role in higher cognition like language and reasoning (e.g. Eerland, Guadalupe, \& Zwaan, 2011). Radical Embodied Cognition (REC) is the claim that the body and the environment are actually part of cognition and as a result, for example, there is no need to have internal representations of the environment (Wilson \& Golonka, 2013; van Dijk, Kerkhofs, van Rooij, \& Haselager, 2008). Enactivism is a strand of REC that stems from the early work in philosophy of biology of Maturana and Varela (1980) and was popularized as an alternative to traditional cognitive science by Varela, Thompson, and Rosch (1991). It shares theoretical commitments with complex systems theory, phenomenology and Buddhist tradition in, on the one hand, grounding cognition on the organizational principles of living systems while at the same time giving a prominent role to the investigation of human experience. Three main principles adopted by enactivism are (1) challenging the dichotomy between internal components of the system and its external conditions, instead stressing the interaction between the two, (2) emphasizing emergent properties on higher levels of organization and (3) viewing the organism as an active autonomous entity that is able to adaptively maintain itself in the environment ${ }^{1}$.

We think enactivism has a lot to offer to the study of cognition because it is an approach that is both naturalistic and

[^3]non-reductionist. However, in this paper we highlight its methodological weakness that might be preventing it from gaining popularity, namely a lack of explicit commitment to how cognitive phenomena are to be explained. We exemplify this issue using a case of enactive accounts of social cognition. We further point out that contemporary cognitive science has two major explanatory frameworks on offer: a deductive-nomological framework, typically associated with REC, and a mechanistic framework, typically adopted by traditional cognitive scientists. We suggest that given the similarity between enactivists and other REC-ers, it is likely that enactivists implicitly subscribe to the deductive-nomological framework. In contrast to this, we argue that the mechanistic framework is not only compatible with enactivism but also preferable. We consider two main objections raised by RECers against mechanistic explanation and show that they rely on a misunderstanding of what such an explanation entails. We end the paper with a preliminary picture of enactive mechanistic explanation of social cognition.

## Enactive Social Cognition

In broadest terms, a non-EC view on social cognition assumes that humans can interact with others successfully only if they are able to see other people as beings with mental states, can infer these states using a so-called 'theory of mind' or simulation and plug in the results of such inferences in planning their own actions. A regular EC view denies the need for such complex representations and inferences emphasizing real-time interaction with other people and perceptual information available in such settings. Certain varieties of simulation accounts of mindreading fit into this framework.

What distinguishes a REC approach is an insistence that the particular dynamics of social interaction themselves play a crucial role in explaining social cognition. This is because "becoming a temporary unit of social action with another person also involves creation of a new perception-action system with new capabilities " (Marsh, Johnston, Richardson, \& Schmidt, 2009, p. 1219). Theoretically this has lead to a claim that there is no need to represent other people or their perspective on the world in order to coordinate with them successfully. Methodologically, it has been suggested that the correct level of analysis in the study of social cognition is the social unit, rather than an exclusive focus on the individuals that comprise it. Instead of searching for internal properties


Figure 1: Perceptual crossing experiment.
A pair of blind-folded participants (call them A and B) are asked to interact in a one-dimensional horizontal field in which they can move using a computer mouse. The field is perceived solely via tactile feedback: encountering a stimulus produces a vibration. Participant A can sense 3 kinds of stimuli: a static object, B's avatar and B's "shadow" that follows B's avatar movement but does not provide $B$ with sensation (the situation is analogous for B). Participants are asked to click when they think to be in contact with the other's avatar. A typical strategy is to move back and forth, especially when a stimulus is encountered. This allows for discriminating between a static and mobile object (if the stimulation changes despite the participant staying in place, the object is not static). The results also show that participants click more often when encountering the other's avatar compared to the other avatar's shadow. However, this increased correct clicking is not due to better recognition (the relative probability of clicking on the avatar is not higher) but rather because the avatars spend more time in front of each other. This effect emerges because the situation of 'sensing the other' while 'being sensed' is more stable than sensing an insensitive shadow. The task is solved globally even if participants are not conscious of this effect and if the solution does not appear in an individual behavioral measure.
of individual independent cognizers, we are to investigate the social interconnectivity that emerges as a result of the interaction and constrains individual-level behavior from the level of a new overarching structure.

One of the most distinctive empirical paradigms that exemplifies this idea is a perceptual crossing (PC) study (Auvray \& Rohde, 2012) presented in Figure 1, in which the task is to distinguish another agent from inanimate objects. A frequent assumption in traditional explanations of social cognition is that such a recognition is accomplished by some special cognitive module (e.g. a module of agency or animacy detection) that is a precondition for interacting successfully. The results of the PC experiment (see Figure 1 caption and the original article for details) have been interpreted to show the reverse: that the social interaction itself and its particular dynamics constitute a solution to such a task. Therefore, in the oftrepeated claim by enactivists, social interaction constitutes social cognition.

Despite a theoretical and empirical research program on social cognition, we believe so far enactivists have not been sufficiently explicit about the explanatory methodology they subscribe to, by which we mean clarity on what is their explanatory target and what constitutes an explanation (Cummins, 2000; Wright \& Bechtel, 2007).

Since social interaction is proposed as an explanation, it cannot be what enactivists are trying to explain. It would mean that the explanandum is perhaps social cognition or experience but these are traditionally understood as individuallevel phenomena ${ }^{2}$. If an explanandum is to be re-construed on a supra-personal level, we need an account of social cognition on that level (without equating it with social interaction). Enactivists could, for example, take more precisely defined types of interaction as their phenomena of interest (cooperation, competition, exchange) and then seek non-individual explanations for their emergence. Our proposal discussed in the last section is to shift to a multi-level explanandum.

Moving on to the model of explanation, an explicit commitment on what constitutes an (good) explanation is important so that any given instance can be judged as to whether it succeeds. In contemporary cognitive science two explanatory frameworks have been discussed most widely: a deductivenomological (DN) and a mechanistic one (Cummins, 2000) ${ }^{3}$.

According to a DN framework, explaining a particular phenomenon proceeds by citing relevant general laws, the details of particular circumstances, and how the phenomenon is to be expected given these two pieces of knowledge. Such an explanation has a form of a deductive argument that derives the explanandum from a certain law taken as a premise (Hempel, 1965). Many REC-ers have explicitly argued for adopting the DN framework for explaining cognition (Walmsley, 2008).

By contrast, in a mechanistic framework, one wants to know not just that a certain regularity holds and what it is but also why it holds and how it is implemented. An answer to this question is sought in identifying a mechanism, where:

A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena (Bechtel \& Abrahamsen, 2005, p. 423).

Enactivists have not explicitly accepted or rejected either of the frameworks. However, given the similarity between them and other proponents of REC, it seems safe to assume that their implicit notion of explanation is likely closer to the DN framework. We now proceed to considering the reasons behind this preference and show that in fact there can be enactive mechanistic explanations.

[^4]
## Enactivist Worries about Mechanistic Explanation

Since space is limited we will focus on just two worries that prevent enactivists from adopting a mechanistic approach. We acknowledge that there are other issues that could be raised (e.g. the role of representations in mechanistic explanation) but our reply to them would be similar in spirit to what we offer here: that the notion of mechanism is richer and more flexible than typical complaints about it presuppose.

## The Decomposability Worry

The main worry enactivists and other REC-ers seem to have with the mechanistic approach is that it allegedly views cognitive systems as decomposable or near-decomposable while in reality they are non-decomposable. For example, Lamb and Chemero (2014) argue that according to the mechanists, producing an explanation requires (1) "decomposition [that] involves developing a model of a system's behavior by identifying discrete component parts and their linear, or weakly non-linear, interactions" and (2) "localization [that] involves mapping those discrete components and interactions onto features of a physical system" (pp. 809-810). What is often added to this charge is that such an explanatory strategy views cognitive systems as component-dominant, i.e. the behavior of the whole is a simple additive result of the behavior of its components, whose properties and functions are rigid and pre-determined (Favela, 2015). Therefore, a single component can be analyzed in isolation as responsible for some particular capacity of the system.

In an opposition to this view on the brain and cognition, REC-ers argue that in fact living cognitive systems are nondecomposable into components and interaction-dominant. That is, the behavior of the whole is more than a simple sum of the parts because interactions between parts are mostly non-linear, the behavior of each part dynamically depends on all other parts of the system and it is not possible to assign any specific task to any component. Therefore, interactions between components are more important than the components themselves (Richardson \& Chemero, 2014) ${ }^{4}$.

If neural and cognitive systems are indeed nondecomposable and mechanistic framework can only be applied to decomposable systems, then obviously enactivists cannot make use of it. However, these arguments betray a misunderstanding of the mechanistic framework and explicit dismissal of the new developments in this field.

First of all, mechanists explicitly argue against mere aggregation of components and place heavy emphasis on their organization (Wimsatt, 1997). It is because the way parts are organized in space and time that they together can exhibit be-

[^5]havior that they cannot exhibit on their own. It is because the parts are on a lower level than the whole they comprise that they cannot have the same properties (cf. the properties of hydrogen and oxygen vs water).

Second, there is no reason to suppose that only linear and sequential modes of organization are allowed in mechanisms. Especially when dealing with biological mechanisms, nonlinear and cyclic modes are ever-present. Such a focus on biology has led mechanists to stress the necessity for dynamic mechanistic explanation because in a system organized nonlinearly "the operations performed by parts of the mechanism vary dynamically, depending on activity elsewhere in the mechanism" (Bechtel, 2011, p. 551). Therefore, an explanation has to include not just a static diagram of components and their organization but also a description of how the functioning of these parts is orchestrated in time, including potential shifts of the overall functional organization. Adding dynamics to a mechanistic explanation does not turn it into a law-based explanation (Bechtel \& Abrahamsen, 2010, 2011).

The general thrust of these extensions of the mechanistic framework is to stress that cognitive systems are likely to lie on a continuum between the extremes of non-decomposable and fully-decomposable. They are, instead, integrated systems, in which it is still possible to identify components but their functions are not necessarily predetermined and fixed. Nor is there a trivial additive relationship between component sub-functions and the overall phenomenon. Rather, their contribution to the operation of the whole might dynamically depend on other parts of the system, the larger context and be variable in time. It does not mean that when studying a mechanism for a particular phenomenon it is impossible to identify these contributions (see also Menary, 2007).

In reply to such arguments, Lamb and Chemero (2014) state that

If a neo-mechanist wishes to discard the condition of decomposability, then she does so at the cost of discarding the feature of neo-mechanistic explanations that makes them distinct from more general accounts of naturalistic explanation (p. 813).

We wish to oppose this complaint. First, it is unreasonable to expect that a certain concept or theory once proposed cannot be developed further. Second, what is distinctive about mechanistic explanations is not decomposability but a concern for causal structure underlying the phenomenon ${ }^{5}$, and for explaining how things work rather than merely stating what are the laws. Finally, specific for the topic of this paper, a REC-er has yet to justify to what extent non-decomposability applies to social cognition, even if it holds for the brain.

[^6]
## The Extended Cognition Worry

The second major worry enactivists have about mechanistic explanation has to do with the claim that social interaction itself constitutes social cognition. This is in line with a general REC view that cognition is not done by the brain alone but by an extended brain-body-environment system. In the case of social cognition, it is rather an extended brain-body-environment-body-brain system (Froese, Iizuka, \& Ikegami, 2013). The fear is that perhaps mechanistic framework somehow precludes such an extended conception of cognition.

The worry is seemingly justified by the following critique by Herschbach (2012). In his article on social cognition sub-titled "A mechanistic alternative to enactivism" (emphasis added), he very acutely points out that enactivists have not been very clear on what they mean by constitution in their claim that "social interaction constitutes social cognition". Constitution is standardly taken to imply a part-whole relationship and if the claim is that supra-personal interaction constitutes individual cognition, then it is somehow a category mistake and a confusion of levels of organization. On the other hand, if constitution is aimed at emphasizing the causal links between agents engaged in the interaction, then enactivists are committing a well-known couplingconstitution fallacy (Adams \& Aizawa, 2010). In this fallacy, frequently ascribed to proponents of extended cognition in general, one points out extensive causal coupling between a cognitive agent and some external factors and then concludes that therefore these factors are part of cognition. Such a conclusion is thought to be unwarranted because coupling and constitutive relations are in general not equivalent.

Herschbach proposes that adopting a mechanistic framework can capture everything that enactivists want to say about social interaction without committing the fallacy. He states that perceptual crossing example would be described by a mechanist as a network composed of interacting agents to be explained by focusing on the agents, their behavior and organization. A mechanist would then move one level down to the internal mechanisms of the agents and how they produce the particular behavior observed in the experiment in response to particular sensory input. The main point of difference between enactivists and mechanists, according to Herschbach, is that while the former would like to say that the environmental input constitutes social cognition, the latter would say that only the agent-internal mechanism constitutes the phenomenon of interest (the behavior exhibited in the experiment) while the environmental input is merely an external influence on that mechanism. That is, the mechanism succeeds only when situated in the appropriate social context of having contact with another agent.

Herschbach grounds this conclusion on the fact that only parts that participate in a self-organized autonomous individual can be truly said to constitute cognition. He follows Bechtel (2009) who has argued that it is the autonomous living system that is the proper "locus of control", differentiated from the environment, because it is the living system
that needs to maintain itself as a unity in constantly changing external conditions.

If adopting a mechanistic framework were to indeed preclude speaking of social interaction playing a constitutive role, enactivists would not be able to accept it. However, we believe there are reasons to oppose Herschbach's conclusion.

The first most obvious reply to Herschbach is that he is replacing the enactive explanandum with his own by switching from the phenomenon of interest being social interaction as a whole to the behavior of the individual. Even though enactivists have not been very clear on their exact explanandum, they would definitely resist this move from the higher to the lower level.

Furthermore, regarding autonomy as a guide to the boundaries of cognition, both Herschbach and Bechtel misunderstand the notion of autonomy adopted by enactivists. Living systems are autonomous in being self-determined rather than being steered from outside (Bechtel's "locus of control"). However, they are also autonomous in being operationally closed, that is, organized in a circular manner, in which the processes and components that constitute the system are themselves constituted by that system. This, however, applies not just to the bio-chemical processes of selfmaintenance, but also to the closure of the sensorimotor loop of the organism. This loop is closed not to the environment but through the environment, which is merely an additional step in the loop, not an input or output external to the system (see Villalobos \& Ward, 2015, for a more detailed argument). The point here is that enactivist autonomy does allow for the constitutive role of the environment in the cognitive process.

Finally, to respond to Herschbach from within a mechanistic framework itself is to point at the recent literature that treats the coupling-constitution fallacy as an instance of a general problem of demarcating the boundaries of a mechanism (Kaplan, 2012). In short, what is required to allow for deciding what constitutes part of the mechanism is an account of constitutive explanatory relevance, i.e. a way to determine which components and processes are relevant to a particular mechanistic explanation (Craver, 2007). This does not need to be a priori based on deciding what cognition really is and whether it really extends beyond the brain. In fact, it is even possible to develop a deflationary (yet still mechanistic) account which shows how certain kinds of dynamic non-linear coupling just are constitutive (Kirchhoff, 2016).

In sum, contra Herschbach (2012), adopting the mechanistic framework does not in fact necessitate abandoning the constitutive role of social interaction in social cognition.

## The Enactive Mechanisms Proposal

We believe enactive mechanistic explanation is possible as there is sufficient basis on both sides of the debate for such a reconciliation. Constructing such an account requires two things. First, it requires disambiguation of the notion of 'composition' involved in mechanistic explanations to recognize its compatibility with enactivist claims about non-


Figure 2: Traditional social cognition.
decomposability of cognitive systems. Identifying components in a cognitive system does not necessarily entail that these components are self-contained, so that the mechanism is a mere linear addition and causal interrelation of them. We can just as well identify components that are defined by their role in the overall whole. This latter, holistic, notion of 'component' is in line with enactivism. Crucially, though, it does leave room for a mechanistic explanation. Secondly, a tailoring of the mechanistic framework is required to fit wider enactivist commitments, such as, for example, making room for cognitive mechanisms that are non-representational and extended. This reorientation towards mechanisms can be advantageous to enactivists for several reasons.

First, it equips enactivism with an explicit and coherent explanatory framework, which comes with specific tools and strategies for constructing explanations of cognitive phenomena. For example, mechanistic literature on mutual manipulability as a guide to constitutive relevance (Craver, 2007) can help make clear what elements of individual cognition and social interaction are essential to particular tasks. Similarly, discussions on how to think of inter-level causation (Craver \& Bechtel, 2007) can help understand the autonomy of the supra-personal level that enactivists consider important.

Second, the claim that social interaction itself should constitute the primary level of explanation in enactive work on sociality to some extent encourages ignoring the individual mechanisms. By contrast, mechanistic emphasis on working parts and their operations highlights the need to provide a distinctively enactive account of what goes on in the individual brains and bodies, i.e. offer a truly multi-level explanation for a multi-level explanandum. Otherwise, a traditional cognitive scientist might well acknowledge the role of interaction but combine that with a non-enactive account of internal mechanisms, thereby defying the whole purpose of constructing an explanatorily complete enactive cognitive science.

Third, enactive mechanistic explanation promotes integration with the rest of cognitive science while at the same time making clear how enactive explanations are different from traditional ones. That is, competing explanations could now be formulated in the same language and compared, instead of two communities adopting completely different explanatory frameworks and talking past each other. This is not to say that dynamical, more law-oriented approaches are to be


Figure 3: Enactive social cognition.
eliminated and we definitely see value in a pluralistic attitude (Dale, Dietrich, \& Chemero, 2009). However, we think at least some intersection is essential for continued progress.

The 'commensurability' of mechanistic enactivism and traditional cognitive science can be illustrated schematically. Figure 2 represents a traditional cognitive science approach to social cognition. All the components of the cognitive mechanisms (differently colored cogs) are located inside the agents' brains. Succeeding in a social task requires one agent to "replicate" the cogs of the other agent inside their own brain, i.e. internally represent the mental states of the other by means of "theory of mind" or simulation. The replicated cogs will not be the same as the original ones (hence the blurriness) but need to be sufficiently close if the agents are to interact successfully. The unfolding of the interaction is then explained in terms of the operations of this internal machinery, giving a strong impression that once all the cogs are in place, the whole process might as well proceed offline.

The contrasting enactive mechanistic view is depicted in Figure 3. Here the explanandum is particular kinds of social interactions in which the individuals participate. The explanation is to be achieved by specifying all the components of the picture that contribute to the realization of such interactions. The components of the cognitive mechanisms (the cogs) are distributed across the brain and the body of both agents and dynamically coupled (the toothed belt), respecting the enactivist rejection of the internal-external dichotomy. The contribution of the individual brains to the overall social interaction is diminished with respect to the previous figure, suggesting a need for an alternative account of such internal mechanisms. The fact that the coupling is a constraint on individual mechanisms rather than an additional cog, expresses the idea that interaction consists of interacting individuals yet allowing for emergent effects. Furthermore, the picture includes the possibility that the coupling might be affected by contextual factors (the tension pulley), such as the layout of the environment in which interaction unfolds, or some sociocultural circumstances.

To restate the point of our paper in terms of the second figure above, we believe the current state of the matters in enactivist theorizing about social cognition is an exclusive focus on the toothed belt. We think the time is ripe to start examining the rest of the picture.

## Acknowledgments

The first author is funded by the Netherlands Organization for Scientific Research (NWO), research project PGW-12-32.

## References

Adams, F., \& Aizawa, K. (2010). Defending the bounds of cognition. The Extended Mind, 67-80.
Auvray, M., \& Rohde, M. (2012). Perceptual crossing: The simplest online paradigm. Frontiers in Human Neuroscience, 6, 181.
Bechtel, W. (2009). Explanation: Mechanism, modularity, and situated cognition. In P. Robbins \& M. Aydede (Eds.), The cambridge handbook of situated cognition (pp. 155170). Cambridge: Cambridge University Press.

Bechtel, W. (2011). Mechanism and biological explanation. Philosophy of science, 78(4), 533-557.
Bechtel, W., \& Abrahamsen, A. (2005). Explanation: A mechanist alternative. Studies in History and Philosophy of Biological and Biomedical Sciences, 36(2), 421-441.
Bechtel, W., \& Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. Studies in History and Philosophy of Science Part A, 41(3), 321-333.
Bechtel, W., \& Abrahamsen, A. (2011). Complex biological mechanisms: Cyclic, oscillatory, and autonomous. In C. A. Hooker (Ed.), Philosophy of complex systems (pp. 257-285). New York: Elsevier.
Craver, C. F. (2007). Constitutive explanatory relevance. Journal of Philosophical Research, 32, 3-20.
Craver, C. F., \& Bechtel, W. (2007, 1 September). Top-down causation without top-down causes. Biology \& philosophy, 22(4), 547-563.
Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil \& R. Wilson (Eds.), Explanation and cognition (pp. 117-145). Cambridge, MA: MIT Press.
Dale, R., Dietrich, E., \& Chemero, A. (2009). Explanatory pluralism in cognitive science. Cognitive Science, 33(5), 739-742.
Eerland, A., Guadalupe, T. M., \& Zwaan, R. A. (2011). Leaning to the left makes the eiffel tower seem smaller: Posturemodulated estimation. Psychological Science, 22(12), 1511-1514.
Favela, L. H., Jr. (2015). Understanding cognition via complexity science. Unpublished doctoral dissertation, University of Cincinnati.
Froese, T., Iizuka, H., \& Ikegami, T. (2013). From synthetic modeling of social interaction to dynamic theories of brain-body-environment-body-brain systems. The Behavioral and Brain Sciences, 36(04), 420-421.
Froese, T., Iizuka, H., \& Ikegami, T. (2014). Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. Scientific Reports, 4(3672).

Hempel, C. (1965). Aspects of scientific explanation and other essays in the philosophy of science. New York: Free Press.
Herschbach, M. (2012). On the role of social interaction in social cognition: A mechanistic alternative to enactivism. Phenomenology and the Cognitive Sciences, 11(4), 467486.

Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. Biology \& Philosophy, 27(4), 545-570.
Kirchhoff, M. D. (2016). From mutual manipulation to cognitive extension: Challenges and implications. Phenomenology and the Cognitive Sciences, 1-16.
Lamb, M., \& Chemero, A. (2014). Structure and application of dynamical models in cognitive science. In Proceedings of the 36th annual meeting of the cognitive science society (pp. 809-814).
Marsh, K. L., Johnston, L., Richardson, M. J., \& Schmidt, R. C. (2009). Toward a radically embodied, embedded social psychology. European Journal of Social Psychology, 39(7), 1217-1225.
Maturana, H., \& Varela, F. (1980). Autopoiesis and cognition: The realization of the living. Boston: D Reidel Publishing.
McGee, K. (2005). Enactive cognitive science, part 1: Background and research themes. Constructivist Foundations, l(1), 19-34.
Menary, R. (2007). Cognitive integration: Mind and cognition unbounded. Palgrave Macmillan.
Richardson, M. J., \& Chemero, A. (2014). Complex dynamical systems and embodiment. In L. Shapiro (Ed.), The routledge handbook of embodied cognition (pp. 39-50). New York: Routledge.
van Dijk, J., Kerkhofs, R., van Rooij, I., \& Haselager, P. (2008). Can there be such a thing as embodied embedded cognitive neuroscience? Theory \& Psychology, 18(3), 297-316.
Varela, F., Thompson, E., \& Rosch, E. (1991). The embodied mind: Cognitive science and human experience. Cambridge: MIT Press.
Villalobos, M., \& Ward, D. (2015). Living systems: Autonomy, autopoiesis and enaction. Philosophy \& Technology, 28(2), 225-239.
Walmsley, J. (2008). Explanation in dynamical cognitive science. Minds and Machines, 18(3), 331-348.
Wilson, A. D., \& Golonka, S. (2013). Embodied cognition is not what you think it is. Frontiers in Psychology, 4, 58.
Wimsatt, W. C. (1997). Aggregativity: Reductive heuristics for finding emergence. Philosophy of Science, 64, 372384.

Wright, C. D., \& Bechtel, W. (2007). Mechanisms and psychological explanation. Philosophy of Psychology and Cognitive Science, 4, 31-79.

# Human Visual Search as a Deep Reinforcement Learning Solution to a POMDP 

Aditya Acharya (axa384@student.bham.ac.uk)<br>School of Computer Science<br>University of Birmingham

# Xiuli Chen (x.chen.1@bham.ac.uk) 

School of Psychology
University of Birmingham

Christopher W. Myers (Christopher.Myers2@wpafb.af.mil)<br>Air Force Research Laboratory<br>Performance and Learning Models Branch, USA

Richard L. Lewis (rickl@umich.edu)
Department of Psychology
University of Michigan,Ann Arbor, USA

Andrew Howes (HowesA @bham.ac.uk)<br>School of Computer Science<br>University of Birmingham


#### Abstract

When people search for a target in a novel image they often make use of eye movements to bring the relatively high acuity fovea to bear on areas of interest. The strategies that control these eye movements for visual search have been of substantial scientific interest. In the current article we report a new computational model that shows how strategies for visual search are an emergent consequence of perceptual/motor constraints and approximately optimal strategies. The model solves a Partially Observable Markov Decision Process (POMDP) using deep Q-learning to acquire strategies that optimise the tradeoff between speed and accuracy. Results are reported for the Distractor-ratio task.


Keywords: Computational Rationality; Deep Reinforcement Learning; Deep Q-Learning; Visual Attention.

## Introduction

One of the many tasks for which people use vision is to search for items in the environment. Visual search might be used to locate a phone on a table, a car in a parking lot or a family member in a crowd. In a typical laboratory visual search task, participants are asked to find a visual target amongst distractors. For example, searching for a Gabor patch in a high contrast noisy background (Najemnik \& Geisler, 2008), or searching for a red coloured letter O in a display that consists of red Xs and green Os (Shen, Reingold, \& Pomplun, 2000). Many, though not all, visual search tasks require a number of fixations and saccades before the target is found.

From a cognitive science perspective, visual search is interesting because data from visual search experiments can be used to inform theories of the underlying constraints on vision (e.g (Geisler, 2011) and also to inform theories of how people adapt eye movement strategies to these constraints (e.g (Najemnik \& Geisler, 2005). Human behaviour is a consequence of both the constraints and the adapted strategies and explanations of behaviour require both (Lewis, Howes,
\& Singh, 2014). In fact, there is a long history of cognitive science research on visual search and there are a number of competing theoretical approaches.

First are the map-based approaches described by (Kowler, 2011), such as salience maps (Itti \& Koch, 2000) and activation maps (Pomplun, Reingold, \& Shen, 2003; Wolfe, 2007), where the perceived visual information is represented as a topological distribution in a graphical map form. The salient area or peaks in the map represent items that significantly differ from their neighbouring items, that may contain attributes of interest. These peaks in the map are then used to guide the eyes through the display using some selection rules, such as a greedy heuristic (Pomplun et al., 2003) or a winner-take-all heuristic (Itti \& Koch, 2000). To summarize, the map based approach assumes that saccades are programmed to move the fovea to those areas in the display that stand out from surroundings.

Second are the Bayes optimal state estimation approaches (Myers, Lewis, \& Howes, 2013; Najemnik \& Geisler, 2008), in which it is assumed that visual information is recorded as a Bayesian estimate of the state of the world. On each fixation the estimated state is updated by optimally integrating information (Bayes rule) from the previous state and from the fovea and from the periphery according to its reliability. The eye movements are then made using these states and applying a heuristic decision rule (e.g., 'Maximum A Posteriori' (MAP)) to navigate. This rule generates a behaviour in which attention is directed to areas which have the highest probability of target present. Alternatively, Najemnik and Geisler (2005) observed that the number, and spatial distribution, of saccades could be better explained by a model in which each saccade was directed to an 'ideal' location (i.e., a location that maximises information gained). Their model was sensi-
tive to known human constraints on vision, i.e., the accuracy of perceiving a feature degrades with eccentricity.

Third are the optimal control approaches (Butko \& Movellan, 2008; Hayhoe \& Ballard, 2014; Nunez-Varela \& Wyatt, 2013; Sprague, Ballard, \& Robinson, 2007), in which it is assumed that the eye movements are not made to estimate some statistics about the world but rather the goal is to maximize the overall performance utility. The maximum reward/utility an individual can attain throughout the task is bounded by the noisy encoding of the visual information by the human brain. In contrast to map-based and optimal state estimation approaches, where prior assumptions about eye movement decisions are made by heuristic rules, the control strategy emerges as a consequence of bounds imposed by the human visual system. To summarize, the optimal control approach assumes that the saccades are programmed to move the fovea so as to maximise task utility/reward.

In the current article we report a novel (approximately) optimal control model of the distractor ratio task. The purpose of this model is to (1) explain phenomena not previously explained as optimal control, (2) to further elucidate the framing of visual search as a Partially Observable Markov Decision Process (POMDP) (Kaelbling, Littman, \& Cassandra, 1998), and (3) to explore the role of deep Q-learning (Mnih et al., 2015) in solving the tractability problems with previous optimal state estimation and optimal control approaches. The model goes beyond the optimal state estimation model of Myers in that it is applied to the full display size used by (Shen, Reingold, \& Pomplun, 2003). The model uses deep Q-learning to solve a POMDP. It attempts to maximise a reward signal given constraints imposed by the human visual information processing system. We compare the performance of the optimal control model to a model that uses MAP-like heuristics. We show that the optimal control model offers higher utility and better fits to the human data than the heuristic model. Lastly, we use the model to explain phenomena associated with the distractor ratio paradigm (Bacon \& Egeth, 1997; Shen et al., 2000; Zohary \& Hochstein, 1989). A phenomena that has previously been explained using the salience-map based approach.

## The Distractor Ratio Task

In the distractor ratio task the display consists of a target object, which is randomly positioned amongst distractor objects each of which shares at least one common feature with the target. The goal is to respond whether the target is present or absent. An example display is shown in Figure 1 where the target is a red letter O . The distractors in this display share either a same-colour or same-shape feature with the target.

In a number of studies it has been observed that people respond more quickly, and with fewer eye movements, for extreme ratios of same colour to same shape distractors (Egeth, Virzi, \& Garbart, 1984; Shen et al., 2003). In Figure 1, the target - a red letter O - can be located easily in display (a) and (c) with ratios 3:45 and 46:2 respectively as compared to

(a)

(b)

(c)

Figure 1: Distractor ratio stimuli with ratio distributions: (a) 3:45, (b) 24:24, (c) 46:2 and target stimuli: red coloured letter O.


Figure 2: (a) Average number of fixations per trial as a function of the number of distractors sharing colour with the search target in target-absent trials and target-present trials for high discriminability condition. (b) Saccadic bias (the difference between the observed frequency and chance performance) as a function of the number of same-colour distractors in target- absent trials for high discriminability condition (Shen et al., 2003)
display (b), for which a response takes a relatively long time. The distractor ratio effect reported by Shen et al. (2003) is shown in Figure 2.

In addition to the distractor-ratio effect, Shen et al. (2003) also observed a saccadic selectivity effect. In Figure 2, the frequency of saccades to same-colour distractors is plotted against the number of same-colour distractors. In the plot, the saccade frequencies are higher for rare features (colour or shape) than should be expected by chance (represented by the horizontal line). When the same-colour distractors are rare in the display, the participants were more likely to make eye movements towards them than when they were common. Conversely, when the number of same-colour distractors was high, the participants were more likely to make eye movement towards same-shape distractors.

## The Model

In the following sections we describe the individual components of the model for performing a 36-element distractorratio task, and provide a walk-through of the model process before presenting the model results.

## External Display

In the model, we represent the display by randomly distributing the target and the distractors in a grid, where each cell consists of either a target object, a distractor object with common colour or a distractor object with common shape. In the display, there is only one target object and the number of distractors are determined by randomly sampling a ratio per trail.

The display is represented by two feature vectors, one for colour and one for shape. The presence or absence of a feature in each cell in the model is represented numerically by the number 1 for presence and 0 for absence. The random distribution of these features in the environment was achieved by sampling randomly from the following set of ratios, $r=R$ (3:33, 6:30, 9:27, 12:24, 15:21, 18:18, 21:15, 24:12, 27:9, 30:6, 33:3).

## Actions

The action space consists of (1) fixate on a cell, (2) respond present and (3) respond absent. In our study there was a grid of $6 \times 6$ coloured shapes and there were therefore a total of 36 possible fixation actions. A trial was terminated by the choice of the present or absent action.

## Reward

A reward was given after choosing a present or absent action. The reward distribution was defined as a value 10 for a correct response, a value of -10 for an incorrect response and a value of -1 for each fixation. The penalty on each fixation imposes a speed-accuracy trade-off. More fixations gives greater accuracy but at a cost.

## Observation Model

Every time the model fixates, it also makes an observation. The observation obtained by the model is constrained by the noise in the human visual system. Two types of noise are added to the signal: spatial smearing noise and feature noise.

1. Feature Noise: The human eye's ability to discriminate and perceive object features degrades with eccentricity according to a hyperbolic function (Strasburger, Rentschler, \& Jüttner, 2011). To model this function we added Gaussian white noise with mean 0 and standard deviation as eccentricity, i.e., a function of visual angle ' $\theta$ ' between the fovea and the given location, and a scalar weight ' $w_{\text {featural }}$ ' to scale the effect of distance to the fovea for feature noise. Therefore, the equation for the observation after adding feature noise at location $j$ given that the eye is focused on location k is as follows,

$$
\begin{gathered}
\delta_{\text {featural }}\left(S_{t}, j\right)=v\left[s_{t}\right]+N\left(\theta, \sigma_{f}\left(\theta_{j k}, w_{\text {featural }}\right)\right) \\
\sigma_{\text {featural }}\left(\theta_{j k}, w_{\text {featural }}\right)=\frac{\theta_{j k}}{\left(w_{\text {featural }}\right)}+c
\end{gathered}
$$

where, $v\left[s_{t}\right]=1$ if the location $s_{t}$ contains a target feature, else $v\left[s_{t}\right]=0, c$ is a constant with value $10^{-4}$ to avoid 0
variance in the model, $\sigma_{f}\left(\theta, w_{f}\right)$ is the variance to simulate the degrading eccentricity and ' $\theta$ ' is the distance between the fixated cell and location $j$.
2. Spatial Smearing: Another source of uncertainty in the human visual system is the localization error (Levi, 2008), where information in the parafovea may erroneously combine features from one location with adjacent locations. Therefore, for each location in the colour and shape vector a weighted sum is calculated for the location and its adjacent eight locations. For example, If a red X is surrounded by green Os in the parafovea then, as a consequence of spatial smearing, the participant would be uncertain whether they are actually looking at a red X or a green O .
In the model, spatial smearing is represented by a weighting function (Gaussian kernel) with standard deviation as a function of visual angle ' $\theta$ ' between the fovea and the given location, and a scalar weight ' $w_{\text {spatial' }}$ ' to scale the effect of distance to the fovea for spatial noise. The weighting function here is a normalised function. As ' $\theta$ ' (distance) increases the acuity decreases and the standard deviation of the Gaussian kernel increases, this means that the percept of the item at a given location suffers greater interference from surrounding items. This encoding is done for each location in the display. Thus, the equation for the observation after adding spatial noise at location $j$ given that the target features are at location $S_{t} \in(1,2, \ldots, n)$ and the eye is focused on location $k$ is as follows,

$$
\begin{gathered}
\delta_{\text {percept }}\left(S_{t}, j\right)=K\left(s, \sigma_{s}\left(\theta_{j k}, w_{\text {spatial }}\right)\right) \times \delta_{\text {featural }}\left(S_{t}, j\right) \\
\sigma_{\text {spatial }}\left(\theta_{j k}, w_{\text {spatial }}\right)=\frac{\theta_{j k}}{\left(w_{\text {spatial }}\right)}+c
\end{gathered}
$$

where, K is the Gaussian kernel with kernel size $s=1$, $\sigma_{s}\left(\theta_{j k}, w_{s}\right)$ is the variance. $\delta_{\text {percept }}\left(S_{t}, j\right)$ is calculated separately for both shape and colour feature vectors. $c$ is a constant with value $10^{-4}$ to avoid 0 variance in the model. Now each percept ( $\delta_{\text {percept }}$ ) (one for colour and one for shape) is represented as a vector of noisy observations for each location. A consequence of introducing the noise is uncertainty in the content of the location.

## State Estimation

At each time step $t$ on which a fixation is made the model receives a noisy observation for each location. The values for perceived colour and shape are then combined (Hadamard product) for each location $[i, j]$. We refer to these combined values as relevance scores, where a higher score in a location signifies high relevance to the task. These scores are then integrated across fixations, using naive Bayesian inference (Kalman filter), to get the current state $B_{t}$ which is a vector of estimated relevance scores across fixations ${ }^{1}$.

[^7]
## Heuristic Control Model

The Heuristic control model makes fixations and observations as described above. In order to decide which fixations to use and when to respond it makes use of two heuristics. The first uses a MAP-like strategy to determine where to fixate next, and the second uses a thresholded stopping rule.

## Optimal Control Model

As we have said, at each point in time, the model observes the external environment through a noisy percept with a high resolution fovea and low resolution parafovea and receives an observation $o_{t}$. The model then extract the high resolution local information from the environment by taking actions $a_{t} \in A$ ( $A$ is the set of actions) to move the fovea (e.g., choose where to move the fovea). Since the environment is only partially observed the model needs to integrate information over time in order to determine how to act and how to make eye movements most effectively. It does this using the Bayesian state estimator described above.

At each step, the model receives a scalar reward $r_{t}$ (which depends on the action taken by the agent), and the goal of the agent is to maximize the total sum of such rewards $R=$ $\mathbb{E}\left[\Sigma \gamma^{t-1} r^{t}\right]$, where $\gamma \in(0,1)$ is the discount factor.

The most important aspect of the Optimal Control model is that rather than using heuristics to choose what to do next, it learns an approximately optimal policy using Deep Qlearning.
Deep Q-learning The Deep Q-learner made use of the following network architecture.

The relevance score estimate $B_{t}$ ( 36 element vector) from the state estimator (above) was taken as the input. This input was connected to a fully connected hidden layer consisting of nodes equivalent to number of elements in the display, i.e., 36 , with rectifier activation function. This is followed by a second fully connected hidden layer consisting of again nodes equivalent to number of elements in the display, i.e., 36 , with sigmoid activation function. Finally, the output layer was a fully connected linear layer of 38 nodes with single output for each action in the task. To avoid over-fitting of the network $l 2$ regularization of the weights was applied with value $10^{-5}$.

During the training process a fixed size batch of transitions $<s, a, r, s^{\prime}>$ were sampled from a replay memory and used for learning. For each time step $(t)$, the deep Q-network (with parameters $\theta$ ) is trained to approximate the action-value $(\mathrm{Q}-$ value) function from the sampled transitions by minimizing the loss functions $L\left(\theta_{i}\right)$ :

$$
L\left(\theta_{i}\right)=\mathbb{E}_{s, a \sim \pi_{\theta}}\left[\left(y_{i}-Q(s, a ; \theta)\right)^{2}\right]
$$

where $y_{i}=r+\gamma \max _{a^{\prime}} Q\left(s^{\prime}, a^{\prime} ; \theta^{\prime}\right)$ is the target Q -value estimated from a target Q-network $\left(\theta^{\prime}\right)$. The parameters of target Q-network $\left(\theta^{\prime}\right)$ is copied over from the learned network ( $\theta$ ) after a fixed number of iterations.

```
Algorithm 1 Deep Q Network Algorithm
    initialize replay memory \(D\), weights of the main network
    \(\theta\) and target network \(\theta^{\prime}\).
    observe the initial state s.
    repeat
        select an action a
            with probability epsilon select \(a\) random action.
            otherwise select \(a=\operatorname{argmax}_{a^{\prime}} Q\left(s, a^{\prime} ; \theta\right)\).
        perform the action \(a\).
        observe the reward \(r\) and new state \(s^{\prime}\) for action \(a\).
        store transition \(<s, a, r, s>\) in the replay memory \(D\).
        sample random transitions \(\langle s, a, r, s\rangle\) from the re-
    play memory D.
        calculate the target value \(t\) for each sampled transi-
    tion.
        if \(s^{\prime}\) is the terminal state then
            \(t=r\)
        else
            \(t=r+\gamma Q\left(s^{\prime}, \max _{a^{\prime}} Q\left(s^{\prime}, a^{\prime} ; \theta\right) ; \theta^{\prime}\right)\)
        end if
        update the network using \((t-Q(s, a ; \theta))^{2}\) as the loss.
        \(s=s^{\prime}\)
        after every fixed steps \(\theta^{\prime}=\theta\)
    until terminal state
```


## Model Results

The Heuristic control model was run for 30,000 trials and 10 regression runs to check for consistency. The Optimal control model was run for 20 million trials. We first tested the accuracy of the models. Accuracy is the proportion of trials on which the model correctly responded either present or absent. The best fitting optimal control model achieved an accuracy of $96 \%$ in its last 50000 trials. In comparison, human participants achieved $98 \%$ accuracy. The accuracy of the best fitting Heuristic control model was $94 \%$. Accuracy and utility of both models is plotted in Figure 3. The plots show a clear advantage of the Optimal control model for all explored parameter settings. In other words, the approximately Optimal control model outperforms the Heuristic control model in all cases.

Plots of fixation frequency versus same colour distractorratio at different levels of spatial and feature noise are shown in Figure 5. The results show that both model Heuristic and Optimal control model generate similar distractor ratio curves to humans (Figure 2) for target absent, where more fixations are required for ratios close to 1 . While the RMSEs for the Heuristic control model were smaller than for the Optimal control model (Optimal: $R M S E=0.81$; Heuristic $R M S E=0.41$ ), the goodness of fit against Human performance for the Heuristic control model was $R^{2}=0.95$ and for the Optimal control model was $R^{2}=0.98$. A weakness of the Heuristic control model was that it produced DR effects for both target present and target absent. In contrast, the Optimal control model predicted a DR effect in the absent condition


Figure 3: (a) Mean accuracy achieved by both models plotted against different noise parameter settings. (b) Mean utility gained by both models plotted against different noise parameter settings. Where, FN is Feature noise, SN is Spatial Noise and TH is the threshold set for heuristic control model.


Figure 4: Saccadic bias as a function of the number of same colour distractors for Target Absent.
only. In terms of the shape of the DR curve and saccadic selectivity curve, the similarity between humans and Optimal control model is greater than the similarity between Heuristic control model and humans (see Figure 2).

The saccadic bias effect is shown in Figure 4. For the explored parameter settings, the Heuristic control model generated higher levels of saccadic bias than generated by the Optimal control model and these levels were nearer to those generated by humans (Optimal: $R M S E=8.93$; Heuristic $R M S E=6.93$ ). However, the Optimal control model explained more of the variance. The goodness of fit of the best fitting Heuristic control model was $R^{2}=.94$. In contrast, the best fitting Optimal control model had a goodness of fit of $R^{2}=0.97$. While the Heuristic control model predicts a magnitude of saccadic bias that corresponds to that of humans at extreme levels of same-color (around 30\%), it is the Optimal control model that has the better fit. This is likely due to the extreme curvature (sinusoidal) of the saccadic bias for the Heuristic model which is not present in the humans.

One of the effects in the human data that is not captured by either the Optimal or the Heuristic control model is the asymmetric effect of shape and colour (see Figure 2). This is very


Figure 5: Number of fixations as a function of same-colour distractors for (a) the Heuristic model with target present, (b) the Control model with target present, (c) the Heuristic model with target absent, (d) the Heuristic model with target present.
likely due to the fact that we used the same noise parameter values for both shape and colour in the model's observation function. Further work is needed to explore the effect of the known differences in acuity functions for shape and colour (Kieras \& Hornof, 2014).

## Discussion and Conclusion

While the results presented here are preliminary, they offer some evidence that the distractor-ratio effect is the consequence of an approximately optimal adaptation to the constraints imposed by the human visual information processing system. Unlike previous work, including Myers et al. (2013), our results are based on a model that makes approximately optimal control decisions to choose fixation locations rather than a model that uses MAP-like heuristics.

Achieving these results required two contributions to cognitive modeling. The first is the novel application of POMDPs to the framing of the distractor-ratio problem, further extending the work of Butko and Movellan (2008). The POMDP framing is important because it provides a rigorous basis for exploring the computationally rational adaptation of human strategies to known information processing constraints (Lewis et al., 2014; Howes, Lewis, \& Vera, 2009). It thereby helps make the crucial link between cognitive mechanism and rationality that supports deep explanations of behaviour.

The second contribution is the novel application of Deep

Q-Learning (Mnih et al., 2015) to determine the optimal policy given a theory of human visual information processing capacities. The role of reinforcement learning based algorithm's have previously been proposed as means of explaining human learning processes (Dayan \& Daw, 2008) and also, as means of deriving rational analyses of what a person should do in particular task (Chater, 2009). Our work is more aligned with the goals of (Chater, 2009). The purpose of our reinforcement learner was not to model the step-by-step learning process, but rather to model the rational outcome of the learning process - an approximately optimal adaptation to information processing limits.

There is a substantial amount of work to be done. While the best fitting Optimal control model explained $98 \%$ of the variance, to be fully confident that it is better than the Heuristic control model, we need to more fully explore the parameter space of both models. For example, for the Heuristic control model, it might be the case that even higher feature noise, and lower spatial noise, might further improve the fit. We also need to find a fit that reduces the RMSE of the Optimal control model.

In conclusion, we have demonstrated that framing the visual search problem as a POMDP and solving this problem with deep Q-learning is a viable approach to explaining effects such as distractor-ratio and saccadic selectivity.

## References

Bacon, W. F., \& Egeth, H. E. (1997). Goal-directed guidance of attention: evidence from conjunctive visual search. Journal of Experimental Psychology: Human Perception and Performance, 23(4), 948.
Butko, N. J., \& Movellan, J. R. (2008). I-pomdp: An infomax model of eye movement. In Development and learning, 2008. icdl 2008. 7th ieee international conference on (pp. 139-144).
Chater, N. (2009). Rational and mechanistic perspectives on reinforcement learning. Cognition, 113(3), 350-364.
Dayan, P., \& Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. Cognitive, Affective, \& Behavioral Neuroscience, 8(4), 429-453.
Egeth, H. E., Virzi, R. A., \& Garbart, H. (1984). Searching for conjunctively defined targets. Journal of Experimental Psychology: Human Perception and Performance, 10(1), 32.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. Vision research, 51(7), 771-781.
Hayhoe, M., \& Ballard, D. (2014). Modeling task control of eye movements. Current Biology, 24(13), R622-R628.
Howes, A., Lewis, R. L., \& Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. Psychological review, 116(4), 717.
Itti, L., \& Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research, 40(10), 1489-1506.

Kaelbling, L. P., Littman, M. L., \& Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. Artificial intelligence, 101(1), 99-134.
Kieras, D. E., \& Hornof, A. J. (2014). Towards accurate and practical predictive models of active-vision-based visual search. In Proceedings of the 32nd annual acm conference on human factors in computing systems (pp. 38753884).

Kowler, E. (2011). Eye movements: The past 25 years. Vision research, 51(13), 1457-1483.
Levi, D. M. (2008). Crowdingan essential bottleneck for object recognition: A mini-review. Vision research, 48(5), 635-654.
Lewis, R. L., Howes, A., \& Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. Topics in cognitive science, 6(2), 279-311.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Humanlevel control through deep reinforcement learning. Nature, 518(7540), 529-533.
Myers, C. W., Lewis, R. L., \& Howes, A. (2013). Bounded optimal state estimation and control in visual search: Explaining distractor ratio effects. In Proc. cogsci.
Najemnik, J., \& Geisler, W. S. (2005). Optimal eye movement strategies in visual search. Nature, 434(7031), 387391.

Najemnik, J., \& Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. Journal of Vision, 8(3), 4-4.
Nunez-Varela, J., \& Wyatt, J. L. (2013). Models of gaze control for manipulation tasks. ACM Transactions on Applied Perception (TAP), 10(4), 20.
Pomplun, M., Reingold, E. M., \& Shen, J. (2003). Area activation: A computational model of saccadic selectivity in visual search. Cognitive Science, 27(2), 299-312.
Shen, J., Reingold, E. M., \& Pomplun, M. (2000). Distractor ratio influences patterns of eye movements during visual search. Perception, 29(2), 241-250.
Shen, J., Reingold, E. M., \& Pomplun, M. (2003). Guidance of eye movements during conjunctive visual search: the distractor-ratio effect. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 57(2), 76.
Sprague, N., Ballard, D., \& Robinson, A. (2007). Modeling embodied visual behaviors. ACM Transactions on Applied Perception (TAP), 4(2), 11.
Strasburger, H., Rentschler, I., \& Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. Journal of vision, 11(5), 13-13.
Wolfe, J. M. (2007). Guided search 4.0. Integrated models of cognitive systems, 99-119.
Zohary, E., \& Hochstein, S. (1989). How serial is serial processing in vision? Perception, 18(2), 191-200.

# Executive function and attention predict low-income preschoolers' active category learning 

Katherine A. Adams ${ }^{1}$ (kat.adams@ nyu.edu), George Kachergis ${ }^{2,3}$<br>${ }^{1}$ Department of Applied Psychology, New York University<br>${ }^{2}$ Department of Artificial Intelligence, Radboud University<br>${ }^{3}$ Donders Institute for Brain, Cognition and Behavior


#### Abstract

Recent studies find that school-age children learn better when they have active control during study. Yet little is known about how individual differences in strategy or cognitive control skills may affect active learning for preschoolers, nor if experimental measures of active learning map onto real-world learning outcomes. The current study assesses 101 low-income 5 -year-olds on an active category learning task, and measures of executive function, attention, and school readiness. We find that preschoolers use an informative sampling strategy for categories defined by stimuli features in 1D and when presented with a distractor dimension (2D). Children accurately classify in 1D, but show mixed performance in 2D. Attention predicts sampling accuracy, and working memory and inhibitory control predict classification accuracy. Performance in the active learning task predicts early math and pre-literacy skills. These findings suggest that trial-by-trial learning decisions may reveal insight into how cognitive control skills support the acquisition of knowledge.


Keywords: active learning; executive function; attention; cognitive development; education

## Introduction

From the enthusiastic preschooler who asks "why?" to the infant who turns her head to attend to a novel toy, children learn by actively exploring the world around them. Experimental studies show that children are engaged problem-solvers and employ strategies such as hypothesis testing, for example playing more with a toy after being shown confounded information about how it works (Schulz \& Bonawitz, 2007).

Research in cognitive science suggests that active information gathering boosts children's performance in learning experiments (Partridge, McGovern, Yung, \& Kidd, 2015; Sim, Tanner, Alpert, \& Xu, 2015). For example, Sim and colleagues (2015) recently found that 7-year-old children learn categories better after self-selecting examples of category membership than when passively presented with a random sequence of examples. Yet little is known about how variation in children's abilities to optimally sample information may affect learning outcomes. Do young children differ in their information sampling strategies? What skills help children be good active learners? How do experimental measures of active learning map onto real-world learning outcomes? These questions have important implications within cognitive science and may inform targeted education interventions, particularly for children from under-resourced backgrounds who are at increased risk for poor academic outcomes (Blair \& Raver, 2014). This study takes a first step in addressing these questions by examining low-income preschool children's active sampling strategies in a category learning task. We then
ask how individual differences in a series of executive function, attention, and school readiness measures relate to active learning performance.

Educational research has long been interested in how young children's abilities to actively attend and engage during learning affect academic outcomes. One set of factors identified are executive functions (EF), higher-order cognitive control skills such as the ability to hold items in working memory, inhibit a prepotent response, and flexibly shift attention. Higher EF is associated with higher socioemotional and cognitive skills, and predicts early math and pre-literacy skills (Blair \& Raver, 2014). Similarly, individual differences in preschool children's sustained and selective attention are important predictors of cognitive and academic skill (Steele, Karmiloff-Smith, Cornish, \& Scerif, 2012). The neural networks that underlie EF and attention undergo tremendous growth during the preschool years. Several successful preschool interventions capitalize on this neurocognitive plasticity by targeting EF and attention as a means to boost school readiness and close income-based academic achievement gaps (Ursache, Blair, \& Raver, 2012). Importantly, these intervention programs promote children's active engagement in learning as a key mechanism to support both EF and academic skills. However, this research is limited both by conceptualizing active learning in global behavior terms and by operationalizing learning outcomes with static standardized assessments.

Active learning paradigms from the cognitive science and machine learning literature offer a higher resolution to examine how young children actively learn and employ cognitive control processes. Here, active learning is defined as allowing learners to generate or make decisions about the information they want to experience trial by trial (Gureckis \& Markant, 2012). Trial-by-trial analyses of active information gathering can reveal meaningful variation in learning strategies. For example, Gureckis and Markant (2009) investigated adult learners' ability to gauge information value during an active search task similar to the children's game battleship. The authors found that participants' information generating behaviors took two forms: one relatively fast and undirected and another slower, more effortful, that exploited local information constraints. Moreover, response time and search efficiency differed across these "modes."

Benefits of active control during information gathering are that learners can ask targeted questions to avoid redundant examples or content too difficult, creating learning situations to
best fill their personal knowledge gaps (Gureckis \& Markant, 2012). Active control also supports learning by enhancing the encoding of episodic representations which increases the likelihood of retrieving information about the experienced stimuli from memory (Markant, Ruggeri, Gureckis, \& Xu, 2016). Adult learners can benefit from even subtle control over timing by coordinating the presentation of new information with their optimal attentional state so they are alert and ready to encode (Markant, DuBrow, Davachi, \& Gureckis, 2014). Moreover, active control during memorization is associated with increased coordination in the neural networks that support executive control, attention, and memory encoding (Markant et al., 2016).

A limitation to active learning is that benefits can vary based on learners' abilities and task demands. For example, learners may be biased when sampling data, creating an uninformative feedback loop. Markant (2016) manipulated the hypothesis generation process in a series of category learning tasks to assess the impact on adults' ability to learn simple and complex rules. Results showed successful active learning depended on a match between the target rule and salient perceptual and abstract features of the task stimuli, and poor learning was due to generation of hypotheses that followed a non-relevant rule. That is, adult participants benefited from active learning opportunities under complex task demands when they were able to shift their attention to relevant rules or dimensions while ignoring others-a central EF skill.

To our knowledge, no studies to date have examined the coordination of EF and attention in children during active learning. Moreover, very little is known about how individual differences in active learning relate to sampling strategies or school readiness. Identifying the mechanisms supporting successful active learning may inform both cognitive science theory and educational interventions to support school readiness. The current study addresses these research gaps by examining active learning in a large sample of low-income preschoolers using a multi-dimensional category learning task, as well as a well-validated battery of EF, attention, and school readiness measures.

## Method

## Participants

One hundred and one preschoolers $(M=61 \mathrm{~m}$; Range $=55-$ 67 m ; Male $=46$ ) were tested as part of a school readiness study run in collaboration with two Head Start preschool centers. Participants came from low-income backgrounds, with an average reported yearly income of $\$ 11,968$ (Range $=733-$ 34,486). The sample was predominantly African American ( $\mathrm{N}=86$ ).

Children were tested in their preschools by trained assessors using a touchscreen laptop. Administration of the tasks was divided over two testing days within a one week period. EF tasks were administered on day 1 , and the category learning task and school readiness assessment were administered on day 2.


Figure 1: Examples of 1D trials varying by size dimension.

## Category Task

Materials The category task (a modified version from Sim et al., 2015) was presented as a multi-slide questionnaire using the Qualtrics survey system. First, Block 1D presented two forests of trees that varied by one stimuli feature. The color row showed 10 images, identical in size and shape, but with varied leaf color, ranging from orange to green (left to right). The size row of 10 images were identical in color and shape, but progressed in size from smallest to largest (left to right). Block 2D presented a 7-by-6 grid of trees that varied horizontally by size (smallest on the left) and vertically by color (orange on top). In this task, worms and snails live in different groups of trees. The goal of the task is to classify trees based on the type of animal that lives there. Small worm and snail icons were displayed above exemplar trees in the sampling phase, and appeared as two larger button choices at the bottom of the screen in the test phase.
Procedure The 1D and 2D blocks each began with a demonstration phase, followed by two testing sequences which switched the dimension of categorization. In the 1D block, the first sequence featured the color row of stimuli and the second used the size row. In the 2D block, the first sequence was categorized by the size dimension and the second by color. Each sequence began with a 2 -trial sampling phase followed by a classification phase (4 1D trials, 82 D trials).
Block 1D Demonstration phase. To introduce the task, children were told to pretend they were scientists and figure out where two types of animals, worms and snails, liked to live. First, children were shown the the 1D color row. To demonstrate an example categorization of the trees, children were shown a small worm or snail icon above each and every tree (see Fig. 1). A red circle appeared around the group of trees with worms and another around the group with snails to em-
phasize that the animals were grouped separately. Children were asked to point to the category boundary, described as the "edge between the trees where the worms and the snails live." Once the child guessed, they were shown the boundary with a red arrow. This sequence was repeated with the 1D size row of trees and a new category boundary. Following the Sim et al. (2015) task design, these practice trials were meant to establish that (1) the animal icon above the tree indicated that the animal lived in that tree, (2) there was an invisible category boundary that divided the trees into two groups, and (3) the category boundary moved with each new forest.

Sampling phase. Children were first presented with the 1D color row of trees. In sampling trial 1, a worm and a snail icon appeared over 2 exemplar trees. The letters A, B, and C appeared under possible trees to sample. One sampling option was informative to find the category boundary, while the other two were non-informative because their category membership could be inferred by the position of the exemplars. By limiting learners to three sampling options, we increased our power to differentiate informative vs. uninformative sampling strategies over fewer trials, reducing noise and task demands for this very young sample.

To complete the sampling selection, the child was prompted, "Here's where a worm lives and here's where a snail lives. If you want to find the edge between the trees, would you want to learn about what lives in tree $\mathrm{A}, \mathrm{B}$, or C?" Once the child touched the sampling tree option of their choice, the selection was automatically logged in the Qualtrics database. Sampling trial 2 revealed the correct worm and snail icons above the three sampling tree options of the previous trial, and three new trees were shown as sampling options (see Fig. 1). Children selected a sampling option and the task advanced to the classification phase.

Classification phase. Children were presented with the 1D color row of trees without exemplars. At the bottom of the screen, a larger image of a worm and snail were shown vertically aligned. To reduce task demands, only 4 of the 10 trees were queried for classification. On each classification trial, the test trial number (1-4) appeared underneath one of the trees as a cue to guess the category membership of that tree. Children indicated their response by touching either the worm or snail response icon. Together, these responses demonstrated where each child believed the category boundary was generally located.

The sequence of sampling and classification phases was repeated for the 1 D size row of trees, with a new category boundary and locations for exemplars and test trials.
Block 2D To introduce the 2D block, children were shown the $6 \times 7$ grid of trees and instructed, "In big forests, you have to find out if the worms and snails live in groups based on the SIZE of the trees or the COLOR of the trees. They only care about the size OR color!" The demonstration phase was identical to that in the 1D block (see Fig. 2). Sampling and classification phases in 2D followed the same procedure as 1D. Children were not told whether color or size was the rel-


Figure 2: Examples of 2D trials with horizontal category boundary (classification trial not shown).
evant dimension for categorization.
The first 2D sequence had a category boundary determined by tree size, following a vertical axis. Sampling trial 1 featured 3 category exemplars and 4 sampling tree options. The location of the exemplars made categorization only possible by the vertical dimension (i.e. by size). One sampling option was informative to the vertical category boundary. The category membership of the three other non-informative options could be inferred by the locations of the exemplars. After 2 sampling trials, children completed 8 classification test items. The number of exemplars, sampling options, and classification trials were increased compared to the 1D block to include a variety of positions across the 2 D grid.

The second 2D sequence had a category boundary determined by tree color, following a horizontal axis (see Fig. 2). This dimensional switch (i.e., requiring attention to horizontal relations between exemplars to infer category boundary, not vertical as in past trials) is a feature of dimensional card sort games, classic EF tasks which require the participant to flexibility shift attention to the new relevant dimension and inhibit response to the old dimension.
Coding Selection trials were coded as correct if the child selected the option informative to finding the category boundary. Aggregate scores were computed for overall task accuracy and overall sampling and classification accuracy, and for sampling and classification accuracy on 1D vs. 2D blocks.

## EF, Attention, and School Readiness Tasks

Working Memory. Digit Span is a widely used executive function task that assesses children's working memory (WM). Children are instructed to repeat number sequences of sequentially longer length in forward and backward conditions. Children in this sample were largely unable to repeat sequences backwards, so only correct responses on the forward condition are reported here.

Attention and Inhibitory Control. In the Continuous Performance Test (CPT), one hundred pictures are randomly presented on a touch screen one at a time for 300 ms followed by blank response screen for 1500 ms . Children are instructed to touch the screen as soon as an animal appears. Stimuli in-
clude 20 presentations of the target stimuli (animals) and 80 presentations of nontarget stimuli (objects). We report reaction time on correct touches to targets, a measure of attention processing speed (APS). We reverse-coded percent of missed responses to targets (omission error) and incorrect touches to distractors (commission error) as indices of sustained attention (SA) and inhibitory control (IC), respectively.

Math and Pre-Literacy Skills. The Woodcock Johnson III Tests of Achievement (WJ-III) is a well-validated assessment of school readiness skills. The Applied Problems subtest assesses children's early mathematical reasoning. The Letter Word subtest requires children to identify letters and words to measure their pre-literacy skills. A sum of the total correct answers is computed for each subtest and then translated into a standardized W-Score.

## Results

## Sampling Performance

We first ask if preschoolers can strategically sample in a category learning task. In the 1D block, children are significantly above chance in accurately choosing the informative sampling option $(\mathrm{M}=.48$, chance $=.33 ; t(99)=4.06, p<0.001)$. Within the 1D block, mean sampling accuracy is not different for color $(\mathrm{M}=.45)$ and size $(\mathrm{M}=.51), t(99)=-1.37, \mathrm{p}=$ .18. Children also chose the informative sampling option in the 2 D block $(M=.35$, chance $=.25 ; t(98)=3.5, p=0.001)$. Within the 2D block, mean sampling accuracy is also not different for color $(\mathrm{M}=.33)$ and size $(\mathrm{M}=.37), t(99)=-1.07$, $\mathrm{p}=.29$. We find that sampling accuracy in 1 D is related to sampling accuracy in 2D, $r=.26, p=.01$.

Figure 3 shows participants' mean accuracy on sampling questions (left panel) and subsequent categorization questions (right panel) by stimulus dimension and dimension of the sampling space.

## Classification Performance

Overall, children are above chance in correctly classifying test items in 1D $(\mathrm{M}=.66$, chance $=.5 ; t(99)=6.3, \mathrm{p}<$ $0.001)$ but not in 2D $(\mathrm{M}=.51$, chance $=.5 ; t(98)=.97$, p $=.337$ ). Mean classification accuracy in the 1D block is significantly higher than mean classification accuracy in the 2D block $(t(97)=5.14, \mathrm{p}<.001)$. Within the 1D block, mean classification accuracy is significantly higher for color ( $\mathrm{M}=$ $.71)$ than for size $(\mathrm{M}=.63), t(99)=2.23, \mathrm{p}=.03$. Within the 2 D block, children are at chance on the size condition $(\mathrm{M}=.47$, chance $=.5 ; t(98)=-1.35, \mathrm{p}=.18)$ but interestingly above chance on the subsequent color condition, which includes a dimension switch on the category boundary ( $\mathrm{M}=$ .55 , chance $=.5 ; \mathrm{t}(98)=2.15, \mathrm{p}=.034)$. Mean classification accuracy is significantly higher for color in the 2 D block than for size $(t(98)=2.24, \mathrm{p}=.03)$. Note that effects of size vs. color dimensions should be interpreted with caution, as the blocks were presented in fixed order.

## Does Sampling Predict Classification?

We next ask if sampling accuracy benefits subsequent classification accuracy, as suggested in previous active category


Figure 3: Accuracy on sampling questions (left) and categorization test (right) for each relevant stimulus dimension (color/size) and dimensionality of the stimulus space. Dotted lines show chance (sampling chance: $1 \mathrm{D}=33 \%, 2 \mathrm{D}=25 \%$ ).
learning studies with both adults and school-age children. Surprisingly, sampling accuracy is not related to classification accuracy. Children who choose the most informative sampling strategy in 1D are not better at 1D classification, $r=.06, p=.58$, nor are 2D sampling and classification accuracy related, $r=-.03, p=.81$. Comparing classification accuracy on 1D of good samplers ( $M_{a c c}>.7$ ) vs. poor samplers $\left(M_{a c c}<.3\right)$ yielded no significant differences, $t(72.9)=.84$, $p=.41$, nor for good vs. poor samplers in $2 \mathrm{D}, t(35.7)=.2$, $p=.84$.

## Relations Between EF and Active Learning

We next examined the role of executive function and attention in predicting active learning performance using a series of exploratory logistic mixed-effects regression models to the item-level with subject as a random factor. Age, sex, EF, and attention measures were fixed predictors. Prior to analyses, we scaled and centered all variables. Table 1 presents descriptives of executive function, attention, and school readiness measures.

First, we predicted overall accuracy, including both sampling and classification trials $(\mathrm{N}=2,632 ; \mathrm{R}$ syntax: $\quad$ Correct $\sim$ age + sex $+W M+A P S+S A+$ IC $+(1 \mid$ Subject $)$ ). There was a significant positive effect for $\mathrm{WM}(\beta=.11, Z=2.11, p=.04)$, showing that participants with higher working memory perform better overall in the task. Next, we predicted accuracy on all sampling trials ( $\mathrm{N}=676$ ), adding overall classification accuracy as an additional fixed predictor (R syntax: Correct $\sim$ age + sex + WM + APS + SA + IC + class Acc + (1|Subject)). There was a significant positive effect

Table 1: Descriptives of Executive Function, Attention, and School Readiness Measures.

|  | Mean (SD) | Range |
| :--- | :---: | :--- |
| Digit Span (\% correct forward) | $44 \%(13)$ | $0-67 \%$ |
| CPT Omission Errors (\%) | $47 \%(28)$ | $0-100 \%$ |
| CPT Commission Errors (\%) | $12 \%(14)$ | $0-78 \%$ |
| CPT Reaction Time (ms) | $824(166)$ | $474-1364$ |
| WJ-III Applied Probs (W-Score) | $407(17)$ | $350-440$ |
| WJ-III Letter Word (W-Score) | $333(20)$ | $276-369$ |

for attention processing speed (APS) $(\beta=.38, Z=2.57$, $p=.01$ ), showing that participants with faster attention processing are more accurate at sampling. We then predicted accuracy on all classification trials $(\mathrm{N}=1956)$, substituting in overall sampling accuracy as an additional fixed predictor. There was a significant positive effect for WM ( $\beta=.12$, $Z=2.14, p=.03$ ), showing that participants with higher working memory are more accurate at classification.

We next ran the models by 1D and 2D blocks. Predicting 1D sampling ( $\mathrm{N}=339$ ) with 1D classification accuracy as the additional fixed predictor revealed no significant effects. For 1D classification trials ( $\mathrm{N}=598$ ) with 1D sampling accuracy as the additional fixed predictor, there was a surprising significant negative effect for sustained attention (SA) ( $\beta=-.37$, $Z=-1.98, p=.048$ ), such that children who were less responsive to targets during a sustained attention task had better classification accuracy in the 1D block.

Predicting to 2 D sampling ( $\mathrm{N}=337$ ) with 2 D classification as the additional fixed predictor also revealed a positive effect of attention processing speed ( $\beta=.357, Z=2.026$, $p=.043$ ). For 2D classification trials with 2D sampling accuracy as the additional fixed predictor, there was a significant effect of inhibitory control (IC) ( $\beta=.131, Z=2.087$, $p=.037$ ), showing that children who are better at inhibiting a prepotent response are more accurate at 2D classification.

## Predictors of School Readiness

How does active learning performance relate to school readiness? To examine this question, we use an exploratory linear mixed-effects model fit by REML (nlme package) at the subject level to predict to math and pre-literacy scores on the WJ-III assessment. First, we predict to math scores using subject as a random effect, and age, sex, EF, attention, and overall sampling and classification accuracy as fixed predictors ( R syntax: math $\sim$ age + sex $+\mathrm{WM}+\mathrm{APS}+\mathrm{SA}+\mathrm{IC}$ + class Acc + Sampling Acc, random= $\sim 1 \mid$ subject). We found significant positive effects for overall sampling accuracy $(t(77)=3.85, p<.001)$, and overall classification accuracy $(t(77)=2.64, p=.01)$, suggesting that children who are better at active learning in the category task are have better early math skills over and above the contributions of EF, attention, and demographics.

We ran the same exploratory linear mixed-effects model
to predict pre-literacy scores but did not find any relations between pre-literacy and sampling or classification accuracy. We modified the predictors, collapsing over sampling and classification trials to examine the effect of overall active learning performance. Here, we find that overall accuracy in the active learning task is a positive predictor of pre-literacy skills $(t(78)=2.03, p=.046)$.

## Discussion

We found that 5-year-olds from low-income backgrounds use an informative sampling strategy in an active category learning task. Preschoolers are able to accurately classify the category membership of test items in 1D, but show mixed performance in the 2D classification blocks. Sampling accuracy across dimensions hangs together: children who choose the most informative option in 1D are also better at sampling in 2D. However, children who are good at classification in 1D are not more likely to be good at classification in 2D. Contrary to past active learning studies, we do not find that better sampling accuracy benefits classification accuracy in either 1D or 2D blocks. However, individual differences in children's EF and attention skills shine a light on potential cognitive control processes that support success in active learning. We found that attention processing speed largely supports sampling accuracy and better working memory is linked to higher accuracy on classification. Notably, better inhibitory control supports classification accuracy when the categories are presented with a distractor dimension (2D).

Previous published work has found active learning benefits in categorization tasks for 7-year-olds (Sim et al., 2015), but little research has examined preschool-aged children. One concern was that younger children would struggle with the metacognitive organization needed to plan and follow an informative sampling strategy. Despite variability, our data show that many preschool children made queries that were informative to finding a category boundary.

It may be that categorization within a 2D space was especially difficult to navigate for children at this age, however we found that children had above chance classification accuracy within the 2D block after the dimensional rule switch from a vertical to horizontal category boundary. Although surprising, one possibility is that children became accustom to categorization within a 2D space over multiple trials, and thus increased accuracy was partially due to practice. A limitation to this study is that the task order was fixed, and thus size vs. color performance is confounded with practice effects. It is interesting to note that practice effects may be present in the 2D condition despite the added complexity of a categorization switch. Additional study is needed to understand how variation in boundary options and stimuli characteristics may affect children's active category learning.

Exploratory analyses examining the role of executive function and attention skills add nuance to understanding children's performance in this task. We found that attention processing speed supports children's sampling strategies. The link between attention processing and active learning benefits
is discussed in Markant et al. (2014), who found that learners benefit from even minimal control of timing during learning by matching the presentation of information to their optimal attentional state. It is possible that those children who were able to maintain a ready state of attention processing were better able to encode information about sampling parameters than those whose attention processing was slower.

We found that higher working memory supports both overall task accuracy and classification across 1D and 2D blocks. Working memory may help children to remember the relative locations of exemplars within the 1D and 2D spaces, which is necessary to infer both the category boundary and test items' class. Inhibitory control significantly predicts classification accuracy in the 2D block. Dimensional shifting requires the learner to inhibit a learned response or rule and attend to new information. To attend to the category boundary and correctly classify exemplars in a 2D space, this task requires children to determine how exemplars related to each other along the relevant dimension, while ignoring relations on the non-relevant dimension. These exploratory findings suggest that inhibitory control may help children attend to relevant features under complex learning demands.

The lack of relation between category learning sampling and classification is at odds with previous studies which found better sampling led to better performance in active learning (e.g., Ruggeri et al., 2016). In fact, our design was meant to decouple sampling performance from categorization performance: the category memberships of all sampling choices are revealed once the selection is made. Thus a child who makes a bad selection is not penalized-they see the exemplar of the informative choice as well as the uninformative choices. Because we were limited to a fixed task across all administrations, this design allowed us to make sure all children's sampling choices were revealed and that they all saw the same information about the boundary leading into the classification phase. Thus, this design removed the necessity of good sampling to support classification accuracy within the task. We found that children's memory skills, likely related to remembering the location of exemplars, appeared to play a more important role in classification accuracy.

Our results also suggest that children's performance in the active learning task is related to their early math and preliteracy skills. Both overall sampling and overall classification accuracy significantly predict math scores, above and beyond demographics, executive function, and attention skills. Overall active learning accuracy predicts pre-literacy skills. These exploratory findings suggest that trial-by-trial performance in a lab-based measure of active learning may be related to children's acquisition or implementation of academic knowledge. Children's development and use of learning processes and problem solving strategies may rely in part on cognitive control skills. The benefits of good active learning skills may cascade overtime to support children's acquisition and practice of domain-specific knowledge. Importantly, these correlational data are only the first step in investigat-
ing active learning in relation to school readiness and additional research should examine this potential link, as it could be highly informative to educational intervention efforts.

While most education and developmental researchers examine EF and learning by way of standardized school readiness tests, the current study's findings suggest that details of children's trial-by-trial learning decisions may reveal important details of how cognitive control skills support the acquisition of knowledge. We note several limitations to this study. First, the sample is low-income and the restricted range of socio-economic status (SES) may lower generalizability, although examining the relations between learning processes and school readiness is of particular importance for this population. We plan follow-up studies including a high income cohort to examine the relations between SES, cognitive control skills, and active learning. Second, this work is both correlational and uses concurrent measures. Our future studies include experimental active learning paradigms for young children that vary aspects of cognitive control processes to better tease apart the role of EF and attention on active learning.

## Acknowledgments

We thank Todd Gureckis and C. Cybele Raver for their support and guidance on this study.

## References

Blair, C., \& Raver, C. C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. PloS One, 9(11), el12393.
Gureckis, T. M., \& Markant, D. B. (2009). Active Learning Strategies in a Spatial Concept Learning Game. In Proc. of CogSci 31.

Gureckis, T. M., \& Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. Perspectives on Psychological Science, 7(5), 464-481. doi: 10.1177/1745691612454304

Markant, D. B. (2016). The impact of biased hypothesis generation on self-directed learning. In A. Papafragou, D. Grodner, D. Mirman, \& J. Trueswell (Eds.), Proc. of CogSci 38. Austin, TX: Cognitive Science Society.
Markant, D. B., DuBrow, S., Davachi, L., \& Gureckis, T. M. (2014). Deconstructing the effect of self-directed study on episodic memory. Memory and Cognition, 42, 1211-1224.
Markant, D. B., Ruggeri, A., Gureckis, T. M., \& Xu, F. (2016). Enhanced memory as a common effect of active learning. Mind, Brain, and Education, 10(3), 142-152.
Partridge, E., McGovern, M. G., Yung, A., \& Kidd, C. (2015). Young children's self-directed information gathering on touchscreens. In R. Dale et al. (Eds.), Proc. of CogSci 37. Austin, TX: Cog. Sci. Society.
Schulz, L., \& Bonawitz, E. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. Developmental Psychology, 43(4), 1045-1050.
Sim, Z. L., Tanner, M., Alpert, N. Y., \& Xu, F. (2015). Children learn better when they select their own data. In R. Dale. et al. (Eds.), Proc. of CogSci 37. Austin, TX: Cognitive Science Society.
Steele, A., Karmiloff-Smith, A., Cornish, K., \& Scerif, G. (2012). The multiple subfunctions of attention: Differential developmental gateways to literacy and numeracy. Child Development, 83(6), 2028-2041.
Ursache, A., Blair, C., \& Raver, C. C. (2012). The promotion of self-regulation as a means of enhancing school readiness and early achievement in children at risk for school failure. Child Development Perspectives, 6(2), 122-128.

# Simulating behavioural interventions for developmental deficits: When improving strengths produces better outcomes than remediating weaknesses 

Hala Alireza (halalireza@yahoo.co.uk)<br>Developmental Neurocognition Lab, Birkbeck, University of London<br>Malet Street, London WC1E 7HX, UK

Anna Fedor (fedoranna@gmail.com)<br>2MTA-ELTE Theoretical Biology and Evolutionary Ecology Research Group<br>Budapest, H-1117, Hungary

Michael S. C. Thomas (m.thomas@bbk.ac.uk)<br>Developmental Neurocognition Lab, Birkbeck, University of London<br>Malet Street, London WC1E 7HX, UK


#### Abstract

Computational models of cognitive development have been frequently used to model impairments found in developmental disorders but relatively rarely to simulate behavioural interventions to remediate these impairments. One area of controversy in practices of intervention is whether it is better to attempt to remediate an area of weakness or to build on the child's strengths. We present an artificial neural network model of productive vocabulary development simulating children with word-finding difficulties. We contrast an intervention to remediate weakness (additional practice on naming) with interventions to improve strengths (improving phonological and semantic knowledge). Remediating weakness served to propel the system more quickly along the same atypical trajectory, while improving strengths produced long-term increases in final vocabulary size. A combination yielded the best outcome. The model represents the first mechanistic demonstration of how interventions targeting strengths may serve to improve behavioural outcomes in developmental disorders. The observed effects in the model are in line with those observed empirically for children with word-finding difficulties.


Keywords: artificial neural networks; developmental disorders; intervention; vocabulary development; wordfinding difficulties

## Introduction

## Theories of deficits versus theories of intervention

In the field of developmental disorders, there are extensive theories about the causes of behavioural deficits. However, these have played a relatively small role in intervention practices. Indeed, theories of treatment have often developed relatively independently of theories of deficit (Laws et al., 2008; Michie \& Prestwich, 2010). The gap between a mechanistic understanding of the causes of deficits and everyday therapeutic practice exists for a number of reasons. Most obviously, the primary focus of intervention is on behavioural outcomes, which do not in themselves necessitate an understanding of underlying cause. In addition, there are diverse real-world constraints influencing the interventions that are selected. And it is
difficult to apply causal principles to the complex therapeutic situation involving treatment of the whole child via a social interaction with the therapist. Nevertheless, it remains an important ambition to narrow the gap between theories of deficit and practices of intervention.

## Improve strengths or remediate weaknesses?

One area of controversy in practices of intervention is whether it is better to attempt to remediate an area of weakness or to build on the child's strengths. For example, in the field of developmental language disorders, Leonard (2014) argues that generally, therapists prefer to work on developing compensatory strategies through targeting the child's strengths rather than trying to improve his or her area of weakness (see also Bishop, Nation \& Patterson, 2014).

To take a more specific case, where children have difficulties in producing words that they already know (so called word-finding difficulties; WFD), therapists may simply require the child to spend more time practicing naming, the area of weakness. Alternatively, they have the option of targeting children's knowledge of word sounds (phonology) or word meanings (semantics). Therapists have found that interventions that elaborate the semantic aspect of words (e.g., McGregor \& Leonard, 1989) or interventions that focus on the phonological component of word finding (e.g., Best, 2005) both alleviate WFD to some extent. In a survey, Best (2003) asked therapists what kind of difficulties they found most often co-occurring with WFD in the children they saw. Phonological problems were reported to co-occur $46 \%$ of the time, while semantic problems cooccurred only $13 \%$ of the time. Nevertheless, when asked which types of intervention they most often used, therapists reported more often using techniques to improve semantic knowledge than phonology ( $79 \%$ of the time compared to $54 \%$ ). In this case, then, therapists frequently sought to buttress areas of strength to improve naming skills.

One explanation for the tendency of therapists to work less on areas of weakness and more often on areas of strength is to improve the child's confidence in a domain
where he or she is struggling. However, from a theoretical perspective, one might ask through what mechanisms could improving a strength serve to remediate a behavioural impairment in a developing cognitive system?

There are at least three ways that improving a strength could remediate a behavioural impairment. First, the 'strength' could represent an alternative cognitive system or pathway to deliver a similar behavioural outcome. Improving a strength then translates to encouraging a compensatory strategy. Second, the target behavior may be delivered by an interactive system in which multiple sources of knowledge combine to drive behavior. Stronger input from one source might then make up for weaker input from another. Third, the target behavior may require mappings to be learned between representations. Improving the structure of the representations might serve to make learning those mappings easier. In this article, we use computational modeling to investigate the third of these options.

## Computational modeling of interventions

Computational modeling, particularly the use of artificial neural networks, has been extensively applied to understanding the mechanisms underlying developmental deficits, in disorders such as dyslexia, Specific Language Impairment, and autism (Thomas \& Karmiloff-Smith, 2002). Relatively few models of developmental deficits have been extended to the simulation of behavioural interventions to remediate these deficits, and the framework for doing so has only recently been laid out (Thomas et al., 2017). Two notable exceptions are Harm, McCandliss and Seidenberg's (2003) simulation of an intervention for dyslexia, and Best et al.'s (2015) model of interventions for productive vocabulary deficits. In both cases, a typical model established the developmental trajectory under normal circumstances; an atypical model was created in which a computational constraint limited development; and a behavioural intervention was simulated by adding further input-output mappings to the model's training set for a discrete period, usually relatively early during training. Here, we adapt Best et al.'s model of vocabulary development to contrast the effects of improving strengths versus remediating weaknesses.

## Connectionist models of vocabulary development

Sentence production involves a sequence beginning with a planned message, followed by selection of major lexical concepts, assigning syntactic functions, assembling phonologically realized words and morphemes into a sentence frame, and programming articulatory processes (e.g., Bock \& Levelt, 1994). Connectionist models of word production have tended to focus on the step involving the retrieval of phonological forms given a semantic specification of the desired lexical item (e.g., Dell, Schwartz, Martin, Saffran, \& Gagnon, 1997). Developmental models have simulated the learning of mappings between pre-specified semantic and phonological codes (e.g., Plunkett et al., 1992), or between semantic and
phonological representations emerging in self-organizing maps (e.g., Li, Zhao \& MacWhinney, 2007; Mayor \& Plunkett, 2010). Best et al. (2015) used a similar approach, but implemented the semantic and phonological components of the model via 3-layer autoassociator networks trained using backpropagation (Rumelhart, Hinton \& Williams, 1986). The hidden unit representations of the semantic component were then mapped to the hidden units of the phonological component, via an intermediate layer of hidden units, to provide a pathway for the development of naming. A reverse pathway simulated comprehension. Naming behavior began to emerge while the semantic and phonological representations were themselves still developing. By restricting learning in the semantic component, the phonological component, or the pathway between them (for example, by reducing hidden unit numbers or the learning rate), Best et al. were able to capture various patterns of atypical naming development observed in a sample of children with WFD.

## Simulations

## Simulation design

Using the same architecture, we followed the Best et al. (2015) model in simulating productive vocabulary development in children with WFD by restricting the computational capacity of the pathway mapping between semantic and phonological representations. The semantic and phonological components themselves developed typically. Early in development - when slow vocabulary growth was already detectable - a behavioural intervention was applied for a limited period during training. Five different interventions were contrasted, of three types: (1) remediating the weakness - the model was provided with additional training on the naming pathway; (2) improve the strength - the model was provided with additional training to improve the semantic representations, the phonological representations, or both at once; (3) both types 1 and 2 were combined into an intervention that sought to simultaneously improve strength and remediate weakness. We observed the immediate effect of intervention, in terms of potentially accelerated vocabulary growth, and the eventual outcome, in terms of the largest vocabulary size achieved following each type of intervention.

The original Best et al. (2015) model used a fairly abstract rendition of semantics and phonology and a training set of only 100 items. Here, we used more realistic semantic and phonological representations, and scaled up the training set slightly to around 400 items. The typically developing model was designed in such a way that it reflected salient properties of vocabulary growth, including a comprehension-production asymmetry (Bloom, 1973) and a vocabulary explosion / exponential growth in vocabulary size (e.g., McMurray, 2014).

The model contained three assumptions not in the Best et al. (2015) model. First, phonological representations were required to be more accurate than semantic outputs to drive
a behavioural response, under the assumption that phonological output needs to drive motor assemblies, while semantic comprehension only requires that the output fall in the correct attractor basin (Hinton \& Shallice, 1991). This assumption generated the production-comprehension asymmetry.

Second, we implemented a sensitive period in the development of the components but not the pathways in the model, through pruning of network connectivity after a given point in development. This created the potential for early training to create enriched lower level representations by utilizing the then-available rich connectivity. Pathways did not experience this pruning, under the view that sensitive periods are characteristic of lower but not higher cognitive systems (Takesian \& Hensch, 2013). The effect of timing of intervention was subsequently evaluated.

Third, plasticity was set higher in the pathways than the components (via the learning rate parameter), so that the development of semantic and phonological representations would be the limiting factor on the development of naming. If the semantic and phonological representations were to quickly reach ceiling before naming had developed, interventions targeting phonology and semantics would have no scope to improve naming performance. The effects of both the second and third assumptions were evaluated by also running the model in their absence.

Finally, we explored whether the five types of intervention would enhance performance in a typically developing model, or whether they only had the potential to improve performance in systems exhibiting delayed development.

## Simulation details

Architecture: The architecture of the vocabulary development model is shown in Figure 1. It comprised four linked backpropagation networks. The semantics component comprised a 3-layer autoassociator with 1029 input and output units and 45 hidden units. The phonology component was an autoassociator with 456 input and output units and 60 hidden units. The naming pathway linked the semantic hidden units with the phonological hidden units via an intermediate layer of 175 units. Naming constituted activating semantic inputs and measuring phonological outputs. The comprehension pathway ran in the other direction and also contained 175 units. In the atypical model, the number of hidden units in the naming pathway was reduced to 90 prior to training.

Additional parameters: The learning rate in the semantic component was .015 and in the phonological component was .025 . In the pathways, the learning rate was .15 . Sigmoid activation functions had a temperature of 1.5 in the components and 1 in the pathways. In the components, after epoch 75, any connection weights with an absolute magnitude of less than .5 had a $5 \%$ chance of being permanently removed each epoch. Initial weights were given random values via a Gaussian distribution with mean 0 and standard deviation 0.5 . Gaussian noise with a standard
deviation of .15 was added to the net input of units in the components, and noise with a standard deviation of .05 in the pathways, to provide a stochastic basis for naming errors in normal functioning. Continuous activation values on the phonological output were converted to responses by finding the nearest legal phoneme in each slot and assessing whether the full phoneme string was the correct name. If the average root mean square error between the activation vector for each phoneme and the nearest legal phoneme code exceeded 0.03 , that phoneme was coded as no response. A nearest neighbor technique was also used to assess the accuracy of semantic outputs. These parameters were selected to calibrate the typical model.


Figure 1: Architecture of the productive vocabulary model, with phonological and semantic components and linking pathways

Training set: The training set comprised 397 words, each with a phonological and a semantic representation. It was generated by combining two sources, a set of 1029 speaker generated semantic feature norms for 456 words collected by Vinson and Vigliocco (2008) from 280 adults; and the Children's Printed Word Database (Masterson et al., 2010), which is an online database of the vocabulary in reading materials used by 5-9 year old children in the UK. The 397 words represent those present in both resources. The semantic representations comprised the 1029 feature set, where a feature was set to 1 if any adult rated it as a characteristic of a given word meaning, and 0 otherwise. The phonological representations used a 19-bit articulatory code for phonemes (Thomas \& Karmiloff-Smith, 2003) and a left-justified slot-based CCCVVCCC syllabic scheme to capture words up to three syllables in length, with $3 \times 8 \times 19=$ 456 phonological features in total.

Training schedule: Networks were trained for 1000 epochs, with random presentation and pattern update. Training of autoassociators and pathways was interleaved. Weights were updated using the backpropagation algorithm with the cross entropy error measure.

Simulation of interventions: For atypical networks, an intervention began at 100 epochs of training and lasted for 100 epochs. For the main condition, at this point TD models had acquired a productive vocabulary size of 67 words, while atypical models had a vocabulary size of 36 words.

For the intervention, one or more components or pathways were trained with 5 times the frequency of the rest of the system. The extra training could be on the semanticsphonology naming pathway, the semantics component alone, the phonological component alone, both semantics and phonological components, or all of these combined.

Conditions: To test the importance of timing, interventions were compared at 100,250 and 750 epochs. To test the effect of plasticity assumptions in the model, the first variant removed connectivity pruning from the components. The second variant removed the higher plasticity of the pathways, setting their value to .025 .

Replications: All conditions were replicated three times with different random seeds. The full design took approximately 100 days of simulation time. Results graphs are shown averaged over replications; individual data are included in the following tables.

## Results

Figure 2 displays developmental trajectories for naming and comprehension in the typical and atypical models. For the typical model, naming lagged behind comprehension, exhibiting the expected comprehension-production asymmetry. Naming itself showed an accelerating rate of development, consistent with a vocabulary explosion. The atypical model with WFD exhibited delayed development in naming but not comprehension (the slightly better comprehension was just a chance difference).


Figure 2: Development of naming and comprehension in the typical and atypical models. The dotting line depicts the point at which the intervention was applied

Figure 3 shows the effects of the intervention to remediate weakness, with extra training on the naming pathway, compared to the impaired model without treatment. The intervention produced accelerated development, but there was no gain in final productive vocabulary level. Figure 4 shows the result of improving strengths - extra training on the otherwise typically developing semantic and phonological representations, which are respectively the inputs and outputs of the naming pathway. These interventions produced slower effects during the intervention itself, but led to long-term (if relatively modest) increases in final productive vocabulary levels. Figure 3 also
contains the combined strength-and-weakness intervention. The combined intervention showed the initial immediate gains of the remediation intervention as well as the longterm elevated final level of the strengths intervention. Figure 5 includes the effects of these interventions on typically developing models. Extra practice in naming accelerated development but did not raise the final level. Extra elaboration of semantic and phonological representations by contrast increased the final productive vocabulary size even for the typically developing networks.

Tables 1 and 2 show the final level performance, split by replication, contrasting the intervention targeting weakness (Naming), the intervention targeting strengths ( $\mathrm{S}+\mathrm{P}$ ), and the intervention targeting both strengths and weaknesses (Both). Table 1(a) contains the data for the above base condition; 1 (b) demonstrates that when plasticity reductions in the components were removed as a model assumption, the same pattern of results held. Table 1(c) shows that without the assumption of greater plasticity in the pathways, the same pattern also held. Table 2 contrasts the effect of interventions at different points in training. The strengths intervention ( $\mathrm{S}+\mathrm{P}$ ) diminished in size and disappeared the later in development it was applied. The weakness intervention (Naming) showed the opposite pattern, increasing in size the later it was applied. The combined (Both) showed a uniform effect across development. Within each condition, the three replications demonstrated a common profile.


Figure 3: The effect of interventions to remediate weakness on naming accuracy, as well as the combined intervention. Shaded region $=$ period of intervention


Figure 4: The effect of interventions to improve strengths on naming accuracy. Shaded region $=$ period of intervention


Figure 5: The effect of interventions on naming accuracy for typically developing (TD) and impaired networks. Shaded region $=$ period of intervention

Table 1: Naming accuracy at the end of training for typical (TYP), atypical (ATYP), and atypical intervened networks: (a) the base condition; (b) removing plasticity reduction in the semantic and phonological components; (c) removing greater plasticity in pathways. $\mathrm{S}+\mathrm{P}=$ strengths intervention. Naming $=$ weakness intervention. Both $=$ combined. Three replications and average are shown
(a) Naming accuracy at the end of training

|  | TYP | ATYP | Intervention |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | S+P | Naming | Both |
| R1 | $84 \%$ | $49 \%$ | $57 \%$ | $50 \%$ | $52 \%$ |
| R2 | $86 \%$ | $52 \%$ | $57 \%$ | $50 \%$ | $57 \%$ |
| R3 | $86 \%$ | $49 \%$ | $54 \%$ | $51 \%$ | $55 \%$ |
|  |  |  |  |  |  |
| Avg | $85 \%$ | $50 \%$ | $56 \%$ | $50 \%$ | $55 \%$ |

(b) Without plasticity reduction in S and P components

|  | TYP | ATYP | Intervention |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  |  |  | S+P | Naming | Both |
| R1 | $94 \%$ | $49 \%$ | $54 \%$ | $48 \%$ | $51 \%$ |
| R2 | $95 \%$ | $52 \%$ | $55 \%$ | $51 \%$ | $53 \%$ |
| R3 | $95 \%$ | $51 \%$ | $55 \%$ | $50 \%$ | $53 \%$ |
|  |  |  |  |  |  |
| Avg | $95 \%$ | $51 \%$ | $55 \%$ | $50 \%$ | $53 \%$ |

(c) Equalized plasticity in pathways and components

|  | TYP | ATYP | Intervention |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | S+P | Naming | Both |
| R1 | $84 \%$ | $55 \%$ | $61 \%$ | $54 \%$ | $60 \%$ |
| R2 | $81 \%$ | $55 \%$ | $64 \%$ | $53 \%$ | $60 \%$ |
| R3 | $82 \%$ | $54 \%$ | $60 \%$ | $50 \%$ | $59 \%$ |
|  |  |  |  |  |  |
| Avg | $82 \%$ | $55 \%$ | $61 \%$ | $52 \%$ | $60 \%$ |

## Discussion

We used an artificial neural network model of impaired vocabulary development to explore the relative merits of a behavioural intervention to remediate weakness versus one to improve strengths. The two interventions yielded contrasting patterns. The intervention to remediate the weakness - more practice on naming itself - produced an
immediate improvement in naming accuracy, but did not raise the ceiling vocabulary size that could be attained by the model. Intervention had served to propel the model more quickly along the same atypical trajectory. This is because the (lower) ceiling level of performance was constrained by the reduced computational capacity of the naming pathway. By contrast, either improving the semantics representations or the phonological representations - which were otherwise developing typically - produced slower changes during the intervention period, but then long-term gains in the size of the productive vocabulary that the model could acquire. Improving both semantic and phonological representations together gave the largest gains. These gains occurred because semantic and phonological representations became more delineated (or less confusable) through additional training, so that a pathway with limited capacity could achieve higher accuracy. Combining intervention on weakness and strengths gave both immediate gains during intervention and a long-term improvement in the vocabulary size that could be attained.

Table 2: Effects of timing: (a) Phonological + Semantic intervention, (b) Naming intervention, (c) Combined intervention at 100, 250, and 750 epochs

| (a) Phonological+Semantic intervention |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| TYP |  |  |  |  | ATYP |
| R1 | $85 \%$ | $50 \%$ | $57 \%$ | $53 \%$ | 750 |
| R2 | $85 \%$ | $51 \%$ | $57 \%$ | $54 \%$ | $50 \%$ |
| R3 | $87 \%$ | $51 \%$ | $54 \%$ | $55 \%$ | $52 \%$ |
|  |  |  |  |  |  |
| Avg | $86 \%$ | $51 \%$ | $56 \%$ | $54 \%$ | $51 \%$ |

(b) Naming intervention

|  | TYP | ATYP | 100 | 250 | 750 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| R1 | $85 \%$ | $50 \%$ | $50 \%$ | $50 \%$ | $53 \%$ |
| R2 | $85 \%$ | $51 \%$ | $50 \%$ | $51 \%$ | $55 \%$ |
| R3 | $87 \%$ | $51 \%$ | $51 \%$ | $51 \%$ | $55 \%$ |
|  |  |  |  |  |  |
| Avg | $86 \%$ | $51 \%$ | $50 \%$ | $51 \%$ | $54 \%$ |

(c) Combined intervention

|  | TYP | ATYP | 100 | 250 | 750 |
| :--- | :---: | :---: | :---: | :---: | :---: |
| R1 | $85 \%$ | $50 \%$ | $52 \%$ | $54 \%$ | $52 \%$ |
| R2 | $85 \%$ | $51 \%$ | $57 \%$ | $55 \%$ | $55 \%$ |
| R3 | $87 \%$ | $51 \%$ | $55 \%$ | $55 \%$ | $55 \%$ |
|  |  |  |  |  |  |
| Avg | $86 \%$ | $51 \%$ | $55 \%$ | $55 \%$ | $54 \%$ |

We included assumptions about plasticity in the model that there would be sensitive periods in the components but not pathways, that the lower plasticity of the components would be the limiting factor on naming development - but neither proved essential for producing the above effect. We also altered the timing of intervention, and here showed that improving strengths yielded greatest gains early in development, while remediating weaknesses yielded the
greatest gains late in development. Combining both produced a uniform effect across development.

This model represents the first mechanistic demonstration of how working on strengths may serve to improve behavioural outcomes in developmental disorders. The observed effects of improving semantic and phonological representations are in line with those observed empirically for children with WFD (Best, 2005; McGregor \& Leonard, 1989). Two further methods by which improving strengths might improve behavioural outcomes remain to be explored: encouraging compensatory mechanisms through intervention, and bolstering convergent sources of information in interactive systems.

The model nevertheless demonstrated relatively modest accuracy gains through intervention - certainly there was no elimination of the deficit (it was reduced from $35 \%$ to $29 \%$ ). This is in line with general arguments made by Thomas et al. (2017): with some exceptions, where deficits arise through neurocomputational constraints in developing systems, behavioural interventions alone are unlikely to be successful in fully alleviating deficits. The conditions of optimal outcome are, however, a fruitful avenue for computational investigations, in the wider context of narrowing the gap between mechanistic theories of deficit and clinical practices of intervention.

## Acknowledgments

This research was supported by UK Economic and Social Research Council (ESRC) grant no. RES-062-23-2721.

## References

Best, W. (2003). Word-finding in children: Finding the right approach. Royal College of Speech Language Therapists Bulletin, September 2003, 5-6.
Best, W. (2005). Investigation of a new intervention for children with word-finding problems. Int. Journal of Language Communication Disorders, 40(3), 279-318.
Best, W., Fedor, A., Hughes, L., Kapikian, A., Masterson, J., Roncoli, S., Fern-Pollak, L., \& Thomas, M. S. C. (2015). Intervening to alleviate word-finding difficulties in children: Case series data and a neurocomputational foundation. Cognitive Neuropsychology, 32(3-4), 133-168
Bishop, D., Nation, K., \& Patterson, K. (2014). When words fail us: insights into language processing from developmental and acquired disorders. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1634), 20120403.
Bloom, L. (1973). One word at a time: The use of single word utterances before syntax. Mouton: The Hague.
Bock, K., \& Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. Gernsbacher (Ed.), Handbook of Psycholinguistics (pp. 945-984). London: Academic Press.
Dell, G., Schwartz, M., Martin, N., Saffran, E., \& Gagnon, D. (1997). Lexical access in aphasic and nonaphasic speakers. Psychological Review, 104, 801-838.

Harm, M., McCandliss, B. \& Seidenberg, M. (2003). Modeling the successes and failures of interventions for disabled readers. Scientific Studies of Reading 7, 155-182.
Hinton, G. E. \& Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. Psychological Review, 98(1), 74-95.
Laws, J., Campbell, C., Roulstone, S., Adams, C. \& Boyle, J. (2008). Mapping practice onto theory: the speech and language practitioner's construction of receptive language impairment. International Journal of Language Communication Disorders, 43(3), 245-263.
Leonard, L. B. (2014). Children with Specific Language Impairment. Cambridge, MA: MIT Press.
Li, P., Zhao, X., \& MacWhinney, B. (2007). Dynamic selforganization and early lexical development in children. Cognitive Science, 31, 581-612.
Masterson, J., Stuart, M., Dixon, M., \& Lovejoy, S. (2010). Children's printed word database. British Journal of Psychology, 101, 221-242.
Mayor, J., \& Plunkett, K. (2010). A neuro-computational account of taxonomic responding and fast mapping in early word learning. Psychological Review, 117(1), 1-31.
McGregor, K. K., \& Leonard, L. B. (1989). Facilitating word-finding skills of language-impaired children. Journal of Speech and Hearing Disorders, 54, 141-147.
McMurray, B. (2014). Defusing the childhood vocabulary explosion. Science, 317, 631.
Michie, S., \& Prestwich, A. (2010). Are interventions theory-based? Development of a theory coding scheme. Health Psychology, 29(1), 1-8
Plunkett, K., Sinha, C., Møller, M. F. \& Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. Connection Science, 4, 293-312.
Rumelhart, D. E., Hinton, G. E., \& Williams, J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart \& J. L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the microstructure of cognition (Vol. 1) (pp. 318-362). Cambridge, MA: MIT Press.
Takesian, A. E. \& Hensch, T. K. (2013). Balancing plasticity/stability across brain development. Progress in Brain Research, 207, 3-34.
Thomas, M. S. C. \& Karmiloff-Smith, A. (2002). Are developmental disorders like cases of adult brain damage? Implications from connectionist modeling. Behavioral and Brain Sciences, 25(6), 727-788.
Thomas, M. S. C. \& Karmiloff-Smith, A. (2003). Modeling language acquisition in atypical phenotypes. Psychological Review, 110(4), 647-682.
Thomas, M. S. C., Fedor, A., Davis, R., Yang, J., Alireza, H., Charman, T., Masterson, J. \& Best, W. (2017). Computational modeling of interventions for developmental disorders. Submitted for publication.
Vinson, D. P., \& Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. Behavior Research Methods, 40(1), 183-190.

# A Bayesian Model of Memory for Text 

Mark Andrews (mark.andrews@ntu.ac.uk)

Department of Psychology, Nottingham Trent University
Nottingham, NG1 4FQ, UK


#### Abstract

The study of memory for texts has had an long tradition of research in psychology. According to most general accounts of text memory, the recognition or recall of items in a text is based on querying a memory representation that is built up on the basis of background knowledge. The objective of this paper is to describe and thoroughly test a Bayesian model of this general account. In particular, we develop a model that describes how we use our background knowledge to form memories as a process of Bayesian inference of the statistical patterns that are inherent in a text, followed by posterior predictive inference of the words that are typical of those inferred patterns. This provides us with precise predictions about what words will be remembered, whether veridically or erroneously, from any given text. We then test these predictions using data from a memory experiment using a relatively large sample of randomly chosen texts from a representative corpus of British English.


Keywords: Bayesian models; Memory; Reconstructive memory; Text memory;

## Introduction

The seminal study on memory for text ${ }^{1}$ is usually attributed to Bartlett (1932). In this now classic work, Bartlett argued that a person's memory for what they read is based on a reconstruction of the information in the text that is strongly dependent on their background knowledge and experiences. From this seminal work, but especially since the widespread adoption of schema based accounts of text memory beginning in the 1970's (e.g., Mandler \& Johnson, 1977; Schank \& Abelson, 1977; Bower, Black, \& Turner, 1979), there has been something close to a consensus on the broad or general characteristics of human text memory. According to this general account - which we can summarize by the following schematic:


- the recognition or recall of items in a text is based on querying a representation of the text that is built up on the basis of background knowledge and experience.

Although some variant of this general account is widely held, it is essentially an informal and untestable theory. Certainly, there has been ample evidence showing that we use our background knowledge to make inferences and associations concerning text content and that these inferences then influence our memory (e.g. Bransford, Barclay, \& Franks,

[^8]1972; Graesser, Singer, \& Trabasso, 1994; Zwaan \& Radvansky, 1998; Rawson \& Kintsch, 2002, to name but a few). However, in most studies, even fundamental concepts such as memory schemas are not formally defined (see, e.g. Ghosh \& Gilboa, 2014), and ostensibly formal models of knowledge influences on text representation, such as the well known work of Kintsch (1988), often require hand-coding of background knowledge and text structures and can only be applied to small and contrived examples. Consequently, there is no formal or computational account of how background knowledge is used to infer a representation of text content and how memories are then derived from this representation that is sufficiently precise to lead to testable empirical predictions.

In this paper, following general principles followed by Hemmer and Steyvers (2009a, 2009b, 2009c) in their studies on memory for visual objects and natural scenes, we describe a probabilistic model that uses Bayesian inference to infer a representation of a text's content on the basis of background knowledge and then uses posterior predictive inference to represent the memories of that text. This provides us with precise predictions about what words will be remembered, whether veridically or erroneously, from any given text. We then test these predictions using data from a memory experiment using a relatively large sample of randomly chosen texts from a representative corpus of British English.

## Probabilistic Model

We begin with the assumption that our background knowledge that is relevant for our memory of text is primarily knowledge of the statistical patterns across spoken and written language. Given any probabilistic language model that specifies these statistical patterns, as we explain below, we may then use Bayes's rule to infer which patterns are inherent in any given text. From this, we may then predict, via posterior predictive inference, which words are and are not typical or compatible with the inferred statistical representation of the text. This effectively serves as the memory of the content of the text. As such, this provides a computational description of the previous schematic, i.e.,


In practical terms, we have many options for our choice of probabilistic language model. However, probabilistic topic
models (see, e.g. Griffiths, Steyvers, \& Tenenbaum, 2007; Steyvers \& Griffiths, 2007; Blei, 2012) have proved highly effective in capturing the statistical patterns that characterize the coarse-grained "discourse topics" across spoken and written language. Here, we use a type of probabilistic topic model known as a hierarchical Dirichlet process mixture model (HDPMM) (Teh, Jordan, Beal, \& Blei, 2006).

A HDPMM is a probabilistic generative model of bag-ofwords ${ }^{2}$ language data. It treats a corpus of language data as a set of $J$ texts $w_{1}, w_{2} \ldots w_{j} \ldots w_{J}$, where text $j$, i.e., $w_{j}$, is a set of $n_{j}$ words from a finite vocabulary, represented simply by the $V$ integers $\{1,2 \ldots V\}$. From this, we have each $w_{j}$ defined as $w_{j}=w_{j 1}, w_{j 2} \ldots w_{j i} \ldots w_{j n_{j}}$, with each $w_{j i} \in\{1 \ldots V\}$. As a generative model of this corpus, the HDPMM treats each observed word $w_{j i}$ as a sample from one of an underlying set of component distributions, $\phi_{1}, \phi_{2} \ldots \phi_{k} \ldots$, where each $\phi_{k}$ is a probability distribution over $\{1 \ldots V\}$. Each $\phi_{k}$ effectively identifies a "discourse topic". For example, here is a sample of 6 topics from an inferred model, where we show the 7 most probable words in each topic:

| theatre | music | league | prison | rate | pub |
| :---: | :---: | :---: | :---: | :---: | :---: |
| stage | band | cup | years | cent | guinness |
| arts | rock | season | sentence | inflation | beer |
| play | song | team | jail | recession | drink |
| dance | record | game | home | recovery | bar |
| opera | pop | match | prisoner | economy | drinking |
| cast | dance | division | serving | cut | alcohol |

The identity of the particular topic distribution from which $w_{j i}$ is drawn is determined by the value of a discrete latent variable $x_{j i} \in\{1,2 \ldots k \ldots\}$ that corresponds to $w_{j i}$. The probability distribution over the possible values of each $x_{j i}$ is given by a categorical distribution $\pi_{j}$, i.e., $\pi_{j}=\pi_{j 1}, \pi_{j 2} \ldots \pi_{j k} \ldots$, where $0 \leq \pi_{j k} \leq 1$ and $\sum_{k=1}^{\infty} \pi_{j k}=1$, that is specific to text $j$. Each $\pi_{j}$ is assumed to be drawn from a Dirichlet process prior whose base distribution, $m$, is a categorical distribution over the positive integers and whose scalar concentration parameter is $a$. The $m$ base distribution is assumed to be drawn from a stick breaking distribution with a parameter $\gamma$. As such, the generative model of the corpus is as follows:

$$
\begin{array}{lll}
w_{j i} \mid x_{j i}, \phi \sim \operatorname{dcat}\left(\phi_{x_{j i}}\right), & x_{j i} \mid \pi_{j} \sim \operatorname{dcat}\left(\pi_{j}\right), \quad i \in 1 \ldots n_{j} \\
& \pi_{j} \mid a, m \sim \operatorname{ddp}(a, m), \quad j \in 1 \ldots J \\
& m \mid \gamma \sim \operatorname{dstick}(\gamma),
\end{array}
$$

where dcat is a categorical probability distribution, ddp is a Dirichlet process, and dstick is a stick breaking distribution. The prior on the component distributions $\phi_{1} \ldots \phi_{k} \ldots$ was a Dirichlet distribution with concentration parameter $b$ and length $V$ location parameter $\psi$.

Having inferred a HDPMM on the basis of a corpus of language data $\mathcal{D}$, given any new text, $w_{j^{\prime}}$, we can use Bayes's rule to infer the posterior probability over $\pi_{j^{\prime}}$, which is the probability distribution over the discourse topics in $w_{j^{\prime}}$ :

$$
\mathrm{P}\left(\pi_{j^{\prime}} \mid w_{j^{\prime}}, \mathcal{D}\right) \propto \mathrm{P}\left(w_{j^{\prime}} \mid \pi_{j^{\prime}}, \mathcal{D}\right) \mathrm{P}\left(\pi_{j^{\prime}} \mid \mathcal{D}\right)
$$

[^9]We may then use the posterior predictive distribution to infer the words that are typical of the topics inherent in $w_{j^{\prime}}$. The predicted probability of word $w_{j^{\prime} i^{\prime}}$ given text $w_{j^{\prime}}$ is given by

$$
\mathrm{P}\left(w_{j^{\prime} i^{\prime}} \mid w_{j^{\prime}}, \mathcal{D}\right)=\int \mathrm{P}\left(w_{j^{\prime} i^{\prime}} \mid \pi_{j^{\prime}}, \mathcal{D}\right) \mathrm{P}\left(\pi_{j^{\prime}} \mid w_{j^{\prime}}, \mathcal{D}\right) d \pi_{j^{\prime}}
$$

Corpus As our language corpus, we used the British National Corpus (BNC) (BNC Consortium, 2007). From the entire BNC, we extracted all sections that were tagged as paragraphs. This gave us a corpus with a total word count of $87,564,696$ words. From this, we created a set of 184,271 texts, each between 250 and 500 words long. These were created by using either single paragraphs in this count range, or concatenating consecutive paragraphs until they were within this range. The total word count of this set of texts was $78,723,408$ words. We then restricted the word types by excluding words that occurred less than 5 times in total, and any words on either of two lists of stopwords, and any words that were not listed in a dictionary of $\approx 60 \mathrm{~K}$ English words. This lead to a final vocabulary of 49,328 word types. For more information, see Footnote ${ }^{3}$.

Inference We used a Gibbs sampler to infer the posterior distribution over the values of latent variables, i.e., $\left\{x_{j i}: j \in\right.$ $\left.1 \ldots J, i \in 1 \ldots n_{j}\right\}$, as well as the hyper-parameters $m, a, b, \psi$, and $\gamma$. For more information, see Footnote ${ }^{4}$

Prediction From the entire set of paragraphs in the BNC, we randomly sampled 50 paragraphs whose length was $150 \pm 10$ words, where at least $90 \%$ of the words are in the aforementioned dictionary of English words, and where at least 75\% of the words were in a set of words for which word association norms exists (see the following section for more details on the word association norms we used). For more information, see Footnote ${ }^{5}$.

For each of the 50 sampled texts, we then used posterior predictive inference, as described above, to obtain the probability distribution over words that are typical or compatible with the topic based representation of each text. As explained above, this distribution effectively provides the inferred model's memory of the content of the text. A Gibbs sampler was used to infer each text's posterior distribution over $\pi$, which is the probability distribution over discourse topics in that text. Two example texts and their posterior predictive inferences are shown in Figure 1. For more information, see Footnote ${ }^{6}$.

[^10]Improve your mood and counteract stress: Ask anyone who exercises regularly and they will tell you that they always feel exhilarated at the end of a session even if they had begun by feeling that they were not in the mood for exercise and had almost forced themselves to continue. Physical fitness also provides considerable protection against stress and the illnesses it can cause. So, however busy your life, perhaps you could try and fit some regular exercise into your day. Let it be something which is in complete contrast to the way you normally spend your time. One word of warning though: if you are someone whose daily life involves a strong competitive element, you would do well to avoid too much in the way of competitive sport (squash, tennis and so on) as your form of exercise as these will only tend to maintain an already high level of stress. routine walk $_{\text {swinming }}$ fit training weight aerobics health yoga anxiety programme rest session fitness increase life running week jogging rate level
 ${ }_{\text {begin }}$ muscles ${ }_{g y m}$ minutes mood heart strength body muscle physical day time

Developmental norms are an attempt to provide an indication of the ages at which one might expect ordinary children to show evidence of certain skills or abilities. Since children vary with respect to the ages at which they demonstrate any particular behaviour, norms represent an average obtained from an examination of the developmental changes occurring in a large number of children. Data from a large sample will show the earliest age at which a child would be expected to gain control of a particular aspect of language, and the age by which 90 per cent or 95 per cent of non-handicapped children might be expected to show evidence of the same ability. If children who have already been diagnosed as suffering from some specific handicapping condition are included, the data will show the expected age delay before this group matches the performance of the normally developing children.


Figure 1: An example of two of the texts used in the memory experiment, and samples from the HDPMM's posterior prediction for each one. The predicted words are scaled as a function of their predicted probability, and we show the 50 most highly predicted words (excluding stopwords and words not in the vocabulary) for each text. Words in italics are predicted words that were not in the text itself. These, in effect, are the model's false memories.

## Comparison models

The focus of our analysis is whether the probability of recognizing or recalling any given word having read a particular text is predicted by our HDPMM's posterior predictive distribution over words for that text. To properly evaluate the model's predictions, it is necessary to compare them to those of other plausible models. Here, we will compare the Bayesian model to predictions made by two associative models. Both of these models predict that the words that are remembered from a text are those that are most associated, on average, with the text's content. Associative models are strong models to compare to the Bayesian model because associative strength has been repeatedly shown to a strong predictor of memory for words in word lists (e.g., Roediger, Watson, McDermott, \& Gallo, 2001; Gallo, 2006).

The statistical co-occurrence probability of two words, $w_{k}$ and $w_{l}$, which we will denote $\mathrm{P}_{\mathrm{C}}\left(w_{k}, w_{l}\right)$, is defined as the empirical probability of observing word $w_{k}$ and $w_{l}$ in the same text ${ }^{7}$ in the language. Here, we calculate $\mathrm{P}_{\mathrm{C}}\left(w_{k}, w_{l}\right)$ using

[^11]the same BNC corpus as was used above, i.e. with the same 184,271 texts each between 250 and 500 words. From this, we can calculate
$$
\mathrm{P}_{\mathrm{C}}\left(w_{k} \mid w_{l}\right)=\frac{\mathrm{P}_{\mathrm{c}}\left(w_{k}, w_{l}\right)}{\mathrm{P}_{\mathrm{C}}\left(w_{l}\right)}
$$
which is the conditional probability of observing $w_{k}$ in any text given that $w_{l}$ has been observed. From this, if text ${ }_{j}=$ $w_{j 1}, w_{j 2} \ldots w_{j n_{j}}$, the predicted association probability of word $w_{k}$ according to text ${ }_{j}$ is
$$
\mathrm{P}_{\mathrm{C}}\left(w_{k} \mid \operatorname{text}_{j}\right)=\frac{1}{n_{j}} \sum_{i=1}^{n_{j}} \mathrm{P}_{\mathrm{C}}\left(w_{k} \mid w_{j i}\right)
$$

We can interpret this value intuitively as the average association between $w_{k}$ and text ${ }_{j}$, with association defined in terms of statistical co-occurrences in the language.

An alternative means to calculate the average association between $w_{k}$ and text ${ }_{j}$ is using word association norms, rather

[^12]than statistical co-occurrences. If $A_{k l}$ is the frequency that word $w_{k}$ is stated as associated with word $w_{l}$, then the conditional probability of word $w_{k}$ given $w_{l}$ is
$$
\mathrm{P}_{\mathrm{A}}\left(w_{k} \mid w_{l}\right)=\frac{A_{k l}}{\sum_{i=1}^{V} A_{i l}},
$$
where $V$ is the total number of words in our vocabulary of response words. Now, given text ${ }_{j}=w_{j 1}, w_{j 2} \ldots w_{j n_{j}}$, we can calculate
$$
\mathrm{P}_{\mathrm{A}}\left(w_{k} \mid \operatorname{text}_{j}\right)=\frac{1}{n_{j}} \sum_{i=1}^{n_{j}} \mathrm{P}_{\mathrm{A}}\left(w_{k} \mid w_{j i}\right)
$$
which we can interpret as the average association between $w_{k}$ and text ${ }_{j}$, with association now defined in terms of word association norms rather than statistical co-occurrences. Though a large set of English word association norms are available from the widely used Nelson norms (Nelson, McEvoy, \& Schreiber, 2004), we used an even larger set that is a prerelease of the English small world of words association norms (De Deyne \& Storms, 2017). This provided word associates, produced by 101,119 participants, to 10,050 word types. For more information, see Footnote ${ }^{8}$.

## Experiment

Our aim in this experiment is to measure participants' memory of the 50 sampled texts described in the previous section. Participants read these texts at their normal reading speed and then their memory for what they have read is tested using both recall and recognition tasks. We will then compare the pattern of results from our participants with the predictions of the models.

## Methods

Participants 216 people ( 113 female, 103 male) participated in the experiment. The ages ranged from 17 to 78 years, with a median of 34 years. Participants were recruited from the student and general populations, with the only restriction being that they be native English speakers.
Design Pre-experiment sample size determination calculations showed that, given the reasonable assumptions of both inter-text and inter-subject variability in memory performance, a relatively large number of texts and participants was necessary. In particular, we showed that there is a high probability of detecting effects, even when these effects are relatively weak, if we have at least 50 texts and at least 150 subjects are used. Importantly, these results hold even when each subject sees only a small subset of total number of texts, and this subset can be as low as 3 texts per each participant. We therefore used all 50 texts described above, and initially aimed for approximately 200 participants, with each participant being tested with a randomly sample of 3 texts.

[^13]Materials The texts used as stimuli for this experiment were the above mentioned 50 texts.

For the recognition tasks, test word lists with 20 words each were created. Of the 20 words in each list, 10 were present in the to-be-memorized text, while the remaining 10 were not present in it. For each text, the list was created as follows. Key words were extracted from each text and also from the surrounding paragraphs to that text in the BNC. This was done by calculating the tfidf (term frequency, inverse document frequency) value for each word, and then applying a threshold to exclude the less informative words. 10 words were then randomly selected from the key words of each text. A further 10 words were randomly sampled from the key words of the surrounding paragraphs excluding any words the in the main text itself. This set of 10 words were therefore not present in the text to be memorized, but given that they were selected from surrounding paragraphs, they were likely to be meaningfully related to it. As such, they would serve a useful items on the recognition memory test as they could not easily be dismissed without a proper search of memory. For more information, see Footnote ${ }^{9}$.


Figure 2: The task diagram of one block in the experiment: Participants read a randomly assigned text, perform a filler task, and then have their memory tested using either a recognition or recall test, with the test type being randomly chosen. This process is repeated three times for each participant.

Procedure Each experiment session proceeded as follows (see also Figure 2):

- After initial information and instructions, which informed participants that they would be engaging in memory tasks, one of the sample texts appeared on screen. Participants were instructed to read this text at their normal reading. The text stayed on screen for a maximum of 90 seconds,

[^14]but after 45 seconds, participants were able to move on the next screen if they so wished.

- On the following screen participants were asked to play the computer game Tetris for exactly 60 seconds.
- At the completion of the game, participants proceeded to the memory task. For each participant and for each text, the memory test was randomly chosen to be either a recognition or a recall task.
- For the recognition test, the 20 test items were presented on screen, one word at a time, with an inter-stimulusinterval of 2 seconds. They remained on screen for 5 seconds or until the subject indicated with a button press whether the word shown was present or absent from the text. No feedback was given after each response.
- If the participant was assigned to the recall test, a screen of a list of small empty text boxes was presented where and they were asked to type as many words as they could remember, one word into each text box. Initially, 10 empty texts boxes were presented, and more boxes could be added with a button press.
- Upon completion of the memory test, participants were given the option of pausing or proceeding to the next test. Each participant performed three tests in total, with the three texts to which they were assigned being always randomly sampled from the set of 50 texts.
The experiments were presented using the Wilhelm ${ }^{10}$ webbrowser based experiment presentation software that was hosted at https://www.cognitionexperiments.org. This software allowed the experiment to be done any web-browser based device, e.g., phones, tablets, laptops and desktops.


## Results

For more information about the results, see Footnote ${ }^{11}$.
Descriptives In the recognition memory tests trials, the overall accuracy rate was $76 \%$. Overall, the false positive rate, i.e. where participants responded "present" to words that were not present in the text they read, was $27 \%$. The false negative rate, i.e. where participants responded "absent" to words that actually were present in the text, was $22 \%$. For the recall tests, the median number of recalled words per each test was 7 , with between 2 and 15 words recalled in $95 \%$ of tests. The overall accuracy of recall was $70 \%$, and thus there was an overall false recall rate of $30 \%$.

Model evaluation For the recognition memory data, we model how well each model predicts the behavioural results using a random effects logistic regression model. In other words, for each of the models being evaluated, we fit the

[^15]recognition memory data using the same random effects logistic regression but using a different predictor variable in each case. The logistic regression model is
$$
\log \left(\frac{p_{i}}{1-p_{i}}\right)=\alpha+\alpha_{s_{i}}+\alpha_{t_{i}}+\left(\beta_{s_{i}}+\beta_{t_{i}}+\beta\right) \phi_{i}+b x_{i}
$$
where $i$ indexes the experiment trial, $p_{i}$ is the probability of the participant responding "present" to the word presented on trial $i, s_{i}$ is the identity of the participant on trial $i, t_{i}$ is the identity of the text on trial $i, \phi_{i}$ is the log of the model's predicted probability of the word on trial $i, x_{i}$ indicates if the word on trial $i$ was present in text $t_{i}$. The random effects regression coefficients are $\alpha_{s_{i}}, \alpha_{t_{i}}, \beta_{s_{i}}, \beta_{t_{i}}$, which are modelled as drawn from zero-mean Normal distributions.

Having fit the logistic regression model using the predictions of the HDPMM topic model, the co-occurrence based model, the association norm based model, and a null model (where $\phi_{i}$ is set to 0 for all $i$ ), we calculate model fit statistics such as BIC, AIC, and Deviance. They are shown in the following table:

|  | HDPMM | Co-occur | Assoc | Null |
| ---: | :---: | :---: | :---: | :---: |
| BIC | 5775.68 | 5824.33 | 6083.58 | 6212.77 |
| AIC | 5715.97 | 5764.62 | 6023.87 | 6186.23 |
| Deviance | 5697.97 | 5746.62 | 6005.87 | 6178.23 |

We will concentrate on the BIC results as the $\log _{e}$ of the Bayes Factor comparing any model $\mathscr{M}_{0}$ to model $\mathscr{M}_{1}$ can be approximated by half the difference of the BIC of models $\mathcal{M}_{1}$ and $\mathcal{M}_{0}$. Thus, the $\log _{e}$ of the Bayes factor comparing the HDPMM predictions to those of the co-occurrence based association model is 24.32 . By any standard, this is overwhelming evidence in favour of the predictions of the HDPMM relative to those of the co-occurrence model. For example, Kass and Raftery (1995) argue that a $\log$ Bayes factor on a $\log _{10}$ scale that is greater than 2.0 is already decisive evidence in favour of the better model. In our case, our $\log _{e}$ result of 24.32 is 10.42 on a $\log _{10}$ scale. As the BIC of the association norm model is even greater than that of the co-occurrence model, there is overwhelming evidence in favour of the HDPMM relative to the comparison models.

For the recall memory task results, each set of recalled words by a participant on any given test $j$, which we will denote by $\omega_{j}=\omega_{j 1}, \omega_{j 2} \ldots \omega_{j n}$, can be reasonably viewed as draws from a subjective probability distribution that is the participant's memory representation of the contents of the text. We can calculate the likelihood of this data according to the probability distribution defined by any of our models, denoted generically by $\psi$, as follows:

$$
\mathrm{P}\left(\omega_{j} \mid \psi\right)=\prod_{i=1}^{n} \prod_{v=1}^{V} \psi_{v}^{\mathbb{I}\left(r_{i}=v\right)}=\prod_{v=1}^{V} \psi_{v}^{r_{j v}}
$$

where $\mathbb{I}(\cdot)$ is an indicator variable that takes the value of 1 if its argument is true, and $r_{j v}$ is the number of times that word $w_{v}$ occurs in $\omega_{j}$, which in this case will be either $r_{j v}=1$ if
word $w_{v}$ was recalled and $r_{j v}=0$ otherwise. The $\log _{e}$ of the likelihood of all the recall memory task data is

$$
\log _{e} \prod_{j=1}^{L} \mathrm{P}\left(\omega_{j} \mid \psi\right)=\log _{e} \prod_{j}^{L} \prod_{v=1}^{V} \psi_{v}^{r_{j v}}=\sum_{j}^{L} \sum_{v=1}^{V} r_{j v} \log _{e} \psi_{v}
$$

These results are presented in the following table:

|  | HDPMM | Co-occur | Assoc |
| ---: | ---: | ---: | ---: |
| logLik | -14109.02 | -15100.94 | -16039.98 |
| Deviance | 28218.03 | 30201.88 | 32079.96 |

Given that the deviance is equal to the BIC plus a constant term, the difference of the deviances is identical to the difference of the corresponding BIC's. Approximating the $\log _{e}$ of the Bayes factor by half this difference, we therefore calculate a $\log _{10}$ Bayes factor for the evidence for the HDPMM predictions relative to those of the nearest model, the co-occurrence based association model, as 430.79 . On the basis of the interpretation described above, this is again overwhelming evidence in favour of the HDPMM.

## Discussion

In this paper, we have proposed - and then tested using a high powered behavioural experiment - a Bayesian account of how we form memories for spoken and written language. This account models how we use our background knowledge to form memories as a process of Bayesian inference of the statistical patterns that are inherent in each text, followed by posterior predictive inference of the words that are typical of those inferred patterns. We have implemented this model specifically as a HDPMM and applied it to an approximately 80 m word corpus of texts taken from the BNC. This allowed us to make predictions of the probability of remembering any given word in each text from a sample of texts taken from the BNC. We tested these predictions in a behavioural experiment with 216 participants. The results of the analysis from both the recognition and recall data provided overwhelming evidence in favour of the Bayesian model relative to non-trivial alternative models.

## References

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press.
Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.
BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). Oxford University Computing Services: http://www.natcorp.ox.ac.uk/.
Bower, G., Black, J., \& Turner, T. (1979). Scripts in memory for text. Cognitive Psychology, 11(2), 177-220.
Bransford, J. D., Barclay, J. R., \& Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. Cognitive psychology, 3(2), 193-209.
De Deyne, S., \& Storms, G. (2017). Small world of words, www.smallworldofwords.org.

Gallo, D. (2006). Associative illusions of memory: False memory research in DRM and related tasks. Psychology Press.
Ghosh, V. E., \& Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. Neuropsychologia, 53, 104-114.
Graesser, A., Singer, M., \& Trabasso, T. (1994). Constructing inferences during narrative text comprehension. Psychological review, 101(3), 371-395.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114(2), 211-244.
Hemmer, P., \& Steyvers, M. (2009a). A Bayesian account of reconstructive memory. Topics in Cognitive Science, 1(1), 189-202.
Hemmer, P., \& Steyvers, M. (2009b). Integrating episodic and semantic information in memory for natural scenes. In Proceedings of the 31th annual conference of the cognitive science society (pp. 1557-1562).
Hemmer, P., \& Steyvers, M. (2009c). Integrating episodic memories and prior knowledge at multiple levels of abstraction. Psychonomic Bulletin \& Review, 16(1), 80-87.
Kass, R. E., \& Raftery, A. E. (1995). Bayes factors. Journal of American Statistical Association, 90(430), 773-795.
Kintsch, W. (1988). The role of knowledge in discourse comprehension - A construction integration Model. Psychological Review, 95(2), 163-182.
Mandler, J. M., \& Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. Cognitive psychology, 9(1), 111-151.
Nelson, D., McEvoy, C., \& Schreiber, T. (2004). The university of south florida word association, rhyme and word fragment norms. Behavior Research Methods, Instruments, \& Computers, 36, 408-420.
Rawson, K. A., \& Kintsch, W. (2002). How does background information improve memory for text content? Memory \& cognition, 30(5), 768-778.
Roediger, H. L., Watson, J. M., McDermott, K. B., \& Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. Psychonomic Bulletin \& Review, 8(3), 385-407.
Schank, R. C., \& Abelson, R. P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum Associates.
Steyvers, M., \& Griffiths, T. (2007). Probabilisitic topic models. In T. Landauer, D. McNamara, S. Dennis, \& W. Kintsch (Eds.), Handbook of latent semantic analysis. Psychology Press.
Teh, Y. W., Jordan, M. I., Beal, M. J., \& Blei, D. M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476), 1566-1581.
Zwaan, R. A., \& Radvansky, G. A. (1998). Situation models in language comprehension and memory. Psychological bulletin, 123(2), 162.

# Numbers Uniquely Bias Spatial Attention: A Novel Paradigm for Understanding Spatial-Numerical Associations 

Lauren S. Aulet (lauren.s.aulet@emory.edu)<br>Department of Psychology, Emory University<br>Atlanta, GA 30322 USA

Sami R. Yousif (sami.yousif@yale.edu)<br>Department of Psychology, Yale University<br>New Haven, CT, 06520 USA

Stella F. Lourenco (stella.lourenco@emory.edu)<br>Department of Psychology, Emory University<br>Atlanta, GA 30322 USA


#### Abstract

Over the past two-and-a-half decades, numerous empirical studies have demonstrated a relationship between numbers and space. A classic interpretation is that these spatialnumerical associations (SNAs) are a product of a stable mental number line (MNL) in the mind, yet others have argued that SNAs are a product of transient mappings that occur in working memory. Importantly, although the latter interpretation has no implications for the representation of number, the former suggests that the representation of number is inherently spatial. Here, we tease apart questions of spatial representation (à la an MNL perspective) and spatial strategy (à la alternative accounts). In a novel place-the-number task, we demonstrate that numbers automatically bias spatial attention whereas other ordinal sequences (i.e., letters) do not. We argue that this is evidence of an inherently spatial representation of number and explore how this work may help answer future questions about the relationship between space and number.


Keywords: spatial-numerical associations (SNAs); mental number line (MNL); automaticity; working memory; polarity correspondence; synesthesia

## Introduction

Since the seminal work of Dehaene, Bossini, and Giraux (1993), the link between space and number has inspired a wealth of research (for recent review, see Fischer \& Shaki, 2014). In the classic paradigm, participants made parity judgments (odd/even) of Arabic numerals using left and right response keys, finding that participants responded faster to smaller numbers when using the left key and faster to larger numbers when using the right key. This general finding has since been replicated using numerous paradigms. One such example is the magnitude comparison task in which participants indicated whether the digit shown is greater than or less than some value (e.g., 5; Fitousi, Shaki, \& Algom, 2009). Later work demonstrated that simply perceiving numbers biases spatial attention: participants were faster to detect a leftward target when primed with a small digit and faster to detect a rightward target when primed with a large digit
(Fischer, 2003; but see, e.g., Zanolie \& Pecher, 2014, for replication failure). Further, changes in spatial attention bias number generation: when asked to randomly generate numbers while making alternating left/right head movements, participants more frequently generate small numbers when their head is oriented to the left and large numbers when their head is oriented to the right (Loetscher, Schwarz, Schubiger, \& Brugger, 2008).

A common theory of the spatial-numerical associations (SNAs) described above is that they are the product of a stable mental number line (MNL), wherein smaller numbers are represented on one side of space and larger numbers are represented on the other (in Western cultures, smaller numbers are represented on the left and larger numbers on the right; e.g., Dehaene et al., 1993). Yet there remain objections to this theory. Proctor and Cho (2006), for example, argued that polarity correspondence (a +/- categorization of stimulus and response) can explain the observed associations. Indeed, many tasks rely on a dichotomous response (e.g., left/right keys, left/right head position), and may be explained in this way. Another view, which has posed an even greater challenge to the MNL account, argues that SNAs are a product of task-specific associations established online within working memory (WM; van Dijck \& Fias, 2011). The crux of this debate is whether the observed SNAs are driven by a stable spatialnumerical link (e.g., MNL) or by transient mappings of number onto space (e.g., polarity correspondence, or a WM account; for further discussion, see Cheung et al., 2015).

In general, those who argue in favor of a WM account argue that the ostensibly transient mappings are, at least in part, a product of task demands. For example, in the classic parity judgment task, participants respond using leftward and rightward oriented keys. One may argue that the relative orientation of these keys is sufficient to induce a spatial mapping (see Viarouge, Hubbard, \& Dehaene, 2014 for discussion on the induction of spatial reference frames in SNA tasks).

Table 1: SNA tasks and their task demands.

| Category | Examples | Task demands |  |  | Ordinal control |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Dichotomous Categorization | Directional prime | Magnitude salience |  |
| Parity Judgment | Dehaene et al., 1993 <br> Marghetis et al., 2013 | $\checkmark$ | $\checkmark$ | $\checkmark$ | $\checkmark$ |
| Mag. Judgment | Fitousi et al., 2009 <br> Marghetis et al., 2013 | $\checkmark$ | $\checkmark$ | $\checkmark$ | X |
| Lat. Comparison | Lavidor et al., 2004 Cheung et al., 2015 | $\checkmark$ | $\checkmark$ | $\checkmark$ | X |
| Numerical Posner | Fischer et al., 2003 <br> Ruiz Fernández et al., 2001 | X | $\checkmark$ | X | X |
| Num. Bisection | Fischer et al., 2001 Calabria \& Rossetti, 2005 | X | $\checkmark$ | X | X |
| Number generation | Loetscher et al., 2008 <br> Cheung et al., 2015 | X | $\checkmark$ | $\checkmark$ | X |
| Eye-tracking | Holmes et al., 2016 <br> Schwarz \& Keus, 2004 | X | X | $\checkmark$ | X |
| Place-the-number |  | X | X | X | $\checkmark$ |

In particular, a polarity correspondence account would be concerned about the use of a dichotomous response. Indeed, many of the SNA tasks mentioned above possess some kind of task demand. Parity judgment tasks and magnitude comparison tasks involve both a dichotomous manual response (left key/right key) and a dichotomous judgment (less than/greater than). Other parity and magnitude judgment tasks have utilized a go/no-go paradigm to circumvent the spatial information provided by the response keys, but these tasks nevertheless depend on a dichotomous response scheme (e.g., Marghetis, Kanwal, \& Bergen, 2013). This dichotomous response, though not spatial, nevertheless lends itself to a polarity correspondence account. Furthermore, even tasks that do not require a dichotomous response still have certain features that may instantiate a left-to-right reference frame. For example, in the work of Fischer and colleagues (2003), a left-to-right frame may be induced by the locations of the target (as either on the left or right side of fixation). The same may be said for the paradigm utilized by Ruiz Fernández and colleagues (2011), wherein, after presentation of a number, they made an arbitrary selection between items construed on the left and right sides of space. In the work of Loetscher and colleagues (2008), a left-to-right frame is being specifically induced by the turning of the head. (For a more complete list of SNA paradigms and their task demands, see Table 1). Thus, it is unclear whether there is any evidence of an SNA (and consequently, a stable MNL) in the absence of any such task demands.

Eye-tracking paradigms have been promising in this regard. For instance, Holmes, Ayzenberg, and Lourenco (2016) had participants play a virtual blackjack game while their eye gaze was being tracked. It was found that both the value of a card on a given trial as well as the overall value
of one's hand at a given time significantly predicted eye gaze in a manner consistent with observed SNAs for Western participants (i.e., smaller magnitudes produced more leftward eye movements and larger magnitudes produced more rightward eye movements). This study provides strong evidence for a left-to-right oriented MNL by demonstrating that number representations bias spatial attention even in the absence of a directional prime and a dichotomous response scheme (see Table 1). Other eyetracking studies have yielded similar results (e.g., Schwarz \& Keus, 2004; Loetscher et al., 2010). Yet in Holmes et al. (2016), the task requires the explicit processing of numerical value (e.g., value of a card or hand). Though the processing of numerical magnitude may not be explicitly required in the other tasks above, they do invoke some property of number (e.g., parity). As such, two questions remain: do numbers automatically bias spatial attention in the absence of a directional prime and even when numerical properties are irrelevant to the task at hand? Furthermore, and critically, is this bias specific to number?

## Automaticity as a criterion for representation

Understanding whether SNAs manifest automatically (i.e., in the absence of task demands) is crucial for understanding the relationship between space and number, in large part because automaticity suggests that the relationship is representational (as an MNL hypothesis would predict) rather than transient (as a polarity correspondence or WM account would predict). Nowhere is this criterion more apparent than in the literature on synesthesia. Automaticity, here, is where many have drawn the line between a relationship that is merely associative as opposed to truly synesthetic (see Grossenbacher \& Lovelace, 2001; Mattingley, 2009). We argue that the spatial-numerical
relationship should be considered in similar terms (see also, Cohen Kadosh \& Henik, 2007). By this criterion, automaticity helps us to understand the nature of the relation between space and number: namely, whether they share representational space (as an MNL hypothesis predicts), or whether the two are only transiently associated with one another (as alternative hypotheses predict).

Here, we present evidence from a novel SNA paradigm -the place-the-number task -- which suggests that numbers do in fact automatically bias spatial attention. Very simply, participants viewed a number on a screen, memorized its location, and, after a delay, placed the number back in its original location. This task revealed a robust spatialnumerical relationship. In two additional control experiments, we found no consistent mapping of letters to space, suggesting that these attentional biases are specific to number and not ordinal sequences (a control which has not always been tested with other paradigms). Predictions made by MNL and WM accounts of SNAs diverge in such conditions: an MNL account predicts that this spatial bias is specific to number whereas a WM account predicts that this bias generalizes to any ordinal sequence (e.g., letters, months, etc.; van Dijck et al, 2014).

## Experiment 1: Place-the-number task

## Method

In this novel paradigm, participants viewed an Arabic numeral (1-9) presented in black font within a rectangle (white fill with black outline; $918 \times 495$ pixels). This task was created in Visual Basic and presented on a 19in computer monitor. Participants sat approximately 65 cm from the monitor. Each digit was presented 20 times, for a total of 180 trials, randomly ordered. On each trial, participants were instructed to remember the location of the digit. The digit remained on screen until participants clicked a button located at the bottom of the screen, at which time the digit disappeared. Participants were then instructed to click the remembered location to place the digit at that location. These instructions were presented in a pop-up dialog box, which also ensured that participants did not fixate on the original location of the digit. Participants could further adjust this initial placement by dragging and dropping the digit. Participants then confirmed their final placement by clicking another button on screen and immediately proceeded to the next trial.

Thirty-seven undergraduates participated in this task for course credit. All participants had normal or corrected-to normal vision. Procedures were approved by the Institutional Review Board (IRB). One participant was excluded from the statistical analyses due to poor accuracy ( $>2.5$ SD from the group mean), where accuracy is calculated as the distance between the digit's original location and the participant's final placement.

## Results

The remaining participants $(N=36)$ had a mean accuracy of 18.76 pixels ( $\mathrm{SD}=13.67$ ). The variable of interest was participants' accuracy along the horizontal axis ${ }^{1}$. For each trial, we calculated accuracy as the difference between the x-coordinate of the participant's final placement and $x$ coordinate of the digit's original location, such that a negative value represents a more leftward placement, in comparison to the original location, and a positive value represented a more rightward placement, in comparison to the original location. For each participant, we calculated the mean accuracy for each digit and calculated a slope by regressing these values onto their corresponding numerical value (See Fig 1). Thus, in this paradigm, a positive slope represents the canonical, left-to-right SNA. Participants' slopes were significantly greater than zero, $t(35)=2.11, p<$ $.05, d=.35$. Furthermore, a significant number of participants $(N=24)$ showed this effect (binomial test, $p<$ .05). In other words, participants placed smaller numbers more leftward than larger numbers, consistent with a left-toright $\mathrm{SNA}^{2}$. To determine whether the SNA shown here is unique to number (as an MNL account predicts) or occurs for any ordinal sequence (as a WM account predicts), we conducted a control experiment (Exp. 2A) with letters.


Figure 1: Scatterplot displaying mean spatial bias (final original placement) for digits 1-9 including the best-fitting regression line.

[^16]
## Experiment 2A: Letter control (A-I)

## Method

The procedure for Experiment 2 A was identical to Experiment 1 except instead of Arabic numerals as stimuli, participants were presented with the first nine letters of the alphabet (A-I). Thirty-eight undergraduates participated for course credit. One participant was excluded from the statistical analyses as they did not complete all trials. One participant was excluded from the statistical analyses for poor accuracy.

## Results

The remaining participants ( $N=36$ ) had a mean accuracy of 14.25 pixels $(\mathrm{SD}=7.62)$. Importantly, unlike Experiment 1, participants' slopes were not significantly different from zero, $t(35)=-.85, p=.40$. These results demonstrate that letters, although ordinal, do not generate a spatial association in this paradigm. However, since the letters used only spanned the beginning of the alphabet, it remains possible this sequence was not comparable to the Arabic numerals used in Experiment 1. The following experiment was designed to address this concern.


Figure 2: Scatterplot displaying mean spatial bias (final original placement) for letters A-I including the best-fitting regression line.

## Experiment 2B: Letter control (A-Z)

## Method

The procedure for Experiment 2 B was identical to Experiment 1 and 2A but instead of Arabic numerals as stimuli, nine letters evenly spaced throughout the alphabet (A, D, G, J, M, P, S, V, Y) were presented, as we hypothesized participants could more easily distinguish between the ordinal position of "A"/"Y" than "A"/"I", for
example. Thirty-eight undergraduates participated for course credit. One participant was excluded from the statistical analyses as they did not complete all trials. One participant was excluded from the statistical analyses for poor accuracy.

## Results

The remaining participants $(N=36)$ had a mean accuracy of 13.65 pixels $(\mathrm{SD}=9.11)^{3}$. Consistent with the findings of Experiment 2A, participants' slopes were not significantly different from zero, $t(35)=-.14, p=.89$, confirming that letters do not generate a spatial association in this paradigm.


Figure 3: Scatterplot displaying mean spatial bias (final original placement) for letters evenly distributed in the alphabet including the best-fitting regression line.

## General Discussion

A primary goal of this study was to demonstrate an SNA in the absence of any sort of task demand. In the place-thenumber task, participants' responses are non-dichotomous, there is no left-right directional prime, and the value of the stimulus is not necessary to complete the task. Yet participants nevertheless exhibited an SNA consistent with past results. Our two control experiments demonstrate that this effect is specific to number and not other ordinal sequences such as letters.

Consistent with our hypotheses, these findings demonstrate that numbers - but not letters -automatically bias spatial attention in accordance with an MNL account. The fact that this effect occurs in the absence of relevant task demands is critical. Those who posit alternative accounts of SNAs often offer explanations that rest on demand characteristics of the tasks themselves, but there are

[^17]no such demands here. Therefore, we conclude that the automaticity of the spatial bias in this task sheds light on the nature of numerical representations - that they are inherently spatial unlike other ordinal sequences. Though other accounts exist, we believe these results provide strong evidence in favor of an MNL account of SNAs.

This interpretation is consistent with recent neural work which has explored the relation between space and number (Harvey et al., 2013). Harvey and colleagues (2013) found evidence of a topographic map for numerosity in the posterior superior parietal lobule, akin to topographic maps for sensorimotor systems (Udin \& Fawcett, 1988). Within this area, medial regions preferred small numerosities and lateral regions preferred large numerosities. Importantly, the location and numerosity preference of this topographic map was consistent across participants. These data support the growing body of evidence not only that number and space are deeply related in the mind, but, additionally, that numerical representations have an inherently spatial organization.

## Spatial representation versus spatial strategy

One question that arises from this interpretation is why others have demonstrated effects of ordinal sequences (e.g., Gevers, Reynovet, \& Fias, 2003; van Dijck \& Fias, 2011). To answer this question, we want to make a critical distinction. On the one hand, we might ask what things we can organize spatially; on the other hand, we might ask what things are inherently spatial. It is only in the context of the latter question that we argue numbers are unique. That we can organize ordinal sequences spatially should come as no surprise. People can organize items in any number of spatial arrangements and this type of spatialization has often been considered important for reasoning (Johnson-Laird, 1983). But the question that concerns the authors here is whether numbers are unique in the sense that they are automatically represented spatially in the mind.

This dichotomy is reminiscent of the so-called "dualprocess" model of SNAs (e.g., Ginsburg \& Gevers, 2015) which entail both long-term SNAs as well as spatialization in working memory. Abrahamse, van Dijck, and Fias (2016) have argued against this view, suggesting that their WM account is more parsimonious - that it "captures the complexity of the empirical database" without the need for long-term associations (p. 7). Indeed, if the mind were constructed for the sole purpose of representing number, then it may have evolved to do so in a parsimonious manner. Yet, Abrahamse and colleagues (2016) ignore the possibility that multiple mechanisms, some of them domaingeneral, may be at play. As we suggest, all ordinal sequences can be represented in space, but only numbers are automatically represented in this manner.

A working memory account of SNAs suggests that we have the propensity to organize sequences spatially in order to minimize the load of maintaining the sequence in the mind at once. This idea is reminiscent of the "method of loci" - a means of improving memory per spatial
visualization - is at least two millennia old (as in Cicero's De Oratore), but it has nothing to do with intrinsic characteristics of the representation. Thus, it becomes important to differentiate questions of spatial representation and spatial strategy. Previous tasks, some rife with potential task demands, failed to make a distinction between these two perspectives, yet have been interpreted as evidence of spatial representation. van Dijck and Fias (2011) have argued, partially on account of these task demands, that SNAs are merely transient mappings that occur in working memory - which, within our framework, falls under the purview of spatial strategy. In pursuing the latter issue of strategy, those who have espoused this WM perspective have overlooked the former, more crucial question of representation. That is, while we have argued that numbers are unique insofar as they are inherently spatial, van Dijck and Fias have succeeded only in showing that other sequences can, in certain contexts, be mapped spatially.

## Empirical horizons

What does the distinction between spatial representation and spatial strategy buy us? As a starting point, it establishes that numbers are in fact unique: they bias spatial attention automatically which suggests that their representation is inherently spatial in a way that other ordinal sequences are not. With this in mind, we are able to ask more nuanced questions about the underlying relationship between space and number. For example, why are numbers unique in this way? Do we come into the world with the propensity to represent numbers spatially, or is it learned? Perhaps more critically: what is the utility of a spatial-numerical mapping? For example, despite the seeming ubiquity and permanence of SNAs, it is unclear whether this spatial-numerical bias is related to math performance, with some studies reporting a positive relationship, some a negative relationship, and some no relationship at all (for review, see Cipora, Patro, \& Nuerk, 2015).

Not only does the place-the-number task play a part in raising these questions, it may also help to answer them. To further understand the phylogenetic and ontogenetic development of these associations, it is necessary to examine them in early childhood as number concepts are still being acquired. This has proven challenging, however, given that many SNA tasks are difficult to administer to children. The place-the-number task alleviates these concerns and might allow for the study of SNAs at a time in development when they have greater utility.

In sum, we have shown that numbers, but not letters, bias spatial attention in a manner that is consistent with an MNL hypothesis of SNAs. We argue that previous work which has posited alternative explanations to this account have been inadvertently answering a separate question - one about spatial strategy rather than spatial representation. Here, we have clarified the difference between these two accounts and suggested how the place-the-number task may be used to guide future research on the deep relationship between space and numbers in the mind.

## References

Abrahamse, E., van Dijck, J. -P., \& Fias, W. (2016). How does working memory enable number-induced spatial biases? Frontiers in Psychology, 7, 977.
Calibria, M. \& Rossetti, Y. (2005). Interference between number processing and line bisection: a methodology. Neuropsychologia, 43, 779-783.
Cheung, C. -N., Ayzenberg, V., Diamond, R. F. L., Yousif, S. R., \& Lourenco, S. F. (2015). Probing the mental number line: A between-task analysis of spatial-numerical associations. Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 357-362). Hillsdale, NJ: Lawrence Erlbaum Associates.
Cipora, K., Patro, K., \& Nuerk, H. -C. (2015). Are spatialnumerical associations a cornerstone for arithmetic learning? The lack of genuine correlations suggests no. Mind, Brain, and Education, 9, 190-206.
Cohen Kadosh, R. \& Henik, A. (2007). Can synesthesia research inform cognitive science? Trends in Cognitive Science, 11, 177-184.
Dehaene, S., Bossini, S. \& Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122, 371-96.
Fischer, M. H. (2001). Cognition in the bisection task. Trends in Cognitive Science, 5, 460-462.
Fischer, M. H., Castel, A. D., Dodd, M. D., \& Pratt, J. (2003). Perceiving numbers causes spatial shifts of attention. Nature Neuroscience, 6, 555-556.
Fischer. M. H., \& Shaki, S. (2014). Spatial associations in numerical cognition- From single digits to arithmetic. Quarterly Journal of Experimental Psychology, 67, 14611483.

Fitousi, D., Shaki, S., \& Algom, D. (2009). The role of parity, physical size, and magnitude in numerical cognition: The SNARC effect revisited. Perception \& Psychophysics, 71, 143-155.
Gevers, W., Reynvoet, B., \& Fias, W. (2003). The mental representation of ordinal sequences is spatially organized. Cognition, 97, 87-95.
Ginsberg, V. \& Gevers, W. (2015). Spatial coding of ordinal information in short- and long-term memory. Frontiers in Human Neuroscience, 9, 8.
Grossenbacher, P. G. \& Lovelace, C. T. (2001). Mechanisms of synesthesia: cognitive and physiological constraints. Trends in Cognitive Sciences, 5, 36-41.
Harvey, B. M., Klein, B. P., Petridou, N., \& Dumoulin, S. O. (2013). Topographic representation of numerosity in the human parietal cortex. Science, 341, 1123-1126.
Hoffmann, D., Hornung, C., Martin, R., \& Schiltz, C. (2013). Developing number-space associations: SNARC effects using a color discrimination task in 5-year-olds. Journal of Experimental Child Psychology, 116, 775-791.
Holmes, K. J., Ayzenberg, V. \& Lourenco, S. F. (2016). Gamble on gaze: Eye movements reflect the numerical value of blackjack hands. Psychonomic Bulletin \& Review, 23, 1974-1981.

Jewell, G. \& McCourt, M. E. (2000). Pseudoneglect: a review and meta-analysis of performance factors in line bisection tasks. Neuropsychologia, 38, 93-110.
Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press.
Loetscher, T., Bockisch, C., Nicholls, M. E. R., \& Brugger, P. (2010). Eye position predicts what number you have in mind. Current Biology, 20, R264-R265.
Loetscher, T., Schwarz, U., Schubiger, M., \& Brugger, P. (2008). Head turns bias the brain's internal random generator. Current Biology, 18, R60-R62.
Longo, M. R. \& Lourenco, S. F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. Neuropsychologia, 45, 14001407.

Marghetis, T., Kanwal, J., \& Bergen, B. K. (2013). Placing numbers in behavioral space: Activity-specific interactions between numbers and space with a single response button. Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 972977). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mattingley, J. B. (2009). Attention, automaticity, and awareness in synesthesia. Annals of the New York Academy of Sciences, 1156, 141-167.
Proctor, R. W. \& Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. Psychological Bulletin, 132, 416-442.
Ruiz Fernández, S., Rahona, J. J., Hervás, G., Vázquez, C., Ulrich, R. (2011). Number magnitude determines gaze direction: Spatial-numerical association in a free-choice task. Cortex, 47, 617-620.
Schwarz, W. \& Keus, I. M. (2004). Moving the eyes along the mental number line: comparing SNARC effects with the saccadic and manual responses. Perception \& Psychophysics, 66, 651-664.
Udin, S. B. \& Fawcett, J. W. (1988). Formation of topographic maps. Annual Review of Neuroscience, 11, 289-327.
van Dijck, J. P. \& Fias, W. (2011). A working memory account for spatial-numerical associations. Cognition, 119, 114-119.
van Dijck, J. P., Abrahamse, E. L., Acar, F., Ketels, B., \& Fias, W. (2014). A working memory account of the interaction between numbers and spatial attention. Quarterly Journal of Experimental Psychology, 67, 15001513.

Viarouge, A., Hubbard, E. M. \& Dehaene, S. (2014). The organization of spatial reference frames involved in the SNARC effect. Quarterly Journal of Experimental Psychology, 67, 1484-1499.
Zanolie, K. \& Pecher, D. (2014). Number-induced shifts in spatial attention: a replication study. Frontiers in Psychology, 5, 987.

# Highly Proficient Bilinguals Maintain Language-Specific Pragmatic Constraints on Pronouns: Evidence from Speech and Gesture 

Zeynep Azar (zeynep.azar@mpi.nl)<br>Radboud University, Center for Language Studies, Nijmegen<br>International Max Planck Research School for Psycholinguistics, Nijmegen<br>Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

Ad Backus (A.M.Backus@uvt.nl)
Tilburg University, Department of Culture Studies
Warendelaan, 5037 AB Tilburg, The Netherlands
Aslı Özyürek (asli.ozyurek@mpi.nl)
Radboud University, Center for Language Studies, Nijmegen
Max Planck Institute for Psycholinguistics, Nijmegen,
Donders Center for Cognition, Nijmegen
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands


#### Abstract

The use of subject pronouns by bilingual speakers using both a pro-drop and a non-pro-drop language (e.g. Spanish heritage speakers in the USA) is a well-studied topic in research on cross-linguistic influence in language contact situations. Previous studies looking at bilinguals with different proficiency levels have yielded conflicting results on whether there is transfer from the non-pro-drop patterns to the pro-drop language. Additionally, previous research has focused on speech patterns only. In this paper, we study the two modalities of language, speech and gesture, and ask whether and how they reveal cross-linguistic influence on the use of subject pronouns in discourse. We focus on elicited narratives from heritage speakers of Turkish in the Netherlands, in both Turkish (prodrop) and Dutch (non-pro-drop), as well as from monolingual control groups. The use of pronouns was not very common in monolingual Turkish narratives and was constrained by the pragmatic contexts, unlike in Dutch. Furthermore, Turkish pronouns were more likely to be accompanied by localized gestures than Dutch pronouns, presumably because pronouns in Turkish are pragmatically marked forms. We did not find any cross-linguistic influence in bilingual speech or gesture patterns, in line with studies (speech only) of highly proficient bilinguals. We therefore suggest that speech and gesture parallel each other not only in monolingual but also in bilingual production. Highly proficient heritage speakers who have been exposed to diverse linguistic and gestural patterns of each language from early on maintain monolingual patterns of pragmatic constraints on the use of pronouns multimodally.


Keywords: bilingualism; heritage speakers; gesture; crosslinguistic influence; pronoun; pragmatics; discourse

## Introduction

The use of subject pronouns by bilingual speakers of a prodrop (e.g. Spanish) and a non-prop language (e.g. English) in contact situations has been a commonly studied test case of cross-linguistic influence. Pro-drop languages habitually drop arguments and use overt pronouns mainly to mark
pragmatic information such as contrast and emphasis (e.g. Enç, 1986). The alternation between overt pronouns and dropped arguments is determined by discourse-pragmatics in those languages unlike in non-pro-drop languages such as English. Studies looking at heritage speakers who had lower proficiency in their pro-drop language than in their non-prodrop language found an increase in the frequency of pronouns or a loss of the pragmatic constraints on the use of pronouns in the pro-drop language (Paradis \& Navarro, 2003; Polinsky, 1995; Silva-Corvalan, 1994). On the other hand, studies looking at heritage speakers who are exposed to the pro-drop language more regularly and who have high proficiency in both languages found no cross-linguistic influence (CerrónPalomino, 2016; Keating, Jegerski \& van Patten, 2016; Montrul, 2004). Most studies, however, have focused on Spanish as a pro-drop language and English as a non-prodrop language in the United States.
In this paper, we look at language contact influence on subject pronouns studying Turkish heritage speakers in the Netherlands. Pronouns are less frequently used in pro-drop Turkish than in non-pro-drop Dutch, and they are pragmatically marked forms in Turkish (Enç, 1986) (similar to Spanish) but not in Dutch. Additionally, unlike previous studies in this domain, we examine not only patterns in the pro-drop language but also in the non-pro-drop language. We ask whether bilingual speakers maintain differences between Turkish and Dutch in terms of pragmatic constraints on the use of pronouns. Furthermore, as a novel contribution to research on cross-linguistic influence on subject pronouns, we extend our investigation to the visual modality of language, i.e. co-speech gestures. Studies of multimodal narratives have shown that speakers' gestures are sensitive to the amount of information encoded in speech. When referents are maintained in discourse, speakers not only reduce content of the referring expression by using pronouns or null forms,
but they also reduce the frequency of gestures related to referents (Azar \& Özyürek, 2015; Perniss \& Özyürek, 2015). Additionally, referents that are uniquely identified in speech are more likely to be accompanied by gestures (So, Kita \& Goldin-Meadow, 2009), suggesting gesture is tightly linked to speech. Whether this link extends to pragmatic marking of pronouns, that is whether languages that mark pronouns pragmatically in speech are more likely to mark them with gestures as well, has not been investigated so far. Furthermore, nothing is known about the multimodal nature of the cross-linguistic transfer in this domain.
As for gestures of bilingual speakers, in particular proficient L2 learners have been reported to show crosslinguistic influence in how frequently they gesture overall (So, 2010; see Cavicchio \& Kita, 2013 who found no crosslinguistic influence) and in their motion verb expressions (Brown \& Gullberg, 2008; Özçalışkan, 2016). Gestural transfer in the contexts of language contact and for differential pragmatic marking of pronouns on the other hand is an unexplored research topic. Thus, as a novel contribution to bilingualism research, we investigate whether heritage speakers who are highly proficient in their two languages maintain pragmatic constrains on the use of subject pronouns in speech and gesture or whether there is cross-linguistic influence in the two modalities.

An earlier study that looked at the use of subject pronouns by adult Turkish heritage speakers in the Netherlands (Doğruöz, 2007) found no cross-linguistic influence in the quantity of subject pronouns in informal interviews, though a few cases of the 1 st person pronoun were attested where monolinguals would not use a pronoun, e.g. in the immediately preverbal positions. We contribute to the literature on the use pronouns by Turkish-Dutch bilingual adult speakers in the Netherlands with a more controlled study (with respect to the discourse content) and in the context of narratives eliciting third-person references. Furthermore, we study not only Turkish narratives but also Dutch narratives produced by the same set of speakers. Finally, we take the multimodal aspects of reference production into account and investigate the use of gestures to mark subject referents by Turkish-Dutch bilinguals for the first time.

## Method

## Participants

20 Dutch monolingual speakers studying in Nijmegen (14 females; age mean $=21.5$ ), 20 Turkish monolingual speakers studying in Istanbul ( 17 females; age mean $=22.2$ ) and 20 bilingual speakers ( 14 females; age mean $=23.3$ ) studying in Nijmegen participated in our study in return for payment or course credits. Note that "monolingual" speakers in our study have some knowledge of English but they speak only one of the two languages that are of interest for this study.

Bilingual participants filled in a survey regarding their language history, current language use, and language proficiency in Turkish and Dutch. All bilingual speakers were born and raised in the Netherlands; their parents immigrated from Turkey to the Netherlands as young adults. Bilinguals were exposed primarily to Turkish at home until they started school at around the age of 4. They reported to mainly speak Dutch at school and mostly mix the two languages at home and among friends. Bilinguals rated their overall reading, speaking and comprehension proficiency higher in Dutch than in Turkish on a 5-point Likert scale (see Table 1). As a measure of oral fluency, we calculated articulation rate (number of syllables/ articulation time) (cf. De Jong, \& Wempe, 2009 for the script) for each participant using samples of around 30 seconds from the narratives we collected (the stimuli and procedure explained below). Bilinguals did not differ significantly from monolinguals in Turkish $t(38)=1.994, p=.053$ or in Dutch $t(38)=0.934, p=$ .356. Bilinguals' articulation rate was not significantly different between their Turkish and Dutch, either, $t(19)=$ 2.047, $p=.954$, suggesting they have similar levels of oral fluency in both languages (see Table 2).

Table 1
Self-rated Bilingual Proficiencies (1 = native; $5=$ beginner), Mean (SD)

|  | Speaking | Comprehension | Overall |
| :---: | :---: | :---: | :---: |
| Turkish | $2.50(1.32)$ | $2.25(0.79)$ | $2.40(1.27)$ |
| Dutch | $1.30(0.47)$ | $1.10(0.31)$ | $1.50(0.76)$ |

Table 2
Monolingual and Bilingual Speakers' Articulation Rates, Mean (SD)

|  | Monolingual | Bilingual |
| :---: | :---: | :---: |
| Turkish | $4.81(0.55)$ | $4.44(0.63)$ |
| Dutch | $4.62(0.71)$ | $4.42(0.57)$ |

## Stimuli

We used two short silent videos (cf. Azar, Backus \& Özyürek, 2016) to elicit narratives. Three characters were engaged in joint activities; cooking in one video and office work in the other. Figure 1 illustrates stills from each video.


Figure 1: Stills from the stimulus videos featuring kitchen (upper row) and office activities (bottom row)

## Procedure

Participants were invited to a quiet room in pairs and were assigned the role of either speaker or addressee (the assignment was random in monolingual sessions). The speaker watched the stimulus videos one by one on a computer screen. Once each video ended, the computer screen turned white and the speaker told the addressee what they had watched. The addressees were instructed that after each narrative, they could ask clarification questions and that they would be given two short written questions about each narrative. The purpose of this was to ensure that the speakers included enough details in their narratives and that the addressees paid attention. Once the instructions were given, the experimenter left the room and came back after each narrative with the questions for the addressee. The bilingual participants repeated the task once in Turkish with a Turkish monolingual addressee and once in Dutch with a Dutch monolingual addressee. The addressees were not confederates and there was at least two weeks between the two sessions. The order of the two videos was counterbalanced across participants. For bilinguals, the order of language was counterbalanced as well. All sessions were videotaped.

## Data Coding

We coded and analyzed speech from the speakers of each pair. We transcribed the video narratives using the standard orthography of each language and coded gestures with the frame-by-frame video annotation software ELAN (cf. Lausberg., \& Sloetjes, 2009).

Speech Coding We divided the narratives into clauses, utterances with a single subject argument and a single predicate. We coded only clauses with an animate subject argument (referring to the human characters in the stimulus videos) and marked whether the subject argument was maintained from the previous clause or not. We analyzed only clauses with maintained subjects since pronouns as reduced forms are used most frequently in those contexts (cf. Azar et al., 2016 for Turkish and Dutch). We further coded each maintained subject argument for one of the three possible referring expression types: noun phrase (NP), pronoun (third person and demonstrative pronouns) and null form. (1b) in Dutch and (2d, 2e) in Turkish illustrates clauses with maintained subjects. Subject arguments are underlined and subscripts index coreferentiality. Following Paradis and Navarro (2003), we coded Turkish subjects for pragmatic marking: contrast (disambiguation between two possible referents) or emphasis (highlighting information). Additionally, we also coded whether pronouns referring to subjects that are marked for emphasis were accompanied by the emphatic marker $d A$ 'also' (as in 2e). This clitic has been suggested to be a focus marker in Turkish (Enç, 1986) and has been shown to accompany pronouns when used for maintained subject arguments by monolingual Turkish speakers (Azar et. al., 2016). We did not code pragmatic marking for Dutch subjects because we expect Dutch
speakers to maintain subjects with pronouns as defaults forms rather than using pronouns to mark pragmatic information due to Dutch being a non-pro-drop language.
(1) a. Een meisje ${ }_{i}$ probeerde een pot open te maken. A irll $_{\mathrm{i}}$ tried to open a jar.
b. Die kreeg hem niet open.

That $_{\mathrm{i}}$ (the girl) did not open it.
(2) c. Ondan sonra $\underline{k} 1 Z k^{k}$ geliyor. Then irl $_{\mathrm{k}}$ is coming.
d. $\emptyset_{\mathrm{k}}$ çocuğa yardım ediyor.
null form
$\left(\right.$ She $_{\mathrm{k}}$ is helping the boy.
e. $\underline{\mathrm{O}}_{\mathrm{k}}$ da kağıtları diziyor.
pronoun

Gesture Coding We coded gestures temporally aligning with maintained subjects in speech, specifically with subject pronouns. We analyze gestures that anchored subjects in gesture space (i.e. index-finger and whole hand points). In Figure 2, the subject in (b) is maintained from (a) and marked with a pronoun in speech in Turkish and with an index-finger pointing gesture. The pronoun in speech is given in bold and the gesture and the character the pronoun refers to are highlighted in pictures.


Figure 2: Index-finger pointing gesture referring to the character in the video (highlighted) and temporally aligning with maintained subject pronoun in speech (in bold)

## Predictions

With regard to monolinguals, we expect speech and gesture to parallel each other in terms of the information they encode and therefore we expect cross-linguistic differences in the frequency of pronouns in speech and frequency of gestures marking pronominalized referents. In speech, we expect to find few pronouns in Turkish and in contexts where subject arguments are pragmatically marked for contrast or emphasis. Considering pronouns are marked forms in Turkish but not in Dutch, we predict that Turkish monolingual speakers will mark subject pronouns with gestures more than Dutch speakers. In terms of bilinguals we can anticipate the following scenarios for speech.

Influence of Dutch on bilingual Turkish: Based on studies that found cross-linguistic influence from non-pro drop English on pro-drop Spanish in subject pronouns of Spanish
heritage speakers in the States (e.g., Silva-Corvalan 1994), we expect bilinguals to have loosened the pragmatic constraints on the use of pronouns. Bilinguals in Turkish might use pronouns also when the subjects are not pragmatically marked and might accompany subjects that are marked for emphasis with the emphatic marker $\mathrm{d} A$ less frequently than monolinguals.

No cross-linguistic influence: Taking into account the literature which did not find cross-linguistic influence on subject pronouns for bilinguals with high proficiency in both languages (e.g. Cerrón-Palomino, 2016; Keating, Jegerski \& van Patten, 2016), we predict that bilinguals will maintain pragmatic constraints on the use of pronouns.

As for gestures, based on theories suggesting that speech and gesture parallel each other in production (Kita \& Özyürek, 2003; So et al., 2009), we expect the crosslinguistic influence on gestures to align with patterns of influence in speech. Alternatively, considering some L2 studies have found cross-linguistic transfer on gesture but not on speech (Özçalışkan, 2016), we may observe crosslinguistic influence on gesture modality only. Speakers may extend the pragmatic marking of pronouns with gestures from Turkish to Dutch and gesture with Dutch pronouns more frequently than Dutch monolinguals. Alternatively, bilinguals might loosen the pragmatic marking of gestures in Turkish as an influence from Dutch and gesture with pronouns less than monolinguals in Turkish.

## Analyses and Results

We performed arcsine transformation on ratio values for analyses though we report untransformed values. We analyzed the data using Linear Fixed Effects Models in IBM SPSS statistics 20 . We started with the simplest model with fixed effects only, and built more complex models by adding random intercepts. We compared each 'more complex' model to the previous simpler one in each step and in case of a significant difference we picked the model with the lower log-likelihood value. Bonferroni correction for multiple comparisons was applied for each model.

## Pronouns in Speech

We calculated the ratio of subject arguments referred to with a pronoun (subject pronouns) out of all maintained subject arguments in narratives per participant. We performed linear mixed model on subject pronouns with the following fixed effects: language type (Turkish vs. Dutch), language status (monolingual vs. bilingual) and the interaction between language type and language status. We started with the fixed effects only, and built more complex models by adding random intercepts and slopes for participants, language type and language status. The model that best described the variance of the data had random intercepts for participants and random intercepts for language type (Turkish or Dutch) varying by participants random slopes.
We found a significant effect for language type $F(1$, $66.657)=316.119, p<.001$ and for language status $F(1$, $45.204)=4.600, p=.037$ and a significant interaction
between the two $F(1,66.657)=4.174, p=0.045$. We further broke down the interaction and performed mixed linear models for Turkish and Dutch with language status (monolingual vs. bilingual) as fixed effect, following the same procedure as before. The model that best explained the variance for both Turkish and Dutch data was the simplest model with fixed effect language status. We did not find a main effect for language status $F(1,40)=0.852, p=.362$ for Turkish but for Dutch $F(1,40)=4.721 . p=.036$. Bilingual speakers used more pronouns in Dutch than monolinguals. Figure 3 illustrates the mean proportions of pronouns referring to subject arguments in monolingual and bilingual narratives by language.


Figure 3: Mean proportions of maintained subject pronouns in monolingual and bilingual narratives across Turkish and Dutch. The error bars represent standard errors of the mean.

Since we did not predict the findings in bilingual Dutch, we compared the use of the other two referring expressions we coded in speech, noun phrase (NP) and null form, across monolingual and bilingual Dutch to understand whether the higher use of pronouns by bilinguals could be driven by the lower use of one of the other two forms. We found that the bilinguals used null forms less frequently (although marginally) than monolinguals in Dutch $t(30.790)=-2.047$, $p=.049(M=0.132 ; 0.246$ respectively $)$.
Next, we looked at whether monolingual and bilingual speakers differed in the pragmatic marking of pronouns in Turkish. Out of all subjects that were encoded as pronouns, $82 \%$ in monolingual and $78 \%$ in bilingual narratives was marked for either emphasis or contrast. In total, there were 49 subject referents in monolingual Turkish and 44 subject referents in bilingual Turkish that were marked for emphasis and referred to with pronouns in speech. $88 \%$ of those pronouns in monolingual Turkish and $84 \%$ in bilingual Turkish was accompanied by the emphatic marker $\mathrm{d} A$. Thus, bilinguals were similar to monolinguals in Turkish in terms of the pragmatic constrains on the use of pronouns in speech.

## Pronouns Marked with Gestures

We calculated the ratio of gesturally marked subject pronouns out of all subject pronouns in speech per participant. We performed linear mixed model on gesturally marked subject pronouns with fixed effects language type, language status and the interaction of the two, following the same procedure as in our speech analyses. The model that best described the variance of the data had random intercepts for participants and language type (Turkish or Dutch) varying by participants random slopes. We found a significant effect for language type $F(1,69.358)=10.062, p$ $=.002$, showing Turkish speakers were more likely to mark pronouns with gestures than Dutch speakers. We did not find a significant effect for language status $F(1,92.697)=0.078$, $p=.781$ and no significant interaction between the fixed effects (language type and language status) $F(1,64.913)=$ $.001, p=.979$, suggesting bilinguals did not differ from monolinguals in terms of marking pronouns with gestures in either language. See Figure 4 for the mean values of gesturally marked pronouns.

Even though we found pronouns were more likely to be gesturally marked in Turkish than in Dutch, both in monolingual and bilingual speech, this could be due to an overall higher frequency of gestures in Turkish than in Dutch rather than an effect modulated by pragmatics. As a control, we looked at whether speakers per language group differed in how likely they are to gesturally mark a noun phrase (NP), the other overt referring expression type that we coded for speech. We performed mixed linear models on the ratio of gesturally marked NPs, following the same procedure as in our pronoun analyses. We did not find a main effect for language $F(1,56)=0.410, p=.525$, suggesting Turkish and Dutch speakers did not differ in how likely they were to mark NPs with gestures, contrary to what we found for pronouns. Turkish monolingual speakers gestured with NPs ( $M=0.33$, $S E=.083$ ) as often as Dutch monolinguals ( $M=0.28, S E=$ .126), suggesting the cross-linguistic difference we found for pronouns can be explained by the difference in the pragmatic status of pronouns across Turkish and Dutch and this effect is sensitive to the referring expression type used in speech. We did not find a main effect for language status $F(1,56)=$ $2.551, p=.116$ or an interaction of language and language status $F(1,56)=1.144, p=.289$. Bilinguals did not differ from monolinguals in Turkish $(M=0.42, S E=.120)$ or in Dutch ( $M=0.31, S E=.135$ ) in terms of how frequently they marked NPs with gestures.

## Discussion

In this study, we investigated whether there is cross-linguistic influence on the use of pronouns in narratives by heritage speakers who have high proficiency in both languages they speak. We specifically focused on the pragmatic constraints on the use of pronouns and we studied both speech and gestures for the first time in this domain looking at narratives


Figure 4: Mean proportions of gesturally marked maintained subject pronouns in monolingual and bilingual narratives across Turkish and Dutch. The error bars represent standard errors of the mean.
of Turkish heritage speakers in the Netherlands. We compared bilingual speech and gesture productions to those of monolinguals in Turkish and Dutch.

We showed that monolingual Turkish speakers used pronouns infrequently to maintain subject referents in narratives and mostly when the referents were pragmatic marked. Additionally, in line with our predictions, Turkish monolingual speakers were more likely to gesturally mark pronouns than Dutch monolingual speakers, suggesting linguistic forms that are pragmatically marked in speech (i.e. pronouns in pro-drop Turkish) are more likely to be marked with gestures as well.
Bilingual speakers did not differ from monolinguals in their pro-drop language, Turkish, in terms of how likely they were to use pronouns to maintain subject referents. Furthermore, we did not find any differences between monolingual and bilingual speakers in Turkish in terms of pragmatic constraints on the use of pronouns. Bilinguals used pronouns in Turkish to maintained referents that were marked for pragmatics, either for emphasis or contrast and they used the emphatic marker $\mathrm{d} A$ in similar ways to monolinguals. Our findings suggest that heritage speakers who were raised bilingually and who have high proficiency in both languages as well as using them daily, seem to have mastered the pragmatic constraints on the use of pronouns and to maintain them.
Although we did not expect any differences between monolingual and bilingual Dutch speech, we found that bilingual speakers used more pronouns and fewer null forms in Dutch than monolingual speakers. We suggest that bilingual speakers might have used coordinated clauses which allows null forms in Dutch less often than monolinguals and therefore dropped referents less often. However, since the use of null forms is not the main focus of our paper, we will not investigate this possibility further.
As for the visual modality, bilinguals maintained pronouns as marked forms in Turkish similar to monolingual speakers.

Bilinguals did not extend Turkish gestural marking to their Dutch narrative productions, either. Our findings are in line with those of Cavicchio \& Kita (2013) who looked at the overall gesture rate in L2 narratives, but differ from others which found cross-linguistic transfer of gesture with regard to the overall gesture rate (So, 2010) or motion verb expressions (Brown \& Gullberg, 2008; Özçalışkan, 2016).
To conclude, we show that speech and gesture parallel each other at the discourse-pragmatic level: Forms that are pragmatically marked in speech are more likely to be marked with gestures as well, extending the literature on crosslinguistic gestural differences in monolingual narratives.
Furthermore, we provide the first evidence that the parallel relation between speech and gesture (cf. So et al., 2009) extends to the domain of crosslinguistic influence in contact situations: When the influence is not evident in speech, it is not observable in gesture as well, at least with regard to pronoun use in the narratives of heritage speakers. Heritage speakers with high proficiency in both languages maintain pragmatic constraints on the use of subject pronouns, both in speech and gesture. Our findings therefore align with the studies that did not find cross-linguistic influence on the speech of highly proficient heritage speakers (e.g. CerrónPalomino, 2016; Keating, Jegerski \& van Patten, 2016). This suggests that proficiency in the heritage language may be an important determinant of the cross-linguistic influence on the use of pronouns in narratives in both modalities of language.
We suggest that studying bilingual gestures in addition to speech, especially in domains that show cross-linguistic influence in speech, will contribute to more complete theories of bilingualism. A better understanding of whether spoken and visual modalities undergo the same processes will provide valuable insight into the scope of cross-linguistic influence and language change beyond what we can learn from studies of speech alone.

## Acknowledgments

This research is funded by the Center for Language Studies, Radboud University Nijmegen, the Netherlands and partially by the Erasmus Staff training grant granted to the first author by the International Office Radboud University. The Max Planck Institute for Psycholinguistics also provided technical support. We thank Dr. Pamela Perniss for the stimulus video featuring cooking activities. We also thank Dr. Ayşe Caner and Dr. Nihan Ketrez for providing the location and participants for data collection in Istanbul, Turkey.

## References

Azar, Z. \& Özyürek, A. (2015). Discourse management: Reference tracking in speech and gesture in Turkish narratives. Dutch Journal of Applied Linguistics, 4(2), 222240.

Azar, Z., Backus, A., \& Özyürek, A. (2016). Pragmatic relativity: Gender and context affect the use of personal pronouns in discourse differentially across languages. In A. Papafragou, D. Grodner, D. Mirman, \& J. Trueswell (Eds.), Proceedings of the 38th Annual Meeting of the Cognitive

Science Society (pp. 1295-1300). Austin, TX: Cognitive Science Society.
Brown, A. \& Gullberg, M. (2008). Bidirectional crosslinguistic influence in L1-L2 encoding of Manner in speech and gesture: A study of Japanese speakers of English. Studies in Second Language Acquisition, 30, 225251.

Cavicchio, F. \& Kita, S. (2013). English/Italian bilinguals switch gesture parameters when they switch languages. Proceedings of TiGeR 2013 (pp. 305-309).
Cerrón-Palomino, A. (2016). (No) English interference on U.S. Southwest Spanish? A look at variable subject expression in Phoenix Spanish-English bilinguals. Sociolinguistic Studies, 10(3), 383-408.
Doğruöz, A. S. (2007). Synchronic variation and diachronic change in Dutch Turkish: A corpus based analysis. Doctoral Dissertation, Tilburg University, the Netherlands.
Enç, Mürvet. (1986). Topic switching and pronominal subjects in Turkish. In D.I. Slobin and K. Zimmer (Eds.), Studies in Turkish linguistics (pp. 195-209). Amsterdam: John Benjamins.
de Jong, N.H. \& Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods \& Instrumentation, 41 (2), 385-390.
Keating, G. D., Jegerski, J. \& van Patten, B. (2016). Online processing of subject pronouns in monolingual and heritage bilingual speakers of Mexican Spanish. Bilingualism: Language and Cognition, 19 (1), 36-49.
Kita, S., \& Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48(1), 16-32.
Lausberg, H. \& Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, Instruments, \& Computers, 41(3), 841-849.
Montrul, S. (2004). Subject and object expression in Spanish heritage speakers: A case of morphosyntactic convergence. Bilingualism: Language and Cognition, 7 (2), 125-142.
Paradis, J. \& S. Navarro (2003). Subject realization and crosslinguistic interference in the bilingual acquisition of Spanish and English: What is the role of input? Journal of Child Language, 30(2), 371-393.
Perniss, P. M., \& Özyürek, A. (2015). Visible cohesion: A comparison of reference tracking in sign, speech, and cospeech gesture. Topics in Cognitive Science, 7(1), 36-60.
Polinsky, M. (I995). Cross-linguistic parallels in language loss. Southwest Journal of Linguistics, 14 (1\&2), 87-123.
Özçalışkan, Ş. (2016). Do gestures follow speech in bilinguals' description of motion? Bilingualism: Language and Cognition, 19 (3), 2016, 644-653.
Silva-Corvalan, C. (1994). Language contact and change. Oxford: Clarendon Press.
So, W. C., Kita, S. \& Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand in hand. Cognitive Science, 33, 115-125.

# Rise and fall of conflicting intuitions during reasoning 

Bence Bago (bencebago@ed-alumni.net)<br>Paris Descartes University - Sorbonne Paris Cité<br>LaPsyDE (CNRS unit 8240)

Wim De Neys (wim.de-neys@ parisdescartes.fr)<br>Paris Descartes University - Sorbonne Paris Cité<br>LaPsyDE (CNRS unit 8240)


#### Abstract

Recent dual process models proposed that the strength of competing intuitions determines reasoning performance. A key challenge at this point is to search for boundary conditions; identify cases in which the strength of different intuitions will be weaker/stronger. Therefore, we ran two studies with the two-response paradigm in which people are asked to give two answers to a given reasoning problem. We adopted base-rate problems in which base rate and stereotypic information can cue conflicting intuitions. By manipulating the information presentation order, we aimed to manipulate their saliency; and by that, indirectly the activation strength of the intuitions. Contrary to our expectation, we observed that the order manipulation had opposite effects in the initial and final response stages. We explain these results by taking into account that the strength of intuitions is not constant but changes over time; they have a peak, a growth, and a decay rate.


Keywords: reasoning; conflict detection; hybrid dual process model

## Introduction

Decades of research in thinking and reasoning has revealed that people are usually subject to errors. Consider for example the following situation:
"There is a party with 1000 people. Jo is a randomly chosen participant from the party. We know that Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer. We also know that 900 people attending the party are women. What is most likely: Is Jo a man or a woman?"

This is a so-called base rate problem. Based on the "normative" ${ }^{1}$ principle that a randomly drawn individual

[^18]will more likely come from the largest group, one should favor the conclusion that Jo is a woman. However, the majority of people tend to err on this problem by going with the presented stereotype (which cues that Jo is a man). Dual process theories provide an explanation for general thinking bias on problems such as the base rate task. They distinguish two types of processing, Type 1 and Type 2. One should note that there are many dual process theories, but in this study, we will focus on the most influential dual process theory, the default-interventionist theory. Type 1 processes (also referred to as intuitive processes) are thought to be completely autonomous, while Type 2 processes (also referred to as analytic processes) are more controlled. Type 1 processing generates responses cued by stereotypes or common beliefs; relying on this intuitive, initial response is what makes people biased in such situations. After Type 1 processing produced a response, in some cases, Type 2 processing gets engaged; this type of processing has the ability to override and correct the response generated by Type 1 processing. In general, it is assumed that Type 2 processing has the ability to generate responses based on logic or probabilities, while Type 1 processing has not been considered to be able to handle information such as logical properties of the task, or probabilities (Kahneman, 2011; Stanovich \& Evans, 2013).

However, recently, conflict detection studies (De Neys, 2012 , 2014) indicated that the assumption that Type 1 processing is not able to handle probabilistic or logical information might not hold. These studies showed that even biased reasoners were able to detect the conflict between intuitive "heuristic" cues (e.g., stereotypes) and "normative" logical and probabilistic principles (e.g., base rate probabilities). These studies usually contrast conflict and no-conflict reasoning problems. In conflict problems, heuristic processing and normative principles cue different responses as in the base rate problem above. In a no-conflict problem normative principles and heuristic processing cues the same response; for example, imagine that the abovepresented base rate problem would state that there are 900 men and 100 women. In this case, both the stereotype and
simplicity, we stick to the traditional labeling. In the same vein, we use the term "logical" as a general header to refer both to standard logic and probability theory.
base rate probabilities would cue the same response (that Jo is a man). In conflict problems, studies showed that even incorrect reasoners (compared to correct reasoners in noconflict problems) showed elevated response times, decreased post-decision confidence, and higher activation in brain areas mediating conflict detection across a range of tasks (for review see De Neys, 2012).

These results made some authors suggest that there occurs some kind of elementary processing of logical/probabilistic information even during Type 1 processing. De Neys (2012) argues that conflict detection happens as a result of two conflicting Type 1 outputs, generated by two kinds of intuitions. He argues that one of these intuitions is based on stereotypes or common beliefs (heuristic intuition) the other one is based on logico-mathematical principles (logical intuition).

Recently, Bago and De Neys (2017a) went a step further and argued that people are not just able to detect the conflict intuitively but some of them are able to give the logically correct response intuitively. Our so-called hybrid dual process model argues that the two different intuitions differ in activation strength (or "salience"), and the actual intuitive response that the person provides will be the one which gained more strength. The relative difference between the strength of the heuristic and logical intuitions defines how pronounced the conflict is; the smaller the relative difference, the more pronounced the conflict will be; the larger the relative difference, the less pronounced it will be.

A key question at this point is to search for boundary conditions; identify cases in which the strength of different intuitions will be more or less pronounced. One way to do so is to manipulate the presentation order of base rate information and stereotypes. Let us explain why. In a previous study, Pennycook, Fugelsang, and Koehler (2015) argued that a "given piece of information is at its most salient just prior to judgement" (Pennycook et al., 2015, p. 57). Pennycook et al. (2015) further argued that this would mean that base rate information is most salient if presented right before the decision was made (after the stereotypical description had been presented). The authors observed that presenting the base rate information at the end of the problem indeed boosted participants' accuracy compared to the condition when it was presented first. To help us explain these results, one could operationalize saliency as the strength of a given intuitive response. Hence, whatever information was presented later, would be the more salient, therefore the intuition cued by this piece of information would be the stronger one.

In this study, we wanted to test the robustness of these findings - will we get the same effects after purely intuitive Type 1 processing? Thus, to test this question, one needs to use a research design which is able to separately measure intuitive Type 1 responses from analytic Type 2 responses. For this reason, we used the two response paradigm (Thompson, Prowse Turner, \& Pennycook, 2011). In the two response paradigm, participants are presented with the same item twice. First, they are asked to give a very quick
intuitive, initial response. Then, the same task is presented again and now they can take as much time as they want before providing their final response. One also needs to be sure that the initial response is truly intuitive; we achieved this by applying a strict response deadline ( 3 seconds) and a secondary task that burdens reasoner's (executive) cognitive capacity during the initial response. With these manipulations we can experimentally knock out Type 2 processing during the initial responding (Bago \& De Neys, 2017a).

Our hypothesis was that if presentation order indeed affects the strength of an intuition, we should observe the same effect after purely intuitive processing as has been observed previously after deliberative thinking. That is, if base rates are presented last, the strength of the base rate intuition should be higher, and therefore more correct responses should be observed both at the initial and final response stages.

## Study 1

## Method

## Participants

In total, 149 participants took part in the experiment ( 86 female, $M=39.3$ year, $S D=12.7$ year). Participants were recruited online, via Crowdflower, and received $\$ 0.25$ for their participation. Subjects were randomly assigned to one of the two conditions. Note that data in the S-BR condition were taken from the study of Bago \& De Neys, (2017b). A total of $44.5 \%$ of participants reported having high school as highest completed educational level, while $52.1 \%$ reported that they have a post-secondary educational degree (3.4\% reported less than high school).

## Materials

Reasoning task. Participants solved a total of eight baserate problems. All problems were taken from Pennycook, Cheyne, Barr, Koehler, and Fugelsang (2014). Participants always received a description of the composition of a sample (e.g., "This study contained I.T engineers and professional boxers"), base rate information (e.g., "There were 995 engineers and 5 professional boxers") and a description that was designed to cue a stereotypical association (e.g. "This person is strong"). Participants' task was to indicate to which group the person most likely belonged.

The problem presentation format we used in this research was based on Pennycook et al.'s (2014) rapid-response paradigm. In this paradigm, the base rates and descriptive information are presented serially and the amount of text that is presented on screen is minimized. Pennycook et al. introduced the paradigm to minimize the influence of reading times and get a purer and less noisy measure of reasoning time per se. Participants received 3 pieces of information in a given trial. First, the names of the two
groups in the sample (e.g., "This study contains clowns and accountants"). This sentence stayed on the screen and was always presented first. Participants were presented with stereotypical descriptive information (e.g., Person 'L' is funny) as well. The descriptive information specified a neutral name ('Person L') and a single word personality trait (e.g., "strong" or "funny") that was designed to trigger the stereotypical association. Participants also received the base rate probabilities. In this experiment, we manipulated the presentation order of the base rate probabilities and stereotypes. So, for one group the base rates were presented first (BR-S), for the other group, the base rates were presented last, after the stereotype (S-BR). Presentation order was manipulated between-subject. The following illustrates the full problem format in the S-BR condition:

## This study contains clowns and accountants. <br> Person 'L' is funny. <br> There are 995 clowns and 5 accountants. <br> Is Person 'L' more likely to be: <br> o A clown <br> o An accountant

Half of the presented problems were conflict items and the other half were no-conflict items. In no-conflict items, the base rate probabilities and the stereotypic information cued the same response. In conflict items, the stereotypic information and the base rate probabilities cued different responses. Three kinds of base rates were used: 997/3, 996/4, 995/5.

Each problem started with the presentation of a fixation cross for 1000 ms . After the fixation cross disappeared, the sentence which specified the two groups appeared for 2000 ms . Then the first information appeared, for another 2000 ms , while the first sentence remained on the screen. Finally, the last information appeared together with the question and two response alternatives. Note that we presented the last information and question together (rather than presenting the last information for 2000 ms first) to minimize the possibility that some participants would start solving the problem during the presentation of the last part of the problem. Once all the parts were presented, participants were able to select their answer by clicking on it. The position of the correct answer alternative (i.e., first or second response option) was randomly determined for each item. The eight items were presented in random order.

Confidence in the correctness of the response was recorded after the initial and the final response stages by asking participants to indicate their confidence level on a scale ranging from $0 \%$ to $100 \%$.
Cognitive load task. We used a concurrent load task - the dot memorization task - to burden participants' executive cognitive resources while they were solving the reasoning tasks. The idea behind the load manipulation is straightforward. One of the defining features of Type 2 processing is that it requires executive (working memory) resources (e.g., Evans \& Stanovich, 2013; Kahneman, 2011).

Hence, if we burden participants' cognitive resources with a secondary load task while they are solving the reasoning problems, we reduce the possibility that they can engage in Type 2 thinking (De Neys, 2006).
In every trial, after the fixation cross disappeared, participants were shown a matrix in which 4 dots were presented in a complex interspersed pattern in a $3 \times 3$ grid for 2000 ms . Participants were instructed to memorize the pattern. Previous studies established that this demanding secondary task successfully burdens executive resources during reasoning (De Neys, 2006). After the matrix disappeared, the reasoning problem was presented as described above and participants had to give their first response. Then participants were shown four matrices with different dot patterns and they had to select the correct, to-be-memorized matrix. Participants were given feedback as to whether they recalled the correct matrix or not. Subsequently, the problem was presented again and participants selected their final response and response confidence. Hence, no load was imposed during the second, final response stage. All trials on which an incorrect matrix was selected ( $9.5 \%$ of trials) were removed from the analysis.
Response deadline. In order to minimize the possibility of Type 2 engagement during the initial response, we used a strict response deadline ( 3000 milliseconds), based on a reading pre-test (see Bago \& De Neys, 2017a). 1000 ms before the deadline, the background turned yellow to alert the participants to the approaching deadline. If participants did not select an answer within 3000 ms they got feedback to remind them that they had not answered within the deadline and they were told to make sure to respond faster on subsequent trials. Obviously, there was no response deadline on the final response, but only on the initial response. All trials where participants did not manage to provide a response were excluded from the analysis (8.7\% of trials).
Procedure. The experiment was run online. People were clearly instructed that we were interested in their first, initial response to the problem. Instructions stressed that it was important to give the initial response as fast as possible and that participants could afterwards take additional time to reflect on their answer. After the instructions, participants were presented with practice problems to familiarize them with the procedure. At the end of the experiment, demographic questions were collected.

## Results

Our main interest concerns the response accuracy analysis. Table 1 gives an overview of the findings. As one can see, we replicated the findings of Pennycook et al. (2015) at the final response stage for the conflict problems: Final accuracies on conflict problems are higher (41.6\%) when the base rates are presented last vs. first ( $24.3 \%$ ). However, contrary to our expectations, we do not observe the same effect at the initial response stage; there is even a trend towards fewer correct responses in the "base rates
last" S-BR condition ( $29.7 \%$ ) vs BR-S ( $31.8 \%$ ) condition. Indeed, the final conflict response accuracies in the S-BR condition were higher than the initial conflict response accuracies, whereas the reverse trend can be observed in the BR-S condition. In other words, the condition with the highest final accuracy (S-BR) was the one with the lowest initial accuracy, while the condition with the lowest final accuracy (BR-S) was the one with the highest initial accuracy.

Finally, as expected, note that accuracies on the noconflict problems were always very high. Not surprisingly, in the absence of conflict, both the stereotype and base-rates can cue the correct response whatever order the information is presented in.

Table 1. Percentage of correct initial and final responses for conflict and no-conflict items in both order conditions.

|  | Response | Order |  |
| :--- | :--- | :--- | :---: |
|  |  | S-BR | BR-S |
| Conflict | Initial | $29.7 \%$ | $31.8 \%$ |
|  | Final | $41.6 \%$ | $24.3 \%$ |
| No-conflict | Initial | $93.4 \%$ | $90.1 \%$ |
|  | Final | $93.7 \%$ | $91.4 \%$ |

Note. $\mathrm{S}-\mathrm{BR}=$ base rates last/ BR-S = base rates first.
We used mixed effect logistic regression (logit) models to analyze the data and entered accuracy as a dependent variable. The order manipulation (S-BR/BR-S), response number (initial/final response), and their interaction were entered as predictors into the model. We also accounted for the random effect (random intercept) of subjects. We concentrated our analysis on the critical conflict problems. Only the interaction improved model fit significantly $\chi^{2}$ (5) $=20.18, p<0.0001, b=1.94$, but not the main effect of order $\chi^{2}(3)=0.19, p=0.66$ or response number $\chi^{2}(4)=$ $0.38, p=0.54$. These results confirm our visual inspection that order affects initial and final accuracies differently.

Table 2. Frequency of each direction of change category (number of trials) for conflict items in both conditions.

| Direction of change | Order |  |
| :--- | :--- | :--- |
|  | S-BR | BR-S |
| 11 | $26.7 \%(54)$ | $19.7 \%(47)$ |
| 00 | $55.4 \%(112)$ | $63.6 \%(152)$ |
| 10 | $3 \%(6)$ | $12.1 \%(29)$ |
| 01 | $14.9 \%(30)$ | $4.6 \%(11)$ |

Note. S-BR = base rates last/ BR-S = base rates first.
For completeness, one could also test the direction of change in every trial (Bago \& De Neys, 2017a). Specifically, people can give correct or incorrect responses on both response stages; this means that one could give two correct (" 11 "), two incorrect ("00"), an initial correct but final incorrect ("10"), or an initial incorrect but final correct (" 01 ") response. The results of the direction of change analysis are summarized in Table 2. In both order
conditions, the most frequent categories were the " 00 " and " 11 " cases. In line with previous observations (Bago \& De Neys, 2017a; Thompson et al., 2011) people rarely changed their initial response (i.e., taken together the " 10 " and " 01 " cases account for $16 \%-18 \%$ of the trials). Interestingly, the direction in which people changed also tended to be reversed; in the S-BR condition most people who did change, changed from an incorrect to correct response (i.e., " 01 " category, $14.9 \%$ vs " 10 " category, $3 \%$ ). However, in the BR-S condition most people who changed their initial response, changed it to an incorrect response (i.e., " 10 " category dominates with $12.1 \%$ vs $4.6 \%$ for the " 01 " category). Hence, this fits with the overall trend towards the higher likelihood of an initial incorrect and final correct response when the base rates are presented last. A Chisquare test of independence revealed that the distribution of the direction of change categories in the two order conditions significantly differed from each other $\chi^{2}(3)=$ $27.56, p<0.0001$.

## Discussion

Contrary to our expectations, we did not observe the expected accuracy effect at the initial response stage; we only observed it in the final response stage. However, we wanted to be sure that the findings were robust before drawing any conclusions. Note that Pennycook et al. (2015) already observed that their order findings were robust against manipulations of the extremity of the base rates. That is, they found the same order effect on (final) accuracies when they used so-called "moderate" base rates (e.g., base rate probabilities of 700 men and 300 women) instead of the "extreme" base rates (e.g., e.g. base rate probabilities of 995 men and 5 women) that were adopted in our (and their) Study 1. In Study 2 we therefore also adopted the moderate base-rates and examined whether the unexpected reversal of the order effect on initial, intuitive responses would still be observed.

## Study 2

## Method

## Participants

In total, 162 participants took part in the experiment (98 female, $M=40.2$ year, $S D=14.6$ year). Participants were recruited online, via Crowdflower, and received $\$ 0.25$ for their participation. Subjects were randomly assigned to one of the two conditions. Note that data in the S-BR condition were taken from the study of Bago and De Neys (2017b). A total of $46.3 \%$ of participants reported having high school as highest completed educational level, while $52.5 \%$ reported that they have a post-secondary educational degree (1.3\% reported less than high school).

## Materials

Reasoning task. The identical experimental design was used as in Study 1. The only difference is that we used moderate base rates instead of extreme ones, namely $700 / 300,710 / 290$ and $720 / 280$. In $16.7 \%$ of the trials participants did not provide the correct response for the dot matrix task, and in $10.5 \%$ of the trials, participants did not manage to produce an initial response within the deadline. These trials were excluded from further analysis. Overall, $24.6 \%$ of the trials were excluded and 977 were analyzed.

## Results and discussion

Table 3 summarizes the accuracy results. As the table indicates, no-conflict response accuracies are again very high overall and we also replicated the conflict problem pattern we observed in Study 1: As Pennycook et al. (2015) found, presenting the base rates last led to increased accuracy on the final response. However, as in Study 1, the opposite trend was observed in the initial response. We also observe again that there were more initial than final incorrect response in the BR-S condition, whereas the opposite trend is observed in the S-BR condition. Statistical analysis on the conflict problems confirmed our visual inspection; neither presentation order $\chi^{2}(3)=0.04, p=0.84$, nor response number improved model fit significantly, $\chi^{2}$ (4) $=0.05, p=0.83$, only their interaction $\operatorname{did} \chi^{2}(5)=9.73, p=$ $0.0018, b=1.4$.

Table 3. Percentage of correct initial and final responses for conflict and no-conflict items in both order conditions.

|  | Response | Order |  |
| :--- | :--- | :--- | :---: |
|  |  | S-BR | BR-S |
| Conflict | Initial | $16.4 \%$ | $18.3 \%$ |
|  | Final | $23 \%$ | $13.2 \%$ |
| No-conflict | Initial | $90.9 \%$ | $90.9 \%$ |
|  | Final | $90 \%$ | $92.5 \%$ |

Note. S-BR = base rates last/ BR-S = base rates first
Table 4 summarizes the results of the direction of change results for conflict items. Here too we observe the same trend as in Study 1. Among the few people who changed their response, the direction in which they changed are reversed as a function of presentation order; in the S-BR condition most people who did change, changed from an incorrect to correct response. But in the BR-S condition, more people changed to an incorrect response. A Chisquare test of independence revealed that the distribution of the direction of change categories in the two order conditions significantly differed from each other $\chi^{2}(3)=$ $18.22, p=0.0004$.

Table 4. Frequency of each direction of change category (number of trials) for conflict items in both order conditions.

| Direction of change | Order |  |
| :--- | :--- | :--- |
|  | S-BR | BR-S |
| 11 | $14.2 \%(32)$ | $8.2 \%(21)$ |
| 00 | $74.8 \%(169)$ | $76.7 \%(197)$ |
| 10 | $2.2 \%(5)$ | $10.1 \%(26)$ |
| 01 | $8.8 \%(20)$ | $5.5 \%(13)$ |

Note. S-BR = base rates last/ BR-S = base rates first

## General Discussion

In this paper, we tested whether manipulating the presentation order of the base rates and stereotypes had the same effect after purely intuitive processing (i.e., initial response) as had been observed previously after deliberative thinking (i.e., final response). In two studies, we replicated the findings of Pennycook et al. (2015) at the final response stage: Final accuracies on conflict problems were higher when the base rates were presented last. However, contrary to our expectations, in both studies this effect consistently reversed at the initial response stage. Why is this the case? We believe that these results draw attention to a simple but somewhat neglected issue in reasoning models, namely that intuitive responses are not generated instantly at full strength.

The hybrid dual process model that we presented in the introduction (e.g., Bago \& De Neys, 2017a) argues that reasoning performance in the initial response stage is determined by the strength of different intuitions, for example. The implicit assumption here is that the strength of these intuitions is "instant" and "constant". That is, the idea is that the intuition is readily generated with full force and maintains this strength level.

However, upon some further reflection, this assumption might be quite naïve. It is reasonable to assume that even a quickly generated intuition needs some time to reach its peak. Keeping this feature in mind might suffice to explain the current findings. Have a look at Figure 1. In this illustration, the strength of two intuitions $\left(I_{1}, I_{2}\right)$ change over time - they have a peak, a growth and a decay rate. The $y$ axis represents the strength, the x -axis represents time, while T 1 and T 2 represent the time of initial and final response, respectively.
$\mathrm{I}_{1}$ and $\mathrm{I}_{2}$ will start gaining strength when the relevant cue is presented (in the $S-B R$ condition $I_{1}$ is the heuristic intuition cued by the presentation of the stereotype, and $\mathrm{I}_{2}$ is the logical intuition cued by the base rate information). So, in the S-BR condition, the stereotype is presented first. When the stereotype is presented, the intuition $\left(\mathrm{I}_{1}\right)$ cued by it starts gaining strength. Subsequently, the presentation of the base rate information cues the logical intuition $\left(\mathrm{I}_{2}\right)$ and its strength will also start rising. Both intuitions grow until they reach their peak. At T1, $\mathrm{I}_{1}$ has already reached its peak, and is stronger than $I_{2}$ (which has not reached its peak yet); as a result, $\mathrm{I}_{1}$ will be the initial response. But after T1, the
strength of $I_{1}$ starts decaying, while the strength of $\mathrm{I}_{2}$ is still increasing, and it reaches its peak at T 2 . At $\mathrm{T} 2, \mathrm{I}_{2}$ will be the stronger intuition, so people will more likely pick I2 as their final response. Hence, the mere growth and decay of an intuition - or it's "rise and fall" as we labelled it in the title implies that (ceteris paribus) the most recently cued intuition will be weaker earlier on in the reasoning process (e.g., initial response stage) and dominate later in the reasoning process (e.g., final response).

Clearly, we have presented and illustrated the most generic and general case in which two intuitions have the same peak level, growth, and decay rate. Obviously, these features might vary. One intuition might have a higher peak than the other, or a faster/slower growth/decay than the other. In addition, we believe that deliberation might also modulate the strength level. For example, one can imagine that one functional consequence of deliberation might be to boost or sustain the peak activation level of one intuition and decrease activation of the other. These more specific features have to be tested and validated in future studies. For example, one could try to test the role of deliberation by examining the impact of cognitive load on the presentation order findings in the second response stage. However, in all these more specific cases the general principle holds that we have to keep in mind that intuitions are not necessarily generated instantly but "rise and fall"; we need to consider their growth and decay. We believe this should motivate further research in the area by trying to determine what the growth and decay functions look like exactly.


Figure 1. Illustration of how the strength of intuitions might change over time. The $y$-axis represents the activation strength while the $x$-axis represents time. $I_{1}$ and $I_{2}$ represent the two cued intuitions. Note that in the BR-S condition $\mathrm{I}_{1}$ is the logical intuition cued by the base rate probabilities, while $\mathrm{I}_{2}$ is the heuristic intuition cued by the stereotypes. Consequently, in the S-BR condition, $\mathrm{I}_{1}$ is the heuristic and $\mathrm{I}_{2}$ is the logical intuition. $\mathrm{T}_{1}$ and $\mathrm{T}_{2}$ represent the time of initial and final response, respectively.

## Acknowledgement

Bence Bago was supported by a doctoral fellowship from Ecole des Neurosciences de Paris Ile-de-France. This research was also supported by a research grant
(DIAGNOR, ANR-16-CE28-0010-01) from the Agence National de la Recherche.

## References

Bago, B., \& De Neys, W. (2017a). Fast logic?: Examining the time course assumption of dual process theory. Cognition, 158, 90-109.
Bago, B., \& De Neys, W. (2017b). Examining the hybrid dual process model. Manuscript in preparation.
De Neys, W. (2006). Automatic-heuristic and executiveanalytic processing during reasoning: Chronometric and dual-task considerations. The Quarterly Journal of Experimental Psychology, 59(6), 1070-1100.
De Neys, W. (2012). Bias and conflict a case for logical intuitions. Perspectives on Psychological Science, 7(1), 28-38.
De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. Thinking \& Reasoning, 20(2), 169-187.
De Neys, W., \& Schaeken, W. (2007). When people are more logical under cognitive load. Experimental Psychology (formerly Zeitschrift Für Experimentelle Psychologie), 54(2), 128-133.
Evans, J. S. B., \& Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. Perspectives on Psychological Science, 8(3), 223-241.
Franssens, S., \& De Neys, W. (2009). The effortless nature of conflict detection during thinking. Thinking \& Reasoning, 15(2), 105-128.
Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., \& Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. Memory \& Cognition, 42(1), 1-10.
Pennycook, G., Fugelsang, J. A., \& Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive Psychology, 80, 34-72.
Stanovich, K. E., \& West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. Behavioral and Brain Sciences, 23(5), 645-665.
Thompson, V. A., Prowse Turner, J. A., \& Pennycook, G. (2011). Intuition, reason, and metacognition. Cognitive Psychology, 63(3), 107-140.

# The Refugees' Dilemma: not all deontological moral choices are of the same kind 

Albert Barque-Duran and Emmanuel M. Pothos<br>Department of Psychology, City, University of London, London EC1V 0HB, UK. albert.barque-duran@city.ac.uk


#### Abstract

The focus of the present work concerns the nature of deontological decisions. We test the hypothesis that it is possible to specify deontological moral choices based on an unemotional rule, norm or principle and that such moral choices can be distinguished from emotion-driven ones. Using a novel paradigm for moral choice that we call The Refugees' Dilemma, we provide evidence for such a rulebased route to moral choice. We show that participants with high scores in a Cognitive Reflection Test (CRT) were more likely to adopt utilitarian or rule-based responses, as opposed to emotional ones. We also found that rule-based respondents reported the highest average psychological distance, more so that even utilitarian respondents. These findings show how emotional and rule-based influences can be separated with the appropriate scenario and challenges the approach of assuming both influences can be combined into a single deontological route in dual-process models.


Keywords: Dual-Process Models; Deontological Ethics; Moral Judgment; Moral Psychology.

## Introduction

Moral decision-making is at the heart of modern democracies. Therefore, understanding the principles underpinning moral judgment is fundamentally important. Consider the recent refugee crisis. The number of forcibly displaced people worldwide reached 59.5 million at the end of 2014, the highest level since World War II. Of these 59.5 million, 19.5 million were refugees, and 1.8 million were asylum-seekers. How do individuals in destination countries form opinions regarding refugees and asylum seekers? At the very least, understanding the influences shaping moral choice should provide individuals with better insight (and possibly control) into their ultimate determinations.

The established theory is that moral decisions are driven by two complementary influences (Singer, 1991; Chaiken \& Trope, 1999). Such so-called dual-process theories contrast utilitarian responses, resulting from controlled cognitive processes, with non-utilitarian (considered deontological) responses, assumed to be driven by automatic/intuitive emotional processes (Greene et al., 2001; Greene \& Haidt, 2002; Greene et al., 2004; Koenigs et al., 2007). Utilitarian/consequentialist judgments are aimed at maximizing benefits and minimizing costs across affected individuals (Mill, 1861/1998), while the deontological perspective (Kant, 1785/1959) emphasizes rights and duties.

We focus on deontological decisions. It seems there is a fundamental inconsistency in current understanding of such
decisions. On the one hand, they are meant to be based on some rule, principle or norm. On the other hand, the deontological route is meant to be automatic and rely on the emotional content of the situation. However, evaluating a decision in terms of consistency to a rule (such as a moral norm) should be an analytic process (Posner \& Snyder, 1975; Sloman, 1996; Kahneman, 2003). There is a corresponding debate, with some researchers arguing that deontological decisions are a confabulation of moral emotions (Greene, 2007; Haidt, 2001) and others rejecting this assumption (Kahane \& Shackel, 2010; Kahane, 2012; Mihailov, 2016). We propose progress this debate, using a novel lab-based moral dilemma, that we call The Refugees' Dilemma. We explore whether it is possible to discriminate between moral decisions based on the emotional content of a situation (emotional decisions) and decisions driven by a prerogative of consistency with a rule (rule-based decisions).

Moral decision-making has been dominated by the Trolley dilemma (Thomson, 1985). However, the footbridge version of this dilemma is ill-suited for the present purposes, because the deontological option (not to push the fat man) is confounded with the emotional choice. The novel Refugees' Dilemma addresses this problem. The Refugees' Dilemma is an adaptation of the Trolley dilemma, but involving three choices: Utilitarian (driven by consequences/outcomes) vs. Emotional (driven by emotions) vs. Rule-based (driven by an unemotional rule).

We employ three tools which may reveal differences between the three postulated routes to moral decisionmaking. First, we measured psychological distance. We assume that participants making rule-based or utilitarian decisions will evaluate a situation with greater psychological distance and conversely regarding emotional decisions. Psychological distance weakens the intensity of people's affective reactions, such as feelings of empathy (Williams et al., 2014). Furthermore, increasing psychological distance leads individuals to construe situations in more abstract terms, which sometimes aligns with more utilitarian decision-making (Trope \& Liberman, 2010) and, we hypothesize, with more rule-driven decisions too.

Second, we tested participants on the Cognitive Reflection Test (CRT) (Frederic, 2005), which distinguishes two modes of cognitive processing, one that is more reflective and slow versus one that is more immediate with little conscious deliberation. Differences in the CRT should align with a propensity to adopt utilitarian vs. rule-based vs. emotional decision making, whereby we assume that
utilitarian and rule-based decisions require greater reflection (since a person needs evaluate how the well-being of one group of individuals is balanced with that of another or consider the applicability of a rule and the consequences of violating it) than emotional ones.

Third, we implemented a time-manipulation, whereby participants would either be told they had unlimited time or that they should respond as quickly as possible (but even in the time pressure condition participants had ample time to respond). Finally, we included measures with a mindset for practical application. Is responding in The Refugees' Dilemma sensitive to religious or political characteristics?

## Method

## Participants

A total of 1508 participants, all of whom were US residents, were recruited on-line and received $\$ 0.80$ for doing the task ( 706 women, 801 men ; M age $=34.6$ years, $\mathrm{SD}=11.17$ ). As this is the first study with The Refugees' Dilemma, no prior power analyses were conducted and instead we decided a priori to limit recruitment to 1500 participants. The City, University of London Psychology Department Research Ethics Committee granted approval for this project (reference PSYETH (S/L) 15/16 238).

## Materials and Procedure

The study was designed in Qualtrics, run on Amazon Mechanical Turk and lasted 10 minutes approximately. We used frequency of Type of Judgment (Utilitarian vs. Emotional vs. Rule-based) as the dependent measure. Time (No Time vs. Unlimited Time vs. Time Pressure) was manipulated between participants and we used the scores from the CRT (Frederic, 2005) to measure thoughtful (high CRT scores) vs. unreflective (low CRT scores) cognitive processes.

After a few preliminary screens (consent form; some basic demographic information), all participants were presented with The Refugees' Dilemma (full text in Online Supplementary Material). They were instructed to read it carefully and had to spend at least 60 seconds reading it before the experiment advanced. The Refugees' Dilemma asks a participant to imagine himself/herself as a security guard in a border control of a hypothetical country, which neighbors three other countries. Participants are told they have to make one last decision before borders close (until further notice) and that there are instructions that entry into their country will be allowed from just one country.

Then, participants were presented with a reinforcementlearning task to ensure that they had been paying attention during the previous screen. Three basic multiple-choice questions regarding The Refugees' Dilemma were presented (e.g., "As a security guard, what is the name of your country?). Feedback was provided and participants had to keep responding until no mistakes were made.

Subsequently, the three moral scenarios were presented (Utilitarian: where ten refugees from another country need help; Emotional: where a refugee orphan child from another country needs medical attention immediately; and Rulebased: where a traveller from your own country wants to go back home and the law from your country specifies that travellers who are citizens from your own country have to take priority when returning). The text for each scenario was supplemented with an illustration (Figure 1). The moral choice was then presented to participants: "Who do you allow to your country? Remember, you can only allow traveller(s) from one neighbouring country". Participants had to choose between Choice 1 (Utilitarian; "The 10 refugees from Beta"), Choice 2 (Emotional; "The refugee orphan child from Gamma"), or Choice 3 (Rule-based; "The traveller who is an Alpha citizen coming from Delta").

Regarding the time manipulation, one third of participants was not provided with any indication of time for making their judgment (No Time). Participants in the Unlimited Time condition were instructed as follows: "You will have unlimited time to answer the question in the next page. Think carefully about your judgment before responding". Participants in the Time Pressure condition were presented with the following instructions: "The question in the next page should be answered as fast as possible. Use your first impression/ gut feeling in order to respond"; these participants had to make their moral choice while a timer (at the top of their screen) kept track of elapsed time.

Participants were next asked to complete a 4-items questionnaire (see Online Supplementary Material), which was intended as a measure of the basis of participants' judgments (e.g., "How much would you say that doing the greater good for the greatest number of people/ emotion/ a principle, norm or rule was the basis for your decision?"). The order of these three questions was presented randomly and participants had to respond moving a slider that went from 0 (not at all) to 7 (for the above question, completely based in doing the greater good for the greatest number of people/ emotion/ a principle, norm or rule). The fourth item of the questionnaire, following the same format, was a measure of psychological distance (Trope \& Liberman, 2010) (e.g., "How distant do you feel yourself from the scenario when making your decision?").

Participants were next presented with three "catch questions", to control for attention and basic comprehension during the task (e.g., "How many refugees there were in the group from Beta?").

Then, participants had to complete a CRT (Frederic, 2005) as a measure of two modes of cognitive processing, quick with little conscious deliberation versus slower and more reflective. The test consisted of three multiple-choice questions (e.g., "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?").

Finally, participants were asked to complete demographic questions regarding their levels of Religiosity (using a 7-
point Likert scale) and Political Views (Liberal, Moderate, Conservative or Something else).


Figure 1: Illustrations and choices used in The Refugees'
Dilemma: (1) Utilitarian (judgment driven by consequences/ outcomes) vs. (2) Emotional (judgment driven by emotions) vs. (3) Rule-based (judgment driven by a rule, principle or norm).

## Results

## Validation of the Experimental Paradigm

We excluded those participants who did not answer the catch questions correctly (92/1508). No other sample trimming was conducted.

We first discuss results which aim to validate the assumptions in the design of The Refugees' Dilemma. We tested if the three different choices presented in the dilemma (Choice 1, Utilitarian, "The 10 refugees from Beta"; Choice 2, Emotional, "The refugee orphan child from Gamma"; and Choice 3, Rule-based, "The traveller who is an Alpha citizen coming from Delta") were indeed aligned with doing the greater good for the greatest number of people, with emotion or with a rule, as assumed (see Figure 2). As expected, participants making the utilitarian choice reported that their decision was mainly based on doing the greater good for the greatest number of people ( $\mathrm{M}=6, \mathrm{SD}=1.3$ ). Participants making the emotional choice reported that their decision was mainly based on emotions $(M=5.7$, $\mathrm{SD}=$ 1.4). Finally, participants making the rule-based choice reported that their decision was mainly based indeed on a rule, principle or norm $(\mathrm{M}=6, \mathrm{SD}=1.5)$. One-way ANOVAs for each group of participants were all significant: $\mathrm{F}(2,308)=56.93, \mathrm{p}<.001, \mathrm{w} 2=.27$ for the utilitarian respondents; $\mathrm{F}(2,1226)=337.787, \mathrm{p}<.001$, $\mathrm{w} 2=.35$ for the emotional respondents; $\mathrm{F}(2,2708)=2511.996, \mathrm{p}<.001$, $\mathrm{w} 2=.65$ for the rule-based respondents. A Tukey post-hoc test for each group revealed significant differences in the expected directions ( $\mathrm{p}<.001$ ). These results are all
consistent with expectation regarding the assumptions motivating the three options in The Refugees' Dilemma.


Figure 2: Mean scores for the basis of judgments, for participants making the utilitarian, emotional, or rule-based choice. Error bars represent standard errors.

As a manipulation check regarding time, we examined the amount of time that participants took to make their judgments. Participants spent more time responding in the Unlimited Time condition (18.74s) than in the No Time condition (14.47s) and than in the Time Pressure condition (8.11s). A one-way between subjects ANOVA for these means was significant $\left(\mathrm{F}(2,1414)=25.017, p<.001, w^{2}=\right.$ .03). A Tukey post-hoc test revealed that all pairwise comparisons between groups were significant $(p=.013)$.

## High vs. Low Cognitive Resources, Psychological

## Distance and Time

We first considered whether results from the CRT influence moral choice (entire sample; Figure 3). We selected participants who reported High vs. Low CRT scores (i.e. 3/3 points and $0 / 3$ points in the CRT, respectively). As expected, high CRT participants were more likely to opt for the utilitarian response $(59.38 \%)$ than low CRT ones ( $40.63 \%$ ). Likewise, emotional answers were more likely for low CRT participants (55.7\%) than otherwise (44.31\%). Importantly, the rule-based response was also more likely for high CRT participants ( $60.73 \%$ vs. $39.27 \%$ ), indicating the rule-based moral choices require a similar route as utilitarian ones. A $3 \times 2$ chi-square test on response counts, with the variables Type of Response (Utilitarian vs. Emotional vs. Rule-based) and CRT score (High vs. Low) was significant, $\chi 2(2, \mathrm{~N}=897)=19.66, \mathrm{p}<.001$.


Figure 3: Percentage of Utilitarian, Emotional and Rulebased responses for participants who followed a slow and more reflective cognitive process (High CRT) or a quicker with little conscious deliberation one (Low CRT). Error bars represent standard errors.

We next examined whether different moral changes reflected the expected differences regarding Psychological Distance (see Figure 4). Participants opting for the rulebased option reported the highest distance ( $\mathrm{M}=3.77$, $\mathrm{SD}=$ 2.1), followed by participants making the utilitarian selection ( $\mathrm{M}=3.18, \mathrm{SD}=2.1$ ), and finally the ones selecting the emotional answer $(\mathrm{M}=2.71, \mathrm{SD}=1.93)$. It is interesting that participants making the rule-based choice reported the highest distance, perhaps because the application of a rule to the dilemma requires a degree of detachment from the specifics of the situation more so than even for utilitarian respondents. A one-way between subjects ANOVA for these means was significant ( $\mathrm{F}(2$, $1414)=38.233, \mathrm{p}<.001, \mathrm{w} 2=.05)$. A Tukey post-hoc test revealed that psychological distance was significantly different between participants making the utilitarian and the rule-based selection ( $\mathrm{p}=.018$ ) and between participants making the emotional and rule-based selection ( $\mathrm{p}<.001$ ). There were no statistically significant differences between the utilitarian and emotional groups $(\mathrm{p}=.089)$.


Figure 4: Mean scores for psychological distance between the different moral choices presented in The Refugees’ Dilemma. Error bars represent standard errors.

Finally, we considered differences in Type of Response (Utilitarian vs. Emotional vs. Rule-based) depending on both the Time manipulation (No Time vs. Unlimited Time vs. Time Pressure) and the CRT score (High vs. Low). The three-way loglinear analysis produced a final model that retained the Type of Response x CRT score interaction, but not the three-way interaction. The likelihood ratio of this model was $\chi 2(12)=11.647, \mathrm{p}=.475$. The Type of Response $\times$ CRT score interaction was significant, $\chi 2$ (2) $=$ $20.225, \mathrm{p}<.001$. This interaction indicates that the relative frequencies of utilitarian, emotional and rule-based responses were different across high, low CRT scores. Of interest, the frequency of rule-based responses increased more dramatically between low, high CRT scores ( 225 to 351), than for utilitarian responses ( 25 to 38 ). As expected, the frequency of emotional responses followed the opposite direction (142-113 ratio of low to high CRT scores). Therefore, the analysis reveals a fundamental difference between the cognitive resources used to reach a specific type of judgment (as also concluded with the other analyses above), but these effects were not influenced by the time manipulation.

## Political Views and Religiosity

We first explored the differences in moral choice, depending on participants' stated Political Views (Figure 5a). Liberals were more likely to opt for the utilitarian response (10.11\%) compared to Moderates (5.3\%) and Conservatives (3.9\%). Liberals were also more likely to opt for the emotional answer (33.23\%) compared to Moderates (28.53\%) and Conservatives (21.19\%). Interestingly, Conservatives were more likely to opt for the rule-driven judgment (74.93\%) compared to Liberals (56.67\%) and Moderates (66.13\%). A corresponding $3 \times 3$ chi-square test of independence was highly significant, $\chi 2(4, \mathrm{~N}=1363)=37.62, \mathrm{p}<.001$. Individual $3 \times 2$ chi-square tests for each category of
respondents (utilitarian, emotional, rule-based) were also significant $\left(\chi^{2}(2, \mathrm{~N}=1363)=15.61, \mathrm{p}<.001 ; \chi^{2}(2\right.$, $\mathrm{N}=1363)=15.64, \mathrm{p}<.001 ; \chi 2(2, \mathrm{~N}=1363)=33.23, \mathrm{p}<$ .001; respectively).

Regarding religiosity, we selected only participants who reported Low vs. High levels of religiosity (i.e. $1 / 7$ points and $3 / 7$ or more points in the 7 point Likert scale, respectively; Figure 5b). Low-Religious participants were more likely to opt for the utilitarian response (8.1\%) compared to High-Religious ones (6.32\%). High-Religious participants were more likely to opt for the emotional answer ( $31.04 \%$ ) compared to Low-Religious ones (25.6\%). Finally, Low-Religious participants were more likely to opt for the rule-driven judgment ( $66.35 \%$ ) compared to HighReligious ones ( $62.64 \%$ ). A $3 \times 2$ chi-square test of independence on Type of Judgment (Utilitarian vs. Emotional vs. Rule-based) against participants' levels of Religiosity (Low vs. High) was significant $\chi 2$ (2, N=994) = 15.36, p $<.001$.
a)

b)


Figure 5: Percentage of Utilitarian, Emotional and Rulebased responses for (a) participants' Political Views (Liberal vs. Moderate vs. Conservative) and (b) participants' levels of Religiosity (Low vs. High). Error bars represent standard errors.

## Discussion

Established theory assumes that deontological moral choices involve a fast, gut-feeling process, driven by the emotional content of the situation (Greene, 2009). There is no doubt that this is sometimes the case, e.g., in cases of moral norms of high emotional content (Valdesolo \& DeSteno, 2006). However, it seems counterintuitive that all deontological moral choices are of this kind. We supported the hypothesis that there are deontological moral choices based on an unemotional rule, which can be distinguished from emotiondriven ones. We provided evidence for a route to moral choice, distinct from the emotional and utilitarian routes, and rather based on a prerogative to adhere by a given rule.

The characteristics of the rule-based influence in moral choice were explored with three manipulations. First, according to Construal Level Theory (CLT), greater psychological distance would go hand-in-hand with lower emotional involvement. We found that rule-based respondents reported the highest average distance, more so that even utilitarian respondents. Such a result is consistent with the nature of the rule provided in The Refugees' Dilemma, since application of the rule forces ignoring most characteristics of the different options. Second, high CRT participants were more likely to adopt utilitarian or rulebased responses, as opposed to emotional ones. This shows how emotional and rule-based influences can be separated and challenges the approach of assuming both can be combined into a single deontological route. Note, other work supports a view of utilitarian judgments as reflecting a greater "need for cognition" (Bartels, 2008), "cognitive reflection" (Hardman, 2008), and working memory capacity (Moore et al., 2008). Third, a time manipulation produced a complex interaction with CRT level. Future work should examine whether perhaps just tracking time might result in reduced cognitive resources for moral decisions, regardless of condition.

We developed a new paradigm for moral choice, The Refugees' Dilemma, which is based on a situation relevant for millions of citizens, especially in Europe and North America. We hope that future work will further explore moral decision situations informed by relevant current affairs or near-future social dilemmas (e.g., Bonnefon et al., 2016). With a mind to such applications, we reported some interesting correspondences between moral choice in The Refugees' Dilemma and participants' political affiliations and religious convictions.

## Acknowledgements

This work was supported by grants AFOSR FA8655-13-13044 and Leverhulme Trust grant RPG-2015-311 and H2020-MSCA-IF-2015 grant 696331.

## References

Singer, P. (1991). A companion to ethics. Oxford, England: Blackwell Reference.

Chaiken, S.; Trope, Y. (1999). Dual-Process Theories in Social Psychology. New York: Guilford Press.
Greene, J. D., Sommerville, R., Nystrom, L., Darley, J., \& Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 21052108.

Greene J, Haidt J. (2002). How (and where) does moral judgment work? Trends Cognitive Science 6: 517523. doi:10.1016/S1364-6613(02)02011-9.
Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD. (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron; 44:389-400.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F. (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446: 908- 911. doi:10.1038/nature05631.
Mill, JS. Utilitarianism. Crisp, R., editor. New York: Oxford University Press; 18611998.
Kant, I. Foundation of the metaphysics of morals. Beck, LW., translator. Indianapolis: Bobbs-Merrill; 17851959.
Posner, MI., Snyder, CRR. (1975). Attention and cognitive control. In: Solso, RL., editor. Information processing and cognition. Hillsdale, NJ: Erlbaum, p. 55-85.
Sloman, S. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 11, 3-22.
Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. Am. Psychol. 58, 697-720.
J.D. Greene (2007). Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development(ed. W. Sinnott-Armstrong). MIT Press, Cambridge, MA.
Haidt J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review.108:814-834. [PubMed: 11699120]
Kahane G., and N. Shackel. (2010). Methodological issues in the neuroscience of moral judgement. Mind \& Language 25(5): 561-582.
Kahane G. (2012). On the wrong track: process and content in moral psychology. Mind \& Language 27(5): 519-545.
Mihailov, E. (2016). Is Deontology a Moral Confabulation? Neuroethics, 9, 1-13. doi: 10.1007/s12152-015-9244-5.
Thomson, J. (1985). The trolley problem. Yale Law, 94: 1395-1415. doi:10.2307/796133.
Williams, Lawrence E. Stein, R., Galguera, L. (2014), The Distinct Affective Consequences of Psychological Distance and Construal Level. Journal of Consumer Research, 40, 1123-1138.
Trope, Y., \& Liberman, N. (2010). Construal Level Theory of Psychological Distance, Psychological Review, 117 (April), 440-63.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 2542. doi: $10.1257 / 089533005775196732$

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. Journal of

Experimental Social Psychology, 45, 581-584.
Valdesolo, P. \& DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. Psychological Science, 17, 476-477.
Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. Cognition, 108(2), 381-417.
Hardman, D. (2008). Moral dilemmas: Who makes utilitarian choices? Unpublished manuscript.
Moore, A., Clark, B., \& Kane, M. (2008). Who shalt not kill?: Individual differences in working memory capacity, executive control, and moral judgment. Psychological Science, 19(6), 549-557.
Bonnefon, J. F., Shariff, A., Rahwan, I. (2016). The Social Dilemma of Autonomous Vehicles. Science. 352(6293):1573-1576.

# I know what you need to know: Children's developing theory of mind and pedagogical evidence selection 

Ilona Bass (ilona.m.bass@rutgers.edu) ${ }^{1}$, Elizabeth Bonawitz (elizabeth.bonawitz@rutgers.edu) ${ }^{1}$, Patrick Shafto (patrick.shafto@gmail.com) ${ }^{2}$, Dhaya Ramarajan (dhayar@gmail.com), Alison Gopnik (gopnik@berkeley.edu) ${ }^{\mathbf{3}}$, Henry Wellman (hmw@umich.edu) ${ }^{4}$<br>${ }^{1}$ Department of Psychology, Rutgers University, Newark, NJ 07102<br>${ }^{2}$ Department of Math and Computer Science, Rutgers University, Newark, NJ 07102<br>${ }^{3}$ Department of Psychology, University of California, Berkeley, CA 94720<br>${ }^{4}$ Center for Human Growth and Development, University of Michigan, Ann Arbor, MI 48109


#### Abstract

Natural pedagogy emerges early in development (Knudsen \& Liszkowski, 2012), but good teaching requires presenting evidence specific to learners' knowledge (Shafto, Goodman, \& Griffiths, 2014). How might the development of Theory of Mind (ToM) relate to the ability to select pedagogical evidence? We present a training study in which we investigated the link between preschool-aged children's false-belief understanding and their ability to select evidence for teaching. Our results suggest that children with more advanced ToM abilities were better evidence selectors, even when controlling for effects of age and numerical conservation abilities. We also found that children who improved more in false-belief understanding from pre- to post-test performed better on the pedagogical tasks over the course of the training. Finally, we report tentative evidence for a link between the pedagogical training and improvements in ToM. Our findings suggest important connections between ToM and evidential reasoning in natural pedagogy in early childhood.


Keywords: pedagogy, theory of mind, evidence selection

## Introduction

The ability to teach and be taught by others is an indispensable human capability. Social transmission of information is one of the key ways in which both children and adults learn about the world, and some have argued that the natural tendency to teach and to be ready to learn from others may be what sets human intelligence apart from other animals (e.g., Moll \& Tomasello, 2007). Indeed, teaching in children emerges at an early age: Three-year-olds spontaneously engage in teaching behavior with their peers (Ashley \& Tomasello, 1998), and infants as young as 12 months selectively point to convey information to naïve (as opposed to knowledgeable) adults (Knudsen \& Liszkowski, 2012). Investigating children's developing ability to teach others may shed insight into the cognitive mechanisms that support natural pedagogy. We will suggest that the factors that support this skill - reasoning about the knowledge states of others and reasoning about evidence - are intertwined.

## Teaching in Early Childhood and Theory of Mind

Children's teaching abilities improve considerably between the ages of three and five years. Davis-Unger and Carlson (2008) had three- to five-year-old children teach a confederate how to play a novel board game, and found that older children 1) taught for longer periods of time, 2) explained
more of the rules, and 3) used a more diverse range of teaching strategies. Similarly, Strauss, Ziv, and Stein (2002) found that five-year-olds taught others by providing verbal explanations, whereas three-year-olds used more demonstrationbased teaching strategies. There is also evidence that older children possess more declarative knowledge about pedagogy in general (Ziv \& Frye, 2004).

What are the fundamental cognitive underpinnings that support the development of children's pedagogical skills? Theory of Mind (the ability to represent others' mental states and to understand that others may experience mental states that are different from one's own) has been proposed as being critical for children's teaching. Intuitively, a relationship between Theory of Mind (ToM) and children's developing teaching skills makes sense: ToM involves monitoring the mental states of others, and effective teaching requires understanding what your student does and does not know. Additionally, ToM undergoes drastic qualitative change between the ages of three and five, the same period during which children's pedagogical skills are developing. Indeed, there is a wealth of empirical work that provides evidence for a link between ToM development and pedagogical skill (Davis-Unger \& Carlson, 2008; Strauss et al., 2002). ToM may thus be an important cognitive mechanism that drives the development of children's ability to teach others.

## Evidence Selection in Teaching

The past work on children's teaching and ToM ability has operationalized teaching ability in various ways, including by the length of the teaching interaction, the types of strategies used, and whether children recognize that some individuals need to be taught while others do not. An additional and perhaps more detailed way of conceptualizing pedagogical skill comes from the distinct but related body of literature on concept learning and pedagogical sampling. Research in this field emphasizes the importance of selecting and presenting a learner with specific evidence that will allow them to infer a particular conclusion (e.g., Gweon, Tenenbaum, \& Schulz, 2010; Shafto et al., 2014). According to this view, being a "good" teacher requires more than just recognizing whether or not someone needs to be taught, or even that some learners need to be taught more than others; rather, good teaching depends on having a deeper understanding of the precise ev-
idence that certain learners may or may not need in order to infer a particular conclusion.

Prior work has shown that children are sensitive to learning goals in pedagogical scenarios. Six-year-olds will select diverse samples to teach a novel concept to a peer, but not to learn a novel concept for themselves (Rhodes, Gelman, \& Brickman, 2010). Preschoolers are even capable of selectively presenting evidence to intentionally deceive learners. Rhodes, Bonawitz, Shafto, Chen, and Caglar (2015) showed three- to six-year-olds a novel toy that activated when any block was placed on it. They then asked children to pick two blocks to either 1) teach a naïve puppet how the toy really worked, or 2) trick her into thinking that only red blocks made it go. Children reliably selected blocks that would best communicate the pedagogical goal, regardless of whether the goal was to teach or to deceive (Rhodes et al., 2015).

There is also an abundance of work demonstrating that when learning from others, children use the evidence presented to them to make inferences about the knowledgeability of their teachers (see Kushnir and Koenig (in press) for a recent example). Pasquini, Corriveau, Koenig, and Harris (2007) showed children videos of adults naming familiar objects with varying rates of accuracy; children were then asked from whom they would prefer to learn the names of novel objects. Three- and four-year-olds preferred to learn from more accurate teachers, suggesting that children use presented evidence in pedagogical scenarios to update their beliefs about whether teachers are knowledgeable or not. Despite this robust preference for accurate teachers, there has also been work showing that children are able to exonerate previously inaccurate teachers whose past inaccuracies occurred for legitimate reasons (Nurmsoo \& Robinson, 2009). Children therefore additionally monitor teachers' epistemic states in conjunction with the evidence they've presented in order to make inferences about their competence.

Together, the works cited in this section suggest that children are developing the ability to reason about evidence in the service of teaching in the early childhood years. However, we are unaware of any work that has investigated the precise relationship between ToM development and children's ability to effectively select pedagogical evidence to teach others. ToM may play an especially important role in supporting this aspect of teaching, because effective evidence selection requires the on-line monitoring of a learner's epistemic state relative to a particular learning goal. The current paper presents a novel experiment that explores the relationship between children's pedagogical evidence selection and ToM development.

## Teaching Training and Theory of Mind Study

We investigated the relationship between children's Theory of Mind ability (as measured by a false-belief battery; Wimmer \& Perner, 1983; Gopnik \& Astington, 1988) and their ability to select evidence to teach another. Assuming this link, we predicted that children with more proficient Theories of Mind would be better at pedagogical evidence selection, and also
that training pedagogical skill might lead to improvements in ToM reasoning abilities. To explore this, we assessed children's false-belief understanding before and after training them on two pedagogical tasks. We also assessed children's understanding of numerical conservation; we wanted to be sure that any improvements we saw due to the pedagogical training was specific to ToM abilities and not to other unrelated domains of cognitive development.

We chose to use false-belief tasks to measure ToM ; between the ages of three and five, children reliably transition from predicting others' actions based on the veridical state of the world to understanding that others' actions are in fact guided by their (sometimes false) beliefs (Wellman, Cross, \& Watson, 2001). Some have argued that implicit falsebelief understanding emerges at much earlier ages (between 10 and 15 months), and that apparent developments in ToM between the ages of 3 and 5 years are actually reflections of task demands (Baillargeon, Scott, \& He, 2010). Nevertheless, there is ample evidence that the changes that occur in children's ToM understanding during the preschool years are critical: This is the time during which children gain the ability to provide explicit causal explanations for others' actions based on epistemic states (e.g., Bartsch \& Wellman, 1989); further, differences in preschoolers' false-belief understanding are predictive of numerous other capabilities, including children's tendency to talk about people in everyday conversation, and their social competence more broadly (see Astington \& Jenkins, 1995; Imuta, Henry, Slaughter, Selcuk, \& Ruffman, 2016), suggesting an important link between performance on these tasks in early childhood and real cognitive development. We therefore used false-belief tasks to measure ToM abilities.

## Methods

## Participants

Sixty-one children $\left(M_{\text {age }}=47\right.$ months, range $=39-55$ months) were recruited from and tested at local preschools.

## Tasks

False-Belief Children's ToM was assessed using two classic false-belief tasks. In the first task, children saw a storybook in which Sally put her cookie in a box and then left the room. While Sally was gone, Anne came in and moved Sally's cookie from the box to a bag. Children were then asked, "When Sally comes back, where will she look for her cookie?" Children earned a point if they correctly reported that Sally would look for her cookie in the box. In the second task, children were shown a crayon box that, it was soon revealed, actually contained some keys. Experimenters asked the children 1) what they thought was inside the box when they first saw it, and 2) what was really inside. Children earned one point if they correctly answered both of these questions. The experimenter then introduced a doll, and asked children the same two questions ("What will the doll think is inside this box? What's really inside?"). Children
again received one point for correctly answering both questions. False-belief scores could thus range from zero to three.
Numerical Conservation Control Task To assess children's understanding of numerical conservation, experimenters showed children two parallel rows of ten objects each, both of which were equal in length. Children were asked if row A or B had more objects, or if they were the same. Then, experimenters lengthened one of the rows, and again asked children if row A or B had more objects, or if they were the same. This process constituted one trial; children had to answer both questions correctly on a given trial to earn one point. Experimenters administered two trials; conservation scores could thus range from zero to two.
Pedagogical Training and Test The pedagogical training entailed a novel word learning task and a causal toy activation task. In the novel word learning task, children were told that a novel word (e.g., "Dax") represented the concept they were trying to learn. They were shown a picture of an object with two discrete features (e.g., a fork that is white), and were told that this picture represented the target concept ("This is a Dax!"). Given the inherent ambiguity in the word's extension, the experimenter explained what the novel word really meant ("Dax means fork."). The experimenter then presented two additional pictures, each of which contained an item that overlapped with exactly one of the original picture's two features (e.g., a white spoon, and a black fork). The experimenter then asked children to teach a confederate what the novel word meant by providing examples using the three pictures, without explicitly telling the confederate what the novel word meant. In order to provide a correct response, children had to present the necessary and sufficient examples to identify the correct rule while ruling out other hypotheses ${ }^{1}$.

In the causal toy activation task, children were presented with a novel toy with two distinct mechanisms (e.g., a wheel and a bell). The experimenter first showed children how to activate the toy, causing it to perform some desirable outcome such as lighting up or playing music ("You need to ring the bell and spin the wheel at the same time to make the toy go."). As in the novel word task, children were then instructed to teach a confederate about the toy by providing examples of which combinations of mechanisms did and did not make the toy go. In order to provide a correct response, children had to demonstrate both necessary and sufficient evidence for the confederate to rule out all alternative explanations and correctly infer which mechanism(s) activated the toy.

For both tasks, if children provided insufficient evidence, the confederate prompted the child by musing aloud about the remaining possible explanations. For example, if the child only showed the confederate that operating both mechanisms simultaneously made the toy go, the confederate might say:

[^19]"Oh, so you showed me both at the same time. It could be that you need to do both at the same time to make it go, or it could be that the wheel by itself could make it go, or that the bell by itself could make it go. Can you teach me?" Note that often children would need to present negative examples to rule out plausible hypotheses (e.g., showing that the wheel by itself did not make it go). The number of prompts children required before providing complete evidence was the primary DV for both pedagogical training tasks; these scores could range from a minimum of zero (i.e., children who provided necessary and sufficient evidence spontaneously) to a maximum of two (i.e., children who required prompting after each demonstration until all evidence had been provided).

There were six different versions of each task: The novel words were fep, dax, modi, toma, wug, and blicket; the causal toys were phone, gear toy, helicopter, shadowbox, red airplane, and purple. Some of the novel words represented just one of the two categories to which the example object belonged ("Dax means fork"), while others represented both categories ("Dax means white fork"). Likewise, some causal toys would activate any time one of the mechanisms was operated ("Any time you ring the bell, it makes the toy go"), while others would only activate if both mechanisms were operated simultaneously ("You need to ring the bell and spin the wheel at the same time to make the toy go"). Varying the stimuli in this way ensured that children would have distinct teaching goals on different trials, and would thus have to select evidence that corresponded to the particular teaching goal of a given trial in order to provide a correct response.

## Procedure

Children's understanding of false-belief and numerical conservation was assessed on a preliminary testing day. Children who scored fewer than two out of three points on the false-belief task were classified as copy theorists (i.e., those who think that beliefs are always consistent with the world), while children who scored two or more points were classified as perspective theorists (i.e., those who understand that beliefs may vary with perspective, and can thus be false; see Goodman et al., 2006). Copy theorist (CT) children ( $N=40$ ) were randomly assigned to either the control or the training condition. Over the course of the following six weeks (beginning on the preliminary testing day), children in the training condition ( $N=22 ; M_{\text {age }}=46$ months) received two training sessions per week on both pedagogical tasks. One version of each task was administered on a given testing session, with the novel word task always being presented first. As there were six versions of both the novel word task and the causal toy task, the experimenter administered the same version of each task across both sessions of a given week. The order in which the different versions of the tasks were presented was randomized across participants. At the end of this six week period, children's understanding of false-belief and numerical conservation were reassessed using the same measures with slightly different stimuli.

CT children in the control condition $\left(N=18 ; M_{\text {age }}=46\right.$

| CT Children: Training | 6-week Training Example Schedule |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Preliminary Assessments <br> - False-belief <br> - Conservation |  | 1 |  | 2 |  | 3 |  | 4 |  | 5 |  | 6 |  | Post-test Assessments <br> - False-belief <br> - Conservation |
|  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |  |
|  |  | Modi | Mod | Bicket | Blicket | Dax | Dax | Toma | Toma | Fep | fep | Wug | Wu |  |
| CT Children: Control |  | $\begin{aligned} & \begin{array}{l} \text { Gear } \\ \text { Tor } \end{array} \end{aligned}$ | $\begin{aligned} & \begin{array}{l} \text { Gear } \\ \text { Tor } \end{array} \end{aligned}$ | Purple | Purple | $\begin{array}{\|c} \hline \text { Shadow } \\ \text { box } \\ \hline \end{array}$ | $\begin{gathered} \text { Shadow } \\ \text { box } \end{gathered}$ | $\begin{aligned} & \text { Heli- } \\ & \text { coper } \end{aligned}$ | $\begin{gathered} \text { Heli- } \\ \text { coper } \end{gathered}$ | $\begin{gathered} \text { Red } \\ \text { Aiplane } \end{gathered}$ | $\begin{gathered} \text { Red } \\ \hline \text { Airplane } \end{gathered}$ | Phone | Phone |  |
| Preliminary Assessments <br> - False-belief <br> - Conservation | 6-week Delay |  |  |  |  |  |  |  |  |  |  |  |  | Post-test Assessments <br> - False-belief <br> - Conservation |
| PT Children | One Training Session |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Preliminary Assessments <br> - False-belief <br> - Conservation | Novel Word |  | $\begin{array}{\|c\|} \hline \text { Wug } \\ \hline \text { Purple } \\ \hline \end{array}$ |  |  |  |  |  |  |  |  |  |  |  |

Figure 1: A schematic of our study design, with examples of possible pedagogical training schedules for CT children in the training condition and PT children.
months) received no pedagogical training, and their falsebelief and conservation understanding was reassessed after a six-week delay. Perspective theorist (PT) children ( $N=21$; $M_{\text {age }}=49$ months) did not receive longitudinal pedagogical training, since they had little to no room for improvement on the false-belief tasks; instead, they received just one session of the pedagogical tasks on the preliminary testing day, allowing us to measure their initial teaching abilities. The versions of the pedagogical tasks used with PT children were randomized across participants. PT children's false-belief and conservation understanding was not reassessed. See Figure 1 for a schematic of our study design.

## Results

One CT child in the training condition did not complete one session of the causal toy task, another CT child did not complete one session of both tasks, and one PT child's numerical conservation abilities were not assessed; these individual data points were treated as missing in subsequent analyses. Otherwise, all children completed all training sessions and assessments. We created a composite pedagogical skill score for each training session by calculating the average number of prompts children required across both tasks in a given session. Lower scores indicated better task performance.

## Initial False-Belief \& Pedagogical Skill

We first investigated the effects of preliminary false-belief understanding on initial (i.e., non-trained) pedagogical skill. An independent-samples t-test compared CT children in the training condition to PT children on the average number of prompts required on the preliminary testing day. We found that PT children $(M=1.05, S D=.57)$ provided complete evidence with significantly fewer prompts than CT children $(M=1.45, S D=.55), t(41)=2.38, p=.022,95 \% \mathrm{CI}_{\text {diff }}=$ [.06, .75]; see Figure 2A. We also looked at the novel word and causal toy tasks separately: While PT children ( $M=.57$, $S D=.68$ ) significantly outperformed CT children ( $M=1.27$, $S D=.78)$ on the causal toy task $(t(41)=3.17, p=.003,95 \%$
$\left.\mathrm{CI}_{\text {diff }}=[.26,1.15]\right)$, there were no significant differences between the two groups on the novel word task (CT: $M=1.64$; PT: $\left.M=1.52 ; p=.584,95 \% \mathrm{CI}_{\text {diff }}=[-.30, .53]\right)$. This disparity may be explained by the seemingly increased difficulty of the novel word task relative to the causal toy task. Indeed, two paired-samples t-tests revealed that both CT and PT children performed better on the causal toy task than on the novel word task on the preliminary testing day (CT: $t(21)=5.05$, $p<.001,95 \% \mathrm{CI}_{\text {diff }}=[.56,1.35] ; \mathrm{PT}: t(20)=2.16, p=.042$, $\left.95 \% \mathrm{CI}_{\text {diff }}=[.01, .71]\right)$; additionally, more children required the maximum of two prompts on the novel word task $(N=29)$ than on the causal toy task $(N=12)$. The novel word task might have been more difficult for children than the causal toy task for several reasons; perhaps children are more generally familiar with toys, or have more experience teaching about toys than about words. Future work could explore the differences between these two tasks.

There are many possible reasons why PT children may have outperformed CT children on the causal toy task, including age or other cognitive factors. To control for this, we ran two between-subjects ANCOVAs, with theorist type (CT vs. PT) predicting performance on the causal toy task; we included preliminary conservation scores as a covariate in one analysis, and age at pre-test in the other. PT children still outperformed CT children on the causal toy task, even when controlling for effects of age $(F(1,40)=6.11, p=.018)$ and conservation scores $(F(1,39)=9.35, p=.004)$, providing stronger evidence for a direct link between false-belief understanding and teaching ability.

## False-Belief Improvement \& Pedagogical Skill

Next, we investigated the relationship between overall aggregate performance on the two pedagogical tasks and improvement on the false-belief task from pre- to post-test. Using data from CT children in the training condition, we ran a correlation between false-belief improvement (i.e., pre-test false-belief scores subtracted from post-test scores) and the


Figure 2: A. PT children provided the confederate with complete evidence after receiving significantly fewer prompts than CT children on the first day of pedagogical testing. B. Children required fewer prompts over the course of the pedagogical training. C. For children who answered all false-belief questions incorrectly at pre-test, those in the training condition significantly improved in false-belief understanding from pre- to post-test, whereas those in the control condition did not. Asterisks denote significance at the $p<.05$ level. All error bars represent two standard errors.
mean number of prompts required across all twelve training sessions. We found a statistically significant negative linear relationship between these two factors, $r(20)=-.43, p=$ .047. In other words, children who required fewer prompts over the course of the training generally improved more in false-belief understanding from pre- to post-test. Two partial correlations revealed that this finding qualitatively persisted when statistically controlling for average age $(r(19)=-.43$, $p=.054$ ) and improvement in conservation understanding $(r(19)=-.41, p=.063)$.

## Effect of Training on False-Belief Understanding

Finally, we evaluated the possible effects of pedagogical training on children's false-belief scores. Our first question was whether the training was actually effective in improving children's pedagogical skills. A repeated-measures ANOVA on the mean number of prompts children required on each of the twelve training sessions revealed a significant effect of session $(F(11,220)=4.96, p<.001)$, as well as a significant linear trend (i.e., a straight line fit the data at better than chance levels; $F(1,20)=20.85, p<.001)$. Children's performance on the pedagogical tasks thus did improve with training (see Figure 2B).

Next, we ran an independent-samples t-test comparing CT children in the training condition $(N=22)$ to those in the control condition ( $N=18$ ) on false-belief improvement. This direct comparison between training and control participants did not yield significant results ( $p=.65$ ). However, CT children who answered one false-belief question correctly at pre-test had less room for improvement. Indeed, looking only at CT children who answered zero false-belief questions correctly at pre-test, we did observe improved false-belief understanding for children in the training condition ( $N=12 ; M_{\text {improve }}=.19$, $S D=.26 ; t(11)=2.55, p=.027$ ), but not for those in the control condition ( $N=8 ; M_{\text {improve }}=.08 ; p=.170$ ); see Figure 2C. Importantly, conservation scores did not differ for either group between pre- and post-test (Training: $p=.551$; Control: $p=.197$ ), suggesting that the training targeted ToM
without necessarily leading to general improvement in cognitive reasoning. Note that this result does not directly compare training to control children, and should be interpreted with caution. However, coupled with our finding that initial false-belief understanding is related to non-trained pedagogical skill, this may suggest an important link between reasoning about others' minds and pedagogical evidence selection.

## Discussion

Past work has shown that children's developing ToM reasoning abilities are related to their pedagogical skill, but has not looked at the precise relationship between ToM development and the ability to select optimal evidence to teach others. Our results suggest that having a more developed ToM is broadly related to being better at evidence selection, even when controlling for age and more general cognitive abilities. Further, we found tentative supporting evidence for the idea that training pedagogical evidence selection may in turn improve children's ToM reasoning abilities. Taken together, our results are consistent with prior work on the relationship between ToM and teaching skills, and provide support for a strong link between pedagogical evidence selection and theory of mind.

Our results speak to existing models of ToM development that postulate genuine conceptual change during the preschool years. Specifically, we found evidence for a link between performance on a false-belief task and the discrete developmental capability of pedagogical evidence selection, suggesting that the changes in false-belief understanding that occur between the ages of 3 and 5 may reflect deeper qualitative changes in children's Theories of Mind. As we noted in the introduction, we recognize that there is a diverse range of perspectives on the course of children's ToM development, and we will not attempt to resolve that debate here. Rather, we simply suggest that our findings cannot be explained in their entirety by false-belief task demands (especially given that our results persist when controlling for effects of age), and may therefore be indicative of some type of conceptual change in ToM during the preschool years.

Our findings also have implications for current theories and models of natural pedagogy and epistemic trust. Shafto et al. (2014) propose a Bayesian model of pedagogical teaching and learning, according to which the evidence that teachers choose to present directly depends on the learner's prior knowledge and the learning goal that the teacher is trying to communicate. This pedagogical model is a special case of the broader model of epistemic trust (Shafto, Eaves, Navarro, \& Perfors, 2012; Eaves \& Shafto, 2012, in press), which explicitly connects developmental changes in reasoning about others' beliefs to interpretation of evidence selection by others. Our results support these models that link evidence selection and reasoning about other minds. We also extend their findings, showing that this link 1) exists even in young children who have not yet been exposed to formal schooling, and 2) is manifest in their selection of evidence for others.

Along with this prior work, our paper also speaks to broader theories of natural pedagogy, and supports a potential link between the uniquely human ability to teach others and the development of the ability to reason about others' minds; this raises questions about whether these skills may even be evolutionarily intertwined. Whatever the case may be, reasoning about other minds, as conceptualized in the field, is composed of multiple interrelated inference problems. Understanding the role of these social inferences in learning requires investigating how children approach several conjoined problems, as we have done. Our work exemplifies and shows the value of that approach.

## References

Ashley, J., \& Tomasello, M. (1998). Cooperative problemsolving and teaching in preschoolers. Soc Dev, 7(2), 143163.

Astington, J. W., \& Jenkins, J. M. (1995). Theory of mind development and social understanding. Cogn Emo, 9(2-3), 151-165.
Baillargeon, R., Scott, R. M., \& He, Z. (2010). False-belief understanding in infants. Trends Cogn Sci, 14(3), 110-118.
Bartsch, K., \& Wellman, H. (1989). Young children's attribution of action to beliefs and desires. Child Dev, 60(4), 946-964.
Davis-Unger, A. C., \& Carlson, S. M. (2008). Development of teaching skills and relations to theory of mind in preschoolers. J Cogn and Dev, 9(1), 26-45.
Eaves, B., \& Shafto, P. (2012). Unifying Pedagogical Reasoning and Epistemic Trust. In F. Xu \& T. Kushnir (Eds.), Advances in child development and behavior (pp. 295319). Elsevier.

Eaves, B., \& Shafto, P. (in press). Parameterizing developmental changes in epistemic trust. Psychon Bull Rev.
Goodman, N. D., Baker, C. L., Bonawitz, E., Mansinghka, V. K., Gopnik, A., Wellman, H., ... Tenenbaum, J. (2006). Intuitive theories of mind: A rational approach to false belief. Proc 28th Annual Conf Cogn Sci Soc.

Gopnik, A., \& Astington, J. W. (1988, February). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. Child Dev, 59(1), 26-37.
Gweon, H., Tenenbaum, J., \& Schulz, L. (2010). Infants consider both the sample and the sampling process in inductive generalization. PNAS, 107(20), 9066-9071.
Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., \& Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. Dev Psychol, 52(8), 1192-1205.
Knudsen, B., \& Liszkowski, U. (2012). One-year-olds warn others about negative action outcomes. J Cogn Dev, 14(3), 424-436.
Kushnir, T., \& Koenig, M. (in press). What I don't know won't hurt you: The relation between professed ignorance and later knowledge claims. Dev Psychol.
Moll, H., \& Tomasello, M. (2007). Cooperation and human cognition: The Vygotskian intelligence hypothesis. Philos Trans R Soc Lond B Biol Sci, 362(1480), 639-648.
Nurmsoo, E., \& Robinson, E. J. (2009). Children's trust in previously inaccurate informants who were well or poorly informed: When past errors can be excused. Child Dev, 80(1), 23-27.
Pasquini, E. S., Corriveau, K. H., Koenig, M., \& Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. Dev Psychol, 43(5), 1216-1226.
Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., \& Caglar, L. (2015). Controlling the message: Preschoolers' use of information to teach and deceive others. Front Psychol, 6, 867.

Rhodes, M., Gelman, S. A., \& Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. Dev Sci, 13(3), 421-429.
Shafto, P., Eaves, B., Navarro, D. J., \& Perfors, A. (2012, February). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. Dev Sci, 15(3), 436447.

Shafto, P., Goodman, N. D., \& Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cogn Psychol, 71, 55-89.
Strauss, S., Ziv, M., \& Stein, A. (2002). Teaching as a natural cognition and its relations to preschoolers' developing theory of mind. Cogn Dev, 17(3), 1473-1487.
Wellman, H. M., Cross, D., \& Watson, J. (2001). Metaanalysis of theory-of-mind development: The truth about false belief. Child Dev, 72(3), 655-684.
Wimmer, H., \& Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1), 103-128.
Ziv, M., \& Frye, D. (2004). Children's understanding of teaching: The role of knowledge and belief. Cogn Dev, 19(4), 457-477.

# Didn't know, or didn't show? Preschoolers consider epistemic state and degree of omission when evaluating teachers 

Ilona Bass (ilona.m.bass@rutgers.edu) ${ }^{1}$, Elizabeth Bonawitz (elizabeth.bonawitz@ rutgers.edu) ${ }^{1}$, Hyowon Gweon (hyo@stanford.edu) ${ }^{2}$<br>${ }^{1}$ Department of Psychology, Rutgers University, Newark, NJ 07102<br>${ }^{2}$ Department of Psychology, Stanford University, Stanford, CA 94305


#### Abstract

The ability to recognize and evaluate reliable informants is a critical skill for effective social learning. Building on prior work showing children's sensitivity to informants who omit relevant information, here we asked whether children's teacher evaluations incorporate information about 1) the epistemic state of the teacher, and 2) the amount and value of information taught. Preschool-aged children rated informants who taught learners about a novel toy with four functions; we systematically varied the number and value of functions the teachers knew and taught. Our results indicate that children exonerated unintentional omissions of teachers who had incomplete knowledge, and provided graded ratings based on the degree of omission. These findings are consistent with the predictions of prior computational work, and suggest that the ability to reason about others' knowledge plays an important role in children's inferences about others' efficacy as informants.


Keywords: cognitive development, pedagogy, social learning

## Introduction

Young children rely heavily on others for their learning. Although children readily explore and learn from their own experience (Schulz, 2012; Bonawitz, van Schijndel, Friel, \& Schulz, 2012; Stahl \& Feigenson, 2015), pedagogy is a powerful, effective way to learn about the world. Recent research suggests that children do more than simply absorb and accumulate information from others; they actively modulate their inferences depending on the social context (Bonawitz et al., 2011), and selectively approach others to request information when help is needed (Gweon \& Schulz, 2011; Goupil, Romand-Monnier, \& Kouider, 2016). However, learning from pedagogy comes with an inherent hazard: being misinformed. Informants may vary in quality - some may be wrong, ignorant, or even deceptive. Thus, the ability to detect and evaluate unhelpful informants is critical for accurate learning. How do young children face this challenge?

Prior research has found that children avoid learning from informants who provide inaccurate information (e.g., Birch, Vauthier, \& Bloom, 2008; Jaswal \& Neely, 2006; Koenig, Clément, \& Harris, 2004; Pasquini, Corriveau, Koenig, \& Harris, 2007). Recent studies further suggest that young children recognize and evaluate a more subtle form of misinformation: providing accurate yet insufficient evidence. Given a teacher who presented one function on a toy, children rated the teacher as more helpful when the toy only had one function than when it had four (i.e., when the teacher omitted 3 of the 4 functions; Gweon, Pelton, Konopka, \& Schulz, 2014a; Gweon \& Asaba, in press). Children as young as four show this sensitivity, although they successfully evaluate underinformative teachers only after observing a fully informative
teacher (Gweon \& Asaba, in press). Thus by the preschool years, children expect teachers to be accurate and fully informative, and penalize those who violate these expectations.

This early-emerging sensitivity to teacher informativeness raises important questions about how children make these evaluations: What are the representations and inferences that allow children to distinguish helpful and less helpful teachers? One possibility is that children learn sets of rules and exceptions that allow them to recognize and avoid undesirable teachers. Prior findings suggest that young children are biased towards trusting adult informants, and may even continue to trust them after discovering their unreliability (Jaswal, Croft, Setia, \& Cole, 2010). Children may also acquire a set of rules akin to Gricean Maxims (Grice, 1975), which prescribe that a helpful, cooperative communicator should provide accurate and relevant information in the right amount. If children are simply using learned heuristics or rules to evaluate informants, it may be difficult for them to make nuanced, context-specific judgments of informant quality, particularly in novel situations. However, another possibility is that these evaluations arise from sophisticated inferences about teacher informativeness; by understanding how unobservable mental states of others (e.g., informants' intent or knowledge) can influence their teaching behaviors, children can draw much more flexible and accurate informant evaluations even in novel contexts.

Previous work on Theory of Mind and moral reasoning suggests that young children readily interpret others' observable actions in light of their unobservable mental states: They evaluate others' actions based on their outcomes and on the actor's underlying intent, exonerating accidental harms (e.g., Cushman, Sheketoff, Wharton, \& Carey, 2013; Wellman, Cross, \& Watson, 2001; Nelson, 1980; Baird \& Astington, 2004). Furthermore, even toddlers exonerate an agent who refused to help another person when the agent was incompetent and thus unable to help the requester (Jara-Ettinger, Tenenbaum, \& Schulz, 2015). Given prior work on children's ability to consider others' mental states in evaluating others, here we ask whether children can consider informants' knowledge and their competence in evaluating their teaching.

Prior computational work describes teacher-learner interactions as based on a set of mutually constraining inferences. The teacher considers the learner's knowledge to select the evidence that would maximally increase the learner's belief in the correct hypothesis. The learner updates his beliefs with the assumption that the teacher is knowledgeable and intends
to provide the best information for the learner (Shafto, Goodman, \& Frank, 2012; Shafto, Goodman, \& Griffiths, 2014). In this framework, a teacher can be evaluated based on how she samples information for the learner, and what the learner can infer from the information.

This allows us to consider two key hypotheses about what might influence children's evaluations of teachers. Consider a teacher demonstrating a device with four functions (some interesting, some humdrum) to a naïve learner. How might a rational observer evaluate the teacher, based on what she demonstrates? First, we might predict that an evaluation of an informant is sensitive to the epistemic state of that informant. For example, consider two teachers, each of whom demonstrates just one of the four functions. One teacher knows that the device has four functions, but the other only knows about the one function she demonstrated. While the learner only learns about one of the four functions in both cases, we might be more inclined to pardon the teacher who didn't know about the additional functions: The ignorant teacher demonstrated everything she knew, and may thus be considered a better teacher than the knowledgeable informant who omitted information. We refer to this as the epistemic pardon hypothesis.

Our second hypothesis pertains to the quality of the taught information. A teacher who knows all four functions of a device will be most helpful if she demonstrates all four, and least helpful if she demonstrates none. Extending this reasoning to partial demonstrations, we would predict evaluations to be modulated by the degree of omission: Even when two teachers both omit information, a teacher who demonstrates two functions is still better than someone who showed just one. Further, if the functions differ in their value (e.g., how interesting they are), we might also expect an effect of the value of demonstrated functions: A teacher who demonstrates two high-value functions and omits two low-value functions would be better than someone who does the opposite. We refer to these predictions as the quality-of-omission hypothesis.

Recent computational work has formalized the two hypotheses posited above, and shown that adults' evaluations of various teachers are highly consistent with these hypotheses (Bass, Hawthorne-Madell, Goodman, \& Gweon, 2015). When adults evaluate informant quality, they readily incorporate information about a teacher's epistemic state, as well as the amount and the value of taught information. Adults' informant evaluations are thus likely based on abstract representations of others' minds rather than a set of rules that dictate what a teacher should or should not do.

Some prior work suggests that children's evaluations of teachers also depend on abstract representations of knowledge states rather than simple heuristics. For instance, children show increased exploration of a toy following a teacher's demonstration of that toy if the teacher had previously committed a sin of omission (Gweon et al., 2014a), suggesting that children use concrete demonstrations to infer abstract qualities of teachers' quality, and adjust their inferences accordingly. Children also understand that omission isn't al-
ways bad: Given a toy with 20 buttons but only 3 that are functional, children prefer a teacher who shows just the 3 functional buttons (as opposed to the one who additionally shows the 17 inert buttons), if the learner already expects only a few of the buttons to work (Gweon, Shafto, \& Schulz, 2014b). Children thus readily consider learners' epistemic states to evaluate teacher helpfulness, and even judge omission as beneficial when partial demonstration is sufficient. However, these studies leave open a critical question: Can children consider the teacher's epistemic state in evaluating the helpfulness of their teaching? Going beyond recognizing that teachers might not know everything (Jaswal \& Neely, 2006), can children actually use this information to exonerate under-informative pedagogy? Because children are surrounded by many adults who are much more knowledgeable than they are, this may be a particularly challenging inference for young children.

## Preschoolers' Evaluations of Teachers

In the current study, we investigate whether preschool-aged children's teacher evaluations reflect the underlying representations of teachers' knowledge and competence; in particular, we ask whether each of our two hypotheses (epistemic pardon, and quality-of-omission) - both of which are consistent with adults' teacher evaluations (Bass et al., 2015) are also consistent with children's ratings of teacher quality. We showed children videos of five different informants who taught learners about a novel toy with four functions. We systematically varied the number and value of functions that the teachers knew and taught, and randomized the order in which the five teachers were seen with one caveat: All children first saw the teacher who knew and taught all four of the toy's functions, and were told that this was an example of excellent teaching. Our decision to anchor children's responses in this way was motivated by prior findings: First, children reliably rate teachers highly when they provide true and complete information (e.g., Gweon et al., 2014a; Koenig \& Harris, 2005); second, although four- and five-year-olds' ability to evaluate under-informative teachers is limited, seeing an example of a fully informative teacher first allows them to successfully evaluate under-informative teachers (Gweon \& Asaba, in press). These results suggest that such contextual support helps children attend more closely to dimensions of teacher informativeness. Since we are interested in children's ratings of several under-informative teachers relative to each other (and not to the fully informative teacher), we anchored children's ratings of this ideal teacher at the top of the scale.

## Methods

## Participants

Thirty-four children $\left(M_{\text {age }}=60\right.$ months, range $=49-72$ months; 15 females) were tested at local preschools.

## Materials

Rating Scale Children used a 0 to 20 point rating scale to evaluate teachers. Children placed a small circular magnet on


Figure 1: All children saw and rated the KA_TA teacher first. The order of the remaining four test trials was counterbalanced across children. Memory cues were adhered to the rating scale as children provided their ratings.
the scale to indicate how good they thought a teacher was.
Novel Toy The novel toy was a square pyramid covered in blue felt with four colorful buttons, each corresponding to a different function. Two functions were low-value: The toy could beep, and it could make a static-like noise. The other two functions were high-value: The toy could play clips of two different children's songs. The relative value of these functions were validated in a separate group of 10 children $\left(M_{\text {age }}=66\right.$ months, range $=49-91$ months $)$, who were asked to rate "how cool" each of the four functions were using the rating scale described above. The two songs ( $M=15.3$, $S D=3.1$ ) were rated significantly higher than the beep and the noise ( $M=8.8, S D=3.7 ; t(9)=3.41, p=.008$ ), with no differences within the value pairs ( $p$ 's $>.33$ ).

Teaching Videos Teaching videos were presented on a 15inch MacBook Pro, and comprised two main phases. In the Exploration phase, the teacher sat down at a table on which the novel toy was placed and explored the toy's functions; then, a naïve learner suddenly entered the room, startling the teacher out of her exploration, and asked her to show him how the toy worked. In the Teaching phase, the teacher demonstrated to the learner some subset of the functions she had discovered during the exploration phase (details follow), after which she said, "That's how this toy works!" thereby clearly ending the demonstration.

There were five versions of the teaching videos, which varied based on the number and value of the functions that the teacher discovered and taught. In the Exploration phase, the teacher either discovered 1) all four functions, or 2) just the low-value "beep" function before the learner entered the room. In the Teaching phase, the teacher either taught: 1) all four functions, 2) both high-value functions, 3) both lowvalue functions, or 4) just one low-value function ("beep"). Crossing these two variables yielded five possible teaching scenarios: KA_TA, in which the teacher Knew All and Taught

All; KA_THH, where she Knew All and Taught 2 Highvalue functions; KA_TLL, in which she Knew All and Taught 2 Low-value functions; KA_TL, where she Knew All and Taught 1 Low-value function ("beep"); and KL_TL, where she Knew 1 Low-value function ("beep") and taught it.
Memory Cues To help children recall precisely what each teacher knew and taught, we created small cards that depicted screenshots of the Exploration and Teaching phases from the teaching videos. Small arrows with adhesive backs were attached to each memory cue (see Figure 1).

## Procedure

Frame Story \& Rating Scale Training Children were told that they would be meeting some people who were in teaching school; the experimenter needed the child's help to figure out how good the different teachers were so that she would know how much more school the teachers needed. The experimenter then introduced the rating scale, and children were briefly trained on how to use it to indicate teacher quality. Children who failed this training did not proceed to the main task and were dropped from analysis (see Results).
Novel Toy Next, the experimenter introduced the novel toy, and encouraged children to try to figure out how it worked. After the child successfully pressed all four buttons, the experimenter noted that they now knew all about the toy. Children were then told that the other day, the teachers from the school had taught some new students about how the novel toy worked, and it was the child's job to watch them teach about the toy and figure out how good each teacher was at teaching. Teacher Evaluations All children were first shown the KA_TA condition. Before watching the video, children were told that this teacher was all done with school, and was therefore already a good teacher. After watching the first video, children were shown the memory cue for the KA_TA teacher,
and were asked to provide a rating. Children who did not place the marker near the top of the rating scale were reminded that this teacher was already done with school. The experimenter adhered the memory cue's arrow to the rating scale where the child had placed the marker.

Children were then shown the remaining four teachers who, they were reminded, were still in teaching school; these constituted the four test trials. The order in which these four teaching scenarios were presented was completely counterbalanced, yielding 24 different orders. While the actors in the videos and the test conditions were fully counterbalanced with respect to each other, such that any potential effects of teaching condition could not be explained by personal characteristics of the actors, the order of the actors was always the same (e.g., "Liz" was always the first teacher, even though the first test condition varied between participants). After watching each video, children were shown the memory cue for the teacher they had just seen, and were asked to provide a rating. The experimenter adhered the memory cue's arrow to the rating scale where the child had placed the marker. Children who merely placed each teacher on the scale in the order of presentation (from most helpful to least helpful) were excluded from analysis. See Figure 1 for a schematic of the procedure for teacher evaluations.

## Results

Prior to analysis, we dropped children who did not pass the rating scale training $(N=1)$, placed teachers in descending order on the rating scale as they saw them $(N=7)$, or gave all teachers the same rating $(N=1)$. One additional child did not want to continue playing after the first KA_TA trial. Our final sample therefore consisted of 24 children ( $M_{\text {age }}=60$ months, range $=49-72$ months; 12 females).

We first asked whether children differentiated between the four teachers in the test trials. An omnibus repeated-measures ANOVA on children's ratings in these four trials revealed a significant main effect of condition $(F(3,69)=3.50, p=$ $.020, \eta_{\mathrm{p}}^{2}=.132$; see Figure 2). We therefore conducted followup analyses to investigate our two stated hypotheses.

## Epistemic Pardon Hypothesis

To investigate the effect of the teacher's epistemic state on children's ratings, we compared the KA_TL condition to the KL_TL condition, thereby holding constant what the teacher taught and only varying what she knew. A paired-samples t -test revealed significant differences between the ratings of these two teachers $\left(t(23)=2.58, p=.017, \eta_{\mathrm{p}}^{2}=.224\right)$, with children giving higher ratings to the teacher who knew only one function ( $M=11.5, S D=5.7$ ) than the teacher who knew all four functions but taught just one ( $M=7.8, S D=5.4$ ).

We also looked at the number of children who placed the KL_TL teacher higher than the KA_TL teacher on the rating scale. Seventy-one percent of participants rated the KL_TL teacher higher than the KA_TL teacher; this proportion differed significantly from chance ( $50 \%, p=.032$ one-tailed),


Figure 2: Average ratings for the KA_TA reference teacher and all four test conditions. Children rated the teacher who knew only one function (KL_TL) higher than the teacher who knew all but taught one (KA_TL), pardoning omission when it occurred for epistemic reasons. Children also showed sensitivity to the degree of omission, rating the teacher who demonstrated two low-value functions (KA_TLL) as better than the teacher who demonstrated one (KA_TL).
providing additional evidence that children considered teachers' epistemic states when making their evaluations, and were even able to exonerate bad teaching when it was explained by limited knowledge.

## Quality-of-Omission Hypothesis

We explored the effect of the degree of teachers' omission of information on children's ratings with a paired-samples ttest, comparing the KA_TLL condition to the KA_TL condition (varying the number of functions taught while holding epistemic state and value constant). We again found significant differences $\left(t(23)=2.54, p=.019, \eta_{\mathrm{p}}^{2}=.218\right)$ : Children gave higher ratings to the teacher who demonstrated two low-value functions ( $M=11.9, S D=6.4$ ) than the teacher who demonstrated just one low-value function ( $M=7.8$, $S D=5.4$ ). As before, we also compared the proportion of children who rated the KA_TLL teacher higher than the KA_TL teacher to chance. This binomial test neared significance ( $p=.076$ one-tailed), with $67 \%$ of children rating the KA_TLL teacher higher than the KA_TL teacher.

Finally, we compared the KA_TLL teacher to the KA_THH teacher to examine the effect of information value on children's ratings. This paired-samples t-test was not significant ( $p=.874$ ): Children did not differentiate between teachers who taught two high-value ( $M=12.1, S D=5.2$ ) versus two low-value ( $M=11.9, S D=6.4$ ) functions. Possible explanations for this null result follow in the discussion. ${ }^{1}$

[^20]
## Discussion

Inspired by computational models of pedagogy and prior behavioral work with adults (e.g., Shafto et al., 2014; Bass et al., 2015), here we investigated how children make nuanced evaluations of helpful and unhelpful teachers; specifically, we asked whether children 1) exonerate partial teaching based on the teacher's epistemic state, and 2) provide graded evaluations based on the amount and value of information taught. We found that, like adults, preschoolers were sensitive to teachers' epistemic states, and accordingly pardoned informants who provided less information when teaching from limited knowledge. Children's ratings were also sensitive to the amount (but not the value) of information taught.

The results from our epistemic comparisons extend prior work showing that children prefer truthful teachers (Koenig et al., 2004; Koenig \& Harris, 2005; Jaswal \& Neely, 2006), and fully informative teachers (Gweon et al., 2014a; Gweon \& Asaba, in press). They are also consistent with more recent findings on children's ability to consider learners' epistemic states (Gweon et al., 2014b) in evaluating teachers. However, our findings are somewhat surprising in light of the idea that many explicit Theory of Mind (ToM) skills are just developing between the ages of three and five (Wellman et al., 2001). Without explicit information about what the teacher knew, preschoolers were able to 1) infer her epistemic state by observing her exploration, and 2) use this representation to pardon her "sin of omission".

This finding thus raises important questions about the relationship between the development of ToM reasoning and social evaluation in pedagogical contexts. If ToM does in fact modulate children's teacher evaluations, children may become more adept at selecting from whom to learn throughout their preschool years. Indeed, Jaswal et al. (2010) found that three-year-olds are almost indiscriminately trusting of informants, while older children are more wary of possible misinformation. It would be interesting to ask whether children who are better at ToM reasoning also consider teachers' epistemic states more readily, leading them to be more willing than children with less proficient ToM abilities to exonerate teachers who were unintentionally under-informative. Critically, given recent findings on the relationship between ToM and children's own teaching skills (Bass et al., in press), such results would support important links between theory of mind, pedagogical skill, and teacher evaluations.

Note that although children did exonerate the KL_TL teacher relative to the KA_TL teacher, no under-informative teacher was rated as favorably as the informant who knew and taught all four of the toy's functions. Intuitively, this makes sense: Children's ratings of an informant's helpfulness will reflect, among other things, how well a learner learned as a consequence of the informant's teaching. Thus while the KL_TL teacher did the best she could given what she knew, she was still not as good of a teacher as the KA_TA teacher because she failed to discover information that could have been useful for the learner. This intuition also naturally arises
in adults' teacher evaluations, and is consistent with Bayesian models of pedagogical reasoning (e.g., Bass et al., 2015). Are there circumstances under which under-informative teaching can be fully exonerated? In ongoing work, we are exploring whether the degree to which children exonerate underinformative teachers is modulated by contexts that explain away the teacher's failure to discover relevant functions and resultant lack of knowledge (e.g., a broken toy).

Our results also show that children did not penalize all omissions equally. Even though all teachers were underinformative, children were sensitive to the "degree of omission," giving lower evaluations to teachers who provided less information. This extends prior work showing that children distinguish fully informative teachers from those who were vastly under-informative (Gweon et al., 2014a; Gweon \& Asaba, in press), and further suggests that children's evaluations of under-informative teachers are based on a more nuanced understanding of teachers' behaviors than a simple binary judgment. This leaves open questions about the nature of the mechanisms that underlie sensitivity to informant quality more generally: How early do they emerge? What other factors can children incorporate into their informant evaluations, and how do these change as children develop?

Our work adds to the growing body of literature on children's ability to draw pragmatic inferences from others' behaviors in both verbal and nonverbal communication. Recent work has demonstrated intriguing parallels between children's evaluations of pedagogical informants and their ability to draw scalar implicature (Gweon \& Asaba, in press). Given prior work on scalar implicature that reveals children's ability to evaluate infelicitous uses of quantifiers (Barner, Brooks, \& Bale, 2011; Katsos \& Bishop, 2011), our results further suggest that children as young as four might have the necessary prerequisites for considering the "degree of sin" in infelicitous scalar expressions (e.g., it is worse to say that the boy drank "a bit" of milk than to say he drank "some" milk, when really he drank almost all the milk in the cup).

Finally, we note that children's ratings in the current study were not moderated by the value of the demonstrated functions: Children rated a teacher who chose to show the two lower-value functions just as highly as the teacher who showed two higher-value functions. These results differ from adults' sensitivity to information value in a highly similar paradigm (Bass et al., 2015). There are several possible explanations for this null finding. First, the relative value of the toy's functions in our study may not have been salient enough to elicit this difference. While we did validate the functions' values in a separate group of participants, those children were explicitly asked to compare and consider the functions' "coolness"; for children in the current study, these subtle value differences may not have been conspicuous enough to differentiate teachers who taught the songs versus the noises. A second possibility is that the ability to consider the value of information in service of making pedagogical evaluations does not emerge until later in development. This would suggest that
although children show remarkable ability in evaluating others, there may be other important factors that young children fail to consider. Third, it is possible that children are capable of considering information value (and that the functions' values were sufficiently salient in our task), but that children spontaneously attributed a reason for why the informant selected these functions; for instance, perhaps the low-value teacher really liked those functions, or thought they would be more important for the learner to know. Future work could tease apart these hypotheses to identify the role of information value in children's informant evaluations.

As we have discussed, there are many unanswered questions concerning the nature of children's reasoning about pedagogical informants that our results do not directly address. Nevertheless, along with prior work, our findings suggest that young children do have abstract representations of what it means to be a good teacher. Understanding the development of children's epistemic trust and its relationship to their growing ability to reason about others' minds will provide further insight into the cognitive mechanisms that support the uniquely human abilities to learn from and teach others.

## References

Baird, J. A., \& Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. New Dir Child Adolesc Dev, 2004(103), 37-49.
Barner, D., Brooks, N., \& Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. Cognition, 118(1), 84-93.
Bass, I., Bonawitz, E., Shafto, P., Ramarajan, D., Gopnik, A., \& Wellman, H. (in press). I know what you need to know: Children's developing theory of mind and pedagogical evidence selection. Proc 39th Cogn Sci Soc.
Bass, I., Hawthorne-Madell, D., Goodman, N., \& Gweon, H. (2015). Not by number alone: The effect of teachers' knowledge and its value in evaluating "sins of omission". Proc 37th Cogn Sci Soc.
Birch, S., Vauthier, S., \& Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. Cognition, 107(3), 1018-1034.
Bonawitz, E., Shafto, P., Gweon, H., Goodman, N., Spelke, E., \& Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. Cognition, 120(3), 322-330.
Bonawitz, E., van Schijndel, T., Friel, D., \& Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. Cogn Psychol, 64(4), 215-234.
Cushman, F., Sheketoff, R., Wharton, S., \& Carey, S. (2013, April). The development of intent-based moral judgment. Cognition, 127(1), 6-21.
Goupil, L., Romand-Monnier, M., \& Kouider, S. (2016). Infants ask for help when they know they don't know. PNAS, 113(13), 3492-3496.

Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and semantics: Speech acts (Vol. 3, pp. 41-58). New York: Academic Press.
Gweon, H., \& Asaba, M. (in press). Order matters: Childrens evaluation of under-informative teachers depends on context. Child Dev.
Gweon, H., Pelton, H., Konopka, J., \& Schulz, L. (2014a). Sins of omission: Children selectively explore when teachers are under-informative. Cognition, 132(3), 335-341.
Gweon, H., \& Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. Science, 332(6037), 1524.
Gweon, H., Shafto, P., \& Schulz, L. (2014b). Children consider prior knowledge and the cost of information both in learning from and teaching others. Proc 36th Cogn Sci Soc.
Jara-Ettinger, J., Tenenbaum, J., \& Schulz, L. (2015). Not so innocent: Toddlers' inferences about costs and culpability. Psychol Sci, 26(5), 633-640.
Jaswal, V., Croft, A., Setia, A., \& Cole, C. (2010). Young children have a specific, highly robust bias to trust testimony. Psychol Sci, 21, 1541-1547.
Jaswal, V., \& Neely, L. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. Psychol Sci, 17(9), 757-758.
Katsos, N., \& Bishop, D. (2011, July). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. Cognition, 120(1), 67-81.
Koenig, M., Clément, F., \& Harris, P. (2004). Trust in testimony: Children's use of true and false statements. Psychol Sci, 15(10), 694-698.
Koenig, M., \& Harris, P. (2005). Preschoolers mistrust ignorant and inaccurate speakers. Child Dev, 76(6), 12611277.

Nelson, S. A. (1980). Factors influencing young children's use of motives and outcomes as moral criteria. Child Dev, 51(3), 823-829.
Pasquini, E., Corriveau, K., Koenig, M., \& Harris, P. (2007). Preschoolers monitor the relative accuracy of informants. Dev Psychol, 43(5), 1216-1226.
Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. Trends Cogn Sci, 16(7), 382-389.
Shafto, P., Goodman, N., \& Frank, M. (2012). Learning from others: The consequences of psychological reasoning for human learning. Perspect Psychol Sci, 7(4), 341-351.
Shafto, P., Goodman, N., \& Griffiths, T. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cogn Psychol, 71, 55-89.
Stahl, A., \& Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. Science, 348(6230), 91-94.
Wellman, H. M., Cross, D., \& Watson, J. (2001). Metaanalysis of theory-of-mind development: The truth about false belief. Child Dev, 72(3), 655-684.

# Modeling human categorization of natural images using deep feature representations 

Ruairidh Battleday<br>Helen Wills Neuroscience Institute; University of California, Berkeley, Berkeley, California, USA<br>Joshua Peterson<br>Department of Psychology, University of California, Berkeley, Berkeley, CA, USA<br>Tom Griffiths<br>Department of Psychology, University of California, Berkeley, Berkeley, CA, USA


#### Abstract

Over the last few decades, cognitive scientists have developed sophisticated formal models of human categorization, and computer vision researchers have achieved increasingly impressive performance in natural image classification. In this paper, we combine the strengths of these approaches, using representations from a convolutional neural network to evaluate cognitive models of categorization against $>300,000$ human judgments of natural images. We find that a prototype model performs best overall, and that an exemplar model performs best when the network's most abstract features are used. Altogether, our results demonstrate that the optimal categorization strategy over a set of stimuli is deeply linked to how they are represented, suggesting that any satisfying characterization of categorization behavior over naturalistic stimuli must consider it the result of a dual process of feature learning and strategy selection. The paradigm we present herein offers one avenue to begin this undertaking.


# Semantic Typology and Parallel Corpora: Something about Indefinite Pronouns 

Barend Beekhuizen<br>Department of Computer Science<br>University of Toronto<br>barend@cs.toronto.edu

Julia Watson<br>Department of Computer Science<br>University of Toronto<br>j.watson@mail.utoronto.ca

Suzanne Stevenson<br>Department of Computer Science<br>University of Toronto<br>suzanne@cs.toronto.edu


#### Abstract

Patterns of crosslinguistic variation in the expression of word meaning are informative about semantic organization, but most methods to study this are labor intensive and obscure the gradient nature of concepts. We propose an automatic method for extracting crosslinguistic co-categorization patterns from parallel texts, and explore the properties of the data as a potential source for automatically creating semantic representations for cognitive modeling. We focus on indefinite pronouns, comparing our findings against a study based on secondary sources (Haspelmath 1997). We show that using automatic methods on parallel texts contributes to more cognitively-plausible semantic representations for a domain.


Keywords: semantic typology; semantic representation; parallel corpora

## Introduction

An important goal of cognitive science is to determine valid semantic representations, e.g., for use in computational cognitive models of language acquisition and processing. Semantic typology - which studies the patterns of crosslinguistic variation in what words and other linguistic elements mean - reveals universal tendencies in how languages carve up the space of a semantic domain (Haspelmath, 2003; Regier, Kemp, \& Kay, 2015). In particular, Bowerman (1993) argues that (all else being equal) the greater the number of languages that label a pair of situations (objects, events, ...) with the same word (called co-categorization), the more conceptually similar these situations are. For instance, many languages co-categorize situations of 'stable support' (see Fig. 1) with those of 'tenuous support', but use a different term for 'containment', reflecting that the first two situations are more semantically similar than the last.

More generally, such crosslinguistic co-categorization patterns can define a geometric semantic similarity space (Levinson et al. 2003). To obtain such a space, we first represent a situation as a vector of terms used to express that situation across languages (cf. the row of terms in Fig. 1). These vectors are then projected into a lower-dimensional space (cf. the distances in the two-dimensional space of Fig. 1). This insight has informed cognitive modeling work on spatial relations (Beekhuizen, Fazly, \& Stevenson, 2014) and color (Beekhuizen \& Stevenson, 2016), where descriptions of situations, elicited from speakers of a number of languages, were used to create vector-based geometric semantic representations. A computational learning model trained on those representations successfully simulated developmental error patterns in word meaning acquisition.

In order to deploy such approaches to additional semantic domains, we need practical and robust methods for semantic typological analysis. Elicitation data (e.g., Berlin \& Kay,


Figure 1: Representing the conceptual distance between situations as the number of languages co-categorizing them.

1969; Bowerman \& Pederson, 1992) - terms describing nonlinguistic stimuli, obtained from informants - allows control in defining the set of stimuli for a domain, but is resource intensive and limited to concrete domains. Expert (Haspelmath, 1997) or automatic (Youn et al., 2016) analyses of secondary sources (such as dictionaries and grammars) don't rely on access to informants across many languages, but are focused on coarser-grained semantic distinctions than are found in elicitation stimuli. Both of these methods lack the frequencies and patterns of actual usages in natural communication.

A complementary source of data recently used in semantic typology is parallel text (e.g., Cysouw \& Wälchli, 2007) i.e., the same text translated into different languages. While parallel text has its own potential disadvantages, such as a risk of "translationese" or mistranslations (Levshina, 2017), it can be applied to abstract domains that are hard to obtain elicitation data for, and has actual usage tokens that can reveal nuances of meaning not captured in dictionaries. This latter point is especially relevant for creating semantic spaces for cognitive modeling, as semantic categories display prototype structures with more and less central members (Rosch, 1973). Deriving semantic representations from actual usages can yield a continuous semantic similarity space which potentially reflects such structures; training computational cognitive models on such representations thus has the potential to better match behavioural data. To exploit this potential of parallel text, we need automatic methods for extracting the co-categorization patterns - the terms used across multiple languages for the same situation (cf. Fig. 1) - that can form the basis for such vector-based representations.

In this paper we first propose an automatic method for extracting crosslinguistic co-categorization patterns from parallel texts, to complement elicitation data and secondary sources. Next, we explore the properties of the resulting data as a potential source for automatically creating semantic spaces for cognitive modeling. We focus on indefinite

| Acronym | Semantic function | Example |
| :--- | :--- | :--- |
| SP-K | specific, known | I want to tell you something. |
| SP-U | specific, unknown | Someone broke into our apartment. |
| NS | irrealis non-specific | I need someone strong for the job. |
| CD | conditional | Let me know if anybody shows up. |
| QU | question | Is anything bothering you? |
| IN | indirect negation | I don't think anything matters. |
| DN | direct negation | Nobody came. |
| CP | comparison | She can run faster than anybody. |
| FC | free choice | You can pick anything! |

Table 1: Haspelmath's 9 functions with examples.
pronouns as an abstract semantic domain for which elicitation would be a difficult method, but for which we have a good understanding of the typology from expert judgments and secondary sources (Haspelmath, 1997). By using parallel texts, we are able to get a fuller picture of the semantic structure of this domain, in particular seeing evidence for gradience in multiple ways: finer-grained semantic functions that show gradient patterns across languages, and gradient relationships (distances) among the semantic functions. We thus show that using automatic methods on this complementary data source can contribute to more cognitively-plausible semantic representations, by fleshing out expert analysis of secondary sources with usage data that reflects the discourse use and frequency of the semantic functions.

## Indefinite pronouns

Indefinite pronouns, such as somebody, anything, and nowhere, are used to express indefinite reference - i.e., introducing a discourse referent which the speaker typically does not intend the hearer to uniquely identify. Reference may be to an entity from any of the major ontological categories such as PEOPLE, THINGS, and Places. Haspelmath (1997) outlines 9 semantic functions that indefinite pronouns can express; see Table 1. To identify the set of functions, he draws on semantic motivation - whether a coherent functional definition can be established for each. Importantly, linguistic evidence is considered for deciding whether two related functions should be merged or split: specifically, if at least one language has a term that can be used for only one of the functions - i.e., if there is a language with a term that does not co-categorize the two - then the two functions are considered distinct.

The identified semantic functions are analogous to stimuli in an elicitation task, although at a coarser grain: each function represents a set of situations that are co-categorized. Like elicitation data, terms in each language are associated with each of the semantic functions they can express, and patterns of crosslinguistic co-categorization can be revealed. These patterns can be visualized in a graphical semantic map: functions (nodes) are connected by edges such that connected subgraphs correspond to sets of functions that can be cocategorized. The semantic map of Haspelmath (1997), in Fig. 2, shows that, in both example languages, the terms carve out different, but in both cases connected, partitionings of the graph.


Figure 2: Semantic map from Haspelmath (1997) with English and Nanay terms.

Despite the insight they provide, semantic maps do not capture certain properties of the underlying semantic space that are important to semantic representation. Two related issues in particular motivate our work here. First, there is no indication of the distance in semantic space that an edge in the map represents, although it is likely that some functions connected to a node may be closer or further semantically than others. For example, although IN connects to both DN and CP by a single edge in Fig. 2, it is likely more similar to DN. Second, the use of a single node for a function assumes (instrumentally) that functions are internally homogeneous. However, functions may display a gradient internal structure - e.g., some cases of DN may be 'better' instances than others. Both of these factors may contribute to the cognitive plausibility of a semantic space for use in computational modeling.

As discussed above, parallel usage data has the potential to address these issues by providing a more continuous representation than secondary data. Actual usage data may reveal how related Haspelmath's various functions are, and how homogeneous they are internally. Such insights are crucial for the use of semantic-typological analyses in cognitive science, e.g., in modeling the acquisition of such terms.

## Method: Translations from Parallel Text

Our goal is to construct geometric semantic spaces through the use of parallel (translated) usage data. We draw on the patterns of how terms are translated across many languages to find co-categorization patterns, which can then be used to derive a semantic space. We propose an automatic method that extracts the translations of each occurrence of a seed word (here, English indefinite pronouns) in every other language in our corpus. These extracted arrays of translations form a vector of terms across languages analogous to those obtained through elicitation data (cf. Fig. 1), and can be used to construct a geometric space.
Corpus and language sample. We extracted our data from a sentence-aligned parallel corpus of subtitles of films and TV series (Lison \& Tiedemann, 2016; www. opensubtitles.org). We selected the 30 (out of 65) languages across 9 language families for which the most


Figure 3: Extraction of situation vectors; see text.
parallel data was available, ${ }^{1}$ and extracted all utterances for which we found a translation into all languages.
Identifying translations across languages. We first obtained automatic alignments of translated words for each pair of languages in our corpus, using the HMM implementation of Liang, Taskar, and Klein (2006) with the default settings; see Fig. 3(a) for an example with four languages. From the pairwise alignments, we created a graph, per utterance, with edges between all words that are aligned with each other, (Fig. 3(b)). From this graph, we extracted the subgraphs that were densely connected (i.e., for which the words are often mutually aligned), ${ }^{2}$ and select those subgraphs that contain one of the indefinite pronouns in English (Fig. 3(c)). Each such subgraph is then linearized to form the vector representation of a situation (Fig. 3(d)). The Table in Fig. 3(d) illustrates the correspondence to semantic typology: Every row contains a 'stimulus', for which the various languages present elicited terms (cf. the table in Fig. 1). Note that sometimes responses are missing, or multiple words form the response.
Extraction of indefinite pronouns. We focus on the two ontological categories PEOPLE and THINGS; other categories (e.g., TIME and PLACE) were too infrequent. To identify indefinite pronoun usages in our corpus, we extracted utterances for which the English expression consists of any of the 9 words combining some-, any-, no- with -thing, -body, -one (cf. rows in bold in Fig. 3(d)). From among these situations, we selected only those that included an expression from each of at least 25 languages, to ensure sufficient linguistic variation for each situation. ${ }^{3}$ The resulting data consisted of 698

[^21]situations - i.e., exemplars of indefinite pronoun usage represented by vectors of terms in 25-30 languages.
Annotation. In order to compare the patterns in our usage data to Haspelmath's (1997) analysis, it was necessary to identify the semantic function (see Table 1) of the indefinite pronoun usages in each of our situations. To do so, three annotators (the authors) labelled the English indefinite pronoun in each situation with its Haspelmath function. Annotators were provided the sentence containing the pronoun, as well as some context before and after. We merged Specific Known and Specific Unknown into one function called Specific (SP), given the uncertainty in this task of judging whether something is known to the speaker. ${ }^{4} 152$ cases consisted of negative English indefinite pronouns like nothing and no one, which we automatically marked as DN. On the remaining 546 exemplars, inter-annotator agreement was satisfactory for a task of this difficulty (pairwise Cohen's $\kappa=[.84, .80, .79]$ ), and the majority annotation was used for each situation.
Further experimental set-up. Although the extracted situations are generally of a high quality, ${ }^{5}$ sometimes mistranslations are extracted. To reduce noise, we only use those terms that are statistically significantly associated with at least one of the annotated functions (using a Fisher Exact test). This way, low-frequency translations that are dispersed over functions are filtered out. To avoid the risk of overinterpreting patterns or overtuning models on the basis of a single sample, we split the data set into a development (dev) and test set. The examples and patterns reported below come from both the dev and test set, but quantitative results are provided for the test set only. We conduct all analyses on PEOPLE and things separately, because we found in exploratory data analysis that PEOPLE and THINGS showed differences in their patterns which have potentially interesting cognitive implications. The full data set, including stemming dictionary, annotation schemas, and all software used for the analyses, can be found at https://github.com/dnrb/indefinite-pronouns

## Results

With the extracted situation vectors, we can now study the semantic space derived from parallel usage data, and see how similar it is to Haspelmath's (1997) semantic typology based (primarily) on secondary sources. In particular, we are interested to see where the parallel usage data reveals characteristics of the semantic space not observed in Haspelmath's map.

## Are all semantic functions equally important?

Table 2 presents the frequency of the semantic functions. We see that most functions in the center of Haspelmath's

[^22]|  | SP | NS | CD | QU | IN | DN | CP | FC |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| PEOPLE | .16 | .20 | .07 | .16 | .05 | .28 | .01 | .08 |
| THINGS | .28 | .15 | .05 | .09 | .02 | .36 | .00 | .06 |
| Overall | .24 | .17 | .06 | .11 | .03 | .33 | .00 | .06 |

Table 2: Distribution of functions given ontological category.

|  | $k=$ | 2 | 3 | $\mathbf{4}$ | 5 | 6 | 7 | 8 | 9 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 10 |  |  |  |  |  |  |  |  |  |
| PEOPLE | .20 | .25 | .41 | .35 | .34 | .34 | .32 | .30 | .32 |
| THINGS | .30 | .38 | .47 | .36 | .35 | .35 | .33 | .39 | .33 |

Table 3: Adjusted Rand index score for PEOPLE and THings with $k$-means clustering, given various values of $k$.
(1997) semantic map are rather infrequent (CD, IN, CP). This may explain Haspelmath's observation that, across languages, there are no terms that solely apply to two functions in the middle of the map: Languages typically co-categorize infrequent functions with one of the more frequent neighboring functions (e.g., NS or DN). It also explains aspects of the graphical structure of the map: low-frequency functions are in the middle of the map because sometimes they share a term with the left side of the map, and sometimes with the right.

A notable exception is FC, located at the edge of the map despite its low frequency. This suggests FC is conceptually different from the other functions (except CP). Many languages co-categorize FC and universal quantification - unlike English, which generally uses any- vs. every- respectively. The use in many languages of a universal quantification term for the semantic function FC may account for its distinctive position in the map despite its low frequency.

## Are the functions at the right level of granularity?

A second issue worth investigating is whether Haspelmath's proposed functions constitute the best way of grouping the usage data into sets with related semantics: actual usage data may reveal that the functions are not well discriminable or have further coherent subdivisions. We explore this through automatic clustering of the parallel usage data. Each of our extracted situations is a vector of mutually-translated indefinite pronouns (see Fig. 3(d)); together they form a vector space within which we can measure situation (dis)similarity. Thus we can determine the optimal partitioning of the data into clusters and see how well those clusters correspond to the gold annotation. Here, we use $k$-means clustering (MacQueen, 1967), an unsupervised technique that partitions the data into $k$ clusters. The input for $k$-means is a distance matrix between all pairs of situations belonging to either PEOPLE or THINGS. The distance $d$ between a pair of situations $s$, $s^{\prime}$ is given by taking the Jaccard index over the sets of terms ${ }^{6}$ $T_{l}(s)$ and $T_{l}\left(s^{\prime}\right)$ used to express each of $s, s^{\prime}$ in each of the languages $l \in L$, and summing over all languages $l$ :

[^23]| Cluster | Function |  |  |  |  |  |  |  | Evaluation |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | SP | NS | CD | QU |  |  | CP | FC | $P$ | $R$ | $F_{1}$ |
| 1 | 18 | 24 | 6 | 3 | 0 | 2 | 0 | 0 | . 91 | . 92 | . 91 |
| 2 | 1 | 0 | 2 | 15 | 1 | 4 | 0 | 2 | . 60 | . 83 | . 70 |
| 3 |  | 0 | 1 | 0 | 5 | 27 | 0 | 0 | . 97 | . 82 | . 89 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 7 | . 80 | 1.00 | . 89 |

Table 4: Correspondence table for the 8 functions with $k=4$ clusters for PEOPLE; rightmost columns present cluster precision $(P)$, recall $(R)$ and $F_{1}$ score for every cluster against function tuples (SP,NS,CD); (QU); (IN,DN); (CP,FC).

$$
\begin{equation*}
d\left(s, s^{\prime}\right)=\sum_{l \in L} \frac{\left|T_{l}(s) \cap T_{l}\left(s^{\prime}\right)\right|}{\left|T_{l}(s) \cup T_{l}\left(s^{\prime}\right)\right|} \tag{1}
\end{equation*}
$$

We assess the relative quality of different numbers of clusters by comparing their fit to the annotations using the adjusted Rand index (Rand, 1971). Table 3 presents the results.

If Haspelmath's set of functions is the best way of describing the data, $k$-means clustering with $k=8$ should be the $k$ with the highest correspondence to the annotated functions, partitioning the data into 8 clusters corresponding to the 8 functions. However, with $k=8$, a relatively poor Rand index score is achieved, and rather than aligning with the semantic functions, the inferred clusters mostly cross-cut them (e.g., there are 2 clusters containing many DN ). The fact that the optimal partitioning cross-cuts functions suggests that there are finer semantic distinctions within the functions that play out in the way languages label these.

Instead, we find that $k=4$ gives the highest correspondence with the manually annotated clusters. The 4 clusters correspond to 4 sets of related functions: (SP,NS,CD), (QU), (IN,DN), and (FC,CP); see Table 4. There is some leakage between the clusters (see the non-boldface numbers in Table 4), but the precision, recall, and $F$ scores using these sets of functions as the target labels for the 4 clusters are very high, showing these sets of related functions have a clear similarity structure. ${ }^{7}$

These results yield two distinct views of the data. On one hand, the typological usage data points to more fine-grained semantic distinctions within some of the 8 functions. On the other hand, we find semantic similarity between the functions that reveals a coarser grouping of the functions than is apparent from the semantic map structure of Haspelmath (1997). These findings point to a key role of gradience in understanding the semantic space of indefinite pronoun usage.

## The perspective of a similarity space

The clustering over the parallel usage data suggests more gradience in the semantic space underlying indefinite pronoun semantics, both within and between functions, than the semantic map of Haspelmath (1997) suggests. We take a more

[^24]

Figure 4: OC plots for indefinite pronouns (best viewed on screen).
direct way of obtaining insight into this space by representing the similarity between all situations in a low-dimensional space. Visualizing this space can also be informative about the use of such a space in a computational cognitive model.

We apply 2-dimensional Optimal Classification (OC), a dimensionality reduction technique useful for typological data (Croft \& Poole, 2008). The input for this algorithm is the list of individual situations, each represented as a set of terms (across all languages) used for that situation. For each term $w$ across all languages, OC creates a cutting line in the 2 dimensional plane, which divides the situations into those expressed with $w$ and those not expressed with it. This way, pairs of situations expressed with similar sets of terms will typically be located close together in the OC space. (In our data, we have $n=303$ cutting lines for PEOPLE and $n=435$ cutting lines for things.) Our data yields very high accuracy (proportion of situations being on the correct side of the cutting line, averaged over all cutting lines) of .94 (PEOPLE) and .95 (THINGS). Because each situation is represented by a set of terms across all languages, this result shows high agreement among the languages in how they carve up the situations into functions - although, as we see next, they exhibit gradience in the gradual shifting of terms for related situations.

The topology of the function annotations in the twodimensional space generally follows Haspelmath's map, despite working from different data and with different methods (see Figures 4 a and 4 b ): in the top-left corner, we find NS and SP followed by CD and QU towards the center; the top-right cluster contains DN and IN, whereas the bottom cluster consists of CP and FC. ${ }^{8}$ However, there are several aspects of the functions that are observable in this continuous space that are not apparent in the graphical semantic map.

First, we observe that not all functions that neighbor each

[^25]| bs | hr | en | $\underset{\text { sl }}{\substack{\text { guage }}}$ | pt | da | Functions |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| išta | išta | anything kaj |  | alguma coisa noget |  | QU |
| nešto nešto |  |  |  | QU, CD |
|  |  | QU, CD |  |  |
|  |  | something nekaj |  |  |  | QU, CD, NS |
|  |  | algo |  |  |  | NS, SP |

Table 5: A gradient for the (SP,NS,CD,QU) region.
other in Haspelmath's map are equidistant in the OC solution: QU has an edge to each of NS and IN in the map, but is closer to the former in the OC projection. The projection furthermore displays gradience among the functions: some QU-labeled situations are closer to NS, whereas others are closer to IN, DN, or FC, suggesting that the functions are more continuous than the graphical map suggests.

Second, the functions display internal gradience. Fig. 4c shows terms in four languages for THINGS annotated as DN. The gradient comes about because of languages whose terms form supersets of each other: Estonian keegi is a superset of Croatian nitko, which is a superset of English nobody, which is a superset of Slovene nihče. Across languages there thus seems to be agreement about a scale of subtypes of DN, but languages vary on the placement of the lexical boundaries.

Finally, we find gradients that cross-cut the function boundaries. Table 5 illustrates a gradient of terms standing in a superset-subset relation to each other that cross-cuts the functions SP, NS, CD, and QU. This gradient was obtained by running a one-dimensional OC on the situations in the (SP,NS,CD,QU) region, which lays out all situations on one line so as to obtain a maximal accuracy in placing cutting points for terms. This analysis yields an accuracy of .96 , which suggests that languages strongly agree on having a single dimension roughly cross-cutting the functions SP, NS, CD , and QU on which they locate their term boundaries.

Visualizing crosslinguistic usages in a continuous space gives further insight into the structure of the underlying semantic domain. The observed gradients call for further analysis and provide predictions for behavioural experiments. In
particular, if the patterns of crosslinguistic variation are indicative of cognitive distinctions in semantic space, we expect to see evidence in both adult behaviour and developmental patterns in children.

## Conclusions

Crosslinguistic patterns of co-categorization yield insight into the semantic space underlying linguistic usages. We deploy parallel usage data in the form of movie subtitles to study the patterns of crosslinguistic variation in the categorization of indefinite pronouns. We find the cross-linguistic usages display a more fine-grained pattern than suggested by a study on the basis of (primarily) secondary data (Haspelmath, 1997). In particular, the frequencies of the identified semantic functions vary, the distances between the functions are not uniform, and within functions, coherent subgroupings could be established. Our findings suggest the parallel usage data captures something about the semantic space that is not represented in the more static secondary sources.

The current method can easily be applied to other domains, but also involves several restrictions. Using pairwise alignments on parallel texts makes the approach computationally intractable beyond 30-50 languages, as a set of alignments has to be extracted for every language pair. We are looking into methods to circumvent this aspect of the method. The inability of the model to go 'below' the word level is also limiting, as many well-established patterns of cross-linguistic semantic variation involve morphology (e.g., case marking, nominalization patterns).

Furthermore, it is crucial to establish the cognitive plausibility of the semantic similarity space independently by seeing if it can predict behavioral experiments such as word usage similarity judgments, or developmental patterns. For example, we must explore whether, as for space and color, the semantic space for indefinite pronouns predicts aspects of the acquisitional pattern of these words: Is English any-, for instance, hard to acquire because it covers a large, rather disjunct region of the semantic space? Are indefinite pronoun systems in languages that follow the typologically more common patterns easier to acquire for first and/or second language learners? We hope these automatic methods for using parallel text in semantic typology can help us further understand patterns of learning and usage in abstract domains of meaning.

## Acknowledgments

We would like to thank Martin Haspelmath for sharing his indefinite pronoun data, and support from NSERC of Canada.

## References

Beekhuizen, B., Fazly, A., \& Stevenson, S. (2014). Learning meaning without primitives: Typology predicts developmental patterns. In Proceedings CogSci.
Beekhuizen, B., \& Stevenson, S. (2016). Modeling developmental and linguistic relativity effects in color term acquisition. In Proceedings CogSci.

Berlin, B., \& Kay, P. (1969). Basic color terms: Their universality and evolution. Berkeley: UC Press.
Bowerman, M. (1993). Typological perspectives on language acquisition: Do crosslinguistic patterns predict development? In E. Clark (Ed.), Proceedings of the 25th annual Child Language Research forum (pp. 7-15).
Bowerman, M., \& Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In Annual report of the MPI for Psycholinguistics (pp. 53-56).
Croft, W., \& Poole, K. (2008). Inferring universals from grammatical variation: multidimensional scaling for typological analysis. Theoretical Linguistics, 1-37.
Cysouw, M., \& Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. Language Typology and Universals, 60, 95-99.
Haspelmath, M. (1997). Indefinite pronouns. Oxford: OUP.
Haspelmath, M. (2003). The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), The new psychology of language (pp. 211-242).
Levinson, S. C., Meira, S., \& The Language and Cognition Group. (2003). 'Natural concepts' in the spatial topological domain - Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. Language, 79(3), 485-516.
Levshina, N. (2017). Subtitles as a corpus: An n-gram approach. Corpora.
Liang, P., Taskar, B., \& Klein, D. (2006). Alignment by agreement. In Proceedings NAACL (pp. 104-111).
Lison, P., \& Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Proceedings LREC.
MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (pp. 281-297).
Palla, G., Derényi, I., Farkas, I., \& Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814-818.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.
Regier, T., Kemp, C., \& Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney \& W. O'Grady (Eds.), The handbook of language emergence (pp. 237-263).
Rosch, E. (1973). Natural categories. Cognitive Psychology, 4, 328-350.
Silverstein, M. (1976). Hierarchy of features and ergativity. In R. Dixon (Ed.), Grammatical categories in Australian languages (pp. 112-171).
Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W., \& Bhattacharya, T. (2016). On the universal structure of human lexical semantics. PNAS, 113, 1766-1771.

# How Relative is the Relative Frame of Reference? Front and back in Norwegian, Farsi, German, and Japanese 

Sieghard Beller (sieghard.beller@uib.no)<br>Andrea Bender (andrea.bender@uib.no)<br>Department of Psychosocial Science, University of Bergen<br>N-5020 Bergen, Norway


#### Abstract

Across languages, people differ in which of the three basic frames of reference (FoRs) they prefer when describing spatial relations: absolute, intrinsic, or relative. But how much variation is there with regard to the relative FoR, which is anchored in the observer and occurs as one of three variants? Is the reflection variant canonical, as assumed by many scholars? And how are objects in a person's back referred to: by turning towards the objects? Results from two studies, one with speakers of Norwegian and Farsi, the other with speakers of German and Japanese, reveal that reflection is not canonical, but that translation and even rotation are used as well. In addition, turning towards objects arranged in a person's back is very rare; what people use instead is a backward projection strategy that goes without rotation.


Keywords: Spatial cognition, frames of reference (FoR), relative FoR, cross-linguistic study.

## Introduction

"Where is the ball in relation to the box?" In order to answer questions like this, we have to establish a coordinate system-a frame of reference (FoR)-that allows us to derive a specific response such as "The ball is in front and to the right of the box." Across languages, people differ in the frame of reference they preferentially adopt. Variation has been documented especially with regard to which of the three basic FoRs is used: the absolute FoR anchored in a superordinate field like the cardinal directions, the intrinsic FoR anchored in a reference object like an arrow, or the relative FoR anchored in an observer (Levinson, 2003; Majid et al., 2004; Senft, 1997). Less attention has been devoted to the variants of the relative FoR, despite the fact that variation in relative referencing has been known since Hill's (1982) comparison of English and Hausa speakers.
This paper adds to a survey exploring variation in the use of the relative FoR in different languages (for results on German, English, Mandarin Chinese, and Tongan, see Beller et al., 2015), by extending the set of sampled languages. In two studies, one with speakers of Norwegian and Farsi, and another with speakers of German and Japanese, we inspected which variant of the relative FoR speakers of these languages apply in frontal and dorsal tasks with objects laid out in front of or behind an observer.

## Variants of the Relative Frame of Reference

Frames of reference are used to describe the position of a figure object F in reference to a ground object G . In contrast to the absolute and intrinsic FoR, the relative FoR requires
to do so from an observer's viewpoint V. As objects can be in front of or behind the observer, the distinction between frontal and dorsal is indispensable.

## FoRs in Frontal Settings

Constructing a relative frame of reference requires the coordinate system that is originally anchored in the observer-his or her FRONT/BACK and LEFT/RIGHT-to be projected onto the ground object G. In frontal settings, this can be done in three ways (Levinson, 2003): The coordinate system can be translated into G so that FRONT is assigned in gaze direction of the observer to the space beyond $G$ (see Figure 1A). It can be reflected in G so that FRONT is assigned to the space between the observer and G (Figure 1B). In both cases, assignment of LEFT and RIGHT remain unaffected. Finally, it can be rotated in G by $180^{\circ}$; in this case, FRONT is, again, assigned to the space between the observer and G, but assignment of LEFT and RIGHT are swapped (Figure 1C).

Of these variants, reflection is often assumed to be the canonical one (Clark, 1973; Grabowski \& Miller, 2000; Janzen et al., 2012). In our cross-linguistic survey (Beller et al., 2015, cf. Table 3, p. 11), such a preference for reflection was found most strongly among speakers of German (89\%) and English (73\%), whereas speakers of Mandarin Chinese and Tongan clearly preferred translation (64\%) over reflection (24\%). The rotation variant was chosen rarely in


Figure 1: Variants of the relative FoR for frontal settings (according to Levinson, 2003); L/R: left/right; F/G: figure/ground; V : viewpoint of the observer.
all four languages. Extending our survey to other languages so as to broaden our knowledge with regard to intra- and cross-linguistic variation in the use of these variants of the relative FoR is the first aim of the current paper.

## FoRs in Dorsal Settings

Research on the relative FoR has focused nearly exclusively on how people describe relations of objects that are laid out in front of an observer-for obvious reasons the most natural situation-although adopting someone's perspective already includes the distinction between what is in front of and what is behind that person. But how, if at all, would objects laid out behind that person be referred to?
One hypothesis put forward by Grabowski and Miller (2000) is that people refrain from referring to objects in their back. Rather, they turn around toward the objects, thereby converting the dorsal into a frontal setting, and then employ the FoR they prefer for frontal settings. However, the first studies on dorsal references (Beller et al., 2015, 2016) provided only weak evidence for this turn hypothesis. Despite participants' preferences for reflection or translation in frontal tasks, only a few responses in dorsal tasks were in accordance with the corresponding strategies turn-reflection and turn-translation. What most participants seemed to do instead was a kind of backward projection of the observer's coordinate system (without rotating the observer's perspective) either in a back translation version (Figure2A) or a reflection "with eyes in the back of one's head" (Figure 2B).
For logical reasons, both of these backward projection strategies lead to the same responses as the turn-rotation strategy (turn the perspective and apply the rotation variant to the resulting frontal setting; see Figure 2C). The reasons for why we assumed that the participants applied backward projection were twofold: First, these strategies do not necessitate two laborious (mental) rotations, and second, participants applied the rotation variant (Figure 1C) only rarely in frontal tasks-why should they do so in dorsal

"The ball is in front and to the right of the box."
Figure 2: Three variants of the relative FoR for dorsal settings (Beller et al., 2015); BP: backward projection.


| The white $\{$ circle, | starfish \} is located $\ldots$ |
| :--- | :--- |
| $\square$ in front of | $\square$ in front and to the left of |
| $\square$ behind | $\square$ in front and to the right of |
| $\square$ to the left of | $\square$ behind and to the left of |
| $\square$ to the right of | $\square$ behind and to the right of |$\quad$| $\ldots$ the black $\{$ square , flower, arrow, scorpion \}. |
| :--- |

Figure 3: Four example items (Beller et al., 2015, p. 6).
tasks? Exploring backward projection further, as compared to the turn-hypothesis, is the second aim of the current paper.

## Study 1

The first study was implemented as a paper-and-pencil survey that followed the design described in Beller et al. (2015) and included two languages from the Indo-European language family: Norwegian from the Germanic branch and Farsi from the Indo-Iranian branch.

## Methods

Materials. The materials were the same as in Beller et al. (2015): twelve items in each of two conditions (frontal and dorsal), six with a non-oriented ground object (three depicting inanimate objects, three depicting living beings) and six with an oriented ground object (again three depicting inanimate objects and three depicting living beings). Participants were asked to indicate for each item the relation between figure F and ground G from the viewpoint V of a depicted observer, by choosing one of eight options (in front of, behind, to the left of, to the right $o f$, and combinations of in front of/behind and to the left/right of). Four example items are shown in Figure 3. All materials were translated into Norwegian and Farsi by bilinguals.

Participants. The Norwegian sample consisted of 64 students from the University of Bergen ( 51 female; age $M=$ 23.3 years, $S D=5.3$ ), and the Farsi sample of 130 participants, most of them students from the Universities of Teheran, Schiraz, and Ghazwin, but also some non-students ( 88 female; age $M=26.8$ years, $S D=15.9$ ).

Design and Procedure. For each of the two conditions (frontal vs. dorsal), two item orders were prepared: The first one started with the six non-oriented items (in a random order) and then proceeded with the six oriented items (also in a random order); the second order was the exact reversal and thus started with the six oriented items. The eight response options were always presented in the same order. A between-subjects design was used. Participants were assigned randomly (but equally) either to the frontal or to the dorsal condition; the two item orders per condition were balanced in each subgroup. Participants were tested individually or in small groups, and were instructed to work on all tasks in the given order.

## Results and Discussion

In the frontal condition, we distinguished between the three variants of the relative FoR: translation, reflection, and rotation (Figure 1). In the dorsal condition, we distinguished between three variants according to the turn-hypothesis: turn-translation, turn-reflection, and turn-rotation, the latter one being equivalent to two backward projection strategies, translation ${ }_{\text {BP }}$ and reflection ${ }_{\text {BP }}$ (Figure 2). For items with an oriented ground object, we also considered the intrinsic FoR.
In a first step, we checked the two samples for differences in the mean number of responses that are not covered by one of these FoRs. Overall, this number of "unexplained responses" was fairly low ( $M=8.6 \%$; Table 1 ). An analysis of variance with two between-subjects factors, language (Norwegian vs. Farsi) and perspective (frontal vs. dorsal), and one within-subject factor ground object (non-oriented vs. oriented) indicated no significant effects (all $F(1,190)$ < $1.53 ; p \geq .218 ; \eta^{2} \leq .008$ ), suggesting that neither the

Table 1: Frequency (\%) of responses that are not covered by one of the FoRs under scrutiny in Study 1 and Study 2.

| $\begin{aligned} & \text { I } \\ & \text { 立 } \\ & \text { in } \end{aligned}$ | Type of item | Norwegian | Farsi |
| :---: | :---: | :---: | :---: |
|  | Frontal, non-oriented | 5.7 | 10.6 |
|  | Frontal, oriented | 8.3 | 10.1 |
|  | Dorsal, non-oriented | 7.8 | 10.7 |
|  | Dorsal, oriented | 7.3 | 8.6 |
| $\begin{aligned} & N \\ & \stackrel{N}{3} \\ & \stackrel{y}{3} \end{aligned}$ | Type of item | German | Japanese |
|  | Frontal, non-oriented | 3.7 | 3.7 |
|  | Frontal, oriented | 4.8 | 3.5 |
|  | Dorsal, non-oriented | 4.0 | 2.6 |
|  | Dorsal, oriented | 4.2 | 3.5 |

Table 2: Individual consistency in FoR adoption (in \% of items) in Study 1 and Study 2.

| $\begin{aligned} & \text { त } \\ & \text { خ } \\ & \text { B } \end{aligned}$ | Type of item | Norwegian | Farsi |
| :---: | :---: | :---: | :---: |
|  | Frontal, non-oriented | 88.5 | 79.5 |
|  | Frontal, oriented | 85.4 | 68.2 |
|  | Dorsal, non-oriented | 87.0 | 85.2 |
|  | Dorsal, oriented | 78.6 | 68.5 |
| $\begin{aligned} & \text { N } \\ & \text { N } \\ & \text { D } \end{aligned}$ | Type of item | German | Japanese |
|  | Frontal, non-oriented | 93.5 | 87.3 |
|  | Frontal, oriented | 89.8 | 82.1 |
|  | Dorsal, non-oriented | 92.6 | 90.7 |
|  | Dorsal, oriented | 91.2 | 83.9 |

unusual dorsal perspective nor the type of ground object influenced the coverage of responses by the FoRs under scrutiny in the two samples alike.

In the next step, we determined whether the individual participants adopted one FoR consistently and, if so, which one. To this end, we counted for each participant how often each FoR variant could be coded in each of the four blocks of six items (frontal non-oriented, frontal oriented, dorsal non-oriented, and dorsal oriented). For example, if reflection could be coded on 6 out of the 6 frontal oriented items, consistency would be $100 \%$ for reflection; if reflection could be coded on 5 items and translation on 1 item, consistency would be $83.3 \%$ for reflection and $16.7 \%$ for translation; etc. We then used the maximum of these values (among the different FoR variants) as estimate of a participant's consistency in FoR adoption across the items of the respective block ( $100 \%$ and $83.3 \%$ in the examples). Mean consistency values are displayed in Table 2.

Overall, FoRs were adopted with a mean consistency of $80.1 \%$ across the two samples. In other words: Participants adopted their individually preferred FoR in 4.81 of 6 items of a block. An analysis of variance with two betweensubjects factors, language (Norwegian vs. Farsi) and perspective (frontal vs. dorsal), and one within-subject factor ground object (non-oriented vs. oriented) indicated three significant effects: Consistency was generally higher for the Norwegian speakers than for the Farsi speakers ( $84.9 \%$ vs. $\left.75.3 \% ; F(1,190)=11.3 ; p=.001 ; \eta^{2}=.056\right)$; it was higher for non-oriented items than for oriented items ( $85.1 \%$ vs. $75.2 \% ; F(1,190)=31.8 ; p<.001 ; \eta^{2}=.144$ ); and there was an interaction of the two factors language $\times$ ground object $\left(F(1,190)=5.6 ; p=.019 ; \eta^{2}=.029\right)$. Thus, the possibility of applying an additional FoR (here: intrinsic) was a source of inconsistency, but to a different extent in the two languages. Interestingly, the unusual dorsal perspective per se did not matter: Consistency did not differ significantly between the frontal and the dorsal condition ( $84.4 \%$ vs. $79.8 \% ; F(1,190)=0.045 ; p=.832 ; \eta^{2}<.001$ ).

Finally, we identified each participant's preferred FoR as the one FoR variant that was coded (a) more often than all others and (b) in at least 4 out of the 6 items of a block (i.e.,

Table 3: Preferred FoR (in \%), adopted in at least 4 out of 6 items of a block (frontal non-oriented, frontal oriented, dorsal non-oriented, and dorsal oriented) in Study 1 and Study 2.

| FoR | Study 1 |  |  |  | Study 2 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Non-oriented G |  | Oriented G |  | Non-oriented G |  | Oriented G |  |
|  | Norwegian | Farsi | Norwegian | Farsi | German | Japanese | German | Japanese |
| Frontal items | ( $N=32$ ) | ( $N=66$ ) | ( $N=32$ ) | ( $N=66$ ) | ( $N=140$ ) | ( $N=109$ ) | ( $N=140$ ) | ( $N=109$ ) |
| Intrinsic | n.a. | n.a. | 3.1 | 19.7 | n.a. | n.a. | 2.9 | 4.6 |
| Translation | 21.9 | 28.8 | 15.6 | 18.2 | 7.1 | 48.6 | 4.3 | 45.9 |
| Reflection | 71.9 | 40.9 | 71.9 | 16.7 | 88.6 | 36.7 | 85.0 | 28.4 |
| Rotation | - | 10.6 | - | 4.5 | 0.7 | 1.8 | - | 0.9 |
| No preference | 6.3 | 19.7 | 9.4 | 40.9 | 3.6 | 12.8 | 7.9 | 20.2 |
| Dorsal items | ( $N=32$ ) | ( $N=64$ ) | ( $N=32$ ) | ( $N=64$ ) | $(N=140)$ | ( $N=109$ ) | ( $N=140$ ) | ( $N=109$ ) |
| Intrinsic | n.a. | n.a. | 6.3 | 18.8 | n.a. | n.a. | 3.6 | 6.4 |
| Turn-translation | 3.1 | - | - | - | - | 1.8 | - | 2.8 |
| Turn-reflection | 3.1 | - | - | - | 5.7 | 2.8 | 6.4 | 2.8 |
| Translation ${ }_{\text {BP }} /$ reflection ${ }_{\text {BP }} /$ /urn-rotation | 81.3 | 82.8 | 71.9 | 40.6 | 88.6 | 89.0 | 85.7 | 78.9 |
| No preference | 12.5 | 17.2 | 21.9 | 40.6 | 5.7 | 6.4 | 4.3 | 9.2 |

Note. BP: backward projection; n.a.: not applicable; modal response printed in bold face.
with a consistency of $\geq 66.7 \%$ ). Participants' preferred FoRs are presented in Table 3. Log-linear analyses of FoR preferences revealed differences between the two languages for the two blocks of frontal items (non-oriented: $G^{2}=12.7$; $d f=3 ; p=.005$; oriented: $G^{2}=33.7 ; d f=4 ; p<.001$ ), and for the block of oriented dorsal items ( $G^{2}=8.9$; $d f=2$; $p=.012)$, but not for the block of non-oriented dorsal items ( $G^{2}=4.7 ; d f=3 ; p=.194$ ).

If adopting the intrinsic FoR was possible, some participants preferred this $\mathrm{FoR}^{1}$, particularly in the Farsi sample (mean percentage across all items with an oriented G for Norwegian: 4.7\%; Farsi: 19.2\%). Other participants seemed to change their referencing strategy itemspecifically, as indicated by the increased number of participants with no clear preference for any FoR variant as compared to the items with a non-oriented ground object.
Among the variants of the relative FoR for frontal tasks, both translation and reflection were adopted, but to a different extent in the two samples. The reflection variant prevailed most strongly among the Norwegian speakers (mean percentage across all frontal items for Norwegian: $71.9 \%$; Farsi: $28.8 \%$ ), while translation was preferred by a substantial proportion of speakers in both samples (Norwegian: $18.8 \%$; Farsi: $23.5 \%$ ); the rotation variant, in contrast, was confined to some Farsi speakers (7.6\%).
Among the variants of the relative FoR for dorsal tasks, one variant clearly stood out, namely the one that is indicative of the application of backward projection and turnrotation (Figure 2). Its frequency (mean percentage across

[^26]all dorsal items for Norwegian: 76.6\%; Farsi: 61.7\%) approximates the sum of translation, reflection, and rotation from the frontal tasks. But since the factor perspective was implemented between-subjects, the frontal and dorsal data cannot be related to one another on an individual basis, which would have provided a stronger argument in favor of this correspondence. In either case, the two FoR variants predicted by the turn-hypothesis-turn the view towards the objects and then apply the FoR preferred for frontal settings (i.e., reflection or translation)-were adopted very rarely.

In sum, Study 1 demonstrated that the reflection variant of the relative FoR is not canonical. While being the most frequent FoR in frontal tasks, the translation variant is adopted as well, and some participants even adopted the rotation variant. Participants' dorsal references suggested backward projection as the main strategy, but the data are not fully conclusive due to the between-subjects design.

## Study 2

In order to allow us to relate a participant's referencing preference in dorsal tasks to that in frontal tasks, the second study included perspective (frontal vs. dorsal) as a withinsubject factor. The study was implemented as an online survey and compared two languages from different language families: German, another Germanic language, and Japanese from the Japonic language family.

## Methods

Materials. The items were the same as in Study 1. The materials were translated from German into Japanese by bilinguals and were implemented as a web-based online questionnaire.

Participants. The German sample consisted of 140 student and non-student participants ( 105 female; age $M=27.3$ years, $S D=10.9$ ), and the Japanese sample of 109 student and non-student participants ( 64 female; age $M=28.5$ years, $S D=10.4$, with 15 not indicating their age).

Design and Procedure. The two perspectives (frontal vs. dorsal) were implemented within-subject. Which came first was assigned randomly for each participant. Within each perspective, non-oriented and oriented items were presented in blocks, and within each block in random order.

## Results and Discussion

The data were analyzed in the same way as in Study 1. In the first step, we checked the two samples for differences in the mean number of responses that are not covered by one of the FoRs under scrutiny. Overall, this number of "unexplained responses" was very low ( $M=3.7 \%$; Table 1) and lower still than for Norwegian and Farsi. An analysis of variance with the between-subjects factor language (German vs. Japanese) and two within-subject factors, perspective (frontal vs. dorsal) and ground object (nonoriented vs. oriented), indicated no significant effects (all $\left.F(1,247)<1.9 ; p \geq .171 ; \eta^{2} \leq .008\right)$. Neither the unusual dorsal perspective nor the type of ground object influenced the coverage of responses by the FoRs under scrutiny in the two samples alike.
Then, we checked how consistently each FoR variant was adopted. Overall, FoRs were adopted with a mean consistency of $88.9 \%$ across the two samples (Table 2). In other words: Participants adopted their individually preferred FoR in 5.33 of 6 items of a block. An analysis of variance with the between-subjects factor language (German vs. Japanese) and two within-subject factors, perspective (frontal vs. dorsal) and ground object (nonoriented vs. oriented), detected the same three effects as in Study 1: main effects of language and ground object, and an interaction of the two factors. Consistency was higher for the German speakers than for the Japanese speakers ( $91.8 \%$ vs. $\left.86.0 \% ; F(1,247)=15.2 ; p<.001 ; \eta^{2}=.058\right)$. It was also higher for non-oriented items than for oriented items ( $91.0 \%$ vs. $\left.86.8 \% ; F(1,247)=32.2 ; p<.001 ; \eta^{2}=.115\right)$, indicating again that the possibility of applying the intrinsic FoR was a source of inconsistency, but to a different extent in the two languages (as reflected in the interaction; $F(1,247)=5.1$; $p=.024 ; \eta^{2}=.020$ ). And, as in Study 1, the unusual dorsal perspective per se did not matter: Consistency was nearly the same for the frontal items as for the dorsal items ( $88.2 \%$ vs. $\left.89.6 \% ; F(1,247)=2.258 ; p=.134 ; \eta^{2}=.009\right)$.

Participants' preferred FoRs are shown in Table 3. Loglinear analyses of FoR preferences indicated differences between the two languages for the same three item blocks as in Study 1: for the two blocks of frontal items (non-oriented: $G^{2}=78.3 ; d f=3 ; p<.001$; oriented: $G^{2}=96.0 ; d f=4$; $p<.001$ ), and for the block of oriented dorsal items ( $G^{2}=10.4 ; d f=4 ; p=.034$ ), but not for the block of nonoriented dorsal items ( $G^{2}=4.6 ; d f=3 ; p=.201$ ).

Table 4: Preferred FoR in dorsal item blocks depending on the preferred FoR in frontal item blocks in Study 2.

| Dorsal <br> preference | Frontal preference |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | Translation | Reflection | Rotation | Other |
| Turn-translation | 3 | - | - | 2 |
| Turn-reflection | 2 | 20 | - | 1 |
| BP/turn-rotation | $106_{\mathrm{BP}}$ | $278_{\mathrm{BP}}$ | 3 | 40 |
| Other | 8 | 16 | 1 | 18 |
| $N=498$ | 119 | 314 | 4 | 61 |

Note. Data are summed over non-oriented and oriented item blocks and the two samples. BP: backward projection (translation ${ }_{B P}$ or reflection ${ }_{B P}$ ); the category other includes participants with no preference (from all tasks) and with a preference for the intrinsic FoR (from tasks with an oriented G). Grey cells: Responses according to the turn-hypothesis.

If adopting the intrinsic FoR was possible, again some participants preferred this FoR (mean percentage across all items with an oriented G for German: $3.2 \%$; Japanese: $5.5 \%$ ), but less so than in the Farsi sample. Some participants also seemed to change their referencing strategy item-specifically, as indicated by the increased number of participants with no clear preference for any FoR variant as compared to the items with a non-oriented ground object.

Among the variants of the relative FoR for frontal tasks, translation and reflection were preferred most often, but again to a different extent in the two samples. The reflection variant prevailed most strongly among the German speakers (mean percentage across all frontal items for German: $86.8 \%$; Japanese: $32.6 \%$ ). This finding replicates data from a German sample collected with a paper-and-pencil questionnaire (Beller et al., 2015), thereby validating the methodological change to an online assessment (see Beller et al., 2015, 2016, for a broader discussion of the paperpencil assessment and other methodological issues). In contrast, the translation variant prevailed among the Japanese speakers (German: 5.7\%; Japanese: 48.6\%). The rotation variant was adopted only by very few participants (German: 0.4\%; Japanese: $1.4 \%$ ).

Among the variants of the relative FoR for dorsal tasks, the variant indicating the application of backward projection and turn-rotation strongly dominated in the two samples alike (mean percentage across all dorsal items for German: $87.1 \%$; Japanese: $83.9 \%$ ).

The implementation of perspective as a within-subject factor in this study allows us to relate each participant's preference in dorsal tasks to his or her preference in frontal tasks and thereby to disambiguate the dorsal response (cf., Beller et al., 2016). To this end, we cross-tabulated participants' preferred FoRs for frontal and dorsal tasks (summed over non-oriented and oriented item blocks and the two samples). The results are reported in Table 4. Of the 498 preference pairs, $26(5.2 \%)$ were indicative of the turnhypothesis (grey cells). Most of these participants adopted the turn-reflection variant in line with the overall higher
prevalence for reflection. This provides some support for the turn-hypothesis. However, the vast majority of pairings (384 or $77.1 \%$ ) pointed at backward projection as the prevailing strategy (translation $_{\mathrm{BP}}$ : $21.3 \%$; reflection ${ }_{\mathrm{BP}}$ : $55.8 \%$ ).

In sum, Study 2 corroborated further that the reflection variant of the relative FoR is not universally adopted. While being the most frequent FoR used for frontal tasks in the German sample, the translation variant predominated in the Japanese sample. In line with the results from Beller and colleagues (2016), participants' dorsal references indicated backward projection as the main strategy.

## General Discussion

The goal of this paper was to broaden our knowledge regarding intra- and cross-linguistic variation in the use of different variants of the relative FoR for spatial references in frontal and dorsal settings. In particular, we asked two questions: Do people have a canonical preference for the reflection variant of the relative FoR in frontal settings, as assumed by some scholars? And do people (mentally) turn around to an object configuration in their back and apply the FoR they prefer for frontal settings (turn-hypothesis)? Our findings indicate that neither is the case.
With regard to the first question, we detected a great deal of intra- and cross-linguistic variation in people's use of the relative FoR in frontal settings. The speakers of Norwegian, German, and Japanese exhibited high intra-individual consistency. Almost all participants applied the same variant of the relative FoR repeatedly for a whole set of tasks. Among the speakers of Farsi, consistency was lower, indicating more task-specific references, particularly in cases where the intrinsic FoR was also possible. With regard to the inter-individual consensus within the samples, we observed high consensus among the German speakers (i.e., most speakers adopted the same FoR variant as everybody else: reflection), moderate consensus among the Norwegian and Japanese speakers (some of which preferred the reflection variant, others the translation variant), and an even weaker consensus among the Farsi speakers (for which the data also indicate a rare but consistent use of the rotation variant). All in all, reflection and translation were the dominant variants of the relative FoR for frontal settings, replicating the general pattern found for German, English, Mandarin Chinese, and Tongan (Beller et al., 2015).
In spite of this diversity in frontal tasks, most participants converged on the very same response in the dorsal tasks. In most cases, this response could be attributed to backward projection strategies that are in line with people's frontal preference for translation or reflection, but get by without (mental) rotation, and are thus quite adaptive given the fact that mental rotation comes with substantial cognitive costs (Duran, Dale, \& Kreuz, 2011; Shepard \& Cooper, 1982).

Finally, the degree of linguistic variation is revealing in yet another regard. The intra-linguistic variation we found reflects the fact that spatial prepositions like "in front of" or "behind" are inherently underspecified. Nothing in these words tells us where exactly FRONT or BACK is. This can
only be established after having adopted a specific point of view, or frame of reference. Yet, which FoR a speaker adopts is either due to his or her individual preference or to conventions within his or her speech community. Viewed in this way, the variation we found is a cultural rather than a purely linguistic phenomenon.

## Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft DFG through a Heisenberg Fellowship to AB (Be 2451/8-1,2) and through a grant for the project Spatial referencing across languages: Cultural preferences and cognitive implications to AB ( $\mathrm{Be} 2451 / 13-1$ ) and SB ( Be 2178/7-1). For data collection, we thank Solveig Dittmann, Aurora Bratterud Helgesen, Anette Sollund, and Tora Standal-Knudsen (Norwegian), Atifeh Esfandiari (Farsi), and Meike Rosemeyer (German and Japanese). We thank Lisa Hüther for assisting AE and MR, and Lisa Hüther and Fiona Wilke for comments on earlier versions of this paper.

## References

Beller, S., Bohlen, J., Hüther, L., \& Bender, A. (2016). Perspective taking in referring to objects behind versus in front of an observer: Frames of reference, intra-individual consistency, and response latencies. The Quarterly Journal of Experimental Psychology, 69, 1384-1408.
Beller, S., Singmann, H., \& Hüther, L., \& Bender. A. (2015). Turn around to have a look? Spatial referencing in dorsal versus frontal settings in cross-linguistic comparison. Frontiers in Psychology, 6: 1283, 1-17.
Clark, H. H. (1973). Space, time, semantics, and the child. In T. E. Moore (Ed.), Cognitive development and the acquisition of language. New York: Academic Press.
Duran, N. D., Dale, R., \& Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive costs. Cognition, 121, 22-40.
Grabowski, J., \& Miller, G. A. (2000). Factors affecting the use of dimensional prepositions in German and American English: Object orientation, social context, and prepositional pattern. Journal of Psycholinguistic Research, 29, 517-553.
Hill, C. A. (1982). Up/down, front/back, left/right. A contrastive study of Hausa and English. In J. Weissenborn \& W. Klein (Eds.), Here and there. Amsterdam: Benjamins.

Janzen, G., Haun, D. B. M., \& Levinson, S. C. (2012). Tracking down abstract linguistic meaning: Neural correlates of spatial frame of reference ambiguities in language. PLoS ONE, 7(2), e30657.
Levinson, S. C. (2003). Space in language and cognition. Cambridge: Cambridge University Press.
Majid, A., Bowerman, M., Kita, S., Haun, D., \& Levinson, S. C. (2004). Can language restructure cognition? The case for space. Trends in Cognitive Sciences, 8, 108-114.
Shepard, R. N., \& Cooper, L. A. (1982). Mental images and their transformations. Cambridge, MA: MIT Press
Senft, G. (Ed.) (1997). Referring to space. Oxford: Clarendon.

# Quantifying Infants' Statistical Word Segmentation: A Meta-Analysis 

Alexis Black (akblack2g@gmail.com)<br>Department of Linguistics, 2613 West Mall<br>Vancouver, BC, Canada V6T 1Z4

Christina Bergmann (chbergma@gmail.com)<br>Laboratoire des Sciences Cognitives, Ecole Normale Supérieure, 29, rue d'Ulm 75005 Paris, France


#### Abstract

Theories of language acquisition and perceptual learning increasingly rely on statistical learning mechanisms. The current meta-analysis aims to clarify the robustness of this capacity in infancy within the word segmentation literature. Our analysis reveals a significant, small effect size for conceptual replications of Saffran, Aslin, \& Newport (1996), and a nonsignificant effect across all studies that incorporate transitional probabilities to segment words. In both conceptual replications and the broader literature, however, statistical learning is moderated by whether stimuli are naturally produced or synthesized. These findings invite deeper questions about the complex factors that influence statistical learning, and the role of statistical learning in language acquisition.


Keywords: language acquisition; statistical learning; word segmentation; meta-analysis

## Introduction

Statistical learning (SL), the ability to extract statistical patterns from a continuous stream of perceptual experiences, is of fundamental theoretical importance. The first evidence that infants can extract statistical information from speech and use it to group syllables was provided by Saffran, Aslin, and Newport in 1996. This seminal paper has since accrued thousands of citations, and spurred a rich literature invoking SL as one foundation for language acquisition (see Newport, 2016) as well as perceptual learning more broadly (see Aslin, 2017). SL mechanisms have furthermore been successfully implemented in a range of computational models (e.g., Pearl, Goldwater, \& Steyvers, 2010; Lloyd-Kelly, Gobet, \& Lane, 2016). In short, statistical learning abilities are of fundamental, crossdisciplinary importance to better understand the computational foundations of cognition.

While many would accept some role for SL mechanisms, the nature and extent of this role remains contested. The more abstract the level of analysis, the more vigorous the debate (e.g., can SL yield syntactic 'rules'?) - but inconsistencies emerge even at the level of tracking transitional probabilities (TPs) as a means of word segmentation. For example, the original effect has failed to replicate under certain conditions (e.g., variable word length: Johnson \& Tyler, 2010; Lew-Williams \& Saffran, 2012) or showed a developmental shift in cue-weighting (e.g., Thiessen \& Saffran, 2003). Finally, a recent meta-
analysis that examined natural speech word segmentation (not determined by TPs) revealed a significant, but small effect (Bergmann \& Cristia, 2016), leading to concerns about the robustness of infants' word segmentation in the absence of TPs.

In the current paper, we use meta-analysis to quantify and contextualize infants' ability to detect regularities in a continuous speech stream. To this end, we have aggregated all available evidence from the published record and present a meta-analysis of infant SL word segmentation studies. A meta-analytic approach helps establish the magnitude of an underlying effect, something single experiments are not equipped to do - and thus has the potential to impact future theory- and model-building. On the practical side, effect sizes are crucial for determining power of future studies, thus increasing the replicability of a line of inquiry and reducing the cost (failed studies, or testing too many participants) for single researchers.

We also take several steps beyond quantifying the underlying effect: Aggregating over studies allows for the identification of moderator variables, which also contributes to theory building and may guide future research. We examine three potential moderators that are relevant to the intersection of theories of infant cognition and statistical learning: (1) age, (2) stimulus naturalness, and (3) non-TP cues. The justification for investigating these particular moderators is described in brief. (1) All studies in the current meta-analysis use looking-time preferences. The direction of preference (to novel or familiar items) is commonly thought to relate to infant age and/or stimulus complexity (e.g., Hunter \& Ames, 1988). We therefore predicted that developmental change might be reflected in a shift of preference (e.g., from a preference for words to one for non-words), or in a stronger effect over time. (2) Given the familiarity preference found in a previous meta-analysis on natural speech (Bergmann \& Cristia, 2016), we hypothesized that it might be that the predominant novelty preference established for SL studies since Saffran et al. (1996) is grounded in methodological choices. The primary difference between these two datasets is in the nature of the stimuli: naturally produced vs highly artificial speech stimuli. Even within the literature of the current dataset, however, stimuli differ along this dimension. We therefore compare SL studies with natural and artificial stimuli. (3) Finally, a number of studies pitted alternative cues (e.g., word-level stress) against TPs. It is therefore important to
examine the impact of these conflicting cues on SL performance compared to no conflict.

We also assess publication bias in the literature; a current topic that is especially important for infant research, considering the high cost of testing participants and the consequent use of small samples (Frank et al., 2017).

## Methods

To collect data, we complemented expert lists with two google scholar searches. We first surveyed papers citing Saffran, Aslin, \& Newport (1996) with the word "infant/infancy", but not "visual" in the title. The second search aimed to cast a wider net; search terms were now "month/s" and not "infant/infancy" or "visual". These two strategies yielded a total of 314 unique papers, which were then screened for inclusion. The criteria were: (1) contains data on infants from (2) behavioral experiments which exposed infants to a familiarization phase of continuous, artificial speech and which measured (3) reactions (typically looking times to unrelated visual stimuli) to both statistical words and non-words (this definition includes part-words).

The final sample encompassed 20 papers ( 10 containing conceptual replications ${ }^{1}$ ) yielding 68 ( 17 replication) effect sizes. Note that one paper often contains several experiments (henceforth: samples) that can yield effect sizes, for example when testing different age groups. In total, we are reporting on experiments testing 1,454 infants between 4.5 and 11.1 months. Children were tested in the headturn preference procedure (Kemler Nelson, et al., 1995; 59 samples) or the central fixation paradigm (Graf-Estes \& Lew-Williams, 2015; 9 samples).

## Effect Size Calculation

All scripts and raw data are available on github. ${ }^{2}$ The effect size we report here is a standardized mean difference of infants' looking behavior when listening to statistical words versus non-words. Since a preference for non-words (novelty preference) is dominant in the literature, positive values reflect this direction of the effect. The larger the effect size, the bigger the observed standardized mean difference between the two types of test trials. In turn, negative values indicate that infants demonstrated a familiarity preference, i.e. they listened longer to statistical words over non-words ${ }^{3}$.

[^27]We computed Hedges' $g$ (Morris, 2010), a variant of Cohen's $d$ (Cohen, 1988) that is preferred in the case of small sample sizes. Effect sizes were calculated based on reported test statistics: for 50 samples we could use means and standard deviations of test trials; for 17 samples $t$-values for the main comparison were available. To ensure consistency in the direction of the effect, we re-coded $t$ values as positive when infants listened longer to statistical non-words and as negative otherwise. We used standard formulae for effect size calculation in within-participant designs (Lipsey \& Wilson, 2001, when means and standard deviations were available; Dunlap et al., 1996, for effect sizes based on $t$-values). One paper reported betweenparticipant results and we computed effect sizes and variances from means and standard deviations accordingly (Lipsey \& Wilson, 2001). When the same infants contributed to multiple effect sizes, we computed the median of all critical values to ensure independent samples (here, 4 effect sizes were derived from 8 non-independent samples). We could not compute effect sizes for 6 additional experiments, due to lack of information.

Only one of the 20 papers included reported correlations between test trials, which capture the dependency between the two data points stemming from the same participants and are necessary for $t$-value based effect size and general effect size variance calculation. We imputed random values based on the distribution of correlations reported in a similar meta-analysis (Bergmann \& Cristia, 2016; updated data available via metalab.stanford.edu). ${ }^{4}$

## Meta-Analysis

To establish the size and variance of the effect, we fitted a multivariate random effects model using the R ( R core team, 2016) package metafor (Viechtbauer, 2010). Random effects models assume that all effect sizes are sampled from a distribution of effect sizes and try to estimate the mean and variance of this distribution. In the multivariate model, the interdependence between effect sizes from the same paper is taken into account, yielding a more robust measure

[^28]of the true effect. To investigate the impact of additional variables, we introduce moderators to this model.

## Bias

We tested for bias in the published literature by assessing funnel plot asymmetry, which is significant when a portion of the expected distribution of effect sizes around the weighted mean is missing, yielding an over-representation of a part of the underlying effect size distribution. We test for asymmetry using the rank correlation test (implemented in metafor; Viechtbauer, 2010).
To further investigate biases, we make use of p-curves to test whether there is an excess in $p$-values just below the significance threshold of .05 and if the distribution of $p$ values indicates an underlying real effect (Simonsohn, Nelson, \& Simmons, 2014). To this end, we enter all exact $t$-values that were reported ( $\mathrm{n}=48$ for the whole dataset).

## Results

## Original Paper

We first calculated the effect size and its variance for the two experiments reported by Saffran, Aslin, and Newport (1996). Hedges' $g$ was $0.4(\mathrm{SE}=0.040)$ for experiment 1 and 0.38 ( $\mathrm{SE}=0.041$ ) for experiment 2. According to Cohen's (1988) criteria this is a small to medium effect.

If experimenters base their sample size decisions on this effect size, they would have to test 53 infants in a paired samples design to achieve $80 \%$ power (computed with the R package pwr; Champely, 2016). The median sample size in our dataset is 22 participants, which would mean a $42 \%$ probability of obtaining a significant result, assuming the effect is of the size reported in the initial study; inversely, $58 \%$ of attempts to replicate this finding should fail.

## Conceptual Replications

First, we report on the experiments that were identified as replications of the original report (Saffran et al., 1996). Seventeen experiments could be included in these analyses.

Meta-Analytic Effect The variance-weighted effect size Hedges' $g$ is $0.21(\mathrm{SE}=0.1)$, which is significantly different from zero ( $95 \%$ CI $[0.02,0.4], p=.03$ ) and indicates a preference for statistical non-words. Note that this effect is smaller than the original report, and typical power is thus only $16 \%$ with 22 participants. Heterogeneity is significant, indicating variance in the data that is not explained by random measurement error $(\mathrm{Q}(16)=71, p<.001)$.

Moderator Analysis: Age We find no significant effect of the moderator centered age in days $(\mathrm{Q}(1)=0.6, \beta=-0.001$, $\mathrm{SE}=0.0015,95 \% \mathrm{CI}[-0.004,0.018], p=.5)$.

Moderator Analysis: Stimuli Naturalness Studies on SL differ in the stimuli; in this dataset, 11 effect sizes came from experiments with synthetically generated speech, 6 were based on experiments with naturally produced speech.


Figure 1: Funnel plot (code adapted from Sakaluk, 2016) showing standard error of the effect size as a function of effect size for 17 conceptual replications. The solid line marks zero, the dashed line the effect estimate, and the grey line indicates the funnel plot asymmetry.

Overall, the moderator test is significant $(\mathrm{Q}(1)=5, p=$ .023 ) with a negative estimate $(\beta=-0.35, \mathrm{SE}=0.16,95 \%$ CI $[-0.66,-0.05]$ ), indicating that infants tend to show less of a novelty preference with stimuli produced by human speakers.

Follow-up analyses focusing on subsets revealed that synthetically produced stimuli lead to a significant positive effect (Hedges' $g=0.32, \mathrm{SE}=0.05,95 \%$ CI [0.2, 0.4$], p<$ .001 ), while those replications relying on naturally-produced speech yield an effect size not different from zero (Hedges’ $g=0.02, \mathrm{SE}=0.2,95 \% \mathrm{CI}[-0.36,0.41], p=.9)$.

Publication Bias The funnel plot shown in Figure 1 displays a greater density of large effect sizes that are of low-precision (lower right quadrant) and some effect sizes that are of high precision but outside the expected distribution (upper left quadrant), which is illustrated further by the linear regression line in grey. This line should be horizontal in the case of an even distribution around the median effect. Nonetheless, asymmetry is not significant with Kendall's $\tau=.26, p=.15$.

The p-curve analysis based on the 6 significant $t$-values available in this dataset indicates a flat distribution of $p$ values, as would be expected when there is no underlying effect ( $\mathrm{Z}=-0.43 ; p=.33$ ). However, these $6 t$-values might not be representative of the 17 studies analyzed here.

## Complete Literature

Meta-Analytic Effect When taking into account all 68 independent effect sizes, the meta-analytic effect size Hedges' $g$ is 0.09 ( $\mathrm{SE}=0.05$ ), which is not significantly different from zero (CI [-0.02, 0.19], $p=.1$ ). This dataset, however, includes a number of samples that explicitly pit TPs against other segmentation cues, and thus may be expected to lead to different effects represented within the same data. Indeed, heterogeneity is significant $(Q(67)=$ $334, p<.001$ ). We thus analyze each of our moderators.


Figure 2: Effect size by participant age for all samples; point size is inverse variance. Black refers to synthetic, grey to natural speech. The dashed line indicates zero.

Moderator Analysis: Age As described in the introduction, more mature infants might show a different direction of preference or larger effect. However, we find no (linear) effect of age $(\mathrm{Q}(1)=0.3, p=.6)$. Follow-up analyses introducing a quadratic term for age confirmed this finding.

Moderator Analysis: Stimuli Naturalness In the full dataset, the use of artificial and natural speech is fairly balanced, with 38 instances of computer-generated stimuli and 30 of human speakers. The moderator test is significant $(\mathrm{Q}(1)=11, p<.001)$, and the results mirror our findings in the conceptual replication dataset. Figure 2 displays all samples, with color encoding natural (grey) vs artificial (black) stimuli. The meta-analytic effect for experiments with artificial stimuli is significantly above zero (Hedges' $g$ $=0.23, \mathrm{SE}=0.06,95 \%$ CI $[0.11,0.35], p<.001)$. In contrast, natural speech yields an effect not different from zero (Hedges' $g=-0.05,95 \%$ CI $[-0.2,0.06], p=.4$ ).

Moderator Analysis: Cue conflict Cues can either be absent ( $\mathrm{n}=20$ ), congruent with TPs (32), or in conflict with statistical information (16). Those cues encompass word stress (8), sentence level prosody (3), duration (2), intensity (2), and co-articulation (1). We predicted that cues that coincide with TPs might strengthen the effect, while those that conflict with TPs may reveal a different, possibly even opposing effect. We therefore introduced a three-leveled moderator. This analysis revealed no significant moderator effect $(\mathrm{Q}(2)=1.9, p=.4)$.

Of the 48 samples that involve additional cues, 24 are based on the effect of a correlate of word-level stress on segmentation. These studies propose that infants will be driven to segment speech using a trochaic stress pattern, in line with their native language. Artificial languages with trochaic stress are therefore congruent with TP cues, and are predicted to lead (as a whole) to novelty preferences; those with iambic stress conflict with TPs, and are predicted to lead (as a whole) to null or familiarity preferences.


Figure 3: Funnel plot of all samples. For details see Figure 1.

A moderator analysis restricted to samples with additional stress-based segmentation cues fails to confirm this prediction $(Q(1)=0.7, p=.4$; Cue conflict [iambic stress]: $\beta$ $=-0.07, \mathrm{SE}=0.08,95 \% \mathrm{CI}[-0.23,0.09])$.

Publication Bias Figure 3 shows an even distribution of effect sizes around the estimated median, the large spread illustrating the unexplained heterogeneity. The ranktest indicates no significant asymmetry (Kendall's $\tau=-.01, p=$ .9; see also grey linear regression line in Figure 3).

The p-curve based on 34 significant $t$-values indicates that the data contain evidential value $(\mathrm{Z}=-2.47 ; p=.007$ for the full p-curve) and there is no excess of "just significant" pvalues. Power based on the p-curve is estimated to be $25 \%$.

## Discussion

In the present paper, we examine infants' ability to track transitional probabilities (TPs) in continuous streams of speech. Experiments replicating the original Saffran et al. (1996) paradigm reveal a significant and reliable effect (Hedges' $g=.21$ ) that is on par with the effect found in the meta-analysis of natural speech segmentation (Hedges' $g=$ .22; Bergmann \& Cristia, 2016), albeit in the opposite direction of preference. An analysis of the whole literature fails to find a significant aggregated effect, but is reliably influenced by naturally vs. synthetically produced speech. There was no evidence for a developmental shift in or strengthening/weakening of preference, nor for a consistent and reliable role of additional cues. Finally, there is no clear evidence for publication bias. Taken together, these results invite deeper consideration of several issues in the future study of SL and theories of language acquisition, discussed in turns below.

## One Mechanism Among Many

The data presented here confirm that infants can track statistically defined patterns and use that information to segment a stream of speech into word-like units. The strength of this capacity, however, may be more fragile than expected. How are we to understand these findings, as we
continue to examine the import of statistical learning in language acquisition?

When aggregating across different studies, we put to the test the idea that researchers can predict the direction of infant looking-time preferences. Most popular theories of infant preference (e.g. Hunter \& Ames, 1988; Kidd, Aslin \& Piantadosi, 2012; 2014) predict an interplay between stimulus complexity and infant readiness to encode this complexity. In the case of TP-based word segmentation, we therefore expected a linear (or quadratic) shift from familiarity to novelty preferences as infants age. We instead find a consistent novelty preference. On the other hand, there is a significant effect of stimuli naturalness: While studies using synthesized speech yield reliable novelty preferences, studies using naturally produced speech fail to find reliable effects. It is likely that natural speech, even when altered to be largely monotonic and lacking syllable co-articulation, is more acoustically complex than synthetic speech. This is supported by the consistent familiarity preference across age groups found by Bergmann \& Cristia (2016). Infants may thus be more likely to show a familiarity preference to natural speech because it may take more time to process (and hence habituate to/learn from) this complex signal. There is some evidence in the SL literature to support this idea: some studies find alternating patterns of looking preference by block (e.g. Graf-Estes \& Lew-Williams, 2015). This, however, is rarely reported. Future investigations based on the meta-analytic data presented here might pursue the role of stimulus complexity by assessing the possible interactions between stimulus type, familiarization duration, age, and direction of lookingtime preference.

Several of the studies in the dataset were designed to test the limits of SL. They have been included because in all cases infants might have opted to segment the language based on TPs alone; we hypothesized that, once taken in sum, these studies might have revealed evidence that TPs drive segmentation even in the face of alternative cues. This did not turn out to be the case - there is no reliable effect for segmentation when all studies are considered together. Moreover, and surprisingly, there is no pattern that unites samples in which cues are congruent with TPs vs those in conflict with TPs. These results, in fact, suggest that infants only succeed at tracking TPs when presented with artificial speech sounds. Given the results of the Bergmann \& Cristia (2016) meta-analysis, we find this unlikely to reflect the true state of the world; rather, we believe it suggests that what does drive performance in the relatively simple paradigm of TP-based word segmentation remains underspecified and requires further theoretical, experimental, and metaanalytical consideration. Future work extending from the current dataset will aim to contribute to this discussion by accruing enough data to be able to examine additional moderators (e.g. familiarization duration) and outcome variables (i.e. effect sizes based on proportions of infants showing the effect, as opposed to standardized means of looking-time differences).

## Practical Implications

There are several points to take into account when planning future SL word segmentation studies. First, assuming an effect size of Hedges' $g=.21$, the power of a typical 22 sample design is a meagre $16 \%$ (note that the pcurve analysis indicates an overall power level of $25 \%$ in the significant portion of the studies). A well-powered study ( $80 \%$ ) would require a sample of 180 infants ( 142 if the direction of the preference can be predicted). This is impractical in the current state of infant research which relies on single labs conducting such studies (but see the alternative collaborative approach outlined by Frank et al., 2017). We do not intend to suggest that SL is not worth investigating - but it does call into question the methods with which we choose to investigate it. Power might, for example, be increased with more robust methods, calling for infant researchers to improve extant paradigms. At this point, we are only beginning to have sufficient power to fully understand the role of methods, stimuli, and test set-up (see e.g., Frank et al., 2017). One possibility lies in adopting more implicit measures of SL such as through neuroimaging, which may be less susceptible to factors affecting the direction of infant looking-preference.

## Limitations

Any meta-analysis is limited by a number of factors, one of which is that the analysis is only as good as the data it contains. In other words, the studies reported here are those that have been published (or made available online) and were findable through our search criteria (see supplementary material for a full list of included studies). Since the effect is small, we expect that a number of failures to replicate the original finding are confined to the file-drawer, simply because they were underpowered. Further, studies showing a familiarity preference might not be published as those are not expected in replications of Saffran et al., (1996). Including such (presumed) file-drawer studies would make our estimates much more reliable and we strongly encourage researchers with unpublished work to contact the authors and contribute these findings (or any published data that may have been regrettably missed).

A second limitation is missing information. For example, in order to compute effect sizes and their variance for within-participant designs, it is necessary to know the correlation between infants' preferences for each test-item type. We have temporarily imputed these figures based on similar data (Bergmann \& Cristia, 2016), and ran additional analyses to confirm that different values result in similar outcomes. However, we hope that authors who can retrieve this data will be willing to enrich our dataset, and recommend to all to include this information in future publications

## Conclusion

This meta-analytic analysis of statistical learning as applied to word segmentation has revealed a reliable but small
effect. We hope that this paper promotes future research that will seek to better characterize infant performance on SL tasks, and will thus contribute to stronger theories and models of infant cognition and behaviour.

## Acknowledgments

The authors thank the Cognitive Science Society for hosting a meta-analysis tutorial in 2016 and for participants of this tutorial, who pre-coded some of the included papers. Alexis Black was supported by the UBC 4 Year Fellowship and NSERC Discovery Grant (Individual) to Carla L. Hudson Kam "Constraints on language acquisition and how they change (or don't) with age." Christina Bergmann was supported by the European Horizon 2020 programme (Marie Skłodowska-Curie grant No 660911), the Agence Nationale pour la Recherche (ANR-10- IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC), and the Fondation de France.

## References

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. Wiley Interdisciplinary Reviews: Cognitive Science, 8, e1373.
Bergmann, C., \& Cristia, A. (2016). Development of infants' segmentation of words from native speech: A metaanalytic approach. Developmental Science, 19, 901-917.
Champely, S. (2016). pwr: Basic Functions for Power Analysis. R package version 1.2-0. https://CRAN.Rproject.org/package $=$ pwr
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
Dunlap, W.P., Cortina, J.M., Vaslow, J.B., \& Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods, 1, 170-177.
Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Yurovsky, D. (under review). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Preprint posted on PsyArXiv at https://osf.io/preprints/ psyarxiv/27b43/
Graf Estes, K., \& Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. Developmental Psychology, 51, 1517-1528.
Hunter, M. A., \& Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. Advances in Infancy Research 5, 69-95.
Johnson, E. K., \& Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. Developmental science, 13(2), 339-345.
Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., \& Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. Infant Behavior and Development, 18, 111-116.

Kidd, C., Piantadosi, S. T., \& Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PloS One, 7, e36399.
Kidd, C., Piantadosi, S. T., \& Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. Child Development, 85, 1795-1804.
Lew-Williams, C., \& Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. Cognition, 122, 241-246.
Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., \& Frank, M. C. (2017). A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis. Preprint posted on PsyArXiv at https://osf.io/preprints/psyarxiv/htsjm/
Lipsey, M.W., \& Wilson, D.B. (2001). Practical metaanalysis. Thousand Oaks, CA: Sage.
Lloyd-Kelly, M., Gobet, F., \& Lane, P. C. (2016). Under Pressure: How Time-Limited Cognition Explains Statistical Learning by 8-Month Old Infants. Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 1476-1480). Austin, TX: Cognitive Science Society.
Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta- analysis on repeated measures. British Journal of Mathematical and Statistical Psychology, 53, 17-29.
Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. Language and Cognition, 8, 447-461.
Pearl, L., Goldwater, S., \& Steyvers, M. (2010). Online learning mechanisms for Bayesian models of word segmentation. Research on Language and Computation, 8, 107-132.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926-1928.
Sakaluk, J. (2016, February 16). 7. Make It Pretty: Forest and Funnel Plots for Meta-Analysis Using ggplot2. [Blog post]. Retrieved from https://sakaluk.wordpress.com/2016 /02/16/7-make-it-pretty-plots-for-meta-analysis/
Simonsohn, U., Nelson, L. D., \& Simmons, J. P. (2014). Pcurve: A key to the file-drawer. Journal of Experimental Psychology: General, 143, 534-547.
Thiessen, E. D., \& Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7to 9-month-old infants. Developmental Psychology, 39, 706-716.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36, 1-48.

# Narrowing of the Cone-of-Direct Gaze Through Reinforcement Learning 

# Sam Blakeman (sam.blakeman.15@ucl.ac.uk) \& Denis Mareschal (d.mareschal@bbk.ac.uk) 

Department of Psychological Sciences, Birkbeck, University of London


#### Abstract

The Cone of Direct Gaze (CoD) is described as the range of eye gaze deviations over which an observer reports gaze as being directed towards them. The CoD has been found to narrow with age across childhood (Mareschal et al. 2016). We investigated whether reinforcement learning, so critical in shaping eye gaze responses in infancy, was able to account for the emergence of a CoD and its narrowing in childhood. To this end, we adapted Triesch et al.'s (2006) reinforcement learning model by (1) defining a topology over object locations, and (2) introducing opponent non-linear reward profiles for looking at objects and caregivers. In Simulation 1 we show that these modifications give rise to a functional CoD in which there is reduced eye gaze following and increased fixation on the caregiver for locations with a small caregiver eye gaze eccentricity. In Simulation 2 we show that the width of this effect reduces with learning, suggesting that developmental decreases in the CoD may be driven by reinforcement learning. In Simulation 3 we explore how changes in model parameters can explain the CoD in high anxiety populations. Finally, the model provides one way of unifying the developmental gaze-following and CoD literatures, until now considered largely independent.


Keywords: Reinforcement Learning; Cone of Direct Gaze; Gaze Following; Development; Social Anxiety; Autism;

## Background

The eyes are a key aspect of social intelligence. From the eyes one can infer an individual's emotions or desires, which can help guide social behavior, and also learn about the surrounding environment (Shepherd et al., 2010). Through joint attention, individuals can alert others to interesting objects in the environment by guiding their attention to that object. Eye gaze following is one form (Scaife and Bruner, 1975). In this seminal study the authors showed that infants were able to interpret the direction of another individual's eye gaze could and use that as a cue to look in the perceived direction. This allows infants to find objects of interest in the environment and learn from experienced caregivers in a non-verbal manner. Indeed, infants have been shown to be sensitive to eye gaze from a very young age, appearing to show a preference for eyes over other parts of the face (Hains \& Muir, 1996).

Infants with autism have a reduced ability to follow eye gaze (Leekam et al., 1997). A recent study by Thorup et al. (2016) found that infants at high risk of developing autism rely disproportionally on directional information from the head as compared to the eyes. This reduced ability to follow eye gaze may be a contributing factor to the deficits in social cognition and communication associated with autism.

While joint attention via eye gaze following appears to be a crucial tool for the developing infant, the perception of
eye gaze direction is not uniform across eye gaze deviations. The Cone-of-Direct gaze (CoD), is defined as the range of gaze deviations that we perceive to be looking directly at us (Gamer and Hecht., 2007). The CoD, therefore, has implications for how we perceive social situations and our interpretation of eye gaze. For example, if a gaze deviation falls inside our CoD then we may perceive it as looking directly at us and not engage in any eye gaze following behavior.

The perception of whether an individual is looking directly at you or not is also of particular interest to those investigating social anxiety disorders (Schulze et al., 2013). For example, a study by Jun et al. (2013) reported a wider CoD for high socially anxious males compared to low socially anxious males. Similarly, Gamer et al. (2011) conducted a study where participants had their CoD measured in response to a virtual head. They found that participants with social phobia had a wider CoD in the presence of a second virtual head that was directed at them. Such studies suggest that a wider CoD is associated with social anxiety and may play a role in the disorder.

Both eye gaze following and the CoD undergo changes during development. Eye gaze following emerges and improves during infant development (Brooks and Meltzoff., 2005, Deak., 2015), while the CoD becomes narrower during childhood (Mareschal et al., 2016). This equates to older children being more reliable at following eye gaze and more accurate at interpreting small eye gaze deviations as not being directed at them. It is possible that these developmental timelines for eye gaze following and the CoD are crucial for infant and child development and may be altered in clinical disorders such as autism and social anxiety. It is therefore important to understand their emergence and developmental trajectory.

Reinforcement learning (Sutton and Barto, 1998) has received much interest in recent years and there is now good evidence of the neurocomputational basis of reinforcement learning (Schultz et al., 1997). It has been proposed as a possible mechanism for the emergence and improvement of eye gaze following in infants. Triesch et al. (2006), describe a reinforcement learning model in which rewards obtained by following a caregiver's gaze led to the reinforcement of eye gaze following behavior. According to this account, infants associates the rewarding object that the caregiver is looking at with the act of following the caregiver's gaze, thereby building a predisposition to follow gaze as a consequence of experience rather than an innate behavior.

While reinforcement learning may account for the emergence of eye gaze following, the mechanism behind the emergence of a CoD is yet to be elucidated. To what extent is the CoD and its development the result of an innate prior
and to what extent is it the result of learning and the external environment? To investigate such a question, we explored whether the reinforcement learning framework could also account for the CoD and its changes through development. If reinforcement learning where to play a role in the emergence of a CoD then it may provide a link between eye gaze following and the CoD. It would also highlight reinforcement learning as a promising target for the therapeutic investigation of disorders such as autism and social anxiety.

## Triesch et al.'s (2016) Model

Triesch et al.'s (2006) model serves as a spring board for this study. The model consists of an infant, a caregiver and an object (Figure 1). Both the infant and caregiver remain in fixed positions while the object is able to move around N discrete locations. Two parameters, T_min and p_shift are responsible for the movement of the object around these locations. T_min specifies the minimum amount of time an object must spend in a location, while $p_{-}$shift specifies the probability of shifting to a new location per time step after T_min. This shifting of the object also determines the shifting of the caregiver's gaze.


Figure 1. Diagram of the gaze following reinforcement learning model proposed by Triesch at al. (2006).

When the object moves to a new location the caregivers gaze is shifted to a new location. The caregiver can look at $\mathrm{N}+1$ potential locations; the N locations the object can reside in plus the location of the infant. The parameter p_valid determines the probability that the new location of the caregiver's gaze is the same as the new location of the object. This probability effectively models two scenarios. The first scenario is when the caregiver's gaze may not be a $100 \%$ predictive of where the object is and the second scenario is when the infant's interpretation of the caregiver's gaze may not be a $100 \%$ accurate. The p_valid parameter accounts for both of these scenarios because both an inaccurate caregiver gaze or a poor interpretation of the caregiver gaze will lead to the infant following the caregivers gaze to an incorrect location.

The infant's behavior is modelled using a reinforcement learning framework whereby it is essentially driven to maximize rewards in the environment. The infant is broken up into two agents, a 'when' agent and a 'where' agent. The when agent is responsible for deciding whether it is time to shift gaze on a time step and the where agent is responsible
for deciding where to shift the gaze. These decisions are driven by the rewards encountered in the environment by the infant. In this environment the infant has four possible views, each of which having an associated reward:

1. An empty location ( $\mathrm{R}_{\text {nothing }}$ )
2. A location containing the object $\left(\mathrm{R}_{\text {object }}\right)$
3. A profile view of the caregiver $\left(\mathrm{R}_{\text {profile }}\right)$
4. A frontal view of the caregiver as they look directly at the infant ( $\mathrm{R}_{\text {frontal }}$ )
For each of these views the infant receives the associated reward $\left(R_{x}\right)$ multiplied by a habituation value. This habituation value exponentially decreases as the infant fixates on a location. The degree of this decrease is controlled by the habituation parameter beta ( $\beta$ ). Equally, the reward value for locations that have been habituated to, but the infant is no longer looking at, recover at the same rate. Habituation is important in a reinforcement learning framework such as this because otherwise the infant could just fixate on a single reward (e.g. the caregiver), and never have the motivation to shift gaze.

Taking the reward structure and habituation into account, the state-space of the when agent becomes two dimensional. The first dimension is how long has the infant been looking at the same location and the second dimension is the reward received by the infant. Representing the state-space with these two dimensions allows the when agent to decide whether to carry on looking at the same location or look somewhere else. If the decision is made to look somewhere else, then the where agent then specifies the location of the new gaze. The state-space of the where agent varies along a single dimension, which represents the gaze of the caregiver. This corresponds to $\mathrm{N}+2$ states. There are N number of states for when the caregiver is looking at each of the N object locations. It is these states that the infant uses to interpret where the caregiver is looking. Another state is for when the caregiver is looking directly at the infant and a final state is for when the gaze of the caregiver is unknown to the infant. While the action-space of the when agent is simply stay or move, the action-space of the where agent is of size $\mathrm{N}+1$. The where agent can decide to shift gaze to one of the N object locations or to look directly at the caregiver.

Both the when and where agents learn using temporal difference (TD) learning and the SARSA algorithm (Rummery and Niranjan, 1994; Equation. 1).

```
\(\delta_{t}=r_{t}+\gamma Q_{t}\left(s_{t+1}, a_{t+1}\right)-Q_{t}\left(s_{t}, a_{t}\right)\)
    \(Q_{t+1}\left(s_{t}, a_{t}\right)=Q_{t}\left(s_{t}, a_{t}\right)+\alpha \delta_{t}\)
    \(s_{t}=\) state at time \(t\)
    \(a_{t}=\) action at time \(t\)
    \(r_{t}=\) reward at time \(t\)
    \(\delta_{t}=\) temporal difference error at time \(t\)
    \(Q_{t}\left(s_{t}, a_{t}\right)=\) state - action value at time \(t\)
    \(\alpha=\) learning rate
    \(\gamma=\) discount factor for future rewards
```

In this framework the temporal difference error is calculated and the parameter gamma $(\gamma)$ is used as the discount factor
for future rewards. This temporal difference error is then used to update the appropriate state-action value and the size of this update is controlled by the learning parameter alpha $(\alpha)$. A softmax function was used to mapped the state-action values to actions. This allowed for a balance between exploration and exploitation, determined by the parameter tau was responsible for this balance. A larger value of tau results in more exploration and less exploitation of the stateaction values.

## Simulation 1 - Emergence of the Cone of Direct Gaze

To investigate whether reinforcement learning could also account for the emergence of a CoD , the reward structure of Triesch et al.'s model was modified to include a spatial topology in the reward space

## Methods

A CoD is inherently a spatial phenomenon that is assessed by having an individual look further and further away from a participant until the participant judges the gaze to no longer be directed at them. For this reason, the N object locations in Triesch et al.'s model were first given a spatial location identity based on the gaze deviation required from the caregiver to look at them (Figure 2). Values stepped four degrees at a time and the N locations were arranged in a linear manner. This alteration to Triesch et al.'s model allowed for the analysis of gaze following behaviour based on the spatial location of the caregiver's gaze.


Figure 2. Layout of the object locations in the model.
Next, the model's reward structure was modified. Specifically, the caregiver profile reward ( $\mathrm{R}_{\text {profile }}$ ) and object reward ( $\mathrm{R}_{\text {object }}$ ) were changed in an opposing manner using Gaussian functions (Figure 3).

Various studies have shown that infants prefer direct gaze over averted gaze (e.g., Farroni et al., 2002), which in the model equates to $\mathrm{R}_{\text {frontal }}>\mathrm{R}_{\text {profile. }}$. In terms of a CoD, small eye gaze deviations are likely to be interpreted as being direct and so $\mathrm{R}_{\text {profile }}$ should have a higher reward value for small eye gaze deviations compared to large eye gaze deviations. A Gaussian function was therefore applied to $\mathrm{R}_{\text {profile }}$ so that it increased in value as the caregiver looked at locations which required smaller eye gaze deviations.

The opposite transformation was applied to $\mathrm{R}_{\text {object }}$ so that it decreased in value as the caregiver looked at locations which required smaller eye gaze deviations. This aimed to represent the fact that objects outside of the infant's current
visual field are more likely to be unexpected ( so informative) and therefore more rewarding than objects that currently reside in the visual field. These two modifications to the reward structure had the net effect of increasing the caregiver's relative reward at smaller eye gaze deviations and increasing the object's relative reward at large eye gaze deviations.


Figure 3. The modified reward structure for the object and the profile view of the caregiver.

To evaluate the effect of these modifications 500 simulations were run for 100,000 learning iterations to establish gaze following. After learning, each simulation was run for 10,000 iterations without any learning to gather stable gaze following measurements, which were then averaged across simulations. The model parameters were as follows: Number of locations ( N ) $=10$; Degree step per location (D) $=4$; Reward for looking at empty location $\left(\mathrm{R}_{\text {nothing }}\right)=1$; Reward for looking at the frontal view of the caregiver $\left(\mathrm{R}_{\text {profile }}\right)=1$; Peak Reward for looking at the object $\left(\mathrm{R}_{\text {object }}\right)=1$; Sigma of the Gaussian applied to the object reward $\left(\mathrm{S}_{\text {object }}\right)=9$; Sigma of the Gaussian applied to the caregiver $\left(\mathrm{S}_{\text {profile }}\right)=9$; Habituation rate $(\beta)=0.5$; Learning Rate $(\alpha)=0.0025$; Discount Factor $(\gamma)=0.8$; Exploration vs. exploitation $(\tau)=0.095$; Minimum fixation time $\left(T_{\min }\right)=4$; Probability of shifting $\left(\mathrm{p}_{\text {shift }}\right)=0.5$; Predictiveness of caregiver gaze $\left(p_{\text {valid }}\right)=0.75$. Unless otherwise stated, these values were used in all simulations

## Results

Two measurements were used to assess the effect of the modified reward structure. The first measurement was the mean time spent by the infant fixating on the caregiver. This represented how long the infant looked at the caregiver before shifting gaze and served as an indirect measure of the probability of shifting gaze. The second measurement was the total number of gaze follows made by the infant. This was a direct measure of eye gaze following behaviour. Both of these measurements were examined as a function of object location.

After implementing the reward structure in Figure 3, the mean time spent fixating on the caregiver was larger when the caregiver was looking at locations that required small eye gaze deviations (Figure 4, left panel). This contrasted with the model's performance when endowed with a flat reward profile. In addition, the total number of gaze follows
was smaller when the caregiver was looking at locations that required small eye gaze deviations (Figure 4, right panel). These findings are consistent with the concept of a CoD . By fixating on the caregiver for longer during small eye gaze deviations the infant acts as if the caregiver is looking directly at them and is unable to follow their gaze to another location. Similarly, the increased number of eye gaze follows for large eye gaze deviations indicates that the infant is correctly classifying them as indirect and can therefore follow them to the object. These findings suggest that a CoD can emerge under a reinforcement learning framework where the caregiver and object rewards act in an opposing manner.


Figure 4. (Left panel) Mean time spent by the infant fixating on the caregiver as a function of the caregiver's gaze. The dotted magenta line represents the results when the object reward and caregiver profile reward are not modified, as in the original Triesch et al. (2006) model. (Right pane) Total Number of gaze follows made by the infant as a function of the caregiver's gaze. The dotted magenta line represents the results when the object reward and caregiver profile reward are not modified.

## Simulation 2 - Developmental Trajectory of the Cone of Direct Gaze

After confirming that reinforcement learning could lead to the emergence of a CoD, we investigated the effect of reinforcement learning on the CoD over time to see if it could also explain known developmental changes.

## Methods

In order to get a measure of the width of the induced CoD , the mean fixation duration and the number of gaze follows were overlaid and their intersects calculated. To achieve this, it was first necessary to rescale the feature so that both measurements were operating on the same scale (Equation 2). Each value had the minimum value subtracted and this was then divided by the range of the values. This produced a final value that ranged between 0 and 1 . Gaussian curves were then fit to both feature scaled measures, with the mean fixation time requiring a single term and the number of gaze follows requiring two terms. Finally, the two intersection points of the fitted Gaussian curves were calculated and the width between the two points
was taken as a proxy for the width of the CoD in the model (Figure 5).

Equation 2

$$
x^{\prime}=\frac{x-\min (x)}{\max (x)-\min (x)}
$$

To observe the change in this width over time, 500 simulations were run for $1,000,000$ learning iterations. At 100,000 learning iteration intervals, learning was halted and 10,000 iterations were run to gather stable gaze following measurements. These results were averaged across all simulations for each break in the learning process.


Figure 5. Quantification of the width of the CoD effect. After feature scaling each measure, Gaussians were fit to both of them. The intersect points of these Gaussians were then calculated and the horizontal distance between the intersection points was taken as a proxy for the width of the CoD effect.

## Results

The CoD width was found to decrease as the number of learning iterations increased (Figure 6). This is consistent with the finding that the CoD decreases during child development (Mareschal et al., 2016) and suggests that reinforcement learning may be one explanation for these changes.


Figure 6. Change in the width of the CoD effect as a function of the number of learning iterations.

## Simulation 3 -High Anxiety Populations

In this simulation, we explore different parameter values in an attempt to capture known differences in the CoD for individuals with social anxiety.

## Methods

The peak reward value for both the frontal and caregiver rewards ( $\mathrm{R}_{\text {frontal }}, \mathrm{R}_{\text {profile }}$ ) were systematically decreased. For each set of $R_{\text {frontal }}$ and $R_{\text {profile }}$ values 500 simulations were run for $1,000,000$ learning iterations. At 200,000 learning iteration intervals, learning was halted and 10,000 iterations were run to gather stable measurements. These results were averaged across all simulations for each break in the learning process. The same method was used for decreasing values of $\beta$ (habituation rate) but results were taken at 100,00 learning iteration intervals. The $\beta$ values allowed for a finer temporal resolution than adjusting the $\mathrm{R}_{\text {frontal }}$ and $\mathrm{R}_{\text {profile }}$ values because when $\mathrm{R}_{\text {frontal }}$ and $\mathrm{R}_{\text {profile }}$ were set to a value of 0.5 the CoD effect broke down at 100,000 iterations.

## Results

Socially anxious individuals have a wider CoD than control individuals (Jun et al., 2013; Gamer et al., 2011). One possible explanation for this is that they are avoiding eye contact (Schneier et al., 2011; Schulze et al., 2013) and therefore may have a reduced 'caregiver' reward compared to non-socially anxious individuals. To test this hypothesis, we reduced the rewards associated with the caregiver ( $\mathrm{R}_{\text {frontal }}, \mathrm{R}_{\text {profile }}$ ) and looked at the effect on the width of the CoD . Simulations were run for longer than previous simulations because the wider CoD for highly anxious individuals should be present during adulthood (Jun et al., 2013; Gamer et al., 2011). Reducing $\mathrm{R}_{\text {frontal }}$ and $\mathrm{R}_{\text {profile }}$ did not result in a wider CoD (Figure 7). On the contrary, the simulations suggest that after 600000 iterations, there is a trend towards a narrower CoD when $\mathrm{R}_{\text {frontal }}$ and $\mathrm{R}_{\text {profile }}$ were reduced.


Figure 7. Change in the width of the CoD effect for different values of $R_{\text {profile }}$ and $R_{\text {frontal }}$ as a function of the number of learning iterations.

Another common theory relating to social anxiety is the hyper-vigilance-avoidance hypothesis (Horley et al., 2004) (Wieser et al., 2009). This hypothesis states that socially anxious individuals are hyper-vigilant towards anxiety provoking stimuli and tend to engage in avoidance by looking away. To investigate whether such a hypothesis could account for the wider CoD in socially anxious individuals we increased the value of the habituation parameter $\beta$. The goal of this modification was to reduce the
infant's gaze fixation time on the caregiver, as would be expected from avoidance. Increasing the value of $\beta$ resulted in a progressively narrower CoD effect (Figure 8). This effect was evident after around 200,000 learning iterations.


Figure 8. Change in the width of the CoD effect for different values of $\beta$ as a function of the number of learning iterations. Larger values of $\beta$ resulted in a narrower CoD effect.

## Discussion

We have demonstrated that a reinforcement learning account of eye gaze following behavior can be extended to account for the emergence and development of a CoD. In addition, the model also captured the developmental narrowing of the CoD (Mareschal et al., 2016). While a preference for direct gaze may be present from birth for example (Farroni et al., 2002), the fact that the CoD appears to narrow under the influence of reinforcement learning, as seen in developing children, suggests that at least some aspects of the CoD are experience dependent.

The fact that the CoD appears to be influenced by experience and learning poses interesting questions for researchers investigating clinical populations of socially anxious individuals. Importantly it suggests that a critical developmental period may exist that could act as a therapeutic window to reduce the occurrence of behaviors such as social anxiety. We used the model to investigate which aspects of the reinforcement learning framework could influence the developmental trajectory of the CoD. One theory for why socially anxious people may have a wider cone than control individuals is because of their aversion to direct eye contact (Schneier et al., 2011; Schulze et al., 2013). To probe this further, we reduced the rewards associated with the caregiver ( $\mathrm{R}_{\text {frontal }}$ and $\mathrm{R}_{\text {profile }}$ ) and looked at the effect on the width of the CoD. Reducing $R_{\text {frontal }}$ and $\mathrm{R}_{\text {profile }}$ resulted in a trend towards a narrower CoD , the opposite of what is seen in highly anxious individuals.

As an alternative to reducing $\mathrm{R}_{\text {frontal }}$ and $\mathrm{R}_{\text {profile }}$, we also investigated the effect of increasing the value of the habituation parameter $\beta$. This was done in an attempt to capture the hyper-vigilance-avoidance hypothesis, which states that socially anxious individuals are quicker to engage and then avoid anxiety provoking stimuli. An increase in the value of $\beta$ however, did not lead to a wider CoD. That said, care must be taken when making conclusions from this result. We used mean the infant's fixation time on the
caregiver as a functional measure of COD. This measure was used because under a CoD an infant should fixate for longer at small eye gaze deviations because it judges the gaze to be direct. However, this poses a problem when we want to model hyper-vigilance and avoidance by increasing $\beta$. An individual exhibiting hyper-vigilance and avoidance will have a reduced mean fixation time on the 'caregiver' even if they perceive the gaze to be directed at them because they would rather shift their gaze to nothing than hold direct gaze. So while increasing the value of $\beta$ may capture this behaviour, the measure of mean fixation duration will be lower for these cases causing the width of the CoD effect to be smaller even if the CoD is actually wider. Therefore, in order to accurately assess the effect of hyper-vigilance and avoidance on the width of the CoD effect, a different measure that captures when the infant perceives the gaze as being direct is needed.

Our findings (and Triesch et al's) have potential implications for several disorders other than social anxiety. The fact that both eye gaze following and the CoD appear linked by reinforcement learning could provide novel opportunities to investigate disorders that produce both characteristic eye gaze following and CoD behavior. One example of this is autism spectrum disorder. Triesch et al. (2006) put forth multiple candidates under the reinforcement learning framework that could produce the reduced eye gaze following described in individuals with autism. These candidates included a reduced learning rate, reduced caregiver reward and increased shifting latency. The fact that reinforcement learning can account for both eye gaze following and the CoD allows us to explore these candidates further. For example, studies have suggested that individuals with autism have a narrower CoD (Matsuyoshi et al., 2014) and so these candidates should be able to account for that. Indeed, in this study we demonstrated that reducing the caregiver rewards resulted in a slightly narrower cone in the later stages of a simulation. This finding lends weight to reduced caregiver rewards being present in autism. A similar approach should be taken for the other candidates to see if they too can account for both the eye gaze following and CoD differences described in autism, thereby either confirming or rejecting their validity.

## Acknowledgments

This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC), UK. We thank Isabelle Mareschal for very helpful input.

## References

Brooks R. \& Meltzoff A.N. (2005) The Development of Gaze Following and its Relation to Language. Dev. Sci, 8:535-543. Deak G.O. (2015) When and Where do Infants Follow Gaze? 5th International Conference on Development and Learning and on Epigenetic Robotics, Providence, RI USA.
Farroni T., Csibra G., Simion F. \& Johnson M.H. (2002) Eye Contact Detection in Humans From Birth. PNAS, 99: 9602-5.

Gamer M. \& Hecht H. (2007) Are you Looking at me? Measuring the Cone of Gaze. JEP: HPP, 33:705-15.
Gamer M., \& Hecht H., Seipp N. \& Hiller W. (2011) Who is Looking at me? The Cone of Direct Gaze Widens in Social Phobia. Cog. \& Emotion, 25:756-64.
Hains S.M.J. \& Muir D.W. (1996) Infant Sensitivity to Adult Eye Direction. Child Development, 67:1940-51.
Horley K., Williams L.M., Gonsalvez C. \& Gordon E. (2004) Face to Face: Visual Scanpath Evidence for Abnormal Processing of Facial Expression in Social Phobia. Psychiatry Res, 127:43-53.
Jun Y.Y., Mareschal I., Clifford C.W. \& Dadds M.R. (2013) Cone of Direct Gaze as a Marker of Social Anxiety in Males. Psychiatry Res, 210:193-8.
Leekam S.R., Baron-Cohen S., Perrett D.I., Milders M. \& Brown S. (1997) Eye-Direction Detection: A Dissociation Between Geometric and Joint Attention Skills in Autism. BJDP, 15:77-95.
Mareschal I., Otsuka Y., Clifford C.W., Mareschal D. (2016) "Are you Looking at me?" How Children's Gaze Judgments Improve with Age. Dev. Psy, 52:695:703.
Matsuyoshi D., Kuraguchi K., Tanaka Y., Uchida S., Ashida H. \& Watanabe K. (2014) Individual Differences in Autistic Traits Predict the Perception of Direct Gaze for Males, but not for Females. Molecular Autism, 5:12.
Rummery G.A. \& Niranjan M. (1994) On-Line Q-Learning Using Connectionist Systems. CUED/F-INFENG/IR 166. Cambridge University, UK.
Scaife M. \& Bruner J.S. (1975) The Capacity for Joint Visual Attention in the Infant. Nature, 253:265-66.
Schneier F.R., Rodebaugh T.L., Blanco C., Lewin H. \& Liebowitz M.R. (2011) Fear and Avoidance of Eye Contact in Social Anxiety Disorder. Compr Psychiatry, 52:81-7.
Schulze L., Renneberg B. \& Lobmaier J.S. (2013) Gaze Perception in Social Anxiety and Social Anxiety Disorder. Front. in Human Neuroscience, 7:872.
Schultz W., Dayan P. \& Montague P.R. (1997) A Neural Substrate of Prediction and Reward. Science, 275:1593-9.
Shepherd S.V. (2010) Following Gaze: Gaze-Following behaviour as a Window into Social Cognition. Front. in Integrative Neuroscience, 4:5.
Sutton R.S.\& Barto A.G. (1998) Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.
Thorup E., Nystrom P., Gredeback G., Bolte S., Falck-Ytter T. \& the EASE Team (2016) Altered Gaze Following During Live Interaction in Infants at Risk for Autism: an Eye Tracking Study. Molecular Autism, 7:12.
Triesch J., Christof T., Deak G.O. \& Carlson E. (2006) Gaze Following: Why (not) Learn it? Dev. Sci, 9:125-47.
Wieser M.J., Pauli P., Weyers P., Alpers G.W. \& Muhlberger A. (2009) Fear of Negative Evaluation and the Hypervigilance-Avoidance Hypothesis: An Eye-Tracking Study. Journal of Neural Transmission, 116:717-23.

# Functionally localized representations contain distributed information: insight from simulations of deep convolutional neural networks 

Nicholas Blauch (nblauch@umass.edu)<br>University of Massachusetts, Amherst<br>Center for the Neural Basis of Cognition, Pittsburgh PA

Elissa Aminoff
Department of Psychology, Fordham University Center for the Neural Basis of Cognition, Pittsburgh PA

Michael J. Tarr<br>Department of Psychology, Carnegie Mellon University Center for the Neural Basis of Cognition, Pittsburgh PA


#### Abstract

Preferential activation to faces in the brain's fusiform gyrus has led to the proposed existence of a face module termed the Fusiform Face Area (FFA) (Kanwisher et. al, 1997). However, arguments for distributed, topographical object-form representations in FFA and across visual cortex have been proposed to explain data showing that FFA activation patterns contain decodable information about non-face categories (Haxby et. al, 2001; Hanson \& Schmidt, 2011). Using two deep convolutional neural network models able to perform humanlevel object and facial recognition, respectively, we demonstrate that both localized category representations (LCRs) and high-level face-specific representations allow for similar decoding accuracy between non-preferred visual categories as between a preferred and non-preferred category. Our results suggest that neuroimaging of a cortical "module" optimized for face processing should yield significant decodable information for non-face categories so long as representations within the module are activated by non-face stimuli.


Keywords: module, localized categorical representation, distributed object-form topography, deep convolutional neural network, virtual electrophysiology

## Introduction

How are mental representations organized in the brain? Do certain brain regions contain functional modules, dedicated to representing and processing a very specific type of information? Or is neural real estate more generally involved in the processing of many different types of stimuli? Evidence from fMRI has been used to propose the existence of functional modules for the processing of certain classes of visual information within the brain. Cortical modularity was proposed first for the visual processing of faces in the socalled Fusiform Face Area (FFA) (Kanwisher et. al, 1997), then for the visual processing of scenes/places in the so-called Parahippocampal Place Area (PPA) (Epstein \& Kanwisher, 1998), and then for the visual processing of body parts in the so-called Extrastriate Body Area (EBA) (Downing et. al, 2001). In each of these studies, preferential activation of a certain class of visual stimuli (e.g. faces) in a certain region of the brain (e.g. fusiform gyrus) was used as evidence for
modular processing within that region, leading to the authors renaming the region in terms of the modular processing (e.g. Fusiform Face Area). Not all authors agreed that preferential activation of a cortical region by a certain stimulus class was convincing evidence of underlying modular processing. Haxby et. al (2001) used multi-variate pattern analysis (MVPA) to demonstrate that putative functional modules for processing of scenes in the PPA and faces in the FFA contain patterns of activation useful for decoding whether a subject is viewing one of two categories not thought to be processed within the module. These authors interpreted their findings in the context of an "object-form topography" model, in which the ventral temporal cortex possesses a distributed, topographical representation of object-form features which underlie all forms of visual recognition. In their account, the large responses found in proposed functional modules are complemented by small responses throughout ventral temporal cortex in computations underlying visual categorization and other aspects of visual cognition.

Later, Spiridon \& Kanwisher (2002) ran a similar fMRI study incorporating greater variability across images within a category (e.g. different viewpoints, exemplars, and image formats) in order to determine whether decodable abstract category information was truly distributed equally throughout ventral temporal cortex, as was argued by Haxby et. al (2001), or whether there might be localized decoding advantages corresponding to the locations of proposed functional modules. This study demonstrated that some abstract categorical information was present for certain categories outside their region of maximal activation (i.e. the location of a proposed module), replicating a main finding of Haxby et. al (2001). However, controlling for the number of voxels used in decoding analysis, this study demonstrated strong advantages in decodable information relating to discrimination between a preferred category (e.g. faces) and a non-preferred category (e.g. houses) in the region of proposed modularity (e.g. FFA). Additionally, in PPA and FFA, distinct disadvantages were found for the decoding of two non-preferred categories (e.g. faces vs. objects and objects vs. houses, respectively). Thus, while abstract
categorical information of certain categories may exist outside the region where modular processing is proposed, such abstract categorical information is by no means equally distributed throughout ventral temporal cortex. The authors thus argued for a more modular account of PPA and FFA, whereby these regions are primarily involved in the processing of a single category of information (scenes/houses and faces, respectively).

To account for both sets of findings, Cowell \& Cottrell (2013) performed multi-variate pattern analysis (MVPA) on a neurocomputational model capable of discriminating between the 6 visual categories used in the analyses of Haxby et. al (2001). The neurocomputational model first applies Gabor filtering of input images to obtain a perceptual representation; it then feeds the activations of many Gabor filters into a self-organizing Kohonen map, which utilizes unsupervised learning to cluster its inputs into a twodimensional representation. While the neurocomputational model contained no modular mechanisms, through unsupervised learning it developed the types of functionally localized stimulus representations used to argue in favor of modular processing, whereby certain patches of the Kohonen grid contained both preferential activation and enhanced decodable information about some categories. The effects were greatest for faces. Because they were able to simulate the data used to argue both for localized and distributed topographical representation with a neurocomputational model of distributed topographical representations that is more parsimonious than one postulating the existence of functional modules, the authors rejected the interpretation of a functional module for face processing based on the data of Spiridon \& Kanwisher (2002).

The result of Cowell \& Cottrell (2013) demonstrates that the evidence used to postulate functional modules may be accounted for by a model employing a distributed representation. However, the model is unable to account for human-level behavioral performance, and rather was constrained only to perform 6-way visual categorization. In the age of biologically-inspired computational systems capable of human-level object categorization (e.g. Krizhevsky et. al, 2012) and face individuation (e.g. Parki, Vedaldi, and Zimmerman, 2015), such behavioral constraints should become standard practice for models of neural representation. The decision not to constrain the computational model to perform human level face individuation, for example, belies the need for a functional module for face processing. Thus, we examine two deep, convolutional neural networks (DCNNs), one trained for large-scale object categorization and one trained for expert face individuation. In these networks, we focus on two types of category-specific representations for analysis. The first type of representation is the localized categorical representation (LCR), found in the final hidden layer of AlexNet (Krizhevsky et. al, 2012), whereby a single unit represents the likelihood of a given category in an image shown to the network. Such a localized categorical representation differs from the topographical object-form
representations proposed by Haxby et. al (2001), in that a single value represents the abstract category information. However, localized category representations receive input from a processing layer which is well-described as a topographical object-form representation; thus, they are not true "modules". The second type of representation is taken as a deep layer of face-specific representations within the faceindividuation network of Parki, Vedaldi, and Zimmerman (2015), VGG-Face, which is optimized for facial recognition only. In our view, VGG-Face in toto is a face-dedicated module; that is, a system optimized on and dedicated to to the processing of faces, only. The deep layer was chosen as a layer with high-level, complex face-specific features useful for recognition, but not explicitly representing individuals. We think that such representations are a reasonable model for what is proposed to be encoded in FFA (see Kanwisher \& Yovel, 2006). We perform "virtual electrophysiology," (Yamins \& DiCarlo, 2016) on both systems in order to determine whether these two types of category-specific representations produce the characteristic signal used to argue for distributed category-general representations: decodable information for non-preferred categories.

## Method

Model simulations were run in the MATLAB programming environment, using the MatConvNet toolbox (Vedaldi \& Lenc, 2015). Both models used in this study are examples of deep convolutional neural networks (DCNNs). Such networks were developed by computer vision researchers as engineering solutions for problems of visual recognition (e.g. LeCun et. al, 1998; Krizhevsky et. al, 2012). DCNNs contain several layers of processing, each of which contains a set of mathematical filtering operations (units or filters) which are convolved across the input, usually followed by a set of fully-connected layers which contain units which apply simple weighted summations of the units at the layer before. In all DCNNs for visual categorization, there exists a final layer of processing containing a set of units whose size is equal to the number of categories to be tested from, where each unit's activation corresponds to the likelihood that a certain category is present in the image; this vector of information is typically transformed via a softmax operation into explicit probabilities that the image may be categorized into each possible category.

The first DCNN model used is AlexNet (Krizhevsky et. al, 2012), pre-trained and uploaded to MatConvNet by Vedaldi \& Lenc (2015). AlexNet was trained to perform 1000-way categorization of visual images on the 2011 ImageNet training set, which contains 1.2 million images evenly distributed across 1000 categories. For simulations, a different set of images not used in training, the 2011 ImageNet validation set, was used as stimuli, containing 50 images for each of 1000 categories. First, we recorded the activation patterns of each unit within AlexNet to each image of the validation set. While the network is said to contain 5 convolutional layers and 3 fully-connected layers, additional intermediate operations (rectification, pooling,
normalization) result in 22 stages of processing with activation values, where the first stage is defined by the RGB coordinates of the image. Full details on AlexNet can be found in the original paper (Krizhevsky et. al, 2012).

In our initial analyses, we consider four representative layers within AlexNet (Conv1, Conv5, FC6, and FC8) to demonstrate how informational content changes with depth in the network (Figure 1). The activation patterns of Conv1 are those which are input directly to Conv2, thus occurring after rectification, max-pooling, and response normalization. The activation patterns of Conv5 are taken after the fifth convolution and rectification, and are the inputs to FC6. The patterns of fully-connected layers FC6 and FC8 are taken after rectification. For each unit of each layer considered, we compute a set of "categorical signal-to-noise ratios" (cSNRs), for each category of ImageNet. For a given unit and category, the cSNR is computed as the signal-to-noise ratio of the unit's activation across all exemplars of the category. To create populations of units sorted by their cSNR for a given discrimination task on a subset of categories, we first create a vector containing the maximum unit cSNRs across all categories in the subset. This vector is then sorted in three ways, keeping the indices of units available: increasing cSNR, decreasing cSNR, and random. For each layer and discrimination task, three populations of size $n$ are created by selecting the first $n$ units from each of these vectors, for several values of $n$, and bootstrapping is performed across random samples of categories. The activation patterns of each population serve as the set of predictors for the classification of the ImageNet validation set images for each category in the discrimination task. Multi-class classification is achieved with a classification-tree based system, using the fitctree function in the MATLAB Statistics and Machine-Learning toolbox. The classifier is cross-validated using 10 -folds of $80 \%$ training, $20 \%$ testing samples (crossval function). Finally, the loss is computed across the several folds of the cross-validated model (kfoldLoss function). Across bootrsapped camples of category, accuracy is reported as $1-$ mean loss, and error bars are the standard error of accuracy.

Next, AlexNet FC8 is examined in more detail (Figure 2). In analyses similar to those conducted by Haxby et. al (2001), we compare the 2 -way classification of preferred and nonpreferred categories. For a given unit, the preferred category is the category which is explicitly represented; all other categories are potential non-preferred categories. Starting with 20 randomly-drawn ImageNet categories, we generate 100 pairs of preferred/non-preferred categories, and 100 pairs of non-preferred/non-preferred categories, with 5 pairs of each type for every category. For 2-way classification, we use a support-vector machine classifier (fitcsvm function, MATLAB Statistics and Machine-Learning toolbox). The same cross-validation and bootstrapping methods described in preceding analyses are used to generate an estimate of mean and standard error of classification accuracy for each pair type.

We perform similar analyses on VGG-Face, a DCNN trained for face-individuation on over 2000 faces (Parki,

Vedaldi, and Zimmerman, 2015). To achieve a representation to serve as a model of FFA representations, we examine the activations in layer 35 , the final layer before activations are condensed to individual face-specific representations. Layer 35 contains 4096 nodes representing high-level, complex information optimized for face individuation. In performing virtual electrophysiology on VGG-Face, we use as stimuli the fMRI localizer stimulus sets for faces, body-parts, objects, and scenes, in addition to the 2011 ImageNet validation set. These localizer sets are used by TarrLab and many other laboratories in the Center for the Neural Basis of Cognition, Pittsburgh PA, for fMRI research in order to localize functionally-defined regions such as the Fusiform Face Area (FFA), Extrastriate Body Area (EBA), Lateral Occipital area (LO), and Parahippocampal Place Area (PPA). Each localizer set contains 80 images of the category used to localize a corresponding functional brain area. 2-way classification tasks are created using pairs of the categories defining each localizer set. Additionally, a sample of 45 pairs taken from 10 randomly drawn ImageNet categories are used to bootstrap an estimate of the ability to predict all pairs of ImageNet categories. All units in VGG-Face layer 35 are used as predictors in a 2 -way SVM classification system akin to that used in Figure 2 and the results are shown in Figure 3.

## Results

The results of initial system-wide analyses of AlexNet representations are shown in Figure 1. In nearly all cases, decoding accuracy increases with the number of units used as predictors in the classifier, and decreases with the number of categories required for discrimination. In Conv1, Conv5, and FC6, for all discrimination tasks, sorting units by their categorical signal-to-noise ratio (cSNR; see methods) allows for improvements in decoding accuracy given the same number of units. However, in FC8, for all discrimination tasks, sorting units by cSNR has no effect on decoding accuracy, suggesting that localized categorical representations in FC8 possess information relevant to decoding between non-preferred categories.

The results of detailed analyses of AlexNet fully-connected layer 8 (FC8) are shown in Figure 2, where categories are organized by whether they are preferred (explicitly represented) or non-preferred (not represented) by a given unit in FC8. The decoding accuracy between the preferred category and a randomly chosen non-preferred category (see Methods) is not significantly greater than the decoding accuracy between randomly-generated pairs of non-preferred categories ( $\mathrm{p}=0.57$ ); both values are significantly greater than chance ( $\mathrm{p}<0.001$ ).

The results of all analyses of VGG-Face are shown in Figure 3. All discriminations involving face as one of two categories yield perfect discrimination accuracy. 2 out of 4 discriminations involving pairs of non-face categories (scene vs. body part; scene vs. object) yield perfect discrimination accuracy. The remaining 2 discrimination tasks (body part vs. object; pairs of ImageNet categories) yield non-perfect but greater than chance discrimination accuracy ( $\mathrm{p}<0.001$ ).

## Figures



Figure 1: Population decoding from specified layers of the DCNN. Populations are selected in three ordering schemes, choosing the units with the highest cSNR (blue), the lowest cSNR (red), or at random (yellow). A large separation between curves indicates a local code, whereas a small separation indicates a distributed code.


Figure 2: Single-unit decoding of FC8 in AlexNet. Each unit has a "preferred" category - the category it represents. All other categories are non-preferred categories. Bootstrapping was performed as described in Methods. Mean accuracies and standard errors across bootstrapping are shown. Accuracy for preferred/non-preferred is not significantly greater than accuracy for non-preferred only $(p=0.57)$. Both values are significantly greater than chance ( $\mathrm{p}<0.001$ ).


Figure 3: Support-vector machine (SVM) classification of layer 35 activations of VGG-Face network, shown for different pairs of categories. Each SVM is cross-validated with 10 folds of $80 \%$ training $/ 20 \%$ test data, and the accuracy is reported as 1 - mean loss across folds. Error bars are shown as 1 - standard error of loss across folds. For ImageNet, standard error is computed across 45 bootstrapped pairs of 10 randomly chosen categories. All accuracies are significantly greater than chance ( $\mathrm{p}<0.001$ ).

## Discussion

First, using AlexNet, an arbitrary deep convolutional neural network (DCNN) for visual object categorization, we demonstrated that units explicitly representing a single visual category, what we deem localized categorical representations (LCRs), provide information allowing for the decoding of non-preferred categories, at a level equal to that for decoding between the represented category and a non-preferred category. Demonstrating that these localized categorical representations contain distributed information related to non-preferred categories suggests that the presence of domain-general decodable information in a putative domain-specific cortical region is not grounds to reject domain-specificity, as was done by Haxby et. al (2001), and Hanson \& Schmidt (2011), in favor of distributed objectform topographical representations. While LCRs differ from distributed object-form topographical representations in that they represent abstract category information in a single value, the LCRs in AlexNet receive their input from a distributed object-form representation of the sort proposed by Haxby et. al (2001), and are by no means well-described as the sort of functional modules rejected by this study and proposed by the likes of Kanwisher et. al, (1997) for visual face processing, Epstein \& Kanwisher (1998) for visual place/scene processing, or Downing et. al (2001), for visual body-part processing, whereby functional modules likely contain several processing stages for fine-grained analysis of exemplars of the preferred category. To ask whether such functional modules might also give rise to decodable information about stimuli outside the domain of modularity, we relied on a second DCNN specialized for face individuation, VGG-Face.

To acquire a representation of maximal similarity to the high-level face-optimized representations thought to be housed in the Fusiform Face Area, we took the layer 35 activations of VGG-Face, a set of 4096 nodes which project to the individual face probability nodes one layer later. We model these layer 35 nodes as a face module akin to the domain-specific interpretation of the Fusiform Face Area (e.g. Kanwisher et. al, 1997; Kanwisher et. al, 2006). We demonstrated that this "face module" contains patterns of activation capable of perfect discrimination between pairs of categories containing a face, and two of four pairs of categories not containing a face; the other two pairs yielded high discrimination significantly above chance. Thus, we find that domain-specific, face-optimized representations yield domain-general decodable information. This result provides support for the idea that activations within a cortical "module" might contain information relevant to decoding between categories for which that module is not specialized to process. This result strengthens our earlier result, demonstrating that it is improper to reject the possibility of a functional module associated with a given brain region on the grounds that the region's activation patterns allow for decoding between stimuli unrelated to the module's proposed primary function.

An important conceptual point is that the interpretation of our results - that modules should not be rejected on the grounds of producing domain-external information - rests on the assumption that a cortical module would be activated by domain-external information. In the case of localized categorical representations for object categories, it seems likely that all categories would be processed and that some activation might reach LCRs not representing the category of viewing. However, in the case of a cortical module for face processing, this point is less clear. Evidence of subcortical face detection mechanisms (for review, see Johnson, 2005) suggest that the brain may be capable of filtering out non-face information from higher processing (i.e., in FFA), via a fast detection process. Though, as we sometimes perceive faces on trees and in other places in which there are not faces, it is likely that non-face information does, on occasion, pass through the face-detector for further processing. It is possible that all information that arises for domain-external stimuli in FFA, for example, comes from images or image parts which contain something that looks enough like a face to pass through an early detection process, into higher regions of the face processing network. Once the visual information is allowed to pass, our results demonstrate that its processing within a face-optimized processor should give rise to decodable information.

Some authors have argued that the Fusiform Face Area (FFA) is better described as a mechanism for expert-level, fine-grained visual discriminations rather than a faceprocessing module, suggesting that neural substrate within FFA is specialized for visual categorization requiring repeated subordinate-level identification, a task which happens to occur most frequently in the context of face processing, thus resulting in the large preference for faces (e.g. Gauthier et. al, 1999; Tarr \& Gauthier, 2000). Indeed, our results add an interesting point to this theoretical framework. Regardless of whether FFA is specialized for faces or expertise, if it develops representations useful for discriminating between individual faces, these representations are also likely to be useful for discriminating other visual objects. Thus, learning a new category of expertise (e.g. birds) might recruit a previously face-specific cortical region, on the basis of that region containing the most useful representations for the expert task, especially if exemplars of these categories have sufficient visual similarity to faces to pass through an early face-detection gate (if one exists). In this sense, FFA would not be a face module, but rather a brain area optimized most strongly for facerecognition, but also recruited for expert subordinate-level visual recognition.

Whether cortical modules exist in the sense motivated most strongly by Kanwisher (2010) remains an open debate. However, should such cortical modules exist, if their representations are activated by non-preferred categories, these "modules" are likely to produce activation patterns which allow for decoding between non-preferred categories, the characteristic result of studies which sometimes claim evidence of distributed, non-modular processing. As such, it
behooves the field to develop more sensitive and diagnostic measures to assess these critical questions regarding the fundamental nature of representation in the brain.

## Acknowledgments

NB is grateful to his family, and especially to his parents Lisa and James, for their endless support; to the summer Undergraduate Program in Neural Computation at CNBC, Pittsburgh, where the majority of this research was performed; to Ying Yang for mentorship on machine learning techniques, and for helpful discussions of early pieces of this work throughout the summer; and finally to Rosie Cowell for introducing him to the modular debate, which spawned the theoretical interpretation of earlier results, as well as for helpful comments on an early version of this manuscript.

## References

Cowell, R. A., \& Cottrell, G. W. (2013). What Evidence Supports Special Processing for Faces? A Cautionary Tale for fMRI Interpretation. Journal of Cognitive Neuroscience, 25(11), 1777-1793. https://doi.org/10.1162/jocn
Downing, P. E., Jiang, Y., Shuman, M., \& Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. Science, 293(September), 2470-2473.
Epstein, R., \& Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598-601. https://doi.org/10.1038/33402
Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., \& Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expert ise in recognizing novel objects. Nature Neuroscience, 2(6), 568-73. Retrieved from
http://www.biac.duke.edu/education/courses/spring03/cog dev/readings/I. Gauthier et al (1999).pdf
Güçlü, U., \& van Gerven, M. a. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. Journal of Neuroscience, 35(27), 10005-10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015
Hanson, S. J., \& Schmidt, A. (2011). High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. NeuroImage, 54(2), 1715-1734. https://doi.org/10.1016/j.neuroimage.2010.08.028
Haxby, J. V, Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., \& Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(September), 2425-2430. https://doi.org/10.1126/science. 1063736
Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. Proceedings of the National Academy of Sciences, 107(25),

11163-70.
https://doi.org/10.1073/pnas. 1005062107
Kanwisher, N., McDermott, J., \& Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex
specialized for face perception. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 17(11), 4302-11. https://doi.org/10.1098/Rstb.2006.1934
Krizhevsky, A., Sutskever, I., \& Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems, 1-9. https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09. 007
LeCun, Y., Bottou, L., Bengio, Y., \& Haffner, P. (1998). Gradient Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11), 2278-2324. https://doi.org/10.1109/5.726791
Spiridon, M., \& Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron, 35(6), 1157-1165. https://doi.org/10.1016/S0896-6273(02)00877-2
Tarr, M. J., \& Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Nature Neuroscience, 3(8), 764-769. https://doi.org/10.1038/77666
Vedaldi, A., \& Lenc, K. (2015). MatConvNet. Proceedings of the 23rd ACM International Conference on Multimedia - MM '15, 689-692. https://doi.org/10.1145/2733373.2807412
Yamins, D. L. K., \& DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience, 19(3), 356-365. https://doi.org/10.1038/nn. 4244
Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., \& DiCarlo, J. J. (2014). Performanceoptimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences of the United States of America, 111(23), 861924. https://doi.org/10.1073/pnas. 1403112111

Parkhi, O. M., Vedaldi A.and Zisserman, A. (2015). Deep face recognition, Proceedings of the British Machine Vision Conference (BMVC).

# Inferential Role Semantics for Natural Language 

Peter Blouw (pblouw@uwaterloo.ca) Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Centre for Theoretical Neuroscience, University of Waterloo<br>Waterloo, ON, Canada N2L 3G1


#### Abstract

Cognitive models have long been used to study linguistic phenomena spanning the domains of phonology, syntax, and semantics. Of these domains, semantics is somewhat unique in that there is little clarity concerning what a model needs to be able to do in order to provide an account of how the meanings of complex linguistic expressions, such as sentences, are understood. To help address this problem, we introduce a treestructured neural model that is trained to generate further sentences that follow from an input sentence. These further sentences chart out the "inferential role" of the input sentence, which we argue constitutes an important part of its meaning. The model is trained using the Stanford Natural Language Inference (SNLI) dataset, and to evaluate its performance, we report entailment prediction accuracies on a set of test sentences not present in the training data. We also report the results of a simple study that compares human plausibility ratings for both ground-truth and model-generated entailments for a random selection of sentences in this test set. Finally, we examine a number of qualitative features of the model's ability to generalize. Taken together, these analyses indicate that our model is able to accurately account for important inferential relationships amongst linguistic expressions.


Keywords: natural language inference; recursive neural networks; language comprehension; semantics

## Introduction

By most accounts, linguistic comprehension is the result of cognitive processes that map between sounds and mental representations of meaning (Christiansen \& Chater, 2016; Pickering \& Garrod, 2013; Smolensky \& Legendre, 2006). An obvious challenge for these accounts is to provide a good theoretical characterization of the relevant representations. Numerous proposals can be found in the literature, but there is no obvious consensus regarding their relative merits.

Arguably, the reason for this lack of consensus is that linguistic comprehension is itself a somewhat vague and illdefined phenomenon. In the context of efforts to model linguistic comprehension, for instance, it is not entirely obvious what a model needs to be able to do in order to provide an account of how people understand complex linguistic expressions such as phrases and sentences.

In this paper, we argue that one thing models of linguistic comprehension need to be able to do is generate predictions about what follows from a given sentence during a conversation. For example, to understand the statement "The dancers parade down the street", one must be able recognize that the dancers are outside, that they are not standing still, that there is likely a surrounding audience, along with various other things. Comprehending a sentence therefore involves drawing inferences that identify the expected consequences of the occurrence of the sentence in the linguistic environment. And since comprehending a sentence involves


Figure 1: Sentence encoding with a dependency tree recursive neural network (DT-RNN). A dependency parser is used to produce the computational graph for a neural network, which is then used to produce a distributed representation of sentence by merging distributed representations of individual words. Figure adapted from Socher et al. (2014).
comprehending its meaning, it follows that meaning of an expression is at least partly determined by the inferences it licenses (Brandom, 1994)

To motivate this inferential approach to semantics, we introduce a neural network model that learns to generate sentences that are the inferential consequences of its inputs. The model functions by first encoding a sentence into a distributed representation, and then decoding this representation to produce a new sentence. The encoding procedure involves dynamically generating a tree-structured network layout of the sort depicted in Figure 1. Once a sentence encoding is produced using this network, it is fed through an "inverse" treestructured network to produce a predicted sentence. Interestingly, different inverse or decoding networks can be used to generate different sentences from a single encoding. To train the model parameters (i.e. the network weights shared across different tree structures) we use the Stanford Natural Language Inference dataset (Bowman et al., 2015).

In what follows, we first describe the model and then empirically evaluate its ability to produce plausible entailments for sentences unseen in the training data. We present experimentally produced plausibility ratings for a random collection of generated sentences, and from these ratings conclude that the model captures something important about the inferential roles of ordinary linguistic expressions. We further contend that the model motivates the view that understanding a linguistic expression is not (as is typically thought) a matter of mapping it onto a representation that somehow constitutes its meaning. Rather, understanding a linguistic expression is a
matter of inferring the expected consequences of its occurrence in the linguistic environment. The reason for drawing this conclusion is that the expected consequences of a sentence cannot be "read off" of any single representation in the model. Instead, these consequences are derived from the global behavior of the model and the processes that it implements.

## Tree-Structured Neural Networks

To build our model, we take advantage of recently developed techniques for using neural networks to define composition functions that merge distributed representations of words into distributed representations of phrases and sentences (Socher et al., 2012, 2014). The core idea behind these techniques is to produce a parse tree for a sentence, and then transform the tree into a neural network by replacing its edges with weights and its nodes with layers of artificial neurons. Activation is then propagated up the tree by providing input to layers that correspond to certain nodes, as shown in Figure 1. The input at each node is typically a distributed representation or "embedding" corresponding to a single word (see Mikolov et al., 2013).

This general method can be applied using arbitrary tree structures, and we adopt a dependency-based syntax in the experiments described below. There are three reasons for this choice (Socher et al., 2014). First, the assignment of different network weights to different dependency relations allows for the creation of networks that are more sensitive to syntactic information. Second, the semantic role of an individual word can often be read off of the dependency relation it bears to a head word, which allows for the creation of networks that are also sensitive to semantic information. Finally, dependency trees are less sensitive to arbitrary differences in word order, which helps to ensure that simple variations of a sentence get mapped to similar distributed representations. The model we adapt - the dependency tree recursive neural network (DTRNN) - is introduced in Socher et al. (2014)

Some formal details concerning the behavior of DT-RNNs are helpful at this point. First, an input sentence $s$ is converted into a list of pairs, such that $s=\left[\left(w_{1}, x_{1}\right),\left(w_{2}, x_{2}\right), \ldots\left(w_{n}, x_{n}\right)\right]$, where $w$ is a word and $x$ is the corresponding word embedding. Next, a dependency parser is used to produce a tree that orders the words in the sentence in terms of parent-child relations. Each node in this tree is then assigned an embedding in a two-step manner. First, all of the leaf nodes in the tree (i.e. nodes that do not depend on other nodes) are assigned embeddings by applying a simple transformation to their underlying word embeddings:

$$
\begin{equation*}
h_{i}=f\left(W_{v} x_{i}+b\right) \tag{1}
\end{equation*}
$$

where $h_{i}$ is the embedding for some leaf node $i$ in the tree, $x_{i}$ is the embedding for the word corresponding to this node, $W_{v}$ is a matrix that transforms word representations, $b$ is a bias term, and $f$ is an element-wise nonlinearity. Second, embed-
dings are recursively assigned to all of the non-leaf nodes by composing the embeddings of their children as follows:

$$
\begin{equation*}
h_{i}=f\left(W_{v} x_{i}+\sum_{j \in C(i)} W_{R(i, j)} \cdot h_{j}+b\right) \tag{2}
\end{equation*}
$$

where $h_{i}$ is again the embedding for some node $i$ in the tree, $x_{i}$ is the embedding for the word corresponding to this node, $j$ is an index that ranges over the children, $C(i)$, of the node $i$, and $W_{R(i, j)}$ is a matrix associated with the specific dependency relation between node $i$ and its $j^{t h}$ child. $h_{j}$ is the embedding corresponding to this child. So, in the example tree in Figure 1, the embeddings for nodes 1,4 , and 6 would be computed first, since these nodes have no children. Then, embeddings will be computed for any nodes whose children now all have assigned embeddings (in this case, nodes 2 and 7). And so on, until an embedding is computed for every node.

Model training is done via backpropogation and requires that a cost function be defined for the sentence embeddings produced at the root of each tree. The free parameters are the weights $W_{v}$ and $W_{r \in R}$, along with the bias term $b$. Word embeddings can also be fine-tuned over the course of training.

## Generating Entailments

Choosing an appropriate cost function for a recursive neural network can be difficult, since it is not always clear what makes for a "good" sentence embedding. It is accordingly common to see these networks applied to narrow classification tasks such as the prediction of sentiment ratings (e.g. Socher et al., 2012). Our goal is define an optimization objective that accounts for the principle that understanding a linguistic expression involves drawing inferences about what follows from it.

To accomplish this goal, we define a model composed of two DT-RNNs, one that encodes an input sentence into a distributed representation, and another that decodes this representation into a new sentence that is entailed by the input sentence. This model is inspired by Iyyer et al.'s (2014) work using DT-RNNs analogously to autoencoders, but introduces a decoding procedure that computes an appropriate response to the input sentence, rather than merely reconstructing it. Other related work is described in (Kolesnyk et al., 2016).

The model is trained on pairs of sentences standing in entailment relations. A dependency parser ${ }^{1}$ is again used to produce a tree-structured network for each sentence, but the network associated with the second sentence is run in reverse, as shown in Figure 2. A word prediction is generated at each node in this second tree using a softmax classifier, which allows us to define a cross-entropy loss function over nodes and trees as follows:

$$
\begin{equation*}
J(\theta)=-\sum_{i} \sum_{j} t_{j}^{(i)} \log p\left(c_{j}^{(i)} \mid s_{i}\right) \tag{3}
\end{equation*}
$$

[^29]

Figure 2: Generating entailments with paired encoder and decoder DT-RNNs. The decoder network computes a probability distribution over words at each node, conditioned on the sentence representation produced by the encoder. The parameters of both the encoder and decoder are trained via backpropogation through structure using error derivatives supplied at each node in the decoding tree. The encoder and decoder trees are dynamically generated for each pair of sentences in the training data.
where $t_{j}^{(i)}$ is the target probability (i.e. 1) for the correct word at the $j^{\text {th }}$ node in the $i^{\text {th }}$ training example, $p\left(c_{j}^{(i)} \mid s_{i}\right)$ is the computed probability for this word given the input sentence $s_{i}$, and $\theta$ is the set of combined parameters for the encoder and decoder DT-RNNs.

We train the model via stochastic gradient descent by backpropogating through both the decoder and encoder tree for each training example. The result of training is a set of weights associated with dependencies for both encoding and decoding, a set of weights for predicting a distribution over words from a node embedding for each dependency, a set of biases (we allow dependency-specific biases), and the input transformation matrix $W_{v}$. When the trained model is used to perform inference using a novel input sentence, the encoder DT-RNN is assembled into a tree using the learned encoding weights. The decoder DT-RNN is then also assembled into a tree using the learned decoding weights, and activation is propagated through the encoder and into the decoder to produce a probability distribution over words at each tree node. The words with the highest probability at each node are then used to construct the predicted entailment for the input sentence. The tree structure for the decoder can either be selected randomly or stipulated ahead of time.

## Experiments

In the remainder of the paper, we describe a number of basic experiments that illustrate how this general modeling framework can be used to illuminate the phenomenon of language comprehension. We first perform a basic evaluation of how well the decoder model is able to generate entailments by measuring the percentage of correct word predictions over all decoding tree nodes in both the training set and an unseen test set. We then present the results of an experiment designed to evaluate the quality of the entailments generated by our model. Next, following Kolesnyk et al. (2016), we iterate the encoding-decoding procedure to generate chains of
entailments from a given input sentence that delineate simple inferential roles. Finally, we analyze the effect of substituting individual words in an input sentence. The goal of this analysis is to evaluate the extent to which the model is able to learn indirect inferential roles for words and appropriately generalize to a wide range of novel sentences that can be substitutionally derived from a single familiar sentence.

## Training Data

To train encoder and decoder networks, we use a subset of the Stanford Natural Language Inference dataset introduced in Bowman et al. (2015). This dataset consists of approximately 570,000 sentence pairs with labeled inferential relationships. Specifically, the first sentence in each pair can either entail, contradict, or be neutral with respect to the second sentence, and since our interest is generating entailments, we restrict our attention to pairs labeled with the entailment relation.

To reduce the amount of noise and complexity in the dataset, we also perform some simple pre-processing steps. First, we screen for misspelled words, ${ }^{2}$ and eliminate all sentence pairs containing a misspelling. Second, we eliminate all sentence pairs containing a sentence longer than 15 words in order to avoid fitting model parameters to a small number of very long sentences that produce highly complex dependency trees. After preprocessing, the data consists of 106,288-pair training set, a 1701-pair development set, and 1666 pair test set. We train on the training set and use the development set for tuning hyperparameters such as the learning rate and the number of training epochs. The vocabulary used during training and testing consists of 22,555 words.

## Quantitative Evaluations

To evaluate the ability of the model to generate plausible entailments, we first measure the proportion of correct wordlevel predictions during decoding in both the training set and

[^30]Table 1: Examples of Entailments Generated From Novel Test Sentences.

Sentence A boy and girl child swing together on a swing set. Entailment Two kids swing on a swing.

A young blond boy is eating cake with a spoon.
A boy is eating a cake.
A surfer is performing a jump stunt in the ocean. A surfer and a surfboard is outside.

Table 2: Word-Level Accuracy for Entailment Generation

| Model | Training Set (\%) | Test Set (\%) |
| :---: | :---: | :---: |
| Chance | 6.0 | 5.9 |
| DT-RNN | 66.7 | 61.8 |

the test set. We provide the tree structure of each entailed sentence during decoding, so inference involves propagating activities through paired trees of the sort depicted in Figure 2 to generate a set of word predictions. Some example entailments produced from sentences drawn from the test set are listed in Table 1. The decoding tree used to produce each entailment is chosen randomly in these examples.

The word vectors that provide input to the encoder are initialized using 300-dimensional Word2Vec embeddings (Mikolov et al., 2013), while biases are initialized as the zero vector. Each set of weights associated with a syntactic dependency is initialized as a $300 \times 300$ identity matrix with mean-zero Gaussian noise for both the encoder and decoder. The word transformation matrix, $W_{v}$, is initialized in the same way. During learning, all of these matrices are updated using stochastic gradient descent, along with the Word2Vec embeddings and the biases. We perform approximately 5 epochs of training using an initial learning rate of $6 \times 10^{-4}$, and we progressively anneal this rate over the course of training.

To collect accuracy measures, we simply tally the proportion of nodes in the decoding trees for which the predicted word is the same as the actual word given in the relevant test set; the decoding tree is determined by a parse of the correct entailment in every case. We compare against a baseline accuracy of chance. As shown in Table 2, The DT-RNN model performs considerably better. It is worth noting that generated sentences containing words not present in the correct entailment may still be appropriate, given that no entailment is uniquely correct. It is also worth noting that prior work involving SNLI has almost uniformly focused on the problem of classifying sentence pairs. Given that our interest is in generation rather than classification, we cannot easily draw comparisons to earlier work, and therefore use novel methods of evaluation.

## Empirical Evaluations

Next, we conduct a simple study in which human subjects are asked to evaluate the plausibility of model-generated sentences. During the study, participants are shown a series of

Table 3: Plausibility Ratings for Inferential Relations.

| Source | Status | Mean Likert Rating (1-5) |
| :---: | :---: | :---: |
| Human | Entailment | $4.05 \pm 0.09$ |
| Model | Entailment | $3.53 \pm 0.12$ |
| Human | Contradiction | $2.05 \pm 0.12$ |

* Margins are bootstrapped $95 \%$ confidence intervals.
sentences introduced as true captions for unseen images. ${ }^{3}$ For each caption, the participants are shown an alternate caption and asked to evaluate the likelihood that it is also true of the corresponding image. Evaluations are recorded using a five point Likert scale that ranges from "Extremely Unlikely" (1) to "Extremely Likely" (5). The original caption in each case is the first sentence in a pair randomly chosen from the SNLI test set, while the alternate captions are either (a) model-generated entailments, (b) human generated entailments drawn from the test set, or (c) human generated contradictions also drawn from the test set. This betweensubjects experimental design is similar to the method used by Bowman et al. (2015) to validate human-generated sentence pairs during the creation of SNLI. The main difference is that we evaluate model-generated sentences in addition to humangenerated sentences.

Seventy-five participants from the United States were recruited through Amazon's Mechanical Turk and split evenly into the three conditions. The main captions were identical across conditions, and each participant was asked to rate 20 caption pairs. ${ }^{4}$ Participants were paid $\$ 1.00$ for their time. Two of the seventy-five participants failed to complete the study and did not have their responses included in the results. Repeat participation was blocked by screening Mechanical Turk worker IDs.

The Likert ratings collected during the study are assessments of the plausibility of the inferential transition from one sentence (the main caption) to another (the alternate caption).

[^31]

Figure 3: A model-generated inferential network around the sentence "A man is outside". Each inferential transition is the result of generating a predicted entailment after encoding the sentence at the beginning of each arrow. The entire network is generated starting with only the four outermost sentences, which are drawn from the SNLI test set.

The transitions involving sentence pairs drawn directly from SNLI offer a kind of gold standard for both good and bad transitions. The results shown in Table 2 indicate that modelgenerated transitions are seen to be almost as plausible as the gold-standard transitions drawn from SNLI. We take this to be preliminary evidence that model is able to capture certain tacit inferential relationships between natural language expressions.

## Qualitative Extensions

In order to further analyze the model's behavior, we examine a number of qualitative features of the inferential relationships it is able to learn. First, we examine iterative applications of the model to its own predictions, following similar work by Kolesnyk, Rocktäschel, and Reidel (2016) that makes use of a sequential LSTM. Next, we examine the model's ability to disciminate the inferential significance of lexical items by performing simple word-by-word substitutions in an input sentence. The point of these analyses is to demonstrate that models of the general class we are proposing are useful tools for both formalizing and learning the inferential roles of a wide variety of linguistic expressions.

## Iterative Inferences

Once an input sentence has been passed through the model to generate an entailment, it is possible to use this entailment as a new input to the model. Repeated applications of the model accordingly make it possible to chart out an "inferential network" around a particular starting sentence. Figure 3 offers a simple model-generated example of an inferential network in which numerous sentences describing men doing things outdoors are eventually mapped to the sentence "A man is outside".

In general, predicted entailments that are shorter than an input sentence tend to be more abstract and general, while predicted entailments that are longer than an input sentence tend to introduce plausible elaborations (Kolesnyk et al., 2016). For instance, the sentence "A bird is in a pond." can be used to generate the sentence "A little bird is outside in a small pond." by using a decoding tree with nodes for two additional adjectives and an additional adverb.

## Substitutional Analysis

If individual words in an input sentence are replaced, it becomes possible to identify the impact of particular words on
the inferences that are licensed by a particular sentence. In Figure 4, for instance, the replacement of a subject noun or the main verb can be seen to have significant effects on the kinds of entailments that are generated. The model is impressively sensitive to sophisticated linguistic cues concerning agreement. For instance, the model correctly infers that "boy" should be paired with the male possessive "his", while "girl" should be paired with the female possessive "her". It is worth emphasizing that all of the sentences that result from substitution are completely novel from the model's perspective. The fact that the model is able to generate reasonable entailments for many of these sentences suggests that it is able generalize beyond the training data quite successfully.

A further application of substitutional analysis involves examining a model's ability to learn about theoretically interesting constructions involving negations, quantifiers, and numerals. For instance, the model exhibits a rudimentary ability to handle numerals appropriately, as is shown by the inference from "A boy and a girl..." to "Two kids..." Negations are a bit more troublesome: the model correctly infers "not outside" from "in a car", but incorrectly infers "not indoors" from "in a store". Quantifiers, finally, are an open question: the model correctly infers "The women" from "Many women", but it is not clear that this is the result of learning a relation between "Many" and the plural forms of nouns. Examining specific linguistic constructions in this substitutional manner is a promising avenue for future research.

## Discussion

Overall, the point of this work is to motivate an approach to semantics based on inferential relationships amongst linguistic expressions (Brandom, 1994). Our use of the encoderdecoder DT-RNN model is designed to illustrate how generalized inferential roles can be learned for arbitrary linguistic expressions from examples of how sentences are distributed as tacit "premises" and "conclusions" in a space of inferences. It is accordingly possible to characterize this work as an extension to the well-known distributional approach to semantics (Turney \& Pantel, 2010), wherein we replace the generic notion of a linguistic context with the more fine-grained notion of an inferential context.

As with most natural language generation systems, many of the sentences produced by our model are defective in some way. As can be seen in the examples in Table 1, our generated entailments are almost always thematically appropriate,


Figure 4: Substitutional analysis using the sentence "A boy in a beige shirt is sleeping in a car". The model is able to predict appropriate entailments for a range of sentences that are similar to this initial sentence shown at the top left. The fact that these substitutionally-derived sentences are not present in SNLI dataset indicates that our model is able to generalize by interpolating between the example inferential transitions found in the training data.
but sometimes contain agreement errors or misplaced words that render the entailment as a whole ill-formed. And, not infrequently, the model produces entailments that are more or less incomprehensible. There are two ways to address these problems. The first involves the use of increased amounts of training data to provide the model with a more points in the "space of inferences" to interpolate between. The second involves the use of more sophisticated network architectures that help the model to learn to more selectively make use of only the input information that is most relevant to generating a good entailment. LSTM network architectures, such as the Tree LSTM (Tai et al., 2015), are likely to provide improvements on this second front.

Finally, an important limitation of our work is that we do not consider the relationship between linguistic expressions and the non-linguistic world. A natural way to account for this relationship is to suppose that a sentence's occurrence in the linguistic environment licenses certain expectations about what can be seen, heard, or otherwise perceived. To return to our initial example, if one understands the statement "The dancers parade down the street", one will expect to see and hear dancers upon going to the relevant street. We accordingly suggest that if an individual can adequately infer all that follows from a given linguistic expression, both linguistically and non-linguistically, then there is nothing further they need to be able to do to count as understanding what the expression means. The main consequence of this view is that inference should be at the core of any theory of semantic cognition.

## Acknowledgments

This work was supported by an Ontario Graduate Scholarship and the Canada Research Chairs program.

## Code

All of the simulations described in this paper were implemented using a neural network library written by the first author, available online at https://github.com/pblouw/pysem.

## References

Bowman, S., Angeli, G., Potts, c., \& Manning, C. (2015). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics.

Brandom, R. (1994). Making it explicit: Reasoning, representing, and discursive commitment. Harvard University Press.
Christiansen, M., \& Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 1-72.
Iyyer, M., Boyd-Graber, J., \& Daume III, H. (2014). Generating sentences from semantic vector space representations. In Nips workshop on learning semantics.
Kolesnyk, V., Rocktäschel, T., \& Reidel, S. (2016). Generating natural language inference chains. arXiv preprint arXiv:1606.01404..
Mikolov, T., Sustkever, I., Chen, K., Corrado, G., \& Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems 28.
Pickering, M., \& Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and Brain Sciences, 36, 329-392.
Smolensky, P., \& Legendre, G. (2006). The harmonic mind: From neural computation to optimality-theoretic grammar (Vol. 1). MIT Press.
Socher, R., Huval, B., Manning, C., \& Ng, A. (2012). Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (p. 1201-1211). Association for Computational Linguistics.
Socher, R., Karpathy, A., Le, Q., Manning, C., \& Ng, A. (2014). Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association of Computational Linguistics, 2, 207-218.
Tai, K., Socher, R., \& Manning, C. (2015). Improved semantic representations from tree-structured long shortterm memory networks. In Proceedings of the 53rd annual meeting of the association for computational linguistics (p. 1556-1566). Association for Computational Linguistics.
Turney, P., \& Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, 141-188.

# Split-Second Detection of Cooperativeness from Faces in the Trust Game 

Jean-Francois Bonnefon<br>University of Toulouse, Toulouse, France<br>Astrid Hopfensitz<br>University of Toulouse, Toulouse, France<br>Wim De Neys<br>Université Paris Descartes, Paris, France


#### Abstract

Economic interactions often imply to gauge the trustworthiness of others. Recent studies showed that when making trust decisions in economic games, people have some accuracy in detecting trustworthiness from the facial features of unknown partners. Here we provide evidence that this face-based trustworthiness detection is a fast and intuitive process by testing its performance at split-second levels of exposure. Participants played a Trust game, in which they made decisions whether to trust another player based on their picture. In two studies, we manipulated the exposure time of the picture. We observed that trustworthiness detection remained better than chance for exposure times as short as 100 ms , although it disappeared with an exposure time of 33 ms . We discuss implications for ongoing debates on the use of facial inferences for social and economic decisions.


# Diversity in a Contrast Set Increases Generalization from a Single-Item Target 

David Bosch<br>New York University<br>David Kalkstein<br>New York University<br>Tali Kleiman<br>The Hebrew University of Jerusalem


#### Abstract

Four experiments explored the effect of diversity of contrasting negative evidence on inductive inferences drawn from a single-item target. In Experiments 1 and 2, we found that increasing the diversity of a contrast set led people to infer that a target exemplar corresponded to a higher level category and led to greater generalization of a novel property associated with the target. Further, we demonstrated two boundary conditions in which the effect only occurred when the contrast set was consistent with a higher level category that both united the contrast exemplars and distinguished them from the target (Experiment 4) and when contrast and target shared an obvious parent category (Experiment 5). Taken together, these findings demonstrate that increasing the diversity of a contrast increases generalization from a target, but only if the contrast set is drawn from a single category that excludes, but shares a common parent with, the target.


# Causal learning from interventions and dynamics in continuous time 

Neil R. Bramley ${ }^{1}$ (neil.bramley@nyu.edu), Ralf Mayrhofer ${ }^{2}$ (rmayrho@gwdg.de)<br>Tobias Gerstenberg ${ }^{3}$ (tger@mit.edu), David A. Lagnado ${ }^{4}$ (d.lagnado@ucl.ac.uk)<br>${ }^{1}$ Department of Psychology, NYU, New York, NY, 10003, USA<br>${ }^{2}$ Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073, Germany<br>${ }^{3}$ Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA<br>${ }^{4}$ Department of Experimental Psychology, UCL, London, WC1H 0DS, UK


#### Abstract

Event timing and interventions are important and intertwined cues to causal structure, yet they have typically been studied separately. We bring them together for the first time in an experiment where participants learn causal structure by performing interventions in continuous time. We contrast learning in acyclic and cyclic devices, with reliable and unreliable causeeffect delays. We show that successful learners use interventions to structure and simplify their interactions with the devices and that we can capture judgment patterns with heuristics based on online construction and testing of a single structural hypothesis.


Keywords: causal learning; intervention; time; causal cycles; structure induction; dynamics.

In a dynamically unfolding world, using actions to uncover causal relationships requires good timing. It is hard to tell whether a new medication is effective if you take it with others, or just as you start to feel better. Likewise, it is hard to tell whether a new law lowers crime if it is introduced just after other reforms or before a major election. Such inferences, having to do with delayed effects and an evolving causal background, can be particularly tough in cyclic systems in which feedback loops make prediction difficult even with complete knowledge (Brehmer, 1992). Thus, for interventions to be effective tools for unearthing causal structure it is important to time and locate them carefully, paying close attention to the temporal dynamics of surrounding events and the possibility of feedback loops.

Previous work has shown that people make systematic use of temporal information, taking event order as a strong cue to causal order (Bramley, Gerstenberg, \& Lagnado, 2014), and making stronger attributions when putative cause-effect delays are in line with expectations (Buehner \& McGregor, 2006) and have low variance across instances (Greville \& Buehner, 2010). Recent work has also developed frameworks for probabilistic causal inference from event timings based on parametric assumptions about cause-effect delays (Bramley, Gerstenberg, Mayrhofer, \& Lagnado, submitted; Pacer \& Griffiths, 2015).

A distinct line of work has shown that people are adept at inferring causal structure from interventions - idealized actions that set variables in a system (e.g., Bramley, Dayan, Griffiths, \& Lagnado, 2017; Coenen, Rehder, \& Gureckis, 2015). This work has not explored the role of temporal information however. While researchers have speculated about the close relationship between temporal and interventional inference (e.g., Lagnado \& Sloman, 2004), our paper is the first to explore interventional causal learning in continuous time.

## The learning problem

We explore the general problem of how people learn about a causal system by interacting with it in continuous time. We focus on abstract causal "devices" made up of 3-4 components (cf. Figure 1). For causally related components, we assume each activation of a cause will tend to bring about a single subsequent activation of its effect after a parametric delay (described below). For example, Figure 1a shows a learner's interactions with a $B \leftarrow A \rightarrow C$ Fork during which time they perform four interventions. Activations of both $B$ and $C$ succeed the interventions on $A$ but with some variability in delays.

We focus on situations where components never spontaneously activate, but where causal relations work stochastically (e.g., are successful with probability $w_{S}$ ). Any pair of components can be connected in either, neither or both directions resulting in a hypothesis space $S$ of 64 possible structures for devices made up of three components, and 4096 for four components. Learners can intervene on the devices by directly activating any component at any moment of their choosing. Interventions are always successful in that they instantaneously activate the targeted component. The downstream causal effects of intervened-on components are the same as those of components that were activated by other components. Thus, we model the consequences of interventions in analogy to the $\operatorname{Do}($.$) operator introduced by Pearl$ (2000), such that interventions provide no information about the causes of the intervened-on component.

## Choosing interventions

Seeing the effects of one's interventions in continuous time provides rich information for causal inference. On the flip side, there are also no completely independent trials. For instance, in Figure 1a, the early interventions on $C$ and $B$ might, in principle, be responsible for the observed effects that happen shortly after the intervention on $A$. In general, one cannot rule out the possibility something that happened earlier is still exerting its influence, or that an effect is yet to reveal itself. Fortunately, interventions provide anchor points. We know that events due to interventions weren't caused by anything else, and that these events only affect the future but not the past (Lagnado \& Sloman, 2004). This means that by intervening, learners can recreate some of the advantages that come with a discrete trial structure. For example, by waiting long enough between interventions to be confident prior effects have dissipated, an otherwise confusing event stream


Figure 1: Examples of using real-time interventions to infer causal structure. Left: True generative causal model with subplots showing delay distributions. Right: Timelines showing an active learners' interactions with each system with a row for each component $A$ (top), $B$ (middle) and $C$ (bottom), and white circles indicating their activations over 45 seconds (x-axis). " + " symbol and incoming hand icon indicate interventions. Dashed gray lines indicate the actual cause-effect relationships.
becomes more palatable and informative about the underlying structure. Figure 1 b gives an example of interventions that are not well chosen. The learner performs four interventions in the same locations as Figure 1a but does so in close succession. It is hard to attribute causal responsibility for these activations, since there are so many similarly plausible candidates. Consequentially, this data is considerably less informative.

In discrete-trial interventional learning, participants exhibit a positive testing strategy - they prefer to intervene on root variables that bring about many effects (Coenen et al., 2015). While often not leading to the most globally informative choice, a positive testing strategy is an effective way of assessing the adequacy of one's current working hypothesis, making it a manifestation of confirmatory testing (Nickerson, 1998). Many other components will be affected if one's hypothesis is right, and few if it is wrong. Repeated positive testing might be more justifiable in the continuous time context because cause-effect delays may play out differently each time, and potential temporal reversals between variable activations will help to rule out candidate structures (Bramley et al., 2014). For example, in Figure 1a the second intervention on $A$ leads to $B$ and $C$ occurring in reversed order, allowing the learner to rule out a $A \rightarrow B \rightarrow C$ Chain structure.

## Causal cycles

The vast majority of causal learning studies have focused on acyclic causal systems in which causal influences flow only in one direction, never revisiting the same component. However, many natural processes are cyclic and people frequently report cyclic relationships when allowed to do so (e.g.

Sloman, Love, \& Ahn, 1998). While there are ways of adapting the causal Bayes net formalism to capture cycles (Rehder, 2016), these generally simplify the problem to influences between fixed time steps (e.g. Rottman \& Keil, 2012), or just to the long-run equilibrium distribution (e.g. Lauritzen \& Richardson, 2002). However, by focusing on continuous time and developing a representation capable of modeling causal dynamics, we are able to directly compare learning in acyclic and cyclic causal systems.

Dynamic systems can be hard to predict even with perfect knowledge. Positive feedback loops can lead to sensitive dependence on initial conditions with very different behavior resulting from small perturbations in starting conditions (e.g., Gleick, 1997). Figure 1c gives an example of interventions on a cyclic causal system (assuming that the connections work $90 \%$ of the time). Interventions initialize looping behavior because of the bidirectional relationship $A \leftrightarrow B$ (e.g., $A \rightarrow B \rightarrow A \rightarrow B \ldots$ ) leading to many subsequent activations of both the loop components and the output component $C$, continuing until either the $A \rightarrow B$ or $B \rightarrow A$ connection fails. Based on simply looking at the timeline, it seems likely that it will be easier to identify which components are either directly involved in cycles, or outputs from cyclic components (due to their recurrent activations), but harder to identify the exact causal relationships (e.g. whether it is $A$ or $C$ that causes $B$ in this example since both tend to recur shortly before $B$ ).

## Normative inference

As a benchmark, we developed a Bayesian model of causal structure inference. We consider the data $\mathbf{d}_{\tau}\left\{d_{X}^{(1)}, \ldots, d_{X}^{(n)}\right\}$ to be made up of all activations (with events indexed in chronological order and $X$ indicating the activated component) conditioned upon the set of interventions $\mathbf{i}_{\tau}=$ $\left\{i_{X}^{(1)}, \ldots, i_{X}^{(m)}\right\}$. Both $\mathbf{d}_{\tau}$ and $\mathbf{i}_{\tau}$ are restricted to the interval between the beginning of the clip and time $\tau$, which we assume to be the moment at which the learner makes the inference. For instance, one might interact with a causal device for 5000 ms , performing interventions on components $A$ and $B$ at 100 ms and 1200 ms respectively: $\mathbf{i}_{5000}=\left\{i_{A}^{(1)}=100, i_{B}^{(2)}=\right.$ $1200\}$, and observing two activations of $C: \mathbf{d}_{5000}=\left\{d_{C}^{(1)}=\right.$ $\left.1500, d_{C}^{(2)}=2800\right\}$.

Normative Bayesian structure inference involves updating a prior over structure hypotheses $P(S)$ with the likelihood $p\left(\mathbf{d}_{\tau} \mid S ; \mathbf{i}_{\tau}, \mathbf{w}\right)$ to get a posterior belief over structures $P\left(S \mid \mathbf{d}_{\tau} ; \mathbf{i}_{\tau}, \mathbf{w}\right)$ given the set of parameters $\mathbf{w}:{ }^{1}$

$$
\begin{equation*}
P\left(S \mid \mathbf{d}_{\tau} ; \mathbf{i}_{\tau}, \mathbf{w}\right) \propto p\left(\mathbf{d}_{\tau} \mid S ; \mathbf{i}_{\tau}, \mathbf{w}\right) \cdot P(S) \tag{1}
\end{equation*}
$$

An immediate issue with calculating the likelihood of an observed set of activations given a candidate model is that there are likely to be multiple potential paths of actual causation that could have produced the data (Halpern, 2016), each

[^32]

Figure 2: Devices tested and results from experiment in a) reliable and b) unreliable delay conditions. Node shading: Intervention choice prevalence by component. Edge shading: accuracy. Note: Ints = average number of interventions performed; Acc = mean accuracy.
of which implying a different likelihood. For example, if the true structure is a $A \rightarrow C \leftarrow B$ Collider, the data above might be produced in two ways. $A$ could have caused the first activation of $C$ and $B$ the later $\left(i_{A}^{(1)} \rightarrow d_{C}^{(1)}, i_{B}^{(1)} \rightarrow d_{C}^{(2)}\right)$. Alternatively, $A$ could have caused the later activation of $C$ and $B$ the earlier $\left(i_{A}^{(1)} \rightarrow d_{C}^{(2)}, i_{B}^{(1)} \rightarrow d_{C}^{(1)}\right)$.

However, as there can only be one true path of actual causation in the set of possible paths $\mathbf{Z}_{s}$, we can sum over these to get the likelihood of the data given a candidate model $s \in S$ :

$$
\begin{equation*}
p\left(\mathbf{d}_{\tau} \mid s ; \mathbf{i}_{\tau}, \mathbf{w}\right)=\sum_{\mathbf{z}^{\prime} \in \mathbf{Z}_{s}} p\left(\mathbf{d}_{\tau} \mid \mathbf{z}^{\prime} ; \mathbf{i}_{\tau}, \mathbf{w}\right) \tag{2}
\end{equation*}
$$

We assume that the actual causal delays (in $\mathbf{Z}_{s}$ ) are Gamma distributed (see also Bramley et al., submitted) with a known expected duration $\mu$ and shape $\alpha$ (i.e., variability). The likelihood of the data given a specific path $\mathbf{z}^{\prime}$, then, is the product of the (Gamma) likelihoods of the observed delays and causal strength $w_{S}$ combined with the likelihoods of (non-)events, the occurrence of which failed either due to the $1-w_{S}$ causal failure rate or due to the effect potentially occurring after $\tau$ (i.e., some time in the future).

With these ingredients the posterior belief over causal structure hypotheses can be determined. However, it is only feasible to enumerate all possible paths of actual causation for a sufficiently small number of events. While for a large number of events the calculations become intractable, we were able to compute the posteriors in the described manner for the data from the current experiment, resorting only in rare cases to an approximation. ${ }^{2}$

## Experiment

Participants' task was to discover the causal connections between the components of several devices in limited time

[^33](see Figure 2). Half of the devices were acyclic (top; no feedback loops) and half were cyclic (bottom; contained a feedback loop). Participants were able to activate any of the components by clicking on them. We were interested in how participants chose where to intervene and when. We examined two delay conditions between subjects, one in which the true cause-effect delays were reliable (Gamma distributed with $\alpha=200, M \pm S D 1.5 \pm 0.1$ seconds) and one where they were unreliable ( $\alpha=5, M \pm S D 1.5 \pm 0.7$ seconds). Following Greville and Buehner (2010), we expected that performance would be better when causal delays were reliable. We also predicted that complex dynamics would lead to worse performance when the true structure was cyclic, and that successful participants would spread their interventions widely over time, thus minimizing the ambiguity of resulting patterns of effects.

## Methods

Participants Forty participants (14 female, aged $32 \pm 9.0$ ) were recruited from Amazon Mechanical Turk (yielding 20 subjects in each delay-reliability condition) and were paid between $\$ 0.50$ and $\$ 3.20(\$ 2.06 \pm 0.39)$ depending on performance (see Methods section). The task took around 20 minutes.
Materials and procedure Each device was represented with a circle for each component and boxes marking the locations of the potential connections (see Figure 3a). ${ }^{3}$ Trials lasted for 45 seconds during which components activated if clicked on or if caused by the activation of another component, with delay and probability governed by the true underlying network (Figure 3b). Causal relationships worked $90 \%$ of the time (i.e., causal strength $w_{S}=0.9$ ) and there were no spontaneous activations. Activated components turned yellow for 200 ms , and intervened-on components were additionally marked by a "+" symbol. Initially, all components were inactive and no

[^34]

Figure 3: Experimental procedure. a) Up to 6 interventions could be performed by clicking on the components during the 45 second trial. b) This would lead to subsequent activations determined by causal connections and delays in the true model. c) Participants marked their beliefs about the structure during the trials by clicking on the edges. d) At the end of each trial they received feedback. Broad gray arrows: ground truth, Green $=$ correct, Red $=$ incorrect.
connections were marked between them.
Prior to the inference tasks, participants were trained on the delays in their condition and how to register structure judgments through interaction with an an example device. They then had to correctly answer comprehension check questions and complete a practice problem, before facing the 12 test devices in random order with randomly orientated and unlabeled components.

In the test phase, participants could perform up to 6 interventions on each trial and register/update their judgments about the causal structure as often as they liked until the 45 seconds for a device ran out (for details see Figure 3). At the end of each trial, they were given feedback showing the true relationships and which of them they had correctly identified. To incentivize proper judgments, bonuses were paid based on connections participants had registered at a randomly chosen point during each trial.

## Results

We analyze participants' judgments by first comparing their accuracy by delay-reliability condition (between subjects: reliable vs. unreliable) and device type (within subject: acyclic vs. cyclic). We then analyze the timing and spacing of participants' interventions and how these relate to the evidence and judgments.
Accuracy Participants updated and confirmed their judgment about the structure $\mathrm{M} \pm \mathrm{SD} 1.6 \pm 1.2$ times per trial on average. Judgment time was not significantly related to accuracy, but within trials, final judgments were slightly more accurate than initial judgments, with participants correctly identifying $69 \% \pm 30 \%$ (chance performance would be $25 \%$ ) compared to $65 \% \pm 28 \%$ of the connections, $t(479)=$ $5.2, p<.001$ (remember that bonuses incentivised making judgments early). Only $4 \%$ of judgment updates decreased the number of connections, $24 \%$ resulting in the same number as before, and $72 \%$ increasing the number of connections.

Focusing on final judgments, participants correctly identified [reliable,acyclic]: $82 \% \pm 29 \%$, [reliable,cyclic]: $68 \% \pm$ $28 \%$, [unreliable,cyclic]: $69 \% \pm 29 \%$, [unreliable,cyclic]: $56 \% \pm 29 \%$ of the connections. A repeated measures analysis revealed a significant effect of delay-reliability condition, $F(1,38)=4.6, p=.04$, and cyclicity, $F(1,38)=39, p<$ .001, but no interaction, with unreliable delays and cyclic structures associated with lower accuracy. Figure 2 shows that participants found the Cyclic 3, 5 and 6 structures hardest to identify on average, struggling in particular with distin-
guishing looping from output components.
Ideal Bayesian inference based on the evidence generated by participants predicts a different pattern. While reliable delays allow greater accuracy than unreliable ones, $F(1,38)=$ $24.3, p<.001$, there is no predicted difference in accuracy between acyclic and cyclic devices, $F(1,38)=0.43, p=.5$. In fact, posterior uncertainty over all possible models, measured by Shannon entropy, was generally lower for evidence generated by a cyclic $.74 \pm 1.26$ than an acyclic $1.95 \pm 1.29$ devices, $F(1,38)=109, p<.001$.
Timing of interventions We hypothesized that spacing interventions out in time would be important for successful learning. Participants waited $7.3 \pm 2.8$ seconds between interventions on average. In a regression including delay condition and total number of interventions as covariates, leaving longer intervals between interventions was positively associated with accuracy, $F(1,36)=14.0, \beta=0.04, \eta_{p}^{2}=$ $.26, p=.001$, with no interaction with condition. The variability of these gaps - measured by their coefficient of variation $C V=\frac{\sigma}{\mu}$ - was also inversely related to accuracy, $F(1,36)=7.9, \beta=-0.5, \eta_{p}^{2}=.18, p=.008$ and this effect was stronger in the unreliable delay condition, $F(1,35)=4.5, \eta_{p}^{2}=.11, p=.04$. We also assessed the intervals participants left after the most recently preceding event (whether this was an intervention or an effect) before performing their next intervention. Again larger intervals, $F(1,36)=7.7, \beta=0.06, \eta_{p}^{2}=.18, p=.008$, and less variation, $\beta=-.25, F(1,36)=5.0, \eta_{p}^{2}=.12, p=.03$, was associated with accuracy with neither measure interacting with delay condition. Both larger intervals between interventions, and between interventions and the most recently preceding effect were also associated with lower posterior entropy, with $\beta=0.05, F(1,36)=9.9, \eta_{p}^{2}=.22, p=0.003$ and $\beta=0.09, F(1,36)=8.1, \eta_{p}^{2}=.18, p=0.007$, respectively. However, there was no evidence for a relationship between entropy and the variability of either interval type.
Positive testing We found evidence of a preference for positive testing, with participants performing $1.2 \pm 0.5$ times as many interventions per root component than per non-root component $t(59)=3.9, p<.001$. This preference was associated with higher accuracy after accounting for condition, $F(1,37)=21, \eta_{p}^{2}=0.37, p<.001$, and did not interact with condition. Degree of root preference, however, was not significantly related to posterior uncertainty from the perspective of an ideal Bayesian learner.

Adaptation to cycles While participants performed fewer interventions on cyclic $(4.1 \pm 1.1)$ compared to acyclic ( $5.4 \pm$ 0.7 ) devices, $t(39)=8.7, p<.001$ (see Figure 2), they still experienced far more effects in the cyclic systems $(29.3 \pm 10)$ compared to the acyclic ones $(4.7 \pm 1.1), t(39)=15.5, p<$ .001. This was due to the reciprocal relationships sustaining activations until one of the links failed. Thus while there was normatively more evidence available in the cyclic trials - as reflected by the generally lower posterior uncertainty - the large number of events resulted in more ambiguous evidence, with many candidate causes per effect and a large number of potential actual causal pathways.
Summary Participants were better at identifying causal relations from interventions when delays were reliable and the true structure was acyclic. Meanwhile, ideal learner accuracy was affected by reliability by not cyclicity. Successful participants spread their interventions out more in time, waited longer after previous events, distributed them more evenly and favored root components. Participants frequently updated their models by adding additional connections but rarely removed connections.

## Modeling heuristic inferences

Participants' deviations from the prediction of an ideal Bayesian learner suggests that they relied on simpler learning strategies. In this section we compare judgment patterns to several heuristic models inspired by work on order-driven (e.g., Bramley et al., 2014) and incremental causal structure learning (e.g., Bonawitz, Denison, Gopnik, \& Griffiths, 2014; Bramley et al., 2017).

Several papers have proposed that human causal learning is based on the adaptation of a single global hypothesis (Bonawitz et al., 2014), which might be achieved incrementally through making local changes as data is observed (Bramley et al., 2017). This seems particularly applicable in a continuous-time context, where normative inference is tough and the evidence arrives continuously. People may learn locally, ignoring dependence on beliefs about surrounding relationships (e.g. Fernbach \& Sloman, 2009), or use their current model as a basis, comparing observations against predictions, only adding new connections to explain events that cannot easily be accommodated by their existing model (Bramley et al., 2017).

The idea that learners might construct their causal hypotheses incrementally can be combined with different degrees of sensitivity to timing as well as the predictions of their current structure hypothesis. This suggests several potential heuristics that adapt a single model belief $b$ as events are experienced. The result in each case is a single structural belief that evolves as events occur (we write $\mathbf{b}=\left\{b^{(0)}, \ldots, b^{(n)}\right\}$, where the sequence of belief indices correspond to the event indices in $\mathbf{d}_{\tau}$ ):

1. Order Only (OO) Heuristic OO attributes each new effect to the most recently preceding event at any different component (either the most recent intervention in $\mathbf{i}_{\tau}$ or activa-
tion in $\mathbf{d}_{\tau}$ ). If the currently held model hypothesis $b^{(t-1)}$ does not contain a respective edge, $b^{(t-1)}$ is augmented with an edge to make $b^{(t)}$. Figure 4 a gives an example of this. Starting from $b^{(t-1)}$ with a single $D \rightarrow B$ connection, the heuristic connects $A$ to $B$ upon observing $B$ 's activation straight after activating $A$, and then $B$ to $C$ when $C$ activates shortly after.
2. Time Sensitive (TS) TS is like OO but with sensitivity to the expected cause-effect delays. It attributes activations to the (previous) event such that the respective delay would be most likely given the knowledge of the true causal delay distribution, and augments $b^{(t-1)}$ with an edge, if there is none yet, to form $b^{(t)}$. In the example (Figure 4 b ), $C^{\prime}$ s activation time is most consistent with $C$ being caused by the intervention on $A$, thus the model adds an $A \rightarrow C$ connection, rather than a $B \rightarrow C$ connection, going into $b^{(t+1)}$.
3. Structure + Time Sensitive (STS) STS is like TS, but it first checks if there is already an adequate explanation in the current model $b^{(t-1)}$. Concretely, it compares the likelihood of the most likely explanation that is already a cause in $b^{(t-1)}$ to the most likely explanation overall (i.e., the one selected by TS). Where these differ, it only adds an edge if the respective delay is substantially more likely than the delay implied by the best existing explanation in $b^{(t-1)}$, where we assume that "substantially more likely" means a likelihood ratio $>\frac{20}{1}$. Figure 4 c gives an example. Unlike TS, this heuristic does not add an $A \rightarrow C$ connection going into $b^{(t+1)}$ because $C$ 's activation can be explained well enough by the existing connection $D \rightarrow C$. While an $\mathbf{i}_{A}^{(2)} \rightarrow \mathbf{d}_{B}^{(1)}$ delay is slightly more probable than a $\mathbf{i}_{D}^{(1)} \rightarrow \mathbf{d}_{B}^{(1)}$ delay, the difference is not substantial enough to warrant the addition of another connection.

## Model comparison procedure

To compare the heuristics to participants' judgments, we simulated belief trajectories bs for all the heuristics based on the evidence generated by all participants, starting each trial with an unconnected model at $t=0$. For TS and STS, we assumed knowledge of true $\mu, \alpha$ and $w_{S}$ as participants had been trained on these during the instructions. We predicted participants' judgments based on what the simulated belief trajectories looked like at judgment time. We then assessed their accuracy in the task (e.g. the proportion of connections marked correctly) and accordance rate (the proportion of connections marked the same as the matched participant's). Additionally, we also compared participants to a Random baseline that marked a new random causal structure on every judgment, and an Ideal learner that always selects the max $P\left(M \mid \mathbf{d}_{\tau} ; \mathbf{i}_{\tau}, \mathbf{w}\right)$ according to the Bayesian inference model.

## Modeling results

The results of these simulations are reported in Table 1. Overall, STS was the most closely accordant with participants but individually participants were almost evenly split between STS and OO, both for all judgments and restricted to the final


Figure 4: Example where proposed heuristics' predictions diverge. $b^{(t-1)}$ : the learner's belief at the start of the period depicted in the timeline. After observing $d_{B}^{(t)}$ the models update $b^{(t-1)}$ to form $b^{(t)}$. Then after observing $d_{C}^{(t+1)}$, they update to form $b^{(t+1)}$. Blue lines indicate probability density for cause-effect delays starting from each event, used to determine the most likely cause of each event (TS), and whether it is sufficiently more likely than any existing causes (STS).
judgments. Participants accuracy $(0.65 \pm 0.19)$ was closest to that of the simplest heuristic OO. Mean participant accuracy by trial was correlated with that of all three heuristics $r_{\mathrm{OO}}=$ $.83, r_{\mathrm{TS}}=0.92, r_{\mathrm{STS}}=0.61$, but negatively correlated with Ideal judgments $r_{\text {Ideal }}=-.45$. Like participants but unlike the Ideal learner, all three heuristics were less accurate at cyclic than acyclic structures OO: $t(39)=9.5, p<.001$, TS: $t(39)=$ $10.6, p<.001$, STS: $t(39)=4.5, p<.001$.

## General Discussion

In our experiment, people used interventions to learn about the causal structure of devices whose dynamics unfolded in continuous time. As we predicted, cyclic structures were harder to learn than acyclic ones even though this was not reflected in the evidence available for an ideal learner, suggesting that the evidence produced by cyclic devices, involving many activations and potential causal paths, was harder for human learners to process. We found that the observed determinants of successful learning - equal spacing of interventions in time and a preference to intervene on root variables - made structure inference easier for a heuristic and

Table 1: Model comparison

| Model | Accuracy (\%) |  | Accordance (\%) |  | N best (/40) |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | All | Final | All | Final | All | Final |
| Random | 25.0 | 25.0 | 25.0 | 25.0 | 0 | 0 |
| OO | 66.2 | 64.7 | 67.2 | 64.9 | 16 | 17 |
| TS | 79.7 | 78.9 | 67.3 | 65.5 | 4 | 5 |
| STS | 87.9 | 90.9 | 69.3 | 69.2 | 15 | 13 |
| Ideal | 91.0 | 95.3 | 66.1 | 68.9 | 5 | 5 |

bounded learning system.
In light of this, we considered several heuristic learning models. Participants' judgments were best explained by assuming that they added connections to a single evolving candidate hypothesis as they observed events. Some subjects appeared to rely on a simple order heuristic (OO) whereas others displayed sensitivity to the delays between events (TS) and whether events were predicted by existing structure beliefs (STS). Participants rarely removed connections during the trials. Given more time to learn, however, it seems likely that they would also sometimes prune connections from their models - e.g., when events predicted by their current model repeatedly fail to occur. In general, positive testing of one's current hypothesis is an effective way for learners that are limited to a single global hypothesis to test its predictions against reality, and tune, refine, or or even abandon it, if necessary.

In sum, rather than grappling with an unmanageable space of possible structures and causal paths, participants seem to naturally follow Yogi Berra's advice: "You don't have to swing hard [to hit a home run]. If you got the timing, it'll go."
Acknowledgments TG was supported by the Center for Brains, Minds \& Machines (CBMM), funded by NSF STC award CCF-1231216. RM is supported by a DFG grant (Ma 6545/1-2) as part of SPP 1516 "New Frameworks of Rationality".

## References

Bonawitz, E. B., Denison, S., Gopnik, A., \& Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. Cognitive Psychology, 74, 35-65.
Bramley, N. R., Dayan, P., Griffiths, T. L., \& Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. Psychological Review, 123(3), 301-338.
Bramley, N. R., Gerstenberg, T., \& Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In Proceedings of the $36^{\text {th }}$ Annual Meeting of the Cognitive Science Society (pp. 236-242).
Bramley, N. R., Gerstenberg, T., Mayrhofer, R., \& Lagnado, D. A. (submitted). The role of time in causal learning.
Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. Acta Psychologica, 81(3), 211-241.
Buehner, M. J., \& McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. Thinking \& Reasoning, 12(4), 353-378.
Coenen, A., Rehder, R., \& Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. Cognitive Psychology, 79, 102-133.
Fernbach, P. M., \& Sloman, S. A. (2009). Causal learning with local computations. Journal of Experimental Psychology: Learning, Memory \& Cognition, 35(3), 678.
Gleick, J. (1997). Chaos: Making a new science. Open Road Media.
Greville, W. J., \& Buehner, M. J. (2010). Temporal predictability facilitates causal learning. Journal of Experimental Psychology: General, 139(4), 756-771.
Halpern, J. Y. (2016). Actual causality. MIT Press.
Lagnado, D. A., \& Sloman, S. A. (2004). The advantage of timely intervention. Journal of Experimental Psychology: Learning, Memory \& Cognition, 30, 856-876.
Lauritzen, S. L., \& Richardson, T. S. (2002). Chain graph models and their causal interpretations. Journal of the Royal Statistical Society, 64(3), 321-348.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175.
Pacer, M. D., \& Griffiths, T. L. (2015). Upsetting the contingency table: Causal induction over sequences of point events. In Proceedings of the $37^{\text {th }}$ Annual Meeting of the Cognitive Science Society.
Pearl, J. (2000). Causality. New York: Cambridge University Press (2nd edition).
Rehder, R. (2016). Reasoning with causal cycles. Cognitive Science, to appear.
Rottman, B. M., \& Keil, F. C. (2012). Causal structure learning over time: observations and interventions. Cognitive psychology, 64(1), 93-125.
Sloman, S. A., Love, B. C., \& Ahn, W.-K. (1998). Feature centrality and conceptual coherence. Cognitive Science, 22(2), 189-228.

# Harmony in a non-harmonic language: word order learning in French children 

Guillaume Braquet (s1577900@sms.ed.ac.uk)<br>Department of Linguistics and English Language, 3 Charles Street<br>Edinburgh, EH8 9AD UK

Jennifer Culbertson (jennifer.culbertson@ed.ac.uk)
Centre for Language Evolution, 3 Charles Street
Edinburgh, EH8 9AD UK


#### Abstract

Recent studies using artificial language learning have argued that the cross-linguistic frequency of harmonic word order patterns-in which heads are ordered consistently before or after dependents across syntactic categories-reflects a cognitive bias (Culbertson, Smolensky, \& Legendre, 2012; Culbertson \& Newport, 2015a). These studies suggest that English speaking adults and children favor harmonic orders of nouns and different nominal modifiers (adjectives, numerals). However, because they target English learners, whose native language is harmonic in the nominal domain (Num-Adj-N), this preference may be based on transfer rather than a universal bias for harmony. We present new evidence from French-speaking children, whose native language is non-harmonic in this domain (Num-N-Adj). Our results reveal clear effects of native language transfer, but also evidence that a harmonic pattern is favored even in this population of learners.


Keywords: cognitive biases; artificial language learning; typology; syntax; word order; French

## Introduction

Cross-category harmony (Greenberg, 1963; Hawkins, 1983) is perhaps the most well-known typological generalization of syntax: while it has been revised many times over the years (Dryer, 1992; Biberauer, Holmberg, \& Roberts, 2014), most linguists continue to assume that it reflects some underlying "law of human behavior" (Greenberg, 1966). Nevertheless, like most typological generalizations, harmony is a statistical tendency, and researchers have pointed to the possibility that correlations in word order across categories may largely reflect the effects of language contact and shared inheritance rather than universal properties of human cognition (Dunn, Greenhill, Levinson, \& Gray, 2011; Ladd, Roberts, \& Dediu, 2014).

To provide evidence of a link to cognition, a number of recent studies have investigated word order harmony using artificial language learning experiments. Culbertson et al. (2012) taught English-speaking adults a miniature artificial language in which simple noun phrases, including a noun and either an adjective or a numeral word, are used to describe a set of pictures. The languages feature variable patterns of nominal word order that tend toward either harmonic or non-harmonic. Learners tended to regularize harmonic input patterns-using the input harmonic order more frequently than it appeared in the input. By contrast, they tended to shift non-harmonic patterns toward harmonic ones rather than regularizing. Culbertson and Newport (2015a) found an even stronger effect in English-speaking children, who almost across the board altered non-harmonic input patterns to
make them harmonic (e.g., by changing the order of one of the modifiers to match the other). Indeed, even when children were taught a completely regular non-harmonic pattern (e.g., adjectives always follow and numerals always precede the noun), they still produced a harmonic output (Culbertson \& Newport, 2015b).

## Cognitive bias or native language influence?

These studies suggest the possibility that learners may preferentially change non-harmonic patterns to harmonic ones, rather than the reverse, offering a potential explanation for a similar typological asymmetry. Frequencies of the relevant nominal word order patterns are reported in the World Atlas of Language Structures Online, shown in Table 1. Harmonic patterns (both modifiers either pre-nominal or post-nominal) outnumber non-harmonic patterns. Note in addition, that within both harmonic and non-harmonic pattern types, the one with post-nominal adjectives is more frequent (i.e., N Adj with N-Num; N-Adj with Num-N).

Table 1: Frequency of harmonic and non-harmonic combinations of noun with adjective and numeral (Dryer, 2013a, 2013b).

|  | Adj-N | N-Adj |
| :--- | :--- | :--- |
| Num-N | 251 | 168 |
| N-Num | 37 | 509 |

Data from English-speaking learners reflects both a harmony bias and a bias for post-nominal adjectives to some extent. For example, English-speaking adults were equally likely to regularize pre- and the post-nominal harmonic patterns, despite the latter having less surface-level overlap with English than either non-harmonic pattern. They were also least likely to regularize the non-harmonic pattern with pre-nominal adjectives (Culbertson et al., 2012). Englishspeaking children did not exhibit any differences among nonharmonic patterns, however, their overall preference for harmonic patterns was in fact skewed toward the post-nominal N-Adj with N-Num (Culbertson \& Newport, 2015a).

Nevertheless, evidence of a harmony preference in English speakers does not unequivocally suggest an explanation for the typological asymmetry seen here. This is of course because English itself exemplifies one of the two harmonic
patterns-Adj-N with Num-N. Thus a preference for harmony in this population may be due to transfer at a more abstract level. In other words, English speakers may find the postnominal harmonic pattern easy to learn because they are used to treating numerals and adjectives similarly in terms of their relative order with the noun. If abstract transfer is responsible for the harmony bias in English speakers, then French speaking learners should behave quite differently. The default nominal word order in French exemplifies the more common non-harmonic pattern, N-Adj with Num-N. Numeral order is strictly pre-nominal, while adjectives are typically postnominal but exhibit some flexibility. Most importantly, there is a small lexically-specified set of adjectives which precede the noun. Examples of default order are shown in (1). ${ }^{1}$
(1)

> a. maison bleue house blue
> 'blue house'
> b. deux maisons
> two houses
> 'two houses'

French speakers could therefore be reasonably expected to prefer non-harmonic patterns over harmonic ones, since their prior language experience provides evidence that these two types of modifiers behave differently with respect to order. On the other hand, if the preference for harmony reflects a universal cognitive bias, then even French speakers-whose native language violates it-may exhibit its effects.

## Experiment 1

Here we explore the potential effects of native language influence and cognitive biases on nominal word order learning in French-speaking children. We follow the general design and procedure of Culbertson and Newport (2015a) and Culbertson and Newport (2015b). Children are taught a variable version of one of four patterns corresponding to those in Table 1. They are trained on simple phrases comprising a noun with an adjective or a noun with a numeral, and are then tested on their production of those phrases. The extent to which learners accurately reproduce and regularize these variable patterns is used to infer their relative preferences.

## Participants

Participants were 48 children ( 24 females), 6-7 years of age (mean $=6 ; 7$, matched with participants in Culbertson and Newport (2015a)). They were recruited from elementary schools in Southwest France, and were native speakers of French, who were either monolingual, or bilingual in French and Occitan (a Romance language spoken in this region, which uses the same nominal word order as French). Parental consent was obtained for all participants. Three additional children were excluded from the analysis due to failure to

[^35]complete the experimental session (2), or extremely low score on vocabulary learning (1).

## Materials

The artificial language consisted of 10 words: 4 nouns corresponding to novel objects, 3 adjectives, and 3 numeral words. Following Culbertson and Newport (2015a), nouns were nonce words which were phonotactically plausible in the participants' native language, and modifiers were pseudononce words resembling the corresponding native language words. This lexicon was used to describe pictures like those shown in Figure 1. Importantly, all of the adjectival modifiers used here appear in the default post-nominal order in French, they are not in the set of pre-nominally ordered adjectives.

Table 2: Artificial language lexicon.

| Nouns | Adjective | Numerals |
| :--- | :--- | :--- |
| [bogi] | [bly] (bleu 'blue') | [doks] (deux, 'two') |
| [sefi] | [tachu] (tacheté, 'spotted') | [tra] (trois, 'three') |
| [voli] | [pølu] (poilu, 'furry') | [kits] (quatre, 'four') |
| [kani] |  |  |



Figure 1: Example visual stimuli.

## Design \& Procedure

Participants were randomly assigned to one of four input word order conditions. Each condition featured a dominant pattern for each modifier type, used in $75 \%$ of utterances. The remaining $25 \%$ used the alternative order. The variation present in the input was unpredictable; it was not conditioned on any particular lexical items. These conditions are illustrated in Table 3.

Table 3: Conditions, according to dominant pattern type (shaded cells are non-dominant.

|  | Adj-N | N-Adj | Num-N | N-Num |
| :--- | :---: | :---: | :---: | :---: |
| Harmonic dominant |  |  |  |  |
| Pre-N | $75 \%$ | $25 \%$ | $75 \%$ | $25 \%$ |
| Post-N | $25 \%$ | $75 \%$ | $25 \%$ | $75 \%$ |
| Non-harmonic dominant |  |  |  |  |
| N-Adj, Num-N | $25 \%$ | $75 \%$ | $75 \%$ | $25 \%$ |
| Adj-N, N-Num | $75 \%$ | $25 \%$ | $25 \%$ | $75 \%$ |

Following Culbertson and Newport (2015b), the experiment consisted of a single session, lasting approximately 30 minutes. During this session participants were trained and
tested on the language with an experimenter present. The experiment was presented using PsychoPy software (Peirce, 2009) on a Macintosh laptop in a quiet corner of the child's classroom. Children were told they would be playing a game to learn an alien language with the help of a friendly alien named Clémy.

The experiment began with a series of games designed to teach children the novel nouns and their meanings. The first game ( 20 trials) involved seeing a grayscale picture of a single object, listening to the label provided by the alien speaker, and repeating the label aloud. The second game ( 20 trials) involved listening to a label and clicking on whichever of the four novel objects it corresponded to. Feedback highlighting the correct picture was given on all trials. A sound indicated whether the child's choice was correct or incorrect. The third game ( 20 trials) tested children's ability to provide the correct label for an object shown on the screen. Feedback in the form of the correct noun label was provided on all trials.

The second part of the experiment consisted of a series of similar games designed to teach children simple phrases in the language. Recall that each phrase consisted of either a noun with an adjective or a noun with a numeral, but never both. Half of all trials featured an adjective, and half a numeral. The first game (48 trials) involved seeing pictures as in Figure 1 above, and hearing a phrase to describe it provided by the alien speaker. Participants were told to repeat the phrase. The second game ( 48 trials) involved hearing a phrase and choosing the corresponding picture from an array of four choices. Feedback was given as described above. Finally, children were shown a picture, and were asked to provide a description for it ( 48 trials). No feedback was given, however if the child had trouble with one of the words, the experimenter would help.

## Results

Here we report results from the final phrase production task. Comprehension of phrases was uniformly high across conditions ( $>85 \%$ ). Figure 2 shows the average proportion of trials in which children used the dominant word order in each of the four word order input conditions. Children used the dominant order almost exclusively in the post-nominal harmonic condition, however dominant order use in the other three conditions was much lower. Children roughly matched the input proportions in the non-harmonic N -Adj, Num- N condition, which is most similar to their native language pattern.

Figure 3 shows use of the dominant pattern broken down by the type of modifier, illustrating more clearly what learners are producing when they don't match the input. While in the post-nominal harmonic condition both modifier types are regularized equally often, there is a clear difference between modifier types for the other three conditions. In the pre-nominal harmonic condition, children tended to use the numeral in the dominant input order more often, however neither modifier type reproduces the input pattern closely. In the two non-harmonic conditions, children were more likely to match the dominant input order for whichever modifier


Figure 2: Average proportion use of dominant input order by condition. The dotted line is the proportion of the dominant order used in the input (75\%). Error bars represent 95\% CIs.
tended to appear post-nominally: the adjective in the Frenchlike N -Adj, Num-N condition, and the numeral in the Adj$\mathrm{N}, \mathrm{N}$-Num condition. In other words, learners' productions in these three conditions tended to move the language away from the input pattern and toward a post-nominal harmonic pattern.


Figure 3: Average proportion use of dominant input order for each modifier type by condition. The dotted line is the proportion of the dominant order used in the input (75\%). Error bars represent 95\% CIs.

These data were analyzed using mixed-effects logistic regression as implemented in the lme 4 R package (Bates, 2010), with condition (input pattern type) and modifier type (adjective or numeral) as fixed effects and participants and items (stimulus picture) as random effects. Modifier type was sum coded; condition was treatment coded with the most French-like pattern ( $\mathrm{N}-\mathrm{Adj}$, Num-N) as the baseline
level. This model revealed a significant difference between the French-like ( $\mathrm{N}-\mathrm{Adj}, \mathrm{Num}-\mathrm{N}$ ) condition and each of the other three conditions: compared to this condition there was significantly less use of the input order in the pre-nominal harmonic condition $(\beta=-3.99 \pm 0.56, p<0.001)$ and the nonharmonic Adj-N, N-Num condition ( $\beta=-3.60 \pm 0.54, p<$ 0.001 ), but significantly more use of the input order in the post-nominal harmonic condition ( $\beta=17.88 \pm 17.03, p=$ 0.03 ). A significant main effect of modifier type was also present, indicating less matching of numeral order overall ( $\beta=-2.28 \pm 0.22, p<0.001$ ). Finally, significant interactions between condition and modifier type were found. As suggested by Figure 3, the main difference between the French-like and post-nominal harmonic conditions is in use of the dominant order for the numeral ( $\beta=-12.50 \pm$ $17.04, p=0.04)$. By contrast the opposite was true for both the pre-nominal harmonic $(\beta=4.45 \pm 0.33, p<0.001)$ and non-harmonic Adj-N, N-Num ( $\beta=4.89 \pm 0.31, p<0.001$ ) conditions where the bigger difference with the French-like condition was in use of the dominant adjective order. ${ }^{2}$

A visualization of the general direction of change from the input in each condition can be seen in Figure 4. This shows the proportion of pre-nominal adjectives and numerals each individual child produced, colored by their input condition. Learners generally cluster in the harmonic post-nominal area of the space, no learners for whom this was the input condition shifted away from this pattern. A smaller cluster appears around the non-harmonic French-like ( $\mathrm{N}-\mathrm{Adj}$, Num-N) area, with children in that input condition plus some from the prenominal harmonic input condition who have switched only the adjective order to post-nominal. Figure 5 summarizes this picture by calculating each child's preferred pattern, determined by the order used in the majority of utterances (greater than $50 \%$ ) for each modifier type. For example, one child in the French-like N-Adj, Num-N condition produced adjectives post-nominally $100 \%$ of the time, and numerals postnominally $88 \%$ of the time. This child was thus classified as having the post-nominal harmonic pattern.

## Discussion

French-speaking children learning variable patterns of harmonic or non-harmonic nominal word order showed a striking pattern of behavior in their productions. When their input featured the post-nominal harmonic pattern $\mathrm{N}-\mathrm{Adj}$, $\mathrm{N}-\mathrm{Num}$ as the dominant order, they regularized this pattern, producing nearly deterministic output. When their input featured a dominant pattern similar to their own native language, N - Adj , Num-N, they matched or regularized the adjective order, but not the numeral order. To the extent that children were fail-

[^36]

Figure 4: Output patterns for each individual participant, colored by condition, defined as proportion pre-nominal order for each modifier type. Points are jittered to prevent overplotting. Larger points outlined in black represent input proportions for each condition. Dashed lines provide a visualization of rough pattern types.


Figure 5: Number of participants in each condition whose preferred pattern (used in $>50 \%$ of utterances) corresponds to each pattern type.
ing to match the pre-nominal numeral order, they were necessarily switching it to post-nominal, in harmony with the adjectives. This same behavioral pattern, of switching the predominantly pre-nominal modifier to match the post-nominal one, is seen to a more extreme degree in the non-harmonic Adj-N, N-Num condition. Children in that condition generally matched the post-nominal numeral order, and showed a very strong tendency to switch the adjective to follow as well. In the pre-nominal harmonic condition, a similar pattern is again found: children generally produced more postnominal phrases than were present in the input for both modifier types, but particularly the adjective. Overall then, most
of the children in this experiment produced a pattern corresponding most closely to the post-nominal harmonic N -Adj, N -Num.

Recall that above we suggested two competing hypotheses which generated different predictions about the behavior of French-speaking learners in our experiment. Under the first hypothesis, influence from the native language is the primary driver of behavior. Previous results with English speaking learners were consistent with abstract level transfer (i.e., a general preference for harmonic patterns). Thus French speakers would be predicted to prefer non-harmonic patterns, though a preference for the specific native language pattern (here N -Adj, Num-N) is also possible. The second hypothesis is that a cognitive bias favoring harmonic patterns is present universally across learners, even learners whose native language actively violates it. This predicts that French-speaking learners will, like English speakers, prefer harmonic patterns. The results reported here are not fully consistent with either hypothesis. Most obviously, learners did not show a preference for the pattern, or pattern type, most similar to their native language. Neither did they prefer harmonic patterns across the board. However, a combination of native language transfer effects and a universal harmony bias provides a coherent explanation. French-speaking children in our experiment exhibited a strong preference for post-nominal adjective order, in accord with their native language. This preference was generalized to numerals under a pressure for harmony.

In order for this explanation to work, we need to rule out a potential alternative: that French-speaking children age 67 years do not distinguish numerals from adjectives, or have not fully mastered the order of numerals in their native language. This could lead to the over-use of post-nominal ordering if children have, by contrast, mastered the default adjective order. There is relatively little work on the acquisition of nominal word order in monolingual French children. Indeed, we are not aware of corresponding work on the acquisition of numeral-noun order. We therefore extracted all instances of noun phrases including one of the numeral words 'two' through 'ten' in the Lyon corpus (Demuth \& Tremblay, 2008). ${ }^{3}$ This is a publicly available corpus of naturalistic parent-child interactions including 5 children, recorded for 1 hour every 1-2 weeks from age 1 to 3 years ( 185 hours of speech total). Children's first noun phrases with a numeral word occurred as early as $1 ; 9$. Out of a total of 258 instances, no word order errors were found. Based on this evidence, it appears likely that by 6-7 years of age, French-speaking children have long since mastered the pre-nominal order of numerals words in their language. Given the variation in adjective ordering in French, it would seem plausible that children may take longer to acquire this aspect of the syntax of their native language. However, evidence from spontaneous speech suggests that French-English bilingual children have

[^37]mastered French adjective order very early as well, by $2 ; 5$ at least (Nicoladis, 2002). This may be accomplished earlier in monolinguals, in line with the early documented acquisition of adjective order in English by the age of 2 (Brown, 1973). In our study then, it does not seem likely that French-learning children's preference for post-nominal order in both adjectives and numerals is the result of a lack of knowledge about the syntax of these categories in their native language.

What remains then, is to understand how the results of this study fit with with the previous findings from Englishspeaking adults and children. As mentioned above, Englishspeaking child learners did not generalize their native prenominal order preferentially, but rather readily produce both harmonic patterns. If anything, there was somewhat stronger preference for the post-nominal harmonic pattern (Culbertson \& Newport, 2015a, 2015b). For English-speaking adults, both harmonic patterns were regularized, and among nonharmonic patterns the one with pre-nominal adjectives (Adj$\mathrm{N}, \mathrm{N}-\mathrm{Num}$ ), was particularly dispreferred (Culbertson et al., 2012). Recall that the typological distribution in Table 1 suggests both a preference for harmonic patterns, and a general preference for post-nominal adjectives. Thus among the harmonic patterns, the post-nominal one is more common, and within the non-harmonic patterns, N -Adj, Num-N is the most common. If these two pressures are at work across both learner populations, English and French, then we expect to see behavior mirror the typology. However, if these two biases are, as suggested above, influenced by learners' native language experience, then a more complex picture emerges. For English-speaking children (and adults), the harmony bias is strengthened by native language experience, while the preference for post-nominal adjectives is weakened. As a result, learners strongly prefer harmonic languages, and only weakly prefer post-nominal over pre-nominal order. For Frenchspeaking children, the opposite holds. The harmony bias is weakened, while the preference for post-nominal adjectives is strengthened. This results in a strong preference for the post-nominal harmonic pattern only, and changes to the input in each other condition which move the language toward that pattern via the adjective. Based on differences found between English-speaking adults and children, we would predict that French adults should show a similar but less dramatic pattern of preferences to children in our study.

## Why these biases?

Both of the biases we have argued to be at work here reflect potentially quite general cognitive mechanisms. Harmonic patterns are simpler, in the sense that they involve fewer, more general rules, which can be generalized across categories. For extensive discussion of this idea, see Culbertson and Kirby (2016). The preference for post-nominal adjectives may reflect a pressure to establish the object of modification first, particularly in cases where the meaning of the noun is predictive in determining the meaning of the adjective, e.g., for gradable adjectives like 'tall' (Kamp \& Partee, 1995; Ramscar, Yarlett, Dye, Denny, \& Thorpe, 2010). Interestingly,

Nicoladis (2006) shows that French-English bilingual children ( $2 ; 11-5 ; 3$ ) produce more reversals of adjective order than their monolingual counterparts in both French and English. These reversal errors were more likely to involve adjectives incorrectly in the post-nominal position than the reverse. In French, this corresponded to placing a typically pre-nominal adjective like grand 'big' after the noun. In English, this corresponded to placing adjectives in an incorrect post-nominal position when they were post-nominal in French (e.g., 'a monkey purple'). This is in line with our findings in the sense that post-nominal adjective order seems to be more readily generalized than pre-nominal order. Overall then, the preference for post-nominal adjective ordering found among languages of the world and reflected in both English and French learners, may be related to general properties of learning and processing.

## Conclusion

This study tested a classic hypothesis in linguistics: that harmonic word order patterns, which maintain a consistent order of syntactic heads relative to modifiers, are preferred to non-harmonic alternatives. Previous studies have shown that English-speaking learners prefer harmony in the nominal domain, however this could reflect abstract transfer since English exemplifies a pre-nominal harmonic pattern. Here, we targeted French-speaking child learners, whose native language is non-harmonic. If a harmony bias is present in this population it would provide a strong indication that this is indeed a universal pressure. By contrast, if transfer is the main driver of learning behavior, then no harmony bias is expected in this population-instead, non-harmonic patterns may be preferred. Our results revealed a strong preference for post-nominal harmonic order, which we argue reflects the effects of a harmony bias in conjunction with a preference for post-nominal adjectives.

## Acknowledgments

We thank the schools, parents and children who participated in this study and the developmental linguistics reading group at the University of Edinburgh for comments on this work.

## References

Bates, D. (2010). lme4: Mixed-effects modeling with R. http://lme4.r-forge.r-project.org/book.
Biberauer, T., Holmberg, A., \& Roberts, I. (2014). A syntactic universal and its consequences. Linguistic Inquiry, 45(2), 169-225.
Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard University Press.
Culbertson, J., \& Kirby, S. (2016). Simplicity and specificity in language: Domain general biases have domain specific effects. Frontiers in Psychology, 6(1964). doi: 10.3389/fpsyg.2015.01964
Culbertson, J., \& Newport, E. L. (2015a). Harmonic biases in child learners: In support of language universals. Cognition, 139, 71-82.

Culbertson, J., \& Newport, E. L. (2015b). Language learning and word order regularities: Children's errors reflect a typological preference for harmonic patterns. Talk given at BUCLD 40.
Culbertson, J., Smolensky, P., \& Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122, 306-329.
Demuth, K., \& Tremblay, A. (2008). Prosodicallyconditioned variability in children's production of French determiners. Journal of Child Language, 35(1), 99-127. (Cited By (since 1996): 1)
Dryer, M. (1992). The Greenbergian word order correlations. Language, 68(1), 81-183.
Dryer, M. (2013a). Order of adjective and noun. In M. Haspelmath, M. S. Dryer, D. Gil, \& B. Comrie (Eds.), The world atlas of language structures online (chap. 87). Munich: Max Planck Digital Library. Retrieved from http://wals.info/chapter/87
Dryer, M. (2013b). Order of numeral and noun. In M. Haspelmath, M. S. Dryer, D. Gil, \& B. Comrie (Eds.), The world atlas of language structures online (chap. 89). Munich: Max Planck Digital Library. Retrieved from http://wals.info/chapter/89
Dunn, M., Greenhill, S., Levinson, S., \& Gray, R. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. Nature, 473(7345), 7982.

Fox, G., \& Thuilier, J. (2012). Predicting the position of attributive adjectives in the french np. In New directions in logic, language and computation (pp. 1-15). Springer.
Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), Universals of language (p. 73-113). Cambridge, MA: MIT Press.
Greenberg, J. (1966). Language universals. The Hague: Mouton.
Hawkins, J. A. (1983). Word order universals. New York: Academic Press.
Kamp, H., \& Partee, B. (1995). Prototype theory and compositionality. Cognition, 57(2), 129-191.
Ladd, D. R., Roberts, S. G., \& Dediu, D. (2014). Correlational studies in typological and historical linguistics. Annual Review of Linguistics, 1 .
Nicoladis, E. (2002). The cues that children use in acquiring adjectival phrases and compound nouns: Evidence from bilingual children. Brain and Language, 81(1), 635-648.
Nicoladis, E. (2006). Cross-linguistic transfer in adjectivenoun strings by preschool bilingual children. Bilingualism: Language and Cognition, 9(01), 15-32.
Peirce, J. W. (2009). Generating stimuli for neuroscience using psychopy. Frontiers in Neuroinformatics, 2(10).
Ramscar, M., Yarlett, D., Dye, M., Denny, K., \& Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. Cognitive Science, 34(6), 909-957.

# How can I help? 24- to 48-month-olds provide help specific to the cause of others' failed actions 

Sophie Bridgers (sbridge), Sara Altman (skaltman), and Hyowon Gweon (hyo@stanford.edu)<br>Department of Psychology, Stanford University


#### Abstract

When young children see others fail to achieve a goal, they spontaneously help. But there are many reasons why someone might fail, and consequently, many ways to help. In order to help effectively, we need to understand why someone is failing, so we can address the cause. One important distinction is whether the failure is due to the agent's own actions or something external to her in the world. Here we show that 24 - to 48 -month-olds can use their past experience to reason about the probable cause of another person's failure and provide help appropriate for that cause. Children's help targeted the world when their prior knowledge suggested that the source of failure was external to the agent, and targeted the person's actions when this source appeared to be internal to the agent.


Keywords: social cognitive development; prosocial behavior; causal reasoning; theory of mind; helping

## Introduction

Imagine a frustrated traveler at a train station, fumbling with a ticket machine. Chances are someone will offer help, but how this person helps might depend on the situation. If the helper sees that the traveler is inserting the bill in the wrong direction, she might kindly re-orient the bill; however, if the helper knows that the ticket machine is out of order, she might direct the traveler to another machine nearby.

Humans are remarkably helpful creatures from an early age (Tomasello, 2009). Although preverbal infants may not yet be able to offer help with ticket machines, they will pick up objects others have dropped and pass them back, and show someone struggling to reach inside a box an easier way to get in (Warneken \& Tomasello, 2006). As in the ticket machine example, however, helping others is not only a decision of whether to help, but also a decision of how to help.

The decision of how to help is sometimes straightforward. When someone is struggling to hold open a door, there is typically only one way to help. In many contexts, however, it is not so clear what kind of help is needed. In order to figure out how to help others, we need to understand why someone is struggling. Critically, more often than not, it is up to the helper, rather than the helpee, to determine exactly what kind of help is needed. If the traveler knew the reason why she was failing to insert her bill, she might have already solved the problem. When deciding how to help, therefore, it is critical to determine the source of an actor's failure: whether the failure is due to the actor herself, or due to the external world.

We routinely make these judgments about our own failed actions. If everyone else bought a ticket from the machine but somehow you cannot, you are likely using the machine incorrectly. However, if everyone experienced occasional failure, perhaps the machine is not reliable. Furthermore, these inferred reasons inform our decisions about what to do next: if
you are the source of your failure, you might seek help; if the machine is broken, you might try a different one.

Previous work suggests that even infants can infer the causes of their own failed actions, and respond appropriately to achieve their goals (Gweon \& Schulz, 2011). In this study, children were given covariation evidence indicating either that a toy sometimes worked and sometimes did not (regardless of the agent), or worked for some agents but not others. In the former case, when children failed to activate the toy, they reached for a new toy, suggesting they attributed their failures to the toy and not to their own actions. In the latter case, they were more likely to hand the toy to their mother, suggesting they inferred that their failures were due to something about their own actions and not the toy. In this case, infants were able to determine the source of their failures because they had observed others' interactions with the toy. Without such information, they would not have been able to determine why they had failed; they might not know what to do without a knowledgeable agent's help. Critically, this agent's help would be most effective is she could directly address the cause of the child's failure.

Here, we ask whether young children can reason about the cause of others' failures to inform their decisions of how to help. There are reasons to believe that even very young children may be able to do this. Studies with toddlers and preschoolers suggest that young children can provide help not only when the helpee's needs are straightforward and observable (e.g., picking up dropped objects, Warneken \& Tomasello, 2007), but also when her needs are more internal and abstract (e.g., beliefs, goals, competence). For instance, 12-month-olds are more likely to point out the location of a dropped object if their social partner has not seen it fall than when she has seen it (Liszkowski, Carpenter, \& Tomasello, 2008); 18-month-olds can use their social partner's prior experience to infer different goals from the same failed action and help her achieve that goal (Buttelmann, Carpenter, \& Tomasello, 2009). Preschoolers (42-month-olds) can reason about other peoples' action capabilities to predict from whom someone else will ask for help (Paulus \& Moore, 2011), and even anticipate that their social partner needs help before she does and pre-emptively intervene to help her achieve her ultimate goal (Bridgers, Jara-Ettinger, \& Gweon, 2016; Martin \& Olson, 2013).

However, inferring the possible causes of others' failed goal-directed actions might be more challenging than reasoning about the goals themselves, as the child must decide between (at least) two competing hypotheses which may or may not be observable in the failed action itself. Furthermore, us-
ing this inferred cause to generate the appropriate helpful behavior is also a nontrivial task, and arguably more difficult than helping in contexts in which there is one clear way to help. Even though prior work suggests that preverbal infants can make these inferences about their own failures and decide what to do to achieve their own goals, applying the same inferences to others' actions and providing the most effective help may be more challenging. Indeed, there is a body of research suggesting that reasoning about one's own actions may precede, and is a necessary precursor to, reasoning about others (e.g., Sommerville \& Woodward, 2005; Sommerville, Woodward, \& Needham, 2005).

The current study investigates 24 - to 48-month-olds' abilities to reason about the causes of others' failed actions and offer help accordingly. Children observed an adult fail to activate a toy because she either (a) used the toy incorrectly, or (b) chose a faulty toy. We then gave children the choice to help by either handing the person a working toy or by demonstrating the correct way to use the toy. Individually, both options were perfectly reasonable ways to help, and within the repertoire of behaviors children have exhibited in prior work (e.g., Warneken \& Tomasello, 2006). The critical question here, however, is whether children will provide help that best addresses the likely cause of this person's failure (i.e., the toy or her own actions).

## Experiment

In our experiment, we created a situation in which children were faced with two ways to help. We manipulated the cause of the helpee's failure, which made one way more effective than the other. We recruited 24- to 48-month-olds, who were slightly older than children in other studies that reported spontaneous helping behaviors (e.g., Warneken \& Tomasello, 2006; Cortes Barragan \& Dweck, 2014), as the ability to help others appropriately based on the cause of their failure might require richer representations about others' goal-directed actions and more sophisticated inferential abilities.

## Methods

Participants Fifty-two 24- to 48-month-olds $(M(S D)=$ $2.78(.48) \mathrm{yrs}, 44 \%$ female) from a museum in Palo Alto, CA participated. An additional 15 children were excluded from analysis due to parental interference ( $n=9$ ), experimenter error $(\mathrm{n}=1)$, shyness ( $\mathrm{n}=3$ ), or lack of video recording $(\mathrm{n}=2)$. We randomly assigned children to one of two conditions: the Broken Toy condition ( $\mathrm{n}=26 ; M(S D)=2.78(.54) \mathrm{yrs}$ ) or the Wrong Action condition ( $\mathrm{n}=26 ; M(S D)=2.78(.43) \mathrm{yrs}$ ).
Stimuli We constructed 3 identical-looking toys. One side of each toy was covered in yellow felt and had a yellow button in the center. The opposite side was covered in red felt and had a red button in the center. The yellow button on two toys played music, while the yellow button on the third toy was inert. On all 3 toys, the red buttons were always inert. The toys were placed on a white-plastic tray and covered with grey felt. See Figure 1 for a schematic of the toys and procedure.

Procedure The experiment began with a warm-up phase in which a confederate and experimenter engaged the child in reciprocal games (e.g., rolling a ball back and forth through a tube) in order to help the child feel comfortable with the researchers, and promote general helping behavior (see Cortes Barragan \& Dweck, 2014). After approximately 5 minutes of warm-up, the confederate excused herself from the room, explaining that she had work to do.

Next came the play phase in which the child gained experience with the toys. The experimenter did not pedagogically demonstrate how the toys worked, but instead behaved as if she were exploring the toys and discovering what they did. She took one toy out at a time and showed it to the child. In the Broken Toy condition, the toys were oriented such that the yellow side was on top. She noticed the yellow button, pressed it, and reacted positively to the music that played. She also encouraged the child to press the yellow button and again reacted positively, saying, "Music! The yellow side plays music!". She then turned the toy around in her hands until she discovered the red button on the opposite side, and expressed mild surprise, as if she did not expect it to be there. She pressed the red button and also encouraged the child to do so, acting perplexed and disappointed that it did not play music. The experimenter then took the second toy out, which she and the child explored in the same way (i.e., the experimenter pressed each button, and then encouraged the child to do so). This second toy was always the broken toy, so neither button played music. This process was repeated with the third toy, which functioned the same as the first (i.e., the yellow button played music, but the red button did nothing). The child and experimenter then explored each toy again, taking turns pressing the buttons. In the Wrong Action condition, everything was the same except that the toys were placed with the red side up, such that the red button was discovered first, and then the yellow. By the end of this phase, all children experienced that pressing the yellow buttons on two of the toys played music (and one was inert), and that none of the red buttons played music.

In the helping phase, the experimenter placed toys back on the tray and covered them with the felt. The toys were placed as they were during the play phase: yellow-side-up in the Broken Toy condition, and red-side-up in the Wrong Action condition. The child sat approximately 6 ft away from the tray, either by him-/herself or with a parent. The experimenter then called the confederate back into the room and explained that she and the child were playing with toys that played music. The confederate said, "I love music!" and knelt down behind the tray, facing the child. She appeared to select a toy at random from behind the felt; the child could not see which toy was chosen.

The confederate then moved the tray (which contained 2 of the 3 toys, covered by the felt) off to one side (counterbalanced) and placed her chosen toy in front of her. She pressed the button on top (the yellow button in the Broken Toy condition; the red button in the Wrong Action condition), and the


Figure 1: Schematic of the play phase and the helping phase and the toys used in both conditions.
toy did not play music. The confederate remarked, "Hmm, no music!" and pressed the button again, expressing disappointment and saying, "Still no music! I really want to play music!" She then put one hand on the tray, and at the same time, slid her toy with the other hand such that it was parallel with the tray but on her opposite side. Once the toy and tray were equidistant from the confederate, she removed her hand from the toy and removed the felt from the tray to reveal the two other toys. She then gestured to both the toy and the tray and asked, "Can you help me play music?" The toy and tray were far enough apart (approx. 2 ft ) and from the child (approx. 5 ft ) that $\mathrm{s} /$ he could not approach both simultaneously.

If the child did not respond, the confederate and experimenter provided planned prompts, waiting 5 seconds in between, until the child responded. The last of these prompts involved the confederate moving closer to the child (approx. 2 ft .) and placing the tray and the toy within the child's reach but still far enough apart that the child could only reach to one location at a time.

In summary, the only difference across conditions was whether the non-obvious button (i.e., the button on the bottom of the toy) that the experimenter revealed to the child during the course of the play phase was non-functional (Broken Toy condition) or functional (Wrong Action condition). In both conditions, the confederate pressed the obvious button (i.e., the button on top) and the toy did nothing.
Coding We were interested in children's first helping response after the confederate's failure to activate the toy (i.e., her first button press). The key dependent measure was the target toy of this behavior, coded as either the "confederate's toy" or the "toys on the tray". All children who responded fell into one of these two categories.

Additionally, we looked at the consequence of children's first helping responses. We coded whether their behavior was "successful or "unsuccessful" in achieving the confederate's goal of playing music. In the Broken Toy condition, a child's first response was coded as "successful" if the child pressed the yellow button on a toy from the tray or directed
the confederate to press it (e.g., telling her to do so; handing or pointing to a toy yellow side up); "unsuccessful" responses included pressing or directing the confederate to press the red button on any toy or the yellow button on the confederate's toy. In the Wrong Action condition, a behavior was coded as "successful" if a child flipped and pressed the yellow button or directed the confederate to do so (e.g., telling her to press it, flipping a toy and handing or pointing to it yellow-sideup). Thus, in the Wrong Action condition, a behavior could be successful regardless of which toy a child's first response targeted, whereas in the Broken Toy condition, only behavior directed toward the toys on the tray could be successful. The first and second author transcribed and coded children's behavior and a researcher blind to the hypotheses coded these transcriptions for reliability; agreement was $100 \%$.

## Predictions and Results

Children in both conditions saw the same set of toys and a confederate fail in the same way (she pressed an obvious button on top of a toy and it did not play music). Furthermore, in response, all children could approach either the confederate's toy or a toy on the tray. What differed across conditions was the likely cause of the confederate's failure. We manipulated the source of failure by varying whether the obvious button on top of the toy was functional on 2 of the 3 toys (i.e., yellow button; Broken Toy condition) or non-functional on all 3 toys (i.e., red button; Wrong Action condition).

We predicted responses to vary across conditions depending on the source of the confederate's failure. In the Broken Toy condition, the likely reason for her failure was the toy and not her own action. Thus, it was more helpful to get a new toy (a toy on the tray) than to act on the confederate's toy. In the Wrong Action condition, however, the likely reason was the confederate's action and not the toy, suggesting that children could help by approaching the confederate's toy to correct her action (there was less need to get a new toy). Thus, we predicted that more children would approach the "toys on the tray" in the Broken Toy condition than in the Wrong Action condition. As predicted, children were significantly more


Figure 2: Proportion of children whose first response was directed to the confederate's toy (blue) or the toys on the tray (red) in the helping phase. Error bars: bootstrapped 95\% CI.
likely to direct their help toward a toy on the tray in the Broken Toy condition than in the Wrong Action condition (73\% vs. $27 \%$; two-tailed Fisher's Exact Test, $p=0.002$ ).

We then looked at children's responses within each condition. In the Broken Toy condition, it was clear that the confederate was acting on the broken toy, and that children could offer help only by approaching a toy on the tray. No action on the confederate's toy could yield music. We thus predicted that children in this condition would preferentially direct their help toward a toy on the tray. Indeed, children were more likely to approach the "toys on the tray" than the "confederate's toy" (19/26; two-tailed binomial test, $p=0.029$ ).

In the Wrong Action condition, children could, in principle, help the confederate by showing the yellow button on either the confederate's toy or a toy on the tray. In fact, unlike in the Broken Toy condition, the outcome of their help was probabilistic, as there was a $33 \%$ chance that any toy children chose to flip over would be broken. However, there were reasons to expect a preference for the confederate's toy in the Wrong Action condition. First, children might have been inclined to approach the toy on which the confederate had just acted. Second, by acting on the object with which she failed, children can guarantee that they are offering help to achieve her specific goal to activate that toy. Finally, while approaching a toy on the tray changes two variables (both the object and the agent), by acting on the confederate's toy, children can more clearly disambiguate the cause of her failure. Thus, we expected that children might show a mild preference for the confederate's toy, although we did not have a strong a priori prediction. The results showed that the majority of children in the Wrong Action condition did approach the "confederate's toy" (19/26, two-tailed binomial test, $p=0.029$ ). See Figure 2 for a summary of children's first responses.

Our secondary measure of interest was the success of children's helping responses (i.e., did their help enable the confederate to achieve her goal of playing music?). Successful behavior in the Wrong Action condition was arguably more
complex than in the Broken Toy condition, as children had to reveal the non-obvious button on the bottom of a toy. In the Broken Toy condition, children simply had to point out another obvious button on a different toy. Despite this difference, children's help did not differ across conditions (twotailed, Fisher's exact test, $p=0.01$ ) and was remarkably successful overall. The majority of children engaged in successful helping behavior ( $44 / 52,85 \%$ ), and this trend was consistent within each condition (Broken Toy: 19/26, 73\%; Wrong Action: 25/25, 100\%). In the Broken Toy condition, children's help could only be successful if they approached the "toys on the tray". Of the children who did this, $100 \%$ of them were successful. In the Wrong Action condition, children's help could be successful if they approached either the "toys on the tray" or the "confederate's toy". One child was dropped from this analysis because the camera angle prevented clear visual access to the nature of her helping behavior, but all children included provided successful help.

Finally, as an exploratory analysis, we re-coded children's first responses as "correct" (Broken Toy: "toys on tray"; Wrong Action: "confederate's toy") or "incorrect". We fit a generalized linear model with correctness as the outcome variable, condition as a categorical predictor variable, and age as a continuous predictor variable. This analysis revealed no difference in children's tendency to behave "correctly" by condition or age (condition: $\beta=-.364, z=-.471, p=.638$; age: $\beta=1.533, z=1.614, p=0.107$ ).

## Discussion

Our results suggest that 2- and 3-year-old children were able to infer the likely cause of another person's failure and offer help that appropriately addressed this cause. Rather than simply helping the confederate with the toy she previously tried but failed to activate, or offering her a new toy across the board, children selectively approached the confederate's toy or a new toy depending on the source of the confederate's failure. More specifically, when children's prior knowledge suggested that the confederate was failing due to something about the world (e.g., a faulty toy), they provided help that changed this external variable (i.e., acting on a new toy in the same way). But when her own action was the likely culprit, children helped by keeping the world constant and showing her the correct action to take (i.e., acting on the same toy but in a different way). Moreover, beyond simply directing their help toward the likely cause, they provided assistance that successfully fulfilled the confederate's goal.

These results support the idea that young children are not just motivated to help (Tomasello, 2009); they are also motivated (and able) to provide help that is appropriate and effective. From a brief training with the causal structure of simple toys, children as young as 2 years of age were able to use their prior experience to infer the cause of the actor's failure, and intervene in a way that specifically targeted this cause.

In the Broken Toy condition, the toy was clearly the cause of the confederate's failure, and the only way to help was
to get her a new toy. In the Wrong Action condition, the confederate's action was clearly the cause of her failure, but there was more than one way to help: you could show her the right action on her toy or another toy. Children in this condition appeared sensitive to this response ambiguity. Although most children approached the confederate's toy (19/26; 73\%) and revealed the functional button on the bottom, some (7/26; $27 \%$ ) approached the toys on the tray and flipped one of these toys over instead. Therefore, although the children who approached the tray may have thought the confederate's toy was broken, their behavior suggests they still attributed the confederate's failure to her action.

Though children in the Wrong Action condition could have helped the confederate by revealing the functional button on the confederate's toy or on a toy from the tray, children still preferentially approached the confederate's toy. While this tendency could be a simple inclination to approach a toy that someone else has chosen before, it could also reflect more sophisticated reasoning about how best to help the confederate. First, showing the confederate the correct action on the same toy directly helps her achieve her specific goal of making that toy play music. Second, by acting on the confederate's toy, children can effectively hold the "toy" variable constant and vary just the "action" variable. Thus, even though the probability of providing effective help is the same (67\%) for any of the toys, the outcome of the action is more informative when the child acts on the confederate's toy. If this action is successful (i.e., the toy plays music), then the confederate's action was wrong; if this action is unsuccessful, the confederate has a broken toy. By contrast, acting in a new way on a different toy can only be informative for the confederate if the action is successful. The exact reasoning underlying children's preference for the confederate's toy remains unclear and is an avenue for future work.

The content of children's helping behavior provides a more nuanced picture of their reasoning about the confederate. Children not only seemed to reason about the confederate's observable, failed action but also her internal mental states. In the Broken Toy condition, all of the children who successfully helped indicated that the confederate should try to press the yellow button on one of the toys on the tray. However, none of these children provided exhaustive information about the functionality of the toys (i.e., they did not reveal the nonfunctional red button on the bottom). This suggests that the children were sensitive to the fact that the confederate's goal was to play music, rather than to learn how the toys worked. This finding is consistent with prior research showing that 4- to 5-year-olds adjust the amount of information they provide depending on whether their social partner wants to know how a toy works or simply wants to see what the toy does: children were more likely to provide information that fully disambiguated the causal system for the former than for the latter goal (Gweon, Chu, \& Schulz, 2014). Would children in our current age range similarly demonstrate the other side of the toy if the confederate expressed a desire to learn how the
toy works? This is an interesting question we might explore in future studies.

In addition to reasoning about the confederate's goal, it is possible that children were reasoning about her knowledge and beliefs. This possibility is particularly salient in the Wrong Action condition. The confederate held the incorrect expectation that the red button played music and expressed frustration upon failure. Thus, a rich interpretation of children's helping behavior (in this case, flipping over the toy) is that they acted on the toy to correct the confederate's false belief. Although considerable evidence from the literature on Theory-of-mind development suggests children are unable to represent others' mistaken beliefs until around age 4 (e.g., Gopnik \& Slaughter, 1991; Wellman, Cross, \& Watson, 2001), some work suggests children the age of our participants might be capable of such belief reasoning, especially in contexts in which they are motivated to help others (Buttelmann et al., 2009; Southgate, Chevallier, \& Csibra, 2010).

Although this is an interesting possibility, it is important to note that it was not necessary to attribute a mistaken belief to the confederate in order to provide appropriate help in our task. Children could have selected the appropriate action by simply attributing ignorance about the functionality of the toys instead of a false belief. Thus, understanding the exact nature of the representation that motivated children's behavior remains an important question for future research.

The absence of age-related trends raises the question of when children might be able to offer help that addresses the cause of others' failed actions. As previously discussed, 16-month-olds are capable of distinguishing between external and internal sources of their own failed actions and will intervene accordingly (Gweon \& Schulz, 2011). Our current work extends these findings in an important direction, suggesting that the causal inference that supports how we respond to our own failed actions may also support how we help others remedy theirs. Thus one might naturally ask: Would 16-montholds also use this reasoning to choose how to help?

In order to succeed in our task, children must (1) have the ability to infer the cause of failure for others' goal-directed actions, (2) have the knowledge to figure out how best to help, and (3) select and execute the more effective action. One possibility is that even though preverbal infants can reason about the cause of agents' failures, they may fail to recruit this reasoning in helping decisions due to constraints in their working memory or executive function. Additionally, prior work suggests that although one-year-olds can provide help when the helping action is constrained, they struggle when the situation is more open-ended (Svetlova, Nichols, \& Brownell, 2010). However, it is possible that we may find similar abilities in infants in a simple paradigm that minimizes such demands. We are currently exploring this possibility.

Finally, our findings have implications for understanding the nature of early instrumental helping. Much of the prior work on the development of helping behavior has focused
on whether and why young children help (e.g., Warneken \& Tomasello, 2006; Cortes Barragan \& Dweck, 2014; Svetlova et al., 2010). Though our experiment instead focuses on the how of early helping, it is still reasonable to ask whether children in our task were really offering help to the confederate or if they responded for another reason, such as a desire to socially interact, or a personal desire to hear the music (see Paulus \& Moore, 2012). Although the selectivity in children's helping responses in our study provides suggestive evidence that children were not simply motivated to interact with the confederate, our experimental design does not allow us to completely disentangle the different possible motivations behind their helping behavior. However, this distinction is not critical for our current purposes. Our main interest in this study was whether children can infer the likely cause of others' failures, and whether such causal reasoning can lead to behaviors that are consequentially effective in helping others achieve their goals.

In fact, children's behaviors in our study suggest that another important motivator for our prosocial behaviors may be our curiosity and desire to understand causal relationships (Gopnik, 1998). When we see someone struggling at the ticket machine, we might want to help not only because we want to help her but also because we want to know why she is failing. Children in our study might have been motivated by similar reasons; such actions would not only help others achieve their goals but also help children themselves learn about the world.

By using what we know, we can better help others. Although deciding how to help in the real-world can be a challenging, open-ended problem, humans can figure out why others fail and the best way to help. Our work suggests that even toddlers are able to solve this problem using their own experience as a guide. While young children are constantly helped and taught by others, the ability to harness this knowledge to figure out how to effectively help others themselves is present early in life.

## Acknowledgments

We thank Fernanda Kramer and Angelina Garron for help with data collection and coding, three anonymous reviewers for their helpful comments, the Palo Alto Junior Museum \& Zoo, families who participated, and our funding sources: John Templeton Foundation Varieties of Understanding grant (HG), NSF Graduate Research Fellowship (SB).

## References

Bridgers, S., Jara-Ettinger, J., \& Gweon, H. (2016). Children consider others' expected costs and rewards when deciding what to teach. In A. Papafragou, D. grodner, D. Mirman, \& J. C. Trueswell (Eds.), Proceedings of the 38th annual conference of the cognitive science society. Austin, TX.
Buttelmann, D., Carpenter, M., \& Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. Cognition, 112(2), 337342.

Cortes Barragan, R., \& Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children's benevolence. Proceedings of the National Academy of Sciences, 111(48), 17071-17074.
Gopnik, A. (1998). Explanation as Orgasm. Minds and Machines, 8(1), 101-118.
Gopnik, A., \& Slaughter, V. (1991). Young children's understanding of changes in their mental states. Child Development, 62, 98-110.
Gweon, H., Chu, V., \& Schulz, L. E. (2014). To give a fish or to teach how to fish? Children weigh costs and benefits in considering what information to transmit. In Proceedings of the 36th annual conference of the cognitive science society (pp. 559-564).
Gweon, H., \& Schulz, L. (2011). 16-Month-Olds Rationally Infer Causes of Failed Actions. Science, 332(6037), 15241524.

Liszkowski, U., Carpenter, M., \& Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. Cognition, 108(3), 732-739.
Martin, A., \& Olson, K. R. (2013). When Kids Know Better: Paternalistic Helping in 3-Year-Old Children. Developmental Psychology.
Paulus, M., \& Moore, C. (2011). Whom to ask for help? Children's developing understanding of other people's action capabilities. Exp. Brain Research, 211, 593-600.
Paulus, M., \& Moore, C. (2012). Producing and understanding prosocial actions in early childhood. Advances in Child Development and Behavior, 42, 275-309.
Sommerville, J., \& Woodward, A. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. Cognition, 95(1), 1-30.
Sommerville, J., Woodward, A., \& Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. Cognition, 96(1), B1-B11.
Southgate, V., Chevallier, C., \& Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. Developmental Science.
Svetlova, M., Nichols, S. R., \& Brownell, C. A. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. Child Dev, 81(6), 1814-1827.
Tomasello, M. (2009). Born (and Bred) to Help. In Why we cooperate (pp. 1-25). MIT Press, Cambridge, MA.
Warneken, F., \& Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. Science, 311(5765), 1301.
Warneken, F., \& Tomasello, M. (2007). Helping and cooperation at 14 Months of age. Infancy, 11(3), 271-294.
Wellman, H., Cross, D., \& Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. Child Development, 72(3), 655-84.

# Watching Non-Corresponding Gestures Helps Learners with High Visuospatial Ability to Learn about Movements with Dynamic Visualizations: An fNIRS Study 

Birgit Brucker (b.brucker@iwm-tuebingen.de) ${ }^{\text {a }}$

Björn de Koning (b.b.dekoning@fsw.eur.nl) ${ }^{\text {b }}$<br>Ann-Christine Ehlis (ann-christine.ehlis@med.uni-tuebingen.de) ${ }^{\text {c }}$<br>David Rosenbaum (david.rosenbaum@med.uni-tuebingen.de) ${ }^{\text {c }}$<br>Peter Gerjets (p.gerjets@iwm-tuebingen.de) ${ }^{\text {a }}$<br>${ }^{\text {a }}$ Leibniz-Institut für Wissensmedien, Schleichstrasse 6, 72076 Tuebingen, Germany<br>${ }^{\mathrm{b}}$ Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands<br>${ }^{c}$ Department of Psychiatry \& Psychotherapy, University Hospital Tuebingen, Calwerstr. 14, 72076 Tuebingen, Germany


#### Abstract

This study investigates whether making and observing (human) gestures facilitates learning about non-human biological movements and whether correspondence between gesture and to-be-learned movement is superior to noncorrespondence. Functional near-infrared spectroscopy was used to address whether gestures activate the human mirrorneuron system (hMNS) and whether this activation mediates the facilitation of learning. During learning, participants viewed the animations of the to-be-learned movements twice. Depending on the condition, the second viewing was supplemented with either a self-gesturing instruction (Y/N) and/or a gesture video (corresponding/non-corresponding/no). Results showed that high-visuospatial-ability learners showed better learning outcomes with non-corresponding gestures, whereas those gestures were detrimental for low-visuospatialability learners. Furthermore, the activation of the inferiorparietal cortex (part of the hMNS) tended to predict better learning outcomes. Unexpectedly, making gestures did not influence learning, but cortical activation differed for learners who self-gestured depending on which gesture they observed. Results and implications are discussed.


Keywords: Learning about movements; dynamic visualizations; human mirror-neuron system; gestures; functional near-infrared spectroscopy.

## Learning from Dynamic Visualizations

In recent years, dynamic visualizations such as animations and videos have become a popular instructional tool to visualize processes and phenomena that are dynamic in nature (e.g., cardiovascular system, lightning formation, fish movements). Obviously, dynamic visualizations are wellsuited for this purpose given that they explicitly depict visuospatial information over time. Nevertheless, research thus far indicates that dynamic visualizations are often not superior to learning from static visualizations (e.g., CastroAlonso et al., 2016; Mayer et al., 2005). It appears that dynamic visualizations are particularly effective for learning about movements when biological movement is involved (Hoffler \& Leutner, 2007) like when learning to tie knots
with the hands (Marcus et al., 2013) or learning to classify fish movements (Brucker et al., 2015). However, so far (1) there is only a handful of studies investigating the instructional potential of dynamic visualizations addressing biological movement and most of them focus on handmanipulative tasks, and (2) it is yet unexplored to what extent learning about biological movements from dynamic visualizations can be enhanced by additional instructional support. These aspects provided the basis for the present study wherein we investigated the value of observing and making gestures for learning to classify fish movement patterns from dynamic visualizations.

## Gestures and Learning

It is by now relatively well-established that making and observing gestures is beneficial for acquiring knowledge about different scientific topics and spatial problem solving (e.g., Chu \& Kita, 2011; Cook \& Goldin-Meadow, 2006). In learning about movements from dynamic visualizations, there is also increasing evidence that showing hands in manual tasks (e.g. origami folding, Marcus et al., 2013) or observing gestures in addition to the learning material improves learning outcomes (Brucker et al., 2015; De Koning \& Tabbers, 2013). It is assumed that this is due to the activation of brain regions (i.e., the human mirrorneuron system [hMNS]; Fogassi \& Ferrari, 2011; Rizzolatti \& Craighero, 2004) involved in the observation, understanding and imitation of other persons' actions. This is in line with the current hypothesis that the stimulation and involvement of this hMNS might be beneficial for learning about complex continuous aspects with dynamic visualizations (Ayres et al., 2009; Van Gog et al., 2009).

Initial evidence for this comes from a study by Brucker et al. (2015) wherein low- and high-visuospatial-ability learners had to learn fish movement patterns from dynamic visualizations whilst observing additional gestures that did or did not correspond to the depicted movements. Results showed better learning outcomes and higher cortical
activation in the inferior-frontal cortex (part of the hMNS) for low-visuospatial-ability learners after watching gestures that corresponded to the to-be-learned fish movements compared to watching non-corresponding gestures. High-visuospatial-ability learners achieved high learning outcomes with both gestures. Unexpectedly, low-visuospatial-ability learners who watched the noncorresponding gestures could also achieve high learning outcomes if they activated their inferior-parietal cortex (also part of the hMNS). These findings provide the first indication that the hMNS is also involved in representing non-human biological or even non-biological movements, if the observer is able to anthropomorphize these movements (cf. De Koning \& Tabbers, 2011). So, drawing on the hMNS by showing learners gestures associated with the learning content seems an effective instructional strategy to improve learning about biological movements from dynamic visualizations.

Based on the notion that learner-generated gestures, as compared to just observing other's gestures, have a more direct and stronger influence on the degree to which the hMNS is activated (e.g., Montgomery, Isenberg, \& Haxby, 2007), asking learners to make gestures related to the movements depicted in a dynamic visualization themselves may be a way to further enhance learning (cf. De Koning \& Tabbers, 2011). Additional advantages of self-performed gestures relate to the manner (e.g., speed, amplitude) in which the gestures are made and the possibility to draw on one's personal experiences (with fish movement) in order to perform the gestures. By embodying the learning content in one's sensory and motor systems based on physical movements (i.e., gestures), the information is coded in a distinct, visuospatial representational format that enriches the way the information is represented, thereby creating a higher-quality mental representation (Paas \& Sweller, 2012). Higher-quality mental representations are associated with better learning (Goldin-Meadow et al., 2001), yielding faster and more accurate performance on learning tests. It is important to note that these anticipated benefits only arise as long as the act of making gestures is not too demanding, complex or distracting (De Koning \& Tabbers, 2013; Skulmowski et al., 2014). Together, by focusing on selfperformed gestures whilst learning about biological movements from dynamic visualizations, we move into a promising but yet unexplored field of research (for an exception see De Koning \& Tabbers, 2013).

## Visuospatial Ability, Gestures, and Learning

As processing continuous changes requires visuospatial ability (cf. Hegarty, 1992), it is likely that learners' visuospatial ability will determine how much they benefit from dynamic visualizations and additional gestures (cf. Hegarty \& Waller, 2005). According to previous research (e.g., Höffler, 2010) learners with higher visuospatial ability outperform learners with lower visuospatial ability during learning with visualizations, and visuospatial ability may moderate the effectiveness of learning with different
instructions and visualization formats. Higher visuospatial ability may compensate for "poor" instructions (i.e., in our case unrelated non-corresponding gestures, cf. Methods section), whereas learners with lower visuospatial ability suffer from such instructions (cf. ability-as-compensator hypothesis; Höffler, 2010). For example, relating this to the Brucker et al. (2015) study, high-visuospatial-ability learners likely possess the skills and resources to see when gestures are in conflict with the depicted content and come up with an own strategy to elaborate on the relevant movements, whereas low-visuospatial-ability learners do not possess these skills and therefore are less able to deal with situations where gestures are in conflict with the dynamic visualizations resulting in lower learning outcomes. Thus, taking into account learners' visuospatial ability is relevant when studying the value of gestures in learning about movements from dynamic visualizations.

## Present Study

This study addresses the question to what extent learning about biological movements from dynamic visualizations can be enhanced by adding information in the form of gestures. We implemented gesture-information in two ways: By making gestures of the learners themselves and by observing gestures displayed on a video. We investigated making gestures (by the learner) by contrasting (1) studying the dynamic visualizations whilst making gestures to (2) studying the visualizations without making gestures. Moreover, we examined observing gestures (that do or do not correspond to the depicted non-human biological movements) by contrasting studying the dynamic visualization whilst (1) observing corresponding versus (2) observing non-corresponding versus (3) not observing additional gestures. Furthermore, functional near-infrared spectroscopy (fNIRS), which is a non-intrusive neurophysiological method to gather data about cortical activation of humans, is used to investigate whether the hMNS is activated during viewing gestures and learning about biological movements from dynamic visualizations. We hypothesize that studying the dynamic visualization with additionally making gestures yields higher learning outcomes than studying without making gestures. Additionally, we hypothesize that studying the dynamic visualizations with additionally observing gestures yields higher learning outcomes than studying without observing gestures. In accordance with Brucker et al. (2015), this pattern is expected to vary as a function of level of gesture correspondence and learner's visuospatial ability: low-visuospatial-ability learners are expected to show higher learning outcomes only on corresponding gestures, whereas high-visuospatial-ability learners are expected to show improved learning outcomes for corresponding and noncorresponding gestures. Furthermore, we hypothesize that the hMNS is more strongly activated with self-performed gestures than with observed gestures, which in turn is more strongly activated than studying without gestures. Moreover, we hypothesize that higher hMNS activation is
associated with higher learning outcomes. This is expected to be particularly true for low-visuospatial learners.

## Methods

## Participants and Design

One hundred and eighteen university students ( $M=24.37$ years, $S D=3.99$; 84 females; 109 right handed) were recruited via an online system (http://www.orsee.org/) and compensated with 10 Euro. They had to learn to discriminate different fish according to their movements based on dynamic visualizations. There were four different to-be-learned movement patterns of fish. The participants saw each movement pattern twice: Firstly, they saw an animation of the specific movement pattern. Secondly, they saw the animation of the specific movement pattern again. But this time depending on the experimental condition, the animation could have been complemented with two additional sources: either a written instruction to self-gesture (making gestures) and/or a video of a person performing gestures with his hands and arms (observing gestures). Depending on this 2-by-3-between subjects design of the study with the two independent factors making gesture and observing gesture there were six conditions in total. Making gesture was varied in two variants: Participants either did or did not get the instruction "Please make your own gestures, that help you to better understand the movement." Observing gesture was varied in three variants: Participants either saw gestures that did correspond or that did not correspond (i.e., were unrelated) to the fish movement patterns or they saw no gesture at all (see Figure 1).

For the observing gestures conditions we used the gestures from Brucker et al. (2015). For the corresponding gestures, an expert regarding fish movements displayed with his hands and arms representations of the respective movements as clearly as possible, whereas for the noncorresponding gestures the (same) expert performed gestures with his hands and arms that were unrelated to the fish movement patterns (i.e., waving, circulating the forearms around each other, drumming, and pointing.

Participants saw the animation of the first fish movement for 30 s . Then a pause of 30 s (black screen) followed before they saw the animation of the first fish movement with its additions (depending on the experimental condition) for 30 s again. Then again a pause of 30 s (black screen) followed before the presentation of the next fish movement started in the same manner. The learners were instructed to relax in the pauses with the intention that the activations of the brain areas of interest were supposed to return to baseline level before the next visualization was displayed.

## Materials

Participants were asked to learn to classify four different fish movement patterns. These fish movement patterns differ in terms of the parts of the body that generate propulsion (i.e., several fins or the body itself) and also in the manner of how these body parts move in the three-
dimensional space (i.e. different paddle-like or wave-like movements). The four different movement patterns were: 1. oscillation of the pectoral fins; 2 . undulation of the body; 3 . undulation of the dorsal and anal fins; and 4. oscillation of the dorsal and anal fins (and undulation of the pectoral fins). During identifying these movement patterns it is very challenging that fish may deploy other movements in addition (e.g., to navigate) and these additional movements can easily be mistaken for movements used for propulsion in another movement pattern. We used the fish animations and gesture videos from Brucker et al. (2015). The movement cycles of the movement patterns were presented in loops in the animations ( 30 s per movement pattern, 25 fps, size: $480 \times 360$ pixels). The gestures were presented in the respective conditions in loops in the videos ( 30 s per movement pattern, 25 frames per s, size: $480 \times 360$ pixels). The presentation of all visualizations was system-controlled.


Figure 1: Six conditions in the 2-by-3-design of the study.

## Measures

Learning Outcomes To assess learning outcomes, we administered a movement pattern classification test comprising 45 dynamic multiple-choice items. These items consisted of underwater videos of real fish performing one of the four to-be-learned movement patterns or a distractor movement pattern. Learners had to identify the body parts relevant for propulsion and their way of moving to choose for each item the kind of movement pattern that was depicted. Each item was visible for 7 s and immediately afterwards participants had 3 s time to choose the correct answer by pressing a corresponding button. Each item was awarded one point for the correct answer ( 0 to max. 45 points). The test items were presented in blocks of 30 s so that 3 items were grouped together. Pauses of 30 s (black screen) followed each block.

Learners' Visuospatial Ability To assess learners' visuospatial ability we used a short version of the paper folding test (PFT, Ekstrom et al., 1976; ten multiple-choice items; total processing time: three minutes). In this task, participants see five options from which they have to choose the correct answer. The stimuli are depictions of papers that are folded stepwise and then were punched in the folded state. The answer options depict unfolded papers with punches being either in the correct or incorrect positions. Each correct answer is worth one point (max. 10 points).

Cortical Activation During viewing the fish animation for the second time in the learning phase, cortical activation was assessed via fNIRS measurements with an ETG-4000 (Hitachi Medical Co.). We used a $2 \times 22$ channel array as probe set that was placed over fronto-temporo-parietal regions and was centered at the T3-T4 and C3-C4 positions (not exactly terminating on these positions because of the fixed interoptode distances) according to the standard locations of the 10-20 system for electrode placement (Jasper, 1958). The fNIRS system measures the change in the product of hemoglobin $(\mathrm{Hb})$ concentration and effective optical path length in human brain tissue. The unit of Hb change is molar concentration ( $\mathrm{mM}=\mathrm{mmol} / \mathrm{l}$ ) multiplied by optical path length (mm). Local increases of Hb are indicators of cortical activity (Obrig \& Villringer, 2003).

## Procedure

Participants were tested individually. After reading a printed overview with information about the procedure of the study, they had to answer the demographics and the PFT. Then, the experimenter placed and adjusted the fNIRS probe set on the scalp of the participants. Subsequently, the computerbased learning materials were presented (learning phase). For each of the four to-be-learned movement patterns, learners were presented with the two presentations of the fish animations (1. fish animation and 2. fish animation plus additional gesture video and/or self-gesturing instruction depending on the experimental condition). Following the learning phase ( 8 min ) learners performed a filler task (about 8 min ), in which they answered some questions on object positions of depicted objects. Subsequently, learners completed the movement classification test ( 15 min ). Participants were instructed to put both their forefingers and both their middle fingers on predefined keys as well as one of their thumbs on the space bar to answer the test items. The predefined keys were labeled on the screen with static screenshots from the learning animations of the four movement patterns and the spacebar was labeled with a grey bar indicating movements that were not part of the learning phase (i.e. distractor items). In total, one experimental session lasted approximately 50 minutes.

## Results

## Learning Outcomes

To analyze learning outcomes, we conducted an ANCOVA (univariate analysis of covariance) with the factors making gesture, observing gesture, and the continuous factor learners' visuospatial ability as a covariate. We inserted all interaction terms in the analysis to investigate the possible interactions. For learning outcomes, results showed no main effect of making gestures ( $F<1$, ns), no main effect of observing gestures $(F(2,106)=1.65, M S E=119.63, p=$ $.20, \eta_{p}^{2}=.03$, ns), but there was a significant main effect for learners' visuospatial ability $(F(1,106)=11.58, M S E=$ 119.63, $p=.001, \eta_{p}^{2}=.10$ ). This effect has to be interpreted in terms of the significant interaction between observing
gestures and learners' visuospatial ability on learning outcomes $\left(F(2,106)=7.93, M S E=119.63, p=.001, \eta_{p}^{2}=\right.$ .13 ; see means and standard errors in Figure 2). There were no other significant interactions or three-way-interactions (all $p \mathrm{~s}>.35$, ns). The significant interaction between observing gestures and learners' visuospatial ability on learning outcomes showed that for participants with high visuospatial ability (defined as one standard deviation above the sample mean) the non-corresponding gesture led to better learning outcomes than the corresponding gesture ( $p$ $=.001$ ) and no gesture $(p=.02)$. For participants with low visuospatial ability (defined as one standard deviation below the sample mean) non-corresponding gestures were worse for learning than no gesture ( $p<.01$ ), whereas there was no significant difference between the corresponding gesture condition and the no gesture condition ( $p=.23$, ns). Thus, the non-corresponding gestures are beneficial for high-, but detrimental for low-visuospatial-ability learners.


Figure 2. Interaction between learners' visuospatial ability and observing gestures on learning outcomes.

## Cortical Activation

To analyze the cortical activation, we defined two regions of interest (ROIs) on the left hemisphere for the hMNS among the respective channels (cf. Rizzolatti \& Craighero, 2004). The two ROIs were the left inferior-frontal cortex (IFC) and the left inferior-parietal cortex (IPC, cf. Figure 3). Cortical activation in these areas was analyzed with two ANCOVAs with the factors making gestures, observing gestures, and learners' visuospatial ability as a covariate. We had to exclude five participants from these analyses because of poor data quality resulting in a total number of 113 participants in these analyses. Even though making gestures did not influence results on learning outcomes, analyses on cortical activation showed tendencies for an interaction between making gestures and observing gestures for both IFC activation $\left(F(2,100)=2.94, M S E=.001, p=.06, \eta_{p}^{2}=\right.$ $.06)$ and IPC activation $(F(2,100)=2.42, M S E=.001, p=$ $.06, \eta_{p}^{2}=.05$ ). There were no other significant main effects or interactions in these analyses (all $p \mathrm{~s}>.104$, ns). Pairwise comparisons revealed that participants observing corresponding gestures showed higher IFC activation if they self-gestured than when they did not self-gesture ( $p=.005$ ). However, participants observing non-corresponding gestures showed higher IPC activation if they self-gestured than when they did not self-gesture $(p=.02)$. This might be an indicator that during watching corresponding gestures the

IFC is more important, whereas during processing noncorresponding gestures the IPC becomes more important at least when the participants were instructed to self-gesture.


Figure 3. Spatial arrangement of the left probe set.

## Effects of Cortical Activation on Learning

To address the question whether higher hMNS activation is directly associated with better learning outcomes, we conducted two ANCOVAs with the factors making gestures, observing gestures, learners' visuospatial ability and cortical activation in terms of IFC activation or IPC activation, respectively. There was a tendency that higher IFC activation lead to higher learning outcomes $(F(1,88)=$ $3.22, M S E=124.85, p=.08, \eta_{p}^{2}=.04$ ). This analysis on IFC activation did also show the main effect for visuospatial ability $\left(F(1,88)=7.58, M S E=124.85, p<.01, \eta_{p}^{2}=.08\right)$ as well as the interaction between observing gesture and visuospatial ability $(F(2,88)=3.93, M S E=124.85, p=.02$, $\eta_{p}^{2}=.08$; both effects reported for learning outcomes, see Figure 2). For IFC activation there were no other significant main effects or interactions (all $p \mathrm{~s}>.27$, ns). The analysis on IPC activation did also show the main effect for visuospatial ability $(F(1,88)=7.18, M S E=128.56, p<.01$, $\left.\eta_{p}^{2}=.08\right)$ and the interaction between observing gesture and visuospatial ability $(F(2,88)=5.18, M S E=128.56, p<.01$, $\eta_{p}^{2}=.11$; both effects reported for learning outcomes, see Figure 2). For IPC activation there were no other significant main effects or interactions (all $p \mathrm{~s}>.189$, ns).

## Discussion

This study investigated whether making and observing additional gestures improves learning about biological movements from dynamic visualizations and to what extent this is related with the cortical activation in areas associated with the hMNS. Regarding learning outcomes, our results indicate that the observation of gestures has different effects for high- and low-visuospatial-ability learners, particularly when dealing with non-corresponding gestures. For high-visuospatial-ability learners, non-corresponding gestures improved learning (even beyond corresponding gestures), whereas for low-visuospatial-ability learners the observation of non-corresponding gestures had detrimental effects on learning. These findings are largely in line with those reported by Brucker et al. (2015) and indicate that particularly when high-visuospatial-ability learners are challenged by a desirable difficulty (cf. Schüler, 2017), in this case by creating a conflict between the visualized fish movements and the (mismatching) gestures, they are stimulated to put more effort in reducing the conflict and come up with a strategy to more elaborately process the relevant movements. This in turn increases the chance that
they properly understand the depicted movement. In contrast, low-visuospatial-ability learners presumably are insufficiently equipped for managing such a situation of conflicting information (e.g., they do not have the resources to identify the mismatch or do not know how to cope with that), and are not able to accurately process the movements and to avoid reduced performance.

In this study, IFC activation tended to predict better learning outcomes. However, compared to the Brucker et al. (2015) study, we did not find the result pattern that IPC activation compensates for missing support of visuospatial ability or non-conflicting gestures. This might be explained by the fact that in the present study participants who neither have visuospatial ability nor non-conflicting gestures at their disposal (i.e. the group of low-visuospatial-ability learners who saw non-corresponding gestures) still could focus on the fish animation. This was possible because in this study the gestures were presented at the same time as the fish, whereas in our prior study the gestures were presented separated in time from the fish animations. However, further research should investigate direct comparisons of sequential and simultaneous presentations of additional gestures.

Another interesting result of this study is that, in contrast to our hypothesis, self-performed gestures did not improve learning outcomes. In line with this, several recent attempts to augment learning about non-human movement (e.g., lightning formation, grammar rules) by instructing learners to make gestures while studying an animation also failed to improve learning performance (e.g. De Koning \& Tabbers, 2013; Post et al., 2013). Collectively, the conclusion from this and other studies is that independent from timing of gestures (during or after learning from dynamic visualizations) and instructional approach (instruct specific ways to perform gestures or let learners decide how to perform gestures) making gestures does not seem to benefit learning from dynamic visualizations involving non-human movement. Importantly, however, making gestures did activate the hMNS. Participants who were instructed to selfgesture activated different parts of the hMNS depending on which gesture they simultaneously observed: with the corresponding gestures there was higher IFC activation, whereas with the non-corresponding gestures there was higher IPC activation. This can be brought in line with our previous findings (Brucker et al., 2015), in which we also found evidence that the IFC plays a role during watching corresponding gestures, whereas the IPC comes into the picture when (conflicting) non-corresponding gestures have to be processed. The IPC is associated with processes of motion analysis and motor imagery, which may both be helpful in the context of identifying the mismatch between the to-be-learned movements and the non-corresponding gestures. However, future research is needed to explore these processes in more detail. Future research should also address one limitation of this study - namely the lack of insight into learners' strategies - by replicating it with think-aloud protocols so that it is possible to discover the strategies learners use when observing and making (non-
corresponding) gestures in learning from dynamic visualizations. Furthermore, it is important to further identify potential neural correlates of (gesture-supported) learning with dynamic visualizations and to further unravel the relations between activation in different parts of the brain and learning outcomes. The present study provides a starting point from which future research endeavors within this emerging field of research can be explored with the goal to incorporate (observing and making) gestures in a way that learning about non-human movements from dynamic visualizations is enhanced. In conclusion, this study shows that observing additional gestures is helpful for learning about movements, but learners need different types of gestures depending on their amount of visuospatial ability. Thus, different types of gestures should be applied: High-visuospatial-ability learners should be challenged with noncorresponding gestures, whereas low-visuospatial-ability learners might be supported with corresponding gestures.

## References

Ayres, P., Marcus, N., Chan, C., \& Qian, N. (2009). Learning hand manipulative tasks: When instructional animations are superior to equivalent static representations. Computers in Human Behavior, 25, 348353.

Brucker, B., Ehlis, A.-C., Häußinger, F.B., Fallgatter, A.J., \& Gerjets, P. (2015). Watching corresponding gestures facilitates learning with animations by activating human mirror-neurons: An fNIRS study. Learning and Instruction, 36, 27-37.
Castro-Alonso, J.C., Ayres, P., \& Paas, F. (2016). Comparing apples and oranges? A critical look at research on learning from statics versus animations. Computers \& Education, 102, 234-243.
Chu, M., \& Kita, S. (2011).The Nature of Gestures' Beneficial Role in Spatial Problem Solving. Journal of Experimental Psychology: General, 140, 102-116.
Cook, S.M., \& Goldin-Meadow, S. (2006). The role of gesture in learning: Do children use their hands to change their minds? Journal of Cognition and Development, 7, 211-232.
De Koning, B.B., \& Tabbers, H.K. (2011). Facilitating understanding of movements in dynamic visualizations: An embodied perspective. Educational Psychology Review, 23, 501-521.
De Koning, B.B., \& Tabbers, H.K. (2013). Gestures in instructional animations: a helping hand to understanding non-human movements? Applied Cognitive Psychology, 27, 683-689.
Ekstrom, R., French, J., Harman, H., \& Dermen, D. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton: Educational Testing Service.
Fogassi, L., \& Ferrari, P.F. (2011). Mirror systems. Wiley Interdisciplinary Reviews: Cognitive Science, 2, 22-38.
Goldin-Meadow, S., Nusbaum, H., Kelly, S.D., \& Wagner, S. (2001). Explaining math: gesturing lightens the load. Psychological Science, 12, 516-22.

Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. Journal of Experimental Psychology: Learning, Memory and Cognition, 18, 1084-1102.
Hegarty, M., \& Waller, D. (2005). Individual differences in spatial ability. In P. Shah, \& A. Miyake (Eds.), Handbook of Visuospatial Thinking. Cambridge University Press.
Höffler, T.N. (2010). Spatial ability: Its influence on learning with visualizations-a meta-analytic review. Educational Psychological Review, 22, 245-269.
Höffler, T.N., \& Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. Learning and Instruction, 17, 722-738.
Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. Electroencephalography and Clinical Neurophysiology, 10, 370-375.
Marcus, N., Cleary, B., Wong, A., \& Ayres, P. (2013). Should hand actions be observed when learning hand motor skills from instructional animations? Computers in Human Behavior, 29, 2172-2178.
Mayer, R. E., Hegarty, M., Mayer, S., \& Campbell, J. (2005). When static media promote active learning: Annotated illustrations versus narrated animations in multimedia instruction. Journal of Experimental Psychology: Applied, 11, 256-265.
Montgomery, K.J., Isenberg, N., \& Haxby, J.V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. Social Cognitive and Affective Neuroscience, 2, 114-122.
Obrig, H., \& Villringer, A. (2003). Beyond the visible Imaging the human brain with light. Journal of Cerebral Blood Flow \& Metabolism, 23, 1-18.
Paas, F., \& Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. Educational Psychology Review, 24, 27-45.
Post, L.S., van Gog, T., Paas, F., \& Zwaan, R. A. (2013). Effects of simultaneously observing and making gestures while studying grammar animations on cognitive load and learning. Computers in Human Behavior, 29, 1450-1455.
Rizzolatti, G., \& Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience, 27, 169-192.
Schüler, A. (2017). Investigating gaze behavior during processing of inconsistent text-picture information: Evidence for text-picture integration. Learning and Instruction, 49, 218-231.
Skulmowski, A., Bunge, A., Kaspar, K., \& Pipa, G. (2014). Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. Frontiers in Behavioral Neuroscience, 8, 426.
Van Gog, T., Paas, F., Marcus, N., Ayres, P., \& Sweller, J. (2009). The mirror-neuron system and observational learning: Implications for the effectiveness of dynamic visualizations. Educational Psychology Review, 21, 2130.

# Learning on Multi-Touch Devices: The influence of the distance between information in pop-ups and the hands of the users 

Birgit Brucker (b.brucker@iwm-tuebingen.de)<br>Lara Scatturin (scalara@gmx.de)<br>Peter Gerjets (p.gerjets@iwm-tuebingen.de)<br>Leibniz-Institut für Wissensmedien, Schleichstrasse 6, 72076 Tuebingen, Germany


#### Abstract

Prior research indicated that information processing is influenced by the proximity of the hands to information: visuospatial processing is fostered near the hands, whereas textual processing might not be affected or even inhibited near the hands. This study investigated how the proximity of the hands to digital information in pop-ups influences learning outcomes on multi-touch devices. Depending on the distance between the information in the pop-ups and the hands of the users there were three conditions: (1) all pop-ups opened near the hands, (2) all pop-ups opened far from the hands, and (3) pop-ups with visuospatial information opened near the hands, whereas pop-ups with textual information opened far from the hands (mixed condition). Results showed better learning outcomes when visuospatial pop-ups are presented near the hands, whereas there was no difference in learning outcomes between near and far presented textual pop-ups. Results and implications for multi-touch designs are discussed.


Keywords: Learning; Information in Pop-Ups; Near-HandAttention; Hand Proximity; Multi-Touch Devices; Design Implementations

## Learning with Multi-Touch Devices

Our hands are the perfect interface of our body to get in contact with the world we live in. They allow us to grasp objects or defend us from potentially harmful things. In our daily life not only real objects, but also digital objects on multi-touch devices become more and more present due to the rapid technological development. One general question that arises, is, how such interaction devices should be designed and implemented to support users during acquiring knowledge with digital objects, such as digital pictures and texts (cf. North et al., 2009). Particularly, the conditions under which interactive multi-touch displays are able to facilitate learning are important. This study investigated how the distance between the hands and fingers of the users to additional digital information in pop-ups influences learning on multi-touch devices. Additionally, this was addressed for different types of information in the pop-ups, namely visuospatial and textual information.

## Near Hand Attention

A few studies investigate the directness of manipulation in terms of hand proximity on multi-touch-tables. For
example, Schmidt, Block, and Gellersen (2009) compared direct input on a multi-touch table display with indirect input where the input is made on the table but the surface on which the action was visible was a separate vertical display. Schmidt et al. (2009) showed that the direct condition led to better results than the indirect version. Moreover, Brucker et al. (2014) and Brucker, Ehrmann, \& Gerjets (2016) showed that direct interaction with visuospatial elements (i.e., pictures of art pieces) was beneficial for learning compared to indirect interaction. This leads to the assumption that hand proximity is particularly beneficial for learning about visuospatial information.

Indeed, prior research on information processing near the hands on computers without multi-touch interaction indicated that the processing of visuospatial information is positively influenced when the hands are near the to-beprocessed stimuli. For example, Reed, Grubb, and Steele (2006) showed that visuospatial processing is enhanced near the hands, because objects that are located near the hands receive higher visual attention than objects that are distant to the hands. There is a large amount of studies pointing in the same direction that visuospatial processing is fostered near the hands (e.g., Abrams et al., 2008; Cosman \& Vecera, 2010; Tseng \& Bridgeman, 2011; Vishton et al., 2007). Tseng and Bridgeman (2011) found evidence that the proximity of the hands lead to deeper and more detailed processing of visual information. Cosman and Vecera (2010) showed facilitated figure-ground-distinctions near the hands. Vishton et al. (2012) showed higher visual precision near the hands (lower Ebbinghaus-illusion).

However, Davoli et al. (2010) showed that not only visuospatial processing is fostered, but that also semantic processing might be impaired near the hands. In their first experiment, participants judged sentences classified by the experimenters as meaningless (e.g., "Tim typed his suitcase to the car" instead of a sentence classified as meaningful, such as "Tim carried his suitcase to the car") more often as meaningful in the near-hand conditions than when the sentences were presented far from the hands (cf. Figure 1). In their second experiment, Davoli et al. (2010) found a reduced stroop-interference in the near-hand condition on a classical stroop-task (naming the color of a word instead of reading it: for example the word "RED" could be written in red [congruent condition] or in green [incongruent
condition]). Thus, participants could better suppress reading the word when their hands were near the stimuli. Davoli et al. (2010) interpreted the results of both studies as impaired semantic processing. According to Graziano and Gross (1994) stimuli that appear outside the border of 20 cm around the hands do not activate bimodal neurons sensitive to both touch and sight and thus, these stimuli can be considered as presented far from the hands. Furthermore, Adam et al. (2012) demonstrated that the proximity of the hands to the stimuli plays also a role while the hands are moving. This is particularly of interest, because during interaction gestures on multi-touch devices the hands have to be moved almost all the time.


Figure 1: Experimental setting in the near-hand (left) and the far-hand condition (right; cf. Abrams et al., 2008).

## Design Implementations of Information in Pop-Ups

During interacting with digital information on mutli-touch devices it is common that touching on the information activates a certain functionality. We address one specific functionality of information depiction on multi-touch devices - namely pop-ups displaying additional information - by addressing how the information processing of the additional information is influenced by the proximity of the hands to the pop-ups. If pop-ups are used on multi-touch devices the question arises where these pop-ups should open on the display in relation to the position of the hands and fingers that activate it. We will shortly introduce three possible distances: near, far, or mixed.

Near distances between the pop-ups and the fingers of the users have the advantage that users easily find the additional information and do not have to run with their eyes over the display to search for it. However, near distances might also lead to coverings because information that is beforehand near to the finger of the user has to be superimposed by the new additional information in the popups. Regarding far distances between the pop-ups and the fingers of the users exactly the opposite advantages and disadvantages occur: Coverings can be prevented, but learners might have to use cognitive resources to find the additional information with their eyes (even though the information might be connected to the position of the finger with a line or an arrow or something similar). This might cause a type of split attention effect (cf. Ayres \& Sweller, 2005) between the users previous focus, where her/his hands and fingers are, and the new additional information in the pop-ups. The third way of presenting additional information in pop-ups - the mixed distances - depends upon the information that is entailed in the pop-up: As abovementioned there is a large amount of studies showing that visuospatial information processing is fostered near the hands. Thus it might be advantageous to depict pop-ups
entailing visuospatial information near the hands (even though coverings might occur). For textual information the result pattern is not that clear: there is one study indicating that it might be better to process textual information further away from the hands (cf. Davoli et al., 2010), but this is not enough evidence to make a strong assumption about the best position for pop-ups entailing textual information.

## Hypotheses

We assumed that information in pop-ups entailing visuospatial information should be learned better when the visuospatial pop-ups are presented near to the hands of the users. For information in pop-ups entailing textual information we did not expect such a difference, although it might be even advantageous if these textual pop-ups are presented far from the hands of the users.

## Methods

## Participants and Design

Fifty-six university students (average age: 24.39 years, SD = 4.58 years; 43 female) from a German university were randomly assigned to one out of three conditions in a between-subjects-design with the factor "pop-up distance" (near versus far versus mixed). Each student received 12 Euro for participating in the study. Art history majors were excluded from the study.

## Materials and Domain

Learning Materials and Multi-Touch Table As instructional domain, art history was chosen. The learning materials consisted of five paintings (cf. Figure 2) from the Herzog Anton Ulrich-Museum in Braunschweig, Germany.


Figure 2: Pictures of the five paintings used in this study.
High quality photographs of the paintings (in the following termed pictures) were displayed on a multi-touch table:

1) "Selbstbildnis" (1547) - Ludger tom Ring d. J. (1522-1584),
2) "Porträt des Reinhard Reiners und seiner Ehefrau Gese" (1569) Ludger tom Ring d. J. (1522-1584),
3) "Frühstücksstillleben" (1642) - Willem Claesz. Heda (1594-168/82),
4) "Die Hochzeit des Peleus und Thetis" (1602) - Joachim Anthonisz. Wtewael (1566-1638), and
5) "Die Heilige Katharina" (around 1620/24) - Bernardo Strozzi (1581/82-1644).

The size of the display of the multi-touch table was $128 \times 135$ cm with a resolution of $1920 \mathrm{px} \times 2160 \mathrm{px}$ via $2-\mathrm{x}-$ Full-HDprojection. We implemented the following interaction
possibilities with the pictures on the multi-touch table. For all five pictures additional information was accessible by touching a " i "-symbol on the bottom right corner. By pressing the " i " the picture turned around and a menu appeared. The menu entailed on the top left a small version of the painting in the middle on the top a short introduction text to the painting, and moreover four thematic index cards with teaser sentences (see Figure 3 for an example).


Figure 3: Menu of "Die Hochzeit von Peleus und Thetis" (1602) with the four thematic index cards.

Each index card could be opened via the " $i$ "-symbol and gave additional information about a certain aspect concerning the painting (e.g., the artist, the story of the painting, details and imagery, space and composition, light and color; see Figure 4 for an example). By touching the " $x$ "-symbol the index card could be closed again to get back to the menu, which could then be itself closed again by touching the respective "x"-symbol on it to go back to the picture. This structure with its three layers (painting - menu - index card) was developed in cooperation with the curators from the museum as well as the computer scientists that implemented the information on the multi-touch table in the context of developing an informal visitor-informationsystem for the museum (Gerjets et al., 2013).


Figure 4: Example of an opened thematic index cars. The arrows mark the pop-ups within the text and the picture.

We decided to stick to three layers at the maximum, however for some aspects there was more additional information that would not have fit the limited space of the respective index card. For this additional information we decided not to open another layer to avoid the user of getting lost (e.g., Conklin, 1987), but instead used pop-ups. These pop-ups appeared by touching highlighted words or
parts of the pictures. The distance in which the pop-ups opened when participants touched them was subject to experimental manipulation (see next subsection).

Hand Proximity of Pop-Up Distance According to the factor "pop-up distance" we compared experimental conditions in which all pop-ups opened near the touching hand/finger of the users (see Figure 5) with conditions in which all pop-ups opened far from the touching hand/finger of the users (at least 25 cm ; this distance was chosen to definitely exceed the peripersonal space of 20 cm around the hands, cf. Graziano \& Gross, 1994; see Figure 6) with conditions in which pop-ups containing visuospatial information opened near, whereas pop-ups containing textual information opened far (at least 25 cm ) away from the touching hand/finger of the users (mixed, see Figure 7).


Figure 5: Near pop-ups condition.


Figure 6: Far pop-ups condition.


Figure 7: Mixed pop-ups condition.

## Measures

The measures comprised a questionnaire on demographics and on participants' familiarity with the domain, a visuospatial ability test, and learning outcome measures.

Demographic Data and Familiarity with the Domain The demographic questionnaire assessed age, gender, body size, need of glasses or contact lenses, major, and study progress. Moreover, this questionnaire assessed participants' familiarity with the domain to determine participants' familiarity with the content domain of this study (i.e., art) and to ensure that all students were novices with respect to this domain. Questions comprised details of the participants' school education in art (e.g., number of courses taken, grades) and their familiarity with and interest in art, for instance, indicated by their visits in museums or galleries within the last year. Participants received points for answers that indicated at least some familiarity with the domain. Depending on the question they could receive only positive points (e.g., 0 to +4 points), whereas for some questions they could also receive negative points (e.g., -2 to +2 points). Depending on these calculations, a participant could receive points within the range of -28 to +40 points.

Learners' Visuospatial Abilities Visuospatial abilities of the participants were assessed with a short version of the paper folding test (PFT, Ekstrom, French, Harman, \& Dermen, 1976). This short version consists of ten multiplechoice items. Participants have to choose the correct answer out of five options for each item. The stimuli are depictions of stepwise folded sheets of paper that were perforated in their folded state. The answer options depict the holes of various unfolded sheets of paper with the holes being either in the correct or incorrect positions. A maximum of three minutes is assigned to work on the items. For each correct answer participants received one point (max. 10 points).

Learning outcome measures Learning outcomes were measured by means of 60 multiple-choice items about the contents entailed in the pop-ups. For each of the 60 implemented pop-ups, there was one multiple-choice item. Most of the items ( $88 \%$ ) had four answer possibilities of which always only one was correct. The remaining items had two answer possibilities. Depending upon the content of the pop-up (visuospatial versus textual), the respective item asked for visuospatial or textual information. Visuospatial items asked for certain details (e.g., depicted objects or parts) from the picture, showed different versions due to color filters or mirroring (see Figure 8 for an example). Textual items asked for specific information that was only given in the texts of the corresponding pop-up (see Figure 9 for an example). For each correct answer participants received one point (max. 60 points).

## Procedure

Participants were tested individually. Subsequently to reading a short overview on the study, they worked on the demographics, the questionnaire on participants' familiarity with the domain, and the PFT. Afterwards, participants were instructed to stand at a fixed position in front of the multitouch table to control the distance from the table. Then, they started with a practicing task on the multi-touch table to get

- depending on the experimental condition - used to manipulating the digital objects and the way they could interact with the depicted information (about four minutes). Subsequently, participants started with the learning phase in which they could - again depending on the experimental condition - freely explore the five pictures of the paintings with the corresponding menus, index cards, and pop-ups (maximal 45 Minutes to explore the five paintings; participants took on average 35.84 Minutes [SD = 7.47]). During the learning phase participants were allowed to zoom and move the digital objects and freely switch between the paintings, their menus, index cards and popups. They were instructed to focus on the information in the pop-ups to ensure that they open preferably all of them to extract the relevant information. Subsequently to the learning phase, the participants answered the 60 multiplechoice items. One session lasted approximately 80 Minutes.


Figure 8: Example of a visuospatial test item.


Figure 9: Example of a textual test item.

## Results

## Learner Prerequisites

To investigate the comparability of the experimental conditions we conducted several analyses of variance (ANOVAs) with the between-subjects-factor "pop-up distance" and the dependent variables participants' familiarity with the domain, age, and visuospatial abilities and a chi-squared-test for gender. There were no differences between the three experimental conditions regarding participants' familiarity with the domain $(F(2,53)=2.00$, $\left.M S E=118.66, p=.15, \eta_{p}^{2}=.07, \mathrm{~ns}\right)$ and participants' age, ( $F<1$, ns). In general, the means indicated that participants' familiarity with the domain was rather low and that it varied a lot across participants (cf. large standard deviations; for means and standard deviations see Table 1). Furthermore, there were no significant associations between the three experimental conditions and participants' gender $\left(\chi^{2}(2,56)\right.$ $=1.58, p=.45$, ns; see Table 1 for the number of females in each condition). Thus, the conditions are comparable with
regard to learners' prerequisites in terms of familiarity with the domain, age and gender.

However, for participants' visuospatial abilities there was a significant difference between the three experimental conditions $\left(F(2,53)=3.76, M S E=5.81, p=.03, \eta_{p}^{2}=.12\right)$. Bonferroni-adjusted post-hoc comparisons showed that only participants in the mixed condition had higher visuospatial abilities than participants in the far condition ( $p=.04$ ). Thus, we calculated an analysis of covariance (ANCOVA) with the between-subjects factor "pop-up distance" and learning outcomes as dependent variable, in which we included visuospatial abilities as a covariate and moreover the interaction term pop-up distance * visuospatial abilities to test whether there was an interaction between pop-up distance and visuospatial abilities. Because this interaction term did not reach statistical significance, we report for reasons of simplicity the ANCOVA with visuospatial abilities included as covariate, but without incorporating the interaction term (pop-up distance * visuospatial abilities).

Table 1: Means and standard deviations (in parentheses) of learner prerequisites and learning outcomes ( $\%$ correct) as a function of hand proximity.

|  | pop-ups <br> near <br> $(n=18)$ | pop-ups <br> far <br> $(n=18)$ | pop-ups <br> mixed <br> $(n=20)$ |
| :--- | :---: | :---: | :---: |
| Domain Familiarity | 6.83 | 2.11 | -0.15 |
| $(-28$ to +40$)$ | $(9.04)$ | $(11.65)$ | $(11.09)$ |
| Visuospatial abilities | 6.61 | 4.89 | 6.90 |
| (PFT 1-10) | $(2.45)$ | $(3.01)$ | $(1.65)$ |
| Age (in years) | 24.06 | 25.00 | 24.15 |
|  | $(5.86)$ | $(4.06)$ | $(3.84)$ |
| Female Participants | 15 | 12 | 16 |
| Learning Outcomes | 58.53 | 52.98 | 59.82 |
| in\% correct | $(10.67)$ | $(8.29)$ | $(9.79)$ |

## Learning Outcomes

An ANCOVA with the between-subjects factor pop-up distance and visuospatial abilities as covariate revealed a significant main effect of pop-up distance for learning outcomes $\left(F(2,52)=3.17, M S E=92.77, p=.05, \eta_{p}^{2}=.11\right.$, achieved power $=0.62$ ), whereas there was no effect of visuospatial abilities on learning $(F(1,52)=1.08$, MSE $=$ 92.77, $p=.30, \eta_{p}^{2}=.02, \mathrm{~ns}$ ). To disentangle the main effect of pop-up distance for the three groups, we calculated contrast analyses in which we compared on the one hand the two conditions with near pictures (pop-ups near and pop-ups mixed) to the condition with far pictures (pop-ups far) and on the other hand the two conditions with far texts (pop-ups far and pop-ups mixed) to the condition with near texts (pop-ups near). These contrast analyses revealed that near pictures lead to better learning outcomes than far pictures $\left(F(1,52)=6.10, M S E=92.77, p=.02, \eta_{p}^{2}=.11\right)$, whereas there were no differences in learning outcomes for near and far texts $(F<1$, ns). This result pattern is in line with our hypothesis (for means and standard deviations see Table 1).

## Discussion

This study addressed how the position of pop-ups in relation to the hands and fingers of the users on multi-touch devices influences learning about the information entailed in the pop-ups. Results showed that learning outcomes are better if pop-ups that contain pictures are presented near the hands, whereas there was no difference for learning outcomes between near-presented and far-presented pop-ups that contain texts. This result pattern is in line with prior findings that visuospatial information is better processed near the hands (e.g., Reed et al., 2006). Moreover it is in line with our hypothesis that visuospatial contents should be presented near the hands of users on multi-touch displays.

Both alternative explanations of split-attention (cf. Ayres \& Sweller, 2005) or possible coverings of important information were not solely valid. A split-attention effect would have favored the near condition because in the far hand conditions the attention of the learners would have been split between the origin of the pop-up where the finger activates it and the location where the pop-up really opens (at least 25 cm distance). A prevent-coverings explanation would have favored the far hand condition, in which the pop-ups open at least 25 cm away from the users fingers, because this implementation prevents coverings of the relevant contents that might have been caused by the opening pop-ups or the fingers, hands or arms of the users.

For textual contents it seems to be indifferent of how near or far these information is presented to the hands of the users. Maybe we were not able to find any differences for near and far texts because we measured recognition of specific details from semantically correct sentences. In contrast to this measure, the only paper that found evidence for differences regarding textual information so far, addressed semantic processing and contrasted meaningful with meaningless sentences (Davoli et al., 2010). However, we did not investigate this more basal type of semantics, but rather a more complex type of semantics. Weidler and Abrams (2014) showed enhanced cognitive control near the hands. Admittedly, they did not address textual processing directly in their studies, but their results indicate that tasks that highly focus on cognitive control should be enhanced near the hands. This result pattern might also explain parts of the Davoli et al. (2010) study, namely the reduced Stroop-inference-effect that can be also explained by higher cognitive control near the hands and not by worse semantic processing. Hence, particularly the complexity of the textual materials might have influenced the information processing. Thus, one might have also assumed that textual information should also be better processed near the hands. However, our results showed neither better nor worse performance for pop-ups containing textual information near the hands. We cannot preclude that both processes - the worse processing of semantic stimuli (cf. Davoli et al., 2010) and the enhanced processing due to higher cognitive control (cf. Weidler \& Abrams, 2014) - might have influenced learning about textual information in this study. Further research is needed to disentangle these concurring explanations.

Another important difference from our study compared to prior research in this field is the direct interaction with the materials. Participants directly manipulated the pop-ups during learning, instead of only holding their hands next to the stimuli as in many prior studies. Thus, in the present study the participants were involved by their active manipulation of the given materials (e.g., freely choosing with which object they want to interact, moving and zooming of objects). Under the evolutionary assumption visuospatial stimuli are potential candidates for manipulation, because grasping of desirable objects or withdrawing the hands in case of dangerous or harmful objects was important in the evolution of human beings. Thus visuospatial stimuli are much more likely to be interacted with than textual stimuli for which no such evolutionary assumption exists. This might also give some hints for the result pattern that we found differences for popups with pictures, but not for pop-ups with texts. Future studies should investigate the importance of the direct manipulation for our result patterns by comparing interactive with non-interactive conditions.

Moreover, in this study the hands and the fingers of the participants, with which they opened the pop-ups, was visible for both the near pop-ups, as well as the far pop-ups, whereas in prior research the hands of the participants were often not visible in the far hand conditions as the hands are for example positioned in the lab of the participants. The visibility of the hands and fingers might have influenced the result pattern even though the far pop-ups opened at least 25 cm away from the finger of the participants.

Further research is needed to replicate our findings. In future studies the exact distance of the pop-ups in the different conditions should be assessed. Moreover, the popup attendance should be gathered as a manipulation check whether all participants really accessed all relevant information by opening all pop-ups. Additionally, eyetracking data would deliver more insights in the question which information the participants really processed during learning. Furthermore, assessing verbal and visual memory skills in addition to visuospatial abilities might contribute to the understanding of the different result patterns for visuospatial and textual contents in the pop-ups.

In sum, the results from this study yield direct implications for designing multi-touch environments: Let pop-ups containing visuospatial information open near the hands, but let pop-ups containing textual information open further away to prevent coverings if the size of the display and the number of users allow for such a far presentation.

## References

Abrams, R. A., Davoli, C. C., Du, F., Knapp, W. H., \& Paull, D. (2008). Altered vision near the hands. Cognition, 107, 1035-1047.
Adam, J., Bovend'Eert, T., van Dooren, F., Fischer, M., \& Pratt, J. (2012). The closer the better: hand proximity dynamically affects letter recognition accuracy. Attention, Perception, \& Psychophysics, 74, 1533-1538.

Ayres, P., \& Sweller, J. (2005). The split-attention-principle in multimedia learning. In R. E. Mayer (Ed.), The Cambridge Handbook of Multimedia Learning (pp. 135146). New York, NY: Cambridge University Press.

Brucker, B., Edelmann, J., Brömme, R., \& Gerjets, P. (2014, August). The proximity of the hands to the objects influences learning on multi-touch devices: Touch pictures, but don't touch words! [Poster] Meeting of the EARLI SIG6 Instructional Design \& SIG7 Learning and Instruction with Computers. Rotterdam, The Netherlands.
Brucker, B., Ehrmann, A., \& Gerjets, P. (2016). Learning on multi-touch devices: Don't cover texts and touch pictures long enough. In J. Désiron, S. Berney, M. Bétrancourt, \& H. Tabbers (Eds.), Proceedings EARLI Special Interest Group Text and Graphics: Learning from Text and Graphics in a World of Diversity (pp. 48-50). Geneva, Switzerland: University of Geneva.
Conklin, J. (1987). Hypertext: An Introduction and Survey. Computer, 20, 17-41.
Cosman, J. D., \& Vecera, S. P. (2010). Attention affects visual perceptual processing near the hand. Psychological Science, 21, 1254-1258.
Davoli, C. C., Du, F., Montana, J., Garverick, S., \& Abrams, R. A. (2010). When meaning matters, look but don't touch: The effects of posture on reading. Memory \& Cognition, 38, 555-562.
Ekstrom, R., French, J., Harman, H., \& Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton: Educational Testing Service.
Gerjets, P., Özbek, O., Blattner, E., Brucker, B., PeifferSiebert, L., \& Edelmann, J. (2013). ARTcard Hypermediale Aufbereitung von Besucherinformationen im Kunstmuseum. Tübingen: IWM.
Graziano, M. S., \& Gross, C. G. (1994). Mapping space with neurons. Current Directions in Psychological Science, 3, 164-167.
North, C., Dwyer, T., Lee, B., Fisher, D., Isenberg, P., Robertson, G., \& Inkpen, K. (2009). Understanding Multi-touch Manipulation for Surface Computing. In Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II (INTERACT '09). Springer-Verlag, Berlin, Heidelberg, 236-249.
Reed, C. L., Grubb, J. D., \& Steele, C. (2006). Hands up: Attentional prioritization of space near the hand. Journal of Experimental Psychology: Human Perception and Performance, 32, 166-177.
Tseng, P., \& Bridgeman, B. (2011). Improved change detection with nearby hands. Experimental Brain Research, 209, 257-269.
Vishton, P. M., Stephens, N. J., Nelson, L. A., Morra, S. E., Brunick, K. L., \& Stevens, J. A. (2007). Planning to reach for an object changes how the reacher perceives it. Psychological Science, 18, 713-719.
Weidler, B. J., \& Abrams, R. A. (2014). Enhanced cognitive control near the hands. Psychonomic Bulletin \& Review, 21, 462-469.

# A Flexible Mapping Scheme for Discrete and Dimensional Emotion Representations: Evidence from Textual Stimuli 

Sven Buechel (sven.buechel@uni-jena.de)

Udo Hahn (udo.hahn@uni-jena.de)<br>Jena University Language \& Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Fürstengraben 27, D-07743 Jena, Germany

http://www.julielab.de


#### Abstract

While research on emotions has become one of the most productive areas at the intersection of cognitive science, artificial intelligence and natural language processing, the diversity and incommensurability of emotion models seriously hampers progress in the field. We here propose kNN regression as a simple, yet effective method for computationally mapping between two major strands of emotion representations, namely dimensional and discrete emotion models. In a series of machine learning experiments on data sets of textual stimuli we gather evidence that this approach reaches a human level of reliability using a relatively small number of data points only.


Keywords: Models of Human Emotion; Representation Mapping; Machine Learning; Natural Language Processing

## Introduction

In the past decades, a multitude of different models have been devised to elucidate the nature of human emotion (Scherer, 2000). A common distinction at the representational level of emotions sets dimensional models apart from discrete or categorical models (Stevenson, Mikels, \& James, 2007).

Dimensional models consider affective states to be best described relative to a small number of independent emotional dimensions (often two or three). Substantial contributions to this line of research are often attributed to Osgood, Suci, and Tannenbaum (1957) as well as Mehrabian and Russell (1974) (Scherer, 2000). Although different labels have been proposed by major proponents of this approach, we here refer to these fundamental dimensions as Valence (the positiveness or negativeness of an emotion), Arousal (a calm-excited scale) and Dominance (the perceived degree of control over a (social) situation) - $V A D$, in short. ${ }^{1}$

Discrete models, on the other hand, often refer to emotions as evolutionary derived response pattern to major environmental events-each with its specific elicitation conditions (Scherer, 2000). Thus, in contrast to dimensional models which tend to focus on the subjective feeling aspect of emotion (and its associated verbal expression) researchers who adhere to the discrete approach rather tend to focus on motor (especially facial) expression and adaptive behavior. Among others, Plutchik (1980), Izard (1994) and Ekman (1992) are most influential for the development of this line of research.

[^38]Although many different sets of such basic emotions have been proposed (typically ranging between 7 and 14 categories), up until now, no consensus has been reached on their exact and complete number (Scherer, 2000). However, most researchers seem to agree on at least five basic categories, namely Joy, Anger, Sadness, Fear, and Disgust.

For dimensional models, a broad variety of stimulus data bases have been developed, predominantly covering lexical stimuli. The Affective Norms for English Words (ANEW) (Bradley \& Lang, 1999a) have been one of the first and probably most important data sets which comprise affective norms for Valence, Arousal and Dominance for 1,034 English words. Complementary lexical affective norms have also been developed for a wide range of other languages, such as German, Spanish or Polish (Võ et al., 2009; Redondo, Fraga, Padrón, \& Comesaña, 2007; Riegel et al., 2015). In addition, larger linguistic units have been considered for emotion assessment moving ratings from lexical items up to sentence and text level (Pinheiro, Dias, Pedrosa, \& Soares, 2017; Bradley \& Lang, 2007), on the one hand, and considering alternative modalities, such as pictures and sounds, on the other hand (Lang, Bradley, \& Cuthbert, 2008; Bradley \& Lang, 1999b). Although these stimulus sets were primarily created for dimensional representations, research activities increasingly covered discrete emotion representations, as well (for all modalities). Consequently, many of the stimuli which have formerly been rated according to affective dimensions only, in the meantime, have also received discrete categorical norm ratings in terms of double encodings (e.g., Stevenson and James (2008), Stevenson et al. (2007) and Libkuman, Otani, Kern, Viger, and Novak (2007); see Table 1 for a list of resources with both dimensional and discrete ratings).

These resources have been highly influential for artificial intelligence (AI) research: Within the broader context of affective computing (Picard, 1997), they have specifically fostered the prediction of affective states from textual stimuli which is-as a subtask of natural language processing (NLP)—most commonly referred to as sentiment analysis (Pang \& Lee, 2008; Liu, 2015; Mohammad, 2016). At the outset, NLP researchers focused on the Valence dimension only, typically trying to assign a piece of text to either a positive or a negative class (Pang, Lee, \& Vaithyanathan, 2002). In the meantime, the interest in more advanced
models of emotions (going beyond positive-negative polarity judgments) has increased considerably. At first, this development was centered around discrete models (Ovesdotter Alm, Roth, \& Sproat, 2005; Strapparava \& Mihalcea, 2007), whereas only very recently the interest in dimensional models rapidly began to rise, as well (Buechel \& Hahn, 2016; Wang, Yu, Lai, \& Zhang, 2016; Sedoc, Preoţiuc-Pietro, \& Ungar, 2017)—a focal change that profoundly benefited from the availability of affective norms developed in psychology labs. Ironically, in NLP, we now face a situation where the enormous interest in analyzing affectively loaded language has led to a proliferation of competing formal representation schemes for affective states whose motivation can be traced in various branches of psychological emotion theory (Valence-only, Valence-Arousal-Dominance, different sets of basic emotions, etc.). Consequently, it has become increasingly difficult to reliably compare the performance of different emotion recognition algorithms (Buechel \& Hahn, 2016).

A possible solution to this dilemma is to elaborate explicit mappings between different representation formats, i.e., to predict the affective norm of a stimulus according to one representation format when the norm is already known in another format (e.g., dimensional and discrete representations; see Figure 1 for a graphical illustration). Not only would this affect formerly incommensurable algorithms but also widely ease the reusability of text collections annotated with different emotional ratings-one of the most important factors for advances given the predominance of training data-dependent supervised machine learning in NLP. In fact, not only computationally focused research would benefit from such mappings but also empirical research in psychology and cognitive science (Stevenson et al., 2007). By that, both the dimensional and the discrete view on emotion would be further integrated so that empirical findings from one view (e.g., regarding priming or memory) could be more directly compared to findings from the other view. Furthermore, existing stimulus sets originally based on one of these approaches could be easily enriched by norms employing other encoding schemes so that researchers could choose from a number of alternative, though mutually translatable emotion representation formats when designing experiments. This outlook becomes even more promising when we take into account the vast number of stimuli sets which bear ratings according to dimensional and discrete formats (see Table 1).

Despite the benefits of transferability, previous work on automatically translating between those formats (in contrast to manual re-annotation) has been relatively rare in the fields of psychology and AI. Stevenson et al. (2007) collected discrete ratings in addition to the dimensional ratings of ANEW. They repeated this effort in a follow-up study for the International Affective Digitized Sounds (IADS) stimulus set (Stevenson \& James, 2008; Bradley \& Lang, 1999b). Performing multiple linear regression between the categories/dimensions of both formats they evaluated the predictive power of the elements of the source representation (dimensions or categories) by the


Figure 1: Affective space spanned by the Valence-ArousalDominance model, together with the position of six basic emotions (as determined by Russell and Mehrabian (1977); figure adapted from Buechel and Hahn (2016)).
statistical significance of their $\beta$-coefficients. They conclude that neither any of the affective dimensions consistently predict (one of the) discrete categories nor can predictions be made the other way round. The findings of Pinheiro et al. (2017) on their own data set of Portuguese sentences, in principle, support this conclusion. The present study differs from these precursors by concentrating on the combined model performance (and not on the contribution of the individual independent variables to $i t$ ).

In contrast with these rather negative interpretations, in AI research, such emotion mappings have already been implemented with quite promising results. Calvo and Kim (2013) presented an algorithm that determines the emotional category of a text based on dimensional word ratings from psychology, using VAD as an interim representation before mapping onto discrete categories. Similarly, Buechel and Hahn (2016) presented a tool for predicting VAD scores from texts which maps their output onto basic emotions using support vector machines. Not only did they achieve highly competitive results regarding their emotion predictions, but they also report on a surprisingly high mapping performance (up to $R^{2}=.944$ when predicting Valence given numerical scores for five basic emotions).

In this contribution, we follow up on this line of research by presenting a series of machine learning experiments that scrutinize the capability of such mapping schemes for textual stimuli. We restrict ourselves to well-known data sets of relatively small size so that the implications of our work can be put into practice without further restrictions (e.g., data limitations in a specific domain). For modeling, we decided to rely on $k$ Nearest Neighbor (kNN) regression because of its simplicity, thus demonstrating that even elementary machine learning methods are sufficient here.

Our experiments fall into three steps. First, we generally demonstrate the feasibility of our approach by examining
the mapping performance between discrete and dimensional emotion formats on two different data sets, an English and a Spanish one. Second, we investigate how well these models generalize over different data sets and languages. In a third step, we examine how well this approach can be ported from psychology to NLP.

## Study A: Mapping within a Stimulus Set

In the first experiment, we examine the capability of machine learning techniques to map dimensional and discrete emotion formats onto each other when training and test data are derived from the same data set.

## Method

Material. We compose two different stimulus sets each receiving dimensional and discrete ratings from individual contributions. The first data set is ANEW which carries norms for Valence, Arousal and Dominance as supplied by Bradley and Lang (1999a); later on Stevenson et al. (2007) added discrete norms for Joy, Anger, Sadness, Fear and Disgust to it. The first half of the second set was originally presented by Redondo et al. (2007) as the Spanish adaptation of ANEW, thus including direct Spanish translations from the original English items. 1,012 of these words overlap with the ones rated by Ferré et al. (2016) according to basic emotions (together forming the second stimulus set). For both the English and the Spanish stimulus set, dimensional ratings were assigned using a 9-point SAM (a set of human-like pictograms displaying different levels of Valence, Arousal and Dominance (Bradley \& Lang, 1994)). For the emotional categories, 5-point scales ranging from not at all to extremely were used. We use mean ratings by all subjects as supplied by the respective authors without performing any further transformation of the data (e.g., re-scaling).
Procedure. We used the R package CARET ${ }^{2}$ to train kNN models in order to map between dimensional and discrete emotion representation schemes. For each dimension/category of the target representation, an individual model was trained given all the dimensions/categories of the source representation as features (e.g., there is one model to predict Anger given Valence, Arousal and Dominance ratings as input). We ran a 10 -fold cross-validation ( $90 \%$ of the data were used for training and hyper-parameter tuning and the remaining $10 \%$ were made available for testing; the process was repeated ten times averaging the results). For the hyper-parameter $k$ a grid search was performed repeating the procedure for each integer in the interval [ 1,100 ]. Consequently, the $k$-values may vary across the individual models. For comparability between different contributions, Pearson's $r$ was used to assess the goodness of the fit.

## Results

Table 2, section "Study A", depicts the results of the crossvalidation (data sets in rows, target dimension/category in

[^39]columns). As can be seen, the results range roughly between $r \approx .73$ up to .97 (both for mapping onto VAD on the English data set). We consider these figures to be surprisingly high, given the small amount of data points we have (from a machine learning point of view) and the elementary model we chose. Henceforth, for comparing correlation coefficients, we use two-tailed $Z$-tests for independent samples (tests for dependent samples are not eligible due to our crossvalidation methodology). We find that mapping from dimensional to discrete ratings performs significantly better on the English data set than mapping the other way round ( $z=2.42$, $p<.05$ ), while the difference in mapping accuracy is not significant regarding the Spanish data $(z=0.74, p \geq .05)$.

Next we compare our model's fit against human reliability. Warriner et al. (2013) replicated the ratings of ANEW finding a correlation of their novel data with the original norms of $r=.953, .759$ and .795 for Valence, Arousal and Dominance, respectively. Thus, on the English data set, computationally mapping discrete emotion norms to dimensional ratings results in a significantly higher correlation with the original values than this replication study regarding Valence and Dominance (Valence: $z=4.1, p<.001$; Dominance: $z=3.1$, $p<.001$ ). For Arousal, the results are not significantly different ( $z=1.72, p \geq .05$ ).

## Study B: Crosslingual Mapping

In our second experiment, we examine in how far the above models generalize over different studies and languages.

## Method

We use the same English and Spanish data sets as for the previous experiments (with both dimensional and discrete ratings). In line with Study A, we train an individual model for each target category/dimension using all dimensions or categories of the source format (discrete or dimensional) as features. The $k$-parameter was chosen according to the highest performance in the 10 -fold cross-validation set-up from the prior experiment. This time, the models were trained on the whole of one data set and then mutually tested on the other one (so that eight models are trained on the English data-one for each dimension/category-and then tested on the Spanish data, and the other way round). Therefore, no cross-validation is necessary. Performance is measured as correlation between predicted and actual values.

## Results

Overall, we find that the models trained on the English data generalize well over the Spanish data and vice versa (see Table 2, Study B). The drops in performance (compared to Study A) regarding the individual dimensions/categories range well below $10 \%$ points. In fact, regarding the average performance of mapping basic emotions onto VAD dimensions, for neither of the two data sets the correlation decreases significantly (comparison relative to the target data; mapping to English: $z=1.46, p \geq .05$, to Spanish: $z=1.6, p \geq .05)$. Regarding the mapping from the dimensional to the discrete

Table 1: Overview of selected stimulus sets bearing ratings according to both dimensional and discrete models.

| Stimuli | Overlap | Dimensional Ratings | Discrete Ratings |
| :--- | ---: | :--- | :--- |
| words | 1,012 | Redondo et al. (2007) | Ferré et al. (2016) |
|  | 1,036 | Bradley and Lang (1999a) | Stevenson et al. (2007) |
| sentences | 1,192 | Buechel and Hahn (2017) | Strapparava and Mihalcea (2007) |
|  | 192 | Pinheiro et al. (2017) | Pinheiro et al. (2017) |
| images | 703 | Lang et al. (2008) | Libkuman et al. (2007) |
| sounds | 111 | Bradley and Lang (1999b) | Stevenson and James (2008) |

Table 2: Results for studies A, B and C in Pearson's relative to the data sets on which the models are trained and tested on (English, Spanish and EmoBank (EmoB.)), and what the input and what the target emotion format for the mapping is (dimensional or discrete). Av: Average over the respective correlation coefficients (dimensional (VAD) or discrete basic emotions (BE)).

| Study | Data | Dimensional $\rightarrow$ Discrete |  |  |  |  |  | Discrete $\rightarrow$ Dimensional |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Joy | Ang. | Sad. | Fear | Dsg. | $\mathbf{A v}_{\text {be }}$ | Val. | Aro. | Dom. | Avvad |
| A | English $\rightarrow$ English | 0.960 | 0.873 | 0.863 | 0.868 | 0.798 | 0.872 | 0.967 | 0.725 | 0.840 | 0.844 |
|  | Spanish $\rightarrow$ Spanish | 0.959 | 0.848 | 0.826 | 0.872 | 0.743 | 0.849 | 0.971 | 0.743 | 0.860 | 0.858 |
| B | English $\rightarrow$ Spanish | 0.948 | 0.791 | 0.807 | 0.829 | 0.698 | 0.815 | 0.966 | 0.740 | 0.808 | 0.838 |
|  | Spanish $\rightarrow$ English | 0.948 | 0.831 | 0.855 | 0.841 | 0.772 | 0.850 | 0.963 | 0.715 | 0.795 | 0.825 |
| C | EmoB. $\rightarrow$ EmoB. | 0.738 | 0.481 | 0.674 | 0.559 | 0.348 | 0.560 | 0.788 | 0.227 | 0.412 | 0.476 |
|  | English $\rightarrow$ EmoB. | 0.643 | 0.411 | 0.637 | 0.518 | 0.301 | 0.502 | 0.682 | 0.156 | 0.360 | 0.400 |
|  | Inter-Rater Reliab. EmoB. | 0.599 | 0.495 | 0.682 | 0.638 | 0.445 | 0.572 | - | - | - | - |

format, losses in performance are significant, however, only by a small margin when mapping from Spanish to English ( $z=1.98, p<.05$; to Spanish: $z=2.52, p<.05$ ). Comparing our models to human reliability (see above), we find that, for Valence, the predictions by the models trained on the Spanish data have still a significantly higher correlation with the original norms by Bradley and Lang (1999a) than the reproduced norms by Warriner et al. (2013) $(z=2.81, p<.001)$. For Arousal, the reproduction yields a significantly higher correlation ( $z=2.19, p<.5$ ) while for Dominance the difference is not significant $(z=0.03, p \geq .05)$.

## Study C: Application to NLP Data Set

In the third experiment, we examine whether the mapping approach from the previous two studies translates to a concrete NLP scenario, given the task to automatically enrich existing emotion data sets with complementary emotion formats.

## Method

We here rely on the recently developed EmoBank data set ${ }^{3}$ (Buechel \& Hahn, 2017) which comprises 10k sentences together with their VAD ratings. To the best of our knowledge, EmoBank is the only NLP resource annotated for multiple emotion formats: A subset of 1,192 sentences (English news headlines) has formerly been annotated for six emotion categories on a $[0,100]$ scale by Strapparava and Mihalcea

[^40](2007). We use this subset, first, to train kNN models in a cross-validation set-up (as in study A), and second, to evaluate the performance of the models previously trained on the English stimulus set on these novel ratings (as in Study B).

## Results

This set-up yields three main results. First, the overall mapping performance drops sharply compared to the former two studies. Comparing the cross-validation performance of our models from the English stimulus set (Study A) with those of the EmoBank data (Table 2, Study C), we find a considerable decrease in correlation of about 35 percentage points (comparing average correlation coefficients for basic emotions and VAD; $z=15.99$ and 16.21, respectively, $p<.001$ ).

In contrast to these mediocre results, the second main finding can be summarized such that our performance does only decrease by a small margin when the models are not trained on EmoBank but on the English stimuli from Study A (comprising words instead of headlines and gathered with a dissimilar methodology; first $v s$. second line of Table 2, Study C). For mapping onto VAD, the drop is still statistically significant ( $z=2.11, p<.05$ ) while for mapping onto BE it is not ( $z=1.82, p \geq .05$ ). This suggests that, although our approach works better for lexical data gathered in psychological settings than for headlines annotated in NLP frameworks, the models still generalize well in the sense that one can apply models trained on the former to the latter without sacrificing a lot of performance.

Even more surprisingly, our third main finding is that our approach still performs very well compared to human reliability (see bottom row of Table 2). Inter-rater reliability is reported by Strapparava and Mihalcea (2007) as the correlation of one rater with the mean judgment of the remaining raters averaged over all raters. Therefore, the output of our models can be cautiously compared against these reliability values. In this setting, we find no significant difference regarding the average over the basic emotions ( $z=0.4, p \geq .05$ ). We carefully interpret this observation to indicate that our output correlates with the aggregated rating of several subjects about as good as an average human does. Thus, consistent with our findings from Study A and B, our approach appears to perform comparably to human subjects and, in fact, even predicts normative Joy ratings significantly better $(z=6.02, p<.001)$. This suggests that the performance drop highlighted as the first main finding might point at different levels of data quality rather than taking this as evidence that our approach might be unsuitable for NLP data (we will get back to this issue in the subsequent discussion section).

## General Discussion

We presented a series of experiments in which we examined the level of performance that can be achieved for mapping emotion ratings onto each other following the dimensional or the discrete representation format for the case of textual stimuli. To make our work more informative in terms of immediate reusability, we limited ourselves to employing relatively small and commonly used data sets, as well as elementary machine learning techniques.

In study A, we took into account two data sets from psychology, an English and a Spanish word stimulus set, each one bearing dimensional and discrete emotion ratings. On both sets, the mapping performance was surprisingly high. When comparing our prediction accuracy to a reassessment study of the English norms with human subjects, we found that our predicted values yielded significantly higher correlation with the original ratings than the novel reproduction regarding two of the three VAD dimension. This astonishing result suggests that given affective ratings in one format, ratings for the complementary emotion format can be computationally induced at a human level of reliability.

Study B goes beyond these considerations by asking how well these models generalize over different data sets with focus on different languages. The observation that the decrease in average mapping performance is only statistically significant in half of the cases suggests that the models generalize well over different (European) languages. However, it must be taken into account that the English and Spanish data sets are direct translations of each other regarding their raw data, possibly boosting the pairwise reusability of the models.

In Study C, we investigated a realistic usage scenario for our approach. Instead of lab data sets typically used in psychology, we here focused on a recently developed corpus of real-world news headlines, again annotated for both emo-
tional dimensions and categories. This set-up yielded three results. First, compared to the former studies, we found a strong decrease in overall mapping performance. Second, the difference between the models directly trained on these data and the ones transfered from Study A were quite small (not even significant for mapping onto BEs). And third, our data suggest that our approach is on par with human annotation performance, despite the overall drop in mapping accuracy.

A possible explanation for this somewhat inconsistent behavior could be that, while the psychological data sets consist of word stimuli with explicit selection criteria, EmoBank comprises "real-world" language data (news headlines instead of individual words). Thus, subjects can interpret these stimuli in a greater number of ways and may also be more strongly affected by biases from, e.g., political orientation or personal biography. In addition, the stimuli from Studies A and $B$ have typically received a greater number of individual ratings which makes their aggregation potentially more reliable (i.e., less noisy in terms of training data).

Besides the above considerations, the results from Study C actually support the flexibility of the approach outlined here. Especially the observation that our models for mapping existing annotations operate about as accurately as a single human rater freshly annotating new raw data suggests that soon we may be able to fully automatically translate affective norms in terms of VAD to basic emotions and vice versa.

In conclusion, the experiments we presented here clearly demonstrate the power as well as the possible impact of our (still rather simple) set-up. The perspective of being reliably able to map back and forth between those popular emotion formats could not only lead to an improved availability of emotionally rated data sets in psychology and NLP. In addition, it may promote the integration of both views on emotion in psychological theory. Despite only presenting evidence from textual stimuli, we suggest that our approach may work for other modalities (and possibly across modalities) as well because no linguistic information was used for the prediction. We will address this conjecture in future work.

## Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful suggestions and comments.

## References

Bradley, M. M., \& Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, 25(1), 49-59.
Bradley, M. M., \& Lang, P. J. (1999a). Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings (Tech. Rep. No. C-1). Gainesville, FL: University of Florida.
Bradley, M. M., \& Lang, P. J. (1999b). International Affective Digitized Sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2). Gainesville, FL: University of Florida.

Bradley, M. M., \& Lang, P. J. (2007). Affective Norms for English Text (ANET): Affective ratings of text and instruction manual (Tech. Rep. No. D-1). Gainesville, FL: University of Florida.
Buechel, S., \& Hahn, U. (2016). Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In ECAI 2016 - Proceedings of the 22nd European Conference on Artificial Intelligence (pp. 1114-1122).
Buechel, S., \& Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In EACL 2017 - Proceedings of the 15th Conference of the European Chapter of the ACL (Vol. 2: Short Papers, pp. 578-585).
Calvo, R. A., \& Kim, S. M. (2013). Emotions in text: Dimensional and categorical models. Computational Intelligence, 29(3), 527-543.
Ekman, P. (1992). An argument for basic emotions. Cognition \& Emotion, 6(3-4), 169-200.
Ferré, P., Guasch, M., Martínez-García, N., Fraga, I., \& Hinojosa, J. A. (2016). Moved by words: Affective ratings for a set of 2,266 Spanish words in five discrete emotion categories. Behavior Research Methods. (Online First Article) doi: 10.3758/s13428-016-0768-3
Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. Psychological Bulletin, 115(2), 288-299.
Lang, P. J., Bradley, M. M., \& Cuthbert, B. N. (2008). International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual (Tech. Rep. No. A-8). Gainesville, FL: University of Florida.
Libkuman, T. M., Otani, H., Kern, R., Viger, S. G., \& Novak, N. (2007). Multidimensional normative ratings for the International Affective Picture System. Behavior Research Methods, 39(2), 326-334.
Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. New York, NY: Cambridge U.P.
Mehrabian, A., \& Russell, J. A. (1974). An approach to environmental psychology. Cambridge, MA: MIT Press.
Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), Emotion Measurement (pp. 201237). Oxford, U.K.: Elsevier.

Osgood, C., Suci, G., \& Tannenbaum, P. (1957). The measurement of meaning. Urbana, IL: Univ. of Illinois Press.
Ovesdotter Alm, E., Roth, D., \& Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In HLT-EMNLP 2005 - Proc. of the Human Language Technology Conference \& Conference on Empirical Methods in Natural Language Processing (pp. 579-586).
Pang, B., \& Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
Pang, B., Lee, L., \& Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques.

In EMNLP 2002 - Proc. of the Conference on Empirical Methods in Natural Language Processing (pp. 79-86).
Picard, R. W. (1997). Affective computing. Cambridge, MA: MIT Press.
Pinheiro, A. P., Dias, M., Pedrosa, J. a., \& Soares, A. P. (2017). Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. Behavior Research Methods, 49(2), 698-716.
Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik \& H. Kellerman (Eds.), Emotion: Theory, research and experience (Vol. 1: Theories of Emotion, pp. 3-33). New York, NY: Academic Press.
Redondo, J., Fraga, I., Padrón, I., \& Comesaña, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). Behavior Research Methods, 39(3), 600-5.
Riegel, M., Wierzba, M., Wypych, M., Żurawski, L., Jednoróg, K., Grabowska, A., \& Marchewka, A. (2015). Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List-reloaded (BAWLR) for Polish. Behavior Research Methods, 47(4), 12221236.

Russell, J. A., \& Mehrabian, A. (1977). Evidence for a threefactor theory of emotions. Journal of Research in Personality, 11(3), 273-294.
Scherer, K. R. (2000). Psychological models of emotion. In J. C. Borod (Ed.), The neuropsychology of emotion (pp. 137-162). Oxford, U.K.: Oxford University Press.
Sedoc, J., Preoţiuc-Pietro, D., \& Ungar, L. H. (2017). Predicting emotional word ratings using distributional representations and signed clustering. In EACL 2017 - Proceedings of the 15th Conference of the European Chapter of the ACL (Vol. 2: Short Papers, pp. 564-571).
Stevenson, R. A., \& James, T. W. (2008). Affective auditory stimuli: Characterization of the International Affective Digitized Sounds (IADS) by discrete emotional categories. Behavior Research Methods, 40(1), 315-321.
Stevenson, R. A., Mikels, J. A., \& James, T. W. (2007). Characterization of the Affective Norms for English Words by discrete emotional categories. Behavior Research Methods, 39(4), 1020-1024.
Strapparava, C., \& Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In SemEval-2007 - Proc. of the 4th Intl. Workshop on Semantic Evaluations (pp. 70-74).
Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., \& Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). Behavior Research Methods, 41(2), 534-538.
Wang, J., Yu, L.-C., Lai, K. R., \& Zhang, X. (2016). Dimensional sentiment analysis using a regional CNN-LSTM model. In ACL 2016 - Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers, pp. 225-230).
Warriner, A., Kuperman, V., \& Brysbært, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45(4), 1191-1207.

# Maintenance of Perceptual Information in Speech Perception 

Wednesday Bushong (wbushong@ur.rochester.edu)<br>Department of Brain and Cognitive Sciences, University of Rochester<br>T. Florian Jaeger (fjaeger@ur.rochester.edu)<br>Department of Brain and Cognitive Sciences, University of Rochester


#### Abstract

Acoustic and contextual cues to linguistic categories (e.g., phonemes or words) tend to be temporally distributed across the speech signal. Optimal cue integration thus requires maintenance of subcategorical information over time. At the same time, previous work suggests that finite sensory memory or processing capacity strongly limits how much subcategorical information can be maintained (or for how long). We argue that previous work might have over-interpreted the role of these limitations. In two perception experiments, we find no limit in the ability to maintain subcategorical information. We also find that maintenance seems to be the default, neither limited to perceptually particularly ambiguous signals, nor a learned strategy specific to our experiment. In contrast, listeners' decision for how long to delay categorization, we find, is a function of perceptual ambiguity. It is therefore crucial to distinguish between in-principle abilities (even when they reflect default processing), and decisions made within the bounds of those abilities.


Keywords: linguistics; cognitive science; speech recognition; language comprehension

## Introduction

One of the most fundamental problems of auditory processing is the transient nature of the acoustic signal; the systems underlying speech perception receive large amounts of information every second. The bounds of sensory memories thus create a pressure to incrementally infer abstract linguistic categories (e.g., phonemes and words) from the auditory signal before that information becomes unavailable.

However, much of the information relevant to inferring a particular part of the auditory signal, for example a segment (phoneme), is not contained on the segment itself. For example, one of the main cues to coda stop voicing is duration of the previous vowel (Klatt, 1976). Thus, in order to successfully resolve the voicing of a coda stop, listeners must maintain information about the previous vowel and integrate it with the evidence they receive later. This is opposed to a scheme where the listener removes information about the previous vowel and only maintains some abstract categorical representation that does not include duration information.

Previous work suggests that listeners can indeed maintain and use subcategorical information at least at short timescales. In a classic study, Ganong (1980) found that lexical constraints can influence the perception of a word-initial sound: sounds varying on the $/ \mathrm{d} /-/ \mathrm{t} /$ continuum are perceived to be more /d/-like when presented before "ash" (dash is a word while tash is not). More evidence that subcategorical information is maintained within a word comes from eyetracking studies: McMurray, Tanenhaus, and Aslin (2002) found that listeners looked to competitor items like "bear"
and "pear" gradiently according to voice onset time (VOT), the cue that distinguishes $/ \mathrm{b} /$ and $/ \mathrm{p} /$, suggesting that subphonemic information is maintained and used in higher-level processes (for a review of related work, see Dahan, 2010).

However, there are other possible sources of information that follow a target segment or word that occur much later downstream. For example, persevatory co-articulation might spread information over following syllables (Magen, 1997). The identity of later segments might also contain information about earlier segements because of phonotactic dependencies within and across syllables. Even context beyond word boundaries regularly contains information that can help to resolve uncertainty about the input.

A small literature has investigated the extent to which listeners can maintain subcategorical information at longer distances across the word boundary. In a classic study, Connine, Blasko, and Hall (1991) tested whether listeners could maintain subcategorical information about a segment 3 syllables or 6-8 syllables downstream. Participants listened to sentences like "When the ?ent in the fender was well camouflaged, we sold the car." and judged whether the word they heard was tent or dent. The ? represents a sound that varied along VOT, the primary cue distinguishing between $/ \mathrm{t} /$ and $/ \mathrm{d} /$. In this example, the later word fender semantically biases interpretation of the target word to be dent. If listeners can maintain information about the identity of the ?-segment, they should integrate the biasing context into their decisions. Connine and colleagues reported two important findings, both of which have recently been revisited.

First, participants maintained subcategorical information about the ?-segment for 3 syllables: responses reflected both the specific VOT of the segment and the contextual bias. After 6-8 syllables responses reflected only VOT, but not biasing context. This finding is often interpreted to demonstrate the limits of subcategorical information maintenance. However, participants were allowed to respond at any point during the sentence; in fact, in the 6-8 syllables condition, participants responded before even hearing biasing context $84 \%$ of the time. This leaves open whether participants could not maintain subcategorical information for longer periods of time, or chose to respond early for other reasons.

Secondly, Connine and colleagues report that the context effect was only reliably present at ambiguous VOTs: sounds that were perceptually unambiguously /t/ or /d/ were not integrated with later context. This has been taken to mean that even information maintenance of up to 3 syllables is limited to the special case of perceptually highly ambiguous per-
cepts. This second conclusion, too, however, has to be interpreted with caution. Connine and colleagues measured the context effect in proportion of $/ \mathrm{t} / \mathrm{vs}$. /d/ responses. This is problematic (see also Jaeger, 2008): a context effect that is identical across all VOTs when measured in log-odds-i.e., equally large for perceptually clear and perceptually ambiguous VOTs-will results in smaller or insignificant context effects for perceptually clear VOTs when measured in proportions. Crucially, there are a priori reasons to believe that the effect should be constant in log-odds (Bicknell, Jaeger, \& Tanenhaus, 2016). The analysis conducted by Connine and colleagues thus leaves open whether subcategorical information maintenance is limited to special cases.

A recent study, Bicknell et al. (2016), revisited both of these problems. Bicknell and colleagues replicated Connine et al. (1991) with one minor change to procedure. Participants were required to wait until the end of the sentence to respond, ensuring that they heard the biasing context. Unlike in the original study, Bicknell et al. (2016) analyzed the log-odds of responding /t/ vs. /d/ and found that listeners maintained subcategorical information for both the 3 and 6-8 syllable conditions (see also Szostak \& Pitt, 2013 for similar results in a different phonetic contrast). This suggests that there may be an important distinction between listeners' ability to maintain subcategorical information and when listeners decide to respond.

The idea of a distinction between in-principle abilities and the decision process motivates the present experiments. Our first goal is to replicate the between-experiment comparison across Bicknell et al. (2016) and Connine et al. (1991) within the same paradigm. Anticipating our result, we indeed replicate the contrast, showing that it is important to distinguish between the ability to maintain information and the decision to provide a categorization response. Given that listeners sometimes choose to make a response before receiving additional semantic information, we ask whether subcategorical information maintenance is a default strategy employed by listeners or is specific to experience in our task. Finally, we ask what influences the decision process by investigating the role of perceptual ambiguity on when participants choose to make a response.

In order to answer these questions, we conducted a webbased experiment that closely followed the paradigm of Connine et al. (1991) and Bicknell et al. (2016). Betweenparticipants we manipulated only one aspect of the procedure, holding everything else constant. In the forced-response group of participants, they were required to wait until the end of the sentence before making a response. In the freeresponse group, they could make a response whenever they wanted during the sentence. The forced-response group gives us insight into the ability to maintain subcategorical information. The free-response group allows us to ask what drives listeners' decisions to categorize.

| Context | Distance | Sentence |
| :--- | :--- | :--- |
| Tent-biasing | Near (3 syllables) | When the $[\mathrm{t} / \mathrm{d}]$ ent in the <br> forest was .. |
| Dent-biasing | Near (3 syllables) | When the $[\mathrm{t} / \mathrm{d}]$ ent in the <br> fender was ... |
| Tent-biasing | Far (6-8 syllables) | When the $[\mathrm{t} / \mathrm{d}]$ fent was <br> noticed in the forest, ... |
| Dent-biasing | Far (6-8 syllables) | When the $[\mathrm{t} / \mathrm{d}]$ ent was <br> noticed in the fender, ... |

Table 1: Example stimuli from the experiment in each biasing context and distance condition.

## Experiment

## Participants

We recruited a total of 96 participants from Amazon Mechanical Turk ( 48 for the forced-response group, 48 for the freeresponse group). Participants were awarded $\$ 3.00$ for their participation in a 30 -minute experiment.

## Materials

Materials were identical across the two participant groups and were modeled on Connine et al. (1991). Table 1 shows example sentences in each context and distance condition. We manipulated context (tent-biasing vs. dent-biasing), distance (near, 3 syllables vs. far, 6-8 syllables), and VOT (10, 40, 50, 60,70 , and 85 ms ). We chose our range of VOTs based on simulation-based power analyses so as to maximize statistical power to assess the size of the context effect across the VOT continuum, while also ensuring that there were a range of perceptually ambiguous and unambiguous sounds (based on the VOT distributions of our recording speaker). Seven different sentence frames were constructed. Each participant heard each sentence frame in each of the context, distance, and VOT condition combinations, resulting in a total of 168 sentences in the experiment.

## Procedure

Participants were instructed to listen to the sentence and report whether they heard tent or dent. In the forced-response group, participants were instructed to wait until the end of the sentence to make a response. In the free-response group, participants were instructed that they could respond whenever they wish during the sentence after hearing the critical word.

## Data Exclusions

We excluded participants who showed no main effect of VOT on their responses from further analysis. That is, these were participants who did not increase their /t/ responses as VOT increased, suggesting that they had faulty audio equipment, did not understand the task, or were otherwise not paying attention. In the forced-response group, this resulted in the removal of nine participants (18.75\%) from analysis. In the free-response group, eleven participants ( $22.92 \%$ ) were removed. These exclusions hold across all analyses below.


Figure 1: Proportion /t/ responses by biasing context condition for both groups of participants. Error bars are 95\% confidence intervals over subject means. See text for discussion of subset vs. all trials.


Figure 2: Size of context effect in log-odds space at each VOT for each group as estimated by our simple effects mixed models. Error bars are $95 \%$ confidence intervals.

## Analysis 1: Limits of Subcategorical Information Maintenance

## Analysis

Analyses 1 and 2 are based on the same mixed-effects regression, analyzing the proportion of $/ \mathrm{t} /$ responses as a function of VOT (a continuous variable), context, distance, trial number, and their interactions. We included random slopes for context and distance by participants and items (due to data sparsity and the consistency of the VOT effect across participants and items, we did not use the maximal random effects structure; see Bates, Kliegl, Vasishth, \& Baayen, 2015). Different predictors in this model answer different questions. In Analysis 1, we focus on the overall effect of context and its interactions with VOT and distance.

In addition to the model described above, in Analysis 1 we also fit a second model which assessed the relative magnitudes of the effect of context at each VOT while removing the potentially problematic assumption that VOTs are related linearly to the log-odds of /t/responses. This was achieved by recoding the model so as to assess the simple effects of context at each level of VOT. This analysis does thus not a priori assume any specific relation between VOTs and /t/ responses.

For each model of the free-response group, we present two analyses. First, we analyzed only the trials on which participants responded at least 200 ms after offset of the biasing context. This allows a direct comparison with the forced-
response group, where participants always responded after hearing the biasing context by design. Second, we also conducted the same analyses using all of the data in the freeresponse group in case these data are more comparable.

## Results

Figures 1 and 2 summarize the context effect results of both groups. We found a main effect of VOT on /t/ responses (forced-response: $\hat{\beta}=0.18, p<0.001$, free-response subset: $\hat{\beta}=0.13, p<0.001$, free-response all trials: $\hat{\beta}=0.16, p<$ 0.001 ). We also found a context main effect (forced-response: $\hat{\beta}=1.11, p<0.001$, free-response subset: $\hat{\beta}=2.08, p<$ 0.001 , free-response all trials: $\hat{\beta}=1.62, p<0.001$ ). In the forced-response group and the subset of trials in the freeresponse group where participants responded after biasing context, there was no interaction between context and distance (forced-response: $\hat{\beta}=-0.09, p=0.57$, free-response: $\hat{\beta}=0.18, p=0.38$ ). When we analyzed all trials of the freeresponse group, there was a context x distance interaction such that the context effect was smaller in the far condition ( $\hat{\beta}=-0.39, p=0.02$ ).

A simple effects analysis revealed that the effect of context was significantly positive at $50 \mathrm{~ms}, 70 \mathrm{~ms}$, and 85 ms VOT in both groups ( $\hat{\beta} s=0.58-1.95, p \mathrm{~s}<0.05$ ). In the freeresponse group, the context effect was also significant at all other VOTs (subset: $\hat{\beta} s=0.93-2.23, p s<0.01$, all trials: $\hat{\beta} s=0.61-1.97, p s<0.05)$. In the forced-response


Figure 3: Interaction between context effect over trial for both groups of participants.
group, the context effect was marginal at 40 ms and 60 ms VOT ( $\hat{\beta} s=0.87,0.47, p s=0.06,0.08$ ), and not significant at $10 \mathrm{~ms} \operatorname{VOT}(\hat{\beta}=-0.42, p=0.5)$.

## Discussion

Replicating both Connine et al. (1991) and Bicknell et al. (2016), we found that listeners have the ability to maintain subcategorical information well beyond the word boundary. When forced to wait, participants' responses reflected both the VOT and the contextual bias even at the longest delay tested (replicating Bicknell et al., 2016; Szostak \& Pitt, 2013). Interestingly, the effect of context seemed more or less constant across the entire range of VOTs tested in both groups. This is exactly as expected by an ideal observer that integrates the perceptual signal with context (Bicknell, Bushong, Tanenhaus, \& Jaeger, in preparation). It also suggests that listeners do not necessarily limit the maintenance of subcategorical information to perceptual inputs that are perfectly ambiguous. Instead, it seems listeners maintain subcategorical information even when the perceptual input is already rather unambiguous ${ }^{1}$.

When participants were free to choose when to respond, however, we found an interaction between context and distance, such that the context effect was smaller at longer timescales. This would suggest that participants were deciding to respond before hearing biasing context: indeed, the free-response group responded before biasing context on $32 \%$ of far trials and $0.5 \%$ of near trials (a point we return to in Analysis 3).

Analysis 1 leaves open whether this tendency to maintain subcategorical information is a strategy participants adopt specifically for this experiment, rather than reflecting a more general property of speech perception. Analysis 2 begins to address this question by investigating the context effect across trials.

[^41]
## Analysis 2: Subcategorial Information Maintenance: Experimental Artifact or Default Behavior?

We analyze changes in the effect of context over the course of the experiment in both groups. If we observe an effect of context from the very beginning of the experiment, this suggests that listeners maintain subcategorical information by default. On the other hand, if we observe no context effect until later in the experiment, this suggests that listeners have learned to maintain subcategorical information.

## Analysis

We used the same logistic regression model from Analysis 1 and focus on the effects of context, trial, and their interaction. Trial was coded so that the coefficient estimate for context reflects the context effect at the very first trial (by subtracting 1).

## Results

Figure 3 shows the context effect over trials in both groups of participants. The context effect was significant from the very first trial of the experiment (forced-response: $\hat{\beta}=$ $1.11, p<0.001$, free-response subset: $\hat{\beta}=2.08, p<0.001$, free-response all trials: $\hat{\beta}=1.62, p<0.001$ ). We found a significant negative interaction between context and trial for both groups of participants (forced-response: $\hat{\beta}=-0.004, p<$ 0.001 , free-response subset: $\hat{\beta}=-0.009, p<0.001$, freereponse all trials: $\hat{\beta}=-0.004, p<0.001$ ).

## Discussion

Particpants in both experiments exhibited clear context effects right from the beginning of the experiment. This suggests that participants have the ability to maintain subcategorical information without requiring extensive exposure to a particular task. We also found a negative context by trial interaction, such that the context effect got smaller over the course of the experiment. This could mean that participants maintain subcategorical information to a lesser extent as time goes on (e.g., because of fatigue or boredom with the task). Alternatively, participants may still be maintaining subcategorical information but may rely less on context during their
decision making process, and use VOT more (e.g., because participants become more certain of the talker-specific VOT distribution, cf. Kleinschmidt \& Jaeger, 2015).

## Analysis 3: Strength of Perceptual Evidence and Decision to Categorize

Although maintenance of subcategorical information seems to be a default strategy among participants, the context by distance interaction in the free-response group in Analysis 1 suggests that participants did not necessarily wait for biasing context to make their responses.

This raises questions about what determines when listeners provide a categorization response. If listeners have enough perceptual evidence to confidently make a categorization, they may tend to respond early rather than waiting for the biasing context that provides additional information about the identity of the segment (note that this leaves open whether listeners maintain subcategorical information beyond this point; we return to this below). To answer this question, we analyze when participants in the free-response group made responses, and whether this was dependent on the perceptual ambiguity of the stimulus.

## Analysis

We used mixed-effects logistic regression to analyze the proportion of responses before biasing context as a function of perceptual ambiguity and distance. For each trial, we coded whether the participant responded before or after having heard biasing context (defined as 200 ms after biasing word offset to account for motor planning). To estimate (subjective) perceptual ambiguity, we compute the distance (in probability space) of each VOT from the maximally unambiguous point based on average response probabilities ${ }^{2}$. If strength of perceptual evidence affects when listeners make a decision before obtaining more information (provided by the biasing context), we should see more responses before biasing context for less ambiguous stimuli.

## Results

Figure 4 shows proportion of responses before biasing context by perceptual ambiguity of the stimulus. We found a significant effect of ambiguity ( $\hat{\beta}=-4.06, p=0.006$ ), such that participants were less likely to respond before biasing context when the perceptual stimulus was more ambiguous. We also found a main effect of distance ( $\hat{\beta}=6.93, p<0.001$ ) such that participants were more likely to respond before biasing context when it occurred 6-8 syllables away from the target word than when it occurred 3 syllables away. We additionally found a main effect of VOT such that participants were less likely to respond before biasing context as VOTs became longer ( $\hat{\beta}=-0.007, p<0.001$ ). There were no other main effects or interactions.

[^42]

Figure 4: Proportion of responses before biasing context by perceptual ambiguity. Error bars are $95 \%$ confidence intervals.

## Discussion

We found that participants were more likely to respond before hearing biasing context when the perceptual signal was less ambiguous, and when biasing context appeared farther away from the target word. We also found a main effect of distance: participants were more likely to respond before biasing context when it occurred farther away from the target word. These results suggest that while listeners have the ability to maintain subcategorical information for unambiguous stimuli over long distances, when given a choice listeners decide to respond earlier when they have stronger perceptual evidence for categorization.

## General Discussion

Together, our results suggest that in principle, listeners can maintain subcategorical information well beyond word boundaries. Listeners seem to do so by default, and both for ambiguous and unambiguous percepts. This suggests that the limits of listeners ability to maintain subcategorical information are less strict than previously assumed (Connine et al., 1991; Christiansen \& Chater, 2016). At the same time, listeners do not wait arbitrarily long for additional informative context. When given the opportunity, listeners responded on $16 \%$ of all trials before additional context could aid recognition. Critically, listeners' decisions to respond early were not arbitrary, but rather systematically conditioned on the ambiguity of the perceptual input: listeners were more likely to respond before biasing context when the perceptual signal was less ambiguous. This strategy seems to vary little across participants.

Three questions stand out to us as requiring further attention. First, importantly, little is known about what kind of information is being maintained. It is possible that listeners retain a rich representation of the original percept, some more abstract representation of their certainty in the identity of the segment, or something in between.

Second, it is unclear what becomes of these representations after listeners make a perceptual decision. It could be the case that the maintenance process and decision-making process are dependent on or independent of each other. The large literature on exemplar-based approach to speech perception suggests that exemplars are stored and used later in speech perception (Hay \& Drager, 2010; Strand \& Johnson, 1996; Goldinger, 1997). The apparent storage of this low-level information in long-term memory is puzzling if there are strict limitations on the amount of information that can be maintained during speech perception-a paradox that has, to the best of our knowledge, received surprisingly little attention.

Third, we found evidence that the maintenance of subcategorical information in the present experiments does not seem to be learned over time in a task-specific manner. It is, however, an open question whether listeners can flexibly adapt the degree to which (or duration for which) they maintain subcategorical information, depending on their goals or the structure of the current task. Such flexibility would suggest that listeners' decisions about at which point to categorize input might more often be constrained by the goal to quickly infer the meaning-bearing message, rather than being constrained by strong limits of perceptual memory. For example, it is possible that the limits (or lack thereof) of maintenance observed in experiments like ours (and a large body of previous work; for review, see Dahan, 2010) reflect participants' beliefs based on previous experience about the expected utility of delaying categorization. In that case, listeners might adapt these beliefs after exposure to stimuli that contain or do not contain helpful contextual information.

## Acknowledgments

This work was partially funded by NSF NRT \#1449828 (graduate stipend to W.B.) and NSF IIS-1150028 and NIHCD R01 HD075797 (to T.F.J.). The views expressed here do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

## References

Bates, D., Kliegl, R., Vasishth, S., \& Baayen, H. (2015). Parsimonious mixed models. arXiv preprint arXiv:1506.04967.
Bicknell, K., Bushong, W., Tanenhaus, M. K., \& Jaeger, T. F. (in preparation). Listeners can maintain and rationally update uncertainty about prior words.
Bicknell, K., Jaeger, T. F., \& Tanenhaus, M. K. (2016). Now or ... later: Perceptual data is not immediately forgotten during language processing. Behavioral and Brain Sciences, 39, 23-24.
Christiansen, M. H., \& Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. Be havioral and Brain Sciences, 39, e62.
Connine, C. M., Blasko, D. G., \& Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. Journal of Memory and Language, 30(2), 234-250.

Dahan, D. (2010). The time course of interpretation in speech comprehension. Current Directions in Psychological Science, 19(2), 121-126.
Ganong, W. F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6(1), 110.
Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. Talker variability in speech processing, 33-66.
Hay, J., \& Drager, K. (2010). Stuffed toys and speech perception. Linguistics, 48(4), 865-892.
Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434446.

Klatt, D. H. (1976). Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. The Journal of the Acoustical Society of America, 59(5), 12081221.

Kleinschmidt, D. F., \& Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. Psychological review, 122(2), 148.
Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in english. Journal of Phonetics, 25(2), 187205.

McMurray, B., Tanenhaus, M. K., \& Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. Cognition, 86(2), B33-B42.
Strand, E. A., \& Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In Konvens (pp. 14-26).
Szostak, C. M., \& Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. Attention, Perception, \& Psychophysics, 75(7), 1533-1546.

# Task-oriented Bayesian inference in interval timing: People use their prior reproduction experience to calibrate time reproduction 

Zhenguang Cai (Zhenguang.cai@uea.ac.uk)<br>School of Psychology, University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR4 7TJ, UK<br>Ruiming Wang (wangrm@scnu.edu.cn)<br>School of Psychology, South China Normal University, Tianhe District, Guangzhou, 510631, China<br>Huihui Liu (1425357597@qq.com)<br>School of Psychology, South China Normal University, Tianhe District, Guangzhou, 510631, China<br>Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk)<br>Department of Experimental Psychology, University College London, London, WC1H 0AP, UK


#### Abstract

The estimation of duration has been shown to follow Bayesian inference, where people use their prior belief to calibrate the estimation. This explains timing biases such as the range bias where a duration is reproduced as longer when previously encountered durations were longer than shorter. However, it is unclear whether prior belief is based on previously perceived or reproduced durations. In 4 experiments, we show that the range bias occurs between short and long reproduction ranges but not between short and long perception ranges. Further analyses also show that the prior is updated by the most recent reproduced (but not perceived) duration. Together these results support a task-oriented Bayesian inference account of time reproduction, where people use the perceived duration and their past reproduction experience to make an inference about how much time to reproduce.


Keywords: time perception; Bayesian inference; memory; psychophysics

## Introduction

The mind is good at estimating quantitative dimensions of the physical world: we are able to estimate how much time has elapsed, how much distance has been traveled, how large an area is, etc. Indeed, our superb capacity to quantify things enables us to better adapt to the environment. However, these quantitative intuitions are not without errors. Systematic biases in human magnitude estimation have been identified (for reviews see Poulton, 1979, and Petzschner, Glasauer, \& Stephan, 2015). Of these, the most robust is probably the bias of central tendency (Hollingworth, 1910), a phenomenon which has also been known by a variety of other names (e.g., contraction, regression effect, regression toward the mean). Central tendency refers to the observation that people tend to make estimates closer to the mean of the magnitudes to be estimated, leading to the (relative) overestimation of lower magnitudes in the stimulus set and underestimation of higher magnitudes. A central tendency bias has been observed in the estimation of distance (Jou et al., 2004; Radvansky et al., 1995), brightness (Fotios \& Cheal, 2007), weight (Jones \& Hunter, 1982), and loudness (Algom \& Marks, 1990), to mention a few. In particular, the central tendency bias has been most often observed in the estimation of time intervals, with the tendency for people to relatively overestimate shorter durations and underestimate longer durations (Bausenhart, Dyjas, \& Ulrich 2014; Gu \& Meck, 2011;

Jazayeri \& Shadlen, 2010; Lejeune \& Wearden, 2009; Moon \& Anderson 2013). More interestingly, as a result of the central tendency bias, a stimulus duration tends to be reproduced as longer if it occurs as a member of a longer range (e.g., 1000 ms in the range of $847-1200 \mathrm{~ms}$ ) than as a member of a shorter range ( 1000 ms in the range of 671 1023 ms ) (Jazayeri \& Shadlen, 2010), a phenomenon which we refer to as the range bias.

Jazayeri and Shadlen (2010) argue that the central tendency and the range bias occur because time estimation follows Bayesian inference: as memories of durations (indeed magnitudes in general) are inherently noisy (Gallistel \& Gelman, 2000), people resort to their prior belief about how likely a duration is in order to calibrate their estimate of the magnitude of a perceived duration. In Bayesian inference, a posterior belief is the product of the "likelihood" (reflecting the variability of perceptions of a given duration) and the prior, and this posterior (the estimated duration) is necessarily pulled toward the mean of the prior distribution (e.g., the midpoint of experienced durations), hence the central tendency and the range bias. Such a Bayesian inference account of timing has been endorsed in many subsequent related research on timing behaviour, e.g., animal timing ( Li \& Dudman, 2014), time prediction (Griffiths \& Tenenbaum, 2011; Di Luca \& Rhodes, 2016) and delay discounting (McGuire \& Kable, 2012) (see Shi, Church \& Meck, 2014, for a review). Indeed, Petzschner et al. (2015) argue that Bayesian inference is used in estimating all kinds of physical magnitudes and can accommodate a wide range of behavioural effects in magnitude estimation, central tendency included.

A common assumption in Bayesian accounts of timing (and indeed magnitude estimation in general) is that people make an optimal estimate about how long the perceived duration is by incorporating their belief about how likely a duration is as a function of previously presented magnitudes; in a time reproduction task, they then reproduce an amount of time to match this estimate (Cicchini et al., 2012; Jazayeri \& Shadlen, 2010; Di Luca \& Rhodes, 2016). These accounts thus assume that the estimation of a perceived duration makes reference to the previously perceived durations.

However, these accounts ignore another source of information that participants can rely upon when trying to optimally reproduce durations, i.e. their past experience with
reproduced durations．According to this alternative account， instead of making an estimate of a perceived duration and then use this estimate to guide time reproduction，people directly make an estimate about how long the reproduced duration should be．In this task－oriented inference，people make use of their past reproduction experience rather than their past perception experience；after all，when one is to reproduce a duration，the history of the reproduced durations may provide a better constraint for optimally determining how much time to reproduce．

The current study investigates whether people resort to previous perception or reproduction experience as the prior in their inferences．To do so，we take advantage of the range bias（a duration is reproduced as longer if it is placed in a context of long than short stimulus durations）．If people use their perception experience to calibrate their time estimation， then we should expect the range bias to occur between contexts of long vs．short perceived durations，even when the context of reproduced durations is kept constant．If instead people use their reproduction experience in their inference， we should expect the range bias to occur between contexts of long vs．short reproduced durations，even when the context of perceived durations is kept constant．

## Experiment 1

## Methods

Participants． 32 volunteers（ 24 females， $20.3 \pm 1.6$ in age） from the South China Normal University community volunteered for a small monetary reward．Participants in this experiment（and indeed in each experiment reported here）did not take part in any other experiment，though participants for all the experiments came from the same participant pool． These participants（and also those in other experiments）had normal or corrected－to－normal vision and received a small monetary reward for their participation

Design．We manipulated the reproduction context while keeping the perception context constant．To do this，we used an alternative－task paradigm where participants perceived a duration and then，upon a cue，either reproduced the stimulus duration or compared it to a new duration．Participants were presented with a set of short durations（ $600-2200$ in steps of 200 ms ）a set of long durations（ $1800-3400$ in steps of 200 ms ），all interleaved．Half of the participants reproduced the short durations and compared the long durations；the other half did the opposite（i．e．reproducing long durations and comparing short durations）．Note that，in such a design，while the two groups of participants differed in their reproduction contexts，they had identical perception contexts．Critically， the two reproduction contexts overlapped in three stimulus durations（ $1800,2000,2200 \mathrm{~ms}$ ），which allowed us to determine whether different reproduction contexts lead to a range bias，even when the perception context was kept constant．

Materials．For each of the 18 durations，a shorter（ 0.1 log shorter）and a longer（ 0.1 log longer）comparison duration was created．Each participant completed 5 blocks of trials．In
each block，every stimulus duration was presented twice， either both as reproduction or comparison trial．Half of the comparison trials used a shorter comparison duration and the other half had a longer one．Trials in each block were presented in an individually randomized order．In total，there were 180 experimental trials．

Procedure．The experiment was run on a desktop using E－ Prime 2．0．Participants sat about 50 cm away from the monitor．The experiment began with a practice session of 4 trials（ 2 reproduction and 2 comparison trials）followed by the main experiment．In a trial，a black cross（Courier New 48）was presented for a stimulus duration，followed by a blank screen of 300 ms ．Then，a cue（an asterisk＂＊＂or the phrase＂第二段时间＂，meaning＂second duration＂）was presented．An asterisk informed participants to reproduce the stimulus duration by holding down the spacebar for the same duration．At the press of the spacebar，the asterisk turned into three asterisks which remained on screen until the release of the spacebar．The phrase＂第二段时间＂informed participants to compare the first（stimulus）duration with an upcoming（comparison）duration．The text cue stayed on screen for 1 s and was replaced by a blank screen of 300 ms ． The comparison duration was then presented with a blue cross（Courier New 48），followed by a blank screen of 300 ms．Then a judgment screen was displayed asking participants to decide whether the first（stimulus）or second （comparison）duration was longer by pressing the＂ F ＂or＂ J ＂ key．


Fig．1．Results of Experiment 1．Each dash represents a participant＇s averaged reproduction for each stimulus duration in the long（red）or short（blue）reproduction context； the squares represent the averaged reproductions at the group level．

## Results and discussion

As the comparison data does not address our theoretical interest，to save space，we only report analyses on the reproductions．We first excluded as outliers reproductions that were $1 / 3$ or 3 times the stimulus duration，leading to the loss of $2 \%$（ 65 reproductions）of the data（this trimming
criterion was adopted for all experiments reported in this study). For the remaining data, a participant's mean reproduction for each stimulus duration was computed. We compared reproductions for the three overlapping stimulus durations ( 1800,2000 , and 2200 ms ) between the two reproduction contexts. An ANOVA with reproduction context as a between-participant factor and stimulus duration as a within-participant continuous variable revealed a significant main effect of reproduction context $\left(F_{(1,30)}=\right.$ 12.41, $p=.001, \eta^{2}=.29$ ): durations were reproduced for longer in the long than the short reproduction context (see Fig. 1). Reproduced durations increased as a function of the stimulus duration $\left(F_{(1,30)}=72.03, p<.001, \eta^{2}=.71\right)$ and did not significantly interact with reproduction context $\left(F_{(1,30)}=\right.$ $\left.1.44, p=.240, \eta^{2}=.05\right)$.

The main effect of reproduction context suggested a range bias in the reproduction of the overlapping durations: even when the perception context was kept constant, durations were reproduced as longer when prior reproductions were longer. Such a reproduction range bias is inconsistent with previous Bayesian inference accounts which posit that, to estimate a duration, people use their memory of the perceived duration and experience of previously perceived durations. Instead, the results suggest that people make use of their experience of previously reproduced durations in order to calibrate their reproductions.

In Experiment 2, we aimed to replicate the reproduction range bias using a within-participant design. In particular, we distinguished the long and short reproduction contexts using different modalities of reproduction: for half of the participants, people reproduce long durations with motor reproduction and short durations with an auditory reproduction (see below); for the other half, the paring was reversed. If reproduction experience calibrates duration reproduction, we should again expect a range bias for overlapping durations between the two reproduction contexts.

## Experiment 2

## Methods

Participants. 20 volunteers ( 10 females, $20.8 \pm 2.5$ in age) took part in the experiment.

Design. This experiment was similar to Experiment 1 except that we replaced the comparison task in Experiment 1 with an auditory reproduction task (and we manipulated reproduction task within-participants). As in Experiment 1, there were two duration ranges (short range: $600-2200$ in steps of 200 ms ; long range: $1800-3400$ in steps of 200 ms ). Two experimental versions were created such that one version had the short range paired with motor reproduction and the long range with auditory reproduction and the other version had the reverse. As in Experiment 1, we were interested whether people would be susceptible to the reproduction range bias when reproducing the overlapping durations (i.e. 1800,2000 and 2200 ms ) under different reproduction contexts.

Materials. As in Experiment 1, there were 5 blocks of trials and each block contained two occurrences of each of the 18 stimulus durations (i.e. 36 trials in each block). Trials in each block were presented in an individually randomized order. For auditory reproduction, a 10 s sine-wave pure tone sampled at a rate of 44100 Hz was created using Audacity.

Procedure. The experimental setting and overall experimental procedure were the same as those in Experiment 1, except that participants always reproduced a stimulus duration. After a cross was presented for a stimulus duration, followed by a blank, an image of a keyboard (as a cue for motor reproduction) or a mouse (as a cue for auditory reproduction) was displayed. For motor reproduction, as in Experiment 1, participants held down the spacebar to reproduce the stimulus duration. For auditory reproduction, participants clicked the mouse (at which point the mouse image disappeared) to initiate a tone and clicked again to terminate it when they felt that tone had been played for the same length as the stimulus duration. The experiment lasted for about 25 min .


Figure 2. Results of Experiment 2. Each dash represents a participant's averaged reproduction for each stimulus duration in the long (red) or short (blue) reproduction range; the squares represent the averaged reproductions at the group level.

## Results and discussion

About 6\% (216 reproductions) of the data were excluded as outliers (the high exclusion rate was due to the fact that participants sometimes accidentally pressed rather than held down the spacebar due to the influence of the auditory reproduction method). For the overlapping durations, an ANOVA using reproduction context and stimulus duration as within-participant variables reveals a significant main effect of reproduction context $\left(F_{(1,19)}=27.73, p<.001, \eta^{2}=.59\right)$, with longer reproductions of the overlapping stimulus durations in the long compared to the short reproduction context. Reproductions also increased as a function of stimulus duration $\left(F_{(1,19)}=131.5, p<.001, \eta^{2}=.87\right)$. There was no significant interaction between reproduction context and stimulus duration $\left(F_{(1,19)}=0.13, p=.724, \eta^{2}=.01\right)$.

The results thus replicated, with a within-participant design, the finding in Experiment 1 that reproduction experience calibrates duration reproductions. That is, a duration was reproduced as longer if it was done in the context of long reproductions compared to short reproductions. They also suggest that people construct taskspecific priors (i.e. past motor vs. auditory reproduction experience in the current experiment) in their time estimation, an issue that awaits further empirical verification.

In Experiment 3, we further explore whether manipulating the perception context alone leads to a range bias. If reproduction experience, but not perception experience, calibrates reproduction, we should not see a range bias in this experiment.

## Experiment 3

## Methods

Participants. 20 volunteers ( 14 females, $20.0 \pm 1.2$ in age) took part in Experiment 3.

Design. We manipulated perception context (short vs. long) within-participants in a blocked design. As in Experiment 1, Experiment 3 used an alternative-task paradigm (reproduction or comparison). The two perception contexts were created using three ranges of durations: the short perception context consisted of 6 short durations ( $600-$ 1600 in steps of 200 ms ) serving as comparison durations and 6 mid durations ( $1200-2200$ in steps of 200 ms ) serving as reproduction durations; the long perception context consisted of the 6 mid durations serving as reproduction durations and 6 long durations ( $1800-2800$ in steps of 200 ms ) serving as comparison durations. Thus, the two perception contexts had the same range of durations to be reproduced (i.e. both had the mid durations for reproduction) but differed in the range of durations to be perceived (long and mid durations for the long perception context but short and mid durations for the short perception context). If the perception context manipulation leads to a range bias, we should expect the mid durations to be reproduced as longer in the long than in the short perception context. Alternatively, if the range bias is driven by reproduction experience only, we should expect the mid durations to be reproduced as equally long between the two perception contexts.

Materials. As in Experiment 1, a shorter ( 0.1 log shorter) and longer ( 0.1 log longer) comparison duration were created for each of the comparison durations. Three blocks of materials were created for both the short and the long perception context. In each block, each stimulus duration was presented twice for reproduction and twice for comparison (once with a longer comparison duration and once with a shorter comparison duration), amounting to 24 trials in each block. Two experimental versions were created: the three blocks of the short perception context preceded those of the long perception context in one version and the order was reversed in the other. A short practice block of 4 trials preceded the first block. In order to prevent possible spillover of the perception context in the first three blocks to the last
three blocks, a compulsory 2-min break was inserted after the first three blocks; additionally, a practice block of 12 trials preceded the $4^{\text {th }}$ block.

Procedure. The experimental setting and the trial structure were identical to those in Experiment 1; that is, after the presentation of a stimulus duration, depending on the ensuing cue, participants either reproduced the stimulus duration or compared it with an upcoming duration. During the 2-min break, participants were allowed to do whatever they liked as long as they remained seated in the test cubicle. The experiment took about 25 min .


Figure 3. Results of Experiment 3. Each dash represents a participant's averaged reproduction for each stimulus duration in the long (red) or short (blue) perception context; the squares represent the averaged reproductions at the group level.

## Results and discussion

The trimming criterion led to the exclusion of 5\% (80 data points) of all the reproduced durations. An ANOVA with perception context as a within-participant factor and reproduction durations $(1200-2200)$ as a within-participant continuous variable revealed no significant main effect of perception context $\left(F_{(1,19)}=0.82, p=.377, \eta^{2}=.04\right)$, suggesting that the reproductions of stimulus durations were similar between the two perception contexts. Reproduction increased as a function of the stimulus duration $\left(F_{(1,19)}=\right.$ 255.3, $p<.001, \eta^{2}=.93$ ). There was no significant interaction $\left(F_{(1,19)}=1.18, p=290, \eta^{2}=.06\right)$.

The failure for the perception context manipulation to induce a range bias suggests that participants did not use their perception experience to infer stimulus duration for their reproductions. In Experiment 4, we changed all the comparison trials in Experiment 3 into reproduction trials so that the long and short perception context had respectively a long and short reproduction range (i.e. we additionally introduced the reproduction context manipulation). If people use reproduction experience to calibrate their time estimation, we should restore the range bias that was missing in Experiment 3.

## Experiment 4

## Methods

Participants. Another 20 volunteers ( 14 females, $19.9 \pm 1.1$ in age) took part in the experiment.

Design, materials and procedure. These were the same as those in Experiment 3 except that the comparison trials in Experiment 3 were changed into reproduction trials. Thus, the short perception context had 6 short durations ( $600-1600$ in steps of 200 ms ) and 6 mid durations ( $1200-2200$ in steps of 200 ms ), all to be reproduced; the long perception context had 6 mid durations ( $1200-2200$ in steps of 200 ms ) and 6 long durations ( $1800-2800$ in steps of 200 ms ), all to be reproduced.


Figure 4. Results of Experiment 4. Each dash represents a participant's averaged reproduction for each stimulus duration in the long (red) or short (blue) perception/reproduction context; the squares represent the averaged reproductions at the group level.

## Results and discussion

We excluded 3\% (106 data points) of all the reproduced durations as outliers. We compared the reproductions of the 6 overlapping stimulus durations ( $1200-2200$ in steps of 200 ms ) between the two perception (and indeed reproduction) contexts. In contrast to the finding in Experiment 3, the ANOVA showed a significant main effect of perception/reproduction context $\left(F_{(1,19)}=10.00, p=.005, \eta^{2}\right.$ $=.34$ ), with longer reproductions of the stimulus durations when they were part of the long than the short perception/reproduction context. Reproductions increased as a function of the stimulus duration $\left(F_{(1,19)}=255.6, p<.001\right.$, $\eta^{2}=.93$ ). The two variables did not interact significantly $\left(F_{(1,19)}=0.44, p=.514, \eta^{2}=.02\right)$, suggesting a central tendency in the reproduced durations in both contexts.

The most striking observation is the return of the range bias for the overlapping stimulus durations when different reproduction ranges were introduced, in contrast to Experiment 3, where the reproduction range was the same between the two perception contexts. Such a finding clearly
suggests that the reproduction experience, rather than the perception experience, drives the range bias.

## Prior updating

A crucial prediction of Bayesian inference in time perception is that the prior is constantly updated. If the prior in time reproduction is based on previously reproduced rather than perceived durations, as our experiments have shown, we should predict the most recently reproduced (but not perceived) duration to have an influence on the prior, and hence on the posterior, such that a longer reproduced (but not perceived) duration in the preceding trial leads to a longer reproduced duration in the current trial.

The comparison trials in Experiments 1 and 3 allowed us to examine the possible influence of the preceding perceived (i.e. stimulus) duration on the prior belief (and the reproduced duration in the current trial). We used linear mixed effects modelling for these analyses, where we included as predictors the stimulus duration of the current trial and the stimulus duration in the preceding trial. For Experiment 1, though reproductions increased as a function of the current trial's stimulus duration ( $\beta=556.28, S E=35.56, t_{(31.0)}=15.65, p<$ .001 ), they were insensitive to the magnitude of the stimulus (i.e. perceived) duration in the preceding comparison trial ( $\beta$ $\left.=-1.75, S E=17.64, t_{(28.1)}=-0.10, p=.922\right)$. The same pattern was also observed in Experiment 3 ( $\beta=241.36, S E=16.73$, $t_{(19.4)}=14.43, p<.001$, for current stimulus duration; $\beta=-$ 36.90, $S E=26.35, t_{(19.1)}=-1.40, p=.177$, for preceding stimulus duration).

We next analyzed reproductions taking into account the preceding reproduced duration (i.e. when the preceding trial was a reproduction trial) in all the 4 experiments. Reproductions always increased as a function of the stimulus duration of the current trial $\left(\beta=570.52, S E=23.86, t_{(31.1)}=\right.$ 23.91, $p<.001 ; \beta=643,41, S E=45.47, t_{(19.0)}=14.15, p<$ $.001 ; \beta=225.05, S E=20.50, t_{(17.6)}=10.98, p<.001 ;: \beta=$ 413.54, $S E=22.89, t_{(19.0)}=18.07, p<.001$; for Experiments $1-4$ respectively) and also of the preceding trial $(\beta=191.46$, $S E=16.99, t(38.2)=11.27, p<.001 ; \beta=89.34, S E=16.64$, $t_{(18.4)}=5.37, p<.001 ; \beta=151.48, S E=19.95, t_{(20.6)}=7.59, p$ $<.001 ; \beta=125.94, S E=15.96, t_{(17.7)}=7.89, p<.001$; for Experiments 1-4 respectively). These findings consistently suggest that the prior was updated by a recent reproduction output such that a recent longer reproduction increased the prior mean, which in turn increased the posterior mean, resulting in a longer reproduction in the current trial.

## General discussion

In four experiments, we showed that people use their past (in particular the most recent) reproduction experience to calibrate their duration reproduction. These results are inconsistent with previous Bayesian inference accounts of timing (and magnitude estimation in general), whereby people use previously perceived durations to calibrate their noisy memory of stimulus duration, which in turn is used to guide reproduction (Jazayeri \& Shadlen, 2010; Di Luca \& Rhodes, 2016; Cicchini et al., 2015; Petzschner et al., 2015).

Such a memory-optimizing account is explicitly spelled out in Petzschner et al. (2015) as a Bayesian inference account of magnitude estimation in general. In this account, people perceive a magnitude (e.g., a duration) and keep a noisy memory of it. Later they use their prior belief to infer an optimal estimate based on this noisy memory, and this estimate represents the inferred stimulus duration that is used to guide subsequent response (e.g., time reproduction).

Instead, the current findings support a task-oriented Bayesian inference account, where people directly use the noisy memory of the stimulus duration and their past reproduction experience to infer a reproduction estimate. Note that, unlike previous accounts, such an estimate is not an updated version of the stimulus duration but should instead be viewed as a planned reproduced duration.

If it is the case that Bayesian inference is task-oriented (i.e. the inference serves the task at hand), then we should expect the source of the prior information to vary across different magnitude tasks. For instance, it is possible that, whereas time reproduction recruits prior reproduction experience, time comparison may instead recruits prior perception experience as the task would involve making inferences about perceived durations. It is also possible, as Experiment 2 suggested, that different reproduction tasks may resort to task-specific priors. These remain to be tested in future studies.

## Acknowledgments

The research was supported by ESRC (ES/L010224/2) and the National Natural Science Foundation of China (31571142).

## References

Algom, D., \& Marks, L. E. (1990). Range and regression, loudness scales, and loudness processing: Toward a rangebound psychophysics. Journal of Experimental Psychology: Human Perception and Performance, 16, 706.
Bausenhart, K. M., Dyjas, O., \& Ulrich, R. (2014). Temporal reproductions are influenced by an internal reference: Explaining the Vierordt effect. Acta Psychologica, 147, 6067.

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., \& Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. Journal of Neuroscience, 32, 10561060.

Di Luca, M., \& Rhodes, D. (2016). Optimal perceived timing: integrating sensory information with dynamically updated expectations. Scientific Reports, 6.
Fotios, S. A., \& Cheal, C. (2007). Evidence for response contraction bias in side-by-side matching tasks. Lighting Research \& Technology, 39, 159-169.
Gallistel, C. R., \& Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. Trends in Cognitive Sciences, 4, 59-65.
Griffiths, T. L., \& Tenenbaum, J. B. (2011). Predicting the future as Bayesian inference: people combine prior
knowledge with observations when estimating duration and extent. Journal of Experimental Psychology: General, 140, 725.
Gu, B. M., \& Meck, W. H. (2011). New perspectives on Vierordt's law: memory-mixing in ordinal temporal comparison tasks. In Multidisciplinary aspects of time and time perception (pp. 67-78). Springer Berlin Heidelberg.
Hollingworth, H. L. (1910). The central tendency of judgment. The Journal of Philosophy, Psychology and Scientific Methods, 7, 461-469.
Jazayeri, M., \& Shadlen, M. N. (2010). Temporal range calibrates interval timing. Nature Neuroscience, 13, 10201026.

Jones, L. A., \& Hunter, I. W. (1982). Force sensation in isometric contractions: a relative force effect. Brain Research, 244, 186-189.
Jou, J., Leka, G. E., Rogers, D. M., \& Matus, Y. E. (2004). Contraction bias in memorial quantifying judgment: Does it come from a stable compressed memory representation or a dynamic adaptation process? American Journal of Psychology, 117, 543-564.
Lejeune, H., \& Wearden, J. H. (2009). Vierordt's The Experimental Study of the Time Sense (1868) and its legacy. European Journal of Cognitive Psychology, 21, 941-960.
Li, Y., \& Dudman, J. T. (2013). Mice infer probabilistic models for timing. Proceedings of the National Academy of Sciences, 110, 17154-17159.
McGuire, J. T., \& Kable, J. W. (2012). Decision makers calibrate behavioral persistence on the basis of timeinterval experience. Cognition, 124, 216-226.
Moon, J., \& Anderson, J. R. (2013). Timing in multitasking: Memory contamination and time pressure bias. Cognitive Psychology, 67, 26-54.
Petzschner, F. H., Glasauer, S., \& Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. Trends in Cognitive Sciences, 19, 285-293.
Poulton, E. C. (1979). Models for biases in judging sensory magnitude. Psychological Bulletin, 86, 777-803.
Radvansky, G. A., Carlson-Radvansky, L. A., \& Irwin, D. E. (1995). Uncertainty in estimating distances from memory. Memory \& Cognition, 23, 596-606.
Shi, Z., Church, R. M., \& Meck, W. H. (2013). Bayesian optimization of time perception. Trends in Cognitive Sciences, 17, 556-564.

# The most efficient sequence of study depends on the type of test 

Paulo F. Carvalho (pcarvalh@andrew.cmu.edu)<br>Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave Pittsburgh, PA 15213 USA<br>Robert L. Goldstone (rgoldsto@indiana.edu)<br>Department of Psychological and Brain Sciences \& Cognitive Science Program, Indiana University, 1100 E. 10th St. Bloomington, IN 47405 USA


#### Abstract

Previous research has shown that the sequence in which concepts are studied changes how well they are learned. In a series of experiments featuring naturalistic concepts (psychology concepts) and naïve learners, we extend previous research by showing that the sequence of study changes the representation the learner creates of the study materials. Interleaved study leads to the creation of relatively interrelated concepts that are represented by contrast to each other and based on discriminating properties. Blocked study, instead, leads to the creation of relatively isolated concepts that are represented in terms of their central and characteristic properties. The relative benefits of these representations depend on whether the test of conceptual knowledge requires contrastive or characteristic information. These results argue for the integrated investigation of the benefits of different sequences of study as depending on the characteristics of the study and testing situation as a whole.


Keywords: study sequence; interleaving; interrelated concepts;

## Introduction

The sequence of study while learning concepts changes what is learned and how well it is learned. Therefore, it is perhaps not surprising that understanding how students should organize their study to promote learning has emerged as a major area of active interest in educational and cognitive science research. Previous research has focused on how different sequences might improve learning (Birnbaum et al., 2013), and how the benefits of different sequences might interact with different study conditions (Carvalho \& Goldstone, 2015), materials (Carvalho \& Goldstone, 2014), individual characteristics (Sana et al., 2016), or selfregulation (Carvalho et al., 2016).

When hard-to-discriminate concepts are studied in an interleaved fashion, by alternating the study of the different concepts, learning is improved compared to when different concepts are studied in separate blocks (Kornell \& Bjork, 2008). However, the benefit of interleaved study is not universal. For example, it has been shown that when studying concepts that have high within-category diversity in their properties (for example, the category mammal which includes bats, cows, and whales), studying each concept in separate blocks can result in better learning (Carvalho \& Goldstone, 2014). This apparent inconsistency lead to the proposal that different sequences of study emphasize different properties of the studied materials and thus might be more appropriate for different types of concept learning tasks
(Carvalho \& Goldstone, 2014). The Sequential Attention Theory (Carvalho \& Goldstone, 2015), proposes a mechanism through which attention and encoding during blocked study are progressively directed towards the similarities among successive items belonging to the same category whereas attention and encoding during interleaved study are progressively directed towards the differences between successive items belonging to different categories. Because of this influence on cognitive processing, Carvalho and Goldstone (2015) propose that the sequence of study can accelerate or delay learning, depending on whether the constraints created by the sequence of study match those of the encoding situation (e.g., interleaved study in situations critically hinging on the encoding of differences between concepts, such as the study of highly similar concepts), or mismatch it (e.g., blocked study in the same situations).

In this work, we aim to extend these results to demonstrate that different encoding experiences will result in different representations that will be more or less appropriate depending on the requirements of the testing situation. Our proposal is as follows: because different information is encoded with different sequences of study, different sequences of study potentiate different representations of what was studied. More specifically, encoding the differences between concepts through interleaved study will tend to lead to the creation of interrelated concepts whose representations are contrasted away from each other by emphasizing or exaggerating their distinctive characteristic relative to each other (Corneille et al., 2006; Goldstone, 1996). Conversely, blocking will tend to lead to encodings of the similarities within each concept that will, in turn, create relatively isolated, stand-alone, representations (Goldstone, 1996).

These different representations, once created, are suited for different uses. Although an interrelated representation of two concepts will be helpful in a new context in which discriminating the previously learned concepts is important, isolated representations of the same concepts may not be as useful. Conversely, an isolated representation of a concept will include more information about all the properties of that concept, whether or not they serve to distinguish it from the other learned concepts, making it ideal for situations in which these details are relevant, such as when the concept must be differentiated from other new concepts possessing new distinctive features.

Students often create flashcards as a study and self-testing tool (Hartwig \& Dunlosky, 2012). These flashcards might
include a definition or an example of a concept on one side of the card and the correct response on the other. When studying using examples, students might choose to study all the cards from one concept in a block or to interleave cards from different concepts. One important question, then, is if different sequences of examples will influence students' performance for different types of tests - a question that, to the best of our knowledge, has not been addressed before. This is not only an important question at the theoretical level - to know the representational differences created by different sequences - but also at the practical level because changing the sequence of study materials is an easy and cheap intervention that might have substantial influences on learning outcomes (Dunlosky et al., 2013). In fact, previous researchers have emphatically advocated presenting information interleaved whenever possible, warning students about the perils of blocked study (e.g., Bjork, Dunlosky, \& Kornell, 2013), and it has been suggested as an important factor of which all new instructors should be aware (Deans for Impact, 2015).

For this purpose, we developed two experiments in which learners studied concepts of psychology (e.g., "Hindsight bias"; Rawson et al., 2015) in one of the sequences and were then tested in different situations, similar to common study practices by students. Importantly, some of the tests required discrimination between different concepts (e.g., multiplechoice test), whereas others required an independent representation of each concept (e.g., writing a definition). We consider writing a definition to require an independent representation because these definitions can be expressed without referring to other learned concepts. For example, a participant could write a definition for "availability heuristic" without having learned or remembered any of the other presented concepts (Goldstone, 1996). We predict that for tests that emphasize isolated, independent knowledge of the properties of each concept, such as writing a definition, participants will perform better following blocked study. Conversely, for tests that require discriminating different concepts, i.e., those that involve choosing between several options, participants will perform better with interleaved study.

## Experiment 1

## Method

Table 1: Participant demographic characteristics for Experiments 1 and 2.

| Experiments 1 and 2. |  |  |
| :---: | :---: | :---: |
| Pair | Exp. 1 | Exp. 2 |
| Mean Age (SD) | $33(10)$ | $36(11)$ |
| Gender (\% Females) | $45.5 \%$ | $68 \%$ |
| Education (\% Bachelor's or | $50 \%$ | $64 \%$ |
| higher) | $0.04(0.21)$ | $0.21(1.13)$ |

Participants. A group of twenty-eight people were recruited through Amazon's Mechanical Turk (https://www.mturk.com/). Data from 6 participants were excluded from analyses because of possible compliance
issues (see below for details). The demographic characteristics of participants in the overall sample are presented in Table 1.
Stimuli. We used a stimulus set of introductory concepts and examples created by Rawson et al. (2015). The stimuli included 10 concepts taught in Introductory Psychology and 10 example situations for each concept, collected from textbooks of Introductory Psychology. The concepts were divided into two groups by relatedness. Each group contained unrelated concepts only, whereas across groups pairs of related concepts existed (see Table 2). Relatedness of the concepts was judged by the authors by comparing the definitions of the concepts and confirmed by analyzing the pattern of errors in multiple-choice questions without feedback in a pilot study. Previous research looking at sequence of study using these materials used this concept grouping as well (Rawson et al., 2015).

Table 2: Groups of concepts used in Experiment 1 and Experiment 2. Each row includes a pair of related concepts.

Columns contain only unrelated concepts.

| Pair | Group A | Group B |
| :---: | :---: | :---: |
| 1 | Availability Heuristic | Representativeness <br> heuristic |
| 2 | Door-in-the-face | Foot-in-the-door |
| technique | technique |  |
| 3 | Hindsight bias | thinfing |
| 4 | Fundamental | Deindividuation |
| 5 | Mertribution exporrore effect | Social facilitation |

Design and Procedure. This Experiment had two conditions manipulated within-subject: Study Sequence (Blocked vs. Interleaved) and Type of Test (Multiple-Choice Test vs. Definition Match Test vs. Write Definitions Test).

The experiment had three phases: pretest, study and test. Participants completed one pretest, two study phases and two tests phases in the following order: Pretest - Study 1 - Test 1 - Study 2 - Test 2. The first and second study phases were the same in every aspect except for the sequence of study and the concepts studied. One study phase was interleaved and the other blocked (order counterbalanced across participants). A different group of to-be-learned concepts was used in each study phase. In the interleaved condition learners studied an example of each concept before studying the same concept again (e.g., ABCABC...). Conversely, in the blocked condition learners studied all examples of each concept before starting a new concept (e.g., AABBCC...). Moreover, the test phase only tested the concepts learned in the immediately preceding study phase. Between each study and test phase participants completed a distractor task by watching a 4-minute video on an unrelated topic and answering a question about that video.

During the pretest phase participants were told that they would be presented with several psychology concepts that they were asked to rate regarding their familiarity/knowledge. Participants were told that not
knowing the concepts was not an issue for the study, would not impact their eligibility or payment, and that they should be honest in their responses. On each trial, the name of a concept was presented and participants had to rate on a scale from 1 ("Not familiar at all") to 7 ("Very familiar") how familiar they were with that concept. Following each rating, participants were asked to provide an example of that concept, or enter "I don't know" if they did not know any. Participants completed the pretest for the ten to-be-studied concepts across both study phases.

Following the completion of the pretest, participants completed the study phase. During the study phase, participants studied examples of situations depicting each of five concepts, one at a time and were asked to choose the name of the concept they thought the example instantiated. Participants were given feedback after each response. During study, participants studied five examples of each of the five concepts.

During the test phase, participants completed three types of tests: Multiple-Choice, Writing Definitions and Match Definitions, always in that order. The Multiple-Choice test used the same procedure as the study phase with new examples and without feedback. In the Writing Definitions, test participants were shown the name of each of the concepts studied one at a time and asked to write the best definition possible for that concept, based on what they had learned in the previous study phase. In the Match Definitions test participants were presented with the textbook definition of each concept, one a time, and asked to identify what concept that definition belonged to by pressing the corresponding button on the screen. The order of trials within each of the tests was randomized across participants. None of the test phase tasks had any time limit.

## Results and Discussion

Because the study was conducted online without experimenter supervision, we first inspected the data in order to identify potential compliance issues. For each participant, we calculated the median response time during both study phases. The sample's median response time to complete the study phase was 10.5 seconds per problem (max: 22.9 sec./problem; min: $0.73 \mathrm{sec} . /$ problem). We calculated the 10th and the 90 th percentiles for the distribution of median response times, $3.3 \mathrm{sec} . / \mathrm{problem}$ and $16 \mathrm{sec} . / \mathrm{problem}$ respectively, and used these values as a measure of noncompliance in the task. Responding too fast (faster than the 10th percentile) is likely due to participants who are not reading the problems and just advancing through the experiment quickly; similarly, longer response times (above that of the 90 th percentile) are likely due to potentially distracted participants. Six participants were identified based on this analysis and their data were excluded from further analyses.

All the analyses below are ANCOVA analyses including average pretest score and counterbalancing condition as covariates.

Pretest. To analyze the data from the pretest we calculated 25th, 50 th and 75 th percentiles of the ratings (see Table 3). As can be seen, most participants showed little or no knowledge of the to-be-studied concepts (mean of approximately 2 in a 1-7 scale). The provided examples further confirmed this interpretation.

| Table 3: Pretest results for Experiments 1 and 2 (1-7 scale). |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | 25 <br>  <br> Percentile | Percentile | 75 <br> Percentile | $M$ | SEM |
| Exp <br> 1 | 1.00 | 1.55 | 2.03 | 1.71 | 0.16 |
| Exp | 1.34 | 1.79 | 2.39 | 1.97 | 0.16 |
| 2 |  |  |  |  |  |

Study Phase. Mean performance during the blocked study phase was $72 \%$ (SEM $=5 \%$ ), whereas during interleaved study it was $67 \%$ (SEM $=5 \%$ ). This difference was not statistically significant, $F(1,20)=1.95, p=.169, \eta^{2}{ }_{G}=.012$. Test Phase. Two trained coders, blind to condition assignment, rated as correct or incorrect each of the written definitions. These two coders agreed $87 \%$ of the time and inter-coder reliability was high, Cohen's Kappa $=.725, p<$ .0001. Disagreements were resolved by a third coder, also blind to condition assignment of the responses.
Performance for the test phase is depicted in Figure 1. As can be seen in the figure, the type of tests varied in their level of difficulty, with participants performing better in the Definitions Match test and worse in the Write Definitions test, $F(2,42)=17.62, p=<.0001, \eta^{2}{ }_{G}=0.151$. Although there was no overall main effect of study sequence $F(2,42)$ $=1.36, p=.256, \eta^{2}{ }_{G}=.006$, there was a significant interaction between type of test and study sequence, $F(2,42)=5.26, p$ $=.022, \eta_{G}^{2}=.022$.


Figure 1: Results for the Test Phase of Experiment 1. Dotted lines represent chance level. Error bars represent standard errors of the mean.

To further investigate this interaction, we compared the effect of type of study sequence on each of the tests by calculating the difference in performance following blocked and interleaved study for each type of test (interleaved blocked). The difference in performance between the two conditions varied across type of test, $F(2,42)=5.10, p=$ $.011, \eta^{2}{ }_{G}=.110$. Planned contrasts using FDR correction indicate that the effect of study sequence was significantly
different when comparing the Write Definitions test ( $M=$ $.15, S E M=.06)$ with the Multiple-Choice test $(M=.009$, SEM $=.03$ ), $\mathrm{p}=.033$, and the Match Definitions test $(M=$ $.02, S E M=.05$ ), $p=.040$, but not when comparing the Multiple Choice and the Match Definitions tests, $p=.844$.

These results are consistent with our proposal that blocked study encourages learners to develop independent, standalone representations rather than highlighting diagnostic features (i.e., those that discriminate between the concepts). Interleaved study emphasizes features that discriminate between concepts, which would be more helpful for a subsequent categorization task than a task that requires generation of stand-alone definition of the concept.

## Experiment 2

Our main proposal in this paper is that blocked study creates relatively independent representations of each concept studied which emphasizes the concept's characteristic features. These independent representations include more details from each concept than what is fostered by the relatively interrelated representations created during interleaved study. In the context of studying examples of different concepts, we proposed that blocked study allows learners to more successfully write definitions of the concepts because a definition requires the type of knowledge that blocked study promotes; it is generally possible to write good definitions for the learned psychology concepts without mentioning other psychology concepts learned at the same time. Consistent with this hypothesis, Experiment 1 showed that following blocked rather than interleaved study, learners were more successful at writing definitions of concepts, but the groups did not differ on classifying examples.

However, when two concepts are highly related (e.g., foot-in-the-door and door-in-the-face technique) their definitions can be aptly construed in relation to each other. If they are studied together, one central feature to include in the definition is the feature that discriminates them. Thus, the fact that in the previous experiments learners studied in the same session concepts that were dissimilar from each other and varied in many properties (see Table 2) might have contributed to the pattern of results seen. Would studying similar concepts together change the pattern of results observed?

Studying related concepts together changes the learning task in at least three critical ways. First, studying similar concepts in the same session might result in the necessity to discriminate between similar situations in order to find the subtle differences between the two types of concepts. It has been shown before that the interrelated representations promoted by interleaved study are likely to improve learning in these situations of learning highly similar concepts (Carvalho \& Goldstone, 2014). Second, the features that discriminate these related concepts are also characteristic features of the concept, unlike what is the case when the concepts are dissimilar (see Table 2). This means that interleaved study could promote representations appropriate for a writing definitions test through identification of
differences between concepts, whereas these differences would not be likely to be highlighted in the previous experiment.

In sum, when similar items are studied in the same session, there are several reasons to believe that performance would benefit from interleaved study, even when the test requires learners to write definitions. However, when similar items are studied in separate sessions, as in Experiment 1, blocked study would promote best performance in a test requiring isolated representations, such as writing definitions.

To test this, we used a procedure similar to how students often organize their study. In most natural situations students are likely to randomly assign the topics to be studied to a study session or to follow the sequence of their textbook or instructor. Therefore, in this experiment we randomly assigned concepts to being studied either interleaved or blocked, instead of using different pre-defined groups of concepts that guarantee low between-category overlap as in the previous experiment. This results in a situation where similar concepts might be studied together or separately. We compare performance on multiple-choice and writing definitions tests following blocked or interleaved study in each one of these situations.

## Method

Participants. A group of 36 people completed the experiment following recruitment through Amazon's Mechanical Turk (https://www.mturk.com/). Data from 3 participants were excluded due to self-reported previous participation in another study with the same materials. Data from an additional 8 participants were excluded from analyses because of possible compliance issues (see below for details). The final sample included 25 participants. Table 1 includes the demographic characteristics of participants in the overall sample.
Stimuli and Procedure. In this experiment, we used the same set of materials as in Experiment 1 and Experiment 2, but concepts were randomly assigned to be studied interleaved or blocked. Thus, in this experiment we did not force related concepts to be studied in separate phases.

The procedure was similar to the procedure used in Experiment 1 except for the following differences. Participants studied only eight concepts, four interleaved and four blocked. During study, participants saw four situations depicting each one of the concepts. After study, participants played a game of Tetris for 30 seconds.

The test phase included only a multiple-choice test and a writing definitions test, always presented in that order. During the multiple-choice test participants saw a total of four novel examples of the concepts studied, presented one at a time, and were asked to indicate which concept it illustrated.

## Results and Discussion

We identified potentially non-compliant participants using the participants' response times during study. The sample's median response time to complete the study phase was 8.4
seconds per problem (max: $22.8 \mathrm{sec} . /$ problem; min: 0.47 sec./problem). The $10^{\text {th }}$ and $90^{\text {th }}$ percentiles for the distribution of median response times were $2.6 \mathrm{sec} . / \mathrm{problem}$ and $16 \mathrm{sec} . /$ problem respectively. Eight participants were identified as outliers based on their falling outside of this range and their data were excluded from further analyses.

In all the analyses presented below, mean pretest score and counterbalancing condition were included as covariates.
Pretest. As in the previous experiments, participants showed little to no pre-training knowledge of the to-be-studied concepts (see Table 3).
Study Phase. Mean performance during the blocked study phase was $79 \%$ ( $S E M=2.5 \%$ ), while during interleaved study it was $73 \%(S E M=4 \%)$. However, this difference was not statistically significant, $F(1,25)=2.65, p=.116, \eta^{2}{ }_{G}=$ 0.04.


Figure 2: Results for the Test Phase of Experiment 2. Dotted lines represent chance level. Error bars represent standard errors of the mean.

Test Phase. Two trained coders, blind to condition assignment, rated as correct or incorrect each of the Written Definitions provided. The two coders agreed $84 \%$ of the time and inter-coder reliability was high, Cohen's Kappa $=.611$, p $<.0001$. Disagreements were resolved by a third coder, also blind to the condition assignment of the responses.

To analyze the results from the two tests used in this experiment we classified each concept based on whether it had been studied blocked or interleaved and whether its related concept (see Table 2) had been studied in the same sequence or in different sequences. When both related concepts were studied in the same phase and in the same sequence (e.g., "foot-in-the-door technique", "door-in-theface technique" studied blocked), they were both classified as "Blocked" and "Same Sequence." However, when only one of the related concepts was studied, or the two related concepts were studied in different phases/sequences, both were marked "Different Sequences."

This classification of the concepts resulted in empty cells for participants who did not have both concepts studied in the Same Sequence and concepts studied in Different Sequences for both interleaved and blocked study. Because traditional repeated-measures ANOVA does not allow for the existence of empty cells and we wanted to maximize the inclusion of all data collected, here we used mixed model analyses and report Wald $F$ tests and respective $p$-values using KenwardRoger's approximation (Kenward \& Roger, 1997). The results are depicted in Figure 2.

As we saw in the previous experiments, overall learners performed better on the Multiple-Choice test than when writing definitions, Wald $F(1,33.842)=91.68, p<.0001$. Similarly, the sequence of study had no overall effect on performance, Wald $F(1,33.948)<1$. No interaction was found between these two variables, Wald $F(1,92.007)<1$.

However, the relatedness between concepts presented in the same sequence influenced performance. Overall, when participants studied the two related concepts in the same sequence their performance was lower ( $M=46.52 \%, S E M=$ $1.50 \%$ ) than when related concepts were not in the same sequence ( $M=62.41 \%, S E M=1.04 \%$ ), Wald $F(1,24.284)$ $=15.60, p=.0006$. Item relatedness also interacted with sequence of study and type of test, Wald $F(1,98.522)=6.83$, $p=.010$.

To further analyze this interaction, we explored the test results for each type of test separately. For the MultipleChoice test, only the effect of relatedness reached statistical significance, Wald $F(1,24.264)=9.43, p=.005$. However, for the Writing Definitions test, in addition to a significant effect of item relatedness, Wald $F(1,25.649)=9.81, p=$ .004 , we also found a significant interaction between item relatedness and sequence of study, Wald $F(1,29.282)=6.71$, $p=.015$ (see right panel of Figure 5). As predicted by the results of Carvalho and Goldstone (2014), the relative relatedness between items modulates the relative benefit of each sequence for the Writing Definitions test. Moreover, consistent with the results of Experiment 1, we see that when similar items are not studied in the same sequence, performance in the Write Definitions test benefits from blocked study, although this effect was only marginally significant, $t(35)=1.90, p=.066, d=0.317$.

## General Discussion

Overall, the results presented here show that the different sequences of study affect performance differently for different types of test. Studying examples of different concepts in a blocked sequence improves performance in a test requiring learners to provide a definition of the concept studied, whereas for other tests there is no difference in performance between the two sequences of study.

Consistently with previous research, we have argued that this pattern of results is related to the acquisition of different knowledge with each sequence. Whereas blocked study results in the creation of a relatively isolated representations (i.e., a stand-alone, independent representation of each concept), interleaved study results in interrelated representations (i.e., focusing on how a concept differs from other(s) studied at the same time; Corneille et al., 2006; Goldstone, 1996). Going one step further, these different representations are likely to be the result of differences in the underlying attentional and encoding processes (Carvalho \& Goldstone 2015). The information attended to and encoded during study will dictate what type of representation is brought to a new situation and therefore what is available at test.

Moreover, we also saw that the effect of study sequence is modulated by whether discrimination based on subtle differences is necessary or not during study or test, such as is the case with the related concepts presented together in Experiment 2. We argued that this is the result of the pressures of the study and testing situation: when studying related concepts, interleaved study (and the interrelated representations it promotes) helps learners determine what discriminates between closely related concepts. This interpretation is consistent with the results of Carvalho and Goldstone (2014) showing that when learners studied similar categories, interleaved study improved learning, whereas when studying dissimilar categories, blocked study improved learning. Although the sample sizes used in the studies reported here might seem small, it is important to note that all critical comparisons were within-subject manipulations which increases the analytic power and that the effect sizes reported here are large and in line with previous similar research.

In sum, the two main contributions of the present work are as follow; first, it goes beyond existing demonstrations that blocking is better/worse than interleaving by showing how sequence affects what is learned by creating different representations given the same content. Second, it provides evidence for the context-dependent nature of learning and how the benefits of each sequence depend on the learning situation. This evidence adds to previous demonstrations that the best sequence of study depends on the type of material being studied (Carvalho \& Goldstone, 2014; Patel et al., 2016), the type of study task (Carvalho \& Goldstone, 2015; Rawson et al., 2015), and whether students actively decide how to organize their study (Carvalho et al., 2016). These results also show the importance of developing theories of why one intervention is better than another. We have proposed a theory based on the similarities of the materials being learned and the nature of the task. When concepts are similar to each other, learners prioritize learning discriminating features. Writing definitions generally benefits from stand-alone representations unless the concepts being defined are similar to each other and benefit by being contrasted. The study of how an intervention interacts with the learning situation, we would argue, has the potential to not only provide a fuller understanding of how learning takes place, but also provide richer, more precise, recommendations for practice (Jonassen, 1982).

## Acknowledgments

We are grateful to Katherine Rawson for sharing the stimuli set with us. Dustin Finch, Kaley Liang, Ashton Moody, Alifya Saify, and Shivani Vasudeva assisted with response coding. Work supported by NSF grant \# 0910218 and IES grant \# R305A1100060 to RG, and Fellowship \# SFRH/BD/78083/2011 from the Portuguese Foundation for Science and Technology to PC.

## References

Birnbaum, M.S., Kornell, N., Bjork, E.L., \& Bjork, R.A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. Memory \& Cognition, 41(3), 392-402.

Bjork, R.A., Dunlosky, J., \& Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. Annual Review of Psychology, 64, 417-444.
Carvalho, P.F., \& Goldstone, R.L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. Memory \& Cognition, 42(3), 481-495.
Carvalho, P.F., \& Goldstone, R.L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. Psychonomic Bulletin \& Review, 22(1), 281-288.
Carvalho, P.F., \& Goldstone, R.L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? Frontiers in Psychology, 6.

Carvalho, P.F., Braithwaite, D.W., De Leeuw, J.R., Motz, B.A., \& Goldstone, R.L. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. PLoS ONE, 11(3).
Corneille, O., Goldstone, R.L., Queller, S., \& Potter, T. (2006). Asymmetries in categorization, cerceptual discrimination, and visual search for reference and nonreference exemplars. Memory \& Cognition, 34(3), 556-567.
Deans for Impact. (2015). The Science of Learning. Austin, TX: Deans for Impact.
Dunlosky, J., Rawson, K.A, Marsh, E.J., Nathan, M.J., \& Willingham, D.T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. Psychological Science in the Public Interest, 14(1), 4-58.
Goldstone, R. L. (1996). Isolated and interrelated concepts. Memory \& Cognition, 24(5), 608-628.
Hartwig, M.K., \& Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? Psychonomic Bulletin \& Review, 19(1), 126134
Jonassen, D.H. (1982). Aptitude-Versus Content Treatment Interactions: Implications for Instructional Design. Journal of Instructional Development, 5(4), 15-27.
Kenward, M.G., \& Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 53(3), 983-97.
Kornell, N., \& Bjork, R.A. (2008). Learning concepts and categories: is spacing the "enemy of induction"? Psychological Science, 19(6), 585-592.
Patel, R., Liu, R., \& Koedinger, K. (2013). When to Block versus Interleave Practice? Evidence Against Teaching Fraction Addition before Fraction Multiplication. In A. Papafragou, D. Grodner, D. Mirman, \& J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 2069-2074). Austin, TX: Cognitive Science Society.
Rawson, K.A., Thomas, R.C., \& Jacoby, L.L. (2015). The Power of Examples: Illustrative Examples Enhance Conceptual Learning of Declarative Concepts. Educational Psychology Review, 27(3), 483-504.
Sana, F., Yan, V.X., \& Kim, J.A. (2016). Study Sequence Matters for the Inductive Learning of Cognitive Concepts. Journal of Educational Psychology.

# Is there an explicit learning bias? Students beliefs, behaviors and learning outcomes 

Paulo F. Carvalho (pcarvalh@andrew.cmu.edu)<br>Elizabeth A. McLaughlin (mimim@cs.cmu.edu)<br>Kenneth R. Koedinger (koedinger@cmu.edu)<br>Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave<br>Pittsburgh, PA 15213 USA


#### Abstract

Learning by doing refers to learning practices that involve completing activities as opposed to explicit learning (e.g., reading). Although the benefits of learning by doing have been described before, it is still relatively uncommon in instructional practice. We investigated how much students employ learning by doing in online courses, and whether it is associated with improved learning outcomes. Spending more time completing activities had a larger impact on learning outcomes than spending more time reading, even in the case of mostly declarative content, such as in a Psychology course. Moreover, learning by doing is more efficient: grade improvements of 1 standard-deviation require $10-20 \%$ less time in learning by doing than reading. Finally, we contrast this evidence with students' a priori intuitions on best study strategies for their online course. Students overestimate the value of explicit learning through reading, and underestimate the value of active learning.


Keywords: learning by doing; retrieval practice; self-regulated learning; doer effect

## Introduction

A lot of instruction is focused on explicit learning (for example, through textbook reading, classroom lectures, and online videos). The underlying assumptions often are (a) that most knowledge we expect students to acquire in our courses is declarative in nature and, (b) perhaps, that even procedural knowledge can initially be acquired this way. Consistent with these beliefs, much emphasis has been devoted to the creation of video-based Massive Open Online Courses (MOOCs) and text-based online courses.

The emphasis on explicit learning is in stark contrast to established phenomena in cognitive psychology, advocating for the use of testing (Roediger \& Karpicke, 2006) and active learning (Wieman, 2014) as better learning tools. The testing effect describes the positive effects of engaging in selftesting, instead of additional passive study (for a review see Roediger \& Karpicke, 2006). This effect has been repeatedly shown in laboratory settings with diverse materials, including word pairs and text passages (e.g., Karpicke \& Blunt, 2011; Karpicke \& Roediger, 2008). The success of the testing effect in the laboratory led to some in-classroom studies looking at its extensibility as a tool to promote students' learning, also with positive outcomes (e.g., McDaniel et al., 2007).

Why might active practice not be used in the classroom? One possibility is that the effect is limited to controlled laboratory contexts in which other aspects of real-world instruction do not vary. There is currently a lack of evidence
from large-scale classroom studies demonstrating the benefits of testing over reading outside the lab. Another possibility is that the advantage of learning by doing is specific to some types of materials (e.g., procedural knowledge), thereby limiting its use by instructors and students. Indeed, there is some evidence showing that, under some circumstances, additional passive reading practice, compared to doing activities, might result in better learning (Sweller \& Cooper, 1985). In the KLI framework, Koedinger and colleagues (2012) postulate that the learning goals and the nature of the materials being studied are the determining forces behind whether reading or doing are better for improving learning.

In sum, conceivably, learning by doing is not used because it is not effective in real-world contexts or across a wide range of knowledge types. Moreover, there is an underlying assumption that when the focus of learning is declarative knowledge, the emphasis should be on reading activities that would foster the formation of connections between concepts and the creation of robust declarative knowledge (Anderson \& Schunn, 2000). Nonetheless, learning by doing is important because most of human expertise involves tacit knowledge of the cues and conditions for deciding when, where, and what knowledge to bring to bear in complex situations (Zhu et al., 1996). In this view, there might be no verbal shortcut to acquiring expertise; it might be best acquired by repeated practice.

In our research, we explore whether learning by doing is a better way of learning across different types of knowledge (i.e., declarative and procedural), and whether it is more efficient. We compare learning outcomes of students enrolled in two online courses as a function of frequency and time spent completing practice activities (doing) vs. reading.

## Students' study behavior and their beliefs about the best study strategy

Even if a class is designed to encourage students to learn by doing, including extensive self-testing and guided practice activities but minimal text, it is an open question whether students a) realize its potential, b) use it, and c) whether selfdirected learning by doing in the classroom is as effective as its guided counterparts in the laboratory. These questions are theoretically and practically important because previous research on other cognitive approaches to improve learning have repeatedly shown a difference in outcomes between when students are in control of their study and when they are not (Carvalho et al., 2016; Ciccone \& Brelsford, 1976), as
well as a lack of awareness by the students on how to best organize their study (Karpicke et al., 2009).

Koedinger et al. $(2015,2016)$ illustrated the power of learning by in the context of online courses used in real classrooms. The Open Learning Initiative (OLI) at Carnegie Mellon University (CMU) is a learning environment that includes several courses each focusing on rich and interactive learn-by-doing activities, aligned with student-centered learning outcomes, and designed around science-based learner models. By analyzing student self-regulated study behavior in online classes taught at different universities using OLI materials, Koedinger et al. $(2015,2016)$ identified a "doer effect" - completing more practice activities is a stronger predictor of student performance than completing more reading activities.

## The present work

The present work builds on early evidence of the "doer effect" and extends it. One explanation for why completing more doing activities has a larger impact on learning than completing more reading activities is that completing doing activities may be more time intensive. If students devote more time to studying, regardless of how they do it, they are more likely to learn more. In other words, more learning results from the time devoted to an activity (e.g., reading or doing), not from the activity itself. Conversely, if learning by doing is more beneficial because it engages students in an active learning process (Wieman, 2014), it should be associated with better learning outcomes even if students spend comparatively less time engaging in that activity. To investigate this question, we compare the time spent reading and doing, and its relative impact on learning outcomes. Is it the case that reading for longer periods results in better learning outcomes than doing for shorter periods?

Additionally, to probe the generalizability of learning by doing even for declarative knowledge, we investigate students' behavior in two courses. An introductory psychology course focusing mostly on declarative knowledge and a computation course focusing on both declarative and procedural (learning how to code) knowledge.

Finally, we investigate students' beliefs about the usefulness of using learning by doing in their study. At the start of each course, as part of an optional unit, students completed a question on what they thought was the best strategy to study for the course. Are students' a priori beliefs on how to study biased towards explicit learning (i.e., reading)?

## The "doer effect" in a Psychology MOOC

## Method

Sample. Our analyses include data from 783 students enrolled in an online "Introduction to Psychology as a Science" MOOC offered by the Georgia Institute of Technology through Coursera. We included in the analyses students registered in OLI for whom pretest, quizzes and the
final exam data were available. For a description of the entire sample see Koedinger et al. (2015).
Description of the course. The course "Introduction to Psychology as a Science" was designed as a 12 -week introductory survey course, and is often taught during the first year of college. For each week of class, the course targeted a major topic area (e.g. Memory, Sense and Perception, Abnormal Behavior). Elements of CMU's Open Learning Initiative (OLI) "Introduction to Psychology" course were incorporated into Georgia Tech's "Introduction to Psychology as a Science" MOOC. OLI materials including text and interactive activities were available to students, in addition to the lectures, quizzes and other Coursera-based activities of the larger course. Each sub-topic was supported by a pre-recorded video lecture ( $10-15$ minutes, with downloadable slides) and included matched modules and learning outcomes in the OLI learning environment. A highstakes quiz assessed students against these outcomes at the end of each week.

The OLI modules included a variety of expository content (text, examples, images, and video clips) and a large number of interactive activities. Broadly, these activities serve two purposes. "Learn By Doing" activities, intended to support student outcome achievement, provide feedback and robust hints to support students. Figure 1 shows an example of a "Learn by Doing" activity from the Personality module covered in week 9 of the course. Another type of activity, "Did I Get This" activities, provided a self-comprehension check for students. These activities were created in conjunction with the OLI text materials and complement it by providing testing ("Did I Get This") or active learning ("Learn by Doing") activities that cover the concepts described in the text.


Figure 1: Screenshot of an OLI "Learn By Doing" activity from the module on Personality.
Research questions, measures, and analysis plan. We explore three main research questions: Q1: Does completing more practice ("doing") activities, compared to completing more reading activities, predict better learning outcomes?; Q2: Does spending more time on practice ("doing") activities, compared to time spent on reading activities, predict better learning outcomes?; Q3: What are students' beliefs regarding "best study strategies" for an online course?

To approach these questions, we created the following analyses plan. First we calculated our dependent measures:
number of reading and doing activities and total time spent in each. We started by identifying for each student the number of doing and reading activities. A doing activity was identified as responding to at least one practice activity in the OLI. A reading activity was identified as opening a text webpage in the OLI. Because opening a text webpage does not necessarily mean a student was reading, we adjusted the total number of activities by removing extremely short reading activities (below the $10^{\text {th }}$ percentile of reading time for all reading activities for that student). Our reasoning is that when students took extremely short amounts of time in a page, they might not have in fact read the associated text.

Table 1: Study strategies students could choose from. The metacognitive activity in the Psychology MOOC course included only the first 4 strategies. The Computing OLI course included all strategies in this table.

| Strategy | Description |
| :--- | :--- |
| Game the <br> System | Do activities without reading text. Select different <br> answer choices until correct. |
| Do-Read | Do first activity and if cannot answer, read relevant <br> text. |
| Read-Do | Read text and complete do activities as they come <br> up |
| Read | Read text, skip doing activities. |
| Do | Complete some activities, then go back to text and <br> look for a similar example to read. |

Because doing activities were contained inside webpages, often surrounded by text, time spent doing and reading for each activity/page was inferred from recorded data as follows. Timestamps were recorded for when a student opened a OLI page, when a student made a choice in each step of an activity (e.g., selected an option in a multiplechoice question), checked their responses in activities, asked for hints in the activities, and closed the page. From these logs, we could infer doing time as the time difference between the initial step and the final step of an activity. All other time spent (from opening to closing) in the page was considered reading time. However, this process does not include the time spent reading the doing activity text before starting the activity. To correct for this, the time spent completing each doing activity also includes a proportion of the time right before completing the first step (the other portion being classified as time spent reading). For reading time, we calculated the difference between the time when a webpage was initially accessed and the time an activity was started, as well as the time between an activity was finished and the another one started or the webpage was closed plus the portion of the time immediately before the first step of each activity. However, initial analyses of the time spent in each page revealed a number of large outliers (several standard deviations above the student mean for the student). These times might be indicative that a student left the webpage opened while completing other activities (potentially not related to the course). To correct for this, in addition to removing very short reading times (for
consistency with the number of reading activities analysis above), we also replaced very long reading times (above the $90^{\text {th }}$ percentile for that student) with the average reading time for that student. This way, we hope to reduce the influence of situations during which the student had the page opened but was not actively reading the text presented.

Our dependent measures included the summed quiz score across the 11 quizzes and the final exam score, all multiplechoice questions. Each quiz was worth 10 points. The final exam had 35 questions (each worth 1 point). To account for differences in student prior knowledge, we entered pretest score as a predictor in the models. The pretest, completed at the start of the course, was composed of multiple-choice questions from content covered in most of the units of the course and was graded from 0-20 points.

We converted the raw scores for the independent measures of student behavior as well as the dependent measures into standardize $z$-scores for ease of comparison across measures. We analyzed the effect of each independent measure on each dependent measure separately using a logistic regression model (in $R$ code):

$$
\begin{aligned}
& \text { zQuiz[zExam] }=\operatorname{lm}(z \text { Pretest }+ \text { zNumDoAct } \\
& [\text { zTotalDoTime }]+\text { zNumReadAct[zTotalReadTime }]+ \\
& \text { zNumDoAct[zTotalDoTime] }{ }^{*} \text { zNumReadAct[zTotalReadTi } \\
& \text { me }], \text { data }=\text { oli_do_read })
\end{aligned}
$$

Finally, to identify students' beliefs regarding best studying strategies (Q3), we took the students' responses in an activity during the "Learning Strategies" module included in the beginning of the OLI course. In this optional module (not included in the other analyses), students were introduced to several key research findings in learning sciences, and "best strategies" to achieve best learning. In one of the activities included in this unit students were asked to choose which of four study strategies they thought would yield best results in the course (see Table 1). To describe students $a$ priori study strategy judgments, we calculated the proportion of students who chose each of these alternatives before starting their study in the course. Each student could choose one or more of the options, out of the four offered: "Game the system", "Do-Read", "Read-Do", and "Read" (see Table 1).

## Results and Discussion

Table 2: Descriptive statistics for the main measures of students' study behavior and independent measures in the Psych MOOC

|  | $M(S D)$ | Median | $25^{\text {th }}$ <br> Prctl. | $75^{\text {th }}$ <br> Prctl. |
| :---: | :---: | :---: | :---: | :---: |
| Read Time <br> (mins) <br> Doing Time <br> (mins) | 9408 <br> $(5377)$ | $833(748)$ | 478 | 244 |
| \#Read Activities | $287(210)$ | 245 | 156 | 384 |
| \#Doing | $435(265)$ | 541 | 152 | 683 |
| Activities | $11(3.5)$ | 11 | 9 | 12265 |
| Pretest | $89(18.3)$ | 94 | 83 | 101 |
| Quizzes | $27(5.8)$ | 28 | 24 | 31 |
| Final Exam |  |  |  |  |

Descriptive measures of student behavior. As it can be seen in Table 2, students spent on average more time reading than doing ( 9000 min vs. 800 min , respectively); conversely students completed more doing than reading activities (435 vs. 287 , respectively). This overall descriptive data is consistent with the nature of the OLI course, which included a large number of short doing activities and text passages.
Q1: More "doing" activities predicts better learning outcomes. Results of the logistic regression predicting Quiz and Exam performance using number of doing and reading activities are presented in Table 3.

The regression analysis showed that higher quiz and exam scores are predicted by completing a larger number of doing activities $(\beta=0.40, p<.0001$ and $(\beta=0.24, p<.0001$, respectively), and by completing more reading activities ( $\beta=$ $0.11, p=.001$ and $\beta=0.11, p=.03$, respectively).

Importantly, the relative benefit of completing more doing activities was 2.4 to 3.6 times larger than completing more reading activities.

Overall, these results support those found by Koedinger et al. (2015, 2016), showing that completing more doing activities predicts better learning outcomes to a greater degree than completing more reading, even when we correct for the existence of very short (potentially off-task) reading events.

Finally, contrary to some intuitive predictions of the complementary nature of the two types of learning activities, their positive effect on learning outcomes are not additive. Completing more doing activities is more beneficial when students completed less reading activities (and vice-versa; $\beta$ $=-0.15, p<.0001$ and $\beta=-0.04, p=.40$, respectively).

Table 3: Results of logistic regression for both courses. Coefficients are standard deviations from the mean (z-scores).

|  |  | Quiz |  |  |  |  | Exam |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Course | DV | $\begin{gathered} \mathrm{Adj} \\ R^{2} \end{gathered}$ | Doing Coef. | Reading Coef. | Interact Coef. | Effect <br> Ratio | $\begin{gathered} \mathrm{Adj} \\ R^{2} \end{gathered}$ | Doing Coef. | Reading Coef. | Interact Coef. | Effect <br> Ratio |
| Psych | Number Activities | . 29 | $\begin{gathered} 0.40 \\ (0.04) \end{gathered}$ | $\begin{gathered} 0.11 \\ (0.04) \end{gathered}$ | $\begin{aligned} & \hline-0.15 \\ & (0.04) \end{aligned}$ | 3.6 | . 14 | $\begin{gathered} 0.24 \\ (0.05) \end{gathered}$ | $\begin{gathered} 0.10 \\ (0.05) \end{gathered}$ | $\begin{aligned} & \hline-0.04 \\ & (0.04) \end{aligned}$ | 2.4 |
| MOOC | Total <br> Time | . 19 | $\begin{gathered} 0.39 \\ (0.04) \end{gathered}$ | $\begin{gathered} 0.11 \\ (0.04) \end{gathered}$ | $\begin{gathered} -0.19 \\ (0.03) \end{gathered}$ | 3.5 | . 11 | $\begin{gathered} 0.19 \\ (0.05) \end{gathered}$ | $\begin{gathered} 0.13 \\ (0.04) \end{gathered}$ | $\begin{gathered} -0.09 \\ (0.03) \end{gathered}$ | 1.5 |
| Computing | Number Activities | . 42 | $\begin{gathered} 0.56 \\ (0.02) \end{gathered}$ | $\begin{gathered} 0.23 \\ (0.02) \end{gathered}$ | $\begin{gathered} -0.16 \\ (0.06) \end{gathered}$ | 2.43 | . 08 | $\begin{gathered} 0.28 \\ (0.02) \end{gathered}$ | $\begin{gathered} 0.05 \\ (0.02) \end{gathered}$ | $\begin{gathered} -0.05 \\ (0.02) \end{gathered}$ | 5.6 |
| OLI | Total Time | . 10 | $\begin{gathered} 0.19 \\ (0.03) \end{gathered}$ | $\begin{gathered} 0.27 \\ (0.03) \end{gathered}$ | $\begin{gathered} -0.03 \\ (0.003) \end{gathered}$ | 0.70 | . 02 | $\begin{gathered} 0.15 \\ (0.04) \\ \hline \end{gathered}$ | $\begin{gathered} 0.08 \\ (0.03) \end{gathered}$ | $\begin{gathered} -0.02 \\ (0.003) \end{gathered}$ | 1.9 |

Q2: More time in doing activities predicts better learning outcomes. The results of the logistic regression predicting Quiz and Exam performance using total time doing and reading are also presented in Table 3. The regression analyses showed that higher quiz and exam scores were predicted by spending more time doing ( $\beta=0.39, p<.0001$ and $\beta=0.19$, $p<.0001$, respectively), as well as reading ( $\beta=0.11, p=.006$ and $\beta=0.13, p=.001$, respectively).

Importantly, because reading requires, on average, more time than doing (see mean and standard deviations in Table 3 ), for each 1 standard-deviation ( 18.3 points, $17 \%$ total score) improvement in the total quiz score, students had to complete only a total of 18.45 hours of doing work during the 12 weeks of the course (or 1.5 hours/week), but 166.22 hours of reading work during the same period (or 13.8 hours/week). Similar improvements in final exam score require 16.16 hours of doing work but 168.77 hours of reading work over the entire course. Finally, similarly to what we saw when analyzing number of activities completed, spending more
time completing doing activities is more beneficial when students spend less time reading (and vice-versa; $\beta=-0.19, p$ $<.0001$ and $\beta=-0.09, p=.005$, respectively). This result further indicates that the benefits of the two types of activity is not additive.

Q3: Students overestimate the benefits of reading. Only a subset of students $(N=389)$ from the original sample described above also completed the "Learning Strategies" module (the module was optional). Table 4 shows the percentage of students who chose each possible study strategy as well as the percentage of students who chose exclusively each option. As it can be seen from the table, the large majority of students ( $93 \%$ ) chose "reading and completing the activities as they appear" ("Read-Do") as the best strategy. In fact, Read-Do was the most popular as the exclusive choice. Did students who chose a strategy focused on learning by doing spend more time doing than reading? To evaluate this question, we looked at the relative time spent doing vs. reading depending on the strategy the student chose.

Table 4: Percentage of students who selected each strategy as best for learning in the course.

| Course | N | Game the system |  | Do-Read |  | Read-Do |  | Read |  | Do |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Only <br> selection | Selected | Only <br> selection | Selected | Only <br> selection | Selected | Only <br> selection | Selected | Only <br> Selection |  |
| Psychology <br> MOOC | 389 | $8 \%$ | $8 \%$ | $32 \%$ | $5 \%$ | $93 \%$ | $61 \%$ | $6 \%$ | $0.05 \%$ | N/A | N/A |
| Computing <br> OLI | 950 | $3 \%$ | $0 \%$ | $36 \%$ | $4 \%$ | $94 \%$ | $40 \%$ | $4 \%$ | $0 \%$ | $36 \%$ | $0.05 \%$ |

For each student, we calculated the difference between total time doing and total time reading (doing-reading). More positive values in this measure indicate more time doing relative to time spent reading. We compared students who chose only the strategy "Do-Read", those who chose that strategy and another strategy, and those who chose any other strategy. Students who chose only the "Do-Read" strategy ( $M$ $=-7202, S D=6139$ ), or that strategy in addition to another ( $M=-7694, S D=4414$ ), spent relatively more time completing doing activities than those who did not choose that option $(M=-9746, S D=4983 ; t(386)=2.238, p=.026$ and $t(386)=3.63, p<.0001$, respectively).

## The "doer effect" in an online Computing Course

One of the goals of this research was to investigate whether engaging in learning by doing is an effective learning strategy for different types of knowledge. To extend the nature of the types of knowledge covered, we ran the same analyses with data from students' study behavior in an online version of a computing course. The content of this course is substantially different from the more expositive nature of an introductory psychology course. The course design followed the same overall principles and was similar to the Psychology course in terms of number of activities available to the students (see Koedinger et al., 2016 for details).
Sample. Our analyses include data from 2261 students enrolled in the online computing course "Information Systems" at University of Maryland University College (UMUC) using the OLI platform. We included in the analyses below students registered in OLI for whom quiz scores and a final grade were available. No pretest was available in this course.
Research questions and analyses plan. The same research questions and analyses plan as for the Psychology MOOC were used. The dependent measures used were the percentage correct across all quizzes and the final grade in number (1-5). Regression models do not include a pretest score.

## Results and Discussion

Descriptive measures of student behavior. Similar to what we found in the Psychology MOOC course, students spent on average more time reading than doing; conversely students completed more doing than reading activities (see Table 5).
Q1: More "doing" activities predicts better learning outcomes. Better quiz and exam scores are predicted by completing more doing activities ( $\beta=0.56, p<.0001$ and $\beta$ $=0.27, p<.0001$, respectively), as well as more reading activities $(\beta=0.22, p<.0001$ and $\beta=0.05, p=.02$, respectively; see Table 3). Moreover, we found similar ratios of benefit of doing over reading (2.43-5.6) as in the Psych MOOC, as well as a counter-intuitive interaction whereby the effect of doing activities is greater for lower amounts of reading, but only when predicting quiz scores (and viceversa; $\beta=-0.16, \mathrm{p}<.0001$ and $\beta=-0.05, p=.02$, respectively).

Table 5: Descriptive statistics for the main measures of students' study behavior and independent measures in the Computing course

|  | $M(S D)$ | Median | $25^{\text {th }}$ <br> Prctl. | $75^{\text {th }}$ <br> Prctl. |
| :---: | :---: | :---: | :---: | :---: |
| Read Time <br> (mins) | 13714 <br> $(21948)$ <br> Doing Time <br> (mins) | 2830 <br> $(7447)$ | 23679 | 613 |
| 19481 |  |  |  |  |
| \#Read Activities | $30(31)$ | 24 | 26 | 1749 |
| \#Doing <br> Activities | $70(55)$ | 64 | 14 | 136 |
| Percent Correct <br> Quizzes | $7.98(2.8)$ | 8.5 | 5.82 | 10.44 |
| Final Grade | 3.89 <br> $(1.25)$ | 4 | 3 | 5 |

Q2: More time in doing activities predicts better learning outcomes. Better quiz and exam scores are predicted by spending more time completing doing activities ( $\beta=0.19, p$ $<.0001$ and $\beta=0.14, p<.0001$, respectively), as well as more time reading $(\beta=0.27, p<.0001$ and $\beta=0.08, p=.01$, respectively; see Table 3). There is also an interaction, whereby the positive effect of more time spent in doing activities is larger when students spend less time reading (and vice-versa; $\beta=-0.03, p<.0001$ and $\beta=-0.02, p<.0001$, respectively). Although for the quiz scores we see a larger impact of more reading time compared to more doing time (as evidence by a ratio smaller than 1 ), for both quiz and exam scores it is clear that spending more time completing doing activities is more beneficial and efficient because it takes on average less time to complete more doing activities and this has an impact on performance. For example, for a 1 standard deviation ( $2.8 \%$ ) improvement in quiz scores, students would have to spend 70.9 hours over the duration of the course completing doing activities, but a whopping 326.13 hours reading - a gain of more than $20 \%$.
Q3: Students overestimate the benefits of reading. Among the subset of students who completed the question on what they believed was the best learning strategy $(N=950)$, the large majority of students indicated that they should read all the text and complete all activities as they show up ( $94 \%$, see Table 4). Only a small number of students indicated that they should focus mostly on the doing activities (36\%). Moreover, the students' a priori strategy preference did not predict their relative time spent doing, $F(2,938)<1, p=.558$, demonstrating that even students who completed more doing activities are probably unware of its benefits.

## General Discussion

The results of this research indicate that self-regulated learning by doing is associated with larger learning gains than learning by reading. More importantly, besides being a desirable learning strategy, it might also be more efficient. Across two different online courses focusing on different types of content, we found that students who completed more doing activities showed larger learning gains in shorter time (between 10 and $20 \%$ less time to achieve similar improvements). This result is important for two reasons: (1)
it emboldens efforts to include more active, doing activities in lessons, as an alternative to reading activities, and (2) it shows the generalizability of learning by doing to different kinds of materials, even materials often thought of as involving declarative, as opposed to procedural, knowledge.

Learning by doing as described here involved effortful (Roediger \& Karpicke, 2006), active engagement and knowledge manipulation by the student (Wieman, 2014), with timely feedback (Roediger \& Karpicke, 2006). All these properties have been associated with better learning outcomes compared to passive learning situations such as reading. Any of these factors might have contributed to the benefits of spending more time completing doing activities. Interestingly, the benefits of learning by doing were larger when students spent less time reading, suggesting that the two types of activity might be non-additive. An interesting hypothesis for future research is whether learning by doing could replace some or all of the learning that takes place from reading. Can effective learning of declarative knowledge be done exclusively by doing with feedback?

Importantly, we found that students do not realize the potential of learning by doing. Students seem to overestimate the value of explicit, verbal, learning and underestimate the value of active learning, as seen by their overwhelming support for strategies that emphasize reading and weak support for strategies that emphasize doing. Similar dichotomies between best learning outcomes and students $a$ priori judgements of best study practices have been described before (see Roediger \& Karpicke, 2006), and underscore the important role of familiarizing students with empirically tested best-practices.

Finally, the naturalistic character of the data and the approach used here have great potential. Natural datasets (such as the two used in this investigation) are increasingly available and allow for a wider investigation of the generalizability, effectiveness and adequacy of learning methods, theories, and approaches developed in the laboratory. This approach can play a key role for the future of learning science because of the novel insights that can only be gained from studying how learning takes place in natural contexts by their natural agents (Jones, 2016). However, admittedly, the research presented here does not allow us to establish causal links or discriminate between alternative theories of why learning by doing is a more efficient learning strategy. It is possible that the differences in learning by completing reading and doing activities presented here are due to a third variable; though previous research suggests that might not be the case (Koedinger et al., 2016). Nonetheless, the research presented here can stimulate future controlled studies that establish causal links, and investigate which characteristics of learning by doing in classroom contexts contribute to its benefits.

## Acknowledgments

This work was supported by a National Science Foundation grant (ACI-1443068) toward the creation of LearnSphere.org and by funding from Google. All data presented here is available from LearnShepere.org's DataShop (https://pslcdatashop.web.cmu.edu).

## References

Anderson, J. R., \& Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. Advances in Instructional Psychology, 5, 1-34.
Carvalho, P.F., Braithwaite, D.W., de Leeuw, J.R., Motz, B.A., \& Goldstone, R.L. (2016). An in vivo study of selfregulated study sequencing in introductory psychology courses. PLoS ONE, 11(3).
Ciccone, D.S., \& Brelsford, J.W. (1976). Spacing repetitions in paired-associate learning: Experimenter versus subject control. Journal of Experimental Psychology: Human Learning and Memory, 2(4), 446.
Jones, M. N. (2016). Developing cognitive theory by mining large-scale naturalistic data. In M. N. Jones (Ed.), Big Data in Cognitive Science. New York: Taylor \& Francis.
Karpicke, J.D., \& Blunt, J.R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. Science, 331(6018), 772-775.
Karpicke, J.D., \& Roediger, H.L. (2008). The critical importance of retrieval for learning. Science, 319(5865), 966-968.
Karpicke, J.D., Butler, A. C., \& Roediger III, H.L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own?. Memory, 17(4), 471-479.
Koedinger, K.R., Corbett, A.T., \& Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science, 36(5), 757-798.
Koedinger, K.R., Kim, J., Jia, J.Z., McLaughlin, E.A., \& Bier, N.L. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In Proceedings of the Second (2015) ACM Conference on Learning@ Scale (pp. 111-120). ACM.
Koedinger, K.R., McLaughlin, E.A., Jia, J.Z., \& Bier, N.L. (2016). Is the doer effect a causal relationship? How can we tell and why it's important. In Proceedings of the Sixth International Conference on Learning Analytics \& Knowledge (pp. 388-397). ACM.
McDaniel, M.A., Anderson, J.L., Derbish, M.H., \& Morrisette, N. (2007). Testing the testing effect in the classroom. European Journal of Cognitive Psychology, 19(4), 494-513.
Roediger III, H. L., \& Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science, 1(3), 181-210.
Sweller, J., \& Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. Cognition and instruction, 2(1), 59-89.
Wieman, C. E. (2014). Large-scale comparison of science teaching methods sends clear message. Proceedings of the National Academy of Sciences, 111(23), 8319-8320.
Zhu, X., Lee, Y., Simon, H. A., \& Zhu, D. (1996). Cue recognition and cue elaboration in learning from examples. Proceedings of the National Academy of Sciences, 93(3), 1346-1351.

# A Model of Cognitive Control in the Wisconsin Card Sorting Test: Integrating Schema Theory and Basal Ganglia Function 

Andrea Caso<br>(andrea@andreacaso.com)

Richard P. Cooper<br>(R.Cooper@bbk.ac.uk)

Centre for Cognition, Computation and Modelling<br>Department of Psychological Sciences, Birkbeck, University of London<br>Malet Street, London, WC1E 7HX, United Kingdom


#### Abstract

We present a schema-based model of a classic neuropsychological task, the Wisconsin Card Sorting Task (WCST), where competition between motor and cognitive schemas is resolved using a variation of a neuroanatomically detailed model of the basal ganglia (Gurney et al., 2001). We show that the model achieves a good fit with existing data at the group level, and correctly identifies two distinct cognitive mechanisms held to underlie two distinct types of error. However, at the individual level, the correlations amongst other error types produced by the model differ from those observed in the human data. To address this, we cluster participant performance into distinct groups and show, by fitting each group separately, how the model can account for the empirically observed correlations between error types. Methodologically, this demonstrates the importance of modelling participant performance at the sub-group or individual level, rather than modelling group performance. We also discuss implications of the model for the WCST performance of elderly participants and Parkinson's patients.


Keywords: schema theory; contention scheduling; basal ganglia; Wisconsin Card Sorting Task; modelling individual performance

## Introduction

Schema theory is a framework based on the idea that behaviour in many areas depends on abstractions over instances, i.e., schemas. In these abstract terms, schema theory is very general. It has been applied in domains ranging, for example, from event memory (Bartlett, 1932) to motor control (Schmidt, 1976). Norman and Shallice (1980) applied the theory in the domain of routine sequential action. Their theory proposes that action schemas work in a cooperative or sequential fashion, but also compete with each other for activation. While schema theory is helpful in representing functional interactions in the action-perception cycle, it is not committed to a specific neural implementation. However, at the neural level the basal ganglia have been proposed as a good candidate for resolving competition between schemas in order to carry out action selection (Redgrave et al., 2001). In part this is because of their recurrent connections with the cortex.

In this paper we present a model of the Wisconsin Card Sorting Task (WCST) where competition between motor and cognitive schemas is resolved using a variation of a neuroanatomically detailed model of the basal ganglia. We
use a genetic algorithm to search the model's parameter space and obtain a good fit for the data. Further analysis of correlations between error types, however, suggests the need to model individual participant data. Yet for reasons of computational efficiency this is impractical. We therefore cluster participant performance into a small number of distinct groups (5) and run separate genetic algorithms to fit the groups individually. The results capture both group performance and correlations between error types across individuals.

## The Task and the Model

In the WCST, participants are required to sort a series of cards into four categories based on binary (i.e., correct / incorrect) feedback. Each card shows one, two, three or four shapes, printed in one of four colours, and there are four shapes (triangle, star, cross, circle). It is therefore possible to sort cards according to colour, number or shape. To succeed, participants must match each successive card with one of four target cards (One Red Triangle, Two Green Stars, Three Yellow Crosses, Four Blue Circles), and use the subsequent feedback to discover the appropriate rule, but once they have discovered the rule (as indicated by a succession of 10 correct sorts), the experiment changes the rule without notice. The task yields a number of dependent measures, including the number of rules obtained (with a deck of 64 cards), the number of cards correctly sorted, the number of perseverative errors (where negative feedback is ignored) and the number of set-loss errors (where the participant fails to stick with a successful rule).

The model comprises three cognitive schemas and four motor schemas (see Fig. 1). ${ }^{1}$ Cognitive schemas represent the selection rules (Sort by Colour, Sort by Number, Sort by Shape) while the four motor schemas represent the acts of putting the stimulus card below each of the four target cards. Each schema has an activation level that varies over time as a function of input from various sources. Cognitive schemas are fed by an external channel that changes by a fixed amount according to external positive/negative feedback. Motor schemas are fed by cognitive schemas, and this signal is rule-dependent. If, for instance, the stimulus card displays

[^43]

Figure 1: Schematic of the model, not showing competition between schemas. Cognitive schemas (top row) send signals to the motor schemas (bottom row)


Figure 2: Schematic of the competition between schemas. The basal ganglia units compute the amount of inhibition that each schema receives given the activation of the others.
three red circles, the colour schema will excite the fourth motor schema (Four Blue Circles), the shape schema will excite the third motor schema (Three Yellow Crosses), and the colour schema will excite the first motor schema (One Red Triangle). Motor schemas are also fed by environmental cues depending on the stimulus card feature. Thus, when cognitive schemas are not strong enough to influence motor schemas, action selection may be driven by stimulus features only.

This simple model is complemented by a mechanism that implements and resolves competition between schemas within each hierarchical level: cognitive and motor schemas feed into two parallel computational mechanisms that each return a signal in the form of inhibition to the individual channels at each level (see Fig. 2 for an illustration at the cognitive level). In the brain, this competition between schemas is thought to be carried out by the basal ganglia


Figure 3: Schematic of the basal ganglia. Legend: Cortex-Thalamic complex (CTX-THAL), Striatum (STR), Subthalamic nucleus (STN), Globus Pallidus Internal/External Segment (GPi and GPe)
(Gurney et al., 2001). Corticobasal loops are mostly segregated (Alexander et al., 1986) and this is reflected in the model through the independence of information processed in the basal ganglia units at the two levels (cognitive and motor).

The model also implements a rudimental learning mechanism. This consists in a fixed change in signal to the cognitive schemas following a reward. Its purpose is to analyse how baseline levels of signal influence schema selection and ultimately, performance on the WCST. Manipulation of the thresholds of saturation functions in cortical units and associated basal ganglia units represent dopamine signalling in the cortex and in the basal ganglia, respectively. Therefore, the mechanism underlying cognitive control is a feedback-driven signal to the cognitive schemas.

## Computation in Individual Units

The model consists of 7 cortical units, 3 of which control cognitive operations and 4 of which control motor operations (see Fig. 1). These units correspond to schemas. Cognitive and motor units send their signal to their respective striatal units (see Fig. 3). Subthalamic units connect all units at the same hierarchical level (cognitive or motor), ensuring that the basal ganglia units act as a competitive suppressor of schemas as a function of the other schemas' outputs.

Individual units are connected as shown in Fig. 3. Their computations are shown below. In all cases, $u_{i}$ represents the entry signal to the unit, $a_{i}$ is the result of integration along the time domain, and $o_{i}$ represents the output of the individual units. The function $\sigma$ computes the sigmoid function of the input, ensuring output values are bounded between 0 and 1. Sigmoid functions have a fixed slope and threshold. Varying the threshold of cortical or striatal units alters the way competition between units is carried out, and can be considered a function of tonic dopamine present in the circuit. (In a separate simulation it has been shown that the level of external dopamine from the substantia nigra pars compacta ( SNpc ) unit can be simulated by varying the threshold of the saturation curve in the striatum $\left(\beta_{c t x}\right)$, without making use of an additional unit.)

## Cortical Units:

$$
\begin{gathered}
\mathbf{u}_{\mathrm{i}} \Leftarrow \sum_{\mathrm{j}} \mathbf{w}_{\mathrm{i}, \mathrm{j}} \cdot \mathbf{u}_{\mathrm{i}}+\mathbf{o}_{\mathrm{ext,i},}+\mathbf{o}_{\text {thal, }, \mathrm{i}} \\
\mathbf{a}_{\mathrm{i}}(\mathbf{t}) \Leftarrow \boldsymbol{\delta} \cdot \mathbf{a}_{\mathbf{i}}(\mathbf{t}-\mathbf{1})+(\mathbf{1}-\boldsymbol{\delta}) \cdot \mathbf{u}_{\mathbf{i}}(\mathbf{t}-\mathbf{1}) \\
\mathbf{o}_{\mathbf{i}} \Leftarrow \boldsymbol{\sigma}\left(\mathbf{a}_{\mathrm{i}}\right)
\end{gathered}
$$

Striatum (D1 and D2):

$$
\begin{gathered}
\mathbf{u}_{\mathbf{i}} \Longleftarrow \mathbf{o}_{\mathrm{ctx}, \mathbf{i}} \\
\mathbf{a}_{\mathbf{i}}(\mathbf{t}) \Leftarrow \boldsymbol{\delta} \cdot \mathbf{a}_{\mathbf{i}}(\mathbf{t}-\mathbf{1})+(\mathbf{1}-\boldsymbol{\delta}) \cdot \mathbf{u}_{\mathbf{i}}(\mathbf{t}-\mathbf{1}) \\
\mathbf{o}_{\mathbf{i}} \Leftarrow \boldsymbol{\sigma}\left(\mathbf{a}_{\text {strD1/D2,i}}\right)
\end{gathered}
$$

## Subthalamic Nucleus:

$$
\begin{aligned}
& \mathbf{u}_{\text {stn,i}}(\mathbf{t}) \Leftarrow \mathbf{w}_{\text {stn }} \cdot \mathbf{o}_{\text {ctx, } \mathbf{i}}+\mathbf{w}_{\text {gpe_stn }} \cdot \mathbf{o}_{\mathbf{g p e}, \mathbf{i}}(\mathbf{t}-\mathbf{1}) \\
& \mathbf{a}_{\text {stn }, \mathbf{i}}(\mathbf{t}) \Leftarrow \boldsymbol{\delta} \cdot \mathbf{a}_{\text {stn, } \mathrm{i}}(\mathbf{t}-\mathbf{1})+(\mathbf{1}-\boldsymbol{\delta}) \cdot \mathbf{u}_{\text {stn, } \mathrm{i}}(\mathbf{t}-\mathbf{1}) \\
& \mathbf{o}_{\mathrm{stn}, \mathrm{i}} \Longleftarrow \sigma\left(\mathrm{a}_{\mathrm{stn}, \mathrm{i}}\right)
\end{aligned}
$$

Globus Pallidus (External Segment):

$$
\begin{gathered}
\mathbf{u}_{\text {gpe }, \mathbf{i}} \Leftarrow \mathbf{w}_{\text {stn_gpe }} \cdot \sum_{\mathbf{i}} \mathbf{o}_{\text {stn,i}}+\mathbf{w}_{\text {strD2_gpe }} \cdot \mathbf{o}_{\text {strD2,i }} \\
\mathbf{a}_{\mathbf{g p e}, \mathbf{i}}(\mathbf{t}) \Leftarrow \boldsymbol{\delta} \cdot \mathbf{a}_{\mathbf{g p e}, \mathbf{i}}(\mathbf{t}-\mathbf{1})+(\mathbf{1}-\boldsymbol{\delta}) \cdot \mathbf{u}_{\text {gpe }, \mathbf{i}}(\mathbf{t}-\mathbf{1}) \\
\mathbf{o}_{\mathbf{g p e}, \mathbf{i}} \Leftarrow \boldsymbol{\sigma}\left(\mathbf{a}_{\text {gpe }, \mathbf{i}}\right)
\end{gathered}
$$

Globus Pallidus (Internal Segment):

$$
\begin{aligned}
& \mathbf{u}_{\text {gpi,i }}(\mathbf{t}) \Longleftarrow \mathbf{w}_{\text {stn_gpi }} \cdot \sum_{\mathrm{i}} \mathbf{o}_{\text {stn,i }}+\mathbf{w}_{\text {gpe_gpi }} \cdot o_{\text {gpe }, \mathbf{i}}(\mathbf{t}-1) \\
& +\mathbf{w}_{\text {strD1_gpi }} \cdot \mathbf{o}_{\text {strD1,i }}(\mathbf{t}-1) \\
& \mathbf{a}_{\mathrm{gpi}, \mathrm{i}}(\mathbf{t}) \Leftarrow \boldsymbol{\delta} \cdot \mathbf{a}_{\mathrm{gpi}, \mathrm{i}}(\mathbf{t}-\mathbf{1})+(\mathbf{1}-\boldsymbol{\delta}) \cdot \mathbf{u}_{\mathrm{gpi}, \mathrm{i}}(\mathbf{t}-\mathbf{1}) \\
& \mathbf{o}_{\mathrm{gp} \mathbf{i}, \mathbf{i}} \Longleftarrow \sigma\left(\mathbf{a}_{\mathrm{gp}, \mathrm{i}, \mathrm{i}}\right)
\end{aligned}
$$

## Thalamus:

$$
\begin{gathered}
\mathbf{u}_{\mathbf{i}} \Leftarrow \mathbf{o}_{\mathrm{gpi}, \mathbf{i}} \\
\mathbf{a}_{\mathbf{i}}(\mathbf{t}) \Leftarrow \boldsymbol{\delta} \cdot \mathbf{a}_{\mathbf{i}}(\mathbf{t}-\mathbf{1})+(\mathbf{1}-\boldsymbol{\delta}) \cdot \mathbf{u}_{\mathbf{i}}(\mathbf{t}-\mathbf{1}) \\
\mathbf{o}_{\mathbf{i}} \Leftarrow-\boldsymbol{\sigma}\left(\mathbf{a}_{\mathbf{i}}\right)
\end{gathered}
$$

## Feedback

Feedback takes place after each trial. If the selected response is correct, the external signals $o_{\text {ext }, i}$ to the cognitive units ${ }^{2}$ that correspond to the matched features are increased by a fixed amount $b_{l}$. If the selected response is incorrect, inputs to those units that correspond to the matched features are decreased by a fixed amount $b_{l}$.

## Simulation of Wisconsin Card Sorting Test

## Simulation of an Individual Task

To simulate the WCST, a virtual deck of 64 cards is produced, shuffled and presented to the model. All the units perform the computation outlined in the previous section. The first motor unit to reach a fixed activation value (measured as the area under the time-curve, rather than simply as a threshold) is selected. After the selection and feedback, a new card is presented. The resulting plot for activation of the cognitive units is shown in Fig. 4.

As can be seen in Fig. 4, when the first card is presented the system must work out that 'colour' is the first correct

[^44]

Figure 4: Activation of cognitive schemas during a complete run (involving sorting all 64 cards). Solid lines represent the actual activation while dashed lines represent the external input due to positive/negative feedback. Processing cycles are represented on the horizontal axis.
sorting criterion. Feedback alone is not sufficient, as the selected card may match more than one feature. Basal ganglia units intervene by supressing the inappropriate cognitive schemas, enabling the correct schema to be permanently selected. When the sorting criterion changes (after 10 correct responses) the system tends to perseverate for a short period of time, before selecting the correct criterion again. Feedback-dependent external activation and resolution of competition both play a role in activating the correct cognitive schemas. Whereas the activation of cognitive schemas is regulated by feedback, the activation of motor schemas is regulated by cognitive schemas and environmental cues.

## Parameters

The model has a number of parameters. One important parameter is the threshold of the saturation curve of the striatum, represented by the threshold of the sigma function applied to the striatal output $\left(\beta_{\mathrm{str}}\right)$. Extreme values of this parameter (substantially greater than or less than 0.5 ) disrupt the competition between schemas. When the threshold is too high schemas are driven by their input values and they undergo increasingly homogenous inhibition from the basal ganglia. This phenomenon is analogous to the Parkinson's disease dopamine depletion in the SNpc (Cooper \& Shallice, 2000).

## Dependent Measures

Performance was scored according to a range of measures as indicated in Heaton (1981). Completed Categories (CC) and Total Errors (TE) measure the overall performance. A Set


Figure 5: Comparison between Simulation and Data from neurologically healthy young participants. $Z$ values indicate the z score of the difference between human and simulated data for each dependent measure.

Loss Error (SL) is counted whenever an incorrect response is selected after 5 or more correct responses, where at least one is unambiguous (i.e., the card matches only one feature). A Perseverative response (PR) is counted whenever a response would have been correct under the previous rule. (A subject can score a perseverative response even before completing the first category: if three consecutive responses are made selecting the same sorting rule, that rule will be the criterion that the subject can perseverate to.) Those perseverative responses that are also incorrect responses are counted as Perseverative Errors (PE). Non-perseverative errors (NPE) are calculated as the Total Errors (TE) minus Perseverative Errors (PE).

## Results

Results for two sets of 48 participants ( 48 healthy young adults and 48 simulated participants) are depicted in Fig. 5. The figure compares the aggregate results from the simulation (Sim) with the aggregate data from the human participants (Data).

A genetic algorithm attempted to find the best parameters that produce low $t$ statistics and low $z$ statistics between data and simulation. Given the presence of a multitude of parameters that influence each other in a non-linear fashion, a perfect fit is unattainable. However, the model appears to do a good job in reproducing group mean and standard errors, as shown by the figure.

## Correlational Analysis

Analysing aggregate data is not sufficient to assess model performance, since a model should also aim to dissociate between psychological constructs (Cassimatis et al., 2008). Therefore, correlational analysis between the most informative variables (TE, PE, SL) was also performed, using bootstrapping and sampling the mean value to obtain 1000 points. Multiple runs of the sampling algorithm produce very similar results. Fig. 6 and Fig. 7 show the


Figure 6: Correlations - Neuropsychological Data


Figure 7: Correlations - Simulation
correlation matrices for these variables in both the human data and the simulation.

The correlation matrices show that the simulation correctly identifies that the mechanism that produces set loss error can be dissociated from the process that causes other types or errors. However, the simulation fails to reproduce the high correlation ( $r=.91, p<.01$ ) between Total Errors and Perseverative Errors. In addition, it displays a weak but significant negative correlation ( $r=-$ .31, $p<.01$ ) that is not present in the empirical data.

## Discussion

The model yields an adequate fit for young participants on the WCST. Computation in the model appears to be stable, in that minimal parameter variations do not disrupt functioning. The model also correctly reflects the independence between Set Loss Errors (SL) and Total Errors (TE) found in the human data, suggesting a dissociation in the cognitive processes that produce those errors.

However, the model is subject to several limitations. The lack of positive correlation between PE and TE in the simulation is both puzzling and concerning. One possibility,
however, is that this apparent failing reflects the implicit assumption that performance of the human participants can be modelled by a single set of parameter values (i.e., by a group of 48 virtual participants with identical cognitive characteristics). We explore this possibility in the following section.

## Grouping Data

## Introduction

In the light of the failure of the model to reproduce the empirically observed correlations between TE and PE, we analyse how data from young participants can be clustered into a small number of groups based on the three critical dependent variables reflecting errors (TE, PE, SL).

These three types of errors have been specifically chosen because they are most representative of performance failures. Data clustering was calculated using a k-means algorithm with $\mathrm{k}=5$ (purely for reasons of computational efficiency). Two points were excluded because they were outliers. The algorithm was initialised based on the observation of the spatial 3D distribution of points. The most distinctive features are the accumulation of points around the origin, the sparseness of points as total and perseverative errors increase, and an isolated cluster of points with SL equal to 1 .

Fig. 8 shows how the clustering of the groups and Table 1 shows mean and standard deviation of the dependent variables in the individual groups.

## Simulation

After clustering the groups, as outlined in Table 1, we run five genetic algorithms separately to determine best-fitting parameter values for each group. In each case, seven model parameters were initially randomised to values within their reasonable ranges, and model errors recorded. A t-value between the simulation's and the original experimental data was computed and its mean used as the inverse of the GA's fitness value. Table 1 shows performance errors of the simulation with the highest fitness and Fig. 9 shows a 3D representation of the individual values.

Table 1: Data Groups

| G | N | TE | PE | SL |
| :---: | :---: | :---: | :---: | :---: |
| $1 \bullet$ | 18 | $8.89(S D=2.03)$ | $6.22(S D=2.03)$ | $0(S D=0)$ |
| $2 \bullet$ | 13 | $14.85(S D=1.77)$ | $8.77(S D=1.92)$ | $0(S D=0)$ |
| $3 \bullet$ | 5 | $28.00(S D=1.73)$ | $18.40(S D=2.30)$ | $0(S D=0)$ |
| $4 \bullet$ | 7 | $14.71(S D=2.63)$ | $9.57(S D=0.53)$ | $1(S D=0)$ |
| $5 \bullet$ | 3 | $22.33(S D=2.08)$ | $11.67(S D=1.15)$ | $0(S D=0)$ |

Table 2: Simulation of the five clusters

| G | N | TE | PE | SL |
| :---: | :---: | :---: | :---: | ---: |
| $1 \bullet$ | 18 | $8.83(S D=1.38)$ | $5.89(S D=1.08)$ | $0.11(S D=0.32)$ |
| $2 \bullet$ | 13 | $14.31(S D=1.55)$ | $9.23(S D=1.17)$ | $0.08(S D=0.28)$ |
| $3 \bullet$ | 3 | $22.00(S D=5.00)$ | $7.33(S D=0.58)$ | $0.67(S D=1.15)$ |
| $4 \bullet$ | 6 | $15.00(S D=2.53)$ | $10.83(S D=1.72)$ | $0.5(S D=0.55)$ |
| $5 \bullet$ | 2 | $18.50(S D=2.08)$ | $9.50(S D=0.71)$ | $0.00(S D=0.00)$ |

## Discussion

Results from the simulation are shown in Table 2. In total, 4 outliers have been excluded from the analysis (2, 1 and 1 from categories 3, 4 and 5, respectively). These outliers may conceivably have been produced by the model's unstable response to increasingly higher parameter values. Clustering the participant data into a small number of more homogenous groups greatly increases the correlation between TE and PE ( $r$ increases from .04 to .50 , compared with the observed value of .92 ) and decreases the correlation between SL and TE/PE, improving the fit of the model in both respects. Fig. 10 displays the new correlation plots worked out combining all of the five simulations together.

## General Discussion

The model we presented combines a variation of the Cooper and Shallice (2000) model of action selection and a variation of the Gurney et al. (2001) model of the basal ganglia. One of the strengths of this combined model is the possibility to generalise it to other cognitive control tasks (e.g. Stroop task, Probabilistic Reversal Learning, Eriksen Flanker Task, etc.) and to accommodate the presence of


Figure 9: Simulated data with five clusters


Figure 10: Correlation between performance errors aggregating the values from five different set of parameters
units representing other brain areas where different computation is performed (e.g., amygdala, cerebellum), enabling the simulation of cognitive tasks in broader contexts (e.g. Emotional Stroop Task, WCST in cerebellar patients). In principle, this enhances the contention scheduling theory with neuroanatomical detail, allowing a more precise localisation of processes in a particular task, and integration with functional neuroimaging data. In addition, this implementation allows for the inclusion of two distinct learning mechanisms in the cortex and the basal ganglia: the current model can potentially be updated to a learning-based model by developing these mechanisms.

With respect to cortical learning, in the model as it stands, the supervisory system that controls how subjects respond to positive and negative feedback is fixed and consequently performance tends to be too robust to basal unit dysfunctions. This might be addressed by incorporating dynamic learning that allows supervisory control to vary according to the schemas' activations, resulting in low or high baseline levels of dopamine in the striatum having a greater impact on cognitive performance.

The present paper makes the case for modelling subgroup data (or, whenever possible, individual data), instead of aggregate results, and presents evidence of how data clustering improves the model overall fit. Clustering is especially advisable for models of higher-order cognition, where subjects tend to have variable attention and may use qualitatively different cognitive strategies.

A final conclusion emerges from two joint observations: First, fitting clusters with increasingly extreme error values becomes increasingly more problematic. Second, another set of simulations (not reproduced here) shows that damaging the cortical and subcortical units threshold does not seem to produce the level of decline in performance found in Parkinson's disease patients without dementia (Paolo et al., 1996). Since healthy older controls have a different performance profile than the younger controls against which the current model was assessed, the loss of dopaminergic
cells in SNpc does not alone explain the inferior performance in the elderly and PD patients ${ }^{3}$.

These two joint findings suggest that the cognitive mechanisms producing perseverative and set loss errors might be independent only for a small number of errors. As that number increases, these two mechanisms might be correlated and possibly causally related. New experimental data to confirm this hypothesis is warranted.

## References

Alexander, G. E., DeLong, M. R., \& Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annual Review of Neuroscience, 9(1), 357-381.
Bartlet, F. C. (1932). Remembering. A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge.
Bezard, E., Gross, C. E., \& Brotchie, J. M. (2003). Presymptomatic compensation in Parkinson's disease is not dopamine-mediated. Trends in Neurosciences, 26(4), 215-221.
Cassimatis, N. L., Bello, P., \& Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher order cognition. Cognitive Science, 32(8), 1304-1322.
Cham, R., Studenski, S. A., Perera, S., \& Bohnen, N. I. (2008). Striatal dopaminergic denervation and gait in healthy adults. Experimental Brain Research, 185(3), 391-398.
Cooper, R. P., \& Shallice, T. (2000). Contention scheduling and the control of routine activities. Cognitive Neuropsychology, 17(4), 297-338.
Gurney, K., Prescott, T. J., \& Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. Biological Cybernetics, 84(6), 401-410.
Hall, H., Sedvall, G., Magnusson, O., Kopp, J., Halldin, C., \& Farde, L. (1994). Distribution of D1-and D2-dopamine receptors, and dopamine and its metabolites in the human brain. Neuropsychopharmacology, 11(4), 245-256.
Heaton, R. K. (1981). A manual for the Wisconsin card sorting test. Western Psychological Services.
Norman, D. A., \& Shallice, T. (1980). Attention to action: Willed and automatic control of behavior (UCSD CHIP Report No. 99).
Paolo, A. M., Tröster, A. I., Blackwell, K. T., Koller, W. C., \& Axelrod, B. N. (1996). Utility of a Wisconsin Card Sorting Test short form in persons with Alzheimer's and Parkinson's disease. Journal of Clinical and Experimental Neuropsychology, 18(6), 892-897.
Schmidt, R. A. (1976). The schema as a solution to some persistent problems in motor learning theory. In G. E. Stelmach (Ed.), Motor Control: Issues and Trends (pp. 41-65). Academic Press: New York.

[^45]
# Distributional learning and lexical category acquisition: What makes words easy to categorize? 

Giovanni Cassani, Robert Grimm, Steven Gillis, and Walter Daelemans<br>Computational Linguistics and Psycholinguistics (CLiPS) Research Center<br>Department of Linguistics, University of Antwerp, 13 Prinsstraat<br>B-2000 Antwerpen, Belgium<br>\{name.surname\}@uantwerpen.be


#### Abstract

In this study, results of computational simulations on English child-directed speech are presented to uncover what distributional properties of words make it easier to group them into lexical categories. This analysis provides evidence that words are easier to categorize when (i) they are hard to predict given the contexts they occur in; (ii) they occur in few different contexts; and (iii) their contextual distributions have a low entropy, meaning that they tend to occur more often in one of the contexts they occur in. This profile fits that of content words, especially nouns and verbs, which is consistent with developmental evidence showing that children learning English start by forming a noun and a verb category. These results further characterize the role of distributional information in lexical category acquisition and confirm that it is a robust, reliable, and developmentally plausible source to learn lexical categories.


Keywords: Distributional bootstrapping; Lexical category acquisition; Statistical learning; Computational psycholinguistics; Language acquisition

## Introduction

Distributional bootstrapping (Maratsos \& Chalkley, 1980) is an influential account of how children start breaking into language, and specifically of how they start grouping words into lexical categories such as nouns and verbs. More specifically, it claims that children use patterns of co-occurrences across linguistic units, such as words and morphemes, to group words that share similar contexts. Several computational simulations have shown that distributional information is a rich, useful, and usable source of knowledge about lexical categories (Mintz, 2003; Redington, Chater, \& Finch, 1998; St. Clair, Monaghan, \& Christiansen, 2010). Moreover, a number of behavioral experiments have confirmed that children use this information to group words together (Mintz, Wang, \& Li, 2014; Reeder, Newport, \& Aslin, 2013).

Research on distributional bootstrapping has mostly focused on investigating which contexts constitute the best cues for the acquisition of lexical categories. Several proposals that have been put forward share the approach of grouping together those words that share similar contexts of occurrence, but differ in the starting assumptions and the types of contexts they evaluate. For example, Mintz (2003) suggested that frequent frames, i.e. trigrams consisting of two words flanking an empty slot $\left(a X_{\_} b\right)$, are a psychologically plausible and highly effective type of context for acquiring lexical categories. St. Clair et al. (2010), on the contrary, provided evidence that better categorization can be achieved by using bigrams $\left(a \_X+X \_b\right)$ that can be readily combined to obtain trigram level information.

This paper aims to explore distributional bootstrapping further and uses computational simulations to answer the following research question: what distributional properties of words make it easier to categorize them on the basis of the contexts they co-occur with? The relation between distributional properties of words and the extent to which these can be easily categorized in terms of lexical categories has been largely neglected in previous research, but characterizing it is important for two main reasons. Firstly, it generates predictions about the effect that several distributional properties of words have on lexical category acquisition in English speaking children: testing them can shed further light on the plausibility of distributional learning as an underlying mechanism for lexical category acquisition. Importantly, it is not enough that a model behaves like humans: a statistical analysis of what drives the model's behavior is necessary to assess whether it is driven by the same factors that affect human behavior. Secondly, it can help to constrain the development of psychologically motivated models of lexical category acquisition, by showing what information children are sensitive to when solving the task of grouping words into lexical categories.

In this work, computational simulations are used to carry out a categorization experiment whose outcome is used as the dependent variable in a regression analysis aimed to uncover the effect of several distributional properties of words on categorization accuracy. Results shed light and generate predictions on the mechanisms underlying distributional learning of lexical categories, and ultimately provide information to guide and constrain the development of psychologically motivated models of bootstrapping in language acquisition.

## Methods

## Corpora and pre-processing

In order to perform the computational simulations, transcribed interactions involving children and caretakers available in the CHILDES database (MacWhinney, 2000) were used. More specifically, the Manchester corpus (Theakston, Lieven, \& Pine, 2001) from the British English part, and the Suppes corpus (Suppes, 1974) from the American English part were selected, since they have both been widely used in previous research on distributional bootstrapping of lexical categories. The Suppes corpus consists of transcripts of one child, Nina, recorded from $1 ; 11$ to $3 ; 3$, while the Manchester corpus contains data of 12 children, recorded for varying periods within the age range $1 ; 8$ to $3 ; 0$. Both come with an au-
tomatic categorization in terms of Part-of-Speech (PoS) tags, which can be accessed on the MOR tier of the CHILDES annotation scheme. The child-directed speech from the corpora was pre-processed to deal with some aspects of the transcriptions. Two dummy symbols, \#start and \#end, were inserted at the beginning and end of each utterance. This manipulation is motivated by evidence that sentence boundaries provide useful distributional information (Freudenthal, Pine, \& Gobet, 2008). It also allows us to exploit every utterance from the corpus, including single word utterances: words occurring in isolation are considered to be occurring in the bigrams \#start_X and $X_{\_} \# e n d$, and in the trigram \#start_X_\#end.

Corpora from individual children were processed separately using a sliding window approach: starting from the first lexical element of the utterance, each word was considered as target, and all bigrams and trigrams occurring next to it were collected. These types of contexts were chosen given that they have been widely explored in previous research (Mintz, 2003; Monaghan \& Christiansen, 2008; St. Clair et al., 2010). As an example, consider the following utterance from the Manchester corpus: \#start are~v you~n going~v to $\sim f u n c t$ put $\sim v$ that $\sim a d v$ one $\sim n$ inside $\sim a d v$ ? \#end. The first target word, are, occurs in two bigrams, \#start_X and X_you~n, and two trigrams, \#start_X_you~n and $X_{-} y o u \sim n_{-}$going $\sim v$. For words in the middle of the utterance, three trigrams are available. The tags after the tilde indicate the lexical category to which each word belongs according to the automatic categorization. The original categories were collapsed to a coarser set, consisting of five categories: nouns ( $n$ ), including pronouns; verbs ( $v$ ), including auxiliaries, copulas, and nonfinite forms ${ }^{1}$; adjectives (adj), adverbs (adv), and function words (funct). The idea is to zoom in on the open classes, conflating the closed class words in a single category given that function words are categorized later in development. No lemmatization is performed, and all information about lexical categories is preserved ${ }^{2}$, although it is only used to evaluate whether categorization has been successful.

In order to minimize both the number of assumptions and that of possible decisions in the design of the experiment, all bigrams and trigrams are considered: some will turn out to be more informative to the categorization task than others, but the analysis of this aspect of the problem falls outside of the scope of this study. Larger $n$-grams are not considered due to the limited size of the corpora: they would be too infrequent to affect categorization.

## Experimental setting

A categorization experiment was carried out, in which words were clustered together based on the similarity of the contexts in which they occurred in corpora of English child-directed speech (Redington et al., 1998). Words that tend to occur

[^46]in the same contexts are considered to be more similar and clustered together: target words are categorized correctly if they are assigned the correct lexical category by the computational simulation. The experiment was performed using Memory-Based Learning (MBL, (Daelemans \& van den Bosch, 2005)), a class of machine learning algorithms which implements an exemplar-based strategy and categorizes new items using retrieval of or similarity to items stored in memory, with no explicit abstraction.

The categorization experiment consists of two main phases, which are referred to as training and testing in the paper. During training, co-occurrence counts between target words and contexts are collected on a portion of the input data and stored in memory. Each word is represented as a vector of counts, with each count indicating the co-occurrence frequency of the corresponding word and context. During testing, a new portion of the input is considered and the same procedure is applied. At the end of this second stage, the learner has created two matrices of co-occurrence counts. Each word from the test matrix is categorized by comparing its vector of co-occurrences with all the vectors from the training matrix, looking for the most similar one; the two are then clustered together. During learning, the model has no access to the correct lexical categories of the words and only groups them together based on their co-occurrence patterns, in an unsupervised way. At the end of the process, the category of two words that were clustered together is inspected: if they share the same lexical category, the word from the test set has been categorized correctly. In this framework, the only factor driving clustering is similarity, which is a well-documented cognitive mechanism in categorization (Sloutsky, 2003).

In order to divide each individual corpus into a training and a test set, utterances of child-directed speech were ordered chronologically and split in two parts: (i) the first $70 \%$ of the utterances were allocated for training; and (ii) the last $30 \%$ of the utterances were used as test set. To evaluate how different distributional properties interact with time, operationalized as a larger exposure to the input language, an incremental training approach was implemented. In detail, training started on the first $40 \%$ of all the utterances, then proceeded on the first $45 \%$, always increasing by 5 percentage points, up to the full training set ( $70 \%$ of the total utterances). The test set was kept constant to make sure that any change in performance came from the knowledge inferred from the training set and not by differences in the test set.

The TiMBL package (Daelemans, Zavrel, van der Sloot, \& van den Bosch, 2009) was used to carry out the simulation, using the default IB1 algorithm (Aha, Kibler, \& Albert, 1991) and cosine as a distance metric, because of its robustness to different frequencies in the co-occurrence vectors, and setting the number of nearest neighbors to 1 . Moreover, no feature weighting based on co-occurrence statistics from the training corpus was applied during the categorization experiment: this allows us to perform the categorization experiment without weighting contexts according to their informativity, avoiding
the effect of supervision on classification, which would be psychologically questionable and bias the results.

Importantly, no claim is put forward that children actually keep track of all available bigrams and trigrams, or that they implement an analogue of the IB1 algorithm with the chosen parameter setting. The interest of the current analysis is purely in the information that supports learning and in the analysis of the effects that distributional properties of words have on categorization, as operationalized using MBL.

## Statistical analysis

Four pieces of distributional information were computed for each word on the test set (last $30 \%$ of utterances of each corpus) and used as predictors in a regression model:

Token frequency: the log-transformed frequency count of each token. The transformation is motivated by evidence from Keuleers, Diependaele, and Brysbaert (2010) that lexical frequency effects are better captured by logtransformed frequency counts. A positive effect of frequency is expected (Ambridge, Kidd, Rowland, \& Theakston, 2015), since more frequent items are typically learned better than less frequent ones.

Contextual diversity: the log-transformed count of how many different contexts a word occurs in. A negative effect for contextual diversity is predicted: if a word occurs in many different contexts, its co-occurrence vector is noisy and it is harder to reliably group it with other words. This is the case, e.g., of function words, like conjunctions and determiners: they occur in all sort of contexts, making it hard to group them with similar words.

Average conditional probability: the average conditional probability of a word given all the contexts it occurs with. Consider a toy example where the context the $X$ occurs 100 times, 15 of which with the word cat: $p($ cat $\mid$ the_ $X)$, is thus 0.15 . Assume also that the word cat occurs 40 times in the context $a_{-} X$, which in turn occurs 200 times: $p\left(c a t \mid a_{-} X\right)$ is 0.2 . In order to obtain the average conditional probability for the word cat, $p(c a t \mid t h e \quad X)$ and $p\left(c a t \mid a_{-} X\right)$ are averaged, yielding 0.175 . This independent variable is predicted to have a negative effect on categorization: high conditional probability means that the contexts in which a target word occurs do not occur with other words, making it hard to find shared contexts of occurrence between the target and other words.

Entropy: the entropy of the co-occurrence vector of a word (Shannon, 1948), normalized by the number of contexts it occurs with, so that entropy lies between 0 and 1 . The entropy of a word is low when it occurs in the same context the majority of the times, while the more even the distribution of co-occurrences for a word, the higher its entropy. Entropy relates to diversity and its effect should go in the same direction: the more a word occurs equally frequently in the contexts it co-occurs in, the noisier its co-occurrence
vector and the harder it is to correctly group it with similar words. Importantly, normalized entropy provides a related but different piece of information than contextual diversity: the normalization ensures that the number of different contexts a word occurs in does not affect entropy.

A further independent variable was considered for both words and contexts, i.e. time, operationalized as the amount of training input on which the computational simulations were trained: time goes from 0 (i.e. $40 \%$ of all utterances in the corpus used as training set) to 6 ( $70 \%$ of all utterances in the corpus used as training set). Time should have a positive effect, since exposing the model to more input language should provide more reliable and robust information about co-occurrence patterns.

The analysis was restricted on words that appeared in all 13 individual corpora ( 12 from the Manchester corpus and 1 from the Suppes corpus), to reduce the effect of idiosyncrasies and focus on general patterns. All words with a token frequency of 1 were also excluded from the analysis, because when this is the case, contextual diversity and entropy are fully determined. If a word occurred only once, then it also occurred in only one context (diversity of 1 ), and its entropy is 0 , because the full probability mass is on the only context the word occurred in.

In order to analyze how easy it is to categorize a word, logistic mixed-effects models (Baayen, Davidson, \& Bates, 2008) were fitted using the "lme4" package in R (D. Bates, Maechler, Bolker, \& Walker, 2015). Random intercepts for corpus ( 13 levels) and word ( 456 levels, i.e. the single words that survived the filtering steps just detailed) were included. The categorization outcome of each word was used as a binary dependent variable, with each correctly categorized word coded as 1 . Covariates were included in a step-wise fashion, according to the improvement in fit measured by the Akaike Information Criterion (AIC, (Akaike, 1973)).

## Results

The best converging logistic mixed-effects model included main effects for average conditional probability, entropy, time, and contextual diversity. Adding a main effect for token frequency resulted in the model not converging. Two-way interactions between time and conditional probability, entropy, and lexical diversity were tested; however, when these were entered, the model did not converge. Table 1 provides the $\beta$ s estimated for this model, expressed on the log-odds scale, while Figure 1 represents the effects graphically, with accuracy expressed as proportion. The final model resulted in a marginal $R^{2}$ of 0.055 and in a conditional $R^{2}$ of 0.913 , suggesting that while the effect of predictors is significant, they do not explain much variance in the data. This is further addressed in the discussion.

As predicted, the average conditional probability of a word given the contexts in which it occurs has a strong negative effect on the estimated accuracy ( $\beta=-12.17, t=-11.56, p<$ 0.001 ), and the same is true for the entropy of the distribu-

Table 1: Mixed-effects model fitted to analyze what distributional properties make words easier to categorize. Estimates (Est.) and standard errors (Std. Err.) are provided on the logodds scale. (Cond. Prob.: average conditional probability of words given contexts; Cont. Div.: contextual diversity.

| Ind. Vars. | Est. | Std. Err. | $z$ | $p$ val. |
| :--- | ---: | ---: | ---: | ---: |
| (Intercept) | 14.185 | 1.298 | 10.928 | $<.001$ |
| Cond. Prob. | -12.170 | 1.053 | -11.560 | $<.001$ |
| Entropy | -11.027 | 1.215 | -9.077 | $<.001$ |
| Time | 0.078 | 0.011 | 6.838 | $<.001$ |
| Cont. Div. | -0.893 | 0.255 | -3.509 | $<.001$ |

tion of co-occurrence counts of a word over all the contexts it occurs in ( $\beta=-11.027, t=-9.077, p<0.001$ ). Time has a significantly positive effect $(\beta=0.078, t=6.838, p<0.001)$, showing that the clustering algorithm is actually exploiting the larger amount of input language to better group similar words together. Finally, contextual diversity has a significant negative effect ( $\beta=-0.893, t=-3.509, p<0.001$ ), suggesting that words are easier to categorize when they occur in fewer contexts, matching the initial hypothesis. As it was reported, adding frequency resulted in convergence issues: this is most likely due to the filtering step. It is possible that surviving words had similar frequency counts, making it impossible for the model to find sufficient variation to estimate the effect of token frequency on categorization accuracy, once contextual diversity already entered the model (since it improved the fit more than token frequency).

## Discussion

The results that have been presented point to a relation between distributional properties of words and the degree to which it is easy to categorize them into lexical category. The easiest words appear to (i) be on average hard to predict given the contexts in which they occur; (ii) have a very skewed distribution of co-occurrence counts with the contexts they occur in, meaning that they tend to occur most often in one or few contexts; and (iii) tend to generally occur in few contexts.

First, being able to predict a word given the contexts it occurs in is detrimental to categorization. This entails that effective categorization depends on some uncertainty in the co-occurrence patterns of words and contexts. Since categorization works on similarity (Sloutsky, 2003), two words can only be grouped together if they occur in the same context, i.e. they have something in common. The negative effect of conditional probability of words given contexts also points to a feature that contexts should have in order to be useful and usable, namely that they need to occur with more than one word. As a matter of fact, the conditional probability of words given contexts is computed by dividing the co-occurrence count of the word and the context by the frequency count of the context itself. For the average conditional probability of word given context to be low, each context must occur with other words


Figure 1: Main effects, with confidence bands, of average conditional probability, entropy, time, and contextual diversity on how easy it is to categorize a word in terms of lexical categories. The order, from top to bottom, reflects the improvement in fit brought by each predictor. The $y$-axis represents probabilities estimated from the log-odds reported in Table 1. Each axis is automatically scaled to provide a clear depiction of the effect. The plots were obtained using the effects package in R (Fox, 2003).
a substantial amount of times. This hypothesis fits evidence provided by Matthews and Bannard (2010) that children find it easier to group words together when these occur in contexts that, in turn, occur with several different words.

The negative coefficients of entropy and contextual diversity complement the negative effect of average conditional probability: the latter indicates that words are easier to categorize when they tend to occur in contexts only a fraction of the times the contexts themselves occur. $\beta$ s for normalized entropy and contextual diversity, on the contrary, tell that words are easier to categorize when they tend to occur most often in one or few contexts. The ideal situation is thus that of a word that always and only occurs in a single context, which however occurs with many other words, which also only occur in that context (to reduce noise). The effects of entropy and contextual diversity indicate that uncertainty in word-context co-occurrence patterns is necessary at the context level but detrimental at the word level: words need to occur in few contexts for effective categorization. This is likely due to the fact that when contextual diversity and entropy are high, the co-occurrence pattern of a word can be very noisy.

The distributional properties that make a word easier to categorize are rather distinctive of content words, especially nouns: knowing a context, e.g. a determiner, it is hard to predict exactly which noun will appear next to it, because many different nouns (and some adjectives) are possible, which translates into a low conditional probability of words given contexts. Moreover, it is likely that a noun occurs with one of the few determiners or possessive pronouns of the English language, thus scoring low on contextual diversity, and that most of the times it occurs with just a couple of specific determiners or possessive pronouns, scoring low on entropy. In order to get a grasp of which lexical categories easier words belonged to, those words that were categorized correctly for at least $80 \%$ of the 13 individual corpora at the last stage of training were selected. This analysis highlighted 127 such words: 2 function words, 101 nouns, and 24 verbs. This shows that the distributional properties of words that make them easier to categorize strongly correlate with lexical categories, and that the same features are a possible candidate to explain why certain lexical categories are formed earlier than others ${ }^{3}$. Furthermore, the majority of the 51 words that are never categorized correctly predominantly consists of function words (26) and adverbs (18), the categories that are learned later in development (E. Bates, Dale, \& Thal, 1995). The observation that nouns are categorized best also relates to the observation that children form a productive noun category earlier than any other category (Tomasello, 2000). The reported evidence lends support to the hypothesis that the so-called noun bias can be traced back to the distributional properties of words belonging to different lexical categories (Cassani, Grimm, Daelemans, \& Gillis, submitted), showing

[^47]that regardless of the fact that the set of target words contained an equal number of nouns and verbs, noun categorization is more effective.

The reported evidence also parallels and complements results about word learning, which suggest children find it easier to learn words (particularly nouns) when they occur in a variety of different contexts (Hills, Maouene, Riordan, \& Smith, 2010). While a comprehensive experiment is still lacking that explicitly contrasts the effect of contextual diversity on word learning and categorization, it emerges that this factor impacts both phenomena, although in opposite directions. While a higher contextual diversity is beneficial for word learning, it is detrimental to word categorization, as appears from the statistical analysis reported here. Further research about the interplay between different frequency effects (Ambridge et al., 2015) is needed to clarify to what extent distributional learning drives and explains language acquisition in its many different aspects and sub-tasks.

Lastly, this study investigated a fully distributional explanation of the developmental pattern of lexical category acquisition. However, the low $R^{2}$ shows that the distributional properties we investigated leave a substantial portion of variance unexplained, calling for further research on which properties affected the machine learner and whether these also influence children during lexical category acquisition. Moreover, current research has highlighted the importance of other sources of information during lexical category acquisition and word learning (Roy, Frank, DeCamp, Miller, \& Roy, 2015), including morphology, phonetics, semantics and prosody (Monaghan \& Christiansen, 2008). The influence of these sources of information should be further analyzed to complement research on distributional bootstrapping.

Summarizing, this study provided evidence about the effect of different distributional properties of words on the acquisition of lexical categories from distributional information. Conditional probability, entropy, and contextual diversity have a negative effect on categorization accuracy. Words with these features tend to be content words, mostly nouns, which also appear to be the words children start grouping earlier and most effectively. Future studies should assess the cross-linguistic validity of these findings, to understand whether the same distributional properties have similar effects in typologically different languages. Moreover, a similar approach - performing statistical analysis on the outcome of computational simulations - could be used to investigate what distributional properties make contexts more useful. Finally, other computational models should be tested, to compare their outcome to developmental data and shed light on which architectures are closer to what children actually do.

## Conclusion

The evidence presented in this study shows that specific distributional properties of words determine how easy it is to cluster them together based on the similarity of their cooccurrence patterns. In detail, words are easier to categorize
(i) when they are hard to predict given the contexts they occur in, (ii) when they generally occur in few contexts, and (iii) when they tend to occur more often in one context, having low entropy. This study extends previous research on distributional bootstrapping by providing evidence that distributional properties also affect which words are categorized more easily and which lexical categories are formed earlier.

## Acknowledgments

This research was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp.

## References

Aha, D. W., Kibler, D., \& Albert, M. K. (1991). Instancebased learning algorithms. Mach Learn, 6(1), 37-66.
Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov \& F. Csáki (Eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR. Budapest, Hungary: Akadémiai Kiadó.
Ambridge, B., Kidd, E., Rowland, C. F., \& Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. J Child Lang, 42(2), 239-273.
Baayen, H. R., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang, 59(4), 390-412.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2015). Fitting linear mixed-effects models using lme4. J Stat Softw, 67(1), 1-48.
Bates, E., Dale, P. S., \& Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher \& B. MacWhinney (Eds.), The handbook of child language (p. 96-151). Oxford, UK: Blackwell.
Cassani, G., Grimm, R., Daelemans, W., \& Gillis, S. (submitted). Distributional bootstrapping and the noun bias: Zooming in on developmental plausibility.
Daelemans, W., \& van den Bosch, A. (2005). Memory-based language processing. Cambridge, UK: Cambridge University Press.
Daelemans, W., Zavrel, J., van der Sloot, K., \& van den Bosch, A. (2009, 31 May 2010). Timbl: Tilburg Memory Based Learner, version 6.3. reference guide. (Tech. Rep.). Tilburg University.
Fox, J. (2003). Effect displays in r for generalised linear models. J Stat Softw, 8(15).
Freudenthal, D., Pine, J. M., \& Gobet, F. (2008). On the utility of conjoint and compositional frames and utterance boundaries as predictors of word categories. In V. Sloutsky, B. Love, \& K. McRae (Eds.), Proceedings of the 30th annual meeting of the Cognitive Science Society (p. 19471952). Austin, TX: Cognitive Science Society.

Hills, T. T., Maouene, J., Riordan, B., \& Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. J Mem Lang, 63(3), 259-273.

Keuleers, E., Diependaele, K., \& Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. Front Psychol, 1, 174.
MacWhinney, B. J. (2000). The childes project: Tools for analyzing talk. the database. (3rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.
Maratsos, M. P., \& Chalkley, M. A. (1980). The internal language of children syntax: The nature and ontogenesis of syntactic categories. In K. E. Nelson (Ed.), Children's language (Vol. 2, chap. 2). New York, NY: Gardner Press.
Matthews, D., \& Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of childdirected speech. Cognitive science, 34(3), 465-488.
Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. Cognition, 90(1), 91-117.
Mintz, T. H., Wang, F. H., \& Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (bigram) parts. Cognitive psychology, 75, 1-27.
Monaghan, P., \& Christiansen, M. H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), Corpora in language acquisition research: History, methods, perspectives (Vol. 6, p. 139-164). Amsterdam: John Benjamins Publishing.
Redington, M., Chater, N., \& Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. Cognitive science, 22(4), 425-469.
Reeder, P. A., Newport, E. L., \& Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. Cognitive psychology, 66(1), 30-54.
Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., \& Roy, D. K. (2015). Predicting the birth of a spoken word. Proc Natl Acad Sci U S A, 112(41), 12663-8.
Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27, 379-423.
Sloutsky, V. M. (2003). The role of similarity in the development of categorization. Trends in cognitive sciences, 7(6), 246-251.
St. Clair, M. C., Monaghan, P., \& Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. $\operatorname{Cog}$ nition, 116(3), 341-360.
Suppes, P. (1974). The semantics of children's language. Am Psychol, 29(2), 103-114.
Theakston, A. L., Lieven, E. V. M., \& Pine, J. M. (2001). The role of performance limitations in the acquisition of "mixed" verb-argument structure at stage i. J Child Lang, 28(1), 127-152.
Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74(3), 209-253.

# Knowledge transfer in a probabilistic Language Of Thought 

Samuel J. Cheyette, Steven T. Piantadosi<br>\{scheyett, spiantad\} @ ur.rochester.edu<br>Department of Brain and Cognitive Sciences<br>University of Rochester, Rochester NY, 14627 USA


#### Abstract

In many domains, people are able to transfer abstract knowledge about objects, events, or contexts that are superficially dissimilar, enabling striking new insights and inferences. We provide evidence that this ability is naturally explained as the addition of new primitive elements to a compositional mental representation, such as that in the probabilistic Language Of Thought (LOT). We conducted a transfer-learning experiment in which participants learned about two sequences, one after the other. We show that participants' ability to learn the second sequence is affected the first sequence they saw. We test two probabilistic models to evaluate alternative theories of how algorithmic knowledge is transferred from the first to second sequence: one model rationally updates the prior probability of the primitive operations in the LOT based on what was used in the first sequence; the other stores previously likely hypotheses as new primitives. Both models perform better than baselines in explaining behavior, with the human subjects appearing to transfer entire hypotheses when they can, and otherwise updating the prior on primitives.


Keywords: Knowledge transfer; Concepts; Language Of Thought; One-shot learning

## Introduction

One of the most remarkable capabilities of human cognition is the ability to rapidly create algorithms that are applicable to a new situation. For instance, an adult can quickly pick up a new card game, absorbing the rules and intuiting the strategy. Yet even the simplest card game is complex: it requires knowledge of basics like moves and turns; and it requires complex reasoning abilities, such as general-purpose strategic maneuvers in games. It seems more generally that humans' capacity to infer a lot from sparse data must be undergirded by a flexible array of useful concepts about many domains developed over a lifetime. For example, knowing about strategy in Texas Hold 'Em makes it possible to quickly pick up many other types of poker without reverting to a novice level, because requisite concepts across types of poker share similarities - betting, bluffing, winning hands. Yet this still leaves open the question: what are the representations and computations that make such effective transfer of abstract knowledge in this and myriad other domains possible?

Part of humans' adeptness in learning about new domains quickly may lie in their ability to map old conceptual structures to new ones, allowing them to infer abstract knowledge. This relational reasoning ability has often been characterized as "analogical" in nature (Markman, 1997), and many theories of analogical inference have been proposed on this basis (e.g. Gick \& Holyoak, 1980; Gentner, 1983; Holyoak \& Thagard, 1989; Hummel \& Holyoak, 1997). Gentner’s 1983 theory of "structure-mapping", formalized later as the "Structure Mapping Engine" (SME) (Falkenhainer et al., 1989), is an influential framework for describing analogical inference. On this account, situations or facts are given descriptions in
predicate logic, the components of which are either objects, relations, or attributes. The goal of a learner when presented with two situations is to make a mapping between these components by finding structural correspondences, and then inferring facts about one situation from the mapping to the other.

A commonality among SME and other theories of analogical transfer is their assumption of static knowledge representations. But structure only captures a limited subset of human knowledge. Other kinds of knowledge, such as learned processes or algorithms are untouched by these theories. In the poker example above, the algorithm of shuffling or bluffing may be transferred whole cloth to a new kind of poker. These abilities may be borrowed and incorporated into the algorithms that reason strategically. This kind of reuse would be much more like a programming language library-a location from which pieces of algorithms can be copied and reusedthan just a recognition of a correspondence of pieces. Indeed, in the same way that SME allows for powerful new inferences based on structure, transfer of algorithmic pieces could be part of the answer to how children eventually acquire algorithmically sophisticated representations: learners who can transfer algorithmic pieces need not construct entirely new representations each time they encounter a new domain.

Here, we experimentally and computationally test transfer of algorithmic components of representations by modeling concept learning as program induction over compositional functions, a system often called a "Language Of Thought" (Fodor, 1975). Under the LOT, a learner's job is to induce simple generative programs from primitive functions that match their observations of the world. In essence, this model treats learning as programming: there are a small set of "built in" operations that must be composed correctly in order to express richer algorithmic knowledge. This family of models has successfully been applied to explain human behavior in many rule-learning domains (Piantadosi \& Jacobs, 2016), including kinship and taxonomies (Kemp et al., 2008; Katz et al., 2008; Mollica \& Piantadosi, 2015), number (Piantadosi et al., 2012), causality, (Goodman et al., 2011), and words (Siskind, 1996; Piantadosi et al., 2008), among others.

Unlike structure mapping theories, LOT models are able to account for concept learning without requiring a significant amount of pre-developed knowledge. On the other hand, LOT models do not provide an account of humans' ability to transfer abstract knowledge between already-learned concepts. In general, it is an open question how LOT models can adapt their inductive biases and primitive representations through experience.

One possibility is that primitives are weighted in their prior according to their past utility as in the "Rational Rules" model
(Goodman et al., 2008). On this account, the prior is computed integrating out the production probabilities, allowing for a reduction in the penalty for repeated use of the same production rule. Among other things, this model has been used to explain selective attention effects, the finding that people tend to focus on as few features as possible to explain an observation.

Another possible way of explaining knowledge transfer in a LOT model is that upon learning a useful program, people store that program as a primitive for later re-use. This approach seems potentially more powerful than only updating priors over primitives themselves, as it could provide a basis for building increasingly complex, hierarchical conceptual structure. Indeed, Dechter et al. (2013) demonstrated how program recombination and re-use can facilitate and improve learning in the domains of both arithmetic and Boolean logic, using program induction over combinatory logic expressions. Others have explored models of sub-program re-use in Probabilistic Context Free Grammars, such as adaptor grammars (Johnson et al., 2006) and fragment grammars (O'Donnell et al., 2011). However, it has yet to be determined empirically if any of these models can explain human transfer of knowledge.

We ran a sequence-learning experiment to test human knowledge transfer, training people on one sequence and then testing them on a transfer sequence. We manipulated the congruity of the sequence pairs, corresponding to the abstract similarity of the training and transfer sequences. The results from our experiment suggest that having seen a congruous sequence in the past has a significant beneficial effect on accuracy. We modeled participants' learning curves in a probability-matching model and three probabilistic LOT models: a Rational Rules-type model that updates the prior of production rules in previously useful concepts; a model that adds previously useful concepts in full to its set of production rules; and a baseline model LOT model that does not update between training and transfer sequences. We compared the fit of each model to human data from our experiment. We found that the LOT model that re-uses high probability hypotheses from training provides the best fit to the data in the congruous condition, and the Rational Rules model provides the best fit in the incongruous condition. These findings suggest that learners transfer entire concepts when they can, and otherwise prefer previously used primitives.

## Experiment

We used a one-shot transfer learning paradigm in which participants were shown pairs of sequences which could either have come from similar LOT programs or not. To determine effects of knowledge transfer, we tested whether participants' overall accuracy on the the second sequence varied as a function of the first.

Participants 360 participants were recruited from Amazon Mechanical Turk, whose ages varied from 20 to 67 . They were paid 50 cents to complete the experiment, which took roughly $3-5$ minutes.

## Make a guess about the next color.



Figure 1: Example of display participants saw in the experiment.

## Method

Design The task involved a repeated binary choice, in which participants had to pick between two colored symbols (orange and blue) 15 times in learning both the training and transfer sequence. There were a total of 12 stimuli of which 6 were designated training sequences and 6 were designated transfer sequences. The manipulation was a fullfactorial between-subjects design with respect to the stimuli, so every possible combination of these sequences was tested, with only two shown to any given subject. An example of the display shown to participants is given in Figure 1. Note that every participant in both conditions saw the exact same training sequences - the differences in stimuli between conditions were only in the transfer sequence (the second of the two).
Stimuli The particular stimuli we chose were partly designed to allow for differing levels of compression in encoding in the LOT model. Some pairs of stimuli involve very simple repetitions, e.g. $\left(\left(A^{2} B\right)^{N}\right)$ and $\left(\left(A^{3} B\right)^{N}\right)^{1}$, which in our model are expressible in short hypotheses. Other patterns are not as efficiently compressible in our model, such as the repetition of $\left((A B)^{2} B\right)$. But, more importantly, they were designed such that the congruous pairs had abstract similarity, such that learning the first might help with learning the second. For instance, a congruous counterpart of the sequence $\left(A^{2} B^{3}\right)^{N}$ is the sequence $\left(A^{2} B^{4}\right)^{N}$, since a simple change to the description of one would result in the other. Every sequence in the first set had a congruous counterpart in the second set. The full set of stimuli is shown in Table 1, with congruous pairs adjacent.
Procedure Participants each saw two sequences, one after the other. Starting with no information about each sequence and ending with the entire sequence displayed on the screen, participants chose the symbol they thought was most likely given the previous values of the sequence they could see. After each guess, feedback appeared on the screen as to whether or not they were correct, and the correct symbol was placed at the end of the sequence on the screen. After participants completed the first sequence, it was erased from the screen, and they then completed the same task for the second sequence, starting from the beginning.

[^48]

Figure 2: This plot shows participants' accuracy in the transfer sequence over all 15 trials in the congruous (blue) and incongruous (red) conditions. The dots on top and bottom represent participants responding correctly or incorrectly at that trial, respectively. The decreasing transparency of squares of dots on top shows increasing numbers of correct responses, and the fact that the blue squares on top are less transparent than the red squares on top represents better learning in the congruous than incongruous condition. The two curves are the best-fit logistic regression predictions for the congruous and incongruous conditions.

| Training | Transfer |
| :--- | :--- |
| $\left(A^{2} B^{3}\right)^{N}$ | $\left(A^{2} B^{4}\right)^{N}$ |
| $B^{5}(A B)^{N}$ | $A^{4}(B A)^{N}$ |
| $\left(A^{2} B\right)^{N}$ | $\left(A^{3} B\right)^{N}$ |
| $\left(A B A B^{2}\right)^{N}$ | $\left(B A B A^{2}\right)^{N}$ |
| $B^{6} A^{N}$ | $A^{4} B^{N}$ |
| $B A^{2} B A^{3} B A^{4} \ldots B A^{i-1} B A^{i}$ | $A B A B^{2} A B^{3} \ldots A B^{i-1} A B^{i}$ |

Table 1: The full set of stimuli in our experiment is comprised of the first 15 symbols of each of these sequences. The congruous pairs are adjacent, and any non-adjacent pair is considered incongruous. The notation $X^{N}$ used here can be understood as N repetitions of sequence X . The bottom-most congruous pair is not as easily expressible in this way, but can be understood as incrementally increasing runs of one symbol interspersed by the other.

Participants were instructed to make their best guess about the next value of each sequence, even if they were unsure. They were told nothing about whether the two sequences were related, only that they both involved strings of colored symbols.

## Results

Our primary concern in analysis is to determine both the effect of sequence step and the effect of congruity on learning the transfer sequence. To determine both in a single analysis, we ran a logistic regression with both factors as fixed effects as well as random subject and sequence intercepts.

The results of this analysis revealed both a main effect of sequence step $(\beta=0.12, z=15.8, p<0.001)$ and congruity ( $\beta=0.70, z=4.69, p<0.001)$. The interaction between congruity and sequence step was not significant $(z=-0.12)$.

The fits from this analysis are shown in Figure 2, with the curves representing the best-fit regression lines for the congruous (blue) and incongruous (red) conditions.

Collapsing over all sequences and sequence steps in the transfer sequence, and just considering the average correct response given condition, those in congruous condition responded correctly more often $(M=0.75)$ than those in the incongruous condition ( $M=0.61$ ). The lack of interaction between congruity and sequence step implies that there is a lingering but constant beneficial learning effect in the congruous condition compared to the incongruous condition, but that the speed of learning in the two conditions is roughly the same.

## Model

The general modeling framework we used is a probabilistic Language Of Thought. In this approach there are a set of primitive, typed, and compositional operations, analogous to the statements that define programming languages (e.g. for Python 'if', 'elif', 'while', 'True', etc... would be considered primitive operations). The set of operations defines the "grammar", and the allowed rules for composing them are the rules for a Probabilistic Context Free Grammar (PCFG). The list of all possible compositions of production rules defines the entire hypothesis space. Since the number of possible hypotheses produced in our grammar is infinite, we use Metropolis-Hastings, a Markov Chain Monte Carlo (MCMC) sampling method, to provide a finite approximation to the entire space.

Each hypothesis H can be assigned a probability for any observed data D , which is computed via Bayesian inference: $P(H \mid D) \propto P(D \mid H) P(H)$. The likelihood, $P(D \mid H)$ is determined by how well the output of the hypothesis matches the
data. The prior probability $P(H)$ is computed according to the prior rule for PCFGs, which is the product of the prior probability of each primitive production rule $R$ composing $H: P(H)=\prod_{R \in H} P(R)$. The highest posterior probability hypothesis is therefore the most concise one that fits the data.

In the likelihood, we assume that hypotheses' output may be slightly noisy, giving each digit in the output sequence a 0.01 chance of being flipped. This likelihood formulation weights generated sequences higher in the likelihood in proportion to their similarity with the observed data. In addition to the intuitive plausibility of a similarity-weighting likelihood metric, this likelihood helps MCMC learn correct hypotheses by providing a graded (non-modal) posterior space. We performed no model fitting, and all parameters were used "out-of-the-box".

We ran a Metropolis-Hastings sampler for 100,000 steps and stored the top 100 hypotheses with the highest posterior found on each incremental prefix of the sequence.

## Hypotheses

In our model, hypotheses output binary sequences, corresponding to the binary colored symbols in the experiment. The production rules - which are the same across models are themselves operations on sequences and integers that return sequences. The production rules we chose were simply chosen to roughly be the minimal set necessary to concisely represent the sequences humans saw:

- $A^{\infty}$. Returns the symbol $A$ repeating unboundedly.
- $B^{\infty}$. Returns the symbol $B$ repeating unboundedly.
- Alternate $\left(I N T_{1}, I N T_{2}\right)$. Returns the sequence of alternations of $I N T_{1}$ and $I N T_{2}$. E.g. Alternate $(2,3) \Rightarrow\left(A^{2} B^{3}\right)^{\infty}$.
- Increment $\left(I N T_{1}\right)$. Returns the sequence of alternating repetitions of increasing length, starting from length $I N T_{1}$. E.g. Increment $(2) \Rightarrow A^{2} B^{3} A^{4} B^{5} \ldots A^{N-1} B^{N} \ldots$
- Append $\left(S E Q_{1}, S E Q_{2}\right)$. Returns $S E Q_{2}$ on $S E Q_{1}$. E.g. Append $\left(A^{2}, B^{2}\right) \Rightarrow A^{2} B^{2}$.
- Weave $\left(S E Q_{1}, S E Q_{2}\right)$. Returns $S E Q_{2}$ weaved between $S E Q_{1}$. E.g. Weave $\left(A^{2}, B^{2}\right) \Rightarrow(A B)^{2}$.
- Take $\left(S E Q_{1}, I N T_{1}\right)$. Returns the first $I N T_{1}$ items from $S E Q_{1}$. E.g. Take $\left((A B)^{5}, 2\right) \Rightarrow A B$.
- Invert $\left(S E Q_{1}\right)$. Returns the inversion of $S E Q_{1}$. E.g. $\operatorname{Invert}\left(B^{3} A\right) \Rightarrow A^{3} B$.

In these rules, $I N T$ could expand to the integers $1 \ldots 10$.

## Models of Learning

We implemented three different LOT models to test various possibilities about human concept learning from experience: a baseline model which does not update; a model that updates the prior of primitives; and a model that adds previous highposterior programs to its set of primitives. Each model was run on all 36 conditions in the experiment. Additionally, we
implemented a unigram model of the sequence to compare against the LOT models. The LOT models all started with the same production rules, which we assumed to have a uniform prior probability. All models were implemented using a freely available software package called LOTlib (Piantadosi, 2014).

Non-Updating Model In the baseline model, the primitives and their priors were fixed between the first and second sequence, and did not change.
Rational Rules Model We implemented a version of the Rational Rules model (Goodman et al., 2008), which updates the priors over primitives according to their posteriorweighted production rule count. This corresponds to a Dirichlet-Multinomial model, in which counts of each production rule in the Maximum A Posteriori (MAP) hypothesis from the training sequence are summed and subsequently used in computing the primitives' priors when learning the transfer sequence. Since a higher count corresponds to a decreased penalty for use in a tree, this is essentially a way of increasing the prior for primitive production rules useful in learning the training sequence. We assumed a uniform prior over production probabilities in the training sequence.
Re-Use Model Upon learning a concept, people may store and re-use this concept as a primitive. The way we captured this idea in our model was by placing the MAP hypothesis from the end of the training sequence as a primitive for generating hypotheses in the transfer sequence. The hypothesis space over primitives was re-normalized such that the primitives retained a uniform prior probability after this primitive was added.
Unigram Model We implemented a unigram model that responds proportionally to the probabilities of previous symbols. More specifically, we modeled this as a beta-binomial over the counts of the digits with a uniform prior. The counts were updated starting on the first sequence and continued through the second sequence. This is a baseline comparison, as it implements (smoothed) probability matching without taking into account any contingency.

## Results

Figure 3 shows the model's performance (with human data for comparison) at each step, collapsed over all sequences. The top panels display performance in the congruous condition and the bottom four show performance in the incongruous condition. It's worth noting again that these are predictions made with no model parameter tuning, but the rankorder speed of learning between models is unlikely to be affected by this. The first interesting thing to note is how well, and how quickly, each of the models learns in the congruous and incongruous conditions. The Re-Use model shows the greatest disparity between conditions, guessing accurately on average $66 \%$ of the time in congruous case and $54 \%$ of the time in the incongruous case, a difference of $12 \%$. This is substantially higher than the difference in the Rational Rules model (4\%), the unigram model (1\%), and the no-updating


Figure 3: Overall model correctness overtime in the congruous (top) and incongruous (bottom) conditions, collapsed overall all sequences. The human data is shown in red in each plot, as comparison. The dashed line is the just the constant of $\mathrm{y}=0.5$, for comparison.

| Condition | Analysis | No Update | Rat. Rules | Re-Use | Unigram |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Congruous | Mean Squared Error $(x 10)$ | 0.57 | 0.37 | $\mathbf{0 . 1 7}$ | 0.65 |
| Congruous | $R^{2}$ | 0.60 | 0.70 | $\mathbf{0 . 7 2}$ | 0.13 |
| Congruous | Log Likelihood | -853 | -796 | $\mathbf{- 7 3 6}$ | -870 |
| Incongruous | Mean Squared Error $(x 10)$ | 0.16 | $\mathbf{0 . 1 2}$ | 0.13 | 0.23 |
| Incongruous | $R^{2}$ | 0.67 | $\mathbf{0 . 7 5}$ | 0.75 | 0.56 |
| Incongruous | Log Likelihood | -4322 | $\mathbf{- 4 2 6 8}$ | -4280 | -4412 |

Table 2: Overall performance measured in Mean Squared Error, $R^{2}$, and Log Likelihood, for each of the models in both the congruous and incongruous condition. The best fit for each metric is bolded. Note that the log likelihoods can be compared like AIC values since there are no free parameters.
model ( $0 \%$ ). This difference in the re-use model is most similar to humans, who responded correctly $75 \%$ of the time in the congruous congruous and $61 \%$ of the time in the incongruous condition, a change of $14 \%$.

To more precisely compare the model and human fits for each sequence, we report the Mean Squared Error (MSE), $R^{2}$, and Log Likelihood to aggregate human responses, for each sequence and condition in Table 2. In both conditions, all the LOT models were significantly better fits than the unigram model. In the congruous condition, the Re-Use model was clearly a better fit than any other LOT model or the unigram model. The reason it out-performs all the other models in this case is primarily that none of the others learn the sequences fast enough. In the incongruous condition, the LOT models in this case perform more similarly than in the congruous condition, but the Rational Rules model provides a slightly better fit of the three according to each metric.

## Discussion

The fact that the Re-Use model has the highest accuracy in the congruous condition (and closest to human-level) suggests that it is a better model of how humans' inferences benefit from helpful experience. The Re-Use model also displays the greatest disparity in accuracy between the two conditions,
though still not quite as large in the gap in human performance between conditions ( $12 \%$ versus humans' $14 \%$ ). Interestingly, the models display much more similar learning curves in the incongruous case. This means that the disparity in performance in the two conditions may be entirely due to the relative benefit of congruous experience - insofar as it changes primitives or their priors beneficially - but not as much to hindrance from incongruous experience. If true, this would predict that humans would perform about as well on the transfer sequence with no training sequence at all as with an incongruous training sequence.

To understand the Re-Use model's performance, it is informative to look at the actual representations that allow it to learn more quickly than the other models in the congruous condition. For each sequence, the MAP hypothesis from the first sequence is used in the MAP representation of the second sequence by the final step. Indeed, it is often orders of magnitude higher in the posterior than any other hypothesis. For instance, consider the case where the model sees:

$$
\left((A B)^{2} B\right)^{3}
$$

as training followed by:

$$
\left((B A)^{2} A\right)^{3}
$$

as transfer. The MAP hypothesis for the training sequence is displayed in orange in Figure 4.

This hypothesis gets added as a primitive, which we can call $M A P_{1}$. The shortest program on the transfer sequence that fits the data by the final step (and before), is simply invert $\left(M A P_{1}\right)$, which is the entirety of the tree in Figure 4. This, of course, generates the inverse sequence generated by $M A P_{1}$, which is a simple and low-cost transformation when treating $M A P_{1}$ as a primitive. The tree representing the MAP hypothesis for the transfer sequence in the Re-Use model is much higher in the prior than the MAP representation both the Rational Rules model and the No-Update model construct, since it only uses two primitives, compared to their use of eight.


Figure 4: The Re-Use model's MAP hypothesis for generating repetitions of $\left((B A)^{2} A\right)$ in the congruous condition. The part in orange is the MAP hypothesis from learning the training sequence $\left((A B)^{2}\right) B$, and the blue is a transformation on it, treating it as a primitive production rule.

It is also interesting that the Rational Rules model provides the best fit in the incongruous condition, closely followed by the Re-Use model. This suggests that even when people can't transfer a whole concept, they still prefer using primitives of past hypotheses. One possibility to explore in the future is combining the Rational Rules and Re-Use models. Another potentially powerful model could account for partial sub-tree re-use. This would reflect the possibility that people not only store useful programs in their entirety, but store useful subprograms. This added flexibility in recombination has been modeled using adaptor grammars (Johnson et al., 2006) and fragment grammars (O’Donnell et al., 2011). But inference in these models is substantially more complicated than models considered in this paper, and the extent of human flexibility in this regard remains an open question.

## Conclusion

Our experiment showed that people benefit in learning a sequence given prior experience with an abstractly congruous sequence. By considering congruity as a function of similarity in LOT program-space, we can understand human knowledge transfer as changes in the representations and biases of LOT models. We showed that a LOT model that treats previously learned programs as primitive rules is the best fit to human data in the congruous condition. On the other hand, we found that the LOT model that rationally updates the prior
on existing production rules is the best fit in the incongruous condition. This provides evidence that people spontaneously transfer knowledge of both whole programs and their subcomponents when learning.
Acknowledgements We would like to thank Jenna Register, Fred Callaway, Frank Mollica, and colala for helpful comments and conversations.

## References

Dechter, E., Malmaud, J., Adams, R. P., \& Tenenbaum, J. B. (2013). Bootstrap learning via modular concept discovery. In 23rd international joint conference on artificial intelligence. Beijing, China.
Falkenhainer, B., Forbus, K. D., \& Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial intelligence, 41(1), 1-63.
Fodor, J. A. (1975). The language of thought. Harvard University Press.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155-170.
Gick, M. L., \& Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychology, 12(3), 306-355.
Goodman, N. D., Tenenbaum, J., Feldman, J., \& Griffiths, T. (2008). A rational analysis of rule-based concept learning. Cognitive Science, 32, 108-154.
Goodman, N. D., Ullman, T. D., \& Tenenbaum, J. B. (2011). Learning a theory of causality. Psychological review, 118(1), 110.
Holyoak, K. J., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive science, 13(3), 295-355.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological review, 104(3), 427.
Johnson, M., Griffiths, T. L., \& Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In Advances in neural information processing systems (pp. 641-648).
Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., \& Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. In Proceedings of the cognitive science society (Vol. 30).
Kemp, C., Goodman, N., \& Tenenbaum, J. (2008). Learning and using relational theories. In Advances in neural information processing systems (Vol. 20, p. 753-760).
Markman, A. B. (1997). Constraints on analogical inference. Cognitive science, 21(4), 373-418.
Mollica, F., \& Piantadosi, S. (2015). Towards semantically rich and recursive word learning models. In Proceedings of the cognitive science conference (Vol. 37).
O'Donnell, T. J., Snedeker, J., Tenenbaum, J. B., \& Goodman, N. D. (2011). Productivity and reuse in language. Proceedings of the 33rd Annual Conference of the Cognitive Science Society.
Piantadosi, S. (2014). LOTlib: Learning and Inference in the Language of Thought. available from https://github.com/piantado/LOTlib.
Piantadosi, S., Goodman, N., Ellis, B., \& Tenenbaum, J. (2008). A Bayesian model of the acquisition of compositional semantics. In Proceedings of the cognitive science society (Vol. 30).
Piantadosi, S., \& Jacobs, R. (2016). Four problems solved by the probabilistic language of thought. Current Directions in Psychological Science, 25(1), 54-59.
Piantadosi, S., Tenenbaum, J., \& Goodman, N. (2012). Bootstrapping in a language of thought: a formal model of numerical concept learning. Cognition, 123, 199-217.
Siskind, J. M. (1996, Oct-Nov). A computational study of crosssituational techniques for learning word-to-meaning mappings. Cognition, 61(1-2), 1-38.

# Cross-situational learning of novel anaphors 

Karen Clothier (kclothi1@jhu.edu)<br>Cognitive Science Department, 3400 N. Charles Street<br>Baltimore, MD 21218 USA

Satoshi Nambu (satoshi.nambu@monash.edu)<br>School of Languages, Literatures, Cultures and Linguistics, 20 Chancellors Walk Clayton, 3800 AUSTRALIA

Hajime Ono (hajime@tsuda.ac.jp)
Faculty of Arts, 2-1-1 Tsuda-machi, Kodaira-shi
Tokyo, 187-8577 JAPAN

Akira Omaki (omaki@uw.edu)<br>Department of Linguistics, Guggenheim Hall Box 352425<br>Seattle, WA 98195 USA


#### Abstract

Word learning research has shown that learners constrain the hypothesis space for word meanings by using multiple sources of information, such as cross-situational regularities of word-context co-occurrences or syntactic cues, like the number of arguments. These studies typically focus on word meaning development where these cues can be helpful but not necessary. As such, it sheds little light on the acquisition of anaphors, which requires tracking syntactic dependencies across situations. To test whether or how learners track this information, we conducted a novel anaphor learning experiment with English and Japanese speakers, manipulating cross-situational regularities in anaphors and their syntactic dependencies. Results show both English and Japanese speakers closely track the frequency of interpretive possibilities for novel anaphors. However, they demonstrate difficulties learning long-distance reflexives, which are compatible with either local or non-local antecedents. This suggests that successful anaphor learning requires more than cross-situational regularities of interpretive possibilities.


Keywords: anaphors; binding; language acquisition; statistical learning; word learning

## Introduction

Sentences (1) and (2) illustrate that the interpretation of local reflexives like himself and pronouns like him rely on structural relationships to their antecedents (the noun phrase that the reflexive or pronoun refers to, e.g. himself $=$ John in (1), but him = Bill in (2)).
(1) Bill said that John kicked himself.
(2) Bill said that John kicked him.

This complementarity suggests that these anaphors respect different locality constraints; informally, English reflexives can only be bound by antecedents in the same clause, whereas pronouns must not be bound by antecedents in the same clause (Chomsky, 1981).

Other languages have different types of anaphors. For example, Japanese has a long-distance (LD) reflexive zibun, and replacing himself with this LD reflexive in the Japanese
translation of (1) results in an ambiguous sentence, as zibun can refer to either Bill or John. This suggests that in Japanese and other related languages, the local domain for zibun is expanded to the whole sentence, thereby allowing both the local and non-local antecedents.

This cross-linguistic variation suggests that the interpretive possibilities for anaphors are not universally determined, and must be learned from language experience. Learning the interpretive possibilities of anaphors requires learners to infer the intended meaning of the utterance based on the utterance context. For example, the nature of the events described by (1) and (2) is different: one is reflexive, while the other is transitive. Thus acquiring anaphors requires the learner to simultaneously track the syntactic relations between the anaphor and its antecedent, as determined by the context of the utterance. Critically, learning that the LD reflexive can take either a local or nonlocal antecedent requires tracking such information across multiple situations in which the LD reflexive is used with either one.

Previous word learning research has investigated how learners use such cross-situational regularities or syntactic structures to learn word meanings. For example, 2.5 yearolds acquired the meaning of a novel verb (e.g. pim) dependent solely on the frequency with which the verb was presented with a particular video of a particular action (Scott \& Fisher, 2012; for similar work on nouns see Smith \& Yu, 2008). While there is an on-going debate over whether learners gradually update word meaning hypotheses across situations, or whether learners instead iteratively test and revise successive hypotheses across situations (see Medina, Snedeker, Trueswell, \& Gleitman, 2011), both lines of work critically point to the fact that information across trials contributes to the word learning process.

Much work has also explored how learners use syntactic frames to constrain the meaning of novel words (syntactic bootstrapping; see Gleitman, 1990). For example, Yuan and Fisher (2009) showed that when infants heard a series of
transitive sentences with two arguments (e.g. Jimmy blicked the cat) in the absence of contextual information, they reliably interpreted the novel verb as referring to a causative event that requires two participants. On the other hand, presenting a series of intransitive sentences led infants to interpret the novel verb as referring to a single participant event. Such work shows that learners can use information from syntactic structures to narrow their hypotheses about word meanings.

In sum, the work reviewed above shows that learners make use of both distributional and linguistic cues. However, few studies have investigated the mechanism of anaphora acquisition, which requires tracking the crosssituational regularities of syntactic relations themselves. Research on anaphora acquisition mechanisms not only fills this empirical gap, but also provides a novel window into the constraints on cross-situational, statistical learning mechanisms.

The present study uses a novel anaphor learning experiment to investigate if and how participants use crosssituational co-occurrences of word form, contextual information, and syntactic structure. During the learning phase, participants are presented with sentence-picture pairs that contextually constrain the intended meaning of the target English sentence like (1) or (2) with a novel anaphor. We used three novel anaphors, following the interpretive possibilities of local reflexive (botu), pronoun (sumu), as well LD reflexive (togu). We included these three anaphor types for three reasons. First, this combination of anaphors is attested in Japanese and other languages, and therefore provides an ecologically valid anaphor system. Second, for English speakers, the Japanese LD reflexive is a new category of anaphor, so this allows us to simulate an actual learning process. Third, the local reflexive and pronoun are expected to be readily learnable for English speakers since both exist in English, and are also presented with a single interpretive possibility across trials. Thus, these anaphors can be used to validate this novel experimental procedure.

In order to probe the effectiveness of distributional information, we manipulated the frequency of two interpretive possibilities for the LD reflexive (i.e., local antecedent, akin to the meaning of (1), or non-local antecedent, akin to the meaning of (2)) by creating three between-subjects learning conditions: a Balanced condition where there were $50 \%$ local and $50 \%$ non-local antecedents; an LD-majority condition where $80 \%$ of the time the antecedent was non-local, and only local $20 \%$ of the time; and a Local-majority condition, where $80 \%$ of the time the antecedent was local, and only non-local $20 \%$ of the time.

If anaphor acquisition relies on accruing information and testing hypotheses across situations, then we would expect that a learner in the Balanced condition would be best able to learn the Japanese-style LD reflexive; encountering both interpretive possibilities should maximize learners' chances to realize the optionality of local and non-local antecedents.

However, there are two additional biases that may affect how learners use distributional information. For example, it
has been proposed that in the absence of clear disambiguating information, learners are often biased to adopt syntactic structures that are easier to process in subsequent comprehension (e.g. Fedzeschkina, Newport, \& Jaeger, 2016; Hawkins, 1999). With respect to processing LD reflexives, it has been found that at least in processing of Chinese LD reflexives, readers are biased to access local antecedents due to constraints on the working memory mechanism (e.g., Dillon et al., 2014). If this bias extends to anaphor acquisition, learners may struggle to learn the availability of the non-local antecedent with the LD reflexive, whereas even a relatively small percentage of input supporting the local antecedent may be sufficient to learn the availability of the local antecedent. Under this account, learners may in fact represent our LD-majority condition as if it were a 'Balanced' condition, because this local binding bias would enhance the availability of the local antecedent while dampening that of the non-local antecedent. If this is the case, then learners in the LDmajority condition - instead of the Balanced - should be best able to realize the optionality of local and non-local antecedents with the LD reflexive.

Another potential source of bias, which may be particularly relevant to the present study, is the influence of anaphors in the participants' native languages. The second language (L2) acquisition literature has found evidence for strong first language influence on the L2 acquisition of anaphors. For example, Yuan (1998) showed that Japanese speakers learning the Chinese LD reflexive were more likely to accept non-local antecedents than English speakers learning Chinese. Conversely, Japanese speakers learning English struggled to rule out non-local antecedents for English reflexives in a similar task (Hirakawa, 1990). Together, these results suggest that learners tend to 'transfer' and expect the same anaphors in their L2. Given that our study explores how adult learners acquire a novel anaphor, anaphors in the native language may constrain what can be learned within a single experimental session.

To investigate the native language influence on anaphor acquisition, we conducted this experiment with English native speakers and Japanese native speakers. If prior knowledge affects novel anaphor acquisition, then we would expect a contrast between the English and Japanese group: English speakers should struggle to learn the optionality in the LD reflexive, whereas Japanese speakers should be able to correctly learn the novel LD reflexive based on their knowledge of zibun in their native language.

## Experiment

## Participants

Fifty-seven native English-speaking members of the Johns Hopkins University community participated in the experiment. They were compensated with course credit or $\$ 10$ cash. According to self-reports, none of the participants knew languages with LD reflexives. Participants were randomly assigned to one of the three learning manipulation
conditions. All participants completed a learning phase and then a test phase in the same session.

In addition, 57 native Japanese-speaking students from Tsuda College participated in the Japanese version of the same experiment. They were compensated $¥ 1000$. Performance in an English cloze task (adapted from Kobayashi, 2002) revealed intermediate syntactic and semantic knowledge in English (maximum possible score: $25 ; \mathrm{M}=15.529, \mathrm{SE}=4.130$ ), suggesting the relative dominance of Japanese over English.

## Materials

Learning Phase Trials The 78 sentences for the Learning Phase were English sentences with clausal embedding (e.g. John $\{$ said/remembered $\}$ that Susan combed $\{b o t u / t o g u / s u m u\}$ ), where botu, togu and sumu were the novel words used as the local reflexive, LD reflexive, and pronoun respectively. These anaphors were not marked for gender. Three main clause verbs (comb, wash, fan) were used once with both said and remembered. Using these six sentence frames, three male and three female character names were permuted to create 24 sentences containing the local reflexive and 24 containing the pronoun. For the local reflexive sentences, one sentence for each main clause verb was replaced with a mono-clausal sentence with the same verb. Taking these local reflexive sentences (including the mono-clausal sentences) and adding 6 additional embedded clause sentences following the same procedure described above created the 30 sentences containing the LD reflexive.

In order to make the intended interpretation of each sentence clear, throughout the experiment pictures were paired with each sentence to form a trial. Local reflexive sentences were depicted with pictures showing the syntactically local noun phrase as the antecedent of the anaphor; pronoun sentences were depicted with pictures showing the syntactically non-local noun phrase as the antecedent of the anaphor (as in Figure 1). Critically, LD reflexive sentences were depicted with either type of picture; referred to as the local antecedent in the former case, and the non-local antecedent in the latter.


Figure 1: An example picture-sentence pair with togu (the novel LD reflexive), which illustrates the non-local antecedent (Adam washing David). The local antecedent version of this trial type would picture Adam washing himself inside the speech bubble.

Learning Conditions The picture-sentence pairs were used to create three different distributions of LD reflexive interpretations. For the Balanced condition, 15 LD reflexive sentences were paired with local antecedent pictures (including the necessarily local mono-clausal sentences, which appeared with a single character performing an action to themselves, with no speech bubble); the remaining 15 sentences appeared with non-local antecedent pictures. For the LD-majority condition, 24 sentences appeared with nonlocal antecedent pictures, while the remaining six sentences (including the three mono-clausal sentences) appeared with local antecedent pictures. For the Local-majority condition, 24 sentences (including the three mono-clausal sentences) appeared with local antecedent pictures, while the remaining 6 appeared with non-local antecedent pictures. This resulted in an $80 \% / 20 \%$ distribution in the two unequally distributed conditions.

Picture Verification Test To create the sentences for the picture verification test, one set of six embedded clause sentence frames for each anaphor was reused with three new main clause verbs (splash, paint, measure). To create the trials, half the sentences were paired with their appropriate picture as described above to make match trials, and the other half was paired with the inappropriate picture (e.g. a non-local picture with a local reflexive sentence) to make mismatch trials, making sure that a roughly equal number of each anaphor, embedding and main clause verbs were in both sets of trial types. ${ }^{1}$

Materials for the Japanese experiment For the Japanese version of this experiment, the Japanese sentence materials were constructed in a very similar way as in the English experiment. However, because Japanese verbs are more selective in their argument structure, we were forced to change the events depicted in order to maintain natural sounding sentences, opting for a construction where the anaphor is marked with the dative particle -ni. The events depicted in the Japanese learning materials were: sticking tape to someone (__ni gamutepu-wo haru), wrapping a ribbon around someone (__ni ribbon-wo makitsukeru), and loading a $\log$ onto someone (__ni maruta-wo noseru). In the test they were: spilling water onto someone (__ni mizuwo kakeru), putting paint on someone (_ni enogu-wo nuru), and pinning an award on someone (__ni bajji-wo tsukeru). All stimuli were presented in Japanese using hiragana and kanji, with novel anaphors spelled in katakana.

## Procedure

The procedure was identical for the two language groups, but the Japanese version was carried out in Japanese by trained native Japanese-speaking research assistants.

[^49]Learning Phase This experiment was implemented in PsychoPy (Pierce, 2007). During the Learning Phase the participant was presented with one picture-sentence trial at a time. They were instructed to read the sentence aloud, and take as long as they needed to figure out the meaning of the novel word. They used the space bar to progress to the next trial, working at their own pace.

Participants were instructed to infer the meaning of novel words they would encounter in the experiment. The instructions explicitly stated that the novel words may or may not correspond to existing words in their native language. After the instructions, there were three blocks of learning trials, consisting of 26 trials each, with eight sentences each for the local reflexive and pronoun, and 10 LD reflexive sentences (including 1 mono-clausal local and one LD reflexive sentence).

In order to motivate participants during the learning phase, these learning blocks were interspersed with two quiz blocks containing 12 picture verification test trials, four for each anaphor. At the start of the first quiz block participants were given brief instructions telling them to respond 'match' if the sentence appropriately described the picture based on what they thought the novel words meant. There were two practice trials. No feedback was provided.

The order of these learning blocks, and of the quiz blocks was counter-balanced across participants to control for any list order effects. The trial presentation order was randomized within each block using PsychoPy's trial randomization function. Participants typically completed this phase in 20-30 minutes.

Test Phase Participants were given the picture verification test as described above for the quiz trials. Again, they were told to indicate if the sentence described the picture by pressing either ' $f$ ' for match or ' $j$ ' for mismatch. The trials were presented randomly, with a break halfway through. Participants typically completed the test phase in 20-30 minutes.

## Results

For statistical analyses of the data, the picture verification responses from each language group were entered into a mixed logit model (Jaeger, 2008) with the acceptance response (i.e. the picture and the sentence description match) as the dependent variable. We used learning condition, picture antecedent and anaphor type, and the twoand three-way interactions as predictors, and participant and item as random effects.

English Figure 2 shows the results for the English-speaking participants. On average, participants in all three learning conditions were more likely to accept the local reflexive with a local antecedent ( $\beta=1.568, \mathrm{z}=13.239, \mathrm{p}<.000$ ) and less likely to accept the pronoun with a local antecedent ( $\beta=$ $-1.531, \mathrm{z}=-12.833, \mathrm{p}<.000$ ). This was an expected pattern given that local reflexives and pronouns exist in English, and the input provided consistent information. This


Figure 2: Mean acceptance rates, split by each anaphor given a particular antecedent, across the three learning conditions, from English speaking participants.
reassured us that the task was achievable given consistent input. We turn now to the optionality in the LD reflexive.

With respect to the acquisition of the LD reflexive, we had initially expected participants in the Balanced condition to show the best ability to accept the LD reflexive in both local and non-local antecedent trials. However, this expectation was not confirmed, as the acceptance rate for the LD reflexive was not high in either local or non-local antecedent trials (Figure 2). In order to explore the impact of distributional regularities, we conducted planned pairwise analyses and compared response rates for the LD reflexive across the three learning conditions. Participants in the Local-majority condition and LD-majority condition showed a reliable preference to accept the LD reflexive with the antecedent that was frequently presented in the input (Local-majority condition: $\beta=1.211, \mathrm{z}=6.336, \mathrm{p}<.001$; LD-majority condition: $\beta=-1.024, \mathrm{z}=-5.201, \mathrm{p}<.001$ ). Finally, when response rates for the LD reflexive were compared across the three conditions, participants only showed a strong preference for an antecedent in the Local condition ( $\beta=1.211, \mathrm{z}=6.336, \mathrm{p}<.001$ ) and the LD condition ( $\beta=-1.024, \mathrm{z}=-5.201, \mathrm{p}<.001$ ), not the Balanced condition. Furthermore, in a model considering only acceptance rates of the LD reflexive in the Balanced condition, there was no significant preference for either antecedent ( $\beta=0.614, \mathrm{z}=1.547, \mathrm{p}=.122$ ), suggesting that participants in the Balanced condition truly were responding at chance on a given LD reflexive trial. This pattern of responses suggests that participants' behavioral response patterns reproduced the distributional regularities they observed in their input.

Moreover, participants in the LD-majority condition were less likely to accept the local antecedent with the LD reflexive compared to those in the Local-majority condition ( $\beta=-0.823, \mathrm{z}=-6.119, \mathrm{p}<.000$ ), or those in the Balanced condition ( $\beta=-0.347, z=-2.593, p=.010$ ). Similarly, participants in the Local-majority condition were more likely to accept the LD reflexive with local antecedents compared to those in the Balanced condition $(\beta=0.340, \mathrm{z}=$
2.775, $\mathrm{p}=.006$ ), and the LD-majority condition ( $\beta=0.823$, $\mathrm{z}=6.119, \mathrm{p}<.000$ ). These patterns suggest that participants in the unequally distributed learning conditions also reproduced the rate at which the local or non-local antecedents appeared in their input.

Japanese Figure 3 shows the results for the Japanesespeaking participants. Overall, Japanese speakers' response patterns were very similar to those of English speakers. Participants across the three learning conditions accepted the local reflexive more often with local antecedents $(\beta=$ $1.196, \mathrm{z}=10.524, \mathrm{p}<.001$ ) and rejected the pronoun with local antecedents $(\beta=-1.275, \mathrm{z}=-10.519, \mathrm{p}<.001)$. Japanese speakers were also sensitive to the distributions in their learning condition. When participants in the Balanced condition were compared to those in the Local-majority condition, there was no significant difference in their acceptance rates for the two antecedents with the LD reflexive $(\beta=0.208, z=1.378, p=.168)^{2}$. But like English speakers, participants in the LD condition compared to the Balanced were less likely to accept the LD reflexive with a local antecedent ( $\beta=-0.637, \mathrm{z}=-4.757, \mathrm{p}<.000$ ).

Pairwise analyses of the LD reflexive data at each group level again reveal the similarity between English and Japanese speakers. Participants only showed a strong preference for a local antecedent in the Local-majority condition ( $\beta=0.787, \mathrm{z}=4.065, \mathrm{p}<.001$ ) and non-local antecedent in the LD-majority condition ( $\beta=-1.062, \mathrm{z}=$ $-5.228, \mathrm{p}<.001$ ), but no clear preference was observed in the Balanced condition. Furthermore, in comparison to participants in the Local-majority condition, those in the LD-majority condition were less likely to accept the LD reflexive with local antecedents ( $\beta=-0.679, \mathrm{z}=-5.228, \mathrm{p}<$ .000). Finally, when response rates for the LD reflexive were compared across the three conditions, participants only


Figure 3: Mean acceptance rates, split by each anaphor given a particular antecedent, across the three learning conditions, from Japanese speaking participants.

[^50]showed a strong preference for an antecedent in the Local condition ( $\beta=0.787, z=4.065, p<.001$ ) and the LD condition ( $\beta=-1.062, \mathrm{z}=-5.228, \mathrm{p}<.001$ ), not the Balanced condition. In the model considering only acceptance rates of the LD reflexive in the Balanced condition, there was again no significant preference for either antecedent ( $\beta=-0.493, \mathrm{z}=-1.215, \mathrm{p}=.224$ ).

Overall, participants across both language groups show strikingly similar response patterns. They were only willing to accept each anaphor following the distribution of antecedent co-occurrences provided in their input, even when the optionality presented there resembled their native language (i.e. Japanese speakers). Results also suggest that the chance responding in the Balanced condition across both language groups was not the result of different individuals preferring different antecedents, rather as a group these participants are simply reproducing the statistical regularities in their input.

## Discussion

The present study used a novel anaphor learning paradigm to investigate cross-situational learning of novel anaphors. Overall, we found that participants in our study appear to track the distribution of syntactic structures across situations to constrain their anaphor acquisition. The fact that learners reproduced the distribution of local vs. non-local interpretations for the LD reflexive indicates that the input distribution was guiding the process of anaphor acquisition. This kind of probability-matching behavior, which has been reported in other statistical learning paradigms (e.g. Hudson Kam \& Newport, 2005), further illustrates humans' ability to reproduce and learn regularities in their input from simple frequency tracking (c.f. Estes, 1976). Furthermore, the fact that there was little difference between the two language groups - specifically the fact that Japanese speakers, like English speakers, treated the LD reflexive as a local reflexive or pronoun based on their learning condition rather than following the interpretive possibilities of their native zibun - suggests that biases to copy and reproduce regularities in the input distribution played a more important role than other potential biases (e.g. processing biases or L1 influence).

One of the main research questions in this study concerned how participants handle the optionality in the interpretive possibilities of the LD reflexive, and how manipulating the distribution of those two options would affect acquisition. The results demonstrate that learners struggle to acquire the optionality of the LD reflexive, regardless of their native language. Participants in the unequally distributed learning conditions appeared to treat the antecedent that only appeared $20 \%$ of the time as noise, ignoring it during learning, and accepting roughly that same rate of 'noise' during test (for all three anaphors). In other words, participants in the LD-majority and Local-majority conditions simply treated the LD reflexive as another form of either the pronoun or local reflexive, respectively.

Participants in the Balanced condition appear to be roughly at chance in accepting or rejecting the provided antecedent.

These results and conclusion raise an important question for future research: if distributional regularities are not useful in acquiring anaphors that allow for more than one interpretation, how do speakers of languages like Japanese acquire the interpretative optionality of their LD reflexives?

One possible explanation is that learners must first be confident about one possible antecedent before allowing optionality. For example, learners in the two unequally distributed learning conditions may have settled on one interpretation, without having received enough evidence to allow optionality. Learners in the Balanced condition, however, were not confident about either antecedent. To test this, an on-going follow-up experiment is exploring the effect of presentation order and the frequency of particular antecedents, e.g. presenting a majority of either local or non-local antecedents before introducing the kind of optionality we presented in the Balanced condition.

Alternatively, this difficulty could be explained by a uniqueness principle, i.e., bias against many-to-one meaning-form mappings. Learners must learn that both multiple semantic and syntactic mappings are acceptable for LD reflexives, so future work should investigate this bias at both levels. For example, highlighting the distinctiveness of the LD reflexive's interpretive optionality may increase learners' acceptance of optionality. To this end, sentences with relative clauses may help learners disambiguate between anaphor types: in "The woman standing next to Susan splashed her/herself," zibun is only interchangeable with herself, not her because Susan does not c-command the anaphor. The current data can only address the acquisition of the locality constraint, and not the c-command requirement on the structural relation between antecedent and anaphor, but learning both in tandem may be critical for successful acquisition of an LD reflexive.

Sentences where the reflexive is in the subject position of an embedded clause could also provide evidence that the LD reflexive differs from the local reflexive in its syntactically constrained interpretive possibilities (e.g. "John said that zibun-wa awesome."). While such a sentence is not possible with a local reflexive, it is grammatical for zibun precisely because the LD reflexive can take an antecedent outside the local clause (in contrast to the local reflexive's more restricted locality constraint). Providing these sentences in the learning input may provide further evidence to learners about the distinctiveness of the LD reflexive, and increase their confidence that optionality is integral to the LD reflexive type anaphor itself and not noise in the input.

In short, findings from the present study provide an important step towards understanding the constraints on cross-situational learning of anaphoric expressions. We suggest that successful acquisition of an LD reflexive may require that learners incrementally acquire interpretive possibilities in sequence, or that they are presented with an additional syntactic cue that unambiguously indicates the availability of the non-local antecedent interpretation.

## Acknowledgments

We thank Akari Ikeuchi, Mei Yoshimoto, Moeko Sato, Morine Kondo, and especially Moeko Kita for help with data collection in Japan. The Japanese study was supported by NSF EAPSI fellowship 1614233 awarded to KC.

## References

Chomsky, N. (1981). Lectures on Government and Binding. Dordrecht: Foris Publications.
Dillon, B., Chow, W. Y., Wagers, M., Guo, T., Liu, F., \& Phillips, C. (2014). The structure-sensitivity of memory access: evidence from Mandarin Chinese. Frontiers in Psychology, 5, 1025.
Estes, W. K. (1976). The cognitive side of probability learning. Psychological Review, 83(1), 37.
Fedzechkina, M., Newport, E. L., \& Jaeger, T. F. (2016). Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. Cognitive Science.
Gleitman, L. (1990). The structural sources of verb meanings. Language acquisition, 1(1), 3-55.
Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. Language, 244-285.
Hirakawa, M. (1990). A study of the L2 acquisition of English reflexives. Second Language Research, 6(1), 6085.

Hudson Kam, C. L., \& Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language learning and development, 1(2), 151-195.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434446.

Kobayashi, M., (2002). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. The Modern Language Journal, 86(4), 571-586.
Medina, T. N., Snedeker, J., Trueswell, J. C., \& Gleitman, L. R. (2011). How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences, 108(22), 9014-9019.
Peirce, JW (2007) PsychoPy - Psychophysics software in Python. Journal of Neuroscience Methods, 162(1-2):8-13.
Scott, R. M., \& Fisher, C. (2012). 2.5-year-olds use crosssituational consistency to learn verbs under referential uncertainty. Cognition, 122(2), 163-180.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Yuan, B. (1998). Interpretation of binding and orientation of the Chinese reflexive ziji by English and Japanese speakers. Second Language Research, 14, 324-340.
Yuan, S., \& Fisher, C. (2009). "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. Psychological Science, 20(5), 619-626.

# Expected Utility in Romantic Relationships: Satisfaction as a Cue for Romantic Partnership Dissolution 

Samantha E. Cohen (cohense@indiana.edu)<br>Psychological and Brain Sciences and Cognitive Science Program, 1101 W. $10^{\text {th }}$ Street Bloomington, IN 47405 USA<br>Peter M. Todd (pmtodd@indiana.edu)<br>Psychological and Brain Sciences and Cognitive Science Program, 1101 W. $10^{\text {th }}$ Street Bloomington, IN 47405 USA


#### Abstract

The choice to enter and leave a romantic relationship can be framed as a decision-making problem based on expected utility of the partnership over time, akin to a forager deciding whether to stay in a particular patch based on the amount of resources it provides. We examined the temporal trajectory of three traits that may correspond to resources in romantic relationships-trust, love, and satisfaction-to determine whether they behave like depleting or replenishing patches from a foraging perspective. All three rise over time in intact relationships-suggesting replenishment-but plateau or fall in dissolved relationships-suggesting depletion. Survival analysis demonstrated that higher ratings of all three quality variables decreased the risk of romantic dissolution. The results suggest that these cues are lower in dissolved relationships, indicating individuals could potentially use them as cues for leaving an unsatisfactory relationship patch via aspiration-level cognitive mechanisms.


Keywords: foraging theory; romantic relationships; survival analysis; relationship dissolution; mate choice

All good things must come to an end, including romantic relationships. In some cases, it is not until death that we part, but many relationships end before that point through someone's active choice. Relationship dissolution is a highintensity and long-lasting stressor (Simpson, 1987; Sprecher, 1994), indicating that avoiding it is important to people. Yet most people seem to enter relationships that eventually end and are often caught off-guard when it occurs, indicating that it can be difficult to predict. Individuals cannot usually determine the expected outcome of a relationship prior to entering the relationship itself, meaning they may continuously evaluate their relationship and decide whether one's relationship is likely to (or should) end. Previous work in this choice domain has tackled the brighter side of romance, such as how individuals choose a relationship partner (e.g. Beckage, Todd, Penke, \& Asendorpf, 2009), but less work has examined how individuals decide to move on. What cues do individuals consider when choosing to end a relationship, and how are they incorporated into choices?

How long to stay with something typically depends on what one is getting out of it. Individuals clearly expect to get something out of romantic relationships, given that many expend significant time, money, and energy searching
for and maintaining them. The best-studied aspects of relationship quality (or utility) include relationship satisfaction, intensity of love, and level of trust. Metaanalyses have found that measures of relationship quality more strongly influenced relationship dissolution than either individual traits (e.g. neuroticism) or external factors (e.g. social network overlap; Le et al., 2010). Daters may thus consider the amount of utility, quality, or other valued attribute produced in a relationship when choosing whether to exit it and attempt to find a better relationship. This utilitarian view departs from the traditional cultural emphasis on the holistic nature and complexities of romance. However, for a significant portion of human history, relationships were designed to be mutually beneficial to a couple and their extended families, and relationships with little expected gain were avoided.

It is not novel to suggest individuals attempt to get the best relationship possible (whatever that might be), but few have connected the study of romantic relationships to theoretical models of choice. Problems involving searching for and maximizing resources are well studied within the foraging literature in biology. Optimal foraging theory (OFT; Stephens \& Krebs, 1986) considers how foragers maximize their total gain (e.g. calories) in an environment. Foragers traverse a landscape filled with patches of some resource (e.g. berry bushes), choosing to enter and obtain (e.g. consume) those resources before leaving to find new or better patches. The relationship-foraging model (Cohen \& Todd, 2017) treats the search for successive relationships as a foraging problem. Entering a relationship could be thought of as entering a "patch" that provides some mix of satisfaction, love and other benefits. Relationship foragers search through a social landscape of potential romantic partners, choosing to pass by some, entering into a relationship with others (and reaping the benefits), and possibly eventually departing in search of another. Depending on just what sort of patch a relationship is like, OFT can make predictions about how long people should stay in the relationship patches they find.

One commonly-studied type of patch is characterized by resources that are depleted over time as the forager consumes them. Berry bushes are a common example of this type of patch, with the ripe berries being continually consumed until the forager decides to leave. For these
bushes, the rate of return (e.g. calories consumed per unit time) is expected to increase rapidly early on as a forager enters a patch, with a diminishing rate of return as resources are consumed. The Marginal Value Theorem (Charnov, 1976) in OFT states that individuals should leave such depleting patches when the rate of return within the current patch is less than the rate of return that is expected from the environment at large, given optimal search behavior and expected search costs.

But not all patches only deplete over time-some may deplete and replenish (such as a berry bush with new berries ripening across weeks), and some may produce a roughly constant output for extended periods. Immobile barnacles settled on a tide-pool rock capturing food floating past are partaking of such long-lasting patches. These and other animals are known as sit-and-wait foragers (Beachly, Stephens, \& Toyer, 1995), as once they have found a patch location of this non-depleting type they can stay there exploiting it for a long time. In such situations, "simply staying put can make good economic sense" (p. 265), but leaving is also predicted when patch quality declines or other circumstances change, such that the forager could get a better rate of return by looking (and exploiting) elsewhere.

How do relationships compare to these types of patches? Some aspects seem typically to start high and deplete over time, such as the novelty of a new partner. But others do not appear to be capped-there is probably no preset amount of satisfaction, love, or trust that one can get in a relationship, unlike the fixed number of ripe berries currently in a bush. This suggests that various measures of relationship quality may not only fall but also can replenish over time, and can even stay level or grow indefinitely.

It is important for specifying the relationship-foraging model to determine whether any aspect of relationship quality follows the assumptions of OFT regarding particular types of patchy resource, whether depleting, or replenishing, or constant. A depleting patch should show high (or rapidly growing) initial rate of return, then falling over time. A replenishing patch would have a rate of return that falls and rises, possibly repeatedly. Finally, a constant patch would have a rate of return that rises (possibly very quickly) to a roughly steady state.

In this work, we assess ratings of three types of quality in a romantic relationship: satisfaction, love, and trust. We examine how these factors change over time, to determine which may be useful in predicting relationship dissolution given different patch definitions in the relationship-foraging model. The data come from self-reports from intact and dissolved romantic relationships.

Previous research has demonstrated that aspects of quality, especially satisfaction, tend to increase dramatically at the start of a relationship (Rusbult, 1983) followed by relative stability or decreasing levels over time (Levenson \& Gottman, 1985; Rusbult, 1983; Sprecher, 1999). In one of the rare longitudinal works, satisfaction and commitment to the relationship increased over time (Rusbult, 1983). Sprecher (1999) measured several aspects of quality at
yearly intervals in couples and found the opposite result: Satisfaction was significantly lower for both genders and love was significantly lower for men at the first yearly interval, but otherwise, ratings were remarkably stable. Only the longest-lasting couples (those lasting the full 4 year study period) reported an increase in quality, with a slight increase in commitment and satisfaction for women in the final year. Thus, we expect that longer-lasting relationships will report greater satisfaction, and possibly other quality variables such as trust, although this pattern will not hold for love beyond the earliest time points.

These trajectories diverge depending on the eventual relationship outcome. Among relationships remaining intact (at least, during the duration of a particular study), satisfaction and commitment stabilize or slowly increase over time (Rusbult, 1983; Sprecher, 1999). Individuals exiting a dissolved relationship report decreased satisfaction but only slight decreases in love, suggesting relationships tend to end due to changes in satisfaction rather than love (Sprecher, 1994; Sprecher 1999). These diverging paths lead to large disparities in commitment and satisfaction between couples remaining intact and those ending (Rusbult, 1983; Simpson, 1987). Hence, we expect that reports on satisfaction and trust, but not love, to be higher for intact than dissolved couples.

In the next section we analyze the time-course of selfreported ratings of different types of relationship quality, to assess their similarity to rates of return for depleting, replenishing, and constant patches. We then use survival analysis to test whether these self-reported aspects of quality influence the likelihood of exiting a relationship (i.e., leaving a relationship patch).

## Method

A survey was used to assess relationship length versus quality of intact and dissolved romantic relationships in undergraduates. All procedures were approved by the Indiana University Institutional Review Board.

## Participants

A sample of 700 undergraduates was collected from the Indiana University psychology subject pool. To qualify, participants needed to have at least one romantic experience, including but not limited to casually going on dates, being in a committed relationship, or getting married. Forty-five participants who indicated they had never been in any sort of romantic relationship or did not indicate whether they were in a relationship currently were not included in the analyses. Of the remaining 655 participants, $62.0 \%$ were currently in a romantic relationship. All 655 participants indicated the type of relationships they were giving us ratings for (whether their current relationship or their most recent dissolved relationship); $50.8 \%$ were describing committed relationships (but not engaged or married), 44.4\% were casual relationships and early relationships without a clear classification, $2.4 \%$ were engaged or married, and the remaining $2.3 \%$ fell outside these
categories. Of participants primarily identifying with a single racial group ( $n=616$ ), $72.1 \%$ identified as white, $14.90 \%$ as Asian, $8.1 \%$ as black, and $4.9 \%$ as Hispanic or Latino. Subjects were primarily heterosexual (91.3\%) females (58.6\%) between the ages of 18-21 (90.2\%). Participants were compensated with course credit.

## Measures

Participants completed a survey for approximately 30 minutes online through Qualtrics. Participants currently in a romantic relationship (intact couples) were asked to describe various qualities of their current relationship. Individuals not currently in a relationship were asked to imagine their most recent romantic relationship prior to its dissolution (dissolved couples). Duration of the relationship was measured from its current state (for intact couples) or its final duration (for dissolved couples) in months ( $n=395$ ) and month- and year-based categorical bins ( $n=655$ ) (see Figure 1).

Participants reported three measures of relationship quality on Likert scales:

- Love: "How in love with your partner are you?"
- Trust: "How much do you trust your partner?"
- Satisfaction: "In general, how satisfied are you with your current relationship with your partner?"

Love was measured only for participants indicating they were in love with their partner (meaning a rating is not available for all participants) and ranged from 1 (not very) to 9 (intensely) ( $n=391, M=7.55, S D=1.50$ ). Trust ranged from 1 (not at all) to 9 (completely) ( $n=655$, $M=6.79, S D=2.05$ ). Relationship satisfaction ranged from 1 (extremely dissatisfied) to 7 (extremely satisfied) ( $n=655, M=5.33, S D=1.75$ ). The upper endpoint for the satisfaction scale was different from the other two scales because satisfaction is typically measured on a bipolar 7point scale (e.g. Simpson, 1987), so scores are compared only via correlations.

## Results

To determine whether love, trust, or satisfaction follow appropriate rate of return curves to be considered depleting, replenishing, or constant resources as found in foraging theory, we plotted their mean trajectories over categorical duration groups (see Figures 2, 3, and 4, respectively). Note that each of these trajectories are constructed across subjects, as we only have zero or one data point on each scale from each subject. Love (Pearson's $\mathrm{R}, r=0.20, n=222$, $p=.003$ ) and trust ( $r=0.11, n=395, p=.02$ ) both weakly but significantly positively correlated with relationship duration, but satisfaction did not ( $r=0.03, n=395, p>.05$ ). This suggests that love and trust are generally higher at longer relationship durations, while satisfaction is relatively stable. The variables were highly intercorrelated, with significant, moderate relationships between love and trust ( $r=0.51$,
$n=391, p<.001)$, love and satisfaction ( $r=0.29, n=391$, $p<.001$ ), and trust and satisfaction ( $r=0.48, n=655, p<.001$ ).

While love showed a dramatic increase over the first few months, in line with our prediction, each variable overall tended to continually increase slightly as duration grew, rather than forming a curve of diminishing returns with a plateau (although the curve of trust beyond one year might be described as a plateau).


Figure 1: Distribution of relationship lengths for both dissolved and intact relationships.

Comparing dissolved and intact couples' ratings of each variable, intact couples generally had equal or greater quality across relationship durations with lower variability. Looking at all time points together, there was also greater quality for intact over dissolved relationships as a whole (Independent Samples T-Test, love: $\mathrm{t}(391)=5.11$, trust: $\mathrm{t}(655)=10.02$, satisfaction: $\mathrm{t}(655)=9.58$, all $p<.001$.

The rapid initial rise in love overall is driven largely by ratings of dissolved relationships, which start at much lower levels before plateauing. Intact relationships vary in love by little more than a point across durations. This is somewhat surprising-were the data longitudinal, we could conclude that individuals experiencing strong feelings of love early in a relationship are more likely to stay together. However, for the present work, we can only say that individuals looking back on dissolved, low-duration relationships report them as low in love, while those who are still in them report more intense feelings of love. Trust shows a slight increase overall for both intact and dissolved couples (the few data points for 3-5 year dissolved couples makes that mean value unreliable). Satisfaction is relatively stable both overall and for intact couples, but varies widely, rising and falling, for dissolved couples.

## Survival Analysis

While the quality-over-time plots are indicative of general trends, they do not accurately reflect the outcome of all relationships. Many of the currently intact relationships in our sample will end after the completion of the study. That is, the eventual outcome for these relationships is unknown (or censored), and we should not assume that they will never end.


Figure 2: Mean intensity of love for individuals currently in an intact relationship, reporting a previous dissolved relationship, and overall. Error bars are $\pm 1$ SE.


Figure 3: Mean intensity of trust for individuals currently in an intact relationship, reporting a previous dissolved relationship, and overall. Error bars are $\pm 1 \mathrm{SE}$.


Figure 4: Mean intensity of satisfaction for individuals currently in an intact relationship, reporting a previous dissolved relationship, and overall. Error bars are $\pm 1$ SE.

To more fully assess how each quality variable interacts with relationship duration while accounting for these possible eventual dissolutions, a survival analysis of relationship duration was run using the lifelines package in Python 3.5 (Davidson-Pilon, 2016). This analysis measures the connection of specified variables with the likelihood of relationship dissolution over time by estimating the number of relationships currently intact that will eventually dissolve at any particular duration. If a variable has no impact on relationship dissolution (and hence inversely duration), then it is not likely to be a resource that matters in terms of relationship foraging. However, if a resource increases or decreases the risk of dissolution (hence, a shorter or longer duration, respectively), it is something to be actively avoided or sought out, respectively, from the perspective of relationship foraging.

A Cox Proportional Hazard model of relationship duration (in months, available for $n=395$ ) was fit individually to satisfaction, trust, and love to determine how they affect the risk of dissolution. Dissolved relationships were coded as observed events, and non-numeric durations (e.g. "more than 24 months") were excluded.

Table 1 shows the coefficients for each factor, which indicate the change each factor causes to the baseline hazard rate of dissolution based on length of the relationship alone. A positive coefficient indicates a heightened risk of dissolution relative to the baseline; a negative coefficient indicates a reduced risk of dissolution.

All three variables were significantly predictive of reduced risk of dissolution ( $p<.001$ ), such that individuals reporting greater love, trust, or satisfaction would have longer relationships on average (with love having the

Table 1: Results of the Cox Proportional Hazards model, where the $\log$ of the coefficient indicates the proportional dissolution risk compared to the baseline hazard rate.

| Variable | Events/Observations | Cox Coefficient $(\beta)$ | $\mathrm{e}^{\beta}$ | $p$ | $95 \%$ CI |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Love | $73 / 221$ | -0.432 | 0.649 | $<.001$ | -0.649 to -0.216 |
| Trust | $184 / 395$ | -0.391 | 0.677 | $<.001$ | -0.512 to -0.270 |
| Satisfaction | $184 / 395$ | -0.285 | 0.752 | $<.001$ | -0.409 to -0.161 |

greatest impact and satisfaction the least). This is in line with our predictions and corroborates the patterns in Figures 2-4 while also considering possible eventual dissolutions beyond the durations in our data.

## Discussion

Examining changes in relationship quality over time revealed that longer relationships had reliably elevated levels of love and trust (but not satisfaction), contrary to our predictions. Our findings of satisfaction stability across relationship lengths mirror those of Sprecher (1999), while our continual rise in trust and love is in line with Rusbult (1983)'s pattern of satisfaction increase.

As we expected, intact relationships showed significantly higher and more consistent rates of return in trust and satisfaction than dissolved relationships, in line with findings by Rusbult (1983). However, contrary to expectations, love was also higher in intact couples than dissolved couples. Sprecher (1994) previously found that love, but not satisfaction, generally remains high for dissolved couples, suggesting dissolution may be based on satisfaction alone. Results from our survival analysis showing that greater love was associated with lower likelihood of dissolution also countered Sprecher's finding. All three variables were thus higher in intact than dissolved relationships, and were predictive of dissolution. This suggests that these dyadic relationship factors may be used as criteria for choosing when to exit a relationship, or may co-vary with other factors predictive of dissolution.

The goal of this work was to examine whether particular aspects of relationship quality follow the time courses expected for depleting, or replenishing, or constant resource patches, to enable characterization of a relationship-foraging model (Cohen \& Todd, 2017). The three qualities of love, trust, and satisfaction show rising levels over time on average for all relationships (intact and dissolved), suggesting that they are replenishing resources. It is important from a foraging perspective though to consider dissolved relationships separately, where a decision to leave the relationship-patch was made by at least one party; there love seems to plateau, suggesting a constant resource, while trust and satisfaction may actually decrease somewhat at the longest duration, suggesting depletion. Here more data is needed to better specify the form of these functions and what kind of patches they may conform to.

How might individuals use different aspects of quality to choose when to leave a relationship patch? If relationships
are like depleting patches, then OFT predicts that people will usually decide to leave and break up the relationship when its quality declines below the level they could expect to get elsewhere, without necessarily having any other relationship to move to. If relationships are like nondepleting sit-and-wait patches, then people would not be expected to leave while quality remains constant (or grows), but if they encounter another potential relationship of possibly higher value this could prompt a shift to that new patch, depending on switching costs. These strategies could make use of key cognitive comparison abilities.

First, while individuals would benefit from being able to accurately track their current relationship quality over time, previous work suggests that they may not be good at it (Sprecher, 1999). Regardless of the actual change in relationship quality, humans have a tendency to say that things have changed for the better. Individuals reporting annual current ratings of love and satisfaction for four years show very little significant differences between temporal samples, but report increases when asked how their feelings have changed over time (Sprecher, 1999). This suggests that individuals are not especially accurate at recalling past rates of return from a relationship, at least explicitly. However, given the relatively stability of quality after the first few months, individuals could conceivably hold a single average of quality and accurately characterize the overall quality of the relationship. Searchers are generally sensitive though to the rates of return from other possible relationships (interdependence theory; Thibaut \& Kelley, 1959). That is, one's satisfaction with a relationship's quality can depend on comparison to the quality one could expect from currently available alternative mates (Rusbult, 1983; Simpson, 1987; Thibaut \& Kelley, 1959). Thus we might expect relationship-leaving strategies to depend on comparison to other current possibilities or to averages of past experienced quality levels.

Second, by comparing the current relationship quality to a single previous average level or to the expected level of another currently possible relationship, individuals can use a simple aspiration value heuristic to assess the value of staying in their current relationship. Previous work in romantic partner choice shows individuals use aspirationlevel heuristics when choosing to enter a new relationship based on the attractiveness of potential mates at speeddating events (e.g. Beckage et al., 2009), so conceivably, this mechanism may apply to choosing to exit a relationship as well. The decision threshold could be based on the gap in
quality between successful and unsuccessful relationships over time. In our data, the size of this gap generally increased over time, which would predict continual dissolutions over time. Individuals may also consider the length of time spent below an aspiration value, given that a large gap between initial and long-term quality (for love, at least) is expected in all relationships in the early stages.

Relationships may be a different type of patch from just depleting or non-depleting: one where the quality level (return rate) increases the more the forager puts effort into it (up to a point). Moreover, the forager does not usually know how high the return rate will grow until they get there. (nor know how low the return may fall if it ever does start to decline). How should an individual decide whether to stay in such a situation? This may depend on how rapidly the rate of return is increasing or decreasing at particular stages of the relationship as a consequence of particular amounts of effort. New models will be needed to explore this, going beyond some of the strong simplifying assumptions made in existing models of non-depleting patch foraging. For instance, rather than having patches switch from being good to being bad instantaneously and searchers switch from not knowing anything about the return rate to knowing it precisely (Beachly et al., 1995), more realistic models for relationship foraging should include gradual change in rates of return and gradual learning of those rates.

A limitation of this work is that each relationship quality variable was measured at only one point, so intra-individual change in quality cannot be examined. Instead, we calculated the expected curves of returns over time using population averages. These patterns of change in quality do though mirror those from related longitudinal work. While not ideal, this replication provides some reassurance that using single observations per individual gives a relatively realistic representation, at least of the overall shapes of these curves, though future iterations of this model should use longitudinal data. Relatedly, there could have been systematic differences in ratings between those participants who were currently in relationships and so were asked to give current ratings, and participants who were not currently in relationships and so were asked to recall ratings from their earlier, now-dissolved relationship. For participants asked to recall their ratings of quality prior to their previously relationship's breakup, it is inevitable that this type of recall could skew ratings of quality, likely negatively. In addition, ratings could differ depending on who initiated dissolution. If the person did the breaking up, then they may rate their relationship quality as decreasing before the breakup; but if the person was on the receiving end, they may not rate quality as declining so much. Getting data from both parties in multiple relationships would help address this issue. We only asked for ratings of love from those participants who indicated they were in love, which may have skewed results positively (although they still showed a significant difference between intact and dissolved relationships). Finally, this work used a college sample, and the results may not generalize to the overall
population, especially given that most college students are not yet old enough to have had many long-term relationships.

Using a new framework, relationship-foraging modelling, and a technique that is relatively uncommon in cognitive science research, survival analysis, we found that some aspects of relationship quality may be considered depleting or replenishing resources from a foraging perspective, depending on how the relationship proceeds. Further research will explore whether thinking about relationships in terms of resource-filled patches that people can exploit and deplete or work to replenish themselves over time can help us understand how and when couples decide to stay or leave to forage for greener pastures.

## References

Beachly, W. M., Stephens, D.W., \& Toyer, K. B. (1995). On the economics of sit-and-wait foraging: Site selection and assessment. Behavioral ecology, 6(3), 258-268.
Beckage, N., Todd, P. M., Penke, L., \& Asendorpf, J. (2009). Testing sequential patterns in human mate choice using speed dating. In Proceedings of the 2009 Cognitive Science Conference, pp. 2365-2370.
Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. Theoretical population biology, 9(2), 129-136.
Cohen, S.E. \& Todd, P.M. (2017). Relationship foraging: Does time spent single predict relationship length? Manuscript submitted for publication.
Davidson-Pilon, C. (2016). Lifelines. Github repository: https://github.com/CamDavidsonPilon/lifelines
Le, B., Dove, N. L., Agnew, C. R., Korn, M. S., \& Mutso, A. A. (2010). Predicting nonmarital romantic relationship dissolution: A meta-analytic synthesis. Personal Relationships, 17(3), 377-390.
Levenson, R. W., \& Gottman, J. M. (1985). Physiological and affective predictors of change in relationship satisfaction. Journal of Personality and Social Psychology, 49(1), 85-94.
Rusbult, C. E. (1983). A longitudinal test of the investment model: The development (and deterioration) of satisfaction and commitment in heterosexual involvements. Journal of Personality and Social Psychology, 45(1), 101-117.
Simpson, J. A. (1987). The dissolution of romantic relationships: Factors involved in relationship stability and emotional distress. Journal of Personality and Social Psychology, 53(4), 683-692.
Sprecher, S. (1994). Two sides to the breakup of dating relationships. Personal Relationships, 1(3), 199-222.
Sprecher, S. (1999). "I love you more today than yesterday": Romantic partners' perceptions of changes in love and related affect over time. Journal of Personality and Social Psychology, 76(1), 46-53.
Stephens, D. W., \& Krebs, J. R. (1986). Foraging theory. Princeton, NJ: Princeton University Press.
Thibaut, N. \& Kelley, H. (1959). The social psychology of groups. New York: Wiley.

# Analogical gestures foster understanding of causal systems 

Kensy Cooperrider (kensy@uchicago.edu)<br>Department of Psychology, 5848 S. University Avenue<br>Chicago, IL 60637 USA

Dedre Gentner (gentner@northwestern.edu)<br>Department of Psychology, 2029 Sheridan Road Evanston, IL 60208 USA

Susan Goldin-Meadow (sgm@uchicago.edu)<br>Department of Psychology, 5848 S. University Avenue<br>Chicago, IL 60637 USA


#### Abstract

Sensitivity to the causal structure underlying phenomena is critical to expert understanding. Fostering such understanding in learners is therefore a key goal in education. We hypothesized that observing analogical gestures-which represent relational information in visuospatial formatwould lead learners to notice and reason about underlying causal patterns, such as positive and negative feedback. Participants watched brief video lectures about the human body and the plant kingdom, which were delivered along with gestures representing either: 1) visuospatial details (iconic gesture condition); or 2) relational structure (analogical gesture condition). In a subsequent classification task, relative to participants who saw iconic gestures, participants who saw analogical gestures were more likely to sort the phenomena described in the videos-as well as novel phenomena-by their causal structure (e.g., positive feedback). The results suggest that analogical gestures can be harnessed to foster causal understanding.


Keywords: analogy; relational reasoning; gesture; learning; complex systems

## Introduction

A deep understanding of any complex phenomenon-from the ebb and flow of the tides, to the rise and fall of blood pressure-requires an understanding of the causal structure that gives rise to it (Lagnado, Waldmann, Hagmayer, \& Sloman, 2007; Mackie, 1980; Sloman, 2005). Yet this is no trivial task. The causal relations that govern phenomena throughout the physical and social world are often embedded in a wealth of concrete, causally irrelevant particulars (e.g., Rottman, Gentner, \& Goldwater, 2012). Thus a key question for cognitive scientists and educators alike is: How do people come to understand causal structure, and how can we foster this understanding?

We focus here on an important arena of causal understanding: namely, causal systems-abstract patterns of causation, such as positive and negative feedback, that occur in a wide range of phenomena (Fernbach \& Sloman, 2009; Rottman, Gentner, \& Goldwater, 2012; see also Day, Motz, \& Goldstone, 2015). Feedback systems can be found in the human body, in household appliances, in economic markets, and in plant physiology, to name just a few domains. For
example, in a simple two-factor positive feedback system, an increase in one causal factor causes an increase in a second factor, which in turn causes an increase in the first factor; in a negative feedback system, an increase in one causal factor causes an increase in a second factor, which then causes a decrease in the first factor. The challenge such systems present for a learner are considerable: To notice, for instance, the abstract causal likeness between human perspiration and a flush toilet-i.e., that both involve negative feedback-one has to look past a host of visuospatial and sensory differences between the two. One way that expertise in noticing such causal patterns comes about is through repeated opportunities to compare examples across domains, as happens over an extended science education (Rottman et al., 2012). Indeed, such a process can be induced in the laboratory by having learners analogically compare disparate examples of causal systems (Goldwater \& Gentner, 2015), much as comparison can promote relational learning more generally (Christie \& Gentner, 2010; Doumas \& Hummel, 2013; Gentner, Loewenstein, Thomspson, \& Forbus, 2009; Gick \& Holyoak, 1983; Jung \& Hummel, 2011; Kotovsky \& Gentner, 1996; Kurtz, Boukrina, \& Gentner, 2013).

Here we explore a less-studied route through which it may be possible to foster causal understanding: analogical gestures. In previous work, we found that, people produce gestures in abundance when explaining feedback systems (Cooperrider, Gentner, \& Goldin-Meadow, 2016). These gestures were analogical in that they used space, not to represent concrete spatial details, but to represent relational structure: they used locations to distinguish causal factors, motion to show increases and decreases to those factors and causal relationships between them, and complex movements to summarize the overall relational structure of the systems. Strikingly, these gestures occurred in abundance even though participants were explaining systems that were not inherently spatial and were, by design, devoid of the kinds of concrete details that usually prompt gestures. These laboratory findings provide insights into how people express spatial analogies in gesture, and also raise an important further question: Might using such analogical gestures during instruction foster understanding of causal patterns?


Figure 1: A selection of gestures from one of the video explanations, describing the phenomenon of anxiety attacks. In the version seen by participants in the iconic condition, all the gestures represented concrete aspects of the phenomenon (panels $\mathrm{a}, \mathrm{b}$, and d); in the version seen in the analogical condition, the gestures represented the two causal factors (e) and the pattern of increase and decrease to those factors ( f and h ). For all phenomena, the contrasting versions of the explanations used the same number of gestures, included beat gestures in the same places ( c and g ), and used an identical audio track.

Gesture is ubiquitous in everyday conversation as well as in the classroom (Goldin-Meadow, 2003). Despite being more implicit than speech (McNeill, 1992), gesture is an important medium for communicating ideas (Hostetter, 2011), including abstract ideas about relational structure (Jamalian \& Tversky, 2012). Moreover, gesture has been found to boost learning in a range of content domains (Ping \& Goldin-Meadow, 2008; Singer \& Goldin-Meadow, 2005; Valenzeno, Alibali, \& Klatzky, 2003). While prior studies have focused on gesture's consequences for young learners, gestures conveying abstract concepts have also been attested in lectures to older students, in disciplines ranging from literary studies to mathematics (Corts, 2006; Mittelberg, 2008; Núñez, 2008).

Importantly, the gestures used in everyday communication and the classroom come in different varieties, and may not all be equally effective in conveying ideas about causal structure. A first type is iconic gestures. These are produced in the course of explaining concrete, visuospatially rich content, and are used to represent size and shape, location, motion, and spatial relationships (Alibali, 2005; Alibali \& Hostetter, 2008; McNeill, 1992). A second type of gesture is more abstract, using location, motion, and spatial relationships to represent ideas and relationships that are not inherently spatial. Such gestures include those described in our prior work on explanations of feedback systems (Cooperrider, et al., 2016), and also a range of other content domains (Cienki \& Muller, 2008; Cooperrider \& Goldin-Meadow, 2017; Goldin-Meadow, 2003). It is this latter type-analogical gestures-that we
predict would lead observers to notice and reason about causal structure. Iconic gestures, by contrast, may have no effect on causal understanding, or may even hinder it by highlighting concrete particulars.
In the present study, we test the idea that analogical gestures can be used to foster understanding of causal systems - patterns which are often buried beneath concrete particulars. To this end, we created two sets of short video lectures: one in which an actor accompanies his explanations of phenomena in the human body and plant kingdom with iconic gestures that depict concrete visuospatial details (iconic gesture condition); another in which the actor accompanies his explanations with analogical gestures depicting relational structure (analogical gesture condition). We hypothesized that participants in the analogical gesture condition would be more likely to notice the underlying causal structure of the phenomena described in the lectures and, moreover, that these participants would be more likely to discern causal structure when encountering novel phenomena.

## Methods

## Participants

60 undergraduate students from Northwestern University participated in exchange for course credit. 15 participants were eliminated for failing a video comprehension check (described below), and two were eliminated for admitting during debriefing that they listened to the audio but did not watch the screen. In all, 43 participants ( 21 iconic condition,

22 analogical condition; 24 men; $M$ age $=18.7$ years) were included in the analyses.

## Materials and procedure

Video Stimuli Drawing on materials from prior studies (Rottman et al., 2012; Smith \& Gentner, 2014), we developed short descriptions of four phenomena: anxiety attacks (a positive feedback system within the domain of the human body); blood pressure regulation (negative feedback, human body); bracken fern growth (positive feedback, plant kingdom); and prayer plant cycles (negative feedback, plant kingdom). The descriptions balanced concrete details (e.g., for the anxiety attacks description: the "heart feeling like it is pounding") with clues to causal structure ("this will lead to even more intense symptoms"). We filmed an actor delivering each of these ( $\sim 45 \mathrm{sec}$ ) explanations in two versions: one with iconic gestures depicting concrete aspects of the phenomenon described, and one with analogical gestures representing the causal factors involved in the phenomenon and the behavior of those factors (Fig. 1). The iconic gestures were based on the actor's intuitions about what would be most natural; the analogical gestures were inspired by the gestures produced spontaneously by participants when describing highly abstract versions of feedback systems (Cooperrider et al., 2016). For each phenomenon, the two versions had the same number of target-i.e., iconic or analogical-gestures (5-7 per explanation), as well as the same number of beat gestures (2-3), which were included to make the explanation more naturalistic. Finally, to control for differences in prosody, the audio track of the actor speaking was identical in the two versions. This was achieved by: 1) recording a primary audio track, 2) filming the actor talk and gesture in sync with the primary track while it played, 3 ) aligning the video with the primary track and removing the secondary audio.

Classification Task We developed a phenomenon classification task inspired by the Ambiguous Sorting Task (AST) used in prior studies (Rottman et al., 2012; Goldwater \& Gentner, 2015). Participants were first presented with written descriptions of three new phenomena: blood clots (positive feedback, human body); spotted knapweed growth (negative feedback, plant kingdom); and internet routers (common cause, technology). These three phenomena served as the "seed" categories into which further descriptions would have to be classified. The central feature of the task, as with the AST variants used previously, is that nature of the categories is up to the participant to decide: the seed phenomena represent three different domains as well as three different causal structures, thus affording both kinds of classification.

After reviewing the seed phenomena, the participant classified eight further written descriptions, one at a time. Four of these were descriptions of novel phenomena; the other four were written versions of the videos watched earlier. Note that only six of the phenomena were "critical" in that classifying by domain and by causal structure were
mutually exclusive: all four novel phenomena and two of the familiar ones (blood pressure, bracken fern). The other two familiar phenomena (anxiety, prayer plants) shared the same combination of domain and causal structure as the seed phenomena (e.g., anxiety attacks and blood clotting are both in the domain of the human body and both positive feedback systems).

Other Assessments Participants completed two further assessments of causal system understanding: a battery of inference questions and a diagram task. Both concerned only the four phenomena familiar from the videos. For the inference battery, participants answered eight questions (two per phenomenon) querying general behaviors of-and predictions about - the systems described. For the diagram task, participants were shown contrasting diagrams of positive and negative feedback, with a detailed explanation of the symbols used (e.g., plus and minus signs). Participants were then asked to match each of the videos seen previously to the correct diagram.

Procedure All video stimuli and assessments were implemented in Qualtrics and displayed on a desktop computer. The experiment started with a brief video introduction by the lecturer, which was the same for both conditions. Participants then watched the four videos in a fixed order, in the versions corresponding to their condition assignment, i.e., for a participant in the iconic condition, the iconic gesture versions of all four explanations. After watching each video, participants answered two multiplechoice questions about their basic content (e.g., which symptoms of anxiety were mentioned). These questions were intended as an attention check, and participants who got one or more wrong were excluded.

After the final video, participants proceeded to the classification task. The seed phenomena were presented as fixed blocks of text on the screen; the to-be-classified phenomena were presented, one at a time, on moveable digital "cards." Participants were instructed to: "Decide which of the descriptions [i.e., the seed phenomena] the card [i.e., the to-be-classified phenomenon] is most similar to, and drag it to that pile." The eight phenomena were presented in a fixed order, with the four novel ones first, followed by the four familiar ones. We had participants sort the novel phenomena first to encourage them to think deeply about the task, rather than sort by first impulse. After the classification task, participants completed the inference question battery and the diagram task, both in a fixed order, and were debriefed. In all, the task took around 20 minutes.

## Results

## Classification Task

Our primary measure was the mean proportion of phenomena that participants classified by domain (human body, plant kingdom, or technology), by causal structure (positive feedback, negative feedback, or common cause),


Figure 2: The mean proportion of critical phenomena classified by domain, by causal structure, or by other criteria in the two gesture conditions. Error bars represent standard error of the mean.
or by other criteria. We first considered only the six critical items-that is, those for which classification by domain or by causal structure were mutually exclusive. Participants in the iconic gesture condition classified a higher proportion of the critical phenomena by domain than did participants in the analogical gesture condition (iconic: $M=.37, S D=.20$; analogical: $M=.19, S D=.16 ; t=-3.2, d f=41, p=.003$, Cohen's $d=-0.98$ ). Conversely, participants in the analogical gesture condition classified a higher proportion of the critical phenomena by causal structure than did participants in the iconic gesture condition (analogical: $M=$ $.64, S D=.18$; iconic: $M=.47, S D=.20 ; t=3.02, d f=41, p$ $=.004$, Cohen's $d=0.92$ ) (Fig. 2). Participants in the two conditions sorted by some other criterion to the same extent (iconic: $M=.17, S D=.15$; analogical: $M=.17, S D=.15 ; t$ $<.001, d f=41, p=1$ ). Importantly, the same pattern of significance holds when looking only at the four novel phenomena (by domain: $p=.01$; by causal structure: $p=$ .01 ; by other: $p=.88$ ). Indeed, for all eight phenomena, a higher proportion of participants in the analogical gesture condition sorted by causal structure than did participants in the iconic gesture condition.

To get a better sense of individual participants' classification behavior, we also zoomed in on the two critical phenomena that were featured in the videos: blood pressure regulation and bracken fern growth. Note, again, that to classify these by causal structure, participants had to resist the temptation to group the blood pressure description with the human body seed (blood clots) and the bracken fern description with the plant kingdom seed (spotted knapweed). Yet not a single participant in the analogical gesture condition classified both these phenomena by domain, compared to six participants in the iconic gesture condition who did (two-tailed Fisher's exact, $p=.009$ )

## Other assessments

Inference Question Battery Participants in both conditions answered the majority of inference questions correctly and at close to ceiling, with no difference between the conditions (iconic: $M=.90, S D=.12$; analogical: $M=.86$, $S D=.13, t=-1.09, d f=41, p=.28)$.

Diagram Task Participants in both conditions answered the majority of diagram questions correctly and at close to ceiling, but with those in the analogical gesture condition performing marginally better (iconic: $M=.80, S D=.19$; analogical: $M=.89, S D=.13, t=1.82, d f=41, p=.08$ ).

## Discussion

The present study investigated the hypothesis that seeing certain types of gestures would lead observers to notice and understand causal structure. Specifically, we expected that analogical gestures, which represent relational structure spatially, would foster understanding of causal structure better than would iconic gestures, which represent concrete visuospatial details. This hypothesis was borne out. Participants in the analogical gesture condition noticed the causal structure of the phenomena described in videos, and were also more likely to notice causal structure in entirely new phenomena. Given the design of our classification task, this was no easy feat. To classify by causal structure, participants had to look past compelling differences of content to find deeper similarities, or look past compelling content similarities to discern deeper differences. Our prior work showed that people spontaneously produce analogical gestures when explaining causal systems; the current study builds on these findings to show that such analogical gestures have important consequences for learning.

Our leading interpretation of the present findings, again, is that observing analogical gestures led participants to notice and reason about causal structure. A second-and not mutually exclusive-possibility is that the concrete gestures in the iconic condition hindered participants from the discerning the underlying relational structure by lavishing them with vivid details. Follow-up studies with a "no gesture" control condition would clarify whether our two gesture conditions are indeed pulling observers in opposite directions or, if not, which gesture type is driving the observed pattern of results. This question is of clear theoretical interest, but we also note that, in teaching contexts, gesture is ubiquitous, perhaps even inevitable. Thus, from a practical perspective, the important question is not whether teachers should gesture about the phenomena they are explaining, but how. Our results suggest that iconic gestures, as natural as they are, may not always be the best choice. It may be that the best instructors already intuit this, using gestures that highlight relational structure when possible.

By what specific mechanism(s) did analogical gestures have their beneficial effects? We hypothesize that these gestures helped convey the causal content of each phenomenon by capturing it in schematic spatial form.

Indeed, space is a familiar and intuitive format in which to represent and reason about relational structure (e.g., Gattis, 2004; Tversky, 2011). But we think there may have also been another important reason for the efficacy of analogical gestures: namely, that they invited comparison and uniform representation across the scenarios. Participants in the analogical gesture condition viewed four videos, all featuring qualitatively similar gestures. For example, all four parsed the phenomena into causal factors by establishing locations in space, and all showed the increases and decreases to those factors as vertical movements. The gestures in the iconic gesture condition also had some commonalities across videos (e.g., the gestures in both human body videos indexed body parts), but they were hardly as schematic and alignable. Prior work has shown that using the same words in superficially different contexts prompts observers to compare those contexts (Clement, Mawby, \& Giles, 1994; Gentner, 2003), and using similar gestures across different examples may have similar effects. The idea that such a mechanism drives the current results is consistent with earlier findings that prompting people to compare examples of feedback systems fosters causal understanding (Goldwater \& Gentner, 2015). Thus, while we refer to the abstract gestures in the present study as "analogical" because they rely on a structured mapping between spatial structure and relational structure (Cooperrider et al., 2016), they are also "analogical" in another sense: they invite observers to form analogies across the different contexts in which they are used. Future studies might assess these mechanisms by comparing a condition in which the analogical gestures are qualitatively similar and thus alignable across lessons-as in the present study-with a condition in which the analogical gestures are more heterogeneous and thus less alignable.

What makes the present findings perhaps surprising is that co-speech gestures are largely implicit (Goldin-Meadow, 2003; McNeill, 1992). People seem to produce gestures spontaneously and unreflectively, and do not always notice the ones that others produce. Indeed, when we queried participants at the end of the present experiment about what they thought of the lecturer's gestures, several participants demurred, saying that they "didn't notice them." And yet, despite this "under the radar" quality, these gestures have clear consequences for learning (for a recent review, see Novack \& Goldin-Meadow, 2015). Another question for further research is whether the implicit nature of gesture is key to its benefits. Would the techniques that speakers use to make gesture more salient-e.g., looking at their own gestures (Cooperrider, 2017)—make gesture even more powerful in instruction? And would more explicit forms of visuospatial communication, such as diagrams (Novick, 2003; Tversky, 2011) or sketches (Forbus, Usher, Lovett, \& Wetzel, 2011), also be effective in fostering causal understanding?

## Conclusion

Sensitivity to causal structure is a hallmark of expert understanding. Discovering how to foster such sensitivity is an important goal for cognitive scientists and educators across the natural and social sciences. Our results suggest that a ubiquitous dimension of communication-gesturemight be harnessed to this end. Analogical gestures like those in the present study have been elicited in the lab (Cooperrider et al., 2016), but their importance in instruction has not been investigated. The present findings offer first steps toward figuring out whether those gestures have consequences for learners and, if so, why.

## Acknowledgments

Funding for this study was provided by the NSF Spatial Intelligence and Learning Center (SBE 0541957, Gentner and Goldin-Meadow are co-PIs), the Office of Naval Research (ONR (N00014-16-1- 2613 to Gentner), and NICHD (R01-HD47450 to Goldin-Meadow).

## References

Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. Spatial Cognition \& Computation, 5(4), 307-331.
Christie, S., \& Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. Journal of Cognition and Development, 11(3), 356-373.
Cienki, A., \& Müller, C. (Eds.) (2008). Metaphor and gesture. Philadelphia: John Benjamins.
Clement, C. A., Mawby, R., \& Giles, D. E. (1994). The Effects of Manifest Relational Similarity on Analog Retrieval. Journal of Memory and Language, 33, 396420.

Cooperrider, K. (2017, in press). Foreground gesture, background gesture. Gesture.
Cooperrider, K. \& Goldin-Meadow, S. (2017, in press). When gesture becomes analogy. Topics in Cognitive Science.
Corts, D. P. (2006). Factors characterizing bursts of figurative language and gesture in college lectures. Discourse Studies, 8(2), 211-233.
Day, S. B., Motz, B. A., \& Goldstone, R. L. (2015). The cognitive costs of context: the effects of concreteness and immersiveness in instructional examples. Frontiers in Psychology, 6(Dec), 1-13. http://doi.org/10.3389/ fpsyg.2015.01876.
Doumas, L. A. A. \& Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. PLOS One, 8 (6).
Fernbach, P. M., \& Sloman, S. A. (2009). Causal learning with local computations. Journal of experimental psychology: Learning, memory, and cognition, 35(3), 678.

Forbus, K., Usher, J., Lovett, A., \& Wetzel, J. (2011). CogSketch: Sketch understanding for Cognitive

Science Research and for Education. Topics in Cognitive Science, 3(4), 648-666.
Gattis, M. (2004). Mapping relational structure in spatial reasoning. Cognitive Science, 28, 589-610.
Gentner, D. (2003). Why we're so smart. In D. Gentner \& S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought. Cambridge, MA: MIT Press.
Gentner, D., Loewenstein, J., Thompson, L., \& Forbus, K. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. Cognitive Science, 3, 1343-1382.
Gick, M. L., \& Holyoak, K. J. (1983). Schema Induction and Analogical Transfer. Cognitive Psychology, 15, 1-38.
Goldin-Meadow, S. (2003). Hearing gesture: How our hands help us think. Cambridge, MA: Harvard U. Press.
Goldwater, M. B., \& Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. Cognition, 137, 137-153.
Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. Psychological Bulletin, 137(2), 297-315.
Hostetter, A. B., \& Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. Psychonomic Bulletin \& Review, 15(3), 495-514.
Jamalian, A., \& Tversky, B. (2012). Gestures alter thinking about time. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Cognitive Science Society. Austin, TX: Cognitive Science Society.
Jung, W., \& Hummel, J. E. (2011). Progressive alignment facilitates learning of deterministic but not probabilistic relational categories. In L. Carlson, C. Hölscher, \& T. F. Shipley, Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Kotovsky, L., \& Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.
Kurtz, K. J., Boukrina, O., \& Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 39(4), 1303-1310.
Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., \& Sloman, S. A. (2007). Beyond covariation: cues to causal structure. In A. Gopnik \& L. Schulz (Eds.), Causal learning: Psychology, philosophy, and computation. New York: Oxford University Press.
Mackie, J. L. (1980). Cement of the universe: A study of causation. Oxford, England: Clarendon Press.
Mittelberg, I. (2008). Peircean semiotics meets conceptual metaphor: Iconic modes in gestural representations of grammar. In A. Cienki \& C. Müller (Eds.), Metaphor and gesture. Philadelphia: John Benjamins.
Novack, M. \& Goldin-Meadow, S. (2015). Learning from gesture: How our hands change our minds. Educational Psychology Review, 27(3), 405-412.

Novick, L. R. (2001). Spatial diagrams: key instruments in the toolbox for thought. In D. L. Medin (Ed.), The psychology of learning and motivation (Vol. 40). San Diego: Academic Press.
Núñez, R. (2008). A fresh look at the foundations of mathematics: gesture and the psychological reality of conceptual metaphor. In A. Cienki \& C. Müller (Eds.), Metaphor and gesture. Amsterdam: John Benjamins.
Ping, R., \& Goldin-Meadow, S. (2008). Hands in the air: Using ungrounded iconic gestures to teach children conservation of quantity. Developmental Psychology, 44(5), 1277-1287.
Rottman, B. M., Gentner, D., \& Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. Cognitive Science, 36(5), 919-32.
Singer, M. A., \& Goldin-Meadow, S. (2005). Children learn when their teachers' gestures and speech differ. Psychological Science, 16, 85-89.
Smith, L. A., \& Gentner, D. (2014). The role of differencedetection in learning contrastive categories. In P. Bello, M. Guarini, M. McShane, \& B. Scassellati (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society. Austin: Cognitive Science Society.
Sloman, S. A. (2005). Causal models: How people think about the world and its alternatives. Oxford: Oxford University Press.
Tversky, B. (2011). Visualizing thought. Topics in Cognitive Science, 3(3), 499-535.
Valenzeno, Laura, Martha W. Alibali, and Roberta Klatzky. 2003. Teachers' gestures facilitate students' learning: A lesson in symmetry.Contemporary Educational Psychology, 28, 187-204.

# Characterizing spatial construction processes: Toward computational tools to understand cognition 

Cathryn S. Cortesa (ccortes4@jhu.edu)<br>Department of Cognitive Science, 3400 N. Charles St. Baltimore, MD 21218<br>Jonathan D. Jones (jdjones@jhu.edu)<br>Department of Electrical and Computer Engineering, 3400 N. Charles St. Baltimore, MD 21218<br>Gregory D. Hager (hager@cs.jhu.edu)<br>Department of Computer Science, 3400 N. Charles St. Baltimore, MD 21218<br>Sanjeev Khudanpur (khudanpur@jhu.edu)<br>Department of Electrical and Computer Engineering, 3400 N. Charles St. Baltimore, MD 21218

Amy L. Shelton (ashelton@jhu.edu)<br>School of Education, 2800 N. Charles St. Baltimore, MD 21218<br>Barbara Landau (landau@jhu.edu)<br>Department of Cognitive Science, 3400 N. Charles St. Baltimore, MD 21218


#### Abstract

Spatial construction-creating or copying spatial arrangements-is a hallmark of human spatial cognition. Spatial construction appears early in development, predicts later spatial and mathematical skills, and is used throughout life. Despite its importance, we know little about the cognitive processes underlying skilled construction. Construction tasks are highly complex but analyses have tended to focus on broad-stroke measures of end-goal accuracy. In this paper we introduce a novel behavioral coding formalism to characterize an individual's entire construction process, examine many individuals' processes in aggregate, and summarize patterns that emerge. The results show high consistency at certain points occurring throughout the construction, but also indicate flexibility in the interim paths that lead to and diverge from these points. Our approach offers a new method that can more precisely describe the behavioral patterns observed during construction in order to reveal the underlying cognitive processes engaged, and capture individual differences in building expertise.


Keywords: spatial skills; spatial cognition; block copying; computational model

## Introduction

Spatial construction-the activity of creating novel spatial arrangements or copying existing ones-is a hallmark of human spatial cognition. These activities naturally occur during childhood and adolescence and are related to later achievements in science, technology, engineering, and mathematics (STEM) fields (Hsi, Linn, \& Bell, 1997; Kell, Lubinski, Benbow, \& Steiger, 2013; Verdine, Golinkoff, Hirsh-Pasek, Newcombe, et al., 2014). Moreover, spatial play during early schooling-including spatial building tasks-contributes to school readiness (Verdine, Golinkoff, Hirsh-Pasek, \& Newcombe, 2014; Wai, Lubinski, \&

Benbow, 2009), developmental of logico-mathematical abilities (Casey et al., 2008; Cheng \& Mix, 2012; Nath \& Szücs, 2014), and math performance in middle and high school (Stannard, Wolfgang, Jones, \& Phelps, 2001; Wolfgang, Stannard, \& Jones, 2003).

Despite the importance of spatial construction skills, little is known about the cognitive processes underlying their origins and development. Part of the reason for this is that spatial construction skills are highly complex, yet the cognitive characterization of these skills and their measurement has been quite limited. For example, although evaluation of block construction tasks has long been recognized as an important assessment of spatial skills (Bailey, 1933), most methods of assessment only evaluate the end product (accuracy), and fail to measure the construction process. Studies have generally reported broad stroke outcome measures such as time to complete a structure (Akshoomoff \& Stiles, 1996; Frick, Hansen, \& Newcombe, 2013), binary measures of block placement as correct or incorrect (Brosnan, 1998; Hoffman, Landau, \& Pagani, 2003; Stiles \& Stern, 2001), or summary ratings for the complexity, planning, or organization of free-play block designs (Caldera et al., 1999; Casey \& Bobb, 2003; StilesDavis, 1988; Stiles \& Stern, 2001). Even studies that aim to characterize development of construction processes or strategies have used analytic categories that are limited in their generality for understanding construction. For example, some have suggested that children start with simple iterative methods (i.e. stacking blocks on top of one another), then move to sequential combinations of methods (i.e. first creating a line of blocks next to one another, then creating a stack), and finally come to flexibly shift between multiple methods (Stiles-Davis, 1988; Stiles \& Stern, 2001; Stiles, Stern, Trauner, \& Nass, 1996). These
characterizations tell us little about the step-by-step processes that the user takes when carrying out a complex construction, nor how the ever-expanding set of outcomes grows over time.

More recent studies have attempted to provide a more precise characterization of the process occurring during construction. Verdine and colleagues characterized children's placement errors, including whether a block was placed in the correct layer, in correct orientation relative to other blocks, and with the correct attachment studs connected (Verdine, Golinkoff, Hirsh-Pasek, Newcombe, et al., 2014; Verdine, Golinkoff, Hirsh-Pasek, \& Newcombe, 2016). Researchers in computer science have generated step-by-step instructions for assembling block models based on physical constraints such as avoiding 'floating blocks' not supported from below (Zhang, Igarashi, Kanamori, \& Mitani, 2016). Both studies begin to characterize the temporal and incremental nature of block construction.

Each of the approaches discussed above provides a description of the accuracy of a block construction at points intermediate to building or at the end; but none provides a characterization of an individual's complete construction process. Yet, variability and/or consistency across individuals' construction processes may reveal much about the underlying cognitive and perceptual abilities and biases that influence the builder's construction choices.

The incremental process of adding blocks to a structure can unfold in many ways, with different strategies leading to the same successful solution. Some of this variation may be unimportant-merely small tweaks in the options one can use to complete a construction. Other aspects of variation are likely to reflect important cognitive processes. For example, limitations of attention and memory make it likely that certain strategies or processes will be preferred as they may reflect more efficient use of available cognitive resources (Ballard, Hayhoe, Pook, \& Rao, 1997). Certain strategies may also reflect the builder's understanding of the physical principles engaged during building. For example, the effect of gravity could bias the builder to construct from the bottom layer upwards (Zhang et al., 2016). Finally, construction strategies may be related to perceptual or semantic groupings of the blocks within the structure. The sub-parts which the builder chooses to construct, and the order in which they are created may be driven by the builder's perceptual parsing of the model being copied. More generally, there may be systematic commonalities in the construction paths that builders use, and these may vary depending on the builder's level of skill.

Understanding the principles underlying construction requires methods that can characterize the builder's full construction process. Ideally, the best analysis would completely describe the entire construction process, capturing any imaginable construction outcome as well as each step of building along the way. This kind of characterization would be as relevant for a simple stack of blocks as it would be for an elaborate castle or an abstract collection of connected pieces.

To our knowledge, such methods have never been reported. Therefore, in this paper, we report a new method for characterizing the precise nature of processes involved when people carry out a relatively simple construction task: using blocks to copy a target model. To do this, we ask adults to carry out a simple building task, using a set of Duplos ${ }^{\mathrm{TM}}$. We describe our new method for coding block construction behavior that uses a novel computer interface. Our method characterizes each partial assembly created during the building process as a step taken along a path from the start to the end of construction. We then evaluate common states and predominant path types traversed by adults as they move through the construction process. Finally, we make inferences about the underlying cognitive mechanisms engaged during block construction.

## Method

## Participants

Twenty-seven healthy adults 18-53 years old ( $M=21 ; 4, S D$ $=6 ; 6$ ) participated in the study. A university ethical review board approved the study's procedures, and all participants provided informed consent.

## Materials

Participants were asked to copy six different block models of varying size, each consisting of 4,6 , or 8 blocks. Each participant copied each of the six models in randomized order, but always began with the two smallest models (models 1 and 2). Figure 1 shows each of the six models.

We used Duplo ${ }^{\text {TM }}$ blocks for the construction copy task. These blocks were chosen for several reasons. First, the attachment mechanism allowed the blocks to be connected to each other in fixed ways. The attachment studs permitted precise specification of the relationships between blocks above and beside one another. In addition, the limited set of colors (red, yellow, green, blue) of each shape ( $2 \times 2$ square, $4 \times 2$ rectangle) were ideal for the precise measurement system we developed.


Figure 1: The block models used in this study. Models 1 and 2 contain four blocks, models 3 and 4 contain six blocks, and models 5 and 6 contain 8 blocks.

We mounted a PrimeSense Carmine RGBD camera in an overhead configuration to record participants' behaviors as
they carried out the construction task, at a rate of 30 frames per second. All videos were coded using our annotation interface. The coder viewed the video recording frame-byframe on a desk-top computer.

## Procedures

Participants were seated at a table marked with an outline of a rectangular area ( $14.75 \times 24.00 \mathrm{in}$.) in which they completed their block construction copy. During data collection, the experimenter observed participants in real time on a video display monitor. A vertical black barrier was placed on the table behind the construction area to obscure the video display monitors from the participant and to avoid distraction. Figure 2 shows the testing equipment setup used for the study.

In the procedure, the experimenter first placed the model at a $45^{\circ}$ angle in the rear left corner of the marked construction area on the table. Each model was presented in a standardized orientation so that the greatest number of model surfaces were visible to the participant. Then, the experimenter placed the corresponding loose blocks on the table in the center of the construction area by emptying them from a small bag. This ensured random starting positions for each of the blocks used to construct the copy. Participants were instructed to take their time and to copy the model, building as efficiently and accurately as possible.


Figure 2: Overhead camera and blocks set up for the block copying task. The model was placed at the rear left of the table, and blocks for the copy construction were placed at the center.

## Analytic Rationale

To account for the broad range of construction behaviors and resulting complex patterns in the copy, we developed a new behavioral coding system, executed in a custom designed computer interface. The video frames for each trial from each participant were coded as a series of actions, each of which culminated in a state. Each action captured the start and end time of a change made to the copy as it was being constructed. Actions could be comprised of a single relationship, such as placing two blocks adjacent on the
table. Other actions included a complex set of simultaneous relationships such as adding a single block in a location that was both above and beside other blocks. Actions could be constructive, such as adding a block or connecting multiple parts, or deconstructive, such as removing a block or separating a structure into two parts.

Each relationship was defined specifically by the set of attachment studs involved. For example, if two rectangle blocks were placed horizontally adjacent to each other along the principal (long) axis, then four studs on each block would be involved in the adjacency. Alternatively, if they were attached along the secondary (short) axis, only two studs on each block would be implicated. Block studs were identified according to the column and row on each block, so the coded data specified the exact relationship between sets of two blocks.

Each action modified the environment to result in a new block state, defined as a set of block attachments present in the copy. Since the construction process occurred over time, each action included a time stamp that allowed block states to be ordered. Here, we refer to the ordered sequence of states over time as a construction path, where actions represent transitions connecting one state to the next. To illustrate, Figure 3a shows six observed states (illustrated as images of block configurations) and 11 observed actions (directed arrows). Any set of arrows that lead from the first null state to the final correct copy state could comprise a construction path.

## Data analysis

One researcher coded all videos. The initial state of the model was always a null state in which no blocks were connected. Each other state along the path to the final copy was attained via a constructive or deconstructive action taken at the preceding state. Since each participant could take any number of actions, construction path length was not balanced across individuals. We also used the coded data to count the number of unique state transitions for all participants in aggregate. Results of the analysis are described below. Principles of the results are true across all six models, but we illustrate using two models as examples.

## Results

Overwhelmingly, the most common actions were correct single-block placements over time. Participants tended to take efficient paths that traversed an average of just over $\mathrm{n}-1$ states for a model that contains n blocks. This held true for all six of the models, including the four-block models 1 and $2(M=3.3, S D=0.9$ and $M=3.1, S D=0.6$, respectively), the six-block-models 3 and 4 ( $M=5.0, S D=0.7$, and $M=$ $5.5, S D=1.8$, respectively), and the eight-block-models 5 and $6(M=7.4, S D=0.9$, and $M=7.9, S D=1.7$, respectively). Strikingly, the observed correct states represented only a small proportion of total possible correct states. For example, for the six-block model 3 (shown in

Figure 3b), 79 possible correct states exist ${ }^{1}$. In aggregate, our sample executed a total of 136 actions, but only created 16 different correct states ( $27 \%$ of all possible correct states). An additional three erroneous states were observed in model 3; these will be discussed later.


Figure 3: Observed paths for models 2 (a) and 3 (b). Paths begin at the top, where the null state represents no blocks connected. Images represent block states; arrows connecting images represent actions. Numbers adjacent to arrows represent the number of times that action was executed.

Of the 16 observed states for model 3 (Fig 3b), some were created by a majority of participants, while others were rare. We found the same pattern in the observed paths, that is, the actions moving from state to state. Though all observed paths led to a correct copy construction, some were highly likely, while others were highly unlikely. For example, the first image in the fourth row in Figure 3b was created by a great majority of participants ( $96.3 \%$ ). The most common

[^51]path to achieve this state led from the fifth image in the third row of Figure 3b, such that 17 of the 26 ( $65.4 \%$ ) people who created the same penultimate state achieved it by placing the green square on the second layer.

The most commonly traversed states created points of convergence. Convergence points represented a single subassembly that results from several different preceding actions. We observe that convergence points tend to be highly likely states, which most or all participants created along the way to a complete construction. As shown in Figure 3b, most participants (66\%) passed through the state in which three blocks are joined with horizontal adjacency to create the base of the copy (second image, second row). The observed frequencies of convergence points are remarkable when one considers that only a fraction of all possible efficient paths go through these states.

We also observed points of divergence. Divergence points represented cases where participants, when presented with identical partial assembly states, chose to proceed with several different actions. For example, about $70 \%$ of those who created the base of model 3 b proceeded to place the green then the blue square in the second layer. The other $30 \%$ instead placed the same two blocks in the opposite order, first placing the blue and then the green square in the second layer. This is illustrated in Figure 3b, in the third and fifth images of the third row.

Our results demonstrated some commonalities across the six different models. The most frequently constructed partial assembly states for all six of the models represented a complete layer. Across all six models, $83.6 \%$ of participants began their copy construction by creating the base layer. For the models with six or eight blocks, $77.4 \%$ and $86.5 \%$ of participants created the complete second layer as a partial assembly, respectively. Across all six models, each complete layer state is visited more frequently than would be expected by chance, even with the most conservative comparison against only other observed states with the same number of blocks (all $p$ 's $<.001$ ).

Many participants' construction paths (75.5\%) traversed all complete layer states, although this is by no means necessary in order to achieve a correct copy. Specifically for the most complex model, model 6 , those participants who traversed each complete layer partial assembly state in their individual construction path also tended to have the shorter path lengths $(t(24)=-2.57, p=.017)$. In other words, when faced with a complex block copying task, building layer by layer appears to be both highly likely and highly efficient. These observations provide insight into the importance of layers, which may be driven by the builder's understanding of physical properties such as gravity and/or perceptual biases that suggest a natural parse in terms of layers.

Although most block placements were correct (i.e. replicated part of the model in the copy), there were some errors-that is, states that did not represent a correct part of the model. These errors contributed to deviations from the main construction paths. If erroneous states are included in our calculation, the number of possible block states in given
a starting set of 4 , 6 , or 8 blocks is vast, but finite. For example, a mathematician recently estimated that there are nearly a billion possible ways to connect six uniform rectangular Lego ${ }^{\text {TM }}$ blocks contiguously (Abrahamsen \& Eilers, 2011). The model and our instructions to participants constrained their behavior such that even though errors occurred, only an extremely small proportion of all possible states were observed.

The errors observed in this sample also provided insight into the cognitive limitations of our adult participants. Two categories of errors were observed. First, spatial errors occurred when a participant utilized the correct block, but placed it in incorrect orientation relative to the rest of the copy. For example, in model 2, one participant placed the yellow square block with incorrect relationship relative to the red and green rectangles underneath, shown in the first image of the first row in Figure 3a. Second, block identity errors occurred when a participant created the correct form in their copy, but used the wrong color block relative to the model. For example, for model 3, one participant used the yellow rectangle to create the base of their copy instead of the green rectangle (first image, first row of Figure 3b).

Overall, our results provided rich detail about the step-bystep process undertaken by our adult participants in the block copying task. We observed only a small portion of all possible correct states, and a yet smaller portion of all possible states including erroneous ones. The distribution of the sample across different construction paths was not uniform, but rather demonstrated commonalities across the six models. Specifically, convergence points were observed corresponding to completed copy layers, and divergence was observed in the order of block placement within a single layer. Most common construction paths involved the sequential construction of horizontal layers, beginning with the base and building upward.

## Discussion

Our study presented a precise, quantitative method for understanding how people carry out a simple block construction task. Using a novel behavioral coding method together with computational modelling, we precisely described the block construction process as a temporal sequence of states. This approach shed light on the cognitive processes that support spatial construction tasks.

A description of state transitions illustrated commonalities among the construction paths that participants used for each of the six models. Convergence points tended to correspond to the completion of a horizontal layer in the model, while divergence points tended to correspond to various orders of placing blocks within a layer. We hypothesize that convergence points can be interpreted as boundaries between perceptual or semantic chunks-that is, they represent sub-goals that builders had in mind as they approached and carried out the task. Although we did not provide any pre-determined conceptual units or clear perceptually-based chunks (such as sub-parts built from same-colored blocks), participants nonetheless created these
chunks in systematic ways. The location of convergence points, for example, at the completion of a horizontal layer, may indicate that participants grouped or chunked the models principally into horizontal layers or "floors".

It is likely that the underlying structure of sub-goals will vary substantially, depending on a variety of factors. For example, a model that is organized to highlight salient perceptual units, such as multiple vertically adjacent blocks of the same color, could induce a construction path that would take most builders through a convergence point organized as a vertical chunk, and not the horizontal layers observed in the present study. In this case, we anticipate that adults would attend to the imposed perceptual units and change their construction strategy to build using sub-goals defined by these color-units. Similarly, incorporating conceptual structure into the models could radically alter people's construction paths-heads and eyes on structures that look like animals, or wheels on structures that look like vehicles could serve as the chunks or sub-goals to be built. The role of conceptual knowledge in the reproduction of complex figures has a long history in the domain of chess, where experts are known to reproduce board configurations using sub-structures that reflect high-level concepts such as attack and defend (Chase \& Simon, 1973).

In our simple construction task, errors were relatively rare. Errors were characterized as either spatial relationship or block identity errors. We hypothesize that spatial errors indicate problems with spatial working memory, in translating information observed in the model into the working copy. Block identity errors, on the other hand, may involve object working memory, or a prioritization of spatial configuration over color information. These error types are likely linked to the relative simplicity of the models we used; analysis of error patterns for more complex models may well reveal more variation.

We see our method as a powerful way to examine the nature of sub-goals and errors, applicable to a variety of visual-spatial construction tasks involving conceptual or perceptual chunks. The extent to which observed construction paths and construction errors change over variations in the target model would provide insight into how building principles change across target types. In addition, our method permits evaluation of variation in construction paths across different participant populations including construction experts compared to novices, and developmental populations of children at different ages.

We believe that our analytic method has great potential for revealing the fine-grained nature of many tasks that require step-by-step actions, which in turn require rich cognitive capabilities, including representation of the goal as well as strategies for moving from a start state to an end state. Such general task requirements are ubiquitous throughout life-from the toddler who learns to operate an iPad to the adult who learns to cook a gourmet meal. Our insight is that understanding complex skills requires a finegrained and precise approach, exemplified by the method we have introduced. Block construction serves as a first
demonstration of the utility of our approach, but it is by no means the end.

## Acknowledgments

This work was supported by NSF grant \#1561278.

## References

Abrahamsen, M., \& Eilers, S. (2011). On the asymptotic enumeration of Lego structures. Experimental Mathematics, 20, 145-152.
https://doi.org/10.1080/10586458.2011.564539
Akshoomoff, N. A., \& Stiles, J. (1996). The influence of pattern type on children's block design performance. Journal of the International Neuropsychological Society, 2, 392-402.
Bailey, M. W. (1933). A scale of block constructions for young children. Child Development, 4(2), 121-139. https://doi.org/10.2307/1125591
Ballard, D. H., Hayhoe, M. M., Pook, P. K., \& Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences, 20, 723-767.
Brosnan, M. J. (1998). Spatial ability in children's play with Lego blocks. Perceptual and Motor Skills, 87, 19-28.
Caldera, M., Culp, A., O Brien, M., Tuglio, R., Alvarez, M., \& Huston, A. (1999). Children's play preferences, construction play with blocks, and visual-spatial skills: Are they related? International Journal of Behavioral Development, 23(4), 855-872. https://doi.org/10.1080/016502599383577
Casey, B. M., Andrews, N., Schindler, H., Kersh, J. E., Samper, A., \& Copley, J. (2008). The development of spatial skills through interventions involving block building activities. Cognition and Instruction, 26(3), 269309. https://doi.org/10.1080/07370000802177177

Casey, B. M., \& Bobb, B. (2003). The power of block building. Teaching Children Mathematics, 10(2), 98-102.
Chase, W. G., \& Simon, H. A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.
Cheng, Y. L., \& Mix, K. S. (2012). Spatial training improves children's mathematics ability. Journal of Cognition and Development, 15(1), 2-11. https://doi.org/10.1080/15248372.2012.725186
Frick, A., Hansen, M. A., \& Newcombe, N. S. (2013). Development of mental rotation in 3- to 5 -year-old children. Cognitive Development, 28(4), 386-399. https://doi.org/10.1016/j.cogdev.2013.06.002
Hoffman, J. E., Landau, B., \& Pagani, B. (2003). Spatial breakdown in spatial construction: Evidence from eye fixations in children with Williams syndrome. Cognitive Psychology, 46, 260-301. https://doi.org/10.1016/S0010-0285(02)00518-2
Hsi, S., Linn, M. C., \& Bell, J. E. (1997). The role of spatial reasoning in engineering and the design of spatial instruction. Journal of Engineering Education, 86(2), 151-158. https://doi.org/10.1002/j.21689830.1997.tb00278.x

Kell, H. J., Lubinski, D., Benbow, C. P., \& Steiger, J. H.
(2013). Creativity and technical innovation: Spatial ability's unique role. Psychological Science, 24(9), 18311836. https://doi.org/10.1177/0956797613478615

Nath, S., \& Szücs, D. (2014). Construction play and cognitive skills associated with the development of mathematical abilities in 7-year-old children. Learning and Instruction, 32, 73-80. https://doi.org/10.1016/j.learninstruc.2014.01.006
Stannard, L., Wolfgang, C. H., Jones, I., \& Phelps, P. (2001). A longitudinal study of the predictive relations among construction play and mathematical achievement. Early Child Development and Care, 167(1), 115-125. https://doi.org/10.1080/0300443011670110
Stiles-Davis, J. (1988). Developmental change in young children's spatial grouping activity. Developmental Psychology, 24(4), 522-531. https://doi.org/10.1037//0012-1649.24.4.522
Stiles, J., \& Stern, C. (2001). Developmental change in spatial cognitive processing: Complexity effects and block construction performance in preschool children. Journal of Cognition and Development, 2(2), 157-187.
Stiles, J., Stern, C., Trauner, D., \& Nass, R. (1996). Developmental change in spatial grouping activity among children with early focal brain injury: Evidence from a modeling task. Brain and Cognition, 31(1), 46-62. https://doi.org/10.1006/brcg.1996.0024
Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., \& Newcombe, N. S. (2014). Finding the missing piece: Blocks, puzzles, and shapes fuel school readiness. Trends in Neuroscience and Education, 3(1), 7-13. https://doi.org/10.1016/j.tine.2014.02.005
Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., \& Newcombe, N. S. (2016). Links between spatial and mathematical skills across the preschool years. Monographs of the Society for Research in Child Development.
Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., Newcombe, N. S., Filipowicz, A. T., \& Chang, A. (2014). Deconstructing building blocks: Preschoolers' spatial assembly performance relates to early mathematics skills. Child Development, 85(3), 1062-1076. https://doi.org/10.1111/cdev. 12165.
Wai, J., Lubinski, D., \& Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. Journal of Educational Psychology, 101(4), 817-835. https://doi.org/10.1037/a0016127
Wolfgang, C. H., Stannard, L., \& Jones, I. (2003). Advanced constructional play with LEGOs among preschoolers as a predictor of later school achievement in mathematics. Early Child Development and Care, 173(3), 467-475. https://doi.org/10.1080/02568540109594958
Zhang, M., Igarashi, Y., Kanamori, Y., \& Mitani, J. (2016). Component-based building instructions for block assembly. In Computer-Aided Design and Applications (pp. 55-59). Vancouver, Canada: Taylor \& Francis. https://doi.org/10.1080/16864360.2016.1240450

# Mathematical invariants in people's probabilistic reasoning 

Fintan J. Costello (fintan.costello@ucd.ie)<br>School of Computer Science, University College Dublin, Dublin 4, Ireland<br>Paul Watts (watts@thphys.nuim.ie)<br>Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland


#### Abstract

Recent research has identified three invariants or identities that appear to hold in people's probabilistic decision making: the addition law identity, the Bayes rule identity, and the QQ identity (Costello and Watts, 2014, Fisher and Wolfe, 2014, Costello and Watts, 2016b, Wang and Busemeyer, 2013, Wang et al., 2014). Each of these identities represent specific agreement with the requirements of normative probability theory; strikingly, these identities seem to hold in people's probability judgments despite the presence of strong and systematic biases against the requirements of normative probability theory in those very same judgments. We assess the degree to which two formal models of probabilistic reasoning (the 'probability theory plus noise' model and the 'quantum probability' model) can explain these identities and biases in probabilistic reasoning.


## Introduction

A fundamental goal of science is to find invariants: constant relationships that hold between different variables. While such invariants occur frequently in the 'hard' sciences, they are rare in behavioural science. Recent work, however, has identified three invariants that appear to hold in people's probabilistic reasoning: the addition law, Bayes rule and ' QQ ' ('Quantum Question') identities (Costello and Watts, 2014, Fisher and Wolfe, 2014, Costello and Watts, 2016b, Wang and Busemeyer, 2013, Wang et al., 2014). Each identity describes a constant relationship that holds between different probabilistic judgments, and each represents specific agreement with the requirements of probability theory. Strikingly, these identities hold in people's probability judgments despite the presence of strong biases, or systematic deviations from probability theory, in those very same judgments. We assess two formal models of probabilistic reasoning (the probability theory plus noise model, Costello and Watts, 2016, and the quantum probability model, Wang et al. 2014) in terms of their ability to explain these invariant identities and biases.

## Identities in probabilistic reasoning

We use the following notation. We take $P(A)$ to represent the normatively correct probability of event $A$. We take $P_{*}(A)$ to represent a subjective estimate of $P(A)$. The QQ identity involves the relationship between probability estimates when questions are presented in specific orders. We take $P_{B A}(A)$ to represent the subjective estimated probability of $A$ when questions are presented in the order $B A$ (when people are asked to estimate $P(A)$ immediately after being asked to estimate $P(B)$ ) and take $P_{A B}(A)$ to represent the estimate when
questions are in the reverse order. Since a subsequent estimate for $P(B)$ cannot affect the results obtained from a prior estimate for $P(A), P_{*}(A)=P_{A B}(A)$ and $P_{*}(B)=P_{B A}(B)$.

## The QQ identity

Consider a situation where people are asked questions in two alternative orders $A B$ or $B A$. This situation is commonly seen in polls; for example, in a Gallup poll conducted in September 1997, half of participants were asked the question "Do you think Al Gore is honest and trustworthy?" followed immediately by the question "Do you think Bill Clinton is honest and trustworthy?", while the other half of participants were asked the same questions in the reverse order (Moore, 2002). People's answers in such situations are often strongly influenced by the order of question presentation $\left(P_{A B}(A) \neq P_{B A}(A)\right)$. In the Clinton-Gore questions, for example, $76 \%$ of participants answered 'yes' to the Gore question when it was asked first $\left(P_{A B}(A)=0.76\right)$, while $66 \%$ answered yes when it was asked second $\left(P_{B A}(A)=0.66\right)$ : the prior presentation of the Clinton question produced a bias, reducing the likelihood of a 'yes' answer to the Gore question. Simultaneously, however, results (both from experimental studies and from polls) show that the following identity tends to hold reliably in sequential question answering:
$P_{A B}(A \wedge B)+P_{A B}(\neg A \wedge \neg B)-P_{B A}(A \wedge B)-P_{B A}(\neg A \wedge \neg B)=0$
(where $A \wedge B$ represents a 'yes' answer to both $A$ and $B$ and $\neg A \wedge \neg B$ represents a 'no' answer to both questions). This expression has a value of -0.003 in answers to the ClintonGore questions, for example. This identity holds for questions across a wide range of different topics in 72 different national representative surveys in the US, and in laboratory studies of the effects of order in question answering, even though these surveys show significant bias due to question order (Wang et al., 2014). This is just as predicted by the quantum probability model, and is seen as providing 'the strongest form of support' for that model (Wang et al., 2014).

## The Addition Law and Bayes Rule identities

A number of identities must hold in standard probability theory. One such identity is the addition law, which requires that

$$
\begin{equation*}
P(A)+P(B)-P(A \wedge B)-P(A \vee B)=0 \tag{1}
\end{equation*}
$$

must hold for all events $A$ and $B$. Two other 'expansion' identities require that

$$
\begin{align*}
& P(A \wedge B)+P(A \wedge \neg B)-P(A)=0  \tag{2}\\
& P(A \wedge B)+P(\neg A \wedge B)-P(B)=0 \tag{3}
\end{align*}
$$

must hold for all events $A$ and $B$. Consider experiments where we ask people to estimate various probabilities $P(A), P(B)$, $P(A \wedge B), P(A \vee B), P(A \wedge \neg B), P(B \wedge \neg A)$ (not in any fixed ordering), and combine those estimates as in the various identities. Results show that, when we combine people's probability estimates for a given pair of events $A, B$ as in the addition law identity, the average value obtained is equal to probability theory's required value of 0 . When we combine the same estimates for the same events $A, B$ as in the two expansion identities, the average value is not equal to 0 ; instead, the average value is positive (typically around 0.25 ) and is similar for both of these expansion identities. In other words, people's probability estimates reliably agree with probability theory for the addition law identity, but deviate from probability theory for the two expansion identities.

The addition law identity applies to direct or marginal probabilities. Similar results hold for identities that involve conditional probabilities. One such identity is the additive form of Bayes rule, which requires that

$$
\begin{equation*}
P(B \mid A) P(A)-P(A \mid B) P(B)=0 \tag{4}
\end{equation*}
$$

must hold for all events $A$ and $B$. Two parallel 'Bayes expansion' identities require that

$$
\begin{align*}
& P(A \wedge B)-P(A \mid B) P(B)=0  \tag{5}\\
& P(A \wedge B)-P(B \mid A) P(A)=0 \tag{6}
\end{align*}
$$

must hold for all events $A$ and $B$. Experimental results for these identities follow those seen above: for the Bayes Identity the average value in people's estimates is equal to 0 , while for the two Bayes expansion identities, the average value is positive (typically around 0.12 , half the value seen for expansion identities in Equations 2 and 3) and is similar for both of these expansion identities (see Table 1). Results for these identities don't just hold when averaging across events: they also hold separately for each individual pair of events $A$ and $B$, and they hold for estimates about familiar everyday events, medical diagnoses, future political or economic outcomes, or personality-description scenarios (Costello and Watts, 2014, Fisher and Wolfe, 2014, Costello and Watts, 2016b).

These patterns of agreement with the addition law and Bayes rule identity and simultaneous violation of the expansion identities (with approximately the same positive value for Equations 2 and 3 and approximately half that value for Equations 5 and 6), are predicted by the probability theory plus noise model. Confirmation of these predictions has been taken as evidence that the probability theory plus noise
model 'may provide a fully general account of the mechanisms by which people estimate probabilities' (Costello and Watts, 2016b).

The quantum probability model, then, accounts for the QQ identity and for biases due to order effects, while the noise model accounts for the addition law and Bayes rule identities and for biases in the expansion identities. Can either model explain all three sets of results? In the next section we show that the quantum model is in principle unable to explain the addition law and Bayes rule results. We then show that the noise model gives a natural account for all these results.

## The quantum probability model

The quantum probability model (Wang and Busemeyer, 2013, Wang et al., 2014) assumes that people's probabilistic reasoning follows the mathematical rules used to calculate event probability in quantum theory. A fundamental aspect of quantum theory is that the probability of two quantum events can depend on the order in which those events are measured. This order dependence allows the quantum probability model to address various order effects seen in people's sequential inference and judgment.

Probability has a geometric interpretation in quantum theory, based on the projection of vectors. We avoid this geometric interpretation here and instead focus on explaining how quantum probability agrees with, and deviates from, standard probability theory. In quantum probability, an observable defines the set of all possible distinct outcomes for a given measurement: the set of possible answers to the question represented by that measurement. The primary theoretical distinction between quantum and standard probability lies in the idea of 'compatible' or 'incompatible' observables. Two observables are compatible if both observables can be measured simultaneously. If two observables are compatible, quantum probability theory reduces exactly to standard probability theory in all cases. This means that if two observables are compatible then all the probability theory identities described above have a value of 0 , and there are no order effects in judgment.

Incompatible observables, by contrast, cannot be measured simultaneously, and measurement outcomes depend on the order of measurement. If all probabilities are measured with the same ordering then again quantum probability theory reduces exactly to standard probability theory (if all probabilities are of the form $P_{A B}()$, for example, then all relationships between those probabilities match the requirements of standard probability theory and all probability theory identities hold). If probabilities are measured with different orderings, however, then quantum probability deviates from standard probability, producing biases in judgment and order effects in sequential question answering such as $P_{A B}(A) \neq P_{B A}(A)$ and $P_{A B}(A \wedge B) \neq P_{B A}(A \wedge B)$.

## Addition law and Bayes rule identities

The addition law and Bayes rule identities apply in cases where questions are not presented in some specific order $A B$

Table 1: Predicted values of the noise model and the quantum model for a series of probability theory identities. Standard probability theory requires these identities to have a value of 0 . Observed average values for these identities are from Costello and Watts (2016b), Experiment 1. Similar average values hold for each individual pair $A, B$ in that experiment and in a range of other experiments.

| identity |  | noise model | quantum model |  |  | observed |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | compatible | incompatible: order $A B$ | incompatible: order $B A$ |  |
| (1) | $P(A)+P(B)-P(A \wedge B)-P(A \vee B)$ |  | 0 | 0 | $\delta_{A}$ | $\delta_{B}$ | 0.01 |
| (2) | $P(A \wedge B)+P(A \wedge \neg B)-P(A)$ | $d$ | 0 | $-\delta_{A}$ | 0 | 0.26 |
| (3) | $P(A \wedge B)+P(\neg A \wedge B)-P(B)$ | $d$ | 0 | 0 | $-\delta_{B}$ | 0.23 |
| (4) | $P(B \mid A) P(A)-P(A \mid B) P(B)$ | 0 | 0 | $\Delta_{A B}$ | $\Delta_{A B}$ | 0.006 |
| (5) | $P(A \wedge B)-P(A \mid B) P(B)$ | $d / 2$ | 0 | $\Delta_{A B}$ | 0 | 0.12 |
| (6) | $P(A \wedge B)-P(B \mid A) P(A)$ | $d / 2$ | 0 | 0 | $-\Delta_{A B}$ | 0.12 |

or $B A$, but instead are order independent. In this situation there are no order effects for simple probabilities (the probability of $A$ is $P_{*}(A)=P_{A B}(A)$ and that of $B$ is $P_{*}(B)=P_{B A}(B)$ ). Order effects for incompatible observables still apply when people are asked to estimate conjunctive or disjunctive probabilities such as $P(A \wedge B), P(A \wedge \neg B)$ or $P(A \vee B)$. For such conjunctions or disjunctions the quantum probability model assumes a particular characteristic ordering for observables that depends on the causal link between those observables. Complex probabilities such as $P(A \wedge B)$ are estimated using this characteristic ordering. This means that the relationship between a simple probability $P_{*}(A)$ and the conjunctive probabilities $P(A \wedge B)$ and $P(A \wedge \neg B)$ will depend on this characteristic ordering. In particular, when the characteristic ordering of observables for conjunctions is $A B$ we have

$$
\begin{equation*}
P_{*}(A)=P_{A B}(A)=P_{A B}(A \wedge B)+P_{A B}(A \wedge \neg B) \tag{7}
\end{equation*}
$$

as in standard probability theory (since the ordering of observables is the same for all three probabilities in this expression, quantum probability reduces to standard probability in this case). When the characteristic order of observables for conjunctions is $B A$, however, we have

$$
\begin{equation*}
P_{*}(A)=P_{A B}(A)=P_{B A}(A \wedge B)+P_{B A}(A \wedge \neg B)+\delta_{A} \tag{8}
\end{equation*}
$$

where $\delta_{A}$ is a 'quantum interference' term for observable $A$. This quantum interference term represents deviation from the requirements of probability theory in the estimate $P_{*}(A)$, and arises from the difference between probabilities measured in the orders $A B$ and $B A$. Note that quantum interference is not an error term here: it is a constant that specifies the relationship between $P_{*}(A)$ and $P_{B A}(A \wedge B)+P_{B A}(A \wedge \neg B)$ for a given participant and a given pair of events $A B$. Parallel results arise for the probability of $B$, where with the characteric ordering $B A$ we have

$$
\begin{equation*}
P_{*}(B)=P_{B A}(B)=P_{B A}(A \wedge B)+P_{B A}(\neg A \wedge B) \tag{9}
\end{equation*}
$$

as in probability theory, while with the ordering $A B$ we have

$$
\begin{equation*}
P_{*}(B)=P_{B A}(B)=P_{A B}(A \wedge B)+P_{A B}(\neg A \wedge B)+\delta_{B} \tag{10}
\end{equation*}
$$

where $\delta_{B}$ is the interference term for the estimate $P_{*}(B)$.

From these expressions for $P_{*}(A)$ and $P_{*}(B)$ we derive the quantum probability model's predictions for values of the addition law and the two expansion identities (Equations 1, 2 and 3 ) in three separate situations: where observables are compatible, where the ordering of observables is $A B$, and where the ordering is $B A$. The first three lines of Table 1 shows these predictions. From Table 1 we see that, if observables are compatible, all three identities have a predicted value of 0 (contrary to experimental results). If observables are measured in the order $A B$ or $B A$, however, one expansion identity has a predicted value of 0 and the addition law and the other expansion identity have values that deviate from 0 by exactly the same magnitude but with opposite signs (contrary to experimental results). The quantum probability model's predictions are inconsistent with the experimental results in all three situations.

In quantum probability theory a conditional probability $P(A \mid B)$ is necessarily measured in the order $B A$ (with the given event occurring first and the conditional event occurring after). This means that the relationships

$$
\begin{align*}
& P_{B A}(A \wedge B)=P(A \mid B) P_{B A}(B)=P(A \mid B) P_{*}(B)  \tag{11}\\
& P_{A B}(A \wedge B)=P(B \mid A) P_{A B}(A)=P(B \mid A) P_{*}(A) \tag{12}
\end{align*}
$$

necessarily hold in quantum probability (since the probabilities in these expressions are all measured in the same order, and so follow the requirements of probability theory). We define

$$
\Delta_{A B}=P_{A B}(A \wedge B)-P_{B A}(A \wedge B)
$$

to represent the effect of order on conjunctive probability judgments $P_{A B}(A \wedge B)$ and $P_{B A}(A \wedge B)$. Then substituting from Equations 11 and 12 into the Bayes rule and 'Bayes expansion' identities (Equations 4, 5 and 6), we derive predictions in three separate situations, as before (see Table 1). Here we see that, if observables are compatible, all three identities have a predicted value of 0 (contrary to experimental results). If observables are measured in the order $A B$, one expansion identity has a predicted value of 0 and the Bayes rule and the other expansion identity have the same values, deviating from zero by $\Delta_{A B}$ (contrary to experimental results). If observables are measured in the order $B A$, one expansion identity has a
predicted value of 0 and the Bayes rule and the other expansion identity have values that deviate from zero by exactly the same magnitude of $\Delta_{B}$ but with opposite signs (again, contrary to experimental results). The quantum probability model's predictions are inconsistent with the experimental results in all three situations.

## The QQ identity

Consider our definition of $\Delta_{A B}$ to represent order effects for the conjunctive probability judgments $P_{A B}(A \wedge B)$ and $P_{B A}(A \wedge B)$. A necessary mathematical consequence of quantum probability is that exactly the same order effects apply to conjunctive probabilities $P_{B A}(\neg A \wedge \neg B)$ and $P_{A B}(\neg A \wedge \neg B)$, and so we have

$$
P_{A B}(A \wedge B)-P_{B A}(A \wedge B)=\Delta_{A B}=P_{B A}(\neg A \wedge \neg B)-P_{A B}(\neg A \wedge \neg B)
$$

and therefore the QQ identity holds for events $A$ and $B$ in the quantum probability model. Wang et al. (2014) estimate the size of the order effect in each of their 72 different polls or experimental studies via the related measure

$$
Z=\max \left\{\begin{array}{c}
\left|P_{B A}(A \wedge B)-P_{A B}(A \wedge B)\right|+ \\
\left|P_{B A}(\neg A \wedge \neg B)-P_{A B}(\neg A \wedge \neg B)\right| \\
\left|P_{B A}(A \wedge \neg B)-P_{A B}(A \wedge \neg B)\right|+ \\
\left|P_{B A}(\neg A \wedge B)-P_{A B}(\neg A \wedge B)\right|
\end{array}\right.
$$

(so that the overall order effect is equal to the summed absolute values of the order effects for $A \wedge B$ and $\neg A \wedge \neg B$, or for $A \wedge \neg B$ and $\neg A \wedge B$, whichever is greater). The greater the value of this measure, the larger the order effect. They find statistically significant order effects in most of these polls or studies, but reliable agreement with the QQ identity. The fact that this QQ identity appears to hold simultaneously with such order effects has been taken as clear evidence that 'human judgments follow quantum rules' (Wang et al., 2014).

## The probability theory plus noise model

The probability theory plus noise model assumes that people estimate probabilities via a mechanism that is fundamentally rational (following standard frequentist probability theory), but is perturbed in various ways by the systematic effects or biases caused by purely random noise or error. This approach follows a line of research leading back at least to Thurstone (1927) and continued by various more recent researchers (see, e.g. Dougherty et al., 1999, Erev et al., 1994, Hilbert, 2012). This model explains a wide range of results on bias in people's direct and conditional probability judgments across a range of event types, and identifies various probabilistic expressions in which this bias is 'cancelled out' and for which people's probability judgments agree with the requirements of standard probability theory (see Costello and Watts, 2014, Costello and Mathison, 2014, Costello and Watts, 2016a,b,c).

In standard probability theory the probability of some event $A$ is estimated by drawing a random sample of events, counting the number of those events that are instances of $A$, and
dividing by the sample size. The expected value of these estimates is $P(A)$, the probability of $A$; individual estimates will vary with an approximately normal distribution around this expected value. The probability theory plus noise model assumes that people estimate the probability of some event $A$ in exactly the same way: by randomly sampling items from their memory, counting the number that are instances of $A$, and dividing by the sample size. Since memory is subject to various forms of random error, the model assumes that items have some probability $d<0.5$ of being counted incorrectly. Given this error, a randomly sampled item can be counted as $A$ in two mutually exclusive ways: either the item truly is an instance of $A$ and is counted correctly (this occurs with probability $P(A)(1-d)$, since $P(A)$ items are truly instances of $A$, and items have a $(1-d)$ chance of being read correctly), or else the item truly is not an instance of $A$ and is counted incorrectly as $A$ (this occurs with probability $(1-P(A)) d$, since $(1-P(A))$ items are truly not instances of $A$, and items have a $d$ chance of being read incorrectly). The expected value for a noisy estimate for the probability of $A$ is thus

$$
\begin{equation*}
P_{*}(A)=P(A)(1-d)+(1-P(A)) d=(1-2 d) P(A)+d \tag{13}
\end{equation*}
$$

and we expect individual estimates $p_{*}(A)$ to vary independently around this expected value. This equation represents a pattern of regressive bias moving probability estimates $P_{*}(A)$ away from the true, objectively correct probability $P(A)$. Reasoning just as above, the model similarly predicts an expected value for the conditional probability $P(A \mid B)$ of
$P_{*}(A \mid B)=\frac{(1-2 d)^{2} P(A \wedge B)+d(1-2 d)[P(A)+P(B)]+d^{2}}{(1-2 d) P(B)+d}$
These expressions account for various observed patterns of bias in people's direct and conditional probability judgment (see Costello and Watts, 2014, 2016b).

## Addition law and Bayes rule identities

This model makes predictions about the values of various probability theory identities. If we substitute Equation 13 into the addition law identity, for example, we get an expected value of

$$
P_{*}(A)+P_{*}(B)-P_{*}(A \wedge B)-P_{*}(A \vee B)=0
$$

and so this model predicts that this expression should have a value of 0 , and just as seen in experimental results (Costello and Watts, 2014, 2016b, Fisher and Wolfe, 2014). Similarly, if we substitute Equation 13 and Equation 14 into the Bayes rule identity, we get an expected value of

$$
P_{*}(B \mid A) P_{*}(A)-P_{*}(A \mid B) P_{*}(B)=0
$$

and again the model predicts a value of 0 , just as just as seen in experimental results.

Agreement with probability theory for the addition law and the Bayes rule identity arises, in this model, despite significant regressive bias due to random noise in individual probability estimates making up these expressions. This is because
in these identities the various biases due to random noise in those individual probability estimates all cancel out. For other probability theory identities, however, this model predicts no cancellation of regressive effects. For example, substituting the expression from Equation 13 into the two 'expansion' identities (Equation 2 and 3), we get

$$
\begin{aligned}
& P_{*}(A \wedge B)+P_{*}(A \wedge \neg B)-P_{*}(A)=d \\
& P_{*}(A \wedge B)+P_{*}(\neg A \wedge B)-P_{*}(B)=d
\end{aligned}
$$

and the model predicts the same positive value for both identities, again just as observed in experimental results (Costello and Watts, 2016b, Fisher and Wolfe, 2014).

For the two 'Bayes expansion' identities (Equation 5 and 6) we get

$$
\begin{aligned}
& P_{*}(A \wedge B)-P_{*}(A \mid B) P_{*}(B) \\
& =d(1-d)-d(1-2 d)[P(A)+P(B)-2 P(A \wedge B)] \\
& P_{*}(A \wedge B)-P_{*}(B \mid A) P_{*}(A) \\
& =d(1-d)-d(1-2 d)[P(A)+P(B)-2 P(A \wedge B)]
\end{aligned}
$$

Since probability theory requires that $0 \leq P(A)+P(B)-$ $2 P(A \wedge B) \leq 1$ for all $A$ and $B$, and since $d<0.5$ by assumption, we see that
$d^{2} \leq d(1-d)-d(1-2 d)[P(A)+P(B)-2 P(A \wedge B)] \leq d(1-d)$
and values for both these identities will be distributed between $d^{2}$ and $d(1-d)$ in a way that depends on $P(A)+$ $P(B)-2 P(A \wedge B)$. The average value for $P(A)+P(B)-$ $2 P(A \wedge B)$ (across uniformly distributed probabilities that are constrained to be consistent with probability theory) is 0.5 , and so the average value for this expression is equal to $d / 2$, the centerpoint of this range. The model thus predicts the same average positive value for these identities; a value half that for the first two expansion identities. Again, this is just as seen in experimental results (Costello and Watts, 2016b).

## The QQ identity and order effects

The probability theory plus noise model, as presented above, assumes that when $P(B)$ and $P(A)$ are estimated sequentially, the value given for $P(A)$ is not influenced by the prior value given for $P(B)$. This is because the model assumes that people estimate some probability $P(A)$ by drawing a sample of items at random from memory, and counting the proportion that are $A$. To allow sequential effects in the noise model, we can relax this assumption, and say that the chance of a given item being sampled from memory is influenced by the degree to which that item is already active or 'primed'. Since the estimation of probability $P(B)$ involves drawing a sample of items and counting the proportion that are $B$, those items that were counted as $B$ are more active (are primed), and so are more likely to be included in the 'random' sample of items drawn when estimating $P(A)$, causing an order effect.

Suppose that the chance of an already primed item being sampled is $s$. Also suppose that $P(B)$ has just been estimated
in a previous sample: $P_{*}(B)$ then represents the proportion of items in that previous sample that were read as $B$. A sample is now drawn to estimate $P(A)$. Each item drawn to make up that new sample has a probability $s P_{*}(B)$ of coming from the primed set of items that were already read as $B$, and a probability $1-s P_{*}(B)$ of being drawn randomly from the set of all items in memory. For the $s P_{*}(B)$ items in our sample that were previously read as $B$, the probability of one of those items being read as $A$ is $P_{*}(A \mid B)$; this is the conditional probability of an item being read as $A$, given that it was read as $B$. For the remaining items that were just sampled randomly from memory, the probability of one of those items being read as $A$ is simply $P_{*}(A)$. Given that we have just given an estimate for the probability $P(B)$, then, the expected value for an immediately following estimate for $P(A)$ will be

$$
P_{B A}(A)=s P_{*}(B) P_{*}(A \mid B)+\left(1-s P_{*}(B)\right) P_{*}(A)
$$

and, substituting from Equations 13 and 14 and simplifying we get

$$
\begin{equation*}
P_{B A}(A)=P_{*}(A)+s(1-2 d)^{2}[P(A \wedge B)-P(A) P(B)] \tag{15}
\end{equation*}
$$

From Equation 15 we see that $P_{B A}(A) \neq P_{*}(A)$ and so $P_{B A}(A) \neq P_{A B}(A)$ will hold in this model in general, with the probability of a 'yes' answer to question $A$ when that question comes first being different from the probability of a 'yes' answer when question $A$ immediately follows question $B$. This model thus produces order effects in question answering, just as seen in experimental data.

Despite these order effects, the QQ identity also holds in this model. To see this, consider that, since $P_{*}(B)$ is the probability of answering 'yes' to a question $B$ and $P_{B A}(A)$ is the probability of answering 'yes' to a question $A$ that immediately follows a question $B$, the probability of answering 'yes' to both questions when presented in the order $B A$ is

$$
\begin{aligned}
& P_{B A}(A \wedge B)=P_{*}(B) P_{B A}(A) \\
& =P_{*}(B) P_{*}(A)+P_{*}(B) s(1-2 d)^{2}[P(A \wedge B)-P(B) P(A)]
\end{aligned}
$$

and the probability of answering 'yes' to both questions in the order $A B$ is

$$
\begin{aligned}
& P_{A B}(S \wedge B)=P_{*}(A) P_{A B}(B) \\
& =P_{*}(A) P_{*}(B)+P_{*}(A) s(1-2 d)^{2}[P(A \wedge B)-P(B) P(A)]
\end{aligned}
$$

and so

$$
\begin{align*}
& P_{B A}(A \wedge B)-P_{A B}(A \wedge B) \\
& =s(1-2 d)^{2}[P(A \wedge B)-P(B) P(A)]\left[P_{*}(B)-P_{*}(A)\right] \tag{16}
\end{align*}
$$

Using the same line of reasoning for the probability of answering 'no' to both questions, we get
$P_{A B}(\neg B \wedge \neg A)-P_{B A}(\neg B \wedge \neg A)$
$=s(1-2 d)^{2}[P(\neg B \wedge \neg A)-P(\neg B) P(\neg A)]\left[P_{*}(\neg A)-P_{*}(\neg B)\right]$
Substituting from Equation 13 and rearranging we have

## References

$$
\begin{aligned}
P_{*}(\neg A)-P_{*}(\neg B) & =(1-2 d)[1-P(A)]+d-(1-2 d)[1-P(B)]-d \\
& =P_{*}(B)-P_{*}(A)
\end{aligned}
$$

and from standard probability theory we have

$$
P(\neg A \wedge \neg B)-P(\neg A) P(\neg B)=P(A \wedge B)-P(B) P(A)
$$

and so

$$
\begin{align*}
& P_{A B}(\neg A \wedge \neg B)-P_{B A}(\neg A \wedge \neg B) \\
& =s(1-2 d)^{2}[P(B \wedge A)-P(B) P(A)]\left[P_{*}(B)-P_{*}(A)\right] \tag{17}
\end{align*}
$$

giving
$P_{A B}(A \wedge B)+P_{A B}(\neg A \wedge \neg B)-P_{B A}(A \wedge B)-P_{B A}(\neg A \wedge \neg B)=0$
and this model satisfies the QQ identity.

## Conclusions

Much research on people's probabilistic reasoning over the last 50 years has focused on the various significant biases seen in probability estimation and judgment. Invariants such as the addition law, the Bayes rule identity, and the QQ identity, which hold simultaneously with these biases, reveal an important fact: they show us that these biases are systematically and quantatitively related and can be explained mathematically. We can see this in the case of the QQ identity, where there are reliable order effects (biases) in responses which nonetheless cancel out when responses are combined in the identity. We also see this in the addition law and Bayes rule identities, where there are reliable biases in probability estimates which again, cancel out when those estimates are combined in those identities.

In this paper we've shown that, unlike the quantum probability model, the probability theory plus noise model is able explain the satisfaction of three invariants in people's probabilistic judgment (the addition law, Bayes rule and QQ identities) alongside the occurence of various forms of systematic bias in those same judgments. These results support the theoretical proposal in that account, which is that human probabilistic judgment is based on a rational process (one that follows frequentist probability theory) that is subject to random noise. It is important to stress that we are not suggesting that people's probability estimates are themselves rational. This is clearly not the case: there is very extensive evidence demonstrating that people's probability estimates are systematically biased away from the requirements of probability theory. We argue that these biases are a consequence of the influence of random noise on the probability estimates generated by an underlying rational process. While this noise is random, it has systematic, directional effects (our noisy model's expected averages for probability estimates are systematically biased away from the 'true' probability values, in a way that seems to match the biases seen in people's estimates) which are cancelled out in these three identities. This model gives a new and useful perspective on the various systematic biases seen in people's probabilistic reasoning.

Boyer-Kassem, T., Duchêne, S., and Guerci, E. (2016). Testing quantum-like models of judgment for question order effect. Mathematical Social Sciences, 80:33-46.
Costello, F. and Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. Psychological Review, 121(3):463-480.
Costello, F. and Watts, P. (2016a). Explaining high conjunction fallacy rates: the probability theory plus noise account. Journal of Behavioral Decision Making. In press, available at http://dx.doi.org/10.1002/bdm. 1936.
Costello, F. and Watts, P. (2016b). People's conditional probability judgments follow probability theory (plus noise). Cognitive Psychology, 89:106-133.
Costello, F. and Watts, P. (2016c). Probability theory plus noise: replies to Crupi and Tentori (2015) and to Nilsson, Juslin and Winman (2015). Psychological Review, 123(1):112-123.
Costello, F. J. and Mathison, T. (2014). On fallacies and normative reasoning: when people's judgements follow probability theory. In Proceedings of the 36th annual meeting of the Cognitive Science Society, pages 361-366.
Dougherty, M. R. P., Gettys, C. F., and Ogden, E. E. (1999). Minerva-DM: A memory processes model for judgments of likelihood. Psychological Review, 106(1):180-209.
Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. Psychological Review, 101(3):519527.

Fisher, C. R. and Wolfe, C. R. (2014). Are people naïve probability theorists? A further examination of the probability theory + variation model. Journal of Behavioral Decision Making, 27(5):433-443.
Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. Psychological Bulletin, 138(2):211-237.
Moore, D. W. (2002). Measuring new types of questionorder effects: Additive and subtractive. The Public Opinion Quarterly, 66(1):80-91.
Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34(4):273.
Wang, Z. and Busemeyer, J. R. (2013). A quantum question order model supported by empirical tests of an a priori and precise prediction. Topics in Cognitive Science, 5(4):689710.

Wang, Z., Solloway, T., Shiffrin, R. M., and Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. Proceedings of the National Academy of Sciences, 111(26):9431-9436.

# Experts are better than novices when imagining wines, but not odors in general 

Ilja Croijmans<br>Centre for Language Studies, Radboud University, Nijmegen, NL<br>Laura Speed<br>Centre for Language Studies, Radboud University, Nijmegen, NL<br>Artin Arshamian<br>Centre for Language Studies, Radboud University, Nijmegen, NL; Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, SW<br>Asifa Majid<br>Radboud University, The Netherlands; Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, NL; Max Planck Institute for Psycholinguistics, Nijmegen, NL


#### Abstract

Olfactory imagery is disputed to exist in novices, but is reported to be easier for smell experts. It plays an important role in wine expertise. Previous research shows experts' superior cognitive abilities do not transfer beyond their domain of expertise. This leads to two questions: do wine experts have more vivid imagery for the multisensory experience of wine? And how general is wine experts' olfactory imagery? Wine experts and novices completed a questionnaire measuring the vividness of imagery for the color, smell, and flavor of wine. In addition, all participants completed a questionnaire on general smell imagery. Wine experts were better than novices at imagining wines in all modalities, but not better at imagining smells in general. Novices reported the strongest imagery for the appearance of wine, but experts showed no difference between the senses. So mental imagery becomes more vivid with expertise; but only for imagery directly expertise related.


# The elusive oddness of or-introduction 

Nicole Cruz ${ }^{1,2}$ (ncruzd01@mail.bbk.ac.uk), David Over ${ }^{3}$ (david.over@durham.ac.uk), \& Mike Oaksford ${ }^{2}$ (m.oaksford@bbk.ac.uk)<br>${ }^{1}$ Laboratoire CHArt (EPHE), 4-14, rue Ferrus, 75014 Paris, France<br>${ }^{2}$ Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK<br>${ }^{3}$ Department of Psychology, Durham University, South Rd, Durham DH1 3LE, UK


#### Abstract

The inference of or-introduction, $p$, therefore $p$ or $q$, is fundamental in classical logic and probability theory. Yet traditional research in the psychology of reasoning found that people did not endorse this inference as highly as other onepremise valid inferences. A radical response to this finding is to claim that or-introduction is in fact invalid. This response is found in the recent revision of mental model theory (MMT). We argue that this revision of the theory leads to a number of logical problems and counterintuitive consequences for valid inferences, and present an experiment extending recent studies showing that people readily accept or-introduction under probabilistic instructions. We argue for a pragmatic explanation of why the inference is sometimes considered odd. The inference is not odd when people reason from their degrees of belief.


Keywords: or-introduction; reasoning; mental models; probabilistic approach

## The New Paradigm and earlier MMT

There has been a paradigm shift in the psychology of reasoning (Oaksford \& Chater, 2013; Over, 2009), from binary approaches focussed on drawing conclusions from arbitrary assumptions, to Bayesian and probabilistic accounts focussed on people's degrees of belief and belief revision and updating in reasoning.

## The probabilistic approach

A central position in the probabilistic approach is that most reasoning in both everyday life and science takes place under uncertainty. This uncertainty cannot be captured in classical binary logic, but it can be in probability theory (Adams, 1998; Coletti \& Scozzafava, 2002).

A basic hypothesis in this new approach is that people's degree of belief in a conditional statement, $\mathrm{P}($ if $p$ then $q)$, does not correspond to the probability of the material conditional of classical logic (which is equivalent to not-p or $q$ ), but instead to the conditional probability, $\mathrm{P}(q \mid p)$. The proposal is that people arrive at this conditional probability by performing a Ramsey test, a mental simulation in which they hypothetically add $p$ to their beliefs, make any changes necessary to preserve consistency, and judge the probability of $q$ on this basis (Evans \& Over, 2004; Ramsey, 1929/1990; Stalnaker, 1968). The identity $\mathrm{P}($ if $p$ then $q)=$ $\mathrm{P}(q \mid p)$ is generally called The Equation (Edgington, 1995) and has received strong empirical support (Evans, Handley,

Neilens, \& Over, 2007; Evans, Handley, \& Over, 2003; Oberauer, Geiger, Fischer, \& Weidenfeld, 2007; Oberauer \& Wilhelm, 2003; Fugard, Pfeifer, Mayerhofer, \& Kleiter, 2011; Barrouillet \& Geauffroy, 2015).

## Probabilities in earlier MMT

Earlier versions of mental model theory (MMT) proposed that people reason by creating a mental representation of the logical possibilities in which the premises of an inference are true (e. g. $p \& q$ is true in one possibility: that in which both $p$ and $q$ are true; whereas $p$ or $q$ is true in three possibilities: when $p$ is true and $q$ false, when $p$ is false and $q$ true, and when $p$ and $q$ are both true). Each of these possibilities is called a model. People then eliminate any models of the premises that contradict one another (e. g. if Premise 1 of an inference is $p$ or $q$ and Premise 2 is not- $p$, then people eliminate the two models of Premise 1 in which $p$ is true). Finally, people formulate an informative conclusion based on any models remaining after eliminating inconsistencies. It was further held that people make errors in reasoning because they tend to represent only what is true in a model, and to leave implicit what is false, and because they tend to leave implicit and then forget entire models.

MMT was originally formulated within the binary approach to reasoning. Hence it focussed on the truth or falsity of a statement, given the truth of some other statements, and proposed the core meaning of conditionals to correspond to the material conditional (Johnson-Laird \& Byrne, 1991, 2002).

However, MMT was early on extended to reasoning with extensional probabilities, representing these as proportions of models or numerical "tags" on models (Girotto \& Johnson-Laird, 2004, 2010; Johnson-Laird, Legrenzi, Girotto, Legrenzi, \& Caverni, 1999; c. f. Geiger \& Oberauer, 2010) in a way consistent with the rules of probability theory.

MMT has also been recently extended to subjective probabilities (Khemlani, Lotstein, \& Johnson-Laird, 2014). However, a problem with this account is the lack of clarity in its computational level specification. For example, it proposes that people intuitively grasp that the "logical relation" $p$ or not-p has a probability of 1 , but also that people intuitively compute $\mathrm{P}(p$ or $q)$ by taking the average of $p$ and $q$ (cf. Juslin, Nilsson, \& Winman, 2009) - even though the logical connective is the same in both cases. The account also provides no means for computing correct
conditional probabilities, and can therefore not account for their occurrence. This contrasts with the earlier proposal in MMT of the subset principle for computing extensional conditional probabilities (Johnson-Laird et al., 1999), which could account for both errors and normative responses.

## New MMT

Johnson-Laird and colleagues recently proposed a radical revision of MMT that claims to integrate logic further with probability (Johnson-Laird, Khemlani, \& Goodwin, 2015). The revision changes the meanings of conditionals, conjunctions and disjunctions. These statement types are still represented using the same models as before, but whereas in previous versions of MMT a statement was true when at least one of its models was true, in the new version of the theory a statement is true when all of its models are possible (Johnson-Laird et al., 2015).

A positive consequence of this revision, in our view, is that the paradoxes of the material conditional are now considered invalid. For example, the inference $q$, therefore if $p$ then $q$ is considered invalid because the model for $q$ does not establish that all three models of the material conditional are possible. But the revision creates a number of logical problems and counterintuitive consequences for other inferences (Baratgin, Douven, Evans, Oaksford, Over, \& Politzer, 2015; Over \& Cruz, in press).

## Logical problems with new MMT

## "Possible", "true", and "valid"

Johnson-Laird et al. (2015) argue that a statement is true when all of its models are possible. But it is not clear what is meant by "possible". It cannot be logical possibility, because logically the four combinations of the truth and falsity of $p$ and of $q$ ( $p \& q, p \&$ not- $q$, not- $p \& q$, not- $p \&$ not- $q$ ) are always possible unless they contain a contradiction (Baratgin et al., 2015). Moreover, with logical possibility the new version of MMT would imply that the tautology $p$ or not $-p$ is false because the $p \&$ not- $p$ model is not possible (Baratgin et al., 2015).

Yet a narrower notion of possibility does not seem to work either. This can be seen if we apply the idea that a statement is true whenever all of its models are possible to statements that have a single model. The theory then implies that a statement $p$ is true when it is possible. But it can be possible, and readily conceivable, for us to sleep a little longer tomorrow, and yet be false if we wake up early instead. Truth does not follow from mere possibility, no matter how it is defined (Over \& Cruz, in press).

A further problem arises when the notion of truth in new MMT is used to assess the validity of an inference. JohnsonLaird et al. (2015) continue to define an inference as logically valid when its conclusion is true in every case in which its premises are true. If a statement is true when all of its models are possible, then by implication an inference is valid when the truth of the premises establishes that all models of the conclusion are possible. This formulation of
the concept of validity is of course difficult to understand without a clear definition of "possible". But one way of operationalising it could be as follows. An inference is valid in new MMT when the models of the premises contain all the models of the conclusion. This operationalisation renders the paradoxes invalid. But it leads to counterintuitive conclusions for other inferences. For example, it implies that the inference $p$ or $q$, therefore $p$ is valid, even though it is counterintuitive and invalid in classical and probabilistic logics. At the same time, the account implies that the inference not- $p$, therefore not-( $p$ \& $q$ ) is invalid, but this is an intuitive inference to make, and it is valid in classical and probabilistic logics.

## Or-introduction

In what follows we focus on the inference of orintroduction, $p$, therefore $p$ or $q$. This inference is valid in classical logic and in the probabilistic approach because it is incoherent to judge that $\mathrm{P}(p)>\mathrm{P}(p$ or $q)$. It was also valid in previous versions of MMT.

Past studies using binary instructions (asking participants to assume the premises to be true, and then to judge whether the conclusion also had to be true) found that people accept the inference less frequently than other valid inferences (Braine, Reiser, \& Rumain, 1984; Orenes \& Johnson-Laird, 2012; Rips, 1983). But the probabilistic approach and previous versions of MMT agreed that the lower acceptance rate can be explained through pragmatic factors: $p$ is a stronger, more informative statement than $p$ or $q$, and so it is pragmatically infelicitous to assert $p$ or $q$ when one has enough information to assert $p$ (Grice, 1989). Orenes \& Johnson-Laird (2012) specified this position further, suggesting that the pragmatic infelicity of the inference comes from the fact that the conclusion $p$ or $q$ includes a model in which the premise $p$ is false. We agree with Grice (1989) that it is potentially misleading to assert $p$ or $q$ in a conversation, suggesting that $p$ is possibly false, after inferring $p$ or $q$ from $p$ (Gilio \& Over, 2012).

However, in the new version of MMT or-introduction is considered invalid for the same reason as the paradoxes: the model for $p$ does not establish that the three models for the disjunction are possible. This revision does not take into account the more recent finding that under probabilistic instructions (asking participants for their degree of belief in the premise, and for their degree of belief in the conclusion given their degree of belief in the premise) or-introduction is accepted to a high degree, indistinguishable from that of other, uncontroversially valid inferences (Cruz, Baratgin, Oaksford, \& Over, 2015; Politzer \& Baratgin, 2016). According to the probabilistic approach, people accept the inference under probabilistic instructions because pragmatic constraints related to what is asserted in a conversation tend to be eliminated or reduced when people are asked directly for their subjective beliefs. Conversational principles about not misleading our hearers (Grice, 1989) do not apply when we are making inferences from our own beliefs as premises to further beliefs in a subjective mental process.

The assumed invalidity of or-introduction also disables central logical and probabilistic principles. For example, if or-introduction is invalid, then so is a version of the inference in which the disjunction is "packed": p, therefore superset of $p$. For example, there is tea, therefore there is tea or coffee can be paraphrased as there is tea, therefore there is a hot beverage. Or-introduction thus enables us to establish basic set-subset relations. If people are unable to establish such relations, then the MMT account of reasoning with categorical syllogisms breaks down. Or-introduction is also used in the proofs of many fundamental theorems of probability theory, such as the theorem of total probability, $P(p)=P((p \& q)+P(p \&$ not $-q))$, which is itself derived from the fundamental logical principle that $p$ is equivalent to $(p \& q)$ or $(p \& n o t-q)$. MMT cannot integrate logic and probability theory while implying that such principles are invalid.

In what follows we analyse in more detail the relation between or-introduction and two further inferences: andelimination, $p \& q$, therefore $p$, and or-MP: if $p$ or $q$ then $r$, $p$, therefore $r$.

## And-elimination = or-introduction

And-elimination appears to be valid in the new version of MMT, because the model of the premise contains all the models of the conclusion. But the validity of andelimination implies the validity of or-introduction, and vice versa, as follows:

$$
\begin{array}{ll}
p \& q \text {, therefore } p & \\
\text { not- } p \text {, therefore not- }(p \& q) & \text { (by reductio ad absurdum) } \\
\text { not- } p \text {, therefore not- } p \text { or not- } q & \text { (by de Morgan) } \\
p, \text { therefore } p \text { or } q & \text { (by substitution of terms) }
\end{array}
$$

Mental model theory could argue that it does not accept this proof because one or more of the rules used in the derivation are itself invalid in the theory. But the invalidity of such elementary logical rules would have counterintuitive consequences for a wide range of further inferences.

## Or-MP: or-introduction through the back door

The inference if $p$ or $q$ then $r, p$, therefore $r$ can be called orMP because it is the short form of a two-step inference, in which one first uses or-introduction to infer $p$ or $q$ from $p$, then then uses $p$ or $q$ together with if $p$ or $q$ then $r$ to infer $r$ through the inference of modus ponens (MP).

Under binary instructions or-MP is endorsed to a degree at least as high as MP (Rips, 1983). The inference also appears to be valid in new MMT because the models of the premises contain all the models of the conclusion. But as outlined above, or-MP includes or-introduction as a component.

Followers of MMT might reply that, in the new version of the theory, or-MP is valid directly without the intermediate step of or-introduction. But the validity of or-MP also implies the validity of or-introduction in a direct way. If we substitute $p$ or $q$ for $r$, the resulting inference is if $p$ or $q$
then $p$ or $q, p$, therefore $p$ or $q$. The first premise is a tautology, which always holds and does not have to be assumed, and the rest is explicit or-introduction (c. f. Over \& Cruz, in press).

We conducted an experiment to test people's intuitions about the validity of or-introduction, and its relation to andelimination and or-MP, using probabilistic instructions.

## Method

## Participants

A total of 121 participants from English speaking countries completed the experiment through the online platform Prolific Academic. After removing cases that failed a test question or included trial reaction times of 3 sec or less, 112 participants remained for analysis. They had a mean age of 29 years (range: 18-73), and a varied formal-educational background. All indicated having at least good English language skills. Participants' percentage rating of task difficulty was on average $48 \%$.

## Design and materials

Participants were shown the 6 inferences of Table 1 (two further inferences investigating other questions are not discussed here due to space constraints). Inferences 1 to 5 were presented three times, with three different premise probabilities $(100 \%, 80 \%$, and $60 \%)$. Inference 6 was only presented with a premise probability of $100 \%$ because one of its premises is a tautology. Premise probability was varied with the aim of generalising the results to different premise probabilities, and was not associated with particular predictions. For each trial, participants' task was to judge how likely the conclusion of the inference can be, given the likelihood of the premise.

Table 1. The inferences investigated.

| Name | Form |
| :---: | :---: |
| 1 or-introduction | $p \therefore p$ or $q$ |
| 2 and-elimination | $p \& q \therefore p$ |
| 3 Paradox 1 | $q \therefore$ if $p$ then $q$ |
| 4 Paradox 2 | not-p $\therefore$ if $p$ then $q$ |
| 5 or-MP (a) | if $p$ or $q$ then $r, p \therefore r$ |
| 6 or-MP (b) | if $p$ or $q$ then $p$ or $q, p$ $\therefore p \text { or } q$ |

$$
\text { Note. " } \therefore \text { " = "therefore". }
$$

The experiment involved three comparisons. The first was between inference 1 (or-introduction) and inference 2 (andelimination). The probabilistic approach predicts people will give similar ratings to these two inferences because the validity of one implies the validity of the other, and asking directly for people's degree of belief in the conclusion is expected to reduce pragmatic factors that may have led to lower acceptance rates of or-introduction using binary instructions (Orenes \& Johnson-Laird, 2012; Rips, 1983). In contrast, new MMT would predict inference 2 to be
accepted to a high degree but inference 1 to be rejected, because inference 2 is valid but inference 1 is invalid in the new version of the theory.

The second comparison was between inference 1 (orintroduction) and inferences 3 and 4 (the paradoxes of the material conditional. The probabilistic approach predicts people will accept inference 1 to a higher degree than inferences 3 and 4 because the first is valid but the latter two invalid. In contrast, new MMT would predict that people will reject all three inferences to a similar degree, because they are considered invalid for the same reason in the theory.

The third comparison was between inference 1 (orintroduction) and inferences 5 and 6 (or-MP). The probabilistic approach predicts people will endorse inferences 1,5 and 6 to a similar degree, because they are all valid. New MMT would predict inference 5 to be accepted as valid (by assuming it to be computed in a direct way without the intermediate step of or-introduction), but inferences 1 and 6 to be rejected because or-introduction is invalid on their account.

On each trial participants saw an inference embedded in a pseudo-naturalistic context story. The context stories changed on every trial, and were randomly allocated to the inferences for each participant. The order of occurrence of the inferences was also varied randomly for each participant. With 8 inferences (two not reported here) and 3 probabilities (and inference 6 only being paired with a probability of $100 \%$ ), there were 22 trials overall, plus two control trials to check whether participants were paying attention.

## Procedure

After going through the instructions and three practice trials involving different inferences to those in Table 1, participants worked through the 24 trials of the experiment. They then provided demographical information and indicated whether they had taken part seriously. The final page provided debriefing information. The experiment took on average 13.2 minutes to complete.

## Results and discussion

To compare the above predictions of the probabilistic approach and the revised version of MMT, two linear mixed model analyses were performed. The procedure for model construction followed the recommendation of Barr, Levy, Scheepers, \& Tily (2013) of implementing the maximum possible random effects structure justified by the design. The models included a random intercept for participants, but random effects for material could not be included because the random allocation of materials to inferences had as a consequence that there were not enough repetitions of the same type of material within each cell of the design. Predictor variables were centred around their grand mean to avoid problems of multicollinearity when including interaction terms. Comparisons of the main $F$-test results with likelihood-ratio tests led to the same pattern of
significant and non-significant effects. Effect sizes were calculated using the formulas suggested by Snijders and Bosker (2012), requiring the use of ML as opposed to reML as estimation method. The results are displayed in Figure 1.

## Analysis 1: Inferences 1 to 5

We first fitted an overall model with inference (1 to 5) and probability $(100 \%, 80 \%, 60 \%)$ as independent variables ${ }^{1}$. Judgments of conclusion probability increased with increasing premise probability, $F(2,1568)=118.76, p$ $<.001, \eta_{\mathrm{p}}^{2}=.074$. Mean probability judgments differed between inferences, $F(4,1568)=269.09, p<.001, \eta_{\mathrm{p}}{ }^{2}=$ .334. The size of the effect of premise probability also differed between inferences, $F(8,1568)=13.06, p<.001$, $\eta_{\mathrm{p}}^{2}=.055$.


Figure 1. Judgments of conclusion probability for inferences 1 to 5 as a function of premise probability. Error bars show 95\% CIs.
${ }^{1}$ Following the notation of Snijders \& Bosker (2012), the equation for measurement $i$ of participant $j$ was given by:
$y_{i j}=\gamma_{00}+\gamma_{10} x_{1 i j}+\gamma_{20} x_{2 i j}+\gamma_{30} x_{1 i j} x_{2 i j}+u_{0 j}+\epsilon_{i j}$ This random coefficients model had 17 parameters: 1 for the fixed effect of the intercept, 4 for the fixed effect of inference, 2 for the fixed effect of premise probability, 8 for the fixed interaction between inference and premise probability, 1 for the variance of the intercept, and 1 for the residual variance. The fact that the predictors were centred is not represented in the equation due to space constraints. The equations for the other linear mixed models computed in the analyses followed the same principles, but are not reported due to limitations of space.

Figure 1 suggests that the interaction between inference and premise probability can be traced back to the lack of an effect of premise probability for inference 4 . Follow-up analyses showed that there was indeed no effect of premise probability for inference $4, F(2,224)=1.06, p=.35, \eta_{\mathrm{p}}^{2}=$ .005. However, the size of the effect of premise probability still varied between inferences $1,2,3$ and $5, F(2,1232)=$ $189.60, p<.001, \eta_{\mathrm{p}}^{2}=.017$.

Follow-up analyses to the effect of inference showed that there was no significant difference in probability judgments for inference $1(M=75.14, S E=1.34)$ and inference $2(M=$ $78.02, S E=.82)\left(F(1,560)=3.23, p=.073, \eta_{\mathrm{p}}{ }^{2}=.005\right)$. This is in accordance with the predictions of the probabilistic approach, because or-introduction and andelimination can be derived from one another as valid inferences. The result is at odds with new MMT, which predicts the second to be valid but the first invalid.

Judgments for inference 1 were higher than those for inference $3(M=59.66, S E=1.77)(F(1,560)=70.028, p<$ $\left..001, \eta_{\mathrm{p}}^{2}=.085\right)$ and than those for inference $4(M=37.58$, $S E=1.70)\left(F(1,560)=424.164, p<.001, \eta_{\mathrm{p}}^{2}=.371\right)$. This is again in accordance with the probabilistic approach, for which or-introduction is valid, but the paradoxes of the material conditional are not. It goes counter to new MMT, in which the three inferences are invalid for the same reason.

Judgments for inference $5(M=83.85, S E=1.03)$ were slightly higher than those for inference $1(F(1,560)=33.47$, $\left.p<.001, \eta_{\mathrm{p}}{ }^{2}=.044\right)$ and than those for inference 2 ( $F(1$, $672)=14.990, p<.001, \eta_{\mathrm{p}}^{2}=.022$ ). Taken by itself this finding is in accordance with both the probabilistic approach and new MMT, because both predict the inference to be valid. Small differences in the acceptance of valid inferences are not a problem for either theory, as long as the difference between responses to valid and responses to invalid inferences is larger, as Figure 1 clearly corroborates. The slightly higher acceptance of or-MP than of orintroduction and and-elimination is in accordance with the fact that or-MP includes MP as a component, and MP tends to be endorsed at ceiling. Responses to inference 5 become more consequential to the questions investigated here when compared to those of inference 6.

## Analysis 2: Inferences 1, 5, and 6

We next fitted a model with inference $(1,5,6)$ as the independent variable, using responses for a premise probability of $100 \%$. Judgments for inference 1 ( $M=85.60$, $S E=1.98$ ) were again slightly lower than those for inference $5(M=92.63, S E=1.98)(F(1,112)=8.34, p=$ $\left..005, \eta_{\mathrm{p}}^{2}=.027\right)$ and than those for inference $6(M=91.32$, $S E=2.03)\left(F(1,112)=4.69, p=.032, \eta_{\mathrm{p}}^{2}=.019\right)$. Judgments for inference 5 did not differ from those for inference $6(F(1,112)=.31, p=.58)$. Thus version (a) of or-MP (if $p$ or $q$ then $r$, $p$, therefore $r$ ) and version (b) of the inference (if $p$ or $q$ then $p$ or $q$, $p$, therefore $p$ or $q$ ) were endorsed to the same degree, even though version (b) explicitly contains or-introduction as a component. This is
in accordance with the probabilistic approach, under which the two inferences are equivalent, but at odds with new MMT, which would predict version (b) of the inference to be rejected because or-introduction is invalid in its account.

An interesting, not anticipated finding concerns the pattern of results for the paradoxes (inferences 3 and 4). For inference 4 there was no effect of premise probability, and judgments of conclusion probability were consistently low. Judgments for inference 3 were also clearly lower than those for inferences 1,2 , and 5 , but they did covary positively with premise probability.

A first account of this difference could be as follows. In the case of inference 4 , not- $p$, therefore if $p$ then $q$, premise and conclusion contain no elements in common, and so the fact that the premise is uninformative about the conclusion is clearly apparent. Without any information about the conclusion, people assign a low probability to it, expressing that it does not follow from the premise. In the case of inference $3, q$, therefore if $p$ then $q$, the premise is also uninformative about the conclusion, and so any response is coherent. But in the absence of further information, it is reasonable to infer that a given probability of $q$ will remain invariant under the assumption of $p$. It therefore makes sense for responses to covary positively with premise probability.

## General discussion

We investigated the inference of or-introduction and its relation to and-elimination and or-MP. Earlier research using binary instructions had found or-introduction to be accepted less frequently than and-elimination (Rips, 1983). The recent revision of MMT argues that or-introduction is in fact invalid. But the assumptions of this revision have inconsistencies and counterintuitive consequences for other inferences. We extended recent findings (Cruz et al., 2015; Polizer \& Baratgin, 2016) using probabilistic instructions and found that or-introduction is accepted to a high degree, indistinguishable from that of and-elimination. People's responses to or-MP were slightly higher than those for orintroduction and and-elimination, even though or-MP contains or-introduction as a component. These findings are in accordance with the predictions of a Bayesian approach to the study of reasoning, but not with those of new MMT.

With the exception of the paradoxes, the findings could also have been accounted for in earlier versions of MMT concerned with extensional probabilities (Girotto \& Johnson-Laird, 2004, 2010; Johnson-Laird et al., 1999; c. f. Orenes \& Johnson-Laird, 2012). These earlier formulations converged with the probabilistic approach in holding that or-introduction is valid, but is sometimes odd for pragmatic reasons. The results of this experiment provide further evidence for the pragmatic explanation. The findings also suggest that people tend to reason using or-introduction in a more logical and less biased way under probabilistic than under binary instructions. People's inferences from their own degrees of belief to further degrees of belief do not seem to be governed by the conventions of conversation for
speakers and hearers in open discussions, but rather by the Bayesian principles of belief revision and updating, as proposed in the new paradigm in the psychology of reasoning.

## Acknowledgments

We are grateful to the Mexican National Council of Science and Technology (CONACYT) for supporting this research.

## References

Adams, E. (1998). A primer of probability logic. Standford, US: CLSI publications.
Baratgin, J., Douven, I., Evans, J. St. B. T., Oaksford, M., Over, D. E., \& Politzer, G. (2015).The new paradigm and mental models. Trends in Cognitive Sciences, 19(10), 547-548.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. Journal of Memory and language, 68, 255-278.
Barrouillet, P., \& Gauffroy, C. (2015). Probability in reasoning: A developmental test on conditionals. Cognition, 137, 22-39.
Braine, M. D. S., Reiser, B. J., \& Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. Psychology of Learnind and Motivation, 18, 313-337.
Coletti, G., \& Scozzafava, R. (2002). Probabilistic logic in a coherent setting. Dordrecht, NL: Kluwer.
Cruz, N., Baratgin, J., Oaksford, M., \& Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. Frontiers in Psychology, 6, 192.
Edgington, D. (1995). On conditionals. Mind, 104, 235-329.
Evans, J. St. B. T., Handley, S. J., \& Over, D. E. (2003). Conditionals and conditional probability. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 321-335.
Evans, J. St. B. T., \& Over, D. E. (2004). If. Oxford, UK: Oxford University Press.
Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., \& Kleiter, G. D. (2011). How people interpret conditionals: Shifts toward the conditional event. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 635-648.
Geiger, S. M., \& Oberauer, K. (2010). Towards a reconciliation of mental model theory and probabilistic theories of conditionals. In In M. Oaksford, \& N. Chater (Eds.), Cognition and conditionals: Probability and logic in human thinking (289-308). Oxford, UK: Oxford University Press.
Gilio, A., \& Over, D.E. (2012). The psychology of inferring conditionals from disjunctions: A probabilistic study. Journal of Mathematical Psychology, 56, 118-131.
Girotto, V., \& Johnson-Laird, P. N. (2004). The probability of conditionals. Psychologia, 47, 207-225.
Girotto, V., \& Johnson-Laird, P. N. (2010). Conditionals and probability. In In M. Oaksford, \& N. Chater
(Eds.), Cognition and conditionals: Probability and logic in human thinking (103-116). Oxford, UK: Oxford University Press.
Grice, H. P. (1989). Studies in the Way of Words. Cambridge, US: Harvard University Press.
Johnson-Laird, P. N., \& Byrne, R. M. J. (1991). Deduction. Hillsdale, US: Erlbaum.
Johnson-Laird, P. N., \& Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. Psychological Review, 109(4), 646-678.
Johnson-Laird, P. N., Khemlani, S., \& Goodwin, G. P. (2015). Logic, probability, and human reasoning. Trends in Cognitive Sciences, 19(4), 201-214.
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., \& Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62-88.
Juslin, P., Nilsson, H., \& Winman, A. (2009). Probability theory, not the very guide of life. Psychological Review, 116, 856-874.
Khemlani, S. S., Lotstein, M., \& Johnson-Laird, P. N. (2015). Naive probability: Model-based estimates of unique events. Cognitive Science, 39, 1216-1258.
Oaksford M., \& Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. Thinking \& Reasoning, 19, 346-379.
Oberauer, K., \& Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 680693.

Orenes, I., \& Johnson-Laird, P. N. (2012). Logic, models, and paradoxical inferences. Mind \& language, 27(4), 357377.

Over, D. E. (2009). New paradigm psychology of reasoning. Thinking \& Reasoning, 15, 431-438.
Over, D. E., \& Cruz, N. (in press). Probabilistic accounts of conditional reasoning. In L. J. Ball, \& V. A. Thompson (Eds.), International handbook of thinking and reasoning. Hove, UK: Psychology Press.
Politzer, G., \& Baratgin, J. (2016). Deductive schemas with uncertain premises using qualitative probability expressions. Thinking \& Reasoning, 22(1), 78-98.
Ramsey, F. P. (1990). General propositions and causality. In D. H. Mellor (Ed.), Philosophical papers (pp. 145163). Cambridge, UK: Cambridge University Press. (Original work published 1929).
Rips, L. (1983). Cognitive processes in propositional reasoning. Psychological Review, 90, 38-71.
Snijders, T. A. B., \& Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sage Publications.
Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), Studies in logical theory (pp. 98-112). Oxford, UK: Blackwell.

# A Dynamic Tradeoff Model of Intertemporal Choice 

Junyi Dai (dai@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development, Lentzeallee 94<br>Berlin, 14195 Germany

Timothy J. Pleskac (pleskac@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development, Lentzeallee 94<br>Berlin, 14195 Germany

Thorsten Pachur (pachur@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development, Lentzeallee 94 Berlin, 14195 Germany


#### Abstract

The delay discounting perspective, which assumes an alternative-wise processing of attribute information, has long dominated research on intertemporal choice. Recent studies, however, have suggested that intertemporal choice is based on attribute-wise comparison. This line of research culminated in the tradeoff model (Scholten \& Read, 2010; Scholten, Read, \& Sanborn, 2014), which can accommodate most established behavioral regularities in intertemporal choice. One drawback of the tradeoff model, however, is that it is static, providing no account of the dynamic process leading to a choice. Here we develop a dynamic tradeoff model that can qualitatively account for empirical findings in intertemporal choice regarding not only choices but also response times. The dynamic model also outperforms the original, static tradeoff model when quantitatively fitting choices from representative data sets, and even outperforms the best-performing dynamic model derived from Decision Field Theory in Dai and Busemeyer (2014) when fitting both choices and response times.


Keywords: intertemporal choice; tradeoff model; dynamic models, random utility, discrimination threshold

## Introduction

Many human decisions, mundane or momentous, involve choices between outcomes that materialize at different times in the future, ranging from dieting and exercising plans to education and saving decisions. Research on such intertemporal choices has a long history and has revealed various behavioral regularities. For example, it was found that large rewards suffer less proportional discounting than small ones do (the magnitude effect; e.g., Green, Myerson, \& McFadden, 1997), and that people's preference between options that have different delays can reverse as time passes (e.g., Green, Fristoe, \& Myerson, 1994). Various descriptive models have been developed to account for these empirical phenomena. Among them, the tradeoff model (Scholten \& Read, 2010; Scholten, Read, \& Sanborn, 2014) currently appears to be one of the most promising models since it provides a unified framework for qualitatively explaining a majority of the empirical findings. Most crucially, it can account for the nonadditivity in delay discounting (e.g.,

Scholten \& Read, 2010; Scholten, Read, \& Sanborn, 2014), which eludes any model built on the notion of delay discounting.

One drawback of the tradeoff model, however, is its static nature. As a result, it lacks an account of the dynamic process leading to the explicit intertemporal choices. Nevertheless, any decision is a result of some process that unfolds in time. The characteristics of the process affect the final decision as well as process-related variables such as response time. Therefore, a static model provides only an incomplete description of intertemporal choice, and an account of the underlying dynamics is required for a more comprehensive understanding thereof.

In this paper, we propose a modified tradeoff model of intertemporal choice that has a dynamic structure and can thus account for both choice and response time data. We show that this dynamic tradeoff model can qualitatively accommodate key findings in the literature regarding both choice patterns and relationship between choices and response times. In two model-comparison analyses, we further show that the dynamic model can even outperform promising competing models when fitting empirical data quantitatively.

## The Tradeoff Model

To account for intertemporal choice, research has for a long time been mainly conceptualized using the notion of delay discounting, according to which the delay of a reward decreases its present subjective value. One major concern of this approach has been to find the most appropriate form of the discount function, which describes how subjective value decreases with increased delay length. A critical assumption in this endeavor is that each option has a discounted utility or present value independent of other competing options. Importantly, this predicts that intertemporal choice should be transitive: if, among three options $\mathrm{X}, \mathrm{Y}$, and Z , one chooses $X$ over $Y$ and $Y$ over $Z$, then he or she should also choose $X$ over $Z$.

A series of studies by Read, Scholten, and colleagues, however, demonstrated that the transitivity of intertemporal choices is sometimes violated for a triple of options $\mathrm{S}, \mathrm{M}$,
and L with increasing money amounts and delay lengths (e.g., Scholten et al., 2014). For example, when facing Option S of receiving \$30 in one week, Option M of receiving $\$ 35$ in two weeks, and Option $L$ of receiving $\$ 40$ in three weeks, some people might choose Option S over Option M, choose Option M over Option L, but choose Option L over Option S. In each pair, one option has a smaller but sooner reward (smaller-but-sooner, or SS, option), and the other has a larger but later reward (larger-but-later, or LL, option). The cyclical choice pattern suggests that such people prefer the SS options for adjacent pairs of options (e.g., Option S vs. Option M), but the LL option for the distant pair (Option S vs. Option L). By contrast, others may instead choose the LL options for adjacent pairs but the SS option for the distant pair. The former cyclical pattern can be accounted for by assuming that the amount of discounting associated with a given difference in delay (i.e., the interval between one and three weeks) is smaller when it is treated as a whole than when it is divided into subintervals (e.g., two subintervals of one week), whereas the latter implies the opposite. Together, these patterns suggest nonadditivity in delay discounting.

To accommodate violation of transitivity in intertemporal choice, several alternatives to discounting models were developed, culminating in the tradeoff model. The key difference between the tradeoff model and previous delay discounting models lies in how attributes are assumed to be processed. In contrast to the notion of alternative-wise delay discounting, the tradeoff model assumes that people process attribute information in intertemporal choice by comparing options within each attribute, and that advantages on one attribute (e.g., reward amount) are traded off against the disadvantages on the other attribute (e.g., waiting time). Such an attribute-based approach has been shown to better capture the empirical data quantitatively than the traditional alternative-based approach reflected in the delay discounting paradigm (Dai \& Busemeyer, 2014).

According to the tradeoff model, when choosing between an SS option with a money amount of $\mathrm{x}_{\mathrm{S}}$ and a delay length of $\mathrm{t}_{\mathrm{S}}$, and an LL option with a money amount of $\mathrm{x}_{\mathrm{L}}$ and a delay length of $t_{\mathrm{L}}$, a decision maker (DM) compares the effective compensation with the effective interval. Let $\mathrm{v}(\mathrm{x})$ denote a value function and $\mathrm{w}(\mathrm{t})$ denote a time weighting function. The effective compensation is defined as the difference in the value of the two money amounts, that is, $\mathrm{v}\left(\mathrm{x}_{\mathrm{L}}\right)-\mathrm{v}\left(\mathrm{x}_{\mathrm{S}}\right)$, and the effective interval is defined as the difference in the weighted delay lengths, that is, $w\left(t_{L}\right)$ $\mathrm{w}\left(\mathrm{t}_{\mathrm{s}}\right)$. In addition, the effective interval is assumed to be weighed against the effective compensation by a tradeoff function $\mathrm{Q}\left(\mathrm{w}\left(\mathrm{t}_{\mathrm{L}}\right)-\mathrm{w}\left(\mathrm{t}_{\mathrm{S}}\right)\right)$ to make the decision. The SS option should be preferred when $\mathrm{Q}\left(\mathrm{w}\left(\mathrm{t}_{\mathrm{L}}\right)-\mathrm{w}\left(\mathrm{t}_{\mathrm{s}}\right)\right)$ is larger than $\mathrm{v}\left(\mathrm{x}_{\mathrm{L}}\right)-\mathrm{v}\left(\mathrm{x}_{\mathrm{S}}\right)$, and the LL option should be preferred when $\mathrm{Q}\left(\mathrm{w}\left(\mathrm{t}_{\mathrm{L}}\right)-\mathrm{w}\left(\mathrm{t}_{\mathrm{S}}\right)\right)$ is smaller than $\mathrm{v}\left(\mathrm{x}_{\mathrm{L}}\right)-\mathrm{v}\left(\mathrm{x}_{\mathrm{S}}\right)$.

In the latest version of the tradeoff model (Scholten et al., 2014), the subjective value of a money amount $x$ is given by

$$
\begin{equation*}
v(x)=\frac{1}{\gamma} \log (1+\gamma x) \tag{1}
\end{equation*}
$$

where $\gamma$ represents diminishing absolute sensitivity to differences in money amount, the time weight of a delay length $t$ is given by

$$
\begin{equation*}
w(t)=\frac{1}{\tau} \log (1+\tau t) \tag{2}
\end{equation*}
$$

where $\tau$ represents diminishing absolute sensitivity to differences in delay length, and

$$
\begin{equation*}
Q\left(w\left(t_{L}\right)-w\left(t_{S}\right)\right)=\frac{\kappa}{\alpha} \log \left(1+\alpha\left(\frac{w\left(t_{L}\right)-w\left(t_{S}\right)}{\vartheta}\right)^{\vartheta}\right) \tag{3}
\end{equation*}
$$

in which $\kappa>0$ represents delay sensitivity, $\vartheta>1$ represents superadditivity, and $\alpha>0$ represents subadditivity. To accommodate probabilistic choice patterns (Dai \& Busemeyer, 2014), it is further assumed that the choice probability of the LL option over the SS option is given by a ratio rule, that is,

$$
\begin{equation*}
\operatorname{Pr}(\mathrm{LL} \mid\{\mathrm{SS}, \mathrm{LL}\})=\left(\frac{v\left(x_{L}\right)-v\left(x_{S}\right)}{Q\left(w\left(t_{L}\right)-w\left(t_{S}\right)\right)}\right)^{1 / \epsilon}, \tag{4}
\end{equation*}
$$

where $\varepsilon>0$ represents response noise. With these assumptions, the models can accommodate a large number of behavioral regularities in intertemporal choice, such the aforementioned magnitude effect, preference reversal, and nonadditivity in delay discounting.

## A Dynamic Version of the Tradeoff Model

One important aspect of intertemporal choice that the tradeoff model cannot explain is the recent finding regarding a relationship between choices and response times in intertemporal choice (Dai \& Busemeyer, 2014). Specifically, it was found that pairs of options that give rise to extreme choice proportions tend to be associated with faster response times than pairs with more moderate choice proportions. We refer to this relationship as the fast-andextreme effect. Because the tradeoff model is silent on the temporal dynamics underlying intertemporal choice, it lacks an account of this finding. Here we present a modification of the model to equip it with a dynamic structure while keeping its key assumption of attribute-based processing.

As the original tradeoff model, we assume that a DM performs intertemporal tradeoffs by comparing the effective intervals with the effective compensations. However, unlike the latest implementation of the model (Scholten et al., 2014), in the modified version we assume a more straightforward comparison that goes without the mediation of the tradeoff function. Specifically, we assume that $v\left(x_{L}\right)$ $\mathrm{v}\left(\mathrm{x}_{\mathrm{S}}\right)$ is directly compared to $\mathrm{w}\left(\mathrm{t}_{\mathrm{L}}\right)-\mathrm{w}\left(\mathrm{t}_{\mathrm{S}}\right)$. To accommodate the probabilistic nature of intertemporal choice, we make two further assumptions. First, both the effective compensation and the effective interval are assumed to be random, denoted as $\mathrm{V}\left(\mathrm{x}_{\mathrm{L}}\right)-\mathrm{V}\left(\mathrm{x}_{\mathrm{S}}\right)$ and $\mathrm{W}\left(\mathrm{t}_{\mathrm{L}}\right)-\mathrm{W}\left(\mathrm{t}_{\mathrm{S}}\right)$, respectively, to reflect the uncertainty in these subjective evaluations. Second, it is assumed that a decision is made when the absolute difference between the two (random) quantities is larger than a positive value; otherwise the DM acquires another sample of the effective compensation and interval without accumulating preferences from previous samples. This process continues until a decision can be made. Note that the first assumption echoes the notion of random utility in economics (e.g., McFadden, 1973), while
the second assumption is built on the concept of discrimination threshold in psychophysics (Fechner, 1860).

To derive quantitative predictions, we assume that the effective compensation and effective interval follow independent normal distributions, with respective variance proportional to the mean of the distribution. As a result, the difference between effective compensation and effective interval is also normally distributed with

$$
\begin{equation*}
\mu=\left[v\left(x_{L}\right)-v\left(x_{S}\right)\right]-\left[w\left(t_{L}\right)-w\left(t_{S}\right)\right], \tag{5}
\end{equation*}
$$

and

$$
\begin{equation*}
\sigma=\sqrt{c\left\{\left[v\left(x_{L}\right)-v\left(x_{S}\right)\right]+\left[w\left(t_{L}\right)-w\left(t_{S}\right)\right]\right\}} \tag{6}
\end{equation*}
$$

in which $c$ is a proportional constant to be estimated from the data. ${ }^{1}$ Given the assumption of non-accumulative sampling until a sufficiently large difference is obtained, the choice probability of the LL option is given by

$$
\begin{equation*}
\operatorname{Pr}(L L \mid\{S S, L L\})=\frac{\Phi\left(\frac{\mu-\delta}{\sigma}\right)}{\Phi\left(\frac{\mu-\delta}{\sigma}\right)+\Phi\left(\frac{\mu-\delta}{\sigma}\right)}, \tag{7}
\end{equation*}
$$

in which $\Phi$ represents the cumulative distribution function of a standard normal distribution, and $\delta$ denotes the smallest positive difference (i.e., the positive discrimination threshold) required to make a decision.

Because the modified model goes without the tradeoff function and the related ratio choice rule (which are critical for the original model to accommodate the nonadditivity in delay discounting), an alternative mechanism is required in order to retain this capability. To this end, we further assume that the discrimination thresholds for choosing the SS and LL options (i.e., $\delta_{\mathrm{S}}$ and $\delta_{\mathrm{L}}$ ) are different, echoing the general idea of decision bias in the literature of choice models (e.g., Busemeyer \& Townsend, 1993). ${ }^{2}$ In this case,

$$
\begin{equation*}
\operatorname{Pr}(L L \mid\{S S, L L\})=\frac{\Phi\left(\frac{\mu-\delta_{L}}{\sigma}\right)}{\Phi\left(\frac{\mu-\delta_{S}}{\sigma}\right)+\Phi\left(\frac{\mu-\delta_{L}}{\sigma}\right)} . \tag{8}
\end{equation*}
$$

To derive predictions on response time distributions for the modified tradeoff model, we assume that the time it takes to assess a sample follows a Gamma distribution with a scale parameter of $\theta$ and a shape parameter of 2 . Because empirical distributions of response time tend to be singlepeaked (rather than monotonously decreasing), we fix the shape parameter at 2 instead of 1 . The total response time is assumed to be the sum of time(s) required for all samples drawn until a decision is made, plus a nondecision time. With these assumptions, we analytically derive joint probability density functions for both choices and response times (see Dai, Pleskac, \& Pachur, 2016). Such analytical solutions are usually not available for other dynamic choice models. This ends our description of the modified tradeoff model (hereafter the dynamic tradeoff model). See Figure 1 for the dynamic structure of the model.

[^52]
## Explanatory Power of the Dynamic Tradeoff Model

Because the dynamic tradeoff model inherits the assumption of attribute-based processing, it can accommodate several key findings in intertemporal choice, including the magnitude effect, the common ratio effect, and the common difference effect. According to the magnitude effect, larger amounts appear to be discounted at a lower rate than smaller ones. For example, if a DM is indifferent between receiving $\$ 100$ now and receiving $\$ 200$ in a year, suggesting an annual discount rate of $50 \%$, then the same person would tend to prefer receiving $\$ 2000$ in a year to receiving $\$ 1000$ now, suggesting an annual discount rate lower than $50 \%$. From an attribute-based perspective, this change in discount rate can be easily explained by noticing that the effective compensation between $\$ 1000$ and $\$ 2000$ is much larger than that between $\$ 100$ and $\$ 200$, whereas the effective intervals for the two choice scenarios are just the same.


Figure 1: The dynamic structure of the modified tradeoff model of intertemporal choice.

The common ratio effect (i.e., the delay duration effect in Dai and Busemeyer [2014]) implies that increasing the delays of both options proportionally would shift people's preference toward the SS option. In this case, the change in attribute values produces a larger effective interval while keeping the same effective compensation, with the observed effect as a natural result. Finally, the common difference effect suggests that postponing both options by the same length would increase people's preference to the LL option. This effect is accounted for by the attribute-based approach together with the nonlinearity of the time weighting function (i.e., $w(t)$ ). The particular form of the function (i.e., Equation 2) entails that increasing both delays by the same length would lead to a smaller effective interval and thus shift preference towards the LL option.

With the assumption of distinct discrimination thresholds for choosing the SS and LL options, it can be shown that the dynamic tradeoff model can produce nonadditivity in delay discounting demonstrated as a violation of transitivity.

Specifically, when the probabilistic nature of choice is considered, a violation of transitivity is usually formalized as a violation of weak stochastic transitivity (WST; Davidson \& Marschak, 1959). WST requires that for a triple of options $\mathrm{X}, \mathrm{Y}$, and Z , if $\mathrm{P}_{\mathrm{XY}} \geq 0.5, \mathrm{P}_{\mathrm{YZ}} \geq 0.5$, then $\mathrm{P}_{\mathrm{XZ}}$ should also be no smaller than 0.5 , in which $\mathrm{P}_{\mathrm{AB}}$ represents the probability of choose option A over option B in a binary choice. In other words, a violation of WST occurs when $\mathrm{P}_{\mathrm{XZ}}$ $<0.5$ given the two preconditions. This is consistent with the dynamic tradeoff model. For example, for the aforementioned triple of options with increasing money amounts and delay lengths, the dynamic tradeoff model would predict $\mathrm{P}_{\mathrm{SM}}=0.53, \mathrm{P}_{\mathrm{ML}}=0.58$, but $\mathrm{P}_{\mathrm{SL}}=0.35$ when $\gamma$ $=0.05, \tau=0.01, \mathrm{c}=1, \delta_{\mathrm{S}}=0.01$, and $\delta_{\mathrm{L}}=2$, violating WST.

Besides showing intransitive intertemporal choices produced by sub- or superadditivity in delay discounting, Scholten et al. (2014) also suggested a more intricate pattern of intransitive intertemporal choice called relative nonadditivity. Specifically, intransitive intertemporal choices tend to show subadditivity when differences between delay lengths are large relative to the differences between money amounts, and show superadditivity when the differences between delay lengths are large relative to those between money amounts. To account for this pattern, Scholten et al. defined additivity in delay discounting in terms of a product rule of choice odds and showed that this definition naturally led to the pattern of relative nonadditivity. According to this definition, subadditivity occurs when

$$
\begin{equation*}
\Omega_{L S}>\Omega_{M S} \times \Omega_{L M} \tag{9}
\end{equation*}
$$

in which $\Omega_{\mathrm{XY}}$ denotes the choice odds of option X over option Y , that is, $\mathrm{P}_{\mathrm{XY}} / \mathrm{P}_{\mathrm{YX}}$, and superadditivity occurs when

$$
\begin{equation*}
\Omega_{L S}<\Omega_{M S} \times \Omega_{L M} \tag{10}
\end{equation*}
$$

According to Scholten et al., the ratio choice rule of the static tradeoff model is the key component for explaining relative nonadditivity.

The dynamic tradeoff model, which goes without the ratio choice rule, can account for the same phenomenon. According to the dynamic model, $\Omega_{\mathrm{MS}}$ and $\Omega_{\mathrm{LM}}$ tend to be smaller than 1 when differences between delay lengths are large relative to the differences between money amounts, so is $\Omega_{\mathrm{LS}}$. Given the choice rule of the dynamic model (i.e., Equation 7 for equal discrimination thresholds for choosing the SS and LL options, or Equation 8 for distinct discrimination threshold), it can be shown that the same conditions tend to render $\Omega_{L S}>\Omega_{M S} \times \Omega_{L M}$. For example, for a triple of options $\mathrm{X}, \mathrm{Y}$, and Z with increasing reward amounts of 10,11 , and 12 dollars, and increasing delay lengths of 5,10 , and 15 days, $\Omega_{\mathrm{MS}}=0.074, \Omega_{\mathrm{LM}}=0.105$, but $\Omega_{\mathrm{LS}}=0.026>0.008=\Omega_{\mathrm{MS}} \times \Omega_{\mathrm{LM}}$ when $\gamma=\tau=0.05$, c $=1$, and $\delta_{\mathrm{S}}=\delta_{\mathrm{L}}=0.05$. To the contrary, with the same set of model parameters but another triple of options $X^{\prime}, Y^{\prime}$, and Z' with increasing reward amounts of 10,20 , and 30 dollars, and the same increasing delay lengths of 5,10 , and 15 days, $\Omega_{\mathrm{MS}}=3.15, \Omega_{\mathrm{LM}}=2.30$, but $\Omega_{\mathrm{LS}}=4.14<7.26=\Omega_{\mathrm{MS}} \times \Omega_{\mathrm{LM}}$. In the first triple, the differences between delay lengths are
large relative to those between reward amounts, whereas in the second triple, the latter are large relative to the former.

Besides accounting for major empirical regularities in choice, the dynamic tradeoff model can also accommodate the fast-and-extreme effect, one robust relationship between choices and responses in intertemporal choice (Dai \& Busemeyer, 2014). According to the model, the more strongly the expected difference between effective compensation and effective interval differs from zero, the higher the probability of obtaining a difference large enough in each sample and the further away ratio of $\Phi\left(\frac{\mu-\delta}{\sigma}\right)$ to $\Phi\left(\frac{-\mu-\delta}{\sigma}\right)$ is from 1. The former leads to faster response times because fewer samples are required to trigger a decision, whereas the latter leads to more extreme choice proportions.

In summary, the dynamic tradeoff model can qualitatively accommodate all the major findings in intertemporal choice that are captured by the static tradeoff model; in addition, it can also qualitatively accommodate the fast-and-extreme effect, a prominent relationship between choices and response times that eludes the static tradeoff model. In the next section, we show further that the dynamic model can also quantitatively fit empirical data better than promising competing models.

## Quantitative Model Comparisons

We conducted two model-comparison analyses to show the power of the dynamic tradeoff model in quantitatively fitting empirical data. First, we compared it with the latest, full version of the static tradeoff model (Scholten et al., 2014) in terms of their performance in fitting choice data only. Second, we compared the dynamic tradeoff model with the best-performing model in Dai and Busemeyer (2014)—which is built on Decision Field Theory (DFT; Busemeyer \& Townsend, 1993)—with regard to their performance in fitting choice and response time data simultaneously. The DFT model assumes a sequential sampling approach and an attention shift mechanism for making intertemporal choices. Specifically, it suggests that a DM attends to either the money or the delay attribute at a time and evaluates the relevant difference between options to update his or her preference. This preference updating process continues over time as the DM switches attention between the two attributes until the preference level of one option reaches a preference threshold to trigger a decision. See Dai and Busemeyer for more details of the DFT model.

## Method

We used data from three representative empirical studies to assess the performance of the models in accounting for individual-level data. The first data set came from Study 1 in Dai (2014), in which half or all the choice questions for each individual had an immediate SS option. The second data set came from Dai (2016), which focused on the nonadditivity in delay discounting and involved only delayed SS and LL options. The third data set came from

Study 3 in Dai and Busemeyer (2014), which examined the magnitude effect, the common ratio effect, and the common difference effect, and again involved only delayed SS and LL options. All three data sets contained participants who showed the fast-and-extreme effect. A total of 138 participants contributed data to the analysis: 61 from the first data set, 40 from the second, and 37 from the third. In all three studies, the choice questions for each participant were adjusted to suit the time preference level of the individual, and each question was presented multiple times. In this way, moderate choice proportions could be induced at an individual level to better distinguish probabilistic models from one another.

The models were fitted to individual data from each data set using the predicted functions of choice probability or joint probability density functions of choices and response times. We used the SIMPLEX algorithm implemented in the fminsearch function of Matlab to find the maximumlikelihood parameter estimates of each model, which was then used to calculate the Bayesian Information Criterion (BIC; Schwarz, 1978). The BIC is a common measure for relative model performance and expresses a model's ability to capture the data, taking into account its complexity (based on the number of free parameters). A lower BIC indicates a better balance between goodness of fit and model complexity and thus a more desirable model.

To evaluate the absolute performance of the dynamic tradeoff model, we compared its predictions with the observed data in terms of the fast-and-extreme effect. Specifically, we categorized all repeatedly presented questions into five equal-interval groups regarding observed choice proportions of the LL options and then calculated the mean observed and predicted response times for each question. The observed and predicted response times within each bin were then averaged to obtain overall measures of the observed and predicted results regarding response time. The fast-and-extreme effect suggests that mean response times associated with moderate choice proportions should be longer than those with extreme choice proportions.

## Results

Table 1 presents the results of comparing the static and dynamic tradeoff models in terms of the numbers of participants whose data were better described by either model when fitting only choice data, whereas Table 2 shows the results of comparing the dynamic tradeoff model with the best-performing DFT model in Dai and Busemeyer (2014) when fitting both choice and response time data. In each comparison, the dynamic tradeoff model outperformed the other model both separately for each data set and aggregated across all data sets. ${ }^{3}$ Furthermore, Figure 2 shows that the dynamic tradeoff model reproduces the observed fast-and-extreme effect, supporting the validity of the model as a descriptive account. The difference in mean response time between questions with extreme choice

[^53]proportions (i.e., $p<0.2$ or $p>0.8$ ) and those with moderate choice proportions (i.e., $0.2 \leq p \leq 0.8$ ) was statistically significant for both observed $(t=-9.83, p<.001)$ and predicted data $(t=-5.08, p<.001)$.

Table 1: Number of Participants Whose Choice Data Were Better Described by the Static or Dynamic Tradeoff Model.

| Data Set | Static model | Dynamic model |
| :---: | :---: | :---: |
| 1 | 10 | 51 |
| 2 | 1 | 39 |
| 3 | 18 | 19 |
| Across | 29 | 109 |

Table 2: Number of Participants Whose Choice and Response Time Data Were Better Described by the BestPerforming DFT Model in Dai and Busemeyer (2014) or the Dynamic Tradeoff Model.

| Data Set | DFT model | Dynamic tradeoff model |
| :---: | :---: | :---: |
| 1 | 18 | 43 |
| 2 | 14 | 26 |
| 3 | 15 | 22 |
| Across | 47 | 91 |



Figure 2. Average mean response times for questions with different observed choice proportions of the LL options.

Error bars show 95\% confidence intervals.

## Discussion

The static tradeoff model (Scholten et al., 2014) represents one of the most successful cognitive models to describe intertemporal choice. However, up to now there have been no attempts to examine how this modeling approach could be extended to also account for the dynamics of the underlying decision process. Here we developed a dynamic modification of the tradeoff model, which can accommodate not only key choice regularities but also the response time data and prominent regularities therein (e.g., the fast-andextreme effect). We also showed that this modified model quantitatively outperforms the original static tradeoff model when fitting choice data and the best-performing DFT model in Dai and Busemeyer (2014) when fitting both
choice and response time data. The model's ability to capture the data both qualitatively and quantitatively underlines the value of developing dynamic accounts of intertemporal choice for a better understanding of this central topic in both psychology and economics.

## A General Framework for Developing Dynamic Models of Choice

When developing the dynamic tradeoff model, we invoked and combined two time-honored concepts: the notion of random utility in economics and the concept of discrimination thresholds in psychology. Combining these concepts seems to offer a promising, but so far neglected, approach to developing dynamic choice models, and we argue that it could be applied to transform existing static models of choice also in other domains. For example, it could be applied to extend static models of risky choice into dynamic ones as long as the corresponding models can reasonably offer a measurement of the relative attractiveness of each option and the variability thereof. With dynamic models, both choice and response time data from empirical studies can be utilized to compare competing models for a more powerful model selection. Dai, Pleskac, and Pachur (2016) conduct a more comprehensive development and analysis of such a random-utility-with-discriminationthreshold (RUDT) framework, and compare it to other dynamic approaches to modeling intertemporal choice.

## Future Directions

In addition to the fast-and-extreme effect discovered in Dai and Busemeyer (2014), recent studies (Dai, 2014) have suggested another striking but less common relationship between choices and response times in intertemporal choice. Specifically, it was found that, within each choice question, the option chosen more frequently also tended to be chosen more quickly than the other option. Unfortunately, this fast-and-frequent effect poses a severe challenge to both the best-performing DFT model in Dai and Busemeyer and the dynamic tradeoff model developed here. Both models predict that the conditional response time distribution given choosing one option should be identical to that given choosing the other option. As a consequence, the option chosen more frequently is predicted to have the same mean response time as the other option, contradicting the fast-andfrequent effect. It is possible, however, to modify the dynamic tradeoff model to accommodate this effect (Dai et al., 2016). Specifically, by assuming that the discrimination thresholds are not fixed across successive samples but converging, it is possible to account for the pattern. To put DFT models of intertemporal choice on equal footing, attempts should be made to improve them as well. Future research should explore alternative forms of the tradeoff model under the RUDT structure and compare them with appropriate competing models to examine the performance of the dynamic tradeoff model.

## Conclusion

Most existing models of intertemporal choice, including the original tradeoff model, are static and thus lack a proper account of the dynamic processes leading to a choice. In this paper, we showed how the static tradeoff model can be modified into a dynamic one with a general structure built on the concepts of random utility and discrimination threshold. The advantages of the dynamic tradeoff model are demonstrated by its capability to qualitatively accommodate empirical findings and its better performance in quantitative model comparisons. Future studies should further explore the capacity of this approach for explaining more phenomena in intertemporal choice and beyond.

## References

Busemeyer, J. R., \& Townsend, J. T. (1993). Decision field theory: A dynamic cognitive approach to decision making in an uncertain environment. Psychological Review, 100, 432-459.
Dai, J. (2014). Using test of intransitivity to compare competing static and dynamic models of intertemporal choice. Doctoral dissertation, Department of Psychological and Brain Sciences, Indiana University, Bloomington.
Dai, J., \& Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. Journal of Experimental Psychology: General, 143, 1489-1514.
Dai, J., Pleskac, T. J., \& Pachur, T. (2016). Dynamic cognitive models of intertemporal choice. Manuscript submitted for publication.
Davidson, D., \& Marschak, J. (1959). Experimental tests of a stochastic decision theory. In C. W. Churchman \& P. Ratoosh (Eds.), Measurements: Definitions and theories. New York, NY: Wiley.
Fechner, G. T. (1860). Elemente der Psychophysik [Elements of Psychophysics]. Leipzig, Germany: Breitkopf \& Haertel.
Green, L., Fristoe, N., \& Myerson, J. (1994). Temporal discounting and preference reversals in choice between delayed outcomes. Psychonomic Bulletin \& Review, 1, 383-389.
Green, L., Myerson, J., \& McFadden, E. (1997). Rate of temporal discounting decreases with amount of reward. Memory and Cognition, 25, 715-723.
McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in econometrics. New York, NY: Academic Press.
Scholten, M., \& Read, D. (2010). The psychology of intertemporal tradeoffs. Psychological Review, 117, 925944.

Scholten, M., Read, D., \& Sanborn, A. (2014). Weighing outcomes by time or against time? Evaluation rules in intertemporal choice. Cognitive Science, 38, 399-438.

# What is Learning? A Definition for Cognitive Science 

Jim Davies (jim@jimdavies.org)<br>2202B Dunton Tower, Institute of Cognitive Science, Carleton University<br>1125 Colonel By Dr., Ottawa, Ontario, K1S 5B6, Canada


#### Abstract

Many intuitive notions of "learning" do not support the diverse kinds of learning across different situations and learners. In this paper I offer a functional definition of learning from a cognitive science perspective, which attempts to account for the presence of learning in different physical substrates. The definition is that a particular event should be considered a good example of "learning" to the degree to which the following characteristics describe it: 1) a system undergoes change to its informational state or processing 2) the change is for the purpose of more effective future action, 3 ) the change is in response what the system experiences, and 4) the system executes the change, rather than some outside force. Episodes are better examples of learning according to how many of these characteristics they have. I discuss benefits and limitations of this characterization.


Keywords: learning; philosophy; conceptual analysis; cognitive science; functionalism; substrate neutrality

## Introduction

According to Daniel Reisberg (Wilson \& Keil, 2001), learning "can be understood as a change in an organism's capacities or behavior brought about by experience." The Oxford Companion to Philosophy defines it as "the acquisition of a form of knowledge or ability through the use of experience." These examples are reasonable and intuitive first passes, but are not defended.

Perhaps the simplest definition of learning would be "the creation of memory," but this merely pushes the definitional difficulty to the term "memory." Nevertheless, this discussion will assume that all memory-creating processes are examples of learning (as a sufficient condition), though I will not use the term in the definition.

In this article, I will present and defend a definition of "learning" for cognitive science. My goals are that this definition will cover all accounts of learning that we observe in natural and artificial systems, and reject cases of change that should not be considered learning.

My approach assumes a version of functionalism as applied to mental concepts: that many entities in our world should be defined not
by their physical properties, but by how they interact in an information processing system.

The definition is that a particular event should be considered a good example of "learning" according to the degree to which the following characteristics describe it: 1) a system undergoes change to its informational state or processing, 2) the change is for the purpose of more effective future action, 2) the change is in response what the system experiences, and 3) the system executes the change, rather than some outside force. This is a "family resemblance" characterization, rather than one of necessary and sufficient conditions, though in this paper I will discuss the characteristics as though they were necessary for purposes of clarifying the benefits and drawbacks of including each one.

## What Can Learn?

Learning is prototypically thought of as something animals do. But some plants have a limited form of memory, and the encoding of this memory can be considered a form of learning. The venus flytrap, for example, has hairs around its trap to detect the presence of food. These hairs have haptic sensors. But the trap will not close immediately upon triggering these sensors, which is good, because closing and opening the trap is expensive in terms of energy and, for opening it again, time. So the plant will only close when another sensor detects touch within 20 seconds of the first touch elsewhere-effectively detecting a bug walking across the plant. This prevents the trap from closing when hit with raindrops, twigs, or other non-food entities (Chamovitz, 2012). This is a very simple, very short-term memory. But even in humans we do not require that encodings be long-term to be considered memory, such as the phonological loop (Baddeley, 1992). Because we classify some explicitly short-term stores as memory in humans suggests that it is
sensible to refer to flytraps as systems that can, because they can create memories, learn.
Perhaps even more surprising are examples of learning recently discovered in single-celled organisms (Boisseau, Vogel, \& Dussutour, 2016). Even slime molds can habituate to stimuli.

Our immune systems effectively remember past experiences to better deal with future infections. Immune system learning behaves a bit like classifier systems in artificial intelligence (Farmer, Packard, \& Perelson, 1986), and can even overlearn, as seen in autoimmune disorders.

Beyond the vast variety of organisms that can learn, we also have an entire field of machine learning to consider: pieces of software created by human beings that learn. Neither immune systems, plants, AIs nor slime mold cells have nervous systems, but they are capable of limited kinds of learning, suggesting that there should not be a biological, let alone a neuronal, condition.

## Human Learning at Different Levels of Analysis

When we examine human learning, we can see it happening at many levels of analysis. I will use a running example of learning to avoid eating food that makes one ill.

Neuroscientists now know a lot about how association and feedback can change neurons and how they communicate. This can happen, for example, through synaptic changes: neurons encode association through long-term potentiation and depression (associative learning; Cooke \& Bliss, 2006), and can engage in supervised learning (Ishikawa, Matsumoto, Sakaguchi, Matsuki, \& Ikegaya, 2014). An immediate nausea response could trigger an instance of supervised learning, "punishing" neurons that were involved with consuming the food, and their relationships to sensing that that food was present.

Synapatic changes in taste receptors allow us to habituate to bitter foods and drinks-children sometimes vomit when they first taste the foods that many adults enjoy. We evolved to dislike bitter foods, generally speaking, because they are more often lacking in nutrition (Sandell \& Breslin, 2006). Eating bitter foods that don't sicken us gradually habituates the sensors in our tongue to the particular taste.

In addition to synaptic changes, the brain learns through creation, movement, destruction, and the changing of the shape of neurons.
Moving up to the information processing level, every major cognitive architecture has a theory of learning. The most popular control system used in cognitive architectures is the production system (e.g., ACT-R, Soar, EPIC, and OpenCog all use them). When something bad happens, recentlyfired productions are "punished," making the system less likely to get itself in the same situation again. In connectionist architectures, learning changes connection weights in neural networks using learning algorithms such as backpropagation (Chauvin \& Rumelhart, 1995).

At the behavioral level, we can describe a person's reluctance to eat a food that previously made them sick with the theory of conditioning.
I have shown how the same event of an individual agent learning not to eat a particular food can be effectively described as learning at different levels, but we might describe other examples of learning in distributed cognitive systems (Hutchins, 1995). A theater company might better learn how to market its performances, or a game development team might better learn how to use feedback from user testing to make better products.
The idea of distributed cognitive systems (and the related notion of extended minds) is controversial (Davies \& Michaelian, 2016), but those who accept their existence would probably consider them capable of learning.

## "Systems" Learn

To conclude this section, it is a mistake to define learning so that only humans and other animals are included. We can see learning in single-celled organisms, artificial intelligences, immune systems, and plants. Nor can we define learning as something only "agents" do. Distributed cognitive systems learn, but these are, perhaps, not best described as "agents" or "organisms."
As such, I suggest the term "system," meaning a complex of elements that engage in information processing in pursuit of goals or preferences, be they explicit (as in a person's desire to be not hungry) or implicit.

## Information Processing

In the proposed definition, the changes to the system need to be changes to the representation or processing of information. For purposes of this definition, information is defined as anything that has a representational use in a system-be it symbolic or subsymbolic. That is, the information stands for something else, be it a physical referent in the world, a utility to the system, an internal category, or anything else.
To make this clearer, I will describe systems that are not information processors. A memoryfoam mattress changes in response to your lying down on it. It does so for purposes of your comfort-so it even has a function. It was designed to adapt to the environment, just as machine learning programs were designed to adapt to theirs. The mattress even has the word "memory" in its label.

According to my definition, the mattress is a poor example, because the change is not informational. The change in topography of the mattress does not mean anything to anybody in its normal use (if you came home and found that nobody was on your mattress but there was a deformity, you might use that deformation to conclude someone had been on it recently. In this case, the mattress deformation becomes a representation (to you), and is arguably a part of some distributed processing information system including you and the mattress.)
Similarly, a knife is not learning when you sharpen it, and your muscles are not learning when they get stronger because of a workout.

But all of these cases are merely physical changes, and in learning, these physical changes are important only because they encode changes to information storage and processing. Changes in knife sharpness and muscle tone are functional changes, but not of information processing systems. Instances of biological plasticity that do not involve information processing (like the growth of a callus) are not considered learning.

## The Purpose of More Effective Action

The intuitive notion of learning is that when the system learns something, it is somehow improved. It either knows something it didn't before, or is
able to do something it couldn't before, or can do it better.
This characteristic poses some immediate problems, because not everything people learn is good for them. If people tell you something that isn't true, and you believe them, then you have learned something false. And even though some false beliefs might help us, we can assume that, in general, false beliefs lead to poorer behavior (mental or physical) in the future.

Some learned behaviors are bad for us. In the case of post-traumatic stress disorder (PTSD), we learn behaviors that are problematic in nontraumatic situations (such as diving beneath the table whenever a helicopter flies by, or having nightmares that plague one for years; see Levin, 2000). I'll refer to learning false things, and the learning of maladaptive behaviors as "bad learning."
For the definition to be able to include bad learning, it is insufficient to say that learning must always leads to better behavior. However, we can avoid the problem by saying that it learns with the purpose of better future behavior.
I will explain with an analogy to digestion. We might describe the purpose of digestion as altering large, insoluble food molecules into smaller molecules that can be used as nutrition. The fact that we can digest poisons and non-nutritional food does not mean that the function of digestion isn't to nourish the organism. A system can be used poorly without removing its function. For the same reason, just because we can engage in bad learning does not mean that the function of learning isn't to promote better future behavior, nor that those bad things aren't learned.

Similarly, we remember lots of true but trivial facts that we might not productively use (or, indeed, even retrieve) ever again. In these cases, too, these declarative memories are not being used for better future action. But they are encoded because they might be useful someday. The mind remembers things without the certainty of what, exactly, will and won't be useful in the future. Will it be important to remember that Jill was wearing a red sweater? Probably not, but if we need to describe her to someone else, that fact might turn out to be useful.

We can see how memory is biased in terms of what it expects will be useful, however. For example, people tend to better remember things likely to be relevant to future events. The existence of a push pin will be better recalled if it is on the floor, where it might be stepped on, than if it is safely in a box (Zwaan, Van den Broek, Truitt \& Sundermeier, 1996). Words related to survival are better remembered than other words (Nairne, 2010).

When an organism gets hit in the head, and suffers some deficit as a result, we would not want to consider this learning. Although brain damage affects the information processing of a system, the purpose of being hit in the head (if there even is one) is not to promote better future behavior, so the definition excludes this.
Another challenging example is the deliberate, direct physical change to a brain. When a doctor performs neurosurgery, or prescribes psychoactive medication, the purpose is better future behavior. If, in the future, we are able to "download" skills directly into our heads, as is done in Matrix films, should this be considered learning? In this account it is a bad example of learning, because the system is not changing itself. However, if, somehow, somebody managed to brain surgery on oneself, then my account would have to accept that as learning, strange as it sounds.

We sometimes deliberately alleviate mental tiredness by taking a rest, drinking coffee, or eating something. These activities have the purpose (among others, perhaps) of better future behavior. And some of these examples are the agent changing itself. Although rest and consuming coffee and food might be best described at a biological level, rather than at an information processing level, it is likely that there is an information processing level of description of how these activities promote better behavior. My definition includes these activities as decent examples of learning. The only characteristic missing is "experience," because the psychological experience of doing brain surgery on oneself or drinking coffee is not what causes the change (beyond placebo effects).

This raises the question of what counts as "experience." A body can experience hair loss at a
barber, arguably, but what we want to capture here should not include experiences irrelevant to a cognitive system. I suggest that we ignore consciousness and say that an experience is limited to what the sensory apparatus of the system can detect. For an immune cell, it has receptors for detecting pathogens. A committee has analogues to sensory apparatus in the sense organs of the people that make it up.

Should the system be required to change itself for it to be considered learning, or are outside forces acting on a system acceptable? I will deal with issues regarding this question next.

## Cultural and Evolutionary Learning

Some might want to describe learning at the sociological level. For example, in Fiji there is a cultural taboo: pregnant and lactating women may not eat certain kinds of fish. It turns out that avoiding consumption of these fish reduces a woman's chances of being getting fish poisoning by $30 \%$ during pregnancy and $60 \%$ during breastfeeding (Henrich \& Henrich, 2010).

It is common for cultural taboos to have practical value that the people in the culture are not aware of. Often these are framed in terms of religion (for an example, see Harris, 1978). These taboos are refined over the course of generations. No single individual need engage in learning for this to happen, though individuals encode the information state of the cultural system. If we look at culture as an evolving entity, and, in particular, the ideas in the culture as undergoing evolutionary selection, we can see how ideas that facilitate reproduction will have a better chance of enduring over the years than others (Richerson \& Boyd, 2008).

What we observe, then, is that the society itself is doing the learning. The society, in this respect, is a cognitive system that is distributed over time, and we can observe the information changes it makes to act better in the future.

One might also look at a species as a system that learns through Darwinian evolution. Sweller and Sweller (2006) suggest that this happens, analogically mapping long-term memory with a genome; learning from other humans with biological reproduction; problem solving with
random mutation, and so on. Although I have not found an analogous argument for culture, it seems that one could easily be made.

But is evolutionary change "for the purpose of better future behavior?" We often can take a design stance to evolutionary processes to help us understand them, but biologists take great pains to make it clear that evolution is not goal-directed. Darwinian evolution is not purposeful (unless it is artificial selection, or is designed by a programmer in a simulation).

Specific behavioral phenotypes can be described as having purposes. As Daniel Dennett describes it, cuckoo chicks push other birds out of the nest. As scientists, we can ask why, and get a description at the level of neurons, but it is also profitable to look at the function of this behavior: to maximize resource acquisition from the cuckolded parent bird (1987). The function is a "free floating rationale." But application of this to the evolutionary process itself is more problematic. The products of evolution might be purposeful, even if evolution itself is not.

My point here is not take a strong stance on whether or not the changes to cultures and species that we see over time should count as learning, but to discuss how different definitions of learning would or would not include them. The definition I'm suggesting in this paper would render these poorer examples, because the changes are not for the purpose of better future behavior in genetic nor in cultural evolution, the changes are (arguably) not occurring through experience (can a culture or species experience something?) and finally because the system is not changing itself (this is clear for the genome, and possibly true for a culture). We still might metaphorically describe them as learning, and doing so might help us understand or teach these concepts.

## Limitations of the Analysis

"Learning" happens to be a word in English, the lingua franca of science. However, we need to be careful not to assume that the existence of a word means that it necessarily refers to a natural kind. Other languages might break up the world in different ways, and ultimately whether learning
exists in a way that happens to be captured by the English word for it is an empirical question.
This paper is in the tradition of a classical-styled conceptual analysis, looking for and suggesting conditions for what would count as an instance of "learning," and this is, admittedly, old-fashioned.

Is there a better way to do it? An earthquake can be described and explained using theories and equations from geology, but it turns out that these same theories apply to quakes that happen elsewhere as well-moons, starts, other planets, etc. Thus it makes sense to suggest that the idea of a "quake" extends beyond those that happen on Earth (United States Geological Survey, 2012).

This makes sense because we have a theory that is broadly, and successfully, applied. Admittedly, this is not happening with learning. Perhaps future descriptions of learning will be more theorybased. That is, we come up with a theory of learning (or a particular kind of learning), and then see to which phenomena in the world the theory can be productively applied. These future investigations might mean that "learning," as we conceive of it in English, isn't a sensible scientific category at all (Churchland, 1989, suggests that no sensible scientific categories should be based on folk psychology).

However, there is no general theory of learning yet, and if we think of cognitive science as the study of cognition independent of the substrate that supports it, it is helpful to have some idea of what we mean by learning. This paper is intended to be a start to the discussion, and more of a stepping-stone for future refinement rather than the final answer.

## Conclusion

We've known for a long time that the search for necessary and sufficient conditions for concepts is often a fruitless task, so the definition should be seen as a list of family-resemblance features. My suggested definition is that an event is a better fit for the category "learning" depending on the degree to which the characteristics in the following list describe it:

1. The change happens to an information processing system
2. The change happens with the purpose of better future action
3. The change happens in response to the system's experience
4. The change is executed by the system itself, rather than some outside influence

This definition covers the intuitive and prototypical instances of learning, but renders as poor examples some processes that we might want to productively talk about as learning, such as evolutionary processes over species and cultures.

With hope, future research will ground the definition of learning in a theory of learning process, in contrast to my attempt to define it from a conceptual analysis.

## Acknowledgments

Thank you for the helpful comments of the attendees of a brownbag talk I gave on this topic at Carleton University in 2016.

## References

Baddeley, A. (1992). Working memory. Science, 255 (5044), 556.
Boisseau, R. P., Vogel, D., \& Dussutour, A. (2016, April). Habituation in non-neural organisms: evidence from slime moulds. In Proceedings of the Royal Society B, 283(1829), 20160446).
Chamovitz, D. (2012). What a plant knows: A field guide to the senses. Scientific American: New York. Pages 50--63.
Chauvin, Y., \& Rumelhart, D.E. (1995). Backpropagation: theory, architectures, and applications. Psychology Press.
Churchland, P. M. (1989). A neurocomputational perspective: The nature of mind and the structure of science. MIT press.
Cooke, S. F., \& Bliss, T. V. P. (2006). Plasticity in the human central nervous system. Brain, 129(7), 1659-1673.
Davies, J. \& Michaelian, K. (2016). Identifying and individuating cognitive systems: A taskbased distributed cognition alternative to agentbased extended cognition. Cognitive Processing. 17(3), 307-319.

Dennett, D.C. (1987) The Intentional Stance, Cambridge, MA: MIT Press.
Farmer, J. D., Packard, N. H., \& Perelson, A. S. (1986). The immune system, adaptation, and machine learning. Physica D: Nonlinear Phenomena, 22(1), 187-204.
Harris, M. (1978). India's sacred cow. Human Nature, 1(2), 28-36.
Henrich, J. \& Henrich, N. (2010). The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. Proceedings of the Royal Society B: Biological Sciences, 277(1701), 3715-3724.
Ishikawa, D., Matsumoto, N., Sakaguchi, T., Matsuki, N. \& Ikegaya, Y. (2014). Operant conditioning of synaptic and spiking activity patterns in single hippocampal neurons. The Journal of Neuroscience. 34(14), 5044-5053.
Levin, R. (2000). Nightmares: Friend or foe? Behavioral and brain sciences, 23(06), 965.
Nairne, J. S. (2010). Adaptive memory: Evolutionary constraints on remembering. Psychology of Learning and Motivation, 53, 1-32.
Richerson, P. J. and Boyd, R. (2008). Not by genes alone: How culture transformed human evolution. Chicago, IL: University of Chicago Press.
Sandell, M. A., \& Breslin, P. A. (2006). Variability in a taste-receptor gene determines whether we taste toxins in food. Current Biology, 16(18), R792.
Sweller, J., \& Sweller, S. (2006). Natural information processing systems. Evolutionary Psychology, 4(1), 434-458.
United States Geological Survey, (2012). Earthquake hazards program. Retrieved 5 April from http://earthquake.usgs.gov/learn
Wilson, R.A., \& Keil, F.C. (2001). The MIT encyclopedia of the cognitive sciences. MIT press.
Zwaan, R.A., Van den Broek, P., Truitt, T.P., \& Sundermeier, B. (1996). Causal coherence and the accessibility of object locations in narrative comprehension. Abstracts of the Psychonomic Society, 1, 50.

# Re-representation in comparison and similarity 

Samuel Day (days@susqu.edu)<br>Jennifer Asmuth (asmuth@susqu.edu)<br>Department of Psychology, Susquehanna University<br>514 University Ave., Selinsgrove, PA 17870 USA


#### Abstract

Re-representation is a crucial component of structure mapping theory, allowing individuals to notice structural commonalities between situations that do not initially have identical relational representations. Despite its theoretical importance, however, this concept has been the subject of very little empirical work. In two experiments, we find that a case's participation in one comparison systematically changes its perceived similarity to new cases, in a pattern consistent with re-representation. Additional work rules out alternative explanations based on relational priming.


Keywords: analogy; re-representation; similarity

## Introduction

Consider how similar you find the following events:

- Nicole finally got out of the bad relationship that had prevented her from pursuing her own interests.
- As the zoo keeper was busy cleaning its habitat, the Burmese python was able to escape its open cage.

If you are like the participants in our research, you were probably willing to call these events fairly similar, at least after a little bit of consideration. In a very literal sense, these cases differ in significant ways-in their settings, their characters, their implications, even in the species of their protagonists. At a more abstract level, however, they share important structural features. Specifically, both situations describe characters who are able to escape from a confining environment.

The dominant model for understanding structured comparisons such as these is Gentner's $(1983,1989)$ structure mapping theory (SMT). According to this model, individual cases involve hierarchically-structured mental representations of labeled relations, each of which may take other relations or entities as arguments. For example, the common relational structure in the sentences above might be conveyed through a proposition such as: ESCAPED_FROM(ESCAPER, CONFINING_ENVIRONMENT). In this formulation, ESCAPED_FROM is a relation: it describes a relationship between multiple entities, and is therefore represented as a predicate that takes multiple distinct arguments. These arguments represent the assignment of entities to the relation's roles.

Comparison, according to SMT, involves a process of mapping in order to establish a structural alignment between the representations. The goal of this process is to define correspondences between the representations while following certain important rules and constraints. For example, although two corresponding objects may be quite dissimilar (e.g., Nicole and the python), relations in two representations will only be
mapped to one another if they are semantically identical. Another constraint, the principle of one-to-one correspondence, states that each element in one representation may be mapped to no more than one element in the other. Additionally, if relations in two representations correspond to one another, those relations' arguments must also correspond (the principle of parallel connectivity).

In the example cases above, the ESCAPED_FROM relations in the two representations would be placed in correspondence, which is allowed because they are identical. In order to maintain parallel connectivity, the arguments of those relations would then be mapped in a role-consistent way, despite their surface differences: Nicole would correspond with the python (they are both "escapers"), and her bad relationship would correspond with the snake's cage (as the confining environments).

Structure mapping theory has been a very successful model for understanding a wide range of cognitive phenomena, including similarity, analogy, classification and knowledge transfer (see Markman \& Gentner, 2001). However, in the basic form described above, it would quickly run into significant problems in the real world. For example, as noted, SMT asserts that relations may only be mapped to one another if they are semantically identical. However, it is not difficult to find cases that are perceived as analogically similar despite having nonidentical relations. For example, people can easily recognize the structural similarity between Bill drove to the store and Bill jogged to the store, even though their relations do not perfectly match (Gentner \& Kurtz, 2006). Or consider the sentences John is taller than George and Martha is shorter than Mary. Despite the conspicuous appropriate mapping (John and Mary are both taller), strict enforcement of the identity requirement would lead to a failed match, since TALLER_THAN and SHORTER_THAN are clearly not the same.

Fortunately, researchers have proposed a way around this problem. Specifically, it is theorized that representations may undergo a process of re-representation, in which structural and conceptual changes occur in order to enable potential relational matches (see Falkenhainer, Forbus \& Gentner, 1989; Holyoak, Novick \& Melz, 1994; Kotovsky \& Gentner, 1996; Kurtz, 2005). A variety of methods for re-representation have been proposed. For example, a cognitive system may store information about the similarity of different relations (e.g., knowing that drive is relatively similar to jog; Holyoak \& Thagard, 1989), or may initiate a search for common superordinate relations (e.g., both drive and jog are examples of move; Falkenhainer, et al., 1989). Another approach is to decompose a relation into its component structure (e.g., Gentner, 1983; Gentner \& Kurtz, 2006). For example, buying a book and taking a book do not initially
contain a match, but their relations can be decomposed into representations such as:

- BUY(book) $\rightarrow$ CAUSE(PAY_FOR(book), OBTAIN(book))
- TAKE(book) $\rightarrow$ CAUSE(PICK_UP(book), OBTAIN(book))
which would reveal an identically-matching predicate: OBTAIN.
Despite the importance of re-representation to the overall theory of structure mapping, however, it has been the subject of very little empirical work. The primary experimental research directly addressing the issue comes from Gentner and Kurtz (2006). Participants in their studies were willing to call two sentences analogous when the verbs were nearly synonymous (Fred reclined on the couch and Carl lay on the couch) or semantically "near" to one another (Fred reclined on the couch and Carl sat on the couch), but not when they were semantically "distant" in meaning (Fred reclined on the couch and Carl sneezed on the couch). Interestingly, response times were significantly longer when judging the "near" verbs than the synonyms. This finding is interpreted as evidence for rerepresentation, which would have required additional processing in order to determine a match. The authors also reported a tendency for participants to use new language (terms not present in either sentence) in their later justifications for their similarity ratings, consistent with a change in how those cases were represented. They acknowledged, however, that this might have reflected processes occurring during the justification task itself rather than the initial comparison.

In our studies, we use similarity ratings to assess potential rerepresentation. Similarity is a fundamental psychological process thought to play a role in everything from stimulus generalization in classical conditioning (Pavlov, 1927; Shepard, 1987) to categorization (e.g., Smith \& Medin, 1981), retrieval (e.g., Hintzman, 1984), inference (e.g., Osherson, et al., 1990) and problem solving (e.g., Ross, 1987). Similarity ratings have also been used successfully in prior research as a measurement of representational change (e.g., Boroditsky, 2007; Goldstone, Lippa \& Shiffrin, 2001). In the present experiments, we examine whether participation in one comparison can alter a case's mental representation in a way that changes its perceived similarity to new cases.

For example, consider the similarity between these cases:

- While testing a network security system, the computer scientist inadvertently released a destructive virus onto the internet.
- As the zoo keeper was busy cleaning its habitat, the Burmese python was able to escape its open cage.

Participants in our studies were able to recognize important structural commonalities between the described events, and responded with fairly high similarity ratings. In this case, the two situations are similar because they both describe someone inadvertently releasing something dangerous. According to SMT, this perceived similarity would require them to establish a common relational representation for the overlap between the cases, such as re-representing both in terms like: RELEASE(AGENT, RELEASED_ENTITY). But consider
what would happen if a participant then compared one of those cases to a new situation, as in:

- As the zoo keeper was busy cleaning its habitat, the Burmese python was able to escape its open cage.
- Nicole finally got out of the bad relationship that had prevented her from pursuing her own interests.

Now the relational structure of the first sentence, established during the prior comparison, would be incompatible with that of the second, because the RELEASE relation is not identical with the ESCAPE relation. (Of course, this depends on some assumptions about participants' mental representations, but ones that are borne out by our data-see General Discussion.) At this point, it is possible that the participant might devote the additional processing effort required to change the representational structure yet again, in search of potential shared relations. However, in most real-world experiences-as in most experimental settings-we believe that individuals will tend to exert a more modest level of processing, in this case typically relying on the representation that has already been created. If so, they would determine that the two cases in the second comparison are simply not very similar to one another.

In these two experiments, we examine whether similarity ratings are reliably higher when one of the compared cases has recently participated in another comparison that involves the same shared relational structure, relative to recent comparisons involving a different structure. For control purposes, the relevant test comparisons were always the same across participants-only the preceding comparison varied between conditions. In Experiment 1, we establish this basic effect, while Experiment 2 both replicates this finding and rules out alternative explanations based on relational priming.

## Experiment 1

Participants Thirty participants were recruited through Amazon's Mechanical Turk in return for $\$ 1.00$ payment.

Materials and Design The study was computer-based, and was administered online. After reading the instructions, each participant read 18 sentence pairs, presented on-screen one at a time. Participants were asked to rate the similarity of each pair by clicking on a horizontal 15 -point scale. Above the scale was a prompt, "How similar are these situations?", and the scale's endpoints were labeled Very dissimilar and Very similar. The entire task took approximately five minutes to complete.

We developed six Standard sentences, each of which could reasonably be construed according to two different relational structures, which we will refer to as Structure $A$ and Structure B (see Figure 1 for a visual depiction of the overall design). For example, the sentence about the python and the zoo keeper described in the Introduction could be represented as an example of "being able to escape from a confining environment," or as an example of "inadvertently releasing something dangerous." Each of these Standards was involved in two consecutive comparisons.


Figure 1: Design of Base-Test comparison pairs in Experiment 1.

We will refer to the first of these comparisons for each Standard as the Base comparison. The other sentence in the Base comparison varied between participants: approximately half of the participants compared the Standard to an unambiguous example of Structure A, while the remainder compared the Standard to an example of Structure B. The subsequent trial was the Test comparison. The sentences in this comparison were the same for all participants: the Standard was compared to a new example of Structure A. Ratings from the six Test comparisons (one for each Standard) provided the relevant measurement in our experiment. While the sentences in the Test comparisons were identical for all participants, they were classified as either Same trials or Different trials, according to whether the preceding Base comparison involved the same relational structure (Structure A) or a different structure (Structure B). Our primary question within this experiment is whether similarity ratings on Same trials would be significantly greater than those on Different trials, consistent with re-representation of the Standard.

In sum, the relevant stimuli included six relevant items sets, each containing one Standard sentence, two analogous examples of Structure A (one for potential use in the Base comparison and one for the second Test comparison), and one example of Structure B (for potential use in the Base comparison).


Figure 2: Results from Experiment 1

Each participant completed three Same trials and three Different trials. The condition (Same vs. Different) of each of the six Test trials was assigned randomly for each participant. The presentation order of the six comparison pairs also varied randomly between participants. Additionally, participants completed six filler comparisons-one at the beginning of the task, and one following each of the Test trials except the lastfor a total of 18 comparisons.

Results and Discussion A paired-samples t-test revealed a significant difference between conditions (see Figure 2; $t(29)=$ 4.99, $p<.001, d=1.17$ ), with Same trials ( $M=9.59, S D=2.41$ ) receiving considerably higher similarity ratings than Different trials ( $M=6.40, S D=2.99$ ). To ensure that these effects were not driven by a small subset of the materials, we also analyzed the data across items. Similarity ratings for all of the six items were higher during Same trials than Different trials, and there was a significant difference between the ratings at the item level $t(5)=$ 4.13, $p=.009, d=2.23$ ). Because the sentences in these Test trials were identical for all participants, these systematic differences must reflect the influence of the Base comparisons that preceded them, the sole variation between conditions.

The observed pattern is consistent with a process of rerepresentation. According to this explanation, the structure and content of the mental representation of the Standard sentence was altered during the Base comparison in order to maximize its similarity to its paired sentence. When the resulting representational structure was also a good match for the sentence in the subsequent Test trial, a straightforward mapping would have been possible and comparison would proceed smoothly. However, when the initial re-representation left the Standard with a structure that mismatched the paired sentence in the Test trial, the perceived similarity between the sentences would be poor unless the participant devoted the additional processing effort necessary to alter the Standard yet again.

## Experiment 2

Changes in perceived similarity represent a straightforward, low-level means of assessing participants' mental representations. However, while our data is consistent with the proposed explanation of re-representation, there is a salient alternative explanation that must also be considered. Our
approach suggests that the changes in similarity ratings were the result of persisting changes in the mental representations of the Standard sentences themselves. However, our data could also be explained by the activation of more abstract representations that are external to the individual sentences, through a process of relational priming.

Consider the example stimuli discussed in the Introduction, with Nicole escaping from her bad relationship, and the python escaping from its enclosure at the zoo. In the course of comparing these situations, participants may be activating an abstract representation of the relation ESCAPED_FROM. In fact, there is considerable empirical evidence that comparison can promote the generation or activation of abstract knowledge structures (e.g., Catrambone \& Holyoak, 1989; Gentner, Loewenstein \& Thompson, 2003; Gick \& Holyoak, 1983). If so, that representation could presumably still be active and influential during the subsequent Test comparison. As such, it would be in a position to alter the perceived similarity in that trial in at least two different ways. First, it could serve to influence and bias the interpretations of each of the comparison sentences independently. For example, prior research has shown that when individuals are primed with traits such as brave and adventurous, they tend to develop more positive impressions of a character who attempts dangerous, exciting tasks, relative to participants who were primed with traits such as reckless and foolish (Higgins et. al, 1977). In other words, the mental availability of a concept appears to bias people's interpretation of novel, ambiguous stimuli. In our example case, priming of a general relational concept such as escape could be biasing participants to interpret subsequent sentences as examples of that schema.

At the same time, priming of the ESCAPED_FROM relation could be influencing participants' assessments of the relationship between the sentences in the Test trial. A large body of literature has shown that individuals give higher ratings on a variety of measures to a stimulus when it is processed more fluently (e.g., Mandler, Nakamura \& Van Zandt, 1987; Whittlesea, 1993). This fluency may be the result of a variety of factors, including physical properties of the stimulus itself, but it is most commonly associated with prior exposure to a stimulus. In our study, all participants might have been able to recognize the relevant relationship between the sentences in the Test trials. However, if that particular relationship was already primed and strongly available, the commonalities might have become easier to process, and this ease of processing may in turn have led to a heightened sense of relevance or meaning. If so, prior research suggests that this sense of fluency (or disfluency, in the Different trials) could have influenced participants' similarity ratings in a pattern similar to that observed in our data. Some prior research is consistent with the idea that relational priming may influence comprehension and interpretation (e.g., Estes, 2003; Estes \& Jones, 2006).

In some ways, the distinction between an explanation based on re-representation and an explanation based on relational priming is subtle. At a theoretical level, however, this distinction is crucial. As discussed, structure mapping theory is a highly influential model that has had a great deal of explanatory success. However, its viability depends upon its ability to accommodate
matches between relationships that are similar but not identicaland this ability depends upon the process of re-representation: structural and/or conceptual changes in one or both of the mental representations. An explanation based on relational priming would not require any changes in the representation of the Standard itself, and therefore would provide no evidence that rerepresentation was occurring. In order to draw any meaningful support for our hypotheses, we therefore need to either rule out a priming explanation, or to demonstrate that re-representation is exerting an influence over and above that of simple priming. In Experiment 2, we add a control condition in order to assess the independent contributions of re-representation.

Participants Sixty participants were recruited through Amazon's Mechanical Turk in return for $\$ 1.00$ payment.

Materials and Design Experiment 2 included two betweenparticipants conditions. In the Repeated Standard condition ( $n=$ 30), the materials and procedure were identical to those used in Experiment 1. The Relational Priming condition ( $n=30$ ), which served as a control, was identical to the Repeated Standard condition with the exception that each of the Standard sentences was compared only once, during the Test trial, rather than in two consecutive trials (the Base and the Test). During the Base trials, participants in this condition were presented with two sentences that were each an example of one of the two relevant relational structures. That is, each participant compared either two examples of Structure A or two examples of Structure B.

For instance, two consecutive trials in the Relational Priming condition might include the following two comparisons:

- When the instructor turned around to write something on the board, Eric slipped out of the boring lecture.
- The rabbit had been cornered by a fox for several minutes, but finally lunged through the weeds and got away safely.
- As the zoo keeper was busy cleaning its habitat, the Burmese python was able to escape its open cage.
- Nicole finally got out of the bad relationship that had prevented her from pursuing her own interests.
Unlike Experiment 1, and unlike the Repeated Standard condition in this experiment, participants in this condition did not see the Standard sentence (about the python at the zoo) until the Test comparison. There was therefore no opportunity for prior re-representation of that situation. There was, however, still an opportunity for relational priming. The two sentences in the Base trial are each unambiguous examples of escape, and that is the most salient commonality between them. According to a priming explanation for our data, that relation would have been highlighted and made more accessible during the Base trial, and would therefore be in a position to influence subsequent comparisons.

If the differences in Experiment 1 were solely the result of relational priming, we would expect no differences between the Repeated Standard and Relational Priming conditions, because the priming effects should be equivalent. If, on the other hand, re-
representation is influencing perceived similarity, effects should be greater for participants in the Repeated Standard condition.

Results and Discussion There were three primary goals in Experiment 2. First, it gave us an opportunity to attempt a replication of the findings from Experiment 1, which is important given the novelty of those results. Second, it allowed us to assess whether relational priming may exert an influence in simple consecutive similarity judgments. This is an interesting question in its own right, as we will explore in the General Discussion. Finally, and most importantly, this experiment allowed us to compare the two between-participants conditions, one of which provided the opportunity for re-representation and one of which did not. Because the two conditions should have been equivalent in terms of potential relational priming, any observed advantages for the Repeated Standard condition would provide strong evidence that re-representation had taken place.

We again found an overall advantage for ratings on the Same trials $(M=8.81, S D=2.49)$ relative to the Different trials ( $M=$ $\left.6.47, S D=2.36 ; F(1,58)=46.17, p<.001, \eta_{p}^{2}=.44\right)$. However, because this includes both of the between-participants conditions, we performed a separate analysis of the Repeated Standard condition (which was identical to Experiment 1) to determine whether the basic pattern from the first study had been replicated. This revealed a pattern of results very similar to Experiment 1 (see Figure 3). Similarity ratings for the Same trials ( $M=8.89$, $S D=2.48$ ) were significantly higher than those for Different trials $(M=5.66, S D=2.46 ; t(29)=5.72, p<.001, d=1.31)$, replicating our initial finding. This pattern also held in a separate analysis across items $(t(5)=7.16, p=.001, d=1.94)$.

Next, we examined whether relational priming might have had an influence on participants' ratings. In a separate analysis of the Relational Priming condition, Same trials ( $M=8.63, S D=2.53$ ) received higher similarity ratings on average than Different trials ( $M=7.28, S D=1.98$ ), across participants $(t(29)=3.67, p=.001$, $d=0.59$ ) and items $(t(5)=3.00, p=.03, d=1.00)$, suggesting that relational priming was indeed having a measurable effect.


Figure 3: Results from Experiment 2

Most important for our theoretical interests, a 2 (Repeated Standard vs. Relational Priming) $\times 2$ (Same vs. Different trials) ANOVA showed a significant interaction between participant condition and item condition $\left(F(1,58)=8.16, p=.006, \eta_{p}^{2}=\right.$ .12). This interaction reflected the fact that the advantage of Same over Different trials was significantly greater in the Repeated Standard condition (mean difference $=3.81, S D=$ 3.84) than the Relational Priming condition (mean difference $=$ $1.36, S D=2.03$ ). This advantage was seen for all six items sets individually, and confirmed by an interaction in an analysis across items $\left(F(1,10)=4.08, p=.021, \eta_{p}^{2}=.43\right)$.
(To ensure that the materials were equally apt in both conditions, we confirmed that ratings for the Base comparisons did not differ between conditions $(t(59)=1.22, p=.23)$. In fact, there was a small numerical advantage for the Relational Priming condition: $M=10.23, S D=1.52$, vs. $M=9.74, S D=1.55$.)

## General Discussion

The results of these two studies are informative in several ways. First and foremost, they provide important evidence for a process of re-representation during comparison. As the dominant model of analogy and structured comparison, structure mapping theory has been used to explore and explain a wide variety of cognitive phenomena. However, its ability to scale up to even very basic real-world situations depends on its ability to flexibly find connections between related but non-identical structures. Re-representation has historically been cited as the underlying explanation for this ability. Despite its theoretical importance, however, direct evidence for re-representation has remained scarce. In our studies, comparing a standard case to one situation systematically changed its perceived similarity to new cases. The pattern of these changes indicates that the representational structure and content of the original standard had been altered in a way that made it more compatible with its compared situation-in other words, it had been re-represented. This effect held even when controlling for potential relational priming effects. By adding support to this critical but under-explored area, our results are able to further bolster the viability of structure mapping theory in general.

Although it was not our primary research focus, another informative contribution of these studies is that Experiment 2 demonstrates a novel form of relational priming. The idea of relational priming - that processing a particular semantic relation in one situation may make it easier to process in the futureseems reasonable, and perhaps even obvious given what we know about priming in other contexts. However, finding evidence to support this phenomenon has not always been straightforward. In one of the earliest experimental attempts, Spellman and colleagues (2001) found no indication of relational priming between word pairs in a lexical decision task, even when participants were explicitly told to focus on the relationships between the presented words. Only when individuals were told to notice that consecutive trials might involve the same relationship was a modest effect observed. Subsequent research has been more successful in finding examples of relational priming, through the use of more tightly controlled stimuli and by having participants engage in tasks that more naturally
involved the activation of relations, such as the interpretation of two-word phrases (e.g., Bendig \& Holyoak, 2009; Estes, 2003; Estes \& Jones, 2006; Gagné, 2001). However, our control condition in Experiment 2 is the first example to our knowledge that demonstrates relational priming through changes in perceived similarity, and the first to find large effects with such naturalistic stimuli.

Finally, our experiments introduce a novel method for assessing mental representation more generally. As with most studies of this type, the stimuli for the present experiments were coded by the authors largely as a function of our intuitions about the semantics of the situations involved. However, as history has repeatedly shown, researcher intuitions can often be wrong. Furthermore, intuitions can vary markedly between individuals. Consider the following two situations: (1) X was victorious over Y, and (2) Y was defeated by X. In our stimuli, we assumed that these reflected two distinct representational structures. Another researcher, however, might reasonably argue that they are simply two different ways of expressing the same underlying proposition. Our experimental method provides a direct way to address this issue. In our studies, comparing a sentence to a clear example of the DEFEAT structure made it subsequently seem significantly less similar to an unambiguous example of VICTORY. In this case, the intuition that these reflect distinct mental representations appears to have been correct, although until that point it was an open empirical question. Examination of the similarity changes resulting from re-representation offers us an intriguing tool for exploring and answering questions about the semantics of mental relations, and therefore provides a potential window into a variety of important mental processes.

## References

Bendig, B. W., \& Holyoak, K. J. (2009). Relational priming of analogical reasoning. In New Frontiers in analogy research: Proceedings of the second international conference on analogy (pp. 30-36).
Boroditsky, L. (2007). Comparison and the development of knowledge. Cognition, 102(1), 118-128.
Catrambone, R., \& Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(6), 1147.
Estes, Z. (2003). Attributive and relational processes in nominal combination. Journal of Memory and Language, 48(2), 304319.

Estes, Z., \& Jones, L. L. (2006). Priming via relational similarity: A copper horse is faster when seen through a glass eye. Journal of Memory and Language, 55(1), 89-101.
Falkenhainer, B., Forbus, K. D., \& Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial intelligence, 41(1), 1-63.
Gagné, C. L. (2001). Relation and lexical priming during the interpretation of noun-noun combinations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(1), 236.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive science, 7(2), 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou \& A. Ortony (Eds.), Similarity and analogical reasoning (pp. 199-241). London: Cambridge University Press. (Reprinted in Knowledge acquisition and learning, 1993, 673-694).
Gentner, D., \& Kurtz, K. J. (2006). Relations, objects, and the composition of analogies. Cognitive Science, 30(4), 609-642.
Gentner, D., Loewenstein, J., \& Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95(2), 393.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive psychology, 15(1), 1-38.
Goldstone, R. L., Lippa, Y., \& Shiffrin, R. M. (2001). Altering object representations through category learning. Cognition, 78(1), 27-43.
Higgins, E. T., Rholes, W. S., \& Jones, C. R. (1977). Category accessibility and impression formation. Journal of Experimental Social Psychology, 13(2), 141-154.
Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. Behavior Research Methods, Instruments, \& Computers, 16(2), 96-101.
Holyoak, K. J., Novick, L. R., \& Melz, E. R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. J. Holyoak \& J. A. Barnden (Eds.), Advances in connectionist and neural computation theory, Vol. 2: Analogical connections. Norwood, N.J.: Ablex.
Holyoak, K. J., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive science, 13(3), 295-355.
Kotovsky, L., \& Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.
Kurtz, K. J. (2005). Re-representation in comparison: Building an empirical case. Journal of Experimental \& Theoretical Artificial Intelligence, 17(4), 447-459.
Mandler, G., Nakamura, Y., \& Van Zandt, B. J. (1987). Nonspecific effects of exposure on stimuli that cannot be recognized. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(4), 646.
Markman, A. B., \& Gentner, D. (2001). Thinking. Annual Review of Psychology, 52(1), 223-247.
Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., \& Shafir, E. (1990). Category-based induction. Psychological review, 97(2), 185.
Pavlov, I.P. (1927). Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. London: Oxford University Press.
Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(4), 629.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317--1323.
Smith, E. E., \& Medin, D. L. (1981). Categories and concepts (p. 89). Cambridge, MA: Harvard University Press.

Whittlesea, B. W. (1993). Illusions of familiarity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(6), 1235.

# Comprehenders Rationally Adapt Semantic Predictions to the Statistics of the Local Environment: a Bayesian Model of Trial-by-Trial N400 Amplitudes 

Nathaniel Delaney-Busch (ndelan02@tufts.edu) ${ }^{1}$, Emily Morgan (emorga08@tufts.edu) ${ }^{1}$, Ellen Lau (ellenlau@umd.edu) ${ }^{2}$, Gina Kuperberg (kuperber@nmr.mgh.harvard.edu), ${ }^{1,3}$<br>${ }^{1}$ Department of Psychology, Tufts University<br>${ }^{2}$ Department of Linguistics, University of Maryland<br>${ }^{3}$ Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital


#### Abstract

When semantic information is activated by a context prior to new bottom-up input (i.e. when a word is predicted), semantic processing of that incoming word is typically facilitated, attenuating the amplitude of the N400 event related potential (ERP) - a direct neural measure of semantic processing. This N400 modulation is observed even when the context is just a single semantically related "prime" word. This so-called "N400 semantic priming effect" is sensitive to the probability of seeing a related prime-target pair within experimental blocks, suggesting that participants may be adapting the strength of their predictions to the predictive validity of their broader experimental environment. We formalize this adaptation using an optimal Bayesian learner, and link this model to N400 amplitudes using an information-theoretic measure, surprisal. We found that this model could account for the N400 amplitudes evoked by words (whether related or unrelated) as adaptation unfolds across individual trials. These findings suggest that comprehenders may rationally adapt their semantic predictions to the statistical structure of their broader environment, with implications for the functional significance of the N 400 component and the predictive nature of language processing.


Keywords: language; prediction; rational adaptation; semantic priming; EEG/ERP; word processing; information theory; Bayesian modeling; surprisal

## Introduction

How a word is processed fundamentally depends on the context. Predictable words are processed more quickly than unpredictable words (Fischler \& Bloom, 1979), with shorter fixations (and more frequent skips) during reading (see Staub, 2015 for review). A similar facilitation pattern is found on the N400 component (Kutas \& Hillyard, 1984), an event-related potential (ERP) that reflects semantic processing (Kutas \& Federmeier, 2011). The degree to which any particular word is facilitated is proportional to the probability of encountering that word given the context (DeLong, Urbach, \& Kutas, 2005; Smith \& Levy, 2013).

Semantic facilitation is observed even when the preceding context is only a single word. For example, the processing of a "target" word is facilitated when it is preceded by a semantically related (versus an unrelated) "prime" word: the so-called "semantic priming effect" (Neely, 1991). Semantic priming is also apparent on the N400 component, with more predictable words eliciting a
smaller (less negative) N400 amplitude than less predictable words (Bentin, McCarthy, \& Wood, 1985).

Importantly, the strength of the behavioral semantic priming effect is sensitive not only to the degree to which the prime and target are semantically related, but to the probability of receiving a related trial in the first place (Brown, Hagoort, \& Chwilla, 2000; Grossi, 2006). This has also been found with ERPs: experimental blocks with a higher proportion of related trials elicit a larger N400 semantic priming effect (Lau, Gramfort, Hamalainen, \& Kuperberg, 2013; Lau, Holcomb, \& Kuperberg, 2013). The semantic priming effect is likely sensitive to predictive validity because participants implicitly track and adapt to changes in the statistical contingencies over time.

Here we utilized data from Lau and colleagues (Lau, Holcomb, et al., 2013) to build and test a quantitative hypothesis of what drives this effect of predictive validity on the N400 semantic priming effect. Specifically, we asked whether the larger N400 priming effect in the high predictive-validity block could have been achieved through a "rational" probabilistic model of trial by trial adaptation. Although there are an infinite number of ways to build such a model, there are some particular theoretical constraints we can start from. Current evidence suggests that (a) prediction in language processing is probabilistic in nature, (b) predictions incrementally adapt to new information (where adaptation should be rapid when the environment changes), and (c) the brain calculates something like prediction error.

## Probabilistic Prediction in Language Processing and Semantic Priming

The role of prediction in language has long been debated, with differing definitions of what a "prediction" actually entails (see Kuperberg \& Jaeger, 2016 for discussion). Here, we view probabilistic prediction as a central feature of language comprehension (DeLong et al., 2005; Federmeier, 2007; Smith \& Levy, 2013), which does not necessarily need to be strategic or even conscious in nature. For any given context, there exists a probability distribution over the words that could be encountered next. A "prediction" is simply the presence of this probability distribution. While there is evidence for probabilistic prediction at multiple different levels of representation (Kuperberg \& Jaeger, 2016), here we focus only on prediction at the lexical level.

Note that, defined in this way, prediction exists even in the absence of a local context. Consider an experiment where words are being serially presented to participants at random. The "prediction" that participants make in such an experiment could be expressed as a probability distribution over words given an average context. This is functionally identical to word frequency, where high frequency words are more probable given a random/average context and low frequency words are less probable given a random/average context (Norris, 2006).

Given these assumptions about the nature of linguistic predictions, we can view the semantic priming effect as a type of probabilistic prediction. If a participant knows that a prime informs the target, they will implicitly generate a probability distribution over possible targets given that prime. Target processing is facilitated proportional to its probability. Though these prime-target transition probabilities are not easy to estimate from corpus studies (people don't often write in prime-target pairs), it can be estimated using production tasks like word association.

## A Rational Model of Adaptation

Probabilistic prediction is only beneficial if it actually approximates the statistical structure of the environment. Bad predictions aren't helping anybody. A number of recent language studies suggest that people rapidly adapt local models based on changes in their environment (Kleinschmidt \& Jaeger, 2015). For example, an environment with a high proportion of typically dispreferred syntactic parses can attenuate or even reverse the so-called "garden-path" effect in ambiguous sentences (Fine, Jaeger, Farmer, \& Qian, 2013). At the phonemic level, participants change their perception of ambiguous phonemes if one of the two competing options was locally repeated (Kleinschmidt \& Jaeger, 2016). And across longer periods, people show signs of adaptation to foreign accents that can generalize between different speakers with that accent (Bradlow \& Bent, 2008).

We can view the predictive validity manipulation in Lau et al (2013) through a similar lens. In experimental contexts with a higher proportion of semantically related word-pairs, participants will rely relatively more on the prime (versus relying more on a random/average context) to inform their predictions of the target. This means that the probability mass assigned to any particular target word depends not only on the associative strength of its prime, but also the likelihood that the prime provides information about the target in the first place (i.e. the predictive validity effect). When the proportion of semantically related and unrelated trials within a block changes, people adapt.

Although there are many ways to quantify adaptive learning, one attractive theoretically motivated implementation is Bayesian updating. This assumes that adaption is "rational" in nature (Anderson, 1990). Here, initial belief about the probability of obtaining a particular
type of trial (i.e. a related versus unrelated prime-target pair) is denoted by the prior probability $\mathrm{p}(\mathrm{h})$. Upon receiving a trial, this prior belief is updated using Bayes Law. This "posterior" is then used as the prior belief for the next trial.

Applying this rational, Bayesian framework to the predictive validity effect (on semantic priming) has a number of advantages. It would allow for prior beliefs from an initial lower predictive validity block to influence expectations for a subsequent higher predictive validity block in a principled way (i.e. the prior). It would adapt incrementally across trials. It would adapt more quickly near the change point, when evidence is low. And beliefs would slowly asymptote to the known true probability of receiving a related trial.

In the present investigation, we will use this type of Bayesian framework as the starting point for explaining how the brain adapts to changes in the statistical contingencies of incoming language input. We will refer to this as a Rational Adapter model.

## The N400 Measures Information Content

So our model should be probabilistic, and it should adapt, but a third component of this model is required before it can be tested: a linking function to actual brain activity. It is not necessarily the case that N400 amplitudes - the brain activity we are attempting to model - need be linearly dependent on the probability of a word given a context. Here, we argue that the N400 is best thought of not as a measure of probability, but as a measure of information (see Rabovsky \& McRae, 2014 for discussion).

In information theory (Shannon \& Weaver, 1949), the amount of information conveyed by an event is "whatever was not known ahead of time". An input that was perfectly predicted does not convey any new information. In contrast, messages that are not very predictable convey a lot of information. The amount of information (in units of "bits") that was not predicted ahead of time is called "surprisal", quantified as $-\log _{2}[p($ word $\mid$ context $)]$. One bit is the amount of information provided by flipping a fair coin once. A halving of the probability (e.g. conveying a sequence of two coin flips instead of one) corresponds with a 1 bit increase in surprisal.

This simple transformation has proved tremendously powerful in explaining language processing data (Hale, 2001; Levy, 2008). Smith and Levy (2013) provide empirical evidence that reading times relate to word probability logarithmically (e.g. as with the surprisal transformation) across six orders of magnitude. More recently, Frank and colleagues (Frank, Otten, Galli, \& Vigliocco, 2015) discuss ERP evidence that the N400 component is sensitive to word surprisal. Given this evidence, for the present investigation, information-theoretic surprisal will be our linking function between the Rational Adapter model and N400 amplitude, rather than probability.

In the present investigation, we aim to build and test a Rational Adapter model of semantic priming against neural N400 component data. Specifically, we use data from Lau and colleagues (2013). Here, the N400 semantic priming effect was measured first in block 1 where $10 \%$ of the trials were related, and then in block 2 where $50 \%$ of the trials were related. We use the block 1 only to inform the prior, which is then fed into the Rational Adapter model to predict N400 amplitudes in block 2, as participants' beliefs about the predictive validity of the prime adapt.

We hypothesize that this Rational Adapter surprisal model will better account for the N400 data than a nonlearning model of N400 data. Specifically, we hypothesize that the size of the N400 effect (the difference in amplitude of the N400 evoked by related and unrelated target words ) will increase rapidly near the beginning of block 2 , as the Rational Adapter shifts from a " $10 \%$ related" prior towards a " $50 \%$ related" asymptote.

## Methods, Modeling and Results

## ERP Data collection

We used data from Lau and colleagues (Lau, Holcomb, et al., 2013). Briefly, 32 right-handed participants (13 men) between age 19-24 were shown sequential prime-target pairs as event-related potentials (ERPs) were recorded and timelocked to the onset of the target. Participants were asked to perform a semantic monitoring task that was not directly related to the experimental manipulation. All participants saw an initial 400 trials (block 1) where $10 \%$ of the stimuli were related (e.g. "ladder... climb"), followed by 400 trials where $50 \%$ of the stimuli were related (block 2 ). The blocks were separated by a short break, but participants were not explicitly told about any changes in the experiment.

80 of the trials in block 2 ( 40 related, 40 unrelated) were critical prime-target pairs that were matched and counterbalanced across participants, alongside 320 fillers. Primes were presented with an SOA of 600 ms and targets had a duration of 900 ms . The N400 component was averaged across a time window of $300-500 \mathrm{~ms}$ over the average of three centro-parietal channels. Extreme outliers in N400 measures were removed (4 standard deviations or more from the mean).

Visualization Lau et al originally reported that the N400 semantic priming effect was larger in block 2 than block 1. For the present investigation, we plotted how the N400 amplitudes of these critical trials changed over the course of block 2, as shown in Figure 1. This was estimated using a loess local regression over N400 amplitudes for related and unrelated words across the ordinal position of critical items in the experiment (this local regression was necessary because not every participant saw critical targets in the same place). As can be seen here, N400 amplitudes for related and unrelated words are initially similar, but then diverge as participants are exposed to more and more of the block.


Figure 1: Block 2 N400 amplitudes over trials.

## The Rational Adapter Word Surprisal Model

Our Rational Adapter model consists of three primary components: a) a Bayesian belief about the probability of receiving a related vs. an unrelated prime-target pair at any given point, b) a mixture of $p$ (word|prime) and $p$ (word|average context) given these beliefs about the trial types, and c) a conversion from these probabilistic predictions to word surprisal (as a linking function to the N400 component). The whole model takes the form:

$$
\begin{gathered}
\text { Word surprisal }=-\log _{2}\left[\lambda * p(\text { word } \mid \text { prime })+(1-\lambda)^{*}\right. \\
p(\text { word } \mid \text { average context })]
\end{gathered}
$$

where $\lambda$ is a point estimate of the probability with which a rational adapter expects a related trial at that point in time.

We use a beta-binomial model to estimate a participant's belief about the probability of seeing related versus unrelated trials. To set a prior on the beta distribution, we assume that participants enter block 2 assuming the proportion will be the same as block 1: a $10 \%$ chance of receiving a related trial. This prior is expressed using a 1:9 ratio of related:unrelated pseudocounts. Though participants see 400 trials in block 1, participants will likely discount their previous experience somewhat (reflecting some uncertainty). As a best-guess approximation, we set the prior going into block 2 at $\operatorname{Beta}(5,45)$, i.e. pseudocounts equivalent to having seen 5 related and 45 unrelated trials. In other words, participants were assumed to put more weight on their experience with the new block (vs their prior given the old block) after about 50 trials.

After each new trial, this Beta distribution is updated by adding the observed trial counts to the prior pseudocounts. For example, after 5 related and 5 unrelated trials in block 2, a participant's beliefs would be modeled as $\operatorname{Beta}(10,50)$. We took the mean of this beta distribution just before each critical trial to reflect the point probability $\lambda$ with which that participant expects a related trial for that event.

This probability, $\lambda$, then provides a weighting term for a mixture model between the two ways that participants might generate more specific predictions about the upcoming target word at any given trial. Given a related trial, we
model these within-trial predictions as p (word|prime), estimated using "forward association strength" (FAS) from the Florida Word Association Norms (Nelson, McEvoy, \& Schreiber, 2004), and then we weight this probability by $\lambda$. Given an unrelated trial, we model these within-trial predictions as p (word|average context), as estimated by word frequency from the SUBTLEX corpus (Brysbaert \& New, 2009), and then we weight this probability by $1-\lambda$.

The mixture of these two terms yields a "word probability", given the prime and beliefs about whether or not it will inform the target at any point in the experiment.

Finally, this word probability is transformed into "word surprisal", or the amount of information that was not predicted ahead of time (in bits), given by $-\log _{2}[\mathrm{p}$ (word probability)].

## Word Surprisal and N400 Amplitudes

To numerically test whether our estimate of surprisal explains variance in the N400 amplitudes evoked by each target word, we conducted a linear mixed-effects regression using the lme4 package in R, with word surprisal as a predictor and centro-parietal N400 amplitude for each trial in block 2 as an outcome. Word surprisal was standardized. The maximal random effects structure across (crossed) subjects and items was used (Barr, Levy, Scheepers, \& Tily, 2013).

Results We found that word surprisal significantly accounted for variance in N400 amplitudes $(\beta=-1.14, t=-$ $5.24, \mathrm{p}<0.001$ ). As word surprisal increased, N400 amplitudes tended to be more negative (i.e. larger).

## Word Surprisal Explains Trial-by-Trial Variance

There is an important caveat to this "rational adapter" word surprisal effect, however: by definition, unrelated words tend to have high surprisal, while related words tend to have low surprisal. As such, the word surprisal effect in block 2 could potentially be attributable to the categorical "Relatedness" effect already reported in the initial study.

To address this possibility, we ran a second linear mixedeffects regression that included both categorical Relatedness and word surprisal as predictors. This would show whether our rational adapter estimate of word surprisal could account for variance in N400 amplitudes above and beyond what could already be explained by the main effect of related vs unrelated trials. Again, the maximal random effects structure for word surprisal was used.

Results We found that word surprisal significantly accounted for variance in N400 amplitudes ( $\beta=-2.21, \mathrm{t}=-$ $2.76, \mathrm{p}=0.006$ ) above and beyond the main effect of Relatedness. This indicates that the surprisal difference between related and unrelated words was not sufficient to account for the way that word surprisal related to the N400 in the first model.

We caution that given the multicollinearity between word surprisal and the relatedness effect (the primary motivation
for running this test in the first place), this $\beta$ estimate is likely inflated. We limit our conclusions to the explanatory power here, not the regression coefficient.

## The "Rational Adapter" Word Surprisal Model Outperforms its Constituent Elements Alone

Another potential concern is that the model we used to estimate word surprisal simply includes more information about the trials. Namely, the word frequency and FAS of each trial are inputs to the Rational Adapter model calculations. These could have explained items-level variance in N400 amplitudes without resorting to adaptation, given that the N400 component is already known to be sensitive to both these factors. In short, perhaps the explanatory power of our rational adapter model is primarily due to the inclusion of trial-specific frequency and FAS information, rather than prediction and adaptation.

To address this possibility, we ran a third linear mixedeffects regression that includes not only word surprisal as a predictor, but also Frequency, FAS, and Relatedness predictors for each trial. This tests whether the particular arrangement of inputs into the "rational adapter" word surprisal model explains variance in N400 amplitudes marginal to the stationary main effect of Relatedness and to its constituent items-level elements. Again, the maximal random effects structure for word surprisal was used (across both items and subjects), and all continuous predictors were standardized.

Results We found that word surprisal significantly accounted for variance in N400 amplitudes $(\beta=-2.30, t=-$ 2.11, $p=0.036$ ) above and beyond Frequency, FAS, and Relatedness. This indicates that the particular way itemslevel features were combined into the our model is an important source of explanatory power, and that the increased fit is not simply due to the fact that our model included additional information about items-level features.

## Finding the Optimal Prior

Our model assumed that the rational adapter should approach $\lambda=0.1$ (the actual block 1 proportion) as they go through the first 400 trials of block 1, regardless of what their expectations were coming into the experiment. Given that our model is explicitly a rational one, we kept constant this 1:9 related:unrelated ratio for the prior for block 2. However, that still leaves the prior strength (i.e. number of pseudocounts) as an assumption that can be explored. For hypothesis testing above, we assumed that participants entered block 2 with a $\operatorname{Beta}(4,45)$ prior, i.e. that participants believed it would have the same $10 \%$ relatedness proportion as block 1 with a weight of 50 pseudotrials. This 50 pseudocount prior weighting, however, was essentially guesswork (we didn't want to bias our hypothesis tests by interrogating many models and selecting the best one). Here, we sought to ensure that our results were not idiosyncratically dependent on having made a "lucky" guess.


Figure 2: Deriving the optimal prior strength.
A low pseudocount prior like $(1,9)$ would cause rational adaptation to proceed very quickly, while a high prior pseudocount like $(40,360)$ would cause rational adaptation to proceed much more slowly, as participants enter block 2 with much more certainty about the environmental statistics. The pseudocount (and thus speed of adaptation) that best explains variance in N400 amplitudes is thus an empirical question: what is the optimal prior strength?

To find the optimal prior, we calculated the word surprisal (for all trials for all participants, as above) for every integer "prior strength" from 1 to 800 pseudocounts. We then ran a separate linear mixed-effects regression model for each prior strength with word surprisal and Relatedness as predictors and a maximal random effects structure. After fitting these 800 regression models, we extracted the loglikelihood of each.

Results These data are shown in Figure 2. The single maximum log-likelihood was obtained with a $\operatorname{Beta}(7.7$, 69.3) prior, a "prior strength" of 77 pseudocounts. However, all pseudocounts between 70 and 85 yielded similar model fits, and performance degrades smoothly on either side.

This indicates that, on average, participants began giving more weight to the new block's data than the previous block's data 70-85 trials into block 2. A rational adapter with a very weak prior (below $\sim 50$ pseudocounts) does not account for N 400 data well because it adapts too quickly. Similarly, a rational adapter entering block 2 with a very strong prior (above $\sim 200$ pseudocounts) also does not account for the N400 data well because it adapts too slowly.

We note that some models had poor fit because they did not converge. None were within the 70-85 range capturing the maximum.

## Discussion

Previously, Lau and colleagues (2013) found that the N400 semantic priming effect shows evidence that adaptation occurred when the predictive validity of the local context changed. In the present investigation, we explored the nature of the adaptation process as it unfolded. Figure 1 shows the trial-by-trial nature of this adaptation over ordinal position in block 2. Our Rational Adapter model provides a
theoretically-grounded quantitative account of how that adaptation may have occurred on an incremental trial-bytrial basis. It was built using three foundational considerations: that contexts can probabilistically inform lexico-semantic expectations for upcoming stimuli, that these expectations adapt rationally (in an optimal Bayesian manner), and that the N400 component is sensitive to units of information rather than units of probability (after the present analyses, we tested this assumption and found that word surprisal significantly accounted for variance in N400 amplitudes $[\mathrm{t}=-2.57, \mathrm{p}=0.011]$ above and beyond word probability and the categorical Relatedness effect.).

In a re-analysis of the original study, we provide empirical evidence that this model is consistent with how brain activity evoked by target words changed over the course of block 2 . We showed that it accounted for variance in N400 amplitudes above and beyond the stationary effect of related versus unrelated trials, suggesting that it was capturing trial-by-trial differences within block 2. Further, we showed that this particular formulation of the rational adapter model accounted for significant variance in N400 amplitudes above and beyond even its own constituent elements, suggesting that the additional explanatory power was not simply due to the inclusion of items-level information (our single trial approach to ERP analysis). These findings extend previous work on rational adaptation to demonstrate that it can account for changes in predictions during lexico-semantic processing.

In addition, we used the rational adapter model to derive the rate of adaptation that best accounted for the ERP data. Even though participants saw 400 trials in Block 1, we estimate that participants adapted as if they had only seen $70-85$ trials of bock 1 by the time they entered block 2 . Although participants were not informed of the changing environmental statistics (and the manipulation was not overtly task-relevant), we speculate that the conspicuous block boundary may have prompted participants to adapt at a faster rate. Additionally, there may be a decay or filtering that occurs for distant exposures, which dynamical models of prediction and adaptation may be able to account for.

While the present study included data from a semantic priming paradigm, we suggest that a similar pattern may hold in comparable experiments with more expansive contexts, like sentences or discourses, as the theoretical underpinnings are functionally the same. For example, in experimental contexts with a high proportion of highly constraining sentences, we might expect participants to learn to predict more strongly. Finally, these data have implications for the functional significance of the N400 component. The N400 is often discussed as being sensitive to probabilities. We suggest that its sensitivity to probabilistic measures like cloze probability, forward association, and even frequency may be best conceptualized it as reflecting units of information rather than probability alone (see also Frank et al., 2015; Rabovsky \& McRae, 2014; Smith \& Levy, 2013).

## Acknowledgments

This project was funded by NIMH-R01-MH071635 to GRK and the Sidney J. Baer Trust to GRK. Thank you to Florian Jaeger, Heather Urry, Eddie Wlotko, and Meredith Brown for input.

## References

Anderson, J. (1990). The Adaptive Character of Thought (Studies in Cognition): Psychology Press.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang, 68(3).
Bentin, S., McCarthy, G., \& Wood, C. C. (1985). Eventrelated potentials, lexical decision and semantic priming. Electroencephalography and clinical neurophysiology, 60(4), 343-355.
Bradlow, A. R., \& Bent, T. (2008). Perceptual adaptation to non-native speech. Cognition, 106(2), 707-729.
Brown, C. M., Hagoort, P., \& Chwilla, D. J. (2000). An event-related brain potential analysis of visual word priming effects. Brain and language, 72(2), 158-190.
Brysbaert, M., \& New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977990.

DeLong, K. A., Urbach, T. P., \& Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature neuroscience, 8(8), 1117-1121.
Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. Psychophysiology, 44(4), 491-505.
Fine, A. B., Jaeger, T. F., Farmer, T. A., \& Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. PLoS ONE, 8(10).
Fischler, I., \& Bloom, P. A. (1979). Automatic and Attentional Processes in the Effects of Sentence Contexts on Word Recognition. Journal of verbal learning and verbal behavior, 18(1), 1-20.
Frank, S. L., Otten, L. J., Galli, G., \& Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. Brain and language, 140, 1-11.
Grossi, G. (2006). Relatedness proportion effects on masked associative priming: an ERP study. Psychophysiology, 43(1), 21-30.
Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 159-166.
Kleinschmidt, D. F., \& Jaeger, F. T. (2015). Robust speech perception: Recognize the familiar, generalize to
the similar, and adapt to the novel. Psychological review, 122(2), 148.
Kleinschmidt, D. F., \& Jaeger, F. T. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? Psychonomic bulletin \& review, 23(3), 678-691.
Kuperberg, G. R., \& Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? Lang Cogn Neurosci, 31(1), 32-59.
Kutas, M., \& Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology, 62, 621-647.
Kutas, M., \& Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature, 307(5947), 161-163.
Lau, E. F., Gramfort, A., Hamalainen, M. S., \& Kuperberg, G. R. (2013). Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. $J$ Neurosci, 33(43), 17174-17181.
Lau, E. F., Holcomb, P. J., \& Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. $J$ Cogn Neurosci, 25(3), 484-502.
Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3), 1126-1177.
Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. Basic processes in reading: Visual word recognition, 11, 264-336.
Nelson, D. L., McEvoy, C. L., \& Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. Behavior research methods, instruments, \& computers : a journal of the Psychonomic Society, Inc, 36(3), 402-407.
Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. Psychological review, 113(2), 327-357.
Rabovsky, M., \& McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. Cognition, 132(1), 68-89.
Shannon, C. E., \& Weaver, W. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.
Smith, N. J., \& Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302-319.
Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. Language and linguistics compass, 9(8), 311-327.

# The Role of Imagination in Exemplar Generation: The Effects of Conflict and Explanation 

Tammy M. Dennis (tdennis@student.unimelb.edu.au)<br>Daniel R. Little (daniel.little@unimelb.edu .au)

Melbourne School of Psychological Sciences, The University of Melbourne, VIC 3010, Australia


#### Abstract

Structured imagination refers to reliance upon prior knowledge when generating novel examples of a provided category. Yet studies supporting this tenet use experimental designs where the stimuli themselves cue exemplars based on culturally relevant items. The present study combined exemplar generation with abstract stimuli as a means of attenuating instructional bias. Participants were shown a group of abstract shapes identified as a single category and instructed to generate another member of this category. We additionally examined whether the introduction of a cognitive conflict (by including an anomalous category member) and self-explanation during generation affected the level of imaginative responses. Contrary to expectations, the presentation of a conflicting category member did not result in more imaginative responses when compared to more homogenous stimuli sets. However, a significantly greater degree of imaginative responses was observed from participants who were required to explain their thinking prior to and whilst constructing their exemplars.


Keywords: imagination; exemplar generation; cognitive conflict; self-explanation; reflective abstraction.
[Ramanujan's equations] must be true because no one would have the imagination to invent them.

G. H. Hardy

## Introduction

How does imagination impact the generation of novel ideas? Theories of structured imagination assert that the generation of ideas, concepts, and objects depends on and is constrained by prior knowledge (Ward, Patterson \& Sifonis, 2004). Prior knowledge limits the set of features, dimensions, relations, functions, and so on under consideration. However, support for this position comes from experimental paradigms where instructions reference prior concept knowledge or examples that cue prior knowledge. Consequently, the influence of prior knowledge on the role of imagination in, for instance, an exemplar generation paradigm remains somewhat unclear.

The exemplar generation paradigm requires participants to generate new category members that could plausibly belong to some presented set (e.g. Jern \& Kemp, 2013; Ward, 1994). The paradigm anticipates that imagination behaves like other cognitive processes in requiring reasoning about the rules (or other requirements) for category membership. The empirical evidence supporting the role of prior knowledge has countered romantic views that imagination stems from some unique unobservable process. For example,

Ward (1994) asked participants to draw an animal from another planet. Most responses contained features typical of Earth animals, (e.g., bilateral symmetry). Hence, knowledge from existing concepts was projected onto the generated exemplars. Evidence for the constraining effect of prior knowledge also comes from the study of cognitive biases in innovation. For instance, functional fixedness (Duncker, 1945) refers to a tendency to focus on an object's most common use. By contrast, McCaffrey \& Krishnamurty (2014) argue that more novel ideas are generated when attention shifts to less frequently noticed attributes of a problem (or in our case, of the stimulus).

In the present study, we attempt to limit reliance on prior knowledge by using sets of abstract stimuli drawn from a continuous multidimensional space. The use of abstract stimuli should increase the reliance on identification of similarity between stimulus features (Tversky, 1977). For example, new exemplars generated to belong to a presented category (see Figure 1) should tend to be similar on features such as color and shape. If only perceptual features are accessible, participants would be prone to adopt feature matching strategies such as replication (e.g., copying one figure directly) or averaging (e.g., generating the mean of the presented examples; see Figure 1). Hence, although the use of abstract stimuli may lessen reliance on prior knowledge, the use of feature matching strategies will likely be enhanced.

The extent to which participants merely replicate or copy one of the presented category members may also depend on whether participants adopt a strong sampling assumption (i.e., that the category members were deliberately chosen as positive examples; perhaps as the only members of that category; Navarro, Dry \& Lee, 2012). Under a weak sampling assumption, participants may view the presented


Figure 1. Examples of sets of abstract stimuli varying in perceptual features; shape, color and feature matching response strategies. Left: replication response strategy. Right: averaging response strategy. The conflict item is the blue category member. The online version is in color.

|  | Dimensionality |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Shape | Color | Both | Shape | Color | Both |
| Variability | Baseline |  |  | Conflict and self-explanation |  |  |
| High | at $x^{2}$ | $\begin{aligned} & x x^{2} \\ & x<x \end{aligned}$ |  | $\begin{aligned} & x x^{x} \\ & x=0 \end{aligned}$ | $\begin{aligned} & * k \\ & k * \end{aligned}$ | $\begin{aligned} & x \\ & x<x \\ & x \end{aligned}$ |
| Med | 28 $v^{2} 0$ | \% ${ }^{2}$ | 20 $x^{2}+8$ | $\begin{aligned} & x x^{2} \\ & x+ \end{aligned}$ | $\begin{aligned} & x x^{k} \\ & x x^{x} \end{aligned}$ | $\begin{aligned} & x \sqrt{2} \\ & x \geqslant \end{aligned}$ |
| Low | $\begin{aligned} & x_{2}= \\ & x^{2}=2 \end{aligned}$ | $\begin{aligned} & \text { it } x \\ & \text { ㅂ }+x \end{aligned}$ | $\begin{aligned} & k x^{2} \\ & 2 k+2 \end{aligned}$ | $\begin{aligned} & k \geqslant k \\ & k \geqslant k \end{aligned}$ | $x x^{2}$ | $\begin{aligned} & x^{2} \\ & x=x \end{aligned}$ |

Figure 2. Examples of Stimulus Variability and Dimensionality for each instruction condition. Baseline condition: top panel. Conflict and self-explanation condition: bottom panel. The online version is in color.
category as being sampled with no restrictions; hence, the presented objects may have implicit or hidden dimensionality with unknown support (i.e., the range of values that exist for a dimension). Our first goal was to examine how variability in the presented set affected the novelty of the generated exemplars.

Ward and Sifonis (1997) varied instructions across three experimental conditions: participants were told to generate (1) an alien animal, (2) an animal wildly different from Earth animals, or (3) a living thing. The second condition produced more unusual creatures than the first condition; while diversity was largest in the third group where many responses did not adhere to standard animal features. The authors' suggested that this was encouraged through ambiguity in the instructions and by participants describing their creature prior to drawing it. A second goal of the present study was to examine whether self-explanation leads to more imaginative exemplars.

The present study aimed to examine the effects of: (i) variability of the presented category, (ii) inclusion of a conflict item, and (iii) prospective explanation and thinking aloud on exemplar generation. We compared a baseline condition to a condition which contained an anomalous exemplar and to a further condition which additionally required participants to engage in self-explanation during the generation process. Cognitive conflict is recognized as a reliable method for promoting the search for new knowledge (e.g. Limon, 2001). For example, a study conducted by Kang, Scharmann and Noh (2004) found conflict recognition promoted the invention of alternative concepts and explanations to account for the disparity caused by a conflicting stimulus. Consequently, presentation of a conflicting item may act to stimulate imagination, highlighting potential options for a new category member. For instance, an anomalous category member could highlight new dimensions from which an exemplar could be sampled.

If the underlying category rules are unclear, selfexplanation may allow for exploration of alternative hypotheses through which to understand the presented


Figure 3. Example of a Single Trial. A) presented category examples, B) black box for drawing generated exemplar, C) color bar and saturation and brightness cube. The online version is in color.
category (Williams \& Lombrozo, 2010). Self-explanation, has a direct influence on how objects are mentally represented and understood. Verbalisation of thought processes during an activity can enable access to cognitive processes which are not directly observable (e.g. Ericsson \& Simon, 1980). Self-explanation can also override the influence of similarity, facilitate generalisation, and promote the integration of novel information with existing knowledge (Lombrozo, 2006). Similar to its influence on intelligence observed in educational psychology models which aim to accelerate cognitive development (Adey, 1992), selfexplanation, therefore, affords a means of engaging imagination in order to resolve conflict.

## Method

## Participants

Participants were 129 University of Melbourne students who received course credit for participation. Of the participants, 16 were excluded as a result of missing data, leaving a total of 113 (91 females, $\mathrm{M}=19.9,3.11$ ). Participants were randomly assigned to one of three conditions; baseline ( $\mathrm{n}=$ $35)$, conflict $(\mathrm{n}=43)$ or self-explanation $(\mathrm{n}=35)$.

## Stimuli

The stimuli were "blob-shaped" radial frequency curves, which varied in shape and color. Four shapes were presented, and participants were asked to generate a new member of the same category by drawing a shape using the mouse and selecting a color. In the baseline condition, the shape of the objects could vary with color identical across all of the category members; the color of the objects could vary with shape fixed, or both shape and color could vary (see Figure 2). The shape was determined by convolving three sine waves of different angles; two degrees, four degrees, and a third angle randomly generated from a normal distribution with a mean of zero and a standard deviation of 10 (low), 12.5 (medium), or 15 (high) ${ }^{1}$.

The color was determined by selecting a starting hue from a set of fully saturated and fully bright hue values between

[^54]|  | Response Strategies Using Perceptual Features |  |  |  |  |  | Response Strategies Not Using Perceptual Features |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Replication |  | Averaging |  | Imaginative Explicit |  | Imaginative Implicit |  |
| Dimensionality | Stimuli | Exemplars | Stimuli | Exemplars | Stimuli | Exemplars | Stimuli | Exemplars |
| Shape | 数 令 3 | $\sum_{3}^{2}$ | $\cdots 3$ |  | 28 28 | $\{3$ | $x+2$ $0 \times 2$ | Nomer |
| Color | $* 3$ $* *$ | 53 | $2 x^{2}$水 |  |  |  | zts to ${ }^{3}$ | $53$ |
| Both |  | $\sqrt{5}$ | $\begin{aligned} & x<2 \\ & x \\ & 23 \end{aligned}$ |  | $\begin{aligned} & 3 \\ & 2 \\ & 2 \end{aligned}$ | $53$ | 082 $x^{2} 20$ |  |

Figure 4．Coding Response Strategies．Exemplars in the black boxes are participant examples of generated responses to presented stimuli．The online version is in color．
zero and 255．These values were converted to RGB values using a look－up table．If color varied，then the hue was adjusted by a normally distributed random adjustment with a mean of zero and standard deviation of 20 （low）， 40 （medium），or 60 （high）${ }^{2}$ ．

In the conflict and self－explanation conditions，three of the items had low levels of variability（as in the baseline condition），and one study item varied markedly from the other three items on shape，color，or on both dimensions（see Figure 2）．The levels of conflict variation were determined in the same way as the baseline condition but with much larger standard deviations．

## Procedure

Participants completed 27 trials．Each trial consisted of presentation of four items displayed on a computer monitor， with instructions indicating that the four items all belonged to a single category．Participants were instructed to use the computer mouse to draw another member of this category in the provided box and to select its color．Colors were selected by choosing a hue from a color bar and then adjusting the saturation and brightness by selecting from the shading box （see Figure 3）．

In the baseline and conflict conditions，after completion of all trials，participants were instructed to provide a retrospective explanation for each trial．For the self－ explanation condition，participants were asked to provide a prospective explanation and then to think aloud whilst drawing their exemplar．Prospective and retrospective verbalizations were recorded on a digital recorder．

Shape and Color Similarity Scoring To provide an objective measure of how closely the generated objects

[^55]matched the presented category exemplars，we developed two measures：（1）The first method used translation，rigid shape rotation，and scaling to find the maximum proportion of overlap between the generated object and each presented exemplar．This measure ranged from 0 and 1 ，with 1 indicating perfect overlap．The shape similarity scores were approximately normally distributed with a mean of 61 （sd＝ 0.13 ）．（2）The second method extracted the color of the generated object and computed the Delta E color difference between the drawing and each of the presented examplars （Wyszeck \＆Stiles，1982）．These scores were positively skewed with a lower bound at 0 （perfect color match）．We log transformed these scores and adjusted the range to 0 and 1 ， and finally subtracted the scores from 1 so that 1 indicated a perfect color match．The final score was the minimum score across all four presented category members．The color similarity scores had a mean of $.18(\mathrm{sd}=.07)$ ．

Expert Coding Note that our shape overlap similarity measure does not really capture the extent of creativity in the generated exemplars；nor does it capture other potentially interesting patterns of exemplar generation．To capture these patterns，we had two experts（one was the first author） classify each drawing into whether the generated object replicated one of the presented exemplars，whether the generated object appeared to be an average of the presented exemplars，or whether the generated object exhibited imaginative characteristics．This latter rating category was further broken down into cases which used the perceptual features of the presented shapes（termed explicitly imaginative；i．e．，imaginative responses which utilized variations in blob－shape or color；e．g．，a butterfly which is blob－shaped）or whether the generated exemplar drew on implicit characteristics（we term this rating category

All values were selected based on pretesting to ensure perceptibly low，medium，and high levels of variation．
implicitly imaginative; i.e., response which utilize features other than blob-shape or color; e.g., a drawing of a space ship - which may have been justified as necessary to interact with


Figure 5. Average shape similarity (top panel) and color similarity (bottom panel) as a function of dimensionality and variability in the baseline condition.
the alien shapes that had been presented as part of the category). There was an $86 \%$ agreement between our expert raters; disagreements were then clarified between raters by drawing on the recordings generated by participants. Figure 4 shows exemplars representative of each response strategy.

## Results

## What is the effect of variability in the presented category on the generated exemplars?

We first examined the effects of dimensionality and variability within the baseline condition on both shape and color similarity. Dimensionality refers to what aspects of the presented category exemplars varied within the presented set (e.g., shape could vary with color fixed across all four category members, color could vary with shape fixed, or both
could vary). Variability refers to the level of variability applied to the dimensions which were not fixed in the presented set. We examined only the baseline condition because variability in the conflict and explanation conditions was instantiated via the anomalous items and is not comparable.

Figure 5 shows that there was little effect of dimensionality or variability on the shape similarity scores. On the other hand, the color similarity scores systematically decreased with increasing variability whenever color varied in the presented set. That is, whenever the presented category exemplars varied in color, people generated colors which were more dissimilar (to the presented colors) than when color was fixed or of lower variability.

A two-way Dimensionality $x$ Variability ANOVA confirmed these results. The largest F-ratio for the shape scores was the effect of variability, $F(2,882)=1.45, p=.23$. For the color scores, there was a main effect of dimensionality, $F(2,882)=151.33, p<.001$, a main effect of variability, $F(2,882)=27.35, p<.001$, and an interaction, $F(4,882)=11.06, p<.001$.


Figure 6. Average shape similarity (top panel) and color similarity (bottom panel) scores in the baseline, conflict and explanation conditions. Error bars are $+/-1$ standard error.

## Does conflict or explanation lead to less similar generated exemplars?

Automated Scores We compared the average shape and color scores for each participant across the baseline, conflict, and explanation conditions. There was a general decreasing trend across these conditions for both shape and color
similarity (see Figure 6). Both effects were significant using a one-way ANOVA: (Shape: $F(2,110)=3.2, p=.044$; Color: $F(2,110)=64.05, p<.001)$.

Expert Ratings The proportion of each drawing type is shown in Figure 7. It is immediately obvious that the main strategy for exemplar generation was to simply replicate one of the presented category members. While all conditions showed some evidence of generating imaginative exemplars, the level of implicitly imaginative exemplars appears to be higher (with a co-occurring decrease in replication) in the explanation condition. This was confirmed by a significant chi-squared test for independence across all three groups, $\mathrm{X}^{2}$ (6) $=63.46, p<.001$, and for the baseline vs explanation comparison, $\mathrm{X}^{2}(3)=49.11, p<.001$, the conflict vs explanation comparison, $\mathrm{X}^{2}(3)=34.59, p<.001$, but not the baseline vs conflict comparison, $\mathrm{X}^{2}(3)=3.02, p=.15$.


Figure 7. Expert rating proportions for the shape of the generated objects for each strategy type: replication, averaging, explicit imaginative, and implicit imaginative.

## Discussion

This experiment examined the effects of stimulus variability, which included adding a conflict item to the presented category set, as well as the impact of prospective selfexplanation and thinking aloud on exemplar generation.

We first showed that variability on a color dimension leads to more novel generation of values for that color dimension. We did not find an effect of variability on shape. There are a few potential explanations for this difference between shape and color similarity. For one, our operationalization of shape similarity as overlap did not adequately capture the more creative responses that were identified by our raters (see Figure 4, for examples). Second, color was easier to generate in our experiment than shape. Shape had to be hand-drawn but color could be selected from a color palatte. Hence, a third difference was that the range of color options was presented to the participant, but the range
of possible shapes was unbounded. The use of a mouse instead of a stylus may have restricted the shapes that were generated; we leave this as a goal for future experiments.

We next showed that including an anomalous exemplar and allowing self-explanation, led to responses which were less similar to the presented category members. Our rating analyses clearly showed that in all conditions, replication was the most utilized strategy. Examination of the drawings revealed that most subjects drew blob like objects and that typically these blobs looked primarily like one of the presented objects. The predominant application of replication strategies in all conditions in this study may have been anticipated through our use of abstract stimuli. The use of abstract stimuli promoted the predilection toward perceptual features of the presented stimuli (Tversky, 1977), and the absence of feedback on task performance made category comprehension more ambiguous. Consequently, a replication strategy represented the simplest approach to meet task parameters, whilst requiring the least amount of cognitive effort. This result may also indicate that participants adopted a strong sampling assumption (Navarro et al., 2012) and simply sampled from the presented set.

Despite stimuli having no direct links to prior conceptual knowledge, some responses drew on knowledge external to the perceptual features of the manufactured categories. This was more prevalent when participants were encouraged to self-explain. This implies that the method of instruction, along with the method of response, plays an active role in both understanding category membership and the subsequent exemplar generation process.

Presentation of a conflicting category member did not result in more or less imaginative exemplars. The implication is that the conflict item proved difficult to assimilate, resulting in replication being favoured as a strategy. The importance of delivering a conflict at the appropriate level so as to sustain interest has been demonstrated in previous research in learning. As noted by Limon (2001), the presentation of contradictory data can only result in a meaningful conflict if it presents a challenge to existing held beliefs. If the basis of the conflict is not understood, it fails to engage the person, and therefore the conflict may be ignored or explained away (Adey, 1992). Our results suggests that conflicting items which maintain the same explicit dimensional structure might limit the recruitment of imagination. We leave it for future research to examine conflicts which signal the implicit dimensionality of the concept.

On the other hand, prospective self-explanation promoted greater use of imaginative strategies. This indicates that cognitive interaction via self-explanation can foster imaginative responses to category conflicts. It appears that self-explanation provided a mechanism for reflective abstraction of the conflict (Adey et al, 2007), and encouraged imagination to resolve the problem. In line with studies into the influence of self-explanation in dealing with anomalous data (e.g., Williams, Lombrozo, \& Rehder, 2011), selfexplanation encourages the greater use of imaginative
strategies in response to a category conflict. It is the combination of both a conflict and a mechanism to explore the reasons behind the conflict which increases the use of imagination.

A limitation of the current study is that it failed to address whether self-explanation increases the likelihood of imaginative responses without the involvement of a conflict category member. It remains unclear whether highly imaginative responses would have arisen if a more complete range of feature variation had not been revealed by the conflict item. Therefore, repeating the current study allowing explanation in the baseline condition would facilitate a better understanding of the importance of self-explanation in promoting imaginative responses. Future studies should also explore the role of self-explanation in the use of replication. When a conflict category member was presented amongst similar stimuli, participants following a replication strategy either chose to replicate one of the three similar items or the minority conflict item. Understanding the patterns which underlies this decision making is important, as it represents the starting point for imaginative responses.

The motivation for the present study was to better understand the role of imagination in exemplar generation. To address the gap in the literature about the impact of instructions on structured imagination, abstract stimuli were used as a means of reducing access to prior knowledge when generating exemplars. Although participants favoured a strategy which leveraged perceptual features, the current study provides evidence of imaginative responses leveraging implicitly related prior knowledge. In addition, selfexplanation was shown to be an effective mechanism in generating imaginative exemplars in the presence of category conflict. This experiment confirms that self-explanation makes structured imagination more flexible when interacting with unexpected categorisation tasks, and represents the starting point for greater exploration into how imagination responds to cognitive challenges.

It is worthwhile to consider how one might develop a computational model of exemplar generation. Clearly an essential mechanism is the ability to retrieve instances or features from members of stored categories and then to combine these retrieved features. This type of mechanism is reminiscent of the echo content mechanisms in Minerva (Hintzman, 1984). Our results suggest a mix of strategies which direct retrieval of one of the presented category members being the most common. The fact that replication is increased in the conflict condition suggests a role for selective attention in determining retrieval. However, in some cases, there appears to be probabilistic sampling not only of the physically presented shape and color dimensions, but also of dimensions which are implicit to the presented category and likely more conceptual than physical (see Figure 4). McCaffrey \& Krishnamurty (2014) propose a taxonomy of different feature types that ranges from physical features such as size, shape, mass, weight, to the identification of object parts, to the types of functions or uses an object has, its super- and subordinate categories,
associated concepts, aesthetic values, and causal relations. In this taxonomy, only certain types of features (or dimensions) are immediately available to sensory perception. Selfexplanation seems to result in an increased probability of sampling from more implicit dimensions.

## Acknowledgments

DL was supported by an ARC Discovery Project grant: DP160102360. We also thank Grace Kilmer for help with rating the responses.

## References

Adey, P. (1992). The CASE results: implications for science teaching. International Journal of Science Education, 14, 137-146.
Adey, P., Csapo, B., Demetriou, A., Hautamaki, J., \& Shayer, M. (2007). Can we be intelligent about intelligence?: Why education needs the concept of plastic general ability. Educational Research Review, 2, 75-97.
Duncker, K. (1945). On problem-solving. Psych Monographs, 270, 58, i-113.
Ericsson, K. A., \& Simon, H. A. (1980). Verbal reports as data. Psych Review, 215-251.
Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. Behav Research Methods, 16, 96-101.
Jern, A. \& Kemp, C. (2013). A probabilistic account of exemplar and category generation. Cog Psych, 66, 85-125.
Kang, S., Scharmann, L. C., \& Noh, T. (2004). Re-examining the role of cognitive conflict in science concept learning. Research in Science Education, 34, 71-96.
Limon, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. Learning and instruction, 11, 357-380.
Lombrozo, T. (2006). The structure and function of explanations. Trends in cognitive sciences, 10, 464-470.
McCaffrey, T. \& Krishnamurty, S. (2014). The obscure features hypothesis in design innovation. International Journal of Design Creativity and Innovation, 3, 1-28.
Navarro, D. J., Dry, M. J. \& Lee, M. D. (2012). Sampling assumptions in inductive generalization. Cognitive Science, 36, 187-223.
Tversky, A. (1977). Features of Similarity. Psychological Review, 84, 327-352.
Ward, T. (1994). Structured Imagination: The role of category structure in exemplar generation. Cognitive Psychology, 27, 1-40.
Ward, T. B., \& Sifonis, C. M. (1997). Task demands and generative Thinking: What changes and what remains the same? The Journal of Creative Behaviour, 31, 245-259.
Ward, T., Patterson, M. J., \& Sifonis, C. M. (2004). The role of specificity and abstraction in idea generation. Cognitive Research Journal, 16, 1-9.
Williams, J. J., \& Lombrozo, T. (2010). The role of explanation in discovery and generalisation: Evidence from category learning. Cognitive Science, 34, 776-806.
Williams, J. J., Lombrozo, T., \& Redher, B. (2011). Explaining drives the discovery of real and illusory patterns. Proceedings of the $33^{\text {rd }}$ annual conference of the Cognitive Science Society (pp. 1352-1357)
Wyszecki, G., \& Stiles, W. S. (1982). Color science (Vol. 8). New York: Wiley.

# A hierarchical Bayesian model of "memory for when" based on experience sampling data 

Simon Dennis(simon.dennis@gmail.com)<br>Hyungwook Yim (hyungwook.yim@newcastle.edu.au)

School of Psychology, University of Newcastle, Australia

Vishnu Sreekumar (vishnu.sreekumar85@gmail.com)<br>Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke National Institutes of Health, USA

Nathan J. Evans (nathan.j.evans@uon.edu.au)<br>Paul Garrett (paul.garrett@uon.edu.au)

School of Psychology, University of Newcastle, Australia
Per Sederberg (sederberg.1@osu.edu)
Department of Psychology, The Ohio State University, USA


#### Abstract

Participants wore a smartphone, which collected GPS, audio, accelerometry and image data, in a pouch around their necks for a period of two weeks. After a retention interval of one week, they were asked to judge the specific day on which each of a selection of images was taken. To account for people's judgements, we proposed a mixture model of four processes - uniform guessing, a signal detection process based on decaying memory strength, a week confusion process and a event confusion process in which the sensor streams were used to calculate the similarity of events. A model selection exercise testing all possible subsets of the processes favoured a model that included only the event confusion model. GPS similarities were found to be the most significant predictors, followed by audio and accelerometry similarities and then image similarities.


Keywords: memory, experience sampling, hierarchical Bayesian model

## Introduction

Friedman $(1993,2004)$ argued that people typically employ one of four strategies to identify when events occurred. On some occasions, people can directly retrieve declarative knowledge about the event. For instance, many people can recall that the attacks on the Twin Towers occurred on September 11th 2001. Friedman argues, however, that such declarative knowledge is quite rare and is reserved for events of global or personal significance. On other occasions, people have relative order information that they can use to make a judgement. If one were asked when George W. Bush initiated military action in Afghanistan, one may not know the date, but one can make an inference. The military action in Afghanistan occurred as a consequence of the September 11th attacks and therefore was most likely to have been in late 2001. There is a natural order in which these events occurred and provided someone has the order information and access to the time of the
original event they can deduce the timing of the subsequent event. Again, Friedman argues that judgements based on relative order information are rare.

More common, according to Friedman, are judgements made using location-based strategies. Location-based processes rely on the retrieval of information associated with the cues that can be used to draw inferences about the timing of an event. For instance, suppose you are asked when you last saw your friend Mary. You might recall that you share a Psychology-101 class with Mary. Furthermore, you know that Psychology-101 occurs on Mondays and Wednesdays at 2 pm . It is now Saturday, so you infer that it was Wednesday at 2 pm when you last saw Mary.

Sometimes, however, the necessary knowledge to make an inference is not available. In these circumstances, Friedman argues one resorts to a distance-based strategy. Distance-based strategies rely on some quality of the memory that changes as a function of time. For instance, one might judge strong memories as having occurred more recently.

There is a substantial literature that has asked people to report on the time at which events occurred (see Friedman, 1993 and Thompson, Skowronski, Larsen \& Betz, 1996, for reviews). Much of this literature has involved the memory for events that occurred outside of the laboratory, but which can be dated because they are part of the public record or have been recorded in personal diaries (Kemp, 1999). Generally, people are very poor at identifying when events occurred showing a bias to report events as being too recent when they occurred remotely in time - forward telescoping (Huttenlocher, Hedges, \& Prohaska, 1988) or too remote when they occurred recently - backward telescoping (Hinrichs \& Buschke, 1968).

Early distance-based theories proposed that the psychological representation of time was logarithmically compressed, much as other psychophysical dimensions are
(Ferguson \& Martin, 1983). These theories are able to account for the decrease in accuracy that occurs with retention interval, but have been discounted because well-known remote events are often dated accurately. If it were the time axis itself that was compressed, dating accuracy should always be directly related to recency (Huttenlocher et al., 1988). Alternatively, the accessibility of the memory trace could be used to infer the time of occurrence. Some evidence suggests that better known events are dated more recently (Brown, Rips \& Shevell, 1985). However, there is substantial subsequent evidence that this does not occur (Thompson et al. 1988). Again, better known events are reported more accurately. Furthermore, people are often capable of accurately reporting the day of the week an event occurred, while struggling to faithfully retrieve the month or year (Friedman \& Wilkins, 1985). If people were using distance-based strategies, accuracy at smaller temporal scales ought to be worse than at larger temporal scales. The existing literature has tended to conclude that location-based strategies are far more commonly employed than distance-based strategies (Friedman, 1993; Thompson et. al. 1996).

There are two main models of time reporting that have been proposed (Huttenlocher, Hedges \& Prohaska, 1988; Kemp, 1999). The first of these, by Huttenlocher, Hedges, and Prohaska (1988), proposes that events are associated with time information, which is unbiased but subject to error that increases with time. Bias is introduced because answers are constrained to lie within the reference period either implicitly or explicitly defined by the memory query thus generating forward and backward telescoping. In addition, the theory posits that memory units are organized hierarchically - (e.g. day, month, year) and that events may be associated with any of these levels. The model provides a good quantitative fit to data they collected on judgements of when movies that were part of a campus initiative were shown.

Kemp's (1999) theory is similar to Huttenlocher's in that the representation of time is not systematically distorted and that reconstruction of this time information is the basis of memory judgements. Rather than suggesting that temporal information is distorted with age, Kemp (1999) proposed that when time information is available it is accurate regardless of age. However, only a small number of memories have stored time information. Events of a similar kind are associated with each other and retrieval proceeds by retrieving similar events until one is found for which time information has been stored. An inference is then made on the basis of this information.

Both the Huttenlocher and Kemp theories are typically construed as location-based theories because they rely on the retrieval of time information (i.e., a location in time) and inference proceeds on the basis of that information. However, another possibility is that the time information to which they refer evolves in a continuous fashion on multiple
time scales simultaneously. This kind of model is commonly employed to account for grouping effects in short term serial recall (e.g., Henson, 1998). Furthermore, it is possible that what is retrieved from memory is a combination of specific content on which conscious inferences can be drawn (location-based) and this more graded hierarchical form of context (distance-based).

Although it has long been argued that memory research that is focused solely on laboratory work is futile (Neisser, 1976), the difficulty has been how to proceed when the experience of the participant before they enter the laboratory cannot be rigorously captured. Today, however, technology provides us with entirely new options. Easy to carry and able to monitor multiple sensor streams, smartphones can provide a convenient and ubiquitous window into the events of the life of an individual. We had participants wear a phone around their necks for two weeks and collected image, audio, GPS and accelerometry data. We then developed a hierarchical Bayesian model to capture distance and location based processes.

## Method

## Participants

A total of 18 adult participants were recruited from flyers posted at the University of Newcastle and received \$100 compensation.

## Procedure

In prior work, we built a system that consists of an Android app, server infrastructure and user interfaces. The app acquires image, time, audio (obfuscated), GPS, accelerometer and orientation information at approximately five minute time intervals. The app runs in the background as a service and users carried the phone in a pouch attached to a neck strap from morning till evening (see Figure 1). Participants could turn off the app anytime they needed privacy. When the phone detects WiFi and is charged, it sends the stored data automatically to a remote server. This usually happened once per day at the end of the day when users charge the phone overnight.

Participants were instructed to wear the phone for two weeks. They returned to the laboratory on the Friday of the 3rd week and were presented with images one at a time and were required to determine on which of the week days each image was taken (participants were informed that the images only came from the week days). Each participant's test was based on images drawn from their own lifelogs. We selected images that came from distinct episodes as much as possible, and also avoided using images that were blurred due to excessive motion. The number of stimuli varied between participants since the available data depended on individual lifestyles. A presented image remained on the screen while they made the day judgment and they could use as much time as they needed to respond.


Figure 1. Android phone worn by a participant during the experience sampling phase of the study.

## Modeling

To account for people's judgements, we proposed a mixture model of four processes - (1) random (uniform) guessing, (2) a signal detection process based on decaying memory strength (distance), (3) a week confusion process (location) and (4) a event confusion process (location) in which the sensor streams were used to calculate the similarity of events. We start by describing the distance and location (sensor) models and then outline the mixture model incorporating all four processes.

## Distance model

Figure 2 depicts the distance based model we employed. Mean memory strength $(\mu)$ elicited at retrieval was assumed to decay exponentially with scale, $\alpha$, asymptote, $\beta$, and rate, $\lambda$. Variability around this mean was assumed to be Gaussian with standard deviation, $\sigma$. The probability of a response occurring given that the presented stimulus was taken on a given day, is given by the probability density that falls between criteria that separate it from the neighbouring days.
The nine criteria that determined the response probabilities for each day were fixed to the mid-point of $\mu$ values of each neighbouring day (to alleviate sampling issues that resulting from attempting to estimate these as free parameters). We used Bayesian hierarchical modeling to fit the model, with each individual's parameters being constrained by a group level distribution. All parameters were sampled on a double infinite scale, meaning that we sampled the inverse Probit of $\alpha$ and $\beta$, and the natural logarithm of $\lambda$ and $\sigma$, and that all group level distributions were, therefore, normal.


Figure 2. The distance based model.

## Location (sensor) model

The location (sensor) model assumed that events were stored in memory and that the likelihood of confusing the representation of the correct event with the stored representation of another event is determined by the similarity of those events. Each day was divided into hour periods and image, GPS, audio and accelerometry representations of those events were calculated. For a given sensor stream the distance of an image's event to a given day for a given stream was taken to be the minimum Jenson Shannon distance from the event to the events of that day (see Figure 3).


Figure 3. The sensor model.
These distance scores for each of the streams entered into a conditional logit model to determine the probability that the participant would respond with a given day. Missing data
were assumed to have a prior of a truncated normal distribution.
Like the distance model, the parameters were estimated in a hierarchical fashion, with the natural logarithm of the weights being estimated, making the group-level distributions normal.

## Mixture model

To estimate the probability of a participant's response we assumed a mixture model of the distance and location (sensor) models described previously as well as a random (uniform) guessing model and a location (week) model that assumed that participants correctly inferred the day of the week on which the event occurred, but had a certain probability of incorrectly determining the week (the kind of model that one might assume if people are relying on their schedules to make judgements).

A model selection exercise testing all possible subsets of the processes was conducted using the common model selection metric WAIC, which attempts to weigh both the goodness of fit to the data and the complexity of the model, in order to approximate the leave-one-out cross validation metric. The preferred model was the location (sensor) model although the location (sensor) + random model also performed well (see Table 1).

Table 1: Models tested and corresponding WAICs

| Model | WAIC |
| :--- | :---: |
| Location (sensor) | -1544 |
| Location (sensor) + Random | -1546 |
| Location (sensor) + Distance | -1557 |
| Location (sensor) + Distance + Random | -1565 |
| Distance + Random | -1649 |
| Location (week) + Random | -1958 |
| Distance + Location (week) | -2161 |
| Random | -2544 |
| Distance | -2688 |
| Distance + Location (week) + Random | -2812 |
| Location (week) | $-\infty$ |

To understand the performance of the models, it is useful to compare the posterior confusion matrices they produce to the data. Figure 4 shows the confusion matrix of responses accumulated over participants. The x -axis show the days on which the events actually occurred, and the $y$-axis shows the participants' responses. The diagonal represents correct responses, while responses off the diagonal are errors. The matrix is dominated by correct responses, with cells close to the major diagonal (representing adjacent days) showing significant mass particularly in week one.


Figure 4. Data confusion matrix


Figure 5. Distance only confusion matrix
The distance only model is able to explain the structure off the diagonal by estimating a large standard deviation for the strength distributions (see Figure 5). However, a large standard deviation prevents the model from capturing the proportion of correct responses on the diagonal.
The distance model performs much better when it is mixed with the uniform distribution (see Figure 6). The structure off the diagonal is captured by the uniform component, while the structure on and adjacent to the diagonal is captured by the distance model. The observation that counts adjacent to the diagonal are larger in week one is accommodated naturally by the model because in the first
week the gradient of the strength is small, which makes it more difficult to distinguish between adjacent days.


Figure 6. Distance + Random confusion matrix

The week only model does poorly. The model assigns no probability to cells that are neither on the diagonal nor exactly a week out (the off diagonals five above and below the main diagonal). As there are observations in those cells, the WAIC is negative infinity. When mixed with the random model, the model does better, but still has a tendency to predict more week out responses than appears in the data (see Figure 7).


Figure 7.: Week + Random confusion matrix


Figure 8. Sensor confusion matrix

The best model is the sensor model (see Figure 8). Unlike the distance and week models the sensor model did not require the random component in order to provide a good fit to the data. In fact, adding the random component decreases the WAIC slightly as the model is penalized for additional complexity (i.e. the mixture probability). While the distance and week models are informed only by the day on which the event occurred, the sensor model constructs a representation of the event that captures where the participant was (GPS), what the participant was hearing (audio), what the participant was doing (accelerometry) and what the participant was seeing (images) and compares it with representations of all other events. The importance of these features can be inferred from the weights associated with each of the data streams.


Figure 9. The posterior distributions of the sensor weights.

Figure 9 shows the posterior distributions of these weights. The GPS stream has the strongest weights followed by the audio stream and the accelerometry stream, which are approximately equal. The image stream has the lowest weights. That the image stream should have the lowest weights is counter intuitive. The participants are presented with the image as a retrieval cue, and so one might have expected the visual information to be salient.

There are multiple possible interpretations of this result. It may be that the image representation that we chose (GIST; Oliva \& Torralba, 2001) does not carry the information that participants rely upon when making memory judgements. Another possibility is that it is the static nature of the images or the fact that they are not synchronized with the direction of gaze that compromised this stream. While head mounted video technologies exist they are currently difficult to deploy for the duration of recording required for the time scales we explore here. Furthermore, they introduce additional ethical hurdles that need to be considered. A third possibility is that the result is not artifactual, but is a reflection of the information employed by the memory system. While the visual domain seems salient perhaps it is other aspects of experience that drive the retrieval and inferential systems that people employ to make location based judgements.

## Conclusions

When people are asked to determine when an event occurred, Friedman (2004) argues that people use a combination of distance based and location based processes, with location based processes being the most common. The current work supports this assertion.

Furthermore, we have demonstrated that it is possible to predict the responses people will make to images taken from their personal experience in the world outside the laboratory on a stimulus by stimulus basis. We believe this work establishes a new benchmark for what models of episodic memory should achieve and provides the promise of a more quantitatively rigorous, ecologically valid and translationally relevant memory science.

## Acknowledgments

We would like to acknowledge the funding from the Australian Research Council (DP150100272) and Air Force Office of Scientific Research (FA9550-09-1-0614).

## References

Chalnick, A., \& Billman, D. (1988). Unsupervised learning of correlational structure. Proceedings of the tenth annual conference of the cognitive science society (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.
Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum \& J. Feldman (Eds.), Computers and thought. New York: McGraw-Hill.

Ferguson, R. P., \& Martin, P. (1983). Long-term temporal estimation in humans. Perception and Psychophysics, 33, 585-592.
Friedman, W. J. (1993). Memory for the time of past events. Psychological Bulletin, 113(1), 44-66.
Friedman, W. J. (2004). Time in autobiographical memory. Social Cognition, 22(5), 591-605.
Friedman, W. J., \& Wilkins, A. J. (1985). Scale effects in memory for the time of past events. Memory and Cognition, 13, 168-175.
Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. Cognitive Psychology, 36, 73-137.
Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. Cognition and Brain Theory, 6, 287-317.
Huttenlocher J., Hedges L., \& Prohaska V. (1988). Hierarchical organization of ordered domains: Estimating the dates of events. Psychological Review, 95, 471-484.
Huttenlocher J., Hedges L., \& Prohaska V. (1992). Memory for day of the week: A $5+2$ day cycle. Journal of Experimental Psychology: General, 121(3), 313-325.
Kemp, S. (1999). An associative theory of estimating past dates and past prices. Psychonomics Bulletin and Review, 6, 41-56.
Matlock, T. (2001). How real is fictive motion? (Unpublished doctoral dissertation), Psychology Department, University of California, Santa Cruz.
Neisser, U. (1976). Cognition and reality: Principles and implications of cognitive psychology. New York: Freeman.
Newell, A., \& Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Ohlsson, S., \& Langley, P. (1985). Identifying solution paths in cognitive diagnosis (Tech. Rep. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.
Oliva, A., \& Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3), 145-175.
Shrager, J., \& Langley, P. (Eds.) (1990). Computational models of scientific discovery and theory formation. San Mateo, CA: Morgan Kaufmann.
Sreekumar, V., Dennis, S., Doxas, I., Zhuang, Y., \& Belkin, M. (2014). The geometry and dynamics of lifelogs: discovering the organizational principles of human experience. PloS one, 9(5), e97166.
Thompson, C. P., Skowronski, J. J., \& Lee, D. J. (1988). Reconstructing the date of a personal event. In M. M. Gruneberg, P. E. Morris, \& R. N. Sykes (Eds.), Practical aspects of memory: Current research and issues: Vol. 1. Memory in everyday life. Chichester: Wiley.
Thompson, C. P., Skowronski, J. J., Larsen, S. F., \& Betz, A. L. (1996). Autobiographical memory: Remembering what and remembering when. Mahwah, NJ: Erlbaum.

# Refixations gather new visual information rationally 

Yunyan Duan (yduan@u.northwestern.edu) and Klinton Bicknell (kbicknell@ northwestern.edu)<br>Department of Linguistics, Northwestern University<br>Evanston, IL 60201 USA


#### Abstract

The standard model is that word identification in reading is a holistic process, most efficient when words are centered in the visual field. In contrast, rational models of reading predict word identification to be a constructive process, where readers efficiently gather visual information about a word, and may combine visual information about different parts of the word obtained across multiple fixations to identify it. We tease apart these accounts by arguing that rational models should predict that the most efficient place in a word to make a second fixation (refixation) depends on the visual information the reader has already obtained. We present evidence supporting this prediction from an eye movement corpus. Computational model simulations confirm that a rational model predicts this finding, but a model implementing holistic identification does not. These results suggest that refixations can be well understood as rationally gathering visual information, and that word identification works constructively.


Keywords: eye movements; reading; word identification; rational analysis; refixation

## Introduction

Reading is a complex information processing task with a goal usually related to comprehending the text. In general, accurate text comprehension requires the identification of many (if not most) of the words in a text. It is not surprising, then, that decades of research on eye movements in reading have established that word identification can be seen as the primary driver of eye movements (Rayner, 1998). A substantial body of work has studied the role in this process of many information sources relevant to word identification in reading, including especially word frequency and in-context predictability, among others. However, although visual information is the primary source of information used to ultimately identify a word, the fundamental way in which visual information is used in word identification remains unresolved.

In the standard model of word identification in reading, word identification is hypothesized to be a holistic process, during which visual information about the word as a whole constrains the efficiency of identification. Eye movement studies have shown that a word presented in isolation is most rapidly identified when fixating approximately at its center (O'Regan, 1990, 1992). It has also been found in natural reading that the fixation position that minimizes gaze duration (the total amount of time spent fixating a word in first pass) and refixation probability (the probability of making more than one fixation on a word in first pass) is on average at or slightly left of the center (Rayner, Sereno, \& Raney, 1996). One explanation for these results is that when the word center is directly fixated, the largest possible part of the word falls in the central high-acuity portion of the visual field (the fovea), yielding the highest-quality visual input of the whole word; as the fixation deviates from the center, more letters of the word fall out of
the fovea and suffer from a rapid drop in acuity, leading to poorer visual information about the overall word. Following this interpretation, it is hypothesized that visual processing efficiency of a word is maximized when fixating at word center, and decreases with increasing distance between word center and fixation position. This standard holistic account is incorporated in dominant eye movement models of eye movement control in reading (e.g. E-Z Reader, Reichle et al., 2009; and SWIFT, Engbert et al., 2005).

Alternatively, word identification may not utilize visual information holistically, especially in natural reading. Unlike in isolated word identification where information about a word comes only from visual input obtained by directly fixating it, in natural reading information about a word comes from more sources. These include contextual information from the preceding text and visual information obtained from fixations close to but not on the current word, which may still yield some visual preview of the word's initial letters. As a result, the most efficient positions from which to obtain useful new visual information about the word can vary from trial to trial, dependent on the information already obtained. Even in such an account, it is still possible that, on average, the most efficient positions are located near the center (as has been found in prior work). This account of word identification is implemented in rational models, which consider reading as a process of combining information from various sources to identify words and making eye movement decisions to maximize identification efficiency (Bicknell \& Levy, 2010, 2012; Legge et al., 1997, 2002). For example, if a reader in this framework is working to identify a particular word, considering all the information that has already been gathered, there may be parts of the word that the reader has already identified relatively well and parts that are still relatively uncertain. It is intuitive in such a situation that identification efficiency will be maximized by moving the eyes next to the part of the word about which the reader is still relatively uncertain. This is because such fixations would obtain fine-grained visual information of a particular part of the word, which can be combined with visual information obtained from previous fixations (as well as linguistic contextual information), and identify the word in a constructive manner. Thus, contrary to the holistic account's view that any fixation landing on a non-central position slows identification efficiency relative to a central fixation, the view from rational models is that the position in the word to move the eyes next to maximize identification efficiency will vary from trial to trial and depend on information already obtained.

A phenomenon that can be used to tease apart these two accounts is that of refixations, cases in which a word is fixated more than once during first-pass reading. The goal of an in-
tended refixation is assumed to be moving the eyes to a position that will maximize identification efficiency of the current word. Despite previous experiments showing that refixation rate varied on average as a quadratic function of the distance between word center and the fixation position (McConkie, Kerr, Reddix, Zola, \& Jacobs, 1989) and was influenced by linguistic properties such as word frequency (Rayner et al., 1996), few studies shed light on where refixations go. The two accounts of word identification make different predictions for this question. The rational model predicts that refixations target the part of the word about which sufficient information has not yet been obtained. Which part of the word this is depends on the visual information already available. ${ }^{1}$ In contrast, the standard model of word identification predicts that refixations should always target the center to maximize the holistic visual processing efficiency of the word, independent of information obtained about different parts of the word.

Naively, then, we could tease apart these two hypotheses by analyzing the relationship between the position of the initial fixation on a word (the 'landing position') and the refixation position. The rational account would predict that if the landing position is at the beginning of the word, a refixation should be at the end, and vice versa, whereas the standard model would predict that all refixations cluster around the center, regardless of the landing position. Empirically, this prediction of the rational models is borne out (Rayner et al., 1996), but the standard model explains this phenomenon in a different way. Specifically, there is a concept of systematic error (McConkie et al., 1988), which suggests that intended saccade sizes become biased toward the overall average saccade size. This means that refixation saccades intended to be short and target the center of the word in the standard model will tend to overshoot their target, landing on the opposite end of the word. Thus, both the standard model combined with systematic error and the rational model predict the effect of landing position on where refixations go.

Analyzing where refixations go as a function of the location of the previous fixation made before fixating a word (the 'launch site'), however, can tease apart these two accounts, when controlling for effects of landing site. If a reader's first eye movement to the word is launched from a position close to the word, then more visual information about the word's beginning should be available (relative to the launch site being further away), holding constant the landing site. Therefore, rational models predict that for closer launch sites, a refixation should be less likely to move the eyes back toward the beginning of the word (Fig. 1, right panel). In contrast, the standard model would not predict such an effect, but predict that an intentional refixation that follows a fixation on the left half of a word should always go forward, while one that follows a fixation on the right half should always go backward, always targeting the word center (Fig. 1, left panel).

[^56]

Figure 1: The standard model and the rational model make different predictions for where refixations go. For the standard model, refixations always target the center of the word, regardless of launch site. For the rational model, refixations target positions where character identity has low confidence (here represented by hypothetical $m(j)$ values). Therefore, closer launch sites, which provide more visual information about the word's initial letters (schematically represented here by grey rectangle) predict refixations are more likely to move forward. The refixation decisions here are based on eye movement policy parameters of $\alpha=.9$ and $\beta=.7$. (See Eye movement policy section for more details.)

In this paper, we empirically evaluate these two competing predictions by performing a statistical analysis of where refixations go in a large eye movement corpus, and we compare these results to simulations from computational models of both accounts. In the next section, we report the results of our statistical analysis of human refixation data, showing that it is as predicted by the rational account. We then confirm that an eye movement model that implements the standard model cannot accommodate this finding by performing simulations with E-Z Reader (Reichle et al., 2009). After that, we describe our rational model of refixations. Finally, we confirm that simulations using it show the same qualitative pattern as the human data, and then conclude.

## Experiment 1: Human data in Dundee corpus

This analysis aims to tease apart the predictions of the rational model and the standard model on where refixations go. Specifically, we use the English part of the Dundee corpus (Kennedy, 2003) of eye movements during natural reading, and analyze the direction of refixation as a function of launch site, statistically controlling for landing site.

## Methods

Data The English part of the Dundee corpus contained eye movement records from 10 native English-speaking participants as they read through newspaper editorials (see Kennedy \& Pynte, 2005 for further details.) We first did a set of screening procedures, according to criteria that are generally applied to eye movement data, to remove fixations involving blinks, non-first-pass fixations, and the first/last two fixations
of the line. After this procedure, the corpus contained 23,854 fixations that were followed by a refixation during first-pass reading ( $18.9 \%$ of first-pass fixations). These data then underwent screening procedures excluding: (a) extremely far launch sites ( $1 \%$ ), leaving the launch sites of fixations in the range $[-16,-1]$ (in terms of number of characters from word beginning); (b) fixations that landed on the space right before the word ( $25.5 \%$ ) or on the last character of the word (4.7\%) to ensure the variability of refixation directions; and (c) fixations on words of which the previous word was skipped to eliminate possible overshootings of the previous word ( $20.9 \%$ ), since these can be followed by corrective saccades. In the end, the data consisted of 7,667 fixations.

Statistical analysis A logistic generalized linear mixedeffects model (GLMM) was used to analyze the direction of refixations (forward vs. backward). Fixed effects included launch site and combinations of word length and landing site, which controlled for arbitrary effects of word length and landing site on refixation direction. Random effects included a random intercept and slope of launch site by subjects. Significance testing was via likelihood ratio test. All statistical analyses were implemented in the R environment, using the glmer function from the lme 4 package (Bates, Mächler, Bolker, \& Walker, 2015) for GLMM implementation. In order to ensure model convergence, word length-landing site pairs for which all refixations (or all but 1 ) moved in the same direction were excluded, leaving 6714 fixations ( $87.6 \%$ ).

## Results and discussion

Fig. 2 shows the effect of launch site on the probability that refixations move forward for each word length-landing site pair. The GLMM showed that nearer launch sites predicted significantly more forward refixations, $\hat{\beta}=0.15, S E=$ 0.03, $\chi_{1}^{2}=13.98, p<0.001,95 \%$ confidence interval (CI) $=[0.10,0.20]$. As reported in the following section, the standard model can accommodate this effect only for landing sites on the right half of the word. To see whether this was also true of the human data, separate analyses were carried out for fixations with landing sites on the left and the right half of the word. For the left half ( 4790 fixations), launch site predicted more forward refixations, $\hat{\beta}=0.16, S E=0.03, \chi_{1}^{2}=$ 10.91, $p<0.001,95 \% C I=[0.09,0.22]$, and the same was true for the right half ( 1362 fixations), $\hat{\beta}=0.14, S E=0.04$, $\chi_{1}^{2}=7.40, p<0.01,95 \% C I=[0.05,0.22]$. These observations that closer launch sites predicted more forward-moving refixations confirm the rational model's predictions. The separate analyses of fixations on the left and right halves of the word indicated that this effect generalized across both.

## Experiment 2: E-Z Reader

This section aims to show that the standard model does not predict the effect of launch site on direction of refixations. To this end, we carry out the same analyses as the previous section on simulation data from E-Z Reader, a computational model of eye movements in reading that incorporates the stan-


Figure 2: Effect of launch site on proportion of forwardmoving refixations on data from Dundee corpus. Each panel contains data from a combination of word length and landing position, and shows a GAM smoother.
dard holistic model of word identification, and always targets refixations to the center of words. In principle, then, all intentional refixations following a fixation on the left half of the word should move forward and those following a fixation on the right half should move backward. Simulations with an implemented version of this model help to ensure that unintentional refixations - saccades intended for another word that happen to become a refixation due to motor error - do not in general change these predictions.

## Methods

Data We used E-Z Reader 10 (Reichle et al., 2009) to generate eye movement data for 100,000 virtual readers reading sentences from the Schilling corpus (Schilling, Rayner, \& Chumbley, 1998) of single English sentences typical of reading experiments. Each virtual reader was a simulation completed using a Monte Carlo run of the model.

The data cleansing procedure was the same as that in Expt. 1. Out of the 20, 189,603 first-pass fixations, $3,417,999$ ( $16.9 \%$ ) of them were followed by a refixation. Excluding extreme launch sites, fixations landing on initial or final letters of a word, and skipping of the previous word left $1,029,801$ fixations. Launch site ranged between $[-15,-1]$.

Statistical analysis A generalized linear model (GLM) with the same fixed effects as that in Expt. 1 was adopted to analyze the effect of launch sites on refixation direction. Random effects were removed from the GLMM used for Expt. 1 since the virtual readers were simply different Monte Carlo runs with no systematic differences. Excluding word length-landing position pairs where all refixations (or all but 1) moved in the same direction left 899,838 fixations ( $87.4 \%$ ).

## Results and discussion

Fig. 3 shows the effect of launch site on the probability for refixations moving forward. The GLM showed that nearer
launch site predicted significantly more forward refixations, $\hat{\beta}=0.08, S E=0.004, \chi_{1}^{2}=386.66, p<0.001,95 \% C I=$ [ $0.07,0.09]$. However, this effect was driven by fixations landing on the right half of the word, $\hat{\beta}=0.10, S E=0.004$, $\chi_{1}^{2}=542.99, p<0.001,95 \% C I=[0.09,0.11]$, while fixations landing on the left half had $99 \%$ refixations moving forward and yielded an opposite effect, $\hat{\beta}=-0.33, S E=0.03$, $\chi_{1}^{2}=147.37, p<0.001,95 \% C I=[-0.39,-0.27]$.

Therefore, E-Z Reader does not in general predict that closer launch sites should lead to refixations being more likely to go forward, contrary to our observations on the human data, although it can accommodate such a prediction for fixations on the right half of the word. Although this effect on the right half of the word may seem surprising, we note that the predictions we described above for this account only hold for intentional refixations. We believe that this effect on refixations on the right half of the word arises from unintentional refixations. Specifically, for a fixation position on the right half of a word, the E-Z Reader model will generally execute one of two behaviors: initiating a saccade to refixate the word or initiating a saccade to move on to the next word. In this case, an intended refixation will target a leftward position (since the center of the word is to the left of fixation) and an intended saccade to the next word will target a rightward position. Which of these two behaviors occurs depends on how quickly the identification (or more technically, $L_{1}$ ) is completed for the current word. Closer launch sites mean that identification of the word will be completed more quickly, which in turn will lead to a greater chance of making a forward saccade intended for the next word. Assuming some of these forward saccades become unintentional forward refixations, this creates exactly the predicted relationship between launch site and refixation direction. For the present purposes, however, the main conclusion here is that the standard model cannot reproduce a general effect of launch site on refixation direction.

## Rational models of reading

In this section, we describe an implemented rational model of refixations, which we will use in the next section to confirm that the intuitively-derived predictions of the rational account for the relationship between launch site and refixations are actually produced by an implemented rational model. Rational models of reading use Bayesian inference to combine visual information with language knowledge (e.g., contextual information). Based on the posterior distribution, eye movements are selected to maximize identification efficiency. The rational model of refixations we describe in this paper also follows this idea, and can be viewed as an application of the more general-purpose rational models of eye movements in reading to the specific situation of refixations. This section introduces the framework of our model.

## Word identification as Bayesian inference

Word identification consists of Bayesian inference, in which a prior distribution over possible identities of the text given by


Figure 3: Effect of launch site on proportion of forwardmoving refixations in data from E-Z Reader simulation. Each panel contains data from a combination of word length and landing position and shows a GAM smoother.
its language model is combined with a likelihood term given by 'noisy' visual input at the position of fixation to form a posterior distribution over the identity of the text given all information sources. Formalized with Bayes' theorem,

$$
\begin{equation*}
p(w \mid \mathcal{I}) \propto p(w) p(\mathcal{I} \mid w) \tag{1}
\end{equation*}
$$

where the probability of the true identity of the word being $w$ given uncertain visual input $\mathcal{I}$ is calculated by multiplying the language model prior $p(w)$ with the likelihood $p(\mathcal{I} \mid w)$ of obtaining this visual input from word $w$, and normalizing.

In general, the prior $p(w)$ represents reader expectations for words conditioned on the context, but for the present paper, we ignore context and use only a word frequency model for simplicity. The visual likelihood is computed similarly to in Bicknell and Levy (2010): each letter is represented as a 26-dimensional vector with a single element being 1 and the rest being 0s. Visual input about each letter is accumulated iteratively over time by sampling from a multivariate Gaussian distribution centered on that letter with a diagonal covariance matrix $\Sigma=\lambda^{-1} I$, where $\lambda$ is the reader's visual acuity for that letter. Visual acuity depends on the location of the letter in relation to the point of fixation, which is a function of the letter's eccentricity $\varepsilon$. In our model, we assumed that acuity is a symmetric, exponential function of eccentricity:

$$
\begin{equation*}
\lambda(\varepsilon)=\int_{\varepsilon-.5}^{\varepsilon+.5} \frac{1}{\sqrt{2 \pi \sigma^{2}}} \exp \left(-\frac{x^{2}}{2 \sigma^{2}}\right) d x \tag{2}
\end{equation*}
$$

with $\sigma=3.075$, the average of two $\sigma$ values for the asymmetric visual acuity function ( $\sigma_{L}=2.41$ for the left visual field, $\sigma_{L}=3.74$ for the right visual field) used in Bicknell and Levy (2010). In order to scale the quality of visual information, we multiply each acuity $\lambda$ by the overall visual input quality $\Lambda$, which is set to 12 in our simulation (see Expt. 3 below).

## Eye movement policy

Based on the posterior distribution on possible identities of the word, eye movement decisions are selected to maximize reading efficiency. For example, the first rational model of reading, Mr. Chips, used this optimizing principle: the model reads input text sequentially, without error, in the minimum number of saccades (Legge et al., 1997, 2002). Specifically, saccades were made to minimize the expected entropy of the current word after the next fixation.

In a more recent rational model of eye movements in reading (Bicknell \& Levy, 2010, 2012), eye movement decisions depend on the uncertainty of the posterior distribution about each letter position. Specifically, given a fixation landing on an unknown character $c$ in position $j$, the marginal probability $m$ of the most likely character under the posterior is

$$
\begin{equation*}
m(j)=\max _{c} p\left(w_{j}=c\right) \tag{3}
\end{equation*}
$$

where $w_{j}$ indicates the character in position $j$. A high value of $m(j)$ indicates relative confidence about the character's identity, and a low value relative uncertainty. The model then decided between four possible actions based on $m(j)$ : continuing to fixate the current landing position, moving backward, moving forward, and ending the reading process.

We use a similar eye movement policy in our refixation model. If the value of the aforementioned statistic $m(j)$ is less than a parameter $\alpha$, the model chooses to continue fixating the current position. Otherwise, if the value of $m(j)$ is less than the parameter $\beta$ for some leftward position, the model initiates a saccade to the closest such position. If no such positions exist to the left, then the model initiates a saccade to the closest position to the right for which $m(j)<\alpha$. Once a refixation is executed, the simulation ends. If all $m(j)$ values to the right (left) are above $\alpha(\beta)$, we decide this word is identified with a satisfactory uncertainty level, and the identification of this word ends. In such a situation, we expect that the eyes move to the next word, which is beyond the current paper's scope of studying refixations.

The actual landing position is the intended fixation position with random motor error: the actual landing position $\ell_{i}$ is sampled from a Gaussian centered on the intended target $t_{i}$ with standard deviation given by a linear function of the intended saccade distance

$$
\begin{equation*}
\ell_{i} \sim \mathcal{N}\left(t_{i},\left(\sigma_{0}+\sigma_{1}\left|t_{i}-\ell_{i-1}\right|\right)^{2}\right) \tag{4}
\end{equation*}
$$

for some linear coefficient $\sigma_{0}$ and $\sigma_{1} .{ }^{2}$ In Expt. 3 in this paper, we follow the SWIFT model in using $\sigma_{0}=0.87, \sigma_{1}=0.084$. A refixation occurs if the actual landing site of the next fixation falls on the same word.

## Experiment 3: Rational model

In this section, we analyze simulated data from our rational model of refixations to verify that it does indeed make the

[^57]prediction that we derived from it intuitively: that refixations would be more likely to move forward for closer launch sites. As described in the previous section, the rational model of refixations we use combines information from previous fixations (including the launch site) to form a posterior distribution on the identity of a word through Bayesian inference. It then makes refixation eye movements to parts of the word about which it is uncertain.

## Methods

Model parameters For the language model component of the word identification model (the prior), we used word frequency information (a unigram model) from the Corpus of Contemporary American English (COCA) (Davies, 2016). For this simulation, we did not optimize the behavior policy parameters to maximize reading efficiency as in Bicknell and Levy (2010), but set them manually to what we surmised might be reasonable values of $\alpha=0.9$ and $\beta=0.7$. Future work will optimize them, but we do not expect the qualitative predictions relevant to this analysis to change.

Data Eye movement data were generated to identify a word. All words were in the most frequent 5,000 words in COCA, and word lengths ranged between [ 3,10$]$. Launch site had a range of $[-10,-1]$. For each word length, each possible landing position, and each launch site, 200 trials were run to model the word identification process as when a fixation landed on that landing position, preceded by a fixation on that launch site. In each trial, a word was randomly selected uniformly from words with the same length.

Procedure Each trial began with a fixation with a duration of 200 time steps on the launch site, in order to represent the visual information obtained prior to fixating the word. Then, the fixation at the landing site began. On each timestep of that fixation, visual information was obtained and integrated with prior information to update the posterior, and then a behavior decision was made: whether to continue fixating, make a refixation, or stop reading (see model description).
Statistical analysis A GLM with the same fixed effects as that in Expt. 2 was adopted to analyze the effect of launch site on refixation direction. Excluding word length-landing position pairs where all refixations (or all but 1) moved in the same direction left 25,636 fixations.

## Results and discussion

Fig. 4 shows the effect of launch site on the probability for refixations moving forward. As expected, the GLM showed that nearer launch site predicted significantly more forward refixations, $\hat{\beta}=0.07, S E=0.005, \chi_{1}^{2}=187.62, p<0.001$, $95 \% C I=[0.06,0.08]$. The same pattern held for both data with landing positions on the left half of the word, $\hat{\beta}=0.04$, $S E=0.008, \chi_{1}^{2}=28.85, p<0.001,95 \% C I=[0.02,0.06]$, and the right half, $\hat{\beta}=0.12, S E=0.009, \chi_{1}^{2}=179.38, p<$ $0.001,95 \% C I=[0.10,0.14]$. These results confirm that an implemented rational model does indeed make this predic-


Figure 4: Effect of launch site on proportion of forwardmoving refixations in data from rational model simulation. Each panel contains data from a combination of word length and landing position and shows a GAM smoother.
tion, which we observed in Expt. 1 to hold of human data.

## General discussion

In this paper, we investigated how visual information is used for word identification during natural reading. We compared two accounts: (1) the standard holistic model, in which visual information about the word as a whole is used in word identification, and processing is always most efficient from the center; and (2) a rational model, in which readers combine information from many sources to identify a word constructively, and the fixation location that maximizes identification efficiency depends on what prior information has been obtained. We suggested that these two models make divergent predictions for the possible effects of launch site on where refixations go. Specifically, only the rational model should predict that refixations are more likely to go rightward for closer launch sites. An analysis of a large human eye movement corpus confirmed that this prediction of the rational account holds in human data. Model simulations confirmed that a rational model does indeed predict it, and that at least one of the implementations of the standard model (E-Z Reader) could not accommodate this finding.

These findings seem strongly inconsistent with models in which all intentional refixations target the center of a word, which in turn suggests that the standard holistic model of word identification in reading may be incorrect. However, it is possible to imagine that other refixation targeting schemes could be used even if the holistic model of word identification in reading is correct. For example, even under the standard model, it might be a useful strategy to target a refixation further forward in a word when that word is closer to being identified. Even if there is an efficiency penalty for being away from the center while that word is finished being identified, that penalty might be outweighed by the benefits of being closer to the next word when the reader's attention
(soon) turns to it.
While it's possible that such eye movement models could be constructed while maintaining the standard model of word identification, our findings are completely consistent with the predictions of rational models of reading, and suggest that these models should be more fully explored. Here, we focused specifically on how visual information already obtained about a word influences where refixations should go, but rational models predict that the interaction of visual and linguistic information is what should ultimately matter. Future work should test these more complex predictions.

## References

Bates, D., Mächler, M., Bolker, B., \& Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48. doi: 10.18637/jss.v067.i01
Bicknell, K., \& Levy, R. (2010). A rational model of eye movement control in reading. In Proceedings of the 48th annual meeting of the Association for Computational Linguistics (pp. 1168-1178).
Bicknell, K., \& Levy, R. (2012). Word predictability and frequency effects in a rational model of reading. In Proceedings of the 34th annual conference of the Cognitive Science Society.
Davies, M. (2016). Corpus of Contemporary American English (COCA). Harvard Dataverse. Retrieved from http://dx.doi.org/10.7910/DVN/AMUDUW
Engbert, R., Nuthmann, A., Richter, E. M., \& Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. Psychological Review, 112(4), 777-813.
Kennedy, A. (2003). The Dundee Corpus [cd-rom]. Psychology Department, University of Dundee.
Kennedy, A., \& Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. Vision Research, 45(2), 153-168.
Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., \& Tjan, B. S. (2002). Mr. chips 2002: New insights from an idealobserver model of reading. Vision Research, 42(18), 2219-2234.
Legge, G. E., Klitz, T. S., \& Tjan, B. S. (1997). Mr. Chips: an ideal-observer model of reading. Psychological Review, 104(3), 524-553.
McConkie, G. W., Kerr, P. W., Reddix, M. D., \& Zola, D. (1988). Eye movement control during reading: I. the location of initial eye fixations on words. Visual Research, 28, 1107-1118.
McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., \& Jacobs, A. M. (1989). Eye movement control during reading: Ii. frequency of refixating a word. Attention, Perception, \& Psychophysics, 46(3), 245-253.
O'Regan, J. K. (1990). Eye movements and reading. Reviews of Oculomotor Research, 4, 395-453.
O'Regan, J. K. (1992). Optimal viewing position in words and the strategy-tactics theory of eye movements in reading. In Eye movements and visual cognition (pp. 333-354). Springer.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372-422.
Rayner, K., Sereno, S. C., \& Raney, G. E. (1996). Eye movement control in reading: a comparison of two types of models. Journal of Experimental Psychology: Human Perception and Performance, 22(5), 1188-1200.
Reichle, E. D., Warren, T., \& McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. Psychonomic Bulletin \& Review, 16(1), 1-21.
Schilling, H. E., Rayner, K., \& Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. Memory \& Cognition, 26(6), 1270-1281.

# A rational analysis of curiosity 

Rachit Dubey (rach0012@berkeley.edu)<br>Department of Education, University of California at Berkeley, CA, USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology, University of California at Berkeley, CA, USA


#### Abstract

We present a rational analysis of curiosity, proposing that people's curiosity is driven by seeking stimuli that maximize their ability to make appropriate responses in the future. This perspective offers a way to unify previous theories of curiosity into a single framework. Experimental results confirm our model's predictions, showing how the relationship between curiosity and confidence can change significantly depending on the nature of the environment.


Keywords: curiosity; rational analysis; computational model

## Introduction

In 1928, upon returning from a vacation, Alexander Fleming, who was a professor of Bacteriology at St. Mary's Hospital in London, noticed how a mold floating in one of his dirty petri dishes held the surrounding bacteria at bay. This peculiar event led him to develop a hypothesis that would be a prelude to the development of penicillin. The history of science abounds with incidents in which an event piqued the curiosity of a scientist thereby leading to important discoveries (other examples include Curie, Faraday, and Planck). For this reason, intellectual curiosity has long been recognized as the essence of science. In fact, Herbert Simon famously titled a 1992 talk given at Carnegie Mellon as 'The cat that curiosity couldn't kill' and described curiosity to be not only the beginning of all science, but also its end (Gobet, 2016).

Considering how important curiosity is to scientific discoveries and many other aspects of cognition, it is surprising that our understanding of curiosity as a psychological phenomenon remains quite limited (Simon, 2001; Gottlieb et al., 2013; Kidd \& Hayden 2015). Encouragingly, the field has seen a revived interest in curiosity in recent years with psychologists and neuroscientists beginning concentrated efforts to study curiosity systematically (Kang et al, 2009; Gruber, Gelman, \& Ranganath, 2013; Law et al., 2016; Walin, O’Grady \& Xu, 2016). However, much previous work on curiosity has either focused on defining a taxonomy for curiosity or providing a mechanistic explanation of curiosity (Kidd \& Hayden, 2015). This means that while we have made some progress describing the psychological processes involved in human curiosity, we have not satisfactorily provided an explanation of the purpose and function of curiosity. For example, a commonly held notion about the function of curiosity is that it motivates learning (Loewenstein, 1994; Kidd \& Hayden, 2015). Although it is easy to say that learning is the goal of curiosity, this is not very precise in its meaning. How does curiosity facilitate learning? Why does it do so?

In light of this, in this paper we present a rational analysis of curiosity in the spirit of Anderson (1990) and Marr
(1982) with the goal of providing a purposive explanation of curiosity. Our work shows that a rational analysis can predict many aspects of curiosity without making assumptions about its mechanisms. We start by defining an abstract representation of the problem that curiosity solves and making a small number of assumptions about the nature of the environment. Following that, we explore the optimal solution to this problem in light of these assumptions. Our theory posits that people are curious about stimuli that maximally increase the usefulness of their current knowledge. Depending on the structure of the environment, the stimuli that maximize this value can either be ones that are completely novel or that are of intermediate complexity. As a consequence, our rational analysis provides a way to unite previously distinct theories of curiosity into a single framework.

The rest of the paper is structured as follows. We first review previous theories of curiosity and then introduce our rational model of curiosity. Following that, we show how our model offers support to previous distinct accounts of curiosity thereby unifying them in a single model. We then conduct a behavioral experiment to test our model's predictions and evaluate how our model accords with human curiosity. We close with a discussion of the implications of our results.

## Models of curiosity

A number of theories have been proposed in the past to describe the psychological processes involving curiosity. In this section, we describe these theories in brief and provide their individual strengths and weaknesses.

Curiosity based on novelty. Several psychological theories have linked curiosity with novelty by hypothesizing that gaining information about novel stimuli is intrinsically rewarding (Berlyne, 1950; Smock \& Holt, 1962). Berlyne (1960) called this "perceptual curiosity" and described it as a driving force that motivates an organism to seek out novel stimuli which diminishes with an increase in exposure. This has also been supported by some neuroscientific studies that show that novel stimuli activate reward-responsive areas in the brain (Ranganath and Rainer, 2003; Düzel et al., 2010). However, a severe limitation of this theory is that it assumes that it is optimal for an individual to explore novel stimuli in all environments. A novel stimulus doesn't necessarily mean that it contains information that is useful or generalizable to an individual. This is also pointed out by previous studies that show that exploration based only on novelty could lead agents to be trapped in unlearnable situations (Gottlieb et al., 2013).

Curiosity based on information-gap. One of the most popular theories of curiosity is the information-gap hypothesis proposed by Loewenstein (1994). According to the information-gap hypothesis, curiosity arises whenever an individual has a gap in information prompting it to complete its knowledge and resolve the uncertainty. Thus, curiosity peaks when one has a small amount of knowledge but it diminishes when one knows too little or too much about the stimuli. A number of studies have supported this prediction and showed that curiosity is an inverted $U$-shaped function of confidence, with people showing the highest curiosity for topics that they were moderately confident about (Kang et al., 2009; Baranes, Oudeyer, \& Gottlieb, 2015). Berlyne (1960) called this form of curiosity "epistemic curiosity" and described it as a drive to acquire knowledge. While this theory has considerable strengths, it is also constrained in that an individual can only be curious about stimuli in known contexts. Thus, if an individual has no prior knowledge about stimuli in the environment then it is not clear how curiosity will function in that environment (as one will not be curious about any stimuli in an environment that it has no prior knowledge of).

Curiosity based on learning progress. A third theory concerning curiosity is guided by the hypothesis that learning progress generates intrinsic reward (Schmidhuber,1991; Schmidhuber, 2010). This hypothesis proposes that the brain is intrinsically motivated to pursue tasks in which one's predictions are always improving. Thus, an individual will not be interested in tasks that are too easy or too difficult to predict but will rather focus on tasks that are learnable. Based on Schmidhuber's framework, a number of papers in developmental robotics have supported this idea showing how an agent can explore in an unknown environment (Oudeyer \& Kaplan, 2006; Oudeyer, Kaplan, \& Hafner, 2007). While this theory can probably describe some forms of curiosity, it is again constrained in explaining curiosity in certain environments. For example, if an agent is ever present in an environment that has many difficult tasks then it is not clear how curiosity will work (as an agent will not be curious about anything within that environment).

Summary and prospectus. While each of these theories have their strengths, we first note that all the above theories are concerned with describing how curiosity functions and how it relates to different psychological factors. However, none of these theories satisfactorily provide an explanation as to why curiosity works the way it does. Second, we believe that all these theories need not be in contention but are all rather special cases of curiosity. As we will describe in the rest of the paper, our rational model supports each of these theories and unifies them in one common model.

## Rational model of curiosity

In this section, we detail our rational model of curiosity. We first consider the abstract computational problem underlying curiosity and then formally derive an optimal solution to this problem.

## Computational problem underlying curiosity

Suppose that an agent is in an environment with $n$ stimuli, each of which provides a reward if the appropriate response is produced. The goal of the agent is to decide what to explore in the environment in order to maximize its knowledge and hence maximize rewards in the future.

The environment determines the probability with which each stimulus occurs in the environment. Let $p_{k}$ denote the "need probability" that a stimulus $k$ will occur in the future (Anderson, 1990). Given this, the agent assigns a confidence value $c_{k}$ to each stimulus in the environment. $c_{k}$ denotes the probability the agent knows the correct response to the $k$ th stimulus. This probability increases at a decreasing rate with respect to the number of exposures $h_{k}$ with that stimulus. $h_{k}$ denotes the number of times the agent has been exposed to the $k$ th stimulus. For convenience, we describe the relationship between $c$ and $h$ by a bounded growth function,

$$
\begin{equation*}
c_{k}=1-e^{-h_{k}} \tag{1}
\end{equation*}
$$

However, our predictions hold for any monotonically increasing function.

Next, the agent computes the value of its overall knowledge. The value, denoted as $V$, is a function of the need probability $p$ and the confidence factor $c$ and is given as follows:

$$
\begin{equation*}
V=\sum_{k} p_{k} \cdot c_{k} \tag{2}
\end{equation*}
$$

According to this equation, the value of an agent's knowledge is simply the chance of successfully responding to the next stimulus computed by summing over all stimuli in the environment.

The goal of the agent is to increase the value of its current knowledge $V$, which it can do so by taking actions to increase $h$ for the various stimuli in the environment. So the computational problem that the agent has to solve is choosing which stimulus to explore further i.e. deciding which stimulus $k$ to increase $h_{k}$ for.

## Deriving an optimal solution

To solve the problem of choosing which stimulus to explore further, the agent can evaluate the change in $V$ as it explores each stimulus in the environment. Thus for every stimulus $k$ in the environment, the agent should compute the change in its knowledge that would result from exploring that stimulus. The stimulus that causes the largest increase in the overall value should be explored first. This computation can be done simply by differentiating $V$ with respect to $h_{k}$,

$$
\begin{equation*}
\frac{\mathrm{d} V}{\mathrm{~d} h_{k}}=p_{k} \cdot \frac{\mathrm{~d} c_{k}}{\mathrm{~d} h_{k}} . \tag{3}
\end{equation*}
$$

An agent operating according to this model will explore stimuli that have a high rate of change of value w.r.t exposure associated with them. This rate of change, i.e. $\frac{\mathrm{d} V}{\mathrm{~d} h_{k}}$, is simply the curiosity the agent has for knowing the $k$ th item, which we


Figure 1: Relationship between need probability $p_{k}$, exposure $h_{k}$, and confidence $c_{k}$ in different environmental structures. Graph 1 shows an environment in which need probability is related to exposure and subsequently confidence. Graph 2 shows an environment in which need probability is independent of exposure and confidence.
denote as $\Omega_{k}$. The agent will explore stimuli in the environment that it is most curious about. In this way, curiosity helps the agent to achieve its goal of maximizing its knowledge.

Under the choice of the form of $c_{k}$ given in Equation 1, we calculate this derivative as follows:

$$
\begin{equation*}
\Omega_{k}=p_{k} \cdot \frac{\mathrm{~d}\left(1-e^{-h_{k}}\right)}{\mathrm{d} h_{k}} . \tag{4}
\end{equation*}
$$

Upon differentiation, we get the relationship of curiosity $\Omega_{k}$, with respect to need probability $p_{k}$ and exposure $h_{k}$ as follows:

$$
\begin{equation*}
\Omega_{k}=p_{k} \cdot e^{-h_{k}} \tag{5}
\end{equation*}
$$

## Relationship to previous models

Having a formal account of curiosity, we now describe how our model relates to previous theories of curiosity.

First, we note that in our rational model framework, two different forms of environmental structure can exist. The first form comes in when the agent is an environment where $p_{k}$ is related to $h_{k}$ (as described in Graph 1, Figure 1). In this environment, stimuli frequently encountered by an agent are more likely to be needed in the future. Thus, the probability that the agent will require a stimulus in the future determines the number of times the agent is exposed to the stimulus which in turn determines the confidence of the agent in knowing that stimulus. The second form comes in when the agent is in an environment where $p_{k}$ and $h_{k}$ are independent of each other (as described in Graph 2, Figure 1). In this environment, the agent can encounter any stimulus in the future regardless of their previous occurrence.

Novelty based curiosity. According to theories that are based on curiosity driven by novelty, an agent is most curious
about stimuli that it is least confident about:

$$
\begin{equation*}
\Omega_{k}=1-c_{k} . \tag{6}
\end{equation*}
$$

According to our rational model, when the agent is in an environment where $p_{k}$ and $h_{k}$ are independent of each other (as described in Graph 2, Figure 1), the relationship between curiosity and exposure will be the one described in Equation 5 where $p_{k}$ is simply a constant value. Thus, curiosity is highest when exposure is lowest and it decreases as exposure increases i.e. curiosity is highest for novel stimuli. The relationship between curiosity and confidence can be rewritten as

$$
\begin{equation*}
\Omega_{k}=p_{k} \cdot\left(1-c_{k}\right) \tag{7}
\end{equation*}
$$

If $p_{k}$ is equal for all $k$, this reduces to Equation 6. Thus, when need probability and exposure are not related to each other, our rational model is similar to the previously proposed novelty based curiosity theory.

Information-gap hypothesis. When the agent is in an environment $p_{k}$ and $h_{k}$ are related to each other (as in Graph 1, Figure 1), then $p_{k}$ is proportional to $h_{k}$ and the relationship between curiosity and exposure given in Equation 5 reduces to

$$
\begin{equation*}
\Omega_{k} \propto h_{k} \cdot e^{-h_{k}} \tag{8}
\end{equation*}
$$

Subsequently, using Equation 1, confidence will be related to curiosity as

$$
\begin{equation*}
\Omega_{k} \propto-\log \left(1-c_{k}\right) \cdot\left(1-c_{k}\right) . \tag{9}
\end{equation*}
$$

Interestingly, this relationship between curiosity and confidence is highly similar to the one described by the information gap hypothesis. Loewenstein used Shannon's (1948) entropy formula to describe the relationship between curiosity and confidence as below:

$$
\begin{equation*}
\Omega_{k}=-\log \left(c_{k}\right) \cdot\left(c_{k}\right) \tag{10}
\end{equation*}
$$

Both the information-gap theory and our model predict that an inverted U-shape relationship exists between curiosity and confidence. In this view, when an agent exists in an environment where need probability is related to exposure, our rational model relates to the information-gap theory.

Relationship to curiosity based on learning progress. According to the learning progress hypothesis, an agent is intrinsically motivated to pursue tasks in which predictions are constantly improving thereby avoiding boring or extremely complicated tasks. An agent operating under this model ends up exploring stimuli of "intermediate complexity".

Our model proposes that an agent will explore stimuli that maximize the value of its current knowledge. In an environment where need probability and exposure are related to each other, then curiosity is highest for stimulus with moderate exposure i.e. intermediate complexity (Equation 8 and 9). Thus,


Figure 2: Relationship between a) curiosity and exposure, and b) curiosity and confidence in an environment where need probability is related to exposure (Graph 1, Figure 1).
in this environment, an agent that aims to maximize its knowledge behaves similarly to an agent whose curiosity is driven by learning progress.

Summary. Whereas previous theories associated curiosity to factors such as novelty, knowledge gap, and learnability, our model shows that depending on the structure of the environment, an agent's curiosity can be driven by any of these factors. In this way, our rational model allows to bridge previous theories related to curiosity in a single framework. In an environment where need probability and exposure are related, our rational model associates with the information gap and learning progress hypothesis. In an environment where need probability and exposure are not related, our model is akin to the novelty-based theory of curiosity.

## Empirical predictions

The rational model presented above makes two different empirical predictions.

Prediction 1. The first prediction arises when the agent is in an environment where the relationship between $p_{k}$ and $h_{k}$ holds true (as in Graph 1, Figure 1). The relationship between curiosity and exposure can be described using Equation 8 and between curiosity and confidence using Equation 9. Thus, Equation 8 predicts that an inverted $U$-shape relationship will exist between curiosity and exposure and Equation 9 similarly predicts that curiosity will be highest when the agent is moderately confident about a stimulus (see Figure 2). We test


Figure 3: Relationship between a) curiosity and exposure, and b) curiosity and confidence in an environment where need probability is independent of exposure (Graph 2, Figure 1).
this prediction in the confidence sampling condition of our behavioral experiment.

Note that this prediction fits the information gap and learning progress hypothesis which also predict an inverted Ushape curve between curiosity and confidence. While several studies have supported the existence of this U-shaped relationship, our model also predicts how to make this effect go away as described in our second prediction.

Prediction 2. Our second prediction comes in when the agent is in an environment where the relationship between $p_{k}$ and $h_{k}$ no longer holds true (as described in Graph 2, Figure 1). Then the relationship between curiosity and exposure will be the one described in Equation 5 and the relationship between curiosity and confidence will be that given in Equation 7. Equation 5 predicts that curiosity is highest when exposure is lowest and it decreases as exposure increases. Similarly, Equation 7 predicts that curiosity will be highest when confidence is the lowest (also shown in Figure 3). We test this prediction in the uniform sampling condition of the behavioral experiment.

While this prediction accords with the prediction of the novelty based hypothesis, that hypothesis can't explain our model's first prediction. On the other hand, while the information gap and learning progress hypothesis were in line with our model's first prediction, both of these theories fail to explain our model's second prediction.

## Testing the model predictions

This section details the behavioral experiment that was conducted in order to test our model predictions. The experiment used two different scenarios - confidence sampling and uniform sampling - to assess whether people's curiosity is affected by changes in the relationship between need probability and confidence. In the confidence sampling condition, we created an environment such that need probability was related to confidence (Graph 1, Figure 1) and in the uniform sampling condition they were independent of each other (Graph 2, Figure 1). Based on our model predictions, we hypothesize that an inverted U-shape relation will exist between confidence and curiosity in the confidence condition and a decreasing relation will exist in uniform sampling condition.

## Participants

We recruited 298 participants from Amazon Mechanical Turk. They earned $\$ 1.50$ for participation with the option of earning an additional bonus of $\$ 0.80$. Participants in the experiment were randomly assigned to one of two conditions: confidence sampling condition (163 participants) and uniform sampling condition (135 participants). Informed consent was obtained using a consent form approved by the institutional review board at Berkeley.

## Stimuli

The stimuli used in the experiment were 40 trivia questions on various topics that were taken directly from Experiment 1 in Kang et al. (2009). According to the authors, these questions
were designed to measure curiosity about semantic knowledge and evoke a range of curiosity levels.

## Procedure

The experiment was divided into two phases - the main round and bonus round. The main round was used to elicit and measure curiosity in participants. Participants were shown 40 trivia questions one after another and were asked to rate their confidence (i.e., probability that they know the correct answer) and curiosity in knowing the correct answer. Curiosity ratings were on a scale from 1 to 7 and the confidence scale ranged from 0 to $100 \%$. Following Kang et al.'s methodology, the raw curiosity ratings were individually normalized and confidence was rescaled to range from 0 to 1 . The order of trivia questions was randomized for each participant. Thus, the main round of the experiment followed the procedure of Kang et al.'s design closely. This part of the experiment took approximately 7-8 minutes to complete.

After the main round, the bonus round began which consisted of two parts. In the first part, all 40 questions from the main round were shown one after another and participants could choose to reveal the answer to those questions. However, each time they chose to reveal an answer, they had to wait an extra 10 seconds for the next question to appear. Findings from Experiment 3 of Kang et al. (2009) showed that participants were more likely to spend time, to wait longer, for the answers that they were more curious about. Thus, requiring participants to spend time to obtain information served as a proxy to measure their curiosity.

In the second part, participants attempted to answer 10 questions that were sampled from the main round ( $\$ 0.08$ bonus for each correct answer). To discourse participants from using Google or other search engines, they were only given 2 minutes in total to answer the questions.

At the beginning of the experiment, participants were randomly assigned to two conditions - the confidence and the uniform condition. Both the conditions had the same main round as described above but used different sampling methods for the bonus round. In the confidence condition, the sampling in bonus round was done based on the confidence ratings provided by the participants i.e. the questions for which participant's confidence rating was higher were more likely to appear in the second part of the bonus round. In the uniform condition, this sampling procedure was completely random i.e. each question was equally likely to appear in second part of the bonus round. Critically, participants were informed about the sampling procedure for their respective condition before the beginning of the bonus round. In a sense, the confidence condition creates a situation in which confidence is related to need probability (Graph 1 in Figure 1) and the uniform condition breaks this relationship (Graph 2 in Figure 1).

According to our model's predictions we should see an inverted U-shape between curiosity and confidence for both the conditions in the main round. However, the curiosity of participants (i.e. the answers they revealed) should be different for both conditions in the bonus round. For the confi-


Figure 4: Relationship of curiosity and confidence in the main round for a) confidence condition and b) uniform condition. The markers indicate mean curiosity at each confidence level and the solid curve is the regression line. Curiosity is an inverted-U function of confidence for both conditions.
dence condition participants' probability of revealing an answer should be highest for questions which they were moderately confident about. On the other hand, in the uniform condition, participants should be most curious about questions for which they were least confident about.

## Results

For all analyses that follow, we removed participants that revealed either too little $(<3)$ or too many answers $(>37)$ in the bonus round. 78 participants were excluded based on this criterion and our final data consisted of 220 participants (118 in the confidence condition and 102 in the uniform condition).

Main Round. Consistent with our prediction, an inverted U-shape exists between curiosity and confidence for both


Figure 5: Probability of participants revealing an answer as a function of confidence in the bonus round. Consistent with our model's prediction, an inverted U-shape exists between curiosity and confidence in the confidence condition and a decreasing relationship exists in the uniform condition.
conditions (Figure 4). Following the method of Kang et al., we fitted curiosity with confidence and uncertainty i.e. confidence $\times$ (1-confidence) for both conditions. For the confidence condition, the model provided $r=0.2$ and a significant coefficient for uncertainty (estimate $=2.01, p<0.001$ ). For the uniform condition, the model provided similar results with $r=0.2$ and significant coefficient for uncertainty (estimate $=2.12, p<0.001$ ). Thus, for both conditions, the model provided a significant quadratic coefficient thereby demonstrating the prevalence of an inverted U-shape between curiosity and confidence for both conditions.

Bonus Round. We first computed the probability of participants revealing an answer conditioned on the confidence rating for both the conditions. As per our model's predictions, an inverted $U$-shape exists for the confidence condition and a decreasing relationship exists for the uniform condition (Figure 5). Similar to the previous analysis, we fitted confidence and uncertainty to both the conditions. For the confidence condition, the model provided $r=0.9$ and a significant coefficient for both confidence and uncertainty (estimates $=-0.15$ and 0.53 respectively with $p<0.05$ for both) thereby showing a U-shape relationship. For the uniform condition, the model provided $r=0.91$ but the coefficient for uncertainty was not significant $(p=0.09)$. On the other hand, the coefficient for confidence was significant (estimate $=-0.23, p<$ 0.001 ), implying a decreasing relationship of curiosity with confidence for the uniform condition.

## Discussion

Curiosity is one of the hallmarks of human intelligence and is crucial to scientific discovery and invention. Models of curiosity have previously explained human curiosity by linking it to various psychological factors such as novelty, information-gap, and learning progress. We have shown that these different models are all special cases of curiosity - depending on the environment, curiosity can be driven by any of these factors. Along with providing a way to unify previous distinct mechanistic accounts of curiosity, our rational model explains human curiosity in various settings.

Our results suggest that human curiosity is not only sensitive to the properties of the stimuli but it is also affected by the nature of the environment. If people are in an environment where need probability influences exposure, then their curiosity is highest for stimuli for which they are moderately confident about. On the other hand, if need probability and exposure are independent of each other then curiosity is highest for novel stimuli, i.e. stimuli for which people have little confidence. This can have important implications in the context of education where researchers are concerned with ways to pique curiosity in students. If we want to make people curious about tasks or activities for which they have little confidence in, perhaps subtle changes in the structure of the environment might be a step towards achieving that. We intend to explore such possibilities in future work, building upon the foundation established in this paper and working towards a
better understanding of how to make people more curious especially in pedagogical settings.

## References

Anderson, J. R. (1990). The adaptive character of thought. Psychology Press.
Baranes, A., Oudeyer, P.-Y., \& Gottlieb, J. (2015). Eye movements reveal epistemic curiosity in human observers. Vision research, 117, 81-90.
Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behaviour1. British Journal of Psychology. General Section, 41(1-2), 68-80.
Berlyne, D. E. (1960). Conflict, arousal, and curiosity.
Düzel, E., Bunzeck, N., Guitart-Masip, M., \& Düzel, S. (2010). Novelty-related motivation of anticipation and exploration by dopamine (nomad): implications for healthy aging. Neuroscience \& Biobehavioral Reviews, 34(5), 660-669.
Gobet, F. (2016). From bounded rationality to expertise. In Minds, models and milieux (pp. 151-166). Springer.
Gottlieb, J., Oudeyer, P.-Y., Lopes, M., \& Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends in cognitive sciences, 17(11), 585593.

Gruber, M. J., Gelman, B. D., \& Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. Neuron, 84(2), 486-496.
Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T. Y., \& Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. Psychological Science, 20(8), 963-973.
Kidd, C., \& Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. Neuron, 88(3), 449-460.
Law, E., Yin, M., Goh, J., Chen, K., Terry, M. A., \& Gajos, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 4098-4110).
Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. Psychological bulletin, 116(1), 75.
Marr, D. (1982). Vision: A computational approach.
Oudeyer, P.-Y., \& Kaplan, F. (2006). Discovering communication. Connection Science, 18(2), 189-206.
Oudeyer, P.-Y., Kaplan, F., \& Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. IEEE transactions on evolutionary computation, 11(2), 265-286.
Ranganath, C., \& Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. Nature Reviews Neuroscience, 4(3), 193-202.
Schmidhuber, J. (1991). Curious model-building control systems. In IEEE International Joint Conference on Neural Networks, 1991. (pp. 1458-1463).
Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). IEEE Transactions on Autonomous Mental Development, 2(3), 230-247.
Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(4), 623-656.
Simon, H. A. (2001). Seek and ye shall find: How curiosity engenders discovery. Designing for science: Implications from everyday, classroom, and professional settings, 5-20.
Smock, C. D., \& Holt, B. G. (1962). Children's reactions to novelty: An experimental study of "curiosity motivation". Child Development, 631-642.
Walin, H., 'Grady, S. O., \& Xu, F. (2016). Curiosity and its influence on children's memory. In Proceedings of the 38th Annual Conference of the Cognitive Science Society.

# Refining the distributional hypothesis: A role for time and context in semantic representation 

Melody Dye (meldye@indiana.edu) ${ }^{1}$, Michael N. Jones (ionesmn@indiana.edu) ${ }^{1}$, Daniel Yarlett (daniel.yarlett@gmail.com) $)^{2}, \mathcal{\&}$ Michael Ramscar (michael.ramscar@uni-tuebingen.de) ${ }^{3}$<br>${ }^{1}$ Department of Psychological \& Brain Sciences, Indiana University, Bloomington<br>${ }^{\mathbf{2}}$ Referral Exchange, San Francisco ${ }^{\mathbf{3}}$ Department of Linguistics, University of Tübingen


#### Abstract

Distributional models of semantics assume that the meaning of a given word is a function of the contexts in which it occurs. In line with this, prior research suggests that a word's semantic representation can be manipulated - pushed toward a target meaning, for example - by situating that word in distributional contexts frequented by the target. Left open to question is the role that order plays in the distributional construction of meaning. Learning occurs in time, and it can produce asymmetric outcomes depending on the order in which information is presented. Discriminative learning models predict that systematically manipulating a word's preceding context should more strongly influence its meaning than should varying what follows. We find support for this hypothesis in three experiments in which we manipulated subjects' contextual experience with novel and marginally familiar words, while varying the locus of manipulation.


Keywords: distributional semantics; vector space models; discriminative learning; word frequency; semantic priming

## Introduction

In the study of human conceptual knowledge, a central theoretical question concerns how semantic representations are learned from the environment. How do speakers acquire knowledge of the meaning of a word and the precise contexts of its use? How are they able to make principled inferences about its senses and its similarity to other words? Inquiries in this domain have focused on two types of converging information sources that are thought to underpin these representations - perceptual and distributional (Andrews, Vigliocco, \& Vinson, 2009; Bruni, Tran, \& Baroni, 2014). Perceptual data derives from experiencing words in relation to the world, in connection with objects, events, and affordances in the immediate physical environment. Distributional data, by contrast, derives from experiencing words in relation to other words. While it is clear that neither data stream alone suffices to explain semantic representation, there appears to be considerable redundancy between them (Louwerse, 2007; Riordan \& Jones, 2010).

Distributional models operate on the assumption that the similarity between two words is a function of the overlap between the contexts in which they occur, a principle commonly known as the distributional hypothesis (Firth, 1957; Miller \& Charles, 1991). For instance, encountering the word violin in the same context as classical and strings supports the inference that these words are semantic neighbors. Such an inference will also be supported for words that occur in closely related musical contexts, such as cello, but not for those that occur in unrelated contexts, such
as jaguar. One of the key advantages of the distributional approach is that it provides an objective and replicable method of quantifying meaning, based solely on the statistical regularities found in large bodies of text.

Since the introduction of Latent Semantic Analysis (LSA; Landauer \& Dumais, 1997) to the cognitive sciences, a variety of different distributional models have been proposed to account for semantic phenomena. Within this class of models, there is considerable variation in implementation (for the latest class, see Baroni, Dinu \& Kruszewksi, 2014). Nevertheless, they share the same core architectural assumption that word meaning is derivable from lexical co-occurrence patterns. Words are represented as vectors within a high-dimensional semantic space, and word meanings as points located within that space. Whereas distributionally similar words tend to cluster together, words that occur in more distinctive contexts are more dispersed. The similarity relations derived from these models can then be used to account for phenomena as diverse as semantic priming (Jones, Kintsch, \& Mewhort, 2006), semantic categorization (Bullinaria \& Levy, 2007), and visual search (Huettig, Quinlan, McDonald, \& Altmann, 2006).

Implicit in these models is the notion that the lexicon is a highly interconnected system. The representation of a given word is neither static nor modular, but changes as a function of lingustic experience, both with that word in particular, and with others within the lexicon. As a demonstration of this principle, McDonald and Ramscar (2001) manipulated readers' semantic representations of marginally familiar and novel words by situating them in paragraph contexts that also contained close associates of a target meaning. For instance, subjects who read about a samovar in paragraph containing words like boil and electric rated it as closer to the meaning of kettle than subjects who read a modified version of the paragraph, which contained associates of an alternative meaning, urn. Even though subjects never directly observed the word kettle in training, their representation of the critical item-samovar-was moved closer to it, simply by virtue of encountering samovar in a similar linguistic context.

## Learning in Time

One question that arises from this, is the extent to which distributional learning about a particular item is influenced by the sequential structure of the context in which it is embedded (Elman, 1990; Jones \& Mewhort, 2007). Language unfolds in time, with one word following another in succession. Thus, the influence that the local context exerts on the critical item might depend on whether it helps predict the occurrence of that item, or is, in turn, predicted
by it - that is, whether the context is encountered before or after the item.

This framing maps naturally onto the the convergent and divergent learning hierarchies described by Osgood (1949). These abstract schemas capture asymmetries in how information is structured in time (Figure 1). In associative learning, convergent hierarchies label a situation in which a variety of cues are associated with a functionally identical outcomes ( $C_{1}, C_{2}, \ldots C_{\mathrm{x}} \Rightarrow O$ ), while divergent hierarchies label the inverse scenario, in which a single cue is associated with varied outcomes $\left(C \Rightarrow O_{1}, O_{2}, \ldots O_{\mathrm{x}}\right)$. Convergent hierarchies have been found to result in greater facilitation and positive transfer in learning, whereas divergent hierarchies yield interference and negative transfer.


Figure 1: Sequential relationships between linguistic regularities. The left side of the figure shows a convergent hierarchy; the right, a divergent one (Ramscar, 2013).

The temporal asymmetries captured by these schemas appear to be ubiquitous in word learning (Ramscar et al., 2010). Consider the problem of learning the relation between a class of things in the world - say, the category [ cat ] - and the word that denotes it - cat. Clearly, a sizable discrepancy exists between the rich array of perceptual features that belong to the class and the comparatively sparse acoustic features of the verbal label. Whereas the flesh and blood exemplars of the category exhibit a wide variety of discriminable features, across various perceptual modalities, the label itself comprises a simple sequence of sounds, which are likely to be perceived categorically (Kuhl, 2000). Accordingly, in a standard category learning paradigm, in which a category exemplar precedes its verbal label, a convergent hierarchy results. However, simply reverse the timing - by placing the label before the exemplar -and the structure becomes divergent.

The terminology used to describe this pair of temporal structures varies by research domain. In the study of categorization, a distinction is commonly drawn between classification, in which subjects predict the class to which an exemplar belongs based on its features, in line with a convergent schema, and inference, in which subjects predict an exemplar's feature values based on its class, in line with a divergent schema (Yamauchi \& Markman, 1988). Likewise, in the study of causal reasoning, predictive reasoning licenses inferences from a variety of possible causes to a shared effect-both rain and sprinklers make grass wet-in line with a convergent schema, whereas diagnostic reasoning licenses inferences from a common cause to its possible effects-rain makes grass wet and
green-in line with a divergent schema (Waldmann \& Holyoak, 1992; Waldmann, 2000).

Learning algorithms can help provide a mechanistic account of how the structure of information in time affects what is learned in these tasks. A critical assumption shared by most models of learning, ranging from classical conditioning to perceptrons, is that learning is scaffolded by the predictions we make about our environments, and powered by the surprise we experience whenever there is a mismatch between expectation and reality. Learning proceeds as a continual process of updating and refining expectations, selectively weighting the most informative predictors to relevant outcomes, while eliminating redundant or potentially misleading cues. When our predictions align with reality, learning asymptotes (Rescorla, 1988).

To examine how convergent and divergent structures affect word learning, Ramscar and colleagues (2010) simulated supervised category learning with the RescorlaWagner rule, while manipulating the sequencing of category exemplars and verbal labels. The findings were striking: The same algorithm, run over the same task, produced remarkably different representations of the learning environment, depending on the temporal sequencing of information: While convergent structures yielded predictive representations, divergent structures yielded veridical ones. Specifically, convergent schemas facilitated competition between the available perceptual features for associative weight, resulting in abstraction of the informative dimensions that best predicted the category label. By contrast, divergent schemas facilitated learning of the actual feature probabilities given the label. (For a closely related result in a different model architecture, see Yamauchi, Love, \& Markman, 2002).

The differences in these representations can be mapped onto the differences between discriminative and generative classifiers in machine learning ( Ng \& Jordan, 2002). In learning a verbal category, the problem is to establish the likelihood of a category label $L$ given some set of perceptual features $F$. To solve this problem, discriminative classifiers learn a direct mapping between features and labels, which yields $\mathrm{p}(L \mid F)$. Generative classifiers solve the same problem indirectly, by learning the joint probability of $\mathrm{p}(L, F)$ and relying on Bayesian inference to calculate the posterior likelihood of $\mathrm{p}(L \mid F)$. While discriminative classifiers are more efficient and better at minimizing error, generative classifiers operate with a more complete picture of the probability space (Levering \& Kurtz, 2014). Convergent schemas yield $\mathrm{p}(L \mid F)$; divergent schemas $\mathrm{p}(L, F)$.

The resultant representations appear to be optimized for different tasks. In studies of human category learning, convergent schemas benefit the learning of categories that require information-integration (Ashby, Maddox, \& Bohil, 2003; Yamauchi et al., 2002), which likely form the majority of natural kinds (Rosch \& Mervis, 1975). However, there are notable drawbacks to categorical responding. As a category structure becomes better learned, stimulus dimensions that are relevant to a particular categorization are selectively attended, such that they acquire distinctiveness, while irrelevant dimensions are
ignored, or down-weighted, maximizing intra-category similarities and inter-category differences (Goldstone, 1994; Lawrence, 1949; Nosofsky, 1986). Accordingly, while convergent schemas support accurate categorization across an array of perceptual domains, they can also systematically alter similarity relations, impairing memory for exemplars seen in training (Davis \& Love, 2010; Dye \& Ramscar, 2009) and distorting judgements of the underlying featural space (Yamauchi \& Markman, 1998). Likewise, in causal inference, whereas predictive reasoning is susceptible to blocking effects, diagnostic reasoning is not (Waldmann \& Holyoak, 1992). The optimal information structure at encoding thus depends on the demands imposed at retrieval (Tulving \& Thomson, 1973).

Previous research has examined the effect of these asymmetric information structures on category learning and causal inference. This paper addresses itself to distributional learning, where what is learned is not the relation between words and physical referents, but rather that of words in relation to each other.

| Study | $\mathbf{1}$ | $\mathbf{2}$ |  |
| :--- | :--- | :--- | :--- |
| Cover Story | Alien <br> Grammar | Man vs. <br> Machine | Semantic <br> Identification |
| Training <br> Design | 10 Train-Test <br> Blocks | 1 Train-Test <br> Block | 1 Train-Test <br> Block |
| Training <br> Length | 8 Associates / <br> Topic | $15 /$ Topic | $15 /$ Topic |
| Critical Item | Pseudoword | LF | LF \& HF |
| Topic <br> Meanings | Random <br> Assignment | Synonyms of <br> Critical Item | Semantic <br> Category |

Table 1: Design differences between studies.

## Studies

In the following three experiments, our aims were first, to build on the original findings of McDonald and Ramscar (2001)-which demonstrated that a pair of words can be moved closer together in semantic space even if they have never been encountered together-and second, to investigate whether readers would attach more weight to the associates that occurred before a word of interest, rather than after, as predicted by previous simulations (Ramscar et al. 2010).

The three studies presented here are all variations on the same principal design. In training, subjects read short passages containing critical words. These passages had been constructed such that the contexts occurring before the critical item were designed to encourage one set of inferences about its meaning, while the contexts occurring after it were designed to encourage a different, competing set of inferences. This design allowed us to measure the relative influence of preceding and succeeding contexts on semantic representation.

Variations on this design were devised to investigate the robustness of the predicted effects of training, and included (e.g.) the choice of cover story, the semantic proximity of
the topic meanings to the critical item and to each other, and the precise organization and length of training and test blocks (Table 1). Detailed descriptions of each experimental design, including counterbalancing and randomization, timing procedures, and lexical controls, are available in the Supporting Materials.

The training phase of each study required a set of critical items, competitor topic meanings, and a set of close lexical associates of each topic. From these materials, a set of triplets was created, each of which consisted of a critical word and two different topic meanings. One of these topic words was designated the preceding topic, and the other, the succeeding topic (Table 2).

| Triplet | topic1 | critical | topic2 |
| :--- | :--- | :--- | :--- |
|  | dream | fugue | music |

Table 2: An example of a training triplet taken from Study 2, in which the critical word fugue has been paired with the competing topics dream and music.

For each topic in a given triplet, corpus data were used to generate a ranked list of its lexical associates. These were used to construct training trigrams, which consisted of the critical item and a pair of its topics' close associates on either side of it (Tables $3 \boldsymbol{\&}$ ). These training trigrams were embedded into larger strings, which subjects were incidentally exposed to in training. The precise number of training trials varied by study.

| Condition 1 |  |  | Condition 2 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| T1 <br> associate1 | critical | T2 <br> associate1 | T2 <br> associate1 | critical | T1 <br> associate1 |
| T1 <br> associate2 | critical | T2 <br> associate2 | T2 <br> associate2 | critical | T1 <br> associate2 |
| $\ldots$ | critical | $\ldots$ | $\ldots$ | critical | $\ldots$ |
| T1 <br> associateN | critical | T2 <br> associate $N$ | T2 <br> associate $N$ | critical | T1 <br> associateN |

Table 3: Abstract representation of the training trigrams for a given critical word and its two topic meanings.

| \{DREAM, critical, MUSIC \} |  | \{MUSIC, critical, DREAM\} |  |  |  |
| :---: | :--- | :---: | :---: | :--- | :---: |
| chasing | fugue | listening | listening | fugue | chasing |
| lucid | fugue | classical | classical | fugue | lucid |
| worthy | fugue | primal | primal | fugue | worthy |

Table 4: Partial training sets in Study 2 for the critical item fugue and its topic synonyms dream and music. In Condition 1 (right), the ordering of associates is reversed from Condition 2 (left).

Post-training, two tests were administered. In the first, a semantic priming task, a prime word was briefly presented on-screen, and subjects were asked to determine whether the following word was a real word in English. Each critical item was tested in combination with its two competitor topics, alternating its position as a prime or target (Table 5).

Subsequently, in a semantic similarity rating task, subjects were asked to rate the similarity of various word pairs on a numerical scale, ranging from "unrelated in meaning" to "identical in meaning". Each critical item was alternately paired with its two topics (Table 6).
$\left.\begin{array}{|l|c|c|c|c|}\hline \text { Semantic }\end{array} \begin{array}{c}\text { topic1 } \rightarrow \\ \text { critical }\end{array} \quad \begin{array}{c}\text { critical } \rightarrow \\ \text { topic1 }\end{array} \quad \begin{array}{c}\text { topic2 } \rightarrow \\ \text { critical }\end{array} \quad \begin{array}{c}\text { critical } \rightarrow \\ \text { topic2 }\end{array}\right]$

Table 5: Example test trials from the semantic priming task.

| Semantic <br> Similarity | topic1 $\mid$ critical | topic2 $\mid$ critical |
| :--- | :---: | :---: |
|  | music $\mid$ fugue | dream $\mid$ fugue |

Table 6: Example test pairs from the semantic rating task.
A key point of difference between studies was the frequency of the critical item: Study 1 employed pseudowords, Study 2, low frequency items, and Study 3, a mix of high (HF) and low frequency (LF) items.

## Hypotheses

Priming Semantic priming is a classic paradigm for studying representation in semantic memory (Neely, 1991). A general finding is that a target item will be processed more efficiently when it is preceded by a semantically related prime, with the degree of facilitation depending on the relatedness of the pair. For instance, bread will be processed more quickly and accurately when it is preceded by butter than when it is preceded by nurse (Meyer \& Schvaneveldt, 1971). When what is studied is the extent to which recently trained associations can facilitate primingas is the case here-the priming is classed as episodic (Hayes \& Bissett, 1998; McKoon \& Ratcliff, 1979). When those associations are indirectly trained, it can be further classed as mediated (Lowe \& McDonald, 2000).

In our studies, a key consideration is that lexical processing is sensitive to temporal contingencies (Deese, 1965). If subjects learn about both the associative and temporal relations between critical items and their topics, then they should be faster and more accurate on lexical decision trials that are consistent with the sequences observed in training. For example, in training sequences in which Topic ${ }_{1} \rightarrow$ Critical $\rightarrow$ Topic $_{2}$, Topic ${ }_{1}$ should be a better prime to the critical item than Topic2, and Topic2 should be better primed by the critical item than Topic ${ }_{1}$.
Similarity Similarity judgments can be affected by the dimensions of alignment that are currently deemed salient to the comparison (Nosofsky, 1986; Tversky, 1977). In the domain of perceptual learning, simulations of convergent and divergent schemas indicate that they develop different feature weights, resulting in correspondingly different representations of the similarity space among exemplars (Ramscar et al., 2010).

If distributional learning is also sensitive to how information is structured in time, then the associative relations the critical item develops with its topics over
training should similarly depend on the positioning of their associates. When multiple lexical associates serve to predict a critical item, the information structure will be convergent; when the critical item serves to predict multiple lexical associates, the structure will be divergent (see Figure 1).

In the convergent case, competition between the lexical associates present in the preceding context should preferentially weight the shared semantic features they have in common with their topic word. By contrast, in the divergent case, weights will be tuned according to cooccurrence rates, which may not select for the most predictive dimensions. Convergent learning should therefore bring the preceding topic into closer alignment in similarity space with the critical item (Dye \& Ramscar, 2009).

## Study 1

Subjects were told that scientists had intercepted an alien communication that they had managed to partially translate, but needed further help in order to fully decode. Participants were presented with a series of these cryptic messages, and instructed to learn as much as they could about the alien word in the middle. That critical item was always a nonsense word.

The experiment was designed such that each subject completed ten short experimental sessions, comprising both training and test, one after the other. This meant that participants learned about each critical item in individual blocks, rather than learning about multiple items simultaneously. The design was fully randomized, such that the specific pairing of topic meanings with a given critical item varied by participant. At the end of the experiment, results were aggregated across all sessions.

Participants Eighteen Stanford University undergraduates participated for course credit.
Results In the semantic priming task, lexical decision accuracy was at ceiling, averaging 98\%. However, differences in response time were apparent. A paired samples t-test revealed that when the critical item served as a prime to one of its topics, subjects were significantly faster at recognizing succeeding topic words than preceding topic words $[t(17)=2.30, p=0.017]$, with a mean 37 ms advantage. However, this advantage was mediated by the prime type: when the topic words themselves served as primes to the critical item, no difference was observed between the preceding and succeeding topics $[t(17)=0.25, p>0.5]$.

After completing the priming task, subjects rated the semantic similarity of each critical item and its competitor topics. A sequential learning account suggests that the preceding topic word should become more similar to the critical item over training. In line with this prediction, subjects rated the preceding topic word as significantly more similar to the critical item than the succeeding topic word $[t(17)=2.27, p=0.018]$. Non-parametric analyses of the data, with the Wilcoxon signed-ranks test, yielded the same pattern of results.

## Study 2

Subjects were told they were taking part in a study testing their ability to distinguish human from artificial intelligence.

On each trial, they were presented with a trigram sequence (Table 4), and asked to judge whether those words had come from a text generated by a human or a computer. In this study, critical items were LF words, whose potential topic meanings were plausible synonyms (e.g., the critical item abscond was matched with the topic words hide and flee). The design was counterbalanced such that the position of each topic word was split evenly across participants. Testing was conducted at the end of the full training session.

Participants 43 undergraduates at Indiana University, Bloomington participated for course credit.

Results The test results of Study 2 replicate the pattern of results in Study 1. In the priming task, lexical decision accuracy averaged $86.4 \%$ overall and $81.8 \%$ for critical items. A dependent samples t-test revealed that when the critical item served as a prime to one of its topics, subjects were faster $[t(42)=-1.73, p=0.046]$ and more accurate $[t(42)=2.45, p=0.009]$ at recognizing topic words that had followed that item, compared to those that had preceded it. A by-items analysis produced a similar pattern for speed $[t(27)=1.53, p=0.068]$ and accuracy $[t(27)=1.85, p=0.038]$. This facilitation pattern was not evident when HF topic words primed LF critical items.

After completing the priming task, subjects rated the semantic similarity of each critical item and its competitor topics. Consistent with Study 1, a dependent samples T-test revealed that preceding topics were rated more similar to critical items, both by subjects $[t(42)=2.99, p=.002]$ and items $[t(13)=2.83, p=0.007]$. Non-parametric analyses, with the Wilcoxon signed-ranks test, confirmed the pattern of results.

## Study 3

Subjects were told they were taking part in a study on reaction time. Words were presented one by one, and subjects were instructed to make a keyboard response every time they saw an item that was either a fruit or a piece of furniture. Training trigrams (Figure 4) were pseudorandomly interspersed throughout this text sequence, with the design counterbalanced such that the position of each topic word was split evenly across participants.

To further assess the extent to which the frequency of the critical item might mediate the predicted effects, both HF and LF critical words were chosen, and each pair of topic meanings was assigned to a pair of unrelated critical items, one in each frequency band (e.g., the critical items jacket and repast were both assigned the same topic pair). Topic meanings were moderately semantically related to each other, but not to either critical item.

As with Study 2, testing was conducted at the end of the full training session.

Participants 26 undergraduates at Indiana University, Bloomington participated for course credit. Two subjects
were dropped from the similarity analyses for selecting the same number for every pair.

Results Study 3 largely replicated the pattern of results in Study 1 and 2. However, in the semantic priming task, the locus of the effect was different: Lexical decision accuracy was at ceiling when HF topic words served as targets ( $98.7 \%$ ). However, when topic words served as primes to the critical items, a 2 (training position) by 2 (critical item frequency) repeated measures ANOVA revealed main effects of item frequency for accuracy $[F(1,25)=16.86$, $p<0.001]$ and RT $[F(1,25)=29.49, p<0.001]$, and of training position for accuracy $[F(1,25)=3.82, p=.061]$ : Subjects were faster and more accurate at recognizing HF targets overall, and more accurate at recognizing critical items that had followed that topic in training, compared to those that had preceded it.

Analysis of the similarity ratings data revealed a main effect of training position $[F(1,22)=5.09, p=.034]$, a main effect of topic frequency $[F(1,22)=10.07, p=.004]$, and a marginally significant interaction between training position and critical item frequency $[F(1,22)=3.88, p=.062]$. Post hoc analyses (Tukey HSD) indicated that, as predicted, LF critical items became more similar to their topic words over training than did HF items. Further, the effect of the training manipulation was mediated by the frequency of the critical item: The preceding topic word was rated as significantly more similar than the succeeding topic word for LF items ( $p<.03$ ), but not for HF items.

## Discussion

Priming Results Speakers appear to be finely attuned to the statistical regularities of their language, allowing them to anticipate upcoming linguistic events based on the current input (Pickering \& Garrod, 2007). This notion is supported by our priming results in Studies 1 and 2, which indicate that when the critical items served as primes, subjects were significantly faster to respond to topic words whose associates had occurred after the critical items in training. This suggests that episodic priming is sensitive not only to temporal contiguity, but also to directionality. ${ }^{1}$

Interestingly, however, when the prime order was reversed, and the topic words served to prime the critical items, the effect disappeared in two of the three studies. The effect thus appears to be mediated by the frequency relationship between primes and targets.

At first blush, the results of Studies 1 and 2 may seem surprising. In semantic priming, a common finding is that while HF targets are responded to more efficiently overall, it is LF targets that typically show greater facilitation from semantically-related HF primes (Becker, 1979)-not HF targets, as in our studies. However, there are important differences between studies that test semantic memory (preexisting semantic associations in long term memory), and those that test episodic memory (associations developed over the course of study), like ours.

[^58]While HF words outperform LF words in semantic tasks, and appear to be more broadly accessible in the lexicon, in episodic paradigms, it is LF words that tend to be better recognized and recalled (Gregg, 1976). This is due, at least in part, to the fact that HF words occur in many more contexts than LF words, making them less associable with any given experimental context (Anderson, 1974; Steyvers \& Malmberg, 2003).

The studies presented here examined the extent to which recently trained semantic and temporal associations facilitate priming. As with other episodic tasks, LF words should develop stronger associations to other experimental items than HF words (the similarity analyses in Study 3 attest to this). The key consideration is that these associations are directional: For a given item, its connections to other words may be distinct from its connections from other words (Nelson \& McEvoy, 2000). It follows sensibly then that in Studies 1 and 2, the LF critical items served as effective cue to the HF topic words, even when the reverse does not obtain (Ramscar et al., 2014).

Similarity Results Across three studies, critical items were rated as more similar to their preceding topics than their succeeding topics, a finding predicted by previous modeling simulations of convergent and divergent learning schemas. As with the priming results, the effect of this training manipulation was modulated by the frequency of the critical item (Study 3).

## General Discussion

Learning is a temporal phenomena, and it can produce asymmetric outcomes depending on how information is structured in time. Such asymmetries have been previously documented in causal reasoning (Waldmann \& Holyoak, 1992) and categorization (Ashby et al., 2002; Ramscar et al., 2010; Yamauchi et al., 2002), and are also attested in sequential learning in non-human animals (Chen et al., 2016). The goal of the present research has been to investigate whether these asymmetric effects might be similarly observable in distributional learning from reading. Across three experiments, our results affirm that they are. An obvious next step is to assess whether models that learn distributed semantic representations of words can replicate these findings (following Jones et al., 2006).

An additional theoretical possibility raised here is that linguistic regularities may play different functional roles depending on whether they participate in convergent or divergent schemas. Suggestive evidence has been offered in artificial language experiments: Whereas stable prefixes and their following nouns are better learned, stable suffixes increase the similarity among those nouns, helping them cohere better as a category (Ramscar, 2013; see also Valian \& Coulson, 1988). Biases toward prefixing or suffixing may thus represent a trade-off between ease of processing and learnability, with suffixes facilitating the discovery of grammatical categories among young learners (St. Clair, Monaghan, \& Ramscar, 2009), and prefixes serving to reduce uncertainty in online comprehension and production (Dye et al., 2017). This proposal is consistent with the finding that in child-directed speech, new words are
preferentially introduced in utterance-final positions (Fernald \& Mazzie, 1991), which appears to promote the best learning outcomes (Fernald, Thorpe, \& Marchman, 2010; Yu \& Smith, 2012). In future research, this framework might be extended to address broader typological questions on the forces at work in language change and evolution.

## Acknowledgments

This research was funded by an NSF graduate fellowship to MD. Many thanks to Colin Allen, Brendan Johns, the Shiffrin lab group, and three anonymous reviewers for comments and discussion.

## References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. Cognitive Psychology, 6(4),
451-474. 451-474.
Andrews, M., Vigliocco, G., \& Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. Psychological Review, 116(3), 463-498.
Ashby, F. G., Maddox, W. T., \& Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. Memory \& Cognition, 30(5), 666-677.
Baroni, M., Dinu, G., \& Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. ACL, 238-247.
Becker, C.A. (1979). Semantic context and word frequency effects in visual word recognition. Journal of
Experimental Psychology. Human Perception Experimental Psychology: Human Perception and Performance, 5(2), 252-259.
Bullinaria, J. A., \& Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. Behavior Research Methods, 39(3), 510-526.
Chen, J., Jansen, N., and Cate, C. (2016). Zebra finches are able to learn affixation-like patterns. Animal Cognition, 19, 65-73.
Conway, C. M., Bauernschmidt, A., Huang, S. S., \& Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. Cognition, 114(3), 356-371.
Davis, T., \& Love, B. C. (2010). Memory for Category Information Is Idealized Through Contrast With Competing Options. Psychological Science, 21(2), 234-242.
Dye, M., Milin, P., Futrell, R., \& Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J.P. Blevins, \& H. Bartos (Eds.) Perspectives on Morphological Organization: Data and Analyses. Brill: Leiden.
Dye, M. \& Ramscar, M. (2009). No representation without taxation: The costs and benefits of learning to conceptualize the environment. Proceedings of the 31st Meeting of the Cognitive Science Society, Amsterdam, NE.
Ellis, N. C. (2006). Language acquisition as rational contingency learning. Applied Linguistics, 27(1), 1-24.
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.
Firth, J.R. (1957). Papers in Linguistics 1934-1951. London: Oxford University Press.
Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), Recall and recognition. London: Wiley.
.
Huettig, F., Quinlan, P. T., McDonald, S. A., \& Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. Acta Psychologica, 121(1), 65-80.
Jones, M.N., \& Mewhort, D.J.K. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 114(1), 1-37.
Jones, M.N., Kintsch, W., \& Mewhort, D.J.K. (2006). High-dimensional semantic space accounts of priming. Journal of Memory \& Language, 55(4), 534-552.
Kuhl, P. K. (2000). A new view of language acquisition. Proceedings of the National Academy of Sciences, 97(22), 11850-11857.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.
Lawrence, D.H. (1949). Acquired distinctiveness of cues: Transfer between discriminations on the basis of familiarity with the stimulus. Journal of Experimental Psychology, 39, 770-784.
Levering, K. R., \& Kurtz, K. J. (2014). Observation versus classification in supervised category learning. Memory \& Cognition, 43(2), 266-282.
Louwerse, M. M. (2008). Embodied relations are encoded in language. Psychonomic Bulletin \& Review, 15(4), 838844.

Lowe, W. \& McDonald, S. (2000). The direct route: Mediated priming in semantic space. Proceedings of the 22 nd cDonal Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.
McDonald, S. \& Ramscar, M. (2001) Testing the distributional hypothesis: The influence of context on judgements of semantic sim
of Edinburgh.
Meyer, D.E., \& Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90(2), 227-234.
Miller, G. A., \& Charles, W. G. (1991). Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1), 1-28.
Neely, J.H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. Basic processes in reading: Visual word recognition, 11, 264-336.
Nelson, D. L., \& McEvoy, C. L. (2000). What is this thing called frequency? Memory \& Cognition, 28(4), 509-522.
Ng, A. Y., \& Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. Advances In Neural Information Processing Systems, 14.
Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-61
Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. Psychological Review, 56, 132-143.
Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. Psihologija, 46(4), 377-396.
Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., \& Baayen, H. (2014). The Myth of Cognitive Decline: Non-Linear Dynamics of Lifelong Learning. Topics in Cognitive Science, 6(1), 5-42.
Ramscar, M., Yarlett, D., Dye, M., Denny, K., \& Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. Cognitive Science, 34(6), 909-957.
Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. The American Psychologist, 43(3), 151$\begin{array}{r}160 . \\ \hline\end{array}$ and distributional models of semantic representation. Topics in Cognitive Science, 3(2), 303-345.
Rosch, E., \& Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7(4), 573-605.
St. Clair, M.C., Monaghan, P., \& Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. Cognitive Science, 33(7), 1317-1329.
Steyvers, M. \& Malmberg, K.J. (2003). The effect of normative context variability on recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(5), 760-766.
Tulving, E., \& Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80(5), 352-373.
Tversky, A. (1977). Features of similarity. Psychological Review, 84(4), 327-352.
Yamauchi, T. \& Markman, A.B. (1998). Category learning by inference and classification. Journal of Memory and
Language, 39, Language, 39, 124-148.
Yamauchi, T., Love, B.C., \& Markman, A. B. (2002). Learning nonlinearly separable categories by inference and
classification. Journal of Experimental Psychology: Learning Memory. classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(3), 585-593.

# Cute Little Puppies and Nice Cold Beers: An Information Theoretic Analysis of Prenominal Adjectives 

Melody Dye (meldye@indiana.edu) ${ }^{1}$, Petar Milin (p.milin@sheffield.ac.uk) ${ }^{2}$, Richard Futrell (futrell@mit.edu) ${ }^{3}$, \& Michael Ramscar (michael.ramscar@uni-tuebingen.de ) ${ }^{4}$<br>${ }^{1}$ Cognitive Science Program, Indiana University, Bloomington ${ }^{2}$ Department of Journalism, University of Sheffield<br>${ }^{3}$ Department of Brain and Cognitive Sciences, MIT ${ }^{4}$ Department of Linguistics, University of Tübingen


#### Abstract

A central goal of typological research is to characterize linguistic features in terms of both their functional role and their fit to social and cognitive systems. One longstanding puzzle concerns why certain languages employ grammatical gender. In an information theoretic analysis of German noun classification, Dye et al. (2017) enumerated a number of important processing advantages gender confers. Yet this raises a further puzzle: If gender systems are so beneficial to processing, what does this mean for languages that make do without them? Here, we compare the communicative function of gender marking in German (a deterministic system) to that of prenominal adjectives in English (a probabilistic one), finding that despite their differences, both systems act to efficiently smooth information over discourse, making nouns more equally predictable in context. We examine why evolutionary pressures may favor one system over another, and discuss the implications for compositional accounts of meaning and Gricean principles of communication.


Keywords: prenominal adjectives; grammatical gender; language comprehension; language production; language evolution; information theory; typology; formal semantics; Gricean conversational maxims

## Introduction

Linguistic typologists work to define similarity and difference across languages, in an effort to establish what invariant 'universal' properties might underpin the fundamental human capacity for language, amidst remarkable diversity (Evans \& Levinson, 2009). This enterprise is complicated by the fact that language is a hybrid system, which is product both of a common biological endowment (shared across languages and peoples) and of a particular ecological niche (specific to a given language).

As languages evolve, they adopt communicative strategies in response to both social and cognitive pressures, strategies which are then refined over generations of cultural transmission (Becker et al., 2009; Tomasello, 2003; Boyd \& Richerson, 2005; Atkinson \& Gray, 2005). In seeking to understand the limits on variation, a typologist has the unenviable task of disentangling biological imperative (Christiansen \& Chater, 2008) from cultural and historical contingencies (Lupyan \& Dale, 2010), such as migrations or language contact.

Information theoretic approaches to language can help clarify this problem, by setting a goalpost that is explicitly functional, rather than biological or cultural (Ramscar \& Baayen, 2013). On this view, language is a communication system like any other, with the same fundamental purpose of
transmitting information. A language's structural features should thus be subject to the selection pressures that govern the design of efficient digital codes (Dahl, 2004). On this read, variation among languages is the result of selective adaptations to variable circumstances, with communicative efficiency the key measure of fitness.

This is not to imply that the solutions that different languages converge on are 'equally' optimal to some prespecified degree. Evolutionary processes achieve localrather than global-optima, and are chained to their particular historical lineage (Simon, 1989). Rather, the idea is to provide an overarching framework in which the host of interacting variables may be arrayed, so as to better understand how the system maintains and restores a functional equilibrium. In particular, it allows us to ask: How are the perturbations in one part of the system balanced by compensating forces in another?

For example, whereas more 'synthetic' languages, like German, rely heavily on morphological devices to convey information, others, like English, leave more to the surrounding context (Lupyan \& Dale, 2010). This mode of typological inquiry can help uncover how languages use different means to nevertheless achieve similar functional ends, and the potential trade-offs-in terms of complexity and efficiency-that these different strategies may incur (Pellegrino, Coupé, \& Marsico, 2011).

## Two Germanic Tongues

One longstanding puzzle for typologists concerns why certain languages employ grammatical gender, which assigns nouns to distinct classes and marks neighboring words for agreement. From a taxonomic standpoint, gender specification can often appear arbitrary, with little obvious correspondence between the semantic properties of a given referent and its noun class (Vigliocco et al., 2005). Historically, gender has thus been viewed as a useless ornament with little apparent rhyme or reason (Maratsos, 1979). In previous work, Dye et al. (2017) offered a possible solution to this puzzle, using an information theoretic lens to clarify the communicative function of noun classification in German.

On their account, grammatical gender marking serves to modulate nominal entropy, making nouns more equally predictable in context. This functionality benefits language processing in multiple ways: 1) by helping speakers avoid the peaks in uncertainty that would otherwise occur over nouns, smoothing entropy over the larger sequence; 2) by reducing competition between nouns that are highly confusable in context; and 3) by facilitating the use of a richer array of lexical items.

These findings raise a further puzzle: If gender systems are so beneficial to processing, languages should tend to maintain or expand them as they evolve. Yet a number of closely related Germanic tongues have followed precisely the opposite trajectory: Swedish, Danish, and Dutch have all consolidated their noun classification systems, while English has dispensed with gender altogether.

Like Modern German, Old English ( $\sim 750-1150$ AD) classified nouns according to three genders (masculine, feminine, and neuter) and all inanimate nouns belonged to one of the three classes (Curzan, 2003). However, in Modern English, aside from a few archaic exceptions, only nouns referring to males and females take gendered pronouns; inanimate nouns are neuter. The gender system in Modern English is thus far simpler than the noun class systems found in Old English and Modern German.

This raises the worrying possibility that English lacks the resources to accomplish the same specificity of expression available in German. However, another possibility, explored here, is that rather than employing a rigid grammatical device, English relies on a more graded, semantically transparent method of entropy reduction: namely, prenominal adjectives.

Like gender markers, adjectives may act to systematically delimit the space of following nouns. For example, massive and moist are likely to have markedly different following distributions. Yet even subtle differences, such as that between great big and very big, could be highly informative in English. To test this proposal, we use tools from information theory to compare gender marking in German (a deterministic system) to prenominal adjective use in English (a probabilistic one).

## Nominal Uncertainty Management

Languages appear to be organized to maintain relatively stable levels of uncertainty across discourse (Genzel \& Charniak, 2002), employing various strategies to make each lexical choice more equally predictable in context, and thereby reducing processing difficulties (Tily et al., 2009; Jaeger, 2010). In information theory, uncertainty is quantified in terms of entropy. Formally, the entropy H over a distribution of lexical items is a measure of the expected value of information ('surprisal') over the full range of items (Shannon, 1948):

$$
H(X)=-\sum_{x \in X} p(x) \log _{2} p(x)
$$

In many languages, like English and German, nouns are the most diverse part of speech. When prior context is ignored, uncertainty should thus be highest at points where a noun occurs. For example, in the following sequence, uncertainty over possible noun continuations (!) will be higher than for possible verb continuations (\#):
I would \# like a ! beer

Unsurprisingly, nouns are among the most common sites for disfluencies, incorrect retrieval, and mishearings (Clark \& Wasow, 1998; Vigliocco, 1997).

Nevertheless, speakers have various resources at their
disposal for making a particular lexical choice more or less predictable in context. One possibility is to rely on the preceding discourse as a form of scaffolding. Noun class is an efficient system for implementing this principle. Consider the German equivalent of (1):
Ich hätte gern ein ! Bier

Grammatical gender markers can significantly ease the lexical access problem by systematically narrowing the set of candidate nouns that follow (Dahan et al., 2000), thereby offloading some of the uncertainty about the upcoming noun onto the determiner.

To evaluate this hypothesis, Dye et al. (2017) examined the entropy of nouns in German, a language with a threeclass gender system. An analysis of the Stuttgart deWaC mega-corpus (Faaß \& Eckart, 2013) revealed that gender markers systematically reduced nominal entropy across all cases. Further, this appeared to benefit lexical diversity: German plurals, which are not gender-marked, showed a reduction in their type/token ratio, suggesting that the presence of a gender marker was catalyzing the use of a wider array of lexical items.

Yet English is not without its own entropy-smoothing resources. Compared to the sparse semantic context provided by (1), the noun beer should be more predictable following the comparatively constraining context provided by (3):

## I would like a nice cold ! beer

This raises an important question: Might prenominal adjectives in English serve a similar function to grammatical gender markers in German?

Suggestive evidence comes from the visual world paradigm, an experimental framework for studying online language processing in which subjects' eye movements over a visual display are monitored as they listen to a concurrent speech stream (Tanenhaus et al., 1995). A common finding is that listeners fixate semantically-related pictures as they become relevant, with patterns of eye movements timelocked to incoming speech. In studies of this kind, prenominal adjectives and gender markers have been shown to play similar functional roles: When French and Spanish speakers encounter a gendered determiner, they rapidly shift their gaze to gender-consistent referents in the display in anticipation of the upcoming noun (Dahan et al., 2000; LewWilliams \& Fernald, 2007). Similarly, when viewing an array of semantically plausible competitors, English speakers interpret prenominal adjectives contrastively, quickly homing in on likely candidates (Sedivy et al., 1999; Fernald, Thorpe, \& Marchman, 2010). Such findings suggest that both prenominal elements serve a predictive, discriminative function.

## Corpus Analysis

To more closely examine this apparent functional similarity between languages, we conducted a comparison of prenominal adjective and determiner usage in written English and German.

## Corpora

Analyses were initially run on manually annotated newswire corpora, and subsequently replicated on larger web-crawled mega-corpora. These corpus types trade off on scale and precision. Due to space constraints, we report one or the other, but not both; in each case, the qualitative nature of the results are the same.

The newswire corpora included the Negra II corpus of German newspapers, (Skut et al. 1997) and the New York Times Gigaword corpus (Graff et al., 2007). The webcrawled WaCky mega-corpus supplied the SdeWaC, a subset of the German section (Baroni, Bernardini, Ferraresi, \& Zanchetta, 2009), comprising more than 850 M word tokens and 1.1 M word types (Faaß \& Eckart, 2013), and the ukWaC, the British English subset, comprising nearly 2 billion word tokens and 3.8 M word types (Ferraresi et al., 2008). It is worth noting that these are collections of written language, which may not reflect the complexities of spoken production (Baayen, Milin, \& Ramscar, 2015).

Additional annotation for fine-grained part-of-speech categories and extraction was carried out with the RFTagger (Schmid \& Laws, 2008) and the Stanford Parser (Klein \& Manning, 2003).

## Determiners

Entropy Reduction In German, grammatical gender serves to subdivide the space of nouns that can legally follow each marker. By markedly reducing nominal entropy, gender facilitates the use of a more diverse-and more informative -set of nouns following gender-marked determiners. Consistent with this thesis, when Dye et al. (2017) compared singular nouns in German (which are marked for gender), with plural nouns (which are not), they found that singular nouns following determiners were significantly more lexically diverse than their plural counterparts.

By comparison, English determiners, which are neither gender nor case-specific, have less potential to be informative about their following nouns. Consider that while the determiner the in English is informative about the type of word that will follow (a noun, most likely), in German, the determiners der, die, das, den, dem, and des convey not only part of speech information, but also delineate the specific set of lexical items that can follow. This suggests that English determiners may not support the same level of lexical diversity available in German.

To examine this possibility, we first compared the conditional entropy of German nouns following articles (which are gender-marked) to that of English nouns following articles (which are not), in the Negra II and NYT Gigaword corpus, respectively. While the average uncertainty following the determiners was similar across languages, German determiners supported much greater entropy reduction than their English equivalent, a result that held across corpus types.

As Figure 1 illustrates, following a definite article, the conditional entropy of English nouns was similar to that of German nouns ( 10.17 vs. 10.55). However, whereas German provided a substantial entropy offset, English provided none at all. In German, removing information
about definite articles-and hence, about noun class-led to a significant increase in entropy (from 10.55 to 11.71 bits). In the simplified model corpus depicted in Figure 1, whereas the baseline entropy difference between marked English and German nouns suggests a usage rate of around $30 \%$ more nouns, the difference between marked and unmarked German nouns is the equivalent of more than $125 \%$ more nouns.

Lexical Diversity This finding suggests that compared to English, German noun usage must be more heterogeneous following determiners (Figure 1). To compare nominal usage across languages, we calculated the type/token ratio of noun lemmas in these contexts in the Negra II and NYT Gigaword corpus, following Dye et al. (2017), and normalizing for corpus size to make the results comparable. Conveniently, type/token ratio is the inverse of average frequency, which means that the greater the diversity of nominal usage, the lower the average frequency. We found that whereas the average frequency of the German noun lemmas in Negra II was 2.12, the average frequency of similar noun lemmas in the English ukWaC sample was 4.93 ( $p<0.001$ ).

These results suggest that noun class allows German speakers to use more 'informative'-and therefore, less frequent and less predictable-nouns after definite articles more often than English speakers do. Or, to put it another way, German speakers appear to use the entropy reduction provided by noun class to choose nouns that are more specific, resulting in greater nominal diversity.


Figure 1: To illustrate the relationship between entropy, probability, and frequency in a corpus of nouns, the $x$-axis above represents the entropy for a given noun as the size of a set of nouns of equal frequency ( $1, y$-axis) increases. As the size of the set of items increases linearly, entropy rises as an exponential function.

## Adjectives

While our results confirm that nominal usage following determiners is more diverse in German than in English, it does not therefore follow that English is lexically impoverished compared to German, or unable to achieve the same degree of specificity. In particular, definite articles are not the only type of word that typically precede nouns-
adjectives are also common prenominally, and may serve a similar function.

To further explore this idea, we compared the adjectivenoun sequences in the ukWaC and the SdeWaC corpora. Both the overall proportion of adjectives ( $t_{p}=1992.336 ; p<$ .0001 ) and the probability of a noun being preceded by an adjective ( $t_{p}=85.088 ; p<.0001$ ) were significantly higher in English than in German. While German nouns are significantly more lexically diverse than their English counterparts, precisely the opposite obtains for adjectives.

These results present us with two different theoretical explanations. One possibility is that English speakers use more adjectives overall to compensate for the use of less varied nouns in communication-i.e., that they make their messages more specific through adjectival syntagmatic choices. Alternatively, it might be that while German speakers use more varied nouns after articles, English speakers use an equally diverse set of nouns, but rely on adjectives-rather than determiners-to facilitate the use of more informative nouns. Fortunately, these accounts make competing predictions, allowing us to distinguish between them empirically.

The first account accords well with the taxonomic assumption that adjectives add semantic detail to nouns, or somehow "modify" their semantic content (Kamp \& Partee, 1995). On this assumption, adjectives should preferentially modify high frequency nouns, which are in greater need of semantic augmentation, over low frequency nouns, which tend to be more specific (Rosch, 1978). For example, dog is less informative than retriever, which is less informative than dachshund; accordingly, dog should be the most frequently modified, and dachshund the least.

However, if prenominal adjectives in English serve a similar role to gendered determiners in German, precisely the opposite prediction should be made regarding frequency. In German, the entropy reduction properties afforded by noun class facilitate the use of more informative (lower frequency) nouns. If, in English, at least some of this functionality is subsumed by prenominal adjectives, then it is low frequency nouns that should be preferentially "modified", not high frequency ones. The relationship between adjectives and noun frequency thus provides an important test case.

In line with the entropy smoothing account, our analysis reveals a negative correlation between a noun's log frequency and its likelihood of being modified ( $r=-0.17, p$ $<0.001$ ). Moreover, our investigation ${ }^{1}$ indicates that in English, adjectives redistribute the relative entropy of nouns, thus serving to balance the degree to which nouns can be predicted in context: More frequent nouns tend to be preceded by adjectives that are (on average) higher frequency and higher entropy ( $e d f=22.06$ : $F=32069$ : $p<$ 0.0001 ). Indeed, a nonlinear interaction between adjective entropy and adjective frequency accounts for fully $94 \%$ of the variance in noun frequency (Figures 3, 4).


Figure 3: Adjective maximum entropy, which provides an upper bound on uncertainty about the preceding adjective, accounted for almost $90 \%$ of the variance in noun frequency (Adjusted $R^{2}=$ 0.891 ), revealing that more frequent nouns are preceded by larger number of different adjectives.


Figure 4: To achieve greater precision, a second interactive model was run, which regressed noun frequency with the tensor product of adjective entropy by adjective average frequency. This interactive model accounted for fully $94 \%$ of the variance in noun frequency (Adjusted $R^{2}=0.941$ ), and achieved better goodness-offit than the max entropy model, as indicated by both the difference in AIC (28603.60), and the Chi-square test of fREML scores $\left(\chi^{2}=\right.$ 14274.832, edf difference $=3.000, \mathrm{p}<0.0001$ ).

## Discussion

In comparing English and German, two closely related Germanic tongues, we found that whereas German nouns are significantly more lexically diverse than their English counterparts, precisely the opposite obtains for adjectives. These results suggest that the difference between German and English does not lie in the 'specificity' of expression, per se, but rather in how specificity is achieved.

German uses gender marking to distinguish between likely lexical competitors, and adjectives to make rarer lexical items more predictable in context. By contrast, in English, which largely lacks gender, adjectives assume both roles. While these findings are compatible with discriminative accounts of language processing (Ramscar et al, 2010), they raise questions about the explanatory adequacy of traditional taxonomic theories.

[^59]
## General Discussion

## To Gender or Not

On an evolutionary scale, languages tend to become more codified over time, as frequently used sequences of words gradually crystallize into more rigid conventions, a process known as grammaticalization (Hopper \& Traugott, 1993). However, at a number of points in its history, English has taken the opposite developmental path.

One such turning point was the invasion and colonization of the British Isles in the 8th and 9th centuries by the Norse, followed by the Norman conquest of England in the 11th century. As a result of the extended interaction between Old English and Norse, much of the information that had been encoded in fixed aspects of the grammar became "optional" - expressed by words rather than fixed grammatical markers. Old English, the language of England at the beginning of this period, looks like Modern German, with relatively complex patterns of inflection for number, gender, and case. However, by the end of this period, Old English had been eclipsed by Middle English, which much more closely resembles the modern tongue: nouns are marked only for number, adjectives are no longer inflected, and demonstratives are reduced in kind (Dawson, 2003). What might explain this trajectory?

Evolutionary Pressures While it is well known that some languages are easier for adult learners to master than others, it is also the case that first languages are acquired at different rates-Russian children, for example, take several years longer than their Turkish neighbors to sort out nominal case marking (Slobin, 2006). However, what is difficult for a child to learn, may not be difficult for an adult, and vice versa; early language acquisition and adult second-language learning are qualitatively different, both in the nature of the task demands, and in the capacities of the learners themselves (Ramscar \& Gitcho, 2007; ThompsonSchill et al., 2009). Likewise, there may be tradeoffs between what is easy to acquire, and what is efficient to process (see Ramscar et al. 2010 on adjective ordering).

In line with this proposal, there is accruing evidence that the structural form of a language is coupled to its population (and history) of adult learners (Johnson \& Newport, 1989; Trudgill, 2002). Support for this comes from a series of indepth analyses of the World Atlas of Language Structures conducted by Lupyan and Dale (2010), who found that languages with "larger speaker populations, greater geographical coverage, and greater degree of contact with other languages" (p. 6) tend to be morphologically simpler, more transparent in their mappings between form and meaning, and more likely to express semantic distinctions through lexical or pragmatic means, rather than encoding them explicitly in the grammar.

On this account, languages strike a balance between early learnability and adult processing that is moderated by their social niche. Thus, while morphologically complex languages provide a rich set of additional cues to scaffold infant learning, this early advantage has significant drawbacks for adult speakers. The same marking
conventions that support young learners, prove nearly impossible for adult learners to master (Johnson \& Newport, 1989), particularly when extrapolating from noisy input (Hudson Kam \& Newport, 2009).

In languages spoken by large populations of adult learners, there is thus both impetus and imperative to simplify the obligatory aspects of the grammar. Moreover, adult speakers are instrumental to how languages evolve-it is skilled language users (not novices) who make and spread innovations (Labov, 1972; Trudgill, 2010), and adult learners readily adapt newly acquired grammars to better meet their communicative needs (Fedzechkina, Jaeger, \& Newport, 2011; Kirby, Cornish, \& Smith, 2010).
From this perspective, the distribution of Modern English can be seen as having developed in response to the selective pressures produced by the conflicting gender systems of Old English and Norse, combined with a large percentage of adults in the population of language learners. These conditions resulted in a shift away from the abstract, grammaticalized entropy management system of Old English gender marking, to the more probabilistic, semantically transparent system based on adjectives found in Modern English. In comparison to German and Old English, Modern English has thus traded efficiency-in communicative terms-for error tolerance, making it more amenable to later learning.

## Adjectives and Overspecification

From a certain perspective, languages with complex inflectional patterns can appear inefficient, in that they obligate the marking of certain distinctions-such as the temporal remoteness of an action or event-that may or may not actually be relevant to the topic at hand (Lupyan \& Dale, 2010). Yet languages with more transparent semantics employ much of the same apparent redundancy: Native English speakers, who are not grammatically obliged to be superfluous, still regularly produce overspecified utterances like "that's a cute little puppy" and "how about a nice cold beer?" (Deutsch \& Pechmann, 1982; Engelhardt et al., 2006) The logic of such productions has proved notoriously difficult to account for: For one, they appear to violate the Gricean Maxim of Quantity, which assumes that speakers provide just enough information to identify a referent, and no more; for another, their combinatorial meaning has defied systematic description (Lahav, 1989), relegated by formal semanticists to the realms of 'context dependence' and 'vagueness' (Kamp \& Partee, 1995).
However, productions like these only appear mysterious if their meanings are assumed to be compositional-i.e., constructed as a function of their syntax and the meanings of their constituent parts (Fodor \& Lepore, 2002; but see Baroni \& Zamparelli, 2010 for a novel approach). Under the alternative model suggested by information theory, utterances are produced so as to iteratively reduce uncertainty, and different languages employ more (or less) conventionalized means of streamlining that process (Baayen \& Ramscar, 2015). While the patterns of adjective use in English are difficult to account for in terms of formal semantics, their communicative function is strikingly clear from an information theoretic perspective.

## Future Directions

One straightforward extension of this work is to the ordering of prenominal adjectives in English. Violations of conventional adjective ordering can make interpretation difficult, as when we compare 'old French red wine crates' with 'red French old wine crates.' Yet adjective ordering cannot be explained by a simple syntactic rule, and while various elaborate semantic hierarchies have been suggested (Table 1), they are not consistent enough to be implemented computationally (Malouf, 2000).

|  | Modification zones for adjectives |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Determiner | Specification | Description | Classification | Head noun |
| The | same | beautiful | italian | actress |
| Her | own | handsome | jazz | musician |
| The | next | big | guardian | dog |
| This | particular | tall | aboriginal | carving |
| The | same | gray | church | tower |

Table 1: A semantic account of adjective order, in which specifying adjectives "single out or quantify the referent", descriptive adjectives "characterize the referent along a variety of semantic parameters", and classifying adjectives "categorize the referent" (Kemmerer et al., 2007: 240).

Ziff (1960) proposed that adjective order was determined according to two closely related heuristics: the adjective's "privilege of occurrence" (i.e., the range of nouns it might modify) and its "definiteness of denotation" (i.e., the extent to which its interpretation depended on the noun being modified). On this account, adjectives that are more privileged and more definite should be slotted closer to the noun. In a related vein, Danks and Glucksberg (1971) argued that adjectives are ordered according to their "discriminative potential", with the most broadly discriminating being placed first. Both of these claims are amenable to further scrutiny in terms of information and prediction.

One possibility is that adjective chains follow the familiar branching structure seen in personal names (Ramscar et al., 2014), with set-size increasing as a function of proximity to the head noun. This would be consistent with the finding that more frequent adjectives tend to precede less frequent ones. However, given that adjectives' appear to smooth entropy, rather than simply reduce it, the precise chaining structure may be closely tied up with the frequency of the noun being 'modified'. This could explain apparent exceptions to this trend (like "witty young lawyer").

In addition to adjective order, similar analyses might help explain the cross-linguistic differences that have been observed in languages with postnominal adjective biases (Percy et al. 2009; see also Lambert \& Paivio, 1956). More ambitious extensions could be made to other parts of speech, such as verbs and adverbs, and for other languages, beyond those studied here.

## Acknowledgments

Many thanks are due to Christian Adam for his heroic feats of data collection in mining the WaCky corpora.

## References

[^60]
# Vanishing the mirror effect: The influence of prior history \& list composition on recognition memory 

Melody Dye (meldye@indiana.edu), Michael N. Jones (mnjones@indiana.edu), \& Richard Shiffrin (shiffrin@indiana.edu)<br>Department of Psychological \& Brain Sciences, Indiana University, Bloomington


#### Abstract

In the study of recognition memory, a mirror effect is commonly observed for word frequency, with low frequency items yielding both a higher hit rate and lower false alarm rate than high frequency items. The finding that LF items consistently outperform HF items in recognition was once termed the "frequency paradox", as LF items are less well represented in memory. However, recognition is known to be influenced both by 'context noise'-the prior contexts in which an item has appeared-and 'item noise'-interference from other items present within the list context. In a typical recognition list, HF items will suffer more interference than LF items. To illustrate this principle, we deliberately manipulated both the contexts in which critical items had been encountered prior to study, and the confusability of targets and distractors. Our results suggest that when noise sources are balanced, the mirror effect disappears.


Keywords: recognition memory; context noise; item noise; prior history; semantic similarity; orthographic similarity; list length; word frequency; mirror effect; differentiation

## Introduction

In a typical episodic memory experiment, subjects are introduced to a new item or list of items within the experimental context, and memory is then tested for that set. In an old-new recognition task, for example, subjects study a list of words, and then are asked to discriminate words seen at study (targets) from non-studied words (foils). What is potentially challenging about the task is that subjects must identify just those items seen at study from all other words encountered in everyday life. In other words, they must discriminate between pre-experimental familiarity with the test items and familiarity that is specific to the task context. Performance at test is assessed by the $d$ ' sensitivity index, common in signal detection, which computes the distance between the means of the hit-rate distribution (the probability of correctly identifying a target) and the false alarm-rate distribution (the probability of misidentifying a lure), normalized by the common standard deviation.

The study of recognition memory has been dominated by global matching models, which are variants on signal detection models. These capture the idea that recognition of a particular item depends not solely on the properties of the item itself, in isolation, but also on other items present in memory (for a review, see Clark \& Gronlund, 1996). When a particular item is tested, the available cues-such as item and context-form a joint probe of memory, which is accessed in parallel. This global search yields a numerical value, which prompts an 'old'response if it exceeds some criterion. The returned value is variously understood as the global familiarity of the test item, the match between the test item and the contents of memory, and the activation strength of memory for that test item. How the value is calculated also depends on the process specified by the model, ranging from the sum of retrieval strengths (Gillund
\& Shiffrin, 1984) to the match between vectors (Murdock, 1982).

A general assumption is that the distribution of familiarity values will have a higher mean for studied items than for unstudied lures. However, interference at retrieval can arise from two sources: item noise (McClelland \& Chappell, 1998; Shiffrin \& Steyvers, 1997) and context noise (Dennis \& Humphreys, 2010). Item noise refers to the probability of a chance match between an item at test and memory traces for other studied items. Context noise refers to the probability of a match between the experimental context and other contexts in which the tested item has occurred.

To study how noise arises in recognition, designs typically manipulate one or more variables of interest, such as the number of items on the list (list length), the number of repetitions or exposure duration of a particular item at study (item strength), and the number of repetitions of the list (list strength). The properties of the items may also be systematically manipulated: For instance, a list might be comprised of an equal proportion of randomly selected high (HF) and low frequency (LF) items (mixed list), or alternately, contain only items selected from one frequency band (pure list).

The study presented here was designed to investigate the extent to which item and context noise affect recognition processes, by systematically manipulating both the prior contexts in which critical items had been encountered (Kinsbourne \& George, 1974; Estes \& Maddox, 1997), and the similarity of items within the list (Hintzman, 1988; Shiffrin, Huber, \& Marinelli, 1995). There are a number of reasons to believe that these manipulations to item and context noise should differentially affect items as a function of their frequency, which we shall now review.

## Word Frequency Effects in Recognition

In studies of recognition memory, a mirror effect is commonly observed in subject performance with regards to item frequency: Compared to HF items, LF items are better discriminated, yielding a higher hit rate (HR) and lower false alarm rate (FAR) (Glanzer \& Adam, 1985; Glanzer et al., 1993). A similar effect is observed in forced-choice recognition paradigms that include (in addition to the usual old-new pairs) old-old and new-new pairs, for which there is no 'correct' answer. Subjects in these studies preferentially choose LF words over HF words for target pairs, and HF words over LF words for foil pairs (Glanzer \& Bowles, 1976).

In assessing word frequency effects (WFE), there are a few wrinkles to consider: For one, the mirror effect is not always perfectly symmetric; the performance gap is typically smaller for hits than false alarms, and there may be differences in criterion as well as sensitivity (Hintzman, Caulton, \& Curran, 1994). For another, recognition performance does not vary monotonically with word frequency. Instead, LF words only appear to benefit when subjects have some familiarity with them (Schulman, 1976; Zechmeister, Curt, \& Sebastian, 1978). Further, when frequency is considered as a continuous variable,

HR follows a U-shape, with the greatest decrements observable in the mid-frequency band (Hemmer \& Criss, 2013).

Clearly, differences in performance on high and low frequency items cannot be reduced to their differential repetition in prior history. Instead, there appear to be multiple, interacting factors at play in producing differences between frequency bands. Some of the key factors include: 1) how well a particular item is differentiated from other items in the lexicon given prior learning history; 2) how discriminable that item is from other items on the present list, given the specific list composition; and 3) the degree to which that item will be associated with the present task context, which should be inversely related to the number of distinct contexts in which it has previously appeared.

## Differentiation over Learning

How do these dimensions differ for high and low frequency items in a standard recognition experiment? A common theoretical assumption is that greater experience with an item over learning leaves it better differentiated in memory-the idea being that repeated exposure acts to increase similarity between the studied item and its memory trace, while decreasing the similarity between its trace and all others (Criss, 2006). This view of repetition falls naturally out of discriminative learning models (Rescorla, 1972; Ramscar et al., 2010), in which cue weights are tuned to produce ever more efficient responding. It is also common to the study of categorization, where it is well known that similarity relations among items change in systematic ways as a function of learning (Nosofsky, 1986).

Models of recognition memory formalize this notion in slightly different ways. In the Retrieving Effectively from Memory (REM) model, each time an item is encountered within a given context, its episodic memory trace is updated, accruing more complete and accurate feature information (Shiffrin \& Steyvers, 1997). Thus, more encoding opportunities lead to a higher probability of self-match and a lower probability of matching an unrelated item, leading to 'differentiation' of the trace. Likewise, in the subjective likelihood model (SLiM), initial experience with an item yields a noisy and underspecified representation of its features, which is refined over learning (McClelland \& Chappell, 1998). A logical inference from these models is that HF items-by virtue of having been experienced more often, and in more contextsshould be better differentiated from one another in long-term memory than are LF items.

If HF items are better learned, why do they not routinely outperform LF items in recognition, as they do in other memory paradigms, like lexical decision and recall? This is known as the frequency paradox (Gregg, 1976).

## Context Noise

To address this question, it helps to consider how memory for a word depends on the contexts in which it has been previously encountered. Events stored in memory are comprised of both information that was central to processing (the item itself) and information that was available in the peripheral environment (the broader context). The contextual information that is encoded may include aspects of the temporal or physical context in which an item is presented, the emotional state of the learner, and so on (Murnane, Phelps, \& Malmberg, 1999; Smith, Glenberg, \& Bjork, 1978).

Because contextual information is stored alongside item information, memory for an item is facilitated when there is a high degree of match between its encoding and retrieval
contexts, a principle known as encoding specificity (Tulving \& Thomson, 1973). However, similarity between contexts can also produce interference in tasks, like recognition, that require discrimination among encoding contexts. In making an accurate recognition judgment, one of the key challenges is in distinguishing between familiarity with the item from the study list and familiarity from previous experiences in everyday life. The more prior contexts in which an item has occurred, and the more confusable those contexts with the study list, the harder the problem.

One way to demonstrate this is by incidentally exposing subjects to critical targets and lures in a familiarization phase prior to study, which shares many contextual features with the recognition task (e.g., the location, time of day, etc). Recognition for pre-exposed items is reliably impaired (Kinsbourne \& George, 1974; Tulving \& Kroll, 1995). Another method is to select list items that vary in their contextual diversity (CD)-i.e., the number of different pre-experimental contexts in which they have appeared. When CD varies, items with higher diversity scores are less well recognized overall, with a lower HR and higher FAR (Jones, Johns, \& Reccia, 2012; Steyvers \& Malmberg, 2003).

These findings establish context noise as an important source of interference at retrieval. Importantly, context noise is also a key dimension on which HF and LF items differ. Not only are HF words experienced more often than LF words, they are experienced in a more variable set of verbal contexts (Adelman et al., 2006; Jones, Johns, \& Reccia, 2012). Given their high frequency of occurrence in text and speech, they are also more likely to have been experienced more recently (Scarborough, Cortese, \& Scarborough 1977; Anderson \& Schooler, 1991). Relative to LF items, the contexts in which HF items are experienced will thus be more confusable with the study list, significantly increasing the difficulty of the recognition task for those items.

## Item Noise

Another clue to the "frequency paradox" concerns how memory for a single item depends on the composition of the surrounding list. As von Restorff (1933) demonstrated in a classic experiment, distinct items fare well on tests of recognition. For example, in a 10 item-list comprised of 9 nonsense syllables and 1 number, the number is recalled with far greater accuracy than the syllables. However, the extent to which a particular item benefits from its distinctiveness-i.e., its dissimilarity from other items-depends crucially on how dissimilar the rest of the items on the list are from each other.

To illustrate this idea, von Restorff placed the lone number on a list with several equally unrelated items, including "a syllable, a color patch, a single letter, a word, a photograph, a symbol, an actual button, a punctuation mark, and the name of a chemical compound" (as reported by Hunt, 1995, p. 109). Unsurprisingly, once all the items were similarly distinct, no advantage for the lone number was found. A benefit only obtained when the other items were clustered in similarity space relative to the critical item. That is, "similarity must establish a context in which difference functions" (Hunt, 1995).

The distinctiveness hypothesis proposes that memory for a given item should vary inversely with its featural overlap with other items at study (Hunt \& Mitchell, 1982). In line with this, when subjects are asked to remember a list of statements that are either congruent or incongruent with their expectations, incongruent-facts tend to be advantaged in recall-but only so long as they comprise a minority of the list (Hastie \& Kumar, 1979). Parallel results have been reported for word recall, where it has been found that orthographically or semantically
unusual items only benefit when presented with common ones (Hunt \& Eliott, 1980; see also Zechmeister, 1972). This is consistent with a surprisal-based account of the von Restorff effect (Green, 1956).

In recognition, it is clear that distinctiveness matters both at encoding and at retrieval. Researchers as far back as Postman (1951) have observed that performance on tests of recognition memory varies inversely with the similarity of items at study and at test, and thus, with the choice of distractor (Anisfield \& Knapp, 1968; Bahrick, Clark, \& Bahrick, 1967). For example, in face recognition, distinctive faces are better recognized than common faces when lures are selected at random, but recognized more poorly when the similarity of lures to targets is controlled (Davidenko \& Ramscar, 2005).

Importantly, distinctiveness is a feature on which low and high frequency items are bound to vary. LF words are, on average, more orthographically distinctive than HF wordscomprised of more rare letters, and more uncommon combinations of letters (Estes \& Maddox, 2002; Malmberg et al. 2002)—and belong to much sparser orthographic neighborhoods, with both fewer and rarer neighbors (Landauer \& Streeter, 1973).

In a random selection of words, LF items will also be more semantically distinctive than their HF counterparts. This is guaranteed by the distributional properties of the lexiconspecifically the fact that LF words are drawn from a much larger sample than their HF counterparts (using a 1 per million word cutoff, $80 \%$ of all words can be classified as low frequency; van Heuven et al. 2014). As a result, LF items will be less semantically similar to one another, on average, than HF items. A variety of measures of semantic richness attest to this: LF words have fewer closer associates (Deese, 1960; Balota et al., 2004), fewer close semantic neighbors (Pexman et al., 2008), and more sparse network connectivity (Steyvers \& Tenenbaum, 2005).

While HF items are better differentiated in memory, they are also drawn from a much more tightly clustered similarity space, both in terms of their surface and semantic features. When presented in a mixed lists of randomly selected items, they should thus be less distinctive at encoding and more confusable at test. ${ }^{1}$

## Study

In standard recognition experiments, there is a significant imbalance between frequency bands. When item selection is random, LF words should tend to be more orthographically and semantically distinctive than HF words, suggesting that they demand more attentional resources at encoding, and are less confusable with frequency-matched distractors at test. At the same time, their occurrence as a list item is less confusable with other, previous occurrences: LF items have been experienced in fewer, less diverse contexts, and are less likely to have been experienced recently.

In this study, our goal is to bring the sources of noise for high and low frequency items more in balance. To accomplish this re-balancing act, we manipulated two key variables: (1) recency of exposure ('context noise') and (2) inter-item similarity ('item noise'). When these noise sources are equalized, HF items, which are better represented in memory, should outperform LF items.

Design In the familiarization phase of the study, subjects completed a simple reading comprehension test in which they were incidentally exposed to a set of critical words. Following a short delay, subjects returned to complete a list recognition task in which they studied a list of words, and at test, were asked to distinguish between studied items (targets) and novel items (foils).

Context noise was manipulated by inserting previously encountered critical words at study and at test. To assess how recent exposure affected recognition, the study counterbalanced both whether a given word was encountered in reading, and whether it occurred as a target or foil. Item noise was manipulated by selecting control words for the recognition task from dense semantic categories (Figure 1). To assess for frequency effects, both critical and control words were evenly divided between high and low frequency bands.


Figure 1: The average semantic similarity of targets to distractors in lists randomly generated from the Exp. 1 control items, as compared to standard episodic word pools (see Dye et al., 2017 for methodology). Drawing items from semantic categories disproportionately increases similarity for LF items.

Participants 54 undergraduate students at Indiana University participated in the experiment for course credit. All were native American English speakers with normal or corrected-to-normal vision. 3 subjects were excluded from the analysis for performing at chance on the reading comprehension portion of the experiment.

Materials Two word lists were constructed (see Appendix), each of which comprised 40 critical words: 20 HF (165 occurrences/million) and 20 LF (1 occurrence/million), frequency matched across lists, using counts drawn from the Corpus of Contemporary American English (COCA: Davies, 2010).

In addition, an inventory of 240 control words was created, drawn from sixteen semantic categories (such as 'music' and 'time'). Half of these semantic categories were comprised of HF items, and half LF items. These control items were included to assess how item noise affects recognition. Introducing

[^61]semantic categories should disproportionately amplify item noise for LF items, by increasing the semantic and orthographic confusability of targets and distractors (Figure 1).

To create reading materials for the comprehension task, short passages were excerpted from the collected works of the notable Columbian author, Gabriel Garcia Marquez. Specifically, for each word on each of the lists, a passage containing that word was identified and paired with a true statement that synthesized the sentence in which the word had occurred. Affirming the statement as true relied on correct comprehension of the word. Additionally, 20 control passages, which contained no critical items, were selected and paired with a false statement. Each critical word appeared in only one of all possible paragraphs, and only once in that paragraph.

To gauge how pre-exposure affected recognition accuracy and response time, four counterbalanced conditions were created, such that across subjects, each critical item was presented as both a target and as a foil, and was either preexposed (encountered once previously in reading) or novel (occurring for the first time in the recognition task).

Study lists comprised 40 critical items and 120 control, and test lists comprised all 160 targets and an additional 160 foils, with the same 1:3 distribution between critical and control items. Here again, controls were evenly split between high and low frequency items, drawn from the same part of the frequency distribution as the critical words.

Procedure In the first stage of the experiment, subjects completed a self-paced reading comprehension task in which they read a series of short passages and, following each paragraph, were presented with a short statement and asked to determine whether it was true or false. Subjects then moved to a different experiment room to complete a 20 -minute distractor task, in which they solved a series of tangram puzzles. They then returned to the original room to complete the list recognition task.

At study, 160 words were presented on a computer monitor for 1 s each, separated by a 100 ms ISI. At test, subjects were presented with a new set of items, and asked to judge whether a given item had been presented at study. Testing consisted in 320 self-paced recognition trials, with up to 5 s to respond. Order of presentation for passages and for list items was randomized.

Results Looking first to the control items, which were drawn from tightly clustered semantic categories, but did not vary in their exposure history: Welch two-sample $t$-tests confirmed that -consistent with the typical finding-LF targets had a significantly higher HR than HF targets, both by items $[t(212.63)=-4.12, p<.0001]$ and by subjects $[t(99.34)=-3.20$, $p=0.002$ ]. However, the FAR for LF and HF foils was not significantly different ( $p>.5$ ), and the speed of correct rejections was slower for LF foils, both by items $[t(234.24)=-5.85, p<.0001]$ and (marginally) by subjects $[t(97.82)=-1.70, p<.0 .092]$.

Performance on control items thus shows a marked departure from the standard mirror effect: The typical FAR advantage for LF items disappears, and LF foils are more slowly rejected than HF foils (Figure 2). The trends captured here are robust over the course of testing (see Figure 6 for contrast). This finding is consistent with the notion that the introduction of semantic categories differentially increases item noise for low frequency items, diminishing the typical LF advantage. However, LF control items still outperformed HF control items overall-the increase in FAR was balanced by the sustained HR advantage.


Figure 2: Control item performance for correct RT (right panel) and $p$ (old) (left panel), shown by frequency and trial type. Error bars are SEM.

An identical pattern is observable for the critical items with no prior exposure (Figures 3, 4). However, for these items, the mirror effect disappears completely following exposure at reading, and overall performance for HF and LF items draws even (Figure 3). This is because while $p$ (old) increases overall, the LF FAR increases sharply, far outstripping that of the HF foils (Figure 4).


Figure 3. The effect of prior reading exposure on critical items, as measured by $d^{\prime}$ (using a $1 / 2 \mathrm{~N}$ correction).


Figure 4: The effect of prior reading exposure on $p$ (old), graphed by frequency and trial type. Error bars are SEM.

Performance on critical items (Figures 4, 5) can be broken down as follows: For targets, there was a main effect of item frequency on accuracy $[F(1,50)=28.42, p<0.001]$, and a main effect of exposure condition both on accuracy $[F(1,50)=3.35$, $p=.073$ ] and correct RT $[F(1,50)=14.36, p<.0005]$. Subjects were more likely to affirm LF targets overall, and to more quickly and (marginally more) accurately affirm targets that had previously been seen in reading.

For foils, the picture was somewhat more complicated, but no less consistent. For response time, there was a main effect of
item frequency on correct RT $[F(1,50)=5.87, p<0.02]$, but no effect of prior context. For accuracy, there were main effects of item frequency $[F(1,50)=3.98, p=0.052]$ and prior context $[F(1,50)=14.82, p<.001]$, modulated by a significant interaction between frequency and context $[F(1,50)=4.66, p<.05]$. Post hoc analyses (Tukey HSD) indicated that previous exposure significantly increased the FAR for LF items ( $p<0.001$ ) but not HF items ( $p>.5$ ), and that the FAR for pre-exposed LF items was significantly higher than for pre-exposed HF items ( $p<0.005$ ).


Figure 5: The effect of prior reading exposure on response latency, graphed by frequency and trial type. Error bars are SEM.


Figure 6: $P($ old $)$ to critical LF items as a function of test position for hits (top) and false alarms (bottom). Trend lines are generated by the glm smoothing method in ggplot2.

To summarize: For HF items, the primary effect of prior exposure was to increase the speed and accuracy of hits. This effect was also observable for LF targets, and the time to execute a hit was similar for HF and LF targets. However, with LF items, the recency manipulation led to an overall bias in $p$ (old), such that both the HR and the FAR were significantly higher than that of HF items. The dramatic increase in FAR, as a result of item and context noise, is mirrored by the finding that correct rejections were significantly slower for LF foils across both exposure conditions.

In this study, the magnitude of the performance drop for LF items is, in part, a function of testing (Annis, Malmberg, Criss, \& Shiffrin, 2013). At the beginning of testing, no effect of prior history was apparent: the HR for exposed and unexposed critical items was identical, as was the FAR. However, while
the HR for LF items uniformly declined over trials, the pattern of false alarms diverged depending on prior exposure (Figure 6): Whereas for unexposed items, the FAR showed a steady downward trend, for previously encountered items, precisely the opposite was true. This suggests that the ability to discriminate prior context decreased with continued testing. By contrast: For HF items, while a similar decline in HR is observable over testing, the FAR remains constant, and exposure condition does not appear to interact with these trends.

## General Discussion

This paper investigates the sources and robustness of the mirror effect for normative word frequency, finding that under the right set of experimental conditions, it disappears. In particular, when noise sources for high and low frequency items are balanced, LF items prove to be more confusable than better-learned HF items.

## Word Frequency Effects

The aim of the present study was to examine how item and context noise interact with word frequency effects. Item noise was manipulated by selecting control items from a small set of semantically cohesive categories, such as 'music' and 'cooking' (Shiffrin, Huber, \& Marinelli, 1995). Context noise was manipulated by incidentally exposing subjects to critical items prior to the recognition task (Kinsbourne \& George, 1974; Tulving \& Kroll, 1995). Both noise sources have been found to impair recognition in a similar fashion: While these manipulations lead to an overall increase in the probability of responding 'old', the increase in hits is slower than the concomitant increase in false alarms, leading to a general decline in discriminability. For instance, when categories of items are present within a recognition list, hits and false alarms increase monotonically with the number of items within each semantic category, such that discriminability decreases as a function of category size (Hintzman, 1988). Similarly, when items are incidentally exposed prior to study, confusability increases as a function of the number of prior exposures (Criss \& Shiffrin, 2004; Chalmers \& Humphreys, 1998), and as the delay between the familiarization and recognition phases decreases (Maddox \& Estes, 1997).

While previous research has tended to focus on how noise affects items from within a single frequency band, our experiment assessed how items were differentially affected as a function of their frequency. A close analysis of the mirror effect for recognition suggests that it derives from the distinctiveness of LF items relative to their HF counterparts. Specifically -in a random selection of items, LF targets will be more distinctive at study, and more distinctive at test compared to foils; in addition, the contexts in which they have previously occurred will be less confusable with the present study context.

In our study, these advantages are systematically mitigated. Introducing verbal categories entails that items will be sampled from a dense semantic space, rather than randomly from the lexicon at large. This selects for LF items that are more similar to each other than HF items, rendering them more confusable at test. Likewise, pre-exposing critical items guarantees that all such items, regardless of frequency, will have recently been experienced in a highly similar, confusable context. If the usual LF FAR advantage is mediated, at least in part, by the greater distinctiveness of randomly selected LF targets relative to potential lures, and by the greater distinctiveness of their prior contexts of occurrence, then these manipulations should diminish or reverse that advantage.

Our findings comport well with this proposal. The item noise manipulation disappeared the LF FAR advantage both for control items and for critical items with no prior exposures: LF foils attracted a similar number of false alarms as HF foils and were rejected significantly more slowly. (A similar result has been reported when orthographic similarity among items is controlled, and lures are orthographically matched to targets; Hall, 1979; Malmberg, Holden, \& Shiffrin, 2004).

The context noise manipulation amplified this effect, fully reversing the FAR advantage in favor of pre-exposed HF items, a trend that intensified over the course of testing. This occurred because while pre-exposure dramatically increased the LF FAR, it had a negligible effect on HF items. These manipulations thus vanished one half of the standard mirror effect, equalizing the overall discriminability of HF and LF items.

Nevertheless, LF items maintained a strong HR advantage. There are a number of possible theoretical explanations for this result: LF items may have garnered more attentional resources at study (Malmberg \& Nelson, 2003), been more easily associable with the present task context (Hirshman, Whelley, \& Palij 1989), or simply been a better match to their own memory traces at retrieval. All these explanations are potentially consistent with the results of the present study, but beyond its scope to establish; further experimental work is needed to distinguish among these accounts.

## Modeling Accounts

Empirical results like those presented here can provide important constraints on representational assumptions in modeling (Criss \& Shiffrin, 2004). For example, to account for the standard mirror effect, the REM model assumes that LF items possess more rare features than HF items, features that are more diagnostic (Steyvers \& Shiffrin, 1997). This implies 1) that LF targets will be a better match to their own memory traces than HF targets (resulting in a higher HR), and 2) that LF foils will be less likely to share features in common with targets than HF foils, resulting in less spurious matches (resulting in a lower FAR). REM thus correctly predicts that when targets are matched with highly similar foils, the FAR for LF items should substantially increase, diminishing or reversing the standard mirror effect. REM can also be modified to account for the finding that increasing the proportion of HF words on a list decreases the FAR for LF items, by assuming that the distinctiveness of the LF items at study leads to better encoding (Malmberg \& Murnane, 2002).

Likewise, virtually all models of recognition memory incorporate the idea that an item's prior contexts of occurrence are a critical source of interference (Dennis \& Humphreys, 2000). A common assumption is that both item and context information are stored at encoding and that similarity between the study context and prior experiences gives rise to interference at retrieval. What varies is how: In some models, the item and context on the current trial form a joint probe of memory (Gillund \& Shiffrin, 1984). In others, the context cue first acts to restrict the subset of activated memory traces to those that match the current context, prior to comparing the item cue to the resulting set (Shiffrin \& Steyvers, 1997).

In future work, it may be profitable to use item representations derived directly from the items themselves, by quantifying the lexical and semantic characteristics of a given list or word pool (Dye et al., 2017). Models can then be constructed and tested against the true properties of the stimulus set, permitting cleaner adjudication between competing accounts.

## Acknowledgments

This research was funded by an NSF graduate fellowship to MD. Many thanks to Michael Ramscar, Brendan Johns, Gregory Cox, and Rui Cao for insightful comments and discussion.

## References

Adelman, J. S., Brown, G. D. A., \& Quesada, J. F. (2006). Contextual diversity, not word frequency, determines wordnaming and lexical decision times. Psychological Science, 17(9), 814-823.
Anderson, J.R. \& Bower, G.H. (1972). Recognition and retrieval processes in free recall. Psychological Review, 79(2), Anderson, J.R. \& Bower, G.H. (1972). Recognition and retrieval processes in free recall. Psychological Review, 79(2),
$97-123$.
Anderson, J. R., \& Schooler, L. J. (1991). Reflections of the Environment in Memory. Psychological Science, 2(6), 396-
Anisfeld, M., \& Knapp, M. (1968). Association, synonymity, and directionality in false recognition. Journal of
Experimental Psycholog,
Experimental Psychology, 77(2), 171-179.
nnis, J., Malmberg, K. J., Criss, A. H., \& Shiffrin, R. M. (2013). Sources of interference in recognition testing. Journal
Annis, J., Malmberg, K. J., Criss, A. H., \& Shiffrin, R. M. (2013). Sources of interference in recognition testing. Journal
of Experimental Psychology: Learning, Memory, and Cognition, 39(5), $1365-1376$
Bahrick, H. P., Clark, S., \& Bahrick, P. (1967). Generalization gradients as indicants of learning and retention of a
recognition task. Journal of Experimental Psychology, 75(4), 464-471.
Balota, D. A.. Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., \& Yap, M. (2004). Visual Word Recognition of
Single-Syllable Words. Journal of Experimental Psycholog, 133(2), 283-316. Single-Syllable Words. Journal of Experimental Psychology, 133(2), 283-316
Chalmers, K. A., \& Humphreys, M. S. (1998). Role of generalized and episode specific memories in the word frequency
effect in recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(3), 610-632.
Clark, S. E., \& Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the
data. Psychonomic Bulletin \& Review, 3(1), 37-60.
Criss, A. H. (2006). The consequences of differentiation
Criss, A. H. (2006). The consequences
Criss, A. H., \& Shiffrin, R. M. (2004). Context Noise and Item Noise Jointly Determine Recognition Memory: A
Comment on Dennis and Humphreys (2001). Psychological Review, 111(3), $800-807$.
Davidenko, N. \& Ramscar, M. (2005) Distinctiveness effects in face memory vanish with well-controlled distractors.
Proceedings of the 27 th Annual Conference of the Cognitive Science Society, Mahwah, NJ.
Proceedings of the 27th Annual Conference of the Cognitive Science Society, Mahwah, N.
Deese, J. (1960). Frequency Of Usage And Number Of Words In Free Recall: The Role Of Association. Psychological
Deese, J. (1960). Frequency Of Usage And Number Of Words In Free Recall: The Role Of Association. Psychological
Reports, 7(2), $337-344$.
Dennis, S., \& Humphreys, M. S. (2001). A context noise model of episodic word recognition. Psychological Review, ennis, S., \& Humphrey
108(2), 452-478.
Dorfman, D. \& Glanz
Dorfman, D., \& Glanzer, M. (1988). List composition effects in lexical decisions and recognition memory. Journal of
Memory and Language, 27, 633-648.
Dye, M., Ramscar, M., \& Jones, M. (2017). Representing the richness of linguistic structure in models of episodic
memory. Proceedings of the 39th Annual Conference of the Cognitive Science Society.
Estes, W. K., \& Maddox, W. T. (2002). On the processes underlying stimulus-familiarity effects in recognition of words and nonwords. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(6), 1003-1018.
Eugene, Z. B. (1972). Orthographic Distinctiveness as a Variable in Word Recognition. The American Jo
Eugene, Z. B. (1972). Orthographic Distinctiveness as a Variable in Word Recognition. The American Journal of
Psychology, 85(3), $425-430$.
Gillund, G., \& Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. Psychological Review, 91(1), 1-
Glanzer, M., \& Adams, J. K. (1985). The mirror effect in recognition memory. Memory \& Cognition, 12, 8-20.
Glanzer, M. \& Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. Journal of
Experimental Psychology: Human Learning and Memory, 2(1),21-31.
Experimental Psychology: Human Learning and Memory, , (1), 21-31.
Glanzer, M., Adams, J. K., Iverson, G. J., \& Kim, K. (1993). The regularities of recognition memory. Psychological
Enzer, M., Adams, J. K., Iverson, G. J., \& Kim, K. (1993). The regularities of recognition memory. Psychological
Review, $100,546-567$.
Hastie, R., \& Kumar, P. A. (1979). Person memory: Personality t
Journal of Personality and Social Psychology, 37(1), $25-38$.
Journal of Personalty and \& Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in
Her recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(6), 1947-1952.
Hintzman, D. L. (1988). Judgments of Frequency and Recognition Memory in a Multiple-Trace Memory Model. Psychological Review, 95(4), 528-551.
intzman, D. L. Caulton, D. A., \& Cu
Experimental Psychology: Learning, Memory, and Cognition, 20, 275-289.
Hirshman, E., Whelley, M. M., \& Palij, M. (1989). An investigation of paradoxical memory effects. Journal of Memory
and Language, 28(5), 594-609.
and Language, 28(5), 594-609.
unt, R. R., \& Mitchell, D. B. (1982). Independent effects of semantic and nonsemantic distinctiveness. Journal of
Experimental Psychology: Learning Memory and Cogition 8(1), 81-87.
Hunt, R. R., \& Eliot, J. M. (1980). The Role of Nonsemantic Information in Memory: Orthographic Distinctiveness
Effects on Retention. Journal of Experimental Psychology: General, 109(1), 49-74.
Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. Psychonomic Bulletin \& Review, 2(1), unt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. Psychonomic Bulle
$105-112$.
Johnson, M. N., Dye, M., \& Johns, B. T. (2017). Context as an organizing principle of the lexicon. In B. Ross (Ed.), The
Psy ones, M. N., Johns, B. T., \& Recchia, G. (2012). The role of semantic diversity in lexical organization. Canadian Journal of Experimental Psychology, 66, 115-124.
Verbal Learning and Verbal Behavior, 13(1), 63-69.
dauer, T. K., \& Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalen
assumptions for theories of word recognition. Journal of Verbal Learning and Verbal Behavior, 12(2), 119-131.
Maddox, W. T., \& Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. Journal
Maddox, W. T., \& Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. Journal of
Experimental Psychology: Learning Memory, and Cognition, 23(3) 539-559.
Experimental Psychology: Learning, Memory, and Cognition, 23(3), 539-55e
Journal of Experimental Psychology: Learning, Memory, and Cognition, 28 (4), $616-630$.
almberg, K.J. \& Nelson, T.O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. Memory \& Cognition, 31, 35-43. memory. Memory \& Cognition, 30(4), 607-613.
memory. Me mory \& Chagnition, 30(4), 607-613.
McClelland, J., \& Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the
effects of experience in recognition memory, Psychololel effects of experience in recognition memory, Psychological Review, 105, $724-760$.
Murdock, B.B. (1982). A theory for the storage and retrieval of item and associative information. Psychological Review,
urnane, K., Phelps, M. P., \& Malmberg, K. (1999). Context-dependent recognition memory: the ICE theory. Journal of
Experimental Psychology: General, 128 (4), 403-415.
Essfy, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental
Psychology: General, 115(1), 39-61. Psychology: General, 115(1), 39-61.
exman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., \& Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. Psychonomic Bulletin \& Review, 15(1),
$161-167$. 161-167.
1951). The generalization gradient in recognition memory. Journal of Experimental Psychology, 42(4), $231-235$.
Ramscar, M., Yarlett, D., Dye, M., Denny, K., \& Thorpe, K. (2010). The Effects of Feature-Label-Order and their Ramscar, M., Yarlett, D., Dye, M., Denny, K., \& Thorpe, K. (2010). The
implications for symbolic learning. Cognitive Science, 34(6), 909-957.
Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. The American Psychologist, 43(3), 151-160.
Scarborough, D. L., Cortese, C., \& Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory carborough, D. L., Cortese, C., \& Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory
Journal of Experimental Psychology: Human Perception and Performance, 3(1), 1-17.
Shiffrin, R. M., \& Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory Psychonomic Bulletin \& Review, 4, 145-166.
Psychonomic Bulletin \& Review, 4, 145-166. (1995). Effects of category length and strength on familiarity in
Shiffrin, R. M., Huber, D. E., \& Marinelli, K. .
recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(2), 267-287. Smith, S. M., Glenberg, A., \& Bjork, R. A. (1978). Environmental context and human memory. Memory \& Cognition,
Smith, S. M., Glenberg, A., \& Bjork, R. A. (1978). Environmental context and human memory. Memory \& Cognition,
64, 342 2-353.
Steyvers, M., \& Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. Journal of
Experimental Psychology: Learning, Memory, and Cognition, 29(5), $760-766$
Steyvers, M., \& Tenenbaum, , B. B. (2005). The large-scale structure of semantic
Steyvers, M., \& Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a
Tulving, E., \& Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. Psychonomic Bulletin
\& Review, $2(3), 387-390$.
\& Review, 2(3), $887-390$. J. J. (1968). Errors in recognition learning and retention. Journal of Experimental
Underwood, B. J., \& Freund, J. S. (1968). Errors in recognition learning and retention. Journal of Experimental
Psychology, $78(1), 55-633$
Zaki, S. R., \& Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perce
recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1022-1041.
recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1022-1041.
chmeister, E. B., Curt, C., \& Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of
word frequency. Bulletin of the Psychonomic Society, 11(6), 371-373
word frequuncy. Bulletin of the Psychonomic Society, 11(6), 371-373.
n Restorff, H. (1933). Uber die Wirkung von Bereichsbildungen im Spurenfeld. Psychologische Forschung, 18,
299-342.
Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. Psychologische Forschung, 18,
299-342.

# Creating words from iterated vocal imitation 

Pierce Edmiston (pedmiston@wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street<br>Madison, WI 53706 USA<br>Marcus Perlman (marcus.perlman@mpi.nl)<br>Max Planck Institute for Psycholinguistics<br>Nijmegen, 6500 AH The Netherlands<br>Gary Lupyan (lupyan@wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street<br>Madison, WI 53706 USA


#### Abstract

We report the results of a large-scale ( $N=1571$ ) experiment to investigate whether spoken words can emerge from the process of repeated imitation. Participants played a version of the children's game "Telephone". The first generation was asked to imitate recognizable environmental sounds (e.g., glass breaking, water splashing); subsequent generations imitated the imitators for a total of 8 generations. We then examined whether the vocal imitations became more stable and word-like, retained a resemblance to the original sound, and became more suitable as learned category labels. The results showed (1) the imitations became progressively more word-like, (2) even after 8 generations, they could be matched above chance to the environmental sound that motivated them, and (3) imitations from later generations were more effective as learned category labels. These results show how repeated imitation can create progressively more word-like forms while retaining a semblance of iconicity.


Keywords: categorization; transmission chain; language evolution

People have long pondered the origins of languages, especially the words that compose them. For example, both Plato in his Cratylus dialogue (Plato and Reeve, 1999) and John Locke in his Essay Concerning Human Understanding (Locke, 1948) examined the "naturalness" of words-whether they are somehow imitative of their meaning. Some theories of language evolution have hypothesized that vocal imitation played an important role in generating the first words of spoken languages (e.g., Brown et al., 1955; Donald, 2016; Imai and Kita, 2014; Perlman et al., 2015); early humans may originally have referred to a predatory cat by imitating its roar, or to the discovery of a stream by imitating the sound of rushing water. Such vocal imitation might have served to clarify the referent of a vocalization and eventually establish a mutually understood word. In this study, we investigate the formation of onomatopoeic words-imitative words that resemble the sounds to which they refer. We ask whether onomatopoeic words can be formed gradually and without instruction through repeated imitation.

Onomatopoeic words appear to be a universal lexical category found across the world's languages (Dingemanse, 2012). Languages all have conventional words for animal vocalizations and various environmental sounds. Rhodes (1994), for example, documented a repertoire of over 100 onomatopoeic


Figure 1: The design of the transmission chain experiment. 16 seed sounds were selected, four in each category of environmental sounds. Participants imitated each seed sound, and then the next generation of participants imitated the imitations and so on for 8 generations.
words in English, which he notes exist along a continuum from "wild" to "tame". People often use more wild vocal imitations and other sound effects during demonstrative discourse, especially when producing quotations (Blackwell et al., 2015; Clark and Gerrig, 1990). Wild words have a more imitative phonology whereas tame words take on more standard phonology of other English words. In some cases, words that begin as wild imitations of sounds become fully lexicalized and integrated into the broader linguistic system, when they behave like more "ordinary" words that can undergo typical morphological processes. Examples are English words like "crack" or the recently adapted "ping".

However, not all researchers agree that vocal imitation has any significant role in language. For instance, Pinker and Jackendoff (2005) suggested that, "Humans are not notably talented at vocal imitation in general, only at imitating speech
sounds (and perhaps melodies). For example, most humans lack the ability (found in some birds) to convincingly reproduce environmental sounds ... Thus 'capacity for vocal imitation' in humans might be better described as a capacity to learn to produce speech." Nevertheless, experiments show that people can be quite effective at using vocal imitation. For example, Lemaitre and Rocchesso (2014) collected imitations and verbal descriptions of mechanical and synthesized sounds. When participants listened to these and were asked to identify the source, they were more accurate with imitations than descriptions. A subsequent study found that vocal imitations tend to focus on a few salient features of the sound rather than a high fidelity representation, which aids identification of the source (Lemaitre et al., 2016).

Thus humans can be effective at communicating with vocal imitation, it can play an important role in narration and discourse, and it appears to be the basis for substantial inventories of sound-imitative vocabulary across languages. But the process by which onomatopoeic words like "meow", "ping" and "buzz" emerge from vocal imitations has yet to be observed. Here we examine whether simple repeated imitations of environmental sounds become more word-like even in the absence of explicit communication intent or the intent to create a word-like token. Alternatively, repeating imitations might never stabilize on a particular wordform, or the limited fidelity of human vocal imitation may simply restrict the formation of stable words through iterated imitations.

To test this, we recruited participants to engage in a large scale online version of the children's game of "Telephone" in which an acoustic message is passed from one person to the next. After obtaining these imitations, we investigated how the imitations changed over generations to determine whether they became more word-like. We investigated the acoustic properties of the imitations as well as the orthographic properties once transcribed into English words. We find that by both measures the imitations become more stable through repetition. In addition to stability, we also find that the imitations can still be matched back to the original sounds at above chance levels for many generations. Finally, we measured how quickly the invented words are learned as category labels in a category learning experiment, and find that later generation imitations are easier to learn as category labels.

## General Methods

In Experiment 1 we collected iterated vocal imitations using the transmission chain design depicted in Fig. 1. We then assessed changes in these imitations over generations in the remaining experiments, which are listed in Table 1. In Experiment 2 we assessed the extent to which each imitation could be matched back to its originating sound. Experiment 3 involved collecting transcriptions of imitations, and these transcriptions were matched back to the original sounds in Experiment 4. In Experiment 5 we selected transcriptions taken from first and last generation imitations as novel labels in a simple category learning experiment.

Table 1: Experiment sample sizes. Participants in Experiments 1-4 were recruited via Amazon Mechanical Turk and paid to participate in an online study. Participants in Experiment 5 were University of Wisconsin-Madison undergraduates who received course credit in exchange for participation.

| $\#$ | Experiment | N |
| :--- | :--- | ---: |
| 1 | Collecting imitations | 94 |
| 2 | Matching imitations to seeds | 752 |
| 3 | Collecting transcriptions | 218 |
| 4 | Matching transcriptions to seeds | 444 |
| 5 | Category learning | 63 |

## Exp 1: Collecting imitations

In Experiment 1 we collected the iterated vocal imitations that served as the basis for the remaining experiments. Our hypothesis was that these vocal imitations would become more stable as they were repeated over generations of speakers.

## Methods

We selected inanimate categories of sounds because they were less likely to have lexicalized onomatopoeic forms already in English, and they were less familiar and more difficult to imitate. Nonetheless, it is possible that lexical knowledge still influenced imitation fidelity-a possibility to be explored in future work. The sounds used here were selected using an odd-one-out norming procedure ( $N=105$ participants) to reduce an initial set of 36 sounds in 6 categories to a final set of 16 "seed" sounds: 4 sounds in each of 4 categories. The four final categories included: water, glass, tear, zipper.

Participants were paid to participate in an online version of the children's game of "Telephone". The instructions informed participants that they would hear some sound and their task is to reproduce it as accurately as possible using their computer microphone. Participants listened to and imitated 4 sounds. Participants received one sound from each of the four categories of sounds drawn at random such that participants were unlikely to hear the same person more than once. Imitations were monitored by an experimenter to catch any gross errors in recording before they were passed on to the next generation of imitators, including blocking sounds that violated the rules of the experiment, e.g., by saying something in English.

Given large differences in recording quality resulting from conducting the experiment online, we were unable to use previously published techniques for calculating acoustic distance (cf. Lemaitre et al., 2016). Instead, we obtained subjective measures of acoustic similarity using a controlled, randomized norming procedure completed by research assistants. Five RAs listened to pairs of imitations while blind to generation and rated their similarity on a 7-point scale where a 1 meant the sounds could never be confused with one another and a 7 meant the sounds were nearly identical.


Figure 2: Increase in acoustic similarity over generations. Points depict mean acoustic similarity ratings for imitations in each category of environmental sounds. The predictions of the linear mixed effects model with random effects for rater and category are shown, with error bands denoting +/- 1 standard error of the model predictions.

## Results

We collected a total of 480 imitations, of which 115 were removed, leaving 365 imitations along 105 contiguous transmission chains for analysis. Imitations from later generations were rated as being more similar to one another than imitations from earlier generations, $b=0.09(0.02), t(4.5)=4.42, p$ $=0.009$ (Fig. 2), suggesting that the imitations are stabilizing through repetition.

## Exp 2: Matching imitations to seeds

Experiment 2 was conducted to determine if the imitations retained some resemblance to the original environmental sound that motivated it (i.e. the seed sound). Participants listened to imitations and guessed which seeds they came from. By varying the relationship between the imitation and the options presented to each participant, we were able to assess the extent to which the imitations retained categorical as opposed to specific, identifying information. On the view that repetition makes the imitations more word-like, we expected later imitations to be better indicators of categories of sounds as opposed to specific sounds within each category.

## Methods

All 365 imitations collected in Experiment 1 were tested in each condition depicted in Fig. 3. On each trial participants listened to an imitation and selected among four possible options as to which option sounded the most like the imitation. They did not receive any feedback on their performance. We tested three types of matching questions that differed according to the relationship between the imitation and the four seed sounds serving as the options in the 4AFC task (Fig. 3).


Figure 3: Types of matching questions depicted in relation to the 16 seed sounds. For each question, participants listened to an imitation (dashed circles) and had to guess which of 4 sound choices (solid circles) they thought the person was trying to imitate. (Top) True seed questions contained the actual seed that generated the imitation in the choices, and the distractor seeds were sampled from different categories. (Middle) Category match questions also used distractor sounds from different categories but the correct seed was not the actual seed, but a different sound within the same category. (Bottom) Specific match questions pitted the actual seed against the other seeds within the same category.

## Results

Matching accuracy for all question types started above chance for the first generation of imitations, $b=1.65$ ( 0.14 ) log-odds, odds $=0.50, z=11.58, p<0.001$, and decreased steadily over generations, $b=-0.16$ (0.04) log-odds, $z=-3.72, p<$ 0.001 . We tested whether this increase in question difficulty was constant across the three types of questions or if some question types became more difficult at later generations.

The results are shown in Fig. 4. Performance decreased over generations more rapidly for specific match questions than for category match questions, $b=-0.05$ ( 0.02 ) log-odds, $z=-2.53, p=0.012$, suggesting that category information was more resistant to loss through transmission. One explanation for this result is that the specific match questions are simply harder than the category match questions. However, performance also decreased more rapidly for the easiest type of question where the correct answer was the actual seed generating the imitation. The advantage for having the true seed among the options decreased over generations, $b=-0.07$ (0.02) log-odds, $z=-2.83, p=0.005$. These results indicate that later generation imitations were more likely to be recognized as identifiers of a particular category than they were of particular exemplars within each category.


Figure 4: Accuracy in matching imitations back to seed sounds. Performance is separated by question type based on the relationship between the imitation and the options in the question (see Fig. 3). Lines depict predictions from the generalized linear mixed effects model along with $+/-1$ standard error of the model predictions.

## Exp 3: Collecting transcriptions of imitations

In addition to assessing stability in the acoustic properties of the imitations, we also measured orthographic agreement. If imitations are becoming more wordlike we would expect orthographic agreement to increase over generations.

## Methods

We selected the first and final three imitations in each transmission chain to be transcribed into English orthography. Participants were instructed to write down what they heard as a word so that the written word would sound like the message.

## Results

We collected a total of 2182 or roughly 21 transcriptions per imitation. All transcriptions containing actual English words were excluded from analysis. Orthographic agreement was measured as the longest contiguous substring match between the most frequent transcription of an imitation and all other transcriptions. Analyzing changes in orthographic agreement over generations paralleled what was observed in the analysis of acoustic similarity: Transcriptions from later generation imitations were more similar to one another in terms of orthographic distance than transcriptions from earlier generations, $b=-0.12$ ( 0.03 ), $t(3.0)=-3.62, p=0.035$ (Fig. 5). This result supports our hypothesis that the imitations were becoming more stable in both acoustic and orthographic forms.

## Exp 4: Matching transcriptions to seeds

Experiment 4 tested whether the transcriptions could be matched back to the original seed sounds.


Figure 5: Average orthographic distance among transcriptions of imitations taken from first and last generations. Each point represents the average distance among all transcriptions for a single imitation. Error bars are +/- 1 standard error of the linear mixed effects model predictions.

## Methods

The top 4 most frequent transcriptions for each imitation transcribed in Experiment 3 were tested in Experiment 4. Participants completed a modified version of the 4AFC described in Experiment 2. Instead of listening to imitations, participants now read a transcription of an imitation, which they were told was an invented word. They were instructed that the word was invented to describe one of four presented sounds, and they had to guess which one. Specific match questions (see Fig. 3) were not collected for transcriptions.

## Results

Participants were able to guess the correct meaning of the transcribed word above chance even after 8 generations of repetition, $b=0.83$ ( 0.13 ) log-odds, odds $=-0.18, z=6.46, p$ $<0.001$ (Fig. 6). This was true both for true seed questions, $b$ $=0.75$ (0.15) log-odds, odds $=-0.28, z=4.87, p<0.001$, and for category match questions, $b=1.02(0.16)$ log-odds, odds $=0.02, z=6.39, p<0.001$. The effect of generation did not vary across these question types, $b=0.05$ ( 0.10 ) log-odds, $z$ $=0.47, p=0.637$.

## Exp 5: Transcriptions as category labels

In Experiment 5 we examined whether there was a learning advantage to the more word-like imitations emerging through iterated repetition as compared to direct imitations of the source of the sound. We hypothesized that transcriptions of the more word-like forms emerging through repeated imitation should be easier to generalize to new category members than transcriptions from direct imitations.


Figure 6: Matching accuracy for transcriptions of imitations taken from first and last generations. True seed questions contained transcriptions of the actual seed generating the transcribed word. Category match questions contained transcriptions of imitations of other seeds from the same category.

## Methods

To determine which transcriptions to test as category labels, we selected transcriptions which were matched above chance in Exp. 2. Of these, transcriptions with fewer than two unique characters or more than 10 characters in length were excluded. The final set comprised first and last generation imitations sampled to control for overall matching accuracy.

Participants learned, through trial-and-error, the names for four different categories of sounds. On each trial participants listened to one of the 16 environmental sounds used as seeds and then saw a novel word-a transcription of one of the imitations. Participants responded by pressing a green button if the label was the correct label and a red button otherwise. They received accuracy feedback after each trial.

The experiment was divided into blocks so that participants had repeated exposure to each sound and the novel labels multiple times within a block. At the start of a new block, participants received four new sounds from the same four categories (e.g., a new zipping sound, a new water-splash sound, etc.) that they had not heard before, and had to associate these sounds with the same novel labels from the previous blocks. The extent to which their performance declined at the start of each block serves as a measure of how well the label they associated with the sound worked as a label for the category.

## Results

When participants had to generalize the meaning of the novel label to new category members (new sounds), they were faster when the label came from transcriptions of later generation imitations than from transcriptions of first generation imitations, $b=-114.13$ (52.06), $t(39.9)=-2.19, p=0.034$ (Fig. 7A). Accuracy improved over generations but did not significantly differ between groups, $p>0.05$. The effect can be further localized within each block. Comparing RTs on the trials leading up to a block transition and the trials immediately after the block transition (6 trials) revealed a reliable interaction between block transition and the generation of the transcribed label, $b=-146.75$ (65.47), $t(1869.7)=-2.24, p=$


Figure 7: (Top) RTs on correct trials by block, showing faster responses when learning category labels transcribed from last generation imitations. (Bottom) RTs on trials leading up to and immediately following the block transition where new category members are introduced.
0.025 (Fig. 7B). This suggests that in addition to becoming more stable both in terms of acoustic and orthographic properties, imitations that have been more repeated may also be faster to learn as category labels.

## Discussion

We show that repeated imitation of an originally imitative vocalization gradually becomes more word-like as it is transmitted along the chain of a "Telephone" game. The first evidence provided showed that imitations became more stable over generations of repetition, both in terms of acoustic similarity as well as in orthographic agreement. But more than just becoming more stable over generations, the imitations also become more word-like in that they served as more effective category labels. Category information was more resilient to transmission decay than specific information identifying a particular exemplar within a category. This category information remained even when the imitations were transcribed into lexical forms, as participants were able to guess the categorical meaning of the word at above chance levels even after 8 generations of repetition. One such consequence of having words is that they make categorization easier. In support of this conclusion, we found that participants were faster and
not less accurate in learning category labels that had emerged through repeated imitation than those who learned from transcriptions of direct imitations of the environmental sounds, completing the transition from nonverbal imitation to a fully lexicalized word form and demonstrating the impact of this transition on communication.

One result that did not fit squarely with imitations becoming more word-like is that with transcriptions, there was no difference over generations between question types. If the results of matching transcriptions back to seed sounds would have mirrored the results of matching imitations we would have expected the difference between True seed questions and Category match questions to decrease over generations, but it did not. Although participants were able to match transcriptions to categories of sounds after 8 generations of repetition, it was easier for them to match a transcription to the actual seed that generated the transcription, meaning that individuating information was retained over and above category information. One possible explanation for this is that by converting the imitations into orthographic representations of phonemes, idiosyncratic features of the sound could become rendered as categorical phonological features. This process could exaggerate the features and facilitate identification of the source. To test this we need to collect match accuracy for transcriptions on Specific match questions to see if transcriptions are able to be matched within-category even when the imitations that generated those transcriptions are not.

Our study focused on the formation of onomatopoeia-sound-imitative words-but in addition to onomatopoeia, many languages have semantically rich systems of ideophones. These words comprise a grammatically and phonologically distinct class of words that are used to express a variety of sensory-rich meanings (Dingemanse, 2012; Voeltz and Kilian-Hatz, 2001). Notably, these words are often recognized by native speakers to be somehow imitative of their meaning. For example, in Japanese, the word 'koron' - with a voiceless [k] refers to a light object rolling once, the reduplicated 'korokoro' to a light object rolling repeatedly, and 'gorogoro' - with a voiced [g] - to a heavy object rolling repeatedly (Imai and Kita, 2014). The iconicity of ideophones was verified by an experiment that tested the ability of nave listeners to guess the meanings of words sampled from five different languages (Dingemanse et al., 2016). Although words for sounds were guessed more accurately than the rest, listeners were better than chance at guessing the synonyms of ideophones that expressed meanings from all five semantic categories tested - color/visual, motion, shape, sound, and texture. In addition, laboratory experiments show that people are able to generate imitative vocalizations for a variety of non-sound concepts, and that these are also understandable to nave listeners (Perlman et al., 2015). Thus vocal imitation has the potential to play a role in word formation that extends beyond just the imitation of sounds.

Our findings from an online game of Telephone suggest that the formation of words from vocal imitation can be a sim-
ple process. The results show how repeated imitation can create progressively more word-like forms while retaining a resemblance to the original sound that motivated it. This raises the possibility that onomatopoeic words can be created simply through repeated imitation.

## References

Blackwell, N. L., Perlman, M., and Tree, J. E. F. (2015). Quotation as a multimodal construction. Journal of Pragmatics, 81:1-7.
Brown, R. W., Black, A. H., and Horowitz, A. E. (1955). Phonetic symbolism in natural languages. Journal of abnormal psychology, 50(3):388-393.
Clark, H. H. and Gerrig, R. J. (1990). Quotations as demonstrations. Language, 66:764-805.
Dingemanse, M. (2012). Advances in the Cross-Linguistic Study of Ideophones. Language and Linguistics Compass, 6(10):654-672.
Dingemanse, M., Schuerman, W., and Reinisch, E. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. Language, 92.

Donald, M. (2016). Key cognitive preconditions for the evolution of language. Psychonomic Bulletin \& Review, pages $1-5$.
Imai, M. and Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1651).
Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., and Susini, P. (2016). Vocal Imitations of Non-Vocal Sounds. PloS one, 11(12):e0168167-28.
Lemaitre, G. and Rocchesso, D. (2014). On the effectiveness of vocal imitations and verbal descriptions of sounds. The Journal of the Acoustical Society of America, 135(2):862873.

Locke, J. (1948). An essay concerning human understanding. In Dennis, W., editor, Readings in the history of psychology. Norwalk, CT.
Perlman, M., Dale, R., and Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. Royal Society Open Science, 2(8):150152-16.
Pinker, S. and Jackendoff, R. (2005). The faculty of language: what's special about it? Cognition, 95(2):201-236.
Plato and Reeve, C. D. C. (1999). Cratylus. Hackett, Indianapolis.
Rhodes, R. (1994). Aural images. Sound symbolism, pages 276-292.
Voeltz, F. E. and Kilian-Hatz, C. (2001). Ideophones, volume 44. John Benjamins Publishing.

# A Model of Event Knowledge 

Jeffrey L. Elman (jelman@ucsd.edu)<br>Department of Cognitive Science, 9500 Gilman Drive<br>La Jolla, CA 92093 USA

Ken McRae (kenm@uwo.ca)<br>Department of Psychology, Social Science Centre<br>London, ON N6A 5C2 Canada


#### Abstract

We present a connectionist model of event knowledge that is trained on examples of sequences of activities that are not explicitly labeled as events. The model learns co-occurrence patterns among the components of activities as they occur in the moment (entities, actions, and contexts), and also learns to predict sequential patterns of activities. In so doing, the model displays behaviors that in humans have been characterized as exemplifying inferencing of unmentioned event components, the prediction of upcoming components (which may or may not ever happen or be mentioned), reconstructive memory, and the ability to flexibly accommodate novel variations from previously encountered experiences. All of these behaviors emerge from what the model learns.


Keywords: events; schema; scripts; prediction; recurrent connectionist model

## Introduction

We know many things about the world. How that knowledge is organized, its content, and how it is stored, accessed, and learned have been the subject of semantic memory research for some time. A long and rich tradition of scholarship has produced a relatively stable set of theoretical constructs that are used for discussing this kind of knowledge, including categories, concepts, and features.

But people also possess another type of knowledge that has been long recognized as extremely important, although it is less clearly understood. This is knowledge about common situations and events, and has been referred to by a range of names, including pragmatic knowledge and world knowledge. Such knowledge appears to serve multiple purposes. It guides our behavior, and helps us interpret the behavior of others. We use this knowledge to anticipate the consequences of events as they unfold. We use this knowledge extensively in language understanding to make inferences about components of situations that may be unstated or incompletely described.

Bartlett (1932) was one of the first psychologists to talk about the role of such knowledge in memory. Later, in the 1970s and 1980s, cognitive psychologists such as Bransford and colleagues demonstrated that event knowledge is important in encoding and retrieving details about situations. Garrod and Sanford, among many others, showed that this kind of knowledge supports inferences in language
comprehension. One assumption that appears to be shared (though was often implicit) was that the use of world/pragmatic/event knowledge in language comprehension occurred at late stages in processing. In large part this reflected theoretical assumptions of the time in linguistics and psycholinguistics, but it is also true that the typical experimental tasks used at the time were off-line, and did not lend themselves to tracking real-time incremental processing.

Over the years, there have been a number of attempts to formalize this kind of knowledge, giving rise to mechanistic explanations involving frames (Minsky, 1974), scripts (Schank \& Abelson, 1977), schema (Norman \& Rumelhart, 1981), and stories (Mandler, 1984), among others. Although the core intuitions motivating these proposals were widely accepted, the actual implementations revealed a number of challenges. Templates were inherently rigid and inflexible. Yet most situations admit a large range of variation and novelty. Moreover, many situations involve blends of multiple events. Symbolic architectures did not lend themselves to dealing with such challenges. Thorny questions were raised and not satisfactorily answered: What is an event (and what is not)? What is the content and detail of event knowledge? Does event knowledge have a structure common across all event types? How is event knowledge accessed and used? How is event knowledge learned? These questions remain open to this day.

Several recent developments have encouraged cognitive scientists to focus more intensely on event knowledge and how best to model it. Our own interest arises from work in language processing using real-time measures to examine processing as comprehenders deal with incrementally presented input. There is now considerable evidence that event knowledge plays a significant role in comprehension very early in processing. Indeed, it guides expectations even in advance of input being received. The time course of how this knowledge is accessed and deployed is now not only of great theoretical interest (insofar as it may constrain our theories about the cognitive architecture underlying language understanding), but has become something that can be measured empirically.

A second development was the emergence of nonsymbolic computational frameworks that demonstrate the ability to capture behaviors that simultaneously reflect
awareness of global abstractions as well as sensitivity to ways in which those abstractions may be graded and affected by subregularities and even idiosyncracies. Connectionist models have these qualities (although Bayesian models often do as well). Our research uses a connectionist model because they exhibit key additional capabilities. They learn by example, and they allow us to probe in (simulated) real time the dynamics of the network's responses to incrementally presented input.

In the remainder of this paper, we present a model and report a set of simulations we have conducted. We begin by explaining the design criteria that guided model development. These criteria were chosen because we believe they are needed to model processes that reflect the use of event knowledge in human behavior. We conclude by discussing what we have learned from the model, and ways in which it might guide future research.

## The Model

## Design Criteria

The model's architecture was developed with the goal that it should have the following four properties.

Learn the components that comprise an activity. We make the assumption that events can be viewed as sequences of activities, where activities occur in the moment and are comprised of various participants, actions, and contexts. Rather than prespecifying a template for necessary or sufficient components, the model must learn which components occur and co-occur across contexts and sequences. These co-occurrences may be statistically variable, and the model must learn these (often high-order) statistical interdependences.

Learn the temporal structure of activity sequences. We also assume that the temporal structure of activity sequences that make up an event may be variable across instances of any given event type. The model must learn this temporal structure, including cases in which that structure is rigid and obligatory as well as cases in which there is a high degree of variability or optionality. The model should be able to use its knowledge of temporal structure to anticipate likely future activities given previously encountered sequences. These expectations should reflect both global contingencies as well as predictions that may reflect more idiosyncractic variants of an activity sequence. In human terms, the model should be able to make predictive inferences.

Learn to generalize from specific examples of events. Although the model will learn from multiple examples of a given event type, it must learn the (often graded) patterns that underly them. It must also learn subregularities and if possible, exceptions.

Fill in missing information. Both during learning and testing, the model may be exposed to activity descriptions in which some highly expected information is omitted. The model should be able to activate missing elements, as appropriate (pattern completion). In human terms, the model should be able to make elaborative inferences.

## Architecture

The architecture of the model is shown in Figure 1.


Figure 1
There is a single network, but the left and right portions play complementary roles. The left portion receives input from the world in the form of (localist) specifications of potential participants, actions, and contexts that might characterize the current activity under description. Each rectangle thus represents a number of possible inputs of the same category (agents, patients, etc.). It should be emphasized, however, that there is no representational status to these groups. As far as the network is concerned, every input node in all of these groups is orthogonal to every other node. If there are similarities in terms of behavior or statistics of privilege of occurrence, the network must discover them. Input nodes are fully connected to nodes in the Hidden Unit layer, and hidden units also connect (with different weights) back to input units. This use of recurrence allows the network not only to learn co-occurrence patterns among input units, but also to implement constraint satisfaction. This means that after the network has learned, it has the potential to activate missing elements in an input pattern, as appropriate. The Next Activity side of the network consists of units that are identical to the Current Activity units, but the job of the Next Activity units is to predict which activity will follow, given the sequence so far. Recurrent connections from the hidden units back to themselves are critical for this function because they provide the network with an internal representation (which must be learned) of the past that can be used for prediction. This architecture builds on elements of prior modeling that has provided a strong foundation for the present approach, including in particular Botvinick and Plaut (2004), Elman (1990), Rogers and McClelland (2004), Rumelhart, Smolensky, McClelland, and Hinton (1986), St. John and McClelland (1990), and Reynolds, Zacks, and Braver (2007).

## Training and Testing

Simulations were conducted using the $r b p$ package from the PDPtool simulator (McClelland, 2016). Weights in the network were initialized with random values between $\pm 0.1$ and adjusted gradually using backpropagation through time (Williams \& Zipser, 2004). Training stimuli were either artificially generated activity sequences (Studies 1-3) presented one activity at a time, or sequences obtained from human norming data (Study 4). After training, testing was conducted by freezing weights and presenting the network with input sequences designed as analogs of stimuli used in human experimental paradigms. Details of the general training regime can be found in http://tatar.ucsd.edu/jeffelman/EventModelTraining.html, and details relevant to each simulation are given below.

## Simulation Results

## Study 1: Pattern completion and elaborative and predictive inferences

Typical language use relies heavily on interlocutors' shared knowledge. This allows speakers to omit information that is assumed to be known by the comprehender, and allows comprehenders to infer unstated information. A frequent distinction is made between elaborative inferences, which involve unstated details regarding an activity currently described, and predictive (or forward) inferences, which involve expectations about what will occur next. Bridging inferences are those in which a comprehender draws on knowledge only as needed to understand a prior statement. The extent to which, and conditions under which, such inferences are drawn remains a topic of debate. Bridging inferences are largely uncontroversial. However, whether, and under which conditions, elaborative and predictive inferences occur is still debated (for review, see Murray, Klin, \& Myers, 1993). In Study 1, we first verify that the constraint satisfaction properties of the network do support inference under optimal conditions. We then examine the fragility or robustness of such inferencing because it has been claimed that discrepant data have arisen from stimulus properties and the sensitivity of behavioral measures.

Simulation 1.1. The network was trained on event sequences that ranged in length from three to six activities. The sequences might be glossed as (1) John goes to a fancy restaurant; (2) John is cutting wood in the forest, using an axe; (3) John (and other people) cut themselves accidentally with a knife, and he bleeds; (4) John (and other people) cut themselves accidentally with an axe and the wound is fatal; (5) Mary and Penny are in the library and Mary asks Penny a question, which Penny answers. Having learned these sequences, the model was then tested on novel sequences. The sequences were novel both in that they omitted critical information, and they involved new combinations of activities that the model had not encountered in the same event. Figure 2 shows activations in the Next Activity units
in response to the input sequence John is in a restaurant; John cuts himself; What happened to John? (the query takes the form of simply presenting John without any specified result, so the network must fill in the information). Figure 3 shows similar activations, but in response to the sequence John is in the forest; John cuts himself; What happened to John?


Figure 2


Figure 3
After receiving the input that John is in a restaurant, a knife is inferred to be present, whereas in the forest, axe is activated. These may be considered elaborative inferences. Then when John cuts himself, with no instrument mentioned, the network immediately begins to predict the result that is consistent with the instrument. These are predictive inferences. Such inferences have not always been found in humans, however. One possibility raised by Murray et al. (1993) is that failures to detect predictive inferences may result from experimental stimuli in which either the forward inference is disrupted, or it is not tested soon enough. In Figure 4, we see what happens when the discourse is disrupted by switching to a new topic, which is


Figure 4
a situation involving Mary, immediately after the cutting activity. The network begins to predict that John will bleed (because he is assumed to have cut himself with a knife, given the restaurant context). However, as soon as Mary is introduced, the activations of all consequences of cut
decrease sharply. Probing for the consequence of John cutting himself subsequent to this topic change would show little or no evidence of the predictive inference, consistent with Murray et al.'s findings.

Simulation 1.2. One open question concerns precisely how far into the future comprehenders predict when processing incrementally presented language. In much of the empirical literature focusing on prediction in language, there has been an implicit assumption that the next word in a sequence is anticipated by comprehenders, but nothing beyond that. However, more recent findings suggest that when language is used to describe an event, comprehenders anticipate event-relevant elements even at points in the discourse where they might not be appropriate (Metusalem, Kutas, Urbach, Hare, McRae, \& Elman, 2012). A simplified stimulus example is the short story: The crowd is in the stands. The crowd looks around. The skater goes to the podium. The audience applauds, The skater receives a $\qquad$ Participants' brain activity was measured while reading the final noun. When an event-appropriate word, such as medal, was presented, the N400 amplitude was small. A word that was completely anomalous (e.g., bleach) elicited a large N400. However, a word that was contextually anomalous but event-appropriate (e.g., podium) produced an N400 with intermediate amplitude. The authors interpreted this as evidence that event elements are activated and available even at times when they might not be expected immediately.
Figure 5 shows the network's activations in the Next Activity units throughout such a stimulus sequence.


Figure 5
By the second activity, the network has already activated two event-appropriate elements, podium and medal. Bleach is not activated at all. As the focus shifts in the fourth activity back to the crowd, both medal and podium are deactivated. However, near the end, the network re-activates both. The re-activation of medal is not surprising because it has been mentioned explicitly. However, the network has also learned that podium is the likely location for awarding a medal and so activates it as well, though at a lower level. There are two lessons from this simulation. First, behavioral evidence for the activation of putatively inferred event elements may depend on the timing of the probe. Second, it may be that only highly sensitive behavioral measures will reveal the presence of partially activated event elements. These elements, even if only partially activated, become more easily accessible should subsequent discourse make reference to them. This in fact was seen in Simulation 1.1.

## Study 2: Novel Events and Blending

In real life, events exhibit not only variability (which the model accommodates, as we see in Figures 1 and 2) but often are combined in novel ways. Fixed templates or rigid structures are ill suited for dealing with this. In the next simulation, we test the model's ability to flexibly respond when events are combined in unusual ways.

Simulation 2.1. The model was trained on sequences that included examples of going to a restaurant (as in Simulation 1.1), and activities corresponding to a romantic relationship between two people (John and Mary), with Mary being married to a third person (Bill). Furthermore, the model was exposed to examples of aggressive behavior between various people (but not including John or Bill). In many of the latter examples, weapons are used. Gun is a more typical weapon, but knives are used occasionally. After training, the model was tested on a sequence that might be glossed as John and Mary are at a fancy restaurant. John and Mary cut steak with a knife. Bill enters the restaurant. Bill attacks John. Activations of relevant nodes are shown in Figure 6.


Figure 6
Two things are apparent. First, as soon as Bill enters the restaurant, the model quickly adjusts its expectations about what it predicts will happen next. Second, and more interesting, is that although the model has learned that gun is the most common weapon used in aggressive behavior, the presence of knife that was established from the outset (even prior to its mention) leads to the knife being the predicted instrument in this new situation. Thus, the model not only adjusts to a change in sequence structure that it has not encountered before, but it also flexibly incorporates relevant components from the first event into the second event. That is, the model produces a novel response to a situation it has never encountered by drawing and integrating knowledge from different events.

## Study 3: Priming

Studies 1 and 2 illustrate examples of priming. There is a large literature showing that event relevant information facilitates processing target elements related to that event. These include typical agents, patients, and instruments priming their event-relevant verbs, priming between eventrelevant nouns, and verbs priming their event-related agents, patients, and instruments (for review, see McRae \& Matsuki, 2009). The model exhibits the same behavior, not shown here because of space limitations. Instead, we
demonstrate an example of priming involving second order dependencies between event elements. Bicknell et al. (2010) found that the patient that is expected to follow a given verb may depend on the agent carrying out the action. Thus, shopper saved... primes money, whereas the lifeguard saved... primes person. (Control conditions established that the priming was not directly between the agent or verb and the patient, but that it required the combination.)

Simulation 3.1. The model was trained on various examples of shoppers and lifeguards (and other people) in events in which saving was one of the activities. Typically, reflecting the corpus analyses carried out by Bicknell et al., shoppers save money whereas lifeguards save people. When probed with the partial description of an activity shopper + saved (Figure 7), the model predicted money as the most likely patient, compared to lifeguard + saved, which led to greater activation of person (Figure 8). However, we also see that that there is an asymmetry in the responses, such that at later stages in processing, lifeguard + saved results in an increased activation of money (though still lower than person).


Figure 7


Figure 8

This reflects asymmetries in the training data that mirror asymmetries in corpus analyses, that is, that save is overall more commonly associated with money than with people. We might test the model's predictions (to our knowledge, as of now untested) by testing whether the timing of the patient probe leads to different degrees of facilitation, depending on when the probe was presented.

## Study 4: Learning from Human Data

In the previous simulations, we used training sets that were designed by hand. The design of the training corpora was controlled to carefully probe the network's behavior under different learning situations. This strategy is similar to that used in many human behavioral experiments. But in real life, people's knowledge of events results from experiences that may involve considerably greater variability. Consolidating such experiences and making sense of commonalities, subregularities, and exceptions is a challenge. Furthermore, temporal structure may vary considerably not only between different event types but even within a single event type. For example, there may be some parts of an event in which the ordering of activities is consistent and even obligatory (eggs must be broken before they are fried), whereas activity sequences in other parts of the event may be optional (one might make coffee before making eggs, or after). To investigate these issues, we
conducted a norming study to sample people's knowledge of types of events.

Norming study 4.1 We used 81 events, drawing on prior literature that has used stimuli that describe events and situations. Some of these events have clear goals and outcomes (e.g., fixing a flat tire). Other events are more situation-like, in that things happen but the goal and outcome are less clear (e.g., going to a picnic). Using Mechanical Turk, participants were asked to list up to 12 activities for each event. Participants saw a random subset of $10-12$ of the 81 events, and each event was presented to 22-24 different participants. Table 1 shows responses from three participants for fixing a flat tire.

Table 1: fixing a flat tire

| Pull over | Get out of car | Pull over |
| :--- | :--- | :--- |
| Get out of car | Loosen lug nuts | Open trunk |
| Open trunk | Jack up car | Get tire iron |
| Get spare tire | Remove lug nuts | Get spare tire |
| Get jack | Remove flat tire | Put on hazard lights |
| Remove flat tire | Put on new tire | Jack up car |
| Put on new tire | Tighten lug nuts | Remove lug nuts |
| Tighten lug nuts | Remove jack | Take flat tire off |
| Put flat tire in trunk |  | Put on new tire |
|  |  | Tighten lug nuts |
|  |  | Lower car |

The data can be visualized using graph analysis, in which nodes represent activities and directed arrows show temporal sequence (size indicates frequency), as in Figure 9.


Figure 9
Some of the sequences are consistently ordered (e.g., jack up car $>$ remove flat tire $>$ put on new tire), undoubtedly reflecting causal constraints. Other sequences may be performed optionally at different times. How does the model deal with such data?
Simulation 4.1. The model was trained on the activity sequences provided by 23 participants for fixing a flat tire. Of particular interest is that although the model responds appropriately to the data it was trained on, its responses also incorporate what it has learned from other participants. The model thus does not slavishly reproduce the individual training data, but detects general patterns that are common across all the data.
Can the model generate activity sequences on its own? We tested this by seeding the model with a reasonable starting
activity, and then using the most strongly predicted elements as the subsequent input. This process iterated until the event was complete. The initial five activities in the network's self-generated sequence are shown in Figure 10 (presenting greater than five makes the figure unreadable). Notably, the network's self-generated activity sequence is not identical to any single participant's sequence. However, it is a completely reasonable abstraction of the sequencing across all participants' descriptions.


Figure 10

## Discussion

Our goal was to develop a model that could learn the structure and temporal dynamics of activity sequences, as well as the co-occurrence properties of participants, activities, and contexts in those sequences. Although we might call these sequences events, the concept of event is not a primitive in the model and events are not pre-defined templates. Rather, what we might call an event is an epiphenomenal consequence of having to learn about activity sequence structure. Having done this, the architecture of the model allows it to perform pattern completion, both in the moment (supporting elaborative inferences) and across time (supporting predictive inferences). The model replicates a wide range of behavioral studies (only a few of which are described herein) for which event knowledge has been hypothesized to play a role. It also produces unanticipated behaviors that can be tested empirically to assess the model.
A great deal remains to be done. The model's inputs serve as cues to event knowledge, but the model itself does not provide those cues. Those cues must come from perceptual or motor evidence from the world as well as a language processor. Nor does the model provide an account for how these various cues can serve to alter focus on different event elements, including adjusting how the temporal contour of the event is understood (e.g., by grammatical aspect). We are guardedly optimistic that these are tractable problems, and that the model we propose here provides a solid framework for understanding how people acquire, represent, and use knowledge of events in the world.

## Acknowledgments

This work was supported by grants from the NIH (CHD053136) and Natural Sciences and Engineering Research Council of Canada (OGP0155704).

## References

Bartlett, F. C. (1932). A Theory of Remembering. Cambridge, UK: Cambridge University Press. Bicknell, K., Elman, J. L., Hare, M., Mcrae, K., \& Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. Journal of Memory and Language, 63(4), 489-505.
Botvinick, M. M., \& Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. Psychological Review, 111(2), 395-429.
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.
Mandler, J. M. (1984). Scripts, Stories and Scenes: Aspects of Schema Theory: Hillsdale, NJ: Erlbaum.
McRae, K., \& Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. Language and Linguistics Compass, 3, 1417-1429.
McClelland, J. L. (2017). PDPTool. Palo Alto: Stanford University. Retrieved from https://web.stanford.edu/group/pdplab/pdphandbook/
McRae, K., Hare, M., Elman, J. L., \& Ferretti, T. R. (2005). A basis for generating expectancies for verbs from nouns. Memory and Cognition, 33(7), 1174-1184.
Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., \& Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. Journal of Memory and Language, 66(4), 545-567.
Minsky, M. (1974). A Framework for Representing Knowledge. Cambridge, MA: MIT Press.
Norman, D. A., \& Rumelhart, D. E. (1981). The LNR approach to human information processing. Cognition, 10(1), 235-240.
Reynolds, J. R., Zacks, J. M., \& Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. Cognitive Science, 31, 613-643.
Rogers, T. T., \& McClelland, J. L. (2004). Semantic cognition: A parallel distributed processing approach. Cambridge, MA: MIT Press.
Rumelhart, D. E., Smolensky, P., McClelland, J. L., \& Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland \& D. E. Rumelhart (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 2, pp. 7-57). Cambridge, MA: MIT Press.
Schank, R. C., \& Abelson, R. P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum Associates.
St. John, M., \& McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. Artificial Intelligence, 46, 217-257.
Williams, R. J., \& Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. Neural Computation, 1, 270-280.

# The role of intentionality in causal attribution is culturally mediated: evidence from Chinese, Mayan, and Spanish populations 

Stephanie Evers (saevers@buffalo.edu) ${ }^{\text {a }}$<br>Erika Bellingham (ebelling@buffalo.edu) ${ }^{\text {a }}$<br>Katharine Donelson (donelson@buffalo.edu) ${ }^{\text {a }}$<br>Jing Du (millydu1019@buaa.edu.cn) ${ }^{\text {b }}$<br>José Antonio Jódar Sánchez (jiodarsa @buffalo.edu) ${ }^{\mathbf{a}}$<br>Fuyin Li (thomasli@buaa.edu.cn) ${ }^{\text {b }}$<br>Jürgen Bohnemeyer (jb77@buffalo.edu) ${ }^{\mathbf{a}}$<br>${ }^{\text {a Department of Linguistics, University at Buffalo - SUNY, } 609 \text { Baldy Hall, Buffalo, NY 14260, USA }}$<br>${ }^{\mathrm{b}}$ School of Foreign Languages, Beihang University, No. 37, Xueyuan Rd, Beijing, China, 100191


#### Abstract

Speakers of Mandarin, Spanish, and Yucatec Maya watched videos of two actors involved in a causal chain initiated by one of them. After watching each video, participants divided 10 tokens into piles indicating their assignment of responsibility for the resulting event. There was a significant interaction between intentionality and population: causer and causee intentionality made a significant difference only for the Spanish and Yucatec participants, but not for the Chinese participants. This is in line with previous findings suggesting that internal dispositions play a lesser role in responsibility attribution in societies in which attention to individual agency is far more common than attention to group agency.


Keywords: causality; agency; responsibility; intentionality; cultural mediation; linguistics; social psychology

## Introduction

Linguistic theories of the mapping between meaning and syntactic form in language have long recognized the key role of agency and causality. In verbal representations of causal chains, causality and agency determine the assignment of grammatical relations such as subject and object, voice (active/passive), case marking, and a host of other properties (e.g., Croft 1987; Dowty 1991; Foley \& Van Valin 1984; among many others). Yet, much of this work implicitly treats causality and agency as universal notions - even in crosslinguistic research (e.g., Comrie 1981; Dixon 2000; Shibatani 2002). Meanwhile, a growing body of work in the field of social psychology calls the universality of these notions very much into question.
Take, for example, the nexus between responsibility and intentionality. Much theoretical work on agentivity in language assumes that prototypical agents are volitional and that nonvolitional causers are either atypical agents or not agents at all (Dowty 1991; Lakoff 1977; Van Valin \& Wilkins 1996; inter alia). [We assume in the following that volitional actions require a choice on the actor's part and intentional actions require a plan; the latter are thus a subset
of the former (cf. Van \& Valin \& Wilkins 1996: 315-316)]. However, cross-cultural research since the 1990s has uncovered evidence suggesting that internal dispositions such as volition and intentions do not play the same role in attributions of causality across cultures. Much of this research has focused on a contrast (treated as binary) between two types of societies that are said to differ from one another in terms of the relative prominence of individual agency and group agency in their members' cognition. American culture has been said to downplay collective agency in favor of individual agency. We assume that our Spanish participants exhibit the same trait. In contrast, group agency is hypothesized to play a relatively more prominent role in Chinese culture.

For example, in one classic study, Morris \& Peng (1994) examined reports of similar crimes in Chinese- and Englishlanguage newspapers, showing that the former paid relatively more attention to explanatory factors in the situational context of the crime while the latter spent more time discussing the perpetrator's presumed disposition. Similar patterns have been reported, with varying theoretical conclusions, by Chiu et al (2000); Choi \& Nisbett (1998), Choi et al (1999), Maddux \& Yuki (2006); Menon et al (1999), and Peng \& Knowles (2003), inter alia.

If the attribution of responsibility and causality is indeed influenced by culture-specific folk theories of agency, then such folk theories may also influence the role of agency in the grammars of different languages. We are currently laying the groundwork for a large-scale crosslinguistic study of the representation of causality in the grammars of languages spoken around the world. In preparation for this effort, we decided to directly investigate the role of intentionality in causal attributions in three populations: Mandarin Chinese speakers from Mainland China, Spanish speakers from Spain, and Yucatec Maya speakers from Mexico. Mayans practice a traditional Mesoamerican horticulturalist society surrounded by a Western-dominated Spanish-speaking society and transitioning into the Age of

Globalization. Their inclusion in our study allows us to open up the investigation beyond the egocentric-sociocentric dichotomy of the debate on 'dispositionalism' and 'situationalism' in social psychology, but also to compare our findings to those reported in Le Guen et al (2015), whose sample includes Yucatecans as well.

Le Guen et al (2015) investigated the role of concepts of chance and coincidence in causal attributions among four populations, comparing Tseltal and Yucatec Mayans, urban Mexicans of non-indigenous descent, and Germans. They based their stimuli on Alicke's (2000) 'Culpable Control Model', which distinguishes three components of responsibility: whether the causer intended their immediate action (' $I \rightarrow A$ '), whether they intended the final outcome (' $\mathrm{I} \rightarrow \mathrm{O}$ '), and whether there is a causal relation between the action and the outcome (' $\mathrm{A} \rightarrow \mathrm{O}$ '). The distinction between $\mathrm{I} \rightarrow \mathrm{A}$ and $\mathrm{I} \rightarrow \mathrm{O}$ informed our stimulus design as well. Le Guen et al found that intentionality had a greater effect on the Mayan participants' attributions of causality than on those of the German and Mexican populations. By way of interpretation, they tacitly point toward the tradition of anthropological research identifying elements of 'magical thinking' in traditional non-Western cultures (e.g., EvansPritchard 1937).

## Method

Speakers of Mandarin, Spanish, and Yucatec Maya watched videos of two actors involved in a chain of events that culminates in a resulting event. In each case, the chain is initiated by one actor, dubbed the Causer in the following. The second actor is affected by the Causer's action and may or may not in turn affect a third, inanimate, entity. This second actor is labeled the Causee After watching each video, participants divided 10 tokens into piles indicating their assignment of responsibility for the resulting event. Piles represented 'Causer', 'Causee', and 'Neither'.

## Participants

12 speakers of Yucatec Maya, 16 Mandarin speakers, and 20 Spanish speakers were recruited from and tested at sites in Barcelona and Murcia, Spain, at Beihang University in Beijing, China, and in the village of Yaxley, Quintana Roo, Mexico. The Chinese participants included 8 women and 8 men aged 19-40 ( $M=27.46, ~ S D=4.98)$. Spanish participants included 12 women and 8 men aged 18-55 ( $\mathrm{M}=$ $28, \mathrm{SD}=12.92$ ). The Yucatec participants included 5 women and 7 men aged 18-76 ( $\mathrm{M}=44, \mathrm{SD}=16.83$ ). Participants completed the tasks in about 45 minutes and were compensated 100 pesos (approximately $\$ 5$ USD), 8 euros (approximately $\$ 9$ USD), and a cup of coffee, respectively (all Mandarin participants were students of the school of foreign languages at Beihang University).

## Materials

The experiment comprised a training phase involving 10 video clips and a test phase with 24 video clips. Four of the
training clips and three of the test items were cut from news reports, a home video show, and a movie. The remaining videos were taped with students and faculty of the University at Buffalo Linguistics Department staging the actions and events. The mean duration of the test videos was 8.05 seconds (SD 4.56s). They were shown to the participants on laptop computers.

The test items are described in Table 1 in terms of the action/event involving the Causee. These actions/events can all in one way or another be understood as caused by the Causer - in some cases via a physical impact on Causee; in others via a reflexive/uncontrolled or deliberate psychological response to the Causer's behavior or as a response to a gestural command by Causer. Three intentionality variables are represented as well: whether Causer intended their action $(\mathrm{I} \rightarrow \mathrm{A})$, whether Causer intended the outcomes of the chain $(\mathrm{I} \rightarrow \mathrm{O})$, and whether Causee acted intentionally/volitionally. ${ }^{1}$

Table 1: Test Phase Video Descriptions

| Clip Description | Causer |  | Causee |
| :--- | :---: | :---: | :---: |
| (CE=Causee) | intentional | intentional |  |
|  | $\mathrm{I} \rightarrow \mathrm{A}$ | $\mathrm{I} \rightarrow \mathrm{O}$ |  |
| CE breaks a plate | Yes | Yes | Yes |
| CE breaks eggs | Yes | Yes | Yes |
| CE collapses a cup tower | Yes | No | No |
| CE collapses a cup tower | Yes | Yes | No |
| CE collapses a cup tower | Yes | Yes | No |
| CE cuts a piece of paper | Yes | Yes | Yes |
| CE falls | Yes | Yes | No |
| CE falls | No | No | No |
| CE falls | No | No | No |
| CE is scared/falls over | Yes | Yes | No |
| CE is startled | No | No | No |
| CE is thrown a distance | Yes | Yes | No |
| CE laughs | Yes | Yes | No |
| CE leaves | Yes | No | Yes |
| CE leaves | Yes | Yes | Yes |
| CE sits down | Yes | Yes | Yes |
| CE swings a swing | Yes | Yes | Yes |
| CE tears a piece of paper | Yes | Yes | Yes |
| CE tears a piece of paper | Yes | Yes | No |
| CE tears a piece of paper | No | No | No |
| CE tears a piece of paper | Yes | Yes | No |
| CE tosses a ball into a box | Yes | Yes | Yes |
| CE wakes | Yes | No | No |
| CE yawns | No | No | No |
|  |  |  |  |

[^62]

Figure 1: Causee leaves when causer sings poorly


Figure 2: Causee knocked into cups by causer with cart
Intentionality was indicated by obvious body language on the part of the actor, including whether or not they looked at the causee or touched them with their hands in a manner that appeared to be controlled. For example, in an 'unintentional' clip, a Causer walks into a room without looking at the causee and loudly sneezes, which causes a startled Causee to tear a piece of paper. In the contrasting 'intentional' clip, the Causer looks at the Causee and deliberately pushes them, causing them to tear the paper. Four of the training items featured scenes that fit the same parameters as the test items. The remaining six items featured actions on which the two actors collaborate, events that occurred without the involvement of either actor, and events in which one actor destroyed an object while the other looked on.

Participants were given 10 identical tokens, which consisted of small glass stones or other objects of similar size. To prevent confusion about the purpose of the task, no tokens resembling currency were used. These tokens represented total responsibility for end results in video clips observed during the task, where each token symbolized $10 \%$ of total responsibility. Participants were also given a sheet of paper with three circles drawn on it. The leftmost circle represented the actor who ended in the left-most position or final frame of the video clip, the center circle represented the other actor, and the right-most circle represented 'neither actor.' Circles were arranged in a horizontal row, or in two rows where the two circles representing actors were next to
one another in the top row and the 'neither' circle was drawn below them.

## Procedure

Preparation. Prior to working with participants, researchers established how to convey the concept of responsibility in the target language. A complication of concern was the potential for negative implications of 'blame'. To avoid participant confusion over assigning blame to a neutral event, researchers explained that participants should think of assigning responsibility in terms of explaining the events to someone who wanted to know what happened and why. For video clips depicting one character involuntarily initiating a causal chain, participants had to decide between prioritizing intentionality or control in the assignment of responsibility. This provided data on cross-cultural differences on how these two factors were weighed.

Before the training phase, the task was explained to participants in their native language. Participants were asked to indicate which actor in each video was responsible for the resulting event and reminded that they could distribute responsibility between all three piles, two piles, or just one pile so long as the distribution of tokens at the end of each trial was proportional to the amount of responsibility of each actor. After this explanation, the participant watched the 10 training videos.

Training. The purpose of the training phase was to allow the participants to gradually familiarize themselves with the rationale of the ratings procedure. For this reason, it was designed to initiate training with six scenes in which the assignment of responsibility seemed straightforward (collaborative action; no involvement of either actor; or one actor involved while the other was not), followed by four items similar in structure to the test items, where responsibility assignment is more competitive, at the end. For each of the first three videos, the experimenter would demonstrate by playing the video, and apportioning the tokens in the appropriate way, and then would explain why they did so. Next, the experimenter would invite the participant to use the tokens to rate responsibility in each of the remaining seven scenes. The experimenter would play a clip, establish which circle on the paper represented each actor in the video, replay the video and ask the participant to distribute the tokens. The experimenter would correct any confusion about allocating the tokens and verified that the participant understood the task.

Testing. The test items were presented in one of four pseudo-randomized orders. Participants were randomly and evenly distributed over these four orders.

During the test phase, participants watched the 24 test clips. After each clip, experimenter and participant established which circle would represent each actor in the video and then played the video a second time. The participant was then asked to distribute responsibility for the final outcome of the clip between the actors. Responses
were recorded in a spreadsheet. Experimenters did not question participant understanding of video clips or correct token distribution during this portion of the experiment. After watching the 24 clips, the participant viewed each clip again and provided a verbal description of the action in the video. The sessions were video recorded in their entirety.

## Results

## Exclusions

One response by a Mayan participant was accidentally omitted from recording. There are no further missing observations.

## Predictions

If it is the case that East Asians pay relatively less attention to internal dispositions of the causer and more to situational factors in their causal attributions compared to Westerners, as suggested by the line of research starting with Morris \& Peng (1994), both Causer intentionality and Causee intentionality should play a less predictive role in the ratings of the Chinese participants than in those of the Spanish participants. On the other hand, Le Guen et al's (2015) findings suggest that Causer intentionality may play an even greater role in the Yucatecans' responsibility assignments than in those of either of the other two groups.

## Analysis

Figure 3 shows the mean Causer responsibility ratings by population, suggesting small but significant differences (Mandarin $\mathrm{M}=7.37, \mathrm{SD}=2.09$; Spanish $\mathrm{M}=5.98, \mathrm{SD}=$ 3.14; Yucatec $\mathrm{M}=6.67, \mathrm{SD}=3.24$ ). Figure 4 presents a breakdown by Causer intentionality, suggesting that the Mayan and Spanish participants, but not the Chinese participants, assigned more responsibility to intentional than to unintentional Causers, as predicted.


Figure 3: Mean Causer responsibility rating by population

A linear mixed effects regression model was fitted, using the lme4 package in R and treating the Causer responsibility rating as dependent variable. The rating was treated as a continuous rather than ordinal (categorical) variable since the participants expressed it through the proportional allocation of the tokens rather than through labeled categories. As fixed factors were included Population, $\mathrm{I} \rightarrow \mathrm{A}$, $\mathrm{I} \rightarrow \mathrm{O}$, Causee Intentionality, and all binary interactions between Population and the intentionality variables. Random intercepts were added for participant and stimulus clip (formula: CR.Responsibility $\sim$ Population $+\mathrm{I} \rightarrow \mathrm{A}+$ $\mathrm{I} \rightarrow \mathrm{O}+$ Intentionality.of.CE + Population $* \mathrm{I} \rightarrow \mathrm{A}+$ Population * $\mathrm{I} \rightarrow \mathrm{O}+$ Population * Intentionality.of. $\mathrm{CE}+(1$ $\mid$ Participant.ID $)+(1 \mid$ Clip.Code $)$ ). The three intentionality variables were coded binarily.


Figure 4: Mean Causer responsibility rating by population and Causer intentionality
Table 2 summarizes the effects. Due to the multitude of models, the confidence level should be Bonferroni-adjusted to $\mathrm{p}<.001$. Effects outside this level should be ignored. There were main effects of population and causee intentionality and significant interactions between population and the $\mathrm{I} \rightarrow \mathrm{A}$ link and between population and causee intentionality. There was neither a main effect of the $\mathrm{I} \rightarrow \mathrm{O}$ link nor did it feature in any significant interaction. Collinearity of factors above . 6 occurred exclusively between the absence of $\mathrm{I} \rightarrow \mathrm{A}$ and the absence of $\mathrm{I} \rightarrow \mathrm{O}$ (to be expected, as in the design of the items, the former entails the latter, i.e., we did not include scenes in which an unintended action accidentally yielded an intended outcome; cf. Table 1) and between some of the interactions and either their component factors or interactions sharing a factor.

## Discussion

The presence of an unintentional (nonvolitional) Causee significantly boosted attribution of responsibility to the causer across populations. This is of course eminently plausible and thus can be seen as very basic support of internal validity.

In line with what Figure 1 suggests, the Spanish and Yucatec participants' ratings were significantly lower than those of the Chinese participants, although the differences
were quite small. This effect is plausibly attributable to the Spanish and Yucatecan participants having paid more attention to the intentionality/volitionality of the Causee than the Chinese participants, in line with the hypothesis that intentionality plays a lesser role in the Chinese participants' attributions.

Table 2. Significant factors in the regression model with

| Factor$(\mathrm{CE}=\text { Causee })$ | Baseline population Chinese Spanish Yucatec |  |  |
| :---: | :---: | :---: | :---: |
|  |  |  |  |
| Chinese | N/A | *** 5.1 |  |
| Spanish | ***-1.2 | N/A | **-1.1 |
| Yucatec |  | ** 1.2 | N/A |
| No I $\rightarrow$ O |  |  |  |
| No I $\rightarrow$ A |  | ***-3.6 | *-2.2 |
| CE unintentional | ** 1.7 | *** 2.8 | ** 1.6 |
| Chinese * No I $\rightarrow \mathrm{O}$ | N/A |  |  |
| Spanish * No I $\rightarrow$ O |  | N/A |  |
| Yucatec * No I $\rightarrow$ O |  |  | N/A |
| Chinese * No I $\rightarrow$ A | N/A | ***3.1 | * 1.8 |
| Spanish * No I $\rightarrow$ A | *** -3.1 | N/A | *-1.4 |
| Yucatec * No I $\rightarrow$ A | *-1.8 | * 1.4 | N/A |
| Chinese * CE unintentional | N/A | **-1.1 |  |
| Spanish * CE unintentional | ** 1.1 | N/A | ** 1.2 |
| Yucatec * CE unintentional |  | **-1.2 | N/A |

Crucial for the evaluation of our predictions are the interactions between population and the causer intentionality variables. In line with our predictions, we found that absence of the $\mathrm{I} \rightarrow \mathrm{A}$ link - i.e., an unintended action on the causer's part - strongly positively interacted with Chinese against Spanish as baseline and vice versa. In other words, the Chinese participants' responsibility scores for unintentional causers were significantly higher than the Spanish participants, in line with predictions. The relevant interactions with Yucatec were not significant by the Bonferroni criterion.

Turning to unintentional causees, this factor showed a significant positive interaction with Spanish against Chinese as baseline. The Spanish participants rated Causer intentionality higher than the Chinese participants when the Causee was unintentional. This suggests that the Spanish participants paid more attention to the intentionality of the causee than the Chinese participants did - again in line with predictions. When Spanish was the baseline, significant negative interactions with both Chinese and Yucatec were found (and when Yucatec was the baseline, of course the inverse positive interaction with Spanish materialized). This suggests that the Spanish participants rated Causer responsibility relatively higher when the Causee was acting involuntarily than did the Yucatec participants. It thus appears that Causee intentionality played a greater role for the Spanish participants than for the Yucatecans.

The difference between the populations was specifically located in scenes that lacked $\mathrm{I} \rightarrow \mathrm{A}$, in other words, scenes in
which an unintended action caused a certain result (which in our stimuli was likewise unintended). In all instances, the relevant actions of the Causer involved spontaneous bodily functions (yawning, sneezing, losing balance, fainting). Such acts caused the Spanish and Yucatec participants, but not the Chinese participants, to rate the Causer's responsibility lower. In contrast, we found no significant effect for scenes in which intended actions had unintended consequences (e.g., causing somebody to leave the room by singing poorly or causing somebody to knock down a cup tower by running into them while dragging a cart backwards into the room).

## General discussion

We did not find greater sensitivity to intentionality among our Mayan participants than among the other two groups. There is thus no apparent evidence of remnants of 'magical thinking' in our traditional non-Western population, contrary to the findings of Le Guen et al (2015). However, as predicted by a line of studies in social psychology, the Chinese participants in our experiment appear to have been less inclined to factor the intentionality of both the Causer and the Causee into their attributions of responsibility than the Spanish participants. To our knowledge, this has been demonstrated here for the first time in terms of relative responsibility distribution between competing actors.

## Future research

Whether our evidence of culture-specificity in causal attributions submits to the explanatory mechanisms in terms of folk theories of group agency vs. individual agency and/or context sensitivity invoked in research tradition that motivated the present study remains to be seen. A crucial test will be the extension of the investigation to other populations of the supposed 'egocentric' vs. 'sociocentric' types. We are currently preparing to test further populations.

A question we intend to take up in the next phase of our investigation is whether the apparent difference in causal attribution also manifests itself in the grammatical means used when members of the different groups talk about causality. It has often been observed that more agentive causal chains tend to be represented more compactly in language than less agentive ones. Thus, Sally made Floyd knock over the cup tower implicates, but does not entail, that Sally acted intentionally, whereas Sally bumped into Floyd and he knocked over the cup tower does not (McCawley 1976). This predicts that members of sociocentric societies may use relatively more compact representations of lowintentionality scenarios than members of egocentric societies. If confirmed, this could suggest a relationship between grammars and folk theories of agency.
For the treatment of agency in linguistic theories, two responses to our findings are conceivable: retain a universal notion of agency, which then plays a variable role in the grammars of different languages, or replace it with culturespecific concepts of agency, which then would stand a chance of playing a more uniform role across languages.

## Acknowledgments

The material presented here is based upon work supported by the National Science Foundation under Grant No. BCS153846 and BCS-1644657, ‘Causality Across Languages'; PI J. Bohnemeyer. We would like to thank the members of the University at Buffalo Semantics Typology Lab for assistance with the creation of the stimuli (in particular, Anastasia Stepanova, Alexandra Lawson, Randi Moore, and Karl Sarvestani) and the members of the Beihang Research Group for Event Representation and Cognition for their assistance in testing the Chinese participants (specifically, Enirile, Hongxia Jia, Fuyin Li, Jinmei Li, Sai Ma, Chenxi Niu , and Mengmin Xu ). We are grateful to Dare Baldwin, Sieghard Beller, Andrea Bender, and Bertram Malle for advice.

## References

Alicke, M. D. (2000). Culpable control and the psychology of blame. Psychological Bulletin, 126(4), 556-574. https://doi.org/10.1037/0033-2909.126.4.556
Chalnick, A., \& Billman, D. (1988). Unsupervised learning of correlational structure. Proceedings of the tenth annual conference of the cognitive science society (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.
Chiu, C.-Y., Morris, M. W., Hong, Y.-Y., \& Menon, T. (2000). Motivated cultural cognition: The impact of implicit cultural theories on dispositional attribution varies as a function of need for closure. Journal of Personality and Social Psychology, 78(2), 247-259. http://doi.org/10.1037//0022-3514.78.2.247
Choi, I., \& Nisbett, R. E. (1998). Situational salience and cultural differences in the correspondence bias and in the actor-observer bias. Personality and Social Psychology Bulletin, 24, 949-960.
Choi, I., Nisbett, R. E., \& Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. Psychological Bulletin, 125(1), 47-63. http://doi.org/10.1037/0033-2909.125.1.47
Comrie, B. (1981). Language universals and linguistic typology: Syntax and morphology. Chicago: University of Chicago Press.
Croft, W. (1987). Categories and relations in Syntax: The clause-level organization of information. Ph.D. dissertation, Stanford University.
Dixon, R. M. W. (2000). A typology of causatives: form, syntax and meaning. In R. M. W. Dixon \& A. Y. Aikhenvald (eds.), Changing valency: Case studies in transitivity. Cambridge: Cambridge University Press. 30-83.
Dowty, D. R. (1991). Thematic proto-roles and argument selection. Language 67: 547-619.
Evans-Pritchard E. E. (1937). Witchcraft, Oracles and Magic among the Azande. Oxford: Oxford University Press.
Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum \& J. Feldman (Eds.), Computers and thought. New York: McGraw-Hill.

Foley, W. A. \& R. D. Van Valin, Jr. (1984). Functional syntax and universal grammar. Cambridge: Cambridge University Press.
Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. Cognition and Brain Theory, 6, 287-317.
Lakoff, G. (1977). Linguistic gestalts. Papers from the Thirteenth Regional Meeting of the Chicago Linguistic Society. 236-87
Le Guen, O., Samland, J., Friedrich, T., Hanus, D., \& Brown, P. (2015). Making sense of (exceptional) causal relations. A cross-cultural and cross-linguistic study. Frontiers in Psychology, 6(OCT), 1-16. http://doi.org/10.3389/fpsyg.2015.01645
McCawley, J. (1976). Remarks on what can cause what. In M. Shibatani (ed.), Syntax and Semantics VI: The grammar of causative constructions. New York, NY: Academic Press. 117-129.
Maddux, W. W., \& Yuki, M. (2006). The "ripple effect": cultural differences in perceptions of the consequences of events. Personality and Social Psychology Bulletin, 32, 669-683. http://doi.org/10.1177/0146167205283840
Matlock, T. (2001). How real is fictive motion? Doctoral dissertation, Psychology Department, University of California, Santa Cruz.
Menon, T., Morris, M. W., Chiu, C., \& Hong, Y. (1999). Culture and the construal of agency: Attribution to individual versus group dispositions. Journal of Personality and Social Psychology, 76(5), 701-717. http://doi.org/10.1037/0022-3514.76.5.701
Morris, M. M. W., \& Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. Journal of Personality and Social Psychology, 67(6), $949-971$. http://doi.org/10.1037/00223514.67.6.949

Newell, A., \& Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Ohlsson, S., \& Langley, P. (1985). Identifying solution paths in cognitive diagnosis (Tech. Rep. CMU-RI-TR-852). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.
Peng, K., \& Knowles, E. D. (2003). Culture, education, and the attribution of physical causality. Personality and Social Psychology Bulletin, 29(1991), 1272-1284. http://doi.org/10.1177/0146167203254601
Shibatani, M. \& P. Pardeshi. (2002). The causative continuum. In M. Shibatani (ed.), The grammar of causation and interpersonal manipulation. Amsterdam: Benjamins. 85-126.
Shrager, J., \& Langley, P. (Eds.) (1990). Computational models of scientific discovery and theory formation. San Mateo, CA: Morgan Kaufmann.
Van Valin, Robert D., and David Wilkins. "The case for "effector": Case roles, agents, and agency revisited." Grammatical constructions: Their form and meaning (1996): 289-322.

# Explaining Enculturated Cognition 

Regina E. Fabry (regina.fabry@psychol.uni-giessen.de)<br>Justus Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Strasse 10 F 35394 Giessen, Germany


#### Abstract

Many of our cognitive capacities are shaped by enculturation. Enculturation is the temporally extended transformative acquisition of cognitive practices such as reading, writing, and mathematics. They are embodied and normatively constrained ways to interact with epistemic resources (e.g., writing systems, number systems). Enculturation is associated with significant changes of the organization and connectivity of the brain and of the functional profiles of embodied actions and motor programs. Furthermore, it has a socio-culturally structured dimension, because it relies on cumulative cultural evolution and on the socially distributed acquisition of cognitive norms governing the engagement with epistemic resources. This paper argues that we need distinct, yet complementary levels of explanation and corresponding temporal scales. This leads to explanatory pluralism about enculturated cognition, which is the view that we need multiple perspectives and explanatory strategies to account for the complexity of enculturation.


Keywords: enculturation; neural plasticity; neural reuse; embodied cognition; cognitive niche construction; cumulative cultural evolution; cultural learning; reading acquisition; explanatory pluralism

## Introduction

Many cognitive processes are shaped and facilitated by our successful acquisition of cognitive practices such as reading, writing, or mathematics. Cognitive practices are evolutionarily recent, embodied interactions with writing systems, number systems, and various other epistemic resources in our local environment (Fabry, 2017; Menary, 2015). Given the evolutionary recency of cognitive practices, with reading, writing, and mathematics dating back to approx. 3000 BC (Donald, 1991), the question arises how their acquisition can be explained from an empirically informed perspective. The purpose of the present paper is to sketch an explanatory framework that can help close the current gap in thinking about the phylogenetic and ontogenetic emergence of cognitive practices. The proposal is that competence in the performance of cognitive practices is the result of enculturation.

Enculturation is defined as the acquisition of cognitive practices during ontogeny. It is a temporally extended process that augments and transforms our overall cognitive capacities. There are two background assumptions that inform the conceptualization of enculturation. First, it is committed to a robust variant of embodied cognition (Menary, 2015). On this view, the embodied interaction with the local environment plays an indispensable functional role
in at least some cognitive processes. In the present context, embodiment is understood as the bodily manipulation of epistemic resources (Menary, 2010; Rowlands, 1999), e.g. by initiating and executing eye movements and hand movements. Second, the present account of enculturation rests on on the assumption that cognitive practices are cases of strong embedded cognition (Menary, 2015). This amounts to the idea that at least some cognitive processes are realized by the integration of cerebral, extra-cerebral bodily, and environmental components. We will see in the course of this paper that the theoretical commitments to strong embodied and embedded cognition are supported by empirical research.

## Shaping the Brain and the Rest of the Body

Enculturation is associated with significant changes to the organization and connectivity of the brain and to the functional profiles of embodied actions and motor programs. Learning driven plasticity (LDP) is a potent principle governing ontogenetic brain development, according to which structural changes of the organization and connectivity of brain areas lead to new neuronal functions (Ansari, 2012; Menary, 2015). LDP is not an open-ended process that leads to the unlimited realization of new neuronal circuits. Rather, it is constrained by the functional biases of certain cortical units that contribute to the development of new neuronal connections. This is suggested by empirical and conceptual work on neural reuse (Anderson, 2010, 2015) and neuronal recycling (Dehaene, 2005, 2010). The idea is that especially evolutionarily recent cognitive practices need to allocate and re-exploit already existing structural and functional connections of brain areas and integrate them into new neuronal circuitry. The scope of neural reuse in each particular case depends on the functional and structural biases of specific brain areas and on the functional proximity of uses to which these areas can be put (Anderson, 2015).

Neural reuse - and its component process of neuronal recycling that is associated with the acquisition of reading, writing, or mathematics - is a guiding principle of LDP. This is especially important in cases of enculturation. The reason is that it helps explain how cognitive practices can be acquired, given that there was not sufficient evolutionary time for the development of brain circuits unequivocally and exclusively dedicated to them.

The assumption that enculturation is defined as the acquisition of embodied cognitive practices gives rise to the idea that LDP is complemented by a genuinely bodily form of transformation. According to the principle of learning driven bodily adaptability (LDBA), new bodily ways to
interact with the socio-culturally structured environment emerge in the course of enculturation. LDBA guides the ontogenetic trajectory of skilled motor action. The resulting development of new motor patterns and action routines is constrained by extra-cerebral bodily biases, e.g., by the functional potential of the overall morphology of human bodies and their constitutive parts (Dounskaia, van Gemmert, \& Stelmach, 2000; Furuya \& Altenmüller, 2015; Phillips, Ogeil, \& Best, 2009). The employment and allocation of bodily resources available to the human organism bring about the embodied adaptation to new cognitive practices in close co-ordination with LDP. In this sense, the functional biases of brain areas are complemented by the biases of functional units of the rest of the body. It is important to note that the overall possibility space of cognitive practices is not only defined, but also delimited by the anatomical and physiological properties of the human body.

## Interacting with the Cognitive Niche

The present account of enculturated cognition is committed to the view that cognitive practices are distributed across the brain, the rest of the body, and the local environment. This leads to the assumption that our understanding of LDP and LDBA needs to be complemented by considerations of the embodied interaction of cognitive systems with the cognitive niche. The cognitive niche can be defined as the incrementally, trans-generationally structured socio-cultural environment that provides human organisms with epistemic resources for the completion of cognitive tasks (Bertolotti \& Magnani, 2016; Clark, 2006, 2008; Kendal, 2011; Sterelny, 2003, 2012; Stotz, 2010). Examples of resources in the cognitive niche include writing systems, number systems, and notational symbol systems. In addition, the cognitive niche is also characterized by socio-cultural institutions like kindergartens, schools, or universities. Cognitive practices are shared by a large number of individuals in the cognitive niche. Therefore, the skilful performance of cognitive practices is constrained by sets of cognitive norms. These norms regulate the interaction with epistemic resources (Menary, 2007, 2016). They need to be learned and automatized in the course of enculturation (Menary, 2013). Since cognitive practices are socio-cultural phenomena, their acquisition is itself a socio-culturally structured process. This process is characterized by scaffolded learning (Clark, 1997; Estany \& Martínez, 2014; Wood, Bruner, \& Ross, 1976). The notion of scaffolded learning refers to the idea that the acquisition of a cognitive practice is a systematic process of novice-expert interaction in the cognitive niche. This interaction is structured by the current developmental stage of the novice and a specific set of skills and knowledge that needs to be acquired in the long run (Vygotsky, 1978).

## Socio-Cultural Learning

It is conceivable that scaffolded learning is the result of evolutionary processes that have shaped specific types of
socio-culturally transmitted human cognitive capacities (Boyd, Richerson, \& Henrich, 2011; Henrich, 2016). According to Kline's (2015) recent framework for the investigation of teaching, direct active teaching is of vital importance for human scaffolded learning. It is defined by the "[...] manifestation of relevant information by the teacher to the pupil, as well as the pupil's interpretation of that information as generalizable" (Kline, 2015, p. 12). In contrast to other forms of teaching and learning that are ubiquitous in the animal kingdom, e.g., social tolerance or opportunity provisioning, direct active teaching is specific to humans. It is likely to have co-evolved with genuinely human ways of cognitive niche construction. If correct, cumulative cultural evolution and scaffolded learning, where the latter might be the result of trans-generationally emerged socio-cultural processes (Heyes, 2012), mutually influence and constrain each other.

In sum, enculturation is a complex phenomenon that requires the synthesis of several explanatory components targeting the cerebral, the extra-cerebral bodily, and the socio-cultural dimensions of cognitive practices. This leads the question how we can combine these components in such a way that we will end up with an explanation of enculturation that is both conceptually coherent and empirically plausible without running risk of committing mereological fallacies. The suggestion of the present paper is to analyse enculturation at three levels of explanation and corresponding time scales.

## Levels and Time Scales of Explanation

Levels of explanation are defined by the conceptual and/or empirical tools, by the research questions, and by the individuation and operationalization that are employed to account for (a component of) a certain target phenomenon (Dennett, 1969; Drayson, 2012, 2014; Metzinger, 2013). In this sense, levels specify the scope and the epistemic tools of explanation. Following Drayson (2012), it is reasonable to include Kim's distinction of the vertical and the horizontal into our meta-theoretical consideration of explanation: "The term 'vertical' is meant to reflect the usual practice of picturing micro-macro levels of a vertical array, with the micro underpinning the macro. In contrast, we usually represent diachronic causal relations on a horizontal line, from past (left) to future (right) [...]" (Kim, 2005, p. 36). This distinction adds a temporal dimension to the individuation of levels of explanation. Accordingly, levels of explanation correspond to specific temporal scales. For the present considerations, we can distinguish three levels of explanation and corresponding time scales.

First, on a sub-personal level of explanation, we can provide an account of the cerebral and extra-cerebral bodily functions that underlie the acquisition of cognitive practices. On this level of explanation, we focus on the corresponding physiological time scale. This temporal scale is defined by time intervals that have a duration of hundreds of milliseconds to several seconds. The time intervals are
determined by the full range of electrophysiological, neuroimaging, and eye-tracking paradigms and the resulting statistical analyses. This explanatory component is concerned with the consideration of LDP and LDBA.

Second, on a personal level of explanation, we can investigate the diachronic unfolding of specific changes of the human organism as a whole that characterize the ontogenetic process of enculturation. This level of explanation corresponds to an organismic timescale. First, we can develop an account of the temporal unfolding of the novice's ongoing interaction with experts in a particular domain. This helps specify the various stages of the acquisition of cognitive practices and the ways in which it relies on the scaffolding by other cognitive agents. Second, from this perspective we can provide an account of the properties of epistemic resources in the cognitive niche as they are relevant for the acquisition and on-going realization of cognitive practices. Finally, we can identify the set of cognitive norms that are acquired and applied in the course of the acquisition of a certain cognitive practice.

Third, the supra-personal level of explanation is also relevant for a full-fledged account of enculturation. The reason is that contemporary cases of enculturation are rendered possible by evolved biological principles and the inter-generational transmission of practices, skills, and epistemic resources. On this view, enculturation is constituted by the interdependence of evolved cerebral and extra-cerebral bodily biases and of the on-going large-scale process of cognitive niche construction. Therefore, it seems reasonable to introduce an additional type of diachronic explanation, namely a supra-personal level of explanation that focuses on an evolutionary time scale comprising hundreds to thousands of years.

## Reading Acquisition: A Paradigm Case of Enculturation

To illustrate the considerations, conceptual distinctions, and the meta-theoretical assessment of the account of enculturation, I will now consider reading acquisition as a paradigm case of enculturation. At first sight, reading poses a challenge to researchers, because it requires an explanation of how we are able to acquire reading, given that there was not sufficient evolutionary time for dedicated brain areas to develop. Dehaene (2010) refers to this as the "reading paradox."

On a sub-personal level of explanation and at a physiological time scale, this paradox can be solved by considering LDP and its guiding principle, i.e., neural reuse. There is now much empirical evidence suggesting that the brain undergoes significant plastic changes in the course of reading acquisition at times $\mathrm{t}_{1}, \mathrm{t}_{2}$, and $\mathrm{t}_{3}$. First, many studies

[^63]and theoretical evaluations emphasize the crucial importance of the left ventral occipito-temporal (vOT) area (Dehaene, 2005, 2010; McCandliss, Cohen, \& Dehaene, 2003; Price \& Devlin, 2003, 2004; Vogel, Petersen, \& Schlaggar, 2014). 1 Recent studies indicate that its activation level peaks in beginning readers and that its decrease, by way of comparison with the level of neuronal activation at $t_{1}$, is associated with reading proficiency (Ben-Shachar, Dougherty, Deutsch, \& Wandell, 2011; Brem et al., 2010; Maurer et al., 2006). Second, there is a significant increase of functional connectivity between the left vOT area and lefthemispheric frontal and temporal areas that are reliably associated with language processing and production (Dehaene et al., 2010; Gaillard, Balsamo, Ibrahim, Sachs, \& Xu, 2003; Turkeltaub, Gareau, Flowers, Zeffiro, \& Eden, 2003). It is in virtue of LDP that new structural and functional connections can be realized as a solution to new and challenging processing needs.

LDP is complemented by LDBA. In the case of reading acquisition, LDBA is mainly realized by the developmental trajectory of eye movements. In general, eye movement patterns in reading are constituted by the alternation between fixations and saccades (Rayner, 2009; Rayner et al., 2001, 2007). Eye movements are necessary because of the acuity limitations of the visual field. The functional biases of the ocular-motor system, e.g., the saccadic latency and the saccadic span, constrain the developmental trajectory of reading. Research paradigms employing eye-tracking methodologies are specifically interested in evaluating the span or amplitude of saccades, the duration of fixations, and the landing positions or locations of fixations with regard to certain target words embedded in syntactically and semantically structured linguistic items. Comparisons of novice and proficient readers reveal that proficient readers display a decrease of fixation durations, refixations (i.e., several fixations targeted at the same word), as well as an increase of saccadic amplitudes (Huestegge et al., 2009; Joseph \& Liversedge, 2013; Seasseau et al., 2013). These findings suggest that ocular-motor patterns adapt to the demands and requirements of processing structured linguistic material.

On a personal level of explanation and at an organismic time scale, reading acquisition is characterized by scaffolded learning and structured novice-teacher interactions. In the case of alphabetic writing systems, reading instruction puts a particular emphasis on phonics instruction (Rayner, Foorman, Perfetti, Pesetsky, \& Seidenberg, 2001). Phonics instruction conveys the alphabetic principle, according to which graphemic units of an alphabetic writing system correspond to phonemic units of the target language (Castles \& Coltheart, 2004; Dehaene, 2010; Snowling, 2000; Ziegler \& Goswami, 2006). In the vast majority of cases, the alphabetic principle can only be understood and applied if

Lipka, 2011). This provides further support for the idea of neural reuse, because it establishes that one particular brain area has a certain bias that makes it suitable to contribute to functionally distinct, yet partly overlapping neural circuits.
novices receive extensive instruction and scaffolded tutorials provided by teachers and other caregivers. These tutorials provide detailed information about the cognitive norms underlying reading, e.g., by progressively increasing the complexity of phoneme-grapheme correspondences in the training materials. In addition, there are other types of metalinguistic awareness that need to be made explicitly available to novice readers. Beginning readers are already proficient speakers of their native language and are able to apply fluently syntactic, semantic, and pragmatic norms in their everyday conversations. However, they usually lack the explicit insight that utterances are made up of sentences and that sentences are constituted by the combination of words (Frith, 1985; Rayner et al., 2001). To novices, these basic properties have to be made explicitly available in order to put them in the position to apply the knowledge about it automatically and fluently at later stages of reading acquisition. Furthermore, novices need to be acquainted with the norm that alphabetic writing systems are decoded from left to right and from the top to the bottom of a page (Dehaene, 2010). In sum, explicit reading instruction is a good example of scaffolded learning and of the socioculturally structured transmission of knowledge and skills.

The history of writing systems is a good example of cumulative cultural evolution (Henrich, 2016). This example needs to be approached on a supra-personal level of explanation and at an evolutionary time scale. The first writing system we know of is the cuneiform system. It dates back to approx. 3000 BC and was pictorial in origin. The cuneiform system was cumulatively refined in the service of an accurate representation of abstract ideas and relations that were especially relevant for trade and the organization of social communities. Furthermore, the functional biases of the brain and the rest of the body constrained the properties of symbols, e.g., the arrangement of lines or inter-letter spacing (Dehaene, 2010). Linearity, the grouping of symbols, and grammatical norms were not pre-given properties of early writing systems. Rather, they gradually evolved over hundreds of years. Tracing back the development of alphabetic writing systems, Donald characterizes the evolutionary trajectory as a "[...] progression from a primarily visual medium, inventing completely new representations like lists of numbers, to a medium which, increasingly, tried to map the narrative products of the language system" (Donald, 1991, 289). In sum, then, contemporary writing systems are the direct result of cumulative cultural evolution. They were afforded by the socially structured need of a system that can represent transactions, relations, genealogies, and so forth.

## Towards Explanatory Pluralism

The previous considerations suggest that we need at least three levels of explanation and corresponding temporal scales to unveil the complex dynamics that give rise to enculturation. This suggestion is at odds with explanatory monism, according to which there will always be one and
only one explanation of a certain target phenomenon or a specific set of target phenomena in the long run (Colombo \& Wright, 2017; Kellert, Longino, \& Waters, 2006). It is informed by unificationism and by reductionism. Unificationism is the idea that there will always be one set of principles that is able to unify previously distinct kinds of explanation targeted at a certain phenomenon. Reductionism about theory formation is the idea that we will gain new knowledge if we discover low-level principles to which previously entertained higher-level explanations can be reduced (Colombo \& Wright, 2017). The present analysis of enculturation and the distinction of complementary levels of explanation and corresponding time scales leads to the view that it is at least unlikely, if not impossible, that unificationism and reductionism are meta-theoretical principles that will lead to a complete and exhaustive account of enculturation. This is the reason why the present account of enculturated cognition is an example of explanatory pluralism about theory formation in the cognitive sciences. Explanatory pluralism is the view that there will always be more than one and only one explanation of a specific target phenomenon (Van Bouwel, Weber, \& de Vreese, 2011; Dale, 2008; de Jong, 2001). This stance towards enculturation promises to arrive at a better understanding of the complexity of the phylogenetic and ontogenetic development of cognitive practices and of the temporal unfolding of enculturated organism-niche interactions. Thus, the positive proposal is that the phylogenetic and ontogenetic components of enculturation on sub-personal, personal, and suprapersonal levels of explanation are required for the prospect of a full-fledged and complete account of enculturation. The consideration of reading acquisition as a paradigm case of enculturation can lend support to the idea that we need the pluralistic explanatory stance in order to account for the vast set of empirical results and empirically informed considerations applying to enculturation.

## Concluding Remarks

Enculturation is a temporally extended process that transforms our overall cognitive capacities. In this paper, I have argued that enculturation is a complex phenomenon that needs to be approached on at least three levels of explanation and corresponding time scales. The reason is that enculturation spreads across the brain, the rest of the body, and the cognitive niche. Explanatory pluralism allows us to do justice to these dynamics, because it provides us with an explanatory strategy that is able to track the ontogenetic and phylogenetic component processes of enculturation. The application of this strategy to the cases of reading acquisition shows that the present account of enculturation has the conceptual resources to connect initially disparate lines of empirical research. The suggestion is that the present account of enculturation promises to provide us with a better understanding of the ways in which our cognitive processes are shaped and re-shaped by the delicate interaction of the brain, the rest of the body, and the cognitive niche.

## References

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. Behavioral and Brain Sciences, 33(4), 245-266.
Anderson, M. L. (2015). After phrenology: Neural reuse and the interactive brain. Cambridge, Mass: MIT Press.
Ansari, D. (2012). Culture and education: New frontiers in brain plasticity. Trends in Cognitive Sciences, 16(2), 93-95.
Ben-Shachar, M., Dougherty, R. F., Deutsch, G. K., \& Wandell, B. A. (2011). The development of cortical sensitivity to visual word forms. Journal of Cognitive Neuroscience, 23(9), 2387-2399.
Bertolotti, T., \& Magnani, L. (2016). Theoretical considerations on cognitive niche construction. Synthese, 1-23.
Van Bouwel, J., Weber, E., \& de Vreese, L. (2011). Indispensability arguments in favour of reductive explanations. Journal for General Philosophy of Science, 42(1), 33-46.
Boyd, R., Richerson, P. J., \& Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. Proceedings of the National Academy of Sciences, 108(Supplement 2), 10918-10925.
Brem, S., Bach, S., Kucian, K., Guttorm, T. K., Martin, E., Lyytinen, H., Brandeis, D., et al. (2010). Brain sensitivity to print emerges when children learn letter-speech sound correspondences. Proceedings of the National Academy of Sciences, 107(17), 7939-7944.
Castles, A., \& Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? Cognition, 91(1), 77-111.
Clark, A. (1997). Being there: Putting brain, body, and world together again. Cambridge, Mass.: MIT Press.
Clark, A. (2006). Language, embodiment, and the cognitive niche. Trends in Cognitive Sciences, 10(8), 370-374.
Clark, A. (2008). Supersizing the mind: Embodiment, action, and cognitive extension. New York: Oxford University Press.
Colombo, M., \& Wright, C. (2017). Explanatory pluralism: An unrewarding prediction error for free energy theorists. Brain and Cognition, 112, 3-12.
Dale, R. (2008). The possibility of a pluralist cognitive science. Journal of Experimental \& Theoretical Artificial Intelligence, 20(3), 155-179.
Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. In S. Dehaene, J.-R. Duhamel, M. D. Hauser, \& G. Rizzolatti (Eds.), From monkey brain to human brain: A Fyssen Foundation Symposium (pp. 133-157). Cambridge, Mass: MIT Press.
Dehaene, S. (2010). Reading in the brain: The new science of how we read. New York: Penguin Books.
Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Filho, G. N., Jobert, A., Dehaene-Lambertz, G., et al. (2010). How learning to read changes the cortical networks for vision and language. Science, 330(6009), 1359-1364.
Dennett, D. C. (1969). Content and consciousness. London, New York: Routledge \& K. Paul.
Donald, M. (1991). Origins of the modern mind: Three stages in the evolution of culture and cognition. Cambridge, Mass: Harvard University Press.
Dounskaia, N., van Gemmert, A. W. A., \& Stelmach, G. E. (2000). Interjoint coordination during handwriting-like movements. Experimental Brain Research, 135(1), 127-140.

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. Philosophical Perspectives, 26(1), 1-18.
Drayson, Z. (2014). The personal/subpersonal distinction. Philosophy Compass, 9(5), 338-346.
Estany, A., \& Martínez, S. (2014). "Scaffolding" and "affordance" as integrative concepts in the cognitive sciences. Philosophical Psychology, 27, 98-111.
Fabry, R. E. (2017). Betwixt and between: The enculturated predictive processing approach to cognition. Synthese, 1-36.
Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. E. Patterson, J. C. Marshall, \& M. Coltheart (Eds.), Surface dyslexia.Neuropsychological and cognitive studies of phonological reading (pp. 301-330). Hillsdale, N.J.: Erlbaum.
Furuya, S., \& Altenmüller, E. (2015). Acquisition and reacquisition of motor coordination in musicians. Annals of the New York Academy of Sciences, 1337(1), 118-124.
Gaillard, W. D., Balsamo, L. M., Ibrahim, Z., Sachs, B. C., \& Xu, B. (2003). fMRI identifies regional specialization of neural networks for reading in young children. Neurology, $60(1)$, 94-100.
Henrich, J. P. (2016). The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter. Princeton: Princeton University Press.
Heyes, C. (2012). Grist and mills: On the cultural origins of cultural learning. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1599), 2181-2191.
de Jong, H. L. (2001). Introduction: a symposium on explanatory pluralism. Theory \& Psychology, 11(6), 731-735.
Kellert, S. H.; Longino, H. E.; Waters, C. K. (2006). Introduction: the pluralist stance. In C. K. Kellert, S., Longino, H., \& Waters (Ed.), Minnesota studies in philosophy of science, vol. 19: scientific puralism (pp. vii-xxix). Minneapolis: University of Minnesota Press.
Kendal, J. R. (2011). Cultural niche construction and human learning environments: Investigating sociocultural perspectives. Biological Theory, 6(3), 241-250.
Kim, J. (2005). Physicalism, or something near enough. New Jersey: Princeton University Press.
Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. Behavioral and Brain Sciences, 38, 1-70.
Ludersdorfer, P., Kronbichler, M., \& Wimmer, H. (2015). Accessing orthographic representations from speech: The role of left ventral occipitotemporal cortex in spelling. Human brain mapping, 36(4), 1393-1406.
Maurer, U., Brem, S., Kranz, F., Bucher, K., Benz, R., Halder, P., Steinhausen, H.-C., et al. (2006). Coarse neural tuning for print peaks when children learn to read. NeuroImage, 33(2), 749-758.
McCandliss, B. D., Cohen, L., \& Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. Trends in Cognitive Sciences, 7(7), 293-299.
Menary, R. (2007). Cognitive integration: Mind and cognition unbounded. Basingstoke, New York: Palgrave Macmillan.
Menary, R. (2010). Dimensions of mind. Phenomenology and the Cognitive Sciences, 9(4), 561-578.
Menary, R. (2013). The enculturated hand. In Z. Radman (Ed.), The hand, an organ of the mind: What the manual tells the mental (pp. 349-367). Cambridge, Mass: MIT Press.
Menary, R. (2015). Mathematical cognition: A case of enculturation. In T. Metzinger \& J. M. Windt (Eds.), Open MIND (pp. 1-20). Frankfurt am Main: MIND Group.
Menary, R. (2016). Pragmatism and the pragmatic turn in cognitive
science. In D. Engel, Andreas K., Friston, K., \& Kragic (Ed.), Where is the action? The pragmatic turn in cognitive science (pp. 219-237). Cambridge, Mass: MIT Press.
Metzinger, T. (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. Frontiers in Psychology, 4.
Phillips, J. G., Ogeil, R. P., \& Best, C. (2009). Motor constancy and the upsizing of handwriting. Human Movement Science, 28(5), 578-587.
Price, C. J., \& Devlin, J. T. (2003). The myth of the visual word form area. NeuroImage, 19(3), 473-481.
Price, C. J., \& Devlin, J. T. (2004). The pro and cons of labelling a left occipitotemporal region "the visual word form area." NeuroImage, 22(1), 477-479.
Purcell, J. J., Jiang, X., \& Eden, G. F. (2017). Shared orthographic neuronal representations for spelling and reading. NeuroImage, 147, 554-567.
Purcell, J. J., Napoliello, E. M., \& Eden, G. F. (2011). A combined fMRI study of typed spelling and reading. Neuroimage, 55(2), 750-762.
Rapp, B., \& Lipka, K. (2011). The literate brain: The relationship between spelling and reading. Journal of cognitive neuroscience, 23(5), 1180-1197.
Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., \& Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. Psychological Science in the Public Interest, 2(2), 31-74.
Rowlands, M. (1999). The body in mind: Understanding cognitive processes. Cambridge: Cambridge University Press.
Snowling, M. J. (2000). Dyslexia (2nd ed.). Malden, MA: Blackwell Publishers.
Sterelny, K. (2003). Thought in a hostile world: The evolution of human cognition. Malden, Mass: Blackwell.
Sterelny, K. (2012). The evolved apprentice: How evolution made humans unique (Vol. 2012). Cambridge, Mass: The MIT Press.
Stotz, K. (2010). Human nature and cognitive-developmental niche construction. Phenomenology and the Cognitive Sciences, 9(4), 483-501.
Turkeltaub, P. E., Gareau, L., Flowers, D. L., Zeffiro, T. A., \& Eden, G. F. (2003). Development of neural mechanisms for reading. Nature Neuroscience, 6(7), 767-773.
Vogel, A. C., Petersen, S. E., \& Schlaggar, B. L. (2014). The VWFA: It's not just for words anymore. Frontiers in Human Neuroscience, 1-10.
Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, Mass: Harvard University Press.
Wood, D., Bruner, J. S., \& Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89-100.
Ziegler, J. C., \& Goswami, U. (2006). Becoming literate in different languages: Similar problems, different solutions. Developmental Science, 9(5), 429-436.

# A Cognitive-Pharmacokinetic Computational Model of the Effect of Toluene on Performance 

Christopher R. Fisher (christopher.fisher.27.ctr@us.af.mil)<br>Christopher Myers (christopher.myers.29@us.af.mil)<br>Reem Hassan (remist3@gmail.com )<br>Christopher Stevens (christopher.stevens.28@us.af.mil)<br>C. Eric Hack (charles.hack.ctr@us.af.mil)<br>Jeffery Gearhart (jeffery.gearhart.ctr@us.af.mil)<br>Glenn Gunzelmann (glenn.gunzelmann@us.af.mil)<br>$711^{\text {th }}$ Human Performance Wing, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH, USA


#### Abstract

We developed a cognitive-pharmacokinetic computational (CPC) model to understand how pharmacoactive substances, such as caffeine and toluene, modulate cognition. In this integrated model, dynamic physiological mechanisms are simulated to predict concentrations of the solvent toluene in the brain, which modulates specific cognitive systems in a doseresponse fashion over multiple hours. We used our CPC model to reanalyze the results from prior research that documented an increase in reaction time following exposure to toluene in several laboratory tasks with no change in accuracy. Our analysis provides tentative evidence that toluene affects motor execution, rather than attention or declarative memory.


Keywords: ACT-R; Toluene; Pharmacokinetic; Computational Models

## Introduction

In cognitive science, there has been a growing trend toward incorporating neurological, physiological, and bodily factors into theories of cognition to improve our understanding of cognition. van Rijn et al. (2016) argued that theories of cognition should span multiple levels of abstraction, such as the computational, algorithmic, and implementational levels proposed by Marr (1982). Integrative models have the advantage of explaining phenomena that are difficult to explain with models cast at a single level, constraining and informing cognitive models, demonstrating physiological/neural plausibility of cognitive models, and enabling prediction at multiple time scales (e.g., milliseconds and hours).

Integrative models take several forms. For instance, embodied cognition encompasses a diverse set of views with a common emphasis on the central role for perceptual-motor systems, bodily states, and the environment in cognition (Wilson, 2002). Many views of embodied cognition posit a reciprocal relationship between cognition and bodily states, or a tight coupling between perception and action, whereas some even consider the environment as part of the cognitive system (Wilson, 2002). Another integrative approach incorporates physiological moderators of cognition, such as circadian
rhythm, to explain changes in cognition due to sleep deprivation (Gunzelmann et al., 2009).

The benefits of integrated models have been demonstrated in several recent studies. For example, Turner et al. (2016) demonstrated that incorporating neural models into a sequential sampling model of inter-temporal decision making improved fit and cross-validation compared to the component models. Integrated models have also been used to understand how cognitive moderators, such as stress and fatigue, alter task performance. For example, Gunzelmann et al. (2009, 2012) developed integrated models to understand the detrimental effects of sleep deprivation on cognition. In the integrated models, the physiological dynamics of circadian oscillation and sleep homeostasis modulated specific cognitive mechanisms. One important insight gained from these integrated models was that sleep deprivation affects procedural memory in simple reaction time (RT) tasks (Gunzelmann et al., 2009) and declarative memory in arithmetic tasks (Gunzelmann et al., 2012). Dancy et al. (2015) used a similar approach to understand how the physiology of a startle response affects cognition. Their integrated model revealed that the startle response increases epinephrine and the behavioral consequences could be explained in terms of fluctuations in the level of noise in declarative memory.

Building upon these integrative approaches, we present a general model that we call a cognitive-pharmacokinetic computational (CPC) model to understand how pharmacoactive substances (PSs) modulate cognition. PSs include toxins (e.g., toluene), pharmaceuticals, and other chemicals (e.g., caffeine) that affect cognition. Although there is a robust literature showing that PSs impact behavior, there is little computational work investigating which cognitive mechanisms are affected by PSs. Cognitive models lack a theory detailing how PSs are metabolized across time, whereas physiological models lack a theory that can link physiological changes to behavioral and cognitive consequences.

By integrating these approaches, we can disentangle the
contributions of multiple cognitive systems to overall task performance and identify which cognitive system is affected by PSs. To test our CPC model, we focus on a chemical called toluene, a colorless and odorless solvent commonly found in products, such as paint and adhesives, and is present in many work environments. Several studies have shown that acute and chronic toluene exposure leads to performance deficits in terms of RT or accuracy on a wide array of cognitive tasks (Rahill et al., 1996; Tang et al., 2011). What remains unclear is which cognitive system is affected by toluene exposure. For example, RT could increase as a result of a slowdown in attentional processing, motor execution, or memory retrieval. We use our CPC model to distinguish between these competing explanations.

## Data and Tasks

We developed a CPC model of the toluene exposure experiment reported in Rahill et al. (1996). In that experiment, six subjects completed a battery of cognitive tasks, once without toluene exposure and again with toluene exposure. The experiment was conducted in a chamber where the atmospheric level of toluene was precisely controlled. A battery of seven cognitive tasks was administered three times throughout each 8 -hour condition: once upon entering the chamber, again at 150 minutes immediately following 15 minutes of exercise and 15 minutes of biological exposure analysis, and one last time at 330 minutes. In the toluene condition, the air concentration of toluene was maintained at 100 ppm for the first 6 hours, after which point no further toluene was released into the chamber. Rahill et al. (1996) found that toluene increased mean RT for six of the seven tasks without impacting accuracy.

We developed CPC models for a subset of the tasks that offered the greatest chance of differentiating between competing accounts of performance decrements due to toluene exposure. The tasks were the procedural memory task, the recognition memory task, and the arithmetic task. Collectively, these tasks form a discriminative test bed for evaluating competing explanations because each task engages cognitive mechanisms in the computational model to varying degrees and thus produce a different pattern of predictions depending on which mechanism is modulated by toluene. Furthermore, a viable model must meet the challenge of producing the following pattern of effects with a single mechanism: an effect of toluene on RT in the procedural and recognition memory tasks, but not the arithmetic task.

## Procedural Memory Task

On each trial of the procedural memory task, the number 1 , 2,3 , or 4 was presented on the screen. Participants were instructed to respond according to the following stimulusresponse mapping: press one button if the number was 1 or 2 and press another button if the number was 3 or 4 . The stimulus disappeared after a maximum of 600 ms . ${ }^{1}$

[^64]

Figure 1: An illustration of the dose-response predictions of the motor CPC model for the procedural memory task. Toluene is rescaled in RT units for illustration.

## Recognition Memory Task

During the learning phase, participants studied a set of six letters displayed simultaneously on the screen until they were confident the letters were committed to memory. Next, on each trial of the test phase, participants indicated whether or not a memory probe was in the studied list. The stimulus disappeared after a maximum of 850 ms .

## Arithmetic Task

On each trial, a set of three single digit numbers were presented on the screen (e.g., $3+4-5$ ) and participants indicated with the appropriate button whether the solution was less than or greater than 5 . The stimulus disappeared after a maximum of $3,500 \mathrm{~ms}$.

## Cognitive-Pharmacokinetic Computational Model

The CPC model spans two levels of abstraction. At the physiological level, the physiologically-based pharmacokinetic pharmacodynamic (PBPK-PD) model describes the physiological dynamics that control the distribution and concentration of toluene. The output of the PBPK-PD model is the concentration level of toluene in the brain at a given point in time. At the cognitive level, the ACT-R cognitive architecture models the interplay of multiple cognitive systems during task performance. ACT-R and the PBPK models are integrated into a single model based on the assumption that toluene affects physiological and neural processes that support cognition, which in turn, affects performance. Figure 1 illustrates how mean RT for the motor CPC model tracks changes in toluene level in a dose-response fashion. We describe the submodels and model integration in the following sections.

## PBPK-PD Model

We used a PBPK-PD model to quantify the concentration of toluene in the brain throughout the exposure period (Tardif et al., 1997). The PBPK-PD model allows us to estimate the

[^65]amount of toluene in the brain and formulate dose-response predictions for task performance. A PBPK-PD model is an in silico representation of the movement of chemicals in the arterial blood, flowing to each major organ or lumped tissue compartment, including the brain.

PBPK-PD models calculate the time-course of PSs in the vascular and body tissues via ordinary differential equations to account for absorption, distribution, metabolism, and excretion processes. The following is an example differential equation of a metabolizing tissue:

$$
\begin{equation*}
\frac{d A_{l}}{d t}=\left(Q_{l} \times C_{a}\right)-\left(Q_{l} \times \frac{C_{l}}{P_{l}}\right)-\left(\frac{V_{M a x} \times C V_{l}}{K M+C V_{l}}\right) \tag{1}
\end{equation*}
$$

where subscripts $l$ and $a$ denote liver and arterial, respectively, $A=$ amount of chemical (mg), $Q=$ blood flow (L/hr), $C=\mathrm{mg} / \mathrm{L}, K M=$ Michaelis-Menten constant $(\mathrm{mg} / \mathrm{L}), \mathrm{P}=$ tissue/blood partition coefficient, CV is venous concentration, and $V_{\max }=$ maximum rate of parent chemical change to metabolite ( $\mathrm{mg} / \mathrm{hr)}$ ).

There are three basic critical components to PBPK-PD models: 1) species-specific physiological parameters, 2) chemical-specific parameters, and 3) experiment-specific details for the studies to be simulated. As per convention, the physiological and chemical parameter values in our model were based on prior empirical measurement (Tardif et al., 1997). Species-specific physiological parameters are the organ weights and blood flow rates for the defined compartments (e.g., organs) in the PBPK-PD model and are derived from the closest like species when not available. Chemicalspecific parameters that are unique for each chemical are the tissue solubility (partition coefficient), metabolism of the parent compound, and plasma and tissue binding characteristics. The specific experimental details pertain to the time of dosing and amount, route of dosing, and whether the subjects are physically active or quiescent. These details were obtained from Rahill et al. (1996).

Figure 1 shows the time-course of toluene concentration in the brain. Toluene increases rapidly from 135 to 150 minutes while the subject engages in exercise. Toluene concentration plateaus during rest then declines rapidly after the end of the 360-minute exposure period.

## ACT-R

ACT-R is a cognitive architecture that specifies how modular cognitive systems interact to produce cognition and overt behavior (Anderson, 2007). Models developed within ACT-R posit a common set of processes and mechanisms, which are instantiated as a computer simulation. Independent modules operate in parallel and include declarative memory, vision, attention, and motor modules. Procedural memory coordinates the behavior of the architecture through a set of production rules. Production rules follow an "IF-THEN" structure that encodes the conditions under which specific actions are taken. ACT-R provides a structure within which potential explanations for the effect of toluene on cognition can
be formalized. For example, toluene might disrupt normal functioning of declarative memory, resulting in a slowdown in the retrieval of task-relevant information. We developed CPC models that formalize three explanations: toluene affects (1) declarative memory, (2) attention, or (3) motor execution. For brevity, we will refer to each of these explanations as the memory CPC, attention CPC, and motor CPC model, respectively.

The memory CPC model formalizes the hypothesis that toluene interferes with memory retrieval. In ACT-R, each fact stored in declarative memory-called a chunk- is associated with an activation value corresponding to the frequency and recency with which it has been used. Higher activation results in faster and more accurate retrieval. The declarative memory system in ACT-R offers several potential mechanisms for toluene modulation. Our criteria for selecting a mechanism were (1) it must be theoretically grounded and (2) it must produce a transient effect. We selected the parameter $F_{d}$ because it produces a temporary decrease in activation and has been successful in accounting for the transient effect of fatigue on declarative memory (Gunzelmann et al., 2012). $F_{d}$ scales base-level activation as follows:

$$
\begin{equation*}
b_{i}=F_{d} \cdot b_{L L} \tag{2}
\end{equation*}
$$

where $F_{d}=[0,1], b_{i}$ is base-level activation for chunk $i$, $b_{L L}$ represents activation associated with life-long learning ( $\approx 2.68$; Gunzelmann et al., 2012). According to this explanation, toluene causes an acute decrease in activation, resulting in longer RTs and more errors. Decay, by contrast, has a destructive effect, which cannot be restored without additional practice.

The attention CPC model formalizes the hypothesis that the time required to attend to a stimulus is longer, resulting in a longer observed RT with no direct change in accuracy. Attentional processing time is controlled by increasing the attention latency parameter. The motor CPC model formalizes the hypothesis that toluene slows down the motor system, which increases RT without affecting accuracy. Motor execution is controlled by increasing the motor latency parameter.
Procedural Memory Model Declarative memory was populated with four chunks that encoded the response mapping. On each trial, the model attended to the number presented on the screen, retrieved a response mapping, and responded with the key specified in the retrieved chunk.
Recognition Memory Model Declarative memory was populated with chunks representing each letter in the alphabet. Once the list of six letters was presented, the model located a new letter starting on the left. After locating the letter, the model attended to the letter, and rehearsed it so as to strengthen its activation in memory. Throughout the course of the learning phase the model studied the list by repeating this cycle of productions. In Rahill et al. (1996), the learning phase was terminated by the subject when he or she was confident that the letters were memorized. However, no informa-
tion regarding the duration of the study phase was reported. We assumed participants studied the list for 10 seconds before proceeding to the test phase, which produced high accuracy found in similar studies (Levinson et al., 2005) with minimal time commitment. When a letter appeared during the test phase, the model attended to it, attempted to retrieve a chunk in memory that matched the letter and was in the study list, and executed a response. The model responded "yes" by key press if the retrieved letter matched the letter presented on the screen. If the letter did not match or no letter could be retrieved, the model responded "no" by key press.

Arithmetic Model Declarative memory in the arithmetic model was populated with chunks representing arithmetic facts. Once the problem was presented (e.g., $3+5-2$ ), the model processed each of the five components starting from left to right. First, the model located the leftmost stimulus (e.g., 3). Next, the model attended to the stimulus and then encoded the stimulus to keep track of the problem state. The model then repeated the procedure on the next stimulus (e.g., + ). After encoding the first two numbers and operator, the model retrieved and then encoded the intermediate solution (e.g., 8). The model processed the remaining stimuli and retrieved the final solution (e.g., 8-2 = 6). Lastly, the model responded whether the solution was less than or greater than 5 via key press. If a math fact could not be retrieved, the model responded randomly.

## Model Integration

The following equations provide a high-level representation of the model integration:

$$
\begin{equation*}
\operatorname{PBPK}(\Lambda, t)=\tau \tag{3}
\end{equation*}
$$

where $\Lambda$ is a set of parameters, $t$ is time since the beginning of the experiment, and $\tau$ is the predicted concentration of toluene in the brain. A high level representation of ACT-R is given by

$$
\begin{equation*}
\operatorname{ACTR}_{m}(\Theta)=(R T, A C C) \tag{4}
\end{equation*}
$$

where $\Theta$ is a set of parameters, $m \in\{$ procedual,recognition memory,arithmetic $\}$ indexes the ACT-R models, and the tuple ( $R T, A C C$ ) is the predicted mean reaction time and accuracy.

A linear link function allows specific ACT-R parameters to vary as a function of toluene level as follows:

$$
\begin{equation*}
\theta_{p}=\beta_{0_{p}}+\beta_{1_{p}} \tau \tag{5}
\end{equation*}
$$

$p \in P=\left\{F_{d}, A, M\right\} \subset \Theta$ indexes the toluene-modulated ACTR parameters, which correspond to fatigue declarative memory $\left(F_{d}\right)$, attention latency (A), and motor latency (M). The intercept $\beta_{0_{p}}$ is the value of parameter $\theta_{p}$ when the concentration level of toluene in the brain is zero. $\beta_{1_{p}}$ is the slope which represents the degree to which $\theta_{p}$ varies as a function of $\tau$. $\beta=\left\{\beta_{0_{p}}, \beta_{1_{p}}\right\}$ represents the set of link function parameters. Let $\hat{\Theta}=\Theta \backslash P$ be the subset of ACT-R parameters that are not determined from Equation 5 (e.g., latency factor).

Table 1: The slopes used in the link function of the CPC models. Slopes were varied over the ranges in brackets with 10 equal interval steps.

| CPC Model | $\beta_{1_{A}}$ | $\beta_{1_{F_{d}}}$ | $\beta_{1_{M}}$ |
| :--- | :---: | :---: | :---: |
| Attention | $[0, .015]$ | 0 | 0 |
| Memory | 0 | $[-.03,0]$ | 0 |
| Motor | 0 | 0 | $[0, .015]$ |
| Baseline | 0 | 0 | 0 |

The CPC model integrates the ACT-R and PBPK-PD models through the linear link function and can be represented as:

$$
\begin{equation*}
\operatorname{CPC}(\hat{\Theta}, \Lambda, \beta, t)=(R T, A C C) \tag{6}
\end{equation*}
$$

We imposed the following parametric restrictions on the CPC models in the interest of parsimony. First, as shown in Table 1, we assumed that toluene affected only one cognitive system: either the attention, memory, or motor system. For example, in the attention CPC model, the slope $\beta_{1_{A}}$ was allowed to vary while the other slopes were fixed to zero. As a basis for comparison, we also included a baseline CPC model in which no parameters were modulated by toluene. Second, we used the same parameterization of the link function across the three tasks. Specifically, when a slope was estimated, the estimated value applied across the three tasks. We also fixed the intercepts to $\beta_{0_{A}}=.085$ and $\beta_{0_{M}}=.05$, which are default values that have emerged as good fitting values across a wide range of studies. Because the intercept $\beta_{0_{F_{d}}}$ does not have a default value, we fixed this parameter to theoretically justified values of $.72,1$, and .83 for the procedural, recognition memory, and arithmetic tasks, respectively, to reflect differences in prior exposure to task-specific stimuli. For example, subjects had more experience with the alphabet used in the recognition memory task than the response mapping used in the procedural memory task.

Third, when possible, we fixed other parameters to default values. For example, we fixed decay to .5. Mismatch penalty and activation noise do not have default values, and as such, were set to 2.8 and .15 for all models under consideration. Fourth, task-specific parameters were fixed across toluene conditions, blocks, and the four CPC models. Specifically, we set the retrieval threshold to .78 in the recognition memory task to control the speed of negative responses. Finally, the parameters of the PBPK-PD model were fixed to values acquired through prior empirical measurement.

## Results

Human RTs (black) are displayed in Figures 2-4 for each task. Each panel represents an exposure condition, and points within each condition represent mean RT for a given block. Human RT increased in the toluene condition for the procedural and recognition memory tasks, but remained relatively constant in the arithmetic task. Although accuracy data were not reported in Rahill et al. (1996), no effect of toluene on ac-

Table 2: RMSE of best fitting models. PMT: procedural memory task, RMT: recognition memory task, AT: arithmetic task

|  |  | RMSE |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| CPC Model | $\beta_{1_{p}}$ | PMT | RMT | AT | Aggregate |
| Attention | .0015 | 10.13 | 28.04 | 92.30 | 56.00 |
| Memory | . .003 | 10.87 | 30.33 | 89.24 | 54.78 |
| Motor | .0045 | 9.59 | 23.40 | 90.17 | 54.07 |
| Baseline | 0 | 12.02 | 30.83 | 88.71 | 54.66 |



Figure 2: RT predictions for the CPC models plotted against the human mean RT for the Arithmetic Task. Bars are standard deviations.
curacy was detected. Based on other studies using the same or similar tasks, we assume that accuracy was $\geq 90 \%$ (Levinson et al., 2005; Vincent et al., 2012).

We fit the four CPC models using the parameter ranges displayed in Table 1 and selected the best fitting models using a two-stage procedure. In the first stage, we selected the subset of results for which accuracy was $\geq 90 \%$ in all blocks to ensure that the predictions were in line with previous studies. In the second stage, we selected the $\beta_{1_{p}}$ with the lowest aggregate RMSE for each model. Table 2 summarizes aggregate RMSE, RMSE broken down by task, and the best fitting $\beta_{1_{p}}$ for each model. The predictions of the best fitting CPC models are compared to the human data in Figures 2-4.

Aggregate RMSE was the lowest for the motor CPC model, suggesting that toluene slows down motor processing. Although the improvement in aggregate RMSE relative to the baseline model may appear small, it hides modest but important improvements in the procedural and recognition memory tasks. Importantly, the motor CPC model was able to capture the qualitative pattern of effects found in the human data: an effect of toluene in the procedural and recognition memory tasks with no effect in the arithmetic task.


Figure 3: RT predictions for the CPC models plotted against the human mean RT for the Procedural Memory Task. Bars are standard deviations.


Figure 4: RT predictions for the CPC models plotted against the human mean RT for the Recognition Memory Task. Bars are standard deviations.

It is also informative to discuss patterns found in some poorly fitting CPC models. For example, when $\beta_{1_{A}}$ increased for the attention CPC model, the RT predictions improved to a similar degree as the motor CPC model in the procedural and recognition memory tasks. In the arithmetic task, however, the attention CPC model greatly over-estimated RT due to the large contribution of attention to the overall RT. This finding provides further evidence against the attention CPC model.

## Discussion

We developed and tested a set of CPC models to understand which cognitive systems are affected by toluene and lead to
the performance decrements reported in the literature. The CPC model integrated the physiological dynamics of toluene concentration into the ACT-R cognitive architecture to produce dose-response predictions over a prolonged period of toluene exposure. The CPC models formally instantiated deficits in memory, attentional, and motor processing as competing explanations for detrimental effects of toluene exposure. Our model comparison provided tentative evidence that performance decrements are driven by a slowdown in motor execution. Furthermore, we also found evidence against attention as a mechanism: when attention was modulated by toluene to the same extent as the motor system, it greatly overestimated RTs in the arithmetic task. Importantly, the motor CPC model produced the pattern of toluene effects in the human data: an effect of toluene in the procedural and recognition memory tasks, and no effect in the arithmetic task.

Our CPC model adds to the growing literature showing that integrated models can yield more accurate predictions and deeper insights compared to non-integrative approaches. The CPC model has several benefits. First, it enabled us to account for data at two different time-scales: on the order of milliseconds as well as hours. Second, the CPC model was powerful yet highly constrained. With the CPC model, we were able to account for the effects of toluene exposure in three tasks using a single mechanism. Moreover, other parameters were either set to default values or otherwise highly constrained. Third, the CPC model is quite general, allowing it to be applied to any PS of interest.

## Limitations

Our findings should be interpreted in light of several limitations. First, research on PSs often has small sample sizes and small exposure manipulations due to restrictions imposed by institutional review boards to limit risk. As a result, discriminating among competing explanations is challenging and our conclusions regarding the motor CPC model remain tentative. Second, we also could not apply the model at the individual level or examine nuanced predictions (e.g., false alarms vs. misses) because only summary data were available. Third, our PBPK-PD model could not examine the possibility of region-specific effects of toluene in the brain. A model with this level of detail would provide tighter integration and more focused hypotheses about the affected mechanisms. Finally, we made assumptions about several unreported or ambiguous methodological details, such as the number of trials, the duration of the study phase in the recognition task, and the use of response deadlines. Nonetheless, when these assumptions were changed, the motor CPC still emerged with the strongest level of support.

## Acknowledgments

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This research was supported by a $711^{\text {th }}$ Human Performance Wing Chief Scientist

Seedling grant to Dr. Myers. Distribution A: Approved for public release; distribution unlimited. Cleared 02/02/2017; 88ABW-2017-0428.

## References

Anderson, J. R. (2007). How can the human mind occur in the physical universe? Oxford University Press.
Dancy, C. L., Ritter, F. E., Berry, K. A., \& Klein, L. C. (2015). Using a cognitive architecture with a physiological substrate to represent effects of a psychological stressor on cognition. Computational and Mathematical Organization Theory, 21(1), 90-114.
Gunzelmann, G., Gluck, K. A., Moore Jr, L. R., \& Dinges, D. F. (2012). Diminished access to declarative knowledge with sleep deprivation. Cognitive Systems Research, 13, 1-11.
Gunzelmann, G., Gross, J. B., Gluck, K. A., \& Dinges, D. F. (2009). Sleep deprivation and sustained attention performance: Integrating mathematical and cognitive modeling. Cognitive Science, 33(5), 880-910.
Levinson, D., Reeves, D., Watson, J., \& Harrison, M. (2005). Automated neuropsychological assessment metrics (anam) measures of cognitive effects of alzheimer's disease. Archives of Clinical Neuropsychology, 20(3), 403408.

Marr, D. (1982). A computational investigation into the human representation and processing of visual information. San Francisco: WH Freeman and Company.
Rahill, A. A., Weiss, B., Morrow, P. E., Frampton, M. W., Cox, C., Gibb, R., ... Utell, M. J. (1996). Human performance during exposure to toluene. Aviation, space, and environmental medicine, 67(7), 640-647.
Tang, C. Y., Carpenter, D. M., Eaves, E. L., Ng, J., Ganeshalingam, N., Weisel, C., ... Fiedler, N. L. (2011). Occupational solvent exposure and brain function: an fmri study. Environmental health perspectives, 119(7), 908.
Tardif, R., Charest-Tardif, G., Brodeur, J., \& Krishnan, K. (1997). Physiologically based pharmacokinetic modeling of a ternary mixture of alkyl benzenes in rats and humans. Toxicology and applied pharmacology, 144(1), 120-134.
Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., \& Steyvers, M. (2016). Why more is better: Simultaneous modeling of eeg, fmri, and behavioral data. Neuroimage, 128, 96-115.
van Rijn, H., Borst, J., Taatgen, N., \& van Maanen, L. (2016). On the necessity of integrating multiple levels of abstraction in a single computational framework. Current Opinion in Behavioral Sciences, 11, 116-120.
Vincent, A. S., Roebuck-Spencer, T., Gilliland, K., \& Schlegel, R. (2012). Automated neuropsychological assessment metrics (v4) traumatic brain injury battery: military normative data. Military medicine, 177(3), 256-269.
Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin \& Review, 9(4), 625-636.

# Thinking about the future: The role of spatial metaphors for time 

Stephen J. Flusberg ${ }^{1}$ (stephen.flusberg@purchase.edu)<br>SUNY Purchase College, Department of Psychology<br>735 Anderson Hill Road, Purchase, NY 10577, USA<br>Teenie Matlock ${ }^{1}$ (tmatlock@ucmerced.edu)<br>Cognitive \& Information Sciences Program and Center for Climate Communication<br>University of California, Merced, 5200 North Lake Rd., Merced CA 95343<br>Paul H. Thibodeau ${ }^{1}$ (pthibode@oberlin.edu)<br>Oberlin College, Department of Psychology<br>120 W. Lorain St., Oberlin, OH 44074, USA<br>${ }^{1}$ Authors listed alphabetically as all contributed equally


#### Abstract

People often use spatial language to talk about time, and this is known to both reflect and shape how they think about it. Despite much research on the spatial grounding of temporal language and thought, little attention has been given to how spatial metaphors influence reasoning about real events, especially those in the future. In a large online study ( $N=2362$ ), we framed a discussion of climate change using spatial metaphors that varied on reference-frame (ego- vs. time-moving), speed of movement (fast vs. slow), and time horizon (near, medium, or far future). We found that describing climate change as approaching (time-moving frame) - versus something we approach - made the issue seem more serious, but also more tractable, at least when the rate of motion was fast (e.g., "it's rapidly approaching"). These findings offer novel insights into the relationship between spatial metaphors and temporal reasoning and how we communicate about uncertain future events.


Keywords: metaphor, space, time, framing, reasoning, future

## Introduction

People often talk about time in terms of space (Clark, 1973; Lakoff \& Johnson, 1980). Two holidays can be described as close together, and deadlines, as rapidly approaching. Spatiotemporal metaphors, which underlie such talk, are ubiquitous across cultures (Boroditsky, 2011; Núñez \& Cooperrider, 2013). What's more, much research has established that we actually mentally represent and reason about time in terms of space as well, and that this happens in a manner that is consistent with the particular language we use (Boroditsky, 2000, 2011; Boroditsky \& Ramscar, 2002; Casasanto, 2005; Casasanto \& Boroditsky, 2008; Núñez \& Cooperrider, 2013; but see Casasanto, 2016).

A popular method for assessing this claim is to manipulate how someone is thinking about space before asking them to reason about time. In one early study, for example, Boroditsky (2000) showed participants spatial primes that depicted an agent moving towards a goal or an object moving toward an agent, and then asked them to answer an ambiguous temporal question: "Next Wednesday's meeting has been moved forward two days. Which day is the meeting now that its been moved?".

English speakers use two spatial reference frames for talking about time: ego-moving, which depicts the agent as actively moving through time-space (e.g., "we are approaching retirement") and time-moving, which depicts the agent as stationary while events in time move toward them (as in, "the holiday season is approaching"). In the ambiguous 'Wednesday's meeting' question, the implied vector of motion (forward) differs depending on which frame you adopt - toward (time-moving) versus away from (ego-moving) the individual - such that the meeting could now be interpreted as falling either on Monday (timemoving) or Friday (ego-moving). Boroditsky found that egoand time-moving spatial primes reliably biased participant responses to the ambiguous question in a metaphorcongruent manner, suggesting that people were relying on active spatial representations to reason about time (see also McGlone \& Harding, 1998).

This basic pattern of results has been replicated and extended in many ways, from the use of more ecologically valid spatial primes (Boroditsky \& Ramscar, 2002), to nonlinguistic measures of temporal reasoning (Casasanto, 2005; Casasanto \& Boroditsky, 2008), to cross-cultural comparisons (Boroditsky, Fuhrman, \& McCormick, 2010). These data offer converging support for the view that people (frequently) represent and reason about time using their knowledge of space, and that the specific spatial relations that are mapped onto the domain of time are shaped by patterns of metaphor in language (along with other factors like writing direction and cultural values; Boroditsky, 2011; Casasanto, 2016; Núñez \& Cooperrider, 2013).

While research has been largely focused on showing links between space and time in the mind, scant attention has been given to whether metaphors influence how people reason about real world events in the future. Do people conceptualize impending events differently when they are described using different spatial metaphors for time?

To address this question, we conducted a large-scale linguistic framing study to assess how people think about negative outcomes associated with climate change. We chose climate change because it is a real-world problem laden with uncertainty. In general, people have a very poor
understanding of what climate change is, what the specific outcomes will be, and what to do about it (see Barnosky et al., 2017). We reasoned that this inherent ambiguity might make it easier to observe the effects of spatiotemporal metaphors on how people think and feel about the issue since people's prior beliefs may be somewhat nebulous.

In our experiment, participants first read a brief article about US efforts to tackle climate change, and then responded to questions about how serious and tractable they viewed the issue. We manipulated whether the report described climate change with the ego- or time-moving reference-frame, whether speed of movement in time-space was fast or slow, and whether US conservation goals were situated in a relatively near, medium, or far future time horizon.

We hypothesized that using a time-moving reference frame would make the effects of climate change seem more urgent and serious, since this perspective represents the individual as fixed in place, unable to control the arrival of a negative future event. This would be consistent with prior research showing that people who spontaneously adopt a time-moving perspective tend to show higher levels of state and trait anxiety and depression, which are associated with a loss in feelings of agency (Richmond, Wilson, \& Zinken, 2012).

However, this increase in feelings of urgency does not imply an increase in pessimism about the tractability of the issue. In fact, it could be the case that the time-moving frame might lead people to view climate change as a more tractable problem, given that the individual is free to engage in their own actions on this construal (since they are not occupied with the task of moving through time). This would resonate with research showing that people who spontaneously adopt the time-moving reference frame procrastinate less and are more conscientiousness than those who spontaneously adopt the ego-moving reference frame. (Duffy \& Feist, 2014; Duffy, Feist, \& McCarthy, 2014). One way of thinking about this is that the decrease in feelings of control (and increase in anxiety) that results from the time-moving frame might lead to a compensatory counter-response, such that people would now be motivated to believe that personal actions are likely to be effective in addressing the problem. In other words, when it feels like you cannot stop a future event from happening, you will feel better if you consequently believe that at least you can deal it when it arrives.

We included the speed manipulation to assess whether the "rate" at which we approach future events (or they approach us) might affect or interact with the temporal reference frame in shaping these attitudes towards climate change. It is plausible, for instance, that faster "motion" would be associated with a greater sense of urgency. The time horizon manipulation was included in part to affect perceptions of whether the US seemed likely to achieve the conservation milestones in the article, which allowed us to assess effects of the other spatial metaphors independently of this judgment (see Flusberg, Matlock, \& Thibodeau, in press).

## Experiment

## Methods

Participants A total of 2400 participants were recruited and paid through Amazon's Mechanical Turk for the study in the Spring of 2016 (Berinsky, Huber, \& Lenz, 2012; Buhrmester, Kwang, \& Gosling, 2011). We restricted our sample to people living in the US who had a good performance rating ( $>90 \%$ ) on previous Turk tasks. Data was not analyzed from 38 participants who did not complete the study (i.e. from participants who did not submit a valid completion code), leaving a sample size of $N=2362$. The sample was $46 \%$ male and had a mean age of 35.2 years ( $S D$ $=11.1$ ).

Materials \& Procedure Participants read a brief fictional article that described US efforts to reduce greenhouse gas emissions. It used (1) an ego- or time-moving frame of reference; (2) temporal language about climate change as a slow or fast process (speed), and (3) identified an outcome on a time horizon in the relatively near (2025), medium (2040), or distant future (2115).

As shown in Figure 1, the report began, "In response to the recent Paris Climate Talks, the Associated Press release the following brief statement." The title was presented below this heading in capital letters. The rest of the passage was an appeal for addressing climate change and identified a specific goal for the US: to reduce greenhouse gas emissions by more than $30 \%$ by 2025,2040 , or 2115 .


Figure 1. Participants read this report, which varied on frame of reference, speed of change and time horizon.

The body of the report for the ego-moving frame of reference condition read (differences by speed and time horizon conditions are noted in the text):

We're \{rapidly / gradually\} approaching the day when it will be too late to prevent the devastating effects of climate change. We will \{quickly / eventually\} find ourselves in a world that includes more extreme weather conditions, more public health problems, as well as severe economic challenges if we don't start \{racing / inching\} towards a solution soon. As a result, the United States has pledged to reduce its carbon footprint in the next few decades, approving dozens of projects as part of an effort to reduce greenhouse gas emissions by more than $30 \%$ by $\{2025,2040,2115\}$. The projects will leverage scientific expertise and individual engagement to improve the energy efficiency of cars and
buildings, reduce personal energy use, and increase the use of renewable energies such as wind and solar. Let's avoid the \{race / slow crawl\} towards disaster!
The body of the report for the time-moving frame of reference condition read (differences by speed and time horizon conditions are noted in the text):

The day is \{rapidly / gradually approaching when it will be too late to prevent the devastating effects of climate change. If a solution doesn't start heading our way \{quickly / eventually\}, more extreme weather conditions, more public health problems, as well as severe economic challenges will \{swiftly / slowly\} appear. As a result, the United States has pledged to reduce its carbon footprint in the next few decades, approving dozens of projects as part of an effort to reduce greenhouse gas emissions by more than $30 \%$ by $\{2025,2040$, $2115\}$. The projects will leverage scientific expertise and individual engagement to improve the energy efficiency of cars and buildings, reduce personal energy use, and increase the use of renewable energies such as wind and solar. Let's watch out for disaster as it \{quickly / slowly\} approaches!
Target and Background Questions. After reading the article, participants answered question about whether they thought the US would achieve its climate reduction goal in the stated time frame (i.e., by 2025,2040 , or 2115 ). Then they answered questions about whether they thought the problems of climate change would be solved, whether the disastrous effects of climate change were inevitable at this point, how urgent it is for the US to implement energy reduction programs, and how much risk they perceived to be associated with climate change. Participants also answered questions about their willingness to change their own behavior to reduce greenhouse gas emissions (see Figure 2).

```
Dependent Measures:
    Goal: The United States will achieve its goal of reducing greenhouse gas emissions by more than 30% by...
        1, Strongly disagree, to 5, Strongly agree
    Solvable: Humans will inevitably solve the problems associated with climate change, preserving the earth
        for future generations.
            1, Strongly disagree, to 5, Strongly agree
    Inevitable: The disastrous effects of climate change are inevitable, and there is nothing we can do to
        prevent them.
            1, Strongly disagree, to 5, Strongly agree
    Urgent: How urgent is it for the US to implement energy reduction programs right away?
            1, Not at all urgent, to 5, Very urgent
    Risk Perception: How concerned are you with the following potential consequences of climate change?
        (e.g., soil erosion, water drought, economic decline; 13 items; }\alpha=.93
            0, Not at all concerned, to 7, Extremely concerned
    Behavioral Intentions: Would you be willing to pay a carbon offset cost on future purchases of items
        derived from fossil fuels? (6 items; \alpha=.87)
            1, Definitely no, to 5, Definitely yes
Background Questions:
    Belief in Global Warming: I believe that burning fossil fuels increases atmospheric temperature to some
        measurable degree.(2 items; }\alpha=.90
            1, Strongly disagree, to 5, Strongly agree
    Political ideology (0, very liberal, 100, very conservative)
    Age, Gender, Education, Political affiliation (Democrat, Independent, Republican, or Other)
```

Figure 2. Dependent measures and background questions.

Most dependent measures were rated on a 5-point scale. One exception was the measure of risk perception, recorded on an 8-point slide bar. The measure of risk perception included 13 items (Cronbach's $\alpha=.93$ ). The measure of behavioral intentions included six items (Cronbach's $\alpha=$ .87). All other dependent measures were a single question.

Participants then answered demographic questions about their age, gender, educational history, political ideology (categorically and on a continuum), and about their belief in global warming (two items: "I believe that burning fossil fuels increases atmospheric temperature to some measurable degree" and "I believe that the burning of fossil fuels on the scale observed over the last 50 years has increased atmospheric temperature to an appreciable degree"; Chronbach's $\alpha=.90$; Lewandowsky, Oberauer, \& Gignac, 2013).

Data Reduction. As expected, the dependent measures were correlated with one another. As shown in Table 1, the six measures clustered into two groups: there was a high correlation (a) between the goal judgment, assessment of whether climate change would be solved, and whether the consequences of climate change were inevitable ( $r \mathrm{~s}>.19$ ), and (b) between the measures of urgency, risk perception, and willingness to change one's behavior ( $r \mathrm{~s} \gg .5$ ).

Table 1. Correlations between the dependent measures. Asterisk indicates statistical significance at the $* p<.001$ level.

|  | 1 | 2 | 3 | 4 | 5 | 6 |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| 1. Goal |  | $.38^{*}$ | $-.19^{*}$ | $.15^{*}$ | $.15^{*}$ | $.15^{*}$ |
| 2. Solvable |  | $-.29^{*}$ | -.04 | -.01 | -.01 |  |
| 3. Inevitable |  |  | $-.13^{*}$ | -.06 | $-.10^{*}$ |  |
| 4. Urgent |  |  |  | $.52^{*}$ | $.56^{*}$ |  |
| 5. Risk |  |  |  |  | $.55^{*}$ |  |
| 6. Behavior |  |  |  |  |  |  |

To further investigate the relationship among dependent measures, we did an exploratory factor analysis: a principal components analysis, using singular value decomposition (Mardia, Kent, \& Bibby, 1980). Principal components analysis (PCA) extracts the common variance in measures that are conceptually and empirically related (Dunteman, 1989). The analysis revealed, based on the Kaiser criterion and an analysis of the Scree plot, two major underlying sources of variance in the data, consistent with the pairwise correlations in Table 1. As shown in Table 2, the first factor loaded most heavily on the measures of urgency, risk perception, and willingness to change. The second loaded most heavily on the goal judgment, whether people thought climate change would be solved, and whether they thought the consequences of climate change were inevitable.

To analyze the data parsimoniously, we created two composite outcome variables based on the clustering of dependent measures found in the pairwise correlations and exploratory factor analysis. We combined the first three questions, using PCA, into a measure of how tractable participants considered the problem of climate change, and the last three questions into a measure of how serious participants considered the problem of climate change (see Table 2 for weights used to create the composite measures). The composite measures captured the majority of the variance in the raw data: the measure of how tractable
participants' considered climate change to be captured 53\% of the variance in the first three questions; the measure of how serious participants' considered climate change captured $70 \%$ of the variance in the last three questions ${ }^{1}$.

Table 2. Factor loadings of exploratory factor analysis and weights used to create composite outcome measures.

|  | PC1 | PC2 | Tractable | Serious |
| :--- | :--- | :---: | :---: | :---: |
| Goal | .27 | $\mathbf{. 5 0}$ | .58 |  |
| Solvable | .11 | $\mathbf{. 6 6}$ | .64 |  |
| Inevitable | -.20 | .$- \mathbf{4 6}$ | -.51 |  |
| Urgent | $\mathbf{5 4}$ | -.18 |  | .58 |
| Risk | $\mathbf{5 3}$ | -.19 |  | .57 |
| Behavioral intentions | $\mathbf{. 5 5}$ | -.18 |  | .59 |

There was a moderate correlation between the two composite measures (tractable and serious), $r(2360)=.12$, $p<.001$ (in contrast, $p c 1$ and $p c 2$ are orthogonal), and both were positively correlated with participants' belief in global warming: tractable, $r(2360)=.14, p<.001$; solvable, $r(2360)=.66, p<001$.

## Results

Analysis We conducted our primary hypothesis tests on the two composite measures (i.e. how tractable and serious people consider climate change to be). We also analyzed the six dependent measures separately in exploratory follow-up analyses. For each analysis, an initial omnibus betweensubjects ANOVA is presented with tests for main effects of and interactions between the three experimental manipulations. We show the results of the two ANOVAs, along with the coefficients from the corresponding linear regression models in Tables 3 and 4. Time horizon (2025, 2040, 2115) was treated as an ordinal variable; frame of

[^66]reference (ego- versus time-moving) and speed (fast versus slow) were treated as factors.

To account for the primary source of variance in the dependent measures, we included participants' belief in the anthropogenic origins of climate change as a covariate, although this did not affect the reliability of the results. To address secondary research questions (e.g., who is affected by manipulating the reference frame?), we tested for 2-way interactions between the experimental manipulations and participants' belief in global warming.

Serious The results of a model in which the Frame of reference (ego- or time-moving), speed (fast or slow), and time horizon (2025, 2040, or 2115), as well as participants' belief in global warming, were used to predict how serious people consider climate change to be (see Table 3). The strongest predictor of seriousness was participants' belief in global warming. People who recognized the anthropogenic origins of climate change thought the issue was more urgent, recognized more risk, and were more willing to change their behavior to reduce greenhouse gas emissions.

The model also revealed a main effect of the frame of reference, qualified by an interaction with participants' belief in global warming. Overall, participants were more likely to think of climate change as an urgent issue with important risks worthy of behavior change in the timemoving condition $(M=.06, S D=1.44)$, compared to the ego-moving condition ( $M=-.05, S D=1.52$ ). This was especially true for people who were skeptical about the anthropogenic origins of climate change, likely due to a ceiling effect on this measure for non-skeptics.

Table 3 Effects of experimental manipulations and belief in global warming on perceptions of the seriousness of climate change. The results of the ANOVA $\left(d f_{I}=1\right.$ and $d f_{2}=2350$ in every case) are shown in the first column; regression coefficients (and standard errors) are shown in the second column.

|  | $F(p)$ | $B(S E)$ |
| :--- | :--- | :---: |
| Intercept |  | $0.00(.02)$ |
| Time horizon | $0.63(.429)$ | $-0.03(.03)$ |
| Speed: Slow | $0.05(.818)$ | $0.00(.02)$ |
| Frame of reference: Time | $4.55(.033)$ | $0.01(.02)$ |
| Belief in global warming | $1775.65(<.001)$ | $0.95(.02)$ |
| Time horizon * Speed | $0.51(.475)$ | $0.02(.03)$ |
| Time horizon * Frame | $0.84(.359)$ | $0.03(.03)$ |
| Speed * Frame | $0.04(.839)$ | $0.01(.02)$ |
| Time horizon * Belief | $1.41(.236)$ | $0.04(.03)$ |
| Speed * Belief | $1.55(.213)$ | $-0.03(.02)$ |
| Frame * Belief | $4.66(.031)$ | $-0.05(.02)$ |
| Time * Speed * Frame | $2.06(.151)$ | $0.04(.03)$ |

Separate analyses on "raw" questions about urgency, risk perception, and behavioral intentions yielded consistent results. For example, people reported perceiving more risk in the time-moving reference frame, $B=.43, S E=.20, p=$ .032. Perceptions of risk were related to beliefs about global warming, $B=1.01, S E=.04, p<.001$. These predictors also
interacted, $B=-.12, S E=.05, p=.029$, suggesting that the effect of the reference frame most strongly affected people who reported skepticism about climate science.

Finally, the exploratory analyses of participants' responses to questions of urgency, risk perception, and behavioral intentions suggested that one effect was obscured by analyzing the composite measure: of speed on perceptions of urgency. Participants reported that global warming was a more urgent issue to address when the language suggested that the climate was changing quickly, $B$ $=.27, S E=.14, p=.049$, regardless of time horizon or frame of reference.

Tractable The results of a model in which the Frame of reference (ego- or time-moving), speed (fast or slow), and time horizon (2025, 2040, or 2115), as well as participants' belief in global warming, were used to predict judgments related to how tractable people consider climate change to be (see Table 4). Consistent with the analysis of how serious people consider the issue, the strongest predictor in the model was participants' belief in global warming. People who recognized the anthropogenic origins of climate change were more optimistic about being able to address the problem, probably due to the fact they are the ones who think it is a problem in the first place. There was also a main effect of the time horizon manipulation: people considered the issue more tractable when the specific goal was situated in the distant, as opposed to the near, future.

Table 4. Effects of manipulations and belief in global warming on perceptions of tractable-ness of climate change. ANOVA results ( $d f_{1}=1$ and $d f_{2}=2350$ in every case) are in the first column; regression coefficients (and standard errors) are in the second column.

|  | $F(p)$ | $B(S E)$ |
| :--- | :--- | :--- |
| Intercept |  | $0.00(.03)$ |
| Time horizon | $8.29(.004)$ | $0.07(.03)$ |
| Speed: Slow | $0.00(.981)$ | $0.00(.03)$ |
| Frame of reference: Time | $1.34(.247)$ | $0.02(.03)$ |
| Belief in global warming | $43.04(<.001)$ | $0.17(.03)$ |
| Time horizon * Speed | $1.94(.164)$ | $-0.04(.03)$ |
| Time horizon * Frame | $0.02(.888)$ | $0.00(.03)$ |
| Speed * Frame | $3.11(.077)$ | $0.05(.03)$ |
| Time horizon * Belief | $6.42(.011)$ | $-0.08(.03)$ |
| Speed * Belief | $0.22(.638)$ | $-0.01(.03)$ |
| Frame * Belief | $0.22(.643)$ | $-0.01(.03)$ |
| Time * Speed * Frame | $0.56(.454)$ | $0.02(.03)$ |

Finally, there was an interaction between participants' belief in global warming and the time horizon. The effect of the time horizon manipulation was driven by participants who were more skeptical about climate science. For example, among participants who reported the most skepticism about climate change (a score less than 2 on the measure of belief in global warming; $n=196$ ), there was a relatively large effect of the time horizon manipulation $\left(M_{2025}=-.75, S D=1.34 ; M_{2040}=-.57, S D=1.44 ; M_{2115}=-\right.$ $.11, S D=1.31), F(1,194)=8.652, p=.004$. In contrast,
among participants who reported the strongest belief in climate science (a score greater than 4 on the measure of belief in global warming; $n=668$ ), there was no effect of the time horizon manipulation $\left(M_{2025}=.10, S D=1.22\right.$; $\left.M_{2040}=.10, S D=1.40 ; M_{2115}=.19, S D=1.35\right), F(1,194)=$ $0.36, p=.551$. In other words, people who reported a belief in climate science tended to be optimistic about the issue, regardless of the time horizon of the goal, whereas people who were skeptical about climate science were more likely to think the issue would be solved on a more distant time horizon.


Figure 3. Participants' perception of the tractability of the problem of climate change, as a function of the reference frame and speed. Error bars denote SEMs.

The omnibus test also revealed a marginal interaction between the speed manipulation (whether the effects of climate change were described as happening quickly or slowly) and the frame of reference (ego- or time-moving). As shown in Figure 3, there was an effect of the reference frame when the report described the effects of climate change as occurring quickly (but not slowly). In the fast speed condition, participants were more optimistic about solving climate change on the time-moving reference frame compared to the ego-moving reference frame, $t(1160)=$ $2.07, p=.039$. In the slow speed condition, there was no effect of the reference frame, $t(1198)=0.38, p=.704$.

## Discussion

Time is highly abstract, but people manage to talk and think about it by drawing on spatial language and knowledge (Boroditsky, 2011). In this paper, we examined how particular spatial metaphors used to describe uncertain future outcomes would affect how people think about an important issue. So, instead of focusing on whether people reason about time using spatial representations, we looked at how different spatial construals would affect how they think about a real-world event. In a large online study, we framed a discussion of US efforts to tackle climate change using spatial metaphors that varied according to reference-frame
(ego- vs. time-moving), speed of movement (fast vs. slow), and time horizon (near, medium, or far future).

People appeared to be more optimistic about solutions for climate change with a more distant time horizon (implying there would be more time to address it). This was especially true for people who were skeptical of climate science, which probably reflects a ceiling effect for those who are more inclined to accept the scientific consensus.

More interestingly, and consistent with our initial hypothesis, climate change seemed more serious and urgent when described with a time-moving metaphor than with an ego-moving metaphor. This supports the view that talking about uncertain future events as approaching of their own accord may be associated with additional anxiety surrounding the issue. Because this effect was actually most pronounced for people who were skeptical about climate science (again suggesting a ceiling effect for non-skeptics), this finding may have important practical applications for policy makers and climate science communicators. We also observed some suggestion that metaphorical speed affects this sense of urgency, such that fast "motion" language makes people think the issue is more urgent.

Also consistent with our initial prediction, the reference frame appeared to affect perceptions of the tractable-ness of the issue of climate change, though only when the process was described as happening quickly: people felt the issue was more tractable on the time-moving reference frame when climate change was said to be occurring rapidly. This may arise from the increased sense of urgency and risk surrounding the issue on the time-moving frame - to effectively compensate for this increase in existential anxiety, people may come to view the problem as something they can actually address through concerted action.

Though preliminary, these findings have shed some new light on how metaphor can affect reasoning - both in general and for an issue with real world consequences. Taking a nuanced approach like this to investigating metaphor has the potential to advance our understanding of how metaphor works in the context of communicating about real world problems.

## References

Barnosky, A., Matlock, T., Christensen, J., Han, H., Miles, J., Rice, R., ... \& White, L. (2016). . Establishing common ground: Finding better ways to communicate about climate disruption. Collabra, 2(1).
Berinsky, A. J., Huber, G. A., \& Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Political Analysis, 20(3), 351-368.
Boroditsky, L. (2000). Metaphoric Structuring: Understanding time through spatial metaphors. Cognition, 75(1), 1-28.
Boroditsky, L. (2011). How Languages Construct Time. In Dehaene and Brannon (Eds.,) Space, time and number in the brain: Searching for the foundations of mathematical thought. Elsevier. ISBN: 978-0-12-385948-8

Boroditsky, L., Fuhrman, O., \& McCormick, K. (2010). Do English and Mandarin speakers think differently about time? Cognition. doi:10.1016/j.cognition.2010.09.010
Boroditsky, L. \& Ramscar, M. (2002). The Roles of Body and Mind in Abstract Thought. Psychological Science, 13(2), 185-188.
Buhrmester, M. D., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science. 6, 3-5.
Casasanto, D. (2005). Perceptual foundations of abstract thought. (Doctoral dissertation). Department of Brain \& Cognitive Sciences, MIT, Cambridge, MA.
Casasanto, D. (2016).. (2016). Cognitive Sciences, MIT, Cambridge, MA.go hand in hand. In B. LewandowskaTomaszczyk (Ed.),wConceptualizations of time onceptualizations of time.), w(EBenjamins.
Clark, H.H. (1973). Space, time, semantics, and the child. In T. E. Moore (Ed.), Cognitive Development and the Acquisition of Language (pp. 27-83). New York, NY: Academic Press.
Duffy, S. E., \& Feist, M. I. (2014). Individual differences in the interpretation of ambiguous statements about time. Cognitive Linguistics, 25(1), 29-54
Duffy, S. E., Feist, M. I., \& McCarthy, S. (2014). Moving Through Time: The Role of Personality in Three Real Life Contexts. Cognitive science, 38(8), 1662-167
Dunteman, G. H. (1989). Principal components analysis (No. 69). Sage.
Flusberg, S. J., Matlock, T., \& Thibodeau, P. H. (in press). Metaphors for the war (or race) against climate change. Environmental Communication
Lakoff, G., \& Johnson, M. (1980). Metaphors we live by. Chicago and London: University of Chicago Press.
Leiserowitz, A. (2006). Climate change risk perception and policy preferences: The role of affect, imagery, and values. Climatic change, 77(1-2), 45-72.
Lewandowsky, S., Oberauer, K., \& Gignac, G. E. (2013). NASA faked the moon landing-therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. Psychological Science, 24(5), 622633.

Mardia, K. V., Kent, J. T., \& Bibby, J. M. (1980). Multivariate analysis. New York: Academic Press.
McGlone, M. S., \& Harding, J. L. (1998). Back (or forward?) to the future: The role of perspective in temporal language comprehension. Journal of Experimental Psychology: Learning, Memory,and Cognition, 24, 1211-1223.
Núñez, R., \& Cooperrider, K., (2013). The Tangle of Space and Time in Human Cognition. Trends in Cognitive Sciences, 17(5), 220-229.
Richmond, J., Wilson J. C. \& Zinken, J. (2012). A feeling for the future: How does agency in time metaphors relate to feelings? European Journal of Social Psychology 42(7), 813-823.

# Early Colour Word Learning in British Infants 

Samuel H. Forbes (samuel.forbes@psy.ox.ac.uk)<br>Department of Experimental Psychology, University of Oxford<br>United Kingdom

Kim Plunkett (kim.plunkett@psy.ox.ac.uk)<br>Department of Experimental Psychology, University of Oxford United Kingdom


#### Abstract

Colour word learning has traditionally been viewed as a difficult task. Previous accounts have focussed on infants' ability to show an adult-like understanding of colour terms. Here we examine whether infants understand colour terms at a basic level, using two different methods: first, evidence from parental reports that British infants can comprehend colour terms early, second from experimental data using eye-tracking. These finding show that colour word learning is a process that begins much earlier than previously thought, and develops slowly as infants learn where the boundaries of each term are located. Due to their abstract properties, colour words present a unique opportunity to assess category learning in infants, as well as the mechanisms that control word learning in general.


Keywords: Word learning; language acquisition; colour

## Introduction

It has long been documented that colour words are difficult for children to learn (Kowalski \& Zimiles, 2006; O’Hanlon \& Roberson, 2006; Soja, 1994; Sandhofer \& Smith, 1999). They are learned late (Heider, 1971; Franklin, 2006) and even when infants do learn to say colour words, they are riddled with errors (Pitchford \& Mullen, 2003). But just how difficult are colour words to learn? Studies to date have had difficulty establishing a time-line for when colour words are learned, leading to various theories about why colour words are more difficult to learn, such as the inability to put categorical terms on a continuous spectrum of colour (Andrick \& Tager-Flusberg, 1986), or their ability to abstract colour as a relevant domain of linguistic meaning (Kowalski \& Zimiles, 2006; O'Hanlon \& Roberson, 2006; Sandhofer \& Smith, 1999). In the present study, it is demonstrated that in fact infants have some degree of colour word knowledge much earlier than previously shown, and that they are able to recognise typical colours when named. This finding suggests that the difficulties that infants have with understanding colour terms correctly may be due more to an inability to recognise the category boundary, rather than a complete lack of understanding of the basic terms.

Research into colour word learning in infants has shown that the age of acquisition has dropped dramatically over time (Franklin, 2006; Shatz, Behrend, Gelman, \& Ebeling, 1996). While early studies had colour words learned as late as $7 ; 0$ (Heider, 1971), more recent studies have concluded that they are learned successfully around three years of age (Pitchford \& Mullen, 2002) or earlier (Mervis, Bertrand, \& Pani, 1995). While there may be an actual drop in age of colour term acquisition due to a rise in coloured plastic goods or other such
environmental influences, it is also possible that an infant's comprehension of colour terms has been underestimated.

It is clear that infants have difficulty grasping an adult-like understanding of the meaning of a colour word, but this does not necessarily imply that they have failed to grasp the meaning of the term. Research into the slow inductive process of colour word learning has shown that it is possible for toddlers from around 30 months to 40 months to comprehend basic colour terms while still making errors (Wagner, Dobkins, \& Barner, 2013; Wagner, Jergens, \& Barner, 2014). This suggests that colour words are similar to most other classes of word, in that children first acquire a partial meaning for the colour terms, and start producing the terms, before then later slowly acquiring a fuller, adult-like meaning. In this case, proper, adult-like comprehension follows only after production, further raising questions about what it is to comprehend in such an abstract category of words. While this account may provide some insight into how colour words are learned, the participants are old enough that they may have already comprehended colour words to a degree. This raises two questions: when are colour words learned, and what is meant in this instance by "learned?"

The present study employs two different means to answer these questions: a parental word-learning survey, and an eye tracking paradigm. While word-learning surveys are particularly useful for measuring production (Fenson et al., 1994), measuring comprehension in this way has been debated (Tomasello \& Mervis, 1994; Houston-Price, Mather, \& Sakkalou, 2007), despite test showing that if anything, they could even be an underestimate of the child's ability (Styles \& Plunkett, 2009). In the case of colour terms, which are a much more abstract category of word than concrete nouns, this doubt might be magnified. The aim of the present study is to investigate when colour word learning is occurring, and in doing so, establish where the process of colour term acquisition might begin. The results of the parental survey are then to be compared to eye-tracking data, in order to confirm the measures of early comprehension, and compare them to measures of production from the survey.

## Parental Reports

## Method

Participants 2692 8-30 month-old participants' details were filled out by parents either on paper or online before visiting the lab. Participants were always visiting either Ply-
mouth Babylab or Oxford BabyLab as part of an experiment. Participants for whom there was incomplete descriptive data of age and gender were not included in the analysis. As a number of participants visited more than once, the total was $N=3413$ samples ( 1653 female). Figure 1 shows participant information by age and gender.


Figure 1: Histogram of participant samples in each age group by gender.

Procedure Parents were asked to fill out the Oxford Communicative Development Inventory (Oxford CDI, Hamilton, Plunkett, \& Schafer, 2000). The Oxford CDI is a British adaptation of MacArthur-Bates CDIs (Fenson et al., 1994), which contains 416 terms. The Oxford CDI differs from the MacArthur-Bates CDIs in that it measures comprehension and production for the full age range of participants. Parents or caregivers were asked to check if their child does not comprehend, comprehends, or comprehends and produces each word. Thus an infant was marked as producing a term only if they were thought to "comprehend and produce" the term, while selecting either "comprehends" or "comprehends and produces" saw them marked as comprehending.

The Oxford CDI contains four colour terms: red, blue, green, and yellow. Only data from these four terms are included in the analysis.

Data was analysed in two separate binomial regressions, one for the comprehension data, and another for the production data, where the outcome was a binary response for the colour term in question (yes or no for either comprehension or production). Both models had fixed effects of which colour term was being recorded, as well as age of participant and the gender of the participant. Gender and colour were both dummy coded in each model.

## Results

A summary of the model coefficients of each model can be seen in Table 1 and Table 2. Effects were added into the model individually and compared by measuring reduction in deviance compared to degrees of freedom, distributed as a
$\chi^{2}$. Colour improved model fit in comprehension $\operatorname{dev}(3)=$ $15.95, p=0.0012$ and in production $\operatorname{dev}(3)=21.58, p<$ 0.0001 , as did gender, both in comprehension $\operatorname{dev}(1)=$ $43.2, p<0.0001$, and in production $\operatorname{dev}(1)=59.89, p<$ 0.0001 .

Table 1: Standardized values of comprehension model effects. Colours are as compared to Blue.

|  | Estimate | Std. Error | $z$ value | $\operatorname{Pr}(>\|z\|)$ |
| :--- | :--- | :--- | :--- | :--- |
| (Intercept) | -8.42 | 0.20 | -41.51 | 0.000 |
| Age | 0.39 | 0.01 | 41.56 | 0.000 |
| Green | -0.25 | 0.07 | -3.75 | 0.000 |
| Red | -0.04 | 0.07 | -0.68 | 0.499 |
| Yellow | -0.11 | 0.07 | -1.67 | 0.095 |
| Male | 0.66 | 0.27 | 2.42 | 0.016 |
| Age:Male | -0.05 | 0.01 | -3.61 | 0.000 |

Table 2: Standardized values of production model effects. Colours are as compared to Blue.

|  | Estimate | Std. Error | $z$ value | $\operatorname{Pr}(>\|z\|)$ |
| ---: | ---: | ---: | ---: | ---: |
| (Intercept) | -9.24 | 0.23 | -39.59 | 0.000 |
| Age | 0.40 | 0.01 | 39.07 | 0.000 |
| Green | -0.34 | 0.07 | -4.65 | 0.000 |
| Red | -0.17 | 0.07 | -2.34 | 0.019 |
| Yellow | -0.19 | 0.07 | -2.55 | 0.011 |
| Male | -0.60 | 0.34 | -1.77 | 0.077 |
| Age:Male | 0.01 | 0.01 | 0.60 | 0.549 |

The results from Table 1 show that there was a large difference between green and the other colours, both in speaking and comprehension, while yellow and red were both slightly behind blue in both areas. Male infants learned the colour words significantly slower than female infants, suggesting that the general advantage that has been seen in word learning for female infants also applies to colour word learning, although it is possible that there is an effect of the higher rates of colour vision problems in males.

The results show that the four basic colour terms measured in this report are learned at a similar age, as reported by parents. All four colour terms are learned mostly within a month of each other, however the data suggests that green may be learned significantly after the other three, with blue generally the first colour term learned. Green may in many cases be learned after the other three colours, but only by a short amount of time. Around three months on average separates learning the colour word and producing it - a gap that suggests that there is a very short time in between basic learning and production.

The colour word learning shown here suggests both comprehension and production at a much earlier age than has been previously shown. By 24 months, around $75 \%$ of infants have learned the four colour terms, and around $50 \%$ were already producing the terms. This result is in stark contrast to previous studies that suggested that at the earliest, some children


Figure 2: CDI estimates of proportion of British infants who understand colour terms at each age.
comprehended colour terms around 30 months (Sandhofer \& Smith, 1999; Pitchford \& Mullen, 2002). Here, parents report that roughly all of the children have learned the terms by this age.
The above results make startling claims about when colour word learning occurs. However, parental surveys have long been called into question, particularly in the case of measuring comprehension (Tomasello \& Mervis, 1994; HoustonPrice et al., 2007). Additionally, the parental reports only contain data from the four basic colour terms, which may not capture the full picture of colour word learning. In order to test the veracity of the claims made by the parental questionnaire, and to expand the scope of colours used from four to six, a second study was designed, using eye tracking data.

## Eye-tracking Study

## Method

Participants $N=146$ participants were recruited for this study, either online or from the local maternity ward. A further $N=23$ participants were excluded for fussiness or for failing to complete the experiment, while $N=5$ participants were excluded from the experiment for failing to complete at least one trial with each colour as both target and distractor. Full participant information can be seen in Table 3. Participants were in 5 age groups, with the oldest age group as a control group who have likely learned colour words (and thus were not selected to be a specific age), and the youngest age group a control group who likely have not yet learned colour terms. All participants were monolingual and were learning English as their first language.

Table 3: Descriptive statistics for participants in Exp 2

| Group | $N$ | Mean Age (months) | SD (months) |
| ---: | ---: | ---: | ---: |
| 12 | 30 | 11.84 | 0.70 |
| 16 | 29 | 15.96 | 0.70 |
| 19 | 31 | 19.69 | 0.73 |
| 24 | 28 | 24.30 | 0.36 |
| 48 | 28 | 53.46 | 18.78 |

Procedure Participants were seated on the lap of the caregiver, around 75 cm from the eye tracker and presentation screen. Participants first completed a nine-point calibration sequence, and following that the trials began. In each trial, participants saw an attractive attention-getter for 2 seconds, and then were presented with two objects. The objects were identical in every way, except for colour. The objects on the screen were coloured one of red, blue, green, yellow, black, or white. All colours were considered to be typical examples of that colour category by three independent observers, and not too dark or light. 2 seconds after the test objects appeared on the screen, participants heard a colour term corresponding to one of the objects, in the format of "look, red," "look at the red one" or "look at the red chair." The objects that made up the stimuli were all regular items that would be seen around the household by the infants, such as furniture or items of clothing. The trial would continue for another 5 seconds after colour label onset.

Each participant saw 18 trials, 3 of each colour, and all trials were left right randomised. Presentation was counter-


Figure 4: Fitted growth curve of looking to each colour after target word onset, separated by age group. Dotted red line indicated target word onset, dotted black line indicates chance looking.


Figure 3: A typical trial as seen by participants.
balanced so that any pairing that participants saw with one of the stimuli (e.g. the red chair) as a target, they would then see again with that as a distractor, and the stimulus that had been the distractor (e.g. the white chair) then became the target. All stimuli were presented on a neutral grey background, and trials were presented in a randomised order. Infant fixations were recorded using a Tobii eye tracker sampling at 120 Hz .

The data was modelled with a hierarchical binomial growth
curve, using the function glmmPQL, with quartic orthogonal polynomials (Mirman, 2014), with a random effect of subject. Age and time were treated as continuous numeric variables, and target colour was dummy coded. Rather than modelling the raw data, which allows infant colour preferences to override the effect of the label, trials were aggregated for each participant, such that the proportion of looks to target was defined by using the target and distractor for the same colour i:

$$
\text { Proportion }=\frac{\operatorname{TargetLooks}(i)}{\text { TargetLooks }(i)+\operatorname{DistractorLooks}(i)}
$$

## Results

The fitted model can be seen in Figure 4. The results indicate that there is very little colour word understanding at either 12 or 16 months, but by 19 months most of the participants understand all six colour words being examined. The effect of the label becomes more defined in the 24 month age group, until almost all participants react to the colour label by looking to the target in the 48 month group. The key fixation period is during the first two seconds after label onset. After that initial effect of the target label, it is likely that the infants return to random looking or looking by preference.

The results of this experiment also show that there is very little difference in the word learning of individual colour terms. Over the three month period between 16 months and 19 months, infants progress from knowing little of any of the colour terms to displaying some abilities in all of them. This shows that any differences in colour word learning must be of a smaller difference than three months.

The model output showed a significant effect of all four polynomials ( $p<0.0001$ ), of age ( $p<0.0001$ ) and of the interaction between age and each of the polynomials ( $p<$ 0.0001 ). The differences between colour terms had no significant effects on the model output (all $p>0.05$ ). This suggests that general patterns of looking at the target vary over the course of the trial by age, but do not really vary by colour.

Finally, the proportion of looks to the target was collapsed so that for each colour and each participant any participant with a proportion of looks to target over 0.55 (above chance) in the first two seconds after target word onset were coded as "looking to target," while any participant looking at a proportion lower than 0.55 were coded as "not looking." These answers were then compared with the responses given by parents in the Oxford CDI for those participants by way of a Chisquared test with Yates' continuity correction. $N=6$ participants were excluded from this analysis as they did not fill out the Oxford CDI. Results showed a significant relationship between CDI answers and infants performance in the eye-tracking task, $\chi^{2}(1, N=560)=22.974, p<0.0001$.

## General Discussion

Two experiments were presented in order to measure colour word learning, the first using parental reports of whether their children were understanding and producing the terms, the second an experimental study using eye-tracking to measure whether infants looked to the coloured object upon hearing the colour label. Both methods have shown that there is early comprehension of colour terms, before 24 months in many participants according to the CDI results. The purpose of the second study was to verify the early comprehension that was evidenced in the parental reports. In the parental reports, the majority of participants were thought to have comprehended all four colour terms in the Oxford CDI by around 24 months. At 19 months, however, only $25 \%$ of participants were reported to have understood the colour terms. It is clear from the eye-tracking data that there is some understanding of the colour terms already by 19 months. Not only does the experimental evidence confirm that colour words are learned early, but it suggests that parents may in fact be underestimating how much their children know with respect to colour terms.

The results of each study showed that although there maybe patterns in the order of acquisition of colour terms in English, the differences between those colour terms are not great - in the parental reports green was seen to be significantly behind the other three colour terms, but only by a difference of around a month, while in the eye-tracking study, none of the colours were learned by the majority of partic-
ipants at 16 months, while at 19 months all were learned. This suggests that there is a great deal of consistency as to when colour words are learned, and follow the assertion of Mervis et al. (1995) that following the learning of the first colour term, others quickly follow.

The results of the present study show that colour word learning occurs as young as 18 months, significantly earlier than has been shown by previous studies (Kowalski \& Zimiles, 2006; O’Hanlon \& Roberson, 2006; Pitchford \& Mullen, 2002; Sandhofer \& Smith, 1999; Soja, 1994). However, both elements of the present study only addressed typical examples of the colours in question, not a full adult-like understanding of colour terms. Previous studies focussed on atypical examples of the colour categories in order to examine colour word learning, but evidence presented here suggests that it is highly likely that this may have led to an underestimation, as the participants may have understood typical examples, but been unable to extend the colour categories in the same way an adult would (Wagner et al., 2013).

In the present study evidence is given for an early comprehension around 19 months of age, in contrast to this, Wagner et al. $(2013,2014)$ have shown that errors are still consistently made by participants around three years of age when presented with atypical examples of the colour category. This suggests that colour words are learned slowly over a long period of time, where an early comprehension precedes production, and that comprehension develops slowly until they have achieved an adult-like understanding (Wagner et al., 2014).

In this sense, colour words behave the same as other classes of word, including concrete nouns (Andersen, 1975), where children begin with typical examples of the word, but take time to learn the words closer to the boundary (Wagner et al., 2014; Yurovsky, Wagner, Barner, \& Frank, 2015). Often the examples closer to the category boundary could be learned after production, while the typical examples will be learned before production. Very similar examples have been seen in the case of time words (Tillman \& Barner, 2015), where children understand the order of time words, but not the exact meaning until much later.

The present study also compared the results of a parental report with those of an experimental eye-tracking study. The eye-tracking data successfully corroborates the results of the parental reports, with results suggesting a close relationship between both measures. Measuring comprehension through the use of questionnaires has long been questioned (HoustonPrice et al., 2007; Tomasello \& Mervis, 1994), despite evidence that it is a successful measure of comprehension at least in the case of object labels (Dale, 1991; Styles \& Plunkett, 2009). The data evidenced here suggests that parents estimate colour word comprehension conservatively, possibly due to comprehension of a colour term being more abstract and thus harder to observe than a more concrete term. Even in the abstract case of colour terms, parental reports provide a useful estimate of comprehension, but parents are sensitive to small improvements that children make in their vocabularies
at a young age, and as such this captures comprehension at an early stage of the process. While parental reports are found here to be largely consistent with CDI reports of their understanding of these individual terms, little is known about how the rest of the colour vocabulary will develop after they learn their first few terms.

Colour words in many respects behave like other classes of words; they are learned early, but it takes infants some time to establish where the boundaries are located, and find an adultlike definition. We have provided strong evidence that colour word comprehension occurs much earlier than thought, preceding production and slowly developing for a number of years. It is worth noting all participants in this study were British monolingual infants and toddlers, learning English. The order and timing of early comprehension of colour words at this stage is only known for British English; it remains to be seen whether the same trends apply to colour word learning globally.

## Acknowledgments

We thank the members of the Oxford BabyLab for assistance in booking participants and gathering CDI data, and to Caroline Floccia and members of the Plymouth Babylab for gathering and sharing CDI data. We are grateful to all the parents and infants who participated in this research. S.H.F. is funded by the Rhodes Trust.

## References

Andersen, E. S. (1975). Cups and glasses : learning that boundaries are vague. Journal of Child Language, 2(September), 79-103.
Andrick, G. R., \& Tager-Flusberg, H. (1986). The acquisition of colour terms. Journal of Child Language, 13(1), 119134.

Dale, P. S. (1991). The validity of a parent report measure of vocabulary and syntax at 24 months. Journal of speech and hearing research, 34(3), 565-571.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in Early Communicative Development. Monographs of the Society for Research in Child Development, 59(5), 174-185.
Franklin, A. (2006). Constraints on children's color term acquisition. Journal of Experimental Child Psychology, 94(4), 322-327.
Hamilton, A., Plunkett, K., \& Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. Journal of Child Language, 27(3), 689-705.
Heider, E. R. (1971). "Focal" color areas and the development of color names. (Vol. 4) (No. 3). US: American Psychological Association.
Houston-Price, C., Mather, E., \& Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. Journal of Child Language, 34(4), 701-724.

Kowalski, K., \& Zimiles, H. (2006). The relation between children's conceptual functioning with color and color term acquisition. Journal of Experimental Child Psychology, 94(4), 301-321.
Mervis, C. B., Bertrand, J., \& Pani, J. R. (1995). Transaction of cognitive-linguistic abilities and adult input: A case study of the acquisition of colour terms and colour-based subordinate object categories. British Journal of Developmental Psychology, 13(3), 285-302.
Mirman, D. (2014). Growth Curve Analysis and Visualization Using $R$ Analysis and Visualization Using R. Boca Raton, FL: Chapman \& Hall / CRC Press.
O'Hanlon, C. G., \& Roberson, D. (2006). Learning in context: Linguistic and attentional constraints on children's color term learning. Journal of Experimental Child Psychology, 94(4), 275-300.
Pitchford, N., \& Mullen, K. (2003). The development of conceptual colour categories in pre-school children: Influence of perceptual categorization. Visual Cognition, 10(1), 51-77.
Pitchford, N., \& Mullen, K. T. (2002). Is the acquisition of basic-colour terms in young children constrained? Perception, 31(11), 1349-1370.
Sandhofer, C. M., \& Smith, L. B. (1999). Learning color words involves learning a system of mappings. Developmental Psychology, 35(3), 668-679.
Shatz, M., Behrend, D., Gelman, S. A., \& Ebeling, K. S. (1996). Colour term knowledge in two-year-olds: evidence for early competence. Journal of Child Language, 23(1), 177-199.
Soja, N. N. (1994). Young Children's Concept of Color and Its Relation to the Acquisition of Color Words. Child Development, 65(3), 918-937.
Styles, S., \& Plunkett, K. (2009). What is 'word understanding' for the parent of a one-year-old? Matching the difficulty of a lexical comprehension task to parental CDI report. Journal of Child Language, 36(4), 895-908.
Tillman, K. A., \& Barner, D. (2015). Learning the language of time: Children's acquisition of duration words. Cognitive Psychology, 78, 57-77.
Tomasello, M., \& Mervis, C. B. (1994). The instrument is great but measuring comprehension is still a problem. Monographs of the Society for Research in Child Development, 59(5), 174-179.
Wagner, K., Dobkins, K., \& Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. Cognition, 127(3), 307-317.
Wagner, K., Jergens, J., \& Barner, D. (2014). Partial Color Word Comprehension Precedes Production. Proceedings CogSci, 1724-1729.
Yurovsky, D., Wagner, K., Barner, D., \& Frank, M. C. (2015). Signatures of Domain-General Categorization Mechanisms in Color Word Learning. Proceedings CogSci.

# Word Identification Under Multimodal Uncertainty 

Abdellah Fourtassi<br>afourtas@stanford.edu<br>Department of Psychology<br>Stanford University

Michael C. Frank<br>mcfrank@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

Identifying the visual referent of a spoken word - that a particular insect is referred to by the word "bee" - requires both the ability to process and integrate multimodal input and the ability to reason under uncertainty. How do these tasks interact with one another? We introduce a task that allows us to examine how adults identify words under joint uncertainty in the auditory and visual modalities. We propose an ideal observer model of the task which provides an optimal baseline. Model predictions are tested in two experiments where word recognition is made under two kinds of uncertainty: category ambiguity and distorting noise. In both cases, the ideal observer model explains much of the variance in human judgments. But when one modality had noise added to it, human perceivers systematically preferred the unperturbed modality to a greater extent than the ideal observer model did.


Keywords: Language; audio-visual processing; word learning; speech perception; computational modeling.

Language uses symbols expressed in one modality (e.g., the auditory modality, in the case of speech) to communicate about the world, which we perceive through many different sensory modalities. Consider hearing someone yell "bee!" at a picnic, as a honeybee buzzes around the food. Determining reference involves processing the auditory information and linking it with other perceptual signals (e.g., the visual image of the bee, the sound of its wings, the sensation of the bee flying by your arm).

This multimodal integration task takes place in a noisy world. On the auditory side, individual acoustic word tokens are almost always ambiguous with respect to the particular sequence of phonemes they represent, which is due to the inherent variability of how a phonetic category is realized acoustically (e.g., Hillenbrand, Getty, Clark, \& Wheeler, 1995). And some tokens may be distorted additionally by mispronunciation or ambient noise. Perhaps the speaker was yelling "pea" and not "bee." Similarly, a sensory impression may not be enough to make a definitive identification of a visual category. ${ }^{1}$ Perhaps the insect was a beetle or a fly instead.

Thus, establishing reference requires reasoning under a great deal of uncertainty in both modalities. The goal of this work is to characterize such reasoning. Imagine, for example, that someone is uncertain whether they heard "pea" or "bee", does it make them rely more on the visual modality (e.g., the object being pointed at)? Vice versa, if they are not sure if they saw a bee or a fly, does that make them rely more on the auditory modality (i.e., the label)? More importantly, when input in both modalities is uncertain to varying degrees, do

[^67]they weight each modality according to its relative reliability, or do they over-rely on a particular modality?

In this paper, we propose a probabilistic framework where such reasoning can be expressed precisely. We characterize uncertainty in each modality with a probability distribution, and we predict ideal responses by combining these probabilities in a optimal way. Our work can be seen as an extension to previous Bayesian models of phoneme identification (e.g., Feldman, Griffiths, \& Morgan, 2009), where, instead of a unimodal input, we model a bimodal one. A few studies have explored some aspects of audio-visual processing in a probabilistic framework (e.g., Bejjanki, Clayards, Knill, \& Aslin, 2011). In these studies, the researchers focused on the specific case of phoneme recognition from speech and lip movement, however, where information is tightly correlated across modalities.

In the present work, we study, rather, the case of arbitrary mapping between sounds and visual objects. We test participants on their ability to process audio-visual stimuli and use them to recognize the underlying word. More precisely we study the case where both the word's form and the word's referent are ambiguous, and we examine the extent to which humans conform to, or deviate from the predictions of the ideal observer model. Moreover, some previous studies on audiovisual processing documented cases of modality preference, when people rely predominantly on the visual modality (e.g., Colavita, 1974) or the auditory modality (e.g., Sloutsky \& Napolitano, 2003). Thus, we will explore if participants in our task show evidence of a modality preference.

The paper is organized as follows. First, we introduce our audio-visual recognition task. We next present the ideal observer model. Then we present two behavioral experiments where we test word recognition under audio-visual uncertainty. In Experiment 1, audio-visual tokens are ambiguous with respect to their category membership. In Experiment 2, we intervene by adding noise to one modality. In both experiments participants show qualitative patterns of optimal behavior. Moreover, while participants show no modality preference in Experiment 1, in Experiment 2 they over-rely on visual input when the auditory modality is noisy.

## The Audio-Visual Word Recognition Task

We introduce a new task that tests audio-visual word recognition. We use two visual categories (cat and dog) and two auditory categories (/b/ and /d/ embedded in the minimal pair /aba/-/ada/). For each participant, an arbitrary pairing is set between the auditory and the visual categories, leading to two audio-visual word categories (e.g., dog-/aba/, cat-/ada/).


Figure 1: Overview of the task

In each trial, participants are presented with an audiovisual target (the prototype of the target category), immediately followed by an audio-visual test stimulus (Figure 1). The test stimulus may differ from the target in both the auditory and the visual components. After these two presentations, participants press "same" or "different."

This task is similar to the task introduced by Sloutsky and Napolitano (2003) and used in subsequent research to probe audio-visual encoding. However, unlike this previous line of research, here participants are not asked whether the two audio-visual presentations are identical. Instead, the task is category-based: They are asked to press "same" if they think the second item (the test) belonged to the same category as the first (target) (e.g., dog-/aba/), even if there is a slight difference in the word, in the object, or in both. They are instructed to press "different" only if they think that the second stimulus was an instance of the other word category (cat-/ada/).

The task also includes trials where pictures were hidden (audio-only) or where sounds were muted (visual-only). These unimodal trials provide us with participants' categorization functions for the auditory and visual categories and are used as inputs to the ideal observer model, described below.

## Ideal Observer Model

The basis of our ideal observer model is that individual categorization functions from each modality should be combined optimally. In each modality, we have two categories: /ada/ ( $A=1$ ) and /aba/ $(A=2)$ in the auditory dimension, and cat $(V=1)$ and $\operatorname{dog}(V=2)$ in the visual dimension. We assume, for the sake of simplicity, that the probability of membership in each category is normally distributed:

$$
\begin{aligned}
& p(a \mid A) \sim N\left(\mu_{A}, \sigma_{A}^{2}\right) \\
& p(v \mid V) \sim N\left(\mu_{V}, \sigma_{V}^{2}\right)
\end{aligned}
$$

In the bimodal condition, participants see word tokens with audio-visual input, and have to make a categorization decision. We define word tokens as vectors in the audio-visual space, $\mathbf{w}=(a, v)$. A word category $W$ is defined as the joint distribution of auditory and visual categories. It can be characterized with a bivariate normal distribution:

$$
p(\mathbf{w} \mid W) \sim N\left(M_{W}, \Sigma_{W}\right)
$$



Figure 2: Illustration of our model using simulated data. A word category is defined as the joint bivariate distribution of an auditory category (horizontal, bottom panel) and a visual semantic category (vertical, left panel). Upon the presentation of a word token $w$, participants guess whether it is sampled from the word category $W_{1}$ or from $W_{2}$. Decision threshold is where the guessing probability is 0.5 .

We have two word categories: dog-/aba/ ( $W_{1}$ ) and cat-/ada/ $\left(W_{2}\right)$. Participants can be understood as choosing one of these two word categories (Figure 2). For an ideal observer, the probability of choosing category 2 when presented with an audio-visual instance $\mathbf{w}=(a, v)$ is the posterior probability of this category:

$$
\begin{equation*}
p\left(W_{2} \mid \mathbf{w}\right)=\frac{p\left(\mathbf{w} \mid W_{2}\right) p\left(W_{2}\right)}{p\left(\mathbf{w} \mid W_{2}\right) p\left(W_{2}\right)+p\left(\mathbf{w} \mid W_{1}\right) p\left(W_{1}\right)} \tag{1}
\end{equation*}
$$

We make the assumption that, given a particular word category, the auditory and visual tokens are independent:

$$
\begin{equation*}
p(\mathbf{w} \mid W)=p(a, v \mid W)=p(a \mid W) p(v \mid W) \tag{2}
\end{equation*}
$$

Under this assumption, the posterior probability reduces to:

$$
\begin{equation*}
p\left(W_{2} \mid \mathbf{w}\right)=\frac{1}{1+(1+\varepsilon) \exp \left(\beta_{0}+\beta_{a} a+\beta_{v} v\right)} \tag{3}
\end{equation*}
$$

with $\beta_{a}=\frac{\mu_{A 1}-\mu_{A 2}}{\sigma_{A}^{2}}, \beta_{v}=\frac{\mu_{V 1}-\mu_{V 2}}{\sigma_{V}^{2}}, \beta_{0}=\frac{\mu_{A 2}^{2}-\mu_{A 1}^{2}}{2 \sigma_{A}^{2}}+\frac{\mu_{V 2}^{2}-\mu_{V 1}^{2}}{2 \sigma_{V}^{2}}$ and $1+\varepsilon=\frac{p\left(W_{1}\right)}{p\left(W_{2}\right)}$ is the proportion of the prior probabilities. If the identity of word categories is randomized, and if $W_{1}$ is the target, then $\varepsilon$ measures a response bias to "same" if $\varepsilon>0$, and a response bias to "different" if $\varepsilon<0$.

In sum, the posterior 3 provides the ideal observer's predictions for how probabilities that characterize uncertainty in
each modality can be combined to make categorical decision about bimodal input.

## Experiment 1

In Experiment 1, we test the predictions of the model in the case where uncertainty is due to similar auditory categories, and similar visual categories. Crucially, the similarity is such that the distributions overlap. To simulate such uncertainty in a controlled fashion, we use a continuum along the second formant (F2) linking the words /aba/ and /ada/, and we use a morph that links a dog prototype and a cat prototype.

## Methods

Participants We recruited a planned sample of 100 participants, recruited from Amazon Mechanical Turk. Only participants with US IP addresses and a task approval rate above $85 \%$ were allowed to participate. They were paid at an hourly rate of $\$ 6 /$ hour. Data were excluded if participants completed the task more than once ( 2 participants). Moreover, as specified in the preregistration (https://osf.io/h7mzp/), participants were excluded if they had less than $50 \%$ accurate responses on the unambiguous training trials (6), and if they reported having experienced a technical problem of any sort during the online experiment (14). The final sample consisted of 76 participants.

Stimuli For auditory stimuli, we used the continuum introduced in Vroomen, van Linden, Keetels, de Gelder, and Bertelson (2004), a 9-point /aba/-/ada/ speech continuum created by varying the frequency of the second (F2) formant in equal steps. We selected 5 equally spaced points from the original continuum by keeping the end-points (prototypes) 1 and 9, as well as points 3,5 , and 7 along the continuum. For visual stimuli, we used a morph continuum introduced in Freedman, Riesenhuber, Poggio, , and Miller (2001). From the original 14 points, we selected 5 points as follows: we kept the item that seemed most ambiguous (point 8 ), the 2 preceding points (i.e., 7 and 6 ) and the 2 following points (i.e., 9 and 10). The 6 and 10 points along the morph were quite distinguishable, and we took them to be our prototypes.

Design and Procedure We told participants that an alien was naming two objects: a dog, called /aba/ in the alien language, and a cat, called /ada/. In each trial, we presented the first object (the target) on the left side of the screen simultaneously with the corresponding sound. The target was always the same (e.g., dog-/aba/). The second sound-object pair (the test) followed on the other side of the screen after 500 ms and varied in its category membership. For both the target and the test, visual stimuli were present for the duration of the sound clip ( $\sim 800 \mathrm{~ms}$ ). We instructed participants to press "S" for same if they thought the alien was naming another dog-/aba/, and "D" for different if they thought the alien was naming a cat-/ada/. For each participant, we randomized the soundobject mapping as well as the identity of the target.

The first part of the experiment trained participants using only the prototype pictures and the prototype sounds (12 tri-


Figure 3: Average human responses in the auditory-only condition (left), and visual-only condition (right). A) represents data from Experiment 1, and B) data from Experiment 2. Error bars are $95 \%$ confidence intervals. Solid lines represent unimodal logistic fits.
als, 4 each from the bimodal, audio-only, and visual-only conditions). After completing training, we instructed participants on the structure of the task and encouraged them to base their answers on both the sounds and the pictures (in the bimodal condition). There were a total of 25 possible combinations in the bimodal condition, and 5 in each of the unimodal conditions. Each participant saw each possible trial twice, for a total of 70 trials/participant. Trials were blocked by condition and blocks were presented in random order.

## Results

Unimodal conditions this is the case where the pictures were hidden, or where the sounds were muted. Average categorization judgments and fits are shown in Figure (3, A). The categorization function of the auditory condition was steeper than that of the visual condition. The fit was done using the Nonlinear Least Squares (NLS) R package, as follows. For an ideal recognizer, the probability of choosing category 2 (that is, to answer "different") when presented with an audio instance $a$, is the posterior probability of this category $p\left(A_{2} \mid a\right)$. If we assume that both categories have equal variances, the posterior probability reduces to:

$$
\begin{equation*}
p\left(A_{2} \mid a\right)=\frac{1}{1+\left(1+\varepsilon_{A}\right) \exp \left(\beta_{a 0}+\beta_{a} a\right)} \tag{4}
\end{equation*}
$$

with $\beta_{a}=\frac{\mu_{A_{1}}-\mu_{A_{2}}}{\sigma_{A}^{2}}$ and $\beta_{a 0}=\frac{\mu_{A_{2}}^{2}-\mu_{A_{1}}^{2}}{2 \sigma_{A}^{2}}$. $\varepsilon_{A}$ is the response bias in the auditory-only trials.

For this model (as well all other models), we fixed the values of the means to be the end-points of the corresponding continuum: $\mu_{A 1}=0$ and $\mu_{A 2}=4$ (and similarly $\mu_{V 1}=0$, and $\mu_{V 2}=4$ ). To determine the values of the bias and the variance, we fit a model for each modality, collapsed across participants. For the auditory modality, we obtained $\varepsilon_{A}=-0.20$ and


Figure 4: Proportion of "different" judgments as a function of auditory distance. Solid lines represent average human responses (left), predictions of the ideal observer (middle), and the bimodal fit (right). Dashed lines represent average human responses in the unimodal conditions. Colors represent values in the visual continuum. A) represents data from Experiment 1, and B) data from Experiment 2.
$\sigma_{A}^{2}=2.04$. For the visual modality, we obtained $\varepsilon_{V}=-0.11$ and $\sigma_{V}^{2}=3.34$.
Bimodal condition We fit a model to human responses in the bimodal condition, collapsed across participants, finding $\varepsilon=-0.32, \sigma_{A b}^{2}=5.00$ and $\sigma_{V b}^{2}=7.27$. The fit explained $94 \%$ of total variance.

Ideal observer model We derived the predictions of the ideal observer model by using the values of the variances derived from the unimodal conditions, and the response bias derived from the bimodal condition, and by substituting these values into the expression of the posterior in Eq. 3. Figure $(4, \mathrm{~A})$ shows participants' responses in the bimodal condition (left), as well as the prediction of the ideal observer (middle), and the bimodal fit models (right).

Response bias We found negative values in all response bias terms, which suggests a general bias toward answering "different." This bias is probably due to the categorical nature of our same-different task: when two items are ambiguous but perceptually different, this could cause a slight preference for "different" over "same".

Modality preferences We next analyzed whether there was a preference for one or the other modality when making decisions in the bimodal condition, beyond that explained by the variance in categories implied by the unimodal responses. This preference would manifest as a deviation from the decision threshold predicted by the ideal observer model. The decision threshold is defined as the set of values in the audio-
visual space along which the posterior (Eq. 3) is equal to 0.5 . The decision threshold takes the following form:

$$
\begin{equation*}
v=-\frac{\sigma_{V}^{2}}{\sigma_{A}^{2}} a+v_{0} \tag{5}
\end{equation*}
$$

If the slope derived from the bimodal fit is greater than the slope of the ideal observer, this finding would suggest a general preference for the auditory modality (similarly, a smaller slope would suggest a preference for the visual modality). The limit cases are when there is exclusive reliance on the auditory cue (a vertical line), and where there is exclusive reliance on the visual (a horizontal line). Figure 5 (top left) shows the decision threshold in the audio-visual space with a constant intercept; the fit to human data (solid black line) was very close to the ideal observer threshold (red line). Nonparametric resampling of the data showed no evidence of a deviation from the slope of the ideal observer ( 5 , bottom left).

## Discussion

Qualitatively, participants' judgments were similar to the predictions of the ideal observer model (remember that the latter was obtained by optimally combining fits to the unimodal data). Consider, for example, the contrast between the auditory-only case (dashed black line in Figure 4, top left) and the bimodal case (solid colored lines). Higher certainty in the visual modality generally influenced responses, with greater visual distance leading to more "different" ratings and less visual distance leading to more "same" ratings. Similar ob-
servations can be made about the contrast between the visualonly case and the bimodal case.

Overall, we found that the ideal observer model explained much of the variance in judgments $\left(r^{2}=0.89\right)$. But although we see a qualitative resemblance between human data and the model, there were quantitative differences. For example, model predictions were more influenced by the visual modality at the auditory midpoint (the point of highest uncertainty) than human judgements, and were more compressed at the endpoints (the points of lowest uncertainty).

Formally, there was an increase in the value of the variance associated with each modality. Whereas the ideal observer model predicted the weights to be proportional to $1 / \sigma_{A}^{2}$ and $1 / \sigma_{V}^{2}$, for the auditory and the visual modalities, respectively (see expression 3), the fit to human data suggested that the real weights were proportional to $1 / \sigma_{A b}^{2}$ and $1 / \sigma_{V b}^{2}$. Our analysis of modality preference showed that the relative values of these variances were almost the same (Figure 5, left). Thus, 1) the bimodal presentation introduced a certain level of randomness in the participants' responses, and 2) this increased randomness did not affect the relative weighting of both modalities, i.e., participants were weighting modalities according to their relative reliability. The latter explains the qualitative resemblance between the predictions of the ideal observer and human data, and the former explains the quantitative discrepancy.

In sum, we found that participants followed the ideal observer model in that they weighted modalities according to their reliabilities. In real life, however, tokens can undergo distortions due to noisy factors in the environment. In Experiment 2, we explore this additional level of uncertainty.

## Experiment 2

Imagine that the speaker generates a target production $t$ from an auditory category $t \mid A \sim N\left(\mu_{A}, \sigma_{A}^{2}\right)$. In Experiment 1, we assumed that the observer could directly retrieve this production token. But if the observer is in a noisy environment, they do not hear exactly this produced target, but the target perturbed by noise, which we assume, following Feldman et al. (2009), that it is normally distributed: $a \mid t \sim N\left(t, \sigma_{N}^{2}\right)$. When we integrate over $t$, we get:

$$
\begin{equation*}
a \mid A \sim N\left(\mu_{A}, \sigma_{A}^{2}+\sigma_{N}^{2}\right) \tag{6}
\end{equation*}
$$

In this experiment, we explored the effect of this added noise ${ }^{2}$ on performance in our task. We tested a case where one modality was ambiguous and noisy (auditory), and where the other modality was ambiguous but not noisy (visual). We were interested to know if participants would treat this new source of uncertainty as predicted by the ideal observer model, and whether noise in one modality would lead to some systematic preference for the non-noisy modality.

[^68]
## Methods

Participants A planned sample of 100 participants was recruited online through Amazon Mechanical Turk. We used the same exclusion criteria as in the previous experiment; the final sample consisted of 93 participants.

Stimuli and Procedure We used the same visual stimuli as in Experiment 1. We also used the same auditory stimuli, but we convolved each item with Brown noise of amplitude 1 using the audio editor Audacity (2.1.2). The procedure was exactly the same as in the previous experiment, except that test stimuli were presented with the new noisy auditory stimuli.

## Results

Unimodal conditions We fit a model for each modality, collapsed across participants. For the auditory modality, our parameter estimates were $\varepsilon_{A}=-0.18$ and $\sigma_{A}^{2}+\sigma_{N}^{2}=4.70$. For the visual modality, we found $\varepsilon_{V}=-0.24$ and $\sigma_{V}^{2}=3.93$. Figure 3 (bottom) shows responses in the unimodal conditions as well as the unimodal fits. In contrast to Experiment 1, auditory responses were flatter (showing more uncertainty).
Bimodal condition We fit a model to human responses in the bimodal condition, collapsed across participants. We estimated $\varepsilon=-0.38, \sigma_{V b}^{2}=5.21$, and $\sigma_{A b}^{2}+\sigma_{N b}^{2}=9.84$. The fit explained $97 \%$ of total variance.
ideal observer model We generated the predictions of the ideal observer model by using the values of the variances derived from the unimodal conditions, and the response bias derived from the bimodal condition, and by substituting these values into the expression of the posterior in Eq. 3. Results are shown in Figure 4 (bottom).

Modality preferences Participants' decision threshold suggested a preference for the visual modality (the non-noisy modality). Indeed non-parametric resampling of the data showed a decrease in the value of the slope ( 5 , right).

## Discussion

We found, similar to Experiment 1, that participants generally showed qualitative patterns similar to the ideal observer model ( $r^{2}=.91$ ). But we also found a similar discrepancy at the quantitative level. The ideal observer model predicted the modality weights to be proportional to $1 /\left(\sigma_{A}^{2}+\sigma_{N}^{2}\right)$ and $1 / \sigma_{V}^{2}$, for the auditory and the visual modalities, respectively. The fit to human data suggested that the empirical weights were proportional to $1 / \sigma_{A b}^{2}$ and $1 / \sigma_{V b}^{2}$. An interesting difference with Experiment 1, however, was that participants had a clear preference for the non-noisy modality, as the values of the relative variances were different (Figure 5, right). This preference affected the relative weighting, where, contrary to Experiment 1, the visual modality had greater weight than what could be expected from its relative reliability alone.

It is important to understand that this preference was not the mere consequence of the added noise increasing the variance of the auditory modality, since this increase was already


Figure 5: Top: decision thresholds in the audio-visual space. Red dotted line is the prediction of the ideal observer. Blue dotted lines are cases where modality preference is twice as strong as the ideal observer. Solid line is the threshold derived from human data. Bottom: comparison of the threshold slope between the ideal observer and the fit to human data. Error bars are $95 \%$ confidence intervals computed via nonparametric bootstrap.
accounted for in the ideal observer model. The preference was, rather, a form of over-reliance on the visual modality.

## General Discussion

Understanding language requires both the ability to integrate multimodal input, and the ability to deal with uncertainty. In this work, we explored a case where both abilities were at play. We studied the case of identifying a word when both its form (auditory) and its referent (visual) were ambiguous with respect to their category membership (Experiment 1), and when the form was perturbed with additional noise (Experiment 2). We introduced a model that instantiated an ideal observer, predicting how information from each modality could be combined in an optimal way. In both experiments, participants showed the qualitative patterns of the ideal observer.

There were, however, quantitative differences. Audiovisual presentation increased the level of randomness in the participants' responses. One possible explanation is that this phenomenon was caused by the arbitrary nature of the formmeaning mapping. Previous studies suggest that while redundant multimodal information improves performance (e.g., determining the frequency of a bouncing ball from visual and auditory cues), arbitrary mappings generally tends to hinder performance (for review, see Robinson \& Sloutsky, 2010).

Interestingly, however, in Experiment 1 this increase in randomness occurred at a similar rate for both the auditory and the visual modality, and thus, it did not affect their relative weighting. The latter was primarily determined by in-
formational reliability. Only when we intervened by adding noise to one modality in Experiment 2, did participants show a systematic preference for the non-noisy modality. One possible explanation for this preference could be that people do not combine cross-modal uncertainties of a similar kind (e.g., ambiguity in both modalities) in the same way they would combine uncertainties of different kinds (e.g., ambiguity in one modality and noise in the other). For instance, it could be that the latter, but not the former, cause the over-reliance on a particular modality.

Overall, in both Experiments, the majority of the variance could be explained by an ideal observer that combined multimodal information optimally. In the light of this main result, we can revisit some previous findings in the literature. For instance, Sloutsky and Napolitano (2003) reported a dominance for the auditory modality in children. This dominance disappears or reverses in adults. Could this difference be driven by changes across development in the level of perceptual noise affecting the intrinsic relative reliability of modalities (by analogy to Experiment 2)? More work is needed to carefully examine this (and other) speculations, and more generally, to determine the extent to which the optimal combination account helps us better understand the mechanisms of word processing and learning.

## Acknowledgements

This work was supported by a post-doctoral grant from the Fyssen Foundation.

## References

Bejjanki, V. R., Clayards, M., Knill, D. C., \& Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. PLoS ONE, 6 .
Colavita, F. B. (1974). Human sensory dominance. Perception \& Psychophysics, 16.
Feldman, N., Griffiths, T., \& Morgan, J. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. Psychological review, 116(4), 752-782.
Freedman, D., Riesenhuber, M., Poggio, T., , \& Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science, 291.
Hillenbrand, J., Getty, L. A., Clark, M. J., \& Wheeler, K. (1995). Acoustic characteristics of american english vowels. Journal of the Acoustical Society of America, 97.
Robinson, C. W., \& Sloutsky, V. M. (2010). Development of cross-modal processing. Wiley Interdisciplinary Reviews: Cognitive Science, 1 .
Sloutsky, V. M., \& Napolitano, A. (2003). Is a picture worth a thousand words? preference for auditory modality in young children. Child Development, 74.
Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., \& Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. Speech Communication, 44.

# Could both be right? Children's and adults' sensitivity to subjectivity in language 

Ruthe Foushee (foushee@berkeley.edu) Mahesh Srinivasan (srinivasan@ berkeley.edu)<br>Department of Psychology, University of California, Berkeley, Berkeley, CA 94704


#### Abstract

While some word meanings, like "spotted," depend on intersubjectively accessible properties of the world, others like "pretty" invoke speakers' subjective beliefs. We explored children and adults' sensitivity to the subjectivity of a range of adjectives, including words like "spotted" and "pretty," but also words like "tall," which are evaluated relative to a standard. Participants saw two speakers who had independently experienced sets of exemplars of a novel object kind disagree about whether a critical exemplar was, e.g., "tall," "pretty," and "spotted." In Experiments 1 and 3, speakers had seen distinct sets of exemplars, while in Experiments 2 and 4, the sets were identical. Adults always judged disagreements over words like "pretty" as faultless-indicating that both speakers "could be right"-and permitted less faultless disagreement for ones like "tall" when the speakers had experienced identical sets of exemplars. Strikingly, children did not respond in an adult-like manner until age 8 or 9 , but their explanations for speakers' conflicting assertions suggested some sensitivity to the kinds of knowledge relevant for evaluating different adjectives.


Keywords: metalinguistic development; theory of mind

## Introduction

What goes into our understanding of what other people say? While comprehension of some words, like "spotted" or "striped," does not appear to rely on our knowledge of others' beliefs and experiences, comprehension of other words, like "pretty" or "tasty," does. Thus, when someone says that they saw a "spotted bird," we can understand what they mean by leveraging our understanding of what "spotted" and "bird" refer to. We understand that the bird must have some spots on it to be called "spotted," and that were we also to see the bird, we would agree with the speaker's description. The meaning of "spotted" is in this sense intersubjective, based on properties in the external world all speakers can access. In contrast, when someone says that they saw a "pretty bird," it is not immediately apparent what property of the bird she is describing, nor that we would agree that the same bird is "pretty." This is because the meanings of words like "pretty" are not intersubjective, but instead depend on the speaker's belief.

The present studies are motivated by the idea that many words-far beyond clearly subjective predicates of personal taste like "pretty," "tasty," and "funny"-are interpreted relative to their speakers. In particular, we consider the cases of relative adjectives like "big" and "cold," which have to be interpreted relative to the nouns they modify: a "big bird" is smaller than a "big lion." We test whether these predicates are also interpreted relative to the prior experiences and beliefs of speakers. Do we adjust the imagined size of a "big apartment" or temperature of a "cold day" depending on the sample of apartments and weather we believe our interlocutor has experienced? Speakers may have different thresholds for calling an apartment "big" or a day "cold" depending on their
prior experiences, such that they may not always agree about whether a specific apartment or day is "big" or is "cold."

As we review below, the fact that speakers may have different things in mind when using words like "big" and "pretty" may pose a challenge for successful communication. This is especially true in light of evidence that children (and even adults) have an overarching tendency toward naive realism; i.e., to behave as though their own perception reflects reality and their judgments are objective (Ross \& Ward, 1996).

## Background

Previous work demonstrates that children have sophisticated knowledge of relative adjectives, but leaves open whether they incorporate information about their interlocutors into their interpretations. Four-year-olds understand that the meanings of words like "big" and "tall" depend on distributions of referents within a given class. For example, they appropriately identify "tall pimwits" as ones at the higher end of the distribution of only pimwit heights, even if that means ignoring other, taller objects (Barner \& Snedeker, 2008). Five-year-olds also understand that the frame of reference for what counts as "high" or "low" varies with the class of object in question (e.g., "high" for a bird is much higher than "high" for a bunny: Smith, Cooney, \& McCord, 1986).

One reason to think that children may have difficulty interpreting the meanings of words that depend on speakers' beliefs is that they appear to begin life heavily influenced by naive realism, and behave as if their own construal of a stimulus will be shared by others. In one study, for example, children were shown an image that was then covered so only an unidentifiable quadrant of it was left visible. Four-year-olds predicted that others would still be able to identify the largely occluded image, seemingly discounting their previous subjective access to it in full (Taylor, Cartwright, \& Bowden, 1991).

In general, the literature suggests that young children might be able to understand why people say the things they say, but still have difficulty thinking that others can have different meanings for words than they do, perhaps due to their fundamental assumptions about language itself. For example, although toddlers recognize that others might want a different snack from one that they themselves find delicious (Repacholi \& Gopnik, 1997), they judge statements of unconventional snack opinions like "ice cream is yucky" as unacceptable well into the preschool years (Holubar, 2015). Thus, preschoolaged children understand that others may have different preferences, but struggle with understanding that an unqualified statement about a preference that they think is wrong can still be "right." Children's eventual success on false belief tasks (e.g., Wellman \& Liu, 2004) demonstrates their understanding that an individual's experience leads to their beliefs. But
it may be more challenging for children to understand that experience might result in speakers of the same language having different meanings for the same word (e.g., such that speakers have different temperatures in mind when they talk about a "cold day"). Consistent with this, studies on children's beliefs about conventionality in language have argued that children expect object labels to be shared by other speakers of their language, even given evidence to the contrary (e.g., when other speakers were absent when a novel object's label was taught: Diesendruck, 2005).

While adults "succeed" on some of these tasks that stump children, they are also not immune to the influence of naive realism, suggesting some form of continuity over development. For example, adults often overestimate the prevalence of their own attitudes in the general population, and are reluctant to attribute those attitudes to their own subjective experience, rather than to objective features of the world (see Ross \& Ward, 1996 for a review). When it comes to language, adults have the metalinguistic knowledge to be able to explicitly judge words and phrases like "pretty shirt" as subjective when they are presented in the absence of a referent that they could evaluate (Scontras, Degen, \& Goodman, 2016). However, to our knowledge there have been no empirical investigations of whether adults permit different word meanings for speakers when the adults themselves are confident of whether the word applies: e.g., when they are confident that a particular shirt is "pretty" or "big." Additionally, no studies have explored whether adults' tolerance of disagreement about different word uses is influenced by their knowledge of a speaker's relevant prior experience.

## The Present Studies

Here, we explore adults' and children's sensitivity to linguistic subjectivity across four experiments. In particular, we ask whether, in addition to considering the real-world distribution of a specific noun's referents along a given dimension, listeners also interpret adjectives like "tall" relative to what they know of the distribution that the speaker has experienced. To test these ideas, we manipulate whether two speakers experience different distributions of exemplars of a novel object kind (Exps. 1 and 3), or identical distributions of exemplars (Exps. 2 and 4), and assess whether this affects adults’ (Exps. 1 and 2) and children's (Exps. 3 and 4) judgments of whether the two speakers can disagree about how to describe a novel target exemplar that they can both see.

Across our studies, the disagreeing assertions that participants judge involve adjective-noun phrases that describe the same target object: e.g., "That's [not] a tall pimwit." We introduce novel nouns, but use familiar gradable adjectives (GAs) that vary in how intersubjective versus subjective they are. We categorize these adjectives into three classes. Following Syrett, Kennedy, and Lidz (2009), we call words like "spotted" absolute GAs. These adjectives require their arguments to possess some minimal degree of a property, and their meaning is largely context-independent. We refer to contextdependent adjectives like "tall" as relative GAs (Syrett et al.,
2009), and refer to adjectives like "pretty" as subjective GAs.

To assess individuals' appreciation of the subjectivity of these different kinds of adjectives, we obtain judgments of faultless disagreement: disagreements where neither person is wrong (Barker, 2013). Such judgments are closely correlated with direct measures of statements' subjectivity (Scontras et al., 2016), and offer a less metalinguistically demanding measure to use with children. In addition, we elicit qualitative explanations from participants to understand the sources of knowledge that they are drawing on when evaluating speakers' utterances. Critically, given that participants maintain visual access to the complete distribution of exemplars observed by both speakers, they are able to form their own evaluation of whether the adjective-noun phrase applies to the target exemplar that is the subject of the speakers' disagreement. Since they share this evaluation with only one of the two disagreeing speakers (e.g., one will call the pimwit "pretty" and one "not pretty"), we can interpret judgments of faultess disagreement as overcoming naive realism.

## Experiment 1

Participants Twenty-five UC Berkeley undergraduates participated in Experiment 1 (18 women, 19.65-27.37, $M=$ 21.24 years, $S D=1.68$ years). All were native speakers of English and received course credit for their participation.

## Stimuli and Methods



Speakers exposed to DISTINCT distributions (Exps. 1 \& 3)


Figure 1: Schematic of experimental setup for Exps. 1-4.

Experimental Setup The stimuli were sets of eleven objects belonging to distinct novel kinds. Critical kinds were pimwits, thin purple cylinders ranging from 0.75 to 6.25 inches in height, and from densely to sparsely spotted, and daxes, blue and yellow spheres ranging from 0.5 to 3 inches in diameter, and from heavily to lightly striped. Each set was divided into two arrays composed of the five smallest and five largest exemplars, with the exemplar in the middle of the size distribution used as the critical target exemplar (Figure 1).

Participants sat across a table from two wooden house-like structures separated by a narrow stage. The experimenter sat behind the display and animated pairs of puppets representing familiar characters from Sesame Street, who she
explained could not see or hear anything that happened beyond their "classrooms" while they were inside them.

The experiment consisted of two blocks of two training trials each, followed by two blocks of three critical trials each and a post-test. In each block, two speakers were independently introduced to distinct arrays of a novel kind in their classrooms by the experimenter, before emerging to view and disagree about a new exemplar (the target) placed by the experimenter in the middle of the stage.

Training Trials The initial training trials familiarized participants with the paradigm, and provided practice with judging disagreements as faultless and not. In them, characters saw distinct sets of exemplars labeled with the same noun, and then 'disagreed' over a target exemplar that shared properties with both sets. In a faultless training trial, Dawn might see five feps that were matte white circles, while Big Bird saw five feps that were sparkly white squares. Dawn would exclaim that the target fep, a sparkly white circle, was "sparkly," while Big Bird would assert that it was "round" (a faultless disagreement). The non-faultless complement would consist of one speaker asserting the target exemplar was "white," while the other said it was "black." Participants received feedback for their answers on only the first block of training trials, and we recorded their judgments prior to feedback.

Critical Trials In critical trials, the characters were introduced to distinct sets of exemplars belonging to the same novel kind (pimwits or daxes). For example, Zoe might see five relatively short and densely spotted pimwits, while Cookie Monster saw five relatively tall and lightly spotted ones. Upon encountering the intermediate target pimwit, Zoe would assert that it was tall, which Cookie Monster would deny. Following the disagreement, participants answered whether each speaker was "wrong" or "could be right," and explained why. Responses where participants answered "could be right" for both speakers were coded as indicating faultless disagreement. For each novel kind, speakers disagreed over an absolute, relative, and subjective GA.

The order of the blocks, which speaker asserted the positive statement, and the block-internal order of the relative versus subjective disagreements were counterbalanced across participants. Disagreements over absolute GAs were always presented last to avoid invalidating one of the speakers. To prevent speakers from being degraded across blocks for wrong assertions, new speakers were introduced each block of trials.

Qualitative Explanations In each critical trial, we collected qualitative explanations of participants' evaluations of speakers' assertions. From explanations collected during piloting, we developed three primary codes to describe participants' responses. Trained coders identified whether each explanation made reference to apparently intersubjective properties of the target exemplar (Object Property-e.g., "It is beautiful;" "There are dots on it"), the distinct arrays
of exemplars the speakers had experienced (SPEAKER EX-PERIENCE-e.g., "He saw tall pimwits and she saw short ones"), or the speaker's subjective evaluation of the object (Speaker Opinion-e.g., "He likes purple and she doesn't like spots."). Explanations could receive multiple codes.

Post-Test We directly assessed participants' own evaluation of the target exemplars in a post-test. Participants saw the entire distribution of exemplars, and answered whether the target exemplars were "spotted," "tall," "pretty," etc.

## Results

Faultless Disagreement Disagreements over relative and subjective GAs were almost always judged as faultless (spotted: $24 \%$, striped: $20 \%$, tall: $100 \%$, big: $100 \%$, pretty: $100 \%$, boring: $92 \%$; see Figure 2). There were no significant differences between the proportions of faultless disagreement for the two adjectives in each class, so we collapse them here.


Figure 2: Adult rates of faultless disagreement judgment during critical trials in Exps. 1-2 by gradable adjective type (ABSOLUTE: "spotted," "striped;" RELATIVE: "tall," "big;" SUBJECTIVE: "pretty," "boring"). Participants in Exp. 1 judged speakers exposed to distinct distributions of exemplars, while participants in Exp. 2 judged speakers who had seen identical ones. Error bars for this and all plots indicate $95 \%$ bootstrapped confidence intervals $(k=1000)$.

Relation to Post-test The relation between participants' post-test evaluations and faultless disagreement judgments qualitatively distinguished the three classes of adjectives. For absolute GAs, participants uniformly answered "yes" when asked whether the target pimwit or dax was "spotted" or "striped," and typically answered that only the speaker who asserted the same could be right. For relative GAs, in contrast, while participants again all said that the target exemplar was "tall" or "big," they also all responded that both speakers could be right. Despite variability in participants' own evaluations of the critical items' beauty ( $91 \%$ said it was "pretty") or tedium ( $17 \%$ said it was "boring"), they almost always judged disagreements over subjective GAs as faultless.

Qualitative Explanations Participants for the most part cited distinct sources of knowledge to explain their judgments of utterances from different adjective classes (Figure 3). They referred to speakers' opinions (SPEAKER OPINION) exclusively when explaining their evaluations of utterances using
subjective GAs (and did so on $80 \%$ of all subjective trials).
We fit separate logit models to the data for the two remaining codes (Object Property and Speaker Experience) that were used in explanations regarding more than one adjective class, using GA type as our sole predictor. Participants were more likely to refer to object properties to explain absolute GA utterances ( $\beta=1.266, p<0.001$ ), and less likely to cite them when explaining subjective GA judgments ( $\beta=-3.462, p<0.001$ ). Participants cited speakers' unique experiences most in explaining judgments of relative GA utterances, and were unlikely to do so to explain judgments of absolute GAs $(\beta=-1.153, p<0.001)$.


Explanation Code ■Object Property 【Speaker Experience■Speaker Opinion

Figure 3: Adult reference to properties of the target exemplar (Object Property), speakers' distinct experiences of the object kind (SPEAKER EXPERIENCE), and speakers' subjective opinions (Speaker Opinion), in explaining their evaluations of assertions. Panels reflect proportions of each code for explanations regarding a given GA type (in columns) in a given experiment (in rows).

## Experiment 2

Having demonstrated that adults readily judge disagreements over relative and subjective GAs as faultless when speakers have experienced distinct distributions of exemplars, we explored the limits of listeners' acceptance of subjective meanings by equating the disagreeing speakers' experiences.

Participants 33 undergraduate adults ( 26 women, 18.1039.83 years, $M=20.91, S D=3.52$ ) participated.

## Stimuli and Methods

The experimental paradigm was identical to that of Experiment 1 with two changes: 1) speakers saw identical distributions of exemplars in their respective classrooms, and 2) we introduced an additional, plain (i.e., not spotted or striped) target exemplar for each novel kind about which the speakers only disagreed using our subjective GAs. ${ }^{1}$

## Results

Faultless Disagreement We fit a logit model to the faultless disagreement judgment data with GA type as a predictor. Par-

[^69]ticipants were highly likely to permit faultless disagreement for subjective $(\beta=7.377, p<0.001)$ and relative $(\beta=3.061$, $p<0.001)$ GAs. They were unlikely to permit faultless disagreement over absolute ones ( $\beta=-2.501, p<0.001$ ).
Relation to Post-test As in Experiment 1, we see differences among the adjective classes in the relation between participants' own assertion of each GA and their permission of faultless disagreement over it. In the post-test, all participants judged the target pimwit with spots "spotted," "tall," and "pretty." $94 \%$ answered "yes" when asked if the plain target pimwit was "pretty." For the dax with stripes, $97 \%$ judged it "striped." $64 \%$ said it was "big," and only $9 \%$ said it was "boring," while $55 \%$ said that the plain dax was. Despite substantial variation in their own evaluations of the critical items, participants almost always permitted faultless disagreement for the subjective GAs. For the absolute GAs, which the vast majority of participants accepted as true of the critical items, participants permitted very little faultless disagreement, but judged disagreements over the relative GAs as faultless between half and three-quarters of the time (Figure 2).

Even when listeners do not have an explanation for speakers' differing standards for relative GAs, they may permit faultless disagreement due to the standard's uncertainty. Participants permitted more faultless disagreement over "big" ( $72 \%$ of participants), which a lower proportion (64\%) agreed was true of the critical dax, and less over "tall" ( $56 \%$ ) which all participants agreed was true of the pimwit.
Qualitative Explanations Fitting logit models to the data for each explanation code and GA type, participants again were likely to refer to object properties in explaining judgments over absolute GAs $(\beta=1.541, p<0.001)$, but not relative $(\beta=-4.616, p<0.001)$ or subjective $(\beta=-4.616$, $p<0.001$ ) ones. Participants were most likely to refer to speakers' experiences-even though they were identical-in their explanations for relative ( $\beta=3.960, p<0.001$ ) and subjective ( $\beta=2.060, p<0.01$ ) GAs, and least likely for absolute ones ( $\beta=-3.466, p<0.001$ ). Finally, participants were unlikely to refer to speakers' opinions in explaining disagreements over absolute GAs ( $\beta=-2.501, p<0.001$ ), but were likely to do so in explaining disagreements over relative GAs ( $\beta=1.096, p<0.05$ ), and highly likely for subjective ( $\beta=3.521, p<0.001$ ) ones as well (Figure 3).

Compared to Experiment 1, adults permitted less faultless disagreement for relative and absolute GAs when speakers had experienced identical distributions of exemplars (Figure 2). This was not the case for subjective GAs, which participants continued to permit faultless disagreement over.

## Experiment 3

Experiment 3 followed up on the previous experiments with adults by exploring the developmental trajectory of linguistic subjectivity. We tested a large age range to span a broad swath of theory-of-mind and metalinguistic development.

Participants Seventy-one children across three age groups participated (24 4-5.5 years: 15 girls, $M=4.83, S D=0.34$; $235.5-7$ years: 8 girls, $M=6.05, S D=0.470 ; 248-9.5$ years: 14 girls, $M=8.90, S D=0.34$ ). Four children were excluded due to experimenter error or broken stimuli $(n=2)$.

## Stimuli and Methods

Experiment 3 used the same method as Experiment 2, except that speakers experienced distinct exemplar distributions as they did in Experiment 1. We also only included one object kind, pimwits, to keep it a more manageable length for children.

## Results



Figure 4: Child rates of faultess disagreement judgment in Exp. 3. Train:AbS and Train:FD trials were training intended to elicit non-faultless and faultless judgments, respectively. Participants made two judgments over the subjective GA "pretty," regarding a spotted as well as a plain pimwit, but there was no significant difference between rates of faultless disagreement between the two.

Faultless Disagreement We fit a logit model to the critical trial faultless disagreement data with GA and age. Children were significantly less likely to permit faultless disagreement for absolute GA "spotted" ( $\beta=-7.646, p<0.001$ ), and significantly more likely for relative and subjective GAs "tall" and "pretty" ("tall": $\beta=2.047, p<0.001$; "pretty" for spotted pimwit: $\beta=1.170, p<0.05$; "pretty" for plain pimwit: $\beta=1.831, p<0.001$ ). In general, they permitted faultless disagreement more with age ( $\beta=0.764, p<0.001$ ).

We can think of children's initial judgment rates on the faultless disagreement training trials as baselines (Figure 4). Even in our oldest age group, rates of faultless disagreement on the critical trials are significantly below those of the faultless training trials (for relative trials: $t=-2.164$, $d f=31.373, p<0.05$; for subjective trials: $t=-3.820$, $d f=70.616, p<0.001$ ). While capable of judging disagreements as faultless, children were reluctant to do so when they themselves agreed with only one of the speakers.

Relation to Post-test $96 \%$ of children judged the target pimwit "spotted." More children judged it "pretty" (85\%) than "tall" ( $49 \%$ ). $55 \%$ answered that the plain pimwit was "pretty." For the absolute and relative GAs, we see roughly
the same qualitative relation between post-test response and faultless disagreement judgments as with adults: greater posttest consensus meant less faultless disagreement.
Qualitative Explanations There appear to be some children across our age range who understood the source of knowledge most relevant for each GA, though children referred to object properties most frequently for all types until our oldest age group (Figure 5). We fit logit models to the data for each explanation code separately, with GA type, age, and their interaction as predictors. Children were highly likely to refer to properties of the objects in explaining absolute $(\beta=2.543, p<0.05)$ and relative $(\beta=3.758$, $p<0.05)$ GA utterances, and less likely to do so for relative $(\beta=-0.800, p<0.01)$ and subjective ones with age ( $\beta=-0.642, p<0.01$ ). They were least likely to refer to speakers' experiences in explaining absolute utterances ( $\beta=-6.608, p<0.01$ ), though more likely to refer to them at all with age $(\beta=0.586, p<0.05)$. Finally, they became more likely to refer to speakers' opinions to explain subjective utterances as they got older ( $\beta=0.872, p<0.05$ ).


Figure 5: Proportion of children's explanations in Exp. 3 receiving each qualitative code, by GA type and age group (in panels).

## Experiment 4

Experiment 4 used the same method as above, with children at the older end of the age range. As in Experiment 2, the distributions that the two speakers saw were identical.

Participants Participants were 24 children 8-9.5 years of age (12 girls; $M=9.09, S D=0.44$ ).

## Results

Faultless Disagreement Children permitted faultless disagreement on $98 \%$ of faultless training trials, and on none of non-faultless training trials. They did so most on subjective trials ( $67 \%$ of the time, $95 \%$ CI:53-80), followed by relative $(38 \%, C I: 21-54)$ and absolute $(1 \%, C I: 0-4)$.

Children were least likely to permit faultless disagreement over absolute GA "spotted" ( $\beta=-3.135, p<0.01$ ), and highly likely for all other adjectives ("tall": $\beta=2.625$, $p<0.05$, "pretty" for spotted pimwit: $\beta=4.052, p<0.001$, "pretty" for plain pimwit: $\beta=3.829, p<0.001$ ).
Relation to Post-Test While all children judged the target pimwit "spotted" and "tall," they resembled adults by still permitting faultless disagreement over "tall" about half the time (and almost never over "spotted"). $79 \%$ and $92 \%$ participants judged the spotted and plain pimwits "pretty," respectively. Despite general consensus over their beauty, children responded more like adults in nonetheless permitting faultless disagreement over them at relatively high, equivalent rates.
Qualitative Explanations Children's explanations for absolute GA "spotted" were highly likely to receive the Овject Property code ( $\beta=2.398, p<0.01$ ), while explanations of relative ( $\beta=-2.734, p<0.01$ ), and subjective ( $\beta=-3.267, p<0.001$ ) utterances were unlikely to. Explanations of absolute GA utterances were also unlikely to be coded as referring to Speaker Experience ( $\beta=-3.135$, $p<0.01$ ), which was highly likely for relative GA utterances ( $\beta=2.625, p<0.05$ ). Lastly, explanations of subjective GAs were likely to be coded as citing Speaker Opinion ( $\beta=4.005, p<0.001$ ), in contrast to explanations about absolute GA utterances ( $\beta=-3.135, p<0.01$ ).

## General Discussion

We tested theoretical claims about faultless disagreement arising when there is uncertainty about how and whether to assess something as, e.g., "pretty" or "tall" (Barker, 2013). We asked in particular whether individuals consider the reference distribution of their interlocutors in interpreting relative gradable adjectives. Adults reliably permitted faultless disagreement over relative and subjective GAs when two speakers had had distinct personal experiences. Rates of faultless disagreement decreased for relative GAs when speakers had experienced identical distributions, but did not disappear altogether, suggesting that adults were instead permitting faultless disagreement out of an understanding of the standard's uncertainty. Together, these findings provide evidence for the consideration of speaker at the level of semantics, as well as adults' sensitivity to the potential for differing standards of more than just explicitly context-dependent adjectives.

The development of sensitivity to linguistic subjectivity appears to be exceptionally prolonged: for the most part, children 'sided' with the speaker who voiced their own evaluations. Two factors might explain the apparent gap between adults and children in our studies. First, previous work sug-
gests that children better grasp subjectivity when they are able to reason about an individual's goals (Holubar, 2015), a dimension that was absent from our experiments. To this end, our ongoing studies explore the effect of goal-oriented contexts (e.g., choosing who you would want to be friends with or learn from), which might be more sensitive to children's nascent understanding of the different implications of being "wrong" about whether something is "spotted" as opposed to "pretty." Second, there may be more continuity between adults and children than it appears. Adults' permission of faultless disagreement and explanation of different GA utterances may reflect social pressures and metalinguistic knowledge, rather than a core belief that their own evaluation is subjective. When it comes to predicates of personal taste, although adults may readily say that the meaning of "good" is subjective, such that both speakers can be right, we have all had the experience of disagreeing about whether a movie or song is "good." Future studies will test for possible continuity between children and adults by examining the contexts in which children may behave more like adults in their metalinguistic judgments of subjectivity, and the contexts in which adults may react similarly to children in their implicit commitment to intersubjectivity.

## References

Barker, C. (2013). Negotiating taste. Inquiry: An Interdisciplinary Journal of Philosophy, 56(2-3), 240-257.
Barner, D., \& Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4 -year-olds interpret tall and short based on the size distributions of novel noun referents. Child Development, 79, 594-608.
Diesendruck, G. (2005). The principles of conventionality and contrast in word learning: An empirical examination. Developmental Psychology, 41(3), 451-463.
Holubar, T. F. (2015). Children's reasoning about unconventional opinions as evidence for naive realism. Unpublished doctoral dissertation, Stanford University.
Repacholi, B. M., \& Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. Developmental Psychology, 33(1), 12-21.
Ross, L., \& Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E. S. Reed, \& E. Turiel (Eds.), Values and knowledge (pp. 103-135). Hillsdale, NJ: Erlbaum.
Scontras, G., Degen, J., \& Goodman, N. D. (2016). Subjectivity predicts adjective ordering preferences. Open Mind.
Smith, L. B., Cooney, N. J., \& McCord, C. (1986). What is "high"? The development of reference points for "high" and "low". Child Development, 57, 583-602.
Syrett, K., Kennedy, C., \& Lidz, J. (2009). Meaning and context in children's understanding of gradable adjectives. Journal of Semantics, 27(1), 1-35.
Taylor, M., Cartwright, B., \& Bowden, T. (1991). Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? Child Development, 62(6), 1334-1351.

# Word Embedding Distance Does not Predict Word Reading Time 

Stefan L. Frank (s.frank@let.ru.nl)<br>Centre for Language Studies, Radboud University<br>P.O. Box 9103, 6500 HD Nijmegen, The Netherlands


#### Abstract

It has been claimed that larger semantic distance between the words of a sentence, as quantified by a distributional semantics model, increases both N400 size and word-reading time. The current study shows that the reading-time effect disappears when word surprisal is factored out, suggesting that the earlier findings were caused by a confound between semantic distance and surprisal. This absence of a behavioural effect of semantic distance (in the presence of a strong neurophysiological effect) may be due to methodological differences between eye-tracking and EEG experiments, but it can also be interpreted as evidence that eye movements are optimized for reading efficiency.


Keywords: reading; eye tracking; N400; distributional semantics; semantic distance; word surprisal

## Introduction

An open question in the study of human language processing is to what extent mere semantic similarity among words within a sentence or text affects the comprehension process. Results from controlled experiments are inconclusive. On the one hand, there is ample evidence for effects on the N400 event-related brain potential (ERP) component: Reading a word that is semantically related to words in the preceding context decreases N400 size, relative to when the context words are not meaning related (Camblin, Gordon, \& Swaab, 2007; Metusalem et al., 2012; Paczynski \& Kuperberg, 2012). A number of behavioural experiments, however, failed to find corresponding effects on word-reading time (Gordon, Hendrick, Johnson, \& Lee, 2006; Traxler, Foss, Seely, Kaup, \& Morris, 2000). In contrast, two studies that analysed reading times on naturalistic texts (instead of taking a controlled experimental approach) did find that words are read faster when they have stronger semantic relatedness to earlier words in the text (Mitchell, Lapata, Demberg, \& Keller, 2010; Pynte, New, \& Kennedy, 2008). In those studies, semantic relatedness measures were obtained from a distributional semantics model, which assigns numerical vectors to words on the basis of the words' co-occurrence patterns in large text corpora. These vector representations are known as word embeddings in the computational linguistics literature. Words that tends to occur in similar contexts receive similar embeddings. Consequently, distances between the words' embedding vectors correspond to semantic distances between the corresponding words.

If semantically related words tend to co-occur, a word's occurrence can (to some extent) be predicted from the presence of related words. Consequently, if one wants to claim that the reading process on word $w_{t}$ is affected by the word's semantic relatedness to the preceding words $\left(w_{1}, \ldots, w_{t-1}\right)$, it is crucial to factor out any effect of the predictability of
$w_{t}$ from its previous context. Otherwise, apparent effects of relatedness could in fact be due to word predictability instead. Frank and Willems (in press) recently showed that N400 effects of semantic distance (as quantified by a distributional semantics model) remain when factoring out the words' (un)predictability as quantified by their surprisal (i.e., $-\log P\left(w_{t} \mid w_{1}, \ldots, w_{t-1}\right)$ ), leaving no room for a confound between predictability and semantic distance. The current paper will show that the same is not true for reading times: Effects of semantic similarity on reading times for naturalistic materials, of the type reported by Mitchell et al. (2010) and Pynte et al. (2008), disappear when surprisal is factored out, provided that surprisal is computed by a powerful enough language model. Hence, semantic similarity between the words of a sentence or text affects N400 size but not reading time.

## Method

## Eye-tracking Data

Word-reading times were extracted from two published sets of eye-tracking data: The UCL corpus (Frank, Monsalve, Thompson, \& Vigliocco, 2013) and the English Dundee corpus (Kennedy \& Pynte, 2005). The UCL corpus comprises data from 42 native English speakers reading 205 individual sentences sampled from three unpublished novels; the Dundee corpus has 10 participants reading newspaper editorials. Frank and Willems (in press) demonstrated strong N400 effects of semantic distance (over and above the effect of surprisal) for the sentences of the UCL corpus. Mitchell et al. (2010) reported reading-time effects of semantic distance in the Dundee data, and similar results by Pynte et al. (2008) were based on the French part of the Dundee corpus, also comprising newspaper texts.

Four measures of reading time will be investigated: firstfixation duration, first-pass duration (the sum of fixation durations on a word before the first fixation on any other word), right-bounded reading time (the sum of fixation durations on a word before the first fixation on a later word), and go-past reading time (the sum of fixations on all words from the first fixation on the current word until the first fixation on a later word). These four measures, in this order, have been argued to reflect increasingly late cognitive processes (Clifton Jr., Staub, \& Rayner, 2007; Gordon et al., 2006).

## Models

Each content word of the UCL and Dundee corpora was assigned a measure of semantic distance to preceding content words, as well as five estimates of word surprisal. The distributional semantics and surprisal models were trained on
the first slice of the ENCOW14 web corpus (Schäfer, 2015), comprising 644.5 M word tokens of 2.81 M types.

Semantic Distance Word embeddings were generated by the word2vec skipgram model (Mikolov, Chen, Corrado, \& Dean, 2013), which is basically a feedforward neural network with one hidden layer. The network learns to associate each input word $w_{t}$ to the $k$ words immediately preceding and following (i.e., the sequence $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k}$ ). After training the network, the vector of connection weights from each input unit to the 300 -unit hidden layer forms the embedding for the word corresponding to the input unit. The 'window size' parameter was set to $k=5$ in the current application of the model.

As explained in the Introduction, the distance between two word vectors quantifies the semantic distance between the two words. A common distance measure used in distributional semantics is the cosine of the angle between the vectors. Here, we require a measure for the distance between the current word's embedding $\vec{w}_{t}$ and its entire previous context (not just a single word). The vector representing the combination of content words from the previous context is defined as simply the sum of the words' individual vectors. Thus, the relevant distance measure becomes

$$
\begin{equation*}
\operatorname{semdist}(t)=-\cos \left(\vec{w}_{t}, \sum_{w \in A_{t}} \vec{w}\right) \tag{1}
\end{equation*}
$$

where $A_{t}$ is a collection of content words that precede $w_{t}$ in the sentence or text. For the individual sentences of the UCL corpus, $A_{t}$ contains all content words preceding $w_{t}$ in the sentence. For the full texts of the Dundee corpus, $A_{t}$ contains the four content words immediately preceding $w_{t}$ in the text (if $w_{t}$ is among the text's first four content words, $A_{t}$ will contain correspondingly fewer words). If $A_{t}$ is empty, word $w_{t}$ has no semantic distance. Semantic distance values on the UCL corpus were identical to those used by Frank and Willems (in press) to analyse N400 ERP effects.

Surprisal Word surprisal was computed by $n$-gram language models, which simplify the full conditional probability $P\left(w_{t} \mid w_{1}, \ldots, w_{t-1}\right)$ to $P\left(w_{t} \mid w_{t-n+1}, \ldots, w_{t-1}\right)$, that is, only the $n-1$ previous words are taken into account when estimating the occurrence probability of $w_{t}$. Model order $n$ was varied from $n=2$ to $n=5$, and the model was generated by SRILM (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen \& Goodman, 1999).

The semantic distance measure defined above is sensitive to content words beyond the $n-1$ previous words that matter to an $n$-gram model. If semantic distance correlates with surprisal, this could yield apparent effects of semantic distance that are in fact due to unpredictability resulting from words outside of the $n$-gram window. To control for this, a 'skipbigram' language model (SBLM) was used to obtain a fifth set of surprisal values:

$$
P_{\mathrm{sblm}}\left(w_{t} \mid A_{t}\right)=\frac{1}{\left|A_{t}\right|} \sum_{w_{i} \in A_{t}} P\left(w_{t} \mid w_{i}\right)=\frac{1}{\left|A_{t}\right|} \sum_{w_{i} \in A_{t}} \frac{P\left(w_{i}, w_{t}\right)}{P\left(w_{i}\right)}
$$

with $A_{t}$ as defined as in Equation 1 and $\left|A_{t}\right|$ the number of words in $A_{t} . P\left(w_{t} \mid w_{i}\right)$ denotes the probability that $w_{t}$ occurs within a distance of 15 words after occurrence of $w_{i}$. That is, the preceding content words $w_{i} \in A_{t}$ are taken as independent cues to the occurrence of $w_{t}$, whose skip-bigram probability is computed by averaging over these individual cues.

The required word-pair probabilities $P\left(w_{i}, w_{t}\right)$ are estimated from co-occurrence frequencies in the training corpus, using the Simple Good-Turing smoothing method (Gale \& Sampson, 1995) to estimate the total probability of all unseen pairs. This total probability $P_{0}$ is divided over the unseen pairs $(v, w)$ in proportion to $P(v) P(w)$, that is, the probability of each particular unseen pair $(v, w)$ is given by:

$$
P(v, w)=\frac{P_{0} P(v) P(w)}{1-\sum_{\left(v^{\prime}, w^{\prime}\right) \in S} P\left(v^{\prime}\right) P\left(w^{\prime}\right)},
$$

where $S$ is the set of all ordered word pairs observed in the training data within a 15 -word distance from each other.

Relation between semantic distance and surprisal Table 1 shows there indeed exists a positive confound between surprisal and semantic distance, which grows stronger as the language model is able to use words from further back in the context.

Frank and Willems (in press) interpolate the 5-gram and skip-bigram models to minimize average surprisal over the UCL corpus and show empirically that the semantic distances do not contain information that can be used to further improve this interpolated language model. Hence, if the semantic distances account for variance in human reading difficulty measures over and above what is already explained by the surprisal values, this cannot be attributed to a confound between semantic relatedness and predictability but must be due to the effect of semantic relatedness itself.

## Data Analysis

Linear mixed-effects regression models were fitted to the logtransformed reading times using as covariates: word position in the sentence, word length (number of characters), word log-frequency in ENCOW14, and a binary factor indicating whether or not the previous word was fixated. To account for the possibility that reading-time effects appear shortly after the point at which they originate (so-called spillover effects), the previous word's length and log-frequency were also included. All two-way interactions between these six factors were also present.

Table 1: Correlation coefficients between semantic distance and surprisal values.

|  | Language model |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Data set | 2-gram | 3-gram | 4-gram | 5-gram | SBLM |
| UCL | .19 | .26 | .27 | .27 | .29 |
| Dundee | .05 | .18 | .20 | .21 | .26 |

The main factor of interest was the word's semantic distance measure. Separate analyses were run using the current and previous word's semantic distance; the latter capturing potential spillover. ${ }^{1}$ In addition to a control condition without any surprisal measure in the regression, five separate analyses were run including $n$-gram surprisal with $n=2,3,4,5$, or both 5-gram and SBLM surprisal (always for both the current and previous word).

Random effects in the regression model were the bysubject and by-word intercept, and by-subject slopes of semantic distance and any surprisal measure that was included as a fixed effect.

Regression models were fitted to each of the four reading time measures from both data sets, making a total of 96 analyses: 4 reading time measures $\times 2$ corpora $\times 2$ semantic distance measures (of current or previous word) $\times$ ( 5 surprisal measures +1 control). Words were excluded from analysis if they were not fixated, were attached to punctuation, contained any non-letter or more than one capital letter, or were the first or last word on a line.

## Results

Figure 1 displays the estimated regression coefficient (i.e., effect size) of the semantic distance predictor in each of the 96 fitted regression models. Note that effect sizes cannot be compared between the analyses investigating the current versus previous word's semantic distance. This is because these analyses apply to different sets of words: All content words when the current word's semantic distance is used, but the words directly following content words (including many function words) when the previous word's semantic distance is the variable under investigation. The same holds for the estimated regression coefficients of the surprisal predictors, plotted in Figure 2. ${ }^{2}$

For the UCL corpus, none of the semantic distance effects reach statistical significance. For the Dundee corpus, there is a clear effect of semantic distance in the expected (i.e., positive) direction when surprisal is not factored out, and it remains present for later reading time measures when surprisal takes only very local context into account (i.e., under a bigram model).

As is clear from Figure 2, words with higher surprisal take longer to read, as is well known from the literature (e.g. Monsalve, Frank, \& Vigliocco, 2012; Smith \& Levy, 2013). Surprisal computed by the novel SBLM language model has an effect over and above 5-gram surprisal, at least for the Dundee corpus, which means that it is not merely the local,

[^70]4-word context that is is taken into account when generating expectations about upcoming words. Rather, long-distance co-occurrence patterns between content words matter as well.

There are a few noticeable difference between the results for the UCL and Dundee data sets, which mirror differences in the text materials of these two corpora. Surprisal effects appear to be more reliable in the Dundee data, in that the zero point falls further outside the confidence intervals. This can simply be explained by the Dundee data set being much larger than the UCL data set ( 134,203 versus 18,178 data points). Interestingly, the UCL corpus results show larger effect sizes (i.e., larger coefficients) which is probably due to these materials having been specifically designed for language model evaluation. Compared to the Dundee corpus texts, the UCL corpus sentences contain fewer low-frequency words (for which surprisal is hard to estimate reliably) and can comprehended more easily without relying on world knowledge (which the language models do not incorporate). Finally, the fact that the SBLM model explained unique variance in reading times from the Dundee corpus only can be explained by the fact that this corpus consists of full texts as opposed to the UCL corpus's individual sentences. Compared to individual sentences, full texts will contain more content words outside of the 5-gram window, making the SBLM model more influential.

## Discussion

Results on the Dundee corpus showed significant, positive effects of semantic distance on all four reading time measures when surprisal was not taken into account. However, factoring out surprisal as computed by anything more powerful than a bigram model made the effects of semantic distance disappear. Apparently, these effects were due to a confound between semantic distance and surprisal, that is, a word is less likely to appear if it has weaker semantic relatedness to earlier words. It was actually a word's unpredictability, rather than its semantic content per se, that resulted in increased reading time.

Indeed, the findings by Pynte et al. (2008) and Mitchell et al. (2010), on the French and English Dundee corpus, respectively, can be attributed to confounds between semantic relatedness and predictability. Pynte et al. (2008) did not factor out surprisal (or even simple transitional probabilities between words) in their analysis of the effect of semantic distance. Mitchell and Lapata's (2009) goal was to show that incorporating semantic distance measures from their own 'simple semantic space model' (as well as from a Latent Dirichlet Allocation Topics model; Griffiths, Steyvers, \& Tenenbaum, 2007) reduces perplexity of a combined $n$-gram and probabilistic phrase-structure grammar. That is, taking these semantic measures into account improves the language model. Consequently, the improved fit to reading time could be due merely to more accurate next-word prediction rather than to semantic similarity per se.

The UCL corpus results showed no effect of semantic dis-


Figure 1: Regression coefficients (with $95 \%$ confidence intervals) of semantic distance predictor, when factoring out different measures of surprisal. The leftmost two panels display results on the UCL corpus; the Dundee corpus results are shown in the rightmost panels. The 2nd and 4th panel show the coefficient of the previous word's semantic distance. Reading time measures are indicated by FF (first fixation), FP (first pass), RB (right-bounded), and GP (go-past).


Figure 2: Regression coefficients (with $95 \%$ confidence intervals) of surprisal predictor, when factoring out semantic distance. "5-gram*" refers to the effect of 5-gram surprisal when SBLM-surprisal is also included as a regressor, and "sblm" refers to the effect of SBLM-surprisal over and above 5-gram surprisal. The leftmost two panels display results on the UCL corpus; the Dundee corpus results are shown in the rightmost panels. The 2nd and 4th panel show the coefficient of the previous word's semantic distance. Reading time measures are indicated by FF (first fixation), FP (first pass), RB (right-bounded), and GP (go-past).
tance on reading times whatsoever, even when surprisal was not taken into account. This is remarkable considering that Frank and Willems (in press) found that N400 effects of the very same semantic distance values are of similar size as and independent from - the effect of surprisal as computed by an interpolated 5-gram and skip-bigram language model. This discrepancy between neurophysiological and behavioral effects is consistent with findings from the controlled experimental studies mentioned in the Introduction. But how can it be explained?

One possible cause is the difference in stimuli presentation method. The eye-tracking methodology allows a natural reading processes whereas in most EEG reading studies, words are presented one at a time for an unnaturally long duration. The EEG data used by Frank and Willems (in press) came from a study with word-length dependent presentations durations of at least 627 ms (Frank, Otten, Galli, \& Vigliocco, 2015), which is much longer than fixation durations in natural reading. Wlotko and Federmeier (2015) showed that using more natural word presentation rates in an ERP reading study can remove particular effects of semantic relatedness on the N400. If semantic distance effects are delayed relative to surprisal effects, this could explain their absence in reading times: By the time they would have appeared, any effect has already been washed out by the processing of several other words. Although Figure 1 indeed shows a trend for the semantic distance effect to be somewhat stronger for the later reading time measures (as was also found by Pynte et al., 2008), the same is true for the surprisal effect (Figure 2) so this cannot explain why reading times are insensitive to semantic distance. Moreover, Frank and Willems (in press) found fMRI effects of semantic distance (as quantified by distributional semantics) during normal speech comprehension, indicating that the presence of a measurable neural response does not rely on unnaturally slow presentation rates.

An alternative, and possibly more interesting explanation of the difference between N 400 and reading time effects is that reading is optimized for speed (Smith \& Levy, 2013). Being faster on more predictable (i.e., lower surprisal) words increases overall efficiency, whereas there is no reason to be faster on merely semantically related words. Hence, we would expect reading times to display effects of surprisal but not of semantic distance. Other dependent variables from eye-tracking, however, could show sensitivity to semantic distance, and this is exactly what Van den Hoven, Hartung, Burke, and Willems (2016) found in a recent analysis of data from a Dutch narrative text reading eye-tracking study: Semantic distance correlated with saccade distance and regression probability but not with reading time after factoring out trigram surprisal. In contrast, the reason why the N400 shows effects of both surprisal and semantic distance could be that it forms an index of the difficulty of retrieving lexical information from long-term memory (Brouwer, Fitz, \& Hoeks, 2012; Kutas \& Federmeier, 2000). As Frank and Willems (in press) argue, this difficulty is reduced both by probabilis-
tic word prediction (surprisal) and by semantic similarity to earlier words (word embedding distance).

## Conclusion

The current results failed to replicate earlier findings of a positive correlation between reading times on naturalistic data and semantic relatedness between words, as quantified by a distributional semantics model. This apparent effect of semantic relatedness appeared to be due to a confound with word predictability. Of course, it is possible that an effect of semantic distance reappears when using a different distributional semantics model, or a more sophisticated technique for combining single word vectors into a sentence context vector (Equation 1). However, it is equally true that improved surprisal models may undo the work of more sophisticated word embedding models. And crucially, the current distributional semantics modelling choices were appropriate for predicting reading times when surprisal was not taken into account, as well as N400 sizes over and above surprisal, so they should also have sufficed for revealing reading time effects of semantic distance that are independent from surprisal, if there had been any.

## Acknowledgements

The work presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

## References

Brouwer, H., Fitz, H., \& Hoeks, J. C. J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. Brain Research, 1446, 127-143.
Camblin, C. C., Gordon, P. C., \& Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. Journal of Memory and Language, 56, 103-128.
Chen, S. F., \& Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 13, 359-394.
Clifton Jr., C., Staub, A., \& Rayner, K. (2007). Eye movements in reading words and sentences. In R. Van Gompel, M. Fisher, W. Murray, \& R. L. Hill (Eds.), Eye movement research: A window on mind and brain (pp. 341-372). Elsevier, Oxford, UK.
Frank, S. L., Monsalve, I., Thompson, R. L., \& Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. Behavior Research Methods, 45, 1182-1190.
Frank, S. L., Otten, L. J., Galli, G., \& Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. Brain and Language, 140, 1-11.
Frank, S. L., \& Willems, R. M. (in press). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. Language, Cognition and Neuroscience.

Gale, W. A., \& Sampson, G. (1995). Good-Turing frequency estimation without tears. Journal of Quantitative Linguistics, 2, 217-237.
Gordon, P. C., Hendrick, R., Johnson, M., \& Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. Journal of Experimental Psychology. Learning, Memory, and Cognition, 32, 1304-1321.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114, 211-244.
Kennedy, A., \& Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. Vision Research, 45, 153-168.
Kutas, M., \& Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences, 4, 463-470.
Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., \& Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. Journal of Memory and Language, 66, 545-567.
Mikolov, T., Chen, K., Corrado, G., \& Dean, J. (2013). Efficient estimation of word representations in vector space. In Proceedings of the ICLR Workshop.
Mitchell, J., \& Lapata, M. (2009). Language models based on semantic composition. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (pp. 430-439). Singapore: Association for Computational Linguistics.
Mitchell, J., Lapata, M., Demberg, V., \& Keller, F. (2010, July). Syntactic and semantic factors in processing difficulty: An integrated measure. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 196-206). Uppsala, Sweden: Association for Computational Linguistics.
Monsalve, I. F., Frank, S. L., \& Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 398-408). Avignon, France: Association for Computational Linguistics.
Paczynski, M., \& Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. Journal of Memory and Language, 67, 426-448.
Pynte, J., New, B., \& Kennedy, A. (2008). On-line contextual influences during reading normal text: A multipleregression analysis. Vision Research, 48, 2172-2183.
Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, \& A. Witt (Eds.), Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora (pp. 28-34). Mannheim, Germany: Institut für Deutsche Sprache.
Smith, N. J., \& Levy, R. (2013). The effect of word pre-
dictability on reading time is logarithmic. Cognition, 128, 302-319.
Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing (pp. 901-904). Denver, Colorado.
Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., \& Morris, R. K. (2000). Priming in sentence processing: intralexical spreading activation, schemas, and situation models. Journal of Psycholinguistic Research, 29, 581-595.
Van den Hoven, E., Hartung, F., Burke, M., \& Willems, R. M. (2016). Individual differences in sensitivity to style during literary reading: Insights from eye-tracking. Collabra, 2, 25.

Wlotko, E. W., \& Federmeier, K. D. (2015). Time for prediction? the effect of presentation rate on predictive sentence comprehension during word-by-word reading. Cortex, 68, 20-32.

# Is Conflict Detection in Reasoning Domain General? 

Darren Frey (darren@post.harvard.edu)<br>Paris Descartes University, Sorbonne Paris Cité CNRS UMR 8240, LaPsyDÉ, Paris, France

Wim De Neys (wim.de-neys@parisdescartes.fr)<br>Paris Descartes University, Sorbonne Paris Cité<br>CNRS UMR 8240, LaPsyDÉ, Paris, France


#### Abstract

A great deal of reasoning research indicates that individuals are often biased by intuitive heuristics. However, contemporary results indicate that individuals seem sensitive to their biases; they seem to detect conflict with reasoning norms. One of the key remaining questions is whether this conflict sensitivity is domain general. To address this question, we administered a battery of five classical reasoning tasks to a large sample of subjects and assessed their conflict detection efficiency on each task by measuring their response confidence. Results indicate that conflict detection is, in most senses, not domain general, though there are compelling exceptions.


Keywords: conflict detection; reasoning; bias; domain generality; decision making

## Introduction

That human reasoning is prone to error comes as no surprise to most everyone. Upon reflection, we discern blunders in our own reasoning in the most ordinary of circumstances: we realize we miscalculated how long a trip would take us; we come to acknowledge our snap judgment of a colleague was mistaken. Likewise, we often witness mistakes in others and throughout history, some of which aggregate in the most atrocious of ways: an innocent man is convicted, judged, tried, and sentenced to more than forty years of solitary confinement based on the shakiest possible evidence, the testimony of a single untrustworthy witness (e.g., the case of Albert Woodford, see Aviv, 2017).

Although mistakes like this seem, at first, entirely unrelated, one could argue they often issue from a uniform set of underlying tendencies. Exploring these tendencies has motivated much of the research in reasoning and decision making throughout the past four decades. One compelling and especially generative account of reasoning mistakes contrasts two types of thinking: fast, associative, heuristic thinking and slower, more demanding, rule-based reasoning (Kahneman, 2011). Returning to the example of misjudging a colleague, the fast type of reasoning (System 1) seems to account for our initially misguided impression, which is then revised upon reflection-and after gathering more evidence-by the slower, more deliberate type of thinking (System 2).

The family of dual process theories that rely on contrasting these two forms of reasoning has provided
countless testable hypotheses and a diverse set of approaches and methods. While most parties agree that heuristics are useful, efficient, and often optimal means of navigating complex environments, investigators disagree about how often and in what contexts they conflict with logical and mathematical principles. Evans (2003, 2010), Kahneman (2011), and Stanovich and West (2000) insist that individuals regularly make reasoning mistakes because of unchecked heuristic inferences, while Gigerenzer (2008), Katsikopoulos (2013) and others emphasize that heuristics are generally ecologically rational and truth preserving. ${ }^{1}$

Until recently, one of the cardinal doctrines of the dual process account of reasoning mistakes relied on the imperceptibility of reasoning conflicts. Prominent scholars have argued that our reasoning mistakes masquerade beneath our awareness, which is, at least partially, what accounts for their ubiquity. Surely-the argument goes-if reasoners were aware of their mistakes they would correct them. However, many contemporary empirical analyses of reasoning bias suggest that individuals are often sensitive to conflicts between heuristics and normative principles even when they err (e.g., Bonner \& Newell, 2010; De Neys \& Glumicic, 2008; Pennycook, Fugelsang, \& Koehler, 2015; Handley \& Trippas, 2015).

Researchers have demonstrated this across a number of diverse reasoning tasks using many different methods. Much of the work relies on contrasting tasks that contain conflicts between intuitively cued heuristics and normative principles with structurally identical tasks containing no such conflict. Standard behavioral markers that index conflict on lower level tasks, like response times (RT) and confidence levels (Yeung \& Summerfield), also indicate people are sensitive to conflict in higher level reasoning tasks (Bonner \& Newell, 2010; De Neys \& Glumicic, 2008; Pennycook, Fugelsang, \& Koehler, 2012). Additionally, people tend to fixate visually on the conflicting elements of the tasks, as evidenced in eye and gaze tracking experiments (De Neys \& Glumicic, 2008; Ball, Phillips, Wade, \&

[^71]Quayle, 2006), and they register heightened levels of arousal on skin conductance recordings (De Neys, Moyens, \& Vansteenwegen, 2010). There is neuropsychological evidence of conflict sensitivity as well, derived from both fMRI (De Neys, Vartanian, \& Goel, 2008; Simon et al,. 2015) and EEG analyses (De Neys, Novitskiy, Ramautar, \& Wagemans, 2010). Despite the diversity of methods and tasks supporting these effects, there are many who find the research problematic, largely because its results seem to imply that individuals have fairly immediate access to logical and probabilistic principles (Mata, Schubert, \& Ferreira, 2014; Pennycook et al., 2012; Singmann, Klauer, \& Kellen, 2014; Travers, Rolison, \& Feeney, 2016). Even the most ardent proponents of the work acknowledge that it is still developing and in need of greater clarification (De Neys, 2012, 2014).

Although conflict detection has been explored with a variety of methods and across various tasks, no previous research examines individuals' tendencies to detect conflict across a range of tasks, giving researchers no clear sense of how or whether conflict sensitivities interact. It is unclear, for example, whether a given person's ability to detect a conflict between an intuitively cued heuristic and a reasoning rule on a particular kind of task is related to her ability to do so on different tasks. In essence, a key open question is whether conflict detection is domain general or task specific.

By further clarifying the precise nature of conflict detection, research of this sort will help characterize emergent dual process theories, especially those that explicitly rely on conflict detection mechanisms. For example, Pennycook et al's (2015) three-stage dual-process model relies on differentiating between successful conflict detection and "cognitive decoupling," which is the more resource intensive process of rejecting a conclusion at odds with reasoning rules even when it has been facilitated by a certain heuristic. Crucially, although conflict detection failures are a prominent feature of this model, it is unclear if conflict sensitivity is a stable individual difference. If conflict detection is domain general, then one would expect the prominence of these failures to extend fairly globally. Apart from further specifying the theory, such a conclusion would offer a partial account of the prevalence of cognitive biases. However, if conflict detection is task specific, then one can suppose that empirically observed detection on a given task is largely unrelated to others, and the prevalence of bias needs to be accounted for in other ways.

To address this issue, we presented a battery of the most intensively studied tasks in the field to a large number of reasoners. This enabled us to assess their conflict detection efficiency by measuring their response confidence. Examining the relationship of detection efficiency across the tasks gives us evidence with which to evaluate whether conflict detection is domain general or task specific.

## Method

## Participants

A total of 318 undergraduates (260 female; average age = 22.32, $\mathrm{SD}=6.11$ ) at Paris Descartes University completed the experiment.

## Materials

The experiment consisted of adaptations of five classic reasoning tasks. For each of the five tasks, participants received two conflict items, two no-conflict items, and one abstract control, resulting in 25 items. The tasks were as follows.

Bat and Ball Items (BB) The conflict items in this set were modeled after the canonical CRT problem (Frederick, 2005): "A bat and a ball together cost $\$ 1.10$. The bat costs $\$ 1$ more than the ball. How much does the ball cost?" The answer that often comes to mind is 10 cents, though the correct answer is 5 cents $(\$ 0.05+\$ 1.05=\$ 1.10)$. Participants likely intuitively substitute the "costs $\$ 1$ more than" phrase with "costs \$1," so to generate no-conflict variants one simply removes this phrase (see De Neys, Rossi, \& Houdé, 2013).

Ratio Bias Items (RB) Also called "denominator neglect" problems, these items consist of asking participants to choose between two trays, a small tray and a large one, containing a mixture of gray and white marbles. The participants' goal is to get a gray marble, but the marble will be drawn from the tray they select at random. In a conflict item, the absolute value of gray marbles in the large tray is greater than the absolute value of gray marbles in the small tray, but the relative value of gray marbles is greater in the small tray (e.g., $19 / 100$ vs. $2 / 10$, $19 \%$ vs. $20 \%$ ). Since the marble is being selected at random, one should choose the tray that maximizes the relative likelihood of getting a gray marble (the small tray), but participants are often intuitively and immediately drawn to the larger tray. To generate noconflict items one aligns the relative and absolute values in a tray, so that the tray most likely to have a gray marblethe one with the highest relative value-is also the most perceptually salient one-the one with the highest absolute value (e.g., 21/100 vs. 2/10, see Bonner \& Newell, 2010).

Syllogism Items (SYL) Syllogisms are fundamental arguments in classical logic that consist of two premises and a conclusion, which necessarily follows from the premises when the argument is valid. When the conclusion is at odds with common beliefs, participants tend to deem it logically invalid even when explicitly told just to evaluate the argument's validity (Markovits \& Nantel, 1989). A conflict item consists of a logically valid (or invalid) argument structure with an unbelievable (or believable) conclusion. Here is an example of an unbelievable but valid argument: All mammals can walk. Whales are mammals. $\therefore$ Whales can walk. No-conflict items are those in which common beliefs and the argument's logical structure both cue the
same response. All problems were based on Markovits and Nantel's material (1989).

Base Rate Items (BR) Base rate items consist of statistics describing a sample from which an individual is randomly selected along with a description of the individual. Here is an example of a conflict item: "In a study 1000 people were tested. Among the participants there were 5 sixteen-yearolds and 995 forty-year-olds. Lisa is a randomly chosen participant of the study. Lisa likes to listen to techno and electronic music. She often wears tight sweaters and jeans. She loves to dance and has a small nose piercing. What is most likely? (A.) Lisa is sixteen. (B.) Lisa is forty." This item creates a conflict by calling to mind a stereotype that is at odds with the statistically most likely outcome. To generate no-conflict items, one aligns the statistics and the intuitively cued heuristic. For example to turn the above item into a no-conflict example, switch the base rates so the the sample consists of 995 sixteen-year-olds and 5 forty-year-olds. All problems were based on De Neys and Glumicic's (2008) material.

Conjunction Items (CON) Modeled on the classic Linda problem, participants received descriptions about individuals that either intuitively prompt a single statement (no-conflict) or a conjunctive statement (conflict), and they are asked to decide which statement is most likely. Since a single statement is always more likely than a conjunctive statement, subjects should always choose the single statement regardless of whether it coheres with the stereotype. Here is an example of a conflict item: "Jon is 32. He is intelligent and punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in languages and art. Which one of the following statements is most likely? (A.) Jon plays in a rock band. (B.) Jon plays in a rock band and is an accountant." Since the description generally cues an accountant stereotype, subjects often wrongly choose the less likely option, B. Noconflict items simply isolate the heuristically cued option. All problems were based on De Neys, Cromheeke, and Osman's (2011) material.

## Procedure

The participants were tested in groups of no more than thirty students in a silent classroom at the beginning of a course. In addition to the conflict and no-conflict items illustrated above, participants answered one abstract neutral problem per task. These were designed to query abstract knowledge of relevant reasoning rules and were variants of the above tasks with no clear, consistent intuitive or heuristic prompts. Accuracy on the neutral control items was high (mean accuracy $81.6 \%$, $\mathrm{SD}=0.18$ ). All analyses were run filtering for controls and they made no significant impact on any of the results. Thus, we will present only our unfiltered data in what follows and will not discuss the control items further.

The overall structure of the experiment, a within subject design, was manipulated in three ways: it was balanced for
conflict content, task order, and conflict presentation order. The conflict and no-conflict contents were balanced across participants, such that half the participants received, for example, the conflict conjunction item above, while the other half received its no-conflict analogue, and vice versa. Additionally, the order in which a given task was presented varied, as did whether an individual first saw a conflict or no-conflict item. A partial Latin square of these factors generated 10 different experiment formats, which were distributed evenly across the participant sample.

All items were presented on their own page. At the bottom of which there was a scale where participants indicated how confident they were in their response on a range from $0 \%$ (not at all confident) to $100 \%$ (completely confident).

## Results

## Accuracies

Table 1 (first two rows) presents averages of accuracy levels on each of the tasks, separated by conflict status. Replicating classical findings, performance on no-conflict items was consistently higher than performance on conflict items. In all cases except for RB, contrasts between performance on conflict and no-conflict items was significant (all BB/SYL/BR/CON t > 10.07, $\mathrm{p}<0.001$; RB: $\mathrm{t}(315)=1.10, \mathrm{p}=0.28)$.

Table 1: Accuracies and Conflict Detection Effects

|  | Conflict |
| :---: | :---: | :---: | :---: | | No-Conflict |
| :---: |$c$| Conflict |
| :---: |
| Task |
| Accuracy (SD) |
| Accuracy (SD) |
| Detection (SD) |

## Conflict Detection

To get a sense of how widely conflict detection efficiency is distributed across tasks, it is useful to look at what proportion of the sample tended to detect conflict on the entire battery. At the aggregate level, averaged across tasks, we observe most individuals (74.70\%) tend to a lowered confidence level on conflict vs. no-conflict items. Across all tasks, this difference amounts to a $9.50 \%$ diminution in confidence on incorrectly solved conflict items compared to correctly solved conflict items, $\mathrm{t}(307)=12.01$, $\mathrm{p}<0.001$. This is roughly reflected in the task by task contrasts, though it is highly variable. For example, in the case of the BB items the confidence diminution was $23.15 \%$, while most others hovered around $10 \%$, and, in contrast with previous findings (Stupple, Ball, Evans, Kamal-Smith, 2011), there was little difference between confidence levels on SYL items ( $0.37 \%$ ). In all cases except for SYL, $\mathrm{t}(177)$
$=0.50, \mathrm{p}=0.61$, the task specific confidence decrease was significant: $\mathrm{BB} / \mathrm{RB} / \mathrm{BR} / \mathrm{CON}$ : all $\mathrm{t}>2.65$, all $\mathrm{p}<0.01$.

## Task Specificity and Domain Generality

With a view to evaluating whether conflict detection is task specific or domain general, we ran four primary kinds of analyses: correlations between conflict items across tasks; correlations between conflict detection effects across tasks; analyses of the distribution of conflict detection effects by individual; and regressions to predict conflict detection effects with a composite meant to uncover diffuse evidence of domain generality.

Table 2 summarizes the results of the first two analyses. The statistics above the diagonal are correlations between conflict detection effects across biased individuals on each of the tasks. The correlations below the diagonal are between accuracies on each of the conflict items across tasks for all participants $(\mathrm{N}=318)$. Accuracies on conflict items were significantly correlated between all tasks (all p < 0.04 ), although most correlations are fairly modest, ranging from 0.12 to 0.32 .

Table 2: Conflict Accuracy and Detection Correlations

|  | BB | RB | SYL | BR | CON |
| :---: | :---: | :---: | :---: | :---: | :---: |
| BB |  | $0.06{ }_{56}$ | -0.09 ${ }_{114}$ | $0.07{ }_{142}$ | $0.04{ }_{145}$ |
| RB | 0.24** |  | $0.22_{55}$ | $-0.02_{76}$ | $0.15{ }_{75}$ |
| SYL | 0.26 ** | $0.23 * *$ |  | $0.14_{150}$ | $0.06_{158}$ |
| BR | $0.17{ }^{* *}$ | 0.12* | 0.11* |  | $0.16{ }^{*}{ }^{27}$ |
| CON | 0.12** | 0.13* | 0.13* | 0.32** |  |

As a reminder, a conflict detection effect is, in this context, a diminution of confidence on an incorrectly solved conflict item relative to a correctly solved no-conflict item. Given the notorious noisiness and subjectivity of confidence measures and the difficulty of interpreting conflict detection effect sizes (Frey, Johnson, \& De Neys, 2017), this is measured in a binary way: one either shows the effect or one does not. In stark contrast to the pattern below the diagonal, correlations between conflict detection effects-those above the diagonal-are almost uniformly insignificant. The only exception is CON \& BR, which was correlated at 0.16 ( $p$ 0.02 ). The correlation between BR \& SYL was marginally significant, $\mathrm{r}=0.144, \mathrm{p}=0.078$. Additionally, Bayes Factors for the correlations were all below 0.52, except CON \& BR, which was 1.59.

The binary correlations of conflict detection effects provide no real evidence of domain generality. However, if there is a general and diffuse signal, why should that be captured by simple, pairwise correlations? Perhaps it is the case that conflict detection on a particular task is better predicted by a non-specific and global sensitivity to conflict across tasks. We ran a regression analysis to address this hypothesis, using the combined predictive power of a participant's responses across all tasks. In particular, we used logistic regressions to determine whether conflict
detection on a given task was predicted by one's tendency to detect conflict on all of the other tasks. For example, to see if we can predict whether an individual shows an effect on the BB items, we tallied how often she showed an effect on all the other items (RB, SYL, BR, and CON) and used the latter as our predictor variable.

Table 3: Predicting Conflict Detection Effects

| Model | Beta | Z | P | Pseudo $\mathrm{R}^{2}$ |
| :---: | :---: | :---: | :---: | :---: |
| BB | 1.02 | 5.46 | $<0.01$ | 0.17 |
| RB | 0.37 | 1.34 | 0.18 | 0.02 |
| SYL | 0.07 | 0.40 | 0.69 | $<0.01$ |
| BR | 0.23 | 1.99 | 0.05 | 0.01 |
| BR2 | 0.12 | 0.61 | 0.55 | $<0.01$ |
| CON | 0.33 | 2.08 | 0.04 | 0.01 |
| CON2 | 0.20 | 1.00 | 0.32 | $<0.01$ |

As is clear from Table 3, the goodness of fit of these first models (BB, RB, SYL, BR, CON) was generally quite low. The pseudo $\mathrm{R}^{21} \mathrm{~s}$ (McFadden's) range from $<0.001$ to 0.02 , except for the BB model which was at the limit of what is considered reasonably good (0.17). The relative goodness of this model is reflected in its higher beta coefficient (1.02), which is significant ( $\mathrm{p}<0.001$ ).

The only other models with significantly predictive coefficients were the BR and CON models. However, given the tight correlation between these items discovered in the first analysis, we wondered if this was driving the effect. Indeed, if one runs a restricted model, omitting the correlate of the predictor (leaving out CON in the BR case, and vice versa), the models (BR2 and CON2 in Table 3) have inferior goodness of fit and coefficients that are both smaller and no longer significant. ${ }^{2}$

## Individual Differences

So far, we have uncovered little evidence of domain generality in conflict detection. However, most of the previous findings rely on averaging effects across reasoners. It might well be the case that there are individuals who show evidence of fairly generalized conflict detection. The concern we address in this section is that important differences between individuals might be lost by aggregating as we have, a concern that echoes theorists who emphasize the importance of examining individual differences in reasoning and decision making (Baron, 2010).

[^72]To address this issue, we present a final means of characterizing the sample's overall conflict sensitivity, which is summarized in Figure 1. We scored every biased participant individually in order to get a sense of the distribution of conflict detection effects. For a given individual, the total number of tasks on which she showed a conflict detection effect was divided by the total number of tasks on which she was biased, giving us a range of detection levels spanning from 0 (showing no effect on any of the tasks on which an individual is biased) to 1 (showing an effect on all of the tasks on which an individual is biased).


Figure 1: Frequency of Reasoners by Detection Level
There are concentrations of individuals who consistently detect (Detection Level 1: $18.09 \%$ of the sample), detect half the time (Detection Level 0.50: 21.28\% of sample), and who consistently do not detect (Level 0: $12.77 \%$ of the sample), with all other participants distributed between these three groups. The observation that up to $13 \%$ of the sample shows a Detection Level 0 is in line with previous findings that suggest there are subsets of reasoners who consistently fail to detect conflict (Frey, Johnson, \& De Neys, 2017, Pennycook et al., 2015). The additional observation that $18 \%$ of the sample shows perfect detection across all tasks also implies that there might be exceptions to the overall trend toward task specificity. Although this distribution is compatible with the few studies that have explored individual differences in conflict detection previously, we cannot confirm the representativeness of this kind of a distribution given our methods, as we could have arrived at it by chance.

## Discussion

While performance on conflict items was consistently correlated, we found no clear indication that conflict detection is similarly correlated. Even using a more liberal measure, one that leverages the predictive power of the entire panel of tasks to anticipate conflict detection on a single task, there was only the faintest signal of generality. Nevertheless, base rate (BR) and conjunction (CON) items
were correlated, and the more liberal regression models relying on them were minimally predictive, as was the model predicting the bat and ball (BB) problems. So we found, additionally, no clear evidence of hard and fast task specificity.

One might classify our findings as "domain specific," where a domain is defined as a set of problems that share similar reasoning rules subject to comparable competing intuitive heuristics. From such a perspective, base rate and conjunction items would be considered to fall within the same domain, as they share similar underlying reasoning structures (statistics and probabilities, respectively) that are in conflict with comparable intuitively prompted heuristics (social stereotypes in both cases), and in indeed both were developed to evaluate biases resulting from the representativeness heuristic.
This hybrid outcome has a number of exciting theoretical features and practical applications. For example, Teovanović, Knežević, \& Stankov (2015) argue against a single, explanatory factor underlying cognitive biases that one can easily relate to general intelligence. The account we present here is commensurate with those findings, as it seems indicative of multiple, often dissociable loci of conflict detection failures. Additionally, one of the implications of our findings is that a conflict detection failure on a given task may be largely dissociable from a conflict detection failure on a distant task. This is a hopeful conclusion, especially given the evidence that at the individual level such failures are a non-negligible source of reasoning bias (e.g., Pennycook et al., 2015). The prominence of conflict detection failures on a certain task need not paint a grim picture of reasoning globally. However, the association within what we are calling a domain indicates that at points detecting on a given task will be related to detection on a different task, a relationship that could be exploited educationally. For example, a reasonable pedagogical strategy might begin by allocating resources to the easier of two related tasks, relying on the shared conflict prompting structures to aid in instructional transfer and facilitate instruction on the second task.
These findings raise many additional questions. Since confidence measures are inherently noisy, our results are necessarily tentative. It will be important to revisit the question of the domain generality of conflict detection with additional measures, especially response times. Additionally, given that we were interested in performance across many tasks, we were only able to use a few items per task, so our findings need to be interpreted cautiously. Another particularly promising research project will be to further characterize those individuals who detect conflict in a domain general manner. For example, it would be particularly instructive to determine whether they share similar general cognitive capacities or tend have related thinking dispositions.

## Acknowledgements

This research was supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence National de la Recherche. Additionally, Darren Frey is supported by the Sorbonne Paris Cité International Grant (INSPIRE).

## References

Aviv, R. (2017, January). How Albert Woodfox Survived Solitary. The New Yorker.
Ball, L. J., Phillips, P., Wade, C. N., \& Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. Experimental Psychology, 53(1), 77-86.
Baron, J. (2010). Looking at Individual Subjects in Research on Judgment and Decision Making (or anything). Acta Psychologica Sinica, 42(1), 88-98.
Bonner, C., \& Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. Memory \& Cognition, 38(2), 186-196.
De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. Perspectives on Psychological Science, 7(1), 28-38.
De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. Thinking \& Reasoning, 20(2), 169-187.
De Neys, W., Cromheeke, S., \& Osman, M. (2011). Biased but in Doubt: Conflict and Decision Confidence. PLoS ONE, 6(1), e15954.
De Neys, W., \& Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. Cognition, 106(3), 1248-1299.
De Neys, W., Moyens, E., \& Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. Cognitive, Affective, \& Behavioral Neuroscience, 10(2), 208-216.
De Neys, W., Novitskiy, N., Ramautar, J., \& Wagemans, J. (2010). What makes a good reasoner?: Brain potentials and heuristic bias susceptibility. In Proceedings of the Annual Conference of the Cognitive Science Society (Vol. 32, pp. 1020-1025).
De Neys, W., Rossi, S., \& Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. Psychonomic Bulletin \& Review, 20(2), 269-273.
De Neys, W., Vartanian, O., \& Goel, V. (2008). Smarter than we think: when our brains detect that we are biased. Psychological Science, 19(5), 483-489.
Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. Trends in Cognitive Sciences, 7(10), 454-459.
Evans, J. S. B. T. (2010). Intuition and Reasoning: A DualProcess Perspective. Psychological Inquiry, 21(4), 313326.

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 25-42.

Frey, D., Johnson, E., De Neys, W. (2017, forthcoming) Individual Differences in Conflict Detection During Reasoning. Quarterly Journal of Experimental Psychology
Gigerenzer, G. (2008). Gut Feelings: The Intelligence of the Unconscious (Reprint edition). Penguin Books.
Handley, S. J., \& Trippas, D. (2015). Chapter Two-Dual processes and the interplay between knowledge and structure: a new parallel processing model. Psychology of Learning and Motivation, 62, 33-58.
Kahneman, D. (2011). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux.
Katsikopoulos, K. V. (2014). Bounded rationality: the two cultures. Journal of Economic Methodology, 21(4), 3 61-374.
Markovits, H., \& Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. Memory \& Cognition, 17(1), 11-17.
Mata, A., Schubert, A.-L., \& B. Ferreira, M. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. Cognition, 133(2), 457-463.
Pennycook, G., Fugelsang, J. A., \& Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? Cognition, 124(1), 101-106.
Pennycook, G., Fugelsang, J. A., \& Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive Psychology, 80, 3472
Simon, G., Lubin, A., Houdé, O., \& Neys, W. D. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. Cognitive Neuroscience, 6(4), 158-168.
Singmann, H., Klauer, K. C., \& Kellen, D. (2014). Intuitive logic revisited: new data and a Bayesian mixed model meta-analysis. PloS One, 9(4).
Stanovich, K. E., \& West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? The Behavioral and Brain Sciences, 23(5), 645-665; discussion 665-726.
Stupple, E. J. N., Ball, L. J., Evans, J. S. B. T., \& KamalSmith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. Journal of Cognitive Psychology, 23(8), 931-941. https://doi.org/10.1080/20445911.2011.589381
Teovanović, P., Knežević, G., \& Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. Intelligence, 50, 75-86.
Travers, E., Rolison, J. J., \& Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. Cognition, 150, 109-118.
Yeung, N., \& Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1594), 1310-1321.

# Sequential effects in prediction 

Dinis Gökaydin (d.gokaydin@warwick.edu.au)<br>Warwick Mathematics Institute, University of Warwick, United Kingdom

## Anastasia Ejova (anastasia.ejova@mq.edu.au)

Department of Psychology, Maqcuarie University, Australia


#### Abstract

We studied a simple binary prediction task and discovered that, when making predictions, humans display sequential effects similar to those in reaction time. Moreover, we found that there are considerable individual differences in sequential effects in prediction, again similarly to reaction time studies. We discuss our results in light of the view that sequential effects are the trace of an attempt at detecting a pattern in the sequence, as well as the possible influence of randomness perception in our results. We conclude that the same pattern detection mechanism is likely to underlie sequential effects in reaction time and prediction.


Keywords: sequential effects; prediction
When responding to a sequence of stimuli, human performance depends on the past sequence of stimuli, often to a larger extent than on the properties of the stimuli (e.g. Bertelson, 1961; Cho et al., 2002). This phenomenon, known as 'sequential effects', is commonly interpreted as the product of an attempt at detecting a pattern in the sequence of trials, and in particular whether it is a repeating or alternating sequence. Take for instance a random sequence with two possible elements: after a repeating run, people tend develop an expectation that the next event will be the same; similarly, after an alternating run, an expectation will develop that the next event will alternate relative to the last one. This is reflected in human reaction times (RTs) which tend to be faster for those events which are expected and shorter for those that are not. For instance, let ' $R$ ' and ' $A$ ' stand for repetitions and alternations of stimuli: after seeing RRRR people react faster if the next event was $R$, and slower if it was $A$; conversely, after seeing AAAA they will react faster to another A and slower to an R. If we plot mean RT for all possible histories of events we obtain a 'profile' of sequential effects. Figure 1 shows a commonly obtained profile of sequential effects, often referred to as 'cost-benefit' in order to highlight the trade-off in RT after a given sequence.

But what if, instead of reacting to each element, a prediction must be made about what the next one will be? It follows from the expectation-based account expounded above that those events which are expected should be predicted the most (i.e,. prediction frequency should be negatively proportional to RT). However, the longer a repeating run is, the more people have been found to predict that the sequence will alternate (Jarvik, 1951), an effect known as the 'gambler's fallacy' (Oskarsson, Van Boven, McClelland, \& Hastie, 2009). At first sight, results from prediction experiments where the gambler's fallacy is observed are incompatible with those of RT experiments - where it is found that RT decreases as a
function of run length (Bertelson, 1961) - since, together, the findings from the two paradigms paradoxically imply that humans predict more and more that which they predict less and less. Also, note that, since people react faster when a pattern is confirmed, they are going with the pattern - i.e. behaving as if the pattern will continue - whereas, when predicting, they seem to be going against the perceived pattern.

One possible explanation for the differences observed between prediction and RT is that people might perceive the sequence to be random in some cases and not random in others (Nickerson, 2002). It stands to reason that if a sequence is random, then any regular pattern encontered must be shortlived; if, on the other hand, a sequence is judged to be structured, then a pattern might be more likely to continue. Differences in randomness perception might influence results, but there is evidence that both phenomena - decreasing RT and increasing proportion of prediction with increase in run length - occur simultaneously, as both have been observed in experiments where subjects were made to predict - as well as react to - each stimulus (Hale, 1967; Perruchet, 1985). At first sight, this finding does not seem compatible with the randomness perception account, since that account would suggest that whether or not the sequence is percieved as random changes within the same trial. Thus, while randomness perception cannot be ruled out as a possible explanation for some results, it is not the full story.

Before a decision can be made to go with or against a pattern, the pattern must be detected in the first place. Parsimony suggests that the pattern detection mechanism underlying sequential effects in both prediction and reaction is the same, but that the information it conveys is being used in different ways, and this forms our first hypothesis. As evidence for this hypothesis, we will take any similarities in the profiles of sequential effects in prediction and reaction time. For instance, should the proportion of times people predict the next event to repeat or alternate (i.e., the prediction probability) be found to resemble the cost-benefit pattern in Figure 1, this would be taken as evidence that the same type of sequential effects can be found in RT and prediction. However, we do not know beforehand whether humans are going with or against the pattern in the sequence when making predictions. If going against the pattern we would expect prediction probability to be proportional to RT, and negatively proportional if going with the pattern. In the latter case the profile of sequential effects in prediction would look like an inverted copy of its RT counterpart (see Figure 1, right panel). Previous evidence points to the fact that humans are going against the pattern


Figure 1: Illustration of working hypotheses. Sequences at the bottom are shown in terms of repetitions (R) and alternations (A) of stimuli and should be read from top to bottom with the last event in bold. Left panel shows reaction time data of a single individual included in the analyses performed by Gökaydin et al. (2016), illustrating the ideal profile of sequential effects, also known as 'cost-benefit'. Right panel illustrates what would be expected if prediction also displayed sequential effects similar to those observed in RT. The solid line on the right panel shows what would be expected if humans were going against the pattern in a prediction task, i.e. predicting more often that the pattern will continue; the dashed line shows what would be expected if subjects were going with the pattern.
when predicting, so our second hypothesis is that prediction probability will be proportional to RT, and that the respective profile of sequential effects will show the same 'polarity' as that of RT (Jarvik, 1951; Hale, 1967; Perruchet, 1985).

## The structure of sequential effects

In order to test our hypotheses we must assess whether sequential effects in reaction and prediction are the same. However, this is not a simple matter of comparing results of two sets of subjects performing a reaction and a prediction task, as it is well known from reaction time studies that there is extensive variation in the profile of sequential effects depending on experimental parameters such as the interval between the stimuli (Soetens, Boer, \& Hueting, 1985; Gökaydin et al., 2016), as well as for different individuals performing the same experiment. In fact, the 'typical' profile of sequential effects shown in Figure 2 (left panel) is the exception rather than the rule (Gökaydin et al., 2016). Therefore, in order to demonstrate that sequential effects in prediction are the same as those in reaction time we will try to show they have the same structure.

There is growing evidence that sequential effects in reaction time can be explained in terms of two separate components, one perceptual and related to the sequence of stimuli and the other motor in origin and related to the sequence of responses (Jentzsch \& Sommer, 2002; Maloney, Dal Martello, Sahm, \& Spillmann, 2005; Wilder, Jones, Ahmed, Curran, \& Mozer, 2013; Gökaydin et al., 2016). Crucially, the relative



Figure 2: Perceptual and motor sequential effects. Both panels show data collected by Jentzsch and Sommer (2002). Left panel shows the evidence for a perceptual component of sequential effects (S-LRP) and the right panel for a motor component (LRP-R). See main text for an explanation of the meaning of these two components.
contributions of the two components of sequential effects are known, and take the form of the profiles shown in Figure 2. Moreover, sequential effects in reaction time - across different participant and experimental conditions - are known to be well approximated by a linear combination of the two components, giving us a simplified working model of sequential effects in reaction time. Applying this model to results from a prediction task gives us a way of testing whether sequential effects in reaction time and prediction are similar.

Figure 2 shows the best evidence available about the two components of sequential effects, from an EEG study conducted by by Jentzsch and Sommer (2002). The authors measured the time between stimulus onset and the occurrence of the lateralised readiness potential (LRP), termed S-LRP; and the time between the LRP and the moment a response occurred, or LRP-R. Since the LRP is thought to separate temporally pre-motor from motor processing, S-LRP and LRP-R are considered to give a measure of pre-motor and motor processing respectively. By measuring both S-LRP and LRP-R as a function of the sequence of stimuli, Jentzsch and Sommer (2002) sought to capture the pre-motor and motor contributions towards sequential effects (see Figure 2). Further evidence from other studies shows that what Jentzsch et al. referred to as pre-motor processing can safely be assumed to be perceptual in nature and associated with the processing of stimuli (Maloney et al., 2005; Wilder et al., 2013). For this reason, we will refer to the two components of sequential effects simply as perceptual and motor, denoting them, respectively, as $P$ and $M$.

## Different ways of looking at the data

Sequential effects are usually studied in the context of twoalternative forced-choice tasks ( 2 AFCs ) where one has to react to each stimulus as quickly as possible (e.g., pressing one button if the stimulus appears on the left and another if the stimulus appears on the right). Error rates in this type of task
tend to be quite low, which means that the sequences of stimuli and responses are very similar and that organising results as a function of one or the other yields the same results. By contrast, in a prediction task with two equiprobable stimuli, the error rate is $50 \%$ by design, and this means that the sequence of responses and of stimuli become uncoupled. This raises the question: should the sequence of stimuli or that of predictions be used in order to study the way in which predictions depend on the sequence? And what information is each type of analysis conveying? One possibility is that analysing predictions as a function of the history of predictions - which involve a motor action - we will recover the motor component of sequential effects in prediction; conversely, predictions as a function of the history of stimuli might yield the perceptual component.

## Method

## Participants

21 subjects ( 11 female, 10 male) participated in the experiment. Subjects were recruited using Amazons Mechanical Turk (Mturk) system. Only participants with at least 500 HITS completed on Mturk and with an approval rate of $95 \%$ or above were accepted and paid $\$ 5$ USD for taking part in the experiment.

## Stimuli

Stimuli consisted of two white dots of radius equal to 30 pixels, horizontally separated by a distance in pixels equal to $20 \%$ of the width of the screen. The dots were white and displayed against a grey (RGB 0.5/0.5/0.5) background. During each trial, the two possible positions of the dots were indicated by two black squares equal in width to the diameter of the dots.

## Procedure

Each trial began with a 2000 ms -long text display above the two black squares: 'Is the next dot on the left or on the right?'. Predictions were made with the ' f ' key for 'left' and ' j ' for 'right'. If a prediction was made during the 2000 ms period, the corresponding black square's border would thicken and further key presses had no effect. Once 2000 ms elapsed, the next dot appeared for 600 ms , together with feedback (green tick for a correct prediction or a red cross for incorrect). If no prediction was made, a warning message 'Don't forget to guess' was displayed before the appearance of the next dot. If no prediction was made for five consecutive trials, the experiment stopped, and the message 'Please remember to respond' was displayed until the space bar was pressed. Each subject performed 500 trials separated into five blocks of 100 each, with an additional 10 practice trials. The sequence of dots was random, with the constraint that the frequency of left and right dots was equal for each block.

## Data analysis

In the sequential effects literature it has been customary to show results as an average of a few participants. However,
recent work has uncovered that individual differences are not only substantial but also meaningful in that they reflect different contributions - perceptual and motor - towards sequential effects (Gökaydin et al., 2016). Thus, average results are not conclusive with respect to demonstrating that sequential effects in prediction are similar to those in reaction time.

In order to calculate the probability of repeating/alternating as a function of the history of stimuli for each participant, each participant's trials were separated according to five-long histories of predictions, with prediction probabilities being calculated simply as the relative frequency with which a repetition or alternation was predicted as the fifth event in each of the 16 possible five-event-long histories presented on the $x$-axis of Figure 1. For instance, denoting frequency by $f($.$) ,$ the probability of alternating after predicting ARA was calculated as $p(A R A A)=f(A R A A) /(f(A R A A)+f(A R A R))$.

We will use $X_{s}, Y_{s}$ to denote the left/right dots and $X_{p}, Y_{p}$ the left/right predictions. In order to calculate the probability of repeating/alternating as a function of the history of stimuli, sequences such as $A R A R$ consisted of $X_{s} Y_{s} Y_{s} X_{s} X_{p}$ and $Y_{s} X_{s} X_{s} Y_{s} Y_{p}$. Probabilities of repeating/alternating were then calculated as above.

As discussed above, in order to assess whether sequential effects in prediction are similar to those in reaction time, we will use a simple model which is known to provide a good description of sequential effects in reaction time and apply it to sequential effects in prediction. Our model will consist of a simple linear combination of the perceptual elements of sequential effects. Our model then reads as $a P+b M$, where $a$ and $b$ are scalar free parameters and $P$ and $M$ are the perceptual and motor components of sequential effects - effectively just the profiles shown in Figure 2.

## Results

We will look primarily at individual results given that we know from reaction time studies that individual differences can be substantial (Gökaydin et al., 2016). Moreover, as discussed above, looking at averaged results is inconclusive with respect to assessing whether prediction and reaction time show the same type of sequential effects. We will discuss two types of analysis: prediction probability as a function of the history of predictions - prediction history profiles for short - and prediction probability as a function of the history of stimuli - or stimulus history profiles. Results from both types of analysis will be shown in turn. Overall, prediction history profiles emerged as having a larger number of individuals with a better fit to the combination of the components model: 17/21 prediction history profiles had an $R^{2}$ greater than 0.5 , compared to $8 / 21$ for stimulus history profiles. Note that a clear profile of sequential effects on one type of analysis was no guarantee that a clear profile emerged for the other type: several subjects displayed clear sequential effects on their prediction history profile but not in their stimulus history profile; conversely, one subject displayed strong sequential effects on the stimulus history profile but not on



Figure 4: Individual sequential effects in prediction as a function of stimulus history. Blue solid lines- empirical mean prediction probability. Red dashed lines - best fitting linear combination of the form $a P+b M$ where $P$ is represented by S-LRP and $M$ is represented by LRP-R. Inset bar plots show coefficients $a(P)$ and $b(M)$. Also shown is the $R^{2}$ value of the fits.

The remaining profiles in Figure 4 (bottom two) show two profiles which are consistent with a mixture of the two components of sequential effects, where the motor component is inverted but not the perceptual, and vice-versa. These types of profile, resembling an almost two-tiered dependence on the last event and whether this was a repetition or an alternation, are common at the individual level in RT studies despite only recently having been described (Gökaydin et al., 2016). Interestingly, no single participant exhibited a good fit to a combination of the two components where both had positive coefficients. Note that the two sets of subjects shown in figures 4 and 3 are different, with the exception of the lower-left panel of both figures, which show both types of analysis for the same individual.

Recall that, based on previous results, we hypothesized that prediction probability would be proportional to reaction time, reflecting the fact that humans predict less that to which they respond the fastest, and vice-versa. In the context of our model, this would imply that the coefficients of perceptual and motor components have the same sign on average in reaction time and prediction. Figure 5 shows the coefficient values of both components for all the individuals with very good fit to the model $\left(R^{2}>0.7\right)$. At first sight our re-


Figure 5: Coefficients of the best fitting linear combination $a P+b M$ for those subjects with $R^{2} \geq 0.7$. Blue triangles show the coefficients of the fit to results as a function of prediction history; red squares show fits to results as a function of stimulus history. The large blue triangle and red square show the respective means.
sults differ from RT experiments in one respect: in prediction profiles at least, the motor component often varies from a strongly positive to a strongly negative sign, whereas in RT it has almost always a positive sign (see Gökaydin et al. (2016), supplementary information). With respect to differences between prediction history profiles and stimulus history profiles - squares and triangles in Figure 5 - there is a hint that perhaps stimulus history shows one or the other component in isolation, whereas prediction history shows a more balanced mixture of the two components, but it is too early to draw any firm conclusions.

## Discussion

Our first hypothesis posited that we would find sequential effects in prediction similar to those in reaction time. The evidence presented here - while falling short of demonstrating that sequential effects in reaction and prediction are the same - does strongly suggest that sequential effects in prediction are similar to those in RT insofar as they are well captured by a combination of the two components of sequential effects in reaction time - perceptual and motor. That such clear profiles of sequential effects were obtained (Figures 3 and 4) was somewhat surprising given the smaller number of trials relative to typical RT tasks, as well as the less constrained nature of the task. Still, results were visibly noisier when compared to reaction time experiments, and many subjects failed to exhibit any appreciable fit to the two-component model. Nevertheless, we cannot rule out that those individuals who did not exhibit a good fit to the model exhibit a new type of sequential effect which is meaningful. In order to firmly establish the nature of sequential effects in prediction an experiment with larger numbers, followed by latent variable analysis - such as principal components analysis (PCA) (Gökaydin et al., 2016) - is necessary. This would allow us to match the structures of sequential effects in prediction and reaction time, rather than just a few individual results.

We also hypothesized that prediction probability would be
directly proportional to RT, rather than negatively proportional as is more intuitive. Again in this case conclusions can only be drawn on average, since there is considerable variation in the sign of the two components in both RT (see Gökaydin et al. (2016)) and prediction (Figure 5). Despite the small sample size, one difference did emerge: the sign of the motor component in prediction profiles ranges from strongly negative to strongly positive, whereas in RT studies the motor component seems constrained to be positive. Another interesting observation is that the motor component with a negative sign and in relative isolation - i.e. not in combination with the perceptual component - occurred in half $(4 / 8)$ of the stimulus profiles with an $R^{2}$ greater than 0.5 . In other words, when analysing results as a function of the history of stimuli, half of the participants were using only the motor component of sequential effects. Moreover, when analysing results as a function of the history of predictions, we obtained clear contributions from both the motor and perceptual components. At first sight, these results at odds with the interpretation of the components of sequential effects as associated with the perceptual and motor systems since - in a task where the sequence of responses and stimuli are de-coupled - we should recover the motor component when looking at the sequence of responses and the perceptual component when looking at the sequence of stimuli. Therefore, our results may force a reinterpretation of the motor/perceptual association of the two components of sequential effects.

There is some debate regarding the computational nature of sequential effects. Some authors argue that sequential effects reflect the tracking different types of statistics in the environment (Wilder, Jones, \& Mozer, 2009), whereas others argue that sequential effects are instead the product of the separate detection of alternating and repeating patterns (Maloney et al., 2005). We will use the latter interpretation in order to guide our discussion, but the different explanations are not incompatible and the ensuing discussion would hold if we interpret sequential effects as tracking different statistics. In the context of the pattern-detection interpretation, the perceptual component is the natural candidate for an alternation detector, whereas the motor component would play the role of a repetition detector (see (Gökaydin et al., 2016) for an explanation of this mapping). A change in sign of either coefficient would therefore imply a change in whether a particular subject is going for or against the respective pattern. For instance, when predicting, a positive sign of the motor component would mean that the the participant is going against a perceived repeating pattern, and the opposite is true for a negative coefficient. ${ }^{1}$ In light of this, we can now see that the variation in the sign of the coefficients of both components of sequential effects (Figure 5) may reflect a differential treatment of repeating and alternating patterns: in some cases subjects are going against both types of pattern - repeating or alternating - and other times against one but with the

[^73]other. The only combination we did not obtain was a negative sign on both coefficients, which would imply going with both patterns.

Earlier we proposed that the subjective perception of randomness might influence the polarity of sequential effects, since whether or not a pattern will continue depends on whether the sequence is random or not. At first sight, the randomness perception account would seem to imply that humans either go against both types of pattern - repeating and alternating - or with both. After all, if we perceive the sequence as being random we should bet against both repeating and alternating sequences continuing, and the opposite if we believe the sequence to be structured. However, it is conceivable that the perception of randomness has a differential effect on repeating and alternating patterns or, somewhat equivalently, that individuals give different weight to repetitions and alternations when judging a sequence to be random. In fact, it is well known from RT studies that there are substantial individual differences with respect to sensitivity to repetitions and alternations (Soetens et al., 1985; Gökaydin et al., 2016). In our experiment we did not bias the participants either way, and it is therefore natural to assume that individual perception of the random nature of the sequence would vary depending on endogenous factors. One way to test the influence randomness perception on sequential effects in prediction would be to conduct the same experiment giving participants a strong hint that the sequence is random, and contrasting these results with a situation where it is implied that the sequence has a pattern.

Some of the participants in our study exhibited a clear prediction history profile, some a clear stimulus history profile, and some both. The implication is that some humans are tracking the sequence of predictions, others the sequence of stimuli, and others both, in order to try and make predictions. What is it that makes some people more sensitive to one or the other type of information? The perception of randomness may yet again play a role in this respect, since a belief that the sequence is random should lead to a dismissal of the sequence of stimuli as uninformative. If participants believe the sequence is random they might try to generate the most random possible sequence of responses by using their repetition and alternation detectors 'in reverse' in order to create a sequence of responses that is poor in repeating and alternating patterns. If this were the case, we should expect to see a positive coefficient on both components in those participants with a clear prediction history profile (blue triangles in Figure 5), which seems to be the case for a few subjects, but not all. Again, we cannot discard the possibility that individual differences in sensitivity to repetitions and alternations might play a role in this case, and that some individuals might put more or less emphasis on repetitions or alternations when generating the most random sequence possible.

## Conclusion

We have shown for the first time that prediction tasks display sequential effects similar in nature to those observed in reaction time. This work goes some way towards unifying the areas of prediction and reaction time in binary decision tasks.

## References

Bertelson, P. (1961). Sequential redundancy and speed in a serial two-choice responding task. Quarterly Journal of Experimental Psychology, 13(2), 90-102.
Cho, R. Y., Nystrom, L. E., Brown, E. T., Jones, A. D., Braver, T. S., Holmes, P. J., \& Cohen, J. D. (2002). Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. Cognitive, Affective, \& Behavioral Neuroscience, 4(2), 283-299.
Gökaydin, D., Navarro, D. J., Ma-Wyatt, A., \& Perfors, A. (2016). The structure of sequential effects. Journal of Experimental Psychology: General, 145(1), 110-123.
Hale, D. J. (1967). Sequential effects in a two-choice serial reaction task. The Quarterly journal of experimental psychology, 19(2), 133-141.
Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. Journal of Experimental Psychology, 41(4), 291-297. doi: 10.1037/h0056878
Jentzsch, I., \& Sommer, W. (2002). Functional localization and mechanisms of sequential effects in serial reaction time tasks. Perception \& Psychophysics, 64(7), 1169-1188.
Maloney, L. T., Dal Martello, M. F., Sahm, C., \& Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. Proceedings of the National Academy of Sciences of the United States of America, 102(8), 3164-3169.
Nickerson, R. S. (2002). The production and perception of randomness. Psychological Review, 109(2), 330-357.
Oskarsson, A. T., Van Boven, L., McClelland, G. H., \& Hastie, R. (2009, March). What's next? Judging sequences of binary events. Psychological Bulletin, 135(2), 262-285.
Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. The Pavlovian Journal of Biological Science, 20(4), 163-170.
Soetens, E., Boer, L. C., \& Hueting, J. E. (1985). Expectancy or Automatic Facilitation? Separating Sequential Effects in Two-Choice Reaction Time. Journal of Experimental Psychology: Human Perception and Performance, 11(5), 598-616.
Wilder, M., Jones, M., Ahmed, A. A., Curran, T., \& Mozer, M. C. (2013). The persistent impact of incidental experience. Psychonomic Bulletin \& Review.
Wilder, M., Jones, M., \& Mozer, M. C. (2009). Sequential effects reflect parallel learning of multiple environmental regularities. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, \& A. Culotta (Eds.), Advances in Neural Information Processing Systems 22 (pp. 2053-2061). Curran Associates, Inc.

# Empirical tests of large-scale collaborative recall 

Monica A. Gates (mgates@berkeley.edu)<br>Jordan W. Suchow (suchow@berkeley.edu)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, Berkeley, USA


#### Abstract

Much of our knowledge is transmitted socially rather than through firsthand experience. Even our memories depend on recollections of those around us. Surprisingly, when people recall memories with others, they do not reach the potential number of items they could have recalled alone. This phenomenon is called collaborative inhibition. Recently, Luhmann and Rajaram (2015) analyzed the dynamics of collaborative inhibition at scale with an agent-based model, extrapolating from previous small-scale laboratory experiments. We tested their model against human data collected in a large-scale experiment and found that participants demonstrate non-monotonicities not evident in these predictions. We next analyzed memory transmission beyond directly interacting agents by placing agents into networks. Contrary to model predictions, we observed high similarity only within directly interacting pairs. By comparing behavior to model predictions in large-scale experiments, we reveal unexpected results that motivate future work in elucidating the algorithms underlying collaborative memory.


Keywords: collaborative memory; collaborative inhibition; network transmission; crowdsourcing; agent-based modeling

Our memories often rely on the people around us: every day we communicate with our colleagues and friends, forming and editing memories in each interchange. People learn to access each other's memories within long-term couples (Wegner, Erber, \& Raymond, 1991), and groups collectively form memories that define their values (e.g., Hirst \& Echterhoff, 2012). As people connect within increasingly larger networks, collaborative memory becomes ever more relevant.

In psychology, collaborative memory has historically been investigated in small-scale, lab-based experiments. Much work on group memory has thus focused on dyads or triads. However, our worlds are more richly connected than can be replicated in a lab setting, and many of the findings from this work may not be applicable to the larger systems of our everyday lives. To address this lack of understanding of large groups, recent efforts have focused on investigating memory abilities using agent-based modeling (Luhmann \& Rajaram, 2015). By analyzing human performance in past memory experiments, researchers can derive putative algorithms that describe human memory recall and embed these algorithms in artificial agents. These "agents" can then participate in novel memory paradigms with hundreds of agents interacting at a time. Agent-based modeling provided a solution to the difficulty of recruiting large numbers of participants and arranging them in the networks required by memory experiments.

However, we have recently developed a novel approach that allows us to overcome the previous impossibility of analyzing collaborative memory abilities at scale. Using new technology interfacing with web-based crowdsourcing tools such as Amazon Mechanical Turk, we can now recruit and
organize hundreds of online participants into real-time interactive chatrooms. Moreover, by considering participants as "nodes" in a network graph, we can assemble participants into arbitrary network structures.

The plan of the paper is as follows. We first validate our approach by replicating established collaborative memory effects in small groups, then investigate collaborative memory at unprecedented scale (Experiment 1). We then explore how memories spread beyond direct communication by examining memory transmission across networks (Experiment 2). We compare our human results to those predicted by agent-based modeling (Luhmann \& Rajaram, 2015) to determine the models' accuracy in describing behavior. We find that participants show memory effects not predicted by the model, illustrating the difficulty of extrapolating findings to larger groups. Within networks, participants also diverge from model predictions, showing reduced similarity in the words they recall beyond direct interactions. These results highlight the importance of large-scale studies in developing predictive models of human interaction, and further our understanding of the complexity of real-world network transmission and memory.

## Collaborative Inhibition

Imagine a group of people recalling a list of words collaboratively. The group would generate more words than any individual trying alone. However, the key comparison is not between the number of words on the group's list and the number of words on any one individual's list- it is between the group's list, and the cumulative list of what all the individuals could have done had they worked alone. This comparison is often made in the well-established "collaborative memory" task. In this task, participants listen to a long list of items (often words) and then recall as many items as possible, either as a group or individually. The number of words recalled by the group is compared to the number of words recalled by the "nominal group": the summed list of an equivalent number of individuals (redundant words removed). In the collaborative memory paradigm, nominal groups routinely outperform collaborative groups, a finding called collaborative inhibition. This effect has been replicated across many studies and variations on the paradigm (see Rajaram \& Pereira-Pasarin, 2010).

The leading theory describing collaborative inhibition is the retrieval disruption hypothesis (e.g., Basden, Basden, Bryner, \& Thomas, 1997; Rajaram \& Pereira-Pasarin, 2010). This hypothesis states that when initially listening to a wordlist, people form idiosyncratic representations of the words. When recalling words alone, participants effectively use their idiosyncratic organizations to recall the words.

However, when placed in groups, other participants can disrupt a participant's recall, leading to reduced performance. This hypothesis predicts that when participants are encouraged to organize information in similar ways, collaborative inhibition will disappear. In fact, when participants are experts in their domain (Meade, Nokes, \& Morrow, 2009) or are exposed to similarly ordered information (Finlay, Hitch, \& Meudell, 2000), inhibition does not occur. In generating model predictions, Luhmann and Rajaram (2015) incorporated the assumptions of the retrieval disruption hypothesis. In Experiments 1 and 2, we design empirical studies to compare behavioral results to these predictions.

## Agent-Based Model

In the model described by Luhmann and Rajaram (2015), agents encode $N$ items (words), where $N=40$. Agents have two representations. The first is an activation vector $\mathbf{A}$ of length $N$. Each entry $\mathbf{A}_{j}$ gives the probability that the given item $j$ will be retrieved. The second representation is an inter-item association matrix $\mathbf{S}$ of size $N \times N$. Each entry $\mathbf{S}_{i j}$ gives item $j$ 's association with item $i$ (associations were not necessarily reciprocal). This matrix would normally contain agents’ prior knowledge about word associations; however, Luhmann and Rajaram (2015) assigned values of $\mathbf{S}$ randomly between -2 and 2 to reflect agnosticism about the semantic relationships between words. (The $\mathbf{S}$ matrix was not used in our empirical studies.)

Each agent in this model has two behaviors. The first is encoding an item $i$. The first step in encoding an item is to reduce the activation of the maximally active item in vector $\mathbf{A}$, where $\beta$ is the learning rate:

$$
\begin{equation*}
\Delta \mathbf{A}_{\max }=-\beta \mathbf{A}_{\max } \tag{1}
\end{equation*}
$$

Next, the agent reduces the activations of items semantically associated with the maximally active item:

$$
\begin{equation*}
\Delta \mathbf{A}_{j}=-\beta \mathbf{S}_{j, \max } \mathbf{A}_{j} \tag{2}
\end{equation*}
$$

Finally, the agent increases the activation of the to-beencoded item $i$, with $\alpha$ acting as the learning rate:

$$
\begin{equation*}
\Delta \mathbf{A}_{i}=\alpha\left[1-\mathbf{A}_{i}\right] . \tag{3}
\end{equation*}
$$

The activation $\mathbf{A}$ vector is then normalized to ensure that its entries can be interpreted as probabilities: $\sum_{i} \mathbf{A}_{i}=1$.

An agent can also retrieve (and "orally state") an item. Agents take turns retrieving items, and on each turn, an agent retrieves an item with probability $\gamma$. The item $i$ that is retrieved is chosen according to the proportions in $\mathbf{A}$, such that items with higher activations are more likely to be retrieved. Then the activation vector is modified. First, the agent decreases activation of items semantically associated with the retrieved item $i$, in line with the theory of retrieval disruption:

$$
\begin{equation*}
\Delta \mathbf{A}_{j}=\beta \mathbf{S}_{j i} \mathbf{A}_{j} \tag{4}
\end{equation*}
$$

Next, if $i$ is not the maximally activated item, the agent reduces the activation of the maximally active item according
to Equation 1, and the activations of items semantically associated with the maximally active item according to Equation 2. Item $i$ is then encoded according to Equation 3. $\mathbf{A}$ is then normalized such that $\sum_{i} \mathbf{A}_{i}=1$. Just as the retrieving agent encodes the item after retrieving it, "listening" agents also then encode the item according to the encoding process described previously. Luhmann and Rajaram (2015) used the following parameter settings: $\alpha=0.2, \beta=0.05$, and $\gamma=0.75$.

In the first set of simulations in Luhmann and Rajaram (2015) comparing collaborative and nominal recall (our Experiment 1), model predictions were generated by presenting agents with wordlists and then having agents recall words via the described procedure. When an agent generated a word, it was shared with every other agent in the network. Agents participated in 20 rounds of retrieval within each simulation. A total of 1000 simulations (comparing 1000 collaborative and 1000 nominal results) were run for each group size.

In their second set of simulations analyzing agent interaction over networks (our Experiment 2), agents participated in 800 "timesteps" rather than rounds. In contrast with the large-scale collaborative simulations, in each timestep, every agent interacted with one other randomly chosen agent who was directly connected to them in the network. These interactions were pairwise, in contrast to previous simulations, during which the agent and their partner both had the opportunity to retrieve a word. This pattern of one-on-one interaction captures a form of organic social interaction in which someone may run into a friend and chat, and then continue on until they happen upon someone new.

Agents were placed in two types of networks: one empirically derived network, Zachary's karate club (Zachary, 1977), and one algorithmically derived network, a small-world network (Watts \& Strogatz, 1998). The karate club network describes the 78 links between 34 members of a club. Smallworld networks are based on the 6-degrees-of-separation phenomenon, the theory that it often takes around 6 links to connect any two individuals (Travers \& Milgram, 1969). To generate small-world networks, Luhmann and Rajaram (2015) used the Watts-Strogatz algorithm with the following parameters: 100 nodes, an average degree of 4 (participants were on average connected to four others), and a rewiring probability of 0.1 . A total of 1000 simulations were run for each network type. Small-world networks were randomly generated for each simulation.

In the network experiments, the measure of interest was similarity across agents. Agent similarity was compared by computing correlations across participants' activation vectors A. To capture the notion that agent similarity should be high both when agents mutually forgot or remembered a word, the absolute value of Pearson's correlation coefficient was used.

## Testing Model Predictions at Large Scale

In Experiments 1 and 2, we design empirical behavioral experiments that align with the specifications of the modeling work as closely as possible. However, the modeling work differed in that agents were allowed to submit any word that
they had not previously retrieved, whereas in the behavioral work, participants were not allowed to recall words that they or any other group members had previously recalled.

In Experiment 1, we first replicate findings from smallgroup experiments, then empirically explore the impact of large group size on collaborative inhibition. Previous work has suggested that collaborative inhibition increases as group size increases from 1 to 4 participants (Basden, Basden, \& Henry, 2000; Thorley \& Dewhurst, 2007), but Luhmann and Rajaram (2015) were the first to scale up to a hundred agents with their agent-based model. Their model predicts that collaborative inhibition rises with group size, peaks at around 8 individuals, and then begins decreasing (Figure 1a). Specifically, collaborative recall continues increasing with group size, but nominal recall hits ceiling at around 8 people as the disruption of idiosyncratic recall strategies is compensated by sheer group size. Since collaborative inhibition is the difference between nominal and collaborative recall, from this point collaborative inhibition begins to decrease. This prediction represents an extrapolation of results from small group sizes, and we tested the assumptions underlying this agentbased model by comparing human performance in the collaborative memory experiment to the model predictions.

In Experiment 2, we turn to memory transmission across networks. One person's behavior can have effects far beyond their direct connections, and viruses, information, and behaviors like smoking can spread over social networks. This transfer of information beyond direct interactions is called "hyperdyadic spread" (Christakis \& Fowler, 2009). Consistent with hyperdyadic spread, memory researchers have found that indirectly connected pairs have more similar memories than unconnected pairs (Yamashiro \& Hirst, 2014) and that distal partners can influence word recall (Choi, Blumen, Congleton, \& Rajaram, 2014). The model from Luhmann and Rajaram (2015) accordingly predicts that agents who never directly interacted, but share neighbors, will be similar (Figure 2a). Moreover, the model also predicts that agent similarity will depend on the networks that agents participate in. Agents in small-world networks were expected to be more similar than agents in karate club networks if they had directly interacted, but the opposite was expected for agents further apart. In Experiment 2, we implemented the agent-based network models with real participants to test these predictions.

## Experiment 1: Small and Large Groups

## Methods

Participants 1138 participants were recruited through Amazon Mechanical Turk. Participants were excluded from the experiment if they did not complete the pre-experiment arithmetic task and they did not contribute words in the main experiment. Sixteen participants were removed from the collaborative experiments for a total of 561 participants. Nominal groups were matched; thus 561 participants participated in the nominal experiments. The average ( $\pm \mathrm{SD}$ ) number of participants in collaborative experiments was 15.2
$\pm 0.7$ for groups of size $16,7.6 \pm 0.7$ for groups of size 8 , $4.0 \pm 0$ for groups of size $4,3.0 \pm 0$ for groups of size 3 , and $2.0 \pm 0$ for groups of size 2.

Participants would occasionally repeat the task, as they could choose to complete the task again on Amazon Mechanical Turk despite written advisement against this. Of the participant data included in this paper (other pilot task versions were also executed), in Experiment 1, 134 participants repeated the experiment more than once ( $14.6 \%$ of participants), and $30.03 \%$ of the data was generated by these participants. Participants participated an average of 1.22 times. The mean proportion of repeaters across group sizes (both nominal and collaborative) was as follows: group size of 2: $0.33 \pm$ 0.35 (SD), group size of $3: 0.39 \pm 0.31$, group size of $4: 0.27$ $\pm 0.21$, group size of $8: 0.27 \pm 0.20$, and group size of 16 : $0.27 \pm 0.09$. The participants who repeated the nominal experiments did not show improvement over time, despite having seen the same wordlists: the correlation between number of repetitions and number of words recalled was $r=-0.03$ (91 data points). Participants did improve across repetitions in the collaborative experiments ( $r=.47,89$ data points). However, the proportion of repeaters within experiments was anticorrelated with group size: the correlation between experiments of group sizes $2,3,4,8$, and 16 and the proportion of repeaters was $r=-0.086$.
Stimuli Participants saw 60 unrelated words, each selected from a different category from Overschelde, Rawson, and Dunlosky (2004). In collaborative experiments, words were presented roughly simultaneously across all participants. The average time ( $\pm \mathrm{SD}$ ) between presentation of a wordlist to the first participant compared to the last participant was as follows: $0.4 \pm 0.5$ seconds for groups of size $2,0.2 \pm 0.3$ secs for groups of size $3,0.9 \pm 0.9$ secs for groups of size $4,2.5 \pm 1.1$ secs for groups of size $8,4.2 \pm 2.6$ secs for groups of size 16 .

Procedure Participants observed wordlists: each word was presented for two seconds. After seeing the list, participants completed a 30 -second-long arithmetic filler task before advancing to the recall task. Participants were placed in chatrooms alone or with other participants, and were encouraged to type as many of the words they had seen as possible. Participants were not told how many other participants were in the chatroom: their responses appeared in blue font, and responses from all others appeared in black. They saw all previous words entered and were not permitted to submit any word that had already been submitted. This choice- that any words already present on the group list were not redisplayed- was made to encourage participants to read others' submitted words, and because it more closely matched the lab-based version of the collaborative memory paradigm, where verbal recall creates social pressure to not repeat words. There was no time limit for the recall task.

Experiments contained group sizes of $2,3,4,8$, or 16 participants. For each recall method (nominal or collaborative),



Figure 1: Effect of group size on collaborative inhibition. (a) Model results, reproduced from Luhmann and Rajaram (2015). (b) Behavioral results. The mean proportion of words ( $\pm \mathrm{SE}$ ) recalled in the nominal (red) and collaborative (blue) conditions are shown; collaborative inhibition (yellow) is the subtraction of these two values. Note that the horizontal axis is logarithmic.

48 groups of 2 were analyzed; 32 groups of 3 were analyzed; 24 groups of 4 were analyzed, and 12 groups of 8 and 16 were analyzed. This was a $2 \times 5$ design, crossing recall method by group size. In the nominal recall condition, participants recalled words alone, and their recall lists (with redundant words removed) were added together according to the appropriate group size. In the collaborative recall condition, groups of participants were placed in chatrooms and recalled together. Recalled words that had not been on the original lists were marked as incorrect and not included.

## Results

The collaborative inhibition effect is most reliably observed in triads (Rajaram \& Pereira-Pasarin, 2010), and we replicated this effect in our behavioral data at group size 3: $t(62)$ $=2.34, p=0.02, d=.60$, independent 2 -sample $t$-test. $(\alpha=$ .05 for all planned comparisons to follow. Unlike previous studies, we investigate the collaborative inhibition effect at multiple group sizes. Had we run separate studies, we would have used $\alpha=.05$, justifying its use here.) Collaborative inhibition in the literature is frequently but not always observed in pairs (Rajaram \& Pereira-Pasarin, 2010), but we did not observe this effect in this group size $(t(94)=0.78, p=0.44$, $d=.16$ ). In the two studies known to the authors examining tetrads (Thorley \& Dewhurst, 2007; Basden et al., 2000), a collaborative inhibition effect was observed, but we did not observe this effect at group size of $4(t(46)=-0.42, p=0.67$, $d=.13$ ).

Given previous results at group sizes of 2,3 , and 4 , it is reasonable to extrapolate and hypothesize that the trend of collaborative inhibition may be expected to continue or even widen at larger group sizes (Luhmann \& Rajaram, 2015). Intriguingly, this was not the pattern of results observed: participants did not show a collaborative inhibition effect at group size of 4 , and continued to not show a collaborative inhibition effect at group size of $8(t(22)=0.25, p=0.80, d=.11)$, contrary to model predictions. A collaborative inhibition effect did reoccur at group size $16(t(22)=2.17, p=0.04, d=.93)$, but variance for the nominal group is likely decreased due to
ceiling effects.
Overall, using a between-participants two-way unbalanced ANOVA, we surprisingly failed to observe a main effect of recall method $\left(F(1,246)=2.03, p=0.16, \eta^{2}=.0045\right)$ : nominal and collaborative groups did not recall significantly different numbers of words when results from all groups were combined (Figure 1b). We did observe the expected main effect of group size, $F(4,246)=49.08, p<0.0001, \eta^{2}=.44$, in that larger group sizes increased word recall. There was no interaction effect between recall method and group size $(F(4,246)$ $=1.15, p=0.33, \eta^{2}=.010$ ).

Contrary to the model predictions from Luhmann and Rajaram (2015), the collaborative inhibition effect became less strong at group sizes greater than 3 . This observation motivates the use of large-scale studies and further experiments testing whether the retrieval disruption and related hypotheses are enough to explain these results, or whether new models of human collaborative recall are necessary.

## Experiment 2: Networks

While people occasionally come together to work in shortterm groups, we often function in long-lasting social networks, communicating occasionally with far-flung friends. These networks are complex and can spread information at a prodigious rate: a secret you tell a close friend one day might be known by the whole community the next. In Experiment 2, we sought to investigate how people share and generate information when communicating across complex networks.

## Methods

Participants After removing inattentive participants, 383 participants were sorted into one of 12 karate club networks; the mean number of participants per network was $31.9 \pm 1.4$ (SD). 390 participants were sorted into 12 different smallworld networks; the mean number of participants per network was $32.5 \pm 0.7$. Removing participants changed the structure of the networks, and path lengths increased accordingly. 81 participants repeated the network experiments more than once ( $12.7 \%$ of participants), and $27.8 \%$ of the data was


Figure 2: Similarity of agents within networks. (a) Model results, reproduced from Luhmann and Rajaram (2015). (b) Behavioral results. Mean similarity ( $\pm \mathrm{SE}$ ) across agents is plotted as a function of the minimum distance between each agent node. Participants were arranged in karate club networks (red) or small-world networks (yellow).
generated by these participants. Participants participated an average of 1.21 times. The mean proportion of repeaters in the karate club experiments was $0.23 \pm 0.10$ (SD) and in the smallworld experiments was $0.32 \pm 0.11$. The participants who repeated the network experiments did not improve in the task: the correlation between number of repetitions and words recalled was $r=-0.18$ (159 data points).

Small-world networks were randomly generated for each experiment. The average time ( $\pm \mathrm{SD}$ ) between presentation of a wordlist to the first participant compared to the last participant was as follows: $6.9 \pm 7.4$ secs for the karate networks and $3.3 \pm 1.2$ secs for the small-world networks.

Procedure To compare our results with model predictions, we sought to replicate the model's paradigm as closely as possible with human participants. Each participant was assigned as a node in a graph with the option to communicate only with individual neighbors, where "neighbor" is defined as nodes participants were directly connected to. Every time a participant generated a word, their word was shared with a randomly chosen neighbor, rather than broadcast to the entire group as in Experiment 1. Networks were generated as described in the model, except that 34 agents were included in the small-world network, to match the karate club numbers.

## Results

Hyperdyadic Spread We first asked whether participants would show evidence of hyperdyadic spread. Luhmann and Rajaram (2015) computed agent similarity by comparing their agents' activation vectors $\mathbf{A}$, but in the non-modeling world we were restricted to externally observable correlates of participants' representations. Thus, we computed the absolute value of correlations for the list of words that participants recalled in the task as our measure of hyperdyadic spread.

Using recalled lists of words, we calculated the similarity between every two agents, and sorted the correlations based on agents' proximity in the network. Specifically, we used "hops" to describe how many connections were necessary to
link an agent node to another. If agents were connected and could directly communicate with each other, they were separated by a hop distance of 1 . If the closest path between agents included one other node, they were separated by a hop distance of 2. In this study, hyperdyadic spread would be observed if there was a non-zero similarity between agents at a hop distance of 2 or greater.

Though modeling predictions suggest agents will show hyperdyadic spread, participants who did not interact (participants separated by more than 1 hop) did not show evidence of hyperdyadic spread. We may have expected agents with shared neighbors to be similar, analogous to the spread of smoking habits (Christakis \& Fowler, 2009). However, habits develop over long time periods and are perhaps more transmissible than individual words, especially in social networks crafted from personal relationships. The advantage of using simplistic stimuli like wordlists is that if effective, we gain access to a reductionist, explainable system for investigation: in this case how memory representations are related. To this end, perhaps if semantically related words had been selected (rather than an unrelated wordlist presented in random order), we would have observed hyperdyadic spread in a new model system. This result highlights the effectiveness of iterating on model-based predictions and behavioral comparison.

Network Structure We next asked whether choice of network affects agent-pair similarity. In this case the model predictions closely align with behavioral results at a distance of 1 hop, though at greater distances the behavioral results exhibit unpredicted non-monotonicities. Specifically, at a distance of 1 hop, participants in small-world networks were more similar than in karate club networks $(t(22)=-6.24, p<0.0001, d$ $=2.66$, independent $t$-test), likely due to the increased local connections in small-world networks compared to the karate club network. Behavior at hop distances greater than 1 exhibited non-monotonicity: networks did not affect agent similarity ( 2 hops: $t(22)=-2.42, p=0.024, d=1.03 ; 3$ hops: $t(22)$ $=-1.17, p=0.26, d=.50 ; 4$ hops: $t(22)=-0.84, p=0.41, d$
$=.36 ; 5$ hops: $t(22)=1.45, p=0.16, d=.62 ; 6$ hops: $t(12)$ $=1.04, p=0.49, d=1.45 ; \alpha=.008$ to account for multiple comparisons).

Accordingly, the behavioral results did not show a main effect of network type across hop distance 1-6 $(F(1,122)=$ $1.44, p=0.23, \eta^{2}=0.0014$, between-participants 2-way unbalanced ANOVA), indicating that overall similarity between agents in karate club networks compared to small-world networks was not significantly different across all hop distances (Figure 2 b ). We observed an expected main effect of hop distance $\left(F(5,122)=163.86, p<0.0001, \eta^{2}=0.81\right)$, describing that agents were less similar when they were further apart. There was an interaction effect between recall method and hop distance $\left(F(5,122)=15.01, p<0.0001, \eta^{2}=0.075\right)$, describing the non-uniform decrease in agent similarity as hop distance increased. In sum, though behavioral results failed to show hyperdyadic spread, similarity between directly interacting agents was dependent on the network structure.

## General Discussion

In an increasingly interconnected world, understanding how our memories are impacted by interacting with others will influence how we organize ourselves, think and remember. In two studies, we examined collaborative memory: how remembering words in groups changes performance compared to recalling alone. We investigated collaborative memory in small and large groups and across network structures, comparing empirical results to the agent-based model predictions developed by Luhmann and Rajaram (2015).

We first replicated the collaborative inhibition effect in triads, the most reliable group size in exhibiting this effect. We then observed that in real participants, collaborative inhibition does not uniformly persevere at larger group sizes, despite what was suggested from small-scale studies (Luhmann \& Rajaram, 2015). One suggestion for why this could be is that factors outside of retrieval disruption affected participants. For example, post-experiment questionnaires indicated that in large groups where participants raced to submit nonrepeated words, some participants felt competitive pressure rather than the cooperation evident in smaller groups. While this issue might have been reduced by using a turn-taking structure rather than free recall, waiting for each participant in a large group to take a turn would introduce a new set of problems. Future models of collaborative memory should incorporate the intrinsic difficulties of organizing large groups of people, especially as people perhaps develop different strategies and algorithms to cope. To this end, large-scale quantitative work would be well complemented by fine-grained analysis of individual differences in strategies, and closer study of the interactions between individual cognition and the medium of interaction.

Model predictions suggested that the collaborative memory paradigm would be a good candidate for examining hyperdyadic spread, but empirically only participants who had directly interacted showed increased similarity. However, this
failure to observe hyperdyadic spread could perhaps be improved if semantically related wordlists were used or if a more sensitive measure than "words recalled" were used as the comparison metric. Moreover, if the behavioral task had been structured such that agents could re-submit words that their neighbors had submitted to other neighbors, we likely would have observed greater spread: future work will have to determine which paradigm structures will best inform our understanding of collaborative recall.

The unexpected results from these studies, in extrapolation to larger group sizes and in network structure, motivate and can inform future models describing the mechanisms underlying memory representations and collaborative interaction.

Acknowledgments. This work was funded in part by NSF grant 1456709 to T.L.G., DARPA Cooperative Agreement D17AC00004 to T.L.G and J.W.S., and N.I.H. U.C. Berkeley Neuroscience Training Program Grant to M.A.G.

## References

Basden, B., Basden, D., Bryner, S., \& Thomas, R. (1997). A comparison of group and individual remembering: does collaboration disrupt retrieval strategies? Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(5), 1176.
Basden, B., Basden, D., \& Henry, S. (2000). Costs and benefits of collaborative remembering. Applied Cognitive Psychology, 14(6), 497-507.
Choi, H.-Y., Blumen, H., Congleton, A., \& Rajaram, S. (2014). The role of group configuration in the social transmission of memory: evidence from identical and reconfigured groups. Journal of Cog nitive Psychology, 26(1), 65-80.
Christakis, N. \& Fowler, J. (2009). Connected: the surprising power of our social networks and how they shape our lives. Little, Brown.
Finlay, F., Hitch, G., \& Meudell, P. (2000). Mutual inhibition in collaborative recall: evidence for a retrieval-based account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(6), 1556.
Hirst, W. \& Echterhoff, G. (2012). Remembering in conversations: the social sharing and reshaping of memories. Psychology, 63(1), 55.

Luhmann, C. \& Rajaram, S. (2015). Memory transmission in small groups and large networks an agent-based model. Psychological Science, 26, 1909-1917.
Meade, M., Nokes, T., \& Morrow, D. (2009). Expertise promotes facilitation on a collaborative memory task. Memory, 17(1), 3948.

Overschelde, J. V., Rawson, K., \& Dunlosky, J. (2004). Category norms: an updated and expanded version of the norms. Journal of Memory and Language, 50(3), 289-335.
Rajaram, S. \& Pereira-Pasarin, L. (2010). Collaborative memory: cognitive research and theory. Perspectives on psychological science, 5(6), 649-663.
Thorley, C. \& Dewhurst, S. (2007). Collaborative false recall in the drm procedure: effects of group size and group pressure. European Journal of Cognitive Psychology, 19(6), 867-881.
Travers, J. \& Milgram, S. (1969). An experimental study of the small world problem. Sociometry, 425-443.
Watts, D. \& Strogatz, S. (1998). Collective dynamics of 'smallworld' networks. Nature, 393(6684), 440-442.
Wegner, D., Erber, R., \& Raymond, P. (1991). Transactive memory in close relationships. Journal of personality and social psychology, 61(6), 923.
Yamashiro, J. \& Hirst, W. (2014). Mnemonic convergence in a social network: collective memory and extended influence. Journal of Applied Research in Memory and Cognition, 3(4), 272-279.
Zachary, W. (1977). An information flow model for conflict and fission in small groups. Journal of anthropological research, 452473.

# Faulty Towers: A hypothetical simulation model of physical support 

Tobias Gerstenberg, Liang Zhou, Kevin A. Smith \& Joshua B. Tenenbaum

\{tger, zhoul, k2smith, jbt\} @mit.edu

Brain and Cognitive Sciences, Massachusetts Institute of Technology


#### Abstract

In this paper we introduce the hypothetical simulation model (HSM) of physical support. The HSM predicts that people judge physical support by mentally simulating what would happen if the object of interest were removed. Two experiments test the model by asking participants to evaluate the extent to which one brick in a tower is responsible for the rest of the bricks staying on a table. The results of both experiments show a very close correspondence between hypothetical simulations and responsibility judgments. We compare three versions of the HSM which differ in how they model people's uncertainty about what would happen. Participants' selections of which bricks would fall are best explained by assuming that hypothetical interventions only lead to local changes while leaving the rest of the scene unchanged.


Keywords: causality; counterfactual; hypothetical; mental simulation; intuitive physics; physical support.

## Introduction

When we look at a physical scene, such as the towers shown in Figure 1, we don't just see a pile of bricks. We also have a sense for how stable the different towers are and what is causing that stability (Battaglia, Hamrick, \& Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, \& Tenenbaum, 2016). In this paper, we look at how people judge the extent to which different bricks carry the responsibility for a tower's stability. We argue that people judge responsibility by imagining what would happen to the tower if the brick were removed, and develop a hypothetical simulation model (HSM) of physical support which captures this process.

We build on previous work in which we have shown how a counterfactual simulation model (CSM) explains people's causal judgments about dynamic collision events (Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2012, 2014, 2015; Gerstenberg \& Tenenbaum, 2016). In these experiments, participants saw collisions between billiard balls, and were asked to evaluate to what extent one ball had caused another ball to go through a gate in a wall (or prevented the ball from going through). The CSM assumes that people reach this judgment by comparing what actually happened with what would have happened in a counterfactual situation in which the candidate cause had been removed from the scene. As predicted by the model, participants' cause and prevention judgments increased the more certain they were that the outcome would have been different if the candidate cause had been removed from the scene. The CSM also captures the cognitive processes by which participants reach their judgments: participants' eye movements reveal how they spontaneously anticipate what would have happened in the relevant counterfactual situation (Gerstenberg, Peterson, Goodman, Lagnado, \& Tenenbaum, in press).

The CSM makes the strong prediction that counterfactual simulation forms a necessary part of how people make causal judgments, and that no adequate account of people's causal judgments about particular events can be developed that does


Figure 1: Experiment 1. Example stimuli. Note: Red bricks that would fall off the table if the black brick were removed (according to ground truth) are marked with a white dot at their center. The dots were not displayed in the actual experiment.
not rely on counterfactuals (cf. Wolff, 2007). Thus far, however, the CSM has only been applied to modeling causal judgments about dynamic collision events. Here, we demonstrate the generality of the account by showing how a model of hypothetical simulation naturally handles judgments about physical support.

Judging physical support is different from judging causation in several ways. First, hypotheticals are different from counterfactuals in that they are future-oriented and don't require going back in time (Beck, 2015). When making causal judgments about dynamic collisions, the observer needs to remember what actually happened, and contrast this with what would have happened in the relevant counterfactual situation. However, when making judgments of physical support in static scenes, like the tower configurations in Figure 1, there is no need to go back in time. We merely need to simulate what a possible future would look like in which certain aspects of the scene were changed.

Second, the mental simulations that are required to imagine the relevant counterfactual or hypothetical are different (cf. Freyd, Pantzer, \& Cheng, 1988; Holmes \& Wolff, 2010). When simulating counterfactuals, we want to stay as close as possible to what actually happened, and only modify the world as little as possible to make the counterfactual true (Gerstenberg, Bechlivanidis, \& Lagnado, 2013; Lewis, 1973; Pearl, 2000). But what do we keep constant in the causal model of the situation, and what do we change? When judging whether a ball would have gone into the goal, we need to simulate what the trajectory of the ball would have been if the collision hadn't taken place. To model people's uncertainty, we can add noise to the simulation of the ball's trajectory (cf. Smith \& Vul, 2013) and keep everything else that we know about the scene as it was (e.g. we wouldn't change the size


Figure 2: Schematic illustration of how different versions of the hypothetical simulation model apply noise when considering what would happen if the black brick were removed.
of the goal in the counterfactual simulation). However, when judging responsibility for a tower's stability, it is less clear what aspects of the scene we should hold constant. We will compare several implementations of the HSM that differ in how they capture people's uncertainty about what would happen.

The road map for the rest of the paper is as follows: We first present in detail how the HSM predicts judgments of physical support. We will test the model in two experiments in which we ask one group of participants to make hypothetical judgments, and another to evaluate causal responsibility. As predicted by the HSM, there is a very close correspondence between hypothetical and responsibility judgments. Heuristic strategies that focus on features of the scene (such as a tower's height, or the number of bricks on top of the brick of interest) cannot explain people's judgments as well. We end by discussing limitations of the current approach and by offering directions for future research.

## Hypothetical simulation of physical support

In our experiments, we ask participants how responsible the black brick is for the red bricks staying on the table. To derive predictions from the HSM we need to determine (1) what hypothetical situation to consider, and (2) how to simulate what would happen in that situation. We assume that when judging responsibility, participants consider a hypothetical situation in which the black brick is removed. Participants then use their intuitive understanding of physics to mentally simulate what would happen in that situation.

Recent work has argued that some aspects of people's intuitive understanding of physics are well-described by assuming we have an approximate simulation engine in our mind that is akin to a physics engine (Battaglia et al., 2013; Lake, Ullman, Tenenbaum, \& Gershman, 2016). Part of what makes these simulation engines "approximate" is that they assume that people's representation of a physical situation is uncertain. This uncertainty can come in many forms, such as perceptual uncertainty about the exact location of objects (Battaglia et al., 2013), dynamic uncertainty about how exactly an object will move (Smith \& Vul, 2013), and uncertainty about latent physical parameters such as friction and mass (Sanborn, Mansinghka, \& Griffiths, 2013).

To investigate whether people's mental simulations incorporate the assumption that only some aspects of the physical scene would directly be affected by the hypothetical intervention, we contrast three implementations of the HSM. These implementations differ in how they capture people's uncertainty about what would happen if the black brick were removed. All models apply noise in the same way: as a small impulse to some of the red bricks immediately after the removal of the black brick. The models differ, however, in which bricks they apply noise to. Figure 2 illustrates how the three different models work. The global noise model applies a small impulse to all the bricks and thus captures a general uncertainty about the scene (cf. Battaglia et al., 2013). The local noise model applies the impulse only to the red bricks that are directly in contact with the black brick. This model captures the assumption that participants will be most uncertain about what would happen in the area around the black brick. The above noise model applies noise only to bricks that are above the black brick and "connected" with it. Any brick that directly contacts and has its center of mass above that of the black brick counts as connected. This definition is then applied recursively. For example, brick 2 in Figure 2c is connected since brick 1 is in contact with and above the black brick, and brick 2 is in contact and above brick 1. This model captures that removing the black brick will affect the other bricks in an asymmetric way. Similar to when we lift a wooden block playing Jenga, this version of the model assumes that we have uncertainty particularly about those parts of the scene that would be affected by this kind of manipulation.

## Experiment 1

In the experiment, participants saw towers of bricks like the ones shown in Figure 1. Depending on the experimental condition, participants were asked to consider what would happen if the black brick weren't there, or evaluate the extent to which the black brick is responsible for the red bricks staying on the table. In line with the HSM, we predicted that there would be a close relationship between hypothetical and responsibility judgments.

## Methods

Design \& Procedure The experiment had three conditions that differed only in terms of the dependent measure. ${ }^{1}$ In the selection condition, participants were asked to "Please click on the red bricks that would fall off either side of the table if the black brick wasn't there." In the prediction condition, participants were asked to answer the question: "How many of the red bricks would fall off the table, if the black brick wasn't there?" Participants provided their answer on a sliding scale ranging from 0 to the number of red bricks present in the scene in steps of 1 . In the responsibility condition, participants were asked to answer the question: "How responsible

[^74]

Figure 3: Experiment 1. Scatter plots showing the relationship between the empirical probability with which each brick was selected and (a) the ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.
is the black brick for the red bricks staying on the table?" Responses were provided on a sliding scale ranging from "not at all" (0) to "very much" (100).

The procedure for all three conditions was identical. Participants first received instructions about the task. They then saw a number of warm-up animations that showed 20 bricks being dropped on the table. These animations were shown to familiarize participants with the relevant properties of the physical scene such as gravity, the friction between the bricks, as well as the table friction. Participants were only allowed to proceed to the next stage once they had watched at least five animations.

After the warm-up, participants saw 42 images of different towers of bricks in randomized order (see Figure 1 for examples). The stimuli varied the number of bricks on the table (range $=7$ to $20, \mathrm{M}=13.7, \mathrm{SD}=3.3$ ), as well as the number of red bricks that would fall off the table if the black brick were removed (range $=0$ to $6, \mathrm{M}=2, \mathrm{SD}=1.9$ ). Participants' tasks differed depending on the condition as described above. Finally, participants were asked to provide open-ended feedback about the task, and provided demographic information.

On average, the experiment took 15.71 ( $\mathrm{SD}=6.49$ ), 9.86 ( $\mathrm{SD}=6.49$ ), and 8.88 minutes $(\mathrm{SD}=8.90)$ in the selection, prediction, and responsibility condition, respectively.

Table 1: Summary of model results for Experiments 1 and 2 as applied to the data in the selection condition.

|  | Experiment 1 |  |  |  | Experiment 2 |  |  |  |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| model | r | RMSE | L | $\boldsymbol{\sigma}$ | r | RMSE | L | $\boldsymbol{\sigma}$ |
| truth | 0.55 | 34.74 | -21374 | 0 | 0.64 | 31.65 | -22279 | 0 |
| global | 0.75 | 20.92 | -9274 | 6.9 | 0.61 | 29.03 | -14034 | 2.5 |
| local | 0.70 | 22.26 | -9727 | 11.2 | 0.66 | 25.35 | -12617 | 7.2 |
| above | 0.87 | 15.34 | -8435 | 14.3 | 0.73 | 22.08 | -11824 | 12.5 |

Note: $\mathrm{r}=$ Pearson correlation, $\mathrm{RMSE}=$ root mean squared error, $\mathrm{L}=$ $\log$-likelihood of the data, $\sigma=\mathrm{SD}$ of the Gaussian from which the noise impulse is drawn that is applied to different bricks depending on the model.


Figure 4: Relationship between the predicted number of red bricks that would fall if the black brick weren't there (prediction condition) and number of selected bricks that would fall (selection condition). Note: The letters refer to the examples shown in Figure 1 for Experiment 1, and Figure 6 for Experiment 2. Error bars in all figures denote bootstrapped $95 \%$ confidence intervals.

Participants 121 participants $\left(\mathrm{M}_{\text {age }}=34, \mathrm{SD}_{\text {age }}=12,47 \mathrm{fe}-\right.$ male) were recruited via Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016) with $N=38$ in the selection condition, $N=42$ in the prediction condition, and $N=41$ in the responsibility condition. We excluded participants from further analysis based on their responses to the catch trial shown in Figure 1a. Eleven participants in the prediction condition were excluded because they predicted that at least one red brick would fall. Six participants in the responsibility condition were excluded because they gave a responsibility rating greater than 15 . No participants were excluded from the selection condition because no participant selected any of the bricks on the catch trial.

## Results

We will discuss the results from the selection, prediction, and responsibility conditions in turn.
Selection condition We tested how well the three different noise models captured participants' selections of which bricks would fall off the table if the black brick weren't there (see Figure 2). For each model, we used maximum likelihood fitting to find the noise parameter which predicts participants' selections best. For each setting of the noise parameter, we ran 100 simulations per stimulus and used the proportion of samples that each brick fell off the table in the noisy simulations to predict the probability that a given brick will be selected to fall by participants. (Figure 8 gives an example for what these predictions look like for stimuli used in Experiment 2.) Overall, the above noise model accounted best for the data (cf. Table 1).
Prediction condition Figure 4a shows the relationship between the number of bricks predicted to fall and the average number of bricks that participants selected in the selection condition. Overall, the two ways of probing participants' hypothetical simulations lead to very similar results. However, participants in the prediction condition predicted that more bricks would fall than participants in the selection condition selected (most of the data points are below the diagonal). The noise model which best accounted for participants' selections, also accurately predicts participants' average judgments about how many bricks would fall with $r=.88, \mathrm{RMSE}=0.84$.


Figure 5: Relationship between the predicted proportion of bricks that would fall if the black brick weren't there and responsibility judgments. Note: The letters refer to the examples shown in Figure 1 for Experiment 1, and Figure 6 for Experiment 2.
Responsibility condition Figure 5a shows the relationship between the proportion of bricks that participants in the prediction condition believed would fall off the table if the black brick weren't present in the scene, and participants' responsibility judgments. As predicted by the HSM, there was a very close relationship between prediction and responsibility judgments $r=.84, \mathrm{RMSE}=6.55$. This suggests that participants evaluated a brick's responsibility by considering what proportion of bricks would fall off the table if the brick weren't there. When we use the proportion of bricks selected in the selection condition to predict participants' responsibility judgments, we get a similarly good fit with $r=.78$, RMSE $=7.65$. A noise-free model that uses the proportion of bricks that actually fall off the table does not account well for participants' responsibility judgments $r=.35$, RMSE $=11.42$.

As an alternative to the HSM, we compared a heuristic model which predicts participants' responsibility judgments based on features of the physical scene. Table 2 shows how well the different features correlated with participants' judgments individually, as well as when combined via a linear regression model. We included features about the whole scene such as the number of bricks, the tower height, the average distance of each brick to the nearest edge of the table, as well as the average height and angle of each brick. We also included features specific to the black brick such as its distance to the nearest edge, its height and angle, as well as the number of bricks above it. To define the number of bricks above, we used the same criterion as the above noise model (cf. Figure 2c). As Table 2 shows, the best individual predictor for participants' responsibility judgments is the average height of each brick in the scene, followed by the number of bricks above the black one. Neither individual feature describes participants' responsibility judgments as well as the predictions
Table 2: Correlation coefficients between features and participants' responsibility judgments in Experiments 1 and 2. Note: The scene features, brick features, and all features columns show how well regressions that combine these features correlate with participants' judgments.

|  | $\begin{gathered} \mathrm{n} \\ \text { bricks } \end{gathered}$ | scene features |  |  |  | black brick features |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | tower height | avg edge distance | avg height | $\begin{gathered} \text { avg } \\ \text { angle } \end{gathered}$ | scene features | edge distance |  |  | n bricks above | brick features | $\left\lvert\, \begin{gathered} \text { all } \\ \text { features } \end{gathered}\right.$ |
| Exp 1 | . 16 | . 55 | . 39 | . 73 | . 21 | . 81 | . 02 | -. 19 | -. 05 | . 61 | . 62 | . 88 |
| Exp 2 | -. 05 | . 21 | -. 10 | . 07 | . 01 | . 26 | . 12 | -. 74 | -. 04 | . 69 | . 79 | . 85 |

Note: $\mathrm{n}=$ number, avg = average.


Figure 6: Experiment 2. Example stimuli. Note: White dots indicate which bricks would fall if the black brick weren't there. There were 6 different configurations of towers ( I through VI), and 7 different positions for the black brick in each tower, see c), d), and h).
(and selections) that participants made in the other two conditions of the experiment. A regression model that combines all features correlates well with participants' responsibility judgments ( $r=.88, \mathrm{RMSE}=5.89$ ), as does a model that only considers the scene features $(r=.81, \mathrm{RMSE}=7.14)$. A model which only includes features about the black brick doesn't fare as well $(r=.62, \mathrm{RMSE}=9.6)$. Even though a model that includes all features explains slighlty more of the variance that the HSM, this is likely due to overfitting; using model selection criteria, we find that the HSM performs better $(\mathrm{AIC}=276.52, \mathrm{BIC}=281.66)$ than the heuristic model ( $\mathrm{AIC}=283.72$, $\mathrm{BIC}=302.57$ ).

## Discussion

The results of Experiment 1 support the predictions of the HSM. Most importantly, there was a very close relationship between the responsibility judgments of one group of participants, and the number of bricks that another group of participants predicted would fall if the black brick weren't there. A heuristic model that does not rely on physical simulations but uses features that can be directly extracted from the scene fared worse when taking into account both variance explained and model complexity. We contrasted three implementations of the HSM which differ in the way in which they capture people's uncertainty about what would happen if the brick were removed. The results show that the above noise model correlates best with participants' selections. This model assumes that participants are particularly uncertain about what would happen to the bricks that are located above the black one.

## Experiment 2

Experiment 1 elicited participants' judgments for a wide array of different situations. In Experiment 2, we chose a more tightly controlled stimuli set, a selection of which is shown in Figure 6. We generated six different tower configurations. For each configuration, we chose seven positions for the black brick such that removing it would result in 0 to 6 red bricks falling off the table. While a heuristic model that used


Figure 7: Experiment 2: Scatter plots showing the relationship between the empirical probability with which each brick was selected and (a) the ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.
global scene features explained responsibility judgments well in Experiment 1, we expected this model to perform poorly here since it doesn't take into account where the black brick is positioned.

In order to better tell apart the different implementations of the HSM, we included tower configurations with disjointed sets of bricks (Tower III and Tower IV). For example, consider the configuration of bricks shown in Figure 6c. While a global noise model predicts that some of the red bricks on the right would fall off the table, the local versions of the model predict that only the bricks on the left side will fall.

## Methods

Design \& Procedure The design, procedure, and questions were identical to those of Experiment 1. Participants saw 43 trials in randomized order whereby one trial served as a catch trial. On average, the experiment took $13.04(\mathrm{SD}=6.87)$, $11.57(\mathrm{SD}=5.24)$ and 7.86 minutes $(\mathrm{SD}=3.48)$ in the selection, prediction, and responsibility condition, respectively.
Participants 129 participants $\left(\mathrm{M}_{\text {age }}=36, \mathrm{SD}_{\text {age }}=11.3,59\right.$ female) were recruited via Amazon Mechanical Turk with $N=42$ in the prediction condition, $N=44$ in the selection condition, and $N=43$ in the responsibility condition. We used the same exclusion criteria as in Experiment 1 based on the same tower shown in Figure 1a. 1 participant was removed in the selection condition, 3 participants in the prediction condition, and 3 in the responsibility condition.

## Results \& Discussion

Selection condition Figure 7 shows the correspondence between participants' brick selections and the predictions according to the ground truth as well as our three noise models as illustrated in Figure 2. Overall, the above noise model accounted best for participants' selections, as in Experiment 1 (cf. Table 1).

Let us look at the two situations shown in Figure 8 in some


Figure 8: Empirical selection percentages for two different stimuli together with the predicted selection probabilities according to the different noise models. The numbers (and color fill) indicate what percentage predicted that a particular brick would fall off the table if the black brick were removed. Red and black frames around a brick indicate that the brick would fall or stay on the table, respectively.
more detail. For the example shown in the top row, participants' selections corresponded closely to the ground truth. Since the global noise model applies an impulse to all the bricks, it incorrectly predicted that participants would select bricks on the right. The local noise model incorrectly predicted selections of bricks underneath the black one. The above noise model best predicted participants' selection in this case. It only assigned a small probability that any of the bricks on the right would be selected (because sometimes the bricks on top of the black brick will fall towards the right), or bricks that are underneath the black one.

The example in the bottom row shows a situation where participants' selections didn't correspond to the ground truth. Here, the majority of participants believed that none of the bricks would fall if the black brick weren't there. When the black brick is removed, the two bricks directly underneath it fall to the left and right, and the one falling to the right pushes the stack of bricks on the right off the table. None of our noise models was able to capture participants' selections in this case. The above noise model did a particularly poor job for the simple reason that it doesn't apply any noise in this case. Since the black brick is on top, its predictions correspond to the ground truth. What this clearly shows is that our noise models don't yet completely capture participants' hypothetical simulations. We will discuss some ideas about how the improve the models in the General Discussion below.
Prediction condition Figure 4 b shows the relationship between the number of bricks predicted to fall and the average number of bricks that participants selected in the selection condition. As in Experiment 1, there was a very close relationship between predictions and selections, and, again, participants predicted that more bricks would fall on average than they selected. The above noise model again best explained participants' predictions with $r=.76, \mathrm{RMSE}=1.41$.
Responsibility condition Figure 5b shows the relationship between participants' predictions and responsibility judgments. Like in Experiment 1, participants' responsibility judgments were well-accounted for by the proportion of
bricks that would fall off the table if the black brick were removed $r=.91$, RMSE $=8.66$. Again, we can also account for participants' responsibility judgments based on the proportion of bricks that were selected in the selection condition $r=.91$, RMSE $=8.67$. A noise-free model again fails to account well for participants' responsibility judgments with $r=.36, \mathrm{RMSE}=19.99$.

Table 2 shows how well different features of the physical scene correlate with participants' responsibility judgments in Experiment 2. Expectedly, global scene features did not correlate well with participants' responsibility judgments this time because these features do not capture the actual position of the black brick. For example, they don't distinguish the configuration shown in Figure 6 c from the one shown in Figure 6h. However, a good predictor of participants' responsibility judgments was the height of the black brick. The lower the black brick was located, the more responsible it was. Unlike in Experiment 1, the average height of the bricks in the tower did not correlate with responsibility judgments. Unsurprisingly, the number of bricks above the black brick was again a good predictor. However, there was no single predictor that accounted as well for participants' responsibility judgments as participants' predictions or selections in the other two conditions did. Even a regression that combines both scene and black brick features ( $r=.85, \mathrm{RMSE}=11.17$ ) does not explain participants' responsibility judgments as well as the HSM does.

## General Discussion

How do people judge physical support? In this paper, we develop and test a hypothetical simulation model (HSM) of physical support. Based on a model of counterfactual simulation which was originally developed to explain causal judgments about collision events (Gerstenberg et al., 2012, 2014, 2015; Gerstenberg \& Tenenbaum, 2016), the HSM predicts that we judge physical support by imagining what would happen if the object were removed. An individual brick is responsible for other bricks staying on a table to the extent that these bricks would fall off the table if that brick were removed. The results of two experiments show that the greater the proportion of bricks that participants predict would fall of the table, the more responsible that brick is seen for the other bricks staying on the table. Simple features of the physical scene such as the height of the tower, or the position of the brick of interest, as well as combinations of these features cannot explain participants' judgments as well.

The central claim of the HSM is that people judge physical support by simulating what would happen to the scene if the object of interest were removed. We contrasted three different implementations of the HSM which differ in how they model participants' uncertainty about what would happen in the relevant hypothetical situation. Similar to how people spontaneously consider counterfactuals when judging causation (Gerstenberg et al., in press), people naturally play "mental Jenga" when judging responsibility for physical support. Participants' selections of which bricks would fall were
best explained by a model that adds noise to the bricks located above the removed brick. While this model does a good job overall, there remain situations that it cannot capture adequately (cf. Figure 8).

We believe that there are at least three sources of uncertainty that affect participants' judgments: first, there is perceptual uncertainty about the exact spatial location of the different bricks (cf. Battaglia et al., 2013). Second, there is uncertainty about the hypothetical intervention itself: would the black brick simply disappear, or would it be removed, thereby affecting the bricks above it. Third, there is dynamic uncertainty about what would happen once the brick is removed (cf. Smith \& Vul, 2013). While the current implementation of the HSM uses a physics engine as a proxy for participants' mental model, we are eager to explore how an approximate simulation model (which doesn't represent each brick individually) might be able to capture participants' judgments (cf. Davis \& Marcus, 2016). Ideally, such a model would help explain when it is that people's physical intuitions are faulty and deviate from the ground truth.
Acknowledgments We thank Sunny Khemlani and two anonymous reviewers for their comments on the paper. This work was supported by the Center for Brains, Minds \& Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333.

## References

Battaglia, P. W., Hamrick, J. B., \& Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45), 18327-18332.
Beck, S. R. (2015). Why what is counterfactual really matters: A response to Weisberg and Gopnik (2013). Cognitive Science, 40(1), 253-256.
Davis, E., \& Marcus, G. (2016). The scope and limits of simulation in automated reasoning. Artificial Intelligence, 233, 60-72.
Freyd, J. J., Pantzer, T. M., \& Cheng, J. L. (1988). Representing statics as forces in equilibrium. Journal of Experimental Psychology: General, 117(4), 395-407.
Gerstenberg, T., Bechlivanidis, C., \& Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In M. Knauff, M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 23862391). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 378-383). Austin, TX: Cognitive Science Society.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, \& B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 523-528). Austin, TX: Cognitive Science Society
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 782-787). Austin, TX: Cognitive Science Society.
Gerstenberg, T., Peterson, M., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (in press). Eye-tracking causality. Psychological Science.
Gerstenberg, T., \& Tenenbaum, J. B. (2016). Understanding "almost": Empirical and computational studies of near misses. In A. Papafragou, D. Grodner, D. Mirman, \& J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 2777-2782). Austin, TX: Cognitive Science Society.
Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. Behavior Research Methods, 48(3), 829-842.
Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., \& Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. Cognition, 157, 61-76.
Holmes, K. J., \& Wolff, P. (2010). Simulation from schematics: dorsal stream processing and the perception of implied motion. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 2704-2709). Austin, TX: Cognitive Science Society.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., \& Gershman, S. J. (2016). Building machines that learn and think like people. arXiv preprint arXiv:1604.00289.
Lewis, D. (1973). Causation. The Journal of Philosophy, 70(17), 556-567.
Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge, England: Cambridge University Press.
Sanborn, A. N., Mansinghka, V. K., \& Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. Psychological Review, 120(2), 411-437.
Smith, K. A., \& Vul, E. (2013). Sources of uncertainty in intuitive physics. Topics in Cognitive Science, 5(1), 185-199.
Wolff, P. (2007). Representing causation. Journal of Experimental Psychology: General, 136(1), 82-111.

# Gaze Shifts between Text and Illustrations are Negatively Related to Reading Fluency in Beginning Readers 

Karrie E. Godwin (kgodwin1@kent.edu)<br>Kent State University, Department of Educational Psychology<br>150 Terrace Drive, White Hall, Kent, OH 44243 USA<br>Cassondra M. Eng (cassonde@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Ave. Pittsburgh, PA 15213 USA

Anna V. Fisher (fisher49@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology<br>5000 Forbes Ave. Pittsburgh, PA 15213 USA


#### Abstract

Learning to read is often considered the most important skill taught in school because reading is a gateway to other learning. Many children struggle to acquire this fundamental skill. Suboptimal design of books for beginning readers may contribute to the difficulties children experience as close proximity between text and illustrations could promote attentional competition hampering literacy skills. The present work utilized eye-tracking technology to examine how beginning readers allocate attention and whether these patterns are related to fluency (Experiment 1) and comprehension (Experiment 2). Results suggest when reading books in which text and illustrations are in close proximity, children frequently shift attention away from the text. This pattern of attention was negatively associated with fluency, but not associated with comprehension. This line of research aims to provide theoretical insights about design principles for reading materials that can be employed to optimize instructional materials and promote literacy development in young children.


Keywords: attention; reading; reading fluency; reading comprehension; illustrations; eye tracking

## Introduction

Reading is often considered the most important skill taught in elementary school: it is not only important in its own right, but it is a key gateway to other learning within and outside of school. Failing to 'learn to read' early in life is followed by failure to 'read to learn' later in life (National Association for the Education of Young Children, 1998). Many children struggle to acquire this fundamental skill for a variety of reasons, including but not limited to neurodevelopmental disorders (e.g., Dyslexia and ADHD), poor pre-reading skills (e.g., phonological awareness), and vulnerabilities in general cognitive functioning (e.g., working memory, processing speed, etc.) (e.g., Armbruster, Lehr, \& Osborn, 2009; Biederman et al., 2004; Dykman, \& Ackerman, 1991; Jacobson et al., 2011). In addition to these factors, empirical and theoretical work in the field of cognitive science may offer insights into how subtle changes in reading materials can affect the process of learning-to-read. It is important to understand how the properties of reading materials may affect children's emerging literacy skills because these factors are
considerably more malleable than factors intrinsic to the child, and thus can be leveraged to improve learning.

The typical layout of books for beginning readers intermix text with illustrations in close proximity (see Figure 1). In many cases illustrations are detailed, colorful, and engaging. There are a number of reasons for including illustrations in books for beginning readers, including: defining the setting and characters, contributing to text coherence, reinforcing the text, providing additional information, or motivating the reader (Carney \& Levin, 2002; Fang, 1996). However, the close proximity of text and illustrations may create competition for attentional resources, a situation that could be particularly problematic for beginning readers.


Figure 1. Examples of books for beginning readers where text and illustrations are in close proximity (i.e., embedded within the illustration).

There are theoretical reasons and related empirical findings that support the notion of competition between illustrations and text. According to the Dual-channel Theory of Multimedia Learning (Mayer et al., 2001), combining text with graphical representations can lead to split attention between the two sources of information. Similarly, the Cognitive Load Theory suggests that illustrations in close proximity to text may increase extraneous cognitive load on the learner thereby reducing the amount of cognitive resources available for text decoding (Torcasio, \& Sweller, 2010; Chandler \& Sweller, 1992).

A number of studies examined the effects of supplementing text with illustrations on a variety of
outcome measures (e.g., comprehension, retention, and problem-solving) relating to the goal of reading-to-learn in college students (presumably, a population of fluent readers). Some studies found that competition between text and graphical representations lead to reduced performance (e.g., Kalyuga, Chandler, \& Sweller, 1998; Mayer et al., 2001), whereas other studies reported facilitatory effects (e.g., Bétrancourt \& Bisseret, 1998; Ginns, 2005; Moreno \& Mayer, 1999). On the basis of this large body of evidence, researchers have formulated a number of principles regarding how to combine text with illustrations in a way that facilitates the process of reading-to-learn depending on the nature of the illustrations (e.g., illustrations that are well or poorly aligned with the text content), the level of difficulty of the text, and individual characteristics of the learners (e.g., learners possessing or lacking background knowledge relevant to the text content) (Carney \& Levin, 2002; Levin \& Mayer, 1993; Mayer, 2014).

However, the multimedia principles of effectively combining text and illustrations for the purpose of reading-to-learn in fluent readers may have limited applicability to the design of reading materials for young children whose goal is learning-to-read. Conceivably, the detrimental effects of competition for attentional resources between difficult-to-decode text and easy-to-interpret illustrations on emerging literacy skills may be more pronounced in beginning readers in whom reading has not yet become an automatized skill.

In contrast to the large body of research investigating the effects of combining illustrations with text for the purpose of reading-to-learn, few studies examined this question in the context of learning-to-read. The ubiquitous practice of combining illustrations and text in materials for beginning readers was first questioned in a handful of studies nearly five decades ago (Braun, 1969; Samuels, 1967). In these studies children were taught to read sight vocabulary with words either presented in isolation or next to corresponding illustrations during the training phase. During the testing phase, printed words were presented without pictures. The results of these studies showed performance was better during the training phase for words that were accompanied by pictures than for words presented in isolation; however, the opposite was true during the testing phase, suggesting that pictures presented alongside printed words interfered with the acquisition of sight vocabulary. In another study, kindergarten-age children were given reading instruction using a storybook in which text was either accompanied by illustrations or presented without illustrations (Samuels, 1970). For more skilled readers, there was no difference in learning gains whether children received reading instruction using a storybook with or without illustrations; however, less skilled readers showed higher gains in the no-pictures condition. More recently, Torcasio and Sweller (2010) reported that reading proficiency in 6- to 7 -year-old children improved more when children practiced reading a story without illustrations compared to reading the same story with illustrations.

The studies above provide suggestive evidence that close proximity of text and illustrations in books for beginning readers may interfere with learning-to-read. However, these studies have two critical shortcomings that limit their impact. First, although prior studies have proposed a mechanism by which illustrations in storybooks may disrupt reading fluency, they have not assessed this mechanism directly. Specifically, Samuels (1970) suggested that pictures may distract children from printed text. Similarly, Torcasio and Sweller (2010) suggest that when text is accompanied by illustrations, children devote working memory resources to processing the illustrations thus having less resources for processing the text. While this possibility is plausible, there is no direct evidence showing that children devote less resources to processing text in the presence of illustrations.

Second, prior studies focused on children's ability to read words quickly and accurately (i.e., a component of reading fluency) but largely did not consider reading comprehension. However, it is possible that the detrimental effect of illustrations for decoding could be offset by the potential beneficial effects of illustrations on reading comprehension. Indeed, instructing children to refer to illustrations to aid comprehension as well as decoding is a common instructional strategy in elementary school (Samuels, 1970) (although we should note the paucity of research on the effectiveness of this strategy). Alternatively, it is possible that by interfering with fluency, illustrations also interfere with comprehension, as Torcasio and Sweller (2010) suggested. When considering possible effects of illustrations on learning-to-read it is essential to assess both reading fluency and comprehension in order to obtain evidence that can have an impact on educational practice.

The present research investigates how beginning readers allocate their attention while reading and explores the possibility that gaze shifts away from the text (hypothesized to be due to the close proximity of text and illustrations) are negatively correlated with reading fluency (Experiment 1) and comprehension (Experiment 2). As noted previously, while we hypothesize that children's gaze shifts are a result of attentional competition induced by the close proximity between text and illustrations, it is also possible that these gaze shifts are an explicit strategy children deploy to aid decoding and comprehension. The present study does not rule out this alternative interpretation, a point we return to in the Discussion section.

## Experiment 1

## Method

## Participants

The sample consisted of 24 children $\left(M_{\text {age }}=7.20\right.$ years, $S D=$ 0.35 years, 10 females, 14 males). Participants were recruited from schools in and around Pittsburgh Pennsylvania. Children were tested individually by trained hypothesis-blind research assistants.

## Design and Procedure

## Book Selection

To maintain a high level of ecological validity, children read commercially available books designed for beginning readers from the I Can Read book series (Children read either Biscuit Wants to Play (2002), or Biscuit Goes to School (2003) written by Alyssa Satin Capucilli).

Children read aloud a commercially available book for beginning readers in which the text and illustrations are in close proximity. A Tobii X2-60 portable eye tracker was used to measure children's patterns of attention allocation indexed by gaze shifts. Prior to reading the story an independent measure of reading fluency was administered (i.e., Word Recognition in Isolation Test). Reading fluency was also assessed while children were reading the story using a Running Record. Additional details regarding each measure are provided below.

## Measures

## Gaze Shifts

Eye gaze is a common measure of attention in a variety of settings and is a particularly appropriate measure in the context of reading, a complex cognitive task in which eye gaze location and the focus of attention are difficult to dissociate as they are thought to overlap (for review see Rayner, 2009). The Tobii X2-60 portable eye tracker was utilized to measure children's eye movements while reading. On each page of the book, text, picture, and white space Areas Of Interest (AOI's) were created. A python script was then used to calculate the number of times a child shifted fixation away from the text AOI's (i.e., to illustrations or white space AOI's) and the average number of gaze shifts per page was then calculated.

## Reading Fluency Measures

Fluency is defined as "accurate and automatic decoding of the words in the text, along with expressive interpretation of the text" (Rasinski, 2004, p. 2). In the present experiment one component of reading fluency, decoding accuracy, was assessed. Two measures of reading fluency were utilized: the Word Recognition in Isolation test and a Running Record. Both measures are described below.
Word Recognition in Isolation Test. Children completed a modified Word Recognition in Isolation (WRI) test, which is a common measure of reading fluency (Morris, 2013). In the WRI, participants were shown leveled lists of words presented individually on a computer screen. The participant is tasked with reading as many words as they can within the allotted time limit (in order to avoid frustration, a ceiling was imposed such that testing ceased if a participant failed to read at least $50 \%$ of the words correctly on a given word list). The child's score was based on the number of words the participant correctly read aloud within the time limit (i.e., number of words read correctly out of 100 total possible words).

Running Record. While participants read the story aloud, the experimenter completed a Running Record (Clay, 1972) in which the experimenter recorded the child's decoding accuracy for each word in the story. The percentage of correct responses was then calculated.

## Results

On average, children switched their point of fixation away from the text 3.68 times (range $=1.50$ to 11.29 times) per page. The average number of words per page was 6.94 ; thus even when the text length was relatively short many children frequently shifted their attention away from the text (see Figure 2).


Figure 2. A sample gaze plot showing multiple gaze shifts away from the text.

Furthermore, children's tendency to switch their point of fixation away from the text while reading was negatively associated with the WRI ( $M=.41, S D=.27$ ), an independent measure of reading fluency, as shown in Figure 3 (Panel A); $r=-.60, p=.002$. Additionally, Running Record accuracy was obtained for a subset of participants ${ }^{1}(\mathrm{n}=13 ; M=.97, S D$ $=.04$ ). Critically, this online measure of fluency was also negatively (and strongly) correlated with the number of gaze shifts away from the text (Figure 3, Panel B); $r=-.96$; $p<.0001$. Thus, less fluent readers (as measured by both the WRI and the Running Record) tended to make more frequent gaze shifts away from the text.

The present results indicate gaze shifts away from text are negatively associated with reading fluency; however, it is an open question as to whether gaze shifts away from the text are also associated with reading comprehension. Experiment 2 begins to explore this possibility.

## Experiment 2

## Method

## Participants

The sample consisted of 17 children $\left(M_{\text {age }}=7.15\right.$ years, $S D=$ 0.59 years, 6 females, 10 males, and 1 child whose sex was not reported by the parents). Due to technological issues, for 1 child no eye tracking data was available. Participants were

[^75]recruited from schools and camps in and around Pittsburgh


Figure 3. Association between gaze shifts away from text and performance on two measures of reading fluency: WRI (Panel A) and Running Record (Panel B). Note only a subset of the sample had Running Record data available; thus, the gaze shift range differs across the two panels.

Pennsylvania. Children were tested individually by trained hypothesis-blind research assistants.

## Design and Procedure

## Book Selection

In order to test the generalizability of the results from Experiment 1, a new book was selected for Experiment 2. We selected a book that had a more complex story line in order to address the question of whether gaze shifts away from the text are also negatively associated with reading comprehension. As in Experiment 1, children read a commercially available book designed for beginning readers in order to maintain a high level of ecological validity. Children read a book from the Hooked on Phonics Learn to Read series entitled Good Job Dennis written by Amy Kraft.

Akin to Experiment 1, children read aloud a commercially available book in which the text and illustrations are in close proximity. A RED250 mobile eye tracker was used to measure children's patterns of attention allocation indexed by gaze shifts. Prior to reading the story an independent measure of reading fluency was administered (i.e., Word Recognition in Isolation Test). Reading fluency was also assessed using a Running Record. After reading the story
children's reading comprehension was assessed. Additional details regarding each measure are provided below.

## Measures

Gaze Shifts
A RED250 mobile eye tracker (SensoMotoric Instruments, Inc.) was used to measure children's patterns of attention allocation. We used the RED250 mobile eye tracker in Experiment 2 (opposed to the Tobi X2-60 utilized in Experiment 1) due to its higher sampling rate which makes it better suited for reading studies. On each page of the book, text, picture, and white space AOI's were created. The SMI BeGaze Eyetracking Analysis Software was then used to calculate gaze shifts away from the text AOI's (i.e., to illustrations or white space AOI's) and the average number of gaze shifts per page was then calculated.

## Reading Fluency Measures

The reading fluency measures (WRI and Running Record) were identical to those described in Experiment 1.

## Reading Comprehension Measures

Retelling Prompt Retelling is a common measure of reading comprehension in elementary school (Nilson, 2008). Children were asked to orally recount the story. Retelling accuracy was scored by calculating the number of key events correctly recounted (out of 5 possible events). Scores are reported as the proportion of correct responses.

Story Questions At the end of the story children were asked 3 questions probing their memory for details from the story. For example, in the story various animals escape from a pet store including cats, dogs, birds, rabbits, and frogs. Children might be asked to recall which pets escaped. Children could receive full credit if in their response the child identified 4 or more animals that escaped and 0 points if they failed to recall the animals that escaped or provided an incorrect response. Children could also receive partial credit if the child recalled correctly a subset of the animals. Accuracy on the Story Questions was scored as the proportion of correct responses (out of 7 possible points).

## Results

On average, children switched their point of fixation away from the text 8.03 times per page (range $=1.75$ to 15.5 times). The average number of words per page was 42.67 . As in Experiment 1, there was a significant correlation between the number of gaze shifts and both measures of reading fluency (WRI and Running Records). Children's tendency to switch their point of fixation away from the text while reading was negatively associated ( $r=-.78, p<.0004$ ) with the WRI $(M=.69, S D=.19)$ as shown in Figure 4 (Panel A) and with their Running Record accuracy ( $M=96 \%$, $S D=5 \%$; Panel B; $r=-.63, p=.009$ ). This pattern of results provides corroborating evidence that gaze shifts away from the text are related to poorer reading fluency scores.


Figure 4. Association between gaze shifts away from text and two measures of reading fluency (Panel A-B) and reading comprehension (Panel C) in Experiment 2.

A composite variable was created for the reading comprehension measures. Retelling score ( $M=.45 S D=.23$ ) and Story Question score $(M=.59, S D=.29)$ were standardized using Z-scores and averaged together to create the composite variable, Comprehension Composite Score. Thus, the comprehension composite score reflects how many standard deviations the child is away from the overall mean. A score of 0 would indicate the child's performance is equal to the overall mean (an average performer on the Story Question and Retelling measures). There was no significant correlation between gaze shifts and the Comprehension Composite, $r=-.13, p=.62$ (See Figure 4

Panel C). In contrast to the results for fluency, the observed pattern of results for comprehension are not consistent with the hypothesis that the close proximity between illustrations and text could impede reading skills. Interestingly, the results also seem to suggest that children are not utilizing illustrations (or perhaps not utilizing illustrations in an effective manner) to aid comprehension, as one might expect to observe a positive correlation between gaze shifts and reading comprehension if referencing illustrations was in fact scaffolding children's reading comprehension. Future research is needed to expound upon this initial work, a point we return to in in the discussion section.

## Discussion

Although the practice of using illustrations in materials for teaching children to read has undergone little change in over 250 years (Samuels, 1970), no research has systematically examined the possible costs and benefits of this practice for children's attention, reading fluency, and comprehension. By leveraging mobile eye tracking technology, the present research found that gaze shifts away from the text (hypothesized to be due to the close proximity of text and illustrations) are negatively associated with reading fluency, an important component of literacy. This work highlights the importance of investigating experimentally more optimal book designs for beginning readers. This work also addresses an important gap in the prior literature, which focused almost exclusively on components of reading fluency. The present work adds to the prior literature by investigating the relationship between attention allocation (indexed by eye gaze) and both reading fluency as well as reading comprehension.

Our working hypothesis is that the majority of gaze shifts away from the text occur as a result of the competition between text and illustrations being resolved in favor of the latter. The findings reported above are consistent with this hypothesis: across two experiments with different samples of children and different reading materials we observed a strong and negative relationship between the frequency of gaze shifts away from text and measures of reading fluency. However, given the correlational nature of the study, other interpretations of these findings are of course possible. For example, it is possible that less proficient readers attempt to use illustrations to help with decoding and comprehension. Future experimental work is necessary to adjudicate between these possibilities. Towards this goal, we are currently collecting data in which we assess children's reading fluency and comprehension while children are reading stories in which the placement of illustrations relative to the text is experimentally manipulated.

Overall, the present study provides evidence that gaze shift patterns are associated with reading performance, and thus highlights the need for further research into the nature of this association. Future research should also expound upon these findings to ascertain whether, illustrations in close proximity to text allow children to bypass the important work of decoding as this "shortcut" may have
cumulative effects on children's literacy acquisition and subsequent learning. Additional research is also needed to systematically examine the role of individual differences. One future direction we are currently pursuing is whether the negative association between illustrations and reading fluency can be offset by modifying the design of books to reduce the competition between text and illustrations through increased spatial separation, a layout we hypothesize to be more optimal for beginning readers.

This work begins to build an important foundation of research that has direct implications for educators and publishers and that aims to ultimately improve children's literacy acquisition.

## Acknowledgments

We thank Kristen Boyle, Lindsay Gorby, Ashli-Ann Douglas, Rebeka Almasi, Hyunji Do, Graciela Garcia, and Clara Lee for their help collecting data and Marcus Godwin for creating the Python script to compute participants' gaze shifts. We also thank the children, parents, and teachers at the CMU Children's School, St. Edmund's Academy, and Sacred Heart Elementary School in Pittsburgh (PA) who made this project possible. This work was partially supported by the ProSEED/Simon Initiative Seed grant from Carnegie Mellon University and by the James S. McDonnell Foundation $21^{\text {st }}$ Century Science Initiative in Understanding Human Cognition - Scholar Award (220020401) to A.V.F.

## References

Armbruster, B. B., Lehr, F., \& Osborn, J. (2009). Put reading first: The research building blocks of reading instruction: Kindergarten through grade 3 ( $3^{\text {rd }}$ ed.). C. R. Adler (Ed.). National Institute for Literacy. Retrieved from https://lincs.ed.gov/publications/pdf/PRFbooklet.pdf
Bétrancourt, M., \& Bisseret, A. (1998). Integrating textual and pictorial information via pop-up windows: an experimental study. Behaviour \& Information Technology, 17(5), 263-273.
Biederman, J., Monuteaux, M. C., Doyle, A. E., Seidman, L. J., Wilens, T. E., Ferrero, F., Morgan, C. L., \& Faraone, S. V. (2004). Impact of executive function deficits and Attention-Deficit/Hyperactivity Disorder (ADHD) on academic outcomes in children. Journal Of Consulting And Clinical Psychology, 72(5), 757-766.
Braun, C. (1969). Interest-loading and modality effects on textual response acquisition. Reading Research Quarterly, 4(3), 428-444.
Carney, R. N., \& Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. Educational Psychology Review, 14(1), 5-26.
Chandler, P., \& Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. British Journal of Educational Psychology, 62(2), 233-246
Clay, M. (1972). The early detection of reading difficulties: A diagnostic survey.
Dykman, R. A., \& Ackerman, P. T. (1991). Attention deficit disorder and specific reading disability: Separate but
often overlapping disorders. Journal Of Learning Disabilities, 24(2), 96-103.
Fang, Z. (1996). Illustrations, text, and the child reader. What are pictures in children's storybooks for? Read. Horizons 37: 130-142.
Ginns, P. (2005). Meta-analysis of the modality effect. Learning and Instruction, 15, 313-331.
Jacobson, L. A., Ryan, M., Martin, R. B., Ewen, J., Mostofsky, S. H., Denckla, M. B., \& Mahone, E. M. (2011). Working memory influences processing speed and reading fluency in ADHD. Child Neuropsychology, 17(3), 209-224.
Kalyuga, S., Chandler, P., \& Sweller, J. (1998). Levels of expertise and instructional design. Human Factors, 40, 117.

Levin, J. R., and Mayer, R. E. (1993). Understanding illustrations in text. In B. K. Britton, A. Woodward, \& M. R. Brinkley (Eds.), Learning from Textbooks: Theory and Practice (pp. 95-113). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.
Mayer, R. E. (Ed.) (2014). The Cambridge handbook of multimedia learning. New York: Cambridge University Press.
Mayer, R.E., Heiser, J., \& Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. Journal of Educational Psychology, 93, 187-198.
Moreno, R., \& Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. Journal of Educational Psychology, 91(2), 358.
Morris, D. (2013). Diagnosis and correction of reading problems ( $2^{\text {nd }}$ ed.). New York, NY: Guilford Press.
National Association for the Education of Young Children (1998). Learning to read and write: Developmentally appropriate practices for young children. A joint position statement of the International Reading Association and the National Association for the Education of Young Children. Young Children, 53(4), 30-46.
Nilson, N. L. (2008). A critical analysis of eight informal reading inventories. The Reading Teacher, 6(17), 526536.

Rasinski, T. V. (2004). Assessing reading fluency. Pacific Resources for Education and Learning. Honolulu: HI. Retrieved from: http://files.eric.ed.gov/fulltext/ED483166.pdf
Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. The Quarterly Journal Of Experimental Psychology, 62(8), 1457-1506.
Samuels, S. J. (1967). Attentional process in reading: The effect of pictures on the acquisition of reading responses. Journal of Educational Psychology, 58(6), 337-342.
Samuels, S. J. (1970). Effects of pictures on learning to read, comprehension, and attitudes. Review of Educational Research, 40(3), 397-407.
Torcasio, S., \& Sweller, J. (2010). The use of illustrations when learning to read: A cognitive load theory approach. Applied Cognitive Psychology, 24(5), 659-672.

# A Cognitive Model of Strategic Deliberation and Decision Making 

Russell Golman (rgolman@andrew.cmu.edu)<br>Carnegie Mellon University, Pittsburgh, PA.

Sudeep Bhatia (bhatiasu@sas.upenn.edu)<br>University of Pennsylvania, Philadelphia, PA.


#### Abstract

We study game theoretic decision making using a bidirectional evidence accumulation model. Our model represents both preferences for the strategies available to the decision maker, as well as beliefs regarding the opponent's choices. Through sequential sampling and accumulation, the model is able to intelligently reason through two-player strategic games, while also generating specific violations of Nash equilibrium typically observed in these games. The main ingredients of accumulator models, stochastic sampling and dynamic accumulation, play a critical role in explaining these behavioral patterns as well as generating novel predictions.


Keywords: Decision making; Game theory; Sequential sampling; Preference accumulation

## Introduction

Game theory studies the behavior of idealized decision makers. The standard solution concept for a strategic game is Nash equilibrium, which relies on common rationality and accurate expectations. Given expectations of others' choices, players behave rationally, and the resulting play conforms to these expectations (Luce \& Raiffa, 1957).

Not surprisingly, human decision makers display numerous systematic departures from Nash equilibrium (see Camerer, 2003 for a review). We present a cognitive model of strategic deliberation and choice in one-shot, two-player games, that is able to accommodate these departures. Our model proposes that decision makers dynamically and stochastically accumulate both their own preferences for available strategies, as well as beliefs about the opponent's preferred strategies. There are bidirectional relationships between preferences and beliefs, so that beliefs about what the opponent will choose influence the decision makers' preferences, and these preferences in turn influence beliefs about the opponent's choices. Ultimately, decision makers can respond to what they think the opponent will do, and also revise these beliefs as they deliberate.

Our model can be seen as an extension of decision field theory (Busemeyer \& Townsend, 1993; also Bhatia, 2014 and Rieskamp, 2006), an existing accumulator-based theory of non-strategic risky choice. Accumulator models rely on two main ingredients: stochastic sampling and dynamic accumulation (see Busemeyer, 2015 for a review). These ingredients are critical in our model for making deliberation subject to intrinsic variability and requiring it to play out over time, and we show that both ingredients have a central
role in capturing the behavioral patterns observed in strategic choice. By demonstrating the relationship between our model and established preference accumulation models, we demonstrate that a single framework can be used to understand choice behavior across a variety of non-strategic and strategic settings.

## Game Theoretic Decision Making

In strategic games, two or more players make choices over a set of strategies. Crucially, the strategies chosen by the players collectively determine the outcomes of the game, so that each player's utility depends on the other's choice as well as on their own. We define a finite-strategy two-player game with a set of pure strategies for each player, $S_{1}=$ $\left\{s_{11}, \ldots s_{1 N}\right\}$ and $S_{2}=\left\{s_{21}, \ldots s_{2 M}\right\}$ respectively, and a pair of payoff functions $u_{1}$ and $u_{2}$ that give each player's utility for each profile of pure strategies $\left(s_{1 i}, s_{2 j}\right)$. Thus if player 1 selects $s_{1 i}$ and player 2 selects $s_{2 j}$ the utility for player 1 is $u_{1}\left(s_{1 i} ; s_{2 j}\right)$ and the utility for player 2 is $u_{2}\left(s_{2 j} ; s_{1 i}\right)$, with $\boldsymbol{u}_{i j}=\left(u_{1}\left(s_{1 i} ; s_{2 j}\right), u_{2}\left(s_{2 j} ; s_{1 i}\right)\right)$. We define the set of best responses for player $\mu$ to an opponent's strategy $s_{-\mu}$ as $\operatorname{BR}\left(s_{-\mu}\right)=\arg \max u_{\mu}\left(s_{\mu} ; s_{-\mu}\right)$. Then a pure strategy Nash equilibrium can be defined as a strategy profile $\left(s_{1 i}, s_{2 j}\right)$ such that $s_{1 i} \in \operatorname{BR}\left(s_{2 j}\right)$ and $s_{2 j} \in \operatorname{BR}\left(s_{1 i}\right)$.

There are a number of settings where Nash equilibrium fails to accurately describe human behavior. For example, Nash equilibrium predicts unraveling when players have incentives to undercut each other. Consider the traveler's dilemma game (Basu, 1994), in which two travelers have lost identical items and must request compensation. The airline accepts the lower claim as valid and pays that amount to both players, and, additionally penalizes the higher claimant with a fee and rewards the lower claimant with a bonus. We represent this game with the strategy sets $S_{1}=S_{2}=\{20,30, \ldots, 90\}$, where $x_{1 i}$ and $x_{2 j}$ correspond to the amounts (in dollars) associated with strategies $s_{1 i}$ and $s_{2 j}$, and we have utilities $\boldsymbol{u}_{i j}=\left(0.01\left(x_{2 j}-\gamma\right), 0.01\left(x_{2 j}+\gamma\right)\right)$ if $x_{1 i}>$ $x_{2 j}, \boldsymbol{u}_{i j}=\left(0.01 x_{l j}, 0.01 x_{2 j}\right)$ if $x_{l i}=x_{2 j}$, and $\boldsymbol{u}_{i j}=\left(0.01\left(x_{l j}+\gamma\right)\right.$, $\left.0.01\left(x_{2 j}-\gamma\right)\right)$ if $x_{1 i}<x_{2 j}$. Here $\gamma$ corresponds to the reward/penalty offered by the airline, and is set so that $10<$ $\gamma \leq 20$. For comparability with other games, we have scaled utilities to lie between 0 and 1 .

The airline's scheme rewards undercutting the other traveler. The best response is always to claim exactly 10 less than the other traveler does (if it is feasible to do so).

As a result, the only Nash equilibrium strategy for both players is to claim 20. In experiments average claims actually are well above the lower bound that Nash equilibrium predicts (e.g. Capra et al., 1999).

Experiments on the traveler's dilemma game also find that claims are higher when the reward/penalty, $\gamma$, is lower. This payoff sensitivity is hard to reconcile with players choosing best responses to the strategies they expect their opponent to play. Nash equilibrium predicts that responses in the traveler's dilemma should be independent of $\gamma$, as changing payoffs without changing best responses should have no effect on choice behavior.

Another setting in which Nash equilibrium fails to appropriately describe behavior involves coordination games. These are games with multiple pure strategy Nash equilibria, in which players are incentivized to choose the same strategy. Due to the presence of multiple equilibria, Nash theory cannot make precise predictions. However, human decision makers are often fairly predictable. Consider the Hi-Lo coordination game, in which decision makers have to choose between two strategies: Hi and Lo. In this game we have: $\boldsymbol{u}_{i j}=(1.0,1.0)$ if both players both choose $\mathrm{Hi} ; \boldsymbol{u}_{i j}=(\gamma, \gamma)$, with $0<\gamma<1.0$, if both plays choose Lo; and $\boldsymbol{u}_{i j}=(0,0)$ if they choose different strategies. Not surprisingly, decision makers almost always successfully coordinate on Hi-Hi to obtain the highest possible rewards in this game (Colman, 2003).

In some games, decision makers do not choose any of the Nash equilibrium strategies when the potential costs of miscoordination are too great. This can be observed in the boobytrap game, which is a standard prisoner's dilemma augmented with a third option that allows decision makers to purchase a "boobytrap" to punish their opponent if he or she defects (Misyak \& Chater, 2014). Particularly, we have $\boldsymbol{u}_{i j}=(0.9,0.9)$ if both players cooperate, $\boldsymbol{u}_{i j}=(0.8,0.8)$ if both players defect, and $\boldsymbol{u}_{i j}=(0.89,0.89)$ if both players choose boobytrap. Additionally, $\boldsymbol{u}_{i j}=(0.7,1)$ if player 1 cooperates and player 2 defects, $\boldsymbol{u}_{i j}=(0.9,0.89)$ if player 1 cooperates and player 2 chooses boobytrap, and $\boldsymbol{u}_{i j}=(0,0.69)$ if player 1 defects and player 2 chooses boobytrap (and vice versa, as the game is symmetric). Nash equilibrium predicts that decision makers should ignore the boobytrap choice, however the presence of the boobytrap greatly increases the rate of cooperation in the game, contradicting the prediction of Nash equilibrium.

Yet another set of findings not accounted for by Nash equilibrium theory involves strategy salience. In many games, strategies with salient labels are more likely to be chosen. This is the case in coordination games offering multiple payoff identical strategies, with one of the strategies circled, underlined, or made salient using some other technique. Here players can coordinate successfully by selecting the salient strategy (Mehta et al., 2004).

## Bidirectional Accumulation

We propose an extension to a preexisting accumulator model of risky choice, decision field theory (Busemeyer \&

Townsend, 1993). As in decision field theory, decision makers use two layers of nodes: one to accumulate preferences in favor of the available choice options, and one to represent the probabilistic events involved in the decision. In the strategic context, the choice options are the strategies available to the decision maker and the events are the possible strategies the opponent may use. Thus, the strength of the connection from the node representing a strategy $j$ for the opponent to the node representing preference for a decision maker's strategy $i$, is proportional to the utility of strategy $i$ for the decision maker, given that the opponent plays strategy $j$. Decision makers sample the events according to the subjective probabilities they assign to their occurrence. Thus, strategies that are more likely to be played by the opponent are sampled more frequently and thereby play a larger role in determining the decision makers' preferences.

Decision field theory assumes that decision makers' beliefs about events (and subsequently sampling probabilities for these events) are fixed. For the most part this is reasonable: decision makers' preferences do not influence the actual probability with which different events occur. This assumption is less reasonable in strategic settings. Sophisticated opponents, who can anticipate decision makers' choices, will adjust their own choices to maximize their reward. We thus assume a bidirectional accumulation process to represent strategic deliberation. At each time period, decision makers sample one of their opponent's strategies based on the activations of the nodes corresponding to these strategies, and update their preferences over their own strategies based on this sample. Decision makers then sample one of their own strategies based on the activation of the nodes, and use this sample to update their beliefs about their opponent's choices. In essence, decision makers have dynamically changing mental representations for not only their own preferences, but also their beliefs about their opponents' preferences, allowing them to deliberate intelligently using perspective taking and a sophisticated theory of mind.


Figure 1: Illustration of bidirectional accumulation model


Figure 2: Simulated distribution of choices in the traveler's dilemma.

Formally, if the decision maker has to choose from the set of strategies $S_{1}=\left\{s_{11}, \ldots s_{1 N}\right\}$, then the preference layer in our model consists of $N$ nodes, with node $i$ representing strategy $s_{1 i}$. The activation of node $i$ at time $t, A_{l i}(t)$ corresponds to the decision maker's preference for strategy $i$ at time $t$. Correspondingly if the opponent has the set of available strategies $S_{2}=\left\{s_{21}, \ldots s_{2 M}\right\}$, then the belief layer in our model consists of $M$ nodes, with node $j$ representing strategy $s_{2 j}$. The activation of node $j$ at time $t, A_{2 j}(t)$ corresponds to the beliefs that the decision maker has about the opponent's preference for strategy $j$, at time $t$. We also denote the salience bias of any strategy $i$ (for the decision maker) or $j$ (for the opponent) as $\sigma_{1 i}$ or $\sigma_{2 j}$. These salience biases $\sigma_{1 i}$ and $\sigma_{2 j}$ are independent of the decision process and are determined by various exogenous factors.

At each time period $t$, the decision maker draws one sample of the opponent's strategies. We assume that a softmax (logit) function, with stochasticity parameter $\lambda>0$, determines the effect of activation strength and the exogenous salience bias on sampling probability. Thus, the probability of sampling strategy $j$ at time $t$ is given by: $p_{j}=$ $e^{\lambda\left(A_{2 j}(t-1)+\sigma_{2 j}\right)} / \sum_{k=1}^{M} e^{\lambda\left(A_{2 k}(t-1)+\sigma_{2 k}\right)}$. If the opponent's strategy $j$ is sampled, then the decision maker observes the utility for each strategy $i$ conditional on the opponent playing this sampled strategy: $u_{1}\left(s_{1 i} ; s_{2 j}\right)$. The decision maker's preferences are then updated based on this calculated utility, so the activation for each strategy $i$ becomes: $A_{1 i}(t)=A_{1 i}(t-1)+u_{1}\left(s_{1 i} ; s_{2 j}\right)$.

As discussed, beliefs about the opponent's strategies are themselves updated based on the utility the opponent would derive conditional on a sample of the decision maker's
strategies. Thus, after updating activation states $A_{l i}(t)$, decision makers draw one sample of their own strategies. The probability of sampling strategy $i$ at time $t$ is given by: $q_{i}=e^{\lambda\left(A_{1 i}(t)+\sigma_{1 i}\right)} / \sum_{k=1}^{N} e^{\lambda\left(A_{1 k}(t)+\sigma_{1 k}\right)}$. After sampling strategy $i$, the updated activation for each opponent strategy $j$ is $A_{2 j}(t)=A_{2 j}(t-1)+u_{2}\left(s_{2 j} ; s_{1 i}\right)$.

The deliberation process begins with nodes having no initial activation: $A_{l i}(0)=0$ for all $i ; A_{2 j}(0)=0$ for all $j$. Activation accumulates according to these equations until a time $t=T$. At this time, the most preferred strategy --that is, the one whose node has the highest activation-- is the strategy that is chosen by the decision maker. The parameter $T$ corresponds to an exogenous time limit on the deliberation process, and represents the amount of time taken by the decision makers to make their choices. The proposed model is illustrated in Figure 1.

## Explaining Behavioral Findings

In order to demonstrate how our model works, we use it to simulate choices in the games we introduced earlier. Our simulations use the same strategy and reward profiles as in examples in the previous section. For each of the games and each set of parameter values, we simulate our model 3000 times and report aggregate choice probabilities. We find that the model is fairly robust to parameter variation in the range $\lambda \in[0.25,4]$ and $T \in[10,30]$, and any combination of parameter values in this range produces behavior consistent with the empirical findings we have reviewed. When not explicitly specified, we set salience to $\sigma_{1 i}=\sigma_{2 j}=0$.

Traveler's Dilemma. In the traveler's dilemma our model predicts a failure of unraveling. This is demonstrated in Figure 2 which plots the probability of selecting
strategies in the set $\{20,30, \ldots, 90\}$ for $\gamma=11$ and $\gamma=19$, with varying values of $\lambda$ and $T$. Instead of predicting that players always claim the lowest possible amount, as in Nash equilibrium, here the model generates a distribution of choices that spreads across the range of strategies available to the decision maker. The model also displays payoff sensitivity. For a larger value of the reward/penalty parameter $(\gamma=19)$, the distribution of choices is smaller.

The intuition behind the model's predictions is appealing. For low rewards/penalties, i.e. low values of $\gamma$, the payoffs when both players make high claims are significantly higher than the payoffs when there is a low claim. The potential cost of missing out on this high payoff dwarfs the cost of making a higher claim than the opponent or the benefit of making a lower claim than the opponent. So, a few samples (or even a single sample) of the opponent playing a high claim will lead to high activation for one's own high claims. As beliefs about the opponent's strategy are updated, there will be more samples of high claims, and strategies involving an additional step of undercutting can accumulate the most utility. The number of steps of undercutting that does occur depends on payoff magnitudes. Increasing the parameter $\gamma$ encourages undercutting. Although it does not affect best responses (that is, the ranking of payoffs in any given sample of play), it does affect the accumulation of payoffs over time, so strategies involving more undercutting can accumulate activation more quickly.

Stochastic sampling plays an important role in the emergence of payoff sensitivity. The magnitudes of payoff differences affect the probabilities of sampling each strategy. The degree of responsiveness to the payoff parameter $\gamma$ that we observe in the predicted choices for this game depends on the logit sampling parameter $\lambda$. Comparing across the columns of Figure 2, we see larger shifts in the distribution of choices from a change in the reward/penalty parameter $\gamma$ as the parameter $\lambda$ increases.

Our model also makes new predictions about the relationship between decision time and the strategy chosen in the traveler's dilemma. Each step of undercutting takes time, and thus both the decision maker's preferred claim and the beliefs about the opponent's claim should thus decrease over time. Comparing across the rows of Figure 2, we observe lower claims when the decision time $T$ is larger. Indeed, experiments have revealed that decision makers take longer to choose the lowest claim than the highest claim (Rubinstein, 2007).

Overall, with reasonable parameter values, the model predicts a failure of unraveling. Indeed, full unraveling, consistent with Nash equilibrium would only occur with very large values of $\lambda$ and $T$, i.e., when poorly performing strategies are rarely sampled and there are many periods of sampling and iterative updating. Assuming deterministic sampling of best responses or unlimited decision time would thus lead to poor behavioral predictions for the traveler's dilemma. Conversely, assuming uniformly random sampling would lead to unreasonably high odds of choosing 80 relative to 70 , underestimating people's ability to put
themselves in their opponents' shoes and think strategically about their responses.

The Hi-Lo Game. Although the Hi-Lo game has two Nash equilibria, our model favors the $\mathrm{Hi}-\mathrm{Hi}$ equilibrium. This is shown in Figure 3, which plots the probability of choosing Hi as a function of the payoff for coordinating on Lo $(\gamma)$ for varying values of $T$ and $\lambda$. Across all parameter values we consider, Hi is the modal choice. When the payoff asymmetry is extreme, i.e., $\gamma=0.1$, Hi is almost certain to be chosen. Still, as the Lo-Lo payoff $\gamma$ increases, so does the probability of choosing Lo.

Predictable coordination in the Hi-Lo game is intuitive. The Hi strategy, which offers higher payoffs in the case of successful coordination, accumulates more activation when it is sampled from the other layer of the network than the low strategy does. This creates a feedback effect, so the model is more likely to think about Hi when forming beliefs about the opponent's choices. Believing that the opponent will choose Hi further reinforces the model's preference Hi .

The positive feedback loop, along with stochastic sampling, actually facilitates the occasional choices of Lo. If Lo is sampled first, it gains an advantage, and it becomes more likely to be sampled again. As the logit sampling parameter $\lambda$ increases, it becomes somewhat more likely (albeit still not very likely) that the model repeatedly samples Lo early on, gets fixated on this strategy, and eventually chooses it. In the extreme case that the sampling parameter $\lambda$ gets unrealistically large, the strategy sampled in the first time period may be sampled forever thereafter, completely determining the path of the deliberation. Since both strategies have the same probability of being sampled in the first period of the deliberation, the model's choice distribution approaches a 50-50 split between Hi and Lo independent of $\gamma$ for very large values of $\lambda$. As can be seen, decision time has little effect on the choice distribution, with longer deliberation only slightly reducing noise and increasing the probability of selecting the modal choice, Hi .

The Boobytrap Game. Our model deviates far from Nash equilibrium in the boobytrap game as well. For $\lambda \in$ $[0.25,4]$ and $T \in[10,30]$, it predicts that players will almost certainly cooperate (cooperation with a greater than $90 \%$ chance for all parameters). Here, a non-Nash strategy is favored due to the high magnitude of its advantage when the other player does not best respond compared to the low magnitude of its cost when the other player does respond rationally. Against the boobytrap strategy, defection is extremely undesirable. The model predicts that players will never choose the boobytrap strategy, because it is dominated by cooperation. However, the model predicts that decision makers usually will contemplate this boobytrap strategy as part of their deliberation, and this causes their preferences for defection to drop strongly.

Again, our model's behavior would be very different with an assumption of deterministic sampling of the most highly activated strategy. With deterministic sampling, the model is confident that the boobytrap strategy will not be played, so it chooses to defect.


Figure 3: Simulated probability of choosing Hi in the Hi-Lo game.


Figure 4: Simulated probability of heads in the simple heads-or-tails coordination game.

Salient Labels. Our model recognizes salience effects, too. In the simple heads-or-tails coordination game with heads being especially salient, such that $\sigma_{1 \mathrm{H}}=\sigma_{2 \mathrm{H}}=\varsigma$ and $\sigma_{1 \mathrm{~T}}=\sigma_{2 \mathrm{~T}}=0$, we find that the probability of choosing heads is increasing in its salience $\varsigma$, as shown in Figure 4. This figure plots the probability of choosing heads in this game as a function of the salience of heads, $\varsigma$, for varying values of $T$ and $\lambda$. As we should intuitively expect, when sampling is less noisy, i.e., when $\lambda$ is greater, the players are more sensitive to salience. Specifically, when near the high end of our range, i.e., $\lambda=4$, if heads is sufficiently salient, it is almost certain to be chosen. (In contrast, with an assumption of uniformly random sampling, our model would not account for any salience effect at all.) Convergence occurs quickly, so we see few effects from increasing the decision time $T$. Higher values of $T$ only slightly reduce noise and increase the choice probability of heads when the logit sampling parameter $\lambda$ is small.

## Discussion

We have proposed a cognitive model of strategic deliberation and decision making. Our model is able to account for violations of Nash equilibrium involving failures of unravelling, payoff sensitivity, predictable coordination, and salience, and we illustrate this by simulating our model on four different games. Note that these violations have also been documented in a number of additional games, including the minimum-effort coordination game, the stag hunt game, the battle of sexes game, the discoordination game, the 11-20 game, the hide and seek game, the matching pennies game, and the Kreps game. Elsewhere we show that our model makes realistic behavioral predictions for all of these games, for $\lambda \in$
$[0.25,4]$ and $T \in[10,30]$, however we exclude these findings from this paper, due to space constraints.

Our model is closely related to existing accumulator theories choice, and we suggest that it can be seen as a direct extension of decision field theory (Busemeyer \& Townsend, 1993; also see Busemeyer, 2015 for a review). The novel element in our model involves the representation of beliefs regarding opponent's choices and the bidirectional updating of both preferences and beliefs over the time course of the decision process. Intuitively, bidirectional feedback in the accumulation process allows decision makers to base their choices on their beliefs about the opponent's choices, but also to update their beliefs as their own preferences evolve. As this updating happens gradually over time, the decision makers' intended choices (and beliefs about the opponent's choices) get increasingly more sophisticated the longer they spend deliberating. Eventually the nodes for the opponent's strategies develop unequal activation, with strategies that are appropriate responses to the decision maker's preferences having higher activation. Highly activated opponent strategies are more likely to be sampled, and the decision maker is subsequently more likely to develop preferences that intelligently respond to the opponent's anticipated choices.

Note that there is considerable evidence that decision makers are able to represent the preferences and beliefs of others separately from their own. Although the nature of these representations is not typically studied in the context of game theoretic deliberation, some experimental work on theory of mind in strategic games does support our proposed model. Hedden and Zhang (2002), for example, find that players in sequential move games have sophisticated beliefs about the opponent's preferences, and that these beliefs are
dynamically modified based on the evidence presented to the decision maker during the decision process. Goodie et al. (2012) also find that players' beliefs about their opponent's preferences are fairly complex, and are formed in response to the players' own preferences.

Our approach is also closely related to cognitive decision modeling (in non-strategic settings) that uses neural networks with recurrent connectivity (Glöckner et al., 2014; Holyoak \& Simon, 1999). Recurrence in these networks is often bidirectional; the activation of cues and decision attributes may influence and be influenced by beliefs and preferences. The bidirectional feedback in the above models and in ours is very similar, implying that our model could be adapted for other cognitive decision modeling applications.

Our bidirectional accumulation model also bears some resemblance to models of behavioral game theory, such as level-k reasoning and logit quantal response equilibrium (McKelvey \& Palfrey, 1995; Nagel, 1995). In both our model and in level-k reasoning, individuals engage in an iterative process of deliberation that terminates before reaching a point of self-consistency. Likewise, in both our model and in logit quantal response equilibrium, individuals use a stochastic logit response rule to consider responses, thereby generating payoff sensitivity. However, unlike these models, our approach implements the deliberation process within a well-established psychological framework. This allows our model to describe salience effects, while also predicting the effects of time pressure and response time. Our model also makes more realistic stochastic choice predictions than either of these two existing theories: It permits trial-to-trial variability in choice, while also avoiding the selection of dominated strategies.

Our approach is also quite parsimonious. There are two parameters in our model: the decision time parameter, $T$, and the stochastic sampling parameter, $\lambda$. Decision time $T$ can be seen as controlling the extent of bidirectional processing one can engage in during deliberation and thus determining one's level of strategic sophistication. Quick decisions involve fairly limited reasoning, with choices responding to simplistic beliefs about the opponent. Decisions that are a product of extended deliberation, in contrast, generate choices based on a more sophisticated theory of mind. As in all accumulator models, decision time also influences the amount of variability in the decision.

The stochastic sampling parameter $\lambda$ can also be seen as affecting the extent of bidirectional processing one engages in. When $\lambda$ is small, strategies are sampled with close to uniform probability, and activation in one layer of the network has little or no effect on the accumulation of activation in the other layer of the network. As $\lambda$ increases, the decision maker becomes more and more likely to sample the most preferred strategies. When $\lambda$ is very large, the most highly activated strategies are almost deterministically sampled, so preferences and beliefs interact more strongly during the deliberation.

Ultimately, the model's key behavioral properties depend critically on its dynamic and stochastic processes. Many
scholars have suggested that behavioral theories of decision making can, with incorporation of these fundamental cognitive processes, describe a wide range of behavior (e.g. Busemeyer, 2015). Our results reinforce these claims by demonstrating the explanatory power of stochastic sampling and dynamic accumulation in strategic choice.

## References

Basu, K.. (1994). The Traveler's Dilemma: Paradoxes of Rationality in Game Theory. American Economic Review, 84(2), 391-395.
Bhatia, S. (2014). Sequential sampling and paradoxes of risky choice. Psychonomic Bulletin and Review, 21(5), 1095-1111.
Busemeyer, J. R., \& Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review, 100(3), 432-448.
Busemeyer, J. R. (2015). Cognitive science contributions to decision science. Cognition, 135, 43-46.
Camerer, C.(2003). Behavioral game theory: Experiments in strategic interaction. Princeton University Press.
Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. Behavioral and Brain Sciences, 26(2), 139-198.
Glöckner, A., Hilbig, B. E., \& Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. Cognition, 133(3), 641666.

Goodie, A. S., Doshi, P., \& Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. Journal of Behavioral Decision Making, 25(1), 95-108.
Hedden, T., \& Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. Cognition, 85(1), 1-36.
Holyoak, K. J., \& Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. Journal of Experimental Psychology: General, 128(1), 3-18.
Luce, R. D., \& Raiffa, H. (1957). Games and Decisions. New York: John Wiley Sons.
Mehta, J., Starmer, C., \& Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. The American Economic Review, 84(3), 658-673.
Misyak, J. B., \& Chater, N. (2014). Virtual bargaining: a theory of social decision-making. Philosophical Transactions of the Royal Society: B, 369(1655)..
Nagel, R. (1995). Unraveling in guessing games: An experimental study. The American Economic Review, 85(5), 1313-1326.
Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(6), 1355.
Rubinstein, A. (2007). Instinctive and cognitive reasoning: a study of response times. The Economic Journal, 117(523), 1243-1259.

# How Can Memory-Augmented Neural Networks Pass a False-Belief Task? 

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths<br>University of California, Berkeley<br>\{eringrant, nematzadeh, tom_griffiths\} @berkeley.edu


#### Abstract

A question-answering system needs to be able to reason about unobserved causes in order to answer questions of the sort that people face in everyday conversations. Recent neural network models that incorporate explicit memory and attention mechanisms have taken steps towards this capability. However, these models have not been tested in scenarios for which reasoning about the unobservable mental states of other agents is necessary to answer a question. We propose a new set of tasks inspired by the well-known false-belief test to examine how a recent question-answering model performs in situations that require reasoning about latent mental states. We find that the model is only successful when the training and test data bear substantial similarity, as it memorizes how to answer specific questions and cannot reason about the causal relationship between actions and latent mental states. We introduce an extension to the model that explicitly simulates the mental representations of different participants in a reasoning task, and show that this capacity increases the model's performance on our theory of mind test.


Keywords: language understanding, question answering, theory of mind, false-belief test

## Introduction

Question answering poses difficulties to artificial intelligence systems because correctly answering a query often requires sophisticated reasoning and language understanding capacities, and so simply memorizing the answer or searching in a knowledge base is not enough. Despite this challenge, recent neural network models that make use of attention mechanisms in combination with an explicit external memory can successfully answer questions that require more complex forms of reasoning than before (e.g., Sukhbaatar, Weston, Fergus, et al., 2015; Henaff, Weston, Szlam, Bordes, \& LeCun, 2017). The benchmark dataset for such tasks has become the Facebook bAbi dataset (henceforth, bAbi) (Weston, Bordes, Chopra, \& Mikolov, 2016), which is a collection of question-answering tasks in the form of simple narrative episodes - termed stories - that are accompanied by questions about the state of the world described in the stories. (See Figure 1 for an example story from this dataset.)

Although bAbi is a start towards enumerating the requirements for human-like reasoning capabilities, it lacks tasks for testing the ability to reason about mental states, which is also necessary for correctly answering questions of the sort that humans encounter regularly. Consider the following:

> Sally and Ann are in the kitchen.
> Sally placed the milk in the pantry.
> Sally exited the kitchen.
> Ann moved the milk to the fridge.

For a model to correctly answer questions such as Where would Sally/Ann search for the milk? it need not only recognize that Sally and Ann have mental representations of the
state of the world but also that these representations are inconsistent: Sally believes that the milk is in the pantry while Ann thinks it is in the fridge.

Psychologists have used a task similar to this scenario termed the false-belief task - to examine children's development of theory of mind: the capacity to reason about the mental states of oneself and others (Premack \& Woodruff, 1978). Most 3-year-old children, after observing such a scenario, answer that Sally would search for the milk in the fridge because they cannot infer Sally's belief about the location of the milk, which is inconsistent with their own knowledge (e.g., BaronCohen, 1989; Baron-Cohen, Leslie, \& Frith, 1985). However, most older children are able to identify, correctly, that Sally's belief is different from theirs in that she thinks that the milk is the pantry.

To answer questions about situations like those that occur in a false-belief task, a model needs to use the observed actions in the scenario to infer the mental states of Sally and Ann. In this work, we investigate whether the End-toEnd Memory Network (henceforth MemN2N), a recent neural question-answering model (Sukhbaatar et al., 2015) that solves most of the bAbi tasks, is able to answer questions of the same structure as a false-belief task. We formulate scenarios to capture different possible causal relations among actions and beliefs, and examine the performance of the model therein. We find that the MemN2N model performs well only in the presence of strong supervision - when the training and test data share the same casual structure. This result suggests that the model is able to memorize the training data but is unable to learn to reason about mental states and how they cause and are caused by actions.

Furthermore, to simulate the (perhaps inconsistent) beliefs of the participants in a story, we extend the MemN2N model to include a separate memory representation for each participant. We show that this extension improves model performance, suggesting that explicitly modeling agents' knowledge in a disentangled manner is in part sufficient for more human-like reasoning on a false-belief task.

## Theory of Mind and the False-Belief Task

A theory of mind is integral for an agent to predict and explain the behavior that is caused by the mental representations of other agents, and therefore succeed on tasks such as the false-belief task. For children, this capacity is acquired gradually over the course of development. In particular, children undergo several milestones before they develop an adult-like theory of mind: By age two, they can distinguish between external states of the world and internal mental states possessed by cognitive agents (e.g., Meltzoff, Gopnik, \& Repacholi,

Mary got the milk there.
John moved to the bedroom.
Sandra went back to the kitchen.
Mary traveled to the hallway.
Q: Where is the milk? A: hallway
Figure 1: An example task from the bAbi dataset.
1999). By age four, they can distinguish between consistent and inconsistent mental states (e.g., Perner, Leekam, \& Wimmer, 1987), which allows them to identify a false belief.

Previous computational works have modeled human performance on the false-belief task. Some focus on modeling the development of theory of mind by instantiating a model that initially fails but eventually passes the false-belief test (Van Overwalle, 2010), while others study the settings in which a model can succeed on the task by varying the input data or the model architecture (O'Laughlin \& Thagard, 2000; Triona, Masnick, \& Morris, 2002; Goodman et al., 2006). However, none of these models use natural language sentences, despite the fact that the psychological false-belief task is usually administered verbally in the form of a natural language reasoning problem.

Furthermore, natural language is known to interact with the development of theory of mind. For example, use of mental state terms in child-directed speech (e.g., Slaughter \& Gopnik, 1996), engagement in pretend play (Youngblade \& Dunn, 1995), storybook reading (Rosnay \& Hughes, 2006), and reflection on events in the child's past (Nelson, 2007) serve to accelerate its developments, while, in turn, a greater grasp of theory of mind leads to increased linguistic ability (Milligan, Astington, \& Dack, 2007). In this work, we examine whether a model can learn from natural language about the causal relationship between actions and beliefs, in order to be able to answer questions that require reasoning about mental states.

## Memory Networks

The MemN2N model of Sukhbaatar et al. (2015) comprises an external memory cache and mechanisms to read and write to this memory. The model is trained to write a sequence of stories into its external memory and to answer questions about the stories by reading its memory and emitting the correct vocabulary item. At test time, the model is evaluated by the extent to which it can correctly answer questions about a held-out set of test stories.

Formally, the model ingests a sequence of input sentences $\left(x_{1}, \ldots, x_{n}\right)$ and produces, for each input item $x_{i}$, both a memory representation $m_{i}$ and a context representation $c_{i}$, which are stored in memory. The model is then presented with a question $q_{k}$ about the story, for which it produces an internal representation $u_{k}$. To answer the question, the model computes a normalized association score $p_{i k}$ between the question representation and each of its stored memory representations:

$$
\begin{equation*}
p_{i k}=\frac{\exp \left\{u_{k}^{\mathrm{T}} m_{i}\right\}}{\sum_{j} \exp \left\{u_{k}^{\mathrm{T}} m_{j}\right\}} \tag{1}
\end{equation*}
$$

This weight can be interpreted as an attention mechanism that defines where in memory the model will look for information relevant to the given question.

The model then produces an output representation by way of a linear combination of its context representations, weighted by the attention computed in Equation (1):

$$
\begin{equation*}
o_{k}=\sum_{i} p_{i k} c_{i} \tag{2}
\end{equation*}
$$

The output representation is combined with the query representation and decoded by some function $f$ to produce the predicted answer $\hat{a}$ :

$$
\begin{equation*}
\hat{a}=f\left(o_{k}+u_{k}\right) . \tag{3}
\end{equation*}
$$

Learning model parameters at training time is done by way of stochastic gradient descent in cross entropy error.

## Simulation 1: MemN2N Model

We evaluate the model introduced in the previous section on a set of novel textual reasoning tasks inspired by the falsebelief task. Our tasks take the form of a sequence of natural language sentences - termed a story - and an associated question about the story.

Since we aim to create tasks that, for humans to solve, involve reasoning about other agents' beliefs, we design various story templates that simulate how different actions give rise to different beliefs, and conversely how different beliefs result in different actions. These stories differ in whether or not the agent who is the subject of the question has observed a change in the state of the world (i.e., the agent has a true belief), or has not (i.e., has a false belief). The stories further differ in whether the belief is observable (i.e., the story explicitly contains sentences such as Sally believes the milk is in the pantry) or whether only actions are observable. When the agent harbors a false belief, and the model is asked to predict the action of the agent without explicit reference to the beliefs of the agent in the story, we recover a simulation of the classic false-belief task.

With this experimental design, we aim to determine whether the MemN2N model can reason about how actions cause beliefs and vice versa, and how much information needs to be revealed to enable the model to succeed.

## Data Generation

To generate stories and corresponding questions, we emulate the bAbi (Weston et al., 2016) dataset generation procedure. We define a world of entities, which are the people and objects described in the stories, and possible predicates that take entities as subject and, optionally, object. Each entity has properties that define the predicates of which it can be subject or object. For example, a world may contain Sally with

## BA

Anne moved the milk to the fridge.
True Sally believes the milk is in the fridge.
Belief Q: Where did Sally search for the milk?
A: fridge
Sally believes the milk is in the pantry.
Sally exited the kitchen.
False Anne moved the milk to the fridge.
Belief Sally entered the kitchen.
Q: Where did Sally search for the milk?
A: pantry

Sally placed the milk in the pantry.
Anne moved the milk to the fridge.
Q: Where does Sally believe the milk is?
A: fridge
Sally placed the milk in the pantry.
Sally exited the kitchen.
Anne moved the milk to the fridge.
Sally entered the kitchen.
Q: Where does Sally believe the milk is?
A: pantry

Sally placed the milk in the pantry. Anne moved the milk to the fridge.
Q: Where did Sally search for the milk? A: fridge

Sally placed the milk in the pantry. Sally exited the kitchen. Anne moved the milk to the fridge. Sally entered the kitchen. Q: Where did Sally search for the milk? A: pantry

Figure 2: Examples of the training data, with the predicates of interest underlined. Note that the true-belief and false-belief test tasks are of the same form as the top and bottom items, respectively, in the last column.
the property is agent and apple with the property is object. Our rules permit Sally to perform the action displace on the apple.

In this work, we consider a restricted set of action and belief predicates. Our actions define simple interactions of an agent with the world (e.g., place, move, enter, exit) and our beliefs correspond to mental state terms (e.g., believe, think), inspired by the terms that children gradually learn to understand and use correctly over the course of development (e.g., Bretherton \& Beeghly, 1982; Johnson \& Wellman, 1980). Our templates manipulate the order of action and belief predicates to test how the model reasons about the causal relations between them.

## Experimental Conditions

Story Template We define a set of templates that correspond to the type of story that we wish to generate. Each template fixes a sequence of predicates and therefore puts constraints on the entities that may fill the template. For example, a template could be the sequence (drop, pick up, exit). Completion of the template entails sampling valid entities from the world to fill the subject and object positions of the predicates, producing, for example, the story (Sally dropped the ball, Sally picked up the ball, Sally exited the room).

We consider three different template types:

- BA: observable beliefs (e.g., Sally believes the milk is in the pantry) give rise to observable actions (e.g., Sally searches the pantry);
- AB: observable actions (e.g., Sally places the milk in the pantry) give rise to observable beliefs (e.g., Sally believes milk is in the pantry); and
- $\mathbf{A}(\mathbf{B}) \mathbf{A}$ : observable actions (e.g., Sally places the milk in the pantry) give rise to observable actions (e.g., Sally searches the pantry) by way of unobserved beliefs (e.g., Sally believes the milk is in the pantry).

Note that the $\mathbf{A B}$ and $\mathbf{A}(\mathbf{B}) \mathbf{A}$ conditions are different in that in $\mathbf{A B}$, the question explicitly asks about Sally's belief; in
$\mathbf{A}(\mathbf{B}) \mathbf{A}$, on the other hand, the question is about Sally's action, which has been brought about by Sally's unobserved belief.
True vs. False Belief In addition to the type of template, for each story we manipulate whether the agent about whom the question is asked (i.e., Sally) has a true belief or a false belief about the state of the world. In the case that the agent has a true belief, the agent observes all changes in the state of the world and thus their beliefs are consistent with the world. On the other hand, in the case that the agent has a false belief, the agent does not observe one or more changes in the state of the world (because, for instance, Sally may exit the room), and thus has a belief that is inconsistent with the world.
Training Conditions We have six possible story types as a results of crossing the template types with the true and false belief story types; examples of each of the story types are given in Figure 2. We sample from these story types to produce our training conditions, in the following manner:

- When the training condition is such that $p($ false belief $)=$ 0 or 1 , we sample only stories with true beliefs or false beliefs, respectively, and when $p($ false belief $)=0.5$, we sample half of our stories with true beliefs and half with false beliefs.
- We sample stories from five different possible groups of templates: $\mathbf{B A}, \mathbf{A B}, \mathbf{A B}+\mathbf{B A}, \mathbf{A}(\mathbf{B}) \mathbf{A}$ and $\mathbf{A B}+\mathbf{B A}+\mathbf{A}(\mathbf{B}) \mathbf{A}$.

The $\mathbf{A B}+\mathbf{B A}$ and $\mathbf{A B}+\mathbf{B} \mathbf{A}+\mathbf{A}(\mathbf{B}) \mathbf{A}$ conditions provide the model with training data that better approximates the variety of possible scenarios in the world. In these cases, the model observes more ways in which actions and beliefs interact, and thus we would expect it to be able to better generalize to new scenarios. Moreover, $\mathbf{A B}+\mathbf{B A}$ provides the model with the opportunity to learn transitive inference - given that an action (e.g., placing milk in the pantry) results in a belief (e.g., the milk is in the pantry), and a belief (e.g., the milk is in the pantry) can cause an action (e.g., searching for milk in the pantry), a model that reasons about actions and beliefs could learn that an action (e.g., searching for milk in the pantry) is


Figure 3: Accuracy in Simulation 1. Test accuracies for the true-belief (TB) and false-belief (FB) tests across training conditions in Simulation 1. We report results for $p($ false belief $)=0.5$, since varying this parameter did not affect results except in the few cases discussed in the text.


Figure 4: Accuracy in Simulation 2. Test accuracies for the true-belief (TB) and false-belief (FB) tests across training conditions in Simulation 2. As in Figure 3, we report results only for $p($ false belief $)=0.5$.


Figure 5: Attention in Simuation 2. Visualisation of the attention weighting over memory caches for the true-belief (TB) and false-belief (FB) tests. We omit the visualization for the $\mathbf{B A}+\mathbf{A B}$ and $\mathbf{B A}+\mathbf{A B}+\mathbf{A}(\mathbf{B}) \mathbf{A}$ training conditions because the test accuracy distribution in Simulation 2 for these conditions is very similar to the $\mathbf{A}(\mathbf{B}) \mathbf{A}$ training condition (see Figure 4).
a consequence of an unobservable belief brought about by a preceding action (e.g., placing milk in the pantry).

Crossing template types $\mathbf{B A}, \mathbf{A B}, \mathbf{A}(\mathbf{B}) \mathbf{A}, \mathbf{A B}+\mathbf{B A}$, $\mathbf{A B}+\mathbf{B A}+\mathbf{A}(\mathbf{B}) \mathbf{A}$ with $p($ false belief $)=\{0.0,0.5,1.0\}$ produces our 15 training conditions. We run 10 simulations for each training condition and for each configuration of parameter settings of the MemN2N model. ${ }^{1}$
Test Conditions We aim to evaluate the model on tasks that require reasoning about latent mental states, in analogy to the classic false-belief task; however, such a capacity should apply not only in cases when an agent has a belief that is inconsistent with the state of the world (i.e., a false belief) but also when they have a true belief about the world. We therefore consider two test conditions: a true-belief (TB) and a falsebelief (FB) task. All examples in both of these test conditions share the $\mathbf{A}(\mathbf{B}) \mathbf{A}$ template type, but the conditions differ in that the true-belief task contains only examples with true beliefs (i.e., $p($ false belief $)=0$ ), and the false-belief task contains only false belief examples (i.e., $p($ false belief $)=1$ ).

[^76]
## Results

As noted by Sukhbaatar et al. (2015), the MemN2N model exhibits large variance in performance across simulations, and so we show performance by plotting the distribution of test accuracies in boxplot format. In Figure 3, we report accuracy on both test conditions (the true-belief (TB) and falsebelief (FB) tasks) across the training conditions, for $p$ (false belief) $=0.5$. The results for $p($ false belief $) \in\{0,1\}$ were similar except in the case of the $\mathbf{A B}$ story template; we compare this case with the BA condition in Figure 6 and discuss in the following. Note that success at test time corresponds to achieving 1.0 accuracy in both the TB and FB test conditions.
Training Condition BA: Beliefs to Actions The model fails on the TB task in the BA training condition, while succeeding on the FB task. This is true no matter the value of $p$ (false belief) (as depicted in Figure 6). To understand why this occurs, consider the following example of a BA training story when the false belief occurs:
Sally believes the milk is in the pantry. Sally exited the kitchen. Anne moved the milk to the fridge. Sally entered the kitchen.

Additionally, consider the BA training story when the false belief does not occur:

Anne moved the milk to the fridge. Sally believes the milk is in the fridge.

To answer the training question Where did Sally search for the milk? the model seems to learn that it should look for the sentence containing Sally and a container entity (i.e., Sally believes the milk is in the fridge).

This strategy works for the false-belief test (see Figure 2, last column, bottom row), because Sally believes that the milk is in the pantry - the location in which she originally placed it - and thus the sentence containing Sally and the identity of a container always proviedes the correct answer. However, this strategy fails on the true-belief test (again, see Figure 2, last column, top row), because Sally observes that the milk has been moved, and so no longer believes that the milk is in fridge. This suggests that the model is unable to infer that an observable action changes the mental state of Sally.
Training Condition AB: Actions to Beliefs The model is unable to achieve good performance on both the TB and FB tests in the AB condition. When the model performs better, it is in cases where the test is very similar to the training condition, i.e., the false-belief test with $p($ false belief $)=1$ in training and true-belief test with $p$ (false belief) $=0$ in training.
Training Condition AB+BA: Transitive Inference The model fails on both test tasks in the $\mathbf{A B}+\mathbf{B A}$ training condition. This is evidence that the model cannot reason about the causal relationships between actions and beliefs to perform transitive inference.
Training Condition $\mathbf{A}(\mathbf{B}) \mathbf{A}$ : Equivalent to TB/FB Test The model achieves best performance on $\mathbf{A}(\mathbf{B}) \mathbf{A}$ in the $p$ (false belief) $=0.5$ condition. This again happens because the test and training conditions are similar: the model observes examples of both the FB and TB test tasks in training, and thus receives supervision to give the correct answer at test. However, the model performs well only on the TB task in the $p$ (false belief $)=0$, and on the FB task in the $p($ false belief $)=1$ condition. This is because the model does not observe examples like one or the other test condition at training time.

Notably, the performance is not high even in the $p$ (false belief) $=0.5$ condition (the median is approximately $55 \%$ on both test tasks), despite the fact that the model is given testlike examples at training time. It is therefore not clear that the model is robustly able to solve a conditional reasoning task in which the correct answer is dependent on whether or not the observer sees the movement of the object and thus has a false or true belief. This, along with the model's failure in the other training scenarios, motivates an extension to the model, which we consider in the next section.

## Simulation 2: Multiple-Observer Model

We now propose a model that is given information about whether each agent in the story observes each sentence in the story. In general, this must also be inferred from context, but here we assume such annotations are available to the model as we simply attempt to investigate the effect of this information on the model's predictions.
(a) $\boldsymbol{p}($ false belief $)=0$ in training.


Figure 6: From Simulation 1. The test accuracy in the AB condition is dependent on the value of $p$ (false belief), but not in the BA condition.

Formally, for a story of $N$ input items that describes a situation with $M$ agents, we provide the model with an $N$-by$(M+1)$ observer annotation matrix $S$ such that $S_{i j}=1$ if input item $x_{i}$ is observable to agent $j$ and 0 otherwise, where we assign the oracle observer (who observes all input items) to the first index. These annotations are used to mask the input such that $M+1$ (possibly different) stories are produced, each of which corresponds to the story that a particular agent observes. Memory representations, attention over each memory cache, and output representations are computed separately for each observer, and so $M+1$ output representations are computed, each corresponding to the output of a distinct observer's memory.

The model then computes an attention weighting over each of the observer memory caches ( $c f$. Equation (1)):

$$
\begin{equation*}
r_{k \ell}=\frac{\exp \left\{u_{k}^{\mathrm{T}} o_{k \ell}\right\}}{\sum_{n} \exp \left\{u_{k}^{\mathrm{T}} o_{k n}\right\}} . \tag{4}
\end{equation*}
$$

This attention over memory caches is used to compute a weighted combination of the output representations that correspond to the memory cache for each agent (cf. Equation (3)):

$$
\begin{equation*}
\hat{a}=f\left(u_{k}+\sum_{\ell} r_{k \ell} o_{k \ell}\right) . \tag{5}
\end{equation*}
$$

Note that the model considered in Simulation 1 is exactly this model extension with $r_{k 0}=1$ and $r_{k m}=0, \forall m \neq 0$ (i.e., attention is given only to the oracle memory cache).

In this extension, the model is given explicit information about which observations in a story are available to each agent, by way of the annotation matrix $S$. However, it must learn to reason about this information in order to arrive at the correct answer, as before with how to write to memory and read from memory, and now with how to select over which observer's knowledge of the story is relevant to answer the question.

## Results

We report results of the model extension on the TB and FB tests in Figure 4, as well as a visualization of the attention weights in Figure 5. Our simulated data is composed of scenarios with only two agents, and therefore the extended model attends over three memory caches (one for the oracle that observes everything, one for Anne, and one for Sally, about whom the question is asked).

The extended model achieves higher accuracy across all training conditions. Notably, the model performs near perfectly (i.e., both TB and FB are close to 1 ) in the $\mathbf{A B}+\mathbf{B A}+\mathbf{A}(\mathbf{B}) \mathbf{A}$ case, meaning that the model can learn to ignore irrelevant training stimuli. This suggests that awareness of agent's knowledge about the state of the world helps in a task of reasoning about latent mental states.

Furthermore, the attention plots show that the model learns to attend to the memory representation of Sally in the FB test, which contains the information about how to answer questions about Sally's actions and beliefs. On the other hand, in the TB test, the model does not attend differently to the different memory caches, because the observations stored in all caches are the same.

## Conclusions

We investigated whether a recent language learning model that succeeds on a suite of textual reasoning tasks is able to succeed in a task that requires reasoning about latent mental states. We found that the model is unable to succeed in a set of simulated true-belief and false-belief tasks unless it has observed at training time situations that have the same structure as the test tasks, even if the diversity of the data is increased. This strongly suggests that the model is not reasoning about the state of the world, nor about mental representations thereof, but is simply memorizing its input. As a consequence, the model will not be able to succeed in a task of reasoning that differs greatly from the situations that it has observed at training time. This is in contrast to the the novelty of situations that people encounter regularly, in which they must reason about the causal relationship between events in the world and latent mental states.

However, incorporating a simple mechanism that informs the model that there may be multiple observers with differing representations of the story allows the model to achieve higher performance on the simulated false-belief and truebelief tasks. Under this modification, the model does not simply memorize the training data but also learns to use
knowledge that agents have (perhaps conflicting) observations about the story in order to answer the question. We could interpret this as analogous to the development of theory of mind in that, when a child is able to reason about others' knowledge of and beliefs about the world, the child succeeds on tests of theory of mind such as the false-belief task. A further direction of research could investigate whether manipulating variables in the training data (e.g., frequency of mental state terms) affects the model's performance in a manner similar to how a child's developmental trajectory would be affected.

## References

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. J. of Child Psychology \& Psychiatry, 30(2), 285-297.
Baron-Cohen, S., Leslie, A. M., \& Frith, U. (1985). Does the autistic child have a theory of mind? Cognition, 21(1), 37-46.
Bretherton, I., \& Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. Developmental Psychology, 18(6), 906.
Goodman, N. D., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In Proceedings of Cog. Sci. (pp. 13821387).

Henaff, M., Weston, J., Szlam, A., Bordes, A., \& LeCun, Y. (2017). Tracking the world state with recurrent entity networks. In Proceedings of ICLR.
Johnson, C. N., \& Wellman, H. M. (1980). Children's developing understanding of mental verbs: Remember, know, and guess. Child Development, 1095-1102.
Meltzoff, A. N., Gopnik, A., \& Repacholi, B. M. (1999). Toddlers' understanding of intentions, desires and emotions: Explorations of the dark ages. In P. Zelazo, J. Astington, \& D. Olson (Eds.), Developing theories of intention: Social understanding and self control (pp. 17-41). Erlbaum.
Milligan, K., Astington, J. W., \& Dack, L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. Child develop., 78(2), 622646.

Nelson, K. (2007). Young minds in social worlds: Experience, meaning, and memory. Harvard University Press.
O'Laughlin, C., \& Thagard, P. (2000). Autism and coherence: A computational model. Mind \& Language, 15(4), 375-392.
Perner, J., Leekam, S. R., \& Wimmer, H. (1987). Three-yearolds' difficulty with false belief: The case for a conceptual deficit. British Journal of Developmental Psychology, 5(2), 125-137.
Premack, D., \& Woodruff, G. (1978, Dec). Does the chimpanzee have a theory of mind? Behavioral and Brain Sci., 1(4), 515-526.
Rosnay, M., \& Hughes, C. (2006). Conversation and theory of mind: Do children talk their way to socio-cognitive understanding? British Journal of Developmental Psychology, 24(1), 7-37.
Slaughter, V., \& Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. Child development, 67(6), 2967-2988.
Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In Proceedings of NIPS (pp. 2440-2448).
Triona, L. M., Masnick, A. M., \& Morris, B. J. (2002). What does it take to pass the false belief task? An ACT-R model. In Proceedings of Cog. Sci. (p. 1045).
Van Overwalle, F. (2010). Infants' teleological and belief inference: A recurrent connectionist approach to their minimal representational and computational requirements. NeuroImage, 52(3), 1095-1108.
Weston, J., Bordes, A., Chopra, S., \& Mikolov, T. (2016). Towards AI-complete question answering: A set of prerequisite toy tasks. In ICLR.
Youngblade, L. M., \& Dunn, J. (1995). Individual differences in young children's pretend play with mother and sibling: Links to relationships and understanding of other people's feelings and beliefs. Child Development, 66(5), 1472-1492.

# Evidence for a facilitatory effect of multi-word units on child word learning 

Robert Grimm, Giovanni Cassani, Walter Daelemans and Steven Gillis<br>Computational Linguistics and Psycholinguistics (CLiPS) Research Center Department of Linguistics, University of Antwerp, 13 Prinsstraat<br>B-2000 Antwerpen, Belgium<br>\{name.surname\}@uantwerpen.be


#### Abstract

Previous studies have suggested that children possess cognitive representations of multi-word units (MWUs) and that MWUs can facilitate the acquisition of smaller units contained within them. We propose that the formation of MWU representations precedes and facilitates the formation of single-word representations in children. Using different computational methods, we extract MWUs from two large corpora of English childdirected speech. In subsequent regression analyses, we use age of first production of individual words as the dependent and the number of MWUs within which each word appears as an independent variable. We find that early-learned words appear within many MWUs - an effect which is neither reducible to frequency or other common co-variates, nor to the number of context words contained in the MWUs. Our findings support accounts wherein children acquire linguistic patterns of varying sizes, moving gradually from the discovery of MWUs to the acquisition of small-grained linguistic representations. ${ }^{1}$


Keywords: multi-word units; age of first production; word learning; language acquisition; computational modeling

## Introduction

Frequently co-occurring word combinations have been investigated in studies examining both child (Bannard \& Matthews, 2008; Arnon \& Clark, 2011; McCauley \& Christiansen, 2014) and adult processing (Arnon \& Snider, 2010), with mounting evidence that children and adults represent such sequences separately from their constituent words. Indeed, given that many English word sequences have idiosyncratic meanings which cannot be derived from the meaning of their constituent words (e.g. pay attention to, leave of absence, you're welcome), it is reasonable to expect language users to store such semantically opaque sequences in memory. Findings from the literature, however, extend beyond this: in addition to non-compositional constructions, people are likely to also lexicalize frequent but semantically transparent formulaic sequences (Wray, 2008). Here, we use the term multi-word unit (MWU) to refer to any sequence of words - compositional or not - which is likely to be lexicalized, and we investigate the role of MWUs in the acquisition of individual words.

More concretely, we expect a facilitatory interaction between the acquisition of MWUs and the acquisition of their constituent words. Provisional evidence for a beneficial impact of MWUs on the acquisition of smaller Linguistic units was collected by Arnon and Clark (2011), who showed that children make fewer inflectional errors on known words if

[^77]the words are contained within frequent MWUs. Usagebased approaches to language acquisition, meanwhile, suggest that children acquire a repertoire of both lexically specific and more abstract multi-word constructions (Tomasello, 2009; Behrens, 2009). Based on this, we propose that children sometimes possess MWU representations before they form representations of the words contained within them, and that these MWU representations then facilitate the acquisition of single-word representations. We dub this the $M W U$ acquisition hypothesis.

With the availability of a growing number of corpora of child-caregiver interactions on the one hand (MacWhinney, 2000) and the development of methods for the extraction of MWUs from corpora on the other hand (McCauley \& Christiansen, 2014; Brooke, Tsang, Hirst, \& Shein, 2014), we are in a position to investigate the kinds of MWUs children are likely to acquire. Concretely, we extract MWUs from two large corpora of transcribed child-directed speech, using (a) a computational model employed by McCauley and Christiansen (2014) to account for findings from the language acquisition literature and (b) an algorithm, developed by Brooke et al. (2014), intended to build a comprehensive lexicon of psychologically plausible MWUs. We view extracted MWUs as an approximation of the types of MWUs children might discover and use the number of MWUs within which a given word is contained as an independent variable.

Throughout, we use the age at which children first produce words (age of first production / AoFP) as an index of word learning: if a word is first produced relatively early in development, we assume that this is in part because it is easy to learn when and how to use it. Given the MWU acquisition hypothesis, we expect a facilitatory effect of the number of MWUs in which a word appears on its AoFP. This effect, moreover, should be uniquely attributable to MWUs - and not to individual word frequency, semantic co-variates, or the number of context words contained in MWUs. Number of cooccurring context words has previously been shown to predict age of acquisition for words (Hills, Maouene, Riordan, \& Smith, 2010); but if our proposal is correct, such an effect should disappear once MWUs are taken into consideration.

## Related Work

## Language Acquisition

MWUs have emerged as an important theoretical concept in usage-based approaches to Language Acquisition (Tomasello, 2009). Within this broad theoretical framework,
learners' linguistic representations are conceived of as continually complexifying entities, with the developed cognitive system containing both lexically specific and more abstract patterns. At early stages in development, most representations are lexically specific, and child language is "(partially) formulaic and item-based" (Behrens, 2009, p. 393). That is, child language development is thought to involve representations which are lexically specific and span multiple words.

Experimental evidence for the existence of children's MWU representations comes from Bannard and Matthews (2008), who presented 2 and 3 year-olds with frequent MWUs like a drink of tea and matched infrequent MWUs like a drink of milk that differed in the last word. 2 and 3 year-olds were faster to repeat frequent MWUs, and 3 year-olds were also faster to repeat the first three words if they formed a frequent MWU with the fourth word. Since the final word and the final bigram (e.g. of tea and of milk) were matched for frequency, the processing advantage for frequent MWUs can only be attributed to the frequency of the entire MWU, rather than to the frequencies of its component words, suggesting that children have access to cognitive representations of MWUs. Bannard and Matthews (2008) argue, furthermore, that their subjects were likely familiar with the words comprising the MWUs, which implies the existence of (partially) independent MWU and single-word representations. ${ }^{2}$

In addition, Arnon and Clark (2011) found that MWUs interact with the acquisition of morphemes: 4;6 year-olds produced more correct irregular plurals after familiar lexically specific frames than after general questions. Subjects were presented with depictions of several objects. The object name was elicited either with a labeling question or with a lexically specific frame. For example, on one particular trial the objects were sheep, the lexically specific frame was Count some - , and the labeling question was What are all these called? $4 ; 6$ year-olds were more likely to complete the lexically specific frame with sheep and would provide relatively more incorrect plural forms - like the over-regularized sheeps - in response to the labeling question. This suggests that MWUs like count some sheep affect the way in which some of the smaller units contained within them are learned.

## Computational Modeling

The above-cited results by Arnon and Clark (2011) and Bannard and Matthews (2008) have been modeled by McCauley and Christiansen (2014). In a comprehension phase, their model segments a corpus of child-directed speech into MWUs. In a production phase, it generates childproduced utterances based on stored MWUs. Given a corpus, MWUs are extracted by comparing the conditional probability of the current word given the preceding word to a running average of all such probabilities, for all words so far encoun-

[^78]tered one position to the left of the current word. If this backward transitional probability (BTP) is larger than the running average, the current and preceding word are part of an MWU. The process continues until the BTP falls below the average, at which point the current MWU is stored in memory.

Extracted MWUs can then be used to re-construct childproduced utterances. McCauley and Christiansen (2011) compared model-derived to child-produced utterances across 13 corpora from the CHILDES database (MacWhinney, 2000). On average, about $60 \%$ of utterances were successfully re-produced - illustrating that a purely MWU-based system can account for a majority of child-produced utterances. Importantly, MWUs discovered by the model can also be used to model results from Bannard and Matthews (2008) and Arnon and Clark (2011). In both cases, stimuli were sequences of words - constructions like a drink of tea in the former and count some sheep in the latter study. McCauley and Christiansen (2014) assigned a chunkedness score to each stimulus by calculating the product of BTPs between the MWUs used by the model to re-produce each stimulus. In each study, differences in scores reflected differences in subjects' performance: stimuli with lower reaction times in Bannard and Matthews (2008)'s study were assigned a larger chunkedness score, as were stimuli which elicited a larger proportion of correctly inflected nouns in Arnon and Clark (2011)'s study.

## Natural Language Processing

McCauley and Christiansen's $(2011,2014)$ model can be situated in a tradition that measures association strength between pairs of words; words are then grouped together if their association strength exceeds a particular threshold. McCauley and Christiansen $(2014,2011)$ use BTP as the measure of association. Other options include pointwise mutual information or log likelihood (cf. Pecina, 2010, for an overview). All association-based methods require an arbitrary threshold for inclusion of words in MWUs. In addition, there is no consensus on which association measure is best. An alternative approach is to identify frequent n -grams - called lexical bundles -, but this requires very high frequency thresholds (Biber, Conrad, \& Cortes, 2004). There is, then, no generally accepted way of extracting MWUs from corpora, nor is it common practice to evaluate whether extracted MWUs correspond to psychologically real entities.

Work by Brooke et al. (2014) has recently begun to address these issues. Their method operates at the token level, identifies MWUs of varying sizes, and relies on two parameters: a frequency threshold and a maximum MWU size. Broadly speaking, the algorithm considers all possible segmentations of a given sentence into $n$-grams that meet a pre-specified frequency threshold. Then, that segmentation is selected which maximizes the predictability of each word within its n-gram. The stated goal of this work is to develop a method for the extraction of an MWU lexicon that would correspond to knowledge of MWUs possessed by native speakers. The system has since been refined by Brooke et al. (2015), who also in-
troduced first steps towards evaluating MWU lexicons.

## Hypothesis

According to the MWU acquisition hypothesis, children sometimes acquire MWU representations before they acquire representations of the individual words contained in MWUs, and access to MWU representations then facilitates acquisition of the words contained in them. ${ }^{3}$ While this hypothesis is grounded in the literature, it is not clear via which mechanisms MWUs might aid the word learning process. Consequently, our goal is to provide evidence that MWUs uniquely facilitate word learning, and not how this process unfolds. Below, we nevertheless sketch two possible scenarios.

One possibility is that children initially acquire MWUs as unanalyzed units. This could result from an initial undersegmentation of the input: words, before their meaning is established, need to be identified from a continuous stream of sound. Early in development, children might sometimes segment multi-word chunks before they begin to segment individual words from within those chunks. Thus, some early fossilized MWUs are likely to be (partially) undersegmented chunks. In this scenario, the more initially undersegmented MWUs contain it, the earlier a given word is going to be segmented. We would then expect this early segmentation to translate into early induction of meaning.

A second possibility is that children discover some words before establishing their meaning. They would then go on to discover MWUs containing those words, at which point they have access to fully-fledged MWU representations without having access to the meaning of each individual word. The more MWUs contain a given word, the more words it is going to be linked to - and the more words will prime its retrieval, making it more salient for the learner. On average, a word with many links will be more easily retrieved than a word with few links. Because of this, we would expect fewer necessary exposures to establish the meaning of a word which forms part of relatively many MWUs, compared to words contained in fewer MWUs.

As mentioned, we do not distinguish between these two and other such possibilities. Instead, we aim to broadly corroborate the $M W U$ acquisition hypothesis by showing that MWUs uniquely facilitate word learning: if, all else being equal, words contained in many MWUs are learned earlier than other words, this would be indicative of a developmental pattern which begins with the formation of MWU representations and then proceeds to the acquisition of individual words.

## Method

Our method is the following: first, we extract MWUs from two corpora of English child-directed speech (CDS) and estimate age of first production (AoFP) for the words produced by the children addressed in the CDS corpora. We then use

[^79]the number of MWUs within which each target word appears (\#MWUs) as an independent variable - next to several covariates - in a linear regression analysis, with AoFP as the dependent variable. If the $M W U$ acquisition hypothesis is true, we expect a unique facilitatory effect of $\# M W U s$ on AoFP.

## Child-Directed Speech

We use two corpora of CDS, which both consist of the adultproduced utterances from several corpora on the CHILDES database (MacWhinney, 2000). Some corpora are based on cross-sectional studies, while others are longitudinal. In addition, subjects vary in age. Regardless, each corpus consists of standardized transcripts, based on recordings of childcaregiver interactions. In order to maximize the amount of data, we ignore possible fine-grained differences between age cohorts and compile a North-American corpus (NA-CDS) from 45 American English corpora ${ }^{4}$ and a British English corpus (BE-CDS) from eight British corpora ${ }^{5}$. Table 1 summarizes statistics.

Table 1: Relevant corpus statistics.

| measure | CDS-BE | CDS-NA |
| :--- | :--- | ---: |
| nr. tokens | $4,681,925$ | $6,389,963$ |
| nr. types | 24,929 | 37,128 |
| median length of utt. | 4 (IQR: 4) | 4 (IQR: 4) |
| nr. adult speakers | 201 | 774 |
| nr. children addressed | 134 | 441 |
| mean child age (months) | 33 (SD: 9) | 41 (SD: 23) |

## Extraction of Multi-Word Units

To extract MWUs from the CDS corpora, we use McCauley and Christiansen's (2014) model as well as Brooke et al.'s (2014) method. McCauley and Christiansen's (2014) model called Chunk-Based Learner (CBL) - processes a given corpus utterance by utterance and word by word. Processing an utterance $u$ is initiated by incrementing the frequency count of the first word $w_{1} \in u$ by 1 and creating a new MWU with $w_{1}$ as its only member. For each subsequent word $w_{i}$ at utterance position $1<i \leq$ length $(u)$, the model keeps track of the number of times $w_{i}$ has been encountered so far, as well as how often the immediately preceding word $w_{i-1}$ has occurred one position to the left of $w$. The model then calculates the backward transitional probability (BTP) of $w_{i}$ and $w_{i-1}$ : $p\left(w_{i-1} \mid w_{i}\right)$. If this probability is larger than the average BTP across all words which have occurred one position to the left

[^80]of $w$ in all utterances so far considered, $w_{i}$ is added to the current MWU. Else, the current MWU is added to a set $M$, and a new MWU is created - again with $w_{i}$ as its only member. In this way, the model discovers MWUs of size 2 or larger, as well as single-word units, collected in $M$. In our analyses, we use all MWUs which occur at least twice in the input corpus.

As a second model, we use the method from Brooke et al. (2014) ${ }^{6}$. We refer to it as Prediction Based Segmenter (PBS), as it splits utterances into n-grams whose component words are maximally predictable. The basic idea is that given an n-gram $w_{1} \ldots w_{n}$, the conditional probability of any word $w_{i}$ given the remaining subsequence $w_{1} \ldots w_{i-1}, w_{i+1} \ldots w_{n}$ should be maximal. In essence, the algorithm splits utterances into $n$ grams such that each word's predictability is maximized, capturing the intuition that words within MWUs are more predictive of one another than words outside of MWUs - but see Brooke et al. (2014) for a more in-depth explanation. Specifying a maximum n-gram length of ten - longer than most utterances in the corpus - , we use the PBS to segment utterances into either single-word units or MWUs with a minimum size of two and a maximum size of ten. As with the CBL, we retain all MWUs which occur at least twice.

Running the models on the two CDS corpora results in four different sets of MWUs, whose distributions are summarized in Table 2. The CBL results in a larger number of shorter MWUs, while the PBS identifies MWUs that are a bit longer. There are generally more MWU types than word types (compare Table 1).

Table 2: Relevant statistics about the distribution of MWUs.

| corpus | measure | CBL | PBS |
| :---: | :--- | ---: | ---: |
| CDS- | MWU tokens | $1,073,037$ | 978,804 |
|  | MWU types | 465,447 | 387,391 |
|  | median length | $4(\mathrm{IQR:} \mathrm{3)}$ | 5 (IQR: 4) |
| CDS- | MWU tokens | $1,40,8614$ | $1,338,173$ |
|  | MWU types | 628,252 | 492,863 |
|  | median length | $4(\mathrm{IQR:} \mathrm{3)}$ | 5 (IQR: 4) |

## Age of First Production

To induce AoFP, we start from a corpus of child-produced utterances, treating a word as having been learned at the earliest developmental stage at which any child within the corpus can produce it. Developmental stage is defined in terms of mean length of utterance (MLU) - the average child utterance length, in tokens, within a transcript. Since transcripts have varying lengths, we estimate MLU for each transcript via statistical bootstrapping, wherein the sampling distribution of the population is approximated by drawing random samples from the data (Davison \& Hinkley, 1997). Each bootstrap is based on 1000 random samples with replacement, with the sample size equal to the number of child utterances

[^81]per transcript. We thus induce MLU rather than AoFP estimates, since MLU is a more robust estimator of development (Parker \& Brorson, 2005): children who are close in age may nevertheless be far apart in terms of language development. For simplicity, we still refer to a word's MLU value as its AoFP. To induce a value for any word, we calculate the set of MLUs $\gamma$ for all transcripts within which the word appears and assign it the smallest value in $\gamma$.

We perform this procedure for each word produced by the children addressed in the two CDS corpora - once for the NA data and once for the BE data, meaning that we end up with two AoFP data sets: 441 children are addressed in the CDSNA corpus and together produce 29,188 different words, each of which is assigned an AoFP value; and 134 children are addressed in the CDS-BE corpus, producing 14,747 different words, again each with its own AoFP value.

## Regression Analyses

In the regression models, we use AoFP as the dependent variable. The first key independent variable is the number of different MWUs within which a given target word appears (\#MWUs). For example, assuming our corpus is CDS-NA and our target words are girl and sit, we count the unique MWUs which contain these two words. To illustrate this, Table 3 shows the five most frequent MWUs, in CDS-NA, containing the two words. Counting all such MWUs, we end up with 113 (PBS) and 230 MWUs (CBL) for girl, and 253 (PBS) and 488 (CBL) MWUs for sit. The second key independent variable is the number of unique context words appearing in all MWUs within which a given target word is contained (\#ctxt). If MWUs aid word learning, we should see a facilitatory effect of \#MWUs on AoFP, and this effect should not be reducible to \#ctxt. If a target word appears within a large number of MWUs, it will also tend to co-occur with a large number of context words. We posit, however, that MWUs - not individual words - are the cognitively relevant units; and we predict, therefore, that it is the number of MWUs - not the number of co-occurring context words which affects learning.

Further, we include the following co-variates: the corpusfrequency of each target word (freq), number of syllables (syl), phonological neighborhood density (phon), and concreteness ratings (con). Given a target word, phon is defined as the number of homophones, plus the number of words that can be derived from the target by either adding, deleting, or substituting a single phoneme. phon, together with $n s y l$, is derived from the CMU pronunciation dictionary ${ }^{7}$. Concreteness ratings for 40,000 lemmas are taken from Brysbaert, Warriner, and Kuperman (2014) ${ }^{8}$, who collected them from over four thousand participants via Mechanical Turk. Since ratings were collected for lemmas, whereas we work with word forms, we assigned the lemma rating to all word forms which correspond to the lemma. Regression analyses are based on

[^82]Table 3: The five most frequent MWUs, found in CDS-NA, for the target words girl and sit. Frequency counts for the MWUs are give in parentheses.

| word | CBL | PBS |
| :--- | :--- | :--- |
| girl | good girl (410) | good girl (440) |
|  | little girl (110) | that's a girl (101) |
|  | a girl (68) | little girl (175) |
|  | that's a good girl (98) |  |
|  | sit down (627) | that's a good girl (59) |
| the little girl (51) |  |  |
| sit | sit up (88) | sit down (846) |
|  | sit here (46) | sit up (141) |
|  | sit over here (46) | you sit (117) |
|  | sit down please (41) | you wanna sit (87) |
| come sit (85) |  |  |

all words for which phon, syl and con estimates are available: 7,265 words in CDS-BE and 5,724 words in CDS-NA. Table 4 shows three example data points.

To increase the generality of this study's implications, we use AoFP from children who were not addressed in the corpus used to estimate \#MWUs, \#ctxt, and frequency. In other words, we use AoFP from the children addressed in the CDSNA corpus for regression models which include \#MWUs, \#ctxt and frequency counts from CDS-BE; and we use AoFP from the children addressed in CDS-BE for regression models which include independent variables from CDS-NA.

Table 4: Example data points from the CDS-BE corpus, with \#MWUs and \#ctxt estimated via the PBS. The phon and nsyl predictors are not shown due to space constraints.

| word | freq | con | \#ctxt | \#MWUs | AoFP |
| :--- | :--- | :--- | :--- | :--- | :--- |
| goes | 3,183 | 2.19 | 430 | 156 | 0.51 |
| lunch | 1,175 | 4.31 | 168 | 57 | 1.29 |
| running | 853 | 4.27 | 86 | 46 | 1.16 |

## Results

Table 5 presents results of four linear regression analyses (2 methods for MWU extraction $\times 2$ CDS corpora). All variables are log-transformed, and \#ctxt as well as \#MWUs are increased by 1 , in order to avoid problems from zero counts. The baseline models with all co-variates (second column) explain between 38 and 44 percent of the variance in AoFP. Freq and con have facilitatory effects, while there are no statistically significant effects for phon and nsyl. Given that increased frequency of exposure is associated with early word learning (Ambridge, Kidd, Rowland, \& Theakston, 2015), the effect of freq is not surprising, while the effect of con implies that words associated with concrete concepts tend to be earlyacquired.

Adding \#ctxt to the baseline models (third column) leads to
a significant increase in $R^{2}$, with a facilitatory effect of \#ctxt. Adding \#MWUs to the baseline models (fourth column) also improves the fit, with a facilitatory effect of \#MWUs. Interestingly, the effect of $\# M W U s$ is stronger than the effect of \#ctxt. Neither effect is reducible to the frequency of target words, their concreteness, their phonological complexity, or the density of their phonological neighborhoods. In models which include the covariates plus \#ctxt and \#MWUs (fifth and sixth columns), \#MWUs continues to exert a facilitatory effect; but importantly, \#ctxt now has an inhibitory effect on AoFP. This pattern suggests that the initial facilitatory effect of \#ctxt is due to collinearity with \#MWUs.

Our results imply that it is involvement in a large number of MWUs - not co-occurrence with a large number of context words - which drives word learning. Furthermore, the effect of MWUs may be limited to MWUs consisting of relatively few words. Hence, when factoring out \#MWUs, co-occurrence with a large number of context words inhibits acquisition of the target words; and when factoring out the effect of context words, the positive effect of \#MWUs persists.

## Conclusions and Future Work

We began this paper with a review of studies which suggest that children acquire representations of MWUs and that MWUs could facilitate the acquisition of smaller linguistic units contained within them. Based on this, we proposed the MWU acquisition hypothesis, according to which the formation of MWU representations precedes and facilitates the formation of individual word representations. The facilitatory effect of $\# M W U s$ on AoFP supports this hypothesis. More broadly, it supports accounts of language development wherein children acquire linguistic units at various levels of granularity, transitioning gradually from MWUs to more small-grained units.

Our results also have implications for a previous finding: Hills et al. (2010) found that the sum of unique context words occurring within a window of five words to the left and right of each target word predicts age of acquisition of the targets. We also observed a facilitatory effect of \#ctxt. However, an inhibitory effect of \#ctxt emerged once \#MWUs was controlled for. Thus, given that their measure is similar to \#ctxt, it is possible that Hills et al. (2010)'s result is due to collinearity with the number of MWUs within which target words appear.

In formulating the hypothesis, we purposefully remained agnostic with respect to the specific mechanisms involved in the facilitatory interaction between the acquisition of MWU and single word representations. Accordingly, our results support a general class of theories wherein MWUs are acquired before single words. These could be usage based approaches to language acquisition (Tomasello, 2009), but also proposals such as Peters' (1983), according to which earlyacquired MWUs are undersegmented chunks which are gradually segmented into smaller units - units which are themselves stored in memory, where they are again subject to segmentation. In future work, we plan to experiment with differ-

|  | Effect ( $\Delta R^{2}$ in \%) |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Data set and corpus | Covariates <br> baseline | Log-\#ctxt | Log-\#MWUs | Log-\#ctxt <br> unique | Log-\#MWUs <br> unique |
| CBL |  |  |  |  |  |
| CDS-BE | $44.85^{* * *}$ | $1.23^{* * *}$ | $1.73^{* * *}$ | $0.34(\mathrm{I})^{* * *}$ | $0.85^{* * *}$ |
| CDS-NA | $38.33^{* * *}$ | $0.87^{* * *}$ | $1.35^{* * *}$ | $0.13(\mathrm{I})^{* * *}$ | $0.61 * * *$ |
| PBS |  |  |  |  |  |
| CDS-BE | $44.85^{* * *}$ | $0.78^{* * *}$ | $1.52^{* * *}$ | $0.55(\mathrm{I})^{* * *}$ | $1.29^{* * *}$ |
| CDS-NA | $38.33^{* * *}$ | $0.47^{* * *}$ | $1.09^{* * *}$ | $0.18(\mathrm{I})^{* * *}$ | $0.79^{* * *}$ |

Table 5: Effects of log-transformed \#ctxt and log-transformed \#MWUs. The effects of \#ctxt and \#MWUs were calculated after those of the co-variates had been included. Unique effects are those with the indicated variable entered last (i.e. \#ctxt after covariates + \#MWUs, or \#MWUs after \#ctxt + covariates). I = inhibitory effect of indicated variable.
ent operationalizations of MWUs, in order to examine what types of MWUs have the strongest potential effect on word learning. This, in turn, may allow us to more closely specify the mechanisms whereby MWUs facilitate word learning.

## Acknowledgments

This research was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp.

## References

Ambridge, B., Kidd, E., Rowland, C. F., \& Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. Journal of Child Language, 42(02), 239-273.
Arnon, I., \& Clark, E. V. (2011). Why brush your teeth is better than teeth-children's word production is facilitated in familiar sentence-frames. Language Learning and Development, 7(2), 107-129.
Arnon, I., \& Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62(1), 67-82.
Bannard, C., \& Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. Psychological Science, 19(3), 241-248.
Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. Linguistics, 47(2), 383-411.
Biber, D., Conrad, S., \& Cortes, V. (2004). If you look at: Lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), 371-405.
Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G., \& Shein, F. (2015). Building a lexicon of formulaic language for language learners. In Proceedings of the 11th workshop on multiword expressions (pp. 96-104). Denver, Colorado: Association for Computational Linguistics.
Brooke, J., Tsang, V., Hirst, G., \& Shein, F. (2014). Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In Proceedings of coling 2014, the 25th international conference on
computational linguistics: Technical papers (pp. 753-761). Dublin, Ireland: Dublin City University and Association for Computational Linguistics.
Brysbaert, M., Warriner, A. B., \& Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. Behavior Research Methods, 46(3), 904911.

Davison, A. C., \& Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge: Cambridge University Press.
Hills, T. T., Maouene, J., Riordan, B., \& Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. Journal of Memory and Language, 63(3), 259-273.
MacWhinney, B. (2000). The childes project: tools for analyzing talk. Hillsdale, NJ: Lawrence Erlbaum Associates.
McCauley, S. M., \& Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The cappuccino model. In L. A. Carlson, C. Hölscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd annual conference of the cognitive science society (pp. 1619-24). Austin, TX: Cognitive Science Socitety.
McCauley, S. M., \& Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. The Mental Lexicon, 9(3), 419-436.
O'Donnell, T. J. (2015). Productivity and reuse in language: A theory of linguistic computation and storage. MIT Press.
Parker, M. D., \& Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (mlum) and mean length of utterance in words (mluw). First Language, 25(3), 365-376.
Peters, A. M. (1983). The units of language acquisition. Harvard university press.
Tomasello, M. (2009). Constructing a language. Harvard university press.
Wray, A. (2008). Formulaic language: Pushing the boundaries. Oxford University Press.

# The nature of quantities influences the representation of arithmetic problems: evidence from drawings and solving procedures in children and adults 

Hippolyte Gros (hippolyte.gros@cri-paris.org) ${ }^{(1,2,3)}$<br>(1) Université Paris Descartes, Sorbonne Paris Cité, 12 Rue de l'École de Médecine, 75006 Paris, France.<br>Jean-Pierre Thibaut (jean-pierre.thibaut @u-bourgogne.fr)<br>(2) LEAD-CNRS, UMR 5022, UB, Université de Bourgogne Franche-Comté, Pôle AAFE - Esplanade Erasme, 21065 Dijon. France.<br>Emmanuel Sander (emmanuel.sander@univ-paris8.fr)<br>(3) Paragraphe Lab, EA 349, Université Paris 8,<br>2 Rue de la Liberté, 93526 Saint-Denis Cedex 02, France.


#### Abstract

When solving arithmetic problems, semantic factors influence the representations built (Gamo, Sander \& Richard, 2010). In order to specify such interpretative processes, we created structurally isomorphic word problems that could be solved with two distinct algorithms. We tested whether a distinction between cardinal and ordinal quantities would lead solvers, due to their daily-life knowledge, to build different representations, influencing their strategies as well as the nature of their drawings. We compared $5^{\text {th }}$ grade children and adults in order to assess the validity of this hypothesis with participants of varying arithmetic proficiency. The results confirmed that the distinction between cardinal and ordinal situations led to different solving strategies and to different drawings among both age groups. This study supports the ontological distinction of cardinal versus ordinal quantities and calls for the consideration of the role of daily-life semantics when accounting for arithmetic problem solving processes.


Keywords: arithmetic problem solving; interpreted structure; semantic encoding; strategy choice.

## Introduction

What are the steps taken from reading an arithmetic word problem to implementing a set of mathematical operations, and how can they be studied? It is well established, since Riley, Greeno and Heller (1983) proposed their typology of additive word problems, that different problem statements lead to different performances. Yet, the reasoning processes and representations accounting for such differences remains controversial.

The schema theory (Kintsch \& Greeno, 1985) proposes that solving a word problem requires to select and to instantiate a schema fitting the problem at hand. For example, any comparison problem will require to retrieve the corresponding schema and to implement it with the available values (Riley et al., 1983). However, it has been argued that this approach underestimates interpretative effects. For instance, Hudson (1983) showed that young children had much more trouble solving a problem stating "there are 5 birds and 3 worms, how many more birds is there than worms?" than they had solving a problem in which the question was "how many birds won't get a worm?". This
result shows that two problems which share the same schema can lead to different performances.

A contrasting approach comes from Johnson-Laird's (1983) theory of mental models. It posits that, during reading, a mental representation is constructed in working memory, and that its structure is analogous to that of the situation depicted in the problem statement (Reusser, 1990). This representation depicts the meaningful relations between the elements of the problem. The idea of a problem-specific representation, integrating conceptual information from the problem statement, can account for the interpretative effects described in the literature. De Corte, Verschaffel and De Win (1985) showed that rewording a problem statement by making the semantic relations more salient facilitates the solving process. Similarly, introducing daily-life situations in the cover stories of word problems contributes to better performance (Stern \& Lehrndorfer, 1992; Vlahović-Štetić, 1999). The use of specific words or sentences can modify the representation constructed by the solvers (Cummins, Kintsch, Reusser \& Weimer, 1988). In a study challenging the predictions of mental model and schema views, Thevenot, Devidal, Barrouillet and Fayol (2007) showed that placing the question at the beginning instead of at the end of a problem statement provided more benefit to the less experienced solvers. This result supported the mental model theory, whereas the schema theory would have predicted the reverse pattern.

## Semantic determinants of problem solving

The semantic determinants of a solvers' mental representations are an important issue. Bassok, Wu and Olseth (1995) showed that the semantic relations which connect a problem's entities influence analogical transfer. They contrasted problems where objects were given to people (OP) and problems were people were assigned to objects (PO). They found that, since in real life objects are usually given to people rather than people being assigned to objects, OP training examples led to better performance with OP transfer problems than PO training examples did with PO transfer problems.

Along this line, Bassok, Chase and Martin (1998) asked participants to create addition or division problems involving specific sets of objects that were provided. They showed that when the objects shared a functionally asymmetric semantic relation (e.g. apples and baskets evoke the contain relation), participants tended to create division problems, whereas they created addition problems when using functionally symmetric sets of objects (such as oranges and apples, that belong to the same superordinate fruit category). These biases are not driven by arithmetic properties but rather by the world semantics. Bassok (2001) developed the semantic alignment framework proposing that solvers abstract an interpreted structure that depends on their world knowledge about the entities described in the problem statement. This interpreted structure integrates the structural role of the entities mentioned in the problem, and can thus lead to an appropriate use of abstract formal knowledge when the relations it describes are semantically aligned with the mathematical relations of the problem (Bassok et al., 1998; Bassok, 2001). Both behavioral (Bassok, Pedigo \& Oskarsson, 2008) and physiological (Guthormsen et al., 2015) measures confirmed that problem solving is easier when daily-life knowledge (world semantics) and knowledge about mathematical concepts (mathematical semantics) are aligned with each other.

## Investigating participants' representations

The semantic alignment framework predicts that representations abstracted from problem statements influence the solver's solving strategies. Yet, the key semantic dimensions influencing the representations and explaining the lack of transfer remain to be elucidated in order to promote methods to help students overcome the incompatibilities posed by a problem.

In this regard, problems with multiple solving strategies are of particular interest to study representations, since the selection of one strategy over another is informative about the representation constructed by the solvers (De Corte, Verschaffel \& De Win, 1985). For instance, Thevenot and Oakhill (2005) worked on a multiple-step problem solving task in which the cognitive load was manipulated through values size (large or small). They showed that depending on the size of the values, participants used different solving algorithms. The issue of the semantic determinants of problem representations can be tackled by using such a paradigm in which different solving strategies are available, and the solver's ability to pick and use one informs us about the abstracted interpreted structures (Hakem, Sander, Labat \& Richard, 2005). For example, Coquin-Viennot \& Moreau (2003) showed that the presence of a grouping element in a problem statement (such as flowers presented within a bouquet instead of separately) could incite participants to use a factorizing rather than a development algorithm.

Another way to study the participants' mental representations is the use of drawings. Vosniadou and Brewer (1992) elicited drawings from $3^{\text {rd }}$ and $5^{\text {th }}$ grade children so as to study the development of their representations of the earth. As for problem solving, studies have highlighted the link
between problem representations and drawings of the situations (Barrios \& Martinez, 2014; Edens \& Potter, 2007). Drawings are thus an accurate way to gather information regarding the solvers representations.

## Encoding ordinal and cardinal quantities

In the following experiment, we will capitalize on problems that can be solved with two different strategies. Previous studies have suggested that an ontological distinction can be drawn between two types of situations involving numerical values: cardinal situations, consisting of sets of unordered elements, and ordinal situations, where units are endogenously ordered and can be represented along an axis, such as a timeline (Gamo et al., 2010; Hakem et al., 2005; Sander \& Richard, 2005). When solving an arithmetic word problem, the authors posited that solvers abstract an interpreted structure that is aligned with either a cardinal or an ordinal representation.

These two types of representations elicit different solving strategies: in ordinal representations, subtractions are seen as calculations performed on a one-dimensional ordered scale, whereas in cardinal representations they are encoded as a difference between a whole and a component part (Hakem et al., 2005). Thus, according to Gamo et al.'s hypothesis, a subtraction could either be perceived as a comparison or as a complementation, depending on the situation described in the problem statement. The paradigm developed by Hakem et al. consisted in problems that admitted two distinct solving strategies implementable for both cardinal (number of people) and ordinal (duration) quantities. Problem statements 1 and 2 below embody this distinction between cardinal problems and ordinal problems:

- Problem 1: "There are 5 people in the Richard family. When the Richards go on holiday with the Roberts, they make a total of 14 people at the hotel. The Roberts are joined on holiday by the Dumas family. In the Dumas family, there are 3 people less than in the Richard family. The Roberts are going on holiday with the Dumas. How many will they be at the hotel?"
- Problem 2: "Antoine took painting classes for 5 years, and stopped at the age of 14 . Jean started at the same age as Antoine, and went to classes 3 years less than him. How old was Jean when he stopped attending painting classes?"


Figure 1: Structure of the problems. This structure can depict both problems and is compatible with both strategies.

Problems 1 and 2 are isomorphs sharing the same deep structure (Figure 1), and can both be solved using either of two strategies: either a 3-step complementation strategy ( 14 $5=9 ; 5-3=2 ; 9+2=11$ ) or a 1 -step matching strategy ( $14-3=11$ ). Yet, the authors hypothesized that (i) because the quantities used are different, the interpreted structures are too, each problem statement consequently favoring the use of one strategy over the other; and that (ii) the unequal distribution of strategies used may be accounted for by the nature of the abstracted representations: problem 1 encoded as a cardinal problem (Figure 2) and problem 2 encoded as an ordinal problem (Figure 3).


Figure 2: Cardinal representation of problem 1. This interpreted structure fosters the calculation of the intersection (part 2) between whole 1 and whole 2, thus favoring the 3 -step complementation strategy.


Figure 3: Ordinal representation of problem 2. This interpreted structure puts forward the fact that the difference between whole 1 and whole 2 is equal to the difference between part 1 and part 3 . The shorter 1 -step comparison algorithm thus becomes available to solve the problem.

In accordance with the authors' hypothesis, the participants who were asked to solve the problems using as few operations as possible found the 1 -step matching strategy on problem 1 in less than $5 \%$ of the cases. On the other hand, problem 2 led to the use of the 1 -step matching strategy in over $60 \%$ of the cases, suggesting that comparisons are indeed made salient in ordinal representations. Hakem et al.'s (2005) study of the solving strategies showed that the two types of problems were underlain by different representations. Yet, the claim that ordinal and cardinal quantities evoke the corresponding ordinal and cardinal representations warrants further empirical support.

## Present study

Our study builds on the work of Hakem et al. (2005) in order to highlight the role of general semantic features on the representations abstracted by the solvers and on the implemented solving strategies. We aimed at providing converging measures of the impact of the cardinal/ordinal distinction on the solvers' ability to solve the problems, and to provide the first empirical test of these effects on children and adults simultaneously. To this end, $5^{\text {th }}$ graders as well as adults were asked to perform two tasks: solving problems involving different types of cardinal and ordinal variables using as few operations as possible, and making a drawing for each problem.

The goal of the experiment was twofold: first, we intended to confirm with both age groups the validity of the ordinal versus cardinal distinction with new material including new types of quantities and using systematically controlled problem statements. This was intended to show that strong semantic effects affect both younger and older - more proficient - participants in arithmetic problem solving. Second, we wanted to show that those effects originate in the representations abstracted from the problems, and translate into the algorithms implemented by the solvers. We predicted that within each group, the mean percentage of the 1 -step matching strategy would be significantly lower on cardinal problems than on ordinal problems, despite the adults achieving a globally higher solving performance than the children. Also, we predicted that for each age group the drawings would reveal a higher ordinal versus cardinal ratio of distinctive features for ordinal than for cardinal problems.

## Experiment

## Participants

We recruited samples from two populations for this study: a group of 59 children in $5^{\text {th }}$ grade ( 27 females, $M=11.00$ years, $S D=0.36$ ), and a group of 52 adults ( 36 females, $M=$ 26.86 years, $S D=9.72$ ). All participants were recruited from the Paris region and spoke French fluently. None had previously participated in any similar experiment.

## Materials and procedure

Each participant was presented with a set of 12 problems, 6 using ordinal values (duration, height or number of floors), and 6 using cardinal values (number of elements, price or weight), according to Hakem et al.'s definition. We considered duration, height and number of floors as ordinal values because their ordinal component is salient in daily life, putting emphasis on successorship relation and on comparison. Similarly, number of elements, price and weight were used as cardinal values because the world semantics attached to such quantities evoke the unordered grouping of elements assigned to values and the partition of a whole into its component parts.

All the problems had the same number of sentences. The numerical values were provided in the same order and both numerical values and problem order were randomized between participants. The problems were printed on 13-page
booklets with the instructions detailed on the first page. The participants were asked (1) to solve the problems using as few arithmetic operations as possible, (2) to write down every operation they made, even those they solved using mental calculation, and (3) to make a drawing for each problem statement that could help someone else understand and solve the problem. Each page was divided into four parts: problem statement, 'draft' area, 'response' area and 'drawing' area. The booklets and instructions were strictly identical for both age groups.

## Coding

The successful strategies were categorized either as correct 1 -step matching strategy, or as correct 3 -step complementation strategy. A problem was considered correct when the expected result was obtained and accompanied by calculations ${ }^{1}$. Regarding the drawings, we designed an 8-item rating scale evaluating to what extent the drawings possessed ordinal versus cardinal characteristics. The scale included 4 cardinal items (Figure 4.a) and 4 ordinal items (Figure 4.b). The items were chosen so that they would either depict unordered elements being grouped in sets, or ordered elements being described as positions on an axis and compared along this axis.


Figure 4: Coding grids for cardinal (a) and ordinal (b) features.

[^83]Each drawing, including those of failed problems, was scored by two independent raters who were not familiar with the theories at play and ignored the hypotheses being tested. They were asked to rate the drawings according to the 8 items scale resulting from the aggregation of Figures 4.a and 4.b. After an initial rating phase, the percent agreement between the two raters was of $91.87 \%$. An inter-rater reliability analysis using Cohen's Kappa statistic showed substantial agreement between raters. $(\kappa=0.61, S E=0.14)$, according to Landis \& Koch's typology (1977). After discussion, the raters reached $100 \%$ agreement. For each drawing, a score was then calculated by subtracting the number of cardinal items to the number of ordinal items, thus creating a scale ranging from -4 (the most cardinal) to +4 (the most ordinal).

## Results

Our first hypothesis was that problems with ordinal quantities should facilitate the use of the matching 1-step strategy compared to problems with cardinal quantities.


Figure 5: Children's and adults' mean rate of use of the two solving algorithms depending on the type of quantity used.

Figure 5 details the participants' use of each strategy depending on the type of quantity featured in the problems. A paired t -test revealed that the mean rate of use of the 1 -step matching strategy was higher on ordinal $(M=0.39, S D=$ 0.31 ) than on cardinal ( $M=0.08, S D=0.17$ ) problems $(t(58)=8.36, p<0.001)$. The same analyses were performed for the adults and showed that the mean rate of use of the 1step matching strategy was also higher on ordinal ( $M=0.457$, $S D=0.33)$ than on cardinal $(M=0.253, S D=0.35)$ problems $(t(51)=4.99, p<0.001)$. This confirmed that the choice of a solving algorithm is influenced by the cardinal versus ordinal nature of the quantities, and that this effect is robust among
adults. Additionally, the 1-step algorithm was significantly less used by children than by adults on cardinal $(t(109)=$ $3.48, p<0.001$, unpaired t-test) but not ordinal $(t(109)=1.10$, $p=0.27$, unpaired t-test) problems, meaning that children had significantly more difficulty than adults using the 1 -step strategy on cardinal, but not on ordinal problems.

To test our second hypothesis, we focused on the drawings made by the participants. Figure 6 details the rating of the drawings depending on the type of quantity used in the problems. The drawing score was significantly lower for drawings depicting problems with cardinal quantities ( $M=-$ $0.55, S D=0.78$ ) than for those describing problems with ordinal quantities $(M=0.06, S D=0.87, t(58)=5.61, p<$ 0.001 , paired t-test), indicating that problems using ordinal quantities led young participants to draw ordinal features (axes, intervals, etc.) at a higher ratio over cardinal features (sets, groups of elements, etc.) compared to the ordinal problems.


Figure 6: Children's and adults' mean drawing score depending on the type of quantity used in the problems. Vertical bars denote 0.95 -confidence intervals.

Similarly, among adult participants, problems with cardinal quantities $(M=-1.48, S D=0.79)$ led to a significantly lower drawing score than problems with ordinal quantities ( $M=$ $0.89, S D=0.86, t(51)=12.44, p<0.001$, paired t-test). In sum, the presence of ordinal (resp. cardinal) quantities seems to result in representations featuring a higher number of ordinal (resp. cardinal) features, in both children and adults. Of note, drawing score was significantly higher among children than among adults on cardinal problems $(t(109)=$ $6.24, p<0.001$, unpaired t -test) but significantly lower among children than among adults on ordinal problems $(t(109)=5.00, p<0.001$, unpaired t -test $)$; in other words, children included significantly less cardinal features than adults while representing cardinal problems, and significantly less ordinal features than adults when representing ordinal problems.

## Discussion

The fact that the cardinal versus ordinal distinction in problem statements influenced both children's and adults' solving strategies confirmed the robustness of these interpretative effects, even with experienced solvers who should not meet any difficulty in solving such simple problems. Indeed, children performed about half as well as adults, yet the distinction between cardinal and ordinal problems remains significant in both populations. Additionally, adults' performances were significantly higher on cardinal, but not on ordinal problems, indicating that when semantically congruent with the 1 -step strategy, world semantics help children achieve adult-like performance on the task.

The elicited drawings provided an empirical confirmation of the importance of the ordinal versus cardinal distinction in both populations. The fact that children produced drawings that had significantly fewer ordinal or cardinal features on ordinal or cardinal problems, respectively, may be attributed to a global lack of details in their drawings, which nevertheless did not prevent a significant distinction between cardinal and ordinal drawings to be revealed among children. Additionally, children may have more difficulties to produce a graphic implementation of ordinal situations, which would explain their poor ordinal score ( 0.06 ) for ordinal problems. This calls for further research on the topic.

Overall, the results of both the drawing and the solving tasks show that the participants' semantic knowledge about the quantities used in the problems (i.e. their experience with counting the number of apples in a bag, adding the price of every item on a bucket list, calculating the arrival time of their train or using the elevator) influence the encoding of arithmetic word problems. The distinction introduced between ordinal and cardinal problem statements was reflected in the representations constructed (as shown by the drawings made by the participants) and led solvers to use different solving algorithms, even when asked specifically to use the shorter strategy they could find. Furthermore, the fact that those effects could be highlighted both with young pupils and adults indicates the robustness of such encoding constraints. The ontological distinction hypothesized between ordinal and cardinal representations was thus confirmed on two complementary tasks.

The use of a double measure of the influence of the solvers' world knowledge allowed us to gather converging clues shedding light both into the abstracted representations and into the subsequently implemented algorithms. By focusing on the role of semantic properties on the initial encoding of a problem, we hope to gain a finer understanding of arithmetic problem solving in its whole, and to pave the way for accounting for the interactions between world semantics, mathematical semantics and algorithms. Understanding the determinants of problem representations is a crucial step to identify the potential pitfalls and dead ends born from semantic incongruence (Gros, Sander \& Thibaut, 2016) as well as to help develop analogical transfer between isomorphic problems by promoting semantic recoding among
the learners (Gamo, Sander \& Richard, 2010; Gros, Thibaut \& Sander, 2015). Doubtlessly, the educational opportunities resulting from a better understanding of the conditions necessary for semantic recoding and analogical transfer between problems are promising.

## References

Barrios, F. M. G., \& Martínez, E. C. (2014). Diagrams produced by secondary students in multiplicative comparison word problems. Journal of Mathematics and System Science, 4(2), 83.
Bassok, M. (2001). Semantic alignments in mathematical word problems. In Gentner, Holyoak and Kokinov (eds.) The analogical mind: Perspectives from cognitive science, (pp.401-433). Cambridge, Ma: MIT Press
Bassok, M, Chase, V. M, \& Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. Cognitive Psychology, 35, 99-134
Bassok, M., Pedigo, S. F., \& Oskarsson, A. T. (2008). Priming addition facts with semantic relations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(2), 343.
Bassok, M., Wu, L., \& Olseth, L. K. (1995). Judging a book by its cover: interpretative effects of content on problemsolving transfer. Memory and Cognition, 23, 354 e 367.
Coquin-Viennot, D., \& Moreau, S. (2003). Highlighting the role of episodic model in the solving of arithmetical problems. European Journal of Psychology and Education, 18, 267-279.
Cummins, D. D., Kintsch, W., Reusser, K., \& Weimer, R. (1988). The role of understanding in solving word problems. Cognitive psychology, 20(4), 405-438.
De Corte, E., Verschaffel, L., \& De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solution. Journal of Educational Psychology, 77, 460-470.
Edens, K., \& Potter, E. (2007). The relationship of drawing and mathematical problem solving: Draw for math tasks. Studies in Art Education, 48(3), 282-298.
Gamo, S., Sander, E., \& Richard, J.-F. (2010). Transfer of strategy use by semantic recoding in arithmetic problem solving. Learning and Instruction, 20(5), 400-410.
Gros, H., Sander, E., \& Thibaut, J.-P. (2016), "This problem has no solution" : when closing one of two doors results in failure to access any., $38^{\text {th }}$ Annual Meeting of the Cognitive Science Society, Philadelphia, USA, 10-13 August 2016.
Gros, H., Thibaut, J.-P., \& Sander, E. (2015), Robustness of semantic encoding effects in a transfer task for multiple strategies arithmetic problems, $37^{\text {hh }}$ Annual Meeting of the Cognitive Science Society, Pasadena, USA, 22-25 July 2015.

Guthormsen, A. M., Fisher, K. J., Bassok, M., Osterhout, L., DeWolf, M., \& Holyoak, K. J. (2015). Conceptual integration of arithmetic operations with real-world knowledge: Evidence from event-related potentials. Cognitive science.

Hakem K., Sander E., Labat J-M., DIANE : a diagnosis system for arithmetical problem solving, Proceedings of international Conference on Artificial Intelligence in Education (AIED2005), 258-265(2005).
Hudson, T. (1983). Correspondences and Numerical Differences between Disjoint Sets. Child Development, Vol. 54, No. 1, 84-90.
Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness (No. 6). Harvard University Press.
Kintsch, W., \& Greeno, J. G. (1985). Understanding and solving word arithmetic problems. Psychological review, $92(1), 109$.
Landis, J. R., \& Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159174.

Reusser, K. (1990). From text to situation to equation: Cognitive simulation of understanding and solving mathematical word problems. In H. Mandl, E. De Corte, N. Bennett, \& H.F. Friedrich (Eds.), Learning and instruction: European research in an international context. Volume 2.2: Analysis of complex skills and complex knowledge domains. Oxford, England: Pergamon.
Riley, M. S., Greeno, J. G., \& Heller, J. 1.(1983). Development of children's problem-solving ability in arithmetic. The development of mathematical thinking, 153-196.
Sander, E., \& Richard, J. F. (2005). Analogy and transfer: encoding the problem at the right level of abstraction. In Proceedings of the 27th Annual Conference of the Cognitive Science Society, Stresa, Italy.
Stern, E., \& Lehrndorfer, A. (1992). The role of situational context in solving word problems. Cognitive Development, 7(2), 259-268.
Vlahović-Štetić, V. (1999). Word-problem solving as a function of problem type, situational context and drawing. Studia Psychologica, 41(1), 49-62.
Thevenot, C., Devidal, M., Barrouillet, P., \& Fayol, M. (2007). Why does placing the question before an arithmetic word problem improve performance? A situation model account. The Quarterly Journal of Experimental Psychology, 60(1), 43-56.
Thevenot, C., \& Oakhill, J. (2005). The strategic use of alternative representations in arithmetic word problem solving. The Quarterly Journal of Experimental Psychology Section A, 58(7), 1311-1323.
Vosniadou, S., \& Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. Cognitive psychology, 24(4), 535-585.

# Does Mandarin Spatial Metaphor for Time Influence Chinese Deaf Signers＇ Spatio－Temporal Reasoning？ 

Yan Gu ${ }^{1}$（yan．gu＠tilburguniversity．edu）<br>Yeqiu Zheng ${ }^{2}$（yeqiu．zheng＠tilburguniversity．edu）<br>Marc Swerts ${ }^{1}$（m．g．j．swerts＠tilburguniversity．edu）<br>${ }^{1}$ Tilburg center for Cognition and Communication（TiCC），Tilburg University，the Netherlands<br>${ }^{2}$ Department of Econometrics and Operations Research，Tilburg University，the Netherlands


#### Abstract

In Mandarin Chinese，the space－time word＂前／qian＂is used to express both the spatial concept of front／forward and the temporal concept of early／before（e．g．，＂前天／qian－ tian＂，literally front day，meaning the day before yesterday）． This is consistent with the fact that Mandarin speakers can gesture to the front of the body to refer to a past event，and more generally can have past－in－front space－time mappings． In Chinese Sign Languages，however，the spatial front／forward and the temporal early／before are signed differently as the sign for spatial front is only used for the spatial concept of forward，and the sign for before／past is directed to the back．In this study we investigate whether the Mandarin sagittal spatial metaphors for time influence Chinese deaf signers＇spatio－temporal reasoning．In two experiments，we found that Chinese deaf signers with higher Mandarin proficiency were more likely to interpret the Mandarin word＂前／qian＂as the temporal conception of past（Study 1），and to perform past－in－front space－time mappings（Study 2）as opposed to signers with lower Mandarin proficiency．The findings of the study not only provide within－culture evidence for the influence of language on thought，but also demonstrate that even cross－ modal space－time metaphors can have an impact on deaf－ signers＇spatio－temporal reasoning．


Keywords：space and time；Chinese deaf signers；language and thought；conceptual metaphor

## Introduction

Across cultures people use spatial representations to think about time（Bottini，Crepaldi，Casasanto，Crollen，\＆ Collignon，2015；Boroditsky，2000；Casasanto \＆ Boroditsky，2008；see reviews Bender \＆Beller，2014； Núñez，\＆Cooperrider，2013）．Most Europeans feel that the future is in front of them and the past is at their back （e．g．，Miles，Nind，\＆Macrae，2010；Ulrich et al．，2012）． Such an intuition matches the human＇s experience of walking in a certain direction，which is usually forwarding to the front，so that the passed－by path is the past and the place ahead represents the future．Interestingly，the future－ in－front and past－at－back mappings are also expressed as such in many languages．For instance，in English，one can say＂We look forward to the New Year ahead，and look back to the hard times behind（e．g．，Clark，1973；Lakoff \＆ Johnson，1980）．
However，the way of conceptualizing the past at the back and the future in the front does not generalise to all languages．For example，speakers of Aymara exhibit the opposite sagittal space－time mappings，with past things in front of them，and the future as yet unseen events behind
them．This conceptual mapping is consistent with the way they produce co－speech gestures，and with the spatial metaphors in their language，as front year in Aymara has the meaning of last year（Núñez \＆Sweetser，2006）． Interestingly，Moroccans also have a strong tendency to place past events in front，even though in Arabic the front／back time metaphors are similar to most future－in－ front languages such as English．It has been argued that the reason for Moroccans＇past－in－front space－time mapping is that，in their culture，tradition and old generations are more valued．Thus space－time mappings in people＇s minds are conditioned by their cultural attitudes towards time（e．g．，with a strong focus on past times and old generations）．It is claimed that the mental space－time mappings are dependent on attentional focus and can be independent from the space－time mappings expressed in language（de la Fuente，Santiago，Román， Dumitrache，\＆Casasanto，2014）．Moreover，a recent study on Mandarin speakers shows that there are both long－term effects of cultural attitudes on the spatialization of time，and immediate effects of lexical cues to space－ time metaphors which can probe people＇s mental representations（Gu，Zheng，\＆Swerts，2016）．

Despite the fact that a growing number of studies have shown that linguistic，cultural and bodily experiences have separate influences on people＇s spatial representation of time（e．g．，Boroditsky，2001；Casasanto，\＆Bottini， 2014；Fuhrman \＆Boroditsky，2010；Núñez \＆Sweetser， 2006；Núñez，Cooperrider，Doan，\＆Wassmann，2012；Saj， Fuhrman，Vuilleumier，\＆Boroditsky，2014；Torralbo， Santiago，\＆Lupáñez，2006），our knowledge on why some communities adopt a future－in－front mapping whereas others a past－in－front mapping for time is still incomplete． For instance，very few studies have researched deaf signers＇spatio－temporal reasoning．

Sign language speakers also tend to use spatio－temporal metaphors to express time．For instance，signers＇bodies are often referred to as a deictic reference of now，and the future is signed to the front（e．g，the American Sign Language，Emmorey，2001）relative to the signer＇s body， and the past to the back（e．g．，the French Sign Language， Maeder \＆Loncke，1996；the Spanish Sign Language Pereiro \＆Soneira，2004）．At first sight，it would seem reasonable to assume that the metaphorical timelines in those sign languages would agree with the way these are used in the corresponding spoken languages．

Interestingly，there are dramatic differences in the deictic sagittal timelines between the Chinese Signed Language（CSL）and Mandarin Chinese．In Mandarin

Chinese，the sagittal space－time word＂前／qian＂indicates both the spatial concept of forward／front and the temporal concept of early／before（Yu，2012）（e．g．，＂前天／qian tian＂， literally：front day，meaning：the day before yesterday）．A case study on gestural behaviour has shown that Mandarin speakers can point to the back or front of their body to refer to the conception of before，depending on whether the language suggests an ego－moving perspective（e．g．， We are running to the future ahead．）or a time－moving perspective（e．g．，The future is coming．）（Chui，2011）． Recent quantitative research，on the other hand，finds that Mandarin speakers are more likely to gesture the past to the front when referring to temporal expressions with the sagittal space－time word（前／qian）（Gu，Mol，Hoetjes，\＆ Swerts，in prep）．Partially due to this lexical effect，some Mandarin speakers even explicitly report to believe the past to be positioned behind and the future in front of them（Gu，Zheng，\＆Swerts，2016）．
In CSLs，however，the spatial forward and the temporal early／before are signed differently，i．e．，the sign of front is only used for the spatial concept of forward，whereas the concept of before／past is signed towards the back（e．g．， Zheng，2009；Wu \＆Li，2012）．In other words，in their lexicon，deaf signers only have the past－at－back space－ time mappings，which is different from Mandarin speakers who additionally have past－in－front mappings （Table1）．As deaf signers learn the spatial concepts earlier than the abstract concepts of time，it is plausible that if a signer has not acquired the Mandarin space－time word（前 ／qian）as a temporal past conception，s／he is likely to interpret the word as a spatial concept of forward，which is consistent with that in CSL（front in the space）．By contrast，if a signer has acquired the space－time word as a temporal past conception，s／he is likely to map the past to the back as suggested by the CLS past－at－back mappings， or s／he may also establish new space－time mappings with the past in the front，similar to Mandarin speakers．

Table 1：Differences between Mandarin Chinese and Standard CSL in sagittal spatio－temporal metaphors．

| Mandarin | Front（space） | The day before <br> yesterday（time） |
| :--- | :--- | :--- |
|  | 前面（front surface） | 前天（front day） |
|  | One hand with the <br> index finger extended， <br> point to the very front． | The ingers point to the |

Note：The spatial concept of front in Mandarin is consistent with that in CSL．Figures of signs are reproduced from the CSL， 2003

Given the cross－linguistic differences in space－time metaphors between Mandarin Chinese and CSL，and given that learning a new category of spatial metaphors for time may influence one＇s mental representation of time（Boroditsky，2001），this paper aims to study（1） whether the differences in space－time metaphors between Mandarin Chinese and CSL influence Chinese deaf signers＇understanding of time；（2）in the context of

Chinese culture，whether the acquisition of Mandarin sagittal spatial metaphors leads Chinese deaf signers to a change in space－time mappings．To this end，we have conducted two studies：study 1 used a clock question to test how Chinese deaf signers interpret the sagittal space－ time word 前／qian（spatial front or temporal before）；study 2 used a temporal diagram task to explicitly examine Chinese deaf signers＇space－time mapping．

## Study 1：The Clock Question

## Method

## Participants

15 deaf signers $(\mathrm{F}=8)$ from Rizhao Special Education School participated in the experiment．They were fluent users of Standard CSL．They studied in different grades at school，ranging from the $4^{\text {th }}$ grade to $9^{\text {th }}$ grade（ $M=7.5$ ）． Their mean age was 17.6 years（ $S D=2.9$ ）．The average hearing loss was moderate－severe，as reported by the signers themselves and their teacher（ $M=3.8,1$－Slight，2－ Mild，3－Moderate，4－Moderately Severe，5－Severe，6－ Profound）．Permission was granted to the investigators to have access to the participants＇Mandarin Chinese exam scores from the record of their last end－term exam．

## Materials and Procedure

Singers were given a questionnaire to fill in personal information and family background．The instructions were not in sign language but in written Mandarin．In the middle of the questionnaire，there was a clock question （Table 2）．The sagittal space－time word 前qian（literally front／forward or temporally before）in this question is somewhat ambiguous in meaning though mainly used as a temporal expression．Most Mandarin speakers will interpret the question as moving the clock one hour beforelearlier，thus answering the question as 12 AM （Lai \＆Boroditsky，2013）．However，if deaf signers think of the space－time word（前／qian）as a spatial front，then they are likely to move the clock one hour forward，thus giving 2 PM as an answer．It is also assumed that deaf signers of higher Mandarin proficiency levels are more likely to interpret the space－time word as a temporal past，as opposed to signers of lower Mandarin proficiency levels．

Table 2：The clock question in Mandarin and English．


## Data Analysis

Data of two participants were excluded from the analysis， as they did not fully complete the questionnaire．As a dependent variable，we counted participants＇responses to the clock question（answer： 12 AM or 2 PM ）．

We would discuss below how those responses were moderated by possible factors．The first and most important factor was participants＇Mandarin proficiency level．It was mainly measured by the school grade level in which a deaf signer was studying（grade），as a deaf signer studying in a higher grade was expected to have a higher Mandarin proficiency level than a signer studying in a lower grade．Second，signers＇Mandarin exam score（exam score）was used to supplement the proficiency measurement，albeit that the exam papers and intrinsic difficulty of tests were different across grades．

Additionally，given that age can influence individual＇s sagittal spatial－temporal reasoning（de la Fuente et al．， 2014），we controlled for age as a possible factor． Participants＇hearing loss and their parents＇deafness （deaf parents）were also considered to be factors that may influence participants＇space－time mappings．

## Results and Discussion

About $70 \%$ of participants（ 9 out of the 13 deaf signers） responded according to the spatial understanding of the word＂前／qian＂（forward），giving 2 PM as an answer．In comparison to the $13 \%$（ 3 out of 24）of Mandarin monolinguals in Lai \＆Boroditsky（2013）＇s study， Chinese deaf signers were significantly more likely to give an answer of 2 PM than Mandarin monolinguals （Fisher exact test，$p=.001$ ，Odds Ratio $=15.75,95 \% \mathrm{CI}=$ ［2．91，85．22］）．Given that these deaf signers have already learned Mandarin temporal conceptions in low grades， this indicated that participants may still be influenced by the spatial sign of front from their CSL．

Furthermore，we tested whether Mandarin proficiency influenced signers＇understanding of the space－time word （前／qian）．The results showed that the factor grade was significant $(\beta=.387, t=3.01, p=.020,95 \% \mathrm{CI}=$ ［．083，．691］），while controlling for the other factors exam score，age，deaf parents and hearing loss（Table 3）．This indicated that those higher graders were more inclined to interpret the space－time word（前／qian）as temporal before （ 12 AM）．Assuming that higher graders are likely to have higher Mandarin proficiency levels than lower graders， the effect of grade suggests that signers＇Mandarin proficiency levels play a role in shaping their understanding of the conceptions of the space－time word （前／qian）．A seemingly contradictory finding is that exam score was not significant $(\beta=-.0002, t=-.020, p=.985$ ， $95 \% \mathrm{CI}=[-.019, .019])$ ，keeping all other variables constant．This might be due to the fact that there was only limited variation in Mandarin proficiency within a grade．

Table 3：Results of the clock question．

| clock | Coef． | $\mathbf{t}$ | $\mathbf{P > t}$ | ［ 95\％CI ］ |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| grade | .387 | 3.01 | $\mathbf{. 0 2} * *$ | .083 | .691 |
| exam＿score | -.0002 | -.02 | .985 | -.019 | .019 |
| age | -.029 | -0.69 | .515 | -.129 | .071 |
| deaf＿parents | -.176 | -0.75 | .479 | -.731 | .380 |
| hearing＿loss | -.231 | -1.84 | .108 | -.528 | .066 |
| ＿cons | -1.20 | -1.76 | .122 | -2.83 | .413 |
| Note：${ }^{*} p<.1,{ }^{* *} p<.05$ |  |  |  |  |  |

The fact that signers with lower Mandarin proficiency levels were more likely to give an answer of 2 PM may be caused by their use of spatial reference of front（primarily be triggered by lexical cues），though this does not necessarily imply that they also explicitly conceptualise the future as in front of them．Study 2 investigated the Chinese deaf signers＇sagittal space－time mappings using a more explicit temporal diagram task．

## Study 2：A Temporal Diagram Task

## Method

## Participants

All participants in study 1 took part in study $2 .{ }^{1}$

## Materials and Procedure

Participants performed a temporal diagram task（de la Fuente et al．，2014，Experiment 1），which has been adapted and used in Gu，Zheng，and Swerts（2016）＇s study． They sat at a table and saw a toy doll（named Xiaoming） with one box behind and one box in front of it． Participants and the character faced the same sagittal direction（Fig．1）．Participants were provided with a written instruction in which they could read that the day before yesterday（前天／qian－tian，tr．front day）Xiaoming went to visit a friend who liked eating apples，and the day after tomorrow（后天／hou－tian，tr．back day）he would be going to visit a friend who likes eating pears．Participants were given an apple and a pear and were instructed to put the apple in the box that corresponded to the past（以前 ／yi－qian，tr．to front）and the pear to the box that corresponded to the future（今后／jin－hou，tr．now back）． The mentioning order of the apple and pear and the way they were paired with the day before yesterday or the day after tomorrow were counterbalanced．Note that there was no ambiguity in the interpretation of the space－time words in this instruction（cf．study 1），e．g．，the concept of the space－time expression＂前天／qian－tian＂，tr．front day can only be interpreted as the day before yesterday．


Figure 1：Schematic setting up of the Experiment reproduced from Gu，Zheng，\＆Swerts（2016）．

Following Gu，Zheng，\＆Swerts（2016）＇s procedure，we asked participants to perform the task with real entities rather than doing it on paper（cf．de la Fuente et al．，2014）．

[^84]This can minimise the potential projection of vertical timelines into the sagittal axis (as Chinese can conceptualise time vertically, mapping the $u p$ and down to the time conceptions of early and late, e.g., Boroditsky 2001; Gu, Mol, Hoetjes, \& Swerts, 2017). Participants were tested individually in Rizhao, China, and all instructions were not in sign language but in written Mandarin Chinese. After all tasks, they were given a small token of appreciation and signed a consent form.

## Data Analysis

In total, data of fourteen participants were used in the analysis (Data from a $4^{\text {th }}$ grader was excluded as she was helped during the task). The dependent variable was participants' responses towards space-time mappings (past-in-front or past-at-back).

As was the case with the previous experiment, we again controlled for possible factors such as participants' exam score, grade, age, hearing loss and deaf parents.

## Results and Discussion

$42.9 \%$ of participants responded according to the past-infront mapping, placing the past event in the box in front of the character and the future event in the box behind it. Although this rate was not significantly different from $50 \%, p=.79$, Odds Ratio $=.75,95 \% \mathrm{CI}=[.18, .71](N=$ 14), we expect it to be significant with a larger sample size. It is unlikely that deaf signers randomly performed the space-time mappings by chance, as shown below.

As we further examined the relationship between signers' Mandarin proficiency and their responses towards space-time mappings, controlling for deaf parents, age and hearing loss, the results showed that grade and exam score were significantly positive (Table 4). Specifically, first, higher graders had a stronger tendency to perform past-in-front mappings ( $\beta=.34, t=2.61, p=.031,95 \%$ $\mathrm{CI}=[.039, .641])$, keeping all other variables constant. Second, those who had higher Mandarin exam scores were more inclined to respond towards past-in-front mappings ( $\beta=.008, t=2.04, p=.075$ (two-tailed), $95 \%$ $\mathrm{CI}=[-.001, .017])$, ceteris paribus. The results indicated that Mandarin proficiency has an effect on signers' spacetime mappings, both between different grades and within a grade. In other words, despite the fact that there are only past-at-back spatio-temporal signs in CSL, deaf signers can gradually establish the Mandarin past-in-front spacetime mappings during their learning process of Mandarin.

Table 4: Results of the temporal diagram task.

| pastfront | Coef. | t | $\mathrm{P}>\mathrm{t}$ | [ 95\% CI ] |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| grade | .340 | 2.61 | $\mathbf{. 0 3 1 * *}$ | .0391 | .641 |
| exam_score | .008 | 2.04 | $\mathbf{. 0 7 5} *$ | -.001 | .017 |
| age | -.047 | -1.53 | .163 | -.119 | .024 |
| deaf_parents | -.477 | -2.04 | $\mathbf{. 0 7 5} *$ | -1.015 | .062 |
| hearing_loss | -.051 | -.37 | .724 | -.372 | .270 |
| _cons | -1.52 | -4.38 | .002 | -2.32 | -.722 |
| Note: ${ }^{*} p<.1, * * p<.05$ |  |  |  |  |  |

Additionally, those signers whose parents were deaf
were less likely to perform past-in-front mappings ( $\beta=$ $.48, t=-2.04, p=.075$ (two-tailed), $95 \% \mathrm{CI}=[-$ $1.015, .062]$ ), ceteris paribus. The results suggested that deaf parents may influence deaf children's space-time mappings. This is plausible, as deaf children may often be exposed to the past-at-back temporal signs performed by their deaf parents. Consequently, they may be more likely to have past-at-back space-time mappings than their counterparts with non-deaf parents.

## General Discussion

In study 1 , we used a clock experiment to examine how Chinese deaf signers interpreted Mandarin spatial metaphor of time. We observed effects of both CSL and learning Mandarin Chinese on their understanding of time. There is a co-activation of signs even in the non-signing linguistic contexts, whereas within the signers' group, those with higher Mandarin proficiency levels were more likely to interpret the space-time word 前/qian as temporal before (like Mandarin speakers). Our results suggest that language transfer occurs across modalities (i.e., a spoken language and a sign language, cf. bimodal bilinguals, Emmorey, Borinstein, Thompson, \& Gollan, 2008).

Alternatively, the results can also be explained in terms of differences in time perspective-taking (e.g., Gentner, Imai, \& Boroditsky, 2002; Moore, 2011; Núñez, Motz, \& Teuscher, 2006; Walker, Bergen, \& Núñez, 2017), which would be consistent with claims of a previous study on Mandarin-English sequential bilinguals (Lai \& Boroditsky, 2013). Similar as in CSL, in English the spatial front usually does not have a meaning of temporal before. That study found that Mandarin-English speakers were influenced by English when answering the clock question in Mandarin, such that Mandarin-English speakers were less likely to answer the clock question as 12 AM , in comparison with what Mandarin monolinguals did According to Lai and Boroditsky (2013), Mandarin speakers mostly take the time-moving perspective (12 AM), whereas English speakers mostly take the egomoving perspective ( 2 PM ). If monolingual signers of CSL mainly take the ego-moving time perspective in deictic time, it is possible that they gradually gain the time-moving perspective after learning Mandarin Chinese.

In study 2, we used a temporal diagram task to explicitly test deaf signers' sagittal space-time mappings. We found that some singers performed past-in-front space-time mappings. Given that Mandarin speakers also have past-in-front mappings (Gu, Zheng, \& Swerts, 2016), the pattern of signers' space-time mappings may be due to a characteristic of the Chinese culture, in which people give more importance to tradition and focus more on the past (Guo, Ji., Spina, \& Zhang, 2012), analogous to what appears to be true for Moroccans. However, within the Chinese culture, we found that the extent to which signers performed past-in-front mappings was positively related to their Mandarin proficiency. Similar to the results of the clock question, we found effects of Mandarin proficiency on Chinese signers' spatio-temporal reasoning, which suggests that learning a novel linguistic spatial metaphor for time may foster a new way of thinking about time (Boroditsky, 2001; Hendricks \& Boroditsky, 2015).

Future studies can further examine this using a non－ linguistic task（e．g．，Fuhrman \＆Boroditsky，2010）．

Alternatively，according to the temporal－focus hypothesis（de la Fuente et al．，2014），cultural attitudes towards time exert an important influence on people＇s space－time mappings．One may argue that the typical Chinese culture is more past－focused than that of the Chinese deaf culture，although this needs a further survey． Given such an assumption，signers may gradually adjust themselves into the mainstream Chinese culture and hence become more similar to the Mandarin speakers．Future study can additionally control for signers＇temporal－focus of attention to corroborate the present findings．
Note that in Standard CSL，there are no sign metaphors that reflect past－in－front space－time mappings but only signs for past－at－back mappings．It would therefore be ideal if we could supplement the current set of results with those obtained from a control group of monolingual deaf signers of CSL to provide stronger evidence that deaf signers indeed think of the past as being situated at the back，though，practically，we can hardly find a group of deaf signers who do not know Mandarin Chinese．Future research may also study illiterate hearing Mandarin speakers to at least examine the effects of written Mandarin proficiency on people＇s spatial－temporal conceptualisations．By contrast，in our study，we found that a certain proportion of Chinese deaf signers put the entity corresponding to the past in the front．Additionally， the effects of exam score and study grade clearly suggest that Chinese deaf signers can gradually＂learn＂to have past－in－front mappings as a function of an improved Mandarin proficiency．This is an intriguing finding as it shows that within the Chinese culture，learning a spatial metaphor in a different modality can still influence people＇s mental representations of time．
Furthermore，the past－in－front mappings performed by the deaf signers in the temporal diagram task can be argued to be a consequence of a direct translation of the spatial conception of front in the Standard CSL，thus characterising the results as merely an effect of language interferences without reference to the differences in spatio－temporal reasoning．For example，participants may simply interpret the sagittal space－time word（前／qian）as front in space rather than understanding the concept of space－time expression（前天／qian－tian，front day）as the past conception of the day before yesterday，though the conception of front day is not ambiguous at all．This is， however，quite unlikely．First，the instructions were checked beforehand by their teacher to ensure that those participants have previously learned all the vocabulary and would be able to understand the sentences and the concept of front day．Second，if deaf signers would have done a direct translation，those signers of lower Mandarin proficiency levels should be more likely to translate the space－time word（前／qian）as front，thus would produce a larger proportion of past－in－front mappings．However， quite on the contrary，we found that deaf signers of lower Mandarin proficiency levels or studying in lower grades were actually more inclined to perform past－at－back mappings，which was consistent with the CSL where the past is signed towards the back．This indicates that
participants even with a low Mandarin proficiency can already understand that front day is a temporal concept． Therefore，needless to say for the higher proficient group， the tendency of having past－in－front space－time mappings likely reflects their spatio－temporal reasoning．

Moreover，it is possible that deaf signers have to rely on their vision heavily as a result of the hearing loss． Consequently，this may trigger a stronger effect to consider things that they have seen in front of them as the past whereas the events that have not seen as the future behind them（cf．Aymara speakers，Núñez \＆Sweetser， 2006）．Apparently，this explanation does not hold for deaf people universally，as deaf users of many other sign languages（e．g．，ASL，FSL）do not exhibit a tendency towards past－in－front space－time mappings．

Additionally，we conducted both studies in Mandarin Chinese rather than in CSL．It would be interesting to ask the deaf signers to fulfil the temporal diagram task with a sign language instruction，the results of which probably can also reveal the effect of Mandarin Chinese on deaf signers＇spatio－temporal reasoning，even when signers think in CSL．Possibly，participants may be visually primed by the spatial movements of the signs in the instruction，for example，for the clock question the CSL will give a strong hint where the clock hand is moving in the signs（either a clockwise or an anti－clockwise movement）．This will not allow us to examine signers＇ authentic interpretation of the sagittal space－time word． Furthermore，the sign for front day（the day before yesterday）is signed as two fingers pointing to the back of the body，hinting a past－at－back space－time mapping．

## Conclusions

In the current study we investigate whether the Mandarin sagittal space－time metaphors influence Chinese deaf signers＇spatio－temporal reasoning．In two experiments， we found that signers with higher Mandarin proficiency were more likely to interpret the Mandarin space－time word（前／qian）as temporal before（Study 1），and to perform past－in－front space－time mappings（Study 2），in comparisons to signers with lower Mandarin proficiency． These findings not only provide within－culture evidence for the influence of language on thought（cf．Boroditsky， 2001；Hendricks \＆Boroditsky，2015），but also demonstrate that even cross－modal space－time metaphors can have an impact on signers＇spatio－temporal reasoning．

## Acknowledgements

The first author received financial support from the NWO Promoties in de geesteswetenschappen（322－89－007）， which is gratefully acknowledged．

## References

Bender，A．，\＆Beller，S．（2014）．Mapping spatial frames of reference onto time：A review of theoretical accounts and empirical findings．Cognition，132，342－382．
Bottini，R．，Crepaldi，D．，Casasanto，D．，Crollen，V．，\＆ Collignon，O．（2015）．Space and time in the sighted and blind．Cognition，141，67－72．

Boroditsky，L．（2000）．Metaphoric structuring： Understanding time through spatial metaphors． Cognition，75，1－28．
Boroditsky，L．（2001）．Does language shape thought？： Mandarin and English speakers＇conceptions of time． Cognitive Psychology，43，1－22．
Cabeza－Pereiro，C．，\＆Fernández－Soneira，A．（2004）．The expression of time in Spanish Sign Language（LSE）． Sign Language \＆Linguistics，7，63－82．
Casasanto，D．，\＆Bottini，R．（2014）．Mirror reading can reverse the flow of time．Journal of Experimental Psychology：General，143，473－479．
Casasanto，D．，\＆Boroditsky，L．（2008）．Time in the mind：Using space to think about time．Cognition，106， 579－593．
China Association of the Deaf and Hard of Hearing （2003）．Chinese Sing Language（中国手语）．Beijing： Huaxia Publishing House．
Chui，K．（2011）．Conceptual metaphors in gesture． Cognitive Linguistics，22，437－458．
Clark，H．（1973）．Space，time semantics，and the child．In T．E．Moore（Ed．），Cognitive Development and the Acquisition of Language（pp．27－63）．New York： Academic Press．
De la Fuente，J．，Santiago，J．，Román，A．，Dumitrache，C．， \＆Casasanto，D．（2014）．When you think about it，your past is in front of you：How culture shapes spatial conceptions of time．Psychological Science，25，1682－ 1690.

Emmorey，K．（2001）．Space on hand：The exploitation of signing space to illustrate abstract thought．In Merideth Gattis（ed．），Spatial Schemas and Abstract Thought（pp． 147－174），Cambridge，MA，US：MIT Press．
Emmorey，K．，Borinstein，H．B．，Thompson，R．，\＆Gollan， T．H．（2008）．Bimodal bilingualism．Bilingualism： Language and cognition，11，43－61．
Fuhrman O．，Boroditsky L．（2010）．Cross－cultural differences in mental representations of time：evidence from an implicit nonlinguistic task．Cognitive Science， 34，1430－1451．
Gentner，D．，Imai，M．，\＆Boroditsky，L．（2002）．As time goes by：Evidence for two systems in processing space time metaphors．Language and Cognitive Processes，17， 537－565．
Gu，Y．，Mol，L．，Hoetjes，M．，\＆Swerts，M．（2017）． Conceptual and lexical effects on gestures：the case of vertical spatial metaphors for time in Chinese． Language，Cognition and Neuroscience，1－16．
Gu，Y．，Zheng，Y．，\＆Swerts，M．（2016）．Which is in front of Chinese people：Past or future？A study on Chinese people＇s space－time mapping．In Papafragou，A．， Grodner，D．，Mirman，D．，\＆Trueswell，J．C．（Eds．）． Proceedings of the 38th Annual Conference of the Cognitive Science Society（pp．2603－2608）．Austin，TX： Cognitive Science Society．
Guo，T．，Ji．，L．，Spina，R．，\＆Zhang，Z．（2012）．Culture， temporal focus，and values of the past and future． Personality and Social Psychology Bulletin，38，1030－ 1040.

Hendricks，R．K．\＆Boroditsky，L．（2015）．New space－ time metaphors foster new mental representations of time．In Noelle，D．C．，Dale，R．，Warlaumont，A．S．，

Yoshimi，J．，Matlock，T．，Jennings，C．D．，\＆Maglio，P． P．（Eds．）Proceedings of the 37th Annual Cognitive Science Society，（pp．902－907）．Austin，TX：Cognitive Science Society．
Lai，V．T．，\＆Boroditsky，L．（2013）．The immediate and chronic influence of spatio－temporal metaphors on the mental representations of time in English，Mandarin， and Mandarin－English speakers．Frontiers in Psychology，4，1－10．
Lakoff，G．，\＆Johnson，M．（1980）．Metaphors we live by． Chicago，IL：University of Chicago Press．
Maeder，C．，\＆Loncke，F．（1996）．Spatial，temporal and temporo－logical notions in French Sign Language： Comparative study of deaf and hearing subjects．Sign Language Studies，90，38－51．
Miles，L．K．，Nind，L．K．，\＆Macrae，C．N．（2010）． Moving through time．Psychological Science，21，222－ 223.

Moore，K．E．（2011）．Ego－perspective and field－based frames of reference：Temporal meanings of FRONT in Japanese，Wolof，and Aymara．Journal of Pragmatics， 43（3），759－776．
Núñez，R．，Cooperrider，K．，Doan，D，Wassmann，J． （2012）．Contours of Time：Topographic construals of past，present，and future in the Yupno valley of Papua New Guinea．Cognition，124，25－35．
Núñez，R．，\＆Cooperrider，K．，（2013）．The tangle of space and time in human cognition．Trends in Cognitive Sciences，17（5），220－229．
Núñez，R．，Motz，B．，\＆Teuscher，U．（2006）．Time after time：The psychological reality of the ego－and time－ reference－point distinction in metaphorical construals of time．Metaphor and Symbol，21，133－146．
Núñez，R．，\＆Sweetser，E．（2006）．With the future behind them：Convergent evidence from Aymara language and gesture in the cross－linguistic comparison of spatial construals of time．Cognitive Science，30，401－450．
Saj，A．，Fuhrman，O．，Vuilleumier，P．，\＆Boroditsky，L． （2014）．Patients with left spatial neglect also neglect the ＂Left Side＂of time．Psychological Science，25，207－214．
Torralbo，A．，Santiago，J．\＆Lupáñez，J．（2006）．Flexible conceptual projection of time onto spatial frames of reference．Cognitive Science，30，745－757．
Ulrich，R．，Eikmeier，V．，de la Vega，I．，Fernandez，S．R．， Alex－Ruf，S．，\＆Maienborn，C．（2012）．With the past behind and the future ahead：Back－to－front representation of past and future sentences．Memory and Cognition，40，483－495．
Walker，E．J．，Bergen，B．K．，\＆Núñez，R．（2017）．The spatial alignment of time：Differences in alignment of deictic and sequence time along the sagittal and lateral axes．Acta Psychologica，175，13－20．
Wu，L．，\＆Li，H．（2012）．Time－space metaphors in Chinese Sign Language（中国手语中的时间空间隐喻）． Chinese Journal of Special Education，150，25－29．
$\mathrm{Yu}, \mathrm{N}$ ．（2012）．The metaphorical orientation of time in Chinese．Journal of Pragmatics，44，1335－1354．
Zheng，X．（2009）．A study of signs for non－visual concepts in Shanghai Variety of Chinese Sign Language（上海手语非视觉概念表达研究）．PhD Dissertation．Fudan University，China．Yu，N．（2012）．

# Language and Spatial Memory in Japanese and English 

Harmen Gudde<br>University of East Anglia

## Kenny Coventry

University of East Anglia


#### Abstract

Demonstratives are among the most frequent words in all languages, but demonstrative systems vary considerably between languages. In two experiments, we tested demonstrative use and the influence of demonstratives on spatial memory in Japanese and English - languages with purportedly very different demonstrative systems. Participants engaged in a 'memory game', tapping their use of demonstratives to describe objects located on a sagittal plane (Experiment 1 ) and the influence of demonstratives on memory for object location (Experiment 2). In addition to distance from speaker, the experiments also manipulated the position of a conspecific (next to or opposite participants). Distance and position of conspecific both affected demonstrative choice and memory in Japanese, with similar effects in English even though English does not explicitly encode the position of a conspecific. We discuss possible universals underlying demonstrative systems and the influence of language on memory.


# A Computational Model for the Dynamical Learning of Event Taxonomies 

Christian Gumbsch (christian@gumbsch.de)<br>Sebastian Otte (sebastian.otte@uni-tuebingen.de)<br>Martin V. Butz (martin.butz@uni-tuebingen.de)<br>Chair of Cognitive Modeling, Department of Computer Science and Department of Psychology, Faculty of Science, Eberhard Karls University of Tübingen, Tübingen, Germany


#### Abstract

We present a computational model that can learn event taxonomies online from the continuous sensorimotor information flow perceived by an agent while interacting with its environment. Our model implements two fundamental learning biases. First, it learns probabilistic event models as temporal sensorimotor forward models and event transition models, which predict event model transitions given particular perceptual circumstances. Second, learning is based on the principle of minimizing free energy, which is further biased towards the detection of free energy transients. As a result, the algorithm forms conceptual structures that encode events and event boundaries. We show that event taxonomies can emerge when the algorithm is run on multiple levels of precision. Moreover, we show that generally any type of forward model can be used, as long as it learns sufficiently fast. Finally, we show that the developed structures can be used to hierarchically plan goaldirected behavior by means of active inference.


Keywords: event models; object interaction; predictive encoding; event segmentation; free energy; active inference; event taxonomy; concept learning

## Introduction

Event segmentation theory (EST) (Zacks \& Tversky, 2001; Zacks, Speer, Swallow, Braver, \& Reynolds, 2007) postulates that humans automatically structure the stream of sensory and sensorimotor information into meaningful events. Events are defined as "a segment of time at a given location that is conceived by an observer to have a beginning and an end" (Zacks \& Tversky, 2001, p. 3). This definition is formulated rather broadly, containing long, abstract events (e.g. 'going on a vacation'), more concrete events (e.g. 'taking a taxi to the airport'), and very short events (e.g. 'grasping something'). Zacks and Tversky (2001) suggest that events can be organized in event taxonomies, where abstract events consist of multiple, more concrete events.

In Gumbsch, Kneissler, and Butz (2016) we proposed a system that learns events and event boundaries from the sensorimotor stream an agent experiences while interacting with its environment. Following EST, we defined predictive events (Zacks et al., 2007) as sets of forward models, i.e. internal models that predict the sensorimotor consequence of an agent's actions. Moreover, event boundaries were defined as the determining features of a situation that typically lead to a transition between events, that is, between active forward models (Butz, 2016). Events and event boundaries are encoded in a Bayesian fashion, whereby learning and updating can be closely related to the free energy minimization principle (Friston, 2009). In addition, though, the detection of event
boundaries, and the consequent possible transition to another event model, was biased towards the detection of transient free energy signals. This is closely related to the transient error principle of Zacks et al. (2007), according to which event boundaries are characterized by transient increases in the perceptual prediction error. The result is an algorithm that learns event boundaries from "surprising" perceptions. We have shown that the algorithm learns a compositional conceptual structure of the experienced environment. Algorithmically, the system monitors the current prediction error of the active forward models and registers surprise when the encountered prediction error surpasses an adaptive confidence threshold. As a result, the set of active models is changed, the relevant features that characterize the encountered event boundary are identified, the event model transitions are memorized, and new event models may be learned. We have shown that the developing model can be used to both predict sensorimotor changes and plan in a goal-directed manner.

In this paper, we generalize the established mechanism, making it more noise robust and more generally applicable beyond linear models and recursive least squares updating. Moreover, we show that the surprise signal and the noise level determine the granularity of the event segmentation. As a result, an event hierarchy can emerge naturally during this process. Finally, we explicitly derive the goal-directed planning process from formalizations of active inference (Friston, 2009; Friston et al., 2015), showing that the developing event and event boundary models are very well-suited for the invocation of inverse, hierarchical, goal-directed behavior.

## Architecture

The system architecture $S$ consists of three continuously interacting main components $\mathcal{S}=\{\mathcal{M}, \mathcal{B}, \mathcal{P}\}$. The set of event models $\mathcal{M}$ comprises all so-far learned temporal forward models. At every point in time, a subset of forward models is active, simulating and predicting the spatiotemporal changes in the environment. These spatiotemporal simulations are determined in the predicted perceptual space $P$. Event transitions are detected based on statistical evaluations of the prediction error, detecting free energy transients. They are stored in the event boundary models $\mathcal{B}$, where each model attempts to identify the critical sensory features that allow the probabilistic prediction of an event transition.

Figure 1 and Figure 2 show the three components in interaction with the outside environment and the internal moti-


Figure 1: Illustration of the system during forward modeling. An active event model predicts the perception based on the next action. The predictions are compared with the real observations and the event model is updated based on the deviation. If this deviation exceeds a threshold, a surprise is detected, the event boundary models are updated, and a transition in the event models may be triggered. Rewards experienced during transitions are registered in the motivational system and are associated with the respective event boundary.
vational system during prediction and during planning mode, respectively. While interacting with the environment, the architecture generates temporal forward predictions and learns from the registered errors (Figure 1). Inversely, the architecture can generate active inference-driven probabilistic planning by activating event boundaries as goals and inversely propagating these goals into $\mathcal{P}$ and $\mathcal{M}$, leading to the generation of goal-directed actions based on the believed knowledge about the environment (Figure 2). Due to the developing event-based predictive environmental model, hierarchical, conceptual planning becomes possible.

## Model components and functionality

Event models $m \in \mathcal{M}$ are encoded as sets of $N$ forward models, given an $N$-dimensional observational space. At a certain point in time $t$ one event model $m(t)$ is active, with $m(t)=\left(m_{1}^{v}, m_{2}^{v}, \ldots, m_{N}^{v}\right)$, where $m_{i}^{v}$ references the currently active forward model component with respect to a particular dimension $i$. Each active forward model $m_{i}^{v}$ predicts the changes in one dimension of the sensory observation $\Delta o_{i}^{\prime}(t)$, given a particular action $a(t)$. After executing $a(t)$, the real observation $o(t+1)$ is compared with the predicted observation $o^{\prime}(t+1)=o(t)+\Delta o^{\prime}(t)$. As a result, the active forward models in $m(t)$ are updated based on the error signal to improve the respective forward models. To maintain minimal additional statistics, each forward model $m_{i}^{v}$ stores the moving average (currently fixed over the last 100 steps) of its prediction error $\bar{e}\left(m_{i}^{v}\right)$, and the moving average of the variance of that prediction error $\bar{\sigma}^{2}\left(m_{i}^{v}\right)$, thus estimating the first three moments of the model's predictions.

Event boundaries $b \in \mathcal{B}$ are represented as top-down generative predictive models, which probabilistically predict the
perceptual features of the environment that are relevant for enabling or causing the transition from one particular event $v$ to another one $\mu$, i.e. $P\left(o \mid m_{i}^{v} \rightarrow m_{i}^{\mu}\right)$. Thus, the models currently assume that all relevant information for the occurrence of an event boundary is observable when an event model transition occurs. In our current implementation, we model these event transitions by means of multivariate Gaussians.

As a result, at any point in time $t$ the generative model of the system is in a particular state $s(t)=$ $\left\{m(t), P(\mathcal{B}), \mathbf{o}(\mathbf{t}), \mathbf{o}^{\prime}(\mathbf{t})\right\}$, where $m(t)$ denotes the current vector of active forward models (winners take all), $P(\mathcal{B})$ denotes the probability mass over possible event boundaries, and $\mathbf{o}(\mathbf{t})$ and $\mathbf{o}^{\prime}(\mathbf{t})$ denote observation densities.

## Simulation and inference-based learning

While interacting with the environment, the system develops its predictive models $\mathcal{M}$ and $\mathcal{B}$, which are updated and improved based on the comparisons between predicted and actually encountered observations in $\mathcal{P}$ (illustrated in Figure 1). Event models are updated by standard gradient descent techniques, seeing that we face a self-supervised learning problem. We evaluate recursive least squares as well as delta-rule based gradient descent-based event models.

Event boundaries are detected by a sudden, significant rise of the prediction error above the tolerated uncertainty. The prediction error $e_{i}(t)$ of sensory dimension $i$ at time step $t$ is considered "surprising" when

$$
\begin{equation*}
e_{i}(t)>\bar{e}\left(m_{i}^{v}\right)+\theta \bar{\sigma}\left(m_{i}^{v}\right) \tag{1}
\end{equation*}
$$

with $m_{i}^{\nu}$ denoting the active forward model of dimension $i$ and $\theta$ the confidence-dependent surprise threshold. This threshold essentially determines when an error is considered significant depending on the currently estimated standard deviation, i.e. the inverse confidence or precision, of the respective forward model. It modulates the granularity of the event segmentation performed by our system, which in turn strongly influences how accurately each ongoing event is predicted.

If a surprise-signal is detected, the system is allowed to switch the active forward models $m(t)$. The system enters a searching period during which the next active models are determined. To do so, all existing forward models $m_{i}$ of dimension $i$ are taken into consideration. For a fixed number of time steps (10 time steps in our simulations) each model predicts the change in observation and is updated. After this search period, if the prediction error is still considered "surprising" by all existing forward models (determined by Eq. 1), a new forward model is generated and added to the possible forward models $m_{i}$. On the other hand, if the prediction error is not surprising for at least one existing model, the forward model $m_{i}^{\mu}$ with the smallest mean error is chosen as the new forward model for dimension $i$.

To summarize, our generative model space essentially consists of a set of temporal forward models $\mathscr{M}$ and a set of probabilistic transition models $\mathcal{B}$, that is, event boundary models. To minimize free energy, the validity of each temporal forward model is optimized by gradient descent - maximizing


Figure 2: Illustration of the system during planning. Based on the system's goal, an event boundary is chosen. The required perception to reach this boundary is compared with the currently predicted perception and the necessary perceptual change is computed. If this change can be achieved by the active event model, a suitable motor command is determined. Otherwise, a new model to fulfill this change is chosen and the event boundary to reach this new model is determined, proceeding recursively.
its predictive accuracy in its applicable subspace. Combined with its current minimal statistics (first three moments), the temporal forward models can be viewed as Gaussian densities, which are optimized approximately optimally based on the incoming sensory feedback (Kneissler, Drugowitsch, Friston, \& Butz, 2015). Additionally, the approach has - as a structural prior - the assumption that temporal forward models will be typically applicable during an extended period of time and that transitions between forward models can be characterized by event boundaries (Butz, 2016). As a result, event boundaries are optimized such that free energy can be minimized equally well in all subspaces of the environment.

## Goal-directed behavior

To invoke goal-directed behavior, we add a simple "motivational" system to the architecture, which associates model states with changes in its internal motivational state (Butz, Shirinov, \& Reif, 2010). In the current implementation, we simplify this part by associating particular motivations with particular event boundaries, such as the disappearance of food due to its consumption. Thus, when goal-directed planning is invoked, a desired event boundary is activated and inversely propagated through the system's generative model.

Formally, we can infer the optimal behavioral policy $\pi$ by assigning a value to all imaginable policies given the current beliefs of the model and choosing the best one:

$$
\begin{align*}
& Q_{\tau}\left(\pi, s_{t}\right)=-E_{\pi}\left[H\left(P\left(o_{\tau} \mid s_{\tau}\right)\right)\right]- \\
& D_{K L}\left[P\left(o_{\tau} \mid s_{\tau}\right)| | P\left(o_{\tau} \mid s_{G}\right)\right] \tag{2}
\end{align*}
$$

which is based on (Friston et al., 2015). This formulation essentially states that a policy $\pi$, starting at believed situational
system state $s_{t}$, will result in the minimization of free energy by minimizing (i) predicted uncertainty over the expected unfolding successive states up to a certain temporal horizon $T$ into the future $\tau=\{t, t+1, \ldots, T\}$ measured by the entropy $H\left(P\left(o_{\tau} \mid s_{\tau}\right)\right)$ over expected states $s_{\tau}$ when pursuing policy $\pi$ and (ii) the divergence, formulated as the Kullback-Leibler divergence $D_{K L}$, between expected future observations and desired future observations given a goal state distribution $s_{G}$, i.e. $P\left(o_{\tau} \mid s_{G}\right)$.

When we fully focus behavior on maximizing reward outcomes, ignoring uncertainties in our expected progression through the environment, it is possible to cancel out the term $E_{\pi}\left[H\left(P\left(o_{\tau} \mid s_{\tau}\right)\right)\right]$ in $D_{K L}$, yielding:

$$
\begin{equation*}
Q_{\tau}^{\prime}\left(\pi, s_{t}\right)=P\left(o_{\tau} \mid s_{\tau}\right) \ln P\left(o_{\tau} \mid s_{G}\right) \tag{3}
\end{equation*}
$$

As a result, the goal is to maximize the overlap between these two probability densities.

In our model's case, $P\left(o_{\tau} \mid s_{G}\right)$ will correspond to a particular model transition $m_{g}^{\nu} \rightarrow m_{g}^{\mu}$ and all other model transitions will be set to zero for all future points in time $\tau$. As a consequence, the model essentially "wants" to maximize

$$
\begin{equation*}
P\left(m_{i}^{\vee} \rightarrow m_{i}^{\mu} \mid s_{\tau}\right) \tag{4}
\end{equation*}
$$

i.e. the probability of the model transition $m_{i}^{\nu} \rightarrow m_{i}^{\mu}$. This corresponds to first maximizing $P\left(m_{i}^{\vee} \mid s_{\tau}\right)$, that is, being in model $m_{i}^{v}$ and then maximizing the likelihood of the transition, which, in turn, corresponds to maximizing the probability of the observation that characterizes the transition, that is, $P\left(\mathbf{o}_{\tau} \mid m_{i}^{v} \rightarrow m_{i}^{\mu}\right)$. Thus, given the current state of the model $s(t)$, the inference process may directly yield motor actions that attempt to maximize $P\left(\mathbf{o}_{\tau} \mid m_{i}^{\nu} \rightarrow m_{i}^{\mu}\right)$ if this seems possible given the current event model state. However, when the event model $m_{i}^{v}$ is currently not active, then a recursive process must start that selects another model transition $P\left(m_{i}^{\nu^{\prime}} \rightarrow m_{i}^{v}\right)$ in order to reach the goal transition by first invoking an intermediate transition (illustrated in Figure 2).

As a result, hierarchical, conceptual goal-directed probabilistic inference-based planning is implemented, which invokes event boundaries as subgoals, given the final goal cannot be reached by the currently available control options. For example, the model can infer that if it wants to drink out of a mug but the mug is not in the hands, the mug first needs to be approached and grasped before it can be transported to the mouth. Note that the recursive planning procedure, in principle, can find any sequence of events and event boundaries, which are believed to lead to the goal. However, time and space as well as precision limitations may generally apply when propagating the active inference signals through the system architecture $\mathcal{S}$.

## Evaluation

To investigate the event segmentation and hierarchical planning capabilities of our system, we have chosen a testing scenario, in which a simulated agent, operating in continuous space, can interact with different objects in multiple ways


Figure 3: The evaluation scenario: Objects (here a sticky object) are generated in the white area and vanish when they enter the black, rectangular mouth area. The blue hand is able to grasp or attach objects. They are detached from the hand once they are inside the red "release-area" cube.
(Figure 3). The agent consists of a hand with three shovellike fingers and a stationary "mouth". The hand is able to move freely through a limited 3-dimensional workspace with a "release-area". Three types of differently colored objects appear in our simulation (two types of "sticky objects" and one type of "marble"). Sticky objects are big and spiky. They automatically attach to the hand upon contact. Once attached they are dragged alongside the hand until they enter the release-area, wherein they detach and drop into the mouth. We use two types of sticky objects in our simulation: light objects do not alter the hand movement when attached to it; heavy ones slow down the hand movement by a factor of $\frac{1}{4}$. Marbles are small spheres that need to be grasped to be transported by the hand. To grasp a marble the hand has to be positioned directly over a marble for a grasping reflex to activate. The fingers open again when the marble is inside the release-area. Carrying a marble is usually far more difficult to predict than dragging a sticky object, since marbles sit loosely between the fingers and shake while being transported. If an object is dropped into the mouth, it is consumed and a new object is generated. In our scenario the system is rewarded once an object is consumed. Thus, to receive rewards, the agent has to attach or grasp the present object, transport it to the release-area, and drop it into the mouth.

In every time step $t$, a motoric action $a(t)$ is performed and a sensory observation $o(t)$ is perceived. In particular $o(t)$ consists of the position of the hand $x_{h}, y_{h}, z_{h} \in[-100,100]$, the position of the object $x_{o}, y_{o}, z_{o} \in[-100,100]$, and the position of the object in a hand-centered frame of reference $x_{o, h}, y_{o, h}, z_{o, h} \in[-200,200]$. Additionally $o(t)$ contains the color of the object and boolean information whether the object has spikes or not. The motor command $a(t)$ determines the change in hand position $\Delta x_{h}, \Delta y_{h}, \Delta z_{h} \in[-1,1]$. Furthermore, $a(t)$ contains the velocity of the object to enable computation of the object's next position even while it is falling.

To evaluate the system, we investigate how the learning capabilities depend on the underlying learning rule. Therefore,
we learn the forward models, currently linear prediction models, both by means of delta rule based gradient descent (learning rate $\eta=0.1$, momentum term $\alpha=0.9$, linear activation function) as well as by means of recursive least squares (RLS) (forgetting factor $\lambda=0.99$ ). RLS essentially implements an adaptive filter that minimizes the sum of squared residuals recursively in an optimal online fashion. Furthermore, we show that both the threshold $\theta$ and sensory noise influence the granularity of the determined event segmentations - allowing the formation of event taxonomies. We performed every simulation 10 times with a different random initialization.

## Results

In a first test we analyzed how the underlying learning rule of the forward models influences the event segmentation and learning accuracy of the system, comparing stochastic gradient descent with RLS. During learning, the motoric action $a(t)$ was determined by an external algorithm, which performed a mixture of random movements and behaviors leading to a new event. Every simulation consisted of 10 training and 10 test epochs, during which no forward model updates took place. In every epoch all three objects were generated once and manipulated by the agent until they were consumed.

The system was able to identify all existing events in the simulation (hand moving normally/slowly, object lying, light/heavy sticky object dragged, marble carried, object dropped) for both types of forward models used. The average prediction error of all active forward models for three different events is shown in Figure 4a. For both learning rules the system improves the prediction accuracy over the training epochs, but RLS-based learning quickly reaches a much better prediction. While this result shows that both forward model learning approaches can be applied, further tests were conducted using RLS-based learning to speed-up the process.

In a second test we analyzed how the surprise threshold $\theta$ influences the granularity of the event segmentation. Thus we performed this test with three different surprise thresholds $(\theta \in\{10,50,100\})$. Since we hypothesized that sensory noise also alters the granularity of segmentation, we additionally varied the amount of Gaussian distributed noise $(\sigma \in\{0.001,0.01,0.05,0.1\})$ that was added to each observation $o(t)$. Additionally, a small amount of Gaussian distributed noise $(\sigma=0.01)$ was added to each action $a(t)$.

The average prediction error for different events at the lowest level of noise tested $(\sigma=0.001)$ is shown in Figure $4 b$. The prediction error of the system greatly varies for the different surprise thresholds. After only a few training epochs, the prediction accuracy of the active forward models for the smallest tested surprise threshold is close to the level of sensory noise. In contrast, the mean prediction error of the system for the largest threshold is only slightly below 1 . For $\theta=50$ the prediction error varies among the different events.

To further analyze this difference in prediction accuracy we examined the number of forward models generated for the different surprise thresholds and determined which forward


Figure 4: Mean prediction error of all active forward models for different events for (a) different forward models without sensory noise; (b) different surprise thresholds $\theta$ using RLSbased forward models and Gaussian distributed sensory noise ( $\sigma=0.001$ ). Three events are shown: object lying $(\diamond)$, light object dragging ( $\square$ ), and heavy object dragging (०).
models were most active for which event. In the following, we first refer to the case with the lowest level of sensory noise ( $\sigma=0.001$ ). Later, we analyze how a variation in sensory noise influences the number of event models. All identified event models are shown taxonomically structured in Figure 5.

For $\theta=100$ two forward models developed for predicting object positions. One of these models was active when the object was lying, the other one was active for the rest of the time. Thus the system differentiated between a moving and a still object in terms of object position. For $\theta=50$ three models were predicting $x_{o}$ and $z_{o}$-position. Here the system further differentiated between slowly carrying a heavy object or transporting a light object at normal speed. For the $y_{o}$-position one additional model was generated, which was active when an object was falling. For $\theta=10$ every type of transportation was represented by a different forward model. Hence, for the smallest surprise threshold, the system even differentiated between slow and normal hand movements, while for larger values of $\theta$ this was not the case.

Sensory noise additionally influences the granularity of the performed event segmentation. While for low noise levels ( $\sigma=0.01$ and $\sigma=0.001$ ) the identified events did not differ, an increase in noise results in a coarser event segmentation. For $\sigma=0.05$ and $\theta=10$ the system used one forward model for every sensory dimension describing hand position and three forward models to predict changes in object position: one for a lying object, one for a transported object, and one for a falling object. The segmentation further coarsened for $\sigma=0.1$, where the system only distinguished between lying and moving objects. If both noise level and surprise threshold were too large ( $\sigma \geq 0.05, \theta \geq 50$ ) the system did not detect event transitions at all, such that only one forward model was generated per sensory dimension.

| $\begin{aligned} & \theta \geq 50 \\ & \theta \geq 50 \end{aligned}$ | object exists |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $\begin{gathered} \theta=10 \\ \theta=100 \end{gathered}$ | object is lying | object moves |  |  |  |
| $\theta=10$ |  | object moves alongside hand |  |  | object is <br> falling |
| $\theta=50$ |  | object moves with hand normally |  | object is dragged slowly |  |
| $\theta=10$ |  | object is dragged | object is carried |  |  |
| Real Events |  |  |  | - |  |

Sensory noise levels: $\square \sigma=0.1, \square \sigma=0.05, \square \sigma \leq 0.01$
Figure 5: The event models regarding object position identified by the system for different surprise thresholds and sensory noise levels. The column on the left states the surprise threshold $\theta$. The associated sensory noise $\in[-\sigma, \sigma]$ is illustrated by color, as explained by the color legend below the table. Each row shows the identified event model for this surprise threshold and noise level. The bottom row illustrates the real, underlying interactions found in our scenario.

In a third test we evaluated the planning capabilities of our system, using its representation of events and event boundaries to generate goal-directed behavior. Furthermore we analyzed how the granularity of the underlying event representations influences planning. Thus, we varied granularity by using three values of the surprise threshold $(\theta \in\{10,50,100\})$ under the low noise condition $\sigma=.001$. Every simulation consisted of 30 training and 30 testing epochs. During each training epoch one sticky object and one marble were generated. The goal of the system was to consume the objects. The system was given a fixed time interval ( 500 simulation steps) to interact with each object. If the system failed to remove an object in the given time period, an external algorithm performed the required movements. During testing, we introduced a new object that closely resembled the marble in its visual characteristics (small, similar color, no spikes). Once grasped, however, the new object behaved like a sticky object and attached to the hand. The challenge of successfully interacting with the new object is thus to grasp the object like a marble (by means of the appropriate event boundary models) but to then transport the object like a sticky object (with the help of the appropriate event models).

Already after the first training epoch, the system consumed $50 \%$ of the novel objects correctly when a fine-grained event segmentation was used $(\theta=10)$. From 25 training epochs onward, the system managed to successfully interact with every new object. For $\theta=50$, the system on average only interacted with $70 \%$ of the new objects correctly after all 30 training epochs. For $\theta=100$, the system did not manage
to interact with any test object successfully. The difference in performance can be explained by examining the identified events (see Figure 5). For $\theta=100$ the system does not distinguish between transporting and dropping an object, such that the 'detachment' event boundary was not learned and thus could not be used as a subgoal. Similarly, for $\theta=50$ the system does not distinguish between transporting a marble or a sticky object, such that the 'grasp' and 'attach' event boundaries were mixed - often leading to unsuccessful grasps.

## Conclusion

Based on event segmentation theory (Zacks et al., 2007) and the principle of free energy minimization (Friston, 2009), we have developed a computational model of hierarchical, behavior-grounded event segmentation. Our system uses a strictly statistical measure of "surprise" to segment the sensorimotor stream, which an agent experiences while interacting with its environment, into events encoded by temporal forward models. In a continuous, noisy simulation, our system was able to identify event models characterizing particular object interactions, e.g. 'carrying an object' or 'dropping an object'. Furthermore, the environment was structured into conceptual event and event boundary encodings, which discriminate the critical features that are crucial for the occurrence of an event, e.g., 'hand contact' is required to manipulate an object. Due to the event-based architecture, the system accomplished to plan hierarchical behavior consisting of multiple subroutines to reach desired goal states.

We showed that a change in the confidence-threshold, which determines when a transient free energy signal is considered "surprising", affects the granularity of the event segmentation. Depending on this threshold, the system accomplished to identify events for concrete object interactions, e.g. 'carrying a marble', or abstract representations of interactions, which subsume several more concrete events, e.g. 'moving an object'. Similarly, an increase in sensory noise entailed a coarser segmentation. Thus, based on these simple statistical principles, a hierarchy can emerge, similar to the event taxonomy proposed by Zacks and Tversky (2001), in which abstract events comprise several more concrete events.

The developing hierarchical organization of event models and consequent event-oriented behavior is closely related to hierarchical reinforcement learning (Botvinick \& Weinstein, 2014; Sutton, Precup, \& Singh, 1999). Similar to our event models, options predict changes in the system's state when performing a sequence of behavior without considering lowlevel steps. While the composition of options has been shown to enable the learning of complex behavior when using a predefined set of goals (Kulkarni, Narasimhan, Saeedi, \& Tenenbaum, 2016), our system determines subgoals by itself by the principle of surprise-detection.

In sum the proposed model offers a general algorithm for online, hierarchical event segmentation learning given continuous sensorimotor experiences. Two main learning biases lead to successful segmentations: first, the partition of the de-
veloping generative model into probabilistic models of events and transitions between events; second, the focused learning of event transitions based on transient free energy signals. Besides the emergence of event taxonomies, we also showed that the developing conceptual structures can be learned to invoke hierarchical, goal-directed planning and behavioral control.

Our current research aims at integrating boundedly complex, non-linear forwards models and recurrent context information. Such enhancements will be necessary to handle non-uniform motion patterns and partially only indirectly observable causes of event transitions robustly. As a result, we hope to be able to show the more general and more scalable applicability of the principles we have introduced herein.

## References

Botvinick, M., \& Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369(1655).
Butz, M. V. (2016). Towards a unified sub-symbolic computational theory of cognition. Frontiers in Psychology, 7(925).
Butz, M. V., Shirinov, E., \& Reif, K. L. (2010). Selforganizing sensorimotor maps plus internal motivations yield animal-like behavior. Adaptive Behavior, 18(3-4), 315-337.
Friston, K. (2009). The free-energy principle: a rough guide to the brain? Trends in Cognitive Sciences, 13, 293-301.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., \& Pezzulo, G. (2015). Active inference and epistemic value. Cognitive Neuroscience, 6, 187-214.
Gumbsch, C., Kneissler, J., \& Butz, M. V. (2016). Learning behavior-grounded event segmentations. In Proceedings of the 38th annual conference of the cognitive science society (pp. 1787-1792).
Kneissler, J., Drugowitsch, J., Friston, K., \& Butz, M. V. (2015). Simultaneous learning and filtering without delusions: a bayes-optimal combination of predictive inference and adaptive filtering. Frontiers in computational neuroscience, 9.
Kulkarni, T. D., Narasimhan, K., Saeedi, A., \& Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In Advances in neural information processing systems (pp. 3675-3683).
Sutton, R. S., Precup, D., \& Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 112, 181211.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., \& Reynolds, J. R. (2007). Event perception: A mind-brain perspective. Psychological Bulletin, 133, 273-293.
Zacks, J. M., \& Tversky, B. (2001). Event structure in perception and conception. Psychological Bulletin, 127, 3-21.

# Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks 

Hyowon Gweon (hyo@stanford.edu) Mika Asaba (masaba@stanford.edu), Grace Bennett-Pierre (gbp@stanford.edu)<br>Department of Psychology, Stanford University, Stanford, CA 94305


#### Abstract

The ability to reason about the difficulty of novel tasks is critical for many real-world decisions. To decide whether to tackle a task or how to divide labor across people, we must estimate the difficulty of the goal in the absence of prior experience. Here we examine adults' and preschoolers' inferences about the difficulty of simple block-building tasks. Exp. 1 first established that building time is a useful proxy for difficulty. Exp. 2 asked participants to view the initial and final states of various block-building tasks and judge their relative difficulty. While adults were near-ceiling on all trials, children showed varying levels of performance depending on the nature of the dimensions that varied across structures. Exp. 3 replicated the pattern. These results suggest that children can reverse-engineer the process of goal-directed actions to infer the relative difficulty of novel tasks, although their ability to incorporate more nuanced factors may continue to develop.


Keywords: Difficulty; Physical reasoning; Social cognition

## Introduction

We often think about how easy or difficult it is to achieve a goal. From a child trying a new jungle gym to a scientist building a research team, the ability to reason about task difficulty is critical for many real-world decisions; it informs decisions about the self (e.g., deciding whether to tackle a task or seek help), about others (e.g., understanding who needs help), and even about groups (e.g., assigning tasks in a collaborative project). Although these decisions might seem "easy", they involve more than simply remembering and retrieving our past experiences; they often require estimates and predictions about novel tasks. Such sophisticated inferences might be especially challenging for young children, who frequently face tasks they have never attempted or completed.

Indeed, having this ability does not mean that our estimates are always accurate. Even as adults, we often under- or overestimate the difficulty of certain tasks, failing to meet deadlines or suboptimally allocating time and effort. Nevertheless, our estimates are usually accurate enough to get by, suggesting that even these inaccurate estimates might be generated in systematic ways. Indeed, our accuracy and precision in estimating the difficulty of a task might improve with experience and knowledge about the task. However, the ability to predict the difficulty of a novel task (i.e., prior to the actual experience with the task) is crucial for making effective decisions about planning, learning, and even interacting with others. What are the cognitive mechanisms that underlie our ability to predict and estimate difficulty, and how does this ability develop in early childhood?

## An intuitive understanding of difficulty

Judgments about perceived task difficulty, or task-related effort, have been mostly studied in terms of its effect on achievement, motivation, and performance attribution in personality and social psychology (e.g., Atkinson, 1957; Weiner, 1966). Early work has operationalized the notion of difficulty as the subjective probability of success or the introspective assessment of required effort (e,g., Atkinson, 1957; Heider, 1958), allowing the notion of difficulty to be measured in quantifiable terms. However, these definitions could easily be intertwined with other agent-dependent concepts such as competence, ability, or intelligence. Prior developmental work has also focused on children's perception of task difficulty and its relationship to motivation and performance in formal educational contexts (Crandall, Katkovsky, \& Preston, 1962; Nicholls, 1978; Nicholls \& Miller, 1983). These studies suggest that although children around age six consider task difficulty in selecting their own goals (Heckhausen, 1967) they still have trouble differentiating objective task difficulty from agent ability (Nicholls \& Miller, 1983).

Some recent work provides indirect support for the idea that children ages 5 to 6 can differentiate objective difficulty from subjective competence. Given information about agents' decisions to pursue goals that vary in costs (i.e., climbing a high hill vs. a low hill) and subjective rewards, children infer agents' competence (subjective costs) (JaraEttinger, Gweon, Tenenbaum, \& Schulz, 2015). Children also reason about the expected costs for discovering a causal mechanism, and prefer to teach someone a toy that would be harder (i.e., require more trial-and-error) for the person to figure out on her own (Bridgers, Jara-Ettinger, \& Gweon, 2016) even though both toys are equally easy for them. These results suggest that children may be able to use the properties of the physical environment to estimate the costs of achieving a goal even without any prior experience.

Indeed, an intuitive understanding of task difficulty does not guarantee adult-like inferences. Numerous studies report children's failure in planning and problem-solving tasks that require sequential representation of task space (e.g., Tower of Hanoi; Klahr \& Robinson, 1981). While children may successfully detect explicit, perceptual cues (e.g., height of hills, number of buttons on toys), average performance of others (Nicholls, 1978), or actual subjective experiences (e.g., solving standardized test problems such as Raven's matrices; Mueller \& Dweck, 1998), they may still fail to infer the difficulty of novel tasks especially when it requires representing or simulating possible states of the world that are not readily
observable. Despite the importance of effort estimation however, children's intuitive concept of difficulty has been rarely studied in its own right. Thus the mechanisms that underlie our ability to reason about difficulty and how they develop in early childhood still remain as important open questions.
Current approach Here we explore adults' and children's ability to estimate the difficulty of novel tasks. Given the early-emerging understanding of physical events (Baillargeon, 2004; Spelke, Breinlinger, Macomber, \& Jacobson, 1992), and the costs of simple goal-directed actions (Liu \& Spelke, 2016; Csibra, 2003), our approach is to ground the basic source of difficulty in agents' interventions on the physical world. We designed a novel task that asked participants to estimate the difficulty of simple engineering goals: building block structures. We explore the idea that humans, even early in life, can estimate the difficulty of novel tasks by reasoning about (1) what physical transitions are involved in the building process, and (2) how an agent might act on the physical states to cause these transitions.

One challenge with eliciting difficulty estimates is that there is no standard metric for measuring the actual difficulty. To establish an objective "ground truth" for our tasks, we used a variable that is often used to capture the lay notion of difficulty: time needed to complete a task. In Exp. 1 we first establish that people's intuitive sense of difficulty is tightly correlated with their estimates of expected time and the actual time. In Experiments 2 and 3, we systematically vary the physical features of the block structures as well as other factors that influence properties of agents' actions in order to examine adults' and preschoolers' ability to judge relative difficulty of various building tasks.

## Experiment 1

In Exp. 1 we had two basic goals for investigating people's ability to estimate task difficulty. First, we wanted to verify that people's difficulty estimates systematically reflect a real-world property of the task that can be measured in standard metric (i.e., time). We thus recruited separate groups of participants to get (1) difficulty estimates and (2) building time estimates of various block structures, as well as their (3) actual building times, and explored the relationships among these variables. Next, we used these estimates to verify that the pairs of block structures (to be used as stimuli in subsequent experiments) varied in their relative difficulty.

## Methods

Participants Separate groups of adults were recruited for the Difficulty Estimation task ( $\mathrm{N}=57$, Age: 20-56), Time Estimation task ( $\mathrm{N}=60$, Age: 21-68), and Build task ( $\mathrm{N}=14$, Age: 18-31). The Difficulty Estimation and Time Estimation tasks were conducted on Amazon's Mechanical Turk (AMT); we excluded participants who gave identical responses on all trials (Difficulty, $\mathrm{N}=3$ ). The Build task was conducted in lab; one participant was dropped due to technical error.

Materials 28 photos (14 each for initial and final states) of various block structures were used for the task. Each structure had a photo of its initial state (e.g., scattered blocks) and final state (completed structure). Blocks were 1" plain, yellow, green, red, or blue wooden cubes. We designed seven pairs of structures that varied in specific dimensions: (1) Numberl ( 3 blocks forming a triangle vs. 10 blocks forming a circle), (2) Number 2 ( 5 blocks forming a small cross vs. 13 blocks forming a larger cross), (3) Stabilityl (10 blocks in a horizontal line vs. 10 blocks stacked vertically), (4) Stability2 (two piles of blocks divided by color (yellow and green) vs. a castle-like structure with levels of yellow and green blocks), (5) Number\&Stability (2 long green blocks stacked vertically vs. 10 plain blocks stacked vertically, height matched), (6) Probability (5 red blocks taken out of a transparent box that contained approximately $85 \%$ red and $15 \%$ blue, or $15 \%$ red and $85 \%$ blue), and (7) Process ( 2 towers of 5 blocks from an initial state that was either near-complete or very incomplete).
Procedure Participants in all three tasks viewed the same initial and final state photos, but responded to different questions depending on the task. In the Difficulty Estimation task, participants were provided examples of very "easy" and "hard" structures in the beginning to anchor them appropriately on the scale $(0-100)$. In each trial, they viewed the initial and the final state photos of a given block structure on the screen (with an arrow pointing from the initial to the final state photo to indicate the physical transition) and answered the question "How difficult would it be to do this?" with a sliding bar. In the Time Estimation task, participants saw the same example structures (presented as structures that take a short or a long time to make) to anchor them on the scale ( $0-100$ seconds); the question in each trial was "How long would it take to make this?". Two structures within a pair were presented sequentially, but the order of presentation was counterbalanced both within each pair and across all pairs.

In the Build task, the experimenter laid out blocks in front of the subject as in the initial state photo in each trial, and asked to use the blocks to create the structure shown in the final state photo. We recorded how long the subject spent building the block structure from start to finish.

## Results

First, we asked whether expected building time can be a good proxy for estimated task difficulty. Even though separate groups participated in the Difficulty Estimation and Time Estimation Tasks, these estimates were highly correlated (Fig. 2 Right: $r=.923, t=8.296, d f=12, p<.001)$. This suggests that people's intuitive sense of difficulty can be directly mapped onto estimates of time, and that actual building times may be an approximate "ground truth" for difficulty. Given this result, we then asked how well people's estimated building times reflect actual building times. Although people generally overestimated the building times (Fig. 2 Left: intercept $=13.647, t=4.227, p=.001)$, the correlation was fairly high $(r=.780, t=4.3167, d f=12, p=.001)$. Sec-


Figure 1: Stimuli used in Expts. 2-3 (final states, except for Process); Exp. 1 stimuli did not include agents. Each pair (shown as columns) had an easier structure and a harder structure. Ratio of difficulty estimates are shown beneath each trial.
ond, we verified that the difficulty of block structures within a pair was significantly different in all pairs (paired t-tests, $p$ 's $<.002$ ); differences in estimated time and actual building times were also significant (paired t-tests, $p$ 's $<.001$ ). These graded measures of difficulty also allowed us to calculate the degree to which one structure was "harder" than the other. We calculated the ratio of estimated difficulty between the two structures (higher value indicates a larger difference) and report these in Figure 1.

Collectively these results suggest that adults can make reliable difficulty estimates of individual block structures in ways that systematically reflect some objective, quantifiable aspect of these tasks (i.e., how long it takes to build the structures). Furthermore, we were able to verify that within a pair of block structures, one was clearly more difficult than the other. Despite all pairs having a clear "answer", the magnitude of the difference between the structures varied across pairs.


Figure 2: Results from Exp.1. Correlation between Actual Build Time and Estimated Build Time (left); correlation between Estimated Difficulty and Estimated Build Time (right).

## Experiment 2

In Exp. 2, we used the stimuli from Exp. 1 to ask whether adults and children can infer the relative difficulty of blockbuilding tasks. Given the results from Exp. 1, we expected adults to show high accuracy in these binary judgments, choosing the structure that was verified as "easier(harder)" in both estimated difficulty and the actual building times.

Our main goal was to examine how children's performance might differ from that of adults. Although adults' estimates indicated that all 7 pairs had a clearly "harder" structure, they varied in why the structures varied in difficulty. In Number and Stability trials, the final structures differed in their observable perceptual properties (size, height). The Number\&Stability trial was matched on these perceptual cues, making the number of actions needed to complete the task the only determining factor for difficulty. To succeed in the Probability trial, children had to understand that relative difficulty is influenced by the availability of the required blocks (thus the ease of acquiring them) even when the final structures are identical. Success in the Process trial required an understanding that the overall difficulty of a task is easier when one starts from a partially complete state.

In light of prior work reviewed above (e.g., Nicholls, 1978, and Liu \& Spelke, 2016), we could consider two extreme possibilities: preschool-aged children might fail to distinguish relative difficulty across the board, or they might successfully detect relative difficulty on all trials. However, a more plausible possibility is that children may succeed in some cases, but selectively fail on other cases. For instance, although it may be easier to detect the differences when a property of the block structures are clearly different (e.g., number, stability), children might struggle in cases where identical structures were built via different processes. In particular, in Probability and Process trials, one cannot rely on the number of blocks used in the structures or their final shapes; one must reason about the agents' actions involved in building the structures. Thus children might struggle selectively in these trials.

Indeed, it is also possible that children have a simple heuristic that difficulty depends entirely on the structure alone. Thus in Exp. 2 we added another trial: two identical sets of towers were built, but one was built by two agents (one tower each) while the other was built by a single agent. Success on this task might speak against the possibility that children fail simply because identical structures were built.

## Methods

Participants Adults ( $\mathrm{N}=45$, Age: 21-59) were recruited on AMT. An additional 13 adults were excluded because they failed the warm-up task $(\mathrm{N}=8)$ or the attention check questions ( $\mathrm{N}=5$ ). Twenty-five preschoolers (17 female, $\operatorname{Mage}(S D)$ : 4.8(.4), Range: 4.1-5.4) were recruited from a laboratory preschool. Seven additional children were excluded due to failure to respond correctly in the warm-up task $(\mathrm{N}=6)$ or experimenter error $(\mathrm{N}=1)$.

Materials The materials were almost identical to those in Exp. 1, except that the photos now showed an agent looking neutrally at the blocks (initial state) or completed structure (final state). Children viewed these photos on a 15 " Macbook Pro (using MATLAB and Psychtoolbox) and indicated their responses by placing their hands on a response pad. Adults viewed the stimuli on Qualtrics. See Fig. 1 for stimuli.

Procedure All children were tested in a quiet room, seated next to the experimenter. Half of the children were always asked to indicate the "easier" one, and the other half were always asked to indicate the "harder" one. A warm-up task ensured children understood the meaning of the word "easier(harder)"; children were first presented with two identical boxes, which the experimenter had them first push and then lift, and were asked "Which one is easier(harder) to push?" and "Which one is easier(harder) to lift?" In the main task, children saw photos of green and yellow blocks presented side by side on the laptop screen. All children were able to identify the green (yellow) blocks by placing their left (right) hand on the response pad. In subsequent test trials, children were told: "Anne and Sally were playing with blocks today." as two initial state photos were presented on the screen. The two final state photos were then revealed below the initial state photos; the experimenter pointed to each photo and said, "This is what Anne made, and this is what Sally made. One of them was easier(harder) to make. Which one was easier(harder) to make?" Agents differed across trials and unique names were used for each agent. Trial order and the side of correct response (L/R) were counterbalanced across trials.

Adults participated in an almost identical task on AMT. Similarly to children, the initial states were presented first and then the final states were revealed below these photos. The only difference was that adults read the questions on the screen and answered by clicking on the correct answer.

## Results

Adults: As expected, adults performed near-ceiling on all trials ( $p<.001$ ). See Fig.3.

4-5 year-olds: Performance did not differ by question type (easier/harder) so we collapsed the responses throughout $\left(\chi^{2}=.784, d f=1, p=.376\right)$. Children showed above-chance performance in 6 of the 8 trials (Number1 ( $77.3 \%, p=.017$ ), Number2 $(73.9 \%, p=.035)$, Stability1 $(81.0 \%, p=.007)$, Stability2 ( $90.5 \%, p<.001$ ), NumStab ( $80.0 \%, p=.004$ ), Cooperation $(81.0 \%, p=.007)$ ), while they did not show

Exp. 2: Adult and Child Forced Choice Data


Figure 3: Exp. 2 results. Average \% correct for each trial (error bars indicate $95 \%$ confidence intervals; $* * * p<.001$, ** $\left.p<.01,{ }^{*} p<.05\right)$.
above chance performance on the remaining two, Process ( $41.7 \%, p=.541$ ) and Probability ( $57.1 \%, p=.664$ ).

We asked whether children's chance-level performance was lower than other trials with similar properties. Performance on Process trial was lower than the Number trials ( $\chi^{2}$ $=14.894, d f=1, p<.001$ ), suggesting that even though the two structures differed in the overall number of actions, children failed if two identical structures were built from diffrent starting points. Similarly, children performed significantly worse in the Probability trial than the Stability trials $\left(\chi^{2}=\right.$ 20.313, $d f=1, p<.001)$; even though the two structures were made of the same number of blocks, children failed when difficulty judgment relied on the process of sampling.

To examine whether children's performance improved with age, we conducted a logistic mixed-effects model with age and trial as fixed effects and subject as a random effect. Age and the Process trial predicted children's accuracy (age: $\beta=1.523, z=2.463, p=.014$; Process: $\beta=$ $-1.765, z=-2.500, p=.013$ ), suggesting that children's accuracy improved with age but they struggled in the Process trial regardless of age. Finally, among trials where children were reliably above chance, children performed worse than adults in Number2 $\left(\chi^{2}(1)=6.98, \mathrm{df}=1, p=.008\right)$ Process $\left(\chi^{2}(1)=12.57, \mathrm{df}=1, p<.001\right)$ and Probability trials $\left(\chi^{2}(1)=17.36, \mathrm{df}=1, p<.001\right)$, and marginally for Number1 $\left(\chi^{2}(1)=3.51, \mathrm{df}=1, p=.061\right)$ and Stability1 $\left(\chi^{2}(1)=\right.$ $3.64, \mathrm{df}=1, p=.057$ ), but not in other trials).

Overall, adults and children were able to judge the relative difficulty of simple physical tasks from just the initial and the final states, without any information about the intermediate processes. It is unlikely that participants had built identical structures in the past and simply recalled their prior experiences to answer these questions. Furthermore, our results suggest that participants did not rely on simple heuristics (e.g, number of blocks, sizes of the structures); their performance was above-chance even when the number and the shape of the structures were identical (Stability1) or their shape and height were matched (Number\&Stability). These results suggest that adults and children were able to reason about the process of the physical transitions between the initial and the
final states. Children were less accurate than adults on some but not on all trials; importantly, they showed a marked diffculty to detect the differences when identical structures were built and the only determining factor was the quality of the actions involved in the building process.

## Experiment 3

Exp. 3 replicated Exp. 2 with separate groups of children and smaller number of trials per child. In addition to successes, we were interested in replicating the failures in Process and Probability trials that presumably tested a more nuanced understanding of the building process.

Participants Thirty-five preschoolers (18 female, $\operatorname{Mage}(S D)$ : 4.7(.4), Range: 4.0 - 5.4) participated in Number, Stability, and Number\&Stability trials (Group1). Another 35 children (15 female, Mage (SD) : 4.2(.7), Range: 4.0-5.8) participated in the Sampling, Process and Cooperation trials (Group2). Across groups, 17 additional children were dropped due to experimenter error ( $\mathrm{N}=7$ ), sibling interference ( $\mathrm{N}=1$ ), not speaking English ( $\mathrm{N}=2$ ), failing the warm-up task $(\mathrm{N}=6)$ or not finishing the game $(\mathrm{N}=1)$.

Materials \& Procedure The task was almost identical to Exp. 2, except that in the warm-up task children were presented with simple line drawings and indicated which was easier(harder) to make, and the photos for Sampling, Process, Cooperation trials were presented on paper ( $8.5 \times 11$ ").

Results Children's performance was highly similar to the pattern in Exp.2: Again, accuracy was above-chance on the same 6 of 8 trials (Number1 ( $77.1 \%, p=.002$ ), Number2 ( $68.6 \%, p=.041$ ), Stability1 ( $85.7 \%, p=.001$ ), Stability2 ( $74.3 \%, p=.006$ ), Num\&Stab $(85.7 \%, p=.001)$, Cooperation ( $77.1 \%, p=.002$ ); children were at chance on Probability $(62.9 \%, p=.176)$ and Process $(57.1 \%, p=.500)$ trials.

Although age did not predict performance in each group (Group $1: \beta=.583, z=1.271, p=.204$, Group $2: \beta=$ $.404, z=.952, p=.341$ ), collapsing across groups (similar in size to Exp.2), we again saw a trending relationship between age and accuracy $(\beta=.522, z=1.676, p=.094$; collapsing across all data, age was a significant predictor of accuracy ( $\beta=.693, z=2.694, p=.007$ ).

## General Discussion

In order to investigate the development of the intuitive sense of difficulty, we designed a concrete, manual activity that even young children enjoy and easily understand: building block structures. Across three experiments, we showed that (1) adults' intuitive sense of difficulty accurately reflects actual measures of difficulty (i.e., building time) in both graded estimates and binary judgments, (2) preschoolers show above-chance performance when the pair of structures varied in the expected number of required actions (due to number of blocks, stability, or the number of agents), but (3) they fail on trials in which identical trials were built, which presumably require them to reason specifically about the pro-
cess of building and the property of actions involved. Collectively, adults and children made systematic judgments about the difficulty of physical tasks from visually observing their initial and final states, without prior experience with the exact building activity or explicit information about the intermediate processes. However, children are still developing these skills throughout the preschool years and possibly beyond.

We found that although adults had a tendency to overestimate the building time, it was strongly correlated with the actual build time, suggesting that these estimates systematically reflected some "ground truth" difficulty of these tasks. Furthermore, these time estimates were tightly linked to adults' difficulty estimates. Indeed, adults' binary judgments reflected the relative difficulty of pairs of structures, resulting in near-ceiling accuracy. This was in stark contrast to children's performance, which was similar to adults in some trials but at chance on some others.

What develops, and what makes us better? One possibility is that the accumulated experience of interacting with physical objects might support a more robust understanding of the underlying physics, increasing the precision of the simulation that might be necessary for generating these intermediate processes (Battaglia, Hamrick, \& Tenenbaum, 2013). Another possibility is that experience improves children's understanding of the dynamics between the physical states and the actions required to cause appropriate transitions between these states. These are not mutually exclusive, and both might lead to more accurate representations of the intermediate processes and the effort (e.g., physical, mental) associated with these transitions. Having self-experience with objects helps infants understand others' goal-directed actions (Sommerville, Woodward, \& Needham, 2005); it is possible that self-experience continues to help adults and children in making these everyday estimates. Future work may explore whether direct experience with these building tasks increases the precision of time and difficulty estimates.

One important question here is how and when children begin to utilize different dimensions of tasks (e.g., process, probability) when making judgments about difficulty. Despite recent work showing an early-emerging sensitivity to statistical distributions of objects (Xu \& Garcia, 2008) and the process by which these objects are sampled by an agent (Gweon \& Schulz, 2011), our results suggest that preschoolers may still fail to incorporate this understanding in reasoning about the relative difficulty of agents' sampling behaviors. Children's failure on Process trials parallels school-aged children's difficulty understanding the relationship between time, speed, and distance concepts (Siegler \& Richards, 1979); when one train started to travel ahead of another train (but they travelled at equal speeds and stopped at the same place), children fail to answer that this train travelled for a shorter time. These observations are consistent with the possibility that children may struggle to discern the differences in difficulty when the tasks are highly similar in their physical properties. While these results suggest the role of a representa-
tional capacity that allows children to simulate multiple intermediate future states sequentially over time, further research is needed to understand the exact nature of their difficulty.

On the other hand, children's robust performance on most trials points to the possibility that the basic inferential ability to estimate difficulty may emerge early. Prior work has found remarkable sophistication in infants' understanding of physical events (e.g., Spelke et al., 1992; Stahl \& Feigenson, 2015), as well as their understanding of agents' actions and interventions on the physical world, both for others (e.g., Liu \& Spelke, 2016; Newman, Lockhart, \& Keil, 2010) and their own (Upshaw \& Sommerville, 2015). Thus it is possible that even younger children have the necessary inferential and representational prerequisites for an intuitive sense of difficulty that may manifest not only in their immediate motor plans but also in their predictions of future events. Due to the verbal demands (e.g., meaning of the words "easy" and "hard"), the current paradigm is unlikely to be useful for children under age 3. Future work might exploit building time (a proxy for difficulty in our tasks) in a predictive looking paradigm to address this possibility. Indeed, a time-consuming task is not always judged as harder than a less time-consuming task. Although here we looked at simple cases in which estimated difficulty directly maps onto time, it would be interesting to further investigate how objective and subjective aspects of physical effort (e.g., height of tower and an agent's building competence) as well as mental effort (e.g., careful placement of blocks) may dissociate time and difficulty estimates.

Difficulty is a difficult concept to investigate scientifically. The current work is a small step to understanding this intuitive yet incredibly complex concept. By first examining how people reason about simple, concrete tasks we may obtain clearer insights on how these intuitions arise, and how they develop into more abstract notions of difficulty that are embedded in people's lay use of this word.

## References

Atkinson, J. W. (1957). Motivational determinants of risktaking behavior. Psychological Review, 64, 359-372.
Baillargeon, R. (2004). Infants' physical world. Current Directions in Psychological Science, 13(3), 89-94.
Battaglia, P. W., Hamrick, J. B., \& Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45), 18327-18332.
Bridgers, S., Jara-Ettinger, J., \& Gweon, H. (2016). Children consider others' expected costs and rewards when deciding what to teach. In Proceedings of the 38th annual conference of the cognitive science society.
Crandall, V. J., Katkovsky, W., \& Preston, A. (1962). Motivational and ability determinants of young children's intellectual achievement behaviors. Child Development, 33, 643-661.
Csibra, G. (2003, March). Teleological and referential understanding of action in infancy. Philos Trans R Soc Lond

B Biol Sci, 358(1431), 447-458.
Gweon, H., \& Schulz, L. (2011, June). 16-Month-Olds Rationally Infer Causes of Failed Actions. Science, 332(6037), 1524-1524.
Heckhausen, H. (1967). The anatomy of achievement motivation. Academic Press: New York.
Heider, F. (1958). The psychology of interpersonal relations. New York: Wiley.
Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., \& Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. Cognition, 140, 14-23.
Klahr, D., \& Robinson, M. (1981). Formal assessment of problem-solving and planning processes in preschool children. Cognitive Psychology, 13, 113-148.
Liu, S., \& Spelke, E. S. (2016, December). Six-month-old infants expect agents to minimize the cost of their actions. Cognition, 160, 35-42.
Mueller, C. M., \& Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. Journal of Personality and Social Psychology, 75(1), 33-52.
Newman, G. E., Lockhart, K. L., \& Keil, F. C. (2010, February). "End-of-life" biases in moral evaluations of others. Cognition, 1-7.
Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. Child Development, 49(3), 800-814.
Nicholls, J. G., \& Miller, A. T. (1983). The differentiation of the concepts of difficulty and ability. Child Development, 54(4), 951-959.
Siegler, R. S., \& Richards, D. (1979, February). Development of Time, Speed, and Distance Concepts. Developmental Psychology, 15(3), 288-298.
Sommerville, J., Woodward, A., \& Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. Cognition, 96(1), B1-B11.
Spelke, E., Breinlinger, K., Macomber, J., \& Jacobson, K. (1992, October). Origins of knowledge. Psychological Review, 99(4), 605-632.
Stahl, A. E., \& Feigenson, L. (2015, April). Observing the unexpected enhances infants' learning and exploration. Science, 348(6230), 91-94.
Upshaw, M. B., \& Sommerville, J. A. (2015). Twelve-monthold infants anticipatorily plan their actions according to expected object weight in a novel motor context. Frontiers in public health, 3(1), 32.
Weiner, B. (1966). Role of success and failure in the learning of easy and complex tasks. Journal of Personality and Social Psychology, 3(3), 339-344.
Xu, F., \& Garcia, V. (2008, April). Intuitive statistics by 8 -month-old infants. Proceedings of the National Academy of Sciences, 105(13), 5012-5015.

# Categorization, Information Selection and Stimulus Uncertainty 

David J. Halpern (david.halpern@nyu.edu)<br>Todd M. Gureckis (todd.gureckis@nyu.edu)<br>Department of Psychology, NYU, 6 Washington Place<br>New York, NY 10003 USA


#### Abstract

Although a common assumption in models of perceptual discrimination, most models of categorization do not explicitly account for uncertainty in stimulus measurement. Such uncertainty may arise from inherent perceptual noise or external measurement noise (e.g., a medical test that gives variable results). In this paper we explore how people decide to gather information from various stimulus properties when each sample or measurement is noisy. The participant's goal is to correctly classify the given item. Across two experiments we find support for the idea that people take category structure into account when selecting information for a classification decision. In addition, we find some evidence that people are also sensitive to their own perceptual uncertainty when selecting information.


Keywords: attention, categorization, information sampling
Categorizing objects or situations into meaningful groups is a critical cognitive ability. Many existing theories of categorization share an unrealistic assumption that information about the features of a to-be-categorized object can be directly and precisely observed (Medin \& Schaffer, 1978; Nosofsky, 1986; Smith \& Minda, 1998; Love, Medin, \& Gureckis, 2004). But this is often not the case; for example, doctors do not always have full access to all the information about patient symptoms (i.e., stimulus features) but instead can only rely on a patient's self-report and medical tests which are subject to selective reporting and noise (perceptual or otherwise). If a doctor orders a cholesterol test, they must take into account that the patient's true levels are somewhat different than those reported by the test due to error or noise. There is often uncertainty not only about the category of an object but also about its specific dimension or feature values.

If noise and uncertainty are the affliction, then the antidote is the fact that categorizers can often select how feature or stimulus measurements are made. In many cases, a good strategy for selecting measurements (e.g., repeating the same measurement when it is known to be noisy) can significantly improve categorization performance. Here we ask what strategies people use to select stimulus measurements in the service of categorization. While previous approaches to this question tend to focus on how people select stimulus features to view without noise before making a categorization decision (e.g., Nelson, McKenzie, Cottrell, \& Sejnowski, 2010 or in eye and mouse tracking studies, Rehder \& Hoffman, 2005; Matsuka \& Corter, 2008; Blair, Watson, Walshe, \& Maj, 2009) here we explore the effect of measurement or perceptual noise on information sampling strategies.

## Categorizing in a Noisy World

We begin by presenting an Ideal Actor analysis of categorization under measurement noise which extends the General

Recognition Theory (Ashby \& Townsend, 1986). We then present a series of experiments exploring how people sample information about category features under conditions of stimulus noise.

Categorizing without perceptual noise The standard model of a probabilistic binary categorization task (Medin \& Schaffer, 1978; Nosofsky, 1986; Smith \& Minda, 1998) assumes that on a given trial $t$, a category $C_{t} \in\{A, B\}$ is drawn randomly and a stimulus $\mathbf{s}_{t}$ is generated from the distribution associated with $C_{t}$ (see Figure 1a). Subjects are assumed to either learn the parameters of these distributions through experience or description. Based on the information in $\mathbf{s}_{t}$, subjects guess which category $C_{t} \in\{A, B\}$ it was generated from. For example, in our first experiment, the stimuli have two dimensions, color and orientation, i.e. $\mathbf{s}_{t}=\left[s_{t_{\text {orientation }}}, s_{t_{\text {color }}}\right]$ and the category distributions are bivariate Gaussian distributions $\mathcal{N}\left(\boldsymbol{\mu}_{C_{t}}, \boldsymbol{\Sigma}\right)$ where the mean is dependent on the category and we can decompose the covariance matrix as $\boldsymbol{\Sigma}=$ $\left[\begin{array}{cc}\sigma_{\text {orientation }}^{2} & \rho \sigma_{\text {orientation }} \sigma_{\text {color }} \\ \rho \sigma_{\text {orientation }} \sigma_{\text {color }} & \sigma_{\text {color }}^{2}\end{array}\right]$. The Ideal Observer decision rule, assuming knowledge of the category distributions, is to use the log-likelihood ratio

$$
\begin{equation*}
l\left(\mathbf{s}_{\mathbf{t}}\right)=\frac{\log P\left(\mathbf{s}=\mathbf{s}_{t} \mid C_{t}=A\right)}{\log P\left(\mathbf{s}=\mathbf{s}_{t} \mid C_{t}=B\right)} \tag{1}
\end{equation*}
$$

This rule responds $A$ when $l\left(\mathbf{s}_{\mathbf{t}}\right)>0$ and $B$ otherwise (Ashby \& Gott, 1988).

Categorizing with perceptual noise In the 1980s, Ashby and colleagues developed General Recognition Theory (GRT), a family of models of multidimensional classification that assumed perceptual noise, in contrast to the models above (Ashby \& Townsend, 1986). Here we discuss a special case of GRT that uses the Ideal Observer decision rule and makes two critical assumptions about perceptual noise: that perceptual noise has a normal distribution and that the noise in perceiving each stimulus dimension is independent. In this model, on trial $t$ the subject perceives a percept $\mathbf{p}_{t}=\left[p_{t_{\text {color }}}, p_{t_{\text {orientation }}}\right]$ with probability $\mathcal{N}\left(\mathbf{p}_{t} ; \mathbf{s}_{t}, \boldsymbol{\Sigma}_{p}\right)$ where $\boldsymbol{\Sigma}_{p}=\left[\begin{array}{cc}\sigma_{\text {Porientation }^{2}}^{2} & 0 \\ 0 & \sigma_{p_{\text {color }}}^{2}\end{array}\right] . \boldsymbol{\Sigma}_{p}$ is diagonal due to the independent noise assumption and $\sigma_{p_{\text {orientation }}}^{2}$ and $\sigma_{p_{\text {color }}}^{2}$ are a subject's perceptual noise for orientation and color respectively. When making their categorization decision, subjects only have access to $\mathbf{p}_{t}$ rather than $\mathbf{s}_{t}$ as before (see Figure 1b). Therefore, the Ideal Observer's decision rule is now based on the $\log$ likelihood ratio $l\left(\mathbf{p}_{\mathbf{t}}\right)$ where $\mathbf{p}_{\mathbf{t}}$ is distributed under each category as $\mathcal{N}\left(\mathbf{p}_{t} ; \boldsymbol{\mu}_{C_{t}}, \boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{p}\right)$

Reducing uncertainty by making measurements One strategy for dealing with uncertainty in the percept $\mathbf{p}_{t}$ is to exert control over the amount of information collected about each stimulus dimension. In our task, the categorizer is given a fixed budget of $\kappa$ noisy measurements $\mathbf{m}_{t}$ of a stimulus made up of two dimensions. These measurements can be allocated so that there are $\lambda_{\text {orientation }}$ measurements of the orientation dimension and $\kappa-\lambda_{\text {orientation }}$ measurements of color with $\lambda_{\text {orientation }} \in[0, \kappa]$. Conceptually, this is like a doctor ordering multiple runs of medical tests and collect more or fewer runs of some tests over others. As in our experiment, orientation measurements are each independently distributed with probability $\mathcal{N}\left(s_{t_{\text {orientation }}}, \sigma_{m_{\text {orientation }}}\right)$ and similarly for color. Because these measurements are all normally distributed, the mean measurement $\overline{\mathbf{m}}_{t}$ is distributed as $\mathcal{N}\left(\mathbf{s}_{t}, \boldsymbol{\Sigma}_{m}\right)$, where $\boldsymbol{\Sigma}_{m}=\left[\begin{array}{cc}\frac{\sigma_{m_{\text {orientation }}}^{2}}{\lambda_{\text {orientaion }}} & 0 \\ 0 & \frac{\sigma_{m_{\text {color }}}^{\mathrm{K}-\lambda_{\text {orientation }}}}{}\end{array}\right]$. We now assume that subjects estimate the mean of several colors or oriented lines with a constant variance, largely supported by the literature on ensemble perception (e.g., Maule \& Franklin, 2016; Dakin \& Watt, 1997). We assume the same perceptual distribution as above except $\overline{\mathbf{p}}_{t}$ represents a subjects estimate of the mean measurement and $\boldsymbol{\Sigma} p$ represents their noise in estimating that mean. Let $\boldsymbol{\Sigma}_{m p}=\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{m}+\boldsymbol{\Sigma}_{p}$. Therefore,

$$
\begin{equation*}
P\left(\overline{\mathbf{p}}=\overline{\mathbf{p}}_{t} \mid C_{t}\right)=\mathcal{N}\left(\overline{\mathbf{p}}_{t} ; \boldsymbol{\mu}_{C_{t}}, \boldsymbol{\Sigma}_{m p}\right) \tag{2}
\end{equation*}
$$

The optimal decision rule depends on the log likelihood ratio $l\left(\overline{\mathbf{p}}_{t}\right)$ as above.
Optimizing stimulus measurement In our most interesting setting (see Figure 1c), we allow subjects to choose $\lambda$ in order to maximize their own categorization performance. This is akin to the doctor trying to optimize the probability of a correct diagnosis while trying to keep the cost of running the medical tests below a budget. To make the dependence on $\boldsymbol{\lambda}$ explicit, we can rewrite $\boldsymbol{\Sigma}_{m p}$ as $\boldsymbol{\Sigma}_{m p}\left(\boldsymbol{\lambda}_{\text {orientation }}\right)=$ $\left[\begin{array}{c}\left(\sigma_{\text {orientation }}^{2}+\frac{\sigma_{\text {morientaion }}^{2}}{\lambda_{\text {Orientation }}^{2}}+\sigma_{P_{\text {orientation }}^{2}}^{2}\right) \\ \rho \sigma_{\text {orientation }} \sigma_{\text {color }}\end{array}\right.$
 expected percent correct of an Ideal Observer with a given $\lambda_{\text {orientation }}$ should be (Anderson, 1958):

$$
\begin{equation*}
\mathrm{EC}\left(\lambda_{\text {orientation }}\right)=\frac{1}{2} \operatorname{erfc}\left(-\frac{\sqrt{\left(\boldsymbol{\mu}_{A}-\boldsymbol{\mu}_{B}\right)^{\prime} \boldsymbol{\Sigma}_{m p}\left(\lambda_{\text {orientation }}\right)^{-1}\left(\boldsymbol{\mu}_{A}-\boldsymbol{\mu}_{B}\right)}}{2 \sqrt{2}}\right) \tag{3}
\end{equation*}
$$

The Ideal Actor should then set $\lambda_{\text {orientation }}$ to $\lambda_{\text {orientation }}^{*}=$ $\arg \max _{\lambda_{\text {orientation }}} \mathrm{EC}\left(\lambda_{\text {orientation }}\right)$.

## Theoretical Predictions

In the following experiment, subjects perform a categorization task with stimuli that vary on color and orientation dimensions. Subjects only receive $\kappa$ noisy signals of what the color and orientation of the stimulus are but choose a number $\lambda_{\text {orientation }}$ of signals to receive of the orientation dimension. By varying the type of category structure (the


Figure 1: (a) Graphical model of the standard categorization model where stimuli $s$ depend on the category $C$. (b) a model analogous to General Recognition Theory where stimuli $s$ depend on the category $C$ but are never directly observable. Instead, the observer has access to $p$ which itself depends on $s$ but is corrupted with noise. (c) An active learning model of categorization under uncertainty. Here $\lambda$ reflects the number of measurements $m$ made of the stimulus $s$. The categorizer has access to $p$ which depends on the category, stimulus, and the information sampling strategy ( $m$ and $\lambda$ ).
two $P(\mathbf{s} \mid C)$ for category A and B) experimentally and using natural variation in perceptual noise, we can test two predictions about how subjects should select signals based on the above Ideal Actor model. We first divide category structures into three groups: 1D-color structures where $P($ orientation $\mid A)=P($ orientation $\mid B), 1 \mathrm{D}-$ orientation structures where $P($ color $\mid A)=P($ color $\mid B)$ and 2D structures otherwise. We predict:

1. The rank ordering of subjects choice of $\lambda_{\text {orientation }}$ should be 1D-color $\leq 2 \mathrm{D} \leq 1$ D-orientation
2. Within 2D categories, subjects choice of $\lambda_{\text {orientation }}$ will be modulated by the subject's measured perceptual noise.

Intuition for the two hypotheses can be seen in the Ideal Actor predictions in Figure 2 which plots expected Ideal Observer percent correct as a function of strategy and category structure in (a) and perceptual noise in (b). In the 1D-color situation accuracy is optimized by collecting zero orientation samples. In the 1D-orientation, accuracy is highest with 10 orientation samples. In the 2-D case, accuracy is optimized with 5 samples (Figure 2a). Furthermore the peak of these functions changes with perceptual noise (Figure 2b).

## Experiment 1

In order to test these predictions, we used a task divided into six phases. In the first two phases, we estimated the free parameters of the Ideal Actor, $\sigma_{p_{\text {orientation }}}^{2}$ and $\sigma_{p_{\text {color }}}^{2}$, the subjects' noise in estimating the mean of a number of colored dots and oriented lines using a 2 AFC task. In the next three phases,


Figure 2: (a) Ideal Observer expected percent correct as a function of condition (described in Table 1) and strategy. The rank order of the max of these curves is 1 D -color $\leq 2 \mathrm{D} \leq 1 \mathrm{D}$-orientation. (b) Ideal Observer expected percent correct if subjects had no perceptual noise (in red). In our experiment, we estimated noise parameters ( $\sigma_{\text {orientation }}$ and $\sigma_{\text {color }}$ ) using a 2AFC task. Shown in blue and purple are Ideal Observers with estimated noise parameters from two different subjects. Perceptual noise has a strong influence on optimal strategy which is indicated by the arrow
subjects underwent extensive training to learn the category and measurement distribution parameters (assumed known by the Ideal Actor). In the final test phase, subjects performed a categorization task where they could decide how allocate a fixed budget of samples to the two stimulus dimensions. All phases were 100 trials except the first category learning phase which was 200.
Participants Sixty one participants were recruited through Amazon Mechanical Turk. Participants received $\$ 8$ for participating in the experiment with a performance based bonus of up to $\$ 10$. Ten trials were selected at random from the entire experiment and participants were awarded a bonus of $\$ 0.25$ for each trial correct. Participants were randomly assigned to the eight conditions described in Table 1.
Stimuli and Procedure All stimuli in the experiment were generated randomly by drawing samples from the generative model. To generate the stimuli, each sample corresponded to either the angle of an oriented line relative to the circle or the color of a dot where the number was the angle on a circle of radius 60 in CIE 1976 (L*, $\mathrm{a}^{*}$, $\mathrm{b}^{*}$ ) color space. The locations of the colored dots on the screen were determined by force layout, an algorithm within the d3 javascript visualization library (Bostock, Ogievetsky, \& Heer, 2011). Examples of these stimuli can be seen in (Figure 3). Throughout the experiment, the "measurement noise" $\sigma_{m}=\sigma_{m_{\text {orientation }}}=$ $\sigma_{m_{\text {color }}}=.6$.
Perceptual Noise Estimation Phases. We adapted a 2AFC task from Jogan and Stocker (2014) designed to estimate subjects' noise in estimating a property of a stimulus. We conducted this task in two phases, one for orientation and one for color (with order counterbalanced across subjects). On each trial of the task, three stimuli $s \in\left\{\right.$ test, reference ${ }_{1}$, reference $\left._{2}\right\}$ were presented. The subject was asked which of the two reference stimuli was closer in terms of the property of interest to the test stimulus and responded by pressing the appropriate computer key. The specific properties of interest here were the average orientation of a set of several lines or the aver-
age color of a set of several dots. The stimuli during these phases looked the same as the stimuli in the later categorization phase (Figure 3b) but with just one feature present. On a given trial $t$, stimuli were generated by drawing $n_{t}$ samples from $\mathcal{N}\left(\mu_{s t}, \sigma_{m}^{2}\right)$ with $n_{t} \in[1,10]$ to keep the range and set sizes of the stimuli the same as in the later categorization experiment. $\mu_{s t}$ was selected on each trial by a Bayesian adaptive procedure (Kontsevich \& Tyler, 1999). Using an Ideal Observer analysis detailed in Jogan and Stocker (2014), we can estimate the perceptual noise parameters for identifying the mean of the stimuli based on subjects' performance in this task.
Category structures/conditions. Throughout the four categorization phases, stimuli were generated from a single pair of categories, chosen from a set of eight categories described in Table 1. For each subject, each dimension was shifted by a random amount chosen from a uniform on $[0,2 \pi]$ to wash out any effect of a specific stimulus range.

| Condition | $\mu_{A}$ | $\mu_{B}$ | $\Sigma$ |
| :---: | :---: | :---: | :---: |
| 1D-orientation | [0, 0] | [1, 0] | $\left[\begin{array}{cc}.2 & 0 \\ 0 & .2\end{array}\right]$ |
| 1D-color | [0, 1] | [0, 0] | $\left[\begin{array}{cc}.2 & 0 \\ 0 & .2\end{array}\right]$ |
| $2 \mathrm{D}_{1}$ | [0, 0] | [.24, 62] | $\left[\begin{array}{cc}.22 & -.2 \\ .2 & .22\end{array}\right]$ |
| $2 \mathrm{D}_{2}$ | [.24, 62] | [0, 0] | $\left[\begin{array}{cc}.22 & -.2 \\ .2 & .22\end{array}\right]$ |

Table 1: Category parameters for Experiment 1. There were eight categories total, with the other 4 being the same but with the entries of $\boldsymbol{\mu}_{A}$ and $\boldsymbol{\mu}_{B}$ swapped. For the main analysis, we collapse the conditions that share the parameters.

Category Learning Phase. In the category learning phase, a category was drawn from a uniform distribution and a bivariate sample was drawn from that category's associated distribution. This sample was converted to the color and orientation stimulus space using the procedure described above (see Figure 3a for an example). Subjects responded by hitting the "Q" key if they thought the stimulus was in category A and the "P" key if the stimulus was in category B. Subjects then received feedback on whether they were correct or incorrect depending on the category structures defined above.
Measurement Noise Learning Phases The measurement noise learning phases were meant to acclimatize the subject to the effects of measurement noise on categorization. The stimuli were created by sampling from the full generative model for the task as described in the theory section and converting samples to the stimulus space as described above. During this phase, subjects did not get to choose the number of measurements of each dimension. Instead, stimuli in the first phase included ten measurements of each dimension ( $\kappa=20$ and $\lambda_{\text {orientation }}=10$ ) and stimuli in the second included a total of ten measurements with a random number of them allocated to orientation $\left(\kappa=10\right.$ and $\left.\lambda_{\text {orientation }} \in[0,10]\right)$ See Figure 3b for an example of the stimuli in this phase. Note that there


Figure 3: Example categorization trials
are multiple lines/dots in this stimulus reflecting the multiple noisy "measurements" made of each dimension. In order to gain an intuitive understanding of the measurement procedure, subjects were told "we showed the color of the stimulus to 10 people and the location of the stimulus to 10 different people. Later each of them had to re-create what they saw from memory. Your task will be to take their recreations and try to guess what category you think the original stimulus belonged to." After every trial, subjects would receive feedback on their categorization judgement as well as feedback about what the true stimulus $\mathbf{s}_{t}$ had been on that trial.
Test Phase During this phase, subjects chose on a slider how many measurements of each dimension they would see on each trial $\left(\lambda_{\text {orientation }}\right)$. Stimuli were then generated in the same manner as in the measurement noise learning phases. Subjects then performed the classification task as in the previous training sections.

## Results

For each subject, we computed the posterior over the $\sigma_{p_{\text {orientation }}}^{2}$ and $\sigma_{p_{\text {color }}}^{2}$ parameters using the analysis described in Jogan and Stocker (2014). In order to check that the Bayesian adaptive procedure converged towards the correct estimate, Jogan and Stocker use a diagnostic called the Boundary Index (BI) a measure of the number of trials that were chosen to be at the boundary of the space. All of our subjects were below the recommended threshold of 0.9.

Using the perceptual noise posterior, we compute a posterior over Ideal Actor strategies for every subject in our experiment. Since each subject only experienced a single set of categories in the test phase, the Ideal Actor only uses a single $\lambda_{\text {orientation }}$ parameter for the entire experiment. However, subjects were able to change their choice of $\lambda_{\text {orientation }}$ parameter on every trial and relatively few subjects used just a single value. In order to compare subjects to the model, we took an average each subjects' setting of $\lambda_{\text {orientation }}$. We chose a priori to average only the second half of test trials to ensure that subjects had stabilized after having experience with using the slider. The results turn out to be unchanged even if we use all of the data from the test phase.

In order to test our first hypothesis we used Kendall's $\tau$, a


Figure 4: (a)Scatter plot of subjects strategies vs. the Ideal Actor (IA) strategy posterior mean. X error bars are + - 1 SD of the IA posterior and Y-error bars are standard error of mean subject strategy. (b)Expected number additional incorrect Ideal Observer (IO) trials relative to the Ideal Actor (IA). We do not include error bars on the actual performance since the sampling distribution is over trial sequences while the sampling distributions on the Ideal Observers are over measurement selections averaged over trial sequences.
common non-parametric rank-correlation method. We found a significant monotonic relationship between category structure and subjects' $\lambda_{\text {orientation }}$ choice with 1D-color $<2 \mathrm{D}<$ 1D-orientation (Kendall $\tau(59)=0.52, \mathrm{p}<1 \mathrm{e}-8$ ).

We also found a significant linear relationship between subjects' $\lambda_{\text {orientation }}$ parameters and the posterior mean Ideal Actor (Pearson $r(59)=0.65, \mathrm{p}<1 \mathrm{e}-8$ ). This result was still significant using just the data from the 2D structure where differences in Ideal Actor strategies are only due to differences in estimated perceptual noise (Pearson $r(30)=0.38, \mathrm{p}=0.03$ ). This provides weak evidence for our second hypothesis.

Given that many subjects did not use just a single $\lambda_{\text {orientation }}$ throughout the whole experiment, what was the cost of their suboptimal choice? Did they know to avoid choices that would lead to significantly worse performance? To answer this, we compare the theoretical performance of the Ideal Actor to what we call the subject Ideal Observer, the theoretical performance of an Ideal Observer who chose $\lambda_{\text {orientation }}$ on every trial as the subject did. The subject Ideal Observer performance is of interest because it isolates the expected decrease in performance solely due to choice of $\lambda_{\text {orientation }}$. In contrast, differences between subjects' actual performance and the Ideal Actor may be for several reasons unrelated to the information selection strategy. In Figure 4b, we compare the Ideal Actor to subject Ideal Observer performance (in blue), subjects' actual performance (in black) and a baseline where the Ideal Observer who chose $\lambda_{\text {orientation }}$ randomly (in purple). Only seven out of sixty-one subjects had subject Ideal Observers that did not perform significantly better than the baseline suggesting that most subjects were sensitive to the costs of choosing $\lambda_{\text {orientation }}$ incorrectly.

## Discussion

We found some preliminary evidence suggesting that people take category structure and perceptual noise into account. While the correlation between the Ideal Actor and subjects strategies was significant, subjects deviated from the Ideal Actor in other significant ways. In particular, subjects of-
ten used multiple $\lambda_{\text {orientation's }}$ throughout the experiment and actual categorization performance did not match the subject Ideal Observer - two substantive suboptimalities. There may be several reasons for this including that subjects might not use the Ideal Observer rule for categorization or subjects did not learn the exact category parameters in the time allotted. It is difficult to assess subject knowledge in this task since the average subject only had $75 \%$ agreement with the Ideal Observer during the last 10 trials of the category learning phase. Also, a MANOVA found that subjects' disagreement with the Ideal Observer and suboptimality was significantly different across category types (Wilk's $\Lambda=.62, \mathrm{~F}(2,58)=7.6, \mathrm{p}=1 \mathrm{e}-4$ ). This suggests that subjects might have significantly different knowledge about the category across conditions. Finally our estimates of $\sigma_{p_{\text {orientation }}}^{2}$ and $\sigma_{p_{\text {color }}}^{2}$ may not be perfect which might bias our Ideal Actor model.

Not learning the category parameters is probably the most serious issue since the Ideal Actor strategy depends heavily on these parameters. In order to address this, we conducted a second study involving only binary-valued stimulus features. With only a finite number of possible stimuli, we can easily check whether subjects have "learned" the category in the sense of having a high agreement with the ideal observer when selecting the category for a given stimulus.

## Experiment 2

Participants Thirty three participants who did not participate in the previous experiment were recruited through Amazon Mechanical Turk. Payment was the same as in Experiment 1.

Categorization Task Subjects in this task were instructed that they needed to help a doctor discover how to categorize patients presenting certain symptoms. Subjects would see the outcome of two medical tests represented as the color of horizontal and vertical lines (blue if positive and red if negative). All four of the possible outcomes can be seen in Figure 5 (a). Based on the stimulus, subjects would have to determine which of two diseases (A or B) the patient had. These diseases (or categories) were defined as bivariate Bernoulli distributions over possible test outcomes. Let 1 denote a positive test and 0 denote a negative test. Then let $P(\mid D)$ be a matrix where each entry with index $[v, h]$ indicates the probability of the vertical test taking on value $v \in(0,1)$ and the horizontal test taking on value $h \in(0,1)$ given that the patient has disease $D \in A, B$. We used category conditions described in Table 2.

In the first phase of the experiment, subjects simply saw the stimuli in Figure 5 with the above probabilities and subjects were told that the tests were performed with no measurement noise. In the later phases, subjects were told that they now would see $\kappa$ tests on every trial with $\lambda_{\text {horizontal }}$ measurements of the horizontal tests. These tests had Bernoulli measurement noise, i.e. the probability of the horizontal test outputting $k$ tests with the true value on each trial was $\left(\begin{array}{c}\lambda_{\text {horizontal }}\end{array}\right) p^{k}(1-p)^{\left(\lambda_{\text {horizontal }}-k\right)}$. In this experiment, we used
$\left.\begin{array}{|c|c|c|}\hline \text { Condition } & P(\mid A) & P(\mid B) \\ \hline \text { 1D-horizontal } & {\left[\begin{array}{cc}.5 & 0 \\ .5 & 0\end{array}\right]} & {\left[\begin{array}{cc}0 & .5 \\ 0 & .5\end{array}\right]} \\ \text { 1D-vertical } & {\left[\begin{array}{cc}.5 & .5 \\ 0 & 0 \\ \hline\end{array}\right]} & {\left[\begin{array}{cc}0 & 0 \\ .5 & .5 \\ 0 & .5 \\ .5 & 0 \\ 0 & .5\end{array}\right]}\end{array}\right]$

Table 2: Category parameters for Experiment 2. There were six conditions total, with the other three being the same but with the category labels swapped. For the main analysis, we collapse the conditions that share the parameters.
a $p$ of .8. Figure 5 shows an example of a noisy stimulus (i.e., multiple measurements of the horizontal or vertical line segment) with it's true value below. The choice of $\kappa$ and $\lambda_{\text {horizontal }}$ in each phase was exactly the same as in Experiment 1 with subjects having a choice of $\lambda_{\text {horizontal }}$ in the last phase as before. The first three phases of this experiment each consisted of 200 trials and the last phase had 100 . While we could not derive a general analytic solution to the Ideal Actor in this case, we can easily compute the strategy by enumerating all of the potential observed measurements. We also assume that in this case the effects of perceptual noise on performance are minimal.


Figure 5: Example experiment 2 stimuli

## Results

According to hypothesis 1 above, the rank order of the subject $\lambda_{\text {horizontal }}$ in each category structure should be [1D-vertical $\leq$ 2D $\leq 1 \mathrm{D}$-horizontal]. We again found a significant monotonic relationship between category structure and subjects' $\lambda_{\text {horizontal }}$ choice in the direction we hypothesized (Kendall $\tau(31)=0.747, \mathrm{p}<1 \mathrm{e}-9)$. In addition, we found a significant linear relationship (Pearson $r(31)=0.855, \mathrm{p}<1 \mathrm{e}-9$ ) meaning that the exact number of samples subjects chose were proportionally similar to the Ideal Actor. One interesting feature of the data was that most of the errors in subjects responses were in the one-dimensional categories, which may be due to a general hesitancy to only sample information about one feature.

We can also perform the same cost analysis as for the previous experiment. Figure 6 b shows that only one subject did not select measurements significantly better than the random


Figure 6: (a)Violin plot of subjects strategies as a function of category structure. (b)Expected number of additional incorrect trials relative to the Ideal Actor
baseline. We can also check whether subjects truly learned the categories: on average subjects had a $96 \%$ agreement with the Ideal Observer in the second half of the category learning phase and only 1 subject was below $90 \%$. In addition, based on a MANOVA, there was no effect of category structure on suboptimal measurement selection or agreement with Ideal Observer (Wilk's $\Lambda=.88, \mathrm{~F}(2,30)=.96, \mathrm{p}=.43$ ) suggesting that none of the conditions were more difficult than the others.

## Conclusion

We developed a new categorization paradigm in order to study people's strategies for information selection. These tasks allowed us to study human information selection in categorization tasks with measurement and perceptual noise, which we argue is the typical situation in everyday categorization. We analytically derived an Ideal Actor model of this task and from that derived two qualitative predictions for human behavior: 1) that subjects would be sensitive to the category structure and 2) their own perceptual noise. In Experiment 1, the predictions for perceptual noise were not fit to the selection task but estimated in a separate psychophysics task. Across two experiments, we demonstrated that most subjects take into account the category structure. The first experiment provided some evidence that subjects take into account perceptual noise as well although the evidence is somewhat weaker.

In order to get a better understanding of people's strategies, future work could address several additional questions including whether people are sensitive to the costs of information collection (see Meder and Nelson (2012) for some evidence that they do not) of different costs for correct or incorrect answers. Another direction might be whether people may be more sensitive to certain features of the categories (such as differences in the mean) than others (like the feature covariance). Finally, information selection has been proposed to be important in several other domains. Feature-based perceptual attention (Scolari, Edward, \& Serences, 2014) can be thought of as a type of information selection and our model has parallels with some existing models in that literature (Palmer, 1990). A future experiment could use our model to investigate how people allocate perceptual resources during categorization. Many economists have recently investigated limited
information as an explanation for many economic phenomena (Caplin, 2015) but have often assumed that people collect information optimally. Using this model and measurement selection task could allow assessment of how people actually select information in choice situations.

Acknowledgments This research was supported by NSF grant BCS-1255538 and a John S. McDonnell Foundation Scholar Award to TMG.

## References

Anderson, T. W. (1958). An introduction to multivariate statistical analysis (1st ed.). John Wiley and Sons, Inc.
Ashby, F. G., \& Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. Journal of experimental psychology. Learning, memory, and cognition, 14(1), 33-53.
Ashby, F. G., \& Townsend, J. T. (1986). Varieties of perceptual independence. Psychological review, 93(2), 154-179.
Blair, M. R., Watson, M. R., Walshe, R. C., \& Maj, F. (2009). Extremely selective attention: eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. Journal of experimental psychology. Learning, memory, and cognition, 35(5), 1196-206.
Bostock, M., Ogievetsky, V., \& Heer, J. (2011). D3 data-driven documents. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2301-2309.
Caplin, A. (2015). Measuring and Modeling Attention. Annual Review of Economics, 8(1), 1-43.
Dakin, S. C., \& Watt, R. J. (1997). The computation of orientation statistics from visual texture. Vision Research, 37(22), 31813192.

Jogan, M., \& Stocker, A. (2014). A new two-alternative forced choice method for the unbiased characterization of perceptual bias and discriminability. Journal of Vision, 14(3:20), 1-18.
Kontsevich, L. L., \& Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. Vision Research, 39(16), 2729-2737.
Love, B. C., Medin, D. L., \& Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. Psychological review, 111(2), 309-332.
Matsuka, T., \& Corter, J. E. (2008). Observed attention allocation processes in category learning. Quarterly Journal of Experimental Psychology, 61(7), 1067-1097.
Maule, J., \& Franklin, A. (2016, Mar). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. J. Opt. Soc. Am. A, 33(3), A22-A29.
Meder, B., \& Nelson, J. D. (2012). Information search with situation-specific reward functions. Judgment and Decision Making, 7(2), 119-148.
Medin, D. L., \& Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85(3), 207-238.
Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., \& Sejnowski, T. J. (2010). Experience matters: information acquisition optimizes probability gain. Psychological science, 21(7), 960-9.
Nosofsky, R. M. (1986). Attention, similarity, and the identificationcategorization relationship. Journal of experimental psychology. General, 115(1), 39-61.
Palmer, J. (1990). Attentional limits on the perception and memory of visual information. Journal of experimental psychology. Human perception and performance, 16(2), 332-350.
Rehder, B., \& Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51(1), 141.

Scolari, M., Edward, F., \& Serences, J. T. (2014). Feature- and Object-Based Attentional Modulation in the Human Visual System. In Oxford handbook of attention (pp. 573-600).
Smith, J. D., \& Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(6), 1411-1436.

# Perceptual contrast and response assimilation in sequential categorization without feedback 

James A. Hampton (hampton@city.ac.uk)<br>Department of Psychology, City, University of London<br>Northampton Square, London EC1V OHB


#### Abstract

Sequential categorization of perceptual stimuli typically shows contrast from one trial to the next. Using familiar categories of animals and faces, contrast effects were dissociated from assimilation effects. Two independent main effects were observed: contrast to the preceding stimulus, and assimilation to the previous response. It is argued that contrast and assimilation may reflect different processes in categorization.


Keywords: categorization, contrast, assimilation

## Introduction

Learning to categorize stimuli along a perceptual dimension (e.g. the pitch of a tone) would appear to require the establishment of a criterion value against which each stimulus can be judged. Above the criterion stimuli fall in one category, and below in the other (Ashby \& Gott, 1988). However Stewart, Brown, and Chater (2002; 2005) argued that people may be unable to keep a consistent criterion in mind through the course of a sequence of such decisions. It is well known from the literature on absolute identification (where each value on the dimension must be identified separately) that learning to make such judgments is very difficult and prone to bias (Garner, 1953; Baird, Green, \& Luce, 1980). In the same way, Stewart et al. (2002) argued, learning and retaining a given boundary on a perceptual dimension such as the pitch of a tone, will also be hard. Instead of making a judgment relative to an absolute value held in memory, people may instead rely on the difference between the current stimulus and the previous one. Since in a categorization task, feedback is normally provided on every trial, people can use the feedback on the previous trial, together with the direction and degree of change between the previous and the current stimulus to arrive at a reasonably accurate judgment. Stewart et al. referred to this way of doing the task as a MAC, or Memory and Contrast, strategy. Such strategies can lead to reasonable accuracy in the task, without recourse to the representation of absolute values in memory.

In this and subsequent research (Stewart \& Brown, 2004, 2005; Stewart, Brown \& Chater, 2005; Stewart \& Morin, 2007), Stewart and colleagues have shown that the use of MAC strategies produces a category contrast effect between one trial and the next. If a sequence of ten tone stimuli is divided into two categories \#1 to \#5 as "low" and \#6 to \#10 as "high", then a given stimulus next to the borderline (e.g. \#5) will be more likely to be correctly called "low" if preceded by a clearly high case (e.g. \#10) than by a clearly low case (e.g. \#1).

Stewart and Brown (2005) proposed that the category contrast effect could be part of a pair of heuristics in which the similarity and dissimilarity of a stimulus to the preceding stimulus is used to inform its categorization. In their Similarity-Dissimilarity (SD-GCM) adaptation of Nosofsky's Generalized Context Model (GCM) (Nosofsky, 1986), they proposed that categorization takes account not only of similarity to the exemplars of a class, as in the GCM, but also dissimilarity. Adopting Nosofsky and Palmeri's (1997) proposal that recently classified exemplars will have greater influence on the decision, the SD-GCM proposes that when a stimulus is very similar to one on the preceding trial or trials, then it should tend to be put in the same category, while when it is very dissimilar, it will tend to be put in the opposite category. The model thus predicts a contrast effect that increases with the dissimilarity between neighbouring stimuli in a sequence, and turns into an assimilation effect as the two become increasingly similar. Jones, Love and Maddox (2006) describe this prediction in terms of a generalization curve from one trial to the next that is positive for similar items (assimilation), but negative for highly dissimilar items (suppressing categorization, and thus generating contrast).

Evidence for the SD-GCM was also found in a series of studies by Hampton, Estes and Simmons (2005). In their paradigm, pairs of ambiguous stimuli were presented simultaneously and participants judged whether both, just one, just the other, or neither were in a category. A strong contrast effect was observed, seen as a bias to judge that just one stimulus was in the category, rather than both or neither.

The aim of the current study is to investigate contrast and assimilation in sequential judgments empirically. To provide a clear picture of the effect of the previous trial however, it is important to be able to separate out the effects of the previous stimulus from the previous feedback (category membership). Jones et al. (2006) pointed out that Stewart et al.'s (2002) procedure of providing correct feedback on every trial means that the two effects are confounded. To remove this confounding, Jones et al. introduced probabilistic category feedback, with the likelihood of a stimulus being categorized in a given category ranging from $90 \%$ to $10 \%$ across the range of stimuli. With this task, it was possible to separate out the effect of the previous stimulus - which they argued would be based on perceptual contrast - from the effect of the previous feedback, which would be a decisional effect. In their study they found that the previous stimulus produced a perceptual contrast effect, while the previous feedback produced assimilation when stimuli were similar, and contrast when they were very different. The heuristics
proposed by the SDGCM were therefore supported by the decisional contrast and assimilation observed, while in addition the effect of one stimulus was shown to lead to contrast with the next. Looking at the parametric properties of these two effects, they suggested that Stewart et al.'s original category contrast effect was probably perceptual in origin, since the decisional contrast only appeared at extreme distances.

The present study aimed to extend the method used by Jones et al. (2006) by differentiating the effects of the stimulus and the prior category in a simpler way. One problem with probabilistic category feedback is that participants may develop higher order models of the task. For example, if the most typical member of Category A is actually categorized by feedback as a B on $10 \%$ of trials, (non-modal feedback) the participant may develop a justified belief that the feedback is erratic and sometimes gives the wrong answer. They may even believe that the stimulus-response mapping has changed (Berg, 1948). Nonmodal feedback is thus likely to cause disturbance to the current response strategy. Indeed, the reported results (Jones et al., 2006, Figure 5) are consistent with this suggestion. When the previous trial gave modal ("correct") feedback, the response curve rose smoothly with position along the physical dimension, asymptoting near 0.95 for the last two or three stimuli. When the previous trial's feedback was non-modal however, the data showed a pattern of similarity based responding, with generalisation of the same category response to similar stimuli, but also a tendency to reverse categorization for distant stimuli, thus leading to the negative generalisation or contrast for large shifts in stimulus values.

Given the possibility of different interpretations of Jones et al.'s results, it is therefore important to find other means of separating the effects of the prior stimulus and feedback to test the generality and reliability of their result. To do this, a categorization task was used in which, because the categories are well known prior to the experiment, no feedback is needed. Hampton, Estes and Simmons (2005) used a categorization task in which 7 images were shown varying along a dimension in which an image of a cat was morphed into that of a dog (see Figure 1). The task was simply to categorize the images as cats or not cats. Since responses to stimuli in the middle of the range will be probabilistic it is possible to break the data down as a function not only of the previous stimulus, but also of the previous response. While this feedback-free procedure has been used in absolute identification (e.g. Mori \& Ward, 1995) it has not been attempted before for categorization.

Based on the results of Jones et al. (2006), assimilation to the previous response/category was predicted, that is, a bias to place a stimulus in the same category as the previous stimulus. Second, this bias should be stronger when the two stimuli are similar, and may even reverse to a contrast effect if they are very dissimilar. Third, it was predicted that there would be a perceptual contrast effect, such that, holding the previous response constant, then a current stimulus would
be more likely to be placed in the opposite category to the previous one.

## Experiment 1

## Method

Participants. Nineteen students from City University London (10 males) took part in the study.

Materials. Seven images (see Figure 1) were taken from Hampton et al. (2005, with thanks to Vladimir Sloutsky). They ranged from a clear picture of a kitten (\#1) through to a clear picture of a puppy (\#7), with equal morphed steps in between. Each image was $200 \times 200$ pixels, measuring 6.6 cm square on the display screen.


Figure 1: Stimuli used in Expt 1
Apparatus. A Dell PC with a 15-inch CRT was programmed in Microsoft Visual Basic.

Design. A repeated measures factor of Preceding Stimulus with 7 levels corresponded to the stimulus (\#1 to \#7) on the preceding trial. For analysis, this was converted to a measure of Relative Distance between the current and preceding stimulus. A post hoc factor of Response differentiated trials on the basis of the response given on the previous trial. The trial sequence provided a balanced pseudo-random sequence maximizing the number of useful trials. There were 6 blocks of 42 trials each. The 3 central images (\#s 3, 4, and 5) where responding would not be at floor or ceiling were taken as "target" stimuli, and evennumbered trials always featured one of these three. Within a block of 42 trials, each target was preceded on oddnumbered trials by each of the 7 images (including itself and the other target stimuli), and the transition from one trial to the next was balanced as in Table 1. Note that there were no transitions between non-target stimuli. Across blocks, the full transition matrix had 6 times as many trials in each cell as in Table 1. Self-terminated breaks occurred after the second and fourth blocks, and filler trials were introduced at the beginning of the experiment and after each break to provide the starting context for the next set of trials.

Procedure. Instructions were displayed below the 7 images ranked from \#1 to \#7. A trial began with the display of an image below which appeared two response boxes labeled CAT and DOG. The image remained on screen until a response was made with the mouse. A centrally located NEXT button then appeared and had to be clicked, to reduce
response perseveration due to mouse position. On each trial the mouse pointer was moved from the same central starting location. To reduce image persistence, and hence the possibility of detecting small changes from trial to trial, the images were displayed alternately to left and right of center.

Table 1: Transition frequency from one trial to the next within each of the 6 blocks of 42 trials. The shaded area shows the critical trials involving a decision about one of the target borderline images.

|  | Stimulus on Current Trial |  |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Trial <br> Before | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Sum |
| 1 |  |  | 1 | 1 | 1 |  |  | 3 |
| 2 |  |  | 1 | 1 | 1 |  |  | 3 |
| 3 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 10 |
| 4 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 10 |
| 5 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 10 |
| 6 |  |  | 1 | 1 | 1 |  |  | 3 |
| 7 |  |  | 1 | 1 | 1 |  |  | 3 |
| Sum | 3 | 3 | 10 | 10 | 10 | 3 | 3 | 42 |

## Results

Preliminary data processing was required to clarify the results. First, one participant chose "cat" on only 6 of the 180 target trials, and was excluded. Next, the rate of responding "cat" to each stimulus was calculated to check for range effects. The cat/dog boundary was not quite at the center of the scale, but lay between \#3 and \#4 on the CAT side of the center. As a result, almost all responses to \#5 were "dog". Since \#5 would therefore show a floor effect, this stimulus was not used as a target, leaving \#3 and \#4 as the target stimuli. The proportion of "cat" responses given to each of these was calculated as a function of the response given on the previous trial, and the relative distance of the previous stimulus from the current target. Two participants lacked data for just one cell each, and missing values were replaced with the relevant mean group.
Figure 2 shows the proportion of "cat" responses to the current target stimulus, averaged over the two target stimuli \#3 and \#4. The top line shows trials where the preceding response was "cat", and the lower shows trials where it was "dog". The horizontal axis shows the relative distance of the current target from the previous stimulus. Thus CAT+2 represents the case where the preceding trial was 2 steps more cat-like than the target and DOG +3 shows the case where the preceding stimulus was 3 steps more dog-like. TARGET represents the case where the preceding stimulus was the same as the current target stimulus. Note that no data are shown for CAT +3 because target \#3 could only have a previous stimulus (\#1) two steps more cat-like. In addition, two other points are missing from the graph because there were too few data points to estimate the mean.

These points corresponded to the prior stimulus three steps more dog-like (DOG +3 ) being called a "cat", and the prior stimulus two steps more cat-like (CAT+2) being called a "dog".
Figure 2 shows two main effects. First, when the previous response was "cat" (the top line), the current target stimulus was more likely to be called "cat" than when the previous response was "dog" (the lower line). Thus there was an assimilation to the previous stimulus showing up as a perseveration of the previous response. Regardless of the distance on the scale between the current target and the previous stimulus, there was a bias to repeat the same response.


Figure 2: Categorization probabilities for Experiment 1
Second, the lines rise from left to right, indicating a contrast from the previous stimulus. As the prior stimulus became more like a dog (moving from left to right along the axis) so the probability of the current target being called "cat" increased in a linear fashion. For example, consider the lower dotted line in the figure where the previous response was "dog"; the same target was called "cat" only $20 \%$ of the time if following the stimulus CAT+1-a stimulus one step more cat-like - but it was called "cat" over $50 \%$ of the time if following stimulus DOG+3 - where the previous stimulus was 3 steps more like a dog.
The data were submitted to ANOVA with factors of Previous Response (Cat, Dog), and Relative Distance of previous stimulus (four levels for which full data were available: CAT +1 , TARGET, DOG +1 , DOG +2 ). Both main effects were highly significant: Previous Response, $\mathrm{F}(1,17)$ $=76.3, \mathrm{p}<.001$, and Relative Distance, $\mathrm{F}(3,51)=5.36, \mathrm{p}<$ .005 , and there was no interaction $(\mathrm{F}(3,51)=1.4, \mathrm{p}>.2)$. Relative Distance had a significant linear trend $(\mathrm{F}(1,17)=$ 14.7, p < .001).

Before discussing the implications of the results, a replication will be described using the same procedure but a different set of stimuli.

## Experiment 2

## Method

Participants. Participants were 23 students (3 male) from City University, London.

Materials. The stimuli used were a set of morphed images created between two celebrity faces, one of Eva Longoria and the other of Victoria Beckham (see Figure 3). Pilot testing established the necessary gradations to achieve a range of ambiguous images.

Apparatus and Procedure. The same apparatus and procedure was employed as in Expt 1. In addition, in a first phase, participants were shown a sheet of images with 18 morphs changing from Eva to Victoria, and were asked to select the image that they felt was the closest to the 50:50 boundary between the two end images. This image was then used as the central target (\#4) in the sequence for that participant with stimuli \#1 to \#7 in equal steps either side. The same sequence of trials was presented as in Experiment 1, with the question "Is this picture more like Eva or Victoria?" and a mouse response. Data were scored as the probability of categorizing the image as Eva.


Figure 3. Morphed Images of Eva Longoria and Victoria Beckham From the Range of 18 Morphs in Experiment 2

## Results

The mean likelihood of responding "Eva" was calculated for the three borderline target faces, as a function of the relative scale position of the previous face, and the response made to the previous trial. As in Experiment 1, there were two clear effects on the probability of choosing the "Eva" response, over and above the effect of the target face itself. First, the previous response had an assimilation effect when the previous response was "Eva" there was an increase from 0.40 to 0.61 in the probability of the current response being "Eva" regardless of the target or the previous stimulus. Second, as in Experiment 1 the effect of the previous stimulus was to produce a contrasting shift in the response.

Because the range of borderline faces was narrower than in Experiment 1, there were insufficient trials in which an ambiguous target followed another ambiguous face. (These trials are critical to applying the method of analysis used in Study 1).

A different method of analysis was therefore used. The effect of the previous response was calculated for the three images in the middle of the scale $3-5$, where there were sufficient responses of each type. A 2 (Previous response) x 3 (Previous Image) x 3 (Target face) repeated measures ANOVA showed a highly significant effect of the Previous Response, $(\mathrm{F}(1,22)=32.3, \mathrm{p}<.001$.)

To assess the contrast effect, for each participant, target face, and previous response, the correlation was calculated between response probability of saying "Eva" and the scale value of the previous stimulus. For example, for face \#4 in Figure 3, the rate of responding "Eva" was calculated based on which of the 7 faces had been presented on the previous trial. A correlation was then calculated to determine whether the rate was increasing or decreasing as the previous stimulus changed. Six correlations were calculated for each participant, based on the 3 target faces $(3,4,5)$ and two previous response possibilities.

Positive correlations indicated a contrast effect. A higher scale value for the preceding stimulus means it is more like Victoria, so that a contrast effect would see a higher probability of saying "Eva" following face \#7, than following face \#1. The correlations were transformed to Fisher Z and t -tests were run across the participants for each of the 3 targets $(3,4,5)$ and 2 previous responses. Eva responses to Target 3 following an Eva response, and to Target 5 following a Victoria response both showed a range effect (the first being towards ceiling and the latter at floor). For the other four conditions, the average correlation ranged from .35 ( $\mathrm{p}<.05$ ) to .49 ( $\mathrm{p}<.002$ ) all showing a strong positive contrast effect of the previous stimulus.

## Discussion

This study set out to separate the effects of the preceding response and stimulus by using a set of images with a vague borderline region that was wide enough to provide trials where a given prior stimulus could be categorized either way.
First, the results clearly demonstrated strong assimilation to the previous response, in line with the findings of Jones et al. (2006). For the central stimuli where sufficient responses of each kind were available, there was a bias of around .20 to .30 in the probability of a given categorization in the direction of the previous response, regardless of the previous stimulus. Unlike Jones et al. however there was no reduction in the response assimilation effect as the stimuli became more different. There are two likely explanations for this effect. First, in paradigms where feedback is provided, people tend to repeat whatever response was reinforced on the previous trial (Jones \& Sieck, 2003), similar to the situation with probability learning (Edwards, 1961, Jarvik, 1951), and consistent with theories of conditioning (Rescorla \& Wagner, 1972). Even though no feedback was provided (the participants in the present study decided for themselves where the boundary lay) they could nonetheless have treated their previous response as an anchor on which to base the next. This is what Mori and

Ward (1995) reported for an absolute identification task when feedback was omitted. A second explanation would be in terms of instability in the classification criterion adopted by participants over the course of the experiment leading to autocorrelation of responses. If the criterion is sometimes set low, then most stimuli will be "cats", and if it then drifts high, they will mostly be "dogs", yielding a greater preponderance of repeated category responses than expected. (Note that this explanation does not work on the basis of participants having different criteria since response rates were calculated separately for each participant before averaging.)

The failure to show a reduction in response assimilation as the stimuli became less similar may be a function of the design. Trials focused on given borderline targets, preceded by different context stimuli. Thus there were no transitions from one extreme of the scale to the other, which is where Jones et al. observed the response-based contrast effect. There was little point in including such trials as almost all responses to the extreme images were at floor or ceiling. As described in the introduction it is uncertain whether the decisional contrast effect reported by Jones et al. (2006) is genuine evidence for negative generalization of the previously reinforced category (as they argue). When nonmodal feedback was provided (i.e. "false" feedback) performance on the next trial was disturbed, and this could have several alternative explanations. Since the argument for negative generalization depends on the use of invalid feedback, it must be treated with caution.

Once the data were analyzed separately depending on both the stimulus and the response on the previous trial, a contrast effect was also seen, confirming Jones et al.'s report of a perceptual contrast effect between neighboring stimuli, independent of the previous response (or in their case feedback). Unlike Stewart et al.'s (2002) contrast effect, the effect here is unlikely to reflect a MAC strategy.

In comparison with previous research using simple tones or rectangles, participants showed little evidence of noticing when a stimulus changed from the previous trial, and if so in what direction. Our procedure of shifting the image left and right between trials may have made it harder to notice when a stimulus had changed from trial to trial and in what direction. For example, if participants had recognized that a stimulus had been repeated, (the point labeled TARGET in Figure 2) then one would naturally expect the previous response to also have been repeated, leading to extreme response rates of 0 if the previous response was Dog and 1 if the previous response was Cat. The data in Figure 1 show no evidence for this strategy. Similarly, if people noticed the direction of change from trial to trial, extreme responding should have been found when the direction of change was further towards the category previously chosen (the Monotonicity Constraint identified by Hampton et al., 2005). If I have just called something a Cat, and this new image is even more cat-like, I would not now call it a dog. In that case, for example, in Figure 1, the rightmost two points on the top filled line should be at 1 , since here the
participant has said that $\mathrm{DOG}+1$ or $\mathrm{DOG}+2$ looks like a cat, and they are now faced with a more cat-like target. This type of behavior was not seen in the data. It is striking that even though participants appeared to make no use of how the stimuli changed from trial to trial, they were still more likely to judge the current borderline image as a cat, when preceded by a more dog-like image, and vice versa, regardless of how they classified that previous image. Perceptual contrast effects such as this can be explained as adaptation level effects (Hampton et al., 2005; Helson, 1964; Treisman \& Williams, 1984). Treisman and Williams proposed a tracking process, whereby category criteria may adjust themselves towards the average of the current stimulus environment, to maintain maximum sensitivity to change, and thus leading to contrast.
The contrast and assimilation effects built in to Stewart and Brown's (2005) SD-GCM may therefore be operating at different levels, something that previous research has not considered. Contrast - the tendency to see one stimulus as more likely to be in the category opposite to the preceding stimulus - comes out as primarily perceptual in the present experimental set up, showing up regardless of how the previous stimulus was categorized. The fact that pairs of trials where a stimulus was repeated showed no increased tendency to repeat the previous response is clear evidence that participants were not using a MAC strategy in this case. Recall that the MAC strategy explains contrast in terms of a participant judging the sign (and possibly magnitude) of the change in the stimulus presented from one trial to the next. A large shift away from the category of the previous stimulus will give rise to contrast.
Assimilation can also have two sources. According to the SD-GCM it is the similarity of two sequential stimuli that leads to perseverance of the response, and hence to assimilation. If you choose to call one stimulus a cat, and the next stimulus is hardly any different, then you call the next stimulus a cat as well. This type of assimilation should then be sensitive to the distance between the two stimuli. Only when a pair of stimuli are highly similar should you get assimilation.
The assimilation observed here was quite different in character. There was no good evidence in the lines in Figure 1 of the slope flattening out or becoming negative as the previous stimulus approached the target. The large assimilation effect was entirely associated with the effect of the previous response. Having called one stimulus a cat or dog, there was an inertia in the response, leading to a constant bias to place the next stimulus in the same category, regardless of the distance between them.

## Acknowledgments

Komal Nanda and Ibhana Islam contributed to Experiments 1 and 2 respectively. The author acknowledges helpful discussions with Paul Egré and Gregory Murphy, and express warm thanks to Zachary Estes for detailed comments on an earlier draft.

## References

Ashby, F. G., \& Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 33-53.
Baird, J. C., Green, D. M., \& Luce, R. D. (1980). Variability and sequential effects in cross-modality matching of area and loudness. Journal of Experimental Psychology: Human Perception and Performance, 6, 277-289.
Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. Journal of General Psychology, 39, 15-22.
Edwards, W. (1961). Probability learning in 1000 trials. Journal of Experimental Psychology, 62, 385-394.
Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. Journal of Experimental Psychology, 46, 373-380.
Hampton, J. A., Estes, Z., \& Simmons, C. L. (2005). Comparison and contrast in perceptual categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 1459-1476.
Helson, H. (1964). Adaptation-level theory. Oxford, UK: Harper \& Row.
Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. Journal of Experimental Psychology, 41, 291297.

Jones, M., Love, B. C., \& Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequence effects in category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 332.
Jones, M., \& Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 626-640.
Mori, S., \& Ward, L. M. (1995). Pure feedback effects in absolute identification. Attention Perception \& Psychophysics, 57, 1065-1079.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.
Nosofsky, R. M., \& Palmeri, T. J. (1997). An exemplarbased random walk model of speeded classification. Psychological Review, 104, 266-300.
Rescorla, R. A., \& Wagner, A. R. (1972). A theory of Pavolvian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H.Black \& W. F. Prokasy (Eds.), Classical conditioning II: Current research and theory (pp. 64-99). New York: Appleton-Century-Crofts.
Stewart, N., \& Brown, G. D. A. (2004). Sequence effects in the categorization of tones varying in frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 416-430.

Stewart, N., \& Brown, G. D. A. (2005). Similarity and dissimilarity as evidence in perceptual categorization. Journal of Mathematical Psychology, 49, 403-409.
Stewart, N., Brown, G. D. A., \& Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 3-11.
Stewart, N., Brown, G. D. A., \& Chater, N. (2005). Absolute identification by relative judgment. Psychological Review, 112, 881-911.
Stewart, N., \& Morin.C. (2007). Dissimilarity is used as evidence of category membership in multidimensional perceptual categorization: A test of the similaritydissimilarity generalized context model. Quarterly Journal of Experimental Psychology, 60, 1337-1346.
Treisman, M., \& Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. Psychological Review, 91, 68-111.

# Radical Embodied Cognition, Affordances, and the (Hard) Problem of Consciousness 

Valerie Gray Hardcastle (VALERIE.HARDCASTLE@uc.edu)<br>Departments of Philosophy, Psychology, and Psychiatry \& Behavioral Neuroscience<br>Cincinnati, OH 45221 USA


#### Abstract

Tony Chemero advances the radical thesis that cognition and consciousness is actually the same thing. He draws this conclusion from his understanding of cognition as an extended process. I question this conclusion because this view expands cognition beyond being the sort of natural kind to which one can tie phenomenal experience. Moreover, because cognition has been radically inflated, despite Chemero's claim to the contrary, embodied cognition does not solve any of the hard problems associated with consciousness.


Keywords: radical embodied cognition; consciousness; perception-in-action; the hard problem.

Novel stimuli capture our attention. This well-known fact forms the basis for several contemporary theories of theories of mind and brain, including Tony Chemero's notion of embodied cognition (Chemero 2009). Chemero holds that noticing unexpected events is key to understanding the larger cognitive system of which our brains are one part. He also believes that expecting some event to occur in the world is somehow tied to our conscious experiences.

In his book Radical Embodied Cognition (2009; see also Silberstein and Chemero 2012, 2015) Chemero (along with Michael Silberstein) advance the radical thesis that cognition and consciousness is actually the same thing. He draws this conclusion from his understanding of cognition as a dynamical, non-linear, relational, and extended process. I question this conclusion. Even if, we are the brain-bodyenvironment synergies that Chemero and others claim we are (e.g., Anderson et al. 2012, Silberstein and Chemero 2012, Kello and van Orden 2009, Kelso 2009), we will not be able to conclude that consciousness and cognition are two sides of the same coin because this view expands cognition beyond being the sort of natural kind to which one can tie phenomenal experience. Moreover, contra Chemero's claim that "the problem of qualia does not arise in radical embodied cognitive science" (2009, Loc 2530/3178), embodied cognition does not solve any of the hard problems associated with consciousness. Nonetheless, some of Chemero's views do help us understand some aspects of conscious experience.

## Radical Embodied Cognitive Science

Most date the recent embodied movement in cognitive science to the work of Rodney Brooks (1991a, 1991b) and Francisco Valera, Evan Thompson, and Eleanor Rosch's book The Embodied Mind (1991), though of course J.J. Gibson's ecological theory is its earliest contemporary incarnation (Gibson 1962,1966,1979). This work was intended to be an antidote for computational views of mind, in which perception, memory, and thought all become manipulations of brain-based mental representations.

Proponents of embodied cognition hold agents' bodies, and often their local environments, are not only physically relevant to cognition, but are also causally constitutive. Moreover, cognition is not the rational and abstract process that computationalists assume, but instead is dedicated to helping our bodies move and act upon our environment.
We evolved as a species to take advantage of our environment, and we do so as we solve so-called cognitive problems. In many cases, what at first appears to be a difficult problem to solve using abstract representations and computations divorced from the physical world turns out to be much easier to resolve if we are allowed to consider our bodies and our environment as cognitive resources. It is easy for a human to learn to move about on land because Mother Nature designed our legs for walking on Earthly terrain (Thelen and Smith 1994). In other words, instead of using our brain to solve problems, we manipulate our bodies and our environments to dissolve them. Moreover, in doing this, we can also alter our bodies and our environments such that the problems we need to solve also change. Gathering food presents a different sort of challenge in a cultivated field, compared to an unspoiled savannah rich with bison.

Hence, instead of thinking of the agent and its environment as two separate entities that occasionally touch each other, it is better to see them a single, interacting, system. It follows from this perspective that cognition is extended into the environment. It also follows that any mathematical model of this larger system should describe how it unfolds dynamically over time, which would require non-linear differential equations.

What makes Chemero's views of embodied cognition radical is that he claims that dynamical systems theory, which we need to be able to model cognition, does not presuppose mental representations, or indeed any representations at all. If brains, bodies, and environment form one unified system, then there is no need for one part of the system to represent another part of the system for everything is always and already connected. Perception, action, and thought itself are all non-representational and non-computational. Let us accept this view as true and see how he might explain consciousness using this understanding of human cognition and action.

## Conscious Cognitive Systems

Silberstein and Chemero (2012) hold that conscious experience is "an essential feature of extended brain-body- environment systems" (p. 36) and that phenomenology and cognition are "inseparable and complementary aspects of coupled brain-bodyenvironment systems....Experience is cognition and cognition is experiential" (pp.40-41). They each "codetermine" the other (p. 41). They get there by positing that some version of neutral monism must be true, if the thesis of radical embodied cognition is true (Silberstein and Chemero 2015). They trace this metaphysics back to William James, who held that there was no actual difference between the so-called objective world out there and the subjective experience of the world in each of us. "What represents and what is represented is here numerically the same" (James 1904, p. 484). Both the subjective and the objective are defined in opposition to one another, and both are ways of understanding the world, which is the "more basic neutral 'stuff' of experience" (Silberstein and Chemero, 2015, p. 186). Hence, one cannot have cognition, an objective process, without also having at the same time, consciousness, the subjective experience.

Computational theories of mind artificially create a problem for consciousness, for they give the impression that one can have mental computation without concomitant conscious thought. We get the problem of consciousness because we can imagine nonhuman, apparently unconscious, machines instantiating a computational theory of mind. That is, we can imagine cognition without consciousness, or so we think.

This is the hard problem of consciousness, a challenge that dates back to at least Gottfried Leibniz in $1714:^{1}$ because we can imagine a something that is

[^85]identical to a human's (or a brain's) physical interactions without also imagining that thing's consciousness, it appears that nothing about any physical interaction should give rise to phenomenal experience. And yet, we are conscious, nonetheless. How can this be? (Leiniz's answer, like David Chalmer's $(1995,1996)$ more contemporary one, is to posit that consciousness is a fundamental part of the ontology of the universe.)
But if computational theories of mind are false, then this particular problem of consciousness does not arise. Cognition is a feature of our extended and embodied system, as is consciousness. By "[refusing] to separate meaningful cognition and phenomenology" (2012, p. 41), Silberstein and Chemero believe that they have eliminated the so-called hard problem of consciousness essentially by definitional fiat. They assert that, "neutral monism properly conceived really does deflate the hard problem once and for all" (2015, p. 182).

Can they do this? To answer this question, we need a more complete picture of what they are envisioning an extended phenomenological-cognitive system to be.

Our nervous system has its own spontaneous and internally generated dynamics, which in turn create transient neural assemblies comprising our sensorimotor capabilities. We are coupled to our environment via these sensorimotor structures, which result in changes to both our internal transient neural assemblies via sensory feedback and in the external environment via behavioral responses. Over time, we become attuned to nuances in the extended brain-bodyenvironment system that complements our sensorimotor sensitivities and external features; this is our niche.
What we perceive in our environment via our sensorimotor structures (and probably other related neural assemblies) are nothing less than Gibsonian affordances, relational features of the brain-bodyenvironment system used to guide our actions and behavior. They are what the environment contains and what we can do. Silberstein and Chemero claim that the "set of affordances" we perceive in our world, "just is the environment as [we] experience it" (2012, p. 43, italics theirs). Hence, "cognition and conscious experience can be understood as a single phenomenon"

[^86](p. 35). Consciousness "is inseparable from cognition, which is the ongoing activity of a nervous system, body, and niche non-linearly coupled to one another" (p. 43).

Just as our neural assemblies are not anatomically hardwired - they are "softly" assembled - so too are the borders of the brain-body-environment system. How far and how much we extend into the environment depend a great deal on what we are trying to do and what barriers or assistance the environment provides to us. The entire system itself then is also a soft assembly whose interaction dynamics determine its structure.

Some call such interaction dominant, softly assembled, systems "synergies." A synergy is a set of structural units that temporarily link together to form a single cohesive functional system. It is maintained or changed on the fly as its dynamics and processes ebb and flow over time (Anderson et al. 2012, Kelso 2009).

We know that we have a synergy when we can measure pink noise associated with it. Pink noise, or $1 / f$ noise, refers to a signal in which the power spectrum density (energy per Hz ) is inversely proportional to the frequency. (It is called pink noise because visible light with this power spectrum looks pink.) We can contrast pink noise with white noise, which has equal energy on every frequency. Pink noise is also a hallmark of fractal timing and appears to be ubiquitous in nature, occurring in everything from cosmic background radiation to flooding patterns of the Nile River (see also Strogatz 2004 for a popular way into these phenomena). It is an indication of nested, selfsimilar structures that occur over time. Guy van Orden and his colleagues (2003) argue that pink noise signifies just the sort of interaction-dominant, softly assembled, system we have been discussing as a model for cognition. (See also Miyazaki et al. 2004.) And with this technical idea in hand, scientists are now able to manipulate and measure our dynamic embodiment experimentally.

For example, Chemero and his colleagues devised an experiment that forces change in our extended cognitive synergy (Dotov et al., 2010, 2017). Undergraduates engaged in a simple video game, using a computer monitor and a mouse. At irregular intervals during each trial, the connection between the mouse and the monitor was disrupted. Interestingly, pink noise is present at the hand- mouse interface until the disruption. Once the disruption is over and the connection returns to normal, the pink noise returns. These measures index changes in the boundary of the extended cognitive synergy. During the normal phase of the task, the mouse is part of the system. During disruption, it is not.

Most important for our purposes, the measures of pink noise correlate with our conscious experiences. When we are engaged in the video game, we are
not aware of the hand-mouse interface per se, but once the connection between mouse and monitor is altered, then the mouse grabs our attention and we become aware of it. The point is, we project ourselves into our environment, and in so doing, we experience the edges of our extended system. We have long known that this is the case, but it is only now that psychologists have been able to develop metrics for measuring changes in our projections.

The pink noise of our cognitive synergy indexes our phenomenal experience. I think that this is the best argument for why the picture Chemero paints might in fact be true. We notice and pay attention to, experience, things that do not match our predictions or expectations. And we experience these things in terms of what we could do, or how we could act. Our experiences are about or of the relationship we have with the world, which is continuously changing and evolving.

These mixes of brain, body, and environment are what Gibson called objective-subjective hybrids. And that fact, Chemero believes, solves the problem of consciousness. For this is how subjectivity exists in an objective world-it exists in the on-going relationship between an agent and the context of its actions. (The relationship itself is neither subjective nor objective. This is the neutral substrate that allows us to define subjectivity and objectivity as two different aspects of the same "monism.")

## The Hard Problem is Hard

What is wrong with this story? Essentially, it is that the sort of non-linear coupling that links us with our environment is found at all sorts of levels of organization. Without further analysis, we cannot identify which level corresponds to cognition; hence, tying consciousness to cognition either means consciousness exists at multiple levels of organization, which strikes most people as improbable, or more work needs to be done to delineate consciousness cognition from other synergies. If we need to do more work, then the problem of consciousness remains to be solved.

To take one example, we find synergies within the brain itself. These are softly assembled, interactiondominant, nonlinear dynamical systems whose behavior strongly resembles prototypical cognition, perception, and action. It is questionable whether these systems are also conscious.
Neuroscientists are now capable of recording responses from thousands of neurons simultaneously. It is becoming clear that neural correlates for things like memory and decision-making are at the population level (Abbott 2012). Modeling studies show that while the action potentials of individual neurons might appear disordered and uncoordinated, population level activity is not.

We know that neurons and neural circuits have to respond quickly and flexibly as contexts change. This means that they need to be able to ignore irrelevant information while reacting to whatever is important to the larger task at hand. Neural responses in the frontal eye fields in monkeys were recorded as they performed a visual discrimination task using noisy stimuli. It turns out that individual neurons simultaneously respond to the motion and color of stimuli, the context, as well as the target itself. However, these signals are separable at the population level through linear regression (Mante et al. 2012). Importantly, stimuli analysis at the population level is integrated with motor choices, just as proponents of embodied cognition would have predicted. We see similar dynamics in the olfactory system of the fly (Luo et al. 2010).

One important facet of nonlinear dynamical systems is that they are nested systems, and their components exhibit the same sort of dynamics as the system as a whole. Individual neurons and the ion channels in neurons also appear to have the same dynamical pink noise properties as the activity of populations of neurons (White et al. 1998, Yu et al. 2005). While most are comfortable believing that monkeys are conscious, it is less clear that we want to assert that flies are, and it is even more problematic to assert that parts of monkey brains or fly brains are conscious.

Chemero perhaps could wriggle out of this problem with his definition of cognition. That is, these might be synergistic systems, but they are not cognitive synergies. He defines cognition as "the ongoing, active maintenance of a robust animal-environment system, achieved by closely coordinated perception and action" (2009, loc. 2696/3178). In other words, the sorts of dynamically coupled systems I have been discussing are necessary but not sufficient for cognition. He is restricting consciousness to the brain-bodyenvironment system's level, or perhaps even to the animal brain-body-environment system. If you move inside the head, while there might exist dynamical systems modeled in identical ways to an animal brain-body-world synergy, there would be no actual cognition. And no cognition would mean no consciousness either, according to Chemero's view.

Of course, this move does not solve the hard problem of consciousness, since it does not explain what might be special about the animal brain-bodyenvironment synergy such that it has consciousness but the olfactory system of the fly brain does not. Indeed, this move echoes the challenge before the computationalist: what is it that is special about human (or primate or animal or whatever is conscious) computations that make them conscious? Prima facie, there is nothing about the computations themselves that should give rise to conscious experience, and there
are certainly many computational systems that we believe are not conscious. Similarly, we can ask: what is it that is special about an animal brain-bodyenvironment interaction that is cognitive and therefore consciousness? Prima facie, there is nothing about being an animal synergy that should give rise to conscious experience. In particular, there does not seem to anything special about an animal brain-body-environment interaction that an animal brain piece-body piece-environment interaction would not also share. Put another way: there does not seem any reason to believe that the neutral monads that comprise our world exist at the level of animalenvironment relations as opposed to animal-partenvironment relations.

Of course, another alternative is that Chemero could bite the bullet here and conclude that fly pieces are indeed conscious, in their own sort of fly-ish way. Being an interaction-dominant, softly assembled, pink noise sort of synergy is both necessary and sufficient for cognition and therefore for consciousness as well. The right sort of dynamics is all you need for cognition; the nested components of the nonlinear systems have all the same properties as the mother system, and this would include cogitating and consciousness.
Perhaps, though, he would not want to do this, since, as Silberstein and he point out, one advantage of tying consciousness to cognition is to "[eliminate] fruitless philosophical discussion of qualia and the so-called hard problem of consciousness" (2012, p. 35). They want to get rid of the challenge of envisioning odd machines as being conscious because they are computing over representations by denying that conscious systems compute at all. But here we are, back discussing an odd machine and whether it has experiences. This time, however, we are wondering whether the system is cognitive after all.

The complaint similar to the one Chemero and Silberstein lodge against traditional consciousness studies can also be lodged against them. While we can define cognition as a particular type of synergy, and we can believe that cognition just is an extended, softly assembled animal brain-body-environment synergy, and we can deny that manipulating representations has anything to do with humans thinking, we still do not escape the fundamental problem with explaining consciousness, that is, explaining why anything at all is conscious.

In other words, there is no reason why a neutral substrate should be conscious. It is, according to their view. But just as we can imagine (or so the story goes) things that are functionally identical to humans but are not conscious, so too can we envision complex brain-body-environment synergies comprised of a neutral substrate that is not conscious. The hard problem remains.

Or, perhaps a more accurate way to describe Silberstein and Chemero's move: consciousness just is an inherent aspect of affordances. Like Chalmers, they try to eliminate the hard problem by making consciousness part of the fundamental structure of the world. But just as with dualism, one needs an argument or evidence for why subjectivity appears where it allegedly does.

## Coda: Consciousness, Projection, and Action

However, there is at least one important difference between the brain-body-environment synergies and the human coupling described above that might give us some insight into conditions for conscious awareness: we do not or cannot project ourselves psychologically onto or through the other individuals. Unlike driving a car, in which we can "feel" the tires on the road, when coupled with another person, we do not "feel" the other person's feet hitting the ground. Whatever sort of system or synergy coupled humans are, the psychological reality of having an animate object in the environment is quite different from having an inanimate object. Inanimate object-environment synergies are transparent to us; human couplings are not.

There has to be something fundamentally different between the two. What is it? I argue that the difference lies in how we perceive the respective affordances. And insofar as how we perceive affordances is tied to how we consciously experience the world, then it could be that Chemero is onto something after all.
For example, there are some odd cases of associative agnosia in which patients are unable to recognize or name living things (like lions or opossums), but they can recognize and name inanimate objects (like forks and radios) without a problem (Satori and Job 1988). If we take a traditional neuro-reductionist point of view, then we should conclude that information about living things is stored in a different place in the brain than information about inanimate objects. Damage to the "living thing" place in the brain results in patients with deficits in recognizing living things and damage to the "inanimate object" place results in patients with deficits in recognizing inanimate objects. But interestingly, and perhaps counter-intuitively, there are very few cases in which a patient cannot recognize inanimate objects, but can recognize living things. We don't get the neat double dissociation that neuropsychologists love.
If there are two separate areas for living and inanimate objects, then why would we find brain damage possible in one area but not the other? Antonio Damasio (1990) suggests that this pattern could be due to a difference in how we perceive living and inanimate objects. In particular, we manipulate inanimate objects, but, for the most part, we do not living ones. As a result, we would be activating a
greater number of brain regions when we perceive inanimate objects than when we do living things: in both cases, we activate the areas associated with visual perception, but in the case of seeing a tool in our environment, we also activate kinesthetic and motor movement brain regions. (Imaging studies bear out Damasio's conjecture, cf., Gerlach et al. 2002, Kellenback et al. 2003.) Hence, it might be the case that we do not have different brain areas that respond differentially to living versus non-living objects, but rather that we just have more regions involved with one type of perception over another. With more regions activated in response to inanimate objects, and therefore more regions that would have to be damaged in order to see the related agnosia, it would not be surprising that we have a hard time finding patients with deficits naming inanimate objects but not living ones.
Here then is the hypothesis: we distinguish objects in our environment based on how we (potentially) interact with them. We perceive living things by their visual features and concomitant affective responses, but inanimate objects based on functional properties. This, of course, is just another way to look at affordances: we see and understand the objects around us in terms how we relate to them, and they to us. But now we can go further: not only are differences in psychological projection between animate and inanimate objects tied to perceptual differences, but also to action-decisions and consciousness itself.
I am claiming that we do not project ourselves through what we perceive to be other living things in our environment. We can only project ourselves through objects that we manipulate functionally (I note that there might be good exceptions to this general rule; for example, when we use a living thing as a tool in our environment. A practiced person might project through a trained seeing-eye dog to the environment beyond.) And, we can only extend our conscious experience into whatever psychological space of projection we have available to us. One limit on our conscious experience is not just the edge of the affordance, so to speak, but it is also the type of affordance we perceive. Functional objects become psychologically transparent to us, such that we project our conscious experiences through them. Animate objects do not.
I conclude: Chemero (and Silberstein) cannot escape the hard problem of consciousness by positing neutral monism. Nevertheless, there is something right about his position. Consciousness is connected to or indexed by or co-occurs with or identical to our perception of affordances, which is intimately tied to how we interact with the objects in our environment. We are aware of what we intend to manipulate in our environment in order to achieve our behavioral goals. Hence, consciousness is not identical to all cognition; it is not even identical to all brain-based cognition.

Instead, it is deeply linked to one very important part of our cognitive processes: perceiving affordances just prior to action.

## Acknowledgments

Early versions of the paper were presented to the Society for Philosophy and Psychology, the University of Waterloo, and at the Conscious Persons workshop at Calvin College. Many thanks to all the participants for their helpful and thoughtful comments. Special thanks are also due to Tony Chemero and Mike Silberstein for engaging with me on this topic, and to the referees for the Cognitive Science Society, whose thoughtful comments made this essay better. This research was partially supported by the Templeton Foundation and the University of Cambridge.

## References

Abbott, L. (2012). The collective wisdom of neurons. Albert and Ellen Grass Lecture. Society for Neuroscience 2012, New Orleans, LA.
Brooks, R. (1991b). Intelligence without reason. Proceedings of 12th International Joint Conference on Artificial Intelligence. 569-595.
Brooks, R. (1991b). Intelligence without representation. Artificial Intelligence, 47: 139-159.
Chalmers, D.J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies 2: 200-219.
Chalmers, D.J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford: Oxford University Press.
Chemero, A. (2009). Radical embodied cognition. Cambridge, MA: The MIT Press.
Damasio, A. (1990). Category-related recognition defects as a cue to the neural substrates of knowledge. Trends in Neuroscience 13: 95-98.
Dotov, D., Nie, L., \& Chemero, A. (2010). A demonstration of the transition from readiness-tohand to unreadiness-to-hand. PLoS ONE, 5: e9433.
Dotov, D., Nie, L., \& Chemero, A. (2017). Readiness-to-hand, unreadiness-to-hand, and multifactality. Journal of Mind and Behavior.
Gerlach, C., Law, I., \& Paulson, O.B. (2002). When action turns into words. Activation of motor-based knowledge during categorization of manipulable objects. Journal of Cognitive Neuroscience 14: 12301239.

Gibson, J. (1962). Observations on active touch. Psychological Review, 69: 477-490.
Gibson, J. (1966). The senses considered as perceptual systems. Boston: Houghton-Mifflin.
Gibson, J. (1979). The ecological approach to visual perception. Boston: Houghton-Mifflin.
James, W. (1904). Does "consciousness' exist? Journal of Philosophy, Psychology, and Scientific Methods 1: 477- 491.

Kellenbach, M.L., Brett, M., and Patterson, K. (2003). Actions speak louder than functions: The importance of manipulability and action in tool representation. Journal of Cognitive Neuroscience 15: 30-46.
Kello, C., \& van Orden, G. (2009). The emergent coordination of cognitive function. Journal of Experimental Psychology: General, 136: 551-568.
Kelso, J.A.S. (2009). Synergies: Atoms of brain and behavior. In D. Sternad (Ed.), Progress in Motor Control. Heidelberg, Germany: Springer, pp. 83-91.
Leibniz, G.W. (1714/1991). Monadologie. Trans. by N. Rescher. Pittsburgh, PA: University of Pittsburg Press.
Luo, S.X., Axel, R., \& Abbott, L.F. (2010). Generating sparse and selective third-order responses in the olfactory system of the fly. Proceedings of the National Academy of the Sciences (USA), 107: 1071310718.

Mante, V., Sussillo, D, Shenoy, K.V., \& Newsome, W.T. (2012). Selection and integration of relevant sensory evidence without gating of sensory inputs. Program No. 175.07. 2012. Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience. Miyazaki, M., Nakajima, Y, Kadota, H., Chitose, K., Ohtsuki, T., \& Kudo, K. (2004). 1/f-type fluctuation in human visuomotor transformation. NeuroReport, 15:1133-1136.
Raafat, R., Chater, N., and Frith, C. 2009. Herding in humans. Trends in Cognitive Sciences 13: 420-428.
Satori, G. and, Job, R. (1988). The oyster with four legs: A neuropsychological study on the interaction of visual and semantic information Cognitive Neuropsychology 5: 105-132.
Silberstein, M., \& Chemero, A. (2012). Complexity and extended phenomenological-cognitive systems. Topics in Cognitive Science 4: 35-50.
Silberstein, M., \& Chemero, A. (2015). Extending neutral monism to the hard problem. Journal of Consciousness Studies.
Strogatz, S. (2004). Sync: How Order Emerges From Chaos In the Universe, Nature, and Daily Life. New York, NY: Hyperion.
Thelen, E., and Smith, L.B. (1994). A dynamic systems approach to the development of cognition and action. Cambridge, MA: The MIT Press.
Van Orden, G., Holden, J., \& Turvey, M.T. (2003). Self-organization of cognitive performance. Journal of Experimental Psychology: General, 132:331-351.
Varela, F., Thompson, E. \& Rosch, E. (1991). The embodied mind: Cognitive science and human experience. Cambridge, MA: The MIT Press.
White, J.A., Klink, R., Alonso, A., \& Kay, A.R. (1998). Noise from voltage-gated ion channels may influence neuronal dynamics in the entorhinal cortex. Journal of Neurophysiology, 80: 262-269.
Yu, Y., Romero, R., \& Lee, T.S., (2005). Preference of sensory neural coding for $1 / \mathrm{f}$ signals. Physical Review Letters,94:108103.

# Convention-formation in iterated reference games 

Robert X.D. Hawkins, Michael C. Frank, Noah D. Goodman<br>\{rxdh, mcfrank, ngoodman\}@stanford.edu<br>Department of Psychology, Stanford University


#### Abstract

What cognitive mechanisms support the emergence of linguistic conventions from repeated interaction? We present results from a large-scale, multi-player replication of the classic tangrams task, focusing on three foundational properties of conventions: arbitrariness, stability, and reduction of utterance length over time. These results motivate a theory of convention-formation where agents, though initially uncertain about word meanings in context, assume others are using language with such knowledge. Thus, agents may learn about meanings by reasoning about a knowledgeable, informative partner; if all agents engage in such a process, they successfully coordinate their beliefs, giving rise to a conventional communication system. We formalize this theory in a computational model of language understanding as social inference and demonstrate that it produces all three properties in a simplified domain.


Keywords: conventions; pragmatics; communication

## Introduction

Just as drivers depend on shared behavioral conventions to safely navigate traffic, successful communication depends on a set of shared linguistic conventions. Speakers of different languages around the world refer to the same object in many different ways, yet when ordering a coffee in San Francisco, one can confidently use the English word "coffee" and be understood. How do these conventions - classically characterized by Lewis (1969) as arbitrary but stable solutions to recurring coordination problems - form in the first place?

While global conventions adopted and sustained throughout a large population of speakers may develop over longer time scales, we also effortlessly coordinate on local conventions - or conceptual pacts (Brennan \& Clark, 1996) within the span of a single dialogue. For example, when discussing possible conditions to use in an upcoming experiment, a team of collaborators might begin the meeting using long descriptions to refer to each condition but end the meeting using conventional terms like "condition A" and "condition B." Since global conventions are hypothesized to emerge through diffuse, repeated interactions of this more local kind (Garrod \& Doherty, 1994), the cognitive mechanisms underlying convention-formation in such interactions are of foundational interest.

In a seminal study by Krauss \& Weinheimer (1964), pairs of participants played a cooperative language game where they were presented with arrays of ambiguous shapes in randomized orders. The players were assigned the roles of director and matcher and allowed to talk freely. The matcher's goal was to rearrange their shapes to match the director's board, and the director's goal was to communicate useful descriptions. Over multiple rounds, descriptions were dramatically shortened: an early description like "upside-down martini glass in a wire stand," became simply "martini" by the
end. Later studies (e.g. Clark \& Wilkes-Gibbs, 1986) refined this paradigm, using larger arrays of tangram-like figures and emphasizing the intricate back-and-forth process through which speakers and listeners negotiate over references. The referring expressions generated by participants in these studies revealed a number of rich qualitative phenomena. Here, we focus on three that are both prescribed top-down by theories of convention-formation and also arise bottom-up as major axes of variation in our data: arbitrariness, stability, and the systematic reduction of utterance length over time.

Arbitrariness is a definitional property of conventions (Lewis, 1969): there must be multiple solutions that would be equally successful as long as both players "agree" (e.g. driving on the left vs. right side of the road). By the final round in a language game, for example, one pair might successfully use the expression 'dancer' to refer to a tangram, while another might use 'skater'. The other definitional property we consider is stability: it is in everyone's best interest to keep using a convention once established. Finally, reduction is more specific to the reference game paradigm and refers to the transformation of longer, complex expressions into simpler expressions over the course of interaction, as Krauss \& Weinheimer (1964) observed. While this broad phenomenon has been replicated many times, exactly what is reduced remains an open empirical question.

Theories of convention-formation differ primarily in the extent to which sophisticated social reasoning and common ground is required. At one extreme, agents use simple heuristic updating rules and do not need to represent or reason about other agents at all (Barr, 2004; Centola \& Baronchelli, 2015; Young, 2015). Simulations elegantly show how arbitrary signaling systems can spread and come to stably dominate large populations. However, due to their 'rich get richer' dynamic, it is not clear how simple heuristic updating mechanisms alone could account for reduction in repeated interaction. At the other extreme are theories in which agents recursively track what information is mutual knowledge, often formalized in a game theoretic setting (Lewis, 1969). Wilkes-Gibbs \& Clark (1992) and others have proposed that agents engage in a collaborative process of actively establishing mutual knowledge, though the mechanisms allowing conventions to emerge under such conditions have not been instantiated in a formal model to our knowledge.

In this paper, we argue for a theoretical position on the spectrum between these poles: conventions form when uncertain agents treat their partners' knowledge as ground truth. In other words, agents assume their partner is knowledgeably and rationally using some conventional lexicon mapping labels to meanings but are themselves initially unsure of its


Figure 1: Example trial in experimental interface. Both players could freely use the chat box, and the matcher could click and drag the tangram images.
identity. Through observing their partner's behavior in repeated actions, agents learn and adopt that lexicon, though their partner in fact begins in the same state of ignorance. When both agents independently adopt such a social learning strategy, they align to one another, coordinating on and implicitly creating shared conventions.

To motivate this theory, we first conduct a large-scale replication of the tangrams task on the web, which has traditionally been limited to relatively small sample sizes in the lab. We use distributions of lexical and syntactic features in the text corpus to operationalize arbitrariness, stability, and reduction, which have been difficult to analyze at a fine-grained level due to the sparseness of existing data. Taking these insights into account, we then formalize our theory in a computational model of communication in repeated reference games based on recent successes capturing language understanding as social inference (Goodman \& Frank, 2016; Goodman \& Stuhlmller, 2013). Finally, we show that this model qualitatively produces all three empirical signatures in a simplified domain inspired by the tangrams task.

## Replication of the Tangrams task

To collect a corpus of reference game dialogue that supports more detailed analyses of convention-formation, we ported the tangrams task used in Clark \& Wilkes-Gibbs (1986) to a real-time, multi-player web environment.

## Methods

Participants 200 participants were recruited from Amazon's Mechanical Turk and paired into dyads to play a realtime communication game using the framework in Hawkins (2015). We excluded games that terminated before the completion of 6 rounds and where participants reported a native language different from English, leaving a corpus of 67 complete games with a total of 9967 utterances.

Stimuli On every trial of the game, both participants were shown a $6 \times 2$ grid containing twelve tangram shapes, reproduced from Clark \& Wilkes-Gibbs (1986). Cells were labeled with fixed numbers from one to twelve in order to help par-
ticipants easily refer to locations in the grid (see Fig. 1).
Procedure After passing a short quiz about task instructions, participants were randomly assigned the role of either 'director' or 'matcher' and automatically paired into virtual rooms containing a chat box and grid of stimuli. Both participants could freely use the chat box to communicate at any time. The director's tangrams were fixed in place, but the matcher could click and drag the shapes to reorder them. The director had to send messages about the locations of different tangrams on their fixed board; the matcher had to identify the corresponding tangram shapes and move them to the correct locations. When the players were satisfied that their boards matched, the matcher clicked a 'submit' button that gave players feedback on their score (out of 12) and scrambled the tangrams for the next round. After six rounds, players were redirected to a short exit survey. We collected the raw text of every message sent and every swapping action taken by the matcher ${ }^{1}$

## Results

Arbitrariness and stability We begin by examining signatures of arbitrariness and stability in our data. We operationalize these concepts using the information-theoretic measure of entropy:

$$
H(W)=\sum_{w} P(w) \log P(w)
$$

where $P(w)$ denotes the distribution over frequencies of word tokens used within a game. Broadly speaking, entropy measures the predictability of a distribution. It is maximized when all elements are equally likely and declines as the distribution becomes more structured, i.e. when the probability mass is concentrated on a small subset of elements.

To derive predictions, we consider a permutation-test null model in which utterances on each of the six rounds are scrambled across games, designed to break any existing structure in each game's idiosyncratic word distributions. The mean empirical entropy should only differ from the null distribution of entropies generated from this scrambling process if both arbitrariness and stability hold.

First, note that if stability did not hold, scrambling would have no effect on the entropy within individual games: a given speaker would already use different words each round, and swapping out the identity of those words would not affect the entropy of the word distribution. There would be no structure to break.

If stability holds but arbitrariness does not, all players would adopt the single optimal (non-arbitrary) way to refer to each tangram. Therefore, the entropy of their word distributions also should not be affected by scrambling: a speaker's real words would be swapped out for the same tokens generated by another speaker. Scrambling wouldn't break the

[^87]|  | $\# 1$ | $\# 2$ | $\# 3$ | $\# 4$ | $\# 5$ | $\# 6$ | $\# 7$ | $\# 8$ | $\# 9$ | $\# 10$ |
| ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| unigrams <br> bigrams | a <br> looks like | like <br> like a | looks <br> a person | the <br> is a | one <br> to the | with <br> with a | to <br> the right | of <br> the left | and <br> the one | on <br> a square |

Table 1: Top 10 unigrams and bigrams with the highest reduction


Figure 2: Reduction phenomena. From left: (1) mean message length in words per tangram, (2) mean number of listener messages, (3) proportion of utterances containing adjectival clauses, (4) proportion of utterances containing subordinate clauses. Error bars are bootstrapped $95 \%$ CIs.


Part of Speech category
Figure 3: Reduction rates for different parts of speech. Error bars are bootstrapped 95\% CIs.
structure of the distribution, because the structure would be the same for all participants.

Finally, if both arbitrariness and stability hold, then different speakers would adopt different referring expressions that persist from round to round. Hence, scrambling should increase the average game's entropy from a relatively low level: each game's idiosyncratic, concentrated distribution of words would be mixed together to form more heterogeneous and therefore high-entropy distributions.

To test this prediction, we computed the average withingame entropy for 1000 different permutations of speaker utterances. Since this permutation scheme keeps the number of messages per participant constant and simply swaps out the content of those messages within each round, it controls for the fact that some speakers sent more messages than others and also that speakers in earlier rounds use more words (see next section). We found that our null distribution lay within the interval [4.88, 4.91], which is significantly higher than the true entropy (averaged across games) of 4.36, $p<0.001$. This pattern is consistent only with signatures of both arbitrariness and stability.
Reduction Next, we turn to a set of analyses examining reduction in utterance length over the course of the experiment. At the coarsest level, we find that the mean number of
words used by speakers decreases over time (see Fig. 2). This decrease replicates a highly reliable reduction effect found throughout the literature on iterated reference games (Brennan \& Clark, 1996; Krauss \& Weinheimer, 1964). Likely due to our purely textual (vs. spoken) interface, participants in our task used significantly fewer words overall than previously reported (e.g. an average of 20 words on the 1 st round, compared to 40 in Clark \& Wilkes-Gibbs (1986)) The following analyses break down this broad reduction into a finer-grained set of phenomena.

The next level of granularity motivating our model approach concerns which kinds of words are most likely to be dropped. Is the speaker adopting a shorthand where they drop uninformative function words, or are they simplifying or narrowing their descriptions by omitting meaningful details (Clark \& Wilkes-Gibbs, 1986)? We used the Stanford CoreNLP part-of-speech tagger (Toutanova, Klein, Manning, \& Singer, 2003) to count the number of words belonging to each part of speech in each message. Fig. 3 shows the percent reduction of different parts of speech from the first round to the sixth round. We find that determiners ('the', ' $a$ ', 'an') are the most likely class of words to be dropped with an $86 \%$ reduction rate, on average. Nouns ('dancer', 'rabbit') are the least likely class to be dropped with only an $62 \%$ rate. Closed-class parts of speech are strictly more likely to be dropped than open-class parts of speech.

While this finding is consistent with the possibility that speakers adopt a shorthand using more fragments as the game proceeds, we find a more complex dynamic by examining the table of unigrams and bigrams most likely to be dropped (see Table 11. Note that alongside dropped articles ('a', 'the'), there are a number of words that form conjunctions ('and') and modifiers ('of', 'with', 'the right'). In other words, it may be more likely that when function words are dropped, it is primarily as part of larger grammatical units that provide additional information in identifying the target.

We explicitly examined this hypothesis by running the Stanford constituency parser (Schuster \& Manning, 2016),
tagging the occurrence of subordinate/adverbial clauses ('sitting facing left') and adjectival clauses ('angel that is praying') $\rrbracket^{2}$ We found that both were reduced over the course of the game (see Fig. 2), lending additional support for the hypothesis that whole meaningful clauses are increasingly omitted. This result prompts a characterization of reduction where, due to uncertainty at the outset about the usefulness of any particular lexical unit, initial phrases pile on multiple partially redundant modifiers and descriptors. As the game progresses and ambiguity of reference decreases, these additional meaningful units become less useful and can be dropped. We return to this characterization more formally within the scope of our model below.

## Model

Here, we present a probabilistic model of language production under uncertainty, which captures several of the signature properties of conventions shown above. This model belongs to the family of Rational Speech Act (RSA) models, which have been successful in explaining a wide range of linguistic phenomena - including scalar implicature, adjectival vagueness, overinformativeness, indirect questions, and nonliteral language use - as arising from a process of recursive social reasoning. Most previous applications of RSA have focused on the listener's problem of language comprehension, but the puzzle of conventionalization is primarily a question of speaker production. An $n$th order pragmatic speaker trying to convey a particular state of affairs $s \in S$ assuming lexicon $\mathcal{L}$ is assumed to select an utterance $u \in \mathcal{U}$ by trading off its expected informativity (with respect to a rational listener agent) against its cost, usually based on length (Goodman \& Frank, 2016):

$$
S_{n}(u \mid s, \mathcal{L}) \propto \exp \left(\alpha \log L_{n-1}(s \mid u, \mathcal{L})-\operatorname{cost}(u)\right)
$$

where $\alpha$ is a soft-max optimality parameter controlling the extent to which the speaker maximizes over listener informativity. The listener, in turn, inverts the speaker model to reason about what underlying state $s$ the speaker is trying to convey, given their utterance $u$ :

$$
L_{n}(s \mid u, \mathcal{L}) \propto P(s) S_{n}(u \mid s, \mathcal{L})
$$

This recursion bottoms out in a literal listener who directly looks up the meaning of the utterance in the lexicon:

$$
L_{0}(s \mid u, \mathcal{L}) \propto \mathcal{L}(u, s) \cdot P(s)
$$

As in several other recent applications of RSA (Graf, Degen, Hawkins, \& Goodman, 2016), we use a graded semantics, where utterances are better or worse descriptions of particular referents. For instance, the utterance "dancer" may initially be expected to apply to a photorealistic image of a ballerina $(\mathcal{L}($ 'dancer', ballerina $)=0.99)$ more than an abstract

[^88]image of one ( $\mathcal{L}($ 'dancer', abstract ballerina $)=0.6)$, but apply to both better than a non-category member like an image of a $\operatorname{dog}(\mathcal{L}($ 'dancer', dog $)=0.05)$.

Our approach to convention-formation begins with the additional assumption of lexical uncertainty (Bergen, Levy, \& Goodman, 2016; Smith, Goodman, \& Frank, 2013). In other words, we assume that instead of having perfect knowledge of $\mathcal{L}$, the listener has uncertainty over the exact meanings of lexical items in the current context (i.e. it is initially unclear which of the ambiguous tangram shapes "the dancer" might refer to). They begin with some prior $P(\mathcal{L})$ about the identity their partner's true lexicon, which may be initially biased toward certain meanings. By conditioning on repeated observations of their partner's behavior, they use Bayes rule to infer this true lexicon:

$$
P_{L_{n}}(\mathcal{L} \mid d) \propto P(\mathcal{L}) \prod_{i} S_{n}\left(s_{i} \mid u_{i}, \mathcal{L}\right)
$$

where $d=\left\{s_{i}, u_{i}\right\}$ is a set of observations of $s_{i}$ and $u_{i}$ coming from previous exchanges $3^{3}$ The listener marginalizes over this posterior when interpreting the speaker's utterance:

$$
L_{n}(s \mid u, d) \propto \sum_{\mathcal{L}} P_{L_{n}}(\mathcal{L} \mid d) L_{n}(s \mid u, \mathcal{L})
$$

The speaker, in turn, considers what utterances would be most informative for such a listener:
$S_{n}(u \mid s, d) \propto \exp \left(\alpha \log \left(\sum_{\mathcal{L}} P_{S_{n}}(\mathcal{L} \mid d) L_{n-1}(s \mid u, \mathcal{L})\right)-\operatorname{cost}(u)\right)$
where the posterior over lexica $P_{S_{n}}(\mathcal{L} \mid d)$, uses the listener likelihood $L_{n-1}$. For the purposes of this paper, we fix the depth of recursion at $n=2$. This model is implemented in the probabilistic programming language WebPPL (Goodman \& Stuhlmller, electronic) ${ }^{4}$ Following Smith et al. (2013), we begin by showing how a random initial choice is taken to be evidence for a particular lexicon and becomes the base for successful communication even though neither party knows its meaning at the outset.

## Results

Arbitrariness and stability Consider an environment with two abstract shapes ( $\left\{s_{1}, s_{2}\right\}$ ), where the speaker must choose between two utterances ( $\left\{u_{1}, u_{2}\right\}$ ) incurring equal cost. Their prior $P(\mathcal{L})$ over the meaning of each utterance is given by a Beta distribution ${ }^{5}$, so on the first round both utterances are

[^89]

Figure 4: (A) Probability of speaker using $u_{1}$ to refer to $s_{1}$, broken out by initial observation: while players are initially ambivalent between the two labels (arbitrariness), the initial mapping is likely to persist (stability). (B) Accuracy rises as speaker and listener align. (C) When conjunctions are introduced into the grammar, utterances get shorter over time (reduction).
equally likely to apply to either shape. If the speaker was trying to get their partner to pick $s_{1}$, then, since each utterance is equally (un)informative, they would randomly sample one (say, $u_{1}$ ), and observe the listener's selection of a shape (say, $s_{1}$ ). On the next round, the speaker uses the observed pair $\left\{u_{1}, s_{1}\right\}$ to update their beliefs about their partner's true lexicon, uses these beliefs to generate a new utterance, and so on. To examine expected dynamics over multiple rounds, we forward sample many possible trajectories.

We observe several important qualitative effects in our simulations. First, the fact that a knowledgeable listener responds to utterance $u$ with $s$ provides evidence for lexicons in which $u$ is a good fit for $s$, hence the likelihood of the speaker using $u$ to refer to $s$ increases on subsequent rounds (see Fig 4 A ). In other words, the initial symmetry between the meanings can be broken by initial random choices, leading to completely arbitrary but stable mappings in future rounds. Second, because the listener is also learning the lexicon from these observations under the same set of assumptions, they converge on a shared set of meanings; hence, expected accuracy rises on future rounds (see Fig. 4B). Third, because one's partner is assumed to be pragmatic, agents can also learn about unheard utterances. Observing $d=\left\{u_{1}, s_{1}\right\}$ also provides evidence that $u_{2}$ is not a good fit for $s_{1}$ by Gricean maxims: if $u_{2}$ were a better fit for $s_{1}$, the speaker would have used it instead (Grice, 1975). Finally, failed references lead to conventions just as effectively as successful references: if the speaker intends $s_{1}$ and says $u_{1}$, but then the listener incorrectly picks $s_{2}$, the speaker will take this as evidence that $u_{1}$ actually means $s_{2}$ in their partner's lexicon and become increasingly likely to use it that way on subsequent rounds.

Reduction in utterance length Finally, we show how our model explains reduction of utterance length over multiple interactions. For utterances to be reduced, of course, they must vary in length. Motivated by our empirical observation that meaningful clauses are the primary unit of reduction, we extend our grammar to include conjunctions. This is one of the simplest ways to constructing longer utterances composition-
ally from lexical primitives, using the product rule:

$$
\mathcal{L}\left(u_{i} \text { and } u_{j}, o\right)=\mathcal{L}\left(u_{i}, o\right) \times \mathcal{L}\left(u_{j}, o\right)
$$

Analogous to our tangram stimuli, which have many ambiguous features and figurative perspectives that may be evoked in speaker descriptions, we consider a simplified scenario where speakers can refer to two different features of the two objects $\left\{o_{1}, o_{2}\right\}$. The speaker has four primitive words at their disposal - two words for shape $\left(\left\{u_{s 1}, u_{s 2}\right\}\right)$ and two for color $\left\{u_{c 1}, u_{c 2}\right\}$ - and has uncertainty over the initial meanings of all four.

While we established in the previous section that conventions can emerge over a reference game in the complete absence of initial preferences, players often bring such preferences to the table. A player who hears 'ice skater' on the first round of our tangrams task is more likely to select some objects more than others, even though they still have some uncertainty over its meaning in the context. To show that our model can accommodate this fact, we allow the speaker's initial prior meanings to be slightly biased. $u_{s 1}$ and $u_{c 1}$ are more likely to mean $o_{1} ; u_{s 2}$ and $u_{c 2}$ are more likely to mean $o_{2}$.

We ran 1000 forward samples of 6 rounds of speakerlistener interaction, and averaged over the utterance length at each round ${ }^{6}$ Our results are shown in Figure 4 C : the expected utterance length decreases systematically over each round. To illustrate in more detail how this dynamic is driven by an initial rational preference for redundancy relaxing as reference becomes more reliable, we walk step-by-step through a single trajectory.

Consider a speaker who wants to refer to object $o_{1}$. They believe their knowledgeable partner is slightly more likely to interpret their language using a lexicon in which $u_{s 1}$ and $u_{c 1}$ apply to this object, due to their initial bias. However, there is still a reasonable chance that one or the other alone actually refers strongly to $o_{2}$ in the true lexicon. Thus, it is useful to produce the conjunction " $u_{s 1}$ and $u_{c 1}$ " to hedge against this possibility, despite its higher cost. Upon observing the

[^90]listener's response (say, $o_{1}$ ), the evidence is indeterminate about the separate meanings of $u_{s 1}$ and $u_{c 1}$ but both become increasingly likely to refer to $o_{1}$. In the trade-off between informativity and cost, the shorter utterances remain probable options. Once the speaker chooses one of them, the symmetry collapses and that utterance remains most probable in future rounds. In this way, meaningful sub-phrases are omitted over time as the speaker becomes more confident about the true lexicon.

## General Discussion

In this paper, we revisited the classic phenomenon of convention-formation in a large-scale, text-based replication of the tangrams task. We argued that several key qualitative patterns in the data - arbitrariness, stability, and the reduction of utterance length over repeated interactions - can be explained by our model of informative communication under lexical uncertainty. This model formalizes a theory where conventions emerge via uncertain agents who assume their partner is knowledgably and informatively using some conventional lexicon. Through repeated observations of their partner's actions, agents learn this lexicon, thereby coordinating and aligning to one another.

Theories of convention-formation vary in the extent to which social reasoning about common ground is required. Our agents lie on a spectrum between the heuristic updating agents of Barr (2004) and the sophisticated agents of Clark \& Wilkes-Gibbs (1986), who collaboratively build up explicit representations of mutual knowledge. Speakers and listeners in our model implicitly coordinate their beliefs through a shared history of observations, which serves as "common ground" in an informal sense. They make critical use of pragmatic, social reasoning in order to learn meanings, but do not explicitly consider the fact that this history is shared, or represent their partner's own uncertainty.

By capturing reduction, which purely heuristic theories have not yet demonstrated, we showed that minimal assumptions of social reasoning go a long way in accounting for key phenomena. Still, our model falls short in some ways. For instance, because we do not provide a mechanisms for the listener agent to respond with confirmation, repair, or followup questions, we cannot make explicit predictions about the reduction in listener messages (as shown in Fig. 23) or the effect of listener input on the conventionalization process. These phenomena require our model to deal with planning over extended dialogues, and more sophisticated speech acts. Similarly, while our model was explicitly designed with linguistic conventions in mind, it remains to be seen whether the same formulation generalizes to broader behavioral conventions. For example, the real-time coordination games used in Hawkins \& Goldstone (2016) may not require players to reason about a structured lexicon with noise, but an action policy representation may play a similar role. While there remain many complex aspects of convention-formation in communication games left for future research, our approach nonethe-
less serves as a lower bound on the degree of social reasoning needed to capture lexical conventions in these games.

## Acknowledgements

We thank Nicole Maslan for her contributions during piloting. This work was supported by ONR grant N00014-13-1-0788 and a Sloan Research Fellowship to NDG. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

## References

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? Cognitive Science, 28(6), 937-962.
Bergen, L., Levy, R., \& Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. Semantics and Pragmatics, 9(20).
Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482.
Centola, D., \& Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. Proceedings of the National Academy of Sciences, 112(7), 1989-1994.
Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22(1), 1-39.
Frank, M. C., Goodman, N. D., \& Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. Psychological Science, 20(5), 578-585.
Garrod, S., \& Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. Cognition, 53(3), 181-215.
Goodman, N. D., \& Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences, 20(11), 818-829.
Goodman, N. D., \& Stuhlmller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. Topics in Cognitive Science, 5(1), 173-184.
Goodman, N. D., \& Stuhlmller, A. (electronic). The design and implementation of probabilistic programming languages. Retrieved from http://dippl.org
Graf, C., Degen, J., Hawkins, R. X. D., \& Goodman, N. D. (2016). Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions. In Proceedings of the 38th annual conference of the Cognitive Science Society.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and semantics (pp. 43-58). New York: Academic Press.
Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. Behavior Research Methods, 47(4), 966-976.
Hawkins, R. X. D., \& Goldstone, R. L. (2016). The formation of social conventions in real-time environments. PLoS ONE, 11(3), 1-14.
Krauss, R. M., \& Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. Psychonomic Science, 1(1-12), 113-114.
Lewis, D. (1969). Convention: A philosophical study. Harvard University Press.
Schuster, S., \& Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In LREC 2016.
Smith, N. J., Goodman, N. D., \& Frank, M. C. (2013). Learning and using language via recursive pragmatic reasoning about other agents. In NIPS (pp. 3039-3047).
Toutanova, K., Klein, D., Manning, C. D., \& Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In NAACL-HLT (pp. 173-180).
Wilkes-Gibbs, D., \& Clark, H. H. (1992). Coordinating beliefs in conversation. Journal of Memory and Language, 31(2), 183-194.
Young, H. P. (2015). The evolution of social norms. Annual Review of Economics, 7, 359-387.

# Sampling frames, Bayesian inference and inductive reasoning 

Brett K. Hayes (b.hayes@unsw.edu.au)<br>Stephanie Banner (steph22banner@gmail.com)<br>Daniel J. Navarro (d.navarro@unsw.edu.au)<br>School of Psychology, University of New South Wales<br>Sydney, 2052, Australia


#### Abstract

We outline and test a Bayesian model of the effects of evidence sampling on property induction. Our model assumes that people are sensitive to the effects of different sampling frames applied to sampled evidence. Two studies tested the model by comparing property generalization following exposure to samples selected because they belong to the same taxonomic category or because they share a salient property. Both studies found that category-based sampling led to broader generalization than property-based sampling. In line with model predictions, these differences were attenuated when a mixture of positive and negative evidence was presented (Experiment 1) and when category-property relations were probabilistic rather than deterministic (Experiment 2).


Keywords: Inductive reasoning; Sampling; Hypothesis testing; Bayesian models; Categorization

## Introduction

Inductive reasoning - the ability to make plausible guesses given inconclusive evidence - is one of the central topics in cognitive science. Much of the traditional work on the topic has emphasized the importance of similarity between premise and conclusion categories (see Hayes \& Heit, 2013, for a review). While undoubtedly useful, the similarity-based approach overlooks a crucial component of induction: people's inductive inferences are strongly influenced by their beliefs about how the evidence was sampled (e.g., Xu \& Tenenbaum, 2007). This phenomenon is referred to as sensitivity to sampling, and there is considerable evidence that human reasoners show exactly this sensitivity

One form of sampling sensitivity occurs when an argument assembled by a knowledgeable and helpful teacher is evaluated quite differently than a set of random facts, even if - by chance - the random process happens to have sampled the same set of facts. In the reasoning literature, this was first discussed by Medin, Coley, Storms and Hayes (2003) in their relevance theory of induction. They suggested that reasoners often make the pragmatic assumption that premise categories are selected to highlight a salient relation, which is then used to guide inference. For example, on learning that zebras and skunks share a novel property, people may infer that the property involves "having stripes" and generalize accordingly. More recently, the formal foundations for pragmatic inference have been established using Bayesian pedagogical sampling models, that model human inductive reasoning by assuming that helpful teachers select informative evidence (Voorspoels, Navarro, Perfors, Ransom, \& Storms, 2015; Ransom, Perfors \& Navarro, 2016;

Shafto \& Bonawitz, 2015). This account is supported by empirical work showing that many inductive phenomena (e.g., premise non-monotonicity, integration of positive and negative evidence) depend on the assumption of a helpful teacher (Ransom et al., 2016; Voorspoels, Navarro, Perfors, Ransom, \& Storms, 2015).

A second kind of sensitivity arises from the so-called "strong versus weak" sampling distinction. Under strong sampling, the learner observes a set of exemplars (e.g., premise categories) that are constrained to possess property p. Under weak sampling, no such constraint exists. Early work highlighted the fact that even this simple constraint can produce substantial changes to how a Bayesian reasoner make inferences (Tenenbaum \& Griffiths, 2001), but many applications of the strong/weak distinction have tended to conflate it with helpful/random sampling (e.g., Xu \& Tenenbaum, 2007), and those that do not have found mixed evidence (e.g., Navarro, Dry \& Lee, 2012). Although there are good reasons to expect helpfully sampled evidence to be similar to strongly-sampled evidence (e.g., Ransom et al., 2016), it is not obvious whether (or when) people are sensitive to sampling assumptions if no helpful teacher is available. Perhaps people are capable of taking a hint from a helpful teacher, but otherwise are largely insensitive to sampling assumptions. Given other evidence that people struggle with conditional probability (e.g, Fiedler, 2012) this is not an implausible idea.

## How sampling frames shape induction

In this paper, we approach the problem from a different perspective, and consider other ways in which data can be sampled in a constrained way. The statistics literature, for instance, emphasizes the importance of a sampling frame (Jessen, 1978): when designing a survey, the researcher may not be able to sample uniformly at random from the entire population of interest, but is instead forced to sample from a restricted subset. When interpreting such data, those properties of the observed data that are attributable to the sampling frame do not require theoretical explanation, as they are deemed an artifact of the sampling process.
The effect of a sampling frame can be substantial. Imagine that you want to learn what plants make you sneeze. The potential search space is large so we apply a sampling frame - we first test a particular category of plant (e.g., sunflowers) - and find that most sunflowers cause us to sneeze. In this situation, the fact that we have never sneezed at a daisy is


Implausible: SP+ not allowed by hypothesis


Figure 1. Schematic of the hypothesis space about the distribution of property $P+$ and its absence ( $P-$ ). Dark quadrants show the hypothesized extension of $P$ across the target category $(S)$ and non-targets ( $L$ ). Dots represent observation of a sample of small birds that has the property $S P+$. For a hypothesis to be plausible it must allow the $S P+$ case to exist.
irrelevant: it can be attributed to the sampling frame. In this context, absence of evidence is not evidence of absence.

Now consider the effect of shifting the sampling frame. Suppose instances are selected because they share the property of interest (e.g., they give positive result on an allergy test). If most of this sample was sunflowers then the absence of daisies might be seen as inductively informative: it suggests that the allergic reaction is limited to the observed category. Despite the fact that neither scenario involves a helpful teacher, the mere presence of a sampling frame allows
the same data to lead to different generalizations (cf. Hsu, Horng, Griffiths, \& Chater, 2016).

There is evidence that people are sensitive to the sampling frame. Lawson and Kalish (2009) presented participants with samples of animals (small birds) that shared a novel property ("has plaxium blood") and manipulated the way exemplars were sampled. In the "category sampling" condition they were told that items were sampled from a taxonomic category (i.e., the frame selects small birds). In the "property sampling" condition people were told that exemplars with plaxium blood were selected. People in the property sampling condition were less likely to generalize the property to other animals. Lawson and Kalish (2009) noted that this result was inconsistent with similarity-based accounts of induction, but they did not explain why the differences occurred.

As it happens, this pattern of results is exactly what one would expect from a probabilistic reasoner who is sensitive to the sampling frame. Later we present a formal model, but the qualitative intuition is simple. Suppose the learner has observed small birds ( $S$ ) with plaxium blood $(P+$ ), and is trying to determine whether large birds $(L)$ also possess plaxium blood. Subject to the constraint that large and small birds both exist, there are six hypotheses consistent with the observations, as shown in Figure 1, and three that are not.

Now consider the plausibility of these six hypotheses under different sampling frames, illustrated by the red rectangles in Figures 2. In category sampling, it is plausible to assume that if any small birds did not have plaxium blood, the $S P$ - case would have been observed. The lack of such observations strengthens three hypotheses and weakens three others. Notably, two remaining hypotheses allow large birds to have plaxium blood $(L P+)$. By contrast, in property sampling it is reasonable to assume that if any large birds had plaxium blood we should have seen the $L P+$ case. The fact that they were not leaves only two viable hypotheses, both of which restrict property $P$ to the target category. Accordingly, generalization is more restricted under property sampling.

Category sampling: frame selects S


Property sampling: frame selects $\mathrm{P}+$


Figure 2. The effect of sampling frame. When the data consist solely of small birds with plaxium (SP+), plausible hypotheses are those for which only $S P+$ is allowed by the sampling frame and the hypothesis. Consequently, $L P+$ is less plausible under property sampling and the learner does not generalize beyond the observed $S P+$ case.

## Experiment 1

Our experimental work replicates the findings of Lawson and Kalish (2009), and extends them in a way that tests our "sampling frames" explanation. In the first experiment, we considered the impact of explicit negative evidence. If a learner encounters non-target category members that lack property $P$, the differences between the two sampling conditions should attenuate. Explicit negative evidence should have a large effect in the category sampling condition, but only a modest effect under property sampling. We expect this difference because property sampling already provides implicit negative evidence, so the added value of the explicit negative evidence is diminished.

Experiment 1 tested these predictions by presenting participants with identical evidence samples obtained via category or property sampling. Half the participants received positive evidence about members of a target category, as per Lawson and Kalish (2009), and half received additional negative evidence about non-target category members. All participants were then asked to judge whether the novel property generalized to other categories.

## Method

Participants. 92 UNSW students (63 female), participated for course credit or payment. The mean age was 20.9 years.
Design and Procedure. The experiment used a $2 \times 2$ between subjects design with equal numbers in each condition.

The procedure for the positive evidence only groups was patterned after Lawson and Kalish (2009). Participants were told they were investigating the properties of animals on a novel island. In the category sampling condition, participants were told that only small birds were sampled from the island. In the property sampling conditions, they were told that only animals with plaxium blood were sampled from the island. Exemplars were revealed as follows: on each of 20 trials, participants could click on one of a large number of on-screen boxes to see an exemplar (each depicted by a unique picture of a small bird), and to learn if the animal had plaxium blood. In the positive evidence condition, all 20 exemplars sampled had plaxium blood.

For the positive+negative evidence groups the procedure was identical, except that there were five trials at the end in which "new" samples from the island were presented. Each of these revealed a single instance from other animal categories (crow, seagull, eagle, squirrel, frog) that did not have plaxium blood. These five trials were always presented, in random order, at the end of sampling phase.

After the learning phase, all participants proceeded to a generalization test. On each of six trials, participants were shown a picture of an animal and asked to estimate the number of such animals from a sample of ten that would have plaxium blood ( $0-10$ ). The test categories included a member of the same target category that was presented during sampling (a novel picture of a sparrow) and five categories that varied in similarity to the target (pigeon, owl, ostrich, mouse, lizard). Test item order was randomized.


Figure 3: Experiment 1. Test phase generalization in each experimental condition.

## Results and Discussion

Generalization scores (out of 10) for all conditions are shown in Figure 3. Visual inspection suggests that the positive-only condition people generalized more narrowly under property sampling (black squares) than under category sampling (black circles). Moreover, this difference is less pronounced when explicit negative evidence is provided (in grey).

More formally, a mixed effects ANOVA revealed that people were less willing to generalize as similarity decreased (left to right in Figure 3; linear trend contrast: $F(1,84)=$ $420.07, p<.001$ ). Generalization to non-target categories was greater following category than property sampling, $F(1,84)=$ $12.36, p=.001$, and when only positive evidence was encountered during sampling, $F(1,84)=39.54, p<.001$. The critical finding, however is the interaction: the difference in generalization between category and property sampling was larger in the positive evidence only condition than in the positive + negative condition, $F(1,84)=5.81, p=.02$.

These results are exactly what we expected: despite the fact that participants in the category and property sampling groups saw exactly the same information, generalization of the novel property was narrower following property sampling. This replicates the main finding of Lawson and Kalish (2009), showing that people's inductive inferences are sensitive to the sampling frame. Moreover, the data supported a novel prediction of our sampling explanation: presentation of negative evidence had greater impact on generalization following category sampling than property sampling.

## Experiment 2

In the next experiment we consider a second manipulation that should - according to the sampling account - attenuate the difference between category and property sampling: ambiguous evidence. In Experiment 1, every member of the target category had the novel property. In Experiment 2, we considered cases where some of the evidence is ambiguous, by including some observations where the plaxium status of the entity was unknown. The qualitative intuition here is that this should introduce uncertainty about the distribution of the property within the target category. Accordingly, the evidentiary value of the data should decrease, leading to a less pronounced difference between the two sampling conditions.

## Method

Participants. 80 UNSW students ( 76 female), participated for course credit or payment. The mean age was 19.4 years.
Design \& procedure. The experiment used a $2 \times 2$ between subjects design, with equal numbers in each condition. The procedure for the deterministic evidence conditions was identical to the positive evidence only conditions in Experiment 1. The procedure for the probabilistic evidence conditions was similar, except that during the sampling phase participants saw an additional five category or property sampling trials. On these trials, additional small birds were presented whose blood type was unknown due to a "machine error". These trials were randomly interspersed with the other trials. The generalization test was the same as Experiment 1.

## Results and Discussion

Generalization scores are shown in Figure 4. As in Experiment 1, generalization of the novel property decreased as similarity to the target category decreased (linear trend contrast: $F(1,76)=117.94, p<.001$. Overall, generalization to non-target categories was greater following category than property sampling, $F(1,76)=8.88, p=.004$. Notably, there was a significant interaction between sampling condition and evidence certainty, $F(1,76)=5.25, p=.03$. Figure 4 shows that the differences in generalization between category and property sampling were relatively large when the evidence was deterministic, but decreased when the observed evidence was probabilistic.

The results for the deterministic evidence condition replicate the earlier finding that property sampling leads to narrower generalization than category sampling. Consistent with the predictions of our model, the difference between sampling conditions was reduced when the relationship between the target property and category was probabilistic.

## Bayesian reasoning with sampling frames

The sampling explanation outlined at the start of the paper provides an intuitive explanation of our results: in this section we provide a more formal account, introducing an inductive reasoning model that accommodates the effect of the sampling frame within the Bayesian framework introduced by Tenenbaum and Griffiths (2001).


Test phase categories

Category Sampling, Deterministic Evidence
Property Sampling, Deterministic Evidence

- Category Sampling, Probabilistic Evidence
Property Sampling, Probabilistic Evidence
Figure 4: Experiment 2. Test phase generalization in each experimental condition.

A Bayesian analysis of the inductive problem proceeds as follows. The test categories consist of items that belong to different taxonomic classes (birds, mammals, reptiles) and vary in size (small, medium, large, and huge). Given this, we define a hypothesis space $H$ by combining these two characteristics. A hypothesis $h$ is admissible if it includes only a single taxonomic class (e.g., birds only) or allows all animals to possess plaxium. Similarly, it is admissible if it specifies a "connected" region on the size dimension (e.g., small-or-medium is allowed, but small-or-huge is not). For simplicity, the Bayesian model assigns equal prior probability $P(h)$ to all hypotheses, with one exception: to account for the fact that people are less willing to generalize across taxonomic classes than across animal sizes, hypotheses that allows all animals to have plaxium blood are only $1 / 5$ as plausible as hypotheses restricted to a single class.
When presented with a set of observations $x$, the learner updates the prior distribution to a posterior via Bayes' rule:

$$
P(h \mid x, f)=\frac{P(x \mid h, f) P(h)}{\sum_{h^{\prime} \in \mathcal{H}} P\left(x \mid h^{\prime}, f\right) P\left(h^{\prime}\right)}
$$

In this expression, the likelihood term $P(x \mid h, f)$ describes the probability of observing the data $x$ if hypothesis $h$ is true and the sampling frame $f$ applies. When determining the probability that a test item $y$ possesses plaxium blood, a Bayesian learner aggregates the posterior probability assigned to those hypotheses $h$ that assign the test item y to the consequential set:


Figure 5. Inductive inferences made by the Bayesian model plotted as a function of test category, sampling condition and evidence type. See main text for details.

$$
P(y \in c \mid x, f)=\sum_{h \mid y \in h} P(h \mid x, f)
$$

The critical feature of this model is the fact that the likelihood term $P(x \mid h, f)$ is sensitive to the sampling frame. Under category sampling, the fact that all observations happen to be small birds is of no evidentiary value: the sampling frame $f$ only admits small birds, and no explanation for this is required. In this sampling regime, a good hypothesis is required to explain the fact that all observations are plaxium positive. If we assume a noisy relationship, where $\theta>.5$ denotes the probability that an animal that falls within the relevant category possesses plaxium blood, then the likelihood becomes:

$$
P(x \mid h, f)=\left\{\begin{array}{cl}
\theta & \text { if } x \in h \\
1-\theta & \text { otherwise }
\end{array}\right.
$$

Under property sampling, this pattern is reversed: the sampling frame admits only plaxium positive observations, and no explanation for this is required. Instead, the data $x$ that the learner must explain is the fact that all the animals are small birds. Again assuming a noisy relationship,

$$
P(x \mid h, f)=\left\{\begin{array}{cl}
\theta /|h| & \text { if } x \in h \\
(1-\theta) /|h| & \text { otherwise }
\end{array}\right.
$$

In this expression, the normalizing term $|\mathrm{h}|$ denotes the "size" of the hypothesis. For a hypothesis that predicts $m$ species to be plaxium positive and $n$ species to be plaxium negative,

$$
|h|=\theta m+(1-\theta) n
$$

Formal details notwithstanding, the main point of these equations is to highlight the fact that the different sampling frames involved ensures that property sampling imposes a size principle (Tenenbaum \& Griffiths 2001) and category sampling does not. When a size principle applies, Bayesian learners will tend to assign more belief to smaller hypotheses, and as a consequence will generalize narrowly. This is illustrated in the top panel of Figure 5 which plots the generalizations made by the Bayesian model when presented
with 20 plaxium positive small birds, setting $\theta=0.6$. As one might expect, the Bayesian model generalizes more narrowly under property sampling.

In Experiment 1, we found that the difference between the two sampling schemes attenuated when participants were presented with plaxium negative observations from nontarget categories, and generalizations narrowed in general. As shown in the middle panel of Figure 5, this is exactly what the Bayesian model does. Regardless of sampling scheme, the negative evidence serves to decrease the plausibility of larger hypotheses (as they are now somewhat inconsistent with the new data), but this has a much smaller effect in the property sampling condition simply by virtue of the fact that these hypotheses were already judged to be somewhat implausible. Accordingly, the Bayesian model produces narrower generalizations and the difference between the two conditions becomes smaller.

In Experiment 2, participants were presented with additional "ambiguous" observations (small birds that may or may not have been plaxium positive). This manipulation is expected to cause people to suspect a noisier relationship between the category and the observed plaxium status, which we operationalize by setting a lower value for $\theta$. When we set $\theta=.55$, we obtain the generalization gradients shown in the right panel of Figure 5. As expected, the Bayesian model produces an attenuated effect of sampling.

## General Discussion

Traditionally, models of property induction (e.g., Osherson et al., 1990) have focused on the similarity between the categories known to possess a property and other categories to which the property might be generalized. Although category similarity is undoubtedly an important component of induction, the current work highlights the additional impact of beliefs about how observed data is sampled. In both experiments, identical sets of observations led to very different patterns of generalization depending on beliefs about how the observations were selected. In the positive evidence condition in Experiment 1 and the corresponding
deterministic condition in Experiment 2, evidence sampling based on shared category membership led to broader generalization of the target property than evidence sampling based on a shared property.

This result shows that people are sensitive to the effects of particular constraints or sampling frames that are imposed on the observations. In category-based sampling, the absence of observations of members of other categories that share a target property is not necessarily seen as evidence of absence. In property-based sampling, the absence of such observations can be seen as evidence that the property does not project beyond the target category. This phenomenon is naturally accommodated by a Bayesian inductive reasoning model. Moreover, this theoretical perspective allowed us to generate two novel predictions. The effect of sampling frame attenuates when explicit negative evidence is added or when ambiguity is introduced to the sample. Both of these effects are captured by the Bayesian model.

Our Bayesian approach suggests additional factors that should moderate the impact of sampling frames. For example, differences in generalization patterns between types of sampling is likely to depend on beliefs about category base rates. In property sampling for example, if members of both the target category (e.g., small birds) and non-target categories (e.g., various types of large birds) are believed to be relatively common, then the fact that the sample of animals with plaxium blood contains no large birds is highly informative. In contrast, if large birds were uncommon, then the absence of large birds with plaxium blood does not license strong conclusions about property generalization.

Previous work (Ransom et al., 2015; Shafto \& Bonawitz, 2015) has shown that inductive inferences are sensitive to intentional factors associated with sample selection (e.g., whether the observations were chosen by a helpful agent to illustrate the breadth of a hypothesis). The current work, together with that of Lawson and Kalish (2009), highlights the importance of a novel factor in induction, namely sensitivity to different types of conditionalization or filtering of the evidence samples on which inferences are based. While this is a new finding in the domain of induction, it bears some resemblance to results observed in probability judgment tasks (see Fielder, 2012 for a review). Fiedler, Brinkman, Betsch and Wild (2000) for example, presented different groups with different types of conditionalized samples. One group saw instances of women who had received a positive breast scan result, and learned whether each woman had breast cancer. Another group saw instances of women with breast cancer and learned whether they had received a positive breast scan. As in the current work, people were sensitive to these different types of sample conditionalization, with the two groups generating very different estimates of the probability that a woman with a positive scan had cancer. In the Fiedler, et al. (2000) study however, the different types of conditionalization led to differences in the characteristics of the instances observed in each sample. The current work goes further, by showing that very different patterns of inference
emerge when identical evidence samples are selected via different types of sampling frames.

## Acknowledgments

This work was supported by Australian Research Council Discovery Grant DP150101094 to the first author. We thank Jeremy Ngo and Minoli Jinasena for assistance in running the experiments.

## References

Fiedler, K., Brinkmann, B., Betsch, T., \& Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. Journal of Experimental Psychology: General, 129, 399-418.
Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In B. H. Ross (Ed.), The Psychology of Learning and Motivation (Volume 57, pp. 1-55). Cambridge, MA: Academic Press
Hayes, B. K., \& Heit, E. (2013). Induction. In D. Reisberg (Ed.) Oxford Handbook of Cognitive Psychology. Oxford University Press, New York, USA.
Hsu, A. S., Horng, A., Griffiths, T. L., \& Chater, N. (2016), When absence of evidence is evidence of absence: Rational inferences from absent data. Cognitive Science, 1-13.
Jessen, R. J. (1978). Statistical survey techniques. New York: Wiley.
Lawson, C. \& Kalish, C. (2009). Sample selection and inductive generalization. Memory \& Cognition, 37, 596607.

Medin, D., Coley, J., Storms, G., \& Hayes, B. (2003). A relevance theory of induction. Psychonomic Bulletin \& Review, 10, 517-532.
Navarro, D., Dry, M. \& Lee, M. (2012). Sampling assumptions in inductive generalization. Cognitive Science, 36(2), 187-223.
Osherson, D., Smith, E., Wilkie, O., López, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.
Ransom, K., Perfors, A., \& Navarro, D. (2016). Leaping to conclusions: Why premise relevance affects argument strength. Cognitive Science, 40, 1775-1796.
Shafto, P., \& Bonawitz, E. (2015). Choice from intentionally selected options. In B. H. Ross (Ed.). The Psychology of Learning and Motivation (Volume 63, pp. 115-139). Cambridge, MA: Academic Press
Tenenbaum J. \& Griffiths T. (2001). Generalization, similarity and Bayesian inference. Behavior \& Brain Sciences, 24, 629-640.
Voorspoels, W., Navarro, D., Perfors, A., Ransom, K., \& Storms, G. (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. Cognitive Psychology, 81, 1-25.
Xu, F., \& Tenenbaum, J. (2007). Sensitivity to sampling in Bayesian word learning. Developmental Science, 10, 288297.

# The adaptive evolution of early human symbolic behavior 

Katrin Heimann<br>Aarhus University<br>Riccardo Fusaroli<br>Aarhus University<br>Sergio Gonzalez de La Higuera Rojo<br>Aarhus University<br>Niels Nørkjær Johannsen<br>Aarhus University<br>Felix Riede<br>Aarhus University<br>Nicolas Fay<br>University of Western Australia<br>Marlize Lombard<br>University of Johannesburg<br>\section*{Kristian Tylén}<br>Aarhus University


#### Abstract

Dating back as far as 100 ka , the Blombos ochre and the Diepkloof ostrich egg engravings are considered among the earliest fossilized evidence of human symbolic behavior. Of special interest to this study is the temporal trajectory spanning more than 30 thousand years from earlier simpler parallel line patterns to later complex cross-hatchings suggesting adaptive compositional development. Through a series of three psychophysical experiments we test the hypotheses that the line engravings at each site evolved to become 1) more salient to the human perceptual system, 2) more discriminable from each other, and 3 ) increasingly associated with symbolic intent. Our findings suggest that just as instrumental tools have been found to undergo cumulative refinements in adaptation to their function, the ochre and egg shell engravings evolved adaptively to become more fit for their cognitive function as signs.


# Children's social referencing reflects sensitivity to graded uncertainty 

Emily Hembacher<br>ehembach@stanford.edu<br>Department of Psychology<br>Stanford University

Benjamin deMayo<br>bdemayo@stanford.edu<br>Department of Psychology<br>Stanford University

Michael C. Frank<br>mcfrank@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

The ability to monitor epistemic uncertainty is critical for selfdirected learning. However, we still know little about young children's ability to detect uncertainty in their mental representations. Here we asked whether a spontaneous information gathering behavior - social referencing - is driven by uncertainty during early childhood. Children ages 2-5 completed a word-learning task in which they were presented with one or two objects, heard a label, and were asked to put the labeled object in a bucket. Referential ambiguity was manipulated through the number of objects present and their familiarity. In Experiment 1, when there were two novel objects and a novel label, the referent was ambiguous; when there were two familiar objects, or only one novel or familiar object, the referent was known or could be inferred. In Experiment 2, there were either two novel objects, two familiar objects, or one familiar and one novel object; in the latter case the referent could be inferred by excluding the familiar object. To further manipulate the availability of referential cues, the experimenter gazed at either the target or the center of the table while labeling the object. In both experiments, children looked at the experimenter more often while making their response when the referent was ambiguous. In Experiment 2, children also looked at the experimenter more when there was one familiar and one novel object, but only when the experimenter's gaze during labeling was uninformative. These results suggest that children's social referencing is a sensitive index of graded epistemic uncertainty.


Keywords: social referencing; help seeking; word learning; uncertainty.

Preschoolers quickly learn new concepts, rules, and language. They also actively explore and ask questions in ways that seem targeted to maximize learning (Chouinard, Harris, \& Maratsos, 2007; Schulz \& Bonawitz, 2007). However, we still have an incomplete understanding of young children's ability to monitor their own mental states, in particular, their epistemic uncertainty (Sodian, Thoermer, Kristen, \& Perst, 2012). Do preschool-aged children monitor uncertainty and actively guide their learning behaviors on the basis of this monitoring, or is early learning better characterized as a process of integrating information that is largely generated externally, for example, by social partners who act as teachers (Csibra \& Gergely, 2006)?

A hallmark of successful uncertainty monitoring is being less confident when the probability of accuracy is lower (Robinson, Johnson, \& Herndon, 1997). This ability includes awareness of complete ignorance, but also of graded evidence in mental representations, which is considered important for predicting outcomes and regulating behavior (Lyons \& Zelazo, 2011). During adulthood, accurately representing one's own learning progress allows for efficient self-directed study and predicts learning outcomes (Dunlosky \& Rawson, 2012). There is mixed evidence about whether young children can
accomplish this type of self-monitoring. For example, 3-year-olds report being equally confident about correct and incorrect responses in memory tasks (Hembacher \& Ghetti, 2014). Preschoolers report being less confident when they are wrong in other tasks, but they are typically overconfident overall (Coughlin, Hembacher, Lyons, \& Ghetti, 2015; Lipowski, Merriman, \& Dunlosky, 2013). However, these studies may underestimate young children's uncertainty monitoring, as they typically rely on explicit metacognitive reports. Children may learn to respond appropriately to uncertainty in everyday learning situations before they can bring it fully into consciousness and report on it.

Several studies have provided evidence that children's spontaneous information-seeking behaviors might track uncertainty. Call and Carpenter (2001) had 2-year-olds choose between several tubes to find a hidden sticker. They found that the toddlers were more likely to peek inside a tube before choosing when they had not seen the baiting of the tubes compared to when they had, suggesting they were aware of their ignorance and managed to delay their response until they were sufficiently confident. In another study, Goupil, Romand-Monnier, and Kouider (2016) found that 20-montholds were more likely to seek help by looking at their parents when they were unable to respond accurately in a memory task. These spontaneous information-gathering behaviors may provide a window into early uncertainty monitoring, and allow us to ask questions about its development.

Here, we focus on the role of uncertainty in guiding social referencing - one form of information gathering - during word learning. Referencing a social partner can provide several types of disambiguating information. For example, children can follow a speaker's gaze direction to infer the referent of a new word, as people tend to look at objects they are referring to. By the second year of life infants follow a speaker's gaze and map labels to objects on the basis of gaze direction (Baldwin, 1991). There is also evidence that infants' propensity for gaze-following predicts later language development (Carpenter, Nagel, Tomasello, Butterworth, \& Moore, 1998), highlighting the importance of this behavior for learning. In addition to monitoring gaze direction, children may reference a social partner's emotional reaction to a stimulus or event, which can help disambiguate the appropriateness of a response (Walden \& Ogan, 1988). Finally, looking at a social partner can be taken as a bid for help (Vredenburgh \& Kushnir, 2015), and may result in explicit instruction.

Social referencing can be an efficient source of disambiguating information, but is it driven by uncertainty during
early childhood? It could be that social referencing is not costly enough to require selectivity, or that uncertainty signals are too weak to drive information-seeking behaviors in young children. Similarly, other learning mechanisms such as the privileging of social information (Ho, MacGlashan, Littman, \& Cushman, 2017) or tracking of regularities in the environment (Yurovsky \& Frank, 2015) may be sufficiently powerful to obviate the need for uncertainty monitoring in preschoolaged children.

The present work asks whether preschoolers reference a speaker more frequently when the referent of the speech is ambiguous. This work adapts a paradigm used by Vaish, Demir and Baldwin (2011) in which 13- to 18-month-olds sat across from an experimenter who produced a label (e.g., "a modi!") in the presence of one or two novel objects. Infants looked towards the experimenter more often when there were two objects present, suggesting that infants' social referencing is driven by referential ambiguity. Here we adapt this procedure for use with preschoolers, who have a richer behavioral repertoire compared to infants, and may not reference social information based on uncertainty for the reasons discussed previously. We ask whether preschoolers look more at a social partner when they are uncertain about the identity of a referent (Experiment 1) and whether they are sensitive to graded uncertainty based on the amount of disambiguating evidence available (Experiment 2).


Figure 1: Study design for Experiments 1 and 2.

## Experiment 1

In Experiment 1, we examined whether children would visually reference a speaker more often when the speaker produced a referentially ambiguous label compared to an unambiguous label. Children sat across from an experimenter who labeled an object on the table between them (Figure 1). The experimenter then asked the child to place the named object in a bucket. Across trials, there were either one or two objects on the table, which were either familiar or novel to the child. This design allowed us to test whether merely having more than one object present is sufficient to increase social referencing (which could not be ruled out by Vaish et al.), or if referential ambiguity (and thus epistemic uncertainty) is the underlying factor. If the latter is true, we expected children to increase their looking to the experimenter only on trials with
two unfamiliar objects, when the object-label mapping was not known and could not be inferred.

We were interested in the amount of social referencing children exhibited across the trial. We considered four different phases of each trial based on the notion that children might expect different social information at different stages of the task. Specifically, we predicted that children might expect the speaker's gaze direction to be informative during the labeling itself, as speakers tend to look at objects they refer to. We predicted that later in the trial, as children reached for an object and placed it in the bucket, they might expect evaluative feedback about their choice (e.g., facial expressions of encouragement or discouragement).

## Methods

Participants We recruited a planned sample of 80 children ages 2-5 years from the Children's Discovery Museum in San Jose, California ${ }^{1}$ The sample included 20 2-year-olds (mean age 31.97 months), 203 -year-olds (mean age 42.65 months), 204 -year-olds (mean age 55.85 months), and 205 -year-olds (mean age 65.21 months). An additional 20 children participated but were removed from analyses because they heard English less than $75 \%$ of the time at home $(n=10)$, because they were unable to complete at least half of the trials in the task $(n=4)$, because of parental interference $(n=1)$, or due to experimenter or technical errors $(n=5)$.
Stimuli and Design Children were presented with one or two objects, heard a label, and were asked to put the labeled object in a bucket. Half of the objects were selected to be familiar to children (e.g., a cow) and half were selected to be novel (e.g., a nozzle). There were four trial types: onefamiliar, one-novel, two-familiar, and two-novel. There were three trials of each type, for a total of twelve trials. Trial types were presented sequentially in an order that was counterbalanced across participants. The assignment of individual objects to trial types was counterbalanced. On familiar trials, the familiar label for the target object was used (e.g., "cow"). On novel trials, a novel label was used (e.g., "dawnoo").

The critical manipulation was of referential ambiguity; on one-familiar and two-familiar trials, there was no referential ambiguity, as children were expected to be certain about the objects and their labels. Similarly, on one-novel trials, children were expected to be certain about the label referent as there was only one option. However, on two-novel trials, the referent was ambiguous, as the novel label could apply to either novel object.

Throughout the task, the experimenter never gazed at the object they were labeling, or responded to children's verbal or non-verbal bids for help by indicating the correct object. Thus, children were expected to remain uncertain about the referent throughout the trial when two novel objects were present.

[^91]Procedure Throughout the study, the child sat at one end of a large circular table, and the experimenter stood at the opposite end. Each trial of the task proceeded as follows: the experimenter placed one or two objects on the sides of the table, out of reach of the child so that the child could not interact with the toys during the labeling event. For oneobject trials, the location of the object (left or right) alternated between trials.

After placing the objects, the experimenter said "Hey look, there's a (target) here." The experimenter gazed at the center of the table rather than the object they labeled (see rationale in Stimuli and Design). The experimenter waited approximately two seconds (based on a visual metronome placed within view) before saying, "Can you put the (target) in the bucket?" They then pushed the object(s) forward within reach of the child, and placed a plastic bucket in the center of the table, also within reach of the child. Prior to the twelve experimental trials, there were two training trials: a one-familiar trial and a two-familiar trial, to acquaint the child with the procedure. A camera placed to the side of the experimenter captured the participant's face, so that looking behavior could be coded from video.

Coding procedure Videos were coded using DataVyu software (http://datavyu.org). For each participant, we coded the number of times they referenced the experimenter across the trial. Because we were interested in the circumstances that elicit social referencing in children, we coded the number of looks that occurred during four phases of the trial: a label phase, which began at the utterance of the label and ended when the experimenter began to slide the objects, a slide phase, in which the experimenter slid the object(s) into the child's reach, a planning phase, which began at the end of the slide and ended when the child touched an object, and a response phase, which began when the child touched an object and ended when the child released the object into the bucket. A second coder independently scored the number of looks for one third of the trials for each participant to establish reliability.

## Results and Discussion

Results of Experiment 1 are presented in Figure 2 Inter-rater reliability for the number of looks in each phase was high, intraclass correlation $r=.97, p<.001$. To test our prediction that referential ambiguity (i.e., having two novel objects) would produce more social referencing, we fit mixed-effects linear regression models separately for each phase with the following structure: number of looks ~ number of objects * familiarity * age in months + (number of objects + familiarity | Subject ID). A single model with phase as a factor did not converge.

We did not find any main or interaction effects of number of objects, familiarity, or age on number of looks during the label or slide phases. Thus, mere novelty or the presence of multiple objects was not enough to increase social referencing. However, we found an interaction effect of number of
objects and familiarity during the planning ( $\beta=0.21, p<$ .001 ) and response phases ( $\beta=0.6, p<.001$ ), such that 2 novel trials were associated with more looking. There was no interaction with age in either phase ${ }^{2}$ In summary, children looked to the experimenter more often when planning and executing a response under uncertainty. These results suggest that children were aware that they did not have sufficient knowledge to answer independently, and referenced the speaker to resolve this uncertainty.

We did not find the expected effect of referential ambiguity in the label phase. It is possible that children failed to predict that they would need more information until later in the trial, when they were actually faced with making a decision. Another possibility is that children's looking was at ceiling during the labeling phase, perhaps because children look at someone who is speaking regardless of the need for referential disambiguation. A third possibility is that this is an artifact of our design, in which the experimenter gazed at the center of the table rather than the referent of the label. Children may have realized that the experimenter's gaze direction during labeling was not informative. Similarly, children may have found it strange to interact with an experimenter who did not gaze at the object they were labeling, which may have produced unnatural patterns of social referencing. Experiment 2 tests these possibilities and examines whether children's social referencing is sensitive to graded uncertainty.

## Experiment 2

Experiment 2 was designed to replicate Experiment 1 and investigate whether children's social referencing is sensitive to uncertainty based on graded evidence about a label's referent. Since we did not observe any difference between one-familiar and one-novel trials, we eliminated single-object trials, leaving the 2 -familiar and 2-novel trials. In addition, we added 1-novel-1-familiar trials. For these trials, we expected that children would be able to infer the referent by excluding the familiar object as a possibility. For example, when a toy lion and a novel item were present, they could exclude that the speaker was referring to the lion as a "blicket" (Markman \& Wachtel, 1988). We predicted that children might be less certain about their choice on these trials compared to when the label and referent were familiar to them (2-familiar trials), but more confident than when there are no cues to reference (2-novel trials).

In addition, we manipulated between participants whether or not the experimenter's gaze during labeling was informative (they gazed at either the referent of their label or the center of the table), allowing us to determine whether children selectively reference gaze during labeling when gaze is expected to be informative. The manipulation of informativity of gaze during labeling also meant that participants in the referential gaze condition had an additional referential cue, which might decrease uncertainty for the remainder of the trial. In Experiment 1, we did not observe an effect of age, so

[^92]

Figure 2: Results of Experiment 1. Number of looks to the experimenter across phases and conditions. Error bars are 95 percent confidence intervals.
we restricted the current sample to 3 - and 4-year-olds.

## Methods

Participants We recruited a planned sample of 80 children ages 3-4 years from the Children's Discovery Museum in San Jose, California ${ }^{3}$ The sample included 40 3-year-olds (mean age 42.89 months) and 404 -year-olds (mean age 53.47 months). An additional 20 children participated but were removed from analyses because they heard English less than $75 \%$ of the time at home $(n=9)$, because they were unable to complete at least half of the trials in the task $(n=7)$, or due to experimenter or technical errors $(n=4)$.
Stimuli and Design The stimuli and design were similar to Experiment 1, except that we eliminated 1-object trials. Instead, we included three trial types: 2-familiar, 2-novel, and 1-novel-1-familiar. There were four of each trial type, totaling twelve trials. In addition, we manipulated the experimenter's gaze behavior between participants. For half of the participants, the experimenter looked at the center of the table while labeling objects; for the remaining half, they looked directly at the objects they labeled.

[^93]Procedure The procedure was identical to Experiment 1, except that there were three practice trials (two familiar trials and one novel trial). We included two familiar trials during the practice so that children would remain motivated to complete the task.

## Results and Discussion

Results of Experiment 2 are presented in Figure 3 Inter-rater reliability for the number of looks in each phase was again high, intraclass correlation $r=.97, p<.001$. To quantify the main and interactive effects of familiarity, gaze informativity, phase, and age on social referencing, we fit a mixed-effects linear regression model with the following structure: number of looks ~ familiarity * age in months * gaze * phase + (familiarity | Subject ID). In contrast to Experiment 1, a model with phase as a predictor converged.

First, do children reference a speaker more often when the objects and label are novel? Phase interacted with familiarity such that the response phase of novel trials was associated with significantly more looks ( $\beta=0.51, p<.001$ ). This result is consistent with our finding from the analysis of the response phase in Experiment 1. However, in contrast to Experiment 1, we did not observe that looking was significantly greater for novel trials in the planning phase.

We were also interested in whether mutual exclusivity tri-


Figure 3: Results of Experiment 2. Number of looks to the experimenter across phases and trial types. Error bars are 95 percent confidence intervals.
als would elicit an intermediate amount of uncertainty. We observed a three-way interaction of familiarity, gaze, and phase, such that the response phase of mutual exclusivity trials in the no-referential-gaze condition was associated with significantly more looks ( $\beta=0.39, p<.01$ ). Thus, mutual exclusivity trials were associated with greater looking only when the experimenter did not provide informative gaze. This finding is intriguing given that children should be able to solve mutual exclusivity trials without gaze information. Instead, they appear to remain relatively uncertain while making a decision if excluding the familiar object is their only cue to reference, but this uncertainty is resolved if the speaker's gaze is informative. On the other hand, informative gaze during labeling did not lessen social referencing for novel trials, suggesting that gaze information alone was not sufficient to reduce uncertainty. Instead, both gaze information and mutual exclusivity provided evidence about a label-object pairing, and children required both types of evidence to feel certain about their response.

Finally, we observed a four-way interaction such that the response phase of novel trials in the gaze condition was associated with more looking with increasing age ( $\beta=0.06, p$ $<.01$ ), suggesting that children may become more selective in their social referencing as they get older. It may be that children improve in their ability to monitor the need for disambiguating information, or they may become more likely to recognize that social information can be a source of disambiguation.

We did not observe social referencing during the label phase, even when referential gaze was available. This result
rules out the possibility that children were less selective during labeling because they learned that gaze direction was not informative.

## General Discussion

During the preschool years, children are increasingly able to actively gather information through help-seeking and exploration (Chouinard et al., 2007; Schulz \& Bonawitz, 2007). Do children monitor their own uncertainty to guide these behaviors, or are they indiscriminate with regard to underlying knowledge states? Here, we examined whether young children's social referencing during a word-learning task was driven by uncertainty about a label's referent.

We found that referential ambiguity strongly predicted children's social referencing. Specifically, we observed this selectivity when children were forced to decide which object the speaker was referring to. We speculate that children referenced the speaker during the decision process because they expected evaluative feedback about their choice, either implicitly through the adult's facial expressions, or through an explicit response. This idea is consistent with other recent research that has found that preschoolers seek help selectively when a problem is difficult or they are less skilled (Vredenburgh \& Kushnir, 2015).

Most intriguingly, we found that children's looking was driven by graded referential evidence. In the case of mutual exclusivity trials, children could solve the problem of reference by excluding the familiar item (Markman \& Wachtel, 1988). Thus, unlike novel trials, they likely had some signal about the correct object-label mapping. If children sim-
ply monitored the presence or absence of such signals, they would have consistently treated mutual exclusivity trials as familiar trials. Instead, their social referencing depended on a combination of cues from mutual exclusivity and gaze informativity, suggesting that they are sensitive to graded evidence and seek disambiguating information only when uncertainty is relatively high. Children's greater social referencing on trials with only one cue to reference (i.e., mutual exclusivity trials with no referential gaze and novel trials with referential gaze) additionally suggests that children may remain uncertain about a new label-object mapping if they have not received confirmation of its accuracy, for example, through explicit feedback or gaze direction.

On the other hand, we found no evidence for selective social referencing as the object was being labeled. One possibility is that young children do not recognize the need for disambiguating information until they need to make a decision. Another possibility is that preschool-aged children spontaneously look at a speaker regardless of ambiguity, and additional looking was not needed or possible. Notably, Vaish et al. observed selective referencing during labeling among infants. Since infants in that study were holding one of the objects during labeling, referencing the speaker would have required them to disengage from that object, and may therefore have been more costly, promoting selectivity. Future research with preschoolers that includes a greater reward trade off between attentional options would help to distinguish among these possibilities. Overall, these results provide evidence that preschool-aged children monitor graded uncertainty in their mental representations and act on that uncertainty through spontaneous information-seeking.

## Acknowledgements

We thank Veronica Cristiano for assisting with data collection. This work supported by a generous gift from Kinedu, Inc.

## References

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. Child Development, 62(5), 875890.

Call, J., \& Carpenter, M. (2001). Do apes and children know what they have seen? Animal Cognition, 3(4), 207-220.
Carpenter, M., Nagel, K., Tomasello, M., Butterworth, G., \& Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the Society for Research in Child Development, 63(4), i-iii-v-vi-1-174.
Chouinard, M. M., Harris, P. L., \& Maratsos, M. P. (2007). Children's questions: A mechanism for cognitive development. Monographs of the Society for Research in Child Development, 72, 1-129.
Coughlin, C., Hembacher, E., Lyons, K. E., \& Ghetti, S. (2015). Introspection on uncertainty and judicious helpseeking during the preschool years. Developmental Sci-
ence, 18(6), 957-971.
Csibra, G., \& Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata \& M. H. Johnson (Eds.), Processes of change in brain and cognitive development (pp. 249-274). Oxford: Oxford University Press: Oxford University Press.
Dunlosky, J., \& Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students learning and retention. Learning and Instruction, 22(4), 271-280.
Goupil, L., Romand-Monnier, M., \& Kouider, S. (2016). Infants ask for help when they know they don't know. Proceedings of the National Academy of Sciences, 113(13), 3492-3496.
Hembacher, E., \& Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. Psychological Science, 25(9), 1-9.
Ho, M. K., MacGlashan, J., Littman, M. L., \& Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. Cognition, 1-16.
Lipowski, S. L., Merriman, W. E., \& Dunlosky, J. (2013). Preschoolers can make highly accurate judgments of learning. Developmental Psychology, 49(8), 1505-1516.
Lyons, K. E., \& Zelazo, P. D. (2011). Monitoring, metacognition, and executive function: Elucidating the role of selfreflection in the development of self-regulation. Advances in Child Development and Behavior, 40, 379-412.
Markman, E. M., \& Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20, 121-157.
Robinson, M. D., Johnson, J. T., \& Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. Journal of Applied Psychology, 82, 416-425.
Schulz, L. E., \& Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. Developmental Psychology, 43(4), 1045-1050.
Sodian, B., Thoermer, C., Kristen, S., \& Perst, H. (2012). Metacognition in infants and young children. In M. J. Beran, J. Brandl, J. Perner, \& J. Proust (Eds.), Foundations of metacognition (pp. 119-133).
Vaish, A., Demir, . E., \& Baldwin, D. A. (2011). Thirteenand 18-month-old infants recognize when they need referential information. Social Development, 20(3), 431-449.
Vredenburgh, C., \& Kushnir, T. (2015). Young children's help-seeking as active information gathering. Cognitive Science, 40(3), 697-722.
Walden, T. A., \& Ogan, T. A. (1988). The development of social referencing. Child Development, 59(5), 1230-1240.
Yurovsky, D., \& Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. Cognition, 145(C), 53-62.

# Investigating the Explore/Exploit Trade-off in Adult Causal Inferences 

Erik Herbst (e.herbst@mail.utoronto.edu)<br>Department of Psychology, University of Toronto, Canada<br>Chris Lucas (c.lucas@ed.ac.uk)<br>School of Informatics, University of Edinburgh, United Kingdom

Daphna Buchsbaum (buchsbaum@psych.utoronto.ca)<br>Department of Psychology, University of Toronto, Canada


#### Abstract

We explore how adults learn counterintuitive causal relationships, and whether they discover hypotheses by revising their beliefs incrementally. We examined how adults learned a novel and unusual causal rule when presented with data that initially appeared to conform to a simpler, more salient rule. Adults watched a video of several blocks placed sequentially on a blicket detector, and were then asked to determine the underlying causal structure. In the near condition the true rule was complex, but could be found by making incremental improvements to the simple and salient initial hypothesis. The distant condition was governed by a simpler rule, but to adopt that rule participants had to set aside their initial beliefs, rather than revising them incrementally. Adults performed better in the near condition, despite this rule being more complex, providing some of the first evidence for an explore-exploit trade-off in inference, analogous to the trade-off in active learning.


Keywords: causality, Bayesian inference, hypothesis search, process model

## Background

Any time we make plans, predict the future, or attempt to understand why events occurred in the past, we are relying on causal knowledge. In acquiring this knowledge, we must draw conclusions from sparse, noisy, and ambiguous evidence. We gain the ability to make sense of this limited information at an early age, with causal thinking showing signs of emergence even in infancy (Sobel \& Kirkham, 2006; 2007; Walker \& Gopnik, 2014). By adulthood, our frameworks for interpreting causal phenomena become much more complex and able to accommodate diverse areas of knowledge (Kemp, Goodman, \& Tenenbaum, 2007).

Despite its usefulness, sometimes our ability to generalize from past causal inferences can lead us astray, as in the case where we encounter a new causal relationship that is rare or strange by the standards of our past experience. For instance, we might expect that either of two switches will turn on a lamp, when in fact the lamp turns on when the switches are in matched positions. While our causal learning process is generally accurate and adaptive (e.g., Griffiths \& Tenenbaum, 2005), in the current paper we claim - in the spirit of previous "rational process" models (e.g. Sanborn, Griffiths, \& Navarro, 2010) - that human causal beliefs are updated in a limited or local fashion that is efficient but subject to systematic failures under certain conditions. This is especially true when the initial hypothesis is at a local optimum - the best hypothesis within reach, but not the best overall - and when the true causal structure is distant from our initial hypothesis
in some hypothesis space. Suppose you break out in a rash every time you buy your favourite candy bar from a vending machine. After searching for the proper cause, you would probably conclude that you are allergic to the candy as soon as it comes to mind. You may be unlikely to consider that you are actually reacting to the coins used to purchase the candy bar, even if this is indeed the case. In this case, discovering the real cause requires abandoning your working hypothesis, rather than just incrementally refining it.

## Bayesian Models of Causal Inference

Several researchers have attempted to explain learning of novel causal relationships using hierarchical Bayesian models of inference (e.g. Griffiths, Sobel, Tenenbaum, \& Gopnik, 2011; Griffiths, Kemp, \& Tenenbaum, 2008). Recent evidence demonstrates that adults and children can successfully modify their causal beliefs in light of new and surprising evidence in a manner that suggests Bayesian inference strategies (e.g., Griffiths, Sobel, Tenenbaum, \& Gopnik, 2011; Lucas, Bridgers, Griffiths, \& Gopnik, 2014). Through this process, learners also create and update higher-level models of how causal relationships operate in general. Regardless of whether human cognition functions exactly this way, hierarchical Bayesian models have accurately predicted human causal learning (Kemp, Goodman, \& Tenenbaum, 2007; Lu, Yuille, Lijeholm, Cheng, \& Holyoak, 2006; Lucas \& Griffiths, 2010; Ullman, Goodman, \& Tenenbaum, 2012).

Although Bayesian models accurately capture many aspects of human causal reasoning, they may not fully account for adults' relative difficulties in learning more unusual types of causal relationships. Specifically, Lucas and colleagues (2014) found that young children were more likely than adults to discover an unusual conjunctive causal relationship. Children and adults were tasked with inferring a causal principle after viewing a machine that activated when certain blocks or block combinations were placed on top of it. Even after viewing evidence that blocks only activated the machine in specific pairs (and not individually), adults had more difficulty than children with generalizing this principle to new blocks. One possibility for this finding is that adults are more biased by prior experiences-as they have observed that conjunctive relationships are relatively rarewhich leads them to demand strong evidence before they infer a conjunctive relationship is present. Indeed, if cogni-
tion operates via Bayesian principles, there are conceivably instances in which rigid commitment to a prior may preclude learners from uncovering the true nature of a causal relationship. However, this may not apply in novel causal situations with which adults have limited experience. Moreover, adults are cognitively different than children beyond simply having more experience, so differences in causal reasoning may in fact be the by-product of some developmental change.

As an alternative to simply having different priors, adults' relative difficulty with conjunctive causal relationships may be explained in terms of the process by which they explore and weigh new hypotheses in light of their current beliefs. It is typically impossible to evaluate all potential hypotheses (of which there may be an infinite number). Bayesian inference is often intractable in practice for complex problems, so human inferences must sometimes depart from the Bayesian ideal. Nonetheless, there is evidence that people may be resource rational observers, making approximately Bayesian inferences in ways that make efficient use of limited time and memory (Bonawitz, Denison, Gopnik, \& Griffiths, 2014; Sanborn, Griffiths, \& Navarro, 2010). As for possible processes underlying these approximations, some empirical phenomena, such as order effects, offer clues. If learners make inferences from a complete set of data, as traditional Bayesian models assume, then they should not be influenced by the order in which stimuli are presented. Nevertheless, humans are sensitive to presentation order (Danks \& Schwartz, 2006; Sanborn, Griffiths, \& Navarro, 2010). One explanation for these order effects is that people arrive at solutions by considering a small number of hypotheses at any single moment in time, and updating or replacing them sequentially with more data - sometimes losing information and leading to small but systematic errors. More recently, Bayesian process models have been proposed to explain these patterns of errors by drawing analogies to Monte Carlo sampling methods that permit tractable and efficient inference in applied statistics and machine learning (Abbott, Hamrick, \& Griffiths, 2013; Shi, Griffiths, Feldman, \& Sanborn, 2010).

Inference techniques are often modelled using Monte Carlo methods that update sequentially and incrementally. These methods allow hypotheses to be revised by sampling from the posterior, without computing the posterior distribution in its entirety. Markov chain Monte Carlo sampling is a popular and efficient subclass of Monte Carlo methods, and it is marked by a degree of stickiness or inertia, in which people hew more closely to their initial hypotheses than a truly optimal Bayesian learner would. This family of models predicts that individuals will tend toward inferences that are similar to their prior beliefs. For example, one study showed that when people made inferences about a causal system, they tended toward solutions that required the fewest single edits to their initial hypothesis, where a single edit is an addition, subtraction, or reversal of a causal link (Bramley, Dayan, Griffiths, \& Lagnado, 2017). Therefore, causal process models can account for multiple limitations on causal learning, and have re-
cently been shown to explain phenomena such as classical anchoring (Lieder, Griffiths, Huys, \& Goodman, 2017). Learners can be constrained not only by priors, but also the similarity of candidate hypotheses to their current beliefs, perhaps precluding them from finding too-distant hypotheses.

## The Explore-Exploit Trade-off in Inference

These findings could reflect a cognitive tradeoff in development that affects how learners search through hypotheses. When presented with a wide range of possibilities, individuals must often decide whether to employ a general, shallow search or a narrow, deep one. This decision is analogous to the explore-exploit tradeoff, whereby decision-makers must allocate cognitive resources to either exploit previous knowledge or explore alternatives (Sutton \& Barto, 1998). Adults may be more inclined to exploit, by searching nearby solutions extensively—and less likely to explore hypotheses that require unusual, low-probability edits to the current hypothesis. With limitations on the number of hypotheses a learner can consider, exploitation-biased adult learners could plausibly benefit from focusing cognitive resources on hypotheses that are refinements of an initial proposal that is plausible and informed by long experience. This will increase efficiency of finding adequate solutions but potentially limit access to distant alternatives. Conversely, exploration-focused learners (young children, perhaps) may spread out their search over a more diverse range of possibilities. Although this approach sacrifices the ability to efficiently refine already-reasonable hypotheses, it may grant access to unusual solutions that would be unreachable with a more conservative search.

Thus, the inferential explore-exploit trade-off may have interesting implications for the process of selecting between competing hypotheses. This selection process has been modelled using Bayesian algorithms for both children and adults (Bonawitz, Denison, Gopnik, \& Griffiths, 2014; Denison, Bonawitz, Gopnik, \& Griffiths, 2013; Lieder, Griffiths, \& Goodman, 2012; Sanborn, Griffiths, \& Navarro, 2010), but relatively little previous work has examined adults' potential tendencies toward exploitation. As one possible example of how hypothesis search may reflect an exploitation bias, Gopnik and colleagues have likened human belief updating to simulated annealing; just as the heating and gradual cooling of metal can increase its malleability, so can a gradual "cooling" of an inference method corresponding to an increasingly conservative search policy lead to better inferences (Gopnik, Griffiths, \& Lucas, 2015; Lucas, Bridgers, Griffiths, \& Gopnik, 2014). For instance, while young children may use hightemperature searches, considering a wide range of hypotheses with relatively equal probability, adults' searches are "cooler" and more narrow in scope. Although commitment to priors may still matter, simulated annealing allows us to examine which types of hypotheses are considered. High-temperature searches are more likely to discard adequate hypotheses, but may allow individuals to escape local optima and discover unlikely solutions that are potentially better. In contrast, lowtemperature searches can quickly converge to good solutions
if fewer low-probability edits are required to get there, but may otherwise get trapped in local optima. With this in mind, adults may have more difficulty discovering unusual causal relationships because their search is too focused and too close to their initial guesses to accommodate distant ideas.

The purpose of our current studies is to test the hypothesis that belief updating in adults is exploitation-biased. To accomplish this, we designed a task encouraging participants to generate a particular initial hypothesis about a novel causal relationship. Evidence that contradicted this hypothesis was then presented, causing participants to modify their beliefs. The true causal structure took one of two forms corresponding to two experimental conditions. In the near condition, the correct causal structure was closer to the initial hypothesis but designed to be relatively complex. In the distant condition, the correct causal structure was simpler but possibly harder to reach when making incremental changes from the initial hypothesis, which is a local optimum. Thus, we hoped to determine the breadth of hypotheses that participants were willing to entertain. If adults' search process is more exploitationbiased, we should expect the near-hypothesis solution would be more easily found than the distant one, even if both rules are a priori equally unlikely. However, if adults' failure to infer unlikely causal relationships is simply due to the low prior probability that they place on these relationships, then they should be equally unlikely to consider either solution.

## Experiment 1: Investigating the Explore-Exploit Tradeoff in Inference

Participants Participants were 90 adult US residents, recruited through Amazon Mechanical Turk and paid a base rate of $\$ 1$ for their time. An additional $\$ 1$ bonus was given to the top $10 \%$ performers as an additional incentive. Participants were divided randomly among near $(\mathrm{n}=45)$ and distant ( $\mathrm{n}=45$ ) conditions. Six participants from the near condition and seven from the distant condition were excluded due to failure to correctly answer attention manipulation tasks.

Materials and Procedure The methods used in this study are similar to those used in previous blicket tasks (e.g. Gopnik \& Sobel, 2000), except that animated video stimuli were presented online using Qualtrics survey software (similar to Buchsbaum et al., 2012). Participants were asked to examine several blocks and determine which blocks are blickets. They were informed that blickets are blocks that activate the blicket detector, and were shown a video of an animated blicket detector activating and not activating. Participants then watched a five-minute animation depicting 20 blocks being consecutively placed onto the blicket detector. If the block was a blicket, the detector lit up and a sound played. The blocks were sorted into blicket/non-blicket categories and left on screen for participants to study.

Whether a block was a blicket depended on specific aspects of the block pattern. Each block had a coloured background (red or blue) and several small red or blue triangles in a fixed pattern (see Figures 1 and 2). The block pattern was such


Figure 1: Examples of blickets in the near condition (left) and the distant condition (right).
that the background colour was the most obvious and visually striking feature. For the first 15 blocks (the initial ruleconsistent blocks), the background colour appeared to determine whether the blocks activated the machine-i.e. blocks with one background colour consistently activated the machine, while the others did not. Inspired by an experimental manipulation in Williams and Lombrozo (2010; 2013), this was designed to lead participants to an initial causal hypothesis based on the objects' most salient feature. The final five blocks (the initial rule-violating blocks), however, violated this initial hypothesis; the blocks that did and did not activate the machine had the opposite background colour as before. Thus participants needed to modify their initial hypothesis to capture the optimal solution.

The true rule separating blickets from non-blickets varied based on condition. This true rule determined whether a block was a blicket $100 \%$ of the time. In the near condition, the background colour was related to whether a block was a blicket, whereas in the distant condition the background colour was unrelated. Each block had five binary features (Figure 1), which could vary by colour on each block (background, corners, centre-left triangle, centre-right triangle, and border), giving a total of 32 different colour combinations. In the near condition, blocks were blickets based on a combination of the background colour and the colour of two secondary features. In the distant condition, only the colour of these two secondary features determined whether a block was a blicket, while the background colour was irrelevant.

Thus, the five features could be labeled as follows: one primary feature (A), two relevant secondary features (B and C), and two irrelevant secondary features (D and E). In the distant condition, the optimal rule for determining whether a block is a blicket-that is, the simplest rule that perfectly explains the data-can be written as $R=(B==C)$, whereas the optimal rule in the near condition can be written as $R=$ $(A \cap \neg B) \cup(\neg A \cap \neg C)$. These rules were designed to seem arbitrary to naïve participants and minimize the role of the participants' prior knowledge. In the near condition, there is a consistently-improving path of single edits to transition from the initial hypothesis, $R=A$, to the correct rule, where a single edit consists of adding or subtracting a variable or chang-
ing an operator (e.g. changing $R=A$ to $R=A \cap \neg B$; Goodman \& Tenenbaum, 2008 use a similar approach for searching a hypothesis space). In the distant condition, the single-edit path to the correct rule requires edits that initially worsen the hypothesis (e.g. removing $A$ as a relevant variable). Participants must therefore ignore the ineffectiveness of these local edits and keep exploring to find the correct solution. Thus, if adults use a Bayesian single-edit search process with an exploit bias, participants should be less likely to abandon $R=A$, and thus should perform more poorly in the distant condition, where $R=A$ is the local optimum.

The lists of blocks seen by participants in the near and distant conditions were generated randomly with the following constraints: a) there were ample block feature combinations that participants did not see, so that they could be tested on these blocks later, and b) the rules and edit paths conformed to the specifications in the previous paragraph. Thus, the final sets of blocks were as follows: near condition participants saw 11 blickets ( 3 initial rule-violating) and 9 non-blickets ( 2 initial rule-violating), whereas distant condition participants saw 10 blickets ( 2 initial rule-violating) and 10 non-blickets (3 initial rule-violating). The differences in block numbers were necessary due to the constraints of the conditions.

Following the presentation of all of the blickets, participants saw a blicket rating task, in which they were asked to judge whether a randomized series of eight blocks were blickets. For each block, participants rated how certain they were that it was, or was not, a blicket, on a seven-point Likert scale ranging from "definitely a blicket" to "definitely not a blicket". Blocks were balanced by background colour, blicket/non-blicket status, and whether they had already been presented in the observation stage. Participants received a score between -3 and 3 for each block based on accuracy and certainty, and the sum of these scores determined their final score for this task. Next, participants completed a forcedchoice task, where they chose which of two blocks was more likely to activate the blicket detector, for a series of four pairs. Blocks were selected randomly such that there were an equal number of initial rule-consistent and initial ruleviolating blocks, and blocks in each pair differed from each other in background colour and whether they were a blicket. Participants received a point for each correct block judgment.

Afterwards, the participants were asked to describe the causal rule they had inferred. They were then told to imagine that a new rule was suggested by a friend, and asked if they preferred this rule over their own. This rule always represented the correct causal structure. The purpose of this question was to ensure that any differences between the two conditions were not due to participants finding the near rule inherently more plausible or likely than the distant one. The participants' rule preference was measured using a seven-point scale. Finally, each participant received questions to test their task comprehension and an instructional manipulation task to control for inattention, similar to the one used by Oppenheimer, Meyvis, and Davidenko (2009).

Results and Discussion If adults' hypothesis search strategy is exploitation-biased, participants in the near condition will perform better on both tasks than those in the distant condition. The results supported our predictions. For the forced-choice task, a $2 \times 2$ ANOVA was run with condition (distant/near) and rule consistency (initial ruleconsistent/violating) as factors (see Table 1 for a score summary). Near condition participants outscored those in the distant condition, $F(1,84)=6.46, p=.01, M S E=0.26$. Participants also scored higher for initial rule-consistent blocks, than for rule-violating blocks, $F(1,84)=226, p<.001, M S E$ $=0.34$. There was no significant interaction effect, $F(1,84)=$ $0.154, p>.69, M S E=0.34$.

For the blicket rating task, a $2 \times 2$ mixed ANOVA (condition x rule consistency) was run (see Table 2 for a score summary). The analysis found that participants were much more likely to confidently identify initial rule-consistent blocks than initial rule-violating blocks $F(1,84)=131, p<.001, M S E=15.32$, suggesting that the salience manipulation was effective and participants were influenced by the background colour. Supporting our forced-choice results, there was a marginally significant effect of condition, $F(1,84)=3.77, p=.06, M S E=$ 11.87, with a mean score of 7.51 for the near condition and 4.63 for the distant condition (scores ranged from - 24 to 24 ).

Intriguingly, and unlike in the forced-choice task, there was also a significant interaction effect, $F(1,84)=3.34, p=.04$, $M S E=15.32$. This is a result of participants in the near condition performing better than those in the distant condition on initial rule-consistent blocks, but equally poorly on initial rule-violating blocks. To assess whether this interaction was due to differences in confidence for some blocks, an additional $2 \times 2$ mixed ANOVA (condition $x$ rule consistency) was run to investigate participants' certainty ratings when evaluating blocks. The analysis showed no main effect of condition, $F(1,84)=2.30, p>.13, M S E=0.69$. Mean confidence ratings were relatively near ceiling in both conditions (greater than 2 out of 3 ), which may partially explain the lack of a main effect. However, participants were more certain of their answers when rating initial rule-consistent blocks than when rating rule-violating blocks, $F(1,84)=22.0, p<.001, M S E$ $=0.32$. There was also a highly significant interaction effect between condition and rule-consistency, $F(1,84)=13.1, p$ $<.001, M S E=0.32$, driven by participants in the near condition having more certainty for initial rule-consistent blocks than for rule-inconsistent blocks, suggesting that while participants in the near condition were better able to correctly categorize both initial rule-violating and initial rule-consistent blocks, they were most confident about the latter.

Additional one-sample t-tests examined whether participants scored better than would be expected by chance. For the forced-choice task, participants correctly classified blocks as blickets and non-blickets significantly better than chance in the near condition, $t(42)=5.82, p<.001$, but not in the distant condition, $t(42)=1.31, p=0.20$. In the blicket rating task, however, participants classified blocks better than

Table 1: Mean scores and SE for forced-choice task. Total scores range from 0 to 4 , and scores for initial rule-consistent and initial rule-violating blocks range from 0 to 2 .

| Condition | Near | Distant |
| :--- | :---: | :---: |
| Total score | $2.53( \pm 0.10)$ | $2.24( \pm 0.12)$ |
| Rule-consistent | $1.90( \pm 0.08)$ | $1.82( \pm 0.07)$ |
| Rule-violating | $0.77( \pm 0.13)$ | $0.42( \pm 0.07)$ |

Table 2: Mean scores and SE for blicket rating task. Total scores range from -24 to 24 , and scores in each sub-category range from -12 to 12 .

| Condition | Near | Distant |
| :--- | :---: | :---: |
| Total score | $8.00( \pm 1.04)$ | $4.87( \pm 1.26)$ |
| Rule-consistent | $9.59( \pm 0.51)$ | $6.39( \pm 0.72)$ |
| Rule-violating | $-1.59( \pm 1.01)$ | $-1.53( \pm 1.06)$ |

chance in both the near condition, $t(42)=7.69, p<.001$, and the distant condition $t(42)=4.13, p<.001$. The at-chance performance of distant condition participants in the forcedchoice task may simply reflect the low number of trials compared to the blicket rating task.

Finally, we looked at participants' preference for the correct rule over their own. Participants in the distant condition significantly preferred the correct friend's rule over their own rule, $\mathrm{t}(42)=4.78, \mathrm{p}<.001$, while participants in the near condition did not, $\mathrm{t}(42)=1.55, \mathrm{p}=.13$. Participants in the distant condition also preferred the friend's rule significantly more than those in the near condition, $t(75)=2.09, p=.04$. This supports our hypothesis that participants in the distant condition had not previously considered the distant rule, rather than that they considered it, but dismissed it as unlikely.

## Experiment 2: A priori rule preference

Although the main study compared the extent to which participants preferred the correct rule over their own, it did not examine the rules in both conditions side-by-side. This study investigated adults' a priori preference for either the near or the distant rule without differentiating data. This was to confirm that differences in causal learning and rule preference between conditions in Experiment 1 were not due to an intuitive preference for the near rule before seeing any data.

Participants Participants were 51 adult US residents, recruited through Amazon Mechanical Turk (MTurk) and paid a base rate of $\$ 0.50$ for their time.
Materials and Procedure As in the previous study, participants were told that blickets were blocks that activated the blicket detector, and saw an animated blicket detector activating and not activating. Unlike the previous study, however, participants only saw one block placed on the machine, causing it to activate. They were then told the two possible
rules, and that both rules accurately described this block, but that only one rule was the correct rule for identifying blocks that activate the machine. Participants were asked to choose which rule they thought was more likely to be correct. These rules were identical to the near rule and the distant rule from the previous study, and the blicket that participants saw was chosen from a set of blocks that conformed to both rules. Finally, after selecting a rule, participants explained why they chose that rule and rated their confidence in their decision, ranging from 1 (just guessing) to 7 (completely certain). This confidence rating was turned into a score ranging from -7 (completely certain the near rule is correct) to 7 (completely certain the distant rule is correct) for statistical analysis.
Results and Discussion Of the 51 participants, 22 preferred the near rule and 29 preferred the distant rule, $\mathrm{p}=.41$, exact binomial test. A one-sample t-test demonstrated that the rule preference scores, $M=0.25, S E=0.50$, did not significantly differ from chance, $t(49)=0.71, p=0.48$. Thus, participants did not prefer one rule over the other, suggesting that it was not an a priori preference for the near rule driving the results of Experiment 1.

## General Discussion

The findings obtained by these studies lend support to the exploitation-biased search hypothesis. We expect that exploitation-biased searches of the hypothesis space will be more likely to discover rules close to the initial hypothesis, and less likely to discover more distant rules, even if they are less complex. As predicted, participants were more accurate at classifying blocks in the near condition than the distant condition. This is especially notable given that participants in Experiment 2 found both rules equally a priori plausible, which supports that the near rule is at least as complex as the distant rule. This in turn makes it less likely that the differences between conditions can be explained by differentlyweighted prior probabilities. Participants performed better in the near condition, where the true rule was arguably more complex, but was comparatively easier to discover from the salient starting point due to the consistently-improving edit path, than in the distant condition, where the true rule was simpler, but where the salient rule was a local optimum. This suggests that adults are searching through their hypothesis space in an exploitation-biased manner.

Nevertheless, participants were better able to identify initial rule-consistent blocks than initial rule-violating ones in both tasks. This suggests that the strength of one's priors may still play a role in conjunction with the exploitation bias. However, this difference in performance suggests intriguing future research avenues-in particular, the finding in the blicket rating task that participants in the near condition scored higher than those in the distant condition on initial rule-consistent but not initial rule-violating blocks. This seems to be driven largely by participants' relative certainty toward initial rule-consistent blocks in the near condition, rather than their accuracy at categorizing the blocks (as mea-
sured by the forced choice task). Future studies might assess how nearness to an initial hypothesis affects the certainty of judgments of causal relationships.

It is still unclear, however, if these difficulties in discovering certain causal relationships are the result of a developmental process. Consequently, we plan to expand this study to directly compare adults with children, to examine whether children possess these same search-related difficulties. If these findings are the result of a developmental shift toward exploitation-based search strategies, then explorationoriented children could perform just as well-if not betterthan adults in tasks such as those in this study. Children should also perform equally well in both experimental conditions, or perhaps even better in the distant condition than in the near one. Particularly, this may be the case if children see the near rule as a priori less likely. When comparing children's and adults' performance, it may also be useful to note differences in time spent on each task, as it might generate additional insights about their hypothesis search process. Although participants in the current studies had unlimited time to complete each task, timing data were not recorded.

In the future, it may be useful to develop a more explicit process model to measure hypothesis distance. Although the near-hypothesis rule is closer to the salient hypothesis, in that adding and subtracting particular predicates improves the hypothesis toward the correct rule, this may not accurately represent how individuals process locality. In other words, we lack a precise model for how people move between rules, and thus exactly how far $R=(B==C)$ is from $R=A$, and how much harder it is to find $R=(A \cap \neg B) \cup(\neg A \cap \neg C)$. In future experiments, this process model will need to be clarified.

Overall, our results demonstrating that adults are able to discover a true causal structure nearer to an initial hypothesis more readily than a distant causal structure of equal or greater complexity provides compelling initial evidence for an explore-exploit trade-off in causal inferences. This may help inform future research on how individuals generate new hypotheses about everyday causal interactions.

## References

Abbott, J., Hamrick, J., \& Griffiths, T. (2013). Approximating Bayesian inference with a sparse distributed memory system. Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 16861691). Berlin.
Bonawitz, E., Denison, S., Gopnik, A., \& Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. Cognitive psychology, 74, 3565.

Bonawitz, E., Denison, S., Griffiths, T.L., \& Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development.
Bramley, N. R., Dayan, P., Griffiths, T. L., \& Lagnado, D. A. (2017). Formalizing Neuraths ship: Approximate algorithms for online causal learning. Psychological review, 124(3), 301.
Buchsbaum, D., Bridgers, S., Whalen, A., Seiver, E., Griffiths, T. L., \& Gopnik, A. (2012). Do I know that you know what you know? Modeling testimony in causal inference. In Proceedings of the 34th annual conference of the cognitive science society.
Danks, D., \& Schwartz, S. (2006). Effects of causal strength on learning from biased sequences. In Proceedings of the 28th annual meeting of the cognitive science society.

Denison, S., Bonawitz, E., Gopnik, A., \& Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. Cognition, 126(2), 285-300.
Gelman, A., Carlin, J. B., Stern, H. S., \& Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). London: Chapman \& Hall/CRC.
Glymour, C. N. (2001). The mind's arrows: Bayes nets and graphical causal models in psychology. MIT press.
Goodman, N. D., Tenenbaum, J. B., Feldman, J., \& Griffiths, T. L. (2008). A Rational Analysis of RuleBased Concept Learning. Cognitive Science, 32(1), 108-154.
Gopnik, A., Griffiths, T. L., \& Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. Current Directions in Psychological Science, 24(2), 87-92.
Gopnik, A., \& Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. Child development, 1205-1222.
Griffiths, T. L., Kemp, C., \& Tenenbaum, J. B. (2008). Bayesian models of cognition. In R.Sun (Ed.), Cambridge handbook of computational psychology (pp. 59100). New York: Cambridge University Press.
Griffiths, T. L., Sobel, D., Tenenbaum, J. B., \& Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. Cognitive Science, 35, 1407-1455
Griffiths, T. L., \& Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive psychology, 51(4), 334-384.
Kemp, C., Goodman, N. D., \& Tenenbaum, J. B. (2007). Learning causal schemata. Cognitive Science Society.
Lieder, F., Griffiths, T. L., \& Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. Advances in Neural Information Processing Systems, 25.
Lieder, F., Griffiths, T., Huys, Q. J., \& Goodman, N. D. (2017). The anchoring bias reflects rational use of cognitive resources. Psychonomic Bulletin \& Review.
Lu, H., Yuille, A., Lijeholm, M., Cheng, P. W., \& Holyoak, K. J. (2006). Modeling causal learning using Bayesian generic priors on generative and preventive powers.
Lucas, C. G., Bridgers, S., Griffiths, T. L., \& Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. Cognition, 131(2), 284-299.
Lucas, C. G., \& Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. Cognitive Science, 34(1), 113-147.
Oppenheimer, D. M., Meyvis, T., \& Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45(4), 867-872.
Sanborn, A. N., Griffiths, T. L., \& Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. Psychological review, 117(4), 1144.
Shi, L., Griffiths, T. L., Feldman, N. H., \& Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. Psychonomic bulletin \& review, 17(4), 443-464.
Sobel, D. M., \& Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. Developmental psychology, 42(6), 1103.
Sobel, D. M., \& Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. Developmental science, 10(3), 298306.

Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction (Vol. 1, No. 1). Cambridge: MIT press.
Ullman, T. D., Goodman, N. D., \& Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. Cognitive Development, 27(4), 455-480.
Walker, C. M., \& Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. Psychological science, 25(1), 161-169.
Williams, J. J., \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. Cognitive Science, 34(5), 776-806.
Williams, J. J., \& Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. Cognitive psychology, 66(1), 55-84.

# Modeling transfer of high-order uncertain information 

Michele Herbstritt (michele.herbstritt@gmail.com) Michael Franke (mchfranke@gmail.com)<br>Department of Linguistics, University of Tübingen<br>Wilhelmstraße 19, 72070 Tübingen, Germany


#### Abstract

Complex uncertainty expressions such as probably likely and certainly possible naturally occur in everyday conversations. However, they received much less attention in the literature than simple ones. We propose a probabilistic model of the use and interpretation of complex uncertainty expressions based on the assumption that their predominant function is to communicate factual information about the world, and that further layers of uncertainty are pragmatically inferred. We collected empirical data on the use and interpretation of these expressions and use it for detailed model criticism.


Keywords: uncertainty; probability; experimental pragmatics; computational modeling

## Introduction

One of the main goals of human linguistic interactions is the exchange of information. However, the information that we want to exchange can be uncertain: we often talk about things that we do not know for sure. As a consequence, it should not surprise us that human languages are equipped with so called "uncertainty expressions" such as epistemic modals (possible, might) and probability expressions (probably, likely).

Simple uncertainty expressions have been extensively investigated in psychology (Beyth-Marom, 1982; Teigen, 1988; Windschitl \& Wells, 1996, 1998) and formal linguistics (Kratzer, 1991; Yalcin, 2010; Egan \& Weatherson, 2011; Lassiter, 2011), where some consensus has recently emerged about the advantage of adopting a formal semantics that uses probability measures (contra the purely qualitative semantics à la Kratzer). Herbstritt and Franke (2016) empirically investigated the production of simple uncertainty expressions (probably, possibly) and propose a pragmatic model of their production. This paper substantially extends the scope of that work: here we investigate complex (or nested) expressions such as probably likely and certainly possible and we model both their production and interpretation in a conversation.

Complex uncertainty expressions have received much less attention in the literature. ${ }^{1}$ Indeed, many foundational issues arise in the attempt to formalize a model of their use and interpretation. Most pressingly are two interrelated concerns: (i) what is the semantic meaning of a complex uncertainty expression? and (ii) what is the communicative goal of a complex uncertainty expression, i.e. what is the pragmatic purpose of communication? In this paper we present a first model that commits itself to what are arguably the most natural answers to (i) and (ii) from the point of view of formal semantics (Swanson, 2006; Moss, 2015) and a rational analysis of communicative practices as efficient transfer of information about the world (Anderson, 1990, 1991).

[^94]This approach enables a straightforward regular and compositional treatment of the meaning of uncertainty expressions: simple and complex uncertainty expressions denote sets of probability distributions over the state space that represents the possible ways in which the world can be. The meanings of simple expressions are always singleton sets. The meanings of complex expressions are derived compositionally in terms of the simple ones and in general they contain more than one distribution (see details below). If we model agents' uncertain beliefs about the world as (sets of) probability distributions over the same state space, then the meaning of a simple or complex uncertainty expression can be seen as a collection of ways to update the agents' beliefs. Figure 1 displays an intuitive representation of this idea.

We incorporate this idea in a probabilistic pragmatic model of language production and interpretation based on the Rational Speech Acts (RSA) model (Frank \& Goodman, 2012). In particular our model can be seen as a conservative generalization of the RSA model proposed by Goodman and Stuhlmüller (2013). The key innovation of our model is to treat uncertain beliefs of agents (and thus the communicative effect of messages) as sets of probability distributions, hence more fine grained than in the usual approach.


Figure 1: Listener's beliefs as complex uncertainty state. Each probability distribution in the listener's beliefs is compatible with the literal meaning of the received message.

The details of the model are spelled out in the next section. In the following section we report on two experiments designed to collect human data about production and interpretation of complex uncertainty expressions. Finally, the predictions of the model are evaluated against experimental data with Bayesian inference and model criticism.

## Pragmatic model

Setup We want to model communication in situations of what we call high-order uncertainty. To illustrate, imagine an urn containing 10 balls of two different colors (e.g., red
and blue). The universe of the discourse, the set of possible states of affairs, can be modeled as the set of natural numbers $S=\{0, \ldots, 10\}$ where each $s \in S$ is a possible quantity of red balls in the urn. The ratio $s / 10$ expresses the objective chance that a randomly drawn ball will be red, and represents first-order uncertainty: even if we know the objective chance, we are uncertain about the color of a randomly drawn ball. The second, high-order level of uncertainty comes into play when we are uncertain about the objective chance too. We model agents that do not have direct access to the objective chance. Instead, one agent (the speaker) can draw a certain number of balls from the urn (referred to as the "access" and denoted with $a$ ) and look at them. The set of possible access values is $A=\{1, \ldots, 10\}$. The number of red balls among the accessed ones is referred to as the "observation" and denoted $o, 0 \leq o \leq a$. We assume that the communication is about the content of the urn: after her observation, the speaker puts all balls back in the urn and makes a prediction about the color of a randomly drawn ball (see Figure 2). This prediction is what the speaker will try to communicate.


Figure 2: Partial observation of the content of the urn.
The probability of observing $o$ red balls when the speaker draws $a$ balls and there are $s$ red balls in the urn is given by the hypergeometric distribution. Assuming that the agent has a prior belief distribution over the state space $S$, we can say that each pair $\langle o, a\rangle$ induces a posterior rational belief distribution over $S$, computed as the Bayes-inverted hypergeometric distribution: ${ }^{2}$

$$
\begin{equation*}
\operatorname{rat} . \operatorname{bel}(s \mid o, a) \propto \operatorname{hypergeom}(o ; a, s, 10) * \operatorname{prior}(s) \tag{1}
\end{equation*}
$$

On the basis of the rational belief resulting from her observation, together with the lexical meaning of the available messages, the speaker chooses the best message to send given her communicative goal.

Messages and semantics The speaker sends messages of the form It is [...] that a randomly drawn ball will be red, choosing from the following 12 expressions to fill the gap:

| likely | possible | unlikely |
| :--- | :--- | :--- |
| certainly likely | certainly possible | certainly unlikely |
| probably likely | probably possible | probably unlikely |
| might be likely | might be possible | might be unlikely |

[^95]Simple messages (likely, possible, unlikely) have a simple threshold semantics:

$$
\begin{aligned}
& \llbracket \operatorname{likely}(p) \rrbracket=\left\{s \in S \mid s / 10>\theta_{\text {likely }}\right\} \\
& \llbracket \operatorname{possible}(p) \rrbracket=\left\{s \in S \mid s / 10>\theta_{\text {possible }}\right\} \\
& \llbracket \operatorname{unlikely}(p) \rrbracket=\left\{s \in S \mid s / 10<1-\theta_{\text {likely }}\right\}
\end{aligned}
$$

The thresholds $\theta_{\text {likely }}$ and $\theta_{\text {possible }}$ are free parameters in the model (more about this below). The variable $p$ can be instantiated with a sentence such as A randomly drawn ball will be red. For example, this semantics states that the meaning of It's possible that a randomly drawn ball will be red is the set of states where the objective probability of the ball being red is bigger than a certain threshold $\theta_{\text {possible }}$.

The semantics of complex messages is stated in a general form as follows:

$$
\llbracket \operatorname{modifier}[\operatorname{simple}](p) \rrbracket=\left\{\langle o, a\rangle \mid \sum_{s \in \llbracket \operatorname{simple}(p) \rrbracket} \operatorname{rat.bel}(s \mid o, a)>\theta_{m}\right\}
$$

where $\theta_{m}$ is the semantic threshold associated with the modifier. ${ }^{3}$ Each state in the meaning of the simple message $\llbracket \operatorname{simple}(p) \rrbracket$ is associated with a certain probability mass according to the rational belief induced in the speaker by each pair $\langle o, a\rangle$; the meaning of the complex message is computed collecting the pairs $\langle o, a\rangle$ where the probability mass of the states in $\llbracket \operatorname{simple}(p) \rrbracket$ is greater than the semantic threshold of the modifier. The semantics of complex messages is rooted in the literal semantics of the simple ones. The difference between the two is that while the meanings of simple messages contain states of affairs, the meanings of complex expressions contains pairs denoting partial observations, i.e. distributions over states. Still, both simple and complex expressions can be linked to sets of probability distributions over world states. Making use of this allows for a uniform grounding of semantic meaning in a model of rational communication.

Beliefs and expected utility On the basis of the literal meaning of each message, we compute their effect on the socalled "literal listener", a theoretical construct modeling the interpretation process of a non-pragmatic agent. Each simple message induces exactly one belief distribution in the literal listener, whereas each complex message induces a set of distributions (one for each pair $\langle o, a\rangle$ in the meaning of the expression). This idea is captured in Equation 2, where the set of distributions lit.bel is defined by cases as a function of messages. ${ }^{4}$

[^96]\[

$$
\begin{align*}
& \text { lit.bel }(m)= \\
& =\left\{\begin{array}{l}
\left\{P \in \Delta(S) \mid \forall s \in S: P(s) \propto \delta_{s \in \llbracket m \rrbracket} * \operatorname{prior}(s)\right\} \quad m \text { simple } \\
\{P \in \Delta(S) \mid \exists\langle o, a\rangle \in \llbracket m \rrbracket: P=\operatorname{rat} . \operatorname{bel}(. \mid o, a)\} m \text { complex }
\end{array}\right. \tag{2}
\end{align*}
$$
\]

We assume that the communicative goal of the speaker is to maximize the information transferred to the listener. Here we formalize this concept as choosing the message which brings the listener's factual beliefs as close as possible to the speaker's, i.e. which minimizes the distance between the probability distributions expressing these beliefs. In general each message is associated with a set of probability distributions over states, according to Equation 2. Idealizing, we assume that the literal listener would uniformly sample from this set of uncertain beliefs upon hearing each message. For this reason the expected utility (EU) of a message $m$ given an observation $\langle o, a\rangle$ is defined as the negative average Hellinger distance between the speaker's belief distribution given an observation and the set of the listener's distributions given a message (Equation 3).

$$
\begin{equation*}
\mathrm{EU}(m ; o, a)=-\operatorname{avg}[H D(\{\operatorname{rat} . \operatorname{bel}(. \mid o, a)\}, \text { lit.bel }(m))] \tag{3}
\end{equation*}
$$

where $H D$ denotes a function computing pairwise Hellinger distances between two sets of discrete distributions. ${ }^{5}$

Production and interpretation Adopting the terminology of rational choice theory, the speaker's behavior is to softmaximize the EU of each message given her observation:

$$
\begin{equation*}
\operatorname{speaker} . \operatorname{prob}(m \mid o, a) \propto \exp (\lambda * \mathrm{EU}(m ; o, a)) \tag{4}
\end{equation*}
$$

EU is multiplied by a rationality parameter $\lambda$ (free in the model) that modulates "how rational" the choice is. ${ }^{6}$ The distribution over messages defined in Equation 4 gives rise to the first half of the set of predictions made by our model, whose fit to the experimental data is discussed below.

A pragmatic listeners reasons about the received message and her model of speaker's behavior in order to infer the most likely interpretation. The pragmatic listener's behavior is modeled as the joint Bayesian inference over the variables of interest:

$$
\begin{equation*}
\text { listener.prob }(s, o, a \mid m) \propto \text { speaker. } \operatorname{prob}(m \mid o, a) * \text { priors } \tag{5}
\end{equation*}
$$

We are interested in the interpretation of uncertainty expressions alongside two axes of their communicative effect. One

[^97]is the objective state of affairs communicated (i.e., the inference of $s$ ). The other is the subjective, high-order, state of uncertainty of the speaker (i.e., the inference of $\langle o, a\rangle$ ). The joint distribution defined in Equation 5 gives rise to the second half of the set of predictions made by our model, whose fit to the experimental data is discussed below.

## Experiments

We conducted two experimental studies, a production task and an interpretation task. The goal of the production task was to collect human data on the use of simple and complex uncertainty expressions under different high-order uncertainty conditions. The goal of the interpretation task was to collect human data on the interpretation of the expressions in terms of inference of $s, o, a$.

Participants 252 self reported English native speakers with USA IP-addresses were recruited via Amazon's Mechanical Turk. 102 participants completed the production task, 150 participants completed the interpretation task.

Material Participants in the production task were exposed to visual stimuli depicting partial observations of the urn. We asked participants to imagine drawing a number of balls (access) and counting the red balls among them (observation). Then they would put the balls back in the urn, and make a prediction about the color of another randomly drawn ball (Figure 2).

The experimental conditions are the different observation/access configurations displayed to the participants. We selected 15 such configuration:

| high | $0 / 2$ | $1 / 4$ | $2 / 4$ | $3 / 4$ | $2 / 2$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| low | $0 / 8$ | $2 / 8$ | $4 / 8$ | $6 / 8$ | $8 / 8$ |
| none | $2 / 10$ | $3 / 10$ | $5 / 10$ | $7 / 10$ | $8 / 10$ |

Each fraction in the table corresponds to a possible partial observation, e.g. $3 / 4$ means accessing 4 balls and observing that 3 of them are red. The fractions are grouped according to their level of high-order uncertainty. Access values smaller than 5 balls are labeled "high" high-order uncertainty, whereas values greater than 5 correspond to "low" high-order uncertainty, and values equal to 10 represent no high-order uncertainty whatsoever.

The set of stimuli for the interpretation task was derived from the 12 expressions assumed in the model.

Procedure Before the experimental phase, participants completed a training phase which contained a cover story introducing an interactive game between two players, a sender and a receiver. Participants in the production task were told that they would play as senders, and that other players would receive their messages and try to guess the content of the urn. Participants in the interpretation task were told that they would play as receivers. The motivation for this setup was to
clarify the purpose of the conversation when producing uncertainty expressions and to prompt participants to reason about the effect of their choices on other agents.

Each participant in the production task completed 12 trials, one for each of 12 conditions randomly picked from the 15 total conditions. In each trial the participant made the partial observation of the urn corresponding to the selected condition and was asked to send a message containing a prediction about the color of a randomly drawn ball. Crucially, this prediction must be expressed by completing a sentence of the form It $[\ldots][\ldots]$ that the next ball will be red, selecting the most appropriate combination of auxiliary/modifier and uncertainty expression from two drop-down menus (Figure 3). ${ }^{7}$


Figure 3: Input menus in the production task.
Each participant in the interpretation task completed 24 trials, 2 for each of the 12 expressions. That is, for each expression there were 2 kinds of trials, perfectly balanced, in random order. Half of the trials ("state" trials) recorded participants' interpretation of the message alongside the objective axis, i.e. their answer to the question "How many red balls do you think there are in the urn?", expressed with a natural number selected with a slider ranging from 0 to 10 . Half of the trials ("observation" trials) recorded participants' interpretation of the message alongside the subjective uncertainty axis, i.e. their answers to the questions "How many balls do you think the sender has drawn? And how many of them do you think were red?", expressed with two natural numbers selected on sliders ranging from 0 to 10 (Figure 4).


Figure 4: Input sliders in observation trials. The picture on the right dynamically visualized the current slider selection in order to provide immediate visual feedback for a selection.

Results Results are visualized in Figures 5 and 6 and will be discussed in the light of the model's predictions below.

[^98]
## Model evaluation and criticism

Model fit The data collected in the production task and the interpretation task are respectively counts of expression choices in each observation condition, and counts of state, access and observation choices for each expression. We used the data to compute credible values for the free parameters of the model, i.e. the semantic thresholds $\theta_{\text {likely }}, \theta_{\text {possible }}$, $\theta_{\text {certainly }}$, the shape parameter of the prior belief distribution $\alpha$, the rationality parameter $\lambda$. We implemented the computational model in JAGS (Plummer, 2003) and approximated the posterior distribution of parameters given the experimental data. We assumed flat prior distributions over the parameters with support $[0 ; 1]$ for the semantic thresholds and $[0 ; 20]$ for $\alpha$ and $\lambda$. We gathered two chains of 2500 samples after an initial burn-in of 2500 . We checked convergence via $\hat{R}$ (Gelman \& Rubin, 1992). Each sample consists of a vector of inferred values for each parameter. The following table summarizes the mean values for the threshold parameters together with their $95 \%$ highest density intervals (HDIs): ${ }^{8}$

|  | $\theta_{\text {likely }}$ | $\theta_{\text {possible }}$ | $\theta_{\text {certainly }}$ |
| :---: | :---: | :---: | :---: |
| mean | 0.531 | 0.214 | 0.979 |
| HDI | $0.511-0.551$ | $0.200-0.236$ | $0.965-0.996$ |

Notice that the model recovers plausible values for thresholds given the data without assuming them from the start.

For each sample vector of parameter values our model generates a set of predictions about speaker's and listener's behavior. In order to evaluate our model we correlated each set of predictions with the set of corresponding experimental count data. The results are collected in vectors of Pearson's correlation scores, whose means and HDIs give us an indication of the overall performance of the fitted model, as summarized in the following table:

|  | expression | state | access | observation |
| :---: | :---: | :---: | :---: | :---: |
| mean | 0.649 | 0.862 | 0.883 | 0.941 |
| HDI | $0.647-0.651$ | $0.857-0.867$ | $0.880-0.886$ | $0.938-0.943$ |

Discussion Correlation scores do not provide detailed information about what aspects of the data the model can and cannot explain. To get a better sense of the performance of the model we compare data and predictions in more detail with posterior predictive checks (PPCs) (Kruschke, 2014).

We begin with the production task (Figure 5). Visual inspection of the plot suggests interesting features of the data. First, the number of observed red balls seems to have an influence on the choice of expressions. For example, with the same access of 8 (middle row of Figure 5), different observation values ( $0,2,4,6$ and 8 ) resulted in different distributions of expressions. This is an intuitive result, and the model correctly predicts the general pattern. Second, the same proportions of red balls but with different access levels seem to result in different expression choices. For example, compare the distributions of expressions observed (and predicted) with

[^99]a proportion of 0 observed red balls and access values equal to 2 and 8 , and similarly with a proportion of 1 and access values equal to 2 and 8 . The distributions are different, and the model seems to predict the patterns.

However, there are also several discrepancies between observed data and the models PPCs. Discrepancies show in Figures 5 and 6 whenever the HDIs of the PPCs do not include the observed frequencies: in these cases the model, being trained on the data, would still be surprised, so to speak, by seeing the data points where observations do not fall in to the HDIs of our PPCs. For example, the model underpredicts choice frequencies of might be possible in favor of possible in the high uncertainty conditions and underpredicts unlikely and likely in the no uncertainty conditions. More in general, the model almost always overpredicts choice of, e.g., certainly possible and might be unlikely. At the same time, whenever PPCs are off for simple expressions, the model underpredicts their choice frequency. This suggests that a crucial ingredient might be missing from the model, namely a baseline preference of some expressions over others.

Turning now to the interpretation task (Figure 6), we observe that in general the patterns displayed in the data seem to be captured relatively well by the model. However, PPCs highlight a number of discrepancies. One clear example concerns the state interpretation for unlikely and its nested versions (left panel, right column): the predictions are visibly shifted to the right compared to the data. Another feature that the model fails to predict is the relatively low counts of access choices of 5 (compared to 4 and 6) for several expressions (middle panel), although this seems to be a puzzling feature of the data rather than an obvious shortcoming of the model.

## Conclusion

Communication under high-order uncertainty raises a number of issues for formal semantics and pragmatics. Our work here is intended as a first but transparent explication of a number of assumptions that allow the formulation of a computational model of the use and interpretation of complex uncertainty expressions. The resulting model captures basic patterns in the data well enough, suggesting that our basic assumptions are not entirely off. Still, detailed model criticism also reveals a number of shortcomings. These point the way to further exploration; we see our main contribution exactly in this pointing. Most importantly, a measure of a differential inclination to produce messages (e.g., in terms of frequency, length, salience) should be included. Also, the artificial restriction on the set of message choices should ideally be relaxed as much as possible. Moreover, it will be telling to see how participants react to contextual manipulations such as of the relative relevance of information about the world state vs. information about the speaker's epistemic state.

## Acknowledgments

Financial support by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63) is gratefully acknowledged.

## References

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.
Anderson, J. R. (1991). Is human cognition adaptive? Behavioral and Brain Sciences, 14(03), 471-485.
Beyth-Marom, R. (1982). How probable is probable? a numerical translation of verbal probability expressions. Journal of Forecasting, 1(3), 257-269.
Egan, A., \& Weatherson, B. (2011). Epistemic modality. Oxford University Press.
Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084), 998-998.
Gelman, A., \& Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). Statistical Science, 7, 457-472.
Goodman, N. D., \& Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. Topics in cognitive science, 5(1), 173-184.
Herbstritt, M., \& Franke, M. (2016). Definitely maybe and possibly even probably: efficient communication of highorder uncertainty. In A. Papafragou, D. Grodner, D. Mirman, \& J. Trueswell (Eds.), Proceedings of the 38th annual conference of the cognitive science society.
Kratzer, A. (1991). Modality. In A. von Stechow \& D. Wunderlich (Eds.), Semantics: An international handbook of contemporary research (pp. 639-650). de Gruyter.
Kruschke, J. (2014). Doing Bayesian Data Analysis, 2nd Edition: A tutorial with R, JAGS, and Stan. Academic Press.
Lassiter, D. (2011). Measurement and modality: the scalar basis of modal semantics. Unpublished doctoral dissertation, NYU Linguistics.
Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. Semantics and Pragmatics, 8(5), 1-81.
Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In K. Hornik, F. Leisch, \& A. Zeileis (Eds.), Proceedings of DSC3 (Vol. 124, p. 125).
Swanson, E. (2006). Interactions with context. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge MA.
Teigen, K. H. (1988). When are low-probability events judged to be 'probable'? effects of outcome-set characteristics on verbal probability estimates. Acta Psychologica, 6(2), 157-174.
Windschitl, P. D., \& Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. Journal of Experimental Psychology: Applied, 2(4), 343.
Windschitl, P. D., \& Wells, G. L. (1998). The alternativeoutcomes effect. Journal of Personality and Social Psychology, 75(6), 1411-1423.
Yalcin, S. (2010). Probability operators. Philosophy Compass, 5(11), 916-37.


Figure 5: Percentages of expression choices in each partial observation condition, together with mean predictions and HDIs.


Figure 6: Counts of state, access and observation choices for each expression, together with mean predictions and HDIs.

# A Priming Model of Category-based Feature Inference 

Laura M. Hiatt (laura.hiatt@nrl.navy.mil)<br>US Naval Research Laboratory<br>Washington, DC 20375 USA


#### Abstract

Categorization has a large impact on how people perceive the world, especially when used to make inferences about uncertain features of new objects. While making these inferences, people tend to draw information from only one possible categorization of a new object; in addition, people are sensitive to pre-existing correlations between features. Here, we explain these trends of feature inference using a priming-based cognitive process model, and show that our model is distinguished in that it can explain not only these two main trends, but also cases where people seem to reverse the first trend and base inferences on information from multiple categories.


Keywords: categorization; priming; spreading activation; inductive inference; cognitive models

## Introduction

Categorization is a fundamental tool in human cognition. One of its main functions is to allow people to more easily understand the world by making inferences about new objects based on existing knowledge that they already have. If one sees a furry animal coming towards it and categorizes it a loose dog, then it would be natural to further infer that the animal is probably friendly.

Systematic research into how these inferences are made has shown two major trends in performance (Nosofsky, 2015; Murphy \& Ross, 1994, 2007; Griffiths, Hayes, \& Newell, 2012). First, people seem to base inferences on a single identified category for an object, even if the object's categorization is uncertain (called the single-category view). So, for example, people would typically infer that the dog is friendly without considering that it might be a fox, which should be avoided. Second, people are sensitive to correlations between features, and are more likely to infer features that are strongly associated with the observed features of the new object. For example, people would be further biased towards inferring the dog is friendly if it were wagging its tail.

While there is a large body of research that supports these two trends, here we consider a series of experiments performed by Murphy and Ross (1994) that comprehensively considered several variants and extensions of the basic inference paradigm. The authors, however, admit that their overall results challenge many of the formal models of categorization and inference (Murphy \& Ross, 1994, 2007), with none fully explaining the results. Recently, Nosofsky (2015) developed a exemplar model of feature inference that does qualitatively capture their results. Notably, however, Nosofsky's (2015) analysis does not discuss an important caveat of the first trend: that responses seem to shift towards a multicategory view, where more than one possible category is considered when making the inference, if participants do not explicitly identify the category before making the feature inference (Murphy \& Ross, 1994; Griffiths et al., 2012).

We present here a priming-based process model of inductive feature inference that explains these two main results, including this caveat. Situated in the cognitive architecture ACT-R/E (Trafton et al., 2013), a critical aspect of our model is that its inferences are based not only on what stimuli have been seen, but also on what the model is currently thinking about (i.e., what is in its working memory). We show our model's ability to account for feature inferences in four main experiments that are particularly indicative of the trends of feature inference: Experiments 1, 5, 6, and 8 from Murphy and Ross (1994).

## Experiments

In the four experiments we consider from Murphy and Ross (1994), participants were shown category structures with differently shaded geometric objects, grouped together and labeled with the category they represent (e.g., Figure 1). Participants were told that the categories represented different children who drew the objects, and that the objects were illustrative of a larger set of drawings by each child. Then, the experimenter told participants about a new drawing, but only shared one feature of it, such as a triangle; this feature, the query feature, was typically chosen to be ambiguous in which child drew it. Participants were then asked what they thought the other feature of the new drawing was (such as the triangle's color). Additionally, in some experiments, participants were asked to categorize the drawing (i.e., say which child drew it) before they inferred the second feature. The most likely category for each query is called the target category.

Experiment 1 focused on whether inferences are made using information from single, or multiple, categories. The categories are shown in Figure $1^{1}$. This experiment had two conditions. In the increasing condition, the query feature was a triangle. The target category for a triangle is Bob, since Bob has the most triangles. The target-category feature, or the feature that would be selected by primarily considering the target category using a single-category view, is black. This condition is called increasing because there is additional evidence outside the target category that the triangle would be black, since Sam and John also sometimes draw triangles, and they also sometimes draw black objects.

Now, consider a new drawing that is a square. Here, the target category is John, and the target-category feature is white. In this condition, the neutral condition, there is no evidence outside of the target category that the square would be white;

[^100]

Figure 1: Category structure for Experiment 1. In the increasing condition, the query feature is a triangle, the target category is Bob, and the target-category feature is black. For the neutral condition, the query feature is a square, the target category is John and the target-category feature is white. Adapted from Figure 1 of (Murphy \& Ross, 1994).
no other child draws both squares and white objects.
Almost all of the 29 participants selected the target category and target-category feature for both the increasing and neutral conditions, and so ceiling effects prevented them from being statistically compared. Participants also, however, provided a probability estimate of their certainty in their response. These probability judgments did not have a ceiling effect, yet provide no evidence of a difference between the two conditions: the average certainty for each condition was $53 \%$. This parity supports the single-category view of feature inference by suggesting that, despite the additional evidence for the target-category feature present in the increasing condition, participants only took the target category into account when making their inference.

Experiments 5 and 6 used a category structure in which the single-category view and multiple-category view suggest different patterns of feature inferences (Figure 2). Further, they considered how the initial step of identifying the target category may affect participants' use of single vs. multiple categories in their inference. Here, the query feature is a triangle, and the target category is Bob, since he drew more triangles than the other children. The single-category view suggests black as the inferred feature; black is thus considered the target-category feature. A multiple-category view, however, suggests that black and white are equally likely.

The results of the experiments support both these views, depending on whether participants were asked to make the initial categorization step. In Experiment 5, where participants did not initially categorize the drawings, $58 \%$ of the 32 participants chose the target-category feature, black, with the majority of remaining responses as white. This difference was not significant, supporting the multiple-category view. In Experiment 6, however, where participants did categorize the drawings before predicting the other feature, $82 \%$


Figure 2: Category structure for Experiments 5 and 6. Here, the query feature is a triangle, the target category is Bob, and the target-category feature is black. Adapted from Figure 3 of (Murphy \& Ross, 1994).
of the 36 participants responded with the target-category feature. Additionally, $88 \%$ of the participants that categorized the drawing into the target category responded with the targetcategory feature. This supports the single-category view and suggests that participants were biased by the target category when they identified it before making their inference.

Experiment 8 focused on exploring how feature correlations may affect predictions. Here, the query features in two conditions were explicitly controlled to have different degrees of correlation with the target-category features. All participants were asked to assign the drawing to a category before responding to the feature queries. Figure 3 shows an example category structure. In the correlated condition, the query and target-category features were perfectly correlated: the query feature was a circle, the target category was " $D$ " and the target-category feature was "vertically striped." In the uncorrelated condition, the features are only weakly correlated, with a query feature of triangle, a target category of "C" and a target-category feature of white.

The results show that $95 \%$ of the 26 participants selected the target category across both conditions. More importantly, more participants selected the target-category feature for the correlated condition ( $94 \%$ ) than for the uncorrelated condition ( $90 \%$ ). This suggests that people are biased towards correlated features when they make inferences.

## Model

We developed a priming-based process model of feature inference given uncertain categorizations, situated within a computational cognitive architecture, ACT-R/E, that allows us to model the processes people undergo as they perform tasks. In this architecture, concepts that are thought about at the same time become associated in memory, and then can prime one another; by using ACT-R/E, we are able to develop a priming-based account of feature inference that is supported by the underlying principles of this existing, well-studied theory of cognition. Here, we first describe the general principles


Figure 3: Category structure for Experiment 8. For the correlated condition, the query feature is a circle, the target category is D and the target-category feature is vertically-striped. For the uncorrelated condition, the query feature is a triangle, the target category is C and the target-category feature is white. Adapted from Figure 5 of (Murphy \& Ross, 1994).
of our model. Then, we give further details of ACT-R/E, and discuss how our model's principles interact with the architecture to make specific predictions about feature inference.

The process model has two phases corresponding to the two phases of the experiment: an initial phase where the model looks at, encodes, and stores the categories and objects in memory; and an inference phase where the model makes the category and feature predictions. During the initial phase, each of the objects becomes associated with its underlying features; both the features and objects, in turn, also become associated with their corresponding category. These associations mean that the concepts prime one another when the model is thinking of them.

Then, during the inference phase, to predict the category of a new object, the model selects the category with the most priming, including priming from the query feature. Consequently, the model's category response is heavily influenced by the presence of the query feature in the category. To perform the feature prediction, the model selects the object in memory that has the most priming, including priming from the query feature and, when applicable, the selected category. The second feature of that object is then considered to be the inferred feature. This means that the predicted feature is heavily influenced by both the correlation between the two features, and the prevalence of that feature within the identified category (when the category is identified).

## Model Architecture

The model was developed within the cognitive architecture ACT-R/E (Trafton et al., 2013), an embodied version of the ACT-R cognitive architecture (Anderson, 2007). At a high level, ACT-R/E is an integrated, production-based system, and models in ACT-R/E capture the core cognitive processes that people go through as they undergo tasks. At its core are the contents of its working memory; working memory indi-
cates, for example, what the model is looking at, what it is thinking, and its current goal. At any given time, there is a set of productions (if-then rules) that may fire because their preconditions are satisfied by the current contents of working memory. From this set, the production with the highest predicted usefulness is selected to fire. The fired production can either change the model's internal state (e.g., by adding something to working memory) or its physical one (e.g., by pressing a key on a keyboard). In our discussion, we abstract over these productions and instead describe processes at a higher level (i.e., we say that we look at an object, instead of discussing the 3-4 productions that must fire to achieve that).

Working memory is represented as a set of limited-capacity buffers that can contain thoughts or memories. In addition to the symbolic information (i.e., factual information) represented as part of these memories, memories have activation values that represent their relevance to the current situation, and guide what memories are retrieved from longterm memory and added to working memory at any given time. Activation has three components, activation strengthening, spreading activation, and activation noise, that together have shown to be an excellent predictor of human declarative memory (Anderson, Bothell, Lebiere, \& Matessa, 1998; Anderson, 1983; Schneider \& Anderson, 2011; Thomson, Harrison, Trafton, \& Hiatt, 2017). Noise is a random component that models the noise of the human brain; since its presence would not affect our results, we ignore noise in the rest of this paper. Activation strengthening is learned over time and is a function of the frequency and recency with which the memory has been in working memory in the past. The predominant role of activation strengthening in this experiment relates to ordering effects, which the experimental stimuli's counterbalancing averages out. Therefore, we primarily focus the rest of our discussion of activation on its third component: spreading activation, or priming.

Priming is a short-term activation that sources from working memory, distributing activation along associations between the contents of working memory and other memories. Memories become associated when they are in working memory at the same time. Once established, an association from memory $j$ to memory $i$ has a strength value that affects the degree to which $j$ primes $i$, and intuitively reflects the probability that memory $i$ is relevant while thinking of memory $j$. This allows spreading activation to capture correspondences between memories that typically co-occur, as well as memories that are semantically related (such as an object and its color and shape). Association strengths are calculated in a Bayesian-like way, and are a non-standard adaptation of ACT-R's Bayesian-based priming mechanisms. We use this adaptation to account for the large numbers of associations and objects needed by the experiments we consider here, which ACT-R's original formulation is unable to do, as well as to capitalize upon its theory that priming stems from working memory; see Hiatt and Trafton (2016) for more information on our priming mechanisms.


Figure 4: Priming for Experiment 1, increasing condition, for the query feature's categorization (left) and the inferred feature (right). Thicker arrows indicate more priming; thinner indicate less. The cumulative priming means that Bob will be selected as the category and black will be the inferred feature.

ACT-R/E models interact with the world using ACT-R/E's built-in functionality. Models can view visual items on a simulated monitor, and can act on the world by pushing keys on a simulated keyboard and clicking a simulated mouse. ACTR/E models are also inherently tied to physical embodiment (i.e., executing models on a robot), but we do not use that functionality in this paper.

## Model Details

Our model for feature inference starts out with only the task knowledge and productions necessary to complete the tasks. It also assumes prior exposure to the category names, since they are names participants would have encountered frequently in their daily lives (i.e., "A", "John", etc.). There are no initial associations; are all learned during the experiment. The model "looks at" the stimuli as the participants did via its simulated monitor.

During the initial experiment phase when the model is looking at the categories and objects, it first finds a category to look at, encodes it and adds it to working memory. While continuing to think of the category, the model then looks at, encodes and adds to working memory each of the objects in that category, while making note of their color and shape. Consequently, as it looks at each object: the object (i.e., "black-triangle") becomes associated with the category (i.e., "Bob"); the object's features (i.e., "black" and "triangle") become associated with both the object and the category; and the features of the object become associated with each other. When it has finished looking at all objects of a category, it repeats this process with the other categories until it has looked at all of the categories and objects on the screen.

During the inference phase, the model first adds the query feature to working memory as part of the process of interpreting the query. When asked to infer the category of an object, the model retrieves the category from memory with the highest activation, including both activation strengthening and spreading activation (i.e., priming), responds with the retrieved category, and leaves the category in working memory. For example, Figure 4, left side, shows the priming when selecting the category for Experiment 1's increasing condition. Then, when asked to infer the object's missing feature,
the model retrieves an object while both the retrieved category (when applicable) and the query feature are in working memory. Again, the object with the highest activation, both activation strengthening and priming, is retrieved; Figure 4, right side, shows this for Experiment 1's increasing condition. The second feature of the retrieved object is given as the response to the query.

## Model Results

In the original experiments, several versions of the basic category structure were created to counterbalance features and category locations. We varied our category structures accordingly, then used our model to simulate data from 500 participants per experiment to allow our results to better converge on the model's true predictions; our reported results are the proportion of the 500 model runs that responded with the target category, target-category feature, etc., for each query.

The model had the same parameters for each experiment. The activation strengthening decay parameter was 0.45 instead of its default of 0.5 . The associative learning rate was 4.8 , representing a moderate rate of learning. There is no real default value for this parameter. All other parameters were set to their default values.

The main experiment and model results are shown in Table 1. For Experiment 1, the model exhibited perfect performance, always selecting the target category, and always selecting the target-category feature for both the increasing and neutral conditions. This is comparable to the experimental results, where almost all participants also selected the target category and target-category features.

In this experiment, however, despite almost all participants selecting the target-category feature, participants' probability judgments of their responses were not as certain, with an average judgment for each condition of $53 \%$ for both the increasing and neutral conditions. While we have no a priori way of extracting probability judgments from the modeling framework we utilize, our model does informally support these results. This is because, from our model's point of view, these conditions' structures are the same. Both include two objects with the query feature and target-category feature in the target category; one object with just the query feature in the target category; and two objects with just the query feature outside of the target category. Thus, in both conditions, while the black-triangle object (or white-square object) is the highest activated object, it only receives about half of the total priming, suggesting a probability judgment of $50 \%$.

For Experiment 5, the model selected the target-category feature $50 \%$ of the time, which moderately reflects the experiment's results. In Experiment 6, the model very strongly matched the experimental data, selecting the target-category feature $80 \%$ of the time, as compared to the experiment's $82 \%$. Additionally, $89 \%$ of the model runs that categorized the drawing into the target category responded with the targetcategory feature, compared to the experiment's $88 \%$.

For Experiment 8, $95 \%$ of model runs selected the target

Table 1: Model Results

| Experiment | Condition/Participant Group | Measurement | Exp. Data | Model |
| :---: | :---: | :---: | :---: | :---: |
| Exp. 1 | increasing | probability judgments | $53 \%$ | $50 \%$ |
|  | neutral | probability judgments | $53 \%$ | $50 \%$ |
| Exp. 5 | all participants | target-cat. feature | $58 \%$ | $50 \%$ |
| Exp. 6 | all participants | target-cat. feature | $82 \%$ | $80 \%$ |
|  | target cat. correct only | target-cat. feature | $88 \%$ | $89 \%$ |
| Exp. 8 | all participants | target category | $95 \%$ | $95 \%$ |
|  | correlated | target-cat. feature | $94 \%$ | $100 \%$ |
|  | uncorrelated | target-cat. feature | $90 \%$ | $91 \%$ |

category, the same as in the experiment. All of the model runs selected the target-category feature for the correlated condition, and $91 \%$ selected the target-category feature for the uncorrelated condition. Again, this strongly corresponds to the experimental results, where there was a significant difference between the two conditions, with $90 \%$ of participants selecting the target-category feature in the uncorrelated condition vs. $94 \%$ for the correlated condition.

## Model Discussion

Recall the two main trends in research on feature inference for uncertain categorizations that are illustrated by the four experiments we consider here. First, people are biased towards the single-category view when making feature inferences; the bias seems to be modulated, however, when they do not categorize the object first. And second, people's inferences are also sensitive to correlations between features, selecting correlated features more often than non-correlated.

The model explains both of these trends via priming between the features, objects and categories. It explains the first trend, and its caveat, because its predictions are based on the sources of priming in working memory, and as such are not inherently based on the consideration of single- or multiplecategories. When making a feature prediction, the model always has the query feature in memory, which primes objects that are associated with it. This serves to provide suggestions compatible with the multiple-category view of what the predicted feature should be. For example, in Experiment 5, where triangle is the query feature and there is no categorization step, triangle equally primes black-triangle and whitetriangle, because there are equal numbers of them. This leads to a roughly an equal likelihood ( $50 \%$ ) of the predicted feature being black or white. While this underestimates the $58 \%$ response rate of the experimental data, given the lack of statistical significance in this experiment, we are comfortable concluding that our model explains this trend.

In conditions where participants categorize the feature before making their prediction, priming stems not only from the query feature but also from the category, which provides suggestions compatible with the single-category view of what the inferred feature should be. In Experiment 6, identical to Experiment 5 but with an added categorization step, when
shown a triangle, the model generally selects Bob as the category (i.e., Figure 4). Bob then strongly primes black-triangle, since it has three of them, and weakly primes white-triangle, since it has only one of them. Combining this category priming with the priming from the query feature, black-triangle overall receives more. Again, this matches the data, where $82 \%$ of participants overall selected black as the inferred feature, and $88 \%$ of participants who identified Bob as the target category selected black as the inferred feature. Overall, then, the model's use of priming in memory allows the model to capture conditions both where participants seem to be biased towards the single-category view, and where they do not - a major contribution of the model.

The model also explains the second main trend of feature prediction, where participants are sensitive to correlations between features. There are two reasons for this. The first is that correlated objects, on average, have slightly higher activation strengthening, since they will be more familiar to participants than objects with less common feature pairings. The second reason is that correlated objects will receive much higher levels of priming from their underlying features because that priming is, in a sense, undiluted by other options. For example, in Experiment 8, where the correlated query feature is circle, the only object primed by circle is vertically-stripedcircle. The target category, D , also spreads a high amount of activation to vertically-striped-circle, since there are three of them in that category, further underscoring the correlated feature as the answer. In contrast, for the uncorrelated query, both sources of priming (the query feature triangle, and the target category C) prime white-triangle in addition to strongly priming the target black-triangle. Thus, the model suggests that for the correlated condition, the target-category feature should almost exclusively be selected, whereas in the uncorrelated condition, the target-category feature should just mostly be selected. These explanations match the data, where the target-category feature was selected for $94 \%$ of correlated, but only $90 \%$ of uncorrelated, structures.

## General Discussion

The authors of the experiments that we model here were ultimately interested in characterizing people's inference behaviors across different manipulations of categories and fea-
tures (Murphy \& Ross, 1994). Recently, as we mentioned, Nosofsky (2015) proposed an exemplar model that qualitatively accounts for the majority of the results. The model is based on an equation that calculates the similarity between feature/category pairs using two parameters: the salience of the feature, and the salience of the category. The probability of inferring the target-category feature is then found by summing the similarity of the query feature/category pair to all displayed feature/category pairs with the target-category feature, and dividing by the summed similarity of the query feature/category pair to all displayed feature/category pairs (irrespective of the target-category feature).

Our view of this promising work, however, is that it does not consider an important result of the experiments: that of the difference in results between experiments where participants explicitly identified the target category, and where they did not (e.g., Experiment 5 vs. Experiment 6). Recall that when participants were asked to identify the target category before making their inference, a large and significant majority responded according to the single-category view; when participants were not asked to identify a target category before making their inference, however, participants' responses greatly shifted towards the multiple-category view. Nosofsky (2015) do not discuss this difference, and considers the results of Experiment 5, instead, as weakly supportive of the singlecategory view that is more strongly suggested by Experiment 6. Although dynamically adjusting the parameter settings depending on the specific queries of the experiment may lead to this difference in predictions, there is no intuition for how this parameter setting change may occur.

Our priming-based process model of feature inference, however, naturally answers that question as part of its core theory. Our model indicates that the difference in results is due to an underlying difference in the way that the experiments are processed by the human mind. It accounts for this difference because it includes the sources of priming in working memory to be a key part of its predictions. It suggests that when a person has explicitly thought about a category, the category is included as part of the inference process, biasing the model towards the single-category view; when a person has not, the model relies only on priming from the query feature, biasing the model towards the multiple-category view. Our model thus explains the same qualitative trends as Nosofsky (2015) while also accounting for this additional aspect of feature inference, and quantitatively matching the data.

Another model that has been proposed for explaining feature inference is the rational model and its associated variants (Anderson, 1991; Sanborn, Griffiths, \& Navarro, 2010). This model, while also rooted in Bayesian-based reasoning, has been shown to have trouble accounting for the breadth of the results we model here (Nosofsky, 2015). A recent promising version of this model was developed by Konovalova and Le Mens (2016), whose rational model is sensitive to uncertainty in categorization; our belief, however, is that it also would have trouble accounting for differences stemming from
the presence or lack of an initial categorization step.

## Acknowledgments

This work was supported by the Office of the Secretary of Defense and the Office of Naval Research. The views and conclusions contained in this paper do not represent the official policies of the U.S. Navy. We thank Greg Trafton, Sunny Khemlani and Tony Harrison for their advice and comments.

## References

Anderson, J. R. (1983). A spreading activation theory of memory. Journal of Verbal Learning and Verbal Behavior, 22(3), 261-295.
Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98(3), 409.
Anderson, J. R. (2007). How can the human mind occur in the physical universe? Oxford University Press.
Anderson, J. R., Bothell, D., Lebiere, C., \& Matessa, M. (1998). An integrated theory of list memory. Journal of Memory and Language, 38(4), 341-380.
Griffiths, O., Hayes, B. K., \& Newell, B. R. (2012). Featurebased versus category-based induction with uncertain categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(3), 576.
Hiatt, L. M., \& Trafton, J. G. (2016). Familiarity, priming and perception in similarity judgments. Cognitive Science. (doi: 10.1111/cogs.12418)
Konovalova, E., \& Le Mens, G. (2016). Predictions with uncertain categorization: A rational model. In Proceedings of the Annual Conference of the Cognitive Science Society.
Murphy, G. L., \& Ross, B. H. (1994). Predictions from uncertain categorizations. Cognitive Psychology, 27, 148-193.
Murphy, G. L., \& Ross, B. H. (2007). Use of single or multiple categories in category-based induction. Inductive reasoning: Experimental, developmental, and computational approaches, 205-225.
Nosofsky, R. M. (2015). An exemplar-model account of feature inference from uncertain categorizations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(6), 1929.
Sanborn, A. N., Griffiths, T. L., \& Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. Psychological Review, 117(4), 1144-1167.
Schneider, D. W., \& Anderson, J. R. (2011). A memorybased model of hick's law. Cognitive Psychology, 62(3), 193-222.
Thomson, R., Harrison, A. M., Trafton, J. G., \& Hiatt, L. M. (2017). An account of interference in associative memory: Learning the fan effect. Topics in Cognitive Science, 9(1), 69-82.
Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello, II, F., Khemlani, S. S., \& Schultz, A. C. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. Journal of Human-Robot Interaction, 2(1), 30-55.

# Quantifying the impact of active choice in word learning 

Shohei Hidaka (shhidaka@jaist.ac.jp) and Takuma Torii (tak.torii@jaist.ac.jp)<br>Japan Advanced Institute of Science and Technology<br>1-1 Asahidai, Nomi, Ishikawa, Japan

George Kachergis (george.kachergis@gmail.com)<br>Dept. of Artificial Intelligence / Donders Institute, Radboud University Nijmegen, the Netherlands


#### Abstract

Past theoretical studies on word learning have offered simple sampling models as a means of explaining real word learning, with a particular goal of addressing the speed of word learning: people learn tens of thousands of words within their first 18 years. The present study revisits past theoretical claims by considering a more realistic word frequency distribution in which a large number of words are sampled with extremely small probabilities (e.g., according to Zipf's law). Our new mathematical analysis of a recently-proposed simple learning model suggests that the model is unable to account for word learning in feasible time when the distribution of word frequency is Zipfian (i.e., power-law distributed). To ameliorate the difficulty of learning real-world word frequency distributions, we consider a type of active, selfdirected learning in which the learner can influence the construction of contexts from which they learn words. We show that active learners who choose optimal learning situations can learn words hundreds of times faster than passive learners faced with randomly-sampled situations. Thus, in agreement with past empirical studies, we find theoretical support for the idea that statistical structure in real-world situations-potentially structured for learning by both a self-directed learner, and by a beneficent teacher-is a potential remedy for the pathological case of learning words with Zipf-distributed frequency.


Keywords: cognitive models of language acquisition; cross-situational word learning; statistical learning

## Child word learning

One of the most prominent differences between human and nonhuman cognition is our language ability. Much research has been dedicated to understanding the human capability for language, with a great deal of discussion focused on the process of language acquisition. A central debate in this conversation considers whether acquisition is based on innate and language-specific mechanisms (Chomsky, 1965; Gleitman, 1990), or bootstrapped from domain-general mechanisms (Smith, 2000; Kachergis, 2012). From the former perspective, humans become competent language users-mastering a complex system of syntax to produce endless semantics-very rapidly, and with relatively little training.

Word learning has been treated as an indicator of language development, and has been compared with a number of other indicators of cognitive abilities, such as memory (Vlach \& Johnson, 2013; Vlach \& Sandhofer, 2012). Although there are multiple empirical estimates of the number of words that children acquire, many studies agree that child's word learning is quite fast. Early
word production starts when the child is 12 months old on average, and by 18 months children can produce 50 words and comprehend 100-150 (Hulit \& Howard, 2002). By 18 years of age, it is estimated we know over 60,000 words (Bloom, 2000). Under the assumption that each child has 8 hours of word learning opportunity everyday, these estimates mean the child learns a new word every learning hour for 18 years of the life.

Given these empirical estimates of word learning, theoretical studies have attempted to account for the quantitative characteristics of word learning. The first question is: What combination of learning mechanisms and structure in the language environment allows children to learn at this rate? This question poses a good necessary condition for any account of child word learning, as it needs to address this quantitative aspect of word learning.

As a first-order approximation, child learning may be modeled as an independent sampling process in which each word is learned independently. To estimate the fastest possible learning rate, (Blythe, Smith, \& Smith, $2010,2016)$ proposed an idealized learning model to address acquiring a full lexicon in the long term: 60,000 words over 18 years. In their model, each word is learned with its first sample - known as fast mapping in the developmental literature. Under the simplifying assumption that each word is independently learned via fast-mapping, and its word frequency is distributed uniformly, their mathematical analysis of the model showed that a cross-situational learner is sufficiently fast to learn all 60,000 words after experiencing a reasonably small number of spoken words.

## Theoretical approach

Blythe et al.'s theoretical estimate has been treated as a theoretical implication that shows learning via independent fast-mapping of words is efficient enough to be a model of child word learning. In this study, we reinspect this theoretical implication by introducing a more realistic word frequency distribution. Our mathematical analysis implies that the learning rate of the independent fast mapping is quite sensitive to the word frequency distribution. More importantly, even fast mapping-the most efficient learning, requiring only a single sample, can be too slow to learn 60,000 words in 18 years, if
word frequency follows Zipf's law (Zipf, 1949) or a fattailed distribution which is often found in natural corpora. Thus, our analysis implies that the independent fast-mapping model cannot be an account for child word learning, if there are many words sampled less frequently. This mathematical implication leads to an empirical test of whether the word distribution in the child-directed speech is uniform or non-uniform such as a Zipf distribution. Thus, in the second study, we analyzed the CHILDES corpora for word distribution in child-directed speech.

Given this result of the mathematical analysis, we explore an extension of the word learning mechanism by additionally assuming that the word learning is more $a c$ tive than that is supposed to be traditionally. Typically, as analyzed in the past studies above (Blythe et al., 2010, 2016), the learning is supposed passive - the learner has no choice but observing samples words and objects from a given probability distribution. This is certainly oversimplified, as actual child word-learners choose when, where and from whom they would like to learn words. Thus, our second analysis estimates the impact of a form of active choice of situations in word learning. Our analysis shows that active learning is likely to have a sufficiently beneficial impact to make word learning fast enough to happen on a realistic timescale.

## Independent fast-mapping learning

## Uniformly distributed word frequency

Blythe, Smith, \& Smith (2010) proposed a mathematical model of word learning, which has a closed-form expression under a certain simplification. In their recent study, Blythe, Smith, \& Smith (2016) analyzed essentially the same model, although slightly modified for analytic convenience. Here we briefly introduce the most recent form (2016) of their model.

Blythe et al. originally consider cross-situational word learning. Suppose there are $W$ words and $O$ objects in the hypothetical world. Further the numbers of words and objects are equal, $W=O$, in their cross-situational learning scheme, and every object has its name and no objects have two names. Namely, there are $W$ correct pairs of words and objects. Without loss of generality, denote the $W$ pairs by $1,2, \ldots, W$, and suppose $k^{\text {th }}$ object is paired with the $k^{\text {th }}$ word.

Given these pairs being unknown, a word learner is to infer correct pairs by going through episodes. In each episode, the learner is exposed to $M \leq W$ words and $M$ objects, without any explicit information on which word is paired with which object. With one episode with $M \geq 2$ objects and words, the learner cannot tell which of $M$ words should be associated to which of $M$ objects.

The most simple model among a series of extended ones is called fast-mapping learning model. In the literature of language development, it is well-known that chil-
dren as young as three years old can quickly generalize a novel name to objects when they hear the novel name given to its fast instance. Due to this one-shot nature of their word learning, it is called fast mapping. Capturing this empirical finding, the fast-mapping learner in the model is supposed to learn a new pair of word and object only with the first experience of it. The fastmapping learner is equivalent to the cross-situational learner, if there is one correct pair of object and word in each episode $(M=1)$.

As fast-mapping learning is the most efficient scheme (at least for independent word learning), it gives a good baseline estimate of the number of samples to learn all the words in a given list. Blythe et al. (2010) model a fast-mapping learner acquiring words independently drawn from a uniform distribution of $W$ words given in each episode. As every episode has one word with probability $1 / W$, this is equivalent to the so-called Coupon Collector's Problem (Blom, Holst, \& Sandell, 1994). In this problem, the expected time $T$ to finish sampling all the words is

$$
E[T]=\sum_{i=1}^{W} E\left[t_{i}\right]
$$

where $t_{i}$ is the time to sample a $i^{\text {th }}$ new word given $(i-1)$ words being learned. Thus,

$$
\begin{equation*}
E[T]=W \sum_{i=1}^{W} 1 / i \approx W \log W \tag{1}
\end{equation*}
$$

Setting the number of words $W=60,000$, which is an empirical estimate of the number of words 18 years old knows on average, $T=660,126$. This estimate is comparable with the "reasonable" number of samples justified by Blythe et al. (2010) which individual children can be exposed to for their 18 years of lives.

## Non-uniformly distributed word frequency

Here we extend this analysis on the fast-mapping learner to the case with word frequency distributed nonuniformly. Our extended analysis will reveal that the estimate based on Equation (1) by Blythe et al. is quite "optimistic", as an estimate with non-uniform word distribution is larger than that in general.

Here let us derive the number of episodes $T$ that, for $0 \leq \epsilon \leq 1$, the $(1-\epsilon)$ of children learned all the $W$ words listed. Suppose a set of $W$ words in which each word $1, \ldots, W$ is drawn from the distribution $p=\left(p_{1}, \ldots, p_{W}\right)$. The proportion of children who finished learning all the words is $(1-\epsilon)$ for $0<\epsilon<1$ requires the number of episodes $T$, which is the root of

$$
\begin{equation*}
\prod_{i=1}^{W}\left(1-\left(1-p_{i}\right)^{T}\right)=1-\epsilon \tag{2}
\end{equation*}
$$

The left hand side of (2) is the probability that every word is present at least once in the $T$ episodes.

Write

$$
f_{W, \epsilon}(x):=\frac{\log \left(1-(1-\epsilon)^{1 / W}\right)}{\log (1-x)}
$$

For the uniform distribution, $p_{i}=1 / W$ for every $i=$ $1, \ldots, W$, the root of (2) is given by

$$
\begin{equation*}
T=f_{W, \epsilon}(1 / W) \tag{3}
\end{equation*}
$$

This $T$ is the number of episodes with which the proportion of children finished learning all the words is $(1-\epsilon)$. Setting $\epsilon=1 / 2$ in (3), we obtain the median of $T$, $f_{W, 1 / 2}(1 / W)$, that is comparable with the mean of $T$ in (1).

Unlike (3) for the uniform distribution, the root $T$ of Equation (2) in general is not closed-form. Thus, let us consider the upper and lower bound for the root instead of the rigorous form of it. For the general word distribution $p=\left(p_{1}, \ldots, p_{W}\right)$, the intermediate value theorem states that there exists a unique constant $c$ holding $\min p \leq c \leq \max p$, with which the root of (2) is expressed as

$$
T=f_{W, \epsilon}(c)
$$

Equivalently, we have inequality

$$
f_{W, \epsilon}(\max p) \leq T \leq f_{W, \epsilon}(\min p)
$$

As we are interested in the worst possible estimate of $T$, this inequality states that the upper bound $T_{+}:=$ $f_{W, \epsilon}(\min p)$ of $T$ is characterized with the probability to sample the least frequent word $\min p$.

This extended mathematical analysis implies that the uniform distribution $q=(1 / W, \ldots, 1 / W)$ of words gives the minimal possible upper bound $T_{+}$among any frequency distribution of $W$ words, as any distribution $\min p$ of $W$ words holds $\min p \leq \min q$. Therefore, the expectation of $T$ in the form of (1) with the uniform distributed words is the most optimistic, which may underestimate the number of episodes required for learning with a realistic word distribution.

For example, let us consider an alternative case that the $W$ word follows the Zipf distribution $p=$ $\left(1^{-1} / H_{W}, 2^{-1} / H_{W}, \ldots, W^{-1} / H_{W}\right)$, where $H_{W}$ is the harmonic number $H_{W}=\sum_{i=1}^{W} i^{-1}$. In this case, the minimal probability is $\min p \approx 1.44 \times 10^{-6}$, and the upper bound $T_{+}$is $1.08 \times 10^{7}$ for $\epsilon=0.01$. This estimate means that learning of Zipf-distributed words requires 16.4 times as many samples as learning of uniformlydistributed words. That means that 206 independent episodes exposed to a word learner every hour (or three episodes every minute), assuming 8 hours of learning everyday of 18 years of life. This estimate cannot possibly be considered "reasonable" with respect to ordinary life of children in any culture.

## Sensitivity to non-uniformity of word frequency distribution

To analyze the sensitivity to the non-uniformity, here we analyze the Zipf distribution with different exponent parameters. Denote the Zipf distribution with the exponent parameter $a \geq 0$ by $p=$ $\left(1^{-a} / H_{W, a}, 2^{-a} / H_{W, a}, \ldots, W^{-a} / H_{W, a}\right)$ where $H_{W, a}$ is the generalized harmonic number $H_{W, a}=\sum_{i=1}^{W} i^{-a}$. It is reduced to the uniform distribution by $a=0$. The larger the exponent $a$ is, the minimal probability $\min p$ is smaller. Thus, here we analyze the upper bound $T_{+}$ as a function of the exponent parameter $a$.

Write $T_{+}=f_{N, \epsilon}(\min p)$, which gives a reasonable estimate of the upper bound of the root $T$ of (2). As a function of the exponent $a$, we have

$$
\frac{\partial \log T_{+}}{\partial a}=\frac{p_{\min }\left(\frac{\partial H_{N, a}}{\partial a} / H_{N, a}+\log W\right)}{\left(1-p_{\min }\right) \log \left(1-p_{\min }\right)}
$$

and further we have

$$
\frac{\partial^{2} \log T_{+}}{\partial a^{2}} \geq 0
$$

This implies the $T_{+}$is a super-exponential monotone function of the exponent $a$. It is also numerically confirmed in Figure 1, in which the numbers of episodes are shown as functions of the exponent for $W=$ 10000,60000 . In this plot, $a=0$ shows an estimate for the uniform distribution, and $a=1$ shows that of the standard Zipfian distribution. It is striking that even the fastest learning such as fast mapping can be quite slow (exponentially as a function of $a$ ) for with distributions with some item with a very small probability.

## Empirical dataset

Given theoretical implication in the previous study, let us analyze an empirical word distribution, which children typically are exposed to. It is difficult to exactly count "episodes" or "pairs of word and object" in a real dataset, due to its ambiguity of definition and it is also up to children's subjective perspective. Here, as a proxy of them, we counted the word frequency based on child-directed speech in the CHILDES corpus (MacWhinney \& Snow, 1990). Figure 2 shows a representative word distribution of 51,446 words aggregated over 4,163 transcripts of all the corpora in CHILDES retrieved in December 2007. The minimal word probability was $1.089 \times 10^{-7}$, which gives the upper bound $T_{+}=$ $f_{51446,0.01}\left(1.089 \times 10^{-7}\right)=1.420 \times 10^{8}$ or the median estimate $T_{+}=f_{51446,0.5}\left(1.089 \times 10^{-7}\right)=1.030 \times 10^{8}$. These estimates of required samples, an order of magnitude larger than the optimistic theoretical estimate, suggest that it is difficult to learn these empirical words with this Zipfian-like frequency distribution.


Figure 1: For $\epsilon=0.01,0.5,0.99$ (broken and solid lines), $N=10000,60000$ and the exponent $a=0,0.25, \ldots, 1.5$, the required number of samples $M$ for a generalized Zipf distribution $p_{k}=k^{-a} / \sum_{k=1}^{N} k^{-a}$ is numerically calculated by the root of Equation (2).

## Active choice of situations

## Formulation

The implication of the mathematical analysis above, which suggested that even fast-mapping may not be efficient enough for non-uniformly distributed words, raises a controversy between past theoretical analyses and empirical findings of quantitative aspects of word learning.

Here, we explore a possibility to reconcile the discrepancy between theory and empirical findings, by considering a further relaxation of past theoretical assumptions about children's word learning. In the conventional theoretical framework, the learner is assumed to be passive, having no choice but to observe and learn from a given context: a randomly-sampled set of of objects, of which a (random) subset are labeled with words. This assumption of a passive learner simplifies the theory, but surely underestimates real learners, who have some choice about which contexts they experience. Here, we consider a type of active learner who is able to choose from which situation/context he or she learns words.

Suppose that there are $N$ word-object pairs and $M$ situations, and that the conditional probability to observe the $i^{\text {th }}$ word-object pair is $p_{i j}$ given the $j^{\text {th }}$ situation. Thus, the active learner has a choice of the situation out of the given $M$ situations from which he or she learns the word-object pairs. Suppose that the active learner chooses the $j^{\text {th }}$ situation by the probability $q_{j}$. Let us denote the $N \times M$ matrix of the conditional probability by $P=\left\{p_{i j}\right\}_{i j}$ and the $N \times 1$ vector of the choice probability by $q=\left(q_{1}, q_{2}, \ldots, q_{M}\right)^{T}$. With this notation, the marginal probability of word-object pairs is given by the vector $P q \in \mathbb{R}^{N}$. According to our mathematical anal-


Figure 2: Word frequency in a corpus aggregated from the CHILDES transcripts.
ysis in the previous section, the minimal probability of objects decides the number of samples required to complete the word learning, the best choice for the active learner is given by the choice probability

$$
\hat{q}=\arg \max _{q} \min (P q) .
$$

This minimal probability, $\min (P \hat{q})$, gives the theoretical upper bound for the minimal number of samples $f_{W, \epsilon}(\min (P \hat{q}))$, as $P$ is not known before empirical learning, and the active leaner also needs to estimate $P$ from the sample. For a given matrix $P$, the optimal $\hat{q}$ can be computed by the iterated linear programming algorithm (See also Appendix for the detail).

As a baseline for the passive learner, we consider the average $\min (P q)$ with the uniform distribution over the vector $q$, whose lower bound is given by the Jensen's inequality

$$
\int_{q \in \mathbb{S}^{N}} \min (P q)(N-1)!\mathrm{d} q \geq \min \left(P \mathbf{1}_{N} / N\right)
$$

where the integral is taken over the $N-1$ dimensional unit simplex $q \in \mathbb{S}^{N}$. For a sufficiently small $x \ll 1$ and $y \ll 1, f_{W, \epsilon}(x) / f_{W, \epsilon}(y) \approx y / x$. Thus, the rate $R=\min (P \hat{q}) / \min \left(P \mathbf{1}_{N} / N\right)$ gives a good estimate for the rate of efficiency $R$, by which the active learning with the optimal probability $\hat{q} R$ times faster than the passive learning with a fixed probability $q$.

## Empirical evaluation

To evaluate the potential impact of the active leaning, we study the SUN database (Xiao, Hays, Ehinger, Oliva, \& Torralba, 2010) as an empirical object distributions in an collection of real-life scenes. The SUN database (retrieved on September 25 th in 2016) has $N=3,458$ objects and $M=1,111$ scenes in it. This data is supposed
to give the $N \times M$ matrix $P$ in which each column is the conditional probability of the objects given each scene. If the scene choice probability is the uniform distribution $q=\mathbf{1}_{N} / N$, the $\min (P q)$ was $8.30 \times 10^{-9}$. Meanwhile, with the optimal $\hat{q}$, the $\min (P \hat{q})$ was $1.95 \times 10^{-6}$, which implies the active learning was approximately $\min \left(P \mathbf{1}_{N} / N\right) / \min (P \hat{q})=235.3$ times faster than the baseline passive learning. The marginal probability distributions of objects for the baseline and optimal $q$ are shown in Figure 3. The difference between the two marginal distributions is visible at their tails - the tail for the uniform $q$ decreases like an exponential function, but that for the optimal $\hat{q}$ decreases as a power function (linear in the double log plot). This empirical evaluation suggests that the active learning of interest can boost the fast mapping a few orders more efficiently.


Figure 3: The marginal probability of objects for the optimal $\hat{q}$ (line) and its baseline (dots).

## Online active learning

The quantification of the efficiency of active learning is based on the optimal $\hat{q}$ with the knowledge of $P$. This gives an optimistic estimate for the active learner, as the matrix $P$ is not fully known in reality. Here we performed a Monte Carlo simulation to quantify the efficiency of an online active leaner who gradually updates knowledge in the matrix $P$ and estimates $q$ on the basis of the sample estimate of $P$. If this online active leaner is comparable with the optimal active learner with $\hat{q}$, we can treat the performance analysis on the optimal active leaner above (a few orders more efficient) as holding for the online active leaner. For this purpose, we generated a $N \times M$ matrix $P$ with $N=1000, M=100$, which has the elements in each column are Zipfian probabilities $P_{\pi(i)} \propto i^{-a}$ with the random coefficients $a \in[1,1.5]$, where $\pi:\{1, \ldots, N\} \mapsto\{1, \ldots, N\}$ is a random permutation. The online active learner has the uniform
choice probability $q_{1}=\mathbf{1}_{N} / N$. For $k^{\text {th }}$ batch of 1000 steps, the online learner samples the objects according to the probability $P q_{k}$, and constructs the sample probability matrix $\hat{P}_{k}$ according to the sample frequency. After the $k^{\text {th }}$ sampling step, the online learner estimates $q_{k}:=\arg \max _{q} \min \left(\hat{P}_{k} q\right)$. In each run of this procedure, we repeat up to $100 \times 1000$ samples, and obtain one sample for the number of required samples to finish learning all the 1000 objects. With 100 runs, we obtain the Monte Carlo estimate of the online learner shown in Figure 4. Figure 4 shows the sample probability distribution of the number of required samples in the Monte Carlo simulation (circles: histogram, line: smoothed estimate), and its comparable median estimate for the optimal learner (green vertical line) and the passive learner with the uniform $q$ (red vertical line). This simulation result shows that the online learner is as fast as the optimal learner, and is likely to be faster than the passive learner.


Figure 4: The probability distribution of the number of required samples to finish learning for passive (red), optimal (green), and online active learner (blue).

## Discussion

This study has provided mathematical analyses of quantitative aspects of word learning that provide key constraints which any theoretical account for children's word learning should satisfy. We reinspected the past theoretical claim by Blythe et al. (2010) that learning via independent fast mapping was efficient enough to account for the average number of words known by 18-year-olds. Our new analysis extends their analysis to fast mapping with non-uniform word frequency distributions, and shows that even learning via fast mapping is not efficient enough to learn words whose distribution has rarely sampled words-including the Zipf (i.e. power-law) distribution, which describes empirical word frequency distributions from natural language.

Given that this new analysis implies learning would be too slow under realistic distributions, we consider a more efficient learning scheme, in which the learner can choose preferred situations from which words are learned. This type of active control over situations or contexts seems natural with respect to general observations of children's behavior, and has been shown to benefit adult word learners (Kachergis, Yu, \& Shiffrin, 2013), but has not been subjected to theoretical analysis as far as we know. We quantify and evaluate the effect of this type of self-directed learning in word learning. As the least probable word in the distribution determines learning efficiency, we analyzed the active choice for the situations maximizing this key parameter. Analyzing an empirical dataset of the words given situations, we estimate that active learning is over two hundred times more efficient in learning time than passive learning. This result suggests that active choice in word learning can resolve the issue that naturalistic non-uniform word distributions greatly slows passive fast mapping.

Our analyses in this paper utilized one of the simplest learning schemes, fast mapping, in order to highlight the effects of varied word frequency distributions, and of active learning. However, we expect the analytic techniques we employed would also allow analysis of other learning algorithms, including many proposed variants of cross-situational learning. In future work, we will report similar analyses for learning schemes with perhaps greater cognitive plausibility. On this path towards ever more realistic assumptions about the language environment and learners' ability to shape it, we expect to make progress toward a general theoretical framework spanning many proposed word learning schemes.

## Acknowledgment

This study is supported by the JSPS KAKENHI Grant-in-Aid for Young Scientists JP 16H05860.

## References

Blom, G., Holst, L., \& Sandell, D. (1994). Problems and snapshots from the world of probability. In (p. 85-87). New York, NY: Springer-Verlag New York.
Bloom, P. (2000). How children learn the meaning of words. Cambridge, MA: MIT Press.
Blythe, R. A., Smith, A. D. M., \& Smith, K. (2016). Word learning under infinite uncertainty. Cognition, 151, 18-27.
Blythe, R. A., Smith, K., \& Smith, A. D. M. (2010, January). Learning Times for Large Lexicons Through Cross-Situational Learning. Cognitive Science, 34(4), 620-642.
Chomsky, N. (1965). Aspects of the theory of syntax. MIT Press.
Gleitman, L. (1990). The structural sources of word meaning. Language Acquisition, 1, 3-55.

Hulit, L., \& Howard, M. R. (2002). Born to talk. Toronto: Allyn and Bacon.
Kachergis, G. (2012). Learning nouns with domaingeneral associative learning mechanisms. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (p. 533-538). Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., \& Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. Topics in Cognitive Science.
MacWhinney, B., \& Snow, C. (1990). The child language data exchange system: An update. Journal of Child Language, 17(02), 457-472.
Smith, L. B. (2000). How to learn words: An associative crane. In R. Golinkoff \& K. Hirsh-Pasek (Eds.), Breaking the word learning barrier (pp. 51-80). Oxford: Oxford University Press.
Vlach, H. A., \& Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. Cognition, 127, 375-382.
Vlach, H. A., \& Sandhofer, C. M. (2012). Fast mapping across time: Memory processes support children's retention of learned words. Frontiers in Developmental Psychology, 3(46), 1-8.
Xiao, J., Hays, J., Ehinger, K., Oliva, A., \& Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In Ieee conference on computer vision and pattern recognition.
Zipf, G. (1949). Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley.

## Appendix: Iterated linear programming

For a $N \times M$ matrix $P$, write its $i^{\text {th }}$ row by $P_{i}$. Let $I=\{1,2, \ldots, N\}$ be the set of all indices. At the initial step, define

$$
K_{0}:=\emptyset, C_{0}:=\mathbb{S}^{M}, q_{0}:=e_{1}
$$

where $e_{1}:=(1,0, \ldots, 0)^{T} \in \mathbb{R}^{N}$. Then for $0<n \leq N$, define

$$
\begin{aligned}
k_{n} & :=\underset{k \in I \backslash K_{n-1}}{\arg \min } P_{k} q_{n-1}, K_{n}:=K_{n-1} \cup\left\{k_{n}\right\} \\
C_{n} & :=\left\{q \in C_{0} \mid \bigwedge_{k \in K_{n}}\left(P_{k_{n}}-P_{k}\right) q \leq 0\right\} \\
q_{n} & :=\underset{q \in C_{n}}{\arg \max } P_{k_{n}} q
\end{aligned}
$$

until $n=m$ such that $\min _{k \in K_{m}} P_{k} q_{m} \leq \min _{k \in I} P_{k} q_{m}$. The algorithm stops the iterative procedure by outputting $q:=q_{m}$.

# Reference Systems in Spatial Memory for Vertical Locations 

Thomas Hinterecker<br>Max Planck Institute for Biological Cybernetics, Tübingen<br>Caroline Leroy<br>Max Planck Institute for Biological Cybernetics, Tübingen<br>Maximilian E. Kirschhock<br>Max Planck Institute for Biological Cybernetics, Tübingen<br>Mintao Zhao<br>Max Planck Institute for Biological Cybernetics, Tübingen<br>Martin Butz<br>University of Tübingen, Computer Science, Cognitive Modeling<br>\section*{Heinrich H. Bülthoff}<br>Max Planck Institute for Biological Cybernetics, Tübingen<br>Tobias Meilinger<br>Max Planck Institute for Biological Cybernetics, Tübingen


#### Abstract

Three experiments investigated the frame of reference used in memory to represent vertical spatial layouts perceivable from a single viewpoint. We tested for the selection of three different reference systems: the body orientation, the visual vertical of the surrounding room, and the direction of gravity. Participants learned and retrieved differently colored objects on a vertical board with body and room orientations varying in relation to gravity and each other systematically. Across all three experiments participants were quicker or more accurate in memory recall when they saw the vertical spatial layout in the same orientation in relation to their body vertical as during learning, irrespective of the direction of gravity or visual room upright. These results indicate that spatial long-term memories for small-scale vertical relations are mainly defined in an egocentric reference system with respect to the body vertical despite the availability of alternative highly salient allocentric reference directions.


# Teaching by Intervention: Working Backwards, Undoing Mistakes, or Correcting Mistakes? 

Mark K Ho (mark ho @brown.edu)<br>Department of Cognitive, Linguistic and Psychological Sciences, Brown University, 190 Thayer Street Providence, RI 02912 USA<br>Michael L. Littman (mlittman@cs.brown.edu)<br>Department of Computer Science, Brown University, 115 Waterman Street<br>Providence, RI 02912 USA<br>Joseph L. Austerweil (austerweil@wisc.edu)<br>Department of Psychology, University of Wisconsin-Madison, 1202 W Johnson Street<br>Madison, WI 53706 USA


#### Abstract

When teaching, people often intentionally intervene on a learner while it is acting. For instance, a dog owner might move the dog so it eats out of the right bowl, or a coach might intervene while a tennis player is practicing to teach a skill. How do people teach by intervention? And how do these strategies interact with learning mechanisms? Here, we examine one global and two local strategies: working backwards from the end-goal of a task (backwards chaining), placing a learner in a previous state when an incorrect action was taken (undoing), or placing a learner in the state they would be in if they had taken the correct action (correcting). Depending on how the learner interprets an intervention, different teaching strategies result in better learning. We also examine how people teach by intervention in an interactive experiment and find a bias for using local strategies like undoing.


Keywords: teaching, intervention, reinforcement learning

## Introduction

When attempting to teach another agent, people have many tools at their disposal. They may choose to explain (Callanan \& Oakes, 1992), give a demonstration (Brugger, Lariviere, Mumme, \& Bushnell, 2007; Buchsbaum, Gopnik, Griffiths, \& Shafto, 2011; Király, Csibra, \& Gergely, 2013), or offer rewards and punishments for taking certain actions (Knox \& Stone, 2015; Ho, Littman, Cushman, \& Austerweil, 2015). Another way in which people teach a learner is by intervening on the learner or the learner's environment. For example, if a puppy urinates on the carpet when a person is trying to teach the puppy to urinate on a pad, a person might move the puppy to the pad or move the pad to the puppy. When teaching another person a skill like tennis, a teacher might intervene on the trainee mid-movement and either adjust their arm to match the target movement or stop them to start over. The space of possible ways in which a teacher could change a learner's situation for pedagogical purposes is large. This raises several questions: First, what is the effectiveness of different intervention strategies? Second, how could learners interpret interventions and how does the interpretation affect a teaching strategy's efficacy? And, finally, what teaching strategies do people tend to use?

In this work, we examine three teaching by intervention strategies from a reinforcement learning perspective (Sutton
\& Barto, 1998). The first, backward chaining, is motivated by algorithms such as value iteration (Bellman, 1957) that solve multi-stage decision-problems by propagating information about rewards to previous states that lead to those rewards. Intuitively, this is akin to teaching a task by "working backwards", first ensuring that the learner knows how to reach a goal from the penultimate state, and then reach the penultimate state from the antepenultimate state, and so on. We consider this a global intervention strategy since it involves changing the learner's state in a manner that reflects the structure of the entire task, rather than a small part of it. The second strategy, undoing, is motivated by the intuition that interventions prevent learners from executing an undesirable action by having them restart from the state they performed the undesirable action. The third strategy, correcting, intervenes on a learner when she executes an undesirable action (like undoing), but places her in the state she would have gone to if she had taken the desired action. Unlike backwards chaining, undoing and correcting involve local changes to an agent's state.

How could a learner interpret an intervention? In a typical reinforcement learning setting, an agent takes an action in a state, and then the environment rewards or punishes her and moves her to a new state (Figure 1). We formalize four ways that an intervention can be interpreted. First, the intervention may simply reset the learner in a new location from which the next action will be taken. Second, the next state that the learner is moved to could be interpreted as part of a transition in the environment. Third, the intervention could be treated as an interruption in a learner's stream of behavior such that the undesirable action just taken never happened. Fourth, the intervention could be treated as a disruption, in which the intervention is experienced negatively. Each of these accounts may interact with a teacher's training strategy in different ways, meaning that the best teaching strategy may be dependent on the learner's intervention interpretation.

The outline of the paper is as follows. First, we review the reinforcement learning framework. Second, we formalize four different ways that a reinforcement learning algorithm


Figure 1: (a) Standard state, action, reward, next state sequence of a Markov Decision Process at a given time step. (b) Four different interpretations of a teacher intervening to place the learner in state $v_{t}$ in response to a learner's action $a_{t}$ from state $s_{t}$. When interventions are interpreted as reset, transition, or disrupt, $r_{t}$ is respectively determined by the environmental next state, $s_{t}^{\prime}$, the teacher's next state, $s_{t+1}$, or the teacher's intervention, $v_{t}$. When the the intervention is treated as interrupt, no reward experienced and no learning occurs for that time step.
could interpret an intervention and three teaching strategies. Third, we conduct simulations to examine how efficacious different teaching strategies are depending on how a learner interprets their interventions. Fourth, we conduct an experiment to investigate how people teach by intervention. We find that undoing, a local intervention strategy, is often effective and that people tend to teach most often by undoing, occasionally correcting, and rarely backward chaining.

## Computational Modeling

In this section we present the standard reinforcement learning (RL) formalism, discuss the four intervention interpretations, and define the three teaching strategies.
Reinforcement Learning RL describes how an agent interacts with an environment and learns reward-maximizing behaviors (Sutton \& Barto, 1998). Formally, an RL algorithm learns to take actions in a Markov Decision Process (MDP), defined by the tuple $\langle S, A, T, R, \gamma\rangle$ : a set of states in the world $S$; a set of actions for each state $\mathcal{A}(s)$; a transition function that maps state-action pairs to a probability distribution over next states, $P\left(s^{\prime} \mid s, a\right)$; a reward function that maps states to scalar rewards, $R: S \rightarrow \mathbb{R}$; and a discount factor $\gamma \in(0,1]$.

At each time step $t$, an RL agent takes an action $a_{t}$ from a state $s_{t}$, which results in moving to next state $s_{t+1}$ and a reward $r_{t+1}=R\left(s_{t+1}\right)$ (Figure 1). Actions are determined by the agent's policy $\pi$ that maps states to distributions over actions. For a policy $\pi$, the value at each state, $V^{\pi}(s)$, is:

$$
\begin{equation*}
V^{\pi}(s)=E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \mid s_{t}=s\right] . \tag{1}
\end{equation*}
$$

The optimal policy, $\pi^{*}$, is one that maximizes the value function in every state, $V^{*}(s)=\max _{\pi} V^{\pi}(s), \forall s \in S$. An agent uses state, action, next state, reward tuples to learn an optimal policy.

Q-Learning One algorithm for learning an optimal policy is Q-learning, which is an off-policy temporal difference control algorithm. Under mild assumptions, Q-learning converges to the true action-value function (Watkins \& Dayan, 1992). Moreover, humans and animals both engage in the type of error-driven reward learning found in Q-learning, making it a useful model with which to test different human teaching strategies (Niv, 2009). We use one form of this algorithm, one-step Q-learning, which is defined by the following update rule given a tuple $\left(s, a, s^{\prime}, r\right)$ :

$$
\begin{equation*}
Q(s, a) \leftarrow Q(s, a)+\alpha\left[r+\gamma \max _{a^{\prime}} Q\left(s^{\prime}, a^{\prime}\right)-Q(s, a)\right] \tag{2}
\end{equation*}
$$

where $\alpha$ is the learning rate. We convert the estimated actionvalue function to a policy using the softmax decision-rule $\pi(a \mid s)=\exp \left\{Q(s, a) / \lambda_{Q}\right\} / \sum_{a^{\prime}} \exp \left\{Q\left(s, a^{\prime}\right) / \lambda_{Q}\right\}$, where $\lambda_{Q}$ is a temperature parameter controlling the probability that an agent takes the action estimated to yield the largest reward depending on the relative rewards she could get by taking other actions.

## Teaching by Intervention

Interpreting Interventions The standard RL formulation does not define how interventions should be interpreted. Thus, we posit four different possible interpretations here, depicted in Figure 1. The four interpretations are motivated by formalizing the following two intuitions in different ways. First, a teacher could be treated as a part of the environment such that her intervention directly changes the next state of the learner (possibly stopping the feedback she would have received had she gone to the next state had the intervention not happened). Second, a teacher is distinct from the standard MDP environment, and intervenes as a direct response to a learner having taken an action and ended up in a next state.

Formally, at a time step $t$, the learner in state $s_{t}$ takes an action $a_{t}$ and ends up in new state $s_{t}^{\prime}$. If the teacher does not intervene, $s_{t+1}=s_{t}^{\prime}$. Otherwise, a teacher intervenes to place the learner in state $v_{t} \in S$. For all intervention types, $s_{t+1}=v_{t}$. However, if the teacher's intervention is interpreted as a reset, then the learner performs a Q-learning update using the tuple $\left(s_{t}, a_{t}, s_{t}^{\prime}, R\left(s_{t}^{\prime}\right)\right)$, meaning that she still receives the reward she would have gotten had she reached $s_{t}^{\prime}$ as the next state. If it is interpreted as a transition, then the learner updates with $\left(s_{t}, a_{t}, v_{t}, R\left(v_{t}\right)\right)$, meaning that she gets the reward had she taken the action that would move her from $s_{t}$ to $v_{t}$. If it is an interruption, then the learner does not update the state-action value function the state-action pair that was intervened on, and she takes her next action in $s_{t+1}=v_{t}$. If it is interpreted as a disruption, then the learner updates with $\left(s_{t}, a_{t}, v_{t},-1\right)$.
Teaching Strategies We discuss three teaching strategies: backward chaining, undoing, and correcting. A teacher using backward chaining has an $n$-length trajectory $J=<$ $\left(s_{1}, a_{1}\right), \ldots,\left(s_{n}, a_{n}\right)>$ that she uses to teach the learner. We denote the states in the trajectory as $S_{J}=\left\{s_{i}: i=1,2,3 \ldots, n\right\}$. The teacher also has a utility function over different interventions, where initially $U_{0}\left(s_{i}\right)=i$ for $i=1,2,3, \ldots, n$ and $U_{0}(s)=-\infty$ for $s \in S \backslash S_{J}$. On each time step, the teacher's utility function is updated as:

$$
U_{t+1}\left(s_{t}\right)=\left\{\begin{array}{l}
U_{t}\left(s_{t}\right)-1 \text { if }\left(s_{t}, a_{t}\right) \in J  \tag{3}\\
U_{t}\left(s_{t}\right) \text { otherwise }
\end{array}\right.
$$

Teachers only intervene when the agent performs an action inconsistent with the trajectory (i.e. $\left(s_{t}, a_{t}\right) \notin J$ ) and place the agent in a next state according to a softmax decision rule over their utilities: $P(v) \propto \exp \left\{U_{t}(v) / \lambda\right\}$, where $\lambda$ is a temperature parameter. The backward chaining teacher is initially more likely to move the agent closer to the end of a target trajectory, but as the agent shows they can perform the target action in a state the utility of moving the agent to that state decreases. Meanwhile, the relative utility of placing the agent in a slightly earlier stage in the trajectory increases.

A teacher using an undoing strategy has a target policy $\pi^{*}$ : $S \rightarrow A$ that it is attempting to teach. On each time step, if an agent's action $a_{t} \neq \pi^{*}\left(s_{t}\right)$, then $v_{t}=s_{t}$. That is, when an agent takes an incorrect action, that action is undone by the teacher and the agent is placed back in the state she took the incorrect action. A teacher using a correcting strategy also has a target policy $\pi^{*}$ that it is attempting to teach. However, if an agent's action $a_{t} \neq \pi^{*}\left(s_{t}\right)$, then $v_{t}=\arg \max _{s} T\left(s \mid s_{t}, \pi^{*}\left(s_{t}\right)\right)$. That is, the teacher will move the agent to the state it would have been in had the agent taken the target action.

## Simulations

To understand the interaction of teaching strategy and learner interpretation, we simulated the performance of a RL agent for each combination in a gridworld task.

Teacher's Reward Function

| $*$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| -10 |  |  |  |  |  | $*$ |
| +10 |  |  |  |  |  |  |
|  | -1 | -1 |  | -1 | -1 |  |
|  | -1 | -1 |  | -1 | -1 |  |
|  |  |  | Start |  |  |  |

Learner's Reward Function


Experiment Interface


Figure 2: Task used for simulations and experiment. Asterisks (*) indicate absorbing states, both providing reward to the learner, whereas the teacher received reward if the learner entered the right door, but was punished if the learner entered the left door. The teacher received a mild punishment whenever the learner entered a garden tile.

## Task

The task we used is shown in Figure 2. It consists of a $7 \times$ 4 gridworld where the learning agent always starts a round in the center tile of the first row. At any given location, a subset of the four cardinal directions is available to the learning agent (e.g. at the bottom edge, "down" is not available as an action). On each episode, the learning agent starts in the bottom-middle tile and the upper-right and upper-left corners of the gridworld are absorbing states.

In our task, the teacher and learner have different rewards for the learner's actions in the MDP. In particular, the two absorbing states (goals) both have $\mathrm{a}+10$ reward for the learner, but for the teacher, only one has +10 while the other has -10 . Additionally, there are several non-absorbing tiles that give the teacher -1 if the learner enters them. These features of the task are visualized in Figure 2.

All simulations used a Q-learning agent with a tabular representation of states $\left(Q_{0}(s, a)=0 \forall s, a, \alpha=.9\right.$, and $\left.\gamma=.95\right)$. Each simulated teacher interacted with the learner for 12


Figure 3: Simulated backward chaining, undoing, and correcting results or different intervention interpretations and intervention probabilities (top three rows). Results of learners trained using participant responses on task (bottom). Total teacher payoff is the net reward of the learner's behavior based on the teacher's reward function during the evaluation phase of each episode.
episodes. Each episode was divided into two phases: a teaching phase and an evaluation phase. During the teaching phase, the simulated teacher interacted with the Q-learner, which selected actions using a softmax rule ( $\lambda_{Q}=.1$ ) and engaged in learning. During the evaluation phase, the learner performed the task without teacher interaction or learning and used a greedy policy. Additionally, the performance was measured with respect to the teacher's payoffs based on her reward function. Each episode phase ended after 25 time steps.

## Teaching Strategies and Interpretations

We tested all combinations of teaching strategy and intervention interpretation (\{backwards chaining, undoing, correcting $\} \times\{$ resetting, transitioning, interrupting, disrupting $\})$. In natural situations, it is not likely that teachers intervene every time a learner takes an incorrect action. Thus, we tested the performance of the models given different probabilities of intervening given that the learner performed an incorrect action: $0.25,0.50,0.75,1.0$. This allowed us to evaluate the robustness of different teaching method and intervention interpretation combinations when feedback is imperfect. Each combination of teaching strategy, intervention type, and intervention probability were simulated 1000 times and teaching performance was based on the evaluation phase.

## Results and Discussion

Simulation results are plotted in Figure 3. When interaction probability is high, undoing is most effective. This is be-
cause interventions act as impassable obstacles to the learning agent, which, combined with a discount rate, makes taking incorrect actions less beneficial than alternative actions that change the state and lead to reward. However, an exception is when the learner interprets interventions as disrupting, where the average performance of the undoing teaching strategy decreases quickly as intervention probability drops. This is because the teacher is less likely to serve as an obstacle, which makes it less likely that the agent will learn that incorrect actions are less efficacious. Across all interpretations, undoing outperforms correcting because undoing implicitly teaches the learner that the garden tiles are negative, whereas, correcting does not. Undoing also leads to more learning experience because correcting allows the agent to progress on the task without actually taking target actions.

When the probability of intervention is high ( $1.0-0.75$ ), the backward chaining strategy performs as well as or worse than the undoing strategy. Unlike undoing, a global strategy like backward chaining's efficacy is robust to less frequent interventions. This is because these interventions ensure that the learning agent has mastered a subset of states and acquired an accurate value representation as opposed to acting as a constraint on transitions in the environment.

The different intervention types also interacted with the teaching strategies in important ways. First, undoing shows identical patterns regardless of whether the intervention type is resetting, transitioning, or interrupting. When it is disrupting, learners reach maximum performance even more


Figure 4: Experimental results. (a) Boxplot of proportion of correcting, undoing, and other interventions performed by individual participants. For many participants, the majority of their interventions were to undo the learner's action. (b) Graphical visualization of teacher-learner interaction during an episode ( $\varepsilon=0.8$ ) illustrating local interventions. Yellow numbers indicate order of interventions. (c) Graphical visualization of participant interventions for actions taken from the same state. For each episode, each participant has one pseudo-count that is divided among all of their interventions in that episode. The number in each tile represents the sum of these pseudo-counts over participants. The intervention probability is the proportion of times that action was subsequently intervened upon.
quickly. Second, for the backward chaining strategy, all strategies but transitioning led to learners acquiring policies that approached the target behavior. This is likely because the transitioning interpretation results in learners using the teacher's interventions as a way to "teleport" to a desirable location on the grid and not properly learn the task.

## Experiment

How do people teach using interventions? Do they use a global strategy like backwards chaining or a local one like undoing or correcting? Our simulations suggest that undoing is the best teaching strategy if teachers intervene when the learner makes a mistake with high probability. However, backwards chaining works better when the teacher intervenes infrequently. Alternatively, it seems intuitive to intervene such that the learner is shown the correct state she should have gone to, and human teachers might use this strategy despite its sub-optimality with Q-learners. To explore these possibilities, we had human teachers interact with agents that were pre-programmed to improve over time. This gave us the opportunity to view how people would teach by intervention independent of the learning mechanism.

## Experimental Design

Participants and materials Thirty-five MTurk participants took a dog training study that used the interface shown in Figure 2. On each trial, the dog would start at a tile and then walk to an adjacent tile. If the participant did not click on the dog at any point during its movement or within 1 s of the dog entering the next tile, the next trial would start. If the participant clicked on the dog, then the dog "paused" and they could drag it to any tile on the gridworld and drop it. The dog then "unpaused" and the subsequent trial would then start at that tile. When the dog reached either "dog bowl," an animated dog treat would appear to indicate that the dog had experienced a reward. Entering either dog bowl tile ended an episode.
Procedure Before the main task, participants completed training trials that taught them how to intervene on the dog's behavior by picking it up. For the main experiment, they were told that they were trying to train a dog to perform a task on its own. The task was for the dog to only go to its own dog bowl, located in the upper-right tile, while avoiding their neighbor's dog bowl, located in the upper-left tile, and also avoiding the two lawns. Thus, the participants' goal in the task maps onto the teacher reward function shown in Figure 2. They had 12
"days" (i.e. episodes) in which they could train the dog, and they were told that each day ended once the dog became tired after 25 steps or became satiated by eating a dog treat. Each trial, the dog was programmed to execute the target policy with a probability of $1-\varepsilon$ and a random action otherwise. $\varepsilon$ started at 1.0 for the first episode and then decreased by 0.1 each subsequent episode until $\varepsilon=0.0$. This gave the impression that the dog was improving over time regardless of the intervention strategy used.

After the task was completed, participants were asked to answer several questions regarding their strategy, how well the dog responded, task difficulty, expected training efficacy, expected efficacy with a real dog, dog ownership, dog training experience, and several demographic questions.

## Results

Intervening People make relatively sparse, local interventions that match the undoing model. Participants intervened on learners' behavior more when the learner performed a nontarget action than when they performed a target action (nontarget: $\mathrm{M}=0.66$, S.D. $=0.22$; target: $\mathrm{M}=0.06$, S.D. $=0.10$; paired t-test: $t(34)=13.77, p<.001)$. Additionally, the proportion of non-target actions that were intervened upon was between 0.5 and 0.75 , the regime where backward chaining and undoing perform comparably. Interventions were also fairly local and close to the final state that resulted in the learner's action (Average Manhattan Distance between next state and intervention: $\mathrm{M}=1.64$, S.D. $=0.49$ ). This indicates that backwards chaining was not often used as a strategy since that strategy requires making more global interventions. Finally, as Figure 4a reveals, many participants performed undoing interventions in which an agent that took a non-target action was placed back into its original position (Correcting: $\mathrm{M}=0.15$, S.D. $=0.14$; Undoing: $\mathrm{M}=0.59$, S.D. $=0.24$; Other: $\mathrm{M}=0.27$, S.D. $\left.=0.19 ; \chi^{2}(2)=335.89, p<.001\right)$.

Teaching Q-learners To compare human and model strategies, we used participants' responses to train Q -learners in the same task. We approximated how participants would have taught real learners by sampling from their responses to a learner's action in the task whenever a simulated learner took the same action. If a particular participant never observed an agent's take a simulated action, the default response was to not intervene. These results are plotted in Figure 3 for comparison with the simulation results.

## Discussion

Our simulations revealed important interactions among teaching strategy, intervention interpretation, and intervention probability. In particular, undoing, which involves local changes to an agent's state, is an especially effective strategy only when interventions are frequent, while backward chaining, which involves state-changes reflecting the global structure of the task, is moderately effective regardless of intervention frequency. Incidentally, when people teach by intervention, they typically engage in undoing, but they do it
less often than they should to train Q-learners (66\%). Generally, people make moderately frequent local interventions.

As this is a preliminary investigation into teaching by intervention, this work has limitations. We use Q-learning as the learner, but other RL algorithms may respond better to human interventions. And given previous work showing that people often teach with communicative intent (Shafto, Goodman, \& Griffiths, 2014; Ho et al., 2015), it may be that the standard RL framework is inadequate for capturing peoples' relatively sparse, local interventions. Future work will also need to test a wider range of MDP tasks. Nonetheless, these simulations and models are a first step towards understanding the everyday phenomenon of teaching by intervention in humans.

## Acknowledgments

MKH is supported by the NSF GRFP (DGE-1058262).

## References

Bellman, R. (1957). Dynamic programming. Princeton, NJ: Princeton University Press.
Brugger, A., Lariviere, L. A., Mumme, D. L., \& Bushnell, E. W. (2007). Doing the right thing: Infants' selection of actions to imitate from observed event sequences. Child Development, 78(3), 802-824.
Buchsbaum, D., Gopnik, A., Griffiths, T. L., \& Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. Cognition, 120(3), 331-340.
Callanan, M. A., \& Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. Cognitive Development, 7(2), 213-233.
Ho, M. K., Littman, M. L., Cushman, F., \& Austerweil, J. L. (2015). Teaching with rewards and punishments: Reinforcement or Communication. In D. C. Noelle et al. (Eds.), Proceedings of the 37th annual meeting of the cognitive science society (pp. 920-925). Austin, TX: Conference Science Society.
Király, I., Csibra, G., \& Gergely, G. (2013). Beyond rational imitation: Learning arbitrary means actions from communicative demonstrations. Journal of Experimental Child Psychology, 116(2), 471-486.
Knox, W. B., \& Stone, P. (2015). Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. Artificial Intelligence, 225, 24-50.
Niv, Y. (2009). Reinforcement learning in the brain. Journal of Mathematical Psychology, 53(3), 139-154.
Shafto, P., Goodman, N. D., \& Griffiths, T. L. (2014, June). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive Psychology, 71, 55-89.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction. MIT press.
Watkins, C. J., \& Dayan, P. (1992). Q-learning. Machine learning, 8(3-4), 279-292.

# Computational and behavioral investigations of the SOB-CS removal mechanism in working memory 

Violette Hoareau ${ }^{1}$ (violette.hoareau@univ-grenoble-alpes.fr), Sophie Portrat ${ }^{1,2}$ (sophie.portrat@univ-grenoble-alpes.fr), Klaus Oberauer ${ }^{3}$ (k.oberauer@ psychologie.uzh.ch), Benoît Lemaire ${ }^{1,2}$ (benoit.lemaire@univ-grenoble-alpes.fr), Gaën Plancher ${ }^{4}$ (Gaen.Plancher@univ-lyon2.fr) and Stephan Lewandowsky ${ }^{5}$ (Stephan.Lewandowsky@bristol.ac.uk)<br>${ }^{1}$ LPNC, University Grenoble Alpes, F-38040 Grenoble, France<br>${ }^{2}$ CNRS, LPNC UMR 5105, F-38040 Grenoble, France<br>${ }^{3}$ University of Zurich, Zurich, Switzerland<br>${ }^{4}$ Université Lumière Lyon 2, Lyon, France<br>${ }^{5}$ University of Bristol and University of Western Australia, Bristol, UK


#### Abstract

SOB-CS is an interference-based computational model of working memory that explains findings from simple and complex span experiments. According to the model's mechanism of interference by superposition, high similarity between memory items and subsequently processed distractors is beneficial because the more a distractor is similar to an item, the more they share similar units, leading to less distortion of the memory item. When time allows, SOB-CS removes interfering distractors from memory by unbinding them from their context. The combination of these two mechanisms leads to the prediction that when free time is long enough to remove the distractors entirely, similarity between items and distractors should no longer be beneficial to memory performance. The aim of the present study was to test this prediction. Adult participants performed a complex-span task in which the free time following each distractor and the similarity between items and distractors were varied. As predicted by the model, we observed a positive effect of the similarity between items and distractors, and a negative effect of pace on the mean working memory performance. However, we did not observe the predicted interaction. An analysis of the errors produced during recall showed that longer free time reduced the tendency of distractors to intrude in recall much less than the model predicted. The SOB-CS model accounted well for the data after a substantial reduction of the removal-rate parameter.


Keywords: Working Memory, SOB-CS model; interference by superposition; removal mechanism

## Introduction

Working memory (WM) is the system responsible for holding information available for ongoing cognition (Baddeley \& Hitch, 1974; Miyake \& Shah, 1999). It is often tested with complex-span tasks (Barrouillet, Bernardin, \& Camos, 2004), which combine an immediate memory test with a concurrent processing demand: some items (e.g., letters or words) are provided one at a time for subsequent recall in order and several distractors are also presented in-between items. The concurrent processing of distractors impairs memory, compared to a simple-span task that consists only of the immediate memory test. There has been ongoing debate about the reasons why distractors affect WM performance. The present study contributes to this question by testing a prediction of the interference-based connectionist model SOB-CS (Oberauer, Lewandowsky, Farrell, Jarrold, \& Greaves, 2012).

According to SOB-CS, forgetting is due to interference between items and distractors. The model is based on a two-
layer connectionist network that associates a distributed item representations with distributed position markers, for instance associating the first item of the sequence with position 1 , through Hebbian learning (Anderson, 1995). Each association is registered in a two-dimensional weight matrix coding for the position and the item representations. During each processing step, SOB-CS assumes that distractors are encoded in the same way as items and associated to the position of the preceding item. In other words, SOB-CS suggests that items and distractors are superimposed in the same weight matrix, leading to a distortion of items by distractors which in turn causes forgetting. In this way, the model is able to reproduce interference between items and distractors according to their similarity: the more a distractor is similar to an item, the more feature values they share, leading to less distortion of the memory item. Therefore, this model predicts that high similarity between an item and the following distractor is beneficial to WM performance.

Oberauer, Farrell, Jarrold, Pasiecznik, and Greaves (2012) reviewed studies that have investigated item-distractor (I-D) similarity effects. They showed that phonological similarity between items and distractors is beneficial if the material was pronounceable non-words (non-words were used in order to ensure that participants do not encode stimuli by their meanings). Oberauer et al. did four experiments in which participants had to remember a list of four non-words items. The two distractors intervening after each item were also nonwords and had to be read aloud. The phonological similarity between the items and the distractors was manipulated. In the first three experiments, distractors were similar to the preceding item whereas in the fourth experiment distractors were similar to the following item. The findings of experiments 1,2 and 3 showed a positive effect of phonological similarity between items and distractors which meshes well with the mechanism of interference by superposition implemented in SOB-CS. Moreover, experiment 4 confirmed the hypothesis of SOB-CS suggesting that distractors are associated to the preceding item and not to the following item: no beneficial effect of I-D similarity was observed when similar distractors preceded, rather than followed, the items to which they were similar.

The SOB-CS model also proposes an explanation for the cognitive load effect (Barrouillet, Portrat, \& Camos, 2011). This effect has been observed in several studies showing that WM performance depends on the proportion of time during which distractors capture attention (Barrouillet, Bernardin, Portrat, Vergauwe, \& Camos, 2007; Barrouillet et al., 2004, 2011). According to decay-based theories (Baddeley, 1986; Towse \& Hitch, 1995; Barrouillet et al., 2011), the cognitiveload effect can be explained as follows: forgetting is mostly due to the time-based decay of the memory traces when distractors are processed. In order to avoid forgetting, memory traces can be reactivated when free time is available between distractors. A sole interference mechanism cannot account for such a positive impact of free time on WM performance. Hence, in SOB-CS, the cognitive load effect is explained as follows: forgetting is counteracted by a removal mechanism in such a way that each distractor that has just been encoded is "removed" during free time. The removal process consists of an unbinding, by Hebbian anti-learning, of the association between each distractor and the study context, thereby rendering the context signal more effective as a retrieval cue for the memoranda. SOB-CS suggests that the strength of removal exponentially depends on the time devoted to it. This mechanism leads to the prediction that the more free time elapses after each distractor, the more time is available for removing the interfering distractor that just has been encoded and hence the better the WM performance.

To sum up, two mechanisms are important in SOB-CS to specifiy the effect of distractors on WM performance. First, according to the mechanism of interference by superposition, distractors which are similar to the preceding item should distort that item less than dissimilar distractors, leading to better performance at recall. Second, the mechanism of "removal" leads to the prediction that distractors are unbound from WM during free time in order to clear memory from irrelevant information. The combination of these two mechanisms gives rise to an interesting hypothesis which is at the heart of the present paper: if free time is long enough to entirely unbind an irrelevant distractor from WM, there is no reason to observe an effect of I-D similarity as distractors would not be present in WM anymore.

## Overview of the experiment

The aim of our experiment was to test this prediction of SOB-CS concerning both the removal and the interference by superposition mechanisms. To do that, our experiment replicates and extends the second experiment presented in Oberauer, Farrell, et al. (2012). It consisted of a verbal complex span task in which items and distractors were pronounceable non-words (i.e. words without semantic meanings). The length of the memory list was constant and set to four items. In the high similarity condition, the similar distractor always immediately followed the items to which they were similar and all the items were dissimilar to each other. Then, as memoranda were non-words, serial recall was done by reconstruc-
tion among a candidate set containing the four list items, four of the distractors and four not presented lures (NPLs). The NPLs were non-words which had never been seen by participants in the current experiment.

In their experiment, Oberauer, Farrell, et al. (2012) only manipulated the similarity between items and distractors. We extended that experiment by adding the manipulation of the pace of the distractor presentation to vary the free time available to remove distractors. We tested participants and the SOB-CS model with three paces (fast, medium, slow). To allow comparisons of our findings with Oberauer, Farrell, et al. (2012), the faster pace of our experiment was the pace used by Oberauer, Farrell, et al. (2012). This extension allowed to test the special prediction that the positive effect of I-D similarity decreases as the free time increases.

## Simulation 1 with SOB-CS

To test the prediction that the positive effect of I-D similarity decreases as the free time increases, we reused and adapted the simulation presented in Oberauer, Farrell, et al. (2012).

## Method

The creation of the stimuli was done similarly as in Oberauer, Farrell, et al. (2012): stimuli were generated and organized in 8 dissimilar sets for each trial, each set containing 10 similar stimuli. From these sets, items, distractors and non-presented lures (NPLs) were selected according to the condition of similarity (i.e. high vs low). The recall candidate sets were composed of the four items of the trial, four distractors and four NPLs. The NPLs were added in order to balance the global attractiveness of the candidate sets between similarity conditions. In the high similarity condition, a distractor is attractive for two reasons: it has been processed and it was similar to the items. In the low similarity condition, distractors are therefore less attractive. In contrast to the distractors, NPLs were dissimilar to the items in the high similarity condition, whereas they were similar to the items in the low similarity condition.

In all the simulations presented in Oberauer, Farrell, et al. (2012), which reproduced well the behavioral data, the encoding duration of the distractor was set to 1000 ms . As the pace of the processing task of all their experiments is 1000 ms , no free time was available to remove the distractor. This means that they did not use the removal mechanism in their simulations. In order to replicate the results in this baseline condition (which is the condition with the faster pace in our experiment), we also set the encoding duration of the distractor to 1000 ms . ${ }^{1}$ For the moderate ( 1800 ms ) or slow (2600 ms ) paces, the removal mechanism was used because there were 800 ms or 1600 ms available free time.

[^101]We ran 1000 simulated subjects, each one completing five trials in each condition as in the experiment. Oberauer, Farrell, et al. (2012) used the SOB-CS default parameter values except for the distinctiveness parameter c , which they lowered from 1.3 to 0.45 to approximately move the overall accuracy into the range of data. This new value of parameter c was justified because non-words are less distinctive than well-known words. In our simulation, we did as Oberauer, Farrell, et al. (2012) except that we lowered the c parameter even more from 0.45 to 0.3 . The reason is that, in our experiment, we ensured that each non-word, as an item, a distractor or a NPL, was seen only once by a participant. In contrast, in Oberauer, Farrell, et al. (2012), 100 trials were performed using a set of 36 non-words only. One trial required 16 items, which is almost half the set of non-words. This means that each participant saw each item more than 40 times during the test, which would make them familiar with the non-words as they go along the test. This difference could make their task easier. This is why, in the simulation, it was justified to set the distinctiveness parameter c to 0.3 instead of 0.45 .

## Results: Simulated data

Correct responses Recall responses were scored as correct when a correct item was chosen in its exact serial position. Figure 1 (panel $\mathrm{B}^{2}$ ) presented the percentage of responses correctly recalled by the model as a function of pace and similarity. As expected, the simulation shows an effect of pace ( $0.38,0.75$, and 0.77 at fast, moderate and slow pace respectively) and an interaction effect according to which the beneficial effect of similarity disappears as the pace slows down (i.e. as free time increases). In fact, we can see that at a fast pace the percentage of correct recall is higher when distractors are similar (compared to dissimilar), to the preceding item ( 0.43 vs. 0.33 ) whereas at moderate and slow paces, the difference between similarity conditions is null ( 0.75 vs. 0.75 and 0.76 vs. 0.78 for moderate and slow paces respectively).

We also analyzed three different kinds of errors. An error could be an intrusion of distractor, an intrusion of NPL or a transposition error (an item from the list in a wrong position).
Distractor intrusions Figure 2 (panel B) presents the proportion of distractor intrusions. First, the simulation showed a strong effect of pace: around $20 \%$ of the responses at fast pace contained distractor intrusions whereas distractor intrusions are negligible (less than $2 \%$ ) at moderate and slow pace. It appears that distractors are sufficiently removed after 800 ms , for not being recalled. No effect of similarity and no interaction were observed.

NPL intrusions Figure 3 (panel B) presents the proportion of NPLs intrusions. Even if NPLs are not encoded into WM, the NPLs can be recalled as they can be confused with the memoranda. The more the WM is distorted by distractors,

[^102]the more we should observe confusion errors at recall. We observed that NPLs intrusion decreased when the free time increased as WM is less distorted. We also observed an effect of similarity: there are more NPLs intrusion in the lowsimilarity condition as the NPLs are similar to the items in this condition. This effect is much stronger at fast pace (when distractors are not removed) than at moderate and slow pace. In fact, we observed that the differences of intrusion rates between low-similarity and high-similarity are $0.09,0.02$ and 0.02 for the fast, moderate and slow pace respectively.

Transposition errors Finally, Figure 4 shows the proportion of transposition errors (order errors) for which a small pace effect was observed. At the fast pace, the proportion of transposition errors was increased by $8 \%$ as compared to the slow and moderate pace. No effect of similarity and no interaction were observed.
Summary of the simulation results In summary, the SOBCS model with its standard parameters (except the c parameter) predicts a beneficial effect of I-D similarity, which is present only when there is no removal of the distractors (i.e. in the fast pace condition). As soon as there is free time (800 ms or 1600 ms ), the similarity effect disappears. The analysis of the different kind of errors show that as soon as there is free time, distractor intrusions is negligible. This finding can explain why we do not observe the similarity effect at a moderate and slow pace: distractors are totally removed according to the SOB-CS model. These predictions will now be compared with human data.

## Experiment

## Method

Participants Participants were 34 students from the University of Bristol. They participated voluntarily in 1-hr session in exchange for course credit. Each participant performed the 6 conditions: three different paces (slow, moderate, fast) $\times$ two similarity conditions (low and high).
Material Participants were presented with four non-words (e.g "zaff") for memorization, each followed by a pair of nonword distractors. The memoranda were presented in red and the distractors in black. Participants were asked to read aloud all the non-words as soon as they appeared but to only memorize the red ones in serial order.

Items and distractors were sampled from a set of nonwords selected from the ARC Nonword Database (Rastle, Harrington, \& Coltheart, 2002). A database of 720 non-words was used to ensure that participants never saw a non-word more than once. Each non-word was pronounceable, composed of one syllable and four letters. The 720 non-words were organized in 240 rhyming groups, each containing three non-words (e.g., "baff, daff, haff" was a rhyming group).

The candidate set for recall was constructed such that its similarity structure was the same for both conditions (low and high similarity). Whatever the similarity condition, partici-


Figure 1: Proportion of correct responses. Error bars are 95\% confident intervals for within-subject comparisons.


Figure 2: Proportion of distractor intrusions. Error bars are $95 \%$ confident intervals for within-subject comparisons.
pants saw four items, four stimuli similar to each item and four stimuli dissimilar to the items. In the high-similarity condition, the four stimuli similar to each item were the distractors whereas in the low conditions they were notpresented lures (NPLs).

In the high similarity condition, one stimulus from four different rhyming groups was chosen at random to be an item, and the other two stimuli of each rhyming group were used as the pair of distractors that immediately followed that item, such that each pair of distractor was similar to their preceding item. The NPLs, for the recall set candidates, were chosen at random from 4 other rhyming groups, such that NPLs did not rhyme with any item or distractors.

In the low similarity condition, four groups were used to create the list of items and NPLs, such that each NPL was similar to an item. Two stimuli from each of four other rhyming groups were chosen at random to serve as distractors. In this way, no pair of distractors rhymed with any item on the low-similarity condition.

For all conditions, we ensured that the four to-bemaintained items were dissimilar to each other.
Procedure A MATLAB program using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) coded by Oberauer and collaborators (2011) was reused with some modifications to display stimuli and record responses. Each trial started with a black centered fixation cross presented during $1,500 \mathrm{~ms}$, followed by a computer-paced presentation of items and distractors. Items always appeared during 800 ms followed by a


Figure 3: Proportion of NPLs intrusions. Error bars are 95\% confident intervals for within-subject comparisons.


Figure 4: Proportion of order errors. Error bars are $95 \%$ confident intervals for within-subject comparisons.

400 ms blank. Distractors appeared at the rate of one stimulus per $1,000 \mathrm{~ms}$ ( 800 ms on, 200 ms off) in the fast condition, $1,800 \mathrm{~ms}$ ( 800 ms on, 1000 ms off) in the moderate condition and $2,600 \mathrm{~ms}$ ( 800 ms on, $1,800 \mathrm{~ms}$ off) in the slow condition. After the last distractor, 12 recall candidates simultaneously appeared on the screen in blue, each in a blue frame. They were displayed at random in a $3 \times 4$ matrix. Participants selected recall choices by clicking inside the items' boxes in the order in which they were presented. A sound indicated that a response has been recorded. They were asked to guess if they could not remember an item. Each participant completed 30 trials, 5 in each condition, in a random order. They were also prompted to take self-paced breaks every six trials. In addition, there were four practice trials at the start with four different conditions: similar/fast, dissimilar/moderate, similar/slow and dissimilar/fast.

## Results: human data and comparisons with simulated data

We ran a two-way ANOVA on the mean proportion of correct responses and different types of errors over trials with similarity (high and low I-D similarity) and pace (slow, moderate and fast) as within-subjects factors.

Correct responses Figure 1 (panel A) shows that, as predicted by the model, there was an effect of similarity $\left[F(1,33)=31.2, p<.001, \eta_{p}^{2}=.48\right]$ and an effect of pace $\left[F(2,33)=12.6, p<.001, \eta_{p}^{2}=.27\right]$. However, contrary to SOB-CS predictions, no interaction was observed
$[F(2,33)<1]$. In fact, even at slow pace, we observed a positive effect of high similarity versus low similarity condition on recall performance.
Distractor intrusions Figure 2, panel A, shows a small effect of pace on distractor intrusions $[F(2,33)=3.18, p=$ $\left..047, \eta_{p}^{2}=.08\right]$ with $27 \%, 26 \%$ and $24 \%$ of distractor intrusions among responses at fast, moderate and slow pace respectively. In contrast, the model predicted a strong effect of the pace with almost no intrusion of distractors at medium and slow pace compare to $20 \%$ at fast pace. This discrepancy suggests that participants did not remove distractors as much as the model did. No similarity effect $[F(1,33)=2.69, p=$ $\left..11, \eta_{p}^{2}=.07\right]$ and no interaction $[F(2,33)<1]$ was observed on the distractor intrusion, as predicted by the model.
NPL intrusions We observed a strong effect of similarity on NPL intrusions $\left[F(1,33)=77.3, p<.001, \eta_{p}^{2}=.70\right]$ and a small pace effect $\left[F(2,33)=4.5, p=.014, \eta_{p}^{2}=.11\right]$ (Fig. 3, panel A). Here again, the model predicted a much stronger pace effect as compared to the experimental data. No significant interaction was found $\left[F(2,33)=2.24, p=.11, \eta_{p}^{2}=\right.$ .06] suggesting that the effect of similarity on NPLs is constant over the pace whereas the model predicts a stronger effect of similarity at fast pace.
Transposition errors No similarity and no pace effect on the transposition errors were found [respectively; $F(1,33)<$ 1 and $\left.F(2,33)=1.27, p=.28, \eta_{p}^{2}=.04\right]$ (Fig. 4, panel A). No significant interaction was found $[F(2,66)=1.27, p=$ $\left..28, \eta_{p}^{2}=.03\right]$. In line with these data, the model predicted no effect of the similarity on the transposition errors. However, the model predicted a small effect of pace between the fast and the two other paces.

## Discussion

The present results replicated the observations found in Experiment 2 of Oberauer, Farrell, et al. (2012) suggesting that forgetting in WM is partly due to interference by superposition. First, whatever the experimental condition, the mean proportion of distractor intrusions was higher than the proportion of NPLs ( 0.25 vs. 0.1 on average). This result demonstrates that distractors, unlike NPLs, are encoded into WM which is a necessary prerequisite for studying distractor interference. Moreover, evidence in favor of the interference by superposition mechanism was provided by replicating the strong benefit of high over low I-D similarity.

However, we observed a discrepancy between some of the model predictions and the data. SOB-CS only fits well the data in the fast condition that is similar to Oberauer, Farrell, et al. (2012)'s experiment, where no removal was used in the simulation. As soon as there is free time and hence removal, the SOB-CS simulation erroneously predicted an interaction between pace and I-D similarity. The error analyses revealed that this difference between model and human seems to be due to an overestimation of the removal strength by the model compared to the experimental findings.

In the following section, we present the results of a grid search on the removal parameter $r$ in order to identify a better $r$ value to reproduce the human data.

## Estimation of the removal parameter

Several experiments (Oberauer, 2001, 2002) estimated that removing part of the contents of working memory takes between 1 and 2 s . In addition to the time devoted to the removal, the strength of removal depends also on a rate of removal controlled by the free parameter $r$. The greater the value of $r$, the faster associations between the distractor and its position are removed. Therefore, in SOB-CS, the removal parameter $r$ was set to 1.5 , which implies that the rate of antilearning for removal has reached $95 \%$ of its asymptote after 2 s. According to the previous experimental analysis, it seems that distractors are not removed as quickly as in the model. To search for a value for the $r$ parameter that would better fit the data, we conducted a grid search on a range between 0 and 1.5 with a step size of 0.1 . The Root-Mean-Square Error (RMSE) was calculated for each parameter value. This measure represents the discrepancy between the model prediction and human data. The lowest RMSE corresponds to the best model. We found that the best model is the one with $r$ equals to 0.1 instead of the standard value 1.5 . If $r$ is set to zero, meaning no removal at all, an important loss of fit is observed as the model does not predict the pace effect anymore. The dashed lines of the panel B of all Figures shows the simulation results with $r$ set to 0.1 . First, the pace and the similarity variables do not interact anymore. Second, the main effect of pace on accuracy and error rates is about as large as in the data. Globally, we observed that the proportion of the different error types fits well the human data. Our result is in contradiction with the conclusion from previous studies (Oberauer, 2001, 2002) that removal takes only 1 to 2 s , as with this new $r$ parameter, removing completely irrelevant information would take about 30 s instead of 2 s .

## General Discussion

In this paper, we aimed to contribute to the debate regarding the reasons why distractors affect working memory performance by testing predictions of SOB-CS. More specifically, we investigated the I-D similarity effect after various amounts of free time. Human results confirm several predictions of SOB-CS. First of all, our results show that distractors are actually encoded in working memory since they were more often recalled than not-presented lures. Then, experimental data reproduced the positive effect of a high similarity between items and distractors originally found by Oberauer, Farrell, et al. (2012). This finding is predicted by the mechanism of interference by superposition of SOB-CS. Finally, we also observed that memory performance increases at a slower pace than predicted by the removal mechanism.

However, results disconfirm one prediction from the model with its standard parameter values: the data show that the ID similarity effect does not diminish with longer free time.

When simulating this experiment with the SOB-CS computational model, the similarity effect disappears when there is 0.8 s or more of free time available, because the model strongly removes distractors. In fact, we observed that at moderate and slow paces, distractors are totally removed according to SOB-CS. The total removing of the distractors cancels the recall difference between similar and dissimilar conditions. Contrary to this expectation, human data still showed distractor intrusions at moderate and slow pace.

Searching for a removal rate able to reproduce the experimental data resulted in a much lower estimate ( $r=0.1$ instead of $r=1.5$ ), which could reproduce the observed similarity effect at all three levels of pace. The removal mechanism was supported, because $r=0.1$ fit better than $r=0$. What are the implications of our removal rate estimate, which is much lower than that in the orignal model? Either, we can consider that $r=0.1$ is the parameter value that holds generally, implying that removal is much slower than thought so far. A way to verify this option would be to simulate other complex span task experiments to test whether their results can be reproduced by SOB-CS with $r=0.1$. Or, there is something particular to delete the material of our experiment that would require a low removal strength. In future research, a comparison of the size of the pace effect across experiments could shed some light on that. In fact, the comparaison of the pace effect of our experiment with all the other experiments can help to determine if our experiment had an exceptional low pace effect or if the removal strength needs to be lowered.

According to decay-based models of working memory, such as TBRS (Barrouillet et al., 2007) or TBRS* (Oberauer \& Lewandowsky, 2011), removal does not exist and free time is used to retrieve and maintain the to-be-remembered items. The maintenance of memory items can be viewed as the strengthening of the item-position bindings of the memory items and also as the strengthening of the representations of individual non-words themselves (i.e., item memory). Decaybased models predict the pace effect which has been observed many times. In fact, the more free time the more opportunity to maintain memory items. In such a model, the positive effect of high similarity between items and distractors, that is not accounted for by decay-based models, can be explained by the assumption that retrieving an item in order to refresh it is easier if it is less distorted by distractors. This process of retrieving would be required whatever the duration of the free time. The effect of similarity on retrieval therefore would lead to a beneficial effect of similar distractors whatever the pace. However, for the moment, decay-based computational models, such as TBRS* do not implement interference by superposition. In the future, it would be interesting to replace the removal mechanism by a mechanism of maintenance in SOB-CS.

## Acknowledgments

We would like to thank Louise Kenney who participated in the data gathering. We also acknowledge the French Regional

Council "Région Rhône-Alpes" which provided a PhD grant to V.H. We acknowledge the Université Grenoble Alpes who gave a grant allowing V.H. to go to the University of Bristol during three months in order to collaborate with S.L.

## References

Anderson, J. A. (1995). An introduction to neural networks. Cambridge, MA: MIT Press.
Baddeley, A. D. (1986). Working memory. New York, NY, US: Clarendon Press/Oxford University Press..
Baddeley, A. D., \& Hitch, G. (1974). Working memory. The psychology of learning and motivation Advances in research and theory, 8, 47-89.
Barrouillet, P., Bernardin, S., \& Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. J Exp Psychol Gen, 133(1), 83-100.
Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., \& Camos, V. (2007). Time and cognitive load in working memory. J Exp Psychol Learn, 33(3), 570-585.
Barrouillet, P., Portrat, S., \& Camos, V. (2011). On the Law Relating Processing to Storage in Working Memory. Psychol Rev, 118, 175-192.
Brainard, D. H. (1997). The psychophysics toolbox. Spatial vision, 10, 433-436.
Miyake, A., \& Shah, P. (Eds.). (1999). Models of working memory: Mechanisms of active maintenance and executive control. New York: Cambridge University Press.
Oberauer, K. (2001). Removing irrelevant information from working memory: a cognitive aging study with the modified Sternberg task. J Exp Psychol Learn, 27(4), 948-957.
Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. J Exp Psychol Learn, 28(3), 411-421.
Oberauer, K., Farrell, S., Jarrold, C., Pasiecznik, K., \& Greaves, M. (2012). Interference between maintenance and processing in working memory: The effect of itemdistractor similarity in complex span. J Exp Psychol Learn, 38(3), 665-685.
Oberauer, K., \& Lewandowsky, S. (2011). Modeling working memory: a computational implementation of the TimeBased Resource-Sharing theory. Psychon B Rev, 18(1), 1045.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., \& Greaves, M. (2012). Modeling working memory: An interference model of complex span. Psychon B Rev, 19, 779819.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.
Rastle, K., Harrington, J., \& Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. Q J Exp Psychol, 55(A), 1339-1362.
Towse, J. N., \& Hitch, G. J. (1995). Is there a relationship between task demand and storage space in tests of working memory capacity? Q J Exp Psychol, 48(A), 108-124.

# Spatial Cognition and Science: The Role of Intrinsic and Extrinsic Spatial Skills from Seven to Eleven Years 

Alex Hodgkiss (alex.hodgkiss.14@ucl.ac.uk)<br>Department of Psychology and Human Development, UCL Institute of Education, 25 Woburn Square London WC1H 0AA, UK<br>Katie A. Gilligan (katie.gilligan.15@ucl.ac.uk)<br>Department of Psychology and Human Development, UCL Institute of Education, 25 Woburn Square London WC1H 0AA, UK<br>Michael S. C. Thomas (m.thomas@bbk.ac.uk)<br>Department of Psychological Sciences, Birkbeck College, Malet Street, London, WC1E 7HX<br>Andrew K. Tolmie (a.tolmie@ucl.ac.uk)<br>Department of Psychology and Human Development, UCL Institute of Education, 25 Woburn Square London WC1H 0AA, UK<br>Emily K. Farran (e.farran@ucl.ac.uk)<br>Department of Psychology and Human Development, UCL Institute of Education, 25 Woburn Square London WC1H 0AA, UK


#### Abstract

The current study investigated the relationship between children's spatial ability and their scientific knowledge, skills and understanding. Children aged $7-11$ years $(N=123)$ completed a battery of five spatial tasks, based on a model of spatial ability in which skills fall along two dimensions: intrinsic-extrinsic; staticdynamic. Participants also answered science questions from standardised assessments, grouped into conceptual topic areas. Spatial scaling (extrinsic static spatial ability) and mental folding (intrinsic dynamic spatial ability) each emerged as predictors of total science scores, with mental folding accounting for more variance than spatial scaling. Mental folding predicted both physics and biology scores, whereas spatial scaling accounted for additional variance only in biology scores. The embedded figures task (intrinsic static spatial ability) predicted chemistry scores. The pattern was consistent across the age range. These findings provide novel evidence for the differential role of distinct aspects of spatial ability in relation to children's science performance.


Keywords: spatial ability; science; STEM; children.

## Introduction

Large-scale longitudinal studies spanning the past 50 years provide convincing evidence that spatial ability in adolescence predicts later science, technology, engineering and mathematics (STEM) achievement; both in academic and career outcomes (Wai, Lubinski \& Benbow, 2009). As well as often cited examples of scientific discoveries resulting from creative spatial thought, a growing body of research with adults and adolescents highlights a link between spatial ability and scientific reasoning (e.g, Kozhenikov \& Thornton, 2006). However, with a few exceptions (e.g Tracy, 1990), the relationship between
spatial ability and science learning in younger children has been largely neglected. This is important to address, given that early science learning involves specific areas of conceptual understanding, and because knowledge of how spatial ability and science relates in younger children has implications for early intervention. The focus of this study was therefore to investigate the relationship between a range of spatial skills and primary-school aged children's scientific knowledge, skills and understanding.

## Spatial ability

Spatial ability, which relates to "the location of objects, their shapes, their relation to each other, and the paths they take as they move" (Newcombe, 2010, p30), has long been recognised as an ability at least partly independent of general intelligence, reasoning and verbal ability (Hegarty, 2014). As well as being apparently distinct from other cognitive abilities, spatial thought itself is generally conceptualised in a multidimensional fashion, consisting of several separate but correlated skills. Two categories of multidimensional models have emerged: ones based in the psychometric tradition (e.g. Carroll, 1993) and other more recent, theoretically driven models (e.g, Uttal et al., 2013).

Psychometric analyses of spatial ability have often resulted in inconsistent findings, with the number of identified factors ranging between two and twelve (Hoffler, 2010). Uttal et al. (2013) argue that some of the inconsistencies result from factor analysis models not being theory-driven. In contrast, Uttal et al.'s (2013) model is based on top-down, theory driven understanding of spatial skills, and draws upon developments in cognitive neuroscience. They propose that spatial skills can be
categorised along two dimensions: static-dynamic and intrinsic-extrinsic. Intrinsic spatial abilities and extrinsic spatial abilities broadly map onto a within-object and between-object classification, respectively. Intrinsic/extrinsic skills can be further categorised as either static or dynamic abilities; dynamic abilities include transformation or movement, whilst static skills do not.

Intrinsic-static skills involve the processing of objects or shapes without further transformation or movement of parts of the object or shape. Tasks that measure this skill often require this processing to occur amidst distracting background information. For example, in disembedding tasks, participants search for a specified 2D shape in a larger distracting image. Intrinsic-dynamic skills involve the processing, and manipulation or transformation of objects or shapes. 2D and 3D mental rotation fit into this category.

Extrinsic-static skills require the processing and encoding of the spatial relations between objects or configurations of objects, without further manipulation or transformation of these relations. The extrinsic-static category includes the ability to find corresponding locations between shapes of equal proportion but differing sizes (e.g. scaling and map use). Extrinsic-dynamic skills involve the apprehension, processing and manipulation of more than one object, or the relationship between objects and frames of reference. Spatial perspective taking, in which a participant visualises what an object would look like from a different viewpoint, is an extrinsic-dynamic skill because it involves the manipulation of the relationship between an object and another frame of reference/viewpoint.

## Spatial ability and science

Spatial skills support understanding and learning of conceptual areas that are very spatial in nature (e.g, astronomy) yet even apparently non-spatial topics are often represented in a spatial format.

Most prior research with adults point to visualisation skills as being related to science learning: the ability to mentally transform spatial information about single objects, assessed through intrinsic-dynamic skills such as mental rotation and mental folding. For example, studies report a link between intrinsic-dynamic spatial skills and conceptual understanding in aspects of biology (e.g. Garg, Norman, Spero \& Mashewari, 1999), chemistry (e.g. Stull, Hegarty, Dixon \& Stieff, 2012) and physics (e.g., Kozhenikov \& Thornton, 2006). In Stull et al. (2012), for instance, 3D visualisation positively correlated with undergraduate students' ability to translate between different diagrammatic representations of chemical structures.

Other spatial skills within Uttal et al.'s (2013) model may play a role in science learning. However, this relationship has been largely neglected to date; the role of extrinsic-static skills such as scaling, for example, has yet to be addressed.

## Spatial ability and science in development

Research relating spatial ability and science learning in younger children is sparse (e.g. Jarvis \& Gathercole, 2003;

Tracy, 1990). Tracy (1990) assessed science performance in a sample of 10 - and 11-year-old students who were split into high and low spatial ability. The study revealed that the high spatial ability group outperformed the low spatial ability group. However, this study did not include any measure of IQ or other cognitive factors, and thus did not discount general ability as an alternative explanation. It also used a composite spatial measure.

One unpublished study that compared the role of different intrinsic-dynamic spatial ability measures on scientific understanding in children, and controlled for verbal ability, found that mental folding, but not mental rotation, predicted five-year-olds' performance in a task involving understanding of force and motion. However, this was still limited to intrinsic-dynamic skills (Harris, 2014).

There is also some mixed evidence to suggest that spatial skills may be more important during the early stages of science learning, rather than later stages (Hambrick et al., 2012). During initial learning a learner may use spatial processing to establish mental maps and models or to problem solve (Mix et al., 2016). With experience, domainspecific knowledge may become more important. Such a hypothesis is supported in the science literature by the finding that visuospatial working memory is less predictive of 14 -year-old versus 11-year-old students' science performance (Jarvis \& Gathercole, 2003), and that mental folding ability predicted children's but not adults' understanding of forces in the previously described study by Harris (2014). The relationship between spatial thinking and science performance, and how this varies between the ages of 7 and 11 , has hitherto remained unclear.

## Current Study

Although prior research indicates that spatial ability predicts aspects of science learning in older populations, little research has been conducted with younger children. Research that has done so, has either focused on visualspatial working memory only, used a composite of spatial ability measures, or has focused only on single object-based manipulation (intrinsic-dynamic spatial skills). Furthermore, no research to date has used and compared a cross-sectional sample to determine if this relationship varies across development.

The aim of the current study was to examine the relationship between 7 and 11-year-olds' performance in a range of spatial ability measures, based on the Uttal et al. (2013) model, and their performance in a science assessment covering aspects of biology, chemistry and physics.

## Methods

## Participants

The initial sample consisted of 127 participants who were recruited from a large, ethnically diverse London primary school. Three pupils did not go on to complete the study
because they were unsuitable for the study due to having a special educational need or an insufficient level of English. Due to missing data, four further participants (one participant per year group) did not have a full set of scores. Three of those participants were missing data from one task only, and so their missing scores were estimated by calculating the mean for their respective year group. The fourth participant, who was missing several variables, was excluded from the analysis. Thus, four participants were excluded in total. The final sample consisted of 123 participants in UK Years 3-6: Year 3 ( $\mathrm{N}=32$, mean (s.d.) age $=8.0$ ( 0.28 ) years), Year $4(\mathrm{~N}=31$, mean (s.d.) age $=9.0$ (0.32) years), Year $5(\mathrm{~N}=31$, mean (s.d.) age $=10.0$ (0.33) years), Year 6 ( $\mathrm{N}=30$, mean (s.d.) age $=11.0$ (0.30) years). Parental consent was obtained for all participants.

## Measures

## Intrinsic-Dynamic Spatial: Mental Rotation

In this mental rotation task (based on Broadbent, Farran \& Tolmie, 2014), children were shown two upright cartoon monkeys, above a horizontal line, on a computer screen, and one monkey below a line which was rotated by varying degrees $\left(0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}, 180^{\circ}\right)$. One monkey above the horizontal line had a blue left hand and a red right hand, and the other monkey had the reverse pattern and was a mirror image of the other. Children were asked which of the two monkeys at the top of the screen matched the monkey at the bottom of the screen, which had been rotated. This task began with six practice items, in which the monkey below was not rotated ( $0^{\circ}$ trials) and then progressed to 36 experimental trials $\left(4 \times 0^{\circ}\right.$ trials, $8 \times 45^{\circ}$ trials, $8 \times 90^{\circ}$ trials, $8 \times 135^{\circ}$ trials and $8 \times 180^{\circ}$ trials).

## Intrinsic-Dynamic Spatial: Mental Folding

This mental folding task (Harris et al., 2013) required children to imagine folds made to a piece of paper, without physical representation of the folding action itself (see Figure 1). Children were shown a shape at the top of a computer screen which contained a dotted line and an arrow. The dotted line represented the imaginary fold line and the arrow indicated where the paper should be folded to. Beneath this item on the screen, children were shown four images of how the item at the top might look after being folded at the dotted lines, only one of which was correct. Children first completed two practice items. Answers to practice questions were checked by the researcher, and if a child had an incorrect answer, they were given one further attempt of that trial. The experimental trials then began, where children had 14 novel items to work through. The test progressed automatically as the child clicked one of the four images at the bottom of the screen. Accuracy and response time were recorded.

## Intrinsic-Static Spatial: Embedded Figures Task

The Children's Embedded Figures Task (CEFT: Witkin et al., 1971) consists of complex figures in which a simple form is embedded (see Figure 1). Children were shown an
image constructed of colourful geometric shapes and asked to locate either a simple house or tent shape 'hidden' within the image. Children were shown the house or tent shape as a cardboard form; it was kept by the child for the first three trials of each block and hidden thereafter. For the first part of the test ( 11 items ) children located a triangular tent shape within each image, the simpler of the two shapes, and for the other half of the test ( 14 items) children located a house shape. Children were given a score of 1 for correctly locating the shape hidden within the figure. Accuracy and response times were recorded using a laptop computer.

## Extrinsic-Static Spatial: Scaling

Our novel spatial scaling task was based on a similar task by Frick and Newcombe (2012). Children were required to find equivalent corresponding locations on two grids, when one was varied in size relative to the other by a predetermined scale factor (see Figure 1). The task was presented to children as a game which involved pirates' treasure maps. Treasure maps were printed on yellow paper and mounted in a large ring bound pad. Each page contained one yellow map with a grid printed in black. Nine (out of 18) items contained grids which separated the map into $6 \times 6$ sections, whereas the other nine items contained grids which separated the map into $10 \times 10$ sections. For each trial, one target section of the printed grid map was coloured in black (the treasure); the target section varied across trials. Participants were also presented with four maps on a touch screen computer, which each had one black square coloured. One computer map contained a black square at a location which corresponded to the printed map; the locations of the black squares on the other three incorrect computer options were systematically chosen. The larger printed maps were either unscaled (1:1), or scaled to either $1: 4$ or $1: 8$, relative to the maps on the computer screen. Participants first completed two practice items after which they completed the main 18 trials of the test. Six items were presented at each scale factor.

## Extrinsic-Dynamic Spatial: Perspective Taking Task

This task was identical to one developed by Frick et al. (2014) which involved spatial perspective taking in which children were required to visualise what photographs would look like when taken from cameras placed at different positions and angles relative to their viewpoint. The child first completed four practice questions involving actual Play Mobil characters, and then one practice question on a computer, which showed a character taking a photograph of two shapes from the same perspective as the child. The child selected the correct option, of four, from below the main image, which showed what the photograph would look like by pressing a touch screen computer. If a child made an error on the practice items, they were given a second attempt. On passing the practice questions, the task continued with the experimental items, where they again chose the correct image out of four options. These varied per the number of objects in the layout (one, two or three)
and the angular difference between the photographer's and the child's perspective $\left(0^{\circ}, 90^{\circ}\right.$ or $\left.180^{\circ}\right)$.

## Science Assessment

The science assessment consisted of two paper-based tests, which children completed in class groups in two separate sessions under the supervision of the researcher. The assessment was curriculum-based and all questions were taken from an online database of past science UK standardised test questions ("Test Base", 2017, January 31 ${ }^{\text {st }}$ ) designed to assess the science curriculum in this age range. The test included approximately equal amounts of biology, chemistry and physics content from a selection of topics appropriate to this curriculum stage.

Each paper had a total possible score of 50 marks leading to a total science mark of 100 . The assessment included questions which varied in difficulty, which again mapped onto curriculum descriptors. Paper one contained questions of low to medium cognitive demand, and paper two contained questions of high cognitive demand. Questions focused on one conceptual sub-topic, e.g., magnetism or changing state. Questions were further sub-divided into items that were either: factual/recall items (e.g. label a diagram; recall a function); problem solving items, which drew on conceptual knowledge; or items that were in the context of hypothetical experiments, thus drawing on procedural skills.


Figure 1: significant spatial predictors (left to right): mental folding task, spatial scaling task ( $1: 8,6 \times 6$ trial), embedded figures task (locate a triangular 'tent' trial).

## Control Variables

The British Picture Vocabulary Scale-III (BPVS-III; Dunn, Dunn, Styles, \& Sewell, 2009) was included as a measure of verbal ability.

## Procedure

All children first completed the two paper-based science assessments, in two sessions administered by the researcher in class groups, within the child's own classroom. Spatial ability was then assessed within two separate sessions. Children were first tested in a computer-based group session of no more than 10 children where they completed the mental folding task and the monkey mental rotation task, in a counterbalanced order. The BPVS, CEFT, spatial perspective taking task and scaling task were then completed in an individual testing session with the researcher, which lasted approximately 30 minutes per child. The order of testing in the individual sessions was also counterbalanced to control for fatigue and order effects.

## Results

A series of multiple regression analyses were conducted to determine the amount of variance in science scores that was accounted for by each of the spatial ability measures. There were no significant gender differences ( $\mathrm{p}>.05$ for all); therefore, participants were treated as one group in the subsequent regression analysis. A separate analysis was run for total science score and for each area of science (physics, biology, chemistry). In each of the models, the covariates (age, BPVS raw score) were added first (steps 1-2). All spatial measures were subsequently entered in a single block, using forward step-wise entry, to determine the best model of spatial ability predictors. Beta values refer to the final model with all variables entered.

Entered in the first step of each model, age in months significantly predicted each science total. Age was most strongly predictive of physics sub-score, with this initial step accounting for $27 \%$ of the variance, $\Delta F(1,121)=45.52$, $p<.001$. Age accounted for $21 \%$ of the variance in total score, $\Delta F(1,121)=31.27, p<.001,13 \%$ of the chemistry scores, $\Delta \mathrm{F}(1,121)=17.75, p<.001$ and $8 \%$ of biology scores, $\Delta F(1,121)=11.03, p<.001$. In the final overall models, after additional variables were entered, age remained as a significant predictor of total score $(\beta=.122, t$ $=2.03, p=.044)$ and physics score $(\beta=.320, t=4.24, \mathrm{p}<$ .001 ), but not biology score ( $\beta=.010, t=.10, p=.916$ ) or chemistry score $(\beta=.102, t=1.23, p=.223)$.

BPVS raw score was entered in the second step of each model and was a significant predictor. BPVS scores were most strongly predictive of total science score, accounting for an additional $37.6 \%$ of the score variance, $\Delta F(1,120)=$ 106.16, $p<.001,25 \%$ of the biology scores, $\Delta F(1,120)=$ $46, p<.001,21 \%$ of the chemistry scores, $\Delta F(1,120)=$ $36.93, p<.001$ and $15 \%$ of the physics scores $\Delta F(1,120)=$ 29.78, $p<.001$. In the final model, BPVS scores remained significant predictors of total science score $(\beta=.567, t=$ $8.9, p<.001$ ), biology score ( $\beta=.443, t=5.38, p<.001$ ), chemistry score $(\beta=.485, \mathrm{t}=5.93, p<.001)$ and physics score $(\beta=.375, \mathrm{t}=4.85, p<.001)$.

The step-wise analysis of spatial measures to predict total science score resulted in mental folding being entered in
step three and spatial scaling being entered in step four of the regression model. Mental folding accounted for an additional $6 \%$ of the variance in total science score, above the previously entered covariates $\Delta F(1,119)=20.62, p<$ $.001, \beta=.211, t=3.54, \mathrm{p}=<.005$. The scaling task accounted for an additional $2 \%$ of the variance in total science scores, $\Delta F(1,118)=6.8, p<.010, \beta=.162, t=2.60$, $p=.010)$.

Step-wise entry of spatial measures predicting biology scores retained mental folding and the scaling task only, in steps three and four. Mental folding accounted for an additional $6 \%$ of the variance in biology scores, $\Delta F(1,119)$ $=11.65, p<.001, \beta=.195, t=2.52, p=.013$ ). Spatial scaling accounted for an additional $3 \%$ of the variance in biology scores, above the previously entered covariates $\Delta F(1,118)=5.50, p<.021, \beta=.190, t=2.34, p=.021)$.

Only the mental folding task was retained as predictor of the physics score $(\beta=.198, t=2.8, p=.006)$ following step-wise analysis, and was entered in step three of the model. It accounted for an additional $4 \%$ of the variance in physics scores $\Delta F(1,119)=7.82, p=.006)$.

The CEFT was the only emerging predictor of chemistry scores $(\beta=.173, t=2.27, p=.025)$. It accounted for an additional $3 \%$ of the variance in chemistry scores $\Delta F(1,119)$ $=5.130, p=.025$ ). Any remaining spatial tasks not reported did not significantly predict any additional unique variance in science ability beyond those spatial measures included in the above models ( $p>.05$ for all).

To determine if age interacted with any of the spatial ability measures, a further four models were constructed in which the covariates were again entered in step 1, followed by the spatial ability measures found to be significant for that science measure, followed by interaction terms (age in months*spatial measure). No significant interactions with age were found ( $p>.05$ for all).

## Discussion

This study revealed that spatial ability predicted children's performance in a curriculum-based science assessment. That is, after controlling for age, gender and verbal ability, spatial ability accounted for $8 \%$ additional variance in total science scores. This builds upon longitudinal research linking spatial ability to STEM outcomes in adults (Wai et al., 2009) as well as correlational research that has associated spatial ability with various aspects of science learning in adults (e.g., Kozhevnikov et al., 2006). It also expands upon the existing findings from child data (Tracy, 1990; Harris, 2014) by investigating a broader range of spatial measures and science topic areas and, also, comparing a wider age range of children within one study.

Mental folding, an intrinsic-dynamic spatial skill, accounted for the most variance in total overall science scores, relative to the other spatial tasks. This is likely to be due to it being a predictor of both physics and biology topics, whereas spatial scaling was not. It is likely that the relationship with physics scores is driven, in part, by
questions on topics such as magnetism, forces and motion, which required visualisation of how objects move and interact; this suggests that children who are more skilled at visualising paper folds are also better at predicting the direction of various types of forces acting on objects.

The role of mental folding in relation to biology scores is less likely to be directly related to the visualisation skill, as in physics, discussed above. One possibility is that the ability to flexibly maintain and manipulate spatial information, as measured through the folding task, may also be related to mental model construction and use. The mental models children possess for the conceptual topics within biology may be spatially-based. For example, when recalling the function of roots, children may recall a spatial mental model of a plant, which is integrated with verbal/propositional information.

Although mental rotation falls into the same category (Uttal et al., 2013) as mental folding (which was a strong predictor), mental rotation accuracy did not feature in any of the final regression models. The two measures correlated only moderately $(r=.294)$. This may be because there are differences between folding and rotation, despite them both being intrinsic/dynamic measures. For example, rotation is a rigid, intrinsic transformation and folding is a non-rigid, intrinsic transformation (Atit, Shipley \& Tikoff, 2013). Further research is needed to investigate this distinction.

Spatial scaling, an extrinsic/static skill, also emerged as a predictor of total science scores, although it contributed less to this model than folding because it was significant only for biology questions. One interpretation of the role of scaling is that it predicted performance because children who perform well on this task are also more able to determine the correspondence between representations of scientific concepts at different scales. Children may, for example, in the classroom, need to determine the correspondence between: an actual plant; scaled-up versions of plant diagrams on an interactive whiteboard; or scaled-down, abstract printed diagrams.

The CEFT, an intrinsic-static spatial skill, was a significant predictor of chemistry scores, but did not feature in the other final regression models. Intrinsic-static spatial skills relate to the processing of objects without further transformation: the arrangements of parts of the object (subparts) as well as the size and orientation objects. This could relate to chemistry items including diagrams which require inspection and discrimination between sub-parts of objects (e.g, 3 beakers with ice cubes, which either have 1,2 or 3 layers of insulation).

An analysis of age-based interactions revealed that the relationships described were steady across development. We had predicted that spatial skills would contribute more to science performance for younger children, than older children, based on prior research (e.g., Hambrick et al., 2012), which would suggest that as domain-specific knowledge increases, spatial abilities play less of a role in science. It may be that, although older children are more experienced in science, the knowledge they have available
for immediate recall may be restricted to topics that they are currently covering.

One limitation of the study was that we only included the BPVS as a control for general level of ability. One might argue that the spatial tasks are also capturing a more general problem solving ability. However, the differential role of various spatial skills revealed in the analyses demonstrate that spatial ability is having an impact on science performance versus it being a general problem solving ability proxy. However, further research should include other measures of general cognitive ability. Second, the nature of the questions, drawn from standardised assessments, meant that it is difficult to determine if the relationships observed relate to scientific knowledge/understanding, or application of knowledge in scientific problem solving. Further studies are planned to systematically include a range of question types for comparison across categories of items.

In summary, the current study provides evidence for a distinctive role for mental folding (intrinsic-dynamic spatial ability), spatial scaling (extrinsic-static spatial ability) and the CEFT (intrinsic-static spatial ability) in children's science learning. The findings have implications for how we can move forward to support children in the science classroom through spatial training and interventions.

## Acknowledgements

This work was supported by an Economic and Social Research Council PhD studentship awarded to the first author.

## References

Atit, K., Shipley, T. F., \& Tikoff, B. (2013). Twisting space: are rigid and non-rigid mental transformations separate spatial skills? Cognitive processing, 14(2), 163-173.
Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge: University Press.
Broadbent, H. J., Farran, E. K., \& Tolmie, A. (2014). Objectbased mental rotation and visual perspective-taking in typical development and Williams Syndrome. Developmental Neuropsychology, 39(3), 205-225.
Dunn, L. M., Dunn, D. M., Styles, B., \& Sewell, J. (2009). The British Picture Vocabulary Scale - 3rd Edition. London: GL Assessment.
Frick, A., Möhring, W., \& Newcombe, N. S. (2014). Picturing perspectives: development of perspectivetaking abilities in 4-to 8-year-olds. Frontiers in Psychology, 5, 386.
Frick, A., \& Newcombe, N. S. (2012). Getting the big picture: Development of spatial scaling abilities. Cognitive Development, 27(3), 270-282
Garg, A., Norman, G. R., Spero, L., \& Maheshwari, P. (1999). Do virtual computer models hinder anatomy learning? Academic Medicine, 74(10), S87-89.

Hambrick, D. Z., Libarkin, J. C., Petcovic, H. L., Baker, K. M., Elkins, J., Callahan, C. N., LaDue, N. D. (2012). A test of the circumvention-of-limits hypothesis in scientific problem solving: The case of geological bedrock mapping. Journal of Experimental Psychology: General, 141(3), 397.
Harris, J. (2014). Where will it go? Concepts of motion in complex events. (Unpublished doctoral thesis). Temple University, Philadelphia, Pennsylvania.
Harris, J., Newcombe, N. S., \& Hirsh-Pasek, K. (2013). A new twist on studying the development of dynamic spatial transformations: Mental paper folding in young children. Mind, Brain, and Education, 7(1), 49-55.
Hegarty, M. (2014). Spatial thinking in undergraduate science education. Spatial Cognition \& Computation, 14(2), 142-167.
Höffler, T. N. (2010). Spatial ability: Its influence on learning with visualizations-a meta-analytic review. Educational Psychology Review, 22(3), 245-269.
Jarvis, H. L., \& Gathercole, E. E. (2003). Verbal and nonverbal memory and achievements on national curriculum tests at 11 and 14 years of age. Educational and Child Psychology, 20(3), 123-140.
Kozhevnikov, M., \& Thornton, R. (2006). Real-time data display, spatial visualization ability, and learning force and motion concepts. Journal of Science Education and Technology, 15(1), 111-132.
Mix, K. S., Levine, S. C., Cheng, Y. L., Young, C., Hambrick, D. Z., Ping, R., \& Konstantopoulos, S. (2016). Separate but correlated: The latent structure of space and mathematics across development. Journal of Experimental Psychology: General, 145, 1206-1227.
Newcombe, N. S. (2010). Picture this: increasing math and science learning by improving spatial thinking. American Educator, 34(2), 29.
Stull, A. T., Hegarty, M., Dixon, B., \& Stieff, M. (2012). Representational translation with concrete models in organic chemistry. Cognition and Instruction, 30(4), 404434.

Test Base. (2016, January 31 ${ }^{\text {st }}$ ). Retrieved from http://www.testbase.co.uk/sec/index.php.
Tracy, D. M. (1990). Toy-playing behavior, sex-role orientation, spatial ability, and science achievement. Journal of Research in Science Teaching, 27(7), 637-649.
Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., \& Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. Psychological Bulletin, 139(2), 352.
Wai, J., Lubinski, D., \& Benbow, C. P. (2009). Spatial Ability for STEM Domains: Aligning Over 50 Years of Cumulative Psychological Knowledge Solidifies Its Importance. Journal of Educational Psychology, 101, 817-835.
Witkin, H. A. (1971). Children's Embedded Figures Test, Consulting Psychologists Press, Palo Alto, CA.

# Moving together: in the body or the mind? 

Lilla Hodossy (Lilla.Hodossy@rhul.ac.uk)<br>Department of Psychology, Royal Holloway University, Egham Hill, Egham TW20 0EX, UK

Jorina von Zimmermann (jorina@eyethink.org)<br>Daniel C. Richardson (dcr@eyethink.org)<br>Department of Experimental Psychology, University College London, Gower St, London WC1E 6BT, UK


#### Abstract

When people move together, as they dance, march or flirt, it increases affiliation between them. But what about 'moving together' produces affiliation: the movements themselves, or the social context of moving 'together'? We instructed pairs of participants to listen to music and move their arms or legs according to shapes appearing on screen. They either carried out the same movements, or when one moved their arms the other moved their legs. They either saw shapes on one laptop, or each had their own laptop. Surprisingly, participants did not like each other more if they carried out the same movements, but affiliation did increase if they danced looking at the same screen. Rather than their movements, instructions, intentions or perceptual experiences, here it is the social context of the actions that produces affiliation, a surprising finding that is not easily accounted for by the dominant theories of mimicry and behavioural synchrony.


Keywords: synchrony; coordination; mimicry; affiliation; joint action

## Introduction

People have danced, marched and moved together across cultures and history (McNeill, 1995). One reason, suggested by the literature, is that mimicry and synchronous movement acts as a form of 'social glue', increasing liking when two people mimic each other's gestures (Chartrand \& van Baaren, 2009), walk in step with each other (Wiltermuth, 2012), tap in synchrony (Hove \& Risen, 2009), or move together to a common beat (Reddish, Fischer, \& Bulbulia, 2013).

Dance as one particularly social form of human coordination (Dunbar, 2012; Tarr, Launay, Cohen, \& Dunbar, 2015), usually takes place in a shared social context: people in the same room, co-present with others, are listening to the same music. Similarly, demonstrations of motor mimicry increasing affiliation also take place in the shared social context of an experiment. What is the contribution of these two factors, a shared social context and similarity in movement, in changing affiliation when people dance together? From research to date this is not clear, as the two factors are confounded in dance as it typically occurs in society, and mimicry as it is typically studied in the laboratory. Which provide the psychological conditions for dancing 'together'?

The recent invention of the "silent discos" separates these factors and raises an interesting question. At these events, each person wears a set of wireless headphones that can be connected to different DJs playing different pieces of music. So two people next to each other may be engaging in similar, synchronised bodily movements, or not. Each person may or may not know if the person next to them is listening to the same music. What conditions will produce affiliation between the dancers: the similarity between their movements, or the knowledge that they are dancing together to the same music? We created an experimental version of this situation to find out. Rather than manipulating shared music, however, we manipulated shared social context.

This question does not just relate to our specific understanding of silent discos, of course, but raises much broader questions about the function of bodily mimicry for the affective states of social relationships. The dominant prediction from the psychological literature is that movement similarity causes affiliation. One proposed mechanism is that action observation activates a representation of a similar motor plan in the observer (Chartrand \& Bargh, 1999). This mirroring has been linked to a particular set of visuo-motor neurones in the brain known as the 'mirror system' (Di Pellegrino, Fadiga, Fogassi, Gallese, \& Rizzolatti, 1992; Mukamel, Ekstrom, Kaplan, Iacoboni, \& Fried, 2010), which are claimed to contribute to social cognition and affiliation (Gallese \& Goldman, 1998; Pineda, 2009). Supported by many experimental findings, these theories predict that what is required to increase affiliation between two silent disco dancers is a match between their bodily movements.

However, there are two reasons to hypothesize that social context may play an important part in the relationship between affiliation and mimicry. Firstly, imitation can be modulated by social factors such as eye contact (Wang, Newport, \& Hamilton, 2010), group membership (Yabar, Johnston, Miles, \& Peace, 2006), or the circumstances under which people meet (Miles, Griffiths, Richardson M., \& Macrae, 2010). Secondly, one lab experiment has shown that affiliation can be increased by action contingency alone. Catmur and Heyes (2013) asked participants to perform either a hand or a foot movement at random. In response, they either saw the same action that they had just performed onscreen, or the opposite one. The actions onscreen either occurred contingently, every time participants acted, or noncontingently, sometimes appearing and sometimes not.


Figure 1. (a) In the joint condition, participants' headphones were plugged into a splitter and they shared a screen. In the parallel condition, participants had two separate screens and their headphones were plugged in separately. In response to a shape appearing on screen, participants either moved the same limb at the same time as each other, or when one moved a hand the other moved a foot (or vice versa). (b) Laboratory layout in the parallel context condition. In the joint condition the screen was placed in the middle.

Participants' pro-social feelings were influenced by the contingency of their actions, but not by the match between the actions they made and the actions they saw.

This leads to the prediction that for our two dancers at a silent disco, what will increase affiliation is the social context that leads them to interpret their actions as linked and contingent upon each other - that they are dancing 'together'. In contrast, the prediction from the behavioural coordination literature is that affiliation will be higher when the participants' movements are the same.

To test these predictions, we instructed pairs of participants to listen to music on headphones and to perform the same or different movements in one of two conditions: in the joint social context, participants danced 'together' looking at a single computer screen that guided their movements. In the parallel social context, they each had their own screens, side by side, showing the same movement instructions. Afterwards, we measured participants' levels of affiliation to tease apart the contribution of social context and movement similarity in producing liking.

## Methods

We performed two experiments manipulating movement similarity and social context between pairs of participants. In the first experiment, as well as using a single screen, the joint social context was additionally established by giving participants the task of first untangling their headphone cables before plugging into their shared display. In our second experiment, we aimed to replicate our methods, but with the untangling task removed, so that the joint social context was established by the shared display alone. Since the experiments and analyses are identical in every other regard, we describe them together here.

## Participants

We estimated an effect size of $d=0.7$, following Lumsden, Miles and Macrae (2014) for affiliation effects arising from mimicry, and an a priori power analysis using $G^{*}$ Power (version 3.1.9.2; Faul, Erdfelder, Lang, \& Buchner, 2007) suggested a sample size of 76 to achieve $85 \%$ power (with $\alpha$ $=0.05$ ). To be conservative, Experiment 1 tested 80 participants ( 58 females; mean age $=24.66$ years; $\mathrm{SD}=6.84$ years, 29 non-UK nationals) and Experiment 2 tested 82 participants ( 62 females; mean age $=21.8, \mathrm{SD}=6.1,45$ nonUK nationals), recruited from the UCL Psychology Subject Pool. No participants were excluded. Participants in both studies were compensated through course credit or a monetary reward of $£ 4$. It was ensured that the members of each dyad did not know each other prior to the experiment.

## Ethics Statement

Ethical approval was obtained from the UCL Research Ethics Committee. All participants provided written informed consent before the beginning of the study and were fully debriefed upon completion.

## Procedure

Experiments had a $2 \times 2$ (social context: joint or parallel; movements: same or different) between-subjects design with pairs randomly assigned to conditions (Figure 1). In Experiment 1, participants in the joint condition were first given the task of untangling headphone cables together then plug them in. In the parallel condition, the headphones were already plugged into separate laptops. In Experiment 2, the untangling task was not included in either condition.

Participants wore headphones and stood 1 m apart from each other, 2.5 m away from a table, in a square marked on the floor (Figure 1). In the parallel social context, there were two screens (diameter: 27 cm , diagonal: 33.78 cm ) 1 m apart on the table that showed identical stimuli throughout the experiment. In the joint social context, there was one screen midway between participants.

In the same movement condition, both participants were given same shape-movement instructions (e.g., circle $=$ leg movement and triangle $=$ hand movement). In the different movement condition, one participant had the mapping reversed. Shapes were presented randomly on the left, right or middle of the display, indicating the direction that participants were to move their limbs. After a practice stage, participants danced for $4: 50 \mathrm{~min}$ to shapes appearing every 1.2 seconds, matching the tempo of the song 'I turn my
camera on' by Spoon. Participants were then led into different rooms. We measured affiliation in two ways. First, participants responded to a 15 item subset of the Subject Impressions Questionnaire (SIQ) from the Intrinsic Motivation Inventory (Ryan, Koestner, \& Deci, 1991). Second, we measured affiliation with the Inclusion of Other in Self (IOS) scale (Aron, Aron, \& Smollan, 1992), in which participants chose between 7 pairs of differently overlapping circles to represent their relationship with the other participant. Finally, as a manipulation check, we asked how much the participants attended to each other, and if they felt like they were dancing 'together'. The experiment lasted for approximately 30 minutes. All measures and manipulations have been reported here, and analysis did not begin in each experiment until we had collected our target of 80 participants in each.

## SOCIAL CONTEXT <br> joint parallel



MOVEMENT
same different



Figure 2. The main effects of social context and movement conditions on two measures of affiliation: SIQ and IOS. There were no significant interactions between the effects. Red and blue lines show the distribution of scores in each condition, with dotted lines giving mean. Grey lines show the Bayesian estimate of distribution of the difference between conditions; grey areas show the $95 \%$ credibility intervals.

## Results

Across two different measures we found strong evidence that affiliation was increased by a joint social context, but was unaffected by movement similarity. We employed a Bayesian analysis of our results, since in addition to avoiding some of the problems of null hypothesis significance testing (Kruschke, 2010), these analyses are able to estimate the relative strength of evidence for and against null and alternative hypotheses (Wagenmakers, Wetzels, Borsboom, \& van der Maas, 2011). Analyses were conducted in R using the BayesFactor package (Morey \& Rouder, 2015) and default parameter values. For our analyses, we used the default Cauchy prior with a scale of $\sqrt{2} / 2$, which is seen as appropriate under a broad array of situations (see Rouder, Speckman, Sun, Morey \& Iverson, 2009), following the emerging practice in this application of Bayesian techniques (e.g. de Moliere \& Harris, 2016), and compared against the null hypothesis that the conditions had no effect.

Figure 2 presents the distribution of SIQ and IOS scores across our manipulations and to the right of each plot, Bayesian credibility intervals (Kruschke, 2010) for the differences between conditions. These analyses suggest that between the social context conditions there is difference between mean SIQ and ISO scores, but no difference between the movement similarity conditions. Since there was no evidence in our analyses for an interaction between social context and movement conditions, the main effects are plotting in Figure 2.

To quantify the strength of these effects further we calculated Bayes factors. On SIQ scores, a Bayesian Type II ANOVA found very strong evidence in favour of a main effect of social context (Bayes factor: 300:1) over the null hypothesis, but evidence against a main effect of movement similarity in favour of the null hypothesis (Bayes factor 6:1) and against there being a difference between the two experiments (Bayes factor 5:1). There was also evidence against any interaction effects between conditions (Bayes factors between 3 and 4:1). A similar pattern of likelihoods was found for IOS scores. There was strong evidence in favour of an effect of social context (Bayes factor 101:1), evidence against an effect of movement condition (Bayes factor 4:1), evidence against a difference between experiments (Bayes factor 5:1), and against any interaction effects (Bayes factors between 3 and $4: 1$ ).

The conclusions we reached from Bayesian analyses were echoed by more orthodox null hypothesis testing. We ran a 2 (movement condition) x 2 (social context) x 2 (experiment) ANOVA. There was a main effect of social context on SIQ scores $\left(F(1,154)=16.58 ; p<.001, \eta_{\mathrm{p}}{ }^{2}=.1\right)$ and on IOS scores $\left.F(1,154)=13.78 ; p<.001, \eta_{\mathrm{p}}{ }^{2}=.08\right)$. But there was not a significant main effect of movement similarity on either measure, and no interaction between social context and movement conditions (all $F \mathbf{s}<1$ ).

Our manipulation check showed that there is strong evidence that participants in the joint social context felt that they were 'dancing together' more (Bayes factor 900:1), but no evidence that this was affected by movement condition
(Bayes factor 0.68:1). There was weak evidence that joint social context resulted in participants paying more attention to each other (Bayes factor 3:1), but evidence in favour of the null hypothesis and against an effect of movement condition on attention (Bayes factor 6:1).

Why was there no effect of movement similarity? One possibility, the 'attention only' account, is that participants' movements did not influence their affiliation because they simply didn't pay attention to each other, but if they had, then movement similarity would have had an effect. Logically, on this account, the more attention participants paid to each other's identical movements, the larger the increase in affiliation would be. And the more they paid attention to each other's dissimilar movements, the larger the decrease in affiliation would be. However, when we looked at the attention participants paid to each other, there was no such pattern of results. In fact, an increase in participants' attention to one another did not affect SIQ and increased it for both similar and dissimilar movements for ISO.

In the case of IOS scores, attention was positively related to affiliation in both the same movement (Bayes factor 70:1) and, crucially, in the different movement condition as well (Bayes factor 50:1). Moreover, the evidence was against a model for ISO scores with attention, movement condition and an interaction between them, over a model that just included attention (Bayes factor 12:1). In the SIQ scores, there was no evidence that attention to others was related to affiliation (Bayes factor $0.35: 1$ ), and strong evidence against the hypothesis that a model with SIQ scores, movement condition and an interaction between them was preferred over the null hypothesis (Bayes factor 50:1).

We ran correlational analyses between our measure of 'attending to others' and the two measures of affiliation overall, and splitting the data according to social context and movement, to see if those relationships changed between conditions. We calculated Zou's (2007) 95\% confidence intervals for differences between conditions. In each case, the CI encompassed 0 , suggesting that the correlations did not differ between conditions, supporting the conclusions drawn from Bayesian analysis.

## Discussion

We found two surprising results. Firstly, participants did not like each other more if they had been performing the same actions, despite the clear prediction from a host of behavioural coordination studies in the literature. Secondly, they did feel closer to each other if they had been moving together in a joint social context. This joint context was established merely by attending to a common display. If participants moved their bodies in the same way, in the same synchronised fashion, but attended to two displays a few degrees apart, then they did not feel increased affiliation towards each other.

Standard mimicry and imitation theories cannot account for these results. Their prediction is that, ceteris paribus, when participants make the same movements, their affiliation should be higher than when they are making different or
incongruent movements. However, our results did not support this prediction, and Baysian analyses strongly suggested that the likelihoods were in favour of there being no effect of movement similarity at all.

How can we explain both our finding that affiliation is dependent on social context, but also past findings that it is caused by motor mimicry? One possibility follows the associative sequence learning model, which holds that 'mirror systems' are the byproducts of learning sensorimotor contingencies in a social context (Catmur, Walsh, \& Heyes, 2009). From infancy, we perceive and perform the same actions as others in the context of rewarding social interactions (Heyes, 2001). These contingencies are learnt, and in adulthood, they continue to produce mimicry, associated with pro-social feelings. Critically, as Cook, Dickinson and Heyes (2012) showed, these sensorimotor contingencies become tied to the context in which they are learned. So, crucially for our results, the sensorimotor contingencies learned from multiple social interactions would only be re-evoked in a social context. This provides a plausible explanation of why only our joint dancing condition affected affiliation: only when participants shared a screen, they perceived themselves to be in a social situation in which their actions were contingent upon one another.

Elsewhere in the literature, it has been shown that forms of joint action and joint attention can have widespread cognitive consequences. For example, there is interference between the stimulus-response mappings of two people engaged in a Simon task together (Sebanz, Knoblich, \& Prinz, 2003). When someone believes that another person is looking at the same stimuli as them, it changes their visual attention (Richardson D., Street, Tan, Hoover, \& Ghane Cavanaugh, 2012) and memory encoding (Shteynberg, 2010; He, Lever, \& Humphreys, 2011). Pro-social feelings are also increased when two people attend to the same stimuli (Fridlund, 1991; Wolf, Launay, \& Dunbar 2015). It seems plausible that our participants who shared a screen cognitively framed their activity in a particular manner (Huhn, Potts, \& Rosenbaum, 2016), as a shared, joint activity, and from this, changes in affiliation were produced.

The interrelated roles of movement similarity and social context cannot be determined from previous results in the literature. Past experimental studies have either confounded social context with movement similarity, failed to manipulate it explicitly, or used reduced, artificial stimulus-response tasks. Indeed, our findings suggest that many past results linking motor mimicry with affiliation may have occurred, in part, because the experimental situation has established a social context in which behavioural coordination is interpreted as contingent. And we would predict that in a silent disco, if two people do not share the same music, and do not interpret their actions as shared and contingent, they will not become friends as quickly.

## Acknowledgments

We thank Aastha Chadha, Michelle Huh and Ashley Kleeberg for their help with collecting data for these experiments. This research was supported in part by an ESRC transformative research grant (ES/M000680/1) awarded to G. Orgs and D. Richardson.

## References

Aron, A., Aron, E., \& Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. Journal of Personality and Social Psychology, 63, 596612. doi: 10.1037/0022-3514.63.4.596

Catmur, C., \& Heyes, C. (2013). Is it what you do, or when you do it? The roles of contingency and similarity in prosocial effects of imitation. Cognitive Science, 37, 15411552. doi: 10.1111/cogs. 12071

Catmur, C., Walsh, V., \& Heyes, C. (2009). Associative sequence learning: The role of experience in the development of imitation and the mirror system. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364, 2369-2380. doi: 10.1098/rstb. 2009.0048

Cook, R., Dickinson, A., \& Heyes, C. (2012). Contextual modulation of mirror and countermirror sensorimotor associations. Journal of Experimental Psychology, 141, 774-787. doi: 10.1037/a0027561
Chartrand, T., \& Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology, 76, 893-910. doi: 10.1037/0022-3514.76.6.893

Chartrand, T. L., \& van Baaren, R., (2009). Chapter 5 Human Mimicry. Advances in Experimental Social Psychology, 41, 219-274. doi: 10.1016/S0065-2601(08)00405-X
de Molière, L., \& Harris, A. J. L. (2016). Conceptual and direct replications fail to support the Stake-Likelihood Hypothesis as an explanation for the interdependence of utility and likelihood judgments. Journal of Experimental Psychology: General, 145, e13-e26.
Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., \& Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Experimental Brain Research, 91, 176-180. doi: 10.1007/bf00230027
Dunbar, R. (2012). Bridging the bonding gap: the transition from primates to humans. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 1837-1846. doi: 10.1098/rstb. 2011.0217
Faul, F., Erdfelder, E., Lang, A., \& Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191. doi: 10.3758/bf03193146

Fridlund, A.J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. Journal of Personality and Social Psychology, 60, 229-240. doi: 10.1037/0022-3514.60.2.229

Gallese V., \& Goldman A. (1998). Mirror neurons and the simulation theory of mindreading. Trends in Cognitive Sciences 2, 493-501. doi: 10.1016/S1364-6613(98)012625.

He, X., Lever, A. G., \& Humphreys, G. W. (2011). Interpersonal memory-based guidance of attention is reduced for ingroup members. Experimental Brain Research, 211, 429-438. doi: 10.1007/s00221-011-2698-8
Heyes, C. (2001). Causes and consequences of imitation. Trends in Cognitive Sciences, 5, 253-261. doi: 10.1016/S1364-6613(00)01661-2

Hove, M., \& Risen, J. (2009). It's all in the timing: interpersonal synchrony increases affiliation. Social Cognition, 27, 949-960. doi: 10.1521/soco.2009.27.6.949
Huhn, J. M., Potts, C. A., \& Rosenbaum, D. A. (2016). Cognitive framing in action. Cognition, 151, 42-51. doi: 10.1016/j.cognition.2016.02.015

Kruschke, J. (2010). Bayesian data analysis. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 658-676. doi: 10.1002/wcs. 72
Lumsden, J., Miles, L., \& Macrae, C. (2014). Sync or sink? Interpersonal synchrony impacts self-esteem. Frontiers in Psychology, 5, 1-11. doi: 10.3389/fpsyg.2014.01064
Morey, R. D., \& Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.10-1. Retrieved from http://CRAN.Rproject.org/package=BayesFactor
McNeill, W. H. (1995). Keeping Together in Time: Dance and Drill in Human History. Cambridge, Massachusetts: Harvard University Press.
Miles, L. K., Griffiths, J. L., Richardson, M. J., \& Macrae, C. N. (2010). Too late to coordinate: Contextual influences on behavioral synchrony. European Journal of Social Psychology, 40, 52-60. doi:10.1002/ejsp. 721
Mukamel, R., Ekstrom, A., Kaplan, J., Iacoboni, M., \& Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. Current Biology, 20, 750-756. doi: 10.1016/j.cub.2010.02.045
Pineda, J. A. (2009). Mirror neuron systems: The role of mirroring processes in social cognition. Totowa: Humana Press. Reddish, P., Fischer, R., \& Bulbulia, J. (2013). Let's dance together: synchrony, shared intentionality and cooperation. Plos ONE, 8, 1-13. doi: 10.1371/journal.pone. 0071182

Richardson, D., Street, C., Tan, J., Kirkham, N., Hoover, M., \& Ghane Cavanaugh, A. (2012). Joint perception: gaze and social context. Frontiers in Human Neuroscience, 6, 1-8. doi: 10.3389/fnhum.2012.00194
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., \& Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic bulletin \& review, 16(2), 225-237.
Ryan, R., Koestner, R., \& Deci, E. (1991). Ego-involved persistence: When free-choice behavior is not intrinsically motivated. Motivation and Emotion, 15, 185-205. doi: 10.1007/bf00995170

Sebanz, N., Knoblich, G., \& Prinz, W. (2003). Representing others' actions: just like one's own? Cognition, 88, B11B21. doi:10.1016/S0010-0277(03)00043-X
Shteynberg, G. (2010). A silent emergence of culture: The social tuning effect. Journal of Personality and Social Psychology, 99, 683-689. doi: 10.1037/a0019573
Tarr, B., Launay, J., Cohen, E., \& Dunbar, R. (2015). Synchrony and exertion during dance independently raise pain threshold and encourage social bonding. Biology Letters, 11, 1-4. doi: 10.1098/rsbl.2015.0767
Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi : The case of psi: Comment on Bem (2011). Journal of Personality and Social Psychology, 100, 426-432. doi: 10.1037/a0022790
Wang, Y., Newport, R., \& Hamilton, A. (2010). Eye contact enhances mimicry of intransitive hand movements. Biology Letters, 7, 7-10. doi: 10.1098/rsbl.2010.0279
Wiltermuth, S. (2012). Synchronous activity boosts compliance with requests to aggress. Journal of Experimental Social Psychology, 48, 453-456. doi: 10.1016/j.jesp.2011.10.007

Wolf, W., Launay, J., \& Dunbar, R. I. (2015). Joint attention, shared goals, and social bonding. British Journal of Psychology, 107 (2), 322-337. doi: 10.1111/bjop. 12144
Yabar, Y., Johnston, L., Miles, L., \& Peace, V. (2006). Implicit behavioral mimicry: Investigating the impact of group membership. Journal of Nonverbal Behavior, 30, 97-113. doi: 10.1007/s10919-006-0010-6

# Modeling Sources of Uncertainty in Spoken Word Learning 

Matthias Hofer (mhofer@mit.edu), Roger Levy (rplevy@mit.edu)<br>Department of Brain and Cognitive Sciences, 43 Vassar St<br>Cambridge, MA 02139 USA


#### Abstract

In order to successfully learn the meaning of words such as bin or pin, language learners must not only perceive relevant differences in the speech signal but also learn mappings from words to referents. Prior work in native (Stager \& Werker, 1997) and second (Pajak, Creel, \& Levy, 2016) language acquisition has found that the ability to perceptually discriminate between words does not guarantee successful word learning. Learners fail to utilize knowledge that they can otherwise use in speech perception. To explore possible mechanisms accounting for this phenomenon, we developed a probabilistic model that infers both a word's phonetic form and its associated referent. By analyzing different versions of the model fitted to experimental results from Pajak et al. (2016), we argue that a mechanism for spoken word learning needs to incorporate both perceptual uncertainty as well as additional, taskspecific sources of uncertainty.


Keywords: word learning, rational model, probabilistic inference, phonological similarity, speech representations

## Introduction

From the perspective of a learner of English, successfully learning the meaning of novel words such as bin or pin requires the ability to perceptually discriminate between similar-sounding words. Creating distinct, non-overlapping representations of the input is necessary because the words need to be mapped onto different classes of referents. This requires perceptual sensitivities to the phonological contrasts critical for the discrimination (Pater, Stager, \& Werker, 2004). In the case of bin and pin this contrast is along the voicing dimension (phonemes $b$ and $p$ differ in voice onset time). Studies in infant native-language (L1) acquisition have shown that while these perceptual abilities are present in 14 month old infants, they do not guarantee successful word learning (Stager \& Werker, 1997; Pater et al., 2004). In a series of experiments conducted by Stager and Werker (1997), infants that were able to perceptually discriminate between two similarsounding words, such as dih and bih, failed to utilize this knowledge during word learning (experiment 4). When first habituated to label/object pairings, infants did not reliably detect when the assignment of words to objects switched (experiment 1) and they failed to notice mispronunciations when an object that had before been introduced as $d i h$ was later referred to as bih (experiment 2). The authors interpreted their findings as infants being unable to attend to fine phonetic detail during word learning and argued that it constitutes a feature of linguistic development.

The same pattern of results has more recently been demonstrated for second-language (L2) learners. A study by Pajak et al. (2016) compared the performance of subjects of two different linguistic backgrounds in a perceptual discrimination and a word learning task. To create a situation paralleling
that of L1 acquisition in infants, the researchers used a miniature language with word pairs at three levels of phonological similarity whose phonology, modeled after Polish, was novel and unfamiliar to all participants: Dissimilar words differed in multiple phonemes (e.g., tala / kenna); similar word pairs differed in one phoneme (e.g., tala / taja); and highly-similar words differed only along a single phonetic dimension, either in length (short vs. long, e.g., tala / talla) or place of articulation (alveolopalatal vs. retroflex, e.g., gotça / gotsa). In order to examine the role of L1-specific differences in task performance, Pajak and Levy (2014) collected data from both Korean speakers, who are sensitive to length contrasts but not to the alveolopalatal vs. retroflex distinction, and from Mandarin speakers, who show the opposite pattern. To ensure that subjects were naive to the stimuli used in the experiment, the researchers tested separate groups of subjects on the perceptual discrimination and the word learning task.

Similarly to results from Stager and Werker (1997), the study found that the ability to perceptually discriminate similar-sounding words in the perceptual task did not successfully translate to the word learning task on a group level, nor did subjects' L1-specific perceptual advantages. Taken together, these findings suggest that the difficulty in learning similar-sounding words, especially during early lexical acquisition, is a general property of learning rather than a developmental stage (Pajak et al., 2016; Perfors \& Dunbar, 2010). At present, little is known about the learning mechanisms that give rise to these difficulties. Here we seek to provide an account of such a mechanism in the form of a probabilistic model developed with the goal of reproducing the results from Pajak et al. (2016)'s original study. While we are not aware of any computational modeling work, the exists prior work addressing these issues on a conceptual level. The failure to utilize perceptual knowledge during word learning has previously been attributed to increased cognitive load (Werker, Fennell, Corcoran, \& Stager, 2002; Stager \& Werker, 1997). While discrimination only requires storage and comparison of perceptual input in phonological short-term memory, word learners must additionally attend to the referent stimulus and integrate label/referent information over the course of learning to infer probable associations. This lowers the resolution of auditory processing (Mattys \& Palmer, 2015), which contributes to the failure of distinctly representing similar-sounding input. We will explore these verbal theories by analyzing which of the proposed components are necessary to account for the observed effects in the study by Pajak et al. (2016), which we will briefly describe in the next section.

## Pajak et al. (2016)'s experiment

In a between-subjects design, ninety subjects, approximately equally split into speakers of Korean and Mandarin, participated in either a perceptual discrimination or a word learning task. Stimuli consisted of 16 bisyllabic consonant-vowel-consonant-vowel (CVCV) nonce words, split into similarity classes as described above. ${ }^{1}$ Perceptual discrimination was tested in an ABX task. Subjects listened to three consecutive words, e.g., talla ${ }_{[A]}$, taja $_{[B]}$, and talla ${ }_{[X]}$, and had to decide which of the first two words sounded more similar to the last one. In the word learning task, each of the 16 labels was associated with a single referent in the form of a visual image and participants' goal was to learn which words were associated with which referents. The experiment consisted of four training blocks (each with 128 trials) and four interleaved testing blocks (each with 64 trials). In each trial, two pictures were presented side by side, corresponding to the referents for labels A and B. Similarly to the ABX task, subjects then heard a label X and had to decide which referent it belonged to. Error feedback was provided to the subjects during the training phase but not during the test phase. The stimulus triplets used in the discrimination task and in the test phase of the word learning task were identical, which made it possible to compare accuracy for triplets across the two tasks.

## Computational Model

A computational analysis of spoken word learning must take into account the goal of the computation, the information available to the learner, and show how this information maps onto appropriate behavioral responses (Anderson, 1990). We suggest that learning novel words requires the learner to perform statistical inference on at least two distinct levels. While the ultimate goal of learning is to infer concepts, or label/referent associations, from a stream of observations, the learner must concurrently infer the label's phonetic form from the acoustic input, since it is not explicit in the speech signal. These two layers of inference give rise to a hierarchical probabilistic model, visually depicted in Figure 1, which we used to model spoken word learning and, using a variation of the generative process, model results from the perceptual discrimination tasks. Model behavior is influenced by three distinct factors: perceptual noise, which affects both discrimination and word learning when processing speech input, taskspecific factors that lower the resolution of auditory representations of speech sounds during word learning (Mattys \& Palmer, 2015), and overall memory capacity.

## Formal characterization of the model

Each word and its corresponding referent define a concept, denoted $c$. To simplify our analysis, we assume that the referent stimulus is observed unambiguously. In our model, observing the referent stimulus is therefore identical to observ-

[^103]ing the concept directly, because of the one-to-one relation between referents and concepts. The a priori probability of choosing any concept is uniform. Corresponding labels are then sampled from the conditional probability $p(\mathbf{l} \mid c)$, whose probability mass is uniform across all possible labels in the language $\mathcal{L}$ and zero otherwise.
\[

\operatorname{Pr}(\mathbf{l} \mid c)= $$
\begin{cases}\frac{1}{N} & \text { if } \mathbf{l} \in \mathcal{L}  \tag{1}\\ 0 & \text { otherwise }\end{cases}
$$
\]

Label $\mathbf{l}$ is a sequence of phonemes of the form $p_{1} p_{2} p_{3} p_{4}$ composed of pre-specified consonant and vowel primitives. Phonemes are represented mathematically as multivariate Gaussian distributions in one of two separate (phonetic) feature spaces, one for consonants and one for vowels. Following prior approaches to the representation of speech sounds (Richter, Feldman, Salgado, \& Jansen, 2016; Bailey \& Hahn, 2005), the feature dimensions of these phonetic spaces correspond to subsegmental features such as manner, place, length, and voicing, or to the first two vowel formants. Distributions are centered around a fixed category mean $\bar{\mu}\left[p_{i}\right]$, a vector of means indexed by the corresponding phoneme. Covariance matrices $\bar{\Sigma}_{w}[i]$, one shared across vowels and one across consonants, are indexed by $i$ only (corresponding to whether the phoneme is of type C or V ).

Conditioned on a choice for label l, we can generate a sequence of phones $s_{1} s_{2} s_{3} s_{4}$, which can be seen as its discrete and noisy realizations of the label's phonemes:

$$
\begin{align*}
\operatorname{Pr}\left(s_{i} \mid p_{i}\right) & =\mathcal{N}\left(\bar{\mu}\left[p_{i}\right], \bar{\Sigma}_{w}[i]\right)  \tag{2}\\
\operatorname{Pr}(\mathbf{s} \mid \mathbf{I}) & =\operatorname{Pr}\left(s_{1} s_{2} s_{3} s_{4} \mid p_{1} p_{2} p_{3} p_{4}\right)  \tag{3}\\
& =\operatorname{Pr}\left(s_{1} \mid p_{1}\right) \operatorname{Pr}\left(s_{2} \mid p_{2}\right) \operatorname{Pr}\left(s_{3} \mid p_{3}\right) \operatorname{Pr}\left(s_{4} \mid p_{4}\right)
\end{align*}
$$

The covariance matrix that determines variability in the realization of speech sounds is specified through the following scalar-matrix-vector multiplication:

$$
\begin{equation*}
\bar{\Sigma}_{\text {task }}[i]=\alpha_{\text {task }} \mathbf{I} \bar{v}[i]^{\text {population }} \tag{4}
\end{equation*}
$$

The scalar $\alpha_{\text {task }}$ allows us to reflect task-specific sources of uncertainty (word learning vs. discrimination). We note that the parameter for word learning, $\alpha_{w}$ can be written as the product of a perceptual 'baseline' acuity parameter for the discrimination task times a constant factor: $\alpha_{w}=c \alpha_{d}$. Assuming pairwise independence across all feature dimensions, the covariance matrix is fully specified by its diagonal elements, encoded in the population-specific diagonal vector $\overline{\mathrm{V}}[i]^{\text {population }}$. Phonetic acuity along those feature dimensions is inversely proportional to variance: the higher phonetic acuity, the lower the variance. This allows us to model differences in L1 background (Korean vs. Mandarin) with respect to perceptual sensitivities along these feature dimensions. For example, for Korean speakers:


Figure 1: Graphical representation of the word learning model. Circles indicate random variables (variables shaded in gray are observed during learning); squares indicate fixed model parameters. To simplify our representation, the model does not include a referent node, which is deterministically generated by sampling from the concept.

$$
\begin{aligned}
\overline{\mathrm{v}}[1]^{\text {Korean }} & =\overline{\mathrm{v}}[3]^{\text {Korean }}=\left(\tau_{F 1_{K}}{ }^{-1} \tau_{F 2_{K}}{ }^{-1}\right)^{T} \\
\overline{\mathrm{v}}[2]^{\text {Korean }} & =\overline{\mathrm{v}}[4]^{\text {Korean }}= \\
& \left(\tau_{\text {length }_{K}}{ }^{-1} \tau_{\text {place }_{K}}{ }^{-1} \tau_{\text {voicing }_{K}}{ }^{-1} \tau_{\text {manner }_{K}}{ }^{-1}\right)^{T}
\end{aligned}
$$

The vowel feature space consists of the first two formants F1 and F2. Consonant space consists of the dimensions voicing, place, manner, and length (Bailey \& Hahn, 2005). All acuity parameters are set to 1 (corresponding to a unit Gaussian variance), except for $\tau_{\text {length }}$ and $\tau_{\text {place }}$, which are population-specific free parameters in the model. As a simple approximation, means in $\bar{\mu}\left[p_{i}\right]$ are evenly spaced across perceptual space. Along each phonetic dimension, we defined a number of subsegmental features (e.g., 'voiced' and 'unvoiced' along the voicing dimension). Phonetic category means can then be written as the vectors composed of these features (the mean of phoneme $f$, for instance, is represented as [unvoiced, labial, fricative, short]). Although not fully accurate in its details, the coarse grained nature of this setup is sufficient with respect to the word pairs used in Pajak et al. (2016)'s experiment. ${ }^{2}$

Word learning model In word learning, subjects engage in consecutive training and test blocks. Each training trial $t$ consist of an observed label/referent pair $\left\{\mathbf{s}_{t}, c_{t}\right\}$. For simplicity we assume that learners discard the negative, second

[^104]exemplar presented to them and only learn from the positive pairing. The learner's goal is to infer probable associations between referents $c$ and labels $\mathbf{I}$, in other words, compute the posterior probability over labels given the observed stimulus and referent $\operatorname{Pr}\left(\mathbf{l}_{t} \mid \mathbf{s}_{t}, c_{t}\right)$ according to:
\[

$$
\begin{equation*}
\operatorname{Pr}\left(\mathbf{l} \mid \mathbf{s}, c, \bar{\Sigma}_{w}\right)=\frac{\operatorname{Pr}\left(\mathbf{s} \mid \mathbf{l}, \bar{\Sigma}_{w}\right) \operatorname{Pr}(\mathbf{l} \mid c)}{\sum_{\mathbf{l}} \operatorname{Pr}\left(\mathbf{s} \mid \mathbf{l}, \bar{\Sigma}_{w}\right) \operatorname{Pr}(\mathbf{l} \mid c)} \tag{5}
\end{equation*}
$$

\]

The output of this computation is then used as a prior for the next trial. To model the difficulty of integrating multiple memory traces over time, one simple approach is to assume an upper bound on the number of memory traces that can be integrated, denoted $m_{c}$. We formalized this intuition by discarding samples from trials $t \geq m_{c}$ (no further updating of probabilistic representations occurs).

After computing $\operatorname{Pr}\left(\mathbf{l} \mid \mathbf{s}_{1}, \ldots, \mathbf{s}_{T}, c_{1}, \ldots, c_{T}, \bar{\Sigma}_{w}\right)$ for the training block, in the test phase, the learner compares two alternative tuples $\left\{c_{A}, \mathbf{s}_{X}\right\}$ and $\left\{c_{B}, \mathbf{s}_{X}\right\}$ to assess which referent is more probable under $\mathbf{s}_{X}$. This is achieved by computing $\operatorname{Pr}\left(c_{A} \mid \mathbf{s}_{X}, \bar{\Sigma}_{w}\right)$ and $\operatorname{Pr}\left(c_{B} \mid \mathbf{s}_{X}, \bar{\Sigma}_{w}\right)$ by integrating over $\mathbf{l}$ :

$$
\begin{equation*}
\operatorname{Pr}\left(c \mid \mathbf{s}, \bar{\Sigma}_{w}\right)=\frac{\sum_{\mathbf{l}} \operatorname{Pr}\left(\mathbf{s} \mid \mathbf{l}, \bar{\Sigma}_{w}\right) \operatorname{Pr}(\mathbf{l} \mid c) \operatorname{Pr}(c)}{\sum_{c} \Sigma_{\mathbf{l}} \operatorname{Pr}\left(\mathbf{s} \mid \mathbf{l}, \bar{\Sigma}_{w}\right) \operatorname{Pr}(\mathbf{l} \mid c) \operatorname{Pr}(c)} \tag{6}
\end{equation*}
$$

Discrimination model In the discrimination task, subjects perceive a stimulus triplet $\left\{\mathbf{s}_{A}, \mathbf{s}_{B}, \mathbf{s}_{X}\right\}$ and decide whether X is more similar to A or to B . We hypothesize that subjects use the generative process outlined above to judge similarity, where they independently determine the likelihood that the stimuli were sampled from two alternative generative models (Tenenbaum \& Griffiths, 2001), described in the following:

$$
\begin{align*}
& \operatorname{Pr}\left(\mathbf{s}_{A}, \mathbf{s}_{B}, \mathbf{s}_{X} \mid \mathbf{l}_{1}, \mathbf{l}_{2}\right)=  \tag{7}\\
& \quad \sum_{\mathbf{l}_{1}}\left[\operatorname{Pr}\left(\mathbf{s}_{A} \mid \mathbf{l}_{1}\right) \operatorname{Pr}\left(\mathbf{s}_{X} \mid \mathbf{l}_{1}\right) \operatorname{Pr}\left(\mathbf{l}_{1}\right)\right] \sum_{\mathbf{l}_{2}}\left[\operatorname{Pr}\left(\mathbf{s}_{B} \mid \mathbf{l}_{2}\right) \operatorname{Pr}\left(\mathbf{l}_{2}\right)\right]  \tag{8}\\
& \operatorname{Pr}\left(\mathbf{s}_{A}, \mathbf{s}_{B}, \mathbf{s}_{X} \mid \mathbf{l}_{1}, \mathbf{l}_{2}\right)= \\
& \quad \sum_{\mathbf{l}_{1}}\left[\operatorname{Pr}\left(\mathbf{s}_{A} \mid \mathbf{l}_{1}\right) \operatorname{Pr}\left(\mathbf{l}_{1}\right)\right] \sum_{\mathbf{l}_{2}}\left[\operatorname{Pr}\left(\mathbf{s}_{B} \mid \mathbf{l}_{2}\right) \operatorname{Pr}\left(\mathbf{s}_{X} \mid \mathbf{l}_{2}\right) \operatorname{Pr}\left(\mathbf{l}_{2}\right)\right]
\end{align*}
$$

The likelihood $\operatorname{Pr}(\mathbf{s} \mid \mathbf{l})$ is the same as in Equation 3 but with covariance matrix $\bar{\Sigma}_{d}[i]$ specific to the perceptual discrimination task.

Response probability Both experimental paradigms use a two-alternative forced choice task (2-AFC) to assess subjects' knowledge. Subject either compare two posterior probabilities over concepts given labels (word learning task) or the likelihoods that the stimulus triple was generated by one of two alternative generative models (discrimination task). In


Figure 2: Comparison of the $M_{+} A_{+}$model to experimental results for the L1 Korean population (top) and for the L1 Mandarin population (bottom). Error bars are standard errors. Accuracy scores are percentage correct in the discrimination task and during the test phase of the word learning task.
both cases, choices are modeled using a Bernoulli distribution and the probability of choosing one alternative over the other is computed using Luce's choice rule (Luce, 1959). Response parameter $\beta$ controls the stochasticity of responses.

## Results

We fitted the model to aggregate subject data from Pajak et al. (2016)'s word learning and perceptual discrimination tasks. Free model parameters included task-specific phonetic acuity for word learning ( $\alpha_{w}$ ) and for discrimination $\left(\alpha_{d}\right)$, four population-specific acuity parameters ( $\tau_{\text {length }_{K}}, \tau_{\text {length }_{M}}$, $\tau_{\text {place }_{K}}, \tau_{\text {place }_{M}}$ ), the response parameter ( $\beta$ ), as well as the memory capacity parameter ( $m_{c}$ ).

Table 1: Results from fitting four versions of our model to the experimental data. Fits are quantified using the product of RMSEs to the word learning and discrimination data across the two speaker populations.

| Model | n.p. | w.l. | w.l.b. | disc. | all |
| :--- | :---: | :---: | :---: | :---: | :---: |
| $M_{-} A_{-}$ | 6 | 0.089 | 0.138 | 0.179 | 0.407 |
| $M_{-} A_{+}$ | 7 | 0.029 | 0.140 | 0.017 | 0.186 |
| $M_{+} A_{-}$ | 7 | 0.076 | 0.098 | 0.172 | 0.347 |
| $M_{+} A_{+}$ | 8 | 0.025 | 0.099 | 0.018 | 0.142 |

To assess which model components are necessary to account for the experimental data, we fitted four alternative ver-
sions of the model that were composed out of two binary factors: the presence $(+)$ or absence $(-)$ of memory constraints $m_{c}(\mathrm{M})$ and the presence of task-specific, non-perceptual uncertainty in the form of separate $(+)$ or shared $(-)$ perceptual acuity parameters across tasks (A), where in the case of shared parameters: $\alpha_{w}=\alpha_{d}$. We also considered separate response rule parameters $\beta$ for word learning and discrimination but found that the improvements were only minimal.

The models were fitted to the data by minimizing the product of six separate error terms. For each group of L1 speakers (Mandarin vs. Korean), we calculated the root mean squared error (RMSE) between model predictions and experimental results, resulting in three separate error terms for (i) overall discrimination accuracy across trial types [disc.], (ii) overall word learning accuracy across trial types [w.l.], and (iii) word learning accuracy across blocks and trial types [w.l.b.]. For each of the four models, Table 1 shows the RMSE for these three scores (averaged across Korean and Mandarin speakers) and their sums [all]. Column [n.p.] indicates the fitted model's number of free parameters.

## The necessity of separate acuity parameters

Pajak et al. (2016)'s main finding was that the difference in accuracy between tasks was mediated by similarity. In other words, performance takes a greater hit from increased perceptual similarity in the word learning task compared to discrimination. The study also found that L 1 -specific perceptual


Figure 3: Figure (a) shows how the $M_{-} A_{+}$model fails to account for the time course of learning. Figure (b) and (c) show comparisons of the $M_{+} A_{+}$model to L1 Korean speakers and L1 Mandarin speakers. Error bars indicate standard errors.


Figure 4: Results for the $M_{-} A_{-}$model (top) and for the $M_{-} A_{+}$model (bottom; both Korean only; results are qualitatively similar for Mandarin speakers)
advantages that are manifest in the discrimination task cannot be utilized in word learning. Model variants $M_{-} A_{-}$and $M_{+} A_{-}$, missing the additional acuity parameter $\alpha_{w}$, were not able to account for these observations. Figure 4 (top) illustrates this point by showing overall results for Korean speakers. One key observation for the $M_{-} A_{-}$model is that, when sharing a single perceptual acuity parameter between the two tasks, the original pattern of findings reverses and the discrimination model performs worse than the word learning model. ${ }^{3}$ The reasons for this are twofold. All other things being equal, there is more uncertainty in the generative process for discrimination (see Equation 7 and 8) than there is in word learning. In discrimination, subjects need to infer the phonetic form of three auditory stimuli, compared to a

[^105]single stimulus in word learning. Moreover, the word learning model can profit from additional information in the form of label/referent representations, gradually sharpening over the course of learning. The fact that perceptual uncertainty alone (in the form of a shared acuity parameter across models) cannot account for the superiority of discrimination performance over word learning suggests that word learning is influenced by additional sources of uncertainty. The bottom of Figure 4, which depicts results for the $M_{-} A_{-}$model, illustrates that adding an additional phonological acuity parameter that is specific to word learning is sufficient to account for both discrimination as well as word learning results.

## Accounting for the time course of word learning

In Pajak et al. (2016), word learning performance only improved over a certain number of trials, resulting in learning curves to asymptote after roughly the second learning block. Reproducing this pattern while simultaneously accounting for the results presented above was an important aspect of our modeling efforts. While model $M_{-} A_{+}$provides an almost ideal fit to aggregate results from discrimination and word learning (see Figure 4), underlining the importance of incorporating acuity factor A into the model, it fails to capture the time course of learning (see Figure 3a).

The only model that fit the entire range of empirical findings was the $M_{+} A_{+}$model. Figure 2 shows that the model is a good fit, both qualitatively as well as quantitatively, to aggregate accuracy scores for both Korean and Mandarin L1 speakers. In particular, simulated data successfully reproduce the lack of L1-specific advantages in word learning compared to discrimination. Figures 3 b and 3 c show that the added memory constraint allows the model to better account for the shape of the learning curve. Models without this component are not able to reproduce this pattern. Table 1 further shows that adding such memory constraints also slightly improves fits to the aggregate word learning data [w.l.] compared to the same model where they are absent.

## Discussion

Recent work by Pajak et al. (2016) suggests that the difficulty of learning label/referent associations for similar-sounding words is a general feature of learning rather than a developmental stage unique to infancy. In working towards a computational theory that could account for this phenomenon, we developed a probabilistic model capable of learning label/referent pairs while at the same time inferring the label's phonetic form. We fitted and compared four versions of the model to data from Pajak et al. (2016), contrasting different factors that are thought to influence performance.

We found that, besides structural differences in the way the generative models for word learning and for discrimination are set up, a single multiplicative factor operating on perceptual uncertainty was sufficient to account for the major differences between perceptual discrimination and word learning. A second additional factor, representing long-term memory constraints, was only necessary to account for the time course of learning.

## Task-related sources of uncertainty

Conceptually, the acuity factor combines sources of uncertainty unique to the word learning task, such as attention to the referent stimulus and the encoding of label/referent exemplars over the course of learning. An interpretation broadly consistent with our model and with previous work (Stager \& Werker, 1997; Mattys \& Palmer, 2015) is that, although originating from post-perceptual sources, the locus of this added uncertainty is perception itself, operating through lowering attention to phonetic detail. On this view, the model's discrimination acuity parameter can be interpreted as representing various sources of perceptual uncertainty, ranging from the transduction of the speech signal at the periphery to phoneme recognition. Word learning-specific sources of uncertainty can be interpreted as a multiplicative factor on perceptual uncertainty, which, multiplied together, constitute the model's word learning acuity parameter. This added factor also accounts for the finding that L1-specific perceptual advantages cannot be utilized in word learning. The overlap of highly-similar word pairs in perceptual feature space is so large that potential advantages along the length and place feature dimensions are washed out.

Another important insight comes from models that lack this separate acuity parameter, which suggest that the discrimination is actually harder than word learning. This makes sense when considering that the generative model for the discrimination task must infer the phonetic form of three stimuli instead of a single stimulus. As a consequence, perceptual uncertainty affects the discrimination task more severely. In the absence of other factors that independently operate on the generative model for word learning, this leads to relative performance benefits in the word learning task.

## Memory constraints

While distinguishing between two major sources of uncertainty might alone be sufficient to account for time-averaged
results, it is not enough to account for the time course of learning in Pajak et al. (2016), which showed that learning stagnates after the second training block. The fact that these performance deficits are specific to word learning suggests that they are due to memory-related processes. We found that incorporating capacity constraints in the form of an upper bound on memory was necessary to fully account for the observed effects.

## Conclusion

Our model is a first step in addressing the question of what are the factors that make the learning of similar-sounding words hard. In particular, the model is consistent with the original explanation given by Stager and Werker (1997); Werker et al. (2002). According to this view, word learning is an inherently hard information processing problem and the difficulties of learning similar-sounding words are a consequence of optimally distributing limited resources across the perceptual and memory-related processes involved in learning.

## References

Anderson, J. R. (1990). The adaptive character of thought. Psychology Press.
Bailey, T. M., \& Hahn, U. (2005). Phoneme similarity and confusability. Journal of Memory and Language, 52(3), 339-362.
Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York: Wiley.
Mattys, S. L., \& Palmer, S. D. (2015). Divided attention disrupts perceptual encoding during speech recognition. The Journal of the Acoustical Society of America, 137(3), 1464-1472.
Pajak, B., Creel, S. C., \& Levy, R. (2016). Difficulty in learning similar-sounding words: a developmental stage or a general property of learning? J Exp Psychol Learn Mem Cogn., 42(9), 1277-99.
Pajak, B., \& Levy, R. (2014). The role of abstraction in non-native speech perception. Journal of phonetics, 46, 147-160.
Pater, J., Stager, C., \& Werker, J. (2004). The perceptual acquisition of phonological contrasts. Language, 384-402.
Perfors, A., \& Dunbar, D. (2010). Phonetic training makes word learning easier. Congnitive Science Proceedings, 2010.

Richter, C., Feldman, N. H., Salgado, H., \& Jansen, A. (2016). A framework for evaluating speech representations. Congnitive Science Proceedings, 2016.
Stager, C. L., \& Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature, 388(6640), 381-382.
Tenenbaum, J. B., \& Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. Behavioral and brain sciences, 24(04), 629-640.
Werker, J. F., Fennell, C. T., Corcoran, K. M., \& Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. Infancy, 3(1), 1-30.

# Knowledge of Cross-Linguistic Semantic Diversity Reduces Essentialist Beliefs about Categories 

Kevin Holmes<br>Colorado College

Erin Luby<br>Colorado College

Sarah Husney
Colorado College


#### Abstract

The words of different languages partition the world in strikingly different ways. Yet many people are unaware of such differences, believing that some of the words of their native language pick out discrete categories based in nature. We investigated whether knowledge of cross-linguistic semantic diversity-putatively inherent to bilingualism-can reduce such essentialist beliefs. In three experiments, we found (a) that bilinguals were less likely than monolinguals to judge membership for animal categories in essentialist terms, (b) that explicit exposure to cross-linguistic semantic diversity, independent of bilingualism, yielded similar effects, and (c) that this manipulation reduced essentialist beliefs about social categories as well. Together, our findings suggest that learning about how languages differ in their semantic systems-a form of metalinguistic knowledge-can lead people to think about categories more flexibly. Implications for research on language and thought, and for ameliorating the negative consequences of social essentialism, are discussed.


# Decisions based on verbal probabilities: Decision bias or decision by belief sampling? 

Hidehito Honda (hitohonda.02@gmail.com)<br>Graduate School of Arts and Sciences, The University of Tokyo<br>3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan<br>Toshihiko Matsuka (matsukat@muscat.L.chiba-u.ac.jp)<br>Department of Cognitive and Information Science, Chiba University<br>1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan


#### Abstract

We examined decisions based on verbal probability phrases, such as "small chance," "likely," or "doubtful" (we call these phrases verbal probabilities). Verbal probabilities have communicative functions called directionality and can be categorized into positive (e.g., "likely" or "probable") or negative (e.g., "unlikely," "doubtful") phrases in terms of their directionality. Previous studies have shown that the directionality of phrases affects decisions. Although such decisions seem biased, we argue that they are not. We hypothesize that since a speaker has the option to choose the directionality used during communication, the selected directionality becomes relevant information to a decision maker, and is taken into account in making decisions. We modeled these processes using the Decision by Belief Sampling (DbBS) model. We found that the observed data could be well explained by our hypothesis, and that the DbBS model could be one of the best potential models for decisions based on verbal probability information.


Keywords: Verbal probabilities; decisions based on verbal probabilities; directionality; decision by belief sampling.

## Introduction

In the research on judgment and decision making, topics pertaining to probabilities, such as probability judgment and decisions based on probability information, have been some of the most studied. In the present study, we shall discuss decision making based on different kinds of probability information.

Probability information can be expressed in various forms. The most basic of these expressions is numerical probability, such as " $20 \%$." Probability information is also expressed with verbal phrases such as "it is likely," "it is doubtful," or "it is certain. ${ }^{11}$ In the present study, we examined decisions based on verbal probabilities and analyzed how the difference in expressions affected cognitive processes.

Particularly, we focus on the communicative functions of verbal probabilities. Teigen and Brun (1995) showed that verbal probabilities have communicative functions, called

[^106]directionality, which change the listeners' focus. Verbal probabilities can be categorized into positive or negative phrases in terms of their directionality. Positive phrases (e.g., "small chance," "likely," "certain") make listeners focus on the occurrence of uncertain events. In contrast, negative phrases (e.g., "unlikely," "doubtful," "uncertain") make listeners focus on the non-occurrence of uncertain events. Previous studies have shown that the directionality affected decision making. Here, we introduce one of the most intriguing studies, the "Marianne study" (Study 1) in Teigen and Brun (1999). This experiment involved a task describing the probable effectiveness of a treatment with either a positive phrase ("there is some possibility that the treatment will be helpful in her case") or a negative phrase ("it is quite uncertain that the treatment will be helpful in her case"). Participants rated how likely they would recommend this treatment to a patient (Marianne) based on these phrasings, using a 4-point scale (1: absolutely yes, 4: absolutely not). Numerical translations for positive and negative phrases answered by different participants were $31.7 \%$ and $31.3 \%$. Based on these results, the two phrases should have conveyed highly similar degrees of certainty for the effectiveness of treatment. However, the participants gave highly different decision ratings depending on the probability expressions. Mean ratings for positive and negative phrases were, 1.78 and 2.78 , respectively (when scores of 1 or 2 were jointly regarded as "Yes" decisions, the proportions of "Yes" decisions for the two phrases were $90.6 \%$ and $32.4 \%$, respectively).

Ostensibly, the results in Teigen and Brun (1999) may have suggested a decision bias produced by the difference in probability expression, perhaps another form of the framing effect (Teigen \& Brun, 2003). In the following sections, we shall claim that the effect of different probability expressions on decisions is not a decision bias, and that such decisions derive from a decision maker's inferences regarding background information based on speaker's choice of directionality. Furthermore, we model this decision processes using a Decision by Belief Sampling model.

## Speaker's choice of directionality and listener's inferences in communication

In communication, speakers will select an expression according to situational factors. McKenzie and Nelson (2003) and Sher and McKenzie $(2006,2008)$ argued that a speaker's reference point affects her/his choice of expression when conveying quantitative information, such as the amount of water in a glass (e.g., "half full" or "half empty"). In a task examining this phenomenon, it was found when the glass (a 500 ml capacity) had 250 ml of water, participants used the description of "half full" when the glass had previously 0 ml of water more often than when the glass had previously 500 ml of water. It was also found that a listener could infer the speaker's original reference point (e.g., the amount of water originally in the glass before more was added or removed) based on the selected expression. Honda and Yamagishi (2017) found analogous tendencies in communication using verbal probabilities. Imagine trying to convey that an uncertain event had a $50 \%$ chance of occurrence using verbal probabilities. Honda and Yamagishi (2017) showed that when a speaker's prior expectation of the event occurrence was lower (higher) than $50 \%$, they tended to prefer positive (negative) phrases. Honda and Yamagishi (2017) also found that listeners could infer the speaker's expectation based on the phrases used. When a positive phrase was presented, listeners tended to infer that speaker's prior expectation of the probability was lower than when a negative phrase was presented.

These findings indicate that the selected phrase implicitly conveys important information for decision making. For example, given that a speaker follows the above regularity, $50 \%$ conveyed by a positive (negative) phrase implies "good (bad)" situation relative to the speaker's expectation. Thus, the findings of Teigen and Brun (1999) (i.e., participants tended to recommend a treatment conveyed by a positive phrase more than one conveyed by a negative phrase) suggest that in making decisions, participants took into account the relevant information (i.e., relatively "good" or "bad" situations) implied by the directionality chosen.

## Model of decision making based on verbal probabilities: Decision by Belief Sampling

As described above, previous findings suggest that different phrases implicitly convey information about the relative status of the decision situation, and people utilize this information in making decisions. In the present study, we model such decision processes based on the Decision by Sampling model (DbS, Stewart, 2009; Stewart, Chater, \& Brown, 2006). In the DbS model, subjective attribute values are constructed by a series of binary, ordinal comparisons to a sample of attribute values that reflect the immediate decision context and real-world distribution. The subjective value for a target is calculated as follows:

$$
\begin{equation*}
\mathrm{r}=\frac{R-1}{N-1} \tag{1}
\end{equation*}
$$

where $\mathrm{r}(0 \leq \mathrm{r} \leq 1)$ denotes the subjective value for a target, and R denotes the rank of the target within the decision


Figure 1. Summaries of DbBS. (A) Probabilistic belief regarding an uncertain event. This is represented by the density function of the beta distribution. (B) Mean of belief. (C) Entropy of belief. (D) Skewness of belief. (E) Subjective value in DbBS. This is represented with the cumulative distribution function of the beta distribution.
sample of N items. In this model, if the decision sample differs, $r$ varies in the relationship between $R$ and the decision sample. Imagine the subjective value for $40 \%$. When decision samples are $10 \%, 20 \%, 30 \%, 30 \%$, and $50 \%$, the subjective value is $\mathrm{r}=(5-1) /(6-1)=0.8$. In contrast, in decision samples of $20 \%, 30 \%, 70 \%, 80 \%$, and $90 \%$, the subjective value is $r=(3-1) /(6-1)=0.4$. That is, even when the target has the same attribute value, the subjective value varies depending on decision samples. Previous studies have shown that decision samples affect an evaluation of the target value and the evaluation for the same target varies depending on the samples (e.g., Stewart, Chater, Stott \& Reimers, 2003; Stewart, Reimers, \& Harris, 2014).

In the present study, we propose a decision model, Decision by Belief Sampling (DbBS), based on the DbS model. Figure 1 summarizes DbBS. Here, we introduce
basic two assumptions: Firstly, the decision maker (DM) refers to memory samples in making decisions, and these samples represent the DM's probabilistic belief of event occurrence. For example, imagine the probable success rates of medical procedures for both a serious disease and appendicitis, respectively. Generally, people believe that the probability of success in treating a serious disease is low compared to the probable success of treating appendicitis (Honda \& Matsuka, 2014). We assume that the DM refers to memory samples according to her/his probabilistic belief. We represent these beliefs using beta distributions. Figure 1 (A) shows four examples of a DM's subjective beliefs regarding uncertain events. We can discuss the features of probabilistic belief based on its statistical characteristics such as mean, entropy (i.e., uncertainties about successes or failures), and skewness of beta distributions (see Figure 1 (B), (C), and (D)). Example 1 represents the belief such that an event will occur or not with high uncertainty and without skewness. Likewise, in Examples 2 and 4, the DM has the belief such that the event will happen with low or high probability with relatively low uncertainty and positive or negative skewness. Example 3 represents the belief that an event has around $50 \%$ of occurrence with low uncertainty and without skewness. Thus, beta distributions can represent extensive kinds of beliefs about uncertain events. Secondly, we assume that a subjective value for a target is constructed by the comparison between the target value and memory samples. Figure 1 (E) shows subjective values calculated by the DbBS model (i.e., equation (1)). Given that beta distributions represent beliefs about uncertain events, subjective values correspond to values in the cumulative distribution functions (CDF) of beta distributions. As is apparent, depending on the beliefs, the subjective values differ even for the same target probability. One of the most notable features in the DbS (or DbBS ) model is that subjective values are highly affected by the skewness of distributions in decision (or memory) samples (Brown \& Matthews, 2011). Therefore, subjective values highly differ between beliefs with high probability and those with low probability (see Examples 2 and 4 in Figure 1).

We believe that the DbBS model can clarify the following points regarding decisions based on verbal probabilities. First, the DbBS model can clarify the implicit assumptions (i.e., beliefs about uncertain events) people have in making decisions. Although Honda and Yamagishi (2017) showed that listeners have different assumptions depending on the directionality of verbal probabilities, it remains an empirical question whether people have such assumptions in making decisions. Using the DbBS model, we can examine this question. Second, the DbBS model will provide a new perspective on phenomena regarding decisions based on verbal probabilities. For example, we can discuss whether the influence of directionality on decisions reflects decision bias.

According to previous findings (Honda \& Yamagishi, 2017) and the assumptions of the DbBS model, our hypothesis is as follows: DMs refer to different memory
samples depending on the directionality of verbal probabilities because the selected directionality become relevant information to DMs. In particular, DMs refer to memory samples with lower probability when presented with positive phrases than when presented with negative phrases. As a result, decision patterns differ between positive and negative phrases. For example, even when DMs think that a probability of an uncertain event is $30 \%$ when presented with a verbal probability, the subjective value for the probability will be higher when presented with a positive phrase than a negative phrase, because DMs have lower memory samples (see Examples 2 and 4 in Figure 1 (A) and (E)).

## Behavioral experiment

In order to examine the above hypothesis, we conducted behavioral experiments about decisions based on verbal probabilities.

## Method

Participants Japanese undergraduates $(\mathrm{N}=60)$ participated as part of their course work.
Tasks, materials, and procedure We conducted two tasks: a decision task and a task measuring the membership function for verbal probabilities. The decision task was based on the Marianne study (Study 1) in Teigen and Brun (1999). The cover story was as follows:

Your friend has periodically been suffering

Table 1. Verbal probabilities used in the experiment.

| Verbal probabilities | $\mathrm{M}_{\text {peak }}$ | $\mathrm{SD}_{\text {peak }}$ |
| :--- | :--- | :--- |
| positive phrases |  |  |
| It is almost certain that * | 0.957 | 0.037 |
| There is a good chance that * | 0.779 | 0.126 |
| It is possible that * | 0.418 | 0.167 |
| It is likely that * | 0.540 | 0.164 |
| There is a small possibility that * | 0.346 | 0.129 |
| There is some possibility that * | 0.232 | 0.116 |
| There is a slight hope that * | 0.121 | 0.115 |
| There is a tiny hope that * | 0.074 | 0.097 |
| negative phrases |  |  |
| There are minor concerns that * | 0.602 | 0.167 |
| It is quite doubtful that * | 0.494 | 0.188 |
| It is not certain that * | 0.532 | 0.165 |
| It is uncertain whether * | 0.466 | 0.178 |
| It is quite unlikely that * | 0.433 | 0.141 |
| There is little hope that * | 0.177 | 0.088 |
| It is unlikely that * | 0.137 | 0.103 |
| It is almost impossible that * | 0.027 | 0.038 |

[^107]from migraine headaches, and is now considering a new method of treatment based on acupuncture. The treatment is rather costly and long-lasting. The friend asks if you think the friend should give it a try. Fortunately, you happen to know a physician with good knowledge of migraine treatment, whom you can ask for advice.
Participants were presented with a verbal probability by the physician (e.g., "It is likely that the treatment will be helpful in that case."). Considering this information, participants were asked to rate how much they would recommend that their friend try this treatment, using a scale that was labeled "do not recommend at all" on the far left and "recommend very much" on the far right. This rating scale contained 101 points $(0-100)^{2}$.

We also measured membership function of verbal probabilities based on Budescu, Karelitz, and Wallsten (2003). Participants were presented with a single verbal probability and 11 probability values ( $1 \%, 10 \%, 20 \%, \ldots$, $90 \%$, and $99 \%$ ) simultaneously and asked to rate the degree (i.e., membership value) to which the verbal probability describes each probability, using a scale that was labeled "not at all" on the far left and "absolutely" on the far right. Therefore, this task measures the degree of certainty attributed to a verbal probability. This rating was recorded with 101 points $(0-100)$.

For these two tasks, we used eight positive and eight negative phrases based on Honda and Yamagishi (2017). Table 1 shows the sixteen phrases used in the experiment.

We conducted the two tasks individually using a computer. In both tasks, a single phrase was randomly presented and participants answered the question. In the decision task, participants answered the question for each phrase once. When measuring membership function, participants answered the question for each phrase twice.

## Results and discussion

Numerical representation of verbal probabilities According to Wallsten, Budescu, Rapoport, Zwick, and Forsyth (1986), we assumed that the degree of certainty attributed to a verbal probability could be represented with a membership function. Peak (the probability with the highest membership value) is one of the most discriminative features of membership functions (Budescu et al., 2003). Accordingly, we assumed that the peak of the membership function represented the degree of certainty for a verbal probability felt by a participant. Since participants rated membership values twice for each phrase, the mean of the membership values was regarded as the membership value for the phrase. Table 1 shows means and SDs of peaks for 16 phrases.
Decision ratings for aggregated data First, we examined the aggregated data. Figure 2 shows the relationship

[^108]between the mean degrees of certainty for phrases (peaks of the membership function) and decision ratings for 8 positive and 8 negative phrases. As is apparent, even though positive and negative phrases were perceived to be analogous in the degree of certainty, decision ratings differed such that participants tended to answer with higher ratings for positive phrases. Therefore, the findings of Teigen and Brun (1999) were essentially replicated in the present study.

Model-based analyses for individual data Next, we analyzed the individual data using the DbBS model. In our DbBS model, we assumed that subjective value of certainty conveyed by a phrase corresponds to the CDF in the beta distribution. Therefore, we estimated two parameters ( $\alpha$ and $\beta$ ) of the beta distribution whose CDF best explains the decision ratings. The two parameters were estimated by a grid search in the range of 0.1 and 10 , with increments of 0.1 . That is, we estimated the parameter using 10000 sets. The parameter set with which the model showed the highest $r^{2}$ between model predictions and decision data was regarded as the best model. We searched the best parameter sets for positive and negative phrases, respectively, for each participant.

We found that the DbBS model generally explained the observed decisions well. The medians of $\mathrm{r}^{2} \mathrm{~s}$ between model predictions and observed data for 60 participants were 0.77 and 0.66 for positive and negative phrases, respectively. In the following analyses, when the model fittings in both positive and negative phrases for a participant showed more than 0.3 in r $^{2}$, we used her/his data. With this criterion, we used data from 45 out of 60 participants ( $75.0 \%$ ). Figure 3 shows five examples of decision ratings and model fittings for positive and negative phrases.


Figure 2. Relationship between subjective degree of certainty for phrases (peak of the membership function) and decision rating.

A


Figure 3. Examples of observed decision rating (points) and model fitting (line) for five participants. (A) shows data for positive phrases. (B) shows data for negative phrases.


Figure 4. Summaries of model-based analyses. (A) Scree plot for within-cluster sum of squares (WSS) in K-Means clustering. (B) Three clusters on decision sample. The black line denotes mean of cluster. The grey line denotes individual data. (C), (D), and (E) show distributions of statistics (mean, entropy, and skewness) in each cluster. (F) Proportions of data in positive and negative phrases that were categorized into each cluster.

Next, we examined participants' memory samples in detail with the following procedures. First, we clustered shapes of beta distributions using probability densities. In particular, patterns of probability densities ${ }^{3}$ for 45 (number

[^109]of participants) $* 2$ (positive and negative phrases) $=90$ data sets were clustered using the K-Means method. We used scree plots for the within-cluster sum of squares (WSS) for each cluster in order to determine the number of clusters (see Figure 4(A)). We adopted three clusters for the following two reasons. Firstly, the reduction of WSS was relatively sharp with up to three clusters. Secondly, since
there were at least 13 data for every cluster, we can assume that each cluster does not necessarily represent rare memory samples.

Figure 4 (B) shows three clusters of memory samples. The black line denotes mean of cluster, and the grey line denotes individual data. Using individual data, we calculated the mean, entropy, and skewness. Figure 5 (C), (D), and (E) show distributions of these statistics in each cluster. Figure $4(\mathrm{~F})$ shows the proportions of data that were categorized into one of the three clusters by positive or negative phrase. The decision patterns were explained with the different assumption between positive and negative phrases. When presented with positive phrases, decision patterns were well explained with the assumption that participants referred to memory samples with low probability (see cluster 1 in Figure 4). In contrast, for the negative phrases, decision patterns were explained under the assumption that participants referred to samples with high probability (see clusters 2 and 3 in Figure 4).

Taken together, we found that the DbBS model generally explained decisions based on verbal probabilities. It was also found that decision patterns were well explained under the different assumptions between positive and negative phrases. For positive (negative) phrases, decision patterns were well explained under the assumption that participants referred to memory samples with low (high) probability. Therefore, our hypothesis was corroborated.

## General discussion

In the present study, we examined decisions based on verbal probabilities. Particularly, we examined whether the DbBS model explained the decision processes. We found that the DbBS model explained the decision patterns well.

Observed differences in memory samples were essentially in accord with our hypotheses based on previous findings about the speaker's choice of directionality in communication (Honda \& Yamagishi, 2017). As previously noted, decisions affected by directionality seem like evidence of decision bias because people make different decisions even when positive and negative phrases convey analogous probabilities. Our present findings answer the question, "Why are people affected by directionality when making decisions?" Our answer is: people take into account the information conveyed by the selected directionality, and as a result refer to different memory samples. Therefore, decisions affected by directionality are not examples of decision bias, but decisions according to different memory samples.

## Acknowledgments

This research was supported by Grants-in-Aid for Scientific Research (A, 16H01725; B, 25280049; B, 16H02835), Innovative Areas (26118002; 16H06569), and Young Scientists (B, 16K16070).

## References

Brown, G., \& Matthews, W. (2011). Decision by Sampling and Memory Distinctiveness: Range Effects from RankBased Models of Judgment and Choice. Frontiers in Psychology, 2, 299.
Budescu, D. V, Karelitz, T. M., \& Wallsten, T. S. (2003). Predicting the directionality of probability words from their membership functions. Journal of Behavioral Decision Making, 16, 159-180.
Honda, H., \& Matsuka, T. (2014). On the role of rarity information in speakers' choice of frame. Memory and Cognition, 42, 768-779.
Honda, H., \& Yamagishi, K. (2017). Communicative functions of directional verbal probabilities: Speaker's choice, listener's inference, and reference points. The Quarterly Journal of Experimental Psychology, 70, 21412158
McKenzie, C. M., \& Nelson, J. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. Psychonomic Bulletin \& Review, 10, 596-602.
Sher, S., \& McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. Cognition, 101, 467-494.
Sher, S., \& McKenzie, C. R. M. (2008). Framing effects and rationality. In N. Chater \& M. Oaksford (Eds.), The probabilistic mind: Prospects for Bayesian cognitive science (pp. 79-96). Oxford: Oxford University Press.
Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. The Quarterly Journal of Experimental Psychology, 62, 1041-1062.
Stewart, N., Chater, N., \& Brown, G. D. A. (2006). Decision by sampling. Cognitive Psychology, 53, 1-26.
Stewart, N., Chater, N., Stott, H. P., \& Reimers, S. (2003). Prospect relativity: How choice options influence decision under risk. Journal of Experimental Psychology: General, 132, 23-46.
Stewart, N., Reimers, S., \& Harris, A. J. L. (2014). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. Management Science, 61, 687-705.
Teigen, K. H., \& Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. Acta Psychologica, 88, 233-258.
Teigen, K. H., \& Brun, W. (1999). The Directionality of Verbal Probability Expressions: Effects on Decisions, Predictions, and Probabilistic Reasoning. Organizational Behavior and Human Decision Processes, 80, 155-190.
Teigen, K. H., \& Brun, W. (2003). Verbal probabilities: a question of frame? Journal of Behavioral Decision Making, 16, 53-72.
Wallsten, T. S., Budescu, D. V, Rapoport, A., Zwick, R., \& Forsyth, B. (1986). Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General, 115, 348-365.

# Semantic diversity, frequency and learning to read: A mini-mega study with children 

Yaling Hsiao<br>University of Oxford

Kate Nation

University of Oxford


#### Abstract

Children who read more tend to be better readers than children who read less. Reading exposure captures not only the number of times words are experienced but also the breadth of the contexts words appear in. Using a large children's corpus of written language, we quantified the former as word frequency and the latter as Semantic Diversity (SemD) (Hoffman et al., 2013). SemD was indexed using Latent Semantic Analysis to calculate the degree of semantic dissimilarity between the contexts in which each appeared. We selected 300 words that varied in SemD for a visual lexical decision and naming task with 9 -year-old children ( $\mathrm{N}=114$ ). Results showed that both frequency and SemD were associated with performance, independently accounting for variation in speed and accuracy. Those words high in frequency and high in SemD were read more efficiently. These findings show that factors beyond frequency are important in determining children's word reading.


# Testing Statistical Learning Implicitly: A Novel Chunk-based Measure of Statistical Learning 

Erin S. Isbilen (esi6@cornell.edu)<br>Cornell University, Department of Psychology, Ithaca, NY 14850 USA

# Stewart M. McCauley (stewart.mccauley@liverpool.ac.uk) <br> University of Liverpool, Department of Psychological Sciences, Liverpool L69 7ZA UK 

Evan Kidd (evan.kidd@anu.edu.au)<br>The Australian National University, Research School of Psychology, Canberra ACT 2601 AU

Morten H. Christiansen (christiansen@cornell.edu)
Cornell University, Department of Psychology, Ithaca, NY 14850 USA


#### Abstract

Attempts to connect individual differences in statistical learning with broader aspects of cognition have received considerable attention, but have yielded mixed results. A possible explanation is that statistical learning is typically tested using the two-alternative forced choice (2AFC) task. As a meta-cognitive task relying on explicit familiarity judgments, 2AFC may not accurately capture implicitly formed statistical computations. In this paper, we adapt the classic serial-recall memory paradigm to implicitly test statistical learning in a statistically-induced chunking recall (SICR) task. We hypothesized that artificial language exposure would lead subjects to chunk recurring statistical patterns, facilitating recall of words from the input. Experiment 1 demonstrates that SICR offers more finegrained insights into individual differences in statistical learning than 2AFC. Experiment 2 shows that SICR has higher test-retest reliability than that reported for 2AFC. Thus, SICR offers a more sensitive measure of individual differences, suggesting that basic chunking abilities may explain statistical learning.


Keywords: statistical learning; chunking; language; language acquisition; implicit learning; learning; memory, serial recall; individual differences

## Introduction

Statistical learning is understood as the process by which individuals implicitly track the distributional regularities in an input, leveraging recurring statistical patterns to facilitate cognitive processing (see Frost, Armstrong, Siegelman \& Christiansen, 2015, for a review). In recent years, validating the theoretical link between the behavior observed in labbased studies of statistical learning and broader aspects of cognition-such as working memory, language processing, and social learning-has garnered extensive interest. However, Romberg and Saffran (2010) noted that although typical tests of statistical learning demonstrate that individuals appear sensitive to statistical structure, such evidence on its own provides little insight into the process of learning, and the nature of the representations that
consequently arise. The lack of a mechanistic understanding of statistical learning was further suggested to complicate attempts to tie this ability to other aspects of cognition, such as language acquisition.

Indeed, endeavors to relate individual variation in statistical learning to other facets of cognitive processing have yielded mixed results. For example, whereas some findings report that statistical learning abilities significantly correlate with verbal working memory and language comprehension (Misyak \& Christiansen, 2012), others find no reliable relationship with language skills (Siegelman \& Frost, 2015). These conflicting reports could suggest either that statistical learning is not meaningfully related to other aspects of cognition, or alternatively, that the measures used to assess statistical learning may not capture its full extent nor the scope of individual variation in this behavior.

In many studies, statistical learning is typically tested using a two-alternative forced-choice task (2AFC), in which learners are presented with pairs of stimuli and are asked to identify which of the two items were present during familiarization. As such, a possible limitation of the 2AFC task is that it is inherently meta-cognitive in nature, requiring the participant to make an explicit response (a button press) based on a "gut feeling" about implicitly acquired statistical regularities. Thus, as suggested by Franco, Eberlen, Destrebecqz, Cleeremans and Bertels. (2015), 2AFC may therefore more accurately reflect explicit decision-making processes rather than the actual underlying statistical learning mechanisms. Relatedly, although the 2AFC task is assumed to serve as an accurate proxy for the learning of statistical structure, the strategy for successful performance on this task may differ from that required for successfully detecting statistical regularities in the input stream (Siegelman, Bogaerts, Christiansen \& Frost, 2017). Lastly, even though 2AFC may yield useful mean estimates of performance at the group level, the additional cognitive complexity associated with 2AFC performance is likely to introduce error variance such that individual scores may not optimally reflect individual differences in statistical learning ability (Siegelman \& Frost, 2015).

Because of these limitations, a unified theoretical framework that situates statistical learning within broader cognitive processing has thus far remained out of reach. In the current paper, we propose a new measure that implicitly tests statistical learning. Our novel task aims to offer more direct insights into what is being learned in statistical learning-based experiments, while at the same time aligning such learning with the wider learning and memory literature.

Recent theoretical considerations suggest that basic abilities for chunking may subserve many aspects of learning and memory, particularly within the domain of language processing (Christiansen \& Chater, 2016). Our perspective builds on classic memory studies demonstrating that the number of items that can be held in memory significantly increases when successfully chunked into larger units (Miller, 1956; Cowan, 2001). This underscores the potential contribution of chunking processes to the successful learning and retention of new material. For example, when tasked with remembering the novel sequence of letters ailcpaphrtleca, preserving the letters in memory poses a considerably greater challenge than successfully recalling the same set of letters chunked into larger coherent units, such as in the sequence catapplechair. Due to our extensive experience with language, the same set of letters can be more easily retained by exploiting our ability to chunk them into words (i.e. "cat", "apple", and "chair"), which in turn can subsequently be deconstructed to retrieve the individual letters. Our novel task takes advantage of similar chunking processes.

Here, we leverage the general capacity for chunking in a statistically-induced chunking recall task (SICR) as a novel implicit measure of statistical learning. We refashion a central tool in the chunking and memory literature-serial recall (e.g., Miller, 1956)—for use in statistical learningbased tasks. Subjects are exposed to six trisyllabic nonsense words using the classic Saffran, Newport and Aslin (1996) paradigm. After training, participants are aurally presented with syllables from the input and asked to recall them out loud. Critically, the experimental items in our task consisted of the concatenation of two words from the input language (Word A + Word B), and control items consisted of the exact same six syllables in a random configuration, like in the example above. Our hypothesis is that if subjects have statistically chunked the syllables in the input stream into words, then recalling a string consisting of two words should yield more accurate recall of the presented syllables than recalling the same set of syllables in a random order. Crucially, our task is scored on a syllable-by-syllable basis rather than assigning a binary 0 or 1 score as in the 2 AFC task, enabling the calculation of subjects' sensitivity to trigrams and serial position. This yields a richer set of performance data than the 2 AFC task, thus providing a more detailed picture of each subject's individual sensitivity to different kinds of information in the input.

In the current paper, we conducted two experiments to determine the efficacy of SICR in capturing statistical learning behavior, and the formation of the word-level
representations from accrued statistics. In Experiment 1, we compare 2AFC performance to SICR, showing that the latter provides a useful, memory-based measure of implicit statistical learning. To be able to relate statistical learning to specific aspects of language and cognition through individual differences studies requires a performance measure that is stable across time. Because recent research has cast doubts on the reliability of the 2AFC task in the context of the classic Saffran-style paradigm (Siegelman, Bogaerts \& Frost, 2016), we conducted a test-retest study of our SICR task in Experiment 2. We conclude with a discussion of the methodological and theoretical implications of SICR, and how future use of this task may help in establishing a definitive relationship between statistical learning and cognition more broadly.

## Experiment 1: Comparing statistically-induced chunking recall (SICR) with 2AFC

Experiment 1 investigated whether chunking might account for the word-level representations gleaned in statistical learning experiments using the classic Saffran et al. (1996) paradigm. In addition to these theoretical considerations, we also sought to assess the methodological efficacy and sensitivity of both the established 2AFC task, and our novel SICR task in assessing statistical learning. Through exposure to the input, we predict that syllables that regularly co-occur in the input will be chunked into words, which should yield higher recall accuracy of the chunked words than the same syllables heard in a random order.

## Method

Participants 69 native English-speaking undergraduates from Cornell University ( 34 females; age: $M=19.78$, $S D=1.62$ ) participated for course credit.

Materials The input language consisted of 18 syllables (bi, bu, di, du, ga, ka, ki, la, lo, lu, ma, mo, pa, po, ri, ta, ti, to), combined into six trisyllabic words: kibudu, latibi, lomari, modipa, tagalu, topoka. Seventy-two randomized blocks of the six words were concatenated into a continuous speech stream using the MBROLA speech synthesizing software (Dutoit et al., 1996). Each syllable was approximately 200 milliseconds long, separated by 75 milliseconds of silence.

For the 2AFC task, six additional foil words were pseudorandomly generated, avoiding the reuse of transitional probabilities from the target words above: dikabi, kigala, lopadu, mamoti, polubu, tatori.

The stimuli for the SICR task consisted of 24 six-syllable items. The twelve experimental items were composed of two adjacent words from the input (e.g., kibudulatibi), and the twelve corresponding foil items consisted of the same set of syllables in pseudorandom order (e.g., kibudulatibi $\rightarrow$ tidubibulaki), avoiding preexisting transitional probabilities from all other syllable combinations in the experiment. Additionally, 125 -syllable practice items were included, which were constructed in the same manner as the 24 items
reported above, but using one full word and the first bigram of a second word.

Procedure The experiment consisted of three distinct tasks. First, subjects were familiarized with the artificial language. To ensure active engagement, a cover task based on Arciuli \& Simpson (2012) was administered. In addition to each of the six words in the experiment, three variants of each word containing a syllable repetition was included in the training stream (e.g., tagalu $\rightarrow$ tatagalu, tagagalu, tagalulu). Participants were instructed to click the space bar when they noticed a repeated syllable. Each of the three variants of the words appeared 4 times, yielding 72 repetitions. In total, training lasted 11 minutes.

After training, participants' knowledge of the artificial language was tested using both the standard 2AFC task, and our SICR paradigm. The order of these two tasks was counterbalanced such that half of the subjects were given 2AFC first, and half were given SICR first. In the 2AFC task, each of the 6 target words were aurally presented with one of the 62 AFC foil words, and subjects were asked to report which of the two trigrams had been present during training. There were 36 2AFC trials in all, in which each target word appeared alongside each foil once.

In the SICR paradigm, 12 five-syllable practice trials were administered prior to the 24 six-syllable items to familiarize subjects with the task, and to ensure that the amount of post-test exposure to the words would be the same regardless of whether subjects did 2AFC first, or SICR first. In this task, participants were told that we would be gauging their ability to recall the syllables from the experiment. Each item was aurally presented, after which subjects were prompted to recite back each syllable in the sequence to the best of their ability. Importantly, at no point in the experiment were subjects informed that they were partaking in a language experiment, nor was their attention directed to the presence of structure.

## Results and Discussion

The mean accuracy of correctly choosing the word over the foil in the 2AFC task was $66 \%(M=.66, S D=0.13)$, which is significantly greater than chance, $t(68)=11.11, p<.001$. These results are comparable with other studies that utilize 2AFC to assess statistical learning, which typically report performance within the range of $60 \%$ (Frost et al., 2015).

Scoring for the SICR task was done on a syllable-bysyllable basis, enabling analysis of both the overall strings, and the individual words composing the strings. When comparing the number of syllables accurately recalled for the experimental items $(M=42.7, S D=10.68)$ to the number of syllables recalled for random items ( $M=31.19$, $S D=10.29$ ), participants accurately recalled significantly more syllables for the experimental items than the random items, $t(68)=13.85, p<.0001$. A similar pattern was observed for trigram performance: participants accurately recalled significantly more of the experimental trigrams ( $M=8.68$,


Figure 1: a) Average SICR performance. Participants recall significantly more syllables when the test items consist of two concatenated input words, and significantly more trigrams within the experimental six-syllable items. b) Serial position curves for experimental and random items.
$S D=4.25$ ) than items consisting of random trigrams ( $M=3.58, S D=3.02$ ), $t(68)=13.72, p<.0001$ (Figure 1a). Additionally, the serial position curves for the experimental and random items can be found in Figure 1b. These results confirm our hypothesis that through exposure to the distributional regularities in the input, individuals appear to have successfully chunked co-occurring syllables into larger units, and the formation of these word-level representations of the input leads to markedly better memory for experimental items.
Interestingly, our analyses revealed no significant correlations between 2AFC and any of our SICR measures $(r(67)=0.21, p=.084$ for experimental items, and $r(67)=0.18$, $p=.4$ for experimental trigrams. For the score distributions of the two tasks, see Figure 2). However, this finding mirrors recent results by Franco et al. (2015), who also found no correlation between 2AFC accuracy and their Rapid Serial Auditory Presentation task (RSAP), a detection task intended to serve as a more implicit measure of auditory statistical learning. Similar to SICR, RSAP works by exposing subjects to an artificial speech stream composed of trisyllabic words, after which subjects were tasked with detecting a target syllable embedded within strings of target words from the training corpus. Unlike explicit measures like 2AFC, RSAP and SICR are implicit measures in which no reference is made to a desired discrimination, and thus may be more sensitive to the acquired statistical regularities, including information about


Figure 2: The distributions of SICR (experimental-random items), 2AFC scores as compared to chance, and syllable recall for experimental items.
which the participant lacks awareness. Thus, 2AFC and SICR may be picking up on different aspects of statistical learning - decision-making processes based on learned information and underlying mechanisms, respectively which may contribute to the low correlation between the two measures.

Notably, our analyses revealed a strong order effect for 2AFC performance: individuals who performed SICR prior to 2 AFC exhibited significantly higher 2 AFC scores, $t(68)=$ $12.06, p<.0001$. Compared to the means of those who completed 2 AFC first, a $7 \%$-point increase in 2 AFC performance was observed for participants who did SICR first. This may account for why our participants on average performed higher on 2 AFC than the $60 \%$ typically reported for this type of statistical learning. By contrast, SICR was unaffected by the order in which it was performed $(t(68)=0.22, p=.59$ for experimental items, $t(68)=-0.22$, $p=.42$ for experimental trigrams). The robustness of SICR is notable given that in both conditions, the amount of postinput exposure was kept the same, ruling out exposure differences as an explanation for the order effects. That is, despite both tasks being granted the same opportunity for post-input learning, only 2 AFC was affected by the additional exposure.

Taken together, several conclusions can be made from the results of Experiment 1. Firstly, our findings support the idea that chunking may serve as the mechanism by which exposure to statistical regularities lead to representational changes in memory. Secondly, our results affirm that SICR can serve as a valid means of testing the acquisition of sequential regularities, with the additional benefit of offering more fine-grained insight into the acquired representations. Finally, the lack of correlation between 2 AFC and SICR may represent fundamental differences between explicit versus implicit measures of learning (Franco et al., 2015). Thirdly, the lack of order effects on SICR performance suggests that it may be a more stable measure of statistical learning ability than 2AFC. To further examine the stability of SICR across time, we assessed its test-retest reliability in Experiment 2.

## Experiment 2: Establishing the test-retest reliability of SICR

To date, varying levels of test-retest reliability for different measures of statistical learning have been found. For instance, using 2AFC as the primary measure, Siegelman and Frost (2015) reported adequate test-retest reliability for auditory verbal adjacent ( $r=0.63$ ), and visual nonverbal adjacent statistical learning ( $r=0.58$ ), and relatively low reliability for auditory nonverbal adjacent ( $r=0.23$ ) and auditory verbal non-adjacent statistical learning ( $r=0.31$ ). The implications of this are twofold: a) that certain types of statistical learning capacities are not stable within individuals and/or b) that certain tasks may lack specificity as to the behavior they aim to capture (Siegelman et al., 2017). Thus, the goals of Experiment 2 were to determine whether SICR provides a reliable measure of individual statistical learning capabilities, and to establish whether the associated hypothesis-that chunking abilities can account for statistical word learning-would replicate.

## Method

The same general method from Experiment 1 was employed, with a few notable exceptions. Subjects were exposed to the same input language, after which SICR was administered to measure word learning. Unlike the previous study, 2AFC was not included in Experiment 2, given existing studies assessing its test-retest reliability. Following the completion of Session 1, participants returned three weeks later and completed the same tasks again in Session 2, mirroring the timespan between test and retest in Siegelman and Frost (2015).

Participants 26 native English-speaking undergraduates from Cornell University (15 females; age: $M=19.31$, $S D=1.32$ ) participated for course credit.

Materials The same input language from Experiment 1 was used. The SICR stimuli consisted of the same 24 sixsyllable items from Experiment 1, half composed of two concatenated words from the input, and the other half their complementary randomized foils.

Procedure The experiment consisted of two tasks. First, subjects were familiarized with the input language, including the same cover task as before. In total, training lasted 11 minutes. The SICR task was identical to Experiment 1, with the exception that participants were given a different randomized input and SICR item order in each session.

## Results and Discussion

As in Experiment 1, participants performed significantly better on the experimental items than on the random items, both in Session 1, $t(25)=5.46, \mathrm{p}<.0001$, and in Session 2, $\mathrm{t}(25)=7.08, p<.0001$. The same results were found for

Table 1: Means and standard deviations of SICR scores

|  | Session 1 |  |  |  | Session 2 |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | M | SD |  | M | SD |  |
| 6-syllable <br> experimental | 36.42 | 12.48 |  | 40.15 | 12.73 |  |
| 6-syllable | 27.04 | 10.71 |  | 28.0 | 10.38 |  |
| random | 6.89 | 4.41 |  | 8.31 | 4.46 |  |
| Trigrams | 3.0 | 2.65 |  | 2.96 | 2.60 |  |
| experimental <br> Trigrams <br> random | 3.96 |  |  |  |  |  |

performance on the trigrams, with participants recalling significantly more experimental trigrams in both Session 1, $t(25)=6.18, p<.0001$, and in Session 2, $t(25)=7.67, p<.0001$. The mean performance on these measures can be found in Table 1. Thus, the results from both sessions replicated the results from Experiment 1.

Between the two sessions, the test-retest reliability of SICR proved to be very strong. SICR performance was highly correlated across the two sessions. Performance on the recall of six-syllable experimental items was $r(24)=0.81$, $p<.0001$ (Figure 2). This exceeds the correlation coefficient of 0.63 reported for 2 AFC in an auditory statistical learning task by Siegelman and Frost (2015). Recall performance on the six-syllable random items was also highly stable, $r(24)=0.85, p<.0001$. Performance on experimental trigrams $r(24)=0.73, p<.0001$ and random trigrams $r(24)=0.82$, $p<.0001$ was also consistent across the two sessions. However, the correlations of the differences scores (performance on experimental minus random items) were slightly lower, yielding $r(24)=0.46 p=.0192$ for six-syllable recall, and $r(24)=0.53 p=.0053$ for trigram recall. These results suggest that performance on the experimental items may be a better measure of individual differences in statistical learning than the difference scores.

In all, the results of Experiment 2 corroborate our findings from Experiment 1, in which experimental items yield significantly better recall. Our results also confirm the


Figure 3: Correlation between Sessions 1 and 2 recall scores for statistically experimental items.
stability of SICR. Taken together, these findings suggest that SICR proves to be both a theoretically valid and methodologically sound measure of statistical learning.

## General discussion

In this paper, we introduced a novel chunk-based method to implicitly test statistical learning-the SICR task-as an alternative to the standard 2AFC task. The results of our experiments demonstrate that through exposure, subjects' implicit chunking of the distributional regularities in the input significantly amplified their baseline working memory abilities (as captured by performance on the random items), and that the formation of multi-syllabic chunked representations of the input markedly boosted recall. Furthermore, these results appear to be strikingly stable over time and are less subject to order effects than 2AFC, which underscores the promise of SICR as a reliable and multifaceted measure of statistical learning faculties.

SICR offers several methodological benefits that circumvent a variety of issues inherent to 2AFC. Because 2AFC relies on overt decision-making processes about the familiarity of stimuli, it is unclear as to whether 2AFC may thus only be reflective of the more explicit meta-cognitive aspects of statistical learning. 2AFC appears to provide more limited sensitivity to individual differences, as it tends to rely on a binary all-or-nothing score. This lack of granularity in the scoring also makes it more difficult to accurately assess the precise extent of learning.

One important difference between explicit tasks like 2AFC and implicit tasks such as SICR is that they may be respectively characterized as 'direct' versus 'indirect' measures of learning (Franco et al, 2015). Whereas direct measures steer participants' attention toward the relevant discriminations they are expected to make, indirect measures that circumvent the need for explicit instruction may be more sensitive to any knowledge the subject has acquired, including material below the threshold of conscious awareness. That is, although direct and indirect measures should exhibit equal sensitivity to consciously known information, direct measures may not be as adept at capturing the accretion of information of which the learner is not yet fully aware. Furthermore, unlike 2AFC and reaction time tasks, SICR requires both immediate comprehension and production on the part of the learner. The task thus provides the means to capture how exposure to statistical regularities can facilitate memory abilities via improved chunking abilities, which in turn may help the learner to overcome the processing pressures deriving from the Now-or-Never bottleneck (Christiansen \& Chater, 2016). As such, SICR may be seen as an ecological measure of the impact of accrued statistics on the online memory processes used to track verbal input, without the need for participants to rely on explicit decision-making.

Whereas 2AFC relies on a binary scoring method, SICR offers a more granular approach by performing scoring on a syllable-by-syllable basis, allowing the evaluation of sensitivity to trigrams and serial position. The richness of
this dataset may also lend itself to acoustic measurements of production durations and analysis of prosody. Because of the sensitivity of SICR to a number of different individual capacities, and findings suggesting that chunking ability serves as a strong predictor of online language processing skills (McCauley \& Christiansen, 2015), SICR may also be employed compare how individual differences in statistical learning may predict other language learning abilities. Indeed, preliminary results from an ongoing study with 5-6-year-old children ( $N=73$ ) indicate that performance on the experimental items in the SICR task correlates significantly with language skill ( $r=0.41, p<.001$ ), whereas 2AFC performance does not ( $r=0.20, p=.096$ ).

More generally, the basic recall methodology upon which SICR piggy-backs has a long pedigree in the domaingeneral memory literature, including serial recall (e.g., Miller, 1956). Of particular importance is the related work on nonword repetition, which has been established as one of the primary predictors of language ability (e.g., Gathercole et al., 1994). Our SICR measure may be seen as a statistical learning-based variation on a nonword repetition task, in which we manipulate the distributional support for the items to be recalled via artificial language exposure. This interpretation of the SICR task dovetails with evidence that nonwords constructed from phoneme sequences that occur frequently in natural language are repeated more accurately than nonwords based on infrequent phoneme strings (Majerus, van der Linden, Mulder \& Peters, 2004). In a similar vein, recall of random digit sequences has also been shown to reflect natural language statistics (Jones \& Macken, 2015).

In addition to the methodological advantages afforded by this novel method, SICR also points toward a theoretical answer to Romberg and Saffran's (2010) concern about the lack of connection between measures of statistical learning and potential underlying processes and representation. Our proposition, given the efficacy of SICR in capturing statistical learning behavior, is that chunking may be seen as the process by which encountered statistics are used to form concrete, discrete units, thereby effectively segmenting a continuous stream into individual words. As such, the output of statistical learning may thus be seen as individual chunks of varying sizes. This notion is corroborated by previous research suggesting that chunking-based processes enable the recoding of incoming information into gradually higher levels of abstraction, from acoustic input, to words, to multiword units and beyond (Christiansen \& Chater, 2016). Thus, SICR provides both a compelling tool to effectively and ecologically appraise statistical learning, and strives to bridge the statistical learning and chunking memory literatures.

## Acknowledgments

We wish to thank Dante Dahabreh, Jess Flynn, Jake Kolenda, Jeanne Powell and Sam Reig for their assistance with collecting and coding data. This work was supported in part by NSF GRFP (\#DGE-1650441) awarded to ESI.

## References

Arciuli, J., \& Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. Cognitive Science, 36(2), 286-304.
Christiansen, M.H. \& Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 39, e62.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24, 87-114.
Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., \& Bertels, J. (2015). Rapid serial auditory presentation. Experimental Psychology, 62, 346-351.
Frost, R., Armstrong, B.C., Siegelman, N. \& Christiansen, M.H. (2015). Domain generality vs. modality specificity: The paradox of statistical learning. Trends in Cognitive Sciences, 19, 117-125.
Gathercole, S. E., Willis, C. S., Baddeley, A. D., \& Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. Memory, 2(2), 103-127.
Jones, G. \& Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. Cognition, 144, 1-13.
McCauley, S.M. \& Christiansen, M.H. (2015). Individual differences in chunking ability predict on-line sentence processing. Proceedings of the 37th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.
Majerus, S., van der Linden, M., Mulder, L., Meulemans, T., \& Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. Journal of Memory and Language, 51, 297-306.
Misyak, J.B. \& Christiansen, M.H. (2012). Statistical learning and language: An individual differences study. Language Learning, 62, 302-331.
Romberg, A. R., \& Saffran, J. R. (2010). Statistical learning and language acquisition. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 906-914.
Saffran, J. R., Newport, E. L., \& Aslin, R. N. (1996). Word segmentation: The role of distributional cues. Journal of memory and language, 35(4), 606-621.
Siegelman, N., Bogaerts, L., Christiansen, M. H. \& Frost, R. (2017). Towards a theory of individual differences in statistical learning. Phil. Trans. R. Soc. B, 372(1711), 20160059.

Siegelman, N., Bogaerts, L. \& Frost, R. (2016). The peculiar tale of ASL: What do we measure when we use the auditory statistical learning task? Talk presented at the conference on First vs. Second Language Learning: From Neurobiology to Cognition conference, Hebrew University, Israel.
Siegelman, N., \& Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. Journal of Memory and Language, 81, 105-120.

# Minimal covariation data support future one-shot inferences about unobservable properties of novel agents 

Julian Jara-Ettinger (julian.jara-ettinger@yale.edu) ${ }^{\mathbf{1}}$, Hyowon Gweon (hyo@stanford.edu) ${ }^{2}$,<br>${ }^{1}$ Department of Psychology, Yale University, New Haven, CT 06520<br>${ }^{2}$ Department of Psychology, Stanford University, Stanford CA, USA


#### Abstract

When we reason about others' behavior, there are often many equally-plausible explanations. If Bob climbs a tree to get an apple, we may be unsure if Bob found climbing difficult but really wanted an apple; if he found climbing easy and was not particularly excited about the apple; or if he found climbing intrinsically fun and just got the apple because it was convenient. Past research suggests that we solve this problem by obtaining repeated observations about the agent and about the world. Here we argue that, beyond allowing us to sharpen our inferences about agents and the world, covariation data also enables us to do one-shot inferences about novel agents. We show that given minimal covariation data, people can infer objective and subjective properties of a new agent from a single event. We show that a model that assumes that agents maximize utilities matches participant judgments with quantitative precision.


Keywords: theory of mind; social cognition; computational modeling

## Introduction

In our everyday social interactions, we easily learn aspects of people that are directly observable. We hear people's names, see what they look like, and recognize their jobs. But getting to know someone means much more: what they like, what they're good at, and even what they think of themselves. We invest much of our social interactions gathering observable evidence about these unobservable qualities of others, and even plan opportunities that serve specifically this purpose such as interviews with applicants or dates with potential partners.

A growing set of studies suggest that when we reason about others we assume that they act to maximize the rewards that they obtain relative to the costs that they incur (see Lucas et al. 2014 and Jara-Ettinger et al. 2016 for review). If, for instance, we watch an agent walk straight to a coffee shop, we can infer that getting coffee is rewarding (explaining why the agent went there) and that walking is costly (explaining why she took the shortest path). Despite its simplicity, this ability to reason about behavior in terms of costs and rewards, called a Nä̈ve Utility Calculus, supports rich explanations, enabling observers to distinguish between highly motivated agents (high rewards) and poorly motivated agents (low rewards), and supporting reasoning about agents who ignore goals because of a lack of competence (costs are too high) and because of a lack of motivation (rewards are not high enough).

Decomposing behavior into costs and rewards, however, means that action-understanding is usually confounded, even in the simplest scenarios. If, for example, an agent jumps over an obstacle to reach an object on the other side, her behavior
can be explained equally well by appealing to different combinations of costs and rewards. The agent may have found jumping very costly, but the outcome even more rewarding. Alternatively, she may have found jumping relatively easy, and the outcome not particularly rewarding. Or she may have even found jumping rewarding, and not cared about the object. To complicate matters further, agents not only incur costs and obtain rewards, but they also have beliefs about their own costs and rewards, and these beliefs guide their behavior. Imagine, for instance, watching a girl pull out a sword from a stone. While it is trivial to see that her goal was to get the sword (and that it was therefore rewarding), it is difficult to determine how much she wanted the sword (was the reward high or low?), how strong she is (is the cost low for her?), how difficult it is to pull the sword out (is the cost high in general?), or what she thought about her own strength before trying (what did she believe about her own costs?).

The problem of confounded explanations is most obvious when we only have access to a single event. But in more realistic situations, we often watch different people pursue the same goal, and we watch the same person pursue different goals (see, e.g., Figure 1). This covariation most directly allows us to learn about the agent we are observing (Kelley \& Michela, 1980). However, it may also enable us to make stronger inferences about new agents. Returning to the example above, what if you knew that several other people had already tried to pull the sword out and failed, and that the girl decided to try anyway? Even though the information about the girl is the same, you might be more confident about your inferences in this second case: the girl probably really wanted the sword (she probably believes that the cost is in general high), she thought she'd be strong enough to succeed (she believes that the cost may be lower for her specifically), and she was right (our observation of her success suggests the cost was indeed lower for her)!

Here we propose that minimal covariation data about the outcomes of agents' goal-directed actions, combined with our commonsense psychology, enable us to make richer inferences about novel agents. We show that even from a brief history of actions, people can make powerful joint inferences about a new agent's desire, competence, and even beliefs about their own competence, all from a single action. Below we briefly review research that motivates our proposal, we present our theory instantiated as a computational model in a Bayesian framework, and we then present two experiments that test our model predictions.

## Agent-dependent and agent-invariant dimensions of costs and rewards

Costs and rewards are partially objective and agent-invariant (e.g., a high hill is more costly to climb than a low hill, and three cookies are more rewarding than one), and partially subjective and agent-dependent (e.g., some are better than others at hill-climbing, and some like cookies more than others). Thus, to effectively explain an event, we not only need to infer the underlying costs and rewards, but we must also uncover what aspects of the costs and rewards are specific to the agent and what aspects of the costs and rewards are properties of the world and apply to all agents. Decomposing costs and rewards into agent-dependent and agent-independent dimensions not only helps us understand the event better. It also helps us understand new events more easily. If we know what costs and rewards are specific to an agent, then we can use this knowledge to explain the agent's behavior in new events (e.g. if learn that someone is strong, this helps us interpret their successes and failures in new events). If we know what aspects of costs and rewards are properties of the environment, then we can use this knowledge to make sense of new agents acting in this familiar situation (e.g. if we learn that a box is heavy, this helps us interpret the success or failure of new agents when trying to lift the box).

## One-shot learning from covariation information with a Naïve Utility Calculus

Based on these intuitions, we propose that people rely on covariation information to break down costs and rewards into their agent-dependent and agent-independent components, and that, with this decomposition at hand, people rely on their Naïve Utility Calculus to make rich inferences from single events. Returning to the example above, if we already understand that getting the sword is difficult to pull out because many have failed, then, if we a new agent succeed, we can be sure that it was not because the sword was easy to pull out, but because the person was strong; an inference that would have been impossible to make the first time we saw this. Similarly, if the successful agent had already watched others try and fail, we can assume that she also knew the sword as difficult to lift, and so she probably thought she has strong enough to succeed; otherwise, she would not have bothered trying. If she succeeds, then we can also be certain that she really was strong.

Recent work suggests that even infants can use covariation information to infer properties of the world and properties of agents. When one agent successfully activated a toy twice but the other failed twice (suggesting one is more competent than the other), infants attributed their own failure with the toy to their incompetence and sought help from others; conversely, when each agent succeeded twice and failed twice on the toy, infants attributed their failure to the toy, seeking a different one instead (Gweon \& Schulz, 2011). Furthermore, older children (4- and 6-year-olds) use covariation information between characters and activities to generate different
causal explanations for their behaviors (Seiver, Gopnik, \& Goodman, 2012). For instance, if Sally and Anne both tried activity A but not B, children were more likely to appeal to the properties of the activities to explain their actions (e.g., A is more fun than B); but when Sally tried both A and B but Anne tried neither, children appealed more to the characters' attributes (e.g., Sally is older). Furthermore, children generalized these explanations to predict whether the characters would try a new activity, or what another character would do on the same activities. These results suggest that humans, even early in life, are sensitive to the covariation information embedded in others' actions: they infer both the relevant properties of people and the physical world (e.g., toys, activities) and readily use it to explain their actions.

Similarly, children have a Naïve Utility Calculus by age five, with some form of it tracing back to infancy. Even infants have some expectation that agents navigate efficiently (Csibra, 2003) and that this expectation reflects some understanding of cost minimization (Liu \& Spelke, 2016). Also before their second birthday, children understand that both competence and rewards vary across agents (Repacholi \& Gopnik, 1997; Jara-Ettinger, Tenenbaum, \& Schulz, 2015). And by age five, children can explicitly explain behavior by inferring the unobservable costs or rewards, given partial information (Jara-Ettinger, Gweon, Schulz, \& Tenenbaum, 2016).

As reviewed above, the two main accounts that our proposal relies on -understanding covariation, and having a Naïve Utility Calculus- are both available early in life. Although our goal here is to explore this possibility with adults, the developmental research suggests that the abilities our account requires are likely central to social reasoning as they can be traced to our first years of life. The next section describes our computational model that formalizes these intuitions. We use the model to obtain quantitative predictions and compare them against empirical data across two experiments. In Experiment 1 we test if our account explains how we jointly infer properties about agents and the world using covariation information, and how this past information, in turn, supports one-shot learning of objective and subjective properties of novel agents. Because participants in Experiment 1 explicitly make judgments about the covariation in formation, in Experiment 2 we test if this step is critical for people to integrate this information when reasoning about new agents.

## Computational modeling

In order to test our predictions more formally, we implemented a computational model of our account and a simple alternative model that ignores the covariation data when inferring properties of the new agent. The principles of our model apply to any situation in which the outcome of events depend jointly on properties of agents and properties of the world; here, we describe it in the context of our experimental paradigm (see Procedure section in Experiment 1 and Figure 1), where agents with different levels of strength attempt to
lift boxes of different weighs in order to obtain rewards.

## Naive Utility Calculus model

Our implementation is a simplified variation of the Naïve Utility Calculus model (Jara-Ettinger, Schulz, \& Tenenbaum, 2015). Whereas the past models were designed to reason about agents navigating in two-dimensional environments, this model is adapted for reasoning about agents making choices without any spatial information (along the lines of (Lucas et al., 2014)). For a single event, the agent's strength and the box's weight are inferred using Bayesian inference:

$$
\begin{equation*}
p(W, S \mid O) \propto p(O \mid W, S) p(W) p(S) \tag{1}
\end{equation*}
$$

where $W$ is the weight of the box, $S$ is the strength of the actor, and $O$ is the observed outcome (success or failure). For simplicity, we use a deterministic likelihood function where agents can successfully lift a box only when their strength is higher than the box's weight. As such, we represent strength and weight using a common scale, using real values ranging from 0 to 1 .

By providing covariation information, where agents interact with different boxes, Equation 1 enables observers to break down events into agent-dependent (strenght) and agent-invariant (weight) components. With this information at hand, when we watch a single event from a new agent (henceforth the one-shot agent), we compute her preference by relying on the assumption that she is attempting to maximize her subjective utilities (see Introduction). An agent's expected utility for any given box is given by the reward associated with the box times the probability that the agent will be able to retrieve it. As such, an agent's choice reflects a trade-off between the magnitude of the reward, and the probability that the agent will be able to get it if she tried. Given a choice $C$, the posterior probability of the agent's underlying preferences is given by

$$
\begin{equation*}
p(P \mid C) \propto p(C \mid P) p(P) \tag{2}
\end{equation*}
$$

where $P$ represents the rewards associated with each option. For simplicity, we assume that the observer has a uniform prior over the agent's preferences $(\mathrm{p}(P))$, and we compute $p(C \mid P)$ by integrating the observer's prior belief over the actor's strength:

$$
\begin{equation*}
p(C \mid P)=\int_{S} p(C \mid P, S) p(S) \tag{3}
\end{equation*}
$$

where $S$ is the agent's strength, and $C$ is the agent's choice. This intermediate term, $p(C \mid P, S)$, integrates the assumption that the agent is attempting to make choices that maximize her utilities. Finally, the one-shot agent's objective strength is also computed using equation 1 .

## Alternative model

To test the role of the past covariation information in the final (one-shot) trial, we implemented a simple alternative model. In this baseline model we assume that participants ignore the
covariation information and make all judgments about the one-shot agent using that event alone. As such, this model is computationally equivalent to the main model, as it relies on Equations 1-3 to reason about the agent, but it does not use the covariation data to sharpen its estimates.

## Experiment 1

To test our hypothesis that people can use past observations of multiple agents to make one-shot inferences about a novel agent, we designed a behavioral experiment where participants received covariation data about three agents, each of whom attempted to lift four different boxes (see Figure 1). Next, participants watched a single agent choose one of the boxes and either succeed or fail to lift it. After this single event, participants were asked to infer three properties of the agent: her preference, her beliefs about her own strength, and her true strength.

## Methods

Participants 100 adults participants (mean age $=35.95$; range: 19-70) from the US (as determined by their IP address) were recruited using Amazon's Mechanical Turk framework. Participants were randomly assigned to one of 10 conditions (see Procedure).
Procedure Participants read a brief story that consisted of two parts. In the first part (Part 1 in Fig.1), participants learned about a game where, if players could successfully lift a box, they were allowed to keep its contents. Next, participants learned about three players (Circle, Rhombus, Triangle) who played with different boxes. There were five boxes, but the agents only had four coins and interacted with just the first four (Candy, Teddy Bear, Rubber Duck, and Baseball boxes); no one interacted with the fifth box (Yoyo box) and no mention was made about it other than stating that it was an option. For each action of each agent, participants learned whether the agent succeeded or failed; the first agent (Circle) sequentially tried the four boxes (in fixed order as shown in Figure 1), followed by the second (Rhombus), and then the third (Triangle). After each attempt, the cumulative outcomes were summarized visually as in Figure 1. After observing this covariation data, participants were asked to determine how heavy each box was and how strong each agent was. Both types of questions were answered on a numerical scale ranging from 0 to 9 . In the weight questions, 0 indicated very light, 5 indicated average, and 9 indicated very heavy. In the strength questions, 0 indicated very weak, 5 indicated average, and 9 indicated very strong.

In the second part of the task (Part 2), participants learned about a fourth agent (Square) who had also watched the other three agents. Participants learned that this final agent only had enough money to play the game just once. Participants were then shown which box (of the five) the agent selected, and whether she succeeded or failed in lifting it. Crossing agent's choice ( 5 boxes) and outcome (success or failure) produced 10 conditions, to which participants were randomly as-

Part 1


## Part 2



Figure 1: Visual summary of the experiment. Participants were introduced to four agents and five boxes. The first three agents (the square, rhombus, and triangle) interacted with the first four boxes (but not the fifth box). In Experiment 1, after observing these trials (and before seeing the final agent), participants were asked to rate the relative strength of these three agents, and the relative weight of the four boxes. In the second part of the experiment, the final agent (the square) chose one of the five boxes and either succeeded or failed to lift it (producing a total of 10 conditions that we tested across participants). Participants were then asked to determine this agent's preference, strength, and beliefs about her own strength when she made her choice.
signed (see Part 2 in Figure 1). Participants were then asked three questions in the following order. First, participants were asked to rate how much the agent wanted the object in the box using a scale from 0 ("not at all") to 9 ("very much"); Preference. Second, they were asked to rate the agents strength on a scale from 0 ("very weak") to 9 ("very strong"); True Strength. Third, participants were asked to rate the agent's beliefs about their own strength on a scale using an identical scale to the one used in the second question (Perceived Strength).

## Results

Participants' responses from the experiment were $z$-scored within response type (preference inferences, weight inferences, and strength inferences) and then averaged across participants.

First, we looked at people's use of covariation data by looking at their inferences about agents' strength and boxes' weights from Part 1. The model provided very high quantitative fits (Figure 2). On the joint inference over strength


Figure 2: Overall results from Experiment 1. The x-axis shows the model predictions and the $y$-axis shows participant judgments. The left plot shows inferences obtained from the covariation data (see Figure 1). The right plot shows inferences made from the one shot event.
and preference for the first set of agents (see Part 1 in Figure 1), the model showed a correlation of $\mathrm{r}=0.99$ with participant data ( $95 \%$ CI: 0.99-1.00).

Having verified that participants attended to the covariation data in Part 1 and accurately inferred the boxes' weight and the agents' strength, we then looked at whether participants made used this information when interpreting the event from the one-shot agent in Part 2.

Qualitatively, the results from the one-shot learning trial were as expected (see Figure 3). Participants judgments about the agents true strength varied both as a function of the box that she chose, and the outcome. Similarly, inferences about the agent's beliefs about her own strength also varied as a function of the box that she chose to lift.

On the joint inferences about the final agents preference, true strength, and perceived strength (Part 2), participant judgments showed a correlation of $\mathrm{r}=0.86$ with participant data ( $95 \%$ CI: 0.67-0.94).

By contrast, our alternative model, which used the same computations but did not learn from the covariation data, failed to predict the one-shot inferences participants made about the novel agent. Because the model ignores the covariation data, it does not make any predictions about the first set of agents; thus we only report the fit between the alternative model and people's responses in Part 2, about the target agent. The model showed a correlation of $\mathrm{r}=0.40(95 \% \mathrm{CI}$ : $-0.06,0.71$ ) against participant judgments.

## Experiment 2

Experiment 1 established that, when given covariation data, people can infer a novel agent's preference, strength, and perceived strength from a single event. In this experiment participants were explicitly asked to think about the covariation data and judge the strength of each agent and the weight of each box. It is possible that people do not naturally decompose preferences and competence into agent-dependent and agent-independent features, and this only happens when par-


Figure 3: Results from Part 2 of Experiment 1. The x-axis shows the box that the protagonist shows and the $y$-axis shows participant's ratings for the agent's perceived strength and the agent's true strength. Red bars show the conditions where the agent failed to lift the box, green bars show the conditions where the agent successfully lifted the box, and the grey bars show the agent's self-perceived strength. Judgments are zscored within participants and averaged and the vertical bars represent $95 \%$ confidence intervals. People inferred lower strength when the agent failed relative to when the agent succeeded, and these inferences depended on the box that the agent chose.
ticipant's attention is drawn to the information they can use. We test this possibility in Experiment 2. Experiment 2 was identical to Experiment 1 with the exception that participants were not asked about the covariation data and were just asked to rate the one-shot agents preference, true strength, and perceived strength.

## Methods

Participants 100 adult participants (mean age $=35.51$; range: 20-70) from the US (as determined by their IP address) were recruited using Amazons Mechanical Turk framework.

Procedure The procedure was identical to Experiment 1 with the exception that people were not asked to judge the weight of each box or the strength of any of the agents in the first part of the story (shown in Figure 1).

## Results

As in Experiment 1, results from the experiment were zscored within response type (preference inferences, weight inferences, and strength inferences) and then averaged across participants.

Figure 4 shows the results from the experiment. As in Experiment 1, the model fit participant judgments with high accuracy (Figure 4a). On the joint inferences about the oneshot agent's preference, true strength, and perceived strength, participant judgments showed a correlation of 0.88 with participant data ( $95 \%$ CI: 0.72-0.95). Consistent with this, participant responses in Experiment 2 resembled the responses


Figure 4: Results from Experiment 2. (a) Model predictions plotted against participant judgments. (b) Comparison of results from Experiment 1 and Experiment 2.
from Experiment 1. Figure $4 b$ shows the comparison between the results in Experiment 1 and the results in Experiment 2. The two sets of data showed a correlation of $\mathrm{r}=0.92$ ( $95 \% \mathrm{CI}$ : 0.80-0.97).

## General Discussion

Inferences about unobservable qualities of others from single observations are often ambiguous. Across two experiments, we showed that people can rely on past knowledge to make strong inferences about new agents from a single action. Consistent with previous research, Part 1 of Experiment 1 showed that people can decompose ambiguous events into properties of agents and properties of the world by relying on the covariation structure in the data (Kelley \& Michela, 1980; Gweon \& Schulz, 2011; Seiver et al., 2012). We also showed that these representations about the agents and the physical world support powerful one-shot inferences in future events. People accurately inferred an agent's preferences, their true strength (competence), and the agent's beliefs about her own strength, all from a single event. In Experiment 2, we replicated these results and showed that people spontaneously make use of covariation data in new events. Even when people were not asked to explicitly reason about the covariation in events, they made the same inferences about the novel agent as the participants in Experiment 1.

To test our proposal, we presented a computational model that jointly infers properties of agents through Bayesian inference over a model of utility maximization. This model enabled us to generate quantitative predictions and test participants' relative judgments holistically. Overall, we found that our formalization predicted participant judgments with high accuracy. In our experiment, inferences about the final agent were tested across participants. As such, each participant only watched a single event. Thus, the graded inferences about the properties of the novel agent (see Figure 3) are not judgments that are relative to each other, but rather absolute estimates relative to past experiences.

In our experiments we clarified that the one-shot agent the square (see Figure 1)- had seen all other agents. This assumption is critical for our model, as its inferences about the agent's mental states -her preference and her perceived
strength- rely on the assumption that the agent herself had some rational estimate of the weight of the boxes.

Intuitively, if the one-shot agent had not seen the covariation information (and was therefore ignorant about the possible weight of the boxes or the strength of the other agents), then her choice would not be as revealing with respect to the strenght of her preference or her beliefs about her own strength. Consistent with this intuition, our model predicts that if the agent did not see the other agents interact with the boxes, participants should continue to infer the agent's true strength as a function of the selected box and the outcome, but they should now infer the same preference independent of the agent's choice, and they should be unable to infer her beliefs about her own strength. Future work may explore this.

Here we focused on cases where participants bring their knowledge about the world (e.g., weight of boxes) to infer properties of a new agent. As discussed in the introduction, people may also bring knowledge about agents they know to infer new properties of the world. Imagine in our paradigm, for example, if people saw the covariation data in Part 1 first, and then in Part 2, one of the agents from Part 1 interacted with a new box. In this case, our account predicts that people should be able to infer the agent's belief about the weight of the box as well as the true weight of the box from that event. Our paradigm can be flexibly adapted to explore this possibility, and future work might test this prediction.

In our experiment, some participants observed the one-shot agent interact with a new box that no one had tried lifting before (the yoyo box). Participants' inferences suggest that they did not have any prior expectations about the weight of this box (see Figure 3). In our experiment, we were clear that all the covariation agents selected the boxes in a fixed order and they only had four coins, explaining why they never tried to lift the yoyo box. If the agents from the covariation stage had freely chosen which box to play with, then their choices would suggest that the yoyo box has a low reward, or that they thought it was too heavy. In future work we may integrate choice reasoning into the covariation stage to test if people can also integrate this information when reasoning about agents.

One open question is whether the type of account that we proposed here is specific to the social domain. Although our model relies on the assumption that agents maximize utilities, much of the model relies on general principles of Bayesian inferences and inductive generalization. The logic behind these inferences -finding the causes of confounded events, and then using this knowledge to infer hidden causes of new eventsis likely to be common in non-social tasks as well (Kemp \& Tenenbaum, 2009).

In sum, our current work provides a window into the richness and the complexity of how people reason about others. Developmental work on Theory of Mind (Wellman \& Cross, 2001), and even tests of Theory of Mind used with adults (Baron-Cohen, Wheelwright, Hill, Raste, \& Plumb, 2001), often rely on inferences about a single, isolated event. How-
ever, it is important to keep in mind that we are constantly observing others' actions and their outcomes in the physical world, and reason about other people who act on the same (or similar) physical world. Exploring the social-cognitive mechanisms that underlie our ability to learn from others to learn better about others is an exciting direction for future research.

## Acknowledgments

This work was supported by the Simons Center for the Social Brain, and Varieties of Understanding grant from the John Templeton Foundation.

## References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., \& Plumb, I. (2001, February). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. J Child Psychol Psychiatry, 42(2), 241-251.
Csibra, G. (2003, March). Teleological and referential understanding of action in infancy. Philos Trans $R$ Soc Lond B Biol Sci, 358(1431), 447-458.
Gweon, H., \& Schulz, L. (2011, June). 16-Month-Olds Rationally Infer Causes of Failed Actions. Science, 332(6037), 1524-1524.
Jara-Ettinger, J., Gweon, H., Schulz, L. E., \& Tenenbaum, J. B. (2016, August). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. Trends in Cognitive Sciences, 20(8), 589-604.
Jara-Ettinger, J., Schulz, L., \& Tenenbaum, J. B. (2015). The naïve utility calculus: Joint inferences about the costs and rewards of actions. In Cogsci.
Jara-Ettinger, J., Tenenbaum, J. B., \& Schulz, L. E. (2015, May). Not so innocent: toddlers' inferences about costs and culpability. Psychological Science, 26(5), 633-640.
Kelley, H. H., \& Michela, J. L. (1980). Attribution theory and research. Annual review of psychology, 31(1), 457-501.
Kemp, C., \& Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. Psychological review, 116(1), 20.
Liu, S., \& Spelke, E. S. (2016, December). Six-month-old infants expect agents to minimize the cost of their actions. Cognition, 160, 35-42.
Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., . . . Hu, J. (2014, March). The Child as Econometrician: A Rational Model of Preference Understanding in Children. PLoS ONE, 9(3), e92160.
Repacholi, B., \& Gopnik, A. (1997). Early reasoning about desires: Evidence from 14-and 18-month-olds. Developmental Psychology, 33(1), 12-20.
Seiver, E., Gopnik, A., \& Goodman, N. D. (2012, September). Did She Jump Because She Was the Big Sister or Because the Trampoline Was Safe? Causal Inference and the Development of Social Attribution. Child Development.
Wellman, H., \& Cross, D. (2001). Theory of mind and conceptual change. Child Development.

# The Semantics and Pragmatics of Logical Connectives: Adults' and Children's Interpretations of And and Or in a Guessing Game 

Masoud Jasbi<br>masoudj@stanford.edu<br>Department of Linguistics<br>Stanford University

Michael C. Frank<br>mcfrank@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

The development of the ubiquitous logical connectives and and or provides a window into the role of semantics and pragmatics in children's linguistic development. Previous research has suggested that adults and children might differ in their interpretation of or in two ways. First, unlike adults, children might interpret or as logical conjunction, akin to and. Second, children might interpret or as inclusive disjunction while adults interpret it as exclusive. We report experimental studies that probe interpretations of and and or in adults and children using truth value judgements as well as children's spontaneous linguistic feedback. Both truth judgements and linguistic feedback showed that four-year-olds do not interpret or as and. While children's truth judgments suggested that they did not derive exclusivity implicatures, however, their corrective feedback showed signs of sensitivity to the implicature, suggesting that the truth value judgement task could have underestimated children's pragmatic competence. More generally, four-yearolds' interpretation of logical connectives may not be as different from adults as previously supposed.


Keywords: language development; semantics; pragmatics; logical connectives; disjunction; conjunction.

## Introduction

An airport sign reads "If you see something, say something." Taken literally, this is a trivial request, but readers infer an interpretation that goes far beyond the literal meanings of the words. How much of what we interpret is due to literal meaning (semantics) and how much due to our general-purpose inferential abilities (pragmatics)? In this paper, we address this question by investigating adults' and children's interpretation of the logical words and and or.

Despite their simple appearance, and and or have been a major source of insight into the contributions of semantics and pragmatics to language interpretation. The meaning of and has always been unambiguously associated with logical conjunction. For example, "There is a cat and a dog in the house." is true when the house has both a cat and a dog but false if only one or neither. The meaning of or, however, has two interpretations: inclusive disjunction and exclusive disjunction. The inclusive interpretation suggests the house has either a cat, a dog, or both. The exclusive one suggests only a cat or dog, not both. Until Grice (1975), it was generally assumed that or is ambiguous between these two meanings.

Grice (1975) argued against this ambiguity account. He maintained that the core meaning of or is inclusive disjunction but we often derive an exclusive interpretation (exclusivity implicature) by reasoning about what the speaker could have said. If the speaker meant to communicate that both a cat and a dog are in the house, s/he could have used the connective and. S/he chose or instead, so s/he did not mean
to communicate that both animals are in the house. In the Gricean account, the exclusivity implicature is not part of or's meaning, but rather the result of our reasoning on speaker's connective choice.

The advent of Gricean pragmatics shifted the focus of research in child language to the differences between adults and children in semantic vs. pragmatic aspects of interpretation. In a series of influential studies, Stephen Crain and colleagues argued that unlike adults who have an implicature-rich exclusive interpretation of or, children as young as three years old, interpret the meaning of or as inclusive disjunction (Chierchia, Crain, Guasti, Gualmini, \& Meroni, 2001; Crain, 2012). They argued that children develop the semantics of or before its pragmatics: they interpret or as inclusive disjunction but fail to enrich it with the exclusivity implicature the way adults do. Therefore, the main difference between children and adults is that children interpret or as inclusive, but adults interpret it as exclusive.

Recent investigations have added a new level of complexity to this line of research. Tieu et al. (2017) and Singh et al. (2016) argued that a large group of children in their studies (30-40\% of the participants) interpreted or as logical conjunction. In other words, these children did not differentiate between and and or. They argue that this conjunctive interpretation of $o r$ is due to non-adult-like pragmatic reasoning: children interpret $A$ or $B$ as $A$ or $B$ or both, but not only $A$, and not only $B$; therefore both $A$ and $B$.

The current paper seeks to fill two gaps in the current literature. First, previous research has focused on children's interpretation of and and or in complex sentences - for example with other logical words such as quantifiers every and none. In this paper we test children and adults' understanding of and and or in simple existential sentences like "There is a cat or a dog." Second, previous research has tested children and adults using the binary truth value judgment task (Crain \& Thornton, 1998). In such tasks participants are asked whether a puppet's statement is right or wrong. In this study, we allow participants to make use of three options: wrong, kinda right, and right. Katsos \& Bishop (2011) argued that ternary judgment tasks are better suited for assessing children's pragmatic competence.

This paper addresses two main questions. First, do children interpret or as logical conjunction (similar to and)? Second, do children understand or as inclusive disjunction, or exclusive disjunction? We conduct two experiments to answer these questions. Experiment 1 tests adults' interpreta-
tions and sets the benchmark for our child study. Experiment 2 investigates children's truth value judgments in a guessing game as well as their spontaneous linguistic feedback in the same task. Considering the first question, neither children's truth value judgement nor their linguistic feedback support the hypothesis that a large group of them interpret or as logical conjunction. For the second question, children's judgments suggest that unlike adults, they do not derive exclusivity implicatures and interpret or as inclusive disjunction. However, children's spontaneous linguistic feedback shows signs of sensitivity to the exclusivity implicature of or.

The next two sections present experiments 1 and 2 and the last section discusses the implications of these findings for theories of semantic and pragmatic development. For further details of the methods as well as the data and statistical analyses, please visit the paper's online repository ${ }^{1}$

## Experiment 1: Adults

## Methods

Participants We recruited 52 English speaking adults online using Amazon's Mechanical Turk (MTurk).

Materials and Design The experimental game included several cards with cartoon images of either one or two animals. The animals included a cat, a dog, and an elephant. Figure 1 shows two example cards. The game also used three types of guesses: simple (e.g. There is a cat), conjunctive (e.g. There is a cat and a dog), and disjunctive (e.g. There is a cat or a dog). Pairing the cards with the guesses resulted in 6 types of card-guess scenarios. Figure 1 shows examples for four critical scenarios. Overall, the animal labels used in the guess and the animal images on the card may have no overlap (e.g. Image: dog, Guess: There is a cat or an elephant), partial overlap (e.g. Image: Cat, Guess: There is a cat or an elephant), or total overlap (e.g. Image: cat and elephant, Guess: There is a cat or an elephant). Crossing the number of animals on the card, the type of guess, and the overlap between the guess and the card resulted in 12 different trial types.

Procedure The experiment had three phases: introduction, instruction, and test. In the introduction, participants saw six sample cards and read that they will play a guessing game with them. Then a blindfolded cartoon character named Bob appeared on the screen and they were told that in each round of the game, they will see a card and Bob is going to guess what animal is on it. We emphasized that Bob cannot see anything. We asked participants to judge whether Bob's guess is wrong, kinda right, or right. In the instruction phase, participants saw a card with the image of a dog and were told that Bob guessed There is a cat on the card. All participants (correctly) responded with wrong.

In the test phase, participants saw one trial per trial type for the total of 12 trials. Within each trial type, the specific card-guess scenario was chosen at random. The order of trial types was also randomized.

[^110]| Guess |  | One Animal (1T) | Two Animals (2T) |
| :---: | :---: | :---: | :---: |
| AND | There is a cat and a dog. | $8$ | $\begin{gathered} 50 \\ 50 \end{gathered}$ |
| OR | There is a cat or a dog. |  | \% |

Figure 1: Critical trials with example cards.

## Results

Here we focus on the results of the critical trial types, pictured in Figure 12 We identify these trials using two features: 1. the connective used for guessing (AND vs. OR) 2. the number of true conjuncts/disjuncts, which corresponds to the number of animals on the card. When only one animal is on the card, only 1 conjunct/disjunct is true (1T) and when two animals are on the card, both conjunct/disjuncts are true (2T).

Adult responses differed both by the connective used and the number of true conjuncts/disjuncts (Figure 2). First, the response pattern in AND trials is different from the one in OR trials. For AND, the responses were on the extremes of right and wrong while for OR, they were distributed on kinda right and right. This pattern suggests that adults interpret and and or differently. Second, the responses were different between the trials where one disjunct/conjunct was true (1T) and those where both disjuncts/conjuncts were true (2T). This difference was greater for conjunction than disjunction. Adults showed a small preference for the use of disjunction when only one disjunct was true. This pattern suggests a small preference for an exclusive interpretation of $o r$ in the guessing game.
Individual Responses In order to understand how participants interpret disjunction, Tieu et al. (2017) and Singh et al. (2016) categorized participants as a function of their responses to the disjunctive trials. Here we perform a similar analysis. In this study, none of the adults considered a disjunctive guess wrong when one or both of the animals were on the card. However, the participants' kinda right and right responses divided them into four categories.

The largest group of participants (23 out of 52) considered the disjunctive guess right when one animal was on the card

[^111]

Figure 2: Adult judgments in critical trials of Experiment 1. Error bars represent $95 \%$ confidence intervals.
(1T), but kinda right with both animals were present (2T). This pattern is consistent with an interpretation of "or" with an exclusivity implicature. The use of disjunction when both disjuncts are true is not wrong but it is nevertheless infelicitous and not completely right. For these participants, kinda right captures the violation of such a pragmatic expectation.

The other 29 participants divided almost equally into three groups. Ten participants rated disjunctive guesses as right in both scenarios where one or two animals were on the card. This pattern is consistent with an inclusive interpretation of or, in which adults do not derive an exclusivity implicature. It is also compatible with some adults being tolerant towards violations of the exclusivity implicature.

Nine other participants rated disjunctive guesses as only kinda right in both one-animal and two-animal trials. In other words, disjunctive guesses were dispreferred regardless of the outcome. This response pattern is consistent with the violation of another pragmatic expectation in the context of a guessing game: the guesser must choose the most specific guess possible. Under this expectation, guesses that cover several possible outcomes are punished. A disjunctive guess never picks a specific outcome and it is possible that for these participants, kinda right captures the violation of this specificity expectation.

Finally, nine participants ( $17 \%$ of participants) reported a disjunctive guess as right when both animals were on the card, but only kinda right when only one of the animals was on the card. In other words, these participants preferred the guess when both disjuncts where true rather than only one. It is possible to interpret such a response profile as some adults interpreting or as logical conjunction. However, it is also possible that these adults considered the goal of the game to be choosing the right animals and did not think the choice of the connective should matter for the purposes of the guessing


Figure 3: Children's judgments for critical trials in Experiment 2. Error bars represent 95\% confidence intervals.
game. In other words, they may have interpreted a right guess as one that picks the correct animals out of the possible set of animals in the game, regardless of the connective used.

The analysis of individual response profiles shows that there is a good deal of variability in the response profiles of adults. However, since we have not systematically manipulated the possible interpretations mentioned above and accounted for noise and chance variation, we remain cautious in our interpretation of participants' response profiles here.

## Discussion

In this study, we tested adult interpretations of the connective words and and or in the context of a guessing game. Adult participants interpreted these words differently and depending on how many disjuncts/conjuncts were satisfied. Overall, a guess with and was considered right if both conjuncts were true and wrong if only one was true. A guess with or was not wrong in either case, yet adults were more likely to consider it as right when only one of the disjuncts was true. Grouping individuals based on their response profiles, we found that some participants dispreferred disjunctive guesses whether one or both disjuncts were true, some considered them better when both disjuncts where true, and some others considered them right in either case.

The results are consistent with the dominant view on the division of labor between semantics and pragmatics in the interpretation of connective words. The semantics of and is captured by logical conjunction and or by inclusive disjunction. And is true when both conjuncts are true and false when only one is true. $O r$ is true in both cases but is not the best option as a connective when both disjuncts are true. In Experiment 2 we examine preschool children's interpretation of these connectives in the context of the same guessing game.

## Experiment 2: Children

## Methods

Participants We recruited 42 English speaking children from the Bing Nursery School at Stanford University. Children were between $3 ; 02$ and $5 ; 02$ years old (Mean $=4 ; 04$ ).

Materials and Design We used the same set of cards and linguistic stimuli as the ones in Experiment 1. The study used 8 trial types and 2 trials per trial type for a total of 16 trials. The trials were balanced to include the same number of one-animal and two-animal cards, the same number of simple and connective guesses, and the same number of expected true vs. false judgments. However, we made a few changes to make the design more suitable for children. Instead of Bob, a puppet named Jazzy played the game with the children. Jazzy wanted to guess what animals were on the cards without seeing them. So he had a sleeping mask on his eyes during the game. Children knew that Jazzy likes guessing but they did not know why Jazzy would choose to guess the way he does; namely, sometimes with simple sentences and sometimes with conjunctions or disjunctions. To introduce a three-valued reward scale similar to the verbal responses wrong, kinda right, and right, we placed a set of red circles, small blue stars, and big blue stars in front of the children. These tokens were used to reward the puppet after each guess.

Procedure The experiment was carried out in a quiet room and the sessions were videotaped. There was a small table and two chairs in the room. Children sat on one side of the table and the experimenter and the puppet on the other side facing the child. The groups of circles, small stars, and big stars were placed in front of the child from left to right. A deck of six cards was in front of the experimenter. Similar to the adult study, participants sat through three phases: introduction, instruction, and test.

The goal of the introduction phase was to show the animal cards to children and make sure they recognize the animals and know their names. The experimenter showed the cards to the children and asked them to label the animals. All children recognized the animals and could label them correctly. In the instruction phase, children went through three example trials. The experimenter explained that he is going to play with the puppet first so that the child can learn the game. He removed the six introduction cards and placed a deck of three cards face-down on the table. From top to bottom (first to last), the cards had the following images: a cat, an elephant, a cat and a dog. He put the sleeping mask on Jazzy's eyes and explained that Jazzy is going to guess what is on these cards. He then picked the first card and asked the puppet: "What do you think is on this card?" Jazzy replied with "There is a dog". The experimenter showed the cat-card to the child and explained that when Jazzy is not right he gets a circle. He then asked the child to give the puppet a circle. Rewards were collected by the experimenter and placed under the table to not distract the child. The second trial followed the same pattern except
that the puppet guessed right and the experimenter invited the child to give the puppet a big star. In the final trial, the puppet guessed that there is a cat on the card when the card had a cat and a dog on it. The experimenter said that the puppet was $a$ little right and asked the child to give him a little star.

In the test phase, the experimenter removed the three instruction cards and placed a deck of 16 randomized cards face-down on the table. In all trials of the study, the face of the card was shown to the child after the puppet's guess. The experimenter explained that it was the child's turn to play with the puppet.
Offline Coding of Linguistic Feedback We also coded children's spontaneous linguistic feedback to the puppet when they saw the card. There were four types of feedback: 1. None, 2. Judgments, 3. Descriptions, and 4. Corrections. None refers to cases where children did not provide any linguistic feedback. Judgments refers to linguistic feedback such as you are right!, yes, nope, you winned. Such feedback expresses whether the puppet was right or not. Descriptions were cases that the child simply mentioned what was on the card with no added lexical item or prosodic stress: cat!, dog and elephant!', There is a cat and a dog! etc. Finally, corrections referred to feedback that provided corrections to what the puppet had said using extra words or prosodic stress. Examples include: cat AND dog (with emphasis placed on and), Both!, The two are!, Just a cat!, Only cat.

## Results

Figure 3 shows the results for the critical conditions in Experiment 2. Comparing the AND and OR trials (Figure 3 rows), we see that children distinguish between and and or in cases where one animal is on the card but not when both are. Given that the one-animal conjunction trials (top left) and the one-animal disjunction trials (bottom left) differ in truth conditions, the difference in response patterns suggests that children at this age have a different semantic knowledge for and and or. The two-animal conjunction and two-animal disjunction trials (top right and bottom right) do not differ in truth values, and the responses also show no difference.

In the one-animal and two-animal trials, children show different response patterns when the guess contains the conjunction word and (top right vs. top left) but not when or is used (bottom right vs. bottom left). Since the truth values of oneanimal and two-animal trials differ for conjunctive guesses but not disjunctive ones, the results suggest that children have different semantic knowledge for and and or. The similarity of the disjunctive guesses in one-animal and two-animal trials (bottom right vs. bottom left) can be interpreted as a lack of exclusivity implicatures in children.
Statistical Modeling We used the R package $\{r \operatorname{stan}\}$ for Bayesian statistical modeling. We fit separate ordinal mixedeffects logistic models for children's and adults' judgments. The response variable had three ordered levels: wrong, kinda right, and right. The trial types One-Animals-OR, Two-


Figure 4: Coefficients capturing the relevant comparisons across conditions across the two experiments (see text). Error bars represent $99 \%$ regions of highest posterior density.

Animals-OR, One-Animal-AND constituted the (dummycoded) fixed effects of the model with Two-Animal-AND set as the intercept. The model also included by-subject random intercepts. The priors over trial types and the random intercepts were set to $\mathcal{N}\left((0,10)\right.$. We also included parameters $C_{1}$ and $C_{2}$, the two cutpoints delimiting the logistic for 1) wrong and kinda right and 2) kinda right and right responses, drawn with the prior $\mathcal{N}(0,1)]^{3}$ All four chains converged after 3000 samples (with a burn-in period of 1500 samples)

We make inferences based on the highest-posterior density (HPD) intervals for the coefficients estimated from each model. Because predictors are dummy-coded, we can examine contrasts of interest by computing the difference between coefficients for pairs of conditions we wish to contrast (Figure 4). Overall, adults' and children's estimated coefficients are similar in sign to one another, though adults are more extreme. The one notable exception to this pattern is for the contrast or, $1 T$ vs. $2 T$, which shows the comparison between the disjunctive trials: both disjuncts are true vs. only one disjunct is true. On average, children are more positive for disjunction on two-animal trials, while adults are more negative. These estimates reflect the exclusivity implicature that adults compute, leading them to judge two-animal trials as more kinda right.

Individual Responses Children showed a wide variety of response profiles for disjunction trials. This was partly because each child responded to two trials per trial type: two one-animal disjunction trials and two two-animal disjunction trials. The largest group ( 10 out of 42) responded with right to all four trials. Six children responded with right to all trials except one one-animal trial that they responded to with kinda right. Six other children responded with kinda right to both one-animal trials and right to both two-animal trials.

However, the main goal of analyzing the response profiles

[^112]was to find children that demonstrated conjunctive readings of or. In order to find such children, we adopted a (lenient) measure: any preference for or when both disjuncts were true was considered a conjunctive profile. More specifically, either the child responded with wrong when one disjunct was true but kinda right or right when both were true; or, the child responded with kinda right when one disjuncts was true but right when both were true. We found 10 children ( $24 \%$ of participants) that matched this profile. In Experiment 1 we found nine adults ( $17 \%$ of participants) who matched such a profile. Furthermore, as explained earlier, such a response profile is also compatible with a different construal of the guessing game in which the goal is to pick the right animals regardless of the logical connective. Therefore, we conclude that the analysis of participants' response profiles did not provide any evidence for the hypothesis that a large group of four-year-old children interpret or as logical conjunction.

Linguistic Feedback We next examined children's linguistic feedback to the puppet (Figure 5). In all critical trials, we found similar proportions of None responses: no comment on the puppet's guess and only rewarding the puppet. However, the proportions of other feedback categories differed between trial types. We performed chi-squared tests of homogeneity to compare the feedback distributions.

In the AND trials, a comparison of the feedback distribution in one-animal and two-animal conditions was statistically significant $(\chi(3,167)=35.99, p<.0001)$, indicating different feedback for true vs. false sentences. In the OR trials, we find a similar significant difference between oneanimal and two-animal trials, suggesting children's sensitivity to the exclusivity implicature of $\operatorname{or}(\chi(3,166)=11.11, \mathrm{p}=$ 0.01 ). In both cases, children's corrective feedback increases for false (AND - one animal) and infelicitous trials (OR - two animals). There was no significant difference between these false and infelicitous trials $(\chi(3,166)=3.19, \mathrm{p}=0.36)$.

The one-animal disjunctive trials (bottom left) showed the highest proportion of Descriptions. These are trials in which the guess is correct but not specific enough: it leaves two possibilities open. These trials were significantly different than the one-animal trials for conjunction $(\chi(3,166)=24.29$, $\mathrm{p}<.0001$ ). Finally, the two-animal conjunctive trials (top right) showed the highest proportion of Judgments such as you are right!. This is not surprising given that in these trials represent the most optimal guessing scenario. These trials had a significantly different feedback distribution from the matching disjunction trials $(\chi(3,167)=42.37, \mathrm{p}<.0001)$.

## Discussion

This study did not find evidence for the hypothesis that a large group of four-year-old children interpret the disjunction word or similar to its conjunctive counterpart and. To the contrary, both children's judgments and their linguistic feedback suggested that they differentiate these two connectives. Instead, children's judgments largely mirrored those of adults. We take this as a sign of children's adult-like semantics for


Figure 5: Children's Linguistic Feedback to Conjunction and Disjunction Trials. Error bars represent $95 \%$ confidence intervals.
and and or. Considering pragmatic inferences with or, children's truth value judgments did not differentiate between trials where one disjunct was true and those where both were true. However, their linguistic feedback to the puppet did differentiate these two trial types. Children provided more corrective feedback when both disjuncts were true, indicating sensitivity to the exclusivity implicature of or.

## General Discussion

We began with two questions. First, do adults/children differentiate or from and? Second, do adults/children interpret or as inclusive disjunction or exclusive disjunction? We presented two studies to address these questions.

For the first question, we reported truth value judgement results as well as results from children's linguistic feedback that suggested both adults and children differentiate or from and. Crucially, children showed different judgments for false vs. true guesses, suggesting that they understand the core semantics of these connectives.

For the second question, adult truth value judgments of or were split between an inclusive and an exclusive interpretation in the guessing game, with a slight advantage for the exclusive interpretation. Children's judgments suggested that they interpret or as inclusive disjunction and do not derive an exclusivity implicature. However, children's spontaneous linguistic feedback in the same task showed signs of sensitivity to the exclusivity implicature of or. In other words, when both disjuncts were true children considered the guess right but corrected the puppet with utterances such as cat AND dog, both!, the two are!.

Based on the truth value judgement results, it is possible to conclude that children, unlike adults, do not derive an exclusivity implicature for or. However, children's spontaneous
linguistic feedback raises another possibility: while the truth value judgement task reflected children's semantic knowledge well, it could have underestimated children's pragmatic competence. We would like to explore this possibility more systematically in a future study.

Overall, our results point to the importance of assessing the semantics and pragmatics that children assign to connectives across a wide variety of contexts and using different measures. Although individual experimental trial types can appear consistent with multiple interpretations, the profile of responses across trial types can be revealing of the underlying representations. More broadly, the investigation of how children acquire semantic representations for logical connectives - and in particular, how they infer an inclusive semantics for or - is an important puzzle for future investigations of early word learning.

## Acknowledgements

We would like to thank Eve V. Clark and Christopher Potts for their generous help and advice on this project. We would also like to thank the staff and children at Stanford's Bing Nursery School. This work was supported by NSF \#1456077 and the Stanford University fellowship to Masoud Jasbi.

## References

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., \& Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In Proceedings of the 25th Boston University conference on language development (pp. 157-168). Cascadilla Press Somerville, MA.
Crain, S. (2012). The emergence of meaning. Cambridge University Press.
Crain, S., \& Thornton, R. (1998). Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics. MIT Press.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and semantics (Vol. 3: Speech Acts, pp. 43-58). Academic Press.
Katsos, N., \& Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. Cognition, 120(1), 6781. Retrieved from http://www.sciencedirect.com/ science/article/pii/S0010027711000758
Singh, R., Wexler, K., Astle-Rahim, A., Kamawar, D., \& Fox, D. (2016). Children interpret disjunction as conjunction: Consequences for theories of implicature and child development. Natural Language Semantics, 24(4), 305352. http://doi.org/10.1007/s11050-016-9126-3

Tieu, L., Yatsushiro, K., Cremers, A., Romoli, J., Sauerland, U., \& Chemla, E. (2017). On the role of alternatives in the acquisition of simple and complex disjunctions in french and japanese. Journal of Semantics, 34(1), 127. http: //doi.org/10.1093/jos/ffw010

# Maintaining Credibility When Communicating Uncertainty: The Role of Communication Format 

Sarah C. Jenkins (s.jenkins.12@ucl.ac.uk) and Adam J. L. Harris (adam.harris@ucl.ac.uk)<br>Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK<br>R.M. Lark (mlark@bgs.ac.uk)<br>British Geological Survey, Keyworth, Nottingham, NG12 5GG, UK


#### Abstract

Research into risk communication has commonly highlighted the disparity between the meaning intended by the communicator and what is understood by the recipient. Such miscommunications will have implications for perceived trust and expertise of the communicator, but it is not known whether this differs according to the communication format. We examined the effect of using verbal, numerical and mixed communication formats on perceptions of credibility and correctness, as well as whether they influenced a decision to evacuate, both before and after an 'erroneous' prediction (i.e. an 'unlikely' event occurs, or a 'likely' event does not occur). We observed no effect of communication format on any of the measures pre-outcome, but found the numerical format was perceived as less incorrect, as well as more credible than the other formats after an 'erroneous' prediction, but only when low probability expressions were used. Our findings suggest numbers should be used in consequential risk communications.


Keywords: verbal probability expressions; numerical probabilities; risk communication; trust; expertise; credibility

## Introduction

Science is suffering from a 'crisis of trust' (House of Lords, 2000); preserving and cultivating the public's trust has never been more important for the scientific community (Nature, 2010). Uncertainty is an inescapable part of any scientific endeavour, but the presence of it creates doubt in the minds of the public and it is often used as a reason to delay taking action (Lewandowsky, Ballard, \& Pancost, 2015). Effectively communicating information regarding risk and uncertainty thus represents a significant problem for scientists.

Methods for communicating risk and uncertainty include using verbal probability expressions (VPEs; e.g. 'possible', 'likely'), numerical expressions (e.g. '20\% likelihood'), or mixed expressions (e.g. 'unlikely [20\% likelihood]'). Budescu and Wallsten (1995) proposed that the choice of format for communicating likelihood information should be governed by the congruence principle: the precision of the communication should be consistent with the degree of certainty that can reasonably be expected for estimates about the event described. Much research has investigated the pitfalls of using VPEs to communicate uncertainty using the 'how likely' translation approach, whereby people are asked to translate a VPE to a corresponding numerical probability

This has highlighted the variability in people's usage and interpretations (e.g., Budescu \& Wallsten, 1985), as well as the influence of other contextual and cultural factors (e.g., Bonnefon \& Villejoubert, 2006; Harris \& Corner, 2011; Harris, Corner, Xu, \& Du, 2013; Teigen \& Brun, 1999, 2003; Weber \& Hilton, 1990). Such variability clearly highlights the potential for a reduction in perceived credibility of the communicator, if there is a disparity between the meaning intended by the communicator and that which is understood by the recipient.
A commonly suggested solution to the problems of miscommunication is to use a dual-scale, mixed format expression to communicate risk and uncertainty, for example 'It is unlikely (less than 33\%)' (e.g., Budescu, Broomell, \& Por, 2009; Budescu, Por, Broomell, \& Smithson, 2014; Harris \& Corner, 2011; Harris et al., 2013; Patt \& Dessai, 2005; Witteman \& Renooij, 2003). Using such a 'verbalnumerical' (V-N) format was found to increase correspondence between people's interpretations and the IPCC guidelines, an effect that replicated across 24 countries (Budescu et al., 2014). However, when shown a histogram of potential outcomes and asked to complete probability statements (e.g., "It is unlikely that the lava flow will extend to a distance of _km'), the so-called 'which outcome' approach to studying VPEs (e.g., Teigen, Juanchich, \& Riege, 2013), participants tended to complete the sentence with a distance that exceeded any represented in the histogram, both for 'unlikely' and 'unlikely ( $20 \%$ chance)' (Jenkins, Harris, \& Lark, 2016; see also Juanchich \& Sirota, 2016). If such phrases are seen as appropriate for communicating an outcome with a $0 \%$ chance of occurring, the mismatch between this and an intended communication of ' $20 \%$ likelihood' could adversely affect confidence in subsequent communications.

Aside from the terms used, a further problem arises from people's general understanding of uncertainty and probability. Uncertainty is often perceived by the public as an 'indicator of ignorance', when in fact it should be seen as a source of actionable knowledge (Lewandowsky et al., 2015). Scientific forecasts are probabilistic (at best) and thus it is, for example, not possible to predict with certainty the probability of a volcanic eruption on a given day. Even if an event is predicted to be 'likely' to occur, the very fact it is not certain means that it might still not happen. In the same vein, even if an event is predicted as 'unlikely' to occur (e.g. 20\% likelihood; Theil, 2002), it does not mean the event will
definitely not occur, given that one in five times it will (on a frequentist interpretation of probability). The expectation of what will happen is largely driven by the directionality of the expression (Teigen \& Brun, 1995, 1999); in that phrases which have negative directionality (e.g. 'unlikely') focus one's attention on the non-occurrence of the event, whereas those with positive directionality (e.g. 'likely') focus on the occurrence of the event. If the outcome is 'opposite' to what was predicted, the predictions are often seen as 'erroneous', which could have a knock-on effect on perceived credibility.

Despite recent calls to use a dual-scale communication format, research has yet to explore the effect of using mixed expressions on the perceived credibility of the communicator. Neither, perhaps more importantly, has it investigated the consequences of 'erroneous' predictions on credibility. Given a major function of risk communication is providing trustworthy information, confidence in the source of the information is key (Kasperson, 2014). After all, even if the information is understood as intended, it is of no use if the communicator is not perceived as credible and thus is not trusted enough to inspire action on the basis of the communication. Indeed, credibility has been found to influence risk perceptions. Trust is negatively associated with perceived risk (Sjöberg, 2001), as well as directly affecting behaviour (Wachinger, Renn, Begg, \& Kuhlicke, 2013).

Longman, Turner, King, \& McCaffery (2012) explored the effect of numerical formats on accuracy of understanding, perceived risk, and source credibility judgements for two different sources of risk information (clinician / pharmaceutical company). The risk estimate was presented either as a either a point ( 20 out of 100), small range ( $16-24$ out of 100) or large range ( $8-32$ out of 100). Range information resulted in reduced understanding and the large range was perceived as more risky compared to a point estimate. Experts using point estimates were viewed as more credible. Gurmankin, Baron and Armstrong (2004) investigated the effect of verbal and numerical statements of risk (percentage / fraction) on trust and comfort in a physician in a hypothetical medical communication. They found subjects were more trusting of, and more comfortable with, numerical versions of the information, though this effect decreased with lowering levels of numeracy, highlighting the importance of including a numeracy measure in the current study.

The importance of investigating the credibility of the communicator cannot be understated. Whilst an accurate understanding of information is clearly desirable, it is people's actions (on the basis of the communication) which matter, given they will have the most consequences for the individual. Therefore an investigation into the effects of communication format should also consider the effect of communication format on people's actions. Doyle, McClure, Paton, \& Johnston (2014) found that fewer people suggested evacuating when the risk of a volcanic eruption was described using verbal terms than when using numerically equivalent terms, suggested to be a result of the fact that VPEs are viewed as more ambiguous, though again the study did not
consider mixed-formats, or the influence of 'erroneous' predictions.

Although previous research has demonstrated the V-N format aids understanding in risk communications (Budescu et al., 2014), it may not be the preferred format for the recipient. Indeed, there may be a discrepancy between what people favour (for instance the preference for receiving information in numerical form, Erev \& Cohen, 1990) and what experts can suitably provide. Using a numerical point estimate (e.g. 15\%) to describe the chance of a natural hazard (which are, by nature, highly uncertain) might be perceived as overly precise according to the congruence principle (Budescu \& Wallsten, 1995) and thus not credible.

A deeper understanding of the effects of using different communication formats and the consequences of 'erroneous' predictions is therefore clearly required, such that the public's trust in science can be built and maintained. We thus sought to examine whether initial perceptions of credibility in the communicator differed according to communication format over two studies featuring low and high probability events. We also investigated whether these perceptions changed after an 'erroneous' prediction (i.e. the 'unlikely' outcome occurred, in Study 1, or the 'likely' outcome did not occur, in Study 2). Ascertaining the effect of these factors is instructive for developing effective risk communication strategies.

## Study 1

## Method

## Participants

300 Native English speakers ( 146 male) aged between 18 $72(M d n=33.5)$ were recruited from Prolific Academic (PA; www.prolific.ac). Participants received $£ 0.75$ for participating.

## Design

A $4 \times 2$ mixed design was used. Communication format was in the low probability domain and had four levels, manipulated between participants: verbal- "unlikely", numerical- " $20 \%$ likelihood", V-N- "unlikely ( $20 \%$ likelihood)" and N-V- " $20 \%$ likelihood (unlikely)." Outcome (pre/post) was a within-participants variable.

Perceptions of trust, expertise, correctness and decision to evacuate were rated on five-point scales. Expertise was operationalised as 'How knowledgeable does the expert seem?' from 1 - 'Not at all knowledgeable' to 5 - 'Extremely knowledgeable'. Trust was operationalised as 'How much do you trust that the expert is giving you complete and unbiased information?' (Dieckmann, Slovic, \& Peters, 2009), from 1 'Not at all' to 5 - 'A great deal'. Decision to evacuate, based on Doyle et al. (2014), was rated from 1- 'Definitely should evacuate today' to 5 - 'Definitely should not evacuate today'. Participants also then had to indicate why they made that decision. Correctness was rated from 1 - 'Not at all correct' to 5 - 'Completely correct'.

## Materials and Procedure

After consenting to participate, participants indicated their age, gender and Prolific ID before reading the introductory text. The introductory text informed participants that they would see a geological scenario and be asked to make a series of judgements about this. On the next screen, participants read a vignette about a current volcanic eruption, in which lava flows were expected. A volcanologist presented a communication about the probability of the lava flows travelling a certain distance:
"Mount Ablon has a history of explosive eruptions that
have produced lava flows. An eruption is currently
underway and lava flows are expected. Volcanologists
from Ablon Geological Centre are communicating
information about the volcano. A volcanologist has
suggested that, given the volcano's recent history, there
is a $\mathbf{2 0 \%}$ likelihood (unlikely) that the lava flow will
extend 3.5 km from the point of eruption."
Participants then provided initial ratings of expertise and trust in the expert's prediction of events. On the subsequent screen, participants were informed that the capital city was at risk of the volcanic eruption and asked to rate whether to evacuate the city today or not (Doyle et al., 2014). A mass evacuation was described as being 'very expensive and extremely disruptive to residents'.

Participants were then informed on the following screen that the unlikely outcome did in fact occur. They were asked to provide further trust and expertise ratings, as well as rating how correct the volcanologist's prediction was in light of the outcome. The next screen then showed a similar communication by a volcanologist about Mount Ablon, set two years on, with participants asked the two evacuation questions, as before.

Finally participants completed a numeracy scale (Lipkus, Samsa, \& Rimer, 2001), with two additional questions from the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, \& Garcia-Retamero, 2012) included to increase variability in scores, given previous studies using PA have found it to be a highly numerate sample. After completing the study, participants were given a code to claim their reward, thanked and debriefed.

## Results

There was a significant correlation between trust and expertise ratings, both pre-outcome, $r=.69, p<.001$ and post-outcome, $r=.74, p<.001$. For ease of exposition, we averaged the measures to create a single measure of credibility. The data were entered into a 4 (communication format) $\times 2$ (outcome) $\times 2$ (numeracy) ANOVA, unless stated otherwise.

Given the highly skewed distribution of responses, participants with scores of eight or under were classed as low numeracy and those with nine or above classed as high numeracy. However, given there was only one effect of (or interaction involving) numeracy across Studies 1 and 2, this variable is only considered further in that single instance.

## Credibility Ratings

Mean credibility ratings, by communication format, are plotted in Figure 1, which suggests that pre-outcome there was little difference between formats. All communication formats suffered from a loss of perceived credibility postoutcome, but there was less of a reduction in the numerical format. Correspondingly, there was a main effect of outcome, $F(1,292)=218.60, p<.001, \eta_{p}^{2}=.43$, and format, $F(3,292)$ $=5.77, p<.01, \eta_{p}^{2}=.06$, but this was qualified by a significant interaction between outcome and format, $F(3,292)=6.87, p$ $<.001, \eta_{p}^{2}=.07$. Simple effects analyses confirmed no effect of format pre-outcome $F(3,296)=0.38, p=.77$, and a significant effect of format post-outcome $F(3,292)=8.02, p$ < .001. It is worth noting, however, that the reduction in credibility was still significant even in the numerical condition, $t(73)=3.66, p<.001, d=0.43$.


Figure 1. Effect of Communication Format on Perceptions of Credibility Before and After an 'Erroneous' Prediction (Error Bars Represent $\pm 1$ SE) - Study 1 - Low Probability.

## Decision to Evacuate

Mean evacuation ratings both pre- and post-outcome, by communication format, are displayed in Table 1, which shows a slight difference between communication formats prior to the outcome. Post-outcome, there was a shift to being more certain about evacuating today. There was a main effect of outcome, $F(1,292)=98.19, p<.001, \eta_{p}^{2}=.25$ and format, $F(3,292)=5.59, p<.01, \eta_{p}^{2}=.05$. Participants were more certain about evacuating today in the verbal condition and least certain decision in the N-V condition. There were no significant interactions (all ps > .12).

## Correctness Ratings

A one-way ANOVA revealed a significant effect of communication format on correctness ratings, $F(3,292)=$ 26.32, $p<.001, \eta_{p}^{2}=.22$, corresponding to the differences in the credibility ratings. From Table 1, the numerical format was seen as 'least incorrect' and the verbal format seen as most incorrect.

## Study 2

## Method

## Participants

299 Native English speakers were recruited from Amazon MTurk. 17 cases were removed for failing the attention check, leaving a final sample of 281 participants ( 138 male) aged between $18-74$ ( $M d n=32$ ). Participants received $\$ 0.60$ for participating.

## Design, Materials and Procedure

As in Study 1, except communication format was set in the high probability domain: verbal - "likely", numerical - " $80 \%$ likelihood", V-N - "likely ( $80 \%$ likelihood)" and N-V-" $80 \%$ likelihood (likely)". In addition, post-outcome, the likely event did not occur.

## Results

Trust and expertise ratings were again correlated (preoutcome: $r=.60, p<.001$; post-outcome: $r=.74, p<.001$ ). We combined the two measures as in Study 1. The data were analysed as in Study 1.

## Credibility Ratings

Mean credibility ratings, by communication format, are plotted in Figure 2, which shows before the outcome there was little difference between formats, as in Study 1. Postoutcome, all communication formats suffered from a loss of perceived credibility, with no notable difference between formats. The outcome and format interaction of Study 1 was not replicated, $F(3,273)=2.53, p=.06$. The main effect of outcome was significant, $F(1,273)=221.23, p<.001, \eta_{p}^{2}=$ .45, and the effect of format was marginally significant, $F$ (3, 273) $=2.59, p=.053, \eta_{p}^{2}=.03$. A post-hoc Gabriel test revealed there were no significant differences between formats (all $p \mathrm{~s}>.08$ ). Highest perceptions of credibility were in the numerical condition ( $\mathrm{M}=3.91, \mathrm{SE}=0.08$ ), and the lowest were in the verbal condition $(\mathrm{M}=3.63, \mathrm{SE}=0.08)$.

## Decision to Evacuate

Mean evacuation ratings for both pre and post-outcome (by communication format) are displayed in Table 1, which shows little difference between formats both pre and postoutcome. Indeed, there was no significant effect of outcome ( $p=.07$ ) nor format $(p=.20)$ on the decision to evacuate. There was a significant effect of numeracy, $F(1,273)=5.08$, $p<.05, \eta_{p}^{2}=.02$, with the high numeracy group more certain about evacuating ( $\mathrm{M}=2.08, \mathrm{SE}=0.10$ ), compared to the low numeracy group ( $\mathrm{M}=2.39, \mathrm{SE}=0.10$ ). There were no significant interactions (all ps > .15).

## Correctness Ratings

Again there was a significant effect of communication format on correctness ratings $F(3,273)=4.90, p<.01, \eta^{2}{ }_{\mathrm{p}}=.05$. As in Study 1, the numerical format was seen as 'least incorrect' and the verbal format seen as most incorrect (see Table 1).


Figure 2. Effect of Communication Format on Perceptions of Credibility Before and After an 'Erroneous' Prediction(Error Bars Represent ${ }_{-} 1$ SE) - Study 2 - High Probability.

## General Discussion

Pre-outcome, people did not perceive any of the volcanologists to be more credible than others using different communication formats, nor was there an effect of format on decision to evacuate. However, post-outcome, credibility was sensitive to an 'erroneous' prediction, with lower ratings in all formats. In Study 1 (low probability), the numerical format was affected least by this, and there was a trend for numerical-led communications (numerical and $\mathrm{N}-\mathrm{V}$ ) to be least affected in Study 2.

It is surprising that there was no initial difference between communication formats on perceptions of credibility in either probability domain, given the findings of Longman et al. (2012) that an expert who used a point estimate was seen as more credible. We would have expected numerical communications to have been rated as more credible, as the decision to use a precise numerical estimate could be thought to reflect a level of confidence and certainty in the prediction. Indeed, people expect experts to provide their knowledge in a precise manner (Shanteau, 1992).

In Study 1, the finding of most interest was the presence of a format $\times$ outcome interaction, whereby the numerical format lost least credibility following the occurrence of the unlikely event. These findings could be partly attributed to the directionality of the expression (Teigen \& Brun, 1995, 1999). Although both V-N and N-V formats featured a negatively directional expression ('unlikely'), it was accompanied by the positively directional phrase ' $20 \%$ likelihood', which may have cancelled out the effect of the negative directionality. Although no significant interaction was observed with high probability expressions, the results followed a similar trend, with numerical and $\mathrm{N}-\mathrm{V}$ expressions least affected.

We were surprised not to replicate Doyle et al.'s (2014) findings that more people chose to evacuate when given a risk communication featuring a numerical expression as opposed to a VPE. Although we found an effect of format in Study 1, it was in the opposite direction to the original study. A large number of responses to the question of 'why' people made their evacuation decision mentioned themes such as 'better to

Table 1. Evacuation and Correctness Ratings for Studies 1 \& 2

| Study | Measure | Outcome | Communication Format- Mean Rating (SE) |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | V | V-N | N | N-V |
| 1 (Low Probability) | Evacuation Decision | Pre | 3.87 (0.09) | 3.87 (0.09) | 3.93 (0.09) | 3.87 (0.09) |
|  |  | Post | 2.41 (0.13) | 2.65 (0.13) | 3.31 (0.13) | 2.83 (0.12) |
|  | Correctness | Post | 1.61 (0.15) | 2.06 (0.14) | 3.35 (0.14) | 2.51 (0.14) |
| 2 (High Probability) | Evacuation Decision | Pre | 1.99 (0.17) | 2.19 (0.16) | 2.42 (0.16) | 2.11 (0.16) |
|  |  | Post | 2.08 (0.16) | 2.34 (0.15) | 2.47 (0.15) | 2.32 (0.15) |
|  | Correctness | Post | 1.80 (0.13) | 2.08 (0.13) | 2.45 (0.13) | 2.31 (0.13) |

be safe than sorry'. There was little cost to the participant to adopt this approach, which could have been a factor in the high proportion of people choosing to evacuate immediately. Whilst Doyle et al. (2014) attributed their results to the ambiguity of VPEs, we argue that our results could also be explained using this reasoning. Participants may have felt that the choice to use a VPE in the risk communication reflected a level of uncertainty in the outcome, with the communicator 'hedging their bets', and thus felt that it was better to adopt a conservative stance and evacuate, 'just in case'. Indeed, this is in line with the appropriate response of increased uncertainty providing an impetus to be concerned and an even greater reason to act (Lewandowsky et al., 2015). Additionally, if an 'unlikely' event were to occur, it would be far more consequential than if a 'likely' event did not occur.

The lack of an influence of numeracy on nearly all of our measures was somewhat unexpected, given the fact that numeracy has been demonstrated to influence effects of communication format (Gurmankin et al., 2004), and information format (e.g. frequencies versus percentages, Reyna, Nelson, Han, \& Dieckmann, 2010).

Further research should seek to explore the effect of the precision of the communication format. Chess, Hance \& Sandman (1988) claimed being open about levels of uncertainty would lead to enhanced credibility and trustworthiness. The current study only explored point numerical estimates (e.g. '20\% likelihood'), rather than more specific point estimates (e.g. '23\% likelihood'). Including range estimates (both small and large) would allow for a better understanding of the benefits of including numbers in risk communications. Whilst Longman et al.'s (2014) findings suggest that range estimates will have a negative effect on understanding and perceived credibility, others have found that range estimates are perceived as more useful and more honest (Dieckmann, Mauro, \& Slovic, 2010; Johnson \& Slovic, 1995).

## Conclusion

This study provides a different perspective to examining the effectiveness of risk and uncertainty communications, moving away from merely how the information is understood. Trust is fundamental to improving these communications (Slovic, 1993), and our work contributes to this somewhat neglected area of research.

The present research provided a systematic comparison of the effect of differing communication formats on the credibility of the communicator in the context of geological risk communications. Identifying instances in which the communication format has a significant impact on the audience's perceptions of the communicator is key to building and maintaining public trust in science, as well as improving the effectiveness of risk communication. Our findings show that the numerical format is viewed as more correct and is most robust against reductions in credibility following an 'erroneous' prediction. The present results thus suggest numbers should be included in these communications wherever possible.

## Acknowledgements

SJ was supported by a UCL IMPACT studentship, half funded by the British Geological Survey (BGS). RML's contribution is published with permission of the Director, BGS. We are grateful to Dr Charlotte Vye-Brown (BGS) for assistance with creating the vignette used in the studies.

## References

Bonnefon, J. F., \& Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. Psychological Science, 17(9), 747-751.
Budescu, D. V, Broomell, S. B., \& Por, H. H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. Psychological Science, 20(3), 299-308.
Budescu, D. V, Por, H. H., Broomell, S. B., \& Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. Nature Climate Change, 4(6), 508-512.
Budescu, D. V, \& Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. Organizational Behavior and Human Decision Processes, 36(3), 391405.

Budescu, D. V, \& Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. Psychology of Learning and Motivation, 32(2), 275-318.
Chess, C., Hance, B. J., \& Sandman, P. M. (1988). Improving dialogue with communities: a short guide for government risk communication. Division of Science and Research, New Jersey Department of Environmental Protection.

Cokely, E., Galesic, M., Schulz, E., Ghazal, S., \& GarciaRetamero, R. (2012). Measuring risk literacy: the Berlin numeracy test. Judgment and Decision Making, 7(1), 2547.

Dieckmann, N., Mauro, R., \& Slovic, P. (2010). The effects of presenting imprecise probabilities in intelligence forecasts. Risk Analysis, 30(6), 987-1001.
Dieckmann, N., Slovic, P., \& Peters, E. M. (2009). The Use of narrative evidence and explicit likelihood by decisionmakers varying in numeracy. Risk Analysis, 29(10), 1473-1488.
Doyle, E. E. H., McClure, J., Paton, D., \& Johnston, D. M. (2014). Uncertainty and decision making: Volcanic crisis scenarios. International Journal of Disaster Risk Reduction, 10(PA), 75-101.
Erev, I., \& Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. Organizational Behavior and Human Decision Processes, 45(1), 1-18.
Gurmankin, A. D., Baron, J., \& Armstrong, K. (2004). The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication. Medical Decision Making : An International Journal of the Society for Medical Decision Making, 24(3), 265-271.
Harris, A. J. L., \& Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. Journal of Experimental Psychology. Learning, Memory, and Cognition, 37(6), 1571-8.
Harris, A. J. L., Corner, A., Xu, J., \& Du, X. (2013). Lost in translation? Interpretations of the probability phrases used by the Intergovernmental Panel on Climate Change in China and the UK. Climatic Change, 121(2), 415-425.
House of Lords. (2000). Science and Technology - Third Report. Retrieved from http://www.publications.parliament.uk/pa/ld199900/ldse lect/ldsctech/38/3801.htm
Jenkins, S., Harris, A. J. L., \& Lark, R. M. (2016). "Unlikely" Outcomes Might Never Occur, But What About "Unlikely (20 \% Chance)" Outcomes? In \& J. . T. A. Papafragou., D. Grodner., D. Mirman. (Ed.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 390-395). Austin, TX: Cognitive Science Society.
Johnson, B. B., \& Slovic, P. (1995). Presenting uncertainty in health risk assessment: initial studies of its effects on risk perception and trust. Risk Analysis: An Official Publication of the Society for Risk Analysis, 15(4), 485494.

Juanchich, M., \& Sirota, M. (2016). How much will the sea level rise? Outcome selection and subjective probability in climate change predictions. Manuscript in progress.
Kasperson, R. (2014). Four questions for risk communication. Journal of Risk Research, 17(10), 12331239.

Lewandowsky, S., Ballard, T., \& Pancost, R. D. (2015). Uncertainty as knowledge. Philosophical Transactions of
the Royal Society A: Mathematical, Physical and Engineering Sciences, 373(2055).
Lipkus, I. M., Samsa, G., \& Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. Medical Decision Making : An International Journal of the Society for Medical Decision Making, 21(1), 37-44.
Longman, T., Turner, R., King, M., \& McCaffery, K. J. (2012). The effects of communicating uncertainty in quantitative health risk estimates. Patient Education and Counseling, 89(2), 252-259.
Nature. (2010). A Question of Trust. Nature, 466(7302).
Patt, A. G., \& Dessai, S. (2005). Communicating uncertainty: Lessons learned and suggestions for climate change assessment. Comptes Rendus - Geoscience, 337(4), 425441.

Reyna, V. F., Nelson, W. L., Han, P. K., \& Dieckmann, N. (2010). How Numeracy Influences Risk Comprehension and Medical Decision Making. Psychological Bulletin, 135(6), 943-973.
Shanteau, J. (1992). Competence in Experts: The Role of Task Characteristics. Organizational Behavior And Human Decision Processes, 53, 252-266.
Sjöberg, L. (2001). Limits of knowledge and the limited importance of trust. Risk Analysis, 21(1), 189-198.
Slovic, P. (1993). Perceived risk, trust, and democracy. Risk Analysis, 13(6), 675-682.
Teigen, K. H., \& Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. Acta Psychologica, 88(3), 233-258.
Teigen, K. H., \& Brun, W. (1999). The Directionality of Verbal Probability Expressions: Effects on Decisions, Predictions, and Probabilistic Reasoning. Organizational Behavior and Human Decision Processes, 80(2), 155190.

Teigen, K. H., \& Brun, W. (2003). Verbal Probabilities: A Question of Frame? Journal of Behavioral Decision Making, 16(1), 53-72.
Teigen, K. H., Juanchich, M., \& Riege, A. H. (2013). Improbable outcomes: Infrequent or extraordinary? Cognition, 127(1), 119-139.
Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis. Journal of Risk Research, 5(2), 177-186.
Wachinger, G., Renn, O., Begg, C., \& Kuhlicke, C. (2013). The risk perception paradox-implications for governance and communication of natural hazards. Risk Analysis, 33(6), 1049-1065.
Weber, E. U., \& Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. Journal of Experimental Psychology: Human Perception and Performance, 16(4), 781-789.
Witteman, C. L. M., \& Renooij, S. (2003). Evaluation of a verbal-numerical probability scale. International Journal of Approximate Reasoning, 33(2), 117-131.

# A modeling link between cognitive and biological homeostasis 

# Gavin Jenkins (gjenkins@sfu.ca) 

Jordan I. Barnes (jordanb@sfu.ca)<br>Paul Tupper (pft3@sfu.ca)

Mark Blair (mblair@sfu.ca)<br>Simon Fraser University<br>8888 University Dr., Burnaby, BC,<br>Canada, V5A1S6


#### Abstract

The problem of stability has long been a limiting factor in developing neural networks that can grow in size and complexity. Outside of particular, narrow parameter ranges, changes in activity can easily result in total loss of control. Human cognition must have reliable means of acting to stay within the stable ranges of sensitivity and activation. Learning is one such mechanism, and population dynamics are another. Here, we focus on another, often overlooked stability mechanism: cellular homeostasis through metabolism dynamics. We ran a visual change detection experiment designed to strain network stability while minimizing any learnable patterns. We fit the data using models with and without cellular energy levels as a factor, finding that the model influenced by its past history of energy use was a closer fit to the human data.


Keywords: Homeostasis; attention; visual change detection; neural modeling

## Introduction

The stability of learning and development depends partly on the stability of the learner's underlying cognitive machinery. A system that is not tethered to a baseline level of activity is vulnerable to being excited out of control or perishing through complete inaction. A number of mechanisms have been uncovered that promote basic stability in neural systems across multiple timescales. For instance, neurons coupled to one another via patterns of local excitation and lateral inhibition, can interact over fractions of a second to stably form and maintain "peaks" of activation around a core value of interest (e.g. Thelen, Schöner, Scheier, and Smith, 2001). As patterns in the short term persist, memory traces can be formed that are able to project the influence of these patterns over much longer timescales, leading to overall stability across similar situations.

Stability can also be driven at a level even below that of populations of neurons. Cellular metabolic processes allow individual neural units to contribute to the stability of a population coded representation by, observing and acting on their own changes in activity, and doing so at multiple timescales. The Hebbian rule, relating changes in correlations in the activity of neurons to their degree of co-activation, under-specifies the adjustment mechanisms needed for higher order stability. Oja (1982) derives an additional term, that, if included in the instantaneous rate of Hebbian weight change, will remain within some stable range of activity while maintaining the correlation. It was suggested that this term could
be thought of as a form of intrinsic "leakage" rate, $\eta$, of the materials available to the synapse.

This initial modification to the Hebbian rule was largely abstracted, however, from the precise biological interpretation. The cellular mechanism would need to toggle the adaptation of a neural unit between "labile" and "stable" dispositions toward changing connection strengths (Bienenstock, Cooper, \& Munro, 1982). One suggested candidate for this process is brain-derived neurotrophic factor (BDNF) (Glaser \& Joublin, 2011). Using Calcium levels as a proxy for the instantaneous levels of change at a synapse, neural units coding for changes in the level of BDNF can dynamically alter the underlying synaptic excitation/inhibition levels of cells. For instance, blocking input channels to the retina using a tetrodotoxin can affect cellular activity in ways that suggest homeostatic forces (Turrigiano, 2011). Strong excitatory inputs, when blocked, allow for a period of higher than normal activity once unblocked; likewise, the opposite is shown when inhibitory inputs are blocked. More generally, energy stores build up in the blocked neurons when receiving lower than normal input (and thus experiencing lower than normal activity themselves), or when allowed to fire higher than normal, stores are depleted. When a baseline level of energy is restored, normal conditions are eventually achieved again, but the effect in the meantime is one of an internal, cellular homeostatic force.

In the present study, we use a visual change detection paradigm to explore the capacity of the cognitive system to adapt to changes in task demands. Our model of the data are more detailed than that described by Oja (1982), but more abstract than a chemically detailed BDNF explanation. Our goal was to create perturbations of cognitive homeostasis that can produce interesting behavioral level data suitable for modeling effects beyond those observable in cellular recordings. The specifics of two computational models are then introduced for capturing the new behavioral effects. Simulations show that the model with a cellular energy term outperforms a version without one. We conclude with a discussion on the merits of including energy terms in basic neural models as a part of establishing a common language of conservation across the brain, even when exploring cognition at the level of behavior.

## Change Detection

The change detection paradigm is a useful way of studying the cognitive operations necessary for several processes relevant to homeostasis. A typical trial in a change detection task consists of a sample set of items to remember, followed by a test set of items and a response prompt. The paradigm is simple yet challenging to model (Johnson, Spencer, Luck, \& Schöner, 2009); so we opted for a highly simplified version with only one feature dimension of change: color. By changing anywhere from zero to six colors per trial, however, we still allowed for a straightforward manipulation of homeostasis across trials, disturbing the balance of expectation, ability, and adaptation constantly throughout our task.

## Experiment

We designed a change detection experiment with the intention of placing participants in situations of rapidly changing cortical energy levels as might be consistent with a BDNF homeostatic response. The task was also designed as a situation where homeostatic control would potentially be beneficial to task performance. Two presentations of colored squares were given to participants per trial, and the number of changes between presentations was reported, changing from zero to six changes unpredictably trial by trial. No overall learnable trial-to-trial pattern was available to participants that would aid them in answering correctly, so that there was no advantage to adopting complex expectations.

## Method

Participants were recruited from the Simon Fraser University psychology department subject pool, where they received course credit for 30 minutes of participation. We asked participants to think of 6 color patches as alien creatures that would change form over the course of a trial. The job of the participant was to correctly classify the number of changes as part of a national effort to better understand these aliens. Of the 33 subjects, 30 were retained for analysis. Three participants were dropped for failing to complete 106 trials of the task.

Each trial involved a masked presentation sequence designed to eliminate any relevance of spatial position of stimuli, so that color alone was the sole feature dimension of change detection. Subjects were instructed to view a fixation cross for two seconds at the start of each trial. A set of six colors was then presented for four seconds, long enough to ensure an ability to briefly encode the colors in memory. The screen was then masked to block effects of afterimages for two seconds, and a second set of colors was presented. The orientation of the colors was a vertical $2 \times 3$ grid rather than the horizontal $3 \times 2$ grid in the first presentation, to remove any clear or objective correspondence between spatial locations in the first and second presentations. All colors were also scrambled in positions, in addition to the display positions being rotated.

The second color display was left on the screen for either $2.5,3.0,3.5$, or 4.0 seconds (counterbalanced across trials
per participant) to allow for different amounts of time for any homeostatic system to adjust cellular activity rates now that the changes, if any, were visible. The intention of the time manipulation was that there would not be sufficient time to make a homeostatic adjustment in one timeframe, but enough in another, and thus that the required time for homeostatic adjustment could be identified by changes in accuracy.

A second mask was then presented for two seconds, whereafter the participant was given any amount of time to report the number of changes on a number line on screen. Participants were given explicit feedback about their answer in the form of a blue mark for the correct answer in addition to the red mark indicating the chosen response. Figure 1 depicts a graphical representation of the task procedure.

Twelve perceptually equidistant colors were assigned to each participant, offset from each other by a random amount. Color coordinates were obtained from a slice of CIEL*a*b* color space, with luminance set to a constant $L=75$. Only coordinates in this color space that have a representation in RGB can be displayed on a monitor, so all but the largest circular portion of the color space satisfying this constraint was removed (Johnson et al., 2009; Lehky \& Sejnowski, 1999). These twelve colors were then used consistently for a participant's entire experiment. Every trial randomly sampled from the participant's set of twelve colors as needed. A "zero changes" trial, for example, only required six colors (the same set of six twice), whereas a "six changes" trial required all twelve colors (two different sets of six).

Trials were counterbalanced with a customized Latin square in order to equally distribute the number of times each possible number of changes in colors (0-6) was the correct answer, while also equally distributing the number of oneback differences between correct answers on the current and previous trial. Figure 2 shows the actual number of total trials across participants of each combination of these variables. The four possible display times for the second color array were evenly distributed within each of these trial types (i.e., within each bar in Figure 2. Display times not shown in Figure).

The partial Latin square was necessary due to mathematical constraints in designing a distribution of trials that attempted to fit both criteria. Each trial is part of a one-back link to its previous trial but also to the next trial, so any change has a cascade of consequences for the options on other trials. Also, certain one-back differences are impossible; for example, if the number of changed colors on a trial is three, then the one-back difference cannot possibly be 6 , because that would mean the correct answer on the previous trial was "3 changes" which is not possible. There were 9 impossible combinations like this overall, forming a triangle of missing bars in Figure 2 (upper right portion of figure).

Trial orders were generated by a Monte Carlo algorithm that simulated many solutions to the overall trial order problem. The algorithm respected the constraints described above, while also introducing randomness in order to limit


Figure 1: The presentation sequence of a trial. A fixation cross appears for 2 seconds. A sample set of 6 colors to be remembered is then presented for 4 seconds. A diffuse colored mask intended to cancel out sensory correlation with the subsequent test array is then presented for 2 seconds followed by a set of 6 color patches rotated by $90^{\circ}$ (and completely scrambled with no correlations in positions before vs. after masking). A final colour mask is then presented for 1 second in advance of the unconstrained response phase. Subjects would click on a value for their estimate of the number of changes, marked in red, and the correct value marked by blue.
trial order confounds between participants. A unique solution was found for each participant. Solutions were defined as trial orders where the total number of trials was nearly equal for each of the seven possible correct answers (so as many trials have an answer of " 4 " as have an answer of " 2 "), and where the number of trials was also nearly equal for each of the seven possible one-back changes in correct answer. The shape depicted in Figure 2 was the algorithm's consistent solution to this problem, with only very minor differences and asymmetries between participants.

## Experiment Results

The average error across all trials and all subjects was 1.5 units (number of color changes, out of 6). A mixed effects model with subject as a grouping factor showed no significant improvement in change detection over the course of the experiment $(t=1.6, p=0.11)$. In accordance with our goal of exploring sensitivity to swings in cognitive energy and activation over time, we also checked for a lag-one correlation in responding over the course of the experiment. Lag-one in this case is being measured as the correlation between responses on trial t and responding on trial t -1, i.e. the correlation between the list of responses and itself shifted one trial sooner. When this correlation is positive, it suggests that answers were given in long "runs" where a high answer would be followed by more high answers and similarly for low. When the lag-one correlation is negative, it suggests a degree of "ping-ponging" back and forth between high and low (number of changes) successive answers more often than would be expected by chance. A lag-one of zero suggests no particular persistent carryover effects of responding from one trial to the next.

For this analysis, it was necessary to control for any lagone correlation that may have been inherent to the trial order itself. Figure 2 shows how, despite equal distribution of changes and differences in changes in colors across trials, patterns exist between these two variables. To control for such patterns during the analysis, we looked for lag-one correlation only in those trials we knew had a symmetric pattern of changes compared to previous trials: ones where the answer was exactly three changes ("Correct Answer $=3$ " set of
bars in Figure 2). Within this restricted data set, we found a positive correlation ( $\beta=0.11, t=2.74, p<0.01$ ) in the responses between trials, indicating a minor preference for repeated "runs" of responding that was not related to any experimental design.

Timing differences in the second color presentation phase of the task did not correspond to significant differences in performance. Differences were expected, but any homeostatic effects may simply be too rapid (or too slow / occurring by memory only during the answering phase) to be distinguished by the difference between 2.5 and 4 seconds of presentation.

## Model

We tested two models against our behavioral data: one biologically inspired model (in line with with the BDNF principles discussed by Glaser and Joublin, 2011) capturing cellular homeostatic principles, and the same model but with the cellular homeostatic term removed. Each was simulated in Matlab using the exact order and content of trials and color values seen by each of the 30 participants in turn. These models provide continuous real number outputs between 0 to 6 , but were forced to choose exact whole number answers for number of changes, as the humans were. Especially for the model that accounted for the C.H. model, the effect of feedback could have contributed differently to the next trial's behavior if whole number answers were not required, so to ensure the most human-like between-trial patterns, model answers were also rounded.

Since our goal was to capture the levels of errors and inaccuracy in human behavior, and the difficulties of the task being between-trial consistency, we fed the signal for the number of changes on a trial to the models directly, so that the target measure of the fit was focused on the specific pattern of errors made by each human on each of their trials. Each model was fit to human responses using the method described below, which tested enough detail to capture these homeostasis straining effects over time, as well as patterns of variance in errors.


Figure 2: Histogram of different trial types in the experiment. Every trial had a correct answer, and all but the first trial had a one-back difference between its correct answer and the correct answer on the previous trial. Some of these combinations were impossible (see text). The shape seen here equally distributes correct answers overall, and also equally distributes changes between trials overall, while avoiding the impossible trial types. Bars are not perfectly symmetrical due to partial randomization between participants to avoid trial order effects. Some small asymmetries between bars are visible due to the necessity of using a Monte Carlo algorithm for this task.

## Cellular Homeostasis Model

The primary model of interest displayed homeostasis as a result of cellular energy resisting extreme response rates by becoming depleted after heavy use or energized after low use. The neuron's energy reserves were its only way of tracking information across trials. Its output on a trial is given by:

$$
O_{t}=a s_{t} E_{t}+b+\varepsilon n
$$

where $s_{t}$ is the stimulus on the trial (the number of changed colors), $\varepsilon$ is normally distributed noise, $a, b$, and $n$ are freely fitted coefficients, and $E$ is cellular energy, calculated per trial as:
$E_{t}=E_{t-1}+\frac{\left(O_{t-1}-3\right) c}{\tau}-\frac{\left(E_{t-1}-1\right) \frac{c}{3}}{\tau}$
The fourth free parameter of the model is $c$ in the energy equation above. $\tau$ was not parameterized and was always set to a constant value of 10 . The $O_{t-1}-3$ represents the fact that three was the most central response out of options 0-6, so any values below this were considered "low" answers that helped relatively gain cellular energy, and any values above this were "high" answers that relatively depleted energy reserves. The third term represents relaxation of the system to neutral energy over time, but $1 / 3$ slower than the rate of energy change from responding (Toyoizumi, Kaneko, Stryker, \& Miller, 2014). Setting $E_{t-1}-1$ causes energy to relax toward a neutral value of 1 , where it would have no effect on the cell's output. All model answers were rounded to the nearest integer from 0-6 - the only valid answers in this task - on a per-trial basis. This was also true of the mathematical model
variant below.
The energy term mimics the predictions of the BDNF chemical cell model, with as much abstraction as necessary to allow for easy application to typical cognitive behavioral data. As in the BDNF model, a decision neuron firing at a relatively low rate ( $0-2$ changes in this task) builds up energy and fires faster than normal for a brief period in response, while a neuron firing relatively energetically (4-6 changes in this task) will progressively deplete energy and fire more slowly for a time in response.

## Simple Mathematical Model

The second model had no biological motivation, but serves as a baseline comparison with the C.H. model. It replicates the C.H. model in structure, but with the energy term removed. This model had no means by which to account for previous trials, as the model had no form of memory/hysteresis. It did, however, still have all of the information needed to perform perfectly at the task according to task instructions (rather than fit to human answers). A perfect score could be achieved with parameter values $a=1, b=0$, and $n=0$. Thus, lower performance by this model at fitting human data would be due to a lesser ability to capture human sources of error and trial to trial effects (theoretically irrelevant and distracting from optimal performance) only. The model is given by:

$$
O_{t}=a s_{t}+b+\varepsilon n
$$

## Fitting Method

We created three dimensional histograms of responses for humans and models to fit and compare results. Every trial across subjects was sorted into histogram bins according to the correct answer on that trial ( 0 to 6 ), the difference between the correct answer on that and the correct answer on the previous trial ( -6 to 6 ), and the participant's (real or simulated) response ( 0 to 6 ). This produced $7 \times 13 \times 7$ set of possibilities ( 637 histogram bins). The first two dimensions represent objective trial types (ones built into the design and based on actual stimuli) and since these were not perfectly evenly distributed due to mathematical constraints (see Figure 2), we weighted the importance of cells from the combinations that had more more data points, using weighted least squares:

$$
\text { Fit }=\sqrt{\frac{1}{637} \sum_{s, d, r}\left(\sum_{s, d}\left(N_{H_{s, d}}\right)\left(H_{s, d, r, \text { model }}-H_{s, d, r, \text { human }}\right)^{2}\right)}
$$

where $H$ is the histogram, and $s, d, r$ are stimuli at time of response ( t ); difference in $\operatorname{stimuli}(\mathrm{t})$ - $\operatorname{stimuli}(\mathrm{t}-1)$; and response, respectively. This method captures information about trial to trial effects, main effects, interactions, general task accuracy, and patterns of variance, all in one measure.

This histogram method was chosen for the objective function to avoid the problem that fitting averaged descriptives like accuracy or standard deviation of responses, which could lead to degenerate model patterns: a mean might be fitted by $100 \%$ of model answers at exactly the mean without realistic variance, for example. Fitting the entire histogram of all relevant measures allowed for the model with the richest detailed pattern of fits across every measure.


Figure 3: Change responses versus change responses on the previous trial, where the current trial had 3 changes. C.H. is the Cellular Homeostatic Model.

The models have differing numbers of free parameters (4 vs. 3), yet due to the dynamical nature of the cellular energy model, its maximum likelihood cannot be easily calculated, and simulations take non-trivial time to perform. Ultimately, the main concern of an overly complex model is failure to generalize, so instead of scoring parsimony, we ruled out over-fitting directly using cross-validation. For each model, we split the subject pool in half, and separately fit each half. We then recorded the fits for each half using only the best parameters found from fitting the other half. All results reported below are exclusively these generalization results, removing our concerns about hidden differences in generalization ability between the three and four parameter models.

We fit a 7 value range for each parameter in a grid covering all of reasonable parameter space for the task, separately for each half of participants as above. We then focused more closely near each best fit at higher precision until precision increases stopped yielding better fits.

## Model fitting results

The best fitting parameters for the cellular homeostasis model were $a=-0.4, b=3.5, c=0.92$, and $n=1.4$. The best fitting parameters for the simple mathematical model were $a=-0.45, b=3.5$, and $n=1.5$.

The average cross-validation weighted least squares error for the cellular homeostasis model was 8.335 , while the average cross-validation weighted least squares error for the simple mathematical model was 9.436 . Since these values already account for the greater potential for over-fitting with four versus three parameters, they can be compared at face value: the energy term meaningfully accounts for human behaviors above and beyond slope, intercept, and noise terms.

Although the magnitude of the effect is somewhat modest,
it is noteworthy to point out again that the simple mathematical model was able to achieve $100 \%$ objective accuracy at the task as per the task instructions by simply fitting parameters $a=1, b=0$, and $n=0$, a combination that was within the tested range of reasonable parameters during fitting. Thus, the lower performance of the simpler model is purely a result of more poorly fitted patterns of human error, possibly error in response to patterns of trials that threw off homeostatic neutrality, since the better fitting energy term in the homeostasis model varies by activity on previous trials.

Ultimately, the exact cause of the better fits of the cellular homeostasis model are unclear. Analysis of the full threedimensional fitting histogram suggested noticeable differences between the two models and between models and human data, but these differences were too diffuse and opaque to easily interpret.

Lag-one correlations on trials with three color changes also fit human data better in the homeostasis model than in the simple mathematical model. Where $\beta_{\text {human }}=0.11$, $\beta_{\text {cellularmodel }}=0.05$ and $\beta_{\text {simplemodel }}=0$ (see Figure 3). These correlations highlight the lag-one effects in particular, but lag-one effects are also built into the 3-dimensional histograms used for the main fitting results.

## Discussion

Behavioral stability is often approached from the perspectives of neural population dynamics or higher-level verbal or executive control theories. Stability is also attainable, however, through more microscopic means, at an intra-cellular or synaptic level. This source of homeostasis in cognition and behavior is, by itself, simple. Activity is most likely stabilized around a static resting level, at least within the timescales afforded by a particular task. This does not necessarily match the flexibility or possible sophistication of higher level stabilizing mechanisms.

Cellular homeostasis is, however, an appreciable effect, especially when studied in a task that eliminates distracting forces and pushes the boundaries of a system's homeostasis. Even in less specialized situations, however, cellular effects are likely continuing to function and can contribute toward an understanding of behavior. This form of homeostasis may generally be playing a silent and under-appreciated role in a wide variety of cognitive activities, providing a small but important level of baseline stability that can act as a foundation for more targeted systems like learning mechanisms to explore more freely without risk of losing control of a system.

Our findings require significant further investigation to establish an exact pattern of behavior that is being captured by the cellular energy term of our model, and followup experiments are necessary to confirm those mechanisms once identified. In the meantime, we suggest cognitive modelers more often consider including cellular energy terms in neural models of not only cellular-level effects, but behavioral effects as well. All cognitive processes involve neurons, so even modest effects of such a cellular system may be of great impor-
tance collectively, for a range of effects at different levels of complexity and abstraction.

## Acknowledgments

We would like to acknowledge the helpful contributions made by all members of the Cognitive Science Laboratory at Simon Fraser University. Gavin Jenkins was supported by an NSERC Discovery Accelerator Supplement. Paul Tupper was supported by an NSERC Discovery Grant and a Tier 2 Canada Research Chair.

## References

Bienenstock, E. L., Cooper, L. N., \& Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience, 2(1), 32-48. doi:10.1371/journal.ppat. 0020109
Glaser, C. \& Joublin, F. (2011). Firing Rate Homeostasis for Dynamic Neural Field Formation. IEEE Transactions on Autonomous Mental Development, 3(4), 285-299. doi:10.1109/TAMD.2011.2138705
Johnson, J. S., Spencer, J. P., Luck, S. J., \& Schöner, G. (2009). A Dynamic Neural Field Model of Visual Working Memory and Change Detection. Psychological Science, 20(5), 568-577. doi:10.1111/j.14679280.2009.02329.x

Lehky, S. R. \& Sejnowski, T. J. (1999). Seeing white: Qualia in the context of decoding population codes. Neural computation, 11(6), 1261-1280. doi:10.1162/ 089976699300016232
Oja, E. (1982). A Simplified Neuron Model as a Principal Component Analyzer. Journal of Mathematical Biology, 1, 267-273.
Thelen, E., Schöner, G., Scheier, C., \& Smith, L. B. (2001). The dynamics of embodiment: a field theory of infant perseverative reaching. The Behavioral and brain sciences, 24(1), 1-34, discussion 34-86. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11515285
Toyoizumi, T., Kaneko, M., Stryker, M. P., \& Miller, K. D. (2014). Modeling the dynamical interaction of Hebbian and homeostatic plasticity. Neuron, 84(2), 1-40. doi:10.1016/j.neuron.2014.09.036
Turrigiano, G. (2011). Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. Annual review of neuroscience, 34, 89-103. doi:10.1146/annurev-neuro-060909-153238

# Viewers' Sensitivity to Abstract Event Structure 

Yue Ji (Jiyue@Udel.Edu)<br>Department of Linguistics \& Cognitive Science, University of Delaware, 125 E. Main Street<br>Newark, DE 19716 USA

Anna Papafragou (Apapafragou@Udel.Edu)<br>Department of Psychological \& Brain Sciences, University of Delaware, 105 The Green<br>Newark, DE 19716 USA


#### Abstract

Bounded and unbounded events differ in whether they include an inherent endpoint (Bach, 1986). Even though this distinction can be important for the way events are identified and processed, the literature on event cognition has not focused on such abstract aspects of event structure. In the present study, we asked whether viewers are sensitive to the distinction between bounded and unbounded events in a category learning task. Our results show that people were more successful in forming the category of bounded events than that of unbounded events. We discuss implications of this finding for event cognition.


Keywords: event structure; endpoint; boundedness

## Introduction

Our experience of the world is intrinsically dynamic. To make sense of the complex flow of changes in our environment, we break continuous streams of experience into separate entities and classify such entities into different types.

Much work has focused on how people segment continuous experience into discrete units, i.e. events. The term "event" refers broadly to a temporal segment that has "a beginning and an ending" (Zacks \& Tversky, 2001). People identify the boundaries of an event through tracking changes in perceptual features such as direction, location, or speed of action (e.g. an arrow hitting a target); more importantly, people encode events based on conceptual features, especially the goal-directedness or causal structure of the corresponding experience (e.g. a person on diet hitting a target; Zacks \& Swallow, 2007). Event boundaries have a privileged status in memory and provide anchors for later learning and describing (Swallow, Zacks, \& Abrams, 2009). In particular, the endpoint is conceptualized as a critical event component. For instance, when comparing two events, the resultant state (e.g. whether a ball knocked over the whole tower or just a few blocks) has more psychological weight than other perceptual features (e.g. the moving direction of the ball) (He \& Arunachalam, 2016). In the well-studied domain of motion events, the goal of motion is more accurately encoded in both language and memory as opposed to other components such as the source (Lakusta \& Landau, 2005, 2012; Papafragou, 2010; Regier \& Zheng, 2007; Wagner, 2009). In addition, people tend to fill the gap between successive events within a causal chain by generating rapid inferences about the endpoint of the first
event. In a study by Strickland and Keil (2011), after watching videos of someone launching an object (e.g. kicking a soccer ball) followed by the object's directed motion (e.g. the ball flying into the goal), participants mistakenly reported that they saw the moment of contact, i.e. the endpoint of the launching event, even when it was actually omitted from the display.

Despite the richness of the literature on event segmentation and the salience of endpoints in event perception, the nature of event endpoints has been less discussed. In most event-segmentation studies, the stimuli are actions by an intentional actor (e.g. a person putting up a tent) and the endpoint is taken to be obvious and welldefined (e.g. the moment the tent is put up). In studies of motion events, the endpoint appears similarly self-evident (and is typically the moment that a moving entity reaches the goal). However, across a broad range of events, the notion of endpoint is not always straightforward. Consider the following situations described by the two sentences in (1):
(1) a. The child played the Moonlight Sonata.
b. The child played the piano.

There is subtle difference between (1a) and (1b). The event in (1a) comes to an end when the last note of the sonata was played. In contrast, it is hard to specify how or when the situation in (1b) ends - the child could stop playing at any point. The endpoint is inherent in the former event but is arbitrary in the latter. Such contrasts have been discussed extensively in the linguistic literature on aspect (i.e., the linguistic encoding of the internal temporal profile of events). In this literature, the distinction between the two sentences in (1) is captured by assuming that (1a) encodes an experience as a "bounded" event but (1b) encodes it as an "unbounded" event (Bach, 1986; Harley, 2003; Jackendoff, 1991). Bounded events have an internal structure with a "built-in terminal point" (Comrie, 1976), "climax" (Vendler, 1957) or "culmination" (Parsons, 1990), while unbounded events are homogenous, lacking internal development (Krifka, 1998). This linguistic distinction presumably has a non-linguistic counterpart in the way events are perceived and understood but to date, this connection has not been explored in detail.

Inspired by the rich linguistic research on how event endpoints are encoded in language (see Filip, 2004; Krifka, 1998 , etc.), one can further identify two major types of consideration that determine whether an event is bounded or
not. First, intuitions about boundedness may be due to the nature of the action. In particular, some actions lead to a change of state in the affected object, such that the endpoint is the resultant state (bounded events); other actions do not affect the object in a perceptible way or the change lacks a well-defined resultant state (unbounded events). The contrast is shown in the following example:
(2) a. The child dressed the teddy bear.
b. The child patted the teddy bear.
(2a) describes a bounded event - the teddy bear was dressed when the child finished. (2b) describes an unbounded event-no predictable result followed from the child's patting. Although both events involve the same object, the difference in actions leads to the contrast in boundedness.

Second, intuitions about boundedness may be due to the nature of the affected object. Sometimes, there is a homomorphism between the affected object and the time course of the event (Dowty, 1991; Krifka, 1989), such that the changes in the object track or "measure out" the way the event develops (Tenny, 1987). When the object itself is quantified, the event is bounded. The contrast can be illustrated by the example below:
(3) a. The child ate a pretzel.
b. The child ate cheerios.
(3a) depicts a bounded event- the event unfolds as the pretzel changes and it ends at the moment when the pretzel is gone. (3b) depicts an unbounded event that lacks an inherent endpoint-the child could stop at any time.

To sum up, bounded and unbounded events differ in whether they have an inherent endpoint. Two major components, i.e. the nature of the action and the affected object, might determine whether an inherent endpoint is available. So far the literature on event perception has not explored the role of boundedness in determining event boundaries, and little is known about whether viewers are sensitive to such abstract aspects of event cognition. One suggestive piece of evidence comes from work focusing on how events are counted. Bounded events are naturally counted in terms of how many inherent endpoints have been achieved. Lacking an inherent endpoint, unbounded events are counted according to spatio-temporal criteria. Returning to the example in (1), imagine that the child paused for a break and then resumed her playing in both situations. The event of playing the Moonlight Sonata still occurred once, but the child played the piano twice. When counting events like (1a), adults look for the inherent endpoint regardless of the pauses (Barner, Wagner \& Snedeker, 2008) but 3-to-5-year-olds tend to over-generalize spatio-temporal criteria and count the number of pauses (Wagner, 2006; Wagner \& Carey, 2003).

In the present paper, we explore viewers' sensitivity to the distinction between bounded and unbounded events (defined in terms of the availability of an inherent endpoint, as in (2) and (3) above). Specifically, we ask whether viewers can group events into the bounded vs. unbounded category in a category learning task. Drawing on the linguistic literature in which the category of unbounded
events is definitionally dependent on the category of bounded events (such that boundedness and unboundedness form a positive-negative pair), we ask whether there is an asymmetric relation between the two types of event in nonlinguistic cognition. If so, the category of bounded events might be learned by observers more easily compared to that of unbounded events.

## Experiment 1

Experiment 1 was a category learning task. Participants were exposed to minimal pairs of bounded and unbounded events (defined by the availability of an inherent endpoint) and had to extract what was shared by different events of the same category and extend this information to new events.

## Method

Participants Forty adults participated in the experiment. All were undergraduates at the University of Delaware and received course credit for participation. Data from an additional group of 2 adults were collected but excluded because these adults were color-blind and failed to identify an important test feature (a red frame) consistently.

Stimuli Twenty pairs of videos were created, such that each pair showed a bounded and an unbounded event (see Table 1). Within each pair, the videos had the same duration (range: $4.5 \mathrm{~s}-13 \mathrm{~s} ; \mathrm{M}=7.98 \mathrm{~s}$ ) and involved the same actor but differed minimally from each other in one of two ways that involved boundedness. For half of the pairs, the bounded and unbounded events within a pair involved the same object but differed in terms of the nature of the action performed on the object: the bounded event displayed an action that caused a clear and temporally demarcated change of state in the object (e.g. fold up a handkerchief) while its unbounded counterpart did not involve such a change (e.g. wave a handkerchief). For the other half of the pairs, the bounded and unbounded events within a pair involved the same action but differed in terms of the nature of the affected object: the bounded event involved a single object (e.g. draw a circle) but its unbounded counterpart involved either an unspecified plurality of objects or a mass quantity (e.g. draw circles).

To ensure that all video stimuli would illustrate the contrast in boundedness presented in Table 1, a new group of 18 adults from the same population was asked to watch a subset of the clips and describe what happened in a full English sentence. For this norming task, the events in Table 1 were split into 2 lists, such that each list included only one member of each pair and an equal number of bounded and unbounded events. Each of the 18 participants was randomly assigned to one of the two lists. Their descriptions were coded for the verb used to describe the action and the noun phrase used to describe the affected object(s). As expected, differences in boundedness within a pair that were due to the nature of the action were reflected in verb choices: bounded stimuli elicited verbs of change of state (e.g. "dress a teddy bear") $98.3 \%$ of the time and unbounded stimuli
elicited verbs denoting activities (e.g. "pat a teddy bear") $93.1 \%$ of the time. Similarly, differences in boundedness within a pair that were due to the nature of the object were reflected in noun phrase choices: bounded events elicited count nouns with definite or indefinite articles (e.g. "eat the/a pretzel") $98 \%$ of the time and unbounded events elicited bare plurals, mass nouns, or related devices (e.g. "eat cheerios") $92.4 \%$ of the time.

For purposes of Experiment 1, the video stimuli were arranged into three basic lists corresponding to the three phases of the experiment (see Table 1). For the initial learning phase, we selected 8 pairs of events ( 4 in which boundedness was due to the Action and 4 in which boundedness was due to the Affected Object) and arranged them into a pseudorandomized presentation list in which a single video was played in the center of the screen and the two videos within a pair appeared in immediate succession (the order of bounded-unbounded events within pairs was counterbalanced within the list).

For the later testing phase, we arranged 8 of the remaining pairs of videos (see Table 1) into 2 lists. Each list contained one video from each pair. We counterbalanced whether the event was bounded or unbounded and whether source of boundedness was the action or the object across lists.

For the final (short) surprise testing phase, we used the last 4 pairs of videos, arranged into 2 lists. The same counterbalancing was used as in the (main) testing phase.

Procedure Participants were randomly assigned to one of two conditions. In the Bounded condition, the videos of bounded events shown in the learning phase were given a red frame while their unbounded counterparts were given a black frame. In the Unbounded condition, the reverse assignment occurred.

In the learning phase for both conditions, participants were asked to watch a few videos and to pay attention to those appearing within a red frame. Their task was to figure out what kind of videos were given the red frame and to decide whether a new video could have the red frame or not.

In the testing phase, participants saw a new set of videos and for each one they were asked: "Could the video have a red frame or not?" (test question) In the surprise testing phase, participants were unexpectedly asked: "Could the video have a black frame or not?" (surprise question) This question was included to probe whether participants formed any hypothesis about the secondary event category present within the experiment, even though it was not the target of the study.

After the end of the session, participants were asked to write down what kind of videos could have a red frame. This was used as an additional source of information about the category that participants had just formed.

Table 1: Videos used in Experiment 1.

| Phase | Boundedness Source | No. | Bounded Events | Unbounded Events |
| :--- | :---: | :--- | :--- | :--- |
|  |  | 1 | fold up a handkerchief | wave a handkerchief |
| Learning | Nature of Action | 2 | put up one's hair | scratch one's hair |
|  |  | 3 | pile up a deck of cards | shuffle a deck of cards |
|  |  | 4 | group pawns based on color | mix pawns of two colors |
|  |  | Nature of Affected | 6 | draw a balloon |
|  | Object | 7 | eat a pretzel | draw circles |
|  |  | 8 | flip a postcard | tie knots |
|  |  | 9 | dress a teddy bear | eat cereal |
|  |  | 10 | roll up a towel | flip pages |
| Testing |  | 11 | fill a glass with milk | pat a teddy bear |
|  |  | 12 | scoop up yogurt | twist a towel |
|  |  | 13 | peel a banana | shake a bottle of milk |
|  |  | 14 | blow a balloon | stir yogurt |

## Results

An ANOVA was performed on the proportion of correct responses to all questions with Source of Boundedness (i.e. Nature of Action vs. Nature of the Affected Object) as a within-subjects factor. No significant difference was found $(F(1,39)=.042, p=.838)$. Therefore, answers to questions targeting the two sources of boundedness were collapsed for further analysis.

Results from Experiment 1 are shown in Figure 1. The proportion of correct responses to test questions was significantly higher in the Bounded $(\mathrm{M}=92.50 \%)$ than in the Unbounded condition ( $\mathrm{M}=76.25 \%$ ) $(t(38)=3.563$, $p$ $=.001)$. No significant difference in the proportion of correct responses to the surprise questions in the two conditions was found $(t(38)=-.831, p=.411)$.

An ANOVA was conducted on the proportion of correct responses with Question Type (Test vs. Surprise) as a within-subjects factor and Condition (Bounded vs. Unbounded) as a between-subjects factor. There was a significant effect of Question Type $(F(1,38)=19.795, p<$ 0.0001 ), no significant effect of Condition $(F(1,38)=$ 2.247, $p=.142$ ), and an interaction between the two factors $(F(1,38)=7.833, p=.008)$. The participants were more accurate in test questions than in the surprise questions in the Bounded condition $(t(19)=6.114, p<.00001)$ but not in the Unbounded condition $(t(19)=1.022, p=.320)$.


Figure 1: Proportion of correct responses in Experiment 1. Error bars represent standard error.

Answers to the last open question asking about the target category focused on 3 aspects of the stimuli-organization, neatness and intention. Organization was the most frequent hypothesis ( 29 out of the 40 answers). Specifically, modifiers such as "organized", or "structured" were used to describe bounded events while "unorganized", or "lacking structure" were given for unbounded events. Neatness was the second most frequent hypothesis ( 15 out of the 40 answers). Words used for bounded events included "neat", "tidy" and "clean" while those for unbounded events included "messy" and "untidy". Lastly, intention was mentioned in 9 out of the 40 answers. Bounded events were depicted as aiming "to achieve a goal", or being "on
purpose" while unbounded events were "lacking an end or purpose", "random".

## Discussion

Performance in test questions directly showed that, given the same contrastive examples in the learning phase, the participants were better at forming the category of bounded events compared to that of unbounded events. Furthermore, in the Bounded condition, learning was focused, with participants being less successful in the surprise compared to the test questions; however, no such asymmetry was found in the Unbounded condition. Further intuitions about boundedness were found in answers to the last open question about the nature of the target (red-frame) stimuli. The most frequent hypotheses referred to the organization of the stimuli. This suggests that participants attended to the internal structure of events when forming hypotheses about the meaning of the to-be-acquired category.

## Experiment 2

In Experiment 1, participants might have benefited from the presentation of paired videos in the learning phase. By showing 2 successive videos with minimal differences, the contrast between bounded and unbounded events was highlighted. Experiment 2 asked whether the category of bounded or unbounded events could be efficiently extracted in a less supportive learning context.

## Method

Participants A new group of forty undergraduates at the University of Delaware were recruited. Data from an additional adult were collected but excluded because he failed to understand the task and did not finish all the questions.

Stimuli and Procedure The stimuli and procedure were identical to those in Experiment 1 with one exception. In the learning phase, the sequence of the 16 videos was pseudorandomized such that any 2 videos within a pair were separated by at least 5 other videos. This made it impossible to detect the contrast between bounded and unbounded events by simply comparing 2 consecutive videos.

## Results

As in Experiment 1, no difference in the proportion of correct responses was found between the two sources of boundedness $(F(1,39)=1.595, p=.214)$. The answers were thus collapsed in the following analysis.

Results from Experiment 2 are shown in Figure 2. Test questions elicited a significantly higher proportion of correct responses in the Bounded ( $\mathrm{M}=84.38 \%$ ) than in the Unbounded condition ( $\mathrm{M}=68.13 \%$ ) $(t(38)=3.365, p$ $=.002$ ). There was no significant difference between the two conditions in the proportion of correct responses to the surprise questions $(t(38)=-1.129, p=.266)$.

An ANOVA conducted with Question Type as a withinsubjects factor and Condition as a between-subjects factor showed a significant effect of Question Type $(F(1,38)=$ $7.095, p=.011$ ), no significant effect of Condition $(F(1,38)$ $=.646, p=.427$ ), and an interaction between Question Type and Condition $(F(1,38)=7.839, p=.008)$. The participants performed better in test questions than in surprise questions in the Bounded condition $(t(19)=4.174, p=.001)$, but not in the Unbounded condition $(t(19)=-.093, p=.927)$.


Figure 2: Proportion of correct responses in Experiment 2. Error Bars represent standard error.

Answers to the open question about the nature of the target category still mainly referred to organization, neatness and intention. These were mentioned in 16,8 and 7 out of the 40 answers respectively. In addition, repetition was used to describe unbounded events in 5 answers. Completion appeared in 3 answers about bounded events.

As is clear from Figures 1-2, performance on the test questions was better in Experiment 1 than in Experiment 2. This was confirmed in an ANOVA that used the proportion of correct responses on the test questions as the dependent measure, and included Condition (Bounded vs. Unbounded) and Experiment ( 1 vs. 2) as between-subjects factors. The analysis showed main effects of Condition, $(F(1,76)=$ 23.940, $p<.0001$ ), and Experiment $(F(1,76)=5.986, p$ $=.017)$, and no interaction between the two factors $(F(1,76)$ $=.000, p=1.000$ ). (Results were similar when accuracy on both test and surprise questions was used as the dependent measure.)

## Discussion

Results from Experiment 2 showed a learning advantage for the category of bounded compared to unbounded events. This pattern was similar to Experiment 1, even though performance in Experiment 2 was worse compared to the earlier study, presumably because of the lack of direct contrast between bounded and unbounded events during the learning phase.

## General Discussion

Our findings provide direct evidence for viewers' sensitivity to the abstract feature of boundedness in event cognition. In that sense, the present data go beyond prior work on how bounded and unbounded events are individuated and counted (Barner, Wagner \& Snedeker, 2008; Wagner, 2006; Wagner \& Carey, 2003). Furthermore, our results demonstrate that there is an asymmetry between bounded and unbounded events, such that it is easier to form the category of bounded compared to unbounded events. Our results raise the possibility that unboundedness is asymmetrically dependent on boundedness during event perception and apprehension, and that bounded - but not unbounded - events form a natural class.

The present data leave several directions open for further research. An important direction concerns the exact nature of the conjectures underlying participants' groupings of events into boundedness categories. The notion of boundedness is broad and can be subject to more abstract considerations than the present discussion has suggested. For instance, the inherent endpoint that defines bounded events can provided by a salient intention (Depraetere, 2007). To take an isolated example, even though the action of warming a soup does not have a clearly defined endpoint, it is often construed as culminating at the point at which the soup has reached someone's favorite temperature. In our study, it seems unlikely that intentionality was the feature responsible for participants' success in the Bounded condition. We asked a new group of 10 people to rate the degree of intentionality for all the videos used in the experiments on a scale from 1 (totally unintentional) to 7 (intentional). There was no significant difference between scores for bounded events $(M=5.829)$ and unbounded events $(M=5.704)(t(9)=1.059, p=.330)$.

Finally, a number of researchers has drawn close parallels between object and event systems from a semantic perspective, such that the property of boundedness in the domain of events has been linked to the issue of quantification in the domain of objects (Bach, 1986; Jackendoff, 1991). In our study, the quantification of the affected object served as a cue for distinguishing bounded events from unbounded ones. It is possible that viewers are better at forming the category of bounded events because it is easier to track a single object compared with an unindividuated substance or objects of a variable number. An interesting further question is whether there is a common notion of boundedness underlying cognitive representations of both events and objects (see Wellwood, Hespos \& Rips, in press) and how distinctions in one domain might generalize to the other.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under grant no. 1632849.

## References

Bach, E. (1986). The algebra of events. Linguistics and Philosophy, 9, 5-16.
Barner, D., Wagner, L. \& Snedeker, J. (2008). Events and the ontology of individuals: Verbs as a source of individuating mass and count nouns. Cognition, 106, 805832.

Comrie, B. (1976). Aspect: An Introduction to the Study of Verb Aspect and Related Problems. Cambridge: CUP.
Dowty, D. R. (1991). Thematic proto-roles and argument selection. Language, 67, 547-619.
Depraetere, I. (2007). (A)telicity and intentionality. Linguistics, 45, 243-269.
Filip, H. (2004). The telicity parameter revisited. SALT, 14, 92-109.
Harley, H. (2003). How do verbs get their names? Denominal verbs, manner incorporation and the ontology of verb roots in English. In N. Erteschik-Shir, \& T. Rapoport (Eds.), The Syntax of Aspect. Oxford: OUP.
He, X., \& Arunachalam, S. (2016). How event endstates are conceptualized [pdf document]. Retrieved from http://tedlab.mit.edu/~mekline/ELC2016/25_ELC2016He\&Arunachalam.pdf
Jackendoff, R. (1991). Parts and boundaries. Cognition, 41, 9-45.
Krifka, M. (1989). Nominal reference, temporal constitution and quantification in event semantics. In R. Bartsch, J. van Benthem, \& P. van Emde Boas (Eds.), Semantics and Contextual Expression, Groningen-Amsterdam Studies in Semantics 11. Dordrecht: Foris Publications.
Krifka, M. (1998). The origins of telicity. In S. Rothstein (ed.), Events and Grammar. Dordrecht: Kluwer.
Lakusta, L., \& Landau, B. (2005). Starting at the end: the importance of goals in spatial language. Cognition, 96, 133.

Lakusta, L., \& Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal. Cognitive Science, 36, 517-544.
Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. Cognitive Science, 34, 1064-1092.
Parsons, T. (1990). Events in the Semantics of English: A Study in Subatomic Semantics. Cambridge, MA: MIT Press.
Regier, T., \& Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. Cognitive Science, 31, 705-719.
Strickland, B., \& Keil, F. (2011). Event completion: Event based inferences distorts memory in a matter of seconds. Cognition, 121, 409-415.
Swallow, K., Zacks, J., \& Abrams, R. (2009). Event boundaries in perception affect memory encoding and updating. Journal of Experimental Psychology, 138, 236257.

Tenny, C. (1987). Grammaticalizing Aspect and Affectedness (Doctoral dissertation). MIT, Cambridge, MA.

Vendler, Z. (1957). Verbs and times. The Philosophical Review, 66, 143-160. Reprinted as Chapter 4 of Z. Vendler (Ed.), Linguistics in Philosophy (1967).
Wagner, L. (2006) Aspectual bootstrapping in language acquisition: Telicity and transitivity. Language Learning and Development, 2, 51-76.
Wagner, L. (2009). Manners and goals in pre-linguistic thought: the origins of aspectual construal. Proceedings of the 33th Annual BU Conference on Language Development (pp. 599-610). Somerville: Cascadilla Press.
Wagner, L., \& Carey, S. (2003). Individuation of objects and events: a developmental study. Cognition, 90, 163-191.
Wellwood, A., Hespos, S. J., \& Rips, L. (in press). The object : substance :: event : process analogy. In T. Lombrozo, S. Nichols, \& J. Knobe (Eds.) Oxford Studies on Experimental Philosophy. New York, NY: OUP.
Zacks, J., \& Swallow, M. (2007). Event segmentation. Current Directions in Psychological Science, 16, 80-84.
Zacks, J., \& Tversky, B. (2001). Event structure in perception and conception. Psychological Bulletin, 127, 3-21.

# Promoting Children's Relational Understanding of Equivalence 

Kristen Johannes (kjohann@WestEd.org)<br>WestEd, 300 Lakeside Dr., Oakland, CA 94612 USA<br>Jodi Davenport (jdavenp@WestEd.org)<br>WestEd, 300 Lakeside Dr., Oakland, CA 94612 USA<br>Yvonne Kao (ykao@WestEd.org)<br>WestEd, 400 Seaport Court, Redwood City, CA 94063 USA<br>Caroline Byrd Hornburg (hornburg.1@nd.edu)<br>University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556 USA<br>Nicole M. McNeil (nmcneil@nd.edu)<br>University of Notre Dame, 102 Haggar Hall, Notre Dame, IN 46556 USA


#### Abstract

Deep understanding of mathematical equivalence is critical for later mathematical understandings. However, research studies and national test results have repeatedly demonstrated that many students fail to develop adequate understanding of equivalence. Recent work from McNeil and colleagues proposes that this failure is partly due to the format of traditional instruction and practice with highly similar problems. Specifically, the change-resistance account (McNeil \& Alibali, 2005) proposes that students struggle with equivalence because they have developed overgeneralized "rules" that affect how they process and approach math problems, (e.g., the operators are always on the left side, the equal sign means to "do something" or "give the answer") and fail to see equations having two separate sides that are being related to one another. Extensive practice with problems in a similar format (e.g., those that present all arithmetic operations on the left side of the equal sign) encourages students to develop ineffective mental models of problem types. We replicate and extend prior work that brings cognitive science research to the classroom. Our findings indicate that applying research-based design principles to arithmetic practice improves student understanding of mathematical equivalence enough to support transfer to novel problem types.


Keywords: Mathematical representations; relational reasoning; mathematics education; randomized control trial

## Introduction

Can a research-based, early elementary intervention help students learn key concepts that may prevent later struggles in algebra? Research suggests that understanding mathematical equivalence is a critical component of algebraic reasoning (Carpenter, Franke, \& Levi, 2003; Charles, 2005; Knuth, Stephens, McNeil, \& Alibali, 2006). However, the majority of US students fail to reason with and apply concepts of equivalence (McNeil \& Alibali, 2005), making encoding errors when remembering mathematical equations (e.g., McNeil \& Alibali, 2004), and interpreting the equal sign to mean "calculate the total"
rather than "two amounts are the same" (e.g., Behr, Erlwanger, \& Nichols, 1980).

Why do so many students lack a relational understanding of the equal sign? McNeil and Alibali (2005) proposed a change-resistance account: traditional arithmetic instruction that focuses on procedures (i.e., solving problems such as $3+4=$ ) promotes a misconception of the equal sign as a request for an answer and interferes with the development of relational understanding. The majority of examples of arithmetic problems in early elementary math curricula show operations (e.g., addition and subtraction) on the left of the equal sign and the "answer" on the right (Seo \& Ginsburg, 2003). Children detect and extract patterns from these examples and ultimately construct long-term memory representations. Although default representations typically speed computation in the problem-solving contexts that children encounter most frequently, these representations may lead to difficulties when patterns are mistakenly transferred to similar, but non-applicable, problem types (e.g., Bruner, 1957).

McNeil and Alibali characterize the representations that develop in early mathematics as "operational patterns" as they reflect an understanding of arithmetic that focuses on the operators (e.g.,,,$+- \times, \div$ ) rather than the relational nature of mathematical equations. Research has identified three types of operational patterns that represent a distorted view of arithmetic and hinder conceptual understanding of the underlying mathematics. First, children learn to expect math problems to have all operations on the left side of the equal sign, with the equal sign immediately before the answer blank on the right, an "operations $=$ answer" problem format (McNeil \& Alibali, 2004). Second, children learn to interpret the equal sign operationally as a symbol to do something (Baroody \& Ginsburg, 1983; Behr et al., 1980). Third, children learn to perform operations on all given numbers in a math problem (e.g., add up all the numbers in an addition problem, McNeil \& Alibali, 2005).

Once entrenched, children rely on these potentially misleading patterns when encoding, interpreting, and solving novel mathematics problems. Students that expect all problems to have operations on the left fail to correctly encode the problem being asked. For instance, after briefly viewing the problem "7 + 4+5=7+_" many children rely on their knowledge of the "operations = answer" problem format and erroneously remember the problem as "7 $+4+5+7=\ldots$ (McNeil \& Alibali, 2004). Students also struggle to interpret what a mathematical problem is asking. When asked to define the equal sign-even in the context of a mathematical equivalence problem-many children treat it like an arithmetic operator (like + or - ) that means they should calculate the total (McNeil \& Alibali, 2005). Finally, entrenched patterns mislead students to solve the problem " $7+4+5=7+\ldots$ " by performing all given operations on all given numbers and put 23 (instead of 9) in the blank (McNeil, 2007; Rittle-Johnson, 2006). These findings support the idea that children's difficulties with mathematical equivalence are partially due to inappropriate knowledge derived from overly narrow experience with traditional arithmetic.

## The ICUE Intervention

Current math practice seems to promote the development of faulty representations, and the change-resistance account's focus on "operational patterns" offers design principles for instruction to improve students' understanding of equivalence. Initially, researchers hypothesized that greater exposure to "non-traditional" arithmetic practice (e.g., presenting operations on the right side of the equation, "_ $=2+4$," [McNeil et al., 2011], organizing practice by equivalent sums [ McNeil et al., 2012], and using relational phrases such as "is equal to" instead of the equal sign in problems [Chesney, McNeil, Petersen, \& Dunwiddie, 2012]). may prevent students from developing operational patterns. Though practice with non-traditional arithmetic in a classroom intervention led to improved outcomes over traditional instruction, a number of students failed to reach proficiency (McNeil, Fyfe, \& Dunwiddie, 2015).

To further promote mastery of equivalence, McNeil and colleagues added additional design features beyond nontraditional arithmetic practice. The current version of the materials, dubbed Improving Children's Understanding of Equivalence (ICUE), consists of second grade student activities that reduce reliance on operational patterns and promote deep understanding of mathematical equivalence through four key components that have independently been shown to be effective:

1. Non-traditional arithmetic practice (Chesney et al., 2012; McNeil et al., 2012, 2015, 2011);
2. Lessons that first introduce the equal sign outside of arithmetic contexts (e.g., " $28=28$ ") before introducing arithmetic expressions (e.g., Baroody \& Ginsburg, 1983; McNeil, 2008);
3. Concreteness fading exercises in which concrete, real-world, relational contexts (e.g., sharing stickers,
balancing a scale) are gradually faded into the corresponding abstract mathematical symbols (e.g., Fyfe, McNeil, Son, \& Goldstone, 2014); and
4. Activities that require students to compare and explain different problem formats and problemsolving strategies (e.g., Carpenter et al., 2003; RittleJohnson, 2006).

## The Current Study: Improving Children's Understanding of Equivalence

A pilot study found the ICUE intervention was successful in improving student understanding of mathematical equivalence (Byrd, McNeil et al. 2015; McNeil, Hornburg, Brletic-Shipley, \& Matthews, under review). The current study sought to replicate the findings with a new population of students and additionally investigate whether the learning transferred to the mathematical practice of generating explanations.

To replicate Byrd et al.'s (2015) pilot study, we compared the full ICUE intervention to a control condition consisting solely of non-traditional mathematical practice and measured students' ability to encode equations, solve problems, and define the relational function of the equal sign.

To test whether the learning transferred to the ability of students to generate mathematical explanations related to arithmetic problems, we gave students performance tasks from the Silicon Valley Mathematics Initiative's (SVMI) Mathematics Assessment Collaborative (MAC). MAC partners with the Mathematics Assessment Resource Service (MARS) to develop tasks that assess core mathematical ideas and practices taught in each grade level. Tasks require students to solve complex math problems as well as give open-ended explanations of their reasoning. For each task, MARS provides scoring rubrics and scorer training procedures, student performance statistics, and examples of common student errors (Foster \& Noyce, 2004).

Our research questions were:

1. Does ICUE promote measurable gains in children's understanding of equivalence?
2. Do the benefits of ICUE activities transfer to generating mathematical explanations?

## Method

## Design

We used a cluster-randomized control trial design to examine the efficacy and generalizability of the ICUE intervention relative to an active control program. Teachers were randomly assigned to use the either the ICUE intervention or Active Control materials. The active control consisted of workbook activities to control for time on task. The active control contained non-traditional arithmetic practice but not the additional components present in ICUE, described above.

Participants. Five second-grade teachers (three treatment, two control) from three California schools used the activities in their classrooms. Class sizes ranged from 21 to 32 , and we analyzed data from 81 students who completed the ICUE activities and measures and 49 students who completed the Active Control activities and measures.

## Procedure and Materials

The procedure for ICUE Treatment and Active Control conditions were identical, differing only in the content of the materials used by teachers and students. Each teacher received training on the study purpose, features of the activities, and strategies for integrating the activities into their typical mathematics curriculum.

Prior to starting the study, participating teachers completed online surveys assessing their mathematics teaching experience and classroom structure and dynamics.

After administering a pretest, teachers used the study materials for approximately 15 minutes twice each week for 16 weeks. In both conditions, teachers were asked to use the study materials to supplement, rather than replace, current instruction, and to limit session duration to 20 minutes.

After completing the 32 sessions, teachers administered the same pretest measure of mathematical equivalence, a proximal transfer measure, two measures of transfer to mathematical explanations, and the Math Concepts subtest of the Iowa Test of Basic Skills.

Active Control. Teachers in the Active Control condition received a set of student workbooks (see Figure 1) and a teacher guide.


Figure 1. Sample workbook page from the Active Control condition materials featuring non-traditional math practice.

ICUE. Teachers in the ICUE Treatment condition received a set of student workbooks (see Figure 2), a teacher guide, and a set of classroom manipulatives including balance scales and flashcards.

## Measures

Pre- and post-test measures of mathematical equivalence. We assessed children's understanding of
mathematical equivalence before and after the interventions using similar measures of equation encoding, equation solving, and defining the equal sign used in previous work by McNeil and colleagues (Byrd et al., 2015; McNeil \& Alibali, 2005; McNeil et al., 2015).


Figure 2. Sample workbook page from the ICUE Treatment condition materials featuring a concreteness fading exercise.

Equation encoding. The encoding measure consisted of recalling four mathematical equivalence problems (e.g., $5+$ $4=3+\ldots$ ) presented one at a time. Each equation was visible for five seconds and students were instructed to remember and write down exactly what they saw after the equation was hidden from view. Responses were coded as correct if the student wrote the equation exactly as shown (i.e., the correct numbers and symbols in the correct order).

Equation solving. The equation solving measure consisted of eight equations with operations on both sides of the equal sign (e.g., $3+5+6=3+$ $\qquad$ ).
Defining the equal sign. The defining the equal sign measure prompted students to write responses to three questions about the equal sign symbol (=): 1) What is the name of this math symbol? 2) What does this math symbol mean? And, 3) Can it mean anything else? Teachers read each question aloud and waited for students to write their responses before moving on to the next question. Responses were coded as relational if the response defined the equal sign as relating two sides of the equation (e.g., two amounts are the same, something is equivalent to another thing). ${ }^{1}$

## Measures of knowledge transfer.

Proximal transfer measure. The proximal transfer measure, used by Byrd et al. (2015), consisted of nine more advanced problems of mathematical equivalence, not strictly aligned with the ICUE intervention. The transfer questions included equations with operations on both sides of the equal sign involving subtraction (e.g., $2+5+3=14$ -
$\qquad$

[^113]larger numbers (e.g., $13+18=\ldots+19$ ), "word problem" items featuring story-to-equation translation, and an "explaining equivalence" problem, which asked students to decide whether the same number should appear in two equations and explain their reasoning.

Distal transfer to mathematical explanations. We selected two MARS items that tested students' understanding of mathematical equivalence, described below. Items were scored by project staff following scorer training, calibration, and reliability procedures established by MARS (Foster \& Noyce, 2004).

Incredible Equations. In this task, students are asked to fill in the missing parts of equations such as "_ $+8+\ldots=$ 16 " and " $11+5=\ldots+8$." Students are asked to explain how they know their answer is correct. When 6,305 students took the task in 2007, the mean score was 6.08 out of 10 with a standard deviation of 2.5 (MARS, 2007).

Agree or Disagree? In this task, students are asked if they agree or disagree with two number sentences: " $8+5=5+8$ " and " $6-4=4-6$ ". Students are asked to explain their answers using words, numbers, or pictures. MARS administered this task to 4,585 second graders in 2004 and found the mean score was 3.10 out of 6 with a standard deviation of 1.94 (MARS, 2004).

Iowa Test of Basic Skills. To make sure any gains in understanding of equivalence do not come at the expense of problem-solving fluency, students completed the Math Concepts subtest of Level 8 of the Iowa Tests of Basic Skills (ITBS), which served as a measure of general mathematical reasoning. Participation in the ICUE Treatment neither helped nor hurt students' performance on this measure, relative to the Active Control group $(t(83)=1.48, n s)$, establishing that the intervention does not improve understanding of equivalence at the expense of general computational fluency.

## Results

## Does ICUE promote measurable gains in children's understanding of equivalence, relative to an Active Control?

We assessed three critical abilities identified by McNeil and colleagues as necessary for success in reasoning about equivalence (Byrd et al., 2015; McNeil et al., under review):

1. Equation encoding: the ability to accurately encode and recreate an equation after seeing it briefly;
2. Equation solving: the ability to solve equations that feature operations on both sides of the equal sign; and
3. Defining the equal sign $(=)$ : the ability to identify " $=$ " as a symbol that signals a relation between two equal numbers or quantities.

Specifically, we examined students' gains in performance on identical pre-intervention and post-intervention tests that assessed the three abilities above. For each of the target abilities, we compared the gains made by students in the

ICUE Treatment condition to those of students in the Active Control condition (Figure 3).

There were no reliable differences between pretest scores for each group, and students in the ICUE Treatment condition made substantially greater gains during the intervention than students in the Active Control condition. The proportion of correct responses for Equation solving items increased by 0.65 for ICUE students, compared to only 0.065 for Active Control students $(t(119)=48.8$, $p<.001$; Cohen's $d>3$ ); the proportion of correct responses for Equation encoding items increased by 0.34 for ICUE compared to 0.26 for Active Control $(t(52)=5.31, p<.001$; Cohen's $d>3$ ); and the proportion of correct definitions of the equal sign increased by 0.38 for ICUE compared to 0.02 for Active Control $(t(125)=8.42, p<.001$; Cohen's $d>3$ ). These results suggest that the ICUE Treatment intervention leads to systematic and measureable gains in children's understanding of and reasoning about mathematical equivalence.


Figure 3. Mean performance gains from pre- to post-test for children in the ICUE and Active Control groups.

## Do the benefits of ICUE activities transfer to more challenging material and generating mathematical explanations?

We explored whether the knowledge that children gained from the intervention activities transferred to problemsolving tasks that were not strictly aligned with the content and goals of the ICUE or Active Control interventions. We first examined performance on a proximal researcherdeveloped measure that included a series of complex equation solving items, word problem items that required translating story content into mathematical equations, and an explaining equivalence item that required students to justify why two sides of an equation were equal (i.e., "Is the number that goes in the the same number in the following two equations? Explain your reasoning."). We compared the performance of ICUE and Active Control students on the measure, which was administered after each group completed all intervention activities (Figure 4).


Figure 4. Mean ICUE and Active Control group performance for researcher-developed transfer items.

Students in the ICUE condition scored reliably higher, on average, than students in the Active Control condition on both complex equation solving items $(t(81)=3.44, p<.01$; Cohen's $d=1.5$ ) and word problem items $(t(129)=5.31$, $p<001$; Cohen's $d=2.5$ ). However, the groups did not differ in their mean performance on the explaining equivalence item $(t(81)=0.15, n s)$.

We also examined transfer to the MARS items. We measured post-intervention performance on the "Incredible Equations" task (scored out of a possible 10 points) and the "Agree or Disagree?" task (scored out of a possible 6 points). As before, we compared performance by students in the ICUE and Active Control conditions, shown in Figure 5. As one teacher from each condition failed to return the MARS posttest materials, results are reported from two treatment teachers and one control teacher.

Students in the ICUE condition performed reliably better than Active Control students on both the Incredible Equations $(t(54)=2.83, p<.05$; Cohen's $d=0.32)$ and Agree or Disagree? tasks $(t(47)=2.36, p<.05$; Cohen's $d=0.43)$.

## Conclusions

A deep understanding of mathematical equivalence is a key building block for later mathematical understandings. However, research studies and national test results have repeatedly demonstrated that many students fail to develop this understanding. The change-resistance account suggests that traditional instruction that relies on extensive practice with problems in a single format may be contributing to students' difficulties by encouraging students to develop ineffective mental models of problem types.

In the current study, we sought to replicate and extend prior work that brings research from the lab into the classroom. The change-resistance account proposes that students struggle with equivalence because they have developed overgeneralized "rules" that affect how they process and approach math problems, (e.g., the operators are always on the left side, the equal sign means to "do something" or "give the answer") and fail to see equations having two separate sides that are being related to one another.


Figure 5. Mean ICUE and Active Control group performance for MARS transfer items.

Overall, our findings indicate that applying researchbased design principles in the form of multiple types of practice improved student understanding of the critical concept of mathematical equivalence.
Our findings replicate Byrd et al. (2015), who found that activities that include the use of the equal sign outside of arithmetic contexts, that start with concrete examples and fade to extractions, and that explicitly prompt students to compare and explain different problem formats and strategies improve student understanding of mathematical equivalence beyond non-traditional arithmetic practice alone.

Students receiving the ICUE materials demonstrated improved performance in equation solving, equation encoding, and providing relational definitions of the equal sign. These improvements did not come at the expense of arithmetic problem-solving fluency, as measured by the ITBS. Further, the learning in ICUE transferred to greater student abilities to solve complex equations and word problems.

Students in both conditions struggled with the researcherdeveloped item that required students to explain equivalence. Their poor performance may reflect confusion with equivalence that persists for more complicated problems with multiple "terms" and different types of operators (both addition and subtraction), a confusion that was reflected in students' explanations of their answers.
The robust improvements on the MARS items supports the possibility that the lack of transfer in the equivalence explanation question was due to confusion regarding multiple terms and operators rather than the ability to generate the explanation. These established items, developed externally, also asked students to explain equivalence, but used blanks, rather than variables, to reflect the unknown entities. On both items, students in the ICUE condition outperformed the students in the active control condition. These findings suggest that the additional practice comparing and explaining different problem formats helped students gain a deeper understanding of not only whether different examples were equivalent, but also why or why not.

Why is it important to test the synergistic effect of research-based design principles? Instructional designers face a large number of decisions in selecting appropriate activities and tasks for students. Though much research seeks to identify how different facets work independently, if research in cognitive science is to extend meaningfully to practice the cumulative effects of using multiple strategies must be tested. Our small-scale cluster-randomized trial suggests that the multi-component ICUE intervention was more effective than an active control of non-traditional arithmetic practice (which in prior work was also more effective than traditional instruction).

Future work, in progress, will test the efficacy of the ICUE intervention in a large-scale cluster-randomized trial with diverse students across the state of California. This work demonstrates how findings in the lab can be successfully implemented in authentic classroom settings to improve student learning outcomes.

## Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150088 to WestEd. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## References

Baroody, A. J., \& Ginsburg, H. P. (1983). The effects of instruction on children's understanding of the "equals" sign. Elementary School Journal, 84, 199-212.
Behr, M., Erlwanger, S., \& Nichols, E. (1980). How
children view the equal sign. Mathematics Teaching, 92, 13-15.
Bruner, J. S. (1957). On perceptual readiness. Psychological Review, 2, 123-152.
Byrd, C. E., McNeil, N. M., et al. (2015). Pilot test of a comprehensive intervention to improve children's understanding of math equivalence. In D. I. C. Francis (Organizer), Explorations in mathematics in the elementary grades. Paper presented at the Annual Meeting of the AERA, Chicago, IL.
Carpenter, T. P., Franke, M. L., \& Levi, L. (2003). Thinking mathematically: Integrating arithmetic and algebra in elementary school. Portsmouth, NH: Heinemann.
Charles, R. I. (2005). Big ideas and understandings as the foundation for elementary and middle school mathematics. NCSM: J. of Mathematics Education Leadership, 7, 9-24.
Chesney, D. L., McNeil, N. M., Petersen, L. A., \& Dunwiddie, A. E. (2012). Arithmetic practice that includes relational words promotes conceptual understanding and computational fluency. Poster presented at the APS Annual Convention, Chicago, IL.
Foster, D. \& Noyce, P. (2004). The Mathematics
Assessment Collaborative: Performance testing to improve instruction. Phi Delta Kappan, 85, 367-374.
Fyfe, E. R., McNeil, N. M., Son, J. Y., \& Goldstone, R. L.
(2014). Concreteness fading in mathematics and science instruction: A systematic review. Educational Psychology Review, 26, 9-25.
Knuth, E. J., Stephens, A. C., McNeil, N. M., \& Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. J. for Research in Mathematics Education, 37, 297-312.
Mathematics Assessment Resource Service. (2004). Tools for teachers: Grade 2. Retrieved from http://www.svmimac.org.
Mathematics Assessment Resource Service. (2007). Tools for teachers: Grade 2, part 3. Retrieved from http://www.svmimac.org.
McNeil, N. M. (2007). U-shaped development in math: 7-year-olds outperform 9 -year-olds on equivalence problems. Developmental Psychology, 43, 687-695.
McNeil, N. M. (2008). Limitations to teaching children $2+$ $2=4$ : Typical arithmetic problems can hinder learning of mathematical equivalence. Child Development, 79, 15241537.

McNeil, N. M., \& Alibali, M. W. (2004). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. Cognitive Science, 28, 451-466.
McNeil, N. M., \& Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. Child Development, 76, 883-899.
McNeil, N. M., Chesney, D. L., et al. (2012). It pays to be organized: Organizing arithmetic practice around equivalent values facilitates understanding of math equivalence. J. of Educational Psychology, 104, 11091121.

McNeil, N. M., Fyfe, E. R., \& Dunwiddie, A. E. (2015). Arithmetic practice can be modified to promote understanding of mathematical equivalence. J. of Educational Psychology, 107, 423-436.
McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A.
E., \& Brletic-Shipley, H. (2011). Benefits of practicing $4=$ $2+2$ : Nontraditional problem formats facilitate children's understanding of mathematical equivalence. Child Development, 82, 1620-1633. McNeil, N. M., Hornburg, C. B., Brletic-Shipley, H., \&

Matthews, J. M. (under review). Improving children's understanding of math equivalence via an intervention designed to reduce reliance on knowledge of traditional arithmetic.
Rittle-Johnson, B. (2006). Promoting transfer: The effects of self-explanation and direct instruction. Child Development, 77, 1-15.
Seo, K.-H., \& Ginsburg, H. P. (2003). "You've got to carefully read the math sentence...": Classroom context and children's interpretations of the equals sign. In A. J. Baroody \& A. Dowker (Eds.), The development of arithmetic concepts and skills: Constructing adaptive expertise (pp. 161-187). Mahwah, NJ: Erlbaum.

# Opponent Uses of Simplicity and Complexity in Causal Explanation 

Samuel G. B. Johnson ${ }^{1}$, J. J. Valenti ${ }^{2}$, \& Frank C. Keil ${ }^{1}$<br>(sgbjohnson@gmail.com, jvalen23@jhu.edu, frank.keil@yale.edu)<br>${ }^{1}$ Department of Psychology, Yale University, New Haven, CT 06520 USA<br>${ }^{2}$ Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218 USA


#### Abstract

People often prefer simpler explanations because they have higher prior probability. However, simpler explanations are not always normatively superior because they often do not fit the data as well as complex explanations. How do people negotiate this trade-off between prior probability (favoring simplicity) and goodness-of-fit (favoring complexity)? Here, we argue that people use opponent heuristics-relying on simplicity as a cue to prior probability but complexity as a cue to goodness-of-fit (Study 1). We also examine factors that lead one or the other heuristic to predominate in a given context. Study 2 finds that people have a stronger simplicity preference in deterministic rather than stochastic contexts, while Study 3 finds that people have a stronger simplicity preference for physical rather than social causal systems. Together, we argue that these cues and contextual moderators act as powerful constraints that help to specify otherwise illdefined hypothesis comparison problems.


Keywords: Causal reasoning; explanation; probabilistic reasoning; heuristics; judgment under uncertainty.

## Introduction

The principle of parsimony has a long and venerable pedigree. It has been discussed since at least Aristotle, who wrote in his Physics that "nature operates in the shortest way possible," and it has since become one of the core tools in our argumentative arsenal as scientists. Of course, this principle was given its most famous formulation given by William of Occam, who advised against "multiplying entities beyond necessity."

Simplicity is not only a core notion in science and philosophy, but may well be an organizing principle of cognition (Chater \& Vitányi, 2003). People prefer simpler causal explanations (Lombrozo, 2007), category assignments (Pothos \& Chater, 2002), and perceptual organizations (van der Helm \& Leeuwenberg, 1996), and more easily learn simple concepts (Feldman, 2000).

This principle is not arbitrary. Other things equal, simpler explanations are more likely to be true because they have higher prior probability. Consistent with this analysis, Lombrozo (2007) found that people use simplicity as a heuristic for estimating prior probabilities. In her experiments, participants performing simulated medical diagnoses would not accept a complex explanation over a simple one unless the prior probabilities favored the complex explanation by a factor of 4 . Further, participants who had a simplicity bias had distorted memories of the disease base rates, recalling the simpler explanations as having had higher prior
probabilities than they in fact did. Thus, people's preference for simple explanations, though sometimes stronger than normatively warranted, appears to track the probabilistic logic favoring simpler explanations.

Yet, simplicity has its limits because a simple and a complex explanation do not always fit the data equally well. There is generally a U-shaped curve in how simple an explanation ought to be. Too complex, and the explanation has a lower prior probability and overfits the data; too simple, and it does not account for the nuance of the phenomenon (Forster \& Sober, 1994). How, if at all, does cognition perform this trade-off?
We propose that people use opponent heuristics to compare a simpler versus a more complex explanation. This view incorporates Lombrozo's (2007) insight that people use simplicity to estimate prior probability-the $\mathrm{P}\left(H_{i}\right)$ terms in Bayesian hypothesis comparison-but couples it with the idea that people also use complexity to estimate likelihoods-the $\mathrm{P}\left(E \mid H_{i}\right)$ terms that measure the goodness-of-fit of the evidence to the data.
For example, if a patient is sneezing and has a stomach ache, the patient could have a cold. This explanation is simple, but fits the data imperfectly. If we took a random sample of the population, a reasonably large fraction would have a cold at any given time-so this explanation has high prior probability. But among those people who have a cold, how many of them would both be sneezing and have a stomach ache? The facts here are complex, and this simple explanation does not fit very well.

In contrast, the patient could have both allergies and a stomach virus. This explanation is more complex, but fits the data neatly. In a random sample of the population, a fairly small number would have both allergies and a stomach virus. Yet, many of those who do have both diseases would likely be suffering from both sneezing and a stomach ache. Even though the prior probability of this complex explanation is low, it fits the data very well.

In this case, simplicity seems to be associated with our estimate of prior probability and complexity seems to be associated with our estimate of likelihood. Of course, this explanation was engineered to produce these intuitions by relying on specific beliefs we have about these diseases. The opponent heuristic account proposes that people also use simplicity and complexity as cues in cases where they cannot estimate probabilities directly from background knowledge. Study 1 tests this possibility.

Initial evidence for this idea comes from studies of intuitive curve-fitting-a superficially distinct but deeply related problem to causal explanation. For any set of scatterplot data, many different trend curves can be drawn
to explain the data, but statistical theory can tell us which curve has the best predictive power, fitting as much of the underlying signal as possible while fitting little of the noise. Yet, people tend to choose curves that are more complex than they normatively should be, rather than curves that are too simple (Johnson, Jin, \& Keil, 2014), as one would expect if people only have a simplicity heuristic but no complexity heuristic. Indeed, these curvefitting studies uncovered direct evidence of a complexity bias, because participants judged the more complex to be literally closer fits to the data, even when the actual fit was held constant. This finding is also consistent with naturalistic studies of everyday verbal explanations drawn from an Internet corpus, for which the best explanations actually tend to be fairly complex (Zemla et al., 2017).

Why is this pair of heuristics useful? Simplicity is just the absence of complexity. How, then, can a pair of heuristics accomplish any more than a single heuristic, when these two heuristics rely on the same cue? While it may seem more parsimonious to assume that people merely use one cue in a U-shaped manner, it is difficult to specify, for any given problem, where the bend in this $U$ should be. Contextual factors (along with background knowledge) must work to calibrate the strength of these two heuristics, in order to produce a unique solution in any given case. Although there is no reason to think that a context-sensitive dual heuristic solution will give an optimal answer, there is reason to think that it may bring the reasoned closer to the right part of the hypothesis space, compared to either heuristic working alone or to any cookie-cutter U-shaped response to simplicity that is not calibrated to the explanatory problem. The current studies look at two possible contextual factors that might modulate the strength of the two heuristics.

First, we consider the determinism of the causal system. In previous studies of simplicity (Lombrozo, 2007), explanations have been produced for deterministic causal systems. In such systems, it is rational to prefer simple explanations. If disease $A$ always causes symptoms $X$ and $Y$, while disease $B$ always causes symptom $X$ and disease $C$ always causes symptom $Y$, the issue of likelihoods or goodness-of-fit simply does not come up: Disease $A$ perfectly explains the evidence, and so do Diseases $B$ and $C$ together. The only issue is which explanation has the higher prior probability, and the simplicity heuristic tells us that, absent any other information, the answer is Disease $A$. Therefore, there is no reason to invoke a complexity heuristic to countervail against the presumption of a simple explanation.

In contrast, when the causal system is stochastic, the likelihoods become a more crucial part of the computation. If disease $A$ sometimes causes $X$ and sometimes causes $Y$, while disease $B$ sometimes causes $X$ and disease $C$ sometimes causes $Y$, it is difficult to evaluate whether the evidence (symptoms $X$ and $Y$ ) are made likelier by disease $A$ or by diseases $B$ and $C$ combined: It depends on the nature of "sometimes." Yet,
in the real world, it is the exception rather than the rule to have precise quantitative information about these likelihoods in stochastic systems. If people rely on a complexity heuristic in such cases, they would judge the likelihood of the evidence to be higher for an individual with two diseases than for an individual with one disease. Study 2 tests whether stochastic contexts therefore lead to a weaker simplicity preference.

Second, we consider the content domain of a causal system. People seem to have different beliefs about the causal textures of different domains. Whereas people tend to identify physical events as having relatively few causes, social events are often thought to have many causes (Strickland, Silver, \& Keil, 2016). This suggests that people may calibrate their prior expectations to more complex explanations in the social domain, compared to the physical domain. Furthermore, people may even deploy different causal concepts across domains (Lombrozo, 2010). Whereas causal claims about physical systems appear to be evaluated in terms of transference and contact, social causal claims appear to be evaluated counterfactually. This too may reinforce the intuition that physical events typically result from highly specified causal factors, whereas social events result from more complex configurations of counterfactual conditions. Since such complex conditions can seldom be known, social systems are often highly unpredictable.

As a consequence of these domain-specific beliefs, people may rely on simplicity as a cue to prior probability to a differing degree across domains. Whereas simplicity is likely to be a potent heuristic for evaluating explanations of physical causation, it may be a weaker cue for evaluating explanations of social causation, if people have a meta-theory that assigns higher prior probabilities to complex social causal explanations, as compared to physical causal explanations. In addition, if social causal systems are seen as more stochastic, this would increase the importance of the complexity heuristic for evaluating explanations of social causation, as compared to physical causation. With a weaker simplicity heuristic and stronger complexity heuristic, people may have a less pronounced bias toward simple explanations in the social domain. Study 3 tests this idea.

## Study 1

To a Bayesian, the key quantities required to compare two hypotheses are the relative prior probabilities of the hypotheses (the prior odds), and the relative fit of each explanation to the data (the likelihood ratio). Study 1 tests whether people use simplicity to estimate these quantities.

Study 1A seeks converging evidence for Lombrozo's (2007) claim that people assign higher prior probabilities to simple hypotheses. Study 1 B tests whether this heuristic favoring simple explanations might be opposed by a heuristic that assigns higher likelihoods to more complex explanations: Do people believe that complex explanations are better fits to the data?

## Method

Participants in all studies were recruited from Amazon Mechanical Turk. Each study included a series of check questions at the end, and participants were excluded from analysis if they answered more than $33 \%$ incorrectly.

Participants ( $N=80,9$ excluded) were randomly assigned to Study 1A (making judgments about priors probabilities) or to Study 1B (making judgments about likelihoods). In both studies, participants completed four items about diseases, similar to the following problem:
There is a population of elves that lives at Gelfert's Glacier. Sometimes the elves have medical problems such as feverish muffets or wrinkled ears.
A Yewlie infection can cause feverish muffets.
A Yewlie infection can cause wrinkled ears.
Hepz's disease can cause feverish muffets.
Aeona's syndrome can cause wrinkled ears.
Nothing else is known to cause an elf's muffets to be feverish of the development of wrinkled ears.

On the same screen, participants completed a series of 10 true/false questions to ensure comprehension.
Participants in Study 1A were then asked to judge the relative prior probabilities ("Imagine that we randomly select an elf from Gelfert's Glacier. Which of the following types of elves do you think we are more likely to have selected?") on a 10 -point scale, with one end corresponding to the simple explanation ("An elf who has a Yewlie infection only") and one end to the complex explanation ("An elf who has both Hepz's disease and Aeona's syndrome"). Participants in Study 1B were asked to judge the likelihoods ("Imagine an elf who has a Yewlie infection only, and another elf who has both Hepz's disease and Aeona's syndrome. Which elf do you think is more likely to develop both feverish muffets and wrinkled ears?") on the same scale.

## Results and Discussion

Data for all studies were recoded so that negative numbers correspond to the simple explanation and positive numbers to the complex explanation.

Participants in Study 1A used a simplicity heuristic, indicating that a randomly selected elf was more likely to have one disease than two diseases $[M=-2.19, S D=$ 1.78; $t(33)=7.19, p<.001, d=1.23]$. This is consistent with Lombrozo's (2007) studies, where overwhelming prior odds were required before participants would favor a complex over a simple explanation in deterministic cases.

However, the story was different for judgments of likelihoods or goodness-of-fit. Here, participants favored the complex explanation $[M=1.41, S D=2.35 ; t(36)=$ $3.65, p=.001, d=0.60]$. This complexity bias in estimating likelihoods was substantial in magnitude ( $d=$ 0.60 ), though smaller than the simplicity bias in estimating priors ( $d=1.23$ ), at least for these stimuli.

These results shows that people do not blindly prefer simple explanations, but instead calibrate their
preferences according to the question asked. Even though the problem did not include any probability information, participants used simplicity and complexity to estimate different probabilistic quantities in opposing ways.

## Study 2

In any causal system where there is uncertainty about which explanation is correct, the prior probabilities of each explanation must be less than 1 , since otherwise there is no reason to observe any data (as it will fail to move the posteriors). However, the likelihoods differ across deterministic and stochastic systems. In deterministic systems, the evidence is always produced with probability 1 by its causes, whereas in stochastic systems, these likelihoods are less than 1.

If explanatory heuristics exist in part because degrees of uncertainty are difficult to estimate and to use in computations, then a simplicity heuristic will always be a useful tool for estimating priors, since they are always uncertain. However, a complexity heuristic is only useful in stochastic systems, where the likelihoods are uncertain. Thus, both heuristics should be at work in stochastic systems (a simplicity heuristic pushing toward simpler explanations and a complexity heuristic pushing toward more complex explanations), whereas only the simplicity heuristic applies in deterministic systems (pushing toward simpler explanations, without an opposing force pushing toward more complex explanations). This leads to the prediction that people should especially favor simple explanations for deterministic systems.

## Method

Participants ( $N=80$, 14 excluded) completed four items corresponding to the cover stories used in Study 1. For one of these items-in the $100 \%$ condition-the causal system was described as deterministic, in that the diseases always led to their symptoms ( $100 \%$ likelihood):

Tritchet's syndrome always ( $100 \%$ of the time) causes both sore minttels and purple spots.
Morad's disaease always ( $100 \%$ of the time) causes sore minttels, but the disease never ( $0 \%$ of the time) causes purple spots.
When an alien has a Humel infection, that alien will always ( $100 \%$ of the time) develop purple spots, but the infection will never ( $0 \%$ of the time) cause sore minttels.

The other three items corresponded to the $90 \%, 80 \%$, and $70 \%$ conditions, which differed only in the causal system being described as stochastic:
Tritchet's syndrome often ([80/65/50]\% of the time) causes both sore minttels and purple spots.
Morad's disaease often (([90/80/70]\% of the time) causes sore minttels, but the disease never ( $0 \%$ of the time) causes purple spots.
When an alien has a Humel infection, that alien will often (([90/80/70]\% of the time) develop purple spots, but the infection will never ( $0 \%$ of the time) cause sore minttels.

After reading this information, participants were asked about their favored explanation ("Which do you think is the most satisfying explanation for Treda's symptoms?") on a scale from 0 (Tritchet's syndrome only) to 10 (both Morad's disease and a Humel infection). The conditions were balanced across the cover stories using a Latin square, and items were completed in a random order.

## Results and Discussion

Participants strongly preferred the simple explanation [ $M$ $=-3.81, S D=1.95 ; t(65)=15.84, p<.001, d=-1.95]$ given deterministic ( $100 \%$ ) likelihoods. This replicates Lombrozo's (2007) finding that people strongly favor simple explanations in deterministic causal systems.

The key question is whether this preference would differ in the stochastic conditions, where a complexity heuristic would be more likely at play for understanding the likelihoods. To keep the likelihood ratio objectively identical across conditions, the likelihood for the simple explanation must equal the product of the likelihoods for the components of the complex explanation (i.e., $90 \% \times$ $90 \% \approx 80 \%, 80 \% \times 80 \% \approx 65 \%$, and $70 \% \times 70 \% \approx 50 \%$ ). This calculation assumes that people believe diseases to cause their symptoms independently-an assumption that Lombrozo (2007) validated for her very similar stimuli.

As predicted by the opponent heuristic account, the simplicity bias was weaker in each of the three stochastic conditions, although participants still had a robust simplicity preference in each of them $[M=-3.00, S D=$ 2.68, $t(65)=9.09, p<.001, d=1.12$ for the $90 \%$ condition; $M=-2.50, S D=2.58, t(65)=7.86, p<.001$, $d=0.97$ for the $80 \%$ condition; $M=-2.48, S D=2.45$, $t(65)=8.24, p<.001, d=1.01$ for the $70 \%$ condition]. The simplicity bias in the stochastic conditions, while large ( $d$ from 0.97 to 1.12 ), was smaller compared to the deterministic condition ( $d=1.95 ; p \mathrm{~s}>.012$ ), as predicted.

However, this design is subject to concerns about demand characteristics and difficulties with probabilities that are unrelated to the proposed mechanisms. In particular, the deterministic condition set all likelihoods to $100 \%$, whereas the stochastic condition had to set different likelihoods for the simple explanation and for each component of the complex explanation. Could people have relied on a strategy such as comparing these numerical likelihoods ( $100 \%$ vs. $100 \%$ and $90 \%$ vs. $80 \%$ for complex vs. simple), favoring the complex explanation in the stochastic conditions merely because it was superficially associated with higher numbers?

If this were the case, people should be increasingly less biased toward the simple explanation as the difference between the simple and complex likelihoods increased. This difference increases not only between the deterministic and stochastic conditions, but also across the stochastic conditions ( $90 \%$ vs. $80 \%, 80 \%$ vs. $65 \%$, and $70 \%$ vs. $50 \%$ ). Thus, on this deflationary account there should be large gaps not only between the deterministic and stochastic conditions, but also among
the stochastic conditions. In contrast, the opponent heuristic account predicts a qualitative shift between the deterministic condition and the stochastic conditions that introduce uncertainty into the likelihoods.

The data are more consistent with the latter prediction, as suggested by the similar effect sizes of the simplicity bias across the three stochastic conditions. There is a significant difference between the $100 \%$ and $90 \%$ conditions, where we shift from deterministic to stochastic $[t(65)=2.61, p=.011, d=0.32]$. However, the difference between the $90 \%$ and $80 \%$ conditions reaches only marginal significance $[t(65)=1.88, p=.064, d=$ 0.23 ] and the difference between the $80 \%$ and $70 \%$ conditions is nowhere near significant $[t(65)=0.04, p=$ $.97, d=0.01]$. The deflationary account would predict equally large differences across these sets of conditions.

Thus, determinism may play a role in striking the balance between the simplicity and complexity heuristics. These results also resolve a puzzle about Lombrozo's (2007) findings. Given that people are reasonably wellcalibrated in evaluating explanations in the real world, it is surprising to see such a striking simplicity bias as one finds in her studies, with prior odds of 4-to-1 required to override a simplicity preference when the evidence is perfectly consistent with either hypothesis. Study 2 found that in more ecologically realistic conditions, where the evidence is not perfectly predicted by any explanation, people are more likely to hedge their bets. People may thus make more accurate explanatory inferences in realistic, stochastic environments.

## Study 3

A second contextual factor that may influence preferences of simple and complex explanations is a system's content domain. People believe that physical events have fewer causes than social events (Strickland, Silver, \& Keil, 2016) and use causal concepts relying on physical transference for physical systems but complex counterfactual conditions for social systems (Lombrozo, 2010). Thus, Study 3 tests the possibility that people would use these expectations to calibrate their explanatory inferences, favoring simpler explanations in physical causal systems compared to social systems.

## Method

Participants ( $N=479$, 89 excluded) read 12 items across four content domains (physics, biology, artifact, and social), which were deterministic for half of participants and stochastic for the other half. These items had the same format as the items used in Study 2, but the content was replaced with various items in physical (ultraviolet waves, subatomic particles), biological (disease, agriculture, dieting), artifact (robots, clocks, toys), and social (team dynamics, child behavior, and romantic attraction) causal systems. Participants then made explanatory judgments on the same scale as Study 2. Items were completed in a random order.

|  | Deterministic | Stochastic |
| :--- | :---: | :---: |
| Physical | $-2.76(2.10)$ | $-2.15(2.40)$ |
| Biological | $-2.59(2.19)$ | $-2.15(2.28)$ |
| Artifact | $-2.32(2.41)$ | $-1.81(2.53)$ |
| Social | $-1.81(2.71)$ | $-1.22(2.59)$ |

Table 1: Means (SDs) in Study 3.

## Results and Discussion

Table 1 shows the effects of both moderators (negative scores reflecting an overall simplicity preference). First, as in Study 2, participants favored the simple explanations more strongly for deterministic than for stochastic systems $[t(388)=2.52, p=.012, d=0.26]$. Thus, the shift seen in Study 2 was not unique to unfamiliar stimuli, or specific to reasoning about diseases. Rather, it is a general pattern used across many content domains.

Second, the ordering of the means across domains was consistent with predictions. Critically, participants had a much stronger simplicity preference in the physical than in the social domain $[t(389)=8.62, p<.001, d=0.38]$. The biological and artifact domains fell in between, with the strongest preference for the physical, followed by the biological, artifact, then social domains. (Keil, Lockhart, \& Schlegel, 2010 find similar patterns in a different task.)

Together, the results of Studies 2 and 3 help to resolve the puzzle of how people could rely on a single cue-an explanation's simplicity-to do two logically independent jobs: estimating the prior and likelihood of an explanation. If contextual moderators can influence the weighting of the simplicity and complexity heuristics, then a reasoner could reach different conclusions about simplicity and complexity in different contexts, in ways which are broadly adaptive.

However, there are lingering puzzles about what determines the strength and even direction of simplicity and complexity preferences. For example, one might have expected inferences to more strongly favor the simple explanations than they did here, given the strong simplicity preferences found for the artificial items in Study 2. The more moderate inferences here may have occurred because the items were seen as more reflective of the real world-where true determinism is rareleading participants to hedge their bets. Alternatively, participants here could be recruiting background knowledge, relying more on memory rather than reasoning. In that case, the strong simplicity preferences found for artificial items in Studies 1 and 2 may better reflect the underlying reasoning processes.

## General Discussion

We set out to understand how people use simplicity to constrain their evaluation of explanations, making tractable an otherwise ill-defined computational problem. Usually, simplicity is a good cue for an explanation's
prior probability (intuitively, simple causes require fewer stars to align in order to occur) while complexity is a good cue for an explanation's likelihood or fit to the evidence (since complex causes have more opportunities to cause each aspect of the evidence). Study 1 found direct evidence for both of these opponent heuristics, directly asking about participants' priors and likelihoods.

However, our explanatory strategies must be definite enough to provide both a unique answer for a given explanatory problem, but also flexible enough to provide different answers to different problems. The opponent heuristics strategy solves this dilemma by modulating the inference depending on context. Study 2 showed that people shift toward complex explanations in stochastic contexts (because such contexts render a complexity heuristic more computationally relevant), and Study 3 showed that people favor simple explanations to varying degrees across domains, in ways that track people's general expectations about the causal textures of these domains: People believe that physical systems are more linear, whereas social systems are more subject to branching, and people correspondingly favor simple explanations to a greater degree for physical systems.

Explanatory Logic. We view these opponent heuristics as one part of a broader explanatory logic-a set of heuristics and strategies that people use for evaluating explanatory hypotheses across a variety of processes in light of our cognitive and informational limitations (Johnson, 2017). Here, we focused on causal explanation and previous work has found similar effects in visual curve-fitting (Johnson, Jin, \& Keil, 2014), both tasks requiring people to evaluate competing hypotheses (causes, trend lines) for available data (effects, data points). However, many other processes also take this form, including categorization (which category best explains the features?), theory-of-mind (which mental state best explains the behaviors?), language (which meaning best explains the utterance?), and memory (which past events best explain the details I recall?).

In ongoing work, we have been looking at simplicity heuristics in some of these other processes. For example, people can belong to several categories simultaneouslyyou can be a feminist bank teller, a Jewish woman, or a gay cognitive scientist. When explaining particular traits, people tend to favor social categorizations that invoke fewer categories simultaneously, but this bias is weaker when the categories are more loosely (i.e., stochastically) associated with the relevant features (Johnson, Kim, \& Keil, 2016). Similarly, people favor mental-state explanations that invoke relatively fewer goals to explain a particular behavior, but this simplicity preference is weaker when the goals are more stochastically associated with the behaviors (Johnson, 2017). Thus, opponent simplicity heuristics appear to pervade cognition.

The Adaptive Value of Opponent Heuristics. Our empirical argument for opponent heuristics has required us to engineer situations where people make errors.

Nonetheless, we maintain that under more ecologically realistic conditions, these heuristics often serve us well and help to make explanatory reasoning possible.

If you have a well-specified prior distribution and likelihood function, then you can do no better than normative Bayesian inference. Our participants fell short of this standard, making inferences that were unreasonably biased toward simple explanations and influenced by normatively irrelevant factors.

Yet, in the real world, we often lack access to substantial information about probability distributions. We often are confronted with novel situations in which we cannot calculate but must simply guess, based on what little we can glean from the immediate problem and what minimal cues we can bring to bear from our previous experience. It may be true that people seldom encounter cases where they must diagnose an elf, deciding among unfamiliar diseases on the basis of make-believe symptoms, but it is true in real-world medical decisionmaking that we are often faced with highly limited information. Doctors have built up a corpus of statistical knowledge about some familiar diseases, and medical scientists may have some evidence to bring to bear on less familiar ones. Yet, no one has joint probability information about all combinations of diseases and symptoms. We must rely on iffy assumptions and fallible heuristics to make any real progress, even in a highly constrained problem domain such as medical diagnosis.

In other cases, probabilities may be even less evident. When making geopolitical forecasts, assessing the reasons for a friend's odd decision, or debating philosophical conundrums, there may be little relevant prior information, and it may be impossible to model the probabilities with any degree of confidence. This is known as radical uncertainty or Knightian uncertainty (Knight, 1921), and some thinkers argue that many sources of uncertainty are not quantifiable using probabilities (e.g., Levi, 1974; Mises, 2008/1949). In cases of Knightian uncertainty, the best we can do is to adopt rules that work reasonably well most of the time, much as David Hume has argued that our inductive habits are justified by habit rather than logic (Hume, 1977/1748). The use of simplicity and other explanatory heuristics appears to be one such adaptive habit.

This is not to claim that our explanatory habits are untethered to the world. On the contrary, simplicity is usually an excellent principle to use because there are often multiple explanations, varying in complexity, which fit the data equally well. In such cases, the priors generally do favor simple explanations, so a simplicity heuristic is reasonable. But when the explanations vary in likelihood, simplicity will lead us astray, as complex explanations are often better fits to the data. Opponent heuristics allow us to harness both of these general facts to our advantage, while avoiding computations that may be intractable and, in Knightian cases, even impossible.

## References

Chater, N., \& Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? Trends in Cognitive Sciences, 7, 19-22.
Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. Nature, 407, 630-633.
Forster, M., \& Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. The British Journal for the Philosophy of Science, 45, 1-35.
Hochberg, J., \& McAlister, E. (1953). A quantitative approach to figural "goodness." Journal of Experimental Psychology, 46, 361-364.
Hume, D. (1977/1748). An enquiry concerning human understanding. Indianapolis, IN: Hackett.
Johnson, S.G.B. (2017). Cognition as sense-making. Unpublished doctoral dissertation.
Johnson, S.G.B., Jin, A., \& Keil, F.C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. In Proceedings of the 36th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Johnson, S.G.B., Kim, H.S., \& Keil, F.C. (2016). Explanatory biases in social categorization. In Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Keil, F.C., Lockhart, K.L., \& Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. JEP: General, 139, 1-15.
Kemp, C., Goodman, N.D., \& Tenenbaum, J.B. (2010). Learning to learn causal models. Cognitive Science, 34, 1185-1243.
Knight, F.H. (1921). Risk, uncertainty, and profit. Boston, MA: Hart, Schaffner, \& Marx.
Levi, I. (1974). On indeterminate probabilities. The Journal of Philosophy, 71, 391-418.
Lombrozo, T. (2007). Simplicity and probability in causal explanation. Cognitive Psychology, 55, 232-257.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. Cognitive Psychology, 61, 303-332.
Mises, L. (2008/1949). Human action: A treatise on economics. Auburn, AL: Ludwig von Mises Institute.
Pothos, E.M., \& Chater, N. (2002). A simplicity principle in unsupervised human categorization. Cognitive Science, 26, 303-343.
Shipley, E.F. (1993). Categories, hierarchies, and induction. In D.L. Medin (Ed.), The psychology of learning and motivation: Vol. 30. San Diego, CA: Academic Press.
Strickland, B., Silver, I., \& Keil, F.C. (2016). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. Memory \& Cognition.
Zemla, J.C., Sloman, S., Bechlivanidis, C., \& Lagnado, D. A. (2017). Evaluating everyday explanations. Psychonomic Bulletin \& Review.

# Principles Used to Evaluate Mathematical Explanations 

Samuel G. B. Johnson ${ }^{1}$, Angie M. Johnston ${ }^{1}$, Marissa L. Koven ${ }^{2}$, \& Frank C. Keil ${ }^{1}$<br>(sgbjohnson@gmail.com, angie.johnston@yale.edu, mlkoven@gmail.com, frank.keil@yale.edu)<br>${ }^{1}$ Department of Psychology, Yale University, New Haven, CT 06520 USA<br>${ }^{2}$ Department of Psychology, Emory University, Atlanta, GA 30322 USA


#### Abstract

Mathematics is critical for making sense of the world. Yet, little is known about how people evaluate mathematical explanations. Here, we use an explanatory reasoning task to investigate the intuitive structure of mathematics. We show that people evaluate arithmetic explanations by building mental proofs over the conceptual structure of intuitive arithmetic, evaluating those proofs using criteria similar to those of professional mathematicians. Specifically, we find that people prefer explanations consistent with the conceptual order of the operations (" $9 \div 3=3$ because $3 \times 3=9$ " rather than " $3 \times 3=9$ because $9 \div 3=3$ "), and corresponding to simpler proofs (" $9 \div 3=3$ because $3 \times 3=9$ " rather than " $9 \div 3=3$ because $3+3+3=9$ "). Implications for mathematics cognition and education are discussed.


Keywords: Mathematics cognition; philosophy of mathematics; explanation; reasoning; concepts and categories.

## Introduction

People track statistical regularities and use these regularities to make sense of the world. Some statistical learning abilities emerge early: Infants use statistics to extract complex visual features (Fiser \& Aslin, 2002) and form categories (Gómez \& Lakusta, 2004). Statistical generalizations are also critical for sense-making in higher cognition. For example, adults and children prefer simpler causal explanations in part because they have higher prior probabilities (Bonawitz \& Lombrozo, 2012; Johnson, Valenti, \& Keil, 2017; Lombrozo, 2007).

Yet, we also seem to track other truths that do not rely on statistical regularities-Platonic, logically necessary regularities such as mathematical truths. From early on, people use mathematical truths to make sense of the world: Even young infants know that if two puppets venture behind a screen, and one comes out, then only one puppet remains behind (Wynn, 1992). Without an understanding of mathematics (i.e., $2-1=1$ ), this eventand many others-would be inexplicable. Mathematical explanation grows even more essential in adulthood, as consumers must account for their spending, programmers must understand the logic of their code, and CEOs must explain their bottom line. For this reason, educators increasingly emphasize the explanatory function of mathematics (Schoenfeld, 1992). For example, the Common Core Standards (2010) state that "one hallmark of mathematical understanding is the ability to justify...why a particular mathematical statement is true or where a mathematical rule comes from" (p. 4).

But to what extent, and by what mechanisms, can
people track such mathematical regularities? Here, we claim that people use a sophisticated set of mechanisms to evaluate mathematical explanations. We argue that people (1) are sensitive to the conceptual structure of arithmetic, (2) construct mental proofs over this structure, and (3) evaluate those proofs using principles that mirror the history, philosophy, and practice of mathematics.
Just as there are intricate connections among concepts in physics and biology, so are mathematical concepts richly structured (Whitehead \& Russell, 1910). For example, geometric facts are grounded in facts about analysis (Bolzano, 1817), and arithmetic facts in set theory (Frege, 1974/1884). More basically, subtraction can be viewed as grounded in addition, multiplication in addition, division in multiplication, and so on (Figure 1; see also Dedekind, 1995/1888; Tao, 2016). Although these concepts need not be viewed asymmetrically, these asymmetries may be psychologically natural. For example, people may follow the principle that more fundamental operations begin with small things and assemble larger things, rather than vice versa. This would make addition more fundamental than subtraction (which breaks larger things into smaller pieces).

We explored the intuitive conceptual structure of mathematics using a simple method-asking people to evaluate mathematical explanations. Consider the explanation " $9 \div 3=3$ because $3+3+3=9$." In one sense, this is a terrible explanation because it is tautological-both facts are necessarily true and logically equivalent. However, we propose that people are willing to evaluate explanations of this sort, and do so as if constructing a mental proof of the explanatory target (here, " $9 \div 3=3$ ") from the putative explanation (" $3+3+3=9$ "), over the conceptual structure in Figure 1. For example, to evaluate " $9 \div 3=3$ because $3+3+3=9$," one would first derive a multiplication fact (" $3 \times 3=9$ ") from the addition fact, and then derive the division fact from that intermediate multiplication fact. We test two principles that people might use for evaluating implicit mental proofs.

First, people may be sensitive to the asymmetric nature of mathematical explanation (Bolzano, 1817; Kitcher, 1975). For example, consider the explanation " $4-2=2$ because $2+2=4$." Although tautological, if this explanation respects a perceived conceptual order, it may be seen as superior to an explanation that does not, such as " $2+2=4$ because $4-2=2$." That is, a statement may be explained in terms of a logically equivalent statement, if that explanation helps to highlight the more conceptually primitive facts grounding it.


Figure 1: Proposed intuitive structure of arithmetic.
Note. Numbers correspond to the proof rules in Table 1, with forward proof rules flowing in the direction of the arrows and reverse proof rules flowing against the direction of the arrows.

| Rule | Input | Output |
| :---: | :---: | :---: |
| Addition/Subtraction Conversion |  |  |
| 1F | $X+Y=Z$ | $Z-X=Y$ |
| 1R | $Z-X=Y$ | $X+Y=Z$ |
| Addition/Multiplication Conversion |  |  |
| 2F | $\Sigma_{Y} X=Z$ | $X \times Y=Z$ |
| 2R | $X \times Y=Z$ | $\Sigma_{Y} X=Z$ |
| Multiplication/Division Conversion |  |  |
| 3F | $X \times Y=Z$ | $Z \div X=Y$ |
| 3R | $Z \div X=Y$ | $X \times Y=Z$ |
| Multiplication/Exponent Conversion |  |  |
| 4F | $X \times X=Z$ | $X^{2}=Z$ |
| 4R | $X^{2}=Z$ | $X \times X=Z$ |
| Exponent/Root Conversion |  |  |
| 5F | $X^{2}=Z$ | $\checkmark Z=X$ |
| 5R | $\sqrt{ } \mathrm{Z}=X$ | $X^{2}=Z$ |

Table 1: Hypothesized rules for mental proofs.
Second, people may prefer explanations that involve fewer steps because such proofs more readily confer understanding (Descartes, 1954/1684; Hardy, 2004/1940; Kitcher, 1983) and are less prone to error (Hume, 1978/1738). For example, " $9 \div 3=3$ because $3+3+3=9$ " might be seen as a worse explanation than " $9 \div 3=3$ because $3 \times 3=9$," since the proof for the former explanation requires two steps (addition to multiplication, multiplication to division) whereas the latter requires only one step (multiplication to division), even though both proofs proceed in the same conceptual order. We test whether people scale their explanatory judgments to proof complexity. If so, this would be evidence not only that people use complexity as a criterion to judge explanatory
quality, but also that people spontaneously construct proofs over the conceptual structure depicted in Figure 1.

Our model assumes that people evaluate these explanations by constructing and evaluating a proof of the explanatory target from the base, using the transformation rules given in Table 1. These correspond to the forward ( F ) and reverse (R) version of each arrow in Figure 1 (see Rips, 1983 for a related idea in propositional reasoning). Proofs are evaluated by assuming a rule cost is incurred for applying each rule, and that the total proof cost is the sum of the costs of the individual rules invoked in the proof. If people are sensitive to proof complexity, then they should prefer proofs with smaller costs. To capture the idea that people prefer explanations consonant with the conceptual order, our model allows forward and reverse rules to have different costs: We predict that reverse rules carry a higher cost than forward rules. That is, a proof has a higher cost to the extent that it uses more rules in general, and more reverse rules in particular. Equivalently, short proofs flowing with Figure 1's arrows would correspond to better explanations than long proofs flowing against the arrows (see examples below).

## Method

We recruited 97 participants from Amazon Mechanical Turk in exchange for a small payment ( $50.5 \%$ female, $\left.M_{\text {age }}=34.0\right)$. Participants were excluded from data analysis if they gave inappropriate answers to the check questions ( $N=6$; see below for details).

Participants rated a series of 30 mathematical explanations. For each explanation, participants were asked "How satisfying do you find this explanation?" on a scale from 0 ("not at all satisfying") to 10 ("very satisfying"). These explanations consisted of all possible pairings of addition, subtraction, multiplication, division, exponent, and root operations, where the constituents were 3 s ; examples are given in Table 2 in the Appendix. For example, across different blocks, participants completed a pair of multiplication/exponent items:

$$
\begin{aligned}
& 3^{2}=9 \text { because } 3 \times 3=9[\text { forward }] \\
& 3 \times 3=9 \text { because } 3^{2}=9[\text { reverse }]
\end{aligned}
$$

Because there are 15 ways of pairing these 6 operations with each other, and two orders (forward and reverse), participants completed a total of 30 items. The forward and reverse items were presented in separate blocks, with the order of the items randomized within each block. The order of the blocks was also randomized.

Check questions were included after the test questions to detect participants who were responding randomly. These always included two items for which one of the equations was false (e.g., " $4+3=7$ because $4+3=2$ " or " $743+259=1,002$ because $743+259=713 "$ ) and two items for which the numbers differed between the two equations (e.g., " $26 \times 47=1222$ because $678-234=444 "$ ). Participants with average answers to these questions that were above the scale midpoint were excluded from data analysis.

## Results

Participants were sensitive to both criteria of conceptual order and proof complexity. We first describe the results relative to the qualitative predictions of the model in order to explain how the model works, and then assess the quantitative fit at both the group and individual levels.

Qualitative model predictions. We anticipated that people would penalize explanations to the extent that the most direct proof requires applying a large number of rules (see Tables 1 and 2), and that application of 'reverse' rules would correspond to a greater penalty. These two principles are captured by (1) computing the shortest distance between the two operations in Figure 1, and (2) penalizing the explanation for each arrow along that shortest path, with arrows in the 'reverse' direction receiving a larger penalty (we call this penalty $R$ ) than arrows in the 'forward' direction (a smaller penalty of $F$ ).

For example, consider explaining a root formula in terms of division (e.g., " $\sqrt{ } 9=3$ because $9 \div 3=3 "$ ). According to our model, people would rate this explanation by producing a mental proof of ' $\sqrt{ } 9=3$ ' from $' 9 \div 3=3$ '. As noted in Table 2, this requires the application of three rules: 3 R (to derive ' $3 \times 3=9$ ' from ' $9 \div 3=3$ '), 4 F (to derive ' $3^{2}=9$ '), and finally 5 F (to derive ' $\sqrt{ } 9=3$ '). Two of these rules are forward and one is backward, so the total penalty is $2 F+1 R$-since this is a relatively high penalty, we would expect this explanation to be rated poorly. In contrast, addition would be seen as an excellent explanation of subtraction, because a subtraction formula (e.g., ' $9-3-3=3$ ') can be derived from addition (' $3+3+3=9$ ') using only one forward rule ( 1 F ), leading to a penalty of only $1 F$. The penalty scores for several of the explanations are given in Table 2 in the Appendix, along with the rules required to perform these proofs.

This model captures several patterns in the means (Table 3 in the Appendix). First, for each operation, we can consider which explanation was rated highest (i.e., the highest mean in each row of Table 3). For the addition operation, which is not conceptually dependent on any of the other operations, its highest rated explanations were subtraction and multiplication-the closest downstream operations. For both subtraction and multiplication, addition is the highest rated explanation, consistent with the topology of Figure 1, wherein both operations depend directly on addition. Similarly, for explaining division and exponentiation, multiplication is highest rated, consistent with Figure 1, in that both operations depend directly on multiplication. Finally, for roots, exponentiation was seen as the best explanation, again consistent with the direct dependence of roots on exponents.

More generally, our model predicts a central role of multiplication and a peripheral role of subtraction. As Figure 1 shows, multiplication is a central node in the conceptual structure of arithmetic-most roads lead to (or from) multiplication-but subtraction is on the periphery. This prediction is borne out by the data. Multiplication is both the most easily explained operation (i.e., the highest
mean in the rightmost column of Table 3) and the operation that explains the most (i.e., the highest mean in the bottom row of Table 3). In contrast, subtraction is least easily explained and explains the least.

Group-level model fitting. We model the results in terms of the sum of the rule costs, shown in Table 2. This analysis assumes that the cost of each rule is determined only by whether it is a forward or reverse rule. Thus, one free parameter $R / F$ is used, reflecting the extent to which $R$ rules were penalized more heavily than $F$ rules.

We modeled the explanation ratings in terms of the summed rule costs, where only the $R / F$ parameter was free to vary. These scores were good predictors of the explanation ratings, $r(28)=-.86, p<.001$. The best fitting value for the $R / F$ parameter was 1.18 , indicating that the explanatory cost of applying reverse rules that go against the conceptual grain of mathematics is $18 \%$ higher than the explanatory cost of applying forward rules. This supports our conjecture that forward explanations (e.g., explaining subtraction in terms of addition) are preferred to their logically equivalent reverse explanations (e.g., explaining addition in terms of subtraction).

This asymmetry between forward and reverse rules is also evident from looking at the means in Table 3. For example, explanations of subtraction in terms of addition were rated more satisfying than explanations of addition in subtraction, since the former grounds an operation in a more psychologically basic operation whereas the latter does the opposite. Since there are five rules in Table 1, there are five directly reversible pairs, as well as four pairs of operations (addition/division, addition/ exponentiation, addition/root, and multiplication/root) that are connected by applying two or more rules in the same direction (see Proof column in Table 2). Averaging across these pairs, the forward explanations were seen as more satisfying than the reverse, $t(90)=3.90, p<.001$.

Individual-level model fitting. Our model also captures individual participants' explanatory judgments. To test the proof complexity factor, we calculated, for each participant, the correlation between the explanatory judgment for each of the 30 items and the number of rules required for that item's proof (i.e., the sum of the $F$ and $R$ columns in Table 2). This parameter-free model captured a substantial amount of the variance within each participant's response pattern, with a mean correlation of -. 46 between number of rules and explanatory judgment (Fisher-transformed to a $z$-score before averaging, and inverse-transformed back to a correlation). Furthermore, almost all participants ( $95.6 \%$ ) had a negative correlation, demonstrating that the excellent model fit at the group level is not due to a small subset of participants, but instead generalizes across almost all participants.

Although this parameter-free model is useful in showing that considerable within-subject variability can be explained via proof complexity, it is less useful for testing asymmetries between the forward and backward rules, since this requires estimating the relative penalties
associated with each rule. To do so, we conducted a linear regression for each participant, using ten dummy-coded variables to represent whether each of the ten rules figures in each item's proof. For example, for the item explaining division in terms of subtraction, the dummy variables for rules $1 \mathrm{R}, 2 \mathrm{~F}$, and 3 F were set to 1 , and all others set to 0 . For each participant, we calculated the regression weights for each rule, reflecting the relative penalty associated with each rule (thus, all regression weights would be expected to be negative), and these weights (averaged across participants) are depicted in Figure 2.


Figure 2: Regression coefficients on each rule.
Note. These coefficients represent the explanatory 'cost' of a given rule appearing in the proof of the explanation. Error bars represent $95 \%$ confidence intervals, calculated over participants.

Mirroring the group-level findings, Figure 2 reveals higher costs for reverse rules than for forward rules, leading to more negative regression coefficients for the reverse rules. This was true for rules 1 F and $1 \mathrm{R}(95 \% \mathrm{CI}$ [ $0.28,1.15]$ for the difference in regression coefficients), rules 3 F and $3 \mathrm{R}(95 \% \mathrm{CI}[0.13,0.81])$, rules 4 F and 4 R ( $95 \%$ CI $[-0.02,0.58]$; marginally significant), and rules 5 F and $5 \mathrm{R}(95 \% \mathrm{CI}[0.17,0.85])$. This difference was not significant for the addition/multiplication rules 2 F and 2 R ( $95 \%$ CI $[-0.41,0.39]$ ), perhaps because repeated addition of the same addends is uncommon except in the context of multiplication. Overall, these findings are consistent with the best-fitting value of the $R / F$ parameter of 1.18 in the group-level analysis, indicating a higher explanatory cost for reverse rules than for forward rules.

## Discussion

Mathematical knowledge is critical for explaining patterns in both the physical and symbolic worlds, and for building an understanding conceptually dependent mathematical facts. Here, we proposed that people evaluate mathematical explanations (e.g., " $9-3-3=3$ because $3+3+3=9$ ") by building a proof from the explanatory base (" $3+3+3=9$ ") to the explanatory target (" $9-3-3=3$ ") using a set of transformation rules (e.g., deriving subtraction from addition). Supporting this idea, participants
preferred explanations that obeyed the conceptual order of mathematics and which required fewer derivational steps.

Where might these intuitions come from? One possibility is that they are rooted in a more basic understanding of the natural numbers (e.g., Carey, 2009; Dehaene, 1997; Rips, Bloomfield, \& Asmuth, 2008) that begins to emerge early in development. For example, addition and subtraction are intimately related to counting, both in development (Rips et al., 2008) and in mathematics (Tao, 2016). This is because the natural numbers are constructed by using the successor function (e.g., 9 is the successor to 8 ). Such psychologically and mathematically primitive mechanisms may underlie lateremerging explanatory intuitions.

Alternatively, could it be possible that people simply parroted explanations as introduced in school? This possibility is unlikely for two reasons. First, multiplication was strongly preferred over subtraction as an explanation. This pattern is consistent with our claims about conceptual structure but conflicts with this alternative account, since subtraction is typically learned before multiplication. Second, we doubt most people have ever heard (for example) division explained in terms of addition, exponential, roots, etc., so differences across these explanations must be due to a chaining mechanism of the type we proposed.

Might analogous results hold beyond arithmetic explanations? Indeed, people have a rich intuitive understanding of other mathematical domains such as geometry (Dillon, Huang, \& Spelke, 2013), suggesting that people have intuitive theories of Platonic regularities across a variety of domains. Moreover, the proof construction and evaluation principles may be the same used in more general deductive reasoning processes (Rips, 1994; but see Johnson-Laird \& Byrne, 1991), in which case our method may generalize. Our studies focused on simple arithmetic operations (e.g., Ashcraft, 1992), but future work could extend this inquiry to other areas of mathematics (such as geometry), populations (such as children or expert mathematicians), or domains (such as dependencies among physics concepts or among mental states) to further map our intuitive theories.

The ontological implications of this work are within the domain of philosophy. For now, we merely contrast two possible views. According to the dominant Platonist view (e.g., Frege, 1974/1884), mathematical truths are 'out there' in the world. On the Platonist view, our results reflect aspects of mathematical structure that have been internalized from the world. However, others with Kantian views argue that mathematical cognition reflects structure imposed on the world by our minds rather than anything intrinsic in the world (Kant, 1998/1781; Mill, 2002/1843; see Lakoff \& Núñez, 2000). On the Kantian view, our results reflect the intrinsic structure of our minds themselves, which we impose on the world.

As for the instructional implications of these findings, we believe mathematics educators are best-positioned to
make the assessment. However, we do make some tentative suggestions. First, mathematical proof may not be intrinsically unintuitive-it may instead be the level of abstraction of many proofs that masks intuitive understanding. If so, introducing simple deductive proofs of simple arithmetic relationships at an earlier educational stage could lay an intuitive foundation for more formal proofs later on (see Carpenter, Franke, \& Levi, 2003). Second, people use the conceptual structure of mathematics to understand derivative concepts in terms of more basic ones. Educators may wish to emphasize these abstract connections, in conjunction with more concrete applications, in order to tap into this intuitive understanding; for example, explaining division both as a way to divide resources and as the inverse of multiplication. Finally, our methodology might be used to assess the explanatory trade-offs between different kinds of examples. Studying explanatory preferences in adults may provide a simple laboratory for testing out explanatory methods that might be used in educational settings, prior to undertaking expensive and risky intervention studies. This method could be used not only to illuminate mathematical understanding, but also the conceptual structure of other domains.

## References

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. Cognition, 44, 75-106.
Bolzano, B. (1817). Purely analytic proof of the theorem that between any two values which give results of opposite sign there lies at last one real root of the equation. (Trans. S. Russ.)
Bonawitz, E. B., \& Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. Developmental Psychology, 48, 1156-1164.
Carey, S. (2009). The origin of concepts. Oxford, UK: Oxford University Press.
Carpenter, T. P., Franke, M. L., \& Levi, L. (2003). Thinking mathematically: Integrating arithmetic and algebra in elementary education. Portsmouth, NH: Heinemann.
Common Core State Standards Inititative. (2010). Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.
Dehaene, S. (1997). The number sense: How the mind creates mathematics. Oxford, UK: Oxford University Press.
Dedekind, R. (1995). What are numbers and what should they be? (H. Pogorzelski, W. Ryan, \& W. Snyder, Trans. and Ed.). Orono, ME: Research Institute for Mathematics. (Original work published 1888.)
Descartes, R. (1954/1684). Rules for the direction of the mind. In E. Anscombe \& P. T. Geach (Trans.), Descartes: Philosophical Writings. London: Pearson.
Dillon, M. R., Huang, Y., \& Spelke, E. S. (2013). Core
foundations of abstract geometry. Proceedings of the National Academy of Sciences, 110, 14191-14195.
Fiser, J., \& Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences, 99, 15822-15826.
Frege, G. (1974/1884). The foundations of arithmetic: A logico-mathematical enquiry into the concept of number. (J. L. Austin, Trans.). Oxford, UK: Blackwell.
Gómez, R. L., \& Lakusta, L. (2004). A first step in formbased category abstraction by 12-month-old infants. Developmental Science, 7, 567-580.
Hardy, G. H. (2004/1940). A mathematician's apology. Cambridge, UK: Cambridge University Press.
Hume, D. (1978/1738). A treatise on human nature. Oxford, UK: Oxford University Press.
Johnson, S.G.B., Valenti, J.J., \& Keil, F.C. (2017). Opponent uses of simplicity and complexity in causal explanation. Proceedings of the 39th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Johnson-Laird, P. N., \& Byrne, R. M. J. (1991). Deduction: Essays in cognitive psychology. Hillsdale, NJ: Erlbaum.
Kant, I. (1998/1781). Critique of pure reason. (P. Guyer, \& A. W. Wood, Trans. and Ed.). Cambridge, UK: Cambridge University Press.
Kitcher, P. (1975). Bolzano's ideal of algebraic analysis. Studies in the History and Philosophy of Science, 6, 229-269.
Kitcher, P. (1983). The nature of mathematical knowledge. Oxford, UK: Oxford University Press.
Lakoff, G., \& Núñez, R. (2000). Where mathematics comes from. New York, NY: Basic Books.
Lombrozo, T. (2007). Simplicity and probability in causal explanation. Cognitive Psychology, 55, 232-257.
Mill, J. S. (2002/1843). A system of logic, ratiocinative and inductive. Honolulu, HI: University Press of the Pacific.
Rips, L. J. (1983). Cognitive processes in propositional reasoning. Psychological Review, 90, 38-71.
Rips, L. J. (1994). The psychology of proof. Cambridge, MA: MIT Press.
Rips, L. J., Bloomfield, A., \& Asmuth, J. (2008). From numerical concepts to concepts of number. Behavioral and Brain Sciences, 31, 623-687.
Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. Grouws (Ed.), Handbook for research on mathematics teaching and learning (pp. 334-370). New York, NY: Macmillan.
Tao, T. (2016). Analysis I (3rd Ed.). Berlin: Springer.
Whitehead, A. N., \& Russell, B. (1910). Principia mathematica (Vol. 1). Cambridge, UK: Cambridge University Press.
Wynn, K. (1992). Addition and subtraction by human infants. Nature, 358, 749-750.

## Appendix

| Operation Explained | Operation Used to Explain | Stimuli | Proof | F | $\boldsymbol{R}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Addition | Subtraction | $3+3+3=9$ because $9-3-3=3$ | 1R | 0 | 1 |
|  | Multiplication | $3+3+3=9$ because $3 \times 3=9$ | 2R | 0 | 1 |
|  | Division | $3+3+3=9$ because $9 \div 3=3$ | 3R, 2R | 0 | 2 |
|  | Exponent | $3+3+3=9$ because $3^{2}=9$ | 4R, 2R | 0 | 2 |
|  | Root | $3+3+3=9$ because $\sqrt{ } 9=3$ | 5R, 4R, 2R | 0 | 3 |
| Subtraction | Addition | $9-3-3=3$ because $3+3+3=9$ | 1 F | 1 | 0 |
|  | Multiplication | $9-3-3=3$ because $3 \times 3=9$ | 2R, 1F | 1 | 1 |
|  | Division | $9-3-3=3$ because $9 \div 3=3$ | 3R, 2R, 1F | 1 | 2 |
|  | Exponent | $9-3-3=3$ because $3^{2}=9$ | 4R, 2R, 1F | 1 | 2 |
|  | Root | $9-3-3=3$ because $\sqrt{ } 9=3$ | 5R, 4R, 2R, 1F | 1 | 3 |
| Multiplication | Addition | $3 \times 3=9$ because $3+3+3=9$ | 2F | 1 | 0 |
|  | Subtraction | $3 \times 3=9$ because $9-3-3=3$ | 1R, 2F | 1 | 1 |
|  | Division | $3 \times 3=9$ because $9 \div 3=3$ | 3R | 0 | 1 |
|  | Exponent | $3 \times 3=9$ because $3^{2}=9$ | 4R | 0 | 1 |
|  | Root | $3 \times 3=9$ because $\sqrt{ } 9=3$ | 5R, 4R | 0 | 2 |
| Division | Addition | $9 \div 3=3$ because $3+3+3=9$ | 2F, 3F | 2 | 0 |
|  | Subtraction | $9 \div 3=3$ because $9-3-3=3$ | 1R, 2F, 3F | 2 | 1 |
|  | Multiplication | $9 \div 3=3$ because $3 \times 3=9$ | 3F | 1 | 0 |
|  | Exponent | $9 \div 3=3$ because $3^{2}=9$ | 4R, 3F | 1 | 1 |
|  | Root | $9 \div 3=3$ because $\sqrt{ } 9=3$ | 5R, 4R, 3F | 1 | 2 |
| Exponent | Addition | $3^{2}=9$ because $3+3+3=9$ | 4F, 2F | 2 | 0 |
|  | Subtraction | $3^{2}=9$ because $9-3-3=3$ | 4F, 2F, 1R | 2 | 1 |
|  | Multiplication | $3^{2}=9$ because $3 \times 3=9$ | 4 F | 1 | 0 |
|  | Division | $3^{2}=9$ because $9 \div 3=3$ | 4F, 3R | 1 | 1 |
|  | Root | $3^{2}=9$ because $\sqrt{ } 9=3$ | 5R | 0 | 1 |
| Root | Addition | $\sqrt{ } 9=3$ because $3+3+3=9$ | 5F, 4F, 2F | 3 | 0 |
|  | Subtraction | $\sqrt{ } 9=3$ because $9-3-3=3$ | 5F, 4F, 2F, 1R | 3 | 1 |
|  | Multiplication | $\sqrt{ } 9=3$ because $3 \times 3=9$ | 5F, 4F | 2 | 0 |
|  | Division | $\sqrt{ } 9=3$ because $9 \div 3=3$ | 5F, 4F, 3R | 2 | 1 |
|  | Exponent | $\sqrt{ } 9=3$ because $3^{2}=9$ | 5F | 1 | 0 |

Table 2: Mental proofs and penalty scores for all explanations.

|  |  | Operation Used to Explain |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Addition | Subtraction | Multiplication | Division | Exponent | Root | Average |
| $\begin{aligned} & \text { D } \\ & \text { 品 } \\ & \text { x } \\ & 0 \\ & 0 \\ & 0 \\ & 0 \\ & 0 \\ & 0 \end{aligned}$ | Addition | - | 6.37 | 7.96 | 5.70 | 5.69 | 4.47 | 6.04 |
|  | Subtraction | 7.14 | - | 4.45 | 4.93 | 4.01 | 3.82 | 4.87 |
|  | Multiplication | 8.11 | 4.20 | - | 7.12 | 8.02 | 6.01 | 6.69 |
|  | Division | 5.76 | 4.65 | 7.46 | - | 5.70 | 5.45 | 5.80 |
|  | Exponent | 6.12 | 3.65 | 8.75 | 5.26 | - | 6.49 | 6.05 |
|  | Root | 5.27 | 3.43 | 7.11 | 5.45 | 7.44 | - | 5.74 |
|  | Average | 6.48 | 4.46 | 7.15 | 5.69 | 6.17 | 5.25 |  |

Table 3: Explanatory ratings for each pair of operations.

# Statistical and Mechanistic Information in Evaluating Causal Claims 

Samuel G. B. Johnson \& Frank C. Keil<br>(sgbjohnson@gmail.com, frank.kei@ @yale.edu)<br>Department of Psychology, Yale University, New Haven, CT 06520 USA


#### Abstract

People use a variety of strategies for evaluating causal claims, including mechanistic strategies (seeking a step-bystep explanation for how a cause would bring about its effect) and statistical strategies (examining patterns of cooccurrence). Two studies examine factors leading one or the other of these strategies to predominate. First, general causal claims (e.g., "Smoking causes cancer") are evaluated predominantly using statistical evidence, whereas statistics is less preferred for specific claims (e.g., "Smoking caused Jack's cancer"). Second, social and biological causal claims are evaluated primarily through statistical evidence, whereas statistical evidence is deemed less relevant for evaluating physical causal claims. We argue for a pluralistic view of causal learning on which a multiplicity of causal concepts lead to distinct strategies for learning about causation.


Keywords: Causal reasoning; concepts and categories; information evaluation; statistical reasoning.

## Introduction

Causal knowledge is crucial for understanding and controlling the world, and strategies for evaluating causal claims are central to gatekeeping that crucial knowledge. Humans seem especially prone to two strategies-a mechanism strategy, on which we consider potential mediating causal links as evidence favoring a causal connection; and a statistical strategy, on which we look for correlations between a cause and effect. For example, Jack is assessing the risk that smoking causes cancer. He can assess this claim mechanistically by considering the plausibility of potential mediating mechanisms that explain relationship between smoking and cancer. Or he can assess the claim statistically by observing whether the frequency of cancer is higher in a population that smokes compared to a population that does not.
There is ample evidence that people use both of these strategies, though different theoretical approaches to causal cognition emphasize different types of information.

According to mechanism-based approaches to causal cognition, we learn about causal relations primarily by searching for generative mechanisms through which causes can produce their effects. Several convergent lines of evidence are consistent with causal relations being represented in terms of underlying mechanisms (see Johnson \& Ahn, in press). Knowledge of underlying mechanisms affects whether discounting or conjunction effects occur in causal attribution (Ahn \& Bailenson, 1996), whether the Markov principle is applied to causal networks (Park \& Sloman, 2013), and whether causal chains are judged to be transitive (Johnson \& Ahn, 2015).
If we learn about causation by searching for plausible
mechanisms, then people would seek out evidence of underlying mechanisms when determining whether one thing causes another. Indeed, people do sometimes assess causal hypotheses by forming a mechanistic narrative that would lead from $X$ to $Y$ and assessing the plausibility of that narrative (e.g., Fernbach, Darlow, \& Sloman, 2011; Kahneman \& Tversky, 1982; Taleb, 2007). For example, Jack might imagine some physiological mechanism by which smoking and lung cancer could be connected, then evaluate the plausibility of these steps. Moreover, people prefer mechanism evidence overwhelmingly in causal attribution - that is, in determining which cause to assign to an effect (Ahn, Kalish, Medin, \& Gelman, 1995).

In contrast, statistics-based approaches to causal learning emphasize the role of statistical knowledge in inferring causal relationships. These theories hold that causal relationships are primarily discovered through information about the co-occurrence of the cause and effect, although individual theories differ in the details of how these inferences work (e.g., Cheng, 1997; Gopnik et al., 2004; Griffiths \& Tenenbaum, 2005). These theories do not necessarily claim that causal relations are represented in terms of statistical patterns, but often hold that causal relations are represented in terms of abstract causal powers underlying the connection between cause and effect, which are then inferred through statistical means (Cheng, 1997; Pearl, 2000). Nonetheless, statistical approaches do claim that causal relations are primarily learned through co-occurrence information, and there is abundant evidence that people are often able to learn from statistical evidence (e.g., Gopnik et al., 2004; Steyvers, Tenenbaum, Wagenmakers, \& Blum, 2003).

Moreover, statistical evidence could be an antidote to the shallowness of people's knowledge of causal mechanisms: Even though people do use mechanism knowledge in evaluating causal relationships when it is available, we do not seem to have extensive knowledge of mechanisms. People greatly overestimate their knowledge of how everyday devices such as flush toilets work, revealing misconceptions and pervasive gaps in understanding (Rozenblit \& Keil, 2002). People's beliefs in mechanisms underlying causal relationships are more likely to take the form of generic or highly unspecified 'placeholders', akin to our beliefs in abstract category essences (Medin \& Ortony, 1989). Such skeletal representations are difficult to square with a strong mechanism view on which people seek a detailed understanding of how causal relationships work and use that understanding to guide inference, but they might seem to be more consistent with statistical approaches to causal thinking on which covariation is used to infer the
existence of abstract underlying mechanisms without being committed to particular mechanistic details.

The mechanism and the statistical approaches, however, need not be in conflict and can be mutually compatible with a third approach known as causal pluralism (Cartwright, 2004; Danks, 2005; Hitchcock, 2003; Lombrozo, 2010; Waldmann \& Mayrhofer, 2016). According to this approach, people might use a multiplicity of causal concepts and a concordant variety of learning strategies in systematic, context-dependent ways. Some prima facie support for the pluralistic position comes from experiments where people used mechanism and statistical evidence in an interactive manner (Fugelsang \& Thompson, 2000; Spellman, 1996).

Yet, little is known about contextual factors that lead each type of evidence to predominate. Here, we look at two dimensions along which causal relations can vary-in its level of abstraction and its domain. Because people seem to use different sorts of causal concepts for representing these relations, we anticipated that people may also use different strategies to learn about these relations. If you need to decide whether something is a banana, the best question to ask would be about its shape, whereas if you need to decide whether something is a peach, the best question would be about its texture. And just as we must consult our concept of 'banana' when deciding whether something is a banana and our concept of 'peach' when deciding whether something is a peach, we must consult our concept of 'cause' when deciding whether a relationship is causal. When we deploy different causal concepts across contexts, this can lead to different learning strategies.
General and Specific Causation. General causal claims refer to generic causal patterns ("Smoking causes a person to get lung cancer"), whereas specific claims refer to concrete occasions when a pattern was instantiated ("Smoking caused Jack to get lung cancer"). The inferences supported by general and particular claims differ in several ways. General claims are associated with more essentialist inferences (Cimpian \& Erickson, 2012) and, in the domain of human behavior, with more neuroscientific rather than psychosocial explanations (Kim, Ahn, Johnson, \& Knobe, 2016). Might these claims also differ in the evidence used for their evaluation?

General claims refer to an entire category of causal relationships (i.e., a set of event pairs), whereas specific claims refer to an instance of that category (one single event pair in that set). Thus, general claims necessarily quantify over multiple instances and intrinsically carry statistical content, whereas specific claims do not. We suggest that this conceptual difference could lead statistical evidence to be privileged more for general rather than specific causal claims.
This pattern of evidence preferences can lead to nonnormative behavior. If we are not privy to the particulars of Jack's case, the only strategy for evaluating a tokencausal claim will be to look for a more general causal
pattern between smoking and cancer-to evaluate the general claim. Thus, evidence relevant for evaluating the general claim would be equally relevant for evaluating the specific claim. Imagine that a tobacco company is being sued under one of two different circumstances: (1) a class action suit (the plaintiffs' lawyers arguing that "Smoking causes a person to get lung cancer"), or (2) Jack's single party action (his lawyers arguing that "Smoking caused Jack to get lung cancer"). In both cases, jurors might be confronted with mechanism evidence, such as a biologist's testimony concerning biochemical mechanisms, or with statistical evidence, such as an epidemiologist's testimony comparing cancer rates across populations. It seems difficult to justify a difference between these two cases in jurors' relative weighing of mechanistic and statistical testimony. Yet, if people rely on different processes for evaluating general and specific claims, then the jurors may well behave differently.

Causation across Domains. People use different intuitive theories of causality across domains. Whereas physical causation is typically conceptualized in terms of force propagating down branching causal chains, social and biological causation are thought of as webs of interconnected influences. People tend to identify physical events as having one cause but many effects, whereas social events are seen as having many causes and many effects (Strickland, Silver, \& Keil, 2016; see also Johnson, Valenti, \& Keil, 2017). Likewise, even young children seem to view biological systems as causally interacting parts in homeostatic balance (Keil, 1989). Thus, the simple, linear causal pathways thought to be at play in the physical world give way to more complex causal structures in the social and biological domains.

Similarly, social (e.g., psychological or economic) causation is often goal-directed (Lombrozo \& Carey, 2006) or equipotential (Heider, 1958)-the same ends can often be brought about through many different means. For this reason, people's causal theories of social (and likely biological) systems often focus on counterfactual dependence (Lombrozo, 2010), whereas their theories of physical systems are characterized more by ideas about physical force and transference of conserved quantities.

Given the relatively linear and force-based conceptions of physical causation, and the relatively web-like and dependence-based conceptions of causality in biological and social causation, people may use a more deterministic concept of physical causation and a more stochastic concept of social and biological causation (Johnson, Valenti, \& Keil, 2017). Thus, people may rely more on mechanistic strategies when learning about physical systems and more on statistical strategies when learning about biological and social systems.

Overview of Studies. Two studies test differences in evidence-seeking between general and specific causal claims, with the studies differing in the framing of the claims. After testing this hypothesis about general versus specific causation, we present an analysis of evidence-
seeking preferences across domains, aggregating across studies. In the General Discussion, we assess the prospects for a pluralistic view of causal learning.

## Experiment 1

In Experiment 1, we tested what kind of information people thought most relevant for assessing general causal claims (e.g., "Eating polar bear liver causes a person to become dizzy") and specific causal claims ("Eating polar bear liver caused Bill to become dizzy"). The mechanism view holds that we learn about causal relationships primarily by searching for underlying mechanisms, leading to a preference for mechanism evidence, whereas the statistical view holds that we learn about causal relationships primarily through contingency information, leading to a preference for statistical evidence. In contrast to both positions, we predicted that, whatever people's baseline preferences for one or the other type of evidence, the preference for statistical evidence would be stronger when evaluating general rather than specific claims.

## Method

We recruited 80 participants from Mechanical Turk, and excluded 5 from data analysis because they incorrectly answered more than $33 \%$ of the check questions.

Participants saw either the general or specific version of each of 24 causal claims, presented in a box. For each item, participants were asked "Which of the following types of evidence would be most helpful to you in determining whether the statement in the box is true?" as a forced-choice. For the polar bear item, the options read:

Statistical: "Measurements of the frequency of dizziness of many people after they eat or do not eat polar bear liver."
Mechanism: "An explanation of why eating polar bear liver would cause a person to become dizzy."
Anecdotal: "Knowing whether there is another occasion on which a person ate polar bear liver and then they felt dizzy."
We assumed that few people would choose the weak anecdotal evidence, and used this option to assess the degree to which participants used poor causal reasoning. The order of the options was randomized for each item, and the items were presented in a random order.

## Results and Discussion

As shown in Table 1, statistical evidence was chosen more frequently when evaluating general compared to specific claims. Due to non-normality, Mann-Whitney $U$ tests were used to compare the number of items for which participants chose each evidence type in each condition.

These tests showed that statistical evidence was chosen for more items when evaluating general claims than when evaluating specific claims [ $U=496.5, p=.028, r=.25$ ]. This corresponded to relatively fewer mechanism responses for the general claims than for the specific claims and fewer anecdotal responses for the general
claims than for the specific claims. Thus, responses shifted relatively more toward statistical evidence for the general than for the specific claims.

This result indicates that people use pluralistic causal learning strategies. Specifically, it appears that the conceptual differences between general and specific claims had downstream consequences for evidenceseeking preferences: Because general claims quantify over instances, statistical evidence is seen as more relevant to evaluating such claims, compared to specific claims, and mechanism evidence is seen as less relevant.

Table 1: Results of Experiments 1 and 2

|  | Statistical | Mechanism | Anecdotal |
| :--- | :---: | :---: | :---: |
| Exp. 1 |  |  |  |
| General | $55.3 \%$ | $36.8 \%$ | $7.9 \%$ |
| Specific | $41.6 \%$ | $47.5 \%$ | $11.0 \%$ |
| Exp. 2 |  |  |  |
| General | $62.0 \%$ | $30.4 \%$ | $7.6 \%$ |
| Specific | $47.6 \%$ | $41.4 \%$ | $11.0 \%$ |
| Domain Analysis |  |  |  |
| Physical | $47.5 \%$ | $46.6 \%$ | $5.9 \%$ |
| Biological | $53.6 \%$ | $38.9 \%$ | $7.5 \%$ |
| Psychological | $51.6 \%$ | $36.0 \%$ | $12.4 \%$ |
| Economic | $53.7 \%$ | $34.7 \%$ | $11.6 \%$ |

Note. Entries indicate the proportion of choices of each evidence type in each experiment. For the domain analysis, the proportion of participants choosing each evidence type was calculated for each item in Experiments 1 and 2, and those proportions were averaged across all items in each domain.

## Experiment 2

Experiment 2 sought to generalize the effect of general versus specific causation to contexts where it is known that the events in the specific causal relationship actually occurred. That is, participants in Experiment 1 evaluated claims such as "Smoking cigarettes caused Jack to get lung cancer" without knowing whether or not Jack in fact smoked and whether or not he had cancer. In such contexts, both statistical and mechanism information may seem irrelevant, since a crucial part of evaluating this claim is establishing first that the cause and effect both occurred. In contrast, Experiment 2 examined contexts where it is known that both cause and effect occurred (e.g., by prefacing the causal claim with the statement "Jack smoked cigarettes, and then Jack got lung cancer"), where the primary concern is distinguishing causation from coincidence (see Cartwright, 2017) and where the available evidence would be seen as more relevant.

## Method

We recruited 80 participants from Mechanical Turk, and excluded 5 from data analysis because they incorrectly
answered more than $33 \%$ of the check questions.
Participants responded to a new set of 24 causal claims. The format of these items differed from Experiment 1 in that contextual information was given for each claim, establishing that the cause and effect occurred. This information was printed above the box containing the claim. For example, one general item read (background information in regular typeface, claim in italics):

Researchers sometimes observe that a person consumes large amounts of meat, and then that the person develops kidney stones.
Consuming large amounts of meat causes a person to develop kidney stones.
The specific version of that item read:
Researchers observed that Tom consumed large amounts of meat, and then that Tom developed kidney stones.
Consuming large amounts of meat caused Tom to develop kidney stones.
The procedure was otherwise identical to Experiment 1.

## Results and Discussion

Although participants preferred statistical information overall, this preference was far stronger when evaluating general than when evaluating specific claims $[U=481.5$, $p=.019, r=.27]$, consistent with Experiment 1. They correspondingly chose mechanism evidence less frequently for general than for specific claims and anecdotal evidence less frequently for general than for specific claims, as shown in Table 1.

These two experiments together are consistent with the idea that people use different learning strategies depending on what causal concept they are consulting. However, there are other differences between general and specific causation that could plausibly account for some of the variance. First, the reference class from which the statistical evidence is drawn may be more relevant for the general than the specific claim, and second, plurality may have been more salient for the general than for the specific claims. We conducted an additional experiment with artificial stimuli to rule out these two alternative explanations, in which both the general and specific claims were prefaced by a statement about the reference class (e.g., "There is a group of 100 Garbotrons"), with the general claim then made about the entire group and the specific claim about an arbitrary member of that group. This equated the reference class and the salience of plurality, yet produced a similar shift across conditions.

These experiments do not fully tease apart whether the difference is due to a statistics preference for general claims or a mechanism preference for specific claims. We conducted two additional studies to answer this question, one in which participants answered an open-ended question about what evidence they would want to use, and another in which participants rated the two types of evidence on independent scales. Consistent with our claim that these differences arise due to more stochastic
representations of general causation, the condition differences were significant for statistical evidence but not for mechanism evidence in both cases.

## Domain Differences

In Experiments 1 and 2, we drew our causal claims from four domains-physical, biological, psychological, and economic-across which causal representations are likely to differ. People typically conceptualize physical causation as flowing in branches, with each event having few causes but many effects, and social (and perhaps biological) causal systems as interconnected webs, in which events have many causes and many effects (Strickland et al., 2016). Similarly, people may use more transference-based (or mechanistic) causal concepts in the physical domain, and more dependence-based (counterfactual or statistical) causal concepts in the social domain (Lombrozo, 2010). Thus, physical systems may be seen as more deterministic and social systems as more stochastic. According to the pluralistic position, these conceptual differences across domains could translate into different learning strategies: We would expect relatively greater reliance on statistical information for social and biological systems and less for physical systems.

We tested this possibility by comparing preferences for statistical evidence across all 48 items used in Experiments 1 and 2, collapsing across the general and specific versions. For each item, a statistics preference score was computed by taking the difference between the proportion of participants choosing statistical evidence for that item and the proportion choosing mechanism evidence for that item. An ANOVA on these scores with domain (physical, biological, psychological, or economic) as a between-items variable uncovered a marginally significant main effect of domain $[F(3,44)=2.22, p=$ $\left..099, \eta_{\mathrm{p}}^{2}=.13\right]$, with the preference for statistics evidence smallest for the physical items $[M=0.01, S D=0.17]$, followed by the biological $[M=0.15, S D=0.19]$, psychological $[M=0.16, S D=0.21]$, and economic $[M=$ $0.19, S D=0.17]$ items. Independent-samples $t$-tests revealed that items from the physical domain had a smaller statistics preference than did items from the combined other domains $[t(46)=-2.56, p=.014, d=$ 0.85 ], while the biological, psychological, and economic domains did not differ from one another [ $t \mathrm{~s}<1, p \mathrm{~s}>.50$ ].
This result further supports the pluralistic position, suggesting that differences in causal concepts used across domains translated into different learning strategies.

## General Discussion

Cognition requires us to attend to and integrate various sources of information into coherent representations of the world. Our representations of causal systems are particularly critical because they allow us to predict and understand events, and to plan interventions on the world to achieve goals. Humans use two distinct strategies for making inferences about causal claims-evaluating the
plausibility of mediating causal mechanisms, and evaluating statistical evidence for contingencies between cause and effect. What factors lead people to favor one strategy over the other?

First, general causal statements, which refer to a category of causal events, are seen as more compatible with statistical evidence than are specific causal statements, which refer to only an individual causal event. We hypothesized that this would occur because representations of general claims intrinsically include statistical content, and people would seek evidence that conforms to their representation of the causal concept.

Second, statistics were seen as more relevant for biological and social systems than for physical systems, whereas mechanistic evidence was more important for physical systems. We predicted this effect because causal representations vary across domains. Whereas physical systems are seen as more linear and force-based, social and biological systems are seen as more branching and counterfactual-based (Lombrozo, 2010; Strickland et al., 2016). Thus, concepts of biological and social causation would be more stochastic than concepts of physical causation, leading people to favor statistical evidence.

Causal Pluralism. Our causal representations subserve a variety of cognitive functions, and exhibit a concordant variety of properties that sometimes appear contradictory (Johnson \& Ahn, in press). For instance, causal representations seem to have many of the properties of associations (Shanks, 1987), yet causal inferences exhibit directional biases that are inconsistent with symmetric associative representations (Waldmann \& Holyoak, 1992). These shortcomings of associative theories have led to the suggestion of causal models or Bayesian networks as the representation over which causal reasoning operates (e.g., Pearl, 2000; Sloman, 2005). Yet, other evidence suggests that people often fail to make the transitive inferences predicted by Bayesian networks (i.e., that $A$ causes $C$, given that $A$ causes $B$ and $B$ causes $C$ ), and that these failures occur when the connection between $A$ and $C$ is not seen as a coherent, schematized mechanism (e.g., sex causes pregnancy, which causes nausea, but sex does not cause nausea; Johnson \& Ahn, 2015). Thus, causal representations appear to have some association-like properties, some network-like properties, and some schema-like properties. Add to this evidence that causal relations are represented with some properties of forces (Wolff, 2007), icons that support mental simulation (Hegarty, 2004), and metacognitive placeholders (Rozenblit \& Keil, 2002), and it becomes clear that people do not represent causation using one unified representation (see Markman \& Dietrich, 2000).

Despite the overwhelming evidence for representational pluralism, it does not follow that people use distinct strategies for learning about different varieties of causal concepts. People may not tailor their learning strategies to the representation at hand, but could instead apply a single learning strategy across all types of causal systems,
such as statistical learning algorithms (Pearl, 2000).
However, the current experiments demonstrate learning patterns that are not only pluralistic, but appear to be tailored to the underlying representation. In the cases of specific causation we used, there is no prior knowledge, so the only option is to learn about the general causal relation anew. If the best strategy for learning about the general claim is statistics, then the best strategy for learning about the specific relation is also statistics. Yet, participants shifted dramatically from statistics when learning about specific claims-a signal that they had applied a heuristic, matching statistical representations of general claims to statistical information. Therefore, any view of causal learning and representation that focuses on a single representation or learning mechanism will fail to capture important aspects of our causal thinking.

In addition to clarifying the debate between mechanism and statistical views of causation, causal pluralism may also be a helpful framework for thinking about debates over causal semantics. Theories of causal semantics embrace diverse accounts based on physical forces (Wolff, 2007), on probability (Good, 1961), and on logic (Lewis, 1973). Teasing these accounts apart has been difficult because they often make similar empirical predictions (Barbey \& Wolff, 2007; Goldvarg \& JohnsonLaird, 2001; Sloman, Barbey, \& Hotaling, 2009).

However, in a pluralistic framework, it may not only be difficult but in fact impossible to capture all of causal semantics using a single representational format. Our causal representations differ not only in reference (general or specific) and domain (physical, biological, social), but along many other dimensions as well, in potentially interconnected ways-among deterministic, chaotic, and indeterministic systems; among the past, present, and future; between observed, unobserved, and unobservable causes and effects; between categorically or continuously valued causes and effects; and among various potential causal structures. A useful strategy going forward may be to investigate the manner in which such variation in causal meaning propagates to causal learning processes.

## References

Ahn, W., \& Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. Cognitive Psychology, 31, 82-123.
Ahn, W., Kalish, C.W., Medin, D.L., \& Gelman, S.A. (1995). The role of covariation versus mechanism information in causal attribution. Cognition, 54, 299352.

Barbey, A.K., \& Wolff, P. (2007). Learning causal structure from reasoning. In Proceedings of the 29th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Cartwright, N. (2004). Causation: One word, many things. Philosophy of Science, 71, 805-20.
Cartwright, N. (2017). Single case causes: What is
evidence and why. In Philosophy of Science in Practice. Cham, Switzerland: Springer International.
Cheng, P.W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104, 367405.

Cimpian, A., \& Erickson, L.C. (2012). The effect of generic statements on children's causal attributions: Questions of mechanism. Developmental Psychology, 48, 159-70.
Danks, D. (2005). The supposed competition between theories of human causal inference. Philosophical Psychology, 18, 259-72.
Fernbach, P.M., Darlow, A., \& Sloman, S.A. (2011). Asymmetries in predictive and diagnostic reasoning. Journal of Experimental Psychology: General, 140, 168-85.
Fugelsang, J.A., \& Thompson, V.A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. Canadian Journal of Experimental Psychology, 54, 15-32.
Goldvarg, E., \& Johnson-Laird, P.N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. Cognitive Science, 25, 565-610.
Good, I. J. (1961). A causal calculus (I). The British Journal for the Philosophy of Science, 11, 305-318.
Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., \& Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. Psychological Review, 111, 3-32.
Griffiths, T.L., \& Tenenbaum, J.B. (2005). Structure and strength in causal induction. Cognitive Psychology, 51, 334-84.
Hegarty, M. (2004). Mechanical reasoning by mental simulation. Trends in Cognitive Sciences, 8, 280-285.
Heider, F. (1958). The psychology of interpersonal relations. Hillsdale, NJ: Erlbaum.
Hitchcock, C. (2003). Of Humean bondage. The British Journal for the Philosophy of Science, 54, 1-25.
Johnson, S.G.B., \& Ahn, W. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. Cognitive Science, 39, 1468-503.
Johnson, S.G.B., \& Ahn, W. (in press). Causal mechanisms. In The Oxford Handbook of Causal Reasoning. Oxford, UK: Oxford University Press.
Johnson, S.G.B., Valenti, J.J., \& Keil, F.C. (2017). Opponent uses of simplicity and complexity in causal explanation. In Proceedings of the 39th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Kahneman, D., \& Tversky, A. (1982). The simulation heuristic. In Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press.
Keil, F.C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: MIT Press.
Kim, N.S., Ahn, W., Johnson, S.G.B., \& Knobe, J.
(2016). The influence of framing on clinicians' judgments of the biological basis of behaviors. Journal of Experimental Psychology: Applied, 22, 39-47.
Lewis, D. (1973). Causation. The Journal of Philosophy, 70, 556-67.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. Cognitive Psychology, 61, 303-32.
Lombrozo, T., \& Carey, S. (2006). Function explanation and the function of explanation. Cognition, 99, 167204.

Markman, A.B., \& Dietrich, E. (2000). In defense of representation. Cognitive Psychology, 40, 138-171.
Medin, D.L., \& Ortony, A. (1989). Psychological essentialism. In Similarity and analogical reasoning. Cambridge, UK: Cambridge University Press.
Park, J., \& Sloman, S.A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. Cognitive Psychology, 67, 186-216.
Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge, UK: Cambridge University Press.
Rozenblit, L., \& Keil, F.C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26, 521-562.
Shanks, D.R. (1987). Associative accounts of causality judgment. Psychology of Learning and Motivation, 21, 229-261.
Sloman, S.A. (2005). Causal models: How people think about the world and its alternatives. Oxford, UK: Oxford University Press.
Sloman, S.A., Barbey, A.K., \& Hotaling, J.M. (2009). A causal model theory of the meaning of Cause, Enable, and Prevent. Cognitive Science, 33, 21-50.
Spellman, B.A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. Psychological Science, 7, 337-42.
Steyvers, M., Tenenbaum, J.B., Wagenmakers, E., \& Blum, B. (2003). Inferring causal networks from observations and interventions. Cognitive Science, 27, 453-89.
Strickland, B., Silver, I., \& Keil, F.C. (2016). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. Memory \& Cognition.
Taleb, N. N. (2007). The black swan: The impact of the highly improbable. New York, NY: Random House.
Waldmann, M.R., \& Holyoak, K.J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. Journal of Experimental Psychology: General, 121, 222-236.
Waldmann, M. R., \& Mayrhofer, R. (2016). Hybrid causal representations. In Psychology of Learning and Motivation (Vol. 65). New York, NY: Academic Press.
Wolff, P. (2007). Representing causation. Journal of Experimental Psychology: General, 136, 82-111.

# Touch Screen Text Entry as Cognitively Bounded Rationality 

Jussi Jokinen<br>Aalto University, Espoo, Finland


#### Abstract

Typing on a smartphone is an everyday activity that involves various cognitive and behavioural processes. This paper models touch screen text entry as cognitively bounded rationality. The model aims to maximise error-free text throughput, while being constrained by its architecture and task environment. Empirical data are used to calibrate the model, which demonstrates adequate fit. The model is used to explore how strategic choices under given constraints affect text entry performance. The preliminary model presented here serves as a confirmation that touch screen text entry can be modelled as cognitively bounded rationality. Future extensions by integration into richer cognitive architectures are outlined.


# Using Measurement Models to Understand Eyewitness Identification 

Matthew Kaesler (matthew.kaesler@adelaide.edu.au)<br>School of Psychology, University of Adelaide<br>Adelaide, SA 5005, Australia<br>John Dunn (john.dunn@uwa.edu.au)<br>School of Psychology, University of Western Australia<br>Perth, WA, 6009, Australia<br>Carolyn Semmler (carolyn.semmler@adelaide.edu.au)<br>School of Psychology, University of Adelaide<br>Adelaide, SA 5005, Australia


#### Abstract

Much research effort has been expended improving police lineup procedures used in collecting eyewitness identification evidence. Sequential presentation of lineup members, in contrast to simultaneous presentation, has been posited to increase witness accuracy, though analyses based in Signal Detection Theory (SDT) have challenged these claims. A possible way to clarify the effect of presentation format on witness accuracy is to develop SDT-based measurement models, which characterise decision performance in terms of psychologically-relevant parameters, particularly discriminability and response bias. A model of the sequential lineup task was developed with a "first-above-criterion" decision rule, alongside a simultaneous model with a "maximum familiarity" decision rule. These models were fit to a corpus of data comparing simultaneous and sequential lineup performance. Results showed no difference in discriminability between the procedures and more conservative responding for the sequential lineup. Future work will examine criterion setting in the sequential lineup and model alternative decision rules.


Keywords: Eyewitness identification; Signal Detection; memory

## Introduction

Typical six-person lineups used in police investigations consist of one suspect, whom police believe may be guilty of a crime, and five known-innocents called variously "fillers' or "foils", selected to resemble the suspect in specified ways (Clark, 2012). A witness may identify (ID) a person from the lineup or reject the lineup ("the person I saw is not here"), and may provide a confidence rating for their choice. In experimental mock-crime studies, a lineup is referred to as target-present (TP) if it includes the person observed by the witness at encoding (the "culprit") or targetabsent (TA) if it is composed entirely of fillers.

Possible decision outcomes are expressed as rates or proportions over a series of trials. From a TP lineup, witnesses may correctly ID the culprit, incorrectly ID a foil or incorrectly reject the lineup. From a TA lineup, witnesses may correctly reject the lineup or incorrectly ID a foil, known as a false alarm.

The members of a lineup may be presented simultaneously, where the witness makes a single decision, or sequentially, where the witness makes a yes/no decision for each member before seeing the next. In experimental studies, the sequential procedure terminates once an ID is made, although variations of the procedure are used in applied settings (Horry, Palmer \& Brewer, 2012). Whether presentation format affects witness accuracy has received significant research attention (Steblay, Dysart \& Wells, 2011). Initial work by Lindsay and Wells (1985) found that the sequential procedure produced a marked reduction in false ID rate, a desirable outcome, and a slight reduction in correct ID rate compared to the simultaneous procedure. Numerous subsequent studies and two meta-analyses (Steblay et al., 2011) have supported this pattern of results and, although the effect has not always been found (Dobolyi \& Dodson, 2013) and seems to have weakened with time (Moreland \& Clark, 2016), evidence for sequential superiority has been persuasive enough for the procedure to be adopted in the United Kingdom and in many jurisdictions in the United States (Clark, 2012).

## Signal Detection Theory Advantage

Recently, researchers have advocated the use of analyses based in Signal Detection Theory (SDT) to evaluate lineup procedures (Mickes, Flowe \& Wixted, 2012). SDT is an approach used to analyse decision performance in a wide variety of areas in which a target, such as an enemy jet on radar or a tumor on an x-ray, must be discriminated from similar non-targets under conditions of uncertainty. In its most basic form, it characterises decision performance as resulting from two sources; discriminability ( $d$ '), related to how well a witness can distinguish targets from non-targets, and response bias or criterion (c), related to overall willingness to make a decision (MacMillan \& Creelman, 2005). Claims of superior performance for the sequential procedure have been based on findings of a higher ratio of correct ID rates to false ID rates - the so called "diagnosticity ratio" - compared to the simultaneous procedure (Steblay et al., 2011). However, Mickes et al.
(2012) have shown that diagnosticity confounds discriminability and response bias. In fact, analysis of lineup data using SDT reveals that there are a range of different diagnosticity ratios associated with the same lineup procedure and that, necessarily, the ratio increases as response bias becomes more conservative, i.e. as people reject more frequently (Wixted \& Mickes, 2012). SDT avoids this problem by computing all empirical or hypothetical correct ID/false ID pairs, which can then be plotted and analysed. Further, formal modelling of the task allows estimation of the entire response curve from estimated parameters - allowing tests of hypotheses about the impact of system variables such as lineup member similarity on theoretically relevant parameters.

## Measurement Models

A measurement model uses theoretically-derived mathematical functions to link observed behavioural data to psychological constructs. Psychologically meaningful model parameters are estimated by fitting the model to observed data (Farrell \& Lewandowsky, 2010).

Palmer and Brewer (2012) sought to address the need for formal modelling of the lineup task by fitting a SDT compound detection model (SDT-CD; Duncan, 2006) to a corpus of studies that compared simultaneous and sequential lineup data. The 'compound' aspect of SDT-CD refers to the fact that a lineup can be decomposed into two decision tasks; target detection (is the target present?) and target identification (if the target is present, which member is the target?). Results showed that the simultaneous and sequential lineups did not significantly differ in terms of discriminability but that the sequential procedure led to more conservative responding.

However, there were critical aspects of the analyses conducted by Palmer and Brewer (2012) that may have affected their results. First, the SDT-CD model was developed to account for simultaneous presentation of stimuli - it does not directly model the sequential procedure. For this reason, we develop and apply a formal model of the sequential procedure. Second, the best-fitting parameter values reported by Palmer and Brewer (2012) may not have been optimal as they appear not to have been estimated by an optimization procedure. For this reason, we fit the two models using a computational optimization routine. Third, given distinct models of simultaneous and sequential procedures, it is important to explore the dependence of each task on the criterion that is set by the witness. Previous research (Horry et al. 2012) has highlighted the vulnerability of the sequential task to criterion setting so we compare the simultaneous and sequential models in terms of their dependence on the decision criteria.

## New Models

Both the simultaneous and sequential models assume an underlying unequal variance SDT model based on prior recognition memory research (Mickes, Wixted \& Wais, 2007). Figure 1 illustrates this model for a single person
lineup. Each member of a lineup is associated with a particular value of memory strength or familiarity. Foil familiarity on both TP and TA trials is modelled as a random draw from a normal distribution (dashed line in Figure 1) with mean zero and standard deviation one. Target familiarity is modeled as a random draw from a normal distribution (solid line in Figure 1), with a mean $d^{\prime}$ and standard deviation $s$.

A decision is made in relation to a decision criterion, $c$. This functions as a 'choice threshold' and reflects response bias. If familiarity is greater than $c$, the witness will identify the lineup member as the culprit. Otherwise, they reject the lineup. The greater the value of $c$, the more conservative is the decision and the less likely that an ID is made.

The following functions define the hit rate ( $h$ ) and false alarm rate ( $f$ ), where $\Phi($.$) is the normal cumulative$ distribution function:

$$
\begin{aligned}
& h(c)=1-\Phi\left(\left(c-d^{\prime}\right) / s\right) \\
& f(c)=1-\Phi(c)
\end{aligned}
$$

For an $n=1$ lineup, $h$ is the correct ID rate and $f$ is the false ID rate.

Figure 1: Basic representation of the unequal variance signal detection model


Simultaneous Model (SDT-SIM). In a simultaneous lineup, there are $n>1$ members where typically, $n=6$. Let $x_{1}, \ldots, x_{n}$ be the familiarity values of each member of the lineup and let $m$ the maximum of these values. The SDTSIM model implements the decision rule to choose member $k$ if $m>c$ and $m=x_{k}$, otherwise to reject the lineup. For a TP lineup, if a choice is made and member $k$ is the target, a correct ID has been made, otherwise, a foil ID has been made. Any selection on a TA lineup is a false alarm.

Sequential Model (SDT-SEQ). In a sequential lineup, the witness makes a decision for each member, presented in a fixed order labelled by indices, from 1 to $n$. Let $K$ be a subset of these indices, such that $x_{i}>c$ for all $i \in K$. If $K$ is empty then the lineup is rejected, otherwise the witness
chooses member $k$ where $k$ is the smallest (i.e. the first) element in $K$. For a TP lineup, if a choice is made and member $k$ is the target then a correct ID has been made, otherwise an incorrect ID is made. Any selection on a TA lineup is a false alarm.

## Summary \& Aims

Despite recent efforts to apply SDT to eyewitness identification, there has been no attempt to model the sequential lineup task taking account of the differences in the decision rule and the importance of criterion setting for response probabilities. We offer, to our knowledge, the first formal measurement model of the sequential lineup. Further, previous research did not use optimization routines to find the best fitting parameters for the data, and thus the conclusions may be different. Finally, the application of formal models in the eyewitness identification domain highlights the most important factors likely to impact on the rates of false identifications of innocent people and failure to detect the presence of a guilty suspect. In summary, the aims of the present study are:

1. To implement a formal model of the sequential lineup procedure (SDT-SEQ).
2. To reanalyse the data reported by Palmer and Brewer (2012), fitting SDT-SIM to the simultaneous lineup results and SDT-SEQ to the sequential lineup results.
3. To compare SDT-SIM and SDT-SEQ in terms of their dependence on parameter values, particularly decision criterion.

## Method

## Studies Analysed

A corpus of 22 studies $(N=3871)$ assembled by Palmer and Brewer (2012) that directly compared simultaneous and sequential presentation, making 44 data sets in all, was reanalysed. Following Steblay, et al.'s (2011) 'full diagnostic design' inclusion criteria the studies all a) manipulated both presentation format and target presence/absence, b) showed ID performance above chance levels and c) involved only adult participants.

## Statistical Analyses

SDT-SIM and SDT-SEQ were fit using optimization of maximum likelihood (implemented using Matlab FMINCON function). This searches parameter space for values of $d^{\prime}$ and $c$ that best characterise observed decision performance. We report goodness-of-fit in terms of the $G^{2}$ statistic which is a function of the maximum likelihood and distributed as chi-squared.

Statistical Considerations. Due to a lack of confidence rating data in many of the studies analysed, the standard deviation of the target distribution ( $s$ ) was not estimated. Instead, the value of $s$ was fixed to one. This is a plausible
assumption in the eyewitness paradigm where each participant encodes a single study item (the culprit). The greater variance of the target distribution observed in recognition memory research may be attributed to encoding variability across a range of study items (Mickes et al., 2007).

Additionally, Palmer and Brewer (2012), following Duncan (2006), used a relative measure of criterion value with the zero point positioned between the target and lure distributions, i.e. $C=c-d^{\prime} / 2$. Both absolute ( $c$ ) and relative $(C)$ criteria are reported here.

## Results

The new models fit the data well; SDT-SIM could not be rejected for 19 of 22 simultaneous data sets, as indicated by non-significant values of $G^{2}$. The model was rejected for data from Carlson, Gronlund and Clark (2008; Experiment 2), Lindsay and Wells (1985), and Rose, Bull and Vrij (2005). SDT-SEQ was also not rejected for 19 of 22 sequential data sets but was rejected for data from Carlson et al. (2008; Experiment 2), Lindsay and Wells (1985), and Pozzulo and Marciniak (2006). The SDT-CD model was also rejected for data from Carlson et al. (2008; Experiment 2), Lindsay and Wells (1985) and Pozzulo and Marciniak (2006), in addition to Greathouse and Kovera (2009).

Taking the parameter values estimated for each dataset, average values of $d^{\prime}, c$ and $C$ were calculated over the corpus of data, weighted according to sample size. Table 1 displays the summary results obtained by Palmer and Brewer (2012) obtained from fitting the SDT-CD model (equivalent to the SDT-SIM model) to data from both simultaneous and sequential lineups, compared to the summary results obtained by fitting the SDT-SIM model to data from both simultaneous and sequential lineups, and fitting the SDT-SEQ model to data from sequential lineups.

Table 1: Mean weighted parameter values from SDT-CD, SDT-SIM and SDT-SEQ

| SDT-CD | $d^{\prime}$ | $c$ | $C$ |
| :--- | :---: | :---: | :---: |
| Simultaneous | 1.64 | -.06 | -.89 |
| Sequential | 1.74 | .44 | -.43 |
| SDT-SIM |  |  |  |
| Simultaneous | 1.37 | 1.21 | .53 |
| Sequential | 1.33 | 1.53 | .87 |
| SDT-SEQ |  |  |  |
| Sequential | 1.40 | 1.55 | .85 |

The first step in our analysis was to fit SDT-SIM to both simultaneous and sequential datasets, attempting to recover a similar pattern of results to those obtained by Palmer and Brewer (2012). Fitting SDT-SIM to all datasets produced a similar pattern of estimates to SDT-CD, with a significantly higher mean weighted $C$ value for the sequential datasets, as indicated by a Welch two-sample weighted $t$-test, $t(36.42)=$ $-3.89, p<.05$, and no significant difference in mean
weighted $d^{\prime}$ values between simultaneous and sequential datasets, $t(41.67)=.34, p=.73$.
The second step was to compare the parameter values recovered by fitting SDT-CD and SDT-SIM/SEQ. Figures 2 and 3 plot the estimated parameter values recovered for each data set when fitting SDT-CD to both simultaneous and sequential datasets and SDT-SIM and SDT-SEQ to their respective datasets.

Figure 2. $C$ vs $d$ ' estimates for all datasets SDT-CD


Figure 3. $C$ vs $d^{\prime}$ 'estimates for all datasets SDT-SIM/SEQ


The difference in $y$-axis range between Figure 2 and Figure 3 indicate that fitting SDT-SIM and SDT-SEQ to their respective datasets produced higher criterion estimates compared to SDT-CD. Welch two-sample weighted $t$-tests indicated that mean weighted $C$ was significantly higher for SDT-SIM, $t(36.74)=-16.41, p<.05$, and SDT-SEQ, $t(38.91)$ $=-9.87, p<.05$, compared to mean weighted $C$ from SDTCD for simultaneous and sequential datasets respectively. There was no difference in mean weighted $d^{\prime}$ for SDT-SIM compared to SDT-CD as fit to simultaneous datasets, $t(40.33)=1.89, p=.06$, however mean weighted $d^{\prime}$ for

SDT-SEQ was significantly lower than SDT-CD as fit to sequential datasets $t(33.76)=2.11, p<0.5$.
The final stage of the analysis was to compare $C$ and $d$ ' values generated by fitting SDT-SIM and SDT-SEQ to their respective data types, as displayed in Figure 3. Examining Figure 3 reveals a cluster of sequential datasets with higher $C$ values than the cluster of simultaneous datasets. A Welch two-sample weighted $t$-test, $t(35.03)=-3.53, p<.05$, indicated that the mean weighted $C$ value of the sequential datasets as estimated by SDT-SEQ was significantly higher than that of the simultaneous datasets as estimated by SDTSIM.
There are no such patterns evident relative to the horizontal axis, with $d^{\prime}$ values for most of the datasets clustered from approximately 1 to 2 . No significant difference between the mean weighted $d^{\prime}$ values for the simultaneous and sequential datasets was found, $t(41.54)=-$ $.28, p=.81$.

## Discussion

The present study developed and fit two SDT-based formal measurement models of the simultaneous (SDT-SIM) and sequential (SDT-SEQ) eyewitness lineup task to a corpus of data collected by Palmer and Brewer (2012) using an optimization procedure, and compared the model's dependence on the parameters discriminability ( $d^{\prime}$ ) and response bias (c) in order to better understand decision performance on the lineup task.
Fitting SDT-SIM to both simultaneous and sequential data, following Palmer and Brewer (2012), produced similar parameter estimates to those generated by fitting SDT-CD, with results suggesting that the sequential procedure encourages more conservative responding but does not differ in discriminability. Fitting SDT-SIM and SDT-SEQ to their respective data types further reinforced this pattern of results, supporting the conclusions reached by Palmer and Brewer (2012).
Compared to SDT-CD, SDT-SIM and SDT-SEQ produced higher criterion estimates for their respective data types while, while SDT-SEQ also produced lower discriminability estimates. The difference in parameter values between Palmer and Brewer (2012) and the results here is likely due to the task and fitting the models using an efficient optimization procedure rather than grid search.
While no discriminability differences were reported here or in Palmer and Brewer (2012), previous studies using ROC analysis to calculate observed discriminability from rating data have shown a discriminability advantage for the simultaneous lineup (Dobolyi \& Dodson, 2013; Mickes et al., 2012). Results here do not necessarily contradict these findings, as simulations can generate different shaped ROC curves from different ID procedures despite holding theoretical $d^{\prime}$ constant between them (Rotello \& Chen, 2016). The relationship between theoretical $d^{\prime}$ as estimated by SDT-SIM and SDT-SEQ and observed "Area Under the Curve" measures of $d^{\prime}$, as used in ROC analysis, will likely be investigated in future work with new data that includes
confidence ratings. Regarding theoretical discriminability, these results challenge the diagnostic feature detection model (Wixted \& Mickes, 2014) that proposes a discriminability advantage for the simultaneous lineup arising from witnesses’ ability to identify diagnostic features between lineup members.

The more conservative decision criterion observed on the sequential lineup can be explained by differences between its "first-above-criterion" identification rule of SDT-SEQ and the "maximum familiarity" rule of SDT-SIM. This difference becomes evident as the decision criterion is made more lenient. In the limit, the most liberal decision criterion (i.e. always choose) in the sequential lineup results in selection of the first lineup member, leading to chance performance; a correct ID rate of $1 / 6$ for TP lineups and a false alarm rate of $1 / 6$ for TA lineups. In contrast, the "maximum familiarity" rule of SDT-SIM means that, for the most liberal decision criterion, the witness always chooses the most familiar lineup member. If $d^{\prime}>0$, the lineup member with maximum familiarity in TP lineups is more likely than other members to be the target, leading to a correct ID rate greater than $1 / 6$, while for TA lineups the false ID rate remains at $1 / 6$. The effect of this difference is that in order for ID performance to be comparable between simultaneous and sequential lineups, the latter requires a more conservative decision criterion.

Based on the present findings, any performance advantage attributed to the sequential procedure is likely due to a stricter decision criterion, not improved discriminability. This suggests that changes in lineup procedure do not alter underlying memory strength. Rather, the quality of memory information available to a witness is largely determined at encoding by factors such as distance, lighting and exposure time (Maclin, Maclin \& Malpass, 2001). The present findings also do not support of the proposal that performance differences are the result of procedural effects on retrieval or reconstructive memory processes taking place during a lineup decision as these are likely to affect discriminability (Ebbesen \& Flowe, 2002).
Wells (2014) acknowledged the mounting body of evidence showing that any perceived sequential lineup advantage is the result of a more conservative decision criterion but contends that it is more useful in applied settings no matter the source of any performance difference. As other researchers have noted (e.g. Clark, 2012), conservative responding reduces both false IDs and correct IDs. If policy makers consider conservative responding in the lineup task desirable, such an affect could be achieved by simpler means than retraining police to administer lineups sequentially, such as instructing witnesses to be very careful in their choosing or by only counting IDs made at high confidence (Wixted \& Mickes, 2012).

## Limitations

Decision Strategy. The present work uses an absolute decision strategy for both the simultaneous and sequential models, despite the simultaneous lineup's long association
with the so-called relative judgement strategy (Wells, 1984). Wells (1984) proposed that the increased innocent suspect ID rate of the simultaneous procedure may be due to the tendency for witnesses to compare across lineup members, selecting the one that most resembles their memory of the perpetrator relative to the other members, rather than comparing each lineup member directly to their memory of the perpetrator as in the absolute decision strategy (cf. Wixted \& Mickes, 2014). The absolute vs. relative distinction has gained some traction in the literature and has received some empirical investigation, although the superiority of one strategy over the other has not been demonstrated (Fife, Perry \& Gronlund, 2014). In line with our present approach, formal modelling of the relative decision strategy could clarify the utility of the absolute/relative distinction to understanding lineup performance.

One option for implementing relative judgement is to use the difference in familiarity between the lineup member with maximum familiarity ( $m$ ) and the next-most-familiar lineup member, which seems to accord with Wells' (1984) description. The rule would be; if this difference score exceeds a criterion, then choose the lineup member with maximum familiarity, otherwise reject the lineup. We are currently developing a formal model based on this rule.

## Future Directions

Criterion Shift in Sequential Lineup. The present work demonstrates that the sequential lineup decisions are critically affected by the placement of the decision criterion. A further question is whether the decision criterion may change over the course of the lineup. In an attempt to forestall such changes, Horry, Palmer and Brewer (2012) investigated the efficacy of "backloading", telling the witness that they will be viewing more photos than there are lineup members. The results indicated that the more photographs the witness was told to expect, the more conservative were their decision criteria. On nonbackloaded lineups, foil choices increased in the later lineup positions. Because Horry, Palmer and Brewer (2012) fit SDT-CD to the data to generate parameter estimates, future research could explore whether these conclusions remain valid after fitting the SDT-SEQ model to these data.

## Conclusion

This study presents two formal measurement models of the simultaneous and sequential lineup tasks, which were fit to a large corpus of data using computational optimization. The development of a sequential model is particularly noteworthy, it being, to our knowledge, the first of its kind in the eyewitness literature. Results show no difference in discriminability between the two procedures and a more conservative decision criterion in the sequential procedure. The models offer a means to investigate the effects of system variables on eyewitness performance in terms of theoretically relevant underlying parameters, demonstrating the value of formal modelling in applied research.

## Acknowledgments

This research was funded by the Australian Government through the Australian Research Council, Discovery Projects DP160101048.

## References

Carlson, C. A., Gronlund, S. D., \& Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. Journal of Experimental Psychology: Applied, 14(2), 118-128.
Clark, S. E. (2012). Costs and Benefits of Eyewitness Identification Reform: Psychological Science and Public Policy. Perspectives on Psychological Science, 7(3), 238259.

Dobolyi, D. G., \& Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. Journal of Experimental Psychology: Applied, 19(4), 345-357.
Duncan, M. J. (2006). A signal detection model of compound decision tasks. (Tech. Rep. No. TR2006-256).
Toronto, ON: Defence Research and Development Canada.
Ebbesen, E. B., \& Flowe, H. D. (2002). Simultaneous v. sequential lineups: What do we really know?. Retrieved December 202016 from https://dspace.lboro.ac.uk/2134/20167
Farrell, S., \& Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. Current Directions in Psychological Science, 19(5), 329-335.
Fife, D., Perry, C., \& Gronlund, S. D. (2014). Revisiting absolute and relative judgments in the WITNESS model. Psychonomic Bulletin \& Review, 21(2), 479-487.
Horry, R., Palmer, M. A., \& Brewer, N. (2012). Backloading in the sequential lineup prevents withinlineup criterion shifts that undermine eyewitness identification performance. Journal of Experimental Psychology: Applied, 18(4), 346-360.
Lindsay, R., \& Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. Journal of Applied Psychology, 70(3), 556-564.
MacLin, O. H., MacLin, M. K., \& Malpass, R. S. (2001). Race, Arousal, Attention, Exposure, and Delay: An Examination of Factors Moderating Face Recognition. Psychology, Public Policy, and Law, 7(1), 134-152.
MacMillan, N., Creelman., C (2005). Detection theory: A user's guide (2 $2^{\text {nd }}$ ed.). Mahwah, New Jersey. Lawrence Erlbaum Associates.
Mickes, L., Flowe, H. D., \& Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. Journal of Experimental Psychology: Applied, 18(4), 361-376.
Mickes, L., Wixted, J. T., \& Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of
recognition memory. Psychonomic Bulletin \& Review, 14(5), 858-865.
Moreland, M. B., \& Clark, S. E. (2016). Eyewitness Identification: Research, Reform, and Reversal. Journal of Applied Research in Memory and Cognition, 5(3), 277283.

Palmer, M. A., \& Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. Law and Human Behavior, 36(3), 247-255.
Pozzulo, J. D., \& Marciniak, S. (2006). Comparing identification procedures when the perpetrator has changed appearance. Psychology, Crime \& Law, 12(4), 429-438.
Rose, R. A., Bull, R., \& Vrij, A. (2005). Non-biased lineup instructions do matter--A problem for older witnesses. Psychology, Crime \& Law, 11(2), 147-159.
Rotello, C. M., \& Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. Cognitive Research: Principles and Implications, 1 (1), 10.
Steblay, N. K., Dysart, J. E., \& Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. Psychology, Public Policy, and Law, 17(1), 99-139.
Wells, G. L. (1984). The Psychology of Lineup Identifications. Journal of Applied Social Psychology, 14(2), 89-103.
Wells, G. L. (2014). Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. Current Directions in Psychological Science, 23(1), 11-16.
Wixted, J. T., \& Mickes, L. (2012). The Field of Eyewitness Memory Should Abandon Probative Value and Embrace Receiver Operating Characteristic Analysis. Perspectives on Psychological Science, 7(3), 275-278.
Wixted, J. T., \& Mickes, L. (2014). A signal-detectionbased diagnostic-feature-detection model of eyewitness identification. Psychological Review, 121(2), 262-276.

# A Biologically Constrained Model of Semantic Memory Search 

Ivana Kajić, Jan Gosmann, Brent Komer, Ryan W. Orr, Terrence C. Stewart, Chris Eliasmith<br>Centre for Theoretical Neuroscience, University of Waterloo<br>Waterloo, ON, Canada N2L 3G1<br>\{i2kajic, jgosmann, bjkomer, rworr, tcstewar, celiasmith\}@uwaterloo.ca


#### Abstract

The semantic fluency task has been used to understand the effects of semantic relationships on human memory search. A variety of computational models have been proposed that explain human behavioral data, yet it remains unclear how millions of spiking neurons work in unison to realize the cognitive processes involved in memory search. In this paper, we present a biologically constrained neural network model that performs the task in a fashion similar to humans. The model reproduces experimentally observed response timing effects, as well as similarity trends within and across semantic categories derived from responses. Three different sources of the association data have been tested by embedding associations in neural connections, with free association norms providing the best match.


Keywords: semantic memory; associations; semantic search; spiking neural network; neural engineering framework

## Introduction

The semantic memory system plays an important role in a variety of cognitive functions. It is essential for language comprehension and understanding, and has been referred to as a mental thesaurus, storing knowledge about words, their meaning and relationships among them (Tulving, 1983).

The advent of neuroimaging techniques and observations from brain lesion studies have allowed more specific localization of the brain regions and networks responsible for semantic representation and processing (Huth, de Heer, Griffiths, Theunissen, \& Gallant, 2016; Quiroga, 2012). In particular, the medial temporal lobe and portions of anterior lobes have been identified as essential to the function of semantic memory. Purely computational semantic network models have successfully explained behavioral data (Collins \& Quillian, 1969; Collins \& Loftus, 1975) and have been purported to reveal principles guiding language formation and organization (Steyvers \& Tenenbaum, 2005). Yet, they have been severely limited in their ability to account for the neural realization of such processes. Our understanding of how networks of millions of neurons perform the computations that underly semantic processing is still extremely limited.

We propose a network of simulated spiking neurons that is able to perform the semantic fluency task in a manner similar to humans. While providing a good match with behavioral data, the model also proposes specific neural mechanisms that may be involved in semantic processes. The components of the model are discussed in terms of functionally and neurologically plausible counterparts found in the human brain.

## Search in the Semantic Space

The semantic fluency task has been used to understand how humans search memory when asked to retrieve items se-
mantically related to a given cue (Thurstone, 1938; Bousfield \& Sedgewick, 1944). In a typical trial, a person is instructed to generate members of a category within a given time limit. One common version of the task requires an individual to list all animals they can think of within a fixed timespan of one or more minutes. Response analysis shows they tend to be grouped into clusters corresponding to subcategories (Troyer, Moscovitch, \& Winocur, 1997), such as pets or farm animals. For example, responses might start with the animals an individual is most familiar with, such as cat, $d o g$, rabbit and then continue with a list of farm animals such as cow, chicken and turkey.

To explain the clustering trend observed in the responses, Hills, Jones, and Todd (2012) suggested that individuals generate responses according to the optimal foraging policy (Charnov, 1976). Animals use such a strategy when searching for food in natural environments: after resources in one area have been depleted, animals continue their search for food in a new patch. In the context of the semantic fluency task, an individual listing animals in a specific sub-category would stop listing animals from that category after being unable to generate new items at a certain rate. Search behavior suggestive of optimal foraging has been reproduced with several different representations and algorithms, including a random walk on a semantic network constructed from free association norms (Abbott, Austerweil, \& Griffiths, 2015). Jones, Hills, and Todd (2015) attribute the simplicity of this particular algorithm to the association norms being a direct result of an experimental design that is very similar to the semantic fluency task. They argue that the fundamental memory retrieval processes and representations are obscured by the data underlying the model and the behaviors that are being explained. However, association data from sources other than association norms, like data learned from natural language, have successfully been used to reproduce human response patterns with random walks (Nematzadeh, Miscevic, \& Stevenson, 2016).

Here, we take a first step towards explaining how the memory retrieval processes and representations described above can be realized by a biologically constrained neural network. The proposed model performs the search based on associative weights encoded within connections between neurons, resembling aspects of a random walk while still conforming to constraints of neural computation. The noise resulting from spiking neurons and the diversity in neuron parameter values lead to the response variability. We show that the search patterns observed in the model responses are consistent with the optimal foraging theory and match human behavioral data.


Figure 1: A: Architecture of the neural network model performing the semantic fluency task. Each box represents a population of spiking neurons. B: Neuronal spiking activity in the model recorded from the population cue. Some neurons are actively spiking when representing words dog, cat and donkey (highlighted area 1), while others only spike when representing words $d o g$ and cat (highlighted area 2). The similarity between these spikes and the ideal spike pattern for each word is shown above.

## Biologically Constrained Representation

Brain imaging studies provide evidence in support of semantic representations distributed across networks of neurons in various brain regions (Huth et al., 2016; Rissman \& Wagner, 2012). While many neurons jointly contribute to representations, single neurons can still exhibit preference for certain input stimuli. For example, neurons in the medial temporal lobe show selective responses for higher-level semantic concepts such as places or people (Quiroga, 2012).

Consistent with the notion of a distributed representation, we employ vector-based representations that can be implemented in a network of spiking neurons by means of the Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003). In the NEF, connection weights between neurons can be analytically computed such that the neural network approximates a desired function.

Given an $n$-dimensional vector representing a preferred stimulus $\boldsymbol{e}$ and some time-varying input $\boldsymbol{x}$, the activity of a single neuron $a_{i}$ can be expressed as

$$
\begin{equation*}
a_{i}=G_{i}\left[\alpha_{i} \boldsymbol{e}_{i}^{\top} \boldsymbol{x}+J_{i}^{\mathrm{bias}}\right] \tag{1}
\end{equation*}
$$

where $G$ represents a spiking neuron model, in this case the Leaky Integrate-and-Fire (LIF) model. The parameter $\alpha$ scales the input and converts the unit of the variable ( $\boldsymbol{x}$ ) to units of current, and $J^{\text {bias }}$ represents background currents.

As a result, if a neuron is driven by an input $\boldsymbol{x}$ that is similar to its preferred direction $\boldsymbol{e}$, the dot product $\boldsymbol{e}^{\top} \boldsymbol{x}$ is larger ( $\boldsymbol{e}^{\top}$ is a transposed vector $\boldsymbol{e}$ ). For a LIF neuron, this translates to a higher input current that drives the neuron to produce a more rapid series of spikes that is transmitted to another neuron. In biological systems, spikes are transmitted across synaptic connections and transformed to post-synaptic current at the site of a receiving neuron. It is important to note that the inputs to the neuron do not have to be characterized as scalar values, as Equation 1 holds for vector inputs.

We can recover the value represented by populations of
neurons by filtering spike trains with a filter $h(t)$ and scaling with decoding weights $\boldsymbol{d}_{i}$ :

$$
\begin{equation*}
\hat{\boldsymbol{x}}=\sum_{i} a_{i} *\left[\boldsymbol{d}_{i} h\right] . \tag{2}
\end{equation*}
$$

The linear filter $h(t)=\tau_{\text {syn }}^{-1} \exp \left(-t / \tau_{\text {syn }}\right)$ models the postsynaptic current. The symbol $*$ denotes convolution, an operation that places such filter at every position where a spike occurs, and sums the result. The decoding weights $\boldsymbol{d}_{i}$ can be analytically computed by a least-squares minimization of the error term $E=\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|$.

To perform a computation, these decoding weights are coupled with the encoding weights $\boldsymbol{e}$ of the receiving neurons. This gives observable connection weights between two neural populations. Specifically, the connections between neurons in the pre-synaptic population $a_{i}$ and the post-synaptic population $b_{j}$ are computed as $w_{j i}=\alpha_{j} \boldsymbol{e}_{j}^{\top} \boldsymbol{d}_{i}$. The group of receiving neurons can also represent a transformed value $f(\boldsymbol{x})$, where $f$ can be a non-linear function. The same optimization method can be used in this case to compute alternative decoding weights $\boldsymbol{d}_{i}^{f}$ to estimate the function.

## Representing Words and Associations

In our model, the vectors $\boldsymbol{x}$ in Equation 1 are 256-dimensional unit vectors that represent animal words. The vectors are generated randomly such that similarity between any two vectors is generally less than 0.1 . This ensures almost orthogonal vectors, with some overlap in representation, meaning the same neurons will be involved in the representation of different words. An example of this representational overlap can be seen in the spike raster plot in Figure 1B, where some neurons fire for all words and some only for a subset. The NEF methods allow us to decode the spiking activity in terms of the words being represented by the neurons with Equation 2 as shown in the upper part of Figure 1B.

Associative relationships between words are represented as linear transformations implemented in the connections be-

Table 1: Utility calculations for different goals and the corresponding actions.

|  | Goal | Utility calculation | Action |
| :--- | :--- | :---: | :--- |
| 1. | Start | goal $\cdot$ start | Set cue to animal, set goal to think |
| 2. | Think | goal $\cdot$ think + response_magnitude -1 | Copy response to cue, add response to responses, set goal to think |
| 3. | Default | 0.4 | Set cue to animal, set goal to think |

tween two groups of neurons. Word vectors are collected row-wise into a single matrix $\boldsymbol{V}$, and associations between pairs of words are encoded into a matrix $\boldsymbol{A}$ such that $A_{i j}$ is the association strength from word $i$ to word $j$. We can then express a new matrix $\tilde{\boldsymbol{A}}=\boldsymbol{V}^{\top} \boldsymbol{A}^{\top} \boldsymbol{V}$ to implement a transformation that multiplies the vector represented by the first group of neurons by the matrix $\tilde{\boldsymbol{A}}$ and transmits the result to the second group. This operation results in a weighted linear combination of vectors that represents words associated with the word represented in the first group of neurons. This method of representing associations is embedded in a large recurrent network to perform the semantic fluency task.
Association Matrices To construct three different association matrices $\boldsymbol{A}$, we use three different sources of associative data: Free Association Norms (FAN; Nelson, McEvoy, \& Schreiber, 2004), BEAGLE (Jones \& Mewhort, 2007) and Google Ngrams (Michel et al., 2011).

The FAN data set has been derived empirically in a free association experiment, where individuals were asked to generate the first word which comes to their mind for given a cue. The data was normed over all participants to yield asymmetric association strengths for over 5,000 words. The Ngram data set contains co-occurrences of sequences of $n$ words extracted from the Google Books Ngram Viewer dataset (Version 2 from July 2012, Michel et al., 2011). This dataset provides occurrence frequencies of n -grams across over 5 million books published up to 2008. We use occurrences of bi-grams to construct an asymmetric association matrix. The BEAGLE dataset has been trained on a 400M-word Wikipedia corpus, yielding unique vector representations for each word. In this data set, similarity between pairs of vectors is computed as cosine similarity, providing a symmetric measure of association strength. We use pre-computed similarities between pairs of animal word-vectors as in Hills et al. (2012).

We take human responses as a reference for the set of animal words and consider only words that are present in all datasets, amounting to 157 animals. The FAN data set contains the smallest vocabulary and is the most restrictive set.

## Proposed Neural Network Model

Using the NEF implemented in the Nengo simulation environment (Bekolay et al., 2014), we constructed a model consisting of approximately 62,000 LIF neurons organized in functional subgroups performing the semantic fluency task. ${ }^{1}$

[^114]The architecture in Figure 1A shows how networks of neurons are organized and connected to perform the task. The model can be divided into two components: the semantic system and the action selection system. In terms of their biological correlates, the semantic system can be mapped to areas of the medial temporal cortex, and the action selection system to the basal ganglia and the thalamus. The action selection system maintains two possible phases: initializing the task and performing the task.

The initialization phase is active only at the beginning of a simulation, where external input is used to drive the goal ${ }^{2}$ population of neurons to represent the vector start. The second phase consists of performing the task itself, and occurs once a cue is provided.

After the task has been initialized, the action selection system (consisting of the basal ganglia BG and thalamus THAL populations) switches to the process of generating word responses within the semantic system. The recurrent action selection system maintains word generation by simultaneously evaluating utilities of actions and selecting the action with the highest utility value. Table 1 shows the mapping between utility calculations and actions utilized by the action selection system. Since the external input initially sets the goal to start, the action selection system will select the first action due to its high utility value. This action will feed the vector animal as input to the population cue, and set the representation in the goal population to think. This action can be interpreted as the instruction "start listing animals".

Next, the semantic system begins to generate associations of the word animal within the association network. The connection between cue and the association network implements the transformation $\tilde{\boldsymbol{A}}$, as described in the previous section.

The association network then represents a vector which is a linear combination of word-vectors associated with animal. For example, there might be a representation corresponding to the vector: $0.5 *$ cat $+0.4^{*} \mathbf{d o g}+0.1 *$ fish. Coefficients represent association strengths between each individual word and the word animal, as derived from the association matrix A. A winner-take-all (WTA) mechanism within the network selects the vector with the largest coefficient, and projects it to the response population. In this example, the response popu-

[^115]lation would now represent the vector cat.
When a response has been generated, the action selection system selects the second action (see Table 1) due to its high utility value. This action projects the word represented in response (e.g., cat) to cue, simultaneously adding it to the representations stored in response memory. The goal continues to be think.

This process within the semantic system continues, with the action selection system selecting the second action most of the time. To prevent the same responses from re-appearing immediately, response memory is implemented as a neural integrator population. It projects inhibitory connections to association network in order to suppress representations of words previously generated as responses.

The last action with a fixed utility value of 0.4 is selected if utilities of all previous actions have evaluated to a lower value. This occurs when the system is unable to come up with a new response (e.g., the WTA mechanism takes too long to decide between two words). While rare, when this situation occurs, the action selection system sets cue to represent the input animal and the goal is set to think.

## Network Parameters

Most parameters in the model have been left at their default values provided by the simulation software Nengo (Bekolay et al., 2014). Table 2 lists the major parameters in the model. Some parameter values (e.g., maximal firing rates) are selected randomly. Each time the model is run, a new set of such parameters are chosen. Such diversity in parameter settings is a first approximation of differences in cognitive processing that may occur across cortical regions of different individuals.

## Results

We ran 141 simulations of the model for each of the three association matrices (Beagle, Ngram, and FAN) and compared them to human data. The number of simulations corresponds to the number of participants in the study by Hills et al. (2012). The simulations were run until the average number of responses produced matched the average number of responses given by human subjects within three minutes.

For each simulation run, we recorded word responses as decoded vector representations in the response population, and inter-item response times (IRT) as times between the onset of the current response and the previous response. Here we consider only relative timings (i.e., the time differences between responses), as mapping to absolute timing (i.e., exact duration of the experiment) would require consideration of the time it takes for other processes to occur, such as visual perception and motor responses, which are not part of this model.

The model responses were evaluated using the same scripts developed for the analysis of the human data, provided in Hills et al. (2012). Each response is assigned an animal category, and the clusters are identified as sequences of responses
within the same category. An animal that could be assigned to two clusters is assigned to both. ${ }^{3}$

The first analysis compares the pairwise similarity of a word and the words preceding it within a cluster (Figure 2A). The similarity is computed as a dot product between two BEAGLE vectors corresponding to the two words in a word pair (Hills et al., 2012). The experimental results in Figure 2A show that the word most similar to the recent word in the patch is the one preceding it, supporting the theory of locality in a memory structure. For the model, this trend is observed with the Ngram and the FAN association matrices, and less so with the BEAGLE association matrix, for which the similarity appears to have a flat trend independent of the position in the cluster.

The second analysis compares the pairwise similarity of subsequent items relative to the position of an item in the cluster (Figure 2B). Human data shows that the lowest pairwise similarity occurs at the cluster transition points, indicated by ' 1 ' on the $x$-axis in the figure. That point shows the similarities between the first word in a cluster and the last word in the preceding cluster. For humans, the mean similarity $\mu$ at the cluster switch is $\mu=0.92$ with standard deviation $\sigma_{\mu}=0.01$. The model using FAN data shows comparable results $\left(\mu=0.93, \sigma_{\mu}=0.01\right)$. For the Ngram and the BEAGLE association matrices this effect is weakly observable $\left(\mu=1.00, \sigma_{\mu}=0.01\right.$ and $\mu=1.01, \sigma_{\mu}=0.01$, respectively), as the word similarity at the transition point remains above the subject's average.

The third analysis concerns the position of a word item within a cluster and the speed of generating a word. The ratio between the average IRT for an item and the participant's mean IRT over the entire task is shown in Figure 2C. Human participants take the most time to produce the first word in a new cluster (reported $t(140)=13.14, p<.001$ ) and least time to produce the second word in a new cluster $($ reported $t(140)=11.92, p<.001)$. This observation is the hallmark prediction of the optimal foraging strategy, suggesting that cluster switches occur when the current IRT increases over the mean IRT value. Figure 2C also shows that the model using the FAN association matrix exhibits the same effects as observed with human responses. It takes significantly more time to generate the first words in a new cluster $(t(140)=4.78, p<.001)$, compared to the second words in the cluster $(t(140)=4.78, p<.001)$.

## Discussion

We have proposed a spiking neural network model that performs the semantic fluency task and shows a good match with human behavioral data. In particular, we embed association data in connections between neurons within a large recurrent network and investigate which source of association information provides the closest match to human performance. Our focus is on identifying plausible, causal neural mechanisms

[^116]Table 2: List of model parameters

| Name | Value (unit) | Explanation |
| :---: | :---: | :--- |
| $d$ | 256 | Dimensionality of word vectors |
| assoc_th | 0.3 (or 0.25) | Default WTA input threshold (Ngram, BEAGLE threshold) |
| $c_{\mathrm{cs}}$ | 3 | Cue to association network connection strength |
| $c_{\mathrm{fs}}$ | 0.2 | Cue feedback connection strength |
| $c_{\mathrm{inh}}$ | -5 | Response memory to association network inhibitory connection strength |
| $\tau_{\mathrm{syn}}$ | 0.1 ms | Synaptic time constant between association network and response |
| $\tau_{\mathrm{syn}}$ | 0.005 ms | Synaptic time constant (default) |
| max_rate | $200-400 \mathrm{~Hz}$ | Range for maximal neural firing rates (default) |



Figure 2: Comparison between model responses for FAN, Ngram and BEAGLE association matrices (blue) and human responses (yellow, reproduced from Hills et al., 2012). A: Pairwise similarity between a word and the words preceding it within the same categorical cluster. B: Pairwise similarity between subsequent words. For example, the bars above ' 1 ' indicate the relative pairwise similarities between the first item in a cluster, and the last item in the previous cluster. $\mathbf{C}$ : Inter-item response times (IRT) between subsequent words. Standard errors of the mean are shown with error bars in all plots.
for performing such tasks. To that end, we have identified computational requirements in terms of processes and relevant neural parameters, and here we discuss how they affect the model's behavior.

The model produces responses in a way that is consistent with predictions made by optimal foraging theory proposed to be used by humans (Hills et al., 2012). It is more likely to switch animal categories when the average similarity of subsequent responses drops below, or gets close to, the overall mean similarity. This effect was observed with all three association matrices, but is most pronounced with the FAN
matrix.
However, the analysis of timing effects allowed us to clearly distinguish between the three matrices. The model using FAN exhibited the same timing effects as observed with human responses. This timing effect was not observed with other association matrices (see Figure 2C). The similarity between cognitive processes involved in free association task and in the semantic fluency task (Jones et al., 2015) is a likely candidate to explain the effectiveness of free norms in matching the experimental data. However, this result could also be seen as support for the plausibility of the proposed neu-
ral mechanisms, as they are able to generate behaviors in accordance with these underlying associations. We expect that a better understanding of cognitive processes involved in free associations could aid understanding of the processes underlying semantic fluency. Our model may prove useful in exploring a variety of possible ways that such associations are neurally realized, as the direct embedding in connection weights as done here is only one possibility.

When building biologically constrained neural models, timing is a highly constrained property of a model. Here, the timing of responses is sensitive to both neural time constants and our characterization of concept representation. This is in contrast to previous models that directly use semantic networks, where timing is a separate and independent parameter. For instance, we identified that a longer synaptic time constant was needed between the association network and the response populations to stabilize the representation. This leads to the prediction that this network will be rich with NMDA receptors in the biological system. These receptors have significantly longer time constants than the more common AMPA receptors. Also, NMDA receptors can be found in the hippocampus, a brain structure located in the medial temporal lobe, whose function has been implicated in semantic and episodic memory.

Our characterization of neural concept representation also has an effect on the timing responses. Specifically, we have observed that the dimensionality of employed vector representations needed to be sufficiently large to achieve experimentally observed timing effects. While we find that $d=256$ suffices for this purpose, a systematic search of dimensionality effects on the performance is needed to see how it affects the behavior. We have tested this model with lower values (e.g., $d=64$ ) and it produced results in support of local search strategy, yet it failed to provide a good match with the timing data. In other work, we have suggested that $d \approx 500$ is necessary for representing human-scale conceptual structures (Eliasmith, 2013), which is consistent with this newer observation.

## Acknowledgments

This work has been supported by AFOSR, grant number FA9550-17-1-0026, the Canada Research Chairs program, the NSERC Discovery grant 261453, Air Force Office of Scientific Research grant FA8655-13-1-3084, CFI and OIT.

## References

Abbott, J. T., Austerweil, J. L., \& Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. Psyc. Rev., 122(3), 558.
Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., ... Eliasmith, C. (2014). Nengo: A Python tool for building large-scale functional brain models. Front. in Neuroinformatics, 7(48).
Bousfield, W., \& Sedgewick, C. (1944). An analysis of sequences of restricted associative responses. The Journal of General Psychology, 30(2), 149-165.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. Theoretical population biology, 9(2), 129-136.
Collins, A. M., \& Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psyc. Rev., 82(6), 407-428.
Collins, A. M., \& Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8(2), 240-247.
Eliasmith, C. (2013). How to build a brain: A neural architecture for biological cognition. New York, NY: Oxford University Press.
Eliasmith, C., \& Anderson, C. H. (2003). Neural engineering: computation, representation, and dynamics in neurobiological systems. Cambridge, MA: MIT Press.
Hills, T. T., Jones, M. N., \& Todd, P. M. (2012). Optimal foraging in semantic memory. Psyc. Rev., 119(2), 431.
Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., \& Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453-458.
Jones, M. N., Hills, T. T., \& Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). Psyc. Rev..
Jones, M. N., \& Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. Psyc. Rev., 114(1), 1.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . others (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176-182.
Nelson, D. L., McEvoy, C. L., \& Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, \& Computers, 36(3), 402-407.
Nematzadeh, A., Miscevic, F., \& Stevenson, S. (2016). Simple search algorithms on semantic networks learned from language use. In 38th Cog. Sci. Proceedings (pp. 13131318). Austin, TX: Cognitive Science Society.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. Nature Rev. Neurosci., 13(8), 587-597.
Rissman, J., \& Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. Annual Rev. of Psyc., 63, 101-128.
Steyvers, M., \& Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cog. Sci., 29(1), 41-78.
Thurstone, L. (1938). Primary mental abilities (No. 1). University of Chicago Press.
Troyer, A. K., Moscovitch, M., \& Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology, 11(1), 138.
Tulving, E. (1983). Elements of Episodic Memory. Oxford, UK; New York: Oxford University Press.

# Mechanisms of overharvesting in patch foraging 

Gary A. Kane (gkane@princeton.edu), Aaron M. Bornstein (aaronmb@princeton.edu), Nathaniel D. Daw (ndaw@princeton.edu), Jonathan D. Cohen (jdc@princeton.edu)<br>Department of Psychology and Neuroscience Institute, Princeton University<br>Princeton, NJ 08544 USA

Robert C. Wilson (bob@email.arizona.edu)
Department of Psychology, University of Arizona
Tucson, AZ 85721 USA

Amitai Shenhav (amitai_shenhav@brown.edu)<br>Department of Cognitive, Linguistic, and Psychological Sciences, Brown University<br>Providence, RI 02912 USA


#### Abstract

Serial stay-or-search problems are ubiquitous across many domains, including employment, internet search, mate search, and animal foraging. For instance, in patch foraging problems, animals must decide whether to stick with a depleting reward vs search for a new source. The optimal strategy in patch foraging problems, described by the Marginal Value Theorem (MVT; Charnov, 1976), is to leave the depleting patch when the local reward rate within a patch matches the overall long-run reward rate. Many species of animals, ranging from birds to rodents, monkeys, and humans, adhere to this policy in important respects, but tend to overharvest, or stick with the depleting resource too long. Here we attempt to determine the cognitive biases that underlie overharvesting in one of these species (the rat). We characterized rat behavior in response to two basic manipulations in patch foraging tasks: to travel time between patches and depletion rate, and two novel manipulations to the foraging environment: the size of reward and length of delays, and placement of delays (pre- vs. post-reward). In response to the basic manipulations, rats qualitatively followed predictions of MVT, but stayed in patches for longer than is predicted. In the latter two manipulations, rats deviated from predictions of MVT, exhibiting changes in behavior not predicted by MVT. We formally tested whether four separate cognitive biases - subjective costs, decreasing marginal utility for reward discounting of future reward, and ignoring postreward delays - could explain overharvesting in the former two manipulations and deviations from MVT in the latter two. All of the biases tested explained overharvesting behavior in the former contexts, but only one bias - in which rats ignore post-reward delays - also explained deviations from MVT due to larger rewards with longer delays and due to introduction of a pre-reward delay. Our results show that multiple biases can explain certain aspects of overharvesting behavior, and, while foraging behavior may be the result of the use of multiple biases, inaccurate estimation of postreward delays likely contributes to overharvesting.


Keywords: foraging; decision-making, subjective utility; delay discounting

## Introduction

Patch foraging refers to situations in which one must decide when to leave a depleting resource patch to search for a
new, likely richer one, that comes at the cost of time and/or effort. The optimal solution in patch foraging is given by the Marginal Value Theorem (MVT; Charnov, 1976): leave when the local reward rate within a patch depletes below the global reward rate across all patches - the average reward rate for the environment. MVT makes two main predictions: i) in patches that contain more reward than average, stay longer to exploit such reward, and ii) when the cost of searching for a new patch is greater (e.g. the time or effort required to travel to a new patch is greater), stay longer in all patches. Many animals, ranging from invertebrates to birds to mammals, qualitatively follow predictions of MVT (Stephens \& Krebs, 1986). However, in most tests, animals, including rats, monkeys, and humans, tend to stay in patches longer than is predicted by MVT (Constantino \& Daw, 2015; Hayden, Pearson, \& Platt, 2011; Nonacs, 1991; Stephens \& Krebs, 1986).

Hypotheses to explain overharvesting include common biases in intertemporal choice, such as i) subjective costs, such as an aversion to leaving the patch (Carter \& Redish, 2016; Wikenheiser, Stephens, \& Redish, 2013); ii) decreasing marginal utility in which large rewards available in a new patch are not viewed as proportionally larger than the smaller, depleted rewards available in the current patch (Constantino \& Daw, 2015); iii) discounting future rewards, in which the value of large rewards available in a new patch are discounted by virtue of being available later, above and beyond the time it takes to travel to the new patch (Blanchard, Pearson, \& Hayden, 2013; Carter \& Redish, 2016; Constantino \& Daw, 2015); and iv) ignoring postreward delays, which causes overestimation of reward rate within the patch due to inaccurate estimation of the time taken to obtain reward (Bateson \& Kacelnik, 1996; Blanchard et al., 2013; Carter \& Redish, 2016; Gallistel \& Gibbon, 2000; Kacelnik, 1997). Although overharvesting is widely observed, there have been few direct investigations into the underlying mechanisms. In this paper, we directly test these hypotheses to rat foraging behavior in an operant chamber based patch foraging task.

First, we characterized rat foraging behavior in response to four manipulations to the foraging environment: to travel
time between patches, rate of reward depletion within patches, scale of reward size and length of delay, and placement of delays (pre- vs. post-reward). Next, we fit formal models representing the four hypotheses to rats' behavior to examine how well each hypothesis explained foraging behavior across all manipulations.

## Methods

## Animals

Adult Long-Evans rats were used (Charles River, Kingston, $\mathrm{NY} ; \mathrm{n}=8$ ). Rats were housed on a reverse $12 \mathrm{~h} / 12 \mathrm{~h}$ light/dark cycle (lights off at 7 A.M.). All behavioral testing was conducted during the dark period. Throughout behavioral testing, rats were food restricted to maintain a weight of $85-90 \%$ ad-lib feeding weight, and were given adlib access to water. All procedures were approved by the Princeton University Institutional Animal Care and Use Committee.

## Operant Foraging Task

This task simulated foraging in a patchy environment, resembling the task used with monkeys by Hayden et al (2011). On a series of trials performed in a standard operant chamber (Med Associates, St. Albans, VT), rats had to repeatedly decide to stay in a patch to continue harvesting a depleting reward source or leave the patch to travel to a new, full patch, incurring a cost of time to travel to a new patch. Rats' decided to harvest from a patch by pressing an activated lever on one side of the front of the chamber, or to travel to a new, full patch by nosepoking at the back of the chamber and then returning to a newly activated lever on the other side of the front of the chamber. To cue the beginning of a trial, lights above the activated lever and the nosepoke illuminated, indicating that the rat could decide to harvest reward from the activated patch (lever press) or to travel to a new patch (nosepoke). The time from start of trial to the decision was recorded as decision time (DT). If the rat pressed the lever to harvest from the activated patch, a cue light turned on in the reward magazine next to the lever following a short handling time delay (HT), and liquid sucrose was delivered when the rat's head entered the magazine. An intertrial interval (ITI) began as soon as the rat entered the reward magazine. To control the reward rate within the patch, the length of the ITI was adjusted based on the DT of the current trial, such that the length of all lever press trials was equivalent. If the rat nosepoked to leave the patch, the lever retracted for a delay period, simulating the travel time, after which, the lever on the opposite side of the chamber extended, representing a new patch from which the rat could harvest.

Each manipulation (travel time, depletion rate, scale, and delay placement) was conducted in separate experiments, with two conditions in each experiment. Rats were trained on each condition for 5 days, and tested for a subsequent 5 days. Conditions within each experiment were
counterbalanced. Details regarding reward size and timing for each experiment can be found in Figure 1. T-tests or ANOVA with repeated measures were used to compare the number of harvests per patch, a proxy for time in the patch, across conditions.

## Models

All models were constructed as Markov Decision Processes. States were represented as trials within patches. A decision to stay in the patch (i.e. harvest from the patch) provided reward for staying in state $s, r_{\text {stay }, s}$, and caused transition to state $s+1$. A decision to leave the patch resulted in travel time delay, $\tau$, followed by the first reward in the next patch, $r_{\text {leave }}$, and associated ITI following the reward, $I T I_{\text {leave }}$. We fit three models based on MVT: a model incorporating a constant subjective cost (subjective cost), a model that accounted for diminishing marginal returns for larger rewards (subjective utility), and a model ignoring postreward delays, as well as a delay discounting model.

For each of the MVT models, we calculated the value of staying in the patch in state $\mathrm{s}, V_{\text {stay }, s}$, as the reward rate within the patch,

$$
\begin{gathered}
T_{s t a y, s}=D T_{s t a y, s}+H T+I T I_{s t a y, s}, \\
V_{s t a y, s}=\frac{r_{s t a y, s}}{T_{s t a y, s}},
\end{gathered}
$$

and the value of leaving the patch in state $\mathrm{s}, V_{\text {leave, } s}$, as the cumulative reward rate across patches,

$$
\begin{gathered}
T_{\text {leave }, s}=D T_{\text {leave } s}+\tau+I T I_{\text {leave }}, \\
R_{s}=p_{\text {stay }, s} * r_{\text {stay }, s}+p_{\text {leave }, s} * r_{\text {leave }}, \\
T_{s}=p_{\text {stay }, s} * T_{\text {stay }, s}+p_{\text {leave }, s} * T_{\text {leave }, s}, \\
V_{\text {leave }}=\frac{\Sigma_{s} p_{s} * R_{s}}{\Sigma_{s} v_{s} * T_{s}}
\end{gathered}
$$

where $R_{s}$ and $T_{s}$ was the average reward and average time for state $s, p_{\text {stay,s }}$ was the probability of choosing to stay in state $s$, and $p_{s}$ was the probability of being in state $s$. Optimal behavior was to leave the patch when $V_{\text {leave }}>=V_{\text {stay }}$ (i.e. when the long-run average reward rate is greater than the local reward rate in the patch). To model rats' behavior, patch leaving distributions were assumed to be normally distributed with respect to $V_{\text {leave }}-V_{\text {stay }}$, with mean $\mu=0$ (i.e. $\left.V_{\text {leave }}=V_{\text {stay }}\right)$ and variance $\sigma^{2}$, a free parameter.

To account for subjective costs, a constant, $c$, representing an aversion to leaving the patch, was added to the model, such that the patch leaving distribution was normally distributed with respect to $V_{\text {leave }, s}-V_{\text {stay,s }}-c$.

For the subjective utility model, the utility for taking action $a$ in state $s$ increased monotonically, but sublinear to the size of the reward, according to a power utility function, dependent on a free parameter, $\eta$,

$$
\begin{gathered}
u_{a, s}=r_{a, s}^{\eta} \\
V_{s t a y, s}=\frac{u_{s t a y, s}}{T_{\text {stay }, s}}, \\
U_{s}=p_{s t a y, s} * u_{\text {stay }, s}+p_{\text {leave }, s} * u_{\text {leave }},
\end{gathered}
$$

$$
V_{\text {leave }}=\frac{\sum_{s} p_{s}^{*} U_{s}}{\sum_{s} p_{s}^{*} T_{s}}
$$

For the ignoring post-reward delays model, delays that occur after receiving reward, but before a decision was made on the next trial (e.g. ITI after reward and DT prior to making next decision), $T_{\text {post }}$, were treated differently than time delays that occured between the decision and receiving a reward (e.g. handling time delay between lever press and reward, or travel time delay between nosepoke and first reward in the next patch). We tested multiple functions for how post-reward delays might have been treated, all in which the increase in perceived time increased monotonically, but sublinear to actual time, including a linear function with slope $<1$, a power function, and an exponential function. The exponential function provided the best fit across all experiments, and was used for further testing:

$$
\begin{gathered}
T_{\text {post }, \text { stay }}=\frac{1-e^{-\alpha^{*}\left(D T_{s t a y, s}+I T I_{\text {stay }, s}\right)}}{\alpha} \\
T_{\text {post,leave }}=\frac{1-e^{-\alpha^{*}\left(D T_{\text {leave }}+I T I_{\text {leave })}\right)}}{\alpha} \\
V_{\text {stay }, s}=\frac{r_{\text {stay }, s}}{H T+T_{\text {post }}}, \\
T_{s}=p_{\text {stay }, s} *\left(H T+T_{\text {post,stay }}\right)+p_{\text {leave }, s} *\left(\tau+T_{\text {post }, \text { leave }}\right) \\
V_{\text {leave }}=\frac{\Sigma_{s} p_{s}{ }^{*} R_{s}}{\sum_{s} p_{s} T_{s} T_{s}}
\end{gathered}
$$

Whereas MVT optimizes all future reward, the delay discounting model, a hyperbolic discounting model, optimizes discounted future reward (i.e. it similarly optimizes future reward, but with less weight to rewards that occur further in the future):

$$
\begin{gathered}
d(t, k)=1 /\left(1+k^{*} t\right), \\
V_{\text {stay }, s}=d(H T, k)\left[r_{\text {stay }, s}+d\left(I T I_{\text {stay }, s}, k\right) \sum_{s^{\prime}} p\left(s^{\prime} \mid \text { stay }, s\right) \sum_{a} p_{a, s^{\prime}} V_{a, s^{\prime}}\right] \\
V_{\text {leave }, s}=d(\tau, k)\left[r_{\text {leave }}+d\left(I T I_{\text {leave }}, k\right) \sum_{s^{\prime}} p\left(s^{\prime} \mid \text { stay }, s\right) \sum_{a} p_{a, s^{\prime}} V_{a, s^{\prime}}\right]
\end{gathered}
$$

where $d(t, k)$ was the hyperbolic discount function of time $t$, with a free parameter, $k . p\left(s^{\prime} \mid a, s\right)$ was the conditional probability of being in future state $s$ ' given action $a$ was taken in state s, $p_{a, s^{\prime}}$ was the probability of taking action $a$ in future state $s$ ', and $V_{a, s^{\prime}}$ was the value of for taking action $a$ in future state $s^{\prime}$.

As the discount parameter, $k$, approached zero (no discounting of future reward), this model converged to MVT; that is, it sought to maximize all future reward. As $k$ increases, future rewards are discounted, such that i) the value of large rewards in a new patch are discounted above and beyond the travel time between patches, and ii) the model sought to maximize reward into the future, but over shorter periods of time.

For all models, one set of parameters was fit to each animal per experiment, to maximize the likelihood of the data from both conditions in that experiment. To test whether the model could explain rat overharvesting behavior in each experiment, we generated predicted patch leaving distributions from the best fit parameters for each
model, then perform a repeated measures ANOVA, to test whether there is an interaction between model predictions and observed behavior (i.e. whether the effect of each experimental manipulation was different between model predictions and observed behavior).

## Results

## Foraging Behavior

Rats were first tested on a manipulation of travel time. With longer travel time, the long-run average reward rate is lower, thus MVT predicts rats should stay in patches longer. Within behavioral sessions, rats encountered three different patch types, which started with varying amount of reward ( 60,90 , or $120 \mu \mathrm{~L}$ ) and depleted by the same rate $(8 \mu \mathrm{~L})$. Between sessions, rats were tested on either a 10 s or 30 s travel time delay following their decision to leave the patch. As predicted by MVT, rats stayed longer in patch types that started with larger reward volume, indicated by more harvests per patch, $\mathrm{F}(2,14)=3145, \mathrm{p}<.001$, and rats stayed longer in all patches with longer travel time, $\mathrm{F}(1,7)=71.4$, $\mathrm{p}<.001$. However, rats overharvested, staying longer in all patches than is predicted by MVT (Figure 1A).

Next, rats were tested on a manipulation of depletion rate. Quicker reward depletion causes the local reward rate to deplete to the long-run average reward rate quicker, such that MVT predicts earlier patch leaving. Within sessions, rats encountered a single patch type (starting volume of 90 $\mu \mathrm{L}$ ), which depleted at a rate of either 8 or $16 \mu \mathrm{~L} /$ trial, tested between sessions. As predicted by MVT, rats left patches earlier when they depleted more quickly, $t(7)=$ $15.835, \mathrm{p}<.001$. But, again, rats stayed in patches longer than is predicted by MVT (Figure 1B).

Rats were then tested on a manipulation of the scale of rewards and time. In one condition, the size of rewards and length of delays was twice that of the other: patches started with 90 or $180 \mu \mathrm{~L}$ of reward, depleted at a rate of 8 or 16 $\mu \mathrm{L} /$ trial, and travel time between patches was 10 or 20 s . Both reward rate within the patch and reward rate across patches were equivalent in the two conditions; thus, MVT predicts no change in behavior. Contrary to predictions of MVT, rats stayed in patches significantly longer when given larger rewards with longer delays, $t(7)=10.039, p<.001$. And, again, rats overharvested in both conditions (Figure 1C).

Lastly, rats were tested on a manipulation of the placement of delays. In one condition, rats experienced no pre-reward delay, and a long post-reward delay (ITI $\sim 10 \mathrm{~s}$, adjusted based on DT). In the other condition, rats experienced a 3 s pre-reward delay, and shorter post-reward delay (ITI $\sim 7 \mathrm{~s}$ ). The duration of each trial did not change, so both the local reward rate within the patch and long-run average reward rate across patches were equivalent between the conditions, and MVT predicts no change in behavior. Rats overharvested in both conditions, but they left patches earlier when part of the delay occurred prior to the reward, $t(7)=7.453, p<.001$ (Figure 1D).


Figure 1: Diagram of each foraging experiment and behavioral data. In diagrams, black boxes represent the start of a trial, at which a decision to lever press or nosepoke must be made. $\mathrm{DT}=$ decision time, $\mathrm{HT}=$ handling time, ITI $=$ intertrial interval. In graphs, black points and lines represent rat data, and red points and lines the optimal behavior predicted by MVT. A) Points represent the mean number of lever presses in each patch from each animal, error bars representing standard error are obstructed by the points. B-D) Each point is the mean number of lever presses in each patch for a single rat, with lines connecting each rats behavior in the two conditions.

## Models of overharvesting

We first tested a model that includes a subjective cost to foraging - a constant that represents a bias towards staying in the patch. Predictions from the model, fit to each rat, are presented in Figure 2. Qualitatively, this model explained rat behavior on the travel time and depletion rate experiments well, producing a predicted number of harvests per patch similar to that exhibited by the rats. However, there was a significant interaction between travel time and predicted vs. observed behavior, $\mathrm{F}(1,7)=7.391, \mathrm{p}=.030$, indicating a difference between how the model vs. the rats responded to the change in travel time. This is likely driven by the model predicting slightly earlier patch leaving in the 30 s travel time relative to rats' behavior. The interaction between depletion rate and predicted vs. observed behavior was not significant, $\mathrm{F}(1,7)=.124, \mathrm{p}=.735$.

As this model only allows for a constant change in the reward rate threshold to leave patches, it is unlikely to account for behavior in which rats select a different threshold between conditions. When rats were given longer rewards with longer delays, they stayed in patches longer, allowing patches to deplete to a lower reward rate before leaving. Similarly, when a pre-reward delay was introduced, rats left patches earlier, at a higher reward rate. The model failed to account for both of these effects (interaction between scale $x$ predicted vs. observed behavior, $\mathrm{F}(1,7)=$ $58.43, \mathrm{p}<.001$; delay x predicted vs. observed behavior, $F(1,7)=48.79, p<.001)$.


Figure 2: Predictions of the subjective cost model for the A) travel time, B) depletion rate, C) scale, and D) pre- vs. post-reward delay experiments. Black points and errorbars represent the mean number of harvests per patch $\pm$ standard error. Colored lines represent the average model predicted number of harvests. The width of the colored line represents the standard error of the predicted number of harvests. There were significant interactions between model predictions and observed behavior in the travel time (A), scale (C), and pre vs. post-reward delay (D) experiments.

We next tested whether diminishing marginal returns could explain overharvesting (Figure 3). Under this
hypothesis, large rewards in a new patch were not valued as proportionally larger to smaller rewards in the current, depleting patch. Predictions from the subjective utility model are presented in Figure 3. As did the subjective cost model, the subjective utility model qualitatively explained overharvesting behavior in the travel time and depletion rate experiments. This was supported by the lack of a significant interaction between travel time and predicted vs. observed behavior, $\mathrm{F}(1,7)=4.501, \mathrm{p}=.072$, although there was a significant interaction between depletion rate and predicted vs. observed, $\mathrm{F}(1,7)=14.12$ ), $\mathrm{p}=.007$.

In the scale experiment, the subjective utility model should estimate a lower reward rate in the environment with larger rewards, and thus predict later patch leaving. However, this model could not explain both general overharvesting, as well as the change in behavior due to scale, $\mathrm{F}(1,7)=112, \mathrm{p}<.001$. Additionally, this model is insensitive to the placement of delays, and failed to predict that rats would leave patches earlier when a pre-reward delay was introduced, $\mathrm{F}(1,7)=77.22, \mathrm{p}<.001)$.


Figure 3: Predictions of the subjective utility model for the A) travel time, B) depletion rate, C) scale, and D) prevs. post-reward delay experiments. There were significant interactions between model predictions and observed behavior in the depletion rate (B), scale (C), and pre- vs. post-reward delay (D) experiments.

Next, we tested whether a delay discounting model that considers future rewards could account for rat overharvesting behavior (Figure 4). As rewards are discounted into the future, the value of the first reward in a new patch was discounted due to the travel time between patches, and the model sought to maximize future rewards over a shorter period of time. The discounting model accurately predicted overharvesting behavior in both travel times; interaction between travel time and predicted vs. observed behavior was not significant, $\mathrm{F}(1,7)=.050$, $\mathrm{p}=$ .830. This model also predicted earlier patch leaving when reward in the patch depleted quicker, but there was significant interaction between depletion rate and predicted vs. observed behavior, $F(1,7)=16.780, p=.005$, indicating that the model-predicted change in behavior is different
from the change in behavior exhibited by rats.
In the scale experiment, when comparing larger rewards with longer delays to smaller rewards with shorter delays, the larger rewards would be discounted to a greater extent. Thus, in this model, the estimate of the overall reward rate would be lower in the environment with larger rewards, predicting that rats would stay longer in this environment. Indeed, this model did predict that rats would stay in patches longer when given larger rewards with longer delays, and the interaction between scale and predicted vs. observed behavior was not significant, $\mathrm{F}(1,7)=.482, \mathrm{p}=.510$. This model also should place lower value on rewards in the patch when there is a longer delay between decision to harvest and obtaining reward. However, there was a significant interaction between pre- vs. post-reward delay conditions and predicted vs. observed behavior, $\mathrm{F}(1,7)=34.650$, $\mathrm{p}<$ . 001 .


Figure 4: Predictions of the delay discounting model for the A) travel time, B) depletion rate, C) scale, and D) prevs. post-reward delay experiments. There were significant interactions between model predictions and observed behavior in the depletion rate (B) and pre- vs. post-reward delay (D) experiments.

Lastly, we tested whether ignoring post-reward delays could explain rats' overharvesting behavior. In this model, time delays that occur after receiving reward, before a decision is made on the next trial (e.g. ITI after reward and DT prior to making next decision), were treated differently than time delays that occur between making a decision and receiving a reward (e.g. handling time delay between lever press and reward, or travel time delay between nosepoke and first reward in the next patch). Time delays that occur after the reward, and before the next decision are assumed to increase monotonically, but sublinear relative to actual time, according to an exponential function. In this model, underestimation of the ITI would cause overestimation of reward rate, and overharvesting. Additionally, in the scale experiment, longer delays would cause greater overestimation of reward rate, and would predict that rats should stay in patches longer when given larger rewards with longer delays. In the pre- vs. post-reward delay
experiment, when the pre-reward delay was introduced, post-reward delays were shorter. In this model, shorter postreward delays would lead to less overestimation of reward rate, and earlier patch leaving.

This model qualitatively explained overharvesting in all four experiments. Additionally, there were no significant interactions between task manipulations and predicted vs. observed behavior (travel time, $\mathrm{F}(1,7)=.416, \mathrm{p}=.539$; depletion rate, $F(1,7)=4.691, p=.067$; scale of reward and time, $\mathrm{F}(1,7)=.047, \mathrm{p}=.835$; pre- vs. post-reward delay, $\mathrm{F}(1,7)=1.985, \mathrm{p}=.202$ ), indicating that there were no differences between rats change in behavior due to experimental manipulation and model predicted change in behavior in all four experiments.


Figure 5: Predictions of the ignore post-reward delays model for the A) travel time, B) depletion rate, C) scale, and D) pre- vs. post-reward delay experiments. Interactions between model predictions and observed behavior were not significant in any of the four experiments.

## Discussion

We characterized patch foraging behavior of one of these species, rats, in a variety of foraging environments, and examined the computational mechanisms of overharvesting. We found that rats, like humans (Constantino \& Daw, 2015), followed the primary qualitative predictions of MVT, leaving patches earlier when the rate of depletion was quicker, and staying longer in patches when travel time was longer. However, as has consistently been observed in other species, they overharvested (or stayed longer in patches than is predicted by MVT). Furthermore, rats deviated from predictions of MVT in other ways, staying longer in patches that provided larger rewards with longer delays, and leaving patches earlier when delays occurred between the decision to harvest from the patch and receiving reward. To examine the cognitive biases that underlie overharvesting, we fit four models to rats foraging behavior in each context: a model including subjective costs, diminishing marginal returns for larger rewards, discounting of future reward, and ignoring post-reward delays, and tested whether predictions of these models were different from rats' behavior. All four models could qualitatively explain rat foraging behavior in response
to a change in travel time and patch depletion rate, but only the 'ignore post-reward delays' model, in which post reward delays are perceived to be shorter than they actually are, could predict both later patch leaving when given larger rewards with longer delays, and earlier patch leaving when a pre-reward delay was introduced. These results suggest that there are multiple cognitive biases that can explain overharvesting in certain contexts, and that foraging behavior may be the result of the use of multiple biases. However, inaccurate estimation of post-reward delays likely contributes to overharvesting.

## References

Bateson, M., \& Kacelnik, a. (1996). Rate currencies and the foraging starling: The fallacy of the averages revisited. Behavioral Ecology, 7(3), 341-352.
Blanchard, T. C., Pearson, J. M., \& Hayden, B. Y. (2013). Postreward delays and systematic biases in measures of animal temporal discounting. Proceedings of the National Academy of Sciences of the United States of America, 110(38), 15491-6.
Carter, E. C., \& Redish, A. D. (2016). Rats value time differently on equivalent foraging and delaydiscounting tasks. Journal of Experimental Psychology: General, 145(9), 1093-1101.
Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. Theoretical Population Biology.
Constantino, S. M., \& Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. Cognitive, Affective, \& Behavioral Neuroscience.
Gallistel, C. R., \& Gibbon, J. (2000). Time, rate, and conditioning. Psychological Review, 107(2), 289-344.
Hayden, B. Y., Pearson, J. M., \& Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. Nature Neuroscience, 14(7), 933-939.
Kacelnik, A. (1997). Normative and descriptive models of decision making: time discounting and risk sensitivity. In Characterizing human psychological adaptations (Vol. 208, pp. 51-66).
Nonacs, P. (1991). State dependent behavior and the Marginal Value Theorem. Behavioral Ecology, 12(1), 71-83.
Stephens, D. W., \& Krebs, J. R. (1986). Foraging Theory. Evolutionary Behavioral Ecology (Vol. 121).
Wikenheiser, A. M., Stephens, D. W., \& Redish, a D. (2013). Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task. Proceedings of the National Academy of Sciences of the United States of America, 110(20), 8308-13.

# Language-users choose short words in predictive contexts in an artificial language task 

Jasmeen Kanwal (jasmeen.kanwal@ed.ac.uk), Kenny Smith (kenny.smith@ed.ac.uk), Jennifer Culbertson (jennifer.culbertson@ed.ac.uk), and Simon Kirby (simon@ling.ed.ac.uk)<br>Centre for Language Evolution, University of Edinburgh<br>3 Charles St., Edinburgh, Scotland EH8 9AD


#### Abstract

Zipf (1935) observed that word length is inversely proportional to word frequency in the lexicon. He hypothesised that this cross-linguistically universal feature was due to the Principle of Least Effort: language-users align form-meaning mappings in such a way that the lexicon is optimally coded for efficient information transfer. However, word frequency is not the only reliable predictor of word length: Piantadosi, Tily, and Gibson (2011) show that a word's predictability in context is in fact more strongly correlated with word length than word frequency. Here, we present an artificial language learning study aimed at investigating the mechanisms that could give rise to such a distribution at the level of the lexicon. We find that participants are more likely to use an ambiguous short form in predictive contexts, and distinct long forms in surprising contexts, only when they are subject to the competing pressures to communicate accurately and efficiently. These results support the hypothesis that language-users are driven by a least-effort principle to restructure their input in order to align word length with information content, and this mechanism could therefore explain the global pattern observed at the level of the lexicon.


Keywords: Information theory; Efficient communication; Artificial language learning; Uniform Information Density

## Introduction

Zipf (1935) observed that word length tends to be inversely proportional to word frequency in the lexicon. He hypothesised that this widespread cross-linguistic pattern was due to the Principle of Least Effort: language-users align formmeaning mappings in such a way that effort is minimised while expressivity is still maintained. However, word frequency is not the only reliable predictor of word length. Using corpora from 11 different languages, Piantadosi et al. (2011) show that a word's predictability in context (where they define context as the two words preceding the target word) is even more strongly correlated with word length than frequency is: words that are, on average, more predictable in context tend to be shorter.

Measuring how predictable or unpredictable a word is in a particular context gives us a way of defining the information content of a word. For example, consider the two sentences:
(1) The early bird catches the worm.
(2) Our early bird special today is a baked-apple worm.

In sentence (1), a well-known proverb, the word worm is entirely predicted by the preceding words. The word itself thus gives us practically no new information, and so it has low information content. In sentence (2), the same word is highly unlikely given the preceding words, and thus we find it surprising. This element of surprise is associated with high information content.

Using these concepts, we can apply Zipf's Principle of Least Effort to hypothesise that a speaker's drive to reduce effort will be directed towards words that are already highly predictable given the context, i.e. have low information content. Words that are more surprising in a particular context will be less likely to be reduced, or more likely to be lengthened. The resulting state in which low-information words are shorter than high-information words, and thus the length of a word is roughly proportional to the amount of information associated with the word, is consistent with the Uniform Information Density (UID) principle (Jaeger, 2010) or the Smooth Signal Redundancy (SSR) hypothesis (Aylett \& Turk, 2004), which state that information is distributed roughly evenly across words in an utterance.

There are many ways to operationalise the information content of a word. One way is to use the N -gram probability of a word, i.e. its probability conditioned on a window of N preceding or following words. This is the method used by Piantadosi et al. (2011). Zipf's word frequency measure is in fact just a limiting case of this N -gram probability, where $\mathrm{N}=0$. Other measures include syntactic probability, a word's probability of appearing in a particular syntactic structure (Jaeger, 2010, e.g.), and givenness, a word's predictability given the semantic context (Aylett \& Turk, 2004).

Both corpus studies and controlled behavioural experiments have linked low information content, operationalised in these different ways, to various types of linguistic reduction. Lieberman (1963); Aylett and Turk (2004); Gahl and Garnsey (2004); Tily et al. (2009); Kuperman and Bresnan (2012), and Seyfarth (2014) show that words with low information content are more likely to undergo various types of phonetic reduction. Bell, Brenier, Gregory, Girand, and Jurafsky (2009) show that each of the four different measures of information content mentioned above may in fact contribute separately to the phonetic duration of a word. Fedzechkina, Jaeger, and Newport (2012) show that case markers are more likely to be omitted on nouns in more probable syntactic roles. Jaeger (2010) shows that that-complementisers are more often dropped when the following word is less surprising in context.

If predictability in context can lead to phonetic reduction, as well as deletion of morphemes and entire words, then these effects might make their way to the overall distribution of form-meaning mappings in the lexicon. However, there is relatively little work directed at understanding how predictability affects this widely observed pattern at the level of the lex-
icon.
One way of investigating the issue is by tracking languageusers' online choices when producing words that are part of a 'clipped pair', i.e. when both a long form and an abbreviated or 'clipped' form exist that have the same or very similar meanings (Mahowald, Fedorenko, Piantadosi, \& Gibson, 2013). E.g. in English, info/information is a clipped pair. Mahowald et al. presented participants with sentences containing a blank and asked them to complete the sentence with either the long or the clipped form corresponding to the relevant meaning. They found that participants were more likely to choose the short form in predictive contexts, which is consistent with the hypothesis that the lexicon-level patterns observed by Piantadosi et al. (2011) may be due in part to a least-effort mechanism, in which speakers balance communicative efficacy with efficiency.

However, because this study uses English sentence frames and target words, we cannot rule out potentially confounding contributions from register, prosody, and participants' learned preferences to their word choice in particular instances. Moreover, we cannot assess whether the effect is really driven by the competing pressures for communicative accuracy and efficiency without manipulating the presence or absence of these different communicative pressures. For instance, in Mahowald et al.'s task, it seems participants clicked on a word rather than typing it in, and thus there was no difference in effort between choosing the long or short form. In addition, participants were told to choose a word based on "which sounded more natural", rather than being directly engaged in a task requiring successful communication.

Here, we present a new artificial language learning study investigating the question of whether language-users align word length with information content when communicating. Our results are consistent with previous findings that language-users tend to use shorter forms in more predictive contexts. Furthermore, the behaviour we observe across different experimental conditions supports the hypothesis that this effect is driven at least in part by a least-effort principle, in which language-users balance the competing pressures for communicative accuracy and efficiency to reshape the lexicon into one where word length is roughly proportional to average information content.

## Method

Artificial language learning studies have previously been used to shed light on the cognitive mechanisms and environmental pressures that shape large-scale linguistic structure. In this paradigm, participants learn an artificial language, and then we observe how they reshape their input as they use the language, in this case to communicate with a partner (e.g., Winters, Kirby, \& Smith, 2015; Kirby, Tamariz, Cornish, \& Smith, 2015; Fehér, Wonnacott, \& Smith, 2016).

## Participants

120 participants ( 53 females, 66 males; one did not report their gender) were recruited and remunerated via Amazon

Mechanical Turk. 108 of these reported themselves as native English speakers, of which 96 were monolingual. A range of other languages were represented across the remaining participants. Ages ranged from 18 to 70 (mean=32.9, $\mathrm{SD}=9.5$ ).

## The Training Language

The study was run online. Participants were trained on two names for each of two plant-like alien objects, by repeatedly being shown pictures of the objects labeled with a simple sentence. The sentence consisted of a framing word followed by the object's name. There were two possible frames, bix and gat. Overall there were 64 training trials, with each object appearing 32 times and each frame appearing 32 times. Crucially, one object appeared seven times more frequently with the frame bix than gat (28 and 4 times, respectively), while the other object appeared seven times more frequently with the frame gat than bix (again, 28 and 4 times, respectively). This meant that each object appeared in both a predictive context and a surprising context; which frame signified which of these contexts was flipped between the two objects.

Furthermore, the object name appeared half the time in its full form, a 7-letter word, and half the time in shortened form, a 3-letter word derived by clipping the last two syllables off the long name. These short and long forms were evenly distributed across both predictive and surprising contexts, ensuring that the input language contained no inbuilt bias towards using one form in any particular context. ${ }^{1}$ A schematic diagram of the object frequencies and labels is provided in Fig. 1 A .

In natural languages, shorter words are subject to greater confusability for a number of reasons: shorter forms have less space for signal redundacy and thus are more likely to be completely lost in noisy signal transmission; and because languages have a finite phoneme inventory, there are more unique possible long strings than short strings, and thus word shortening often results in ambiguity. Indeed, shorter words are more likely to be polysemous and homophonous (Piantadosi, Tily, \& Gibson, 2012). To model this fact in our miniature lexicon, we designed the names such that the short name for both objects was identical (zop), while the long names were unique (zopekil and zopudon).

## Procedure

Participants were assigned to one of four conditions, where we manipulated the presence of pressures to communicate accurately and quickly in a between-subjects $2 \times 2$ design (Kanwal, Smith, Culbertson, \& Kirby, 2017). Each experiment consisted of two phases: training and testing. The training phase was uniform across conditions, while the testing phase varied by condition.

[^117]

Figure 1: (A) The input frequencies of the objects and framing sentences presented during training trials in all four experimental conditions. (B) A sample training trial. (C) A sample director trial in the Combined condition (top) and a matcher trial followed by feedback (bottom).

Training phase On each training trial, an object was presented on screen alone for 700 ms . The appropriate sentence then appeared beneath the object for a further 3000 ms , yielding a total trial duration of 3700 ms . A blank screen showed for 500 ms between trials. The 64 training trials were presented in a different randomised order for each participant.

Testing phase After the training phase, testing procedures varied depending on the experimental condition. In the Combined condition, participants were under a pressure to communicate accurately and efficiently, as according to the Principle of Least Effort, it is balancing these competing pressures that leads language-users to distribute word length inversely to word predictability. The remaining three conditions removed one or both of these accuracy and time pressures.

Condition 1: Combined In the testing phase of this condition, participants were paired with a partner to play a communication game, using the method developed for running two-player online experiments in Kanwal et al. (2017). On
each trial, the 'director' was shown an object on the screen with a framing word followed by a blank. The director was instructed to choose a name for the object to complete the sentence, and once the name was entered, the sentence would be transmitted to the 'matcher'. The director could choose one of two options to complete the sentence: the unique long name for the object or the (ambiguous) short name. Once the chosen name was selected by clicking on the appropriately labeled button, it had to be entered into the blank space by pressing and holding the mouse as each letter appeared one after the other at 1200 ms intervals. Only after all the letters in the name had appeared in the box was the completed sentence transmitted to the matcher. This belaboured method of production, in which the long name was significantly slower to produce than the short name, was introduced to model the difference in effort and speed associated with producing long versus short utterances.

Once the director completed their description, it was transmitted to the matcher, who was asked to choose which of the two objects they thought the director was referring to. Both players were then given feedback as to whether the matcher's choice was correct.
The players alternated roles after every trial, with the matcher becoming the director and the director becoming the matcher, until both completed 32 director trials and 32 matcher trials. The proportion of times each object appeared with each frame in each player's director trials matched those of the training proportions: one object appeared seven times more frequently with the frame gat than bix, and the other appeared seven times more frequently with bix than gat. The order of each participant's 32 director trials was randomly shuffled.

To model the pressures in spoken communication to be both efficient and accurate, pairs were told at the beginning that they would be rewarded a bonus payment of $\$ 1$ if they were the pair to complete the game in the quickest time with the highest number of correct match trials. Time was only counted when the director was entering a name into the blank, and the total time count was displayed next to the blank during this process, to emphasise the time pressure. Screenshots of sample director and matcher trials are shown in Fig. 1C.

In this condition, with pressures to be speedy yet accurate, we expected participants to converge on an optimal strategy in which the short name is used for an object when it appears in its predictive context, and the long name otherwise. In predictive contexts, the framing word already provides a lot of information to the matcher about which object is likely under discussion, and thus participants can minimise effort by using the short form. Conversely, in surprising contexts, the full object name is required to ensure disambiguation.

In order to establish a causal link between these purported mechanisms and the behaviour we observe, we included three further experimental conditions, described below, for a full $2 \times 2$ manipulation of the pressures for accuracy and efficiency.

Condition 2: Accuracy In this condition, participants were paired to play a communication game as described above, but in the director trials, there was no intermediate step between the director choosing a name to complete the sentence and the matcher receiving the sentence; the names were entered instantaneously, thus removing any difference in effort between producing long or short names. Pairs were told that the goal of the game was simply to have their partner make as many correct guesses as possible. No bonus prize was offered in this condition, as we expected many pairs to hit ceiling as they did in Kanwal et al. (2017)-however, fewer than expected actually did so here.

In this condition, we predicted that participants would be more likely to use the long names for both objects across all contexts, as the long names are less confusable, and without a pressure to be efficient, there is little reason to shorten.
Condition 3: Time In this condition, communication was taken out of the game entirely; participants played a oneplayer game consisting of 64 director trials. In each trial, participants completed the sentence with either the long or short name for the object shown, but there was no subsequent communicative task. The name was simply entered as in the Combined condition, by pressing and holding the mouse in the blank space, with each letter appearing at 1200 ms intervals, while a timer displayed the total time count. The next trial began once all the letters had appeared in the box. Participants were told at the beginning of the game that they would be rewarded a bonus payment of $\$ 1$ if they were the player with the shortest total time count.

Here, we expected participants to use the short name for both objects across all contexts: with no communicative purpose attached to the transmissions, and an incentive to be as quick as possible, using the short name in every trial is the best strategy.
Condition 4: Neither The fourth and last condition contained neither a pressure for efficiency nor a pressure for accuracy. As in the Time condition, participants played a oneplayer game with no explicit communicative element. Additionally, there was no time difference associated with transmission; once a label was chosen to complete a sentence, it was instantaneously recorded and the player advanced to the next trial. We included this condition to provide a baseline for participants' behaviour from which to assess the effects of the accuracy and time pressures in the other three conditions.

In this condition we expected that participants would either probability-match-i.e. use the long and short forms for both objects with equal frequency, as in the training trials (Hudson Kam \& Newport, 2005)—or their behaviour would reveal prior biases language users bring to the task, such as a preference against using ambiguous forms, as observed in Kanwal et al. (2017).

## Results

Fig. 2 shows the proportion of trials in which the short name was produced by each participant or pair of participants in


Figure 2: The proportion of trials in which the short name was used in predictive contexts versus the proportion of trials in which it was used in surprising contexts. For the Combined and Accuracy condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time and Neither condition, each data point corresponds to an individual player's productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. These results demonstrate that behaviour consistent with the principles of UID or SSR-using short forms in predictive contexts and long forms in surprising contexts, generating systems that fall in the bottom right corner of each graph-only reliably arises in the Combined condition.
predictive versus surprising contexts. Our predictions were borne out by the results in all four conditions. In the critical Combined condition, in which participants were subject to the combined pressures for accuracy and efficiency, pairs of communicating participants produced systems in which the short name was used in predictive contexts and the long name in surprising contexts. Crucially, only when both pressures were present did participants reliably produce systems where word length was conditioned on context in this way. In the Accuracy condition, participants tended to use the long name for both objects regardless of context, and in the Time condition, they used the short name for both objects regardless of context. In the Neither condition, some participants stuck with the long name or the short name throughout the trials regardless of context, as in the Accuracy or Time conditions; however, most participants probability-matched.

A logistic regression model was fit to the full dataset in R using the lme4 package, with short name use (as contrasted with long name use) as the binary dependent variable; context (predictive or not), experimental condition, and their interaction as fixed effects; and by-participant random slopes


Figure 3: This figure shows the extent to which participants' name choices are conditioned on context (lefthand graph) and object (righthand graph). The dotted line in the lefthand graph represents the mutual information $\left(\mathrm{MI}_{\mathrm{C}}\right)$ associated with the 'optimal' language in least-effort terms-the language in which the short form is used only in predictive contexts, and the long form only in surprising contexts. $\mathrm{MI}_{\mathrm{C}}=0$ for the input language. In the righthand graph, mutual information $\left(\mathrm{MI}_{\mathrm{o}}\right)$ can range from 0 (same name fixed for both objects) to 1 (distinct names fixed for each object). $\mathrm{MI}_{\mathrm{O}}=0.5$ for the input language, marked by the dotted line. Data from only the second half of testing trials is shown in this figure, as participants were more likely to have converged on a stable mapping by this time.
and intercepts for context. The model was sum coded, setting the grand mean as the intercept, to which each level was then compared. The results yielded a significant positive interaction of context in the critical Combined condition ( $\beta=0.619, S E=0.158, p<0.001$ ), indicating that in this condition, participants were significantly more likely to use the short name in predictive contexts. The only other significant effects found were as follows: a positive overall effect in the Time condition ( $\beta=2.187, S E=0.292, p<0.001$ ), indicating that participants were more likely to use the short form in this condition regardless of context; a negative overall effect in the Accuracy condition ( $\beta=-1.470, S E=0.233, p<$ 0.001 ), indicating that participants were less likely to use the short form in this condition regardless of context; and finally a negative interaction effect of context in the Accuracy condition ( $\beta=-0.490, S E=0.161, p=0.002$ ), indicating that in fact participants were even less likely to use the short form in the predictive context in this condition.

An analysis of how participants conditioned the variation in their name usage sheds further light on the differing patterns of behaviour seen across conditions. We calculated the average mutual information between name produced and context (predictive or not) in each participant's output language $\left(\mathrm{MI}_{\mathrm{C}}\right)$. The more reliably participants are conditioning their use of the long and short names on context, the higher we would expect the value of $\mathrm{MI}_{\mathrm{C}}$ to be. The distributions for all four conditions are plotted on the lefthand graph of Fig. 3.

We also calculated the average mutual information between name produced and object (the blue fruit or the red stalk) in each participant's output language ( $\mathrm{MI}_{\mathrm{O}}$ ). This measure allows us to determine whether some participants are us-
ing fixed names for each object, regardless of context. The results are plotted by condition in the righthand graph of Fig. 3. If participants are using a distinct name for each object, $\mathrm{MI}_{0}$ will be close to 1 ; if they are using the same name for both objects, $\mathrm{MI}_{\mathrm{O}}$ will be close to 0 . The former pattern is what we see in the Accuracy condition: most participants use the unique long name for each object, regardless of context. The latter pattern is what we see in the Time condition: most participants use the ambiguous short form for both objects, regardless of context.

In the Combined and Neither conditions, $\mathrm{MI}_{\mathrm{O}}$ hovers around that of the input language. Based on this graph alone, participants may be probability matching in both these conditions, or perhaps reliably conditioning their output on other factors. Looking back at $\mathrm{MI}_{\mathrm{C}}$ disambiguates: it is significantly higher in the Combined condition than in any other condition. A linear regression on $\mathrm{MI}_{\mathrm{c}}$ with condition as predictor variable (fit to the second half of testing trials, as in Fig. 3) yielded a significant negative effect of the Accuracy ( $\beta=$ $-0.081, S E=0.033, p=0.016)$, Time $(\beta=-0.184, S E=$ $0.041, p<0.001$ ), and Neither ( $\beta=-0.128, S E=0.041, p=$ 0.002 ) conditions, with the Combined condition set as the intercept. This result is consistent with what we saw in Fig. 2: in the Combined condition, many participants are optimally conditioning their responses on context, generating systems that fall in the bottom right corner of the graph; in the other conditions, almost no data points fall in this region.

## Discussion

There is mounting evidence that utterance length is linked to information content (Lieberman, 1963; Aylett \& Turk, 2004; Gahl \& Garnsey, 2004; Tily et al., 2009; Bell et al., 2009; Jaeger, 2010; Piantadosi et al., 2011; Kuperman \& Bresnan, 2012; Fedzechkina et al., 2012; Seyfarth, 2014). The explanation put forth in much of this previous work is that speakers are driven by pressures much like those outlined in Zipf's Principle of Least Effort: the competing demands for accurate and efficient communication lead speakers to converge on an optimal system in which information content is spread roughly uniformly across the utterance, resulting in low-information units being shorter than high-information units. This resultant effect appears to have made its way into the structure of the lexicon as a whole: shorter words appear on average in more predictive contexts than longer words (Piantadosi et al., 2011). But is this effect really due to the proposed mechanism? Can speaker choice lead to the reshaping of a lexicon to align it with the principles of Uniform Information Density and Smooth Signal Redundancy?

Here, we presented the first study that concretely addresses these questions. Previous studies either lacked a manipulation of the communicative pressures operating in the task, or lacked a communicative element entirely. In our study, by observing participants' online behaviour in a task in which the pressures to communicate accurately and efficiently were manipulated across four experimental conditions, we have
shown that participants use shorter words in more predictive contexts only when both competing pressures were acting on them. When these pressures were isolated or removed entirely, participants failed to reliably condition their word choices on context.

Furthermore, because our study employed an artificial language learning paradigm, our findings avoid potential confounds from factors such as register, prosody, and participants' learned preferences in their native or second languages. Our results are nevertheless consistent with previous findings that language-users tend to use shorter forms in more predictive contexts when using their native language.

Our results serve as a proof of concept that the lexiconlevel effect observed by Piantadosi et al. (2011) could be driven at least in part by a least-effort principle in which language-users balance the competing pressures for communicative accuracy and efficiency to reshape the lexicon into one where word length is roughly proportional to information content. However, there is a crucial step between what we have observed here-language-users alternating between long and short variants for a single meaning depending on context-and what Piantadosi et al. (2011) observed in the lexicon of different languages, where most meanings don't correspond to both a long and a clipped variant, but rather map to a single fixed form. For these cases, which make up the majority of the lexicon, the length of the form is strongly correlated with the average predictability-in-context of the meaning, across all its different occurrences. We can hypothesise a link between these two phenomena: as a word appears in increasingly more predictive contexts, a reduced variant may come into use. If speakers use the reduced variant in predictive contexts, then this reduced form will consequently become much more frequent than the long form, leading to the long form eventually dying out altogether. This would end in a scenario where a short word, with no alternative variants currently in use, appears on average in a high number of predictive contexts, and thus has a low average information content. Though this story sounds reasonable, a precise mechanistic explanation of how this preference for short forms in more predictive contexts leads to permanent shifts in formmeaning mappings has yet to be thoroughly investigated. We hope this topic is given more attention in future work.

## References

Aylett, M., \& Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and speech, 47(1), 31-56.
Bell, A., Brenier, J. M., Gregory, M., Girand, C., \& Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. Journal of Memory and Language, 60(1), 92-111.
Fedzechkina, M., Jaeger, T. F., \& Newport, E. L. (2012, October). Language learners restructure their input to facilitate efficient communication. Proceedings of the National

Academy of Sciences, 109(44), 17897-17902.
Fehér, O., Wonnacott, E., \& Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. Journal of Memory and Language, 91, 158-180.
Gahl, S., \& Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. Language, 748775.
Hudson Kam, C. L., \& Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language learning and development, 1(2), 151-195.
Jaeger, T. F. (2010, August). Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61(1), 23-62.
Kanwal, J., Smith, K., Culbertson, J., \& Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. Cognition. (In press.)
Kirby, S., Tamariz, M., Cornish, H., \& Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. Cognition, 141, 87-102.
Kuperman, V., \& Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. Journal of Memory and Language, 66(4), 588-611.
Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. Language and speech, 6(3), 172-187.
Mahowald, K., Fedorenko, E., Piantadosi, S. T., \& Gibson, E. (2013, February). Info/information theory: Speakers choose shorter words in predictive contexts. Cognition, 126(2), 313-318.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108(9), 3526-3529.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122(3), 280-291.
Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. Cognition, 133(1), 140-155.
Tanimura, S., Kuroiwa, C., Mizota, T., et al. (2006). Proportional symbol mapping in R. Journal of Statistical Software, 15(i05).
Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., \& Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. Language and Cognition, 1(2), 147-165.
Winters, J., Kirby, S., \& Smith, K. (2015). Languages adapt to their contextual niche. Language and Cognition, 7(3), 415-449.
Zipf, G. K. (1935). The psycho-biology of language (Vol. ix). Oxford, England: Houghton Mifflin.

# Preparatory Effects of Problem Posing on Learning from Instruction 

Manu Kapur<br>ETH Zurich, Switzerland


#### Abstract

A randomized-controlled study compared the preparatory effects of problem-posing on learning from subsequent instruction. Students engaged in problem-posing either with solution generation (where they generated problems and solutions to a novel situation) or problem-posing without solution generation (where they generated only problems) prior to learning a novel math concept. Problem-posing with solution generation prior to instruction resulted in significantly better conceptual knowledge, without any significant difference in procedural knowledge and transfer. These findings suggest that although solution generation prior to instruction plays a critical role in the development of conceptual understanding, and generating problems can further enhance transfer.


# Comparing Individual and Collaborative Problem Solving in Environmental Search 

Franziska Keilmann<br>Max Planck Institute for Biological Cybernetics<br>Stephan de La Rosa<br>Max Planck Institute for Biological Cybernetics<br>Stephan Schwan<br>Leibniz-Institut für Wissensmedien<br>Ulrike Cress<br>Leibniz-Institut für Wissensmedien<br>Betty J Mohler<br>Max Planck Institute for Biological Cybernetics<br>Heinrich H Bülthoff<br>Max Planck Institute for Biological Cybernetics<br>Tobias Meilinger<br>Max Planck Institute for Biological Cybernetics


#### Abstract

Collaborative spatial problem solving is an important yet not thoroughly examined task. Participants navigated individually and in dyads through virtual cities of varying complexity. They only saw the environment part visible from their current location from a bird's eye view map perspective. We recorded missed target locations, overall trajectory length and search time per person until self-indicating whole coverage. Our results show a general increase in missed locations, trajectory length, and search time with the complexity of the environment. These increases differed due to individual and collaborative search. For complex, but not for simple environments individual participants navigated shorter distances, finished earlier, but also missed more target locations than when searching the same environments in collaboration. These results indicate that in complex environments collaborative search is less error prone than individual search, but takes longer. Such initial findings will constrain future theorizing about collaborative spatial problem solving.


# Iterated Teaching Can Optimise Language Functionality 

Vera Kempe (v.kempe@abertay.ac.uk), Kamil Cichon (1301137@abertay.ac.uk)<br>University of Abertay, Dundee, Scotland<br>Nicolas Gauvrit (ngauvrit@me.com)<br>Université Paris 8 \& École Pratique des Hautes Études, Paris, France

Monica Tamariz (m.tamariz@hw.ac.uk)
Heriot-Watt University, Edinburgh, Scotland


#### Abstract

Experimental studies of the cultural evolution of language have focused on how constraints on learning and communication drive emergence of linguistic structure. Yet language is typically transmitted by experts who adjust the input in ways that facilitates learning by novices, e.g. through child-directed speech. Using iterated language learning of binary auditory sequences, we explored how language change is affected by experts' intention to teach the language to novices. Comparison between teaching chains and simple transmission chains revealed that teaching was associated with a greater rate of innovation which led to emergence of more expressive languages consisting of shorter signals. This is the first study to show that during cultural transmission, teaching can modify, and potentially optimise, functional characteristics of language.


Keywords: Teaching; iterated language learning; cultural transmission; algorithmic complexity; compositional structure; combinatorial structure

## Introduction

Cultural transmission of knowledge proceeds via the social learning mechanisms of imitation, emulation and teaching Boyd \& Richerson, 1985; Richerson \& Boyd, 2005). Of these, the adaptive value of teaching has recently received increased attention (Cavalli-Sforza \& Feldman, 1981; Kline et al. 2013; Kline, 2015; Csibra \& Gergely, 2009), highlighting in particular the ostensive use of language in the transmission of technological knowledge required for production of tools and other cultural artifacts (Caldwell \& Millen, 2009; Morgan, Uomini, Rendell, Chouinard-Thuly et al., 2015). In its most general sense, teaching can be defined as any kind of behaviour, intentional or not, that promotes learning by narrowing the range of inferences or behavioural options that another individual can pursue (Kline, 2015). Teaching is especially important for transmission of cognitively opaque cultural traits and traditions, i.e. those whose function is not immediately obvious, thereby contributing to cumulative culture (Mesoudi 2011). While transmission of simpler cultural traits may not benefit from additional teaching (Caldwell \& Millen, 2009), once culture becomes more complex, teaching delivers additional benefits for the transmission process (Morgan et al., 2015).

In contrast, studies of the cultural transmission of language have mainly researched how constraints that
operate on observational learning drive the emergence of structural features like learnability, expressivity (i.e. lack of ambiguity), combinatorial and compositional structure, that support effective communication (Kirby, Cornish and Smith, 2008; Kirby, Tamariz, Cornish \& Smith; 2015; Verhoef, Kirby \& de Boer, 2014). These studies use the iterated language learning method whereby the result of learning in one generation of learners serves as input for the next generation. The picture that emerges from these studies is that cognitive capacity constraints in human learners promote compressibility of individual signals and entire languages, and that the requirements of efficient communication drive languages to be expressive. In tandem, these constraints - the need for transmission efficiency and for referential efficiency - lead to emergence of combinatorial linguistic structure, i.e. systematic associations of components of the signal with dimension of meaning (Kirby et al., 2015).

However, research on language development in children has presented substantial evidence that child language learners receive input that is specifically tailored to support learning, in the form of child-directed speech (e.g. Burnham, Kitamura, \& Vollmer-Conna, 2002; Kempe \& Brooks, 2005; Soderstrom, 2007). While there is considerable debate about whether child-directed speech is universal (e.g. Broesch \& Bryant, 2013; Falk, 2004; Schieffelin, 1985), whether it constitutes intentional teaching or whether it predominantly supports affective bonding and emotion regulation (e.g. Singh, Morgan \& Best, 2002; Uther, Knoll \& Burnham, 2007), functionally it qualifies as a behaviour that not only provides local enhancement by directing the learner's attention to relevant information (Kline, 2015) but also pre-samples the input in a way that can support correct learner inferences about language (Eaves, Feldman, Griffiths \& Shafto, 2016). For learning to occur, no ostensive cues or direct feedback are required as long as the statistical properties of the modified input ensure improved learning. In this study, we explore how tailoring the input in such a way for the learner shapes language structure over the course of language transmission.

If teaching leads to modification of the input that promote correct inferences about language then we can make predictions about the directions in which iterated teaching will change the structure of the emerging system, compared to simple transmission that is only constrained by cognitive
limitations of the learner. Previous iterated language learning studies have highlighted a number of structural features that emerge because languages need to be both learnable and communicatively efficient. We predict that emergence of these features should be facilitated under conditions of teaching: First, if teaching accommodates learnability constraints one would expect transmission fidelity to improve even faster in chains where teaching takes place, compared to simple transmission. Second, as languages evolve to be efficient signalling systems individual signals acquire combinatorial structure by which smaller meaningless subcomponents are recombined to improve signal discriminability, a feature that improves transmission across noisy channels (Verhoef, 2012; Verhoef et al., 2014; Roberts, Lewandowski \& Galantucci, 2015) and should be enhanced through teaching. Third, in order to be communicatively efficient, languages also need to maintain expressivity by avoiding under-specification and ambiguity, another feature we expect to emerge faster under conditions of teaching. Fourth, languages become more systematic and self-similar, a property that, akin to phonotactic rules, supports learnability by reducing combinatorial freedom. This feature should also be enhanced by teaching. Finally, emerging compositional structure serves to systematically link meaningless components of signals to underlying dimensions of their meanings. Although compositional structure requires communicative pressure in addition to language transmission (Kirby et al., 2015; Nowak \& Baggio, 2016), to isolate the effect of teaching, and in the interest of feasibility, we decided to start our exploration with a simple transmission study that did not impose communicative pressure on learners. Despite the lack of communicative pressure, it is still conceivable that teachers modify the input so as to enhance compositional structure in order to highlight this functional aspect of language.

## Method

The present experiment compares transmission of an 'alien' language along chains of learners where each learner's output generated during testing is faithfully represented as input to the learner in the next generation (Simple Transmission condition) with transmission of the language through chains of learners who, after training, are asked to teach the language to the learner in the next generation (Teaching condition). Teaching in this set-up constitutes demonstration of the language to the next learner without any verbal explanation or instruction. The crucial question is whether teachers modify the input when presenting the language to the next learner in a way that goes beyond those modifications that are due to constraints on learning and reproduction.

Participants: Sixty undergraduate students were recruited at the University library for participation in a transmission chain study. Participants were assigned numbers corresponding to their slot in one of six chains, and
were called into the test area when the previous participant had finished the training phase. Performance of these participants was compared to that of sixty other participants who had been tested in a different simple transmission study conducted earlier (Kempe, Gauvrit, Gibson \& Jamieson, submitted). The 'alien' language used in this experiment consisted of high and low tones assembled into 6 or 8 -tone sequences. This signalling system was developed to eliminate any familiarity with signals that could bias learners towards preferences for specific aspects of combinatorial structure.

Materials: Two 500 ms sine-wave tones (high: $440 \mathrm{~Hz}=$ musical note $a$; low: $293.7 \mathrm{~Hz}=$ musical note $d$ ) were synthesised and recorded onto differently coloured answer buzzer of 9 cm diameter each. The fixed tone duration made it impossible to modify the length of the tones if pressing the buzzers for longer periods of time thus eliminating duration as a property of the signals. Seed languages consisted of random sequences of high vs. low tones, which were either six or eight tones long. These binary sequences instantiating the 'words' in the 'alien language' were paired with eight coloured objects differing in shape (spiky 'kiki'type vs. fluffy 'bouba'-type), size ( $2 \times 2 \mathrm{~cm}$ vs. $4 \times 4 \mathrm{~cm}$ ) and brightness ( $25 \%$ vs $75 \%$ saturation) (see Figure 1), which were printed on laminated cards sized $5 \times 8 \mathrm{~cm}$. All objects also had unique properties due to differences in specific shapes and hues.


Figure 1: Meanings associated with the signals (binary sequences) in the 'alien buzzer language'.

Procedure: After signing a consent form, participants in both conditions were told they would learn an 'alien' language used by a species of aliens that had no mouth and therefore used buzzers for communication. Participants were then shown six of the eight cards one at a time to familiarise them with the 'alien' objects. Training proceeded in an incremental fashion: In the Simple Transmission condition, participants were given a demonstration of the binary buzzer sequence for each card, and were asked to repeat it. Demonstration and practice were then immediately repeated resulting in incremental training consisting of two consecutive trials per card, before proceeding to the next card. Order of cards was randomised for each participant by shuffling the cards. After training, participants were shown the cards one at a time, and asked to produce the 'alien buzzer words' to the best of their abilities. Their responses were videotaped, coded and then presented unaltered to the next participant. To prevent the languages from degenerating, a 'homonymy filter' (Kirby et al., 2008) was applied by which up to two identical sequences (i.e. homonyms) were withheld and only six items were
presented during training. In case of no homonyms, two cards were withheld at random; in case of just one pair of homonyms, the card corresponding to one of the homonyms was withheld at random along with one other randomly chosen card. This manipulation was used to encourage productivity and to prevent languages from degenerating into ambiguous systems.

In the Teaching condition, chains were also seeded with six out of eight cards, to maintain compatibility with the Simple Transmission condition. In Generation 1, training proceeded in exactly the same way as in the simple transmission condition. However, from Generation 1 onwards, after training, participants were asked to 'teach' the language to the next participant in the chain who was called into the testing area at that time. When teaching, participants were instructed not to provide any verbal comments or instructions but to simply demonstrate the buzzer sequences twice to the next participant, allowing them to repeat the sequence after each demonstration.

## Results

Participant buzzer responses were videotaped and coded for further analyses. Inter-coder reliability, determined for $17 \%$ of trials, was $94 \%$. All dependent variables were analyzed using Growth Curve Analyses (GCA). To see whether trends were linear or tended to level off we included a quadratic term of Generation following Beckner, Pierrehumbert \& Hay (2017). Thus, our model contained fixed effects of Condition (Simple Transmission vs. Teaching) and linear and quadratic effects of Generation, and random intercepts of Chains as well as random slopes of Generation (Winter \& Wieling, 2016), resulting in a model of the structure Condition + Generation + Generation ${ }^{2}$ + Condition*Generation + Condition*Generation ${ }^{2} \quad+$ (l|Chain) $+\quad$ (0+Generation $\mid$ Chain $) \quad+$ ( $0+$ Generation ${ }^{2} \mid$ Chain). In all cases, the quadratic model provided a better or the same fit to the data compared to the linear model as determined by likelihood-ratio tests.

Expressivity: Languages in the teaching condition contained fewer homonyms than languages in the simple transmission condition. The outputs in the transmission condition were more prone to degenerate into underspecified, more ambiguous languages, as indicated by an interaction of Condition with the linear, $\beta=-0.44, \mathrm{t}=$ 2.50, $\mathrm{p}<.05$, and the quadratic effect of Generation, $\beta=$ $0.05, \mathrm{t}=2.78, \mathrm{p}<.01$. These findings are depicted in Figure 2. Note that in this and all subsequent figures error bars correspond to one S.E.M.


Figure 2: Number of unique sequences (out of 8 ) in Simple Transmission and Teaching chains.

Transmission Accuracy: We used length-normalised Levenshtein edit distance (LED) as an inverse measure of transmission accuracy. LED decreased faster in the Simple Transmission condition (Figure 3), as evidenced by a significant interaction between Condition and the linear effect of Generation, $\beta=-0.03, \mathrm{t}=-2.13, \mathrm{p}<.05$. In other words, when participants were asked to teach they introduced more innovations than when they simply tried to reproduce the binary sequences.


Figure 3: Mean length-normalised Levenshtein edit distance in Simple Transmission and Teaching chains.

Self-similarity: Average pairwise length-normalised LED between all pairs of sequences in a language served as an inverse measure of within-language similarity. This selfsimilarity increased (i.e. LED decreased) overall as indicated by a main effect of Generation, $\beta=-0.06, t=-$ $3.80, \mathrm{p}<.001$. The significant quadratic term suggests that increase of self-similarity was mainly due to the drop from the seed language and levelled off in subsequent generations (Figure 4).


Figure 4: Mean inverse self-similarity (within-language LED) in Simple Transmission and Teaching chains.

Length: Sequences had started out with an average length of 7 tones in the seed language at Generation 0 . In the Teaching condition, sequences remained of roughly the same length, which was significantly shorter than in the Simple Transmission condition, $\beta=-0.81, t=-2.60, p<.05$. The interaction between Condition and the linear effect of Generation, $\beta=0.66, \mathrm{t}=2.46, \mathrm{p}<.05$, confirmed that sequence length increased only during simple transmission (Figure 5).


Figure 5: Mean sequence length in Simple Transmission and Teaching chains.

Combinatorial structure: Structure of individual signals was operationalised as algorithmic complexity, using an estimate developed for short binary strings based on the coding theorem method (Gauvrit, Soler-Toscano, Zenil \& Delahaye, 2014; Zenil, Soler-Toscano, Delahaye \& Gauvrit, 2015). This measure provides an inverse estimate for the amount of structure of a given sequence relative to the variation in structure possible for all sequences of the same length (Figure 6). It captures the intuition that sequences like adadadad or aaaadddd, where $a$ represents the high and
$d$ represents the low note, are more structured than sequences like aadaddda. GCA did not yield any significant effects although the interaction between Condition and the linear effect of Generation, $\beta=-0.13, \mathrm{t}=-1.82, \mathrm{p}=.08$, fell short of significance, suggesting that there may have been a trend for algorithmic complexity to decrease somewhat during simple transmission.


Figure 6: Mean length-normalised algorithmic complexity in Simple Transmission and Teaching chains.

Compositional Structure: To determine compositional structure for each language, we calculated the Pearson product-moment correlations between differences in the three meaning dimensions of all meaning pairs and differences between associated signals pairs within each language, using 10,000 iterations of a Monte Carlo process to obtain a standardised score. This measure remained below the value associated with $\mathrm{p}=.05$, and did not differ between conditions and generations indicating that no compositional structure had emerged (Figure 7).


Figure 7: Mean compositional structure in Simple Transmission and Teaching chains. The horizontal line indicates $z=1.96, p=.05$.

## Discussion

We compared six teaching chains with six simple transmission chains to explore the effect of teaching during cultural transmission of language. As this was an exploratory study, we did not include a requirement to engage in referential communication. Thus, it was not unexpected that compositional structure did not emerge (Kirby et al., 2015), and we found no evidence that teachers would introduce it spontaneously.

What we found was that although transmission accuracy increased overall, it was significantly lower in the Teaching condition, counter to our expectations. Thus, considerably more innovations were introduced into the signals when participants were asked to teach rather than just to reproduce what they had learned. We suggest that these innovations served to stabilise certain features of the languages that the teachers considered crucial. The most notable change from Simple Transmission performance achieved through innovation in the Teaching condition was to maintain expressivity of the language: The number of different sequences within the taught languages remained high thus preventing these languages from degenerating by accumulating homonyms. This is an interesting result because previous research had demonstrated that without a strong incentive to communicate, capacity constraints of the learners drive languages towards under-specification and ambiguity (Kirby et al., 2015). What our findings suggest is that teaching can override this tendency, presumably due to strong biases about the functional destination of language, which is to be expressive, i.e. referentially efficient. It is noteworthy that the expressivity advantage in the Teaching condition arose even though we applied a homonymy-filter in the Simple Transmission condition to prevent the languages from degenerating. Without this filter, languages in the Simple Transmission condition would have accumulated even more homonyms (Kirby et al., 2008), presumably further deviating from the Teaching condition.

Another feature that remained stable in the Teaching condition was sequence length: Teachers managed to maintain sequence length at around the original 7 tones while in the Simple Transmission condition, sequence length increased dramatically. Stabilising or even reducing length is a strategy that can ensure learnability and transmission accuracy by keeping the form of signals within the limits imposed by working memory constraints. As this brevity constraint operated only in the Teaching condition it may reflect a cooperative adjustment on the part of the teacher designed to aid the learner.

We observed little further increase of self-similarity of languages beyond an initial gain following exposure to the initial random binary sequences. Self-similarity can be thought of as a measure of systematicity that is somewhat akin to phonotactic rules. If teachers attempted to resolve the trade-off between expressivity of the language and brevity of the signals, they would be more likely to use the full space of distinct binary sequences of shorter lengths, which restricts opportunity to achieve self-similarity. In line
with this conjecture, the trend towards self-similarity was less pronounced in the Teaching condition, although the difference between conditions did not reach statistical significance.

We also had hypothesised that teaching would lead to a faster increase in combinatorial structure to improve transmission efficiency. For the binary sequences used as signals in this study, introducing combinatorial structure would entail establishing subcomponents (e.g. ad or aad) that can be recombined using operations like repetition or mirroring, as in strings like adadad ( -0.75 [numbers in parentheses are the associated values of length-normalised algorithmic complexity) or aadaadaad (-1.43). In contrast, complex strings like ddaaad (1.43) or ddaaaadda (2.02) do not contain combinations of discernible subcomponents. According to our hypothesis, the Teaching condition should have given rise to more sequences of the former than the latter type. However, our data showed exactly the opposite trend: Although not significantly so, length-normalised algorithmic complexity tended to be higher in the Teaching condition, indicating less combinatorial structure than in the Simple Transmission condition. One possible explanation for this finding is that when trying to produce as many different sequences as possible while maintaining brevity of the signals, teachers sample more densely from the distribution of shorter sequences thereby inevitably utilising more sequences of higher complexity. On the other hand, when the brevity constraint is relaxed, learners in the Simple Transmission condition may produce longer sequences yet processing capacity limitations will force them to settle for more structured ones, which are made up of a limited repertoire of subcomponents.

This pattern of results shows that when teaching, which in this study entailed knowingly serving as input-generating models, participants changed their behaviour to adjust the input so as to constrain learner hypotheses in accordance with their own tacit knowledge about how languages function. Specifically, they were negotiating a trade-off between referential efficiency and transmission efficiency by introducing innovations that allowed them to generate unique sequences for each meaning, to prevent languages from degenerating into under-specified systems, while at the same time facilitating transmission fidelity by stabilizing the length of these sequences. It can be argued that the biases that shaped this teaching behaviour reflect participants' knowledge about the functionality of language as it was acquired through their native language use and, thus, these biases may not be informative about of the role teaching may have played in language evolution. However, learners in the simple transmission condition had access to exactly the same knowledge yet without the motivation to teach those biases were overridden by the drive towards compressibility. Thus, whatever the origins of the knowledge about the functionality of language are, our findings suggest that this knowledge affects teaching.

To summarise, the results of this study support the idea that teachers modify the input to learners in ways that reflect
their biases about the functional utility of a cultural trait. Applying this idea to the study of the cultural evolution of language means that theories of language transmission need to include teaching into the suite of transmission mechanisms under consideration. We hope that our findings will inspire more detailed explorations of the role of teaching in the cultural evolution of language in the future.

## References

Beckner, C., Pierrehumbert, J. B., \& Hay, J. (2017) The emergence of linguistic structure in an online iterated learning task. Journal of Language Evolution, ahead of print.
Boyd, R. \& Richerson, P. J. (1985). Culture and the evolutionary process. Chicago: University of Chicago Press.
Broesch, T. L., \& Bryant, G. A. (2015). Prosody in infantdirected speech is similar across western and traditional cultures. Journal of Cognition and Development, 16(1), 31-43.
Burnham, D., Kitamura, C., \& Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. Science, 296(5572), 1435-1435.
Caldwell, C. \& Millen, A.E. (2009) Social learning mechanisms and cumulative cultural evolution: Is imitation necessary? Psychological Science, 20, 1478-83.
Cavalli-Sforza, L. L., \& Feldman, M. W. (1981). Cultural transmission and evolution: A quantitative approach. Princeton: Princeton University Press.
Csibra, G. \& Gergely, G. (2009). Natural pedagogy. Trends in Cognitive Science, 13, 148-153.
Eaves Jr, B. S., Feldman, N. H., Griffiths, T. L., \& Shafto, P. (2016). Infant-directed speech is consistent with teaching. Psychological Review, 123(6), 758.
Falk, D. (2004). Prelinguistic evolution in early hominins: Whence motherese? Behavioural and Brain Sciences, 27, 491-541.
Gauvrit, N., Soler Toscano, F., Zenil, H., \& Delahaye, J.-P. (2014). Algorithmic complexity for short binary strings applied to psychology: A primer. Behaviour Research Methods, 46(3), 732-744
Kempe, V. \& Brooks, P. J. (2005). The role of diminutives in the acquisition of Russian gender: Can elements of child-directed speech aid in learning morphology? Language Learning, 55, Supplement: The Best of Language Learning, 139-176.
Kempe, V., Gauvrit, N., Gibson, A. \& Jamieson, M. (submitted) Adults are more efficient in creating and transmitting novel signalling systems than children.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105(31), 10681-10686.
Kirby, S., Tamariz, M., Cornish, H., \& Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. Cognition, 141, 87-102.

Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behaviour in humans and other animals. Behavioural and Brain Sciences, 38, E31.
Kline, M. A., Boyd, P. \& Henrich, J. (2013). Teaching and the life history of cultural transmission in Fijian villages. Human Nature, 24, 351-374.
Mesoudi, A. (2011). Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences. Chicago and London: University of Chicago Press.
Morgan, T. J. H., Uomini, N. T., Rendell, L. E., ChouinardThuly, L., Street, S. E., Lewis, H. M., ... \& Whiten, A. (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. Nature Communications, 6.
Nowak, I., \& Baggio, G. (2016). The emergence of word order and morphology in compositional languages via multigenerational signalling games. Journal of Language Evolution, lzw007
Richerson, P. J., \& Boyd, R. (2005). Not by genes alone. Chicago: University of Chicago Press.
Roberts, G., Lewandowski, J., \& Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. Cognition, 141, 52-66.
Schieffelin, B. B. (1985). The acquisition of Kaluli. In D. I. Slobin (ed.), The crosslinguistic study of language acquisition: Vol. 1: The data (pp. 525-594). Hillsdale, NJ: Lawrence Erlbaum Associates.
Singh, L., Morgan, J. L., \& Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? Infancy, 3(3), 365-394.
Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. Developmental Review, 27(4), 501-532.
Uther, M., Knoll, M., \& Burnham, D. (2007, January). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. Speech Communication, 49(1), 27.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. Language and Cognition, 4(4), 357-380.
Verhoef, T., Kirby, S., \& de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. Journal of Phonetics, 43, 57-68.
Winter, B., \& Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. Journal of Language Evolution, 1(1), 7-18.
Zenil, H., Soler-Toscano, F., Delahaye, J.-P., \& Gauvrit, N. (2015). Two-dimensional Kolmogorov complexity and an empirical validation of the coding theorem method by compressibility. PeerJ Computer Science, 1, E23.

# Dynamic Effects of Conceptual Combination on Semantic Network Structure 

Yoed N. Kenett (yoedk@sas.upenn.edu)<br>Sharon L. Thompson-Schill (sschill @ psych.upenn.edu)<br>Department of Psychology<br>University of Pennsylvania, Philadelphia, PA 19104 USA


#### Abstract

The generative capacity of language entails an ability to flexibly combine concepts with each other. Conceptual combination can occur either by using an attribute of one concept to describe another (attributive combination) or by forming some relation between two concepts to create a new one (relational combination). Prior research has addressed whether common or distinct processes support these two putatively different types of combinations. We turn the question around and ask whether the consequences of these combination types on our conceptual system might differ, by comparing semantic memory networks before and after participants perform either attributive or relational conceptual combinations. We find a general effect on the semantic networks: the structure of network decreases after participants conceptually combine some of the concepts in the network. However, the relational combination manipulation has a greater effect. Furthermore, only the relational combination manipulation leads to an increase in the network's connectivity.


Keywords: Conceptual combinations; Semantic Networks

## Introduction

Language generation involves the ability to combine concepts into novel combinations (Boylan, Trueswell, \& Thompson-Schill, 2017). Investigating how individuals combine concepts can shed unique light on different aspects of conceptual knowledge, including the cognitive mechanisms that enable the generative and flexible use of language. For example, consider the noun-noun combination robin hawk: while some interpret this combination as "a red-breasted hawk", applying the attribute "red-breast" of the robin to the hawk; others interpret this combination as "a hawk that preys on robins", applying a thematic, relational role between robins and hawks (Wisniewski, 1996). While these two types of conceptual combination mechanisms-attributive and relational-are studied via behavioral and neurocognitive means (Boylan et al., 2017; Estes, 2003), whether these two mechanisms are similar or distinct remain an open question. Furthermore, the effect of these mechanisms on semantic memory structure has not yet been studied. In this paper, we apply a computational network science methodology to examine the effects of attributive and relational mechanisms on semantic memory structure. Specifically, we will focus on conceptual combinations in noun-noun compounds.

Noun-noun compounds contain a modifier noun followed by a head noun. The modifier noun can be either "attributive" (as in zebra clam, where zebra denotes the attribute "striped") or "relational" (as in mountain lake,
where "mountain" is an object bearing a spatial relation with "lake"). An attributive based conceptual combination involves applying an attribute from the modifier noun to describe the head noun, such as zebra clam ("a clam that has stripes") A relational based conceptual combination, however, cannot be paraphrased this way - tennis ball is not "a ball that is tennis", but rather "a ball for playing tennis" (Downing, 1977; Gagné \& Shoben, 1997).

An open theoretical issue is whether attributive and relational mechanisms are similar or distinct; and if distinct, how these mechanisms are applied (Estes, 2003; Gagné, 2000; Gagné \& Shoben, 1997; Rogers \& McClelland, 2004). We address this issue from a novel perspective: we apply a network science methodology to represent and compare semantic memory networks before and after participants conceptually combine some of the concepts in the network with other concepts in either an attributive or a relational manner. Such an approach allows us to examine, for the first time, in what way conceptual combinations affect semantic memory structure, and how it differs based on attributive or relational mechanisms. We posit that such a conceptual combination manipulation will have restructuring effects on the semantic network, by changing or creating new connections between concepts in the network.

Recent studies have used computational network science to represent the structure of semantic memory (memory for knowledge and facts, Jones, Willits, \& Dennis, 2015), using network science tools, as a semantic network and analyze its properties (for a review, see Borge-Holthoefer \& Arenas, 2010). A semantic network comprises a set of nodes and edges, where nodes correspond to words or concepts and edges connect pairs of nodes and signify some sense of relations between the connected nodes. Of the various network models developed in network science theory, the network model that has been widely used to examine complex systems is the Small World Network (SWN) model. A SWN is a network that is characterized by both high local connectivity and short global distances between nodes, allowing for efficient transfer of information. This network type is known as a small world network because every node is relatively close to other nodes. Analyses of different languages have consistently shown how different linguistic systems exhibit such SWN characteristics, characteristics which are now considered fundamental in facilitating efficient and quick retrieval of linguistic information (Borge-Holthoefer \& Arenas, 2010). Common parameters of network structure include - the networks
clustering coefficient (CC), the average shortest path length (ASPL), and the modularity index (Q).

The CC measures the network's connectivity. It refers to the probability that two neighbors of a node will themselves be neighbors (i.e., a neighbor is a node $i$ that is connected through an edge to node $j$ ). The ASPL and Q index measure the global structure of the network. The ASPL measure refers to the average shortest number of steps needed to be taken between any two pair of nodes. The $Q$ measure examines how a network breaks apart (or partitions) into smaller sub-networks. The larger the modularity measure, the more the network comprised of sub-networks (Newman, 2006). A SWN is characterized by having a high CC and a short ASPL. To examine whether a specific network is a SWN, the statistical properties of empirical data are compared to those of a random null network with the same number of nodes and edges.

Previous work has conducted such analysis to examine cognitive phenomena such as language development, bilingualism, memory search and retrieval, and creative ability (Borge-Holthoefer \& Arenas, 2010). For example, Kenett et al. (2014) found that low and high creative individuals show different semantic network structure. The semantic network of high creative individuals exhibited lower ASPL and Q values, and higher CC values compared to that of the low creative individuals. This was the case despite both networks having an equal number of nodes, edges and average number of edges per node. Thus, semantic networks analysis can be applied to examine differences in semantic memory structure related to different conditions such as attributive or relational combinations.

Some current theories of semantic memory posit that conceptual representations are not invariable across people or across time, but rather dynamically change contingent on context (e.g., task demand, stimulus modality) and individual differences (e.g., processing preferences), with short- and long-term effects on the structure of semantic memory (Yee \& Thompson-Schill, 2016). Such a dynamic perspective describes an experienced-based, distributed, semantic memory system that allows for flexible, generative language. We apply semantic network analysis to examine how the process of combining concepts changes the semantic network and whether such effects depend on the different mechanisms (attributive or relational) applied in such combinations (see also Schilling, 2005).

Here, we present preliminary results of an on-going study where we examine and compare the structure of semantic memory networks before and after an attributive or relational conceptual combinations task. We operationalize the effects of the different conceptual combination mechanisms on semantic memory structure as differences in quantitative measures of the semantic network before and after the conceptual combination task. Specifically, we focus on global measures of the network's structure (ASPL and Q ) and connectivity (CC). We predict that any possible differences between these two mechanisms will be manifested in the post-manipulation networks.

## Materials \& Methods

## Participants

Participants ( $\mathrm{N}=26$ ) were recruited from the University of Pennsylvania as part of a larger on-going research study on conceptual combinations and semantic memory structure. Participants were $55 \%$ female, average age of 22.6 years ( $\mathrm{SD}=3.9$ ) and with an average 16.4 years of education (SD $=3$ ). Participants were randomly assigned to the attributive combinations (AC) or relational combinations (RC) conditions ( $\mathrm{N}=13$ in each group). This study was approved by the University of Pennsylvania Institutional Review Board.

## Design Overview

We characterized the semantic network of participants using their free association responses obtained twice, before and after completing a conceptual combination task that was biased (using both detailed instructions and a priming manipulation) to elicit either attributive or relational interpretations. With this procedure, we were able to assess the main effect, within subjects, of conceptual combination (by comparing the structure of the semantic networks at both time points) as well as the interaction, between subjects, of the type of conceptual combination on network change. We also collected a number of measures of cognitive ability that will be used in planned analyses of individual differences in these effects. We will first describe the conceptual combination task we used to manipulate the type of combination process (attributive or relational). We will then describe the method we applied to represent the semantic networks (before and after the conceptual combination task).

## Conceptual Combination Task

Participants were presented with 25 noun-noun combinations and were required to come up with an interpretation for each combination (Wisniewski \& Love, 1998). They were also asked to indicate how familiar they were with each combination and how hard it was for them to retrieve the interpretation they gave. In order to examine the effect of attributional and relational combination mechanisms on semantic memory, we used ambiguous noun-noun combinations-combinations that can either have an attributive or relational interpretation-and we primed the participants to generate either attributional or relational interpretations. This was achieved both by an instruction manipulation and by an initial noun-noun combination priming phase (Wisniewski, 1996). Our task comprised the following parts: instruction manipulation, priming phase, and ambiguous conceptual combination task. Participants performed this task between two sessions of semantic network estimation (a week apart; see Semantic Network Estimation). This allowed us to examine the effect of the different conceptual combination mechanisms on semantic memory structure.


Figure 1: 2D visualization of the pre- and post- AC and RC semantic networks

During the instructions stage, both participant groups received the same general description of the task. They were then told that there are different strategies that people use to combine concepts together and were given either attributional or relational instructions. In the attributional instructions, participants were told that one such strategy entails applying one dominant attribute of one word to explain the other. In the relational instructions, participants were told that one such strategy entails relating both words in some way. Participants read three ambiguous noun-noun combinations in which the specific required interpretation was emphasized. For example, participants read ant apple. In the attributive instructions participants were told that this could mean 'a small apple' but not 'apple with ants on it'. In the relational instructions participants were told that this could mean 'apple with ants on it' but not 'a small apple'.

In order to increase the difference between the two experimental conditions, we followed these instructions with an attributive or relational priming phase in the conceptual combination task, following from Wisniewski (1996), which showed that noun-noun combinations can be primed to generate either attributive or relational interpretations. We presented participants with ten modifierhead noun-noun combinations, where the head noun
remained constant but the modifier noun either primed an attributive combination or a relational combination (e.g. razor insult for the attributive combination condition vs. girlfriend insult for the relational combination condition).

Finally, participants completed the ambiguous nounnoun conceptual combination task for 25 word-pairs. In order to select stimuli not only that were ambiguous (in that they elicited attributive and relational interpretations across subjects) but also that were flexible (in that the percentage of attributive and relational interpretations could be affected by an instructional manipulation), we conducted a norming study via Amazon's Mechanical Turk (AMT). The AMT surveys were conducted on a larger pool of 50 noun-noun pairs, divided into two surveys of 25 noun-noun pairs each. We conducted three different variations of these surveys with 20 AMT participants in each survey. In the first variation (baseline condition), participants were presented with the noun-noun pairs and asked to generate an interpretation to it. These interpretations were then classified as either attributive or relational by two independent judges (inter-rater agreement > .8). This variation allowed identifying ambiguous noun-noun pairs, classified as pairs that ranged from $30 \% / 70 \%$ to $50 \% / 50 \%$ interpretations. The second and third AMT variations
(biased conditions) manipulated task instruction, as described above, to examine how much these noun-noun pairs could be "pushed" into one type of interpretation. These interpretations were similarly rated as attributive or relational by two independent judges (inter rater agreement $>.8)$. Finally, we examined the effect of the instruction manipulation on biasing the interpretations. We calculated a percent signal change, which quantified the percentage change in interpretations of an ambiguous word-pair from the baseline condition to the biased interpretation condition. This was calculated for both types of interpretations for all noun-noun combinations.

Based on the AMT surveys, 25 noun-noun combinations were chosen. These combinations were chosen so that the modifier-nouns were comprised from five different semantic categories (animals, fruits and vegetables, nature, food, and home). All of the modifier nouns, and none of the head nouns, were included in the semantic network analysis as described below. The average ambiguity of these word pairs was $54 \% / 46 \%$ attributive/relational interpretations. Percent signal change from baseline to biased attributive interpretations was $28 \%$ and biased relational interpretations was $42 \%$. No significant differences were found between the percent signal change for attributive vs. relational interpretations ( $p<.4$ ).

## Semantic Network Estimation

The semantic networks of the AC and RC groups were computed using the computational approach developed by Kenett et al. (2011). Participants in both groups performed a continuous free association task twice, once before and once after the conceptual combination task. Participants were presented with a cue word and had one minute to generate as many associative responses they could for that cue word. Participants generated free associations to a list of 50 cue words. These 50 cue words consisted of five categories used in the conceptual combination task, including the five modifier nouns for each category and five other category members. Thus, the a priori structure of the semantic network consists of five (category) communities.

The semantic network of these 50 cue words was computed and compared between the pre- and post- AC and RC conditions: First, the data were preprocessed to standardize responses and fix any spelling mistakes. Second, the associative correlation between any pair of cue words was calculated using Pearson's correlation. This resulted in a $50 \times 50$ matrix where each cell denotes the association correlation between node $i$ and node $j$. Finally, the planar maximally filtered graph filter was used to remove spurious correlations (Kenett et al., 2014). This produced an adjacency (connectivity) matrix that represents the associative correlations between any pair of nodes. As our focus is on the structure of the networks, the association correlations were binarized to equal one. Thus, the resulting semantic networks are unweighted (all weights equal one) and undirected (symmetrical relations). Constructing semantic networks for different groups (pre- and post- AC
and RC) that are comprised from the same nodes (50 cue words) and with an equal number of edges ( 288 edges) allows comparing between them. Furthermore, the average degree, the average amount of edges per node in all networks was equal (average of 5.76 edges per node).

Analyses were performed with the Brain Connectivity Toolbox for Matlab (Rubinov \& Sporns, 2010). The clustering coefficient (CC; measuring network connectivity) and the average shortest path length (ASPL; measuring global distances) were calculated (Boccaletti, Latora, Moreno, Chavez, \& Hwang, 2006). The network's CC and ASPL were evaluated qualitatively against the equivalent parameters in a random network with the same number of nodes and edges ( $\mathrm{CC}_{\text {rand }}$ and ASPL ${ }_{\text {rand, }}$, respectively). Lastly, the modularity (Q) index was calculated (Newman, 2006). In order to assess the reliability (i.e., statistical significance) of observed differences across time points and across subject groups, we used a bootstrap method (Efron, 1979) to simulate and then compare partial networks for each of the conditions. We reasoned that if the networks differed from each other, then any partial network consisting of the same nodes in the networks should also be different. Furthermore, the bootstrap method makes it possible to generate many simulated partial semantic networks, allowing for statistical examination of the difference between them. The bootstrapping procedure involves random selection of half of the nodes comprising the networks. Partial networks were constructed for each condition (pre- and post- AC and RC) separately for these selected nodes. This method is known as the without-replacement bootstrap method (Bertail, 1997). Finally, for each partial network, the CC, ASPL, and the Q index were computed. This procedure was simulated with 10,000 realizations. The difference between the bootstrapped partial networks on each network parameter was then tested using a mixed model analysis of variance (group [AC, RC] x time [pre, post]).

## Procedure

Participants completed all tasks using the Qualtrics software on two different sessions a week apart. In the first session, participants completed the free-association task. In the second session, participants first completed the conceptual combination task and then the free association task. In the free association task, participants were instructed to generate, in one minute, as many different responses they could think of to a cue word. In each trial, the cue word was presented in the center of the screen with a response box below it, where participants typed their responses. Below the response box appeared a timer, counting down from 60 seconds. After 60 seconds elapsed, a new trial immediately began. Cue words were presented randomly and after 25 cue words participants had a short break. In the conceptual combination task, participants were first instructed on the task with the task manipulation instruction (attributive or relational). Next, a short practice was conducted with the experimenter, who gave feedback on the participant's interpretations. Stimuli used in the practice were not used in
the task itself. In each trial the noun-noun combination appeared in the center of the screen with a response box below it. Participants were instructed to write their interpretations in the response box. Underneath the response box the participant had to choose how familiar s/he was with the noun-noun compound on a five point Likert scale (ranging from extremely familiar to not familiar at all) and how easy it was for them to generate the interpretation on a seven point Likert scale (ranging from extremely easy to extremely difficult). Participants were randomly assigned to the attributive and relational conditions. The stimuli were randomly presented.

## Results

We computed the semantic networks for the pre- and postAC and RC conditions based on the procedure outlined above. Next, we computed and compared the different network measures for all four networks (Table 1). To visualize the networks, we used the force-directed layout of the Cytoscape software (Shannon et al., 2003) to plot the graphs (Figure 1). In these 2D visualizations, nodes (cue words) are represented as circles and links between them are represented by lines. Since these networks are unweighted and undirected, the links merely convey symmetrical relations between two nodes. The grayscale of the nodes relate to the five semantic categories used in our study.

The network analysis revealed both general and specific differences between the pre- and post- AC and RC networks. In regard to structural properties of the networks, ASPL and Q, the post session led to lower ASPL and Q values, which was stronger for the RC network. In regard to connectivity property of the network, CC , the post session led to different effects in the AC and RC networks: while the AC network had a lower CC value, the RC network had a higher CC value, compared to the first session (Table 1).

Table 1: Network measures for the pre- and post- AC and RC networks

|  | AC-Pre | AC-Post | RC-Pre | RC-Post |
| :---: | :---: | :---: | :---: | :---: |
| CC | .702 | .699 | .697 | .701 |
| ASPL | 2.930 | 2.814 | 3.223 | 3.034 |
| $\mathbf{Q}$ | .578 | .565 | .583 | .560 |
| CC $_{\text {rand }}$ | .103 | .125 | .131 | .176 |
| ASPL $_{\text {rand }}$ | 2.331 | 2.341 | 2.339 | 2.338 |

The bootstrapping analysis revealed a significant main effect of time (pre, post) for ASPL and Q, due to decreased values for the post-session (all $p$ 's $<.001$ ). This analysis also revealed for all measures a significant interaction between group and time (all $p$ 's $<.001$ ). For ASPL and Q, this effect resulted from a stronger effect for the RC group (all p's < .001) and for the CC resulted from an increase in CC for the RC group and a decrease in CC for the AC group in the post-session (all $p$ 's $<.001$ ).

## Discussion

In this work, we applied a computational network science approach to examine the dynamic effects of conceptual combination mechanisms on the structure of semantic memory. We found general and specific effects on the network: In both groups, the post manipulation network exhibited lower structural properties of global distances and modularity, which was more pronounced in the RC group. Furthermore, while the AC post-manipulation network exhibited lower connectivity, the RC post-manipulation network exhibited higher connectivity. Thus, our results indicate that the relational combination manipulation has a greater effect on semantic memory structure than an attributive combination manipulation.

Notably, both networks have the same nodes, amount of edges, and average degree (number of edges per node). Thus, these differences reflect both a global task-induced effect on semantic networks and a local effect of relational combination manipulation on semantic memory structure. Both lower ASPL and Q have been related to higher creative ability (Kenett et al., 2014), thus indicating the creative effect of conceptual combinations on semantic memory. This stronger effect, combined with higher CC, in the RC group, suggests that relational combinations may demand the generation of novel contexts in which both nouns relate to each other, thus leading to higher restructuring of the network. More fine grained examination is needed in order to test specific effects on these networks.

Our findings are in line with current theories of semantic memory, which view it as a dynamic system (Schilling, 2005; Yee \& Thompson-Schill, 2016). Such theories argue that both context (task demands) and individual differences (processing style) lead to short- and long-term changes in semantic memory structure. Our current study applies semantic network analysis to examine how a conceptual combination task affects the structure of semantic memory and whether it is affected differently based on a specific conceptual combination mechanism. We show how manipulating concepts in the semantic network (through a conceptual combination manipulation) changes the structure of the network. We will also examine how individual differences affect the structure of semantic memory, based on the behavioral measures we are collecting in our ongoing study. Our findings are also related to recent studies investigating how relational versus attributive based categories differentially effect cognitive processing, such as typicality effects and learning (Asmuth \& Gentner, 2017; Gentner \& Kurtz, 2005; Rein, Goldwater, \& Markman, 2010). For example, Asmuth and Gentner (2017) show how relational nouns are more "mutable" (affected by context) in memory than entity nouns. Thus, our approach offers a quantitative method to examine such behavioral findings.

Finally, there are a few limitations to this study. First, our study currently has a small sample size, which can affect the reliability of our results. We are currently continuing to collect data to conduct these analyses with a larger sample size in each group to strengthen our results. Furthermore,
our research computed semantic networks aggregated at the group-level. It is possible that within these aggregated group-based networks there are further individual differences that relate to semantic memory structure and conceptual combinations. Future research needs to examine the effects of conceptual combinations on semantic memory structure at the individual-level (Benedek et al., 2017).

In conclusion, the work reported here is a first step at harnessing computational network science to investigate the effects of different conceptual combination mechanisms on semantic memory structure. We plan to continue and increase sample size and examine how our findings relate to various behavioral measures we are also collecting, such as creative ability, intelligence and personality traits. Overall, our results demonstrate that semantic networks can be applied to study group-level effects of different conceptual combination mechanisms and contribute to the growing body of literature demonstrating their efficacy in understanding high-level cognition.

## Acknowledgments

This research was funded by an NIH Award to Sharon Thompson-Schill (R01 DC015359-02).

## References

Asmuth, J., \& Gentner, D. (2017). Relational categories are more mutable than entity categories. The Quarterly Journal of Experimental Psychology, 70(10), 2007-2025.
Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., \& Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: a network science approach. Thinking \& Reasoning, 23(2), 158-183.
Bertail, P. (1997). Second-order properties of an extrapolated bootstrap without replacement under weak assumptions. Bernoulli, 3(2), 149-179.
Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., \& Hwang, D. U. (2006). Complex networks: structure and dynamics. Physics Reports, 424(175-308).
Borge-Holthoefer, J., \& Arenas, A. (2010). Semantic networks: Structure and dynamics. Entropy, 12(5), 12641302.

Boylan, C., Trueswell, J. C., \& Thompson-Schill, S. L. (2017). Relational vs. attributive interpretation of nominal compounds differentially engages angular gyrus and anterior temporal lobe. Brain and Language, 169, 8-21.
Downing, P. (1977). On the creation and use of English compound nouns. Language, 810-842.
Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics, 1-26.
Estes, Z. (2003). A tale of two similarities: Comparison and integration in conceptual combination. Cognitive Science, 27(6), 911-921.
Gagné, C. L. (2000). Relation-based combinations versus property-based combinations: A test of the CARIN theory and the dual-process theory of conceptual combination. Journal of Memory and Language, 42(3), 365-389.

Gagné, C. L., \& Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(1), 71-87.
Gentner, D., \& Kurtz, K. J. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, \& P. W. Wolff (Eds.), Categorization inside and outside the lab (pp. 151-175). Washington DC: APA.
Jones, M. N., Willits, J., \& Dennis, S. (2015). Models of semantic memory. In J. Busemeyer \& J. Townsend (Eds.), Oxford Handbook of Mathematical and Computational Psychology (pp. 232-254). Oxford, UK: Oxford University Press.
Kenett, Y. N., Anaki, D., \& Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. Frontiers in Human Neuroscience, 8(407), 1-16.
Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., \& Faust, M. (2011). Global and local features of semantic networks: Evidence from the Hebrew mental lexicon. PLoS ONE, 6(8), e23912.
Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences USA, 103(23), 8577-8582.
Rein, J. R., Goldwater, M. B., \& Markman, A. B. (2010). What is typical about the typicality effect in categorybased induction? Memory \& Cognition, 38(3), 377-388.
Rogers, T. T., \& McClelland, J. L. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. Cambridge, MA: M.I.T. Press.
Rubinov, M., \& Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. NeuroImage, 52(3), 1059-1069.
Schilling, M. A. (2005). A "small-world" network model of cognitive insight. Creativity Research Journal, 17(2-3), 131-154.
Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., \& Ideker, T. (2003). Cytoscape: A software for integrated models of biomolecular interaction networks. Genome Research, 13(11), 2498-2504.
Wisniewski, E. J. (1996). Construal and similarity in conceptual combination. Journal of Memory and Language, 35(3), 434-453.
Wisniewski, E. J., \& Love, B. C. (1998). Relations versus properties in conceptual combination. Journal of Memory and Language, 38(2), 177-202.
Yee, E., \& Thompson-Schill, S. L. (2016). Putting concepts into context. Psychonomic Bulletin \& Review, 23(4), 1015-1027.

# The provenance of modal inference 

Sangeet Khemlani ${ }^{1}$, Thomas Hinterecker ${ }^{\mathbf{2}}$, P.N. Johnson-Laird ${ }^{\mathbf{3}, 4}$<br>sangeet.khemlani@nrl.navy.mil, thomas.hinterecker@tuebingen.mpg.de, phil@princeton.edu<br>${ }^{1}$ US Naval Research Laboratory, Washington, DC 20375 USA<br>${ }^{2}$ Max Planck Institute for Biological Cybernetics, Tübingen, Germany<br>${ }^{3}$ Princeton University, Princeton NJ 08540 USA<br>${ }^{4}$ New York University, New York, NY 10003, USA


#### Abstract

People reason about possibilities routinely, and reasoners can infer "modal" conclusions, i.e., conclusions that concern what is possible or necessary, from premises that make no mention of modality. For instance, given that Cullen was born in New York or Kentucky, it is intuitive to infer that it's possible that Cullen was born in New York, and a recent set of studies on modal reasoning bear out these intuitions (Hinterecker, Knauff, \& Johnson-Laird, 2016). What explains the tendency to make modal inferences? Conventional logic does not apply to modal reasoning, and so logicians invented many alternative systems of modal logic to capture valid modal inferences. But, none of those systems can explain the inference above. We posit a novel theory based on the idea that reasoners build mental models, i.e., iconic simulations of possibilities, when they reason about sentential connectives such as and, if, and or (Johnson-Laird, 2006). The theory posits that reasoners represent a set of conjunctive possibilities to capture the meanings of compound assertions. It is implemented in a new computational process model of sentential reasoning that can draw modal conclusions from non-modal premises. We describe the theory and computational model, and show how its performance matches reasoners' inferences in two studies by Hinterecker et al. (2016). We conclude by discussing the model-based theory in light of alternative accounts of reasoning.


Keywords: mental models, modal reasoning, possibilities, reasoning, probability logic

## Introduction

The word "possibility" is fraught with ambiguity, because philosophers distinguish between different sorts of possibility. An "alethic" possibility is any description that is not self-contradictory. "Deontic" possibilities are those that are permissible (e.g., instances of drinking alcohol when over the legal age restriction), and impossibilities are those that are prohibited (e.g., drinking while under the age restriction). Deontic possibilities can be violated, whereas alethic possibilities cannot (cf., Bucciarelli \& JohnsonLaird, 2005; Bucciarelli, Khemlani, \& Johnson-Laird, 2008). The present paper focuses on a different notion of possibility: "epistemic" possibilities concern possibilities that are consistent with a reasoner's personal knowledge. Reasoning based on possibilities is referred to as "modal" reasoning, because when you assert that something is "possible", you qualify its occurrence. Conventional systems of logic cannot take into account the logical properties of modals to draw conclusions, because they
concern unqualified propositions that are either true or else false. Consider these two assertions:

1a. Sarah is Egyptian.
b. Possibly, Sarah is Egyptian.

Assertion (1a) is unqualified: it asserts a fact about Sarah. If it is true then Sarah is indeed an Egyptian, and if it is false, she is not an Egyptian. Assertion (1b) is subtler. In addition to facts and their negations, it introduces possibilities. Logicians have historically analyzed modal assertions as referring to a set of "possible words" (see Kneale \& Kneale, 1962; Portner, 2009). To say that something is possibly the case is to say that it is true in at least one possible world, and to say that something is necessarily the case is to say that it is true in all possible worlds.

Many different systems of modal logic exist (Kaufmann, Condoravi, \& Harizanov, 2006). Each adopts a different set of axioms that affect which inferences can be proved. Different axiom systems affect which modal inferences are valid and which are not (see, e.g., Kripke, 1963). An inference is valid if it yields a conclusion that is true in every case in which the premises are true (Jeffrey, 1981, p. 1). In principle, an infinite number of modal logics exists, but logicians tend to focus on the axioms themselves, which run in parallel with semantic assumptions about the accessibility of one possibile world from another (Kripke, 1963). For instance, the axiom:

$$
\square \mathrm{A} \rightarrow \mathrm{~A}
$$

where ' $\square$ ' is a symbol that stands for the logical notion of necessity, $A$ is any proposition whatsoever, and ' $\rightarrow$ ' denotes material implication. The axiom asserts that the necessity of $A$ material implies $A$. The axiom does not hold in the modal logic "system K" (for "Kripke"), but it does hold another logic, "system T", and it corresponds to the assumption that accessibility is reflexive, i.e., if a proposition is necessary in a world then it holds in that world.

Here is a set of inferences that are invalid in all systems of modal logic:

2a. A or B or both.
b. Therefore, possibly A.
c. Therefore, possibly B.
d. Therefore, possibly A and B.

These inferences seem intuitively reasonable, but the conclusions ( $2 \mathrm{~b}-\mathrm{d}$ ) are invalid in any modal logic. Suppose that $A$ is impossible but $B$ is true. In logic, the premise is true, but (2b) is false. So, the inference is invalid. Similar suppositions show that all the inferences are invalid, and so, no modal logic permits them. Why, then, are the inferences compelling - almost "obvious" - for humans?

Despite some investigations into reasoning about possibilities (e.g., Bell \& Johnson-Laird, 1998; Byrnes \& Beilin, 1991; Goldvarg \& Johnson-Laird, 2000; Inhelder \& Piaget, 1958; Piéraut-Le Bonniec, 1980; Osherson, 1976; Sophian \& Somerville, 1988), no comprehensive theory of human reasoning exists that explains how humans integrate reasoning about facts with reasoning about possibilities. The fundamental mystery is: where do the possibilities come from? Anecdotally, some researchers find that when participants are allowed to write or type out their own responses to a set of reasoning problems, they spontaneously qualify their inferences, e.g., by noting that a conclusion "is possible" or "could be true" or "might follow." These responses are, of course, different ways of expressing modal conclusions. Other research finds that reasoners are capable of carrying out various modal reasoning tasks systematically, e.g., given a set of premises, they are able to determine whether a conclusion is necessary or possible (e.g., Bell \& Johnson-Laird, 1998; Khemlani, Lotstein, Trafton, \& Johnson-Laird, 2015; Newstead \& Griggs, 1983). But only recently have researchers examined reasoners' tendency to endorse modal conclusions from non-modal premises. Hinterecker and colleagues (2016) gave participants a battery of problems in which participants had to endorse or reject different conclusions from modal premises. Contrast this inference:

## 3. A or B or both. <br> Therefore, possibly A and B.

with this one:

## 4. A or B, but not both.

Therefore, possibly A and B.
Reasoners responded sensibly: they accepted (3) most of the time ( $82 \%$ of trials) but they accepted (4) on only a small minority of trials ( $10 \%$; see Hinterecker et al., 2016, Experiment 1). But, both (3) and (4) are invalid in logic.

## Probabilistic logic

Can Hinterecker and colleagues' findings be explained by an alternative theory? The two inferences above may be treated more sensibly in probabilistic logic, hereafter, "plogic", which is a formal system devised by Adams (1975; 1988). P-logic reinterprets validity on probabilistic terms: a conclusion is probabilistically valid (p-valid) only if in any consistent assignment of probabilities its conclusion is at least as probable as its premises. Hence, in (3), the conclusion, possibly ( $A$ and $B$ ), does not rule out any cases,
i.e., it can be true independent of whether $A$ or $B$ are true. The premise, $A$ or $B$ or both, in contrast, rules out the situation in which both $A$ and $B$ are false. And so, the conclusion has a probability greater than that of the premise in (3), and it is p -valid. In a similar vein, (4) is p-invalid because the probability of the conclusion, possibly( $A$ and $B$ ), is 0 given the premise. And so, no matter what probability is assigned to the premise, the inference is p invalid. P-logic is central to recent probabilistic accounts of human reasoning known colloquially as the "new paradigm" (see, e.g., Evans, 2012; Oaksford \& Chater, 2007; Over, 2009; Johnson-Laird, Khemlani, \& Goodwin, 2015).

But, p-logic does not always make sensible predictions. For instance, it predicts that the following inference is p valid:

## 5. A or B, but not both. <br> Therefore, A or B, or both.

The probability of the conclusion in (5) is greater than or equal to that of the premise, and so p-logic predicts that reasoners should make it. (The inference is always valid in logic.) Yet, participants rejected it on $97 \%$ of trials in the aforementioned study by Hinterecker and colleagues. Perhaps a deeper problem with probabilistic accounts is that they do not explain the provenance of modal conclusions, e.g., "possibly A", from non-modal premises. Hence, an alternative account of reasoning is needed to explain modal inferences.

## A model-based theory of modal inference

The mental model theory of reasoning - hereafter, the "model theory" - posits that reasoners draw conclusions by building and scanning iconic representations of possibilities, i.e., mental models (Johnson-Laird, 2006; Johnson-Laird \& Byrne, 1991). The theory assumes that interpreting compound assertions such as those linked by the connectives and, or, and if, yields a set of discrete possibilities. Models mimic the structure of what they represent, i.e., they are iconic (Peirce, 1931-1958, Vol. 4). But, they can also contain abstract tokens, such as symbols denoting negations (Khemlani, Orenes, \& Johnson-Laird, 2012). They can represent temporal sequences of events as multiple models unfold in time the way events do (Bucciarelli, Mackiewicz, Khemlani, \& Johnson-Laird, under review; Khemlani, Mackiewicz, Bucciarelli, \& Johnson-Laird, 2013).

The theory posits two primary systems for reasoning (see, e.g., Johnson-Laird \& Steedman, 1978): a fast system builds mental models and scans them without the use of working memory. A slower system revises models and fleshes them out to yield a set of fully-explicit models. It also searches for alternative models consistent with the premises. It can correct the errors and biases that the fast system yields, but it is subject to the limitations of working memory. The difference between mental models and fully-explicit models is clear when reasoning about disjunctions, e.g., He has the
soup or the salad or both. Mental models abide by a "principle of truth", i.e., they represent what is true in a compound clause, and not what is false. They can flesh out the initial mental models to yield a set of fully-explicit models, i.e., possibilities that specify both what is true and what is false. The mental models of the disjunction above can be depicted in the following schematic diagram:

| soup | salad |
| :--- | :--- |
| soup | salad |

Each row in the diagram denotes a different possibility. Hence, the first row denotes the possibility in which he has the soup. In contrast, a fully-explicit model represents both what is true in each possibility, as well as what is false:

$$
\begin{array}{lr}
\text { soup } & \neg \text { salad } \\
\neg \text { soup } & \text { salad } \\
\text { soup } & \text { salad }
\end{array}
$$

Three primary findings support the model theory. First, inferences from one model are easier than inferences from multiple models (e.g., Johnson-Laird, Byrne, \& Schaeken, 1992). Second, because reasoners tend to build mental models instead of fully-explicit models, they are prone to systematic errors (see Khemlani \& Johnson-Laird, 2017, for a review). Third, reasoners rely on counterexamples to correct erroneous inferences (e.g., Johnson-Laird \& Hasson, 2003).

But, the theory has at least two serious shortcomings. First, it does not integrate facts and possibilities. As a result, it cannot explain the "obvious" inferences in example (2) above. Indeed, no theory of reasoning adequately integrates facts and modal reasoning, but the problem is particularly acute for the model theory, as the theory is based on the representation of possibilities, and so modal reasoning is within its purview. And second, its various computer implementations do not make quantitative predictions (Johnson-Laird \& Yang, 2008). To rectify these shortcomings, we describe a novel assumption about the representation of mental models below, and then we present a new computational model capable of delivering quantitative predictions by varying how models are built and revised.

## The principle of conjunctive possibilities

We amend the model theory to explain where possibilities come from in inferences that make no mention of them with the following principle:

The principle of conjunctive possibilities: By default, compound assertions between clauses refer to conjunctions of possibilities. A clause can be evaluated as possible if it is affirmed in at least one possibility of the conjunctive set. It can be evaluated as necessary if it can be affirmed in all possibilities. And it is deemed factual if it is affirmed in a set of only one possibility.

The principle posits that a disjunction, He has the soup or the salad or both, refers to a set of possibilities, i.e.:

```
possible( soup & \negsalad ) &
possible(-soup & salad ) &
possible( soup & salad )
```

The addition of the principle solves two mysteries of modal reasoning: first, it explains why reasoners are apt to make modal inferences from non-modal assertions. If compound assertions refer to possibilities, then reasoning about possibilities is the default instead of an extension to more basic reasoning patterns (cf. Inhelder \& Piaget, 1958). Second, because the principle is that possibilities are related through conjunction, it allows reasoners to conclude that any of the separate possibilities can be concluded as possible. An immediate consequence of the assumption is that modal inferences are the default, and reasoning about facts is a special case of reasoning about possibilities.

The principle is presaged by recent ideas due to Zimmerman (2000), who proposed that disjunctions refer to lists of alternatives in a "possible worlds" semantics, and Geurts (2005) who extended the idea to disjunctions that concern facts. The principle we propose, however, applies to all sorts of sentential connectives, including disjunctions, conjunctions, conditionals, and even causal relations, e.g., causes, enables, and prevents (Johnson-Laird \& Khemlani, in press; Khemlani, Barbey, \& Johnson-Laird, 2014).

The principle maintains the separation between mental models and fully-explicit models. Hence, it makes all of the same predictions as previous versions of the model theory. It also predicts that reasoners should deem (5) invalid, which we repeat here:

## 5. A or B, but not both. Therefore, A or B, or both.

Both a truth-functional analysis in logic and the notion of pvalidity in p-logic treats (5) as valid. But, if reasoners represent the exclusive disjunction as a conjunction of possibilities, i.e.:

```
possible( soup & \negsalad ) &
possible(\negsoup & salad )
```

then the conclusion does not follow from the representation, because nothing yields the possibility in which both cases hold.

Nevertheless, the previous predictions are qualitative, not quantitative. A veridical simulation of human reasoning needs to provide a quantitative simulation of the extant data. To do so, we developed a novel computational implementation of the model theory, and we tested it against two experiments by Hinterecker et al. (2016). We now describe the computational model and its simulation of data.

## A computational implementation of the model theory

We developed a computational theory of sentential reasoning that integrates reasoning about facts and reas-


Figure 1. A schematic diagram of the computational model of reasoning. The system operates by parsing premises in natural language, constructing mental models and scanning them to formulate initial conclusions (system 1), and then searching for counterexamples and building fully-explicit models to interrogate initial inferences (system 2).
oning about possibilities. It implements the principles of the mental model theory of reasoning (see, e.g., Khemlani \& Johnson-Laird, 2013) and the principle of conjunctive possibilities introduced here. Figure 1 provides a schematic of the program. The computational model is structured around three general systems:
a) A linguistic system uses a grammar and lexicon to parse verbal assertions.
b) An intuitive system (System 1) uses the parse to construct an initial mental model, i.e., a conjunction of possibilities. It also scans the model to formulate initial inferences.
c) A deliberative system (System 2) can flesh out the mental model and search for alternative models. This system can manipulate and update the representations created in System 1, and it can modify conclusions, but it too can fail when a problem calls for more working memory than it has (Khemlani \& Johnson-Laird, 2017).
In the computational model, system 1 does not have access to working memory, and so it can construct only one mental model at a time. It can flesh out the mental model to make it explicit. The probability of doing so is governed by a parameter, $\varphi$. System 2, however, has access to working memory. As a result, the operations of system 1 are faster and more prone to err than system 2. System 2 can operate on multiple models at a time, search for counterexamples, and construct a set of fully explicit models. The probability of calling system 2 is governed by a separate parameter, $\sigma$. In principle, the size of working memory could also be governed by a parameter in order to model individual differences in reasoning.

The system is capable of carrying out a number of inferential tasks, but for brevity, we consider just two:
assessing that a given conclusion is possible, and assessing that it is necessary. In order to assess an inference, the computer model checks that each possibility in the conjunction corresponding to the conclusion is supported by the premises. If they all are, then the conclusion follows of necessity, and if at least one is, then the conclusion is a possibility. The concept of necessity may vary from participant to participant and from problem to problem: some problems may encourage reasoners to check that the models of the premises hold in all models of the conclusion (a strong notion of validity), and some problems may encourage reasoners to check only that all the possibilities to which the conclusions refer hold in the premises (a weaker notion of validity). Consider how you might respond to problems such as this one:
6. Suppose that: A or B, or both.

Does it follow that: A or B, but not both?
The conjunctive possibilities to which the first premise refers are:

```
possible( A & -B ) &
possible(\negA & B ) &
possible( A & B )
```

And the possibilities to which the second premise refers are:

```
possible( A & -B ) &
possible(\negA & B )
```

Reasoners with a strong notion of validity should judge that (6) is invalid, because the models of the premise do not all hold in the models of the conclusion. Reasoners with a weaker notion of validity should assess that (6) is valid, since the models of the conclusion hold in all the models of the premises. To simulate this non-determinism of human reasoning, we built a third and final parameter into the system, $\gamma$, that denotes the probability of the system adopting a weak version of validity.

We applied the computational model to simulate recent data on modal reasoning. The simulations show a close fit between the predictions of the system and humans' inferential behavior after conservative parameter searching.

## Simulations of Hinterecker et al. (2016)

We sought to use the computational implementation of the model theory to simulate participants' performance in Experiments 1 and 3 in Hinterecker et al. (2016), as those two studies are directly pertinent to how reasoners draw modal conclusions from non-modal premises. In Experiment 1, Hinterecker and colleagues gave participants a battery of diagnostic problems that involve disjunctions. Two of those problems tested the tendency to draw modal conclusions from disjunctions that make no mention of possibilities, and two tested the ability to infer an inclusive disjunction from an exclusive one, and an exclusive





Figure 2. Observed (histograms) data and predicted (circles) proportions of accepting the conclusion for different inferences. Top panel: inferences from Hinterecker et al. (2016, Experiment 1). Bottom panel: inferences from Hinterecker et al. (2016, Experiment 2). For each of the problems, the assertion on top denotes the premise and the assertion on the bottom denotes the conclusion.
disjunction from an inclusive one. Figure 2 summarizes the proportion of participants to accept the varying conclusions given the single premise.

The only parameter that could have affected the system's simulations on the problems in Experiment 1 was the $\gamma$ parameter, which dictates how probable it is for participants to make use of a weak notion of validity. An exhaustive exploration of the parameter space yielded an optimal $\gamma$ value of .75 , i.e., the system optimally modeled the data when it stochastically applied weak validity to $75 \%$ of simulated problems. We generated synthetic data by running 1000 simulations of the four inferences in Experiment 1. Figure 2 (top panel) shows the proportion of correct responses in the observations (histograms) and predictions (circles) in the study as a function of the inference. The
computer model matched the participants' performance in the experiment well ( $\mathrm{r}=.99, \mathrm{RMSE}=.10$ ). The predictions of the computer model were in the $99^{\text {th }}$ percentile relative to hypothetical datasets (Khemlani \& Trafton, 2013).

Hinterecker's et al. (2016) Experiment 3 was a more stringent test of reasoners' ability to infer modal conclusions from an inclusive disjunction, $A$ or $B$ or both. For each problem in the experiment, participants assessed the disjunction and then accepted or rejected the one of the following four conclusions: possibly $A$ and $B$, possibly $A$, possibly $B$, possibly not- $A$ and not- $B$. None of these inferences is valid in any known logic, but as Figure 2 (bottom panel) shows, reasoners endorsed three of the four conclusions. We disabled all of the parameters to see how the computer model matched the participants' performance; it did so extremely well ( $\mathrm{r}=.99$, RMSE $=.12$ ), and additional parameter manipulations would have resulted in only nominal changes to the fit.

In sum, the computational model implementing the model theory and the principle of conjunctive possibilities yielded a close fit to the data from Hinterecker et al. (2016).

## General discussion

Reasoners have no difficulty drawing modal conclusions from compound assertions that make no mention of modality. The ability to do so often seems "obvious"; only experts are likely to realize that this inference is invalid in logic:
7. A or B or both.
Therefore, possibly A.

No known logical system designed to deal with modalities, i.e., a modal logic concerning what is possible or necessary, permits the inference above. Reasoners naive to logic may also be surprised to find that both orthodox logic and probabilistic logic render the following inference valid:

## 8. A or B, but not both. <br> Therefore, A or B, or both.

On our account, reasoners are justified in feeling that the invalidity of (7) and the validity of (8) are counterintuitive and incorrect. The model theory of reasoning, which is based on possibilities (Johnson-Laird, 2006), treats compound assertions, such as conjunctions, conditionals, and the disjunctions in (7) and (8), as conjunctions of possibilities. Hence, reasoning about possibilities is fundamental. Reasoners represent possibilities directly, and so modal reasoning is a natural consequence of the way people represent assertions.

The principle of conjunctive possibilities characterizes the inferences in (6) and (7) more intuitively: it predicts that (6) should be deemed valid and (7) should be deemed invalid. And a computational implementation of the principle makes identical predictions, which are validated by recent experiments on modal reasoning by Hinterecker et al. (2016).

At present, no alternative theory of reasoning, whether based on mental logic (e.g., Rips, 1994) or on the probability calculus (e.g., Oaksford \& Chater, 2007), can explain these phenomena of modal reasoning. Moreover, no computational model of reasoning, whether in psychology, artificial intelligence, or logic, characterizes the inferences in the same manner as the system we outlined above. The reason, as we argue, is that everyday reasoning is based on possibilities, not probabilities or truth-functions.

## Acknowledgments

This work was supported by a grant from the Office of Naval Research to SK. We are also grateful to many colleagues for help and advice. They include Aron Barbey, Paul Bello, Monica Bucciarelli, Ruth Byrne, Nick Chater, Nicole Cruz, Alan Garnham, Gernot Kleiter, Markus Knauff, Philipp Koralus, Paolo Legrenzi, Max Lotstein, Robert Mackiewicz, Juan García-Madruga, Geoff Goodwin, Todd Gureckis, Greg Murphy, Mike Oaksford, Jane Oakhill, Klaus Oberauer, Isabelle Orenes, Dan Osherson, Barbara Partee, Guy Politzer, Cristina Quelhas, Marco Ragni, Célia Rasga, Bob Rehder, Carlos Santamaría, Walter Schaeken, Mark Steedman, and Greg Trafton.

## References

Adams, E. W. (1975). The logic of conditionals. Dordrecht, the Netherlands: Reidel.
Adams, E. W. (1998). A primer of probability logic. Stanford, CA: CSLI Publications.
Bell, V., \& Johnson-Laird, P.N. (1998). A model theory of modal reasoning. Cognitive Science, 22, 25-51.
Bucciarelli, M., \& Johnson-Laird, P.N. (2005). Naive deontics: a theory of meaning, representation, and reasoning. Cognitive Psychology, 50, 159-193.
Bucciarelli, M., Khemlani, S., \& Johnson-Laird, P. N. (2008). The psychology of moral reasoning. Judgment and Decision Making, 3.
Bucciarelli, M., Mackiewicz, R., Khemlani, S., \& Johnson-Laird, P. N. (under review). Recursion in children's comprehension and formulation of algorithms. Manuscript under review.
Byrnes, J. P., \& Beilin, H. (1991). The cognitive basis of uncertainty. Human Development, 34, 189-203.
Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. Thinking \& Reasoning, 18, 5-31.
Goldvarg, Y., \& Johnson-Laird, P. N. (2000). Illusions in modal reasoning. Memory \& Cognition, 28, 282-294.
Geurts, B. (2005). Entertaining alternatives: disjunctions as modals. Natural Language Semantics, 13, 383-410.
Hinterecker, T., Knauff, M., \& Johnson-Laird, P. N. (2016). Modality, probability, and mental models. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42, 1606-1620.
Inhelder, B., \& Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. London: Routledge \& Kegan Paul.
Jeffrey, R. J. (1981). Formal logic: Its scope and limits (2nd Ed.). New York: McGraw-Hill.
Johnson-Laird, P. N. (2006). How we reason. New York, NY: Oxford University Press.
Johnson-Laird, P. N., \& Byrne, R. M. J. (1991). Deduction. Hillsdale, NJ: Erlbaum.
Johnson-Laird, P. N., Byrne, R. M. J., \& Schaeken, W. S. (1992). Propositional reasoning by model. Psychological Review, 99.
Johnson-Laird, P. N., \& Hasson, U. (2003). Counterexamples in sentential reasoning. Memory \& Cognition, 31, 1105-1113.

Johnson-Laird, P. N. \& Khemlani, S. (in press). Mental models and causation. In M. Waldmann (Ed.), Oxford Handbook of Causal Reasoning.
Johnson-Laird, P. N., Khemlani, S. S., \& Goodwin, G. P. (2015). Logic, probability, and human reasoning. Trends in Cognitive Sciences, 19, 201-214.
Johnson-Laird, P. N., \& Steedman, M. J. (1978). The psychology of syllogisms. Cognitive Psychology, 10, 64-99.
Johnson-Laird, P. N., \& Yang, Y. (2008). Mental logic, mental models, and computer simulations of human reasoning. In Sun, R. (Ed.) Cambridge Handbook of Computational Psychology. Cambridge: Cambridge University Press.
Kaufmann, S., Condoravdi, C. \& Harizanov, V. (2006). Formal approaches to modality. In Frawley, W. (Ed.). The Expression of Modality. Berlin, New York: Mouton de Gruyter.
Khemlani, S., Barbey, A., \& Johnson-Laird, P. N. (2014). Causal reasoning with mental models. Frontiers in Human Neuroscience, 8.
Khemlani, S., \& Johnson-Laird, P. N. (2013). The processes of inference. Argument \& Computation, 4, 1-20.
Khemlani, S., \& Johnson-Laird, P. N. (2017). Illusions in reasoning. Minds \& Machines, 27, 11-35.
Khemlani, S., Lotstein, M., Trafton, J. G., \& Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. Quarterly Journal of Experimental Psychology, 68, 2073-2096.
Khemlani, S., Orenes, I., \& Johnson-Laird, P. N. (2012). Negation: a theory of its meaning, representation, and use. Journal of Cognitive Psychology, 24.
Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., \& Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. Proceedings of the National Academy of Sciences of the United States of America, 110, 16766-16771.
Khemlani, S. \& Trafton, J. G. (2014). Percentile analysis for goodness-of-fit comparisons of models to data. In Proceedings of the 36th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Kneale, W., \& Kneale, M. (1962). The development of logic. Oxford, England: Oxford University Press.
Kripke, S. (1963). Semantical considerations on modal logic. Acta Philosophica Fennica, 16, 83-94.
Newstead, S. E., \& Griggs, R. A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. Journal of Verbal Learning and Verbal Behavior, 22, 535-546.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. New York, NY: Oxford University Press.
Osherson, D. N. (1976). Logical abilities in children, Vol. 4: Reasoning and concepts. Hillsdale, NJ: Lawrence Erlbaum Associates.
Over, D. (2009). New paradigm psychology of reasoning. Thinking \& Reasoning, 15, 431-438.
Peirce, C. S. (1931-1958). Collected papers of Charles Sanders Peirce. 8 vols. C. Hartshorne, P. Weiss, and A. Burks, (Eds.). Cambridge, MA: Harvard University Press.
Piéraut-Le Bonniec, G. (1980). The development of modal reasoning: Genesis of necessity and possibility notions. New York: Academic Press.
Portner, P. (2009). Modality. Oxford, UK: Oxford University Press.
Rips, L. J. (1994). The psychology of proof. Cambridge, MA: MIT Press.
Sophian, C., \& Somerville, S. C. (1988). Early developments in logical reasoning: Considering alternative possibilities. Cognitive Development, 3, 183-222.
Zimmermann, T. E. (2000). Free choice disjunction and epistemic possibility. Natural Language Semantics, 8, 255-290.

# Numerical and Non-numerical Magnitude Estimation 

Dan Kim<br>The Ohio State University<br>John Opfer<br>The Ohio State University


#### Abstract

Despite a heated debate regarding a cognitive mechanism of magnitude representation, little has been done to directly compare numerical and non-numerical estimation and provide a unified account of the two processes. In the current study, we examined estimation of numerical and non-numerical quantities on a continuum using various psychophysical functions. Inconsistent with the proportion reasoning and measurement skills accounts, estimates of both numerical and non-numerical quantities were better predicted by the logarithmic-linear model than by cyclic power models. Furthermore, individual differences in the degree of logarithmic compression was highly correlated over tasks, whereas bias measures from competing models did not show such associations. These findings suggest that estimation of both numerical and non-numerical magnitude is processed via shared representation systems that are logarithmically or linearly constructed.


# Variables Involved in Selective Sustained Attention Development: Advances in Measurement 

Jaeah Kim (jaeahk@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue Pittsburgh, PA 15213 USA

Anna Vande Velde (avandeve@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue Pittsburgh, PA 15213 USA

Erik D. Thiessen (thiessen@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue Pittsburgh, PA 15213 USA

Anna V. Fisher (Fisher49@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue Pittsburgh, PA 15213 USA


#### Abstract

Selective sustained attention (SSA) is an important cognitive process that enables everyday functioning and task performance by allowing us to: 1) choose components of our environment to process at the exclusion of others and 2) maintain focus on those components over time. Although SSA is known to undergo rapid and marked changes during the preschool and early primary school children years, there has been a paucity of behavioral data on these years of development due to a lack of child-appropriate testing paradigms. TrackIt is a paradigm that was recently developed to fill the previously existing measurement gap for SSA in these years. In this study, we analyzed errors that children (aged 3-7) make when performing TrackIt, to better understand what factors drive improvement in their performance over age. In addition, we manipulated parameters within TrackIt to place varying levels of demand on children's SSA, and measured behavioral performance over age, with the goal of measuring and characterizing developmental trends during these years. Since TrackIt is still a recent paradigm, our results also help suggest appropriate parameter settings for calibrating the task to different age groups.


Keywords: selective sustained attention; TrackIt

## Introduction

Selective sustained attention (SSA) is an important cognitive process that enables everyday functioning and task performance by allowing us to: 1) choose components of our environment to process at the exclusion of others and 2) maintain focus on those components over time. SSA is known to rely on both endogenous factors (e.g., internal goals) as well as exogenous factors (e.g., stimulus salience) (O’Connor \& Manley, 2004) -- studying specifically how these factors interact and work together in guiding attention contributes to a growing understanding of SSA's
mechanisms. Task paradigms that allow simultaneous investigation of both exogenous and endogenous factors of SSA have been available for adults and infants but not for preschool and early primary school children ( $\sim 3-7$ years) until recently (for review, see Fisher \& Kloos, 2016). These years are particularly important from a research standpoint because data from infants and adults suggests that SSA develops significantly during these intermediate years (Oakes, Kannass, \& Shaddy, 2002). TrackIt, a paradigm developed specifically to fill this measurement gap, is designed to be appropriately challenging for a range of developmental years including the preschool years, with varying parameters for adjustment of difficulty across ages (Fisher et al., 2013).

Prior studies with TrackIt demonstrated that children improve on the task between 3 and 5 years of age (Fisher et al., 2013, Erickson et al., 2015), consistent with the overall developmental pattern of improvement in SSA with age. In order to investigate this improvement more closely, the current study looked at 1) what factors tend to drive the failures (errors), and what, of those, consequently improve to drive the overall performance improvement (see "Factors driving improvement" below), and 2) what the behavioral trajectories representing this improvement look like across an expanded age range. To delve into this issue, we manipulated parameters of TrackIt to place varying levels of demand on children's selective sustained attention to achieve a coarse mapping of behavioral performance in several parameter combinations over ages 3 to 7. In contrast to prior studies using TrackIt that focused on the analysis of correct responses (Fisher et al., 2013; Erickson et al., 2015), this study also examined the patterns of errors as a function of task difficulty and age.

## Factors Driving Improvement

This study introduces a new addition to the TrackIt program: error analysis. This functionality (a software update that is now part of the TrackIt program that is freely available to interested researchers) adds, to the behavioral output, information on the types of errors that participants make. Analyzing the types of errors that children make over development may provide greater insight into what factors constitute the overall improvements that we see in children's SSA performance. For example, some error types help to distinguish between behavioral errors due to failure of SSA and those due to insufficient visuo-spatial resolution. Also, finding that a significant proportion of errors can be related to failure of SSA would help validate TrackIt as a task assessing attention. Thus, the first goal of this study was to present preliminary analyses of error type breakdown over age and difficulty.

## Behavioral Trajectories

Mapping out age-related changes in performance within the multi-dimensional parametric space of variables (i.e., number of distractors, grid size, speed of objects, type of distractors) involved in visual attention serves two important purposes: 1) it begins to fill in the empirical gap in characterizing children's visual SSA development within a single consistent measurement framework, and 2) it suggests initial parameter selection ranges for age groups, to guide researchers using TrackIt (Doebel et al., 2015). Hence, the second goal of this study is to present preliminary findings on parameter space mappings.

## Method

## Participants

Participants were 144 typically developing children (71 female, $M_{\text {age }}=5.08$ years) recruited from local preschools, day care centers, and elementary schools in Pittsburgh, PA. See Table 2 below for a breakdown of participant age statistics.

## Materials and Apparatus

Stimuli were presented on a Lenovo touchscreen laptop with physical screen dimensions $19.1 \mathrm{~cm} \times 34.2 \mathrm{~cm}$ and pixel dimensions 1920x1080 pixels. Participants were seated at a desk facing the screen with their heads about 2 feet away from the screen.

## TrackIt Task

In this task (freely available for download at http://www.psy.cmu.edu/~trackit/), participants were asked to visually track a single target object as it moved on a grid among moving distractor objects. For each trial, the target and distractor objects were randomly picked without replacement from a set of unique objects spanning 9
different shapes with 9 different color possibilities (81 objects in total). ${ }^{1}$ See Figure 1 for examples.

At the beginning of each trial, the objects appeared on the grid, centered in distinct grid cells, and the target object was indicated by a red circle around it. The initial positions of the objects were randomized. The experimenter started each trial with a button press after ensuring the participant was ready to begin.

Upon starting the trial, the red circle disappeared, and the objects began to move in linear trajectories from grid cell to grid cell at a constant speed. At the end of each trial, all objects disappeared from the screen, and the participants were asked to indicate with their finger (on the touch screen) which grid cell the target object was last in before it disappeared.

The sequence of positions in the path of each objects was randomized, with one restriction for just the target: the target had to be in the center of a grid cell at the end of a trial, to reduce ambiguity for the participant in determining its final location. Due to this restriction, the length of trials was not fixed, but varied slightly from trial to trial (to allow the target to reach the center of a grid cell. The minimum trial length was set to 10 seconds. The parameters -- grid size, number of distractors, and speed of objects in pixels per second -- were determined from prior testing in TrackIt with a separate group of 3- to 5-year old children (Fisher et al., 2013), and organized according to participant age and difficulty level as seen in Table 1. Object motion display was set to 30 frames per second.

Table 1: TrackIt parameter combination used in each difficulty level.

|  | Age <br> Group <br> (years) | Grid <br> Size | \# of <br> Distractors | Object <br> Speed <br> (pix/s) |
| :--- | :---: | :---: | :---: | :---: |
| Difficulty | $3-5$ | $4 \times 4$ | 4 | 500 |
| Level 1 | $4-6$ | $6 \times 6$ | 6 | 500 |
| Level 2 | 7 | $6 \times 6$ | 8 | 800 |
| Level 3 |  |  |  |  |

Note: pix/s = pixels/second.

We assessed three different difficulty levels, administered to different age groups, as shown in Table 1. Separate groups of participants were tested in each difficulty level. The sample size per age and difficulty level is presented in Table 2. It should be noted that, ultimately, we aim to obtain a large-scale representative sample of

[^118]Table 2: Sample sizes and age statistics of each age group, for each difficulty level

| Age Group (years) | Difficulty Level 1 |  | Difficulty Level 2 |  | Difficulty Level 3 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\mathrm{n} / \mathrm{m} / \mathrm{f}$ | $\begin{gathered} \text { Age } \\ \text { Mean (Std) } \end{gathered}$ | $\mathrm{n} / \mathrm{m} / \mathrm{f}$ | $\begin{gathered} \text { Age } \\ \text { Mean (Std) } \end{gathered}$ | n/m/f | $\begin{gathered} \text { Age } \\ \text { Mean (Std) } \end{gathered}$ |
| 3 | 25/13/12 | 3.64 (0.21) | - | - | - | - |
| 4 | 17/9/8 | 4.64 (0.27) | 19/6/13 | 4.69 (0.22) | - | - |
| 5 | 26/14/12 | 5.38 (0.23) | 41/19/22 | 5.43 (0.27) | - | - |
| 6 | - | - | 6/4/2 | 6.34 (0.25) | - | - |
| 7 | - | - | - | - | 10/8/2 | 7.27 (0.16) |

Note: $n / m / \mathrm{f}=$ sample size/\# male/\# female.
participants at each age and difficulty level; the present paper reports the initial findings from this study.

Memory Check At the conclusion of each trial, children were presented with 4 shapes that could have served as target objects in this task (one of which was actually the target) and asked to point to the shape they had been tracking (see Figure 1). The responses to memory check questions were recorded by the children's touch screen responses. The memory checks were introduced to help discriminate between two possible reasons why a participant may fail to correctly report the location where the target object disappears. The first possibility is that encoding of the identity of the target object may be insufficiently strong to persist through an entire trial - this would indicate an encoding failure. The second possibility is that a child may track distractors for a part of the trial despite remembering which object was supposed to be watched - this would indicate the failure of selective sustained attention. The target was colored as in the trial, while the remaining 3 shapes and their colors were sampled without replacement from the remaining 8 shapes and colors.

## Design and Procedure

The experimenter administered the TrackIt task to participants in a quiet room or hallway. At the beginning of the task, participants were told that: 1) the objects will start moving around the grid when the experimenter presses a button; 2) the goal is to follow the target object with their eyes; 3) at some point the objects will suddenly disappear, and their job is to point to where the target object was when it disappeared. Each trial was followed by a baseline screen displaying a smiley face, a memory check screen, and a second smiley face baseline (in that order). Participants were told that the smile did not indicate a correct answer and rather meant we were happy they were playing our game. See Figure 1 for a diagram of the task sequence.

Participants completed 11 trials of the task. The first trial was a practice trial and was completed with assistance from the experimenter who traced the moving target with their index finger. The first trial was accordingly omitted from analysis. Participants were then told that they would need to complete the rest of the task by themselves, tracking the target with their eyes only.


Figure 1. The TrackIt task pipeline. A single trial, followed by smiley face, memory check, and smiley face.

Error Analysis Next, we were interested in better understanding what types of tracking errors participants were making. Tracking errors were any answers in the main TrackIt task that weren't the correct cell that the target ended in. Tracking errors were thus further classified based on the incorrect grid cell response indicated by the participant, in relation to the final positions of the target and distractors on the grid. Specifically, in addition to a correct response, we considered 5 types of errors:

Spatial Resolution: The response was a cell adjacent to the correct grid cell, but was not also adjacent to a distractor.

Distractor: The response was a cell that contained a distractor.

Distractor Spatial Resolution: The response was a cell that did not contain a distractor, was adjacent to a cell that contained a distractor, and was not also adjacent to the correct grid cell.

Uncategorizable: The response was a cell that did not contain a distractor, and was adjacent to both the correct grid cell and a cell that contained a distractor.

Other: None of the above. That is, the response was a cell that did not contain a distractor, and was adjacent to neither the correct grid cell, nor a cell that contained a distractor.

In the above, "adjacent" cells are defined as those within one horizontal, vertical, or diagonal step of a given cell. (Cells are not considered adjacent to themselves.) For example, corner cells have 3 adjacent cells, edge cells have 5 adjacent cells, and other cells have 8 adjacent cells.

Note that finding that a significant proportion of errors are distractor-related errors (distractor, distractor spatial resolution, or uncategorizable) would help to validate TrackIt as a task assessing attention; if we find that kids make many spatial resolution errors, it may indicate that the performance is limited by kids' visuo-spatial acuity. In contrast, if we find that kids make predominantly "other" errors, it could suggest that they lose interest in the task entirely or do not understand the task and respond randomly.

## Results

## Memory Check

Responses to individual memory check questions (i.e., which object were you supposed to watch?) were averaged over the 10 experimental trials to yield a Memory Accuracy score for each participant. Memory Accuracy data are presented in Figure 2.


Figure 2. Tracking accuracies for each age and difficulty level, both with and without memory-incorrect trials.

In all conditions and age groups Memory Accuracy was above chance ( $25 \%$ given four response options, all one-sample $t \mathrm{~s}>6.2, p \mathrm{~s}<0.0001$ ). To investigate possible effects of age and difficulty, memory accuracy scores were submitted to a 2-way ANOVA with both age as difficulty level as between-subject factors. This analysis indicated a
main effect of age $(F(2,128)=32.2, p<0.0001) .{ }^{2}$ There was no effect of difficulty and no age-by-difficulty interaction (both $F \mathrm{~s}<1.34$, $p s>0.24$ ). Therefore, any differences in object tracking accuracy between difficulty levels were unlikely to stem from differences in the strength of encoding of the target objects.

## Error Analysis

For this analysis, we excluded trials in which the participant failed the memory check, as encoding errors were a separate type of error that we analyzed separately. We compared the rate of each error type to chance, assuming that the participant response was randomly distributed over the incorrect squares of the grid. Chance was estimated by simulating final states of 10,000 TrackIt trials for each level. Chance levels are given in Table 3.

Table 3: Chance probability of each error type assuming the participant response is uniformly distributed over the grid.

| Error Type | Level 1 | Level 2 | Level 3 |
| :---: | :---: | :---: | :---: |
| Correct | 0.0625 | 0.0278 | 0.0278 |
| Spatial Resolution | 0.0434 | 0.0426 | 0.0277 |
| DSR | 0.3606 | 0.1514 | 0.4814 |
| Distractor | 0.2131 | 0.4449 | 0.1962 |
| Uncategorizable | 0.2114 | 0.1004 | 0.1084 |
| Other | 0.109 | 0.2329 | 0.1585 |

Note: DSR = Distractor Spatial Resolution.
Given that the participant made a tracking error, the average portions of error that were distractor errors was consistently significantly above chance in Level 1 (3-year olds: one-sample $t=3.352, \quad p<0.005 ; 4$-year olds: $t=4.117$, $p<0.001$; 5-year olds: $t=4.756, p<0.0001$ ), Level 2 (4-year olds: $t=4.831, p<0.0001$; 5-year olds: $t=3.805, p<0.001$; 6 -year olds: $t=9.869, p<0.0001$ ), and Level 3 (7-year olds: $t=14.065, p<0.0001$.

In order to understand how error types change with increasing age, we regressed each error type proportion over age. The $\beta$ coefficients and $F$ - and $p$-values for each error type and difficulty level are given in Table 4. In particular, note that only the Distractor, Distractor Spatial Resolution, and Uncategorizable errors in Level 1 show significant decreases with age.

## Tracking Accuracy

[^119]For analyzing tracking accuracy, we included all trials (even those for which the memory check was failed), because we are interested in the true performance of subjects in order to calibrate TrackIt. Furthermore, as shown by a plot of tracking accuracy both including and excluding incorrect

Table 4: Linear regression results from regressing error type proportions over age

| Difficulty Level 1 |  |  |  |
| :---: | :---: | :---: | :---: |
| Error Type | $\beta$ | $F(1,66)$ | $p$ |
| Spatial Resolution | -0.0203 | 1.33 | 0.253 |
| DSR | -0.0706 | 6.96 | $0.010^{*}$ |
| Distractor | -0.0653 | 7.54 | $0.0078^{* *}$ |
| Uncategorizable | -0.0444 | 4.81 | $0.0318^{*}$ |
| Other | -0.0101 | 0.977 | 0.327 |
| Difficulty Level 2 |  |  |  |
| Error Type | $\beta$ | $\mathrm{F}(1,65)$ | p |
| Spatial Resolution | -0.00915 | 1.07 | 0.304 |
| DSR | 0.0125 | 0.185 | 0.669 |
| Distractor | -0.0229 | 0.644 | 0.425 |
| Uncategorizable | -0.0320 | 1.63 | 0.206 |
| Other | -0.0142 | 0.540 | 0.465 |

Note: DSR = Distractor Spatial Resolution.

* $\mathrm{p}<0.05$. ** $\mathrm{p}<0.01$
memory response trials (see Figure 3), filtering by memory check had little effect on the tracking accuracy scores. For all difficulty levels in all age groups, tracking accuracy was significantly above chance (chance is $1 / 16$ for Level 1 and $1 / 36$ for Levels 2 and 3, $t s>3.9, p s<0.0005$ ).

For each of the first two difficulty levels, we saw a significant upward trend effect by an F-test on linear regression ( $\beta=0.2302, F=38.33, p<0.0001$ for Level 1 and $\beta=0.1427, F=7.605, p<0.01$ for Level 2). We could not assess a trend for difficulty Level 3 because we only had one age group for that level.


Figure 3. Tracking accuracies for each age and difficulty level, both with and without memory-incorrect trials.

For difficulty Level 1, tracking accuracy of 3-year olds was significantly below that of 4 -year olds (two-sample $t=-5.05, p<0.0001$ ), but tracking accuracy of 4-year olds was not significantly below that of 5 -year olds (two-sample $t=-1.02, p=0.315$ ). Similarly, for difficulty Level 2 , tracking accuracy of 4 -year olds was significantly below that of 5-year olds (two-sample $t=-2.18, p<0.033$ ), but tracking accuracy of 5-year olds was not significantly below that of 6 -year olds (two-sample $t=-0.88, p=0.382$ ).

In the two age groups that performed two difficulty levels (4-5 year olds), two-sample t-tests revealed that performance differences between difficulty levels were not significant ( $t s<1.11, p s>0.11$ ).

## Discussion

The first purpose of this study was to gain insight into the factors driving improvement by investigating the types of errors made by children. A second purpose was to explore the multidimensional parameter space available within TrackIt, with the goal of identifying both developmental milestones in terms of TrackIt performance as well as appropriate settings for use with children.

## Memory Accuracy

Memory accuracy results indicate that encoding error is more prominent in younger children and improves significantly over age. On the other hand, memory accuracy did not differ significantly across difficulty levels, nor was there an age-difficulty interaction effect. Both of these results are encouraging because they suggest that encoding error does not become a confound when using TrackIt with different difficulties across age groups.

## Error Analysis

As discussed above, the proportion of distractor errors was consistently significantly above chance in every age group and difficulty. In Level 1 difficulty, distractor, distractor spatial resolution, and uncategorizable errors (all distractor-related errors) significantly decreased over age. Noting that uncategorizable errors indicate a combination of spatial and distractor spatial resolution errors, these together suggest that distractors' effect on performance decreases with increasing age.

On the other hand, the reduction in both spatial and distractor spatial resolution errors may also stem from a reduction in errors due to visuospatial resolution. While this was a known confound when analyzing the improvement in tracking accuracy over age, our analysis enables us to partially isolate these two sources of improvement by showing more specifically that distractor errors decrease over time. Since distractor errors are associated only with SSA, and not spatial resolution, this provides a stronger
suggestion (as compared to previous results showing only improvement in TrackIt performance) that the improvement in TrackIt performance over age indeed reflects SSA development.

As with previous analyses, we found greatest improvements in performance between 3- and 4-year olds (see Figure 3), which may explain why the significant change in distractor, spatial resolution, and distractor spatial resolution errors was observed only in difficulty Level 1, the only difficulty level at which we tested 3-year olds. We hypothesize that one possible cause of these results, given that changes in the proportion of distractor-related errors occur primarily between ages 3 and 4, is that these ages may be an especially critical period of rapid SSA development.

## Tracking Accuracy

In our tracking accuracy results, we observed significant developmental upward trends with age in difficulty Levels 1 and 2, as shown in Figure 3. However, more specific analyses of each difficulty level revealed ceiling effects. These suggest that the parameter combinations for Level 1 and Level 2 may be appropriate settings for assessing 3- and 4-year olds, respectively, insofar as they avoid ceiling effects, but more difficult parameter combinations may be necessary for sensitive measurement with older children.

Since performance of 4- and 5-year olds did not drop significantly from Level 1 to Level 2, a linear increase in number of distractors and grid size with age does not seem to be enough to preserve difficulty across age groups.

## Limitations and future directions

Our study did not include 2 year-olds and had limited samples of 6- and 7-year olds. Since significant improvement was observed between 3 and 4 years of age, it may be important to look at 2-year olds also.

The behavioral output of TrackIt is limited in that it records only the participant's response at the end of the trial. In particular, we do not know if participants are continuously attentive to the target throughout the trial (on correct trials) or when participants cease to attend to the target (on incorrect trials). Currently, studies are being run in the lab which combine eye-tracking technology with TrackIt and make this information accessible, potentially giving us a more complete picture of how participants behave during the TrackIt task.

## Conclusion

The findings of this study lay the foundation for further work using TrackIt to study SSA development over a range of ages by a) identifying parameter combinations appropriate for certain age groups, b) discounting reduction in encoding errors as a confounding source of performance improvement over age, and c) enriching the behavioral output of TrackIt with information about the types of errors children make, and hence the sources of their performance
improvements over time. Because of its parametric flexibility, TrackIt can assess SSA across a wide range of ages in the same basic task, lowering the risk that changes measured across age are due to different tasks. Additionally, TrackIt has good psychometric properties in general (test-retest reliability, predictive validity, and now a moderate degree of mapping of parametric space). TrackIt thus provides a practical and novel way of measuring attention in an age-range where we know rapid changes occur, but which we haven't had a task to assess with any degree of sensitivity.

## Acknowledgements

We thank the children, parents, and teachers who made this work possible: CMU Children's School, Sacred Heart Elementary School, Glenn Ave Preschool, Winchester Thurston (North Hills Campus), Beth Shalom Early Learning Center, Mt. Assisi Academy Preschool, and Campus School of Carlow University. The work reported here was supported by the National Science Foundation through a grant awarded to A.V.F. and E.D.T. (BCS-1451706).

## References

Doebel, S., Barker, J. E., Chevalier, N., Michaelson, L., \& Munakata, Y. (2015). Getting ready to use control: Advances in the measurement of proactive control in young children. Poster presented at the Biennial Meeting of the Cognitive Development Society. Columbus, OH.
Erickson, L. C., Thiessen, E. D., Godwin, K. E., Dickerson, J. P., \& Fisher, A. V. (2015). Endogenously and exogenously driven selective sustained attention: contributions to learning in kindergarten children. Journal of experimental child psychology, 138, 126-134.
Fisher, A., \& Kloos, H. (2016). Development of selective sustained attention: The Role of Executive Functions. In L. Freund, P. McCardle, and J. Griffin (Eds.), Executive Function in Preschool Age Children: Integrating Measurement, Neurodevelopment and Translational Research (pp. 215-237). Washington, DC: APA Press.
Fisher, A., Thiessen, E., Godwin, K., Kloos, H., \& Dickerson, J. (2013). Assessing selective sustained attention in 3- to 5-year-old children: Evidence from a new paradigm. Journal of Experimental Child Psychology, 114(2), 275-294.
Oakes, L., Kannass, N., \& Shaddy, J. (2002). Developmental changes in endogenous control of attention: The role of target familiarity on infants' distraction latency. Child Development. 73:1644-1655.
O'Connor, C., Manly, T., Robertson, H., Hevenor, J., Levine, B. (2004). An fMRI study of sustained attention with endogenous and exogenous engagement. Brain and Cognition, 54(2), 113-135.

# Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why? 

Max Kleiman-Weiner ${ }^{1}$ (maxkw@mit.edu), Alex Shaw ${ }^{2}$ (ashaw1@uchicago.edu) \& Joshua B. Tenenbaum ${ }^{1}$ (jbt @mit.edu)<br>${ }^{1}$ Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139<br>${ }^{2}$ Department of Psychology, University of Chicago, Chicago, IL 60637


#### Abstract

Successful and repeated cooperation requires fairly sharing the spoils of joint endeavors. Fair distribution is often done according to preferences for equitable outcomes even though strictly equitable outcomes can lead to inefficient waste. In addition to preferences about the outcome itself, decision makers are also sensitive to the attributions others might make about them as a result of their choice. We develop a novel mathematical model where decision makers turn their capacity to infer latent desires and beliefs from the behavior of others (theory-of-mind) towards themselves, anticipating the judgments others will make about them. Using this model we can construct a preference to be seen as impartial and integrate it with preferences for equitable and efficient outcomes. We test this model in two studies where the anticipated attribution of impartiality is ambiguous: when one agent is more deserving than the other and when unbiased procedures for distribution are made available. This model explains both participants' judgments about the partiality of others and their hypothetical decisions. Our model argues that people avoid inequity not only because they find it inherently undesirable, they also want to avoid being judged as partial.


Keywords: fairness, social cognition, theory-of-mind, decision making, Bayesian models

## Introduction

From the distribution of wealth across society to the distribution of dessert at the end of a dinner party, humans seem uniquely capable of enlarging the size of the pie and sharing it fairly (Tomasello, 2014). We make these decisions guided by normative principles such as efficiency, which says to maximize the total utility of the group and fairness, which says in part that distributions should be both equitable and impartial. We also use these principles intuitively when judging whether others' decisions are fair when considered from an impartial or objective perspective (Rawls, 1971; Nagel, 1986).

In the real world where resources aren't perfectly divisible, these principles can often come into conflict. It is well known that efficient allocations of resources are often inequitable and equitable allocations of resources are often inefficient - they leave some of the pie on the table. For example, if Alice has one apple and Bob has none and we take Alice's apple and throw it out, Alice and Bob are in a more equitable state but the total welfare (efficiency) is reduced. This is called inefficient equity. Even young children prefer inefficient equity: they prefer to destroy a resource rather than distribute it inequitably (Blake \& McAuliffe, 2011; Shaw \& Olson, 2012). Preferences for equity and efficiency are often captured quantitatively by directly deriving them from the outcomes. For instance, efficiency might correspond to the total or average outcome among a group of agents and inequity might correspond to the differences between the outcomes of different agents (Adams, 1965; Fehr \& Schmidt, 1999).

While early work focused on whether a given outcome is perceived as fair (Adams, 1965; Fehr \& Schmidt, 1999), there is now growing evidence that decision makers are sensitive to what their choice signals about themselves. Specifically, inequity created without showing partiality can be fair. If both Alice and Bob are equally deserving but there is only one apple, a decision maker might avoid giving it to either one in order to avoid an outcome that is neither equitable nor impartial. For instance, if the decision maker decided to give the apple to Alice an observer would infer that the decision maker is partial to Alice. However, if the decision maker can flip a coin or access another source of randomness and use the chance outcome to determine who should get the apple, the decision maker can create inequity but without worrying about others attributing partiality (Shaw \& Olson, 2014; Choshen-Hillel, Shaw, \& Caruso, 2015).

Both adults and children adjust their distributional preferences depending on whether they are the ones choosing or not. For instance, people are usually dissatisfied with receiving less than an equally worthy counterpart, but when they created the inequity themselves they were more likely to find this acceptable (Choshen-Hillel \& Yaniv, 2011). Adults and children are willing to create inequity that disadvantages themselves but are less willing to create inequity that could be interpreted as favoritism or nepotistic preferences (ChoshenHillel et al., 2015). These results are incompatible with explanations of social preferences that only consider an aversion to inequitable outcomes or other preferences that are directly derived from outcomes. Understanding how to combine these conflicting perspectives (efficiency vs. equity and equity vs. impartiality) is a challenge that we can address with computational modeling. Specifically, how might a flexible preference for these normative values be integrated together and flexibly applied?

Computationally, preferences like impartiality are significantly more sophisticated than just evaluating expected outcomes. We propose that an aversion to partiality is an aversion to having ones actions appear partial to others. Thus to evaluate whether an action will appear partial requires anticipating how one's actions will be interpreted by others. This requires a mentalistic theory-of-mind: the capacity to interpret behavior as being driven by beliefs, desires and intentions (Dennett, 1989). The same choice made in a different context or from a different set of alternatives might be evaluated differently as it will carry different information about the underlying goals and desires that drove the choice. For instance, if a decision maker can choose to give his colleague


Figure 1: An influence diagram (ID) is a directed acyclic graph over three types of nodes: state nodes (circles), decision nodes (rectangles), and utility nodes (diamonds). Directed edges between nodes determine causal dependencies. State and utility nodes take values that depend on the values of their parent nodes. The total utility to the decision maker is the sum over the utility nodes. Green and red utility nodes correspond to rewards and costs respectively. The value of decision nodes is freely chosen by the decision making agent according to equation (4). (a) ID of the Base Decision Maker. Merit corresponds to $\boldsymbol{\gamma}$ and the Inequity and Efficiency nodes corresponds to the first and second components of equation (3) (b) ID of the Judge which infers whether a base decision maker was partial given an observation of her action, $P($ partial $\mid a)$. (c) The Constructed Social Preference recursively builds on the Base Decision Maker adding an aversion to appearing partial $\left(U_{P}\right)$. (d) Simulated results when the decision maker can allocate $\$ 1,000$ to one agent and $\$ 100$ to another or the value on the x -axis to both agents when both agents are equally meritorious. The Constructed Social Preference is more likely to select the wasteful equal option to avoid an attribution of partiality.
either $\$ 100$ or $\$ 1,000$ and chooses to give him $\$ 1,000$ we might infer that he likes his colleague. However if his choices were to give either $\$ 1,000$ or $\$ 2,000$, giving $\$ 1,000$ signals a dislikes for his colleague. Thus the same action requires a different interpretation depending on the unchosen option. Furthermore, the capacity for theory-of-mind can affect distributional preferences: previous work found that children with a more developed theory-of-mind were more likely to give fair offers in the ultimatum game (Takagishi, Kameshima, Schug, Koizumi, \& Yamagishi, 2010).

In this work, we propose that preferences over the beliefs others will form are constructed by turning theory-of-mind inward, anticipating the evaluations others will make about the actions one might take. With the knowledge of how one's actions will be judged before deciding, a decision maker can calibrate her actions to send the right signals (Baumeister, 1982; Bénabou \& Tirole, 2011). We note that we do not believe agents to be necessarily intentionally signaling impartiality to others. Instead agents may strive to maintain a desired image of themselves from an objective viewpoint or "self-signal" (Nagel, 1986; Bodner \& Prelec, 2003; Bénabou \& Tirole, 2011).

In this paper we develop a computational framework for capturing the above intuitions. We use influence diagrams as a structural representation of a rational actor and Bayesian inference over influence diagrams to enable theory-of-mind inferences about whether an action will be perceived as partial. While the framework we will present is a general way of constructing preferences from the anticipated judgments of others, we focus specifically on constructing distributional preferences with the desire to be perceived as impartial (Shaw, 2013; Shaw \& Olson, 2014; Dungan, Waytz, \& Young, 2014;

DeScioli, 2016). We first present a mathematical model that integrates preferences for efficient and equitable outcomes with an aversion to appear partial. We then test our model empirically in two parameterized allocation games with many conditions that allow us to test some of the fine-grained predictions of the model. Finally, we conclude by sketching how our model can be extended to capture other social desires constructed from a decision maker's preference to appear positively in the minds of others.

## Computational Analysis

In this work we aim to model both the way participants act in resource allocation games as well the judgments they make about the resource allocations of others. We start from the simpler preferences for efficiency and equity which are based on outcomes and build towards constructing a social preferences for impartiality which are implicitly intentional.

We define a resource allocation game as follows. Let $\mathcal{A}$ be the set of actions available to the decision maker. For each action $a \in \mathcal{A}$ there is a probabilistic transition function $P(R \mid a)$ which maps an action to a vector of rewards $R$ where each $r_{i} \in R$ is the amount of reward given to agent $i$. In a resource allocation game, the decision maker picks an action (a) such that the expected reward to the other agents $(R)$ achieves the desires of the decision maker.

We now define the desires of the Base Decision Maker as components of a utility function. These desires will determine how Base Decision Maker distributes resources. We consider two base desires. The first is a relative preference over the rewards received by specific agents. To realize this preference, we include the reward received by each of the other agents as weighted components of the decision maker's
own utility. Depending on the value of these weights, an agent might impartially value others or might be partial towards certain individuals. Formally, let $\alpha_{i} \in \boldsymbol{\alpha}$ be the weight that the decision maker places on the reward given to agent $i$. When $\alpha_{i}>0$, the decision maker gains utility proportional to the reward received by $i$, when $\alpha_{i}<0$ the decision maker loses utility proportional to the reward received by $i$ and when $\alpha=0$ the decision maker is indifferent to the reward received by $i$. By expressing different $\alpha$ over different agents the decision maker can express partiality (or aversion) towards specific agents. Including the rewards received by all others as positive elements $(\alpha>0)$ in the decision maker's own utility creates a preference for Pareto efficient allocations, a form of efficiency where the reward distributed cannot be increased by taking other actions without making one of the receiving agents worse off.

The second base desire implements a form of proportional equity, the idea that those who contribute more to a joint endeavor should reap a larger share of the rewards or "justdesserts". A well studied way to capture proportional equity quantitatively is to constrain the relative reward $\left(r_{i}\right)$ given to each agent to be proportional to their relative effort or merit ( $\gamma_{i}$ ) (Adams, 1965):

$$
\begin{equation*}
\frac{r_{1}}{\gamma_{1}}=\frac{r_{2}}{\gamma_{2}}=\ldots=\frac{r_{N}}{\gamma_{N}} \tag{1}
\end{equation*}
$$

We transform these constraints into a measurement of inequity:

$$
\begin{equation*}
I(R, \boldsymbol{\gamma})=\sum_{i \in N} \sum_{\substack{j \in N \\ j>i}}\left|\gamma_{j} r_{i}-\gamma_{i} r_{j}\right| \tag{2}
\end{equation*}
$$

With a notion of efficiency and equity in place, we can define the allocation preferences for the Base Decision Maker. The expected utility (EU) to the decision maker of choosing $a$ is:

$$
\begin{equation*}
\mathrm{EU}_{\mathrm{base}}[a]=-\alpha_{I A} \mathrm{E}_{a}[I(R, \boldsymbol{\gamma})]+\sum_{i \in N} \alpha_{i} \mathrm{E}_{a}\left[r_{i}\right] \tag{3}
\end{equation*}
$$

where $\mathrm{E}_{a}[I(R, \boldsymbol{\gamma})]$ is the expected amount of inequity created by action $a$ and $\alpha_{I A} \in \boldsymbol{\alpha}$ is the weight the decision maker places on inequity aversion. $\mathrm{E}_{a}\left[r_{i}\right]=\sum_{r_{i}} r_{i} P\left(r_{i} \mid a\right)$ is the expected reward for $i$ when the decision maker takes action $a$. Decision making follows probabilistically by sampling from the soft-max of expected utility:

$$
\begin{equation*}
P(a \mid \boldsymbol{\alpha}) \propto \exp (\beta * \mathrm{EU}[a]) \tag{4}
\end{equation*}
$$

with higher values of $\beta$ leading to a higher probability of selecting the action with the highest expected utility.

Influence diagrams are a natural choice for structurally representing this model since they can flexibly capture decision problems with multiple factors and recursive sources of value. Furthermore, they can be used to reason about the latent mental states of a decision maker from just a sparse and noisy observation of behavior (Jern \& Kemp, 2015; Kleiman-Weiner, Gerstenberg, Levine, \& Tenenbaum, 2015). The utility of the

Base Decision Maker which is defined in equation (3) can be expressed graphically as the influence diagram shown in Figure 1a. The first term of equation (3) corresponds to the $U_{I}$ node and the second term corresponds to the $U_{E}$ node.

We now consider a Judge who makes inferences and judgments about the underlying preferences of the Base Decision Maker following an observation of behavior. Specifically, in the Base Decision Maker the $\boldsymbol{\alpha}$ encode the preferences of the agent and so for the Judge these $\boldsymbol{\alpha}$ become the target of inference. For our purposes, the Judge is interested in the extent that the Base Decision Maker is partial to one or more agents. The Judge's prior is that the Base Decision Maker is partial (a binary variable) with probability 0.5 . If partial, one of the $\alpha_{i}=\alpha_{\text {partial }}$ ( $i$ chosen uniformly at random) and the other $\alpha_{-i}=-\alpha_{\text {partial }}$. Otherwise, if the agent is not partial, all $\alpha_{1 \ldots N}=1$. The Judge also has some prior uncertainty on the degree that the Base Decision Maker cares about inequity so $\alpha_{I A} \sim \operatorname{Exponential}(\boldsymbol{\lambda})$. With these priors over the types of preferences a Base Decision Maker might have, a Judge can use Bayesian inference to compute the extent that an agent was partial based on just a single observed allocation:

$$
\begin{equation*}
P(\text { partial }, \boldsymbol{\alpha} \mid a) \propto P(a \mid \boldsymbol{\alpha}) P(\boldsymbol{\alpha} \mid \text { partial }) P(\text { partial }) \tag{5}
\end{equation*}
$$

where $P(a \mid \boldsymbol{\alpha})$ is the model of action shown in equation (4) and the $\boldsymbol{\alpha}$ are then marginalized out to obtain a posterior on $P$ (partial $\mid a$ ). Figure 1 b shows how the judge does inference over the parameters of the influence diagram representing the Base Decision Maker.

A Constructed Social Preference inherits from and recursively builds upon both the Base Decision Maker and the Judge. In particular, the Constructed Social Preference has an additional preference to appear impartial. Since this is a preference over the beliefs others will form as a result of her decision, the preference to appear impartial is a preference over the posterior $P($ partial $\mid a)$. The Constructed Social Preference integrates these belief based preferences with the preferences for equity and efficiency of the Base Decision Maker:

$$
\begin{equation*}
\mathrm{EU}_{\text {constructed }}[a]=\mathrm{EU}_{\mathrm{base}}[a]-\alpha_{P A} P(\text { partial } \mid a) \tag{6}
\end{equation*}
$$

where $\alpha_{P A}$ is the extent that the Constructed Social Preference cares about whether other agents view her as impartial or not. This equation and the influence diagram in Figure 1c show how the Constructed Social Preference is built on top of the Judge and Base Decision Maker.

The Constructed Social Preference goes beyond preferences over outcomes like those in the Base Decision Maker. Instead, it anticipates the inferences other agents will make about its actions and optimizes its actions so that others have desirable beliefs. Figure 1d shows a simulated example where a decision maker had to choose between allocating either $\$ 1,000$ to one agent and $\$ 100$ to another equally meritorious agent or giving a smaller but equal value to both. The Constructed Social Preference is more likely to select the equal


Figure 2: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where both of the agents were equally meritorious. Trials with no gray bar indicate the model predicted near 0 . Error bars are the standard error of the mean.
option since it implies lower partiality even though both the Base Decision Maker and the Constructed Social Preference care equally about avoiding inequity.

In order to compare the model with human participants, we used maximum-likelihood estimation to optimize the free parameters to human judgments. The five parameters used for all simulations were: $\beta=0.003, \alpha_{\text {partial }}=6, \lambda=0.7$, $\alpha_{P A}=1350$. If agent $i$ was more meritorious than agent $j$ then $\frac{\gamma_{i}}{\gamma_{j}}=4$. Importantly, the parameters used to model the partiality data were constrained to be the same as those used to model participants' decisions.

## Experiments and Results

We test the predictions of this model in two parametric behavioral experiments that measure participants' decisions in a hypothetical resource allocation game as well as judgments about the partiality of another agent who made an allocation. Both experiments were run on Amazon Mechanical Turk. For each condition we compare the average responses with the predictions of the model.

## Experiment 1: Proportionality and Impartiality

In experiment 1 we investigate how equity and merit affect choices in an allocation game. We presented two groups of participants with the following vignette which describes an allocation game that took place in an everyday office setting:

[^120]bonus and Alex the $\$ 100$ bonus / Give them both a bonus of [\$0 / \$100 / \$500 / \$1000 / \$1,100])
Participant group 2: Max decides to [give Alex the \$1,000
bonus and Josh the $\$ 100$ bonus / give Josh the $\$ 1,000$ bonus and Alex the $\$ 100$ bonus / give them both a bonus of (\$0 / \$100 / \$500 / \$1000 / \$1,100)]. Who do you think Max likes better? (Definitely Alex $=-1$, Equal $=0$, Definitely Josh $=1$ )

The bold text shows the different variants of the vignettes. On different trials the value of the equal option varied between $\$ 0$ and $\$ 1,100$. On some trials both employees received equal work evaluations and on some trials one employee received a better work evaluation. The names of the employees changed on each trial but were always a high frequency male name.

We first report the results for when both employees were equally meritorious (Figure 2). We found high rates of inequity aversion that led to highly wasteful bonus allocations (Choices: $\mathrm{N}=89$; Judgments: $\mathrm{N}=104$ ). When the equal sized bonus was $\$ 0$, almost $50 \%$ of participants chose to allocate nothing, wasting a total of $\$ 1,100(\$ 1,000+\$ 100)$ rather than allocating unequal bonuses. When the bonus was $\$ 100$, over $75 \%$ of participants wasted the $\$ 1,000$ bonus in favor of two equal $\$ 100$ bonuses. These allocations were highly wasteful and were Pareto dominated since the unequal allocation would have made at least one of the employees better off without making the other employee worse off.

The partiality judgments made by a second set of participants is consistent with the idea that the aversion to creating unequal outcomes stems in part from a desire to appear impartial. We transformed judgments of liking into a partiality index by measuring absolute difference from 0 . Even when the alternative equal allocation required wasting the entire bonus, a person who allocated the large but unequal bonus was judged as highly partial (towards the person who received the higher bonus). Our computational model corroborates this interpretation and captures both participants' judgments of partiality and then uses those judgments to explain the strong aversion to an unequal outcome. The full model closely follows the pattern of decision making.

We now turn to the trials where one of the two employees received a better evaluation at work than the other and was thus more meritorious (Choices: $\mathrm{N}=89$; Judgments: N $=104)$. Figure 3 shows that this difference was sufficient to drive participant choices away from the wasteful equal bonus towards giving the large but unequal bonus to the employee who was more meritorious. This shift is consistent with equity (the more deserving employee got a greater share of the rewards). However, this also resulted in a novel type of wasteful decision making: the option to allocate $\$ 1,000$ or more to both employees was forgone over $70 \%$ of the time by the Pareto dominated unequal option that maintains equity based on merit.

Surprisingly, participants attributed the lowest partiality to employees who selected the equal bonus even though one of the receiving employees was more deserving than the other. This points to a possible difficulty in achieving equitable dis-


Figure 3: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where one of the agents was more meritorious than the other. Trials with no gray bar indicate the model predicted near 0 . A "fair bonus" was when the decision maker gave the large bonus to the agent with more merit. An "unfair bonus" was when the decision maker gave the large bonus to the agent with less merit. Error bars are the standard error of the mean.
tributions. Even when some agents might be more deserving than others, inferences of partiality are still readily made when observing an unequal distribution. Here equity and impartiality work against each other. Since the equal bonus led to a lower attribution of partially, as the size of the equal bonus grows, the model slowly shifts to the efficient equal bonus.

## Experiment 2: Procedural Fairness and Impartiality

In a second experiment we repeated the equal merit condition of experiment 1 but also included the possibility that the employee making the decision could flip a fair coin to decide who gets $\$ 1,000$ and who gets $\$ 100$ (Choices: $\mathrm{N}=54$; Judgments: $\mathrm{N}=158$ ). Besides the addition of this coin the vignette was identical to the vignette in experiment 1 . This is a key test of the impartiality hypothesis since when the size of the equal bonus is low, an inequitable but efficient allocation can be given without signaling partiality towards either of the employees by flipping a coin (Shaw \& Olson, 2014; Choshen-Hillel et al., 2015).

Consistent with the model predictions shown in Figure 4, participants did not judge employees who flipped the coin to be partial towards either of the employees. When the value of the equal bonus was low ( $\leq \$ 100$ ) participants no longer wasted resources like they did in experiment 1 . Instead they flipped the coin in order to allocate the full bonus without signaling partiality.

Combining the two experiments, we quantify the overall model performance across all of the conditions in the two experiments. Figure 5 shows the quantitative correlation of the model predictions with the average judgments of participants. Overall, participant judgments and decisions were highly correlated $\left(R^{2}=0.94\right)$ with the model predictions. This suggests that the model is capturing some of the fine grained structure of how people attribute both partiality and use it to make allocations of welfare.

Finally, we compare the full model presented here against a lesioned model that includes inequity aversion but does not reason about partiality and hence corresponds to the Base Decision Maker (i.e., $\alpha_{P A}=0$ ). The parameters in the lesioned model were directly fit to the choice data and were not constrained to fit the judgments. This model fit the data less well than the full model $\left(R^{2}=0.82\right)$. However, this lesioned model has less parameters than the full model. To test
for the possibility that the full model is overfitting the data we performed cross-validation using randomly chosen subsets of half the data to fit the free parameters and then tested against the held-out half. The held-out cross-validation correlation between the model and participants was $R^{2}=0.93$ which suggests that the full model is robust and is not overfitting. In contrast, the lesioned model performed much worse ( $R^{2}=0.74$ ) under cross-validation. When the full model was applied only to the choice data it captured nearly all of the variance ( $R^{2}=0.97$ ) and was still robust when evaluated on only held-out trials $\left(R^{2}=0.96\right)$.

## Discussion

We introduced a new computational model for constructing preferences by modeling rational agents which care about what others will infer about them from their actions. In this model, the machinery of theory-of-mind is turned inward to simulate how an action will likely be perceived or judged by others. Agents then use the perceptions and judgments they anticipate others will form to construct rich preferences over socially desirable traits such as impartiality. We tested key components of the model in two behavioral experiments that were designed to contain conflict between efficiency, equity and partiality and measured both participants' hypothetical resource allocations and the judgments they made about the partiality of others who had acted. The predictions of the model were closely correlated with both allocation decisions as well as partiality judgments. Finally, we note the best fit parameters had a high value for $\alpha_{P A}$ which suggests that partiality aversion was playing an important role in the model fit for predicting choices. A lesioned model that did not contain this parameter failed to predict participants' judgments in both experiments.

We now briefly describe qualitatively some of the other predictions this model can make without any structural extension. Our model predicts that when the decision maker and one of the agents have a previous relationship (such as old friends or a reciprocal relationship in a different context) there will be a greater probability of inferring partiality since this previous relationships will manifest itself on the prior over partial. With a greater probability of others inferring partiality a decision maker will be even less likely to give their friend a larger reward than another person. This reasoning might explain why nepotism and cronyism is judged as unfair


Figure 4: Empirical results and model predictions of (a) choices and (b) judgments of partiality for experiment 2 which introduced the option to flip a fair coin to decide the allocation of the unequal bonuses. Trials with no gray bar indicate the model predicted near 0 . Error bars are the standard error of the mean.


Figure 5: Quantification of model performance. Each point represents the model prediction and participant judgment for a single condition. For better fitting models the points will lie close to the $y=x$ diagonal. (left) The full model compared including both decision and judgment data. (middle) The full model compared only on the decision data. (right) Lesioned model that did not include partiality compared only on the decision data.
and avoided (Dungan et al., 2014). Other procedural tools such as the delegation of the decision to a third party may also be important to avoid the attribution of partiality. Under the model we have presented, if an attribution of partiality can be made less likely, the decision maker might be more likely to participate in nepotism and favoritism.

In future work we would like to investigate how other forms of social preferences can be constructed by placing preferences over anticipated judgments. For instance, people might desire to appear as trustworthy and generous or avoid appearing selfish or envious. Ultimately we suspect that an agent who carefully manipulates their image so that all others think she is a great person - will end up behaving quite similar to a person who is truly good. However, her behavior will be less robust - when she suspects her actions are unobserved or can only be interpreted ambiguously, the constructed social preferences disappears along with the altruistic or fair behavior (Dana, Weber, \& Kuang, 2007). By constructing social preferences such as impartiality, a key component of fairness, from the anticipated judgments of others, we quantitatively predict the fine-grained structure of both participants' decisions concerning the allocation of resources and participants' judgments about those who make distribution decisions. Our model makes clear that the power of theory-of-mind is not necessarily limited to understanding the beliefs and desires of other intentional agents. It can also be pointed inward to strategically shape beliefs and desires in others.

Acknowledgement This work was supported by a Hertz Foundation Fellowship, NSF-GRFP, the Center for Brains, Minds and Machines (CBMM), NSF STC
award CCF-1231216 and by an ONR grant N00014-13-1-0333.

## References

Adams, J. S. (1965). Inequity in social exchange. Advances in experimental social psychology, 2(267-299).
Baumeister, R. F. (1982). A self-presentational view of social phenomena. Psychological bulletin, 91(1), 3.
Bénabou, R., \& Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. The Quarterly Journal of Economics, 126(2), 805-855.
Blake, P. R., \& McAuliffe, K. (2011). "I had so much it didnt seem fair": Eight-year-olds reject two forms of inequity. Cognition, 120(2), 215-224.
Bodner, R., \& Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. The psychology of economic decisions, 1, 105-26.
Choshen-Hillel, S., Shaw, A., \& Caruso, E. M. (2015). Waste management: How reducing partiality can promote efficient resource allocation. Journal of personality and social psychology, 109(2), 210.

Choshen-Hillel, S., \& Yaniv, I. (2011). Agency and the construction of social preference: Between inequality aversion and prosocial behavior. Journal of personality and social psychology, 101(6), 1253.

Dana, J., Weber, R. A., \& Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Economic Theory, 33(1), 67-80.
Dennett, D. C. (1989). The intentional stance. MIT press.
DeScioli, P. (2016). The side-taking hypothesis for moral judgment. Current Opinion in Psychology, 7, 23-27.
Dungan, J., Waytz, A., \& Young, L. (2014). Corruption in the context of moral trade-offs. Journal of Interdisciplinary Economics, 26(1-2), 97-118.
Fehr, E., \& Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. The quarterly journal of economics, 114(3), 817-868.
Jern, A., \& Kemp, C. (2015). A decision network account of reasoning about other people's choices. Cognition, 142, 12-38.
Kleiman-Weiner, M., Gerstenberg, T., Levine, S., \& Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In Proceedings of the 37th annual conference of the cognitive science society.
Nagel, T. (1986). The view from nowhere. Oxford University Press.
Rawls, J. (1971). A theory of justice. Harvard university press.
Shaw, A. (2013). Beyond "to share or not to share" the impartiality account of fairness. Current Directions in Psychological Science, 22(5), 413-417.
Shaw, A., \& Olson, K. (2014). Fairness as partiality aversion: The development of procedural justice. Journal of Experimental Child Psychology, 119, 40-53.
Shaw, A., \& Olson, K. R. (2012). Children discard a resource to avoid inequity. Journal of Experimental Psychology: General, 141(2), 382.
Takagishi, H., Kameshima, S., Schug, J., Koizumi, M., \& Yamagishi, T. (2010). Theory of mind enhances preference for fairness. Journal of experimental child psychology, 105(1), 130-137.
Tomasello, M. (2014). A natural history of human thinking. Harvard University Press.

# Explaining Guides Learners Towards Perfect Patterns, Not Perfect Prediction 

Elizabeth Kon (ellie.kon@berkeley.edu), Tania Lombrozo (lombrozo@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, 3210 Tolman Hall,<br>Berkeley, CA 94720 USA


#### Abstract

When learners explain to themselves as they encounter new information, they recruit a suite of processes that influence subsequent learning. One consequence is that learners are more likely to discover exceptionless rules that underlie what they are trying to explain. Here we investigate what it is about exceptionless rules that satisfies the demands of explanation. Are exceptions unwelcome because they lower predictive accuracy, or because they challenge some other explanatory ideal, such as simplicity and breadth? To compare these alternatives, we introduce a causally rich property explanation task in which exceptions to a general rule are either arbitrary or predictable (i.e., exceptions share a common feature that supports a "rule plus exception" structure). If predictive accuracy is sufficient to satisfy the demands of explanation, the introduction of a rule plus exception that supports perfect prediction should block the discovery of a more subtle but exceptionless rule. Across two experiments, we find that effects of explanation go beyond attaining perfect prediction.


Keywords: explanation; learning; causal reasoning

## Introduction

"The great tragedy of science - the slaying of a beautiful hypothesis by an ugly fact." T. H. Huxley (1870)

The best explanations account for all the data we invoke them to explain. But in science and in life, explanations often have exceptions. Even when exceptions fail to "slay" our explanatory hypotheses, they certainly diminish them. What is it about exceptions that threatens the quality of explanations?

One possibility is that exceptions are threatening because they offer evidence against the truth of the explanation in question. To the extent our explanation fails to predict an anomalous observation, we might hold out for a better alternative - one that predicts the observation with greater probability, such that the observation provides greater evidential support for that alternative explanation.

A second possibility is that exceptions diminish the quality of explanations not because they reveal predictive failures, but because they reveal that an explanation is deficient with respect to some other explanatory ideal. Across philosophy and science, we praise explanations for their simplicity, breadth, generality, and ability to unify a diverse range of phenomena. Exceptions may diminish the quality of explanations because they threaten these ideals.

In the current experiments, we test these alternatives by investigating how the process of explaining affects learning (for reviews, see Fonseca \& Chi, 2011; Lombrozo, 2012). Prior work has found that when learners are prompted to explain, they're more likely to discover regularities that support "good" explanations (Lombrozo, 2016). In particular, Williams and Lombrozo (2010) found that when learning to classify robots from novel categories, those participants who
were prompted to explain why each exemplar might belong to its respective category were significantly more likely to discover a subtle classification rule that accounted for all eight items (the $100 \%$ rule), as opposed to settling for a more salient classification rule that only accounted for six ("the $75 \%$ rule"), leaving two exceptions.

The results of Williams and Lombrozo (2010) support the idea that explaining encourages learners to find an exceptionless pattern, but do not reveal what it is about exceptions that makes the $75 \%$ rule less good than the $100 \%$ rule. If explaining drives learners away from exceptions because they decrease predictive accuracy, then a rule with non-arbitrary exception - that is, with exceptions that can be reliably identified a priori, such that predictive accuracy can reach $100 \%$ - should rival an exceptionless rule. In contrast, if exceptions are undesirable because they threaten some other explanatory virtue, such as simplicity or breadth, then even a rule with non-arbitrary exceptions should be dominated by a $100 \%$ rule that classifies all items in a unified way.

To test these predictions, we had participants learn novel relationships while prompted to explain or write down their thoughts, and where the exceptions to the $75 \%$ rule were either arbitrary (as in prior work) or meaningful (in the sense that they supported perfect prediction by representing a "rule plus exception" on the basis of two features). If prompting learners to explain pushes them to find a simple, exceptionless pattern, then the two conditions should yield similar results, whether or not the exceptions are meaningful. On the other hand, if explainers are satisfied by a rule that supports perfect prediction, then discovery of the relatively salient $75 \%$ rule with meaningful exceptions should block discovery of the more subtle $100 \%$ rule. We test these competing predictions in Experiment 1 using a sequential training procedure, and in Experiment 2 using a prediction task.

Our task and stimuli go beyond prior work in a second way, as well. Instead of using a classification task in which participants explain category membership by appeal to arbitrary features, we use a causally-rich property explanation task. Prior work suggests a preference for exceptionless, single-feature rules in classification (e.g., Norenzayan et al., 2002; but see Murphy, Bosch, \& Kim, 2016); explanation could simply heighten this classification-based preference. In the current studies, rather than explaining category membership, participants explain why novel creatures eat flies or eat crabs, where both the $75 \%$ and $100 \%$ rules reflect plausible causal explanations. If prompting learners to explain still promotes discovery of a $100 \%$ rule with these modified stimuli, it would suggest that previously-documented effects of explanation on learning are not restricted to classification tasks (see

| Arbitrary <br> Exception | Meaningful <br> Exception | Both sets | Both sets | Both sets |
| :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |

Figure 1: All Stimuli. The top row creatures all eat flies, and the bottom row eat crabs. For the arbitrary exception exceptiontype, participants saw the creatures in the first column, and for the meaningful exception exception-type, they saw the creatures in the second column. Both stimulus sets included creatures in columns three through five.
also Walker et al., 2017). This finding would also help bridge the gap between laboratory studies involving artificial materials and educational materials such as biology texts, where effects of explanation have been documented and inform curricula (Fonseca \& Chi, 2011).

## Experiment 1

Experiment 1 investigates whether engaging in explanation encourages learners to seek simple, exceptionless rules, or to instead find rules that allow for perfect predictive accuracy. To test this, we created two stimulus sets: one with a "meaningful exception rule" (a $75 \%$ rule with exceptions identifiable by the presence of a second feature), and another with an "arbitrary exception rule" (a $75 \%$ rule with exceptions that do not share a common feature). The meaningful exception rule was relatively easy to discover and supported perfect prediction, but not on the basis of a single feature. If prompting learners to explain makes them persist in seeking an exceptionless single-feature rule, we would predict comparable results for the meaningful exception stimuli and the arbitrary exception stimuli, with learners prompted to explain significantly more likely than those in a control condition to discover the more subtle $100 \%$ rule. On the other hand, if perfect predictive accuracy satisfies the demands of explanation, we would expect discovery of the more salient meaningful exception rule to block discovery of the $100 \%$ rule, yielding an attenuated effect of explanation on $100 \%$ rule discovery, and a boost in discovery of the meaningful exception rule.

## Method

Participants Participants were 443 adults recruited from the Amazon Mechanical Turk marketplace. Of these, 124 failed attention or memory checks (described below) or left questions blank and were therefore excluded from analyses. The statistical significance of results are unchanged when these participants are included.

Materials The stimuli consisted of two sets of eight "creatures" each, four of which ate flies and four of which ate crabs (see Figure 1). For each set, participants could use two possible rules to determine whether a creature ate flies or crabs. The first accounted perfectly for all eight creatures (the " $100 \%$ rule"): all four creatures that ate flies had snouts pointing up; all four creatures that ate crabs had snouts pointing down. The second rule accounted for six of the eight creatures (the " $75 \%$ rule"): three of four creatures that ate flies were on land; three of four creatures that ate crabs were underwater. Importantly, both features of interest (snout direction and habitat) supported plausible causal explanations for why a creature eats flies versus crabs, e.g., "It eats flies because its snout is pointed up, so it can reach flies" or "It eats flies because it lives on land, where flies are found."

The two stimulus sets differed in the nature of the exceptions to the $75 \%$ rule. For participants in the arbitrary exceptions condition, the exceptions to the $75 \%$ rule did not share a meaningful, plausible characteristic on the basis of which they could be identified as exceptions. For participants in the meaningful exceptions condition, the exceptions to the $75 \%$ rule were "newborns"- they were green and shown with eggs in a nest. We refer to this manipulation as "exception-type."
Procedure The task consisted of a study phase followed by a reporting phase and a rule rating phase.

At the start of the study phase, participants were randomly assigned to one of four conditions, which were created by crossing two prompt-types, Explain or Write Thoughts, with two exception-types, arbitrary or meaningful.

In the study phase, all participants were told to study the creatures, and that after the study phase they would be asked questions about how to determine which food a creature eats. To provide context and help participants interpret the images, they were told that the creatures were: "from the planet ZARN: the adults of all of these creatures eat either flies or crabs. Newborn creatures look exactly like their adult forms except that they are green because they photosynthe-
size. There are different subspecies of this animal with different properties. However, they all have a mouth on an inflexible snout, and an ear that sticks up. They are all tailless, born from eggs and have a 4-chambered heart." Participants were presented with a randomized array of the eight creatures corresponding to their condition's exception-type (arbitrary or meaningful). They were then prompted to focus their attention on each creature, individually, in a random order, with a prompt determined by the experimental condition to which they were randomly assigned. Participants in the explain conditions were told to "try to explain why creature X eats flies/crabs." Participants in the write thoughts conditions were told to "Write out your thoughts as you learn that creature X eats flies/crabs." Participants were given 50 seconds to respond to each prompt, at which time their responses were recorded and the prompt for the next item appeared.

In the reporting phase, participants were told that "we're interested in any patterns that you noticed that might help differentiate creatures that eat flies and creatures that eat crabs. For example, did most or all of the fly-eaters you studied tend to have one property, and most or all of the crab-eaters you studied have another property? We're going to ask you to list all of the patterns (differences between fly-eaters and crabeaters) that you noticed, one at a time. PLEASE REPORT ANY PATTERNS THAT YOU NOTICED, EVEN IF THEY WEREN'T PERFECT AND EVEN IF YOU DON'T THINK THEY'RE IMPORTANT." This language, adapted from Edwards, Williams, and Lombrozo (2013), was employed to encourage participants to report the $75 \%$ rule even if they thought it was incidental or superseded by the $100 \%$ rule. In addition to describing the rule they discovered in a freeresponse box, participants were asked how many of the eight items followed the rule.

After finishing the reporting phase, participants were again presented with all eight creatures as well as four candidate explanations (presented in a random order) for "why creatures A-D eat flies (as opposed to crabs)." They were forced to stay on the page for at least 15 seconds to ensure that they read the explanations (there was no upper time limit). Along with an inaccurate explanation included as a control, the explanations provided for rating were:

- $100 \%$ rule: "Because creatures A-D have snouts that point up, and creatures E-H have snouts that point down."
- $75 \%$ rule: "Because creatures A-D live on land, and creatures E-H live in the water."
- $75 \%$ rule + exception associated with their exception-type:
- with arbitrary exceptions: "Because creatures A-D live on land, and creatures E-H live in the water (with some exceptions)."
- with meaningful exceptions: "Because creatures A-D live on land, and creatures E-H live in the water (with the exception of newly-hatched creatures, who are born in the opposite environment)."

Ratings were collected on a 7-point scale with anchors at 1 ("Very Poor Explanation") and 7 ("Excellent Explanation").

Before concluding the experiment, participants completed an attention and memory check question that served as the basis for participant exclusion. They were asked to "look at the following images and select the one that you have studied in previous questions. In the text box next to that image, please also type in whether you think that it eats flies or crabs. It is important for us to know whether our participants are paying attention and are reading all of the instructions, so if you are reading this, what we actually want you to do is to select "None of these objects look familiar," and in the corresponding text box to write in whether the image you recognize from the other options eats flies or crabs." By selecting the instructed button, participants indicated they had been reading instructions, and by correctly reporting the diet of the creature they recognized, participants indicated that they attended to the stimuli in the primary task.

## Results

Overall, participants reported finding an average of 1.25 patterns $(S D=0.96, \min =0, \max =4)$ that they reported accounted for an average of 5.94 exemplars $(S D=1.8, \mathrm{~min}=0$, $\max =8$ ). Reported patterns were coded for mention of the $100 \%$ rule and/or the $75 \%$ rule.
$\mathbf{1 0 0 \%}$ rule reporting: To test whether explanation prompts affected $100 \%$ rule discovery, and whether effects differed across exception-type, we conducted a logistic regression predicting whether participants discovered $100 \%$ rule (yes vs. no) by prompt-type (explain vs. write thoughts) $\times$ exception-type (arbitrary vs meaningful). This revealed a significant effect of prompt-type on reporting the $100 \%$ rule, collapsed over exception-types $\left(\chi^{2}=6.64, p=0.01\right.$; see Figure 2). The interaction term between prompt-type and exception-type was not significant $\left(\chi^{2}=0.28, p=0.6\right)$.


Figure 2: Proportion of Participants Reporting the $100 \%$ Rule in Experiment 1

The results of this analysis are consistent with the hypothesis that what people seek when explaining are rules high in explanatory virtues such as simplicity and breadth: the opportunity to employ a rule + meaningful exception (which was both easy to discover and afforded perfect prediction) did not block participants in the explain condition from seeking an alternative that accounted for all items with a single feature. However, this conclusion should be accepted with some caution: when analyzed alone, there was not a significant effect of prompt-type within the meaningful exceptions conditions ( $\chi^{2}=1.93, p=0.16$ ), but there was in the arbitrary exceptions condition ( $\chi^{2}=5, p=0.03$ ).
$\mathbf{7 5 \%}$ rule reporting: Previous studies have found that prompting participants to explain can decrease $75 \%$ rule reporting relative to a control condition (e.g., Edwards et al., 2013; Williams \& Lombrozo, 2010, 2013). In this study, the proportions of participants reporting the $75 \%$ rule were: $51 \%$ for explain/arbitrary; $46 \%$ for write thoughts/arbitrary; $69 \%$ for explain/meaningful; and $63 \%$ for write thoughts/meaningful.

To analyze these data we ran another logistic regression: discovered $75 \%$ rule (yes vs. no) by prompt-type (explain vs. write thoughts) $\times$ exception-type (arbitrary vs. meaningful). The effect of prompt-type was not significant ( $\chi^{2}=1.15, p$ $=0.28$ ). The effect of exception-type was significant $\left(\chi^{2}=\right.$ $10.08, p<0.01)$. However, the interaction between prompttype and exception-type was not significant $\left(\chi^{2}=0.02, p=\right.$ $0.9)$. So while people were more likely to report the $75 \%$ rule when the exceptions were meaningful, this effect was not moderated by prompt-type. Few participants reported both the $100 \%$ and $75 \%$ rules: $16 \%$ for explain/arbitrary; $9 \%$ for write thoughts/arbitrary; $23 \%$ for explain/meaningful; and $17 \%$ for write thoughts/meaningful.

Rule Rating: To confirm that the manipulation of exception-type had some effect on perceived explanation quality, we compared explanation ratings for the $75 \%$ rule + exception as a function of exception type. Indeed, a t-test revealed higher ratings when the exception was meaningful $t(309)=-4.3, p<0.01$ (see Table 1 for all mean ratings).

Table 1: Average Rule Rating by Exception-type

| Condition | $100 \%$ rule | $75 \%$ rule | $75 \%$ rule + <br> exception | Bad Rule |
| :--- | :---: | :---: | :---: | :---: |
| Arbitrary <br> Exceptions | $5.50(2.23)$ | $3.11(1.95)$ | $4.87(2.05)$ | $1.81(1.54)$ |
| Meaningful <br> Exceptions | $5.19(2.4)$ | $3.97(1.92)$ | $5.80(1.79)$ | $1.55(1.16)$ |

## Discussion

On balance, the results from Experiment 1 support the idea that when it comes to the effects of explanation on learning,
an explanation that supports perfect prediction can still be deficient if it fails to account for all observations in a unified way. The experiment also suggests that the original effects reported in Williams and Lombrozo (2010) are not restricted to explicit classification tasks with arbitrary features: we successfully reproduced effects of explanation in a property explanation task where explanations were causally meaningful.

Introducing a rule with meaningful exceptions did have significant effects: participants were more likely to report discovering the $75 \%$ rule when the exceptions were meaningful (regardless of prompt), and they evaluated the explanation containing a $75 \%$ rule to be more satisfactory when the exceptions were meaningful. However, introducing the 75\% rule with meaningful exceptions did not block participants prompted to explain from discovering the $100 \%$ rule: they seemed to persevere in looking for an exceptionless, singlefeature rule rather than settling for a rule that supported perfect prediction on the basis of multiple features. This conclusion is supported by the significant effect of prompt-type on $100 \%$ rule discovery, which was not qualified by a further interaction with exception-type. At the same time, we note that when restricting analysis to the meaningful exceptions condition, the effect of explanation was not significant. The results of Experiment 1 are therefore somewhat inconclusive, and we revisit the contrast between arbitrary and meaningful exceptions in Experiment 2.

## Experiment 2

Because the results from Experiment 1 were somewhat inconclusive, we ran a new variant of the task. The task used in Experiment 2 was designed to heighten the value of perfect prediction: rather than receiving labelled exemplars at each step, participants attempt to predict the food that each creature eats, receiving feedback as they proceeded. If explanatory judgments track perfect prediction, then participants prompted to explain in this task should be satisfied with a $75 \%$ rule when it involves meaningful exceptions, thereby supporting perfect prediction and blocking or attenuating the effect of explanation on $100 \%$ rule discovery.

## Method

Participants For this study, 164 adults were recruited from the Amazon Mechanical Turk marketplace. Of these, 61 failed the attention and memory checks described above. We note any cases in which relaxing these exclusion criteria affected conclusions regarding statistical significance.
Materials Stimuli were the same as in Experiment 1.
Procedure This task consisted of a study phase and a reporting phase. As in Experiment 1, participants were randomly assigned to one of four conditions, which were created by crossing two prompt-types, Explain or Write Thoughts, with two exception-types, arbitrary or meaningful.

In the study phase, participants were presented with the same introductory text as in Experiment 1. They were then
given 5 seconds to look over all eight creatures together before being shown the creatures individually in a random order.

When presented with each of the eight creatures individually, participants were asked to determine whether the creature eats crabs or flies. Based on the accuracy of their response, they were then taken to a screen that said either "CORRECT This item does eat flies/crabs" or "INCORRECT This item eats flies/crabs." They were then given 45 seconds to respond to their condition-specific prompt; either "This creature eats flies/crabs. Try to explain why this creature eats flies/crabs." or "This creature eats flies/crabs. Write down whatever you are thinking." After cycling through all eight creatures, participants went through them a second time, again in a random order, with 30 seconds to respond.

The reporting phase was identical to that of Experiment 1.

## Results

Overall, participants reported finding an average of 0.95 patterns $(S D=0.96, \min =0, \max =6)$ which they reported accounted for an average of 6.35 exemplars ( $S D=1.53$, min $=0$, $\max =8$ ). Reported patterns were coded for mention of the $100 \%$ rule and/or the $75 \%$ rule.
$\mathbf{1 0 0 \%}$ rule reporting: To analyze $100 \%$ rule discovery (see Figure 3), we ran a logistic regression of discovered $100 \%$ rule (yes vs. no) by prompt-type (explain vs. write thoughts) $\times$ exception-type (with arbitrary exceptions vs. with meaningful exceptions). The interaction between prompt-type and exception-type was not significant ( $\chi^{2}=$ $0.23, p=0.63$ ).

However, there was a significant effect of explanation (collapsed across the two stimulus sets) $\left(\chi^{2}=4.15, p=0.04\right)^{1}$. These findings suggest that the presence of a salient rule that supported perfect prediction in the meaningful exceptions condition was insufficient to block discovery of the $100 \%$ rule, and therefore support the proposal that explainers preferentially seek simple, exceptionless patterns, not merely perfect predictability.

Again, to see whether the effect of explanation held within the meaningful exceptions condition, we ran a logistic regression predicting discovered $100 \%$ rule (yes vs. no) by prompttype (explain vs. write thoughts) using only the results from the meaningful exceptions condition. We found that there was again no significant effect of explanation when restricting analysis in this way, $\left(\chi^{2}=2.94, p=0.09\right)$.
$\mathbf{7 5 \%}$ rule reporting: The proportions of participants reporting the $75 \%$ rule were: $36 \%$ for explain/arbitrary; $32 \%$ for write thoughts/arbitrary; $61 \%$ for explain/meaningful; and $59 \%$ for write thoughts/meaningful.

To analyze these data we ran a logistic regression predicting discovered $75 \%$ rule (yes vs. no) by prompt-type (explain vs. write thoughts) $\times$ exception-type (arbitrary vs. meaningful). Again, the effect of prompt-type was not significant $\left(\chi^{2}=0.06, p=0.8\right)$, the effect of exception-type was

[^121]

Figure 3: Proportion of Participants Reporting the $100 \%$ rule in Experiment 2
significant: $\left(\chi^{2}=7.11, p=0.01\right)$, and the interaction between prompt-type and exception-type was not significant $\left(\chi^{2}=0.02, p=0.9\right)$. Few participants reported both the $100 \%$ and $75 \%$ rules: $4 \%$ for explain/arbitrary; $0 \%$ for write thoughts/arbitrary; $4 \%$ for explain/meaningful; and $4 \%$ for write thoughts/meaningful.

Prediction Performance: As a check to ensure that the $75 \%$ rule with a meaningful exception indeed improved predictability, we additionally analyzed prediction performance in the second block of the task. Specifically, we compared the proportion of exception items that were correctly classified (of 2) as a function of exception-type (arbitrary vs. meaningful) for the 45 participants who reported discovering the $75 \%$ rule, but not the $100 \%$ rule. A t-test revealed that prediction accuracy was indeed higher when exceptions were meaningful ( $M=1.39, S D=0.83$ ) than when they were not ( $M=0.53$, $S D=0.72), t(38)=-3.68, p<0.01$.

## Discussion

The results of Experiment 2 support the proposal that explainers strive for simple, exceptionless patterns rather than settling for perfect predictability. Even though the presence of meaningful exceptions did improve performance on the prediction task, it did not decrease discovery of the $100 \%$ rule differently for participants who explained and for participants who wrote their thoughts.

## Discussion

Across two experiments, we find support for the proposal that when explaining, people prefer rules that are high in explanatory virtues (such as simplicity and breadth) over alternative rules that allow for perfect prediction, but that are deficient in these virtues. The threat posed by exceptions therefore appears to be rooted in their disruption of explanatory ideals and not only predictive accuracy. This result is consistent with the
observation from science and philosophy that the most predictive models are often not the most explanatory. Additionally, by using a causally-rich property explanation task rather than an arbitrary categorization task, we find support for the claim that effects of explanation on the discovery of exceptionless patterns are not restricted to classification contexts.

Despite these promising results, many questions remain open. First, we found a weaker effect of explanation on $100 \%$ rule discovery in the meaningful exceptions conditions than in the arbitrary exceptions condition. This suggests that the presence of a $75 \%$ rule that afforded perfect prediction attenuated $100 \%$ rule discovery. However, the three-way interaction between $100 \%$ rule discovery, prompt type, and exception type did not reach significance, even when pooling results across studies. It thus remains a possibility that introducing meaningful exceptions has a small but real effect on $100 \%$ rule discovery; this is worth revisiting with a larger sample and more varied stimuli and learning tasks. Second, our results speak to the consequences of engaging in explanation, but not to the mechanisms by which explaining generates these consequences. The possibility we have advanced is that by virtue of explaining, participants are more likely to reject working hypotheses as they encounter exceptions, and therefore persevere in looking for a pattern that supports a good explanation, where a "good" explanation goes beyond predictive accuracy. Given that participants approach these problems with a host of prior beliefs, future studies should investigate this process more directly, including how learners go about generating hypothesis, seeking information, and updating their beliefs in light of new information.

The fact that explaining can be beneficial in learning is influencing educational systems from online learning environments (e.g. Williams et al., 2014) to college chemistry courses (Teichert \& Stacy, 2002). However, as demonstrated here, explanation privileges rules that are simple and exceptionless, and not all learning contexts involve this kind of structure. In fact, previous work has found that prompting learners to explain is sometimes detrimental (e.g. Berthold et. al., 2011; Kuhn \& Katz, 2009; Rittle-Johnson \& Loehr, 2016; Williams \& Lombrozo, 2013; see also Nokes et al., 2011). This underscores the importance of understanding when and why engaging in explanation will and will not promote particular learning outcomes; our current findings provide an additional step towards achieving this understanding.

## Acknowledgements

We thank the CoCo Lab, particularly Daniel Wilkenfeld. This work was supported by an NSF Career Grant awarded to T.L. (DRL-1056712), as well as a McDonnell Foundation Scholar Award in Understanding Human Cognition.

## References

Berthold, K., Röder, H., Knörzer, D., Kessler, W., \& Renkl, A. (2011). The double-edged effects of explanation prompts. Computers in Human Behavior, 27(1), 6975.

Edwards, B. J., Williams, J. J., \& Lombrozo, T. (2013). Effects of explanation and comparison on category learning. In Proceedings of the 35th annual conference of the cognitive science society (pp. 406-411).
Fonseca, B., \& Chi, M. T. H. (2011). Instruction based on self-expanation. In R. Mayer \& P. Alexander (Eds.), Handbook of research on learning and instruction. Routledge.
Huxley, T. H. (1870). Biogenesis and abiogenesis. Collected Essays of Thomas H. Huxley, 8, 256.
Kuhn, D., \& Katz, J. (2009). Are self-explanations always beneficial? Journal of Experimental Child Psychology, 103(3), 386-394.
Lombrozo, T. (2012). Explanation and Abductive Inference. In K. J. Holyoak \& R. G. Morrison (Eds.), The oxford handbook of thinking and reasoning. Oxford, UK: Oxford University Press.
Lombrozo, T. (2016). Explanatory Preferences Shape Learning and Inference. Trends in Cognitive Sciences, 20(10), 748-759.
Murphy, G. L., Bosch, D. A., \& Kim, S. (2016). Do Americans Have a Preference for Rule-Based Classification? Cognitive Science.
Nokes, T. J., Hausmann, R. G. M., VanLehn, K., \& Gershman, S. (2011). Testing the instructional fit hypothesis: The case of self-explanation prompts. Instructional Science, 39(5), 645-666.
Norenzayan, A., Smith, E. E., Kim, B. J., \& Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. Cognitive Science, 26(5), 653-684.
Rittle-Johnson, B., \& Loehr, A. M. (2016). Eliciting explanations: Constraints on when self-explanation aids learning. Psychonomic Bulletin \& Review, 1-10.
Teichert, M. A., \& Stacy, A. M. (2002). Promoting understanding of chemical bonding and spontaneity through student explanation and integration of ideas. Journal of Research in Science Teaching, 39(6), 464-496.
Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., \& Gopnik, A. (2017). Explaining Constrains Causal Learning in Childhood. Child Development, 88(1), 229-246.
Williams, J. J., Kovacs, G., Walker, C. M., Maldonado, S., \& Lombrozo, T. (2014). Learning online via prompts to explain. In Chi'14 extended abstracts on human factors in computing systems (pp. 2269-2274).
Williams, J. J., \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. Cognitive Science, 34(5), 776-806.
Williams, J. J., \& Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. Cognitive Psychology, 66, 55-84.

# Selective Information Sampling and the In-Group Heterogeneity Effect 

Elizaveta Konovalova (elizaveta.konovalova@upf.edu)

Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

# Gael Le Mens (gael.le-mens@upf.edu) 

Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain


#### Abstract

People often perceive their in-groups as more heterogeneous than their out-groups. We propose an information sampling explanation for this in-group heterogeneity effect. We analyze a model in which an agent forms beliefs and attitudes about social groups from her experience. Consistent with robust evidence from the social sciences, we assume that people are more likely to interact again with in-group members than with outgroup members. This implies that people obtain larger samples of information about in-groups than about out-groups. Because estimators of variability tend to be right-skewed, but less so when sample size is large, sampled in-group variability will tend to be higher than sampled out-group variability. This implies that even agents that process information correctly - even if they are naive intuitive statisticians - will be subject to the in-group heterogeneity effect. Our sampling mechanism complements existing explanations that rely on how information about in-group and out-group members is processed.


Keywords: Information Sampling, Judgment Bias, Perception of Variability.

## Introduction

A large amount of research has shown that people frequently perceive their groups as more heterogeneous than groups to which they do not belong (Boldry, Gaertner, \& Quinn, 2007; Rubin \& Badea, 2012; Ostrom \& Sedikides, 1992). For example, Park and Judd (1990) found that students majoring in one subject judged students with other major as less variable on such characteristics as extroversion, impulsiveness, and how analytical and reserved they are. This "in-group heterogeneity effect" has received several classes of explanations. One class of existing explanations rely on differences in how information about in-groups and out-groups is processed (Ostrom \& Sedikides, 1992; Ostrom, Carpenter, Sedikides, \& Li, 1993; Park \& Rothbart, 1982) or encoded (Linville, Fischer, \& Salovey, 1989; Linville \& Fischer, 1998; Judd \& Park, 1988; Park \& Judd, 1990). Another explanation takes as a premise that heterogeneity is seen as a positive feature of social groups and that people want to have a positive view of their in-groups (this is the much studied "in-group outgroup bias", see Hewstone, Rubin, and Willis (2002)). People would thus be motivated to see in-groups as more heterogeneous than out-groups (Ostrom \& Sedikides, 1992; Rubin \& Badea, 2012).

Here, we propose a distinct, sampling-based, explanation for the in-group heterogeneity effect. We note that people tend to obtain larger samples of information about in-groups than about out-groups. For example, people can avoid interacting again with an out-group if they had a bad experience
with members of this group. By contrast, people have to keep interacting frequently with members of the in-group even if they had negative experiences with those. Avoidance of the in-group is thus less likely (there is a large literature on this differential 'adaptive sampling' behavior, see Denrell, 2005; Fazio, Eiser, \& Shook, 2004; Fiedler \& Juslin, 2006).

The second premise of our explanation is that the variability of samples of information tends to increase with the size of the sample. Consider for example the variance of a sample of $k$ independent standard normal variables (with mean $\mu=0$ and variance $\sigma=1$ both unknown). This is a random variable that can be written $\hat{\sigma}_{k}=Q /(k-1)$ where $Q$ is a distributed according to a chi-squared distribution with $k-1$ degrees of freedom $\chi_{k-1}^{2}$. The mean of $Q$ is $k-1$. Two features of chisquared distribution are noteworthy: $Q$ is right-skewed (the probability that the sample variance is lower than the mean is higher than $50 \%$ ) and the skewness is decreasing in $k$ (the skewness is equal to $\sqrt{8 /(k-1)}$ ). Overall this means that the sample variance tends to underestimate the true variance ( $\sigma=1$ ) and that the extent of the underestimation decreases as the sample size increases.

These two premises jointly imply that the experienced variability of in-groups will tend to be higher than the experienced variability of out-groups. Under the assumption that people's subjective perception of group variability is closely related to the variance of the sample of information they obtained about this group (Kareev, Arnon, \& Horwitz-Zeliger, 2002; Weber, Shafir, \& Blais, 2004; see Boldry et al., 2007 for a review), this implies that people will tend to perceive in-groups as more variable than out-groups.

This explanation for the in-group heterogeneity effect operates at a different level than existing explanations. Whereas existing explanations focus on how the mind processes information, our explanation focuses on the properties of the samples of information to which the mind has access. We emphasize how the structure of the environment can lead to systematic sampling asymmetries, which in turn imply systematic judgment asymmetries. As such, our explanation fits in the 'sampling approach' to human judgment (Denrell, 2005; Fazio et al., 2004; Fiedler \& Juslin, 2006; Le Mens \& Denrell, 2011).

## Model

Consider a setting where one agent forms attitudes and beliefs about two groups ( $g=i n$, out ). The agent belongs to one
of the two groups - the in-group. In this simple model, we assume that the agent observes two dimensions of the groups: an attitudinal dimension, $A$ and another dimension $X$. Here we assume that the dimension $X$ is not stereotypical in the sense that it does not serve as the basis for categorization. In each period, the agent samples the group or not. When the agent samples a group she observes both dimensions $A$ and $X$ of one of its members.

Belief Updating Let $A_{t, g}$ denote attitude of the agent toward group $g$ at the end of period $t$. If she samples the group in period $t$, two things happen.

- She updates are attitude toward the group. Her new attitude is a weighted average of her previous attitude and the new observation $a_{t, g}$ :

$$
\begin{equation*}
A_{t, g}=(1-b) A_{t-1, g}+b a_{t, g}, \tag{1}
\end{equation*}
$$

where $b \in[0,1]$. We assume that $a_{t, g}$ is normally distributed (with mean 0 and variance 1 ). This attitude updating rule has been found to provide good fit to experimental data on sequential choice under uncertainty (see Denrell, 2005 for a review).

- She obtains an observation $x_{t, g}$ of the non-attitudinal dimension. We assume that $x_{t, g}$ is normally distributed (with mean 0 and variance 1)

If the agent does not sample the group, her attitude does not change ( $A_{t, g}=A_{t-1, g}$ ) and she does not obtain any additional observation of the $X$ dimension.

Perception of Variability Consistent with the sampling approach tradition, we assume that the agent processes sampled information correctly. Let $V_{t, g}$ denote the perceived variability on dimension $X$ at the end of period $t$. Here, we assume that this is given by the standard unbiased sample variance estimator. (In the next section, we consider other estimators of variability.)

$$
\begin{equation*}
V_{t, g}=\frac{1}{n_{t, g}-1} \sum_{k=1}^{t}\left(x_{k, g}-\bar{x}_{t, g}\right)^{2} I_{k, g} \tag{2}
\end{equation*}
$$

where $I_{k, g}$ is an indicator variable equal 1 if group $g$ is sampled in period $k$ (and equal to 0 otherwise), $n_{t, g}$ is the number of samples $\left(n_{t, g}=\sum_{k} I_{k, g}\right), \bar{x}_{t, g}$ is the mean of the sampled observations on the $X$ dimension at the end of period $t$, and $x_{k, g}$ is the observation in period $k$.

Sampling Rule To ensure that variability estimates exist for both groups, we assume that the agent has sampled 2 observations from each group before the first period. In the subsequent periods, the sampling rule follows that used in Denrell (2005). In each period the agent samples the in-group or the out-group based on the current attitude towards that group. Note that sampling rule does not depend on observations on


Figure 1: Model with unbiased variance estimator: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability $\left(P\left(V_{t, i n}>V_{t, \text { out }}\right)\right)$ as a function of time. Each point is based on $10^{5}$ simulations with $b=0.5, r=0.5, s=3$.
dimension $X$. The probability that the agent samples the ingroup is given by the exponential version of the Luce choice rule (Denrell, 2005):

$$
\begin{equation*}
P_{t+1, \text { in }}=r+(1-r) \frac{e^{s A_{t, \text { in }}}}{e^{s A_{t, \text { in }}}+e^{s A_{t, o u t}}}, \tag{3}
\end{equation*}
$$

Here $s$ is a parameter that regulates the sensitivity of the sampling probability to the current attitude, and $r \in[0,1]$ is a parameter that corresponds to the sampling 'bias' in favor of the in-group. The higher $r$ is, the higher is the baseline probability that the agent will sample the in-group. When $r$ is high, the agent is likely to frequently sample the in-group even if she has a negative attitude toward it ( $A_{t, i n}$ is low). This sampling 'bias' in favor of the in-group implies that the agent will gather larger samples of information about the in-group than about the out-group.

## Analysis

We ran computer simulations of the above model. The parameter values that were used in all simulations are $b=0.5, r=$ $0.5, s=3$. These values are similar to estimated parameter values in sequential choice experiments (Denrell, 2005).

Figure 1 displays the likelihood that the estimate of the ingroup variability is higher than the estimate of the out-group variability, $P\left(V_{t, \text { in }}>V_{t, \text { out }}\right)$, as a function of the number of periods. It is higher than 0.5 for all periods after period 1. In other words, the in-group tends to be perceived as more variable than the out-group. The likelihood that the in-group is perceived as more variable than the out-group first increases quickly and then decreases very slowly with the number of periods. It is equal to 0.54 after 50 periods and 0.53 after 100 periods. This suggests that this asymmetry persists even after many periods.


Figure 2: Model with unbiased variance estimator. Distribution of the sample sizes of the two groups after 15 periods. Based on $10^{5}$ simulations with $b=0.5, r=0.5, s=3$.

To develop an intuition for this result, we focus on the end of period 15. First note that the in-group is sampled more times than the out-group (Figure 2). This is because of the assumed sampling advantage of the in-group (eq. 3). Second, note that the distributions of sampled variabilities for the two groups are right skewed but to a different extent (see Figure 3). The distribution of the sampled variability for the in-group $V_{15, \text { in }}$ is less skewed than the distribution of the sampled variability for the out-group, $V_{15, \text { out }}$. By contrast, the mean sample variabilities are the same: ${ }^{1}$ $E\left(V_{15, \text { in }}\right)=E\left(V_{15, \text { out }}\right)=\sigma^{2}=1$. Overall, this implies that $V_{15, \text { in }}$ tends to be larger than $V_{15, \text { out }}$.

More generally, the distribution of variability estimates for a group is skewed, but the skewness decreases with sample size. If $V_{t, g}$ is based on $n$ observations, it is a random variable with a $\chi^{2}$ distribution with $n-1$ degrees of freedom (the mean is assumed to be unknown to the agent). The $\chi^{2}$ distribution is skewed to the right, therefore the probability that the variance estimate is lower than the true variance $\left(\sigma^{2}=1\right)$ is higher than 0.5 . Consider the probability mass below 1 for sample sizes $5,10,15,20$ and 50 . The probability masses are $0.59,0.56,0.55,0.54$, and 0.53 , respectively. In all cases, it is higher than $50 \%$, but it goes down as sample size increases and converges to 0.50 as the sample size becomes large. Because our assumptions about the sampling process imply that the sample collected about the in-group tends to be larger than the sample collected about the out-group, the distribution of $V_{t, i n}$ is likely less skewed than the distribution of $V_{t, o u t}$. This implies that $V_{t, \text { in }}$ is likely larger than $V_{t, o u t}$. In other words, an in-group heterogeneity effect emerges.

[^122]

Figure 3: Model with unbiased variance estimator. Distributions of the variability estimates (according to eq. 2) $V_{15, \text { in }}$ and $V_{15, \text { out }}$ after 15 periods. Black vertical line denotes the true variance on the $X$ dimension $(\sigma=1)$. Based on $10^{5}$ simulations with $b=0.5, r=0.5, s=3$.

## Other Estimators of Variability

A Bayesian Estimator of Variability An alternative implementation of our assumption that information is processed 'correctly' is to assume that the agent is a Bayesian processor of information, with correct priors about the true variance. For simplicity, we assume that the mean on the $X$ dimension is known and equal to $0 .{ }^{2}$ The true variance is drawn from an inverse gamma distribution with parameters $\alpha$ and $\beta$. The inverse gamma distribution is a conjugate distribution: the posterior also follows an inverse gamma distribution.

$$
\begin{equation*}
f\left(\hat{\sigma}_{t, g}^{2}\right) \sim I G\left(\alpha+\frac{\sum_{k=1}^{t} I_{k, g}}{2}, \beta+\frac{\sum_{k=1}^{t}\left(x_{k, g}-\bar{x}_{t, g}\right)^{2} I_{k, g}}{2}\right) \tag{4}
\end{equation*}
$$

We simulated the model by assuming the same attitude updating rule and sampling rule as before, but with the Bayesian estimator of variability in equation 4 . We assumed $\alpha=15 / 2$ and $\beta=\alpha$. ${ }^{3}$

Figure 4 displays the evolution of $P\left(V_{t, \text { in }}>V_{t, \text { out }}\right)$. In this Bayesian setting as well, the in-group tends to be perceived as more variable than the out-group. This result is important, because it demonstrates that even a rational processor of information will tend to perceive the in-group as more variable than the out-group in settings where the in-group is more likely to be sampled again (i.e. $r>0$ ). A similar pattern would emerge if the agent were also updating her attitudes $\left(A_{t, g}\right)$ using Bayes' rule, provided that the sampling rule implies that larger samples are obtained about the in-group than about the out-group.

[^123]

Figure 4: Model with Bayesian estimator of variability. Likelihood that the posterior estimate of in-group variability is higher than the posterior estimate of out-group variability $\left(P\left(V_{t, \text { in }}>V_{t, \text { out }}\right)\right.$ ) as a function of time. Based on $10^{5}$ simulations with $b=0.5, r=0.5, s=3, \alpha=15 / 2, \beta=\alpha$.

Sample Variance In the main analysis, we assumed that the agent estimated the group variabilities using a statistically unbiased estimator (eq. 2). We did so because we wanted to demonstrate that an asymmetry in the perception of variability could emerge even in this case. Several papers have assumed that people use the standard sample variance estimator (e.g. Juslin, Winman, \& Hansson, 2007; Kareev et al., 2002):

$$
\begin{equation*}
V_{t, g}=\frac{1}{n_{t, g}} \sum_{k=1}^{t}\left(x_{k, g}-\bar{x}_{t, g}\right)^{2} I_{k, g}, \tag{5}
\end{equation*}
$$

This estimator is biased for small samples, and the size of the bias is stronger for smaller samples. Unsurprisingly, simulations based on this estimator lead to a stronger asymmetry than in the above analyses. For example, after 15 periods, the likelihood that the estimate of in-group variability is higher than the estimate of out-group variability is $P\left(V_{15, \text { in }}>V_{15, \text { out }}\right)=0.62$. This number was 0.56 when the unbiased variance estimator was used.

We did not find empirical evidence that suggests that people's intuitive variability estimates are closer to the unbiased or the biased estimators. Qualitatively, this is not an issue for our argument, because the asymmetry in perception of groupvariability emerges in both cases. Further work should investigate this issue. This would allow for quantitative predictions about the size of the in-group heterogeneity effect.

## Relation to Existing Literature on In-group Heterogeneity

Most prior explanations of the in-group heterogeneity effect have invoked differences in how information about in-group and out-group is processed. Here we discuss how our ex-

[^124]planation differs from this prior work (we use a taxonomy similar to Ostrom and Sedikides (1992)).

Several explanations rely on motivated cognition (Kunda, 1990). The first mechanism invokes people's desires for positive identities. Those who want a positive social identity are motivated to view their in-groups more positively than other groups (Tajfel, 1982). At the same time, heterogeneity is frequently perceived as a positive feature of social groups (Ostrom \& Sedikides, 1992). Therefore, people are motivated to perceive the in-group as more heterogeneous than out-groups. A related mechanism invokes people's desire for distinct identities. A more heterogeneous in-group allows people to see themselves as unique within the in-group. Thus, people are motivated to see their in-groups as heterogeneous (Pickett \& Brewer, 2001). Yet another explanation based on motivated cognition notes that it is easier to dehumanize more homogeneous groups (Haslam, 2006; Brewer, 1999). Therefore, if the out-group is perceived as less variable than the ingroup, it is easier to justify negative attitudes and even cruel actions towards out-group members.
Another explanation notes that people tend to have prior beliefs that the out-group is more homogeneous. Park and Hastie (1987) showed that if participants first observed exemplars from a group followed by a description of its general characteristics, they perceived this group as more variable compared to when they observed that information in reversed order. This suggests that the prior about homogeneity affects how information is encoded. This finding implies an in-group heterogeneity effect under a (reasonable) assumption that people often learn descriptions of out-groups before interacting with some of their members (e.g. through stereotypes communicated by others in their environment) whereas they learn about natural in-groups by direct observations.

A third class of explanations notes that the self is part of the in-group (Park \& Judd, 1990). Since the self is often perceived as particularly differentiated and unique, this would contribute to an impression that the in-group is more heterogenous than the out-group.
A fourth class of explanations suggests that information about different groups is encoded and retrieved in different fashions. For example, Ostrom et al. (1993) found that information about in-group members is stored in categories related to individual information whereas the information about the out-group members is stored in categories related to stereotypical attributes. Therefore, when the information is recalled, the in-group tends to be associated with more individuating information compared to stereotype based homogeneous information about the out-group. In terms of recall, Park and Judd (1990) suggested that participants recall more extreme exemplars about in-groups than about outgroups. This suggests that memory search processes might differ across in-group and out-group.

These four classes of explanations emphasize features of information processing. By contrast, our explanation focuses properties of the sample of information on which the mind
operates. Because our explanation focuses on a different level than explanations that focus on information processing, it does not contradict these. Rather, it complements them. Our analyses and the experimental findings discussed above suggest that both types of mechanisms likely play a role in explaining why people see their in-groups as more heterogeneous than their out-groups.

The most closely related paper to ours is a paper by Linville et al. (1989). It analyzes an exemplar model that describes how information about groups is encoded, stored and recalled. The authors argue that higher familiarity with the in-group than with the out-group is the cause of the bias. Familiarity in this case is the number of exemplars of each group a person has encountered. They model the encoding, storage and recall of the information using a set of parameters and show that the strength of the bias depends on the information processing. They also consider the case of perfect memory (perfect encoding, no forgetting, and perfect recall). They find an asymmetry in expected variability estimates $\left(E\left(V_{\text {in }}\right)>E\left(V_{\text {out }}\right)\right)$. Their argument is similar to the logic of our model, but their analysis focuses on the asymmetry produced by reliance on the biased variance estimator we discussed above (see eq. 5). Our results differ, because they demonstrate that a systematic tendency to perceive the ingroup as more heterogeneous can emerge even when people use an unbiased estimator of variability. In some sense, our result is stronger because the asymmetry in expected variability implied by the biased variance estimator goes down very quickly with sample size. The asymmetry based on the skewness of the distribution of estimators of variability persists even as sample size becomes somewhat large (although it disappears for very large samples). Another difference is that our model focuses on sampling from the environment whereas this prior paper focused on sampling within the mind.

## Discussion

## Sensitivity Analysis

The magnitude of the in-group heterogeneity effect produced by our model depends on the model parameters.

The baseline probability of sampling the in-group ( $r$ in eq. 3) has a large effect on the magnitude of the bias. For $r$ values close to zero, the likelihood that the in-group is perceived as more variable becomes close to 0.5 (e.g., 0.51 for $r=0.05$ ), but when the advantage of the in-group is high ( $r=0.9$ ), the likelihood that the in-group is perceived as more variable can be as high as 0.64 (see Figure 5). The baseline probability of sampling from the in-group reflects the difficulty of obtaining information about the members of the outgroup. Its value depends on the empirical setting. For example, depending on the social group and socioeconomic structure of a country, the probability can range from small values (for fairly international and integrates societies) to very high values (in isolated homogeneous societies).

The other parameters, $b$ (the weight of the new attitude, see eq. 1) and $s$ (the slope parameter in the sampling rule, see


Figure 5: Model with unbiased variance estimator: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability $\left(P\left(V_{t, \text { in }}>V_{t, o u t}\right)\right.$ ) after 15 periods, as a function of the baseline probability of sampling the in-group $(r)$. Based on $10^{5}$ simulations with $b=0.5, s=3$.
eq. 3) have a positive effect on the strength of the in-group heterogeneity effect, but the effect is not strong.

A different but related model to ours would not assume an inherent sampling advantage for the in-group (take $r=0$ ). In this case, our model does not predict any in-group heterogeneity effect if the groups are equally attractive (i.e., $a_{t, i n}$ and $a_{t, \text { out }}$ are drawn from the same distribution). But suppose that the in-group is more attractive. It is possible to model this by assuming that the mean of the distribution of $a_{t, i n}$ is higher than the mean of the distribution of $a_{t, \text { out }}$ (for simplicity, we assume the variances are equal). In this case, the agent will obtain larger samples about the in-group than about the out-group and an in-group heterogeneity effect will emerge if $s$ and $b$ are high enough.

## In-Group Homogeneity

Several papers have documented instances of an in-group homogeneity effect that seems to contradict the dominant finding of an in-group heterogeneity effect (Simon \& Pettigrew, 1990; Rubin \& Badea, 2007). Our sampling mechanism can acomodate some of these findings.

An in-group homogeneity effect has been found when the feature under consideration is used as a basis for categorization. That is, the value of that feature defines whether the person is categorized into the in- or out-group (Rubin \& Badea, 2007). In this case, the true variability of the in-group might be smaller than the variability of the out-group on the focal feature. Our model can be adapted to this setting by relaxing the assumption that the true variances are the same for the two groups. We can assume instead $\sigma_{i n}^{2}<\sigma_{o u t}^{2}$. Our model implies that the variabilities of both groups will tend to be underestimated, but that the in-group variability will be un-
derestimated to a lower extent than the out-group variability. If the difference in the extent of underestimation is smaller than the difference in true variabilities, our model implies the emergence of an in-group homogeneity effect, in line with the true difference in variabilities. But if the difference in true variabilities is small, our model can lead to the emergence of an in-group heterogeneity effect.

Although most prior research conceptualized the 'ingroup' and the 'out-group' as specific groups, some papers have conceptualized the out-group as 'everyone but the ingroup' (e.g. Alves, Koch, \& Unkelbach, 2016). In this case, the true variability of the out-group is likely much larger than the true variability of the in-group. This setting is a special case of the setting discussed in the previous paragraph.

Finally, there is some evidence that when the in-group is a minority it tends to be judged as more homogeneous than the out-group (Simon \& Pettigrew, 1990). Our model can be adapted to this setting as well. Consider a situation where the in-group is smaller than the out-group and, furthermore, the number of in-group members is smaller than the number of periods. The agent will sample all the in-group members but the sample size will remain small (bounded by the number of members). To illustrate this, let us return to our example where the number of periods is $t=15$ and let us also assume that the sizes of the in-group and out-group are 5 and at least 10 , respectively. Then the sample size for the in-group will not exceed 5 whereas the size of the out-group members sample will be at least 10. This sample size asymmetry in favor of the out-group implies an in-group homogeneity effect.

## Conclusion

People frequently obtain larger samples of information about in-groups than about out-groups. Because estimators of variability tend to be more strongly right-skewed when samples are smaller, this implies that people will be likely to perceive in-groups as more variable than out-groups. In this paper, we showed that this in-group heterogeneity effect emerges even when people process information correctly - even if they are naive intuitive statisticians. Our theory complements existing explanations that proposed that information about in-group and out-group members was processed in different fashions.

## Acknowledgments

We are grateful for comments by Jerker Denrell, Robin Hogarth, participants in the Behavioral and Management Research Breakfast at UPF and UPF Behavioral Workshop, and the JDM 2016 meeting. This research was funded by grants AEI/FEDER UE-PSI201675353 and RYC-2014-15035 from the Spanish Ministry of Science and Innovation and grant IN[15]_EFG_ECO_2281 from the BBVA Foundation to Gaël Le Mens.

## References

Alves, H., Koch, A., \& Unkelbach, C. (2016). My friends are all alike the relation between liking and perceived similarity in person perception. Journal of Experimental Social Psychology, 62, 103-117.
Boldry, J. G., Gaertner, L., \& Quinn, J. (2007). Measuring the measures a meta-analytic investigation of the measures of outgroup
homogeneity. Group Processes \& Intergroup Relations, 10(2), 157-178.
Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate. Journal of social issues, 55(3), 429-444.
Denrell, J. (2005). Why most people disapprove of me: experience sampling in impression formation. Psychological review, 112(4), 951.

Fazio, R. H., Eiser, J. R., \& Shook, N. J. (2004). Attitude formation through exploration: valence asymmetries. Journal of personality and social psychology, 87(3), 293.
Fiedler, K., \& Juslin, P. (2006). Information sampling and adaptive cognition. Cambridge University Press.
Haslam, N. (2006). Dehumanization: An integrative review. Personality and social psychology review, 10(3), 252-264.
Hewstone, M., Rubin, M., \& Willis, H. (2002). Intergroup bias. Annual review of psychology, 53(1), 575-604.
Judd, C. M., \& Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. Journal of Personality and Social Psychology, 54(5), 778-788.
Juslin, P., Winman, A., \& Hansson, P. (2007). The naive intuitive statistician: a naive sampling model of intuitive confidence intervals. Psychological review, 114(3), 678.
Kareev, Y., Arnon, S., \& Horwitz-Zeliger, R. (2002). On the misperception of variability. Journal of Experimental Psychology: General, 131(2), 287.
Kunda, Z. (1990). The case for motivated reasoning. Psychological bulletin, 108(3), 480.
Le Mens, G., \& Denrell, J. (2011, April). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. Psychological Review, 118(2), 379-392.
Linville, P. W., \& Fischer, G. W. (1998). Group variability and covariation: Effects on intergroup judgment and behavior. Intergroup cognition and intergroup behavior, 123-150.
Linville, P. W., Fischer, G. W., \& Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. Journal of personality and social psychology, 57(2), 165.
Ostrom, T. M., Carpenter, S. L., Sedikides, C., \& Li, F. (1993). Differential processing of in-group and out-group information. Journal of Personality and Social Psychology, 64(1), 21.
Ostrom, T. M., \& Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. Psychological Bulletin, 112(3), 536.
Park, B., \& Hastie, R. (1987). Perception of variability in category development: Instance-versus abstraction-based stereotypes. Journal of Personality and Social Psychology, 53(4), 621-635.
Park, B., \& Judd, C. M. (1990). Measures and models of perceived group variability. Journal of Personality and Social Psychology, 59(2), 173.
Park, B., \& Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. Journal of Personality and Social Psychology, 42(6), 1051.
Pickett, C. L., \& Brewer, M. B. (2001). Assimilation and differentiation needs as motivational determinants of perceived in-group and out-group homogeneity. Journal of Experimental Social Psychology, 37(4), 341-348.
Rubin, M., \& Badea, C. (2007). Why do people perceive ingroup homogeneity on ingroup traits and outgroup homogeneity on outgroup traits. Personality and Social Psychology Bulletin, 33(1), 31-42.
Rubin, M., \& Badea, C. (2012). They're all the same... but for several different reasons a review of the multicausal nature of perceived group variability. Current Directions in Psychological Science, 21(6), 367-372.
Simon, B., \& Pettigrew, T. F. (1990). Social identity and perceived group homogeneity evidence for the ingroup homogeneity effect. European Journal of Social Psychology, 20(4), 269-286.
Tajfel, H. (1982). Social psychology of intergroup relations. Annual review of psychology, 33(1), 1-39.
Weber, E. U., Shafir, S., \& Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. Psychological review, 111(2), 430.

# Exploring the decision dynamics of risky intertemporal choice 

Emmanouil Konstantinidis (em.konstantinidis@gmail.com) ${ }^{1}$, Don van Ravenzwaaij (d.van.ravenzwaaij@rug.nl) ${ }^{2}$, \& Ben R. Newell (ben.newell@unsw.edu.au) ${ }^{1}$<br>${ }^{1}$ School of Psychology, University of New South Wales, Sydney, NSW 2052, Australia<br>${ }^{2}$ Department of Psychology, University of Groningen, 9712 TS Groningen, The Netherlands


#### Abstract

Previous research on the effects of probability and delay on decision-making has focused on examining each dimension separately, and hence little is known about when these dimensions are combined into a single choice option. Importantly, we know little about the psychological processes underlying choice behavior with rewards that are both delayed and probabilistic. Using a process-tracing experimental design, we monitored information acquisition patterns and processing strategies. We found that probability and delay are processed sequentially and evaluations of risky delayed prospects are dependent on the sequence of information acquisition. Among choice strategies, directly comparing the values of each dimension (i.e., dimension-wise processing) appears to be most favored by participants. Our results provide insights into the psychological plausibility of existing computational models and make suggestions for the development of a process model for risky intertemporal choice.


Keywords: Risky Intertemporal Choice; Process Tracing; Path Dependency; Sequential Processing; Decision Strategies

## Introduction

While research on risky choice and intertemporal choice have separately provided significant insights into the effects of probability and delay, decisions which involve both elements have received less scrutiny. Two unresolved questions are whether probability and delay are processed sequentially (and if yes, which dimension is considered first), and whether evaluation of risky delayed prospects is path dependent. Öncüler and Onay (2009) found that the order in which participants processed risky delayed prospects affected the final evaluations of these prospects. Using a process-tracing design, they found that amount-related (i.e., money) information was acquired first most often, followed by information about delay and then probability. Interestingly, when participants were required to process delay first, they provided higher evaluations of the same prospect compared to when they processed probability first, supporting the view of path dependency in risky intertemporal choice.

Despite being central to the characterization of choice behavior, no other studies have utilized process-tracing methods in the domain of risky intertemporal choice. Processtracing methods can provide insightful observations about the processes that take place before the actual decision, such as search, integration, and processing of available information (e.g., Reisen, Hoffrage, \& Mast, 2008). In risky intertemporal choice they can provide information about the order in which participants integrate amount, delay, and probability information as well as choice strategies adopted. Accordingly, process data can set the foundations for the development of com-
putational models and offer testable predictions regarding the choice process.

The experimental results relating to path dependency and sequential evaluation are not readily explained by traditional expected discounted utility models. These models focus on predicting choice outcomes (descriptive as-if models), hence arguably not accounting for the underlying psychological processes that are responsible for choice behavior, or simplifying strategies (e.g., heuristics) that people may employ in their decision-making. For instance, some expected utility models assume that people integrate probability and delay information into a common metric of psychological distance (e.g., Baucells \& Heukamp, 2012; Vanderveldt, Green, \& Myerson, 2015). However, Öncüler and Onay (2009) observed in their process data that this strategy was not favored (i.e., transitions between probability and delay information were the least frequent), thus rendering the "common psychological distance" account less likely among competing explanations.

The main purpose of the current work is to extend Öncüler and Onay's (2009) investigation from a pricing task, in which participants had to indicate the present certainty equivalent or pCE of a risky delayed prospect (the minimum amount of money that one is willing to accept instead of a delayed gamble) to a choice task in which participants choose between two risky delayed prospects. We then examine the predictions for path dependency and sequential evaluation in both tasks. This comparison allows us to ask whether choice is also characterized by path dependency and whether the characteristics of this dependency are similar between choice and pricing. The identification of such characteristics and processing strategies can also inform the development of models that rely on psychologically plausible accounts of choice behavior (i.e., psychological process models). Such models have become increasingly popular in many areas of decisionmaking (e.g., Koop \& Johnson, 2013), and often assume that decision-making follows simple rules of information processing, such as dimension-wise evaluation, sequential processing, and partial integration of available information (see e.g., Brandstätter, Gigerenzer, \& Hertwig, 2006).

## Method

## Participants

We tested a total of 63 undergraduate students ( 42 female; Age: $M=19.02, S D=1.56$ ) at the University of New South Wales who participated in return for course credit.


Figure 1: Schematic representation of the experimental design in the pricing (A) and choice (B) tasks. In the pricing task, participants could open as many boxes as they wanted before they gave the pCE of the delayed lottery. There was no limit about the time inspecting a box. The position of each dimension on the screen (Amount, Probability, and Delay) was randomized across trials. The association between colored boxes and dimensions remained invariant throughout each experimental session but was randomized across participants. The design was identical in the choice task and participants had to choose between two delayed lotteries. In this screenshot, the mouse opens an Amount box in the pricing task (i.e., \$120) and a Delay box of Option 1 in the choice task (i.e., 13 months).

## Task and design

We used a process-tracing design (i.e., similar to a MouseLab information board; see Payne, Bettman, \& Johnson, 1993) to monitor information acquisition strategies and processing steps. The experiment consisted of two parts: the first part was a pricing task, where participants had to indicate the pCE of 68 delayed lotteries (presented sequentially and in random order). For each delayed lottery, there were 3 colored boxes in the center of the screen, each containing the numerical value of each lottery's dimensions: Amount of money (in \$), Probability (in \%), and Delay (in months; see Figure 1A). However, this information was hidden and revealed only upon clicking on each corresponding box. When participants clicked on a box, it stayed active (i.e., showing its value) as long as the mouse cursor was within the borders of the box. When they moved the mouse out of the box, it returned to its default state (i.e., hidden). There was no limit in the amount of clicks or the time inspecting a box. In addition, participants could return to an already seen box if they wanted to. The position of each box was randomized across trials.

The second part was a choice task (Figure 1B), which always followed the pricing task and involved a choice between two delayed lotteries. Unbeknownst to the participants, the choice dyads were formed using pairs of prospects from the pricing task ( 34 choice pairs from 68 delayed lotteries). The procedure of acquiring information about each delayed lottery was identical to the pricing task.

## Procedure

Participants sat in front of a computer screen and were given instructions about the task (e.g., details about the information acquisition in the pricing task and what pCE represents). There was also a practice stage prior to the main task where
participants could familiarize themselves with the processtracing character of the task. For the pricing task, there was a box where they could type in their evaluation (Figure 1A). For the choice task, they were told that the task is exactly the same as before, with the only differences being that there was an extra option on the screen and they had to choose between the two, by clicking on the corresponding option label (Figure 1B).

## Results

## Pricing Task

All participants completed the experiment. We excluded one participant because they never acquired probability and delay-related information. Our initial objective was to explore the basic properties of information acquisition in the pricing task (see Figure 2): the frequency that each dimension was inspected, the frequency that each dimension was inspected first (i.e., at the beginning of each trial), last (i.e., before participants provided the pCE value), and intermediate (i.e., excluding first and last inspection items), and the mean inspection time for each dimension. For the analysis of frequency data, we used a linear multilevel model with dimension as fixed-effect and random intercepts for each participant. We applied a square root transformation for the frequency data. ${ }^{1}$ As Figure 2A suggests, participants acquired more amount-related information, followed by probability and delay, and this pattern was present in all categories of interest (All: $\chi^{2}(2)=269.07, p<.001$; First: $\chi^{2}(2)=$ 60.67, $p<.001$; Intermediate: $\chi^{2}(2)=77.69, p<.001$; Last:

[^125]$\left.\chi^{2}(2)=96.18, p<.001\right) .{ }^{2}$ The same pattern is observed in the mean inspection time (Figure 2B): participants spent more time looking at amount information, followed by probability and delay $\left(\chi^{2}(2)=67.20, p<.001\right.$; significant differences between each dimension). Also, the relative preference for inspecting each dimension does not seem to change over time as can be seen in Figure 2C: block $\times$ dimension interaction, $\chi^{2}(6)=1.19, p=0.98$.


Figure 2: Graphical representation of information acquisition in the pricing task: A) Relative frequencies of opened boxes of each dimension (All: Overall; First \& Last: First and Last boxes opened in a trial; Intermediate: Excluding First and Last boxes). B) Mean inspection time for each dimension. C) Pattern of acquisition items across blocks of trials (17 trials/lotteries each).

The next step in our analysis was to inspect transitions between consecutive ( $n \leftrightarrow n+1$ ) information items. This analysis can provide us with information about the sequential nature of risky intertemporal choice. For example, based on the adjacency principle ("information used in temporal proximity should be acquired in close proximity"; see Johnson, Schulte-Mecklenbeck, \& Willemsen, 2008, p. 264), if the Amount $\leftrightarrow$ Probability transition occurs more often and temporally precedes the Amount $\leftrightarrow$ Delay transition, it means that participants pay more attention to the amount and probability aspects of the prospect and probability discounting (or processing of probability) occurs prior to delay discounting (or processing of delay). Table 1 suggests that the Amount $\leftrightarrow$ Probability transition not only occurs more often than any other transition (All column), but it seems to precede any other transition (First column), and be considered more often before the final evaluation of the lottery (Last column). The relatively low proportion of Delay $\leftrightarrow$ Probability transitions

[^126]suggests that participants are not attempting to create a common metric of psychological distance by integrating these two dimensions (cf. Öncüler \& Onay, 2009).

Table 1: Transitions between dimensions in Experiment 1. The $\leftrightarrow$ symbol indicates all transitions from one dimension to the other.

| Transition | All: $N$ | All: \% | First: \% | Last: \% |
| :--- | :---: | :---: | :---: | :---: |
| Amount $\leftrightarrow$ Probability | 8,834 | 0.46 | 0.47 | 0.50 |
| Delay $\leftrightarrow$ Amount | 5,117 | 0.27 | 0.28 | 0.27 |
| Delay $\leftrightarrow$ Probability | 4,335 | 0.23 | 0.22 | 0.19 |

Note: Relative frequencies do not add up to 1 because transitions between the same dimension (e.g., Amount $\leftrightarrow$ Amount) are not included in the table.

We then explored the concept of path dependency as suggested by Öncüler and Onay (2009) by comparing the final evaluations of lotteries when Amount $\leftrightarrow$ Probability or Amount $\leftrightarrow$ Delay was the first occurring transition. Öncüler and Onay found that when participants followed the Amount $\leftrightarrow$ Delay path they gave higher evaluations of the same prospect compared to the Amount $\leftrightarrow$ Probability path. Our results replicate this effect: when examination of delay preceded that of probability, participants gave higher evaluations for the majority of trials ( $70 \%$ ). However, it is not clear how subsequent transitions in our experiment might have affected the final evaluation of the prospect. We try to address this issue along with the issue of imbalance in transitions (which emerges due to the higher frequency of Amount $\leftrightarrow$ Probability transitions) in a following experiment.

## Choice Task

Figure 3 presents information acquisition for each dimension in the choice task, aggregated across the two choice options. The pattern of results looks similar to the pricing task with a few exceptions: First, looking at the overall trend of dimension inspection, there is no difference between amount and probability (Figure 3A; $b=-0.01, t=-0.84, p=.40$ ), but both differ with respect to delay (pairwise contrasts, $p<$ .001). A similar pattern is observed in the intermediate inspection items (no difference between amount and probability, $b=0.07, t=0.34, p=.73$, but they both differ from delay, $p<.001$ ). This presents a difference between the two methods of preference elicitation, indicating that in a choice setting amount and probability may have the same degree of influence on choice. As in the pricing task, the first dimension considered followed the amount $>$ probability $>$ delay scheme, $\chi^{2}(2)=37.23, p<.001$, but there was no difference between dimensions regarding the last information item, $\chi^{2}(2)=2.68, p=.26$. The mean time spent at each dimension (Figure 3B) was not different between amount and probability ( $p=.95$ ), but both were higher than delay ( $p<.001$ ). Regarding selection of each dimension across time, Figure 3C suggests that it does not change between the two halves of the choice task, $\chi^{2}(2)=2.67, p<.001$.


Figure 3: Graphical representation of information acquisition in the choice task, aggregated across the two options: A) Relative frequencies of opened boxes of each dimension (All: Overall; First \& Last: First and Last boxes opened in a trial; Intermediate: Excluding first and last). B) Mean inspection time for each dimension. C) Pattern of acquisition items across blocks of trials ( 17 trials/choices each).

Next, we examined transitions between consecutive information items ( $n \leftrightarrow n+1$ ) which are informative of the strategies that participants use. Assuming a $2 \times 3$ information board/grid where rows represent choice alternatives and columns dimensions (as in Figure 1B), transitions between items can be categorized as dimension-wise (or intradimensional: when the transition examines the same dimension between the two choice options, e.g., Amount in Option $1 \leftrightarrow$ Amount in Option 2), alternative-wise (or interdimensional: when the transition moves between different dimensions of the same option, e.g., Amount in Option $1 \leftrightarrow$ Delay in Option 1), diagonal (i.e., when the transition moves from one dimension of one option to a different dimension of the other option, e.g., Probability in Option $2 \leftrightarrow$ Delay in Option 1), and same (i.e., two consecutive inspections of the same dimension in the same option, e.g., Probability in Option $1 \leftrightarrow$ Probability in Option 1). Table 2 presents the frequency of each of the categories of transitions in the choice task: a first inspection of all transitions (All column) suggests that participants equally combined dimension and alternative-wise strategies. One of the most commonly used strategy indices (SI; Payne, 1976) suggests that participants equally used both strategies to make decisions. The SI is a ratio of the difference between alternative and dimension-wise transitions and it is defined as $\mathrm{SI}=\left(r_{a}-r_{d}\right) /\left(r_{a}+r_{d}\right)$, where $r_{a}$ is the total number of alternative-wise transitions and $r_{d}$ is the total number of dimension-wise transitions. It ranges between -1 to +1 , with negative numbers suggesting more dimension-wise processing and positive numbers suggesting more alternativewise processing. For our data, the $S I$ equaled 0.06 , indicating
roughly equal use of both strategies.
Table 2: Categories of transitions in the choice task. Arrows indicate the direction of the transition within the information board (see Figure 1B).

| Transition |  | All: $N$ | All: $\%$ | First: $\%$ | Last: $\%$ |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Dimension | $\uparrow \downarrow$ | 7,115 | 0.39 | 0.58 | 0.54 |
| Alternative | $\rightleftarrows$ | 8,072 | 0.44 | 0.39 | 0.39 |
| Diagonal | $\nearrow$ | 3,117 | 0.17 | 0 | 0.07 |
| Same | - | 146 | 0.01 | 0.02 | 0.01 |

However, Böckenholt and Hynan (1994) argued that the SI is a biased measure of strategy use when there is a different number of alternatives and dimensions. Specifically, if the number of dimensions is larger than the number of the alternatives ${ }^{3}$, then a positive SI is to be expected, indicating more alternative-wise processing. Böckenholt and Hynan developed an index (strategy measure; SM) which takes into account all possible transitions (e.g., including Diagonal and Same in Table 2):

$$
\begin{equation*}
S M=\frac{\sqrt{N}\left[\left(\frac{A D}{N}\right)\left(r_{a}-r_{d}\right)-(D-A)\right]}{\sqrt{A^{2}(D-1)+D^{2}(A-1)}} \tag{1}
\end{equation*}
$$

where $N$ is the total number of all types of transitions, $A$ is the number of alternatives, $D$ is the number of dimensions, and $r_{a}$ and $r_{d}$ denote frequency for alternative-wise and dimensionwise transitions, respectively. As with the SI, negative values of the SM indicate more dimension-wise processing, as can be seen in Figure 4. Specifically, dimension-wise processing becomes more prevalent as time progresses, as indicated by the linear decrease of the $S M$ value.


Figure 4: Strategy measure ( $S M$ ) averaged across participants for each trial/choice of the task.

## Experiment 1B: Constrained sequential search

Method The purpose of this experiment was to examine in detail the effect of path dependency observed in the pricing task. Participants $\left(N=40, M_{\text {age }}=19.08\right)$ provided the pCE of risky delayed prospects sequentially (i.e., in two stages;

[^127]see also upper panel in Figure 6): in the first stage, they could see either probability or delay (amount was always visible on the screen), and give the present value of the prospect (if delay was presented first), or the certainty equivalent value (if probability was presented first). The value they provided in the first stage appeared in the second stage along with the numerical value of the unseen dimension, and participants had to provide a second and final value for the prospect. We manipulated (three experimental parts) the way that participants acquired probability and delay-related information: a) a free search part where participants could select to see either probability or delay in the first stage, and b) two constrained search parts where either probability or delay was presented to participants first. Hence, participants were presented with the same risky delayed prospect three times.
Results We first examined search patterns in the free search part of the experiment: we found that in $68.50 \%$ of all trials, participants chose to see probability first, replicating the effect we observed in the pricing and choice tasks, that is a preference for inspecting and integrating probability information before delay information. We also examined search patterns as a function of the amount offered (amount was always visible on the screen). Figure 5 presents an interesting pattern: participants' tendency to inspect the probability dimension first increases as amount increases. Despite the overall preference for acquiring probability first (even in the lowest amount category, [ 50,175 ), it is $64.38 \%$ ) the difference between the lowest and highest amount categories is about $10 \%$ ( $74.16 \%$ in the last category).


Figure 5: Proportion of trials in which participants chose to see probability first as a function of Amount (in \$; binned in five equal categories).

Regarding path dependency, we examined the final pCEs in the constrained search parts of the task. Figure 6 shows the proportion of participants that gave a higher final evaluation when they were constrained to inspect probability first (as compared to delay first) as a function of the numerical values of each gamble's probability (lower panel A) and delay (lower panel B). For example, the leftmost data-point in Figure 6A indicates that for the same risky delayed prospect (which has a probability of $2 \%$ ) about $40 \%$ of all participants gave a higher final pCE when they were presented with probability information first than when they were presented
with delay information first (see also the table in the upper panel). Even though Figure 6 essentially ignores interactions between each dimension and collapses across all amount, probability, and delay values, it shows some interesting patterns. First, as the probability in a prospect increases, the proportion of participants who gave a higher evaluation when they were presented with probability first increases, as shown by a multilevel logit regression with probability as fixed effect and participant-specific random intercepts (standardized $b=0.40, z=4.44, p<.001)$. Second, there is a similar trend in the delay panel (as temporal distance increases, the proportion of participants that gave a higher evaluation of the same gamble when they were presented with probability first increases) but it is not as pronounced as in the probability panel (standardized $b=0.24, z=2.70, p=.007$ ). Interestingly, this pattern seems to apply to small values of probability and delay, as when we constrain our analysis in the upper half of both scales (i.e., $50 \%$ to $90 \%$ for probability; 16 to 24 months for delay) the effect disappears (both multilevel logit regressions, $p>.05$ ). Overall, our results replicate the path dependency patterns in Öncüler and Onay (2009) and suggest that path dependency is not stable, but is moderated by the numerical values of each dimension.

|  | Probability First |  |  |  | Delay First |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Stage 1 | \$450 | $2 \% \rightarrow$ | $X_{P F}$ |  | \$450 | $21 \mathrm{M} \rightarrow$ | $X_{D F}$ |
| Stage 2 | $X_{P F}$ | $21 \mathrm{M} \rightarrow$ | $Y_{P F}$ | > | $X_{D F}$ | $2 \% \rightarrow$ | $Y_{D F}$ |



Figure 6: Upper panel: Representation of the task (M: Months). Lower panel: Proportion of participants who gave a higher final pCE (i.e., $Y_{P F}>Y_{D F}$ ) for the same gamble when they were presented with information about probability first across Probability (A) and Delay levels (B).

## Discussion

We set out to uncover the strategies and information acquisition patterns that people use when they evaluate and make decisions about risky delayed prospects. Using three differ-
ent process-tracing tasks to elicit preferences, we observed systematic patterns relating to search, integration, processing, and strategy-use. First, participants acquired more amountrelated information, followed by probability and delay in the pricing task, whereas in the choice task amount and probability appeared to have the same degree of influence on determining choice. Our results are in accordance with recent studies in intertemporal risky choice which found that probability might play a more important role than delay (e.g., Konstantinidis, van Ravenzwaaij, Güney, \& Newell, 2017; Vanderveldt et al., 2015), but are at odds with Öncüler and Onay (2009) who found that participants preferred to acquire delay information before and more frequently than probability information.

Second, Amount $\leftrightarrow$ Probability transitions were more frequent and preceded any other transition in the pricing task. This pattern of results suggests that evaluation of risky delayed prospects is subject to sequential processing. Also, the integration of probability and delay into a common psychological distance measure seems less likely as the Probability $\leftrightarrow$ Delay transition occurs less frequently and temporally follows other types of transitions.

Third, regarding path dependency and sequential processing, our constrained search experiment revealed that the final evaluation of risky delayed prospects is not only dependent on the path taken (i.e., integrating probability information before delay information, and vice-versa), but on the numerical values of each dimension. For example, when participants were first presented with low probability values, they largely discounted the final value of the same prospect as compared to when they saw delay-related information first about the same prospect. We found that the effect of path dependency observed in Öncüler and Onay (2009), that is, the Delay $\rightarrow$ Probability path generating higher values than the Probability $\rightarrow$ Delay path, is only observed with small probabilities.

Fourth, examination of transitions in the choice task reveals that participants employ dimension-wise strategies more frequently than alternative-wise strategies to make decisions in risky intertemporal choice settings. Even though there was no reliable difference between dimension and alternative-wise processing regarding the total number of transitions, taking into account different measures of strategy use (e.g., search indices, overall, first and last inspection items, and transitions between items), we found that dimension-wise processing may be more prevalent among participants, supporting recent studies which found that dimension-wise models in the domains of risky choice and intertemporal choice outperform their alternative-wise counterparts (e.g., Dai \& Busemeyer, 2014).

Lastly, even though individual information acquisition patterns might reflect noisy and idiosyncratic use of strategies, we identified systematic processing strategies and information acquisition patterns that a process model (or any other type of model) in the field of risky intertemporal choice should take into account. Our results also provide testable
grounds for psychological assumptions in models of risky intertemporal choice: we found little evidence that participants treat probability and delay as representing a common factor of psychological distance, or that probability can be translated into delay, and vice-versa.

## Acknowledgments

This research was supported by an Australian Research Council Discovery Project Grant (DP140101145). We thank Emma Williams and Dominic Tran for their help in collecting the data, and Ashley Luckman for helpful discussions.

## References

Baucells, M., \& Heukamp, F. H. (2012). Probability and time trade-off. Management Science, 58, 831-842.
Böckenholt, U., \& Hynan, L. S. (1994). Caveats on a processtracing measure and a remedy. Journal of Behavioral Decision Making, 7, 103-117.
Brandstätter, E., Gigerenzer, G., \& Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. Psychological Review, 113, 409-432.
Dai, J., \& Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. Journal of Experimental Psychology: General, 143, 1489-1514.
Johnson, E. J., Schulte-Mecklenbeck, M., \& Willemsen, M. C. (2008). Process models deserve process data: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). Psychological Review, 115, 263-272.
Konstantinidis, E., van Ravenzwaaij, D., Güney, S., \& Newell, B. R. (2017). Now for sure or later with a risk? Modeling risky inter-temporal choice as accumulated preference. Manuscript submitted for publication.
Koop, G. J., \& Johnson, J. G. (2013). The response dynamics of preferential choice. Cognitive Psychology, 67, 151-185.
Öncüler, A., \& Onay, S. (2009). How do we evaluate future gambles? Experimental evidence on path dependency in risky intertemporal choice. Journal of Behavioral Decision Making, 22, 280-300.
Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. Organizational Behavior and Human Performance, 16, 366-387.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1993). The adaptive decision maker. Cambridge, England: Cambridge University Press.
Reisen, N., Hoffrage, U., \& Mast, F. W. (2008). Identifying decision strategies in a consumer choice situation. Judgment and Decision Making, 3, 641-658.
Vanderveldt, A., Green, L., \& Myerson, J. (2015). Discounting of monetary rewards that are both delayed and probabilistic: Delay and probability combine multiplicatively, not additively. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41, 148-162.
Willemsen, M. C., \& Johnson, E. J. (2011). Visiting the decision factory: Observing cognition with MouselabWEB and other information acquisition methods. In M. SchulteMecklenbeck, A. Kühberger, \& R. Ranyard (Eds.), A handbook of process tracing methods for decision research (pp. 21-42). New York, NY: Psychology Press.

# Consistent Probabilistic Simulation Underlying Human Judgment in Substance Dynamics 

James Kubricht ${ }^{\star 1}$<br>kubricht@ucla.edu<br>Demetri Terzopoulos ${ }^{3}$<br>dt@cs.ucla.edu<br>${ }^{1}$ Department of Psychology

Yixin Zhu ${ }^{\star 2}$<br>yixin.zhu@ucla.edu<br>Song-Chun Zhu ${ }^{2,3}$<br>sczhu@stat.ucla.edu<br>${ }^{2}$ Department of Statistics

Chenfanfu Jiang ${ }^{\star 3}$<br>cffjiang@cs.ucla.edu

Hongjing Lu ${ }^{1,2}$
hongjing@ucla.edu

* Equal Contributors University of California, Los Angeles


#### Abstract

A growing body of evidence supports the hypothesis that humans infer future states of perceived physical situations by propagating noisy representations forward in time using rational (approximate) physics. In the present study, we examine whether humans are able to predict (1) the resting geometry of sand pouring from a funnel and (2) the dynamics of three substances-liquid, sand, and rigid balls-flowing past obstacles into two basins. Participants' judgments in each experiment are consistent with simulation results from the intuitive substance engine (ISE) model, which employs a Material Point Method (MPM) simulator with noisy inputs. The ISE outperforms ground-truth physical models in each situation, as well as two data-driven models. The results reported herein expand on previous work proposing human use of mental simulation in physical reasoning and demonstrate human proficiency in predicting the dynamics of sand, a substance that is less common in daily life than liquid or rigid objects.


Keywords: Intuitive physics; mental simulation; substance representation; prediction

## Introduction

Consider KerPlunk, a children's game in which marbles are suspended in the air by a lattice of straws within a cylindrical tube. The goal of the game is for each player to take turns removing straws while minimizing the number of marbles that fall through the lattice. The task requires players to reason about the interaction between rigid bodies and obstacles in 3D space. But what if the marbles were replaced by balls of liquid or sand? Could humans predict how those substances would move? Would those predictions agree with a generative model based on ground-truth, Newtonian physics?

Recent computational evidence has demonstrated that human predictions do agree with Newtonian physics, given noisy perception and prior beliefs about spatially represented variables: i.e., the noisy Newton hypothesis (Bates, Yildirim, Tenenbaum, \& Battaglia, 2015; Battaglia, Hamrick, \& Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2015; Hamrick, Battaglia, Griffiths, \& Tenenbaum, 2016; Kubricht et al., 2016; Sanborn, 2014; Sanborn, Mansinghka, \& Griffiths, 2013; K. Smith, Battaglia, \& Vul, 2013). The hypothesis suggests that humans rationally infer the values of physical variables and utilize normative conservation principles (approximately) to make predictions about future scene states. Computationally, this is achieved by sampling the initial locations, motions from noisy sensory input, and sampling physical attributes in a physical scene, propagating these variables forward in time according to approximated physical principles, and aggregating queries on the final scene states to form predicted response distributions.

Bates et al. (2015) extended the noisy Newton framework from block tower judgments (Battaglia et al., 2013) to liquid dynamics using an intuitive fluid engine (IFE). In their IFE, ground-truth physics was approximated using smoothed particle hydrodynamics (SPH (Monaghan, 1992), a particlebased computational method for simulating non-solid dynamics. Their model predictions matched human judgments about future fluid states and outperformed alternative models that did not employ probabilistic simulation or account for physical uncertainty. Furthermore, the authors found that their participants' predictions were sensitive to latent fluid attributes (stickiness and viscosity), suggesting that humans have rich knowledge about the intrinsic properties of liquid.

The present study argues for the same general class of model as Bates et al.'s (2015) IFE and extends their work by examining (1) whether human predictions about future states of multiple substances (i.e., rigid balls, liquid, and sand) differ, and (2) whether those differences can be consistently modeled using approximate, probabilistic simulation based on a hybrid particle/grid simulator adapted from previous work (Kubricht et al., 2016). Although granular materials (e.g., sand) are encountered in everyday life, they are far less common than liquid; can humans accurately predict how sand will interact with obstacles and support surfaces? We present two experiments exploring the human capacity to predict the dynamics of substances varying in familiarity and physical properties, examining how human judgments and model predictions vary for different substances. Experiment 1 examines human predictions about the resting composition of sand after pouring from a funnel. In Experiment 2, participants make predictions about the flow of liquid, sand, and rigid balls past obstacles using a design similar to Bates et al.'s (2015) study.

## Computational Models

## MPM Physical Simulator

The Material Point Method (MPM) (Sulsky, Zhou, \& Schreyer, 1995) is commonly used in computer graphics to simulate the behavior of solids and fluids. The MPM has produced physically accurate and visually realistic simulations of the dynamics of liquid (Jiang, Schroeder, Selle, Teran, \& Stomakhin, 2015) and sand (Klár et al., 2016), in addition to general continuum materials such as stiff elastic objects (Jiang, Schroeder, Teran, Stomakhin, \& Selle, 2016).

The Appendix presents a mathematical overview of our MPM simulator, which provides a unified, particle-based simulation framework that handles rigid balls, liquid, and
sand with essentially the same numerical algorithm, albeit with appropriately differing material parameters. The MPM method is physically accurate, numerically stable, and computationally efficient, enabling us to synthesize a large set of stimuli in a short amount of time by simply varying material parameters and the locations of the initial objects and colliding geometries. Running all the simulations in the same framework for the purposes of the present study also enables fair comparisons among the three types of substances, since we avoid potential inconsistencies in the numerical accuracies of multiple simulators specialized to particular materials.

## Intuitive Substance Engine

Although the MPM simulator provides accurate and stable kinematics and dynamics for liquid, sand, and rigid balls using a unified framework, this high-precision, deterministic process does not account for the variability of human judgments in various intuitive physics tasks. Inspired by previous implementations of the noisy Newton framework (e.g., Bates et al., 2015; Battaglia et al., 2013), we combined our MPM simulator with noisy inputs, yielding an Intuitive Substance Engine (ISE) that accounts for uncertainty in human perception and reasoning in physical situations involving the three substances examined in this study. Details on how noisy perceptual inputs are defined and sampled are provided in the Model Results section of each experiment.

It is important to note that our ISE (employing MPM simulation) is roughly equivalent to Bates et al.'s (2015) IFE (employing SPH simulation) in that both models apply the noisy Newton framework to substance dynamics. Indeed, SPH is a viable method for simulating the dynamics of both granular materials and liquids, although MPM provides a more efficient and accurate means of doing so. We do not envision that the predictions of the two methods would differ substantially from one another when applied to a given set of stimuli.

## Data-Driven Models

Two data-driven models based on statistical learning methods were constructed as competing models-the generalized linear model (GLM) (McCullagh, 1984) and Extreme Gradient Boosting (XGBoost) (Chen \& Guestrin, 2016). GLM is a classic machine learning method, commonly expressed by $\mathbf{Y}=\mathbf{X B}+\mathbf{U}$, where $\mathbf{X}$ is the feature input matrix, $\mathbf{B}$ is the parameter matrix (learned using a training dataset), and $\mathbf{U}$ is the error between the ground truth matrix $\mathbf{Y}$ and prediction $\mathbf{X B}$.

XGBoost is a recently-published machine learning method which has been utilized by multiple research teams to achieve outstanding performance in several Kaggle competitions. Essentially, it is a type of tree ensemble model: i.e., a set of classification and regression trees (CART). Formally, $\hat{y}_{i}=$ $\sum_{k=1}^{K} f_{k}\left(x_{i}\right)$, where $K$ is the number of trees, $f_{k}$ is a function in the functional space $\mathbf{F}$ comprising the set of all possible CARTS. The objective function is defined as $R(\theta)=$ $\sum_{i=1}^{n} l\left(y_{i}, \hat{y}_{i}\right)+\sum_{k=1}^{K} \Omega\left(f_{k}\right)$, where $\theta$ includes the model parameters to be learned during training, $l$ is the loss function, which measures the cost between ground truth $y_{i}$ and prediction $\hat{y}_{i}$, and $\sum_{k=1}^{K} \Omega\left(f_{k}\right)$ is a regularization term that prevents the model from over-fitting the training data.


Figure 1: Intermediate frames from the demonstration video in Experiment 1 from the (A) zoomed-out and (B) zoomed-in perspective. (C) Sand pile choices in Experiment 1's judgment task.

## Experiment 1

The first experiment was designed to determine whether humans are able to predict the resting geometry of sand after it is poured from a funnel onto a surface, and whether dynamic visualizations of the pouring behavior facilitate mental simulation of sand-surface interactions.

## Participants

A total of 108 undergraduate students ( 81 females), of mean age $=20.2$ years, were recruited from the University of California, Los Angeles (UCLA), Department of Psychology subject pool and were compensated with course credit.

## Materials and Procedure

Participants first viewed a demonstration video of sand falling from a funnel suspended 10 cm above a level surface. The pouring event was viewed three times from a zoomed-out perspective (Fig. 1A) and then a zoomed-in perspective (Fig. 1B). The duration of the video was 29 sec. After viewing the demonstration video, participants were presented with a sand-filled funnel suspended $1 / 2,1,2$, and 4 cm above the surface in a randomized order.

Forty-three participants were assigned to the Static Condition and viewed a static image (zoomed-out) in which the funnel was positioned at a particular height. Sixty-five were assigned to the Dynamic Condition and viewed a video (zoomed in and out; looped three times; 35 sec duration) of sand pouring from a funnel that was positioned at different heights above the surface. In the Dynamic Condition, the region of the surface where the sand fell was occluded by a gray rectangle.

After viewing each situation, participants were asked to indicate which of four sand piles would result from the sand pouring from the funnel at the indicated height (Fig. 1C). For each trial, the stimulus images (for the Static Condition) and final video frames (Dynamic Condition) remained on the screen until a response was made. The pile choices were shown from the zoomed-in perspective and represented the


Figure 2: Model prediction results compared to human judgments. (Upper) Static Condition. (Lower) Dynamic Condition. Each bar, 1, 2, 3, and 4, corresponds to testing trials with funnel height $1 / 2,1,2$, and 4 cm , respectively.
ground-truth resting geometries resulting from each situation: i.e., Piles $1,2,3$, and 4 correspond with the pile resulting from funnels suspended $1 / 2,1,2$, and 4 cm above the surface, respectively. The experiment consisted of 4 trials. The stimulus videos can be viewed at https://vimeo.com/216585992.

## Human Results

At each funnel height, the proportion of participants choosing each sand pile did not differ between the Dynamic and Static Conditions: $\chi^{2}(3)=2.21,2.34,2.41$, and 1.13 for funnel heights of $1 / 2,1,2$, and 4 cm , respectively. These results suggest that dynamic visualizations of sand pouring from the funnel in each situation did not alter participants' judgments about the sand's resting geometry. However, the participants' pile choices did vary across different heights $\left(\chi^{2}(9)\right.$ $=176.54$ ), indicating that funnel height influenced their predictions on the resting geometry of falling sand.

As shown in Fig. 2, participants' pile choices shifted toward higher-numbered, flatter piles as funnel height increased. These results indicate that participants' predictions were sensitive to funnel height, but inconsistent with groundtruth resting states. In the next section, predictions from the three computational models (ISE, GLM, and XGBoost) are compared to human performance to determine whether the noisy Newton framework can account for participants' deviations from ground-truth judgments.

## Model Results

ISE Predictions: The input variables for our ISE in Experiment 1 were funnel height (i.e., initial sand height) with perceptual uncertainty and sand friction angle with mental simulation uncertainty. Given the ground-truth values of initial funnel height and friction angle $\left(H_{i T}, \theta_{i T}\right), N=10,000$ noisy samples $\left\{\left(H_{i}, \theta_{i}\right), \quad i=1, \ldots, N\right\}$ were generated and passed to our MPM simulator, which returned the final height of the
sand pile for each sample. Instead of choosing from 4 piles (i.e., the task presented to the participants), the MPM simulator compares the estimated height of the final sand pile, formally $D\left(H_{i}, \theta_{i}\right)=H_{p} \in \mathbb{R}>0$, with the heights of the 4 pile options given to human participants. The pile option with the minimum height difference was chosen as the predicted judgment for each sample. Finally, by aggregating predictions across the 10,000 samples, our ISE outputs a predicted response distribution for each trial.

To model physical uncertainty in participants' mental simulations, our ISE sampled funnel heights and friction angles from noisy distributions. Gaussian noise ( 0 mean, $\sigma_{H}^{2}$ variance) was added to the ground-truth funnel height in each situation. Gaussian noise was also added to the ground-truth friction angle $\theta_{i T}$, but in logarithmic space (see Sanborn et al., 2013): $\theta_{i}=f^{-1}\left(f\left(\theta_{i T}\right)+\varepsilon\right)$, where $\theta_{i T}$ is the ground truth value of the initial sand height, $f\left(\theta_{i T}\right)=\log \left(\omega \cdot \theta_{i T}+k\right)$, and $\varepsilon$ represents Gaussian noise with 0 mean and $\sigma_{\varepsilon}^{2}$ variance. The results reported herein used the following model parameters: $\sigma_{H}=0.12 H_{i T}, \sigma_{\varepsilon}=0.6, \omega=0.8$ and $k=1.5$.
Data-Driven Predictions: To predict human judgments, both GLM and XGBoost were tested on the $i$ th pile ( $i=$ $1,2,3,4)$ and trained on the remaining three piles. During training, 10,000 samples were drawn for each remaining pile ( 30,000 samples) and passed to our MPM simulator. Samples were generated using the sampling method described in the previous section. After training on the 30,000 samples, both data-driven models were tested on another 10,000 samples generated from noisy input based on the configuration of pile $i$. The final distribution was formed by aggregating the predictions across the 10,000 samples.

Table 1: Root-mean-square deviation (RMSD) values for the ground-truth (GT), ISE, GLM, and XGBoost models for Experiments 1 and 2. Lower values of RMSD indicate better model fits.

|  | GT | ISE | XGBoost | GLM |
| :--- | :---: | :---: | :---: | :---: |
| Experiment 1 (Static) | 0.458 | 0.101 | 0.267 | 0.171 |
| Experiment 1 (Dynamic) | 0.445 | 0.104 | 0.237 | 0.148 |
| Experiment 2 (Liquid) | 0.145 | 0.081 | 1.382 | 0.077 |
| Experiment 2 (Sand) | 0.170 | 0.080 | 1.422 | 0.120 |
| Experiment 2 (Balls) | 0.186 | 0.102 | 2.067 | 0.191 |

Model Comparisons: Fig. 2 depicts the predictions of the ISE, XGBoost, and GLM models compared to human judgments. All four models achieved high correlations with human performance (Static: $r(12)=0.91,0.84$, and 0.27 ; Dynamic: $r(12)=0.88,0.88$, and 0.30 for ISE, XGBoost, and GLM, respectively). Human performance was much less correlated with ground-truth predictions (Static: $r(12)=0.17$; Dynamic: $r(12)=0.19$ ). The ISE model predictions were more correlated with the human data than the competing datadriven model predictions in the Static condition but were only slightly more correlated than XGBoost predictions in the Dynamic condition. Hence, this paper uses The root-mean-square deviation (RMSD) between human responses and model results to compare the model fits. We found that RMSD between human responses and ISE predictions for the 4 judgment trials was less than that between ground-truth pre-
dictions in both Static and Dynamic Conditions (see Table 1). We also examined modeling performance using the Bayesian information criterion (BIC) to account for the different number of free parameters in each model. We found that the ISE provides a better fit to the human data than the groundtruth and data-driven models in both conditions. For groundtruth, ISE, XGBoost, and GLM models, Static BIC $=-25.0$, $-62.3,-31.2,-45.4$, and Dynamic BIC $=-25.9,-61.3$, $-35.0,-50.0$, respectively. The model with the lowest BIC value is preferred.

Although XGBoost captures most of the trends in the human judgments, it appears to over-fit the data in some cases. In the Static Condition, XGBoost's predicted response proportion for Pile 1 in the Trial $1(1 / 2 \mathrm{~cm}$ funnel height) is greater than the proportion in Trial $2(1 \mathrm{~cm}$ funnel height), which is consistent with human judgments. In the Dynamic Condition, however, XGBoost's predicted response proportion for Pile 1 is greater in Trial 1 than in Trial 2, which is inconsistent with trends in human performance. Alternatively, GLM showed very poor performance, predicting an increasing probability of Pile 1 choices for larger funnel heights. This trend is in the opposite direction of that observed in the human data, most likely due to the small number of training trials used to make each prediction.

## Experiment 2

Our results from the first experiment indicate that humans are able to predict the resting geometry of sand piles, even though they may not have very rich experience interacting with sand in daily-life. The second experiment was conducted to determine 1) whether humans can reason about complex interactions between sand and rigid obstacles and 2) whether their predictions about the resting state of sand in novel situations differ from predictions about other substances, such as liquid and rigid balls.

## Participants

A total of 90 undergraduate students ( 66 females), mean age 20.9, were recruited from the UCLA Department of Psychology subject pool, and were compensated with course credit.

## Materials and Procedure

The procedure in Experiment 2 was similar to the design in Bates et al.'s (2015) experiment: i.e., participants viewed a volume of a substance suspended in the air above obstacles and were asked to predict the proportion that would fall into two basins separated by a vertical divider below (Fig. 3). The present experiment differed from previous work in that participants reasoned about the resting state of one of three different substances: liquid, sand, or sets of rigid balls. Also, whereas the previous study used polygonal obstacles, those in the present study were circles varying in size. Depth information was also not present in the rendered situations. The stimulus videos can be viewed at https://vimeo.com/216585992.

Situations were generated by sampling between 2 and 5 obstacle locations from a uniform distribution bounded by the width and height of the chamber. The diameter, $d$, of each obstacle was sampled from a uniform distribution bounded by [ $0.15,0.85]$ relative to the randomly-generated center points.


Figure 3: Initial (top) and final (bottom) state of liquid (left), sand (middle), and a set of rigid balls (right) for a testing trial in Experiment 2 with 5 obstacles. The percentages indicate the amount of each substance that fell into the left and right basins. Only the initial state of each substance was shown in the testing trials.

The center points were generated by uniformly sampling the entire space. If the generated obstacles were placed outside the boundary, the configuration was rejected and re-sampled. Our MPM simulator was used to determine the ground-truth proportion of each substance in the left and right basins for each of the generated situations. For each substance, forty testing trials ( 10 trials with $2,3,4$, and 5 obstacles) were chosen from the generated set such that the ground-truth proportion of substance in the left basin was approximately uniform across trials. The testing trials were the same for each substance.

Participants were randomly assigned to either the liquid, sand, or rigid balls condition. Thirty participants were assigned to each condition in a between-subjects experimental design. Prior to the testing trials, participants completed five practice trials with two obstacles in each situation in a randomized order. After answering 1) which basin the majority of the substance would fall into and 2) the expected proportion that would fall into the indicated basin, participants viewed a video (13 second duration) of the situation unfolding and were told the resulting proportion in the ground-truth simulation. After completing the practice trials, participants completed 40 testing trials in a randomized order by answering the same two questions in each trial. No feedback was given following the completion of each testing trial.

## Human Results

Participants' predicted proportions in the testing trials were strongly correlated with ground-truth predictions in the liquid, sand, and rigid balls conditions $(r(38)=0.86,0.82$, and $0.88 ;$ RMSD $=0.145,0.170,0.186$, respectively). The deviation for each trial was calculated by subtracting the groundtruth proportion from each participant's proportion response. The deviation differed significantly between the three substance conditions ( $F(2)=3.64, p=0.03$ ), indicating that the difference between human predictions and the groundtruth status varied according to the substance type. To determine whether participants' response proportions differed between substances, a random factor ANOVA was conducted for a chosen set of trials. The chosen set excluded those trials where the majority of each substance fell into the same basin


Figure 4: Model prediction results compared to human predictions. From left to right: Ground-truth (GT), ISE, GLM, and XGBoost.
(left or right) according to the ground-truth simulation. We found that the response proportions showed significant differences depending on substance type $(F(2)=8.43, p<0.01$ ). The next section examines whether an ISE and two datadriven models can capture differences in human performance between the three substances.

## Model Results

ISE Predictions: In Experiment 2, the observable input variables for our ISE for each substance were 1) the initial, horizontal position of the substance, and 2 ) the positions of the circular obstacles in each situation. The latent substance attributes accepted by the engine were viscosity, friction angle, and restitution coefficient for liquid, sand, and the rigid balls, respectively. Gaussian noise was added to the substance's (ground-truth) horizontal position ( 0 mean, 0.35 variance) and the obstacles' (ground-truth) positions in 2D space ( 0 mean, 0.4 variance). Logarithmic Gaussian noise was added to each substance's ground-truth attribute value via the logarithmic transformation specified in Experiment 1. The results reported here utilized the following model parameters for all three substances: $\sigma_{\varepsilon}=0.5, \omega=0.8, k=1.2$. Two thousand samples ( 40 situations $\times 50$ noisy samples) were used for each substance.
Data-Driven Predictions: Similar to Experiment 1, both GLM and XGBoost were tested. The training data were randomly generated situations with basin proportions calculated using resting state output from our MPM simulator. Input features were the collection of both the observable input variables and latent substance attributes used in the ISE prediction. In total, 6000 samples were used for training.
Model Comparisons: Fig. 4 depicts the comparison between human and model basin predictions from the groundtruth (GT), ISE, GLM, and XGBoost models, and Table 1 depicts the root-mean-square deviation (RMSD) of each model's predictions from human ones. The human data were highly consistent with ISE predictions $(r(38)=0.93,0.93$, 0.93 ; $\mathrm{RMSD}=0.081,0.080,0.102$ for liquid, sand, and rigid balls, respectively). The ISE model predictions deviated from the human data to a lesser degree than the GT model predictions $(r(38)=0.87,0.85,0.88 ;$ RMSD $=0.145,0.170,0.186$ for liquid, sand, and rigid balls, respectively), indicating a superior account of human predictions across a range of substances. In comparison, GLM and XGBoost predictions were
less consistent with human predictions (GLM: $r(38)=0.77$, $0.78,0.65$, RMSD $=0.077,0.120,0.191$; XGBoost: $r(38)=$ $0.67,0.74,0.71, \operatorname{RMSD}=1.382,1.422,2.067$ for liquid, sand and rigid balls, respectively). As in the previous experiment, we compared each model's BIC measure in each condition to account for the number of free parameters in each model. We found that the BIC values for the ground-truth, GLM, and XGBoost models (GT: BIC $=-154.5,-141.8$, -134.6 ; GLM: BIC $=-194.0,-158.6,-121.4$; XGBoost: BIC $=36.9,39.2$, 69.2 for liquid, sand, and rigid balls, respectively) were consistently greater than the values for the ISE model (BIC $=-190.0,-191.0,-171.6$ for liquid, sand, and rigid balls, respectively), further reinforcing the superior performance of our simulation-based model.

It is worth noting that our ISE achieved consistent performance across all three substances, whereas GLM and XGBoost were less capable of predicting human judgments about rigid balls and liquid. In addition, our ISE used only one third of the training samples that XGBoost and GLM needed, demonstrating that a generative physical model with noisy perceptual inputs is capable of learning with a smaller number of samples than data-driven methods.

## Discussion

Results from Experiments 1 and 2 provide converging evidence that humans can predict outcomes of novel physical situations by propagating approximate spatial representations forward in time using mental simulation. This stands in contrast to early research in rigid-body collisions suggesting that human physical predictions do not obey ground-truth physics, instead relying on heuristics (e.g., Gilden \& Proffitt, 1994; Runeson, Juslin, \& Olsson, 2000). ISE predictions entailing the noisy Newton framework outperformed both ground-truth and data-driven models in both experiments, further confirming the role of perceptual noise and physical dynamics in human intuitive physical predictions.

Previous work has demonstrated that humans spontaneously employ mental simulation strategies when reasoning about novel physical situations (Clement, 1994; Hegarty, 2004; Schwartz \& Black, 1996). Recent fMRI results suggest that intuitive physical inferences are made using an internal physics engine encoded in the brain's "multiple demand" network (Fischer, Mikhael, Tenenbaum, \& Kanwisher, 2016). Although our ISE employed herein accounted for perceptual uncertainty in each situation, the simulations themselves closely approximated normative physical principles. Adding "stochastic noise" to physical dynamics, however, has been shown to increase model performance when predicting human responses in simple physical situations (K. A. Smith \& Vul, 2013). While dynamic uncertainty can easily be built into rigid-body collisions, employing this strategy in the present physical simulations would preclude stable numerical evaluation. Thus, future computational work should explore methods for adding dynamic uncertainty into complex physical simulations while preserving their accuracy and stability.

Results from the present study demonstrate that human predictions about substance dynamics can be accurately predicted by a unified simulation method with uncertainty im-
plemented into underlying physical variables. It is unlikely, however, that the human brain numerically evaluates partial differential equations to discern whether physical quantities (e.g., mass and momentum) are conserved, nor is it likely that the brain stores the locations of vast numbers of particles to form physical predictions and judgments. Instead, our results provide evidence that humans approximate the dynamics of substances in a manner consistent with ground-truth physics but succumb to biases invoked by perceptual noise when inferring future environmental states. It remains unclear, however, whether the dynamics of rigid objects, liquids, and granular materials are approximated using separable mechanisms or a single cognitive architecture with different assumptions and constraints. The success of our unified simulation model across different substance-types supports the latter perspective.
Acknowledgments Support for the present study was provided by a NSF Graduate Research Fellowship, NSF grant BCS-1353391, DARPA XAI grant N66001-17-2-4029, DARPA SIMPLEX grant N66001-15-C-4035, ONR MURI grant N00014-16-1-2007, and DoD CASIT grant W81XWH-15-1-0147.

## Appendix: Details of Our MPM Simulator

The governing partial differential equations utilize the principles of conservation of mass and momentum:

$$
\begin{equation*}
\frac{D \rho}{D t}+\rho \nabla \cdot \mathbf{v}=0, \quad \frac{D \mathbf{v}}{D t}=\nabla \cdot \sigma+\rho \mathbf{g}, \tag{1}
\end{equation*}
$$

where $\sigma$ is the stress imparted on a particle, $\mathbf{g}$ is the gravitational acceleration, and $\frac{D}{D t}$ is the material derivative with respect to time. The equations are discretized spatially and temporally with a collection of Lagrangian particles (or material points) and a background Eulerian grid. The material type of the simulated substances is naturally specified from the constitutive model, which defines how a material exerts internal stress (or forces) as a result of deformation.

Rigid balls are simulated as highly stiff elastic objects with the neo-Hookean hyperelasticity model, described through the elastic energy density function

$$
\begin{equation*}
\Psi(\mathbf{F})=\frac{\mu}{2}\left(\operatorname{tr}\left(\mathbf{F}^{T} \mathbf{F}\right)-d\right)-\mu \log (J)+\frac{\lambda}{2} \log ^{2}(J), \tag{2}
\end{equation*}
$$

where $d$ is the dimension ( 2 or 3 ), $\mathbf{F}$ is the deformation gradient (i.e., the gradient of the deformation from undeformed space to deformed space), $J$ is the determinant of $\mathbf{F}$, and $\mu$ and $\lambda$ are Lamé parameters that describe the material's stiffness.

Liquid is modeled as a nearly incompressible fluid, with its state governed by the Tait equation (Batchelor, 2000):

$$
\begin{equation*}
p=k\left[\left(\frac{\rho_{0}}{\rho}\right)^{\gamma}-1\right] \tag{3}
\end{equation*}
$$

where $p$ is the pressure, $\rho$ and $\rho_{0}$ are the current and original densities of the particles, $\gamma=7$ for water, and $k$ is the bulk modulus (i.e., how incompressible the fluid is). Through this Equation-of-State (EOS), the stress inside a non-viscous fluid is given by $\sigma=-p \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We further adopt the Affine Particle-In-Cell method (APIC) (Jiang et al., 2015) to greatly reduce numerical error and artificial damping. This enables us to simulate fluids with better accuracy compared to alternative computer graphics methods.

The motion of dry sand is largely determined by the frictional contact between grains. In the theory of elastoplasticity, the modeling of large deformation (e.g., frictional contact) can be based on a constitutive law that follows the Mohr-Coulomb friction theory. Following (Klár et al., 2016), we simulate dry sand based on the Saint Venant Kirchhoff (StVK) elasticity model combined with a Drucker-Prager non-associated flow rule. Plasticity models the material response as a constraint projection problem, where the feasible
region (or yield surface) of the final material stress is restricted to be inside

$$
\begin{equation*}
\operatorname{tr}(\sigma) c_{F}+\left\|\sigma-\frac{\operatorname{tr}(\sigma)}{d}\right\|_{F} \leq 0 \tag{4}
\end{equation*}
$$

where $d$ is the dimension and $c_{F}$ is the coefficient of internal friction between sand grains. The stress (and thus deformation gradient) of each sand particle is projected onto the yield surface so as to satisfy the second law of thermodynamics.

## References

Batchelor, G. K. (2000). An introduction to fluid dynamics. Cambridge university press.
Bates, C. J., Yildirim, I., Tenenbaum, J. B., \& Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. In Proceedings of the 37th annual conference of the cognitive science society.
Battaglia, P. W., Hamrick, J. B., \& Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45), 18327-18332.
Chen, T., \& Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving.
Fischer, J., Mikhael, J. G., Tenenbaum, J. B., \& Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. Proceedings of the national academy of sciences, 113(34), E5072-E5081.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In Proceedings of the 37th annual conference of the cognitive science society.
Gilden, D. L., \& Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. Perception \& Psychophysics, 56(6), 708-720.
Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., \& Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. Cognition, 157, 61-76.
Hegarty, M. (2004). Mechanical reasoning by mental simulation. Trends in cognitive sciences, 8(6), 280-285.
Jiang, C., Schroeder, C., Selle, A., Teran, J., \& Stomakhin, A. (2015). The affine particle-in-cell method. ACM Transactions on Graphics (TOG), 34(4), 51.
Jiang, C., Schroeder, C., Teran, J., Stomakhin, A., \& Selle, A. (2016). The material point method for simulating continuum materials. In Acm siggraph 2016 course (pp. 24:1-24:52).
Klár, G., Gast, T., Pradhana, A., Fu, C., Schroeder, C., Jiang, C., \& Teran, J. (2016). Drucker-prager elastoplasticity for sand animation. ACM Trans Graph, 35(4), 103:1-103:12.
Kubricht, J. R., Jian, C., Zhu, Y., Zhu, S. C., Terzopoulos, D., \& $\mathrm{Lu}, \mathrm{H}$. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In Proceedings of the 38th annual conference of the cognitive science society.
McCullagh, P. (1984). Generalized linear models. European Journal of Operational Research, 16(3), 285-292.
Monaghan, J. J. (1992). Smoothed particle hydrodynamics. Annual review of astronomy and astrophysics, 30, 543-574.
Runeson, S., Juslin, P., \& Olsson, H. (2000). Visual perception of dynamic properties: cue heuristics versus direct-perceptual competence. Psychological Review, 107(3), 525-555.
Sanborn, A. N. (2014). Testing bayesian and heuristic predictions of mass judgments of colliding objects. Frontiers in psychology, 5.

Sanborn, A. N., Mansinghka, V. K., \& Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. Psychological review, 120(2), 411.
Schwartz, D. L., \& Black, J. B. (1996). Analog imagery in mental model reasoning: Depictive models. Cognitive Psychology, 30(2), 154-219.
Smith, K., Battaglia, P., \& Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In Proceedings of the 35th conference of the cognitive science society (pp. 3426-3431).
Smith, K. A., \& Vul, E. (2013). Sources of uncertainty in intuitive physics. Topics in cognitive science, 5(1), 185-199.
Sulsky, D., Zhou, S., \& Schreyer, H. (1995). Application of a particle-in-cell method to solid mechanics. Comp Phys Comm, 87(1), 236-252.

# Synchronization Assessment for Collective Behavior 

Maurici A. López-Felip (maurici.lopez_felip@uconn.edu)<br>Center for the Ecological Study of Perception and Action<br>Department of Psychological Sciences, 406 Babbidge Road, Unit 1020<br>Storrs, CT 06269, USA<br>Tehran J. Davis (tehran.davis@uconn.edu)<br>Center for the Ecological Study of Perception and Action<br>Department of Psychological Sciences, 406 Babbidge Road, Unit 1020<br>Storrs, CT 06269, USA<br>Till D. Frank (till.frank@uconn.edu)<br>Center for the Ecological Study of Perception and Action<br>Department of Psychological Sciences, 406 Babbidge Road, Unit 1020<br>Storrs, CT 06269, USA


#### Abstract

Team cognition can be defined as the ability that humans have to coordinate with others through a complex environment. Sports offer exquisite examples of this dynamic interplay requiring decision making and other perceptual-cognitive skills to adjust individual decisions to the team selforganization and vice-versa. Considering players of a team as periodic phase oscillators, synchrony analyses can be used to model the coordination of a team. Nonetheless, a main limitation of current models is that collective behavior is context independent. In other words, players of a team can be highly synchronized without this corresponding to a meaningful coordination dynamics relevant to the context of the game. Considering these issues, the aim of this study was to develop a method of analysis sensitive to the context for evidence-based measures of team cognition.


Keywords: Team Cognition; Synchronization; Ecological Dynamics;

## Introduction

Central to the definition of a team are the interactions amongst its components (McNeese, Cooke, Fedel \& Gray, 2016). When players cooperate together as a team, the resulting collective behaviors rarely are expressed in terms of the simple summation of the individuals' activities. That is, the team's activity emerges from the coordination of actions and often nonlinear interactions of its players. For example, to be successful in European football (soccer), players must coordinate their actions with others across many different spatial and temporal scales. While recent research has focused on elucidating the mechanisms that facilitate such large-scale coordination, the identification of the fundamental, self-organizing principles that underlie team dynamics remains an unresolved matter (see e.g., Memmert, Lemmink \& Sampaio, 2016; Folgado, Duarte, Fernandes \& Sampaio, 2014). Indeed, techniques to measure collective
emergent behavior are still in an early stage of development (Araujo, Silva \& Ramos, 2014), while many attempts to measure team work have typically focused on measuring outcome performance rather than team dynamics. However, recent attempts to study the dynamics of multi-agent activity have benefitted from concepts and tools from Dynamical Systems Theory (DST) (e.g., Duarte, Araújo, Correia, Davids, Marques, \& Richardson, 2013). While DST provides suitable techniques for modeling living systems, it makes no direct claims about their status nor provides a theoretical basis for understanding goal directed behavior. Amongst the broad range of DST tools, one of the most common approaches used by students of perception, action and cognition is the study of synchronization.

## Measuring Synchronization

Measures of synchrony are used for describing phenomena that obey recurrent, dynamical laws; and have been applied for a wide range of phenomena coming from substantially different fields of study as natural sciences, engineering or even social life (Pikovsky, Rosenblum \& Kurths, 2001).

Whereas in physical, nonliving systems synchrony is often mediated via mechanical coupling (e.g., Huygens famous observations regarding the synchronization of two clock pendulums, (1673/1986), psychological and social systems often synchronize via informational (e.g., visual) coupling (Schmidt, Carello \& Turvey, 1990). Although most research on the synchrony and coupling between actors has focused on dyads, a recently developed method, cluster phase analysis (CPA, Frank \& Richardson, 2010), has been used to extend synchrony measures to groups larger than two people. CPA has been used, for example, to assess the degree to which a group of people successfully synchronized their intentional, oscillatory rhythmic movements with rocking chairs; with synchrony measured using an adaptation of the Kuramoto
order parameter ( $a k a$ cluster amplitude, $\dot{r}$ where high synchronization $=1$ ). Similar methods have been used to characterize teams' phase synchrony in football (see e.g., Duarte et al., 2013; Duarte, Travassos, Araujo \& Richardson, 2014). Here, separate measures of team synchrony are derived using players' displacements along either the latitudinal or longitudinal axis, where a common result is that synchrony is higher in longitudinal displacements than in lateral displacements (Duarte, et al., 2013). Using this method, researchers have also noted that the observed degree of synchrony was not subject to possession of the ball (see e.g., Pinto, 2014; Duarte, et al. 2013), presumably one of the key factors of team organization during the match. However, it may be argued that the technical aspects of this methodological approach do not consider relevant contextual features of the game that are key to self-organizing principles in team sports. This lack of situational context is a consequence of 1) the behavioral variable submitted to the model and 2) the constraints that presents the mathematical model. Behavior is measured in a time-series of displacements along one dimension; however, the Kuramoto model requires phase angles as its input. To overcome this incompatibility, the displacement time-series are transformed to instantaneous phase angles by using the Hilbert Transform (see Pikovsky, et al. 2001 for details). However, this method is limited in that high synchrony can be a consequence of the players simply being very close to each other within that onedimensional space (e.g., $x$-dimension), whereas, conversely if players are far apart within that dimension, synchrony would be low.

Considering these issues (technical and contextual) we aimed to further extend the use of CPA by using insights from a recently developed framework that applies the ecologicaldynamics approach to perception and action in football (López-Felip, 2014). Our model parameters were defined by two situational variables of the game: such as players' orientation-to and distance-from the goal of interest (i.e. the goal being actively attacked by the offense and defended by the defense). Our main hypothesis was that when accounting for these two contextual variables, team synchrony would be dependent on ball possession. This result would suggest the need for further exploring context dependent analyses for evidence-based measures of team cognition.

## Method

## Participants

Twenty-two male elite football players from two European clubs played a friendly game during the pre-season 20162017. Participants ranged in age from 17 years to 34 years (average $26.5 \pm 0.4$ years). At the time of data collection, neither of the teams had initiated their regular competitions, however, the away team was a member of what is typically considered to be a superior league. The entire first half of the match was registered with no injuries or substitutions.

## Instruments

Player position data were collected via GPS (sampling rate of 15 Hz ) for an entire half of forty-five minutes plus extra time. These GPS monitors could reliably capture positional raw data (2D) based on the latitude and longitude positions of all players throughout the match.

## Procedures

The positional raw data were subsequently matched to corresponding events throughout the competition (captured via video). This allowed us to asses when during each timeseries a team was in possession of the ball (i.e. attacking role), as well as identify any prolonged periods of stoppage
(e.g., from injury assistance, goals, etc.) to remove from analysis. The criteria for determining ball possession was based on Reis, Duarte, Araújo, Folgado, \& Frias (2013).

## Data Analysis

Starting with the positional raw data, exocentric coordinates were used to define the state space in which trajectories of players were captured. Then, the goals were represented as specific variables of this state space to create a new variable, angle of the direction to the goal $\left(\theta_{\mathrm{g}}\right)$ :

$$
\begin{equation*}
\theta_{\mathrm{g}}=\tan ^{-1}\left(\frac{\mathrm{x}_{\text {goal }}-\mathrm{x}_{\text {player }} \mathrm{i}}{\mathrm{y}_{\text {goal }}-\mathrm{y}_{\text {player }}}\right) \tag{Eq.1}
\end{equation*}
$$

This measure provided a metric of each player's orientation with respect to the goal. Relative angles were submitted to CPA, creating a time-series of Kuramoto parameter values describing each team's synchrony at every time step.
(Eq. 2)

$$
\dot{r}^{\prime}\left(t_{i}\right)=\left|\frac{1}{n} \sum_{k=1}^{n} \exp \left(i \theta_{k}\left(t_{i}\right)\right)\right|
$$

To account for the distance of a team to the goal of interest, each team's center of mass was assessed at each time step. Distance of the center of mass ( $d_{C O M}$ ) was measured as the mean longitudinal position of all team members over time.
(Eq. 3)

$$
C O M=\frac{1}{N} \sum_{\text {player }=1}^{n} x_{\text {player }}
$$

To simplify our analysis, $d_{\text {COM }}$ values were categorized into four quadrants each spanning 25 m ; where Q1 contained distances closest to the goal of interest and Q4 contained distances furthest away (see Figure 1).


Figure 1: Black disc represents the attacking team and black triangles represent the defending team. Dashed lines represent each player's goal angle. Vertical black lines divide the field in 4 equidistant quadrants. Q1 is the quadrant closest to the active goal and Q 4 the furthest. Then, $x-y$ axes represent the longitude and latitude coordinates from where positional raw data were collected.

Then, provided our research question, we assessed changes in each team's synchrony as 1 ) a function of ball possession (whether teams were attacking or defending) and 2) the distance between the team's center of mass to the goal of interest, $d_{\text {Сом }}$. To do so, each point in the time series of Kuramoto parameter values was independently evaluated as a function of the corresponding $d_{\text {СОМ }}$ quadrant. The resulting mean values for each quadrant were submitted to further analysis, resulting in values reported in Figure 2.

## Results

As determined using the cluster amplitude analysis the overall degree of synchronization of teams were between 0.55 and 0.99 . When phase synchrony was assessed for each team as a function of the playing role in the game (i.e., attacking or defending) and the $d_{C O M}$ to the goal of interest, synchronization differed. Figure 2 shows how the mean tendency of synchrony varies as a context of where and when a team is attacking or defending. That is, mean synchrony decays as teams' $d_{\text {Сом }}$ approaches Q 1 .

Table 1: Mean synchrony of each team as a function of playing role and $d_{\text {COM }}$.

| Team | Quadrants | $\dot{r}=$ Attacking | $\dot{r}=$ Defending |
| :---: | :---: | :---: | :---: |
| Home | Q1 | $N A$ | 0.86 |
| Away | Q1 | 0.84 | 0.89 |
| Home | Q2 | 0.92 | 0.95 |
| Away | Q2 | 0.93 | 0.97 |
| Home | Q3 | 0.97 | 0.98 |
| Away | Q3 | 0.97 | 0.98 |
| Home | Q4 | 0.99 | $N A$ |
| Away | Q4 | 0.99 | 0.99 |

Overall, synchrony increased as the teams moved farther from the goal of interest. At the same time differences in synchrony depending on the team's role became more pronounced as the teams moved closer to the goal of interest. A similar pattern of effects [role: $F(1,15488)=31, p<.001$; quadrant: $F(3,15488)=6484, p<.001$; role $\times$ quadrant: $F(3$, $15488)=622, p<.001$ ] was observed for the home team (note that the $d_{\text {СОм }}$ of the home team never entered Q1 when attacking or Q4 when defending; likely due to the away teams dominance of the match).


Figure 2: Mean synchrony of each team based on field location and role.

## Discussion

The purpose of this study was to develop a method of analyzing team coordination that is sensitive to the context in which team actions unfold over the course of a game. It was hypothesized that by using a measure of phase synchrony sensitive to the contextual circumstances of ball possession, synchrony of a team would change. To test this, two variables that were relevant to the context of the game such as footballer's angle relative to the direction of the active goal and COM of each team on the field were used.

With respect to the experimental hypothesis, a linear effects mixed model showed that team synchrony is dependent on team role and distance from the active goal. Significant effects were found for role and quadrant, qualified by a role $\times$ quadrant interaction in both teams.

Measures of team synchrony showed higher mean values when a team was in defense. These data suggest that individuals tend to coordinate their movements together relative to the goal more in those instances in which they are defending, than the ones in which they are attacking. At the same time, lower synchrony values were found in those instances in which a team was closer to the opponent's goal. This is not surprising for the attacking team, because behavior of a football team when attacking is to spread out and create as many open spaces as possible to the opponent team. Interestingly, the team in defense showed also low values of synchronization in Q 1 . One possibility is that this
may be due to the driving-driver effect (Step \& Turvey, 2010). According to this, the team in defense would try to anticipate the actions of the team in possession of the ball, reflected in the drop of mean synchrony of the defense team in Q1. This conjecture remains an open question.

Although these data showed differences of synchrony in terms of ball possession, the levels of whole team synchrony were, overall, high. All the mean values of cluster amplitude for the angle to the direction of the active goal ranged between 0.84 and 0.99 . These values are similar to those found in football (Duarte et al., 2013) or in intentional oscillatory rhythmic movements of rocking chairs (Frank \& Richardson, 2010).

## Implications for Measuring Synchronization

Based on the approach of previous studies, the present work assessed synchrony by means of an adaptation of the Kuramoto Order Parameter. As explained in the introduction section, when using time-series of displacements in the $x, y$ or $z$ axes to assess synchronization, there is the need to calculate the instantaneous phase angle of the time-series (usually done by the Hilbert Transform). By following these steps, synchrony may remain high and unchanged due to the limitations of the methodology as explained earlier in the introduction.

Hence, the present work, approaches the assessment of synchrony via an alternative methodology. First, we considered that we could explore the possibility of using an angle that was not limited to a one-dimensional plane. Simply because representing dynamics of collective behavior at one dimension did not seem to lead us to our purposes (i.e., provide contextual meaning to assessments of collective behavior). Second, based on previous research, displacements from a time-series have not been able to discriminate between synchrony levels and ball possession during the game. Thus, our approach attempted to link a behavioral variable to the final target that a team aims (i.e., scoring a goal). For example, in models of steering and obstacle avoidance (see e.g., Fajen \& Warren, 2003; Warren, 2006), one of the variables taken in their assessments is the relative angle of the performer's position relative to the goal or obstacle. Here, using a similar variable and clustering the angle of each player relative to the active goal, allowed us to model team dynamics at a 2-dimensional plane and do it relative to the final purpose of the game.

Taking this approach to using the Kuramoto, is not a final model. This is just a preliminary step towards developing a more robust model of synchronization in collective behavior that aims to be sensitive to the context in which team activity occurs.

## Conclusions

This study investigated the degree to which ball possession impacts team synchrony as a function of the team's $d_{\text {COM }}$. López-Felip and Porter (2015) argued that both variables were proposed as proper parameters to include when
modeling football team behavior as a dynamical system. Our finding suggest that appropriately modeling team behavior must take into account variables that capture the meaningful current state of affairs of the game-such as players' orientation and location relative to key points of interest. Future research in this domain should seek to identify and incorporate additional, meaningful aspects (e.g., tactics) to addressing team coordination.

More broadly, these findings may be understood in the claim that efforts to model living systems and their actions should account for context. Understanding the functional, context dependent relationship that exists between organismenvironment and situation could serve to guide and constrain future dynamical analyses and mathematical modeling of team systems (Turvey, 1992; Turvey \& Shaw, 1995).

## References

Araujo, D., Silva, P., \& Ramos, J. P. (2014). Affordancebased decisions guide team synergies during match performance. Research in Physical Education, Sport and Health, 19-26.
Duarte, R., Araújo, D., Correia, V., Davids, K., Marques, P., \& Richardson, M. J. (2013). Competing together: Assessing the dynamics of team-team and player-team synchrony in professional association football. Human Movement Science, 32, 555-566.
Duarte, Travassos, Araujo \& Richardson (2014). The influence of manipulating the defensive playing method on collective synchrony of football teams. In Proceedings of the IX World Congress on Performance Analysis of Sport, Publisher: Routledge, (Eds.), D. Peters, \& P. O'Donoughue
http://doi.org/10.1016/j.humov.2013.01.011
Fajen, B. R., \& Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection. Journal of Experimental Psychology: Human Perception and Performance, 29, 343-362.
http://doi.org/10.1037/0096-1523.29.2.343
Folgado, H., Duarte, R., Fernandes, O., \& Sampaio, J. (2014). Competing with Lower Level Opponents Decreases Intra-Team Movement Synchronization and Time-Motion Demands during Pre-Season Soccer Matches. PLOS ONE, 9, e97145-9. http://doi.org/10.1371/journal.pone. 0097145
Frank, T. D., \& Richardson, M. J. (2010). On a test statistic for the Kuramoto order parameter of synchronization: An illustration for group synchronization during rocking chairs. Physica D, 239, 2084-2092. http://doi.org/10.1016/j.physd.2010.07.015
Huygens, C. (1673). Horologium Oscillatorium. Apud F. Muguet, Paris, France,. English translation: The
Pendulum Clock, Iowa State University Press, Ames, 1986.

López-Felip, M. A. (2014). A scale to measure the complexity and perceptual-cognitive skills in soccer. MSc Thesis, Southern Illinois University.

López-Felip, \& M.A., Porter, J. (2015). An assessment of complexity and perceptual-cognitive skills in soccer. In Favero, T., Drust, B., and Dawson, B. (Eds.), International Research in Science and Soccer II. New York, NY: Taylor \& Francis Group.
McNeese, N., Cooke, N.J., Fedele, M., \& Gray, R. (2016). Perspectives on Team Cognition and Team Sports. In Raab, M. Wylleman, P., Seiler, R., Elbe, A.M., and Hatzigeordiadis, A. (Eds.), Sports and Exercise Psychology Research: From Theory to Practice. American Press.
Memmert, D., Lemmink, K. A. P. M., \& Sampaio, J. (2016). Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. Sports Medicine, 47, 1-10. http://doi.org/10.1007/s40279-016-0562-5
Mohammed, S., Ferzandi, L., \& Hamilton, K. (2010). Metaphor no more: a 15-year review of the team mental model construct. Journal of Management, 36(4), 876910.

Pikovsky, A., Rosenblum, M., \& Kurths, J. (2001). Synchronization: A Universal Concept in Nonlinear Sciences. Cambridge University Press
Pinto, C. (2014). The emergence of team synchronization during the soccer match: understanding the influence of the level of opposition, game phase and field zone. Dissertação de Mestrado. Universidade de Lisboa. Faculdade de Motricidade humana.
Reis, M., Duarte, R., Araújo, D., Folgado, H., \& Frias, T. (2013). Spatial interaction tendencies of football players captured by Voronoi diagrams. Proceedings of the Mathematical Methods in Engineering International Conference. (pp. 279-287). Porto: Instituto Superior de Engenharia do Porto.
Silva, P., Garganta, J., Araujo, D., Davids, K., \& Aguiar, P. (2013). Shared knowledge of shared affordances? Insights from an ecological dynamics approach to team cognition in sports. Sports Medicine, 43, 765-772.
Schmidt, R.C., Carello, C. \& Turvey, M.T. (1990). Phase Transitions and Critical Fluctuations in the Visual Coordination of Rhythmic Movements Between People. Journal of Experimental Psychology: Human Perception and Performance, 16, 227-247.
Step, N. \& Turvey, M.T. (2010). On strong Anticipation. Cognitive Systems Research, 11, 148-164.
Turvey, M. T. (1992). Affordances and prospective control: An outline of the ontology. Ecological Psychology, 4, 173-187.
Turvey, M. T., \& Shaw., R. E. (1995). Toward an ecological physics and a physical psychology. In R. L. Solso, \& D. W. Massaro (Eds.), The Science of the Mind: 2001 and Beyond. New York: Oxford University Press.
Warren WH. (2006). The dynamics of perception and action. Psychological Review, 13, 358-389

# More Siblings Means Lower Input Quality in Early Language Development 

Catherine Laing<br>Duke University<br>Elika Bergelson<br>Duke University


#### Abstract

Previous research has suggested that first-born infants acquire words faster than their later-born peers (Berglund et al., 2005), but may have some disadvantages in other aspects of syntactic and socio-communicative development (e.g. Hoff, 2006). Here we analyzed infants' early lexical development alongside their caregiver input from 6-18 months, in relation to how many siblings they have. We find that having more siblings (rather than being first- or later-born) has a gradient and negative relationship with infants' language development. This affect appears to be manifested in caregiver input: across three different measures of input quality/quantity, disadvantages were found for infants with more siblings. Having a larger number of siblings diminished the quality of the input and led to slower overall lexical development. Implications for language development and learning within dyadic and multi-member contexts are discussed.


# Uncovering visual priors in spatial memory using serial reproduction 

Thomas A. Langlois* (thomas.langlois@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94720 USA<br>Nori Jacoby* (kj2413@Columbia.edu)<br>Center for Science and Society, Columbia University<br>New York, NY 10027 USA<br>Jordan Suchow (suchow@berkeley.edu)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94720 USA<br>*These authors contributed equally to this work


#### Abstract

Visual memory can be understood as an inferential process that combines noisy information about the world with knowledge drawn from experience. Biases can arise during encoding of information from the outside world into internal representations, or during retrieval. In this work, we use the method of serial reproduction, in which information is passed along a chain of participants who try to recreate what they observed. We apply this method to the study of visual perception in the context of spatial memory biases for the remembered position of dots inside different geometric shapes. We present the results of non-parametric kernel density estimation of the end result of serial reproduction to model visual biases. We confirm previous findings, and show that memory biases revealed with our method are often more intricate and complex than what had previously been reported, suggesting that serial reproduction can be effective for studying perceptual priors.


Keywords: Vision; spatial memory; inductive biases; serial reproduction; iterated learning;

## Introduction

Retrieving detailed visual information from memory requires efficient representations of often complex and noisy visual scenes. In Bayesian accounts of reconstruction from visual memory, the memory system integrates sensory information with knowledge acquired from previous experience ("priors"). Effective use of this information may reduce variability in visual memory and improve overall reconstruction accuracy (Weiss et al., 2002). Using priors is usually advantageous because they capture regularities in the structure of the world that are innate or observed over our lifetimes. However, this can lead to substantial biases during reconstruction. This is because prior information may deviate significantly from our observations, especially when a visual scene is unexpected given previous experience.

In many cases, visual priors are categorical (or prototypical), represented in memory as schematic or simplified objects (Huttenlocher et al., 1991). In one experimental paradigm that reveals categorical effects, participants are asked to remember the location of a dot presented within a circle or other bounding shape. After a brief presentation and a delay, participants reproduce the dot's location by placing
it in the recalled position (see Figure 1). Huttenlocher et al. (1991) found that participants tend to misplace dots toward a central (prototypical) location in each of the quadrants of a circle. Following these results, Wedell et al. (2007) tried to predict prototypical positions in spatial memory for dots presented inside a variety of geometric shapes (circle, square, triangle, vertical oval, horizontal oval, and pentagon). In the study, participants were shown thirty-two dots aligned along two concentric circles within each shape. A parametric model with four components (prototypes) was fitted to the remembered positions of the dots, confirming that visual memory of these shapes shows substantial categorical effects.

The approach to characterizing categorical biases used in Huttenlocher et al. (1991) has a number of limitationsspecifically, a relatively small number of to-be-remembered locations (32) and a weak measurement of the biases, resulting in limited resolution for capturing the locations of the categories. In addition, Wedell et al. (2007) used a parametric model with a fixed number of categories. The choice of the model, and the number of categories that were used were not fully justified, requiring certain a priori assumptions. Here, we propose to use a paradigm based on serial reproduction to characterize visual memory biases without needing to rely on parametric modeling and with substantially better resolution and accuracy.

## The method of serial reproduction

Serial reproduction has a long history in experimental psychology, where it has been used to study how various biases distort information when it is transmitted from person to person (Bartlett, 1932). Figure 2 shows a schematic illustration of the experimental paradigm: a participant views a stimulus, such as a dot presented within a bounding shape, and is then asked to reproduce the stimulus as accurately as possible from memory. Critically, the reproduction created by the first participant is used as the stimulus for the second participant, who is then asked to do the same. At each iteration, the reconstruction produced by the previous participant becomes


Figure 1: Illustration of prototype effects in memory for points in a circle. The red crosses represent prototypes, and the small points are typically misremembered as being closer to those prototypes.
the stimulus for the next participant to reproduce. Famous early results include the transformation of an owl-like Egyptian hieroglyph into a small cat after ten iterations of a serialreproduction drawing task (Bartlett, 1932). This result was interpreted in terms of inductive biases in memory: as veridical information from the input becomes degraded following successive iterations, the reconstruction of the ambiguous image is pulled towards a prototypical object with similar visual properties.

Serial-reproduction experiments have long been used to simulate phenomena in cultural transmission, evolutionary biology, anthropology, and cognitive science (Kirby et al., 2008; Claidière et al., 2014), but it wasn't until recently that a rational analysis of serial reproduction considered how information should change as it is transmitted along a chain of rational agents (Xu \& Griffiths, 2010). Under a rational analysis, reconstruction from memory is defined as the problem of inferring the most accurate state of the world from noisy data, such as an imperfect memory trace and perceptual noise during encoding of the image. This problem is modeled using the framework of Bayesian statistics. Previous experience is captured by a prior distribution over possible states (a hypothesis space of world states). A posterior is computed, based on the likelihood, which indicates the probability of observing that information, given some hypothesis about the true state of the world. Xu \& Griffiths (2010) examined the predictions of this Bayesian account of reconstruction from memory for serial reproduction. They found that serial reproduction by Bayesian agents defines a Markov chain with the following transition probabilities:

$$
p\left(x_{n+1} \mid x_{n}\right)=\int p\left(x_{n+1} \mid \mu\right) p\left(\mu \mid x_{n}\right) d \mu
$$

where $x$ is a noisy stimulus (such as an imperfect memory trace) and $\mu$ is the true state of the world that generated that stimulus (in this case, the veridical image that impinged on the visual system). This Markov chain captures the probability of a new stimulus $x_{n+1}$ being created as a reconstruction of a previously seen stimulus $x_{n}$ in each iteration in the serial reproduction chain, and has a stationary distribution, called the
prior predictive distribution, which defines the probability of observing a stimulus $x$ when $\mu$ is sampled from the prior:

$$
p(x)=\int p(x \mid \mu) p(\mu) d \mu
$$

This process approximates a Gibbs sampler for the joint distribution on $x$ and $\mu$ defined by multiplying $p(x \mid \mu)$ and $p(\mu)$. This finding is significant because it provides a mathematical formalism for describing the consequences of serial reproduction: assuming that participants share common inductive biases, the transmission chain will converge to a sample from their shared prior.

In this paper, we explore spatial memory priors in a task where participants were asked to remember the position of a small black dot inside a variety of geometric shapes. Operating under the assumption that people share the same inductive biases, or spatial memory priors, we show that serial reproduction appears to converge on these priors remarkably quickly, revealing patterns that are consistent with some established findings, although in many cases revealing new and intricate patterns that were previously unknown. Finally, we demonstrate the advantages of using a non-parametric kernel density estimation procedure to characterize the prior.

## Methods

## Participants

Participants were recruited online using Amazon Mechanical Turk. All gave informed consent. The experimental protocol was approved by The Committee for the Protection of Human Subjects (CPHS) at the University of California, Berkeley. Each experiment required approximately 70-100 participants. A total of 570 participants took part in Experiment 1 and an additional 590 took part in Experiment 2.

## Stimuli

All images were approximately $400 \times 400$ pixels in size. Each shape was a 6 -pixel-wide black outline over a white background. The sizes and colors of the backgrounds for the images were intended to ensure that the images would be clearly visible in any standard browser window (unlikely to become occluded), and such that the boundaries of the images would be invisible.

## Procedure

We carried out a series of serial reproduction experiments. Participants were presented with timed displays (a shape outline with a dot initialized somewhere within the boundaries of the shape), and were instructed to reproduce the exact location of the dot inside of the shape. Once complete, their response was sent to another worker (again, as a timed display), who was instructed to reconstruct this display from memory, and so on. A total of ten iterations were completed for each chain. See Figure 2 for a schematic diagram of the serial reproduction procedure.

Practice trials. Participants completed ten practice trials in order to become familiar with the user interface. During these practice trials, they were presented with a circle (a black outline of a circle over a white screen), with a dot initialized somewhere within its boundaries. This display was presented for 4000 ms , followed by a blank screen lasting 1000 ms . Next, the circle was presented without the dot and remained on the screen until the participant positioned the dot in the location that they remembered. As soon as the participant clicked, the dot appeared under the mouse cursor. Participants could reposition the dot as many times as they needed. Once done, they pressed a button to proceed to the next trial.

Experimental trials. Following the ten practice trials with the circle, there were ninety-five experimental trials with exactly one of the shapes. In Experiment 1, the shape could be a circle, triangle, square, vertical oval, horizontal oval, or a pentagon. In Experiment 2, the shape could be a regular polygon with more than five vertices. For each of the 95 experimental trials, the presentation time was reduced to 1000 ms . As with the practice trials, the position of the shape on the screen was randomized somewhere inside a larger canvas in order to control for participants resorting to tracking the position of the dot by trivially marking its absolute position on their computer screens. In addition, participants were given trial-by-trial feedback regarding their accuracy. If their responses were within eight percent of the width and height subtended by the shape on the screen, they were told that their response was accurate (a message in green font: "This was accurate"), and received a small monetary bonus. If not, they received no bonus beyond the basic payment for the HIT, in addition to any bonuses accrued from the previous trials, and were presented with a red message ("this was not accurate"). These trials were discarded from the experiment. Participants could not provide multiple responses within a chain.

Experiment 1: We used the same six shapes as Wedell et al. (2007): A circle, equilateral triangle, square, vertical and horizontal ovals, and a regular pentagon. For each shape, we initialized the position of five hundred dots within its boundaries (for the circle, we initialized four hundred dots).

Experiment 2: Because our method revealed a variable number of peaks (prototypes) in the prior for the angular shapes in Experiment 1, and that these appeared to be due in large part to the number of vertices in the polygons (all were regular polygons-an equilateral triangle, square, and pentagon), we wanted to determine the point at which the prototypes begin to merge into the four prototypes in the prior for the circle. We did this by conducting the same experiment with polygons containing increasingly more vertices (approximating a circle more closely as vertices were added).

## Results

Our results are presented in two parts. First, we present all our results for Experiment 1, in which we used the same shapes as Wedell et al. (2007). We demonstrate that using a serial reproduction paradigm, as well as non-parametric kernel density


Figure 2: Serial reproduction chain for one trial in the memory task, illustrating the iterative process for a single dot being remembered. The trial in black represents a standard memory paradigm. In red are additional iterations of the task using the result from the previous iteration as the new stimulus, which form the nodes in the serial reproduction process.
estimation, replicates some (but not all) of their key findings. In particular, we find four prototypes arranged in the centers of the four quadrants of the circle, as well as the horizontal and vertical oval shapes, as they did. However, we also show new and intricate patterns in the priors for the angular shapes (triangle, square, and pentagon). We evaluate the predictions of the model by Wedell et al. (2007) on the data we obtained following one iteration, for all the shapes, and compare them to predictions that we obtain from our estimates of the prior following all ten iterations. In addition, we show quantitative evaluations of the change in copying accuracy for the equilateral triangle. Second, we show the results for Experiment 2, where we illustrate the effect of adding vertices to regular polygons on the prior, revealing hitherto unknown grouping effects of the prototypes in spatial memory that occur as regular polygons begin to approximate a circle.

## Measuring spatial memory priors

Serial reproduction results. Figure 3 shows visualizations of the estimates that we obtained following ten iterations of the serial reproduction experiment using four hundred initial seeds for the circle. Each panel shows the results for each of the ten iterations, including the initial seeds. Notice that the prototypes begin to emerge in as early as the fourth iteration. For the panel showing the results of the tenth iteration, we show an estimate of the prior using our non-parametric kernel density estimate in lieu of plotting the points. Figure 4 shows the results that we obtained for the equilateral triangle. Notice the emergence of bimodal peaks near each of the vertices. This finding suggests that for this shape, there are a total of six prototypes in the prior, grouped in pairs at each corner.


Figure 3: Serial reproduction of 400 dots presented in a circle, for ten generations (iterations) of the process. The top left scatterplot shows the positions of the original seeds (sampled from a uniform distribution) inside the circle shape. The remaining subplots show scatterplots of the results of the serial reproduction chain for iterations 1-10. The subplot of the tenth iteration, in the bottom right, also shows the kernel density estimate. Note that from one iteration to the next, points that were originally scattered uniformally within the circle boundary begin to converge on each of the nearest prototypes at the center of each of the four quadrants in the circle. By the tenth iteration of the process, four clusters are clearly discernable.

Simple shapes. In Figure 5 (panels B and D), we show the kernel density estimates that we obtained for all the shapes. In the case of the circle, vertical oval, and horizontal oval, our results are consistent with past findings (shown in panels A and C). However, we discover bimodal peaks in the vertices of the angular shapes (prototype pairs clustered at each of the corners). This result is particularly striking for the triangle and the square shapes. The same result is present for the pentagon shape, although unlike the peaks in the prior for the triangle and square, those in the pentagon are not quite rotationally invariant, although all three geometric shapes are, suggesting that the shapes and orientations of the modes in the priors are not a simple function of the presence of edges, or the angles at these edges.

Convergence analysis. For the triangle results, we completed a convergence analysis (See Figure 4, panels B and C), using the Jensen-Shannon divergence (JSD). To estimate the variability of these JSDs, we generated 100 bootstrapped data sets sampled from the original data (with replacement). For each one, we computed the JSDs of consecutive iterations (see panel B). The JSD between the initial distribution and iteration 1 was significantly larger than that between the two final iterations ( $p=0.02$ ) and there were no significant differences between the distance between iterations 9 and 10 compared with iterations 8 and $9(p=0.43)$.


Figure 4: Results we obtained for each iteration in the chain for the triangle shape. A. Kernel Density Estimate (KDE) for the initial distribution and the 10 iterations. B. Convergence analysis using the Jensen-Shannon divergence (JSD) between consecutive iterations. C. JSD between all iterations and the last iteration. Note that both measures decrease with the number of iterations, and suggest that convergence occurs at or near the tenth iteration in the serial reproduction chain.

As another measure of convergence, we also computed the JSD between all iterations and the last iteration (Jacoby \& McDermott, 2017) (see panel C). The distance between the last two iterations was significantly smaller than the distances between iteration 10 and each of the remaining iterations ( 0 through 8). The distance between iteration 10 and 9 was marginally larger than the distance between iteration 10 and 8 ( $p=0.041$ ). These analyses suggest that convergence occurs at or near the tenth iteration. To test if the responses of participants became more "prototypical" over the course of the experiment (as they progressed through their trials), we used the estimate of the prior from the final iteration to measure the average log-likelihood of their responses. We used data from the $83 \%$ of the participants who performed more than $80 \%$ of the trials within the accepted criteria (responses within $8 \%$ of the height and width of the shape on the screen). We found that the log-likelihood significantly improved when comparing the first and second half of their responses $(\mathrm{t}(49)=-2.47$,
$p=0.008$ ), and when comparing the first 10 trials to the last 10 trials of each of the subjects $(\mathrm{t}(49)=-2.04, p=0.046)$.


Figure 5: Kernel density estimates for the priors were estimated for all six shapes using the tenth iteration of the serial reproduction chain. A. \& C. Original result by Wedell et al. (2007). B \& D. Kernel density estimates with serial reproduction. E. Boxplots showing model comparisons. We computed the log likelihood difference for the two models as explained in the main text. In all cases the serial reproduction model was significantly better ( $p<0.01$ for all shapes except vertical oval ( $p=0.03$ ) resulting in positive log-likelihood ratios.

Model comparisons. Using a combination of nonparametric kernel density estimation and serial reproduction lets us uncover intricacies in the prior for angular shapes (including bimodal peaks at the vertices) that paint a nuanced picture of human spatial memory priors. In addition, our approach enables us to obtain more than just point estimates of the locations of prototypes in spatial memory. Nevertheless, we provide a comparison between point estimates obtained from our method to those obtained from the model by Wedell et al. (2007), for each shape, using the same number of pa-
rameters. The model describes the remembered position for a dot $i$ (a response vector $\vec{R}_{i}$ ) as a weighted average of the actual location at which the dot was presented, which they refer to as the "fine-grain memory representation", and the weighted sum of the prototype locations, using the following equations:

$$
\begin{gather*}
\vec{R}_{i}=w \vec{S}_{i}+(1-w) \sum_{j=1}^{4} v_{i j} \vec{P}_{j}  \tag{1}\\
v_{i j}=\frac{e^{-c\left\|\vec{S}_{i}-\vec{P}_{j}\right\|}}{\sum_{k=1}^{4} e^{-c\left\|\vec{S}_{i}-\vec{P}_{k}\right\|}} \tag{2}
\end{gather*}
$$

where $\vec{S}_{i}$ and $\vec{R}_{i}$ are vectors in $\mathbb{R}^{2}$ containing the $x$ and $y$ coordinates for each point $i$ in the stimulus phase (iteration 0 ), and in the first response phase (iteration 1), respectively. The $\vec{P}_{j}$ terms correspond to the four prototype vector coordinates being estimated by the model, in addition to weights $w$ that correspond to the relative strength of the veridical memory (as opposed to the strength of a prototype in the prior).The $v_{i j}$ capture the relevance weight of each of the four $j$ prototypes for each point $i$. In other words, the strength of the influence of prototype $j$ for each point $i$. The parameter $c$ corresponds to a "sensitivity" parameter that models the sharpness of the prototype boundaries.

We generated 100 split-half samples of the points for iteration 0 (inital seeds), iteration 1 , and iteration 10 . Next, for each sample, we obtained estimates of the prototype locations for four prototypes (the same number used by Wedell et al.) by running their model using the training half of iteration 0 and the same points in iteration 1. In order to ensure a fair comparison, we sampled four points under local maxima from the Kernel Density Estimate (KDE) fit to the same points in iteration 10. This gave us four prototype estimates from the Wedell et al. (2007) model, and four points corresponding to local maxima in the KDE we fit to the points in the training half of iteration 10 (which can only be obtained from our paradigm), for each training split half. We evaluated the accuracy of these two sets of four prototype estimates by computing the sum of the negative-log-likelihood values from a KDE that we fit to the remaining points in the testing half of iteration 10. Next, we computed the log likelihood difference for the two models, for each of the shapes. In all cases, the serial reproduction model performed significantly better ( $p<$ 0.01 for all shapes except the vertical oval $(p=0.03)$ resulting in positive log-likelihood differences. Boxplots showing all the results are displayed in Figure 5E.

Grouping of prototypes. The apparent increase in peaks in the prior for more complex regular shapes afforded the opportunity to consider changes to the prior in the limit, as the shapes begin to approximate a circle. We computed the entropy of the obtained KDEs to quantify their complexity. Complexity increased with the number of vertices (going from a triangle to a heptadecagon, or seventeen-sided regular polygon). However, the prior for a icosihenagon (twenty-one sided regular polygon) begins to reveal the transformation of
the corner peaks into one of the quadrant peaks. Entropy further decreases for the icosipentagon ( $p<0.001$ ), revealing a prior that appears nearly identical to the prior for a circle, and with similar entropy ( $p=0.68$ ) (see Figure 6).


Figure 6: Grouping, and complexity of prior estimates. A. KDEs for regular polygons of increasing complexity. B. Entropy of the last iteration computed for all shapes. Entropy increases steadily with shape complexity ( 3 to 17 vertices). After the number of vertices exceeds 21, entropy stabilizes, and peaks start grouping toward the nearest quadrant center (as with the circle). We used the Bonferroni correction for multiple comparisons.

## Discussion

In this paper, we made a preliminary foray into exploring spatial memory priors using serial reproduction: a process in which information being transmitted through successive participants leaves behind only a signature of the transformation process itself: the perceptual and reconstructive biases of those participants. This iterative process provides an effective tool for greatly amplifying biases in perception and memory.

We used a serial reproduction paradigm in the context of a spatial memory task. KDEs of the dots' final positions revealed detailed structure in priors over location. We found that the priors for circles and ovals show peaks at the center of each of their four quadrants, but also discovered that angular shapes show bi-modal peaks at the vertices in the prior. The modes appear on either side of each vertex, and do not seem to be a simple function of the angle at each vertex, since they are not rotationally invariant in all cases. We provided quantitative comparisons between the performance of a parametric model, and point estimates derived from the KDEs we obtained following ten iterations of the chain. These comparisons demonstrated that our estimates were significantly better than those obtained from the parametric model (we used the same number of parameters-four prototype estimates, even though our method yields kernel density estimates that clearly reveal more than four in some cases). In future work, we intend to determine if priors differ across individuals, by repeating the experiments so that each participant completes a subset of chains in their entirety
(within-subject design). While some studies show differences between within and between-subject designs (Claidière et al., 2014), most studies showed high agreement between these versions (Xu \& Griffiths, 2010; Jacoby \& McDermott, 2017).

Our results suggest that our approach may provide an opportunity to uncover complex priors for a wide range of perceptual phenomena that would otherwise elude traditional experimental approaches, and parametric models. We plan to use it to measure memory biases when there is more than one point to be remembered (Lew \& Vul, 2015), and to probe for structured priors in memory for local orientation (Wei \& Stocker, 2016). Finally, we intend to uncover perceptual biases in spatial memory using natural complex images, and maps, to explore the effect of higher-order visual features and semantic content on spatial memory biases, and to probe for the emergence of geographic landmarks.

## Acknowledgments

This work was funded in part by NSF grant 1456709 to T.L.G., and DARPA Cooperative Agreement D17AC00004 to T.L.G and J.W.S. We would like to thank Thomas Morgan for his help with online experiments.

## References

Bartlett, F. C. (1932). Remembering: An experimental and social study. Cambridge: Cambridge University.
Claidière, N., Smith, K., Kirby, S., \& Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. Proceedings of the Royal Society of London B: Biological Sciences, 281(1797), 20141541.
Huttenlocher, J., Hedges, L. V., \& Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. Psychological review, 98(3), 352.
Jacoby, N., \& McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. Current Biology.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105(31), 10681-10686.
Lew, T., \& Vul, E. (2015). Structured priors in visual working memory revealed through iterated learning. In Cogsci.
Wedell, D. H., Fitting, S., \& Allen, G. L. (2007). Shape effects on memory for location. Psychonomic Bulletin \& Review, 14(4), 681-686.
Wei, X.-X., \& Stocker, A. A. (2016). Mutual information, fisher information, and efficient coding. Neural computation.
Weiss, Y., Simoncelli, E. P., \& Adelson, E. H. (2002). Motion illusions as optimal percepts. Nature Neuroscience, 5(6), 598-604.
Xu, J., \& Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. Cognitive Psychology, 60(2), 107-126.

A core-affect model of decision making in simple and complex tasks Othalia Larue (Othalia.Larue@Wright.Edu), Alexander Hough (hough.15@Wright.Edu), and Ion Juvina (Ion.Juvina@Wright.Edu)<br>Wright State University<br>Department of Psychology, 3640 Colonel Glenn Hwy<br>Dayton, OH 45435 USA


#### Abstract

When it comes to decision making, the dominant view suggests that engaging in a detailed analytical thought process is more beneficial than deciding based on one's feelings. However, there seems to be a tradeoff, as the complexity and amount of elements on which to base the decision increases, decisions based on affect seem to be more accurate than decisions based on a thorough analytical process in specific contexts. In those last cases, an affective modulation of memory may help to make better decisions in complex tasks that exceed human's limited cognitive capacities. Some dual process accounts, "deliberation-without-attention" hypothesis (Dijksterhuis et al., 2006), oppose a cognitive (i.e., conscious) route to an affective (i.e., unconscious) route. Since most dual process accounts suggest one type of process is better than the other, the interaction and integration of affective and more conscious analytical processes in decision making have been understudied. To address this issue, we propose an explanation of the dynamics and interaction of cognitive (i.e., explicit) and affective (i.e., implicit) encoding and retrieval of elements in memory, using a unified theory based on core affect (Russell, 2003), in the shape of a cognitive model in the ACT-R cognitive architecture.


Keywords: Core affect; ACT-R; decision making; dual process theory; memory modulation; implicit strategy

## Introduction

In a set of experiments, Dijksterhuis et al. (2006) and Mikels et al. (2011) show how being focused on the details of provided information, rather than feelings, affects accuracy in a decision making task. According to these results, being feeling focused in a more complex, memoryoverloading task proves to lead to better performance.

Until recently, the influence of emotion has been neglected in the judgment and decision making literature, with the focus initially being put on the biases emotions enable (Kahneman \& Tversky 1979). Gradually, the focus has shifted toward the positive role of emotions in decision making, as suggested by neurological evidence (Damasio, 1994; Bechara, Damasio, \& Damasio, 2000). In parallel, core affect theory (Russell, 2003) in emotion research, and the somatic marker hypothesis (Damasio, 1994) in decision making research, have emerged to explain how emotion can guide behavior towards a positive outcome.

In this paper we suggest that the results from Dijksterhuis et al. (2006) and Mikels et al. (2011) (i.e., being feelingfocused in a more complex task leads to better performance) can be explained with a core-affect model. Core affect (Russell, 2009) is a neurophysiological state accessible to consciousness as a simple non-reflective feeling and can be
described through the valence (i.e., negative or positive) and arousal (i.e., intensity) values. Our hypothesis here is that core affect modulates memory. The modulation would place emphasis on the objective value of an attribute (i.e., good or bad) rather than details (e.g., higher than average gas mileage), simplifying the information and allowing for more efficient use of cognitive resources. The core affect experienced by participants while implicitly considering options cumulates and later leads to a decision illustrating the emotion-cognition interaction. This, we think, proves to be a better strategy when the limit of memory capacity is reached (e.g., complex task). This hypothesis was tested using a cognitive architecture based on a unified theory of cognition, ACT-R. We previously used this core affect model to explain the impact of affective valence and arousal on memory and memory decay using participant's memory of negative and positive emotion words after different time periods (Juvina \& Larue, 2016). However, here the focus is on the role of affect in decision making. This allows for an explanation of how core affect and cognitive mechanisms are meshed.

## Background

The concept of emotion has been a subject of interest for quite some time. However, theories have only recently attempted to explain their role in cognitive processes using empirical research. Appraisal theories (Lazarus \& Folkman, 1984; Ortony, Clore and Collins, 1990) have emerged as the dominant approach to emotions in the last few decades. Appraisal has been defined as the personal meaning and significance to well-being that is constructed from evaluations of situational factors and knowledge. While this trend of theories clarifies the route by which humans evaluate their environments (e.g., in a bottom-up way), they do not clarify how ongoing affect influences the encoding and retrieval of information in an implicit manner.

In response to this incompleteness, core affect theory (Russell, 2003), "feeling is for doing" (Zeelenberg \& Pieters, 2006), and the somatic marker hypothesis (Damasio, 1994) have attempted to bridge the gap between emotion and behavior. The latter particularly addresses the domain of decision making.
Russell (2009) believes that most phenomena attributed to emotions can be explained in more simple terms (e.g., core affect) without the need for emotion. Core affect is a visceral state that happens before the emotion is consciously identified: feeling good or bad, lethargic or energized (Russell, 2009). Russell's core affect theory suggests
underlying values for valence and arousal are more important than emotion, which he believes is socially constructed. The core affect is the central notion of this theory. Previous events change the core affect, which can occur before the event is actually consciously perceived by the subject and persists during the episode. It also influences the other elements of the emotional episode.
In the domain of decision making, some researchers (Gigerenzer \& Selton, 2002) view heuristics, not only
affective ones, as strategies that lead to sufficient decisions. Implicit strategies for decision making have previously been studied in ACT-R with Instance Based Learning (Gonzalez, Lerch \& Lebiere, 2003). In this paper, another type of implicit strategy involved in decision making - an affective strategy - is investigated.

Existing computational models of affect and emotion are based on appraisal theories and tend to be pre-programmed and hardwired based on the specifications of a particular theory (e.g., Marsella \& Gratch, 2009; Marinier, Laird, \& Lewis, 2009). Previous attempts have been made in ACT-R to add biological roots of emotions (Dancy et al., 2015) effect of emotion on learning and decision making (Belavkin, 2003) and stress (Ritter, Reifers, Klein, \& Schoelles, 2007) by overlaying the architecture.

Since core affect is implicit, more primitive, and more general than the construct of emotion (Russell \& Feldmann Barrett, 1999; Russell, 2009), it could be particularly adapted to be included in a cognitive architecture. When meshed with existing cognitive mechanisms, it could add to existing unified theories of cognition. The resulting model would increase the explanatory power of the core affect theory by clarifying different aspects of the emotioncognition interaction.

## Core affect and memory: theory and implementation

## ACT-R and memory

To capture the core affect modulation of memory and its impact on decision making, we support our model with ACT-R (Adaptive Control of Thought - Rational; Anderson, 2007), a unified theory of human cognition. ACT-R is also a cognitive architecture that is used to develop computational models of various cognitive tasks. ACT-R is composed of various modules: goal, imaginal, visual, aural, manual, vocal, and two memory modules: declarative memory (i.e., facts) and procedural memory (i.e., how to do things). The declarative memory module, which stores facts (i.e., know-what), is the one the core affect directly modulates. Declarative memory includes both symbolic structures (i.e., memory chunks) and sub-symbolic quantities that control the operation of the symbolic structures in the equations. The valuation and arousal values, which help to define the core affect, are subsymbolic quantities added to the current sub-symbolic equations of ACT-R.

## Core affect and memory

We present a summarized version of the core affect mechanism to facilitate the understanding of our model. An extended version of the core affect mechanism can be found in Juvina and Larue (2016). The original equation (Anderson, 2007) that computes the activation of a declarative memory chunk is:

$$
\begin{equation*}
A_{i}=B_{i}+S_{i}+P_{i}+\varepsilon_{i} \tag{1}
\end{equation*}
$$

- $\quad A_{i}$ is the activation of the chunk $i$.
- $\quad B_{i}$ is the base-level term and reflects the recency and frequency of use of chunk $i$.
- $\quad S_{i}$ is the spreading term and reflects the effect of the context on the retrieval of chunk $i$.
- $\quad P_{i}$ is the partial matching term and reflects the degree to which the chunk $i$ matches the specification of the retrieval request.
- $\varepsilon_{i}$ is a noise or variability component.

Activation of a chunk reflects its use, and decays over time if the chunk is not used. Retrieval time and the retrieval probability of a chunk are determined by activation (e.g., chunks under a certain retrieval threshold cannot be retrieved). However, the selection process is impacted by noise. The chunk with the highest activation has the highest probability of being selected, but other chunks get the opportunity as well allowing some exploration behavior in ACT-R.
In the current ACT-R architecture, reward based learning affects procedural memory. However, subjective values (e.g., pleasant or unpleasant) might actually be carried by declarative memory as affectively charged representations, which are easier/harder to retrieve according to these values (Smith, Most, Newsome, \& Zald, 2006). A new ACT-R module called "Valuation" was developed to add valuation and core affect capabilities into ACT-R. In theory, core affect is a diffuse affective state that is not necessarily linked to any specific event and is characterized as a point in a two-dimensional space, where the two underlying dimensions are valence (i.e., pleasure-displeasure) and arousal (Russell, 2009). In our implementation, core affect is defined as two accumulators called core-affect-valuation (Valuation) and core-affect-arousal (Arousal), which are sub-symbolic quantities computed by the "Valuation" module. It also maintains the parameters and history information that are needed for these computations. Both values affect the probability that a chunk can be retrieved from declarative memory. Valuation is an indicator of the affective valence of a particular stimulus or fact learned through interaction with the environment. Arousal is an indicator of the importance or priority that is given to a particular stimulus or fact; it is the absolute value of valuation. Relying on the existing memory mechanisms from the ACT-R theory, valuation and arousal are just two separate terms added to the general activation equation previously introduced:

$$
\begin{equation*}
A_{i}=B_{i}+S_{i}+P_{i}+V_{i}+A r_{i}+\varepsilon_{i} \tag{2}
\end{equation*}
$$

- $\quad V_{i}$ is the valuation term and reflects the rewards received by the model after referencing chunk $i$.
- $A r_{i}$ is the arousal term which reflects the importance of chunk $i$ and is computed as the absolute magnitude of the valuation term.

The learning of valuations occurs when a reward is triggered: the valuations of all chunks that been referenced within a time window are updated. This is compatible with findings of overlapping neural substrates between the attribution of subjective value to stimuli and reward-based learning (Paton, Belova, Morrison, \& Salzman, 2006). The effective reward of a chunk $i$ is the reward value received at time $n$ minus the time since the last reference of chunk $i$.

The learning of valuations for a chunk $i$ is controlled by the following equation:

$$
\begin{equation*}
V_{i}(\mathrm{n})=V_{i}(\mathrm{n}-1)+\alpha v\left[R_{i}(\mathrm{n})-V_{i}(\mathrm{n}-1)\right] \tag{3}
\end{equation*}
$$

- $\quad V_{i}(\mathrm{n})$ is the valuation of chunk i after its nth update.
- $V_{i}(\mathrm{n}-1)$ is the valuation of chunk i prior to its nth update.
- $\alpha v$ is the learning rate for valuations.
- $\quad R_{i}(\mathrm{n})$ is the effective reward value received by chunk i before its nth valuation update.
- $V_{i}(0)$ is determined based on initial parameter settings.

Reward signals allow the model to learn valuation and arousal values for elements according to what is presented in the environment.
Additional parameters make it possible to weight valuation and arousal independently in the equation. Values used in this paper can be seen in Table 1:

- Valuation weight (:vw) is a scale parameter for the valuation term in the general activation equation.
- Arousal weight (:aw) is a scale parameter for the arousal term in the general activation equation.
- Valuation time window (:vtw) is a time window over which to update the valuations. It determines how many chunks are updated.

In the architecture, core affect is the weighted accumulation of valuation and arousal values for all retrievable chunks, and weights are probabilities of retrieval reflecting chunk activations. This value is implicitly maintained by the architecture.

In this implementation, affect phenomena are not hardwired in the cognitive architecture but learned from the interaction among various architectural components and between architecture and environment. Only primitive affective mechanisms: valuation (i.e., valence obtained through interactions) and arousal, were included in the cognitive architecture. Valuation and arousal are added as terms in the general activation equation and influence the probability of a chunk to be retrieved. This is consistent
with the core affect theory (Russell \& Feldmann Barrett, 1999; Russell, 2009).

Our hypothesis is that this is all that is necessary to include at the architectural level to model the interaction between cognition and affect.

## Model

## Conditions

The procedure used here was derived from an experiment by Dijksterhuis et al. (2006) and replicated by Mikels et al. (2011). During these experiments, participants were given information about four different car options (i.e., Car A, Car B, Car C, and Car D) and were instructed to choose which car they believed to be the best choice. Simple attributes, framed as either positive or negative (e.g., this car gets good/bad gas mileage), were provided one at a time for each car option. The best choice was defined as the car with the most positive attributes. The best choice had $75 \%$ positive attributes, two cars had $50 \%$ positive attributes, and one car had $25 \%$ positive attributes. The design consisted of one dependent variable (i.e., car choice) and two independent



Figure 1. Experiment procedure and memory representations across task in detail-focus vs feeling-focus conditions
Participants were split into four conditions based on the two independent variables (i.e., feeling-focus simple, feeling-focus complex, detail-focus simple, and detail-focus complex). Those in the feeling-focus conditions were instructed to rate how they felt about each attribute and make their choice while focusing on their feelings. In the detail-focus conditions, participants were instructed to rate how well they were remembering the attributes and make their decision based on the details of the attributes. Simple conditions had four attributes for each car option, whereas complex conditions had 12 attributes per car option.

All conditions completed a memory recall task at the end of the trial. Results from a chi-square analysis indicated that participants in the detail-focus simple condition performed better than participants in the feeling-focus simple condition, although this difference did not reach significance. However, participants in the feeling-focus complex condition significantly outperformed those in the detail-focus complex condition. There was no difference between focus conditions for memory recall, but there was a difference between simple and complex conditions. Both Dijksterhuis et al. (2006) and Mikels et al. (2011) concluded
that focusing on your feelings leads to better complex decisions compared to more deliberate thinking.

Table 1. Model parameters

| :rt | -2.4 |
| :--- | :--- |
| :vw | 1.0 |
| :aw | 2.0 |
| :av | 0.2 |
| :vtw | 0.5 |

## Encoding across conditions

Encoding mechanisms used are the same, but:

- In the detail-focus, the strategy used makes you consider all features and ratings associated to those features.
- In the feeling-focus condition, the strategy used gives more value to the ratings (i.e., good/bad) than their features..

Table 2. Model strategy in the detail-focus condition

| Step | Strategy in the detail-focus condition |
| :--- | :--- |
| Presentation phase |  |
| 1 | "Car-feature-value" triplet is displayed |
| 2 | See car |
| 3 | Encode car |
| 4 | See feature |
| 5 | Encode feature |
| 6 | See value |
| 7 | Encode value |
| 8 | Clear imaginal (enter chunk in declarative <br> memory) |
| 9 | Go back to step 1 until all cars have been <br> displayed |
| Evaluation phase |  |
| 10 | Pick a car that has not been evaluated yet |
| 11 | Retrieve triplet (car-feature-value) with a "good" <br> judgment |
| 12 | Count the positive values for this car |
| 13 | Go back to step 10 until there are no cars left |
| 14 | Decide the car with the highest count |

Detail-focus condition (Table 2). Stimuli consisting of three elements (car - feature - rating) are presented one at a time to the model The model looks at each element separately and encodes them as a memory chunk of the following association: car - feature - rating. Car is also a memory chunk (Presentation phase in "Detail-focus" in Figure 1).

When all stimuli have been presented to the model, it proceeds to the evaluation through a tallying heuristic (Gigerenzer, 2016): by interrogating its memory on features for each car, counting all chunks for which it can retrieve an association with a "good" rating for a feature. The car with the highest overall number of "good" rating-feature-car associations that could be retrieved is the one that is named
by the model as the best car choice (Evaluation phase in "Detail-focus" in Figure 1). The significantly higher number of reasoning steps in the detail-focus condition (Table 2) results from the thorough analytical process that participants were assumed to engage in during this condition.

Table 3. Strategy in the feeling-focus condition

| Step | Strategy in the feeling-focus condition |
| :--- | :--- |
| Presentation phase |  |
| 1 | "Car-feature-value" triplet is displayed |
| 2 | See car |
| 3 | Retrieve chunk car |
| 4 | See value |
| 5 | Trigger reward depending on value |
| 6 | Update valuations |
| 7 | Go back to step 1 until all cars have been displayed |
| Evaluation phase |  |
| 8 | Retrieve car with highest activation |
| 9 | Decide (highest valuation car is the one with the <br> most "good" features) |

Feeling-focus condition (Table 3). The same stimuli are presented to the model randomly; but the model is going to follow a different strategy. It only looks at the car and rating, as shown in Table 3 (Presentation phase in "Feelingfocus" in Figure 1). The model retrieves the car chunk associated with the presented car, and according to the rating "good" or "bad", sends a reward signal. This reward affects the valuation of this specific car without it being necessary to encode all the features of the car. When the reward signal is sent, all the valuations of chunks that were retrieved in this time window are updated. Recall the explanation in the previous section (detail-focus condition) that the memory representation includes the car chunks. Thus, if the car to which this rating was attached is in the time window, it gets a valuation update.

When all stimuli have been presented to the model, it proceeds to the evaluation by retrieving one of the previously presented cars (Evaluation phase in "Feelingfocus" in Figure 1). The retrieved car is the one with the highest activation, which likely has the highest rating because valuation was updated positively during the first stage of car presentations.

## Results and Discussion

Results in Figure 2 are shown for 50 runs of the model (stable performance, based on cumulative standard deviation, was reached after 43 runs).
There was a significant difference in accuracy between the two feeling focus conditions (i.e., simple and complex), $t(97.89)=2.67, p<0.01$. A significant difference in accuracy was also found between the detail-focus complex condition and feeling-focus complex condition, $t(83.1)=$ $6.001, p<.001$. These same differences were observed in the original experiments.

The model also captured participant's better performance in the detail-focus simple condition compared to the feelingfocus simple condition (difference is not significant like in the original results). However, while still within the standard error range, the detail-focus complex condition appears lower than in the original experiments. This could be explained by something that is not captured in the strategy of our model. When uncertain, human participants could have guessed more accurately (compared to random guesses by the model) based on prior knowledge. For instance, a participant may have eliminated options based on memory that certain options had fewer positive attributes.


Figure 2. Accuracy in feeling-focus vs. detail-focus in complex and simple conditions for the original experiments and our model

## Model dynamics in detail-focus condition

In the simple condition, the model has a performance close to the feeling-focus condition. However, in the complex condition more features are memorized for each car. In this condition, more words are forgotten as the experiment is longer and there are more words to remember. Activation of those unused chunks decay over time. Therefore, when going through all the cars and remembering the features, there are more chances of memory retrieval failures.

The forgetting time is also amplified by the length of the recall strategy, contributing to future retrieval failures. The model may be forgetting features of the next car while listing the elements of the current car. This explains the poorer performance of the model in the complex condition. It is important to note that the model does not account for possible confusions in the car and feature-rating associations.

## Model dynamics in feeling-focus condition

The model performs better in the complex condition than in the simple condition. In the complex condition, while the proportion of good features is the same, the overall number of features per cars is higher. This gives the model more opportunity for rewards. In the simple condition, there are
fewer features and less reward opportunities. The activation equation has a noise parameter. Due to this noise the chunk with the highest valuation might not be the one with the highest activation (thus, not the one retrieved). Therefore, when retrieval from memory is initiated, decay and noise might make the activation number obtained through the activation equation close but higher for another car than the one who received the highest rating. This happens more often in the simple condition where you will have chunks of very close valuations. Figure 3 illustrates the differences in valuation between the chunk representations of the car options. The gap in valuation between options is more visible in the complex condition.

In contrast to the detail-focus condition, retrieving the highest rated car in the Feeling-condition is a very simple and fast process. It only requires one retrieval of the car with the highest activation (no features retrieval involved), therefore there is less decay of activation for the chunks and therefore less ground for retrieval mistakes.


Figure 3. Evolution of the chunk valuation of each cars in feeling-focus condition across rounds, simple vs complex conditions

## General Discussion and Conclusion

In this paper we presented a mechanism for core affect in ACT-R. This mechanism specifies how affect modulates memory (e.g., reducing information or emphasizing the positive or negative value) compared to attempting to remember the entire set of attributes in the detail-focus condition (i.e., high memory load). It also shows that implicit decisions might lead to better decisions than explicit decisions in certain contexts.

We model affect in a cognitive architecture as a phenomenon, which emerges from the dimensions of the core affect theory (i.e., valence and arousal) and is learned through the interaction with the environment. We interpret valence as valuation. Valuation is a sub-symbolic quantity for chunks learned through interactions with the world. Arousal is the absolute value of valuation. Core affect is the weighted accumulation of valuation and arousal values for all retrievable chunks, and weights are probabilities of retrieval reflecting chunk activations. Parameters given to the ACT-R architecture, existing reward mechanism of the architecture, and usage information about the chunks are used to compute valuation and arousal values. The core affect values are implicitly maintained by the architecture. Valuation and arousal are added as terms in the general activation equation and influence the probability of a chunk
to be retrieved. We rely on the existing general activation equation of ACT-R to integrate our model in a unified theory of cognition.

While the core affect theory has been present in theories of emotion, and the role of emotions has been considered in the domain of decision making, very little has been done to connect the work on core affect theory to decision making.

We hypothesized and demonstrated that those mechanisms that allow for an affect phenomenon to emerge, were sufficient to account for the behavior encountered in Mikels et al. (2011)'s experiment. Engaging in a detailed analytical thought process might be as beneficial as deciding based on your feelings in a simple environment (i.e., low cognitive load). However, there is a tradeoff. As the amount of elements on which to base your decision increases, exerting a high load on your declarative memory, decisions based on affect seem to be more accurate than decisions based on a thorough analytical process in those complex environments. We demonstrated that an affective modulation of memory by core affect, which simplifies the amount and complexity of information, could explain this phenomenon. Therefore, core affect may help individuals make better decisions in complex tasks, which exceed limited cognitive capacities by reducing the need to memorize each element included in the decision. Instead, the interaction with the elements a decision is supposed to be based on, can be implicitly processed in conjunction with affect, and the resulting decision can be based on those affects. Furthermore, we showed that an implicit mechanism (core affect) allows us to make an efficient decision.

## Acknowledgments

This work was supported in part by The Air Force Office of Scientific Research grant number FA9550-14-1-0206 to IJ.

## References

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York: Oxford University Press.
Bechara, A., Damasio, H., \& Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. Cerebral cortex, 10(3), 295-307.
Belavkin, R. V. (2003b). On emotion, learning and uncertainty: A cognitive modelling approach. PhD Thesis, The University of Nottingham, Nottingham, NG8 1BB, United Kingdom.
Damasio, A. (1994). Descartes' error: Emotion, reason and the human mind. New York: Putnam Press.
Dancy, C. L., Ritter, F. E., Berry, K., \& Klein, L. C. (2015). Using a cognitive architecture with a physiological substrate to represent effects of psychological stress on cognition. Computational and Mathematical Organization Theory, 21(1), 90-114.
Dijksterhuis, A., Bos, M. W., Nordgren, L. F., \& van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. Science, 311, 10051007.

Gigerenzer, G., \& Selten, R. (2002). Bounded rationality:
The adaptive toolbox. MIT press.
Gigerenzer, G. (2016). Towards a Rational Theory of
Heuristics. In Minds, Models and Milieux (pp. 34-59). Palgrave Macmillan UK.
Gonzalez, C., Lerch, J. F., \& Lebiere, C. (2003). Instancebased learning in dynamic decision making. Cognitive Science, 27(4), 591-635.
Juvina, I. \& Larue, O. (submitted 2016) Modeling core affect in a cognitive architecture: The impact of arousal and valence on memory. Cognitive Systems Research.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: an analysis of decision under risk. Econometrica: Journal of the Econometric Society, 263-291.Lazarus, R. S., \& Folkman, S. (1984). Stress, appraisal, and coping. New York: Springer Publishing Company.
Lazarus, R. S., \& Folkman, S. (1984). Stress, appraisal, and coping. New York: Springer Publishing Company.
Marinier, R. P., Laird, J. E., \& Lewis, R. L. (2009). A computational unification of cognitive behavior and emotion. Cognitive Systems Research, 10, 48-69.
Marsella, S. \& Gratch, J. (2009). EMA: A Process Model of Appraisal Dynamics. Journal of Cognitive Systems Research, 10, 70-90.
Ortony, A., Clore, G. L., \& Collins, A. (1990). The cognitive structure of emotions. Cambridge university press.
Paton, J. J., Belova, M. A., Morrison, S. E., \& Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. Nature, 439(7078), 865-870.
Ritter, F. E., Reifers, A. L., Klein, L. C., \& Schoelles, M. J. (2007). Lessons from defining theories of stress for cognitive architectures. Integrated models of cognitive systems, 1, 254.
Russell, J.A. \& Feldman Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology 76(5), 805-819.
Russell, J. A. (2003). Core affect and the psychological construction of emotion. Psychological review, 110(1), 145.

Russell, J.A. (2009). Emotion, core affect, and psychological construction. Cognition and Emotion, 23(7), 1259-1283.
Smith, S. D., Most, S. B., Newsome, L. A., \& Zald, D. H. (2006). An emotion-induced attentional blink elicited by aversively conditioned stimuli. Emotion, 6(3), 523.
Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. Current directions in psychological science, 10(6), 224-228.
Zeelenberg, M., \& Pieters, R. (2006). Feeling is for doing: a pragmatic approach to the study of emotions ineconomic behavior. In D. DeCremer, M. Zeelenberg \& J. K. Murnighan (Eds.), Social psychology and economics (pp. 117-137). Mahwah, NJ: Erlbaum.

# Generic and Universal Generalisations: Contextualising the 'Generic Overgeneralisation' Effect 

Dimitra Lazaridou-Chatzigoga (d1518@cam.ac.uk)<br>Institut für Anglistik und Amerikanistik, Humboldt-Universität zu Berlin/University of Cambridge Unter den Linden 6, 10099, Berlin, Germany

Napoleon Katsos (nk248@cam.ac.uk)<br>Department of Theoretical and Applied Linguistics, University of Cambridge<br>9 West Road, Cambridge, CB3 9DP, UK

## Linnaea Stockall (l.stockall@qmul.ac.uk)

School of Languages, Linguistics and Film, Queen Mary, University of London Mile End Road, London, E1 4NS, UK


#### Abstract

In this study, we focused on the Generic Overgeneralisation (GOG) effect (Leslie, Khemlani, and Glucksberg 2011) and tested the relevance of context and an explanation based on quantifier domain restriction for the pattern of judgement data observed. Participants judged generic majority characteristic statements like tigers have stripes or statements with universal quantifiers that have different sensitivity to context ('all', 'all the', 'each') preceded by one of three levels of context: a) neutral, where the information in the context does not interact with the truth value of the critical statement, b) contradictory, where it presents an exception which should rule out a universally quantified statement, and c) supportive. Our results suggest that proponents of the generics-as-default view ruled out context prematurely and that in fact context is a viable alternative explanation for much of the so-called GOG effect.


Keywords: context; generalisation; genericity; quantification; quantifier domain restriction

## Introduction

Quantificational generalisations, as in (1), are expressed in quantitative, statistical terms, while generic generalisations, as in (2)-(3), make general claims about kinds of entities and refer to a property that is characteristic of the kind in question, but not necessarily statistically prevalent, as in (3):
(1) Some lions live in cages.
(2) Lions roar.
(3) Lions have manes.

Generic generalisations have long been studied in formal semantics, within which genericity is frequently viewed as a species of quantification. Even though generics have been studied since the seventies (see Dahl 1975; Carlson 1977), how to characterise their semantic interpretation and how to model their truth conditions remains a controversial topic (see discussion in Mari, Beyssade, and del Prete 2013). Within formal semantics, modal logic and probabilistic approaches are most prominent, both of which treat genericity as akin to quantification. According to the modal approach, which is the most widely adopted formal analysis
of genericity, generic meaning is obtained as the effect of an underlying operator or quantifier dubbed 'GEN', which is not phonologically realised but which is active in the composition of the sentence meaning and is an unselective variable binding operator similar to adverbs of quantification like usually, typically, always, as analysed in Lewis (1975). This operator is sentential and is represented by a tripartite structure as in (4) (Krifka et al. 1995). Thus, the logical form of (2) may be given as follows in (5):
(4) GEN [restrictor] [matrix]
(5) $\operatorname{GEN}_{\mathrm{x}}[$ Lions (x)] [Roar (x)]

Generic generalisations can be made using a wide range of different grammatical means, including definite and indefinite singulars and bare plurals in English, but no language has a unique, unambiguous marker of genericity equivalent to a quantifier or determiner. It is important to note that none of the analyses that posit a 'GEN' operator offer an explanation for this, a point that a recent psychological approach to generics, capitalises upon.

This growing body of experimental and developmental psychological work on the topic proposes that genericity is categorically different from (and significantly simpler than) quantification (Leslie 2007, 2008, Gelman 2010). This latter hypothesis, called the Generics-as-Default view (GaD view henceforth), treats generics as an innate and default mode of thinking. This idea is linked to the 2 -system view of cognition argued for by Kahneman and Frederick (2002) among others, which includes a distinction between System 1, a fast, automatic, effortless lower-level system, and System 2, a slower, more effortful higher-level rulegoverned system. According to this view, the fact that no language has a dedicated overt 'GEN' operator does not come as a surprise: given that generics are the most primitive default generalisations, children do not need to learn anything in order to acquire them.

The GaD approach argues that the fact that there is no overt generic operator in any known language is because generics are the unmarked, System 1, case. On this view, only effortful, non-default quantificational generalisations
require overt linguistic exponence. However, while assigning generics to a more basic, unmarked System 1, mode of thinking may sound intuitive at some level, it rests on a vague and undefined notion of markedness. Leslie refers to Chomsky (2000), but she gives no definition of markedness. Intuitively, it seems that what is at stake is surface level overt realisation (third notion of markedness of Haspelmath 2006).

A challenge for both types of approach is to determine which properties can be used in generic statements. Generic generalisations range from definitional statements that do not tolerate any exceptions (triangles have three sides) to statements that involve characteristic properties and are true of the majority of the kind with only a few exceptions (tigers have stripes) - which are called 'majority characteristic' by Leslie et al. (2011) - through statements that are true of a minority of the kind yet are characteristic (ducks lay eggs) - which are called 'minority characteristic'to statements that have low prevalence but involve a property that is noteworthy in some way (sharks attack people) - which are called 'striking'. A further complication is that statistical prevalence is neither a necessary nor a sufficient condition to license generic generalisations, as statements like books are paperbacks may be true of the majority of the kind, but are typically judged as false and thus fail as generic generalisations ('false generalisations').

In Lazaridou-Chatzigoga, Katsos and Stockall (2015) we juxtaposed the linguistic approach to genericity and the experimental research investigating the GaD hypothesis and we highlighted the significant challenges for each approach. We concluded that interdisciplinary work, integrating the tools and perspectives of both strands of investigation, is needed in order to advance our understanding of genericity.

In Lazaridou-Chatzigoga, Stockall and Katsos (2017) we focused on the effect called 'Generic Overgeneralisation' (GOG) (Leslie et al. 2011), which has been used to support the GaD view on generics. The Generic Overgeneralisation (henceforth 'GOG') effect is "the tendency to overgeneralise the truth of a generic to the truth of the corresponding universal statement" (Leslie et al. 2011:17). In that paper, we discuss a set of four non-mutually exclusive explanations for the GOG effect: a) ignorance of the relevant facts, $b$ ) subkind (taxonomic) interpretation, c) the atypical behaviour of all and d) Quantifier Domain Restriction (QDR), which will be explained in more detail in the next section. We proposed that all these factors play a role in explaining the attested behaviour by adults. These factors are independently attested and known to interact with the interpretation of generic and quantified statements. We suggested that even the name of the GOG effect might be misleading. The effect mainly tries to capture the behaviour observed with the quantifier all, which supposedly gets a generic interpretation as a result of an overgeneralisation bias. Thus, perhaps a better name for that effect would be 'Quantifier Reanalysis' effect, because this term would direct the focus where we believe it belongs: on the interpretation of all, or more generally of quantifiers,
rather than the interpretation of generic statements. The overall aim in that paper was to showcase the role of linguistic factors (both semantic and pragmatic) in the interpretation of generic and quantified statements, and to underscore the relevance of linguistically motivated explanations.

In this paper, we address the effect of context on generic and universally quantified generalisations empirically.

## The GOG effect

The first detailed investigation of the GOG effect is found in Leslie et al. (2011). In their experiment 1, participants performed a truth-value judgement task on sentences that were presented in one of three forms: generic, universal (all), or existential (some). The statements involved different kinds of properties: quasi-definitional (triangles have three sides), majority characteristic (tigers have stripes), minority characteristic (ducks lay eggs), majority non-characteristic (cars have radios), striking (pit bulls maul children), and false generalisations (Canadians are right-handed). The authors report that adults sometimes judge universal statements as true, despite knowing that they are truth-conditionally false. For example, participants judged a quantified statement like all tigers have stripes as true, even though they know it is false given that there are albino tigers. The authors claim that the participants made this 'error' because they relied on the corresponding generic statement, which is true. They find that the GOG effect is restricted to characteristic properties and that it occurs in more than half the trials: 78\% for majority characteristic and $51 \%$ for minority characteristic statements.
The authors argue that these elevated acceptance rates are due to participants interpreting the 'false' universally quantified statements as if they were their 'true' generic counterparts, and are thus a clear case of GOG. Before reaching that conclusion, the authors acknowledge three alternative explanations, which they argue are ruled out with subsequent experiments: a) ignorance of the relevant facts, namely, that participants do not know that male ducks do not lay eggs, which they ruled out by administering a knowledge test that confirmed knowledge of the relevant facts (their experiment 3), b) subkind interpretation, namely, that participants interpret all ducks lay eggs as 'all kinds/types of ducks lay eggs', which is a true statement, which they discarded through a paraphrase task (their experiment 2 b ), where only $1 \%$ of the paraphrases explicitly mentioned subkinds, and c) Quantifier Domain Restriction (henceforth QDR), to which we turn in the next paragraph.
Under an explanation based on $Q D R$ participants might interpret a statement like all ducks lay eggs as applying only to a relevant subset of ducks, namely the mature fertile female ducks. Quantified statements are interpreted within a context, which may restrict the scope of the quantifier (see Stanley and Szabó 2000). Thus, the reason why people accept the all statement is because (they believe) it is true once the quantifier has been restricted to the relevant subset of ducks. The authors addressed this alternative explanation
in their experiment 2 a , where they provided the participants with a background context, which was presented before each statement. These contexts included artificial population estimates of the following form:
(6) 'Suppose the following is true: there are 431 million ducks in the world. Do you agree with the following: all ducks lay eggs."

This information was supposed to prime quantification over every individual duck in the world, and thereby make it difficult/impossible to interpret all as restricted to only the ducks that are presupposed by lay eggs. If acceptance of all ducks lay eggs in the first experiment was driven by QDR the authors predicted that it would disappear in the context of population information of the kind above. However, the GOG effect still occurred on a substantial portion of trials for statements with all, with a $60 \%$ acceptance rate for majority characteristic statements and $30 \%$ for minority characteristic statements - less than when the statements appeared with no preceding context ( $78 \%$ and $51 \%$ respectively), but still a high percentage. The authors thus concluded that domain restriction could not be the sole explanation for the GOG effect.

On the above grounds, Leslie et al. (2011) rejected all three alternative explanations and argued they had found evidence for a strong generic bias, according to which people sometimes treat universally quantified statements as if they were generic.

## Overview of the present study

In the present study, we addressed QDR as an explanation for the GOG effect building on a design used by LazaridouChatzigoga and Stockall (2013). We chose to focus on the relevance of context and QDR given that the latter is a pervasive phenomenon affecting quantifiers and their interpretation within a context, and is routinely invoked in quantification (Heim 1991). According to QDR, the domain of a universal quantifier can be restricted in the following sense: in a discourse like 'There was rhubarb pie for dessert, Everyone developed a rash' (example modified from von Fintel 1994) a quantifier like everyone does not quantify over all the individuals in the world, but rather over the contextually restricted set of individuals. Furthermore, listeners are known to be charitable (Grice 1975). Thus, in a conversation one assumes that speakers take the most sensible positions and make the most plausible assertions. Under this view, interpreting everyone as quantifying over all the individuals in the world seems a rather unlikely intended interpretation and moreover one that is not charitable to the speaker because it renders her utterance false, whereas interpreting everyone with respect to the available set of individuals is not only plausible but also charitable to the speaker.

We hypothesised that if we could show that the amount of GOG behaviour can be altered by carefully manipulating different levels of contextual information preceding the
critical utterance, we would have evidence that the observed tendency to accept universally quantified statements as true can be explained through independently motivated mechanisms and that there is no need to appeal to GOG.
Rather than the population statistics contexts used by Leslie et al. (2011), which only had a moderate effect on participant behaviour, we decided to use three different types of contexts. Furthermore, because of the design we adopted, we decided to focus only on majority characteristic statements ('tigers have stripes') leaving minority characteristic statements ('ducks lay eggs') for future investigation. We varied the context preceding the critical utterance as follows: a) neutral, where the information in the context does not interact with the truth value of the critical statement; b) contradictory, where exceptions which should rule out a universally quantified statement are made salient, and c) supportive, where a paraphrase of the critical property is given, which makes its generality salient. Two of the three context types (contradictory and supportive) made the relevant domain for QDR salient, while the neutral context served as a baseline measure. The contradictory and supportive contexts turned the implicit restriction to 'all normal' individuals to an explicit one by either highlighting some abnormal individuals (contradictory) or by stating that the relevant individuals had the relevant property, i.e. they were normal individuals (supportive).
In addition to manipulating context, however, a compelling test of the QDR view also requires testing whether the GOG effect is observed only with all or whether different universal quantifiers would show such an effect. There are reasons to believe that all should not be treated as a representative universal quantifier. It has been argued to a) participate in fallacious reasoning (Jönsson and Hampton 2006), b) be prone to hyperbolic/loose use similar to 'almost all' (Claridge 2011), and c) be ambiguous between distributive and collective interpretation (Beghelli and Stowell 1997). Thus, using different types of universal quantifiers is essential to test the scope of the GOG effect. Furthermore, a study that specifically addresses the relevance of QDR should include universal quantifiers with different sensitivity to QDR. More specifically, QDR is less likely if the universal quantifier used does not require linking with a set under discussion, as is the case with all, compared to each and all the, which have to be interpreted as D(iscourse)-linked (Partee 1995, Pesetsky 1987).
To recapitulate, (a) the contextual manipulations used were expected to make the implicit domain restriction explicit and salient to the participants and (b) this manipulation was expected to influence truth-value judgements by showing a decrease in acceptance rates in the contradictory condition and an increase in acceptance rates in the supportive condition.

## Method

## Participants and procedure

120 volunteers ( 49 male, 70 female, 1 other; aged 19-67; mean age 37.28; SD 13.06) participated in the experiment
over the Internet. Participants were recruited through Amazon's Mechanical Turk system for human interface tasks. All spoke English as their first language and lived in the United States. The study was presented in the online platform Qualtrics. Each trial consisted of three displays. In the first display participants read a background context, in the second display they read a statement and in the third display they were asked to judge whether they agreed with the statement they just read. Their response was recorded by selecting keyboard keys (' A ' for yes and ' K ' for no).

## Materials and design

Participants were presented with 84 statements, including 48 fillers presented in a randomised order. The 12 test items consisted of majority characteristic statements like tigers have stripes and horses have four legs. We included 26 control items, 12 definitional statements like ants are insects and 12 false generalisations like books are paperbacks to get baseline measures and to (semi)counterbalance the percentage of expected True/False responses. All the contexts and items were normed beforehand by English native speakers, who did not take part in the experiment. This was done to make sure that the context manipulations worked as intended. Most of the experimental items used are a subset of the items used by Leslie et al. (2011). The two conditions we manipulated for the majority characteristic items were: a) determiner type: bare plural generic/all/all the/each, and b) context type: neutral/contradictory/supportive. The statements were presented in one of the determiner forms and were preceded by one of three types of context: a) neutral, b) contradictory, or c) supportive, examples of which are given below:
(7)
a. neutral context: Linton $Z o o$ is home to three tigers, Tibor, Baginda and Kaytlin, whose playful games visitors love to watch and photograph.
b. contradictory context: Linton Zoo is home to three tigers, Tibor, Baginda and Kaytlin, whose fur is all white due to a recessive gene that controls coat colour.
c. supportive context: Linton Zoo is home to three tigers, Tibor, Baginda and Kaytlin, whose black and orange coats visitors love to photograph.

Given the 4 determiners (generic/all/all the/each) we created 4 lists with 3 sublists each that varied with respect to the pairing of the items with context type, which gave us 12 lists in total. There were 10 participants in each sublist, who were assigned randomly. Here is a sample of a trial of a statement with all after a neutral context:
(8)

DISPLAY 1:
Background:
Linton Zoo is home to three tigers, Tibor, Baginda and Kaytlin, whose playful games visitors love to watch and photograph.

DISPLAY 2:
Statement: All tigers have stripes. DISPLAY 3:
Do you agree with the statement?
o Yes (A) o No (K)
The definitional and false generalisations were in the generic form in all lists. Fillers served to ensure the percentage of expected True/False responses was similar. The definitional and false generalisations, as well as the fillers, were preceded by a context that did not vary across conditions. The materials can be viewed at http://www.dimitra-lazaridou-chatzigoga.com/cogsci-paper/

## Results and discussion

The final model used included 116 out of the 120 participants. 4 participants were excluded based on their responses to definitional statements; we excluded subjects that responded correctly at fewer than 10 out of 12 items.

## Acceptance rates

Table 1 summarises the proportion of 'TRUE' responses to the TVJ question for the test items (majority characteristic statements) in each condition. We report proportion of 'TRUE' responses rather than the actual number of responses to facilitate comparison with Leslie et al.'s (2011) results.

Table 1: Mean Proportion (SE) of 'TRUE' responses as a function of context and determiner type.

|  | Context |  |  |
| :--- | :--- | :--- | :--- |
| determiner | neutral | contradictory | supportive |
| GEN (ø) | $99.14(3.12)$ | $87.07(0.86)$ | $100(0)$ |
| all | $80.56(3.82)$ | $48.15(4.83)$ | $87.96(3.14)$ |
| all the | $78.33(3.78)$ | $37.50(4.43)$ | $90(2.76)$ |
| each | $79.17(3.72)$ | $30.83(4.23)$ | $85.83(3.2)$ |

As we see above, generics were accepted at higher rates overall than universals, as expected, given that we had chosen items that were true in generic form. Both in the neutral and the supportive condition acceptance rates for generics were at ceiling ( $99 \%$ and $100 \%$ respectively) and were only lower in the contradictory condition (87\%). With universals, the picture is more complicated. All three universals (all, all the, each) were accepted at similar rates in both the neutral and the supportive condition, showing only a small increase in the supportive condition. In the neutral condition, all-statements were accepted $81 \%$ of the time, all the-statements $78 \%$ of the time and eachstatements $79 \%$ of the time. In the supportive condition, allstatements were accepted $88 \%$ of the time, all thestatements $90 \%$ of the time and each-statements $86 \%$ of the time. Universals after a contradictory context yielded fewer acceptances overall, as expected: all-statements were accepted $48 \%$ of the time, all the-statements $38 \%$ of the time and each-statements $31 \%$ of the time. Acceptance rates
for both generics and universals differed significantly between the neutral and the contradictory condition.

On the surface, we do get many 'TRUE' responses to universal quantifiers, as in the GaD literature, which might look like a GOG effect. We predicted that contradictory context should decrease acceptances across the board, while supportive context should increase them. We expected a smaller effect for generics because the generic statements were constructed so as to be true and because of their resistance to contextual restriction (i.e. we expected ceiling effects) and we predicted differences between the universal Qs depending on their sensitivity to QDR/D-linking. Nevertheless, we had specific predictions about the relative rates depending on the level of context, which according to Leslie et al. should not differ. In order to appreciate the relative effect of context on acceptance rates, we subtracted the average means of the contradictory condition from the average means of the neutral condition, as well as the average means of the supportive condition from the average means of the neutral. We took acceptance rates in the neutral condition as our baseline plotted at 0 . Negative values mean fewer acceptances and positive values mean more acceptances. We interpreted the rates obtained as the relative effect of context on acceptance rates plotted in figure 1 below.


Figure 1: The relative effect of context.
We get the effect for the contradictory context exactly as predicted for the universal Qs. The relative effect is bigger for those quantifiers that require QDR because of their semantics (all the, each) than for the one that allows but does not require QDR to the relevant subset (all). We also find that context does affect GEN in the contradictory context.
The prediction about the supportive context was not borne out because of ceiling effects in the neutral condition. Adding explicit information supporting the statement hardly mattered, as acceptance rates did not rise significantly. The ceiling effect might be due to participants being charitable and/or exceptions not being immediately salient.

We used R (R Core Team, 2012) and the lme4 package (Bates et al. 2015) to perform a generalised mixed-effects linear analysis of the relationship between determiner and context, specifying a binomial family. Responses were treated as a dummy coded categorical variable and were modelled with glmer. First, we fitted a full model with det.type and context.type as fixed effects (with an interaction term) and with random intercepts for subjects and items. We performed a likelihood ratio test of the full model with an interaction term against a model without the interaction term and the comparison proved non-significant $\left(\chi^{2}(6)=8.3455, p=0.2139\right)$. Thus, including an interaction term did not improve model fit, so we used the model without the interaction term for all subsequent analyses/comparisons.
We then fitted versions of the full model, in which a single effect was removed and we compared the reduced model to the model without interaction. To test the main effect of context, we removed context. A likelihood ratio test of the model without interaction against the model without context proved highly significant $\left(\chi^{2}(2)=311.81, \mathrm{p}<2.2 \mathrm{e}-16\right)$. Thus, we concluded that there was a main effect of context. To test the main effect of determiner, we removed determiner. A likelihood ratio test of the model without interaction against the model without determiner proved highly significant $\left(\chi^{2}(3)=58.183, \mathrm{p}=1.436 \mathrm{e}-12\right)$. Thus, we concluded that there was a main effect of determiner.

## General Discussion

We set out to explore one of the alternative explanations for the judgement data that concern universally quantified statements, which have been used as evidence of a GOG effect. The study presented here provides experimental evidence for the relevance of a QDR-based explanation of the purported GOG effect. In our study, context did not only affect acceptance rates for all and other universal quantifiers (all the, each), but it further predicted the levels of QDR depending on the level of context. The effect of context was greater for all the and each, two quantifiers that require QDR, while it was smaller for all, whose domain is only optionally restricted. Leslie et al. had ruled out the relevance of context and hence predicted no differences in acceptances across contexts for all. Furthermore, even though they only discuss all they make general claims about (universal) quantification being prone to the GOG effect. We argue that drawing conclusions about universal quantification (and by extension about genericity) requires more subtle manipulations. The differences we found between the different universal quantifiers are predicted according to the QDR view advanced here, but ought to be inconsistent with the GaD view, were they to discuss them.
We also find that context matters for generics too, a fact that bears further investigation, but is in line with recent work that claims that generics display some context sensitivity (Sterken 2015). This might be more consistent with an analysis in which GEN also involves (some form of) quantification rather than one that treats GEN as
categorically/ontologically different from universal quantifiers.
Overall, we argue that there exist alternative explanations for big portions of the supposed GOG effect. The study discussed here did not address all the alternatives, but so far in the literature it has been shown that at least pure error, ignorance and now context all significantly affect acceptance rates. In work in progress, we address crosslinguistic variation in the realisation of generic and universal generalisations. The general thrust of this work is that, rather than being under the influence of a default bias, adults are simply sensitive to the subtle interplay of quantifier semantics and pragmatics on the one hand, and context on the other. This approach has the advantage of accounting for data without postulating ad-hoc mechanisms such as GOG just for generics.

## Acknowledgments

This work was supported by a British Academy/Leverhulme Trust Small Research Grant (SG-132271). We are grateful to Faidra Faitaki and Valentina Hu for their help with experimental materials and data entering, to Anne Beshears, Patrick Elliott, Shivonne Gates, David Hall, Jessica Soltys, Tom Stanton, Nikola Vukovic and Elspeth Wilson for their help with norming and the experimental software, and to Stephen Politzer-Ahles for advice on the statistics.

## References

Bates, D., Maechler, M., Bolker, B. \& Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Beghelli, F., \& Stowell, T. (1997). Distributivity and negation. In Ways of Scope Taking, edited by A. Szabolcsi, 71-107. Dordrecht: Kluwer.
Carlson, G. (1977). Reference to kinds in English. PhD dissertation, University of Massachusetts, Amherst.
Chomsky, N. (2000). New Horizons in the Study of Language and Mind. Cambridge: CUP.
Claridge, C. (2011). Hyperbole in English: A corpus-based study of exaggeration. Cambridge, UK: CUP.
von Fintel, K. (1994). Restrictions of quantifier domains. PhD dissertation: University of Massachusetts, Amherst.
Gelman, S.A. (2010). "Generics as a window onto young children's concepts". In F.J. Pelletier (Ed.), Kinds, things, and stuff: The cognitive side of generics and mass terms. New directions in cognitive science, 100-123. New York: OUP.
Grice, H.P. (1975). "Method in Philosophical Psychology: From the Banal to the Bizarre", Proceedings and Addresses of the American Philosophical Association (1975), 23-53.
Haspelmath, M. (2006). Against markedness (and what to replace it with). Journal of Linguistics 42, 25-70.
Heim, I. (1991). Artikel und Definitheit. In A. von Stechow and D. Wunderlich (eds.), Semantik: Ein internationales Handbuch der zeitgenössischen Forschung, 487-535. Berlin: de Gruyter.

Jönsson, M. L., \& Hampton, J.A. (2006). "The inverse conjunction fallacy." Journal of Memory and Language 55, 317-334.
Kahneman, D. and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.) Heuristics and Biases: The Psychology of Intuitive Judgment. CUP. New York, 49-81.
Krifka, M., Pelletier, F., Carlson, G., ter Meulen, A., Chierchia, G. and Link, G. (1995). Genericity: An Introduction. In G. Carlson and F. J. Pelletier (Eds.) The Generic Book. Chicago. Chicago University Press, 1-125.
Lazaridou-Chatzigoga, D., and Stockall, L. (2013). Genericity, exceptions and domain restriction: experimental evidence from comparison with universals. In Proceedings of Sinn und Bedeutung 17, edited by E. Chemla, V. Homer, and G. Winterstein, 325-343. École Normale Supérieure, Paris.
Lazaridou-Chatzigoga, D., Katsos, K. and Stockall, L. (2015). Genericity is easy? Formal and experimental perspectives. RATIO 28(4), 470-494.
Lazaridou-Chatzigoga, D., Stockall, L. and Katsos, N. (2017). A new look at the 'Generic Overgeneralisation' effect. Inquiry. Advance online publication. doi: http://dx.doi.org/10.1080/0020174X.2017.1285993
Leslie, S.J. (2007). Generics and the structure of the mind. Philosophical Perspectives 21, 375-405.
Leslie, S.J. (2008). Generics: Cognition and acquisition. The Philosophical Review, 117(1), 1-49.
Leslie, S.J., Khemlani, S. and Glucksberg, S. (2011). All Ducks Lay Eggs: The Generic Overgeneralization Effect. Journal of Memory and Language 65(1), 15-31.
Lewis, D. (1975). Adverbs of Quantification. In Formal Semantics in Natural Languages, edited by E. Keenan, 315. Cambridge University Press.

Mari, A., Beyssade, C. and del Prete, F. (2013). Genericity. Oxford: OUP.
Partee, B. (1995). Quantificational structures and compositionality. In E. Bach, E. Jelinek, A. Kratzer and B. Partee (eds.). Quantification in Natural Languages. Dordrecht: Kluwer, 541-601.
Pesetsky, D. (1987). Wh-in-Situ: Movement and Unselective Binding. In the Representation of (In)definiteness, Eric J. Reuland \& Alice G. B. ter Meulen, eds. MIT Press: Cambridge, Mass.
R CoreTeam (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Stanley, J. and Szabó, Z.G. (2000). On Quantifier Domain Restriction. Mind and Language 15 (2-3), 219-61.
Sterken, R. (2015). Leslie on Generics. Philosophical Studies, 172(9), 2493-2512.

# Children's semantic and world knowledge overrides fictional information during anticipatory linguistic processing 

Ruth Lee (rj.lee@mail.utoronto.ca)<br>OISE, University of Toronto<br>Craig G. Chambers (craig.chambers@utoronto.ca)<br>Department of Psychology, University of Toronto<br>Falk Huettig (falk.huettig@mpi.nl)<br>Department of Psychology of Language, Max Planck Institute for Psycholinguistics<br>Patricia A. Ganea (patricia.ganea@utoronto.ca)<br>OISE, University of Toronto


#### Abstract

Using real-time eye-movement measures, we asked how a fantastical discourse context competes with stored representations of semantic and world knowledge to influence children's and adults' moment-by-moment interpretation of a story. Seven-yearolds were less effective at bypassing stored semantic and world knowledge during real-time interpretation than adults. Nevertheless, an effect of discourse context on comprehension was still apparent.


Keywords: discourse; children; sentence comprehension; eyetracking; semantics; cognition; fantastical fiction

## Real-time interpretation of fantastic fiction

Linguistic processing requires listeners to identify relevant thematic relationships between the entities and events evoked in a sentence. Studies of visually-situated language processing have shown that comprehenders use such relations to predict upcoming linguistic input, and in turn direct their attention to compatible referents in the visual world (e.g., Altmann \& Kamide, 1999; Kamide, Altmann, \& Haywood, 2003). For example, when hearing the sentence 'The boy eats the big cake' while looking at a scene containing a cake and a bird, adults and children as young as 2 years of age look to the cake while 'eats' is unfolding (Mani \& Huettig, 2012). Children as young as 3 years of age can also use their prior knowledge of the relationships between actions and agents to generate more sophisticated predictions, e.g. anticipating 'bone' upon hearing "The dog hides" (Borovsky, Elman, \& Fernald, 2012).

In adults, comprehension is also rapidly influenced by higher-order meaning created by physical, functional and situational relations between entities and events (Chambers \& San Juan, 2008; Sedivy, 2003; Tanenhaus, SpiveyKnowlton, Eberhard, \& Sedivy, 1995). Situation-specific factors, including a fictional context, can in fact override lexical and semantic relationships based on stored knowledge (Cook \& O’Brien, 2014; Filik, 2008; Nieuwland \& Van Berkum, 2006). However, prior research with grade school children has shown that they tend to privilege lexical
information over situation-specific knowledge (Snedeker \& Trueswell, 2004; Trueswell, Sekerina, Hill, \& Logrip, 1999). Children may therefore find it difficult to rely on a fictional context to inform real-time language processing, particularly in fantastical fiction where described events (e.g., a person flying) strongly depend on information in the narrative, and are at odds with the nature of the real world. Conversely, it is possible that the incongruent actions and salient contrast between the real and narrative worlds involved in a fantastical narrative may strengthen children's mental simulation of story information, and thus support their ability to rely upon contextual information to anticipate upcoming language input. Preschool children are already becoming competent comprehenders of discourse; they become sensitive to its causal structure (Lynch et al., 2008) and begin to make inferences connecting the events evoked in narrative with world knowledge (Barnes, Dennis, \& Haefele-Kalvaitis, 1996). Preschool children also understand that events in fiction can contradict their knowledge of the real world, and involve systematic rules governing what can and cannot happen within the context of that world (Sharon \& Woolley, 2004; Van de Vondervoort \& Friedman, 2014).

Thus, is not clear how effectively young children can use fantastical facts introduced in a narrative to interpret the story as it unfolds. Notably little work has investigated the real-time processes and underlying mechanisms involved in children's interpretation of fictional discourse. Most investigations of children's narrative comprehension have instead relied on offline measures, such as the verbal production of story elements (e.g. Paris \& Paris, 2003). By using implicit measures such as eye movements we can gain additional insights into children's moment-by-moment and automatic consideration of possible referents as language unfolds in real time.

In the current study, we investigate children's real-time language processing in discourse contexts that present fantastical protagonists and improbable events, using a spoken language eye tracking methodology. The goal of the study is to compare children's and adults' use of fictional
information that contradicts lexical and world knowledge to constrain predictions about upcoming language input. Can young children rely upon fantastical facts introduced in the prior discourse to predict upcoming referents, and what is the time course of this process? In other words, how does a fantastical discourse context compete with information in children's stored representations of semantic and world knowledge to constrain their understanding of the situations being described? The competition between semantic and real-world knowledge and discourse information that violates that knowledge can be explored by presenting children with a discourse-final sentence in which the protagonist acts on an object in an unusual way (e.g., wearing boxes on her feet). By examining eye movements in the window of time between the onset of a verb that semantically constrains the referent (e.g., 'putting on') and the onset of the following noun, we can gain insights into children's interpretation of the unfolding sentence. If children rely upon the fantastical discourse to interpret the sentence, we would expect them to demonstrate more anticipatory eye movement to objects that are congruent with the discourse than to objects that are congruent with stored event semantics and world knowledge, but incongruent with the discourse.

We also explore the relationship of predictive language processing in fantastical contexts with other mental functions. Anticipation of events consistent with a fantastical fictional world is likely to require the suppression of stored knowledge based on the stable semantic relationships of the real world. We might therefore expect it to be positively predicted by inhibitory control and negatively predicted by receptive vocabulary and semantic fluency, which reflect strong, well-defined networks of semantic relationships. Constraints on working memory may also limit children's performance by limiting the ability to maintain concurrent interpretations. Studies measuring event-related potentials (ERPs) have shown working memory to predict adults' ability to use rich contextual information to build a message-level representation of linguistic input (Huettig \& Janse, 2016; Wlotko \& Federmeier, 2012), perhaps because it binds knowledge to linguistic and semantic knowledge in space and time.

## Method

## Participants

Sixty-four 7-year-old children (range: 7;0-7;11, Mage: $7 ; 4$ ) and 68 adults (range: 18-35 years; Mage: 25) participated in the current experiment. Seven-year-olds were chosen because they are highly experienced with narratives; because younger (3-5-year-old) children are more often willing to attribute unconventional behaviour to humans (Boerger, 2011); and because pilot testing revealed that they could attend through 16 consecutive eye-tracking trials, and complete the full one-hour procedure without signs of fatigue. Inclusionary criteria were normal or corrected-tonormal vision and no history of diagnosis or treatment of
cognitive, speech, language, hearing, or attentional issues. Children heard English spoken at home more than 75\% of the time. Adults were native speakers of English. Data from 27 additional participants were collected, but not used due to: unsuccessful calibration ( 3 adults, 1 child), no trials that captured eye movements more than $50 \%$ of the time ( 7 adults, 5 children), failed pre-test ( 3 children), lack of attention ( 2 adults, 3 children), and misunderstanding the task ( 1 adult, 2 children).

## Materials

Cartoon images were taken from open-source resources, and the displays accompanying each narrative depicted agents and objects against a white background. Sentences were prerecorded by a female, native Canadian-English speaker.

## Norming of stimuli

Offline tasks with a separate group of 4- and 5-year-old participants were conducted to establish that even younger children could recognize the objects and used the verbs to identify referents in the expected manner. Children were tested at the Ontario Science Centre. The experimenter showed the child a four-object display, provided a label, and asked her to identify the relevant object. $100 \%$ of children recognized all the images presented during the critical sentences ( $\mathrm{N}=8$ per target image). Children were then introduced to pictures of agents (e.g. 'This is Chloe the fairy'), following which they were presented with a fourobject display, and asked (e.g.) 'What will Chloe eat?' Fourand 5-year-olds selected the only referent that was semantically plausible following the verb over $90 \%$ of the time across trials ( $\mathrm{N}=16$ per target image).


Figure 1. Example of stimuli used for the sentence comprehension task.

## Procedure

Sentence comprehension task Participants sat in a stationary chair in front of a computer with a 1920x1200 LCD display. Eye movements were recorded using a Tobii X120 eye-tracker. A nine-point calibration procedure was used to set up tracking of both eyes. In the description phase of the experimental condition ( $\mathrm{N}=327$-year-olds and $\mathrm{N}=34$
adults), participants saw a centrally presented picture of a fantastical agent (e.g., a superhero or monster), and simultaneously heard a story. Story sentences contained referents that were semantically congruent with a preceding verb, and referents that were unusual patients of the preceding verb. For instance: 'Chloe the fairy doesn't have cake for her snack. She has snow for her snack! And Chloe doesn't wear shoes on her feet. She wears boxes on her feet! What is Chloe going to do?' Participants then saw a central fixation cross. In a subsequent test phase, participants saw a display comprising the four mentioned items (e.g. cake, snow, shoes, and box), one placed in each corner of the screen, and heard the critical sentence. In 4 of the 8 experimental trials, the verb in the critical sentence was semantically constraining (e.g., 'Chloe is eating up the snow'). Thus, the verb narrowed to 1 the number of referents in the display that were coherent with children's stored semantic knowledge (henceforth, semantically congruent referent, or SCR: e.g., cake) as well as the number of referents that were coherent with the story information (discourse-congruent referent, or DCR: e.g., snow). In the other 4 experimental trials, the verb did not constrain the referent: e.g. 'Chloe is looking at the snow'. Half of the participants heard a critical sentence based on the first part of the story (e.g., Chloe eating up the snow), and half heard a story based on the second part of the story (e.g., Chloe putting on the box). In 8 filler trials, participants heard that agents 'sometimes' performed expected actions and 'sometimes' performed actions that violated world knowledge, breaking up the pattern in the content and outcomes of the stories and reducing the risk of strategic adjustments. Counterbalancing was in place for the portion of the story that was referenced during the critical sentence, the order in which typical and atypical verb patients were mentioned, the pairing of stories with constraining and neutral verbs, and the location of the DCR on the screen. The location of other objects was randomized.

To confirm that children could recall simple discourse of the type used in the experiment, the sentence comprehension task was preceded by two offline pre-trials in which children were asked a comprehension question in place of the critical sentence. For instance, children heard 'Gordon the gnome doesn't bang on a drum. He bangs on a pillow! And Gordon doesn't dig with a shovel. He digs with a toothbrush!' Once the array of possible referents was displayed, children were asked 'What does Gordon bang on?' Only three children failed to identify the target during one or both of the two comprehension trials, and were excluded from the analysis.

A separate set of 327 -year-olds and 34 adults participated in a control condition in which no story discourse preceded the critical sentence. In this condition, during the description phrase participants saw the picture of the agent, but in place of the story they only heard (e.g.) 'This is Chloe the fairy', followed by the critical sentence.

Following the sentence comprehension task participants completed several individual difference measures. These were drawn largely from the National Institutes of Health

Toolbox (NIH TB) Cognition Battery (McDonald, 2014), which is administered in a computerized adaptive format. Each of the tasks in the Toolbox has been normed and validated for ages 3-85.

Inhibitory control The inhibitory control measure was the NIH TB Flanker Inhibitory Control and Attention Test. The test requires participants to focus on a specific stimulus while inhibiting attention to other stimuli flanking it. Sometimes the middle stimulus is congruent (pointing in the same direction as the flankers) and sometimes incongruent (pointing in the opposite direction). Scores reflect both accuracy and reaction time.

Working memory Working memory was measured using the NIH TB List Sorting Working Memory Test, which involves both storage and manipulation of items in memory. Images of animals and foods are displayed with accompanying audio and written text (e.g., "horse"). The participant is asked to repeat back the items in size order from smallest to largest, within a single dimension (either animals or foods: 1-List) and then on 2 dimensions (foods, then animals: 2-List). The score is equal to the number of items that are both recalled and sequenced correctly.

Receptive vocabulary The receptive vocabulary measure was the NIH TB Picture Vocabulary Test. Participants hear a word and simultaneously see four photographic images on the computer screen. Participants were asked to point to the picture that most closely matches the meaning of the word.

Semantic fluency Semantic fluency was measured using two components of the NEPSY word generation task (Korkman, Kirk, \& Kemp, 2007). Participants were given one minute to produce as many members of a semantic category as they were able. Categories were 'animals' and 'foods and drinks'. Participants received one point for every correct item. Incorrect words and repetitions were excluded.

## Data scoring and analysis

The proportion of time that participants spent looking to each referent was calculated separately for three timewindows corresponding to different speech landmarks, namely the pre-naming window ( 1000 ms prior to verb onset to verb onset), verb window ( 1280 ms prior to noun onset to noun onset) and noun window ( 233 ms after noun onset to 2000 ms after noun onset). Average looking time within these windows was calculated separately for constraining and neutral verb trials, based on gaze position measures assessed every 50 ms .

Raw scores for receptive vocabulary, inhibitory control and attention, and working memory were downloaded from the NIH Toolbox Assessment Center. Two coders viewed video recordings of the semantic fluency (word generation) task. There was excellent agreement between coders' judgments, $r(126)=1, p=<.001$. Disagreements were resolved by a third coder.

Looking time data failed to fit a normal distribution following log and empirical logit transformations (Barr, 2008.) Therefore, all analyses were bootstrapped in SPSS 21 , using 1000 case resamples with replacement from the original dataset and a $95 \%$ percentile confidence interval.

## Results

Control condition In the control condition with no story, both children and adults looked at chance to the DCR ( $p$ $>.05$ ), and above chance to the SCR (children: $t(27)=5.61, p$ $=.001$, adults: $t(29)=4.64, p=.001)$, as expected (Figure 2).

Constraining verb trials Recall that the displays contained two distractor objects in addition to the DCR and SCR. We first ascertained whether participants looked preferentially to the DCR and SCR. Both adults and children did so at a rate significantly above chance (children: $t(31)=8.1, p=.001$; adults: $t(33)=7.82, p=.001$ ). A one-way ANOVA demonstrated that adults' and children's looking behaviour was similar $(F(1,65)=.00, p$ $>.05$.) Thus, both children and adults discounted distractors from their interpretation of the unfolding sentence following verb onset, as seen in Figure 3.

In order to discover how hearing a constraining verb influenced children's and adults' anticipatory processing, we then examined the proportion of time that participants spent looking to the DCR and SCR before and after verb onset. Since constraining and neutral verb trials were identically structured prior to verb onset, we collapsed looking time in the pre-verb window across trial types (constraining and neutral) for this analysis. Paired t-tests demonstrated that upon hearing a constraining verb, the proportion of both children's and adults' looking time to the semantically congruent referent rose relative to its pre-verb baseline (children: $t(31)=-3.54, p=.005$ ); adults: $t(32)=-$ $2.16, p=.04)$. The proportion of adults' looking time to the discourse-congruent referent also rose following the onset of the constraining verb $(t(32)=-3.32, p=002)$; the proportion of children's looking time did not ( $p>.05$ ). Thus, hearing a constraining verb caused adults, but not children, to increase their consideration of the discourse-congruent referent.

We then examined fixation patterns within the verb window separately for the DCR and SCR to determine whether children's and adults' proportions of looking time to these referents differed. They did not significantly differ for either referent (both $p>.05$ ), nor did proportion of looking time to the SCR differ between children and adults in the no-story control condition ( $p>.05$ ). Next, we examined children's and adults' rates of looking against chance, calculated at .2 to account for looks to blank space on the screen, to establish whether both semantic coherence and discourse context influenced participants' interpretations of the sentence prior to hearing the noun. Both children and adults looked to the DCR at a rate above chance during the verb window, suggesting that both age groups relied to some extent on the discourse context to
interpret an unfolding sentence (children: $t(31)=2.59, p=$ .024 ; adults: $t(33)=3.81, p=.002$ ). Children, but not adults, also looked to the SCR at a rate above chance $(t(31)=4.75, p$ $=.001$ ). Thus, taking the verb window as a whole, both children and adults anticipated the DCR as the patient of the constraining verb, while only children anticipated the SCR.


Figures 2\&3. Time-course plots of proportion of looking time to potential referents on experimental trials containing constraining verbs (Figure 2), and on control trials with no discourse context containing constraining verbs (Figure 3).

## Neutral verb trials

Recall that neutral verb trials did not contain semantically congruent objects because the verb (e.g., "look at" was by definition compatible with all display objects. They instead contained two "discourse-congruent" objects in the sense that each story presented the character carrying out two unusual actions. For neutral verb trials, we therefore collapsed the proportion of looking time to both DCRs. As predicted, neither children nor adults made anticipatory looks to the DCRs during the verb window of the critical sentence. Adults and children looked similarly to the DCRs, and neither adults nor children looked to the DCRs at a rate above chance (all $p>.05$ ). This confirmed that patterns in the constraining verb conditions were not simply due to
attentional capture or interest in the images used as the SCR and DCR.

## Individual differences

Pearson correlations between individual difference measures and proportion of looking time to the DCR and SCR in the verb window were examined separately for adults and children. On control trials containing no story, none of the correlations were significant for children or adults (all $p$ $>.05$ ). On experimental trials, none were significant for adults on constraining or neutral verb trials, nor for children on neutral verb trials (all $p>.05$ ). Children's looking time on constraining verb trials was not correlated with inhibitory control, nor with semantic fluency (both $p>.05$ ). Contrary to expectation, children's working memory was negatively correlated with their looking time to the DCR $(r(30)=$ $-.50, p=.004$ ), and children's receptive vocabulary was positively correlated with their looking time to the SCR ( $r(30)=.46, p=.011$ ).

Linear regressions were conducted on the proportion of children's looks to DCR and SCR in the verb window. Working memory significantly predicted children's looking time to the DCR, $b=-.343, t(1,28)=-3.11, p=.004$, and explained approximately $26 \%$ of variance in children's looking time to the $\mathrm{DCR}, \mathrm{R} 2=.256, \mathrm{~F}(1,28)=9.57, p=.004$. Receptive vocabulary significantly predicted children's looking time to the $\mathrm{SCR}, \mathrm{b}=-.031, t(1,30)=2.82, p=.004$, and explained approximately $21 \%$ of variance in children's looking time to the $\mathrm{SCR}, \mathrm{R} 2=.210, \mathrm{~F}(1,30)=7.99, p=.004$.

## Discussion

This study demonstrates that while 7-year-old listeners to fantastical fiction find it difficult to override semantic congruence in favour of discourse congruence, the discourse context nevertheless competes with semantic relationships based on stored knowledge to direct their interpretation of fantastical fictional events. The results also demonstrate that the importance of different types of predictive information appears to change between grade school and adulthood.

In the absence of a story, children as well as adults generated expectations for the object that served as the most typical patient of an immediately preceding verb, as expected (e.g. Mani \& Huettig, 2012). Given a fantastical story, however, both children and adults used the discourse context to guide their appraisal of appropriate verb patients: they anticipated the DCR, which was congruent both with the discourse and with a constraining verb. However, children's anticipation of the discourse-congruent referent diminished over the time course of the verb phrase, whereas adults' anticipation of this object increased over the same time window, suggesting that children began to discount the early expectations that had been generated for a discoursecongruent noun.

Seven-year-olds had difficulty overriding an interpretation of the critical sentence based on stored semantic relationships and real-world knowledge,
generating expectations for the referent that was congruent with their semantic and world knowledge. Adults did not, although some late consideration of the semantically congruent referent is clearly apparent from an examination of the latter half of the verb window in Figure 2. This is not unexpected, as active prediction is often accompanied by a certain degree of thematic priming even when these effects are incongruent with sentence and discourse information (Kukona, Fang, Aicher, Chen, \& Magnuson, 2011).

Children with strong pre-existing networks of semantic relationships, as indexed by receptive vocabulary, found it difficult to override these networks in favour of the discourse context. On average, they showed less consideration of the discourse-congruent referent than did children with smaller receptive vocabularies. It is also possible that children who possess a large vocabulary have a well-developed sense of the need for a clear conceptual basis for any new semantic relationship. In future research, we will ask whether longer and more causally rich stories than those presented in the current study may improve such children's performance. However, on a different measure of semantic network strength (semantic fluency), there was no relationship with the extent to which the DCR was considered. This may be because this kind of word generation task also places demands on executive control: the inhibition of irrelevant information, and the deployment of strategic planning. Given that we did not find a relationship between our measure of inhibitory control and children's looking behaviour, it is possible that in this task, the process of suppressing semantic knowledge may not require inhibition of the prepotent semantically congruent interpretation. Rather, it may involve maintaining representations of both the semantically congruent and discourse-congruent interpretation, and discounting the latter relative to the former.

Contrary to expectation, the poorer children's working memory, the more they relied on the discourse-congruent referent to interpret the unfolding critical sentence. If this finding can be replicated, several possible explanations should be explored in future research. It is possible that weak representations of the discourse entail relatively more attention to the discourse-congruent referent in attempt to support effort towards recall of the role of the object in the story. It is also possible that the stronger the discourse information in children's working memory, the greater the co-activation in memory of stored semantic information, which then remains relatively highly activated in children in comparison to adults.

The real-time processing of fantastical discourse speaks to the interaction of several abilities and knowledge types stored semantic knowledge, vocabulary, working memory, and the moment-by-moment identification of thematic relationships - all of which influence children's mental representations of the events they hear about. This topic provides a rich opportunity to characterize the information processing skills underlying children's language comprehension at the discourse level.

## Acknowledgments

This work was supported by a Discovery Grant from the Natural Sciences and Humanities Research Council of Canada, 'The influence of fantastical discourse context on young children's on-line sentence comprehension', to P. Ganea. R. Lee is supported by an Ontario Graduate Scholarship. The authors thank the Ontario Science Centre for supporting collection of norming data. Thanks also to Naiqi Gabriel Xiao for assistance with data processing, to Amanda Tallon for assistance with testing, to Rafiya Asad, Reem Draz, Abdullah Malik, Angela Nyhout, Thahmina Rahman, and Amanda Tallon for assistance with stimuli and data coding, and to Begüm Özdemir, Gabrielle Strouse, and others in the Language and Learning Lab for helpful feedback.

## References

Altmann, G., \& Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3), 247-264.
Barnes, M. A., Dennis, M., \& Haefele-Kalvaitis, J. (1996). The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. Journal of Experimental Child Psychology, 61(3), 216-241.
Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. Journal of memory and language, 59(4), 457-474.
Boerger, E. A. (2011). "In fairy tales fairies can disappear": Children's reasoning about the characteristics of humans and fantasy figures. British Journal of Developmental Psychology, 29(3), 635-655.
Borovsky, A., Elman, J. L., \& Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. Journal of Experimental Child Psychology, 112(4), 417-436.
Chambers, C. G., \& Juan, V. S. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. Cognition, 108(1), 26-50.
Cook, A. E., \& O’Brien, E. J. (2014). Knowledge Activation, Integration, and Validation During Narrative Text Comprehension. Discourse Processes, 51(1-2), 2649.

Filik (2008). Contextual override of pragmatic anomalies: evidence from eye movements. Cognition, 106(2), 10381046.

Huettig, F., \& Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. Language, Cognition and Neuroscience, 31(1), 8093.

Kamide, Y., Altmann, G., \& Haywood, S. L. (2003). The time-course of prediction in incremental sentence
processing: Evidence from anticipatory eye movements.
Journal of Memory and Language, 49(1), 133-156.
Korkman, M., Kirk, U., \& Kemp, S. (2007). NEPSYSecond edition (NEPSY II). San Antonio: The Psychological Corporation.
Kukona, A., Fang, S.-Y., Aicher, K. A., Chen, H., \& Magnuson, J. S. (2011). The time course of anticipatory constraint integration. Cognition, 119(1), 23-42.
Lynch, J. S., van den Broek, P., Kremer, K. E., Kendeou, P., White, M. J., \& Lorch, E. P. (2008). The Development of Narrative Comprehension and Its Relation to Other Early Reading Skills. Reading Psychology, 29(4), 327-365.
Mani, N., \& Huettig, F. (2012). Prediction during language processing is a piece of cake-But only for skilled producers. Journal of Experimental Psychology: Human Perception and Performance, 38(4), 843-847.
McDonald, Skye (Ed.) (2014). Special series on the Cognition Battery of the NIH Toolbox. Journal of International Neuropsychological Society (JINS), 20(6), 487-651.
Nieuwland, M. S., \& Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. Journal of Cognitive Neuroscience, 18(7), 1098-1111.
Paris, A. H., \& Paris, S. G. (2003). Assessing narrative comprehension in young children. Reading Research Quarterly, 38(1), 36-76.
Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. Journal of Psycholinguistic Research, 32(1), 3-23.
Sharon, T., \& Woolley, J. D. (2004). Do monsters dream? Young children's understanding of the fantasy/reality distinction. British Journal of Developmental Psychology, 22(2), 293-310.
Snedeker, J., \& Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexicalbiases and referential scenes in child and adult sentence processing. Cognitive Psychology, 49(3), 238-299.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., \& Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632-1634.
Trueswell, J. C., Sekerina, I., Hill, N. M., \& Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. Cognition, 73(2), 89-134.
Van de Vondervoort, J. W., \& Friedman, O. (2014). Preschoolers can infer general rules governing fantastical events in fiction. Developmental psychology, 50(5), 15941599.

Wlotko, E. W., \& Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of messagelevel meaning. NeuroImage, 62(1), 356-366.

# The impact of descriptions and incentives on the simultaneous underweighting and overestimation of rare events 

Garston Liang (garston.liang@unsw.edu.au),<br>Emmanouil Konstantinidis (em.konstantinidis@gmail.com), Aba Szollosi (aba.szollosi@gmail.com)<br>Christopher Donkin (christopher.donkin@gmail.com), \& Ben R. Newell (ben.newell@unsw.edu.au)<br>School of Psychology, UNSW, Sydney, NSW 2052, Australia


#### Abstract

We replicate and extend work demonstrating that choice and probability estimation can be dissociated through the coexistence of contradictory reactions to rare events. In the context of experience-based risky choice, we find the simultaneous underweighting of rare events in choice and their overestimation in probability judgement. This tendency persisted in the presence of accurate descriptions of rare event occurrence (Experiment 1), but was attenuated by incentivizing accurate probability estimates (Experiment 2). The implications of these results for popular models of risky choice are briefly discussed.


Keywords: Decisions from experience; probability estimation; risky choice; underweighting.

## Introduction

Decision-making research often uses a description format to present risk information. In the experimental context of monetary gambles, participants are (usually) presented with two options. One option is deemed "safe" as it provides a sure payoff (e.g., a loss of $\$ 3$ with certainty), while the other option (e.g., a loss of $\$ 4$ with 0.8 probability; no loss otherwise) is deemed "risky" as it bears the risk of a rare event (i.e., the 0.2 probability of no loss). In the description format, the participant is explicitly given this risk information and makes a single choice between the two options. For the aforementioned gamble, Kahneman and Tversky (1979) found that participants preferred the risky option. This has been explained by assuming that participants choose as though they subjectively overweight the probability of the rare event. When the same gamble is presented in the gain domain, participants prefer the safe option (\$3 with certainty), consistent with the explanation of overweighting the $20 \%$ chance of receiving nothing.
In experience-based tasks, participants are presented with the same two choices without the aid of written descriptions. In order to learn about the outcome distributions, participants must repeatedly choose between the two options and observe the outcomes over successive rounds of choices. When risk information is presented in this format, choice preferences are consistent with underweighting the rare event. That is, participants prefer the safe option in the above example because the probability of the rare event ( 0.2 probability of no loss) is subjectively underweighted.

## Probability judgement and underweighting

Fox and Hadar (2006) first proposed that erroneous probability judgements could be responsible for underweighting in experience-based choice. If participants judged the probability
of the rare event to be lower than the objective probability, then underweighting in choice would reflect this erroneous judgement. To examine this judgement error hypothesis, a number of studies have asked participants to estimate the probability of the rare event at the end of the experiment (e.g., Hau, Plesak, Kiefer, \& Hertwig, 2008). Using this retrospective method, studies have generally found probability estimates of the rare event to be accurate.

However, retrospective probes of this nature can create disparities between the information used during the task and judgements formed at the end of the task. Camilleri and Newell (2009) found that retrospectively generated probability estimates failed to predict participant choices during the experiment. One remedy for this is to prompt judgement probes throughout the task. In studies assessing awareness, Newell and Shanks (2014) suggested that assessments should be made as immediately as possible as to avoid forgetting and interference from other cognitive processes. This immediacy criterion can be applied to probability judgements as the large number of trials in experience-based tasks (e.g., 50 trials in Hau et al., 2008) may exacerbate any differences between retrospectively and immediately generated judgements.

## The Coexistence Hypothesis

Barron and Yechiam (2009) used a novel trial-by-trial design to examine probability judgements in an experience-based task. In their Experiment 1, participants were repeatedly presented with a choice between a safe option ( -3 points with certainty) and a risky option ( -20 points with $15 \%$ probability; 0 points otherwise). Note that the expected value of both options was equal so a preference for the risky option would be indicative of underweighting the rare event. After an initial phase of choices alone, participants were prompted to estimate the probability of the rare event for the next trial following each choice.

Barron and Yechiam (2009) found that while participants preferred the risky option in their choices, consistent with underweighting the rare event, they overestimated the probability of the rare event (i.e., the probability of the -20 outcome). This inconsistency occurred at the individual level such that the majority of the participants $(15 / 24)$ simultaneously demonstrated underweighting and overestimation of the rare event; a result that was named the coexistence hypothesis.
Barron and Yechiam (2009) also found evidence of opposing recency effects on choices and probability estimates. Following the observation of a rare event, participants were less likely to
select the risky option. In these same trials though, probability estimates of the rare event were lower, demonstrating a reasoning process akin to the gambler's fallacy. This paradoxical result was called the contingent recency effect.

The coexistence hypothesis and the contingent recency effect are problematic for traditional decision models, such as the Two-Stage Choice Model (Fox \& Tversky, 1998), because they present a dissociation between probability judgements and behavior. Most decision models like the Two-Stage model predict that choices can be mapped onto subjective representations of probability following a transformation according to the weighting function of Prospect Theory (Kahneman \& Tversky, 1979). This makes the coexistence hypothesis an interesting anomaly given that overestimation is inconsistent with underweighting in choices.

## Descriptions in decisions from experience

The existing literature separately compares decisions from experience to decisions from description. However, everyday decision making often utilizes a combination of both information sources (e.g., a doctor informed by both her clinical experience and empirical findings).
The limited number of studies examining decisions from experience in the presence of descriptions have produced inconsistent results. While some have found that the presence of descriptions influence choice in decisions from experience (Jessup, Bishara, \& Busemeyer, 2008) others have contended that the descriptions are neglected (Lejarraga \& Gonzalez, 2011). Aiming to resolve these contradictory accounts, WeissCohen et al. (2016) found that participants predominantly relied upon experience to inform their choices but could be influenced by descriptions that provided novel information. This dovetails with the recent finding that the additional presence of descriptive summaries increased underweighting behavior in experience based tasks (Yechiam, Rakow, \& Newell, 2015).

Taken together, these studies demonstrate that descriptions that explicitly provide probability information can influence choice in an experience-based task. However, less is known about how participants represent probability information when both sources of information are available. Given that risky choices are informed by probability judgements when each information source is presented separately, it is important to examine how individuals reconcile probability information in the presence of both descriptions and experience.

Experiment 1 examined the relationship between probability judgement and risky choice. We presented descriptions in an experience-based task that prompted trial by trial probability estimates similar to Barron and Yechiam (2009). We expected that participants given experience alone would show behavior consistent with the coexistence hypothesis (underweighting in choice and overestimation in judgement). Given that the descriptions explicitly stated the veridically experienced rate of rare events, we expected participants given description and experience to accurately estimate the rare event. However, as experience is primarily relied upon to inform choice patterns, we expected underweighting in choice would still emerge.

## Experiment 1

## Method

Participants Eighty undergraduate students ( $M_{\text {age }}=18.90$ years; $S D=1.66,53$ females, 1 other) from UNSW participated in exchange for course credit, and an incentive payment contingent on the participant's choices during the task ( $M=\$ 3.21$ AUD, $S D=0.18$ ).

Materials Participants were presented with two options associated with either a safe (S) or risky (R) distribution as follows:

Safe (S): -3 points with certainty
Risky (R): -20 points with probability $0.15 ; 0$ points otherwise
The expected values were matched, and so a preference for the risky option would be indicative of underweighting. The risky distribution used random sequences of 120 outcomes, of which exactly $15 \%$ (18 outcomes) were rare events. Each sequence was presented to one participant in each condition.

Design Experiment 1 used a between-subjects design with 2 risk information conditions $(n=40)$. The descriptionexperience (DE) condition completed the repeated choice task with descriptions that stated the outcome distribution of each option. For example, the description for the risky option read " $15 \%$ chance of $-20 ; 85 \%$ chance of 0 ". These descriptions remained visible for the duration of the experiment. The experience-only condition (E) completed the task without descriptions.

Procedure Participants were randomly assigned to the E or DE condition and presented with instructions on a computer screen. Participants were told their show-up payment of $\$ 5.00$ AUD had been converted into 1000 points and that their task was to retain as many points as possible. They were prompted to make a choice between the S and R options on the screen (the locations of which were counterbalanced). Following each choice, full feedback for both the selected and forgone option was provided, which remained visible until participants proceeded to the next trial.

After 40 choice-only trials, participants completed an additional probability estimation task following each choice. Specifically, they were asked "What is the probability that -20 will appear in the next round?", and inputted an integer between 0-100 representing a percentage. Feedback for the current trial remained visible during the estimation task. This choice-then-estimation pattern was repeated for the remaining 80 trials after which the participants were debriefed and paid according to the rate of 2 points $=\$ 0.01$.

## Results

Coexistence Hypothesis We found evidence of underweighting in choice in both the DE and E conditions.

[^128]

Figure 1. Risky choice and probability estimates data in Experiment 1. Error bars represent $\pm 1$ SEM.
(A) Mean PR rates (left) and probability estimates (right) by block of 10 trials.
(B) PR rates as a function of preceding trial outcome.
(C) Violin plots of the mean probability estimates in each condition as a function of preceding trial outcome.

Overall, participants preferred the risky option as indicated by mean PR rates (the proportion of risky choices per block) being significantly greater than 0.5 in E condition, $M=0.70, t(39)=$ $6.50, p<.001$, and the DE condition, $M=0.71, t(39)=6.79, p$ $<.001$. As the expected values of the options were equal, this preference for the risky option (i.e., PR rates $>0.5$ ) demonstrates that participants underweighted the rare event (Figure 1A, left panel).
We examined for differences between the conditions using a $2 \times 12$ mixed ANOVA with condition as the between factor and block (of 10 trials) as the within factor2. The main effect of condition was not significant, $F(1,78)=.42, p=.84$. This result indicates that we failed to find a significant difference in PR-rates between the experience-only condition and experience with explicitly stated descriptions.
Given that probability information was explicitly available, we expected that participants would accurately estimate the rare event. However, we found evidence of overestimation. Mean probability estimates were significantly greater than the objective probability of 0.15 in the E condition, $M=0.30, t(39)$ $=6.64, p<.001$, and the DE condition, $M=0.28, t(39)=5.65$, $p<.001$. The effect of condition was non-significant, $F(1,78)=$ $.22, p=.64$ (Figure 1A right panel).

Taken together, these results show the coexistence of inconsistent reactions to rare events in choice and probability judgement even in the presence of accurate descriptive information. Whilst underweighting suggests the rare event had less subjective impact than warranted on choices, overestimation suggests oversensitivity to the rare events’ appearance. Moreover, the majority of participants (27/40 in

[^129]the $E$ condition, 28/40 in the DE condition) exhibited this inconsistent pattern.

Contingent Recency Effect We separated participant responses into trials following a rare event (-20 outcome) and trials following a common event ( 0 outcome). This allowed us to assess for the impact of the most recently observed outcome, using a $2 \times 2$ mixed ANOVA with the outcome of the preceding trial as the within factor and condition as the between factor for both choices and probability estimates.

For choices, we found a significant interaction effect, $F(1$, $78)=4.41, p=.04$ (Figure 1B). This was qualified by a simple effects analysis which revealed a significant effect of preceding trial outcome for the E condition, $F(1,39)=8.25, p=.01$, with a non-significant effect for the DE condition, $F(1,39)=.10, p$ $=.76$. This result shows that in the E condition people were less likely to select the risky option after observing a rare event compared to the common event (PR|-20 = . 60 and $\mathrm{PR} \mid 0=.72$ ), whereas in the DE condition this effect was not significant $(\mathrm{PR} \mid-20=.70$ and $\mathrm{PR} \mid 0=.71)$.
For probability estimates (PE), we found a significant effect of preceding trial outcome, $F(1,78)=22.98, p<.001$ (Figure 1C). Averaged over conditions, participants estimated the probability of rare event to be lower following a rare event compared to following a common event ( $\mathrm{PE} \mid-20=.20$ and $\mathrm{PE} \mid 0$ $=.30)$. This is evidence of negative recency, which suggests that participants believed the chances of rare events in succession were unlikely. The effect of condition was not significant, $F(1,78)=1.00, p=0.32$.

Taken together, these two results demonstrate an inconsistent reaction to the appearance of the rare event at least in the E condition. On the trials following a rare event, participants were less likely to select the risky option. However,
participants simultaneously estimated the rare event to be less likely to occur. This contradictory pattern demonstrates a dissociation between choices and probability estimates, which replicates and extends the pattern observed by Barron and Yechiam (2009). (Note however, that Figure 1C shows different bimodalities in responding - a pattern that awaits further examination).

## Experiment 2

The novel contribution of Experiment 1 is evidence of overestimation even in the presence of descriptions that explicitly state the probability of rare events. This suggests that overestimation emerges from factors related to experience.
Experiment 2 examined two hypotheses about the emergence of overestimation in experience-based choice. The first is that overestimation of the rare event is due to the anticipation of the loss of points. In Experiment 1, participants were incentivized on the outcomes of their choices and so each observation of the rare event was paired with a tangible loss of their incentive payment. This aversive experience may have led participants to overweigh rare events in memory. If this is the case then accuracy might improve if probability estimation was decoupled from experiencing the consequential outcome.

The second hypothesis is that overestimation arises from psychophysical limitations with probability processing. The suggestion here is that although individuals may be proficient in tracking the frequency of events (e.g., Hasher \& Zacks, 1979), they may have difficulties in expressing this information as probabilities (cf. Gigerenzer \& Hoffrage, 1995). This explanation would predict that overestimation is not affected by consequential outcomes but arises due to an inherent incapability of accurately estimating probabilities. Therefore, even if risky choices were divorced from any loss of points, overestimation would still occur.
We tested these hypotheses by separating consequential choices from probability estimation using different incentive schemes. Specifically, participants were incentivized on i) choices only, ii) only the accuracy of probability estimates, or iii) both choices and the accuracy of probability estimates.

## Method

Participants In Experiment 2, 132 students ( $\mathrm{M}_{\mathrm{age}}=19.52$ years; $S D=1.64,100$ females) participated in exchange for course credit, and a performance-based incentive payment.

## Materials As in Experiment 1.

Design Experiment 2 used a between subjects design with 3 groups $(n=44)$ differing in incentive structure.
The choice-incentive group was incentivized on the outcomes of their choices, replicating the E condition from Experiment 1. The probability-incentive condition was incentivized on the accuracy of their probability estimates. In this condition, participants did not choose between the options but instead pressed a separate button that revealed the outcomes from both options simultaneously. The outcomes were financially inconsequential and their task was to track the
outcomes in order to accurately estimate the rare event. Accuracy was calculated as the percentage point deviation from the experienced probability of rare events on each trial. These deviations were tallied at the end of the experiment rather than during the experiment. This was done to avoid giving any feedback about the accuracy of the estimates which could have influenced responses during the probability estimation task. In the dual-incentive condition, participants were incentivized on both the outcomes of their choices and the accuracy of their probability estimates.

Procedure Participants were provided with instructions on a screen that explained the incentive scheme and the objectives of their respective conditions. Across conditions, participants were given a $\$ 5.00$ show-up payment from which an amount would be deducted contingent on their performance in the task.
In the initial stage, participants in the choice-incentive and dual-incentive conditions made repeated choices while those in the probability-incentive condition tracked the outcomes of both options. After 40 trials, all participants also completed the probability estimation task on each trial. Participants were then debriefed and paid accordingly. Choice incentive participants converted their remaining points at a rate of 2 points $=\$ 0.01$ ( $M=\$ 3.19, S D=.19$ ). Probability-incentive participants were penalized 0.05 cents per percentage point deviation ( $M=$ $\$ 4.44, S D=.31$ ). For the dual-incentive participants, their show-up payment was divided in half with each half paid according to different incentive structures. The "choice" half of the payment was calculated by converting their remaining points at a rate of 1 point $=0.5$ cents. The "probability estimate" half was penalized at 0.025 cents per percentage point deviation.

## Results

Coexistence hypothesis PR-rates were again significantly greater than .5 in the choice-incentive condition, $M=.72, t(43)$ $=7.28, p<0.001$, and the dual incentive condition, $M=.76$, $t(43)=12.98, p<0.001$ (Figure 2A, left panel). This suggests participants underweighted the rare event in their choices. A mixed $2 \times 12$ ANOVA, with condition as the between and block as the within factor, found that the effect of condition was not significant, $F(1,86)=1.46, p=.23$.
Participants overestimated the rare event in the choiceincentive, $M=0.33, t(43)=8.99, p<.001$, and dual-incentive conditions, $M=0.26$, $t(43)=4.50, p<.001$, shown by mean probability estimates significantly greater than the objective probability of 0.15 (Figure 2A, right panel). By comparison, mean estimates in the probability-incentive condition did not differ significantly from the objective probability of $0.15, M=$ $0.16, t(43)=1.09, p=.28)$. Using a $3 \times 12$ mixed ANOVA, we found a significant main effect of condition, $F(2,129)=17.51$, $p<0.001$. Post-hoc Scheffé comparisons showed that all groups significantly differed from each other. The interaction effect was not significant, $F(10.85,699.63)=1.14, p=.33$.

In summary, mean probability estimates were highest in the choice-incentive condition where participants were incentivized for the outcomes of their choices alone. However,


Figure 2. Risky choice and probability estimate data in Experiment 2. Error bars represent $\pm 1$ SEM. (A) PR rates (left) and probability estimates (right) by block of 10 trials.
(B) PR rates as a function of preceding trial outcome.
(C) Violin plots of the mean probability estimates in each condition as a function of preceding trial outcome.
if participants were additionally incentivized for accuracy as in the dual-incentive condition, estimate accuracy was improved. Mean probability estimates were most accurate when consequential outcomes were removed all together.

Taken together, these results replicate the coexistence hypothesis of underweighting in choice and overestimation in probability judgement. We found that 34/44 participants in the choice-incentive condition, and 26/44 participants in the dualincentive condition demonstrated this pattern of responding.

Contingent Recency Effects A $2 \times 2$ mixed ANOVA found a significant effect of preceding trial outcome for risky choices (Figure 2B). This showed that averaged over conditions, PRrates were lower following a rare outcome than following a common outcome (PR|-20 = . 66 and $\mathrm{PR} \mid 0=.76$ ), $F(1,86)=$ 18.91, $p<.001$.

The mean probability estimates contingent on the preceding trial outcome are presented in Figure 2C. We used a $3 \times 2$ mixed ANOVA with preceding trial outcome as the within factor to examine for recency effects. A significant effect of preceding trial outcome was found, $F(1,129)=4.42, p=.04$, which was qualified by a significant interaction between condition and preceding trial outcome, $F(2,129)=5.34, p=$ .01. Simple effects analysis revealed probability estimates in the choice-incentive condition were lower after observing a rare outcome than after observing a common event ( $\mathrm{PE} \mid-20=$ .25 and $\mathrm{PE} \mid 0=.34), F(1,129)=13.33, p<0.001$.

By comparison, the effect of preceding trial outcome was not observed in the probability-incentive condition (PE|-20 $=.18$ and $\mathrm{PE} \mid 0=.16), F(1,129)=0.90, p=0.35$, or the dualincentive condition $(\mathrm{PE} \mid-20=.23$ and $\mathrm{PE} \mid 0=0.25), F(1,129)=$
$0.88, p=0.35$. Taken together, the negative recency effect of preceding trial outcome was only found when participants were not incentivized to estimate accurately. (Note again however, that Figure 2C shows bimodality in responding - clustered around 50 and 0 - a pattern that awaits further examination).

## General Discussion

Across two experiments, we found the coexistence of underweighting in choice and overestimation in probability judgements at the individual level. Furthermore, inconsistency is evident in the trials immediately following rare events. Experiment 1 replicated the coexistence hypothesis in the presence of accurate descriptions. We failed to find a difference in choices and probability estimates between participants that received descriptions and those that did not. Experiment 2 used incentive schemes to show that overestimation emerges in the presence of consequential outcomes. We postulate that attention to the probabilities attenuated the degree of overestimation.

The results of our experiment suggest consequential outcomes biased attention away from probability tracking. Kahneman (1973) defined attention as a limited resource that is allocated according to the demands of the task. Overestimation may have occurred in anticipation of the loss of points associated with the rare event, driving attention towards the outcomes themselves. With respect to the dual-incentive condition, the presence of consequential outcomes in the choice task meant fewer attentional resources could be allocated to probability tracking. This competition for attentional resources between the two tasks would explain why
the degree of overestimation was attenuated, but still remained in the presence of consequential outcomes.

Our overestimation results are incompatible with the judgement error hypothesis (Fox \& Hadar, 2006). Therefore, the coexistence of overestimation and underweighting suggests that probability judgement and choice may reflect separate cognitive processes. The distinction between choice and judgement resembles the comparison between online and memory-based strategies. Hastie and Park (1986) distinguished between two types of strategies based on how information is processed to form a judgement. Online strategies involve step-by-step information processing whereby a judgement is continually updated with new information. By comparison a memory-based strategy involves a discrete instance in which all relevant information is recalled from memory to form a judgement (Haberstroh \& Betsch, 2002).

We propose that the choice process resembles an online strategy that involves continually updating a small mental sample of outcomes with recently observed outcomes. This is compatible with the explanation that underweighting emerges from small mental samples of outcomes (i.e., calculating the expected value of the last 5 observed outcomes, cf. Erev \& Roth, 2014). Concurrently, probability judgements resemble a memory-based strategy where aversive rare events are overweighted in memory, resulting in overestimation (e.g., Tversky \& Kahneman, 1973).

Separate processes for choice and judgement would be consistent with our findings of the contingent recency effects. In trials following the rare event, participants were less likely to choose the risky outcome yet, they paradoxically estimated a lower probability of the occurrence of the rare event. An online strategy for choices would involve a trial-by-trial updating process that incorporates each newly observed outcome into the decision process. Whilst the risky option usually provided the more attractive outcome (i.e., no loss of points), the occasional appearance of the rare event meant that the small sample from the risky option was momentarily less attractive than the safe option. This explains the reduced tendency to select the risky option after observing a rare event. In these very trials, a memory-based strategy that employs the gambler's fallacy would explain the lower probability estimates. Given a rare event on the preceding trial, participants may have reasoned that "lightning does not strike twice". Therefore, the gambler's fallacy may have served as a memory heuristic to simplify the more cognitively demanding memory-based estimation process.

In summary, we have shown the impact of descriptions and incentives on the simultaneous overestimation and underweighting of rare events.

## Acknowledgments

This work was supported by the Australian Research Council (DP140101145; DP160101186).

## References

Barron, G., \& Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the
contingent recency effect. Judgment and Decision Making, 4, 447-460.
Camilleri, A. R., \& Newell, B. R. (2009). The role of representation in experience-based choice. Judgment and Decision Making, 4, 518-529.
Erev, I., \& Roth, A. E. (2014). Maximization, learning, and economic behavior. Proceedings of the National Academy of Sciences, 111,10818-10825.
Fox, C. R., \& Hadar, L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber \& Erev (2004). Judgment and Decision Making, 1, 159-161.
Fox, C. R., \& Tversky, A. (1998). A belief-based account of decision under uncertainty. Management Science, 44, 879895.

Gigerenzer, G., \& Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.
Haberstroh, S., \& Betsch, T. (2002). Online strategies versus memory-based strategies in frequency estimation. ETC. Frequency Processing and Cognition, 205-220.
Hasher, L., \& Zacks, R. T. (1979). Automatic and effortful processes in memory. Journal of Experimental Psychology: General, 108, 356-388.
Hastie, R., \& Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. Psychological Review, 93, 258-268.
Hau, R., Pleskac, T. J., Kiefer, J., \& Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. Journal of Behavioral Decision Making, 21, 493-518.
Jessup, R. K., Bishara, A. J., \& Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. Psychological Science, 19, 1015-1022.
Kahneman, D. (1973). Attention and Effort. Englewood Cliffs, NJ: Prentice-Hall.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 263-291.
Lejarraga, T., \& Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. Organizational Behavior and Human Decision Processes, 116, 286-295
Newell, B.R. \& Shanks, D.R. (2014). Unconscious influences on decision making: A critical review. Behavioral and Brain Sciences, 37, 1-63
Tversky, A., \& Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207-232.
Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., \& Harvey, N. (2016). Incorporating conflicting descriptions into decisions from experience. Organizational Behavior and Human Decision Processes, 135, 55-69.
Yechiam, E., Rakow, T., \& Newell, B. R. (2015). Super-underweighting of rare events with repeated descriptive summaries. Journal of Behavioral Decision Making, 28, 67-75.

# An automatic method for discovering rational heuristics for risky choice 

Falk Lieder ${ }^{1}$ (falk.lieder@berkeley.edu)<br>Paul M. Krueger ${ }^{1}$ (pmk@berkeley.edu)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology,University of California Berkeley, Berkeley, CA 94720 USA<br>${ }^{1}$ These authors contributed equally.


#### Abstract

What is the optimal way to make a decision given that your time is limited and your cognitive resources are bounded? To answer this question, we formalized the bounded optimal decision process as the solution to a meta-level Markov decision process whose actions are costly computations. We approximated the optimal solution and evaluated its predictions against human choice behavior in the Mouselab paradigm, which is widely used to study decision strategies. Our computational method rediscovered well-known heuristic strategies and the conditions under which they are used, as well as novel heuristics. A Mouselab experiment confirmed our model's main predictions. These findings are a proof-of-concept that optimal cognitive strategies can be automatically derived as the rational use of finite time and bounded cognitive resources.


Keywords: Decision-Making; Heuristics; Bounded Rationality; Strategy Selection; Rational Metareasoning

## Introduction

Some situations require us to decide quickly whereas others call for careful consideration of all available options and potential consequences. People seem to master this challenge by choosing adaptively from a toolbox of diverse decision strategies (Payne, Bettman, \& Johnson, 1988; Gigerenzer \& Selten, 2002). This toolbox is assumed to include fast-andfrugal heuristics (Gigerenzer \& Goldstein, 1996) as well as slower and more effortful strategies. Fast-and-frugal heuristics include Take-The-Best (TTB), which chooses the alternative that is favored by the most predictive attribute and ignores all other attributes, satisficing (SAT) (Simon, 1956), which chooses the first alternative whose expected value exceeds some threshold, and random choice; slower strategies include the Weighted-Additive Strategy (WADD), which computes all gambles' expected values based on all possible payoffs. Except for WADD, all of these strategies are heuristics: they solve some problems very efficiently but err on others.

The systematic errors that result from people's use of heuristics are inconsistent with classic notions of rationality such as logic, probability theory, and expected utility theory (Tversky \& Kahneman, 1974). Making good decisions is remarkably constrained: decisions have to be made in a finite amount of time, people's cognitive resources are limited, and maximizing expected utility entails intractable computational problems. This makes expected utility theory an unrealistically high bar for human rationality. According to a more realistic normative standard, people should decide in a way that makes the best possible use of their limited cognitive resources (Griffiths, Lieder, \& Goodman, 2015). Previous research has applied this resource-rational approach to numer-
ical estimation (Lieder, Griffiths, \& Goodman, 2012), availability biases (Lieder, Hsu, \& Griffiths, 2014), and strategy selection (Lieder, Plunkett, et al., 2014). However, this approach has not been applied to the domain in which heuristics have perhaps been studied in greatest detail: multi-alternative risky choice. Work on risky choice suggests that people adaptively switch between multiple different strategies depending on how much time is available and whether one of the outcomes is much more likely than the others (Payne et al., 1988). Yet, it remains unclear how people's decision processes compare to resource-rational behavior.

To answer these questions, we model the decision process as a sequence of costly computations and formalize the optimal decision process as the solution to a meta-level Markov decision process. We combine this theory with an algorithm for approximating the optimal solution to create a computational method that can automatically derive optimal cognitive strategies. These rational heuristics can be interpreted as a fair normative standard for human decision making that takes into account that people's time is costly and that their cognitive resources are bounded. We are optimistic that this novel approach will lead to new insights about how decisionmakers cope with limited time and bounded computational resources, and advance the debate about human rationality.

We illustrate our approach in multi-alternative risky choice and test its predictions using the Mouselab paradigm that is widely used to study decision strategies (Johnson, Payne, Bettman, \& Schkade, 1989). Two known heuristics, TTB and random choice, emerged from our theory as resourcerational strategies for low-stakes decisions with high and low dispersion of their outcome probabilities, respectively. In addition, our computational method discovered a novel heuristic that combines TTB with satisficing. Our experiment demonstrated that people do indeed use the newly discovered heuristic and confirmed our rational model's predictions of when people use which strategy: people used simple heuristics more frequently when the stakes were low, employed fast-and-frugal heuristics less frequently when all outcomes were almost equally likely (low dispersion), and invested more time and effort when the stakes were high. This is the first demonstration that rational meta-reasoning can be used to automatically discover decision strategies used by people.

## Background

We will formulate our theory using the mathematical frameworks of Markov decision processes, bounded optimality, and
rational metareasoning, introduced in this section.

## Markov Decision Processes

Each sequential decision problem can be modeled as a Markov Decision Process (MDP)

$$
\begin{equation*}
M=\left(\mathcal{S}, \mathcal{A}, T, \gamma, r, P_{0}\right) \tag{1}
\end{equation*}
$$

where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $T\left(s, a, s^{\prime}\right)$ is the probability that the agent will transition from state $s$ to state $s^{\prime}$ if it takes action $a, 0 \leq \gamma \leq 1$ is the discount factor on future rewards, $r\left(s, a, s^{\prime}\right)$ is the reward generated by this transition, and $P_{0}$ is the probability distribution of the initial state $S_{0}$ (Sutton \& Barto, 1998). A policy $\pi: \mathcal{S} \mapsto \mathcal{A}$ specifies which action to take in each of the states. The expected sum of discounted rewards that a policy $\pi$ will generate in the MDP $M$ starting from a state $s$ is known as its value function

$$
\begin{equation*}
V_{M}^{\pi}(s)=\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} \cdot r\left(S_{t}, \pi\left(S_{t}\right), S_{t+1}\right)\right] \tag{2}
\end{equation*}
$$

The optimal policy $\pi_{M}^{\star}$ maximizes the expected sum of discounted rewards, that is

$$
\begin{equation*}
\pi_{M}^{\star}=\arg \max _{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{\dagger} \cdot r\left(S_{t}, \pi\left(S_{t}\right), S_{t+1}\right)\right] \tag{3}
\end{equation*}
$$

## Bounded optimality and rational metareasoning

People and robots have to make decisions in a limited amount of time and with bounded cognitive resources. Given that these resources are scarce, which strategy should a decisionmaker employ to use its resources most effectively? The theory of bounded optimality and rational metareasoning (Russell \& Wefald, 1991; Russell \& Subramanian, 1995) was developed to answer this question for rational agents with limited performance hardware. It frames this problem as selecting computations so as to maximize the sum of the rewards of resulting decisions minus the costs of the computations involved.

Concretely, the problem of choosing computations optimally can be formulated as a meta-level MDP (Hay, Russell, Tolpin, \& Shimony, 2012). A meta-level MDP

$$
\begin{equation*}
M_{\mathrm{meta}}=\left(\mathcal{B}, \mathcal{C}, T_{\mathrm{meta}}, r_{\mathrm{meta}}\right) \tag{4}
\end{equation*}
$$

is a Markov decision process whose actions $\mathcal{C}$ are cognitive operations, its states $\mathcal{B}$ represent the agent's probabilistic beliefs, and the transition function $T_{\text {meta }}$ models how cognitive operations change the agent's beliefs. In addition to a set of computations $C$ that update the agent's belief, the cognitive operations also include the meta-level action $\perp$ that terminates deliberation and translates the current belief into action. The meta-level state $b_{t}$ encodes the agent's probabilistic beliefs about the domain it is reasoning about.The meta-level reward function $r_{\text {meta }}$ captures the cost of thinking (Shugan, 1980) and the reward $r$ the agent expects to receive from the environment when it stops deliberating and
takes action. The computations $C$ do not yield any external reward. Their only effect is to update the agent's beliefs. Hence, the meta-level reward for performing a computation $c \in C$ is $r_{\text {meta }}\left(b_{t}, c\right)=-\operatorname{cost}(c)$. By contrast, terminating deliberation and taking action $(\perp)$ does not update the agent's belief. Instead, its value lies in the anticipated reward for taking action, that is

$$
\begin{equation*}
r_{\text {meta }}\left(b_{t}, \perp\right)=\arg \max _{a} b_{t}^{(\mu)}(a) \tag{5}
\end{equation*}
$$

where $b_{t}^{(\mu)}(a)$ is the expected reward of taking action $a$ according to the belief $b_{t}$.

## Adaptive strategy selection in risky choice

Consistent with rational metareasoning, people flexibly adapt their decision processes to the structure of the problem they face. Concretely, Payne et al. (1988) found that people use fast-and-frugal heuristics, like TTB, more frequently when they are under time pressure and when one outcome is much more likely than the others. In this research, participants were given the choice between gambles $g_{1}, \cdots, g_{n}$. Each gamble was defined by the payoffs it assigns to each of four possible outcome whose probabilities are known $(P(O))$. Participants could inspect a payoff matrix $V_{o, g}$ with one row for each outcome $o$ and one column for each gamble $g$. Critically, each payoff is only revealed when the participant clicks on the corresponding cell of the payoff matrix using a mouse; this task is hence referred to as the Mouselab paradigm (see Figure 1).

The adaptiveness of people's strategy choices in the Mouselab paradigm suggests that their decision processes are efficient and effective. But it is difficult to test whether they are optimal, because it is unclear what it means to decide optimally when one's time is valuable and one's cognitive resources are limited. To clarify this, the following section develops a normative theory of resource-bounded decision making in the Mouselab paradigm.

## Boundedly-optimal decision-making

To model the meta-decision problem posed by the Mouselab task, we characterize the decision-maker's belief state $b_{t}$ by probability distributions on the expected values $e_{1}=\mathbb{E}\left[v_{O, g_{1}}\right], \cdots e_{n}=\mathbb{E}\left[v_{O, g_{n}}\right]$ of the $n$ available gambles $g_{1}, \cdots, g_{n}$. Furthermore, we assume that for each element $v_{o, g}$ of the payoff matrix $V$ there is one computation $c_{o, g}$ that inspects the payoff $v_{o, g}$ and updates the agent's belief about the expected value of the inspected gamble according to Bayesian inference. Since the entries of the payoff matrix are drawn from the normal distribution $\mathcal{N}\left(\bar{v}, \sigma_{v}^{2}\right)$, the resulting posterior distributions are also Gaussian. Hence, the decision-maker's belief state $b_{t}$ can be represented by $b_{t}=\left(b_{t, 1}, \cdots, b_{t, n}\right)$ with

$$
\begin{equation*}
b_{t, g}=\left(b_{t, g}^{(\mu)}, b_{t, g}^{\left(\sigma^{2}\right)}\right) \tag{6}
\end{equation*}
$$

where $b_{t, g}^{(\mu)}$ and $b_{t, g}^{\left(\sigma^{2}\right)}$ are the mean and the variance of the probability distribution on the expected value of gamble $g$ of the belief state $b_{t}$.

Given the set $O_{t}$ of the indices $\left(k_{o}^{(1)}, k_{g}^{(1)}\right), \cdots,\left(k_{o}^{(t)}, k_{g}^{(t)}\right)$ of the $t$ observations made so far, the means and variances characterizing the decision-maker's beliefs are given by

$$
\begin{align*}
& b_{t, g}^{(\mu)}=\sum_{(o, g) \in O} p(o) \cdot v_{o, g}+\sum_{(o, g) \notin O} p(o) \cdot \bar{v}  \tag{7}\\
& b_{t, g}^{\left(\sigma^{2}\right)}=\sum_{(o, g) \notin O} p(o)^{2} \cdot \sigma_{v}^{2} . \tag{8}
\end{align*}
$$

The meta-level transition function $T\left(b_{t}, c_{o, g}, b_{t+1}\right)$ encodes the probability distribution on what the updated means and variances will be given the observation of a payoff value $V_{o, g}$ sampled from $\mathcal{N}\left(\bar{v}, \sigma_{v}^{2}\right)$. The meta-level reward for performing the computation $c_{o, g} \in C$ encodes that acquiring and processing an additional piece of information is costly. We assume that the cost of all such computations is an unknown constant $\lambda$. The meta-level reward for terminating deliberation and taking action is $r_{\text {meta }}\left(b_{t}, \perp\right)=\max _{g} b_{t}^{(\mu)}(g)$.

## Approximating the optimal meta-level policy: Bayesian value function approximation

Unfortunately, computing the optimal policy for the metalevel MDP defined above is intractable. However, it can be approximated using methods from reinforcement learning. We initially used the semi-gradient SARSA algorithm (Sutton \& Barto, 1998) with limited success. We therefore developed a new algorithm that replaces the gradient descent component of that algorithm by Bayesian linear regression.

Our algorithm learns a linear approximation to the metalevel Q-function

$$
\begin{equation*}
Q_{\mathrm{meta}}(b, c) \approx \sum_{k} w_{k} \cdot f_{k}(b, c) \tag{9}
\end{equation*}
$$

whose features $\mathbf{f}$ include a constant, features of the belief state $b_{t}$, and features of the computation $c_{t}$. The features of the belief state were the expected value of the maximum of the gambles' expected values $\left(\mathbb{E}\left[\max _{g} E_{g} \mid b_{t}\right]\right)$ and the decision-maker's uncertainty about it ( $\left.\sqrt{\operatorname{Var}\left[\max _{g} E_{g} \mid b_{t}\right]}\right)$. The largest posterior mean $\left(\max _{g} b_{t, g}^{(\mu)}\right.$ ) and its associated uncertainty $\left(\sqrt{\mu_{t, g^{\star}}^{\left(\sigma^{2}\right)}}\right.$ where $g^{\star}=\arg \max _{g} b_{t, g}^{(\mu)}$ ), the second largest posterior mean and the decision-maker's uncertainty about it, and the expected regret $\mathbb{E}\left[\operatorname{regret}(g) \mid b_{t}\right]$ that the decision-maker would experience if they chose based on their current belief (where $\operatorname{regret}(g)=\max _{g} E_{g}-\max _{g} b_{t, g}^{(\mu)}$ for $E_{i} \sim \mathcal{N}\left(b_{t, i}^{(\mu)}, b_{t, i}^{(\sigma)}\right)$ for all gambles $\left.i\right)$. The features of the computation $c_{o, g}$ were its myopic value of computation $\left(\operatorname{VOC}\left(b_{t}, c_{o, g}\right)\right.$; see Russell \& Wefald, 1991), the current uncertainty about the expected value of the inspected gamble $\left(b_{t, g}^{(\sigma)}\right)$, the probability of the inspected outcome, the difference between the largest posterior mean and the posterior mean of the inspected outcome, a binary variable indicating whether the computation acquired new information, and the expected reduction in the expected regret $\operatorname{ER}(b)$ minus its cost (i.e. $\mathbb{E}\left[\operatorname{ER}\left(B_{t+1}\right) \mid b_{t}, c\right]-\operatorname{ER}\left(b_{t}\right)-\lambda$, where $B_{t+1}$ is the
unknown belief state resulting from performing computation $c$ in belief state $b_{t}$ and $\left.\operatorname{ER}\left(b_{t}\right)=\mathbb{E}\left[\operatorname{regret}\left(\arg \max _{g} b_{t, g}^{(\mu)}\right) \mid b_{t}\right]\right)$.

The weights $\mathbf{w}$ are learned by Bayesian linear regression of the bootstrap estimate $\hat{Q}(b, c)$ of the meta-level value function onto the features $\mathbf{f}$. The bootstrap estimator is

$$
\begin{equation*}
\hat{Q}\left(b_{t}, c_{t}\right)=r_{\text {meta }}\left(b_{t}, c_{t}\right)+\hat{w}_{t}^{\prime} \cdot \mathbf{f}\left(b_{t+1}, c_{t+1}\right) \tag{10}
\end{equation*}
$$

where $\hat{w}_{t}$ is the posterior mean on the weights $w$ given the observations from the first $t$ trials, and $\mathbf{f}\left(b_{t+1}, c_{t+1}\right)$ is the feature vector characterizing the subsequent belief state $b_{t+1}$ and the computation $c_{t+1}$ that will be selected in it.

Given the learned posterior distribution on the feature weights $\mathbf{w}$, the next computation $c$ is selected by contextual Thompson sampling (Agrawal \& Goyal, 2013). Specifically, to make the $t^{\text {th }}$ meta-decision, a weight vector $\tilde{w}$ is sampled from the posterior distribution of the weights given the series of meta-level states, selected computations, and the resulting value estimates experienced so far, that is

$$
\tilde{w} \sim P\left(\mathbf{w} \mid\left(b_{1}, c_{1}, \hat{Q}\left(b_{1}, c_{1}\right)\right), \cdots,\left(b_{k-1}, c_{k-1}, \hat{Q}\left(b_{k-1}, c_{k-1}\right)\right)\right) .
$$

The sampled weight vector $\tilde{w}$ is then used to predict the Q-values of each available computation $c \in \mathcal{C}$ according to Equation 9. Finally, the computation with the highest predicted Q -value is selected.

## Application to Mouselab experiment

As a proof of concept, we applied our approach to the Mouselab experiment described below. The experiment comprises $50 \%$ high-stakes problems and $50 \%$ low-stakes problems. Since participants are informed about the stakes, we learned two separate policies for high-stakes and low-stakes problems, respectively. Half of each of those problems had nearly uniform outcome probabilities ("low dispersion") and for the other half one outcome was much more likely than all others combined ("high dispersion"). The parameters of the simulated environment were exactly equal to those of the experiment described below. Our model assumed that people play each game as if they receive the payoff of the selected gamble. We estimated the cost per click to be about $\lambda=3$ cents. This value was selected to roughly match the average number of acquisitions observed in the experiment.

To approximate the optimal meta-decision policy for this task, we ran our feature-based value function approximation method for 4000 low-stakes training trials and 4000 highstakes training trials, respectively.

## Model predictions

The meta-level MDP described above formalizes the costs and benefits of acquiring and processing additional pieces of information: acquiring additional information can improve the decision that will be taken later on but also incurs an immediate cost. Hence, the optimal solution approximated by our computational method executes a cognitive operation or sequence of operations if and only if the resulting improvement in decision quality is larger than cost of those


Figure 1: The Mouselab paradigm, showing an example sequence of clicks generated by the SAT-TTB strategy, which was discovered through approximate rational metareasoning.
operations. Intuitively, this means that the decision process prescribed by our model achieves the optimal tradeoff between decision quality versus decision time and mental effort. This tradeoff depends on the stakes of the decision such that higher stakes usually warrant more deliberation. Likewise, since processing probable outcomes is more likely to improve the quality of the resulting decision than processing improbable outcomes, we expect our model to prioritize probable outcomes over less probable outcomes-especially in high-dispersion trials.

Our computational method automatically discovered strategies that people are known to use in the Mouselab paradigm as well as a novel strategy that has not been reported yet. Our method rediscovered TTB, WADD, and the random choice strategy. In addition, it discovered a new hybrid strategy that combines TTB with satisficing (SAT-TTB). Like TTB, SAT-TTB inspects only the payoffs for the most probable outcome. But unlike TTB and like SAT, SAT-TTB terminates as soon as it finds a gamble whose payoff for the most probable outcome is high enough. On average, this value was about $\$ 0.15$ when the payoffs ranged from $\$ 0.01$ to $\$ 0.25$ (i.e., low-stakes trials). Figure 1 illustrates this strategy.

Furthermore, our model makes intuitive predictions about the contingency of people's choice processes on stakes and outcome probabilities. First, our model predicts that people should use fast-and-frugal heuristics more frequently in highdispersion trials. This is intuitively rational because high dispersion means that one outcome is much more likely than all others and fast-and-frugal heuristics ignore all outcomes except for the most probable one(s). Concretely, our model generated TTB as the strategy of choice for $100 \%$ of the highdispersion problems with low-stakes, but for low-dispersion problems with low-stakes the model considered the random choice strategy to be optimal in the majority ( $56 \%$ ) of cases; it used the SAT-TTB hybrid strategy for $24 \%$ of such trials, and it indicated the TTB strategy only for the remaining $20 \%$.

Second, our model predicts that people should use simple heuristics, like TTB, SAT-TTB, and random choice, primarily when the stakes are low. This, too, is intuitively rational because fast and frugal heuristics tend to be faster but less ac-
curate than more effortful strategies. Our model used these heuristics for $100 \%$ of the low-stakes problems. But for highstakes problems, the model never used any of these or other frugal strategies. Instead, the model typically inspected the vast majority of all cells ( $24.8 / 28$ for low-dispersion problems and $23.7 / 28$ for high-dispersion problems). The few cells that it did not inspect were mostly the payoffs of lesslikely outcomes of the best gamble when its inspected payoffs for the most likely outcome(s) were high enough to guarantee that it would be optimal.

Third, our model predicts that when the stakes are high people should invest more time and effort $(F(1,396)=$ $9886.8, p<0.0001$ ) to reap a higher fraction of the highest possible expected payoff $(F(1,339)=135.24, p<0.0001)$. This, too, is consistent with the rational speed-accuracy tradeoff inherent in our theory. When the stakes were low the model inspected only 4.3 payoffs on average and reaped only $87 \%$ of the possible reward; but when the stakes were high the model inspected 24.3 of the 28 possible payoffs and reaped $99 \%$ of the best expected payoff on average. In $97 \%$ of these trials, the model achieved this near-maximal performance while being more efficient and more frugal than the WADD strategy which it employed for only $3 \%$ of these problems.

## Experimental test of novel predictions

To test the predictions of our model, we conducted a new Mouselab experiment that manipulated the stakes and dispersion of outcome probabilities within subjects in an identical manner to the model simulations.

## Methods

Participants We recruited 200 participants on Amazon Mechanical Turk. The experiment took about 30min. Participants received a base pay of $\$ 1.50$, and one of their twenty winnings was selected at random and awarded as a bonus to motivate them to take each trial seriously (avg. bonus $\$ 3.53$ ).

Procedure Participants performed a variation of the Mouselab task (Payne et al., 1988). Participants played a series of 20 games divided into two blocks. Figure 1 shows a screenshot of one game. Every game began with a $4 \times 7$ grid of occluded payoffs: there were seven gambles to choose from (columns) and four possible outcomes (rows). The occluded value in each cell specified how much the gamble indicated by its column would pay if the outcome indicated by its row occurred. The outcome probabilities were described by the number of balls of a given color in a bin of 100 balls, from which the outcome would be drawn. For each trial, participants were free to inspect any number of cells before selecting a gamble, with no time limit. The value of each inspected cell remained visible onscreen for the duration of the trial. Upon selecting a gamble, the resulting reward was displayed.

Experimental design The experiment used a $2 \times 2$ within subjects design. Each block of ten trials was either low-stakes
or high-stakes, with block order randomly counterbalanced across participants. In games with low-stakes, the possible outcomes ranged from $\$ 0.01$ to $\$ 0.25$, while in high-stakes games, outcomes ranged from $\$ 0.01$ to $\$ 9.99$. The payoffs were drawn from a truncated normal distribution with mean $\frac{r_{\text {max }}+r_{\text {min }}}{2}$ and standard deviation $0.3 \cdot\left(r_{\text {max }}-r_{\text {min }}\right)$. Within each block, there were five low-dispersion trials and five highdispersion, ordered randomly. In low-dispersion trials, the probability of each of the four outcomes ranged from 0.1 to 0.4 , whereas in high-dispersion trials, the probability of the most likely outcome ranged from 0.85 to 0.97 .

Strategy identification We identified six different decision strategies, in humans and in simulations, using the following definitions: TTB was defined as inspecting all cells in the row corresponding to the most probable outcome and nothing else. SAT occurs when one gamble's payoffs are inspected for all four outcomes, potentially followed by the inspection of all outcomes of another gamble, and so on, but leaving at least one gamble unexamined. The hybrid strategy, SATTTB, was defined as inspecting the payoffs of 1 to 6 gambles for the most probable outcome and not inspecting payoffs for any other outcome. TTB2 was defined as inspecting all fourteen cells of the two most probable outcomes, and nothing else. WADD was defined as inspecting all 28 cells column by column.Random decisions mean zero samples were taken.

## Results

Our process tracing data confirmed that people do indeed use the SAT-TTB strategy discovered by our model. Table 1 shows the frequency of various decision strategies, for each of the four different types of trials. Out of 4000 trials across all participants, TTB was the most common strategy overall, accounting for $25.3 \%$ of all trials. SAT-TTB was the second most common strategy among those we examined: participants employed this strategy on $10.7 \%$ of all trials. In $8.0 \%$ of trials participants chose randomly without making any observations-mostly during low-stakes games. Interestingly, we also observed a second novel strategy that we call Take-The-Best-Two (TTB2). This strategy inspects all gambles' payoffs for the two most probable outcomes, and was used in $6.3 \%$ of trials. The WADD strategy occurred in 4.5\% of trials. Finally, the SAT strategy was used in $3.1 \%$ of games.

Consistent with our model's first prediction, people used TTB more frequently when the dispersion was high $\left(\chi^{2}(1)=\right.$ $897.9, p<0.0001)$. Consistent with our model's second prediction, participants used simple heuristics more frequently when the stakes were low: the frequency of the random choice-the simplest heuristic-increased significantly from $4.2 \%$ on high-stakes problems to $19.9 \%$ on low-stakes problems ( $\chi^{2}(1)=88.2, p<0.0001$ ), and so did the frequency of the second simplest heuristic, SAT-TTB $\left(\chi^{2}(1)=86.3, p<\right.$ 0.0001 ), and the third simplest heuristic, TTB $\left(\chi^{2}(1)=\right.$ $20.0, p<0.0001)$. The frequency of SAT also increased from high- to low-stakes games $\left(\chi^{2}(1)=3.4, p<0.05\right.$, one-

|  | Frequency |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Strategy | Total | HS-HD | HS-LD | LS-HD | LS-LD |
| TTB | 1012 | 392 | 64 | 449 | 107 |
| SAT-TTB | 412 | 68 | 54 | 140 | 150 |
| Random | 320 | 41 | 42 | 111 | 126 |
| TTB2 | 251 | 34 | 94 | 25 | 98 |
| WADD | 178 | 33 | 84 | 19 | 42 |
| SAT | 89 | 14 | 22 | 23 | 30 |

Table 1: Frequency of strategy types for each type of trial.
tailed). Finally, consistent with our model's third prediction, the frequency of the most effortful and most accurate strategy, WADD, increased with the stakes $\left(\chi^{2}(1)=19.3, p<0.0001\right)$.

Together, the strategies reported in Table 1 account for only about half ( $48.6 \%$ ) of all trials. To test our model's predictions on all of the trials, we quantified people's decision style by four metrics introduced by Payne et al. (1988): the number of inspected cells (acquisitions), the proportion of those inspections that pertained to the most probable outcome (prioritization), the degree to which subsequent acquisitions inspected the payoffs of different gambles for the same outcome versus the payoffs of the same gamble for different outcomes (outcome-based processing: $\frac{n_{\text {same outcome }}-n_{\text {same gamble }}}{n_{\text {same outcome }}+n_{\text {same gamble }}}$ ), and the average ratio of the expected value of the chosen gamble over the expected value of the optimal choice (relative performance). To further test our model's predictions, we ran a 2-way mixed-effects ANOVA for each of these four metrics.

As shown in Figure 2, the effects of the stakes and outcome probabilities on the four metrics confirmed the model's predictions. Our model's first prediction that high dispersion promotes the use of fast-and-frugal heuristics was confirmed by a decrease in the number of acquisitions $(F(1,3798)=$ $78.24, p<0.0001$ ) in conjunction with an increases in outcome-based processing $(F(1,3432)=68.31, p<0.0001)$ and prioritization $((F(1,3478)=280.1, p<0.0001))$. The increase in prioritization was especially striking: while only $40.4 \%$ of participants' clicks inspected the most probable outcome when dispersion was low, they focused $70.6 \%$ of their acquisitions on the most probable outcome when dispersion was high. Our model's second prediction that the higher stakes should decrease people's reliance on fast-andfrugal heuristics was confirmed by a significant increases in the number of acquisitions $(F(1,3798)=281.47, p<$ 0.0001 ) which was accompanied by a decrease in prioritization $(F(1,3478)=62.42, p<0.0001)$ and an increase in relative performance $(F(1,3798)=47.62, p<0.0001)$. Consistent with the model's third prediction, the average outcomebased processing metric was lower for high stakes but this effect was not statistically significant $(F(1,3432)=2.45, p=$ 0.06 , one-tailed). Our model's third prediction that highstakes increases time, effort, and performance, was confirmed by a significant increases in the number of acquisitions $(F(1,3798)=281.47, p<0.0001)$ and relative performance $(F(1,3798)=47.62, p<0.0001)$ with high stakes.


Figure 2: People's decision style by stakes and dispersion of the outcome probabilities.

Despite these qualitative agreements, there were quantitative differences. Most notably, our model predicted a more pronounced effect of the stakes on the number of acquisitions than we observed in people ( +19.6 vs. +3.4 ); the smaller effect in people might reflect their concave utility function.

## Discussion

In summary, our resource-rational theory of multi-alternative risky choice predicted some of the main strategies people use in the Mouselab paradigm and the conditions under which they are selected. In addition to automatically discovering known strategies and contingencies, our computational approach also discovered a novel, previously unknown heuristic that integrates TTB with satisficing (SAT-TTB), and our experiment confirmed that people do indeed use SAT-TTB on a non-negligible fraction of problems-especially when the stakes are low.

Tajima, Drugowitsch, and Pouget (2016) solved meta-level MDPs to derive boundedly optimal drift-diffusion models. The strategy discovery method presented here generalizes this approach to more complex decision mechanisms that can process and generate evidence in many different ways.

One limitation of the current work is that we do not know how closely our algorithm approximated the optimal policy, and it is possible that a more accurate approximation would yield somewhat different predictions. Future work will systematically evaluate the accuracy of our approximation method on smaller problems for which the optimal metalevel policy can be computed exactly. Another limitation of the present work is that the cost of computation had to be fit to the participants' responses. Future work will control the cost per click and measure it independently. This will enable a direct comparison of the time and effort people invest against the optimal amount of deliberation. However, a thorough answer to this question will require a more detailed model of people's cognitive architecture including a model of working memory. Another direction for future work is to characterize
the decision strategies the model employed on the vast majority of high-stakes problems where it did not use WADD.

Our proof-of-concept study suggests that formulating the problem of making optimal use of finite time and limited cognitive resources as a meta-level MDP is a promising approach to discovering cognitive strategies. This approach can be leveraged to develop more realistic normative standards of human rationality. This might enable future work to systematically evaluate the extent to which people are resourcerational. In the long term, our approach could be used to improve human reasoning and decision-making by discovering rational heuristics and teaching them to people.

Acknowledgments. This work was supported by grant number ONR MURI N00014-13-1-0341 and a grant from the Templeton World Charity Foundation.

## References

Agrawal, S., \& Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In Proceedings of the 30th international conference on machine learning (pp. 127-135).
Gigerenzer, G., \& Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. Psychological review, 103(4), 650.
Gigerenzer, G., \& Selten, R. (2002). Bounded rationality: The adaptive toolbox. MIT Press.
Griffiths, T. L., Lieder, F., \& Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Topics in Cognitive Science, 7, 217-229.
Hay, N., Russell, S., Tolpin, D., \& Shimony, S. (2012). Selecting computations: Theory and applications. In N. de Freitas \& K. Murphy (Eds.), Uncertainty in artificial intelligence: Proceedings of the twenty-eighth conference. P.O. Box 866 Corvallis, Oregon 97339 USA: AUAI Press.
Johnson, E. J., Payne, J. W., Bettman, J. R., \& Schkade, D. A. (1989). Monitoring information processing and decisions: The mouselab system (Tech. Rep.). DTIC Document.
Lieder, F., Griffiths, T. L., \& Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. Advances in Neural Information Processing Systems, 2699-2707.
Lieder, F., Hsu, M., \& Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. In Proceedings of the 36th annual conference of the cognitive science society.
Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., \& Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, \& K. Weinberger (Eds.), Advances in neural information processing systems 27 (pp. 2870-2878). Curran Associates, Inc.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1988). Adaptive strategy selection in decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3), 534.
Russell, S. J., \& Subramanian, D. (1995). Provably boundedoptimal agents. Journal of Artificial Intelligence Research, 2, 575-609.
Russell, S. J., \& Wefald, E. (1991). Principles of metareasoning. Artificial Intelligence, 49(1-3), 361-395.
Shugan, S. M. (1980). The cost of thinking. Journal of consumer Research, 7(2), 99-111.
Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological review, 63(2), 129.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA, USA: MIT press.
Tajima, S., Drugowitsch, J., \& Pouget, A. (2016). Optimal policy for value-based decision-making. Nature communications, 7.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.

# Sequential Effects in the Garner Tasks 

Deborah J. Lin (djlin@student.unimelb.edu.au)<br>Daniel R. Little (daniel.little@unimelb.edu.au)<br>Melbourne School of Psychological Sciences, University of Melbourne, Australia


#### Abstract

The distinction between integral and separable dimensions is of central importance to understanding how humans integrate information from multiple stimulus sources. One approach to characterizing stimulus integrality is through a set of speeded categorization tasks most closely associated with the work of Wendell Garner. These tasks demonstrate that integral dimensions result in marked speed up or slow down in responding when there is correlated or irrelevant variation, respectively, compared with a baseline task. Little, Wang \& Nosofsky (2016) recently found that the slow down or interference can be largely explained by a reduction in the number of direct repetitions in a modified Garner filtering task. In this paper, we examine a large sample of subjects tested on either separable or integral dimensions to determine the extent of and individual differences in the overall and sequential effects in the standard Garner tasks.


Keywords: Categorization; Response Times; Sequential Effects

## Introduction

In the study of perceptual decision-making, it is fundamental to understand the distinction between integrality and separability, as different processing architectures appear to underlie performance with integral and separable dimensions. Information from integral dimensions, which cannot easily be selectively attended to, is best explained as a pooling of information into a single, coactive processing channel (Little et al., 2013). On the other hand, separable dimensions, which can be easily selectively attended to, have been shown to be processed independently in serial or parallel (Fifić et al., 2010). Hence, the notion of integrality and separability must be taken into account in the formal model of categorization and decision making more broadly.

## Garner's (1974) Speeded-Categorization Tasks

One classic approach to understanding integrality is Garner's (1974) set of speeded-categorization tasks (see also Algom \& Fitousi, 2016, for a review). In these tasks, participants categorize stimuli into two categories as quickly and accurately as possible on each trial. Category membership in these tasks is determined by the stimulus' value on a single relevant dimension. The three major task conditions -control, correlated, and filtering - vary in the structure of the stimulus space, as shown in Figure 1.


Figure 1. Garner's (1974) control, correlated, and filtering conditions

In the control condition, there are two stimuli which only vary along the single relevant dimension (i.e., dimension X in Figure 1). In the correlated condition, there are two stimuli which vary along both the relevant dimension and a second irrelevant dimension. In the filtering condition, there are four stimuli with all possible combinations of relevant and irrelevant dimension values. In all conditions, participants should attend primarily to the relevant dimension while ignoring variation in the irrelevant dimension in order to perform the categorization task accurately and quickly.

For integral dimensions, a robust finding is that subjects have shortest response times (RTs) in the correlated task and the longest RTs in the filtering task. This suggests a correlated-facilitation and filtering- or Garner-interference effect, respectively (Garner, 1974). However, for separable dimensions, RTs across control, correlated, and filtering tasks are relatively invariant (Garner, 1974).

These patterns of RTs arise due to a difference in the ability to selectively attend and process information for integral and separable dimensions (Garner, 1974). When dimensions are separable, participants are easily able to selectively attend to relevant dimension, and as a result, the psychological representation of the stimulus space in all three conditions are collapsed to the single relevant dimension such that the correlated and filtering conditions are isomorphic to the control condition. However, when integral dimensions are used, participants are unable to selectively attend to the relevant dimension, and thus have different psychological representations of the stimulus space for each condition. For instance, as the stimuli vary along both dimensions in the correlated task, when the information from these dimensions are pooled and processed in a single channel, psychological discriminability between stimuli may be increased compared to when the stimuli only vary along one dimension in the control condition. With increased discriminability between stimuli, categorization becomes easier and more efficient resulting in
shorter RTs in the correlated task - the correlated-facilitation effect. There are several potential explanations for Garner interference. For one, there are more items in the filtering task than in the control task which may encourage more conservative responding, especially if the stimuli are highly confusable. Alternatively, the increase in the number of items might increase the perceived variability which would act to slow RTs (Nosofsky \& Palmeri, 1997).

In a recent paper using a modified version of the Garner task (see Figure 2), Little et al. (2016) showed that one explanation for filtering interference was the reduction of direct sequential repetitions in the filtering condition. That is, with more items, the probability of any one item repeating is reduced compared to the control condition. Repetitions have been show to produce very fast RTs; consequently, the reduction in repetitions results in slower responding (Fletcher \& Rabbitt, 1978; Krueger \& Shapiro, 1981). Investigations of decomposition (i.e., into sequential effects) of the standard Garner effects (Burns, 2016; Dyson \& Quinlan, 2010) have concluded that repetition effects can not be the sole explanation for Garner interference. However, two limitations of these papers are that only a small number of participants was tested ( $\mathrm{N}=16$; Dyson \& Quinlan (2010); $\mathrm{N}=30$; Burns (2016)) and there was no comparison to sequential effects in separable dimension stimuli in either case. Given that the sequential effects in our modified task were highly pronounced (Little et al., 2016), were also present for separable dimensions in the same modified task (Lin \& Little, 2017), and that we found considerable individual variability in our modified task, we sought in the present paper to conduct a larger replication of the standard Garner task to examine this decomposition using both integral and separable dimensions.

## Sequential Effects

Sequential effects arise due to a reliance on a relative comparison of the current stimulus to the preceding stimulus (or stimuli). These types of effects have been observed in a large variety of categorization tasks (Stewart et al., 2002, see e.g.,) but also in identification (Brown et al., 2007, see e.g.,) and simple choice tasks (Luce, 1986; Jones et al., 2013). One such effect that has been widely studied is the repetition effect, where subjects have higher accuracy and shorter RTs when the current stimulus is identical to the immediately preceding stimulus (Felfoldy, 1974; Lockhead et al., 1978). In their modified task, using integral dimensions, Little et al. (2016) showed that there are complex sequential effects that arise across the control, correlated, and filtering conditions.

1. Repetition Effect: Items which were adjacent to category boundary were categorized faster and more accurately when preceded by the same item than when preceded by another item.
2. Far same category pushing effect: When the near boundary item was preceded by a far item from the same category, RTs were slower and errors higher than when the near item was preceded by another item.


Figure 2. Schematic diagram of the modified Garner-task paradigm using stimuli varying on integral dimensions brightness and saturation - where the relevant dimension is brightness.
3. Adjacent opposite category pulling effect: When the near boundary item was preceded by an adjacent item from the opposite category, RTs were slower and errors higher than when the near item was preceded by another item.
4. Irrelevant dimension change: Finally, in the filtering task, the repetition effect was attenuated and the pushing and pulling effects were enhanced when the irrelevant dimension changed (i.e., when there was only repetition of the relevant but not the irrelevant dimension value). This effect emphasizes the role of previous item distance (i.e., from the current item) in determining the magnitude of the sequential effects. This was also evident in the attenuated pushing and pulling effects in the correlated condition (i.e., since the between category items are further apart in that category).

We have recently demonstrated with separable dimensions that the same repetition, pushing, and pulling effects arise even when there was no overall average RT difference between conditions (Lin \& Little, 2017). There is no effect of changing the irrelevant dimension in the filtering task with separable dimensions consistent with the notion that attention acts to collapse the separable conditions across the irrelevant dimension.

While there have been some investigations of sequential effects in the standard Garner task (Felfoldy, 1974; Lockhead et al., 1978), there have been few comparisons of sequential effects between integral and separable dimensions. Additionally, there is value in collecting a large replication sample in the standard Garner task, as the magnitude and variability of the standard Garner effects and sequential effect are currently unclear. For instance, not all subjects showed the standard Garner ordering (i.e., correlated RT ; control RT ; filtering RT) in a modified Garner task. Thus, the present study seeks to quantify the size and variability of the standard Garner effects and several decompositions of those effects (including sequential effects; from Dyson \& Quinlan, 2010) using a hierarchical Bayesian analysis.

## Method

Two sets of experiments following the general procedure outlined in Garner (1974) were conducted. Experiment 1 used integral dimensions; Experiment 2 used separable dimensions.

Participants In Exp 1, 100 University of Melbourne undergraduates were randomly assigned to either the brightness ( $N$ $=50)$ group or saturation $(N=50) .{ }^{1}$ One saturation participant was excluded due to an overwritten data file. In Exp 2, 99 students were randomly assigned to either the saturation ( $N$ $=49)$ or line-position $(N=50)$ group. All received course credit for participation.
Exp 1: Integral Stimuli Stimuli were color squares (100 $\times 100$ pixels each; Munsell hue 5R) that varied in brightness (value) and saturation (chroma). The set of four stimuli was created by combining two levels of brightness (values 5, 6) and two levels of saturation (chroma 6,8). The stimuli were presented on a monitor resolution of $1280 \times 1024$.
Exp 2: Separable Stimuli Stimuli were colored rectangles ( $170 \times 255$ pixels) with a black outline and with a small inset black vertical line positioned along the base of the rectangle. The color was selected from the Munsell hue 5R with a brightness value of 5 while the saturation was varied. The line varied by position along the base of the rectangle from the left side of the rectangle. The full set of stimuli was created by combining two levels of saturation (chroma 8,10 ) and two line positions ( 60,80 pixels from the left). The stimuli were presented on a monitor resolution of $1280 \times 1024$.

## General Procedure

In both experiments, participants each completed a one-hour categorization task. At the outset, participants were presented with an instruction screen with examples of the stimuli and were told to categorize each stimulus as accurately and quickly as possible. Participants then completed 5 blocks of 24 practice trials followed by 120 experimental trials, and a $6^{\text {th }}$ block of 120 experimental trials.

The control task and correlated tasks were presented over two blocks. In both tasks, only two stimuli of the full set were presented to the participant on each trial. For the subsequent block of the control task, the irrelevant dimension value was switched. For the subsequent block of the correlated task, the relevant and irrelevant dimension values of the two stimuli were both switched.

The filtering task was presented over two consecutive blocks without practice trials for the second block. The blocks of tasks were counterbalanced and the order of presentation of individual stimuli on each trial was randomized anew within each block.

On each trial, a fixation cross was presented for 1500 ms , followed by the stimulus. The participant then decided

[^130]whether the stimulus belonged to category A or B. Response choice and response time (RT) were recorded via button press of a customized RT box Li et al. (2010). The stimulus remained on screen until a button press was made or until the 5000 ms response deadline. Full feedback (i.e., "right", "wrong") was provided for the 24 practice trials; only incorrect response feedback was provided for experimental trials. If a response was not made before the response deadline, feedback "too slow" was given. The feedback remained on screen for 2000 ms .

## Data Analysis

We applied two hierarchical Bayesian models. For the first model, we found the posteriors for a single group distribution for each of the items in the control, correlated, and filtering task in each of the integral and separable experiments. For the second model, we found the posteriors for distributions of each sequential order for each condition across both experiments. That is, we estimated the posterior for when the relevant dimension value repeated and the irrelevant dimension value repeated (hereafter, RR), for when the relevant dimension changed but the irrelevant dimension repeated (CR), when the relevant dimension repeated but the irrelevant dimension change (RC), and for when both the relevant and irrelevant dimensions changed (CC). The control task only contains the RR and CR conditions, the correlated task contains the RR and CC conditions, and the filtering task contains all four conditions.

For each experiment $i$, each subject $j$, and each task (or sequence condition) $k$, we estimated the rt as a lognormal distribution, $r t_{i j k} \sim \log N\left(\mu_{i k}, \sigma_{i k}\right)$. The prior over the subject means was a normal distribution, $\mu_{i k} \sim N\left(M_{k}, S_{k}\right)$, and the prior over the subject precision $\left(1 / \sigma_{i k}\right)$ was a gamma distribution, $1 / \sigma_{i k} \sim \operatorname{Gamma}\left(a_{k}, b_{k}\right)$. Hyperpriors were relatively non-informative, $M_{k} U(0,7), S_{k} \sim U(0,500), a_{k} \sim$ $U(.5,100)$, and $b_{k} \sim U(.5,100)$, where $U(x, y)$ is a uniform distribution over the range $[x, y]$. The models were implemented in JAGS (Plummer, 2003) for which we collected 1000 samples after 1000 burn-in samples from two MCMC chains. Plots of these chains indicated good convergence.

## Results

The estimated rt means and variances are on a logarithmic scale and not the scale of the original RT data. Hence, to summarize the effects, we converted the posterior group logNormal distribution means, M, and standard deviation, to the RT scale using the following transformation:

$$
\begin{gathered}
\tilde{M}_{k}=\exp \left(M_{k}+\frac{S_{k}^{2}}{2}\right) \\
\tilde{S}_{k}=\sqrt{\exp \left(2 M_{k}+S_{k}^{2}\right) \exp \left(S_{k}^{2}-1\right)}
\end{gathered}
$$

## Overall Condition Analysis

We first analysed the overall difference between condition by taking the difference between the Control and Correlated posterior estimates (left panel, Figure 3) and between the Control


Figure 3. Posterior distributions for the difference between control and correlated overall mean RTs (left panel), and the difference between between control and filtering overall mean RTs (right panel). The solid line shows the distribution for the integral posterior and the dotted line shows the posterior for the separable condition.
task and the filtering task (right panel, Figure 3). We note that there were no strong qualitative individual differences; only quantitative variation.

For the comparisons to the correlated condition, positive values would indicate shorter RTs in the correlated condition than the control condition. Analogously, for the comparison to the filtering condition, negative values indicate longer RTs in the filtering condition than in the control condition. As shown in Figure 3, the posterior distributions for the separable conditions have substantial density over 0 indicating no overall effect of condition. For the integral conditions, the distributions had the most density over positive and negative difference values for the correlated and filtering comparisons, respectively. Hence, we've replicated the standard Garner result and have shown that all subjects in our experiment show this pattern of results.

## Sequential Item Analysis

Figure 4 shows the posterior distributions for each of the item conditions. For the separable dimensions condition, posterior distributions for item conditions appear to be relatively invariant across the control, correlated, and filtering tasks, indicating little or no sequential effects. The posterior distributions for the integral dimension condition reveals a more complex pattern of item condition effects. In the control task, the posterior distributions for RR and CC indicate no sequential effects. In the correlated task, the posterior distribution for RR lies slightly lower than CC, suggesting a repetition effect. In the filtering condition, posterior RTs are markedly slower for irrelevant dimension changes (i.e., RC and CC), and quickest when the stimulus is repeated (i.e., RR).
We summarized these distribution by computing several effect decompositions derived by Dyson \& Quinlan (2010).

Filtering interference Note that overall filtering interference can be decomposed into sequential components as: $\left[R R_{\text {filt }}+R C_{\text {filt }}+C R_{\text {filt }}+C C_{\text {filt }}\right] / 4-\left[R R_{\text {cont }}+C R_{\text {cont }}\right] / 2$,


Figure 4. Posterior distributions for the transformed $\operatorname{logNor}$ mal groups means for the Control condition (RR \& CR; Left panel), Correlated condition (RR \& CC; Middle panel), and Filtering condition (CC, CR, RC, CC; Right panel)


Figure 5. Posteriors distributions for irrelevant feature variation (left panel) and stimulus uncertainty (right panel) components of filtering interference for both integral and separable dimensions.
which "filt" refers to the filtering condition and "cont" to the control condition. This overall measure can be further decomposed into the following two components:

1. A measure of irrelevant feature variation, which is positive if there is a cost when the irrelevant dimension changes:
$\left[R R_{\text {filt }}+R C_{\text {filt }}+C R_{\text {filt }}+C C_{\text {filt }}\right] / 4-\left[R R_{\text {filt }}+C R_{\text {filt }}\right] / 2$
2. A measure of stimulus uncertainty, which is positive if there is a cost associated with having more stimuli in the filtering condition controlling for changes in the irrelevant dimension: $\left[R R_{\text {filt }}+C R_{\text {filt }}\right] / 2-\left[R R_{\text {cont }}+C R_{\text {cont }}\right] / 2$

These two effects are shown in Figure 5. For these figures, negative values indicate RT benefits (i.e., shorter RT) while positive values indicate RT costs (i.e., longer RT) for the respective effect. The posterior distribution for both effects for separable dimensions have substantial density over zero, indicating no irrelevant feature variation or stimulus uncertainty effects. In contrast, the posterior distributions for integral dimensions have substantial density over positive values, indicating RT costs as a result of irrelevant variation and stimulus uncertainty. Furthermore, stimulus uncertainty appears to contribute to filtering interference more than irrelevant variation for the integral dimensions condition.

Correlated benefit In the correlated condition we conducted a corresponding decomposition (again following


Figure 6. Posterior distributions for redundancy repetition (left panel) and redundancy change (right panel) components of correlated facilitation for both integral and separable dimensions.

Dyson \& Quinlan 2010) as: $\left[R R_{\text {corr }}+C R_{\text {corr }}\right] / 2-\left[R R_{\text {cont }}+\right.$ $\left.C R_{\text {cont }}\right] / 2$.

This overall measure can be further decomposed into:

1. The effect of redundancy repetition, which indexes the effect of changing both dimensions: $\left[R R_{\text {corr }}-R R_{\text {cont }}\right] / 2$
2. The effect of redundancy change, which indexes the effect of additional irrelevant dimension variation in the correlated condition compared to the control condition: $\left[C_{\text {corr }}-\right.$ CR $\left._{\text {cont }}\right] / 2$.

These effects are shown in Figure 6. For separable dimensions, the posterior distributions for both redundancy repetition and redundancy change have substantial density over zero, indicating no overall correlated facilitation effect. For integral dimensions, both redundancy repetition and redundancy change have substantial value over negative values, indicating RT benefits. In addition, the components appear to contribute approximately equally to the overall correlated facilitation effect.

Repetition Effect Finally, for all three conditions we computed the effect repeating an item compared to switching an item (i.e., in the control and correlated conditions; in the filtering condition, we compared repetition to the average of the other three item RTs). ${ }^{2}$ This repetition measure is computed as:

$$
\begin{aligned}
\text { Control Repetition } & =R R_{\text {cont }}-C R_{\text {cont }} \\
\text { Correlated Repetition } & =R R_{\text {corr }}-C C_{\text {corr }} \\
\text { Filtering Repetition } & =R R_{\text {filt }}-\left[C R_{\text {filt }}+R C_{\text {filt }}+C C_{\text {filt }}\right] / 3
\end{aligned}
$$

This measure can also be interpreted to indicate a repetition effect (i.e., shorter RTs as a result of repetition in both dimensions). A negative value indicates a repetition effect

In the control condition, the distributions for both integral and separable dimensions are centered around zero, suggest-

[^131]

Figure 7. Posterior distributions for the repetition effect in control, correlated, and filtering tasks for both integral and separable dimensions.
ing no repetition effect. In the correlated condition, the posterior distributions for both integral and separable dimensions have a substantial density over negative values, indicating a slight repetition effect. The repetition effect for integral dimensions also appears to be marginally stronger than for separable dimensions; though even here, both distributions have $95 \%$ highest posterior density intervals which overlap 0 . In the filtering condition, the distribution for separable dimensions has substantial density over zero, indicating no repetition effect. However, the distribution for integral dimensions lies mainly over negative values, indicating the presence of a strong repetition effect.

## Discussion

Overall, the hierarchical Bayesian approach in the present study revealed reliably strong standard Garner effects, showing correlated facilitation and filtering interference with integral dimensions but not with separable dimensions. A further decomposition of the Garner effects into sequential item conditions, following Dyson \& Quinlan (2010), provide further insight into the underlying mechanisms of perceptual decision-making. Notably, we found little evidence for any individual differences.

One notable finding is that no sequential effects were found with separable dimensions in the standard Garner task. This result is in contrast to the sequential effects found with separable dimensions in the modified Garner task (Lin \& Little, 2017). One potential explanation could be that the presence and magnitude of sequential effects depends on task complexity. For example, Bentin \& McCarthy (1994) found that immediate repetition provides a relatively larger advantage in lexical decision and face recognition tasks compared to simpler discrimination tasks, as it eliminated the need for more complex processes such as accessing semantic memory. Similarly, as the standard Garner task has a much smaller stimulus space compared to the modified Garner task, repetitions may provide a large benefit for the modified task but a much smaller or no RT benefit, and as a result, no sequential effects arise in the standard task. On the other hand, we have only examined the effects of a single preceding item; in simple RT tasks (i.e., with two stimuli), there are complex sequential ef-
fects extending up to five items back reflecting the influence of repetitions and alternations (Jones et al., 2013).

Another important result is that stimulus uncertainty contributes to filtering-interference more than irrelevant feature variation. An explanation could be that the lack of interference from irrelevant feature variation can be attributed to the integrality of dimensions. If dimensions are less integral and easier to selectively attend to, then the irrelevant variation would not contribute to interference, for example, in the separable dimensions case. It should also be noted that it is difficult to isolate stimulus uncertainty and irrelevant feature variation in the standard Garner task, as an increase in the number of irrelevant dimensions is associated with a larger number of stimuli. Even though these measures attempt to isolate trials where only stimulus uncertainty or irrelevant variation changes, it is unclear whether the larger context of the task has no impact. Burns (2016) attempted to disentangle these two components by introducing a 3-dimensional Garner task where irrelevant variation could be increased without affecting stimulus uncertainty, and demonstrated that irrelevant variation alone can increase interference substantially. In order to further evaluate the components underlying filteringinterference, promising avenues for future work might be to measure these decomposition effects with a variety of different dimensions varying on integrality or to carefully manipulate stimulus uncertainty and irrelevant variation within Burns's (2016) 3-dimensional Garner task.

## Acknowledgments

DL was supported by an ARC Discovery Project grant: DP160102360.

## References

Algom, D., \& Fitousi, D. (2016). Half a century of research on garner interference and the separability-integrality distinction [Journal Article]. Psychological Bulletin.

Bentin, S., \& McCarthy, G. (1994). The effects of immediate stimulus repetition on reaction time and event-related potentials in tasks of different complexity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(1), 130.

Brown, S., Marley, A., \& Lacouture, Y. (2007). Is absolute identification always relative? comment on stewart, brown, and chater (2005).
Burns, D. M. (2016). Garner interference is not solely driven by stimulus uncertainty. Psychonomic bulletin \& review, 23(6), 1846-1853.

Dyson, B. J., \& Quinlan, P. T. (2010). Decomposing the garner interference paradigm: Evidence for dissociations between macrolevel and microlevel performance. Attention, Perception, \& Psychophysics, 72(6), 1676-1691.
Felfoldy, G. L. (1974). Repetition effects in choice reaction time to multidimensional stimuli. Attention, Perception, \& Psychophysics, 15(3), 453-459.

Fifić, M., Little, D. R., \& Nosofsky, R. (2010). Logicalrule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. [Journal Article]. Psychological Review, 117, 309-348.
Fletcher, B., \& Rabbitt, P. M. A. (1978). The changing pattern of perceptual analytic strategies and response selection with practice in a two-choice reaction time task [Journal Article]. The Quarterly Journal of Experimental Psychology, 30, 417-427.
Garner, W. R. (1974). The processing of information and structure [Book]. Psychology Press.
Jones, M., Curran, T., Mozer, M. C., \& Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. Psychological review, 120(3), 628.
Krueger, L. E., \& Shapiro, R. G. (1981). Intertrial effects of same-different judgements [Journal Article]. The Quarterly Journal of Experimental Psychology, 33, 241265.
Li, X., Liang, Z., Kleiner, M., \& Lu, Z.-L. (2010). Rtbox: A device for highly accurate response time measurements [Journal Article]. Behavior Research Methods, 42(1), 212225.

Lin, D. J., \& Little, D. R. (2017). Further tests of sequential effects in a modified garner task using separable dimensions [Journal Article]. Manuscript in preparation.
Little, D. R., Nosofsky, R. M., Donkin, C., \& Denton, S. E. (2013). Logical-rules and the classification of integral dimensioned stimuli [Journal Article]. Journal of Experimental Psychology: Learning, Memory and Cognition, 39, 801-820.
Little, D. R., Wang, T., \& Nosofsky, R. M. (2016). Sequence-sensitive exemplar and decision-bound accounts of speeded-classification performance in a modified garner-tasks paradigm [Journal Article]. Cognitive Psychology, 89, 1-38.
Lockhead, G., Gruenewald, P., \& King, M. (1978). Holistic vs. attribute repetition effects in classifying stimuli. Memory \& Cognition, 6(4), 438-445.
Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization [Book]. New York: Oxford University Press.
Nosofsky, R. M., \& Palmeri, T. (1997). Comparing exemplarretrieval and decision-bound models of speeded perceptual classification [Journal Article]. Perception \& Psychophysics, 59, 1027-1048.
Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In Proceedings of the 3 rd international workshop on distributed statistical computing (p. 125-132).
Stewart, N., Brown, G. D., \& Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli [Journal Article]. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 3.

# Error-Based Learning: A Mechanism for Linking Verbs to Syntax 

Yi Lin (yilin5@illinois.edu)<br>Department of Psychology, 603 E. Daniel Street<br>Champaign, IL 61820 USA

Cynthia Fisher (clfishe@illinois.edu)<br>Department of Psychology, 603 E. Daniel Street<br>Champaign, IL 61820 USA


#### Abstract

Children and adults are guided by verb-specific syntactic likelihoods, or verb bias, in language comprehension and production. Recent reports showed that verb bias can be altered by new linguistic experience. We investigated the mechanisms underlying this verb bias learning or adaptation. Specifically, we asked whether verb bias learning, like abstract syntactic priming, is driven by error-based implicit learning. We report three experiments in which we altered the biases of familiar dative verbs in children's and adults' sentence production, via training trials that induced participants to produce each verb consistently in either double-object or prepositional-object dative structures. Participants' syntactic choices in later test trials reflected the expected adaptation of verb bias to the training experience. In addition, the magnitude of the training effect varied with the likelihood of each sentence structure and with pre-existing verb bias: Unexpected verb-structure combinations resulted in larger training effects, suggesting the operation of error-based implicit learning.


Keywords: language acquisition; verb bias; implicit learning; error-based learning; surprisal

## Introduction

Verbs are choosy about the sentence structures they occur in. Transitive but not intransitive verbs can appear in transitive sentences, with two noun-phrase arguments (She saw Sue; *She slept Sue); only certain verbs permit dative structures, with three arguments (She showed the book to Sue; *She saw the book to Sue). In addition to these all-or-none licensing restrictions, the linking of verbs with syntax is constrained by syntactic likelihoods. Most verbs license multiple structures, but may occur much more often in one structure than another. To illustrate, many dative verbs license both the prepositional-object (PO: He showed/passed the book to her) and the double-object dative structure (DO: He showed/passed her the book), but the verb show occurs much more often in the DO structure than does pass. These verbspecific likelihoods are known as verb bias. Verb bias knowledge guides language processing in children and in adults, affecting how we link verbs and syntax in production, and online expectations about likely sentence structures in comprehension (e.g., Peter et al., 2015; Snedeker \& Trueswell, 2004).

Verb bias effects emerge early in acquisition (Peter et al., 2015; Tomasello, 1992), but continually adapt to ongoing linguistic experience in children and adults. Recent reports show that the biases of even well-known verbs can be altered by
new linguistic experience (Coyle \& Kaschak, 2008; Lin \& Fisher, 2016; Qi, Yuan \& Fisher, 2011; Ryskin, Qi, Duff \& Brown-Schmidt, 2016). For instance, Lin and Fisher (2016) asked children and adults to describe videos by repeating and completing sentence stems provided by an experimenter. Training stems (10 per verb) induced participants to produce one verb only in DO structures (Dora gave Boots__), and another verb only in PO structures (Minnie showed the clock___). Test stems ended at the verb, allowing participants to choose either dative structure (Piglet gave__ The teacher showed___). This brief training changed the biases of a wide range of familiar verbs in adults' and 4-year-olds' sentence production. In unconstrained test trials, participants produced more DO descriptions with verbs trained in DO than in PO structures. Similar verb-bias training effects have been found in children's and adults' comprehension of sentences with a prepositional phrase attachment ambiguity (e.g., Tickle/choose the frog with the feather; Qi et al., 2011; Ryskin et al., 2016). These findings tell us that learners keep track of the statistics of verb-structure combinations in the linguistic environment, and adapt their language-processing systems accordingly.

In the present study, we explored error-based learning as a potential mechanism for verb bias learning. To do so, we explored parallels between verb bias learning and abstract syntactic priming. Syntactic priming is the tendency to reuse a previously encountered syntactic structure. For example, a talker who has recently read a sentence in the DO structure (The governess made the princess a pot of tea) is more likely to choose the same structure to describe an unrelated picture (The boy is handing the singer a guitar; Bock, 1986). Syntactic priming is abstract-it spans different verbs, as in the example just given. Syntactic priming can be measured in children and adults, and in comprehension and production (Rowland et al., 2012; Thothathiri \& Snedeker, 2008). The priming effects are long-lasting (Bock \& Griffin, 2000), suggesting that they reflect long-term learning about abstract syntax. Taken together, the literature on syntactic priming, and recent reports of verb-bias learning, suggest that learners adapt to the statistics both of abstract syntactic structures, and of verbstructure combinations (e.g., Wonnacott, Newport, \& Tanenhaus, 2008).

Of particular interest here, syntactic priming shows 'inverse preference' or 'surprisal' effects (e.g., Bernolet \& Hartsuiker, 2010; Jaeger \& Snider, 2013; Peter et al., 2015).

That is, priming effects are larger if the structure of the prime sentence is unexpected. To illustrate, Bernolet and Hartsuiker (2010), in a study of adult sentence production, reported that (a) DO prime sentences exerted a larger priming effect (relative to baseline) than did prime sentences in the more frequent PO structure, and (b) the magnitude of syntactic priming depended on verb bias: The effect of a DO prime was larger if a PO-biased verb (one that rarely appears in this structure) appeared in the prime sentence. Children show similar effects of verb bias on the magnitude of syntactic priming (Peter et al., 2015). This pattern points to error-based implicit learning as a mechanism for syntactic priming: We expect likely structures, and thus learn more from the unexpected, continuously adapting the language-processing system to a changing linguistic environment (Chang, Dell, \& Bock, 2006; Jaeger \& Snider, 2013).

Could verb bias learning result from the same error-based learning mechanisms that support syntactic priming? If so, then verb bias training effects should vary with training-sentence surprisal. We tested this prediction by adapting the materials of Lin and Fisher (2016) to vary both training structure (DO- vs. PO-training) and pre-existing verb bias (DO-biased vs. PO-biased verbs).

In three experiments, participants watched videos depicting simple transfer events, and were prompted to describe each one by repeating and completing a sentence stem provided by an experimenter (Fig-1). As before, each participant received training trials that induced them to produce DO structures with one verb (DO-training), and PO structures with a second verb (PO-training). Crucially, one of the restricted verbs was chosen to be already DO-biased (e.g., show), while the other was PO-biased (e.g., pass). The assignment of verbs to training conditions varied between subjects, resulting in two list conditions: In the with-bias list, both verbs were trained in the structure that matched their pre-existing biases (e.g., PO-training for PO-biased pass, DO-training for DO-biased show). In the contra-bias list, both verbs were trained in the structure that mismatched their preexisting biases (PO-training for DO-biased show, DO-training for PO-biased pass). Following this training, participants received test trials in which the sentence stems to be completed ended at the verb. Participants' structural choices in these unconstrained test trials provided our measure of verb


Figure 1. Example of contra-bias training in Experiment 1a.
bias learning.
We work through our predictions for test-trial performance in Fig-2. Each panel shows the expected rate of DO-structure responses (as a proportion of DO and PO responses) under different experimental outcomes, plotted by within-subjects training condition (PO-training vs. DO-training) and be-tween-subjects list condition (with-bias vs. contra-bias).

Based on previous results we expected pre-existing verb bias to affect the rate of DO responses at test. Fig-2a shows the data pattern that would result from baseline verb bias alone: DO responses should be much more common for the DO-biased than for the PO-biased verb. Assuming no training effect, the difference between the two verbs (indicated by the equal-sized arrows in Fig-2a) would not vary with training condition.

We also expected to find a training effect. Fig-2b shows the data pattern that would result if a uniform verb-bias training effect, one that does not vary with training-sentence surprisal, were added to the effect of pre-existing verb bias. As Fig-2b shows, PO-training would decrease the rate of DO responding (relative to baseline), and DO-training would increase the rate of DO responses. Given a uniform training effect, the difference between the two verbs, reflecting pre-existing verb bias, would again remain unchanged.

Fig-2c shows the predicted data pattern if verb bias training effects vary with training-sentence surprisal. Trainingsentence surprisal should reflect both the likelihood of the training structure itself, and its fit with the pre-existing bias of the verb. We expected DO-training to exert a larger effect


Figure 2. Predicted rate of DO responses, by training condition (PO- vs. DO-training) and list condition (with-bias vs. contra-bias). Fig-2a shows an effect of baseline verb bias only; Fig-2b shows the influence of both pre-existing verb bias and a uniform training effect; Fig-2c shows the key surprisal predictions. In 2c, the rate of DO responses reflects preexisting verb bias, and an effect of training that is influenced by training-sentence surprisal.
than PO-training, because the DO structure is a non-canonical structure. For example, the DO structure imposes discourse constraints on its use: It is typically used to place dis-course-given recipients in post-verbal position (Show her the picture; Stephens, 2015). The PO structure, in contrast, has no strong discourse constraints (Brown, Savova, \& Gibson, 2008). In our task, without a discourse set-up establishing the recipient as given, the DO structure should be an unexpected choice. We also expected the training effect to vary with preexisting verb bias: the effect of DO-training should be strongest for PO-biased verbs, those that rarely occur in the DO.

Accordingly, as shown in Fig-2c, DO-training should considerably increase the rate of DO responses for the PO-biased verb (e.g., pass), but should have relatively little effect on the rate of DO responses for an already DO-biased verb (e.g., show). In the PO-training condition, we should see relatively little change due to training for either verb, preserving the large difference between verbs that reflects their baseline biases. Notice the key difference between Fig-2b and Fig-2c: A training effect that varies with training-sentence surprisal should reduce the difference between the two verbs in the DO-training condition relative to the PO-training condition.

We tested this prediction with 4- and 5-year-olds in Experiments 1a and 1b, and with adults in Experiment 2.

## Experiment 1a

## Methods

Participants Forty-eight four- and five-year-old children (Mean $=4 ; 8$; Range $=4 ; 0-5 ; 11$ ) participated; all were native speakers of English. Data from 4 additional children were excluded due to low training compliance (see below).

Materials and Procedures The materials were 46 5-s animated video clips depicting transfer events designed to be described by dative verbs, and 49 filler videos that did not depict transfer events. Children watched and described all 95 (critical and filler) videos by repeating and completing a sentence stem (Fig-1). The task was adapted from Lin and Fisher (2016), described in the Introduction. The task took about 30 to 40 minutes, and was made engaging for children by embedding it in a scavenger hunt for which game-tokens were discovered at intervals.

The task included a training and a testing block, with no boundary between them from the child's perspective. The key manipulation involved artificially restricting particular verbs to particular dative structures (only DO or PO) in training. As shown in Fig-1, training stems ended with a post-verbal noun, biasing children to produce either a DO or a PO sentence. Test stems ended at the verb. Show and pass were the two verbs that were restricted in training. These verbs differ in their pre-existing biases, as revealed in a separate norming study. Show is used more often in the DO, and pass in the PO dative structure. Children were randomly assigned to the with-bias or the contra-bias condition. Recall that in the withbias condition, both verbs were trained in the structure that matched their pre-existing bias, whereas in the contra-bias
condition, both verbs were trained in the structure that mismatched their pre-existing bias.

A third dative verb, give, was unrestricted, appearing equally often in the DO and PO structures during training. Unrestricted give (a DO-biased verb) was included to increase children's baseline rate of DO responses in the task; note that children tend to prefer the PO structure in tasks like ours (Peter et al., 2015; Stephens, 2015).

Children received 10 training trials per verb ( 30 training trials total) in the training block. In the test block children received 4 unconstrained test trials per restricted verb (show, pass), and 8 test trials for the unrestricted verb give. The three verbs were interleaved in training and test, and each child heard equal numbers of DO and PO training stems across verbs, ensuring that any effect of training reflected verb-bias learning rather than abstract syntactic priming.

The main task was preceded by a naming game in which children named the familiar characters and objects involved in the events. The video-description task then began with two filler trials to demonstrate the task.

Children's responses were transcribed and coded as DO, PO, or Other, following Rowland et al.'s (2012) criteria. Children who produced fewer than $80 \%$ training-compliant responses in each training condition (e.g., $80 \%$ DO responses for their DO-trained verb) were replaced. The 48 included children showed a compliance rate of $94 \%$ in training.

Our main analyses concerned the responses in the 8 test trials with experimentally restricted verbs. Of the 384 test responses, 8 were Other responses, leaving 376 . The dependent measure was the proportion of DO responses (out of DO and PO responses only), compared across the training and list conditions.

## Results and Discussion

Fig-3a shows children's proportion of DO responses in the test trials with restricted verbs, by within-subjects training condition (PO- vs. DO-Training) and between-subjects list condition (with- vs. contra-bias). As predicted, children showed a training effect, producing more DO responses in the DO-training condition ( $38 \%$ ) than in the PO-training condition (27\%). They also showed a clear effect of pre-existing verb bias, producing more DO responses with the DO-biased verb show (43\%) than with the PO-biased verb pass (22\%), averaged across training conditions.

Crucially, the effect of training varied with the likelihood of the training structure and its fit with pre-existing verb bias. Fig-3a shows that DO-training dramatically increased the rate of DO responses for the PO-biased verb pass, but had little effect on the rate of DO responses for the already DObiased verb show. This asymmetrical training effect reduced the difference between the two verbs in the DO-training condition relative to the PO-training condition. In the PO-training condition, Fig-3a shows a large difference in the rate of DO responses for the PO-biased verb pass, versus the DObiased verb show; this difference straightforwardly reflects


Figure 3. Mean (se) proportion DO responses (out of DO and PO responses) in the test phrase of Experiment la (Figure 3a), Experiment $1 \mathrm{~b}(3 \mathrm{~b})$ and Experiment $2(3 \mathrm{c})$, by training condition (PO vs. DO-training) and list (with-bias vs. contra-bias).
the verbs' baseline verb biases. The effect of training was larger for DO- than for PO-training, reflecting the likelihood of each structure; the effect of DO-training was larger for a PO-biased verb, reflecting the likelihood of verb-structure combinations. This pattern of responses closely resembles the predictions shown in Fig-1c, suggesting both an effect of training and an influence of training-sentence surprisal on the magnitude of the training effect.

This pattern was supported by a two-way mixed-model ANOVA on the proportion of DO responses (arcsine transformed) that revealed a main effect of training $(F(1,46)=$ $5.24, p<.05)$, and an interaction of training and list $(F(1,46)$ $=12.22, p<.01)$. Separate t -tests revealed that the difference between the two verbs was significant in the PO-training condition $(t(47)=3.32, p<.01)$ but not in the DO-training condition $(t(47)<1)$, consistent with our surprisal predictions.

In Experiment 1a we reproduced the verb-bias training effect in young children's language production documented in prior work (Lin \& Fisher, 2016). Experience producing a verb repeatedly in one syntactic structure modified the structural biases of that verb, rendering children more likely to use the verb in the same structure in later sentences. We also found the first evidence that the magnitude of this training effect depends on the likelihood of the training sentences. In Experiment 1 b we sought to extend this effect to different verb sets, exploring the robustness of the surprisal effect.

## Experiment 1b

## Methods

Participants A new group of forty-eight four- and five-yearold children (Mean $=4 ; 7$; Range $=4 ; 0-5 ; 8$ ) participated, all native English speakers. Data from 4 additional children were excluded due to low training compliance (3), or too few dative responses in the test trials (1).

Materials and Procedures Materials and procedures were identical to those of Experiment 1a, except that send was the PO-biased verb for half of the children and throw for the other half. We chose send and throw, two other PO-biased verbs, to seek evidence of surprisal effects with different verb sets. We retained show as the DO-biased verb to avoid reducing
the verb-bias difference between our restricted verbs in Experiment 1 b : Given children's overall preference for the PO structure, our norming study with children identified few strongly DO-biased verbs. As in Experiment 1a, each participant was randomly assigned to the with- or contra-bias list.

Children who produced fewer than $80 \%$ training-compliant responses in each training condition were replaced. We also excluded one child who did not produce at least one dative response in the test block for each restricted verb. The included children produced training-compliant responses in $95 \%$ of training trials. Of the 384 possible responses in the test trials with restricted verbs, 7 were coded as Other trials, leaving 377 DO and PO responses.

## Results and Discussion

Fig-3b shows children's proportion of DO responses in the restricted-verb test trials, by within-subjects training condition (PO- vs. DO-Training) and between-subjects list condition (with- vs. contra-bias). The pattern of responses closely resembles that found in Experiment 1a. Children showed a training effect, producing more DO responses in the DOtraining condition (35\%) than in the PO-training condition (20\%), and an effect of pre-existing verb bias, producing more DO responses for the DO-biased verb show (34\%) than for the PO-biased verbs ( $21 \%$ ).

As before, the effect of training varied with the likelihood of the training structure and its fit with the pre-existing bias. In Fig-3b, DO-training greatly boosted the rate of DO responses for the PO-biased verbs, but had little effect on rate of DO responses for the DO-biased verb show. As a result, the difference between the two verbs in the DO-training condition was much smaller than in the PO-training condition. In the PO-training condition, Fig-3b shows a large difference in the rate of DO responses for the PO-biased verbs versus the DO-biased verb, reflecting these verbs' pre-existing verb biases. Therefore, as before, the data bear out our surprisal predictions: the effect of training was larger for DO- than for PO-training, reflecting the likelihood of each structure, and the effect of DO-training was larger for a PO-biased than for a DO-biased verb. The same pattern emerged for both verb sets (not shown in the figure).

These observations were borne out by an ANOVA on the proportion of DO responses (arcsine-transformed) that again
revealed a main effect of training $(F(1,46)=8.16 ; p<.01)$ and an interaction of training and list $(F(1,46)=5.83 ; p<$ .05 ). Separate $t$-tests revealed that children produced significantly more DO responses for the DO-biased verb show than for the PO-biased verbs send or throw in PO-training $(t(47)=$ 2.7, $p<.01$ ) but not in DO-training $(t(47)<1)$.

Experiment 1 b thus reproduced the key findings of Experiment 1a, varying the verb sets. Again, the pattern of results suggested that the magnitude of verb-bias training depended on training-sentence surprisal. In Experiment 2 we sought evidence of the same surprisal effect in adults, again varying the verb sets.

## Experiment 2

## Methods

Participants Forty-eight college-aged adults participated, all native English speakers. Data from 2 additional adults were excluded due to low training compliance (1), or too few dative responses in the test block (1).

Materials and Procedures Materials and procedures were identical to those of Experiment 1a, except that send and hand were the restricted verbs for half of the participants and pass and show were the restricted verbs for the other half. Send and pass were both PO-biased verbs in our norming data; hand and show were both DO-biased. As in Experiment 1a, each participant was randomly assigned to either the withbias or the contra-bias list condition.

Adults who produced fewer than $80 \%$ training-compliant responses in each training condition were replaced. We also excluded one adult who did not produce at least one dative response in the test block for each restricted verb. The included participants produced training-compliant responses in $99 \%$ of training trials. Of the 384 possible responses in the test trials with restricted verbs, 8 were coded as Other responses, leaving 376 DO and PO responses.

## Results and Discussion

Fig-3c shows adults' proportion of DO responses in the re-stricted-verb test trials, by within-subjects training condition (PO- vs. DO-Training) and between-subjects list (with- vs. contra-bias). Adults, like children, showed an effect of training, producing more DO responses in the DO-training condition (54\%) than in the PO-training condition (34\%). They also showed effects of pre-existing verb bias, producing more DO responses with the DO-biased verbs ( $53 \%$ ) than with the PO-biased verbs ( $35 \%$ ), averaged across training conditions.

The pattern of data shown in Fig-3c again suggests that the effect of training varied with the likelihood of the training structure and its fit with the pre-existing verb bias. The difference between the (pre-experimentally) PO- vs. DO-biased verbs was reduced in the DO-training condition relative to the PO-training condition. This is just what we would predict based on training-sentence surprisal: DO-training strongly increased the rate of DO responding for the PO-biased verbs.

As in Experiment 1b, the same pattern emerged for both verb sets (not shown in the figure).

This pattern was supported by an ANOVA on the proportion of DO responses (arcsine transformed) that revealed a main effect of training $(F(1,46)=16.73, p<.001)$ and an interaction of training and list $(F(1,46)=12.92, p<.01)$. Separate t -tests revealed that the difference between the two verbs was significant in the PO-training condition $(t(47)=$ 2.35, $p<.05$ ) but not in the DO-training condition $(t(47)<$ 1 ), consistent with the surprisal predictions.

## General Discussion

In three experiments, we found that children and adults produced more double-object (DO) sentences for verbs trained in the DO structure than for verbs trained in the PO structure. This difference between training conditions replicates previous reports that the biases of familiar verbs can be altered by new verb-structure patterns in the input (Coyle \& Kaschak, 2008; Lin \& Fisher, 2016).

We also found the first evidence that the magnitude of the verb-bias training effect depended on the prior likelihood of the training sentences. The key result was that, as predicted, DO-training reduced the difference in DO responses between pre-experimentally DO- and PO-biased verbs. After DOtraining, a familiar PO-biased verb such as pass became almost as likely to be used in the DO structure as a familiar DO-biased verb such as show. In contrast, after PO-training, DO-biased verbs were still used much more often in the DO structure than were PO-biased verbs. This pattern supports the hypothesis that training-sentence surprisal affects verbbias learning. PO-training, which linked verbs with what is arguably the default dative structure, produced little change in the rate of DO responses relative to the verbs' pre-existing biases; DO-training, which linked verbs with a less canonical structure, led to sizable increases in the rate of DO responses, but did so mostly for PO-biased verbs, reducing the difference between the PO- and DO-biased verbs. This pattern was observed with 4 -year-olds (Experiments 1a and 1b) and with adults (Experiment 2).

These findings highlight a strong parallel between verb bias learning and syntactic priming. Prior evidence shows that the magnitude of syntactic priming depends on prime sentence surprisal: The largest priming effects are found when the prime structure is uncommon, or is unexpected given the verb in the prime sentence (Bernoulet \& Hartsuiker, 2010; Jaeger \& Snider, 2013; Peter et al., 2015). Here we saw strikingly similar effects for verb-bias learning. In both syntactic priming and verb-bias learning, children and adults learn more from unexpected sentences. This similarity suggests that syntactic priming, which involves learning about abstract syntactic structure, and verb bias learning, which involves linking verbs to syntax, depend on similar learning mechanisms and representations.

This conjecture fits well with the predictions of Chang, Dell and Bock's (2006) Dual-Path model of syntax learning. The model learns to link syntax and semantics without predefined syntactic representations, in a system that yokes a
syntactic sequencing system to a separate message system representing the meaning of input sentences. A key feature of the model is that the syntactic sequencing system is linked to abstract event-role slots in the message system, but not to the word-meanings bound to those event roles. This "Dual-Path" architecture keeps lexical semantics out of the syntax, ensuring that the model creates abstract syntactic representations. Accordingly, the model creates syntactic representations that support abstract syntactic priming, but the model can also learn about the syntactic biases of particular verbs under some circumstances (Chang, Janciauskas, \& Fitz, 2012). Because the model learns via error-based learning, it learns the most from input sentences that are unexpected given the model's prior experience. This model therefore provides one possible account of our findings-sentence surprisal affects verb bias learning as well as syntactic priming because the same error-based implicit learning mechanism underlies learning about abstract syntax and verb bias.

Our results leave open many questions for future research about the nature of the representations that were modified by verb-bias training. For example, participants could have strengthened the link between each verb and an abstract representation of sentence structure or between a verb and a thematic role ordering (Twomey, Chang, \& Ambridge, 2016). Training could have also highlighted the semantic difference between caused possession and caused motion, changing the prominence of recipient vs. theme. Note, though, that adapting the syntax and meanings of verbs are not mutually exclusive (Gleitman et al., 2005).

The verb bias learning studies reported here shed new light on a fundamental question in language acquisition: How do we coordinate abstract syntactic knowledge with our intricate knowledge of words? Our results suggest error-based implicit learning mechanisms help us track the likelihood of both abstract syntactic structures and the linking of those structures with particular verbs. The same learning mechanisms may underlie learning at both levels, creating both abstract and verb-specific syntactic knowledge throughout development.

## Acknowledgements

This research was supported in part by grants from the NSF (BCS - 1348522) and the NICHD (R01 HD054448). The stimulus videos were adapted from materials generously shared with us by Caroline Rowland.

## References

Bernolet, S., \& Hartsuiker, R. J. (2010). Does verb bias modulate syntactic priming? Cognition, 114, 455-461.
Bock, K. (1986). Syntactic persistence in language production. Cognitive Psychology, 18, 355-387.
Bock, K., \& Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? Journal of Experimental Psychology: General, 129(2), 177-192.
Brown, M., Savova, V., \& Gibson, E. (2012). Syntax encodes information structure: Evidence from on-line reading comprehension. Journal of Memory and Language, 66, 194-
209.

Chang, F., Dell, G., \& Bock, K. (2006). Becoming syntactic. Psychological Review, 113(2), 234-272.
Chang, F., Janciauskas, M., \& Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. Language and Linguistic Compass, 6(5), 259-278.
Coyle, J. M., \& Kaschak, M. (2008). Patterns of experience with verbs affect long-term cumulative structural priming. Psychonomic Bulletin \& Review, 15(5), 967-970.
Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., \& Trueswell, J. C. (2005). Hard words. Language Learning and Development, 1(1), 23-64.
Jaeger, F., \& Snider, N. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. Cognition, 127, 57-83.
Lin, Y., \& Fisher, C. (2016). Connecting verbs to syntax: Modifying verb bias. Poster presented at The $29^{\text {th }}$ CUNY Conference on Human Sentence Processing.
Peter, M. Chang, F., Pine, J. M., Blything, R., \& Rowland, C. F. (2015). When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. Journal of Memory and Language, 81, 1-15.
Qi, Z., Yuan, S., \& Fisher, C. (2011). Where does verb bias come from? Experience with particular verbs affects online sentence processing. Proceedings of the $35^{\text {th }} \mathrm{BUCLD}$.
Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M., \& Lieven, E. V. M. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. Cognition, 125, 49-63.
Ryskin, R. A., Qi, Z., Duff, M. C., \& Brown-Schmidt, S. (2016). Verb biases are shaped through lifelong learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(5), 781-794.
Snedeker, J., \& Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. Cognitive Psychology, 49, 238-299.
Stephens, N. (2015). Dative constructions and givenness in the speech of four-year-olds. Linguistics, 53(3), 405-442.
Thothathiri, M., \& Snedeker, J. (2008). Syntactic priming during language comprehension in three- and four-year-old children. Journal of Memory and Language, 58, 188-213.
Tomasello, M. (1992) First Verbs: A Case Study of Early Grammatical Development. Cambridge University Press.
Twomey, K. E., Chang, F., \& Ambridge, B. (2016). Lexical distributional cues, but not situational cues, are readily used to learn abstract locative verb-structure associations. Cognition, 153, 124-139.
Wonnacott, E., Newport, E. L., \& Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. Cognitive Psychology, 56, 165-209.

# Repeated Interactions Can Lead to More Iconic Signals 

Hannah Little (hannah.little@mpi.nl), Marcus Perlman (marcus.perlman@mpi.nl)<br>Language and Cognition Department, Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands

Kerem Eryilmaz (kerem@ai.vub.ac.be)<br>Artificial Intelligence Laboratory, Vrije Universiteit Brussel, 1050 Brussels, Belgium


#### Abstract

Previous research has shown that repeated interactions can cause iconicity in signals to reduce. However, data from several recent studies has shown the opposite trend: an increase in iconicity as the result of repeated interactions. Here, we discuss whether signals may become less or more iconic as a result of the modality used to produce them. We review several recent experimental results before presenting new data from multi-modal signals, where visual input creates audio feedback. Our results show that the growth in iconicity present in the audio information may come at a cost to iconicity in the visual information. Our results have implications for how we think about and measure iconicity in artificial signalling experiments. Further, we discuss how iconicity in real world speech may stem from auditory, kinetic or visual information, but iconicity in these different modalities may conflict.


Keywords: Iconicity; Modality; Artificial Language Experiment; Communication; Conventionalisation

## Introduction

Roughly 7000 languages are spoken around the world, and dozens more are signed. Over the course of human history, according to one rough estimate, hundreds of thousands of languages may have passed in and out of existence (Pagel, 2000). The number of words that have cycled through human languages, then, is enormous, perhaps in the order of billions. Imagine that we could trace these symbols back to their origins. How did people create the first words and signs?

One hypothesis is that the first words were created using iconicity (Fay, Ellison, \& Garrod, 2014; Imai \& Kita, 2014; Perlman, Dale, \& Lupyan, 2015; Perniss, Thompson, \& Vigliocco, 2010). Iconicity is a quality of a signal that, regardless of modality or medium, exhibits a degree of resemblance between its form and its meaning. For example, a person can communicate the idea of a 'rounded' shape by drawing a picture that resembles it, by molding their hands to reflect the shape, or by vocalising a 'round' word like "bouba". Iconicity can function to jump-start a new communication system because it enables a communicator to create new signals that are, to some extent, understandable to a partner without a shared system of conventional symbols.

The hypothesis that the first words were iconic improvisations is supported by evidence from natural signing systems. Traces of iconic creation are apparent in many of the signs of signed languages, and when signers lack a name for a referent, they tend to create an iconic sign for it (Klima \& Bellugi, 1979). Further, when deaf children are raised without native signers and deaf peers, they create iconic gestures that ground the development of home sign systems that they use with hearing adults (Goldin-Meadow, 2003). Experimental
studies where participants communicate using unfamiliar signalling systems also demonstrate extensive use of iconicity to ground novel signals, for example with drawing (Garrod, Fay, Lee, Oberlander, \& MacLeod, 2007), slide whistles (Verhoef, Roberts, \& Dingemanse, 2015), and non-linguistic vocalisations (Perlman et al., 2015).

In comparison to signed languages, the role of iconicity in the creation of spoken languages is obscure. It is widely assumed that spoken languages have markedly less iconicity than signed languages. Yet, it is unclear why this is the case. One widely argued reason is that the vocal-auditory modality affords little iconicity to represent a rich array of meanings (Armstrong \& Wilcox, 2007). This argument is supported mainly by comparing impressions of the iconicity of gesture and sign with vocalisations and speech, and also by experimental studies finding that gestures were more effective than non-linguistic vocalisations at communicating different meanings (Fay, Arbib, \& Garrod, 2013; Fay, Lister, Ellison, \& Goldin-Meadow, 2014). A second possible reason that spoken languages have so little iconicity is their extremely ancient origins. Over so many generations, the original iconicity of spoken languages has mostly degraded. This alternative assumes a process of conventionalisation in which the high level of iconicity characteristic of novel signals decays uni-directionally over time until it eventually disappears.

## Iconicity and conventionalisation

Is it actually the case that the iconicity of novel signals necessarily decays over time as the signal becomes conventionalised? In signed languages, the iconicity of signs does appear to fade over time as the forms become more regularised and systematic (Frishberg, 1975). Although mature signed languages are still iconic to a large extent, they are nevertheless much younger than spoken languages, and we do not know what might happen to their iconicity with further development. Graphic systems may provide a clearer case of how iconicity diminishes over time. For example, early records of written Sumerian, early Egyptian and ancient Chinese show that they originated from more detailed, iconic depictions that have became conventionalised into an increasingly abstract code (Gelb, 1952; Sampson, 1985; Vaccari \& Vaccari, 1964). A smaller-scale, but comparable, process for graphic systems has been demonstrated in the laboratory where drawings lose their iconicity and become more symbolic and arbitrary over repeated interactions (Caldwell \& Smith, 2012; Garrod et al., 2007; Theisen, Oberlander, \& Kirby, 2010).

However, recent experimental studies have found that sig-
nals may sometimes gain iconicity over repeated interactions, even as they otherwise show evidence of conventionalisation. In Perlman et al. (2015), pairs of participants took turns over ten rounds creating non-linguistic vocalisations for different meanings (e.g. big, rough, up). Accuracy within the game increased to ceiling, and vocalisations showed signs of conventionalisation, becoming shorter in duration and more stable in form. To measure how iconicity changed over this process, they tested the ability of naïve listeners to guess the meaning of vocalisations from rounds 1,5 , and 10 . Vocalisations from round 1 were guessed with the lowest accuracy, suggesting they were the least iconic, but in later rounds, vocalisations were guessed with higher accuracy. Verhoef, Kirby, and de Boer (2014) also found that signals increased in iconicity over repeated interactions and iterations. Participants used digital slide whistles to communicate different left or right facing animals. The results showed that participants only encoded the direction of these animals after 2 or 3 generations in an iterated chain.

On the surface, these findings may seem at odds with the idea that the function of iconicity is to bootstrap the formation of a conventional signal. How can signals become initially more iconic and then maintain their iconicity over time, even as they became more conventionalised? One explanation for this result is that the creation of iconic signals in vocalisation is more challenging than in modalities like drawing or gesture. Thus, partners may initially need to explore the signal space and negotiate their shared intuition for a meaningful vocalisation. Over interactions, as signals become streamlined, the strongly iconic features that are found to be effective in distinguishing its meaning tend to be enhanced, while more idiosyncratic features are shed.

## Experiments

## Stimuli

Stimuli for the experiments presented in this paper come from a previous experiment (Little, Eryılmaz, \& de Boer, in press). In this experiment, participants produced signals for meanings varying in shape, colour, and texture, which were designed to have no shared features (explained in Little et al., in press). Figure 1 shows the 15 meanings used in the experiment. Theremin-like signals were created using a "Leap Motion" controller: an infrared sensor that detects hand position (see Eryılmaz \& Little, 2016 for details of the paradigm). Participant's hand position determined the pitch of audio signals. Left to right hand-positions created low to high pitches respectively with a non-linear, exponential relationship between hand-position and pitch. Participants were given this audio feedback in real-time as they produced the signals and participants could not see each other as they produced signals. These signals were used because they share some qualities with speech: they are auditory, continuous and restrict the use of iconicity. At the same time, they are non-linguistic and so minimise possible interference from pre-existing linguistic knowledge and conventions.


Figure 1: The meanings used in the experiment.

The stimuli were created in two experimental conditions: an 'individual' condition, where one person produced signals and received their own signals in batches of 5, and a 'communication' condition where two participants took it in turns to produce and receive signals. When receiving signals, participants were asked to identify their referent from an array of 4 meanings. Feedback was given on the correct answer immediately after each response in both conditions.

The meaning space expanded throughout the experiment: by 5 meanings at a time in every block in the individual condition, and by 2 meanings at a time in the communication condition. In the communication condition, the meaning space only expanded once the participants had agreed on signals for existing meanings (by communicating them correctly twice).

For the experiments in this paper, signals from "early" in the individual condition were taken from the first phase (5 signals) and "late" signals were taken from the last phase (15 signals). In the communication condition, no pair managed to finish the experiment before time ran out, and so all of the data from the "last phase" in the current paper is referring to the last phase participants got to in their particular experiment. "Early" signals from the communication condition were for the first 2 meanings seen.

## Experiment 1: Audio playback experiment

We conducted a playback experiment to examine how the iconicity of signals changed over repeated interactions in the experiment above. Naïve listeners, without knowledge of a signal's development, guessed the meanings of the signals produced in the individual and communication conditions at both the beginning and end of the game. We took listeners' ability to match the signal with its intended referents as a measure of iconicity. This method for measuring iconicity has been used previously in a number of studies (e.g. Garrod et al., 2007; Perlman et al., 2015). The experiment tested two hypotheses, though it should be noted that both hypotheses could work in tandem, or represent different stages of emergence of a communication system.

Hypothesis 1 The first hypothesis is that in the communication condition, repeated interaction between two participants
will lead to initial signals that are high in iconicity, but then become less iconic over interactions. This would follow the results of experiments such as Garrod et al. (2007), that used drawings. Their results also suggest that we should not see a loss of iconicity in the individual condition as conventionalisation requires interaction between communicators.

Hypothesis 2 The second hypothesis is that iconicity will go up in the communication condition, in line with the findings of Perlman et al. (2015). If iconicity is not present from the beginning, or is very idiosyncratic, then interaction may act as a way for signals to adapt to be more transparently iconic. However, without an interlocutor, one would not expect there to be a pressure for transparency in any iconicity present, meaning signals in the individual condition should also not increase in iconicity under this hypothesis.

## Method

Procedure 391 participants were recruited on social media. Each participant was sent to a webpage which redirected randomly to one of several signal sets on its own webpage. A signal set was typically 15 signals. Signals were mp3 files which were playable by the participants by clicking on them. Under each mp3 file was a set of 4 images of possible meanings including the correct referent and 3 others chosen at random. Participants were asked to click on the meaning of the four that "you think the sound refers to". They could change their mind as many times as they liked, and their responses were only recorded after they pressed "submit" at the bottom of the page.

## Results and Discussion

The following results are all produced using a linear mixed effects analysis, from accuracy data that had been binnned by meaning. We included time phase (early or late) and condition (individual or communication) of production as fixed effects. The intended image was controlled for as a random effect with by-meaning random slopes for the effect of time phase and condition. Likelihood ratio tests were used to compare the model against a null model that did not include the variable of interest. The condition in which the signals were produced - individual or communication - did not appear to affect the iconicity of the signals $\left(\chi^{2}(1)=0.1, p=0.74\right)$. Listeners correctly matched signals with their referents with nearly the same level of accuracy in both conditions (around $35 \%$ ). The time phase in which the signal was produced also did not significantly affect guessing accuracy $\left(\chi^{2}(1)=2.3\right.$, $p=0.13$ ). However, there was a significant interaction between condition and time produced $\left(\chi^{2}(1)=5.9, p=0.015\right)$. In the graph (Figure 2), we show that naïve listeners were much better at matching signals that were produced later in the communication condition. In the individual condition, the signals went down slightly in their iconicity, though this difference was not significant.

The results from the audio playback experiment suggest that the iconicity of signals created at the start of the original


Figure 2: The percentage of signals correctly matched with their meanings by naïve listeners. The percentage for behaviour at chance levels is $25 \%$.
communication game is nearly the same in both the individual and communication conditions. Naïve listeners were able to guess their meanings with nearly equal accuracy. This is not indicative of participants not attempting to be iconic (indeed, accuracy was above chance), but it may be that their attempts to be iconic start as being relatively idiosyncratic. In further support of this account, in the individual condition, the iconicity of signals did not change from the early to the late phase of the communication game. It may be that with an individual participant, there is no selection pressure to enhance the strongly iconic features of signals or to discard more idiosyncratic ones.

In contrast to the individual condition, we found that in the communication condition iconicity increased significantly from the early to the late phase, confirming hypothesis 2. In media that afford less iconicity, the presence of another person might cause ongoing pressure to enhance the iconicity of signals, making them increasingly transparent to naïve listeners. However, because of the multi-modal nature of the signals in the initial experiment (i.e. gesture generating audio signals), it is also a possibility that participants were just becoming accustomed to being iconic using the audio feedback (the only thing transmitted between participants), rather than using iconicity in the visual modality.

In the original study, Little et al. (in press) found that signals in the individual condition became longer and more complex over the course of the experiment. However, there was no evidence of signals changing in complexity in the communication condition. Within the experiment, participants were better at correctly matching signals with their correct referent in the individual condition ( $85.6 \%$ correct) than in the communication condition ( $74.4 \%$ correct). In the individual condition, participants improved at recognising the signals correctly throughout the experiment, but they got worse in the communication condition.

## Experiment 2: Visual playback experiment

In this experiment, we look for evidence that participants were distinctly adapting the iconicity of their signals to be optimal for the communication modality. Many people's first instinct is to draw the shapes in the air, but this did not necessarily translate to an optimally iconic signal with respect to the auditory feedback that was generated. This auditory representation was the only information transmitted between participants as they could not see each other. Therefore, participants might have adapted their signals to be more iconic by sound, while at the same time discarding distracting features that turned out to be less iconic.

To examine more specifically how participants in the communication condition adapted their signals over the course of the experiment, we ran a second playback experiment where participants matched visual representations of the signals, instead of auditory ones. If signallers adapted their signals to enhance iconicity for the communication medium, but shed features that are less iconic, then naive guessing accuracy with the visual signals should not increase as it did for the auditory signals. Alternatively - as the visual signals include exactly that same information as the auditory signals, just mapped onto a spatial dimension - visual iconicity might increase along with auditory iconicity.

## Method

Stimuli The stimuli were the same signals used in the audio playback experiment for the communication condition but transformed into visual representations. Signals were small (200x200px) videos of the hand trajectory used to produce the audio signals. A black square moves left and right in real time with how participants' hands moved to produce the signals. These videos were produced using only information from the x -axis of the hand trajectory. We only used information from the x -axis because only the x -axis affected the pitch of signals. This gave the naïve participants in the visual playback experiment the same amount of information as in the audio experiment, making them more directly comparable.
Procedure 97 participants were recruited on social media. Again, each participant was linked to a webpage that redirected them to a webpage with one of several possible signal sets. The procedure was the same as in Experiment 1, except that the stimuli were presented as videos instead of as audio files. Participants were asked to watch 15 videos each and choose the meaning that "you think the video refers to" for each signal.

## Results and Discussion

We compared the results of the visual playback with the results from audio playback in Experiment 1. Again, these results are produced using a linear mixed effects analysis using data binned by meaning. For this experiment, time produced (early or late) and modality (audio or visual) were the fixed effects in the model. Meaning was controlled for as a random effect and the model had by-meaning random slopes for the
effect of time phase and modality. Likelihood ratio tests were used to compare the model against a null model that did not include the variable of interest.

Guessing accuracy in both modalities is shown in figure 3. The modality of the signals - visual or auditory - did not affect the overall accuracy of selecting the correct image $\left(\chi^{2}(1)=1.17, p=0.28\right)$. The time phase in which the signal was produced also did not significantly affect guessing accuracy across modalities $\left(\chi^{2}(1)=1.4, p=0.24\right)$. However, there was a significant interaction between modality and time phase ( $\chi^{2}(1)=5.9, p=0.015$ ). In the early phase, guessing accuracy was statistically equivalent in both modalities. However, in the later phase, while accuracy increased in the auditory condition, in the visual condition it dipped slightly (but not significantly).


Figure 3: The percentage of both visual and audio signals from the communication game matched with their meanings by naïve participants. The percentage for behaviour at chance levels is $25 \%$.

The results from the visual playback experiment demonstrate that signals produced at the beginning of the communication game exhibited a comparable level of iconicity in the auditory representations and the visual representations of the signals. However, as the iconicity appears to have increased in the auditory signals over the phases of the game, the iconicity of the visual transformations did not. This was the case even though the visual signals included the same information as the auditory signals. These results suggest that signallers in the communication game adapted their signals to be more iconic in ways that were particularly suited to the auditory communication channel. Features that may have been more iconic in a visual medium were not enhanced.

## General Discussion

In the first playback experiment, we found that naïve listeners were more accurate at guessing the meanings of signals produced in the later phases of the experiment, but only in the communication condition. The pressure to became more iconic was only present when signals were being negotiated in interaction between individuals. One possible confound
here was that the meaning space expanding more quickly in the individual condition. Given that the signal space has only a limited amount of information to iconically encode meanings without ambiguity, it may be the the expansion meant a loss of iconic information across the whole meaning space. However, this consideration does not account for the fact that the meaning space also expanded in the communication condition where iconicity rose.

In a second playback experiment, we found that iconicity appeared to be enhanced particularly for the auditory communication medium, for which participants may have had weaker intuitions for iconicity compared to a visual medium. Together our findings demonstrate how, under certain conditions, the iconicity of signals can increase over repeated interactions, perhaps especially in a modality that affords less potential for iconicity. This may happen as partners initially explore the signal space and negotiate their shared intuition for a meaningful signal. Over interactions, as signals becomes streamlined with conventionalisation, the strongly iconic features may be agreed upon and enhanced, while more idiosyncratic features are back-grounded. This may only be an initial step in grounding a communication system, in running the experiment for longer, the signals may very well tend towards losing their iconicity again.

The results of this study have implications for semiotics experiments using artificial communication modalities and how iconicity is understood and measured in these studies. Many studies now use continuous auditory feedback as a result of some kinetic input, such as slide whistles (Verhoef, Kirby, \& Boer, 2015), digital slide whistles (Verhoef, Roberts, \& Dingemanse, 2015) and the Leap Motion paradigm (Eryilmaz \& Little, 2016). Iconicity has been measured in signals generated from all of these paradigms, but not always in the same way. In Verhoef, Roberts, and Dingemanse (2015), the iconicity is measured by correlating the direction of stimuli (left or right facing animals) with the direction of pitch in a signal. Little, Eryılmaz, and de Boer (2015) measures iconicity by comparing the similarities between meanings with the similarities between signals, using information from the hand positions, rather than transformed values representing the auditory feedback. Verhoef, Kirby, and Boer (2015) asked naïve participants to rate how well signals "fit" the meanings they were paired with using auditory information alone. Importantly, none of these studies incorporate information from both the auditory and visual aspects of the signal in their measures for iconicity. Of course, there is a perfect correlation between movement and auditory feedback in all of these paradigms, but the results we present here suggest that iconicity may be perceived in very different ways depending on either the visual or auditory information. Some experiments using artificial continuous signal spaces do not have auditory feedback and are purely visual in nature (Galantucci, 2005; Verhoef, Walker, \& Marghetis, 2016). These visual signals are treated as a proxy for a human communication systems in the same way that the paradigms with auditory feed-
back above are. However, it may be important to examine whether the results from such paradigms may, in some cases, be modality-specific.

Of course, caution is required in considering how our findings might generalise to languages and other natural communication systems. There are several reasons for reservation: the linguistic knowledge of our participants, the constrained signal and semantic space, the limited nature of the interaction, and the short time-scale of the experiment. Nevertheless, one interesting point of comparison may be the multimodality of our signals. In real-world communication, multimodality comes not only in the combination of speech and gesture, but also in the auditory and visual information that is conveyed by speech alone (Massaro, 1998; McGurk \& MacDonald, 1976). This multi-modal nature of speech may impact how iconicity is encoded in speech. For instance, one common example of iconicity in spoken language is the /i/ phone for diminutive, as in words like teeny, itty-bitty (Ohala, 1994). This association has been found reliably across languages (Blasi, Wichmann, Hammarström, Stadler, \& Christiansen, 2016). But what features of the /i/ make it iconic? Is it that the high pitch of the second and third formants corresponds with the high-pitched vocalisations of small animals? Is it the kinesthetic feel of articulating the sound, which is produced by contracting the oral cavity? Or might it be the visual features of the vowel, such as the speaker's retracted lips which resemble a submissive facial expression? These are difficult questions to answer, but future experiments might examine multi-modal signals and how iconicity is differentially informative across different modalities.

## Further Work

The main reason for running the playback experiment with visual signals was the observation that participants were inclined to draw in the air as a starting point for novel signals. However, it is possible that this form of iconicity would only be evident from information from both the $x$ - and $y$ - axes. Though the $y$-axis did not affect the auditory feedback in any way, there was nothing to stop participants moving their hand vertically in the experiment. As a next step, we plan to create videos of participant's movements on both the $x$ - and $y$-axes to see if such representations would exhibit a higher level of iconicity as a starting point, which might then decay.

## Conclusion

In conclusion, we would like to challenge the oft-cited notion that languages consistently lose their iconicity over time. The work presented here and elsewhere (Perlman et al., 2015; Verhoef, Roberts, \& Dingemanse, 2015) demonstrates the dynamic nature of iconicity in the evolution of symbol systems, which may adapt to the communication modality and the context in which it is used. Thus the multitude of morphemes cycling through languages may not always be drifting towards arbitrariness. In some cases, words and signs may become more iconic with time. The lexicons of natural languages, whether spoken or signed, exist in a bal-
ance between iconicity and arbitrariness (Dingemanse, Blasi, Lupyan, Christiansen, \& Monaghan, 2015; Perniss et al., 2010).

## Acknowledgements

Thanks to Gareth Roberts for feedback on the first playback experiment that lead to the implementation of the second. The original experimental work was financed by the ERC starting grant project ABACUS, grant number 283435.

## References

Armstrong, D. F., \& Wilcox, S. (2007). The gestural origin of language. Oxford University Press.
Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., \& Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Sciences, 201605782.
Caldwell, C. A., \& Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. PloS One, 7(8), e43807.
Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., \& Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. Trends in Cognitive Sciences, 19(10), 603-615.
Eryılmaz, K., \& Little, H. (2016). Using leap motion to investigate the emergence of structure in speech and language. Behavioral Research Methods, doi:10.3758/s13428-016-0818-x.
Fay, N., Arbib, M., \& Garrod, S. (2013). How to bootstrap a human communication system. Cognitive science, 37(7), 1356-1367.
Fay, N., Ellison, M., \& Garrod, S. (2014). Iconicity: From sign to system in human communication and language. Pragmatics \& Cognition, 22(2), 244-263.
Fay, N., Lister, C. J., Ellison, T. M., \& Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. Frontiers in Psychology, 5, 354.
Frishberg, N. (1975). Arbitrariness and iconicity: historical change in american sign language. Language, 696-719.
Galantucci, B. (2005). An experimental study of the emergence of human communication systems. Cognitive Science, 29(5), 737-767.
Garrod, S., Fay, N., Lee, J., Oberlander, J., \& MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? Cognitive Science, 31(6), 961-987.
Gelb, I. J. (1952). A study of writing: The foundations of grammatology. The University of Chicago Press.
Goldin-Meadow, S. (2003). The resilience of language: What gesture creation in deaf children can tell us about how all children learn language. Psychology Press.
Imai, M., \& Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369(1651), 20130298.

Klima, E., \& Bellugi, U. (1979). The signs of language. Cambridge: Harvard University Press.
Little, H., Eryılmaz, K., \& de Boer, B. (2015). Linguistic modality affects the creation of structure and iconicity in signals. In D. C. Noelle et al. (Eds.), The 37th annual meeting of the cognitive science society (cogsci 2015) (p. 13921398). Austin, TX: Cognitive Science Society.

Little, H., Eryılmaz, K., \& de Boer, B. (in press). Conventionalisation and discrimination as competing pressures on continuous speech-like signals. Interaction Studies.
Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle (Vol. 1). Mit Press.
McGurk, H., \& MacDonald, J. (1976). Hearing lips and seeing voices. , 746-748.
Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols, \& J. J. Ohala (Eds.), Sound symbolism (pp. 325-347). Cambridge: Cambridge University Press.
Pagel, M. (2000). The history, rate and pattern of world linguistic evolution. The evolutionary emergence of language: social function and the origins of linguistic form, 391-416.
Perlman, M., Dale, R., \& Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. Royal Society open science, 2(8), 150152.
Perniss, P., Thompson, R., \& Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. Frontiers in psychology, 1, 227.
Sampson, G. (1985). Writing systems: A linguistic introduction. Stanford University Press.
Theisen, C. A., Oberlander, J., \& Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. Interaction Studies, 11(1), 14-32.
Vaccari, O., \& Vaccari, E. E. (1964). Pictorial chinesejapanese characters: a new and fascinating method to learn ideographs. Vaccari's Language Institute.
Verhoef, T., Kirby, S., \& Boer, B. (2015). Iconicity and the emergence of combinatorial structure in language. Cognitive Science.
Verhoef, T., Kirby, S., \& de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. Journal of Phonetics, 43, 57-68.
Verhoef, T., Roberts, S. G., \& Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In D. C. Noelle et al. (Eds.), The 37th annual meeting of the cognitive science society (cogsci 2015) (pp. 2481-2487). Austin, TX: Cognitive Science Society.
Verhoef, T., Walker, E., \& Marghetis, T. (2016). Cognitive biases and social coordination in the emergence of temporal language. In A. Papafragou, D. Grodner, D. Mirman, \& J. C. Trueswell (Eds.), The 38th annual meeting of the cognitive science society (cogsci 2016) (pp. 2615-2620). Austin, TX: Cognitive Science Society.

# What's worth the effort: Ten-month-old infants infer the value of goals from the costs of actions 

Shari Liu<br>Harvard University<br>Tomer Ullman<br>Massachusetts Institute of Technology<br>Josh Tenenbaum<br>Massachusetts Institute of Technology<br>Elizabeth Spelke<br>Harvard University


#### Abstract

Infants understand that people act in order to achieve their goals, but how can they tell what goals people find worthwhile? Here, we explore the thesis that human infants solve this problem by building a mental model of action planning, taking into account the costs of acting and the rewards actions bring. Consistent with this thesis, we found that 10 -month-old infants, after viewing an agent approach two objects equally often, inferred that the agent preferred the object whose attainment required a costlier action. Infants' responses generalized across changes in perceptual variables that distinguished one action from another (e.g. path length, angle of incline), suggesting that an abstract cost metric based on force or effort supported their judgments. These findings suggest that infants' knowledge about agents may be expressed as a generative model for action planning, which can then be inverted to identify the probable hidden causes for observed actions.


# Why Does Higher Working Memory Capacity Help You Learn? 

Kevin Lloyd (klloyd@gatsby.ucl.ac.uk)<br>Gatsby Computational Neuroscience Unit<br>25 Howland Street, London, W1T 4JG, UK

Adam Sanborn (A.N.Sanborn@warwick.ac.uk)
Department of Psychology, University of Warwick
University Road, Coventry, CV4 7AL, UK

David Leslie (d.leslie@lancaster.ac.uk)<br>Department of Mathematics and Statistics, Lancaster University<br>Lancaster, LA1 4YF, UK

Stephan Lewandowsky (stephan.lewandowsky@bristol.ac.uk)<br>School of Experimental Psychology and Cabot Institute, University of Bristol Clifton, BS8 1TU, UK


#### Abstract

Algorithms for approximate Bayesian inference, such as Monte Carlo methods, provide one source of models of how people may deal with uncertainty in spite of limited cognitive resources. Here, we model learning as a process of sequential sampling, or 'particle filtering', and suggest that an individual's working memory capacity (WMC) may be usefully modelled in terms of the number of samples, or 'particles', that are available for inference. The model qualitatively captures two distinct effects reported recently, namely that individuals with higher WMC are better able to (i) learn novel categories, and (ii) flexibly switch between different categorization strategies.


Keywords: Bayesian inference; particle filter; working memory; category learning; knowledge restructuring

## Introduction

Humans often behave in a manner consistent with Bayesian principles (Chater \& Oaksford, 2008) yet how they achieve this is unclear. Though simple in principle, exact Bayesian calculations are frequently intractable in real-world settings, leading to a need for approximations. In statistics and computer science, this challenge has been met through the development of powerful, general-purpose techniques for approximate Bayesian inference, such as Monte Carlo methods, which allow practical application of Bayesian methods in complex domains. The practical success of these techniques has naturally prompted an interest in whether people deal with uncertainty in an analogous manner (Griffiths, Vul, \& Sanborn, 2012). Importantly, such algorithms can approximate probabilistic inference arbitrarily well when sufficient time and memory are available, thereby providing a benchmark for ideal performance, but also display systematic deviations from the normative solution when resources are limited. These latter 'qualitative fingerprints' may be particularly illuminating when considering human cognition, where constraints on information-processing capacity are typically assumed. A salient example is provided by limits on working memory capacity (WMC; Cowan, 2001). While the exact nature of these limits remain the subject of debate, one prominent conception is that they reflect a limited resource which is
shared across representations and processes in working memory (e.g., Just \& Carpenter, 1992).

In the current work, we consider WMC limits within the context of Bayesian inference, asking whether WMC may be usefully modelled as a constraint on inferential resources. In particular, we model the learning process as one of particle filtering, in which a series of probability distributions is represented by a limited set of samples ('particles') which are sequentially updated over time (Griffiths et al., 2012). Higher WMC is then assumed to be implemented as a greater number of particles. This approach is applied to two recent experiments which indicate positive effects of higher WMC on two distinct aspects of categorization: (i) the facility with which novel categories are learned (Lewandowsky, 2011); and (ii) the ability to flexibly switch between different category representations or response strategies, referred to as knowledge restructuring (Sewell \& Lewandowsky, 2012). We show that both of these effects are qualitatively captured by a single model in which WMC is equated with the number of particles available for inference - i.e., the number of hypotheses about category structure that an individual can concurrently entertain.

## WMC and Category Learning

Lewandowsky (2011) measured participants’ WMC before testing category learning performance on the six classical problem types of Shepard, Hovland, and Jenkins (1961) (henceforth 'SHJ'). Each involves learning to assign a set of stimuli to category $A$ or $B$ based on their values on binary dimensions, but the problem types vary in the number of stimulus dimensions required to correctly perform classification. Consistent with the classical results, participants generally learned the Type I problem fastest, Type VI the slowest, and Types II-V at an intermediate rate. Crucially, WMC score was found to be positively correlated with category learning performance: higher WMC individuals tended to make fewer errors across all problem types.

## WMC and Knowledge Restructuring

Sewell and Lewandowsky (2012) assessed the relationship between WMC and performance in a knowledge restructuring (KR) task. Participants were guided to use one particular categorization strategy in a binary classification task before being instructed to switch to an alternative, equally-effective strategy (Fig 1A). The stimuli, rectangles of varying height with a vertical bar located at different locations along their base, belonged to category $A$ or $B$ depending on their position in category space (Fig 1B). Crucially, training stimuli (filled circles) were clustered into two separate regions of category space (as indicated by different colours), with categories arranged so that partial category boundaries (solid lines) could not be integrated in a coherent manner; neither partial boundary could be extended so as to allow accurate classification of all stimuli in the other cluster. A third, binary 'context' dimension was systematically mapped onto the two training clusters so that stimuli belonging to distinct clusters appeared in different colours (see example stimuli, lower Fig 1B).

At the task outset, participants were given information designed to guide them towards using one of two different strategies for co-ordinating partial categorization rules: (1) a knowledge partitioning (KP) strategy was encouraged by imparting that the context variable (colour) could be used to determine which dimension to use (rectangle height or bar position) for categorization; (2) a context-insensitive (CI) strategy was instead encouraged by highlighting that bar position could be used to determine which partial boundary to apply (i.e., regardless of context). Both strategies could support perfect performance but predicted different patterns of generalization when applied to new stimuli (open squares, Fig 1B) in a transfer test, thereby revealing which strategy was in use (Fig 1C). A summary 'context sensitivity' (CS) measure was applied to participants' test patterns to quantify the degree to which they generalized in a manner consistent with the KP (high CS) or CI (low CS) strategy (Fig 1D).

Critically, Sewell and Lewandowsky found evidence that individuals with higher WMC were more adept at switching between these different categorization strategies when instructed to do so, as measured by how much their CS scores changed between tests. This was interpreted in terms of greater 'knowledge restructuring', i.e., ability to coordinate different category representations or response requirements.

## Modelling Approach

Our model comprises three parts: 1) assumptions about how participants represent categories, specified in terms of an explicit generative process; 2 ) a procedure by which participants are assumed to infer categories in light of prior assumptions and experimental stimuli; and 3) a means for translating participants' beliefs into choice (i.e., a predicted category label). Our description focuses on how the modelling approach is applied to the KR task; the SHJ tasks are simpler and easily modelled with only minor modifications.


Figure 1: (A) Knowledge restructuring (KR) task design. (B) Experimental stimuli, depicted in category space: position of a vertically-oriented bar ( $x$-axis) vs. height of rectangle ( $y$ axis). Filled circles denote training stimuli; open squares denote test stimuli; solid lines indicate the partial rule boundaries. Two example stimuli are shown underneath. (C) 'Ideal' predicted response profiles given exclusive use of a contextinsensitive (CI; top row) or knowledge-partitioning (KP; bottom row) strategy during test. Darker shading indicates a higher probability of classifying as category $A$. (D) Average context-sensitivity (CS) scores across participants during transfer tests, indicating use of CI (low) or KP (high) strategy. Figures B-D adapted from Sewell and Lewandowsky (2012).

## Category Representation

A number of representational formats for categories have been discussed in the literature. Here, we opted to use classification and regression tree (CART) models (Breiman, Friedman, Olshen, \& Stone, 1984). Firstly, these are well-suited to cases in which categories are readily described in terms of simple rules, particularly if an ordering on these rules is suggested (as in the KR task). Secondly, the classification boundaries generated by CART models lead naturally to 'axis-aligned' generalization patterns like those observed in the KR task (participants' response profiles were very similar to those shown in Fig 1C), whereas producing this behaviour is non-trivial for other category models.

Briefly, CART models provide a flexible method for specifying the conditional distribution of a binary category label $y$ given a $p$-dimensional stimulus feature vector $\mathbf{x}=$ $\left(x_{1}, x_{2}, \ldots, x_{p}\right)$. In the KR task, for a given stimulus on trial $t$, we have $y_{t} \in\{A, B\}$ and a 3-dimensional input $\mathbf{x}_{t}=$ $\left(x_{t, 1}=\right.$ bar position ${ }_{t} \in \mathbb{R}, x_{t, 2}=$ height $_{t} \in \mathbb{R}, x_{t, 3}=$ context $_{t} \in$ $\{0,1\})$. The models work by recursively partitioning the input space into axis-aligned cuboids (similar to the partial boundaries in Fig 1B) and applying a simple conditional model to each region (e.g., probability that category label $=$
A). The sequence of partitions can be represented as a binary tree (Fig 2).

Formally, a binary tree structure T consists of a hierarchy of nodes $\eta \in T$. Nodes with children are internal nodes, while nodes without children are leaf nodes (Fig 2A). Each node is associated with a block $B(\eta) \subseteq \mathbb{R}^{p}$ of the input space as follows: the root node is associated with the entire input space, while each further internal node splits its block into two halves by selecting a single dimension $\kappa(\eta)=\{1, \ldots, p\}$ and location $\tau(\eta)$ on which to split (Fig 2B). The block of input space associated with a node $\eta$ is determined by the ranges on each dimension $j$ which it covers, and we denote the corresponding range $R_{j}^{\eta}=\left[R_{j}^{\eta,-}, R_{j}^{\eta,+}\right]$. We call the tuple $\mathcal{T}=(\mathrm{T}, \kappa, \tau)$ the decision tree.


Figure 2: (A) Simple binary tree with (internal) root node $\eta$ which splits into two 'leaf' nodes, $\eta_{L}$ and $\eta_{R}$. (B) Corresponding split of a two-dimensional input space. The root node $\eta$ is associated with the full input space, $B(\eta)$. Here, node $\eta$ is split on dimension $1, \kappa(\eta)=1$, at a location $\tau(\eta)$. This splits the input space into two blocks, $B\left(\eta_{L}\right)$ and $B\left(\eta_{R}\right)$, associated with the leaf nodes $\eta_{L}$ and $\eta_{R}$.

In addition to a decision tree $\mathcal{T}$ with $K$ leaf nodes, a parameter $\Theta=\left(\theta_{1}, \theta_{2}, \ldots, \theta_{K}\right)$ associates parameter value $\theta_{k}$ with the $k$ th leaf node. If a stimulus $\mathbf{x}$ lies in the region of the $k$ th leaf node, then $y \mid \mathbf{x}$ has distribution $f\left(y \mid \theta_{k}\right)$ for some parametric family $f$. It is typically assumed that, conditional on $(\Theta, \mathcal{T}), y$ values within a leaf node are i.i.d. and that $y$ values across leaf nodes are independent. Thus, letting $n_{k}$ denote the number of observations assigned to the $k$ th leaf node and letting $y_{k, i}$ denote the $i$ th observation of $y$ assigned to leaf $k$,

$$
\begin{equation*}
p\left(y_{1: n} \mid \mathbf{x}_{1: n}, \boldsymbol{\Theta}, \mathcal{T}\right)=\prod_{k=1}^{K} \prod_{i=1}^{n_{k}} f\left(y_{k, i} \mid \theta_{k}\right) \tag{1}
\end{equation*}
$$

where $n=\sum_{k=1}^{K} n_{k}$ is the total number of observations.
Prior beliefs about category structure can be formalized as a prior distribution on decision trees, specified via a stochastic generative process. Following Chipman, George, and McCulloch (1998), we set the prior probability of a node $\eta$ in tree structure T being split into children nodes to

$$
\begin{equation*}
p_{\mathrm{SPLIT}}(\eta, \mathrm{~T})=\frac{\alpha}{\left(1+d_{\eta}\right)^{\beta}} \tag{2}
\end{equation*}
$$

where $d_{\eta}$ denotes the depth of the node, and $\alpha<1$ and $\beta \geq$ 0 are parameters controlling expected tree size. Under this specification, the probability $p_{\text {SPLIT }}$ is a decreasing function of node depth, and decreases more steeply for large $\beta$.

In addition to this prior on tree structure T , we generally
assume that the probability of splitting on each dimension is equal,

$$
\begin{equation*}
p(\kappa(\eta)=j)=1 / p, \quad j=1, \ldots, p \tag{3}
\end{equation*}
$$

and that split location is then drawn uniformly from the node's range,

$$
\begin{equation*}
\tau(\eta) \mid \kappa(\eta)=j \sim \mathcal{U}\left(R_{j}^{\eta,-}, R_{j}^{\eta,+}\right) \tag{4}
\end{equation*}
$$

However, in the KR task, participants were guided towards a particular strategy by being told in the first instance that stimulus colour (KP-first condition) or bar position (CI-first condition) reliably indicated whether height or bar position was diagnostic of stimulus category. To incorporate this additional information, we assume a bias term $b \leq 1$ which assigns higher probability to splitting the root node $\eta_{0}$ on the dimension $j^{*}$ highlighted by instruction:

$$
p\left(\kappa\left(\eta_{0}\right)\right)= \begin{cases}b & \text { if } \kappa\left(\eta_{0}\right)=j^{*}  \tag{5}\\ \frac{1-b}{2} & \text { otherwise }\end{cases}
$$

The generative model is completed by the conditional probabilities of stimulus labels given the tree structure, $p\left(y_{1: t} \mid \mathbf{x}_{1: t}, \mathcal{T}\right)$. We assume that the $k$ th leaf node has an associated probability $\theta_{k}$ of generating label $A$,

$$
\begin{equation*}
p\left(y_{t} \mid \theta_{k}, \mathbf{x}_{t}\right)=\theta_{k}^{y_{t}}\left(1-\theta_{k}\right)^{1-y_{t}} \tag{6}
\end{equation*}
$$

and that this probability is an i.i.d. draw from a Beta distribution, $\theta_{k} \stackrel{\text { iid }}{\sim} \operatorname{Beta}\left(a_{0}, b_{0}\right)$. Standard analytical simplification then yields the marginal likelihood
$p\left(y_{1: t} \mid \mathcal{T}, \mathbf{x}_{1: t}\right)=\left(\frac{\Gamma\left(a_{0}+b_{0}\right)}{\Gamma\left(a_{0}\right) \Gamma\left(b_{0}\right)}\right)^{K} \prod_{k=1}^{K} \frac{\Gamma\left(n_{k A}^{t}+a_{0}\right) \Gamma\left(n_{k}^{t}-n_{k A}^{t}+b_{0}\right)}{\Gamma\left(n_{k}^{t}+a_{0}+b_{0}\right)}$,
where $n_{k A}^{t}$ and $n_{k}^{t}$. are respectively the number of instances of category $A$ and the total number of data points in the partition of leaf $k$ up to trial $t$. Note that for a given tree, this likelihood is higher for leaves assigned observations with homogenous labels, and these are exactly the partitions that constitute 'good' solutions to the categorization problem.

## Inference

Participants are assumed to approximate the sequence of posterior distributions $\left\{p\left(\mathcal{T} \mid \mathbf{x}_{1: t}, y_{1: t}\right)\right\}_{t=1}^{T}$ over trials. Given the implausibility of enumerating all possible trees, participants are assumed to represent a relatively small number of samples, i.e. hypotheses, from these posterior distributions which can be updated over time. In other words, we assume participants perform particle filtering.

Two aspects of the inference process which we now describe draw parallels with working memory. Firstly, similar to the idea of a limit on the number of items that can be held in working memory (Cowan, 2001), we assume there is a bounded number of hypotheses about category structure in this case, the particles which correspond to specific tree structures - that can be entertained at a given time. Secondly, similar to the notion that working memory is active (Baddeley, 1992), involving manipulation rather than merely
passive storage of items, we assume that inference involves a continual process whereby local transformations to current hypotheses are proposed, and which may be accepted or rejected. The latter process promotes diversity in the hypothesis set and continuous exploration of the hypothesis space.

In detail, we assume that on trial $t$, a participant's beliefs are represented by a small set of $L$ possible trees $\left\{\mathcal{T}^{(l)}\right\}_{l=1}^{L}$ with associated importance weights $\left\{w_{t}^{(l)}\right\}_{l=1}^{L}$. This set of trees constitutes the limited set of hypotheses putatively maintained in a working memory of capacity $L$. With the observation of the stimulus and category label on the next trial $t+1$, a proper reweighting of the $l$ th tree is given by the following update (Chopin, 2002):

$$
\begin{equation*}
w_{t+1}^{(l)} \propto w_{t}^{(l)} p\left(y_{t+1} \mid \mathcal{T}^{(l)}, \mathbf{x}_{t+1}, y_{1: t}\right) \tag{8}
\end{equation*}
$$

As standard within particle filtering methods, this reweighting process is alternated with a resampling stage in which very unlikely trees, i.e., those with very low weights, are discarded and replaced by replicates of more probable trees. A simple way of doing this is to sample $L$ times with replacement from the set $\left\{T^{(l)}\right\}$ with probabilities proportional to the updated weights $\left\{w_{t+1}^{(l)}\right\}_{l=1}^{L}$ (Gordon, Salmond, \& Smith, 1993). Following this resampling step, all particle weights are equalized to $1 / L$.

Additionally, this resampled particle set can then be rejuvinated (Chopin, 2002), reintroducing diversity and allowing continuous exploration of alternative solutions. This is the 'active' step which, we suggest, recalls conceptions of working memory as involving active manipulation of currentlystored items. Specifically, we may, without altering the targeted posterior distribution, propose transformations of trees from a Markov chain transition kernel $q_{t+1}\left(\cdot \mid \mathcal{T}^{(l)}\right)$ with appropriate stationary distribution $p\left(\mathcal{T} \mid \mathbf{x}_{1: t+1}, y_{1: t+1}\right)$. Closely following the transition kernel suggested by Chipman et al. (1998), we consider the scheme where for each tree $\left\{\mathcal{T}^{(l)}\right\}$, a new tree $\mathcal{T}^{(l) *}$ is proposed by randomly choosing among 3 possible transformations: (1) grow: randomly select a leaf node, then draw a splitting dimension and location from the prior; (2) prune: randomly select an internal node, then turn it into a leaf node by deleting all nodes below it; or (3) change: randomly select an internal node, then reassign it a splitting dimension and location by a draw from the prior. The proposed tree $\mathcal{T}^{(l) *}$ is then accepted with probability
$\alpha\left(\mathcal{T}^{(l)}, \mathcal{T}^{(l) *}\right)=\min \left\{\frac{p\left(\mathcal{T}^{(l) *} \mid \mathbf{x}_{1: t+1}, y_{1: t+1}\right) / q_{t+1}\left(\mathcal{T}^{(l) *} \mid \mathcal{T}^{(l)}\right)}{p\left(\mathcal{T}^{(l)} \mid \mathbf{x}_{1: t+1}, y_{1: t+1}\right) / q_{t+1}\left(\mathcal{T}^{(l)} \mid \mathcal{T}^{(l) *}\right)}\right\}$,
as per the standard Metropolis-Hastings algorithm.
We also need to model the effect of an instruction to switch categorization strategy. We assume that the effect is to change the prior distribution over trees, which is then combined with past observations to produce an updated posterior distribution. This update can be implemented via a simple reweighting operation on the set of trees.

To see how this works, consider the specific example where a participant has initially been guided to use the CI strategy
and after $t$ training sessions has in mind the set of weighted trees $\left\{\mathcal{T}^{(l)}, w_{t}^{(l)}\right\}_{l=1}^{L}$ approximating the target distribution under the prior appropriate to the CI strategy. We denote this target distribution $p_{C I}\left(\mathcal{T} \mid \mathbf{x}_{1: t}, y_{1: t}\right)$. The experimenter then instructs the participant to change to using the KP strategy. Assuming that the set of trees remains fixed, the associated tree weights now need to be changed to reflect the new target distribution $p_{K P}\left(\mathcal{T} \mid \mathbf{x}_{1: t}, y_{1: t}\right)$. This can be achieved by an importance weighting step, treating $p_{C I}\left(\mathcal{T} \mid \mathbf{x}_{1: t}, y_{1: t}\right)$ as the importance distribution. In particular, denoting a particle's weight before and after the instruction to switch as $w_{t}^{(l)-}$ and $w_{t}^{(l)+}$, respectively, the relevant reweighting is

$$
\begin{equation*}
w_{t}^{(l)+} \propto w_{t}^{(l)-} \frac{p_{K P}\left(\mathcal{T}^{(l)} \mid \mathbf{x}_{1: t}, y_{1: t}\right)}{p_{C I}\left(\mathcal{T}^{(l)} \mid \mathbf{x}_{1: t}, y_{1: t}\right)} \tag{9}
\end{equation*}
$$

To switch in the reverse direction - from the KP to CI strategy - the appropriate reweighting instead uses the ratio $p_{C I}\left(\mathcal{T}^{(l)} \mid \mathbf{x}_{1: t}, y_{1: t}\right) / p_{K P}\left(\mathcal{T}^{(l)} \mid \mathbf{x}_{1: t}, y_{1: t}\right)$.

## Choice

Participants are assumed to predict category labels based on their current hypotheses. Assuming a newly-resampled particle set with equal weights $1 / L$, a sample-based approximation to the predictive probability that a stimulus $\mathbf{x}_{t+1}$ has label $y_{t+1}=A$ is given by

$$
\begin{align*}
p\left(y_{t+1}=A \mid \mathbf{x}_{1: t+1}, y_{1: t}\right) & \approx \frac{1}{L} \sum_{l=1}^{L} p\left(y_{t+1}=A \mid \mathbf{x}_{1: t+1}, y_{1: t}, \mathcal{T}^{(l)}\right) \\
& =\frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\theta_{k} \mid \mathbf{x}_{1: t+1}, y_{1: t}, \mathcal{T}^{(l)}}\left[\theta_{k}\right] \tag{10}
\end{align*}
$$

Thus, an approximation to the predictive probability is given by an unweighted average of posterior means for $\theta_{k}$, where $k$ for the $l$ th particle is the index of the leaf node relevant to the input $\mathbf{x}_{t+1}$ in $\mathcal{T}^{(l)}$. In our case, the posterior mean is

$$
\begin{equation*}
\mathbb{E}_{\theta_{k} \mid \mathbf{x}_{1: t+1}, y_{1: t}, \mathcal{T}^{(l)}}\left[\theta_{k}\right]=\frac{n_{k A}^{t}+a_{0}}{n_{k}^{t}+a_{0}+b_{0}} \tag{11}
\end{equation*}
$$

## Results

## Rate of Learning

Lewandowsky (2011) found that WMC was positively correlated with category learning performance. We hypothesized that a greater number of particles, i.e. increasing $L$, would have a similar effect since, on average, one might expect the search for a 'good' (i.e., more probable) category structure to progress faster, and with less chance of getting stuck in local maxima, with a higher number of particles.

Figure 3 displays average simulated learning curves for the SHJ tasks when the number of particles is increased from 1 (Fig 3A) to 100 (Fig 3B). Though the effect is subtle, there is a general steepening of learning curves and a downward shift in initial error rate for problem Type I. A more systematic gauge of the effect is obtained by fitting exponential functions to such learning curves and comparing the size of the fitted coefficients as the number of particles is increased (a


Figure 3: Increasing the number of particles leads to faster category learning. Simulated learning curves for (A) 1 particle, and (B) 100 particles. Learning curves are averages over 100 simulations with other model parameters fixed ( $a_{0}=b_{0}=$ $1 ; \alpha=0.95, \beta=1$ ). (C) Learning rate as a function of number of particles. For each setting, the model is run 100 times and exponential curves fit to each individual learning curve. The resulting coefficients are averaged over both simulation runs and problem types to yield an aggregate 'learning rate'.
larger coefficient indicates a steeper learning curve). Figure 3C shows that the learning rate does increase with more particles, though the effect is small beyond $\approx 20$ particles.

Note that even without fitting the model parameters, the basic SHJ pattern of results - Type I easiest, Type VI hardest, and Types II-V clustered in between - is reproduced. Briefly, this results from the preference for simpler, or more parsimonious, hypotheses that arises naturally within the Bayesian framework. An advantage for the Type II problem relative to types III-V is not produced by the model here, but we note that any such advantage was extremely marginal in Lewandowsky (2011), and that the effect may arise only under specific conditions (cf. Kurtz, Levering, Stanton, Romero, \& Morris, 2013).

## Knowledge Restructuring

Sewell and Lewandowsky (2012) found a positive association between WMC and knowledge restructuring. In the model, increasing the number of particles also has a beneficial effect on the average degree of knowledge restructuring (Fig 4A), with an increased probability of being able to successfully switch strategy (Fig 4B).

This result arises from an enhanced ability to accurately represent the posterior distribution with a greater number of particles. Recall that strategy-switching was modelled by a change in posterior distribution, driven by the different priors underlying the distinct strategies; a simple way to track this change was by reweighting particles according to the new dis-


Figure 4: (A) In both the context-sensitive (CI)-first (left) and knowledge-partitioning (KP)-first (right) conditions, increasing the number of particles $L$ leads to a greater change in context sensitivity (CS) score on average when prompted to change strategy ( 1500 simulation runs per condition). (B) This is due to an increased probability $P$ (switch) of a successful switch ( $\Delta C S>0.5$ ). Lower inset: with fewer particles ( $L=20$ ), it will frequently occur that the model completely fails to switch $(\Delta C S=0)$. Upper inset: with more particles ( $L=100$ ), such failures are unlikely ( 3000 simulation runs; $b=0.9, a_{0}=b_{0}=1, \alpha=0.95, \beta=1$ ).
tribution (Eq. (9)). However, the success of this will depend on how well the particle set covers the support of the updated distribution. With a sufficiently large number of particles, at least some should be allocated to (previously) lower probability regions; if the new strategy corresponds to such a region, then appropriate reweighting can be applied. However, with a decreasing number of particles, representation of the posterior distribution may be so impoverished that such regions of low probability may not contain any particles at all, and so switching is not immediately possible.

## Discussion

Experiments suggest that higher WMC benefits learning of novel categories (Lewandowsky, 2011) and the ability to coordinate different category representations or response strategies (Sewell \& Lewandowsky, 2012). We framed such tasks in terms of inference, where individuals seek to infer the most probable category structure(s) given their prior assumptions and experimental observations/instructions. Further, we assumed that individuals approximate inference by representing and manipulating in working memory a relatively small number of hypotheses - samples, or 'particles' - about possible category structures. Our principal hypothesis was that by linking WMC with the number of such particles, we would observe similarly positive effects of higher WMC on performance. Simulation results were consistent with this hy-
pothesis: more particles in the model enhanced both category learning performance and the ability to switch between different categorization strategies.

These effects respectively arise due to increased search efficiency and what we might call 'representational adequacy'. Conceptualized in terms of search for more probable categories, the more resources (i.e., particles) available to search this space - i.e., the greater the number of hypotheses that one can entertain and manipulate within working memory - then the more likely it is that one will quickly discover good solutions, a process which draws natural parallels with the broader topic of problem-solving (Hambrick \& Engle, 2003; Newell \& Simon, 1972). Furthermore, a greater number of particles generally means that the posterior distribution over categories is more accurately represented - including those assigned lower probability - and this pluralism means that the model can more easily express alternative hypotheses when instructed to switch strategy, as operationalized by a reweighting of particles. This source of flexibility may also be relevant to so-called 'insight' problem-solving (Murray \& Byrne, 2005; Ohlsson, 1992).

The current work is preceded by a number of related lines of research. The HyGene model (Dougherty, Thomas, \& Lange, 2010; Thomas, Dougherty, Sprenger, \& Harbison, 2008), which emphasizes the importance of hypothesis generation and testing, includes the assumption that the number of hypotheses that can be entertained at a given time is limited by working memory constraints. Similarly, in their study of 'garden path' effects in sentence processing, Levy, Reali, and Griffiths (2008) suggested that difficulties in parsing such sentences correctly may be explained by constraints on the resources (i.e., number of particles) available for incremental parsing; their demonstration that a decreasing number of particles increases the probability of parse failure is exactly analogous to the mechanism suggested here in relation strategyswitching.

There are a number of avenues for future investigation. We have focused on qualitative effects here, but fitting the model to individual participants will be necessary for a more quantitative assessment; the obvious prediction is that highWMC individuals should tend to be fit best by a larger number of particles. Decomposing the relative contributions of particular features of the model, such as resampling, should also be explored, and quality of fit directly compared with 'single-particle' approaches (e.g., Bramley, Dayan, Griffiths, \& Lagnado, 2017). How the approach fares in domains beyond category learning is also of clear interest. More generally, Monte Carlo methods provide a rich source of ideas for psychological models - exploring how such methods may succeed or fail to illuminate aspects of human cognition is a substantial task for future research.

## Acknowledgments

This work was supported by the Gatsby Charitable Foundation (KL), and by EPSRC grant EP/I032622/1 (DL).

## References

Baddeley, A. (1992). Working memory. Science, 255, 556-559.
Bramley, N., Dayan, P., Griffiths, T., \& Lagnado, D. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. Psychological Review, 124(3), 301-338.
Breiman, L., Friedman, J., Olshen, R., \& Stone, C. (1984). Classification and regression trees. Belmont, CA: Wadsworth.
Chater, N., \& Oaksford, M. (Eds.). (2008). The probabilistic mind: Prospects for Bayesian cognitive science. Oxford University Press.
Chipman, H., George, E., \& McCulloch, R. (1998). Bayesian CART model search. Journal of the American Statistical Association, 93(443), 935-948.
Chopin, N. (2002). A sequential particle filter method for static models. Biometrika, 89(3), 539-552.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24(1), 87-114.
Dougherty, M., Thomas, R., \& Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. In B. Ross (Ed.), The psychology of learning and motivation (Vol. 52, pp. 299-342). Burlington: Academic Press.
Gordon, N., Salmond, D., \& Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. Radar and Signal Processing, IEE Proceedings F, 140(2), 107-113.
Griffiths, T., Vul, E., \& Sanborn, A. (2012). Bridging levels of analysis for probabilistic models of cognition. Current Directions in Psychological Science, 21(4), 263-268.
Hambrick, D., \& Engle, R. (2003). The role of working memory in problem solving. In J. Davidson \& R. Sternberg (Eds.), The psychology of problem solving (pp. 176-206). Cambridge University Press.
Just, M., \& Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99, 122-149.
Kurtz, K., Levering, K., Stanton, R., Romero, J., \& Morris, S. (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(2), 552-572.
Levy, R., Reali, F., \& Griffiths, T. (2008). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, \& L. Bottou (Eds.), Advances in Neural Information Processing Systems 21.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(3), 720-738.
Murray, M. A., \& Byrne, R. M. (2005). Attention and working memory in insight problem solving. In B. Bara, L. Barsalou, \& M. Bucciarelli (Eds.), Proceedings of the xxvii annual conference of the cognitive science society (pp. 1571-1575). Lawrence Erlbaum Associates.
Newell, A., \& Simon, H. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Ohlsson, S. (1992). Information processing explanations of insight and related phenomena. In M. Keane \& K. Gilhooly (Eds.), Advances in the psychology of thinking. London: HarvesterWheatsheaf.
Sewell, D., \& Lewandowsky, S. (2012). Attention and Working Memory Capacity: Insights From Blocking, Highlighting, and Knowledge Restructuring. Journal of Experimental Psychology: General, 141(3), 444-469.
Shepard, R. N., Hovland, C. I., \& Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs: General and Applied, 75(13), 1-42.
Thomas, R., Dougherty, M., Sprenger, A., \& Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. Psychological Review, 115(1), 155-185.

# Social Network Limits Language Complexity 

Matt Lou-Magnuson<br>Nanyang Technological University, Singapore, Singapore<br>\section*{Luca Onnis}<br>Nanyang Technological University, Singapore, Singapore


#### Abstract

Natural languages vary widely in the degree to which they make use of hierarchical composition in their grammars, in particular, the degree to which syntactic versus morphologi- cal means of composition are utilized. Languages historically spoken in small communities develop much deeper levels of morphological embedding than those spoken by larger groups, an observation confirmed by a statistical analysis of the World Atlas of Language Structures. However, beyond population alone, social networks change in topological structure as they grow, and it may be the pattern of connectivity rather than number of speakers driving these differences. To examine mechanistically this connection between social and linguistic structure, we propose an agent-based model of grammatical change using complex network methods. We identify global transitivity as a physical parameter of social networks critical for developing morphological structure, and hubs associated with scale-free networks as inhibitory, encouraging syntactic composition instead.


# Learning induced illusions: Statistical learning creates false memories 

Yu Luo (yuluo@psych.ubc.ca)<br>Department of Psychology<br>University of British Columbia

Jiaying Zhao (jiayingz@psych.ubc.ca)<br>Department of Psychology<br>Institute for Resources, Environment and Sustainability<br>University of British Columbia


#### Abstract

The cognitive system readily extracts regularities in terms of object co-occurrences over space and time through statistical learning. However, how does learning such relationships influence the memory representations of individual objects? Here we used a false memory paradigm to examine the impact of statistical learning on memory representations of individual objects. Observers were exposed to a temporal sequence (Experiment 1) or spatial arrays (Experiment 2) of objects which contained object pairs (e.g., A-B). In a subsequent recognition phase, observers viewed a sequence or an array containing only one member of the original pair, and judged whether either the presented object or the missing object in the original pair was present. We found that statistical learning not only sharpened the detection of the presented object, but also induced a false memory of the missing object. This reveals a novel consequence of statistical learning: learning of regularities can create illusory memories.


Keywords: Statistical learning; false memory; implicit learning; regularities;

## Introduction

A remarkable ability of the cognitive system is to detect and learn the relationships among objects in the environment. Statistical learning is one mechanism that extracts the statistical relationships between individual objects in terms of object co-occurrences over space and time (Fiser \& Aslin, 2001; Saffran, Aslin, \& Newport, 1996). This process occurs incidentally, without conscious intent or explicit awareness, produces knowledge about object associations that people are not explicitly aware of (Turk Browne, Jungé, \& Scholl, 2005; Turk-Browne, Scholl, Chun \& Johnson, 2009), and can operate in multiple sensory modalities and feature dimensions (Conway \& Christiansen, 2005; Fiser \& Aslin, 2001; Saffran et al., 1996; Turk-Browne, Isola, Scholl, \& Treat, 2008). In addition, several cognitive consequences of statistical learning have been identified, such as the compressing of information (Brady, Konkle, \& Alvarez, 2009; Zhao \& Yu, 2016), attentional prioritization of cooccurring objects (Yu \& Zhao, 2015; Zhao, Al-Aidroos, \& Turk-Browne, 2013; Zhao \& Luo, 2017), and enhanced memory representation (Kim, Lewis-Peacock, Norman, \& Turk-Browne, 2014; Otsuka \& Saiki, 2016).

One important but unexplored question is: how does learning statistical associations influence the representation of individual objects? Initial evidence comes from studies on
false memories of semantically related objects. One pioneer work by Roediger and McDermott (1995) shows that after memorizing a list of words (e.g., nurse, sick, medicine, etc.) that are highly related to a target word that was never present (e.g., doctor), people falsely remember seeing the target word, and label it as an "old" word of the list in the recognition task. This finding was replicated using a visual paradigm, where participants viewed a stereotypical scene (e.g., classroom), and falsely recalled and recognized a target object that was never present (e.g., chalkboard, Miller \& Gazzaniga, 1998). One explanation for this phenomenon is that seeing one object can automatically activate other associated objects based on semantic memory (Roediger, Balota, \& Watson, 2001).

Here, we provide a new explanation behind this old phenomenon which focuses on a simpler mechanism: learning the co-occurrences of objects can create the false memories of non-present object when only its partner is present. We propose that the mere statistical co-occurrence of two objects can produce false memory, independent of semantic associations. Thus, the goal of the current study was to examine whether statistical learning alters the representations of individual objects.

## Experiment 1

This experiment examined how statistical learning alters the representations of individual objects in a temporal context, by first exposing participants to a sequence of paired objects and then testing them on whether seeing an individual object in the pair can produce the false memory of the non-present object in the pair.

## Participants

A total of 120 undergraduates ( 96 female; mean age $=20.6$ years, $\mathrm{SD}=2.8$ ) from University of British Columbia (UBC) participated in the experiment for course credit. Participants reported normal or corrected-to-normal visual acuity and provided informed consent. The protocol was approved by the UBC Research Ethics Board.

## Stimuli

The stimuli consisted of eight real-world objects (Fig.1a) which were selected from a stimulus set in a previous study
(Brady, Konkle, Alvarez, \& Oliva, 2008). All objects were converted to grayscale, and were adjusted to a mean brightness of 84 . Each object subtended $2.8^{\circ}$ of visual angle. The eights objects were randomly assigned into four pairs for each participant and remained constant throughout the experiment (Fig.1a). In each pair, the first object was always followed by the second object. The random assignment of objects into pairs ensured that there was no systematic semantic relation between two objects in a pair, but rather the two objects were associated with co-occurrences. Each pair was repeated 50 times to form a single continuous temporal sequence of objects in a pseudorandom order with a constraint where no single pair could repeat back-to-back.

## Apparatus

Participants in all experiments were seated 50 cm from a computer monitor (refresh rate $=60 \mathrm{~Hz}$ ). Stimuli were presented using MATLAB and PsychophysicsToolbox (http://psychtoolbox.org).

## Procedure

The experiment consisted of two conditions. In the structured condition, the eight objects were grouped into four pairs. In the random condition, the eight objects appeared in a random order in the sequence. Participants were randomly assigned to one of the two conditions ( $N=60$ in each). The experiment contained three phases: exposure phase, recognition phase, and test phase. During the exposure phase, one object appeared at the center of the screen for 500 ms followed by a 500ms inter-stimulus interval (ISI) in each trial (Fig.1b). Participants performed a 1-back task where they judged as quickly and accurately as possible whether the current object was the same as or different from the previous object (by pressing the "/" or " $z$ " key for same or different, respectively, key assignment counterbalanced). For the 1-back task, each object had a $20 \%$ chance of repeating the previous object, producing 480 trials in total. This 1-back task served as a cover task which was irrelevant to learning the object pairs, in order to conceal the true purpose of the study. This ensured that learning of the object pairs was incidental. Participants were not told anything about the object pairs.

After exposure, participants performed a recognition phase (Fig.1c). In each trial, participants viewed a continuous sequence of objects first and then judged whether a certain object was present in the sequence. In the structured condition, there were three types of trials. The first type was missing trials: the sequence contained all four pairs, except for one pair, one member was missing, and observers judged whether the missing member was present in the sequence. The missing trials measured the false alarm rate for the missing object. The second type was presented trials: the sequence contained all four pairs, except for one pair, one member was missing, but this time observers judged whether the presented member was present in the sequence. The presented trials measured the hit rate for the presented object. The third type was baseline trials: the sequence contained all four pairs, and observers judged whether one member in a
pair was present in the sequence. The baseline trials measured the hit rate for the presented object. In the missing trials, each member of an original pair was missing for once, resulting in 8 trials. In the presented trials, the presented object of an original pair was tested once, resulting in 8 trials. In the baseline trials, each member of a pair was tested once, resulting in 8 trials. The 24 trials were repeated twice, producing 48 trials in total (order of the trials was randomized). In the random condition, the trials were the same, except the objects in the sequence appeared in a random order, so the sequence contained no pairs. Each object was presented for 500 ms followed by a 500 ms ISI. After the sequence was presented, a 3000 ms blank screen followed. After the blank screen, an object was presented on the screen as a probe, and participants judged whether the object was presented in the previous sequence (by pressing the " 1 " or " 0 " key for "yes" or "no", respectively). The object remained on the screen until response.

## (a) Pairs


(b) Exposure phase: 1-back task over a sequence of pairs

(c) Recognition phase:

(d) Test phase: Which pair looks more familiar?


Figure 1. Experiment 1. (a) Four object pairs (e.g., A-B). (b) Exposure phase: 1-back task. (c) Recognition phase: missing trials, presented trials, and baseline trials. (d) Test phase: two-alternative forced-choice task.

To examine whether they had successfully learned the object pairs, participants in the structured condition completed a surprise two-alternative forced choice test phase following the recognition phase (Fig.1d). In each trial, participants viewed two sequences of objects. Each object appeared for 500 ms followed by a 500 ms ISI, and each sequence was separated by a 1000 ms pause. Participants judged whether the first or second sequence looked more familiar based on what they saw in the exposure phase (by pressing the " 1 " or " 0 " key for "first sequence" or "second sequence", respectively). One sequence was a pair (e.g., AB), and the other was a "foil" (e.g., A-D) composed of one object from an original pair (e.g., A-B), and the other from a different pair (e.g., C-D), while preserving the temporal positions in the pairs (Fig. 1d). Each pair was tested against each foil twice, which resulted in 16 trials in total (4 pairs $\times$ 2 foils $\times 2$ repetitions). Importantly, each pair and foil were presented the same number of times at test. Thus, to discriminate the pair from the foil, participants needed to know which two specific objects followed each other. The order of the trials was randomized, and whether the pair or foil appeared first was counterbalanced across trials. Participants in the random condition was not tested, since there were no pairs in the sequence.

A debriefing session was conducted at the end of the experiment, where participants were asked if they had noticed any objects that appeared one after another. For those who responded yes, we further asked them to specify which objects followed each other.

## Results and Discussion

At the test phase, pairs were chosen over foils on $60.0 \%$ ( $\mathrm{SD}=19.3 \%$ ) of the time, which was reliably above chance (50\%) $[t(59)=4.01, p<.001, d=0.52]$. Thus, learning of the object pairs was successful. During debriefing, 12 participants reported noticing the pairs, but none could correctly report which specific objects followed each other. This suggests that participants had no explicit awareness of the object pairs.

The false alarm rate (FA) and the hit rate in the recognition phase were presented in Fig. 2 and analyzed with a 2 (condition: structured vs. random; between-subjects) $\times 3$ (trial type: missing vs. presented vs. baseline; withinsubjects) mixed-effects ANOVA.

There was a main effect of condition $[F(1,118)=12.38$, $\left.p<.001, \eta_{p}{ }^{2}=.09\right]$ and trial type $[F(2,236)=425.06, p<.001$, $\left.\eta_{p}{ }^{2}=.78\right]$, but no reliable interaction between condition and trial type $\left[F(2,236)=0.73, p=.48, \eta_{p}^{2}=.006\right]$. Tukey's HSD post-hoc test showed that the FA rate of the missing trials was reliably higher in structured condition ( $27.9 \%$ ) than in random condition ( $20.4 \%$ ), $p=.03$; the hit rate of the presented trials was reliably higher in structured condition (72.5\%) than in random condition ( $62.60 \%$ ), $p<.001$; and a marginal difference in the hit rate of baseline trials between structured $(72.5 \%)$ and random condition ( $65.9 \%$ ), $p=.09$.

These findings suggest that statistical learning not only sharpens the memory of the object within the pairs, but also induces the false memory of the missing object.


Figure 2: The false alarm (FA) rate and the hit rate in recognition phase (error bars reflect $\pm 1 \mathrm{SEM} ;{ }^{\dagger} p<.1,{ }^{*} p<.05,{ }^{* * *} p<.001$ ).

To examine the relationship between statistical learning and recognition performance, we found that there were no correlations between learning of the pairs at the test phase and the FA rate or the hit rate. However, in structured condition there was a weak correlation between the FA rate in the missing trials and the hit rate in the presented trials. There was a moderate correlation between the FA rate in the missing trials and the hit rate in the baseline trials, but no correlations in random condition (Table 1).

Table 1: Correlations between learning of the pairs at the test phase and the false alarm rate or the hit rate

| Condition | Correlation | Correlation results |
| :--- | :--- | :--- |
| Structured | Learning vs. missing | $r(58)=-.15, p=.26$ |
| $(N=60)$ | Learning vs. presented | $r(58)=.21, p=.11$ |
|  | Learning vs. baseline | $r(58)=-.03, p=.79$ |
|  | Missing vs. presented | $r(58)=.28, p=.03$ |
|  | Missing vs. baseline | $r(58)=.31, p=.02$ |
|  | Presented vs. baseline | $r(58)=.48, p<.001$ |
| Random | Missing vs. presented | $r(58)=.17, p=.21$ |
| $(N=60)$ | Missing vs. baseline | $r(58)=.20, p=.12$ |
|  | Presented vs. baseline | $r(58)=.63, p<.001$ |

## Experiment 2

This experiment aimed to generalize the findings in Experiment 1 from the temporal context to a spatial context.

## Participants

A new group of 68 undergraduates ( 51 female, mean age $=20.2$ years, $\mathrm{SD}=2.4$ ) from UBC participated in the experiment for course credit.

## Stimuli

The stimuli were identical to those in Experiment 1, except that in structured condition the four pairs were grouped into horizontal, vertical, and diagonal spatial configurations (Fig.3a). Each array contained all four pairs, and was placed on an invisible $4 \times 4$ grid (subtending $8.2^{\circ} \times 8.2^{\circ}$ ) with the constraint that one pair was adjacent to at least another pair. This was to prevent participants from learning the pairs based on spatial segmentation cues other than object cooccurrences. In random condition, the eight objects were randomly assigned to one of cell on the grid, with the constraint that each object neighbored at least one other object. Thus, the only difference between structured and random condition was the presence or absence of the pairs.

## Procedure

As in Experiment 1, there were two conditions (i.e., structured vs. random, $N=34$ in each) and three phases (i.e., exposure, recognition, and test). In the exposure phase, participants in both conditions viewed arrays of objects, and performed a duplicate detection task where they judged as quickly and accurately as possible whether there were two identical objects in a single array (by pressing the "/" or "z" key for yes or no, respectively, key assignment counterbalanced, Fig. 3b). This duplicate detection task served as a cover task irrelevant to statistical learning, to ensure that learning of the object pairs was incidental. Participants were not told anything about the object pairs. Each array was presented on the screen for 1000 ms followed by a 1000 ms ISI in each trial. There were 480 trials in total, and $20 \%$ of the trials ( 80 trials) contained a duplicate object in the array.

The recognition phase was identical to that in Experiment 1, except that objects were presented all at once on the screen. In each trial, participants viewed an array for 800 ms followed by a 3000 ms pause, and judged whether the probe object was presented in the array. The display time was increased to 800 ms , as it required more time for participants to view all eight objects at once. As before, there were three types of trials: (1) missing trials, where one member in the pair was missing, and the missing object was tested; (2) presented trials, where one member in the pair was missing, but the presented object in the pair was tested; and (3) baseline trials, where all pairs were presented, and one object was tested (Fig. 3c).

After the recognition phase, participants in the structured condition completed the surprise two-alternative forced choice test phase to see whether they had successfully learned the object pairs (Fig. 3d). In each trial, one set of objects was presented on the left and another on the right side of the screen for 1000 ms . Participants judged whether the left or right set of objects looked more familiar based on what they saw in the exposure phase (by pressing the " 1 " or " 0 " key for "left" or "right", respectively). The foils were created following the same logic as in Experiment 1. Participants in the random condition was not tested, since there were no pairs in the array during exposure.

A debriefing session was conducted after test, where participants were asked if they had noticed any objects that appeared with one another. For those who responded yes, we further asked them to specify which objects appeared adjacent to each other.

## Results and Discussion

At the test phase, pairs were chosen over foils on $52.2 \%$ ( $\mathrm{SD}=10.3 \%$ ) of the time, which was not reliably above chance $(50 \%)[t(33)=1.25, p=.22, d=0.21]$. This suggests that participants failed to learn the spatial co-occurrences between the two objects in the pairs. During debriefing, four participants reported noticing the pairs, but none could correctly report which specific objects appeared with each other. This suggests that participants had no explicit awareness of the pairs.
(a) Pairs

(b) Exposure phase: duplicate detection task

(c) Recognition phase:


Figure 3. Experiment 2: (a) Four pairs in four different spatial configurations. (b) Exposure phase: duplicate detection task. (c) Recognition phase: missing trials, presented trials, and baseline trials. (d) Test phase: two-alternative forced-choice task.

The FA rate and the hit rate in the recognition phase were analyzed with a 2 (condition: structured vs. random; between-
subjects) $\times 3$ (trial type: missing vs. presented vs. baseline; within-subjects) mixed-effects ANOVA. There was a main effect of trial type $\left[F(2,132)=137.32, p<.001, \eta_{p}^{2}=.68\right]$, but no main effect of condition $\left[F(1,66)=0.004, p=.95, \eta_{p}{ }^{2}=.00\right]$ and no significant interaction between condition and trial type [ $\left.F(2,132)=0.45, p=.63, \eta_{p}{ }^{2}=.007\right]$. Tukey's HSD post-hoc test showed that the FA rate of the missing trials was not different between the structured condition $(30.7 \%)$ and the random condition ( $30.5 \%$ ), $p=.99$, the hit rate of the presented trials was not different between the structured condition (57.5\%) and the random condition ( $59.4 \%$ ), $p=.99$, and no difference in the hit rate of baseline trials between the structured $(66.7 \%)$ and the random conditions ( $64.3 \%$ ), $p=.97$ (Fig. 4).


Figure 4: The false alarm rate and the hit rate in recognition phase (error bars reflect $\pm 1$ SEM).

We found no correlation between learning of the pairs at the test phase and the FA rate or the hit rate. But in both the structured and random condition, there was a correlation between the FA rate, and the hit rate in the presented trials and in the baseline trials (Table 2).

Table 2: Correlations between learning of the pairs at the test phase and the false alarm rate or the hit rate

| Condition | Correlation | Correlation results |
| :--- | :--- | :--- |
| Structured | Learning vs. missing | $r(32)=.24, p=.17$ |
| $(N=34)$ | Learning vs. presented | $r(32)=.22, p=.22$ |
|  | Learning vs. baseline | $r(32)=.16, p=.35$ |
|  | Missing vs. presented | $r(32)=.66, p<.001$ |
|  | Missing vs. baseline | $r(32)=.57, p<.001$ |
|  | Presented vs. baseline | $r(32)=.40, p=.02$ |
| Random | Missing vs. presented | $r(32)=.44, p=.009$ |
| $(N=34)$ | Missing vs. baseline | $r(32)=.53, p<.001$ |
|  | Presented vs. baseline | $r(32)=.62, p<.001$ |

The lack of memory difference between the structured condition and the random condition could be due to the lack of learning of object pairs in the spatial context.

To further explore whether learning of spatial pairs changed the representation of individual objects in the pairs,
we separated participants who successfully learned the pairs (those who chose pairs over foil above chance, $N=15$ ), and those who failed to learn the pairs (those who chose pairs over foil at or below chance, $N=19$ ) in the structured condition. Among participants who showed learning, pairs were chosen over foils on $61.2 \%$ ( $\mathrm{SD}=6.8 \%$ ) of the time, which was reliably above chance (50\%) $[t(14)=6.44, p<.001, d=1.66]$. Only one participant reported noticing the pairs, but could not correctly report which specific objects appeared with each other. A 2 (group: learners vs. non-learners; betweensubjects) $\times 3$ (trial type: missing vs. presented vs. baseline; within-subjects) mixed-effects ANOVA revealed a main effect of trial type $\left[F(2,64)=68.40, p<.001, \eta_{p}^{2}=.68\right]$, but no main effect of group $\left[F(1,32)=1.25, p=.27, \eta_{p}{ }^{2}=.04\right]$ and no significant interaction between group and trial type $\left[F(2,64)=0.42, p=.66, \eta_{p}{ }^{2}=.01\right]$. Although the results were not reliably different between the two groups, the learners consistently showed numerically greater FA rate and hit rate than the non-learners (Fig.5), a pattern that was consistent with the findings in Experiment 1.


Figure 5: The false alarm rate and the hit rate of learners and nonlearners in recognition phase in the structured condition (error bars reflect $\pm 1$ SEM) .

## General Discussion

The goal of this experiment was to examine whether statistical learning alters the memory representations of individual objects. We found that after learning the temporal co-occurrences of objects, participants showed a reliably higher false alarm rate of seeing a missing object, and a reliably higher hit rate of seeing a presented object (Experiment 1). When the objects co-occurred over space, participants did not successfully express learning of pairs, and therefore did not show differential false alarm and hit rates (Experiment 2). However, with a more detailed analysis, participants who successfully learned the spatial pairs showed numerically higher false alarm rate of the missing object and numerically higher hit rate of the presented object than those who failed to learn the pairs. The current findings suggest that statistical learning not only sharpens the
detection of the objects within the pairs, but also induces a false memory of the missing object.

Induced false memory of the missing object can be explained by the automatic statistical association between the missing object and the presented object in the pair. Once the pairs were learned over repeated exposures even implicitly, one member in the pair could serve as a cue to signal the presence of its partner (Turk-Browne, et al., 2009). Thus, participants may have automatically brought the missing object to mind when seeing its partner in the sequence, thus false recalling that the missing object was present. This suggests that the automatic activation of the missing object was possible by merely co-occurring with its partner previously.

Alternatively, the two co-occurring objects may be unitized after learning. Previous studies have demonstrated that regularities compress information (Brady et al., 2009) and reduce perceived numerosity of the objects (Zhao \& Yu, 2016), which suggests that the co-occurring objects could be grouped and encoded as one single unit. Seeing a member of the unit could trigger the illusion that the entire unit was presented, and therefore inducing the false memory of the missing partner.

The enhanced hit rate of the presented member in the pair could be due to the possibility that statistical regularities automatically draw attention (Zhao et al., 2013). Given that attention plays an important factor in the recognition task, participants in the structured condition may have prioritized processing of the paired objects, and therefore showed a better hit rate compared to the random condition.

Another account for the enhanced memory is that it may be easier to memorize the objects that were present in the sequence, because statistical learning increases the working memory capacity to encode objects (Brady et al., 2009). The better memory performance of the paired objects in the baseline condition was consistent with previous finding that statistical learning enhances memory of structured objects (Otsuka \& Saiki, 2016).

In conclusion, we discovered a novel consequence of statistical learning: it not only enhances the detection of the object within the regularities, but also creates a false memory of the missing object.

## Acknowledgement

This work was supported by NSERC Discovery Grant (RGPIN-2014-05617 to JZ), the Canada Research Chairs program (to JZ), the Leaders Opportunity Fund from the Canadian Foundation for Innovation (F14-05370 to JZ), and Elizabeth Young Lacey Graduate Scholarship (to YL).

## References

Brady, T. F., Konkle, T., \& Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. Journal of Experimental Psychology: General, 138, 487-502.
Brady, T. F., Konkle, T., Alvarez, G. A. and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for
object details. Proceedings of the National Academy of Sciences, USA, 105, 14325-14329.
Conway, C. M., \& Christiansen, M. H. (2005). Modality constrained statistical learning of tactile, visual, and auditory sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 24-39.
Fiser, J., \& Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychological Science, 12, 499-504.
Kim, G., Lewis-Peacock, J. A., Norman, K. A., \& Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. Proceedings of the National Academy of Sciences, 111, 8997-9002.
Miller, M. B., \& Gazzaniga, M. S. (1998). Creating false memories for visual scenes. Neuropsychologia, 36, 513-520.
Otsuka, S., \& Saiki, J. (2016). Gift from statistical learning: Visual statistical learning enhances memory for sequence elements and impairs memory for items that disrupt regularities. Cognition, 147, 113-126.
Roediger III, H. L., Balota, D. A., \& Watson, J. M. (2001). Spreading activation and arousal of false memories. The nature of remembering: Essays in honor of Robert G. Crowder, 95-115.
Roediger III, H. L., \& McDermott, K. (1995). Creating false memories - remembering words not presented in lists. Journal of Experimental Psychology-Learning Memory and Cognition, 21, 803-814.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926-1928.
Turk-Browne, N. B., Isola, P. J., Scholl, B. J., \& Treat, T. A. (2008). Multidimensional visual statistical learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 399-407.
Turk-Browne, N. B., Jungé, J. A., \& Scholl, B. J. (2005). The automaticity of visual statistical learning. Journal of Experimental Psychology: General, 134, 552-564.
Turk-Browne, N. B., Scholl, B. J., Chun, M. M., \& Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. Journal of Cognitive Neuroscience, 21, 1934-1945.
Yu, R., \& Zhao, J. (2015). The persistence of attentional bias to regularities in a changing environment. Attention, Perception, \& Psychophysics, 77, 2217-2228.
Zhao, J., Al-Aidroos, N., \& Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. Psychological Science, 24, 667-677.
Zhao, J., \& Luo, Y. (2017). Statistical regularities guide the spatial scale of attention. Attention, Perception, \& Psychophysics, 79, 24-30.
Zhao, J., \& Yu, R. (2016). Statistical regularities reduce perceived numerosity. Cognition, 146, 217-222.

# An information-seeking account of eye movements during spoken and signed language comprehension 

Kyle MacDonald ${ }^{1}$ (kylem4@stanford.edu), Aviva Blonder ${ }^{2}$ (aviva.blonder@oberlin.edu), Virginia Marchman ${ }^{1}$ (marchman@stanford.edu), Anne Fernald ${ }^{1}$ (afernald@stanford.edu),<br>Michael C. Frank ${ }^{1}$ (mcfrank@stanford.edu)<br>${ }^{1}$ Department of Psychology Stanford University, ${ }^{2}$ Department of Psychology Oberlin College


#### Abstract

Language comprehension in grounded contexts involves integrating visual and linguistic information through decisions about visual fixation. But when the visual signal also contains information about the language source - as in the case of written text or sign language - how do we decide where to look? Here, we hypothesize that eye movements during language comprehension represent an adaptive response. Using two case studies, we show that, compared to English-learners, young signers delayed their gaze shifts away from a language source, were more accurate with these shifts, and produced a smaller proportion of nonlanguage-driven shifts (E1). Next, we present a well-controlled, confirmatory experiment, showing that English-speaking adults produced fewer nonlanguagedriven shifts when processing printed text compared to spoken language (E2). Together, these data suggest that people adapt to the value of seeking different information in order to increase the chance of rapid and accurate language understanding.


Keywords: eye movements; language processing; information-seeking; American Sign Language

## Introduction

The study of eye movements during language comprehension has provided fundamental insights into the interaction between conceptual representations of the world and the incoming linguistic signal. For example, research shows that adults and children will rapidly shift visual attention upon hearing the name of an object in the visual scene, with a high proportion of shifts occurring prior to the offset of the word (Allopenna, Magnuson, \& Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, \& Sedivy, 1995). Moreover, researchers have found that conceptual representations activated by fixations to the visual world can modulate subsequent eye movements during language processing (Altmann \& Kamide, 2007).

The majority of this work has used eye movements as a measure of the output of the underlying language comprehension process, often using linguistic stimuli that come from a disembodied voice. But in real world contexts, people also gather information about the linguistic signal by fixating on the language source. Consider a speaker who asks you to "Pass the salt" but you are in a noisy room, making it difficult to understand the request. Here, comprehension can be facilitated by gathering information via (a) fixations to the nonlinguistic visual world (i.e., encoding the objects that are present in the scene) or (b) fixations to the speaker (i.e., reading lips or perhaps the direction of gaze).

But, this situation creates a tradeoff where the listener must decide what kind of information to gather and at what time.

How do we decide where to look? We propose that people modulate their eye movements during language comprehension in response to tradeoffs in the value of gathering different kinds of information. We test this adaptive tradeoff account using two case studies that manipulate the value of different fixation locations for language understanding: a) a comparison of processing sign vs. spoken language in children (E1), and $b$ ) a comparison of processing printed text vs. spoken language in adults (E2). Our key prediction is that competition for visual attention will make gaze shifts away from the language source less valuable than fixating the source of the linguistic signal, leading people to generate fewer exploratory, nonlanguage-driven eye movements.

## Experiment 1

E1 provides an initial test of our adaptive tradeoffs account. We compared eye movements of children learning ASL to children learning a spoken language using parallel real-time language comprehension tasks where children processed familiar sentences (e.g., "Where's the ball?") while looking at a simplified visual world with 3 fixation targets (a center stimulus that varied by condition, a target picture, and a distracter picture; see Fig 1). The spoken language data are a reanalysis of three unpublished data sets, and the ASL data are reported in MacDonald et al. (under review). We predicted that, compared to spoken language processing, processing ASL would increase the value of fixating on the language source and decrease the value of generating exploratory, nonlanguagedriven shifts even after the target linguistic item began unfolding in time.

To test this prediction, we present traditional behavioral analyses of first shift Accuracy and RT. We also present two model-based analyses. First, we use an exponentially weighted moving average (EWMA) method (Vandekerckhove \& Tuerlinckx, 2007) to categorize participants’ gaze shifts as language-driven or random. In contrast to the standard RT/Accuracy analysis, the EMWA allows us to quantify differences in the accuracy of gaze shifts as a function of when that shift occurred in time. Next, we use drift-diffusion models (DDMs) (Ratcliff \& Childers, 2015) to quantify differences in the underlying psychological variables that might drive behavioral differences in Accuracy and RT. For example, the DDM uses the shape of both the correct and incorrect RT distributions to provide a quantiative estimate of whether higher accuracy is driven by more cautious responding or by more efficient information processing.


Figure 1: Stimuli for E1 and E2. Panel A shows the layout of the fixation locations for all tasks: the center stimulus, the target, and the distracter. Panel B shows the five center stimulus items: a static geometric shape (Bullseye), a static image of a familiar object (Object), a person speaking (Face), a person signing (ASL), and printed text (Text).

## Method

Participants Table 1 contains details about the age distributions of children in all of four samples.

Spoken English samples. Participants were 80 native, monolingual English-learning children divided across three samples. Participants had no reported history of developmental or language delay.

ASL sample. Participants were 30 native, monolingual ASL-learning children (18 deaf, 12 hearing). All children, regardless of hearing status, were exposed to ASL from birth through extensive interaction with at least one caregiver fluent in ASL and were reported to experience at least $80 \%$ ASL in their daily lives. The ASL sample included a wider age range compared to the spoken English samples because this is a rare population.
Stimuli ASL linguistic stimuli. We recorded two sets of ASL stimuli, using two valid ASL sentence structures for questions: 1) Sentence-initial wh-phrase: "HEY! WHERE [target noun]?" and 2) Sentence-final wh-phrase: "HEY! [target noun] WHERE?" Two female native ASL users recorded several tokens of each sentence in a child-directed register. Before each sentence, the signer produced a common attention-getting gesture. Mean sign length was 1.25 sec , ranging from 0.69 sec to 1.98 sec .

| Task | Mean_Age | Min_Age | Max_Age | n |
| :--- | ---: | ---: | ---: | ---: |
| ASL | 27.90 | 16 | 53 | 30 |
| Face | 26.00 | 25 | 26 | 24 |
| Object | 31.90 | 26 | 39 | 40 |
| Bullseye | 26.10 | 26 | 27 | 16 |

Table 1: Age distributions of children in Experiment 1. All ages are reported in months.

English linguistic stimuli. All three tasks (Object, Bullseye, and Face) featured the same female speaker who used natural child-directed speech and said: "Look! Where's the (target word)?" The target words were: ball, banana, book, cookie, juice, and shoe. For the Face task, a female native English speaker was video-recorded as she looked straight ahead and said, "Look! Where's the (target word)?" Mean word length was 0.79 sec , ranging from 0.6 sec to 0.94 sec .

ASL and English visual stimuli. The image set consisted of colorful digitized pictures of objects presented in fixed pairs with no phonological overlap (ASL task: cat-bird, car-book, bear-doll, ball-shoe; English tasks: bookshoe, juice-banana, cookie-ball). Side of target picture was counterbalanced across trials.
Design and procedure Children sat on their caregiver's lap and viewed the task on a screen while their gaze was recorded using a digital camcorder. On each trial, children saw two images of familiar objects on the screen for two seconds before the center stimulus appeared (see Fig 1). Then they processed the target sentence - which consisted of a carrier phrase, a target noun, and a question - followed by two seconds without language to allow for a response. Participants saw 32 test trials with several filler trials interspersed to maintain interest.

Coding. Participants' gaze patterns were coded ( $33-\mathrm{ms}$ resolution) as being fixated on either the center stimulus, one of the images, shifting between pictures, or away. To assess inter-coder reliability, $25 \%$ of the videos were re-coded. Agreement was scored at the level of individual frames of video and averaged $98 \%$ on these reliability assessments.

## Results and Discussion

Analysis plan First, we present behavioral analyses of First shift accuracy and Reaction Time (RT). RT corresponds to the latency to shift away from the central stimulus to either picture measured from target-noun onset. Accuracy was the mean proportion of first gaze shifts that landed on the target picture out of the total number of shifts. We log transformed all RTs and used the lme4 R package (Bates, Maechler, Bolker, \& Walker, 2013) to fit mixed-effects regression models that included a random intercept for each participant and item. Since children's age varied across conditions, we included age in months as a covariate in all models. All analysis code can be found in the online repository for this project: https://github.com/kemacdonald/speed-acc.

Next, we present two exploratory model-based analyses to quantify differences in eye movements across the four samples. First, we use an EWMA method to model changes in accuracy as a function of increases in RT. For each RT, the model generates two values: a "control statistic" (CS, which captures the running average accuracy of first shifts) and an "upper control limit" (UCL, which captures the pre-defined limit of when accuracy would be categorized as above chance level). Here, the CS is an expectation of random shifting to either the target or the distracter image (nonlanguage-driven shifts), or a Bernoulli process with probability of success 0.5 .


Figure 2: First shift accuracy and RTs from E1. Panel A shows a boxplot representing the distribution of RTs for correct (orange) and incorrect (blue) shifts for each center stimulus type. Panel B shows the distribution of mean first shift accuracy scores for each center stimulus type. The solid lines represent median values, the boundaries of the box show the upper and lower quartiles, and the whiskers show the full range of the data excluding outliers.

As the RTs get longer, we assume that participants have gathered more information and should become more accurate, or a Bernoulli process with probability success $>0.5$. Using this model, we can quantify and compare: a) the cutoff point when the CS exceeds the UCL, indicating that participants started to generate language-driven shifts and $b$ ) the proportion of shifts that the model categorizes as language-driven vs. nonlanguage-driven.

Finally, we took the shifts that were categorized as language-driven by the EWMA and fit a hierarchical Bayesian drift-diffusion model (HDDM) to quantify differences in the speed and accuracy of language-driven eye movements. We chose to implement a hierarchical Bayesian version of the DDM using the HDDM Python package (Wiecki, Sofer, \& Frank, 2013) since we had relatively few trials from child participants and recent simulation studies have shown that the HDDM approach was better than other DDM fitting methods for small data sets (Ratcliff \& Childers, 2015). The model assumes that people accumulate noisy evidence in favor of one alternative with a response generated when the evidence crosses a pre-defined decision threshold. Here we focus on two parameters of interest that map onto meaningful psychological variables: boundary separation, which indexes the amount of evidence gathered before a response (higher values suggest more cautious responding) and drift rate, which indexes the amount of evidence accumulated per unit time (higher values suggest more efficient processing).

Behavioral analyses $R T$. Visual inspection of the Fig 2, panel A suggests that there was a speed accuracy tradeoff in the ASL, Face, and Bullseye conditions, with incorrect shifts tending to be faster than correct shifts. To quantify differences across the groups, we fit a linear mixedeffects regression predicting first shift RT as a function of center stimulus type, controlling for age, and including user-defined contrasts to test specific comparisons of interest: Log (RT) ~ center stimulus type + age + (1 | subject) + (1 | item). We found that (a) ASL learners generated slower RTs compared to all of the spoken English samples ( $\beta=-0.97, p<.001$ ), (b) ASL learners' shifts were slower compared directly to participants in the Face task ( $\beta=$ $-0.42, p<.001$ ), and (c) participants in the Face task shifted slower compared to participants in the Object and Bullseye tasks ( $\beta=-0.73, p<.001$ ).

Accuracy. Next we compared the accuracy of first shifts across the different tasks by fitting a mixed-effects logistic regression with the same specifications and contrasts as the RT model. We found that (a) ASL learners were more accurate compared to all of the spoken English samples ( $\beta=$ $-0.78, p<.001$ ), (b) ASL learners were more accurate when directly compared to participants in the Face task ( $\beta=-0.62$, $p=0.001$ ), and (c) participants in the Face task were numerically more accurate compared to participants in the Object and Bullseye tasks $(\beta=-0.73)$ but this effect was not significant ( $p=0.089$ ).
Model-based analyses $E W M A$. Figure 3 shows changes in the control statistic (CS) and the upper control limit (UCL) as a function of participants' RTs. Each CS starts at chance performance and below the UCL. In the ASL and Face tasks, the CS value begins to increase with RTs around 0.7 seconds after noun onset and eventually crosses the UCL, indicating that responses $>0.7 \mathrm{sec}$ were on average above chance levels. In contrast, the CS in the Object and Bullseye tasks never crossed the UCL, indicating that children's shifts were equally likely to land on the target or the distracter, regardless of when they were initiated. This result suggests that first shifts in the Bullseye/Object tasks were not language-driven and may instead have reflected a different process such as gathering more information about the referents in the visual world.

Next, we compared the EWMA output for participants in the ASL and Face tasks. We found that ASL learners generated fewer shifts when the CS was below the UCL ( $\beta=-1.61$, $p<.001$ ), indicating that a larger proportion of their initial shifts away were language-driven (see the differences in the red shaded area in Fig 3). We did not find evidence for a difference in the timing of when the CS crossed the UCL ( $\beta=$ $-0.04, p=0.387$ ), indicating that both groups began to generate language-driven shifts about the same time after noun onset.

HDDM. Using the output of the EWMA, we compared the timing and accuracy of language-driven shifts for participants
in the ASL and Face tasks 1 We found that ASL learners had a higher estimate for the boundary separation parameter compared to the Face participants (ASL boundary $=1.77, \mathrm{HDI}=$ $[1.64,1.9]$; Face boundary $=1.35, \mathrm{HDI}=[1.21,1.49])$, with no overlap in the credible values (see Fig 4). This suggests that ASL learners accumulated more evidence about the linguistic signal before generating an eye movement. We found high overlap for estimates of the drift rate parameter, indicating that both groups processed the linguistic information with similar efficiency (ASL drift $=0.64, \mathrm{HDI}=[0.44,0.83]$; Face drift $=0.57, \mathrm{HDI}=[0.33,0.83])$.

Taken together, the behavioral analyses and the EWMA/HDDM results provide converging support that ASL learners were sensitive to the value of eye movements, producing fewer nonlanguage-driven shifts and prioritizing accuracy over speed, but accumulating information at roughly the same rate. This behavior seems reasonable since the potential for missing subsequent linguistic information is high if ASL users shifted prior to gathering sufficient information. It is important to point out that these were exploratory findings and that there were several, potentially important differences between the stimuli, apparatus, and populations. In E2, we set out to perform a well-controlled, confirmatory test of our adaptive tradeoffs account.

## Experiment 2

In E2, we attempt to replicate a key finding from E1: that increasing the competition between fixating the language source and the nonlinguistic visual world reduces nonlanguage-driven eye movements. Moreover, we conducted a confirmatory test of our hypothesis that also controlled for the population differences present in E1. We tested a sample of English-speaking adults using a withinparticipants manipulation of the center stimulus type. We used the Face and Bullseye stimulus sets from E1 and added two new conditions: Text, where the verbal language information was accompanied by a word-by-word display of printed text (see Fig 1), and Text-no-audio, where the spoken language stimulus was removed. We chose text processing since, like sign language comprehension, the linguistic information is gathered via fixations to the visual world.

Our key behavioral prediction is that participants in the Text conditions should produce a higher proportion of language-driven shifts as indexed by the EWMA model output. We did not have strong predictions for the DDM parameter fits since the goal of the Text manipulation was to modulate participants' strategic allocation of visual attention and not the accuracy/efficiency of information processing.

## Method

[^132]

Figure 3: Output for the EWMA guessing model in E1. The black curve represents the evolution of the control statistic (CS) as a function of reaction time. The grey curve represents the upper control limit (UCL). The vertical dashed line is the median cutoff value (point when the control process shifts out of a guessing state). The grey shaded area represents the $95 \%$ confidence interval around the estimate of the median cutoff point. And the shaded areas represents the proprotion of responses that were flagged as guesses (red) and languagedriven (green).

Participants 25 Stanford undergraduates participated (5 male, 20 females) for course credit. All participants were monolingual, native English speakers and had normal vision.

Stimuli Audio and visual stimuli were identical to the Face and Bullseye tasks in E1. We included a new center fixation stimulus type: printed text. The text was displayed in a white font on a black background and was programmed such that only a single word appeared on the screen, with each word appearing for the same duration as the corresponding word in the spoken language stimuli.
Design and procedure The design was nearly identical to E1, with the exception of a change to a within-subjects manipulation where each participant completed all four tasks (Bullseye, Face, Text, and Text-no-audio). In the Text condition, spoken language accompanied the printed text. In the Text-no-audio condition, the spoken language stimulus was removed. Participants saw a total of 128 trials while their eye movements were tracked using automated eye-tracking software.

## Results and Discussion

Behavioral analyses $R T$. Visual inspection of Figure 5, panel A suggests that there was a speed-accuracy tradeoff for all conditions: incorrect gaze shifts tended to be faster than


Figure 4: Posterior distributions for the boundary and drift rate parameters for children in E1 (Panel A) and adults in E2 (Panel B).
correct shifts. We fit a linear mixed-effects regression with the same specification as in E1, but we added by-subject intercepts and slopes for each center stimulus type to account for our within-subjects manipulation. We did not find evidence that RTs were different across conditions (all $p>.05$ ).

Accuracy. Next, we modeled accuracy using a mixedeffects logistic regression with the same specifications (see Panel B of Fig 5). We found that adults' first shifts were highly accurate, and, in contrast to the children in E1, their responses were above chance level even in the Bullseye condition when the center stimulus was not salient or informative. We also found that participants tended to be less accurate in the Text conditions compared to conditions without text ( $\beta$ $=1.18, p=0.002$ ). We did find not any other statistically significant differences.
Model-based analyses $E W M A$. For all four conditions, the CS crossed the UCL (see Fig 6), suggesting that for all tasks some proportion of adults' shifts were language-driven. Interestingly, we found a graded effect of condition (see the shift in the vertical dashed lines in Fig 5) on the point when the CS crossed the UCL such that the Text-no-audio condition occurred earliest $\left(M_{\text {text-no-audio }}=0.39\right)$, followed by the Text and Face conditions that were not different from one another ( $M_{\text {text }}=0.44, M_{\text {face }}=0.45, p>.05$ ), and finally the Bullseye condition $\left(M_{\text {bullseye }}=0.54\right)$. We also found the same graded difference in the proportion of shifts that occurred while the CS was below the UCL (see the red vs. green shaded area in Fig 5), indicating a higher proportion of first shifts were language-driven in the Text conditions, with the highest proportion in the Text-no-audio condition when tested against the three other conditions ( $M_{\text {text-no-audio }}=3.88, \beta=1.74, p$ $<.001$ ). These results provide strong evidence for our key


Figure 5: Behavioral results from E2. All plotting conventions are the same as in Figure 2.
prediction: that increasing the value of fixating the language source reduces exploratory gaze shifts to the nonlinguistic visual world.
$H D D M$. Using the output of the EWMA, we fit the same HDDM as in E1. There was high overlap of the posterior distributions for the drift rate parameters (see Fig 4, panel B), suggesting that participants gathered the linguistic information with similar efficiency. We also found high overlap in the distribution of credible boundary separation estimates for the Bullseye, Text, and Text-no-audio conditions. Interestingly, we found some evidence for a higher boundary separation in the Face condition compared to the other three center stimulus types (Face boundary $=1.72$, $\mathrm{HDI}=[1.47,1.97]$; Bullseye boundary $=1.42, \mathrm{HDI}=[1.21,1.65]$; Text boundary $=1.38, \mathrm{HDI}=[1.16,1.6]$; Text-no-audio boundary $=1.36$, $\mathrm{HDI}=[1.15,1.58])$, suggesting that adults higher accuracy in this condition was driven by accumulating more information before generating a response.

Together, these results suggest that adults were sensitive to the tradeoff between gathering different kinds of information. When processing text, people generated fewer nonlanguagedriven shifts (EWMA results) but their processing efficiency of the linguistic signal itself did not change (HDDM results). Interestingly, we found a graded difference in the EWMA results between the Text and Text-no-audio conditions, with the lowest proportion of early, nonlanguage-driven shifts occurring while processing text without the verbal stimuli. This behavior makes sense; if the adults could rely on the auditory channel to gather the linguistic information, then the value of fixating the text display decreases. In contrast to the children in E1, adults were highly accurate in the Bullseye condition, perhaps because they construed the Bullseye as a center fixation that they should fixate, or perhaps they had better encoded the location/identity of the two referents prior to the start of the target sentence.


Figure 6: EWMA model output for E2. All plotting conventions are the same as Figure 3.

## General Discussion

Language comprehension can be facilitated by fixating on relevant features of the nonlinguistic visual world or on the speaker. But how do we decide where to look? We propose that eye movements during language processing reflect a sensitivity to the tradeoffs of gathering different kinds of information. We found that young ASL-learners generated slower but more accurate shifts away from a language source and produced a smaller proportion of nonlanguage-driven shifts compared to spoken language learners. We found the same pattern of behavior within a sample of English-speaking adults processing displays of printed text compared to spoken language. These results suggest that as the value of fixating on a location to gather information about the linguistic signal increases, eye movements to the rest of the visual world become less useful and occur less often.

Our work here attempts to synthesize results from different populations and stimuli in a single framework, but it has several limitations that we hope will pave the way for future work. First, we have not performed a confirmatory test of the DDM findings: both ASL-learners (E1) and adults processing language from a person (E2) prioritize accuracy over speed. So these findings, while interesting, are preliminary. Second, we do not know what might be driving the population differences in E1. It could be that ASL-learners' massive experience dealing with competition for visual attention leads to changes in the deployment of eye movements during language comprehension. Or, it could be that the in-the-moment constraints of processing a visual language cause different fixation behaviors. Finally, we used a very simple visual world, with only three places to look, and very simple linguistic stimuli, especially for the adults in E2. Thus it remains an open question how these results might scale up to more com-
plex language information and visual environments.
This work attempts to integrate top-down, goal-based models of vision (Hayhoe \& Ballard, 2005) with work on language-driven eye movements (Allopenna et al., 1998). While we chose to start with two case studies - ASL and text processing - we think the account is more general and that there are many real world situations where people must negotiate the tradeoff between gathering more information about language or about the world: e.g., processing spoken language in noisy environments or at a distance; or early in language learning when children are acquiring new words and often rely on nonlinguistic cues to reference such as pointing or eye gaze. Overall, we hope this work contributes to a broader account of eye movements during language comprehension that can explain fixation behaviors across a wider variety of populations, processing contexts, and during different stages of language learning.

## Acknowledgements

We are grateful to the families who participated in this research. This work was supported by an NSF GRFP to KM and an NIDCD grant to Anne Fernald and David Corina (DC012505).

## References

Allopenna, P. D., Magnuson, J. S., \& Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38(4), 419-439.
Altmann, G., \& Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. Journal of Memory and Language, 57(4), 502-518.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2013). Lme4: Linear mixed-effects models using eigen and s4. r package version 1.0-5.
Hayhoe, M., \& Ballard, D. (2005). Eye movements in natural behavior. Trends in Cognitive Sciences, 9(4), 188-194.
MacDonald, K., LaMarr, T., Corina, D. and, Marchman, V., \& Fernald, A. (under review). Real-time lexical comprehension in young children learning american sign language.
Ratcliff, R., \& Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. Decision, 2(4), 237-279.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., \& Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632.
Vandekerckhove, J., \& Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. Psychonomic Bulletin \& Review, 14(6), 1011-1026.
Wiecki, T. V., Sofer, I., \& Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in python. Frontiers in Neuroinformatics, 7, 14.

# Preschoolers appropriately allocate roles based on relative ability in a cooperative interaction 

Rachel W. Magid (rwmagid@mit.edu), Mary DePascale (mdepascale@wesleyan.edu) \& Laura E. Schulz (lschulz@mit.edu)<br>Department of Brain and Cognitive Sciences, MIT<br>Cambridge, MA 02139 USA


#### Abstract

In cooperative activities, all parties have a shared goal but may not have the same set of skills. The current study considers whether preschoolers are sensitive to probable differences in individuals' competence when allocating roles. We found that 3.5 - to 5.5 -year-olds use relative competence, as indexed by the age of their intended partner, to determine who should do the harder and easier of two tasks in a cooperative interaction. A second experiment demonstrated that children allocate roles differently in a competitive context. Young children infer differences in others' ability and can divide labor efficiently to achieve their goals.


Keywords: cooperation; self/other knowledge; planning.

## Introduction

Cooperation is a foundation of human culture, observed in activities as diverse as governing, hunting, fishing, building, and playing (Tomasello, 1999). Young children begin cooperating in problem-solving tasks and social games by their first birthday, and the sophistication of their cooperative interactions increases over the first few years of life (Brownell \& Carriger, 1990; Cooper, 1980; Warneken, Chen, \& Tomasello, 2006). Children cooperate by sharing food and toys (Brownell, Svetlova, \& Nichols, 2009; Hay, 1979), pointing to inform others (Liszkowski, Carpenter, Striano, \& Tomasello, 2006; Liszkowski, Carpenter, \& Tomasello 2008), and assisting in goal-directed actions (Warneken \& Tomasello, 2007). Children also appear to expect cooperation: when adults disengage from cooperative interactions, they protest (Ross \& Lollis, 1987).

Across species, the most sophisticated forms of cooperation involve collaboration: cases in which individuals flexibly adjust their behavior to accomplish a goal - as in when some individuals pursue prey and others block its escape (Boesch \& Boesch, 1989). In laboratory tasks, children as young as 3.5 engage in this kind of collaboration, flexibly dividing labor, reversing roles when necessary, and coordinating on tasks involving different sub-goals (e.g., as when one child lifts a lever and another pulls a handle to achieve a joint goal; Ashley \& Tomasello, 1998; Fletcher, Warneken, \& Tomasello, 2012). Moreover, older preschoolers divide labor appropriately with respect to available
resources: when the participant has both the tools needed to achieve a joint goal while their partner has only one, five-year-olds (though not 3-year-olds) appropriately delegate to their partner the task corresponding to their partner's tool (Warneken, Steinwender, Hamann, \& Tomasello, 2014). Preschoolers also collaborate to achieve goals by considering what action the other partner has already selected (Warneken et al., 2014).

Critically, previous research has focused on cases in which both partners are, in principle, equally capable of performing both roles. However, people differ not just with respect to the availability of external resources, but also with respect to their physical, cognitive, and emotional resources. This is advantageous for living in social groups, given that collaboration among individuals with different skills might lead to more efficient and effective actions, and better problemsolving (e.g., Azmitia, 1988; Dyer \& Singh, 1998). However, to capitalize on diverse skills, role allocation should correspond to individuals' differing capabilities.

Dividing labor in this way requires integrating several pieces of information. Even in simple twoparticipant scenarios, the individual must represent both her own and her partner's ability to perform the different tasks or components of the task, compare the two, and allocate roles so the person relatively more capable of each task performs it. Thus an adept collaborator should take on a relatively easier task when partnered with someone she regards as more capable than she is, and a relatively more difficult task when partnered with someone she believes is less capable than she is. Here we ask whether preschoolers effectively allocate roles in collaborative tasks by considering their own abilities relative to a partner's.

Previous research, in addition to the work reviewed above, provides grounds for believing that children might succeed at this kind of division of labor. Three and four-year-olds acknowledge and comment on the fact that different people have different abilities (Mostache \& Bragonier, 1981), and are sensitive to differences in others' knowledge, competence and reliability (e.g., Jara-Ettinger, Tenenbaum, \& Schulz, 2015; Koenig, Clément, \& Harris, 2004; Sobel \& Kushnir, 2013). Such evaluations influence children's helping behavior; children as young as 3 who master a problem-solving task spontaneously tutor learners they
know are naïve (Johnson, Pynn \& Nisbet, 2002).
Preschoolers' ability to accurately represent their own strengths and weaknesses is somewhat more controversial. Some work suggests that preschoolers are (excessively) optimistic about their abilities (Cimpian, 2010; Smiley \& Dweck, 1994; Schneider, 1998), and thus resilient in the face of negative feedback (Boseovski, 2010; Droege \& Stipek, 1993; Ruble, Parsons, \& Ross, 1976). To the degree that children misjudge their own abilities, they would be relatively incapable of efficient division of labor.

However, other work suggests that children begin to regard themselves as good or bad at tasks even in very early childhood (e.g., Gunderson et al., 2013; Heyman, Dweck, \& Cain, 1992; Smiley \& Dweck, 1994). Moreover, children begin to evaluate their own performance relative to their peers as young as 3 (Butler, 1988; Cimpian, 2010; Magid \& Schulz, 2015; Rhodes \& Brickman, 2008). For the current purposes, note that even if children are relatively poor judges of their abilities in an absolute sense, they might be able to judge whether one task is easier for them than another, and whether they are more or less capable than a peer. If so, children might recognize that they should take the easier task if they believe their partner is more capable than they are, and the harder task if they believe their partner is less capable.

## Experiment 1

In the current study, we test this by introducing children to two carnival style games: a ring toss and ball toss. Each game had an easy and a hard version. (See Figure 1.) Any individual child got the easy version of one game and the hard version of the other (counterbalanced across participants). Children were not told that one game was "easy" and the other was "hard" but they were allowed to try each game four times to get a sense of their own ability to succeed on each task. Children were then told that another child was going to come to play with them. They were told that they should choose one game for their partner, and one game for themselves, and that if they both succeeded - so that a ring went on a pole and a ball went in the box-a special machine would light up.

How might children infer others' capabilities on a novel task? Considerable work suggests that children play differently with peers of different ages (Brody, Stoneman, \& MacKinnon, 1982; Edwards \& Lewis, 1979; French, 1984) thus here we manipulated the age of the (fictitious) peer to see whether children would use this to infer their peers' competence relative to their own and allocate roles accordingly. In one condition, children were told that the partner would be younger (Younger Other condition); in the other condition they were told that their partner would be older than the
participant (Older Other condition).
There are a number of possible results. If children are poor judges of task difficulty, they should choose at chance. If children judge the tasks accurately, but try only to maximize their own chances of success (and ignore the joint, collaborative nature of the task) they should always choose the easy task for themselves and the hard task for their partner. Alternatively, if children tend to overestimate themselves (or underestimate their partners) they should always choose the hard task for themselves and the easy task for their partner. However, if children's role allocation in cooperative tasks is sensitive to relative ability (as indexed by age), they should choose the easier game for their partner if their partner is younger, and the easier game to themselves if their partner is older.


Figure 1: Each participant saw only one setup (top or bottom). Participants practiced each game before allocating roles.

## Method

Participants and Materials. All procedures and the analysis plan for this study were pre-registered on the Open Science Framework (osf.io/aq246). Assuming a large effect size (Cramer's $\mathrm{V}=.50$ ), a power analysis indicated that 44 participants were required to reach a power of .90 . All participants were recruited from an urban children's museum and randomly assigned to one of two conditions: Younger Other or Older Other. Forty-four children (mean age $=54$ months; range 4366 months) were included in the final sample ( $n=22$ per condition). Ten additional children did not pass the inclusion criteria. (See Procedure for details). An additional five children were tested but excluded due to parental interference $(n=3)$ or failing to provide a response to the test question $(n=2)$.

A felt mat ( $132 \times 94 \mathrm{~cm}$ ) was placed on the floor for game play. The mat was marked with three tape Xs and a line ( 16 cm in front of the Xs) to indicate where participants should stand. Participants stood on the left and right Xs to play games and the center X to answer questions. Children played two games: a ring toss and a ball toss. Each game had two versions-one easier (Easy Rings, Easy Balls) and one harder (Hard Rings, Hard Balls). The ring toss used a plastic pole on a black circular plastic base. The easier version used a shorter pole ( 22 cm with a 5 cm red tip) and was closer to the tape line ( 13 cm away); and the harder version used a taller pole ( 40 cm with a 5 cm red tip) and was farther from the tape line ( 65 cm away). The ring toss game was played with blue rings ( 16 cm diameter). Each ball toss game used a gray fabric box placed on top of a blue plastic crate $(24 \times 24 \times 41 \mathrm{~cm})$ and was played with yellow plastic balls ( 26 cm circumference). The easier version used a larger box ( $29 \times 14 \times 10 \mathrm{~cm}$ ) with a cardboard backboard ( $17 \times 28 \mathrm{~cm}$ ) and was placed at the front of the crate, closer to the tape line $(53 \mathrm{~cm})$. The harder version used a smaller box ( $14 \times 14 \times 10 \mathrm{~cm}$ ) elevated on a black box of the same size, and was placed at the back of the crate, farther from the tape line ( 77 cm ). Half the participants played the Easy Rings and Hard Balls; half the Hard Rings and Easy Balls. Laminated cards $(23 \times 6 \mathrm{~cm})$ showed photographs of Older Other or Younger Other children. Children depicted in the photographs were either two-year-olds ( 10 cm tall) or six-year-olds ( 15 cm tall), based on the condition. A laminated card of the same size had the word "YOU" printed in the center and was used to represent the participant. A remote-controlled LED light machine $(12 \times 13 \times 12 \mathrm{~cm})$ was used for the joint task.

Procedure. All children were tested individually in a quiet room at a children's museum. Children were shown two games (either Easy Rings and Hard Balls, or Hard Rings and Easy Balls) and given the chance to practice each game four times. The game played first (rings or balls), the location of each game (right or left), and the version of each game (easy or hard) were counterbalanced across children. After children practiced, the experimenter introduced the light machine and explained that players of the two games could work together to achieve a single joint goal: if the ball went in the box and a ring went on the pole at the same time, then the machine would light up. The experimenter introduced the participant to the fictional Other child, named Jamie, by explaining that she had talked with the other child earlier that day and that $\mathrm{s} / \mathrm{he}$ wanted to come play the games together with the participant. The experimenter then showed children a card with a picture of the Other child and said that they were either a toddler (Younger Other) or a first-grader (Older Other). The experimenter then asked children
their own age, specifying that the Other child was younger or older, by condition. The Other child was matched by gender to the participant. For each category (Younger boy, Older boy, Younger girl, Older girl) two pictures were used to reduce the possibility that ancillary features of any picture might influence children's choices or perceptions of the Other child's abilities. The photographs represented a diversity of races and ethnicities. The experimenter then asked children to allocate roles by choosing which game the Other child should play, placing the Other child's picture next to the game chosen for them and a card with "YOU" written on it next to the game the participant chose for themselves. One game was designed to be easier than the other, however differences in motor skills or experience might lead different children to different conclusions, thus to ensure that the role allocation matched children's judgment of the relative difficulty of the two games, we asked children "Which game was easier?" As a followup, children were asked why they chose the game they picked for the Other child. Finally, we asked children if the Other child was older or younger to ensure that they had understood the task. This last question was used as an inclusion criterion: children who did not answer correctly were not included in the analysis. ${ }^{1}$ Following these questions, the experimenter left the room briefly ( $15-30$ seconds) and returned saying that she couldn't find the Other child. The experimenter then played the games with the child to turn on the light machine.

## Results

In response to, "Which game was easier?" 37 of the 44 children ( $84 \%$ ) responded that the game we had designed to be easier was easier for them. Children's self-reported judgment was used in all analyses (consistent with the pre-registered design).

As predicted, children's role assignments differed by condition $\chi^{2}(1)=7.62, p=.006, V=.462$. In the Younger Other condition, $64 \%$ of children assigned their partner the Easy Game. In the Older Other condition, only $18 \%$ of children assigned their partner the Easy Game. Collapsing across conditions, $73 \%$ of children assigned roles in a way corresponding to the difficulty of fulfilling each role in the joint task, $p=.004$ by binomial

[^133]test. Children allocated roles in a way most likely to their joint success. Given previous work showing that five-year-olds, but not 3 -year-olds allocate roles based on available resources (Warneken et al., 2014), we also asked whether the likelihood of participants allocating roles based on ability increased with age. We found no evidence of an age effect in the present task, $\beta=-$ $.004(.75), p=.995$, suggesting that even children as young as 3.5 years can allocate roles in a cooperative interaction given inferred differences in ability.

Although there was no significant difference in children's ability to allocate roles effectively in each condition, Fisher's Exact Test, $p=31$, it is intriguing that twice as many children (eight) in the Younger Other condition misallocated the hard game to the toddler whereas only four children in the Older Other condition misallocated the hard game to themselves. Both the Easy Game and the Hard Game were fairly difficult for the preschoolers. Children scored a zero out of four practice trials $32 \%$ of the time across both games. As such, the decision of some children in the Younger Other condition to allocate the Easy Game to themselves may make sense: given that a toddler is unlikely to do better, and the joint goal may thus seem out of reach, it is reasonable for children to choose the game at which they themselves are more likely to succeed. Indeed, when partnered with a younger child, the majority of preschoolers opted for a game that, while increasing the probability of achieving the joint goal, decreased the probability of their own success.

## Discussion

Results from Experiment 1 suggest that children appropriately consider their own and their partner's relative abilities in allocating roles in a cooperative interaction. However, the results raise questions about the extent to which preschoolers simply assign harder games to older children and easier games to younger children without regard for context in which they are making this decision. Here children's explanations provide some insight. Recall that we asked children why they chose one of the games for the Other child and the other for themselves. Nine children did not answer, and eleven gave uninformative answers. However, 24 children referred to the difficulty of the activities and/or alluded to relative ability (e.g., "She is older and can get the balls in"; "Because it's easier for him (in context, this meant "than the other game" rather than "easier for him than me"). Anecdotally, children's spontaneous behavior also provided some evidence that children think about role allocation dynamically: one child in the Younger Other condition who had assigned the Easy game to her partner asked to switch roles when she learned that the Experimenter, not a toddler, would be her partner in the interaction.


Figure 2: Proportion of children who chose the Easy Game or Hard Game for their partner by condition in Experiment 1: Joint Goal Context.


Figure 3: Proportion of children who chose the Easy Game or Hard Game for their partner by condition in Experiment 2: Competitive Context.

## Experiment 2

Testing the sophistication of children's role allocation requires seeing if children allocate the roles differently if they are not in a cooperative context. In Experiment 2 , we tested children in a competitive condition: in this context, children should assign their partner the harder game regardless of the other child's ability.

## Method

All participants were recruited from an urban children's museum and randomly assigned to one of two conditions: Younger Other or Older Other. Forty-four children (mean age $=54$ months; range 42-65 months) were included in the final sample ( $n=22$ per condition). Seven additional children did not pass the inclusion criteria. (See Experiment 1 for details). Two additional
children were tested but excluded due to parental interference. Materials were the same as Experiment 1.

Children were introduced to and practiced the two games as in Experiment 1. After children practiced, the experimenter introduced the light machine and explained that the person who gets a ball in the box or a ring on the pole before the other person wins and gets to turn on the machine to establish a competitive context. The introduction of other child, Jamie, and the questions (asking children which game Jamie should play, which game was easier, and how old Jamie is) were identical to Experiment 1.

## Results

In response to, "Which game was easier?" 34 of the 44 children $(77 \%)$ responded that the game we had designed to be easier was easier for them. As predicted, children's role assignments did not differ by condition $\chi^{2}(1)=1.03, p=.31, V=.204$. In the Younger Other condition, $64 \%$ of children assigned their partner the Hard Game, see Figure 3. In the Older Other condition, only $82 \%$ of children assigned their partner the Hard Game. Collapsing across conditions, $73 \%$ of children assigned the harder game to the Other child, $p=.004$ by binomial test. These results suggest that children do not allocate the harder game to the older child and the easier game to the younger child independent of context. Instead, participants took into account the competitive context of the interaction and assigned roles accordingly.

## Conclusions

In the current study we found that young children allocate roles appropriately in 1) a cooperative interaction, deciding that the less competent partner should take on the easier task while the more competent partner takes on the harder task and 2) a competitive interaction, deciding that the partner should take on the harder task. Note that we did not label the tasks as easy or difficult prior to when children allocated roles.

Past work looking at children's ability to use social comparison information has focused on how children compare themselves to others to evaluate their abilities, or to plan future actions (Butler, 1998; Magid \& Schulz, 2015; Rhodes \& Brickman 2008; Ruble, et al., 1994). The current study shows that relative ability appraisals are also involved in planning joint interactions. Although one could imagine that preschoolers would simply choose which games to play based on how much they like playing each game, these results suggest they consider the games as sub-goals in a cooperative task and consider their own and others’ competence in allocating roles. In future work, we plan
to ask how other contexts affect role allocation. Consider for instance that one goal of an interaction might be to allow the other partner to develop her skills. In this case, less competent, and younger individuals might be asked to do harder parts of a task. Note also that the current study asks a single child to allocate roles for herself and one other child, and to plan for a task occurring immediately. How children allocate roles among multiple individuals, in real time requires negotiating myriad other factors that influence successful cooperation. Additionally, we note that age is a coarse proxy of ability: younger individuals can be more skilled than older ones and in some contexts, most likely are. Moreover, individuals of exactly the same age may have special competencies and expertise in particular areas. Studies suggest that children are sensitive to these differences in ability and know who to ask for help for particular kinds of tasks (Koenig \& Jaswal, 2011; Kushnir, Vredenburgh, \& Schneider, 2013; Lutz \& Keil, 2002). Future work might ask whether children also use such knowledge to allocate roles appropriately. However, these results suggest that at least some of the core skills underlying teamwork and collaborative problem-solving are in place in early childhood.

## Acknowledgements

We thank the Boston Children's Museum and the families who participated in this research and members of the Early Childhood Cognition Lab for helpful comments and discussion. This material is based upon work supported by the Center for Brains, Minds, and Machines, funded by NSF STC award CCF-1231216 and a NSF Graduate Research Fellowship to R.M.

## References

Ashley, J., \& Tomasello, M. (1998). Cooperative problem-solving and teaching in preschoolers. Social Development, 7, 143-163.
Azmitia, M. (1988). Peer interaction and problem solving: When are two heads better than one? Child Development, 87-96.
Boesch, C., \& Boesch, H. (1989). Hunting behavior of wild chimpanzees in the Tai National Park. American Journal of Physical Anthropology, 78, 547-573.
Boseovski, J. J. (2010). Evidence for "rose-colored glasses": An examination of the positivity bias in young children's personality judgments. Child Development Perspectives, 4, 212-218.
Anastasi, J. S., \& Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. Psychonomic Bulletin \& Review, 12, 10431047.

Brody, G. H., Stoneman, Z., \& MacKinnon, C. E. (1982). Role asymmetries in interactions among school-aged children, their younger siblings, and their friends. Child Development, 53, 1364-1370.
Brownell, C. A., \& Carriger, M. S. (1990). Changes in cooperation and self-other differentiation during the second year. Child Development, 61, 1164-1174.
Brownell, C., Nichols, S. R., \& Svetlova, M. (2013). Converging developments in prosocial behavior and self-other understanding in the second year of life: the second social-cognitive revolution. Navigating the Social World: What Infants, Children, and Other Species Teach Us, 385-390.
Butler, R. (1998). Age trends in the use of social and temporal comparison for self-evaluation: Examination of a novel developmental hypothesis. Child Development, 69, 1054-1073.
Cimpian, A. (2010). The impact of generic language about ability on children's achievement motivation. Developmental psychology, 46, 13331340.

Cooper, C. R. (1980). Development of collaborative problem solving among preschool children. Developmental Psychology, 16, 433-440.
Droege, K. L., \& Stipek, D. J. (1993). Children's use of dispositions to predict classmates' behavior. Developmental Psychology, 29, 646-654.
Dyer, J. H., \& Singh, H. (1998). The relational view: Cooperative strategy and sources of interorganizational competitive advantage. Academy of Management Review, 23, 660-679.
Edwards, C. P., \& Lewis, M. (1979). Young children's concepts of social relations: Social functions and social objects. In M. Lewis et al., (Eds.) The child and its family (pp. 245-266). New York, NY: Springer US.
Fletcher, G. E., Warneken, F., \& Tomasello, M. (2012). Differences in cognitive processes underlying the collaborative activities of children and chimpanzees. Cognitive Development, 27, 136-153.
Frey, K. S., \& Ruble, D. N. (1985). What children say when the teacher is not around: Conflicting goals in social comparison and performance assessment in the classroom. Journal of Personality and Social Psychology, 48, 550.
Gunderson, E. A., Gripshover, S. J., Romero, C., Dweck, C. S., Goldin-Meadow, S., \& Levine, S. C. (2013). Parent praise to 1-to 3-year-olds predicts children's motivational frameworks 5 years later. Child Development, 84, 1526-1541.
French, D. C. (1984). Children's knowledge of the social functions of younger, older, and same-age peers. Child Development, 55, 1429-1433.
Heyman, G. D., Dweck, C. S., \& Cain, K. M. (1992). Young children's vulnerability to self-blame and
helplessness: Relationship to beliefs about goodness. Child Development, 63, 401-415.
Johnson-Pynn, J. S., \& Nisbet, V. S. (2002). Preschoolers effectively tutor novice classmates in a block construction task. Child Study Journal, 32, 241-256.
Koenig, M. A., \& Jaswal, V. K. (2011). Characterizing children's expectations about expertise and incompetence: Halo or pitchfork effects? Child Development, 82, 1634-1647.
Kushnir, T., Vredenburgh, C., \& Schneider, L. A. (2013). "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. Developmental Psychology, 49, 446453.

Liszkowski, U., Carpenter, M., Striano, T., \& Tomasello, M. (2006). 12-and 18-month-olds point to provide information for others. Journal of Cognition and Development, 7, 173-187.
Liszkowski, U., Carpenter, M., \& Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. Cognition, 108, 732-739.
Lutz, D. J., \& Keil, F. C. (2002). Early understanding of the division of cognitive labor. Child Development, 73, 1073-1084.
Magid \& Schulz (2015). Quit while you're ahead: Preschoolers persistence and willingness to accept challenges are affected by social comparison. In Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society.
Rhodes, M., \& Brickman, D. (2008). Preschoolers' responses to social comparisons involving relative failure. Psychological Science, 19, 968-972.
Ruble, D. N., Parsons, J. E., \& Ross, J. (1976). Selfevaluative responses of children in an achievement setting. Child Development, 990-997.
Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. Developmental Science, 1, 291-297.
Smiley, P. A., \& Dweck, C. S. (1994). Individual differences in achievement goals among young children. Child development, 65, 1723-1743.
Tomasello, M. (2009). The cultural origins of human cognition. Cambridge, MA: Harvard University Press.
Tomasello, M. (2010). Origins of human communication. Cambridge, MA: MIT Press.
Warneken, F., Chen, F., \& Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. Child Development, 77, 640-663.
Warneken, F., Steinwender, J., Hamann, K., \& Tomasello, M. (2014). Young children's planning in a collaborative problem-solving task. Cognitive Development, 31, 48-58.

# Intuitive psychophysics: Children's exploratory play quantitatively tracks the discriminability of alternative hypotheses 

Rachel Magid<br>Massachusetts Institute of Technology, Cambridge, Massachusetts, USA<br>Max Siegel<br>Massachusetts Institute of Technology<br>Josh Tenenbaum<br>Massachusetts Institute of Technology<br>Laura Schulz<br>Massachusetts Institute of Technology


#### Abstract

Studies suggest that children's exploratory behavior is sensitive to uncertainty; however, few have approached this with sufficient precision to model quantitatively. Across three experiments, children (mean age $=70$ months) were asked to shake a box to identify which of two sets of marbles, differing in numerosity, were hidden inside. The sets' numerosities varied in their discriminability indices - the degree to which listeners can distinguish the sets based on the acoustic information generated. The time children spent shaking the box varied systematically with the discriminability of the alternative hypotheses they were asked to distinguish, even though they heard only one set for each contrast. This suggests that children represent the uncertainty in their own perceptual discrimination abilities (an ability we refer to as an intuitive psychophysics) and their exploratory behavior is precisely calibrated to their degree of uncertainty about alternative hypotheses that might explain unobserved causes of perceptual data.


# Leveraging Response Consistency within Individuals to Improve Group Accuracy for Rank-Ordering Problems 

Brent Miller (brent.miller@vanderbilt.edu)<br>2301 Vanderbilt Place<br>Vanderbilt University<br>Nashville, TN 37240

Mark Steyvers (mark.steyvers@uci.edu)<br>2201 Social \& Behavioral Sciences Gateway Building<br>University of California<br>Irvine, CA 92697


#### Abstract

Averaging the estimates of a number of individuals has been shown to produce an estimate that is generally more accurate than those of the individuals themselves. Similarly, averaging responses from a single individual can also lead to a more accurate answer. How can we best combine estimates within and between individuals to create an accurate group estimate? We report empirical results from a general knowledge rankordering experiment and demonstrate that individuals that provide more consistent answers across repeated elicitations are also more accurate. We develop a consistency weighting heuristic and show that repeated elicitations within an individual can be used to improve group accuracy. We also develop a Thurstonian cognitive model which assumes a direct link between the process that explains the accuracy of an individual and response consistency and show how the model can infer accurate group answers.


Keywords: Bayesian Modeling; Rank Ordering; Knowledge; Recall; Wisdom of Crowds; Within; Expertise; Uncertainty; Coherence; Consistency.

## Introduction

There has been a lot of interest recently regarding how the judgments of individuals can best be combined to make group estimates that are as accurate as possible. When there is a ground truth - one single, verifiable correct answer the group average is often more accurate than most or all of its constituent individual judgments (Davis-Stober, Budescu, \& Broomell, 2014; Wallsten, Budescu, Erev, \& Diederich, 1997; Yaniv \& Foster, 1997) even if the correct answer is unknowable at the time of questioning (Lee, Steyvers, de Young, \& Miller, 2012). When repeated judgments are averaged within one individual as opposed to across individuals, a similar phenomenon occurs. For example, when a single person produces two estimates for the same underlying quantity, the average of the two estimates is generally less erroneous than the individual estimates (Vul \& Pashler, 2008; Herzog \& Hertwig, 2009; Ariely et al. 2000). A standard explanation for these averaging benefits is that random error associated with probabilistic mental representations and processes partially cancel out in the average. A larger averaging benefit is typically found when averaging judgments across as opposed to within subjects (Rauhaut \& Lorenz, 2011;

Müller-Trede, 2011) presumably because differences in mental representations and associated random error is larger across individuals.

In order to improve the accuracy of the group average, many approaches have been developed to identify and upweight more expert or accurate judgments in the group average, including performance or contributor weighting (Budescu and Chen; Cooke, 1991; Bedford \& Cooke, 2001; Aspinal, 2010), consensus (Shanteau et al. 2002; Wang et al. 2011; Batchelder \& Romney, 1988; Batchelder \& Anders, 2012; Lee, Steyvers, de Young \& Miller, 2012; Lee, Steyvers, \& Miller, 2014) as well as subjective confidence and metacognitive judgments (Koriat, 2012; Prelec, 2004).

We will focus on the role of response agreement within subjects as an indicator for expert judgment. Previous research has shown that expert judgments tend to be more consistent over time (Einhorn, 1972, 1974) and that intrasubject reliability can be used as a proxy for expertise (Shanteau, Weiss, Thomas, \& Pounds, 2002; Weiss \& Shanteau, 2003; Weiss, Brennan, Thomas, Kirlik, \& Miller, 2009). This work has focused on the idea of highly specialized expertise and across-question consistency for tasks such as perception and categorization (Weiss \& Shanteau, 2003; Weiss, Brennan, Thomas, Kirlik, \& Miller, 2009). As opposed to previous research, we focus on tasks where expertise may be question-specific; subjects may have knowledge for some questions, but not for others, making their question level consistency more informative about their expertise than the overall domain consistency.

One challenge for using intra-subject consistency as an indicator for expert judgment is that other factors can contribute to response agreement, including decision strategies and episodic recall (Vul \& Pashler, 2008; Hourihan \& Benjamin, 2010). For example, in Vul and Pashler's experiment, subjects were prompted for a second estimate either in the same experimental session or after a delay of three weeks. The intra-subject averages were most accurate after a delay of three weeks, suggesting that subjects were less likely to simply recall the first answer after a long delay. The requirement of a long temporal delay between repeated questions to avoid episodic recall might not be practical in scenarios where subject judgments need to be aggregated over a short interval.

In this paper, we focus on rank-ordering questions where the task is to rank-order a set items such as Presidents by terms in office or US cities by population size (Miller, Hemmer, Steyvers, Lee, 2009; Lee et al. 2012; Lee, Steyvers, Miller, 2014). In contrast to simple yes/no or percentage estimation question involving single quantities, rank-ordering questions involve the retrieval and coordination of many pieces of information, making it less likely that a subject can explicitly remember a previous response. In the absence of easily available episodic strategies, subjects can be asked for a second response almost immediately after their first, eliminating the need for multiple conditions and removing any question anchoring effects.

Our contribution in this paper is threefold. First, we show that the crowd within an individual effect observed by Vul and Pashler exists for rank-ordering tasks, indicating that there is a degree of statistical independence between repeated elicitations for rank-ordering judgments. Second, we demonstrate that the agreement between the first and second response is related to each subjects' response accuracy. We present a simple consistency weighting heuristic where rank-ordering judgments from individuals that are consistent across repeated questions are given larger weight in the group average. We demonstrate that this consistency weighting heuristic significantly improves group accuracy. Finally, we introduce a new repeatedelicitation variant of a Thurstonian model for rank-ordering that has been explored elsewhere (see Steyvers et al., 2009 \& Lee et al., 2012). We compare the performance of the repeated-elicitation model and the original variant, and demonstrate that accounting for the variance in an individual's responses improves overall group aggregation performance

## Experiment

## Method

The experiment was composed of 8 rank ordering questions, and an additional 3 distracter questions; the distracter questions were included to increase the delay between subject responses. Increased delay between responses has been shown previously to increase response independence and effect size (see Vul \& Pashler, 2008). Subjects were 120 undergraduate students between the ages of 18 and 22 at the University of California, Irvine who were compensated with course credit.

Selection for the non-distracter questions was based on difficulty, as determined by the accuracy of subjects in previous experiments (Steyvers et al., 2009; Miller et al., 2011). Approximately one third of questions were selected for being easier (U.S. Holidays, U.S. Presidents, Book Release Dates), three for being moderately difficult (Country Landmasses, U.S. Cities, European Cities), and one two for being particularly difficult (10 Amendments, World Cities). All were general knowledge questions that subjects were likely to have had exposure to. For the
distracter questions, subjects were asked to rank teams for the NFL and NBA based on what they thought their final season standing would be.

Subjects were given the eight knowledge questions in a random order, and items for each question were initially placed in random positions. Subjects were then given the distracter questions. Subjects were then prompted to give responses for the eight questions again, in the same order they appeared in the first elicitation, but with a new random initial placement of the items for each question.

All questions had a ground truth obtained from Pocket World in Figures and various online sources. An interactive interface was presented via a web browser on computer screens. Subjects were instructed to order the presented items (e.g., "Order these books by their first release date, earliest to most recent"), and responded by dragging the individual items on the screen using the computer mouse and "snapping" them into the desired locations in the ordering, as in previous experiments. Transitions between question blocks were marked by a holding page reminding subjects of the instructions for the tasks. At no point were subjects informed that they would be answering the same questions twice.

## Results

Assessing Accuracy Performance was measured relative to the ground truth using Kendall's tau distance $\tau$. This metric is used to count the number of pair-wise disagreements between the reconstructed and correct ordering (lower is better). The larger the distance, the more dissimilar the two orderings are. Values of $\tau$ range from: $0 \leq \tau \leq \mathrm{N}(\mathrm{N}-1) / 2$, where N is the number of items in the order (ten for all of our questions). A value of zero means the ordering is exactly right, a value of one means that the ordering is correct except for two neighboring items being transposed, and so on up to the maximum possible value of forty-five (indicating that the list is completely reversed). An average score of 22.5 is expected for random performance.
Averaged Responses We first evaluated whether or not averaging the responses within each individual reduced the error relative to the individual responses, indicating statistically independent error of the sort observed in the simple recall tasks of Vul and Pashler (2008). Table 1 shows the median Kendall's tau distance for individual rank-ordering problems for the first and second response as well as the combined first and second response using the Borda aggregation method (see modeling section for Borda details). Subjects' error on the first and second responses were not significantly different, on average, and varied according to question difficulty. The averaged first and second responses of each subject (combined column in Table 1) was less erroneous than the first and second responses $-\mathrm{t}(120)=2.16, \mathrm{p}<.05$ and $\mathrm{t}(120)=2.87, \mathrm{p}<.01$ respectively - replicating the findings of Vul and Pashler (2008) for rank ordering tasks.

Table 1: Subject response error (Kendall's tau) across individual rank-ordering problems.

| Problems | $1^{\text {st }}$ | $2^{\text {nd }}$ | Combined |
| :---: | :---: | :---: | :---: |
| Landmass | 9 | 10 | 8 |
| Holidays | 7 | 8 | 7 |
| Presidents | 7 | 7 | 6 |
| Books | 11 | 11 | 10 |
| Euro Cities | 15 | 16 | 14 |
| US Cities | 16 | 14 | 14 |
| World Cities | 21 | 21 | 20 |
| 10 Amendments | 16 | 15 | 15 |
| AVERAGE | 12.5 | 12.7 | 11.9 |

Response Consistency and Accuracy If subject response consistency is correlated to the precision of an individual's knowledge, then multiple independent responses should be further apart from each other the less knowledgeable a subject is. We quantified (inverse) response consistency as the Kendall's tau distance between subjects' responses. Subjects with a larger distance between their first and second judgment should show a higher tau distance to the ground truth. Figure 1 illustrates this relationship separately for the first and second response. The correlation between each subject's response disagreement, and the error of their first and second responses, is $\rho=.51$ and $\rho=.55$ respectively. This correlation is observed not only across all questions, but also for each individual question. The correlation between response disagreement and accuracy appears to scale linearly with overall subject accuracy for the problem.

## Modeling

While averaging across a given individual's responses yields answers that are more accurate, the improvement is far smaller than averaging two responses across subjects (Miller et al., 2011). Given a large number of subjects, it is unclear whether repeated elicitations would improve group responses aggregation if they are merely treated as extra subjects. Can within-subject response consistency be integrated into a between-subject aggregation model to improve overall accuracy? To test this, we evaluate two models - a heuristic approach based on Borda aggregation method and a Thurstonian cognitive model of subject behavior.

## Borda Aggregation

In order to assess if incorporating within-subject response consistency can improve between-subject estimates for rank ordering tasks, we used a modified version of Borda count aggregation that incorporates subject weighting. Borda aggregation is a representative aggregation heuristic that has been used widely elsewhere (see Miller et al., 2009). In traditional Borda count aggregation, all items are assigned


Figure 1: Correlation between response disagreement and accuracy for the first answer (top panel) and second answer (lower panel).
points based upon their location in a given response: 1 point for being in position 1, 2 points for being in position 2, up to 10 points for a list of 10 items. In a standard Borda aggregation method, the points are added across all rankorderings provided by subjects and the items are ordered according to the sum totals for each item. In our modified Borda aggregation method, we add a weighting factor for each individual subject in order to upweight subjects that are more consistent. Specifically, we calculate the point total $S_{k}$ for each item $k \in\{1, \ldots, K\}$ by:

$$
s_{k}=\sum_{j=1}^{N} r_{j, k} w_{j}
$$

where $r_{j, k}$ is the rank of item $k$ for subject $j \in\{1, \ldots, N\}$, and $w_{j}$ is the weight given to subject j . As in a standard Borda method, the sums of these points for each item are then ranked from smallest to largest to determine the final Borda aggregate rank ordering

For an unweighted aggregate rank-ordering, the subject weights were set to the same value for all participants. We used this as the baseline for comparison. For the aggregate rank-ordering weighted by response consistency, we use the inverse of the tau disagreement between the first and second rank-ordering:

$$
w_{j}=1 /\left(\tau_{j}+1\right)
$$

where we add one to the distance in order to avoid zero division. Therefore, the rank-orderings of participants with larger response consistency have a stronger influence on the aggregate rank ordering.

Figure 2 shows the aggregation results. As we found previously (Miller et al., 2009), unweighted Borda
aggregation outperforms the average subject for all eight questions. Additionally, the weighted Borda model performs as well as or better than the unweighted model for all but two of the questions. The weighted Borda model performed worse for the Presidents question because most subjects performed so well that weighting over-penalized the many subjects with near-correct responses. Similarly, the model performed poorly on the European Cities question because there were so few subjects that performed well. Aggregation for the unweighted Borda model was performed across both trials so as not to give the weighted model the advantage of extra subject responses. This superior performance in reconstructing the ground truth ordering demonstrates that response consistency can be used to improve group accuracy for rank ordering tasks. Next we explore whether a cognitive model of the rank-ordering task can better describe subject behavior and more accurately reconstruct the ground truth.

## Thurstonian Model

Given that subject response consistency is clearly related to accuracy in rank-ordering tasks, what kind of mechanism might be responsible for this observed behavior? We developed a probabilistic model based upon a Thurstonian approach. In a Thurstonian representation, the latent ground truth ordering for a specific problem is represented by coordinates on an interval scale. As Figure 3a illustrates, each item $k$ is represented as a latent coordinate $\mu_{k}$ on an interval dimension. Note that this represents not the actual ground truth but the latent truth as perceived by a group of individuals. The one-dimensional representation of items is appropriate as all problems in our study involve onedimensional relative judgments (e.g. the size of items and the timing of events).

Each individual $i$ is assumed to have access to all of the ground truth latent coordinates $\mu$, but without precise knowledge about their exact locations. This uncertainty is represented with normal distributions that are centered on the shared latent ground truth locations and with a subjectlevel $\sigma_{i}$ that represents the uncertainty of the individual about the item locations. Note that for a given subject, all items have the same standard deviation which is a strong assumption but simplifies the model considerably.

As Figure 3b shows, the subject draws mental samples from these item distributions. Repeated elicitations are modeled simply by repeating the sampling process which leads to a new set of samples. The rank-ordering produced by a subject is then based on the order of the mental samples.

As illustrated in Figure 3c, different subjects can have different uncertainty $\sigma_{i}$, and this influences not only the response accuracy but also the response consistency. For example, the larger uncertainty associated with the subject illustrated in Figure 3c leads to more transposition errors in the mental samples associated with a given response - it becomes more likely that samples of nearby distributions are out of order (relative to the latent ground truth) which


Figure 2: Aggregation performance of unweighted and weighted Borda aggregation across first and second responses, compared to the average subject performance.
lowers accuracy. In addition, the larger uncertainty also leads to increased differences in orderings between different responses. Therefore, the model assumes an inherent connection between response consistency and accuracy they are both driven by a latent parameter $\sigma_{i}$ that represents the (inverse) expertise level of a subject for a particular

## a. Latent Ground Truth



## b. Ranking Items via Distributional Sampling



Figure 3: Illustration of the Thurstonian Model for repeated elicitations. (a) The latent ground truth is represented as a set of coordinates on an interval scale (b) Uncertainty about the latent ground truth is represented by Gaussian noise and responses are created by sampling latent values from each item distribution (c) Example of a subject with larger uncertainty about the ground truth and larger variability in the item samples across the first and second response
question.
This multiple-elicitation model is different from previous Thurstonian models that we have presented, where subjects only give a single response per question (Steyvers et al., 2009; Miller \& Steyvers, 2011). This extended model allows us to examine whether accuracy and response consistency can be described with the same underlying mechanism.

We apply Bayesian estimation techniques to infer the group representation from individual orderings. Figure 4 shows the Thurstonian model for a single question across subjects using graphical model notation (see Koller, Friedman, Getoor, \& Taskar, 2007; Shiffrin, Lee, Kim, \& Wagenmakers, 2008, for statistical and psychological introductions). Each node represents a model variable, and the graph structure is used to indicate the conditional dependencies between these variables. Stochastic and deterministic variables are indicated by single-and doublebordered nodes ( $\mu, \sigma, x$ and $y$ respectively), and observed data are represented by a shaded node (y). The plate represents independent replications of the graph structure, which corresponds to multiple elicitations from each individual $i$ and across individuals for each question $j$.

To explain how these data are generated, the model begins with the underlying ground truth location of the items, given by the vector $\mu$. The latent ground truth $\mu$ is given a flat prior such that all item locations are equally likely a priori. Each individual has an associated uncertainty parameter $\sigma_{j} \sim \operatorname{Gamma}\left(\lambda \sigma_{0}, 1 / \lambda\right)$ where $\lambda$ is a hyperparameter that determines the variability of the expertise levels across individuals. We set $\lambda=3$ in the current model.

To determine the order of items for the $i$ th repetition, the $j$ th individual samples a location $x_{i j k}$ for each item $k$ where $x_{i j k} \sim \operatorname{Normal}\left(\mu_{k}, \sigma_{j}\right)$. The sample $x_{i j k}$ represents the realized mental representation for the individual at that particular time. The ordering for each individual is determined by the


Figure 4: Graphical model of the Thurstonian model for repeated elicitations.


Figure 5: Aggregation performance of weighted Borda, traditional Thurstonian, and repeated Thurstonian models.
ordering of all of their mental samples $y_{i j}=\operatorname{Rank}\left(\boldsymbol{x}_{i j}\right)$.
While the generative model is relatively straightforward, the inference is challenging because the observed data $y_{i j}$ is a deterministic ranking. We utilized MCMC procedures originally developed by Yao and Böckenholt (1999), which allowed us to estimate the posterior distribution over the latent variables $x_{i j k}, \sigma_{j}$, and $\mu$ given the observed orderings $y_{i j}$. We use Gibbs sampling to update the mental samples $x_{i j k}$, and Metropolis-Hastings updates for $\sigma_{j}$ and $\mu$.

Figure 5 shows the accuracy of three aggregation models, and demonstrates that the repeated elicitation Thurstonian model performed best overall. It outperformed the weighted Borda model and also outperformed a Thurstonian model that is given both the first and second response of participants but treats the second responses as coming from a new set of participants. Additionally, the repeated elicitation Thurstonian model matched or exceeded other models' performance for each individual question.

The advantage of the repeated elicitation Thurstonian model over the Thurstonian model where the first and second responses are not linked to the same subject is not due to the fact that it has access to additional response information (it uses the same set of subject responses), but because the model simultaneously infers a subject's uncertainty based upon their disagreement with other subjects and their disagreement with themselves. In this way, we have some confidence in the Thurstonian representation of individual-level uncertainty for subject item recall, both as a generative model and as a means of yielding more accurate group estimates for rank ordering tasks.

## Conclusions

In this paper, we have shown that repeated elicitations for general knowledge rank-ordering tasks exhibit statistically independent error, and the variance of that error is correlated to the accuracy of subject responses for easy and difficult questions. Additionally, we have shown that this response consistency can be used to improve group
aggregate accuracy in reconstructing the ground truth answer for rank ordering knowledge tasks. These findings might also be applicable to tasks that do not have a known ground truth, as we have discussed elsewhere (Lee et al., 2012). Finally, we introduced a cognitive model of rankorder judgement wherein a subject-level uncertainty parameter accounted for both subject response accuracy and response consistency, and found that it was best able to capture subject behavior and reconstruct the original ground truth ordering for each of our questions. This lends credence to the idea of a combined probabilistic mechanism for consistency and accuracy underlying the subject behavior observed in these complex knowledge recall tasks.

## References

Aspinall, W. (2010). A route to more tractable expert advice. Nature, 463(7279), 294-295.
Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. Journal of Experimental Psychology: Applied, 6(2), 130-147.
Batchelder, W. H., \& Anders, R. (2012). Cultural Consensus Theory: Comparing different concepts of cultural truth. Journal of Mathematical Psychology, 56(5), 316-332.
Batchelder, W. H., \& Romney, A. K. (1988). Test theory without an answer key. Psychometrika, 53(1), 71-92.
Bedford, T., \& Cooke, R. (2001). Probabilistic Risk Analysis: Foundations and Methods. Cambridge University Press.
Budescu, D. V., \& Chen, E. (2014). Identifying Expertise to Extract the Wisdom of Crowds. Management Science, 61(2), 267-280.
Cooke, R. M. (1991). Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press.
Davis-Stober, C. P., Budescu, D. V., Dana, J., \& Broomell, S. B. (2014). When is a crowd wise? Decision, 1(2), 79101.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. Organizational Behavior and Human Performance, 7(1), 86-106.
Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. Journal of Applied Psychology, 59(5), 562-571.
Herzog, S. M., \& Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. Psychological Science, 20(2), 231-237.
Hourihan, K. L., \& Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory span individuals benefit more from multiple opportunities for estimation. Journal of Experimental Psychology, Learning, Memory, and Cognition, 36(4), 1068-1074.

Koller, F., \& Friedman, N. (2007). Graphic models in a nutshell. In L. Getoor \& B. Taskar (Eds.), Introduction to statistical relational learning (pp. 13-56). MIT Press.
Koriat, A. (2012). When Are Two Heads Better than One and Why? Science, 336(6079), 360-362.
Lee, M. D., Steyvers, M., de Young, M., \& Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. Topics in Cognitive Science, 4(1), 151-163.
Lee, M. D., Steyvers, M., \& Miller, B. (2014). A Cognitive Model for Aggregating People's Rankings. PLOS ONE, 9(5), e96431.
Miller, B. J., Hemmer, P., Steyvers, M., \& Lee, M. D. (2009, July). The wisdom of crowds in rank ordering tasks. In Proceedings of the 9th international conference of cognitive modeling.
Miller, B. J., \& Steyvers, M. (2011, July). The wisdom of crowds with communication. In Proceedings of the 33rd annual conference of the cognitive science society.
Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. Judgment and Decision Making, 6(4), 283.
Prelec, D. (2004). A Bayesian Truth Serum for Subjective Data. Science, 306(5695), 462-466.
Rauhut, H., \& Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. Journal of Mathematical Psychology, 55(2), 191-197.
Shanteau, J., Weiss, D. J., Thomas, R. P., \& Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. European Journal of Operational Research, 136(2), 253-263.
Shiffrin, R. M., Lee, M. D., Kim, W., \& Wagenmakers, E.J. (2008). A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. Cognitive Science, 32(8), 1248-1284.
Vul, E., \& Pashler, H. (2008, July). Measuring the crowd within: Probabilistic representations within individuals. Psychological Science, 19 (7), 645-647.
Wang G., Kulkarni, S. R., Poor, H. V., \& Osherson, D. N. (2011). Aggregating Large Sets of Probabilistic Forecasts by Weighted Coherent Adjustment. Decision Analysis, 8(2), 128-144.
Wallsten, T. S., Budescu, D. V., Erev, I., \& Diederich, A. (1997). Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making, 10 (3), 243-268.
Weiss, D. J., Brennan, K., Thomas, R., Kirlik, A., \& Miller, S. M. (2009). Criteria for performance evaluation. Judgment and Decision Making, 4(2), 164-174.
Weiss, D. J., \& Shanteau, J. (2003). Empirical assessment of expertise. Human Factors, 45(1), 104-116.
Yaniv, I., \& Foster, D. P. (1997). Precision and accuracy of judgmental estimation. Journal of Behavioral Decision Making, 10 (1), 2132.
Yao, G., \& Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. British Journal of Mathematical and Statistical Psychology, 52(1), 79-92.

# A Rational Constructivist Account of the Characteristic-to-Defining Shift 

Francis Mollica, Shirlene Wade, Steven T. Piantadosi<br>\{fmollica | swade | spiantadosi\} @mail.bcs.rochester.edu<br>Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627 USA


#### Abstract

A widely observed phenomenon in children's word-extensions and generalizations is the characteristic-to-defining shift, whereby young children initially generalize words based on typical properties and gradually transition into generalizing words using abstract, logical information. In this paper, we propose a statistically principled model of conceptual development grounded in the trade-off between simplicity and fit to the data. We run our model based on informant-provided family trees and the real-life characteristic features of people on those trees. We demonstrate that the characteristic-to-defining shift does not necessarily depend on discrete change in representation or processes. Instead, the shift could fall out naturally from statistical inference over conceptual hypotheses. Our model finds that the shift occurs even when abstract logical relations are present from the outset of learning as long as characteristic features are informative but imperfect in their ability to capture the underlying concept to be learned-a property of our elicited features.


Keywords: characteristic-to-defining shift; concept learning; development; computational modeling

## Introduction

Children can often comprehend a word and use a word without having a full grasp of its meaning. Consider the following scenario from Keil and Batterman (1984, pp. 226): "This smelly, mean old man with a gun in his pocket came to your house one day and took your colored television set because your parents didn't want it anymore and told him that he could have it." While adults have a strong sense that the man in the scenario is not a robber, young children are willing to label the man a robber. Across multiple domains, young children have been shown to initially privilege perceptuallyobservable, characteristic information in concept learning. Eventually, children transition to more abstract, conceptually-aligned-upon meanings. This phenomena has been termed the characteristic-to-defining shift (Keil \& Batterman, 1984).

Previous research has suggested that perceptual similarity (e.g., shape) plays a strong role in young children's early word-concept mappings (e.g. Landau, Smith, \& Jones, 1988). As children age, they begin to use deeper, relational properties for concept learning (Imai, Gentner, \& Uchida, 1994; Keil \& Batterman, 1984). For example, Keil and Batterman (1984) probed kindergartners', second graders', and fourth graders' definitions for several words using a scenario task. In some scenarios, characteristic features of a term were presented without the defining features of the term; whereas, other scenarios provided the defining features of the term without the typical characteristic features associated with the term. Younger children extended a word's meaning to more scenarios lacking defining features-but possessing many characteristic features-than older children.

While the characteristic-to-defining shift is commonly observed in concept acquisition, the process by which this occurs is unclear. One possibility is that the characteristic-todefining shift is a stage-like transition that occurs in the representational system (Werner, 1948; Bruner, Olver, Greenfield, et al., 1966). For example, the shift could be explained by a transition from representing concepts wholisticallyi.e., using all the features of objects, to representing concepts analytically-i.e., narrowing in specific relevant features of objects (Kemler, 1983). Neural network models of conceptual classification inherently capitalize on this idea when demonstrating a shift (e.g., Shultz, Thivierge, \& Laurin, 2008). Another possibility is that there is a change in the mechanism by which one learns concepts. For example, concept learning might change from storing exemplars to constructing prototype or rule-based representations. These hypothetical changes in representation or processing might be maturational in nature, such as the development of abstraction (Piaget \& Inhelder, 1969). Alternately, they may be driven by inductive inference mechanisms operating over observed data, a la rational constructivism ( $\mathrm{Xu}, 2007$ ).

From the outset we can narrow down this space of hypotheses. The conceptual to defining shift is most likely a function of data, not maturation (Keil, 1983). One prediction of a maturational-shift is that at a single time-point, children should represent all words using characteristic features or defining features, whereas a data-driven shift predicts that both adults and children should have more exemplar-based representations in unfamiliar domains, and more rule-based representations in familiar domains. The former does not explain children's behavior-children seem to possess characteristic representations and defining representations of different words at a single time point. The prediction of the latterthat individuals have more exemplar-based representations in unfamiliar domains and more rule-based representations in familiar domains, is observed in adults (Chi, Feltovich, \& Glaser, 1981) and in children (Chi, 1985).

All of the aforementioned hypotheses require a discrete shift in representation or process. However, it is unclear whether a representational or mechanistic shift is entirely warranted. To date, no model has tested whether a characteristic-to-defining shift could be a natural by-product of the continuous data-driven construction of concepts. We evaluate this proposal in the task of learning kinship concepts. While "mommy" and "daddy" are some of a child's earliest produced words, children actually spend many years mastering kin relations (e.g. Haviland \& Clark, 1974; Benson \& Anglin, 1987; Keil \& Batterman, 1984). For example, 7- and 8 -year-olds are still unable to provide adequate definitions for
a number of kinship terms (Haviland \& Clark, 1974).

## The Acquisition of Kin Terms

Kinship is an ideal domain for studying the characteristic-todefining shift because it easily lends to logical representations (e.g. Kemp \& Regier, 2012); the domain of kinship is familiar to young children; and the characteristic and defining features behind kinship terms are fairly intuitive and straightforward. Furthermore, several semi-structured interviews attempting to uncover children's knowledge of kinship demonstrate considerable variation in children's definitions. For example, the following is an interview with a six-year-old from Benson and Anglin (1987, p. 48):

I: What is an uncle?
S: A man that's related to ya.
I: Tell me everything you know about an uncle.
S: He knew you when you were a baby...Sometimes they work to build houses... Sometimes they join in for the army.
I: Can you tell me anything else about an uncle?
S: They help you. That's all I know.
I: What kind of a thing is an uncle?
S: A man that's related to you.
Based on children's definitions, researchers have proposed theories weighing the importance of conceptual simplicity (Haviland \& Clark, 1974) and the role of sufficient data (Benson \& Anglin, 1987) in the acquisition of kinship terms. To explain the order of acquisition of kinship terms, Haviland and Clark (1974) proposed a semantic complexity hypothesis. In this account, the simplicity of a concept is defined as the fewest number of base relations (e.g., up one node on the family tree) required to explain a relationship on a kinship tree with a penalty on the variety of base relations used. Children use these base relations to build concepts in a piecewise fashion. By this logic, adult-like kinship concepts are acquired for semantically simpler terms before semantically complex terms. Haviland and Clark (1974)'s original hypothesis is a learning model whereby children first develop perceptual features to construct a concept and only over time develop abstract, relational features. This formalism is entirely consistent with the formalisms used in Mollica and Piantadosi (2015), which we also adopt and describe below. Furthermore, simplicity, in general, is an empirically grounded principle underlying concept construction (Feldman, 2000). More specifically, the role of simplicity and communicative efficiency in kinship terms has been demonstrated across a variety of the world's languages (Kemp \& Regier, 2012).

In addition to simplicity, researchers have proposed that the amount and quality of the observed data drive word learning and conceptual development both in kinship (Benson \& Anglin, 1987; Danziger, 1957) and in other domains (e.g., Weisleder \& Fernald, 2013). For example, Benson and Anglin (1987) found that the order of acquisition of kinship terms was best predicted by children's experience with their relatives. In his rejection of stage theories, Danziger (1957) proposed that conceptual development is primarily driven by opportunities provided by the environment. To account for the influence of data, we incorporate assumptions about plausible data distributions in our model. Further, we trade off
the influence of data with semantic complexity by placing a simplicity weighted prior against a data-driven likelihood.

## Our Approach

We approach this problem at the computational level of analysis (Marr, 1982) to demonstrate how an ideal learner would manifest a characteristic-to-defining shift. We start with the model of Mollica and Piantadosi (2015), which demonstrates how a learner could use cross-situational word-referent occurrences to infer the concept that licenses how a word should be extended. We extend the Probabilistic Context-Free Grammar (PCFG) in their model to construct both characteristic and defining hypotheses for kinship terms. We then collected data about the characteristic and logical relationships from two naive informants' own family trees. This is important because the characteristic and logical relationships of real people allows us to test if natural data will contain perceptual and experiential features informative enough to observe a characteristic to defining shift. We ran the model on the informant-provided trees and a simulated tree to generate possible characteristic and defining hypotheses for four kinship concepts: BROTHER, GRANDMA, MOTHER and UNCLE. These hypotheses were then scored using Mollica and Piantadosi (2015)'s Bayesian model according to their simplicity and ability to explain simulated word-referent data. We analyzed (1) whether an ideal learner is most likely to entertain characteristic or defining hypothesis given an amount of data and (2) the accuracy of the hypotheses in explaining the data as a function of the amount of data observed.

We expect that an ideal learner (without any maturational factors) should demonstrate a characteristic-to-defining shift only if the elicited features (both perceptual and experiential) are informative but imperfect in their ability to capture the underlying concept. If the elicited features accurately capture a concept, an ideal learner should never shift from generating characteristic hypotheses to defining hypotheses. On the contrary, if the elicited features are uninformative, and thus poor at capturing a concept, an ideal learner might predominately generate defining hypotheses, predicting either no shift or an implausibly rapid shift from characteristic to defining hypotheses. Therefore, it is crucial that we collect data about the characteristic and logical relationships of real people to test if natural data will contain features within the range of informativity that will show a characteristic-to-defining shift.

## Data Collection

To simulate data for the learning model, two informants, who were blind to the experiment, drew their family tree, ranked each member in terms of how frequently they interacted with them as a child (e.g., see Figure 1), and provided ten oneword adjectives for each family member. For each informant, the unique adjectives were used to construct a binary feature matrix (adjective by family member; e.g., see Figure 2). Each informant was presented with the feature matrix and asked to indicate if each feature applied to each family member. Informants made a response to every cell of the matrix: zero if


Figure 1: Distance-ranked family tree from Informant One.

| START $\rightarrow$ CHAR | START $\rightarrow$ DEF |
| :--- | :--- |
| CHAR $\rightarrow$ union(CHAR, CHAR) | DEF $\rightarrow$ union(DEF, DEF) |
| CHAR $\rightarrow$ intersection(CHAR, CHAR) | DEF $\rightarrow$ intersection(DEF, DEF) |
| CHAR $\rightarrow$ set_difference(CHAR, CHAR) | DEF $\rightarrow$ set_difference(DEF, DEF) |
| CHAR $\rightarrow$ complement(CHAR) | DEF $\rightarrow$ complement(DEF) |
| CHAR $\rightarrow$ feature(VAL) | DEF $\rightarrow$ up_node(DEF) |
| VAL $\rightarrow\{$ Yes, No $\}$ | DEF $\rightarrow$ down_node(DEF) |
|  | DEF $\rightarrow$ lateral_node(DEF) |
|  | DEF $\rightarrow$ male(DEF) |
|  | DEF $\rightarrow$ female(DEF) |
|  | DEF $\rightarrow X$ (i.e., the speaker) |

Table 1: The Probabilistic Context Free Grammar used to generate kinship concepts
the feature did not apply; one if the feature did apply. The informants provided 109 and 88 unique features respectively including both experiential features (e.g., strict) and perceptually observable features (e.g., blonde). Additionally, we simulated data from the extended family tree used in Mollica and Piantadosi (2015). To sample from the extended tree, we ranked distance using Euclidean distance and constructed a feature matrix for the tree based on 29 perceptually observable features following the principles of Mendelian genetics ${ }^{1}$.

## Extending the model

The model incorporates a PCFG prior with uniform rule probabilities to measure the simplicity of any composition of logical or perceptual features. In the PCFG (see Table 1), we include set theoretical primitives-i.e., union, intersection, set_difference and complement-for both characteristic and defining hypotheses. For defining hypotheses, we include gender primitives-i.e., male and female-and graph theoretical primitives that mimic the abstract primitives proposed by Haviland and Clark (1974): up_node, down_node and lateral_node. The terminal for a defining hypothesis is an argument for the speaker $X$, as we assume that the kinship term should be processed relative to the speaker. For characteristic hypotheses, we include a primitive for each feature, which takes a binary indicator variable and returns the set of family members with or without that feature. Using a PCFG as a prior penalizes complex hypothetical meanings and, thus, builds in a simplicity bias as discussed above. It is important to note that our PCFG generates characteristic hypothesesi.e., only containing characteristic information, and defining hypotheses-i.e., only containing logical information (and gender). We leave the exploration of a hybrid characteristicdefining hypothesis space for future research.

Data for the learning model was noisily sampled from a

[^134]family tree such that $90 \%$ of the time the data reflected accurate use of the true concept and $10 \%$ of the time the data was entirely random. To construct a data point, which took the form of a speaker-referent pair $\{s, r\}$, we first sampled a speaker $s$ from a Zipfian distribution over all members of the family tree ordered by reported distance from the learner. Consequently, data from speakers ranked closer in distance to the learner were more likely to be sampled than data from speakers ranked distant to the learner, which is in line with the intuition that most input a child receives comes from her immediate family. We then sampled a referent $r$ from the Zipfian distribution conditioned on the speaker and word. Given all possible referents the speaker could be correctly referring to when using the word, referents that are closer to the learner are more likely to be talked about than the learner's more distant relations. This reflects the intuition that a child is more likely to hear about her immediate family than distant relatives. Both intuitions are supported by Benson and Anglin (1987)'s survey of children's experience with kinship terms and relations. During learning, we compute the likelihood of the data under the same model used to simulate the data.

Together the prior and the likelihood specify a model for all possible hypotheses constructed from the PCFG:

$$
\begin{equation*}
P\left(h \mid\{s, r\}_{i}^{N}\right) \propto \prod_{i}^{N} P\left(r_{i} \mid s_{i}, h\right) \cdot P(h) \tag{1}
\end{equation*}
$$

With this model we can score the probability of a hypothesis conditioned on simulated data. We then investigate the conditions under which a characteristic-to-defining "shift" will naturally emerge as children learn kinship concepts without positing discrete change.

## Methods

Discovering the most likely hypotheses considered by an ideal learner according to Equation 1 is a complex inference problem because the PCFG specifies an infinite set of possible hypotheses. We solved this problem with Markov-Chain Monte-Carlo (MCMC) methods, which provided us with samples from the posterior distribution by walking around the space of hypotheses. In the limit these walks provably draw samples from the true posterior distribution $P\left(h \mid\{s, r\}_{i}^{N}\right)$. We implement our model using LOTlib (Piantadosi, 2014).

At different amounts of data, we expected an ideal learner to favor different hypotheses. Therefore, we explored the space varying the amount of data between 10 data points and 250 data points by 10 point increments. At each increment of data, we ran eight chains per hypothesis type for one million steps. We stored the top 1000 hypotheses from each chain and combined the hypotheses discovered across chains to form a finite hypothesis space representing the posterior distribution over hypotheses. We normalized all hypotheses by calculating the likelihood over the same set of 1000 data points generated using the same procedure used to generate data for individual chains. We then divided this value by the amount of data (i.e., 1000) to get a measure of each hypothesis' average log likelihood per data point.

## Results

The upper panels of Figure 4 show the posterior probability of entertaining either a characteristic or defining hypothesis ( y -axis) as a function of the amount of data observed ( x -axis). For all of the words, we observe the characteristic-to-defining shift-i.e., the probability of entertaining a characteristic hypothesis is initially greater than the probability of entertaining a defining hypothesis. This means that a simple conceptual learning model shows a characteristic-to-defining shift when given real data about logical relations and characteristic features.

The lower panels of Figure 4 show the posterior weighted $F_{1}$ score conditioned on hypothesis type (characteristic or defining). The $F_{1}$ score is the harmonic mean of precision (i.e., the pressure to extend without over-extending) and recall (i.e., the pressure to extend to all the correct referents). An $F_{1}$ score of 1 reflects perfect performance. Notably in Figure 4, the model successfully learns BROTHER, GRANDMOTHER, and MOTHER-i.e., posterior weighted $F_{1}$ scores all reach 1. With 250 data points, the model does not successfully learn UNCLE yet there still is a shift from characteristic to defining hypotheses on a larger timescale ${ }^{2}$ (Note the $x$-axis in the upper panels).

To help build intuitions about how the model works, Figure 3 presents the three most likely hypotheses an ideal learner trained on Informant One's data would consider for GRANDMA at three time points. Before observing data, an ideal learner should prefer the simplest hypotheses, which often generalize to many referents. In this example the three most likely hypotheses are defining hypotheses that select the speaker $X$, a male speaker and everybody but the speaker. After observing three data points, the hypotheses considered become much more plausible and shift to characteristic features. At this time, the three best hypotheses for grandma are that grandmas are either outgoing, nosy or small. In general the model is shifting from simple hypotheses that generalize broadly to hypotheses that narrow in a bit more, yet still over-extend. Immediately post-shift (i.e., 13 data points), we observe a mixture of characteristic and defining hypotheses. The best hypothesis is the speaker's parents' parents, which misses the female component of GRANDMA. The next best hypothesis is that grandmas are outgoing. The third best hypothesis is actually the definition of a GRANDMA-i.e., the female parents of the speaker's parents. This glimpse at the hypotheses just after the shift illustrates that without a sufficient amount of data, even the correct hypothesis is unlikely because it is more complex in the prior. As we observe more data, the imprecision of the two leading hypotheses decreases

[^135]their posterior probability relative to the correct hypothesis, which will make the correct hypothesis the maximum a posteriori (MAP).

The range of hypotheses are similar between the different trees. Across all trees, characteristic hypotheses have very low posterior weighted $F_{1}$ scores compared to the defining hypotheses. In other words, characteristic hypotheses mislabel referents more than defining hypotheses. Yet, the posterior probability of characteristic hypotheses suggests that characteristic hypotheses are clearly favored at low amounts of data. Given the perspective that the emergence of defining hypotheses is delayed due to the development of abstraction, it is particularly important to note that even in a model with abstraction available from the beginning, we observe a characteristic-to-defining shift. Further, compared to a neural network model where all features are initially considered (Shultz et al., 2008), a characteristic-to-defining shift is still observed in our model where it is initially more likely to only consider only a few features.

Taken together, this pattern of results demonstrates that the characteristic-to-defining shift could naturally fall out of a single statistical inference process with a single representational language ${ }^{3}$. It is not necessary to propose a discrete change in representation or processing. Characteristic hypotheses are favored early because with little data the prior dominates inference-they generalize well to small data amounts and are comparatively less complex in the prior than the best defining hypotheses. Only when there is enough data to warrant additional complexity will defining hypotheses come to dominate inference.

## Discussion

In this paper, we tested whether a characteristic-to-defining shift would emerge naturally in a statistically principled learning model without positing a discrete mechanistic or representational shift. In general, the model successfully learns kinship terms and demonstrates a characteristic-to-defining shift using a single representational language of thought and a single statistical inference mechanism. Therefore, while a discrete shift in mechanism or process is possible, it is not necessary to observe a characteristic-to-defining shift during concept learning.

In our model, kinship concepts are developed through statistical inference over word-referent data and observed kinship structures, which could plausibly be developed from statistical learning of structure (Katz, Goodman, Kersting, Kemp, \& Tenenbaum, 2008; Kemp \& Tenenbaum, 2008). When an ideal learner only observes data about a few referents, there are simple characteristic hypotheses based on perceptual observations that will explain the data; however, as more data is observed, these hypotheses fail to adequately fit the data and warrant a prior-likelihood trade-off, such that more complicated defining hypotheses (which are unlikely

[^136]Informant One
(109 features; 31 family members)

(29 features; 37 family members)


Figure 4: For each tree, the top panel displays the posterior probability of using a characteristic (solid line) or a defining (dashed line) hypothesis as a function of the amount of data observed. The bottom panel displays the posterior weighted $F_{1}$ score conditioned on hypothesis type (characteristic as solid line, defining as dashed line) as a function of data.
in the prior) are substantially more likely in explaining the data and thus come to dominate the posterior. Put simply, the characteristic-to-defining shift can be a by-product of datadriven learning.

There are two interesting implications/predictions of our model. First, our model predicts that the ideal learner will shift from characteristic to defining hypotheses even when she is capable of using abstraction from the outset of learning. This suggests that characteristic hypotheses may be useful, and that the observation that children accept and generate characteristic hypotheses at a young age does not preclude their ability to use abstraction or generate logical/defining hypotheses. Second, our model predicts that if there is a characteristic-to-defining shift, the relevant characteristic features should not capture a concept as well as defining features capture the concept; however, in order for a characteristic-to-defining shift to occur, the characteristic features must be informative to a certain degree. If characteristic features are completely uninformative, defining hypotheses should dominate across all amounts of data.

In our initial stab at the problem, we have made several simplifications. For one, the grammar generated hypotheses to be purely characteristic or purely defining. This simplification is reasonable given how adults would extend a kinship term. For example, if you meet a friend's family for the first time at a neighborhood BBQ, you would presumably extend the term uncle to their parent's male siblings and not the neighbors, who might share several characteristic features with your friend's uncles. This is not to say that competent adult speakers do not maintain characteristic information about kinship terms (e.g., grandmothers are typically nice, old ladies). In the same vein, our characteristic and defining features did not share the same formalism (i.e., feature matrices vs. graph-theoretical functions). A future version of the model should permit characteristic and defining primitives within the same hypothesis and possibly within the same formalism (e.g., a feature matrix containing both characteristic and defining features). This model should also be extended beyond the kinship domain. Lastly, the model is sensitive to the structure of the PCFG in determining the prior. Further research should characterise the robustness of the model to variation in the prior.

## Conclusion

In summary, the widely observed characteristic-to-defining shift falls out naturally from a rational data-driven process. Our simulations show that a data-driven inference mechanism (1) demonstrates a characteristic-to-defining shift in the task of concept learning without positing a change in cognitive representations or processes and (2) succeeds at learning most kinship words from a data distribution based on natural language statistics. We find that an ideal learner will demonstrate a shift even when more accurate abstract/logical representations are possible from the onset of learning provided that characteristic features are informative but imperfect in
their ability to capture the underlying concept. While we address the problem of concept learning within the kinship domain, the model framework can be extended to explain concept learning across multiple domains using different representational formalisms.

## References

Benson, N. J., \& Anglin, J. M. (1987). The child's knowledge of english kin terms. First Language, 7(19), 41-66.
Bruner, J. S., Olver, R. R., Greenfield, P. M., et al. (1966). Studies in cognitive growth. Wiley.
Chi, M. T. (1985). Interactive roles of knowledge and strategies in the development of organized sorting and recall. Thinking and learning skills, 2, 457-483.
Chi, M. T., Feltovich, P. J., \& Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive science, 5(2), 121-152.
Danziger, K. (1957). The child's understanding of kinship terms: A study in the development of relational concepts. The Journal of genetic psychology, 91(2), 213-232.
Feldman, J. (2000). Minimization of boolean complexity in human concept learning. Nature, 407(6804), 630-633.
Haviland, S. E., \& Clark, E. V. (1974). this man's father is my father's son: a study of the acquisition of english kin terms. Journal of Child Language, 1(01), 23-47.
Imai, M., Gentner, D., \& Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. Cognitive Development, 9(1), 45-75.
Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., \& Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. In Proceedings of thirtieth annual meeting of the cognitive science society.
Keil, F. C. (1983). On the emergence of semantic and conceptual distinctions. Journal of Experimental Psychology: General, 112(3), 357.
Keil, F. C., \& Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. Journal of Verbal Learning and Verbal Behavior, 23(2), 221-236.
Kemler, D. G. (1983). Exploring and reexploring issues of integrality, perceptual sensitivity, and dimensional salience. Journal of Experimental Child Psychology, 36(3), 365-379.
Kemp, C., \& Regier, T. (2012). Kinship categories across languages reflect general communicative principles. Science, 336(6084), 1049-1054.
Kemp, C., \& Tenenbaum, J. B. (2008). The discovery of structural form. Proceedings of the National Academy of Sciences, 105(31), 10687-10692.
Landau, B., Smith, L. B., \& Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive development, 3(3), 299-321.
Marr, D. (1982). Vision: A computational approach. Freeman \& Co., San Francisco.
Mollica, F., \& Piantadosi, S. T. (2015). Towards semantically rich and recursive word learning models. In Proceedings of the 37th annual meeting of the cognitive science society (pp. 1607-1612).
Piaget, J., \& Inhelder, B. (1969). The psychology of the child. Basic Books.
Piantadosi, S. T. (2014). LOTlib: Learning and Inference in the Language of Thought. available from https://github.com/piantado/LOTlib.
Shultz, T. R., Thivierge, J.-P., \& Laurin, K. (2008). Acquisition of concepts with characteristic and defining features. Proceedings of the 30th Annual Conference of the Cognitive Science Society, 531-536.
Weisleder, A., \& Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. Psychological Science, 24(11), 2143-2152.
Werner, H. (1948). Comparative psychology of mental development. Follett Pub. Co.
Xu, F. (2007). Rational statistical inference and cognitive development. The innate mind: Foundations and the future, 3, 199-215.

# An incremental information-theoretic buffer supports sentence processing 

Francis Mollica, Steven T. Piantadosi<br>\{mollicaf | spiantado\} @gmail.com<br>Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627 USA


#### Abstract

People have the capability to process text three times faster than they would naturally read it, yet many current theories of sentence processing rely on natural reading times as a proxy for processing difficulty. How can people read material so quickly in spite of information processing limitations suggested by sentence processing theories? One possibility is that surprisal effects on reading time, the hallmark of processing difficulty under sentence processing theories, might arise from perceptual processing, implying no relation between surprisal and sentence processing difficulty. In this paper, we conducted a novel self-paced rapid serial visual presentation (RSVP) experiment, which controlled perceptual processes to probe for sentence processing related surprisal effects. We further tested how readers might compensate for information processing limits during RSVP. We find support for sentence processing related surprisal effects, the pattern of which is consistent with a First-In, First-Out (FIFO) buffer model.


Keywords: language processing; linguistic memory; RSVP

## Introduction

One of the remarkable feats of language processing is the quick pace at which various forms of information are seamlessly integrated (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, \& Sedivy, 1995). Researchers have noticed that language can be processed much faster than how we naturally engage with it (e.g., Potter, 1984). For example, by presenting text faster than the human eye would naturally fixate, people are able to read a story $3-4$ times faster than natural. These observations have led to several apps training people how to read faster-although the efficacy of these apps has been contested (see Rayner, Schotter, Masson, Potter, \& Treiman, 2016). Even though researchers have noted that information processing in reading is faster than people's natural reading pace, psycholinguistics has relied upon the assumption that natural reading times are a reflection of processing difficulty. Using reading time as an index of processing difficulty has provided the data behind several theories in sentence processing (e.g., Grodner \& Gibson, 2005; Levy, 2008). Recently, this assumption has been formalized in terms of information theory: words that carry more information tend to increase reading times relative to words that contain less information, suggesting a fixed processing rate that is measured in bits of information per unit time (e.g., Hale, 2001; Levy, 2008). The amount of information conveyed by a word, colloquially referred to as surprisal, is equal to the negative $\log$ probability of the word given its context.

Given the importance of reading times as a measure of processing difficulty, it is surprising that people can read several times faster than they naturally would and it poses a very important question: Are the effects of surprisal observed in reading times reflective of linguistic information processing
limitations, or do they arise from some alternate perceptual process? We conducted a novel rapid serial visual presentation (RSVP) experiment in which we controlled perceptual processing to test if surprisal effects reflect language processing. Our experiment was also designed to probe how readers might be compensating for linguistic information processing limits when faced with rapidly presented text. Specifically, we consider two hypotheses that would reconcile information processing limits with rapid presentation of text. First, language users might suspend incremental information processing to store information in a buffer, until it can be processed. Given the quick response times in RSVP experiments, this would suggest fast rates of information processing, giving us real reason to check our assumption that reading time is a valid measure of sentence processing difficulty. Second, language users might compensate for RSVP by utilizing an incremental First-In, First-Out (FIFO) memory buffer, where information is immediately copied into a buffer and processed out of the buffer at a fixed rate. Many researchers using the RSVP paradigm have proposed the use of a buffer (e.g., Mitchell, 1984); however, buffer models have never been formalized in terms of information processing. In the remainder of the introduction, we motivate the perceptual and linguistic sources of surprisal effects and our buffer models.

## Perceptual Information Maintenance

When the reader manages visual presentation time, it is unclear if we see surprisal effects because readers maintain the visual input while they are syntactically processing the word or if we see surprisal effects for perceptual processing reasons. For example, both Bayesian models of word identification (Norris, 2006) and rational models of eye movements in reading (Bicknell \& Levy, 2012) posit that readers maintain information until they reach a level of confidence in a word's identity, resulting in log effects of word predictability-i.e., numerically the same value as surprisal. This hypothesis is corroborated by work on individual differences in eye-tracked reading that demonstrate a relationship between word-identification ability and reading speed (Kuperman \& Van Dyke, 2011). In our experiment, we control for a perceptual explanation of surprisal effects by using RSVP, which prohibits readers from influencing presentation time. Importantly, the same processes underlying natural reading are still thought to be at play in RSVP (Potter, 1984). Therefore, if surprisal effects were a result of readers maintaining the visual input until they were confident in the word's identity or had finished syntactically processing the word, we would expect that RSVP would eliminate all surprisal effects.


Figure 1: In 5-RSVP, each word is presented for 147 ms and immediately masked by the next word. The last word in each 5 word chunk is masked with random lowercase letters remaining until the participant presses a button to continue.

## Linguistic Information Processing Limits

One of the foundational questions in sentence processing research is how do people construct the correct syntactic parse for a sentence. As mentioned above, the primary method for teasing apart sentence processing theories is making predictions about processing difficulty for the constituent words of a sentence. Information theory has provided one way to formalize the amount of processing required for each word in a sentence. Hale (2001) first noticed that word reading times are proportional to the total probability of all syntactic structures that are no longer possible after observing a word, which happens to be the word's surprisal. Levy (2008) showed that surprisal can be given an alternative interpretation in terms of parallel-resource allocation. As each new word is processed, the syntactic parses under consideration are re-ranked and resources are allocated accordingly. The processing difficulty of that word is reflected in the amount of re-ranking required as measured by Kullback-Leibler divergence, which simplifies to the word's surprisal. Either way, a word's surprisal is a measure of the amount of processing, and surprisal effects in reading suggest processing difficulty caused by information processing limits. In our experiment, we have controlled for a perceptual origin of surprisal effects so that if we still see surprisal effects, we can be more confident in interpreting them as stemming from language processing.

If surprisal effects reflect information processing limits, the contradiction between readers being capable of reading rapidly presented text but possessing hard information processing limits still remains. Our novel five word self-paced RSVP task presents text in five word RSVP chunks for each button press (see Figure 1), allowing us to consider two hypotheses that would reconcile these observations. First, readers might focus on buffering linguistic information during rapid presentation and delay processing until the perceptual barrage has ended. Under this account, we would expect a small uniform profile of surprisal weights for each of the five words presented together (see middle panel of Figure 2). In RSVP tasks, readers only require a one second break between sentences to perform as well as control groups in recall tasks (Potter, 1984), which suggests an information processing rate roughly 100 bits/second ${ }^{1}$. This rate is too fast to predict sur-

[^137]

Figure 2: Predicted surprisal weights (beta values for surprisal) in the 5-RSVP task as a function of word position for our hypotheses.
prisal effects at naturalistic reading paces, which makes this an important hypothesis to rule out for expectation based sentence processing theories.

Second, readers might incrementally process linguistic information, but perception and language processing might not be as tightly coupled as previously thought. Perception might quickly place information into a buffer, where language processing then removes that information at a slower rate, but still very quickly. This allows perception to move ahead while language processing is ongoing. In this FIFO buffer model, the expected profile of surprisal weights depends on the rate of information processing. Roughly, the model yields three different profiles of surprisal weights (see Figure 3). At slow rates of information processing, multiple words will be seen before the processing of the initial word has ended. As a result, the majority of processing time for all of the words is reflected in reading time post visual presentation, which gives rise to a large uniform profile of surprisal weights. At rates of information processing closer to the average amount of information conveyed in a word, sometimes a word is completely processed before the next word appears and sometimes a word's processing briefly carries over to when the next word appears. As a result, early words contribute less to post-presentation reading time than later words, and the profile of surprisal weights increases with a word's presentation position. At fast rates of information processing, each word is completely processed before the next word appears. Therefore, there should be no influence of surprisal on postpresentation reading time, corresponding to a uniform zero profile of surprisal weights. Our a-priori prediction for the FIFO buffer assumed a rate of processing near the average amount of information (see right panel of Figure 2).

It is important to note that the FIFO buffer model shows the traditional surprisal effect with the rate of information processing parameter. The profile of surprisal weights in the FIFO buffer model do not necessarily reflect processing difficulty as in the typical surprisal effect; rather, the surprisal weights reflect each word's contribution to the information remaining in the buffer post-presentation.

## Experiment

Our plan of attack is to see if surprisal effects are present in a masked word-by-word self-paced reading (mSP) task and a novel 5-word chunk self-paced RSVP task (5-RSVP), which holds perceptual information constant but allows the reader unlimited time to process their input. As a result of the per-

[^138]

Figure 3: Predicted surprisal weights from an incremental FIFO buffer model for each word position (w1-w5) as a function of information processing rate. At slow rates, there is a large uniform surprisal effect. At very fast rates, there is no effect of surprisal. In the middle, surprisal weights increase as a function of their position in a chunk.
ceptual controls, we can interpret an effect of surprisal as reflecting linguistic information processing limits. Further, if there is an effect of surprisal, the profile of surprisal weights for each word in the 5-RSVP condition can distinguish between two hypotheses reconciling readers' ability to read rapidly presented text in spite of incremental processing limits (see Figure 2). Specifically, a uniform profile of surprisal weights would suggest a buffer model in which readers suspend incremental linguistic processing until after presentation ends. Whereas, a profile of surprisal weights that increases as a function of a word's position in the five word chunk would suggest a a FIFO buffer model (formalized below). In addition, participants completed a self-paced reading task (SP) to replicate the standard surprisal effect with our materials. Participants also read an additional story at their natural pace to provide a baseline accuracy and reading rate, which ensured our presentation rate was faster than our participants natural reading rate.

Participants. Sixty-four participants were recruited from the University of Rochester community. Participants were compensated for their participation, which lasted approximately 45 minutes. All participants were screened for normal or corrected-to-normal vision, and English as their native language.

Materials and Design ${ }^{2}$. Participants read four excerpts from articles published in 2008 issues of The New Yorker. To equate the length of the excerpts while preserving the naturalness of the reading material, articles were truncated to end with the paragraph containing the $1000^{\text {th }}$ word of the article. Word frequency and bigram probabilities with Laplace smoothing ( $\alpha=0.1$ ) were estimated using ZS (Smith, n.d.) to access Google N-grams (Michel et al., 2011). For each story, ten yes-no questions were constructed (5-yes, 5-no) to serve as a comprehension check. Four lists were constructed such that each list contained a story in every presentation con-

[^139]dition (Baseline, SP, mSP and 5-RSVP) and across lists each story occurred with every condition. To ensure precise presentation timing, stimuli were presented using a monitor with a 144 Hz refresh rate. Responses and reading times were collected using the keyboard. Stimulus presentation and timing and response collection was controlled by Psychopy (Peirce, 2007).

Procedure. Sixteen participants were randomly assigned to each list. Participants were told that they would be reading and answering questions about four short stories. All participants began the experiment with the baseline reading condition. In the baseline condition, the full body of text for one story was displayed on the screen in 17.5 point Free Sans Bold font. Participants could scroll through the text using the up and down arrow keys. Participants were instructed to press the space bar when they had finished reading the text, which started the comprehension check.

Following the baseline reading task, participants were introduced to the mSP condition with three practice trials. On the first trial, each word was presented in the center of the screen for 245 ms and then immediately masked with a random string of 14 lower-case letters. Participants were instructed to press the space-bar to reveal the next word. For the next two practice trials, the presentation duration of each word was decreased to 196 ms and 175 ms , respectively. Before each trial and story, a fixation cross appeared for 147 ms to orient the participant to the center of the screen. After the practice trials, the order of the remaining conditions was randomized for each participant.

In the SP condition, text was presented similar to the practice trials with one exception: each word was not masked and remained visible to the participant until they pressed the space-bar to advance. In the mSP condition, text was presented similar to the practice trials, except the duration for each word was set at 147 ms . In the 5-RSVP condition, each button press triggered the presentation of the next five words in the text. Each word was flashed on the screen for 147 ms . The first four words were masked by the next word's appearance. The fifth word was masked by a 14 character lowercase letter string.

After each story, the comprehension check for that story began immediately. For each of the ten questions, participants registered their answer by pressing the Y (yes) or N (no) key on the keyboard, which prompted the next question. The order of the questions was randomized for each participant. After the comprehension check, participants were instructed to find their experimenter to advance to the next story. Participants were strongly encouraged to take breaks in between stories and often did. In line with our experience and intuitions, our participants reported that reading in these presentation conditions is taxing ${ }^{3}$.

[^140]A


B


Figure 4: A. Average participant accuracy across presentation conditions. B. Surprisal weights from a linear mixed-effect regression of reading time in the SP, mSP and 5-RSVP conditions (as a function of each word's order in the 5-RSVP condition). Red diamonds represent the FIFO buffer model predictions with a best fit rate of $123 \mathrm{bits} / \mathrm{sec}$. All intervals reflect $95 \%$ bootstrapped confidence intervals.

## Results

To check that our presentation rate of $147 \mathrm{~ms} /$ word was less than participant's natural reading rate, we calculated each participant's average reading time per word in the baseline condition. The mean and median natural reading rate was 300 $\mathrm{ms} /$ word (SEM=11). Three participants had average natural reading rates faster than $147 \mathrm{~ms} /$ word, we retained their data in the remainder of the analysis ${ }^{4}$. Accuracy data was analysed using a mixed effect logistic regression with random intercepts for participants, item, story and list. Across all conditions, participant accuracy was on average greater than chance ${ }^{5}$ (see Figure 4A). The baseline accuracy was $74 \%$ suggesting there was no ceiling effect. In the mSP and 5RSVP conditions, accuracy was significantly lower than baseline; however, the effect size is small ( $\mathrm{mSP} O R=0.70$, 5 RSVP $O R=0.48$ ). To account for spurious button presses and abnormally long reading times, we removed the top and bottom $2.5 \%$ of the reading time data for each participant, resulting in a $5 \%$ data loss overall.

We analysed the reading times for each condition separately. With the SP condition, we aimed to replicate the effect of surprisal reported in the literature when presentation is central and perceptual information management is still in the reader's hands. Reading times were analysed using a mixed effect linear regression with surprisal, log frequency, zero-centered length, and position as fixed effects and random intercepts for participant and story. Replicating the prior literature, there was a significant increase in reading time as surprisal increased ( $\beta=0.54, t=4.60, p<0.05$ ).

With the mSP condition, we aimed to test if there would be an effect of surprisal when perceptual information management is no longer in control of the reader. Reading times were analysed using a mixed effect linear regression with the same effect structure as in the SP condition. Even though perceptual control was removed from the reader, there was a significant increase in reading time as surprisal increased ( $\beta=1.24, t=8.60, p<0.05$ ), suggesting that there are sentence processing related surprisal effects without perceptual

[^141]control.
Having observed a surprisal effect in mSP, we analyzed the 5-RSVP condition to tease apart whether readers compensate for information processing limits during rapid presentation by buffering information and suspending information processing entirely-resulting in a uniform non-zero profile of surprisal weights, or by incrementally buffering and processing linguistic information-resulting in an increasing profile of surprisal weights across subsequent words. To test these hypotheses, we fit two mixed effect linear models: one with a single beta weight for surprisal and one with a unique beta weight for each word's surprisal. Both models contained random intercepts for participant and story and fixed effects for zero-centered length, position and a frequency term for each of the five words presented in a chunk. We find a significant surprisal effect in both models (single beta model: $\beta=0.73, t=3.93, p<0.05$ ), which suggests that the observed surprisal effects are reflective of language processing. As can be seen in the profile of the individual beta weights for surprisal in Figure 4B, the profile is more consistent with a non-uniform profile of weights than the uniform profile $\left(\chi^{2}(4)=11.7, p<0.05\right)$.

## First-In, First-Out (FIFO) Buffer Model

To further explore the idea of an incremental linguistic information buffer, we formalize a FIFO buffer model with one free parameter, i.e., the rate of information processing, and fit the model to our data to provide us an estimate of the rate parameter. Our model operates in five 147 ms windows matching our presentation duration for each of the five words. In each window, the new word's information is added to the buffer, and the amount of information that could be processed in that window according to our rate parameter is removed from the buffer. As a result, if a word could not be completely processed in one window, its processing carries over to the next window. If all the information in the buffer is processed, the buffer is empty when the next word appears. The information left in the buffer after the fifth word's window is the information whose processing should contribute to our measured reading time. We use the rate of information processing to predict the expected reading time for that five word chunk. We repeat this process for every chunk in our stories. To arrive at predictions for the surprisal weights, we analyse
our expected reading times using linear regression with separate surprisal weights for each of the five words. We do not include an intercept value as we expect the entire reading time to be a function of each chunk's surprisal values. Predicted surprisal weights for a variety of processing rate parameters are shown in Figure 3. Our a-priori prediction was that surprisal weights should increase across word positions, as early words will tend to be fully processed so that later words have a larger effect on post-presentation reading time. As can be seen in Figure 3, our a-priori prediction holds true for a large range of plausible rate parameters.

We fit our buffer model to the observed pattern of surprisal weights using gradient descent to minimize squared error. We find that the best fit rate of information processing is 121 bits/second. The best model fit profile of surprisal weights is illustrated in Figure 4B as the red diamonds.

## Discussion

To address the apparent contradiction between the speed at which readers can comprehend text and the information processing limits proposed by sentence processing theories, we suggested that surprisal effects observed in reading tasks might actually be a result of perceptual processing (e.g., Norris, 2006) rather than linguistic information processing. If this were the case, we would expect to see no surprisal effects in the mSP and 5-RSVP conditions, where perceptual control was stripped from the reader. Instead, we found effects of surprisal, suggesting that surprisal effects in RSVP reading have a language processing origin.

Having established the existence of an information processing bottleneck in RSVP, we proposed two possible alternatives ${ }^{6}$ that reconcile information processing limits and rapid text comprehension. First, readers may compensate for RSVP by postponing information processing until presentation has ended to ensure all the input is copied into a buffer. Under this account, we would expect a small uniform profile of surprisal weights in our 5-RSVP condition. Second, readers may compensate for RSVP by incrementally buffering and processing linguistic information, allowing perceptual processing and information processing to occur on quick but separate timescales. Our a-priori prediction for the FIFO buffer, in line with the best fit prediction from our model, was that early words should contribute less to the postpresentation reading time than later words, resulting in an increasing profile of surprisal weights in our 5-RSVP condition. As can be seen in Figure 4B, the data are more consistent with the predictions of a FIFO incremental buffer model, suggesting that surprisal effects are present in RSVP and readers compensate for the rapid presentation of text by incrementally buffering and processing linguistic information.

While our experiment was designed to control for a perceptual origin of surprisal effects, it also rules out two deci-

[^142]sion/motor origins. First, surprisal effects would be expected under Hick (1952)'s law if for each word, readers were choosing the identity of the word from multiple alternatives. Presumably, such decision processes would be disrupted by our 5-SP-RSVP condition. Second, the programming of saccadic eye movements is related to the log predictability of the target (Carpenter \& Williams, 1995). RSVP and central presentation would not require eye movements nor provide sufficient time to plan them. It is unclear if our experiment rules out the optimal preparation model of Smith and Levy (2008), which predicts surprisal effects. To summarise their argument, people respond faster to expected events and slower to unexpected events, suggesting that they are preparing for future events according to their predictability. If this finding holds true for all levels of linguistic processing-i.e., under a scale free assumption, the effect of predictability must be on a $\log$ scale so as to satisfy the multiplicative nature of joint probabilities with the additive nature of reaction times. As a result, if a word is optimally prepared for, its processing time should be a multiple of its surprisal. Unfortunately, the model does not specify if there is a time cost to preparation or minimum required preparation time. If there is a cost to optimal preparation, our experiment might not have provided sufficient time to prepare. Similarly, the predictions of Smith and Levy (2013)'s highly incremental processing account of surprisal are unclear for our experiment.

Our results have several implications for sentence processing research. First, even with a linguistic information buffer the best fitting rate of information processing is large-i.e., $121 \mathrm{bits} / \mathrm{sec}$, which corresponds to a word processing rate of 12 words $/ \mathrm{sec}$. This rate is consistent with the upper limit on comprehension in RSVP tasks (Potter, 1984). Interestingly, even under the suspended processing buffer model the rate of information would be large ${ }^{7}$. So if readers are processing information at such a fast rate, why do surprisal effects show up at all in natural reading tasks? One possibility is that even though readers can process information at these rates, they might prefer to maintain the information they are processing before proceeding. Whether this processing is perceptuali.e., waiting for some level of certainty in the percept of the word, linguistic-e.g., parallel resource allocation in syntactic parsing, or optimal preparation is still an open question, with the important implication that surprisal effects in natural reading times might not be a measure of syntactic processing difficulty. Another important possibility is that rate of information processing is not consistent across different tasks. In our opinion, RSVP is a demanding task and as such the rate of information processing might differ from natural rates of processing. In this case, surprisal effects in natural reading could reflect a real processing bottleneck, inspiring the new question, why don't we adjust our processing rates to alleviate the bottleneck?

[^143]Second, our buffer model suggests that there is a decoupling of perceptual and linguistic information processing ${ }^{8}$, which is potentially relevant for two sentence processing phenomenon: spill-over effects and right context effects. In selfpaced reading and eye-tracking data (e.g., Smith \& Levy, 2013; Shvartsman, Lewis, \& Singh, 2014), researchers sometimes discover spill-over effects-i.e., properties of previously fixated/presented words are reflected in the reading time of the current fixated/presented word. Usually spillover effects are realized as increased reading times following a word thought to be difficult to process (or conveying a large amount of information). Spill-over effects could be explained as perceptual processes continuing to advance through the sensory input while being sensitive to the information processing slightly lagging behind in a buffer. For example, if a buffer had a fixed capacity, perceptual processing might stall on words further in the input than is currently being processed until the buffer has room for more information.

Right context effects occur when information further in the sensory input influences previously perceived information. This plays out differently depending on modality. Readers maintain perceptual uncertainty about word identities (Levy, Bicknell, Slattery, \& Rayner, 2009) and their regressive eye movements can be linked to future input increasing uncertainty about past input (Bicknell \& Levy, 2010). Eavesdroppers, on the other hand, do not have the luxury of playback. In speech processing, listeners maintain uncertainty about words (Bicknell, Tanenhaus, \& Jaeger, 2015) and have to hope that the future context will disambiguate the signal for them. The current proposal is that the processing of a segment of speech operates beyond the duration of the speech segment (Dahan, 2010). Arguably, the maintenance of unprocessed information implicates a linguistic buffer. Future research should look into using buffer models to account for these phenomenon.

Our results, in line with previous RSVP studies, show that readers can process text faster than they would process text naturally. Both the surprisal effect in our mSP condition and the information processing rate parameter in our model of the 5-RSVP condition are the first pieces of evidence for surprisal effects in RSVP. Our data are consistent with a FIFO buffer model suggesting that when readers are quickly bombarded with information, they store linguistic information in a buffer and immediately begin processing that information serially at a fixed rate. Our buffer model suggests a looser temporal coupling between perceptual processing and linguistic processing than had previously been theorized. Our initial analyses using a FIFO buffer model prompt further research on the nature of the buffer and how the buffer may be implicated in other sentence processing phenomena.

## References

Bicknell, K., \& Levy, R. (2010). A rational model of eye movement control in reading. In J. Hajivc, S. Carberry, S. Clark, \&

[^144]J. Nivre (Eds.), Proceedings of the 48th annual meeting of the association for computational linguistics (acl) (pp. 1168-1178). Uppsala, Sweden: Association for Computational Linguistics.
Bicknell, K., \& Levy, R. (2012). Word predictability and frequency effects in a rational model of reading. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society.
Bicknell, K., Tanenhaus, M., \& Jaeger, T. (2015). Listeners can maintain and rationally update uncertainty about prior words. Manuscript submitted for publication. [KB].
Carpenter, R., \& Williams, M. (1995). Neural computation of log likelihood in control of saccadic eye movements. Nature, 377(6544), 59.
Dahan, D. (2010). The time course of interpretation in speech comprehension. Current Directions in Psychological Science, 19(2), 121-126.
Grodner, D., \& Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. Cognitive science, 29(2), 261-290.
Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies (pp. 1-8).
Hick, W. E. (1952). On the rate of gain of information. Quarterly Journal of Experimental Psychology, 4(1), 11-26.
Kuperman, V., \& Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. Journal of memory and language, 65(1), 42-73.
Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3), 11261177.
Levy, R., Bicknell, K., Slattery, T., \& Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. Proceedings of the National Academy of Sciences, 106(50), 21086-21090.
Lewis, R. L., \& Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. Cognitive science, 29(3), 375-419.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., \& Brockman, W. (2011). The google books team, joseph p. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden, 176-182.
Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. New methods in reading comprehension research, 6989.

Norris, D. (2006). The bayesian reader: explaining word recognition as an optimal bayesian decision process. Psychological review, 113(2), 327.
Peirce, J. W. (2007). Psychopypsychophysics software in python. Journal of neuroscience methods, 162(1), 8-13.
Potter, M. C. (1984). Rapid serial visual presentation (rsvp): A method for studying language processing. New methods in reading comprehension research, 118, 91-118.
Rayner, K., Schotter, E. R., Masson, M. E., Potter, M. C., \& Treiman, R. (2016). So much to read, so little time how do we read, and can speed reading help? Psychological Science in the Public Interest, 17(1), 4-34.
Shvartsman, M., Lewis, R. L., \& Singh, S. (2014). Computationally rational saccadic control: An explanation of spillover effects based on sampling from noisy perception and memory. ACL 2014, 1.

Smith, N. J. (n.d.). Zs: A file format for efficiently distributing, using, and archiving record-oriented data sets of any size.
Smith, N. J., \& Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In Proceedings of the cognitive science society (Vol. 30).
Smith, N. J., \& Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302319.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., \& Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632.

# Different processes for reading words learned before and after onset of literacy 

Padraic Monaghan (p.monaghan@lancaster.ac.uk) Ya-Ning Chang (y.chang5@lancaster.ac.uk)<br>Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK<br>Stephen Welbourne (stephen.welbourne@googlemail.com)<br>Neuroscience and Aphasia Research Unit (NARU), University of Manchester, Manchester, M13 9PL, UK


#### Abstract

Learning to read has a substantial effect on the representations of spoken and meaning forms of words. In this paper we assess literacy effects beyond representational changes, focusing on adaptations to the architecture of the reading system that maps between these representations. We present a connectionist model of reading that predicted distinct processing of pre- and post-literacy acquired words. For reading for meaning, words learned prior to literacy were processed more indirectly via phonological representations, whereas for post-literacy acquired words, processing was more direct along the orthography to semantics pathway. This more computationally intensive route was prioritised because indirect phonology to semantics mappings were unavailable. Such an effect was less apparent for naming, because learning direct orthography to phonology mappings is less computationally intensive. These results were confirmed in an analysis of naming and lexical decision behavioural data. The effect of literacy onset remains an observable artefact in adult reading.


Keywords: literacy; age of acquisition; language development; reading fluency; reading comprehension; computational modelling.

## Effects of literacy on reading

There are multiple influences on readers' speed and accuracy of reading, and these have been extensively documented in the literature over the last 50 years of reading research. For instance, higher-frequency words tend to be accessed more quickly and accurately than lower-frequency words, and early-acquired words tend to be responded to faster and more accurately than later-acquired words, referred to as an "age of acquisition" (AoA) effect (Brysbaert, \& Ghyselinck, 2006; Cortese \& Khanna, 2007; Juhasz, 2005; Monaghan \& Ellis, 2002).

Theories of the origin of the AoA effect on reading are two-fold. One view is that early acquired words result in prioritised lexical semantic representations, because they enter first of all into the lexical semantic associative network, and subsequently learned words are then connected to previously acquired words (Brysbaert \& Ghyselinck, 2006). Analyses of semantic associations by Steyvers and Tenenbaum (2005) confirmed that early acquired words do have more words associated with them than later acquired words, and they demonstrated that small-scale illustrative versions of this growing semantic associative network could prioritise early acquired words in semantic processing.

An alternative perspective is that AoA effects are instead found in the mappings between representations, rather than the representations themselves (Monaghan \& Ellis, 2010).

Early acquired words are learned when the neural network supporting the mappings among print, sound and meaning is plastic and able to acquire mappings effectively. Mappings for later acquired words are required to fit around the previously learned mappings, when the neural network has lower plasticity, resulting in prioritisation for early over later acquired words. Such AoA effects are predicted to be greater for arbitrary mappings, such as between meaning and sound, rather than for (quasi-)regular mappings such as between print and sound, because learning arbitrary mappings is more computationally intensive and therefore affected more by reduced plasticity (Lambon Ralph \& Ehsan, 2006). However, AoA effects ought still to be observed even for regular mappings because of the smaller, but still present, effect of reducing plasticity in learning the mappings.

These predictions have been supported by meta-analyses of behavioural studies (Brysbaert \& Ghyselinck, 2006) which have investigated AoA effects for naming and for lexical decision. It is generally assumed that for naming, semantic representations of words are minimally involved in producing the phonological form of a word from its orthographic form (Harm \& Seidenberg, 1999). However, lexical decision appears to implicate semantic representations to a greater degree (Chang et al., 2016; Plaut, 1997), in that semantic properties of words, such as imageability or concreteness, account for more variance in lexical decision or picture naming responses and little for written word naming (Balota et al., 2004; Catling \& Johnston, 2009). Brysbaert and Ghyselinck (2006) showed that AoA effects were much greater for tasks involving semantics, including lexical decision, than for tasks involving production of phonology (see also Cortese \& Khanna, 2007). However, the fact that AoA does still account for some variance in naming indicates the effects of plasticity in the quasi-regular print to sound mapping for English (see Lambon Ralph and Ellis, 2000, and Monaghan and Ellis, 2010, for computational illustrations of this).

Conversely, the size of the AoA effect can be used to indicate the extent to which the pathways to and from lexical semantics in the reading system are involved in reading. If the AoA effect is large, then semantics is likely to be involved, if the effect is small then semantics is less likely to be involved. Chang et al. (2016) implemented a triangle computational model of single word reading, and varied the point at which words were presented to the model, to simulate different AoA of words. For simulations of naming, AoA had a significant effect, but for simulations
of lexical decision, AoA accounted for a substantially larger proportion of variance.

One absence from these theoretical and implemented models of reading, however, is the role not only of AoA but also of different modes by which words are acquired. Literacy is known to have profound effects on language processing, resulting in changes to phonological awareness (Hulme, Bowyer-Crane, Carroll, Duff, \& Snowling, 2012; Morais, Cary, Alegria \& Bertelson, 1979), changes to phonological processing of words (Smith, Monaghan, \& Huettig, 2014), as well as semantic fluency (Kosmides, Tsapkini, Folia, Vlahou, \& Kiosseoglou, 2004), and even visual processing (Szwed, Ventura, Querido, Cohen, \& Dehaene, 2012).

However, less studied are the potential effects of literacy on the architecture of the reading system in terms of pathways employed between different representations of words. Prior to literacy, the learner acquires mappings between sound and meaning representations of words, through listening and comprehending words, and speaking words for others' comprehension. However, once the child begins to learn to read for these already known words, mappings will be generated from print to the stored sound and meaning representations. But for new words, the print form will be mapped onto newly acquired sound and meaning representations, where the mappings between sound and meaning are not available in advance.

In terms of the operation of the reading system, this difference between pre-literacy and post-literacy acquired words is likely to be profound. In the triangle model of reading (Seidenberg \& McClelland, 1989) there are two routes by which a printed word can be pronounced. This can occur directly, through learned mappings between print and sound, or indirectly from print via semantics to sound (see Figure 1). Similarly, for reading comprehension, the mapping from print can be directly to meaning, or indirectly, from print to sound to meaning. For pre-literacy acquired words, the indirect route is more likely to be available, because the sound to meaning routes are already acquired, whereas for post-literacy words, the indirect route requires two mappings to be acquired.

Furthermore, the properties of the mappings from print to sound and meaning will also contribute to the extent to which direct and indirect mappings are utilised. Regular mappings, such as between print and sound in English, are easier to acquire than arbitrary mappings, such as between print and meaning. Thus, the direct route is more likely to be prioritised for print to sound mappings than the indirect route, and the indirect route is more likely to be prioritised for print to meaning mappings than the direct route, because the indirect route is more easily acquired, at least for words acquired pre-literacy, where the sound to meaning mapping is already in place in the language processing system.

Based on this theory, we predict that there is likely to be a distinction between pre-literacy and post-literacy processing of words' print to meaning mappings, as in lexical decision. Pre-literacy, the indirect route is more likely to have a
greater influence on processing. Post-literacy, the direct route is likely to have a greater influence. Whereas for print to sound mappings, as in word naming, we predict no difference between pre- and post-literacy processing, because both will be mapped via fast-acquired direct print to sound mappings, which will have an equal influence on reading.

In this paper, we first provide a computational test of the extent to which the triangle model of reading predicts different processing routes for words pre- and post-literacy. We then test whether the predictions of the model are observed in behavioural data on word naming and lexical decision response times. For both the simulation and the behavioural data, we use the size of the AoA effect as an index of the extent to which direct mappings from orthography to semantics are implicated in the reading system. For naming, a larger AoA effect indicates greater use of indirect mappings via semantics for reading tasks, for lexical decision a larger AoA effect indicates greater use of direct mappings from orthography, where arbitrary mappings between orthography and semantics are implicated. A smaller AoA effect for lexical decision indicates that the indirect quasi-regular mapping from orthography to phonology is being prioritised. It is the case that mappings between phonology and semantics are also arbitrary, but these mappings would exert a smaller AoA effect than that observed for the newly acquired mappings because they are intensively trained, and acquired earlier in acquisition, thus reducing distinctions between words due to greater plasticity of resources for early-learned mappings (see e.g., Ellis \& Lambon Ralph, 2000; Monaghan, Chang, Welbourne, \& Brysbaert, 2017).


Figure 1. The architecture of the triangle model of reading.

## A computational model of literacy effects on reading processes

## Method

## Network Architecture

The model is based on the connectionist triangle model of reading (Harm \& Seidenberg, 2004; Seidenberg \& McClelland, 1989), as shown in Figure 1. The critical property of the model is that it incorporates three
representations of single words - print, sound, and meaning. Each of these representations is interconnected by sets of hidden units that permit the mappings between representations to be acquired as a consequence of exposure. The sound and meaning layers were also each connected to a set of attractor units that enable the model to develop highfidelity phonological and semantic representations of words.

Also included was a context layer to enable disambiguation of the meaning of homonyms (e.g., /beIs/ as the instrument bass and as the location base, see Chang et al., 2016, for more details).

## Representations

The representations of words were derived from Harm and Seidenberg's (2004) version of the triangle model. Printed words were represented across 14 letter slots, with each letter slot comprising 26 units relating to one letter of the alphabet. If a letter was present in a slot, then the unit corresponding to the letter had activity 1 , otherwise the units were inactive. Spoken words were represented in terms of segmental phonological features, across 8 phoneme slots of 25 binary phonological feature units, with distinct phonemes represented in terms of overlapping subsets of the units representing the features. Finally, lexical meaning representations were constructed from semantic features in WordNet (Miller, 1990). Each word activated a subset of the 2446 semantic features in the semantic layer of the model, with activity 1 if the semantic feature was associated with the word.

The model was eventually trained to read 6229 monosyllabic words, which were presented during reading training according to log-compressed frequency, where frequency was taken from the Wall Street Journal corpus (Marcus, Santorini, \& Marcinkiewicz, 1993), to be consistent with Harm and Seidenberg's (2004) implementation of the triangle model.

## Training Procedures

The training process had two phases: a pre-literacy and a post-literacy phase. In the pre-literacy training, the model learned to map between phonology and semantics on a subset of words from the entire training set, that children are more likely to have learned before beginning reading. In the post-literacy phase, the model was trained to learn to map all words from orthographic forms onto phonology and semantics.

In pre-literacy training, the model was trained on oral language tasks, including a speaking task (mapping from semantic to phonological representations), a hearing task (mapping from phonological to semantic representations), as well as tasks that assisted in developing stable attractors at phonology and semantics (mapping from phonological to phonological representations, and from semantic to semantic representations). For the speaking task, the semantic input pattern for a selected word was clamped for eight time steps, then in the last two time steps, the model was required to reproduce the phonological form for the word. The difference between the model's actual production and the target phonological production was backpropagated through
the network and connections were adjusted to reduce error. Similarly, for the hearing task, the phonological input and the context were clamped for 8 time steps, and the model was required to produce the target semantic form at the output. For the stable attractor tasks, the input was presented then activation cycled for 6 time steps, before the model was required to reproduce the originally inputted phonological or semantic representation. For pre-literacy training, the four tasks were interleaved, with $40 \%$ of trials each for the speaking and hearing tasks, and $10 \%$ each for the phonological and semantic attractor trials. There were 600,000 trials altogether.
For pre-literacy training, the model was exposed to 2,973 monosyllabic words, which were selected to be the most common words occurring in reading materials before age 18 , and therefore those words most likely that children come across prior to literacy onset. Words were presented randomly, but selected according to their frequency. The model was trained with a learning rate of 0.05 using backpropagation through time, and cross-entropy error was computed. No adjustments to weights were made if the model was within 0.1 of the target for each output unit.
In the post-literacy training, the model was given printed word forms, and required to learn to map onto phonological and semantic representations. Words were presented to the model incrementally, according to the reading-age at which words occurred. Similar to Monaghan and Ellis (2010), reading developed cumulatively, over 14 reading stages reflecting reading materials experienced from age 5 to 18 , determined from the educator's word frequency guide (Zeno et al., 1995), see Chang et al. (2016) for more details.

For each word, the model cycled for 12 time steps of activation after which the model had to generate the phonological and semantic representations of the word. These reading trials were interleaved with hearing and speaking trials, and phonological and semantic attractor trials, to ensure that the pre-literacy mappings were maintained during reading training. There were 1.74 million post-literacy training trials altogether.

Critically, by the end of training, the model had been exposed to all words, but some of these had been acquired prior to literacy onset, and others were acquired from print. We refer to these words as pre- and post-literacy words.

## Testing Procedures

To measure pre-literacy oral language skills, the model was tested on its productions for the speaking and hearing tasks. For semantics, if the model was closer to the target word than any other word, then it was judged to be accurate. For phonology, if the model was closer to the target phoneme at each phoneme slot then it was judged to be correct.
For the analysis of reading performance, we interpreted orthographic to phonological representations to be analogous to behavioural naming responses (Chang, Furber, \& Welbourne, 2012), and orthographic to semantic mappings to relate to lexical decision responses (see, e.g., polarity measure in Plaut, 1997, and Chang et al., 2016).

## Results

At the end of pre-literacy training, of the words to which the model had been exposed prior to onset of literacy, the model was able to speak $90.7 \%$, and comprehend $91.7 \%$ correctly. After reading training, the model was accurate for $99.4 \%$ of phonology and $93.3 \%$ of semantics for the reading task.

To assess whether literacy changed patterns of processing in the model, multiple regression analyses were conducted for the model's simulations of word naming and lexical decision tasks. The mean square error of the model's productions was taken as the dependent variable, and a set of psycholinguistic variables were included as predictors, to relate to previous regression analyses of behavioural data (e.g., Balota et al., 2004; Cortese \& Khanna, 2007). These variables were cumulative frequency (CF), orthographic neighbourhood size (OrthN) (Coltheart, 1977), word length (Len), consistency (Cons) (which was the proportion of words with the same pronunciation of the orthographic rime, e.g., "gave/save" versus "have"), and AoA, which was the reading stage during training for the model. Error scores were log transformed and all the predictor variables were centred.

To examine the effect of literacy onset on the model's performance, hierarchical regression analyses were conducted. At step 1, all psycholinguistic variables were entered, then at step 2 whether the word appeared pre- or post-literacy was entered as a variable interacting with AoA. If processing changes from pre- to post-literacy, then the effect of AoA at the point of literacy onset should change, as an index of the involvement of semantics - reflected in a significant interaction. It was not possible to include literacy onset as a separate variable because it is highly correlated with the interaction term. The results for naming and lexical decision are shown in Table 1.

Table 1. Results from the regression analysis for naming and for lexical decision in the computational model.

|  |  | Naming | Lexical Decision |
| :--- | :---: | :---: | :---: |
|  |  | $\beta$ | $\beta$ |
| Step 1 | CF | $-0.179^{* * *}$ | $-0.107^{* * *}$ |
|  | OrthN | $-0.256^{* *}$ | 0.012 |
|  | Cons | $-0.247^{* * *}$ | -0.016 |
|  | Len | $-0.071^{* * *}$ | $-0.127^{* * *}$ |
|  | AoA | $0.198^{* * *}$ | $0.452^{* * *}$ |
| Step 2 | AoA x | $0.219^{* * *}$ | $0.501^{* * *}$ |
|  | Literacy <br> onset | $\Delta R^{2}=0.37 \%$ | $\Delta R^{2}=1.96 \%$ |
| ${ }^{* * *} p<.001 ;{ }^{* *} p<.01 ; \beta$ is a standardized beta value. |  |  |  |

Literacy onset was a significant predictor of changes in the model's performance - at the point of literacy onset, the regression gradient for the AoA effect changed, such that words acquired pre-literacy demonstrated a smaller change in response times associated with increasing AoA compared to words acquired post-literacy. This effect was substantially larger for lexical decision than for naming responses, suggesting that processing for pre-literacy acquired words used the indirect route from orthography to semantics via phonology, whereas the post-literacy acquired words used the direct orthography to semantics route.

We next tested whether a similar change in processing was associated with literacy onset in naming and lexical decision behaviour.

## Testing the literacy effect in word processing

## Method

The data were a subset of responses from the English Lexicon Project (Balota et al., 2007), comprising naming and lexical decision response times from a set of 816 young adult participants from a range of universities. We acquired data for 2,536 monosyllabic words, for which all the psycholinguistic variables could be generated.

Word-form frequency, orthographic neighbourhood size, and word length were taken from the CELEX database (Baayen, Pipenbrock, \& Gulikers, 2005). These three measures were taken from the same dataset to ensure consistency across these measures. AoA was taken from Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). Consistency of words was determined in the same way as for the computational simulation.

## Results

We first aimed to replicate the results of Balota et al. (2004, 2007) in determining the role of frequency, word length, neighbourhood size, consistency, and AoA in a linear regression on naming response times and lexical decision response times.

Then, we measured whether there was an effect of onset of literacy in the behavioural data through adding an interaction between AoA and literacy onset. Age of literacy onset could not be included a priori as with the simulation, however, we assumed that if there is an effect of onset of literacy, then this should occur somewhere close to the age of 5 . Onset of literacy was thus determined iteratively between the age of $3,4,5,6$, and 7 years in order to assess whether there is a discontinuity in response times predicted by AoA that changes around the age children begin formal literacy. We took as an indicator of discontinuity a significant interaction between AoA and literacy onset, though see Baayen, Feldman, and Schreuder (2006) for an alternative means of measuring discontinuities (note they were unable to test AoA because of small sample size).

For naming and lexical decision response times, the results of the multiple regression are shown in Table 2. For naming, adding the interaction between onset of literacy and

AoA for any of the ages 3 to 7 did not significantly improve the model fit (Bonferroni corrected).

For lexical decision response times, there were significant effects of literacy onset found at ages 5, 6, and 7, with the largest effect for age 6. Figure 2 shows the effect of this discontinuity in predicting response times for lexical decision when the onset of literacy is implemented at age 6. The same Figure illustrates no statistically significant discontinuity effect for naming response times.

Table 2. Results from the regression analysis for naming and for lexical decision in the behavioural data.

|  |  | Naming <br> $\beta$ | Lexical Decision $\beta$ |
| :---: | :---: | :---: | :---: |
| Step 1 | Log-frequency | $-0.156^{* * *}$ | $-0.305^{* * *}$ |
|  | OrthN | $-0.255 * * *$ | -0.001 |
|  | Cons | $-0.115^{* * *}$ | -0.032* |
|  | Len | $0.165^{* * *}$ | $-0.062 * * *$ |
|  | AoA | 0.174*** | $0.440^{* * *}$ |
| Step 2 | AoA x Literacy onset age 3 | -3.220 | 1.572 |
|  | AoA x Literacy onset age 4 | -0.653 | 0.369 |
|  | AoA x Literacy onset age 5 | 0.150 | 0.387* |
|  | AoA x Literacy onset age 6 | 0.151 | $0.348^{* * *}$ |
|  | AoA x Literacy onset age 7 | 0.154 | $0.310^{* * *}$ |

## General Discussion

Onset of literacy has a profound effect on cognition, but generally these effects have been assessed on the representations involved in reading, rather than the pathways involved in mapping between these representations (Hulme et al., 2012; Morais et al., 1979; Smith et al., 2014). In this paper, we show that onset of literacy likely has a long-standing impact on the architecture of the reading system. For words that are in the learner's vocabulary prior to onset of literacy, reading can proceed via two routes - directly, by newly learned mappings from orthography to semantics, or orthography to phonology, or can instead exploit indirect pathways that incorporate learned mappings between phonology and semantics that the learner already has cemented in their language system.


Figure 2. Interaction between AoA and onset of literacy at age 6 in lexical decision but not in naming responses.

For naming tasks, the use of this prior semantics to phonology knowledge has a minimal effect, because the quasi-regularity of orthography to phonology mappings is relatively easy to acquire. The greater difficulty of learning an arbitrary mapping from orthography to semantics, then using this semantic representation to activate the previously acquired phonological representation for the known word, means that this indirect processing is unlikely to be involved differentially for words learned pre- versus post-literacy.
For lexical decision, or other tasks involving activation of semantic representations, the role of literacy onset appears to be quite different. The computational model predicted that when prior knowledge about phonological and semantic associations is available, as it is for pre-literacy acquired words, then an indirect route is likely to be involved in mapping from orthographic to semantic representations. For words learned post-literacy, this prior knowledge is not available, and so the reading system has to proceed via generating either a new mapping from orthography to semantics, or a new mapping from phonology to semantics. Thus, a distinct pattern of response is likely to be observed for lexical decision of pre- and post-literacy words.

The behavioural results provide support for the computational predictions of different pathways used in reading pre- versus post-literacy. Even though literacy onset was several years before the participants in the lexical decision study were tested, the vestiges of literacy onset appear to be still observable in reading behaviour. We acknowledge that literacy onset is not a sudden change, as some new words will still be acquired aurally even after reading training has commenced, and proficient reading is not immediate, but requires extensive, sometimes strenuous, training (e.g., Seidenberg, 2017). Nevertheless, we have shown that literacy onset changes the use that the reader makes of the language system, and this differential use of the system survives to be observed in behavioural responses
even after decades of reading practice

## Acknowledgments

This research was supported by ESRC grant RES-000-224049.

## References

Baayen, R. H., Feldman, L. B., \& Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. Journal of Memory and Language, 55, 290-313.
Baayen, R.H., Pipenbrock, R. \& Gulikers, L. (1995). The CELEX Lexical Database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., \& Yap, M. (2004). Visual word recognition of single-syllable words. Journal of Experimental Psychology: General, 133(2), 283-316.
Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., \& Treiman, R. (2007). The English lexicon project. Behavior Research Methods, 39(3), 445-459.
Brysbaert, M., \& Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. Visual Cognition, 13(7-8), 992-1011.
Catling, J. C., \& Johnston, R. A. (2009). The varying effects of age of acquisition. $Q J$ Exp Psychol, 62(1), 50-62.
Chang, Y. N., Furber, S., \& Welbourne, S. (2012). "Serial" effects in parallel models of reading. Cognitive Psychology, 64(4), 267-291.
Chang, Y. N., Monaghan, P., \& Welbourne, S. (2016). Effects of experience in a developmental model of reading. Proceedings of the 38th Cognitive Science Society Conference.
Coltheart, M., Davelaar, E., Jonasson, J. T., \& Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), Attention and Performance VI (pp. 535-555). Hillsdale, NJ: Lawrence Erlbaum Associates.
Cortese, M. J., \& Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: an analysis of 2,342 words. Quarterly Journal of Experimental Psychology, 60(8), 1072-1082.
Ellis, A. W., \& Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. Journal of Experimental Psychology-Learning Memory and Cognition, 26(5), 1103-1123.
Ghyselinck, M., Lewis, M. B., \& Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. Acta Psychologica, 115(1), 43-67.
Harm, M. W., \& Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. Psychological Review, 111(3), 662-720.

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., \& Snowling, M. J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read combining intervention studies with mediation analyses. Psychological Science, 23(6), 572-577.
Juhasz, B. J. (2005). Age-of-Acquisition effects in word and picture identification. Psychological Bulletin, 131(5), 684712.

Kosmidis, M. H., Tsapkini, K., Folia, V., Vlahou, C. H., \& Kiosseoglou, G.. (2004). Semantic and phonological processing in illiteracy. Journal of the International Neuropsychological Society, 10(6), 818-827.
Kuperman, V., Stadthagen-Gonzalez, H., \& Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, 44(4), 978-990.
Lambon Ralph, M. A., \& Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. Visual Cognition, 13, 292248.

Marcus, M. P., Marcinkiewicz, M. A., \& Santorini, B. (1993). Building a large annotated corpus of English: the penn treebank. Computational Linguistics, 19(2), 313-330.
Miller, G. A. (1990). WordNet: An on-line lexical database. International Journal of Lexicography, 3, 235-312.
Monaghan, J., \& Ellis, A. W. (2002). What exactly interacts with spelling--sound consistency in word naming? Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(1), 183-206.
Monaghan, P., Chang, Y.N., Welbourne, S., \& Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. Journal of Memory and Language, 93, 1-21.
Monaghan, P., \& Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. Journal of Memory and Language, 63(4), 506-525.
Morais, J., Cary, L., Alegria, J., \& Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7(4), 323-331.
Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. Language and Cognitive Processes, 12(5-6), 765-805.
Steyvers, M., \& Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive Science, 29(1), 4178.

Szwed, M., Ventura, P., Querido, L., Cohen, L., \& Dehaene, S. (2012). Reading acquisition enhances an early visual process of contour integration. Developmental Science, 15(1), 139-149.
Zeno, S. (Ed.). (1995). The educator's word frequency guide. Brewster, NJ: Touchstone.

# Multiple variable cues in the environment promote accurate and robust word learning 

Padraic Monaghan (p.monaghan@lancaster.ac.uk)<br>James Brand (j.brand@lancaster.ac.uk)<br>Rebecca L.A. Frost (r.frost1@lancaster.ac.uk)<br>Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK<br>Gemma Taylor (g.taylor4@salford.ac.uk)<br>Department of Psychology, Salford University, Salford M5 4WT, UK


#### Abstract

Learning how words refer to aspects of the environment is a complex task, but one that is supported by numerous cues within the environment which constrain the possibilities for matching words to their intended referents. In this paper we tested the predictions of a computational model of multiple cue integration for word learning, that predicted variation in the presence of cues provides an optimal learning situation. In a cross-situational learning task with adult participants, we varied the reliability of presence of distributional, prosodic, and gestural cues. We found that the best learning occurred when cues were often present, but not always. The effect of variability increased the salience of individual cues for the learner, but resulted in robust learning that was not vulnerable to individual cues' presence or absence. Thus, variability of multiple cues in the language-learning environment provided the optimal circumstances for word learning.


Keywords: word learning; multiple cues; strategies; gesture; prosody; cross-situational learning.

## Cues for word learning

Learning how words relate to objects, actions, properties, or relations in the world is a complex task. One of the key difficulties is that word learning provides few explicit constraints on which words can relate to particular aspects of the environment (Quine, 1960). Thus, in acquiring vocabulary, children must resolve a many-to-many (possibly even an infinite-to-infinite) mapping between words in utterances and elements of the environment around them. So how do children solve this task?

There are two proposals for how learning word-referent mappings can be constrained. The first is that children have internal biases that apply to language learning situations that limit possible referents to words (Markman, 1994). For instance, mutual exclusivity refers to the assumption in word learning situations that each referent has only one name, leading children to pair an unnamed object with a novel word (Markman \& Wachtel, 1998). In terms of limiting referents, children seem to be biased to linking a word with a whole object rather than a part of an object (Macnamara, 1982), and may more readily form categories of objects with similar shape which are referred to by the same word (Baldwin, 1992).

The alternative proposal for resolving the many-to-many mapping problem in word learning is that the environment, rather than the learner, contains many properties that assist
in constraining possible mappings (MacWhinney, 1991). Though a single learning situation contains many possible words and many possible referents for those words, over multiple situations, children may observe that there are cooccurrences between particular words and particular elements of the environment. Yu and Smith (2007) showed that learners are able to exploit such cross-situational statistical relations between words and referents. However, the statistical associations are noisy in real-world childdirected speech settings (Yu \& Ballard, 2007), and so additional cues in the environment are likely to assist further in constraining learning.

One possibility is distributional information in terms of co-occurrences between words. In English child-directed speech, determiners reliably precede nouns in complex utterances (Monaghan \& Mattock, 2012), and these distributional cues can assist the child in knowing which potential words in an utterance are likely to refer to objects in their environment (Fitneva, Christiansen, \& Monaghan, 2009). Other distributional cues that are readily available to children can also provide information about verb categories, and function versus content word distinctions (Childers, 2011; Christiansen \& Monaghan, 2016).

Prosodic information is another cue to assist in reducing the many-to-many mapping problem, not only providing information about different grammatical categories (Christiansen \& Monaghan, 2016) but also indicating speaker focus in a learning situation: Messer (1981) found that approximately $50 \%$ of child-directed utterances with a learning goal had the referring word reaching the highest amplitude.

For further reducing the possibilities for the intended referent, gestural cues provide additional cues to constrain word learning, with $15 \%$ of child-directed speech utterances accompanied by gestures that guided the child to the object being referred to (Iverson, Capirci, Longobardi, \& Caselli, 1999).

## Combining cues for word learning

Individually, then, cues appear to be noisy but informative sources of information about intended referents. Thus, combining cues is likely to result in yet more robust and faster learning. There are several models for how multiple cues may interact for word learning.

First, cues may be additive, such that more information provides cumulative evidence about word-referent mappings. For instance, in a computational model, Yu and Ballard (2007) demonstrated that mapping accuracy improved with the addition of distributional cues.

However, an alternative model for how multiple cues may support learning is provided by Bahrick, Lickliter, and Flom's (2004) intersensory redundancy hypothesis. In this theoretical model, multiple cues that indicate the same structure in language (such as multiple cues indicating the word-referent mapping, for instance) enable the learner to realise that this relation is not random, but carries information about the stimuli. Consequently, cues that are correlated increase in saliency and are attended to more as learning proceeds.

However, this view of increased saliency from redundant cues only applies when there are overlapping cues to structure, and the distribution of cues in the learning environment may be very different. Monaghan et al. (2007) examined cues to grammatical categories of words across a range of languages. They found that distributional information provided, unsurprisingly, valuable information about the role of words in each language - for instance, in English words that belonged to the verb category tended to succeed "you" and precede "the", whereas words that belonged to the category of nouns tended to succeed "the", and precede "to". But, in addition, Monaghan et al. (2007) also found that phonological coherence also applied to these grammatical categories - though there is substantial variation, nouns tend to sound like other nouns and verbs tend to sound like other verbs, in terms of a range of phonological and prosodic properties.

Yet, it was the interplay of these cues that was striking: when distributional information was a weak indicator of grammatical category, Monaghan et al. (2007) found that phonological cues were more reliable, and vice versa. Thus, there was not so much a redundant overlap of cues, but rather a serendipitous arrangement of cues across situations to provide useful information (Christiansen \& Monaghan, 2016).

An alternative perspective, then, is that multiple cues for language structure enable robust learning, but not due to intersensory redundancy, but rather due to providing a safety net that is resistant to variation of their presence in the environment. In Monaghan (2017) this idea of degeneracy was implemented in a connectionist model that took as input multiple information sources to support learning of crosssituational statistical regularities between an object in vision and a word in auditory input, when both the object and the word occurred alongside others. The model was able to learn the cross-situational statistical regularities, but this learning was boosted when additional cues were added to the model's learning environment. One was distributional information (where the referring word was preceded by a marker word, such as "the" preceding a noun). Another was a prosodic cue, where the referring word in the utterance was emphasised in the auditory input. The final cue was a
gestural cue, where attention was drawn to the object that was being referred to in the utterance. In each case, adding the cue improved the model's learning. Furthermore, adding all the cues improved performance still further.

The second set of simulations in Monaghan (2017) tested what effect individually unreliable cues would have on learning. The presence of each of the three cues varied between $33 \%$ and $100 \%$ of the time, but note that in most learning situations, at least one of the cues was likely to be present. The reduction of reliability of multiple cues reduced the speed of learning, however, following training, the ability of the model to respond correctly to word-object mappings when they were presented with no additional cues in the environment was more robust when cues were individually unreliable. The presence of noise in the environment, when that environment provides an unreliable constellation of individual cues, meant that the model was better able to recognise words when the environment was momentarily impoverished. Consider a language instructor who always pointed to the object to which they were referring. That is likely to be helpful for constraining the potential referents for words that the learner hears. But what would happen when the instructor is distracted - or a new instructor with different habits arrives - and does not provide the gestural cue? If the cue was previously $100 \%$ reliable, then this would become a crutch that was relied upon for determining the speaker's intention, and the referent would not be identifiable if not gestured towards.

A computational approach with a similar outcome is Srivastava et al.'s (2014) dropout model, where hidden units in a model are stochastically deactivated to prevent the model overlearning one aspect of the input - to resist relying only on the most reliable information stream in the environment, and consequently preventing effective generalisation. This switching off meant that the model maximised use of information from the environment. However, critically for our purposes, the learning system does not selectively prevent attention to environmental information. The noise in the language environment provides this function. Far from being a problem for learning, environmental noise enabled effective, reliable, and robust learning to take place, providing a positive perspective on poverty of the stimulus (Chomsky, 2005). Indeed, stimulus poverty resulted in rich learning.

However, the benefit of multiple, noisy cues is a prediction of the degeneracy model (Monaghan, 2017) but has not yet been tested empirically. Here, we provide a behavioural test of whether the presence of multiple, variable cues promotes robust word-referent learning. We constructed a cross-situational learning task, with each situation presenting learners with two objects and a set of words (see Monaghan \& Mattock, 2012, for similar outline of the cross-situational word learning design). One of the words always referred to one of the objects, but the other object and the other words varied. Over multiple trials, participants may come to recognise that certain words and objects always co-occurred. We measured the extent to
which additional cues in the environment assisted in learning - implementing gestural, distributional, and prosodic cues to support learning, but we varied the extent to which these cues were present. The degeneracy model predicted that (very) noisy cues should slow learning, but that there may be an optimal level of variability at which learning is more accurate than perfect information conditions when all cues are present. We examined three levels of variability as well as no variability, where cues were present $25 \%, 50 \%, 75 \%$, or $100 \%$ of the time. We measured performance during training exposure, and we also measured whether learning was robust to omission of cues - by testing participants after learning on trials where no cues were present. Based on the predictions of the degeneracy model (Monaghan, 2017), we anticipated that learning would be resistant to omission of cues in all conditions, but that omission of cues may be least affected when those cues were variable during exposure.


## "tha FINTOOM noo chatten"

Figure 1. Example of a learning trial, containing distributional, prosodic (i.e., fintoom is emphasised in the speech), and gestural cues.

## Testing the effect of multiple, variable cues for word learning

## Method

## Participants

Participants were 72 native English speaking adults, mean age $=19.8$ years $(\mathrm{SD}=2.46)$, who were students at Lancaster University. Participants were paid $£ 3.50$ for participating, or received course credit. Participants were assigned to one of four conditions ( $\mathrm{N}=18$ per condition) which varied the extent to which cues were reliably present during training ( $25 \%, 50 \%, 75 \%$, or $100 \%$ of the time).

## Materials

The materials comprised a set of abstract objects and a set of novel words with which the objects were paired during learning. We took 10 arbitrary shape pictures from Fiser and Aslin (2002) (see Figure 1 for examples). For the speech, we generated 22 nonsense words. Ten of the words each referred to one of the object shapes. An additional 10 words did not refer to any shape. A final two words were also generated to act as distributional marker words. Words were read by a female native English speaker in monotone, and
were also read in emphasized form, with the speaker imagining they were speaking the word to a child. Emphasised words had higher mean pitch, greater pitch variation, longer duration, and greater intensity than monotone words (all $t(19)>8.98, p<.001$ ).

Each learning trial comprised an utterance containing a referring word and a non-referring word. When the distributional cue was present, the two words were preceded by marker words that distinguished the referring and nonreferring word. When the prosodic cue was present the referring word was emphasised, otherwise both words were monotonic. When the gestural cue was present, a finger pointed to the intended referent. In the example trial shown in Figure 1, "tha" indicates the following word is the referring word and "fintoom" refers to one of the pictures (in this case, the picture on the left). Cues were randomly selected individually according to the variability condition (e.g., for the $25 \%$ cue, there was a $1 / 4$ chance that each cue was present or absent, such that there were trials where 3,2 , 1, or no cues were present).

An additional training block was constructed from 6 novel shapes and 12 novel words, but these new training data are not reported further here.

## Procedure

Participants were instructed to try to learn which object was referred to by the speech. There were 6 blocks of training, each of which contained 30 trials, where for each trial an utterance was played through headphones and two objects were presented on a computer screen simultaneously. One of the objects was the target and always co-occurred with the referring word, the other object was selected from the remaining nine objects. Within each block of training, objects appeared an equal number of times as target and as foil, and were counterbalanced for appearing on the left or the right of the screen. Presence or absence of cues was manipulated between conditions by randomly selecting whether each cue was present or absent in $25 \%, 50 \%, 75 \%$, or $100 \%$ of trials.

Participants responded by pressing " 1 " or " 2 " for left object or right object, respectively, on a computer keyboard. No feedback was provided on accuracy of performance.

After training, participants were tested for their knowledge of word-referent mappings when all cues were absent, to determine whether learning was robust, or required presence of cues for accurate performance.

## Results

We conducted four separate analyses exploring how learning was affected by the variability of cues. In each analysis, a series of generalized linear mixed-effects models (GLMER) were performed, predicting the dependent variable of response accuracy (correct or incorrect). The models were built up incrementally, adding in fixed effects and performing likelihood ratio tests after the addition of each new fixed effect term (following Barr, Levy, Scheepers \& Tily, 2013). Random effects of participant and experiment version were included in all reported analyses.

First, we analysed learning during training. The effect of block (1-6) significantly improved model fit $\left(\chi^{2}(1)=314.1\right.$, $p<.001$ ), indicating that over the course of training, there was a significant increase in participant's response accuracy. Including variability condition ( $25 \%, 50 \%, 75 \%$ and $100 \%$ ) also significantly improved model fit $\left(\chi^{2}(3)=21.259, p<\right.$ .001). Crucially, there was a significant improvement to model fit when the interaction term of block x condition was added $\left(\chi^{2}(3)=71.113, p<.001\right)$, indicating that performance over the course of training varied by reliability condition. See Table 1 for the final model summary, which indicates that the $75 \%$ condition resulted in more rapid learning than the other conditions (see Figure 2).


Figure 2. Learning trajectories for the word-object mapping cross-situational learning task with multiple cues of different reliabilities.

Table 1. GLMER model summary predicting accuracy from training data.

| training data. |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| Fixed effects | est. | $S E$ | $z$ | $p$ |
| (Intercept) | $\mathbf{. 4 7}$ | $\mathbf{. 2 1}$ | $\mathbf{2 . 2 4}$ | $\mathbf{. 0 2 5}$ |
| Block | $\mathbf{. 3 0}$ | $\mathbf{. 0 3}$ | $\mathbf{1 0 . 9 6}$ | $<. \mathbf{0 0 1}$ |
| Condition $(25 \%-100 \%)$ | -.39 | .29 | -1.36 | .175 |
| Condition $(50 \%-100 \%)$ | -.17 | .29 | -0.58 | .559 |
| Condition $(75 \%-100 \%)$ | -.10 | .29 | -0.33 | .739 |
| Block*Condition(25\%-100\%) | $\mathbf{- . 1 7}$ | $\mathbf{. 0 4}$ | $\mathbf{- 4 . 8 1}$ | $<. \mathbf{0 0 1}$ |
| Block*Condition(50\%-100\%) | $\mathbf{- . 1 7}$ | $\mathbf{. 0 4}$ | $\mathbf{- 4 . 6 5}$ | $<. \mathbf{0 0 1}$ |
| Block*Condition(75\%-100\%) | $\mathbf{. 0 9}$ | $\mathbf{. 0 4}$ | $\mathbf{2 . 1 1}$ | $\mathbf{. 0 3 5}$ |

For the second analysis, we investigated the effect that variability of cues had on sensitivity to the individual cues during training, by measuring the effect of presence of individual cues on learning. In this analysis, only trials where at least one cue was present were included (see Figure 3).

The addition of variability condition significantly improved model fit $\left(\chi^{2}(3)=16.199, p=.001\right)$, indicating that there was a difference in overall accuracy across conditions, with performance in the $100 \%$ condition being significantly greater than the $25 \%$ and $50 \%$ conditions (both $p<.01$ ), but not the $75 \%$ condition ( $p>.05$ ). Next, the addition of cue type also significantly improved model fit $\left(\chi^{2}(2)=32.083, p<.001\right)$. This result indicates that there was a significant increase in accuracy when gesture cues were present, compared with when distributional and
prosodic cues were present, both $p<.001$. Importantly, there was a significant improvement to model fit when the interaction term of variability condition x cue type was added $\left(\chi^{2}(6)=23.665, p<.001\right)$. See Table 2 for the final model summary, which indicates that when variability is at $75 \%$, the salience of gesture cues was increased compared to the $100 \%$ condition, when cues were always present. Variability had the effect of emphasising the contribution of gesture. The benefit of gesture over the other cues was also present for $25 \%$ and $50 \%$ cues, but only when variability was at $75 \%$ was accuracy greater than the $100 \%$ condition.


Figure 3. Performance during training trials by variability condition and cue type.

Table 2. GLMER model summary predicting accuracy from trials when at least one cue was present.

| Fixed effects | est. | $S E$ | $z$ | $p$ |
| :--- | :---: | :---: | :---: | :---: |
| (Intercept) | $\mathbf{. 0 1}$ | $\mathbf{. 0 1}$ | $\mathbf{7 . 0 8}$ | $<.001$ |
| Condition (25\%-100\%) | $\mathbf{. 0 1}$ | $\mathbf{. 0 1}$ | $\mathbf{- 2 . 9 9}$ | $\mathbf{. 0 0 3}$ |
| Condition (50\%-100\%) | $\mathbf{. 0 1}$ | $\mathbf{. 0 1}$ | $\mathbf{- 2 . 5 2}$ | $\mathbf{. 0 1 2}$ |
| Condition (75\%-100\%) | .01 | .01 | 0.69 | .491 |
| Cue(dist-gesture) | .01 | .01 | 0.00 | .99 |
| Cue(dist-prosody) | .01 | .01 | 0.00 | .99 |
| Cue(dist-gesture)*Condition(25-100\%) | $\mathbf{. 0 1}$ | $\mathbf{. 0 1}$ | $\mathbf{3 . 5 3}$ | $<. \mathbf{0 0 1}$ |
| Cue(dist-gesture)*Condition(50-100\%) | $\mathbf{. 0 1}$ | $\mathbf{. 0 1}$ | $\mathbf{3 . 2 3}$ | $\mathbf{. 0 0 1}$ |
| Cue(dist-gesture)*Condition(75-100\%) | $\mathbf{. 0 1}$ | $\mathbf{. 0 1}$ | $\mathbf{2 . 9 8}$ | $\mathbf{. 0 0 3}$ |
| Cue(dist-prosody)*Condition(25\%-100\%) | .01 | .01 | 0.01 | .898 |
| Cue(dist-prosody)*Condition(50\%-100\%) | .01 | .01 | 0.62 | .535 |
| Cue(dist-prosody)*Condition(75\%-100\%) | .01 | .01 | 0.39 | .696 |

During the training trials, the number of cues available to the learner varied from 0 to 3 in the variability conditions. In order to determine the effect of number of cues present, we tested the number of cues present in terms of improvement to model fit. We found that they did $\left(\chi^{2}(1)=66.342, p<\right.$ .001 ), indicating that as the number of cues present increased, accuracy improved (see Figure 4). Further, the interaction of number of cues x variability condition also improved model fit ( $\chi^{2}(2)=14.309, p<.001$ ).

In order to determine how variability affected use of cues, we examined accuracy when all cues were present,
comparing across variability conditions. Importantly, when all three cues were present, accuracy in the $75 \%$ condition was significantly greater than the $100 \%$ condition (estimate $=.86, S E=.41, z=2.11, p=.035)$. Thus, $75 \%$ variability improved the accuracy of performance when all cues were present.

Finally, we determined whether learning was robust under conditions of cue variability, and how variability affected performance during the test trials when none of the cues were present. The addition of variability condition significantly improved model fit $\left(\chi^{2}(3)=11.357, p=.010\right)$, with significant differences between the $100 \%$ condition when compared to the $25 \%$ and $50 \%$ conditions (both $p<$ .05 ), but no significant difference between the $100 \%$ and $75 \%$ conditions $(p=.556)$. Importantly, this reflects the pattern of results found in the final block of training (see Figure 2), where performance improved as reliability of cues increased. See Figure 5 for results. Thus, in all conditions learning was robust to absence of cues.


Figure 4. Test of performance for different number of cues present during training.


Figure 5. Performance on all words after training for test trials, when no cue was present.

## Discussion

The main aim of this study was to test the effect of variation of multiple cues in the language environment for supporting word learning. We predicted, based on the degeneracy model of learning (Monaghan, 2017), that optimal performance would be a consequence of variable presence of multiple cues that aid learning. This was because the learner can exploit multiple information sources, without relying on any one cue, or coming to ignore the contribution of other highly correlated cues.

The results of the behavioural study of learning wordreferent mappings supported the degeneracy model, in that learning was faster and more accurate when distributional, prosodic, and gestural cues occurred in $75 \%$ of trials during training, than when cues were present $100 \%$ of the time.

However, greater variability - $25 \%$ and $50 \%$ occurrence of individual cues - reduced accuracy compared to the $75 \%$ condition, indicating that, for learning a small number of words, the optimal conditions were with cues present more than half the time, but not all the time. In natural language learning situations, reliability of individual cues to support word learning seems to be substantially lower. For instance, the prosodic cue of highest amplitude as an indicator of the referring word occurs in $50 \%$ of learning situations (Messer, 1981), and explicit gestural cues occur substantially less often - even as low as $15 \%$ of learning situations (Iverson et al, 1999). However, these are situations where the vocabulary is much greater than the 10 word-object mappings of the current learning situation, and additional cues to word-referent mappings when the possibilities for those mappings are exponentially higher may have a greater effect even when they occur more rarely. For instance, the model of Monaghan (2017) was trained on 100 words, and under those circumstances $50 \%$ variability was found to be optimal for learning. Scaling up the current language to larger vocabularies will be an important further test of the principles of variation in multiple environmental cues.

Analysis of the trials where individual cues were present or absent indicated that the benefit of variability in presence of cues was greatest for the gestural cue, with variability enhancing the use made of this cue when it occurred (Figure 3 ). Such a result is consistent not only with the degeneracy model of multiple cues, but also with the intersensory redundancy hypothesis (Bahrick et al, 2004), such that correlated cues increase in salience, but with the exception that the redundancy should not be absolute: if cues are perfectly correlated then their salience does not increase, as in the $100 \%$ condition.

The results from analyses of different numbers of cues present showed that combining cues boosted learning (Figure 4), indicating that the learner was exploiting information present from each of the individual cues. It was not the case, for instance, that participants learned to only attend to particular cues, as their confluence resulted in greater improvement. Indeed, when those cues were variable but all present, performance was best of all - again, the $75 \%$ variability condition outperformed the $100 \%$ condition when
all three cues were available in the trial.
In all variability conditions, learning was shown to be robust to absence of individual cues. This is an important result, because it demonstrates that though cues can support learning, they do not over-shadow the cross-situational statistical relations between particular words and objects cooccurring. This was the case even when cues were always present, thus, even if multiple cues are always present they do not result in brittle learning of statistical relations. It may be that individual cues, if occurring with high reliability could interfere with robust learning (e.g., Srivastava, 2014), and this is a topic for future investigation.

We know that the language environment is noisy, but replete with numerous multimodal cues that point in different ways to the same language structures (Whitacre, 2010; Winter, 2014; Yurovsky, Smith, \& Yu, 2013). We have shown that learners are able to exploit these multiple cues, and also their variability, to support word learning.

## Acknowledgments

This work was supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (UK) [ES/L008955/1]. Thanks to Katy Rudd for conducting pilot testing and contributing to the construction of materials.

## References

Bahrick, L. E., Lickliter, R., \& Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. Current Directions in Psychological Science, 13, 99-102.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.
Childers, J. B. (2011). Attention to multiple events helps two-1/2-year-olds extend new verbs. First Language, 31, 3-22.
Christiansen, M.H., \& Monaghan, P. (2016). Division of labor in the vocabulary. Topics in Cognitive Science, 8, 610-624.
Chomsky, N. (2005). Three factors in language design. Linguistic Inquiry, 36, 1-22.
Fiser, J., \& Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences, USA, 99, 1582215826.

Fitneva, S., Christiansen, M.H., \& Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. Journal of Child Language, 36, 967-997.
Iverson, J. M., Capirci, O., Longobardi, E., \& Caselli, M. C. (1999). Gesturing in mother-child interactions. Cognitive Development, 14, 57-75.
Macnamara, J. T. (1982). Names for things: A study of human learning (p. 4). Cambridge, MA: Mit Press.

MacWhinney, B. (1991). A reply to Woodward and Markman. Developmental Review, 11, 192-194.
Markman, E. M. (1994). Constraints on word meaning in early language acquisition. Lingua, 92, 199-227.
Markman, E. M., \& Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20, 121-157.
Messer, D. J. (1981). The identification of names in maternal speech to infants. Journal of Psycholinguistic Research, 10, 69-77.
Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. Topics in Cognitive Science, 9, 21-34.
Monaghan, P., Christiansen, M. H., \& Chater, N. (2007). The Phonological Distributional coherence Hypothesis: Cross-linguistic evidence in language acquisition. Cognitive Psychology, 55, 259-305.
Monaghan, P. \& Mattock, K. (2012). Integrating constraints for learning word- referent mappings. Cognition, 123, 133-143.
Quine, W.V.O. (1960). Word and object. Cambridge, MA: MIT Press.
Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., \& Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.
Yu, C. \& Ballard, D.H. (2007). A unified model of early word learning: Integrating statistical and social cues. Neurocomputing, 70, 2149-2165.
Yu, C. \& Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18, 414-420.
Yurovsky, D., Smith, L. B. \& Yu, C. (2013). Statistical word learning at scale: The baby's view is better. Developmental Science, 16, 959-966.

# The ecological rationality of children's option generation and decision making 

Lisa Musculus (l.musculus@dshs-koeln.de)<br>Institute of Psychology, Department of Performance Psychology, German Sport University Cologne, Germany

Azzurra Ruggeri (ruggeri@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development, Berlin, Germany
Markus Raab (raab@dshs-koeln.de)
Institute of Psychology, Department of Performance Psychology, German Sport University Cologne, Germany and London South Bank University, United Kingdom
and Babett Lobinger (lobinger@dshs-koeln.de)
Institute of Psychology, Department of Performance Psychology, German Sport University Cologne, Germany


#### Abstract

In everyday life, before deciding what to do, one has to think about what could be done. We investigate option generation from a developmental perspective, testing the predictions of the Take-The-First-heuristic (TTF). Moreover, we examine the influence of time limitation on decision-making processes. Using soccer as a testbed, 6 - to 13 -year-old children $(N=97)$ were tested in a video-based option-generation paradigm. Children's performance was aligned with predictions of TTF: Children generated a mean of 2.21 options, did so in a meaningful way and selected the first as final option in $74 \%$. With shorter time, children generated fewer and higher quality options, selected better options and more often the first option as final decision. Further, with age, an increase of the number of options generated and an increase in quality of the final decisions emerged. This age effect was more pronounced with shorter time. Implications for real-life decision-making are discussed.


Keywords: option generation; decision making; heuristics; ecological rationality; development.

## Introduction

Imagine being a young, talented soccer player. You are running, alone, through the middle field towards the goal, dribbling one opponent after the other. You are now 20 meters from the goal, facing the opposing defense rapidly closing on you. What should you do? Maybe you should try to dribble the defense, get closer to the goal and shoot from a shorter distance? Should you try to shoot at goal from where you are now? Or should you pass to one of your team members - maybe Jack, approaching from the right? Or Mike, right behind you?

Most of the time, in everyday life, before deciding what to do, one has to think about what could be done. In this paper, we investigate option generation from a developmental perspective using sport as a testbed. Moreover, we examine the influence of time limitation on option-generation and decision-making processes.

## Option generation

A decision-making strategy usually consists of a search, a stopping, and a decision rule, all together defining how and how much information has to be collected before being able to make a decision (Gigerenzer \& Todd, 1999). However, most real-world situations require us to generate alternative options before making a decision, rather than selecting one from a set of options pre-defined and generated by an experimenter (Payne, Bettmann, \& Johnson, 1988).

Very little is known about how people generate options (for an exception see e.g., Johnson \& Raab, 2003), as most research on decision-making focuses on the other three building blocks of decision making. The Take-The-First heuristic (TTF) is a cognitive model that captures option generation and decision making in familiar, yet ill-defined tasks (Johnson \& Raab, 2003; Raab, 2012; Raab \& Johnson, 2007). In the TTF the building blocks are formally defined as follows: A search rule, suggesting that alternative options are generated in order of validity meaning that subjectively better options are generated earlier; A stopping rule, according to which the generation phase should stop after two to three options have been generated; A decision rule, assuming that people should choose one of the initial options generated (Johnson \& Raab, 2003). In this sense, people would generate a few options and select one of those, rather than exhaustively generating and processing all possible options. However, because those options were generated in order of validity, the decision, although fast and frugal, would tend to be accurate.

Studies with adults and adolescents testing the predictions of the TTF model have previously been conducted in sports (Belling, Suss, \& Ward, 2015; Raab, 2012; Ward, Ericsson, Williams, \& Williams, 2013). Indeed, because of its naturally occurring dynamics (e.g., decisions to be made under time pressure; many potential alternative actions to be considered), sport is the ideal domain to test whether people use fast-and-frugal decision-making heuristics, such as TTF. These studies show that the performance of experienced handball, basketball, and soccer players is pretty accurately predicted by the TTF model: Players tended to generate
alternative options (e.g., shoot at the goal or pass to their teammate) in order of validity and selected as their final decision the first option they had generated.

As for adults, most decision-making research with children focused on information search (see Davidson, 1991, 1996; Gregan-Paxton \& Roedder John, 1995, 1997; Howse, Best, \& Stone, 2003; Ruggeri \& Katsikopoulos, 2013; Ruggeri, Olsson, \& Katsikopoulos, 2015) or investigated cue-based decision strategies (Horn, Ruggeri, \& Pachur, 2016; Mata, von Helversen, \& Rieskamp, 2011). However, to our knowledge, option generation in children has never been studied before.

## Time-limitation effects on option generation and decision making

According to the ecological rationality framework (Todd, Gigerenzer, \& ABC Research Group, 2012), no strategy is always optimal, because the efficiency of a strategy depends on the environmental structure. In this sense, people should be adaptive and modify their strategies depending on how effective they are in a given environment (de Oliveira, Lobinger, \& Raab, 2014). In many real-life situations, as in sports, decisions have to be taken under limited time, and we know that adults adapt to time limitation by using faster and simpler strategies (Ben Zur \& Breznitz, 1981; Payne et al., 1988). In particular, previous studies examining the effects of time limitation on decision-making processes have found that, under pressure, adults tend to increase their information processing speed (e.g., Ben Zur \& Breznitz, 1981; Payne et al., 1988) and use more non-compensatory strategies (e.g., Payne et al., 1988). On the same line, in a study with adult soccer players, Belling and colleagues (2015) found that time limitation reduced the number of task-relevant options generated, although it did not impact the quality of players' decisions.

What about the effects of time limitation on children's decision-making? We know that children are ecological learners, able to adapt their learning strategies to the characteristics (e.g., the statistical structure) of the task at hand (Horn et al., 2016; Nelson, Divjak, Gudmundsdottir, Martignon, \& Meder, 2014; Ruggeri \& Lombrozo, 2015), and they do so already by age four (Ruggeri, Sim, \& Xu, 2017). However, Davidson (1996) investigated the influence of time limitation on children's (7- to 10-year-olds) information search behavior and found that time pressure promoted faster, but generally not more selective searching. In this sense, it is unclear whether children would adapt their option generation and subsequent decision-making strategies depending on the time available.

## The present study

In the present study we use soccer as a testbed for a dynamic, real-life decision-making situation children have experience with. In particular, we extend previous research in two ways. First, we investigate for the first time children's (6- to 13-year-olds) option generation process, testing the predictions of the TTF model. In general, children have been shown to
use simple, non-compensatory information-search strategies (Davidson, 1991; Ruggeri \& Katsikopoulos, 2013), and specifically adolescent handball players have been shown to act according to TTF (Johnson \& Raab, 2003), we expect children to make use of the TTF heuristic. Moreover, in line with studies that investigated decision-making from a developmental perspective and showed an increase of selective, non-compensatory strategy use with age (Davidson, 1991, 1996), we further expect older children to rely more on the TTF heuristic than younger children.

Second, we explore whether and how time limitation influences the option generation and decision making of children. In particular, based on prior research, it is unclear whether children would adapt their option generation and decision-making strategies under time limitation.

## Method

## Participants

Ninety-seven children, all male, participated in this study ( $M_{\text {age }}=10.49$ years, $S D=1.98$ years; ranging from 6.67 to 13.50 years). All participants were recruited from a professional soccer academy in Germany. Prior to beginning the study, written informed parental consent, and local ethical review board approval of the study protocol, were obtained.

## Materials

We used 21 video scenes of live soccer match footage (three for the practice trials, 18 for the test trials). After a short display of buildup play, the scenes suddenly stopped right before the player in possession of the ball had to make a decision. The duration of the video scenes ranged between seven and eleven seconds, and video duration was unrelated to the study variables (all $p>.05$ ). We adopted the same task and materials as in Belling et al. (2015) with one difference: Instead of using an occlusion image that displayed field lines and the location of the ball on a blank white screen, we used real play footage that ended in a frozen frame such that all players are visible and the player with the ball needs to decide (see Figure 1). We chose to end the video with a frozen frame to provide participants with a constant, non-memory based game situation allowing the same condition during the entire option-generation test. Materials were presented to children on a touchpad (size: 8.9'').


Figure 1: Decision-making test procedure. a) The video stopped right before the player in possession of the ball had to make a decision and ended with a frozen frame. b) The option generation phase in which children generated options. c) Option selection phase in which children saw their generated options and subsequently selected the option they thought was the best.

## Design and Procedure

The task was administered collectively in groups of five to nine same-aged children. Within one age group, children
were randomly assigned to the testing sessions. Children were asked to sit at individual desks where a tablet was positioned. They were then introduced to the task procedure via a standardized instructional video (duration: 2:51 min) showing a person conducting the decision-making test for one exemplary soccer scene. The instructional video showed exactly which steps children were required to do on their tablets throughout the testing procedure.

The test proceeded as follows: After viewing each of the 21 videos (see above), videos stopped and held on with a frozen-frame, which gave the children time to generate a maximum of six options directly marking them onto the field via touch-pad (see Figure $1 \mathrm{a}, \mathrm{b}$ ). The first three videos were practice trials, used to familiarize participants with the test. During this familiarization phase children could ask clarifying questions to the experimenter. The other 18 video scenes were used as test trials and were randomly assigned to either the short-time ( 9 videos) or the long-time ( 9 videos) condition. In the long-time trials, children were given 30sec to generate options, whereas in the short-time trials participants were given 7.5 sec to generate. The order of presentation of the test videos was randomized, irrespectively of to which condition they were assigned. Afterwards, participants were asked to select among the options they had generated the one they thought was the best (see Figure 1 c ).

## Results

Results were analyzed with respect to developmental differences on four outcomes: (1) the mean number of options generated across all 18 tests; (2) quality of the generated options; (3) quality of the selected options; (4) participants’ dynamic inconsistency, which is the rate at which children selected as the best option the one they had generated at first. Dynamic inconsistency rates were computed as the relative frequency that the first option was not selected by players to be their final decision: Number of videos minus the frequency of the first generated option being the final decision divided by the total number of videos. Finally, we tested the effect of the time limitation manipulation on above-mentioned outcome variables.
To assess the quality of (2) the options generated and (3) option selected, two experienced youth soccer coaches independently generated options for the 18 test videos presented and rated the quality of each option they had generated on a 10-point scale (from 0, 'not at all good', to 10 'very good'). Overall, coaches generated a total of 104 options for the 18 test videos. That is, a total of 52 options were generated by each coach ( $M=2.89$ options per coach per video). Of the 52 options generated independently, 42 identical options were generated by both coaches indicating an $81 \%$ overlap between coaches. The quality of the options generated only by one coach or not generated by the coaches at all was scored with 0 'not at all good'. Based on the moderate inter-rater agreement for the quality of all options generated (ICC $=.52, p=.01$ ), quality scores for each option were calculated by averaging coaches' quality ratings.

## Option Generation, Decision Making, and Developmental Effects

Children generated a mean of 2.21 options ( $S D=0.65$ ). Overall, the mean quality of the first option children generated was $4.72(S D=0.99)$. We conducted a repeated measures MANOVA with the within-subject factor serial position of option (1-6) and quality as a dependent variable. This analysis revealed that children generated options in a meaningful way indicated by a significant decline of option quality across the serial position, Greenhouse-Geisser corrected $F(2.70,259.43)=859.56, p<.001, \eta_{p}{ }^{2}=.90$.

The mean quality of the option selected as the best was 4.45 ( $S D=1.06$ ). Most importantly, children selected the first generated option as the best one in $74 \%$ of the cases ( $\mathrm{SD}=$ 18.59). Compared to options generated at later serial positions, the first option generated was selected to be the final decision more frequently, $\chi^{2}(5)=4411.70, p<.001$. This was reflected in a dynamic inconsistency rate (i.e., the mismatch between the first option generated and final decision) of $0.26(S D=.19)$. This is a relatively low value, considering that a random selection would have resulted in a dynamic inconsistency rate of 0.55 , resulting from: $1-(1 /$ 2.21). The more options children generated, the higher was the dynamic inconsistency of their decisions, $r=.555, p<$ . 001 .

Separate linear regression analyses revealed that children's age was a significant positive predictor of all optiongeneration and most decision-making variables, except for dynamic inconsistency ( $R^{2}=.02, p=.138$ ). The older the children, the more options they generated ( $R^{2}=.06, p=.019$, $\beta=.24$ ), and the higher was the quality of the first option generated $\left(R^{2}=.19, p<.001, \beta=.44\right)$ as well as that of the option selected as the best $\left(R^{2}=.10, p=.002, \beta=.31\right)$.

## Time Limitation Effects

To explore whether and how time limitation influenced the option generation and decision making of children, we performed a multivariate analysis of variance (MANOVA) with one within subject factor time limitation (short-time vs. long-time condition) and four dependent variables (number of options generated, quality of the first option, quality of selected option and dynamic inconsistency). The repeatedmeasures MANOVA showed a significant multivariate effect, Wilks's Lambda $\lambda=.20, F(5,92)=34.50, p<.001$, $\eta_{\mathrm{p}}{ }^{2}=.62$. Follow-up univariate effects were further inspected for each decision-making variable separately.

In the short-time condition, as compared to the long-time condition, children generated fewer options, $F(1,96)=$ 127.51, $p<.001, \eta_{\mathrm{p}}{ }^{2}=.57$, generated first options with higher quality, $F(1,96)=15.19, p<.001, \eta_{\mathrm{p}}^{2}=.14$, and selected options of higher quality as their final, best decisions, $F(1$, $96)=16.55, p<.001, \eta_{\mathrm{p}}^{2}=.15$. With shorter time, dynamic inconsistency was less apparent than in the long-time condition, $F(1,96)=14.39, p<.001, \eta_{\mathrm{p}}^{2}=.13$.

Table 1: Effect of time limitation on the considered option generation and decision making variables.

|  | Short-time |  |  | Long-time |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  | condition |  |  | $\frac{\text { condition }}{M}$ |  |
|  | $M$ | $S D$ |  | $M$ | $S D$ |
| Total number of options | 1.84 | 0.56 | 2.59 | 0.86 |  |
| Quality of first option | 4.99 | 1.32 |  | 4.45 | 1.09 |
| Quality of final decision | 4.75 | 1.34 |  | 4.16 | 1.21 |
| Dynamic inconsistency | 0.22 | 0.20 | 0.29 | 0.21 |  |

Children generated their first option in a meaningful way: in the short-time condition, Greenhouse-Geisser corrected $F(2.35,225.77)=567.40, p<.001, \eta_{\mathrm{p}}^{2}=.86$, and in the longtime condition, $F(3.09,296.81)=489.89, p<.001, \eta_{p}^{2}=.84$, a significant decline of option quality across the serial position was apparent. Most importantly, children selected their first option to be the final decision in $71.1 \%(n=621)$ of the decisions in the long-time condition and significantly more often in $77.8 \%(n=679)$ of the decisions in the shorttime condition, $\chi^{2}(1, N=97)=11.60, p=.001, r=.16$.

In both conditions, the first option generated was selected to be the final decision more frequently, in the short-time, $\chi^{2}(5, N=97)=1982.61, p<.001, v=.67$, and in the longtime condition, $\chi^{2}(5, N=97)=2444.15, p<.001, v=.75$. For both time-limitation conditions, children's decision making was more dynamically inconsistent the more options they generated (short-time condition: $r=.448$, no time limitation: $r=.581$ ). Further, separate linear regression analyses were conducted for each time-limitation condition. Results revealed that the total number of options children generated predicted the degree of dynamic inconsistency in both conditions: The more options children generated in the shorttime, $\beta=.58, t(95)=6.95, p<.001, R^{2}=.33$, or in the longtime condition, $\beta=.45, t(95)=4.88, p<.001, R^{2}=.19$, the more dynamic inconsistent were their decisions.

We tested further whether age was differentially predictive when time is limited. In the short-time condition, children's age was a significant positive predictor of all optiongeneration and most decision-making variables, except for dynamic inconsistency ( $R^{2}=.01, p=.245, \beta=.12$ ). With time limitation, the older the children, the more options they generated $\left(R^{2}=.11, p=.001, \beta=.33\right)$ and the higher the quality of the first option generated ( $R^{2}=.13, p<.001, \beta=$ .36) as well as that of the option selected as best ( $R^{2}=.10, p$ $=.002, \beta=.31$ ). For the long-time condition, no age effect was found on the number of options generated ( $R^{2}=.02, p=$ $.153, \beta=.15$ ), the quality of the final option selected as best ( $R^{2}=.04, p=.057, \beta=.19$ ) and on dynamic inconsistency ( $R^{2}=.02, p=.130, \beta=.16$ ), but the older the children the higher the quality of the first option in the short-time condition ( $R^{2}=.13, p<.001, \beta=.36$ ).

## Discussion

Little, if anything, is known about how children generate options about which actions can be taken in real-life situations. To address this question, we tested the option
generation and decision making of children based on the predictions of the TTF heuristic. In addition, the influence of time limitation on option generation and decision making was explored. This allowed deepening our understanding of the adaptive relation between time limitation as one relevant environmental factor and the decision-making process predicted by TTF heuristic as argued from an ecological rationality perspective. In an experiment, children between the age of six and 13 years were tested in a video-based soccer decision-making task involving a within-subject timelimitation manipulation.

First, we investigated children's (6- to 13-year-olds) option generation process for the first time. Testing the TTF model revealed that predictions of the TTF heuristic also hold for children. In the option generation phase, as expected, children generated between two and three options, did so in a meaningful, non-random way and selected their first option as the final decision in more than $50 \%$ of the cases. Further, children's option generation influenced their decisions making: the more options children generated, the more dynamically inconsistent they decided. The pattern of results in children mainly matches option-generation and decisionmaking processes that have previously been demonstrated in adolescents and adults (Belling et al., 2015; Johnson \& Raab, 2003). The results are also consistent with findings showing that already schoolchildren use decision heuristics that match the task at hand (e.g., Horn et al., 2016).

Second, we explored whether and how time limitation influenced the option generation and decision making of children. Because of its naturally occurring dynamics, the sports domain is the ideal testbed to investigate situational, real life influence like time limitation. Our results revealed that time limitation affected all decision-making variables. In response to limited time, children generated fewer options, were less inconsistent in their decisions, generated higher quality first options and selected higher quality options as final decisions. This last result differs from what found with adult soccer players, whose quality of option generation and selection was not enhanced in response to time limitation (Belling et al., 2015). However, the positive effect of time limitation on children's option quality demonstrated in the present study theoretically matches predictions of the TTF heuristic and fits with the ecological rationality perspective (Johnson \& Raab, 2003; Todd et al., 2012). Compatible with the notion of "less-is-more", the time constraint prompted the generation of fewer but better options. Our results are also consistent with studies demonstrating that children are indeed ecological learners and speak for an adaptation of strategy use to the situation or task at hand (Horn et al., 2016; Ruggeri \& Lombrozo, 2015).
Finally, we found consistent developmental effects on both option generation and decision making: The number of options generated increased with age, but only in the shorttime condition. This short-time specific age effect hints at a developmental advantage for older children. With increasing age, children seem to adapt to time limitation by speeding up their generation to still produce a valid amount of options
they can choose from $\left(M(S D)_{\text {time }}\right.$ limitation $\left.=2.00(0.46)\right)$, whereas younger children do not $\left(M(S D)_{\text {time }}\right.$ limitation $=$ $1.67(0.60)$ ). In addition, older children seem to focus more on relevant, high-quality options early in the generation, irrespective of time limitation. This was indicated by the quality of the first option generated increasing with children's age irrespective of time limitation. In line with results showing that the information-search behavior of younger children (7- to 10-year-olds) was not more selective (Davidson, 1996), this study showed that selectivity for highquality information during generation seems to emerge later in childhood. For the quality of the final decision, as for the number of options, an increase with age was only apparent under limited time. This could be interpreted as a stronger adaption to time limitation by applying a strict, selective decision rule or applying it, according to the TTF heuristic, more accurately (Johnson \& Raab, 2003). Summing up, we showed that children adapted to the situational demands of time limitation by relying more on the simple TTF heuristic.

In conclusion, the present study shows that in familiar situations children tend to use simple, intuitive option generation and decision-making strategies. In particular, results support that TTF as a cognitive model can account for the option generation and decision making of children between the age of six and 13 years. Further, the study indicates that time limitation was an important situational factor impacting children's decision-making processes. Future studies should, therefore, explore other potentially relevant situational factors. Deepening our understanding of environmental or situational influences would also provide a concrete anchor for interventions targeting children's options and choices. Dynamic decision environments could, for example, be manipulated by the speed, distance or amount of stimuli provided. In particular, effects of situational factors on children's decision-making processes could be integrated into computer-based or real-life interventions and tested in a randomized control trial. Based on that, knowledge should be incorporated into prevention (e.g. traffic education) and training (e.g., sports, physical education) programs in a second step.

## Acknowledgments

We thank our colleagues in the Department of Performance Psychology for providing critical and constructive feedback on this project. We would also like to thank Sven Krüger, Silvana Höft, Elena Steen, Matthias Reubold, Lukas Hombach, Sinikka Heisler, Justin Klandermann, Taner Memis, Lukas Semmlinger, Estelle Schell, Matthias Rißmaier, and Benedikt Kosak for supporting data collection and coding.

## References

Belling, P. K., Suss, J., \& Ward, P. (2015). Advancing theory and application of cognitive research in sport: Using representative tasks to explain and predict skilled anticipation, decision-making, and option-generation behavior. Psychology of Sport and Exercise, 16, 45-59.
http://doi.org/10.1016/j.psychsport.2014.08.001
Davidson, D. (1991). Children's decision-making examined with an information-board procedure. Cognitive Development, 6(1), 77-90. http://doi.org/10.1016/0885-2014(91)90007-Z
Davidson, D. (1996). The effects of decision characteristics on children's selective search of predecisional information. Acta Psychologica, 92(3), 263-281. http://doi.org/10.1016/0001-6918(95)00014-3
de Oliveira, R. F., Lobinger, B. H., \& Raab, M. (2014). An adaptive toolbox approach to the route to expertise in sport. Frontiers in Psychology, 5, 709. http://doi.org/10.3389/fpsyg.2014.00709
Gigerenzer, G., \& Todd, P. (1999). Fast and frugal heuristics: The adaptive toolbox. Simple Heuristics That Make Us Retrieved from http://pubman.mpdl.mpg.de/pubman/item/escidoc:210 2905/component/escidoc:2102904/GG_Fast_1999.pdf
Horn, S. S., Ruggeri, A., \& Pachur, T. (2016). The development of adaptive decision making: Recognition-based inference in children and adolescents. Developmental Psychology, 52(9), 14701485. http://doi.org/10.1037/dev0000181

Johnson, J. G., \& Raab, M. (2003). Take The First: Optiongeneration and resulting choices. Organizational Behavior and Human Decision Processes, 91(2), 215229. http://doi.org/10.1016/S0749-5978(03)00027-X

Laborde, S., \& Raab, M. (2013). The tale of hearts and reason: the influence of mood on decision making. Journal of Sport \& Exercise Psychology, 35(4), 33957. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23966445
Mata, R., von Helversen, B., \& Rieskamp, J. (2011). When Easy Comes Hard: The Development of Adaptive Strategy Selection. Child Development, 82(2), 687700. http://doi.org/10.1111/j.1467-8624.2010.01535.x

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., \& Meder, B. (2014). Children ' s sequential information search is sensitive to environmental probabilities, 130, 74-80. http://doi.org/10.1016/j.cognition.2013.09.007
Payne, J. W., Bettmann, J. R., \& Johnson, E. J. (1988). Adaptive Strategy Selection in Decision Making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3), 534-552. http://doi.org/10.1037/0278-7393.14.3.534
Raab, M. (2012). Simple heuristics in sports. International Review of Sport and Exercise Psychology, 5(2), 104120. http://doi.org/10.1080/1750984X.2012.654810

Raab, M., \& Johnson, J. G. (2007). Expertise-based differences in search and option-generation strategies. Journal of Experimental Psychology: Applied, 13(3), 158-170. http://doi.org/10.1037/1076-898X.13.3.158
Ruggeri, A., \& Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. Cognition, 143, 203-216.
http://doi.org/10.1016/j.cognition.2015.07.004

Todd, P., Gigerenzer, G., \& ABC Research Group. (2012). Ecological rationality: Intelligence in the world. New York: Oxford University Press.
Ward, P., Ericsson, K. A., Williams, M. A., \& Williams, A. M. (2013). Complex Perceptual-Cognitive Expertise in a Simulated Task Environment. Journal of Cognitive Engineering and Decision Making, 7(3), 231-254. http://doi.org/10.1177/1555343412461254

# Multitasking Capability Versus Learning Efficiency in Neural Network Architectures 

Sebastian Musslick ${ }^{1, *}$, Andrew M. Saxe ${ }^{2}$, Kayhan Özcimder ${ }^{1,3}$, Biswadip Dey ${ }^{3}$, Greg Henselman ${ }^{4}$, and Jonathan D. Cohen ${ }^{1}$<br>${ }^{1}$ Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.<br>${ }^{2}$ Center for Brain Science, Harvard University, Cambridge, MA 02138, USA.<br>${ }^{3}$ Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA.<br>${ }^{4}$ Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.<br>*Corresponding Author: musslick@princeton.edu


#### Abstract

One of the most salient and well-recognized features of human goal-directed behavior is our limited ability to conduct multiple demanding tasks at once. Previous work has identified overlap between task processing pathways as a limiting factor for multitasking performance in neural architectures. This raises an important question: insofar as shared representation between tasks introduces the risk of cross-talk and thereby limitations in multitasking, why would the brain prefer shared task representations over separate representations across tasks? We seek to answer this question by introducing formal considerations and neural network simulations in which we contrast the multitasking limitations that shared task representations incur with their benefits for task learning. Our results suggest that neural network architectures face a fundamental tradeoff between learning efficiency and multitasking performance in environments with shared structure between tasks.


Keywords: multitasking; cognitive control; capacity constraint; learning; neural networks

## Introduction

Our limited capability to execute multiple tasks at the same time highlights one of the most fundamental puzzles concerning human processing, which must be addressed by any general theory of cognition (Shenhav, Botvinick, \& Cohen, 2013; Kurzban, Duckworth, Kable, \& Myers, 2013; Anderson, 2013): Why, for some tasks, is the human mind capable of a remarkable degree of parallelism (e.g., navigating a crowded sidewalk while talking to a friend), while for others its capacity for parallelism is radically limited (e.g., conduct mental arithmetic while constructing a grocery list)?

Early theories of cognition, that have continued to be highly influential, assert that the ability to multitask - that is, to carry out a set of tasks concurrently ${ }^{1}$ - can be understood in terms of a fundamental distinction between automatic and controlled processing, with the former relying on parallel processing mechanisms (that can support multitasking) and the latter assumed to rely on a serial processing mechanism with limited capacity (Posner \& Snyder, 1975; Shiffrin \& Schneider, 1977) that can only support processing of a single task at

[^145]a time. In this view, the constraints on the number of controldependent tasks that can be executed at one time reflect an intrinsic property of the control system itself. However, alternative ("multiple-resource") accounts (Allport, 1980; Meyer \& Kieras, 1997; Navon \& Gopher, 1979; Salvucci \& Taatgen, 2008) have suggested that multitasking limitations arise from local processing bottlenecks. That is, if two tasks share the same local resources (i.e. representations required to perform the tasks), then executing them simultaneously can lead to cross-talk and degraded performance. It has been argued that the very purpose of cognitive control is to prevent such cross-talk by limiting the number of active task processes engaged (Cohen, Dunbar, \& McClelland, 1990; Botvinick, Braver, Barch, Carter, \& Cohen, 2001). In this view, constraints in multitasking reflect the consequences of control doing its job, rather than limitations intrinsic to the mechanisms of control itself. This line of argument suggests that, to better understand the conditions under which multitasking is and is not possible, it is necessary to understand the extent to which the task processes involved share representations, and are thus subject to potential interference and the intervention of control to limit processing. This, in turn, raises the question of whether there are general principles of neural architectures that determine the use of shared representation, and how these interact with learning and processing.

One may argue that the constraints that shared representations impose on multitasking are negligibly small in a processing system as large as the human brain. However, simulation studies (Feng, Schwemmer, Gershman, \& Cohen, 2014), followed by analytic work (Musslick et al., 2016) have found that the multitasking capability of a network can drop precipitously as a function of overlap between task processes (i.e. number of shared representations), and that this effect is relatively insensitive to the size of the network.

The findings above suggest that maximal parallel processing performance is achieved through the segregation of task pathways, by separating the representations on which they rely. This raises an important question: insofar as shared representation introduces the risk of cross-talk and thereby limitations in parallel processing performance, why would the brain prefer shared task representations over separate ones? Insights gained from the study of learning and representation in neural networks provide a direct answer to this question:

Shared representations across tasks can support inference and generalization (Caruana, 1997). These benefits are strongly linked to the ability of neural networks to carry out "interactive parallelism", that is, the ability to learn and to process complex representations by simultaneously taking into account a large number of interrelated and interacting constraints (McClelland, Rumelhart, \& Hinton, 1986).

In this study, we examine the tension between interactive parallelism that promotes learning efficiency through use of shared representations, on the one hand, and "independent parallelism" (i.e. the ability to carry out multiple processes independently), on the other hand. That is, we are interested in studying biases that promote shared representations over multitasking performance. We first demonstrate that the wellrecognized (and valued) emergence of shared representations (Hinton, 1986) in response to extrinsic biases (i.e. shared structure in the task environment) leads to constraints in multitasking performance. In the second part, we introduce a formal characterization of a tradeoff between learning efficiency and multitasking performance and examine how intrinsic biases of the network toward the use of shared representations can expose this tradeoff in neural network simulations. The source code for all simulations is available at github.com/musslick/CogSci-2017.

## Neural Network Model

For the simulations described in the paper we focus on a network architecture that has been used to simulate a wide array of empirical findings concerning human performance (e.g. Cohen et al., 1990; Botvinick et al., 2001), including recent work on limitations in multitasking (Musslick et al., 2016). In this section we lay out the architecture of this network, its processing, as well as the task environments used to train it.

## Network Architecture and Processing

The network consists of two input layers, one of which represents the stimulus presented to the network and another that encodes the task that the network is instructed to perform on the stimulus. Stimulus input features can take any real value between 0 and 1 and can be grouped into stimulus dimensions that are relevant for a particular task. The network is instructed to perform a single task by clamping the corresponding task unit in the task layer to 1 while all other task units are set to 0 . These stimulus and task input values are multiplied by a matrix of connection weights from the respective input layer to a shared associative layer, and then passed through a logistic function to determine the pattern of activity over the units in the associative layer. This pattern is then used (together with a set of direct projections from the task layer) to determine the pattern of activity over the output layer. The latter provides a response pattern that is evaluated by computing its mean squared error (MSE) with respect to the correct (task-determined) output pattern. Similar to stimulus features, output units can be grouped into response dimensions that are relevant for a particular task. Note that the
weight projections from each task unit can act as control signals that bias processing towards task-relevant stimulus information represented at the associative and output layer.

In order to represent the task environment described below, the stimulus layer is compromised of 45 input units (features) and the task layer of nine task units. The output layer consists of 15 units and is organized into three response dimensions (with five units per response dimension.). The number of units in the associative layer is set to 100 .


Figure 1: Feedforward neural network used in simulations. The input layer is composed of stimulus vector $\overrightarrow{x_{s}}$ and task vector $\overrightarrow{x_{t}}$. The activity of each element in the associative layer $y_{h} \in \overrightarrow{y_{h}}$ is determined by all elements $x_{s}$ and $x_{t}$ and their respective weights $w_{h s}$ and $w_{h t}$ to $y_{h}$. Similarly, the activity of each output unit $y_{o} \in \overrightarrow{y_{o}}$ is determined by all elements $y_{h}$ and $x_{t}$ and their respective weights $w_{o h}$ and $w_{o t}$ to $y_{o}$. A bias of $\theta=-2$ is added to the net input of all units $y_{h}$ and $y_{o}$. Blue shades in the input and output units (circles) correspond to unit values of $>0$ and illustrate an example input pattern with its respective output pattern: The second task requires the network to map the vector of values in the first five stimulus input units to one out of five output units (yellow shade).

## Task Environment

Each task is defined as a mapping between a subspace of five stimulus features (referred to as a task-relevant stimulus dimension) onto five output units of a task-specific response dimension, so that only one of the five relevant output units is permitted to be active (see Fig. 1). The value of each stimulus feature is drawn from a uniform distribution $U[0,1]$. The rule by which 5 relevant stimulus features of any task-relevant stimulus dimension are mapped onto one of the 5 output units of the task-relevant response dimension corresponds to a nonlinear function that was randomly generated ${ }^{2}$ with a separate "teacher" network (cf. Seung, Sompolinsky, \& Tishby, 1992), and is the same across tasks. However, tasks are considered to be independent in that they differ which stimulus dimension is linked to which response dimension.

The task environment across all simulations encompasses nine tasks. As illustrated in Fig. 2 groups of three tasks map

[^146]onto the same response dimension. However, similarity between tasks could be varied by manipulating the overlap between their relevant stimulus dimensions. At the extremes, task environments can be generated such that tasks of different response dimensions relate to separate stimulus features (no feature overlap, Fig. 2a), or the same stimulus features (full feature overlap, e.g. tasks 1-3 in Fig. 2b).


Figure 2: Task environments. For each task, the network was trained to map a subset of 5 stimulus features onto a subset of 5 output units. At the extremes tasks that were mapped onto different response dimensions (e.g. tasks 1-3) could either (a) rely on separate stimulus features or (b) completely overlap in terms of their relevant stimulus features.

Networks are initialized with a set of small random weights and then trained on all tasks using the backpropagation algorithm ${ }^{3}$ (Rumelhart \& Geoffrey E. Hinton, 1986) to produce the task-specified response for each stimulus.

## Multitasking Limitations Due to Shared Structure in the Task Environment

A key feature of neural networks is their ability to discover latent structure in the task environment, exploiting similarity between stimulus features in the form of shared representations (Hinton, 1986; Saxe, McClelland, \& Ganguli, 2013). In this section we explore how the emergence of shared representations as a function of structural similarities between tasks can impact the multitasking performance of a network.

## Simulation Experiment 1: Shared Task Representations as a Function of Feature Overlap

In order to investigate the effect of structural similarities between tasks we generated task environments with varying overlap between task-relevant stimulus features. We define feature overlap as the number of relevant stimulus features that are shared between any pair of tasks linked to different response dimensions (see Fig. 3a). That is, two tasks involving two different response dimensions could either share

[^147]no relevant stimulus features (cf. Fig. 2a), all five stimulus features (cf. Fig. 2b) or any whole number of features in between, resulting in 6 different task environments. We trained 100 networks in each of the environments. The networks were trained on all nine tasks with the same set of 50 stimulus samples until the network achieved an MSE of 0.01.


Figure 3: Effects of task similarity. (a) Networks were trained in task environments with varying degrees of feature overlap. Yellow and red shades highlight task-relevant stimulus features for two tasks involving different response dimensions. (b) Final multitasking accuracy of the network as a function of the learned similarity between tasks involving different response dimensions. Colors indicate the degree of feature overlap present in the task environment as illustrated in (a).

In order to assess the similarity of learned task representations we focus our analysis on the weights from the task units to the associative layer, insofar as these reflect the computations carried out by the network required to perform each task. For a given pair of tasks we compute the learned representational similarity between them as the Pearson correlation of their weight vectors to the associative layer.

We measured multitasking performance for pairs of tasks (of different stimulus and response dimensions) by activating two task units at the same time and evaluating the concurrent processing performance in the response dimensions relevant to the two tasks. The accuracy of a single task $A_{\text {Single }}$ can be computed as

$$
\begin{equation*}
A_{\text {single }}=\frac{a_{c}}{\sum_{i=1}^{5} a_{i}} \tag{1}
\end{equation*}
$$

where $a_{i}$ is the activation of the $i$ th output unit of the taskrelevant response dimension and $a_{c}$ is the activation of the correct output unit. The multitasking accuracy is simply the mean accuracy of both engaged single tasks.

The simulation results confirm well-known explorations in neural networks (Hinton, 1986; McClelland \& Rogers, 2003; Saxe et al., 2013) that task similarities in the environment can translate into similarities between learned task representations. Critically, this extrinsic bias toward the learning of shared representations negatively affected multitasking performance (Fig. 3b). To illustrate this, consider the simultaneous execution of tasks 1 and 5 in an environment as depicted in Fig. 2b. If the network learns similar representations at the associative layer for tasks 1 and 2 (note that both tasks rely on the same stimulus features), then executing task 1 will implicitly engage the representation of task 2 which in turn causes interference via its link to the response dimension of task 5.

## Multitasking Limitations due to Intrinsic Learning Biases

In addition to environmental biases that shape the learning of shared task representations there may be factors intrinsic to the neural system that can regulate the degree to which such representations are exploited in learning. In this section we introduce a formal analysis of how such biases can affect the tradeoff between learning efficiency and multitasking performance. We then use weight initialization as a learning bias in simulations to establish a causal relationship between the use of shared representations on the one hand, and resulting effects on learning and multitasking, on the other hand.

## Formal Intuitions on the Tradeoff between Learning Efficiency and Multitasking Capability

To gain formal intuition into the tradeoff between multitasking ability and learning speed, we consider a stripped-down version of the introduced network model that is amenable to analysis. In the full model, nonlinear interactions between the task units and the stimulus units occur in the associative layer. Here we assume a gating model in which these nonlinear interactions are carried out through gating signals that can zero out parts of the activity in the associative and output layers, or pass it through unchanged. The choice of which parts of each layer are gated through on each input is left to the designer (not learned, as in the full model).

We study the scheme depicted in Fig. 4 consisting of $M$ input and response dimensions with full feature overlap (cf. Fig. 2b). For the output layer, we assume that the gating variables automatically zero all but the task-relevant response dimensions. For the associative layer, we separate the hidden units into dimensions, one for each input dimension, and make the gating variables zero all representations except the one coming from the task-relevant input dimension (Fig. 4a).

Crucially, when the gating structure is known on a specific example, the output of the network is a linear function of the neurons that are on. Given this setting, the learning dynamics can be solved exactly using methods developed by Saxe, McClelland, and Ganguli (2014). The key advantage afforded by the gating scheme is depicted in Fig. 4a: the input-to-hidden weights for one input dimension can be shared by all tasks that rely on that input dimension. This leads to a factor $\sqrt{M}$ speedup in learning relative to learning a single task by itself (proof omitted due to space constraints).

However, with this gating system, multitasking is not possible: gating another task through to the output will lead to interference. To counteract this, the gating scheme must be changed: response dimensions can be divided into $Q$ groups, each with a dedicated set of hidden units (Fig. 4b). This allows tasks that use response dimensions in different output groups to be performed simultaneously. Hence a maximum of $Q$ tasks can be performed simultaneously, but weight sharing is reduced across tasks by a factor $Q$, slowing learning.

This analysis provides, at least in a simplified system, a quantitative expression of the fundamental tradeoff between


Figure 4: Gating model used for formal analysis. (a) Task information directly switches on or off task-relevant dimensions in the output and associative layers. This allows input-to-hidden weights to be shared across the $M$ different tasks corresponding to different response dimensions, increasing learning speed by a factor $\sqrt{M}$. However, two tasks that rely on different input dimensions cannot be multitasked due to crosstalk at the output (convergent red and green arrows). (b) Multitasking ability can be improved by separating response dimensions into $Q$ groups, each with a dedicated set of units in the associative layer. Gating now permits one task from each group to operate concurrently (red and green arrows no longer converge). However, weight sharing is limited to the group, yielding a learning speed of $\sqrt{M / Q}$.
learning speed and multitasking ability. Let $t$ be the number of iterations required to learn all tasks, $Q$ the maximum number of concurrently executable tasks, and $M$ the number of input/response response dimensions. Then

$$
\begin{equation*}
t^{2} \propto Q / M \tag{2}
\end{equation*}
$$

where the proportionality constant is related to the statistical strength of the input-output association for one task, the learning rate, and the error cut-off used to decide when learning is complete (Saxe, Musslick, \& Cohen, 2017).

Due to the tradeoff in Eqn. (2), gating schemes that share more structure will learn more quickly. Hence generic, randomly initialized nonlinear networks will tend to favor shared representations, as shown in Simulation Experiment 1.

## Simulation Experiment 2: Effects of Learning Biases for Shared Representations

In Simulation 2 we focus on a bias intrinsic to the neural system, i.e. the initialization of the weights from the task layer. We use this factor to systematically examine how the use of shared representations facilitates the discovery of similarity structure while diminishing multitasking performance. To do so, we focus initially on a training environment in which tasks are maximally similar, as this is the condition in which there is most opportunity for exploiting shared representations. We then examine environments with $80 \%$ and $0 \%$ feature overlap between tasks, to test the generality of the observed effects.

To manipulate the bias towards shared task representations, we initialized the weights from the task units to the associative layer, varying the similarity among the weight vectors across tasks with the rationale that greater similarity should produce a greater bias toward the use of shared representations in the associative layer. Weight vectors for tasks relying on the same stimulus input dimensions were randomly initialized to yield a correlation coefficient of value $r$. The correlation value $r$ was varied from 0 to 0.975 in steps of 0.025 and was used to constrain initial weight similarities for 100 simulated networks per initial condition. The weight vectors for tasks of non-overlapping stimulus dimensions were uncorrelated. Finally, all task weights to the associative layer were scaled by a factor of 5 to enhance the effects of different initial task similarities. The networks were trained using the same parameters as reported for Simulation Experiment 1.

Simulation results indicate that networks with a higher similarity bias tend to develop more similar representations at the associative layer for those tasks (in terms of their final weight vector correlations), whereas a lower similarity bias leads to more distinct task representations at this layer. In environments with high feature overlap between tasks, stronger initial biases toward shared representations lead to increased learning speed (i.e. less iterations required to train the network), as similarities between tasks can be exploited (Fig. 5a). Critically, this comes at the cost of multitasking performance. Learning benefits gained from shared representations are less prevalent in environments with less feature overlap between tasks. However, effects of weight similarity biases on multitasking impairments remain (Fig. 5b).


Figure 5: Effects of weight similarity bias. Mean multitasking accuracy (for two tasks simultaneously) plotted against the mean number of iterations required to train the network. Data points represent the mean measures across networks initialized with the same task similarity (constrained by task weight vector correlation) for tasks relying on the same stimulus dimensions. Effects are shown for (a) environments with $100 \%$ feature overlap between tasks, as well as (b) across environments with different feature overlap. Different data point clusters correspond to different training environments.

## General Discussion and Conclusion

The limited ability to perform multiple control-dependent tasks at the same time is one of the most salient characteristics of human cognition, and is universally considered a defining feature of cognitive control. Despite these facts, the
sources of this capacity constraint associated with control remain largely unexplored. Here, we build upon the observation that multitasking limitations can arise from shared representations between tasks (Feng et al., 2014; Musslick et al., 2016), and use a combination of formal analysis and neural network simulations to examine biases towards shared representations that incur such costs in multitasking.

In the first part of this study, we build upon early insights of connectionism that shared representations emerge as a function of task similarities in the environment and demonstrate the deleterious consequences for multitasking performance. It has been shown that networks are capable of extracting similarities from a hierarchically structured input space (Hinton, 1986). Recent analytic and empirical work in the domain of semantic cognition paints a similar picture: neural systems may gradually discover shared structure in the task environment with a bias towards the initial formation of shared, lowdimensional representations (Saxe et al., 2013; McClelland \& Rogers, 2003). Our simulation results are in line with these observations showing that shared task representations emerge as a function of high stimulus feature overlap between tasks and furthered the insight that such similarities in the task environment lead to multitasking limitations.

In the second part, we examined how intrinsic learning biases towards shared or separate representations (by means of weight initialization) can be used to expose a tradeoff between learning efficiency and multitasking performance. Early work in machine learning suggests that learning biases towards a particular representation can be understood as biases of the learner's hypothesis space (Baxter, 1995), that is, the set of all hypotheses a learner may use to acquire new tasks. We formalized this hypothesis space in terms of the amount of shared representations between tasks and showed how this mediates an inverse relationship between learning efficiency and interference-free multitasking. Our neural network simulations confirmed these analytical predictions, showing that a weight initialization bias towards shared representations enables faster learning if shared structure in the environment can be exploited, but incurs a cost for multitasking. A promising direction for future research may be to explore another prediction: our formalism suggests a role for such biases in regularizing the representational complexity of the network, thereby promoting generalization performance.

Our analyses indicate that neural learning systems, whether natural or artificial, are subject to a tension between "interactive parallelism" on the one hand, which exploits the fine grained structure of representations and similarity in the service of learning, and "independent parallelism" that supports concurrent processing of distinct tasks, on the other hand. A similar tension can be found in the domain of learning and memory. The complementary learning systems hypothesis proposes two separate learning systems, one system that relies on shared representations to support inference, as well as another system that uses separate representations to support independent encoding and retrieval of information
(McClelland, McNaughton, \& O'Reilly, 1995). The latter system supports a form of independent parallelism for associational processes that is similar to the form of independent parallelism for executional processes described in this paper.

Altogether our results suggest that the brain may be confronted with balancing multitasking capability against extrinsic and intrinsic biases towards shared representations. A major goal for the development of artificial systems may be to systematically configure the balance between interactive and independent parallelism, as well as to exploit the relative advantages of each. Most efforts in complex neural architectures have focused predominantly on the discovery of shared representations for the purpose of inference and generalization (Bengio, Courville, \& Vincent, 2013). However, one of the future challenges will be to explore the tension between learning efficiency and multitasking in networks with higher complexity (i.e. deep networks), as well as in more naturalistic task environments. We hope that this work will help inspire a proliferation of efforts to further explore this area.

## References

Allport, D. A. (1980). Attention and performance. Cognitive psychology: New directions, 1, 12-153.
Anderson, J. R. (2013). The architecture of cognition. Psychology Press.
Baxter, J. (1995). Learning internal representations. In Proceedings of the eighth annual conference on computational learning theory (pp. 311-320).
Bengio, Y., Courville, A., \& Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.
Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., \& Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychological Review, 108(3), 624.
Caruana, R. (1997). Multitask learning. Machine learning, 28(1), 41-75.
Cohen, J. D., Dunbar, K., \& McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. Psychological Review, 97(3), 332-361.
Feng, S. F., Schwemmer, M., Gershman, S. J., \& Cohen, J. D. (2014). Multitasking vs. multiplexing: toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. Cognitive, Affective, \& Behavioral Neuroscience, 14(1), 129-146.
Hinton, G. E. (1986). Learning distributed representations of concepts. In Proceedings of the 8th confererence of the Cognitive Science Society (pp. 1-12). Hillsdale, NJ: Lawrence Erlbaum Associates.
Kurzban, R., Duckworth, A., Kable, J. W., \& Myers, J. (2013). An opportunity cost model of subjective effort and task performance. The Behavioral and Brain Sciences, 36(6), 661-679.
McClelland, J. L., McNaughton, B. L., \& O'Reilly, R. C. (1995). Why there are complementary learning systems
in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102(3), 419.
McClelland, J. L., \& Rogers, T. T. (2003, April). The parallel distributed processing approach to semantic cognition. Nature reviews. Neuroscience, 4(4), 310-322.
McClelland, J. L., Rumelhart, D. E., \& Hinton, G. E. (1986). The appeal of parallel distributed processing. Cambridge, MA: MIT Press.
Meyer, D. E., \& Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. Psychological review, 104(1), 3.
Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., \& Cohen, J. D. (2016). Controlled vs. automatic processing: A graph-theoretic approach to the analysis of serial vs. parallel processing in neural network architectures. In Proceedings of the 38th annual conference of the Cognitive Science Society (pp. 1547-1552). Philadelphia, PA.
Navon, D., \& Gopher, D. (1979). On the economy of the human-processing system. Psychological Review, 86(3), 214.

Posner, M., \& Snyder, C. (1975). attention and cognitive control,. In Information processing and cognition: The loyola symposium (pp. 55-85).
Rumelhart, D. E., \& Hinton, G. E. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536.
Salvucci, D. D., \& Taatgen, N. A. (2008). Threaded cognition: an integrated theory of concurrent multitasking. Psychological Review, 115(1), 101.
Saxe, A. M., McClelland, J. L., \& Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks. In Proceedings of the 35th annual meeting of the cognitive science society (pp. 1271-1276).
Saxe, A. M., McClelland, J. L., \& Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Y. Bengio \& Y. LeCun (Eds.), International conference on learning representations. Banff, Canada.
Saxe, A. M., Musslick, S., \& Cohen, J. D. (2017). A formal tradeoff between learning speed and multitasking ability in a simple neural network. http://www.people .fas.harvard.edu/~asaxe/multitasking.html. (Retrieved May 13, 2017)
Seung, H., Sompolinsky, H., \& Tishby, N. (1992). Statistical mechanics of learning from examples. Physical Review A, 45(8), 6056.
Shenhav, A., Botvinick, M. M., \& Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron, 79(2), 217-240.
Shiffrin, R. M., \& Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review, 84(2), 127.

# Analogical Inferences in Causal Systems 

Matthew Myers (matthewmyers2020@u.northwestern.edu)<br>Department of Psychology, 2029 Sheridan Road<br>Evanston, IL 60208 USA

Dedre Gentner (gentner@northwestern.edu)<br>Department of Psychology, 2029 Sheridan Road<br>Evanston, IL 60208 USA


#### Abstract

Analogical and causal reasoning theories both seek to explain patterns of inductive inference. Researchers have claimed that reasoning scenarios incorporating aspects of both analogical comparison and causal thinking necessitate a new model of inductive inference (Holyoak, Lee, \& Lu, 2010; Lee \& Holyoak, 2008). This paper takes an opposing position, arguing that features of analogical models make correct claims about inference patterns found among causal analogies, including analogies with both generative and preventative relations. Experiment 1 demonstrates that analogical inferences for these kinds of causal systems can be explained by alignment of relational structure, including higher-order relations. Experiment 2 further demonstrates that inferences strengthened by matching higher-order relations are not guided by the transfer of probabilistic information about a cause from base to target. We conclude that causal analogies behave like analogies in gen-eral-analogical mapping provides candidate inferences which can then be reasoned about in the target.


Keywords: analogy; causality; structure mapping theory; inductive inferences

## Introduction

The current paper challenges recent claims that standard theories of analogy (such as structure-mapping theory; Gentner, 1983, 1989) cannot explain analogical inferences that incorporate causal relations. Holyoak and colleagues (Holyoak et al., 2010; Lee \& Holyoak, 2008) contend that causal analogies require a different kind of process from typical analogies. Specifically, they claim that structure-mapping theory (SMT) -and more broadly, all extant models of analogy—fail to predict people's inferences for causal analogies that involve both generative and preventative causal relations. In their view, causal analogies require models of analogy that incorporate the basic elements of causal models (Lee \& Holyoak, 2008).

We maintain that in causal analogies, the mapping between analogs is done by the same structure-mapping processes as in other domains. Assuming that the mapping yields candidate inferences in the target, normal causal reasoning processes then occur in the target to arrive at further conclusions. Specifically, we show that this division of labor holds for the kinds of materials used by Holyoak and colleagues (Holyoak et al., 2010; Lee \& Holyoak, 2008): analogical processes inform the construction of causal models in the target analog, after which causal reasoning processes are used to draw further inferences in the target. We believe our account provides a better explanation of people's reasoning at the level of representation and, more broadly, offers a more parsimonious
description of analogical reasoning.

## Inference and Similarity

Similarity plays an important role in SMT. While simple physical or property-based similarities can serve as cues to engage in analogical comparison, relational matches are more central to the content of analogical inferences (Gentner \& Markman, 1997). Relational similarity is assessed by a process of structural alignment in which components of the two analogs are placed in correspondence based on a maximal (or near-maximal) match in relational structure. Alignments with deeply embedded relational structures-in which higher-order constraining relations govern lower-order rela-tions-are perceived as more similar than those with shallow structures (Gentner, Rattermann, \& Forbus, 1993) and provide a better basis for candidate inferences to the target (Clement \& Gentner, 1991). Thus, the perceived structural similarity and inferential strength between two analogs typically exhibit a positive correlation.


Targets


Figure 1: The causal systems in Lee \& Holyoak (2008). G and $P$ represent generative and preventative causes. The effect is the outcome feature. Dotted elements in the targets represent information not given. In descending order, similarity ratings between base and target was $G_{1} G_{2} P_{1}, G_{1} G_{2}$, and $G_{1} P_{1}$. The order of inductive strength ratings was $G_{1} G_{2}$, $G_{1} G_{2} P_{1}$, and $G_{1} P_{1}$.

However, Lee and Holyoak (2008) found that these mea-
sures can be disassociated. In Experiment 1, they presented participants with a description of a fictional animal with four notable features. Three of the four features were described as causally related to the fourth: two were generative causes while the other was a preventative cause ( $G_{1}, G_{2}$, and $P_{1}$ ). This animal served as the base analog (see Figure 1). Participants were further given a description of a secondary target animal that possessed one of three combinations of the antecedent features described in the base (i.e., $G_{1} G_{2} P_{1}, G_{1} G_{2}$, and $G_{1} P_{1}$ ); however, they were given no information about how those features affected the outcome. The base animal (Animal A) and target animal (Animal B) were described in the following manner:

Animal A has dry flaky skin, muscular forearms, a weak immune system, and blocked oil glands.
For Animal A, dry flaky skin, tends to PRODUCE blocked oil glands; muscular forearms tend to PRODUCE blocked oil glands; a weak immune system tend to PREVENT blocked oil glands.

Animal B has dry flaky skin, muscular forearms, and a weak immune system.

Participants' task was to either rate the similarity of the base and target, or to estimate the likelihood that the effect would occur in the target. Lee and Holyoak reported that while participants' similarity ratings roughly corresponded with the number of structural relations shared by base and target ( $G_{1} G_{2} P_{1}$ was highest), their inferences did not. In descending order, the observed strength of the effect inference in the target was $G_{1} G_{2}, G_{1} G_{2} P_{1}$, and $G_{1} P_{1}$. The authors contend that these results are problematic for SMT and other models of analogy, for two reasons. First, SMT cannot account for the systematic non-correspondence observed between similarity and inference strength; and second, SMT cannot account for the transfer of probabilistic information from the base because it does not permit the transfer of nonrelational properties of the higher-order relations ${ }^{1}$.

In response to the first issue, Colhoun and Gentner (2009) note that the measure of inductive strength used in the previous experiment focused on a single variable: people's belief about the likelihood of the effect. This ignores a large number of other inferences that must be made between the analogs. Participants are told only that certain factors (e.g., $G_{1}$, dry flaky skin) exist in the target, but they are not told that these factors are causally connected to the effect $E$; these causal links must all be inferred from the base.

[^148]Arguing that the inductive strength of an analogy should be measured by all of its candidate inferences, and not solely a single effect inference, Colhoun and Gentner (2009) used Lee and Holyoak (2008) original stimuli and asked participants to rate their confidence in each of the required inferences in the target ( $G_{1}$ tends to produce $E$, etc.). The result was that the ordering of inductive strength ratings for inferences within the target closely matched the ordering of perceived similarity. In sum, this experiment showed that when the appropriate inferential questions are asked, there is no conflict between perceived similarity and perceived inferential strength.

## Inference and Structure

The second claim against SMT is that it would require the transfer of non-relational properties of higher-order relations, such as propensities for causal antecedents to produce or prevent an effect (as in $G_{1}$ tends to produce $E$ ). Lee and Holyoak (2008) claim that avoiding such a violation of systematicity would require a model to incorporate a kind of mapping process in which degrees of belief in the inferred property of a target are mutually informed by both analogical mapping procedures and causal strength assessments.

However, Colhoun and Gentner (2009) proposed an alternative solution. They suggest that the pattern of relations in the base allows people to infer that the pattern of $G_{1}, G_{2}$ and $P_{1}$ is sufficient to produce $E$. Specifically, participants are told that the base contains $G_{1}, G_{2}$, and $P_{1}$, that $G_{1}$ and $G_{2}$ both tend to produce $E$, that $P_{1}$ tends to prevent $E$, and that $E$ is present. Thus in encoding the base, the presence of the effect $E$ allows people to infer that the combined causal strength of $G_{1}$ and $G_{2}$ exceeds that of $P_{1}$. To test this, in Experiment 2, they presented participants with a base analog consisting of generative and preventive causes ( $G_{1}$ tends to produce $E$, etc.), but varied whether the effect $E$ was actually stated to be present in the base. Participants' task was to generate effect inferences for different targets. The result was that participants gave stronger effect inferences in the target when given a base analog in which the effect was clearly stated to be present than when given a base in which the effect might or might not be present.

Colhoun and Gentner (2009) argue that the presence of an effect in the base provides evidence of the relative strength of the antecedent causes. In other words, people are attending to and transferring a higher-order qualitative relation-that $G_{1}$ and $G_{2}$ are causally stronger than $P_{1}$. This suggests that, as in other areas of analogical reasoning, encoding processes occur in the base, followed by structure-mapping processes that align the base relational structure with that of the target and project candidate inferences to the target. Once these new inferences are projected, causal reasoning processes in the target can produce further inferences.

But there is an alternative account. Perhaps, consistent with the idea that causal propensities are intertwined with the analogical mapping process (Holyoak et al., 2010), the differences in observed inference ratings might simply have been the consequence of a probability calculation computed
in the base and subsequently mapped to a matching target. For example, people may have derived a series of probabilistic propensities for the causal antecedents to bring about the effect. Participants would then project those probabilistic properties of the higher-order predicate relations onto the target (e.g., predicates related by CAUSE- $10 \%$, CAUSE- $50 \%$, CAUSE- $93 \%$, etc.). We explore this possibility in the following two studies.


Figure 2: Base and target systems for Experiment 1 (Panel A) and Experiment 2 (Panel B). $G$ and $P$ represent generative and preventative causes. The experimental manipulations for Experiment 1 and 2 were the bases and targets, respectively.

## Experiment 1

Hypothetically, there are two ways people may calculate probabilistic likelihoods for an effect in a base analog: a global posterior calculation, in which all causal antecedents are factored into a probabilistic estimate; and a local posterior calculation, in which only those antecedent causes that occur in the target are analyzed in the base. Experiment 1 examines a scenario where predictions for global or local posterior calculations are pitted against predictions made by standard mapping theory. Figure 2 A illustrates the scenario used in this experiment. In this study, we varied the causal structure in the base, keeping the target $\left(G_{1} P_{1}\right)$ constant in both conditions. Further, in both conditions, participants are told that the effect $E$ occurs in the base. If participants generate a global posterior calculation for the effect in either the $G_{1} G_{2} P_{1}$ condition [i.e., $P\left(\right.$ Effect $\left.\mid G_{1} G_{2} P_{1}\right)$ ] or the $G_{1} P_{1} P_{2}$ condition $\left[P\left(\right.\right.$ Effect $\left.\left.\mid G_{1} P_{1} P_{2}\right)\right]$, then we should see no difference in inference ratings between conditions. Since people's probability estimates are calculated conditionally on the aggregate influence of all causal antecedents, there is no unitary piece of information that can inform them as to how likely the effect will be when transferred to a $G_{1} P_{1}$ target $^{2}$.

[^149]But suppose instead that participants use the base to generate a local posterior calculation (considering only the relations that match with those in the target)-e.g., $P$ (Effect $\left.G_{1} P_{1}\right)$. In this case, they have no basis for a difference in inference ratings between $G_{1} G_{2} P_{1}$ and $G_{1} P_{1} P_{2}$. Because the strength of the individual causal antecedents is unknown, they would have no information about the degree to which the unmapped cause in the base either prevents $\left(P_{2}\right)$ or contributes $\left(G_{2}\right)$ to the effect. Therefore there is no reason to expect that a systematic difference in effect strength estimates between the two conditions.

In contrast, suppose participants utilize the type of encoding and structure-mapping techniques as described above. In this case, when given the $G_{1} P_{1} P_{2} \rightarrow E$ base, participants should recognize that the effect of $G_{1}$ is stronger than that of both $P_{1} P_{2}$ and is therefore stronger than either of them alone. This relative strength relation is a higher-order relation which takes the causal relations as arguments. When this system is projected to the target $\left(G_{1} P_{1}\right)$, participants should assume that the effect $E$ will occur. In contrast, when given the other base, $G_{1} G_{2} P_{1} \rightarrow E$, participants have no reason to infer that either $G_{1}$ or $G_{2}$ is stronger than $P_{1}$ (since generative relations outnumber preventative relations). Thus the prediction is that the effect inference will be stronger for $G_{1} P_{1} P_{2}$ than for $G_{1} G_{2} P_{1}$ bases.

## Methods

Participants 40 undergraduate students from Northwestern University participated in the study for course credit. One student was excluded because of missing data points and seven were excluded for failing a comprehension check. In all, 32 participants were analyzed in this study.

Materials and Procedure The animal features and structures were those used in Experiment 1 by Lee and Holyoak (2008). Participants were given information about the two animals, a $G_{1} G_{2} P_{1}$ animal and a $G_{1} P_{1} P_{2}$ animal, which served as the different bases. They were also given a target, a $G_{1} P_{1}$ animal (Figure 2A). The bases were described using different sets of features, and therefore the target's features were unique to the given base. In both conditions, participants rated the likelihood of the effect in the target. Inferences were framed as suppositional queries asking participants to predict the number of animals that would exhibit the effect given 100 instances of the target. Furthermore, for the sake of completeness, similarity ratings between base and target were also obtained. These were assessed on a scale of zero to ten, with zero indicating "completely different" and ten indicating "identical". Similarity ratings always preceded inference ratings.

The experiment was conducted on a PC using the software program Qualtrics. Each participant received both $G_{1} G_{2} P_{1}$ and $G_{1} P_{1} P_{2}$ conditions. The order of the two base conditions were counterbalanced. Participants were first given the de-

[^150]scription of a single base and the target animal. Similarity and inference question were placed directly below the description on the same page. Once participants had recorded an answer, they were prompted to continue to the next page. After leaving the page, participants were unable to review the content and answers from the previous page.


Figure 3: Mean outcome (effect) inferences and similarity ratings between target and base. The base categories represent the type of causes each cause had. Error bars represent 1 standard error of the mean.

All participants in Experiment 1 also participated in Experiment 2, which directly proceeded Experiment 1. At the end of the Experiment 2, all participants were given a multiple choice comprehension test that consisted of two questions about the content of stories they were given. If either question was answered incorrectly, that participant's data was removed from analysis for both experiments.

## Results and Discussion

The mean results for both similarity and inference ratings can be seen in Figure 3. The analysis consisted of a paired-sample $t$-test for each rating type. There was a significant difference in mean inference ratings between the conditions: the target in the $G_{1} P_{1} P_{2}$ condition ( $M=62.28, S D=19.84$ ) was rated as significantly more likely to exhibit the effect $E$ compared to the target in the $G_{1} G_{2} P_{1}$ condition ( $M=47.28, S D=17.49$ ), $t(31)=3.44, p<.005$. This suggests that participants' inferences were informed by the higher-order relation observed in the $G_{1} P_{1} P_{2}$ base. There was no significant difference in similarity ratings between the $G_{1} P_{1} P_{2}(M=6.25, S D=1.59)$ and the $G_{1} G_{2} P_{1}(M=5.78, S D=1.75)$ condition $t(31)=1.54, p=.13$. However, the relative similarity ratings were in the same direction as ratings for inferential strength.

These results show that if we assume that people inferred a higher-order relation of relative strength among the causal relations while encoding the base analog, then this relation will be mapped to the target, where it can be used to make causal inferences about the effect $E$. As previously discussed, this pattern of results would be highly unlikely if participants were mapping either a global or local posterior calculation from the base to target. Thus, a higher-order relationship between the generative and preventative causes in the $G_{1} P_{1} P_{2}$
base gave information about the relative strength of $G_{1}$ in the $G_{1} P_{1}$ target. This was not the case for the $G_{1} G_{2} P_{1}$ base. In conjunction with the findings of Colhoun and Gentner (2009), our results suggest that mapping of relational structure, including higher-order relations computed in the base, can account for inferential strength ratings made among analog causal systems.

## Experiment 2

The previous study leaves open another possibility. Perhaps when given the base $G_{1} P_{1} P_{2} \rightarrow E$, participants recognized that $G_{1}$ was stronger than the combined $P_{1} P_{2}$, but simply inferred the absolute strength of $G_{1}$. That is, they inferred that $G_{1}$ was extremely likely to produce effect $E$. To test this, we use the same $G_{1} P_{1} P_{2} \rightarrow E$ base condition that had previously elicited increased inference ratings in Experiment 1. However, this time, the base is constant while the targets vary (see Figure 2B). In both conditions, participants receive the same generative cause (i.e., $G_{1}$ ) observed in the base analog, but also an additional preventative cause that differs by condition. In the Familiar $P$ condition $\left(G_{1} P_{1}\right)$, they receive the same $P_{1}$ feature found in the base. In contrast, in the Novel $P$ condition $\left(G_{1} P_{3}\right)$, they receive a novel preventative feature (i.e., $P_{3}$ ) that has no corresponding relation in the base. If participants simply infer extremely strong causal strength for $G_{1}$ (i.e., $G_{1}$ overpowers the preventative causes), then we should observe no difference between the two conditions. However, if participants are transferring a higher-order relational structure from the base to target, then we should find that people rate the effect to be more probable in the Familiar $P$ condition compared to the Novel $P$ condition.

## Method

Participants The same 40 Northwestern students who participated in Experiment 1 also participated in Experiment 2. Furthermore, the same eight participants whose data was removed from analysis were likewise removed for Experiment 2. A total of 32 participants were therefore analyzed.

Materials and Procedures The animal features for Experiment 2 were taken from Colhoun and Gentner (2009). All participants were run in both the Familiar $P$ and Novel $P$ conditions; order was counterbalanced between participants. They were given the same similarity and inference tasks as in Experiment 1. As before, similarity queries always preceded inference ratings. Participants began the experiment immediately after finishing Experiment 1.

## Results and Discussion

Figure 4 shows the mean inference and similarity ratings. As before, a paired-sample $t$-test was conducted for both measures. Consistent with the hypothesis, mean inference ratings for the likelihood of effect $E$ were significantly greater in the Familiar $P$ condition ( $M=62.31, S D=21.18$ ) than in the Novel $P$ condition $(M=48.59, S D=17.59), t(31)=3.71, p<.001$. Indeed, in the Novel $P$ target, the estimates of likelihood of $E$


Figure 4: Mean outcome (effect) and similarity ratings between target and base. The target categories represent the type of preventative cause present in the target. Error bars represent 1 standard error of the mean.
did not differ from chance (50\%). Similarity ratings were also significantly greater for the Familiar $P$ condition ( $M=6.44$, $S D=1.54$ ) than for the Novel $P$ condition ( $M=4.19, S D=1.67$ ), $t(31)=6.95, p<.001$. This is to be expected, because in the Familiar $P$ condition, there are two shared factors ( $G_{1}$ and $P_{1}$ ), while in the Novel $P$ condition only one factor is shared $\left(G_{1}\right)$.

These findings run contrary to the idea that participants are simply transferring absolute information about the strength of $G_{1}$ from the base to target. Had people simply transferred the strength of the generative relation, the effect inference would have been equally strong in both targets. Instead, participants only inferred that the effect occurs in the target when they could map higher-order relative strength relations from the base to the target. In sum, these findings suggest that participants' inferential strength ratings for the effect in a target can be accounted for by standard analogical mapping models.

## General Discussion

Holyoak and colleagues (Holyoak et al., 2010; Lee \& Holyoak, 2008) argue that causal analogies cannot be modeled in the same way as other analogies and instead require the creation of a specialized system. Specifically, they believe that most existing models, including SMT, cannot accommodate the probabilistic dynamics of causal systems. The evidence provided here suggests otherwise. Across two experiments, we demonstrate that the pattern of analogical inferences observed among various causal systems correspond with predictions made by SMT. Experiment 1 found that stronger effect inferences occurred when the causal relations in the base were united by a higher-order relation that took causes as arguments. In Experiment 2, we tested whether the results could be predicted by assuming the transfer of the individual causal strength of the generative relation from the base to target. On the contrary, the results suggest that a consistent relational structure is required in order for people to infer the effect in the target.

There has been immense progress in analogy research in the last few decades. The evidence suggests that analogy is
a domain-general process that applies across physical causality (Goldwater \& Gentner, 2015), mathematics (Mix, 2008; Rittle-Johnson \& Star, 2007), politics (Spellman \& Holyoak, 1992), spatial scenes (Doumas \& Hummel, 2013; Kurtz \& Gentner, 1993; Richland, Morrison, \& Holyoak, 2006; Richland et al., 2006; Markman \& Gentner, 1993; Sagi, Gentner, \& Lovett, 2012), and scientific reasoning (Gentner, 2002; Pearl, 1992). Our findings here support the idea that analogy is a domain-general process that supports alignment and inference both within and across domains.

## Acknowledgments

We thank Erin Anderson, Kensy Cooperride, Nathan Couch, Christian Hoyos, Sandy LaTourrette, Francisco Maravilla, Ruxue Shao, and Nina Simms for their comments and suggestions. This research was supported by a grant from the Office of Naval Research [award number N00014-16-1-2613] to Dedre Gentner.

## References

Clement, C., \& Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. Cognitive Science, 15.

Colhoun, J., \& Gentner, D. (2009). Inference processes in causal analogies. In New frontiers in analogy research: Proceedings of the second international conference on analogy (p. 82-91). Sofia, Bulgaria: New Bulgarian University Press.
Doumas, L., \& Hummel, J. (2013). Comparison and mapping facilitate relation discovery and predication. PLOS One, 8 , e63889.
Falkenhainer, B., Forbus, K., \& Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial Intelligence, 41, 89-132.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Gentner, D. (2002). Analogy in scientific discovery: The case of johannes kepler. In L. Magnani \& N. Nersessian (Eds.), Model-based reasoning: Science, technology, values. New York: Kluwer Academic/ Plenum Publisher.
Gentner, D., \& Markman, A. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 45-56.
Gentner, D., Rattermann, M., \& Forbus, K. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive Psychology, 25, 524-575.
Goldwater, M., \& Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. Cognition, 137, 137153.

Holyoak, K., Lee, H., \& Lu, H. (2010). Analogical and category-based inference: A theoretical integration with bayesian causal models. Journal of Experimental Psychology: General, 139, 702-727.
Kurtz, K., \& Gentner, D. (1993). Detecting anomalous features in complex stimuli: The role of structured compar-
ison. Journal of Experimental Psychology: Applied, 19, 219-232.
Lee, H., \& Holyoak, K. (2008). The role of causal models in analogical inference. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 1111-1122.
Markman, A., \& Gentner, D. (1993). Structural alignment during similarity comparisons. Cognitive Psychology, 25, 431-467.
Mix, K. (2008). Children's equivalence judgments: Crossmapping effects. Cognitive Development, 23, 191203.

Pearl, J. (1992). Conceptual revolutions. Princeton, NJ: Princeton University Press.
Richland, L., Morrison, R., \& Holyoak, K. (2006). Childrens development of analogical reasoning: Insights from scene analogy problems. Journal of Experimental Child Psychology, 94, 249-273.
Rittle-Johnson, B., \& Star, J. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? an experimental study on learning to solve equations. Journal of Educational Psychology, 99, 561-574.
Sagi, E., Gentner, D., \& Lovett, A. (2012). What difference reveals about similarity. Cognitive Science, 36, 1019-1050.
Spellman, B., \& Holyoak, K. (1992). If saddam is hitler then who is george bush? analogical mapping between systems of social roles. Journal of Personality and Social Psychology, 62, 913-933.

# Individual Differences in Transfer Mediated by Conceptual Priming 

Jairo Navarrete (jnavarrete@ubiobio.cl)<br>Departamento de Ciencias Básicas, Universidad del Bío-Bío<br>Avda. Andrés Bello 720, Chillán, Chile


#### Abstract

Research in analogical transfer suggests a simple type of transfer that occurs due to the activation of key relational concepts. Analysis on mental structured representations indicates that this transfer may act differently depending upon structural and perceptual features of the priming task. Two hundred eight participants were assigned to three experimental groups where they received a structure-priming, tested once and afterwards they received a perceptual-priming and tested again. As predicted, the effect of structure-priming was found across conditions whereas the effect of perceptualpriming (a six-second animation) was detected only in subjects with high levels of cognitive reflectiveness. These individual differences are interpreted as evidence that only highly reflective subjects were able to process visuospatial cues in the animation and to extract their structural features, hence activating relational concepts that influenced their interpretations of subsequent tasks.


Keywords: Analogy, transfer, priming, cognitive reflection.

## Introduction

In the fields of problem solving and analogy research, certain transfer effects have been linked to the low level cognitive process of priming. In one study, subjects were confronted with a biochemistry problem and learned that an inhibitory enzyme decreased virus reproduction (Schunn \& Dunbar, 1996). The next day, the same subjects confronted a genetics problem involving the concept of an inhibitory gene. Although the two problems were not analogous, subjects exposed to the key concept of inhibition in the first session were more likely than control subjects to develop a solution based on the concept of inhibition for the transfer problem. Similarly, Day and Goldstone (2011) showed that subjects familiarized with a simulated physical system of motion were able to transfer the notion of "oscillatory motion" to interpret transfer tasks posed in a context of urban planning. This transfer arises not from a systematic mapping from source to target, but it is due to the activation of one or more key concepts (i.e. priming) and thus it has been called a "piecemeal transfer" (Holyoak, 2012, p. 246).

Similar effects were observed by Burns (1996) when researching transfer between episodes of analogical reasoning. The experimental tasks involved four-term letterstring analogies such as the one depicted in Figure 1. Burns had participants justify a given answer to a first analogy prior to having them propose solutions to a second (nonanalogous) one. Half of the subjects had the two analogical problems presented in the reversed order, and the differences between the two groups on the relative frequencies of all defensible solutions suggest order effects that are consistent with the hypothesis that concepts activated during the first episode effectively biased the solution strategies followed
by participants during the second episode. Burns explained these order effects as the formation of analogical mappings during the first analogical episode which afterwards are transferred to the second analogical episode.

> abc : abd :: xyz : ???

Figure 1: This analogical problem can be stated as "if you changed abc to abd, how would you change xyz in the same way?"
However, Burns' theoretical explanation cannot account for the "piecemeal transfer" observed in the aforementioned studies because the referred problems are not analogical episodes: Within the biochemistry problem, there are no analogical mappings that can be transferred to the genetics problem. Similarly, no analogical mappings in the physical system can be transferred to the domain of urban planning. Hence, this study adopts the view that the activation of key concepts (i.e. priming) helps subjects to spontaneously interpret subsequent tasks according to the primed concept. The research presented here contributes by identifying differences among two kinds of priming referred here as structural priming and perceptual priming. The former being akin to the effect documented by Burns whereas the latter more akin to the one observed by Day \& Goldstone.

A proper outline of the method and predictions for this study requires an analysis of the mental representations underlying tasks such as the one in Figure 1. The next section provides such analysis and proposes mechanisms by which priming may affect the interpretation of these tasks.

## Representations of Four-term Analogies

Four term analogy problems have the form A : B :: C : ? where A and B comprise the source of the analogy whilst C along with the unknown constitute the target of the analogy. To solve this problem, one must interpret the source domain and then look for a solution D such that the relations between A and B can be mapped to the relations between C and D . This kind of problems promote the identification of the source's structure (Indurkhya, 1989) and thus they will be used in the present study to induce structural priming.

Additionally, these tasks will be used here as a manner to measure the priming effect by assessing the interpretation given to the source domain when solving these problems. This is possible because four term analogies in letter-strings (see Figure 2) can be represented by propositions (Burns, 1996). For example, the "append" concept may be represented by the schema " + " in a way that " $+(\mathrm{ab}, \mathrm{cd})$ " represents "abcd". And certain notion of reflection can be represented by the schema "mirror" so that "mirror(xyzw)" becomes a representation of "wzyx".

Relational schemas (such as the ones mentioned above) play an important role in cognitive research because there is
evidence that they allow encoding structural knowledge as a combination of schemas and primitive elements (Halford, Bain, Maybery, \& Andrews, 1998). For example, the string "aab12aab" can be seen as emerging from two primitive elements, namely "aab" and " 12 ". And the source domain of the problem in Figure 2 (i.e. the first two terms) can be represented as it is depicted at the top of Figure 3. In such case, dominant theories of analogical mapping (Gentner, 1983; Holyoak \& Thagard, 1989) predict that "cde12ede" would be the response to this analogical problem.

> aab12aab : aab12baa :: cde12cde : ????????

Figure 2: An ambiguous four term analogy problem
However, the structure depicted in Figure 3 is not the only one that fits the problem, since the base analog could be alternatively conceptualized by a "shift" schema, namely an operation that takes the last letter and moves it to the first position (e.g. shift(xyzw) $=$ wxyz). Although shift and mirror schemas represent different concepts, they sometimes cannot be differentiated from their action on one particular instance. As an illustration, notice that

$$
\operatorname{mirror}(a a b)=\operatorname{shift}(a a b)=\text { baa } .
$$

This makes the analogy in Figure 2 an ambiguous problem: replacing all the "mirror" occurrences in Figure 3 by "shift" leads to another structured representation which predicts a different answer, namely "cde12ecd". This provides a basis to detect the priming effect by assessing the interpretation used by subjects in resolving the problem: if they conceptualized it through the mirror schema, they would prefer "cde12edc" as an answer, whereas if they used the shift schema, they would rather prefer "cde12ecd".

The present study tests the hypothesis that tasks requiring the generation of structured representations can prime a schema in a different manner than those tasks that do not require such structured representations. Several studies from analogical transfer research emphasize differences between abstract, structural aspects of an episode and its perceptual, concrete aspects. The evidence indicates that both aspects influence analogical transfer but that structural aspects have a greater impact than perceptual ones (Blanchette \& Dunbar, 2000; Holyoak \& Koh, 1987). Structural and perceptual features of a priming task may influence transfer as follows: a structure-based task (e.g. the one in Figure 2) might force subjects to create structured representations, thus activating the involved relational schemas. In contrast, a perceptionbased task (e.g. describing a dynamic visual stimulus) may evoke structured representations only in specific individuals, thus leaving open the possibility that schemas remain deactivated while performing the task. More precisely, these individuals may extract structural features from visuospatial cues in the perception-based tasks thus activating relational schemas that influence their interpretations of subsequent tasks. To address for this possibility, the cognitive reflection test (CRT, Frederick, 2005) was taken into account because research in decision making has shown that people who score highly on this test are more likely to engage in
rational, analytic thinking (Shah, Michal, Ibrahim, Rhodes, \& Rodriguez, 2017) whereas subjects with lower scores are less sensitive to notice abstract aspects of a task (Toplak, West, \& Stanovich, 2011).


Figure 3: (Top) The domain "aab12aab : aab12baa" is represented as a combination of schemas-"append" and "mirror"-and primitive elements-"aab" and " 12 ". (Bottom) The answer "cde12edc" for the problem in Figure 2 results from applying the same combination of schemas to the target domain by mapping primitive elements.

Summarizing, structural-tasks should force the generation of structured representations, meaning that the effect of this structural-priming on interpreting a subsequent episode is independent of participant's cognitive reflection. In contrast, perceptual-tasks may evoke structured representations only in subjects with high cognitive reflectiveness, meaning that the effect of perceptual-priming on interpreting subsequent episodes is expected to be effectively modulated by participant's cognitive reflection.

## Method

## Participants

An initial sample of 231 undergraduate students at a Chilean university (Age range $19-29$ years, $M=20.6 S D=2.1$ ) participated in the study for course credits. They were randomly assigned to one of three conditions: two priming conditions and one control condition.

## Procedure and Design

Subjects in the "first mirror then shift" condition (M-S) received a structure-priming favoring a "mirror" schema and were presented with a first test. Afterwards they received a perceptual-priming favoring the "shift" schema and were presented with a second test. Similarly, subjects in the "first shift then mirror" condition (S-M) received a structurepriming favoring the "shift" schema and were presented with the first test. Afterwards they received a perceptualpriming favoring the "mirror" schema and were presented with the second test. This design (detailed below) permits assessing the desired effect of priming on interpreting a transfer episode as follows: the effect of structure-priming can be assessed by comparing the experimental groups in terms of the scores collected in the first test. The effect of perceptual-priming will be assessed through comparing the change of scores (from the first test to the second test) experienced by each experimental group. The no-priming condition (NP) was taken as a control condition: subjects in this condition were primed to activate different schemas than the mirror and shift ones.

The experiment was administered in small groups at the computer laboratory of the university, with each participant working individually at her own pace. Participants took
between $15-20 \min (M=17.6 ; S D=4.6)$ to complete the experiment. The Qualtrics online platform was used to build a questionnaire comprising the following phases: (1) Introduction, (2) Structure priming, (3) Test 1, (4) Cognitive Reflection Test, (5) Perceptual priming, (6) Test 2 and (7) Test 3 (a hinted repetition of Test 2). As noted above, the between-subjects manipulation was restricted to phases 2 and 5; the other phases were identical across conditions. A detailed description of each phase is provided now.


Figure 4: Display for the acquisition of preference scores.

Introduction: In order to familiarize participants with 4-term analogies, the computer screen presented the analogy "Chile:Santiago::Argentina:???", along with its meaning i.e. "Chile is related to Santiago in the same way that Argentina is related to which city?". Three alternative answers were given (Buenos Aires, Mendoza and Paris) each one coupled with a sliding bar ranging from 1 to 100 (see Figure 4). A text below explained that the three scores must add up to 100 , and that they should reflect how good each answer seemed to be. The options were preset at scores of 90,9 and 1 , respectively, with an accompanying text stating that while Mendoza might be a plausible answer (score of 9) because the fact that it belongs to Argentina parallels the fact that Santiago belongs to Chile, a much better answer should be Buenos Aires (score of 90), because its being the capital of Argentina matches the "capital of" relation that holds between Santiago and Chile. Finally, the instructional text stated that Paris is not a good answer (score of 1) because it is hard to find a relation between Paris and Argentina that parallels some relation between Santiago and Chile.

Structure-priming: This phase was intended to activate particular schemas in participants' minds through a priming method similar to the one presented in Burns (1996). Participants in the M-S condition received the mno678 : onm876 :: def234 : ??? problem. According to my analysis, this problem allows only one acceptable solution that involves projecting onto the rightward term of the analogy the mirroring operations that transform the base structures "mno" and " 678 " into "onm" and " 876 ", respectively. The above transformations could only be
conceptualized in terms of mirroring operations, and thus it was expected participants to massively assign high scores to fed432 (derivable via mirroring) and low scores to both edf324 and 4def23 (not derivable via such operation). This should lead to an increased activation of the mirror-schema in relation to other possible transformations. Participants in the S-M condition received the mno678: omn867 :: def234 : ??? problem, an analogy whose only acceptable solution involved applying the "shift" operation that transform the base structures "mno" and " 678 " into "omn" and " 867 ", respectively. We expected participants in this group to assign high scores to fde423, and low scores to both edf324 and 4def23 alternatives. Participants in the NP condition received the human:lungs::fish:??? analogy, followed by the alternatives "gills", "spine" and "fins". Given that this analogy should be solved by evoking the relation "X breathes through $\mathrm{Y}^{\prime \prime}$, we expected this control condition to prime neither a mirroring nor a shifting operation. In these and all subsequent analogical problems, three competing solutions were presented in random order, with their corresponding sliding bars preset to one.

Test 1 To assess whether the structure-priming received in the previous phase can bias subsequent processing, participants of all conditions received the ambiguous problem aab12aab : baa12baa :: cde12cde : ???. This analogy is solvable by applying either mirroring or shifting operations (options edc12edc and ecd12ecd, respectively). The remaining option (dec 12dec) could not be derived from the leftward term of the analogy, and thus was expected to receive low scores regardless of condition.

Cognitive Reflection Test (CRT) This stage was presented to participants as a problem solving section. It comprised three algebra problems whose correct solution does not require complex calculations, but requires participants to suppress an "impulsive" solution that easily comes to mind ${ }^{1}$. As an example, the first item of the CRT consisted of the following problem: "A bat and a ball cost $\$ 1.10$. The bat costs $\$ 1.00$ more than the ball. How much does the ball cost?" Participants had no time limit to answer the problems. No particular criterion was taken into account in order to place the CRT here (between the two main measurement stages). After completing the third item, a yes/no question queried participants about whether they were familiar with any of the problems prior to the experimental session.

Perceptual priming: After being presented with a web video player, participants were asked to run a (six seconds)

[^151]video animation. They were told that they would be able to watch the animation just once, and that they should pay careful attention in order to answer one brief question about the animation (answer limited to 200 characters). This question was aimed to provide a control mechanism to assess whether participants attended to the animation. Depending on the condition, the animation displayed a geometrical array whose dynamic movement was either compatible with a mirroring operation (S-M), with a shift operation (M-S) or unrelated to both (NP).
For the S-M condition, the animation showed how a transparent panel containing three horizontally arranged card figures performed a $180^{\circ}$ turn along the middle vertical axis ${ }^{2}$. As the left-to-right order of the figures changed from "club, diamond, heart" to "heart, diamond, club", this animation was a visuospatial representation of the "mirror" operation. Participants were asked about how the spatial configuration of the club changed during the animation.

For the M-S condition, the animation showed a hammer imparting a rightward motion to the leftmost of three horizontally arranged geometrical figures; this rightward motion was transmitted to the middle figure and ultimately transmitted to the rightmost figure, making it slide-off through a circular circuit that ended up relocating it in the leftmost position ${ }^{3}$. As the left-to-right order of the figures changed from "circle, square, rhombus" to "rhombus, circle, square", this animation was intended to convey a visuospatial representation of the "shift" operation. Participants were asked about how the spatial configuration of the rhombus changed during the animation.

For the NP condition, the animation displayed a transparent rectangle containing three horizontally-arranged card figures (club, diamond and heart) which performed a $360^{\circ}$ turn along its middle horizontal axis, thus leaving the left-to-right ordering of the figures unchanged ${ }^{4}$. As this rotational movement was unrelated to either the mirror or shift operations, it was intended to avoid the activation of the mental representations associated to the two crucial operations. Subjects were asked how the spatial configuration of the club changed during the animation.

Test 2 To assess whether the visuospatial animations received during the previous phase altered subsequent processing, participants of all conditions received the ambiguous problem pq89pq : qp89qp :: xyz89xyz : ???. According to our analysis, this analogy is solvable by applying either mirroring or shifting operations (options zyx89zyx and zxy89zxy, respectively). As in Test 1, the remaining option (xyz89xyz) could not be derived from the leftward term of the analogy, and thus was expected to receive low scores regardless of condition. Upon assigning scores to each of the presented alternatives, subjects were asked to answer a yes/no question about whether the watched video had spontaneously popped up into their

[^152]minds while reading the analogy and/or evaluating the presented alternatives.

Test 3 This is a control measure aimed to assess the extent to which participants were potentially able to use the information contained in the animation for solving the problem presented in Test 2. Participants received the analogy and the same solution options as in Test 2, but it was preceded by an explicit hint, namely, to take into account the animation for assigning scores to the presented alternatives.

## Data Analysis

Participants were classified as having low cognitive reflection if their CRT score was equal to zero, and as having high cognitive reflection in the other cases. Since this study is based on priming effects, I discarded data from 23 participants (4 in the M-S condition and 19 in the S-M condition) who failed to assign high scores ( $>=80$ ) to the only defensible solution to the unambiguous problem of phase 2 , which was meant to operate as a structure-priming for the following phase. A preliminary analysis of the data revealed that a non-negligible proportion of participants in the control condition assigned high scores to the "incorrect" solution for the ambiguous problem presented during Test 1 , thus lessening preferences for the meaningful alternatives. To prevent this unanticipated behavior of the NP group from engendering spurious correlations, raw preference scores were converted to normalized scores which reflect the proportion of preference assigned to the mirror-alternative in relation to the total amount of preference assigned to the two competing and meaningful alternatives:

$$
N S P M=100 * \frac{P_{m}}{P_{m}+P_{s}}
$$

As an example, if a subject assigned a preference of 10 to the mirror-alternative, 40 to the shift-alternative and 50 to the incorrect alternative, its NSPM would be $20 \%$, reflecting that one fifth of the total amount of (the relevant) preference was assigned to the mirror-alternative whereas the remaining $80 \%$ was assigned to the shift-alternative. Due to the ratio form of NSPM scores ${ }^{5}$, I report geometric means (GM) obtained by computing arithmetic means on the logarithm of NSPM scores (see Table 1).

[^153]Table 1
NSPM scores at each stage of this study
Log of NSPM scores Geometric Means GM Ratios

| Condition | N | Log of NSPM scores |  |  |  |  |  | Geometric Means |  |  | GM Ratios |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Test 1 |  | Test 2 |  | Test 3 |  | Test 1 Test 2 Test 3 |  |  | T2/T1 T3/T1 |  |
|  |  | M | SD | M | SD | M | SD | GM | GM | GM |  |  |
| Low CRT |  |  |  |  |  |  |  |  |  |  |  |  |
| NP | 34 | 3,52 | 1,59 | 3,82 | 1,48 | 3,45 | 1,57 | 32,8 | 44,6 | 30,5 | 1,36 | 0,93 |
| M-S | 47 | 4,01 | 1,15 | 4,24 | 0,71 | 2,5 | 2,15 | 54,1 | 68,4 | 11,2 | 1,26 | 0,21 |
| S-M | 32 | 3,23 | 1,66 | 3,12 | 1,89 | 3,54 | 1,83 | 24,3 | 21,6 | 33,5 | 0,89 | 1,38 |
| High CRT |  |  |  |  |  |  |  |  |  |  |  |  |
| NP | 28 | 4,09 | 1,21 | 4,18 | 1,01 | 3,95 | 1,43 | 58,7 | 64,4 | 50,9 | 1,10 | 0,87 |
| M-S | 35 | 3,98 | 1,06 | 3,84 | 1,28 | 1,86 | 2,12 | 52,5 | 45,5 | 5,4 | 0,87 | 0,10 |
| S-M | 32 | 2,71 | 1,94 | 3,64 | 1,31 | 3,91 | 1,38 | 13,9 | 37,1 | 48,4 | 2,67 | 3,49 |

Table 1: The rows of the table report the (within-subjects) change of NSPM scores across the three measurement phases. The effect of structural priming is reported in the columns associated to Test 1 . The column T2/T1 reports the ratio of geometric means associated to Test 1 and Test 2 and suggests a between-groups change of preferences detected only in subjects with higher levels of cognitive reflectiveness.

## Results

Structure-priming A two-way ANOVA was conducted to examine the effect of condition (S-M, M-S and NP) and CRT level (low vs. high) on the NSPM scores collected during Test 1 . Main effects emerged for condition $F(2,203)$ $=10.09, \eta_{\mathrm{p}}^{2}=.085, p<.0001$. Cognitive reflection does not affect the interpretation of the problem in Test 1 since the main effect of cognitive reflection, $F(1,203)=.0001, \eta_{\mathrm{p}}{ }^{2}=$ $0, \mathrm{p}=.99$ and the interaction condition x cognitive reflection, $F(2,203)=2.3, \eta_{\mathrm{p}}^{2}=.022, p=.10$ were not statistically significant. Planned comparisons confirmed differences on NSPM scores between the two competing conditions in each one of the two CRT levels. For high-CRT participants, the S-M condition produced significantly lower NSPM scores $(G M=13.85)$ than those of the M-S condition $(G M=52.41), t(47.16)=3.31, p=.0017, d=.58$. For lowCRT participants, the S-M condition produced significantly lower NSPM scores $(G M=24.18)$ than the $\mathrm{M}-\mathrm{S}$ condition $(G M=54.09), t(50.99)=2.32, p=.024, d=.19$. As a base line for comparison, NSPM scores of subjects in the control condition (NP) are reported in Table 1. These results are consistent with those obtained by Burns (1996), and extend those findings by confirming that the effect of structuralpriming on interpreting subsequent analogical tasks takes place regardless of whether participants exhibit (or not) a natural propensity to engage in rational, analytic thinking.

Perceptual-priming In order to detect possible small effects of perceptual priming, this experiment was designed to measure the extent to which a perceptual-priming can counteract the effect of a prior structure-priming. Therefore, participants whose structure-priming favored the mirror scheme later received a perceptual-priming favoring the shifting schema, and vice-versa. The eventual effect of perceptual-priming was assessed by analyzing the change of NSPM scores within subjects (from Test 1 to Test 2) i.e. the change of subject's appraisal for the mirror-compatible solution. To investigate this change of preferences at the level of individual participants, a $3 \times 2 \times 2$ ANOVA was conducted with condition (S-M, M-S and NP) and cognitive reflection (High vs. Low) as between-subjects factors and session (Test 1 vs. Test 2) as a within-subjects factor. Main
effects emerged for condition $F(2,197)=42.42, \eta_{\mathrm{p}}{ }^{2}=.069$, $p<.0001$, but neither for cognitive reflection $F(1,197)=$ $0.32, \eta_{\mathrm{p}}^{2}<.001, p=.57$ nor for session $F(1,197)=3.152$, $\eta_{\mathrm{p}}{ }^{2}=.0011, p=.077$. As expected, the three-way interaction was significant $F(2,197)=3.06, \eta_{\mathrm{p}}^{2}=.017, p=.045$. To further understand this interaction, planned comparisons in each CRT-level were conducted. In the high-CRT group, the change of mirror-preferences was in agreement with the perceptual-priming: changes in the $\mathrm{M}-\mathrm{S}$ condition ( $-7 \%$ ) were significantly different from the change experienced in the S-M condition $(23.2 \%), t(63.36)=-2.63, p=.011, d=$ 0.644. In contrast, the low-CRT group presented changes of mirror-preferences incompatible with the perceptualpriming: changes in the M-S condition (14.3\%) were not significantly different from the changes in the S-M condition $(-2.7 \%), t(48.31)=.75, p=.46, d=.177$. Data in Table 1 suggests that the control group increased its appraisal for the mirror-alternative from Test 1 to Test 2, which indicates an inherent bias in the experimental design. Still, the results are consistent with the expectations: within participants with high reflectiveness, those who watched the visuospatial representation of the shift-scheme tended to lower their appraisal for the mirror-alternative and those who watched the visuospatial representation of the mirrorscheme increased their appraisal for the mirror-alternative. This is in line with the idea that participants with high reflectiveness were able to process the visuospatial primingtask and extract its relational features thus activating a schema that lessened the effect of the one activated in the previous structure-priming phase. In contrast, participants with low reflectiveness completed the visuospatial primingtask without noticing its structural features, thus not activating any schema and leaving "untouched" the effect of the schema activated in the structure-priming phase.
The differential effect of visuospatial representations between the two experimental groups was not due to the relative proportions of subjects consciously recalling the animations: A chi-squared test showed that the low-CRT and high-CRT groups had similar proportions of participants stating to recall the video while solving the analogy presented during Test $2(21.5 \%$ vs. $28.3 \%$, respectively), $\chi^{2}(1, N=146)=0.58, p=.45$. Only a minor proportion of participants spontaneously recalled the animation while solving the analogy in Test 2. This evidences that such stimuli have primed-rather than consciously induced-participants' appraisal of the mirroralternative since a lack of conscious awareness represents a definitional feature of priming.
As cognitive reflection is correlated with general intellectual abilities (Frederick, 2005), an alternative interpretation for why participants with lower levels of cognitive reflection were not sensible to perceptual priming is that they are intrinsically less capable of translating the perceptual content of the animations to the more structured domain of letter-string problems. To assess this possibility, preferences on the two experimental groups was investigated when subjects were explicitly asked to recall
the videos and use them into solving the analogy problem presented in Test 3 (same problem presented in Test 2). A two-way ANOVA was conducted to examine the effect of condition (S-M, M-S and NP) and cognitive reflection (low vs. high) on NSPM scores collected during Test 3. Main effects emerged for condition $F(2,203)=17.88, \eta_{\mathrm{p}}^{2}=.147$, $p<.0001$; but neither for cognitive reflection $F(1,203)=$ $.08, \eta_{\mathrm{p}}^{2}=0, p=.77$, nor for the condition x cognitive reflection interaction, $F(2,203)=2.21, \eta_{\mathrm{p}}{ }^{2}=.021, p=.11$. This confirms that the uneven effect of perceptual animations as a function of cognitive reflection did not originate in a higher intrinsic ability of the higher cognitive reflection group to understand the correspondences between the visuospatial animation and the subsequent analogical activity.

## General Discussion

Theoretical analyses suggest a simple mechanism whereby transfer mediated by conceptual priming may occur, namely the activation of relational schemas-organized in structured representations-that influence the interpretation of subsequent tasks. The fact that analogical transfer is influenced both by structural and superficial features suggests that transfer mediated by priming should be also subjected to these two aspects: I expected differences of priming effects among tasks requiring structured representations and tasks requiring perceptual descriptions of visuospatial animations. The results confirm this hypothesis and link the disparity of effects to individual differences of subjects: structure-priming effects were found across conditions, whereas only those subjects with higher propensity to engage in analytical thinking were sensitive to perceptual-priming. These individual differences cannot be accounted for by group differences in neither the proportion of subjects consciously recalling the visual stimulus nor the ability to understand the correspondences between the visuospatial prime and the transfer task. Hence, the rationale is that highly reflective subjects unconsciously extract structural features from visuospatial cues thus activating relational schemas that influence their interpretations of subsequent tasks.

This seems to be consistent with certain evidence in literature. For example, students were asked to solve algebraic equations superimposed on a vertically oriented grating continuously moving either to the left or to the right (Goldstone, Landy, \& Son, 2010). This was meant to investigate the effect of these background motions on the "spatial transposition strategy" e.g. moving the number 8 from the left to the right of the equality in $4 * y+8=24$. The study found that the compatibility of the background motion and the motion of numbers implicated by the spatial transposition strategy affects accuracy. The analysis and results presented here indicate that participants with higher propensity to engage in analytical thinking should be more affected by this compatibility because they are more likely to activate a motion-schema from the background motion which, depending on the condition, either conflicts or agrees with the "spatial transposition strategy". Interestingly, the
aforementioned study reported that participants who have taken advanced courses of mathematics were indeed more affected by the compatibility between the two motions.

The results presented in this study are in line with claims in literature suggesting that deep structural aspects are more influential than perceptual-concrete aspects in achieving learning and transfer. But although these results must be viewed as preliminary given the specificity of this study's materials and scope, they open a question for further research: Can this general dichotomy between perceptual and structural aspects be linked to individual differences between subjects?

## Acknowledgements

This research was supported by the Universidad del Bío-Bío grant 124009 3/D and FAPEI Fund. Thanks to Máximo Trench for constructive and helpful discussions. And thanks to various anonymous reviewers for useful comments.

## References

Blanchette, I., \& Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. Memory \& Cognition, 28(1), 108-124.
Burns, B. D. (1996). Meta-analogical transfer: Transfer between episodes of analogical reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(4), 1032.
Day, S. B., \& Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(3).
Fleming, P. J., \& Wallace, J. J. (1986). How not to lie with statistics: the correct way to summarize benchmark results. Communications of the ACM, 29(3), 218-221.
Frederick, S. (2005). Cognitive reflection and decision making. The Journal of Economic Perspectives, 19(4), 25-42.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155-170.
Goldstone, R. L., Landy, D. H., \& Son, J. Y. (2010). The Education of Perception. Topics in Cognitive Science, 2(2), 265-284.
Halford, G. S., Bain, J. D., Maybery, M. T., \& Andrews, G. (1998). Induction of relational schemas: Common processes in reasoning and complex learning. Cognitive Psychology, 35(3), 201-245.
Holyoak, K. J. (2012). Analogy and relational reasoning. The Oxford Handbook of Thinking and Reasoning, 234-259.
Holyoak, K. J., \& Koh, K. (1987). Surface and structural similarity in analogical transfer. Memory \& Cognition, 15(4), 332-340.
Holyoak, K. J., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive Science,13(3),295-355.
Indurkhya, B. (1989). Modes of Analogy (pp. 217-230). London, UK, UK: Springer-Verlag.
Schunn, C. D., \& Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning. Memory \& Cognition, 24(3), 271-284.
Shah, P., Michal, A., Ibrahim, A., Rhodes, R., \& Rodriguez, F. (2017). What Makes Everyday Scientific Reasoning So Challenging? In Psychology of Learning and Motivation (Vol. 66, pp. 251-299). Elsevier.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory \& Cognition, 39(7).

# When extremists win: On the behavior of iterated learning chains when priors are heterogeneous 

Daniel J. Navarro (d.navarro@unsw.edu.au)<br>School of Psychology, University of New South Wales

Amy Perfors (amy.perfors@adelaide.edu.au)<br>School of Psychology, University of Adelaide

Arthur Kary (art.kary@gmail.com)<br>School of Psychology, University of New South Wales

# Scott Brown (scott.brown@newcastle.edu.au) 

School of Psychology, University of Newcastle
Chris Donkin (christopher.donkin@gmail.com)
School of Psychology, University of New South Wales


#### Abstract

How does the process of information transmission affect the cultural products that emerge from that process? This question is often studied experimentally and computationally via iterated learning, in which participants learn from previous participants in a chain. Much research in this area builds on mathematical analyses suggesting that iterated learning chains converge to people's priors. We present three simulation studies suggesting that when the population of learners is heterogeneous, the behavior of the chain is systematically distorted by the learners with the most extreme biases. We discuss implications for the use of iterated learning as a methodological tool and for the processes that might have shaped cultural products in the real world.


Keywords: Iterated learning; language evolution; cultural evolution; inductive biases; Bayesian cognition

Which aspects of our language or culture are shaped by the inductive biases possessed by people, and which aspects are shaped by the process of transmission from one learner to the next? A key framework for thinking about and disentangling these factors is known as iterated learning, shown schematically in Figure 1. Iterated learning is a particular kind of cultural transmission in which behavior arises in one individual (or generation) by learning from the observations of the previous person (generation), forming a chain of learners.

An appealing characteristic of iterated learning is that the behavior of iterated learning chains can be characterized mathematically: under certain assumptions, iterated learning chains with Bayesian learners will converge to a distribution that depends on the learners' priors and the size of the bottleneck (Griffiths \& Kalish, 2007; Rafferty, Griffiths, \& Klein, 2014). These results have allowed researchers to explore inductive biases in different tasks, including function learning (Kalish, Griffiths, \& Lewandowsky, 2007), visual working memory (Lew \& Vul, 2015), reasoning about everyday events (Lewandowsky, Griffiths, \& Kalish, 2009), and category learning (Canini, Griffiths, Vanpaemel, \& Kalish, 2014). They have been especially useful in studying language evolution (Kirby, Griffiths, \& Smith, 2014).


Figure 1: Schematic illustration of a typical iterated learning paradigm, which assumes that learner $n$ learns on the basis of the data provided by learner $n-1$.

Importantly, the theoretical proofs about how iterated learning chains converge depend critically on the assumptions made. For example, if learners select the hypothesis with the highest posterior probability rather than sample from their posterior, an iterated learning chain will tend to exaggerate the prior (Kirby, Dowman, \& Griffiths, 2007). Similarly, we use language to talk about things and events in the world. If one changes the mathematical assumptions to reflect this insight, then the stationary distribution of the chain more closely resembles the posterior distribution (Perfors \& Navarro, 2014). In this paper we consider the role played by individual differences. Such differences are robustly observed in many areas of cognition, yet theoretical results typically assume that all learners share the same biases.

When individual differences exist, what should we expect to observe? One possibility is that the chain converges to a distribution that reflects the "average prior belief" in some sense. For instance, if $10 \%$ strongly believe in hypothesis A and $90 \%$ of people strongly believe in hypothesis B, one might hope that an iterated learning chain reflects $10 \% \mathrm{~A}$ and $90 \%$ B hypotheses. Alternatively, perhaps the chain will produce some other reasonable compromise between A and B that weights each learner in equal proportion. Our findings indicate that neither of these situations necessarily occurs: if people do not share the same priors, iterated learning is not guaranteed to converge to the prior in any meaningful sense. Instead, the distribution to which it does converge is disproportionately influenced by the most biased learners. We illustrate this using three simulation studies.

## Case study 1: Language evolution

Do all learners have equal influence on the process of language evolution? Consider the pressures on a language to incorporate a particular grammatical rule or not. Some learners may have STRONG opinions about a particular rule or construction, whereas others might have WEAK opinions. Exactly who has which might might vary with the particular linguistic context and construction involved: for instance, children may to have a bias for regularization that adults do not share (Hudson Kam \& Newport, 2005), but adult secondlanguage learners may have biases based on transfer from their first language while children do not (Ellis, 2015). We are fairly agnostic at this point about what such biases might be; all that matters for the present purposes is that it is plausible that there are individual differences in at least some language learning biases. Our question is what effect this might have on the nature of the evolved language.

To study this, consider the following experimental design. Participants are presented with sentences in an artificial language that may incorporate a construction (e.g., pluralization rule, morphological marking, etc). After training, participants are asked to produce new sentences, which are presented as the input to the next learner in the chain. This is a relatively typical design, and a simple Bayesian model for this learning problem can be constructed as follows.

If $\theta$ denotes the probability that the grammatical rule should be followed, a Bayesian learner specifies a prior distribution $P(\theta)$. For simplicity we assume a $\operatorname{Beta}(a, b)$ distribution in which $P(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$. In our simulations we assume that some learners enter with a STRONG bias about the grammatical rule, formalized via a $\operatorname{Beta}(1,10)$ prior. In contrast, a WEAK learner might have the opposite bias, but not a strong one, which can be formalized with a $\operatorname{Beta}(2,1)$ prior. Regardless of the biases the learner possesses, it is assumed that belief updating follows Bayes' rule. After a training session in which $x$ of $n$ sentences follow the rule, the posterior distribution $P(\theta \mid x)$ is

$$
\begin{equation*}
P(\theta \mid x) \propto P(x \mid \theta) P(\theta) \tag{1}
\end{equation*}
$$

where $P(x \mid \theta) \propto \theta^{x}(1-\theta)^{n-x}$ is the probability of observing $x$ out of $n$ rule-consistent cases if the true probability is $\theta$. Under these assumptions, the posterior over $\theta$ is a $\operatorname{Beta}(a+x, b+n-x)$ distribution. When asked to generate a novel sentence, a Bayesian learner might sample a value of $\theta$ from their posterior, and their output satisfies the rule with probability $\theta$. The number of rule-consistent sentences $y$ generated by the learner is thus sampled from the posterior predictive distribution $P(y \mid x)$ :

$$
\begin{equation*}
P(y \mid x)=\int_{0}^{1} P(y \mid \theta) P(\theta \mid x) d \theta \tag{2}
\end{equation*}
$$

This kind of model is often used to study regularization in iterated learning designs (Ferdinand, Thompson, Kirby, \& Smith, 2013; Reali \& Griffiths, 2009).


Figure 2: Simulating an iterated learning investigation of language evolution. When the learners all share the same bias (left and middle columns) the average proportion of responses converges to the prior mean (top row), and the distribution of responses converges to the prior distribution (bottom row). When the chain is a mixture of STRONG and WEAK learners, the average proportion of responses does not correspond to the average prior expectation, nor does the distribution converge to the average prior in the population.

## Simulation

We simulate the results of three different kinds of iterated learning experiments. In all cases, the first person is taught ten sentences in an artificial language, five consistent with a grammatical rule; they then generate ten sentences used as input to the next learner. In the first experiment all learners have a STRONG bias about the rule, and in the second experiment all of them have a WEAK bias in the opposite direction. In the third experiment, half of the learners have STRONG biases and half have WEAK opposing ones. In each case results are aggregated across 100,000 simulated iterated learning chains.

The results are shown in Figure 2. As predicted by previous work, in both of the homogeneous cases iterated learning experiment transparently reveals the learner biases: the chain converges to the prior. However, when we consider the iterated learning experiment conducted with a mixed population (right panels of Figure 2) we observe a strikingly different result. In this situation - where half of the learners are STRONG and half are WEAK - the average bias in the population is to expect $38 \%$ of sentences to be rule-consistent. Yet, as the top right panel shows, the iterated learning chain converges to a smaller number, with only $27 \%$ of responses following the rule. More importantly, as the bottom right panel reveals, the distribution of responses bears very little resemblance to the underlying population biases. One might have hoped that, when learners bring different priors to an iterated learning experiment, the chain would converge to a weighted average of their priors. In this case, this weighted average would be a 50-50 mixture of the priors of STRONG learners and WEAK learners (plotted as a histogram). As the figure illustrates, the iterated learning chain (lines) does not converge to anything even remotely similar to this mixture distribution.


Figure 3: Distribution of responses in a mixed chain plotted as a function of the type of learner generating the response.

## Discussion

Why does the iterated learning procedure behave this way when the population is heterogeneous? The answer can be found by separating the responses on the last iteration by learner type, shown in Figure 3. As is clear from inspection, the WEAK bias learners (left) are greatly influenced by the STRONG bias learners: their responses are ruleconsistent $36 \%$ of the time, rather than $67 \%$ as one might expect given their $\operatorname{Beta}(2,1)$ prior, and the distribution of responses (lines) deviates markedly from their prior (histogram). The opposite effect occurs too (right panel), but it is much smaller: the STRONG bias learners increase the proportion of rule-consistent responses from the $9 \%$ rate implied by the $\operatorname{Beta}(1,10)$ prior to $17.5 \%$ in the iterated learning chain. Similarly, their distribution of responses is not markedly different from their prior.

As this example illustrates, when individual differences exist an iterated learning procedure is not guaranteed to reveal the inductive biases of the learner. The Strong learners apply a strong inductive bias, and these learners require a lot of evidence before they are willing (or able) to apply the grammatical rule in question. As a consequence, data generated by a WEAK learner will have minimal ability to sway such a person. The reverse does not hold: the WEAK learners in this scenario are very responsive to external input. As a result, a WEAK bias participant makes a much larger adjustment from the prior than does a STRONG bias one, with the consequence that the overall behavior of the mixed chain is much more heavily driven by the group with the strongest bias.

## Case study 2: Group decision making

Groups of people often arrive at beliefs that seem to lack any evidentiary basis, famously described by the "groupthink" phenomenon (Janis, 1982). How do these false beliefs arise? Do they necessarily reflect a bias shared by all reasoners, or can an entire community be misled by a small number of highly biased learners?

To examine this question, we consider a scenario in which a jury of 12 people begin their deliberations with a straw poll. A notepad is passed around the room, with each person writing down whether they would decide in favor of the plaintiff before removing their sheet of paper and passing the pad to the next juror. Unfortunately, each juror can read the inden-
tations left by the previous one, forming an iterated learning chain. A Bayesian juror might reason about this by considering two hypotheses, namely that the evidence favors the plaintiff $(e=1)$ or the defendant $(e=0)$. The trial evidence sets the juror's prior belief that $P(e=1)=\theta$, which is updated when the vote $v$ of the preceding juror is revealed. The juror unconsciously assigns a reliability value $r$ to this information, such that $P(v=1 \mid e=1)=P(v=0 \mid e=0)=r$. If the preceding juror voted for the plaintiff, the juror's posterior degree of belief that the verdict should favor the plaintiff becomes

$$
\begin{equation*}
P(e=1 \mid v=1)=\frac{r \theta}{r \theta+(1-r)(1-\theta)} \tag{3}
\end{equation*}
$$

and the posteriors are calculated similarly when the earlier vote favored the defendant. For simplicity, we assume that jurors generate their vote probabilistically by sampling from the posterior.

As these equations illustrate, when $r=0.5$ the current juror completely ignores the vote provided by the previous one and the posterior probability is identical to the prior. This arises naturally when the current juror is confident that their existing beliefs incorporate all relevant information about the case, and as such the opinions of other jurors can have no influence upon their own beliefs. We refer to such a juror as a GOAT someone who forms their own view and is not led to conclusions by the opinions of others. In contrast, suppose the juror is underconfident or unsure about their beliefs, perhaps suspecting that other jurors have access to different information. Such a juror will set $r>0.5$, because they attribute evidentiary value to the opinions of others. We refer to this kind of a juror as a SHEEP because they are more likely to adjust their vote to agree with the votes of others.

## Simulations with homogeneous chains

We consider three scenarios. In the first scenario all jurors are GOATS who set $r=0.5$ and have a modest opinion in favor of the defendant $(\theta=0.4)$. In the second scenario all jurors are SHEEP who set $r=0.95$ and have a modest opinion favoring the plaintiff $(\theta=0.6)$. Finally we consider a situation where half of the jurors are SHEEP and the other half are GOATS. To illustrate what happens in these situations we simulated each scenario 100,000 times. The results are plotted in Figure 4. Not surprisingly, because the GOAT jurors ignore the input and generate responses directly from their own prior beliefs, the "chain" starts at their prior (on average, $40 \%$ of jurors vote for the plaintiff) and the total number of votes in favor of the plaintiff follows a binomial distribution.

What should we expect to see if all jurors are SHEEP? One reading of the literature suggests that, since iterated learning chains of Bayesian learners converge to the prior, and since the first SHEEP samples from their own prior, we should see a result not dissimilar to the one we see for Goats. That is - while we might expect to see non-independence among successive jurors - we should find that on average a SHEEP juror should vote for the plaintiff $60 \%$ of the time, in accordance with their priors. However, as the middle column of


Figure 4: The jury straw poll. The top row plots the probability that each juror votes for the plaintiff, as a function of their position in the chain (the dashed line plots the population average prior), and the bottom row plots the distribution of votes for the plaintiff. The left and middle plots show juries composed entirely of GOATS and SHEEP respectively. The plots on the right depict a scenario when $50 \%$ of jurors are SHEEP and $50 \%$ are GOATS.

Figure 4 illustrates, this is not what happens. The first juror votes in accordance with their priors, but by the time the 12th juror is polled, the probability of voting for the plaintiff has risen to $67 \%$. Moreover, it is simple to prove that this reflects the true stationary distribution of the chain. To see this, let $p=P\left(v_{i}=1 \mid v_{i-1}=0\right)$ denote the probability that the $i^{\text {th }}$ juror in the chain votes for the plaintiff given that the previous juror voted for the defendant, and similarly let $d=P\left(v_{i}=0 \mid v_{i-1}=1\right)$ denote the probability that the $i^{\text {th }} \mathrm{ju}-$ ror switches the other direction. The transition matrix for the strawpoll is thus

$$
\boldsymbol{T}=\left[\begin{array}{cc}
1-p & p  \tag{4}\\
d & 1-d
\end{array}\right]
$$

A chain with this transition matrix converges to a stationary distribution $\boldsymbol{\pi}$ in which the (marginal) probability of voting for the defendant and plaintiff is proportional to $d$ and $p$ respectively. To verify this, note that

$$
\begin{align*}
\boldsymbol{\pi} \boldsymbol{T} & \propto[d, p]\left[\begin{array}{cc}
1-p & p \\
d & 1-d
\end{array}\right] \\
& =[d(1-p)+p d, d p+p(1-d)] \\
& =[d, p] \propto \pi \tag{5}
\end{align*}
$$

For a SHEEP juror, the probability of switching the vote from the plaintiff to the defendant is $d=(.1 \times .4) /(.1 \times .4+.9 \times$ $.6)=.069$, and similarly the probability of switching the vote towards the plaintiff is $p=(.1 \times .6) /(.1 \times .6+.9 \times .4)=.142$. In the long run, a chain of SHEEP converges to a $67 \%$ probability of voting for the plaintiff even though each individual SHEEP only assigns a $60 \%$ prior probability to the plaintiff.

On the surface, the SHEEP result seems at odds with the convergence proof in Kalish et al. (2007) - Bayesian learners sampling from their posterior do not (in this instance) converge to the prior. To that end, it is useful to note that the

SHEEP chain violates the assumptions of the original proof, because the SHEEP jurors use the wrong likelihood function for the learning problem. The SHEEP juror assigns evidentiary value to the opinions of other jurors when they should not, because all jurors have seen the same facts at trial. This miscalibration creates the "groupthink" behavior: the SHEEP jurors "double count" the evidence, and the iterated learning chain exaggerates their prior bias.

## Simulations with mixed chains

Now consider what happens when SHEEP and GOATS are mixed together in equal proportions (Figure 4, right). The SHEEP assign prior probability of 0.6 to the plaintiff, whereas the GOATS assign prior 0.4 , so the population average prior is 0.5 . Alternatively, if we consider the behavior of the two homogeneous iterated learning chains, the SHEEP on their own would be expected to converge to 0.67 and the GOATS would converge to 0.4 , so the average of these two long run probabilities is 0.54 . If one did not know the detail of the models, it would be reasonable to expect a mixed chain to produce an average probability of voting for the plaintiff somewhere between $50 \%$ and $54 \%$. Unsurprisingly, it does nothing of the sort. Because GOATS are insensitive to the opinions of others and SHEEP are highly sensitive, the GOATS dominate the mixed chain, and the long run behavior converges to a $43 \%$ probability of voting for the plaintiff. That is, the SHEEP "learn" to mimic GOATS but the GOATS make no such accommodation.

## Discussion

The implications of the jury scenario are twofold. First, the SHEEP-only chain illustrates that it is possible for an iterated learning chain to exaggerate biases even when Bayesian learners sample hypotheses from the posterior. The result complements an earlier result by Perfors and Navarro (2014), which showed that the convergence of iterated learning chains is affected when there is an additional input to the chain (i.e., the world passes new information to learners). In the SHEEP chain we find that convergence is even influenced when learners mistakenly believe there is additional information being passed into the chain. This miscalibration drives a kind of groupthink, in which a collection of individually underconfident learners becomes overconfident as a group.

Second, the behavior of a heterogenous chain is not easily predicted by considering the behavior of the corresponding homogeneous chains, or the priors of individual learners. The mixed chain of SHEEP and GOATS is mostly driven by the GOATS, even though a homogenous chain of GOATS produces a much less extreme outcome than the a chain of pure SHEEP. The reason for this is obvious when we consider the decision making strategies used by the two learner types, but we rarely have access to such information in real life.

## Case study 3: Categorization

Our third case study considers a categorization problem with non-Bayesian learners. We consider stimuli that vary along


Figure 5: Categorization with eight items that vary along one dimension (top panel). Items can be organized into categories that are coherent (left panel) or incoherent (right panel).
a single dimension, with 8 exemplars spaced evenly across the range (i.e., at $x=1, \ldots, 8$ ): an example is shown at the top of Figure 5. Each stimulus can be assigned to one of two categories (A or B), and we are interested in the inductive biases that people bring to this categorization problem.

An iterated learning design can be used to explore these biases. During category learning, each learner is shown training items that consist of four exemplars and their category labels, selected randomly subject to the constraint that there must be one exemplar of each category in the training set. During the test phase the learner must classify the remaining four exemplars. An iterated learning chain is constructed by using a random subset of responses from one learner as the training data for the next, again subject to the constraint that the learner must be shown at least one example of each category.

In our simulations we assume each participant applies the Generalized Context Model (GCM: Nosofsky, 1986). In the GCM, the probability of assigning a test item located at $y$ to category A, given training items $\boldsymbol{x}=\left(x_{1}, \ldots, x_{n}\right)$ with labels $\boldsymbol{l}=\left(l_{1}, \ldots, l_{n}\right)$ is proportional to the summed similarities between $y$ and the category A exemplars:

$$
\begin{equation*}
P(y \in A \mid x, \boldsymbol{l})=\frac{\sum_{i \mid l_{i}=A} S\left(x_{i}, y\right)}{\sum_{i \mid l_{i}=A} S\left(x_{i}, y\right)+\sum_{i \mid l_{i}=B} S\left(x_{i}, y\right)} \tag{6}
\end{equation*}
$$

where similarity decays exponentially with distance, $S(x, y)=$ $\exp (-\lambda|x-y|)$. This model has one free parameter: the specificity parameter $\lambda$ that describes how rapidly similarity decays. When $\lambda$ is large, similarity falls away very quickly with distance, and when $\lambda$ is small it diminishes more slowly.

## Category coherence bias

Although not framed as a Bayesian model, the GCM imposes biases on how learners categorize, and these biases depend on $\lambda$. For instance, the GCM prefers "coherent" categories that assign similar items to the same category. A simple measure of "coherence" counts the number of times that adjacent items are assigned to the same category: the categories on the left of Figure 5 have maximal coherence of six, whereas the incoherent categories on the right have coherence zero. To investigate GCM biases, we simulated the iterated learning experiment described above 100,000 times using different values of $\lambda$, assuming that all learners in a chain have the same $\lambda$. The results (Figure 6, left) show that the GCM bias for coherent categories is strongest for large values of $\lambda$.

Given that individual differences in categorization exist, we ran a second simulation study (Figure 6, right). This time

Category Coherence


Figure 6: Exploring the "category coherence" bias using iterated learning. The $y$ axis plots category coherence (defined in main text). Left panel: Category coherence assuming all participants share the same prior $(\lambda)$. Here there are three chains each reflecting one of the three $\lambda$ values. As $\lambda$ grows higher, iterated learning produces more coherent categories. The grey dashed line reflects the average of the three chains on iteration 15 . Right panel: When there are individual differences within participants, the learners all become somewhat more similar to one another but the effect is small.
we mixed learners that varied in their $\lambda$ values (sampling uniformly at random from $0.1,1$ and 10 ) into a single chain to investigate the effect heterogeneity has on each learner type. Unlike our previous simulations, the heterogeneity of the chain did not distort any of the three GCM learner types to a large extent: the right hand side of Figure 6 is not too dissimilar to the left. Based on this, one might conclude that the heterogeneity of the population has done very little to distort the categorization schemes produced by the various different learners. Unfortunately, this conclusion is unwarranted.

## Category size bias

Categorization is complex, and even this simple problem involves multiple biases. A preference for coherent categories is one kind of bias that a learner might express, but one might be just as interested in exploring the extent to which learners prefer categories to be of similar size. Does the GCM have a bias to split items evenly or unevenly? Does it depend on $\lambda$ ?

To that end, we counted the number of exemplars assigned to the smaller category in our previous simulations. Figure 7 plots this for the three homogeneous chains (left) and the single heterogeneous chain (right). The left panel shows that the GCM has a bias to prefer unevenly sized categories: this bias is weak when the learner generalizes narrowly $(\lambda=10)$, and strong when the learner generalizes widely $(\lambda=0.1)$. Unfortunately, almost none of this differentiation is evident when we look at the heterogeneous chains: the average response is substantially different from when the three learner types were taken separately, and there are almost no individual differences to be found, with all three learner types producing similar responses. With respect to the category size bias, mixing different learners into the iterated learning chain has almost completely erased their differences.

## Category Size



Figure 7: Exploring the "category size" bias using iterated learning. The $y$ axis plots the number of items assigned to the smaller category. Left panel: Homogenous iterated learning chains when all learners use the same value of $\lambda$. The three plots in the figure are quite dissimilar: when $\lambda$ is small the GCM strongly prefers an unequal allocation of items to categories, but when $\lambda$ is large the preference is weak. The grey dashed line reflects the average of the three chains on iteration 15. Right panel: When the same GCM learners are mixed into a heterogenous iterated learning chain, most of this variation is suppressed (the curves are close to each other), and the average size of the smaller category (grey dashed line) has substantially decreased.

## General discussion

The three case studies all display the same pattern. When all learners bring the same inductive bias to the problem, iterated learning behaves in the way that previous theoretical proofs suggest it should (Griffiths \& Kalish, 2007). In particular, when learners are Bayesians with identical priors and correctly specified likelihoods, iterated learning reveals those priors. For a non-Bayesian learner an analogous inductive bias is uncovered. However, when learners bring different biases to the problem there is no guarantee that the responses of any one participant genuinely reflects their prior biases, nor is there any guarantee that the average responses reflect the average bias in the population. To the contrary, our case studies suggest that those learners with the most extreme biases exert a disproportionate influence on the chain. We briefly consider the implications if this pattern holds more generally.

Iterated learning leads a double life within the psychological literature. As a theoretical tool, the underlying dynamics of the chain provide valuable insights into how cultural and linguistic evolution works. From that perspective, our results open up new questions: for instance, does language evolution reflect the cognitive biases of all speakers, or do some subpopulations (e.g., children) exert stronger influences on the process? Similarly, learners with the most confidence in their own beliefs will exert a disproportionate influence on others, providing a justification for expressing overconfidence: if the goal is to have cultural influence rather than be correct, strong biases are better than weak ones. Regardless, the effect of heterogeneity in this context need not be a reason for concern so much as a reason to ask new questions.

On the methodological side, iterated learning has often
been used as a tool for exploring the inductive biases of individuals. Based on formal results suggesting that the stationary distribution of an iterated learning chain is the prior, researchers in cognitive science have sometimes used these designs as a form of elicitation task, in which the (betweensubject) distribution of responses is taken to be reflective of some (within-subject) latent mental representation of the world. In this context, our results suggest that some care is required. When people bring different priors to a task, there is no inherent reason to think that the stationary distribution of an iterated learning chain reveals those priors. The distortions are both systematic and difficult to predict. The latter point is especially troublesome from a methodological perspective. In our third case study, it was not obvious to us that heterogeneity among category learners would produce a large distortion of "category size" biases, but almost no distortion to the bias for "coherent" categories. In this context, we suggest that the interpretation of iterated learning experiments is difficult when individual differences exist.

## References

Canini, K., Griffiths, T., Vanpaemel, W., \& Kalish, M. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. Psychonomic Bulletin \& Review.
Ellis, R. (2015). Understanding second language acquisition (2nd). Oxford University Press.
Ferdinand, V., Thompson, B., Kirby, S., \& Smith, K. (2013). Regularization behavior in a non-linguistic domain. In Proceedings of the 35th Annual Conference of the Cognitive Science Society.
Griffiths, T. \& Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. Cognitive Science, 31(3), 441480.

Hudson Kam, C. \& Newport, E. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. Language Learning and Development, 1(2), 151-195.
Janis, I. L. (1982). Groupthink: psychological studies of policy decisions and fiascoes. Houghton Mifflin Boston.
Kalish, M., Griffiths, T., \& Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. Psychonomic Bulletin \& Review, 14(2), 288-294.
Kirby, S., Dowman, M., \& Griffiths, T. (2007). Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences, 104(12), 5241-5245.
Kirby, S., Griffiths, T., \& Smith, K. (2014). Iterated learning and the evolution of language. Current Opinion in Neurobiology, 28C(108-114).
Lew, T. \& Vul, E. (2015). Structured priors in visual working memory revealed through iterated learning. In Proceedings of the 37th Annual Conference of the Cognitive Science Society, Austin, TX. Cognitive Science Society.
Lewandowsky, S., Griffiths, T., \& Kalish, M. (2009). The wisdom of individuals: exploring people's knowledge about everyday events using iterated learning. Cognitive Science, 33, 969-998.
Nosofsky, R. M. (1986). Attention, similarity, and the identificationcategorization relationship. Journal of Experimental Psychology: General, 115(1), 39-57.
Perfors, A. \& Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. Cognitive Science, 38(4), 775-793.
Rafferty, A., Griffiths, T., \& Klein, D. (2014). Analyzing the rate at which languages lose the influence of a common ancestor. Cognitive Science, 38, 1406-1431.
Reali, F. \& Griffiths, T. (2009). The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. Cognition, 111, 317-328.

# Calculating Probabilities Simplifies Word Learning 

Aida Nematzadeh<br>Dept. of Psychology<br>Univ. of Calif., Berkeley<br>nematzadeh@berkeley.edu

Barend Beekhuizen<br>Dept. of Computer Science<br>University of Toronto<br>barend@cs.toronto.edu

Shanshan Huang<br>Dept. of Computer Science<br>University of Toronto<br>sunny.huang@mail.utoronto.ca

Suzanne Stevenson<br>Dept. of Computer Science<br>University of Toronto<br>suzanne@cs.toronto.edu


#### Abstract

Children can use the statistical regularities of their environment to learn word meanings, a mechanism known as crosssituational learning. We take a computational approach to investigate how the information present during each observation in a cross-situational framework can affect the overall acquisition of word meanings. We do so by formulating various in-the-moment learning mechanisms that are sensitive to different statistics of the environment, such as counts and conditional probabilities. Each mechanism introduces a unique source of competition or mutual exclusivity bias to the model; the mechanism that maximally uses the model's knowledge of word meanings performs the best. Moreover, the gap between this mechanism and others is amplified in more challenging learning scenarios, such as learning from few examples. Keywords: cross-situational word learning; computational modeling; word learning biases


## Introduction

How do people acquire the meanings of words as they begin to learn a language? A well-supported proposal is crosssituational learning (e.g., Pinker, 1989), which suggests that people are sensitive to the regularities that repeat in different situations, and use such evidence to identify the commonalities, from which they can infer word meanings. As an example, when a child hears what a cute kitty, be nice to the kitty, etc., she/he could infer that the word kitty refers to the common referent in all these situations, i.e., a cat. Recent word learning experiments confirm that both adults and infants keep track of cross-situational statistics across learning trials, and infer the correct word-meaning mappings even in highly ambiguous conditions (e.g., Yu \& Smith, 2007; Smith \& Yu, 2008; Yurovsky, Fricker, Yu, \& Smith, 2014).

Despite empirical evidence for statistical cross-situational learning, the exact mechanisms in play are still not fully understood. In this paper, we focus on the first step of a crosssituational framework - the learning that occurs on each observation of a word, which we call in-the-moment learning. Given the words in an utterance and their potential meanings in the accompanying situation, there are many possible ways to associate words and meanings, but only some of these associations are correct. We refer to these in-the-moment associations of words and meanings as alignments, and consider different strategies for assessing the strength of these alignments, drawing on the evolving knowledge of word meanings. We note that previous research has considered "hard" (or binary) in-the-moment learning strategies, where an alignment is either considered by the learner or not (e.g., Trueswell, Medina, Hafri, \& Gleitman, 2013); we instead examine "soft" strategies where alignments have strengths between zero and one.

We formulate various in-the-moment learning mechanisms that introduce different kinds of competition - i.e., the way in which the strength of a word-meaning alignment depends on and interacts with other possible alignments. Each mechanism corresponds to certain statistics of the word learning input, such as the weighted frequency of word-meaning pairs or their conditional probabilities. We show that the different types of competition lead to various kinds of mutual exclusivity behaviours. Mutual exclusivity has been proposed as an explicit bias, in which children assume each word has a single meaning (e.g., Markman, 1987; Markman \& Wachtel, 1988). Here, mutual exclusivity of words and/or meanings arises from competition in a way that focuses learning.

We take a computational modeling approach to investigate the effectiveness of these mechanisms in overall acquisition of word meanings in various long-term word learning scenarios. Using a computational model enables us to explore the impact of different learning mechanisms in a variety of conditions, and to examine the role of one factor (e.g., frequency) while controlling for another one (e.g., utterance length). We find that the mechanism that maximizes the use of the accumulated knowledge of learned meanings performs the best. Interestingly, the performance gap between this mechanism and others is most significant in more difficult learning conditions, such as learning of low frequency words given long utterances. This shows that using conditional probabilities (as opposed to counts) and introducing competition (leading to a mutual exclusivity bias) improves overall word learning and might be necessary to guide learning in the presence of ambiguity or little data.

## A Cross-situational Word Learning Framework

There has been an increased interest in the last decade in developing computational models as tools to study word learning in people. Of particular interest are cross-situational learners that are incremental (e.g., Siskind, 1996; Fazly, Alishahi, \& Stevenson, 2010; Kachergis, Yu, \& Shiffrin, 2012), which is necessary in studying developmental learning patterns. Notably, the model of Fazly et al. $(2008 ; 2010)$ (henceforth FAS) is the first probabilistic model that robustly predicts a range of observed behavior in child word learning. Moreover, this model has been adopted and extended by a series of successive work (e.g., Nematzadeh, Fazly, \& Stevenson, 2012a; Grant, Nematzadeh, \& Stevenson, 2016), demonstrating its robustness in accounting for empirical data. We adopt the FAS word learning framework to examine various in-the-moment learning mechanisms.

## The FAS Model

Word learning input and output. The model's input is a sequence of utterance-scene pairs simulating what the child hears and perceives, respectively. Each utterance is a set of words (ignoring their order), and the corresponding scene is a set of semantic features that represents possible meanings of words in the utterance (see Ex. 1). Word meanings are represented by multiple features, which exposes the model to naturalistic commonalities among the words.

> Utterance: $\{$ Joel, eats $\}$
> Scene: $\{$ PERSON, JOEL, ACT, CONSUME, $\ldots\}$

The output of the model, at each step in learning, is the current representation of the meaning of each word $w$ as a probability distribution, $p(\cdot \mid w)$, over all possible semantic features $f$ that the model has observed in the input scenes.
The word learning problem. Given a corpus of utterancescene pairs, the goal of the model is to learn the meaning probability distribution, $p(\cdot \mid w)$, for all words $w$. Prior to receiving any input, all features $f$ are equally likely for a word. As the model processes each input pair, the probability is adjusted to reflect the cross-situational evidence in the corpus, in two steps: (a) in-the-moment learning on this input pair and (b) update of the word meaning probabilities using the accumulated evidence over all inputs.
In-the-moment learning. Given an utterance and a scene, which features in the scene are part of a word's meaning? There are different possible ways to determine whether a semantic feature is associated with a word in the input pair, and the corresponding strength of that association. FAS assumes that each feature $f$ in scene $S_{t}$ at time $t$, independently of the other features, is aligned to all the words $w$ in the utterance $U_{t}$ with a particular strength (see Figure 1a):

$$
\begin{equation*}
a_{t}(w \mid f)=\frac{p_{t}(f \mid w)}{\sum_{w^{\prime} \in \mathrm{U}_{t}} p_{t}\left(f \mid w^{\prime}\right)} \tag{2}
\end{equation*}
$$

The alignment strength between a feature $f$ and word $w$ depends on the current probability that $f$ is part of the meaning of $w$-i.e., $p_{t}(f \mid w)-$ as well as the probabilities that $f$ is part of the meaning of other words in the utterance (the denominator above).

In this way, Eqn. (2) has words in the utterance "compete" to be associated with a given feature: a higher alignment strength of one word with a feature necessarily results in a lower alignment strength for other words with that feature. This can be interpreted as a directional mutual exclusivity bias: the alignment formulation limits the number of words a feature can be strongly associated with, but does not directly limit the number of features a word can be associated with.
Updating the word meanings. How is the information learned from an input pair incorporated into a learner's longterm knowledge of word meanings? The learner incrementally accumulates the alignment strengths between each $w$
and $f$ in an overall association score $\operatorname{assoc}(w, f)$, which is updated at each time $t$ that $w$ and $f$ co-occur in an input pair:

$$
\begin{equation*}
\operatorname{assoc}_{t}(w, f)=\operatorname{assoc}_{t-1}(w, f)+a_{t}(f \mid w) \tag{3}
\end{equation*}
$$

where $\operatorname{assoc}_{t-1}(w, f)=0$ if $w$ and $f$ have not co-occurred prior to $t$.

After updating the association scores, the meaning probability $p(\cdot \mid w)$ of each word $w$ in the current input is adjusted using a smoothed version of this formula:

$$
\begin{equation*}
p_{t+l}(f \mid w)=\frac{\operatorname{assoc}_{t}(f, w)}{\sum_{f_{j} \in \mathcal{M}} \operatorname{assoc}_{t}\left(f_{j}, w\right)} \tag{4}
\end{equation*}
$$

where $\mathcal{M}$ is the set of all features observed up to time $t$.
In Eqn. (4), the probability of a feature given a word is a normalization of their association score over all possible features, which introduces another source of competition, this time, among features for a given word. This competition can be thought of as a mutual exclusivity bias in the reversed direction of the alignment score in Eqn. (2); here a word can only be strongly associated to a limited number of features.

## Using Sets of Features as Referents

In FAS, an input scene is the set union of all meaning features for all words in the corresponding utterance. This representation lacks information that would be apparent to a child, namely that each set of meaning features belongs to a single entity or event - e.g., PERSON and JOEL, or ACT and CONSUME in Ex. 1. However, replacing such sets of features with a single symbol corresponding to the word meaning would prevent the model from learning semantic similarities among the words (e.g., Nematzadeh, Fazly, \& Stevenson, 2012b). Instead, following Alishahi, Fazly, Koehne, and Crocker (2012), we simply maintain each set of semantic features corresponding to the meaning of each word in the utterance, and we call these sets of features referents, as in Ex. 5: ${ }^{1}$

```
Utterance { Joel, eats, an, apple }
    Scene: {{PERSON, JOEL}, {ACT, CONSUME, ...},
    {SINGULAR, INDEFINITE, DETERMINER, ... },
    {APPLE, FRUIT, FOOD, ...} }
```

A scene is now a set of referents, each of which is a set of semantic features corresponding to the meaning of a word. In the FAS model, calculation of alignment strength between a word $w$ and feature $f$ at time $t$ uses the meaning probability $p_{t}(f \mid w)$. Now, aligning words with referents (as in 5) requires consideration of strength of alignment of a word with a set of features. In calculating alignment strength for a word $w$ and a referent $r$ at time $t$, we change the FAS model to consider $\operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}(r)\right)$, the cosine similarity between the word's current meaning representation and the representation of the referent, where $\mathrm{v}(r)$ and $\mathrm{v}_{t}(w)$ are vectors in which the elements are meaning features. For $\mathrm{v}_{t}(w)$, the value for

[^154]each component feature $f$ is $p_{t}(f \mid w)$. (I.e, $\mathrm{v}_{t}(w)$ is a vector corresponding to $p_{t}(\cdot \mid w)$.) For $\mathrm{v}(r)$, the element values are 1 for features present in the definition of $r$ and 0 otherwise. In this way, alignment strength for word $w$ and referent $r$ is influenced by the strength of the meaning probabilities $p(f \mid w)$ for all features $f$ that are part of the representation of $r$. In the remainder of the paper, we explore variations in how the alignment process actually does this, in ways that implement different types of mutual exclusivity biases.

## In-the-Moment Learning Mechanisms

Competition in the model. We observed above that the alignment strength calculation in Eqn. (2) instantiates a form of mutual exclusivity bias, because words are competing to be strongly associated with a feature during this in-the-moment learning process. With the change of aligning words to referents instead of to features, we have the opportunity to explore various ways to formulate competition in determining the strength of alignments. The three alignment formulations explored here implement (1) no competition among words or referents, (2) competition of referents for a word (as in Alishahi et al., 2012), and (3) competition of words for a referent (analogous to the competition of words for a feature in FAS). Each of these ways of viewing competition implements a different approach to mutual exclusivity in the model, and we will explore the resulting impact on word learning in the results.
No competition. The no-competition mechanism (henceforth, no-comp) serves as a baseline for comparison to the other two. It assumes no mutual exclusivity bias - all the alignments between words and referents are calculated independently, and the value of one alignment does not effect any of the others (see Figure 1b). We formulate such an alignment between a word $w$ and a referent $r$ as simply the similarity between $\mathrm{v}_{t}(w)$ and $\mathrm{v}(r)$ as described above:

$$
\begin{equation*}
a_{t}(w, r)=\operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}(r)\right) \tag{6}
\end{equation*}
$$

This formulation can be seen as a simple weighted count, where each feature relevant to $r$ (valued 1 in $\mathrm{v}(r)$ ) contributes to the overall alignment strength proportionally to the model's prior knowledge of its meaning probability with that word.
Referent competition. Here we adopt the alignment formulation of Alishahi et al. (2012), which we call "ref-comp" because referents compete for alignment with a word. This mechanism implements a directed mutual exclusivity bias in which each word has a preference to be strongly associated with one referent. In other words, referents in the scene compete for a given word, while the alignments of words are independent of each other (see Figure 1c). This preference can be implemented by normalizing the $\operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}(r)\right)$ over all the referents in the scene:

$$
\begin{equation*}
a_{t}(r \mid w)=\frac{\operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}(r)\right)}{\sum_{r^{\prime} \in \mathrm{S}_{t}} \operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}\left(r^{\prime}\right)\right)} \tag{7}
\end{equation*}
$$

By normalizing the weighted count of $\operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}(r)\right)$, this alignment formulation can be interpreted as the conditional probability of $r$ given $w$, rather than a simple count.
Word competition. Here, we consider a competition that is instead analogous to the competition of words for a feature in FAS; "word-comp" is the reverse of ref-comp, because here words compete for a referent. This leads to a directed mutual exclusivity bias, but in the opposite direction to ref-comp. The word-comp mechanism asserts a preference for each referent to be strongly associated with a single word, by having words compete for a referent, while the alignments of referents are independent of each other (see Figure 1d). This bias is formulated by normalizing the $\operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}(r)\right)$ over the words in the utterance (as FAS did):

$$
\begin{equation*}
a_{t}(w \mid r)=\frac{\operatorname{sim}\left(\mathrm{v}_{t}(w), \mathrm{v}(r)\right)}{\sum_{w^{\prime} \in \mathrm{U}_{t}} \operatorname{sim}\left(\mathrm{v}_{t}\left(w^{\prime}\right), \mathrm{v}(r)\right)} \tag{8}
\end{equation*}
$$

This formulation also yields a conditional probability, but here of $w$ given $r$.
The association score. We note one final change to the FAS model to deal with referents: We must modify Eqn. (3) to keep track of associations between a word $w$ and all the features of a referent $r$. Since a feature $f$ can occur in more than one referent in scene $S$, which can have multiple alignment scores, we use the maximum alignment score of a referent that contains the feature in updating the feature's association score:

$$
\begin{equation*}
\operatorname{assoc}_{t}(w, f)=\operatorname{assoc}_{t-1}(w, f)+\max _{r^{\prime} \in S: f \in r^{\prime}} a_{t}\left(w, r^{\prime}\right) \tag{9}
\end{equation*}
$$

The meaning probabilities in the model continue to be calculated between individual features and a word. Recall that the meaning probability distribution $p(\cdot \mid w)$, as a conditional probability over semantic features, enforces a competition among them for the probability mass.

## Experiments

## Set-up

The utterances in the input are child-directed speech taken from the Manchester corpus (Theakston, Lieven, Pine, \& Rowland, 2001) in CHILDES (MacWhinney, 2000). To create the associated scene representations, each word in the corpus is entered into a gold-standard lexicon with a set of semantic features representing its gold-standard meaning, following the procedure of Fazly et al. (2008). The referents shown in Ex. 5 correspond to the gold-standard meanings of each of those words. (The word-mapping in the lexicon is only used to generate scenes, and is not seen by the model.) The model is trained on 20 K utterance-scene pairs, at which point behaviour is stable.

In the following experiments, we examine the quality of the individual learned word representations in two ways: the average acquisition score of all words observed by the model, and the proportion of observed words that is learned. The


Figure 1: Types of alignment mechanisms. Lines of the same color/style compete simultaneously. Thickness indicates varying strength of alignment during a competition.


Figure 2: Developmental plots
acquisition score of each word $w$ is obtained by comparing the word meaning representation $\mathrm{v}(w)$ with a gold standard representation of the word gold $(w)$ using cosine similarity:

$$
\begin{equation*}
\mathbf{a c q}(w)=\operatorname{sim}(\mathrm{v}(w), \operatorname{gold}(w)) \tag{10}
\end{equation*}
$$

where $\operatorname{gold}(w)$ is a vector over all semantic features, with value 1 for features part of the gold-standard meaning of $w$ and 0 otherwise. An observed word counts as "learned" if its acq score is higher than some threshold $\theta$, here set to 0.7 based on the experiments of Fazly et al. (2010).

## Results

## Overall Learning Patterns

Over time, all models converge to high average acq scores (Figure 2a) and proportions of words learned (Figure 2b), but with substantial differences between them. Notably, we find that on the average acq score, the word-comp formulation performs better than the original FAS (. 96 vs. .86), while the ref-comp and no-comp models do not learn the representations as well (both .83).

Two factors may underlie the varying performance of the models: the semantic grouping of features into referents (distinguishing our models from FAS), and the type of in-themoment competition (and resulting type of mutual exclusivity). For the first factor, the word-comp mechanism provides the most direct comparison to FAS: it uses the same direction of bias - in which words compete to align with the elements of the scene - but using referents instead of features. The grouping into referents appears to improve learning. When
aligning features individually as in FAS, the correct features for a word may be aligned more or less strongly (depending on competition for each from other words), so that the overall meaning probability vector may not converge as easily to the full set of correct features. By contrast, when a word has a strong alignment with the correct referent - which corresponds to the gold-standard meaning of the word - all features of the referent are boosted in the meaning probability of the word, yielding improved learning in word-comp over FAS.

Second, we find an interesting asymmetry between the two mechanisms involving competition between the words for a referent (word-comp) and between the referents for a word (ref-comp). Each imposes a conditional probability formulation of competition, but word-comp performs much better, with ref-comp behaving no better than the no-comp model. In fact, the advantage of using referents instead of individual features is completely eliminated in both the no-comp and the ref-comp mechanisms, as both perform worse than FAS.

The source of this asymmetry, we believe, is the deployment of learned knowledge by the model. In both the nocomp and the ref-comp model (Figure 1(b), (c)), a learned word meaning is compared to the referents in isolation from the learned meanings of the other words in the utterance. In this set-up, the knowledge of other word meanings cannot help to guide the model to determine how good a word's alignment to some referent is. By contrast, the word-comp model (Figure 1(d)) tunes the alignments by comparing how similar various learned word meanings are to a referent.

One might expect that mutual exclusivity in the reverse direction (as in the ref-comp model) would achieve the same effects: Tuning the similarity between a word meaning and a referent by the similarity between that word meaning and all other referents should guide the model to correct associations more quickly than not doing so. However, we do not find this effect. We will return to the reason for this lack of effect in the section on the role of frequency.

Competition is clearly important in focusing alignments and facilitating learning, but only in the context of appropriately constraining information: the most effective learning occurs when the competition draws on the maximal amount of learned knowledge in the model, in the form of the developing meaning probabilities. In what follows, we consider the impact of increased ambiguity in forming alignments, or decreased knowledge about words, to see how these factors


Figure 3: Average acquisition score after 20K input items.
impact these various mechanisms. Because the proportion of words learned shows similar relative behaviours to the acq score, in the remaining analysis we focus on comparing acq scores of each of the models after 20 K inputs.

## The Role of Frequency

Children are able to learn word meanings in various conditions, sometimes after only a few observations. Previous research suggests that children use biases such as mutual exclusivity to guide their learning. Learning low-frequency words is also a challenge for computational models, and understanding the mechanisms that improve learning from little evidence can shed light on how children address this issue. The type of competition in our various models plays an important role in their performance on low-frequency words. Figure 3a shows that for the two models with competition over words - the FAS and word-comp models - there is no decrease in performance for words of low frequency $(<5)$ compared to high frequency ( $>10$ ), while for the other two models, no-comp and ref-comp, there is a dramatic drop off in learning.

Specifically, the competition among words in the FAS and word-comp models - which maximizes the use of learned knowledge in focusing alignments - appears to play a crucial role in enabling these models to learn low-frequency words. Comparing the alignments in Figure 1c and Figure 1d in the face of a novel word and its novel referent (as an extreme case of low frequency) will clarify the utility of the learned meaning probabilities. In the word-comp model (Figure 1d), the meaning probabilities of previously-seen words competing for a new referent will not have a very good match to the feature vector for the new referent (since their probabilities will have been adjusted to better fit referents they have been seen with). The novel word will have uniform meaning probabilities that will enable it to better match the new referent, and thus will have a stronger alignment than previously-seen words. By contrast, in the ref-comp model (Figure 1c), the uniform probabilities of the new word will equally match all the referents competing for it, whether they have been seen before or not. There is no prior knowledge in the model in this competition that indicates the previously-seen referents have a better fit with other words. Thus a competition among words works well for novel or low-frequency words by drawing on the fact that previously-seen words will not compete as strongly for a new(er) referent. In short: a new word can
in principle go equally well with any referent in the situation, but a new referent not equally well with any word in the utterance.

## The Role of Utterance Length

Above, we found that the different types of competition gave more pronounced results for low-frequency words than for high-frequency ones. Similarly, we can explore whether there is a differential impact of utterance length on the different models. To simulate this, we manipulated the input generation procedure so that the model was trained only on utterances of length 5 or higher (long-corpus), or 3 and lower (short-corpus). Looking at Figure 3b, we observe that the acquisition scores are globally lower when the models are trained on long sentences only, likely due to the fact that there is more uncertainty about which words and which referents belong together.

Here we see that the word-comp model is the only one to not substantially decline in performance when comparing learning on the short-corpus and long-corpus. While the competition over words seems to help the FAS and wordcomp models equally in dealing with low-frequency words, here the bundling of features into referents as in word-comp is also necessary for performance to be robust to the added ambiguity of long utterances. The FAS model cannot "scale up" to deal with the very long unstructured lists of features in the long-corpus input. This also explains why the model of Alishahi et al. (2012) (the ref-comp approach) worked well in their experiments but not here: the utterances they used all had two words, unlike the naturalistic data we train on above, indicating that ref-comp also cannot scale effectively. Interestingly, as shown in Figure 3c, the word-comp model is particularly robust to the challenge of learning low-frequency words in the corpus of longer utterances, with a very small decrease in performance compared to the other models.

## The Role of Referential Uncertainty

To explore the impact of referential uncertainty - the occurrence of many more potential referents in a scene than there are words - we create a subcorpus that uses every $i^{t h}$ utterance from our full corpus, and uses the utterances in between those to generate "extra" referents in the scenes for utterances in the subcorpus. Here we report results on 20 K inputs with referents added to each scene $S_{i}$ from 0,1 , or 2 utterances in addition to referents taken from utterance $U_{i}$. Figure 4 presents


Figure 4: Average acquisition score after 20K input items, split over different amounts of referential uncertainty.
the results for no referential uncertainty, along with the two added levels of uncertainty. As we expect, the learning performance of all models degrades with higher referential uncertainty. However, in contrast to our previous results, here there is little benefit from either word-based competition or feature bundling. The high degree of ambiguity introduced by these levels of referential uncertainty may be better dealt with by attentional mechanisms that focus joint attention on a likely subset of relevant referents prior to alignment.

## Conclusions and Future Work

Previous research shows that children are sensitive to the cross-situational statistics of their environment: i.e., they can use the regularities across different situations to learn word meanings. However, the detailed mechanisms responsible for cross-situational word learning are still not fully understood, such as precisely what information is used from each observation in identifying the correct word meaning, and how this information is incorporated in the accumulated knowledge about a word. Moreover, children are good at learning word meanings in a variety of situations: they can learn a novel word from a few example and also acquire words from ambiguous/noisy conditions. Previous research has suggested that children are equipped with biases that guide them in word learning by reducing the difficulty/ambiguity of a learning situation. The necessity of these biases in children, and whether they are innate or learnable, are issues that have been debated among cognitive scientists.

Here, we show that one such bias - the mutual exclusivity bias that limits the number of meanings a word takes can be modeled as a competition mechanism during in-themoment learning. The competition exists when the model assesses possible word and referent associations with conditional probabilities as opposed to counts. In other words, the bias or competition is a learning mechanism that is able to condition in-the-moment learning to the learned knowledge of word meanings. We observe that the role of the bias is particularly significant when the learning is more challenging: for example, for learning low-frequency words or from longer utterances. Previous research has investigated how cognitive processes such as memory and attention interact with crosssituational word learning (e.g., Nematzadeh et al., 2012a). Future work should study how these cognitive processes affect the in-moment-learning.

## References

Alishahi, A., Fazly, A., Koehne, J., \& Crocker, M. W. (2012). Sentence-based attentional mechanisms in word learning: evidence from a computational model. Frontiers in psychology, 3.
Fazly, A., Alishahi, A., \& Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In CogSci Proceedings.
Fazly, A., Alishahi, A., \& Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. Cognitive Science, 34(6), 1017-1063.
Grant, E., Nematzadeh, A., \& Stevenson, S. (2016). The interaction of memory and attention in novel word generalization: A computational investigation. In CogSci Proceedings.
Kachergis, G., Yu, C., \& Shiffrin, R. (2012). An associative model of adaptive inference for learning word-referent mappings. Psychonomic Bulletin and Review, 1-8.
MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (3rd ed., Vol. 2: The Database). Erlbaum.
Markman, E. M. (1987). How children constrain the possible meanings of words. In U. Neisser (Ed.), Concepts and conceptual development: Ecological and intellectual factors in categorization (Vol. 1, pp. 255-287). CUP.
Markman, E. M., \& Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20(2), 121-157.
Nematzadeh, A., Fazly, A., \& Stevenson, S. (2012a). A computational model of memory, attention, and word learning. In CMCL Proceedings (pp. 80-89).
Nematzadeh, A., Fazly, A., \& Stevenson, S. (2012b). Interaction of word learning and semantic category formation in late talking. In CogSci Proceedings (pp. 2085-2090).
Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure. Cambridge, Mass.: MIT Press.
Siskind, J. M. (1996). A computational study of crosssituational techniques for learning word-to-meaning mappings. Cognition, 61, 39-91.
Smith, L. B., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Theakston, A. L., Lieven, E. V., Pine, J. M., \& Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. Journal of Child Language, 28, 127-152.
Trueswell, J. C., Medina, T. N., Hafri, A., \& Gleitman, L. R. (2013). Propose but verify: Fast mapping meets crosssituational word learning. Cog. Psych., 66(1), 126-156.
Yu, C., \& Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414-420.
Yurovsky, D., Fricker, D. C., Yu, C., \& Smith, L. B. (2014). The role of partial knowledge in statistical word learning. Psychonomic Bulletin and Review, 21(1), 1-22.

# Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words 

Aida Nematzadeh, Stephan C. Meylan, and Thomas L. Griffiths<br>University of California, Berkeley<br>\{nematzadeh, smeylan, tom_griffiths $\}$ @ berkeley.edu


#### Abstract

Vector-space models of semantics represent words as continuously-valued vectors and measure similarity based on the distance or angle between those vectors. Such representations have become increasingly popular due to the recent development of methods that allow them to be efficiently estimated from very large amounts of data. However, the idea of relating similarity to distance in a spatial representation has been criticized by cognitive scientists, as human similarity judgments have many properties that are inconsistent with the geometric constraints that a distance metric must obey. We show that two popular vector-space models, Word2Vec and GloVe, are unable to capture certain critical aspects of human word association data as a consequence of these constraints. However, a probabilistic topic model estimated from a relatively small curated corpus qualitatively reproduces the asymmetric patterns seen in the human data. We also demonstrate that a simple co-occurrence frequency performs similarly to reduced-dimensionality vector-space models on medium-size corpora, at least for relatively frequent words.


Keywords: word representations; vector-space models; word associations

## Introduction

Finding good representations of the meaning of words is a fundamental problem in cognitive science and related disciplines. Vector-space models of semantics represent words as points in an $N$-dimensional Euclidean space where words with similar meanings are expected to be close together. These models have been successful in both modeling human semantic processing (e.g., Landauer and Dumais, 1997) and natural language processing applications (for a review, see Turney and Pantel, 2010). However, relating the similarity between words to their distance in a vector space means that these representations are subject to certain geometric constraints. Previous research has criticized this property of spatial representations because aspects of human semantic processing do not conform to these same constraints (e.g., Tversky, 1977). For example, people's interpretation of semantic similarity does not always obey the triangle inequality, i.e., the words $w_{1}$ and $w_{3}$ are not necessarily similar when both pairs of $\left(w_{1}, w_{2}\right)$ and $\left(w_{2}, w_{3}\right)$ are similar. While "asteroid" is very similar to "belt" and "belt" is very similar to "buckle", "asteroid" and "buckle" are not similar (Griffiths et al., 2007).

Recent work has resulted in significant advances in vectorspace models of semantics, making it possible to train models on extremely large datasets (Mikolov et al., 2013a; Pennington et al., 2014). The resulting vector-space modelsWord2Vec and GloVe-achieve state-of-the-art results for a wide range of tasks requiring machine representations of word meanings. However, the similarity between words in these models is typically measured using the cosine of the
angle between word vectors (e.g., Mikolov et al., 2013b; Pennington et al., 2014).

In this paper, we examine whether these constraints imply that Word2Vec and GloVe representations suffer from the same difficulty as previous vector-space models in capturing human similarity judgments. To this end, we evaluate these representations on a set of tasks adopted from Griffiths et al. (2007) in which the authors showed that the representations learned by another well-known vector-space model, Latent Semantic Analysis (Landauer and Dumais, 1997), were inconsistent with patterns of semantic similarity demonstrated in human word association data. We show that Word2Vec and GloVe suffer from similar problems. Recent probabilistic interpretations of Word2Vec (Levy and Goldberg, 2014; Arora et al., 2015) provide a way to construct a conditional probability from vector-space representations, although we show that this does not result in a significant improvement in performance over cosine similarity.

A probabilistic topic model performs less well than these vector-space models in predicting overall associations, but provides a better fit to human data on tasks where vectorspaced models are subject to geometric constraints. However, two advantages of the recent models are that they can produce word representations for very large vocabularies (millions of types) and can be trained on very large corpora (hundreds of billions of tokens). We investigate whether the performance of co-occurrence frequency-easily obtainable from large corpora-is comparable to the recent models. We find that vectors of simple co-occurrence frequency provide comparable performance to the above models, suggesting that dimensionality reduction may not be necessary feature for machine representations of words.

## Vector-Space Models

We first provide high-level descriptions of two recent vectorspace models that have received significant attention in the machine learning, natural language processing, information retrieval, and cognitive science communities.

## Word2Vec

Word2Vec (Mikolov et al., 2013b) is a shallow neural network model with a single hidden layer that learns similar vector representations for words with similar distributional properties. They present two variants: continuous bag of words or $C B O W$, in which a word token is predicted from its unordered context, and skipgram, in which a given word token is used to predict words in its context. Both variants perform well predicting associations, analogies, and can be used to
identify idiomatic multi-word phrases. We focus here on the skipgram formulation given its higher obtained performance in a variety of natural language processing tasks.

The objective of a Word2 Vec model is to maximize the average log probability of each word's context following

$$
\begin{equation*}
J=\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p\left(w_{t+j} \mid w_{t}\right), \tag{1}
\end{equation*}
$$

where $T$ is the number of training words and $c$ is the number of context words. $p\left(w_{t+j} \mid w_{t}\right)$ is given by the softmax function,

$$
\begin{equation*}
p\left(w_{o} \mid w_{i}\right)=\frac{\exp \left(v_{w_{o}}^{\prime \top} v_{w_{i}}\right)}{\sum_{w=1}^{W} \exp \left(v_{w}^{\top} v_{w_{i}}\right)}, \tag{2}
\end{equation*}
$$

where $W$ is the number of unique words (type) in the corpus $w_{1} \ldots w_{T}$, and $v_{w}$ and $v_{w}^{\prime}$ are the input and output vector representations of word $w$.

Computing the normalizing term in the softmax is prohibitively expensive for large datasets in that the cost of the computation is proportional to $W$ (which may be in millions), thus an approximation is obtained through hierarchical softmax (Morin and Bengio, 2005), Noise Contrastive Estimation (Gutmann and Hyvärinen, 2012), or a related novel technique they introduce, negative sampling. In negative sampling, the model updates the representations of a small number of words such that the network predicts an observed "positive" word pair (e.g., chicken salad), and does not predict any of a number of "negative" pairs that are unlikely to be observed in the text (e.g. chicken battleship or chicken advantageously). The negative pairs are drawn from an explicitly specified noise distribution, typically a unigram model. Because a small number of negative samples are used-usually fewer than 20-a relatively small number of weights need to be adjusted each time the model updates the representation of a word. Mikolov et al. find additional performance gains by sampling less from high frequency words.

Performance of Word2Vec model thus depends on the number of hidden units (typically 50-600), the size of the context window, the degree to which frequent words are undersampled, and the choice of approximation to the full softmax; if negative sampling is used then the number of negative samples can have a significant effect on performance.

## GloVe

GloVe (Pennington et al., 2014) is a weighted bilinear regression model that uses global co-occurrence statistics to derive a real-valued vector representation of each word. Like Word2Vec, GloVe learns similar vector representations for words that appear in similar contexts, however the latter model differs significantly in that it fits co-occurrence frequencies from an entire corpus rather than iterating through local context windows. GloVe exhibits particularly strong performance in analogy tasks, but also performs well on similarity tasks and named entity recognition (NER).

In GloVe, the best word representations $W$ and $\widetilde{W}$ are found by minimizing a least squares objective:

$$
\begin{equation*}
J=\sum_{i, j=1}^{V} f\left(X_{i j}\right)\left(w_{i}^{T} \widetilde{w}_{j}+b_{i}+\widetilde{b}_{j}-\log X_{i j}\right)^{2} \tag{3}
\end{equation*}
$$

where $V$ is the vocabulary, $i$ and $j$ pick out words in the vocabulary, $\mathrm{f}\left(X_{i j}\right)$ is a weighting term (explicated below), $w_{i}$ is the representation of the $i$ th word, $\widetilde{w}$ is the representation of the $j$ th word, $b_{i}$ and $\widetilde{b}_{j}$ are bias terms, and $\log X_{i j}$ is the co-occurrence count of words $i$ and $j$. If $X$ is symmetric, $W$ and $\widetilde{W}$ are equivalent (differing only according to their random initialization). GloVe additionally introduces a weighting into the cost function of the model to avoid $\log 0$ errors and to dampen the effect of high frequency co-occurrences:

$$
f(x)=\left\{\begin{array}{l}
\left(x / x_{\max }\right)^{\alpha} \text { if } x<x_{\max }  \tag{4}\\
1 \text { otherwise }
\end{array}\right.
$$

where $x$ is the co-occurrence count, and $\alpha$ allows for an exponential weighting of for counts between 0 and the threshold $X_{\max }$. The performance of a GloVe model thus depends on the dimensionality of the word vector (typically 50-300), $X_{\text {max }}, \alpha$, and the size of the window used to compute cooccurrence statistics around each word.

## Co-occurrence Frequency

We also consider a baseline model that simply uses normalized co-occurrence frequencies of words to measure their similarity. In other words, given sufficient data, is a term-byterm matrix sufficient to predict human association norms? We note that this baseline is used by previous work to model human semantic and syntactic processing, as well as in infomration retrieval (e.g., Burgess and Lund, 1997; Azzopardi, 2005).

## Shortcomings of Spatial Models

Similarity between two words in a vector-space model is usually computed using the cosine of the angle or the Euclidean distance between the vectors representing the words. While intuitive, this approach has at least one significant shortcoming: cosine and Euclidean distance cannot capture the observed asymmetries in human similarity judgments because they are inherently symmetric measures. Tversky (1977) famously argued that spatial representations cannot capture human similarity judgments because the latter often violate the metric axioms. For example, elicited word (or phrase) similarity is asymmetric: when queried, most participants considered "North Korea" to be very similar to "China," while the reverse relationship was rated as significantly less strong ("China" is not very similar to "North Korea").

Griffiths et al. (2007) extended this argument to spatial representations of the semantic relationships between words, showing that similar violations of the metric axioms can be demonstrated for vector-space representations. We now revisit these analyses, examining the extent to which they are problematic for new vector-space models.

One of the properties of metric spaces is that the distance between each 3 -word tuple must satisfy the triangle inequality: given three points $x, y$, and $z, d(x, z) \leq d(x, y)+d(y, z)$, where $d()$ is a distance function. This inequality constrains the possible distance values among the vector representations for each of the three words: if distances between the words in two of the pairs are very small, the distance between the words in the third pair is also expected to be small.

After demonstrating that cosine-as a monotonic function of the angle between two vectors-satisfies an analogue of the triangle inequality Griffiths et al. (2007) studied to what extent this is true among the cue-target pairs in the Nelson norms. For the words $w_{1}, w_{2}$, and $w_{3}$, they plot the distribution of $p\left(w_{3} \mid w_{1}\right)$ when both $p\left(w_{2} \mid w_{1}\right)$ and $p\left(w_{3} \mid w_{2}\right)$ are greater than a given threshold $\tau$. They observe that even for large values of $\tau$, there are a lot of very small values of $p\left(w_{3} \mid w_{1}\right)$. Consistent with the intuition that human similarity judgments are not always transitive, they find many cases where two of the pairs $\left(w_{2}-w_{1}\right.$ and $\left.w_{3}-w_{2}\right)$ in a tuple ( $w_{1}$, $w_{2}$, and $w_{3}$ ) are highly similar, but the words in the third pair ( $w_{1}$ and $w_{3}$ ) are not.

As a result, by using cosine (or any distance measure more generally) on vector-space representations, we cannot replicate the asymmetric patterns of similarities observed in human judgments. To enable word representations derived with vector space models to account for a greater range of phenomena, we propose an elaborated, non-metric similarity measure for vector-space representations. Following recent work that provides a probabilistic interpretation of Word2Vec (Levy and Goldberg, 2014; Arora et al., 2015) we calculate the conditional probability for a given pair of words $w_{1}$ and $w_{2}$ using a softmax function:

$$
\begin{equation*}
p\left(w_{2} \mid w_{1}\right)=\frac{\exp \left(\mathbf{w}_{\mathbf{2}} \cdot \mathbf{w}_{\mathbf{1}}\right)}{\sum_{\mathbf{w}_{\mathbf{j}}} \exp \left(\mathbf{w}_{\mathbf{j}} \cdot \mathbf{w}_{\mathbf{1}}\right)} \tag{5}
\end{equation*}
$$

where $\mathbf{w}_{\mathbf{j}}$ is the vector representation of $w_{j}$ and $w_{2} . w_{1}$ is the dot product of the two vectors. Using this probabilistic measure, we can now examine how well Word2Vec and GloVe representations perform on tasks that do not satisfy the geometric constraints, i.e., triangle inequality and asymmetries in similarity judgments.

## Evaluating Vector-Space Representations

In this section, we describe the evaluation data and explain the tasks that we use to examine how well vector-space representations predict human word associations.

## Data: Nelson Association Norms

Following Griffiths et al. (2007), we use the association norms from Nelson et al. (1998) as our gold-standard evaluation data. Nelson et al. (1998) performed an extensive free association experiment where they asked 6000 participants to record the first word they can think of given a cue word. The experiment resulted in a set of 5018 cues, the target words produced in response of each cue (associates), and the probability of producing each target word for a given cue. Ap-
proximately $45 \%$ of the target words are present as cues in the dataset. The Nelson norms are well-suited for the evaluation of semantic similarity because unlike most gold-standard similarity lexicons (e.g., Hill et al., 2015), word associations obtained in this way potentially encode asymmetric relations: the Nelson association norms encode for many words both how likely people are to produce $w_{1}$ when cued with $w_{2}$, as well as $w_{2}$ when cued with $w_{1}$.

## Evaluation Tasks

We evaluate the word representations found by these models on four tasks to assess whether they capture empirical phenomena of interest in the Nelson norms. The first two, coefficient of correlation and median rank of associates, test whether these representations capture the strength of associations between each cue-target pair. The remaining two, the triangle inequality and ratio of asymmetries specifically test whether these representations can account for human behavior on tasks with asymmetric associations.
Coefficient of correlation. Computing the correlation between two list of scores is a standard way for measuring their similarity (Budanitsky and Hirst, 2006). We created a gold-standard list of similarity scores that, for each cuetarget pair in the norms, includes $p$ (target|cue). We then retrieved a list of similarities for the same cue-target pairs from the representations under study, measuring similarity as either cosine $\left(\mathbf{w}_{\text {target }}, \mathbf{w}_{\text {cue }}\right)$ or $p\left(\mathbf{w}_{\text {target }} \mid \mathbf{w}_{\text {cue }}\right)$, where $\mathbf{w}_{x}$ is the vector representation of $x$. To assess the extent to which these representations can predict human similarity judgments of semantic associations, we calculated the Spearman's rank correlation coefficient ( $\rho_{\text {assoc }}$ ) between these two lists.
Median rank of associates. We also assess the quality of the representations by checking whether they produce similar rankings of target words (associates) for each cue in the Nelson norms. For each cue, we rank all its associates based on their conditional probabilities (given the cue) from the Nelson norms, and also get a similar ranking for each cue in the model. For the first associate of each cue, i.e., the one with the highest probability per the Nelson norms ranking, we check its rank in the model list. We take the median rank of the first associate across all the cues from the Nelson norms, and repeat this process for second and third associates.
Triangle inequality. We extend the analysis in Griffiths et al. (2007) to the evaluate whether word representations satisfy the triangle inequality. For every $w_{1}, w_{2}$, and $w_{3}$ such that similarity of $w_{1}-w_{2}$ and $w_{2}-w_{3}$ are greater than a threshold $\tau$, we plot the distribution of similarity values of $w_{1}-w_{3}$. For the Nelson norms, similarity of words in a pair is their conditional probability; for other models similarity is given by the cosine or conditional probability. We select thresholds $(\tau)$ such that for each threshold, the number of pairs selected for each model is similar to that of the Nelson norms; The thresholds for the norms are taken from Griffiths et al. (2007).
Asymmetry ratio. Griffiths et al. (2007) show that the similarity of more than $85 \%$ of cue-target pairs in Nelson norms
are asymmetric by the criterion of at least an order of magnitude difference between $p(w 2 \mid w 1)$ and $p(w 1 \mid w 2)$. However, distance measures are inherently symmetric and for any distance function $d()$, we have $d\left(w_{1}, w_{2}\right)=d\left(w_{2}, w_{1}\right)$. To measure the performance of vector-space representations in predicting the asymmetries, for each cue-target pair in the Nelson norms, we calculate the ratio of asymmetry as follows:

$$
\begin{equation*}
\operatorname{asym}\left(w_{1}, w_{2}\right)=\frac{p\left(w_{2} \mid w_{1}\right)}{p\left(w_{1} \mid w_{2}\right)} \tag{6}
\end{equation*}
$$

We then calculate the Spearman's rank correlation coefficient between the asymmetry scores of these similarities and those from the Nelson norms.

## Corpora and Model Training

To support comparison with Griffiths et al. (2007) we trained GloVe, Word2Vec skipgram, and collected co-occurence frequencies on TASA, the Touchstone Applied Sciences Corpus (Landauer and Dumais, 1997). This corpus consists of approximated 8 M tokens taken from reading materials appropriate for a high school English students. In addition to TASA, we trained Word2Vec skipgram and GloVe, and collected cooccurrence frequencies on English Wikipedia (3.91B tokens). This corpus is too large for training a Latent Dirichlet Allocation (LDA) topic model using Gibbs Sampling. While we tried to replicate the LDA results for TASA with more scalable variational methods (Hoffman et al., 2010), the resulting topics produced associations that were significantly worse than those obtained through Gibbs sampling or either of the vector space models.

Preprocessing was matched to the extent possible across model inputs. All words were translated to their nearest lowercase ASCII equivalent. For both TASA and Wikipedia we discarded function words using the Python stopwords package. For TASA we removed the same set of lowinformation words and enforced the same frequency cutoff as Griffiths et al. (2007). For Wikipedia, we removed words that appeared on too many pages or too few, and retained only the top 100k most frequent remaining words.

To evaluate the performance of the Word2Vec skipgram model we trained 20 models across a range of hyperparameter settings, varying the size of the embedding vector (50, $100,200,300$ or 400 hidden units), the choice of optimization method (hierarchical softmax or negative sampling), and for models with negative sampling the number of negative examples $(5,10,15)$. Words with unigram probability higher than .001 are downsampled following Mikolov et al. (2013b).

Because of an implementation error, we were unable to explore a large parameter space with GloVe, and report only the results with the default parameters $\left(X_{\max }=10, \alpha=.75\right.$, 50-dimensional vectors, and a 7 -word symmetric window on either side of the target word). This leaves open the possibility that GloVe may exhibit even higher performance on TASA and Wikipedia with appropriate parameter settings.

We also compute association using the LDA results (sampled document-topic and topic-word assignments) from Griffiths et al. (2007).

Finally, we used large-scale pre-trained models distributed by the authors of Word2Vec and GloVe. These largestavailable models often exhibit best-in-class performance because they reflect extensive parameter search, proprietary corpora, and distributed implementations that can handle more training data than publicly-distributed single-machine implementations. For Word2Vec we used a pre-trained 300dimensional model obtained by using the continuous bag of words architecture (CBOW) on a corpus of 100 billion words from Google News. For GloVe we used a 300-dimensional model trained by Pennington et al. (2014) using a 2014 export of Wikipedia and the Gigaword 5 corpus, consisting of approximately 6 billion tokens in total. ${ }^{1}$

## Results

Overall associations. We first look at the coefficient of correlation that shows how the various models perform in predicting the overall associations. We find that using conditional probability in place of cosine results in slightly better performance in predicting the semantic associations when the models are trained on medium or large corpora (see cosine ("cos.") and conditional probability ("cond. pr.") columns in Table 1). We also observe that given small and medium corpora (first and second row of Table 1), the Word2Vec skip-gram has the highest correlation with human word associations; but, given the largest corpora, GloVe performs slightly better than the Word2Vec model. Interestingly, given the small and medium corpora, simple co-occurrence frequencies perform similarly to or better than the Word2Vec CBOW and GloVe representations. Looking at the second measure of associations, the median rank of the associates (Table 2), we observe that the LDAmodel and co-occurrence frequencies perform similar to Word2Vec on TASA and Wikipedia; both models exhibit better performance than GloVe. The representations of the pretrained GLoVe model (on the largest corpus) have the lowest (best) median ranks.
Geometric constraints. The results for the triangle inequality analysis using the conditional probability measure are shown in Figure 1 (cosine results are omitted as they cannot produce the pattern). We observe the expected pattern for the Nelson norms, the LDA model, and co-occurrence frequency (see Figure $1 \mathrm{a}-\mathrm{c}$ ): even for large values of $\tau$, there are a lot of pairs that have probabilities close to zero. However, as shown in Figure 1 d-f, we do not see pairs with very small values of similarity when examining large thresholds for any of the recent vector-space representations. Our results reveal that even with a probabilistic measure, Word2Vec representations cannot predict the triangle inequality: for very high thresholds on the similarity of $w_{1}-w_{2}$ and $w_{2}-w_{3}$, there are

[^155]Table 1: The Spearman's rank correlation coefficient ( $\rho_{\text {assoc }}$ ) between gold-standard association scores from Nelson norms and different models of word representations. "cos." and "cond. pr." refer to cosine and conditional probability, respectively. [*] Data unavailable or infeasible to compute given current resources.

|  | Word2Vec CBOW |  | Word2Vec skip-gram |  | GloVe |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | cos. | cond. pr. | cos. | cond. pr. | cos. | cond. pr. | LDA | Co-occurrence |
| Small (TASA) | .22 | .21 | $\mathbf{. 2 5}$ | $\mathbf{. 2 5}$ | .21 | .20 | .20 | .21 |
| Medium (Wikipedia) | .22 | .22 | .23 | $\mathbf{. 2 4}$ | .16 | .19 | $\left[{ }^{*}\right]$ | .20 |
| Largest available | .25 | .26 | $[*]$ | $\left[{ }^{*}\right]$ | .24 | $\mathbf{. 2 7}$ | $[*]$ | $\left[{ }^{*}\right]$ |

Table 2: The median rank of first, second, and third associates ( $1^{s t} / 2^{n d} / 3^{r d}$ ) for different models of word representation using conditional probabilities. The number of possible targets is 3951 for all corpora.[*] Data unavailable or infeasible to compute given current resources.

|  | Word2Vec CBOW | Word2Vec skip-gram | GloVe | LDA | Co-occurrence |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Small (TASA) | $48 / 112 / 160$ | $26 / 72 / 106$ | $56 / 138 / 215$ | $23 / 69 / 103.5$ | $21 / 58 / 122$ |
| Medium (Wikipedia) | $23 / 48 / 75$ | $21 / 46 / 74$ | $52 / 92 / 129$ | $\left[{ }^{*}\right]$ | $23 / 48 / 70$ |
| Largest available | $13 / 29 / 47$ | $\left[^{*}\right]$ | $\mathbf{1 1 / 2 5 / 4 0 . 5}$ | $\left[^{*}\right]$ | $\left[{ }^{*}\right]$ |

no $w_{1}-w_{3}$ pairs with low similarity. These results suggest that using a probabilistic measure do not address the limitations of the vector-space models with respect to the triangle inequality.

Finally, we examine whether the representations capture the observed asymmetry in human similarity judgements as calculated in Eqn. (6). Note that we can only use conditional probabilities in this analysis because the cosine measure is symmetric. This probabilistic measure of similarities in both Word2Vec and GloVe to some extent predicts the asymmetric patterns of similarity observed in the Nelson norms (Table 3). We observe that the performance of the LDA model is comparable to the GloVe representations trained the largest corpora. The GloVe models performs significantly better than the Word2Vec models, which we believe is a result of its objective function-it uses the ratio of conditional probabilities of word pairs in training.

Table 3: The Spearman's rank correlation coefficient ( $\rho_{\text {asym }}$ ) between asymmetry scores of Nelson norms and representations from the models. In our data, there are 7096 cue-target pairs for which target-cue also exits. [*] Data unavailable or infeasible to compute given current resources.

|  | Word2Vec <br> CBOW | Word2Vec <br> Skipgram | GloVe | LDA |
| :---: | :---: | :---: | :---: | :---: |
| Small | .18 | .01 | .32 | $\mathbf{. 4 9}$ |
| Medium | .20 | .19 | .43 | $\left[^{*}\right]$ |
| Largest avail. | .20 | $[*]$ | .48 | $\left[^{*}\right]$ |

## Discussion

The selection of models, corpora, and tasks presented above suggests that LDA and co-occurrence frequencies have certain advantages when compared with the vector-space representations produced by Word2Vec and GloVe. We expound on a few key points below to contextualize our results and set the stage for future research.

Most of the targets and queues analyzed here are of rel-
atively high frequency rank. In future work we would like to investigate exactly how robust each of these models are to sparsity to test the hypothesis that reduced-dimensionality models are better at generalizing, such that they better predict associations for low frequency words.

The two vector-space models investigated here were both developed with the explicit objective of capturing meaningful linguistic difference in the linear substructure of the model (e.g., the vector produced by king - man + woman is closest to queen). As such, these models show strong performance on analogy tasks, while LDA typically fairs poorly. One question is thus whether a single representation could predict word associations, while preserving linear substructure.

## Conclusion

We show that representations from two new vector-space models, Word2Vec and GloVe, suffer from the same geometric constraints as predecessors, and are consequently unable to predict some of the characteristics of human similarity judgments, such as asymmetric similarity relations between two words or triangle inequality. Besides performing well in the above task, word representations derived from LDA topic modeling show remarkable predictive power with respect to human judgments given that they are learned from a dataset two orders of magnitude smaller than comparably performing vector-space models.

## References

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Randwalk: A latent variable model approach to word embeddings. arXiv preprint arXiv:1502.03520, 2015.
M. M. Azzopardi, L.;Girolami. Probabilistic hyperspace analogue to language. In Proceedings of the 28th Annual ACM Conference on Research and Development in Infomration Retrieval (SIGIR 2005), pages 575-576, 2005.
A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1):13-47, 2006.


Figure 1: The triangle inequality histograms on TASA: conditional probability for the third pair of words in a tuple ( $w_{1}-w_{3}$ ) when the first two pairs $\left(w_{1}-w_{2}\right.$ and $\left.w_{2}-w_{3}\right)$ are above the given threshold.
C. Burgess and K. Lund. Modelling parsing constraints with high-dimensional context space. Language and cognitive processes, 12(2-3):177-210, 1997.
T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. Psych. Rev., 114(2):211, 2007.
M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research, 13:307-361, 2012.
F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 2015.
M. Hoffman, D. M. Blei, and F. R. Bach. Online learning for latent dirichlet allocation. In Advances in Neural Information Processing Systems, pages 856-864, 2010.
T. K. Landauer and S. T. Dumais. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, pages 211-240, 1997.
O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In Advances in Neural Information Processing Systems 27, pages 2177-2185. 2014.
T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. 2013a.
T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111-3119. 2013b.
F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In R. G. Cowell and Z. Ghahramani, editors, Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, pages 246-252. Society for Artificial Intelligence and Statistics, 2005.
D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The university of south florida free association, rhyme, and word fragment norms. 1998.
J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing, pages 1532-1543, 2014.
P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37(1):141-188, 2010.
A. Tversky. Features of similarity. Psychological Review, 84 (4):327, 1977.

# Listeners integrate speech, gesture, and discourse structure to interpret the temporal structure of complex events 

Andie Nishimi', Esther Walker ${ }^{2}$, Benjamin K. Bergen ${ }^{2}$<br>\{anishimi, e1walker, bkbergen\}@ucsd.edu<br>${ }^{1}$ Psychology and ${ }^{2}$ Cognitive Science, University of California, San Diego

Tyler Marghetis<br>tmarghet@indiana.edu<br>Psychological and Brain Sciences, Indiana University, Bloomington


#### Abstract

Human communication has a remarkable capacity to describe events that occurred elsewhere and at other times. In particular, when describing complex narratives, speakers must communicate temporal structure using a mixture of words (e.g., "after"), gestures (e.g., pointing rightward for a later event), and discourse structure (e.g., mentioning earlier events first). How do listeners integrate these sources of temporal information to make sense of complex narratives? In two experiments, we systematically manipulated gesture, speech, and order-of-mention to investigate their respective impacts on comprehension of temporal structure. Gesture had a significant effect on interpretations of temporal order. This influence of gesture, however, was weaker than the influence of both speech and order-of-mention. Indeed, in some cases, order-of-mention trumped explicit descriptions in speech; for instance, if 'earlier' events were mentioned second, they were sometimes thought to have occurred second. Listeners integrate multiple sources of information to interpret what happened when.


Keywords: time; gesture; iconicity; multimodal communication; memory.

## Introduction

Human communication stands out among naturally occurring communication systems in its ability to convey information about events occurring in other places and at other times, a feature known as displacement (Hockett, 1960). This includes the concrete details of displaced events-who did what to whombut also when things occurred. If you observed a woman receiving the winning lottery ticket and also getting her purse stolen, then you would want to be clear about which event occurred first. While temporal order is an abstract feature of a complex event, it is often critical for communicative success.

To communicate about temporal order (and to communicate in general), speakers have several
strategies to deploy. The first and most obvious is in their choice of words, like "before" or "after," "earlier" or "later." Second, speakers also communicate about temporal order using visible and systematic motion of their bodies (Cooperrider \& Nunez, 2009; Casasanto \& Jasmin, 2012). Spontaneous co-speech gestures produced by North-American native English speakers often indicate relative temporal order by locating events along an imagined spatial timeline, with earlier events placed to the left and later events placed to the right. Finally, speakers encode temporal order in the structure of their larger discourse. Earlier events are typically expressed earlier in an utterance, while later events are expressed later ("order-of-mention," a.k.a. temporal iconicity, Jakobson, 1971). For instance, if somebody first went to the gym and then stopped for coffee, it would be most natural for them to say, "I went to the gym and stopped for coffee," rather than the reverse; the order in which the events are mentioned can stand in for the order in which they occurred.

During real-world communication, all three of these strategies can be deployed at the same time, complementing each other. For instance, if a speaker were to describe a series of events that occurred on a recent vacation, they might use expressions like "first," "and then," and "finally" to express explicitly, using lexical resources, the temporal order of events. In coordination with these expressions, they might point along the left-to-right spatial axis to convey the temporal order of the events. And, at the same time, they might choose to describe the events in the same order in which they actually occurred.

While we know that speakers do this, less is known about whether listeners actually care. Temporal terms are notoriously hard for children to acquire (Tilman et al., 2017; Shatz et al., 2010); the words "before" and "after," for instance, continue to be confused by most children until they are five years old (Clark, 1971). Listeners are also known to rely on order-of-mention to infer the order in which events occurred (Jakobson, 1971), although past work has focused primarily on contexts where temporal order is ambiguous in speech and gesture.

It's also currently unknown whether listeners rely on temporal gestures to make inferences about the abstract concept of temporal order. By contrast, listeners are demonstrably sensitive to concrete information expressed in gesture. For example, concrete, iconic gestures boost comprehension (Thompson, Driscoll \& Markson 1998) and can even add information not otherwise present in speech (Church et al. 2007; Singer \& Goldin-Meadow 2005). Less is known, however, about the communicative impact of gesture on the interpretation of temporal structure. Speakers use gesture to express a range of information about time, including duration and sequential order (Cooperrider \& Núñez, 2009). There is mixed evidence that, when speech is ambiguous (e.g., 'the meeting was moved forward,' which can mean earlier or later), observers use gesture to determine how the speaker's metaphorical conceptualization of time (Jamalian \& Tversky, 2012), although perhaps only when communication is copresent and not computer-mediated (Lewis \& Stickles, 2016). As far as we know, no previous research has investigated whether gestures about temporal order are actually communicative.

What about when these sources of information are not aligned but contradictory? Sometimes, we mention a later event first, perhaps because of the event's salience. When this occurs, the conflict is reflected in the listener's neural response to the sentence as they resolve the conflict (Münte et al, 1998). Previous research has assumed that, in cases of conflict, speakers will default to the information expressed lexically, overriding the temporal order suggested by order-of-mention. However, discourse comprehension involves probabilistic judgments about how best to integrate potentially contradictory information (e.g., Gibson et al, 2014). Under some circumstances, therefore, speakers are likely to rely on order-of-mention, overriding or ignoring the temporal order conveyed explicitly in speech.

To investigate the communication of complex temporal structure during multimodal discourse, we conducted two studies in which we systematically manipulated how information about temporal order was expressed in speech, gesture, and order-of-mention. Participants viewed brief videos in which a speaker described a complex series of events. Descriptions varied in the use of explicit temporal terms (e.g. "earlier" or "later") and temporal gesture (e.g., a leftward pointing gesture to indicate an earlier event) to order the events in the sequence. Within these descriptions, moreover, pairs of events were sometimes mentioned in the same order as they occurred, so that order-of-mention was a helpful guide to temporal order, but other times the order-ofmention did not align with their actual temporal order. All three sources of information-temporal terms,
temporal gesture, and order-of-mention-were thus fully crossed within subjects.

We foresaw a number of possible outcomes. On the one hand, temporal terms are so explicit and unambiguous that they may overwhelm information from any other source, including gesture and order-ofmention. On the other hand, a complex situation can involve multiple interrelated events, outstripping the relatively simple binary distinctions that are most common in speech (before/after, earlier/later). Under these circumstances, temporal gestures may be especially beneficial, as they allow a listener to track the relative ordering of multiple events. A series of three gestures, for instance, can use relative spatial location to order events without any of the ambiguity that can plague speech. Lastly, order-of-mention may sometimes trump both speech and gesture. First, it uses time (of mention) to represent time (of occurrence)-a direct mapping that may be difficult for a listener to ignore when constructing their discourse model. Second, we know that memory for specific words isn't great as a delay period increases (Sachs, 1967)—when listeners are trying to reconstruct the order of events after the fact, the may be more likely to recall the order in which they were mentioned than the specific words used to describe their order. Thus, there are good reasons to predict that all three sources of information may dominate listener's interpretations of multimodal communication about temporal order.

## Experiment 1

The main purpose of Experiment 1 was to investigate how participants use the temporal information available in multiple communicative resources to construct a temporal narrative of events. We were especially interested in how individuals reconcile situations in which different sources provide conflicting temporal information. And finally, because of past work suggesting that gesture effects on comprehension are amplified over a delay (Church et al. 2007), we added a Memory Condition (Immediate or Delayed) to investigate whether the temporal resources participants use to order a sequence of events changes over a delay.

## Methods

Participants: Forty undergraduate students $(\mathrm{N}=31$ female) participated in this study in exchange for course credit. Sample size was determined in advance on the basis of similar studies of gesture (e.g., $N=45$ in Church et al. 2007).

Materials: We filmed 16 vignettes in which a woman narrated a brief story consisting of four events. The events in a given sequence had all already occurred or all were going to occur. That is, half of the vignettes discussed future events (i.e. planning a hiking trip,
preparing to go camping) while the other half of the videos discussed a sequence of events that had already occurred (i.e. recalling a vacation, recapping a day at work).

Videos were filmed from the neck downwards, ending at the top of the narrator's legs, with the arms clearly visible. Because we manipulated whether or not temporal information is delivered through explicit temporal terms in speech, we cut the narrator's head and neck out of the frame to avoid giving participants visual clues (i.e. voice box movements) from which to draw temporal information. All of the video stimuli were generated originally and contained arbitrarily related events to ensure that participants could not determine the temporal order of events by relying on causal or canonical relationships (i.e. hiking up a cliff generally precedes jumping down a waterfall).

Procedure: Participants watched short video clips that described a four-event vignette. For example, a participant could hear, about an upcoming climb up Mount Everest, that "I should probably replace my old hiking boots and then pick up some snow gear for encountering snowy conditions. I also cannot forget to first get a hiking permit and then purchase an airline ticket to Nepal." In this case, most of the clauses contain explicit temporal terms that disambiguate the actual temporal order. Within each sentence, order of mention also indicates the correct temporal order (e.g., the narrator intends to get a hiking permit before purchasing an airline ticket to Nepal); by contrast, the two sentences' order of mention conflict with the order in which the events occurred. There were sixteen different vignettes in total. The videos of each vignette were randomly presented, and participants saw each of the unique vignettes one time. The video of each vignette, however, was played twice, back to back, to the participant.

Participants also completed a set of seven comprehension questions after viewing a video stimulus. For half of the vignettes, participants received the comprehension task immediately following the presentation of that particular vignette video (our Immediate memory condition). For the other half of the vignettes, participants received the corresponding set of comprehension questions following a 10 minute delay (our Delayed memory condition). During this 10 minute delay, participants completed multiplication and long division problems.

Each comprehension question was presented in 2AFC format (i.e. "Do I need to buy more winter gear before or after getting an airline ticket?") with a 10 second response window. Four of these seven questions tested the temporal relationship of events in the story (target questions). The remaining three questions in the comprehension set were unrelated to temporal content and probed the basic content of each video (filler questions). Question order was randomized for each
video and for each participant. At the end of the experiment, participants filled out a debrief questionnaire.

## Analysis

Before analyzing the data, we removed filler questions, trials with a response time faster than 200 ms , and trials that were two and a half standard deviations faster or slower than each participant's mean response time on each vignette. We excluded participants whose accuracy on the comprehension task was below $50 \%$ (chance) when considering trials where temporal term, gesture information, or both were present, as these individuals were below chance performance even when explicit ordering information was available to them. We also excluded responses for participants who failed our debrief point-of-view item. In this question participants were asked, "Which of the following gestures would the narrator use to accompany the word 'earlier'?" They were given two short video clips, one with the narrator making a rightward (from her point of view) gesture stroke, and one with her making a leftward gesture stroke from which to respond. Participants who chose the rightward gesture stroke-which appears as a leftward stroke from their mirrored perspective-as accompanying the word "earlier," were considered to have failed the debrief and were excluded from this analysis since we wanted to ensure they were interpreting the gestures in the videos the way we intended.

Our primary dependent measure was participants' response (before vs. after), as a function of the information expressed in order-of-mention, speech, or gesture (before, nothing, or after). When neither speech nor gesture include explicit temporal information, there is no 'ground truth' about the order of events. Each resource was dummy coded for its temporal information ("before" $=-1$, no info $=0$, "after" $=1$ ). All analyses used generalized mixed-effects models with a logistic link function, with centered predictors and the maximal converging effects structure justified by the design (Barr et al, 2013).

## Results

## Effects of language, gesture, and order-of-mention

We first examined how comprehension of temporal sequences is affected by temporal terms in speech, temporal gesture, and order-of-mention. All three sources of information had a significant effect. Temporal terms reliably influenced participants' interpretation of temporal order $(b=0.95 \pm 0.17 \mathrm{SEM}, z=5.45, p<$ 0.001 ), with more before interpretations after the use of the word "before" ( $M=77 \%$ ) but more after interpretations after the use of the word "after" ( $M=$ $72 \%$ ). Similarly, order-of-mention had a significant, if smaller, impact $(b=0.72 \pm 0.15$ SEM, $\mathrm{z}=4.94, \mathrm{p}<$
0.001 ), with more before interpretations when the event was mentioned first, but more after interpretation when the event was mentioned second. And temporal gestures, too, had a significant impact ( $\mathrm{b}=0.48 \pm 0.12$ SEM, $\mathrm{z}=$ $3.94, \mathrm{p}<0.001$ ), with more before responses after leftward "past" gestures ( $M=63 \%$ ), and more after responses are rightward "after" gestures ( $M=67 \%$ ). Participants thus were sensitive to all three of the semiotic resources available during multimodal comprehension, with larger influences of temporal terms and order-of-mention, and a smaller but significant influence of gesture.

The only other significant effect we observed was an interaction between temporal terms and temporal gesture ( $b=-0.38 \pm 0.16$ SEM, $z=-2.21, p=0.0271$ ). This was driven by a large effect of gesture when temporal terms were absent entirely from speech ( $b=0.54 \pm 0.17$ SEM, $z=3.19, p=0.0014)$, but a much smaller effect of gesture when accompanied by explicit temporal terms ( $b=0.72$ $\pm 0.21$ SEM, $\mathrm{z}=3.47, \mathrm{p}=0.00541$ ).

## Effects of recall

We were also interested in the effect a delay period would have on participants' comprehension of temporal events. Specifically, we wondered whether we would see the effects of particular resources (i.e. temporal gesture) strengthen over time, as previous research with iconic gesture has found (Church et al., 2007). Our results indicated that participants actually performed the same on the comprehension task regardless of whether it was completed immediately following the video vignette or after a 10 minute delay period. We did not find evidence of interactions between any of the resources and our memory condition factor.

## Discussion

Our study aimed to investigate how people draw on and integrate multiple sources of temporal information during comprehension of complex temporal sequences. We found that participants are independently influenced by the information available through temporal terms, gesture, and order-of-mention.

The presence of temporal terms and temporal gesture each influences participants to respond according to the order presented through these resources perhaps in explicitly conveying temporal order. Order-of-mention is also largely influential as a listener builds a temporal narrative, perhaps because of the salience of the iconicity (i.e. letting the order events are uttered in speech stand in for the order events occur in time).

Interestingly, we found an interaction between temporal terms and gesture, mediated by whether or not information from temporal speech is present-when ordering information from temporal terms is already
present, we see less of an impact of temporal gesture than when it is absent.

We were additionally surprised to not detect an effect of memory condition given our predictions that the effects of gesture are strengthened over time. Perhaps our delay was not long enough to elicit a difficult recall situation, in which temporal ordering information would decay over time. Creating that kind of recall situation is important to reveal any effects that our temporal resources may selectively provide over time.

## Experiment 2

Experiment 2 was designed to replicate and extend slightly the results of Experiment 1. The slight extension was to address our unexpected finding that the relative impact of gesture did not differ after a delay. This appeared to contradict previous evidence that the impact of gesture increases with the passage of time (Church et al, 2007). Based on this, we predicted that, as more time passed after observing an utterance, the relative contributions of explicit terms, gesture, and order-ofmention should change-with, in particular, an increased reliance on gesture.

However, participants' recall was not severely impacted after the delay, suggesting that this delay may not have been sufficiently long enough to observe a shift in importance between temporal terms, temporal gesture, and order-of-mention. We thus increased this delay from 10 minutes to 30 minutes.

## Methods

Participants: Adults ( $N=50,33$ women) participated in exchange for partial course credit.
Materials: The same as in Experiment 1.
Procedure: The same as in Experiment 1, except that we extended the delay period from 10 to 30 minutes.

## Analysis

All exclusionary criteria and data cleaning procedures used in Experiment 1 were also applied for Experiment 2. Analyses again used logistic mixed-effects models, with centered predictors and the maximal converging effects structure justified by the design (Barr et al, 2013).

## Results

Effects of language, gesture, and order-of-mention Experiment 2 replicated the main findings of Experiment 1. Participants reliably drew on information presented through temporal terms $(b=0.86 \pm 015$ SEM, $z=5.68$, $\mathrm{p}<0.001$ ), with more before interpretations after the use of the word "before" ( $M=72 \%$ ) but more after interpretations after the use of the word "after" ( $M=$ $71 \%$ ). They also use the information available via order-of-mention ( $b=0.62 \pm 0.12$ SEM, $z=5.00, p<0.001$ ), with more before interpretations when the event was
mentioned first, but more after interpretation when the event was mentioned second. And similarly, participants also rely on the information present in temporal gesture ( $\mathrm{b}=0.28 \pm 0.14, \mathrm{z}=2.00, \mathrm{p}<0.05$ ), with more before responses after leftward "past" gestures ( $M=62 \%$ ) but more after responses are rightward "after" gestures ( $M=$ $59 \%$ ). These findings replicate the results of our previous experiment: participants are sensitive to the information from temporal terms and order-of-mention in particular, and less reliably applying the information gleaned from gesture (Fig. 1).

We did not, however, replicate the interaction of temporal terms and temporal gesture ( $b=-0.10 \pm 0.14, \mathrm{z}$ $=-0.77, p=0.44)$.

## Effects of recall

We next turned our attention to the effect of the delay period, and how the available temporal resources would be deployed over time. Our results did not reveal an effect in line with our prediction that the impact of gesture would increase over time (Gesture x Memory Condition, $b=0.27 \pm 0.17 \mathrm{SEM}, \mathrm{z}=1.47, \mathrm{p}=0.14$ ), even with a longer delay period.


Figure 1. Impact of each resource on participants' response. Color indicates the log-odds of interpreting an event to have occurred after (vs. before). All three resources had an impact. As gesture went from expressing 'before' to 'after' (i.e., moving rightward), 'after' responses increased (i.e., shift from red to blue). Similarly, as speech went from 'before' to 'after' (ie., bottom to the top), 'after' responses increased. And when order-of-mention suggested that the event occurred after (i.e., right panel), there was a higher proportion of 'after' responses (i.e., shift toward blue).

## Discussion

Experiment 2 replicated the main findings of Experiment 1. We found that in multimodal communication, participants reliably glean information from explicit temporal terms in the utterance, order-of-mention, and to a lesser extent temporal gesture. The only effect that did not replicate was the interaction between gesture and temporal terms, which suggests that this effect is either small and fickle or potentially a false-positive from Experiment 1.

Even with a 30 -minute delay, the benefits of these metaphorical temporal gestures did not increase with the passage of time, unlike past findings for concrete representational gestures (Church et al., 2007). One explanation is that 30 minutes is insufficient to elicit the selective benefits of gesture.

## General Discussion

We set out to investigate how we communicate about the temporal structure of complex events. Multimodal communication offers a range of resources for expressing temporal order: words, gestures, discourse structure. Across both studies, we found that listeners made use of all three of these sources of information, integrating them to make sense of the temporal structure of complex narratives.

## The ephemeral and spatial nature of gesture

A central finding of these studies is that gestures that encode temporal order are genuinely communicative. Gestures are ephemeral, disappearing as soon as they are produced, and are only intermittently interjected into the the speech stream. Despite this, listeners made reliable use of gesture to interpret complex narratives.

Perhaps temporal gesture is especially useful in that it can help create a schematic "bird's eye" view of a complex event by laying out all of the events in their temporal order. A single temporal gesture can depict a pairwise relation between two events by placing them in space - but a sequence of gestures can construct a schematic representation of an entire narrative, including multiple subevents. By enacting a spatial timeline, temporal gestures supply an object for joint attention, available to both speaker and listener.

The utility of gesture may depend on the listener's perspective on the speaker. Because time recruits lateral space, assigning meaning to the right (future) and left (past) sides of space, interlocutors who face each other are confronted with an added challenge: adopting the perspective of their partner. The fact that participants in our study were able to interpret the speaker's lateral temporal gestures - despite the fact that the speaker was both head-on and video recorded - is a testament to the centrality of gesture to human communication.

## Integrating the complementary and contradictory

While we began by considering reasons that one source of information might dominate listeners' comprehension of temporal order, both studies found that all three sources of information make independent and reliable contributions. While the impact of gesture was less pronounced than that of speech or discourse structure, it was nonetheless robust across both studies. All three sources of information appear to make independent contributions to the interpretation of temporal structure.

One avenue for future research is whether there are individual differences in the reliance on these sources of information. Are some listeners especially sensitive to when an event is mentioned in the discourse, while others are more sensitive to how that event is gestured relative to others? If such individual differences exist, and we suspect they do, then these may lead to radically different interpretations when different sources of temporal information come into conflict - when the first event mentioned actually occurred last, or when the speaker gestures left but accidently says 'and then afterward....' Understanding these dynamics will help us understand how miscommunications occur - and are repaired.

Similarly, these three sources of information may have differential impacts in different communicative settings or under different task demands. For instance, Lewis and Stickles (2016) reported that gestures expressing metaphors for time were communicative-but only when the speaker was co-present with the listener, rather than appearing on video. Gestures expressing temporal order may also decrease in importance during videomediated communication, with speech and order-ofmention weighted more heavily. This may account for gesture's relatively smaller effect size in the current studies.

## Conclusion

We began by asking how listeners understand the temporal structure of complex narratives by integrating information from various sources: words, gesture, order-of-mention. We found that each of these resources made independent contributions to the comprehension of temporal order. In particular, these results demonstrate that metaphorical gestures can communicate complex temporal relations. The power of human communication may lie in its use of multiple strategies to communicate abstract information.

## Acknowledgments

Thanks to R. Núñez and M. Kutas for helpful feedback.

## References

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. Frontiers in psychology, 4, 328.
de Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., \& Streri, A. (2014). Representations of space, time, and number in neonates. PNAS, 111, 4809-4813.
Casasanto, D., \& Jasmin, K. (2012). The hands of time: Temporal gestures in English speakers.
Church, R., Garber, P., Rogawlski, K. (2007). The role of gesture in communication and social memory. Gesture, 7(2), 137-158.
Clark, E. V. (1971). On the acquisition of the meaning of before and after. Journal of verbal learning and verbal behavior, 10, 266-275.
Cooperrider, K., \& Núñez, R. (2009). Across time, across the body: Transversal temporal gestures. Gesture, 9(2), 181-206.
Gibson E., Bergen \& Piantadosi (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. PNAS, 110, 8051-8056.
Hockett, C. D. (1960). The origin of speech. Freeman.
Jakobson, R. (1971). Language in relation to other communication systems. Selected writings, 2, 570579.

Jamalian, A., \& Tversky, B. (2012). Gestures alter thinking about time. Proceedings of the 34th Annual Conference of the Cognitive Science Society.
Lewis, T. N., \& Stickles, E. Gestural modality and addressee perspective influence how we reason about time. Cognitive Linguistics.
Münte, T., Schiltz, K., \& Kutas, M. (1998). When temporal terms belie conceptual order. Nature, 395, 71-73.
Sachs, J. S. (1967). Recogpition memory for syntactic and semantic aspects of connected discourse. Attention, Perception, \& Psychophysics, 2(9), 437442.

Shatz, M., Tare, M., Nguyen, S. P., \& Young, T. (2010). Acquiring non-object terms: The case for time words. Journal of Cognition and Development, 11(1), 16-36.
Singer, M. A., \& Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. Psychological Science, 16(2), 85-89.
Tillman, K. Marghetis, T., Barner, D., \& Srinivasan, M. (in press). Today is tomorrow's yesterday: Children's acquisition of deictic time words. Cognitive Psychology.
Thompson, L. A., Driscoll, D., \& Markson, L. (1998). Memory for visual-spoken language in children and adults. Journal of Nonverbal Behavior, 22(3), 167187.

# Dynamics of Affordance Actualization 

Patric C. Nordbeck (nordbepc@mail.uc.edu)<br>Department of Psychology, 47 Corry Boulevard, Edwards Center 1<br>Cincinnati, OH 45221 USA

Laura K. Soter (soterl@carleton.edu)
Department of Philosophy, Carleton College
Northfield, MN 55057 USA

Rachel W. Kallen (kallenrl@ucmail.uc.edu)<br>Department of Psychology, University of Cincinnati Cincinnati, OH 45221 USA

Anthony P. Chemero (chemeray @ucmail.uc.edu)<br>Departments of Philosophy and Psychology, University of Cincinnati<br>Cincinnati, OH 45221 USA

Michael J. Richardson (richamo@ucmail.uc.edu)<br>Department of Psychology, University of Cincinnati<br>Cincinnati, OH 45221 USA


#### Abstract

The actualization of affordances can often be accomplished in numerous, equifinal ways. For instance, an individual could discard an item in a rubbish bin by walking over and dropping it, or by throwing it from a distance. The aim of the current study was to investigate the behavioral dynamics associated with such metastability using a ball-to-bin transportation task. Using time-interval between sequential ball-presentation as a control parameter, participants transported balls from a pickup location to a drop-off bin 9 m away. A high degree of variability in task-actualization was expected and found, and the Cusp Catatrophe model was used to understand how this behavioral variability emerged as a function of hard (time interval) and soft (e.g. motivation) task dynamic constraints. Simulations demonstrated that this two parameter state manifold could capture the wide range of participant behaviors, and explain how these behaviors naturally emerge in an under-constrained task context.


Keywords: affordances; dynamic systems; cusp catastrophe; dynamic modeling; simulations; constraints;

## Introduction

Reorganizing one's activity in relation to changing task demands is a ubiquitous aspect of everyday life and is often required to ensure task success. In order to solve everyday perception-action tasks, human (and animal) behavior is functionally (re)organized in relation to the affordances that define a given task context. Here, the term affordance simply refers to the action possibilities that characterize a given agent-environment system (Gibson, 1979).

Starting with Warren's (1984) seminal work on the perception of climb-ability, affordance perception research has demonstrated that affordances are defined by dimensionless ratios (termed pi-numbers) that capture the intrinsic, or body-scaled "fit" between the relevant aspects
of an environmental surface or object and an intentional agent's perception-action capabilities (i.e. effectivities). For instance, a stair riser is perceived to afford (comfortable) climbing if the ratio of the perceivers leg-length with respect to the height of the riser is less than approximately $\pi=.85$.

Similar body-scaled ratios are known to define a wide range of action possibilities, from the stand-ability and sitability of surfaces (e.g., Fitzpatrick, Carello, Schmidt \& Corey, 1994; Mark, 1987), to the pass-through-ability of apertures (Warren \& Whang, 1987), and the reach-ability and grasp-ability of objects (e.g., Carello, Grosofsky, Reichel \& Solomon, 1989; Cesari \& Newell, 1999; Richardson, Marsh \& Baron, 2007). In each case, this research has demonstrated how individuals correctly detect affordance boundaries (i.e., the boundary between when an action is or is not possible) by means of intrinsic bodyscaled information (e.g., eye-height information) and organize or reorganize their behavioral activity accordingly (Carello et al., 1989). For instance, individuals are able to correctly perceive when an object is reach-able or not by extending their arm when seated, or by bending their torso and extending their arm, or by standing up and walking over to the object and organize their behavior accordingly.

It is important to appreciate, however, that in most task contexts the different ways in which an affordance can be actualized are not organizationally discrete, but overlap. For instance, an object that is reachable by extending the arm, is also often reachable by bending the torso and extending the arm. Similarly, an object that is graspable with one hand, is also likely graspable with two hands. Furthermore, a small, light ball could be gripped with the fingers or grasped with the whole hand, and then carried or thrown to its final destination. In this sense, afforded task goals often entail a nested structure of multiply realizable action possibilities.

In dynamical systems terms, a nested affordance structure corresponds to a multi-stable system, whereby two or more states or modes of behavioral order are simultaneously stable (and could be actualized). Regarding the perception and actualization of nested affordances, a sign of multistability is hysteresis (Kelso, 1995; Turvey, 1990). This occurs when an individual transitions between two different behavioral modes or states at different body-scaled ratios depending on its history of previous performance. For example, an individual will typically transition from onehand to two-hand grasping at a larger arm-span/object-size ratio (i.e., $\pi=.8$ ) as object size is increased, compared to when transitioning from two-hand to one-hand grasping (i.e., $\pi=.65$ ) as object size is decreased (e.g., Frank, Richardson, Lopresti-Goodman \& Turvey, 2009; van der Kamp, Savelsbergh \& Davis, 1998). Although hysteresis has been observed in numerous affordance studies, other dynamic patterns of behavior have also been observed. For instance, in many task contexts individuals exhibit a fixed or critical point transition between different affordances or modes of affordance actualization. That is, individuals exhibit a nonlinear phase transition between different affordances or behavioral modes at the same body-scaled ratio irrespective of whether it is scaled up or down (Richardson et al., 2007; van der Kamp, et al., 1998). Enhanced contrast or negative hysteresis has also been observed and is defined by individuals transitioning between different behavioral modes in a prospective or anticipatory manner (Richardson et al., 2007; LoprestiGoodman et al. 2013). While these transitions show distinct changes in the actualization of an affordance over time, they are still stable solutions in terms of the task goal, or metastable. For example, in order to successfully grasp planks as plank size is increased, a transition (that varies inter-individually) between one-handed and two-handed grasping is necessary to maintain the task goal. Of course, fixed state behavior has also been observed, whereby an individual will enact the same affordance even if other behavioral modes are more effective or stable. For example, in an object grasping task a pair of individuals may choose to pick up objects together, even when it is more efficient and stable to pick up smaller objects separately (Richardson et al., 2007).

Of particular relevance here, is that the varied manner in which individuals are known to actualize a given affordance or transition between different affordances implies that affordance actualization is not determined by body-scaled ratios alone, but rather is determined by a more complex set of behavioral and contextual constraints. For example, the amount of time an individual has to perform a given task, an individual's motivation, and an individual's perceived ability for achieving task success are known to play a determining role in how a particular affordance is actualized (e.g., Lopresti-Goodman, Richardson, Marsh, Carello, \& Baron, 2007; Wilson, Weightman, Bingham, \& Zhu, 2016).

In an attempt to better understand how differing task constraints influence affordance actualization, Fajen (2007)
has proposed a distinction between hard versus soft constraints. Briefly stated, hard constraints are constraints that define a clear line between task success and failure. For example, when driving there is a minimum distance in which a driver would need to start braking in order to avoid colliding into a car stopped in front of them. The boundary between stopping and colliding thus corresponds to a hard task constraint, and if crossed will result in rather dramatic and potentially deadly task failure. However, even in this situation, successfully stopping could entail breaking close to this hard constraint or well before it. Of course, what determines which successful type of breaking behavior a driver chooses to actualize will depend on many different factors, such as the time of day, mood, or the degree to which a given driver prefers a large or small margin of safety. It is these kinds of latter constraints that correspond to soft constraints (Fajen, 2007; also see Harrison, Turvey \& Frank, 2016).

## Current study

The aim of the current study was to examine and model the effects of hard and soft constraints on affordance actualization, for a ball-to-bin transportation task. Of particular interest was the role that temporal constraints play on shaping the behavioral dynamics of an under- or softlyconstrained affordance actualization task. To achieve this, individuals were instructed to transport balls from a starting location to a target bin located 9 meters away. The interval between sequentially presented balls was manipulated by increasing or decreasing number of seconds between 2 and 14 (or vice versa) in 1 second steps every fourth ball. Importantly, individuals could complete the ball-to-bin transportation task in any way they wished; by walking/running and dropping the balls into the target bin or by throwing the balls into the bin from whatever distance they liked. Of interest was the distance that individuals chose to move prior to releasing the ball and the degree to which time-interval, as a control parameter, operated as a constraint on the observed behavioral dynamics.

Given the under-constrained nature of the task and the fact that it was impossible for individuals to achieve complete task success, the expectation was to observe a variety of behavioral dynamics. More specifically, using distance moved prior to attempting to throw the ball into the bin as the dependent variable, the expectation was that participants would exhibit one of four general classes of behavior as time-interval was increased or decreased across a continuous sequence of 52 balls: (I) fixed large distance moved (essentially always walking or running nearly the complete distance to the target bin); (II) fixed small distance moved (essentially always throwing from the ball pickup location); (II) gradual transition from large to small distance moved (or vice versa); and (IV) a non-linear transition between large and small distance moved.

It was also expected that the variation in the behavioral dynamics observed could be modeled using a two parameter, bifurcation or catastrophe model (namely, the
cusp catastrophe, in which the first parameter was represented by time interval and the second represented the collective approximation of the unknown soft constraints that influenced a given individual's behavioral dynamics. Before explicating this modeling endeavor however, the method and data analysis employed for the experimental study is detailed.

## Method

## Participants

Sixty-nine undergraduates from the University of Cincinnati participated in the experiment for partial course credit.

## Materials

At the starting area, seven-inch plastic playpen balls were put through an angled PVC pipe (marked '4' in Figure 1) that protruded into the pick-up area located on a wooden table (marked ' 3 ', dimensions: 40 cm wide, and 26.5 cm deep). The mouth of the PVC pipe extended back through an opaque curtain (marked ' 2 ') and a large wooden bin (marked ' 1 ', dimensions: 110 cm wide, 55 cm deep, and 120 cm high) was positioned at nine meters from the back edge of the ball pickup area. A computer program was used to visually signal an experimenter ('E') positioned behind the pickup location curtain (2) when to release the balls (one by one). A video-camera was used to record participants' (' P ') movements and actions throughout the experiment.


Figure 1: General experimental setup.

## Task and Procedure

Participants were told that the task involved transporting plastic playpen balls from a pickup area to a wooden bin located on the other side of the laboratory room. They were instructed they could only use one ball at a time and that the task was to get the balls into the bin, while at the same time not letting multiple balls stack up at the pickup location. They were told that the time between ball presentations would change from fast to slow or slow to fast (depending on sequence condition). They were also told that if they drop a ball accidentally then it could be picked up, however, if there was an attempt to get the ball into the bin but they missed, then they should ignore it and move on to the next one. Finally, they were instructed to solve the task in any way they liked as long as they followed the rules. (There were no consequences if rules were broken, and no incentive was given for performance).

Participants completed two trial series, with each series including the sequential presentation of fifty-two balls. Thirty-five participants started their first series at an increasing rate: beginning with a 14 second interval of ball presentation, this interval was decreased by 1 second after four balls down to 2 seconds (i.e., four balls were presented sequentially at each time interval). A small break was provided and then the second series began with the control parameter scaled in the reverse direction (i.e., from 2 to 14 seconds). The other thirty-four participants completed these same two trial sequences in the reverse order (i.e., 2 to 14 second sequence, followed by the 14 to 2 second sequence).

## Data Analysis and Behavioral Classification

The distance that participants moved prior to releasing (throwing or dropping) the ball was determined from the video recordings, along with task success (i.e., whether participants successfully got the ball into the bin or not). Although not reported here, the number of balls left within the pickup area at the time that the participant was attempting to get their current ball in was also recorded.

The movement distances were analyzed using Matlab 2016a (MathWorks, MA), with the behavior exhibited by participants in each temporal series (i.e., 2 to 14 second and 14 to 2 second series) graphically classified into one of four different types of dynamics (see below for more details). Prior to classification, the movement distances were averaged over each change in time interval, i.e. the average distance moved prior to releasing the ball was calculated over the four balls that had a fourteen second interval, then the average distance moved prior to releasing the ball was calculated over the four balls within thirteen second intervals, etc. This resulted in thirteen averaged movement distances for each 52-ball sequence. From these behavioral time-series, two descriptive statistics, namely mean distance moved ( $D_{m}$ ) and largest change in distance moved across a change in time interval ( $\Delta D$; i.e. the maximum of the differentiated 13-point behavioral time-series) were used to classify each behavioral time series as follows:

- Stable (fixed) small distance ( $\boldsymbol{s t} \boldsymbol{D}_{\text {small }}$ ) moved, whereby participants essentially always throw the ball from the pickup location or near the pickup location. More specifically, $D_{m}<4.8$ meters and $\Delta D<1.58$ meters.
- Stable (fixed) large distance ( $s t \boldsymbol{D}_{\text {large }}$ ) moved, whereby participants essentially always moved across nearly the complete distance to the target bin prior to releasing the ball. More specifically, $D_{m}>4.8$ meters and $\Delta D<1.58$ meters.
- Gradual change (phase transition) in distance ( $\boldsymbol{p t} \boldsymbol{D}_{\text {gradual }}$ ) moved, whereby participants gradually increased or decreased the distance as time interval decreased or increased, respectively (i.e., an inverse relationship between distance moved and time interval). More specifically, $1.58<\Delta D<3.8$ meters.
- Nonlinear change (phase transition) in distance ( $\boldsymbol{p} \boldsymbol{t} \boldsymbol{D}_{\text {nonlinear }}$ ) moved whereby participants exhibited a large
or nonlinear change in distance moved across a small change in interval. More specifically, $\Delta D>3.8$ meters.

Example time-series of each behavioral type are provided in Figure 2 for both increasing and decreasing interval sequences.


Figure 2: Two examples each of participant (full line, square markers) and simulated (dotted line, triangle markers) trajectories: $s t D_{\text {small }}$ trajectories (top), $s t D_{\text {large }}$ (second), $\boldsymbol{p t} \boldsymbol{D}_{\text {gradual }}$ (third) and $\boldsymbol{p t} \boldsymbol{D}_{\text {nonlinear }}$ (bottom).

## Modeling and Simulation

The possible emergence of the above four types of behavioral dynamics was modeled using a two-parameter task manifold defined by the Cusp Catastrophe (Thom, 1975) equation

$$
\begin{equation*}
\dot{x}=a+b x-x^{3} \tag{1}
\end{equation*}
$$

where $x$ represented that state or dependent variable, i.e., the distance (rescaled) moved prior to releasing or throwing the ball, the parameter $a$ represented (normalized) time interval from $(-2.5=2$ seconds to $+2.5=14$ seconds), and the parameter $b$ represented the collective state of (unknown) soft constraints that might be influencing a participant's behavior at any point during the task (i.e., motivation, intention, perceived ability, learned helplessness, etc.). The manifold in Figure 3 represents the fixed points of $x$, for different parameter settings of $a$ and $b$. That is, each point on the manifold can be understood as representing the distance moved prior to releasing the ball for each $(a, b)$
setting, where $x$ is rescaled (normalized) as a function of $b$ (e.g., from $-2=0$ meters moved to $+2=9$ meters walked when $b=2.8$ and from $-1=0$ meters moved to $+1=9$ meters walked when $b=-1.8$ ).

As can be seen in Figure 3, this manifold includes both mono-stable and bi-stable (multi-stable) regions and predicts the same four patterns of behavioral dynamics defined above depending on the values of $a$ and $b$. More specifically, as $a$ is scaled up or down, larger values of $b$ can result in behavioral trajectories qualitatively consistent with $\boldsymbol{s t} \boldsymbol{D}_{\text {large }}$ and $\boldsymbol{s} \boldsymbol{t} \boldsymbol{D}_{\text {small, }}$, depending on the initial condition of $x$. For $-.5<b<3$, however, the manifold predicts varying degrees of $\boldsymbol{p t} \boldsymbol{D}_{\text {nonlinear }}$ type behavior as $a$ (time interval) is scaled up or down. Finally, when $b<-.5$ the manifold predicts $\boldsymbol{p} t \boldsymbol{D}_{\text {gradual }}$ as $a$ (time interval) is scaled up or down.

It is worth noting at this point that Eq. (1) or the Cusp Catastrophe model has been employed to abstractly capture a wide range of natural bifurcation phenomena, including human anxiety and performance, organizational order, decision-making and dating behavior (e.g., Guastello, 1995; Hardy, 1996; Hardy \& Fazey, 1987; Richardson, Dale \& Marsh, 2014; Tesser, 1980). Typically, the $b$ parameter is fixed and the different behaviors that Eq. (1) can produce are explored by scaling $a$. In fact, this is how the exemplar trajectories plotted on the manifold in Figure 3 were generated (i.e., by fixing the value of $b$ and then scaling $a$ for a given initial condition $x_{0}$ ). In the current task context, this would be equivalent to assuming that although the soft constraints that influence a participant's behavior might change across trial sequences, they remain fixed over a ball sequence. However, there is no reason to assume that this is the case for the current task, rather it seems more likely that the various soft constraints that influence participant behavior change both during and across sequences. For instance, an individual's motivation or goal intention may have been continuously modulated during the task. Thus, at each interval change (or individual ball), the resulting distance moved may reflect a continuous (or discrete) change in both $a$ and $b$.

With the latter point in mind, a range of behavioral trajectories were simulated along the cusp catastrophe manifold by scaling $a$ in interval steps consistent with the time interval steps employed in the experimental study (i.e., from 2.5 to -2.5 in 13 steps), as well as scaling $b$ recursively by adding a number from a unimodal random distribution, with a mean of -.6 (when increasing interval, +. 6 when decreasing) and a standard deviation of 1.65 . The mean of $\pm .6$ was employed as the experimental data revealed that participants had a preference for higher movement (see results section for details). Two sets of 70 trajectories were simulated, with the initial condition $x_{0}$ set at +2 for simulation set one and a normal distribution with $50 \%$ chance of being above 0 for simulation set two (again inspired by participant behavior). The simulated data that resulted was rescaled to the distances of the real (human) experimental data ( $\sim .75$ meters to $\sim 8.75$ meters).


Figure 3: Cusp Catastrophe Model manifold. Blue points represent an exemplar $\boldsymbol{p t} \boldsymbol{D}_{\text {gradual }}$ behavioral trajectory. Red points represent two exemplar $\boldsymbol{p t} \boldsymbol{D}_{\text {nonlinear }}$ behavioral trajectories. The black and green points represent exemplar $\boldsymbol{s t} \boldsymbol{D}_{\text {large }}$ and $\boldsymbol{s t} \boldsymbol{D}_{\text {small }}$ behavioral trajectories, respectively.

## Results

As can be seen from an inspection of Figures 2 and 4, and Table 1, participants produced all four of the behavioral dynamics expected. The variability within and across participants and ball sequences is most easily discerned from an inspection of Figure 4, in which the behavioral dynamics classification is plotted as a function of mean distance moved ( $D_{m}$ ) and maximum change in distance moved across a change in time-interval ( $(\Delta D)$.

Table 1: Distribution of trajectories per type of data.

| Trajectory | Simulated | Actual |
| :--- | :---: | ---: |
| $\boldsymbol{s t} \boldsymbol{D _ { \text { small } }}$ | $8.57 \%$ | $9.42 \%$ |
| $\boldsymbol{s} \boldsymbol{t} \boldsymbol{D}_{\text {large }}$ | $33.57 \%$ | $35.51 \%$ |
| $\boldsymbol{p} \boldsymbol{D _ { \text { gradual } }}$ | $22.14 \%$ | $23.91 \%$ |
| $\boldsymbol{p t D _ { \text { nonlinear } }}$ | $35.71 \%$ | $31.16 \%$ |

The simulated trajectories also produced a comparable set of behavioral trajectories and classifications. Again, this can be seen from an inspection of Figures 2 and 4 and Table 1. The classification system was verified using a K-means Nearest Neighbor (KNN) classifier (in Matlab, with ten number of neighbors, Euclidian distance and squared inverse distance weights) finding $99.28 \%$ correspondence between initial classification and KNN classification of real data.

A curve estimation analysis was conducted on the total frequencies of each distance across all data-points, revealing a linear increase in frequency as distance increased ( $\beta=.85$, $t(34)=9.43, p<.01$, where $x=$ distance moved). A twotailed, bivariate correlation analysis was run to investigate the relationship between distance moved and success (hit) versus failure (miss), revealing a positive association ( $r=$ $.64, p<.01$ ) in that, as distance moved increased so did the probability of success.


Figure 4: Participant (top) and simulated data (bottom) behavioral classification as a function of mean distance moved $\left(D_{m}\right)$ and maximum change in distance $(\Delta D)$.

## Discussion

The current study was designed to explore the effects of hard and soft constraints on the manner in which a task goal was actualized. As expected, a variable range of behavioral dynamics was observed, reflecting the under-constrained nature of the task goal. Furthermore, simulations using a two-parameter Cusp Catastrophe manifold illustrated how the wide range of participant behaviors observed naturally emerged due to an under- or softly-constrained task context. That is, by the continuous modulation of soft constraints during ongoing task performance.

The significance of the current findings with regard to understanding human, affordance-based behavior is twofold. First, the current results highlight how both steady state linear and nonlinear behavioral patterns, as well as metastable and transient behavioral patterns, can all result from the same task dynamic, further emphasizing how complex and context sensitive determinism underlies the emergent (re)organization of ongoing human behavior.

Second, the current results illustrate the need for appropriately identifying what and how soft constraints modulate the actualization of nested affordances or multistable behavioral modes. While there was no attempt to specifically identify what soft constraints guided task performance in the current study, the experimental and modeling methodology developed here could be employed
to identify these constraints in future research. Different hard constraints could be imposed or manipulated, or the saliency of soft constraints within the task context could be explicitly defined. For example, one could introduce the hard (goal) constraint that a participant would fail completely (and need to redo the task) if there is ever more than one ball in the pickup area. This would likely see the elimination of $s t \boldsymbol{D}_{\text {large }}$ behavior. Furthermore, if the salience of a soft constraint were also increased, say by adding motivation in terms of a points or monetary reward system that empathized getting balls in the bin, then one would also predict the (near) elimination of $\boldsymbol{s t} \boldsymbol{D}_{\text {small }}$, with participants predominately producing $\boldsymbol{p t} \boldsymbol{D}_{\text {gradual }}$ or $\boldsymbol{p t} \boldsymbol{D}_{\text {nonlinear }}$ behavior.

It is also possible that task success or failure on each ball throw could have modulated the collective motivational state of participants. The general relationship between longer distance and higher success rate speaks to this point, although it does not apply as motivation to all participants equally. (If this were applicable on an individual basis, there would likely be no $\boldsymbol{s t} \boldsymbol{D}_{\text {small }}$ trajectories.) However, a confounding variable here is the general preference across the entire dataset for longer distances. The interaction of this preference with the individually different effect of timeinterval on distance moved, needs to be examined further in future research.

## Acknowledgments

This research was supported, in part, by the National Institutes of Health (R01GM105045).

## References

Carello, C., Grosofsky, A., Reichel, F. D., \& Solomon, H. Y. (1989). Visually perceiving what is reachable. Ecological Psychology, 1, 27-54.
Cesari, P., \& Newell, K. M. (1999). The scaling of human grip configurations. Journal of Experimental Psychology: Human Perception and Performance, 25, 927-935.
Fajen, B. R. (2007). Affordance-based control of visually guided action. Ecological Psychology, 19, 383-410.
Fitzpatrick, P., Carello, C., Schmidt, R. C., \& Corey, D. (1994). Haptic and visual perception of an affordance for upright posture. Ecological Psychology, 6, 265-287.
Frank, T. D., Richardson, M. J., Lopresti-Goodman, S. M., \& Turvey, M. T. (2009). Order parameter dynamics of body-scaled hysteresis and mode transitions in grasping behavior. Journal of Biological Physics, 35(2), 127-147.
Gibson, J. J. (1979). The ecological approach to visual perception. Boston, MA: Houghton-Mifflin.
Guastello, S. J. (1995). Chaos, catastrophe, and human affairs: Applications of nonlinear dynamics to work, organizations, and social evolution. Mahwah, NJ: Lawrence Erlbaum.
Hardy, L. (1996). Testing the predictions of the cusp catastrophe model of anxiety and performance. The Sport Psychologist, 10, 140-156.
Hardy, L. \& Fazey, J. (1987). The inverted-u hypothesis: A catastrophe for sport psychology? Paper Presented at the

Annual Conference of the North American Society for the Psychology of Sport and Physical Activity. Vancouver. June.
Kelso, J. A. S. (1995). Dynamic patterns. Cambridge, MA: MIT Press.
Lopresti-Goodman, S., Richardson, M. J., Marsh, K. L., Carello, C., \& Baron, R. M. (2009). Task constraints on affordance boundaries. Motor Control. 13, 69-83
Lopresti-Goodman, S. M., Turvey, M. T., \& Frank, T. D. (2013). Negative hysteresis in the behavioral dynamics of the affordance "graspable." Attention, Perception, \& Psychophysics, 75, 1075-1091.
Mark, L. S. (1987). Eyeheight-scaled information about affordances: A study of sitting and stair climbing. Journal of Experimental Psychology: Human Perception and Performance, 13, 361-370.
Newell, K. M., Scully, D. M., McDonald, P. V., \& Baillargeon, R. (1989). Task constraints and infant grip configurations. Developmental Psychobiology, 22, 817831.

Richardson, M. J., Dale, R., \& Marsh, K. L. (2014). Complex dynamical systems in social and personality psychology. Handbook of research methods in social and personality psychology, 253.
Richardson, M. J., Marsh, K. L., \& Baron, R. M. (2007). Judging and actualizing intrapersonal and interpersonal affordances. Journal of Experimental Psychology: Human Perception and Performance, 33, 845-859.
Tesser, A. (1980). When individual dispositions and social pressure conflict: A catastrophe. Human Relations, 33, 393-407.
Thom, R. (1975), Structural Stability and Morphogenesis. Reading, MA: W. A. Benjamin.
Turvey, M. T. (1990). Coordination. American Psychologist, 45, 938-953.
van der Kamp, J., Savelsbergh, G. J. P., \& Davis, W. E. (1998). Body-scaled ratio as a control parameter for prehension in 5- to 9 -year-old children. Developmental Psychobiology, 33, 351-361.
Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. Journal of Experimental Psychology: Human Perception and Performance, 10, 683-703.
Warren, W. H., \& Whang, S. (1987). Visual guidance of walking through apertures: Body-scaled information for affordances. Journal of Experimental Psychology: Human Perception and Performance, 13, 371-383.
Wilson, A. D., Weightman, A., Bingham, G. P., \& Zhu, Q. (2016). Using task dynamics to quantify the affordances of throwing for long distance and accuracy. Journal of Experimental Psychology: Human Perception and Performance, 42(7), 965-981.

# Broadening the Scope of Recognition Memory 

Justin M. Olds (justin.olds@unil.ch) \& Julian N. Marewski (julian.marewski@unil.ch)<br>Department of Organizational Behavior, University of Lausanne<br>1015 Lausanne, Switzerland


#### Abstract

Within the literature of psychological and decision sciences, there is a critical difference in the way recognition is defined and studied experimentally. To address this difference, the current experiment examines and attempts to disentangle the influence of two recognition judgment sources (from within an experiment and from an individual's prior life experiences) upon two different recognition judgments. By presenting participants with a set of related stimuli that vary naturally in environmental occurrence and by manipulating exposure within an experimental context, this experiment allows for a broader and more ecologically valid assessment of recognition memory. Contrasting with the typical wordfrequency effect, the results reveal an overall bias to judge high-frequency items as studied on an episodic recognition test. Additionally, the results underscore the role of context by showing that a single study exposure increases the probability that individuals will judge stimuli as presented outside the laboratory.


Keywords: Recognition memory; decision-making; ecological validity

## Recognizing the Difference between Recognition and Recognition

In general, recognition refers to the experience wherein upon encountering a stimulus an individual has a sense that she has encountered that stimulus before. A recognition judgement, in turn, is when an individual explicitly claims that a stimulus was previously encountered. Within the literatures of psychological and decision sciences, there is a critical difference in the way recognition is studied experimentally. While one set of research focuses on an individual's capacity to recognize stimuli presented previously within an experimental episode (episodic recognition memory), the other set focuses on an individual's capacity to recognize stimuli as previously encountered during the individual's prior life experiences before beginning an experiment (pre-experimental recognition memory). Although these types of recognition are typically studied independently, the sources of experience related to both are inherently interconnected. Indeed, the experience of recognition is influenced by an individual's prior life experiences as well as by the experiences she has during an experiment. The current work provides a framework for studying recognition memory in a way that more readily relates to these two intertwined factors.

In what follows, we first broadly describe the lines of research related to both types of recognition judgements,
including prior work that has examined their interconnected nature. Within this review, we note criticisms of each line of research. Following this, we (1) describe a research methodology that draws upon both lines of work to address these concerns, (2) present the results of an experiment adopting this approach, and (3) discuss implications of these results and considerations for future work.

## Episodic Recognition in Memory Research

Recognition memory has been studied extensively with listlearning experiments. Here, stimuli, such as words or pictures, are presented individually in the form of a study list. After a delay ranging from a few seconds to multiple days, participants are presented with a recognition memory test list, and are asked to discriminate targets (stimuli from the study list) from foils (new items). Episodic recognition memory has been the focus of decades of extensive research and has been noted as an increasingly prevalent research paradigm (e.g., Hintzman, 2011).
A major strength of the episodic recognition memory line of research is experimental control. This is, in part, achieved by minimizing the role of individual stimuli, such as by presenting mixed lists of unrelated and uncommon concrete nouns. This approach follows in the footsteps of pioneering memory researcher Ebbinghaus (1885), who used nonsense stimuli to avoid the influence of everyday exposure.
The advantage of striving for experimental control in this way is also a disadvantage when it comes to understanding how memory judgments operate within everyday decisionmaking. For example, Neisser (1976) argued that memory research should strive toward ecological validity. Drawing upon work in perception by Brunswik (1957) and Gibson (1979), ecological validity refers to applicability outside the laboratory. The importance of ecological validity is underscored by research on eyewitness testimony. For instance, when participants study mixed lists of unrelated words, a positive relationship between accuracy and confidence is typically found (e.g., Dallenbach, 1913; Dunlosky \& Metcalfe, 2009). This intuitive finding dovetails with the 1972 and 1976 U.S. Supreme Court rulings suggesting that highly confident eyewitness identification is likely accurate. This pattern, however, does not hold for lists of similar material (i.e., categorized lists, see e.g., DeSoto \& Roediger, 2014), such as description details of suspects presented to an eyewitness (e.g., Smith, Kassin, \& Ellsworth, 1989). It is disconcerting to consider how other memory research findings might also be misleading due to a similar lack of ecological validity.

Departing from the convention of presenting participants with mixed lists of uncommon, unrelated stimuli would inherently introduce an additional factor, which might interact with or overshadow other experimental manipulations. Specifically, a by-product of presenting participants with a more ecologically valid set of related stimuli is that the individual items will vary based on how often each occurs outside the laboratory. The influence of environmental frequency upon episodic recognition judgements has been the focus of extensive research (e.g., Dennis \& Humphreys, 2001; Estes, 1994; Glanzer \& Adams, 1985; Lohnas \& Kahana, 2013; McClelland \& Chappell, 1998; Shiffrin \& Steyvers, 1997). In these experiments, study and test lists are composed of unrelated words that are sampled randomly from a range of low linguistic frequency words and high linguistic frequency words. The typical word-frequency finding on an episodic recognition memory test is that low-frequency target words are more accurately judged to be "old" than high-frequency target words and low-frequency foil words are more accurately judged to be "new" than high-frequency foils. Although this line of research does begin to reintroduce the influence of pre-experimental exposure into a laboratory setting, these experiments still tend to favor experimental control over ecological validity in numerous ways. First, the study lists in these experiments are typically composed of stimuli that are sampled from opposite poles along a continuum of environmental frequency-either extremely low-frequency or extremely high-frequency (but see Lohnas \& Kahana, 2013, for an exception), essentially transforming the naturally continuous variable of environmental frequency into a dichotomized factor. Second, the study and test lists used in these experiments are often composed of unrelated words. That is, the stimuli belong to many disparate categories and, thus, do not related to one another regarding any real-world inferences, such as person details in relation to culpability.

It may be the case that, similar to experiments investigating the word-frequency effect (e.g., Glanzer \& Adams, 1985), a mnemonic advantage for low-frequency items emerges if participants are tested with a related set of stimuli sampled with varying degrees of environmental frequency. There is some evidence, however, suggesting that this pattern of superior recognition accuracy for lowfrequency items might not persist. Specifically, in an experiment by Jacoby, Woloshyn, and Kelley (1989), which investigated the influence of environmental frequency by composing study and test lists of famous and nonfamous names, the results of an episodic recognition test revealed that famous names (i.e., high-frequency items) were more likely to be judged as presented on the study list than nonfamous ones (i.e., low-frequency items), regardless of whether or not they were actually studied. Thus, it may turn out that the low-frequency item advantage, borne out of memory research favoring experimental control over ecological validity, may not hold when participants are tested with sets of related stimuli.

## Pre-Experimental Recognition in Decision-Making Research

Numerous researchers have investigated how individuals use pre-experimental recognition, a sense of prior exposure to a stimulus outside the laboratory, during decision-making (e.g., Dougherty, Gettys, \& Ogden, 1999; Erdfelder, Küpper-Tetzel, \& Mattern, 2011; Goldstein \& Gigerenzer, 2002; Hertwig, Herzog, Schooler, \& Reimer, 2008; Marewski \& Schooler, 2011). A sense of pre-experimental recognition has been shown to influence a wide array of judgments, including about population size (e.g., Marewski \& Schooler, 2011), fame (Jacoby, Woloshyn, \& Kelley, 1989), and company revenue (Hertwig, et al., 2008), to name but a few. This work has shown that the frequency of occurrence for a given stimulus in the environment (which can be estimated by counting how often a stimulus, say a city name, occurs on the Internet or in the print media), allows for modeling how likely and how quickly the stimulus is to be recognized (e.g., Hertwig et al., 2008). Overall, this body of research underscores the intuitive notion that a sense of recognition is paramount for making many everyday decisions.

The influence of pre-experimental recognition upon decision-making is typically studied within probabilistic inference experiments, in which individuals are assumed to use known attributes of a stimulus as cues to make inferences about an unknown or future criterion. These experiments typically include a recognition task and a paired-comparison inference task. In the recognition task, participants are shown a list of related stimuli, such as city names, and asked to judge if they recognize each item from their prior life experience. Additionally, some experiments ask individuals to report if they have additional knowledge beyond a sense of pre-experimental recognition for each object. Reponses from this task and their respective response times are later used, to predict judgments on the paired-comparison inference task. In the paired-comparison inference task, individuals are shown two items at a time from a related set of stimuli and asked to infer which alternative is higher or lower on some judgement criterion, such as which of two cities has a larger population size.

In part because of its rigid simplicity, one decision strategy in particular, the recognition heuristic (Goldstein \& Gigerenzer, 2002), has been the focus of much research and debate. This strategy assumes that, stemming from an existing relationship in the world between environmental occurrence and a given criterion (e.g., population size), a sense of pre-experimental recognition can readily guide decisions in a straightforward way. Specifically, the recognition heuristic assumes that on a paired-comparison inference task, if one decision alternative is recognized and the other is not, individuals will judge the recognized alternative to have a higher value on the criterion.

In general, this line of research is aptly commended for showcasing and exploring how memory is employed for everyday decisions. One major criticism of this research, however, is that the assumptions about recognition memory
fail to appropriately draw upon theory from the abundance of related recognition memory research (e.g., Dougherty, Franco-Watkins, \& Thomas, 2008; Newell \& Fernandez, 2006). It appears that much of the work on recognitionbased decision-making assumes that pre-experimental recognition is a fixed commodity, whereas research concerning episodic recognition memory has revealed many ways in which a sense of recognition is influenced by contextual conditions. For example, Jacoby et al. (1989) found that presenting nonfamous names within an experiment study phase increased the likelihood that these nonfamous names would be incorrectly judged as famous later (see also Hertwig et al., 2008). Related to this, Pohl and colleagues (Pohl, Erdfelder, Michalkiewicz, Castela, \& Hilbig, 2016), point out two typical experimental procedure choices that fail to consider how a sense of pre-experimental recognition might be influenced by experimental conditions. Both entail how participants, during an earlier part of the experiments, are often exposed to the stimuli that they are later asked to consider during a paired-comparison inference task. First, to obtain a large number of paired-comparison inference trials, items are often paired repeatedly with different items, such that each item appears numerous times during the task. Additionally, the order of the two tasks is often counterbalanced across participants, such that some participants have the recognition task first and others have the paired-comparison inference task first. These two typical methodology choices may influence the sense of preexperimental recognition and respective recognition speed that individuals might use during decision-making. Although Pohl and colleagues (2016) focused on recognition speed specifically in relation to these methodology concerns, context conditions such as exposure within an experiment might also influence the probability that participants will judge stimuli as recognized from outside the laboratory. This is one of the concerns the current approach addresses.

## Experiment

The purpose of the current experiment is to investigate the influence of two fundamental sources of experience, one stemming a person's prior life experiences before entering the laboratory and another stemming from the experiences within an experimental context, upon recognition judgements. Specifically, the memory source related to the experimental context here is a single study exposure and the memory source related to prior life experiences is preexperimental exposure (estimated with web frequencies). The influence of both sources is examined for both episodic recognition (e.g., "Was this city name presented earlier during the experiment?") and pre-experimental recognition (e.g., "Have you ever heard of this city name before beginning the experiment?"). In line with ecological validity, instead of informing participants that stimuli presented during the study phase would be presented during a later memory test, incidental study exposure was adapted from Hertwig et al. (2008).

In relation to previous work, the current experiment also addresses the following two questions. First, does the typical low-frequency item advantage for episodic recognition memory (e.g., Glanzer \& Adams, 1985) occur when individuals are tested with a more ecologically valid set of stimuli (i.e., stimuli from a related set that vary based on their natural occurrence outside the laboratory)? Second, to what extend does a single incidental study exposure influence judgements of pre-experimental recognition, and, if so, does this influence depend on the environmental frequency of stimuli?

## Method

Participants A total of 63 individuals (mean age $=21.5$ years, $56 \%$ female) recruited from the University of Lausanne were paid roughly 26 Swiss francs each (depending on performance) for participating in the experiment. They were tested individually.
Design The study was conducted as a within-subjects design with one independent variable, study status, one pseudoindependent variable, environmental frequency, and two dependent variables (measured using within-subjects blocks), episodic recognition judgements and preexperimental recognition judgements. Study status was manipulated within-subjects by presenting half of the to-betested stimuli within a preceding study phase. Preexperimental frequency was estimated for each stimulus using Wikipedia page occurrences as a proxy for environmental occurrence.
Stimuli The stimuli presented during the experiment were from a set of 200 city names from North American and Western European countries (Canada, England, France, Germany, Italy, Spain, and USA). Additionally, eight extra city names were used at the beginning and end of the study task to minimize primacy and recency study effects. The city names used in the experiment were sampled such that the entire set would include cities from each country that varied in both population size (population statistics obtained from www.citypopulation.de) and environmental occurrence (as approximated by Wikipedia page occurrences) and excluded capital cities. All stimuli were counterbalanced such that each occurred roughly equally often in all study and test conditions.
Procedure The experiment was conducted on a computer using E-Prime experimental software (Psychology Software Tools). Participants were informed that there would be three separate tasks and payment would depend of their performance in each task. First, all participants had an incidental study task, in which half of the city names (100) were presented. Specifically, participants were shown each city name individually and asked to count the number of vowels in each city name. At the beginning of each trial a fixation cross (+) appeared on the screen for 2 s along with a reminder of the task. Participants were informed that the fixation cross would occur immediately before each city name was presented to help them prepare to respond. Afterwards, a city name replaced the fixation cross and
participants were prompted to press the appropriate number key on the keyboard corresponding to the number of vowels in the city name. Participants were given up to 4 s to make their response and a blank screen was presented for 2 s between trials.

The following test phase consisted of two separate tasks, an episodic recognition task ("Was this city name presented during the vowel counting task") and a pre-experimental recognition task ("Have you ever heard of this city name before beginning the experiment?"). Participants were asked to respond with one of two keyboard keys to respond "yes" or "no" for each test trial. Participants were given as long as needed to make their response in both tasks. Each task block consisted of 100 trials, which included half studied and half unstudied city names. Importantly, city names were not repeated across test tasks, but were counterbalanced between-subjects such that each city name would occur roughly equally often in each test phase. The order of test tasks was counterbalanced between-subjects. Similar to the study phase, each trial began with a 2 s fixation cross and was followed by a 2 s intertrial interval during both recognition tasks.

## Results

The data from all 63 participants was analyzed, excluding trials for one city name due to a clerical error. Wikipedia page occurrence values were log transformed to approximate a linear relationship across city names (e.g., Marewski \& Schooler, 2011). Given the two binary dependent variables, separate logistic regressions were planned for both episodic recognition and pre-experimental recognition judgments.

Episodic Recognition Results from the episodic recognition task are presented in Figure 1. From visual inspection of Figure 1, two patterns are apparent. First, the probability that participants judged city names as presented during the study phase was higher for studied city names than for unstudied city names. Second, the probability of judging city names as studied increased as a function of environmental frequency. Moreover, the difference in recognition probabilities between studied and unstudied city names did not seem to vary as a function of environmental frequency. To test the influence of both factors (study status and environmental frequency), a multilevel logistic regression analysis was conducted to fit the episodic recognition data. A test of the model against a constant only model indicated that the predictors as a set provided an improved fit $\chi^{2}=980, d f=2, p<.001$, Nagelkerke's $R^{2}=$ .20). The Wald criterion demonstrated both factors, study status ( $z=5.38$ ) and environmental frequency $(z=6.82)$, contributed to the model fit $(p<.001$ for both). Additionally, by comparing the model to another which included an additional interaction term of the two factors, evidence of an interaction between study status and environmental frequency was not found ( $\chi^{2}=.27, d f=1, p=$ .61). To control for the variance associated with the random factor of repeated measurements from individual
participants, follow-up generalized linear mixed models were conducted. The same pattern emerged: both factors of study status $(z=5.22, p<.001)$ and environmental frequency ( $z=6.77, p<.001$ ) contributed to the model fit, and there was no evidence of an interaction between the two found ( $\chi^{2}=.78, d f=1, p=.38$ ).


Figure 1: Mean episodic recognition rates. The left side depicts the mean influence of study exposure for all city names with standard error bars. The right side depicts the influence of both study exposure and environmental frequency for each city name with a moving average for both studied and unstudied city names across environmental frequency.
Comparison to previous results The results from the episodic recognition task contrast with the typical wordfrequency effect (e.g., Glanzer \& Adams, 1985), which reveal an interaction between study and environmental frequency, such that low-frequency targets are recognized more accurately (i.e., higher hit rate and lower false alarm rate) than high-frequency targets and low-frequency lures are correctly rejected more accurately than high-frequency lures (e.g., Lohnas \& Kahana, 2013). Instead, the current results show that when participants are tested with a set of related stimuli, high-frequency items are more likely to be judged as studied regardless of if they were studied or unstudied (i.e., higher hit and false alarm rates for highfrequency items). Why did this pattern of results differ from the typical word-frequency effect? Although further work is required to better address this question, we can provide some speculation. The potentially stronger association with the experiment context for low-frequency items, perhaps due to item distinctiveness, may have been relatively diluted in the current experiment for a number of reasons. First, one important difference to remark upon is the overall low episodic recognition accuracy from the current experiment compared to previous word-frequency experiments (e.g., Lohnas \& Kahana, 2013). We suspect this difference can be attributed to the difficulty inherent in testing sets of related stimuli and stemming from incidental study (e.g., Criss \& Shiffrin, 2004). This increased task difficulty may have led participants be more influenced by the pre-existing associations for high-frequency items, since these
associations are less contingent upon the study conditions than the associations between the context and studied items.

Pre-Experimental Recognition Results from the preexperimental recognition task are presented in Figure 2. From visual inspection of Figure 2, two patterns are apparent. First, the probability of judging city names as recognized from outside the laboratory increased as a function of environmental frequency. Second, the probability that participants judged city names as recognized from outside the laboratory was slightly higher for studied city names than for unstudied city names. Additionally, it appears that the influence of a study exposure was relatively consistent across varying degrees of environmental frequency. To test the influence of both factors (study status and environmental frequency), a multilevel logistic regression analysis was conducted to fit pre-experimental recognition judgements using study status and environmental frequency as predictors. A test of the model against a constant only model indicated that both predictors as a set provided an improved fit $\left(\chi^{2}=1739, d f=\right.$ 2, $p<.001$, Nagelkerke's $R^{2}=.323$ ). The Wald criterion demonstrated both factors, study status $(z=4)$ and environmental frequency $(z=32)$, contributed to the model fit ( $p<.001$ for both). Additionally, by comparing the model to another which included an additional interaction term for the two factors, evidence of an interaction between study status and environmental frequency was not found ( $\chi^{2}$ $=.621, d f=1, p=.431)$. Evidence for the same pattern was suggested by a generalized linear mixed model with the categorical variable of participant included as a random factor.
Comparison to previous results Similar to previous work (e.g., Marewski \& Schooler, 2011), the current results support the use of web frequencies as a reasonable predictor of pre-experimental recognition. The results from the preexperimental recognition task also converge with previous work showing that an experimental exposure increases the probability of inferring an item to be higher on a criterion related to pre-experimental exposure, such as the fame of individuals (Jacoby et al., 1989) or population size of cities (Hertwig et al., 2008). Unlike previous work, however, the current experiment examines the relationship of a single incidental study exposure across items varying in environmental frequency continuously from extremely infrequent to extremely frequent. Importantly, the current work examines pre-experimental recognition instead of inference judgments, which are assumed to be influenced by a sense of pre-experimental recognition. By focusing on this more basic memory judgment, the current approach and respective data reveal that the presentation of stimuli within an experimental context influences a sense of preexperimental recognition that is core to much research on memory-based decision research (e.g., Goldstein \& Gigerenzer, 2002). One novel finding is that, because of the lack of an interaction between study and environmental frequency, it appears a single study exposure results in a
relatively constant increase in the probability of preexperimental recognition across all items, regardless of environmental frequency.


Figure 2: Mean pre-experimental recognition rates. The left side bar graph depicts the mean influence of study exposure for all city names with standard error bars. The right side depicts the influence of both study exposure and environmental frequency for each city name with a moving average for both studied and unstudied city names across environmental frequency.

## Discussion

The main purpose of the current work is to showcase a broad approach to studying recognition memory-one that considers how different types of recognition emerge as a function of the interconnected factors of (1) an individual's prior life experiences and (2) an individuals' recent and current experiences within an experimental context. This approach was designed to better translate results from memory experiments into real-world situations. This was achieved by testing recognition judgements for related stimuli, which, in turn, we assume more readily relate to everyday recognition judgments and memory-based inferences. For instance, one might be asked to identify or corroborate which colleagues were present at a company meeting or holiday party. This would entail gauging exposure within a context for a related set of stimuli (e.g., co-workers) that vary based on their environmental frequency (i.e., some are more well-known than others). This kind of everyday memory task contrasts sharply with the typical kind of recognition memory task used in psychology experiments, in which stimuli are unrelated and environmental occurrence is either constrained or dichotomized into highly disparate factor levels (extremely low and high-frequency bins).

The importance of adopting a broader and more ecologically valid approach to understanding recognition memory is underscored by the current experiment results. In contrast to the typical word frequency effect (increased episodic recognition accuracy for low-frequency items) found in many previous experiments, (e.g., Glanzer \& Adams, 1985; Lohnas \& Kahana, 2013), the current
experiment results showed an increased tendency to judge high-frequency items from a set of related stimuli as studied.

The current experiment results also provide evidence suggesting that context conditions, such as a single incidental study exposure, influence pre-experimental recognition judgements. This finding suggests that researchers examining memory-based inferences should strongly consider how often and in what manner stimuli are presented within an experiment. Related to this concern, although separate sets of stimuli were presented during both recognition test phases (episodic and pre-experimental) of the current experiment, we reanalyzed both sets of data with the inclusion of test task order as a factor to help rule out the influence of task demands upon the results. For both recognition tasks, the same main effects (study exposure and environmental frequency) and lack of interaction were supported. Importantly, these results did not interact with task order and a main effect of task order was not found.

There are numerous possible extensions of the current work. One is to incorporate the influence of context factors into models of memory-based decision-making. Additionally, the influence of list composition (e.g., Malmberg \& Murnane, 2002) upon both episodic and preexperimental recognition can be explored with sets of related stimuli, such as city names. Although the current experiment included related stimuli that varied widely on environmental frequency and the stimulus set was somewhat balanced, in that half of the city names were typically recognized, it remains largely unexplored to what degree a sense of recognition is influenced by the composition of study and test lists of related sets of stimuli. Similar to list composition effects, testing other stimulus materials, such as eyewitness-related description details, may help reveal the influence of varying environmental occurrence patterns.

## Author's Note

Correspondence: justin.olds@unil.ch. We would like to thank Ulrich Hoffrage and the four anonymous reviewers for their helpful comments on a previous version of this article. This work was supported by a grant from the Swiss National Science Foundation [159822].

## References

Brunswik, E. (1957). Scope and aspects of the cognitive problem. The essential Brunswik: Beginnings, explications, applications, 300-312.
Criss, A. H., \& Shiffrin, R. M. (2004). Interactions between study task, study time, and the low-frequency hit rate advantage in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 778.
Dallenbach, K. M. (1913). The relation of memory error to time interval. Psychological Review, 20, 323.
Dennis, S., \& Humphreys, M. S. (2001). A context noise model of episodic word recognition. Psychological Review, 108, 452.
Roediger III, H. L., \& DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. Memory, 22, 76-91.
Dougherty, M. R., Franco-Watkins, A. M., \& Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental
models and the fast and frugal heuristics. Psychological Review, 115, 199.
Dougherty, M. R., Gettys, C. F., \& Ogden, E. E. (1999). MINERVADM: A memory processes model for judgments of likelihood. Psychological Review, 106, 180.
Dunlosky, J., \& Metcalfe, J. (2009). Metacognition. New York, NY: Sage
Ebbinghaus, H. (1885). Über das gedächtnis: untersuchungen zur experimentellen psychologie. Duncker \& Humblot.
Erdfelder, E., Küpper-Tetzel, C. E., \& Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. Judgment and Decision Making, 6(1), 7.
Estes, W. K. (1994). Classification and Cognition. New York: Oxford University Press.
Gibson, J. J. (1979). The ecological approach to visual perception, New York: Houghton Mifflin.
Glanzer, M., \& Adams, J. K. (1985). The mirror effect in recognition memory. Memory \& Cognition. 12, 8-20.
Goldstein, D. G., \& Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. Psychological Review, 109, 75.
Hertwig, R., Herzog, S. M., Schooler, L. J., \& Reimer, T. (2008). Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 1191.
Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth". Perspectives on Psychological Science, 6, 253-271.
Jacoby, L. L., Woloshyn, V., \& Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. Journal of Experimental Psychology: General, 118, 115.
Lohnas, L. J., \& Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 1943.

Malmberg, K. J., \& Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 616.

Manson v. Brathwaite, 432 U.S. 98,97 S.Ct. 2243 (1976).
Marewski, J. N., \& Schooler, L. J. (2011). Cognitive niches: an ecological model of strategy selection. Psychological Review, 118, 393.

McClelland, J. L., \& Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effect of experience in recognition memory. Psychological Review, 105, 724760.

Neil v. Biggers, 409 U.S. 188 (1972).
Neisser, U. (1976). Cognition and Reality. San Francisco: W.H. Freeman.
Newell, B. R., \& Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. Journal of Behavioral Decision Making, 19, 333.
Pohl, R. F., Erdfelder, E., Michalkiewicz, M., Castela, M., \& Hilbig, B. E. (2016). The limited use of the fluency heuristic: Converging evidence across different procedures. Memory \& Cognition, 44, 1114-1126.
Shiffrin, R. M., \& Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. Psychonomic Bulletin \& Review, 4, 145-166.
Smith, V. L., Kassin, S. M., \& Ellsworth, P. C. (1989). Eyewitness accuracy and confidence: within-versus between-subjects correlations. Journal of Applied Psychology, 74, 356.
Tversky, A., \& Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207232.

# A Computational Logic Approach to Human Syllogistic Reasoning 

Ana Oliveira da Costa* and Emmanuelle-Anna Dietz Saldanha* and Steffen Hölldobler* and Marco Ragni**1<br>* International Center for Computational Logic, TU Dresden, Germany, ana.oli.costa@gmail.com, $\{$ dietz,sh $\} @ i c c l . t u-d r e s d e n . d e$<br>${ }^{* *}$ Cognitive Computation lab, University of Freiburg, Germany, ragni@cs.uni-freiburg.de


#### Abstract

A recent meta-analysis (Khemlani \& Johnson-Laird, 2012) about psychological experiments of syllogistic reasoning demonstrates that the conclusions drawn by human reasoners strongly deviate from conclusions of classical logic. Moreover, none of the current cognitive theories predictions fit reliably the empirical data. In this paper, we show how human syllogistic reasoning can be modeled under a new cognitive theory, the Weak Completion Semantics. Our analysis based on computational logics identifies seven principles necessary to draw the inferences. Hence, this work contributes to a computational foundation of cognitive reasoning processes.


Keywords: Human Reasoning; Syllogistic Reasoning; Logic Programming; Three-valued Łukasiewicz Logic; Abduction

## Introduction

The way of how humans ought to reason correctly about syllogisms has already been investigated by Aristotle (Smith, 1989). A syllogism consists of two premises and a conclusion. Each of them is a quantified statement using one of the four quantifiers all (A), no (E), some (I), and some are not $(\mathrm{O})^{2}$ about sets of entities which we denote in the following by $a, b$, and $c$. An example of a syllogism is:

Some $b$ are a
No bare $c$
In experiments, participants are normally expected to complete the syllogism by drawing a logical consequence from the first two premises, e.g. in this example 'some a are not $c^{\prime}$. The response of the participant - the conclusion - is evaluated as true if it can be derived in classical first-order logic (FOL), otherwise as false. The four quantifiers and their formalization in FOL are given in Table 1. The entities can appear in four different orders called figures as shown in Table 2. Hence, a problem can be completely specified by the quantifiers of the first and second premise and the figure. The example discussed above is denoted by IE4.

Altogether, there are 64 syllogisms and, if formalized in FOL, we can compute their logical consequence in classical logic. However, a meta-analysis by Khemlani and Johnson-Laird (2012) based on six experiments has shown that humans do not only systematically deviate from the predictions of FOL but from any other of 12 cognitive theories. In the case of IE4, besides the above mentioned logical consequence, a significant number of humans answered 'no valid conclusion', which does not follow from IE4 in FOL, as 'some a are not c' follows from IE4 as can be seen in the Venn diagram in Figure 1: $X$ has the property $a$ but not the property $c$.

[^156]Table 1: The four moods and their logical formalization.

| Mood | Natural language | First-order logic | Short |
| :--- | :--- | :--- | :---: |
| affirmative universal all $a$ are $b$ | $\forall X(a(X) \rightarrow b(X))$ | Aab |  |
| affirmative existential some $a$ are $b$ | $\exists X(a(X) \wedge b(X))$ | lab |  |
| negative universal | no $a$ are $b$ | $\forall X(a(X) \rightarrow \neg b(X))$ | Eab |
| negative existential | some $a$ are not $b \exists X(a(X) \wedge \neg b(X))$ | Oab |  |

Table 2: The four figures used in syllogistic reasoning.

|  | Figure 1 | Figure 2 | Figure 3 | Figure 4 |
| :---: | :---: | :---: | :---: | :---: |
| First Premise | a-b | b-a | a-b | b-a |
| Second Premise | $\mathrm{b}-\mathrm{c}$ | $\mathrm{c}-\mathrm{b}$ | $\mathrm{c}-\mathrm{b}$ | $\mathrm{b}-\mathrm{c}$ |



Figure 1: 'some a are not $c$ ' follows from IE4.

Recently, a new cognitive theory based on the Weak Completion Semantics (WCS) has been developed (Hölldobler, 2015). It has its roots in the ideas first expressed by Stenning and van Lambalgen (2008), which unfortunately had some technical mistakes. These were corrected by Hölldobler and Kencana Ramli (2009a), by using the three-valued Łukasiewicz (1920) logic. WCS has been successfully applied - among others - to the suppression task (Dietz, Hölldobler, \& Ragni, 2012), the selection task (Dietz, Hölldobler, \& Ragni, 2013), the belief-bias effect (Pereira, Dietz, \& Hölldobler, 2014a, 2014b; Dietz, 2017), to reasoning about conditionals (Dietz \& Hölldobler, 2015; Dietz, Hölldobler, \& Pereira, 2015) and to spatial reasoning (Dietz, Hölldobler, \& Höps, 2015). Hence, it was natural to ask whether WCS can also model syllogistic reasoning.

We introduce seven principles motivated by findings made in Cognitive Science and Computational Logic, and show how syllogisms together with these principles can be encoded in logic programs. After that we compare the predictions under WCS with the results of FOL, the syntactic rule based theory PSYCOP (Rips, 1994), the Verbal Model Theory (Polk \& Newell, 1995) and the Mental Model Theory (Johnson-Laird, 1983). ${ }^{3}$ The two model-based theories performed best in the meta-analysis (Khemlani \& Johnson-Laird, 2012).

The predictions of the theories FOL, PSYCOP, Verbal, and Mental Models for the syllogisms OA4, IE4, and IA2 and those of a significant percentage of the participants are depicted in Table 3, where the participants were 156

[^157]Table 3: Conclusions drawn by a significant percentage of participants are highlighted in gray and compared to the predictions of FOL, PSYCOP, Verbal, and Mental Models theory for OA4, IE4 and IA2. NVC denotes no valid conclusion.

| Participtants |  | FOL | PSYCOP | Verbal Models | Mental Models |
| :---: | :---: | :---: | :---: | :---: | :---: |
| OA4 | Oca | Oca | Oca, | Oca, | Oca, |
|  |  |  | Ica, lac | NVC | Oac, NVC |
| IE4 | Oac, | Oac | Oac, | NVC, | Oac, NVC |
|  | NVC |  | lac, Ica | Oac | Eac, Eca, Oca |
| IA2 | lac, Ica | NVC | NVC | NVC, Ica | lac, Ica, NVC |

high school or university students. A conclusion is drawn by a significant number of participants, if the number of participants chosing this particular conclusion is statistically too high for the conclusion to be chosen at random. The interested reader is referred to Khemlani and Johnson-Laird (2012) for more details.

FOL and the other three cognitive theories make different predictions. Each theory provides at least one prediction which is correct with respect to FOL and provides an additional prediction which is false with respect to FOL. Currently, the best results are achieved by the Verbal Models Theory which predicts $84 \%$ of the participants responses, closely followed by the Mental Model Theory with $83 \%$, whereas PSYCOP predicts $77 \%$ of the participants responses.

## Weak Completion Semantics

## Logic Programs

A (logic) program $P$ is a finite set of clauses of the form $A \leftarrow \top, A \leftarrow \perp$ or $A \leftarrow B_{1} \wedge \ldots \wedge B_{n}, n>0$, where $A$ is an atom, $B_{i}, 1 \leq i \leq n$, are literals, and $\top$ and $\perp$ denote truth and falsehood, respectively. Clauses are assumed to be universally closed. $A$ is called head and $\top, \perp$ as well as $B_{1} \wedge \ldots \wedge B_{n}$ are called body of the corresponding clause. Clauses of the form $A \leftarrow \top$ and $A \leftarrow \perp$ are called facts and assumptions, respectively. ${ }^{4} \neg A$ is assumed in $\mathcal{P}$ iff $\mathcal{P}$ contains an assumption with head $A$ and no other clause with head $A$ occurs in $\mathcal{P}$. We restrict terms to be constants and variables. We assume for each program $\mathcal{P}$ that the underlying alphabet consists precisely of the symbols occurring in $\mathcal{P}$ and that non-propositional programs contain at least one constant.
$\mathrm{g} \mathcal{P}$ denotes the set of all ground instances of clauses occurring in $\mathcal{P}$, where a ground instance of clause $C$ is obtained from $C$ by replacing each variable occurring in $C$ by a constant. A ground atom $A$ is defined in $g \mathcal{P}$ iff $g \mathcal{P}$ contains a clause whose head is $A$; otherwise $A$ is said to be undefined. $\operatorname{def}(A, \mathcal{P})=\{A \leftarrow \operatorname{Body} \mid A \leftarrow \operatorname{Body} \in \mathrm{~g} P\}$ is called definition of $A$ in $\mathcal{P}$. To clarify the definitions, consider $\mathcal{P}_{\text {ex }}$ :
$p(X) \leftarrow q(X) \wedge \neg r(X) \wedge s(X) . \quad q(a) \leftarrow \top . \quad r(a) \leftarrow \perp$.

[^158]Table 4: The truth tables for the connectives under three-valued Łukasiewicz logic. Note that for the unknown holds: $(\mathrm{U} \leftarrow \mathrm{U})=T$.

The second and the third clause is a fact and an assumption, respectively. $\mathrm{g} \mathcal{P}_{\text {ex }}$ is as follows:

$$
p(a) \leftarrow q(a) \wedge \neg r(a) \wedge s(a) . \quad q(a) \leftarrow \top . \quad r(a) \leftarrow \perp
$$

$p(a), q(a), r(a)$ are defined and $s(a)$ is undefined in $g \mathcal{P}_{\mathrm{ex}}$. Classical logic and logic programs are discussed in detail in e.g. (Hölldober, 2009; Lloyd, 1984).

## Three-Valued Lukasiewicz Logic

We consider the three-valued Łukasiewicz (1920) logic for which the corresponding truth values are true $(\top)$, false $(\perp)$ and unknown $(\mathrm{U})$. A three-valued interpretation $I$ is a mapping from the set of formulas to the set $\{\top, \perp, \mathrm{U}\}$. The truth value of a given formula under $I$ is determined according to the truth tables in Table 4. We represent an interpretation as a pair $I=\left\langle I^{\top}, I^{\perp}\right\rangle$ of disjoint sets of ground atoms, where $I^{\top}$ is the set of all atoms that are mapped to $\top$ by $I$, and $I^{\perp}$ is the set of all atoms that are mapped to $\perp$ by $I$. Atoms which do not occur in $I^{\top} \cup I^{\perp}$ are mapped to $U$. Let $I=\left\langle I^{\top}, I^{\perp}\right\rangle$ and $J=\left\langle J^{\top}, J^{\perp}\right\rangle$ be two interpretations: $I \subseteq J$ iff $I^{\top} \subseteq J^{\top}$ and $I^{\perp} \subseteq J^{\perp} . I(F)=\top$ means that formula $F$ is mapped to true under $I . \mathcal{M}$ is a model of $\mathcal{P}$ if it is an interpretation, which maps each clause occurring in $\mathrm{g} \mathcal{P}$ to $\top . I$ is the least model of $\mathcal{P}$ iff for any other model $J$ of $\mathcal{P}$ it holds that $I \subseteq J$.

## Least Models under the Weak Completion

For a given program $\mathcal{P}$, consider the following transformation: (1) For each ground atom $A$ which is defined in $g \mathcal{P}$, replace all clauses of the form $A \leftarrow \operatorname{Bod}_{1}, \ldots, A \leftarrow$ Body $_{m}$ occurring in $g \mathcal{P}$ by $A \leftarrow \operatorname{Bod}_{1} \vee \ldots \vee \operatorname{Body} y_{m}$. (2) Replace all occurrences of $\leftarrow$ by $\leftrightarrow$. The obtained set of formulas is called weak completion of $\mathcal{P}$ or $w c \mathcal{P} .{ }^{5}$

It has been shown by Hölldobler and Kencana Ramli (2009b) that programs as well as their weak completions admit a least model under three-valued Łukasiewicz logic. Moreover, the least model of $w c \mathcal{P}$ can be obtained as the least fixed point of the semantic operator $\Phi$, which is due to Stenning and van Lambalgen (2008). Let $I=\left\langle I^{\top}, I^{\perp}\right\rangle$ be an interpretation, then $\Phi_{\mathcal{P}}(I)=\left\langle J^{\top}, J^{\perp}\right\rangle$, is defined by:

$$
\begin{aligned}
J^{\top}= & \{A \mid A \leftarrow \operatorname{Bod} y \in \operatorname{def}(A, \mathcal{P}) \text { and } I(\operatorname{Bod} y)=\top\}, \\
J^{\perp}= & \{A \mid \operatorname{def}(A, \mathcal{P}) \neq \emptyset \text { and } \\
& I(\operatorname{Body})=\perp \text { for all } A \leftarrow \operatorname{Bod} y \in \operatorname{def}(A, \mathcal{P})\} .
\end{aligned}
$$

Weak Completion Semantics (WCS) is the approach to consider weakly completed programs, to compute their least

[^159]models, and to reason with respect to these models.
We write $\mathcal{P} \models_{w c s} F$ iff formula $F$ holds in the least model of $w c \mathcal{P}$. Consider again $\mathcal{P}_{\text {ex }}$. Starting with $\langle\emptyset, \emptyset\rangle$, we obtain:
$$
\Phi_{\mathcal{P}_{\mathrm{ex}}}(\langle\emptyset, \emptyset\rangle)=\langle\{q(a)\},\{r(a)\}\rangle=I_{1}=\Phi_{\mathcal{P}_{\mathrm{ex}}}\left(I_{1}\right) .
$$
$I_{1}$ is the least model of $w c \mathcal{P}_{\text {ex }}$.

## Integrity Constraints

An integrity constraint is an expression of the form $\mathrm{U} \leftarrow$ Body, where Body is a conjunction of literals and $U$ denotes the unknown. An interpretation $I$ maps an integrity constraint $\mathrm{U} \leftarrow$ Body to $\top$ iff $I(\operatorname{Body}) \subseteq\{\perp, \mathrm{U}\}$. Given an interpretation $I$ and a finite set IC of integrity constraints, $I$ satisfies IC iff all clauses occurring in IC are true under $I$.

## Abductive Logic Programming

In Logic Programming, abduction is the reasoning process of searching for explanations given a program and some observations, which do not follow from the program (Peirce, Hartshorne, \& Weiss, 1974). Explanations are usually restricted to certain formulas called abducibles. Let $\mathcal{P}$ be a ground program. The set of abducibles of $\mathcal{P}$ is
$\begin{aligned} \mathcal{A}_{\mathcal{P}} & =\{A \leftarrow \top \mid A \text { is undefined or } \neg A \text { is assumed in } \mathcal{P}\} \\ & \cup\{A \leftarrow \perp \mid A \text { is undefined in } \mathscr{P}\} .\end{aligned}$
An abductive framework consists of a program $\mathcal{P}$, a finite set $\mathcal{A}$ of abducibles, a finite set IC of integrity constraints, and an entailment relation. Let $\left\langle\mathcal{P}, \mathcal{A}\right.$, IC,$\left.\models_{w c s}\right\rangle$ be an abductive framework, $\mathcal{E} \subseteq \mathcal{A}$, and $O$ a non-empty set of literals called observation. An observation $O=\left\{o_{1}, \ldots, o_{n}\right\}$ is explained by $\mathcal{E}$ given $\mathcal{P}$ and $I C$ iff $\mathcal{P} \cup \mathcal{E} \models_{w c s} o_{1} \wedge \ldots \wedge o_{n}$ and $\mathcal{P} \cup \mathcal{E} \neq_{\text {wcs }}$ IC. $O$ is explained given $\mathcal{P}$ and $I C$ iff there exists an $\mathcal{E}$ such that $O$ is explained by $\mathcal{E}$ given $\mathscr{P}$ and IC. We prefer subset-minimal explanations. An explanation $\mathcal{E}$ is minimal iff there is no explanation $\mathcal{E}^{\prime}$ such that $\mathcal{E}^{\prime} \subset \mathcal{E}$.

Consider the program $\mathcal{P}=\{w \leftarrow g, w \leftarrow s, g \leftarrow r\} .{ }^{6}$ Suppose we observe $O=\{w\}$. Because the least model of wc $\mathcal{P}$ is $\langle\emptyset, \emptyset\rangle$ the observation does not follow. However, $O$ can be explained by $\mathcal{E}=\{s \leftarrow \top\}$. Starting with the empty interpretation we obtain:

$$
\begin{aligned}
& \Phi_{\mathcal{P} \cup \mathcal{E}}(\langle\emptyset, \emptyset\rangle)=\langle\{s\}, \emptyset\rangle \\
& \Phi_{\mathcal{P} \cup \mathcal{E}}(\langle\{s\}, \emptyset\rangle)=\langle\{s, w\}, \emptyset\rangle=\Phi_{\mathcal{P} \cup \mathcal{E}}(\langle\{s, w\}, \emptyset\rangle) . \\
& \langle\{s, w\}, \emptyset\rangle \text { is the least model of } w c(\mathcal{P} \cup \mathcal{E}) \text {. It entails } O \text {. } \mathcal{E} \text { is } \\
& \text { minimal, whereas } \mathcal{E}^{\prime}=\{s \leftarrow \top, r \leftarrow \top\} \text { is not. }
\end{aligned}
$$

## Seven Principles in Reasoning with Quantifiers

We will apply seven principles in developing a logical form for the representation of syllogisms.

## Licenses for Inferences (licenses)

Stenning and van Lambalgen (2008) propose to formalize conditionals by licenses for inferences. For example, the conditional for all $X$, if $p(X)$ then $q(X)$ is represented by the program $\{q(X) \leftarrow p(X) \wedge \neg a b(X), a b(X) \leftarrow \perp\}$. Its first

[^160]clause states that for all $X, q(X)$ holds if $p(X)$ and $\neg a b(X)$ holds, i.e. nothing abnormal for $X$ is known. This in turn is stated by the second clause. Clauses are assumed to be universally closed and, hence, the universal quantifier can be omitted.

## Existential Import and Gricean Implicature (import)

Humans seem to understand quantifiers differently due to a pragmatic understanding of language. For instance, in natural language we normally do not quantify over things that do not exist. Consequently, for all implies there exists. This appears to be in line with human reasoning and has been called the Gricean Implicature (Grice, 1975). Several theories like the theory of mental models (Johnson-Laird, 1983) or mental logic (Rips, 1994) assume that the sets we quantify over are not empty. Likewise, Stenning and van Lambalgen (2008) have shown that humans require existential import for a conditional to be true. Furthermore, as mentioned in Khemlani and Johnson-Laird (2012), the quantifier 'some a are b' often implies that 'some a are not $b$ ', which again can be explained by assuming the Gricean Implicature: Someone would not state 'some a are $b$ ' if that person knew that 'all a are $b$ '. As the person does not say 'all $a$ are $b$ ' but instead 'some a are $b$ ', we have to assume that 'not all a are $b$ ', which in turn implies 'some a are not $b$ '.

## Unknown Generalization (unknownGen)

Humans seem to distinguish between 'some $y$ are $z$ ' and 'all $y$ are $z$ ', as we have already explained in the previous paragraph. Accordingly, if we observe that an object $o$ belongs to $y$ and $z$ then we do not want to conclude both, 'some $y$ are $z$ ' and 'all $y$ are $z$ '. In order to prevent such unwanted conclusions we introduce the following principle: if we know that 'some $y$ are $z$ ' then there must not only be an object $o_{1}$ which belongs to $y$ and $z$ (by Gricean implicature) but there must be another object $o_{2}$ which belongs to $y$ and for which it is unknown whether it belongs to $z$.

## Converse Interpretation (converse)

Although there appears to be some evidence that humans seem to distinguish between 'some y are $z$ ' and 'some $z$ are $y$ ' (see the results reported in Khemlani \& Johnson-Laird, 2012) we propose that premises of the form lab imply lba and vice versa. If there is an object which belongs to $y$ and $z$, then there is also an object which belongs to z and y .

## Block Conclusions by Double Negation (doubleNeg)

Consider the following two negative sentences (i.e. including negation) and the positive one: ' If not $a$, then $b$. If not $b$ then $c . a$ is true.' The program representing these sentences is $\mathcal{P}=\{b \leftarrow \neg a, c \leftarrow \neg b, a \leftarrow \top\}$. The weak completion of $\mathscr{P}$ is wc $\mathcal{P}=\{b \leftrightarrow \neg a, c \leftrightarrow \neg b, a \leftrightarrow T\}$. Its least model is $\langle\{a, c\},\{b\}\rangle$, and thus $a$ and $c$ are true: $a$ is true because it is a fact and $c$ is true by the negation of $b . \quad b$ is derived to be false because the negation of $a$ is false. This example shows that under WCS, a
positive conclusion ( $c$ being true) can be derived from two clauses, which include negation. Considering the results of the participants' responses in (Khemlani \& Johnson-Laird, 2012), they seem not to draw conclusions through double negation. Accordingly, we block them through abnormalities.

## Search Alternative Conclusions to NVC (abduction)

Our hypothesis is that when participants are faced with a NVC conclusion ('no valid conclusion'), they might not want to accept this conclusion and proceed to check whether there exists unknown information that is relevant. This information may be explanations to facts in our program, and we model such repair mechanism using abduction. Facts in our programs come either from an existential import or from unknown generalization. We use only the first as source for observations, since they are used directly to infer new information.

## Negation by Transformation (transformation)

A negative literal cannot be the head of a clause in a program. In order to represent a negative conclusion $\neg p(X)$ an auxiliary atom $p^{\prime}(X)$ is introduced together with a clause $p(X) \leftarrow \neg p^{\prime}(X)$ and the integrity constraint $\mathrm{U} \leftarrow p(X) \wedge p^{\prime}(X)$. This is a widely used technique in logic programming. Together with the principle licences for inferences, the additional clause is $p(X) \leftarrow \neg p^{\prime}(X) \wedge \neg a b(X)$.

## Representation of Quantified Statements

Based on the principles presented in the previous section, we discuss the representation of three examples.

OA4 The two syllogistic premises of OA4 are as follows:
Some $b$ are not $a$. (Oba)
all $b$ are $c .(A b c)$
The program $\mathcal{P}_{\mathrm{OA} 4}$ representing OA4 consists of:

$$
\begin{array}{lr}
a^{\prime}(X) \leftarrow b(X) \wedge \neg a b_{\text {bna }}(X) . & \text { (transformation \& licenses) } \\
a b_{b n a}\left(o_{1}\right) \leftarrow \perp . 亡 . \\
a(X) \leftarrow \neg a^{\prime}(X) \wedge \neg a b_{\text {naa }}(X) . & \text { (transformation \& licenses) } \\
(\text { trenses) } \\
b\left(o_{1}\right) \leftarrow T . & \text { (import) } \\
\left.a o_{2}\right) \leftarrow T . & \text { (unknownGen) } \\
a b_{\text {naa }}\left(o_{1}\right) \leftarrow \perp . & \text { (doubleNeg \& licenses) } \\
a b_{\text {naa }}\left(o_{2}\right) \leftarrow \perp . & \text { (doubleNeg \& licenses) } \\
c(X) \leftarrow b(X) \wedge \neg a b_{b c}(X) . & \text { (licenses) } \\
a b_{b c}(X) \leftarrow \perp . & \text { (licenses) } \\
b\left(o_{3}\right) \leftarrow T . & \text { (import) }
\end{array}
$$

In addition, we have the following integrity constraint:

$$
\mathrm{U} \leftarrow a(X) \wedge a^{\prime}(X) .
$$

(transformation)

IE4 The two syllogistic premises of IE4 are as follows:
Some bare a. (lba) no bare c. (Ebc)
The program $\mathcal{P}_{\text {IE4 }}$ representing IE4 consists of:

$$
\begin{aligned}
& a(X) \leftarrow b(X) \wedge \neg a b_{b a}(X) . \\
& a b_{b a}\left(o_{1}\right) \leftarrow \perp . \\
& b\left(o_{1}\right) \leftarrow \mathrm{T} . \\
& b\left(o_{2}\right) \leftarrow T .
\end{aligned}
$$

(licenses)
(licenses\&unknownGen)
(import)
(unknownGen)

$$
\begin{array}{lr}
b(X) \leftarrow a(X) \wedge \neg a b_{a b}(X) . & \text { (converse\&licenses) } \\
a b_{a b}\left(o_{3}\right) \leftarrow \perp . & \text { (converse\&licenses\&unknownGen) } \\
a\left(o_{3}\right) \leftarrow T . & \text { (converse\&import) } \\
a\left(o_{4}\right) \leftarrow T . & \text { (converse\&unknownGen) } \\
c^{\prime}(X) \leftarrow b(X) \wedge \neg a b_{b n c}(X) . & \text { (transformation \&licenses) } \\
a b_{b n c}(X) \leftarrow \perp . & \text { (licenses) } \\
c(X) \leftarrow \neg c^{\prime}(X) \wedge \neg a b_{n c c}(X) . & \text { (transformation \&licenses) } \\
b\left(o_{5}\right) \leftarrow T . & \text { (import) } \\
a b_{n c c}\left(o_{5}\right) \leftarrow \perp . & \text { (licenses\&doubleNeg) }
\end{array}
$$

In addition, we have the following integrity constraint:

$$
\mathrm{U} \leftarrow c(X) \wedge c^{\prime}(X) . \quad \text { (transformation) }
$$

IA2 The two syllogistic premises of IA2 are as follows:

## Some $b$ are $a$. (lba)

all $c$ are $b$. (Acb)
Program $\mathcal{P}_{\mathrm{A} 22}$ consists of the following clauses:

| $a(X) \leftarrow b(X) \wedge \neg a b_{b a}(X)$. | (licenses) |
| :--- | ---: |
| $a b_{b a}\left(o_{1}\right) \leftarrow \perp$. | (licenses\&unknownGen) |
| $b\left(o_{1}\right) \leftarrow T$. | (import) |
| $b\left(o_{2}\right) \leftarrow \mathrm{T}$. | (unknownGen) |
| $b(X) \leftarrow a(X) \wedge \neg a b_{a b}(X)$. | (converse\&licenses) |
| $a b_{a b}\left(o_{3}\right) \leftarrow \perp$. | (converse\&licenses\&unknownGen) |
| $a\left(o_{3}\right) \leftarrow T$. | (converse\&import) |
| $a\left(o_{4}\right) \leftarrow \mathrm{T}$. | (converse\&unknownGen) |
| $b(X) \leftarrow c(X) \wedge \neg a b_{c b}(X)$. | (licenses) |
| $a b_{c b}(X) \leftarrow$. | (licenses\&unknownGen) |
| $c\left(o_{5}\right) \leftarrow \mathrm{T}$. | (import) |

## Entailment of Syllogisms

Khemlani and Johnson-Laird (2012) do not report a formal definition for the entailment of syllogisms. They use FOL as a normative theory and test if the conclusions drawn by the participants are correct with respect to a first-order representation of a syllogism. In FOL, a set of formulas $\mathcal{K}$ entails a formula $F(\mathcal{K} \models F)$ if all interpretations which map $\mathcal{K}$ to true map $F$ to true as well. We believe that the entailment relation should reflect the principles on which the representation is based. In the following, an entailment relation regarding WCS is presented, where $y z$ is to be replaced by $a c$ or $c a$.

- $\mathcal{P} \models A y z$ iff there exists an object $o$ such that $\mathcal{P} \models \models_{w c s} y(o)$ and for all $o$ we find that if $\mathcal{P} \models_{w c s} y(o)$ then $\mathcal{P} \models_{w c s} z(o)$.
- $\mathcal{P} \models E y z$ iff there exists an object $o$ such that $\mathcal{P} \models_{w c s} z(o)$ and for all $o$ we find that if $\mathcal{P} \models_{w c s} z(o)$ then $\mathcal{P} \models_{w c s} \neg y(o)$.
- $\mathcal{P} \models I y z$ iff there exists an object $o_{1}$ such that $\mathcal{P} \models_{w c s} y\left(o_{1}\right) \wedge z\left(o_{1}\right)$ and there exists an object $o_{2}$ such that $\mathcal{P} \models_{w c s} y\left(o_{2}\right)$ and $\mathcal{P} \not \models_{w c s} z\left(o_{2}\right)$ and there exists an object $o_{3}$ such that $\mathcal{P} \models_{w c s} z\left(o_{3}\right)$ and $\mathcal{P} \not \models_{w c s} y\left(o_{3}\right)$.
- $\mathcal{P} \models O y z$ iff there exists an object $o_{1}$ such that $\mathcal{P} \models_{w c s} y\left(o_{1}\right) \wedge \neg z\left(o_{1}\right)$ and there exists an object $o_{2}$ such that $\mathcal{P} \models_{w c s} y\left(o_{2}\right)$ and $\mathcal{P} \not \models_{w c s} \neg z\left(o_{2}\right)$.

In case we can not conclude any of these moods, then no valid conclusion is entailed, denoted as $\mathcal{P} \models$ NVC.

## Accuracy of Predictions

We evaluate our approach the same way as the other theories have been evaluated (Khemlani \& Johnson-Laird, 2012): There are nine different answers for each of the 64 syllogisms that can be ordered in a list: Aac, Eac, lac, Oac, Aca, Eca, Ica, Oca, and NVC. We consider two lists: a list for WCS predictions and a list for participants' answers. In the WCS list, we assign 1 to an answer if it is predicted under WCS; else we assign 0 . In the list with the participants' answers we use a threshold function and assign 1 to answers that were drawn by more than $16 \%$ of the participants; else we assign 0 . Both lists can then be compared for their congruence as follows, where $i$ is the $i$ th element of both lists:

$$
\operatorname{Comp}(i)= \begin{cases}1 & \text { if both have the same value for } i \text { th element } \\ 0 & \text { otherwise }\end{cases}
$$

The matching percentage of this syllogism is then computed by $\sum_{i=1}^{9} \operatorname{COMP}(i) / 9$. Note that the percentage of the match does not only take into account when WCS correctly predicts a conclusion, but also whenever it correctly rejected a conclusion. The average percentage of accuracy is then the average of the matching percentage of all 64 syllogisms.

## OA4 - Perfect Match ( $\mathbf{1 0 0 \%}$ )

Consider $\mathcal{P}_{\mathrm{OA} 4}$. The least model of wc $\mathcal{P}_{\mathrm{OA} 4}, I=\left\langle I^{\top}, I^{\perp}\right\rangle$, is:
$I^{\top}=\left\{b\left(o_{1}\right), b\left(o_{2}\right), b\left(o_{3}\right), a b_{c a}\left(o_{1}\right), a^{\prime}\left(o_{1}\right), c\left(o_{1}\right), c\left(o_{2}\right), c\left(o_{3}\right)\right\}$, $I^{\perp}=\left\{a b_{b n a}\left(o_{1}\right), a b_{\text {naa }}\left(o_{1}\right), a b_{b c}\left(o_{1}\right), a b_{b c}\left(o_{2}\right), a b_{b c}\left(o_{3}\right), a\left(o_{1}\right)\right\}$. This model entails only the conclusion 'some $c$ are not $a^{\prime}$ (Oca): There exists an object, viz. $o_{1}$, such that $\mathcal{P}_{\mathrm{OA} 4}==_{w c s} c\left(o_{1}\right) \wedge \neg a\left(o_{1}\right)$ and there exists another object, viz. $o_{2}$, such that $\mathcal{P}_{\mathrm{OA} 4} \models_{w c s} c\left(o_{2}\right)$ and $\mathcal{P}_{\mathrm{OA} 4} \not \vDash_{w c s} \neg a\left(o_{2}\right)$.

## IE4 - Partial Match (89\%)

Consider $\mathcal{P}_{\text {IE } 4}$. The least model of wc $\mathcal{P}_{\text {IE } 4}, I=\left\langle I^{\top}, I^{\perp}\right\rangle$, is

$$
\begin{aligned}
I^{\top} & =\left\{a\left(o_{i}\right) \mid i \in\{1,3,4\}\right\} \cup\left\{b\left(o_{i}\right) \mid i \in\{1,2,3,5\}\right\} \\
& \cup\left\{c^{\prime}\left(o_{i}\right) \mid i \in\{1,2,3,5\}\right\} \\
I^{\perp} & =\left\{a b_{b a}\left(o_{1}\right), a b_{a b}\left(o_{3}\right), a b_{n c c}\left(o_{5}\right), a b_{b n c}\left(o_{5}\right)\right\} \\
& \cup\left\{a b_{b n c}\left(o_{i}\right) \mid i \in\{1,2,3,4,5\}\right\} \cup\left\{c\left(o_{i}\right) \mid i \in\{1,2,3,5\}\right\} .
\end{aligned}
$$

This model entails only 'Some a are not $c$ ' ( Oac ).

## IA2 - Explain NVC: Perfect Match (100\%)

Consider $\mathcal{P}_{\mathrm{AA} 2}$. The least model of wc $\mathcal{P}_{\mathrm{A} 2}, I=\left\langle I^{\top}, I^{\perp}\right\rangle$, is

$$
\begin{aligned}
& I^{\top}=\left\{a\left(o_{1}\right), a\left(o_{3}\right), a\left(o_{4}\right), b\left(o_{1}\right), b\left(o_{2}\right), b\left(o_{3}\right), b\left(o_{5}\right), c\left(o_{5}\right)\right\}, \\
& I^{\perp}=\left\{a b_{b a}\left(o_{1}\right), a b_{a b}\left(o_{3}\right)\right\} \cup\left\{a b_{c b}\left(o_{i}\right) \mid i \in\{1,2,3,4,5\}\right\} .
\end{aligned}
$$

This model entails 'no valid conclusion' (NVC). However, a significant percentage of participants answered lac and Ica, despite IA2 being an invalid syllogism in classical FOL. According to the sixth principle, abduction, we believe that these participants might have searched for alternatives to NVC. We model this by applying skeptical abductive reasoning.

Each head of an existential import generates a single observation. We apply abduction sequentially to each of them. To prevent empty explanations we remove from the current program the fact that generated the observation.

For each observation and each of its minimal explanations we compute the least model of the weak completion of the program extended with the explanation and collect all entailed syllogistic conclusions. Observations that cannot be explained are filtered out. The set Answers consists of all entailed conclusions for the observations left. For the final conclusions, we apply skeptical reasoning, i.e., the final answer to the current syllogism is given by FinalAnswer $=\bigcap_{A \in \text { Answers }} A$. In the case that FinalAnswer is empty, we entail the NVC conclusion.

Reconsider IA2, where the observations are $O_{1}=\left\{b\left(o_{1}\right)\right\}$, $O_{2}=\left\{a\left(o_{3}\right)\right\}$ and $O_{3}=\left\{c\left(o_{5}\right)\right\}$. If we examine $O_{i}=\{o\}$ with $i \in\{1,2,3\}$, then we will try to find an explanation for $O_{i}$ with respect to $\mathcal{P}_{\mathrm{A} 2} \backslash\{o \leftarrow T\} .^{7}$ The set of abducibles is:

$$
\begin{aligned}
& \cup\left\{c\left(o_{i}\right) \leftarrow \top, \quad c\left(o_{i}\right) \leftarrow \perp \quad \mid i \in\{1,2,3,4\}\right\} \\
& \cup\left\{a b_{c b}\left(o_{5}\right) \leftarrow \top \quad \mid i \in\{1,2,3,4,5\}\right\} \\
& \cup\left\{a b_{b a}\left(o_{1}\right) \leftarrow \top, a b_{a b}\left(o_{3}\right) \leftarrow \top\right\} .
\end{aligned}
$$

$\mathcal{E}_{1}=\left\{c\left(o_{1}\right) \leftarrow \top\right\} \quad$ and $\quad \mathcal{E}_{2}=\left\{c\left(o_{3}\right) \leftarrow \top, a b_{b a}\left(o_{3}\right) \leftarrow \perp\right\}$ are the minimal explanations for $O_{1}$ and $O_{2}$, respectively. Note that for $O_{3}$ there is no explanation.

Consider the observation $O_{1}=\left\{b\left(o_{1}\right)\right\}$ and the program $\mathcal{P}_{\mathrm{IA} 2}^{1}=\left(\mathcal{P}_{\mathrm{AA} 2} \backslash\left\{b\left(o_{1}\right) \leftarrow \top\right\}\right) \cup \mathcal{E}_{1}$. The least model of wc $\mathcal{P}_{\mathrm{IA} 2}^{1}$ is $\left\langle I^{\top} \cup\left\{c\left(o_{1}\right)\right\}, I^{\perp}\right\rangle$ where $\left\langle I^{\top}, I^{\perp}\right\rangle$ is the least model of wc $\mathcal{P}_{\mathrm{A} 2}$, as defined before. Thus, $c\left(o_{1}\right)$ is newly entailed to be true after applying abduction. This model entails what participants concluded, namely lac and lca. lac is entailed as there exists an object, viz. $o_{1}$, such that $\mathcal{P}_{\mathrm{IA} 2}^{1} \models_{w c s} a\left(o_{1}\right) \wedge c\left(o_{1}\right)$ and there exists another object, viz. $o_{4}$, such that $\mathcal{P}_{\mathrm{AA} 2}^{1}=_{w c s} a\left(o_{4}\right)$ and $\mathcal{P}_{\mathrm{AA} 2}^{1} \not \models_{w c s} c\left(o_{4}\right)$, and there exists another object , viz. $o_{5}$, such that $\mathcal{P}_{1 \mathrm{~A} 2}^{1} \models_{w c s} c\left(o_{5}\right)$ and $\mathcal{P}_{\mathrm{IA} 2}^{1} \not \vDash_{w c s} a\left(o_{5}\right)$. Analogously, 'some c are a' (Ica) holds.

For the observation $O_{2}=\left\{a\left(o_{3}\right)\right\}$ we consider the program $\mathcal{P}_{\text {IA } 2}^{2}=\left(\mathcal{P}_{\text {IA } 2} \backslash\left\{a\left(o_{3}\right) \leftarrow \top\right\}\right) \cup \mathcal{E}_{2}$. The least model of $\mathcal{P}_{\text {IA } 2}^{2}$ also entails the conclusions lac and Ica.

Answers $\left(\mathcal{P}_{\mathrm{IA}_{2}}\right)=\{\{|a c| c a\},,\{|a c| c a\}$,$\} is the collection$ of all conclusions. FinalAnswer $\left(\mathcal{P}_{\mathrm{A} 2}\right)=\{|a c| c a$,$\} consists of$ the skeptically entailed conclusions, i.e. it is the intersection of all conclusions, which in this case are 'some a are c' (lac) and 'some c are $a$ ' (Ica).

## Overall Accuracy of $89 \%$

The results of the three examples formalized under WCS are summarized and compared to FOL, PSYCOP, the Verbal, and the Mental Model Theory in Table 5. For some syllogisms the conclusions drawn by the participants and WCS are identical and for some syllogisms the conclusions drawn by the participants and WCS overlap. Combining the syllogistic premises representation and entailment rules for all 64 syllogistic premises and applying abduction when NVC was entailed (which happened in 43 cases), we accomplished an average of $89 \%$ accuracy in our predictions. In 18 cases we have a perfect match, in 30 cases the match is $89 \%$, in 13

[^161]Table 5: The conclusions drawn by a significant percentage of participants are highlighted in gray and compared to the predictions of the theories FOL, PSYCOP, Verbal, and Mental Models as well as WCS for the syllogisms OA4, IE4, and IA2.

cases the match is $78 \%$, and in the remaining three cases the match is $67 \%$. We achieve the best performance compared to the other state-of-the-art cognitive theories with the current best performance of $84 \%$ (Verbal Model Theory).

## Conclusions

We developed seven principles for modeling a logical form for the representation of quantified statements in human reasoning, mostly motivated from findings in Cognitive Science. We show how these principles can be encoded within a computational logic approach, the Weak Completion Semantics. After that we discuss the predictions of three examples under WCS and compare them to the conclusions humans draw from in (Khemlani \& Johnson-Laird, 2012). The result with respect to all 64 syllogistic premises under WCS shows that we achieve the best results with a prediction of $89 \%$, compared to the results of other cognitive theories.

## References

Dietz, E.-A. (2017). A computational logic approach to the belief bias in human syllogistic reasoning. In P. Brézillon, R. Turner, \& C. Penco (Eds.), 10th int. and interdisciplinary conference on modeling and using context LNAI (Vol. 10257, p. 691-707). Springer.
Dietz, E.-A., \& Hölldobler, S. (2015). A new computational logic approach to reason with conditionals. In F. Calimeri, G. Ianni, \& M. Truszczynski (Eds.), Logic programming and nonmonotonic reasoning - 13th int. conference, LNAI (Vol. 9345, pp. 265-278). Springer.
Dietz, E.-A., Hölldobler, S., \& Höps, R. (2015). A computational logic approach to human spatial reasoning. In IEEE symposium series on computational intelligence (pp. 1627-1634). IEEE.
Dietz, E.-A., Hölldobler, S., \& Pereira, L. M. (2015). On conditionals. In G. Gottlob, G. Sutcliffe, \& A. Voronkov (Eds.), Global conference on artificial intelligence, Epic Series in Computing, 36 (pp. 79-92). EasyChair.
Dietz, E.-A., Hölldobler, S., \& Ragni, M. (2012). A computational logic approach to the suppression task. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proc.
of the 34th annual conference of the Cognitive Science Society (pp. 1500-1505).
Dietz, E.-A., Hölldobler, S., \& Ragni, M. (2013). A computational logic approach to the abstract and the social case of the selection task. In 11th int. symposium on logical formalizations of commonsense reasoning.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. L. Morgan (Eds.), Syntax and semantics (Vol. 3). Academic Press.
Hölldober, S. (2009). Logik und Logikprogrammierung (Vol. 1: Grundlagen). Synchron Publishers.
Hölldobler, S. (2015). Weak completion semantics and its applications in human reasoning. In U. Furbach \& C. Schon (Eds.), CEUR WS proc. on bridging the gap between human and automated reasoning (pp. 2-16).
Hölldobler, S., \& Kencana Ramli, C. D. (2009a). Logic programs under three-valued Łukasiewicz semantics. In P. M. Hill \& D. S. Warren (Eds.), 25th int. conference on logic programming, LNCS 5649 (pp. 464-478). Springer.
Hölldobler, S., \& Kencana Ramli, C. D. (2009b). Logics and networks for human reasoning. In int. conference on artificial neural networks, LNCS, 5769 (pp. 85-94). Springer.
Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Harvard University Press.
Khemlani, S., \& Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. Psychological Bulletin, 427-457.
Lloyd, J. W. (1984). Foundations of logic programming. Springer.
Łukasiewicz, J. (1920). O logice trójwartościowej. Ruch Filozoficzny, 5, 169-171.
Peirce, C., Hartshorne, C., \& Weiss, P. (1974). Collected papers of charles sanders peirce. Belknap Press of Harvard University Press.
Pereira, L. M., Dietz, E.-A., \& Hölldobler, S. (2014a). An abductive reasoning approach to the belief-bias effect. In C. Baral, G. D. Giacomo, \& T. Eiter (Eds.), Principles of knowledge representation and reasoning: Proc. of the 14th int. conference (p. 653-656). AAAI Press.
Pereira, L. M., Dietz, E.-A., \& Hölldobler, S. (2014b). Contextual abductive reasoning with side-effects. In I. Niemelä (Ed.), Theory and practice of logic programming, 14 (p. 633-648). Cambridge University Press.
Polk, T. A., \& Newell, A. (1995). Deduction as verbal reasoning. Psychological Review, 102(3), 533.
Rips, L. J. (1994). The psychology of proof: Deductive reasoning in human thinking. MIT Press.
Smith, R. (1989). Aristotle's prior analytics. Hackett.
Stenning, K., \& van Lambalgen, M. (2008). Human reasoning and cognitive science. MIT Press.

# Speakers' gestures predict the meaning and perception of iconicity in signs 

Gerardo Ortega ${ }^{1,2}$ (gerardo.ortega@mpi.nl), Annika Schiefner ${ }^{1}$ (a.schiefner@student.ru.nl), \& Aslı Özyürek ${ }^{1,2,3}$ (asli.ozyurek@mpi.nl)<br>${ }^{1}$ Centre for Language Studies, Radboud University,<br>${ }^{2}$ Max Planck Institute for Psycholinguistics, and ${ }^{3}$ Donders Centre for Cognition<br>Wundtlaan 1, 6525XD Nijmegen, The Netherlands


#### Abstract

Sign languages stand out in that there is high prevalence of conventionalised linguistic forms that map directly to their referent (i.e., iconic). Hearing adults show low performance when asked to guess the meaning of iconic signs suggesting that their iconic features are largely inaccessible to them. However, it has not been investigated whether speakers' gestures, which also share the property of iconicity, may assist non-signers in guessing the meaning of signs. Results from a pantomime generation task (Study 1) show that speakers' gestures exhibit a high degree of systematicity, and share different degrees of form overlap with signs (full, partial, and no overlap). Study 2 shows that signs with full and partial overlap are more accurately guessed and are assigned higher iconicity ratings than signs with no overlap. Deaf and hearing adults converge in their iconic depictions for some concepts due to the shared conceptual knowledge and manual-visual modality.


Keywords: iconicity; gesture; sign language; embodied cognition

## Introduction

A question that has puzzled psychologists and linguists for decades is to what extent sign iconicity is accessible to individuals with no knowledge of a sign language. Iconicity, defined as the direct relationship between a linguistic form and its referent, is a ubiquitous property of sign languages observable at many of their linguistic levels of organisation (Cuxac, 1999; Perniss, Thompson, \& Vigliocco, 2010; Pietrandrea, 2002). Sign-naïve adults can accurately guess the meaning of only a small proportion of signs (Griffith, Robinson, \& Panagos, 1981; Grosso, 1993; Klima \& Bellugi, 1979; Pizzuto \& Volterra, 2000), but it has been hard to establish what factors allow them to map certain features of a sign to its correct referent. In an attempt to shed light on this question, we look at the iconic gestures produced by hearing non-signers. Given that iconic gestures are expressed through the same (manual-visual) modality, and importantly, they also share the property of iconicity (Kendon, 2004; McNeill, 1992), we entertain the hypothesis that non-signers may rely on their own gestural repertoire to make form-meaning judgements about signs.

## Iconicity in gesture and sign

Gestures are a fundamental aspect of human communication and are present in all ages and cultures (Kendon, 2004; McNeill, 1992). Gestures are holistic units highly integrated with speech that together convey unified semantic information of a multimodal utterance (Kelly, Creigh, \&

Bartolotti, 2010; McNeill, 1992). Sign languages, in contrast, occur independently from speech, and critically, they have the same levels of linguistic organisation as those reported in spoken languages (Sandler \& Lillo-Martin, 2006).

One point of intersection between sign and gesture is iconicity. Speakers can depict through iconic gestures the visual form of a concept and integrate them with speech as part of a multimodal message. For instance, when a speaker says 'I'll be outside' while producing the gesture of smoking it is clear to the interlocutor that she is going for a cigarette. On the other hand, a large proportion of a signed lexicon has iconic motivation (Pietrandrea, 2002), and crucially, signs may have overlapping structures as gestures (e.g., the sign TO-SMOKE depicts a person smoking a cigarette).

The similarities between sign and gesture were overlooked for many decades, but in recent years scholars have begun systematically comparing both modes of manual communication to shed light on their differences and similarities (e.g., Cormier, Schembri, \& Woll, 2013; Goldin-Meadow \& Brentari, 2015; Perniss, Özyürek, \& Morgan, 2015; Quinto-Pozos \& Parrill, 2015). Given the growing body of evidence showing that gestures and signs share more forms and functions than previously assumed (arguably due to the shared manual-visual modality) (Perniss et al., 2015), we investigate whether non-signing adults fall back on their own gestural repertoire to make judgements about conventionalised signs. The aim of the present study is therefore to investigate whether the overlap in form between signs (i.e., linguistic structure) and gestures (i.e., iconic depictions) predicts non-signers' ability to guess the meaning of signs and assign iconicity ratings.

## Perception of sign iconicity

Iconicity and the extent to which sign-naïve adults can understand the meaning of iconic signs has been a central focus of attention in sign research. The first investigations on the topic demonstrated that iconicity is not easily accessible to non-signers and that the meaning of signs is very difficult to access. In their seminal study, Klima and Bellugi (1979) asked hearing adults without any knowledge of a sign language to guess the meaning of a set of signs. When signs were presented in isolation and when they had to select the correct meaning out of five plausible candidates, participants showed a very low success rate (less than $10 \%$ ). They showed significant improvement, however, when they were presented the sign along with its English translation, and were asked to explain the iconic relationship
between the sign and its meaning. Participants showed overall agreement in that they were able to accurately describe the iconic motivation of more than $50 \%$ of the signs (e.g., most participants agreed that the sign VOTE depicted a person putting a ballot in a box). This study set a benchmark in sign language research and convincingly argued that iconicity is difficult to access by hearing nonsigners and that the notion of iconicity is better understood as a property that lies in a continuum with the meaning of some signs being more transparent than others.
Another study highlights the possibility that similarities between signs and the gestures used by the hearing community may assist sign-naïve participants in guessing the meaning of signs. Grosso (1993) showed a set of iconic and arbitrary signs in Italian Sign Language (LIS) to hearing non-signing adults and asked them to guess their meaning. Participants could not provide a correct response for a large proportion of signs ( $76 \%$ ) but they were very accurate for a considerable number of items on the list (24\%). A detailed analysis of the correctly guessed items revealed that these signs resemble the emblems commonly used by Italian speakers (e.g., the sign GOOD has the same form and meaning as the emblem used by hearing Italians). Emblems have a conventionalised, culture-specific form and meaning (Kendon, 1995, 2004) so when non-signing adults are confronted by signs that overlap in structures, they rightly assume that they also share the same meaning. This study is one of the first to suggest that non-signers' ability to guess the meaning of signs is based on the structural similarities between conventionalised (linguistic) signs and the gestures produced by the surrounding speaking community.
A limitation of this study is that it presupposes that only emblems facilitate the accurate guessing of the meaning of signs but does not say how other types of gestures may also be recruited. Emblems have highly conventionalised hand configurations, are used for specific pragmatic purposes (Kendon, 1995, 2004), and have mental representations akin to those of abstract words (Gunter \& Bach, 2004), so they are retrievable gestural entities that can be compared with convetionsalised signs. However, other types of iconic gestures may also be used as a basis to make judgments about the meaning of signs. In this study, we turn to the systematic iconic gestures shared in a community of speakers to investigate how overlap in form with conventionalised signs influences meaning-based judgements about signs.

## Systematicity in iconic gestures

The form of iconic gestures has been assumed to be variable, with their structure depending on the context in which they are used, the interlocutor, and the communicative intent of the speaker. It has been assumed that individuals tailor their gestures to the main focus of a conversation and as a result they vary in form and meaning from one conversation to the next (Müller, 2013). However, recent studies have found that contrary to this received knowledge, the iconic gestures produced by hearing adults exhibit a high degree of systematicity, and tend to represent very similar forms across individuals.

For instance, it has been found that the iconic co-speech gestures used in object descriptions are highly systematic and their form depends on the physical properties of the referent (Masson-Carro, Goudbeek, \& Krahmer, 2015). Objects that can be manipulated with the hands (e.g., a pen) are represented with gestures mimicking how the object is held; while objects with low manipulability affordances (e.g., a sink) are represented through gestures outlining their shape. A striking degree of systematicity has also been reported in elicited silent gestures (i.e., pantomimes). When asked to express concepts in pantomime, participants tend to systematically differentiate actions from tools through distinct gestural forms (i.e., re-enactment of bodily movements for verbs and handshapes representing the form of objects for nouns) (Padden et al., 2013; Padden, Hwang, Lepic, \& Seegers, 2015). More recently, high degree of systematicity in the structure of pantomimes has also been found across different semantic domains and for geographically unrelated cultures. Ortega and Özyürek (2016) elicited pantomimes from Dutch and Mexican adults and found that both groups employ remarkably similar strategies to depict referents. Through the implementation of specific types of iconic representations and their combinations, participants systematically represent concepts across different semantic domains. These pantomimes bare strong resemblance with the structures of recently discovered sign languages (Safar \& Petatillo, in preparation), so it has been argued that pantomimes reveal some of the cognitive dispositions that give rise to a signed lexicon in emerging sign languages.

The relevance of these studies is two-fold: first, they demonstrate that iconic gestural depictions are not as variable as previously assumed, but rather are deployed systematically to represent concrete concepts within specific semantic domains. Second, such systematicity results in shared knowledge about some manual forms across a community of speakers. As a consequence, individuals are likely to have expectations of how a concept should be represented in the manual-visual modality - at least for a set of referents. This has important implications for the perception of sign iconicity by non-signers. Non-signing adults confronted by conventionalised signs for the first time will not make judgements about their meaning in a vacuum. Rather, they are likely to fall back on their gestural knowledge to make judgments about the meaning of iconic signs.

## The Present Study

Based on evidence that many iconic gestures are highly systematic across individuals (Masson-Carro et al., 2015; Ortega \& Özyürek, 2016; Padden et al., 2013, 2015; van Nispen, van de Sandt-Koenderman, Mol, \& Krahmer, 2014) it is possible to assume that non-signing adults have at their disposal a cohort of shared gestures with specific forms and meanings on which they may base their judgment about signs. In order to test this hypothesis, we carried out two studies. In Study 1 we elicited pantomimes from nonsigning adults to determine which gestures were the most systematic across participants. Once these pantomimes were
selected, we compared them to signs from Sign Language of the Netherlands (NGT) and looked for signs that overlapped in form to different degrees (full, partial, or no overlap). These signs served as stimulus materials for Study 2. In this study, a different group of participants were presented with the signs and were asked, first, to guess their meaning. After they gave their response, they were given the correct translation, and then were asked to give iconicity ratings. The prediction is that when signs map directly to their gesture non-signing adults will be more accurate at guessing their meaning and will assign higher iconicity ratings (e.g., the gesture and the NGT sign TO-SMOKE represent a person smoking a cigarette so participants are likely to be very accurate and give high iconicity ratings). The expected results will lend credence to the hypothesis that sign-naïve adults base their responses not only on their emblems (Grosso, 1993), but also on other types of (iconic) gestures that are systematic within a community.

## Methodology

## Study 1: Pantomime generation task Participants

Twenty native speakers of Dutch ( 10 females, age range: 21-46, mean: 27 years) living in the area of Nijmegen, the Netherlands, took part in the study.

## Procedure

Participants were seated in front of a computer and were asked to produce a gesture that conveyed exactly the same meaning as the word on the screen. They were explicitly told that it was not allowed to speak or to point to any object in the room and that they could say 'pass' if they were unable to generate a pantomime. Two cameras were positioned on each side of the participant to record their gestural productions. Trials started with a fixation cross $(500 \mathrm{~ms})$ followed by the target word ( 4000 ms ) time during which they had to produce their gesture. After the 4000 ms ended, the next trial began. The motivation behind this strict timing was to elicit participants' most intuitive response. Participants were presented a total of 273 words.
Pantomimes were coded according to the description parameters proposed by Bressem (2013), which are based on the phonological parameters handshape, location, and movement of sign languages. Based on these features, we looked at the gestures that exhibited the same structure across a large number of participants. If the same gesture was produced by at least 12 out of 20 participants, it was considered the default gesture for that concept. These resulted in a total of 119 pantomimes that were consistent across a large proportion of the pool of participants (mean number of participants producing the same pantomime: 15.14).

These default gestures were compared to their NGT translation on each phonological parameter (i.e., handshape, location and movement) to select items with different degrees of form overlap. This comparison resulted in three categories of signs. 1) Full overlap ( $\mathrm{N}=36$ ): gesture-sign pairs did not differ in any parameter (Figure 1A). 2) Partial overlap ( $\mathrm{N}=56$ ): this category includes signs in which only
one parameter differed from the gesture (Figure 1B). 3) No overlap ( $\mathrm{N}=54$ ): signs in which two or more parameters differed. This category consisted of 27 signs that did not overlap at all with the elicited gesture, plus an additional 27 signs for which no default gesture could be established (Figure 1C). These three groups of NGT signs ( $\mathrm{N}=146$ ) were the stimuli for Study 2.

## Study 2: Open-cloze and iconicity rating Participants

The participants of this study were a different group of 20 hearing native speakers of Dutch ( 14 female, mean age $=$ 21.8 years) with no knowledge of NGT or any other sign language. None of them took part in the pantomime generation task.

## Stimuli

The stimuli consisted of videos of the 146 NGT selected from Study 1 (i.e., signs with full, partial, and no overlap with gesture). Videos were produced by a deaf signer with neutral face and without mouthings to avoid giving away cues about the meanings of the signs.

$$
\text { GESTURE } \quad \text { SIGN }
$$



Figure 1: Examples of sign-gesture pairs with different degrees of overlap. A) TO-CUT shares all the components (handshape, location, movement) between sign and gesture. B) TO-SAW differs in only one parameter (handshape). C) In LAPTOP, sign and gesture have no overlap.

## Procedure

At the beginning of each trial, an NGT sign in citation form was presented. After the video had played in full and disappeared from the screen, a new screen was presented instructing participants to guess the meaning of the sign and write its meaning in one word (typed). Participants were required to type in an answer for every item but they were also allowed to skip items if they could not come up with a
meaning. After participants had entered an answer, a new screen of instructions came up. Here participants were given the actual meaning of the sign and were asked to judge how well the sign represented its meaning. The sentence read: 'The meaning of the sign is [translation equivalent]. How much does the sign look like its meaning?' The screen displayed a 7-point Likert and participants were required to type in their rating ( 1 representing the lowest similarity and 7 the highest).

## Analysis

Participants gave a response for a large proportion of the signs with passes representing only $6.5 \%$ of responses. Despite being instructed to write only one word, many responses were phrases, but they were still included in the analysis. Based on the Dutch version of the Boston Naming Task (Roomer, Hoogerwerf, \& Linn, 2011), answers were coded as correct and incorrect. Answers were coded as correct if they matched exactly the expected answer (e.g., sign: TO-PULL; response: to pull) or if they were synonyms of each other (e.g., sign: TO-PHONE; response: to ring). This category also included answers that were not the same part of speech as the target sign, but where the answer was specific to the target concept (e.g., sign: TO-PHONE; response: telephone) ${ }^{1}$. We also included phrases containing a verb and the correct argument depicted in the sign (e.g., sign: BANANA; response: to peel a banana). Responses that did not fit into any of these categories were classed as incorrect answers.

Incorrect answers were subdivided into responses that were semantically related and unrelated to the sign. Semantically related answers included responses that belonged to the same semantic domain (e.g., sign: DUCK; response: penguin); as well as answers that were lacking the appropriate abstraction to the target concept (e.g., the sign MONKEY, which re-enacts how a primate scratches the sides of its torso, was often labelled as scratching).
The semantically unrelated category included responses that were plainly wrong, or answers derived from visual information of the sign, but that had no relationship with the concept (e.g., the sign MOUNTAIN describes the outline of two horizontal bumps, but it was often interpreted as a camel).

For the open cloze, the proportions of correct, semantically related, and semantically unrelated answers were calculated for every item, thereby collapsing across participants' answers. Missing answers were discarded for this analysis and did not contribute to the proportions. For the iconicity ratings, all values were averaged across participants to obtain the mean ratings for each of the 146 signs.

[^162]
## Results

Performance on the open cloze was highly variable across participants and items. While only nine items ( $6.2 \%$ ) were correctly identified by all participants, half of the signs (73 signs) were correctly identified by at least $25 \%$ of participants. For 26 items (17.8\%), all answers were semantically related to the target meaning, suggesting that participants were able to correctly identify some aspect of the sign but did not make the full abstraction to the target meaning (e.g., sign: TO-FLY; response: bird). Regarding the iconicity ratings, participants were able to give a response for all items. In order to establish to what extent sign-gesture overlap contributes to guessing the meaning of a sign and assign iconicity ratings, we considered the following variables in the statistical analysis.
Independent variable: Degree of overlap (full, partial, and no overlap)

## Dependent variables:

i. Proportion of correct answers (open cloze)
ii. Proportion of semantically related answers (open cloze)
iii. Proportion of semantically unrelated answers (open cloze)
iv. Mean iconicity rating

A multivariate ANOVA was run to determine the relationship between type of gestural overlap (full, partial and, no overlap) and the dependent variables of the open cloze and the iconicity ratings. Using Pillai's Trace we found a significant overall effect of the degree of overlap, $V$ $=0.541, F(6,230)=14.205, \eta^{2}=.27, p<.001$. The following sections will describe the between-subjects effects for each dependent variable.
i) Turning to the proportion of correct answers in the open cloze, tests of between-subjects revealed a significant effect of degree of overlap, $F(2,116)=24.168, \eta^{2}=.194, p<$ .001. Planned contrasts revealed an increase of correct answers from no overlap items ( $M=0.12, S E=.03$ ) to partial overlap $\left(M=0.46, S E=.05, \Delta=-0.31, S E_{\Delta}=.06, p\right.$ $<.001, \mathrm{BCa} 95 \%$ CI $[-0.45,-0.18]$ ), but no significant difference between partial and full overlap ( $M=0.61, S E=$ $.06, p=.209$ ). The proportion of correctly identified items was thus higher for items with full and partial overlap than for those with no overlap (Figure 2).
ii) Regarding the proportion of incorrect answers that were semantically related to the sign, a test of betweensubjects effects revealed no significant effect of the degree of overlap between gestures and signs, $p=.305$. That is, wrong answers in the open cloze were equally distributed across the three types of signs (full, partial, and no overlap).
iii) Turning to the proportion of incorrect answers that were semantically unrelated to the target concept, tests of between-subjects effects revealed a significant effect of the degree of overlap, $F(2,116)=26.909, \eta^{2}=.317, p<.001$. Signs with no overlap were significantly less likely to be guessed correctly ( $M=0.75, S E=.05$ ) than those with partial overlap $(M=0.41, S E=.05, \Delta=0.34$, BCa $95 \%$ CI [0.21, 0.47], $p<.001$ ). Signs with full overlap were
significantly more likely to be guessed accurately than signs with partial overlap ( $M=0.21, S E=.04, \Delta=0.192, \mathrm{BCa}$ $95 \%$ CI $[0.05,0.33], p=.009$ ). In other words, the less similar a sign is from a gesture, the more likely it is to be guessed inaccurately.
iv) When we look at iconicity ratings, we found an association with the degree of overlap between sign and gesture $F(2,111.836)=54.13, \eta^{2}=.483, p<.001$. Planned contrasts revealed a significant increase of mean iconicity ratings from no overlap ( $M=3.18, S E=0.22$ ) to partial overlap $(M=5.34, S E=.17, \Delta=-2.13$, BCa $95 \%$ CI [2.617, -1.642$], p<.001$ ) but not from partial to full overlap ( $M=5.92, S E=.15, p=.07$ ). These results suggest that when signs have greater overlap in form with their gestures they perceive signs as more iconic (see Figure 3).


Figure 2: Mean proportion of correctly guessed answers as a function of gesture overlap with the target sign


Figure 3: Mean iconicity ratings as a function of gesture overlap with the target sign

## Discussion

These data expands on previous research by showing that the gestural repertoire of non-signing adults is recruited to make judgments about the meaning of lexical signs. We showed that signs that overlap in form with their gestures are guessed more accurately and are judged as more iconic. The proportion of correct answers and iconicity ratings were higher for signs that overlapped in form with gestures, but there was no additional improvement between full and
partial overlap. This suggests that despite their slight structural differences, these two types of signs bear enough resemblance to participants' gestures to make an association between them.

Signs and gestures share the same physical constraints to express a concept in the manual modality, with the referent shaping to some extent the features than can be expressed with the hands (Masson-Carro et al., 2015). It is therefore not surprising that signs and gestures converge in the strategies to depict the visual characteristics of many concepts. If signs and gestures have similar structures for some concepts, it means that deaf and hearing adults share conceptual knowledge about these concepts (i.e., visual, semantic, perceptual, sensorimotor representations). When there is sufficient overlap between signs and gesture, nonsigning adults may tap into these schemas to make judgements about the meaning of signs. These findings also relate to research showing that humans - as well as other primates - understand and evaluate the correctness of others' actions through the activation of brain regions engaged when they perform the same actions themselves (Koelewijn, van Schie, Bekkering, Oostenveld, \& Jensen, 2008; Rizzolatti, Fadiga, Gallese, \& Fogassi, 1996).

The errors produced by participants, however, clearly show that if gesture and sign mismatch, or if the meaning of signs departs slightly from the features they depict, participants are unable to estimate accurately the meaning of a sign. As a result, they will also rate the sign as less iconic. Non-signers have a very limited scope to assign meanings to signs and seem to be inclined to describe only what is directly encoded in them. While they are capable of extracting some visual information from the signs they often fail to respond with the correct metonymic associate (e.g., they respond scratch instead of monkey). This goes to show that despite their similarities, sign languages have established linguistic conventions not shared with gestures and thus are inaccessible to non-signing adults.

This study adds to the body of research investigating how modality shapes linguistic/communicative structures (Perniss et al., 2015).

## Acknowledgments

This work was supported by a Veni grant by the Netherlands Organisation for Scientific Research (NWO) awarded to the first author. We would like to thank the members of the Multimodal Language and Cognition lab at Radboud University for their helpful comments, in particular to Swen Schreiter, Linda Drijvers, and James Trujillo for their support.

## References

Cormier, K., Schembri, A., \& Woll, B. (2013). Pronouns and pointing in sign languages. Lingua, 137, 230-247. http://doi.org/10.1016/j.lingua.2013.09.010
Cuxac, C. (1999). French sign language: proposition of a structural explanation by iconicity. Gesture-Based Communication in Human-Computer, 1739(1), 165184.

Goldin-Meadow, S., \& Brentari, D. (2015). Gesture, sign
and language: The coming of age of sign language and gesture studies. Behavioral and Brain Sciences, 8(4),

1-82. http://doi.org/10.1017/S0140525X15001247.
Griffith, P. L., Robinson, J. H., \& Panagos, J. M. (1981). Perception of iconicity in American sign language by hearing and deaf subjects. The Journal of Speech and Hearing Disorders, 46(4), 388-97. http://doi.org/dx.doi.org/10.1044/jshd. 4604.388
Grosso, B. (1993). Iconicity and Arbitrariness in Italian Sign Language: An Experimental Study. University of Padua, Italy.
Gunter, T. C., \& Bach, P. (2004). Communicating hands: ERPs elicited by meaningful symbolic hand postures. Neuroscience Letters, 372(1), 52-6.
Kelly, S. D., Creigh, P., \& Bartolotti, J. (2010). Integrating speech and iconic gestures in a Stroop-like task: evidence for automatic processing. Journal of Cognitive Neuroscience, 22(4), 683-694. http://doi.org/10.1162/jocn.2009.21254
Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. Journal of Pragmatics, 23(3), 247-279. http://doi.org/dx.doi.org/10.1016/0378-2166(94)00037-f
Kendon, A. (2004). Gesture: Visible action as utterance. Cambridge: Cambridge University Press.
Klima, E., \& Bellugi, U. (1979). The Signs of Language. Harvard: Harvard University Press.
Koelewijn, T., van Schie, H. T., Bekkering, H., Oostenveld, R., \& Jensen, O. (2008). Motor-cortical beta oscillations are modulated by correctness of observed action. NeuroImage, 40(2), 767-775. http://doi.org/10.1016/j.neuroimage.2007.12.018
Masson-Carro, I., Goudbeek, M., \& Krahmer, E. (2015). Can you handle this? The impact of object affordances on how co-speech gestures are produced. Language, Cognition and Neuroscience, 1-11. http://doi.org/10.1080/23273798.2015.1108448
McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.
Müller, C. (2013). Gestural modes of representation as techniques of depcition. In C. Müller, A. Cienki, S. Ladewig, D. McNeill, \& J. Bressem (Eds.), Body Language - Communication: An International Handbook on Multimodality in Human Interaction (pp. 1687-1701). Berlin: De Gruyter Mouton.
Ortega, G., \& Ozyürek, A. (2016). Generalisable patterns of gesture distinguish semantic categories in communication without language. In A. Papafragou, D. Grodner, J. Mirman, \& J. Trueswell (Eds.), Proceedings of the 38th Annual Meeting of the Cognitive Science Society (pp. 1182-1187). Austin, TX: Cognitive Science Society, Inc.
Ortega, G., \& Özyürek, A. (2016). Generalisable patterns of gesture distinguish semantic categories in communication without language. In A. Papafragou, D. Grodner, D. Mirman, \& J. Trueswell (Eds.), Proceedings of the 38th Annual Meeting of the

Cognitive Science Society (CogSci 2016) (pp. 11821187). Austin, TX.: Cognitive Science Society, Inc.

Padden, C., Hwang, S.-O., Lepic, R., \& Seegers, S. (2015). Tools for Language: Patterned Iconicity in Sign Language Nouns and Verbs. Topics in Cognitive Science, 7(1), 81-94. http://doi.org/10.1111/tops. 12121
Padden, C., Meir, I., Hwang, S.-O., Lepic, R., Seegers, S., \& Sampson, T. (2013). Patterned iconicity in sign language lexicons. Gesture, 13(3), 287-305.
Perniss, P., Özyürek, A., \& Morgan, G. (2015). The influence of the visual modality on language structure and conventionalization: Insights from sign language and gesture. Topics in Cognitive Science, 7(Special Issue), 2-11.
Perniss, P., Thompson, R. L., \& Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. Frontiers in Psychology, 1(227), 1664-1678. http://doi.org/dx.doi.org/10.3389/fpsyg.2010.00227
Pietrandrea, P. (2002). Iconicity and arbitrariness in Italian Sign Language. Sign Language Studies, 2(3), 296321. http://doi.org/dx.doi.org/10.1353/sls.2002.0012

Pizzuto, E., \& Volterra, V. (2000). Iconicity and transparency in Sign Languages: A cross-linguistic cross-cultural view. In K. Emmorey \& H. L. Lane (Eds.), The signs of language revisited: An anthology to Honor Ursula Bellugi and Edward Klima (pp. 229250). Mahwah, N. J.: Lawrence Erlbaum Associates.

Quinto-Pozos, D., \& Parrill, F. (2015). Signers and Cospeech Gesturers Adopt Similar Strategies for Portraying Viewpoint in Narratives. Topics in Cognitive Science, 7(Special Issue), 1-23. http://doi.org/10.1111/tops. 12120
Rizzolatti, G., Fadiga, L., Gallese, V., \& Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3(2), 131-141. http://doi.org/10.1016/0926-6410(95)00038-0
Roomer, E. K., Hoogerwerf, A. C., \& Linn, D. E. (2011). Boston benoem taak 2011. Utrecht.
Safar, J., \& Petatillo, R. (n.d.). Strategies of noun-verb distinction in Yucatec Maya Sign Languages. In O. Le Guen, J. Safar, \& M. Coppola (Eds.), Emerging Sign Languages of the Americas. Sign Language Typology Series. Berlin: Mouton de Gruyter.
Sandler, W., \& Lillo-Martin, D. (2006). Sign Language and Linguistic Universals. Cambridge: Cambridge University Press.
van Nispen, K., van de Sandt-Koenderman, M., Mol, L., \& Krahmer, E. (2014). Pantomime Strategies: On Regularities in How People Translate Mental Representations into the Gesture Modality. In Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014) (pp. 30203026). Austin, TX: Cognitive Science Society, Inc.

# A Formal Approach to Modeling the Cost of Cognitive Control 

Kayhan Özcimder ${ }^{1,2, *}$, Biswadip Dey ${ }^{2, *}$, Sebastian Musslick ${ }^{1, *}$, Giovanni Petri ${ }^{3}$, Nesreen K. Ahmed ${ }^{4}$, Theodore L. Willke ${ }^{4}$, Jonathan D. Cohen ${ }^{1}$<br>${ }^{1}$ Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA.<br>${ }^{2}$ Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA.<br>${ }^{3}$ ISI Foundation, Via Alassio 11/c, 10126 Torino, Italy.<br>${ }^{4}$ Intel Labs, Santa Clara, CA 95054, USA.<br>*Equal Contribution, Corresponding Author: ozcimder@princeton.edu


#### Abstract

This paper introduces a formal method to model the level of demand on control when executing cognitive processes. The cost of cognitive control is parsed into an intensity cost which encapsulates how much additional input information is required so as to get the specified response, and an interaction cost which encapsulates the level of interference between individual processes in a network. We develop a formal relationship between the probability of successful execution of desired processes and the control signals (additive control biases). This relationship is also used to specify optimal control policies to achieve a desired probability of activation for processes. We observe that there are boundary cases when finding such control policies which leads us to introduce the interaction cost. We show that the interaction cost is influenced by the relative strengths of individual processes, as well as the directionality of the underlying competition between processes.


Keywords: cognitive control; multi-tasking; intensity; identity

## Introduction

A long standing focus in cognitive research has been towards understanding the ability to execute tasks/processes ${ }^{1}$ that demand cognitive control. In this context, cognitive control is defined as the set of mechanisms required to pursue a goal, especially when distraction or strong competing responses (interferences) must be overcome (Posner \& Snyder, 1975; Shiffrin \& Schneider, 1977; Cohen, Dunbar, \& McClelland, 1990). Earlier work (Posner \& Snyder, 1975; Shiffrin \& Schneider, 1977; Cohen et al., 1990; Botvinick \& Cohen, 2014) has argued that the processes demanding control can be distinguished from automatic processes in terms of the strength of the associations in the pathways underlying processing: automatic processes are characterized by pathways with associations strong enough to resist interference from competing processes, whereas controlled processes are weaker, and therefore rely on input from control mechanisms to support their execution against interference.

Another longstanding observation is that the allocation of cognitive control is costly (often discussed in terms of "mental effort" (Posner \& Snyder, 1975; Botvinick \& Braver, 2015; Shenhav et al., 2017)). This cost has been interpreted in physical terms (such as metabolic demands (Muraven, Tice,

[^163]\& Baumeister, 1998)) or in terms of an opportunity cost reflecting the allocation of a limited resource (Kurzban, Duckworth, Kable, \& Myers, 2013). Elsewhere (Feng, Schwemmer, Gershman, \& Cohen, 2014; Musslick et al., 2016), we have proposed that limitations in the capacity for controldependent processing reflect the purpose of control to diminish interference rather than any intrinsic limitation in the mechanism responsible for control. This view suggests that the architecture of the processing system as a whole constrains the opportunities for control-dependent processing, resulting in opportunity costs associated with allocating control to any particular task(s).

Here, we build on a closely related proposal, by Koechlin and Summerfield (2007), to define the cost of control in terms of internal representational requirements to insure that a given stimulus (or a set of stimuli) produces the desired response (or a set of responses), given the intrinsic architecture of the system. Their work focused on a single task. Here, we extend this to consider an arbitrary number of tasks and thus accommodate their possibility for, and costs of, multitasking (i.e.parallel processing of task pathways). To do so, we follow the framework proposed by Shenhav, Botvinick, and Cohen (2013) that distinguishes two components of control signals: intensity and identity. Specifically, Shenhav et al. (2013) defined the intensity of a control signal as the strength of the signal needed to insure performance of a particular task, and the identity as which control signal should be selected to achieve a desired objective given environmental conditions. Here we build on that distinction to define two corresponding components of the cognitive control costs - a cost associated with intensity, and a cost associated with interaction. Furthermore, we define the interaction cost to capture the level of interference between the processes in a network.

In this paper, we begin by introducing formal constructs for intensity and interaction costs by using the graph theoretic representation of a neural network and terms/notions adopted from probability theory. We describe an intensity cost that represents the control signals (as biases infused into a neural network), above and beyond the specified strength of the signal (stimulus) itself. This is achieved by developing a formal relationship between the probability of successful execution of desired processes and the control signals. In turn, this defines an optimization problem, which can then be solved to find optimal control signals that achieve a specified activation for desired processes. However, we observe that there


Figure 1: Illustration of a single-layered, feed-forward network with 3-input and 3-output layer components, wherein the individual features are scalar in nature.
are boundary cases in which this optimization problem can not be solved. These cases reflect situations in which the simultaneous execution of the processes is not feasible due to interference. Hence, interaction cost analysis motivates an additional investigation towards finding a proper metric that continuously measures the level of interference between processes. To achieve this we introduce the definition of interaction cost associated with process mappings in a network configuration. Specifically, it measures the level of interference introduced by competing processes that interfere with the tasks of interest. In their study, Koechlin and Summerfield (2007) have already used information theoretic terms to measure cognitive control. However, in order to apply these metrics to neural networks while considering parallelism, we augment some of these measures, and through simulations we will demonstrate how interaction cost can be used to predict interference in neural network architectures. Finally, we will discuss general research directions revealed by the analysis presented here.

## Intensity: the cost of control

Cognitive control is defined as the underlying mechanism that biases the processing of a task in order to maximize the reward (Botvinick, Braver, Barch, Carter, \& Cohen, 2001; Botvinick, Cohen, \& Carter, 2004; Bogacz, Brown, Moehlis, Holmes, \& Cohen, 2006; Botvinick, 2007). Here, we adapt the notion of intensity cost from Shenhav et al. (2013) as a function of the amount of control bias that cognitive control applies to the system. However, Shenhav et al. (2013) described this function in qualitative rather than quantitative terms. In this work we provide an explicit characterization of the cost of cognitive effort in terms of a set of physically meaningful parameters, which allow the manipulation of the response of a cognitive architecture.

Following earlier works (Feng et al., 2014; Musslick et al.,
2016), we consider a single-layered, feed-forward network with $N$ input and $M$ output layer components to formalize the notion of intensity cost in a cognitive control context (Fig. 1 shows a simple example of such a network). In this framework, each component represents an input/stimulus or output/response dimension (vector subspace), and the connection from an input to an output component constitutes the processing pathway for a given task. This allows us to define an abstraction of the network as a directed bipartite graph $\mathcal{G}_{B}=(\mathcal{V}, \mathcal{E})$, wherein the set of vertices $\mathcal{V}$ can be partitioned into two disjoint sets $\mathcal{V}_{\text {in }}$ and $\mathcal{V}_{\text {out }}$, representing the input and output layer components respectively. Moreover, a directed edge $(i, j) \in \mathcal{E} \subseteq \mathcal{V}_{\text {in }} \times \mathcal{V}_{\text {out }}$ represents a connection from the vertex $i$ in the input layer to vertex $j$ in the output layer (i.e., a task). In this setting, we represent the processing pathway by introducing a weight matrix $W$ with elements $w_{i j}$. As we will see later, this abstraction plays an important role in formalizing the interaction cost of cognitive control.

In this setting, we assume that control signals bias the processing of a stimulus towards a specified response at two different levels, i.e. $g_{i}$ and $b_{j}$, which we refer to as preinteraction and post-interaction control biases, respectively. This complies with early computational models of cognitive control in which control signals act as an increase in gain of non-linear processing units (Cohen et al., 1990; Botvinick et al., 2001) and allows us to treat such control biases as key contributing factors towards the intensity cost for cognitive control. It is worth noting here that for simplicity the only sources of nonlinearity in this setting are the logistic ${ }^{2}$ activation functions which act upon the linearized output vector $\tilde{\mathbf{y}}=\left[\tilde{Y}_{1}, \tilde{Y}_{2}, \ldots, \tilde{Y}_{M}\right]$. Without loss of generality, in what follows we consider the individual features to be scalar, and carry out a formal investigation on how these control biases $\mathbf{g}=\left[g_{1}, \ldots, g_{N}\right]$ and $\mathbf{b}=\left[b_{1}, \ldots, b_{M}\right]$ influence the response from this cognitive architecture. In our formulation, the corresponding magnitude, i.e. $\|\mathbf{g}\|^{2}+\|\mathbf{b}\|^{2}$, can be treated as a measure of control intensity applied to the system, and therefore the cost for cognitive control.

We begin our analysis by assuming the vector of features $\mathbf{s}=\left[S_{1}, S_{2}, \ldots, S_{N}\right]$ to be an $N$-dimensional multivariate Gaussian random variable with mean $\mu^{S}$ and covariance $\Sigma^{S}$. (The assumption of Gaussianity is motivated by the technical tractability). With this assumption, the vector $\left[X_{1}, X_{2}, \ldots, X_{N}\right]$ becomes an N -dimensional multivariate Gaussian random variable with a shifted mean and same covariance. Furthermore, as all the transformations (before the nonlinear logistic activation function) are linear in nature, $\left[Y_{1}, Y_{2}, \ldots, Y_{N}\right]$ also remains a multivariate Gaussian random variable whose mean and covariance are given by $\mu^{Y}=W\left(\mu^{S}+\mathbf{g}\right)$ and $\Sigma^{Y}=$ $W \Sigma^{S} W^{T}$, respectively. Similarly, the vector of linearized outputs $\left[\tilde{Y}_{1}, \tilde{Y}_{2}, \ldots, \tilde{Y}_{N}\right]$ is also a multivariate Gaussian, with a shifted mean and the same covariance.

[^164]Hence, each individual linearized output $\tilde{Y}_{i}$ is itself a Gaussian random variable with

$$
\begin{aligned}
\text { mean: } & \mu_{i}^{\tilde{Y}}=\sum_{j=1}^{N} w_{j i}\left(\mu_{j}+g_{j}\right)+b_{i} \\
\text { and variance: } & \sigma_{i i}^{\tilde{Y}}=\sum_{j=1}^{N} \sum_{k=1}^{N} w_{j i} w_{k i} \sigma_{k j},
\end{aligned}
$$

where $\mu_{j}$ is the mean of stimulus $S_{k}$ and $\sigma_{k j}$ is the covariance between stimuli $S_{k}$ and $S_{j}$. As a consequence, the corresponding output (response) will have a logit-normal distribution, and this leads us to our key result in this section.

As outlined by Shenhav et al. (2013), the response $O_{i}$ should overcome a specified threshold in order to execute the corresponding process (task). Then, by letting $\alpha_{i} \in(0,1)$ represent this activation threshold associated with output $O_{i}$, the corresponding probability of task execution (probability of the output $O_{i}$ surpassing the threshold $\alpha_{i}$ ) is expressed as

$$
P\left[O_{i} \geq \alpha_{i}\right]=\frac{1}{2}-\frac{1}{2} \operatorname{erf} \underbrace{\left(\frac{\log \left(\frac{\alpha_{i}}{1-\alpha_{i}}\right)-b_{i}-\sum_{j=1}^{N} w_{j i}\left(\mu_{j}+g_{j}\right)}{\sqrt{2 \sum_{j=1}^{N} \sum_{k=1}^{N} w_{j i} w_{k i} \sigma_{k j}}}\right)}_{f\left(\alpha_{i}, b_{i}, w, \mu, \mathbf{g}, \Sigma^{S}\right)} .
$$

Here we have exploited the monotonicity of the logistic function to compute its inverse. Then the result follows from the cumulative distribution function of $\tilde{Y}_{i}$.

We characterized the activation probability of a given network in terms of the pre-interaction and post-interaction control biases. This is crucial because it provides new directions to incorporate the cost of control into the design of a cognitive network architecture. For example, the problem of allocating a limited amount of cognitive control into different components of the network to maximize the associated probability of activation can be formulated as an optimization problem in which the goal becomes minimizing $f\left(\alpha_{i}, b_{i}, w, \mu, \mathbf{g}, \Sigma^{S}\right)$ over $\mathbf{g}$ and $\mathbf{b}$ subject to the constraints $\sum_{i=1}^{N} g_{i}^{2} \leq C_{g}$ and $\sum_{i=1}^{M} b_{i}^{2} \leq C_{b}$, where $C_{g}$ and $C_{b}$ define the maximum amount of control that can be applied. Alternatively, in this setting, we can also approach the problem of minimizing the cost of control, while still maintaining a desired value of probability of activation.

One can consider the joint distribution of the processes of interest to incorporate the effects of interaction between tasks. To be consistent with (Feng et al., 2014; Musslick et al., 2016), it is reasonable to begin with a focus narrowed to the situation where the choice of interaction weights and the prior distribution of the stimuli render the interactions undesirable. Then the effect of multitasking can measured by introducing a suitable distance metric (for example, the Kullback-Leibler divergence (Ortega \& Braun, 2013)) between the joint distributions of relevant processes and the product of corresponding marginals, and one can attempt to minimize this distance
at the expense of a limited amount of cognitive control. However, as one might expect, this optimization problem can have an empty solution set under certain values of the interaction weights and activation thresholds, meaning that certain network configurations strictly prohibit successful multitasking performance (Musslick et al., 2016). Before approaching this computation in detail, it would be beneficial to investigate how the interaction structure influences the solution space, and that leads us to our next section wherein we introduce the notion of interaction cost.

## Interaction: the cost of mapping

In this section, we will introduce a detailed formalism of the interaction cost associated with process mappings in a network configuration to accommodate the possibility for multitasking. In our earlier work (Musslick et al., 2016), we have formalized three distinct types of interference (Fig. 2).

Convergent interference (Fig. 2a) occurs when two inputs/stimuli (e.g. $S_{1}$ and $S_{2}$ ) compete to determine a common output (e.g. $O_{1}$ ). We also consider divergent interference in our analysis (Fig. 2b). Although this does not pose an impediment to performance, i.e. it is possible to generate two distinct outputs (e.g. $O_{1}$ and $O_{2}$ ) to the same input (e.g. $S_{1}$ ), it represents a restriction on the number of independent stimuli (and therefore the number of tasks) that the system can process at once, and thus was treated formally as a type of interference due to this dependency in our analysis of parallel processing capability. Finally, we consider a third, indirect interference that supervenes on the first two (shown in Fig. 2c and Fig. 2d). In this case, the two tasks with strengths $w_{11}$ and $w_{22}$ in question do not directly interfere with one another. However, their simultaneous execution would necessarily engage a third task with strength $w_{21}$ (also possibly a fourth task with strength $w_{12}$ ) that would produce interference in output $O_{1}$ (and $O_{2}$ ). While Musslick et al. (2016) treated these three types of interference identically in terms of their effect on the overall parallel processing capability of a network, the proposed interaction cost will also distinguish between these three types of interference.

In interaction cost analysis, we will assume that a stimulus is of value 1 when it is active, and 0 otherwise. Moreover, to increase tractability, we will consider linear activation at the output level, which also implies that without loss of generality the pre- and post-interaction biases can be assumed to be zero. A more detailed version of the interaction cost analysis, involving the strength of stimuli, as well as the nonlinear activation function, will be discussed in subsequent publications.

To introduce the interaction cost, we take an approach similar to the one adopted by Koechlin and Summerfield (2007). In their work, Koechlin and Summerfield (2007) proposed a metric for selecting a single action among multiple alternatives. Here, we will refine this metric to introduce the interaction cost for neural network architectures. Towards this objective, we first leverage the assumptions discussed earlier in


Figure 2: The illustration of convergent, divergent, asymmetric, and symmetric interference.
the section, and abstract out the network configurations presented in Fig. 2 from the network shown in Fig. 1. We also assume that the strength of a task $i j$ from stimulus $S_{i}$ to output $O_{j}$ is represented by its non-negative weight $w_{i j} \geq 0$.

Let us first consider the case shown in Fig. 2a. It is obvious that the response in the output component $O_{j}$ is completely determined by the stimulus if either $S_{1}$ or $S_{2}$ is activated in the network (executing a single process). However, activating both stimuli $S_{1}$ and $S_{2}$ simultaneously creates a conflict, since the output can not respond to two distinct stimuli simultaneously (as the activations are linear, the network will always have a response in the output level). In order to measure the level of this competition between stimuli, we define a random variable $a_{1}$ associated with the output $O_{1}$ such that $a_{1} \in\{1,2\}$ (Fig. 2a). This implies that $a_{1}=1$ or $a_{1}=2$ when the output $O_{1}$ is driven completely by $S_{1}$ or $S_{2}$, respectively. Since a stronger task will have a higher probability of being selected to generate the response, we consider the relative strengths of the task pathways (with associated strengths $w_{11}$ and $w_{21}$ ) in order to define the probability of the possible outcomes, i.e. the probability of $a_{1}=1$ and $a_{1}=2$ when both $S_{1}$ and $S_{2}$ activated. Hence, we compute the probability as

$$
P\left[a_{1}=1\right]=\frac{w_{11}}{w_{11}+w_{21}}, \quad \text { and, } \quad P\left[a_{1}=2\right]=\frac{w_{21}}{w_{11}+w_{21}} .
$$

Next, we extend this framework to consider networks with $N$ stimuli and $M$ outputs, wherein an output $O_{j}, j \in$ $\{1, \ldots, M\}$ responds to a set of stimuli. Let us assume that there are $n \leq N$ incoming edges to a particular output $O_{j}$, and each edge is originated from a distinct stimulus $S_{i}, i=$ $i_{1}, \ldots, i_{n}$, where $i_{k} \in\{1, \ldots, N\}$. Then the probability of the event that output $O_{j}$ is responding to stimulus $S_{i}$ is given by,

$$
\begin{equation*}
P\left[a_{j}=i\right]=\frac{w_{i j} \mathbb{1}\left(S_{i}\right)}{\sum_{k=1}^{n} w_{i_{k} j} \mathbb{1}\left(S_{i_{k}}\right)}, \tag{1}
\end{equation*}
$$

where $\mathbb{1}\left(S_{i}\right)$ is the indicator function that represents the activation of stimulus $S_{i}$ such that $\mathbb{1}\left(S_{i}\right)=1$ if stimulus $S_{i}$ is active and $\mathbb{1}\left(S_{i}\right)=0$, otherwise. Then, by building upon the ideas proposed by (Koechlin \& Summerfield, 2007), we define the interaction cost as

$$
\begin{equation*}
\Psi\left(a_{j}=i\right)=-\log \left(P\left[a_{j}=i\right]\right) \tag{2}
\end{equation*}
$$

where the logarithm is with respect to base 2.

Equation 1 implies the $P\left[a_{j}=i\right]=1$ when only the relevant stimulus $S_{i}$ associated with task $i j$ is activated in the network. Hence the interaction cost is computed as $\Psi\left(a_{j}=\right.$ $i)=0$ which implies that there is no interaction cost. Moreover, when multiple processes are competing due to the activation of multiple stimuli, $P\left[a_{j}=i\right] \rightarrow 0$ as the competition increases, and as a consequence the interaction cost $\Psi\left(a_{j}=i\right) \rightarrow \infty$.

We further extend equation 1, to encapsulate the probability associated with parallel processing of task pathways in the network. Thus, we introduce the joint probability of distinct output components responding to a set of stimuli. For instance, let us consider the parallel processing of tasks with strength $w_{11}$ and $w_{21}$ in Fig. 2a, and calculate $P\left[a_{1}=1, a_{1}=2\right]$. This is the probability of output component $O_{1}$ responding to both $S_{1}$ and $S_{2}$, and by definition we know that this probability is zero. For the case illustrated in Fig. 2b, the joint probability $P\left[a_{1}=1, a_{2}=1\right]=1$ since activation of $S_{1}$ will activate both processes with strengths $w_{11}$ and $w_{12}$, and there is no competition in outputs $O_{1}$ and $O_{2}$. This result is parallel to the observation made by (Musslick et al., 2016), who stated that divergent interference is not actually an interference but a dependency on the stimuli.

Now let us consider the case introduced in Fig. 2c which can be thought of as the composition of the two cases presented in Fig. 2a-b. We compute the interaction cost of parallel processing the tasks with strengths $w_{11}$ and $w_{22}$. This requires simultaneous activation of $S_{1}$ and $S_{2}$, which indirectly engages the task with strength $w_{21}$, and initiates a competition in the output $O_{1}$. Thus, the interaction cost of parallelism between tasks represented by $w_{11}$ and $w_{22}$ is given by

$$
\begin{aligned}
\Psi_{1}\left(a_{1}=1, a_{2}=2\right) & =-\log \left(P\left[a_{2}=2\right] \cdot P\left[a_{1}=1 \mid a_{2}=2\right]\right) \\
& =-\log \left(1 \cdot \frac{w_{11}}{w_{11}+w_{21}}\right)
\end{aligned}
$$

Here $P\left[a_{2}=2\right]=1$ since task with weight $w_{22}$ is not competing with any other process in the output $O_{2}$. The competition, however, takes place in $O_{1}$, and the interaction cost associated with $w_{11}$ for this case has already been computed when we discussed the case in Fig. 2a.

In a similar way, we can compute the interaction cost of parallelism between tasks represented by $w_{11}$ and $w_{22}$ in Fig. 2d, and we have

$$
\begin{aligned}
\Psi_{2}\left(a_{1}=1, a_{2}=2\right) & =-\log \left(P\left[a_{2}=2\right] \cdot P\left[a_{1}=1 \mid a_{2}=2\right]\right) \\
& =-\log \left(\frac{w_{11}}{w_{11}+w_{21}} \cdot \frac{w_{22}}{w_{22}+w_{12}}\right)
\end{aligned}
$$

In this case, simultaneous activation of $S_{1}$ and $S_{2}$ causes competition in both outputs $O_{1}$ and $O_{2}$. Thus, by revealing further insight about the strength and directionality of interference, the interaction cost serves as an extension of the interference definition presented by (Musslick et al., 2016). For instance, for the same values of $w_{11}, w_{21}, w_{22} \geq 0$ in both configurations in Fig. 2c-d, and given $w_{12} \geq 0$ for the configuration in

Fig. 2d, it is obvious that

$$
\Psi_{1}\left(a_{1}=1, a_{2}=2\right) \leq \Psi_{2}\left(a_{1}=1, a_{2}=2\right)
$$

## Neural Network Simulation

In order to investigate the effect of directionality for the third indirect interference during parallel processing, we implemented a synthetic neural network simulation ${ }^{3}$ identical to our earlier work (Musslick et al., 2016). The neural network used for this simulation maps stimulus input encoded at a stimulus layer via a non-linear associative layer to nonlinear response layer. A separate task input layer encodes the current task to be performed with respect to that stimulus and projects to both the associative layer and response layer. Units in the stimulus layer were grouped into six stimulus dimensions with three units per dimension. Similarly, units in the response layer was grouped into six response dimensions with three units per dimension. The network was trained on 12 tasks, where each task corresponds to a one-to-one mapping between a subset of three input features in a stimulus layer to a subset of three response units in an output layer.

We then used the methods described in Musslick et al. (2016) to extract a bipartite task graph from single representations encoded at the associative layer. The representations associated with each task can be characterized by calculating, for each unit in the associative and output layers, the mean of its activity over all of the stimuli for a given task; this mean pattern of activity can then be used as a representation of the task. Correlating these patterns of activity across tasks yields a task similarity matrix that can be examined separately for the associative and output layers of the network. This can then be used to assess the extent to which different tasks rely on similar or different representation within each layer of the network. Tasks that have similar representations over the associative layer can be inferred to rely on the same input dimension that is, they share an input component in the bipartite graph representation of the network and tasks that are similar at the output layer can be inferred to share an output component. Accordingly, a bipartite graph can be constructed by measuring the patterns of activity observed in the network while it performs each individual task.

The extracted bipartite graph can be used to extract interference patterns between pairs of tasks (cf. Fig. 2). We use this to extract all possible learned task-pairs involving no interference case (Fig. 3a), and two distinct cases of interference as shown in Fig. 3b and Fig. 3c from single-task representations. Fig. 3 shows the activation patterns of the output units for the simultaneous execution of two tasks, averaged across the patterns of all task pairs for a given interference structure. That is, no interference (Fig. 3a) leads to very accurate response patterns (i.e. the current activation shown in orange) is very close to the desired activation pattern shown in grey). For the case in Fig. 3b, the response pattern of task $\left(S_{1}: O_{1}\right)$ is primarily impaired due to the interference aris-

[^165]ing from task $\left(S_{2}: O_{1}\right)$. However, for the case of interference illustrated in Fig. 3c, the response patterns for both tasks $\left(S_{1}: O_{1}\right)$ and ( $S_{2}: O_{2}$ ) are impaired as observed by the activation patterns. These simulation results reflect the influence of the directionality of interference between tasks as predicted by the proposed interaction cost.


Figure 3: This figure illustrates the performance of a task-pair for a given interference pattern. Each tasks maps a subset of three stimulus input units onto three response units (see text). The orange color in the bar plots indicates unit activation of response units relevant to the depicted tasks, while gray indicates desired response pattern of those units.

## General Discussion and Conclusion

In this study, we have introduced two new measures to determine costs associated with intensity and interaction for the demand on control. First, we quantify the intensity cost as a function of the amount of control bias that is supplementary to stimulus-specific processing in order to achieve a desired response from the network. Doing so, we formalize the probability of achieving a desired task given the stimulus, weights and biases infused to the network. Since the stimuli and weights are considered as network properties, the intensity cost to achieve desired response is defined as the amount (value) of control biases required to be injected to the input and output components of the network. The detailed analysis of intensity cost revealed an interesting optimization problem to maximize the probability of surpassing a specified activation for a given budget of resources (i.e. an upper bound on the control biases). The existence of a solution of this optimization problem implicitly reveals whether the desired objective is feasible. However, as it can be foreseen that under certain circumstances the solution does not exist due to interference between the involved processes. Such boundary conditions motivated the second metric introduced in our paper in which we formalize the interaction cost to measure the level of interactions/interference between processes by means of their type of connections and weights.

With the introduced characterization of intensity and interaction costs, it is possible to formally define whether a process is considered a reflex, automatic or controlled. Concretely, a process is considered a reflex if the underlying weight guarantees a successful execution. In other words, a reflex can be successfully executed without any intensity or interaction costs. We assume that the execution of both controlled and automatic processes carries with it an intensity cost as some amount of control bias is needed to elicit a response. However, unlike the former, controlled processes are
subject to interference and thus yield interaction costs large than zero.

The metrics proposed here can also be used towards further understanding cognitive effort as well as synthetic neural networks designed to achieve goal-driven tasks. By using the intensity cost, which reveals the interrelationship between control bias and probability of achieving a desired objective, we will investigate the limitations of any given neural network architecture by allocating a budget of control bias. The intensity cost can also be used to investigate the feasibility of achieving a desired objective defined by the set of processes of interest in a network.

In the interaction cost analysis, we have assumed that there exist a response in the output for any stimulus activation and this may not be the case for a nonlinear activation in the output components. Hence, one major research direction is the detailed analysis for the classification of processes with nonlinear activation in output. Another possible direction for future work is to further analyze the interaction cost in order to capture the properties of the overall network (not only a subset of tasks of interest). This will allow one to use the interaction cost as an objective function for network training. Another possible direction is to explore the interrelationship between intensity and interaction cost. In our work (Musslick et al., 2016), we noticed a fundamental trade-off between shared representations in a network and its parallel processing capability (separated representations). Intuitively, we envision this separation will decrease interaction cost while increasing the likelihood of successful execution for a given budget of control bias.

## Acknowledgements

We would like to thank Zahra Aminzare, Adam Charles, Jonathan Pillow and Vaibhav Srivastava for their help in the formalism of the problem.

## References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., \& Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review, 113(4), 700-765.
Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. Cognitive, Affective, \& Behavioral Neuroscience, 7(4), 356-366.
Botvinick, M. M., \& Braver, T. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. Annual Review of Psychology, 66(1), 83-113.
Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., \& Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychological Review, 108(3), 624-652.
Botvinick, M. M., \& Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. Cognitive Science, 38(6), 1249-1285.

Botvinick, M. M., Cohen, J. D., \& Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. Trends in Cognitive Sciences, 8(12), 539-546.
Cohen, J. D., Dunbar, K., \& McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. Psychological Review, 97(3), 332-361.
Feng, S. F., Schwemmer, M., Gershman, S. J., \& Cohen, J. D. (2014). Multitasking vs. Multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. Cognitive, Affective, \& Behavioral Neuroscience, 14(1), 129-146.
Koechlin, E., \& Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. Trends in Cognitive Sciences, 11(6), 229-235.
Kurzban, R., Duckworth, A., Kable, J. W., \& Myers, J. (2013). An opportunity cost model of subjective effort and task performance. The Behavioral and Brain Sciences, 36(6), 661-679.
Muraven, M., Tice, D. M., \& Baumeister, R. F. (1998). Selfcontrol as a limited resource: Regulatory depletion patterns. Journal of Personality and Social Psychology, 74(3), 774-789.
Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., \& Cohen, J. D. (2016). Controlled vs. Automatic Processing: A Graph-Theoretic Approach to the Analysis of Serial vs. Parallel Processing in Neural Network Architectures. In Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 15471552). Philadelphia, PA.

Ortega, P. A., \& Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. Proceedings of Royal Society A, 469(2153), 20120683.

Posner, M. I., \& Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), Information Processing and Cognition: The Loyola Symposium (p. 55-85). Lawrence Erlbaum.
Shenhav, A., Botvinick, M. M., \& Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron, 79(2), 217-240.
Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., \& Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. (submitted to Annual Reviews of Neuroscience)
Shiffrin, R. M., \& Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review, 84(2), 127-190.

# Optimization of American English, Spanish, and Mandarin Chinese over time for efficient communication 

John K Pate (johnpate@buffalo.edu)<br>Department of Linguistics, 609 Baldy Hall<br>Buffalo, NY 14260 USA


#### Abstract

Frequent words tend to be short, and many researchers have proposed that this relationship reflects a tendency towards efficient communication. Recent work has sought to formalize this observation in the context of information theory, which establishes a limit on communicative efficiency called the channel capacity. In this paper, I first show that the compositional structure of natural language prevents natural language communication from getting close to the channel capacity, but that a different limit, which incorporates probability in context, may be achievable. Next, I present two corpus studies in three typologically-diverse languages that provide evidence that languages change over time towards the achievable limit. These results suggest that natural language optimizes for efficiency over time, and does so in a way that is appopriate for compositional codes.


Keywords: Communicative efficiency; Uniform Information Density; Smooth Signal Hypothesis; Noisy channel

## Introduction

Natural language researchers have long been interested in the prospect that natural language is organized for efficient communication: frequent words tend to be shorter than rare words, allowing talkers to produce shorter word-forms on average. More recent work (Plotkin \& Nowak, 2000; Genzel \& Charniak, 2002; Aylett \& Turk, 2004; Levy \& Jaeger, 2007; Jaeger, 2010; Piantadosi, Tily, \& Gibson, 2011; Seyfarth, 2014) has sought to formalize this work in the context of information theory by proposing that natural language communicates information close to a limit called the channel capacity, although it has left open the question of how closely the limit is approached.

In this paper, I first show that it is not possible for a compositional code like natural language to transmit information close to the channel capacity. Specifically, average signal lengths must exceed the entropy of the distribution over messages by at least the Kullback-Leibler divergence of the true probability distribution over messages from a fully-factorized probability distribution in which every component of the message is statistically independent. Natural language communication must then underperform the channel capacity by at least this Kulback-Leibler divergence.

However, in light of recent work (Piantadosi et al., 2011) that showed a stronger relationship between a word's length and its average probability in context than its unigram probability, I investigate the possibility that language changes over time so that the optimal length of a word, as computed from its average probability in context, better matches its actual length. I present two corpus studies over 350 years in American English, Spanish, and Mandarin Chinese that compare
actual word lengths to optimal word lengths in context and in isolation. The first 'backward-looking' study computes optimal word lengths using modern-day language data, and finds that the mismatch between optimal and actual word lengths is smaller for older words. Moreover, the mismatch drops more rapidly, as a function of word age, when the optimal word length is computed relative to a context-sensitive trigram model than when it is relative to a unigram model.

The second 'forward-looking' study divides the 350-year period into 25-year partitions, uses language data from each partition to compute optimal word lengths for each partition, and trains a regression model to predict whether the wordform appears in the next 25-year partition as a function of the mismatch between the word's actual length and its optimal length. This study finds that words with larger mismatches are less likely to 'survive' to the next partition, and moreover finds a stronger effect of mismatch relative to the contextsensitive trigram model than relative to the unigram model. Together, these two studies provide evidence that natural language lexicons change over time in a way that reflects communicative efficiency pressures on a compositional code.

I start by presenting previous work on informationtheoretic approaches to language production, along with the minimal technical background necessary for this paper. I then show why natural language does not approach informationtheoretic bounds, and use this result to suggest a new bound for compositional codes that may be achievable by considering probability in context. Finally, I present two corpus studies that find evidence that three typologically-diverse languages have changed to approach this new bound.

## Background

Linguists have proposed that language is adapted for communication in a general sense for decades. Zipf (1949) proposed the 'Principle of Least Effort' to explain the observation that frequent words tend to be short: frequent words tend to be short so that talkers usually only have to say short words. Lindblom (1990) proposed Hyper- \& Hypo-articulation theory to explain the observation that vowels in careful speech tend to be less centralized in formant space: talkers provide more distinct vowels when they believe errors are more likely.

Plotkin and Nowak (2000) proposed an explicit model of word formation over the course of language change in an information-theoretic framework, and showed, analytically and via simulation, that it approached information-theoretic bounds as the vocabulary size increases. However, their model considered words in isolation, but natural language ut-
terances consist of sequences of words. This paper will show that codes with the compositional structure that characterizes natural language cannot approach these information-theoretic bounds, and focus on optimizing sentence lengths.

Subsequent work has mainly consisted of corpus studies that show that synchronic samples of natural language exhibit the correlations on would expect, under an informationtheoretic account, between different measures of word length or distinctiveness and word probability, both overall and in context (Aylett \& Turk, 2004; Frank \& Jaeger, 2008; Bell, Brenier, Gregory, Girand, \& Jurafsky, 2009). Piantadosi et al. (2011) revisited Zipfian distributions, and compared both log word probability and average $\log$ word probability in context, operationalized as a trigram model, with word lengths, operationalized as the length of the word's spelling in letters. They found that, of the two probability measures, average log probability in context exhibited a stronger correlation with word lengths. At first glance, this result would seem to contradict an information-theoretic approach: the optimal length of a word in isolation simply is its negative log probability, with an appropriate choice of base for the logarithm. However, I soon show that average probability in context is a more appropriate measure for optimizing sentence lengths.

In this paper, I return to Plotkin and Nowak's (2000) proposal that language adapts towards these bounds over time, using a corpus-based methodology and an emphasis on the relationship between words and their sentential contexts.

## Information theory

In the information-theoretic framing of language, a talker has a message $m$ that is a sequence of message characters $m_{i}$ from some alphabet of message characters $\mathcal{M}$. For example, $\mathcal{M}$ may be the set of lexical entries or, as in CCG, a set of (syntactic category, semantic category) pairs. The message cannot be transmitted directly, so the talker encodes it into a signal $s$ that is a sequence of signal characters $s_{i}$ from some alphabet of signal characters $\mathcal{S} .{ }^{1}$ For example, $\mathcal{S}$ may be the inventory of syllables or phonemes of the language.

An efficient code has two properties. First, it is short: the number of signal characters per message character, on average, is low. Second, it is robust: the probability that the listener fails to identify the correct message is low. The length of the shortest possible code depends only on the probability distribution over messages $P(\boldsymbol{m})$, and is given by the entropy of that distribution:

$$
\begin{equation*}
H(\boldsymbol{m})=\sum_{m \in \boldsymbol{m}} P(m) \log _{b}\left(\frac{1}{P(m)}\right) \tag{1}
\end{equation*}
$$

where $\boldsymbol{m}=\mathscr{M}^{+}$is the set of all messages (i.e. all sequences of message characters). The log term is called the Shannon information of $m$, and the entropy is just the expected Shannon information under the probability distribution $P(m)$. If we set

[^166]$b$ to be the size of the signal alphabet $|S|$, then the Shannon information of $m$ is the optimal signal length in signal characters for message $m$. Adjusting signal lengths to match the Shannon entropy, called source coding eliminates redundancy in the code, and achieves property 1 : short codes.

Listeners may encounter noise in real-world situations, due to slips of the tongue on the part of the talker, distraction or cognitive overload on the part of the listener, dialect differences, environmental noise, or other sources of noise. Noise can be countered by adding redundancy to the signal. For example, a word may differ from all other words by several phonemes, allowing the listener to recover the intended word even if some phonemes are mis-perceived or masked by environmental noise. While the resulting code is more robust, it is also longer, and we might worry that signals will have to become arbitrarily long to drive the error rate toward zero.

The Noisy Channel theorem shows that an arbitrarily low error rate can be achieved with signals that are not arbitrarily long, as long as they do not exceed the channel capacity (Shannon, 1948). The channel capacity depends on both $H(\boldsymbol{m})$ and the uncertainty about the intended signal, given the received signal. Adding redundancy, such as pronouncing words more slowly, that anticipates likely noise is called channel coding, and makes signals robust but still short.

For our purposes, the crucial observation is that it is not possible to get arbitrarily close to the channel capacity if it is not possible to obtain a source code that is arbitrarily close to the entropy of the distribution over messages. The next section shows that, for compositional codes like natural language, optimal source coding is not possible. ${ }^{2}$

## Compositional codes and optimality

This section shows that optimal source coding is not possible for a compositional code like natural language. If a code is optimal and compositional, then it follows that the components of every message are statistically independent. However, this is not true for natural language, since, e.g., transitive verbs tend to appear with at least two noun phrases.

By 'compositional,' I mean only that natural language messages consist of components that are realized the same way across different messages, and that the length of the signal for a message is the total length of the signal for each component of that message. Setting $l_{m}$ to be the length of the signal for message $m$ and $l_{m_{i}}$ to be the length of the signal for component $m_{i}$, compositionality provides:

$$
\begin{equation*}
l_{m}=\sum_{i=1}^{|m|} l_{m_{i}} \tag{2}
\end{equation*}
$$

For example, if a message is a sequence of lexical entries, and a signal is a sequence of phones, then Equation 2 says that the length of a sentence in phones is the sum of the lengths

[^167]of the phonological forms of the lexical entries in that sentence. ${ }^{3}$ Equation 2 is not trivial. Arithmetic codes, for example, encode each message as a number between 0 and 1 that is determined by the conditional probability of each message character given the previous message characters; an individual message character is not directly expressed in any part of the signal, and $l_{m_{i}}$ is not even defined.

Now assume that the length of the sentence $l_{m}$ for each message $m$ is optimal. Because the optimal signal length for a message $m$ is $-\log _{b}(P(m))$, the probability distribution over messages $P_{\boldsymbol{m}}$ can be recovered from $l_{m}$ by exponentiating:

$$
\begin{equation*}
P_{\boldsymbol{m}}(m)=b^{-l_{m}} \tag{3}
\end{equation*}
$$

Since only sentence lengths are assumed to be optimal, component signal lengths $l_{m_{i}}$ may not be negative $\log$ probabilities. They do, however, assume an implicit distribution for which they are least sub-optimal (MacKay, 2003, Ch. 5):

$$
\begin{equation*}
Q_{m}\left(m_{i}\right)=\frac{b^{-l_{m_{i}}}}{z} ; z=\sum_{m_{i} \in \mathcal{M}} b^{-l_{m_{i}}} \tag{4}
\end{equation*}
$$

Equations 2, 3, and 4 imply that each message component is statistically independent:

$$
\begin{align*}
P_{\boldsymbol{m}}(m) & =b^{-l_{m}}=b^{\Sigma_{i=1}^{|m|}-l_{m_{i}}}=b^{\Sigma_{i=1}^{|m|} \log _{b}\left(z Q_{m}\left(m_{i}\right)\right)}  \tag{5}\\
& =z^{|m|} \prod_{i=1}^{|m|} Q_{m}\left(m_{i}\right) \stackrel{\text { def }}{=} Q_{\boldsymbol{m}}(m) \tag{6}
\end{align*}
$$

There does not seem to be any notion of message in natural language that allows for statistically independent message components. For example, messages may be high-level event representations, but such messages that include transfer tend to include at least three entities (a giver, a receiver, and a thing being transferred), and such messages that include edible entities tend to include entities that can eat. Alternatively, messages may be syntactic analyses, but such messages with a determiner tend to have at least one noun, and such messages with a complementizer tend to have at least two main verbs. While other framings are possible, they do not appear to satisfy the independence assumption above. Thus, natural language is not information-theoretically optimal.

More specifically, the average signal length of the best compositional source code must exceed the entropy of the true distribution over messages $P_{\boldsymbol{m}}$ by at least the KullbackLeibler divergence of the fully factorized distribution $Q_{m}$ from the true distribution:

$$
\begin{equation*}
H\left(P_{\boldsymbol{m}}\right)+\mathrm{KL}\left(P_{\boldsymbol{m}} \| Q_{\boldsymbol{m}}\right) \tag{7}
\end{equation*}
$$

Intuitively, language uses at least an extra $\operatorname{KL}\left(P_{\boldsymbol{m}} \| Q_{\boldsymbol{m}}\right)$ signal characters per message character because it incorrectly assumes the message characters are statistically independent.

[^168]
## Optimizing towards the new bound

While the bound in Equation 7 shows that natural language does not approach the channel capacity, natural languages may still adapt over time for communicative efficiency towards the less efficient bound. In fact, the findings of Piantadosi et al. (2011) suggest that languages adapt to minimize $\mathrm{KL}\left(P_{\boldsymbol{m}} \| Q_{\boldsymbol{m}}\right)$. Piantadosi et al. examined how a word's length (in orthographic letters) relates to its unigram probability and its probability in context (operationalized as a smoothed trigram model). Across all eleven languages they examined, word lengths had a stronger relationship with average probability in context than unigram probability.

I propose the following interpretation of their result. Message characters are lexical entries, signal characters are orthographic letters, and probability in context is $P_{m}$. While $Q_{m}$ is determined by word lengths, $P_{m}$ is determined by a stochastic grammar and lexicon together with typical real-world situations. Their results suggest that, as a speech community gains experience with the use of lexical entries in real-world situations, the grammar, including the lexicon, adapts so that $P_{m}$ is better approximated by $Q_{m}$. This adaptation could be achieved by adjusting the grammar, narrowing or broadening word meanings, or deleting lexical entries whose length often differs substantially from their optimal in-context length.

The next two sections present corpus studies that look at adaptation of this sort over centuries in three languages.

## Corpus studies

I now present two corpus studies that find evidence of optimization relative to probability in context over time for English, Spanish, and Mandarin Chinese. The first 'backwardslooking' study relates a word's mismatch with its optimal length to its age. If the lexicon evolves over time for efficient communication, the lengths of oldest words should most closely match their optimal lengths. Moreover, to the extent that efficiency pressures respect sentence length, there should be a stronger relationship between a word's age and its mismatch with optimal lengths under a trigram model than between a word's age and its mismatch under a unigram model.

The 'forwards-looking' study uses a sequence of language models, estimated in 25-year partitions, to predict whether a word appears in the next partition based on how well its length matches its optimal length under each language model. If language change reflects efficiency pressures, words with many extra characters should be less likely to remain in use; and if efficiency pressures respect sentence lengths, the effect of mismatches under trigram models should be stronger.

## Corpus study 1 - Looking backwards

In this study, I used a large dataset containing texts from about 1990 to about 2010 for each of the three languages to compute synchronic unigram and trigram language models for each language. The language models are used to compute optimal lengths for each word in and out of context by subtracting the optimal lengths from the actual lengths to quantify extra characters. I used Google books, a dataset of scanned books, to

Table 1: Study 1 dataset sizes

|  | Language Model | Regression |  |
| :--- | :--- | :--- | :--- |
| Dataset | Tokens | Unigrams | Trigrams |
| English | $71,531,906$ | 81,742 | $24,965,851$ |
| Spanish | $279,744,284$ | 545,708 | $74,144,973$ |
| Chinese | $26,800,660$ | 3,182 | $13,642,166$ |



Figure 1: Heatmaps of actual length minus optimal word length for trigram (left) and unigram (right) models, as a function of the word's earliest appearance in Google books. The blue line is a GAM fit.
estimate when each word of the synchronic time slice first appeared, and perform a regression of the extra character measure against year of first appearance, probability model type, and their interaction, to identify how word length inefficiency varies as a function of word recency and probability model type. A positive coefficient for year of appearance will indicate that more recent words are longer than they should be, and a negative coefficient for the interaction will indicate a stronger relationship between year of appearance and inefficiency relative to a trigram model than between year of appearance and inefficiency in isolation.
Data I approximated a word's year of first appearance as the first year that it appeared in the Google Books unigram records in each language (Michel et al., 2010).

To estimate the American English language models, I used the spoken portion of the Corpus of Contemporary American English (CoCA) (Davies, 2008), which contains news broadcasts from 1990 to 2012. To estimate the Spanish and

Chinese language models, I used 'story' documents from the $3^{\text {rd }}$ edition of the Spanish Gigaword dataset of newswire text from 1993 to 2010 (Ângelo Mendonça, Jaquette, Graff, \& DiPersio, 2011) and the Tagged Chinese Gigaword version 2.0 dataset of newswire text from 1991 to 2004 (Ren Huang, 2009), respectively. While written Chinese does not separate words with whitespace, this dataset is segmented into words.

For each language, I discarded punctuation and words that contained a symbol that was not part of the usual character set for that language, estimated unsmoothed trigram and unigram probabilities. The datasets for regression were obtained by discarding words that did not appear in Google Books after 1650, producing datasets with sizes as reported in Table 1.

These particular languages were chosen because they appeared in Google Books, allowing us to obtain an estimate of word age, and because they use words in very different ways. Spanish has relatively rich derivational and inflectional morphology, with agreement for person and number for verbs and number and gender for adjectives. While English also has relatively rich derivational morphology, it has little inflectional morphology with few agreement constraints. Mandarin Chinese occupies a morphological extreme, with no inflectional morphology or agreement.

Method For each word token in CoCA and Gigaword datasets, I computed the optimal length of the word under its unigram probability and probability in context, operationalized as its trigram probability. The numbers of 'extra' letters relative to each model $e_{\text {uni }}$ and $e_{\text {tri }}$ are then the actual length minus the optimal length:

$$
\begin{aligned}
e_{\mathrm{uni}}(w) & =l(w)-\left(-\log _{b}(P(w))\right) \\
e_{\mathrm{tri}}\left(w_{i} \mid w_{i-2}, w_{i-1}\right) & =l\left(w_{i}\right)-\left(-\log _{b}\left(P\left(w_{i} \mid w_{i-2}, w_{i-1}\right)\right)\right)
\end{aligned}
$$

where $b$ is the size of the signal alphabet. English and Spanish both have a mostly alphabetic orthography, with roughly one letter per sound, so I simply set $b$ to the number of distinct letters in these datasets. For English $b=27$ (a-z plus hyphen), and for Spanish, $b=33$ (with some additional accented letters). Chinese orthography has one character per syllable, and so similarly provides a good indication of word length, but the alphabet size is more complicated. The strict phonotactics of spoken Chinese lead to a syllabic inventory of about 1,500 syllables, but our Chinese dataset contained 6,780 distinct characters (many characters are homophonous). I set $b=1,518$, the number of distinct syllables in the CEDict pronouncing dictionary, to reflect the size of the 'syllable alphabet' for spoken Chinese (CC-CEDICT, 2016). ${ }^{4}$

I performed linear regressions of extra letters against the word's year of first appearance, probability model type, and an interaction between the word's first appearance and probability model type. To make the regressions easier to interpret, I subtracted 1650 from the year of first appearance, so that the oldest words had a year of first appearance of zero.

[^169]Table 2: Coefficients of one linear regression each for American English, Mandarin, and Spanish, of extra letters (English, Spanish) or extra characters (Mandarin) against first appearance, a main effect of probability model type (with unigram coded as 1), and an interaction between first appearance and probability model. All coefficients are significant ( $p<0.01$ ).

|  | Am. English | Spanish | Mandarin |
| :--- | :--- | :--- | :--- |
| Intercept | 3.289 | 4.408 | 1.067 |
| Year of first appearance (since 1650) | 0.006481 | 0.007399 | 0.00214 |
| Which language model | -0.491 | -2.506 | -1.288 |
| Years of first appearance (since 1650) $\times$ Which language model | -0.001465 | -0.001438 | -0.000349 |

Results Figure 1 presents hexagram-binned heatmaps with a Generalized Additive Model fit for each language of extra letters against year of first appearance, separated by language model. All cases show a broad trend where older words have fewer extra letters. The trend is roughly linear except for the latest decades; information-theoretic pressures may be different for recently-coined words.

Table 2 presents coefficients from the three regressions of extra characters against a word's year of first appearance, the probability model used, and their interaction. Each intercept expresses the number of predicted extra letters or characters under the trigram model for words that first appeared in 1650. Adding the 'Which Language Model' coefficient to the intercept obtains the predicted extra letters or characters under the unigram model for words that first appeared in 1650.

The 'Year of first appearance' coefficient expresses how many extra characters we expect a word to have for each year that it is younger than the oldest words. For all three languages, this coefficient is positive, indicating that younger words tend to be longer, than older words, relative to their ideal length under the trigram model. Dividing this coefficient into 1 obtains how old we expect a word to be before an additional letter or character has been 'optimized' away. American English optimizes one letter every 154 years, Spanish optimizes one letter every 135 years, and Mandarin Chinese optimizes one character every 467 years. ${ }^{5}$

Finally, the interaction between year of first appearance and model type expresses the effect of a word's year of appearance under the unigram model minus the effect of a word's year of appearance under the trigram model. The coefficients are negative but smaller in absolute magnitude than 'Year of first appearance,' which indicates that first appearance still has a lengthening effect relative to the unigram model, but a weaker one. American English optimizes an extra letter relative to the unigram model only every 199 years, Spanish optimizes an extra letter only every 168 years, and Mandarin optimizes an extra character only every 558 years.

These results show that words that first appeared in books recently tend to be further from their informationtheoretically optimal lengths than words that first appeared in books several decades ago, and so provide evidence of optimization of the lexicon towards efficiency bounds.

[^170]Moreover, the extra characters relative to the trigram model decreased faster than the extra characters relative to the unigram model. This is a remarkable finding, since it is much harder to optimize for the trigram model - there are many trigram contexts but only one unigram 'context,' and, under this operationalization of 'word,' a word has only one length. However, as previously discussed, there are good reasons to optimize towards a context-sensitive probability model. Communicative efficiency ultimately depends on sentence lengths, not word lengths directly, so considering context can make sentences shorter even if it does not minimize the typical length of individual words.

## Corpus study 2 - Looking forwards

This corpus study looks for evidence that a word is less likely to remain in use if it has more extra characters. For each language, I divided the 350 years of Google Books data described above into 14 partitions of 25 years each, and estimated a unigram and a trigram language model for each of the first 13 partitions to compute extra characters for each word and trigram in each partition under each probability model. To guard against OCR errors in Google Books, I computed extra characters only for words that also appeared in the language model datasets from Study 1. I then performed a logistic regression that predicted whether each word that appeared in partition $n$ also appeared in partition $n+1$ using the extra characters measure, probability model type, and an interaction between extra characters and the probability model type.

Results Table 3 presents strikingly consistent regression results across the three languages. The large intercepts indicate that most words carry over from one partition to the next. As the unigram model is again coded as 1 , the negative main effect of extra characters indicates that words with more extra letters relative to a given partition's trigram model are less likely to persist in the next 25-year partition. Moreover, the positive coefficient of the interaction indicates that the effect of extra letters relative to the unigram model is weaker: the coefficients suggest the effect of unigram mismatch is about half the effect of trigram mismatch in English, two-thirds in Spanish, and about one-third in Mandarin.

## Conclusion

This paper has answered an important question about natural language communication, whether talkers approach information-theoretic limits on efficiency, in the negative. Be-

Table 3: Coefficients of a logistic regression each for American English, Mandarin, and Spanish, of appearance in the next 25year partition against extra letters or characters, a main effect of probability model type (unigram coded as 1), and an interaction between extra characters and probability model. All coefficients are significant ( $p<0.01$ ).

|  | Am. English | Spanish | Mandarin |
| :--- | :--- | :--- | :--- |
| Intercept | 11.772 | 7.477 | 10.415 |
| Extra letters or characters | -0.316 | -0.320 | -2.360 |
| Which language model | -1.015 | -0.852 | -2.415 |
| Extra characters $\times$ Which language model | 0.165 | 0.101 | 1.713 |

cause language is compositional and natural language messages are highly interdependent, natural language cannot approach information-theoretic limits on efficiency. I have used this result to propose a new bound that appreciates probability in context, and interpreted a previous result as evidence that languages optimize for this more appropriate bound.

I then performed two corpus studies that examined how the mismatch between a word's actual length and its optimal length relates to its preservation over the course of language change. The first 'backwards-looking' study found, using optimal lengths computed using fairly homogenous modern-day corpus data, that present-day words more closely match their optimal lengths if the word has been in use for a long time. Moreover, this first study found that the mismatch according to probability in context decreased more rapidly as words age. The second 'forwards-looking' study found that if a word's length more closely matches its optimal length under a language model computed in one 25 -year partition, it is more likely to be retained in the next 25 -year partition. Moreover, extra letters relative to probability in context was a stronger predictor than extra letters relative to a unigram model. Together, these results indicate that natural language lexicons develop over time towards an information-theoretic efficiency bound that is appropriate for compositional codes.

## Acknowledgements

I thank Rui Chaves for feedback on a very early draft, and anonymous reviewers for helpful comments.

## References

Ângelo Mendonça, Jaquette, D., Graff, D., \& DiPersio, D. (2011). Spanish Gigaword Third Edition LDC2011T12 [Computer software manual]. Web download. Philadelphia, PA.
Aylett, M., \& Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech, 47(1), 31-56.
Bell, A., Brenier, J. M., Gregory, M., Girand, C., \& Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language, 60, 92-111.
CC-CEDICT. (2016). http://www.mdbg.net/chindict/. (Accessed: 2016-30-05)
Davies, M. (2008). The Corpus of Contemporary American

English: 520 million words, 1990 - present. (Available online at http://corpus.byu.edu/coca/)
Frank, A., \& Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In Proceedings of CogSci (pp. 933938).

Genzel, D., \& Charniak, E. (2002). Variation of entropy and parse trees as a function of sentence number. In Proceedings of the Association for Computational Linguistics.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61, 23-62.
Levy, R., \& Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In Proceedings of NIPS.
Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H \& H theory. In W. Hardcastle \& A. Marchal (Eds.), Speech production and speech modelling (pp. 403-439). Kluwer Academic Publishers.
MacKay, D. J. C. (2003). Information theory, inference, and learning algorithms. Cambridge University Press. (http://www.inference.phy.cam.ac.uk/mackay/itila/)

Michaelis, J., \& Kracht, M. (1996). Semilinearity as a syntactic invariant. In Proceedings of LACL.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Brockman, W., ... Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books. Science, 331.
Piantadosi, S., Tily, H., \& Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108(9), 3526.
Plotkin, J. B., \& Nowak, M. A. (2000). Language evolution and information theory. Journal of Theoretical Biology, 205, 147-159.
Ren Huang, C. (2009). Tagged Chinese Gigaword Version 2.0 LDC2009T14 [Computer software manual]. Web download. Philadelphia, PA.
Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. Cognition, 133(1), 140-155.
Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423.
Zipf, G. K. (1949). Human behavior and the principle of least effort. Addison-Wesley.

# Relational Concept Learning via Guided Interactive Discovery 

John D. Patterson ${ }^{1}$ (jpatter4@binghamton.edu)<br>David Landy ${ }^{2}$ (dlandy@indiana.edu) Kenneth J. Kurtz ${ }^{1}$ (kkurtz@binghamton.edu)<br>${ }^{1}$ Department of Psychology, Binghamton University, 4400 Vestal Parkway East, Binghamton, NY 13905 USA<br>${ }^{2}$ Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street Bloomington, IN 47405 USA


#### Abstract

A key goal in both education and higher-order cognition research is to understand how relational concepts are best learned. In the current work, we present a novel approach for learning complex relational categories - a low-support, interactive discovery interface. The platform, which allows learners to make modifications to exemplars and see the corresponding effects on membership, holds the potential to augment relational learning by facilitating self-directed, alignably-different comparisons that explore what the learner does not yet understand. We compared interactive learning to an identification learning task. Participants were assessed on their ability to generalize category knowledge to novel exemplars from the same domain. Although identification learners were provided with seven times as many positive examples of the category during training, interactive learners demonstrated enhanced generalization accuracy and knowledge of specific membership constraints. Moreover, the data suggest that identification learners tended to overgeneralize category knowledge to non-members - a problem that interactive learners exhibited to a significantly lesser degree. Overall, the results show interactive training to be a powerful tool for supplementing relational category learning, with particular utility for refining category knowledge. We conclude with implications of these findings and promising future directions.


Keywords: relational categories; structural alignment; discovery learning; category learning; generalization

## Introduction

A key aim of higher-order cognition research is to understand the mechanisms that undergird the ability for humans to acquire and use abstract, complex categories. The literature in concepts and categories research has primarily been devoted to the study of attribute categories categories whose members possess a set of independent features by which they can be classified. Research on attribute category learning has unequivocally advanced our understanding of human concept acquisition and its many facets.

However, much of the category knowledge we possess is not reducible to knowledge of specific attributes - myriad concepts such as positive feedback loop are abstract, attribute-agnostic, and dependent on relationships rather than features. Fittingly, an increasing amount of empirical attention has been granted to the study of relational categories (Gentner \& Kurtz, 2005; Markman \& Stilwell, 2001). Relational categories are rule-like categories whose members share a common relational structure characterized by extrinsic relationships between objects and/or attributes (e.g., protection, sibling, reciprocity). Because relational
categories need only share a relational structure to belong, members of a category can be quite featurally disparate (e.g., your sibling and your dog's sibling hopefully don't look alike). In this way, relational category members share analogical similarity. It should be noted that relational categories are not an idiosyncratic facet of category learning - roughly half of the 100 highest frequency nouns are relational (Asmuth \& Gentner, in press). Thus, to understand human category learning generally, it is critical to understand relational category learning.

A question that bears both theoretical and applied import is: how do we come to acquire relational category knowledge? Previous research has explored the potential for comparison to promote relational discovery and transfer. This work follows from a large body of research showing the benefits of comparison to analogical transfer (Gick \& Holyoak, 1983; Loewenstein, Thompson, \& Gentner, 1999; see also Alfieri, Nokes-Malach, \& Schunn, 2013 for a metaanalysis and review). Studies of comparison with relational categories have largely corroborated findings from the analogical transfer literature; presenting same-category pairs (Patterson \& Kurtz, 2015) or a mixture of same- and different-category pairs (Kurtz, Boukrina, \& Gentner, 2013) during training leads to enhanced learning and transfer over sequential item presentations. The power of comparison can be understood through a process of structural alignment (Markman \& Gentner, 1993). Comparing instances facilitates the alignment of their parallel relational predicates. This serves to highlight common relational structure that is not salient when either instance is considered in isolation. Additionally, comparison facilitates abstraction, which promotes later analogical retrieval and transfer.

As many of the core concepts taught in educational settings are relational in nature (e.g., evolution by natural selection, Newton's laws), relational categories represent a key bridge between cognitive and educational research (Goldwater \& Schalk, 2016). Thus, investigating how relational categories are best learned can both palpably advance educational techniques and further basic, theoretical understandings. In the present work, we draw on an innovative area of education research that serves as a promising avenue for enhancing relational category learning: discovery learning. Discovery learning generally refers to unsupported learning where the learner actively constructs their understanding of some target information using only a set of materials or a task environment. Though many flavors of discovery have been the subject of study, a
common theme in the literature is that completely unassisted discovery approaches are not effective for learning (Mayer, 2004; for a meta-analysis see Alfieri, Brooks, Aldrich, \& Tenenbaum, 2011). Among other reasons, the large cognitive load incurred by needing to generate and explore hypotheses (Sweller, 1988) while metacognitively maintaining an idea of what is known and what needs to be known (Kirschner, Sweller, \& Clark, 2006) can present challenges for the approach. However, when some guidance is introduced (such as direct instruction - e.g., Chen \& Klahr, 1999), discovery learning can be a highly effective tool (Alfieri et al., 2011).

Discovery learning has the clear potential to augment the learning of complex relational concepts in educational settings - particularly when the target category is abstract or when classroom instruction is subpar. With a basic understanding of the target category, an interactive environment that enables learners to freely create or modify category exemplars and receive dynamic category membership feedback ought to enhance category knowledge, notably through three mechanisms. First, it would allow learners to engage in self-directed exploration that is specifically catered to what they do not understand or need further clarification on. The opportunity to select exemplars for study has been shown to confer benefits on rule-based category learning (e.g., Markant \& Gureckis, 2014). Second, the dynamic membership feedback provided by the task interface would implicitly encourage explanations about the causes underlying the effects of learners' modifications. Such self-explanation has been demonstrated to be a powerful facilitator of concept acquisition (e.g., Chi, de Leeuw, Chiu, \& LaVancher, 1994). Third, critically, a learning environment such as this should strongly engage analogical processing faculties. In modifying an exemplar and receiving membership feedback, the learner effectively creates a temporally juxtaposed comparison between the item's new state ( $s$ ) and $s-1$. Modifications that do not break membership create alignably-different, same-category comparisons. These comparisons should promote highlighting of common relational structure and facilitate abstraction. Conversely, modifications that do break membership create alignablydifferent different-category comparisons, which critically should serve to highlight membership-relevant relations.

In the present work, we explore the efficacy of a lowsupport, interactive discovery learning tool to promote the learning of complex relational categories. To avoid effects of domain knowledge, we created an artificial, multiconstraint category that served as the target of learning. Advised by the discovery learning literature and pilot data, we gave participants some support to reduce cognitive load. This support was a clear, but quite abstract, definition of the category that was given to all learners immediately prior to the learning phase. In the interactive condition, participants were given a computerized interface where they could engage in self-directed exploration of three examples of the category. We contrasted interactive training with an
identification learning control in which learners were exposed to a larger number of exemplars in the context of a member identification task. To evaluate the effectiveness of the interactive learning mode, we compared the two learning conditions on their ability to generalize category knowledge to novel exemplars. We predicted interactive learning would lead to enhanced generalization performance.

## Method

## Participants

Seventy undergraduates from Binghamton University participated to partially fulfill a course requirement.

## Materials

The training and generalization stimuli consisted of arrangements of blocks that varied in their size (small, medium, or large), color (white, gray, or dark brown), border color (black or distinctive blue), and spatial location (see Figure 1 for examples). The 'matched containment' concept instantiated by these blocks was quite complex. Category members were characterized by the presence of three or more blocks that obeyed all of the following constraints: (1) the blocks were aligned vertically or horizontally, (2) two of the involved blocks were special by sharing a distinctive blue border color, (3) the special blocks were exactly matched in their attributes, (4) the special blocks contained/flanked at least one additional 'normalbordered' block in the lineup, and (5) all of the contained, normal blocks matched the special blocks on at least one attribute (i.e., color, size, or both).

Twenty-one category members and 21 non-members were used as the stimuli for the identification condition. All members contained the category-defining core - constituted by either three (Length 3 [Len3]; two flanking, one flanked) or four (Length 4 [Len4]; two flanking, two flanked) objects - and one additional distracter block, such that all examples had length +1 blocks. The category core and distracter block were varied in their attributes (i.e., orientation of core [vertical, horizontal], spatial location, color, size) across examples to ensure an attribute-based solution was not available.

The members were comprised of six item types, each which instantiated the special-normal match constraint in a unique way. For both Len3 and Len4 stimuli, there were items whose flanked object(s) matched based on (1) color, (2) size, (3) or both color and size. For Len4 stimuli there were also items whose flanked objects consisted of (4) one color and one size match, (5) one both and one color match, or (6) one both and one size match. The item breakdown can be seen in Table 1. Since the Len4 stimuli included matching types that were distinct from those present in the Len3 stimuli, the Len4 examples were weighted on types 46 to ensure comprehensive coverage of the category for identification learners. The non-member set used was programmatically generated by randomly sampling and arranging blocks, with the constraint that two of the blocks
had to possess the distinctive border. Number of blocks was matched between the members and non-members. The interactive condition was given considerably fewer examples: one Len 3 color match and two Len 4 examples, each which had one color match and one size match.

To evaluate participants' ability to generalize their knowledge, a distinct set of 30 members and 30 nonmembers was created. The members were sampled from several match types. Critically, non-members consisted of items that violated the constraints of membership in several focal ways (see Table 2 for generalization item breakdown). As knowledge of the specific constraint that was violated was necessary to get each of these items correct, they served as a stringent test of category knowledge.


Figure 1: Eight example stimuli from the training (identification) and generalization phases.

## Design and Procedure

Participants were randomly assigned to either identification ( $\mathrm{n}=39$ ) or interactive $(\mathrm{n}=31)$ learning conditions in a between-subjects design. Due to a spreadsheet error, condition assignments were slightly imbalanced.

In the pre-training instructions, all subjects were first informed they would be learning about something called a 'Togging situation' - the arbitrary category label - before being provided with an abstract definition of the category: "A Togging situation occurs when (1) there are two matching special objects with other objects in the space between them; and (2) all the objects in the space between have at least one thing in common with the special objects." Subjects were then told they were to gain a full and clear understanding of Togging situations by engaging in the upcoming learning experience.

Table 1: Number of category members by length and type for identification training.

| Special-Normal Match <br> Type | Length 3 Items | Length 4 Items |
| :--- | :---: | :---: |
| Color | 3 | 1 |
| Size | 3 | 1 |
| Both Size and Color | 3 | 3 |
| One Both, One Color | -- | 2 |
| One Both, One Size | -- | 2 |
| One Size, One Color |  | 3 |
| Total | 9 | 12 |

Condition-specific instructions for the interactive group informed learners they would receive (1) an 'exploration zone' that would tell them if the objects inside were currently in a Togging situation and (2) a set of 'exploration tools' that could be used to modify the objects/attributes in informative ways. They were then told they could gain an understanding by paying attention to modifications that break the Togging situation and by trying to create novel Togging situations. To combat confirmation bias - as piloting revealed this to be a considerable impediment to learning - subjects were also told to try to prove their ideas about Togging wrong by fully testing them. Lastly, subjects were informed they would be tested later and that they would have seven minutes with the learning task. Following these general task instructions, interactive learners then read a brief tutorial that described the ways they could modify the examples with the exploration tools.

Table 2: Number of exemplars by length, type, and membership in the generalization assessment.

| Members: | Length 3 Items | Length 4 Items |
| :--- | :---: | :---: |
| Special-Normal Match Type | 8 | 1 |
| Color | 7 | 2 |
| Size | 4 | 0 |
| Both Size and Color | - | 1 |
| One Both, One Color | - | 7 |
| One Size, One Color | 19 | 11 |
| Total | 8 |  |
| Non-member Type | 4 | 3 |
| No Special-Normal Match | 3 | 3 |
| No Special-Special Match | 2 | 0 |
| Only One Special Block | - | 0 |
| Special Objects not Aligned |  | 7 |
| Only One Special-Normal Match | 17 | 13 |
| Total |  |  |

Instructions in the identification condition informed learners they would receive a series of frames with objects inside, some of which would have a Togging situation. They were told they could gain an understanding by paying attention to the frames and feedback they received and by learning to identify which frames contained a Togging situation. The identification learners also received an analogous instruction to try to prove their ideas about Togging wrong by fully considering the frames and feedback on each trial. Finally, identification learners were informed of the upcoming test.

To remind participants, and guide learning, both conditions were again given the abstract category definition immediately prior to the learning task.

Training - Interactive Condition The training interface can be seen in Figure 2. In the center of the interface was an 'exploration zone' that dynamically checked whether the constraints of category membership were met by the objects inside. The zone's border color turned green if the constraints were met, and red if not. A textual notification above the zone regarding membership mirrored the color feedback. The exploration zone started with a positive example of the category, randomly selected from the three
positive examples that were provided to the interactive group.

Participants could freely engage the training interface in five distinct ways. First, clicking the 'new' button cycled through the three positive examples. The button allowed learners to reset the exploration zone to a positive example if they became lost with their current discovery path and also get experience with the three different instantiations of the category. Second, double clicking a block would change its color - cycling in order through the three colors with each double click. Third, clicking and dragging the bottom right corner of a block diagonally allowed participants to stretch or shrink it to one of the three discrete sizes. Fourth, clicking and dragging elsewhere on a block allowed participants to change its spatial location. Lastly, participants could add or remove objects from the exploration zone. To the left of the zone was a space containing additional normal blocks that varied in their size and color. Participants could bring additional blocks into the exploration zone or remove any of the blocks from the zone to this space. This allowed participants to: (1) swap blocks to change attributes, (2) simplify the example in the exploration zone, and (3) create more elaborate examples of the category that involved more objects.


Figure 2: Visual of the interactive workspace.
Since there were many ways participants could interact with the interface, they were provided with a 'how to' cheatsheet to the right of the exploration zone. During the task, participants had seven minutes for self-directed investigation of the category. A timer in the upper left corner of the interface showed how much time remained. To encourage participants to stay on task, a query was presented below the text notification of membership. The query corresponded to the current state of the objects in the exploration zone. When category constraints were met, the query asked participants if they could, "find a way to break the Togging situation." When the constraints were not met, it asked if they could "find a way to make a Togging situation again." Besides this general query, no additional direction was given during the task.

Training - Identification Condition Learners in the identification group were directly provided with 21
examples of the category - seven times as many as were provided to interactive learners. These 21 positive cases were combined with the 21 negative cases in a random order for each participant. Participants made one pass through the set. On each trial, the participant was presented with an item and two response buttons ("Togging situation", "Not a Togging situation"). The amount of time to study the example and make a response was unconstrained. Participants selected their response using the mouse and were given feedback indicating if they were (in)correct and whether the item was (not) an example of a Togging situation. Feedback was presented for 2.5 s before moving on to the next trial.

Generalization Assessment Following the learning phase, all participants were given a generalization assessment to assay their ability to both identify category members and correctly reject near-miss non-members. The 60 generalization items were presented in a randomized order for each participant. The trial structure of the generalization phase was identical to that of identification training, except no feedback was given.

## Results

## Training

All except three learners in the identification condition ( $M=$ $.83, S E=.02$ ) performed reliably above chance. Data from these non-learners were retained in the subsequent analyses for two reasons: (1) the general pattern of results did not change when their data were excluded, and (2) there was not a comparable basis for excluding interactive learners. Identification training took 3-8 minutes ( $M=3.89$ minutes, $S E=.14$ ). Though there was a wide range, time spent during training did not predict generalization accuracy in a trialwise logistic regression $(\beta=-0.005, S E=0.01, Z=-.44, p=$ .66).

Interactive learners made between 151 and 321 manipulations $(M=227.21, S E=6.60)$. Number of manipulations, however, did not predict generalization accuracy $(\beta=-0.0001, S E=0.001, Z=-.12, p=.91)$, suggesting that the quantity of manipulations was not critical. However, higher rates of crossover - the proportion of the manipulations that switched the state from member to non-member (or vice versa) - were associated with higher generalization accuracy ( $\beta=3.05, S E=0.86, Z=3.54, p<$ .001), suggesting that generating alignably-different different-category comparisons is key for getting the most out of the platform.

## Generalization Accuracy

Trial-wise accuracy data were modeled with logistic regressions. Using condition as the lone predictor, the main analysis yielded the key finding that interactive learning ( $M$ $=.73, S E=.01$ ) significantly augmented generalizable category knowledge over identification learning ( $M=.67$, $S E=.01) ; \beta=0.29, S E=0.07, Z=4.27, p<.001$.

To further probe the effect of condition, we conducted a follow-up analysis to see how each condition performed on members and non-members. To this end, we used condition, item membership ( 1,0 ), and their interaction as predictors. Interestingly, the regression revealed a highly reliable crossover interaction between condition and item membership (see Figure $3 ; \beta=1.35, S E=0.18, Z=7.70, p<.001$ ). The interaction was marked by a reliable enhancement for the identification group on category members (identification: $M$ $=.93, S E=.01$; interactive: $M=.87, S E=.01 ; \beta=-0.63, S E$ $=0.15, Z=-4.20, p<.001$ ), but a reliable enhancement for the interactive group on non-members (identification: $M=$ $.41, S E=.01$; interactive: $M=.59, S E=.02 ; \beta=0.72, S E=$ $0.09, Z=8.01, p<.001)$. It should be noted that average accuracy on non-members was generally low. This is directly attributable to their more challenging nature. Contrasted with the member set, on which it was possible to successfully identify all items using knowledge of any single relational constraint, the non-member set consisted of items that each focally violated a constraint of membership. To perform successfully on these, participants required knowledge of the specific constraint that was violated in each instance. Thus, performance on the non-members serves as a proxy for learners' understanding of the category's composite constraints. While learners still had much to learn about the category, low means should not be interpreted to mean that performance was at chance or random in nature. Rather, the high accuracy observed for members suggests that learners took a limited understanding of the constraints of membership and overgeneralized it to non-members.

Given the curious reversal in the effect of condition between levels of item membership, we were prompted to explore the possibility that identification learners were more likely to overgeneralize category knowledge, which ostensibly would explain this pattern of results. We used two signal detection theory measures to this end: $d^{\prime}$ and $\beta$. $d^{\prime}$ is a measure of sensitivity to the signal when present that reflects hit rate on signal trials while adjusting for false alarm rate on noise trials. A higher $d^{\prime}$ indicates a greater sensitivity to the underlying signal (category members). $\beta$ is a likelihood ratio that reflects response bias. A $\beta$ of 1 indicates learners were neither biased towards nor against extending the category label, whereas $\beta$ below or above 1 indicates a bias towards extending or not extending the label, respectively. $d^{\prime}$ and $\beta$ were computed for each subject and the values for each were then predicted by condition in separate linear regressions. Despite showing increased accuracy for members, identification learners were not more sensitive, owing to a significantly increased false alarm rate (identification: $M=0.58, S E=.04$; interactive: $M=0.41$, $S E=.04 ; \beta=-0.17, S E=0.05, t(68)=-3.19, p<.01)$. In fact, a numerical advantage in $d^{\prime}$ favored interactive learners but did not reach significance (identification: $M=1.46, S E$ $=.09$; interactive: $M=1.71, S E=.21 ; \beta=0.24, S E=0.21$, $t(68)=1.16, p=.25)$. Additionally, identification learners were found to be significantly more biased towards
endorsing items as members - showing lower $\beta$ than their interactive counterparts (identification: $M=0.34, S E=.06$; interactive: $M=0.61, S E=.09 ; \beta=0.27, S E=0.11, t(68)=$ $2.50, p<.05)$. Collectively, these measures indicate that the identification group's enhanced accuracy for members was not the result of greater sensitivity. Instead, it appears to be a byproduct of a liberal extension of a limited understanding of the category, relative to interactive learners.


Figure 3: Generalization performance by condition and item membership. Error bars represent $+/-1$ SE.

## Discussion

The primary goal of this study was to evaluate the potential for a novel, interactive discovery platform to facilitate the acquisition of a complex relational concept. Consistent with our hypothesis, our findings resolutely show that interactive training is an effective way to affect relational category knowledge. Compared to identification training - a learning mode organic to both category learning experiments as well as common educational practices - interactive learners exhibited an enhanced ability to generalize and enriched knowledge of specific membership constraints.

The results of this study inform both basic and applied interests. Our data suggest that our interactive platform can aptly supplement learning when complex, abstract relational categories are the target of learning. On an intriguing note, this paradigm appears to possess a distinct utility for combating overgeneralization by helping learners to explore and refine the boundaries of membership. It should be noted these advantages accrued despite the minimalistic support that was given (compared to other guided discovery approaches; e.g., Chen \& Klahr, 1999), the short amount of time allotted for learning, and the transfer appropriate processing advantage granted to identification learners in the shared task between training and test.

A limitation of this study is the use of randomly generated non-members in the identification training condition. As a function of the random generation, they tended to be slightly more entropic than the positive examples. This exposes a possible deflationary account of these findings - that identification learners may have simply learned to differentiate more and less entropic examples from each
other, which might explain poorer generalization performance. However, this account is unlikely for two main reasons. First, learners were provided a definition of the relational concept not once, but twice, prior to training. A basic understanding of the category should have guided learners to seek information that extended that understanding, not part with it altogether. Second, if learners acquired and used an entropy strategy during training, the effects of this should have been notable in the generalization data. Unlike the training set, non-members in the generalization phase were orderly. If learners adopted an entropy strategy, they would likely use it before realizing, later in the generalization phase, that there were not any entropic cases - at which point they might shift to the principle-relevant knowledge they acquired through the definition and learning experience. If this occurred, we should expect better performance later in the generalization phase. To investigate this possibility, we compared performance on the first 30 trials to the second 30 trials of generalization for identification learners. The difference was non-significant ( $p=.81$ ), suggesting identification learners engaged the task the way we intended. Nevertheless, planned research using yoked controls will provide more definitive evidence.

Further work will be necessary to specify the cognitive processes behind the benefits of interaction in relational category learning. Consistent with Markant \& Gureckis (2014), the effect of actively selecting modifications that supplement one's current understanding is likely to be critical. However, our next main pursuit in developing this platform is to more deeply explore the potential for analogy and comparison to serve as the engine for interactive relational category learning. Much of the power of this learning paradigm likely follows from its facilitation of informative, user-created comparisons with alignable differences - a possibility echoed by the higher generalization accuracy associated with higher rates of category crossover. To the extent that this underlies its utility, providing learners with co-presented exemplars that are dynamically linked in their manipulations should promote enhanced generalization and transfer, and possibly serve to shorten acquisition time. Contrasting this interactive approach with static comparisons and other educational tools, such as the explicit elicitation of selfexplanations, will be integral to the evaluation of this tool's potency in upcoming research.

## Acknowledgments

We greatly thank David Brokaw for his help with programming and the Learning and Representation in Cognition (LaRC) lab for their valuable input.

## References

Alfieri, L., Brooks, P. J., Aldrich, N. J., \& Tenenbaum, H.
R. (2011). Does discovery-based instruction enhance
learning?. Journal of Educational Psychology, 103(1), 118.

Alfieri, L., Nokes-Malach, T. J., \& Schunn, C. D. (2013). Learning through case comparisons: a meta-analytic review. Educational Psychologist, 48(2), 87-113.
Asmuth, J., \& Gentner, D. (2017). Relational categories are more mutable than entity categories. The Quarterly Journal of Experimental Psychology, 70(10), 2007-2025.
Chi, M. T. H., DeLeeuw, N., Chiu, M., \& LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.
Chen, Z., \& Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. Child Development, 70, 1098-1120.
Gentner, D., \& Kurtz, K. J. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, \& P. W. Wolff Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin. Washington, DC: American Psychological Association.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15(1), 138.

Goldwater, M. B., \& Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. Psychological Bulletin, 142(7), 729-757.
Kirschner, P.A., Sweller, J., \& Clark, R.E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41, 75-86.
Kurtz, K. J., Boukrina, O., \& Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(4), 1303-1310.
Loewenstein, J., Thompson, L., \& Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. Psychonomic Bulletin \& Review, 6(4), 586597.

Markant, D. B., \& Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. Journal of Experimental Psychology: General, 143(1), 94.
Markman, A. B., \& Gentner, D. (1993). Structural alignment during similarity comparisons. Cognitive Psychology, 25(4), 431-467.
Markman, A. B., \& Stilwell, C. H. (2001). Role-governed categories. Journal of Experimental \& Theoretical Artificial Intelligence, 13, 329-358.
Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning?. American Psychologist, 59(1), 14-19.
Patterson, J.D., \& Kurtz, K.J. (2015). Learning mode and comparison in relational category learning. Proceedings of the 37th Annual Conference of the Cognitive Science Society.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12, 257-285.

# The Use of Iconic Words in Early Child-Parent Interactions 

Marcus Perlman (marcus.perlman@mpi.nl)<br>Language \& Cognition Department<br>Max Planck Institute for Psycholinguistics, Nijmegen, 6500 AH The Netherlands

Riccardo Fusaroli (fusaroli@cc.au.dk)
Department of Communication and Culture \& the Interacting Minds Center, Aarhus University, Jens Chr Skous Vej 2, Denmark

Deborah Fein (deborah.fein@uconn.edu) \& Letitia Naigles (letitia.naigles@uconn.edu)<br>Department of Psychological Sciences<br>University of Connecticut, Storrs, CT


#### Abstract

This paper examines the use of iconic words in early conversations between children and caregivers. The longitudinal data include a span of six observations of 35 children-parent dyads in the same semi-structured activity. Our findings show that children's speech initially has a high proportion of iconic words, and over time, these words become diluted by an increase of arbitrary words. Parents' speech is also initially high in iconic words, with a decrease in the proportion of iconic words over time - in this case driven by the use of fewer iconic words. The level and development of iconicity are related to individual differences in the children's cognitive skills. Our findings fit with the hypothesis that iconicity facilitates early word learning and may play an important role in learning to produce new words.


Keywords: iconicity; vocabulary development; child-directed speech; sound symbolism

## Introduction

How do young children learn to understand and use their first words? Philosophers have pointed out the inductive challenge of learning to associate the sound of a word with its meaning (Quine, 1960), and developmental psychologists observe that this challenge is compounded for an infant (Imai \& Kita, 2014). Not only must infants isolate the meaning of a word from a noisy, complex environment and learn to generalize it to new contexts, they must grasp the very concept of a symbol in the first place.

One cue infants may utilize to facilitate early word learning is iconicity - a degree of resemblance between the form of a signal and its meaning. For example, a person might represent 'small size' with an index finger-to-thumb pinching gesture, or in speech, by raising the pitch of their voice as they articulate the small-sounding word "itty-bitty". One proposal for how iconicity can help early word learning in spoken languages is the sound symbolism bootstrapping hypothesis (Imai and Kita, 2014). (The term "sound symbolism" is often used to refer to iconicity in spoken words.) On this idea, children are biologically endowed with a bias to recognize various cross-modal correspondences between sound and phenomena of the other senses - for example, between pitch and size or brightness, or between
the duration of a sound and the visual extension of a line (see Spence, 2011 for a review of crossmodal correspondences). These biases, which may also be learned by experience, might bootstrap children into the connection between the sounds of iconic words and their corresponding meanings (Imai \& Kita, 2014; Perniss \& Vigliocco, 2014). The sound of a word could help children identify its intended referent from a complex scene, recognize the invariance of its meaning across contexts, and apply it productively to new contexts. By helping children connect spoken words with their meanings, iconicity may help children gain the "referential insight" that speech sounds refer to entities and properties in the world.

## Laboratory studies of iconicity in word learning

Laboratory studies show that young children are indeed sensitive to iconicity in spoken words, and some studies suggest further that this iconicity can facilitate word learning. Maurer et al. (2006) found that 2.5 -year-old children were more likely to match nonsense words containing rounded vowels (e.g. bouba) with rounded shapes, and words containing unrounded vowels (kiki) with pointed shapes. Another study, using a preferential looking paradigm, found that infants as young as four months are sensitive to this bouba-kiki-type iconicity (Ozturk et al., 2012). There is also evidence that sound-shape correspondence can more directly facilitate word learning in infants (Imai et al., 2015). 14-month-old infants were habituated to combinations of a novel word and a picture of an object, in either an iconically matching or mismatching condition. When the children were then presented with the novel words along with a picture of the correct object and a distractor, they looked more at the correct object when its sound and shape were matching.

The sensitivity of infants to sound-shape iconicity has also been demonstrated using electroencephalography (Asano et al., 2015). Event related potentials were measured as 11 -month-old infants were presented with pictures of shapes followed by novel words that matched or mismatched the shape in iconicity. With mismatching wordshape pairs, subjects showed a response similar to the N400 effect, typically an index of difficulty with semantic
integration. Analysis of brain oscillations found an increase in early $\gamma$-band oscillations in the matching condition, which might indicate increased cross-modal integration between the sound of the word and its visual referent.

Importantly, children can benefit from iconicity in learning to associate sounds and meanings in domains outside of shape. Imai et al. (2008) presented 25-month-old Japanese toddlers with novel words along with two video clips of people performing two different manners of walking. Norming with adult Japanese and English speakers had determined that one of the videos, but not the other, was iconically congruent with the verb. Similarly, when the children were asked to select the manner of walking to which they thought the word referred, they were more likely to choose the one that matched the verb. A subsequent experiment found that 3 -year-old Japanese children were able to correctly generalize the iconic verbs to new agents, but were unable to generalize non-iconic verbs. Further work replicated the finding with 3-year-old English learners, indicating at least some level of universality in the iconicity of the sound-referent pairings (Kantarzis et al., 2011). A study by Yoshida (2012) found similar results with Japanese- and English-speaking toddlers.

## Iconicity in natural word learning

These laboratory studies show that young children are sensitive to certain forms of iconicity in spoken words, and under some conditions, they can learn iconic words faster than non-iconic words. However, if iconicity does play an important role in early word learning in the wild, then there should be evidence of this in natural language learning. In particular, children should learn more iconic words earlier, and caregivers should be more inclined to use iconic words with early learners.

A recent study of English and Spanish found that children do tend to learn more iconic words earlier (Perry et al., 2015). Perry et al. collected native speaker iconicity ratings of roughly 600 English and Spanish words that are learned earliest by children according to the respective MacArthurBates Developmental Inventories (MCDIs). They asked participants to rate the degree to which the words "sound like what they mean". The age of acquisition (AoA) of words was indexed by the proportion of children using the word at 30 months according to the MCDI database. Over multiple experiments in both languages, the results showed that words rated as more iconic were acquired earlier by children.

Notably, this relationship held after controlling for the systematicity of words as measured by Monaghan et al. (2014). Systematicity is an index of the degree to which similar meanings have similar forms in the lexicon of a language. Monaghan et al. found that, from the age of 2 to $13+$ years, children tended to learn more systematic words earlier. According to Monaghan et al., in theory, iconicity and systematicity are orthogonal properties. Words can be systematic but not iconic - a point they illustrate with the English consonant cluster sl- (e.g. slime, slow, slur, slum),
which systematically refers to negative or repellent properties, but bears no clear resemblance to this meaning. Yet, it is questioned whether spoken languages afford sufficient articulatory freedom for words to be iconic but not systematic. For instance, Monaghan et al. offers the example of onomatopoeic words for the calls of small animals (e.g. peep, cheep) compared to calls of big animals (roar, grrr), which both iconically and systematically reflect the animals' size. However, Perry et al.'s (2015) finding that the iconicity-AoA relationship held after controlling for systematicity shows some limited support for the independence of these properties in English - at least for the roughly 300 words for which these measures overlapped. In further support for their independence, Winter et al. (in press) examined the relationship between iconicity ratings and Monaghan et al.'s systematicity index for 1,104 words, and found only a weak correlation of $r=0.06$.

Following the study of Perry et al. (2015), a couple of subsequent studies have found comparable results. Massaro and Perlman (2017) used the same procedure to collect new iconicity ratings for the English MCDI words. They used these to examine the relationship between iconicity and the frequency with which children used the words from 6 to 47 months of age. The study found a gradual decrease in the influence of iconicity on children's production vocabulary with increasing age. An analysis of children's receptive and productive vocabularies with respect to increasing vocabulary size showed that the average iconicity of their vocabulary declined with increasing size.

Massaro and Perlman (2017) observed that the very first words that children produced were especially high in iconicity and included a relatively high number of onomatopoeic words. This observation is consistent with some other studies of early vocabulary, which indicate that a high proportion of children's first words are onomatopoeic or mimetic words. As reported by Laing (2015), a study of children's first five words across various languages found that about $20 \%$ were onomatopoeic (Menn \& Vihman, 2011). Another study found that 3 -year-olds used more mimetics to describe motion events than 5 -year olds (Kita et al., 2010).

In addition to the high level of iconicity in children's early spoken words, Perry et al. (in press) examined whether parents use more iconic words in child-directed speech. The study included iconicity ratings for approximately 2000 English words including Perry et al.'s (2015) prior ratings. First, the study replicated the finding of a relationship between the iconicity of words and their age of acquisition, which was indexed by norms based on adults' subjective ratings (Kuperman et al., 2012). Second, the study examined how iconicity influenced the frequency of words used in children's speech, as well as in the child-directed and adultdirected speech of adults. Word frequencies in childproduced and child-directed speech were calculated from the Child Language Data Exchange System (CHILDES), and word frequencies of adult-adult speech were from the American National Corpus. The results showed that younger
children tended to use higher iconicity words more frequently, but with age, they increasingly favored lower iconicity words. Analyses of adult speech found that adults used higher-iconicity words more frequently when talking to children, but used lower iconicity words more frequently with adults. Thus, as children became older and their vocabulary grew in size, their speech became more adultlike in the use of iconic words.

The finding that parents used more iconic words in child directed speech fits with some previous observations of Japanese-speaking parents. Imai and Kita (2014) described two studies published in Japanese that found that caretakers used more mimetics and onomatopoeia with younger children, but used these words less frequently as children became more language proficient (Saji \& Imai, 2013; Suzuki, 2013). However, caution should be applied in generalizing across languages: Japanese-speaking parents have been shown to use more onomatopoeia and mimetics than English speaking parents, who nevertheless do use some onomatopoeia and especially sound effects (Fernald \& Morikawa, 1993; Yoshida, 2012).

## Current study

In the current study, we examined how the use of iconic words in early conversations between children and caregivers develops within individual children-parent dyads, while controlling for production setting and individual differences in the children's verbal and nonverbal skills. The longitudinal data include about six observations for each of 35 dyads within the same semi-structured activity across children and ages. We examine the change in the use of iconic words in children's speech in comparison to the speech of their parents, allowing us to investigate how parents adapt their level of iconicity to their children's language proficiency. We also explore how individual differences of the children (verbal and nonverbal IQ) relate to the use of iconic words by children and parents over development.

## Methods

## Participants

As part of an ongoing longitudinal study investigating language acquisition in young children with autism spectrum disorder (Goodwin, Fein, \& Naigles, 2012; Tek, Mesite, Fein, \& Naigles, 2014), we recruited 35 typically developing children ( 6 girls, mean age at onset $=20.27$ months, $95 \%$ CIs: 19.78 20.93). All children were monolingual English learners. Their average verbal IQ at visit 1 (as measured by MSEL-EL, cf. below) was 19.89 ( $95 \%$ CIs: 18.421 .76 ) and perceptual IQ at visit 1 (as measured by MSEL-VR, cf. below) 26 ( $95 \%$ CIs: 24.83 27.09).

## Iconicity, systematicity, and concreteness

Iconicity ratings were taken from two previous studies, which collected ratings for 3001 English words (Perry et al.,

2015; Winter et al., in press). Approximately 600 of the words were selected from the MCDI (Fenson et al., 1994), and additional words were chosen to maximize coverage with relevant psycholinguistic datasets of age of acquisition and concreteness norms. Complete methodological details can be found in Perry et al. Native English speakers on Amazon Mechanical Turk rated the iconicity of the words on a scale from -5 to 5 , where 5 indicated that a word was highly iconic and sounds like what it means, -5 that it sounds like the opposite of its meaning, and 0 that it is completely arbitrary. The words "slurp" and "teeny" were suggested as examples of highly iconic words, "cat" and "dog" as arbitrary words (Pinker \& Bloom, 1989), and "whale" and "microorganism" as opposite-sounding words (Hockett, 1960). Each word was rated by at least 10 participants. The average iconicity rating across all words was 0.92 ( $\mathrm{SD}=1.13$ ).

Systematicity measures were taken from Monaghan et al. (2014). This study computed systematicity for a large set of monosyllabic English words by measuring the overall correlation between the degree of similarity between the forms of any two words in the set, and the degree of similarity between their corresponding meanings. Ratings for individual words consisted of their contribution to the overall form-meaning correlation across all the words. Concreteness ratings were taken from Brysbaert et al. (2014), which were collected for 40 thousand English words and short phrases via Amazon Mechanical Turk. Words were rated 1 'abstract' to 5 'concrete'.

## Speech data

The data were collected across six home visits, each separated by 4 months. For six children, data at one visit were missing. This generated a total of 204 visits. During each visit, children engaged in a $30-\mathrm{min}$ semi-structured parent-child play session. All sessions were transcribed at the word-level. The 204 transcripts analyzed consisted of 465,474 words ( 99,210 of children's speech, CS) and 4143 unique words ( 2185 in CS). 1334 unique iconicity-rated words were found ( 899 in CS), covering $32 \%$ of all unique words used ( $41 \%$ in CS) and $78 \%$ of all used words ( $72 \%$ in CS). The systematicity coding covered 699 employed unique words ( 515 in CS): $17 \%$ of all employed unique words ( $12 \%$ in CS) and $26 \%$ of all employed words ( $24 \%$ in CS). Altogether, the iconicity, systematicity, and concreteness coding covered 698 employed unique words (514 in CS): $17 \%$ of all employed unique words ( $12 \%$ in CS) and $26 \%$ of all employed words ( $24 \%$ in CS).

## Analysis

Separately for child-produced and child-directed speech, we produced mixed effects growth curve models assessing the development of iconicity over time in 4 steps. First we used iconicity as outcome measure, visit (linear and squared) as fixed factor, and child ID as random effect, including visit as random slope. Second, we controlled whether the results were preserved when controlling for measures of word
concreteness and systematicity by adding to the first model these measures as fixed effects. Third, we controlled for effects of verbal and nonverbal initial IQ: to the first model we added MSEL-EL and MSEL-VR at visit 1, both as main effects and as interactions with visit. Due to the high colinearity between verbal and nonverbal IQ ( $\mathrm{r}=-0.712$ ), high caution should employed in interpreting the results. Finally, we tested whether significant changes of iconicity over time were due to changes in iconic or in arbitrary words. Iconic words were defined as words with an iconicity score above 1 (531 unique words), arbitrary as words with a score between -0.5 and 0.5 ( 421 unique words). We employed a mixed effects Poisson regression with number of iconic words as outcome, visit (linear and squared) as fixed factor, overall number of words produced by the speaker as offset and child ID as random factor including visit as random slope. This analysis was repeated for unique iconic words and for overall and unique arbitrary words. All analyses were run employing R 3.3.2, RStudio 1.0.136, lme4 1.1-12 and tidyverse 1.1.0 (Bates et al., 2014; RCoreTeam, 2016; RStudioTeam, 2016; Wickham, 2017).

## Results

## Children's speech

First, we analyzed whether the iconicity in children's speech changed over the six visits. We found that iconicity decreases significantly over time (linear: $b=-0.63, \mathrm{SE}=0.11$, $p<0.001$; quadratic: $b=0.43, \mathrm{SE}=0.09, p<0.001$ ), cf. Figure 1 . When controlling for systematicity and concreteness, we still observed analogous results. Iconicity decreases over time (linear: $b=-0.19, \mathrm{SE}=0.05, p<0.001$; quadratic: $b=0.14$, $\mathrm{SE}=0.04, p<0.001)$. Concreteness is a significant predictor of iconicity $(b=0.06, \mathrm{SE}=0.007, p<0.001)$, but not systematicity ( $b=0.004, \mathrm{SE}=0.007, \mathrm{p}=0.565$ ).

Next, we examined whether individual cognitive skills play a role in the iconicity of children's speech. When verbal and nonverbal IQ were added to the model, we observed analogous effects, but modulated by these individual variables. Iconicity decreases over time (linear: $b=-0.63$, $\mathrm{SE}=0.1, p<0.001$; quadratic $b=0.44, \mathrm{SE}=0.09$, $p<0.001$ ). The higher the verbal IQ, the lower the iconicity ( $b=-0.11, \mathrm{SE}=0.03, p<0.001$ ), with only a marginal interaction with visit (linear: $b=0.23, \mathrm{SE}=0.12, p=0.07$; quadratic: $b=-0.11, \mathrm{SE}=0.11, p=0.3$ ). Nonverbal IQ did not have a main effect on iconicity ( $b=-0.002, \mathrm{SE}=0.03$, $\mathrm{p}=0.95$ ), and it marginally modulated the effect of time: the higher the nonverbal IQ, the bigger the linear iconicity decrease $(b=-0.22, \mathrm{SE}=0.12, \mathrm{p}=0.07)$ and the bigger the quadratic slowdown ( $b=0.19, \mathrm{SE}=0.11, \mathrm{p}=0.08$ ).

Finally, we examined whether the decrease in iconicity over time resulted from a decrease in the use of iconic words or an increase in the use of more arbitrary words. We observed no significant change over time in the frequency of iconic words used, either overall uses or by unique tokens ( $p$ 's>0.17). However, there was a significant increase over time in the overall use of arbitrary words $(b=1.07, \mathrm{SE}=0.1$,
$p<0.001$ ), though not in unique arbitrary words used ( $b=0.15, \mathrm{SE}=0.09, p=0.08$ ), cf. Figure 2.

## Adults' speech

Next we examined how iconicity in parents' speech changed over the six visits. Similar to children's speech, we found that iconicity decreased significantly over time (linear: $b=-$ 0.05 , $\mathrm{SE}=0.03, p=0.002$; quadratic: $b=0.03, \mathrm{SE}=0.02$, $p=0.08$ ), cf. Figure 1 . When controlling for systematicity and concreteness, we still observed analogous results. Iconicity decreases over time (linear: $b=-0.07, \mathrm{SE}=0.02$, $p<0.001$; quadratic: $b=0.04, \mathrm{SE}=0.02, p<0.001$ ). Like with children's speech, concreteness is a significant predictor of iconicity ( $b=0.1, \mathrm{SE}=0.003, p<0.001$ ). However, unlike children's speech, systematicity is also a significant predictor of iconicity ( $b=-0.06, \mathrm{SE}=0.003, p<0.001$ ).

Next, we examined the role of children's individual cognitive skills in the iconicity of parents' speech. With verbal and nonverbal IQ added to the model, we observed analogous effects, but modulated by the children's individual cognitive skills. Iconicity decreases over time (linear: $b=-0.06, \mathrm{SE}=0.009, p<0.001$; quadratic: $(b=0.03$, $\mathrm{SE}=0.009, p=0.001$ ). Verbal IQ did not affect the general level of iconicity ( $b=-0.004, \mathrm{SE}=0.005, \mathrm{p}=0.44$ ), but it interacted significantly with time: the higher the verbal IQ, the stronger the linear decrease in iconicity ( $b=-0.03$, $\mathrm{SE}=0.01, p=0.02$ ) and the smaller the slowdown ( $b=0.03$, $\mathrm{SE}=0.01, p=0.004$ ). Nonverbal IQ did not seem to affect iconicity ( $p$ 's>0.4).


Figure 1 - Average iconicity across visits in children and adults' speech


Figure 2 - Frequency across visits of highly iconic and arbitrary words in children and adults' speech

Finally, different from children, we observed a significant decrease over time in the overall frequency of iconic words ( $b=-0.7, \mathrm{SE}=0.03, p=0.012$ ), and a marginal decrease in the number of unique iconic words used ( $b=-0.09, \mathrm{SE}=0.05$, $p=0.058$ ), with no significant quadratic components ( $p$ 's>0.3). In contrast, we observed no significant change over time in the frequency of arbitrary words, either overall or by number of unique tokens ( $p \prime s>0.195$ ), cf. Figure 2.

Table 1 shows words with iconicity ratings of 1.5 or higher from the 100 most frequent words used by children and parents during visits 1 and 6 . Children produced 28 high-iconicity words during visit 1 and 18 during visit 6 . Parents produced 13 such words during visit 1 and 10 during visit 6 .

Table 1. Most frequent high-iconicity words.

| Group | Visit | Words |
| :---: | :---: | :---: |
| Children | 1 | no (2.8), baby (2.2), bye (1.6), one (1.8), vroom (3.5), cup (1.5), snake (2.0), balloon (1.7), help (1.5), pop (4.1), my (1.5), block (2.4), roar (3.9), beep (4.4), boom (3.8), fall (1.5), dump (2.9), shake (2.6), three (1.6), bowl (1.5), hello (2.1), mine (1.5), star (1.6), yes (2.2), yum (2.8), crash (3.7), look (1.8), bee (1.5) |
|  | 6 | no (2.8), one (1.8), three (1.6), snake (2.0), baby (2.2), my (1.5), look (1.8), elephant (2.1), yes (2.2), vroom (3.5), balloon (1.7), beep (4.4), knock (3.1), stop (2.5), pop (4.1), home (2), hiss (4.2), off (1.9) |
| Parents | 1 | look (1.8), one (1.8), baby (2.2), no (2.8), push (2.3), bye (1.6), block (2.4), snake (2.0), work (1.7), knock (3.1), vroom (3.5), help (1.5), hello (2.1) |
|  | 6 | look (1.8), one (1.4), no (2.8), baby (2.2), my (1.5), elephant (2.1), work (1.7), help (1.5), knock (3.1), block (2.4) |

Note. Words ranked in order of frequency. Iconicity ratings in parentheses. Underlined words among the most frequent for that group during the particular visit, but not for the other group. Italicized words among the most frequent during that visit for the particular group, but not for the other visit. E.g. "yum" was used frequently by children during visit 1 , but not by parents during that visit, nor by children during visit 6 .

## Discussion \& Conclusion

Developmental psychologists have proposed that iconicity may facilitate early word learning, helping children to bridge the sounds of words with their meanings. To investigate this hypothesis, we used iconicity ratings for a large set of English words to examine iconicity in the speech of children and parents.

We found that iconicity decreases over language development in both child-produced and child-directed speech. These patterns held after controlling for concreteness and systematicity of the words. In children, the overall decrease in iconicity is driven by an increase in the use of more arbitrary words, rather than a decrease in iconic words. This contrasted to parents, who decreased their use of iconic words, but maintained the frequency of arbitrary words. Our analysis of individual verbal and nonverbal IQ showed that children's level of cognitive ability modulated their transition to the more frequent use of arbitrary words. The results also suggest that parents may adapt their
iconicity more to the children' actual cognitive skills than to their age.

These findings, along with several other studies, show a robust relationship between the iconicity of spoken words as garnered from native speaker ratings, and their prevalence in early communication between children and caregivers (Massaro \& Perlman, 2017; Perry et al., 2015; Perry et al., in press; also see Thompson et al., 2012 for similar results with British Sign Language). They support the hypothesis that iconicity plays a role in facilitating early vocabulary learning. Additionally, they highlight the possible role of iconicity in children's production. Iconicity of words may not just facilitate comprehension of their meaning, but also foster "thinking for speaking" during the beginning phases of learning to produce meaningful words (cf. Slobin, 2006). However, it is also important to note that young children and their parents clearly use a high proportion of arbitrary words too, even as they show a relatively higher inclination to use iconic words.

Our findings suggest several directions for future research on iconicity in early word learning. One important question is whether iconic words actually help children gain the referential insight, which would then ease the way for learning more arbitrary words. Alternatively, iconic words might simply be more readily acquired and put to use, which potentially could still facilitate subsequent word learning by bolstering early vocabulary. A second direction for future research is to investigate the more fine-grained temporal dynamics of iconicity in the unfolding interaction. Do parents and children adapt to each other's level of iconicity, and if so, do they do that on a turn-by-turn base, or at a more general level? Finally, future research might examine how iconic words are used with other iconic devices, such as prosody and iconic gesturing. For example, parents might modulate the prosody of their speech in iconic ways, which could help children with comprehension and word learning (Nygaard et al., 2009).

The current study adds to accumulating research showing iconicity in the lexicons and grammars of spoken and signed languages alike (Dingemanse et al., 2015; Perniss et al., 2010). This research suggests that iconicity is a fundamental property of languages - a complement to arbitrariness. Our findings show how iconicity may play an important role in children's earliest conversations, even in a spoken language like English that lacks a large inventory of widely recognized iconic words.

## Acknowledgments

We are grateful for funding from the Language Evolution and Interaction Group at the Max Planck Institute for Psycholinguistics, and from the Interacting Minds Center (Aarhus university). We thank Padraic Monaghan for sharing his data on the systematicity of English words.

## References

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., \& Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. Cortex, 63, 196-205.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version, 1(7)
Brysbaert, M., Warriner A. B., \& Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46, 904-911.
Dingemanse, M. Blasi, D. E., Lupyan, G., Christansen, M. H., \& Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. Trends in Cognitive Sciences, 19, 603-615.
Fenson, L, Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., et al. Variability in early communicative development. Monographs of Social Research in Child Development, 59, i-185.
Fernald A, Morikawa H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. Child Development. 64, 637-656.
Goodwin, A., Fein, D., \& Naigles, L. R. (2012). Comprehension of wh-questions precedes their production in typical development and autism spectrum disorders. Autism Research, 5(2), 109-123.
Hockett, C. F. (1960). The origin of speech. Scientific American, 203, 88-96.
Imai, M., Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. Philosophical Transactions of the Royal Society b, 369, 20130298.

Imai, M., Kita, S., Nagumo, M., \& Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109, 5465.

Imai, M., Miyazaki, M., Yeung, Henny, Hidaka, S, Kantartzis, K, Okada, H., and Kita, S. (2015) Sound symbolism facilitates word learning in 14-month-olds. PLOS ONE.
Kantartzis, K., Imai, M., \& Kita, S. (2011). Japanese soundsymbolism facilitates word learning in English-speaking children. Cognitive Science, 35, 575-586.
Kita S, Ozyürek A, Allen S, Ishizuka T. (2010). Early links between iconic gestures and sound symbolic words: evidence for multimodal protolanguage. In Proc. 8th Int. Conf. on Evolution of Language, Utrecht, The Netherlands, 14-17 April 2010 (eds ADM Smith, M Schouwstra, B de Boer, K Smith), pp. 429-430. Singapore: World Scientific.
Kuperman, V., Stadthagen-Gonzalez, H., \& Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, 44(4), 978-990.
Laing, C.E. (2014). A phonological analysis of onomatopoeia in early word production. First Language, 34, 387-405.
Massaro, D.W. \& Perlman, M. (2017). Quantifying iconicity's contribution during language acquisition: Implications for vocabulary learning. Frontiers in Communication, 2:4.
Maurer D., Pathman T., Mondloch C.J. 2006 The shape of boubas: sound-shape correspondences in toddlers and adults. Developmental Science, 9, 316-322.
Menn, L., \& Vihman, M. (2011). Features in child phonology: Inherent, emergent, or artefacts of analysis? In Where do phonological features come from? The nature and sources of phonological primitives (Eds. N. Clements \& R. Ridouane), pp. 261-301. Amsterdam: John Benjamins Publishing Company.
Monaghan, P., Shillcock, R.C., Christiansen, M.H., \& Kirby, S. (2014). How arbitrary is language? Philosophical Transactions of the Royal Society B, 369, 20130299.

Nygaard, L.C., Herold, D.S., \& Namy, L.L. (2009). The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning. Cognitive Science, 33, 127-146.
Ozturk O, Krehm M, Vouloumanos A. (2012) Sound symbolism in infancy: evidence for sound-shape cross-modal correspondences in 4-month-olds. Journal of Experimental Child Psychology, 114, 173-186.
Perniss P, \& Vigliocco G. (2014). The bridge of iconicity: from a world of experience to the experience of language. Philosophical Transactions of the Royal Society B, 369; 20130300.

Perniss, P., Thompson, R.L., \& Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. Frontiers in Psychology, 1.
Perry, L.K., Perlman, M., Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. PLoS ONE, 10, e0137147.
Perry, L.K., Perlman, M., Winter, B., Massaro, D.W., \& Lupyan, G. (in press). Iconicity in the speech of children and adults. Developmental Science.
Pinker, S. \& Bloom, P. (1990). Natural language and natural selection. Behavioral and Brain Sciences, 13, 707-727.
Quine, W.V.O. (1960). Word and object: An inquiry into the linguistic mechanisms of objective reference. Cambridge, MA: MIT Press.
RCoreTeam. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014.

RStudioTeam. (2016). RStudio: Integrated Development for R. RStudio. Boston, MA. Retrieved from http://www.rstudio.com/
Saji, N. \& Imai, M. (2013). Goishutoku ni okeru ruizousei no kouka no kentou. In Onomatope kenkyu no shatei- chikadzuku oto to imi (Eds. K. Shinohara \& R. Uno), pp. 151-166. Tokyo, Japan: Hituji Syobo.
Slobin, D.I. (2006). The child learns to think for speaking: Puzzles of crosslinguistic diversity in form-meaning mapping. Studies in Language Sciences, 7.
Spence, C. (2011). Crossmodal correspondences: A tutorial review. Attention, Perception, \& Psychophysics, 73, 971-995.
Stone, W.L., Coonrod, E.E., \& Ousley, O.Y. (2000). Brief report: screening tool for autism in two-year-olds (STAT): development and preliminary data. Journal of autism and developmental disorders, 30(6), 607-612.
Suzuki Y. (2013). Intarakushon no naka de tsukawareru 'onomatope fl suru' doushi. In Onomatope kenkyu no shatei chikadzuku oto to imi. (Eds. K. Shinohara \& R. Uno), pp. 167181. Tokyo: Hituji Syobo.

Tek, S., Mesite, L., Fein, D., \& Naigles, L. (2014). Longitudinal analyses of expressive language development reveal two distinct language profiles among young children with autism spectrum disorders. Journal of autism and developmental disorders, 44(1), 75-89.
Thompson, R.L., Vinson, D.P., Woll, B., Vigliocco, G. (2012). The road to language learning is iconic: evidence from British Sign Language. Psychological Science, 23, 1443-1448.
Wickham, H. (2017). Tidyverse: Easily Install and Load 'Tidyverse' Packages: R package version 1.1.0. Retrieved from https://CRAN.R-project.org/package=tidyverse
Winter, B., Perlman, M., Perry, L.K., \& Lupyan, G. (in press). Which words are most iconic? Iconicity in English sensory words? Interaction Studies.
Yoshida, H. (2012). A cross-linguistic study of sound symbolism in children's verb learning. Journal of Cognition and Development, 13, 232-265.

# Evidence for the size principle in semantic and perceptual domains 

Joshua C. Peterson (peterson.c.joshua@gmail.com)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology, University of California, Berkeley


#### Abstract

Shepard's Universal Law of Generalization offered a compelling case for the first physics-like law in cognitive science that should hold for all intelligent agents in the universe. Shepard's account is based on a rational Bayesian model of generalization, providing an answer to the question of why such a law should emerge. Extending this account to explain how humans use multiple examples to make better generalizations requires an additional assumption, called the size principle: hypotheses that pick out fewer objects should make a larger contribution to generalization. The degree to which this principle warrants similarly law-like status is far from conclusive. Typically, evaluating this principle has not been straightforward, requiring additional assumptions. We present a new method for evaluating the size principle that is more direct, and apply this method to a diverse array of datasets. Our results provide support for the broad applicability of the size principle.


Keywords: size principle; generalization; similarity; perception

## Introduction

In the seminal work of Shepard (1987), the notion of stimulus similarity was made concrete through its interpretation as stimulus generalization. It was shown that, across species (including humans), generalization probabilities follow an exponential law with respect to an internal psychological space. Specifically, the probability that $y$ is in some set $C$ that contains $x$ (what Shepard terms a "consequential subset") is an exponentially decreasing function of distance ( $d$ ) in psychological space:

$$
\begin{equation*}
s_{x y}=e^{-d(x, y)} \tag{1}
\end{equation*}
$$

Shepard termed this phenomenon the Universal Law of Generalization, in that it should apply to any intelligent agent, anywhere in the universe. This result has been used in numerous cognitive models that invoke similarity (e.g., Nosofsky, 1986; Kruschke, 1992).

In spite of this, one could argue that generalization from a single stimulus to another does not adequately describe the full scope of human behavior. Indeed, in a concept learning task, people are asked to generalize from multiple examples of a concept. To capture this, Tenenbaum and Griffiths (2001) extended Shepard's original Bayesian derivation of the law to rationally integrate information about multiple instances. The resulting model defines the probability of generalization (that $y$ is in $C$ ) as a sum of the probabilities of all hypotheses $h$ about the true set $C$ that include both $x$ and $y$,

$$
\begin{equation*}
p(y \in C \mid x)=\sum_{h: y \in h} p(h \mid x) . \tag{2}
\end{equation*}
$$

The posterior probability of each hypothesis is given by Bayes' rule,

$$
\begin{equation*}
p(h \mid x)=\frac{p(x \mid h) p(h)}{p(x)} \tag{3}
\end{equation*}
$$

The prior $p(h)$ represents the learner's knowledge about the consequential region before observing $x$. The likelihood $p(x \mid h)$ depends on our assumptions about how the process that generated $x$ relates to the set $h$. The key innovation over Shepard's model is the use of the likelihood function

$$
p(x \mid h)=\left\{\begin{array}{cc}
\frac{1}{|h|} & x \in h  \tag{4}\\
0 & \text { otherwise }
\end{array}\right.
$$

where $|h|$ is the number of objects in the set picked out by $h$. The motivation for this choice of likelihood function is the size principle, which uses the assumption of random sampling to justify the idea that smaller hypotheses should be given greater weight (see Tenenbaum and Griffiths, 2001, for a demonstration of this when $x$ represents multiple examples).

The value of the size principle lies in the fact that it allows for the benefit of multiple examples of a concept to influence generalization. Assuming samples are drawn independently, the likelihood of a hypothesis for $n$ samples is simply the likelihood of that hypothesis for a single sample to the power of $n$. From this, it can be shown that generalization tightens as the number of examples increases, consistent with human judgments (see Tenenbaum \& Griffiths, 2001).

The size principle thus plays an important role in understanding generalization, placing equal importance on determining whether it actually describes human similarity judgments in a wide range of settings. If the size principle is disconfirmed, an alternative augmentation of Shepard's model is needed to explain generalization from multiple instances. If it can be confirmed to hold broadly, it is a good candidate for a second law of generalization, or an amendment of the original law. In this paper we build on previous work evaluating the evidence for the size principle (Navarro \& Perfors, 2010), providing a novel and more direct methodology and a broader empirical evaluation that includes rich perceptual feature spaces.

## Evaluating the Size Principle

In this section we describe previous work evaluating the size principle and provide the details of our approach.

## Previous work

Navarro and Perfors (2010) made three important contributions towards understanding the scope of the size principle. First, they made explicit a link between the Bayesian model of generalization and a classic model of similarity judgment. The similarities between a set of objects can be summarized in a similarity matrix $\mathbf{S}$, where the entry in the $i$ th row and $j$ th column gives the similarity $s_{i j}$ between objects $i$ and $j$.


Figure 1: Feature Size/Weight Relationships in Semantic Dataset Group 1.

The additive clustering model (Shepard \& Arabie, 1979) decomposes such a similarity matrix into the matrix product of a feature-by-object matrix $\mathbf{F}$, its transpose, and a diagonal weight matrix $\mathbf{W}$,

$$
\begin{equation*}
\mathbf{S}=\mathbf{F W F}^{T} \tag{5}
\end{equation*}
$$

The feature matrix $\mathbf{F}$ is binary and can represent any of a broad set of structures including partitions, hierarchies, and overlapping clusters, and can either be inferred by a number of different models or generated directly by participants. The individual entries of $\mathbf{S}$ are defined as

$$
\begin{equation*}
s_{i j}=\sum_{k=1}^{N_{f}} w_{k} f_{i k} f_{j k} \tag{6}
\end{equation*}
$$

Navarro and Perfors (2010) pointed out that each feature could be taken as a single hypothesis $h$, as it likewise picks out a set of objects with a common property. Having made this link, the degree of generalization between objects $i$ and $j$ predicted by the Bayesian model can be put in the same format as Equation 6: a weighted sum of common features (a similar point was made by Tenenbaum \& Griffiths, 2001). The equivalence can be seen if we let $w_{k}$ represent the posterior $p(h \mid x)$ and $f_{k}$ be the $k^{\text {th }}$ hypothesis $\left(h_{k}\right)$, since $f_{i k} f_{j k}$ selects only the features that contain both objects. If the prior probabilities of the different hypotheses are similar, the likelihood (and the size principle) will dominate and

$$
\begin{equation*}
w_{k} \propto \frac{1}{\left|h_{k}\right|} . \tag{7}
\end{equation*}
$$

Using the link between hypotheses and features, Navarro and Perfors (2010) made their second contribution: an alternative derivation showing that the relationship predicted by the size principle can hold even in the absence of random sampling. They argued that learners encode the similarity structure of the world by learning a set of features $\mathbf{F}$ that efficiently approximate that structure. Under this view, a "coherent" feature is said to be one for which all objects that possess that feature exhibit high similarity. If a learner seeks a set of features that are high in coherence, the size principle emerges even in the absence of sampling since the variability in the distribution of similarities between objects sharing a feature is a function of $\left|h_{k}\right|$.

The third contribution that Navarro and Perfors (2010) made was to evaluate this prediction using data from the Leuven Natural Concept Database (LNCD; De Deyne et al., 2008). This database consists of human-generated feature matrices for a large number of objects, as well as pair-wise similarity ratings for those objects. Navarro and Perfors (2010) observed that, under some simplifying assumptions, the size principle predicts that the similarity between objects that share a feature will be inversely related to the size of that feature. They showed that this prediction was borne out in 11 different domains analyzed in the LNCD.

## Directly testing the size principle

The method adopted by Navarro and Perfors (2010) depends on the derived relationship between the similarity of objects that share a feature and the number of those objects. However, the link that they established between Bayesian clustering and


Figure 2: Feature Size/Weight Relationships in Semantic Dataset Group 2.
the additive clustering model (Equations 5-7) can also be used to directly test the size principle. Since both models take the same mathematical form, we can directly test the size principle by estimating the weights $w_{k}$ for a set of features $\mathbf{F}$ and verifying that Equation 7 holds. If we take the logarithm of both sides of this equation, we obtain the linear relationship

$$
\begin{equation*}
\log w_{k}=-\log \left|h_{k}\right|+c \tag{8}
\end{equation*}
$$

which can be evaluated by correlating $w_{k}$ with the number of objects that possess feature $k$. Given a feature matrix $\mathbf{F}$ and similarity matrix $\mathbf{S}$, the weights $w_{k}$ can be obtained through linear regression (Peterson, Abbott, \& Griffiths, 2016). In additive clustering the weights are often constrained to be nonnegative. To obtain such weights, we employ a non-negative least squares algorithm (Lawson \& Hanson, 1995). We can thus directly test the size principle in any domain where a feature matrix and a corresponding similarity matrix are available. In the remainder of the paper we consider two different sources of such matrices: semantic feature norms and perceptual neural networks.

## Semantic Hypothesis Spaces

We first evaluate evidence for the size principle using two groups of datasets in which people judge the similarity of words. Both datasets contain human similarity ratings of noun pairs corresponding to concrete objects (e.g., "zebra" and "lion") and lists of binary feature labels associated with each object that can be filtered by frequency of mention. We
call these features "semantic" because they are linguistic descriptions of general concepts that exclude perceptual-level information associated with actual instances of that concept or brought to mind when the instance is perceived.

## Semantic Dataset Group 1

Following Navarro and Perfors (2010), the first evaluation dataset is comprised of similarity and feature matrices from the Leuven Natural Concept Database (De Deyne et al., 2008). It includes data for 15 categories (Kitchen Utensils, Clothing, Vegetables, Professions, Fish, Sports, Birds, Fruit, Reptiles, Insects, Tools, Vehicles, Musical Instruments, Mammals, and Weapons), each containing $\sim 20-30$ exemplars. Binary feature matrices for each category contain $\sim 200-300$ unique features each. The feature descriptions are much broader than merely visually apparent features (e.g., "has wings", "eats fruit", "is attracted by shiny objects").

## Semantic Dataset Group 2

The second dataset group consisted of 17 similarity matrices from a variety of sources throughout the literature. The experimental contexts and methodologies differed considerably compared to the those in group 1 . All but one of these datasets (SIMLEX) were taken from the similarity data repository on the website of Dr. Michael Lee (http://faculty.sites.uci.edu/mdlee/similarity-data/). SIMLEX was taken from a larger word similarity dataset (Hill, Reichart, \& Korhonen, 2016). The majority of the datasets (Birds, Clothing1, Clothing2, Fish, Fruit1, Fruit2, Furniture1,

Furniture2, Tools, Vegetables1, Vegetables2, Weapons1, and Weapons2) are from Romney, Brewer, and Batchelder (1993). For dataset pairs such as (Vegetables1, Vegetables2), the first contains more prototypical items than the second. Since none of these datasets contain corresponding object-feature data, we matched objects from each set to the feature norms reported in McRae, Cree, Seidenberg, and McNorgan (2005).

## Analysis \& Results

For each dataset, we computed the element-wise multiplication of each pair of rows in $\mathbf{F}$ and used non-negative least squares to regress this matrix onto the corresponding empirical similarity values. We then computed the $\log$ of all nonzero weights, as well as the log of the feature sizes (column sums of the $\mathbf{F}$ matrix for which there was a corresponding non-zero weight). The resulting $\log$ weights and $\log$ feature sizes are z -score normalized and plotted in Figures 1 and 2 for each category in each subgroup. Red lines indicate perfect -1 slopes as predicted by the size principle, whereas black lines are best fitting lines to the actual data. The corresponding correlation coefficients are reported in Tables 1 and 2 along with a number of other statistics to be discussed.

Average Pearson and Spearman correlations were -0.43 and -0.47 , respectively for group 1 , and -0.63 and -0.61 for group2. For nearly all individual datasets in all groups, coefficients are consistently negative, with the exception of Animals11 in group 2, which along with Animals5 are the only datasets with no published method. Correlations were generally stronger for group 2. All correlations in group 1 were significant at the $\alpha=0.05$ level except for the Reptiles and Mammals datasets. In contrast, virtually no correlations were significant in group 2 given the small number of features with non-zero coefficients, however one-sample $t$-tests confirmed that the mean slopes were significantly less than 0 for both Pearson $(t(16)=-7.65, p<0.0001)$ and Spearman $(t(16)=-7.65, p<0.0001)$ correlations.
The $F R$ (feature ratio) column indicates how many coefficients were positive out of the total possible. Although there were many more features overall in group 1 , the average percentage of features with non-zero weights was comparable ( $28 \%$ and $23 \%$ respectively).
Finally, model performance in predicting similarity $\left(R^{2}\right)$ is reported in the $R_{M P}^{2}$ column, and indicates the degree to which the feature sets are sufficient to accurately predict human similarity judgments. ( $R^{2}$ ) values for group 1 are markedly higher than group 2 (which have many fewer features) and match reliability ceilings reported in the original experiments.

## Perceptual Hypothesis Spaces

While evidence for the size principle seems apparent from studies of semantic hypothesis spaces, there has been no work attempting to verify the operation of the principle for concrete objects, especially with complex, real-world instances of these objects such as natural images. The featural representations of such instances are complex and include innumerable details not contained in semantic descriptions of the
general case, rendering explicit feature descriptions difficult. Here, we offer a method to overcome this challenge by leveraging representations learned from deep neural networks.

## Perceptual Features

Recent work (Peterson et al., 2016) has provided evidence that deep image feature spaces can be used to approximate human similarity judgments for complex natural stimuli. For our analysis, we extracted image features from an augmented version of Alexnet with a binarized final hidden layer ( Wu , Lin, \& Tang, 2015). This allows for a comparison both to non-

Table 1: Correlations between feature size and feature weight (Semantic Datasets Group 1)

| Set | Pearson | Spearman | $\boldsymbol{F} \boldsymbol{R}$ | $\boldsymbol{R}_{\boldsymbol{M P}}^{2}$ |
| :--- | ---: | ---: | :---: | :---: |
| K. Utensils | -0.64 | -0.67 | $94 / 328$ | 0.84 |
| Clothing | -0.42 | -0.47 | $84 / 258$ | 0.71 |
| Vegetables | -0.41 | -0.43 | $91 / 291$ | 0.68 |
| Professions | -0.48 | -0.51 | $73 / 370$ | 0.76 |
| Fish | -0.48 | -0.49 | $43 / 156$ | 0.80 |
| Sports | -0.58 | -0.65 | $85 / 382$ | 0.81 |
| Birds | -0.36 | -0.37 | $72 / 225$ | 0.75 |
| Fruit | -0.23 | -0.37 | $78 / 233$ | 0.74 |
| Reptiles | -0.23 | -0.20 | $45 / 179$ | 0.94 |
| Insects | -0.49 | -0.52 | $52 / 214$ | 0.73 |
| Tools | -0.61 | -0.61 | $62 / 285$ | 0.74 |
| Vehicles | -0.52 | -0.57 | $97 / 322$ | 0.93 |
| M. Instruments | -0.50 | -0.56 | $72 / 218$ | 0.90 |
| Mammals | -0.19 | -0.22 | $84 / 288$ | 0.85 |
| Weapons | -0.36 | -0.38 | $49 / 181$ | 0.88 |

Table 2: Correlations between feature size and feature weight (Semantic Datasets Group 2)

| Set | Pearson | Spearman | $\boldsymbol{F} \boldsymbol{R}$ | $\boldsymbol{R}_{M P}^{2}$ |
| :--- | ---: | ---: | :---: | :---: |
| Animals11 | 0.01 | 0 | $7 / 37$ | 0.31 |
| Animals5 | -0.15 | -0.08 | $8 / 37$ | 0.35 |
| Birds | -0.94 | -0.95 | $4 / 24$ | 0.10 |
| Clothing1 | -0.68 | -0.78 | $6 / 28$ | 0.10 |
| Clothing2 | -0.42 | -0.53 | $12 / 35$ | 0.11 |
| Fish | -1.00 | -1.00 | $2 / 17$ | 0.18 |
| Fruit1 | -0.24 | -0.29 | $12 / 38$ | 0.19 |
| Fruit2 | -0.53 | -0.43 | $4 / 42$ | 0.14 |
| Fruit3 | -0.52 | -0.64 | $11 / 42$ | 0.25 |
| SIMLEX | -0.99 | -1.00 | $3 / 151$ | 0.24 |
| Tools | -0.95 | -0.87 | $5 / 13$ | 0.18 |
| Vegetables1 | -0.20 | -0.08 | $9 / 31$ | 0.31 |
| Vegetables2 | -0.46 | -0.57 | $9 / 31$ | 0.26 |
| Vehicles1 | -0.97 | -0.71 | $5 / 24$ | 0.06 |
| Vehicles2 | -0.87 | -0.82 | $5 / 23$ | 0.03 |
| Weapons1 | -0.85 | -0.87 | $8 / 32$ | 0.16 |
| Weapons2 | -0.96 | -0.74 | $4 / 30$ | 0.03 |



Figure 3: Feature Size/Weight Relationships for Convolutional Neural Network Representations.


Figure 4: Examples of stimuli from each of the 5 natural image categories
perceptual binary feature sets (i.e., features from the previous section) and non-binary perceptual feature sets (i.e., previous work on similarity prediction).

## Stimuli \& Data Collection

We obtained pairwise image similarity ratings for 5 sets of 120 images (animals, fruits, furniture, vegetables, vehicles) using Amazon Mechanical Turk, following Peterson et al. (2016). Examples of images in each dataset are given in Figure 4. The image sets represent basic level categories, with 20-40 subordinate categories in each.

Subjects rated at least 4 unique pairs of images and we required that at least 10 unique subjects rate each possible pair. Each experiment yielded a $120 \times 120$ similarity matrix.

## Analysis \& Results

As before, we computed the pairwise multiplication of each pair of rows in $\mathbf{F}$ (120 images $\times 4096$ neural features) and regressed this matrix onto the corresponding empirical similarity values. The resulting weights and feature sizes are plotted in Figure 3 for each category, and the corresponding correlations are reported in Table 3.

Like the previous semantic datasets, only a small portion of the total features obtained non-zero weights, although the average percentage was much smaller ( $\sim 4 \%$ ). Given that the

Table 3: Correlations between feature size and feature weight (Perceptual Dataset)

| Set | Pearson | Spearman | $\boldsymbol{F R}$ | $\boldsymbol{R}_{M P}^{2}$ |
| :--- | ---: | ---: | :---: | :---: |
| Animals | -0.32 | -0.34 | $122 / 4096$ | 0.56 |
| Fruits | -0.43 | -0.44 | $302 / 4096$ | 0.41 |
| Furniture | -0.48 | -0.51 | $170 / 4096$ | 0.38 |
| Vegetables | -0.41 | -0.34 | $295 / 4096$ | 0.45 |
| Automobiles | -0.46 | -0.49 | $125 / 4096$ | 0.31 |

full feature set is meant to characterize 1000 mostly qualitatively distinct categories from which they were learned (Deng et al., 2009), whereas features from the semantic datasets were relevant only to the objects in each group, this discrepancy is to be expected.

In all five datasets, correlation coefficients are moderate, negative, and significant at the $\alpha=0.001$ level. Average Pearson and Spearman correlation was 0.42 in both cases. Variance explained in similarity matrices was comparable to previous work on predicting similarity from deep features, but was generally reduced given the constraint of binary features and non-negative weights.

## Discussion

We have attempted to provide a direct evaluation of the size principle in both semantic and perceptual hypothesis spaces. In some cases, the correlations we report using our method are weaker overall than those reported in past work (Navarro \& Perfors, 2010), but are consistently negative nonetheless. If anything, this discrepancy serves as a caution to trusting a single method for evaluating the size principle.

Across all datasets, variance explained in similarity judgments ranged from .03 to .94 , however these fluctuations don't appear to vary systematically with the magnitude of the size principle effect, This may indicate that the size principle should emerge with respect to both "good" and "bad" feature sets, so long as they are related to the objects and vary in their inclusiveness.

Furthermore, it appears that the size principle can be shown to operate in more ecologically valid stimulus comparisons
such as visual image pairs. In cases such as these, the specific visual details of the image are relevant, and our feature sets derived from convolutional neural networks included only these features. There may be hundreds of small visual details that are only present in novel instantiations of familiar objects that we encounter on a daily basis and that actually represent the more abstract concepts used in semantic datasets. These results may also have implications for the method of estimating human psychological representations recently proposed by Peterson et al. (2016). In this work, it was shown that human similarity judgments for natural images can be estimated by a linear transformation of deep network features, and the current results imply that this transformation is perhaps partly accounted for by the size principle. This finding may lead to better methods for approximating complex human representations based on psychological theories.

It is apparent from the $F R$ columns of each table that few of the total features were used in the actual models. This may be due in part to useless features, or features associated with too many or too few objects. It may also be due to multicollinearity in our feature matrices (some columns are linear combinations of others). These are unique consequences of using a regression model. For this reason, our method may be less susceptible to over-representing certain features that are redundant. On the other hand, the size principle is meant to address the problem of redundant hypotheses directly, and it may be an undesirable property of our model that these hypotheses are eliminated through other means, which is perhaps the cost of direct estimation of the weights in the additive clustering framework. In any case, this variability in the amount of non-redundant features does not appear to co-vary with the size principle in any systematic way.

The only notable discrepancy between our results and the predictions of the size principle is the variation in the magnitude of the negative slopes obtained, which does not appear to depend on model performance, number of features, or even aspects of the dataset groups or individual datasets. Semantic dataset group 2 had more large slopes (e.g., SIMLEX) than group 1, but also had many small slopes. Similar datasets from group 1 (e.g., Fruit and Vegetables) had fairly dissimilar slopes, and nearly identical datasets from group 2 (e.g., Fruit1 and Fruit2) had widely varying slopes. Prototypicality doesn't seem to matter either, since Fruit1 and Vegetables1 have smaller slopes than Fruit2 and Vegetables2, but Clothing1 and Vehicles1 have larger slopes than Clothing 2 and Vehicles2. Furthermore, we can find examples of both natural and artificial stimuli with comparable slopes. For these reasons, it is unclear what the source of these deviations could be. It is possible that certain experimental contexts encourage a focus on certain featural comparisons that can be represented by a weighting of our feature sets, and so still allow for good model fit. Alternatively, it may be an artifact of the weight estimation algorithm, in which case it will be useful to compare alternative methods.

Our results provide broad evidence for the size principle
regardless of the assumptions that are employed to derive it. Thus, the size principle is a good candidate for a second universal law of generalization, and can be motivated both by rational theories based on strong sampling and feature learning. Further, a $\frac{1}{|h|}$ law can provide a solid basis for generalizing from multiple instances, a behavior that we should expect to find in any intelligent agent, anywhere in the universe.

## References

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., \& Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. Behavior research methods, 40(4), 1030-1048.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., \& Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on (pp. 248-255).

Hill, F., Reichart, R., \& Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics.
Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. Psychological review, 99(1), 22.
Lawson, C. L., \& Hanson, R. J. (1995). Solving least squares problems. SIAM.
McRae, K., Cree, G. S., Seidenberg, M. S., \& McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior research methods, 37(4), 547-559.
Navarro, D. J., \& Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. Acta Psychologica, 133(3), 256-268.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of experimental psychology: General, 115(1), 39.
Peterson, J., Abbott, J., \& Griffiths, T. (2016). Adapting deep network features to capture psychological representations. In Proceedings of the 38th annual conference of the cognitive science society. Austin, TX.
Romney, A. K., Brewer, D. D., \& Batchelder, W. H. (1993). Predicting clustering from semantic structure. Psychological Science, 4(1), 28-34.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237(4820), 1317-1323.
Shepard, R. N., \& Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. Psychological Review, 86(2), 87.
Tenenbaum, J. B., \& Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. Behavioral and brain sciences, 24(04), 629-640.
Wu, Z., Lin, D., \& Tang, X. (2015). Adjustable bounded rectifiers: Towards deep binary representations. arXiv preprint arXiv:1511.06201.

# Modeling the Ellsberg Paradox by Argument Strength 

Niki Pfeifer (niki.pfeifer@Imu.de)<br>Munich Center for Mathematical Philosophy, LMU Munich, Germany

Hanna Pankka (hanna.pankka@helsinki.fi)<br>Department of Philosophy, History, Culture and Art Studies, University of Helsinki, Finland


#### Abstract

We present a formal measure of argument strength, which combines the ideas that conclusions of strong arguments are (i) highly probable and (ii) their uncertainty is relatively precise. Likewise, arguments are weak when their conclusion probability is low or when it is highly imprecise. We show how the proposed measure provides a new model of the Ellsberg paradox. Moreover, we further substantiate the psychological plausibility of our approach by an experiment $(N=60)$. The data show that the proposed measure predicts human inferences in the original Ellsberg task and in corresponding argument strength tasks. Finally, we report qualitative data taken from structured interviews on folk psychological conceptions on what argument strength means.


Keywords: argument strength; coherence; Ellsberg paradox; probability logic

## Introduction

## Measuring Argument Strength

Probabilistic models of argumentation became popular in cognitive science and its subfields including psychology, philosophy, and computer science in recent years (see, e.g., Hahn \& Oaksford, 2006; Haenni, 2009; Zenker, 2013). Like logicbased nonmonotonic approaches for defeasible argumentation (see, e.g. Prakken \& Vreeswijk, 2002), probabilistic approaches allow for dealing with exceptions and retracting conclusions in the light of new evidence. However, in contrast to qualitative logical approaches, probability allows for managing degrees of belief in the sentences involved in common sense argumentation. Moreover, degrees of belief can be used to model the strength of arguments (Hahn \& Oaksford, 2006; Oaksford \& Hahn, 2007; Pfeifer \& Kleiter, 2006; Pfeifer, 2007, 2013b).
The concept "argument" is ambiguous. In logic, it denotes a triple consisting of a (possibly empty) premise set, a conclusion indicator, and a conclusion set. Consider, for example, the following argument, which is an instance of modus ponens:
(P1) If I take the train at five ( $T$ ), I'll be home at six $(H)$.
(P2) I take the train at five $(T)$.
(C) Therefore, I'll be home at six $(H)$.

Here, (P1) and (P2) are the premises, "Therefore" the conclusion indicator and the sentence "I'll be home at six" is the conclusion. In argumentative contexts, "argument" may also denote a premise which speaks for or against a conclusion. For example "The train conductors are on strike", can serve as an argument for concluding that it is better to take the bus. In what follows, however, we will focus on arguments in the logical sense only.

How can we measure the strength of an argument? There are at least two formal approaches to study (probabilistic) argument strength. In the first approach argument strength is based on uncertain consequence relations, i.e., by presupposing that the conclusion follows to some degree from the premises. Usually, this is modeled by a conditional probability of "the conclusion given (some combination of) the premises" of the argument (see, e.g. Hahn \& Oaksford, 2006; Oaksford \& Hahn, 2007). As pointed out by Osherson, Smith, Wilkie, López, and Shafir (1990), measures of confirmation can serve as models for argument strength (for an overview of measures of confirmation see Crupi, Tentori, \& Gonzales, 2007). Measures of confirmation and previous attempts to model argument strength by uncertain consequence relations are problematic when arguments involve conditionals, like the modus ponens above (see premise ( P 1 )): it is far from clear to give a precise meaning of conditionalizing on a combination of premises, when the premise set contains conditional events. There is ample formal and experimental evidence that uncertain conditionals are best modeled by conditional probabilities (see, e.g., Evans \& Over, 2004; Oaksford \& Chater, 2007; Over \& Cruz, in press; Pfeifer, 2014, 2013a). Therefore, conditionals should be modeled by conditional probabilities. However, this requirement would imply to measure the uncertainty of a conclusion given (some combination of) the premises. Unfortunately, satisfactory semantics of expressions like $\overbrace{C}^{\text {conclusion }} \mid \overbrace{(A \text { and }(C \mid A))}^{\text {premises }}$ do not exist yet. Such semantics would, however, be necessary to capture the underlying logical structure of the modus ponens (for an approach which deals with nested conditionals and which avoids Lewis' triviality results, see Gilio, Over, Pfeifer, \& Sanfilippo, 2017; Sanfilippo, Pfeifer, \& Gilio, 2017). Modus ponens is just a relatively simple example here: there are, of course, many other argument forms involving conditionals. The inability to deal with conditionals seems to us to be one of the main reasons, why currently no formally satisfactory measure of argument strength exists within the first approach: measures based on uncertain consequence relations do not seem to be able to deal with the logical form of the argument.

In this paper, we advocate the second approach to argument strength. It satisfies the requirement of doing justice to the logical form of arguments involving conditionals (Pfeifer, 2007, 2013b). Specifically, we define argument strength based on the following ideas: (i) keep the consequence relation deductive, (ii) assign probabilities to the premises, and
then (iii) define the measure of argument strength based on the propagated coherent lower and upper probability bounds on the conclusion (Pfeifer, 2007, 2013b). Probability propagation from the premises to the conclusion is governed by coherence based probability logic (see, e.g. Coletti \& Scozzafava, 2002; Pfeifer \& Kleiter, 2009; Gilio, Pfeifer, \& Sanfilippo, 2016). The coherence approach to probability was originated by Bruno de Finetti (de Finetti, 1970/1974). It conceives probabilities as subjective degrees of belief. Conditional probabilities $(p(C \mid A))$ are primitive. This allows for zero probabilities of the conditioning event $(A)$. Note that in standard approaches to probability, $p(C \mid A)$ is undefined if $p(A)=0$, which is problematic in many argument forms (see, e.g. Pfeifer, 2014; Gilio et al., 2016). Moreover, coherence allows for managing imprecise probabilities (set-valued probabilities involving lower and upper probability bounds), which is relevant for formalising arguments under incomplete probabilistic knowledge. The above mentioned modus ponens, for example, is formalised as follows (see, e.g. Pfeifer \& Kleiter, 2009, Example 1, p. 209):
(P1') $\quad p(H \mid T)=x$
(P2') $\quad p(T)=y$
(C') Therefore, $z^{\prime} \leq p(H) \leq z^{\prime \prime}$, where $z^{\prime}=x y$ and $z^{\prime \prime}=x y+1-y$ are the best possible coherent probability bounds on the conclusion.
Following Pfeifer (2013b), we define the measure of argument strength $\mathfrak{s}$ on an argument $\mathcal{A}$ as follows:

Let $z^{\prime}$ and $z^{\prime \prime}$ denote the coherent lower and upper probability bounds, respectively, on the conclusion of argument $\mathcal{A}$. Then,

$$
\begin{equation*}
\mathfrak{s}(\mathcal{A})={ }_{\text {def. }} \overbrace{\left(1-\left(z^{\prime \prime}-z^{\prime}\right)\right)}^{\text {precision }} \times \overbrace{\frac{z^{\prime}+z^{\prime \prime}}{2}}^{\text {location }} . \tag{1}
\end{equation*}
$$

Intuitively, measure $\mathfrak{s}$ combines the precision and the location of the coherent conclusion probability interval. Specifically, strong arguments are arguments with low imprecision of the conclusion probability (measured by the one-complement of the distance between the upper and the lower probability bounds, $\left.1-\left(z^{\prime \prime}-z^{\prime}\right)\right)$ and with conclusion probabilities close to one (measured by the mean of the lower and upper probability bound, $\left.\left(z^{\prime}+z^{\prime \prime}\right) / 2\right)$. Weak arguments are characterized by a large conclusion interval (i.e., high imprecision) or by a low-probability conclusion (i.e., the center point of the conclusion interval is close to 0 ). For a discussion of how logical validity relates to whether the degree of belief in the conclusion is constrained by the assessment of the premises see, e.g., Pfeifer and Kleiter (2009). Of course, precision and location could be modeled differently (e.g., by using the geometric or the harmonic mean instead of the arithmetic mean). Moreover, in contexts where the location is more important than the precision of the conclusion probability interval (or vice versa), adding suitable weights to formula (1) can adjust the measure for such cases. However, for the purpose of our paper it is sufficient to keep the measure as simple as possible.

Measure $\mathfrak{s}$ has a number of plausible consequences: it ranges always from zero to one (i.e., $0 \leq \mathfrak{s} \leq 1$, since $z^{\prime}$ and $z^{\prime \prime}$ are probability values, which are also in the unit interval, $[0,1]$ ). The extreme " 0 " denotes weak arguments and " 1 " denotes strong arguments. Arguments with conclusion probability 1 , are strong arguments, since $\mathfrak{s}=1$ if $z^{\prime}=z^{\prime \prime}=1$. Arguments with conclusion probability 0 (i.e., $z^{\prime}=z^{\prime \prime}=0$ ) are weak arguments, since $\mathfrak{s}=0$. Likewise, probabilistically non-informative arguments (i.e., $z^{\prime}=0$ and $z^{\prime \prime}=1$ ) are weak arguments, since $\mathfrak{s}=0$.

Interestingly, measure $\mathfrak{s}$ also provides a new solution to the Ellsberg paradox (Ellsberg, 1961), ${ }^{1}$ which we describe in the next section.

## Modeling the Ellsberg Paradox by Measure $\mathfrak{s}$

Ellsberg described the following situation (Ellsberg, 1961):
An urn contains 90 balls, of which 30 are red $(R)$ and 60 are black or yellow. The ratio of the black and yellow balls is unknown-there might be anything between 0 to 60 black (or yellow) balls. One ball is drawn at random from the urn and you are asked to choose a bet between two bets. If you take Bet $\mathbf{1}$, you will win $\$ 100$, if the ball drawn from the urn is red. If you take Bet 2, you will win $\$ 100$, if the ball drawn from the urn is black.

Ellsberg predicted that most people choose Bet $\mathbf{1}$ when asked to decide which of the two bets they prefer. Then, considering again the same urn, Ellsberg predicted that people will choose Bet 4, when they are asked to decide between the following two alternative bets:

If you take Bet 3, you will win $\$ 100$, if the ball drawn from the urn is red or yellow. If you take Bet 4, you will win $\$ 100$, if the ball drawn from the urn is black or yellow.

Ellsberg's predictions create a well-known paradox as they violate the independence axiom of rational choice (see, e.g., Briggs, 2016). Moreover, Ellsberg's predictions were experimentally confirmed in many studies (see, e.g., Becker \& Brownson, 1964; Slovic \& Tversky, 1974; MacCrimmon \& Larsson, 1979).
We propose to frame the Ellsberg paradox in terms of probability logical arguments. Specifically, the premises represent the probabilistic information given in the description of the urn, and the conclusions represent the respective bets involved in the Ellsberg paradox. Thus, we obtain four arguments. Each argument speaks for choosing the corresponding bet. The associated argument to Bet 2, for example, is argument $\mathcal{A}_{2}$ (where " $\vee$ " denotes disjunction ("or") and $R, B$ and $Y$ are mutually exclusive):
$p(R)=.33$
$p(B \vee Y)=.67$
Therefore, $0 \leq p(B) \leq .67$ is coherent.

[^171]The strength of this argument is denoted by $\mathfrak{s}\left(\mathcal{A}_{2}\right)$ and by applying equation (1) equal to . 11 (i.e., $\mathfrak{s}\left(\mathcal{A}_{2}\right)=.11$ ). Table 1 lists the conclusions and the argument strengths $\mathfrak{s}$ for each argument for the corresponding four bets involved in the Ellsberg paradox.

Table 1: Conclusions and normative strengths (s) of Arguments $\mathcal{A}_{1}, \ldots, \mathscr{A}_{4}$ associated with the four bets involved in the Ellsberg paradox. The premises are always $p(R)=.33$ and $p(B \vee Y)=.67$.

|  | Conclusion | Argument strength |
| :--- | :---: | :---: |
| Bet 1 | $p(R)=.33$ | $\mathfrak{s}\left(\mathcal{A}_{1}\right)=.33$ |
| Bet 2 | $0 \leq p(B) \leq .67$ | $\mathfrak{s}\left(\mathcal{A}_{2}\right)=.11$ |
| Bet 3 | $.33 \leq p(R \vee Y) \leq 1$ | $\mathfrak{s}\left(\mathcal{A}_{3}\right)=.22$ |
| Bet 4 | $p(B \vee Y)=.67$ | $\mathfrak{s}\left(\mathcal{A}_{4}\right)=.67$ |

The four argument strength values in Table 1 induce the following preference orders in the classical Ellsberg task: Bet $1 \succ$ Bet 2, since $\mathfrak{s}\left(\mathscr{A}_{1}\right)=.33>\mathfrak{s}\left(\mathscr{A}_{2}\right)=.11$, and Bet $4 \succ$ Bet 3, since $\mathfrak{s}\left(\mathscr{A}_{4}\right)=.67>\mathfrak{s}\left(\mathcal{A}_{3}\right)=.22$ (where $X \succ Y$ denotes $X$ is preferred over $Y$ ). This preference order corresponds to Ellsberg's predictions and matches the data (see, e.g., Becker \& Brownson, 1964; Slovic \& Tversky, 1974; MacCrimmon \& Larsson, 1979). The functions of the four arguments can be understood in an epistemic and in a persuasive sense. The epistemic function of the arguments is to gain knowledge about which bet should be preferred. The persuasive function of the arguments is to convince someone which bet should be preferred.

In the following section we further investigate the psychological plausibility of $\mathfrak{s}$ by an experiment.

## Method

## Participants

In this experiment 60 university students (mean age 25.9 years ( $S D=5.6$ ), 48 females, 12 males) participated for a compensation of $15 €$. All of the participants were Finnish native speakers and none of them had studied psychology, mathematics, statistics or philosophy as their major.

## Design and Materials

We used three target task types: argument ranking tasks, argument rating tasks, and the (original) Ellsberg tasks. The argument ranking tasks first instructed the participants to rank the strength of arguments $\mathcal{A}_{1}$ and $\mathcal{A}_{2}$ (see Table 1). Second, the participants were instructed to rank the strength of arguments $\mathcal{A}_{3}$ and $\mathcal{A}_{4}$. The argument rating tasks instructed the participants to rate the strength of each of the four arguments. In the original version of the Ellsberg task, participants had to rank which bets they preferred as described in the Introduction. We investigated the following questions which relate argument strength to the Ellsberg problem:

- Do the results of the argument strength rating tasks predict the responses in the Ellsberg tasks?
- Do the results of the argument strength rating tasks predict the responses in the argument strength ranking tasks?

Moreover, we explored empirically, whether argument strength formulated in epistemic or in persuasive terms impacts participants' reasoning. Finally, we systematically manipulated the information conveyed in the argument rating and in the argument ranking tasks by the following independent variables: (i) only the uncertainty of the conclusion was presented, (ii) only the uncertainties of the premises were presented, and (iii) uncertainties of the premises and the conclusion were presented. The instructions introduced the following symbol for marking not conveyed information in the respective conditions which correspond the variables (i) and (ii): $\mathbb{4}$. By using a $2 \times 3$ between-participant design we fully crossed epistemic versus persuasive formulations and the manipulated information conveyed in the arguments. In the epistemic booklets we used knowledge-oriented phrasings like "Which argument is stronger to know which bet to choose?", whereas in the persuasive booklets we used according phrasings like "Which argument convinces stronger which bet to choose?". The experimental conditions are explained in Table 2.

Table 2: Experimental conditions ( $\mathrm{Cd} 1-\mathrm{Cd} 6 ; N=60$ ).

| Presented probabilities | Epistemic | Persuasive |
| :--- | :--- | :--- |
| Premise \& conclusion | $\operatorname{Cd} 1\left(n_{1}=10\right)$ | $\operatorname{Cd} 2\left(n_{2}=10\right)$ |
| Conclusion only | $\operatorname{Cd} 3\left(n_{3}=10\right)$ | $\operatorname{Cd} 4\left(n_{4}=10\right)$ |
| Premise only | $\operatorname{Cd} 5\left(n_{5}=10\right)$ | $\operatorname{Cd} 6\left(n_{6}=10\right)$ |

Argument ranking tasks In these tasks, the participants were instructed to imagine two friends arguing about which bet the participant should choose. Then, argument $\mathcal{A}_{1}$ for Bet 1, and argument $\mathcal{A}_{2}$ for Bet 2 were presented to the participant, e.g.:

## Argument 2 for Bet 2

I am $\times \%$ sure that the ball drawn from the urn is red.
I am $\times \%$ sure that the ball drawn from the urn is black or yellow.
Therefore, I am at least $0 \%$ and at most $67 \%$ sure that the ball drawn from the urn is black

The participants were then presented with the question "Which argument is stronger to know which bet to choose?" (Kumpi argumentti on vahvempi sen tietämiseen, kumpi veto kannattaisi valita?) in the epistemic condition. In the persuasive condition, they were asked "Which argument convinces you stronger which bet to choose?" (Kumpi argumentti vakuuttaa sinut vahvemmin siitü, kumpi veto kannattaisi valita?). Then, the participants were instructed to indicate
their choice by ticking the respective box for Argument 1 (i.e., $\mathcal{A}_{1}$ ) or Argument 2 (i.e., $\mathcal{A}_{2}$ ). Finally, the participants ranked Argument 3 (i.e., $\mathcal{A}_{3}$ ) and Argument 4 (i.e., $\mathcal{A}_{4}$ ).

Argument rating tasks In these tasks participants were presented with the same four arguments as in the argument ranking tasks. They were asked to carefully reconsider each. Instead of using forced choice response formats, each argument was followed by a question, e.g., "How strong is Argument 2 for choosing Bet 2?" (Kuinka vahva Argumentti 2 on Vedon 2 valitsemiseksi?; original epistemic formulation) or "How strong is Argument 2 for convincing to choose Bet 2?" (Kuinka vahva Argumentti 2 on vakuuttamaan Vedon 2 valitsemisesta?; original persuasive formulation). The participants were asked to mark their responses on a ten point rating scale (see Figure 1).


Figure 1: Answer scale used in the argument rating tasks.

Ellsberg tasks Here, as explained in the introduction, the participants had to choose which rankings among bets they preferred (Bet 1 or Bet $\mathbf{2}$ and Bet $\mathbf{3}$ or Bet 4). All participants were presented with the same Ellsberg tasks.

## Procedures

Participants completed the booklets individually in a quiet room. At the beginning of the testing, participants were informed to take as much time as needed for completing the tasks. Furthermore, they were instructed not to look back on their previous responses. After reading the introduction the participants worked on tasks which differed from the Ellsberg problem (and which are not in the scope of the present paper). After that, the target tasks were presented in the following order: (i) argument ranking tasks, (ii) argument rating tasks, and (iii) the Ellsberg tasks. The easier argument ranking tasks (rankings require less cognitive effort than ratings) appeared prior to the argument rating tasks to further help participants to familiarize themselves with the task materials. To avoid any influences of the Ellsberg tasks on the argument strength tasks and to see whether our samples replicate the findings in the literature, the Ellsberg tasks were presented at the end of the booklet. Then, the participants filled in demographic data and rated the difficulty and clearness of the tasks. Participants used 9.6 minutes $(S D=2.8)$ on the average to work on the booklets. Each session concluded by an interview to further explore argument strength from a qualitative point of view: we asked how the participants solved

Table 3: Percentages of argument preferences in the argument ranking tasks $(\operatorname{rnk}(\mathcal{A}))$ and in the Ellsberg tasks $(N=60)$.

| \% | $\operatorname{rnk}(\mathcal{A})$ | Ellsberg | \% | $\operatorname{rnk}(\mathcal{A})$ | Ellsberg |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Bet1 | 73,3 | 93,3 | Bet3 | 25,0 | 23,3 |
| Bet2 | 26,7 | 6,7 | Bet4 | 75,0 | 76,7 |

the tasks and what they thought determined the strength of an argument.

## Results and Discussion

We performed Fisher's exact tests to compare the impact of the different booklets on the response frequencies in the argument ranking tasks and in the Ellsberg tasks. Moreover, we tested influences of the different conditions in the argument rating tasks by analyses of variance. After performing Holm-Bonferroni corrections we did not observe any significant differences. We therefore pooled the data for further analysis $(N=60)$.

Ellsberg's predictions The majority of responses in all three types of tasks (i.e., argument ranking, argument rating and Ellsberg task) are consistent with Ellsberg's predictions. Our findings also replicate empirical findings reported in the literature (see, e.g., Becker \& Brownson, 1964; Slovic \& Tversky, 1974; MacCrimmon \& Larsson, 1979). Moreover, our data suggest that classical findings in Ellsberg tasks carry over to (isomorphic) problems formulated in terms of argument strength.

Table 3 shows how the participants ranked the arguments in the argument ranking tasks and how they ranked the bets in the Ellsberg tasks. Bet 1 (resp., argument $\mathcal{A}_{1}$ supporting Bet 1 ) is more frequently chosen than Bet 2 (resp., $\mathcal{A}_{2}$ supporting Bet 2). Likewise, Bet 4 (resp., argument $\mathscr{A}_{4}$ supporting Bet 4 ) is more frequently chosen than Bet 3 (resp., $\mathcal{A}_{3}$ supporting Bet 3).

Moreover, we constructed the underlying preference orders of the argument strengths and the bets from the participants' responses in all the three task types. This allows one to see which choice strategies were most commonly used. In all tasks, strategies consistent with the independence axioms of rational choice were less frequently preferred, as can be seen in Table 4. For constructing the preference orders based on the responses in the argument strength ratings tasks, we made the following assumption: if the strength of an argument $\mathcal{A}_{x}$ was rated higher than the strength of an argument $\mathcal{A}_{y}$, then the corresponding Bet $x$ is preferred over Bet $y$. Again, our findings replicate the predictions of Ellsberg and the previous empirical findings (see, e.g., Becker \& Brownson, 1964; Slovic \& Tversky, 1974; MacCrimmon \& Larsson, 1979).

Table 5 shows the mean argument strength rating responses. As predicted by measure $\mathfrak{s}$, the mean argument strength ratings reflect the Ellsberg predictions, i.e.,

Table 4: Percentages of responses consistent with Ellsberg's predictions $(E)$, the independence axiom of rational choice $(I)$. The preference order $R$ can be interpreted as a reversed version of $E$. " $(x, y) \succ(u, v)$ " means "arguments (resp. bets) $x$ and $y$ are preferred over arguments (resp. bets) $u$ and $v$ ". Preference order responses consistent with $\mathfrak{s}$ are in bold.

| Preference | Tasks $(N=60)$ |  |  |
| :---: | :---: | :---: | :---: |
| Order | $\mathcal{A}$ Ranking | Ellsberg | $\mathcal{A}$ Rating |
| $(1,4) \succ(2,3)^{E}$ | $\mathbf{5 6 . 6 7}$ | $\mathbf{7 1 . 6 7}$ | $\mathbf{5 6 . 1 0}$ |
| $(2,3) \succ(1,4)^{R}$ | 8.33 | 1.67 | 4.88 |
| $(1,3) \succ(2,4)^{I}$ | 16.67 | 21.67 | 21.95 |
| $(2,4) \succ(1,3)^{I}$ | 18.33 | 5.00 | 17.07 |

mean rating $\left(\mathcal{A}_{1}\right)>$ mean rating $\left(\mathcal{A}_{2}\right)$ and mean rating $\left(\mathcal{A}_{4}\right)>$ mean rating $\left(\mathcal{A}_{3}\right)$.

Table 5: Means and standard deviations (SD) of the argument strength ratings on a scale from 0 ("extremely weak") to 10 ("extremely strong"; $N=60$ ).

|  | $\mathcal{A}_{1}$ | $\mathcal{A}_{2}$ | $\mathcal{A}_{3}$ | $\mathcal{A}_{4}$ |
| :---: | :---: | :---: | :---: | :---: |
| Mean | 5,20 | 3,98 | 5,77 | 6,95 |
| $S D$ | 2,64 | 2,58 | 1,74 | 1,87 |

Consistency among the data Based on the argument strength ratings, we predicted the participants' choices in the ranking and in the Ellsberg tasks. The data support our predictions: the argument strength rating responses predict the ranking responses in the Ellsberg tasks. The rating responses also predict the responses in the argument strength ranking tasks (see Table 6 and Table 7).

As some participants had rated the arguments for the bets equally strong, no predictions could be derived in these cases. When taking into account only those cases, in which making predictions was possible, the responses of roughly $3 / 4$ of the participants were consistent with their responses in the ranking tasks. In the argument strength ranking tasks, $77.3 \%$ of the participants chose as predicted between the first two bets and $75.0 \%$ chose as predicted between the second two bets. For the Ellsberg tasks, we observed similarly high percentages (i.e., $75.0 \%$ and $70.8 \%$ of the participants, for the first and the second bet rankings, respectively). This is again experimental support for the psychological plausibility of measure $\mathfrak{s}$.

Finally, we discuss qualitative data taken from structured interviews on folk psychological conceptions on what argument strength means.

Interview results After the participants completed the paper and pencil tasks, we collected folk psychological conceptions on what "argument strength" (argumentin vahvuus)

Table 6: Predictions of bet rankings in Ellsberg tasks based on responses in the argument strength rating tasks $(N=60)$.

|  | Ranking |  |
| :--- | :---: | :---: |
| $\%$ | Bet 1 vs. Bet 2 | Bet 3 vs. Bet 4 |
| Chose as predicted | 55.00 | 56.67 |
| Did not choose as predicted | 18.33 | 23.33 |
| No prediction made | 26.67 | 20.00 |

Table 7: Predictions of argument strength rankings based on the responses in argument strength rating tasks $(N=60)$.

| $\%$ | Ranking |  |
| :--- | :---: | :---: |
|  | $\mathcal{A}_{1}$ vs. $\mathcal{A}_{2}$ | $\mathcal{A}_{3}$ vs. $\mathcal{A}_{4}$ |
| Chose as predicted | 56.67 | 60.00 |
| Did not choose as predicted | 16.67 | 20.00 |
| No prediction made | 26.67 | 20.00 |

means by structured interviews. We asked the participants how they would define argument strength in their own words. Participants who had received the persuasive booklets, we hypothesized, mentioned persuasive aspects (like how convincing arguments are) more frequently than those of the epistemic condition. Moreover, participants who had received the epistemic booklets focused more on epistemic aspects (like truth and knowledge) than those of the persuasive condition. However, the interview responses do not confirm these hypotheses.

The responses to the interview question concerning the meaning of "argument strength" reflected features of our measure $\mathfrak{s}$. Specifically, the location of the coherent conclusion probability interval was referred to by almost all of the participants. For many participants the location seemed to be more important than the precision of the coherent conclusion probability interval. They had, for example, focused solely on the lower probability bound of the interval and ignored the upper bound or responded based on the mean value of the interval.

However, a few participants also referred to the precision of the coherent conclusion probability interval by sentences like:
"The size of this gap between 33 [\%] and 100 [\%] is so big that it increases the uncertainty." (Epävarmuиtta lisää se, että väli 33:n ja 100:n välillä on niin suuri)

Some participants also talked about the truth or correctness of the probability bounds of the conclusion. For them, the arguments were strong, when the probabilities in the conclusions were correct, almost regardless of the values in them.

The interview responses provide folk psychological evidence for using location and precision of conclusion probability intervals for evaluating the strength of uncertain argu-
ments. Location and precision are the key ingredients of our argument strength measure $\mathfrak{s}$.

Finally, we note that Bet $\mathbf{1}$ is usually not compared directly against Bet 3 in the traditional Ellsberg task. The corresponding $\mathfrak{s}\left(\mathscr{A}_{1}\right)$ is a bit higher compared to $\mathfrak{s}\left(\mathcal{A}_{3}\right)$ : epistemically, this makes sense since the conclusion of $\mathcal{A}_{3}$ is highly imprecise while the conclusion of $\mathcal{A}_{1}$ is perfectly precise (see Table 1 above). However, it seems plausible to assume that people would prefer Bet $\mathbf{3}$ over Bet 1. To accommodate $\mathfrak{s}$ for this prediction, one could reduce the impact of the precision by adding suitable weights to the definition of $\mathfrak{s}$.

## Concluding Remarks

Based on the location and the precision of the conclusion's probability interval, we proposed a formal measure of argument strength $\mathfrak{s}$ and showed how $\mathfrak{s}$ predicts responses in Ellsberg tasks. Specifically, we framed choices among bets in terms of probability logical argument forms. Our data support the hypothesis that Ellsbergs predictions can be justified by argument strength rankings and argument strength ratings.

Since the proposed measure exploits tools available in coherence-based probability logic and since it is based on a deductive consequence relation, it allows for dealing with arguments involving conditionals. The proposed measure has many plausible consequences, which calls for future formalnormative and experimental research for modeling also other argument types, like the conditional syllogisms.

Understanding argument strength is important for theories about reasoning and argumentation in general. Our paper sheds formal and experimental light on what argument strength can mean.

## Acknowledgments DFG project PF740/2-2 (SPP1516).

## References

Becker, S. W., \& Brownson, F. O. (1964). What price ambiguity? or the role of ambiguity in decision-making. Journal of Political Economy, 72(1), 62-73.
Briggs, R. (2016). Normative theories of rational choice: Expected utility. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Winter 2016 ed.). http://tinyurl.com/hjwmajw.
Coletti, G., \& Scozzafava, R. (2002). Probabilistic logic in a coherent setting. Dordrecht: Kluwer.
Crupi, V., Tentori, K., \& Gonzales, M. (2007). On Bayesian measures of confirmation. Philosophy of Science, 74, 229252.
de Finetti, B. (1970/1974). Theory of probability (Vols. 1, 2). Chichester: John Wiley \& Sons.
Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. The Quarterly Journal of Economics, 75(4), 643-669.
Evans, J. St. B. T., \& Over, D. E. (2004). If. Oxford: OUP.
Gilio, A., Over, D. E., Pfeifer, N., \& Sanfilippo, G. (2017). Centering and compound conditionals under coherence. In
M. B. Ferraro et al. (Eds.), Soft methods for data science (pp. 253-260). Berlin, Heidelberg: Springer.
Gilio, A., Pfeifer, N., \& Sanfilippo, G. (2016). Transitivity in coherence-based probability logic. Journal of Applied Logic, 14, 46-64.
Haenni, R. (2009). Probabilistic argumentation. Journal of Applied Logic, 155-176.
Hahn, U., \& Oaksford, M. (2006). A normative theory of argument strength. Informal Logic, 26, 1-22.
MacCrimmon, K. R., \& Larsson, S. (1979). Utility theory: Axioms versus 'paradoxes'. In M. Allais \& O. Hagen (Eds.), Expected utility and the Allais paradox (Vol. 1979, pp. 333-409). Dordrecht: Reidel.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. Oxford: Oxford University Press.
Oaksford, M., \& Hahn, U. (2007). Induction, deduction, and argument strength in human reasoning and argumentation. In A. Feeney \& E. Heit (Eds.), Inductive reasoning (pp. 269-301). Cambridge: Cambridge University Press.
Osherson, D. N., Smith, E. E., Wilkie, O., López, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97(2), 185-200.
Over, D. E., \& Cruz, N. (in press). Probabilistic accounts of conditional reasoning. In L. Macchi, M. Bagassi, \& R. Vialem (Eds.), International handbook of thinking and reasoning. Hove Sussex: Psychology Press.
Pfeifer, N. (2007). Rational argumentation under uncertainty. In G. Kreuzbauer, N. Gratzl, \& E. Hiebl (Eds.), Persuasion und Wissenschaft (pp. 181-191). Wien: Lit Verlag.
Pfeifer, N. (2013a). The new psychology of reasoning: A mental probability logical perspective. Thinking \& Reasoning, 19(3-4), 329-345.
Pfeifer, N. (2013b). On argument strength. In F. Zenker (Ed.), Bayesian argumentation. The practical side of probability (pp. 185-193). Dordrecht: Synthese Library (Springer).
Pfeifer, N. (2014). Reasoning about uncertain conditionals. Studia Logica, 102(4), 849-866.
Pfeifer, N., \& Kleiter, G. D. (2006). Inference in conditional probability logic. Kybernetika, 42, 391-404.
Pfeifer, N., \& Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. Journal of Applied Logic, 7(2), 206-217.
Prakken, H., \& Vreeswijk, G. (2002). Logic for defeasible argumentation. In D. M. Gabbay \& F. Guenthner (Eds.), Handbook of philosophical logic (2nd ed., Vol. 4, pp. 219318). Dordrecht: Kluwer.

Sanfilippo, G., Pfeifer, N., \& Gilio, A. (2017). A generalized probabilistic version of modus ponens. https://arxiv.org/abs/1705.00385.
Slovic, P., \& Tversky, A. (1974). Who accepts Savage's axiom? Behavioral Science, 19(6), 368-373.
Zenker, F. (Ed.). (2013). Bayesian argumentation: The practical side of probability. Dordrecht: Synthese Library (Springer).

# Causation and norms of proper functioning: Counterfactuals are (still) relevant 

Jonathan Phillips (phillips01@g.harvard.edu) \& Jonathan Kominksy (jkominsky@g.harvard.edu)<br>Department of Psychology, Harvard University, 33 Kirkland St. Cambridge, MA, 02138


#### Abstract

Causal judgments are well-known to be sensitive to violations of both prescriptive moral and descriptive statistical norms. There is ongoing discussion as to whether both effects are best explained through changes in the relevance of counterfactual possibilities, or if moral norm violations should be independently explained through a potential polysemy whereby 'cause' may simply mean 'is morally responsible for'. In support of the latter view, recent work has pointed out that moral norm violations affect judgments of agents, but not inanimate objects, and that these effects are moderated by agents' knowledge states. We advance this debate by demonstrating that judgments of counterfactual relevance exhibit precisely the same patterns, and that judgments of inanimate objects are actually highly sensitive to whether the object violated a prescriptive norm by malfunctioning. The latter finding is difficult to account for through polysemy, but is predicted by changes in the relevance of counterfactual alternatives. Finally, we show that direct (non-moral) interventions on the the relevance of counterfactual alternatives affect causal judgments in precisely the same way as functional and moral norm violations.


Keywords: causation; norms; counterfactuals; morality; teleology

## Introduction

A central question in research on causal cognition concerns the role of norms. It is well-known that both descriptive norms (e.g., the probability of an event occurring) and prescriptive norms (e.g., the morality of an event occurring) influence judgments of actual causation, that is, a judgment that some event, $e$, was the cause of some outcome, $o$ (Alicke, 2000; Gerstenberg \& Tenenbaum, in press; Hitchcock \& Knobe, 2009; Kominsky, Phillips, Gerstenberg, Lagnado, \& Knobe, 2015). Specifically, people are more inclined to judge that $e$ was the cause of $o$ if $e$ was either very unlikely to happen or morally prohibited. Despite widespread agreement on the existence of the phenomenon, there has been little corresponding agreement on how these effects should be explained.

Most researchers take the impact of descriptive statistical norms to reveal part of the basic underlying processes that support causal reasoning (e.g., Gerstenberg \& Tenenbaum, in press; Icard, Kominsky, \& Knobe, in press; Samland \& Waldmann, 2016). They differ, however, in whether they treat the impact of prescriptive moral norms similarly, or argue that it arises from a fundamentally different set of processes.

On one side, researchers have argued that the impact of both descriptive and prescriptive norms is best explained by changes in the relevance of counterfactual possibilities. These accounts propose that when a norm violation occurs, it increases the relevance of counterfactual alternatives wherein the norm violations are replaced by norm-conforming events (e.g., Halpern \& Hitchcock, 2015; Kominsky et al., 2015; Bello, 2016). In support of this account, recent work demonstrated that norm violations affect explicit assessments of
counterfactual relevance in precisely the same way that they affect causal judgments (Phillips, Luguri, \& Knobe, 2015).

On the other side, other researchers have argued for separate explanations of the two effects. The most recent approach has suggested that the term 'cause' is polysemous: It can be used to talk about whether an agent is morally responsible for an outcome, or it can be used to talk about whether some event causally contributed to an outcome (Samland \& Waldmann, 2016). On this approach, the impact of violations of moral norms can instead be accounted for by arguing that participants are more likely to interpret the word 'cause' as being about moral responsibility in cases where moral norms have been violated.

Advancing this debate, Samland and Waldmann (2016) (S\&W hereafter) reported two important new data points: First, the violation of moral norms selectively influences causal judgments about whether agents caused an outcome, but not causal judgments of whether inanimate objects used by the same agents caused the outcome. Second, factors that affect the moral responsibility of the norm violator (such as their knowledge states) also affect causal judgments (see also Samland, Josephs, Waldmann, \& Rakoczy, 2016). S\&W suggest that these findings are best accounted for by assuming that participants were interpreting the causal question to be about moral responsibility when asked about an agent, but about simple causal contribution when asked about an object.

In arguing that these results provide evidence against a unified counterfactual account, S\&W rely on the assumption that when a norm violation occurs, people consider counterfactual alternatives to the event in its entirety. That is, they consider a counterfactual alternative that involves both the agent who violated a norm and the inanimate object used by that agent. If this assumption is correct, then a polysemy account seems to better capture S\&W's results, since a unified counterfactual account would predict that causal judgments of the agent and the inanimate object would both be affected.

At the same time, though, it is possible that the counterfactual alternatives people represent are more granular. That is, when a moral norm violation occurs, people may consider a counterfactual alternative that involves the norm-violating agent, but not the inanimate object used by that agent. If this turns out to be correct, then S\&W's findings should be understood as perfectly compatible with a unified counterfactual explanation, as this accounts would then predict that causal judgments of the agent, but not the object, would be affected.

To distinguish these possibilities, we begin by asking whether the effects uncovered by S\&W also arise in participants' assessments of which counterfactuals are relevant. One possibility is that, because moral norms apply to agents but not inanimate objects, participants will regard counter-
factual alternatives to what the agent did as relevant, but not alternatives to what the inanimate object did. If so, it would suggest that they are represented somewhat independently of one another. Furthermore, changes to agents' mental states may affect both the agent's moral responsibility and similarly whether it is relevant to consider counterfactual alternatives to their actions, which may help explain why changes to agents' mental states affect causal judgments (see, e.g., Lombrozo, 2010 on how intentions affect causal judgments in doubleprevention scenarios).

## Experiment 1

## Methods

Participants. 610 participants $\left(M_{\text {age }}=37.28, S D_{\text {age }}=\right.$ 12.14; 338 females, 1 unreported) from Amazon Mechanical Turk participated for a $\$ 0.25$ compensation. Participant recruitment was automated through TurkPrime (www.turkprime.com) to prevent repeat participation and limit recruitment to participants with a high approval rating.

Stimuli and procedure. This experiment was nearly identical to S\&W's Experiment 4, but with an additional DV. The overall design was 4 (norm condition) x 3 (question) and administered fully between-subjects. Participants read one of four vignettes (see Supplementary Materials available at https://github.com/phillipsjs/stillRelevant). In all conditions, Tom owns a garden and has two gardeners, Alex and Benni, who each take care of $1 / 3$ of the plants on their own, and jointly tend to the remaining $1 / 3$. Additionally, Alex and Benni always use two fertilizers "A-X200®" and "B-Y33®". Tom reads that fertilizers are good for plants, but using more than one kind of fertilizer could damage his plants, so Tom decides he wants both gardeners to use only fertilizer A-X200. In all cases, however, Alex applies fertilizer A-X200 and Benni applies fertilizer B-Y33, and the plants cared for by both of them are damaged.

The four conditions varied the reason that Benni used BY33. In the Standard norm-violation condition, Benni simply decides to use B-Y33; in the Unintended norm-violation condition, Benni believed he was applying A-X200, but accidentally applied B-Y33; in the Ignorant norm-violation condition, Tom neglects to tell Benni to use only A-X200, and he uses B-Y33 instead; and in the Deceived norm-violation condition, Alex deliberately lies to Benni about which fertilizer he is supposed to use to get him in trouble. We additionally varied the focus of the questions. Participants were either asked questions that focused on the two agents ("Alex" and "Benni"), the two actions ("the application of fertilizer by Alex" and "the application of fertilizer by Benni"), or the two chemicals ("the application of chemical A-X200" and "the application of chemical B-Y33").

After reading the vignette, participants were asked whether it was relevant to consider counterfactual alternatives to some aspect of the event, following Phillips et al., (2015). For example, in the Agent condition, participants indicated both whether they thought it was relevant or irrelevant to consider
what Alex could have done differently and also whether it was relevant or irrelevant to consider what Benni could have done differently. Subsequently, as in S\&W, participants were asked to judge who or what caused the plants to dry up (again depending on the Question condition). In the Agent condition, participants indicated both whether they thought Alex was a cause and also whether they thought Benni was a cause. Because the causal question was simply a replication of S\&W (who did not include a counterfactual question), the counterfactual question was always presented first and mirrored S\&W's causal question as closely as possible.

Following these question, participants received two check questions that tested their understanding of which chemicals were applied by which gardener, and which chemicals Tom wanted each gardener to use. Following S\&W, they were also asked to estimate the proportion of the flowers that dried when (1) only fertilizer A-X200 was applied, (2) only fertilizer B-Y33 was applied, and (3) both were applied.

## Results

We excluded participants who did not answer both of the check questions correctly, and analyzed the remaining 439 participants' judgments. (Note that here and throughout the following experiments, all of the key results remain when these exclusion criteria are relaxed.) To examine the effects of our manipulation on both causation and relevance judgments, we categorized participants' responses as assigning causal responsibility (or counterfactual relevance) to (1) only the norm-violating agent, (2) both agents, or (3) only the normconforming agent, and then subjected both kinds of judgments to a proportional odds logistic regression using the probit function in the MASS package in R. For causal judgments, we observed an effect of the norm-condition $(L R T=20.49$ $[d f=3], p<.001)$, an effect of question $(L R T=44.53$ $[d f=2], p<.001)$, and critically, a norm-condition $\times$ question interaction effect $(L R T=19.94[d f=6], p=.003)$. This precisely replicates the pattern of data observed in S\&W (the complete information on the replication of the key statistical tests reported in S\&W is available in the Supplementary Materials). Importantly, this direct replication of S\&W provides evidence that answering the counterfactual question first did not unduly influence participants' causal judgments.

We next analyzed participants' relevance judgments, and observed a highly similar pattern of results: an effect of norm-condition $(L R T=13.93[d f=3], p=.003)$, an effect of question $(L R T=73.34[d f=2], p<.001)$, and a norm-condition $\times$ question interaction effect $(L R T=14.15$ $[d f=6], p=.028$. Critically, because participants answered this question first, the observed pattern cannot have been influenced by participants' causal judgments. All the same, at the level of each participants' responses, judgments of the causal responsibility were highly correlated with judgments of whether it was relevant to consider alternatives to the agents' actions. This was true both for judgment of the norm-violating agent/action/object, (Pearson's $r=0.553$, $p<.001$ ), and for the norm-conforming agent/action/object
(Pearson's $r=0.406, p<.001$ ), and moreover, held whether participants were making judgments about agents (Pearson's $r=0.651, p<.001$ ), actions (Pearson's $r=0.262, p<$ .001), or simply inanimate objects (Pearson's $r=0.280, p<$ .001).The similarity in the overall pattern of these judgments across all of the conditions can be seen in Figure 1.


Figure 1: Depiction of the relationship between participants' causal ratings and relevance ratings. Judgments related to the norm-conforming agent are marked with a ' C '; Judgments related to the norm-violating agent are marked with a ' V '.

## Discussion

In sum, we replicated S\&W's causal judgments, and found a corresponding pattern for which counterfactual alternatives participants regarded as relevant. These findings mirror those in observed in Phillips et al. (2015), which used a continuous rather than dichotomous measure of counterfactual relevance, and more importantly are predicted by a unified counterfactual account of the impact of norms on causal judgments.

## Experiment 2

We next investigated whether causal judgments of inanimate objects are sensitive to violations of prescriptive norms of proper functioning (e.g., a machine malfunctioning). According to a unified counterfactual account, when a machine malfunctions, it should become more relevant to consider the counterfactual possibility that the machine could have instead functioned as intended, and thus the machine should be judged as more causal. Polysemy accounts do not predict such an effect, as participants should not interpret the word 'cause' to mean 'morally responsible' when discussing an inanimate object. We test these two predictions.

## Methods

Participants. 403 participants $\left(M_{\text {age }}=34.96, S D_{\text {age }}=\right.$ 11.90; 205 females, 1 unreported) from Amazon Mechanical Turk participated for a $\$ 0.25$ compensation. Participant recruitment was again automated through TurkPrime.

Stimuli and procedure. This experiment used a 3 (Norm violation; norm-conforming vs. moral violation vs. malfunction) x 2 (Question: agent vs. object) design, administered fully between-subjects.

Participants read one of three vignettes involving a vending machine in an academic department. In every condition, the machine has three levers (red, black, and white): two produce pencils and one produces an eraser but frequently malfunctions and also produces a broken pencil. There were also two agents: an administrative assistant, and Professor Smith (a recent hire who did not know about the malfunctioning lever). Prof. Smith always pulls the red lever, and the assistant always pulls the black lever. This later results in a problem for a student who needs a pencil to take a test but cannot get one.

In the norm-conforming condition, the red lever and black lever both produce pencils, and the white lever produces erasers (but also consistently malfunctions). Additionally, both administrators and faculty were allowed to take pencils from the machine. Both agents request pencils using the black and red levers, which both function appropriately. The moral violation condition was identical to the normconforming condition, except that the faculty are not allowed to get pencils from the machine (but administrative assistants are allowed), and this rule was known by Prof. Smith. Lastly, the malfunction condition was identical to the normconforming condition except that it was the red lever that produced erasers (and malfunctioned), and Prof. Smith wanted an eraser, so Prof. Smith pulled the red lever and got an eraser and a broken pencil.

Participants were then asked a question about the relevance of counterfactual alternatives and a causal question in random order on separate pages. The relevance of alternatives question was worded and presented the same way as Experiment 1, and either focused on the agents (Prof. Smith, administrative assistant) or the objects (red lever, black lever). The causal question similarly asked either who caused the problem (agent condition) or what caused the problem (object condition), and participants could select one or both.

These were followed by three comprehension check questions and two additional manipulation-check questions. The comprehension questions ensured that participants understood the key facts about the levers, agents, and outcome of the scenario. Additionally, participants rated, on a 0 100 scale, how likely the malfunction was to occur, in order to verify that participants did not think the malfunction also violated a descriptive (statistical) norm. Finally, participants rated their agreement with the statement "It was morally wrong for Prof. Smith to pull the red lever" on a 7-point Likert scale, with the expectation that ratings should be higher in the moral violation condition than the other two conditions, which should not differ from each other. The predictions for both manipulation-check questions were overwhelmingly confirmed (see Supplementary Materials). Thus, any effect of the functional norm violation cannot be explained by appealing to statistical or moral norms.

## Results

We excluded participants who did not answer all three of the check questions correctly, and analyzed the remaining 258 participants' judgments. To facilitate comparison of participants' judgments, we computed a measure of participants' preference for selecting the norm-violating event as a cause. Participants who selected only the norm-violating event as a cause were assigned a score of 1 ; participants who selected both or neither events as causes were assigned score of 0 ; and participants who selected only the norm-conforming event were assigned a score of -1 . We then analyzed participants' causal preference scores with a 2 (Causal Question: Agent vs. Object) $\times 3$ (Norm condition: Immoral vs. Malfunction vs. Normal) proportional odds logistic regression, as in Study 1. This analysis revealed a main effect of Norm condition, $(L R T=71.49[d f=2], p<.001)$, no main effect of Causal question $(L R T=0.045[d f=1], p=.832)$, and critically a Norm condition $\times$ Causal question interaction effect ( $L R T=31.42[d f=2], p<.001$ ).

We decomposed this interaction effect by separately analyzing participants' causal preference scores for each of the different conditions. When the relevant norm was moral and thus applied to the agent but not the object, participants tended to prefer the norm-violating agent as a cause, but did not similarly prefer the norm-violating object as a cause ( $L R T=15.33[d f=1], p<.001$ ). When the relevant norm was functional, and thus the norm applied to the object but not the agent, this pattern was reversed: participants tended to prefer the norm-violating object as a cause, but did not similarly prefer the norm-violating agent as a cause $(L R T=12.36$ $[d f=1], p<.001)$. When there was no norm that applied to either the agent or the object, there was small and nonsignificant preference for the norm-conforming agent but not the object ( $L R T=1.13[d f=1], p=.288$ ).

We next analyzed participants' judgments of the relevance of counterfactual alternatives in exactly the same way. Just as with participants' causal judgments, we observed a main effect of Norm condition, $(L R T=40.53[d f=2], p<.001)$, no main effect of Relevance question $(L R T=0.10[d f=1]$, $p=.747$ ), and critically a Norm condition $\times$ Relevance question interaction effect ( $L R T=33.70[d f=2], p<.001$ ). We decomposed this interaction effect by separately analyzing participants' counterfactual preference scores for each of the different conditions. When a moral norm was salient, participants tended to prefer counterfactuals for the agent, but not the object $(L R T=16.63[d f=1], p<.001)$. When the relevant norm was functional, this pattern was reversed: participants preferred counterfactuals for the object, but not the agent $(L R T=11.20[d f=1], p<.001)$. When there was no norm violation that applied to either the agent or the object, there was a small and significant preference for the normconforming agent, but not the object $(L R T=4.48[d f=1]$, $p=.034)$. A similar pattern is found when only participants' first responses are analyzed, allowing for a between-subjects analysis (see Supplementary Materials).


Figure 2: Average preference score for the norm-violating event in causal judgments (top) and counterfactual relevance judgments (bottom), as a function of which norms were relevant (split into panels). Grey bars depict responses to questions about agents; Red bars depict responses to questions about inanimate objects. Error bars depict +/- 1 SEM.

## Discussion

Experiment 2 found that judgments of inanimate objects are sensitive to violations of prescriptive norms of proper functioning, even though they are not sensitive to violations of moral norms. Specifically, we found that when an inanimate object violated a functional norm, participants' thought it was relevant to consider counterfactual alternatives to that malfunction, and that this effect was mirrored by a corresponding change in participants' causal judgments. This pattern is uniquely predicted by a unified counterfactual accounts of the impact of norms on causal judgments, and is not predicted by an account on which the term 'cause' is polysemous.

## Experiment 3

Previous research (Phillips et al., 2015) has demonstrated that causal judgments are also sensitive to more direct counterfactual manipulations: participants tend to judge an event to be more causal after they generate relevant alternative ways that the event could have occurred. In Experiment 3, we extend this method by asking participants to generate alternatives to one particular aspect of the causal structure that contributes to the outcome (i.e., to the agent or to the inanimate object). We then measure how their causal judgments are affected by this manipulation. This allows us to test a precise prediction of a unified counterfactual account: participants causal judgments should be affected by the generation of counterfactual alternatives primarily for the part of the causal structure that the counterfactual alternative focused on.

## Methods

Participants. 601 participants $\left(M_{\text {age }}=35.96, S D_{\text {age }}=\right.$ 15.58; 304 females, 2 unreported) from Amazon Mechanical Turk participated for $\$ 0.35$ in compensation. Participant recruitment was again automated through TurkPrime.
Stimuli and procedure. This experiment used a 3 (AgentCounterfactual vs Object-Counterfactual vs No Counterfactual) $\times 2$ (Agent Question vs Object Question) design. Counterfactual condition was manipulated between-subjects and Question was manipulated within-subjects.

All participants read the norm-conforming condition from Experiment 2, where the red lever and black lever both produce pencils, and the white lever produces erasers (but also consistently malfunctions). In this scenario, both the administrators and the faculty are allowed to take pencils from the machine, both the administrative assistant and Professor smith request pencils using the black and red levers respectively, and both levers function appropriately to produce pencils. A problem then arises from a lack of pencils.

After reading the vignette, participants underwent the counterfactual manipulation. In the Agent-Counterfactual condition, for example, participants were asked to think about Professor Smith's decision to take a pencil from the vending machine, and then to consider and describe one relevant way that things could have gone differently such that the professor would not have taken one of the pencils from the vending machine. In the Object-Counterfactual condition, by contrast, participants were instead asked to consider and describe a relevant way in which the red lever could have functioned differently such that it didn't produce a pencil from the vending machine. In the No Counterfactual condition, participants were simply asked to describe the story they read.

After completing this task, they rated their agreement (on a scale from 0 ('Completely disagree') to 100 ('Completely agree') with a statement that the Professor caused the problem, and separately with a statement that the red lever caused the problem. The statements were presented in counterbalanced order and on separate pages. Participants then completed a series of control questions that asked them about


Figure 3: Agreement with the causal statement concerning the agent (left bars) and the object (right bars) as a function of Counterfactual condition). Error bars depict +/- 1 SEM.
which levers were actually pulled and about who actually received a pencil in the original story.

## Results

We excluded participants who did not answer both of the check questions correctly, and analyzed the remaining 423 participants' judgments. First, we analyzed the agreement with the two causal statements by comparing a series of linear mixed-effects models using the lme4 package in R (Bates, Maechler, Bolker, Walker, et al., 2014). This analysis revealed a main effect of Question $\left(\chi^{2}(1)=53.135\right.$, $p<.001)$ and a main effect of Condition $\left(\chi^{2}(2)=13.492\right.$, $p=.001$ ). Critically, however, these were qualified by a significant Question $\times$ Condition interaction $\left(\chi^{2}(2)=23.04\right.$, $p<.001$ ). We decomposed this interaction using a series of planned comparisons. These analyses revealed that participants strongly agreed that Professor Smith was a cause of the problem when they considered alternatives to Professor Smith's action ( $M=32.99, S D=33.33$ ), but in comparison, agreed significantly less both when they did not consider counterfactual alternatives ( $M=18.22, S D=27.28$ ), $t(282.48)=4.12, p<.001, d=0.482$, and when they only considered alternatives to the way the lever functioned ( $M=$ $24.43, S D=29.12$ ), $t(279)=2.27, p=.024, d=0.272$.

We also observed a corresponding pattern in participants' agreement with the statement that the red lever caused the problem: participants agreed that the lever was more of a cause when they considered alternatives to the way the lever functioned ( $M=20.11, S D=33.34$ ), than when they did not generate any relevant counterfactual alternatives, $(M=8.62$, $S D=19.64), t(211.21)=3.42, p<.001, d=0.421$, or when they considered alternatives to what Professor Smith $\operatorname{did}(M=10.05, S D=20.59), t(213.65)=2.99, p=.003$, $d=0.367$.

## Discussion

In short, we found that directly manipulating the relevance of counterfactual alternatives affected participants' causal judgments. Moreover, in line with the predictions of a unified counterfactual account, we found this effect was specific to the factor that was altered in the counterfactual alternative.

## General Discussion

The results of these three experiments favor a counterfactual relevance account of the impact of norms on causal judgments. Experiment 1 replicated S\&W's finding that moral norm violations primarily affect causal judgments of intentional agents and not inanimate objects. Experiment 2 further found that violations of norms of proper functioning primarily affect judgments of inanimate objects but not intentional agents. In both experiments, judgments of counterfactual relevance tracked the impact of different norm violations on causal judgments for both intentional agents and inanimate objects. Finally, Experiment 3 demonstrated that nonnormative manipulations of counterfactual relevance produce an analogous pattern in participants' causal judgments. These results support work on causal cognition that provides a central role for counterfactuals (Gerstenberg \& Tenenbaum, in press; Kominsky et al., 2015; Icard et al., in press).

The extant literature on causal judgment now provides evidence for three distinct types of norms that all show similar effects: descriptive statistical norm violations (e.g., Kominsky et al., 2015), prescriptive moral norm violations (e.g., Hitchcock \& Knobe, 2009), and prescriptive functional norm violations (demonstrated here). The demonstration of additional norms that have similar a impact on causal judgments makes a parsimonious explanation increasingly desirable. To extend the polysemy account, for example, one would now have to propose three independent explanations for three qualitatively similar effects. By contrast, an account that appeals to the relevance of counterfactual alternatives provides a unified explanation and predicts that these various norms should all have a qualitatively similar impact.

At the same time, many aspects of the relationship between counterfactual representation and causal cognition remain poorly understood. For example, a critical insight which arises in both $\mathrm{S} \& \mathrm{~W}$ and in the current studies is that norms have a highly specific effect on causal judgments: they preferentially affect causal judgments of the entities to which the norm applies and typically do not extend to other aspects of the same event. Across three experiments, we find a similar pattern in participants' reasoning about counterfactual alternatives. Collectively, these findings suggest that, rather than representing a counterfactual alternative to an event in its entirety, participants' causal and counterfactual cognition represents events more granularly.

Not only does this shape our interpretation of S\&W's original result, it opens an exciting new frontier in the study of causal cognition. How events are represented in causal and counterfactual cognition, and which aspects of an event are
represented as distinct variables, are almost completely unexplored topics (e.g. Halpern \& Hitchcock, 2015 explicitly acknowledge this issue). However, as emerging research makes clear, it will be difficult to make precise predictions about the impact of norms without a more well worked-out theory of how events are represented in causal reasoning.

This opportunity cuts in both directions. These results are, to our knowledge, the first empirical investigation of which events are represented as distinct variables. Yet, as much as we need to build precise theories of how these events are represented in order to understand how norms will affect causal judgments, we can also use the effect of norms on causal judgments to determine which causes are distinct. We look forward to exploring these questions in future work.

## Acknowledgments

We would like to thank Harvard's Moral Psychology Research Lab, D. Lagnado, T. Gerstenberg, and J. Knobe for helpful discussion. This work was supported by Grant N00014-14-1-0800 from the Office of Naval Research to JP.

## References

Alicke, M. D. (2000). Culpable control and the psychology of blame. Psychological bulletin, 126, 556.
Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). lme4: Linear mixed-effects models using eigen and s4. $R$ package version, 1.
Bello, P. F. (2016). Machine ethics and modal psychology. In Computing and philosophy (pp. 245-258). Springer International Publishing.
Gerstenberg, T., \& Tenenbaum, J. B. (in press). Intuitive theories. In M. Waldmannn (Ed.), Oxford handbook of causal reasoning. Oxford University Press.
Halpern, J. Y., \& Hitchcock, C. (2015). Graded causation and defaults. The British Journal for the Philosophy of Science, 66, 413-457.
Hitchcock, C., \& Knobe, J. (2009). Cause and norm. Journal of Philosophy, 106, 587-612.
Icard, T. F., Kominsky, J., \& Knobe, J. (in press). Causality, normality, and sampling propensity. Elsevier.
Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., \& Knobe, J. (2015). Causal superseding. Cognition, 137, 196-209.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. Cognitive Psychology, 61, 303-332.
Phillips, J., Luguri, J., \& Knobe, J. (2015). Unifying moralitys influence on non-moral judgments: The relevance of alternative possibilities. Cognition, 145, 30-42.
Samland, J., Josephs, M., Waldmann, M. R., \& Rakoczy, H. (2016). The role of prescriptive norms and knowledge in childrens and adults causal selection. Journal of Experimental Psychology: General, 145, 125.
Samland, J., \& Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. Cognition, 156, 164176.

# Assessing the Linguistic Productivity of Unsupervised Deep Neural Networks 

Lawrence Phillips (Lawrence.Phillips@pnnl.gov)<br>Pacific Northwest National Laboratory

Nathan Hodas (Nathan.Hodas@pnnl.gov)<br>Pacific Northwest National Laboratory


#### Abstract

Increasingly, cognitive scientists have demonstrated interest in applying tools from deep learning. One use for deep learning is in language acquisition where it is useful to know if a linguistic phenomenon can be learned through domain-general means. To assess whether unsupervised deep learning is appropriate, we first pose a smaller question: Can unsupervised neural networks apply linguistic rules productively, using them in novel situations? We draw from the literature on determiner/noun productivity by training an unsupervised, autoencoder network measuring its ability to combine nouns with determiners. Our simple autoencoder creates combinations it has not previously encountered and produces a degree of overlap matching adults. While this preliminary work does not provide conclusive evidence for productivity, it warrants further investigation with more complex models. Further, this work helps lay the foundations for future collaboration between the deep learning and cognitive science communities.


Keywords: Deep Learning; Language Acquisition; Linguistic Productivity; Unsupervised Learning; Determiners

## Introduction

Computational modeling has long played a significant role within cognitive science, allowing researchers to explore the implications of cognitive theories and to discover what properties are necessary to account for particular phenomena (J. L. McClelland, 2009). Over time, a variety of modeling traditions have seen their usage rise and fall. While the 1980s saw the rise in popularity of connectionism (Thomas \& McClelland, 2008), more recently symbolic Bayesian models have risen to prominence (Chater \& Oaksford, 2008; Lee, 2011). While the goals of cognitive modelers have largely remained the same, increases in computational power and architectures have played a role in these shifts (J. L. McClelland, 2009). Following this pattern, recent advances in the area of deep learning (DL) have led to a rise in interest from the cognitive science community as demonstrated by a number of recent workshops dedicated to DL (Saxe, 2014; J. McClelland, Hansen, \& Saxe, 2016; J. McClelland, Frank, \& Mirman, 2016).

As with any modeling technique, DL can be thought of as a tool which is best suited to answering particular types of questions. One such question is that of learnability, whether an output behavior could ever be learned from the types of input given to a learner. These types of questions play an integral role in the field of language acquisition where researchers have argued over whether particular aspects of language could ever be learned by a child without the use of innate, language-specific mechanisms (Smith, 1999; C. D. Yang, 2004; Chater \& Christiansen, 2010; Pearl,
2014). The success of a domain general learner does not necessarily imply that human learners acquire the phenomenon in a similar fashion, but it does open the possibility that we need not posit innate, domain-specific knowledge.

The crux of these learning problems typically lies in making a particular generalization which goes beyond the input data. One major type of generalization that DL models would need to capture is known as linguistic productivity. A grammatical rule is considered productive when it can be applied in novel situations. For example, as a speaker of English you may never have encountered the phrase a gavagai before, but you now know that gavagai must be a noun and can therefore combine with other determiners to produce a phrase such as the gavagai. Before DL might be applied to larger questions within language acquisition, the issue of productivity must first be addressed. If DL models are not capable of productivity, then they cannot possibly serve to model the cognitive process of language acquisition. On the other hand, if DL models demonstrate basic linguistic productivity, we must explore what aspects of the models allow for this productivity.

## The Special Case of Determiners

For decades, debate has raged regarding the status of productive rules among children acquiring their native language. On the one hand, some have argued that children seem hardwired to apply rules productively and demonstrate this in their earliest speech (Valian, Solt, \& Stewart, 2009; C. Yang, 2011). On the other, researchers have argued that productivity appears to be learned, with children's early speech either lacking productivity entirely or increasing with age (Pine \& Martindale, 1996; Pine, Freudenthal, Krajewski, \& Gobet, 2013; Meylan, Frank, Roy, \& Levy, 2017). Of particular interest to this debate has been the special case of English determiners. In question is whether or not English-learning children have acquired the specific linguistic rule which allows them to create a noun phrase (NP) from a determiner (DET) and noun ( N ) or if they have simply memorized the combinations that they have previously encountered. This linguistic rule, $\mathrm{NP} \rightarrow$ DET N, is productive in two senses. First, it can be applied to novel nouns, e.g. a gavagai. Second, consider the determiners $a$ and the. If a singular noun can combine with one of these determiners, it may also combine with the other, e.g. the wug.

This type of rule seems to be acquired quite early in acquisition, making it appropriate to questions of early productivity, and provides an easy benchmark for a DL model. Yet answering such a simple question first requires addressing
how one might measure productivity. Most attempts to measure productivity have relied on what is known as an overlap score, intuitively what percentage of nouns occur with both $a$ and the (C. Yang, 2011). This simple measure has been the source of some controversy. C. Yang (2011) argues that early attempts failed to take into account the way in which word frequencies affect the chance for a word to "overlap". Because word frequency follows a Zipfian distribution, with a long tail of many infrequent words, many nouns are unlikely to ever appear with both determiners. He proposes a method to calculate an expected level of overlap which takes into account these facts. Alternatively, Meylan et al. (2017) propose a Bayesian measure of productivity which they claim takes into account the fact that certain nouns tend to prefer one determiner over another. For instance, while one is more likely to hear a bath than the phrase the bath, the opposite is true of the noun bathroom which shows a preference for the determiner the (Meylan et al., 2017).

The literature is quite mixed regarding whether or not children show early productivity. Differences in pre-processing have lead researchers to draw opposite conclusions from similar data, making interpretation quite difficult (C. Yang, 2011; Pine et al., 2013). Indeed, most corpora involving individual children are small enough that Meylan et al. (2017) argue it is impossible to make a statistically significant claim as to child productivity. For analyzing whether or not text generated by a DL model is productive or not, we thankfully do not need to fully address the problem of inferring child productivity. Ideally, the model would demonstrate a similar level of overlap to the data it was exposed to. We make use of the overlap statistic from Yang because it is more easily comparable to other works and has been better studied than the more recent Bayesian metric of Meylan et al. (2017).

## Deep Learning for Language Acquisition

Deep learning, or deep neural networks, are an extension of traditional artificial neural networks (ANN) used in connectionist architectures. A "shallow" ANN is one that posits a single hidden layer of neurons between the input and output layers. Deep networks incorporate multiple hidden layers allowing these networks in practice to learn more complex functions. The model parameters can be trained through the use of the backpropogation algorithm. The addition of multiple hidden layers opens up quite a number of possible architectures, not all of which are necessarily applicable to problems in cognitive science or language acquisition more specifically.

While the most common neural networks are discriminative, i.e. categorizing data into specific classes, a variety of techniques have been proposed to allow for truly generative neural networks. These generative networks are able to take in input data and generate complex outputs such as images or text which makes them ideal for modeling human behavior. We focus on one generative architecture in particular known as a deep autoencoder (AE) (Hinton \& Salakhutdinov, 2006).

While AEs have been used for a variety of input data types,
most prominently images, we describe their use here primarily for text. The first half, the encoder, takes in sentences and transforms them into a condensed representation. This condensed representation is small enough that the neural network cannot simply memorize each sentence and instead is forced to encode only the aspects of the sentence it believes to be most important. The second half, the decoder, learns to take this condensed representation and transform it back into the original sentence. Backpropogation is used to train model weights to reduce the loss between the original input and the reconstructed output. Although backpropagation is more typically applied to supervised learning problems, the process is in fact unsupervised because the model is only given input examples and is given no external feedback.

AEs have been shown to successfully capture text representations in areas such as paragraph generation (Li, Luong, \& Jurafsky, 2015), part-of-speech induction (Vishnubhotla, Fernandez, \& Ramabhadran, 2010), bilingual word representations (Chandar et al., 2014), and sentiment analysis (Socher, Pennington, Huang, Ng, \& Manning, 2011), but have not been applied to modeling language acquisition. While any number of DL architectures could be used to model language acquisition, the differences between ANNs and actual neurons in the brain make any algorithmic claims difficult. Instead, DL models might be used to address computationallevel questions, for instance regarding whether or not a piece of knowledge is learnable from the data encountered by children. Before this can be done, however, it remains to be seen whether DL models are even capable of creating productive representations. If they cannot, then they do not represent useful models of language acquisition. This work attempts to address this not by creating a model of how children acquire language, but by using methods from the psychological literature on productivity to assess the capability of DL to learn productive rules.

## Methods

## Corpora

To train our neural network, we make use of child-directed speech taken from multiple American-English corpora in the CHILDES database (MacWhinney, 2000). In particular, we make use of the CDS utterances in the Bloom 1970, Brent, Brown, Kuczaj, Providence, Sachs, and Suppes corpora (Bloom, 1970; Brent \& Siskind, 2001; Brown, 1973; Kuczaj, 1977; Demuth \& McCullough, 2009; Sachs, 1983; Suppes, 1974). The combined corpora contain almost 1 million utterances and span a wide age range, including speech directed to children as young as 6 months and as old as 5 years. Relevant information about the used corpora can be found in Table 1.

Because we are interested in seeing what the AE can learn from data similar to that encountered by children, we train the model only on child-directed utterances. These can be produced by any adult in the dataset, including parents and researchers. Although a comparison with child-produced text


Dense Softmax Layer
GRU Layer (20 dim)
Latent Representation (20 dim)
GRU Layer (20 dim)
Embedding Layer (30 dim)

Figure 1: Visual representation of the autoencoder model.
holds great interest, it is not clear whether child-produced speech is rich enough to support robust language learning on its own. It therefore provides a poor basis upon which to train the AE.

Text from the various corpora is processed as a single document. Child-directed utterances are cleaned from the raw files using the CHILDESCorpusReader function of the Python Natural Language Toolkit (NLTK). Utterances from all nonchildren speakers are included and not limited just to the primary caregiver. Each utterance is split into words according to the available CHILDES transcription and then made lowercase. The model represents only the most frequent 3000 words, while the remainder are represented as a single out-of-vocabulary (OOV) token. This step is taken both to reduce computational complexity but also to mimic the fact that young children are unlikely to store detailed representations of all vocabulary items encountered. Because the neural networks require each input to be of the same length, sentences are padded to a maximum length of 10 words. Sentences that are longer than this are truncated, while short sentences are prepended with a special $P A D$ token.

| Corpora | Age Range | N. Utterances |
| :--- | ---: | ---: |
| Bloom 1970 | $1 ; 9-3 ; 2$ | 62,756 |
| Brent | $0 ; 6-1 ; 0$ | 142,639 |
| Brown | $1 ; 6-5 ; 1$ | 176,856 |
| Kuczaj | $2 ; 4-4 ; 1$ | 57,719 |
| Providence | $1 ; 0-3 ; 0$ | 394,800 |
| Sachs | $1 ; 1-5 ; 1$ | 28,200 |
| Suppes | $1 ; 11-3 ; 3$ | 67,614 |
| Overall | $0 ; 6-5 ; 1$ | 930,584 |

Table 1: Descriptive statistics of CHILDES corpora. Ages are given in (year;month) format and indicate the age of the child during corpus collection.

## Neural Network Architecture

Our autoencoder model was implemented using Keras and Tensorflow. The words in each sentence are input to the model as a one-hot vector, a vector of 0 s with a single 1 whose placement indicates the presence of a particular word. This is
an inefficient representation because it assumes all words are equally similar, e.g. that $d o g$ is equally similar to dogs as it is to truck. To deal with this, the model passes the onehot vector to an embedding layer. Neural word embeddings, as popularized by the word2vec algorithm (Mikolov, Chen, Corrado, \& Dean, 2013), are a way to represent words in a low-dimensional space without requiring outside supervision. Words are placed within the space such that words that are predictive of neighboring words are placed closer to one another. Because our training data is relatively small, we keep the embedding dimensionality low, at only 30 . Standard embeddings trained on much larger NLP corpora tend to use 100 or 200 dimensions.

Once each word has been transformed into a 30dimensional embedding vector, the sequence of words is passed into a gated-recurrent unit (GRU) layer (Cho et al., 2014). The GRU is a type of recurrent (RNN) layer which we choose because it can be more easily trained. RNN layers read in their inputs sequentially and make use of hidden "memory" units that pass information about previous inputs to later inputs, making them ideal for sequence tasks such as language. As such, the model creates a representation of the sentence which it passes from word to word. The final representation is the output of the encoder, a latent representation of the full sentence.

This 20-dimensional latent vector serves as the input to the decoder unit. The first layer of the decoder is a GRU layer of the same shape as in the encoder. For each timestep, we feed into the GRU the latent vector, similar to the model proposed in Cho et al. (2014). Rather than producing a single output, as in the encoder, the decoder's GRU layer outputs a vector at each timestep. Each of these vectors is fed into a shared dense softmax layer which produces a probability distribution over vocabulary items. The model then outputs the most likely word for each timestep.

The model loss is calculated based on the model's ability to reconstruct the original sentence through categorical crossentropy. Model weights are trained using the Adam optimzer over 10 epochs. During each epoch the model sees the entire training corpus, updating its weights after seeing a batch of 64 utterances. While this process does not reflect that used by a child learner, it is a necessary component of training the neural network on such a small amount of data. If the network had access to the full set of speech that a child encounters such a measure likely would not be necessary. Future work might also investigate whether optimizing the dimensionality of the network might lead to better text generation with higher levels of productivity.

## Baseline Models

Because the AE is learning to reproduce its input data, one might wonder whether similar results might be achieved by a simpler, distributional model. To assess this, we also measure the performance of an n-gram language model. We train bigram and trigram language models using the modified Kneser-Ney smoothing (Heafield, Pouzyrevsky, Clark,
\& Koehn, 2013) implemented in the KenLM model toolkit to estimate the distributional statistics of the training corpus. Sentences are generated from the n-gram language model by picking a seed word and then sampling a new word from the set of possible $n$-grams. The smoothing process allows for the model to generate previously unseen n-grams. Sampling of new words continues for each utterance until the end-ofsentence token is generated or a maximum of 10 tokens is reached (the same maximum size as for the AE).

Since the AE is able to generate sentences from a latent representation, it would be inappropriate to generate $n$-gram sentences from random seed words. Instead, for every sentence in the test set we begin the n-gram model with the first word of the utterance. While this allows the model to always generate its first token correctly, this does not directly impact our measure of productivity as it relies on combinations of tokens.

## Productivity Measures

We measure the productivity of our autoencoders through the overlap score described in C. Yang (2011). Words both in the child-directed corpus and the autoencoder-generated output are tagged using the default part-of-speech tagger from NLTK. The empirical overlap scores are simply calculated as a percentage of unique nouns that appear immediately after both the determiners $a$ and the. The expected overlap score is calculated based off of three numbers from the corpus under consideration, the number of unique nouns $N$, the number of unique determiners $D$, and the total number of noun/determiner pairs $S$. The expected overlap is defined as in Equation 1:

$$
\begin{equation*}
O(N, D, S)=\frac{1}{N} \sum_{r=1}^{N} O(r, N, D, S) \tag{1}
\end{equation*}
$$

where $O(r, N, D, S)$ is the expected overlap of the noun at frequency rank $r$ :

$$
\begin{equation*}
O(r, N, D, S)=1+(D-1)\left(1-p_{r}\right)^{S}-\sum_{i=1}^{D}\left[\left(d_{i} p_{r}+1-p_{r}\right)^{S}\right] \tag{2}
\end{equation*}
$$

$d_{i}$ represents the probability of encountering determiner $i$, for which we use the relative frequencies of $a$ and the calculated from the training corpus ( $39.3 \%$ and $60.7 \%$, respectively). The probability $p_{r}$ represents the probability assigned to a particular word rank. The Zipfian distribution takes a shape parameter, $a$ which C. Yang (2011) set equal to 1 and which we optimize over the training corpus using least squares estimation and set at 1.06 :

$$
\begin{equation*}
p_{r}=\frac{1 / r^{a}}{\sum_{n=1}^{N}\left(\frac{1}{n^{a}}\right)} \tag{3}
\end{equation*}
$$

It should be noted that Zipfian distributions are not perfect models of word frequencies (Piantadosi, 2014), but assigning empirically-motivated values to the determiner probabilities
and Zipfian parameter $a$ represents an improvement upon the original measure.

## Results

We analyze our overlap measures for the adult-generated (i.e. child-directed) as well as the autoencoder and n-gram modelgenerated text and present these results in Figure 2. We analyze overlap scores across 10 training epochs with three levels of dropout, $10 \%, 20 \%$, and $30 \%$. Dropout is typically included in neural models to encourage the model to better generalize. We hypothesized that a certain level of dropout would encourage the model to generate novel combinations of words that might lead to higher overlap scores. We find that with only two training epochs the AEs have already begun to near their maximum overlap performance. The $30 \%$ dropout AE achieves the highest level of performance, matching the empirical overlap score of the original corpus. The $10 \%$ and $20 \%$ dropout models perform somewhat worse suggesting that high levels of dropout may be necessary for good text generation.

In Table 2, we present the results for the final epoch of the AE models as well as for the adult-generated and $n$ gram generated text. We note that the expected overlap measure consistently overestimates the productivity of all learners, including the adult-generated text. It is unclear why this should be the case, but could be a result of capping the model vocabularies, resulting in lower $N$ values. In particular, the autoencoders tend to produce a relatively limited set of nouns. Looking at empirical overlap measures, the worstperforming models are the bigram and trigram models with overlap scores below $30 \%$. The AEs fair much better all producing overlap scores over $50 \%$. The $30 \%$ dropout AE is actually able to match the overlap score of the original adultgenerated corpus ( $59.4 \%$ vs. $59.3 \%$ ).

Looking at the number of unique nouns following a determiner $(N)$ and the total number of determiner-noun pairs $(S)$, it becomes clear there are large differences between the n-gram and AE models. The n-gram models tend to produce very few determiner-noun pairs (low $S$ ) but are likely to choose from any of the nouns in the corpus, leading to high $N$. This fact accounts for the low overlap scores that they achieve. In contrast, the AEs follow a pattern which mirrors the adult corpus with few unique nouns but a large number of noun-determiner pairs. In all cases, however, the AEs produce both fewer unique nouns and fewer noun-determiner pairs than the original corpus.

One possible problem for calculating the expected overlaps comes from the difficulty of part-of-speech tagging text generated by the neural network. Whereas adult-generated speech follows set patterns that machine taggers are built to recognize, the neural network does not necessarily generate well-formed language. Examples of AE-generated text can be found in Table 3. In some cases, the tagger treats items that occur after a determiner as a noun regardless of its typical usage. For example, in the generated sentence let put


Figure 2: Empirical overlap scores. Adult-generated speech is marked by the solid black line while autoencoder-generated speech is marked by the dashed colored lines. Results are presented for three levels of dropout, $10 \%, 20 \%$, and $30 \%$. The x -axis represents the training epoch of the model.

|  | $\boldsymbol{N}$ | $\boldsymbol{S}$ | Exp. Over. | Emp. Over. |
| :--- | ---: | ---: | ---: | ---: |
| Adult | 1,390 | 34,138 | $77.5 \%$ | $59.3 \%$ |
| AE 10\% | 861 | 29,497 | $88.4 \%$ | $53.3 \%$ |
| AE 20\% | 870 | 28,817 | $87.6 \%$ | $53.4 \%$ |
| AE 30\% | 816 | 31,181 | $90.8 \%$ | $59.4 \%$ |
| Bigram | 1,780 | 5,177 | $17.6 \%$ | $28.6 \%$ |
| Trigram | 2,506 | 4,595 | $11.2 \%$ | $22.1 \%$ |

Table 2: Expected and empirical overlap scores for adultand autoencoder-generated language with varying levels of dropout. Expected overlap scores were calculated as in Yang (2011). Empirical overlap was calculated as the percent of unique nouns that appeared immediately following both $a$ and the.
put the over over here, the phrase the over is tagged as a $\mathrm{DET}+\mathrm{N}$ pair. These type of errors are further evidenced by the fact that the trigram language model produces a larger set of words tagged as nouns than the original adult-generated corpus ( 2,506 vs. 1,390 ).

Another explanation for the difference between expected and empirical overlaps may come from deviation from a true Zipfian distribution of word frequencies. If word frequencies are Zipfian, we should expect a perfect correlation between log ranks and log counts. C. Yang (2011) report a correlation of 0.97, while our larger corpus deviates from this with $r^{2}=0.86$. Although we attempt to take this into account by fitting the Zipfian distribution's shape parameter, this divergence clearly indicates that further work is needed.

The success of the AE model in generating productive text serves as a confirmation that unsupervised neural models might be used in future work to investigate other cognitive phenomena. This work does not directly address the question of how infants might learn to produce productive speech, it does represent one possible approach. AEs can, for instance, be thought of as information compression algorithms which learn to represent high-dimensional data into a low-
dimensional latent space (Hinton \& Salakhutdinov, 2006). If the brain likewise attempts to find efficient representations of the stimuli it encounters then it may prove fruitful to investigate how these representations compare to one another.

| Adult | Autoencoder |
| :--- | :--- |
| falling down | down down |
| you're playing with | you're playing with <br> the head |
| your bus | what what you say say <br> why did OOV say what's <br> wrong with these apples |
| say with the dada |  |

Table 3: Example adult and AE-generated language. The AEgenerated text is from the final epoch of the AE with $20 \%$ dropout. In bold is a $\mathrm{DET}+\mathrm{N}$ combination that does not appear in the AEs input.

## Conclusion

While there is great interest regarding the inclusion of deep learning methods into cognitive modeling, a number of major hurdles remain. For the area of language acquisition, deep learning is poised to help answer questions regarding the learnability of complex linguistic phenomena without access to innate, linguistic knowledge. Yet it remains unclear whether unsupervised versions of deep learning models are capable of capturing even simple linguistic phenomena. In this preliminary study, we find that a simple autoencoder with sufficient levels of dropout is able to mirror the productivity of its training data, although it is unclear whether this proves productivity in and of itself.

Future work will need to investigate whether more complex models might be able to generate text with higher productivity as well as further investigating how particular model choices impact performance. It would also be worthwhile to compare AEs against simpler models such as a basic LSTM language model. While additional work needs to be done to motivate the use of deep learning models as representations of how children might learn, this preliminary work shows how one might combine techniques from deep learning and developmental psychology.

## Acknowledgments

The authors thank the reviewers for their thoughtful comments and Lisa Pearl for initial discussion regarding productivity.

## References

Bloom, L. (1970). Language development: Form and function in emerging grammars. MIT Press.
Brent, M., \& Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. Cognition, 81, 31-44.
Brown, R. (1973). A first language: The early stages. Harvard University Press.

Chandar, S., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., \& Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In Advances in Neural Information Processing Systems (pp. 1853-1861).
Chater, N., \& Christiansen, M. H. (2010). Language acquisition meets language evolution. Cognitive Science, 34(7), 1131-1157.
Chater, N., \& Oaksford, M. (2008). The probabilistic mind: Prospects for bayesian cognitive science. Oxford University Press.
Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., \& Bengio, Y. (2014). Learning phrase representations using rnn encoderdecoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
Demuth, K., \& McCullough, E. (2009). The prosodic (re)organization of children's early english articles. Journal of Child Language, 36, 173-200.
Heafield, K., Pouzyrevsky, I., Clark, J. H., \& Koehn, P. (2013). Scalable modified kneser-ney language model estimation. In Proceedings of the Association of Computational Linguistics conference (pp. 690-696).
Hinton, G. E., \& Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.
Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. Journal of Verbal Learning and Verbal Behavior, 16, 589-600.
Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical bayesian models. Journal of Mathematical Psychology, 55(1), 1-7.
Li, J., Luong, M.-T., \& Jurafsky, D. (2015). A hierarchical neural autoencoder for paragraphs and documents. arXiv preprint arXiv:1506.01057.
MacWhinney, B. (2000). The CHILDES project: The database (Vol. 2). Psychology Press.
McClelland, J., Frank, S., \& Mirman, D. (Eds.). (2016). Contemporary neural network models: Machine learning, artificial intelligence, and cognition.
McClelland, J., Hansen, S., \& Saxe, A. (Eds.). (2016). Tutorial workshop on contemporary deep neural network models.
McClelland, J. L. (2009). The place of modeling in cognitive science. Topics in Cognitive Science, 1(1), 11-38.
Meylan, S. C., Frank, M. C., Roy, B. C., \& Levy, R. (2017). The emergence of an abstract grammatical category in children's early speech. Psychological Science, 0956797616677753.

Mikolov, T., Chen, K., Corrado, G., \& Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Pearl, L. (2014). Evaluating learning-strategy components: Being fair (commentary on Ambridge, Pine, and Lieven). Language, 90(3), e107-e114.

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin \& Review, 21(5), 1112-1130.
Pine, J. M., Freudenthal, D., Krajewski, G., \& Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. Cognition, 127(3), 345-360.
Pine, J. M., \& Martindale, H. (1996). Syntactic categories in the speech of young children: The case of the determiner. Journal of Child Language, 23(02), 369-395.
Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. In K. Nelson (Ed.), Children's language (Vol. 4). Lawrence Erlbaum Associates.
Saxe, A. (Ed.). (2014). Workshop on deep learning and the brain.
Smith, L. B. (1999). Do infants possess innate knowledge structures? The con side. Developmental Science, 2(2), 133-144.
Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., \& Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the conference on empirical methods in natural language processing (pp. 151-161).
Suppes, P. (1974). The semantics of children's language. American Psychologist, 29, 103-114.
Thomas, M. S., \& McClelland, J. L. (2008). Connectionist models of cognition. Cambridge handbook of computational cognitive modelling, 23-58.
Valian, V., Solt, S., \& Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. Journal of Child Language, 36(04), 743-778.
Vishnubhotla, S., Fernandez, R., \& Ramabhadran, B. (2010). An autoencoder neural-network based low-dimensionality approach to excitation modeling for HMM-based text-tospeech. In IEEE International Conference on Acoustics Speech and Signal Processing (pp. 4614-4617).
Yang, C. (2011). A statistical test for grammar. In Proceedings of the 2nd workshop on Cognitive Modeling and Computational Linguistics (pp. 30-38).
Yang, C. D. (2004). Universal grammar, statistics or both? Trends in Cognitive Sciences, 8(10), 451-456.

# Opinion Cascades and Echo-Chambers in Online Networks: A Proof of Concept Agent-Based Model 

Toby D. Pilditch (t.pilditch@ucl.ac.uk)<br>Department of Experimental Psychology, University College London<br>WC1E 6BT, Gower Street, London, United Kingdom


#### Abstract

In online networks, the polarization of opinions (e.g., regarding presidential elections or referenda) has been associated with the creation of "echo-chambers" of likeminded peers, secluded from those of contrary viewpoints. Previous work has commonly attributed such phenomena to self-regarding preferences (e.g., confirmation bias), individual differences, and the pre-dispositions of users, with clusters forming over repeated interactions. The present work provides a proof of concept Agent-Based Model that demonstrates online networks are susceptible to echo-chambers from a single opinion cascade, due to the spatiotemporal order induced by lateral transmission. This susceptibility is found to vary as a function of degree of interconnectivity and opinion strength. Critically, such effects are found despite globally proportionate levels of opinions, equally rational agents (i.e. absent conformity, confirmation bias or pre-disposition architecture), and prior to cyclical interactions. The assumptions and implications of this work, including the value of Agent-Based Modelling to cognitive psychology, are discussed.


Keywords: Information cascades; opinion dynamics; belief updating; Agent-Based Models

## Introduction

As online networks, such as social media, have developed and increased in popularity, research regarding the spread of false information, the polarization of opinions (Dandekar, Goel, \& Lee, 2013), and echo-chamber phenomena (Del et al., 2016) have become increasingly pertinent topics. Such phenomena pose a problem to society, and democracy as a whole, given the average user's exposure to only the information and opinions of their local (i.e. direct) network, leading to a break-down in informed debate and consensus.

Recently, questions regarding how individuals on a network receive new information and form or adopt opinions has come to the fore. Whether on topics of national referenda, deciding between presidential candidates, or interpreting news events (e.g., who is at fault in the annexation of Crimea, the shooting down of passenger aircraft, the political correctness of a reported comment or behavior), it has become more and more common for such information to be ascertained via social media ${ }^{1}$. In this way, an agent's source of information comes through a filter of network-acquaintances, presenting an unprecedented degree

[^172]of lateral, peer-to-peer dissemination of information. Such peer-to-peer transference of information, in a time where the information itself (whether "fake news", political memes, or posted opinion) carries a bias in its view of the world, presents a problem that psychology and multi-agent modelling is well-placed to answer.

The purpose of the present paper is two-fold: Firstly, this work provides a novel demonstration of the dangers of lateral propagation of opinions in online networks, based solely on the level of interconnectivity and the inherent way in which interpreted events (i.e. opinions) travel through them. This results in high levels of false consensuses and echo-chambers on a local level within the network. Critically, such localized clustering is shown to occur before any repeated interaction behaviors, and is robust to both different opinion strengths and propensities to communicate. Secondly, this work presents an argument that cognitive science is readily placed to lend insight into these interactive, societal level phenomena, and the superaggregate behaviors. Such insight can be lent by the ready application of cognitive models taken from individual-based empirical work and theory, to multi-agent simulations, known as Agent-Based Models (ABMs), so as to uncover and explain phenomena beyond the scope of individualbased experiments.

## Cascades and Opinions

How information is communicated between individuals on a societal (or multi-agent) scale, and its consequences, has been formally approached in two main areas; information cascades and opinion dynamics.

Research in information cascades has focused on the way in which information is spread through a system. This has included how networks may be resistant to cascading influence, such as the spread of cultural fads (Watts, 2002). Such work has typically characterized "information" as a singular, memetic entity that is propagated or hindered by either the properties of individuals within the network (such as the proportion of "easily influenced individuals", see Watts \& Dodds, 2007), or the structure of the network itself (e.g., hierarchical influencers; see Watts, 2002). This work has illustrated power law effects in information propagation across networks, an effect akin to percolation theory in physics (for a review, see Essam, 1980), wherein the clustered structure of a system leads to a critical singularity event (i.e. cascade). These cascades result in cluster size distribution effects, where smaller, more numerous clusters
occur as systems become more interconnected (Amar \& Family, 1995; Meakin, Vicsek, \& Family, 1985).

Research in opinion dynamics has instead focused on the cyclical interactions of individuals within a network. In particular, it has looked at the ways in which individuals and groups interact so as to either reach a consensus (Acemoglu \& Ozdaglar, 2011; Hegselmann \& Krause, 2002) or segregate into polarized sub-groups of homogenous opinion-holders (Dandekar et al., 2013; Duggins, 2016; Zanette \& Gil, 2006). Critically, this research has focused on groups of pre-existing opinion-holders. This work has yielded insights into belief-updating via repeated interaction (such as through the use of the Bounded Confidence Model; Deffuant, Neau, Amblard, \& Weisbuch, 2000), along with psychologically based models of behaviors including network pruning (Ngampruetikorn \& Stephens, 2015), which provides a plausible pruning mechanism of network contacts, based on a confirmation bias (self-regarding) principle.

The present paper interweaves elements of these two lines of literature, in conjunction with cognitive architecture brought forth from models of learning and communication in cognitive psychology. In particular, agents are encoded with three pieces of cognitive machinery: attention (detecting the public declarations of others); learning (incorporating a communication into a belief-state, and evaluating it against evidence); and decision-making (each choosing whether to make their opinion public based on a decision rule). In this way, all agents within the network are equally rational.

By focusing on universal cognitive architecture on the part of agents (and instead introducing stochasticity to the evidence against which an opinion is evaluated), this work argues that echo-chambers may result solely from the way in which networks are structured, and the spatiotemporal order of lateral opinion transference (i.e. an opinion cascade).

The semi-random way in which networks are structured (my relational position to the global network is random, but my method of forming my proximal (direct) connections is rule-based (those whom I know)), runs parallel to work on "small-worlds" (Watts \& Strogatz, 1998), which have shown susceptibility to cascades and synchronizability. As such, echo-chambers may occur without reliance on repeated interaction (Acemoglu \& Ozdaglar, 2011; Duggins, 2016), or individual differences encoded in agents, such as differences in susceptibility, or pre-dispositions towards an opinion (Watts \& Dodds, 2007) or hierarchy (see Quattrociocchi, Caldarelli, \& Scala, 2014).

## Agent-Based Modeling

ABMs are multi-agent, dynamic simulations which use combinations of three central components; agents, patches, and links. Agents are the individual actors within a model, and in the present paper, represent individuals within a network. Agents may be encoded cognitive rules (e.g., learning models), simple behaviors (e.g., signaling to
neighboring agents, movement), and values (e.g., prior beliefs, physical positioning). Agents are ascribed various forms of heterogeneity (such as occupying different positions within a network), as multiple agents are generated within the system. As the simulation runs, agents enact behaviors and update their values according to the specific rules ascribed to them, interacting with other agents and the environment accordingly.
Similarly, both links, which represent connections between agents, and patches, which represent the environment, may be encoded with behaviors and values, and the capacity to dynamically interact and update as the simulation runs. In the present paper, links are used to represent the connections between individuals within a network, and are thus used for signaling between agents. Given the network representation (requiring only agents and the links between them), the present model does not require the use of patches.

ABMs have been used to explore and assess how behaviors on an individual level, when placed within a dynamic, multi-agent, heterogeneous system, can lead to societal level, super-aggregate behaviors (Epstein, 1999, 2006; Schelling, 2006). For example, by encoding a preference in individuals to be neighbors with others who are similar (whether, on racial, socio-economic, or cultural lines), and assuming some stochasticity in signaling such similarity, Thomas Schelling (1971) was able to show the evolution of segregation on a community, and even citywide level. In a similar manner, the previously mentioned research on information cascades and opinion dynamics (Duggins, 2016) has used this technique to demonstrate a number of phenomena, with relatively few assumptions, that are difficult with traditional, equation-based cognitive modelling.

## A Model of Opinion Cascades

The aim of the current model is to provide a proof that the inherent structure of an online network is susceptible to high degrees of opinion segregation (i.e. false consensuses or echo-chambers). Critically, this segregation does not require repeated interaction, and can instead occur as a consequence of a single "cascade" across a network of rational agents (i.e. assuming no individual differences in cognitive architecture), despite equal proportions of opinion-holders on a global level.

A network of agents is created whereby agents are randomly assigned an XY coordinate, and each outfitted with the cognitive architecture and values outlined below. Each agent then forms links with its neighbors based on proximity in terms of Euclidean distance - representative of relational proximity in online networks (see Duggins, 2016). The number of links agents form is manipulated, and based on the percentage of the total number of agents in the system, from $.5 \%$, to $50 \%$. This is calculated by dividing the number of links per agent by the total number of agents in the network. Thus, given a population of 1000 agents, for an interconnectivity of $.5 \%$, all agents form links with their
nearest 5 neighbors; for $10 \%$, the nearest 100 agents, and so on. Accordingly, given a fixed population size across simulations, interconnectivity is manipulated via the number of links each agent possesses. In a similar manner, a neutral "event" node is placed in the geographical center of the simulation, and connected to the nearest agents according to the above rules for interconnectivity. Thus, increasing interconnectivity beyond $50 \%$ serves no purpose, given that every other agent will have been exposed to the neutral event (i.e. is $1^{\text {st }}$ generation), and thus no cascade can occur beyond two time points. Similarly, in the current model, interconnectivity below $.5 \%$ (i.e. 5 links per Agent) starts to risk fracturing the network into separate entities.

## Cognitive Architecture

Each agent is outfitted with simple cognitive architecture that can be classified into three branches: attention, learning, and propagation.

All agents within the network attend to their linkedneighbors, in that they are sensitive to the first of their neighbors to "declare" an opinion. Having seen such a declaration, the agent then moves into a learning phase to evaluate it.

The communicated opinion thus forms a prior for the evaluating agent. As mentioned previously, the opinions in the model are categorized into a binary division (Opinion A, Opinion B). Thus, from a neutral prior (.5), moving towards Opinion A is assigned a positive direction, whilst moving towards Opinion B a negative direction. In this way, a prior indicating Opinion A should shift the neutral recipient agent positively (e.g., $0.5+0.1=0.6$ ), and negatively for Opinion B. The strength of this shift is accordingly manipulated as a proxy of opinion strength / influence.

To represent the relationship between the strength of an opinion and the likelihood of a recipient adhering to that opinion, a learning model is used that allows agents to evaluate the opinion against stochastic evidence.
Specifically, a reinforcement learning model is used (Rescorla \& Wagner, 1972), in which agents evaluate an opinion in light of new evidence, such that the prediction error ( $\delta$ ), multiplied by the learning parameter ( $\beta$ ), is added to the value associated with the opinion (prior) for the current trial $(\mathrm{Q}(\mathrm{t}))$, leading to an updated opinion value $(\mathrm{Q}(\mathrm{t}$ $+1)$ ).

$$
\begin{equation*}
Q(t+1)=Q(t)+\beta \delta(t) \tag{1}
\end{equation*}
$$

Such models have been adapted (with added complexity) successfully to model the impact of instruction in reinforcement learning (Doll, Jacobs, Sanfey, \& Frank, 2009; Staudinger \& Büchel, 2013) and are thus considered a suitable placeholder for the proof of concept model. To evaluate the belief, agents then experience a number of evidence trials (arbitrarily set to 10), where evidence values are binary $\{0,1\}$, and are drawn with equal likelihood (i.e. $\mathrm{P}(\mathrm{E}=1)=.5)$. To reiterate, the learning process herein serves as a representation for the relationship between prior strength, and its likelihood of acceptance/rejection. Thus, if
the communicated opinion is represented by a weaker prior, it is more likely to be rejected by the learning / evaluation process. Similarly, increasing the amount of available evidence has the equivalent effect of converging the agent to the .5 (neutral) true state of the event (i.e. reducing the likelihood of passing on the original opinion). In this way, stronger opinions make the cascade more deterministic. Further, using a stochastic sampling process to dictate opinion uptake serves as a useful baseline model, to which complexity may be added directly to learning processes.

Having evaluated, agents declare for one of the two opinions, based on a decision rule: if Q (posterior) > .5, hold Opinion A; if $<.5$, hold Opinion B. This declaration is then made public (and thus may act as a prior to attending linkedneighbors) with a probability that is manipulated between systems. For example, a $\mathrm{P}($ Declaration $)$ of 1 means all agents will make their opinions public, whilst a P (Declaration) of .1 means there is a $10 \%$ probability of agents making their opinion public. This $\mathrm{P}($ Declaration $)$ bears a parallel to Watts and Dodds (2007) "individual threshold", found to impact spreading phenomena.

## Dynamics

Given the above architecture has been established, simulations commence with the initial, neutral "event" being witnessed by a portion of the network (based on manipulated interconnectivity). These agents (termed $1^{\text {st }}$ generation) start with a neutral prior, and so, based on the stochastic nature of the evidence, half should arrive at each opinion post-evaluation. From this point, if an agent of the $1^{\text {st }}$ generation makes their opinion public (based on manipulated $\mathrm{P}($ Declaration), their attentive (presently neutral) linked-neighbors ( $2^{\text {nd }}$ generation) then take this opinion as a prior, and evaluate it according to the procedure above. This $2^{\text {nd }}$ generation agents, having come to $a$ decision, then similarly each choose whether to make their opinion public (based on $\mathrm{P}\left(\right.$ Declaration)), and thus the $3^{\text {rd }}$ generation is exposed. This process continues until there has been no change in the number opinion-holders (of either type) for two consecutive time periods (i.e. if no one has made an opinion public, and thus the opinions have "died out", or if the network is now completely saturated).

Importantly, for the proof of concept model, having decided upon an opinion, an agent is no longer attentive to further information. This is purposeful to prevent cyclical effects beyond an initial cascade, as the goal of the present paper is to show the susceptibility of interconnected neutral agents to an opinion cascade, without resorting to explanations of homophily (Dandekar et al., 2013) and localized consensus reaching (Ngampruetikorn \& Stephens, 2015).

For the purpose of the present paper, the behaviors of interest are constrained to two, related measures. Firstly, the global proportion of opinions across the system (i.e. the proportion of agents with Opinion A , and the proportion with Opinion B) is of interest before inferring anything about localized clustering. For example, whether localized
clustering is simply a by-product of a dominant, networkwide opinion. This leads to the second measure: the average percentage of likeminded neighbors an agent possesses. In other words, of an agent's visible network, what percentage are in agreement with the agent (e.g., $50 \%$ indicates agents directly linked to equal proportions of each opinion-type). The manipulated variables are summarized in table 1 below:

Table 1. System variables

| Variable | Description | Levels |
| :--- | :--- | :--- |
| Interconnectivity (\%) | (Links per Agent / Total | $0.5,1,1.5$, |
| Opinion Strength | Agents in Network) $* 100$ <br> Added to (or subtracted from) | $\ldots 50$ |
|  | neutral agent prior $(\mathrm{P}(\mathrm{H})=.5)$ |  |
| P (Declaration) | Probability of making opinion <br> public | $0.1,0.5,1$ |

## Central Findings

The above model was implemented in NetLogo (5.2.1). Each system specification (Interconnectivity (100) x Opinion Strength (3) x P(Declaration) (3)) was run independently 100 times, taking an average set of values for each specification. The total number of agents in each simulation was set to 1000 .
Figs. 1a \& 1b show example outcomes of opinion cascades (A in red, B in blue) across a sparsely connected ( $1 \%$ interconnectivity) and a more densely connected (10\% interconnectivity) system, respectively.


Figures 1a and 1b: Sparsely and densely connected networks, post cascade (grey represents unused links).

Importantly, as Fig. 2 illustrates, irrespective of opinion strength, $\mathrm{P}($ Declaration ), or interconnectivity, the global proportion of different opinion holders consistently approximates 50/50.


Figure 2: Proportion of opinion holders across network


Figure 3: Degree of Clustering. Calculated as the average percentage of like-minded neighbors an agent possesses (panels represent $\mathrm{P}($ Declaration) conditions).

The degree of clustering (Fig. 3) can be broken down into several key findings. First, and central to the present paper, localized clustering increases as a function of decreasing interconnectivity and opinion strength, with stronger opinions and low interconnectivity ( $<1 \%$ ) resulting in the local (directly visible) networks of agents consisting of $>90 \%$ likeminded individuals ${ }^{2}$. Second, this effect occurs irrespective of the propensity for individuals within the network to make their opinions public ${ }^{3}$. In other words, whether $\mathrm{P}($ Declaration ) is at $100 \%$ or $10 \%$, localized clustering occurs regardless.
Finally, localized clustering is mitigated by the degree of stochasticity (i.e. as opinion strength moves towards neutral, thus having no communicative impact) and increasing interconnectivity. However, it is important to note that to prevent local clustering requires either no opinion impact or moving towards high (and arguably unrealistic) levels of interconnectivity.

## Discussion

The central finding of the present paper is that the fundamental way in which networks are constructed, when combined with the temporal dynamics of how information travels through them, and the cognitive representation of opinions as a prior, inherently leads to false consensus effects and echo-chambers. Thus, the more deterministic peer-to-peer communications are (i.e. how likely is a recipient to adopt the opinion of a sender), and the lower the relative density of connections within the network, the greater the impact of the spatiotemporal order (i.e. the larger the cascade sequence) on clustering. ${ }^{4}$

[^173]Critically, this effect occurs prior to any repeated interactions between agents, separating the present work from opinion dynamic literatures (Acemoglu \& Ozdaglar, 2011; Allahverdyan \& Galstyan, 2014), and without assuming individual differences on the part of agents (e.g., differences in susceptibility) or singular information types, common to information cascade literatures (Watts, 2002). Further, work in these areas including social network pruning (Ngampruetikorn \& Stephens, 2015) and polarization effects (Dandekar et al., 2013; Duggins, 2016), when looking at cyclical interactions, illustrate that repeated interaction is likely to only exacerbate the already high levels of localized clustering.

## False Consensus and Echo-chambers

The effects described in the present work are found to be broadly independent of the propensity to communicate, and robust across the degree of interconnectivity (requiring approximately $50 \%$ interconnective density to negate, something unfeasible in online networks approaching billions). Putting this into concrete terms, Facebook has an estimated 1.79 billion active users ${ }^{5}$. The average (median) number of "friends" or links is approximately $200^{6}$, meaning the average user is connected to $.000011 \%$ of the overall network. To fully negate the effects demonstrated here would require either the severance of lateral transmissions (or decreasing the deterministic capacity of communications sufficiently), or having each user share direct connections with approximately 900 million other users.

The formation of echo-chambers and the polarization of opinions is typically attributed to repeated interaction with a self-regarding preference (Ngampruetikorn \& Stephens, 2015) or a signaling of like-mindedness (e.g., trust; see Li, Scaglione, Swami, \& Zhao, 2013). This work instead shows that the structure of the network, and the way in which opinions emanate across it, are sufficient to result in false consensus effects and echo-chambers. To put this in more pragmatic terms; regardless of who you know, why you know them, or how you have repeatedly interacted / pruned your network, the fact that you do not, and arguably cannot know enough people, no matter who they are, is sufficient to leave you highly susceptible to echo-chambers.

It should be noted that this proof of concept model carries with it several assumptions. Most notably, opinions are classified in a binary fashion, so as to replicate the target
> opinion strength, then clustering severity is reduced. However, this relies on the strong assumption that there is independence of opinions across a self-selecting network. If one incorporates instead a degree of dependence in neighbouring opinion-holders, then one has in effect shifted echo-chamber formation to precede opinion transmission, and have thus "baked-in" the result.
${ }^{5}$ Figure taken from monthly active users as of the $3{ }^{\text {rd }}$ quarter of 2016. Source: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
${ }^{6}$ Figure taken from Pew Research Center survey of Facebook users in 2014. Source: http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/
opinion types under investigation, and associated with echochambers (e.g., referenda, or political campaigns). Future work is proposed to incorporate variance as they move across a network (i.e. do they dissipate, or become stronger). Secondly, agents attend and evaluate based on the first exposure to an opinion in their immediate network (i.e. those they are directly connected to). Although future work is suggested to incorporate the influence of multiple sources (e.g., via social conformity), such architecture is initially precluded to avoid "baking in" localized clustering effects.

Finally, the present model assumes a flat hierarchy of individuals. Although the argument can be made that fixing the level of interconnectivity for all individuals in a network is an artificial constraint, in terms of the degree of interconnectivity in target systems (e.g., Facebook) the functional difference in interconnectivity among users is between approximately $.000011 \% ~(200$ friends) and $.00028 \%$ ( 5000 friends; Facebook user limit). Although structural hierarchy, such as media influencers, may have an impact on dissemination (along with their own motives, such as following pre-existing opinion trends; see Quattrociocchi et al., 2014), the present work serves to illustrate that localized clustering can result from the spatiotemporal order of lateral transmission across a network.

## Further Work

The present work, in serving as a proof of concept for an increasingly important phenomenon, and providing some initial assumptions to illustrate the effects in a straightforward manner, leaves the door open for further, more psychologically informed modelling opportunities.

Further work should start to incorporate additional complexity on the part of agent (cognitive) architecture, such as the inclusion of social conformity (Latané, 1981), which is predicted to increase clustering tendencies (and feasibly increase the strength of opinions as they spread throughout the system. Similarly, work on confirmation bias suggests a similarly exacerbating role (Allahverdyan \& Galstyan, 2014; Doll et al., 2009; Nickerson, 1998; Staudinger \& Büchel, 2013). Finally, the inclusion of Bayesian models of source credibility (Harris \& Hahn, 2009; Harris, Hahn, Madsen, \& Hsu, 2015; Madsen, 2016) are of interest (Bayesian models of social learning have already started being applied to opinion dynamics; see Acemoglu \& Ozdaglar, 2011), given the way in which people form networks (i.e. we tend to select those we know / trust / like when forming our "direct" network). These suggestions are by no means exhaustive, but serve as examples of the promising (and readily applied) further additions to the framework laid out in the present work.

The present work purposefully precludes such psychological elements, which are expected to exacerbate the effects illustrated in this proof of concept model. This choice was made both for reasons of parsimony, and to provide a demonstration that the effects herein do not rely on such processes or explanations.

In conclusion, the present paper demonstrates that rational agents (i.e. absent special functionality of cognition or individual differences), through the way in which online networks are structured, are intrinsically susceptible to high levels of localized clustering (i.e. echo-chambers) when opinions are transmitted laterally. Further, it is hoped that the present paper serves as an example of how psychological principles taken from the individual level may be applied to a societal level through the use of AgentBased Models.

## Acknowledgments

This paper has benefitted from comments and discussions with Christos Bechlivanidis and Jens Koed Madsen.

## References

Acemoglu, D., \& Ozdaglar, A. (2011). Opinion Dynamics and Learning in Social Networks. Dynamic Games Applications, 1, 3-49.
Allahverdyan, A. E., \& Galstyan, A. (2014). Opinion dynamics with confirmation bias. PloS One, 9(7), e99557.
Amar, J. G., \& Family, F. (1995). Critical Cluster Size : Island Morphology and Size Distribution in Submonolayer Epitaxial Growth. Physics Review Letters, 74, 2066-2069.
Dandekar, P., Goel, A., \& Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. Proceedings of the National Academy of Sciences of the United States of America, 110(15), 5791-6.
Deffuant, G., Neau, D., Amblard, F., \& Weisbuch, G. (2000). Mixing beliefs among interacting agents. Advances in Complex Systems, 3(01n04), 87-98.
Del, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... Quattrociocchi, W. (2016). The spreading of misinformation online. PNAS, 113(3), 554-559.
Doll, B. B., Jacobs, W. J., Sanfey, A. G., \& Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. Brain Research, 1299, 74-94.
Duggins, P. (2016). A Psychologically-Motivated Model of Opinion Change with Applications to American Politics. ArXiv.
Epstein, J. M. (1999). Agent-based computational models and generative social science. Complexity, 4(5), 4160.

Epstein, J. M. (2006). Generative social science: Studies in agent-based computational modelling. Princeton University Press.
Essam, J. W. (1980). Percolation theory. Reports on Progress in Physics, 43, 833-912.
Harris, A. J. L., \& Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: incorporating the role of coherence. Journal of Experimental Psychology. Learning, Memory, and Cognition, 35(5), 1366-1373.

Harris, A. J. L., Hahn, U., Madsen, J. K., \& Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. Cognitive Science, 39(7), 1-38.
Hegselmann, R., \& Krause, U. (2002). Opinion Dynamics and Bounded Confidence. Simulation, 5(3), 2.
Latané, B. (1981). The psychology of social impact. American Psychologist, 36(4), 343-356.
Li, L., Scaglione, A., Swami, A., \& Zhao, Q. (2013). Consensus, Polarization and Clustering of Opinions in Social Networks. IEEE Journal on Selected Areas in Communications, 31(6), 1072-1083.
Madsen, J. K. (2016). Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions. In Proceedings of the 38th Annual Meeting of the Cognitive Science Society (pp. 165-170).
Meakin, P., Vicsek, T., \& Family, F. (1985). Dynamic cluster-size distribution in cluster-cluster aggregation: Effects of cluster diffusivity. Physical Review, 31(1), 564-569.
Ngampruetikorn, V., \& Stephens, G. J. (2015). Bias, Belief and Consensus: Collective opinion formation on fluctuating networks. arXiv Preprint arXiv:1512.09074.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2, 175-220.
Quattrociocchi, W., Caldarelli, G., \& Scala, A. (2014). Opinion dynamics on interacting networks: media competition and social influence. Nature, 4, 1-7.
Rescorla, R. A., \& Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory (2nd ed., pp. 64-99).
Schelling, T. C. (1971). Dynamic models of segregation. The Journal of Mathematical Sociology, 1(2), 143186.

Schelling, T. C. (2006). Micromotives and Macrobehaviour. W.W. Norton \& Company.

Staudinger, M. R., \& Büchel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. NeuroImage, 76, 125-133.
Watts, D. J. (2002). A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences of the United States of America, 99(9), 5766-5771.
Watts, D. J., \& Dodds, P. S. (2007). Influentials, Networks , and Public Opinion Formation. Journal of Consumer Research, 34(4), 441-458.
Watts, D. J., \& Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. Nature, 393, 440-442.
Zanette, D. H., \& Gil, S. (2006). Opinion spreading and agent segregation on evolving networks. Physica $D$, 224, 156-165.

# Make-or-break: chasing risky goals or settling for safe rewards? 

Pantelis P. Analytis ${ }^{1}$ (pp464@cornell.edu),<br>Charley M. Wu² (cwu@mpib-berlin.mpg.de), Alexandros Gelastopoulos ${ }^{3}$ (gelas@bu.edu)<br>${ }^{1}$ Department of Information Science, Cornell University<br>${ }^{2}$ Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development<br>${ }^{3}$ Department of Mathematics and Statistics, Boston University


#### Abstract

Humans regularly invest time towards activities characterized by dramatic success or failure outcomes, where criti-cally, the outcome is uncertain ex-ante. How should people allocate time between such make-or-break activities and other safe alternatives, where rewards are more predictable (e.g., linear) functions of time? We present a formal framework for studying time allocation between these two types of activities, and explore (optimal) behavior in both one-shot and dynamic versions of the problem. In the one-shot version, we illustrate the striking discontinuous relation between peoples skill and optimal time allocation to the make-or-break task. In the dynamic version, we formulate both fully rational and boundedly rational strategies, both defined by a giving up threshold, which adaptively dictates when one should cease pursuit of the make- or-break goal. Comparing strategies across environments, we investigate the cost of sidestepping the computational burden of full rationality.


## 6-page submission removed at author's request

6-page submission removed at author's request

6-page submission removed at author's request

## 6-page submission removed at author's request

6-page submission removed at author's request

# A computational decomposition of task-irrelevant perceptual learning 

Angelo Pirrone (apirrone1 @sheffield.ac.uk)<br>Department of Psychology \& Department of Computer Science, University of Sheffield, UK<br>Tom Stafford (t.stafford@sheffield.ac.uk)<br>Department of Psychology, University of Sheffield, UK


#### Abstract

Change in motion discrimination was assessed after seven days training on a rapid serial visual presentation training task, which included exposure to below-threshold coherent motion that was irrelevant to the task the participant was involved in. Post-training, participants had improved sensitivity for supra-threshold motion discrimination, which was specific to the direction exposed during training. A computational decomposition of the effect shows that the improvement is a combination of (i) an increase in rate at which participants accumulate evidence for the direction to which they were exposed and (ii) a decrease in their criterion for a response. Together with these differences (consistent across participants) other cognitive processes vary non-systematically between the pre-test and the post-test session making an analysis only based on accuracy or reaction times potentially misleading. Our analysis shows the benefits of isolating the different processes that are involved in perceptual decision making and are affected by perceptual learning.


Keywords: task-irrelevant perceptual learning; drift diffusion model; speed-accuracy trade-off

## Introduction

During task-irrelevant perceptual learning (TIPL), participants learn to better discriminate stimuli to which they are merely exposed but that are irrelevant to the purpose of the experiment (Watanabe et al., 2001; Seitz \& Watanabe, 2008). The first example of TIPL comes from Watanabe et al. (2001), in which authors, after testing participants on a motion discrimination task, exposed participants to many days of a rapid serial visual presentation (RSVP) task, on the background of which was presented a random-field motion stimulus with a below threshold coherence in a constant direction across all days of training. Watanabe et al. (2001) found that participants showed an improvement in a post-test motion discrimination task, only for the supra-threshold coherence level in the direction to which they were exposed. Results from TIPL research suggest that the brain, even in the presence of subliminal exposure, adapts to specific features of the environment. As discussed in Watanabe et al. (2001), this result has important implications for modern everyday life in which we are constantly subject to high amounts of information that we try to ignore. TIPL research suggests that such ignored information could still affect our behaviour.

Focusing only on RT or accuracy for a perceptual task has some limitations, since different components contribute to those aspects of the decision which can be eas-
ily measured, RT and accuracy: (1) the difficulty of the decision, (2) the decision criteria adopted by the participant, which can be more or less conservative, (3) the nondecision time, which includes time to encode the stimulus and to execute the motor response, (4) the bias towards a response, (5) across trials variation in the above mentioned mechanisms. If the participant is performing a task on many different days, since the experimenter is interested in the effect of learning over different sessions, it is reasonable to expect that all the above mechanisms could also vary across different days on the basis of factors not related to the experiment (e.g., on one day the participant could be more tired or less collaborative). In some cases, an approach focusing only on RTs or accuracy has even been misleading in generating theories from data (for a discussion and examples, see Ratcliff \& McKoon, 2008).

Here, we use the Drift Diffusion Model (DDM; Ratcliff \& McKoon, 2008) as a tool to isolate different components of the processes that contribute to a decision. The DDM (Ratcliff \& McKoon, 2008) is a computational model of decision making that, over the years, has been applied in a wide variety of tasks (for a review see Ratcliff \& McKoon, 2008). In the DDM the decision maker accumulates evidence supporting two alternatives until a threshold for a decision is reached. In its arguably most used formulation (Ratcliff \& McKoon, 2008), the DDM is composed of seven parameters. The first, denoted by $a$, is the boundary separation and it describes the distance between the two decision boundaries. This parameter is related to the speed-accuracy trade-off. When the boundaries are near, the decision is faster but less accurate; conversely, when the distance between the two boundaries increases, the decision is slower but more accurate. The rate at which noisy information is accumulated over time within a trial is defined by the drift rate, $v$. This parameter reflects the difficulty of the task with respect to the sensitivity of the observer, with easier conditions (and/or more sensitive observers) resulting in faster and more accurate decisions. When the decision maker starts to integrate difference in evidence at an equal distance from the two decision boundaries, the process is defined as unbiased. However, the decision maker can start to integrate evidence nearer one of the two boundaries so that a third parameter affect the decision, the starting point
of evidence accumulation, $z$. When the decision maker is biased towards one of the two boundaries, fast RTs for that boundary but slow RTs for the opposite boundary are predicted, and at the same time the decision is more likely to end at the nearer boundary. The decision time is the time required to cross one of the two boundaries; however, for each reaction time it is present also a component that incorporates the non-decision time component of a decision, ter, which is the time to encode the stimulus and execute the motor response (e.g., pressing a button on the keyboard). Three further parameters of the DDM are: inter-trial variability in drift rate defined as $e t a$, in the starting point, $s z$, and in the non-decision time, st.

Despite widespread success in other domains, the DDM has not been consistently applied in the domain of perceptual learning. In Petrov et al. (2011), the authors performed a DDM decomposition of a fine motion discrimination task. In their study, authors found that perceptual learning was best explained by an increase in drift rate, a decrease in boundary separation and a decrease in both the non-decision time component and its inter-trial variability. In Liu \& Watanabe (2012), participants performed a three day perceptual learning coherent motion direction task (i.e., is the random dot kinematogram all noise or is there some signal?) and authors found an improvement in drift rate but with boundary separation decreasing across the days of training. In Dutilh et al. (2009), participants performed a 5 days lexical decision task and authors found that the learning led to an increase in drift rate, a decrease in boundary separation, as well as a significant decrease in non-decision time. As it is clear from these investigations, RTs and accuracy alone cannot give a full description of the cognitive processes that are most likely to have generated the data. Considering both measures and their distribution can lead to a better understanding of the cognitive processes involved in such tasks. In particular, if perceptual learning is associated with a decrease in the boundary separation, as other studies of task-relevant perceptual learning have found, then assessing perceptual learning via measurement of accuracy will systematically underestimate the true size of perceptual learning (since decreased boundary separation will tend to decrease accuracy, all other things being equal).

Although task-relevant perceptual learning has been decomposed by using the DDM, to our knowledge, no studies have focused on a DDM decomposition of TIPL. In our study, we ran an experiment similar to that presented in Watanabe et al. (2001), and we performed a DDM decomposition of TIPL. Because of the aforementioned multi-component nature of perceptual decisionmaking, our expectation is that use of the DDM will allow a more accurate assessment of perceptual learn-
ing than attention to solely RT or accuracy. Further, the DDM allows us to isolate the component of decision making that reflects a true change to stimulus sensitivity - a change in the drift rate parameter. Because of the potential for non-stimulus related parameters to alter across sessions due to non-experimentally caused factors (such as fatigue or motivation) and because, by their nature, perceptual learning experiments involve testing participants on different days or even weeks, we isolate perceptual learning as an increase in the drift rate for the exposed stimulus (for one participant also the drift of the not exposed direction increased but the increase was greater for the exposed stimulus than for the not exposed stimulus). In this way, we use each participant as their own control, testing them for changes in perceptual decision making for both exposed and not exposed stimuli and thus accommodating non-training related changes in decision parameters.

## Material and methods

Participants Four right-handed healthy university students (2 females, ages: 30, 21, 20, 22 years), with no history of neurological or psychiatric disorders, with normal vision and naïve to the purpose of the study participated voluntarily in the experiment and received a compensation of $£ 50$ for their participation. The experiment was approved by the University of Sheffield, Department of Psychology Ethics Sub-Committee, and carried out in accordance with the University and British Psychological Society ethics guidelines. Participants gave their informed consent.

Apparatus The stimuli were generated on a personal computer using PsychoPy (Peirce, 2009). During the whole experiment, participants had to put their head on a chin rest at a viewing distance of 57 cm from a SONY Multiscan CPD-200ES 17" monitor with a resolution of $1280 \times 1024$ pixels at a refresh rate of 60 Hz . The experiment was conducted under binocular viewing conditions and participants used a keyboard to make a response.
Motion-Direction Stimuli We used stimuli similar to those adopted by other studies on task-irrelevant perceptual learning (Watanabe et al., 2001; Seitz \& Watanabe, 2008): on a grey background, within a black annulus aperture of $1^{\circ}-10^{\circ}$, white dots with a size of $2 \times 2$ pixels were moving with a speed of $6 \%$ and a density of 16.7 $d o t s / d e g^{2} / s$ on a black background. Signal dots were randomly chosen in each frame, and on each frame, noise dots had a random position. Dots had a limited lifetime of three frames after which they were redrawn in random locations. If any of the signal dots were to move out of the annulus, they were replaced randomly in the stimulus field. The stimuli were generated in real time and two non-cardinal directions were employed in this study, $45^{\circ}$
and $135^{\circ}$.
Procedure The experiment consisted of nine sessions; a pre-test to measure sensitivity for various strengths of motion coherence in the two directions, then seven training sessions consisting of a RSVP task with on the background a random dot motion, and finally a post-test that measured sensitivity for various coherence levels in the two directions, that was equal to the pre-test. Participants came on different days for each session, and could take a maximum of three-days break between sessions.

Pre/post Motion-Direction Sensitivity Tests Participants were instructed to pay attention to the stimulus that would be presented for 500 ms and then report as quickly and accurately as possible if the coherent motion was towards up-left ( $45^{\circ}$ on the left with respect to an imaginary vertical reference line) or up-right ( $45^{\circ}$ on the right with respect to a imaginary vertical reference line) by button press. They were instructed to use their right hand index finger to press left on the keyboard for 'up-left', and their middle finger to choose 'up-right'. Participants were instructed that there was always a correct response and were required to fixate the cross at the centre of the screen during the whole task and minimise as much as possible eye movements. In each trial, a fixation cross in the central circle was presented for 333 ms , followed by the presentation of moving dots for 500 ms , followed by two arrows showing the possible direction of the dots and the text 'Answer:' presented on top of the screen until participants made a response. Each test stage consisted of 10 blocks x 2 directions ( $45^{\circ}$ and $135^{\circ}$ ) x 10 motion coherence levels ( $5 \%, 10 \%, 20 \%, 30 \%, 40 \%, 50 \%, 60 \%$, $70 \%, 80 \%, 90 \%) \times 9$ repetitions for a total of 1800 trials and took about 1 hour to complete. Coherence levels were chosen so that for each direction we would have accuracy levels that range from chance to ceiling based on the results of previous pilot studies. During each block the order of presentation of trials was randomised and no accuracy feedback was given to the participants. After a block of 180 consecutive trials participants were required to take a self-paced break to rest before continuing with this task.

Training sessions In the training sessions, participants performed a RSVP character identification task and were asked to report in order of presentation two white capital letters (height about $.9^{\circ}$ ) in a sequence of 10 capital letters presented in the central circle. All letters of the alphabet were used. Distractors consisted of eight capital black letters. The first and second white letters were presented in one of the first five serial positions and in one of the second five serial positions, respectively. They were determined randomly in each trial. Within the annulus aperture of $1^{\circ}-10^{\circ}$ participants were presented a mo-
tion stimulus in one of the two directions, constant across all training sessions, at a coherence level 5\% below their chance level at pre-test, in order to ensure a level reasonably below threshold. For each participant, we computed the motion strength at $50 \%$ accuracy by interpolating the psychometric curve predicted by the model free estimation of the psychometric curve described in Zchaluk \& Foster (2009), and using MATLAB scripts made available by those authors. In each trial, a fixation cross in the central circle was presented for 333 ms followed by the presentation of the stimulus for 500 ms followed by a grey screen and the text 'Type in the two white letters' presented on top of the screen until participants responded. Each test stage consisted of 10 blocks x 108 repetitions for a total of 1080 trials and took about 45 minutes to complete. No accuracy feedback was given to the participants. After a block of 108 consecutive trials, participants were required to take a self-paced break to rest before continuing with this task.

## Results

## Behavioural analyses

The performance of participants in the RSVP was mostly stable across the seven days of training. We did not perform any analysis on the RSVP task as our interest is in the TIPL, hence in the change in performance between pre-test and post-test for the exposed and not exposed directions.

For all analyses, both behavioural analyses and model fitting, we removed, for each participant, the $2.5 \%$ of slowest responses, given that a first inspection of data showed the presence of extremely slow outliers. In the following analyses, each subject is analysed separately. Participants 1 and 2 were exposed to $45^{\circ}$ while participants 3 and 4 were exposed to $135^{\circ}$. T-tests were conducted to investigate overall differences for each participant between the pre-test and the post-test in mean RT and accuracy levels for the exposed and the not exposed directions. Bonferroni corrections were applied on the p-values.

All subjects had a significant decrease in RTs between the pre-test and post-test, for the exposed and the not exposed direction ( $\mathrm{p}<.001$ in all cases). Participant 1 did not have a change in accuracy for the exposed direction between the pre-test and post-test $(\mathrm{t}(9)=-0.34, \mathrm{p}=1)$ while all other participants had a significant increase ( $p$ $<.001$ in all cases). Regarding accuracy of the not exposed direction, there was not a significant change between pre-test and post-test for any subject ( $\mathrm{p}>.07$ in all cases).

## Model fitting

For fitting the diffusion model to RT distributions and proportion of correct and incorrect responses, we used
the Diffusion Model Analysis Toolbox (DMAT; Vandekerckhove \& Tuerlinckx, 2007) for MATLAB. Parameters were estimated by using as the objective function a multinomial likelihood function, which expresses the likelihood of observing a certain proportion of responses in a given number of RT bins and is maximised in order to find the parameter estimates. We decided to represent the reaction time distributions of responses in terms of the classical $.1, .3, .5, .7$ and .9 quantiles that divide the RT distribution. For each participant we fitted a model in which the drift rates were free to vary across all conditions while all other parameters were fixed across conditions within the pre-test and the post-test but could vary between pre-test and post-test. In this way, for each session, a model with 26 parameters was fitted for each individual. However, the high number of parameters, rather than over-fitting, reflects the high number of conditions of this experiment. In fact, the only stimulus-contingent parameter allowed to vary is the drift, which varies according to the difficulty of each coherence level and according to the direction (exposed vs not exposed). Since participants were presented with trials in random order they could not adjust their boundary separation or their starting point of evidence accumulation before the presentation of each trial, hence the assumption of constant boundary and starting point parameters within each session is reasonable, together with stimulus-independent variability in starting point across trials. We assumed as constant the non-decision time component (i.e., stimulus encoding and motor response) between the two directions since it is unlikely that subject would have higher non-decision time (e.g., pressing a button on the keyboard) for one direction compared to the other. DMAT allows to calculate estimates of the parameters' standard errors. For each participant, we performed Wald tests for the difference in parameters between pre-test and posttest using the parameter estimates and their standard errors. Also here, results are presented by participant, Figure 1. Here we focus on within-subjects variation, hence for visibility we do not report consistent scaling and range across the same parameters fitted to different subjects.

Participant 1 had a significant decrease in boundary separation ( $Z=-7.38, p<.001$ ), in non decision time ( $\mathrm{Z}=-11.46, \mathrm{p}<.001$ ) and in starting point $(\mathrm{Z}=-10.36$, $\mathrm{p}<.001$ ). Regarding drift rates, t -tests showed an increase in drift for the exposed direction $(t) 9)=-2.75$, $\mathrm{p}=.045$ ), while the drift of the not exposed direction did not vary between pre-test and post-test $(t)(9)=-.42$, $p=$ 1). Participant 2 had a significant decrease in boundary separation ( $Z=-13.14, p<.001$ ), non-decision time ( $Z=$ $-6.64, \mathrm{p}<.001$ ) and starting point ( $\mathrm{Z}=-7.15, \mathrm{p}<.001$ ). This subject had an increase in drift for the exposed direction $(\mathrm{t}(9)=-3.36, \mathrm{p}=.02)$, while the drift of the not
exposed direction did not vary significantly between the pre-test and the post-test $(\mathrm{t}(9)=-1.15, \mathrm{p}=.56)$. Participant 3 had a significant decrease in boundary separation ( $\mathrm{Z}=-9.18, \mathrm{p}<.001$ ), an increase in non-decision time ( $Z$ $=1.98, \mathrm{p}=.02)$ and a decrease in starting point $(\mathrm{Z}=-$ $5.46, \mathrm{p}<.001$ ). Between the pre-test and the post-test, the drift of the exposed direction increased significantly $(\mathrm{t}(9)=-4.83, \mathrm{p}=.002)$ while the drift of the not exposed direction stayed the same $(\mathrm{t}(9)=-1.61, \mathrm{p}=.28)$. Participant 4 had a significant decrease in boundary ( $\mathrm{Z}=$ $-5.81, \mathrm{p}<.001$ ), an increase in non decision time ( $\mathrm{Z}=$ 2.33, $\mathrm{p}=.01$ ) and a decrease in starting point. $(\mathrm{Z}=-1.37$, $\mathrm{p}=.08$ ). The drift of the exposed direction increased significantly $(\mathrm{t}(9)=-6.25, \mathrm{p}=<.001)$ as well as the drift of the not exposed direction $(\mathrm{t}(9)=-3.85, \mathrm{p}=.01)$. Although the drift rates for both directions increased between pre-test and post-test, the relative change between pre-test and post-test for the exposed direction was significantly higher than the relative change between pretest and post-test for the not exposed direction $(t)(9)=$ $3.72, \mathrm{p}=.01$ ). None of the participants had a significant change between the two sessions in the parameters capturing across trials variability in drift, starting point or non-decision time. The goodness of fit of the model was assessed graphically through quantile-probability plots (not shown for brevity). The quantile probability plots showed that the model on which our analyses are based fits the data well and without mismatches.

## Discussion

Here, using the DDM (Ratcliff \& McKoon, 2008), we have modelled for the first time the processes underlying TIPL in healthy individuals. The results indicate that: (i) TIPL affects the drift rate at which participants accumulate evidence for the exposed direction meaning that their sensitivity for the exposed direction is increased, (ii) TIPL affects the conservativeness of participants' response (iii) non-systematic variations (i.e., the direction of the change varied between subjects) in parameters between the two sessions (e.g., variations in non-decision time, variations in starting point) do not allow a direct comparison of the decision process only based on accuracy and/or RT.

These findings have important implications for the interpretation of perceptual learning data, both taskrelevant and task irrelevant, and, we hope, for the analysis of data collected on different days or for which learning is involved. First, every decision is a mixture of different cognitive processes that can be isolated by this analysis for a more principled interpretation of results. Interpreting learning in terms of latent cognitive variables allows for a more precise investigation of its effect and gives a proper measure of 'true' perceptual learning - change in the drift rate which is related to the qual-


Figure 1: DDM parameters for each participant: top-left (Participant 1), top-right (Participant 2), bottom-left (Participant 3), bottom-right (Participant 4). For each participant, in order from top left to bottom right: boundary separation (a), non decision time (ter), starting point ( z ) and drift rates (v). The x -axis for the drift rate reports, for both the exposed and the not exposed directions, the 10 coherence levels by decreasing difficulty. Bars are standard errors of parameters' estimates.
ity of input information - while weighting for systematic or random variations in other parameters. In particular, caution should be exercised when comparing data across different sessions. As in previous studies of perceptual learning, participants in our study showed evidence of a change in their speed-accuracy trade-off. Not taking this factor into account can lead to wrong conclusions from data. In theory, decreased boundary should result in decreasing accuracy for the not exposed direction. In our investigation, participants had to view the stimuli for 500 ms before giving their response; given this constraint, participants had a relatively long time window to make a decision and as a consequence this might have obscured a decrease in accuracy between the pre-test and the posttest that should result from a decreased boundary. For future investigation, we recommend using a shorter presentation of the stimuli (e.g., 200 ms ) that is more likely to reveal stronger variations in accuracy and RTs of the not exposed direction between the pre-test and the posttest.

Our results show the risk of directly comparing sessions performed on different days without considering the role of each single parameter. Take for example participant 1: by testing for differences in accuracy, a researcher may be tempted to conclude that this subject did not have any TIPL since there is not a difference be-
tween the accuracy of the first and second session both for the exposed and the not exposed direction. However, the model fitting shows that this participant had higher drift rates for the exposed direction, which is likely to be the signature of TIPL, which is accompanied by a decrease in boundary and variations in the bias towards a response. An increase in drift (faster and more accurate responses) accompanied by a decrease in boundary (faster and less accurate responses) can have as output that accuracy levels stay the same.

Previous studies have shown that perceptual learning is associated with a decrease in boundary separation (Petrov et al., 2011; Liu \& Watanabe, 2012; Dutilh et al., 2009) and we replicated this result also here for TIPL showing consistency across four participants. It has been proposed (Liu \& Watanabe, 2012) that this decrease in boundary separation is due to the fact that participants are trying to maximise their reward rate, operatlionalised as the proportion of correct responses divided by the average time between them. In other words, if the quality of information increases (i.e., hence the task becomes 'easier') participants can decrease the time spent on each decision in order to finish the task sooner without sacrificing accuracy too much given the increase in drift rate. To the best of our knowledge, this is the first study to report a DDM decomposition of TIPL and the first study
to show the systematic parameter variations associated with TIPL. Regarding other parameters there is not consistency in the literature regarding the effects of learning, and also here we do not observe a clear pattern across participants. For example, regarding the non-decision time component, in previous studies investigating perceptual learning, Petrov et al. (2011) and Dutilh et al. (2009) found a decrease associated with learning, while Liu \& Watanabe (2012) found that, although not significant overall, some participants showed an improvement. Here we did not find a consistently decreasing non-decision time component between the two sessions, given that only two out of four participants have a decrease in non-decision time. Our only consistent result is that of decreasing boundary related to learning and an effect on the drift; result that shares some similarities with that by Liu \& Watanabe (2012). Regarding the drift, we show that there is an increase in the drift of the exposed direction, compared to the pre-test, and compared to the drift of the not exposed direction of the post-test when the drift of the not exposed direction increases as well in the post-test. Ideally, we would expect that the drift of the not exposed direction would not vary between pre and post-test. For one participant however the drift of the not exposed direction varies as well; this is unlikely to be an effect of TIPL but rather a random variation in participants' performance that further highlights the importance of a DDM decomposition of learning data.

Although the sample size $(\mathrm{N}=4)$ is low, this is in line with similar studies that have performed a DDM decomposition of learning data (e.g., Dutilh et al., 2009), and it is common practice in perceptual learning research. Furthermore the consistency in results across participants reassures us about our conclusions. It is to be mentioned that the training that our participants performed is relatively 'short' if compared with the usual TIPL training of about 20 days, during which TIPL reaches its asymptotic level (Watanabe et al., 2001). Future work, employing more participants and longer training regimes is clearly warranted in order to quantify the rate at which each component is affected by learning, and to quantify the distortion that focusing only on accuracy could lead to; if the days of training increase, the effect on boundary and drift reported here might increase and have even stronger consequences on accuracy and RT.

Here we used the DDM as a model of decision making and in order to analyse our data, but it is to be mentioned that other models of decision making could have been applied (Bogacz et al., 2006). However, given the model mimicry between models of decision making (Bogacz et al., 2006), we predict that our results are likely to be replicated if another model is applied.

Overall, a consideration of the different components in decision making shows that the two components which
are found to vary systematically all have independent effects on speed and/or accuracy. Whilst increased drift will tend to increase speed and accuracy, decreased boundary separation will tend to decrease both. For these reasons, a decomposition of decision making from these observed variables allows us not only to focus on the different effects of perceptual learning individually, but allows us a more accurate assessment of the extent of increased stimulus sensitivity in perceptual learning. Our study is the first to show this increased sensitivity in taskirrelevant perceptual learning, and does so demonstrating that the other components of decision making are affected in a similar way to as in task-relevant perceptual learning.

## Acknowledgments

A.P. was supported by the University Studentship Network in Neuroeconomics. This work appeared in an extended version in A.P. PhD dissertation.

## References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., \& Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review, 113(4), 700.
Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., \& Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. Psychonomic Bulletin \& Review, 16(6), 1026-1036.
Liu, C., \& Watanabe, T. (2012). Accounting for speedaccuracy tradeoff in perceptual learning. Vision research, 61, 107-114.
Peirce, J. W. (2009). Generating stimuli for neuroscience using psychopy. Frontiers in Neuroinformatics, 2, 10.
Petrov, A. A., Van Horn, N. M., \& Ratcliff, R. (2011). Dissociable perceptual-learning mechanisms revealed by diffusion-model analysis. Psychonomic bulletin \& review, 18(3), 490-497.
Ratcliff, R., \& McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. Neural computation, 20(4), 873-922.
Seitz, A. R., \& Watanabe, T. (2008). Is task-irrelevant learning really task-irrelevant? PloS one, 3(11), e3792.
Vandekerckhove, J., \& Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. Psychonomic bulletin \& review, 14(6), 1011-1026.
Watanabe, T., Náñez, J. E., \& Sasaki, Y. (2001). Perceptual learning without perception. Nature, 413(6858), 844-848.
Zchaluk, K., \& Foster, D. H. (2009). Model-free estimation of the psychometric function. Attention, Perception, \& Psychophysics, 71(6), 1414-1425.

# Inferential Pitfalls in Decoding Neural Representations 

Vencislav Popov (vencislav.popov@gmail.com)<br>Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Markus Ostarek (markus.ostarek@mpi.nl)
Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
Caitlin Tenison (ctenison@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA


#### Abstract

A key challenge for cognitive neuroscience is to decipher the representational schemes of the brain. A recent class of decoding algorithms for fMRI data, stimulus-feature-based encoding models, is becoming increasingly popular for inferring the dimensions of neural representational spaces from stimulusfeature spaces. We argue that such inferences are not always valid, because decoding can occur even if the neural representational space and the stimulus-feature space use different representational schemes. This can happen when there is a systematic mapping between them. In a simulation, we successfully decoded the binary representation of numbers from their decimal features. Since binary and decimal number systems use different representations, we cannot conclude that the binary representation encodes decimal features. The same argument applies to the decoding of neural patterns from stimulus-feature spaces and we urge caution in inferring the nature of the neural code from such methods. We discuss ways to overcome these inferential limitations.


## Introduction

A key challenge for cognitive neuroscience is to decipher the representational schemes of the brain, to understand the neural code that underlies the encoding and representation of sensory, motor, spatial, emotional, semantic and other types of information. To address these issues researchers often employ neuroimaging techniques like functional magnetic resonance imaging (fMRI), which measures the blood oxygenation leveldependent (BOLD) activation in the brain that is elicited when participants engage with different stimuli. A common assumption has been that the underlying neural representation of each stimulus has measurable but complex effects on the BOLD activation patterns. In order to understand what those patterns of activity can tell us about how the brain processes and represents information, researchers have used various analytical tools such as univariate subtraction methods, multivariate pattern (MVP) classification, representational similarity analysis (RSA) and, recently, explicit stimulus-feature-based encoding and decoding models (for reviews, see Davis \& Poldrack, 2013, Haxby, Connolly, \& Guntupalli, 2014, or Naselaris, Kay, Nishimoto, \& Gallant, 2011). Despite their differences, these methods aim to quantify how changes in task conditions and the properties of the stimuli relate to changes in BOLD activation and vice versa. Where these methods differ is in how they achieve that mapping and in what inferences they allow us to draw.

In this article, we review some of the known inferential limitations of existing fMRI analysis methods and we highlight a previously unrecognized issue in interpreting results from stimulus-feature-based encoding and decoding models. The latter are steadily becoming the de facto gold standard for investigating neural representational spaces (Haxby et al. 2014, Naselaris \& Kay, 2015).

## Univariate vs. multivariate analysis

Before the advent of the more advanced techniques we review below, the main fMRI analysis tool was based on comparing how activity in a single voxel or averaged activity in a contiguous area of voxels differs between task conditions or stimuli. These univariate subtraction methods have been informative about the relative engagement of certain brain areas in specific tasks. Unfortunately, the coarse nature of this method precludes fine-grained inferences about the underlying representational content and computations that give rise to the observed BOLD signal. By ignoring the possibility that information might be represented in a distributed manner across voxels, the assumptions underlying univariate subtraction methods limit their use in understanding neural representations. In addition, these methods cannot tell us whether changes in activation are due to representational preferences, processing differences, or attentional variation among conditions (Coutanche, 2013).

In contrast, multivoxel pattern analysis (MVPA) techniques have attempted to overcome this limitation by looking at how various categories of stimuli or task conditions lead to differences (i.e. MVP classification) or similarities (i.e. RSA) in distributed patterns of activity over multiple voxels. These methods have become popular because they allow researchers to study neural representational spaces with increasing sensitivity and resolution. For example, a seminal study by Haxby et al. (2001) found that visual object categories can be classified based on the pattern of activation that their exemplars elicited in the ventral temporal cortex. The classification was successful despite the lack of overall activation differences in that region. Similar methods have been used to show that concepts have language-invariant representations in the anterior temporal lobe (Correia et al., 2014), that very similar visual scenes can be discriminated in the hippocampus (Bonnici et al., 2012) and that during their retrieval from memory, the shape, color and identity
of visual objects can be differentially decoded across several cortical areas (Coutanche \& Thompson-Schill, 2015).

Despite early enthusiasm that MVPA methods could be used to understand the structure of the neural code and the nature of the underlying representations (Norman, Polyn, Detre, \& Haxby, 2006), conventional MVP classification and RSA techniques share one of the same fundamental inferential limitations of univariate methods. Successful classification or careful inspection of confusions/similarity matrices can indicate that some relevant information about the stimulus class is present in the population of analyzed voxels, but it cannot identify exactly what that information is, or how it is represented and organized (Naselaris \& Kay, 2015; Poldrack, 2011; Tong \& Pratte, 2012). Because neural data is correlational, many different properties of the stimuli might lead to successful classification of the stimulus category, the task condition, or the brain state in question. For example, successfully categorizing whether a word represents an animate or an inanimate object does not necessarily mean that the region of interest encodes that category distinction. There are many differences between animate and inanimate objects, such as differences in their sensory and functional features (Farah \& McClelland, 1991) that could be responsible for the successful classification.

Another limitation of conventional MVP classifiers is that they cannot generalize and predict behavioral responses to novel types of stimuli or task conditions. To understand why, we can conceptualize classifiers in terms of types and tokens. An MVP classifier is usually trained on stimuli that are tokens from several types. For example, the stimuli tokens might be different category exemplars, and the classifier is trained to predict the type of category to which they belong. Alternatively, the tokens might be multiple presentations of the same word in different modalities or languages and the types are the unique words themselves. In the first case, the classifier can only be used to predict category membership of words that belong to one of the categories on which it was trained. In the second case even though the classifier could be used to predict exemplars in novel languages or modalities, it is again restricted only to exemplars of the words on which it was trained in the first place. In general, while the tokens being tested might be novel, they will be potentially decoded only if they are exemplars of a type that has already been trained on.

For example, if one trains a classifier to predict the color of objects and trains it on yellow and orange objects (Coutanche \& Thompson-Schill, 2015), one will not be able to predict the color of novel objects that are green. This methodological limitation is important - just as understanding how the decimal system represents numbers allows people to understand and manipulate numbers they have never seen before, a complete understanding of any neural representational system should allow researchers to use the neural pattern associated with novel stimuli to predict their identity, even if those stimuli are not exemplars of the types on which a particular model was trained on.

## Stimulus-feature-based encoding models

To overcome this limitation many researchers are turning to a novel analysis method that is known by a few different names voxelwise modelling (Naselaris \& Kay, 2015), stimulus-model based encoding and decoding (Haxby et al., 2014), voxel-based encoding and decoding models (Naselaris et al., 2011), and forward models (Brouwer \& Heeger, 2009; Fernandino, Humphries, Conant, Seidenberg, \& Binder, 2016). This approach can decode the identity of novel types of stimuli from neural activity by predicting activity not for the stimuli themselves, but for a set of simpler features into which they can be decomposed. In a seminal study, Mitchell et al. (2008) predicted the neural activity associated with individual novel words based only on the activation of other words. To achieve that, they decomposed each word into a vector of weights on 25 sensory-motor semantic features (verbs such as "eat", "taste", "run", "fear", etc.). The weights were estimated from cooccurrence statistics of the word with each verb feature in a large corpus. They trained a classifier to predict the neural activity associated with each constituent feature of a training set of words, which resulted in separate neural activation maps for each feature. Neural activity for novel test words was then predicted highly accurately as a linear combination of the semantic feature activation maps weighted by the association of the word with each feature. Based on these results, Mitchell et al. (2008) concluded that the neural representation of concrete nouns might be based on sensory-motor features.

Similar approaches have been used to predict the neural response to novel natural images using Gabor filter features (Kay, Naselaris, Prenger, \& Gallant, 2008), to novel colors based on color tuning curve features (Brouwer \& Heeger, 2009), to novel music clips based on acoustic timbre features (Casey, Thompson, Kang, Raizada, \& Wheatley, 2012), to natural sounds based on frequency, spectral and temporal modulations (Santoro et al., 2014), to novel faces based on a PCA decomposition of face features (Lee \& Kuhl, 2016), to novel words based on subjective sensory-motor ratings (Fernandino et al., 2016). The motivating question behind the majority of these studies has been about the nature of the representations used by the brain in encoding the experimental stimuli, and the results have been used to argue that the neural representation is based on the constituent features of the stimuli used in the model.

To summarize, stimulus-feature encoding models generally use the following analysis procedure: 1) Specify a set of features and dimensions that hypothetically underlie the representation of a stimulus set in brain. 2) Decompose a set of stimuli into vectors of weights for each feature. 3) Select a region of interest (ROI) in the brain from which to analyze neural activation. 4) Train a model to predict activity in each voxel for a training set of stimuli, using the weights of their features as predictors. 5) Derive activation pattern maps (e.g. regression coefficients) associated with each feature. 6) Predict neural activity in the ROI for novel stimuli, based on their feature weights and the activation pattern maps for each feature. 7) Compare predicted neural activity for each novel stimulus with their observed neural activity and derive a measure of fit and accuracy. In essence, stimulus-feature-based encoding models attempt to map a
stimulus feature representational space, where each feature is a separate dimension, and each stimulus is a point in that space, to a neural activation space, where each voxel is a separate dimension, and the activation pattern elicited by each stimulus is a point in that space.

## What can we infer about neural representations?

What can a successful mapping between a stimulus feature space and a neural activation space tell us about the nature of the representation used by the brain? A common inference in these studies has been that if you can predict the identity of novel stimuli based on that mapping, then the neural representation is likely based on the feature set used by the model. Put formally, the inferential claim goes as follows:

1) We can represent certain stimuli as a combination of lower-level features
2) We can show that it is possible to predict the neural pattern caused by a novel stimulus in brain area A from an encoding model based on these features
3) Therefore, brain area A encodes those features and uses a representational scheme based on them.
This claim has been made to different degrees both in theoretical and methodological papers reviewing the approach (e.g., Haxby et al., 2014; Naselaris \& Kay, 2015; Naselaris et al., 2011; Norman et al., 2006; Tong \& Pratte, 2012), as well as in empirical studies that use it to address representational questions (Fernandino et al., 2016; Kay et al., 2008; Mitchell et al., 2008; Santoro et al., 2014; although some are more cautionary, e.g. Lee \& Kuhl, 2016). If this inference is valid, then encoding models could be an extremely powerful tool for understanding the nature of neural representations.

A useful illustrative example of this inference in practice comes from a recent study by Fernandino et al. (2016). The authors wanted to understand how conceptual information is represented in a set of higher-order non-modality-specific brain regions in General Semantic Network (Binder, Desai, Graves, \& Conant, 2009). An encoding model based on subjective ratings for 5 sensory-motor features of training words ("color", "motion", "sound", "shape", "action") was used to predict activation patterns related to novel individual words. The model successfully predicted above chance the brain activity patterns for concrete words in the semantic network regions ( $61 \%$ mean accuracy), but not in a set of control regions associated with visual word form processing. Based on this finding, Fernandino et al. (2016) suggested that "the brain represents concepts as multimodal combinations of sensory and motor representations" and that "heteromodal areas involved in semantic processing encode information about the relative importance of different sensory-motor attributes of concepts, possibly by storing particular combinations of sensory and motor features".

[^174]Here lies the problem - this inference is not formally valid. We need to consider what the data would have looked like if the underlying neural representation was actually different (Mahon, 2015). In this example, the successful decoding of conceptual identity in the GSN based on an encoding model of sensorymotor features does not necessitate the representational format in the GSN to be sensory-motor in nature. The results might be obtained even if the GSN uses amodal representations, as long as there is a non-arbitrary mapping between representations in the GSN and sensory-motor features. To illustrate, let us hypothetically assume that the GSN literally encodes word cooccurrence statistics. As co-occurrence statistics correlate with sensory-motor feature ratings, it would be possible to predict GSN activity patterns based on these features, even if they are not driving the activity patterns. In contrast, successful decoding would be impossible if the mapping between the GSN representations and sensory-motor features was arbitrary. Thus, Fernandino et al.'s (2016) results constitute evidence against the possibility that conceptual representations in heteromodal areas bear an arbitrary relation to sensory-motor features, as has been argued by some proponents of symbolic systems (Fodor \& Pylyshyn, 1988), but should not be taken as conclusive evidence that the GSN encodes multimodal sensory-motor information.

This issue is not limited to the specific study discussed above. To put the claim more generally, we argue that information in one representational system might be decoded based on features from another, even if they use different representational schemes, as long as there is at least a partially systematic mapping between them. Specifically, while such encoding models should be able to predict the neural activation from the features of a stimulus if the brain uses a representational scheme based on those features, the reverse is not guaranteed ${ }^{1}$. A successful prediction can also occur when the stimulus feature space is systematically related to the features that underlie the neural representational scheme. However, that relationship need not be one of equivalence. There are at least three ways in which mappings between representational systems can be made and successful prediction can occur in two of those cases.

## Types of mappings

Arbitrary mappings between representations. First, items from two representational systems might be related in an entirely arbitrary way. For example, the meaning of words is mostly unrelated to their orthographic features ${ }^{2}$, and the geographic locations of countries are not predictive of their names, etc. More generally, consider two unordered sets of items, $A=$ $\left\{A_{1}, A_{2}, \ldots, A_{n}\right\}$ and $B=\left\{B_{1}, B_{2}, \ldots, B_{n}\right\}$. An arbitrary mapping between these two sets exists when the mapping from a specific item in set $A$ to a corresponding item in set $B$ is unrelated to the mappings between the remaining items in the sets. In the context of encoding models and the brain, decoding of novel items from one set would be impossible based on a feature model from the other set, if these two sets are arbitrarily related.
in this domain (e.g., Monaghan et al., 2014), word meanings cannot be systematically predicted based on their orthography and vice versa.

Table 1 Examples of studies that use feature encoding models

| Source | Item | Features | Response vector |
| :---: | :---: | :---: | :---: |
| Mitchell et al., (2008) | Concrete words (dog) | Co-occurrence statistics with 25 sensory-motor verbs | Pattern of activation in all cortical voxels |
| Fernandino et al., (2016) | Concrete words (dog) | 5 sensory-motor relevance ratings | Pattern of activation in the GSN (Binder et al., 2009) |
| Current simulation | Numbers (3497) | 5 decimal digits [030497] | 17 binary digits $[000000000000$ 00000000110110101001 ] |

Sets that use the same representational format. In contrast, a successful prediction can occur if the two sets use the same representational format. Consider the set of multi-digit numbers in the decimal system, $A=\{10,11, \ldots, 427, \ldots\}$, and the set of 10 digits in the decimal system, $B=\{0,1,2,3,4,5,6,7,8,9,10\}$. These sets use the same representational format to represent quantities (the decimal system), and there is a systematic linear mapping from the features (the digits), to the multi-digit numbers, such that:

$$
\begin{gathered}
\overline{d_{n} d_{n-1} \ldots d_{1} d_{0}}=\sum_{i=0}^{n}\left(d_{i} \times 10^{i}\right) \\
3491=3 \times 1000+4 \times 100+9 \times 10+1 \times 1
\end{gathered}
$$

When we have such systematic mappings between systems that use the same representational format, knowing the mapping function allows us to decompose any item from set A as a combination of features from set B . An example of such a mapping would be Fernandino et al.'s (2016) suggestion that the general semantic network encodes multimodal combinations of sensory-motor features by integrating information from modality-specific sensory-motor areas. If this were true, then you could predict the neural pattern of novel items from their featural representations, which is what that study found as well.

Sets that use different but systematically related representational formats. However, there is an alternative, which would also allow you to make a successful prediction from encoding models. Two sets can use different representational schemes, while maintaining a systematic mapping between themselves that allows us to predict the mapping of any one pair of items from knowledge of the mapping function. Within the context of conceptual representations in the brain, higher-level heteromodal areas might use a representational code that is different from the one used by sensory-motor cortices, but there might be a systematic mapping between representations in each system ${ }^{3}$.

For a simplified example, consider the relation between the decimal and the binary systems for representing numeric values. A binary represented value can be transformed into a decimal number by applying the following formula:

[^175]\[

$$
\begin{gathered}
\left(\overline{d_{n} d_{n-1} \ldots d_{0}}\right)_{2} \rightarrow\left(\sum_{i=0}^{n}\left(d_{i} \times 2^{i}\right)\right)_{10} \\
10011_{2} \rightarrow 1 \times 2^{4}+0 \times 2^{3}+0 \times 2^{2}+1 \times 2^{1}+1 \times 2^{0} \\
=16_{10}+2_{10}+1_{10}=19_{10}
\end{gathered}
$$
\]

Clearly, there is a systematic but non-linear mapping between the decimal and the binary system, and yet, these two systems use different codes to represent numbers. If our argument is correct then it should be possible to predict the binary representation of a number based on a decimal feature encoding model. Below we present a simulation that achieves this by applying the encoding model approach often used in neuroimaging studies. Within the simulation, binary vectors are analogous to voxel activation patterns, and the encoding model is based on decimal representations (Table 1).

## Simulation: Decoding binary representations with a decimal feature encoding model

As detailed previously, encoding models predict stimulus identity from brain activation by modelling the relationship between the constituent features of the training stimuli and their corresponding BOLD activation in a group of voxels. Then they use that relationship to estimate the expected neural activation patterns for novel test items based on their feature representations. The predicted activation pattern for each stimulus is compared to the observed patterns for all test stimuli. For the following simulation, let us consider the numbers from 0 to 99999 as our stimulus set. They can be decomposed into 5dimensional feature vectors where each feature is a decimal digit (e.g., 3497 can be decomposed as [0 349 7]. These features can be considered analogous to the 5 sensory-motor relevance ratings of words used by Fernandino et al. (2016) or to the cooccurrence statistics with sensory-motor verbs used by Mitchell et al. (2008). Further, let us consider the binary representation numbers as 17-dimensional vectors (e.g. [000000000000 00000000110110101001 ], to be analogous to the BOLD activation pattern in a set of 17 voxels in an ROI under investigation. The correspondence between these patterns and actual neuroimaging studies using this approach is demonstrated in Table 1.

We trained an encoding model to predict the binary activation pattern for a given number, based on its 5-dimensional decimal
limitations we will briefly cover this issue in the general discussion, but a more in-depth treatment is needed
feature representation. The modelling followed 4 steps: 1) splitting the stimuli into a training ( $90 \%$ ) and a test ( $10 \%$ ) set, 2) fitting multiple linear regression models on the training set with the 17 binary features as response variables, and the 5 decimal features as predictors, 3 ) calculating predicted activation pattern (predicted maps, PMs) for each test item from its decimal features and the multivariate regression model, 4) comparing the PMs with the actual binary patterns for all test items (observed maps, OMs). In the comparison stage, we computed the Euclidean distance between each PM and the OMs for all test items, and we calculated the percentile rank of the similarity between the PM and the OM of each item. For example, if the PM for the number 29782 were most similar to OM for that number, then the percentile rank for it would be $10000 / 10000$ $=1$. However, if it were more similar to the OMs of 1000 other items, then its percentile rank would be $9000 / 10000=0.9$.

The encoding model was successful in decoding the binary representation of untrained items based only on their decimal features. The prediction accuracy of the linear regression model was $0.7(\mathrm{SD}=0.24)$ and a wilcoxon signed rank test showed that it was above chance ( $p<.0001$ ). Since by definition binary and decimal number systems use different representational formats, we cannot conclude that the representation of binary numbers encodes decimal features. By analogy, the successful decoding of patterns of neural activation based on a stimulus feature space, cannot be used to infer that the brain encodes information about these features or that its neural representational space is organized along the dimensions of that feature space.

## Discussion

Stimulus-feature based encoding models (Haxby et al., 2014, Naselaris et al., 2011) are a powerful new tool for studying how the constituent features of stimuli relate to the neural activation patterns elicited by these stimuli. They represent a significant methodological advance over more traditional MVPA methods because they allow us to predict neural activation for novel items and because they can be used to decode the identity of such items from neural data alone. While this is an impressive feat and an incredibly useful tool, we have to be cautious in interpreting what such successes mean for our understanding of the representational system of the brain. Both theorists (e.g., Haxby et al., 2014; Naselaris \& Kay, 2015; Naselaris et al., 2011; Norman et al., 2006; Tong \& Pratte, 2012) and practitioners (e.g. Fernandino et al., 2016; Kay et al., 2008; Mitchell et al., 2008; Santoro et al., 2014) have suggested that we can infer that the brain uses a certain set of features to encode information, if we can successfully decode the activity of novel items from such features. However, as we have argued here, this inference is not formally valid. Successful decoding might be the result of a systematic relationship between the representational system of the brain and the stimulus feature set, even if those utilize different representational schemes.

How do we know whether two representational systems are truly different? It could be argued that in our example, both

[^176]binary and decimal number systems share many properties, and that they are merely different implementations of the same fundamental representation. For example, both systems use the position of a digit to encode its magnitude, and as a result, all arithmetic procedures that can be performed with decimal numbers can be applied to binary numbers as well. We propose that the key issue in determining whether two representations are the same is whether you can establish a one-to-one mapping relation between features at different levels of representation in each system. For example, if you substitute each decimal digit with a unique letter, the resulting system would appear different from the decimal system only on the surface, but the relation between multi-digit numbers and their features would be the same in both cases ${ }^{4}$ In contrast, decimal and binary features have a qualitatively different relation to the numbers they represent. Despite this, binary representations can be decoded based on decimal features, illustrating the inferential problem of encoding models we address here.

It is important to clarify that the "one-to-one" mapping is an abstract requirement. We are not claiming that to establish representational equivalence between the brain and a certain set of features that it is necessary to find a one-to-one mapping between the basic feature components of stimuli and activation in individual voxels or groups of voxels. The brain does not compute and represent information at the voxel level - voxel activations are the result of averaged activity over hundreds of thousands of neurons. The general lack of access to large-scale neural level activity in the living human brain makes it even more important to not only discover analytical tools that helps us relate voxel activation to possible representations, but also to understand the limitations of those tools and what they can and cannot tell us.

An important question that naturally arises from the caveats we discussed is how one can maximize confidence in the outcome of a forward encoding model approach, or conversely, guard oneself against unjustified inferences. We propose that it is crucial to compare the performance of several possible encoding models. To that aim, it is not sufficient to use a "baseline model" that is unrelated to the domain of interest (i.e., compare a semantic feature model to a low-level visual word form model). Instead, one or several alternative representational models should be tested that are derived from competing theories (i.e., semantic model A vs. semantic model B). To illustrate, an elegant comparison of a sensory-based vs. non-sensory-based semantic model was achieved by Anderson et al. (2015). These authors contrasted a visual model with a word cooccurrence model to investigate which brain regions represent modality-specific visual features, and which do not (using differential correlation in RSA rather than an encoding model). The relative superiority of a particular model at predicting activation patterns in a brain region makes it more likely that the brain is using the representational scheme of the better performing model rather than the alternative. However, it is important to keep in mind that such comparisons only provide
perfect decoding accuracy; compare that to the 0.7 decoding accuracy for the decimal-to-binary model
evidence for the relative likelihood of each model, but, due to the limitations discussed above, still do not allow us to infer that the winning model is the "true" model.

For that reason, besides the assessment of relative model performance based on model comparison, a second crucial step is to evaluate absolute prediction performance. In particular, the observed decoding accuracy can be compared to the "noise ceiling", or to the "upper limit of prediction accuracy" (Naselaris et al., 2011), reflecting the maximal performance that can be feasibly achieved given the noise present in the signal. The gap between the two can be thought of as the variance that is not explained by the current model, which should motivate and guide the search for an improved or alternative version of the model. Until such maximal performance is obtained, we should be careful in making strong representational inferences about the brain from the currently available analytic methods.

Ultimately, many of these inferential caveats exist because fMRI data is correlational. Comparing alternative models and evaluating absolute prediction performance might eventually converge on the true underlying feature model, but this is not guaranteed. We propose that an even better way to test representational hypotheses might be to introduce experimental manipulations that affect the hypothesized representational dimensions. For example, one could prime participants to weight some features of the stimuli more than others. If that leads to changes in the performance of a classifier based on the primed features, this would constitute much stronger evidence that these features underlie the neural representational scheme in question. This proposal is logical but it has not been experimentally tested yet, and we look forward to seeing how it will fare in practice.

## References

Anderson, A. J., Bruni, E., Lopopolo, A., Poesio, M., \& Baroni, M. (2015). Reading visually embodied meaning from the brain: visually grounded computational models decode visual-object mental imagery induced by written text. NeuroImage, 120, 309-322.
Binder, J. R., Desai, R. H., Graves, W. W., \& Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and MetaAnalysis of 120 Functional Neuroimaging Studies. Cerebral Cortex, 19(12), 2767-2796.
Bonnici, H. M., Kumaran, D., Chadwick, M. J., Weiskopf, N., Hassabis, D., \& Maguire, E. A. (2012). Decoding representations of scenes in the medial temporal lobes. Hippocampus, 22(5), 1143-1153.
Brouwer, G. J., \& Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. The Journal of Neuroscience, 29(44), 13992-14003.
Casey, M., Thompson, J., Kang, O., Raizada, R., \& Wheatley, T. (2012). Population Codes Representing Musical Timbre for High-Level fMRI Categorization of Music Genres. In Machine Learning and Interpretation in Neuroimaging (pp. 34-41). Springer Berlin Heidelberg.
Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., \& Bonte, M. (2014). Brain-Based Translation: fMRI Decoding of Spoken Words in Bilinguals Reveals Language-Independent Semantic Representations in Anterior Temporal Lobe. The Journal of Neuroscience, 34(1), 332-338.
Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and
mean activation: why, how, and what does it tell us? Cognitive, Affective \& Behavioral Neuroscience, 13(3), 667-673.
Coutanche, M. N., \& Thompson-Schill, S. L. (2015). Creating Concepts from Converging Features in Human Cortex. Cerebral Cortex, 25(9), 2584-2593.
Davis, T., \& Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. Annals of the New York Academy of Sciences, 1296(1), 108-134.
Farah, M. J., \& McClelland, J. L. (1991). A computational model of semantic memory impairment: modality specificity and emergent category specificity. Journal of Experimental Psychology: General, 120(4), 339.
Fernandino, L., Humphries, C. J., Conant, L. L., Seidenberg, M. S., \& Binder, J. R. (2016). Heteromodal Cortical Areas Encode SensoryMotor Features of Word Meaning. Journal of Neuroscience, 36(38), 9763-9769.
Haxby, J. V., Connolly, A. C., \& Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annual Review of Neuroscience, 37(1), 435-456.
Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., \& Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425-2430.
Kay, K. N., Naselaris, T., Prenger, R. J., \& Gallant, J. L. (2008). Identifying natural images from human brain activity. Nature, 452(7185), 352-355.
Lee, H., \& Kuhl, B. A. (2016). Reconstructing Perceived and Retrieved Faces from Activity Patterns in Lateral Parietal Cortex. Journal of Neuroscience, 36(22), 6069-6082.
Mahon, B. Z. (2015). The Burden of Embodied Cognition. Canadian Journal of Experimental Psychology, 69(2), 172-178.
Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., \& Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. Science, 320(5880), 1191-1195.
Monaghan, P., Shillcock, R. C., Christiansen, M. H., \& Kirby, S. (2014). How arbitrary is language? Phil. Trans. R. Soc. B, 369(1651), 20130299.

Naselaris, T., \& Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. Trends in Cognitive Sciences, $19(10), 551-554$.
Naselaris, T., Kay, K. N., Nishimoto, S., \& Gallant, J. L. (2011). Encoding and decoding in fMRI. NeuroImage, 56(2), 400-410.
Norman, K., Polyn, S., Detre, G., \& Haxby, J. (2006). Beyond mindreading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences, 10(9), 424-430.
Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? Trends in Cognitive Sciences, 10(2), 59-63.
Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron, 72(5), 692697.

Santoro, R., Moerel, M., Martino, F. D., Goebel, R., Ugurbil, K., Yacoub, E., \& Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. PLOS Computational Biology, 10(1), e1003412.
Tong, F., \& Pratte, M. (2012). Decoding Patterns of Human Brain Activity. Annual Review of Psychology, 63(1), 483-509.

# The Relational Luring Effect: False Recognition via Relational Similarity 

Vencislav Popov<br>Carnegie Mellon University, Pittsburgh, PA, USA<br>Penka Hristova<br>New Bulgarian University


#### Abstract

We present evidence for a novel relational luring effect (RLE) in recognition memory. Participants performed a continuous associative recognition task in which they had to discriminate between new, old and recombined word pairs. Participants made more false alarms and responded more slowly to lures (TABLE CLOTH) that were relationally similar to studied pairs (FLOOR CARPET). RTs and false alarms for lures increased linearly as the number of previously studied different exemplars of the relation increased (e.g., 0 to 4 previous exemplars). The RLE effect was stronger for relations that were represented by exemplars that were more typical of the relation. These results suggest that semantic relations exist as independent representations in LTM, and that during associative recognition these representations can be a spurious source of familiarity. The RLE has implications for models of semantic and episodic memory, unitization in associative recognition, analogical reasoning, and constructive memory research.


# Target-to-distractor similarity can help visual search performance 

Vencislav Popov (vencislav.popov@gmail.com)<br>Lynne Reder (reder@cmu.edu)<br>Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA


#### Abstract

We found an unexpected positive effect of target-to-distractor similarity (TD) in a visual search task, despite overwhelming evidence in the literature that TD similarity hurts visual search performance. Participants with no prior knowledge of Chinese performed 12 hour-long sessions over 4 weeks, where they had to find a briefly presented target character among a set of distractors. At the beginning of the experiment, TD similarity hurt performance, but the effect reversed during the first session and remained positive throughout the remaining sessions. We present a simple connectionist model that accounts for that reversal of TD similarity effects on visual search and we discuss possible theoretical explanations.


Keywords: visual search; learning; similarity; connectionist model; neural network

## Introduction

Intuitively, the more similar two objects are to each other, the more difficult it should be to say whether they are the same object or not. Research with the visual search task has confirmed this intuition repeatedly - when a target is more similar to distractors in the search array accuracy decreases and response times (RTs) increase (Treisman \& Gormican, 1988; Duncan \& Humphreys, 1989; Treisman, 1991), and more errant saccades are made to the highly similar distractors (Bichot \& Schall, 1999). Despite the ubiquity of this negative target-to-distractor (TD) similarity effect, in a recent experiment that explored how frequency of exposure affects a variety of tasks, including a visual search task, we discovered by accident a positive TD similarity effect in visual search (Reder, Xiaonan, Keinath \& Popov, 2016). We found that greater TD similarity eventually lead to greater accuracy and faster RTs.

The visual search task was performed with Chinese characters over 12 hour-long sessions and the participants were US undergraduates with no previous knowledge of Chinese characters. Interestingly, during the initial stages of the visual search task we observed a negative TD similarity effect, as is expected from prior research, but this effect reversed quickly. After a single training session, higher TD similarity lead to better performance. Since this result was not reported in Reder et al. (2016), we will first describe the experiment and the key results with respect to frequency and similarity.

[^177]
## Method

## Participants

Twenty U.S. college students with no prior experience learning Chinese participated in this experiment.

## Materials

The stimuli for the visual search task were 64 Chinese characters. We grouped the characters based on their visual similarity in 16 sets of four characters. Characters within a set had a higher similarity with each other compared to characters from other sets. This was determined by a native Chinese speaker and was subsequently confirmed by analyzing orthographic vector representations of the characters (Xing et al, 2004; Yang et al 2009) ${ }^{1}$. We used highly similar distractors in order to force participants to encode the entire character rather than a subset of diagnostic features. For each participant, half of the sets were randomly assigned to the high-frequency condition and were presented 20 times more often during the visual search task.

## Procedure

The visual search task was performed over 12 sessions. There were three session per week and each lasted for about 1 hour. Each trial began with a sample character presented in the middle of the screen for 2 seconds. The sample character was followed by a display of 3 to 5 characters. On half of the trials, the display included the target character and participants were to respond whether the target was present. Three of the characters were from the same similarity set as the target character. Additionally, 0-2 characters from different sets of the same frequency class as the target were also present as distractors. After participants made their response, they received immediate accuracy feedback.

## Results and Discussion

We analyzed the accuracy data via logistic mixed-effects regressions and RTs via linear mixed-effects regressions, both with participants and items ${ }^{2}$ as random intercept effects. All effects discussed below were significant ( $\mathrm{p}<.05$ ) as determined by likelihood ratio tests that compared alternative regression models with and without each effect. Most results concerned with effects of frequency are described in Reder et al. (2016; see also Reder et al., 2007); here we focus primarily on the role of similarity.

[^178]

Fig 1. Accuracy and RTs to search the display as a function of target presence or absence, week of training and frequency of exposure.

Figure 1 shows the effect of frequency of exposure on accuracy and RTs for finding the target character. Overall, accuracy was greater and RTs faster for characters from high frequency sets. There was a two-way interaction between frequency and whether the target was present or absent. For accuracy, the effect of frequency was evident only when the target was absent. In Reder et al. (2015), we proposed that frequent exposure facilitates the development of unitized representations of each character. That is, a character seen less often has a weaker chunk representation and is more likely to be encoded as a configuration of some of its features rather than as a single higher-level unit. Thus, when a participant is searching for a LF character the probability of partially matching some of the target's features with features of the distractors is much greater compared to HF characters. This leads to more false alarms in the absent condition, but does not affect the present condition. The interaction was also evident in RTs, although there was still a small effect of frequency in the present trials, likely reflecting the differential efficiency of encoding high and low frequency characters.

A number of previous versions of this experiment had failed to show the hypothesized frequency effects. In those experiments, the distractors in each search array were chosen at random and thus were not very similar to the targets on each trial. In contrast, in the current experiment we ensured that targets were paired with highly similar distractors. We believe that the discrimination required in the prior versions of our visual search task was too easy and as a result, participants were able to perform the task by noting and remembering individual features that distinguish the target from its distractors. As a result, participants did not have to develop stable chunks for each character.

If that is the correct interpretation, then we expected to see an analogous effect within this experiment based on the discriminability (similarity) of the target character to its distractors in the search array. We should see that greater TD similarity leads to a better performance over time, because the increased difficulty in discriminating the target from the distractors forces people to develop stronger and more stable representations of each character as a whole unit/chunk. Note that this prediction is contrary to an intuitive and classic result in the visual search literature - usually, the more similar a target to its distractors, the more difficult it is to perform the task (Duncan \& Humphreys, 1989).

TD similarity was calculated based on vector representations obtained from Yang et al. (2009). Each character was represented as a vector of 270 binary features for five dimensions - simple features, shapes, structure, position and strokes. These vector representations are based on an orthographic analysis of the characters and prior behavioral work (Xing et al, 2004). These representations have been already used successfully to model print-to-sound mappings in Chinese (Yang et al, 2009) with a connectionist model similar to those used in modeling English print-tosound mappings (Harm \& Seidenberg, 1999). For each search array, we calculated the mean Euclidean distance between the target and each distractor. Low and high similarity groups were defined as being below or above one SD around the mean similarity of all search arrays.

Figure 2 shows that our prediction was confirmed. During the first session, initially greater TD similarity lead to slower RTs. However, by the end of the first session the effect had reversed and throughout the remaining sessions high similarity between the target and distractors lead to faster RTs and greater accuracy.


Fig 2. Accuracy and RTs for visual search as a function of similarity of target to distractors within the search array. Right panel shows performance over time during the first session. Left panel shows performance over all 12 sessions.

Contrary to our findings, visual search tasks usually show that high TD similarity leads to lower accuracy and slower response times. Why is it that we found exactly the opposite? A trivial explanation would be that TD similarity in our study was confounded with distractor-to-distractor similarity. The latter consistently shows positive similarity effects. We discounted this explanation by showing that the positive effect of target-to-distractor similarity remained even after controlling for distractor-to-distractor similarity in the regression model.

A theoretical explanation is that most visual search studies use simple stimuli that have pre-existing representations in long-term memory and no additional learning is required. Our study instead used Chinese characters, which are a complex configuration of features for participants who do not know Chinese. Since these characters did not have preexisting representations, participants had to develop them while doing the visual search task. We suggest that those representations were influenced by the demands of the task - to make highly similar patterns more distinct from one another so as to be better suited to support future performance. In essence, we argue that over time when the target is presented along highly similar distractors, the cognitive system builds more distinctive and stable representations of these targets.

Additional support from this argument comes from the fact that (as it was with frequency) the similarity effect is mostly observed in the absent condition (Figure 3, left panel). That is, the benefit from gaining more distinct and stable representations is mostly to prevent the partial matching of shared features between the target and distractors on absent trials.


Fig. 3. Effect of TD similarity on visual search accuracy as a function of training session and whether the target was absent (left panel) or present (right panel).

## A connectionist model

The reversal of the similarity effect is something of a challenge from a modeling perspective. How exactly are more distinct representations built over time and what mechanism drives that differentiation?

In order to capture the reversal of similarity effects, we decided to apply a novel connectionist model that will be able to perform a visual search task while continuously modifying its internal representations of the stimuli. Connectionist
models that represent stimuli as distributed patterns of activity are well suited for exploring the time-dependent changes in the structure of conceptual representations that consist of multiple features. In line with our theoretical interpretation of the data, we expected that in the beginning of training, the model will behave similarly to our human participants and will make more errors for highly similar distractors. However, since this initial behavior would lead to more errors, over time the error-dependent learning might cause the model to alter its internal representations of each stimulus so as to make them more distinct from one another.

In this way, the problem that a connectionist model of this task has to solve is akin to the XOR problem. Specifically, how should the representation of the input patterns be transformed so that similarity is reversed through the transformation? One possibility is for our model to have at least one layer that intervenes between the input layer and a layer that computes similarity between patterns. After a number of failed attempts using a single hidden layer we tried two separate hidden layers, which allowed the network to more gradually change the similarity structure in the input.

The visual search task here requires that participants are able to initially encode the sample character and keep it active in short term memory while comparing it in turn with each candidate character in the search array. To model the task as fully as possible, a model was implemented with a single input layer that can send activation through two different pathways - either to a working memory (WM) module (implemented as a kind of a long short-term memory module), or directly to the comparison layer. This dual pathway represents two ways to use information coming through the senses. One pathway can store the representation of the target in short-term memory and then manipulate it in the absence of the stimulus itself. The other pathway can directly use the incoming information (i.e. the candidates for comparison in the search array).

We assumed that the visual search is performed serially, because the RTs increased linearly with the search array size, and because the slope in the absent condition was twice as large as the slope in the present condition (Treisman \& Gelade, 1980). As a simplification, the model presented below will deal only with this serial search case.

## Architecture

The network consisted of the following layers (the architecture is presented in Figure 4):

- Input: 20 units
- Hidden1: 15 sigmoid units
- Hidden2: 10 sigmoid units
- LSTM module 1
- LSTM_Input: 10 linear units
- LSTM_Buffer: 10 linear units
- LSTM_Context: 10 linear units
- LSTM_Output: 10 linear units
- LSTM_Input_gate: 1 input unit
- LSTM_Context_gate: 1 input unit
- LSTM_Output_gate: 1 input unit
- Direct_output: 10 linear units
- Direct_output_gate: 1 linear unit
- Comaprison: 20 units
- Response: 2 softmax units (output layer)

The input was connected in a feedforward manner to hidden1, which in turn was connected to hidden2. We expected the two hidden layers to progressively extract higher order features of the input. Initial weights between these layers were randomized with a mean of 0 and $\operatorname{sd}$ of 0.5 . Each unit of the hidden 2 layer was connected with the corresponding unit in the input layer of the LSTM module, as well as with the direct_output layer with a frozen weight of 1. The same applied to the connections from LSTM_Input to LSTM_Buffer and from LSTM_Buffer to LSTM_Output. Thus, the output of the hidden2 layer was copied forward to the output layer of the LSTM module, and to the direct_output layer. The LSTM module also had a recurrent context layer that was connected bi-directionally to the LSTM_Buffer layer.

The purpose of the four gates was to control the flow of activation through these two modules. There was a fixed negative bias of -1 to the LSTM_input and LSTM_output layers, and a fixed positive connection of 1 with their corresponding gates. Since they were all linear units that were cropped at 0 and 1 , when a gate was off, no activity was copied to corresponding and subsequent layers. When a gate was on, it negated the bias and the layer copied the output of the preceding layer and passed it forward.

## Network functioning

Each example trial was composed of four events (i.e., presenting different input patterns):

1. Presentation of the target
2. First candidate from the search array
3. Second candidate from the search array
4. Final candidate from the search array

When the target was presented to the input, only the LSTM_input gate was on. Thus, the activity in hidden 2 layer that corresponds to the sample input was copied to the LSTM1_input, LSTM_buffer and LSTM1_context layer. All other gates were off, thus preventing the sample input from transferring to the direct_output layer.

When each candidate from the search array appeared, the LSTM_input gate was off, preventing the candidate representation from entering LSTM module. All the other gates were on. This meant several things happened. The candidate representation on hidden 2 was copied to the direct_output layer. The representation of the sample that was encoded in the LSTM_context layer on the previous time step was transfered back on to the LSTM_buffer, and from there it was transfered to the LSTM_output. At this point, the network had the hidden 2 representation of the sample instantiated on the LSTM_output layer, while the hidden2 representation of the first candidate was active on the direct_output layer.

Both the LSTM_output and the direct_output layers were connected with free random weights with sd 0.1 to the comparison layer, which integrated the representations of the sample and the first candidate input. The comparison layer was connected to the response layer, which consisted of two units - 1 for responding that the two representations are the same, and the other for responding that they are different.


Fig 4. The network architecture. Arrows with stars (*) represent copy connections, where each unit in the sending layer is connected with a single connection with fixed strength 1 to the corresponding unit in the receiving layer.

## Training

To mimic the experiment's stimuli, 64 input patterns of length 20 were created with binary values that were grouped into 16 sets, which had greater similarity within sets than between sets. On average, $50 \%$ of the features in each input vector were "on". The randomization and conditions were equivalent to those in the experiment.

Mean similarity in a set was calculated using Euclidean distance. The groups in the lower $25 \%$ quantile of the distance distribution were designated as "Low distance / High similarity" sets, while groups in the higher $25 \%$ quantile were designated as "High distance / Low similarity" sets.

When a distractor was present, the network was trained to activate the "mismatch" response unit, while when a target was present it was trained to activate the "match" response unit. Therefore, the goal of the network was to discover a suitable combined representation in the comparison layer such that it will be able to discriminate when the LSTM_output and the direct_output layers had the same or different patterns of activation.

We used a back-propagation training algorithm with a learning rate of 0.01 and a momentum descent with a momentum rate of 0.9 . The network was trained for 4000 passes through the training set and the weights were updated at the end of each pass. After every 100 updates, we recorded the output activation of the hidden1, hidden 2 and the response layer.

## Results and discussion

Frequency effects. The main results of the simulation are presented in Figures 5 and 6, which show the activation of the "match" response unit over training time. Since the response layer had softmax units this value can be directly interpreted as the proportion of "match" responses the network would give in response to a pattern. In Figure 5 we can see that the training patterns that were presented more often lead to greater accuracy.


Fig 5. Activation of the Match output unit as a function of training time, stimulus type and frequency of the input pattern.

Several things should be noted about this pattern. We can see that initially the network deactivated the match response unit for both types of stimuli. We can also see that overall the effect of frequency was much greater on target stimuli compared to distractor stimuli, which is exactly the opposite effect than the one we found in the behavioral data. This was probably because there were 5 times more distractor items than target items ( 3 in absent conditions, 2 in the present conditions). In the actual experiment, this too was the case, but participants got feedback only for their final response, thus they had equal amounts of "present" and "absent" feedbacks. On the other hand, the network was trained as if each individual comparison required a response, which causes the discrepancy between distractor and target stimuli.

Thus, while the network captures the overall effect of frequency, its training regimen causes it to miss the specific pattern of frequency for different types of stimuli. This could
possibly be solved by considering the current response layer to be an internal response, reflecting whether there is a match or not. Then a secondary motor response layer can be added which outputs a 'present' response if the internal match response is higher than a threshold, or stays inactive until all candidates have been compared. If by the last one none of them had elicited a match response, it produces an 'absent' response. In this way the network would reflect the actual behavior more closely, and weight updating would be affected only by the final response in each example.

Similarity effects. As can be seen from Figure 6, initially the network performance is better for input stimuli that are less similar to their distractors. This is a normal behavior of connectionist networks, and it is also what is expected by previous behavioral data from the visual search paradigm (Duncan \& Humphreys, 1989). However, after about 2300 weight updates the effect reverses and stimuli that are closer to each other in the input space lead to better performance. Importantly, this reversal happens very shortly after the behavior of the network starts to approximate the behavioral result levels ( $\sim 70 \%$ accuracy), which is exactly the pattern we have seen from the behavioral session - greater similarity impairs performance during the first session of training, but the effect reverses by the end of that session. Indeed, if we limit our attention to the window between updates 2200 and 3000, which is immediately after the pre-training, and before the performance saturates at ceiling, there is a close correspondence between the network performance both in terms of frequency and similarity structure.


Fig 6. Activation of the 'Match' output unit as a function of training time, type of stimulus and Euclidean distance between the target and distractor input in each array.

What could be causing this reversal of the distance/similarity effect? A possible answer comes from examining the input-output mappings, as well as the hidden representations the network develops during training.

If we split the candidate input patterns into targets, similar distractors and dissimilar distractors, then the network is supposed to produce the following outputs. For targets, which are identical to the sample item (thus 0 distance or perfect correlation) the network has to produce a match response, but for distractors that are highly similar as well, it has to produce mismatch responses. Thus, a major conflict during training comes from the fact that when distance is high, the network has to produce only one type of response, but when it is low,
it has to either respond with a match or a mismatch. One way to achieve these contradictory goals would be to develop such hidden representations of the input that cause highly similar patterns to be represented as less similar to each other.

To test this explanation we looked at the distance between the sample item and its distractors in activation patterns in each of the two hidden layers, split by the distance in the input layer. In right panel of Figure 7, we can see that in the first hidden layer the distance structure in the input has been preserved. In the second hidden layer, however, in the beginning there is no difference in distance due to the two layers of random weights and the sigmoid nature of the stimuli. As training progresses, stimuli that were low in distance in the input and the first hidden layer become more distant to one another, compared to stimuli that were highly distinct to begin with.


Fig 7. Distance between the hidden layers representations of the target and the distractors in each training set as a function of training time. Left panel shows distance in the first hidden layer, right panel shows the second hidden layer.

## General Discussion

The current paper present preliminary data on a novel counter-intuitive finding that the usual target-to-distractor similarity effect in visual search reverses after a short training with previously unfamiliar Chinese characters. Namely, while targets that are highly similar to distractors in a search array are usually more difficult to detect, when the stimuli are complex visual objects, this effect reverses after about 20 repetitions of each object as a target. We propose that visual discrimination and learning interact in such a way that greater difficulty in discriminating the stimuli causes the development of more distinct and stable representations.

To test this idea of differentiation in the character representation over time, we fit a novel connectionist model. When it comes to frequency, the network successfully captured the overall effect that more frequently exposed stimuli led to better performance (although see the preceding discussion for some limitations). Theoretically, this was presumably because low frequency made it more likely that people depend on representing the characters as a configuration of features, rather than on its weak chunked representation. This caused them to be more likely to partial match constituent features and confuse distractors with targets. In contrast, the network showed exactly the opposite
effect, because distractors were present 5 times more than targets and had a greater influence over the weight updates.

The most interesting aspect of the model is that it was able to successfully capture the reversal of the similarity effect on visual search performance. It achieved this by transforming the input through multiple hidden layers, which allowed it to change the similarity structure in the input so that highly similar distractors became more and more differentiated in the second hidden layer as training progressed.

This explanation was further supported by a model that involved direct connections from the input to the comparison layer without hidden layer representations (not shown here). This model did not show the similarity reversal effect. This model is analogous to performing the task without having to develop novel representations. One novel prediction from the comparison of these two models and task versions would be that people who learned the Chinese characters under a visual search task would rate highly similar characters as less similar after the training.

Finally, while we simulated the input patterns in this model to resemble as closely as possible how our stimuli were structured, the simulation results might be specific to the interaction between the model architecture and the generated stimuli. Initial modeling results using the actual 270 -length vector representations of the Chinese characters show the same pattern as the simplified model presented here.

## References

Bichot, N. P., \& Schall, J. D. (1999). Effects of similarity and history on neural mechanisms of visual selection. Nature Neuroscience, 2(6), 549-554.
Duncan, J., \& Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological Review, 96(3), 433-458.
Reder, L. M., Paynter, C., Diana, R. A., Ngiam, J., \& Dickison, D. (2007). In B. Ross \& A. S. Benjamin (Eds.), The Psychology Of Learning And Motivation (pp. 271312). New York, NY: Academic Press.

Reder, L. M., Liu, X. L., Keinath, A., \& Popov, V. (2016). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. Psychonomic Bulletin \& Review, 23(1), 271-277.
Treisman, A. (1991). Search, similarity, and integration of features between and within dimensions. Journal of Experimental Psychology: Human Perception and Performance, 17(3), 652-676.
Treisman, A. M., \& Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12(1), 97-136.
Treisman, A., \& Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. Psychological Review, 95(1), 15-48.
Xing, H. B., \& Li, P. (2004). The acquisition of Chinese characters: Corpus analyses and connectionist simulations. Journal of Cognitive Science, 5(1), 1-49.
Yang, J., McCandliss, B. D., Shu, H., \& Zevin, J. D. (2009). Simulating Language-specific and Language-general Effects in a Statistical Learning Model of Chinese Reading. Journal of Memory and Language, 61(2), 238-257.

# Timing Time: Why Early Vision is Cognitively Impenetrable 

Athanassios Raftopoulos (raftop@ucy.ac.cy)<br>Department of Psychology, University of Cyprus, P.O. Box 20537<br>Nicosia 1678, Cyprus


#### Abstract

Newen and Vetter (2016) argue that cognitive penetration (CP) of perceptual experience is the most possible account of the evidence. They target both the weak impenetrability thesis that only some early visual processes are cognitively impenetrable (CI), and the strong impenetrability thesis that all perceptual processes are CI. Since I agree that perceptual processing as a whole is CP, I will concentrate on their arguments against the weak CI thesis. In attacking weak CI, the authors take aim at Raftopoulos' arguments supporting the CI of early vision. Their main argument comes from studies that, Newen and Vetter think, show that early vision is CP by demonstrating the existence of cognitive effects on early vision. I examine the same empirical evidence that Newen and Vatter discuss and argue this same evidence strongly supports the view that early vision is CI.


## 1. Introduction

Newen and Vetter (2016) argue that the CP of perceptual experience is the most possible account of the available evidence. They target both the weak impenetrability thesis that only some early visual processes are CI, and the strong impenetrability thesis that all perceptual processes are CI. Since I agree that perceptual processing, as a whole, is CP, I concentrate on weak CI. The authors criticize Raftopoulos' (2001, 2014) arguments in support of the view that early vision, defined by temporal criteria, is CI. They maintain that top-down influences from higher-level cognitive areas to early visual cortex occur very early. Thus, it is unjustified to conclude from an early timing of a visual process that it is unaffected by top-down cognitive influences.

The authors' main argument against weak CI, given that weak CI holds that during the timing of early vision no direct cognitive effects modulate the ongoing perceptual processes, comes from studies that, Newen and Vetter think, demonstrate the existence of cognitive effects on visual processing that occur within the time frame of early vision.

Furthermore, Newen \& Vetter $(2016,5)$ argue that even if it exists a stage of perceptual processing that is unaffected by cognition, it lasts for so few ms that it could not be a plausible candidate for a perceptual module and it would be of almost no importance to philosophical discussions about the CP of perception. In other words, even if such a CI stage exists, the information it processes and outputs would be too poor to be categorized as a properly speaking perceptual content.

Here, I concentrate on Newen and Vetter's arguments from temporal processing. The main reason is that if it turns out that this evidence does not support the CP of early vision, the rest of the arguments against weak CI are moot.

I argue that the evidence Newen and Vetter adduce if properly examined strongly supports weak CI. The evidence that emerges from these studies shows that all visual brain areas at some time are affected by top-down cognitive signals that usually drive spatial or object/feature based attention. They also show that within the time frame of early vision there is a confluence of top-down, lateral, and bottom up interactions. These interactions, however, do not involve any cognitive signals. These studies do not show that during the first $120-140 \mathrm{~ms}$ of perceptual processing there are direct cognitive top-down effects on early vision. What Newen and Vetter consider as evidence for top-down cognitive effects on early vision is, in effect, evidence for top-down and bottom up interactions that do not involve any cognitive effects.

In the first section, I define early vision and examine most of the studies cited by Newen and Vetter (2016) and argue that they do not show that early vision is CP. I also take up Newen and Vetter's claim that even if a CI stage of visual processing exists it is too impoverished to be deemed a perceptual stage and claim that the output of early vision has a rich structure. In the second section, I examine the rest of the evidence used by Newen and Vatter and argue that in effect it supports the claim that early vision is CI. All recurrent processing during early vision is restricted within the visual areas of the brain and does not involve any causal influences from cognitive states.

## 2. Early Vision and why it is CI, Part 1.

Early vision includes a feed forward sweep (FFS) in which signals are transmitted bottom-up. In visual areas (from LGN to FEF) FFS lasts for about 100 ms . Early vision also includes a stage at which lateral and recurrent processes that are restricted within the visual areas and do not involve signals from cognitive centers occur. Recurrent processing starts at $80-100 \mathrm{~ms}$ and culminates at $120-150 \mathrm{~ms}$. Lamme (2003) calls it local recurrent processing (LRP). The unconscious FFS extracts high-level information that leads to categorization, and results in some initial feature detection. LRP produces further binding and segregation. The LRP is needed because,
owing to the small receptive fields of the neurons in V1 and V2, only local information can be coded at this level. The segmentation and recognition of the objects in a visual scene requires a more global analysis of the visual scene that can be achieved in higher areas, such as V4 or MT/V5, where the neurons' receptive fields o are larger and integrate information across longer distances in the visual field.

The feedback projections provide this global analysis that allows object segmentation, figure/ground separation, and object recognition. In the case of the MT/V5 feedback to V1, there is evidence (Pomp et al, 2015) that this feedback increases the responsiveness of the neurons in V1 especially for low-salience, small signals, which means that the recurrent signals from MT/V5 may serve to disambiguate sub-optimal visual input with respect both to the spatial location and motion of the sub-optimal signals and to their content. In addition, the feedback signals may be used to inform V1 where a change has happened in the visual scene. By not involving signals from the cognitive areas of the brain, FFS and LRP are cognitively impenetrable since the transmission of signals within the visual system is not affected by top-down signals produced in cognitive areas.

The processes of early vision retrieve from the environment information that allows the perception of a visual scene with as much accuracy as possible. To do so, early vision gradually constructs representations of increasing complexity (from variations in light intensities it extracts edges, from edges blobs, from blobs it extracts twodimensional surfaces, and from these the $21 / 2$ sketch). The output of early vision consists in the dynamic structural description of a visual scene corresponding to Marr's 21/2 sketch to which one could add the affordances of objects.

Current research on predictive coding sheds light on the nature of the processes implicated in vision. Applying this to early vision, one gets the following. The top-down and lateral effects in early vision aim to test hypotheses concerning the putative distal causes of the sensory data encoded in the hierarchically lower neuronal assemblies. In this testing, predictions made on the basis of hypotheses about the sensory information that the lower levels should encode assuming that the hypotheses are correct, are tested against the actual sensory information encoded at the lower levels. The hypothesis that best matches the sensory data is selected.

To form hypotheses concerning the probable cause of the sensory data at a certain level, at a specific spatial, and temporal scale, the neuronal assembly at the next level uses information not only about the sensory data at the previous level (or, to be precise, information regarding its prediction error) that is transmitted bottom-up, but also higher-level information that is transmitted either laterally, that is, from neuronal assemblies at the same level (neurons in V1 processing wave-lengths inform other neurons in V1 processing shape information), or top-down from levels higher
in the hierarchy (neurons in V4 are informed about the color of incoming information from neurons in IT). This lateral and top-down processing provides the context in which each neuronal assembly constructs the most probable hypothesis that would explain the sensory data at the lower level.

Since $90 \%$ of the information transmitted by neurons is transmitted within the first 100 ms of the neurons' activation as a response to a stimulus, information to neurons transmitted from other assemblies can affect their activity only if it arrives within the 100 ms time frame (Bullier 2001, 98). In order for the recurrent signals to modulate the activity at the reentered sites, they should reenter them during these crucial 100 ms . Thus, for some signal from V4 or MT/V5, which receive feedforward signals from V 1 , to reenter V 1 in time to influence the activation of the V1's neurons, the loop consisting of feedforward signals from V1 to V4 or MT/V5 and the recurrent signals from V4 or MT/V5 back to V1 must have been completed in less than 100 ms .

To put things into perspective, let us revisit Bullier's (2001) 'reintroinjection' view as it pertains to early vision and involves MT/V5 and its interaction with the lower visual areas V1 and V2. Low spatial frequency (LSF) signals precede high spatial frequency (HSF) signals. LSF information is transmitted through fast magnocellular pathways, while HSF information is transmitted through slower parvocellular pathways; the information transmitted through M-channels reaches V1 from LGN 20 ms earlier than the information transmitted from LGN to V1 through P-channels.

The mean activation latency of the neurons in MT/V5 of the brain is 75 ms after stimulus presentation respectively. Signals arrive at these areas at about the same time as, or a bit later than, they arrive in V1 ( $50-80 \mathrm{~ms}$ ) and V2 ( 85 ms ) and much earlier than they arrive in V4 despite the fact that MT/V5 is anatomically higher than V4 (Bullier (2001, 98). MT/V5 (and FEF) are parts of the 'fast brain' and belong to the dorsal system. MT/V5 is situated in the parietal cortex. Signals from V1 can reach the MT/V5 at about the same time they reach V2, that is, within 1-2 ms. It takes less than 20 ms for the recurrent signals from MT/V5 to affect the activation of neurons in V1 and V2. So, when HSF information transmitted through the P-channels reaches V1, 20 ms after LSF information transmitted through M-channels had reached V1, the responses of the V1 neurons have been modified as a result of the top-down signals from MT/V5 that had received earlier LSF information. In addition to the fast transmission of signals through the M-channels, MT/V5 also receives fast signals directly from LGN bypassing V1 through the koniocellular pathway. Thus, under certain conditions, MT/V5 could be activated earlier that V1 (Foxe \& Simpson 2002).

Bullier $(2001,100)$ concludes "the first wave of activity that invades the visual cortex following a visual stimulus appearing in the visual scene is carried by the M channel . . . the characteristics of the M channel are well suited for such a
first-pass analysis of the visual scene." This entails that even the earliest ERP component, C1, which is elicited between 4060 ms , is not an indice of the activity of V1 alone but also likely reflects top-down influences to V1 from areas as high as MT/V5. These bottom-up and top-down interactions take place within early vision (they have latencies up to 140 ms ), but all these studies bear evidence to the existence of topdown flow of information within early vision that involves stimulus driven signals reprocessed in higher visual areas and no evidence for cognitive effects in early vision.

The reason that the picture described thus far bears no evidence to support the existence of cognitive influences at these early latencies is that the top-down signals are transmitted from MT/V5 and are part of the processing along the dorsal system. The picture presented thus far posits early latencies of the signals arriving to MT/V5 from LGN either directly or through V1, that is, it dictates a bottom-up early activation of MT/V5 that, in turn, sends feedback to V1 affecting the activations of the neurons there. At these latencies there are no top-down cognitive signals to MT/V5 and, therefore, there are no cognitive signals affecting V1.

This conclusion is reinforced by the results of a study by Plomp et al (2015). The researchers combined EEG sourceimaging and Granger-causal modeling with high temporal resolution to "investigate whether recurrent and top-down interactions between visual and attentional brain areas can be identified and distinguished at short latencies in humans." Their results confirm the fast interaction between V1 and MT/V5 reported by Bullier (2001). Their results also show that the C1 ERP waveform ( 50 ms (onset)- $80 / 90 \mathrm{~ms}$ (peak)) reflects both V1 activity and also activity in highly distributed areas situates at the occipital, parietal, and frontal lobes (FEF is in the pre-frontal lobe in an area where the dorsal pathway projects). This activity reflects the early bottom-up and topdown interactions described by Bullier that includes the V1/MT feedback loop. In contradistinction to this early recurrent activity, the parietal cortex and FEF (the later cycle of activity there), which are known to modulate perceptual processing so as to help adapt behavior to the demands of a task and context, affect posterior activity around the latency of N1 ( 170 ms after stimulus onset). Thus, top-down interactions that reflect task-specific processing of the stimuli arise at longer latencies after stimulus onset. Pomp et al. (2015, 4-5) synopsize their results as follows "at the N1 latency, driving from MT no longer showed a stimulus effect, indicating that stimulus-specific driving from MT is confined to earlier latencies, in line with its fast response properties."

As Plomp et al., $(2016,1)$ write "stimulus-evoked activity at latencies before 100 ms is traditionally considered a bottom-up process. Even at these short latencies, however, there is mounting evidence of fast recurrent interactions between visual areas, obtained from direct recordings of neural activity in animal models." This agrees with Bulier's
conclusion that during the early interaction between MT/V5 and V1 (latencies earlier than 100 ms ) the signals are stimulus driven (since the signals entering MT and processed there originate from the stimulus only), while only in the later interactions that involve cognitively driven attention, whose commands are issued according to the task demands, do cognitive factors modulate the activation of neurons in V1.

Suppose that early vision is CI. Newen and Vetter could be right that the perceptual processes within this narrow time frame produce states with such poor contents that they are not properly speaking perceptual states; they could be at most sensory states. As we have seen, however, FFS and local RP allow, in about $120-140 \mathrm{~ms}$ after stimulus onset, the construction of fairly complex representations of stimuli. There is some form of perceptual organization, which certainly includes information regarding the presence of discrete objects in a scene, their orientations, sizes, shapes or forms, motions; these features determine the structural description of objects. Thus, the output of early vision consists of information about spatio-temporal and surface properties, 3D shape viewed from the perspective of the viewer, color, texture, orientation, motion, and affordances of objects, in addition to the representations of objects as bounded, solid entities that persist in space and time (Raftopoulos 2014). I disagree, thus, with Newen and Vetter that the content of CI perceptual states is related to the perception of impoverished black and white pictures. Early vision retrieves from the visual scene an extensive range of information.

## 3. Timing the Cognitive Effects: Why Early Vision is CI, Part 2

Let us examine the evidence that Newen \& Vetter employ to substantiate their claim.

Time-resolving electrophysiological evidence showed that visual cortex is activated within 50 ms and prefrontal areas within 80 ms after visual stimulus onset. This leaves plenty of time for iterative top-down processing between 'cognitive", e.g. frontal and parietal, areas and sensory, e.g. occipital, areas, within the first $100-200 \mathrm{~ms}$ after visual stimulation (Foxe \& Simpson, 2002). Thus, complex high level and reiterative processing can happen very fast and can influence visual processing very early on (. . . Plomp, Hervais-Adelman, Astolfi, \& Michel, 2015). (Newen \&Vetter 2016, 4-5)
Newen and Vetter talk about recurrent signals that involve cognitive activity affecting visual areas at latencies 100-200 ms . Thus, they accept that the available evidence suggests that cognitive effects on visual areas are registered after 100 ms post stimulus. We discussed Plomp et al. (2015) work and their conclusion that the recurrent early activity (before the elicitation of N1) is restricted within visual areas and only after that latency does recurrent activity involving cognitive
centers register in visual areas. Thus, the early recurrent activity (up to about 170 ms ) does not involve cognitive signals and this supports the view that the processes of early vision (that lasts up to about $120-140 \mathrm{~ms}$ ) are not affected by cognitive signals. Thus, when Newen and Vetter conclude "complex high level and reiterative processing can happen very fast and can influence visual processing very early on", this very early on is not early enough to be within early vision.

Indeed, as Foxe and Simpson $(2002,139)$ state
There is clearly sufficient time for multiple iterations of interactive processing between sensory, parietal, and frontal areas during brief (e.g., 200 ms ) periods of information processing preceding motor output . . . These data strongly suggest that activity represented in the "early" ERP components such as P1 and N1 (and possibly even C1) is likely to reflect relatively late processing, after the initial volley of sensory afference through the visual system and involving topdown influences from parietal and frontal regions.
Notice, first, that the reference is to the time frame up to 200 ms and that the recurrent interactions at earlier latencies that Foxe \& Simpson report concern interactions within visual areas. In addition, the top-down signals that are generated in the higher visual areas and reenter the early visual areas within these earlier latencies result from the processing of sensory signals that arrive very quickly, through PM-channels or the koniocellular pathway, to the higher areas; "The rapid activation of prefrontal cortex following initial visual activation (within 30 ms ) suggests that this input is mediated through the faster dorsal visual stream" (Foxe \& Simpson 2002,147-148). This is in line with Bullier's views that we examined earlier. All these suggest that the higher areas at these latencies have not received as yet any signals from cognitive areas and, in this sense the signals that constitute the feedback loop are bottom-up sensory signals and top-down reprocessed and modified sensory signals. In fact, it could hardly be otherwise; since all this very early recurrent activity involves the dorsal system, there is up to date no evidence to support the existence of any cognitive effects on the dorsal system when it functions on line to support fast action.

Furthermore, and in reference to Foxe and Simpson's mention of parietal and frontal regions involved in the early recurrent processing, which may be taken as evidence for the existence of cognitive influences, MT/V5 is in the parietal cortex and FEF is in the prefrontal cortex. Our discussion concerning the role of MT/V5 shows that there are no cognitive effects in the early latencies we discuss and, as our examination of the role of FEF will show, neither are such cognitive effects found in the early activation of FEF and in its role in the early stages of perceptual processing. Foxe and Simpson $(2002,146)$ confirm this analysis by concluding that "multiple visual areas begin to contribute substantially to the surface potential and C 1 begins to reflect contributions from a
number of visual areas other than, but is likely also to include V1 (emphasis added)." Moreover, Foxe and Simpson (2002, 147), after their claim that "that sustained activation patterns within cortical areas are consistent with feedback modulation of 'lower' visual areas by 'higher' areas, as well as local intrinsic processing", add that their findings conform with the findings of Lamme (1995) and Lamme et. al. (1998) about the time frame of feedback modulation in figure-ground segregation studies with monkeys. It is well known that these studies confirm that the recurrent processes that occur at early latencies do not involve cognitive signals.

Newen and Vetter $(2016,5)$ argue that in the visual system there is strong evidence for fast top-down processing within the first 50 ms after stimulus onset, certainly between motion area V5 and primary visual cortex V1 during motion perception (Silvanto, et al., 2005). The reference to the interaction between V5 and V1 during motor perception brings into mind the foregoing discussion of LRP that is restricted within the visual areas. Let use examine these studies to see if this assumption is substantiated.

Silvanto et al., (2005) studied the role of V1 in the visual awareness of motion. Their experiments show that backprojections from extrastriate cortex influence the activations of neurons in V1 and that it is the activation in V1 that determines which information reaches awareness. Since our interest is in the latencies at which the back projections affect V1 and the sites of origin of the top-down signals, I will ignore the findings concerning motion awareness. Silvanto et al., (2005) applied TMS on V1 and V5 at different times to examine the perception of phosphenes. When subthreshold TMS (that is, TMS producing no phosphene on its own) was applied over V5 followed by a subthreshold pulse to V1, subjects did not report any phosphene. When a subthreshold pulse was applied over V5 followed $10-40 \mathrm{~ms}$ later by a suprathreshold pulse over V1, subjects reported a phosphene, which was not merely the suprathreshold V1 phosphene. Instead, it acquired features of a suprathreshold V5 phosphene since subjects reported the perception of movement, and the shape and size of their percept was a mixture of V1 and V5 phosphenes. This shows that activity in V5, which on its own is insufficient to induce a moving percept, can produce such a percept if the level of induced activity in V1 is high enough.

Silvanto et al. $(2005,143)$ conclude that the fact "that moving phosphenes are perceived only when suprathreshold V1 stimulation follows, but not precedes, subthreshold V5 stimulation, together with the gradual increase in motion perception from the $10-50 \mathrm{~ms}$ period, precludes a simple feedforward summation account and points instead to a critical time of backprojection arrival in V1." They also note that the narrow time window for V5-V1 interaction ( $10-50 \mathrm{~ms}$ ) is consistent with previous reports of extrastriate-striate feedback interactions in motion during this time interval. Indeed, this accords with Bullier's (2001) and Plomp et al. (2016) finding
that there is an early (up to 100 ms ) phase of recurrent activity between V1 and MT/V5, but as in these studies, so in Silvanto et al. (2005) report, there is no evidence to suggest top-down cognitive effects at these early latencies, because the recurrent signals from MT/V5 are stimulus driven, or, to use Plomp et al. (2016) term, they are a stimulus-evoked activity.

Next, Newen and Vetter $(2016,5)$ examine the interaction between FEF and V1.
[T]he frontal eye fields (FEF), a higher-level area in frontal cortex involved in motor planning of eye movements, exerts its influence to V5 within 30 ms (Silvanto, Lavie, \& Walsh, 2006). Therefore, a feedback loop from a frontal region to an early occipital region can take as little as 80 ms or less . . . when the task requires face recognition, FEF signals are sent to face-sensitive regions and when the task requires motion discrimination, FEF signals are sent to motion area V5, both within a time frame of $20-40 \mathrm{~ms}$ after FEF activity (Morishima et al., 2009).
FEF is situated in the prefrontal cortex at a site that is heavily interconnected with the parietal cortex and is considered a part of the dorsal system. The mean activation latency of the neurons in FEF is 70 ms after stimulus presentation. Signals arrive at FEF with a slight time delay time with respect to the signals arriving at V1 (50-80 ms) and V2 ( 85 ms ) and much earlier than they arrive at V4 despite the fact that FEF is anatomically higher than V4 (Taylor and Nobre 2007). FEF TMS affects the detection of targets in arrays of distractors and these effects are apparent when pulses are applied early ( 40 and 80 ms ) after presentation of the visual array (O'Shea et al. 2004). HSF signals from V1 can reach FEF in 50-100 ms.

FEF contains visual and movement neurons. Studies (see O'Shea et al. $(2004,1060)$ for a discussion) show that there are two dissociated processing operations in FEF; the target selection by FEF visual neurons and saccade programming by movement neurons. Studies (O'Shea et al., 2016; Silvanto et al., 2006; Taylor \& Nobre 2007) show that some FEF responses are independent of saccades to targets and respond to the visual stimuli. Some of the FEF feedback signals play a role in the perception of a visual scene by affecting in a topdown manner the earlier visual areas. FEF plays a crucial role in visual target discrimination that is independent of saccade programming, as TMS applied to FEF impairs performance in target discrimination tasks if applied between $40-80 \mathrm{~ms}$ after stimulus onset (O'Shea et al. 2004). In addition, these visual neurons of FEF are thought to be associated with top-down or endogenous attention (Taylor and Nobre 2007).

Accepting O'Shea et al. (2004) early latencies of FEF neurons in discriminating targets from non targets (100-120 ms ), in view of the fact that, as Newen and Vetter $(2016,5)$ also accept, FEF exerts its influence to V5 within 30 ms and, therefore, a feedback loop from FEF to an early occipital
region can take as little as 80 ms or less, the total time it takes for the FEF neurons that have distinguished the targets from non targets to affect via top-down feedback projections the early visual areas is about $180-200 \mathrm{~ms}$, considering that the target discrimination in FEF reported by O'Shea t al. (2004) occurs at $100-120 \mathrm{~ms}$. This means that the FEF effects the activation of the neurons in early visual areas with a latency that places these effects outside early vision.

Concerning the finding that FEF neurons effectively discriminate targets from non targets as early as $100-120 \mathrm{~ms}$ after stimulus onset, one could argue that since this discrimination is task relevant and involves cognitive factors, cognition affects a visual area, FEF, within the timing of early vision. O'Shea et al. (2004), think it very likely that the early latency they report is the result of feature pre cueing, which means that the early activity in FEF occurs as the result of a cognitive demand issued before the appearance of the stimulus. I have argued (Raftopoulos 2014) that the cognitive effects on perception through pre-cueing are not cases of CP because they do not affect directly early vision and do not affect its epistemic role in grounding empirical beliefs.

Silvanto et al., (2006), whose study is cited by Newen and Vetter (2016) as showing that early vision is CP, found that stimulation applied to FEF $20-40 \mathrm{~ms}$ prior to the stimulation of MT/V5 decreases the intensity of the MT/V5 stimulation required to elicit phosphenes, which entails that the activity of MT/V5 is modulated by the activity in FEF. FEF has also been found to modulate top-down V4. Silvanto et al (2006, 944) claim that the content of to-down control may be either spatial or feature related, which means that they think that FEF affects the control of top-down attention; "an area involved in control would be expected to be active early and by responding to target features, the FEF could increase the sensitivity of extrastriate neurons to task relevant parameters." (Silvanto et al. 206, 944) With regards to how FEF exerts topdown control, it is possible that FEF activity occurs prior to sensory stimulation as opposed to rapid responses to visual stimuli since FEF neurons may also play a role in visual priming (Silvanto et al. 2006, 944). Thus, as Taylor and Nobre (2007), so Silvanto et al., (2006) think that FEF controls the allocation of top-down attention prior to stimulus presentation.

The discrimination between targets and non-targets depends on the task at hand and is cognitively driven. Thus, the top-down effects that result from this discrimination are also cognitively-driven and the visual processes that are thus affected are clearly CP. Accepting O'Shea et al. (2004) early latencies of FEF neurons in discriminating targets from non targets, since FEF exerts its influence to V5 within 30 ms and, therefore, a feedback loop from FEF to an early occipital region can take as little as 80 ms , the total time it takes for the FEF neurons that have distinguished the targets from non targets to affect via top-down feedback projections MT/V5 is $130-150 \mathrm{~ms}$, and the effects on the early visual areas is about

180-200 ms, considering that the target discrimination in FEF in O'Shea t al. (2004) occurs at $100-120 \mathrm{~ms}$. All of these are outside the timing of early vision and do not entail its CP.

Finally, Newen and Vetter (2016, 5), appeal to a study by Drewes et al. (2016) that shows that object recognition involves recurrent processing with a time constant of 60 ms . Drewes et al. (2016) examine the view that since the visual system extracts from object information, for example, the shape of objects, very fast this entails that the underlying cortical processing should be strictly feedforward. Against this, their study suggests that in shape perception there is a recurrent circuit, which is not an attentional cueing effect but reflects "the time course of feedback processing underlying the rapid organization of shape." (Drewes et al. 2016, 185)

In their introduction, they mention work by Heinen et al. (2005) suggesting that the figure-ground segregation requires two distinct periods of information processing in the early visual areas, an early one around $130-160 \mathrm{~ms}$ and a later one around $250-280 \mathrm{~ms}$ after stimulus onset, and by Wokke et al. (2012) showings that recurrent processing engages the early visual areas (V1/V2) to participate in more complex visual tasks. In an early time window (96-119 msec), detection of figure stimuli and of neural correlates of figure border detection and border ownership occurs. Later (236-259 msec) V1 and V2 participate in surface segregation. Drewes et al. (2016) accept these latencies as a general framework.

Drewes et al. $(2016,190)$ claim that "the extent of facilitation between two shape stimuli depends nonmonotonically on the delay between their presentations, peaking at a delay of 60 ms ." This suggests a recurrent circuit underlying shape processing in the cortical object pathway. They remark that in Wokke's et al. (2012) study TMS was applied to the occipital pole to disrupt processing in V1/V2 or to the lateral occipital lobe to disrupt processing in the LOC. TMS disrupted performance at both locations but at different latencies. In LOC, TMS disrupted processing when the pulse occurred $100-122 \mathrm{~ms}$ post stimulus, while in V1/V2, processing was disrupted when the pulse was applied 160-182 ms post stimulus. This shows a feedback process in the grouping of contour fragments to form shape with a one-way feedback time constant (LOC to V1/V2) of $40-80 \mathrm{~ms}$. Given the 60 ms time constant, the top-down signals reenter V1 and V2 at latencies outside early vision.

## 4. Conclusion

I examined the evidence Newen and Vetter (2016) adduce to support the claim that early vision is CP. None of it supports the existence of direct cognitive effects on early vision. Finally, concerning the claim that a stage of visual processing that is CI is so impoverished that it would not be worthy to be called a stage of perception, I claimed that early vision delivers a rich structure.

## References

Bullier, J. (2001). Integrated model of visual processing. Brain Research Reviews, 36, 96-107.
Drewes, J., Goren, G., Zhu, W., \& Elder, J/H. (2016). Recurrent processing in the formation of percept shapes. The Journal of Neuroscience, 36(1), 185-192.
Heinen, K., Jolij, J., \& Lamme, V. (2005). Figure-ground segregation requires two distinct periods of activity in V1: a transcranial magnetic study. .Neuroreport, 16(13), 14831487.

Lamme, V. A. F. (1995). The neurophysiology of figureground segregation in primary visual cortex. Journal of Neuroscience, 15, 1605-1615.
Lamme, V. A. F. (2003). Why visual attention and awareness are different. Trends in Cognitive Sciences, 7 (1), 12-18.
Lamme VA, Zipser K, Spekreijse H (1998). Figure-ground activity in primary visual cortex is suppressed by anesthesia. Proccedings National Academy Science USA, 95, 3263-3268.
Morishima, Y., Akaishi, R., Yamada, Y., Okuda, J., Toma, K., \& Sakai, K. (2008). Task-specific signal transmission from prefrontal cortex in visual selective attention. Nature Neuroscience, 12(1), 85-90.
Newen, A. \& Vetter, P. (2016). Why cognitive penetration of our perceptual experience is still the most plausible account. Consciousness and Cognition, http://dx.doi.org/10.1016/j.concog.2016.09.005.
O'Shea, J., Muggleton, N.G., Cowey, A., \& Walsh, V. (2004). Timing of target discrimination in human front eye fields. Journal of Cognitive Neuroscience, 16(6), 1060-1067.
Plomp, G., Hervais-Adelma, A. Astofli, L., \& Michel, C. M. (2015). Early recurrence and ongoing parietal driving during elementary visual processing. Nature, Scientific Reports, 5:18733, doi: 10.1038/srep 18733.
Raftopoulos, A. (2001). Is perception informationally encapsulated? The Issue of the Theory-Ladenness of Perception. Cognitive Science, 25, 423-451.
Raftopoulos, A. (2014). The cognitive impenetrability of the content of early vision is a necessary and sufficient condition for purely nonconceptual content. Philosophical Psychology, 27 (5), 601-620.
Silvanto, J., Cowey, A., Lavie, N., \& Walsh, V. (2005). Striate cortex (V1) activity gates awareness of motion. Nature Neuroscience, 8(2), 143-144.
Silvanto, J., Lavie, N., \& Walsh, V. (2006). Stimulation of the human frontal eye fields modulates sensitivity of extrastriate visual cortex. Journal of Neurophysiology, 96(2), 941-945.
Taylor, P.C.J., \& Nobre, A. (2007). FEF TMS affects visual cortical activity. Cerebral Cortex, 17, 391-399.
Wokke, M., Sligte, I., Scholte, H., \& Lamme, V.A.F. (2012). Two critical periods in early visual cortex during figureground segregation. Brain and Behavior, 2(6), 763-777.

# The Wason Selection Task: A Meta-Analysis 

Marco Ragni (ragni@,informatik.uni-freiburg.de)<br>Cognitive Computation Lab, University of Freiburg, 79110 Freiburg, Germany<br>Ilir Kola (kola@informatik.uni-freiburg.de)<br>Cognitive Computation Lab, University of Freiburg, 79110 Freiburg, Germany

P. N. Johnson-Laird (phil@princeton.edu)<br>Princeton University, Princeton NJ 08540 USA and New York University, New York, NY 10003, USA


#### Abstract

In Wason's selection task, participants select whichever of four cards could provide evidence about the truth or falsity of a conditional rule. As our meta-analysis of hundreds of experiments corroborates, participants tend to overlook one of the cards that could falsify the rule. 15 distinct theories aim to explain this phenomenon and others, but many of them presuppose that cards are selected independently of one another. We show that this assumption is false: Shannon's entropy for selections is reliably redundant in comparison with those of 10,000 simulated experiments using the same four individual probabilities for each real experiment. This result rules out those theories presupposing independent selections. Of the remaining theories, only two predict the frequencies of selections, one (due to Johnson-Laird \& Wason, 1970a) provides a better fit to the experimental data than the other (due to Klauer et al., 2007). We discuss the implications of these results.


Keywords: Conditional reasoning; Entropy; Falsity, Selection task; Mental models.

## Introduction

Human beings are able to evaluate whether assertions are true, and to select evidence relevant to such evaluations. The late Peter Wason (1968) carried out a paradigmatic study to test whether naive individuals grasped the relevance of falsification. In his original "selection" task, the experimenter explains to the participants that there is a pack of cards in which each card has a letter on one side and a number on the other side. Four cards are taken at random from the pack, and placed in front of the participant (see Fig. 1). The experimenter then presents the rule:

If there is a $D$ on one side of a card, then there is a 3 on the other side.
The participants' task is to select just those cards that, if turned over, would show whether or not the rule is true or false of the four cards. The task is a demonstration, not an experiment, because it has no independent variable.

Participants tend to select the $D$ card alone, or the $D$ and 3 cards, but rarely the $D$ and 7 cards. Yet, if the 7 has D on its other side, the rule would be false. This failure to falsify was shocking. Perhaps as a consequence more than 300 experiments investigating the task have been published over the last 50 years.

In order to try to understand performance, psychologists developed various versions of the task. They explored rules of different sorts, such as disjunctions and rules framed


Rule: If a card has a $D$ on one side, then it has a 3 on the other side
Figure 1. The four cards in Wason's selection task. Each has a number on one side and a letter on the other side. The participants' task is to select just those cards that, if turned over, would show whether or not the rule shown above holds for the four cards. The letters $p, q$, etc. are added for illustrative purposes as the rule is of the sort, if $p$ then $q$.
with "every" in place of "if" (Wason \& Johnson-Laird, 1969; Wason \& Shapiro, 1971), cards with all the information on one side but partly masked, choices of just two cards (e.g., Johnson-Laird \& Wason, 1970b), or choices of multiple cards, with repetitions of one or more cards (e.g., Oaksford \& Chater, 1994). But, two main versions elicited better performance than abstract rules, such as the one in Fig. 1. One version used everyday rules, such as one about destinations and modes of transport (Wason \& Shapiro, 1971). The other version switched the task around so that participants had to select those cards representing individuals who might be violating a deontic rule (e.g., Griggs \& Cox, 1982), such as:

If a person is drinking beer, then the person must be
over 19 years of age.
The efficacy of some deontic rules, such as one about the amount of postage on letters (Johnson-Laird, Legrenzi, \& Legrenzi, 1972), depended on the participants' familiarity with them, but not all do so.

As the number of experimental studies grew, so too did the number of theories. By our reckoning, there are at least 15 distinct theories of the selection task including ones based on the meaning of conditionals, on formal rules of inference for them, on heuristics such as "matching" in which participants merely select those cards referred to in the rule (Evans, 1977), on content-specific rules of inference, and on the probabilities with which the various items on the cards occur in reality (Oaksford \& Chater, 1994). Given that the selection task has been under investigation
for half a century, the existence of 15 theories about it is embarrassing for cognitive science. Our aim in what follows is therefore to describe meta-analyses of the experiments that aimed to eliminate as many theories as possible.

## Meta-analyses

## The reliability of the results

We searched the literature for experiments on the selection task with the proviso that they used a conditional rule of the sort: if $p$ then $q$, and that they reported at least the frequencies of the four canonical selections of $\mathrm{p}, \mathrm{pq}$, $\mathrm{pq} \overline{\mathrm{q}}$, and $\mathrm{p} \bar{q}$, which the early studies had reported. Henceforth, we abbreviate selections in the preceding way, stating which of the 4 cards they included, e.g., pq denotes a selection of the p and q cards (see Fig. 1). We divided the resulting experiments into three categories according to the nature of the rule they used: abstract, everyday, or deontic. We also classified them according to whether they reported the frequencies of only the 4 canonical selections and a category of "other" selections, or the frequencies of all 16 possible selections. The studies can be found at http://www.cc.uni-freiburg.de/data.
Because the first studies were carried out half century ago and subsequent ones in many countries, their results might be too heterogeneous for an informative test of the theories. We assessed the overall homogeneity of the results for the three categories of task from the reliability of the rank orders of the frequencies of their canonical selections. Table 1 reports Kendall's coefficient of concordance, W, which ranges from 0 for no consensus to 1 for perfect consensus, for the three categories of task. The results show a reasonable and robust consensus over the experiments. Table 2 presents the overall percentages of each of the four canonical selections for the three sorts of selection task. It shows why the deontic task yielded a greater concordance, W: the majority of participants selected cards denoting potential violations of the rule.

Table 1. The concordance across different experiments examining the three main sorts of selection tasks as assessed with Kendall's coefficient of concordance, W, and stating its $\chi^{2}$ and $p$ values.

| Three sorts of <br> selection task | Number of <br> experiments | Kendall's | $\chi^{2}$ and p value |
| :---: | :---: | :---: | :---: |
| Abstract | 104 | $\mathrm{~W}=.34$ | $107, \mathrm{p}<.001$ |
| Everyday | 44 | $\mathrm{~W}=.25$ | $33, \mathrm{p}<.001$ |
| Deontic | 80 | $\mathrm{~W}=.54$ | $29, \mathrm{p}<.001$ |

## The redundancy of the selections

Many studies of the selection task report only the four separate probabilities with which participants selected each of the cards (e.g., Evans, 1977). These results, however, make sense only if the selection of each card is independent of the others. Some investigations have reported this independence (e.g., Evans, 1977). But, others have refuted it by establishing correlations between the selections (Pollard,

1985; Oaksford \& Chater, 1994). Correlations, however, are only among pairs of cards in selections. A better assessment would take into account each selection as a whole

Table 2. The percentages of each of the four canonical selections for the three sorts of selection task

|  | The canonical selections |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | p | pq | $\mathrm{pq} \overline{\mathrm{q}}$ | $\mathrm{p} \overline{\mathrm{q}}$ |
| Abstract | 36 | 39 | 5 | 19 |
| Everyday | 23 | 37 | 11 | 29 |
| Deontic | 13 | 19 | 4 | 64 |

and all the selections made in an experiment. We therefore introduced a new procedure that combines Shannon's measure of entropy (or informativeness) with the computer simulation of thousands of experiments. The underlying intuition is straightforward. Suppose the selections in an experiment are more redundant - more predictable - than a prediction made solely from the frequencies of selecting each of the four individual cards in the experiment. It follows that the cards in selections are, not independent of one another, but interdependent. And some aspect in the process of selecting cards yields the redundancy.

The first step in our procedure is to compute the amount of information in the selections in an experiment, i.e., the difficulty of predicting them. We use Shannon's measure of entropy:

$$
\mathrm{H}=-\Sigma \mathrm{P}_{\mathrm{i}} \log _{2} \mathrm{P}_{\mathrm{i}}
$$

for the set of selections, where $P_{i}$ denotes the probability of the i-th selection, and $\log _{2}$ denotes a logarithm to the base 2. In general, the greater the number of different selections, and the more evenly distributed the frequencies over them, so the value of H increases, and it is harder to predict the selections. If participants chose each card independently of the others, the value of H for the experiment would not differ reliably from its value for selections derived from sampling according to the four probabilities for selecting each card. But, if the value of H for the selections in the experiment is reliably smaller than this theoretical value, then we can reject the null hypothesis of independent selections. In other words, the redundancy reflected in a smaller value of H reflects interdependence in the selections.

As an illustrative example, consider the selections in Experiment 2 of Stahl et al. (2008), which we choose because of its large number of participants: 351 . Here are the frequencies of the selections, in which 6 participants selected none of the cards:

$$
\text { p 92, pq 99, pq } \bar{q} 2, p \bar{q} 20, p \bar{p} q \bar{q} 19, \bar{p} q 6, p \bar{p} \bar{q} 2,
$$

$\bar{p} q \bar{q} 2, \bar{q} 18, \bar{p} \bar{q} 22, p \bar{p} 7, q \bar{q} 6, \bar{p} 7, q 43$, none 6 .
They show that the probabilities of selecting each of the four cards were as follows:

$$
\mathrm{p} 0.69, \mathrm{q} 0.49, \overline{\mathrm{q}} 0.26, \overline{\mathrm{p}} 0.19 .
$$

The value of H for the selections in the experiment is 2.8 bits. Could this value have occurred by chance? We used a resampling procedure to find out its chance probability (see, e.g., Good, 2001). We ran a computer program to carry out 10,000 simulated experiments based both on the
number of participants in the original study and on its probabilities above of selecting the four individual cards. The resulting mean value of H was 3.13 , which shows that the observed selections in the experiment have a redundancy of 0.33 . More important, however, is that not one of the simulated experiments yielded an entropy as low as 2.8 bits, and so the difference is statistically significant ( $\mathrm{p}<.0001$ ). The redundancy in the original experiment did not occur by chance. In summary, a statistically significant degree of redundancy in selections in an experiment is evidence for their interdependence.

We programmed an algorithm based on the same idea. Its key difference from our analysis of Stahl's data above is that it concerns only the four canonical selections. This constraint is necessary because so many experimental reports state the results only for them. Four selections have a maximum entropy of 2 bits if they are each equiprobable. The mean over the 228 experiments (in Table 1) is 1.27 bits (with a standard deviation of 0.48 ). The input to the program states the number of participants and the frequencies of the four selections for each experiment in the set. Its main steps are as follows. For each experiment:

1. Compute N , the number of participants, and the probabilities with which each of the 4 cards occurred in the experiment's selections.
2. Compute Shannon's entropy H for the experiment.
3. Carry out 10,000 simulated experiments based on the probabilities of selecting each card, assigning a selection to each of the N participants.
4. Return the number of simulated experiments with a higher entropy than the actual experiment and the number of them with the same or a lower entropy.

Table 3. The mean entropies (in bits) of 228 experiments on three sorts of selection task, the mean entropies of sets of 10,000 simulations of each experiment, and Wilcoxon's tests ( W , and its p -value) of the difference between them.

| The three <br> sorts of selec- <br> tion task | Mean entropy <br> of experiments of sets of <br> simulations | Mean entropy | Wilcoxon's <br> W and p-value |
| :--- | :---: | :---: | :---: |
| Abstract | 1.32 | 1.42 | $\mathrm{~W}=469, \mathrm{p}<.001$ |
| Everyday | 1.51 | 1.66 | $\mathrm{~W}=28, \mathrm{p}<.001$ |
| Deontic | 1.06 | 1.21 | $\mathrm{~W}=68, \mathrm{p}<.001$ |

Table 3 presents the mean entropies of the 228 experiments investigating the three sorts of selection task, the mean entropies of each of their 10,000 simulations, and the results of Wilcoxon's W test and its p -value comparing the pairs of means. These results allow us to reject the null hypothesis of independent selections. The redundancy shown in the smaller entropies of real experiments over simulated ones shows that the cards in selections are not selected independently of one another. They are selected in an interdependent way. This result eliminates any theory that predicts that selections are independent.

## Theories of the selection task

Some theories of the selection task are informal and make only qualitative predictions about selections (e.g., Wason, 1968). Some predict only whether the correlations between selecting the possible pairs of cards are positive or negative (Oaksford \& Chater, 1994). Some predict only the probabilities of selecting each of the four cards (Evans, 1977; Hattori, 2002; Oaksford \& Wakefield, 2003). We discount all of these theories as insufficiently powerful to make quantitative predictions about the frequencies of the canonical selections, let alone all 16 possible selections. There remain just two theories, which we now outline.

## The insight model

The first algorithms to model the mental processes underlying the selection task were due to Johnson-Laird and Wason (1970a). Their principal algorithm posits three levels of insight into the importance of falsification: no insight, which implies that reasoners select only cards referred to in the rule - an anticipation of "matching" bias (Evans, 1972); partial insight, which implies that reasoners consider all the cards, adding any further cards that verify the rule, or, failing that, that falsify the rule; and complete insight, which implies that reasoners select only cards that can falsify the rule. The algorithm was published as a flow chart, but not implemented, because of a lack of access to a main-frame computer. We recently programmed it, replacing its use of truth tables with mental models and fully explicit models, simplifying its processes, but keeping its original functionality so the program makes the identical predictions to the original version.

Given a rule of the sort if $p$ then $q$, the program begins by compiling a list of cards to select, and its first step is to scan its mental model of the conditional, and as a result to put $p$ on this list. If the program also scans the model in the opposite direction, it adds q to the list. With no insight into the task, these selections verify the rule. However, the program implements two interrelated levels of insight. Partial insight is to assess all the cards, and to add any further card that verifies the rule, or, if none does, to add any that falsifies the rule. So, if $q$ is already in the list, partial insight adds $\bar{q}$, because it can falsify the rule, yielding the selection $\mathrm{pq} \overline{\mathrm{q}}$. Complete insight is to select only cards that can falsify the rule, and yields the selection $\mathrm{p} \overline{\mathrm{q}}$. Complete insight occurs only if all the cards are examined. An explicit biconditional as an input yields a selection of all four cards in certain cases, e.g., when it scans its model in both directions with partial insight.

Fig. 2 presents, not the algorithm, but a tree diagram summarizing its parameters and its predictions for conditionals and biconditionals. As it illustrates, the algorithm produces the same selections as a result of different processes, and it is not deterministic, i.e., nothing in the algorithm determines the level of insight (pace Evans, 1977, who took the algorithm to be deterministic). The predictions in Fig. 2 explain why selections should be interdependent, e.g., verifying cards include q only if they include p and falsifying cards include $\overline{\mathrm{q}}$ if and only if they include
p. The only exceptions to the algorithm's outputs should be the result of guessing or haphazard errors. In fact, these exceptions occur at a rate less than chance in the 288 experiments.


Figure 2. The predictions of the insight model (JohnsonLaird \& Wason, 1970a) as a binary decision tree. Each decision is controlled in its recent implementation by a parameter (see text). Participants with no insight select only cards referred to in the rule. Those with partial insight consider all cards, selecting any further card that can verify the
rule, or, failing that, that can falsify it. Participants with complete insight select only cards that can falsify the rule.

Our implementation of the algorithm contains three probabilistic parameters in the unit interval from 0 to 1 . The first parameter, $c$, is the probability of scanning the model in both directions as opposed to scanning in only one direction. The second parameter, $e$, is the probability of examining all four cards, and if the result fails to add any card that verifies the rule, adding any card that falsifies it. This corresponds to partial insight. The third parameter, $f$, is the probability of complete insight, which makes only a falsifying selection.

## The inference-guessing model

Klauer et al. (2007) proposed a set of related theories, including one with a heuristic component allowing for guessing, and an inferential component. There is no algorithm the implements the theory's underlying processes, but its predictions were modeled in a binary tree. This model has 10 parameters, which are each the probability that one sort of process occurs rather than another, and so each is in the unit interval from 0 to 1 . The model's first parameter is the probability that the inference governs the selection as opposed to guessing. The guessing component makes independent selections of each of the four cards according to four parameters that are the respective probabilities of selecting each of them independently as a result of guessing or any heuristic factor such as "matching" (Evans, 1977). The theory assumes that selections are governed, not by the meaning of the rule, but by inferences from the rule. The particular inferences depends on five parameters:

1. The probability that the rule, if $p$ then $q$, is interpreted as a conditional as opposed to a biconditional.
2. The probability that the inference is forwards from the if-clause: modus ponens (MP) or denial of the antecedent (DA), as opposed to backwards from the then-clause: modus tollens (MT) or affirmation of the consequent (AC).
3. Given the biconditional interpretation, the probability that the interpretation is bidirectional, if $p$ then $q \&$ if $q$ then $p$, as opposed to a case distinction, if $p$ then $q$ \& if not- $p$ then not- $q$. With the bidirectional interpretation, the distinction between forwards and backwards inferences does not apply - both are made, but with a case distinction interpretation, the distinction still applies.
4. The probability that an inference from a conditional or a bidirectional biconditional is a sufficient one as opposed to a necessary one. Normally, $p$ is judged sufficient to infer $q$ from if $p$ then $q$, but sometimes $p$ is judged necessary to infer $q$, as when the conditional is interpreted as stating an enabling condition akin to only if $p$ then $q$. A forward sufficient inference is MP, whereas a forward necessary inference is DA; and a backward sufficient inference is AC, whereas a backward necessary inference is MT.
5. The probability that inferences are made only about the visible sides of cards as opposed to the invisible sides of cards too, i.e., individuals can envisage items on them.

The model contains 10 parameters but the data are the frequencies of the four canonical selections. Hence, to ensure that the process of fitting model to data converges and does not overfit the data, we implemented a restricted in-ference-guessing model that makes the four canonical selections. Fig. 3 summarizes the predictions of this restricted inference-guessing model. The reasoning component in the original model makes no more than two inferences on a trial, and so it cannot make the canonical selection of three cards: $\mathrm{pq} \overline{\mathrm{q}}$. We therefore changed the original guessing component to make this selection.


Figure 3. A restricted version of the binary decision tree of the inference-guessing model (Klauer et al., 2007) for the 4 canonical selections. Each decision is controlled by a parameter (see text).

The two models are based on the only theories that we could find in the literature that can be programmed with parameters that fit data about the frequencies of selections.

## An evaluation of the two models

We evaluated the insight model with 3 parameters (John-son-Laird \& Wason, 1970a) and the restricted inferenceguessing model with 4 parameters (cf. Klauer et al., 2007). Their respective predictions can be represented as trees of
binary decisions (see Fig. 2 and Fig. 3). Both models invoke alternative sequences of processes depending either on three decisions in the insight model or four decisions in the inference-guessing model. Because each model's predictions correspond to a tree of decisions, we evaluated each of them as a multinomial processing tree (MPT) in which the probability of a particular cognitive state is estimated from the observed frequencies of selections (Riefer \& Batchelder, 1988). A program fitted each of the two models to the frequencies of the canonical selections of the three sorts of selection task: 104 experiments with the abstract task, 44 experiments with everyday task, and 80 experiments with deontic task (see Tables 1-3 above). We used the maximum-likelihood method from the R-package for multinomial processing trees (the MPTinR of Singmann \& Kellen, 2012). We calculated three measures to compare the goodness of fits of the two theories:-

- The root mean square errors (RMSEs) of the fits.
- The Bayesian information criterion (BIC), which indicates how much information is lost when a model represents the process that generates the data, taking into account both its goodness of fit and its number of parameters. It penalizes models according to the number of their parameters, and the smaller its value, the better the fit between a model and the data.
- The Bayes factor (BF; Schwarz, 1978), which is a Bayesian method to compare different models. It uses an approximation of the difference between the BIC value of model 1 and BIC value of model 2 as computed by MPTinR. The higher its value between 30 and 100 , the stronger the support for model 1 over model 2 (Wagenmakers et al., 2011).

Table 4 presents the three measures for each of the two models. As it shows, the insight model with three parameters has a closer fits, and lower BIC values, than the restricted inference-guessing model. The Bayesian factor likewise shows stronger evidence for the insight model than for the restricted inference-guessing model. The insight model has the advantage of fewer parameters. As a theory, it is simpler because it relies on the meaning of the rule rather than inferences from it, and because it has no machinery to account for selections that occur at a rate less than chance. But, it is not a paragon, and we explain why below.

## General Discussion

Half a century of research and over 300 articles should have led to a single unique theory of a cognitive task rather than to 15 different theories. That was the situation for Wason's selection task. The present research, however, has eliminated all but one theory. And it did so using the following strategy. It established a large but representative set of experiments investigating rules of the sort if $p$ then $q$ that had a reliable concordance in their results (Table 1). These results established the rarity of falsifying selections,
$p \bar{q}$, except when they violate a deontic rule Table 2). The four canonical selections ( $\mathrm{p}, \mathrm{pq}, \mathrm{pq} \overline{\mathrm{q}}$, and $\mathrm{p} \overline{\mathrm{q}}$ ) are reliably redundant in most experiments in comparisons of each experiment's entropy (informativeness) with the entropy of its

Table 4. The insight model's and the restricted inferenceguessing model's goodness of fit with the individual canonical selections for 288 experiments overall and for the three sorts of selection task: the root mean square errors (RMSE) for their predictions, their Bayesian information criteria (BIC), and the Bayes factors for the better-fitting model.

| The 3 sorts <br> of the sele- <br> ction task | Cognitive <br> model | RMSE | Bayesian <br> Information <br> Criterion (BIC) | Bayes <br> factor |
| :--- | :--- | :---: | :---: | :---: |
| Overall | Insight | 2.69 | 27.7 |  |
|  | Inference- <br> guessing | 19.35 | 37.0 | 99.5 |
| Abstract | Insight | 1.97 | 25.7 | 73.7 |
|  | Inference- <br> guessing | 3.28 | 34.3 | 47 |
| Everyday | Insight | 1.7 | 23.2 | 49.4 |
|  | Inference- <br> guessing | 2.18 | 30.9 | 23.5 |
| Deontic | Insight 0.8 31.4  <br>  Inference- <br> guessing 1.05  |  |  |  |

10,000 simulations based on its four probabilities of selecting each card (Table 3). Not all experiments yield redundant selections, but the vast majority do. This result ruled out theories that imply that selections of cards are independent of one another. Above all, theories therefore need to predict the frequencies of the canonical selections. Perhaps surprisingly, this criterion rules out nearly all the remaining theories. Klauer et al. (2007) had programmed an MPT of their inference-guessing model using 10 parameters to make predictions for the frequencies of all 16 possible selections - most of which do not occur more often than chance. More than twice as many experiments reported the frequencies only of the four canonical selections than reported them for all 16 selections. Hence, we produced an MPT for a restricted version of the model that used four parameters to predict the frequencies of the canonical selections. To do so, we reduced the original parameters for guessing to one, which made a selection of three cards, otherwise impossible for the model to select. For the insight model, we programmed an algorithm that carried out its processes (Johnson-Laird \& Wason, 1970a), and we used it to construct an MPT model with three parameters. The insight model yielded a better fit with fewer parameters (Table 4).

The story of the selection task does not end here. But, the success of the insight theory tells us that we have returned to how it was conceived after only a handful of studies. Naive individuals focus on those cards mentioned in the rule, and select them if they can verify the rule. With a little bit of insight, they consider all the cards, and may select additional cards. With complete insight, they select only cards that can falsify the rule (Johnson-Laird \& Wason,

1970a). We now know that various factors - the competence of participants, the contents of the rule, and the framing of the task - can all enhance insight. An account along these lines seems to be correct, except perhaps when experiments implicate probabilities in their contents or framing (e.g., Oaksford \& Chater, 1994).

The excellent fit of the insight model must be viewed with caution. The number of parameters in a model is a measure of our ignorance. Those for guessing seem to be dispensable. Indeed, some selections are very odd, as we saw earlier in our analysis of the results from Stahl et al. (2008). They are so odd that they must count as irrational on any criterion: the participants erred or guessed. Introducing parameters to model guessing has no theoretical value other than to index the difficulty of a task. The insight theory has three essential parameters, and the original infer-ence-guessing model has five. The difference reflects an crucial distinction: whether people determine the truth value of an assertion based on its meaning (the insight model) or based on inferences from it (the inference-guessing model). Therein may lie the advantage of the insight model. But, we are bound to ask what mechanisms might replace its parameters. We now know that the insight to make falsifying selections depends on various factors, including intellectual ability (e.g., Stanovich \& West, 1998). Hence, it may be feasible to replace the parameter for the probability of complete insight with a measure of ability. It is even conceivable that the parameter of partial insight might reflect a lesser but above average intellect. The parameter for scanning a model of the conditional in both directions is more problematic. It may depend on the processing capacity of working memory. These speculations in no way rule out the possibility of some quite different theory of the selection task outperforming the insight model.
If our research has any general moral, it is an old one: cognitive theories should be effective procedures (JohnsonLaird, 1983, p. 6). They should be programmable.

## Acknowledgements

This research was supported by a DFG-Heisenberg fellowship DFG RA 1934/3-1, and DFG projects RA 1934/2-1 and $4-1$. We thank Linden Ball, Ruth Byrne, Nick Chater, Jonathan Evans, Keith Holyoak, Sangeet Khemlani, Ken Manktelow, Mike Oaksford, Klaus Oberauer, Lance Rips, Carlos Santamaría, Walter Schaeken, Walter Schroyens, Dan Sperber, Christoph Stahl, and Valerie Thompson, for their help. We thank Christoph Klauer for his advice and sending us his algorithm,. We dedicate this paper to the memory of Peter Wason (1924-2003), a unique and most extraordinary researcher.

## References

Evans, J. S. B. (1972). Interpretation and matching bias in a reasoning task. The Quarterly Journal of Experimental Psychology, 24, 193-199.
Evans, J. S. B. (1977). Toward a statistical theory of reasoning. Quarterly Journal of Experimental Psychology, 29, 621-635.

Good, P. I. (2001). Resampling methods: A practical guide to data analysis. NY: Birkhauser Boston.
Griggs, R. A., \& Cox, J. R. (1982). The elusive thematic materials effect in Wason's selection task. British Journal of Psychology, 73(3), 407-420.
Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. Quarterly Journal of Experimental Psychology: Section A, 55, 1241-1272.
Johnson-Laird, P. N. (1983). Mental models. Cambridge, MA: Harvard University Press.
Johnson-Laird, P. N., Legrenzi, P., \& Legrenzi, M. S. (1972). Reasoning and a sense of reality. British Journal of Psychology, 63, 395-400.
Johnson-Laird, P. N., \& Wason, P. C. (1970a). A theoretical analysis of insight into a reasoning task. Cognitive Psychology, 1, 134-148.
Johnson-Laird, P. N., \& Wason, P. C. (1970b). Insight into a logical relation. Quarterly Journal of Experimental Psychology, 22(1), 49-61.
Klauer, K. C., Stahl, C., \& Erdfelder, E. (2007). The abstract selection task: new data and an almost comprehensive model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 680-703.
Oaksford, M., \& Chater, N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review, 101, 608-631.
Oaksford, M., \& Wakefield, M. (2003). Data selection and natural sampling: Probabilities do matter. Memory \& Cognition, 31(1), 143-154.
Pollard, P. (1985). Nonindependence of selections on the Wason selection task. Bulletin of the Psychonomic Society, 23(4), 317-320.
Riefer, D. M., \& Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. Psychological Review, 95, 318-339.
Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.
Singmann, H., \& Kellen, D. (2012). MPTinR: Analysis of multinomial processing tree models in R. Behavioral Research Methods, 45, 560-575.
Stahl, C., Klauer, K.C., \& Erdfelder, E. (2008). Matching bias in the selection task is not eliminated by explicit negations. Thinking \& Reasoning, 14, 281-303.
Stanovich, K. E., \& West, R. F. (1998). Cognitive ability and variation in selection task performance. Thinking \& Reasoning, 4, 193-230.
Wagenmakers, E. J., Wetzels, R., Borsboom, D., \& van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). Journal of Personality and Social Psychology, 100, 3, 426-432
Wason, P. C. (1968). Reasoning about a rule. The Quarterly Journal of Experimental Psychology, 20, 273-281.
Wason, P. C., \& Johnson-Laird, P. N. (1969). Proving a disjunctive rule. Quarterly Journal of Experimental Psychology, 21, 14-20.
Wason, P. C., \& Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. Quarterly Journal of Experimental Psychology, 23, 63-71.

# Mental Algorithms in the Historical Emergence of Word Meanings 

Christian Ramiro (chrisram@berkeley.edu)<br>Cognitive Science Program<br>University of California, Berkeley<br>Barbara C. Malt (barbara.malt @lehigh.edu)<br>Department of Psychology<br>Lehigh University<br>Mahesh Srinivasan (srinivasan@berkeley.edu)<br>Department of Psychology, Cognitive Science Program<br>University of California, Berkeley<br>Yang Xu (yang xu_ch@berkeley.edu)<br>Department of Linguistics, Cognitive Science Program University of California, Berkeley


#### Abstract

Words frequently acquire new senses, but the mental process that underlies the historical emergence of these senses is often opaque. Many have suggested that word meanings develop in non-arbitrary ways, but no attempt has been made to formalize these proposals and test them against historical data at scale. We propose that word meaning extension should reflect a drive towards cognitive economy. We test this proposal by exploring a family of computational models that predict the evolution of word senses, evaluated against a large digitized lexicon that dates back 1000 years in English language history. Our findings suggest that word meanings not only extend in predictable ways, but also that they do so following an historical path that tends to minimize cognitive cost - through a process of nearestneighbor chaining. Our work contributes a formal approach to reverse-engineering mental algorithms of the human lexicon.


Keywords: Word meaning; semantic change; polysemy; chaining; nearest neighbor algorithm; lexicon

Over history, words have frequently acquired new senses, and become polysemous (Bréal, 1897). But the mental process that underlies the historical emergence of word senses is often opaque. Wittgenstein's notion of family resemblance (Wittgenstein, 1953, p31-32) highlights the challenge for researchers, showing that the many senses of the word game form "a complicated network of similarities overlapping and criss-crossing" with nothing identifiably in common (as for board games, card games, ball games, Olympic games, and so on). The network is presumably a reflection of the complex path the word game took in the historical development of its meaning. Decades of research have suggested possible ways that word meanings might be mentally structured or extended over time, but none has been tested formally against historical data at scale. We propose that word meanings should develop historically in ways that minimize cognitive effort, hence reflecting a drive towards cognitive economy (Zipf, 1949; Rosch, 1975). We test this proposal by formalizing previous theories in computational models that predict how word senses might emerge over time, contributing a principled approach to reverse-engineering mental algorithms of the human lexicon.

Our starting point is a set of influential ideas from cognitive science and linguistics suggesting that word meanings or categories might be structured in non-arbitrary ways. For example, pioneering work by Rosch (Rosch, 1975) showed that common semantic categories signified by words such as bird and furniture tend to exhibit a prototype structure, such that certain members of a category are more representative than others (e.g., robins and sparrows are more representative as birds than penguins or bats are). Although this theory has since been adapted to describe how word meanings might be structured (Lakoff, 1987) or extended over time (Geeraerts, 1997), it has not been computationally specified or evaluated broadly in accounting for historical patterns in how word senses emerge. A prominent alternative proposal is exemplar theory (e.g., Medin \& Schaffer, 1978; Nosofsky, 1986), which suggests that all encountered members of the category are stored and used in categorization judgments, although different members may be weighted differently. This proposal has also been used to describe how language might change over time, particularly concerning phonological and semantic representation (Bybee, 2006). To our knowledge, however, there has been no formal comparison of prototype and exemplar theories with respect to their ability to explain the historical emergence of word senses.

A critical addition to this theoretical terrain is the idea of chaining - popularized by Lakoff and other scholars (Lakoff, 1987; Malt, Sloman, Gennari, Shi, \& Wang, 1999) - as a possible mechanism that constrains word meaning extension. Chaining operates by linking an emerging idea (an incipient word sense) to a highly-related, already lexicalized word sense. When this process repeats over time, a chained structure in meaning space results. Recent work by Xu et al. (2016) has explored a preliminary version of this proposal via a nearest-neighbor model in a single semantic domain - household containers - but no systematic formalization or evaluation of chaining has been applied to explain the historical emergence of word senses more broadly. Further, al-
though chaining seems plausible as a mechanism, its theoretical value has been limited in two respects: 1) No work has formally specified why chaining might be a preferred mechanism for the development of word meanings; 2) No largescale assessment of chaining vs. alternative mechanisms has been performed against historical records of word sense extension, leaving open how chaining fares with respect to alternatives. These issues leave open the question of whether the evolution of word meanings follows a cognitively predictable path, and if so, what principles explain this process.

In the current work, we hypothesize that the emergence of word meanings should follow an historical path that minimizes collective cognitive effort. In particular, we propose that chaining should be a preferred algorithm for extending word meanings across history because it tends to minimize the cognitive cost of linking novel ideas with existing ones - a critical property not previously considered with regard to historical sense extension. To test the validity of this argument, we motivate nearest-neighbor chaining with tree-based computer algorithms that minimize edge lengths in a graph. We then formalize the process of chaining as a cognitively economical strategy for encoding novel ideas into an existing lexicon (cf. Xu, Malt, \& Srinivasan, 2016).

We critically assess our proposal by developing a family of computational algorithms of word meaning extension - inspired by the previous literature that described above - and evaluate them against a large historical database of wordmeaning records in English, spanning over 1,000 years. Our research extends a growing body of work which suggests that structures of language conform to efficient design principles (Zipf, 1949; Rosch, 1975; Piantadosi, Tily, \& Gibson, 2011; Kemp \& Regier, 2012; Kirby, Tamariz, Cornish, \& Smith, 2015), by bringing the perspective of cognitive economy to bear on the evolution of polysemy.

## Modeling the emergence of word meanings

## Computational formulation

We present here a formulation of five cognitive algorithms that might predict the historical emergence of word meanings, along with a null model. Given the initial, progenitor meaning of a word, each non-null algorithm postulates a distinct chaining mechanism by which novel word senses might emerge over time by "attaching to" existing meanings. Each algorithm generates a prediction of the historical order through which the meanings for any given word should emerge, which we then test against the historical record. In effect, we reverse-engineer the mental mechanisms of sense extension.

Table 1 summarizes the full set of proposed algorithms. Here $m$ stands for meaning or word sense, and $t$ stands for time. Each algorithm infers the word sense that emerges at time $t+1\left(m_{t+1}\right)$, based on existing senses of a word up to time $t\left(m_{1}, \ldots, m_{t}\right)$. The inferred sense is drawn from the candidate pool of senses (denoted by $m^{*}$ ) that appear after $t$ for a given word. A perfect model would fully recapitulate the

Table 1: Proposed models of word meaning extension.

| Name | Description |
| :--- | :--- |
| Random (null) | $m_{t+1} \sim \operatorname{random}$ draw $m^{*}$ |
| Exemplar | $\left.m_{t+1} \sim E_{m_{i}} \operatorname{sim}\left(m^{*}, m_{i}\right)\right]$ |
| Prototype | $m_{t+1} \sim \operatorname{sim}\left(m^{*}\right.$, prototype $\left.\left(m_{1}, \ldots, m_{t}\right)\right)$ |
| Progenitor | $m_{t+1} \sim \operatorname{sim}\left(m^{*}, m_{1}\right)$ |
| Local | $m_{t+1} \sim \operatorname{sim}\left(m^{*}, m_{t}\right)$ |
| Chaining | $m_{t+1} \sim \max _{i=1}^{t} \operatorname{sim}\left(m^{*}, m_{i}\right)$ |

historical emerging order of all senses of a word. All of our models are parameter-free and thus make minimal assumptions in the computational formulation.

1. The random algorithm - or null model - predicts the historical emergence of a word's senses to be random. This would only be plausible if word senses emerge purely based on immediate communicative needs with no further cognitive constraints.
2. The exemplar algorithm adapts from work by Nosofsky (1986), whereby the emerging sense at $t+1$ is predicted to be the one that bears the highest semantic similarity on average (or the highest sum of semantic similarities, which is equivalent in our case) with existing senses of a word at time $t$. We define semantic similarity identically in all algorithms, and we defer its formal definition to a later section.
3. The prototype algorithm is adapted from work by Rosch (1975) and Geeraerts (1997) and predicts the emerging sense at $t+1$ to be the one that bears the highest semantic similarity with the prototypical sense at time $t$. The prototype at $t$ is defined as the sense that bears the highest semantic similarity with existing senses of a word prototype $\left(m_{1}, \ldots, m_{t}\right) \leftarrow$ $\max _{i} \sum_{j \neq i} \operatorname{sim}\left(m_{j}, m_{i}\right)$. Thus, this algorithm allows the most representative sense of a word to change as a function of time, as more word senses develop.
4. The progenitor algorithm is a variant of the prototype model that assumes a fixed prototype that is always the initial, progenitor word sense (i.e., the earliest sense recorded in history). It predicts the emerging sense at $t+1$ to be the one that bears the highest semantic similarity (among all candidate senses) with respect to the progenitor sense.
5. The local algorithm assumes that word meanings emerge in a temporal linear chain, where the emerging sense at $t+1$ is the one that bears the highest semantic similarity with the sense that appears at time $t$. Critically, senses that appear prior to $t$ have no influence on the emerging sense at $t+1$ on this model. This algorithm posits that sense extension will yield minimal cost locally between consecutive time points, as opposed to yielding globally minimal cost (described below).
6. The chaining (or nearest-neighbor) algorithm is closely related to Prim's algorithm for constructing a minimal spanning tree (Prim, 1957) - but with a fixed (as opposed to random) starting point, i.e., it always begins with the progenitor sense of a word. In essence, this algorithm predicts the





Figure 1: Simulation of the proposed algorithms of word sense extension. The solid red circle symbolizes the progenitor sense of a word. The blue circles represent emerging word senses, and the arrows indicate the predicted path that each algorithm makes about order of emergence. The time labels indicate the predicted sequence of emergence. The cost is the aggregated Euclidean distances traversed by the arrows.
emerging sense at $t+1$ to be the one that bears the highest semantic similarity to any of the existing senses up to $t$, hence rendering a chain that connects nearest-neighboring senses over time. In contrast with the other algorithms described above, this chaining algorithm is also similar to single linkage clustering (Gower \& Ross, 1969) which tends to yield a tree (i.e., each tree node is a sense in this case) with minimal edge lengths among nodes of a graph (i.e., the graph is a network of senses of a word, developed in history). Due to this property, the chaining algorithm assumes the least cumulative historical cognitive effort for the extension of word senses (where effort is inverse to the degree of association between emerging and existing senses of a word), providing the computational implementation of our hypothesis.

## Simulation of sense extension algorithms

To illustrate how nearest-neighbor chaining would yield a near-minimal-cost historical path, we provide a simulation for the proposed algorithms of sense extension as follows.

We generated 15 randomly placed points in a twodimensional plane that represents the meaning space for a hypothetical word (see Figure 1). We took Euclidean distance between-points as a proxy for semantic distance (or inverse semantic similarity) between two senses. We also designated the bottom-right point in the space as the progenitor sense, i.e., it is the earliest seeding sense for the word that is a given to any algorithm. We then applied the family of sense extension algorithms to the remaining data points and visualized the path of emerging senses predicted by each algorithm. Figure 1 shows that these algorithms yield distinct typologies and paths in the simulated meaning space. Specifically, the exemplar algorithm links novels senses to all existing senses based on average distances between them (illustrated by chains that develop from spaces between senses as opposed to those that
stem off directly from senses). The prototype algorithm predicts a dynamic radial structure (Lakoff, 1987), where temporal chains are established by linking novel senses to prototype senses, while allowing the prototype to change over time. The progenitor algorithm predicts a strict radial structure where all senses stem from the earliest progenitor meaning. The local algorithm predicts a linear temporal chain of senses by attaching each emerging sense to the existing sense that appears one time point earlier. Finally, the chaining algorithm renders a tree structure that branches off as needed to preserve nearest-neighbor relations between emerging and existing senses. Importantly, the chaining algorithm yields the minimal aggregated edge lengths, hence rendering a minimal cost in semantic space. This result is robust to variations in simulation parameters and is a consequence of the close link between the nearest-neighbor chaining algorithm and the concept of a minimal spanning tree.

Below, we test the extent to which these algorithms can recapitulate the emergence of word senses, as recorded in a large historical lexicon of English.

## Treatment of data

## Historical lexicon

To evaluate our proposed algorithms, we used the Historical Thesaurus of English (HTE) (Kay, Roberts, Samuels, Wotherspoon, \& Alexander, 2015) - a unique large-scale historical lexicon constructed from the Oxford English Dictionary. This database includes approximately 800,000 word forms and their senses, dated and recorded over a span of over 1,000 years - from Old English to the present day. Each word sense in the HTE is annotated with the date of its emergence (and, where applicable, obsolescence) and part of speech, and is structured in a fine-grained semantic hierarchy that features about a quarter of a million concepts. Consecutive tiers of the hierarchy typically follow a IsA or Partof relation. For example, one sense of the word game under the HTE code "01.07.04.04" is defined in terms of a four-tier hierarchy: The world (01) $\rightarrow$ Food and drink (01.07) $\rightarrow$ Hunting (01.07.04) $\rightarrow$ Thing hunted/game (01.07.04.04).

## Measure and validation of semantic similarity

To quantify similarity between word senses, we defined a measure using the semantic hierarchy in the HTE and then validated it against human judgments. Specifically, we approximated psychological similarity between a pair of word senses $\operatorname{sim}\left(m_{i}, m_{j}\right)$ by a common measure of similarity used in psychology that is bounded in the range of $(0,1)$ (Nosofsky, 1986; Shepard, 1987):

$$
\begin{equation*}
\operatorname{sim}\left(m_{i}, m_{j}\right)=e^{-d\left(m_{i}, m_{j}\right)} \tag{1}
\end{equation*}
$$

Here $d\left(m_{i}, m_{j}\right)$ represents thesaurus-based conceptual distance between two meanings, which we defined by the inverse of a conceptual similarity measure $(s(\cdot, \cdot))$, commonly used in
natural language processing (Wu \& Palmer, 1994; Jurafsky \& Martin, 2009):

$$
\begin{equation*}
d\left(m_{i}, m_{j}\right)=1-s\left(m_{i}, m_{j}\right)=1-\frac{2 \times|p|}{l\left(m_{i}\right)+l\left(m_{j}\right)} . \tag{2}
\end{equation*}
$$

Here $|p|$ is the number of parent tiers shared by senses $m_{i}$ and $m_{j}$, and $l(\cdot)$ is the depth of a meaning in the semantic hierarchy. This measure gives 1 if two meanings are identical, and 0 if they have nothing in common. Table 2 illustrates the calculation of this measure with a concrete example.

Table 2: Illustration of conceptual similarity based on two senses of game recorded in the HTE. Since the two senses share two parent tiers (i.e., The social world $\rightarrow$ Leisure) in the hierarchy, the conceptual similarity is $s(\bullet, \star)=\frac{2 \times 2}{5+6}=\frac{4}{11}$.


We validated this measure of semantic similarity via standard techniques in natural language processing, by evaluating its performance in predicting human judgments of word similarities. Following Resnik (1995), we approximated word similarity by using the pair of senses for the two words that results in maximum sense similarity, defined as follows: $\operatorname{wordsim}\left(w_{i}, w_{j}\right)=\max _{m_{i} \in \operatorname{senses}\left(w_{i}\right), m_{j} \in \operatorname{senses}\left(w_{j}\right)} s\left(m_{i}, m_{j}\right)$.

Our measure of semantic similarity yielded a Spearman's correlation of 0.43 ( $p<0.001$ ) on Lex-999 (Hill, Reichart, \& Korhonen, 2015), which provides a well-known data set of human word similarity judgments. The performance of our measure of semantic similarity is better than the corpusbased skipgram (Word2Vec) model, which has been trained on 1 billion words of Wikipedia text (Mikolov, Chen, Corrado, \& Dean, 2013) and roughly on par with the same model trained on 300 billion words (Faruqui \& Dyer, 2015). In addition, our measure of semantic similarity obtained a Spearman's correlation of $0.52(p<.001)$ on Sim-353 (Finkelstein et al., 2001), another common data set of human word relatedness judgments, which is comparable to the state-of-the-art

GLOVE word vector model, which has been trained on 6 billion words (Faruqui \& Dyer, 2015; Pennington, Socher, \& Manning, 2014).

Having validated our measure of semantic similarity, we used it to assess the mental algorithms described above.

## Choices of words

We focused our analyses on explaining word sense extension in a set of the most common English words. Specifically, we worked with the most frequent 6318 words in the British National Corpus (BNC). Some of the word forms are duplicated in this set because one word can function in multiple part-of-speech categories. However, our results were robust regardless of whether we collapsed these words by form or distinguished them by part-of-speech.

## Model evaluation and results

We used model likelihood to assess the performance of each proposed algorithm. ${ }^{1}$ We defined likelihood as a probability function that specifies the degree to which a model accounts for the entire sequence of senses that historically emerged for a given word. To be concrete, for a sequencce of senses $m_{1}, m_{2}, m_{3}, \ldots, m_{t}$, the likelihood $\mathcal{L}$ is the joint probability of observing such a sequence under a certain model $\mathcal{M}$ :

$$
\begin{equation*}
\mathcal{L}_{\mathcal{M}}=p\left(m_{1}\right) p\left(m_{2} \mid m_{1}\right) p\left(m_{3} \mid m_{1}, m_{2}\right) \ldots p\left(m_{t} \mid m_{1}, \ldots, m_{t-1}\right) \tag{3}
\end{equation*}
$$

We assumed that the progenitor sense is always given, so $p\left(m_{1}\right)=1$. For all remaining emerging senses, the set of algorithms can be evaluated by calculating likelihood based on their specifications in Table 1. For example, the progenitor model would yield a likelihood for the emerging sense at $t=2$ (conditioned on that appeared at $t=1$ ) as follows:

$$
\begin{equation*}
p\left(m_{2} \mid m_{1}\right)=\frac{\operatorname{sim}\left(m^{*}, m_{1}\right)}{\sum_{m^{*} \in\left\{m_{2}, \ldots, m_{t}\right\}} \operatorname{sim}\left(m^{*}, m_{1}\right)} . \tag{4}
\end{equation*}
$$

The algorithm then steps through each point in time and the likelihood correspondingly calculates the degree to which the algorithm predicts the true emerging sense at that point, among a candidate pool of senses that appear after.

Because our null hypothesis is that there exists no predictability in how word senses develop in history, we evaluated each cognitive algorithm against the random null algorithm, using the log likelihood ratio $(L L R)$ - a standard metric for model comparison in statistics: $L L R=\log \left(\mathcal{L}_{\mathcal{M}} / \mathcal{L}_{\text {null }}\right)$. This quantity should be greater than 0 if a given model accounts for word sense extension better than the null, and the converse if the null does better. For any given word, the likelihood function of the null can be determined theoretically, and it is simply the inverse of factorial of $N-1$ for a word with $N$ senses: $\mathcal{L}_{\text {null }}=1 \times \frac{1}{N-1} \times \frac{1}{N-2} \times \ldots \times \frac{1}{1}=\frac{1}{(N-1)!}$. Thus the log likelihood ratio indicates whether a model predicts the sequence of emerging word senses better than chance.

[^179]

Figure 2: Summary of model performances against the null. " 0.0 " on the y -axis indicates performance of the null model. Bar height indicates the mean log likelihood ratio averaged over the entire pool of most common words from the BNC corpus. Error bars indicate $95 \%$ confidence intervals.

Figure 2 summarizes the results. The bar plot shows that each of the proposed algorithms accounts for the historical data that we examined significantly better chance ( $p<0.001$ from 1-tailed $t$-tests), reflected in the positive log likelihood ratios. This observation suggests that the null hypothesis can be rejected: The emerging order of word senses in the English lexicon is not purely random.

Critically, the nearest-neighbor chaining algorithm yielded the highest overall likelihood among all models, and this result was statistically significant according to paired $t$-tests between the chaining model and each of the remaining models ( $p<0.001$ in all four comparisons). This observation provides evidence that word senses emerge in cognitively efficient ways by approximating a minimal spanning tree over the course of history. As such, these data support our hypothesis about nearest-neighbor chaining as the dominant mental algorithm for the historical emergence of word senses.

To illustrate the nearest-neighbor chaining process, we visualized the predicted chaining path for the English word game. Figure 3 shows a low-dimensional projection (via multi-dimensional scaling with a random starting point) of all senses of game as a noun, taken from the HTE database. As can be seen, the chaining algorithm forms a minimal spanning tree among the senses of game, by linking neighboring nodes that are semantically close. Importantly, this process of meaning extension tends to support branching and the formation of local clusters, identified roughly in this case by the three sense groups of "hunting game" (upper-left cluster), "scheme" (middle cluster), and "sports and entertainment" (upper-right cluster) in Figure 3. This offers a computational basis for family resemblance (Wittgenstein, 1953) and polysemy, by allowing words to develop both related and distinct
senses over time.

## Conclusions

We presented the first large-scale computational investigation of the mental algorithms that determine how words evolve new senses over time. We found that the historical emergence of word senses in English is not arbitrary; Instead, it has exhibited a high degree of predictability over the past millennium. Our findings indicate that the order in which word senses emerge can be best accounted for by a process of nearest-neighbor chaining, which supports the view that the historical development of the lexicon follows a trajectory that tends to minimize cognitive effort. Our current analysis focuses on sense extension within individual word forms, but it would be useful to extend our analysis to examine how different words compete to express novel meanings. Our exploration of the mental algorithms that underlie historical sense extension opens new, interdisciplinary venues for reverse engineering the evolution of the human lexicon.

## Acknowledgments

We thank Marc Alexander for helping with data licensing, and the University of Glasgow and the Oxford University Press for making the Historical Lexicon of English available. We thank Charles Kemp and Terry Regier for discussions on chaining and the Computational and Experimental Methods group in Linguistics at UC Berkeley for constructive comments. This project was partly funded by NSF award SBE-1041707 to the Spatial Intelligence and Learning Center (SILC) and NSF award SBE-16302040 to MS.

## References

Bréal, M. (1897). Essai de sémantique: Science des significations. Paris: Hachette.
The British National Corpus, version 3 (BNC XML Edition). (2007). (Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/)
Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. Language, 82(4), 711-733.
Faruqui, M., \& Dyer, C. (2015). Non-distributional word vector representations. arXiv preprint arXiv:1506.05230.
Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., \& Ruppin, E. (2001). Placing search in context: The concept revisited. In Proceedings of the 10th international conference on world wide web (pp. 406414).

Geeraerts, D. (1997). Diachronic prototype semantics: A contribution to historical lexicology. Oxford: Oxford University Press.
Gower, J. C., \& Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. Applied Statistics, 18(1), 54-64.
Hill, F., Reichart, R., \& Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics.


Figure 3: Historical chaining in the English word game. The two-dimensional space is generated by multi-dimensional scaling based on sense similarities. The solid red circle marks the earliest meaning. The arrows indicate the predicted path from the chaining algorithm. The annotations include a gloss for the sense and its recorded period of emergence in the HTE.

Jurafsky, D., \& Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd ed.). New Jersey: Pearson Education.
Kay, C., Roberts, J., Samuels, M., Wotherspoon, I., \& Alexander, M. (2015). The historical thesaurus of english, version 4.2. Glasgow: University of Glasgow.
Kemp, C., \& Regier, T. (2012). Kinship categories across languages reflect general communicative principles. Science, 336(6084), 1049-1054.
Kirby, S., Tamariz, M., Cornish, H., \& Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. Cognition, 141, 87-102.
Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press.
Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., \& Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. Journal of Memory and Language, 40(2), 230-262.
Medin, D. L., \& Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85(3), 207.
Mikolov, T., Chen, K., Corrado, G., \& Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39.
Pennington, J., Socher, R., \& Manning, C. D. (2014). Glove: Global vectors for word representation. In Emnlp (Vol. 14, pp. 1532-1543).
Piantadosi, S. T., Tily, H., \& Gibson, E. (2011). Word lengths
are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108(9), 3526-3529.
Prim, R. C. (1957). Shortest connection networks and some generalizations. Bell Labs Technical Journal, 36(6), 13891401.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmplg/9511007.
Rosch, E. H. (1975). Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104(3), 192.
Shepard, R. (1987). Toward a universal law of generalization for psychological science. Science, 237(4820), 13171323.

Wittgenstein, L. (1953). Philosophical investigations. Oxford: Basil Blackwell.
Wu, Z., \& Palmer, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on association for computational linguistics (pp. 133-138).
Xu, Y., Malt, B. C., \& Srinivasan, M. (2016). Evolution of polysemous word senses from metaphorical mappings. In Proceedings of the 38th annual meeting of the cognitive science society.
Xu, Y., Regier, T., \& Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. Cognitive Science, 40, 2081-2094.
Zipf, G. K. (1949). Human behavior and the principle of least effort. Boston: Addison-Wesley.

# A cognitive analysis of deception without lying 

Keith Ransom (keith.ransom@adelaide.edu.au)<br>School of Psychology, University of Adelaide

# Wouter Voorspoels (wouter.voorspoels@ kuleuven.be) 

Brain \& Cognition, University of Leuven

Amy Perfors (amy.perfors@adelaide.edu.au)<br>School of Psychology, University of Adelaide

Daniel J. Navarro (dan.navarro@unsw.edu.au)<br>School of Psychology, University of New South Wales


#### Abstract

When the interests of interlocutors are not aligned, either party may wish to avoid truthful disclosure. A sender wishing to conceal the truth from a receiver may lie by providing false information, mislead by actively encouraging the receiver to reach a false conclusion, or simply be uninformative by providing little or no relevant information. Lying entails moral and other hazards, such as detection and its consequences, and is thus often avoided. We focus here on the latter two strategies, arguably more pernicious and prevalent, but not without their own drawbacks. We argue and show in two studies that when choosing between these options, senders consider the level of suspicion likely to be exercised on the part of the receiver and how much truth must be revealed in order to mislead. Extending Bayesian models of cooperative communication to include higher level inference regarding the helpfulness of the sender leads to insight into the strategies employed in non-cooperative contexts.


Keywords: deception; Inductive inference; communication; pragmatics

## Introduction

"You can tell he's lying because his lips are moving."
If only detecting lies were that simple! Despite its importance, people generally perform at chance when classifying liars and truth tellers (C. Bond \& DePaulo, 2006). Indeed, most verbal and nonverbal cues have only marginal diagnostic value (DePaulo et al., 2003). Instead of focusing on cues to deception, a promising new approach considers how the cognitive processes involved in deception may differ from telling the truth. It has been suggested, for example, that deception imposes higher cognitive demands on liars, who may find it more difficult to furnish details when interviewed (Vrij \& Granhag, 2012). A good understanding of the cognitive mechanisms underlying deception, taking into account the complexities of the strategies employed, would be a tremendous asset (G. D. Bond, 2012; Blandon-Gitlin, Fenn, Masip, \& Yoo, 2014).

In the present research we analyse the challenge faced by would-be deceivers seeking to conceal the truth. We begin with a brief analysis of the deceiver's perspective, identifying the main deceptive strategies, and outline a preliminary study which illustrates people's preferences for these strategies. We then present two studies where we ask people to conceal the
truth, manipulating the level of suspicion of the hypothetical receiver and the information content of the available message options.

## How to deceive in ten steps (with pictures)

Communication relies on principles of cooperation (Grice, 1989). The intended meaning of a sender rarely coincides perfectly with the "literal" content of a message, but by making assumptions about why the sender chose that particular message, the receiver may infer the intended meaning. By assuming that a sender is cooperative and produces messages that follow the Gricean maxims (described below), a receiver can increase the speed and strength of the inferences they draw from those messages (Horn, 2004).

But what if the sender is not trying to be cooperative? In that case, there are three main strategies the sender can rely on, each corresponding to different violations of the Gricean maxims. Consider the following scenario: You shot your neighbour's hamster with your shotgun while she was away for the weekend. Obviously, you'd prefer that she didn't learn the truth. However, you were given the key to her house to take care of her pet, so you are definitely a person of interest in her investigation. How can you conceal the truth from her?

- Lying: You might try an outright lie: "I did not shoot your hamster". Lying involves communicating a proposition to the receiver with the full knowledge that it is false. From a Gricean perspective, lying is a violation of the supermaxim of quality, stating that your contribution should be true.
- Being uninformative: It seems very sensible to be uninformative. Neighbour: "Did you shoot my hamster?" You:"Have you heard the new Justin Bieber album? It's fantastic!" ${ }^{1}$ With this kind of utterance, it would seem that the receiver can infer nothing beyond her prior beliefs. But this violates the Gricean maxims of relevance and quantity, and these violations can themselves be informative about the sender's intentions even if not the actual facts of the matter.
- Misleading: A third option is to mislead your neighbour by implicature. "I was not at home when your hamster was shot!" You tell her a truth very relevant to the issue at hand, but from which you believe a false conclusion will be drawn (you were not at home when you shot the hamster; you took it with you to a nearby park for target practice). Misleading involves covertly violating the maxims of quantity, but may be harder to detect.

[^180]In a preliminary study, we asked 96 first year psychology students ( 87 women) at the University of Leuven to imagine seven different scenarios like the following:

> A man arrives home after a weekend in Vegas, during which he won $\$ 2000$ playing Poker, but lost $\$ 500$ at the slot machines and $\$ 4000$ on Blackjack. When he returns, he does not know how to tell his wife. His wife knows that he has had gambling problems in the past, but is convinced that they are resolved. Their relationship is currently strained, so the man would rather not cause any additional problems. His wife asks him directly if he has gambled. Which of these answers would you give if you were in his situation?

Participants selected a response from seven options comprised of two lies (e.g. "I didn't gamble"), two uninformative statements (e.g. "there were a lot of people gambling"), two misleading statements (e.g. "I won $\$ 2000$ playing Poker"), and the truth. Figure 1 presents the preference of the participants (collapsed across scenarios and equivalent response options).


Figure 1: When choosing how to communicate in a variety of different scenarios with a clear motivation to deceive, participants showed a strong tendency to mislead rather than be uninformative.

Two important conclusions emerge from Figure 1. Firstly, people were uncomfortable with deception: $37 \%$ of responses involved telling the full truth and only $10 \%$ were outright lies: a surprising number perhaps given that each scenario provided a clear motivation to deceive. However, this finding is consistent with previous research showing that in general people avoid lying through concerns regarding self-image, guilt, and anxiety (Aquino \& Reed, 2002). Of more relevance to our present purposes, we found that among those who chose not to tell the truth, people showed a clear preference for misleading over lying or being uninformative (37\%, 10\% and $15 \%$ respectively). This finding is consistent with earlier work on the topic (Montague, Navarro, Perfors, Warner, \& Shafto, 2011; Rogers, Zeckhauser, Gino, Schweitzer, \& Norton, 2014).

## Balancing suspicion and information

So, why do people seem to prefer to actively mislead rather than be entirely uninformative? At first glance, it seems rational to be as uninformative as possible: the receiver cannot revise her beliefs on the basis of your utterances. Misleading on the other hand, involves salting your statements with a grain of truth - something which the receiver may build upon to infer the whole truth.

An important motivation for choosing a misleading utterance over a strictly uninformative one is because the latter
raises suspicion. Consider the likely response of choosing to be uninformative, as in the Las Vegas scenario:

```
Spouse: Did you gamble?
Gambler: Where shall we go for dinner? I'm hungry.
Spouse: You lost money didn't you?
Gambler: Some of the guys won big.
Spouse: How much did you lose?
```

As Sperber et al. (2010) points out, people have a toolbox of cognitive mechanisms for epistemic vigilance that reduce the risk of being deceived. One such tool supports tracking the cooperation of others; as a result, obvious departures from that cooperation are noted (e.g., Mills, 2013). Responding in an uninformative way violates the principle of cooperation so blatantly that the deception is revealed.

A deceiver, sensitive to the epistemic vigilance of his counterpart may prefer instead to provide truthful but misleading utterances, but in doing so faces a delicate trade-off. Chosen well, such utterances may not only allay the receiver's suspicion, but by virtue of the inferential boost accorded to cooperative speakers, the receiver may be led to a false conclusion, terminating the search for further information. Yet if suspicion is already raised, the receiver is unlikely to fall for the false implicature, which relies on her assumption of cooperation (Dynel, 2011), and may use the information to get closer to the truth.

This analysis points to two opposite forces, balanced in the selection of one strategy over another. On one hand, the knowledge that the receiver may engage in inference about the helpfulness of the statement may lead the sender to opt for a misleading yet informative statement. On the other hand, if the sender considers that the receiver will be suspicious a priori, he may resort to being uninformative. We examine these factors in two experiments.

## Experiment 1: The deception game

As the basis for our empirical investigation, we use a simple two person communication game in which the interests of the sender and receiver are opposed. In the game, the sender (cast in the role of a pirate) and the receiver (cast as an explorer) see four alternative maps, only one of which is genuine. Each map purports to show the true extent of a contiguous region where treasure is buried. The pirate, who knows the identity of the genuine map and must seek to protect it, is required to provide a hint to the explorer in the form of a small number of locations that lie within the region. The explorer must use this information to guess the identity of the genuine map. Both players know that lying is not allowed; so the pirate can only reveal locations where treasure is actually buried.

An example trial faced by participants, who took the role of the pirate in this experiment, is shown in Figure 2. Figure 2(a) illustrates the four maps shown to the pirate and (hypothetical) explorer for an example trial. In providing a hint to the explorer, participants were restricted in their choice to one of three kinds of hints : one uninformative, one mislead-

(a) Four maps provided to the pirate and explorer.

(b) The pirate chooses one of three hints.

Figure 2: Experiment 1: The deception game.(a) Both the pirate and explorer see the same four maps; the shaded area marks the region where treasure be buried. Only one of the four maps is genuine: the pirate seeks to conceal its identity, the explorer seeks to discover it. (b) The pirate must provide a hint to the explorer, and is given three hints from which to choose: a MisLEADING, Uninformative, or Helpful hint. Each hint marks a subset of locations where treasure is buried (blue dots). Since lying is not permitted all marked points lie within the shaded area when overlaid on the genuine map. The hints vary in their potential to drive inference. The MisLEading hint appears to clearly point out the wrong map, despite being consistent with all of them. The Uninformative hint neither points to nor excludes any of the maps. The Helpful hint excludes all of the maps except the genuine one.
ing, and one genuinely helpful. Figure 2(b) shows examples of the three options, and illustrates the likely effect that each hint would have on the inference of a trusting explorer.

The informativeness of a hint was manipulated by varying the number of treasure maps (out of four) that were excluded by the hint. When maximally informative (Helpful), a hint excludes all treasure maps, except the true one (bottom row of Figure 2b). When maximally Uninformative, it excludes no treasure maps (middle row). The Misleading hints were designed to closely resemble one of the three false maps (top row), and to exclude none (as in the example shown), one or two of the treasure maps.

People's beliefs about the suspicion level of the receiver (the hypothetical explorer) were manipulated by changing the proportion of deceivers in the population that were supposedly providing information. Participants were told that they were part of a crew of six sailors providing a hint to the explorer. Participants were also told that the explorer knew that the hint came from an unknown (but randomly selected) crew member and knew the proportion of deceptive crew members. Varying the number of deceivers in the crew from one to five was intended to raise the perceived suspicion level of the explorer. Our question was whether the pirate would track and use this information when providing hints, being more likely to mislead when the suspicion level was lower. When facing
a trusting receiver, we expect people to be more inclined to mislead, provided that the amount of information disclosed is acceptable. But if the receiver is likely to be suspicious or too much information would otherwise be revealed, we expect people to be uninformative.

Participants were 120 undergraduates from the University of Leuven ( $86 \%$ female, ages 18-24, median 18) participating for course credit. Participants faced 30 trials in all: six sets of four maps were presented in conjunction with each of five crew configurations.

## Results and discussion

Our first question was whether people's decision to mislead or not depended on how much information the misleading option gave away. To address this, we examined whether the proportion of participants choosing each kind of hint depended on the number of treasure maps that were excluded by the Misleading one. A chi-square test confirmed the dependency shown in Figure $3\left(\chi^{2}(4,3600)=93.31, p<\right.$ 0.001 ): as the amount of information revealed by the MisLEADING hint increased, people were less willing to select it, preferring instead to choose the Uninformative option. Indeed, in contrast to the pilot data in Figure 1, the UninFORMATIVE option was the most favoured in this task. This is unsurprising perhaps, since a hint that excludes even one of four maps is extremely informative.


Figure 3: Experiment 1: Informativeness. The $y$-axis shows the number of times each type of hint (given by column label) was selected as a function of the number of treasure maps excluded by the misleading hint ( $x$-axis): $0 \%$ means none were excluded, $25 \%$ means one was, and $50 \%$ means that two were. People were sensitive to informativeness: when the misleading hint was more informative, people were less likely to mislead.

Our second question was whether people's decision was affected by their estimate of how suspicious the receiver (the hypothetical explorer) was likely to be. In light of this, we examined the relationship between participants' choice of hints and the number of deceptive crew members providing hints for the explorer. Curiously, as Figure 4 illustrates, there was no evidence for a relation between the level of suspicion and the type of hint selected $\left(\chi^{2}(8,3600)=12.96, p=0.11\right)$. Even when people knew that the explorer thought that five out of the six possible senders was acting deceptively, they did not alter their selection of hints.


Figure 4: Experiment 1: Suspiciousness. The $y$-axis shows the number of times each type of hint (given by column label) was selected as a function of the implied suspicion level of the explorer. The numbers on the $x$-axis represent the number of deceptive members in the crew that provides the explorer with a hint (more members suggests that the explorer should be more suspicious). There was no dependency between choice of hint and number of deceptive crew members.

## Experiment 2: Increasing suspicion

Experiment 1 found convincing evidence for the influence of informativeness on the decision to mislead. If too much information would be revealed with a true but misleading statement, people are more inclined to be uninformative. Surprisingly, we did not find an effect of suspicion. What might be going on here?

One possible explanation is simply that our manipulation was ineffective. Perhaps changing the number of deceptive crew members was not salient enough or required too much effort for participants to interpret or keep in mind.

Experiment 2 was identical to Experiment 1, but rather than have people infer how suspicious the explorer might be from the composition of the crew, we instead gave participants explicit information about the explorer's beliefs. In the Low SUSPICION condition, participants were told that the explorer suspected that the hint came from a teammate, whereas in the High Suspicion the explorer suspected the hint came from an opponent. The experiment was similar in all other respects except for a control condition in which participants were asked to help the explorer (used to identify participants who were not trying or did not understand the task). There were also a number of filler items in which there was no obviously misleading option.

Participants were 98 adults recruited via Amazon Mechanical Turk and paid $\$ 1.25$ USD for 15 minutes participation. Data from 22 participants who failed to demonstrate a sufficient understanding of the experiment were excluded from subsequent analysis. ${ }^{2}$ The remaining 76 participants were $46 \%$ female and aged 20-63 (median age 28.5). Participants faced 30 trials in all: 10 map sets (six experimental, four fillers) were presented in each of the three condition blocks.

[^181]

Figure 5: Experiment 2: Informativeness. The $y$-axis shows the number of times each type of hint (given by column label) was selected as a function of the number of treasure maps excluded by the misleading hint ( $x$-axis). As in Experiment 1, people were sensitive to informativeness: when the misleading hint was more informative, people were less likely to mislead. Data from the control condition are excluded.


Figure 6: Experiment 2: Suspiciousness. The $y$-axis shows the number of times each type of hint (given by the column label) was selected as a function of the implied suspicion level of the explorer. The $x$-axis reflects whether people were told that the explorer was expecting a hint from a member of another team (the HIGH SUSPICION condition) or from a teammate (the LOW SUSPICION condition). When participants knew that the explorer was apt to be suspicious of them, they were less inclined to be misleading, opting instead to be uninformative.

## Results and discussion

As before, our first question was whether people were sensitive to informativeness when choosing which hint to provide. Once again, there was a significant effect of informativeness $\left(\chi^{2}(4,912)=18.04, p=0.001\right)$. As Figure 5 shows, the more maps the misleading option excluded, the less inclined people were to select it, favouring instead the uninformative option.

In light of the null effect in Experiment 1, a perhaps more interesting question is whether people were sensitive to the suspicion level of the explorer when deciding what to tell them. As Figure 6 reflects, when the suspicion level of the explorer is made more obvious, people are indeed sensitive to it. Although the Uninformative hint was still the most popular overall, the MisLEADING option was chosen far more when the explorer was expecting a hint from a trusted teammate $\left(\chi^{2}(2,912)=85.95, p<0.001\right)$. This suggests that people acting as senders are indeed attentive to the level of trust presumed by the receiver; although, taken together with the results from Experiment 1, tracking suspicion may be too cognitively challenging where it is not especially salient.

## Towards a computational model

Experiments 1 and 2 manipulated two important factors: the information content of the messages that deceivers could
choose and their beliefs about the degree of suspicion with which their messages would be received. Taken together, our results show that both factors were important considerations. In this section we present a computational model whose goal is to aid our understanding of these results and generate new testable predictions. While an in-depth analysis of the model is beyond the scope of this paper, here we briefly describe the relevant features.

A convenient starting point for a model of the deception game employed here - and for communication in general is rational inference (e.g., Goodman \& Frank, 2016). In this framework, a receiver faces the challenge of updating her beliefs on the basis of information disclosed by a sender. The sender, for his part, selects information designed (according to his goal) to help or hinder the receiver in her efforts.

We first evaluate things from the perspective of the receiver, who is confronted with a hint $x$ (or, more generally, an utterance). The receiver is assumed to update her beliefs $h$ according to:

$$
\begin{equation*}
P_{\text {Receiver }}(h \mid x) \propto \sum_{s \in \mathcal{S}} P_{\text {Sender }}(x \mid h, s) P(h) P(s) \tag{1}
\end{equation*}
$$

where $s$ represents a sampling strategy employed by the sender and $S$ represents the range of such strategies considered. As a simplifying assumption, we assume that the receiver considers the sender's sampling strategy to be independent of the true hypothesis.

This inference thus depends on the sender, who selects information according to a sampling strategy:

$$
\begin{equation*}
P_{\text {SENDER }}(x \mid h) \propto\left(P_{\text {RECEIVER }}(h \mid x)\right)^{\alpha} \tag{2}
\end{equation*}
$$

where $\alpha$ reflects the goals of the sender, and $P_{\text {Receiver }}(h \mid x)$ the sender's assumptions about how the receiver updates her beliefs. A sender who wishes to reveal the truth to the receiver (i.e., to increase the receiver's posterior probability for the correct hypothesis $h$ ) will have an $\alpha$ with a positive value; one who wishes to conceal the truth has a negative $\alpha$; one who behaves somewhat randomly has an $\alpha=0$. There are other ways to capture conflicting goals, like assigning separate utility functions for the sender and receiver with regard to truth-predicated action, but we chose this for its relative simplicity.

To capture the patterns observed in our deception game, both equations have to be considered simultaneously. That is, both sender and receiver must recursively consider the assumptions and strategies used by the other party. Importantly, from the receiver's perspective the inferential potential of a message depends not only on the information as such, but also on the "sampling strategy" of the sender, which reflects the sender's goals and assumptions about the receiver. For example, sampling procedures that follow the principle of cooperation and the Gricean maxims have a stronger inferential potential (e.g., Bergstrom, Moehlmann, \& Boyer, 2006; Shafto, Goodman, \& Griffiths, 2014; Voorspoels, Navarro, Perfors, Ransom, \& Storms, 2015). Crucially, the receiver


Figure 7: Model predictions for sender actions. Model predictions for the preference of a sender in the deception game for the misleading, uninformative and helpful message options. From left to right, the panels present scenarios with increasingly suspicious receivers (modelled through different kinds of inference about the sender's goals and assumptions). The $x$-axes indicate how informative the misleading option is (in terms of the proportion of hypotheses excluded by it). The model predicts a decrease in preference for the misleading option as it becomes increasingly informative, as well as an increase in preference for misleading when the receiver is less suspicious.
not only updates her beliefs about what is true, but simultaneously makes inferences about the sender's sampling strategy: learning whether the sender is helpful and knowledgeable play a critical role in epistemic vigilance, and has a substantial impact on how rational agents reason (e.g., Shafto, Eaves, Navarro, \& Perfors, 2012).

In Equations (1) and (2), the universe of sampling strategies $S$ evaluated by the receiver is defined in terms of two things that she presumes about the sender: what does the sender assume about her (reflected by $P_{\text {Receiver }}(h \mid x)$ ), and what are his goals (reflected in $\alpha$ ). Many scenarios may be modelled in this way, but here we consider three. If the receiver is TRUSTING, this means that she is performing inference over two possibilities: either the sender is trying to be helpful ( $\alpha=1$ ), or he is inattentive and thus not selecting information with care $(\alpha=0)$. If the receiver is UNTRUSTING, this means that she believes that the sender is trying to conceal the truth from her $(\alpha=-1)$, under the mistaken assumption that he is trusted. Lastly, if the receiver is SUSPICIOUS, this means that she is performing inference over whether to be trusting or untrusting.

How well does this approach capture the main qualitative patterns in the deception game? To answer this, we simulate outcomes for the three scenarios we have outlined. In the leftmost panel of Figure 7, the receiver trusts the sender, but is not sure how attentive he is: he may be acting helpfully ( $\alpha=1$ ) or he may be providing poor but not actively misleading data, perhaps due to lack of motivation, attention, or information ( $\alpha=0$ ).

If the receiver updates her beliefs (concerning the true treasure map) at the same time as her assumptions about the helpfulness of the information received ( $\alpha$ ) then there is reason for the sender to choose a hint that seems informative. That
the message appears informative supports the receiver's assumption that it has been carefully selected, which further fuels inference. This recursive process may lead the receiver to draw a misleading conclusion if the sender is not actually helpful (as in our experiments). However, as the information content of the hint increases, so too does the risk that the receiver will inadvertently arrive at the truth. Consequently, the model captures the fact that the sender's preference for the misleading hint declines with its information content.

In the rightmost panel of Figure 7, in contrast, the receiver is certain that the sender is not to be trusted. If the sender is aware of this, there is little to be gained by attempting to mislead, and so the uninformative hint is preferred. The extent of this preference is, once again, moderated by the information content of the misleading option.

In many situations, a receiver will not be predisposed to regard the sender with complete trust, nor complete distrust, but rather will remain open to either possibility. We model this case by assuming that the receiver is performing inference about whether the sender should be trusted ( $\alpha=1$ or 0 ) or not $(\alpha=-1)$. The preferences of an antagonistic sender facing a suspicious receiver are shown in the center panel of Figure 7. The two conflicting forces are most pronounced here, dividing the sender's preference between the two strategies. On the one hand, the sender may convince the receiver that he is actually trying to help by appearing informative, yet the (real) information can be used by the receiver to rule out previously plausible (but false) hypotheses.

Overall, there are two clear patterns that were found in our experiments and were also predicted by our model. Firstly, as the information content of the misleading option increases, there is an increasing preference for choosing the uninformative hint. Secondly, the more trusting the receiver is assumed to be, the more popular the misleading option becomes. This pattern of results is consistent with the idea that people may be performing some kind of recursive inference over how suspicious their interlocutor is when deciding how to deceive. Furthermore, our results are consistent with the notion posited here and elsewhere (e.g., Goodman \& Frank, 2016; Shafto et al., 2012), that receivers (from the sender's perspective at least) perform joint inference over the goals of the sender and the truth of the matter at hand given the information received.

## Conclusion

"There is nothing more deceptive than an obvious fact." - Arthur Conan Doyle

In two studies we have demonstrated that people's preference for a deceptive strategy hinges on their assumption of whether cooperative norms are expected to apply. In situations where high levels of trust and cooperation are warranted, deceivers are more inclined to actively mislead than to simply withhold information. In this scenario, the deceiver seeks to leverage the inferential boost of cooperative communication. In contrast, when the deceiver believes the false implication will not
be inferred - when the receiver is already suspicious - then preference shifts towards limiting the information disclosed.

## Acknowledgments

We thank Paulien Van Rijmenant for help with data collection. KR was supported by an Australian Government Research Training Program Scholarship. WV is a postdoc at KU Leuven. AP was supported by ARC grants DP110104949 and DP150103280. DJN received salary support from ARC grant FT110100431.

## References

Aquino, K., \& Reed, . I. (2002). The self-importance of moral identity. Journal of Personality and Social Psychology, 83(6), 1423-1440.
Bergstrom, B., Moehlmann, B., \& Boyer, P. (2006). Extending the testimony problem: evaluating the truth, scope and source of cultural information. Child Development, 77(3), 531-538.
Blandon-Gitlin, I., Fenn, E., Masip, J., \& Yoo, A. (2014). Cognitiveload approaches to detect deception: searching for cognitive mechanisms. Trends in Cognitive Science, 18(9), 441-444.
Bond, C., \& DePaulo, B. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10(3), 214-234.
Bond, G. D. (2012). Focus on basic cognitive mechanisms and strategies in deception research (and remand custody of 'wizards' to Harry Potter movies). Journal of Applied Research in Memory and Cognition, 1(2), 128-130.
DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K., \& Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129(1), 74-118.
Dynel, M. (2011). A web of deceit: A neo-gricean view on types of verbal deception. International Review of Pragmatics, 3, 139167.

Goodman, N. D., \& Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences, 20(11), 818-829.
Grice, H. P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.
Horn, L. (2004). Implicature. In L. Horn \& G. Ward (Eds.), Handbook of pragmatics (p. 2-28). Blackwell Publishing.
Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. Developmental Psychology, 49(3), 404-418.
Montague, R., Navarro, D., Perfors, A., Warner, R., \& Shafto, P. (2011). To catch a liar: The effects of truthful and deceptive testimony on inferential learning. In L. Carlson, C. Holscher, \& T. Shipley (Eds.), Proceedings of the 33th annual meeting of the cognitive science society. Austin, TX: Cognitive Science Society.
Rogers, T., Zeckhauser, R., Gino, F., Schweitzer, M., \& Norton, M. (2014, september). Artful paltering: The risks and rewards of using truthful statements to mislead others (HKS Working Paper No. RWP14-045). Harvard Kennedy School. Retrieved from https://ssrn.com/abstract=2528625 doi: http://dx.doi.org/10.2139/ssrn. 2528625
Shafto, P., Eaves, B., Navarro, D. J., \& Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. Developmental Science, 15, 436-447.
Shafto, P., Goodman, N. D., \& Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by and learning from examples. Cognitive Psychology, 71, 55-89.
Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., \& Wilson, D. (2010). Epistemic vigilance. Mind \& Language.
Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., \& Storms, G. (2015). How do people learn from negative evidence? nonmonotonic generalizations and sampling assumptions in inductive reasoning. Cognitive Psychology, 81, 1-25.
Vrij, A., \& Granhag. (2012). Eliciting cues to deception and truth: What matters are the questions asked. Journal of Applied Research in Memory and Cognition, 1(2), 110-117.

# Connecting stimulus-driven attention to the properties of infant-directed speech Is exaggerated intonation also more surprising? 

Okko Räsänen (okko.rasanen@aalto.fi)<br>Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13000, 00076, AALTO, Finland<br>Sofoklis Kakouros (sofoklis.kakouros@aalto.fi)<br>Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13000, 00076, AALTO, Finland<br>Melanie Soderstrom (M_Soderstrom@umanitoba.ca)<br>Department of Psychology, University of Manitoba, P404 Duff Roblin Building, Winnipeg, MB R3T 2N2, Canada


#### Abstract

The exaggerated intonation and special rhythmic properties of infant-directed speech (IDS) have been hypothesized to attract infant's attention to the speech stream. However, studies investigating IDS in the context of models of attention are few. A number of such models suggest that surprising or novel perceptual inputs attract attention, where novelty can be operationalized as the statistical predictability of the stimulus in a context. Since prosodic patterns such as F0 contours are accessible to young infants who are also adept statistical learners, the present paper investigates a hypothesis that pitch contours in IDS are less predictable than those in adultdirected speech (ADS), thereby efficiently tapping into the basic attentional mechanisms of the listeners. Results from analyses with naturalistic IDS and ADS speech show that IDS has lower overall predictability of intonation across neighboring syllables even when the F0 contours in both speaking styles are normalized to the same frequency range.


Keywords: language acquisition; infant-directed speech; statistical learning; attention; stimulus predictability

## Introduction

Infant-directed speech (IDS) is a speaking style that talkers often use when interacting with young infants. In contrast to adult-directed speech (ADS), IDS tends to have exaggerated intonational contours with higher fundamental frequency (F0) and larger frequency range (e.g., Grieser \& Kuhl, 1988), hyperarticulated vowels (Kuhl et al., 1997; but see also Martin et al., 2015), and shorter utterances with a higher token/type ratio (Phillips, 1973). In addition to serving as language input tuned to the developmental stage of the listener (Snow, 1977), one hypothesized role of the exaggerated nature of IDS is that it may engage infants' attention to the speech stream more efficiently than ADS (e.g., Garnica, 1977; Fernald, 1989; see Soderstrom, 2007, for an overview), thereby facilitating language learning from speech.

Although the exaggerated intonation of IDS is often implicitly assumed to be the cause for higher attentional attractiveness, according to our knowledge, no study has systematically evaluated properties of IDS in the context of what is known about perceptual mechanisms for stimulusdriven attention. Instead, the evidence for higher attentional capture of IDS largely comes from behavioral studies that show that infants prefer to listen to IDS over ADS (Fernald,

1985; Cooper \& Aslin, 1990; Pegg, Werker \& McLeod, 1992). In addition, based on acoustic analyses and their perceptual correlates, IDS is often characterized as more salient or prominent than ADS, therefore also potentially being more interesting to the listeners (e.g., Garnica, 1977; Fernald, 1989). Since stimulus-driven attention and prominence of the perceived speech input seem both to be driven by unpredictability of the stimuli in the given context (see the next sub-section; but see also Kidd et al., 2012), the existing knowledge suggests that IDS might be more attractive to the listeners simply because it has different predictability properties over time than ADS. For instance, larger variability of F0 in IDS already implies, but does not guarantee ${ }^{l}$, higher uncertainty regarding the realization of the intonation at any moment in time. However, no study has systematically compared the prosodic predictability of IDS and ADS from a statistical learning point of view, even though infants are known to be sensitive to statistical regularities in their perceptual experience (c.f., Saffran et al., 1996; Soderstrom et al., 2009, and references therein) and to the prosodic structure of their native language already from an early age (e.g., Nazzi et al., 1998).

In the present paper, a quantitative investigation is carried out in order to test whether IDS is indeed not just more variable, but also less predictable than ADS, thereby being in line with the recent predictability-based accounts of perceptual attention. Importantly, we assume that the listener is able to learn the typical behavior of intonational contours from speech experience and this creates the basis for prosodic expectations for new speech input. In order to do this, a straightforward computational model of statistical learning is applied to F0 trajectories of naturalistic IDS and ADS and tested in its ability to predict intonational contours on speech utterances from both speaking styles.

## Stimulus-driven attention and statistical learning

A number of models for stimulus-driven perceptual attention suggest that attention is drawn to stimuli that are low-probability, or unpredictable, in the given context (Itti \& Baldi, 2009; Zhang et al., 2008; Tsuchida \& Cottrell,

[^182]2012; Zarcone et al., 2016), basically enabling the perceptual system to focus on aspects of the environment with the highest information content (Shannon, 1948), i.e., input that is not yet learned and thereby accurately predicted by the brain. However, infants are also known to prefer stimuli that are surprising or novel only as long as the input is not too unlikely in the given context, also known as the Goldilocks effect (Kidd et al., 2012). This suggests that the input should still be structured enough to support learning, thereby providing the basis for statistical expectations and evaluation of the relative information value of the inputs.

Earlier work with prosody perception suggests that lowprobability intonation patterns in the context of otherwise predictable prosody are associated with higher perceptual prominence of the concurrent words (Kakouros \& Räsänen, 2016a) and alter semantic processing of speech (e.g., Magne et al., 2005), having the same consequences as lowprobability words in the given context (see Kakouros et al., submitted, for a discussion). Recent evidence also suggests that adult listeners are sensitive, and rapidly adapt, to changing statistical properties of the intonation patterns, leading to experience-based expectations for prosody whose violations give rise to the subjective impression of prominence (Kakouros \& Räsänen, 2016b; Kakouros et al., submitted). Overall, the earlier research indicates that auditory attention and perceptual prominence are connected to the predictability of the prosodic patterns, and this may play a role also in the perception of IDS.

Importantly, the concept of predictability necessitates some type of mechanism for learning regularities from experience, thus connecting attention and prominence with the concept of statistical learning. The most parsimonious assumption would be that the prosodic learning utilizes the same statistical learning mechanisms hypothesized to play a role in other aspects of language acquisition, but now operating at the level of prosodic features such as F0 contours and energy envelopes instead of the phonemic units of the language. Since infants are known to be adept statistical learners, and since prosodic cues are perceptually accessible to them (e.g., Hawthorne, Mazuka \& Gerken, 2016), it is likely that infants are sensitive to statistical regularities present at the prosodic level similarly to adults.

If predictability of the stimulus in a given context is a major factor in controlling stimulus-driven attentional orientation, as also exemplified by the widely used preferential head-turn or looking-time paradigms to probe infants' learning, we would expect IDS to have different predictability properties than ADS. In the present study, we will look into one specific aspect of IDS, namely, intonation, and test how well F0 contours can be predicted over time for the two speaking styles in question.

## Data

The speech material used in the present experiments comes from the ManyBabies study that aims to replicate IDS preference across a large number of labs (The ManyBabies Consortium, 2017). In the context of that study, naturalistic
speech from female caregivers to their infants or from caregivers to other adults was recorded in central Canada and Northeastern US. All caregivers had infants aged 122250 days. The recordings were carried out in an infantfriendly greeting area/testing room using lapel clip-on microphones connected to smartphones. The task involved describing a closed set of labeled objects by asking the mother to take each object out of a bag one at a time and talk about it to her baby (IDS) or to an experimenter (ADS). In addition, there were two types of objects: those supposedly familiar to the infants (e.g., a ball or a block) and those considered as novel (e.g., a sieve or a whisk). After rough manual segmentation of the recordings into utterances, the utterances were also classified into three categories: utterances containing the familiar object word, those containing the unfamiliar object word, and utterances without naming of the object.

In the present study, we used the Canadian section of the recordings, containing speech from a total of 11 mothers. The US recordings ( 4 mothers) were excluded due the significant presence of room reverberations that could have impacted automatic F0 estimation. All utterances shorter than 1 s or with less than five syllables (see Methods) were discarded, leading to a total of $N=882$ utterances (504 IDS, 378 ADS) with an average of $80.2 \pm 29.9$ utterances per talker. Average utterance length was $4.0 \pm 2.5$ seconds (3.1 $\pm 1.1$ for IDS, $5.2 \pm 3.1 \mathrm{~s}$ for ADS).

## Methods

The overall goal of the analysis was to compare predictability of F0 trajectories in the IDS and ADS utterances using a statistical model. This was done by first syllabifying and estimating F0 trajectories for all speech, parametrizing F0 trajectories during each syllable, clustering the syllable-specific parameters into a finite number of categories ("F0 shapes") in an unsupervised manner, and then modeling the temporal evolution of these F0 states across time. By training the predictive model from a set of utterances and then computing the likelihoods of F0 trajectories on a set of held-out utterances, measures of F0 predictability can be estimated from the data. Fig. 1 shows a schematic picture of the processing pipeline for an individual utterance. All experiments were conducted in MATLAB unless mentioned otherwise.

## Pre-processing of F0 trajectories

F0 trajectories were estimated at a $100-\mathrm{Hz}$ sampling rate with YAAPT-algorithm (Zahorian \& Hu, 2008; version 4.0), constraining F0 estimates to the range of $100-600 \mathrm{~Hz}$ and using YAAPT's ptch_fix() tool for post-processing of the pitch tracks for potential estimation errors and for interpolation of the trajectories across unvoiced regions. For the predictability analysis, utterance-level F0 tracks were zscore normalized to zero mean and unit variance in order to focus on temporal behavior instead of the absolute mean or range of the pitch. In addition, the original non-normalized F0 contours were used as baseline features in the analyses.


Figure 1: A schematic view of the F0 predictability analysis. The output is the probability of F0 in syllable $s$ given the observed F0 in $m$ preceding syllables (after training the statistical model on a number of training utterances).

All utterances were syllabified using a sonority envelopebased automatic syllabifier described in Räsänen, Doyle and Frank (submitted; see also Räsänen, Doyle \& Frank, 2015, for an earlier but similar version). All syllables without any frames with reliable voicing (as determined by YAAPT) and syllables shorter than $50-\mathrm{ms}$ were merged with the neighboring syllables, leading to a total of 8056 syllables in the data set. Note that although this type of acoustic syllabification is not perfectly accurate in terms of the phonological rules of the language, it still provides systematic chunking of speech into syllable-like units with each unit consisting of a sonorous peak surrounded by lesssonorous onsets and coda (see also, e.g., Villing, Ward \& Timoney, 2006, and references therein). Importantly, such acoustic-signal based chunking can be argued to better match the syllabification capabilities of pre-linguistic infants that also must rely on non-phonological acoustic cues in their perception of speech before they master the sound system of their native language (Räsänen et al., submitted).

Following the syllabification, F0 trajectories during each syllable were parametrized by fitting a second order polynomial to the trajectory in time (Fig. 2) and using the polynomial coefficients without the constant term as a parametric description of the F0 during the syllable. Parameters across all syllables in the data were then vector quantized into $Q$ discrete categories using standard k-means clustering with random initialization. In practice, these $Q$ shapes correspond to different F0 patterns with varying curvature and rate of change as a function of time, larger $Q$ simply meaning more fine-grained distinction between F0 patterns that occur during the syllables.

## Temporal modeling of $\mathbf{F 0}$ state sequences

As a result of the pre-processing, the F0 trajectory of each utterance was described as a sequence of discrete states $q_{s} \in$ $Q$, one state per syllable $s$. In order to quantify the predictability of F0, a mixed-order Markov chain model, or MOCM, was trained for the sequences (Saul \& Pereira, 1997). Instead of computing $n$-gram statistics for different $n$-gram orders and then choosing and/or merging the models with best predictive capability, MOCM allows modeling of varying order Markov chains with a single set of model parameters. In MOCM, the probability of an F0 shape $q_{s}$ in syllable $s$, given the preceding $m$ syllables, is calculated as


Figure 2: An example of syllable-wise $2^{\text {nd }}$ order polynomial approximation of the F0 trajectory. Top: The original speech waveform. Bottom: YAAPT-estimated and z-score normalized F0 trajectory with interpolation across unvoiced segments (blue solid line) and the corresponding $2^{\text {nd }}$ order polynomial least-squares fit for F0 during each syllable. Syllable boundaries are shown with vertical lines.

$$
\begin{align*}
& P\left(q_{s} \mid q_{s-1}, \ldots, q_{s-m}\right)= \\
& \sum_{k=1}^{m} \lambda_{k}\left(q_{s-k}\right) \mathbf{M}_{k}\left(q_{s-k}, q_{s}\right) \prod_{j=1}^{k-1}\left[1-\lambda_{j}\left(q_{s-j}\right)\right] \tag{1}
\end{align*}
$$

where lag-specific transition matrices $\mathbf{M}$ and transition weights $\boldsymbol{\lambda}$ are estimated from training data using the Expectation Maximization (EM) algorithm (Saul \& Pereira, 1997). In the context of the present study, $\mathbf{M}_{k}$ describes the transition probabilities between syllabic F0 contours at different lags $k$ while $\lambda$ weighs these probabilities from different distances based on the reliability of the probability estimates in the context of the observed shapes.

In the experiments, a third order $(m=3)$ MOCM model was trained using the syllabic F0 sequences from $90 \%$ of the combined pool of IDS and ADS utterances. This was followed by syllable-by-syllable estimation of F0 likelihoods on the remaining held-out utterances using Eq. (1). The procedure was repeated in a 10 -fold manner until all utterances had been used in the training and test sets. The division of utterances into training and testing sets was purely random, and therefore both contained speech from the same 11 unique talkers. We decided not to use speakerspecific models for F 0 due to the modest number of utterances per talker that would have caused data sparsity issues in the model estimation. As a result, the obtained probability estimates describe how expected is the F0 behavior in the given context given a preceding exposure to a large number of F0 trajectories, low probability reflecting unexpected and thereby attention capturing intonation.

Note that the choice of $Q$, the number of quantization levels for the F0 shapes, contains an inherent tradeoff between the resolution of the F0 trajectory modeling and the amount of data required for model estimation. Although there is no a priori reason to consider any $Q$ specifically favoring IDS or ADS due to the z-score normalization of all F0 values, we wanted to minimize the impact of $Q$ in our
findings. Therefore the simulation was conducted for $Q=6$, 12 , and 24 with syllable-specific likelihood estimates averaged across all these runs. In addition, all likelihoods were averaged across five runs of the entire experiment to diminish variation caused by random initialization of the k means clustering process, even though the k-means clustering results for the two dimensional features were found to be highly consistent across individual runs.

## Data analysis

Five utterance-level statistical descriptors, namely, the mean, SD, min, max, and range ( $\max -\min$ ) were calculated for the F0 likelihoods across all syllables and for the original F0 trajectories in Hz in each utterance. Talker and style-specific (IDS vs ADS) means for the descriptors were then averaged across all the utterances from the given talkers. Before any statistical analysis, the statistical descriptors for F0 likelihoods were corrected for the variable amount of matching training data for the speaker and speaking style in question. This was done by first fitting a speaker-independent linear regression model from the number of matching training samples to the statistical descriptors, and then subtracting the prediction from the original values, basically decorrelating the measures with respect to the amount of training data.

In order to test differences between IDS and ADS, the normalized descriptors for F0 predictability and descriptors for the original F0 values were then compared between the IDS and ADS conditions using the paired t-test with significance level of $p<0.05$ (Holm-Bonferroni corrected for the ten comparisons and $d f=10$ for all reported stats).

## Results

Fig. 3 shows a summary of the results together with markers and $t$-statistics for significant differences between IDS and ADS. As expected, the mean frequency of F 0 in the utterances is higher in IDS $(210.9 \mathrm{~Hz} \pm 29.0 \mathrm{~Hz})$ than in ADS ( $189.9 \pm 23.9 \mathrm{~Hz}$ ). In addition, the average utterancelevel maximum and minimum F0 are significantly higher in IDS, but the overall variability and absolute range (in Hz ) are not different between the speaking styles.

As for the predictability, the mean predictability of F0 in IDS was significantly lower than in $\operatorname{ADS}(t(10)=4.82$, Cohen's $d=1.93$ ). In addition, maximum predictability during each utterance was also lower $(t(10)=5.46, d=2.10)$ and so was the range of predictability values across the syllables in the utterances $(t(10)=5.19, d=1.88)$. In contrast, variability of predictability across the utterances was not different between IDS and ADS. Notably, the average F0 probabilities are within a similar range to what was found to be optimal stimulus complexity for attentional capture in the visual perception experiments of Kidd et al. (2012) and significantly above chance-level ( $p=0.0972$ ). This suggests that the F0 trajectories might be in a suitable complexity region for triggering novelty preference, enabling predictive learning but also leaving room for unpredictable patterns and events.


Figure 3: Top: Utterance-level statistical descriptors of F0 predictability, averaged across all ADS/IDS utterances. Middle: F0 predictability after controlling for the amount of matching training data (speaker \& style) for each utterance. Bottom: Utterance-level descriptors of original F0 in Hz, averaged across all utterances. Error bars denote $\pm 1 \mathrm{SE}$ across all talkers. Significant differences between IDS and ADS are denoted with asterisks and related t -values (paired t-test, $d f=10$, and using significance level $p<0.05$ with Holm-Bonferroni correction for the ten comparisons).

We also repeated the entire analysis but now using linear instead of the 2 nd order model for the syllabic F0 contours (i.e., encoding only the average direction and rate of change in F0 during the syllable). This replicated all the main findings (significantly lower mean, max, and range for the predictability of F0 in IDS; not shown separately). We also tested whether there were differences in the predictability descriptors between the three sentence types (familiar object, unfamiliar object, no labeling) but none of the tests were significant after controlling for multiple comparisons. In addition, the predictability difference is not simply due to a larger quantization error for IDS parameters, since the reported pattern of results persists also if only the IDS data are used for the k-means codebook creation leading to lower quantization errors (RMSE) for the IDS F0 trajectories.

Overall, the main result confirms the hypothesis that the intonation contours in IDS are less predictable than in ADS, at least for the present data set in question.

As a follow-up validation of the findings, we also ran binary logistic regression to classify all the individual utterances into IDS or ADS classes using the utterance-level descriptors for probabilities and raw F0 values as features and using likelihood ratio as the criterion for forward stepwise feature selection (using SPSS version 23.0, IBM Corp., Armonk, NY). The resulting model achieved IDS/ADS utterance classification rate of $74.8 \%$ using a final
set of four features: SD of likelihood (Wald statistic $=$ $31.28, p<0.001 ; d f=1$ for all features), mean likelihood ( $W$ $=23.08, p<0.001$ ), likelihood range $(W=88.34, p<$ 0.001 ), together with max of original F 0 in $\mathrm{Hz}(W=66.24, p$ $<0.001$ ). This further shows that the predictability differences of intonation in IDS and ADS do not simply appear as aggregate measures across a large number of utterances, but can be also used to effectively classify individual utterances into ADS or IDS.

Finally, a subset of the utterances used in the present study had been previously rated for their IDS-likeness using a low-pass version of the recordings as part of the ManyBabies project (see The ManyBabies Consortium, 2017, for details). These utterances were rated on a 7-point Likert scale by several naïve raters recruited from Amazon's mechanical Turk. We therefore calculated the correlation between all the utterance-level F0 descriptors and the human IDS-likeness ratings for all the IDS utterances with ratings $(N=442)$. The human judgments of IDS-likeness correlated with the mean (Spearman's $r=0.25, p<0.001$ ), SD ( $r=0.31, p<0.001$ ), min $(r=0.154, p=0.002)$, max $(r$ $=0.35, p<0.001)$, and range $(r=0.32, p<0.001)$ of the original F0 values, i.e., with all of them. Surprisingly, all the descriptors of F0 trajectory likelihoods were uncorrelated with the human ratings ( $p>0.05$ for all comparisons).

Since predictability was nonetheless a reliable cue in our classification of utterances into IDS and ADS based on the original study labels, the finding with the naïve ratings data suggests a dissociation between perceptual correlates of IDS-like speech in naïve listeners (e.g., high and variable pitch) and the lower predictability of intonation in IDS as a potential attractor of listeners' attention. Notably, an earlier study by Singh, Morgan \& Best (2002) has also reported that higher and more variable pitch alone was not sufficient to capture infants' attention when pitted against affective speech. This suggests that the properties that make an utterance sound IDS-like to a naïve listener may be unrelated to those that lower the predictability of IDS. How those properties relate to the attentional attractiveness of IDS is currently unclear and requires further investigation.

## Discussion and conclusions

This study aimed to test whether the exaggerated intonation in IDS also translates into less predictable prosody over time. The results support this idea, even when the actual mean and range of F0 values in the predictive analysis was normalized between the IDS and ADS utterances. In addition, while IDS intonation is less predictable than ADS, it is still relatively structured as indicated by the mean predictability that is significantly above the chance-level given the analyzed quantization levels. These findings provide initial support to the idea that IDS may be more attentionally attractive simply because it is more surprising without being too chaotic (c.f., Kidd et al., 2012), thereby tapping to the basic attentional mechanisms causing orientation towards unfamiliar inputs.

In addition, some evidence for a dissociation between human ratings of IDS-likeness and predictability of the utterances was also discovered, warranting further research in the issue. In fact, a dissociation between F0 variability and F0 predictability is expected on the basis of predictability-based accounts to prominence and attention in speech. More specifically, it has been argued that the perceptual system should allocate processing resources to the aspects of the input that are not yet predicted by the brain independently of the physical magnitude or other absolute property of the input. In contrast, highly predictable inputs, by definition, have low information value and are therefore low priority targets for sensing and learning even if they have large magnitude on some scale such as loudness or pitch (e.g., Kakouros \& Räsänen, 2016b; Kakouros et al., submitted; see also, e.g., Friston \& Kiebel, 2009). In the context of speech, the talker can control the listener's attention by freely using non-canonical prosodic forms on any word or words of choice without changing the literal meaning of the utterance (Kakouros et al., 2016b; Kakouros et al., submitted). The present study suggests that caregivers may (implicitly) utilize a similar strategy to maintain infants' attention on the speech stream or highlighting certain segments of speech.

However, the present work only provides an initial investigation into the predictability aspects of IDS using a certain modeling approach. Much more work is needed to understand the underpinnings of IDS and how it relates to learning and attention mechanisms of the human cognition. This also includes the need to replicate the present investigation on different speech data and also preferably with alternative approaches to quantifying suprasegmental statistical structure. In addition, prosody is much more than F0 trajectories, and therefore aspects such as timing, utterance duration, and intensity should be investigated from the predictability point of view independently and in conjunction with F0.

## Acknowledgments

O.R. was funded by Academy of Finland project titled "Computational modeling of language acquisition" (274479). All MATLAB scripts for syllabification, mixedorder Markov chains, and reported analyses can be found at https://git.io/vDT0N. The original recordings are available from The ManyBabies Consortium at https://osf.io/re95x/. The authors would like to thank Michael C. Frank for his help with the ManyBabies data.

## References

Cooper, R. P., \& Aslin, R. N. (1990). Preference for infantdirected speech in the first month after birth. Child Development, 61, 1584-1595.
Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. Infant Behavior and Development, 8, 181195.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? Child Development, 60, 1497-1510.
Friston, K. \& Kiebel, S. (2009). Cortical circuits for perceptual inference. Neural Networks, 22, 1093-1104.
Garnica, O. K. (1977). Some prosodic and paralinguistic features of speech to young children. In C. E. Snow \& C. A. Ferguson (Eds.): Talking to children: Language input and acquisition (pp. 63-88). Cambridge, UK: Cambridge University Press.
Grieser, D., \& Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: support for universal prosodic features in motherese. Developmental Psychology, 24, 14-20.
Hawthorne, K., Mazuka, R., \& Gerken, L. (2015). The acoustic salience of prosody trumps infants' acquired knowledge of language-specific prosodic patterns. Journal of Memory and Language, 82, 105-117.
Itti, L., \& Baldi, P. (2009). Bayesian surprise attracts human attention. Vision Research, 49, 1295-1306.
Kakouros, S., \& Räsänen, O. (2016a). Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features. Cognitive Science, 40, 1739-1774.
Kakouros S. \& Räsänen O. (2016b). Statistical Learning of Prosodic Patterns and Reversal of Perceptual Cues for Sentence Prominence. Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA, pp. 2489-2494
Kakouros, S., Salminen, N., \& Räsänen, O. (submitted). Making predictable with style - Behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech. Submitted for publication.
Kidd, C., Piantadosi, S. T., \& Aslin, R. N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. PLoS ONE, 7(5), e36399.
Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... \& Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. Science, 277, 684-686.
Magne, C., Astèsano, C., Lacheret-Dujour, A., Morel, M., Alter, K., \& Besson, M. (2005). On-line processing of "pop-out" words in spoken French dialogues. Journal of Cognitive Neuroscience, 15, 740-756.
Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., \& Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. Psychological Science, 26, 341-347.
Nazzi, T., Bertoncini, J., \& Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. Journal of Experimental Psychology, 24(3), 756-766.
Pegg, J. E., Werker, J. F., \& McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech:

Evidence from 7-week-old infants. Infant Behavior and Development, 15, 325-345.
Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. Child Development, 44, 182-185.
Räsänen O., Doyle G. \& Frank M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. Proc. Interspeech2015, Dresden, Germany, pp. 3204-3208.
Räsänen, O., Doyle, G., \& Frank, M. C. (submitted). Prelinguistic rhythmic segmentation of speech into syllablelike units. Submitted for publication.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926-1928.
Saul, L. \& Pereira, F. (1997). Aggregate and mixed-order Markov models for statistical language processing. In Proc. 2nd Conf. Empirical Methods Natural Language Processing, Providence, RI, USA, Aug. 1997, pp. 81-89.
Shannon, C. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423.
Singh, L., Morgan, J. L., \& Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? Infancy, 3, 365-394.
Snow, C. E. (1977). The development of conversation between mothers and babies. Journal of Child Language, 4, 1-22.
Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. Developmental Review, 27, 501-532.
Soderstrom, M., Conwell, E., Feldman, N., \& Morgan, J. (2009). The learner as statistician: three principles of computational success in language acquisition. Developmental Science, 12, 409-411.
The ManyBabies Consortium (2017). Quantifying the sources of variability in infancy research using the infantdirected speech preference. Manuscript under review. https://osf.io/re95x/
Tsuchida, T., \& Cottrell, G. W. (2012). Auditory saliency using natural statistics. Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci2012), Sapporo, August 1-4 (pp. 1048-1053).

Villing, R., Ward, T., \& Timoney, J. (2006). Performance limits for envelope-based automatic syllable segmentation. Proc. ISSC-2006, Dublin, Ireland, pp. 521526.

Zahorian, S. \& Hu H. (2008). A spectral/temporal method for robust fundamental frequency tracking. The Journal of The Acoustical Society of America, 123, 4559-4571.
Zarcone, A., van Schijndel, M., Vogels, J., \& Demberg, V. (2016). Salience and attention in surprisal-based accounts of language processing. Frontiers in Psychology, 7, article no. 844.
Zhang, L., Tong, M. H., Marks, T. K., Shan, H., \& Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision, 8, 1-20.

# When does a 'visual proof by induction' serve a proof-like function in mathematics? 

Josephine Relaford-Doyle (jorelafo@ucsd.edu) Rafael Núñez (rnunez@ucsd.edu)<br>Department of Cognitive Science, University of California - San Diego<br>9500 Gilman Drive, La Jolla, CA 92093 USA


#### Abstract

A proof by mathematical induction demonstrates that a general theorem is necessarily true for all natural numbers. It has been suggested that some theorems may also be proven by a 'visual proof by induction' (Brown, 2010), despite the fact that the image only displays particular cases of the general theorem. In this study we examine the nature of the conclusions drawn from a visual proof by induction. We find that, while most university-educated viewers demonstrate a willingness to generalize the statement to nearby cases not depicted in the image, only viewers who have been trained in formal proof strategies show significantly higher resistance to the suggestion of large-magnitude counterexamples to the theorem. We conclude that for most university-educated adults without proof-training the image serves as the basis of a standard inductive generalization and does not provide the degree of certainty required for mathematical proof.


Keywords: mathematical reasoning; proof; mathematical induction; visual proof; induction; generalization

## Introduction

Mathematics has been defined as "the science which draws necessary conclusions" (Peirce, 1881). To this end, proofs are indispensible to formal mathematics. A mathematical proof uses deductive logic to establish the truth of a general theorem - for instance, that a property holds for all triangles, or all natural numbers. Relative to the logical system being used, the conclusion of the proof is certain; if the premises are true, then the conclusion is necessarily true as well. The certainty of results obtained through formal, deductive proof is a defining feature of mathematics.

Mathematical induction, despite what its name suggests, is a well-established deductive proof method that can be used to prove that a theorem holds for all natural numbers. It has been suggested that some general theorems that can be formally proved using mathematical induction may also be proved using specially designed images known as 'visual proofs' (Brown, 2010). The claim that a 'visual proof by induction' can prove a general theorem is an interesting one, since any image is necessarily finite and thus can only display a particular set of cases of the theorem. Case-based argumentation falls under the umbrella of inductive reasoning, which does not provide certain conclusions and is not accepted in formal mathematical justification. However, visual proofs contain additional structure that could be leveraged to demonstrate that a theorem necessarily holds in all cases, even those not depicted in the image. Thus, it is possible that a visual proof, despite displaying only a finite number of cases, could serve a proof-like function for some viewers.

Although the status of visual proofs is at the center of a debate in the philosophy of mathematics (see, e.g., Brown, 2010; Doyle, Kutler, Miller, \& Schueller, 2014; Folina, 1999), they have been largely ignored within Cognitive Science and little is known about the nature of reasoning with these images. How do viewers reason with a visual proof by induction? Do they consider the conclusions to be certain, as in mathematical induction, or only likely, as in standard inductive reasoning?

## Induction in Mathematics, Mathematical Induction, and Visual Proofs

The distinction between certain, necessary conclusions and probable or likely conclusions is of central concern in mathematics. Proofs - deductive arguments which provide certain conclusions - are exalted. The writing of proofs, however, comprises only a small part of mathematical practice, and it is widely acknowledged that inductive reasoning plays an important role in mathematics (see Polya, 1954 for an account of induction in mathematics). A commonly held view is that inductive reasoning is an essential part of mathematical discovery, while deduction is required for formal mathematical justification (i.e., proof).

Consider the expressions in Figure 1(a). One might notice a pattern in these examples, namely, that when one adds consecutive odd numbers starting at 1 , the resulting sum seems to be the square of the number of terms being added. We might guess that this pattern holds for other numbers; for example, we might predict that the sum of the first 8 odd numbers is 64 . However, while these six examples allowed us to discover a possible relationship, the examples themselves do not prove that the general theorem is true for all natural numbers. Without a formal proof, any conjecture we have is uncertain and remains open to the possibility of counterexamples. A formal proof of our theorem using mathematical induction is given in Figure 1(b).

Figure 1(c) shows a visual proof of the same theorem (from Brown, 1997). In the image, consecutive odd numbers of dots are arranged in layers, beginning with 1 in the lower left-hand corner. When the dots in the first $n$ layers are considered together the resulting array forms a square, and so the total number of dots in the array is given by $n^{2}$. While the image displays only the first six cases of the general theorem, a viewer might be inclined to guess that the pattern will continue to hold as more layers are added, and therefore be convinced that the general theorem is true. Indeed, images such as these have been described as "rapidly and deeply convincing" (Doyle et al., 2014).

| Theorem: The sum of the first $\boldsymbol{n}$ odd numbers is equal to $\boldsymbol{n}^{\mathbf{2}}$. |  |  |
| :---: | :---: | :---: |
| (a) Six cases | (b) Proof by mathematical induction Theorem: $1+3+\ldots+(2 n-1)=n^{2}$ | (c) Visual Proof |
| $1=1^{2}$ | Base case: $n=1 \rightarrow 1=1^{2}$ | 000000 |
| $1+3=4=2^{2}$ | Inductive step: Assume $1+3+\ldots+(2 k-1)=k^{2}$, |  |
| $1+3+5=9=3^{2}$ | for some fixed number $k$. Adding the next odd | 00000000 |
| $1+3+5+7=16=4^{2}$ | number $2 k+1$ to both sides of the equation, we | 00000000 |
| $1+3+5+7+9=25=5^{2}$ | have: $1+3+\ldots+(2 k-1)+(2 k+1)=k^{2}+(2 k+1)$ | $\begin{array}{llll} 000 & 0 & 0 \end{array}$ |
| $1+3+5+7+9+11=36=6^{2}$ | Re-writing the last odd term and factoring the right side gives us: | $\begin{array}{lll\|l\|l\|l} \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \end{array}$ |

Figure 1: Varying forms of evidence for a general theorem. (a) Specific cases suggest, but do not prove, the general theorem. (b) Formal proof of the general theorem by mathematical induction. (c) Visual proof (from Brown, 1997).

It is unclear, however, the exact nature of the conclusions drawn from the visual proof. As many have pointed out, any image is necessarily finite and thus can only display particular cases of a general theorem (Doyle et al., 2014; Rips \& Asmuth, 2007). This would suggest that the image in Figure 1(c), like the cases presented in 1(a), would serve as the basis for a standard inductive generalization - the image might convince the viewer that the theorem likely holds for all natural numbers, but cannot provide certainty. On the other hand, the image contains structure that is not available in the same cases presented numerically, and which could be exploited in order to demonstrate that the property would necessarily continue to hold for values not depicted in the image. Such an argument would need to demonstrate that the square shape is preserved if and only if the next layer contains the next consecutive odd number of dots. For example, if we start with an $n \times n$ square, the $(n+$ $1)^{\text {st }}$ layer could be constructed by copying the $n^{\text {th }}$ layer and translating the copy up one unit and right one unit (Figure $2 \mathrm{a}, \mathrm{b})$. This results in two vacant positions that must be filled in order to maintain the square shape (Figure 2c, d). Thus, every layer must contain exactly two more dots than its predecessor. Since the difference between any two consecutive odd numbers is 2 , we can conclude that the new layer must contain the next consecutive odd number of dots.


Figure 2: Rigorous image-based argument for the general theorem

Though not a traditional deductive proof, it could be argued that an argument such as this does establish the truth of the general theorem for all natural numbers, and that this conclusion meets the level of certainty required for mathematical proof. It is unknown how accessible such arguments are to viewers, and, more generally, how closely the conclusions drawn from the visual proof resemble the conclusions drawn from a formal proof. In this study we seek to assess the extent to which a visual proof may serve a proof-like function. Specifically, we ask two key questions: (1) Given the visual proof, do viewers generalize the theorem to cases not depicted in the image? (2) If so, is that conclusion considered certain, as in mathematical induction, or only likely, as in standard inductive reasoning?

Finally, to properly address these questions we consider who is viewing the image and in what context the image is viewed. In this study we compare two groups of viewers, one drawn from the general population of university students and one drawn from a group that has received university-level training in formal proofs, including mathematical induction. Additionally, to address the key theoretical difference in mathematics between discovery (in which inductive reasoning is acceptable) and justification (in which it is not), we manipulate the context in which the image is viewed by varying the amount of information provided to the viewer.

## Method

## Participants

Two groups were drawn from distinct populations. The first group ( $n=25$ ) was recruited through the university subject pool. None of these participants had taken a university-level course on mathematical proofs, and so we refer to this group as "proof-untrained" (PU). The second group ( $n=24$ ) was recruited through the mathematics department and consisted of individuals who had received at least a B- in "Mathematical Reasoning", a university-level mathematics course on formal proof strategies including mathematical induction. We refer to this group as "proof-trained" (PT). PT participants had taken significantly more university-level math classes than had PU participants (mean PT $=6.67$,
mean $\mathrm{PU}=2.8, \mathrm{t}=6.58, \mathrm{df}=44.63, \mathrm{p}<0.001)$. No PU participants had taken the "Math Reasoning" course on proofs (but four of the 25 indicated familiarity with mathematical induction).

## Materials

Each participant received one of three tasks. All three tasks included the visual proof in Figure 1(c), which was designed to prove the target statement "The sum of the first $n$ odd numbers is equal to $n^{2 " \prime}$, but varied in the amount of information that was provided to the reader.
Condition A - Justification: The participant was given the full target statement and the visual proof. They were asked to explain how the picture shows that the statement is true.
Condition B - Supported Discovery: The participant was given the visual proof and a fill-in-the-blank version of the target statement ("The sum of the first $n$ odd numbers is equal to $\qquad$ ."). They were asked to fill in the blank and explain how they got their answer.
Condition C - Full Discovery: The participant was given the visual proof and told that a mathematician drew the picture while trying to prove a statement about the sum of odd numbers. They were asked to guess what the mathematician was trying to prove and explain their answer.

Additionally, participants completed a background questionnaire in which they provided the number and names of university math courses they had taken and their level of familiarity with mathematical induction.

## Procedure

Each participant received one of the three task sheets, and the researcher explained that their ultimate task was to create a short tutorial video in which they would explain their response as clearly as possible to potential third-party viewers. A camera was set up directly above the participant's workspace, recording their writing, speech, and manual gestures (Figure 3). Before filming their video each participant was given as much time as they needed to think and plan their response. During this time participants had access to pencils, highlighters, and blank paper, and were free to add any markings to the sheet that might be helpful in explaining their response. The participant indicated they


Figure 3: Screenshot from video footage of a participant's workspace
were ready to start their tutorial video by placing a sign under the camera, and then filmed their explanation. The planning and filming were entirely self-paced and occurred without the researcher present.

When the participant indicated that they had finished their tutorial video, the researcher returned to the room and conducted a semi-structured interview with the participant. To assess whether the participant had generalized the target statement to cases not depicted in the image, any participant who demonstrated understanding of the target statement was asked two questions: "Do you think the statement is true in all cases?" (Q1) and "What would be the sum of the first 8 odd numbers?" (Q2). If the participant indicated generalization to nearby cases (by answering "yes" and " 64 ", respectively), the researcher raised the possibility that large-magnitude counterexamples to the statement may exist and asked the participant what they thought about that suggestion. Any participant who resisted the suggestion of counterexamples was asked how they would argue against such a possibility. After the interview all participants completed the background questionnaire.

## Analysis

Two coders scored each video for six distinct outcomes. The participant's tutorial video received three scores:
(a) Mathematical Statement: Rated whether the participant demonstrated understanding of the target statement.
(b) Explanation Strategy: Rated whether the participant gave a case-based explanation (using the image to show particular cases of the statement), or a pattern-based explanation (describing a general pattern in the image).
(c) Relevant Features of Image: Reflected which features of the image the participant identified as relevant, including odd numbers in layers, square shape, possibility of pattern extension, and necessity of pattern extension.

An additional three scores were given based on the interview portion of the study.
(d) Generalization: Rated whether the participant demonstrated generalization of the target statement to nearby cases as assessed by questions Q1 and Q2.
(e) Resistance to Counterexamples: Rated the participant's resistance to the possibility of large-magnitude counterexamples, ranging from no resistance (0) to complete rejection (5).
(f) Image-Based Argument: Rated whether the participant provided a rigorous image-based argument comparable to the argument represented in Figure 2.

For nominal criteria (a)-(d) and (f), the two coders showed 96.2\% agreement (Cohen's Kappa=0.96). Criterion (e) was was scored on 1-5 scale and also showed high reliability between coders (Krippendorff's alpha $=0.894$ ).

## Results

The two groups differed with respect to the ability to demonstrate understanding of the target statement across conditions. While proof-trained (PT) participants systematically demonstrated such understanding, proof-
untrained (PU) participants' ability to do so varied significantly across conditions (Fisher Exact test, $p=0.012$; Figure 4). All PU participants showed understanding of the target statement when it was provided (Condition A), but only $6 / 8$ participants provided the target response of $n^{2}$ in Condition B, and $3 / 9$ participants generated the complete target statement in Condition C. PT participants, on the other hand, did not significantly vary in their likelihood to demonstrate understanding of the target statement across conditions (Fisher Exact test, $p=0.3$; Figure 4).


Figure 4: Proportion of participants who demonstrated understanding of target statement.

Of the 17 total PU participants across all conditions who demonstrated understanding of the target statement, 10 (59\%) used case-based explanations in their tutorial video (Figure 5); notably, none of these 10 participants referred to the square shape as a relevant feature of the image during their explanation. Explanation strategy varied significantly across conditions, with participants in Condition A showing a stronger preference for case-based strategies, while participants in Condition C were more likely to generate pattern-based explanations (Fisher Exact test, $p=0.026$ ) and more likely to mention the square shape ( $p=0.026$ ). As we found no other significant effects of task context, in the following analysis we group participants across all conditions who demonstrated understanding of the target statement, keeping PU and PT groups separate.


Figure 5: Explanation Strategy: PU participants were significantly more likely than PT participants to give casebased (rather than pattern-based) explanations.

Proof-Untrained (PU) Group In the interview, 16 of the 17 (94.1\%) PU participants who demonstrated understanding of the target statement indicated a willingness to generalize the statement to nearby cases (Figure 6a). Only 5 (31\%) of these participants indicated a high degree of
resistance to large-magnitude counterexamples (characterized by a resistance score of 4 or higher), and only one stated that counterexamples were impossible (Figure 6b). Notably, questionnaire responses revealed that three of the five PU participants who showed high resistance were familiar with mathematical induction. When asked for an argument against counterexamples, only two PU participants were able to generate a rigorous argument based on the image. PU participants had taken significantly fewer university-level math courses than had PT participants; however, number of math courses was not significantly correlated to any study outcomes for the PU group.

Proof-Trained (PT) Group Across all conditions there were 22 PT participants who demonstrated understanding of the target statement. These participants were significantly more likely than PU participants to provide pattern-based explanations (Fisher Exact test, $p=0.026$ ), with only $2 / 22$ (9\%) relying on case-based strategies (Figure 5). PT participants were significantly more likely than PU participants to mention the square shape as a relevant feature of the image ( $21 / 22$; Fisher Exact test $p<0.001$ ). In the interview, all 22 PT participants who demonstrated understanding of the target statement indicated a willingness to generalize the statement to nearby cases not depicted in the image. The likelihood to generalize did not differ between PT and PU participants (Figure 6a); however, PT individuals were significantly more likely to indicate a high degree of resistance to the suggestion of large-magnitude counterexamples (17/22; Fisher Exact test, p=0.008; Figure $6 b)$. When considering all participants who demonstrated understanding of the target statement, PT participants were significantly more likely to provide a rigorous image-based argument against counterexamples than PU participants (8/22; Fisher Exact test, p $=0.035$ ).


Figure 6: (a) Participants in both groups generalized the target statement to nearby cases. (b) However, PT participants showed significantly higher resistance to largemagnitude counterexamples than did PU participants.

## Discussion

The present study explored the conditions in which a visual proof by induction may serve a proof-like function, characterized by generalization to all natural numbers and a belief that this conclusion is necessarily true. Our findings reveal significantly different outcomes for the proof-trained and proof-untrained participants. Specifically, while both groups demonstrated a willingness to generalize to nearby cases, the PU participants showed relatively low resistance to the suggestion that large-magnitude counterexamples may exist. This suggests that for these viewers the visual proof serves as the basis for a standard inductive generalization, and does not provide certainty. Further evidence for this analysis comes from the observation that PU participants were significantly more likely to provide case-based explanations, using the image to demonstrate one or more particular cases of the general theorem. PT participants, on the other hand, showed higher resistance to counterexamples and were more likely than PU participants to provide a rigorous argument using the image. Thus, it seems that the image can serve a proof-like function for viewers who have been trained in formal proof methods. The significant differences between the PT and PU groups contradict claims that visual proofs by induction are equally convincing to all viewers regardless of their knowledge of mathematical induction (Brown, 2010), or that interpreting the image as a proof requires only "basic secondary school knowledge of mathematics" (Jamnik, 2001).

We were surprised to find that the PU participants highly educated adults enrolled at a prestigious university often overlooked key features of the visual proof. Less than $60 \%$ of the PU participants who demonstrated understanding of the target statement mentioned the square shape as a relevant feature of the image. Furthermore, many participants who re-drew the image during their explanation did so in a way that violated the row-column structure of the square array (Figure 7), indicating that they were truly unaware of its importance. However, failure to notice the relevance of the square shape does not explain the PU group's low resistance to counterexamples, as mentioning the square shape was not significantly related to high resistance (Fisher Exact test, $\mathrm{p}=0.59$ ) within this group.


Figure 7: Work of PU participants who re-drew the image in a way that violated the row-column structure and square shape of the array

There were 5 PU participants who expressed a high degree of resistance to the suggestion of large-magnitude counterexamples, two of whom were unfamiliar with mathematical induction. We cannot conclude, however, that the image was serving a proof-like function for these viewers. Prior research has shown that adults do not reliably distinguish between inductive and deductive mathematical arguments and often accept case-based arguments as valid proofs of statements about infinite sets (Eliaser, 2000; Martin \& Harel, 1989). Thus, even if the image functions as a basis for a standard inductive generalization, we would nonetheless expect to see a group of participants who find it highly convincing.

PT participants were more likely than PU participants to show high resistance to counterexamples, and subsequently more likely to provide a rigorous image-based argument for the general theorem. What accounts for these differences? One possibility is that PT participants had been exposed to significantly more university-level mathematics than PU participants. However, the number of university math classes taken prior to participation in the study was not related to any outcome for either group. This suggests that the differences between the two groups cannot be explained simply based on differing amounts of exposure to general mathematics. Instead, it seems that training in proof-writing - a specific and highly technical mathematical practice may make viewers more likely to draw certain conclusions from the image.

Based on our data, exposure to proof-writing could make certain conclusions more likely in at least three ways (not mutually exclusive). First, it could be that some aspect of the task reminds PT participants of the specific proofmethod of mathematical induction (indeed, $75 \%$ of PT participants mentioned mathematical induction at least once during their video and/or interview). These viewers might then recognize that they could use mathematical induction to prove the target statement, and perhaps even complete the proof (as did $25 \%$ of our PT participants). Thus, knowledge of the formal proof could provide an alternate means of acquiring certainty about the conclusion; once achieving this certainty, participants may be more likely to attempt to generate an alternate argument based on the image. However, it cannot be the case that knowledge of mathematical induction is a necessary condition for such an argument, as we observed one participant who was not familiar with mathematical induction produce a rigorous image-based justification of the general theorem.

Second, in addition to gaining familiarity with mathematical induction, training in proof-writing would also expose individuals to a set of general mathematical norms which may lead these participants to demonstrate a higher degree of certainty. All participants who demonstrated understanding of the target statement were asked if they believed the statement to be "true in all cases"; however, the two groups likely interpreted this question differently. For PT participants, "all" (when used in a mathematical context) is a technical term, which by
definition implies the impossibility of counterexamples. PU participants may have been operating with an everyday use of "all", in which the term is considered synonymous to "generally" or "usually" (e.g., "All Californians love the beach."). In this light, the two groups' differing responses to the suggestion of counterexamples may be revealing of their different conceptualizations of the term "all".

Finally, in the practice of writing proofs one learns standard ways of representing general mathematical objects, and these representations may be useful in interpreting visual proofs. We observed that PT participants often invoked the fact that odd numbers are of the form $(2 n+1)$ to explain how they knew that the layers (symmetric legs extending from a single corner dot) would always contain an odd number of dots. Fewer PU participants offered this argument, perhaps because they were not familiar with algebraic representations of parity. Future studies are necessary to determine whether knowledge of algebraic representations of parity allow viewers to exploit structure available in an image.

We also explored how conclusions drawn from the image differed between contexts of justification (Condition A) and discovery (Conditions B/C). We observed only three effects of task context. First, PU participants' ability to demonstrate understanding of the target statement varied significantly between conditions. Specifically, PU participants - while perfectly capable of understanding the target statement when it was provided - were highly unlikely to "discover" the full target statement based only on the image (with only $25 \%$ able to do so in Condition C). These results suggest that, even for most highly-educated viewers, the image must be accompanied by the statement it is intended to prove (or at least a substantial hint, as in Condition B). Next, we observed that PU participants in contexts of full discovery (Condition C) who generated the target statement were more likely to provide pattern-based justifications and more likely to mention the square shape than PU participants who were given the full target statement. This is not surprising, since the target statement was unknown to these participants at the outset of the task and was only discovered if the participant noticed a pattern in the image.

We find it interesting that these three results were the only effects of task context for either group. All participants who demonstrated understanding of the target statement were likely to generalize it to nearby cases, regardless of the justification/discovery context in which they had seen the visual proof. Subsequent resistance to large-magnitude counterexamples - relatively low for PU participants, and high for PT participants - did not vary significantly between task contexts. This suggests that certainty of the conclusion has more to do with the viewer's exposure to mathematical proof-writing than with the justification/discovery context in which the image is viewed. The lack of any effect of task context for PT participants suggests that the sharp distinction between justification and discovery, of such theoretical importance in mathematics, is less prevalent in advanced mathematical practice.

## Conclusion

In this study we investigated the reasoning underlying a visual proof by induction and the nature of the conclusions drawn from the image. A visual proof by induction displays a particular set of cases of a general theorem, yet it also contains structure that could be used to construct a rigorous argument that the theorem is necessarily true for all natural numbers. We found that, while most viewers are willing to generalize the theorem to nearby cases not displayed in the image, viewers who have been exposed to formal proof methods (including mathematical induction) show significantly higher resistance to the suggestion that largemagnitude counterexamples to the theorem are possible, and are significantly more likely to provide a rigorous imagebased argument against counterexamples. For participants without proof-training, conclusions drawn from a visual proof resemble a standard inductive generalization and do not display the level of certainty associated with mathematical proof. These results are consistent between contexts of justification and discovery, indicating that the certainty of conclusions drawn from a visual proof by induction are primarily dependent on the viewer's exposure to proof-writing, rather than the viewing context.

## Acknowledgments

We are grateful to Lillian Chang and Emmanuel Gygi for their assistance on this project.

## References

Brown, J. R. (1997). Proofs and pictures. The British Journal for the Philosophy of Science, 48(2), 161-180.
Brown, J. R. (2010). Philosophy of mathematics: A contemporary introduction to the world of proofs and pictures. Routledge.
Doyle, T., Kutler, L., Miller, R., \& Schueller, A. (2014). Proofs without words and beyond. Convergence. Retrieved from http://www.maa.org/press/periodicals/ convergence/proofs-without-words-and-beyond.
Eliaser, N. M. (1998). What constitutes a mathematical proof? (Doctoral dissertation, Northwestern University).
Folina, J. (1999). Pictures, proofs, and 'mathematical practice': Reply to James Robert Brown. The British Journal for the Philosophy of Science, 50(3).
Jamnik, M. (2001). Mathematical reasoning with diagrams. University of Chicago Press.
Martin, W. G., \& Harel, G. (1989). Proof frames of preservice elementary teachers. Journal for Research in Mathematics Education, 20(1).
Peirce, B. (1881). Linear associative algebra. American Journal of Mathematics, 4(1).
Polya, G. (1954). Induction and Analogy in Mathematics. Princeton, NJ: Princeton University Press.
Rips, L., \& Asmuth, J. (2007). Mathematical Induction and Induction in Mathematics. In A. Feeney \& E. Heit (Eds.), Inductive Reasoning: Experimental, Developmental, and Computational Approaches. Cambridge University Press.

# Interpreting Asymmetries in Speech Perception with Bayesian Inference 

Jie Ren (Jie_Ren@Brown.Edu)<br>Department of Psychiatry and Human Behavior, Brown University, 1011 Veteran Memorial Pkwy, Riverside, RI 02915 USA<br>Joseph L. Austerweil (austerweil@wisc.edu)<br>Department of Psychology, University of Wisconsin - Madison, 1202 W. Johnson Street Madison, WI 53706 USA


#### Abstract

This paper proposes a Bayesian account of asymmetries found in speech perception: In many languages, listeners show greater sensitivity if a non-coronal sound $(/ \mathrm{b} /, / \mathrm{p} / \mathrm{l} / \mathrm{g} /$, $/ \mathrm{k} /$ ) is changed to coronal sounds ( $/ \mathrm{d} /, / \mathrm{t} /$ ) than vice versa. The currently predominant explanation for these asymmetries is that they reflect innate constraints from Universal Grammar. Alternatively, we propose that the asymmetries could simply arise from optimal inference given the statistical properties of different speech categories of the listener's native language. In the framework of Bayesian inference, we examined two statistical parameters of coronal and non-coronal sounds: frequencies of occurrence and variance in articulation. In the languages in which perceptual asymmetries have been found, coronal sounds are either more frequent or more variable than non-coronal sounds. Given such differences, an ideal observer is more likely to perceive a non-coronal speech signal as a coronal segment than vice versa. Thus, the perceptual asymmetries can be explained as a natural consequence of probabilistic inference. The coronal/non-coronal asymmetry is similar to asymmetries observed in many other cognitive domains. Thus, we argue that it is more parsimonious to explain this asymmetry as one of many similar asymmetries found in cognitive processing, rather than a linguisticspecific, innate constraint.


Keywords: speech perceptual asymmetry; Bayesian inference; natural statistics; category variability; nature vs. nurture; domain-general vs. domain-specific

## Introduction

Listeners from a variety of language backgrounds have shown asymmetric sensitivity to different directions of sound changes in speech perception. We focus on one particular asymmetry: consonants with a coronal place of articulation (/d/, /t/, /n/, /1/) are more tolerant to changes or mispronunciations than consonants with non-coronal places of articulation ( $/ \mathrm{g} /, / \mathrm{k} / \mathrm{/} / \mathrm{b} / \mathrm{l} / \mathrm{p} /$ ). For example, the German word for "railway", bahn primed the word for "train" when mispronounced as bahm. However, the word for "tree", baum, did not prime the word for "bush" when mispronounced as baun (Lahiri \& van Coillie, 1999). This indicates that listeners can accept a labial sound as the correct form of a coronal sound but not vice versa. ERP findings corroborate this phenomenon: At temporally early stages of speech perception, German-speaking adults displayed asymmetric discrimination for mispronunciations of familiar words with coronal vs. non-coronal onsets (Friedrich, Lahiri, \& Eulitz, 2008) and internal consonants
(Friedrich, Eulitz, \& Lahiri, 2006; Cornell, Lahiri \& Eulitz, 2013). Similar effects have also been observed with English-speaking (Roberts, Wetterlin \& Lahiri, 2013) and Bengali-speaking adults (Lahiri \& Marslen-Wilson, 1991).

To explain this asymmetric bias in speech perception, the predominant hypotheses have been derived from phonological theories (Kiparsky, 1982) in the framework of Universal Grammar (UG). In particular, the Featurally Underspecified Lexicon (FUL) model (Lahiri \& MarslenWilson, 1991; Lahiri \& Reetz, 2002) suggested that the place of articulation for coronal consonants (/d/ in $\underline{d u c k}$ ) is not stored (underspecified) in phonological representations. Consequently, mispronunciations in the onset of $\underline{d} u c k$, such as $\boldsymbol{g} u c k$, are still compatible with the representation of $d u c k$, and such mispronunciations should have minimal effects on the lexical activation of the word duck. By contrast, the place of articulation for a non-coronal consonant is fully stored (specified) in the phonological representation (the place of articulation of $/ \mathrm{g} /$ in goose is stored as [+velar]), so mispronunciations of the onset of goose, such as doose, will be incompatible with the representation of goose and thus will disrupt lexical activation of the word goose.

Studies with infants and toddlers also support predictions of the FUL model. 6-month-old Dutch-learning infants were habituated to repeated taan or paan tokens and then tested on their ability to discriminate trials in which the stimulus repeated versus trials in which the two stimuli alternated. Whereas infants habituated to paan discriminated the two types of trials, infants habituated to taan did not (Dijkstra \& Fikkert, 2011). The findings were interpreted as support for the FUL model: When the standard of comparison was taan, the place of articulation of the onset /t/ was not specified, so both $\underline{t} a a n$ and paan were compatible with the standard. But when the standard of comparison was paan, place of articulation of the onset / p / was specified, so only paan was compatible with the standard, and paan and taan were discriminable. 4- and 6-month-old Dutch- and Japaneselearning infants were also tested using the same procedure on their discrimination of labial (omba) and coronal (onta) sounds (Tsuji, Mazuka, Christia, \& Fikkert, 2015). Infants habituated to the labial sound omba discriminated the two types of trials, but infants habituated to the coronal sound onta did not, regardless of their language background.

Linguistic hypotheses, such as underspecification, provide one source of explanations for the observed asymmetry in speech perception. Consistent with the FUL model, Fennel
(2007) showed that infants detected a labial-to-coronal switch but failed to detect a coronal-to-labial switch. However, inconsistent with this hypothesis, a follow-up study (Fennell, van der Feest, \& Spring, 2010) showed that 14-month-olds were better able to detect a coronal-to-velar switch than a velar-to-coronal switch. To explain such discrepancy, they analyzed the onset formant frequencies of all the $/ \mathrm{b} /$, $/ \mathrm{d} /$ and $/ \mathrm{g} /$ tokens in their experimental stimuli, and discovered that $/ \mathrm{b} /</ \mathrm{d} /</ \mathrm{g} /$ in acoustic variability. Thus, the authors concluded that the asymmetries they observed might be better explained by acoustic variability than phonological specification.

We examine the same asymmetries in speech perception in the framework of Bayesian inference: Asymmetric perception of coronal and non-coronal places of articulation may arise from differences in the statistical properties of the coronal category and the non-coronal category in competition (e.g. within word minimal pairs). As presented later, coronal consonants as a category are more frequent and/or variable than non-coronal consonants in languages where asymmetric perceptions have been found. These statistical properties yield an asymmetric posterior distribution: Given a speech signal equidistant from a prototypical coronal and noncoronal articulation, the signal is more likely to be a coronal consonant.

## The Model

## Theoretical Overview

Following the tradition of categorical perception of speech sounds by Liberman et. al. (1957), Clayard et. al. (2008) and Feldman, Griffiths, and Morgan (2009), we interpret the asymmetry as the result of statistical inference of speech categories from a noisy speech signal. Listeners utilize available information from a variety of sources to achieve such a goal, including their prior knowledge of native speech categories and the acoustic properties of the speech signal.

A speech category is defined in the model as a distribution over acoustic dimensions. According to the model, when speakers articulate a sound, they first choose a speech category and then articulate a sound exemplar from that category. Each sound exemplar of the speech category varies from one another due to many factors, such as coarticulation, affect, focus and grammatical intonation. Although speaker's articulation over acoustic dimensions is multidimensional, for mathematical simplicity we assume articulations of a speech category can be reduced to a Gaussian distribution over a single acoustic dimension. Thus, the inventory of native speech categories is represented as a set of Gaussian distributions in the model. Different speech categories differ in the location of their means and in how much their articulation varies over the acoustic dimension (variance). In addition, categories may differ in frequency of occurrence with some categories used more frequently than others. The frequency of occurrence of each category is represented by its prior probability.

Listeners assume that the perceived signal was generated by a speech category that is masked by noise, including environmental noise and perceptual errors. The listeners' task is to recover the speech category that is most likely to have produced the speech signal. If there are two categories that could have generated the speech signal, listeners should take into account of both categories by weighing the statistical properties of each category. Suppose that in a hypothetical language, coronal and non-coronal categories have equal variance and equal frequency of occurrence. Then each of the two categories has an equal posterior probability to have generated a speech signal equidistant from the mean of the coronal and non-coronal distributions. However, in real languages, the coronal category is often higher in variances and/or frequencies of occurrence (the tip and blade of the tongue are more flexible and more variable). An ideal observer should take these factors into account, which may result in the posterior probability of the equidistant speech signal to be larger for the coronal than the non-coronal category.

## Mathematical Formulation

Here we formulate a Bayesian model of speech perception. Although we apply the model to the asymmetric perception of coronal and non-coronal consonants, the model may apply to any domain where a person observes a noisy signal from categories, with each category's exemplars being approximately Gaussian distributed ${ }^{1}$.

We consider speech perception as probabilistic inference. Listeners infer the category membership $C_{i}$ of a noisy signal $S$, as denoted by the conditional probability $p(S$ $\left.\mid C_{i}\right)$. We denote $i=1$ for coronal membership and $i=2$ for non-coronal membership. The posterior probability $p\left(C_{1} \mid S\right)$ that an observed noisy signal $S$ is a coronal sound can be obtained by Bayes' Rule:

$$
\begin{equation*}
P\left(C_{1} \mid S\right)=\frac{p\left(S \mid C_{1}\right) P\left(C_{1}\right)}{p\left(S \mid C_{1}\right) P\left(C_{1}\right)+p\left(S \mid C_{2}\right) P\left(C_{2}\right)} \tag{1}
\end{equation*}
$$

$P\left(C_{1}\right)$ in Equation 1 is the prior probability of the coronal category and $p\left(S \mid C_{I}\right)$ is the likelihood of observing stimulus $S$ given it was generated by a coronal category.

Now we derive a closed form solution to the posterior probability that a signal $S$ is coronal, $P\left(C_{I} \mid S\right)$. Suppose that the speaker articulates an exemplar $E$ of the coronal category $C_{l}$ and $E$ is Gaussian distributed with mean $\mu_{c 1}$, the prototype of the coronal category $C_{l}$. Exemplars within a category vary with variance $\sigma_{c l}{ }^{2}$. Therefore,

$$
E \mid C=C_{l} \sim N\left(\mu_{c 1}, \sigma_{c l}^{2}\right)
$$

The speaker's articulation, the speech signal passing through the environment, and the perceptual system of the listener all add noise to the exemplar. These sources of noise

[^183]combined into $\sigma_{S}{ }^{2}$. Therefore, the conditional distribution of $S \mid E$ is
$$
S \mid E \sim N\left(E, \sigma_{S}^{2}\right)
$$
where $\sigma_{S}{ }^{2}$ represents the random noise that is not due to within-category variability ${\sigma_{c l}}^{2}$. Due to conjugacy, $E$ can be marginalized out to form the likelihood $p\left(\mathrm{~S} \mid C_{1}\right)$, which is Gaussian distributed:
\[

$$
\begin{equation*}
S \mid C_{l} \sim N\left(\mu_{c 1}, \sigma_{c l}{ }^{2}+\sigma_{S}^{2}\right) \tag{2}
\end{equation*}
$$

\]

The likelihood's variance is the sum of two components: the category variance ${\sigma_{c l}}^{2}$, and random, environmental, and perceptual noise $\sigma_{s}{ }^{2}$. Plugging in the parameter values into a Normal distribution, Equation 2 can be written as:

$$
\begin{equation*}
p\left(S \mid C_{1}\right)=\frac{1}{2 \pi \sqrt{\sigma_{c 1}^{2}+\sigma_{S}^{2}}} \exp \left\{-\frac{\left(S-\mu_{c 1}\right)^{2}}{2\left(\sigma_{c 1}^{2}+\sigma_{S}^{2}\right)}\right\} \tag{3}
\end{equation*}
$$

Following the same logic, the likelihood of the non-coronal category $p\left(S \mid C_{2}\right)$ is

$$
\begin{equation*}
p\left(S \mid C_{2}\right)=\frac{1}{2 \pi \sqrt{\sigma_{c 2}^{2}+\sigma_{S}^{2}}} \exp \left\{-\frac{\left(S-\mu_{c 2}\right)^{2}}{2\left(\sigma_{c 2}^{2}+\sigma_{S}^{2}\right)}\right\} \tag{4}
\end{equation*}
$$

Plugging Equations 3 and 4 into Bayes Rule (Equation 1), we can rewrite the posterior probability of the coronal category given the perceived speech signal $S$ as

$$
\begin{equation*}
P\left(C_{1} \mid S\right)=\frac{1}{1+\beta_{1} \exp \left\{\beta_{2} S^{2}+\beta_{3} S+\theta\right\} * \frac{P\left(C_{2}\right)}{P\left(C_{1}\right)}} \tag{5}
\end{equation*}
$$

where
$\beta_{1}=\frac{\sqrt{\sigma_{c 1}^{2}+\sigma_{S}^{2}}}{\sqrt{\sigma_{c 2}^{2}+\sigma_{S}^{2}}}, \quad \beta_{2}=\frac{\left(\sigma_{c 2}^{2}-\sigma_{c 1}^{2}\right)}{2\left(\sigma_{c 2}^{2}+\sigma_{S}^{2}\right) *\left(\sigma_{c 1}^{2}+\sigma_{S}^{2}\right)}$
$\beta_{3}=\frac{-2\left(\mu_{c 1}\left(\sigma_{c 2}^{2}+\sigma_{S}^{2}\right)-\mu_{c 2}\left(\sigma_{c 1}^{2}+\sigma_{S}^{2}\right)\right)}{2\left(\sigma_{c 2}^{2}+\sigma_{S}^{2}\right) *\left(\sigma_{c 1}^{2}+\sigma_{S}^{2}\right)}$
The closed form solution for the posterior is given by Equation 5. We explore how the relative differences in variability and frequency between coronals and noncoronals affect the posterior probability of coronals $P\left(C_{1} \mid S\right)$. Then, we analyze the natural statistics of coronals and non-coronals to determine whether perceptual asymmetries would arise from them in an ideal observer.

## Quantitative Evaluation

Suppose that in a hypothetical language, the coronal category and the non-coronal category are equally frequent - have equal priors, $P\left(C_{1}\right)=P\left(C_{2}\right)$. Also suppose that the categories are equally variable, as encoded by $\sigma_{c l}{ }^{2}=\sigma_{c 2}{ }^{2}$. In these circumstances, Figure 1 depicts the posterior probability for a noisy speech signal to be perceived as a coronal sound, $P\left(C_{l} \mid \mathrm{S}\right)$.

Given equal variance and equal frequency of occurrence, the category boundary divides the perceptual space into two equal parts. This indicates that a noisy signal equidistant
from the category prototypes has an equal probability of being perceived as a coronal or a non-coronal. We now examine how heterogeneity of category variances (i.e. if $\sigma_{c l}{ }^{2}$ $\neq \sigma_{c 2}{ }^{2}$ ) and unequal frequency (i.e., if $P\left(C_{1}\right) \neq P\left(C_{2}\right)$ ) affect the posterior probability of the coronal category $P\left(C_{I} \mid S\right)$.


Figure 1: Posterior Probability (category boundary) of the coronal category given equal variance and equal frequency

Effect of Category Variance In many languages where perceptual asymmetries are found, exemplars of the coronal category are more variable than exemplars of the noncoronal category. For example, using a modified Levenshtein distance metric, Cohen-Priva (2012) aligned the underlying (dictionary) forms and phonetic realizations in the Buckeye Natural Speech Corpus (Pitt, Johnson, Raymond, Hume \& Fosler-Lussier, 2007). He created an articulatory confusion matrix for English segments in the corpus. Of the 43,915 coronal stop tokens, 21,576 (49\%) were pronounced either as allophonic variants or as some other phonemes, whereas of the 64,288 noncoronal stop tokens, only 2,997 ( $5 \%$ ) were pronounced as allophonic variants or as an alternative phoneme. Such analyses indicate that coronal stops are about 10 times more variable than noncoronal stops. Moreover, coronals (9\% of all coronal segments; $20 \%$ of coronal stops were deleted) were also more likely to be deleted than noncoronals ( $5 \%$ of all noncoronal segments, $4 \%$ of noncoronal stops were deleted).

The differences in natural language statistics of the within-category variances between coronal and non-coronal categories in the Buckeye corpus provide the following constraint: $\sigma_{c l}{ }^{2}>\sigma_{c 2}{ }^{2}$. Suppose that $\sigma_{c l}{ }^{2}=5 \sigma_{c 2}{ }^{2}$ (approximately the difference in the segment deletion rates for the English data in the Buckeye corpus), the posterior probability for a noisy speech signal to be perceived as a coronal sound is displayed in Figure 2 (with the case where the within-category variances are equal for reference).

As the red dashed curve shows in Figure 2, the category boundary has shifted towards the non-coronal category, leaving a larger posterior probability for a noisy signal equidistant between the categories to be perceived as a coronal sound. Suppose that the posterior probability for the speech signal [g] to be perceived as a coronal sound /d/ is 0.1 given equal variance (blue curve). The shift of category boundary (dashed red line) leads to an increase in the posterior probability for [g] to be perceived as /d/ (to a value around 0.2 ). Thus, due to the higher variance of the coronal
category, an ideal listener is more likely to perceive $/ \mathrm{g} /$ as an exemplar of the /d/ category. Now we examine the reversed direction, i.e. a coronal ([d]) signal is changed to a noncoronal sound (/g/) given unequal variance.


Figure 2: Posterior probability (category boundary) of the coronal category as a result of unequal variance.


Figure 3: Posterior probability (category boundary) of the non-coronal category as a result of unequal variance.

Figure 3 shows the posterior probability for a speech signal to be perceived as non-coronal. The blue curve in Figure 3 depicts the posterior probability given equal within-category variances. The red dashed curve shows the posterior probability given that the coronal category has larger variance. As the red dashed curve in Figure 3 shows, the category boundary has shifted towards the non-coronal category, producing a smaller posterior probability for a noisy signal to be perceived as a non-coronal sound. Suppose that the posterior probability for the speech signal [d] to be perceived as a non-coronal sound /g/ is 0.1 given equal within-category variances (blue curve). The shift of category boundary (red dashed curve) leads to a decrease in the posterior probability for [d] to be perceived as $/ \mathrm{g} /$ (to a value of approximately 0 ).

To summarize Figures 2 and 3, increasing the variance of the coronal category causes an ideal listener to be more likely to perceive a non-coronal signal ( $[\mathrm{g}]$ ) as an exemplar of the coronal category ( $/ \mathrm{d} /$ ), and less likely to perceive a coronal signal ([d]) as an exemplar of the non-coronal category (/g/). [g] can be a /d/but [d] cannot be a $/ \mathrm{g} /$.

Effect of Frequency of Occurrence Coronals also occur more often in natural speech than non-coronals. Table 1 (adapted from Ren, Cohen-Priva \& Morgan, under review) shows the frequencies of occurrence of the coronal category in three languages from typologically distinctive families.

Coronal segments in these languages are at least twice as frequent as either labial or velar segments (Japanese velar stops are exceptional and we will discuss this case later).

Frequency is represented by prior probabilities in the model. $P\left(C_{1}\right)$ and $P\left(C_{2}\right)$ are the prior probabilities of the coronal category and the non-coronal category, respectively. Suppose that $P\left(C_{1}\right)=2 P\left(C_{2}\right)$ (approximately the relative frequency in Table 1).

Table 1: Frequencies of segments in CALLHOME transcripts by place of articulation

|  | Consonant |  | Labial | Coronal | Velar |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Language | Segments |  |  |  |  |
| Arabic | All |  | 91,409 | 222,774 | 94,624 |
|  | Stops |  | 25,592 | 54,279 | 38,544 |
| Japanese | All |  | 57,513 | 236,813 | 99,760 |
|  | Stops |  | 15,854 | 62,241 | 79,117 |
| Spanish | All |  | 101,717 | 320,167 | 53,483 |
|  | Stops | 44,366 | 62,961 | 43,005 |  |

Figures 4 and 5 show the posterior probability of the coronal category $P\left(C_{l} \mid S\right)$ and the non-coronal category $P\left(C_{2} \mid S\right)$, respectively. The category boundary (green dashed curve) has shifted towards the non-coronal category due to the larger prior probability of the coronal category. For comparison, the posterior probability given equal prior probabilities of coronals and non-coronals is plotted as the blue curve. This results in a larger posterior probability for a noisy non-coronal signal ([g]) to be perceived as a coronal sound (/d/) (Figure 4) and a smaller posterior probability for a noisy coronal signal ([d]) to be perceived as a non-coronal sound (/g/) (Figure 5).


Figure 4: Posterior probability (category boundary) of the coronal category as a result of unequal frequency


Figure 5: Posterior probability (category boundary) of the non-coronal category as a result of unequal frequency

Thus, similar to the effect of variance, larger frequency of occurrence of the coronal category also causes an ideal listener to be more likely to accept that the noisy signal [g] is an exemplar of the /d/ category but not vice versa, predicting the same pattern of asymmetry.

Japanese velar sounds provides an interesting test for an ideal listener model. As Table 1 shows, velar stops (/k/ and $/ \mathrm{k}^{\mathrm{y}} /$ ) occur more often than coronal stops (/t/) in Japanese. Our model predicts that assuming equal variance, the pattern of asymmetries should be reversed for Japanese listenersthey should be less sensitive to sound changes from noncoronal to coronal than vice versa. Japanese studies (see Tsuji et. al, 2015) so far have only tested infant listeners with labial (omba) and coronal (onta) but not velar (/k/ and $/ \mathrm{k}^{\mathrm{y}} /$ ) phonemes. Future experimental studies should examine this prediction with these velar and coronal phonemes.

Effect of Variance and Frequency In many languages both the prior distribution and the variance of the coronal category are larger than those of the non-coronal category. Here we examine how prior and variance interact.


Figure 6: Posterior probabilities of the coronal category as a result of unequal frequency and/or unequal variance


Figure 7: Posterior probabilities of the non-coronal category as a result of unequal frequency and/or unequal variance

Figure 6 shows the posterior probabilities of acoustic signals to be perceived as a coronal sound and Figure 7 shows the posterior probabilities of acoustic signals to be perceived as a non-coronal sound with differing assumptions regarding the relative frequency and variance of coronal and non-coronal sounds. The grey line shows category boundary shift as the result of both the larger prior and variance of the coronal category. As shown in both figures, the pattern of asymmetry remains the same, but there is an even larger posterior probability for a noncoronal signal to be perceived as a coronal (Figure 6), and
an even smaller posterior probability obtained for a noncoronal signal to be perceived as a coronal sound (Figure 7). Thus, with larger variance and larger frequency of occurrence, an ideal listener is even more likely to perceive a non-coronal signal ( $[\mathrm{g}]$ ) as a coronal sound (/d/), but even less likely to perceive a coronal signal ([d]) as a non-coronal sound (/g/).

## General Discussion

We presented an alternative account for the asymmetry in perceiving coronal and non-coronal consonants in speech processing: They arise due to Bayesian inference given the natural statistics of coronals and non-coronals. Listeners make use of their represented category frequency and variance to make inference about the category membership of a perceived signal. Asymmetry occurs when the two speech categories in competition (e.g. within a word minimal pair) are not equal in variance and/or frequency of occurrence.

Our approach diverges from the currently predominant approach in linguistics, which explains the asymmetric perception as due to underspecification of the coronal place of articulation. This theory relies on the special phonological status of coronal sounds only. Conversely, our approach accounts for the asymmetry as due to the relative statistical properties of different speech categories. The underspecification hypothesis is a language-specific, innate constraint, whereas our account is experience- (learning-) based and domain-general. For example, Quinn, Eimas \& Rosencrantz (1993) found that 4-month-olds habituated to pictures of cat faces could easily detect a change to a picture of a dog face. However, infants habituated to dog faces failed to detect a change to a cat face. A series of follow-up studies investigating this asymmetry confirmed that dog stimuli are more variable in appearance and that when variability was equated across categories the asymmetry disappeared (Eimas, Quinn \& Cowan, 1994; Mareschal, French \& Quinn, 2000). In music perception, Delbé, Bigand \& French (2006) examined effects of variability by training non-musician adults with two distributions of pitch sequences differing in variability and then testing them on sensitivities to the two directions of changes. Results indicated that changes from the less variable category to the more variable category are more detectable than vice versa.

Our account derives predictions to test in future work. First, category frequencies and within-category variances are learned from early language exposure. In languages where non-coronals are more frequent and/or vary more within-category, the model predicts that the asymmetry should be reversed. Second, at any developmental stage when listeners have stable representations of the corresponding frequency of occurrence and variance of two competing categories, perceptual asymmetries may occur in speech processing. Third, frequency and variance should have independent effects on speech processing. Further, for mathematical simplicity we assumed that a speech category is a Gaussian distribution over a single acoustic dimension.

Thus, the model does not differentiate between different sources/dimensions of variability (e.g., contextual effect, Ganong, 1980). Assuming we can control these factors experimentally and test the posterior probability of coronal and non-coronals in a fine-grained manner, the model makes quantitative predictions as to the precise form of the asymmetry. None of these predictions arises from UG but from the statistics of speech input exposure.

It is also worth noting that not all asymmetries in speech or other cognitive domains are caused by category natural statistics. Findings on vowel (Polka \& Bohn, 2003) and face (Corneille, Goldstone, Queller \& Potter, 2006) perception, for example, have suggested that similar asymmetric patterns could also be due to stimulus saliency and experimental training. Future studies may examine how these factors interact with frequency of occurrence and variances in category perception.

In conclusion, we have provided a novel explanation for the asymmetry between coronal and non-coronal sounds in speech perception. Whereas phonological specification as a hypothesis could be useful for linguistic purposes, it is not necessary to account for asymmetries in speech perception.

## References

Clayards, M., Tanenhaus, M. K., Aslin, R. N., \& Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108, 804-809.
Corneille, O., Goldstone, R. L., Queller, S., \& Potter, T. (2006). Asymmetries in categorization, cerceptual discrimination, and visual search for reference and nonreference exemplars. Memory \& Cognition, 34, 556-567.
Cornell, S. A., Lahiri, A., \& Eulitz, C. (2011). "What you encode is not necessarily what you store": Evidence for sparse feature representations from mismatch negativity. Brain Research, 1394, 79-89.
Delbé, C., Bigand, E., French, R. M. (2006). Asymmetric categorization in the sequential auditory domain. Proceedings of the 28th Annual Cognitive Science Society Conference.
Dijkstra N. \& Fikkert P. (2011). Universal constraints on the discrimination of place of articulation? Asymmetries in the discrimination of paan and taan by 6-month-old Dutch infants. In N. Danis, K. Mesh, \& H. Sung (Eds.), Proceedings of the 35th Annual Boston University Conference on Language Development (pp. 172-180). Somerville, MA: Cascadilla Press.
Eimas, P., Quinn, P., \& Cowan, P. (1994). Development of exclusivity in perceptually-based categories of young infants. Journal of Experimental Child Psychology, 58, 418-431.
Feldman, N. H., Griffiths, T. L., \& Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. Psychological Review, 116, 752-82.
Fennell C. T. (2007). Asymmetries in infants' ability to notice mispronunciations. Poster presented at The

Conference of the Canadian Society for Brain, Behaviour, and Cognitive Science. June 2007, Victoria, Canada.
Fennell C. T., van der Feest S. V. H. \& Spring M. (2010). Perceptual asymmetries of consonants at 14-months: Implications for phonological acquisition. Presented at International Conference on Infant Studies, Baltimore, Maryland, March, 2010.
Friedrich C. K., Eulitz C., \& Lahiri A. (2006). Not every pseudo word disrupts word recognition: An ERP study. Behavioral and Brain Functions, 2, 36-45.
Friedrich C. K., Lahiri A., \& Eulitz C. (2008). Neurophysiological evidence for underspecified lexical representations: asymmetries with word initial variations. Journal of Experimental Psychology: Human Perception and Performance, 34, 1545-1559.
Ganong, W. F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.
Kiparsky, P. (1982). From cyclic phonology to lexical phonology. The structure of phonological representations, 1, 131-175.
Lahiri, A., \& Marslen-Wilson, W. D. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. Cognition, 38, 245-294.
Lahiri, A., \& Reetz, H. (2002). Underspecified recognition. In C. Gussenhoven \& N. Warner (Eds.), Laboratory phonology 7 (pp. 637-676). Berlin: Mouton.
Lahiri, A., van Coillie, S. (1999). Non-mismatching features in language comprehension. Unpublished manuscript, University of Konstanz.
Liberman, A. M., Harris, K. S., Hoffman, H. S., \& Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal ofexperimental psychology, 54(5), 358-368.
Mareschal, D. French, R. M., \& Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. Developmental Psychology, 36, 635645.

Polka, L., \& Bohn, O. S. (2003). Asymmetries in vowel perception. Speech Communication, 41, 221-231.
Quinn, P., Eimas, P., \& Rosenkrantz, S. (1993). Evidence for representations of perceptually similar natural categories by 3- and 4-month-old infants. Perception, 22, 463-475.
Ren, J., Cohen-Priva, U. \& Morgan, J. L. (under editorial review). Underspecification in Toddler's and Adult's lexical representations.
Roberts, A. C., Wetterlin, A., \& Lahiri, A. (2013). Aligning mispronounced words to meaning: Evidence from ERP and reaction time studies. The Mental Lexicon, 8, 140163.

Tsuji, Sho, Reiko Mazuka, Alejandrina Cristia, and Paula Fikkert (2015). "Even at 4 months, a labial is a good enough coronal, but not vice versa." Cognition 134: 252-256.

# Quantitative Models of Human-Human Conversational Grounding Processes 

Clayton D. Rothwell (crothwell@infoscitex.com)<br>Infoscitex Corp., 4027 Colonel Glenn Hwy<br>Wright State U. Department of Psychology, 3640 Colonel Glenn Hwy Dayton, OH 45435 USA<br>Valerie L. Shalin (valerie.shalin @wright.edu)<br>Wright State U. Department of Psychology and Kno.e.sis, 3640 Colonel Glenn Hwy Dayton, OH 45435 USA<br>Griffin D. Romigh (griffin.romigh@us.af.mil)<br>Air Force Research Laboratory, 2610 Seventh St<br>Wright-Patterson AFB, OH 45433


#### Abstract

Natural language dialogue between multiple participants requires conversational grounding, a process whereby interlocutors achieve a shared understanding. However, the mechanisms involved in the grounding process are under dispute. Two prominent models of dialogue between multiple participants are: interactive alignment, a simpler model that relies on automatic priming processes within individuals, and interpersonal synergy, a more complicated model emphasizing coordinated interaction across participants. Using recurrence analysis methods, Fusaroli and Tylén (2016) simultaneously evaluated both models and showed that alignment is an insufficient explanation for grounding or for the teams' task performance. However, their task and resulting dialogues lack the typical complexity of conversations or teamwork. Furthermore, the interpersonal synergy model was not clearly differentiated from other coordination-focused models of grounding with explicit foundations in strategy and intentionality (i.e., audience design, joint activity, perspective taking). Here we test recurrence-based models in a collaborative task that stressed the grounding process. Results support a coordination model of dialogue over the alignment model as a predictor of performance. Content-based mediation analyses showed that the coordination recurrence model includes critical aspects of strategic design and is not purely interpersonal synergy.


Keywords: Communication; Dialogue; Conversational Grounding; Multi-person Cognitive Models

## Introduction

The grounding process is a key focus of human dialogue research (Clark \& Wilkes-Gibbs, 1986; Clark \& Brennan, 1991; Branigan, Pickering, \& Cleland, 2000; Pickering \& Garrod, 2004). Conversational grounding is the process whereby interlocutors determine that they have understood one another, and results in additions to shared knowledge and understanding. Grounding underlies successful collaboration by developing a shared context, supporting immediate feedback of actions, and allowing for incremental progress in conveying intent (Brennan, 1998). Grounding processes influence communication effectiveness and resulting performance metrics such as laboratory task completion time (Clark \& Wilkes-Gibbs, 1986; Clark \& Krych, 2004; Reitter \& Moore, 2014), due in part to the requirements for shared understanding in collaborative tasks.

The proposed models of dialogue use multiple conversational participants as the unit of analysis, but suggest differ-
ent mechanisms for grounding processes (Horton \& Gerrig, 2005; Louwerse, Dale, Bard, \& Jeuniax, 2012; Schober \& Brennan, 2003). One prominent model is interactive alignment (Pickering \& Garrod, 2004). Alignment (also entrainment, convergence or imitation) refers to the increasing similarity of the interlocutors through adoption of each other's phonetic, prosodic, lexical, or syntactic content (Branigan et al., 2000). Alignment credits this to priming, an automatic, covert mechanism by which recent experiences influence the likelihood of future contributions. Alignment at lower linguistic levels presumably propagates to the semantic level and the situation model of the interlocutors, which forms the basis of a shared understanding of each other and of the world. Thus alignment provides an appealing, conceptually straightforward explanation of grounding phenomena.

Some researchers question the sufficiency of alignment to explain common ground and grounding of new material. The prominent alternative models of human grounding processes emphasize coordination and complementarity such as adjacency pairs in the interaction, rather than similarity. Coordination models separate into two categories: strategic design and interpersonal synergy. In strategic design, speakers appear to design utterances in light of their audience's knowledge. The knowledge may concern the audience's culture, group membership, spatial perspective, or previous conversational interactions (Clark \& Marshall, 1981; Schober, 1993). Strategic design is marked by intentionality-goaldirected conversational behavior that seeks and displays evidence of understanding. These goals invoke an additional layer of exchange concerning the collaborative management of the conversation, called Track 2 dialogue (Clark, 1996). Track 2 dialogue includes: acknowledgements of understanding, displays of non-understanding, and requests for clarification. Interpersonal synergy is a recent and relatively less examined theory. The coordination from interpersonal synergy either does not require intentionality (Fusaroli, RaczaszekLeonardi, \& Tylén, 2014), or redefines it (see also Gallagher \& Miyahara, 2012). Interlocutor coordination emerges from a complex dynamical system achieving stability in a specific context, and becomes cemented in interaction routines. The
introduction of new interlocutors into established interaction routines disrupts communication (Fusaroli et al., 2014).

## Quantification of Recurrence

Separate bodies of research have investigated interactive alignment and coordination models, but to the authors' knowledge only one study has attempted to examine the two theories with competing models for the same performance data (Fusaroli \& Tylén, 2016). Recent advancements in the analysis of dyadic dialogue utilize the non-linear analysis methods of recurrence quantification analysis (RQA) and cross recurrence quantification analysis (CRQA). RQA and CRQA originated from the study of dynamic systems and were developed to examine recurrence in chaotic systems, i.e., repetition of states in time series data. RQA seeks recurrence within one time series (analogous to autocorrelation) and CRQA seeks recurrence between two time series (analogous to cross-correlation). These methods reveal and quantify order and organization that is not readily apparent. Originally built for continuous data, these methods have been adapted for categorical data and used in the analysis of lexical content (e.g., Orsucci et al., 2013) and syntax (e.g., Dale \& Spivey, 2006).

## Grounding Process Models

Fusaroli and Tylén (2016) created models for alignment and coordination using recurrence analyses, and discriminated between these models by their relationship to task performance. They argued that the coordination recurrence model specifically represented interpersonal synergy, though strategic design is an unexamined alternative. Their approach is illustrated in Figure 1. The same time series contents appear in each panel but different outlined patterns reflect the different recurrence sensitivities. The alignment model detects patterns transferred from one speaker to the other, such as 'XYY' from speaker A to speaker B (though not illustrated, patterns that go from B to A will also be detected). The coordination model detects speaker-independent patterns, including patterns across speakers such as adjacency pairs. For instance, the pattern 'YXZXY' occurs between A and B and later B and A. In the self-consistency baseline model, recurrence of patterns within A and within B were tested separately selecting for analysis whoever had the higher recurrence rate.

The dialogue in Fusaroli and Tylén (2016) resulted from two participants performing a visual detection task. Each participant made an independent judgment of whether the target signal appeared in the first interval or the second interval of the stimulus. Dialogue only occurred when their judgment disagreed-they discussed the stimulus and came to a collaborative judgment. Collaborative benefit was computed as the ratio between joint performance and the highest individual's performance, where ratio values greater than 1 indicated a benefit from the joint decision. Recurrence values for lexical choice, pauses, and prosody were calculated according to each theory and then used as predictors in separate regression
models with collaborative benefit as the outcome, thereby relating each grounding model to task performance.

Both the alignment and coordination models were related to task performance, but coordination was a better predictor of performance for the lexical level and the speech/pause level. The two models were similar for the prosodic level.

This quantitative approach provides a promising beginning to the direct comparison between grounding models. However, the task and dialogue content was very limited. The task used simple visual psychophysics stimuli and required a simple choice between two intervals. The vocabulary and conceptualizations that appeared in the dialogues, though not reported, were most likely very limited. The importance of these results for more complex dialogues in a more complicated task setting was not established. In addition, the analysis failed to differentiate between interpersonal synergy and strategic design. Although similar in their emphasis on coordination, these models maintain important distinctions regarding the characterization of cognitive mechanisms. Intentionality (and goal-directed behavior) is one way to differentiate between the two models but their correspondence to the coordination recurrence model is not intuitive. Additional analyses must distinguish between synergy on the one hand and design and intentionality on the other.

## Current Study

The current study examined a team task that stressed the grounding process and applied the RQA and CRQA models for coordination and alignment on two lexical levels: the morpheme level used in Fusaroli and Tylén (2016) and the word level. As discussed below, word-level analysis facilitated an additional mediation analysis of recurrence model results. The task and resulting team dialogue resulted in rich, long dialogue with numerous and diverse content to stress grounding processes. Consistent with Fusaroli and Tylén, we hypothesized that the recurrence metrics calculated based on the coordination model would have a stronger relationship to performance than the alignment model. In addition, we sought to investigate what the coordination recurrence model is measuring, and its relationship to the strategic design model of grounding. We created a lexicon of Track 2 dialogue (described below) and tested if the coordination model statistically mediates the relationship between Track 2 dialogue and performance. Mediation can identify a process that underlies an observed relationship (Baron \& Kenny, 1986). We used mediation to see if the variance in performance explained by Track 2 dialogue is reduced by the coordination model. Such mediation demonstrates that the recurrence model for coordination captures aspects of strategic design in addition to, or possibly instead of, interpersonal synergy.

## Methods

## Uncertainty Elicitation Task Corpus

We used materials from the Uncertainty Elicitation Task corpus (Romigh, Rothwell, Greenwell, \& Newman, 2016).


Figure 1: Illustration of the recurrence tests for alignment, coordination, and baseline (adapted from Fusaroli \& Tylén, 2016). Alignment models were sensitive to patterns transferred between speakers. Coordination models were sensitive to patterns independent of speaker, which included patterns across speakers as illustrated here. Baseline models were sensitive to patterns within one speaker (i.e., self-consistency). (Figure used with permission from John Wiley and Sons).

Like Fusaroli and Tylén (2016), this was a symmetric dialogue task-no one speaker had the answer, so the conversational dynamics were flexible and negotiable. Partners had many unlabeled pictures of various real world scenes from both an overhead perspective and street-level perspectives, that they had to match with each other. This led to conceptually complex and diverse dialogues. Partners discussed: house features (e.g., siding, roof, windows, garage, porch, columns, 1 or 2-story), lot features (e.g., trees, yard, fence, driveway, garden, sidewalk, corner lot, playground, pool), street/neighborhood features (e.g., presence of stopsign, power lines, presence of alleyways, nearby parks), and car features (e.g., number of cars, type of vehicle: truck/van/sedan, color).

On each trial, two people sat in separate rooms and worked together to locate street-level pictures of different houses on an overhead map (Figure 2). Street-level images and satellite images were obtained from Google Maps with labels (e.g., street names) removed. The overhead map was the same for both participants and had 12 numbered buildings (1-12). The participants each had street-level pictures of 6 of those buildings on the right hand side of their screens. The participants were given street-level images from different points of view and they had to determine that they were discussing the same building. Their task was to relate the street-level views to the overhead map by labeling the street-level with a number $1-12$, and the trial ended when all street-level images had correct number labels. ${ }^{1}$ As accuracy was held constant, completion time was the performance metric (shorter times indicated better performance). Performance in the task was expected to be related to conversational grounding because participants needed to communicate effectively-make definite references to unlabeled street-level views of houses, share the information from their unique street-level views, and discuss the similarity between street-level and overhead imagery. Five teams of 2 people and each team completed 8 trials for a total of 40 trials. ${ }^{2}$

[^184]

Figure 2: Screen shot from the Uncertainty Elicitation Task. Building numbers appear on the overhead map and participants labeled street-level images using the drop-down boxes centered on each row of images.

## Recurrence Analyses

Our analysis examines recurrence at two lexical levels (the word level and the morpheme level) in search of the model that best predicts task-specific performance.

Prior to calculating the recurrence plot, RQA and CRQA require a number of parameters. We used values keeping with Fusaroli and Tylén (2016) and other categorical analyses of transcript data (Orsucci et al., 2013). The radius value was set to 0 , meaning only an exact match would be counted as a recurrence, which is appropriate for nominal data. For example, for the word-level analysis each word was given an arbitrary unique numerical identifier. The threshold for a line (i.e., recurrence patterns that are parallel to the positive diagonal) was set at 2 . Time delay was set to 1 . The word-level analysis used single words as the unit of analysis, which was specified by an embed value of 1 . The morpheme level ${ }^{3}$ used a 3-letter unit of analysis (i.e., a letter trigram), which was specified by an embed value of 3 .

[^185]The three models in Figure 1 were tested. The alignment model was represented by CRQA of a time series of Speaker A with a time series of Speaker B. To preserve the time sequence and phase information of the entire dialogue, added codes in each time series replaced the other speaker's contributions. The coordination model was represented by RQA of the time series for the entire block (Speakers A \& B). A baseline self-consistency model was represented by performing RQA of each speaker's time series with his/herself and using the recurrence plot with the highest recurrence rate.

The three separate recurrence models output separate recurrence metrics for different regression models, in order to assess the relationship of each recurrence model to task performance. This analysis process differs from a typical regression procedure where predictors are added or removed from a single regression model. Here, three different regression models with the same four predictors were based on different recurrence calculations. Individual recurrence metrics of recurrence rate, determinism, average line length, and line entropy were calculated for each of the three models (i.e., alignment, coordination, baseline), for each trial in the Uncertainty Elicitation Task corpus. The recurrence metrics quantify how much recurrence occurs (Recurrence Rate, $R R$ ), the proportion of recurrence that appears in longer sequences (Determinism, DET), the average length of recurrence sequences (Line Length, L), and the variety in recurrence lengths (Line Entropy, ENT). Recurrence metrics then functioned as predictors for a linear regression model of the performance scores (i.e., completion times) for each trial. Linear models were evaluated using $A d j R^{2}$ values. These models follow the analysis procedures from Fusaroli and Tylén (2016) assuming data from a between-subjects design. Subsequent tests address the repeated measures (i.e., within team) nature of our data. Analyses were performed in R, using the crqa package (Coco \& Dale, 2014).

We also created a lexicon (i.e., word list) for Track 2 dialogue using the Linguistic Inquiry and Word Count 2015 (LIWC) text analysis program (Pennebaker, Boyd, Jordan, \& Blackburn, 2015). Our analysis relied on two separate lexicons that may capture Track 2 issues of dialogue management: Assent (e.g., agree, OK, yes) and Certainty (e.g., indeed, always, never). We reasoned that Assent may capture an addressee's acceptance of a speaker's installment and Certainty may capture confusion regarding an installment. We tested how well the LIWC categories accounted for performance by using the LIWC counts (frequencies of words in the Assent and Certainty lists) for each trial as predictors of performance. We then tested if the parameters from the coordination recurrence model statistically mediated the relationship between the LIWC categories and performance, following the 'Causal Steps' procedure (Baron \& Kenny, 1986). This involved three "Steps" where the LIWC categories were treated as independent variables (IVs) and the recurrence parameters were treated as mediators (Ms): 1) the IVs and performance, 2) the IVs and the Ms, and 3) the (IVs +Ms ) and
performance. Multiple linear regression was used for Steps 1 and 3 while MANOVA was used for Step 2 in order to test for a relationship between multiple LIWC categories and multiple recurrence-parameter mediators.

## Results

First we show that the observed recurrence was not due to chance. Next we show that the coordination model has stronger relationships to task performance than the alignment model for both the word-level and the morpheme-level analyses. Moreover, the coordination model accounts for variance in performance after controlling for team differences whereas the alignment model does not. Mediation analysis shows that the coordination model reflects aspects of Track 2 dialogue.

## Chance Analysis

The structure of recurrence represented by these metrics was not due to chance. We compared the outputs of the recurrence analyses of the data to outputs using a shuffled time series (i.e., randomly ordering the words in the time series).

Paired $t$-tests indicated that recurrence structure is significantly different from shuffled controls for all models for all values (all $=p<.0001$ ), except for word-level recurrence rate. For the word-level test, the recurrence rates are exactly the same because shuffling does not add or remove words.

## Word-Level Analyses

The linear regression models for word-level analyses are shown on the left side of Table 1. The coordination model accounted for more of the variance in completion times (AdjR ${ }^{2}$ $=0.66)$ than the alignment model and the baseline model $\left(\operatorname{Adj} R^{2}=0.14 \& 0.51\right.$, respectively $)$. The baseline model accounted for more variance than the alignment model.

## Morpheme-Level Analyses

The linear regression models for morpheme-level analyses are shown on the right side of Table 1. The pattern of results was the same as the word level. The coordination model accounted for more of the variance in performance $\left(\operatorname{Adj} R^{2}=\right.$ 0.76 ) than the alignment model or the baseline model ( $\operatorname{Adj} R^{2}$ $=0.32 \& 0.64$, respectively). The baseline model accounted for more variance than the alignment model as well.

## Controlling for Team Differences

Space precludes a complete presentation, but controlling for team differences was necessary for the within-subjects design and Team ID was a significant predictor of performance ( $F(4$, $\left.35)=6.04, p<.001, \operatorname{Adj} R^{2}=0.34\right)$. Using statistical controls that removed the variance between teams, we tested if each recurrence model could explain the residual variance. The alignment model did not $\left(F(4,31)=0.53, p=.72, \Delta R^{2}\right.$ $=0.04)$ whereas the coordination model $\operatorname{did}(F(4,31)=8.21$, $p<.001, \Delta R^{2}=0.30$ ).

Table 1: Word-level analyses (left) and Morpheme-level analyses (right)—linear regression models for alignment, perspective taking, and baseline. Predictors were recurrence rate (RR), determinism (DET), average line length (L), and line entropy (ENTR). $\left({ }^{*} p<.05, * * p<.01, * * * p<.001\right.$ )

| Theory | Word-Level |  |  |  | Morpheme-Level |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Contents | $p$-value | Contents | $p$-value | Contents | $p$-value | Contents | $p$-value |
| Alignment* | AdjR ${ }^{2}=0.14$ | 0.051 |  |  | $\boldsymbol{A d j} \mathrm{R}^{2}=0.32$ | $<.01$ |  |  |
|  | $\mathrm{RR}_{A}$ | 0.53 | $\mathrm{L}_{A}{ }^{*}$ | $<.05$ | $\mathrm{RR}_{A}$ | 0.21 | $\mathrm{L}_{A}$ * | $<.05$ |
|  | $\mathrm{DET}_{A}$ | 0.13 | $\mathrm{ENTR}_{A}$ ** | $<.01$ | $\mathrm{DET}_{A} * * *$ | < 001 | $\mathrm{ENTR}_{A}$ * | $<.05$ |
| Coordination*** | $\operatorname{Adj}^{2}=\mathbf{0 . 6 6}$ | < . 001 |  |  | $A d j R^{2}=0.76$ | < 0001 |  |  |
|  | RRs** | <. 01 | $\mathrm{L}^{\text {S }}$ *** | $<.001$ | $\mathrm{RR}_{S}$ | 0.26 | $\mathrm{L}^{\text {s }}$ *** | < 0001 |
|  | $\mathrm{DET}_{S}$ | 0.48 | ENTRS | 0.07 | $\mathrm{DET}_{s} * * *$ | < 001 | $\mathrm{ENTR}_{S}$ *** | <. 001 |
| Baseline*** | $\operatorname{Adj}^{2}{ }^{2} \mathbf{0} .51$ | < . 001 |  |  | $A d j R^{2}=0.64$ | < 0001 |  |  |
|  | $\mathrm{RR}_{B}$ | 0.17 | $\mathrm{L}_{B}$ | 0.15 | $\mathrm{RR}_{B}{ }^{* * *}$ | $<.001$ | $\mathrm{L}_{B}{ }^{* * *}$ | < . 001 |
|  | $\mathrm{DET}_{B}{ }^{* *}$ | < . 01 | $\mathrm{ENTR}_{B}$ | 0.52 | $\mathrm{DET}_{B}{ }^{*}$ | <. 05 | ENTR $_{B} * *$ | < 01 |

## Mediation Analysis

Linear regression results for the LIWC categories Assent and Certainty appear at the top of Table 2. These categories significantly predicted task completion times $\left(A d j R^{2}=0.43\right)$. The coefficients for Assent and Certainty were both negative. More instances of these words resulted in faster completion times (i.e., better performance). Additional mediation tests used the word-level coordination model's recurrence parameters. The MANOVA in Step 2 showed that the LIWC lists were related to these recurrence parameters $(F(4,34)=6.62$, $p<0.001$, and $F(4,34)=9.70, p<0.001$, respectively). Step 3 showed that the recurrence parameters mediated the relationship between task completion times and LIWC categories (Table 2 bottom portion) by eliminating their significance.

Table 2: Mediation analysis for LIWC—See text for details. (*p<.05, **p<.01, ***p<.001)

| Step 1- LIWC's relation to performance |  |  |  |
| :---: | :---: | :---: | :---: |
| Assent*** | < . 001 | Certainty** | < . 01 |
| Step 2-LIWC's relation to Coordination |  |  |  |
| Assent*** | < . 001 | Certainty*** | < . 001 |
| Step 3-LIWC's \& Coordination's relation to performance |  |  |  |
| Assent | 0.30 | Certainty | 0.89 |
| $\mathrm{RR}_{S}{ }^{*}$ | <.05 | $\mathrm{L}_{S}{ }^{* *}$ | $<.01$ |
| $\mathrm{DET}_{S}$ | 0.32 | $\mathrm{ENTR}_{S}$ | 0.32 |

## Discussion

Findings clearly supported the coordination model over the alignment model for both levels of analysis. At the word level, coordination accounted for $52 \%$ more of the variance in task completion times than alignment, and $44 \%$ more at the morpheme level. Although the baseline model performed better than Fusaroli and Tylén (2016), the pattern of findings for
alignment and coordination was similar. Moreover, the relationships between coordination and performance found here were larger than those shown by Fusaroli and Tylén, despite the longer, more complex dialogues. While the coordination model accounted for performance above team differences, the alignment model did not.

Recent research agrees with these findings that communication processes are more complicated than priming-based alignment. Rather than repeating content, interlocutors' contributions provide new content that compliments past contributions (Tenbrink, Andonova, \& Coventry, 2008). Many studies of alignment do not include performance outcomes (e.g., Branigan et al., 2000) and therefore may not identify these insufficiencies. Alignment may still occur over longer time scales, which has been shown to predict task performance (Reitter \& Moore, 2014). The alignment recurrence model used here did not distinguish between short-term and long-term alignment, so it is possible that long-term alignment is responsible for the relationship between alignment and performance.

Beyond support for a general coordination model, the coordination recurrence model appears to contain aspects of strategic design. Indeed, Track 2 dialogue alone accounted for more variance in performance than alignment did at both the word level and morpheme level $\left(A d j R^{2}=0.43\right.$ vs. $0.14 \&$ 0.34 , respectively). Recent research supports the importance of design-utterances often reflect different perspectives and interlocutors appear to keep track of multiple perspectives at the same time (Brennan, Schuhmann, \& Batres, 2013).

## Conclusion

In this paper, we quantitatively modeled conversational grounding processes between two interlocutors. We tested two models for this process, alignment and coordination, in a complex collaborative grounding task. The results clearly discount an alignment model as a sufficient model of the conversational grounding process. Results also indicated that the coordination recurrence model is closely related to Track 2
dialogue and therefore strategic design models of the conversational grounding process must be considered. Our future research will examine whether strategic design accounts for these findings in addition to or to the exclusion of interpersonal synergy.

## Acknowledgments

This research was partially supported by the Air Force Research Laboratory's Human Interface Research Technologies Contract \#: FA8650-14-D-6501. We thank W. Sid Horton for valuable discussions and Kno.e.sis for use of the Linguistic Inquiry and Word Count software.

## References

Baron, R. M., \& Kenny, D. A. (1986). The ModeratorMediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology, 51(6), 1173-1182. doi: 10.1037/0022-3514.51.6.1173
Branigan, H. P., Pickering, M. J., \& Cleland, A. A. (2000). Syntactic co-ordination in dialogue. Cognition, 75(2), B13-25. doi: 10.1016/S0010-0277(99)00081-5
Brennan, S. E. (1998). The Grounding Problem in Conversations With and Through Computers. In S. R. Fussell \& R. J. Kreuz (Eds.), Social and cognitive approaches to interpersonal communication (pp. 201-225). Hillsdale, NJ: Lawrence Erlbaum.
Brennan, S. E., Schuhmann, K. S., \& Batres, K. M. (2013). Collaboratively Setting Perspectives and Referring to Locations Across Multiple Contexts. In Proceedings of the PRE-CogSci 2013 Workshop on the Production of Referring Expressions (Vol. 1, pp. 1-6). Berlin, Germany.
Clark, H. H. (1996). Using Language. Cambridge, UK: Cambridge University Press.
Clark, H. H., \& Brennan, S. E. (1991). Grounding in Communication. In L. Resnick, J. Levine, \& S. Teasley (Eds.), Perspectives on socially shared cognition (pp. 127-149). Washington, D.C.: American Psychological Association. doi: 10.1037/10096-006
Clark, H. H., \& Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language, 50(1), 62-81. doi: 10.1016/j.jml.2003.08.004

Clark, H. H., \& Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, \& I. A. Sag (Eds.), Elements of discourse understanding (pp. 10-63). Cambridge, UK: Cambridge University Press.
Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1-39. doi: 10.1016/0010-0277(86)90010-7

Coco, M. I., \& Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: an R package. Frontiers in Psychology, 5(June), 1-14. doi: 10.3389/fpsyg. 2014.00510

Dale, R., \& Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coor-
dination between children and caregivers in conversation. Language Learning, 56(3), 391-430. doi: 10.1111/j.14679922.2006.00372.x

Fusaroli, R., Raczaszek-Leonardi, J., \& Tylén, K. (2014). Dialog as interpersonal synergy. New Ideas in Psychology, 32(1), 147-157. doi: 10.1016/j.newideapsych.2013.03.005

Fusaroli, R., \& Tylén, K. (2016). Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. Cognitive Science, $40(1), 145-171$. doi: 10.1111/cogs. 12251
Gallagher, S., \& Miyahara, K. (2012). Neo-pragmatism and Enactive Intentionality. In J. Schulkin (Ed.), Action, perception and the brain (pp. 117-146). Basingtoke, UK: Palegrave-Macmillan.
Horton, W. S., \& Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. Cognition, 96(2), 127-142. doi: 10.1016/j.cognition.2004.07.001

Louwerse, M. M., Dale, R., Bard, E. G., \& Jeuniax, P. (2012). Behavior Matching in Multimodal Communication Is Synchronized. Cognitive Science, 36(8), 1404-1426. doi: 10.1111/j.1551-6709.2012.01269.x

Orsucci, F., Petrosino, R., Paoloni, G., Canestri, L., Conte, E., Reda, M., \& Fulcheri, M. (2013). Prosody and synchronization in cognitive neuroscience. EPJ Nonlinear Biomedical Physics, 1(1), 1:6. doi: 10.1140/epjnbp13
Pennebaker, J. W., Boyd, R. L., Jordan, K., \& Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015 (Tech. Rep.). Austin, TX: University of Texas at Austin. doi: 10.1068/d010163
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-190; discussion 190-226. doi: 10.1017/S0140525X04000056

Reitter, D., \& Moore, J. D. (2014). Alignment and task success in spoken dialogue. Journal of Memory and Language, 76, 29-46. doi: 10.1016/j.jml.2014.05.008
Romigh, G., Rothwell, C., Greenwell, B., \& Newman, M. (2016). Modeling uncertainty in spontaneous speech: Lexical and acoustic features. The Journal of the Acoustical Society of America, 140(4), 3401-3401. doi: 10.1121/1.4970912

Schober, M. F. (1993). Spatial perspective-taking in conversation. Cognition, 47(1), 1-24. doi: 10.1016/0010-0277(93)90060-9
Schober, M. F., \& Brennan, S. E. (2003). Processes of Interactive Spoken Discourse: The Role of the Partner. In A. C. Graesser, M. A. Gernsbacher, \& S. R. Goldman (Eds.), Handbook of discourse processes (pp. 123-164). Routledge.
Tenbrink, T., Andonova, E., \& Coventry, K. (2008). Negotiating spatial relationships in dialogue: The role of the addressee. In Proceedings of LONDIAL - The 12th SEMDIAL Workshop (pp. 193-200).

# Comprehenders Model the Nature of Noise in the Environment 

Rachel Ryskin<br>Massachusetts Institute of Technology<br>Richard Futrell<br>Massachusetts Institute of Technology, Cambridge, MA, USA<br>Edward Gibson<br>MIT


#### Abstract

Recent work suggests that language understanding is the result of rational inference over a noisy channel. Upon perceiving a sentence, listeners decode the speaker's intended sentence from the prior probability that a speaker would say that sentence and the probability that it would be corrupted to the perceived sentence by noise. Here we examine the listener's noise model. Readers were asked to correct sentences if they thought they contained an error. We manipulated context such that participants corrected exposure sentences containing either deletion, insertion, swap, mixed, or no errors (e.g., swap: A bystander was rescued by the fireman in the time of nick.). Test sentences were syntactically licensed but implausible (e.g., The bat swung the player). On test sentences, participants' corrections differed by exposure condition. This suggests participants track the type of errors that have a higher likelihood and make inferences about the intentions of the speaker accordingly.


# Is the strength of regularisation behaviour uniform across linguistic levels? 

Carmen Saldana (C.C.Saldana@sms.ed.ac.uk), Kenny Smith, Simon Kirby \& Jennifer Culbertson<br>Centre for Language Evolution, School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK


#### Abstract

Human languages contain very little unconditioned variation. In contexts where language learners are exposed to input that contains inconsistencies, they tend to regularise it, either by eliminating competing variants, or conditioning variant use on the context. In the present study we compare regularisation behaviour across linguistic levels, looking at how adult learners respond to variability in morphology and word order. Our results suggest similar strengths in regularisation between linguistic levels given input languages whose complexity is comparable.


Keywords: artificial language learning; statistical learning; regularisation; variation; complexity; morphology; word order

## Introduction

While languages exhibit variation at all linguistic levels, in the form of paraphrases, synonyms, allomorphs and allophones, that variation tends to be predictable: the choice of variant is (at least partially) conditioned by some aspect of the social or linguistic context. Occasionally, language learners are exposed to input that involves inconsistencies, for instance, when new variants are introduced into an established system, or when conventions are still not established, as in emerging languages (Senghas \& Coppola, 2001; Siegel, 2004). Learners under those circumstances tend to reduce or remove such inconsistencies, i.e. they regularise their input. This can be achieved either by removing competing variants, or conditioning variant choice on the context (Ferdinand, Kirby, \& Smith, 2017).

Regularisation has been documented extensively across linguistic levels (i.e. phonology, morphology, syntax and the lexicon) in natural language; e.g. in language acquisition, language change, and in emerging languages (Senghas \& Coppola, 2001; Siegel, 2004; van Trijp, 2013). Experimental studies involving artificial language learning and statistical learning techniques report regularisation behaviour during the learning and production of probabilistic unconditioned variation in different linguistic units, across different linguistic levels (Culbertson, Smolensky, \& Legendre, 2012; Fehér, Wonnacott, \& Smith, 2016; Hudson Kam \& Newport, 2005, 2009; Wonnacott \& Newport, 2005). Nevertheless, it still remains an open question whether regularisation behaviour applies with uniform strength across linguistic levels and to what extent level-specific biases interact with regularisation during language learning and use.

## Level-specific effects in regularisation behaviour

Research in second language acquisition and pidgin and creole studies has highlighted different developmental paths for morphology and syntax cross-linguistically (Good, 2015; Slabakova, 2013). Studies in pidginisation suggest that, in
periods when pidgins are highly inconsistent, linguistic levels might behave differently: morphologically complex traits such as inflectional morphology seem to be highly simplified whilst syntactic traits such as word order tend to reproduce the input complexity more closely (Good, 2015; Siegel, 2004). Good (2015) argues that this asymmetry is given by a break in transmission from source languages for morphological traits, which are only successfully transmitted if an entire contrasting paradigm is available to the learner, which is not the case in periods of linguistic instability. However, word order variation can be contrastive as well (e.g. S-Aux inversion to distinguish illocutionary forces). Alternatively, a more parsimonious hypothesis we could entertain is that a general tendency for pidgins to comprise highly simplified morphological traits and more conservative word order is rooted in the differing complexity of these traits in the source languages; Hudson Kam and Newport (2009) show that learners are more likely to regularise complex systems of variation.

Recent experimental studies have separately explored the effect of learning biases on typological asymmetries found in morphology and word order respectively. In morphology for example, St Clair, Monaghan, and Ramscar (2009) provide evidence of a preference for suffixing over prefixing, mirroring the cross-linguistic preference for suffixing. In word order, Culbertson et al. (2012) show that learners prefer consistent harmonic word order patterns (i.e. all modifiers either pre-nominal or post-nominal), also found more commonly in the world's languages. Moreover, Culbertson et al. (2012) show that this bias leads to different regularisation behaviour for different word order patterns. Nevertheless, no study has hitherto tried to systematically compare regularisation behaviour across linguistic levels. Uncovering differences in regularisation behaviour across linguistic levels could shed light on the intriguing asymmetry found in pidgin languages: morphological paradigms seem to be highly simplified whilst input complexity is more closely reproduced in word order.

In the present study we combine artificial language learning and statistical learning techniques to systematically compare the strength of regularisation of inflectional morphology and word order, controlling for asymmetries in the complexity and variability of the input languages.

## Experiment 1

We utilise the methodology developed in Culbertson et al. (2012); Hudson Kam and Newport (2005). Adult learners are exposed to a miniature artificial language featuring an inconsistent mixture of synonymous variants. We are interested in how learners restructure the probabilistic unconditioned variation in the input languages, and to what extent that
restructuring is comparable across linguistic levels (specifically, morphology and word order).

## Method

Participants Fifty-six native-English speakers (aged between 18 and 41, mean $=23.2$ ) were recruited from the University of Edinburgh's Careers Service database of vacancies. Each was compensated $£ 6$. Twenty-six participants were assigned to the Morphology condition, and 26 to the Word Order condition; the data from a further 4 participants (all in the morphology condition) were excluded as they either failed to learn the noun lexicon or failed to learn the associations between phrases and pictures.

Input languages We designed two novel languages which contained probabilistic unconditioned variation either in morphology or word order. Their respective probabilistic grammars are shown in Table 1. Both languages were used to describe simple pictures featuring one of two objects. Each object appeared either singly or in a pair; and could appear either in greyscale or coloured in blue. Descriptions were noun phrases composed of a Noun plus a Num(eral) and/or Adj(ective) modifier, which were presented orthographically and aurally to participants during the experiment.

All lexical items were 5 graphemes/phonemes long and had a neighbourhood density of 0 in the English lexicon. Nouns and modifiers differed in their syllabic structure; while all were bisyllabic, nouns (i.e. "mokte" and "jelpa") conformed to a CVC.CV pattern, and modifiers to CV.CCV (based on English phonotactics and the Maximal Onset Principle).
Procedure Participants worked through a six-stage training and testing regime.

Stage 1, noun familiarisation Participants were trained on the two bare nouns that corresponded to pictures of the two different objects in the artificial language. During this phase, participants underwent a block of training consisting of 6 exposure trials and 4 picture-selection comprehension trials (in that order) -each noun-picture pair appeared 5 times (order randomised). Common to all training blocks to follow, on each exposure trial participants were presented with a picture (in this block always of a single object in grey-scale) and a corresponding description in the language (in this block, a bare noun), displayed both visually and aurally. On comprehension trials, participants were asked to select a picture out of an array of four (in this stage, the two objects seen during training plus two distractors) that corresponded to the displayed description in the alien language, and received feedback on their accuracy.

Stage 2, one-modifier training In Stage 2 participants were trained on one-modifier NPs, i.e. a Noun plus either Num or Adj only. Pictures contained any of the two objects presented either in blue and singly (Adj only) or in greyscale and in pairs (Num only). For each picture, a variant was selected randomly from the grammar assigned to the participant. Both grammars contained majority variants with an

Table 1: Probabilistic input languages in the Morphology and Word order conditions. Languages contain probabilistic unconditioned variation in inflectional morphology or word order respectively. All morphological variation resides in the suffixation of the modifiers. All word order variants conform to constituent structure [Num [Adj N]]. There are three types of NPs: Num Only (single Num modifier) refer to objects in pairs and in grey-scale, Adj Only (single Adj modifier) refer to a single object coloured in blue, and two-Mod(ifier) NPs (with both Num and Adj modifiers) correspond to objects in pairs coloured in blue. Languages include two different nouns (each corresponding to a different object) and thus comprise a total of 16 NPs ( 8 per noun) that correspond to a total of 6 pictures (1 per NP type, 3 per object).

| NP TYPE | MORPHOLOGY CONDITION | WORD ORDER CONDITION |
| :---: | :---: | :---: |
| NUM Only | $0.6 \mathrm{NP} \rightarrow \mathbf{N}$ nefri | $0.6 \mathrm{NP} \rightarrow \mathbf{N}$ nefri |
|  | $0.4 N P \rightarrow$ N nezno | $0.4 N P \rightarrow$ nefri N |
| ADJ <br> OnLy | $0.6 \mathrm{NP} \rightarrow \mathbf{N}$ kogla | $0.6 N P \rightarrow \mathbf{N}$ kogla |
|  | $0.4 N P \rightarrow \mathrm{~N}$ kospu | $0.4 N P \rightarrow$ kogla N |
| Two MOD | $0.6 \mathrm{NP} \rightarrow \mathbf{N}$ kogla nefri | $0.6 \mathrm{NP} \rightarrow \mathbf{N}$ kogla nefri |
|  | $0.1 \overline{3} N P \rightarrow$ N kogla nezno | $0.1 \overline{3} N P \rightarrow$ nefri kogla N |
|  | $0.1 \overline{3} N P \rightarrow$ N kospu nefri | $0.1 \overline{3} N P \rightarrow$ nefri N kogla |
|  | $0.1 \overline{3} N P \rightarrow \mathrm{~N}$ kospu nezno | $0.1 \overline{3} N P \rightarrow$ kogla N nefri |

empirical probability of $P=0.6$, and minority variants with $P=0.4$, as shown in Table 1. This phase comprised 40 trials in total, divided in 2 blocks of 20 trials; each block consisted of 15 exposure trials followed by 5 picture-selection trials. Participants saw each of the four different one-modifier pictures 5 times per block (order randomised).

Stage 3, one-modifier testing Stage 3 of the experiment tested the participants' knowledge of the language. They saw the same pictures used in Stage 2 without accompanying text or audio and were asked to type in an appropiate description. They had to describe 20 pictures in total; each of the four different one-modifier pictures was presented 5 times in random order.

Stage 4, full training In Stage 4 participants were trained on a mix of one-modifier (a noun plus Adj or Num) and twomodifier NPs (a noun plus both Num and Adj). Two-modifier NPs were used to describe pairs of blue objects. For onemodifier phrases, variants were chosen in the same way as in Stage 2. For two-modifier phrases, variants were also selected randomly from the grammars assigned, with empirical probabilities of $P=0.6$ and $P=0.1 \overline{3}$ for the majority and the three minority variants respectively (see Table 1). This stage comprised 100 trials ( 20 Num Only, 20 Adj Only and 60 two-Mod), divided into 4 block of 25 ( 15 exposure train-
ing trials followed by 10 picture-selection trials). Participants saw each of the four one-modifier pictures 10 times, and each of the two two-modifier pictures 30 times.

Stage 5, full testing Stage 5 tested participants' knowledge of the whole language. They saw all pictures they had been trained on and were asked to type in appropriate descriptions. They had to describe 52 pictures in total: 10 Adj Only (5 per object), 10 Num Only ( 5 per object), 30 two-modifier ( 15 per object), and additionally, 2 pictures of bare objects by themselves and in grey-scale (1 per object).

## Results

Output variability Figure 1 shows the entropy of participants' production systems for both the Morphology and Word Order conditions. Analyses are run on Stage 5's testing exclusively, i.e. participants' final production sets. Words in the productions were corrected for typos (and only typos). Shannon entropy measures how variable participants' productions are; the higher the scores, the more variable and the lower the scores, the more regular. The Shannon entropy (H) of phrase use for participant is given by

$$
\begin{equation*}
H(X)=-\sum_{i=1}^{n} P\left(x_{i}\right) \log _{2} P\left(x_{i}\right) \tag{1}
\end{equation*}
$$

where the sum is over the different variants, and $P\left(x_{i}\right)$ is the empirical probability of variant $x_{i}$ in the set of a participant's productions, $X$. We treated the two nouns for the different objects as the same variant when we calculated the entropy of the phrase variants such that no variability is introduced by the correct use of the different nouns. Entropy lower- and upper- bounds will vary depending on the number of required and possible variants as well as on the number of production trials. The most regular expressive language contains only one-to-one picture-phrase mappings and therefore only three different variants, one Num Only (e.g. N nefri), one Adj Only (e.g. $N$ kogla) and one two-modifier (e.g. $N$ kogla nefri). The final production phase consisted of 50 trials (excluding the two bare noun trials), divided up into 20 one-modifier trials (half Num Only and half Adj Only) and 30 two-modifier trials: the entropy lower bound for the language overall is thus 1.37 bits, and 0 bits for each of the NP types.

Figure 1 shows the entropy scores for the set of all participants' productions (i.e. the overall language), as well as those for the production sets for specific NP types in isolation: onemodifier Num (Num Only), one-modifier Adj (Adj Only), and two-modifier (two-Mod) NPs. Entropy lower bounds and input entropies are represented as solid and dotted vertical lines respectively. A visual inspection of the Morphology and Word Order conditions in Figure 1 suggests that in many cases participants failed to reproduce the full variability of the input languages; entropy scores are generally lower.

We used the stats and lme4 packages developed in R (Bates, Mächler, Bolker, \& Walker, 2015; R Core Team, 2015) to run a linear mixed effects regression model (which we will call Model 1) to explore the effect of condition on

Table 2: Central tendencies of the proportion of majority input variants in production by condition and NP type. From left to right, the mean, median and mode(s).

|  | Proportion Majority Input Variant in Production |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
|  |  | mean | median | mode $(s)$ |
| Morphology | Num Only | 0.704 | 0.8 | 0.919 |
|  | Adj Only | 0.669 | 0.7 | 0.916 |
|  | two-Mod | 0.609 | 0.65 | 0.843 |
| Word Order | Num Only | 0.580 | 0.65 | $0.094 \& 0.96$ |
|  | Adj Only | 0.585 | 0.7 | $0.104 \& 0.947$ |
|  | two-Mod | 0.442 | 0.33 | $0.089 \& 0.92$ |

regularisation behaviour (dependent variable: entropy). As fixed effects we entered Condition (two levels: Morphology as reference, and Word Order), NP Type (reverse Helmert coded with the 3 ordered levels: Num Only, Adj Only and two-Mod) and System (two levels: Input as reference, and Output). We also entered all interactions between fixed effects. As random effects, we included intercepts for Subject as well as by-Subject slopes for the effects of NP Type and System type. P-values were obtained through the lmerTest package (Kuznetsova, Bruun Brockhoff, \& Haubo Bojesen Christensen, 2015). Results show a significant effect of System ( $\beta=-0.346, S E=0.085, p<.001$ ), suggesting that participants did indeed regularise their input in their output productions. We also found a significant interaction between System and Condition ( $\beta=-0.284, S E=0.119, p=.021$ ), suggesting that participants regularised their input significantly more in the Word Order condition. Results show the expected effect of higher input entropies in two-Mod NPs ( $\beta=0.21, S E=0.024, p=<.001$ ), and no significant interactions between NP Type and System (largest: $\beta=0.027$, $S E=0.028, p=.324$ ) or between NP Type, System and Condition (largest: $\beta=-0.041, S E=0.039, p=.299$ ). These results suggest that participants regularised their input systems across conditions and NP types, and that participants in the Word Order condition regularised them more than those in the Morphology condition.
Variant production Table 2 provides the central tendencies for proportion use of the majority input variant for each NP type. We observe that all distributions in the Word Order condition are bimodal, with modes of the distributions of majority variant use at $P \leq 0.1$ and $P>0.9$ across NP types, suggesting two opposite trends amongst participants: one towards the over-production of the majority input word order variants and another, towards their under-production.

Participants under-producing the majority word order variant in one-modifier NPs are necessarily producing modifiers pre-nominally. Figure 2 shows the overall proportions of the variants produced for two-Mod NPs by all participants. The input proportions are represented by the yellow vertical lines. The word order produced the most is the majority input variant N Adj Num. Although the three remaining input variants (below the grey solid line division) were equally frequent in the input language, the Num Adj N word order is overall


Figure 1：Entropy scores of participants＇production systems．From top to bottom，scores for the Morphology（green）and Word Order（red）conditions in Experiment 1 and for the NoL1 Word Order condition（orange）in Experiment 2．From left to right， entropies of participants＇full production sets as well as entropies by NP type：one－modifier Num（Num Only），one－modifier Adj（Adj Only），and two－modifier（two－Mod）NPs．Input entropy scores are indicated by dashed vertical lines．Minimum entropy scores are indicated by solid vertical lines．Minimum entropy is always 0 for each NP type in isolation but 1.37 for the overall system as it necessitates a minimum of 3 variants，one per NP type．


Figure 2：Box plot displaying the output proportions of two－ modifier variants in the Word Order condition with individual participants＇data points overlaid．Seen（bottom）and unseen （top）variants during training are divided by a solid grey line． Vertical yellow lines indicate input proportions．
more frequently used（although only by a minority as indi－ cated by the median value 0 ）．Only $30 \%$ of participants pro－ duced systems with both harmonic variants（Num Adj N and N Adj Num）－and only $19 \%$ produced both variants more than once，suggesting that although both harmonic orders are preferred overall，they do not generally coexist within the pro－ ductions of a single participant．

We ran a logistic regression model，which we will call Model 2，to explore the average difference between the pro－ portions of Num Adj N variants in input and output linguistic systems．We entered System（two levels：Input as reference， and Output）as the only fixed effect．Random intercepts for Subject as well as by－Subject random slopes for the effect of System were also included．Results show that the Num Adj N variant is produced significantly less in output languages than in the input language（ $\beta=-7.641, S E=1.943, p<.001$ ）． Only a minority of participants overproduced this variant，the majority of participants were in fact under－producing it．On top of the observed preference for harmonic order，these re－ sults confirm a tendency to avoid systems with two opposite N－peripheral variants，i．e．N Adj Num and Num Adj N．

## Discussion of Experiment 1

Our results provide evidence that learners regularise proba－ bilistic unconditioned variation in both morphology and word order．Regularisation behaviour is in line with an overarching simplicity bias argued to be at play in language learning and use（Culbertson \＆Kirby，2016）．Though the input languages were similar in terms of overall system complexity，regulari－ sation behaviour was slightly stronger in the Word Order con－ dition than in the Morphology condition．A close analysis of the variant usage in the Word Order condition suggests that this difference is driven by a bias in favour of harmonic N Adj Num and Num Adj N variants but against their coexis－ tence within a system．This bias could be the result of L1 transfer；participants may have overproduced the Num Adj

Table 3: Probabilistic input language in the NoL1 Word order condition in contrast to the Word Order condition in Experiment 1. Changes in the variant set are indicated with boxes.

| NP TYPE | WORD ORDER | NOL1 WORD ORDER |
| :--- | :--- | :--- |
| NUM | $0.6 N P \rightarrow \mathbf{N}$ nefri | $0.6 N P \rightarrow \mathbf{N}$ nefri |
| ONLY | $0.4 N P \rightarrow$ nefri N | $0.4 N P \rightarrow$ nefri N |
|  |  |  |
| ADJ | $0.6 \mathrm{NP} \rightarrow \mathbf{N}$ kogla | $0.6 \mathrm{NP} \rightarrow \mathbf{N}$ kogla |
| ONLY | $0.4 N P \rightarrow$ kogla N | $0.4 N P \rightarrow$ kogla N |
|  |  |  |
|  | $0.6 N P \rightarrow \mathbf{N}$ kogla nefri | $0.6 N P \rightarrow \mathbf{N}$ kogla nefri |
| TWO | $0.1 \overline{3} N P \rightarrow$ nefri kogla N | $0.1 \overline{3} N P \rightarrow$ N nefri kogla |
| MOD | $0.1 \overline{3} N P \rightarrow$ nefri N kogla | $0.1 \overline{3} N P \rightarrow$ nefri N kogla |
|  | $0.1 \overline{3} N P \rightarrow$ kogla N nefri | $0.1 \overline{3} N P \rightarrow$ kogla N nefri |
|  |  |  |

N word order because it is the most common order in their L1 grammar. To minimise the possible effects of this levelspecific word order bias, Experiment 2 investigated learning in a second word order condition, removing the English-like two-modifier harmonic pattern from the input.

## Experiment 2

Experiment 2 follows the same design as the Word Order condition described in Experiment 1, with one difference: the set of two-modifier NP input variants. As illustrated in Table 3, we replaced the Num Adj N variant with the N Num Adj pattern, maintaining the number of harmonic word orders (two, i.e. N Adj Num and N Num Adj) but eliminating the L1 variant and the presence of opposite N-peripheral patterns. For ease of reference, we call Experiment 2 the NoL1 Word Order condition. We expect the change in the input language to mitigate the effect of L1 transfer and to increase the coexistence of both harmonic patterns.

Participants Twenty-eight native-English speakers (aged between 18 and 35 , mean $=24.8$ ) were recruited via the University of Edinburgh's Careers Service advertisement database. Participants received $£ 6$. Only the data from 26 participants were fit for analysis as two participants either failed to learn the noun lexicon or failed to learn the associations between phrases and pictures.

## Results

Entropy scores obtained in the NoL1 Word Order condition are shown in Figure 1 (coloured in orange). We ran a linear mixed effects model as in Experiment 1 to explore the effect of condition on regularisation behaviour (dependent variable: entropy), including the conditions in Experiment 1 plus NoL1 Word Order. The mixed-effects structure was the same as in the Model 1 but with reverse Helmert coding of Condi-


Figure 3: Box plot displaying the output proportions of twomodifier variants in the NoL1 Word Order condition with individual participants' data points overlaid. Divided by a solid grey line, seen (bottom) and unseen (top) variants during training. Vertical light brown lines indicate input proportions.
tion such that NoL1 Word Order was directly compared to the Morphology condition from Experiment 1, and the Word Order condition was compared to the average of the Morphology and NoL1 Word Order conditions. Results show a significant effect of System ( $\beta=-0.483, S E=0.051, p<.001$ ) and a significant interaction between Word Order and System ( $\beta=-0.073, S E=0.036, p=.046$ ), ratifying the results in Model 1. However, we did not find a significant interaction between NoL1 Word Order and System ( $\beta=-0.063, S E=$ $0.063, p=.317$ ), suggesting that participants in the Morphology and the NoL1 Word Order conditions regularised their input to similar degrees, and on average they regularised it less than participants in the Word Order condition in Experiment 1. As in Model 1, we did not find significant interactions between NP Type and System (largest: $\beta=0.016$, $S E=0.015, p=.288$ ) or between NP Type, System and Condition (largest: $\beta=-0.015, S E=0.011, p=.168$ ). These results suggest that participants regularised their input systems across conditions and NP types, and that whilst participants in the Word Order condition regularised more than those in the Morphology condition, participants in the Morphology and the NoL1 Word Order conditions regularised their input to similar degrees. Excluding the Num Adj N variant in the input language thus eliminated the difference between levels. In other words, participants do not regularise probabilistic unconditioned variation in word order more than in morphology.

Figure 3 shows the overall proportions of the variants produced for two-Mod NPs in the NoL1 Word Order condition. We observe that the most produced word order is the majority input variant N Adj Num, and that the harmonic N Num

Adj word order is overall more frequent than any other minority input variant. Unlike in the Word Order condition where systems with both Num Adj N and N Adj Num patterns were not common, $65 \%$ of participants produced systems with both N Adj Num and N Num Adj harmonic variants in the NoL1 Word Order condition. We ran a logistic regression model to test the difference between the proportions of N Num Adj variants in input and output linguistic systems across participants. We used the same mixed-effects structure as in Model 2. Results suggest that the proportion of N Num Adj variants in the output languages is not significantly different from the input proportion across participants ( $\beta=-0.594, S E=0.546, p=.277$ ).

## Discussion

Our experimental results reveal regularisation behaviour in the production of complex systems of variation in morphology and word order. They also suggest that regularisation behaviour is of similar strength between these linguistic levels given input languages with comparable initial complexities. In Experiment 1 we found higher levels of regularisation in word order than in morphology, apparently due to the specific properties of the set of variants in the input languages. When both harmonic pre-nominal and post-nominal two-modifier variants were included, the coexistence of both variants in a single production system was rare. Although a preference for harmonic order and consistent head position may have been at play, the interference of L1 transfer cannot be categorically rejected. Indeed previous research suggests that L2 learners tend to access their L1 knowledge if it matches the novel input (Weber, Christiansen, Petersson, Indefrey, \& Hagoort, 2016). In Experiment 2, we showed that eliminating opposite N -peripheral positions in the subset of two-modifier variants by replacing Num Adj N with N Num Adj eliminates the difference in regularisation between levels. Our results do not suggest general level-specific learning biases that could straightforwardly predict a typological asymmetry between the strength and speed of regularisation in morphology and word order hinted at in pidgin and creole studies (Good, 2015). Instead, they suggest that asymmetries in regularisation processes in language formation ought to be sought in asymmetries in the input complexity of traits across levels, also taking into account the overlap of features between contributing languages.

## Conclusion

Our results suggest similar strengths of regularisation between linguistic levels given input languages with comparable initial complexities. Nevertheless, preferences for certain patterns within a linguistic level might in fact vary the strength of regularisation behaviour within a given level.

## References

Bates, D., Mächler, M., Bolker, B., \& Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48.

Culbertson, J., \& Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. Frontiers in Psychology, 6, 1964.
Culbertson, J., Smolensky, P., \& Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122(3), 306-329.
Fehér, O., Wonnacott, E., \& Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. Journal of Memory and Language, 91, 158-180.
Ferdinand, V., Kirby, S., \& Smith, K. (2017). The cognitive roots of regularization in language. arXiv preprint arXiv:1703.03442.
Good, J. (2015). Paradigmatic complexity in pidgins and creoles. Word Structure, 8(2), 184-227.
Hudson Kam, C. L., \& Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language learning and development, 1(2), 151-195.
Hudson Kam, C. L., \& Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. Cognitive psychology, 59(1), 30-66.
Kuznetsova, A., Bruun Brockhoff, P., \& Haubo Bojesen Christensen, R. (2015). Imertest: Tests in linear mixed effects models [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lmerTest
R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Senghas, A., \& Coppola, M. (2001). Children creating language: How nicaraguan sign language acquired a spatial grammar. Psychological science, 12(4), 323-328.
Siegel, J. (2004). Morphological simplicity in pidgins and creoles. Journal of Pidgin and Creole languages, 19(1), 139-162.
Slabakova, R. (2013). What is easy and what is hard to acquire in a second language. Contemporary approaches to second language acquisition, 9, 5.
St Clair, M. C., Monaghan, P., \& Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. Cognitive Science, 33(7), 1317-1329.
van Trijp, R. (2013). Linguistic assessment criteria for explaining language change: A case study on syncretism in german definite articles. Language Dynamics and Change, 3(1), 105-132.
Weber, K., Christiansen, M. H., Petersson, K. M., Indefrey, P., \& Hagoort, P. (2016). fmri syntactic and lexical repetition effects reveal the initial stages of learning a new language. Journal of Neuroscience, 36(26), 6872-6880.
Wonnacott, E., \& Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In Bucld 29: Proceedings of the 29th annual boston university conference on language development (pp. 663-673).

# Converting Cascade-Correlation Neural Nets into Probabilistic Generative Models 

Ardavan S. Nobandegani ${ }^{1,3} \quad$ Thomas R. Shultz ${ }^{2,3}$<br>\{ardavan.salehinobandegani@ mail.mcgill.ca, thomas.shultz@mcgill.ca\}<br>${ }^{1}$ Department of Electrical and Computer Engineering, McGill University<br>${ }^{2}$ School of Computer Science, McGill University<br>${ }^{3}$ Department of Psychology, McGill University


#### Abstract

Humans are not only adept in recognizing what class an input instance belongs to (i.e., classification task), but perhaps more remarkably, they can imagine (i.e., generate) plausible instances of a desired class with ease, when prompted. Inspired by this, we propose a framework which allows transforming Cascade-Correlation Neural Networks (CCNNs) into probabilistic generative models, thereby enabling CCNNs to generate samples from a category of interest. CCNNs are a wellknown class of deterministic, discriminative NNs, which autonomously construct their topology, and have been successful in accounting for a variety of psychological phenomena. Our proposed framework is based on a Markov Chain Monte Carlo (MCMC) method, called the Metropolis-adjusted Langevin algorithm, which capitalizes on the gradient information of the target distribution to direct its explorations towards regions of high probability, thereby achieving good mixing properties. Through extensive simulations, we demonstrate the efficacy of our proposed framework. Importantly, our framework bridges computational, algorithmic, and implementational levels of analysis.


Keywords: Deterministic Discriminative Neural Networks; Probabilistic Generative Models; Markov Chain Monte Carlo

## 1 Introduction

A green-striped elephant! Probably no one has seen such a thing-no surprise. But what is a surprise is our ability to easily imagine one. Humans are not only adept in recognizing what class an input instance belongs to (i.e., classification task), but more remarkably, they can imagine (i.e., generate) plausible instances of a desired class, when prompted. In fact, humans can generate instances of a desired class, say, elephant, that they have never encountered before, like, a greenstriped elephant. ${ }^{1}$ In this sense, humans' generative capacity goes beyond merely retrieving from memory. In computational terms, the notion of generating examples from a desired class can be formalized in terms of sampling from some underlying probability distribution, and has been extensively studied in machine learning under the rubric of probabilistic generative models.

Cascade-Correlation Neural Networks (CCNNs) (Fahlman \& Lebiere, 1989) are a well-known class of discriminative (as opposed to generative) models that have been successful in simulating a variety of phenomena in the developmental literature, e.g., infant learning of word-stress patterns in artificial languages (Shultz \& Bale, 2006), syllable boundaries (Shultz \& Bale, 2006), visual concepts (Shultz, 2006),

[^186]and have also been successful in capturing important developmental regularities in a variety of tasks, e.g., the balancescale task (Shultz, Mareschal, \& Schmidt, 1994; Shultz \& Takane, 2007), transitivity (Shultz \& Vogel, 2004), conservation (Shultz, 1998), and seriation (Mareschal \& Shultz, 1999). Also, CCNNs exhibit several similarities with known brain functions: distributed representation, self-organization of network topology, layered hierarchical topologies, both cascaded and direct pathways, an S-shaped activation function, activation modulation via integration of neural inputs, longterm potentiation, growth at the newer end of the network via synaptogenesis or neurogenesis, pruning, and weight freezing (Westermann, Sirois, Shultz, \& Mareschal, 2006). Nonetheless, in virtue of being deterministic and discriminative, CCNNs have so far lacked the capacity to probabilistically generate examples from a category of interest.

In this work, we propose a framework which allows transforming CCNNs into probabilistic generative models, thereby enabling CCNNs to generate samples from a category. Our proposed framework is based on a Markov Chain Monte Carlo (MCMC) method, called the MetropolisAdjusted Langevin (MAL) algorithm, which employs the gradient of the target distribution to guide its explorations towards regions of high probability, thereby significantly reducing the undesirable random walk often observed at the beginning of an MCMC run (a.k.a. the burn-in period). MCMC methods are a family of algorithms for sampling from a desired probability distribution, and have been successful in simulating important aspects of a wide range of cognitive phenomena, e.g., temporal dynamics of multistable perception (Gershman, Vul, \& Tenenbaum, 2012; Moreno-Bote, Knill, \& Pouget, 2011), developmental changes in cognition (Bonawitz, Denison, Griffiths, \& Gopnik, 2014), category learning (Sanborn, Griffiths, \& Navarro, 2010), causal reasoning in children (Bonawitz, Denison, Gopnik, \& Griffiths, 2014), and accounting for many cognitive biases (Dasgupta, Schulz, \& Gershman, 2016).

Furthermore, work in theoretical neuroscience has shed light on possible mechanisms according to which MCMC methods could be realized in generic cortical circuits (Buesing, Bill, Nessler, \& Maass, 2011; Moreno-Bote et al., 2011; Pecevski, Buesing, \& Maass, 2011; Gershman \& Beck, 2016). In particular, Moreno-Bote et al. (2011) showed how an attractor neural network implementing MAL can account for multistable perception of drifting gratings, and Savin and Deneve (2014) showed how a network of leaky integrate-andfire neurons can implement MAL in a biologically-realistic
manner.

## 2 Cascade-Correlation Neural Networks

CCNNs are a special class of deterministic artificial neural networks, which construct their topology in an autonomous fashion-an appealing property simulating developmental phenomena (Westermann et al., 2006) and other cases where networks need to be constructed. CCNN training starts with a two-layer network (i.e., the input and the output layer) with no hidden units, and proceeds by recruiting hidden units one at a time, as needed. Each new hidden unit is trained to maximally correlate with residual error in the network built so far, and is recruited into a hidden layer of its own, giving rise to a deep network with as many hidden layers as the number of recruited hidden units. CCNNs use sum-of-squared error as an objective function, and typically use symmetric sigmoidal activation functions with range -0.5 to +0.5 for hidden and output units. ${ }^{2}$ Some variants have been proposed: Sibling-Descendant Cascade-Correlation (SDCC) (Baluja \& Fahlman, 1994) and Knowledge-Based Cascade-Correlation (KBCC) (Shultz \& Rivest, 2001). Although in this work we focus on standard CCNNs, our proposed framework can handle SDCC and KBCC as well.

## 3 The Metropolis-Adjusted Langevin Algorithm

MAL (Roberts \& Tweedie, 1996) is a special type of MCMC method, which employs the gradient of the target distribution to guide its explorations towards regions of high probability, thereby reducing the burn-in period. More specifically, MAL combines the two concepts of Langevin dynamics (a random walk guided by the gradient of the target distribution), and the Metropolis-Hastings algorithm (an accept/reject mechanism for generating a sequence of samples the distribution of which asymptotically converges to the target distribution).

We denote random variables with small bold-faced letters, random vectors by capital bold-faced letters, and their corresponding realizations by non-bold-faced letter. The MAL algorithm is outlined in Algorithm 1 wherein $\pi(\mathbf{X})$ denotes the target probability distribution, $\tau$ is a positive real-valued parameter specifying the time-step used in the Euler-Maruyama approximation of the underlying Langevin dynamics, $N$ denotes the number of samples generated by the MAL algorithm, $q$ denotes the proposal distribution (a.k.a. transition kernel), $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$, and $\mathbb{I}$ denotes the identity matrix. The sequence of samples generated by the MAL algorithm, $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots$, is guaranteed to converge in distribution to $\pi(\mathbf{X})$ (Robert \& Casella, 2013). It is worth noting that work in theoretical neuroscience has shown that MAL, outlined in Algorithm 1, can be implemented in a

[^187]```
Algorithm 1 The Metropolis-Adjusted Langevin Algorithm
    Input: Target distribution \(\pi(\mathbf{X})\), parameter \(\tau \in \mathbb{R}_{+}\), num-
    ber of samples \(N\).
    Output: Samples \(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(N-1)}\).
    Pick \(\mathbf{X}^{(0)}\) arbitrarily.
    for \(i=0, \ldots, N-1\) do
    Sample \(\mathbf{u} \sim\) Uniform[0,1]
    Sample \(\mathbf{X}^{*} \sim q\left(\mathbf{X}^{*} \mid \mathbf{X}^{(i)}\right)=\mathcal{N}\left(\mathbf{X}^{(i)}+\tau \nabla \log \pi\left(\mathbf{X}^{(i)}\right), 2 \tau \mathbb{I}\right)\)
        if \(\mathbf{u}<\min \left\{1, \frac{\pi\left(\mathbf{X}^{*}\right) q\left(\mathbf{X}^{(i)} \mid \mathbf{X}^{*}\right)}{\pi\left(\mathbf{X}^{(i)}\right) q\left(\mathbf{X}^{*} \mid \mathbf{X}^{(i)}\right)}\right\}\) then
                \(\mathbf{X}^{(i+1)} \leftarrow \mathbf{X}^{*}\)
        else
            \(\mathbf{X}^{(i+1)} \leftarrow \mathbf{X}^{(i)}\)
        end if
    end for
    return \(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(N-1)}\)
```

neurally-plausible manner (Savin \& Deneve, 2014; MorenoBote et al., 2011). ${ }^{3}$ In the following section, we propose a target distribution $\pi(\mathbf{X})$, allowing CCNNs to generate samples from a category of interest.

## 4 The Proposed Framework

In what follows, we propose a framework which transforms CCNNs into probabilistic generative models, thereby enabling them to generate samples from a category of interest. The proposed framework is based on the MAL algorithm given in Sec. 3. Let $f\left(X ; W^{*}\right)$ denote the input-output mapping learned by a CCNN, and $W^{*}$ denote the set of weights for a CCNN after training. ${ }^{4}$ Upon termination of training, presented with input $X$, a CCNN outputs $f\left(X ; W^{*}\right)$. Note that, in case a CCNN possesses multiple output units, $f\left(X ; W^{*}\right)$ will be a vector rather than a scalar. To convert a CCNN into a probabilistic generative model, we use the MAL algorithm with its target distribution $\pi(\mathbf{X})$ being set as follows:

$$
\begin{align*}
\tilde{\pi}(\mathbf{X}) & \triangleq p\left(\mathbf{X} \mid \mathbf{Y}=L_{j}\right) \\
& =\frac{1}{Z} \exp \left(-\beta\left\|L_{j}-f\left(\mathbf{X} ; W^{*}\right)\right\|_{2}^{2}\right) \tag{1}
\end{align*}
$$

where $\|\cdot\|_{2}$ denotes the $l_{2}$-norm, $\beta \in \mathbb{R}_{+}$is a damping factor, $Z$ is the normalizing constant, and $L_{j}$ is a vector whose element corresponding to the desired class is +0.5 (i.e., its $j^{\text {th }}$ element) and the rest of its elements are -0.5 s . The intuition behind Eq. (1) can be articulated as follows: For an input instance $\mathbf{X}=X$ belonging to the desired class $j,{ }^{5}$ the output of

[^188]the network $f\left(X ; W^{*}\right)$ is expected to be close to $L_{j}$ in $l_{2}$-norm sense. In this light, Eq. (1) is adjusting the likelihood of input instance $X$ to be inversely proportional to the base- $e$ exponent of the said $l_{2}$ distance.

For a reader familiar with probabilistic graphical models, the expression in Eq. (1) looks similar to the expression for the joint probability distribution of Markov random fields and probabilistic energy-based models, e.g., Restricted Boltzman Machines and Deep Boltzman Machines. However, there is a crucial distinction: The normalizing constant $Z$, the computation of which is intractable in general, renders learning in those models computationally intractable. ${ }^{6}$ The appropriate way to interpret Eq. (1) is to see it as a Gibbs distribution for a non-probabilistic energy-based model whose energy is defined as the square of the prediction error (LeCun, Chopra, Hadsell, Ranzato, \& Huang, 2006). Section 1.3 of (LeCun et al., 2006) discusses the topic of Gibbs distribution for nonprobabilistic energy-based models in the context of discriminitive learning, computationally modeled by $p(\mathbf{Y} \mid \mathbf{X})$ (i.e., to predict a class given an input), and raises the same issue that we highlighted above regarding the intractability of computing the normalizing constant $Z$ in general. In sharp contrast to (LeCun et al., 2006), our framework is proposed for the purpose of generating examples from a desired class, as evidenced by Eq. (1) being defined in terms of $p(\mathbf{X} \mid \mathbf{Y})$. Also crucially, the intractability of computing $Z$ raises no issue for our proposed framework due to an intriguing property of the MAL algorithm according to which the normalizing constant $Z$ need not be computed at all. ${ }^{7}$

Due to Line 4 of Algorithm 1, MAL's proposal distribution $q$ requires the computation of $\nabla \log \tilde{\pi}\left(\mathbf{X}^{(i)}\right)$, which essentially involves computing $\nabla f\left(\mathbf{X}^{(i)} ; W^{*}\right)$ (note that the gradient is operating on $\mathbf{X}^{(i)}$, and $W^{*}$ is treated as a set of fixed parameters). The multi-layer structure of CCNN ensures that $\nabla f\left(\mathbf{X}^{(i)} ; W^{*}\right)$ can be efficiently computed using Backpropagation. Alternatively, in settings where CCNNs recruit a small number of input units (hence, the cardinality of $\mathbf{X}^{(i)}$ is small), $\nabla f\left(\mathbf{X}^{(i)} ; W^{*}\right)$ can be obtained by introducing negligible perturbation to a component of input signal $\mathbf{X}^{(i)}$, dividing the resulting change in the network's outputs by the introduced perturbation, and repeating this process for all components of input signal $\mathbf{X}^{(i)}$. It is worth noting that although the idea of computing gradients through introducing small perturbations would lead to a computationally inefficient approach for learning CCNNs, it leads to a computationally efficient approach for generation, as the number of input units are typically much fewer than the number of weights in CCNNs (and artificial neural networks in general). It is crucial to note that the normalizing constant $Z$ plays no role in the computation of $\nabla \log \tilde{\pi}\left(\mathbf{X}^{(i)}\right)$.

[^189]
## 5 Simulations

In this section we demonstrate the efficacy of our proposed framework through simulations. We particularly focus on learning which can be accomplished by two input and one output units. This permits visualization of the input-output space, which lies in $\mathbb{R}^{3}$. Note that our proposed framework can handle arbitrary number of input and output units; this restriction is solely for ease of visualization.

### 5.1 Continuous-XOR Problem

In this section, we show how our proposed framework allows a CCNN, trained on the continuous-XOR classification task, to generate examples from a category of interest. The output unit has a symmetric sigmoidal activation function with range -0.5 and +0.5 . The training set consists of 100 samples in the unit-square $[0,1]^{2}$, paired with their corresponding labels. More specifically, the training set is comprised of all the ordered-pairs starting from $(0.1,0.1)$ and going up to $(1,1)$ with equal steps of size 0.1 , paired with their corresponding labels (i.e., +0.5 for positive samples and -0.5 for negative samples); see Fig. 1(top-left). After training, a


Figure 1: A CCNN trained on the continuous-XOR classification task. Top-left: Training patterns. All the patterns in the gray quadrants are negative examples with label -0.5 , and all the patterns in the white quadrants are positive examples with label +0.5 . Red dotted lines depict the boundaries. Topright: The input-output mapping, $f\left(x_{1}, x_{2} ; W^{*}\right)$, learned by a CCNN, along with a colorbar. Bottom: The top-down view of the curve depicted in top-right, along with a colorbar.

CCNN with 6 hidden layers is obtained whose input-output mapping, $f\left(x_{1}, x_{2} ; W^{*}\right)$, is shown in Fig. 1 (top-right). ${ }^{8}$

[^190]

Figure 2: Generating example for the positive category, under various choices for MAL parameter $\tau$ and damping factor $\beta$. Contour-plot of the learned mapping, $f\left(x_{1}, x_{2} ; W^{*}\right)$, along with its corresponding colorbar is shown in each sub-figure. Generated samples are depicted by red dots. $N$ denotes the total number of samples generated by MAL, and $A R$ denotes the corresponding acceptance rate. (a) $\tau=5 \times 10^{-5}$ leads to a very slow exploration of the input space. (b) $\tau=5 \times 10^{-3}$ leads to an adequate exploration of the input space, however, $\beta=1$ is not penalizing undesirable input regions severely enough. (c) A desirable performance is achieved by $\tau=5 \times 10^{-3}$ and $\beta=10$.

Fig. 2 shows the efficacy of our proposed framework in enabling CCNNs to generate samples from a category of interest, under various choices for MAL parameter $\tau$ (see Algorithm 1) and damping factor $\beta$ (see Eq. (1)); generated samples are depicted by red dots. For the results shown in Fig. 2, the category of interest is the category of positive examples, i.e., the category of input patterns which, upon being presented to the (learned) network, would be classified as positive by the network. Because $\tau$ controls the amount of jump between consecutive proposals made by MAL, the following behavior is expected: For small $\tau$ (Fig. 2(a)) consecutive proposals are very close to one another, leading to a slow exploration of the input domain. As $\tau$ increases, bigger jumps are made by MAL (Fig. 2(b)). ${ }^{9}$ Parameter $\beta$ controls how severely deviations from the desired class label (here, +0.5 ) are penalized. The larger the parameter $\beta$, the more severely such deviations are penalized and the less likely MAL moves toward such regions of input space. Acceptance Rate (AR), defined as the number of accepted moves divided by the total number of suggested moves, is also presented for the results shown in Fig. 2. Fig. 2(c) shows that for $\tau=5 \times 10^{-3}$ and $\beta=10$, our proposed framework demonstrates desirable performance: virtually all of the generated samples fall within the desired input regions (i.e., the regions associated with hot colors, signaling the closeness of network's output to +0.5 in those regions; see Fig. 1(bottom)) and the desired regions are adequately explored (i.e., all hot-colored input regions being visited and almost evenly explored).

Fig. 2 depicts all the first $N=2000$ samples generated

[^191]by MAL, without excluding the so-called burn-in period. In that light, the result shown in Fig. 2(c) nicely demonstrates how MAL-by directing its suggestions toward the direction of gradient and therefore moving toward regions with high likelihood-could alleviate the need for discarding a (potentially large) number of samples generated at the beginning of an MCMC which are assumed to be unrepresentative of equilibrium state, a.k.a. the burn-in period. Fig. 3 shows the performance of our framework in enabling the learned CCNN to generate from the category of negative examples, with $\tau=5 \times 10^{-3}$ and $\beta=10$.


Figure 3: Generating example for the negative category, with $\tau=5 \times 10^{-3}, \beta=10$. Generated samples are shown by blue dots. Total number of samples generated is $N=2000$, with $A R=65.13 \%$.

### 5.2 Two-Spirals Problem

Next, we show how our proposed framework allows a CCNN, trained on the famously difficult two-spirals classification task (Fig. 4), to generate examples from a category of interest. The output unit has a symmetric sigmoidal activation
function with range -0.5 and +0.5 . The training set consists of 194 samples ( 97 samples per spiral), in the square $[-6.5,6.5]^{2}$, paired with their corresponding labels $(+0.5$ and -0.5 for positive and negative samples, respectively). The training patterns are shown in Fig. 4(top-left); cf. (Chalup \& Wiklendt, 2007) for details. After training, a CCNN with 14 hidden layers is obtained whose input-output mapping, $f\left(x_{1}, x_{2} ; W^{*}\right)$, is depicted in Fig. 4(top-right).


Figure 4: A CCNN trained on the two-spirals classification task. Top-left: Training patterns. Positive patterns (associated with label +0.5 ) are shown by hollow circles, and negative patterns (associated with label -0.5 ) by black circles. Positive spiral is depicted by a dashed line, and negative spiral by a dotted line. Top-right: The input-output mapping, $f\left(x_{1}, x_{2} ; W^{*}\right)$, learned by a CCNN, along with a colorbar. Bottom: The top-down view of the curve depicted in topright, along with a colorbar.

Fig. 5(top) and Fig. 5(bottom) show the efficacy of our proposed framework in enabling CCNNs to generate samples from the positive and negative categories, respectively. Although similar patterns of behavior observed in Sec. 5.1 due to increasing/decreasing $\beta$ and $\tau$ are observed here as well, due to the lack of space such results are omitted. The results in Fig. 5 depict all the first $N=15000$ samples generated by MAL, without excluding the burn-in period. In that light, these results again demonstrate the efficacy of MAL in alleviating the need for discarding a (potentially large) number samples generated at the beginning of an MCMC run.

Interestingly, our proposed framework also allows CCNNs to generate samples subject to some forms of constraints. For example, Fig. 6 demonstrates how our proposed framework enables a CCNN, trained on the continuous-XOR classification task (see Sec. 5.1), to generate examples from the positive category, under the following constraint: Generated


Figure 5: Generating example for the positive and negative categories, with $\beta=20$ and $\tau=0.7$. Contour-plot of the learned mapping, $f\left(x_{1}, x_{2} ; W^{*}\right)$, along with its corresponding colorbar is shown in each sub-figure. $N$ denotes the total number of samples generated by MAL, and $A R$ denotes the corresponding acceptance rate. Top: Generated example for the positive category, with $N=15000$ and $A R=40.69 \%$; generated samples are depicted by red dots. Bottom: Generated example for the negative category, with $N=15000$ and $A R=40.28 \%$; generated samples are depicted by blue dots.
samples must lie on the curve $x_{2}=0.25 \sin \left(8 \pi x_{1}\right)+0.5$. To generate samples from the positive category while satisfying this constraint, MAL adopts our proposed target distribution given in Eq. (1), and treats $x_{1}$ as an independent and $x_{2}$ as a dependent variable.

## 6 General Discussion

Although we discussed our proposed framework in the context of CCNNs, it can be straightforwardly extended to handle some other kinds of artificial neural networks, e.g. Multilayer Perceptron and Deep Convolutional Neural Networks. Furthermore, our proposed framework, together with recent work in theoretical neuroscience showing possible neurallyplausible implementations of MAL (Savin \& Deneve, 2014; Moreno-Bote et al., 2011), suggests an intriguing modular hypothesis according to which generation could result from two separate modules interacting with each other (in our case, a CCNN and a neural network implementing MAL). This


Figure 6: Generating examples for the positive category, under constraint $x_{2}=0.25 \sin \left(8 \pi x_{1}\right)+0.5$ (dash-dotted curve), with $N=5000$ and $A R=39.82 \%$. Contour-plot of the learned mapping, $f\left(x_{1}, x_{2} ; W^{*}\right)$, along with its corresponding colorbar is depicted. Generated samples are shown by red dots, which appear mainly as solid red curves due to high density.
hypothesis yields the following prediction: There should be some brain impairments which lead to a marked decline in a subject's performance in generative tasks (i.e., tasks involving imagery, or imaginative tasks in general) but leave the subject's learning abilities (nearly) intact. Studies on learning and imaginative abilities of hippocampal amnesic patients already provide some supporting evidence for this idea (Hassabis, Kumaran, Vann, \& Maguire, 2007; Spiers, Maguire, \& Burgess, 2001; Brooks \& Baddeley, 1976).

According to Line 4 of Algorithm 1, to generate the $i^{\text {th }}$ sample, MAL requires access to a fine-tuned, Gaussian noise with mean $\mathbf{X}^{(i)}+\tau \nabla \log \pi\left(\mathbf{X}^{(i)}\right)$ for its proposal distribution $q$. Recently Savin and Deneve (2014) showed how a network of leaky integrate-and-fire neurons can implement MAL in a neurally-plausible manner. However, as Gershman and Beck (2016) point out, Savin and Deneve leave unanswered what the source of that fine-tuned Gaussian noise could be. Our proposed framework may provide an explanation, not for the source of Gaussian noise, but for its fine-tuned mean value. According to our modular account, the main component of the mean value, which is $\nabla \log \pi\left(\mathbf{X}^{(i)}\right)$, may come from another module (in our case, a CCNN) which has learned some input-output mapping $f\left(X ; W^{*}\right)$, based on which the target distribution $\pi\left(\mathbf{X}^{(i)}\right)$ is defined (see Eq. (1)).

The idea of sample generation under constraints could be an interesting line of future work. Humans clearly have the capacity to engage in imaginative tasks under a variety of constraints, e.g., when given incomplete sentences or fragments of a picture people can generate possible completions (Sanborn \& Chater, 2016). Also, our proposed framework can be used to let a CCNN generate samples from a category of interest at any stage during CCNN construction. In that light, our proposed framework, along with a neurally-plausible implementation of MAL, gives rise to a self-organized generative model: a generative model possessing the self-constructive property of CCNNs. Such self-
organized generative models could provide a wealth of developmental hypotheses as to how the imaginative capacities of children change over development, and models with quantitative predictions to compare against. We see our work as a step towards such models. Last but not least, our framework strongly suggests that, contrary to conventional wisdom, the boundary between discriminative and generative models is blurry-perhaps they are just two sides of the same coin!

Acknowledgments. This work is funded by an operating grant to TRS from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Kevin Da Silva Castanheira for helpful comments on an earlier draft of this work.

## References

Baluja, S., \& Fahlman, S. E. (1994). Reducing network depth in the cascade-correlation learning architecture. Technical Report \# CMU-CS-94-209, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
Bonawitz, E., Denison, S., Gopnik, A., \& Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. Cognitive Psychology, 74, 35-65.
Bonawitz, E., Denison, S., Griffiths, T. L., \& Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. Trends in Cognitive Sciences, 18(10), 497-500.
Brooks, D., \& Baddeley, A. (1976). What can amnesic patients learn? Neuropsychologia, 14(1), 111-122.
Buesing, L., Bill, J., Nessler, B., \& Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLOS Comput Biol, 7(11), e1002211.
Chalup, S. K., \& Wiklendt, L. (2007). Variations of the two-spiral task. Connection Science, 19(2), 183-199.
Dasgupta, I., Schulz, E., \& Gershman, S. J. (2016). Where do hypotheses come from? Center for Brains, Minds and Machines (CBMM) Memo No. 056.
Fahlman, S. E., \& Lebiere, C. (1989). The cascade-correlation learning architecture. In Adv. in Neural Information Processing Systems, pp. 524-532.
Gershman, S. J., \& Beck, J. M. (2016). Complex probabilistic inference: From cognition to neural computation. In Computational Models of Brain and Behavior, ed A. Moustafa (Hoboken, NJ: Wiley-Blackwell).
Gershman, S. J., Vul, E., \& Tenenbaum, J. B. (2012). Multistability and perceptual inference. Neural Computation, 24(1), 1-24.
Hassabis, D., Kumaran, D., Vann, S. D., \& Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. Proceedings of the National Academy of Sciences, 104(5), 1726-1731.
LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., \& Huang, F. (2006). A tutorial on energy-based learning. Predicting Structured Data, 1, 0.
Mareschal, D., \& Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. Connection Science, 11(2), 149-186.
Moreno-Bote, R., Knill, D. C., \& Pouget, A. (2011). Bayesian sampling in visual perception. Proceedings of the National Academy of Sciences, 108(30), 1249112496.

Pecevski, D., Buesing, L., \& Maass, W. (2011). Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. PLoS Comput Biol, 7(12), e1002294.
Robert, C., \& Casella, G. (2013). Monte Carlo statistical methods. Springer Science \& Business Media.
Roberts, G. O., \& Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(1), 255-268.
Roberts, G. O., \& Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. Bernoulli, 341-363.
Sanborn, A. N., \& Chater, N. (2016). Bayesian brains without probabilities. Trends in Cognitive Sciences, 20(12), 883-893.
Sanborn, A. N., Griffiths, T. L., \& Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. Psychological Review, 117(4), 1144.
Savin, C., \& Deneve, S. (2014). Spatio-temporal representations of uncertainty in spiking neural networks. In Adv. in Neural Information Processing Systems.
Shultz, T. R. (1998). A computational analysis of conservation. Developmental Science, 1(1), 103-126.
Shultz, T. R. (2006). Constructive learning in the modeling of psychological development. Processes of Change in Brain and Cognitive Development: Attention and Performance, 21, 61-86.
Shultz, T. R., \& Bale, A. C. (2006). Neural networks discover a near-identity relation to distinguish simple syntactic forms. Minds and Machines, 16(2), 107-139.
Shultz, T. R., Mareschal, D., \& Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. Machine Learning, 16(1-2), 57-86.
Shultz, T. R., \& Rivest, F. (2001). Knowledge-based cascade-correlation: Using knowledge to speed learning. Connection Science, 13(1), 43-72.
Shultz, T. R., \& Takane, Y. (2007). Rule following and rule use in the balance-scale task. Cognition, 103(3), 460-472.
Shultz, T. R., \& Vogel, A. (2004). A connectionist model of the development of transitivity. In Proceedings of the 26th Annual Conference of the Cognitive Science Society (pp. 1243-1248).
Spiers, H. J., Maguire, E. A., \& Burgess, N. (2001). Hippocampal amnesia. Neurocase, 7(5), 357-382.
Westermann, G., Sirois, S., Shultz, T. R., \& Mareschal, D. (2006). Modeling developmental cognitive neuroscience. Trends in Cognitive Sciences, 10(5), 227-232.

# Mental Representations and Computational Modeling of Context-Specific Human Norm Systems 

Vasanth Sarathy (vasanth.sarathy @tufts.edu) and Matthias Scheutz (matthias.scheutz@tufts.edu)<br>Department of Computer Science, Tufts University, Medford, MA, USA<br>Yoed Kenett (yoedkenett@gmail.com)<br>Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA<br>Mowafak M. Allaham (mallah5@uic.edu)<br>Department of Psychology, University of Illinois at Chicago, Chicago, IL, USA

Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, University of Wisconson, Madison, WI, USA
Bertram F. Malle (bfmalle@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA


#### Abstract

Human behavior is frequently guided by social and moral norms; in fact, no societies, no social groups could exist without norms. However, there are few cognitive science approaches to this central phenomenon of norms. While there has been some progress in developing formal representations of norm systems (e.g., deontological approaches), we do not yet know basic properties of human norms: how they are represented, activated, and learned. Further, what computational models can capture these properties, and what algorithms could learn them? In this paper we describe initial experiments on human norm representations in which the context specificity of norms features prominently. We then provide a formal representation of norms using Dempster-Shafer Theory that allows a machine learning algorithm to learn norms under uncertainty from these human data, while preserving their context specificity.


Keywords: social cognition, moral psychology, computational modeling, machine learning

## Introduction and Motivation

Someone's cell phone begins to ring in the library. The person quickly answers it by whispering "hold on," then leaves the library and takes the call in a normal voice outside. The person understands that taking a phone call in the library is not socially acceptable, though briefly whispering is. Somehow, the situation activated a set of norms in this person's mind, including: "when someone calls you, you should answer the phone"; "when in a library, you must not talk on the phone"; "when in a library, you may briefly whisper."

Humans living in social communities function more effectively and peacefully when their actions are guided by a shared set of norms (Bicchieri, 2006; Ullmann-Margalit, 1977). The ability to represent and follow norms has many advantages: Norm-consistent actions increase multi-party coordination and cooperation and thus benefit the community as a whole. Norms also simplify people's action selection and standardize behaviors across time and generations. And norm-consistent actions are more predictable and understandable (Malle, Scheutz, \& Austerweil, 2017).

But how does the human mind represent norms, and how are they activated and learned? Surprisingly, there are few cognitive science approaches to the central phenomenon of norms. Logical and specifically deontological approaches have been proposed to formally represent a system of norms (Bringsjord, Arkoudas, \& Bello, 2006; Scheutz \& Malle, 2014; Pereira \& Saptawijaya, 2009; Beller, 2010). These are important starting points, but their formalizations do not necessarily correspond to how norms are represented in the human mind. By contrast, a cognitive science approach would aim at an account of how norms are cognitively represented, how they are activated in relevant situations, and how they are learned in the first place. Here we take a first step toward such an account, following a recent theoretical proposal (Malle et al., 2017). We introduce a basic formal representation of norms that allows us to examine the mentioned cognitive properties of norms (representation, activation, and learning), and we ask what computational models can capture these properties, and what algorithms could learn norms.

Our paper has three main parts. In the first, we present a novel belief-theoretic norm representation format that explicitly captures the context-specificity of norms and incorporates uncertainty associated with norm representations, using Dempster-Shafer Theory (Shafer, 1976). In the second part, we introduce experimental data on human norm representation and activation that underscore the context-specificity of norms and community members' strong but imperfect agreement (uncertainty) over norm applications. In the third part we use our formal norm representation to ask how such imperfect norms systems can be learned by a computational algorithm that honors several of the critical features of norms, including their context specificity and uncertainty.

## Part 1: A Representation Format for Norms

We begin by briefly outlining our norm representation format in first-order logic and provide some intuitions as to how context and uncertainty are accounted for in the format. The
purpose is to introduce some terminology and a minimal degree of formalism in the proposed approach, which will later be useful in developing an algorithm that can learn norms.

Consider a first-order alphabet $\mathcal{L}$, in which we have all the standard symbols (variables, predicates, functors) and logical connectives. In a deontic alphabet, we further include $\mathbb{O}, \mathbb{F}, \mathbb{P}$ that denote modal operators (generally, $\mathbb{D}$ ) for obligatory, forbidden and permissible, respectively. In this alphabet, we define a norm, as follows:

Definition 1 (Norm). A norm is an expression of the form:

$$
\mathcal{N}:=C_{1}, \ldots, C_{n} \Longrightarrow(\neg) \mathbb{D}\left(A_{1}, \ldots, A_{m}\right)
$$

where C represents context conditions and A represents actions or states. The norm expression states that when the contextual atoms $C_{i}$ are true then the Actions or States $A_{j}$ are either obligatory, forbidden or permissible, or their negation.

This type of norm definition follows an approach to normative reasoning and norm formalism that some of us have taken previously (Malle et al., 2017; Bringsjord et al., 2006; Scheutz \& Malle, 2014).

In this paper, we expand the above representation format by explicitly accounting for uncertainty of a norm as follows:
Definition 2 (Belief-Theoretic Norm). A belief-theoretic norm is an expression of the form:

$$
\mathcal{N}:=[\alpha, \beta]:: C_{1}, \ldots, C_{n} \Longrightarrow(\neg) \mathbb{D}\left(A_{1}, \ldots, A_{m}\right),
$$

where $[\alpha, \beta]$ represents a Dempster-Shafer uncertainty interval, with $0 \leq \alpha \leq \beta \leq 1$.

Example 1 Consider an example of an agent reasoning about actions it can perform or states it can enter in a library. We can represent this scenario as a Belief-Theoretic Norm System, $\mathcal{T}$, as follows:
$\mathcal{N}_{1}:=[0.9,1]::$ in $($ library,$X) \Longrightarrow$ (1) state $(X$, quiet $)$
$\mathcal{N}_{2}:=[0.8,0.95]::$ in $($ library,$X) \Longrightarrow \mathbb{P}$ action $(X$, reading $)$
$\mathcal{N}_{3}:=[0.9,1]::$ in $($ library,$X) \Longrightarrow \mathbb{F}$ action $(X$, yelling $)$
$\mathcal{N}_{4}:=[0,0.3]::$ in $($ library,$X) \Longrightarrow \mathbb{O}$ action $(X$, talking $)$
$\mathcal{N}_{5}:=[0.3,0.6]::$ in $($ library,$X) \Longrightarrow \mathbb{F} \operatorname{action}(X$, talking $)$
The norms in this example have intuitive semantics. They generally state that when agent $X$ is in the library (i.e., in $($ library,$X)$ ), then the norm is activated and the agent is obligated to enter a certain state (e.g., state $(X, q u i e t))$ or prohibited from performing a certain action (e.g., action( $X$, talking). The location of the center of the uncertainty interval generally suggests the degree of truth of the norm applying and the width of the interval generally suggests the level of support or evidence for that norm. So norms $\mathcal{N}_{1}, \mathcal{N}_{2}$, and $\mathcal{N}_{3}$ have tight uncertainty intervals close to 1 indicating a confident support for their truth. Norm $\mathcal{N}_{4}$ states that the action of "talking" is obligatory in libraries. Although the uncertainty interval for this norm is tight, the center is closer to zero indicating confident support for the falsity
of the norm. Finally, in rule $\mathcal{N}_{5}$ the question of whether talking is forbidden in a library may be more uncertain, generating a wider interval centered close to 0.5 , indicating support for both truth and falsity, but a general lack of confidence in the evidence. ${ }^{1}$

A belief-theoretic norm system of this form allows the separation of evidence from the norms themselves. The evidence may come in different forms across different modalities and from different sources. The norm system, however, displays the agent's current level of belief about a set of norms that are influenced by the evidence.

In any given situation, the agent may not be reasoning with every norm in a norm system. Instead, the agent may consider a subset of the system, perhaps including only norms that are applicable to the current situation. We capture this intuition in a norm frame, defined below.
Definition 3 (Norm Frame). A norm frame $\mathcal{N}_{k}^{\Theta}$ is a set of $k$ norms, $k>0$, in which every norm has the same set of context predicates and corresponds to the same deontic operator. Thus, in Example 1, norms $\mathcal{N}_{3}$ and $\mathfrak{N} 5$ would constitute a norm frame.

We define a norm frame in this way because it allows for cognitive modeling in a situated manner-that is, reasoning about behavior relevant to a specific situation. This contextspecificity provides a convenient constraint that can help simplify computation and better capture human norm representations, as introduced next.

## Part 2: Norm Representation and Activation in Human Data

We are currently engaged in an empirical research program that tests a number of novel hypotheses about the cognitive properties of norms (Malle et al., 2017). Here we summarize two experiments that illustrate some of these properties and provide the learning data for the norm learning algorithm we introduce in Part 3. In the first experiment, participants generated norms relevant to a variety of contexts; in the second experiment, participants detected norms relevant to those contexts.

## Methodology

In the generation experiment (Kenett, Allaham, Austerweil, \& Malle, 2016), participants ( $n=100$ recruited from Amazon Mechanical Turk, AMT) inspected four pictures, one at a time, that depicted an everyday scene (e.g., library, jogging path; see Figure 1 for examples). While inspecting each picture, they had 60 seconds to type as many actions as came to mind that one is "allowed" to perform in this scene

[^192](Permissions), or is "not allowed" to perform (Prohibitions), or is "supposed" to perform (Prescriptions). This betweensubjects manipulation of norm type was constant across pictures so that each participant answered the same question (e.g., "What are you permitted to do here?") for all four pictures they encountered.


Figure 1: Four sample scene pictures used to elicit norms
To increase generalizability at the stimulus level, the total number of scenes used in the experiment was in fact eight, four that previous participants had tended to describe as locations (e.g., library, cave), and four that they had tended to describe as activities (e.g., jogging outdoors, serving in a restaurant). Each participant was randomly assigned to receive either the "location" set or the "activity" set. Item set made no difference in the results.

The resulting verbal responses were lightly cleaned for spelling and grammatical errors and responses identical in meaning were assigned the same response code, using a conservative criterion so that variants such as "listening" and "listening to music" were counted as distinct. The resulting data structures were then analyzed for consensus (i.e., how many people generated a given response for a given scene) and context distinctiveness (i.e., whether a response generated for one scene was also generated in a different scene).

In the detection experiment, we presented participants ( $n$ $=360$ recruited from AMT) with the same pictures, four per participant. Along with each picture, we presented 14 actions (randomly ordered, one at a time) that a person might perform in this context. Any given participant's task was the same for each of their four pictures: to consider the particular scene and judge whether each of the 14 actions is either permitted, or prescribed, or prohibited. This norm type factor was again a between-subjects manipulation and hence constant across pictures. In addition, to increase generalizability, we used two different formulations for each norm type, summarized in Table 1. Formulation made no difference in the results.

The 14 actions assigned to a given scene under a given norm type (e.g., Library/permitted) consisted of seven "local" and seven "imported" actions. Local actions were the

Table 1: Eliciting Probes for Three Norm Types

| Norm Type | Probe formulations |
| :--- | :--- |
| Permission | Are you allowed to do this here? <br> Are you permitted to do this here? |
| Prohibition | Are you not allowed to do this here? <br> Are you forbidden to do this here? |
| Prescription | Are you supposed to do this here? <br> Should you do this here? |

seven most frequently generated actions for the given scene and norm type in the above generation experiment-for example, the seven actions most frequently mentioned to be permitted in the library. Imported actions were comprised of top-seven actions generated for other scenes (but under the same norm type). Thus, imported actions were still frequent responses to the same norm probe, but in different contexts. ${ }^{2}$ Table 2 provides an illustration of this selection process.

Table 2: Origin of Selected Actions for Library Scene

| Action | Origin |
| :--- | :--- |
| Local, permitted |  |
| reading | from top 7 of Library |
| studying | from top 7 of Library |
| sitting | from top 7 of Library |
| checking out a book | from top 7 of Library |
| learning | from top 7 of Library |
| being quiet | from top 7 of Library |
| using computers | from top 7 of Library |
| Imported, permitted |  |
| eating | from top 7 of Beach |
| walking | from top 7 of Cave |
| listening | from top 7 of Boardroom |
| filling boxes | from top 7 of Harvesting |
| washing hands | from top 7 of Public Bathroom |
| running | from top 7 of Jogging |
| talking | from top 7 of Restaurant |

## Experimental Results

We begin by highlighting three findings from the generation experiment. ${ }^{3}$ First, even though people were entirely unconstrained in their norm-guided actions, they showed a great deal of consensus on the most central norms for each scenario. Table 3 displays (in column Consensus) the seven most frequently mentioned permission norms in two representative scenarios, Library and Jogging, with consensus computed as the percentage of participants who mentioned the particular

[^193]action as permitted in the scenario. (The patterns are consistent across other scenarios.) Second, the most consensual norms are mentioned early on; in other words, what comes to mind first is likely to be a consensual norm. Table 3 shows (in column Position) the average rank position ( $1=$ first, $2=$ second, etc.) at which each action was generated, whereby the expected position under a random distribution would be 4.2 for Library and 4.6 for Jogging. Third, the norms generated for the eight scenarios showed remarkable context specificity. Not only do the two illustrated scenes have no norm in common among their top seven, but of the 56 permitted actions that were mentioned in the top- 7 in each of the 8 scenes, only 5 appeared in more than one scene.

Table 3: Permission Norms for Library and Jogging Scenes in the Norm Generation Experiment

| Library |  |  |
| :--- | :---: | :---: |
| Permitted Action | Consensus | Position |
| reading | $84 \%$ | 2.1 |
| studying | $68 \%$ | 1.8 |
| sitting | $47 \%$ | 3.1 |
| checking out books | $47 \%$ | 4.4 |
| using computers | $32 \%$ | 5.3 |
| learning | $32 \%$ | 6.0 |
| being quiet | $32 \%$ | 7.5 |
| Jogging |  |  |
| walking | $87 \%$ | 1.4 |
| running | $87 \%$ | 1.9 |
| jogging | $53 \%$ | 4.8 |
| talking | $53 \%$ | 5.1 |
| listening to music | $33 \%$ | 4.3 |
| biking | $27 \%$ | 4.7 |
| looking at birds | $27 \%$ | 6.2 |

Two main results stand out from the detection experiment. First, people showed very high consensus in affirming the permissibility of the seven local actions for their respective scenes. For both Library and Jogging, this rate was 99\%; and across all scenes, the number was $97.2 \%$. That is, even though some of these local actions were actively generated as "permissible" by only a third or half of previous participants (see Table 3), when directly confronted with these actions, people almost uniformly recognized their permissibility. (Moreover, this recognition was fast, taking only about 1100 ms on average.)

Second, participants clearly distinguished between the local and the imported actions, accepting the latter as permissible at a significantly lower rate. For Library, this rate was $43 \%$; for Jogging, it was $75 \%$; and across all scenes, it was $66.1 \%$ (all statistical comparisons to local actions $p<.001$, signal detection discrimination parameter $d^{\prime}=1.49$ ). That is, for a given context on average, $34 \%$ of presented actions were judged to be not permitted even though they were explicitly deemed permissible in other contexts.

These results suggest that norms can be activated by static
photographs, and people show high agreement in explicitly recounting these norms (generation experiment). In a more implicit setup (detection experiment), people are fast and almost unanimous in affirming the most important norms of a given context and differentiate them well from norms originating from a different context. Thus, both explicit and implicit judgments show substantial context sensitivity. If these are some of the properties of human social and moral norms, how can they possibly be learned, by humans and machines?

## Part 3: Learning Norms

## How Do People Learn Norms?

In learning social and moral norms, people deal with multiple different norm types (permissions, prescriptions, prohibitions), using many different learning mechanisms, and taking input from many different sources. Here we focus on the process of learning permission norms from simple observation, using responses from a sample of community members described earlier in the detection experiment. Our main goal is to put our proposed computational framework to a test. In the future we will develop further applications (e.g., learning of obligations or learning from instruction)

Consider a person who has never spent time in a library. Upon entering one for the first time, he observes several people reading, studying, and a few whispering. Some sit at computers, one is eating while sitting in an armchair, although there is a sign that says "No food or drink in the library." Our observer also sees several people at the check-out counter, subsequently exiting the library, where another sign says "Don't forget to check out." Briefly, a younger person runs alongside the stacks but then sits down next to an adult.

The number of people performing each behavior, their age, expertise, appearance, perhaps responses from others, and the meaning and force of various physical symbols will all contribute to the speed and confidence with which our protagonist learns the norms of a library. Below we offer a data representational format that incorporates these and other properties of the norm learning process, a format that can also accommodate partial information and unknown prior probability distributions and that can be extended to other learning mechanisms, such as verbal instruction or trial and error.

## Data Representation Format of Norm Learning

Consider a set $S=\left\{s_{1}, \ldots, s_{n}\right\}$ of $n$ evidence sources. For example, an evidence source $s_{i}$ could be a student in the library, the librarian, or a sign at the entrance. To simplify, we are interested in learning about a norm frame $\mathcal{N}_{k}^{\Theta}$ comprising $k$ norms (out of a larger possible set) that all share the same deontic type (here, permissions) and the same general context precondition (here, library).

Let an endorsement $e_{i, j}$ be the $i^{\text {th }}$ data source's endorsement of the $j^{\text {th }}$ norm, where $e \in\{0,1, \varepsilon\}$. The value $e_{i, j}$ is a form of truth assignment, indicating whether the source endorses the norm to be true (1), false (0) or unknown ( $\varepsilon$ ). For example, an observation that a student is reading can be in-
terpreted as showing that this student endorses the norm $\mathcal{N}_{2}$ to be true in this context, hence $e_{i, \mathscr{N}_{2}}=1$. The set $\Phi_{s_{i}}$ represents a given source's finite set of endorsements within a given norm frame, such that $\left|\Phi_{s_{i}}\right|=k$.

Informally, for a set of norms in a given context and for a particular source, we can learn about that source's endorsement of each norm; if we also assign a weight (e.g., reliability, expertise) to the source, we form a data instance. Multiple data instances (i.e., evidence from multiple sources) form a data set. More formally:

Definition 4 (Data Instance). $A$ data instance $d=$ $\left(\mathcal{N}_{k}^{\Theta}, s_{i}, \Phi_{s_{i}}, m_{s_{i}}\right)$ is a tuple comprising a norm frame $\mathcal{N}_{k}^{\Theta}$, a specific source $s_{i}$, a set of endorsements $\Phi_{s_{i}}$ provided by that source, and a mass assignment $m_{s_{i}}$ corresponding to the amount of consideration or reliability placed on source $s_{i}$.
Definition 5 (Dataset). A dataset $\mathcal{D}$ is a finite set of $n$ data instances $\left\{d_{1}, \ldots, d_{n}\right\}$.

Some of the desirable properties of the proposed data representation format are that we can accommodate various types of sources (e.g., behavior, verbal responses, signs and symbols), differential source reliability (mass), order effects (updates can be tuned, if necessary, to the order of received data), missing and imprecise information (we use $\varepsilon$ to represent ignorance), lacking prior probability distributions (we do not require any priors), and varying norm dependencies (e.g., we can capture a correlation between the prohibition to yell and the prohibition to talk).

## Algorithmic Learning of Experimental Data

We can now apply this representation format to the detection data we introduced earlier. The detection experiment featured, for each scene, a norm frame $\mathcal{N}_{k}^{\Theta}$ with $k=14$ potentially permissible actions, where half of the potential actions had been specifically identified as permitted in this scene and the other half as permitted in other scenes (see Table 2). Each participant, $s_{i}$, indicated whether each of 14 actions was in fact allowed in this scene, providing responses of yes (1) or no (0) or no response ( $\varepsilon$ ), thus forming a set of endorsements $\Phi_{s_{i}}$, with $|\Phi|=14$. In this particular case we treat all sources as equally reliable, hence carrying identical $m_{s_{i}}$ weights.

With these representations in hand we can formally define the norm learning problem within our framework and set the stage for an algorithm to analyze evidence and derive a norm structure for a given context in a given community. We remind the reader that, according to Definition 2, any norm (e.g., with respect to reading in a library) has an uncertainty interval $\left[\alpha_{1}, \beta_{1}\right]$ associated with it, which reflects the quality and consistency of the evidence for a given norm to hold. The learning problem thus becomes a parameter learning problem for discovering the values of the uncertainty interval for each norm in a norm frame:

Definition 6 (Norm Learning Problem). For a norm frame $\mathcal{N}_{k}^{\Theta}$ and dataset $\mathcal{D}$, compute the parameters $\alpha_{1}, \ldots, \alpha_{k}, \beta_{1}, \ldots, \beta_{k}$ of that norm frame.

As noted earlier, each data instance $d$ represents a potential arrangement of true and false values for each of the norms in a frame. Setting aside the possibility that $e_{i, j}=\varepsilon$, each data instance thus provides a $k$-length string of 1 s and 0 s (a given participant's response string in the detection experiment). This string is a sample of the normative endorsements in the given community. The norm learning algorithm represents each string as a hypothesis in a set of hypotheses (termed Frame of Discernment in Dempster-Shafer theory) and assigns uncertainty parameters to each norm, updating those values as it considers each new data instance. Algorithm 1, displayed below, achieves this form of norm learning from a human dataset.

```
Algorithm 1 getParameters \(\left(\mathcal{D}, \mathcal{N}_{k}^{\Theta}\right)\)
    \(\mathcal{D}=\left\{d_{1}, \ldots, d_{n}\right\}\) : Dataset containing \(n\) data instances for a
    norm frame
    \(\mathcal{N}_{k}^{\Theta}\) : An unspecified norm frame containing \(k\) norms \(\mathcal{N}\)
    Initialize DS Frame \(\Theta=\left\{\theta_{1}, \ldots, \theta_{2^{k}}\right\}\)
    \(m(\Theta)=1\)
    for all \(d \in \mathcal{D}\) do
        for all \(\mathcal{N} \in \mathcal{N}_{k}^{\Theta}\) do
            Set learning parameters \(p_{1}\) and \(p_{2}\)
            \(\operatorname{Bel}(\mathcal{N} \mid d)=\frac{\operatorname{Bel}(\mathcal{N} \cap d)}{\operatorname{Bel}(\mathcal{N} \cap d)+P l(d \backslash \mathcal{N})}\)
            \(P l(\mathcal{N} \mid d)=\frac{P l(\mathcal{N} \cap d)}{P l(\mathcal{N} \cap d)+\operatorname{Bel}(d \backslash \mathcal{N})}\)
            \(\operatorname{Bel}(\mathcal{N})_{\text {new }}=p_{1} \cdot \operatorname{Bel}(\mathcal{N})_{\text {prev }}+p_{2} \cdot \operatorname{Bel}(\mathcal{N} \mid d)\)
            \(P l(\mathcal{N})_{\text {new }}=p_{1} \cdot P l(\mathcal{N})_{\text {prev }}+p_{2} \cdot \operatorname{Pl}(\mathcal{N} \mid d)\)
        end for
        Set frame \(\Theta\) with \(\operatorname{Bel}(\mathcal{N})_{\text {new }}\) and \(\operatorname{Pl}(\mathcal{N})_{\text {new }}\)
    end for
    for all \(\mathcal{N} \in \mathcal{N}_{k}^{\Theta}\) do
        \(\alpha_{\mathcal{X}} \leftarrow \operatorname{Bel}(\mathcal{N})\)
        \(\beta_{\mathcal{N}} \leftarrow P l(\mathcal{N})\)
    end for
    return \(\alpha_{1}, \ldots, \alpha_{k}, \beta_{1}, \ldots, \beta_{k}\)
```

The algorithm iterates though each data instance in the data set (line 6) and, per instance, through each norm in the norm frame (line 7). For each iteration, we first set the hyperparameters $p_{1}$ and $p_{2}$ (line 8 ) that specify how much weight the algorithm will place on previous learned knowledge $\left(p_{1}\right)$ and on each new data instance ( $p_{2}$ ). These hyper-parameters are then used to compute a conditional belief and plausibility for a norm given that particular instance of data (lines 9,10 ). The conditional beliefs and probabilities then yield an updated belief and plausibility for each norm (lines 11, 12). Finally, the algorithm updates the uncertainty interval for each norm with the new belief and plausibility values.

The result is a set of belief-theoretic norms (norms accompanied with uncertainty intervals), where the width of the uncertainty interval indicates the amount of support for the norm (which may vary, for example, as a function of number of respondents in the human data sample) and the center position of the interval should correspond to the level of agreement in the human respondents' endorsement of the norm.

To put this algorithm to the test, we selected, from our detection experiment, a norm frame of 6 (out of 14) actions for
the context of Library and a frame of 6 (out of 14) actions for the context of Jogging Path. However, we wanted to capture the context specificity of norms and constructed the frames such that 4 actions (running, sitting, walking, and washing hands) were the same in each frame, albeit differentially endorsed in the two contexts (e.g., running was clearly not permissible in Library but very much permissible in Jogging). Thus, the algorithm had to track the norm value of a given action not in general, but conditional on the specific context. If the algorithm succeeds it should recognize which actions people consider permissible and which ones they consider impermissible, for each of the two contexts, and even for those actions that occur in both contexts.

Figure 2 illustrates this success. We display single runs of the algorithm across the dataset. In the single runs, the algorithm considers each data instance (each of 30 participants' judgments) in each context once (in a fixed order), leading to wide uncertainty intervals at first, but narrower ones as the number of data instances increases (up to the maximum of 30) . We also performed iterative runs (not shown), in which the algorithm considers the dataset multiple times, each time randomly selecting a possible order of instances, and converging on an optimal estimate of the norm endorsements in the given community. These estimates are highly comparable to the end points of single runs after 30 data instances.


Figure 2: Single run of learning across two contexts. The narrowing shaded regions indicate converging uncertainty intervals as new data instances are processed. Filled circles represent the descriptive statistics from the experimental data, indicating actual norm endorsement averages among participants-the proportion of participants who answered yes to the question: "Is this action allowed here?" The algorithm displays convergence towards the descriptive statistics (which it was not given), while maintaining a level of uncertainty reflecting the imperfect agreement within the data.

## Conclusion

In this paper we presented a formal representation of norms using first-order logic and Dempster-Shafer theory. The representation captures the context specificity of norms that our experimental data suggest are strongly present in humans. Using a data representation format that incorporates several properties of human norm representation and learning, we then developed a novel algorithm for automatically learning context-sensitive norms from the human data. Because the data format is highly generalizable, norms could be learned from different types of evidence sources in different contexts, and explicitly captures uncertainty due to variations in the source's reliability and the quality of the evidence. The proposed representation and learning techniques provide a promising platform for studying, computationally, a wide array of cognitive properties of norms.

## Acknowledgment

This research was funded in part by grants from the Office of Naval Research (ONR), N00014-14-1-0144, and from the Defense Advanced Research Projects Agency (DARPA), SIMPLEX 14-46-FP-097. The opinions expressed here are our own and do not necessarily reflect the views of ONR or DARPA.

## References

Beller, S. (2010). Deontic reasoning reviewed: psychological questions, empirical findings, and current theories. Cognitive Processing, 11(2), 123-132.
Bicchieri, C. (2006). The grammar of society: The nature and dynamics of social norms. New York, NY: Cambridge University Press.
Bringsjord, S., Arkoudas, K., \& Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems, 21(4), 38-44.
Kenett, Y. N., Allaham, M. M., Austerweil, J. L., \& Malle, B. F. (2016, November). The norm fluency task: Unveiling the properties of norm representation. (Poster.). In Poster presented at the 57th Annual Meeting of the Psychonomic Society, Boston, MA, November 2016.
Malle, B., Scheutz, M., \& Austerweil, J. (2017). Networks of social and moral norms in human and robot agents. In $A$ world with robots (pp. 3-17). Springer.
Pereira, L. M., \& Saptawijaya, A. (2009). Modelling morality with prospective logic. International Journal of Reasoningbased Intelligent Systems, 1(3-4), 209-221.
Scheutz, M., \& Malle, B. F. (2014). "Think and do the right thing"-A plea for morally competent autonomous robots. In Ethics in science, technology and engineering, 2014 IEEE international symposium on (pp. 1-4).
Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton University Press.
Ullmann-Margalit, E. (1977). The emergence of norms. Oxford: Clarendon Press.

# Attractor Dynamics in Delay Discounting: A Call for Complexity 

Martin Schoemann (martin.schoemann@tu-dresden.de)<br>Department of Psychology, Technische Universität Dresden, 01062 Dresden, Germany

Stefan Scherbaum (stefan.scherbaum@tu-dresden.de)
Department of Psychology, Technische Universität Dresden, 01062 Dresden, Germany


#### Abstract

The outcomes of intertemporal choices indicate that people discount rewards by their delay. These outcomes are well described by discounting functions. However, to fully understand the decision process one needs models describing how the process of decision-making unfolds dynamically over time. Here, we validate a recently published attractor model that extends discounting functions through a description of the dynamics leading to a final choice outcome within and across trials. We focus on the decision dynamics across trials. We derive qualitative predictions for the inter-trial dynamics of sequences of decisions that are unique to this type of model. We test these predictions in a delay discounting game where we sequentially manipulated subjective values of options across all attribute dimensions. Results confirm the model's predictions. We discuss future challenges on integrating attractor models towards a general attractor model of delay discounting to enhance our understanding of the processes underlying delay discounting decisions.


Keywords: decision making; delay discounting; process dynamics, attractor dynamics; hysteresis; neural attractor model

## Introduction

Many everyday choices involve options that pose a conflict between immediate, but small gains, and delayed, but larger or more beneficial gains. This conflict occurs on many time scales. For example, you might wonder whether to enjoy spending your money now or saving it for a pension. Or you might be tempted to take the tasty pizza - which is immediately very tasty - instead of the light salad - which might be better for your cardiovascular system in the longterm. In such intertemporal choices (for a review, see Frederick, Loewenstein, \& O’Donoghue, 2002), humans discount the offered gain by the delay of delivery. This delay discounting is well described by utility discounting models which assume that the greater the delay in delivery of a reward, the more the utility of a reward is discounted. Hence, these discounting models represent the subjective value of a reward as a function of its delay (see Doyle, 2013 for an overview). While these mathematical models offer a good description of the average outcome of the decision process - the final choice - they mostly leave open how the exact decision process unfolds in time. Decoding this
process, though, is necessary in order to fully understand the way decisions are made. To fill this gap, recent developments aim to uncover the process dynamics leading to a final decision in delay discounting (Dai \& Busemeyer, 2014; Rodriguez, Turner, \& McClure, 2014; Scherbaum et al., 2016). Specifically, the attractor model approach (Figure 1) has recently been proved useful to uncover the process dynamics leading to a final choice outcome on different time scales, that is within and across sequential intertemporal choices (Scherbaum et al., 2016).

In this study, we will use the attractor model and the experimental paradigm as proposed by Scherbaum et al. (2016) to derive and validate qualitative predictions on the inter-trial dynamics of sequences of intertemporal decisions. More detailed, the attractor model of decision making in


Figure 1: Sketch of possible attractor layouts given different values of the control parameter c. This parameter depends on the relative difference in subjective value (attractiveness) of the options for a subject and hence configures the system for each potential combination of SS and LL: An increase in attractiveness for the LL option results in a negative control parameter which, in turn, increases the depth of the attractor representing the LL option (left panel). In contrast, an increase in attractiveness for the SS option results in a positive control parameter which, in turn, increases the depth for the attractor representing the SS option (right panel). Inherently, the control parameter $c$ is primarily dependent on the values and delays of the presented options, but also on a subject's tendency to discount. Within this potential landscape, the current system state (marked by a red dot) tends to move to the bottom of the potential wells and travels through all intermediate states on its way to a stable final choice.


Figure 2: Inter-trial dynamics in the attractor model. Choosing the LL option in a first trial leads to a bias in a second trial due to slow relaxation (e.g. inertia) of the system state during the inter trial interval (ITI, in this study 1.3 seconds).
delay discounting assumes that the depth of the attractors and hence the stability of its end-states is determined by a combination of each option's reward value and delay. By varying either both or one of those properties, the depth of the attractors can be manipulated. To that end, the depth of the attractors represents the relative attractiveness of each option, and that is, the respective subjective (discounted) value, within the attractor model (Figure 1). Hence, the difference in relative attractiveness between the two options determines the systems preference towards either option and is summarized by the control parameter, which we will call $c$. Figure 1 depicts three kinds of possible attractor layouts given three prototypical specifications of $c$. The attractors itself picture stable neural representations of the available options. So, the left and right panel of Figure 1 reflect almost exclusive activation of one option's representation, and hence illustrate configurations of the system with a preference towards one option $(c \neq 0)$. Accordingly, the special case where $\mathrm{c}=0$ (Figure 1 middle panel) reflects varying amounts of concurrent activation in which the system has not settled into a decision yet, and hence represents a decision in which both options receive an identical input and are thus equally attractive. In this special case scenario, a neutral starting state would keep the system indifferent until slight differences in input (or random noise) tips the system to one side or the other, resulting in a more or less arbitrary decision which was not driven by the systems preference. A major advantage of attractor models is that the decision is not only determined by the current attractor layout, which is in turn determined by the currently offered options, but also through the history of the system's previous decisions (Scherbaum, Dshemuchadse, \& Kalis, 2008; Townsend \& Busemeyer, 1989). This is due to the genuine assumption that the attractors are formed by the offered options and, hence, these attractors are not present between trials. The inertia of neural systems causes the system to temporally recline in the area where it ended up previously-in the vicinity of the vanished attractor representing the recent choice-and to relax only slowly to the neutral start state under no input. For example, if the
model chose one option in a first decision trial, it would remain in the vicinity of this option's attractor in the inter trial interval. In a second decision trial, it would hence start the decision with a bias to the previously chosen option, even if this trial comprises the other option being more attractive (see Figure 2).

Scherbaum et al. (2016) used this premise to predict and validate hysteresis effects (Tuller, Case, Ding, \& Kelso, 1994), which are also known as path-dependence, in intertemporal choice. Hysteresis or path-dependence occur, when the decision for one option biases the next decision in favor of the same option (see Figure 2). Hence, in a series of choices in which the initially unchosen option becomes increasingly more attractive (i.e. sequential manipulation of the difference in the relative attractiveness), people stick to the initially chosen option and switch to the now more attractive option much later than they would if their choices were unbiased. However, the sequential manipulation of the difference in the relative attractiveness was merely operationalized by variation of the delay, though the attractor model predicts the same hysteresis effects when the manipulation is realized through a variation of the reward value or even a combination of delay and value. We hence hypothesized that the emergence of hysteresis effects is independent from the attribute dimension which is used to sequentially manipulate the difference in the relative attractiveness between both options (intervals, value difference, or both together).

To provide an insight into hysteresis effects in delay discounting, we applied the same non-verbal delay discounting task as used in the original study. This task redresses the problem that in standard intertemporal choice tasks the sequential manipulation of reward values or delays is simply too obvious (Scherbaum et al., 2016; Scherbaum, Dshemuchadse, Leiberg, \& Goschke, 2013). In this task, subjects collect coins of different reward values with an avatar which they move on a checkered playing field by clicking with the computer mouse (Figure 3). The playing field stays constant across trials-except the options which change from trial to trial-and the avatar started each trial from the position of the previously chosen option. The goal is to collect as much reward as possible in the allotted amount of time. In each trial of the task, subjects have to choose between two reward options of different magnitude (small vs. large) at different distances (near vs. far fields). Therefore, this task translates delays into distances, which allows for a more implicit sequential manipulation of the relative attractiveness of options.

To implement the sequential manipulation of different attributes, we used this task in a modified, two-step procedure: In the first part, we measured the individual amount of discounting (the measurement block). Based on this amount of discounting, we created individually tailored sequences of decision to study hysteresis in the following part (the manipulation block). We expected the hysteresis effect to be present in all variants of sequential manipulations.

## Methods

## Subjects

43 students ( $65 \%$ female, mean age $=22.98$ years) of the Technische Universität Dresden took part in the experiment that lasted approximately 50 minutes. All subjects had normal or corrected to normal vision. Three out of 43 subjects were excluded from any subsequent analysis due to individual discounting behavior in the measurement block not allowing for a sufficient hysteresis manipulation in the manipulation block. Subjects gave informed consent to the study and received a $2.50 €$ show-up fee and the money they collected within the experiment $($ Mean $=3.17, \mathrm{SD}=0.39)$.

## Apparatus and Stimuli

The experiment was presented on a 17 -inch screen ( 1280 x 1024 pixels, 85 Hz ). As presentation software, we used Psychophysics Toolbox 3 (Brainard, 1997; Pelli, 1997) in Matlab 2010b (the Mathworks Inc.), running on a Windows XP SP2 personal computer. Responses were carried out by moving a high precision computer mouse (Logitech Laser Mouse USB).

Subjects moved an avatar on a playing field divided into $20 \times 20$ fields (Figure 3). To move the avatar, subjects clicked with the mouse in one of four horizontally or vertically adjacent movement fields, as signaled by a white border surrounding the fields. On each trial two reward options were presented as coins on fields marked with a red border: One reward was near but small, the other reward was far but large. The two options' positions were always chosen so that the first move into one direction decreased the distance to one option but increased the distance to the other option. This way, the first move of the avatar already represented a clear preliminary decision for one option and against the other option.

For both options, a number posed within each coin represented the reward value and the horizontal and vertical distance of the reward field to the field of the avatar represented the distance of the option. Reward values ranged from 1 to 99 credits and distances ranged from two to fifteen fields. For better comprehensiveness in the context of intertemporal choice, we maintain in the following the standard description of the time dimension using "soon",


Figure 3: Detail of the dynamic delay discounting paradigm.
"late", "delay", and "interval", although in our scenario time delay is represented by spatial distance. The relation between the two reward values can be characterized as the ratio of the higher and smaller reward value and will be denoted by "difference".

Above the avatar (Figure 3) subjects could see the remaining time within one block, as well as below the collected credits in Euro ( 1 credit $=1 / 10 €$ cent), but only in the very moment when either reward was collected.

## Procedure

Subjects' task was to collect as much reward as possible within the allotted time limit. In each trial, they had to choose between two reward options (one soon but small, $S S$, one late but large, $L L$; see design). They collected the selected reward by moving their avatar with the mouse across the playing field.

A trial started with an inter trial interval (ITI) of 1.3 seconds. Within this interval, the mouse cursor was locked in the center of the field containing the avatar. After the ITI, the two options were presented. As soon as the two options appeared, participants could click on the adjacent movement fields to move their avatar towards the chosen option (Figure 3). When the avatar reached one option, both options disappeared, the value of the selected option was added on the collected credits, and the next trial started.

The experiment consisted of four blocks, with one block lasting eight minutes. Between blocks, subjects were informed about the credits collected and were instructed to rest briefly before the self-paced start of the next block.

Before the start of the experimental blocks, subjects worked through a test block of two minutes to get used to the virtual environment, handling of the mouse, as well as the range of spatial distances and reward values.

## Design

The experiment consisted of four blocks with the first block (measurement block) being conceptually different from the three subsequent blocks as its aim was to measure the subjects' individual discounting behavior. In each of the three subsequent blocks (manipulation blocks) we realized a unique adaptive hysteresis manipulation constituting an interval block, a value block, and a combined block. Each subject's session started with the measurement block followed by the manipulation blocks. The sequential arrangement of the manipulation blocks was fully varied and balanced between subjects.

In the measurement block, reward values ranged from 11 to 99 and distances from three to 15 . That was given by orthogonally varying the intervals ( $1,4,8$, and 12 fields), the differences $(20,50,70,80,88,93,97$, and $99 \%)$, and the delay of the sooner option ( 2 and 3 fields). Additionally, the reward values of the late option were randomly chosen from a discrete uniform distribution between 55 and 99 credits. The combination of 8 differences, 2 distances of the SS options and 4 intervals between the SS and the LL option yielded a complete set of 64 trials. We generated 5 such
sets, with a randomized order of trials within each set. The measurement block's time limit ensured that subjects could work through the complete design matrix, that is one of those 5 sets, at least one time.

To realize the adaptive hysteresis manipulation, we calculated the subjects' individual discounting curve from which we adaptively derived trials compatible to the respective hysteresis manipulation (see Results). The structure of an adaptive hysteresis manipulation is to sequentially change subjects' preference from the SS towards the LL option, or vice versa. In our adaptive trial sequences, we aimed to change subjects' preference in 12 steps as indicated by the differences between the subjective value ratio (SS/LL) in the trials and the indifference points $(-0.3000,-0.2455,-0.1909,-0.1364,-0.0818,-0.0273$, $0.0273,0.0818,0.1364,0.1909,0.2455,0.3000)$, that is the manipulation points. ${ }^{1}$ It is imperative that a negative manipulation point indicates a preference for the LL option and a positive manipulation point a preference for the SS option. Furthermore, it applies that the higher the absolute manipulation point, the more distinct are the relative attractiveness of both options. Hence, a manipulation point of zero represents no preference, that is the indifference point. Please note that the interpretation of the manipulation point is analog to the interpretation of the control parameter.

We then applied this manipulation in three different subblocks. First, in the interval block we consecutively increased or decreased the delay of the LL option to the avatar while keeping all other factors constant within the sequence. For each sequence the delay of the sooner option and the reward value of the late option were randomly chosen from discrete uniform distributions between 2 and 3 fields, and 55 and 99 credits, respectively. The reward value of the sooner option was randomly drawn from the uniform distribution between subjects' two indifference points at the intervals 6 and 7. Furthermore, we varied the direction of these sequences (direction $=$ ascending or descending) and created eight sequences for each direction. This resulted in 16 possible sequences, and hence 192 trials.

Second, in the value block we consecutively increased or decreased the reward value of SS option while keeping all other factors constant within the sequence. Again, for each sequence, the delay of the sooner option and the reward value of the late option were randomly chosen. The delay of the LL option to the avatar was drawn randomly between all intervals at which subjects' indifference point was positioned in such a way that all 12 manipulation points were valid, that is, did not exceed 1 or fall below a value of 0 . For each trial within the sequence, the reward values of the SS option were then calculated. Again, we also varied the direction of these sequences and created eight sequences

[^194]

Figure 4: Subjects' indifference points in the measurement block, depicting the decrease in subjective value of the latelarge option as a function of intervals between the two options. Indifference points are the subjective value ratio (SS/LL) at which subjects chose indifferently between the two options, i.e., the probability of choosing LL over SS is $50 \%$. Note: Error bars indicate standard errors. The curve displays the fitted hyperbolic functions.
for each direction. This resulted in 16 possible sequences, and hence 192 trials.

Third, in the combined block we consolidated the former manipulations and varied both the delay of the LL option to the avatar and the reward value of the SS option in such a way that the manipulation points consecutively increased or decreased within the sequence. Again, for each sequence the delay of the sooner option and the reward value of the late option were randomly chosen. For each trial within the sequence, the delay of the LL option was randomly chosen from the set of intervals in which the respective manipulation point was valid. The reward values of the SS option were then calculated. Again, we also varied the direction of these sequences and created eight sequences for each direction. This resulted in 16 possible sequences, and hence 192 trials.

In sum, we applied a 2 (direction: ascending, descending) x 3 (manipulation type: interval, value, combined) full factorial within-subjects design.

## Results

On average, subjects completed 134 trials $(S D=23)$ in the measurement block. Hence, subjects ran through at least two out of five sets of 64 trials. The aim of the measurement block was to measure subjects' individual discounting behavior indicated by subjects' indifference points as depicted by Figure 4. As an estimate of the indifference point, the point of inflection of a logistic function was fitted to the individual choices as a function of increasing value


Figure 5: Average hysteresis plots between manipulation types. Plots depict subjects' mean response pattern over intervals (panel a) or manipulation points generated by variation of rewards only (panel b) and a combination of rewards and intervals (panel c). Note: Error bars indicate standard errors. The separate colors indicate whether mean responses were derived from ascending or descending sequences. The blue line represents descending sequences ( $\mathrm{LL} \rightarrow \mathrm{SS}$ ). The red line represents ascending sequences $(\mathrm{SS} \rightarrow \mathrm{LL})$.
differences was determined. ${ }^{2}$ To evaluate subjects' discounting behavior in one parameter, we extracted the $k$ parameter by fitting a hyperbolic function to each subject's indifference points over the different intervals. Data revealed an average $k$-parameter of the hyperbolic discounting curve with $M(S D)=0.23(0.19)$, bootstrapped $95 \% C I=[0.18,0.30]$, indicating a very strong discounting behavior. The hyperbolic model had a good fitting

[^195]performance over all subjects, indicated by a high average $R^{2}, M(S D)=.87(.10) .{ }^{3}$

In the manipulation blocks, subjects completed 387 trials ( $S D=67$ ) on average. Hence, on average, subjects ran through 32 hysteresis sequences ( $S D=6$ ), consisting 16 ascending $(S D=3)$ and 16 descending $(S D=3)$ sequences. The SS option was chosen in $48.37 \% ~(S D=22.19)$ of the trials, indicating only a slight decision bias which was not predicted by the model.

The core prediction of the model was that subjects show identical hysteresis effects irrespective of the specific hysteresis manipulation, that is, whether the sequential manipulation of the attractiveness of both options was realized through varying intervals, differences or a combination of both. Figure 5 depicts the hysteresis effect for each manipulation type. The plots indicate that the hysteresis effects are very similar between manipulation types, but show the qualitatively best pattern for the interval manipulation (Figure 5, panel a). In order to test model's predictions, we conducted a two-factorial Repeated Measures ANOVA (direction $x$ manipulation type) on subjects' mean choice. As expected, we solely found a main effect of direction $\left(F(1,39)=17.44, p<.001, \eta^{2}=0.31\right)$, indicating that hysteresis emerged irrespectively of manipulation types. Thus, neither the main effect of manipulation type $\left(F(2,78)=3.03, p=.054, \eta^{2}=0.07\right)$ nor the interaction $\left(F(2,78)=1.22, p=.302, \eta^{2}=0.03\right)$ were statistically significant. In order to focus the analysis on the hysteresis effect, that is, eliminating the variance of the absolute level of LL choices, we summarized hysteresis effects into one hysteresis parameter. The hysteresis parameter was given by calculating the differences between subjects' mean choice in ascending and descending hysteresis sequences for each manipulation type. An additional one-factorial Bayesian Repeated Measures ANOVA on the hysteresis parameter revealed that the data show substantial evidence in favor of the null hypothesis $\left(\mathrm{BF}_{01}=4.64\right)$ claiming that the hysteresis effect does not vary systematically between all three manipulation types. Therefore, we consider the predictions of the model as confirmed.

## Discussion

In this study, we tested predictions of the attractor model of delay discounting in a recent developed non-verbal delay discounting paradigm. Our results validated the model in such a way that its predictions concerning hysteresis effect in delay discounting were confirmed. Specifically, when sequentially varying the attractiveness of both options from a very strong preference towards the SS option to a very strong preference towards the LL option, and vice versa, hysteresis effects occur irrespectively of how the

[^196]attractiveness of any option is varied within the sequence. Therefore, the current study both replicated and added empirical evidence for the validity of the attractor model of delay discounting.

One might object that the predictions of the model were merely derived through a qualitative, argumentative manner. This is obviously true, but not a weakness of the current study. First, concerning the interval manipulation, it was already shown that the exact same predictions can be derived by means of computational simulation based on a competitive neural-network (Scherbaum et al., 2016), hence running a computational simulation with the same model would not provide any new information. Second, and this point is genuine, the model does not allow for reasonable separate simulations of all manipulation types. This is due to the fact that the model merely uses subjective values for each option. The emergence of those subjective values, however, is not covered within the model.

Leaving the emergence of subjective values open the model proves to be useful for predicting intra- and inter-trial dynamics in delay discounting, when a specific discounting function is already given, but it does not explain the emergence of discounting functions. This gap has also been argued for recently by others, reasoning that intertemporal choice consists of two processes (Rodriguez et al., 2014): First, the process of delay discounting, and second, the process of choice. This gap between the two processes could be closed by connectionist models, which have already been used to explain how different discounting functions emerge by linking discounting behavior with aspects of self-control (Scherbaum, Dshemuchadse, \& Goschke, 2012).

The two models provide insights into the dynamics of delay discounting and the dynamics of choice, respectively. Integrating these two models into one general connectionist model of delay discounting could provide insights into the interacting process dynamics of preference (delay discounting) and choice. Such an integration could therefore enhance our understanding of the processes underlying delay discounting decisions and, hence, complement our knowledge about decision outcomes.

## Acknowledgments

This research was partly supported by the German Research Council (DFG) (grant SFB 940/2 2016, project A8). We would like to thank Paula Marie Henneberg, Judith Herbers, Marie Gotthardt, and Natalja Schmidt for their efforts in data collection and management. We gratefully acknowledge helpful comments on this article by Ulrike Senftleben.

## References

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433-436. http://doi.org/10.1163/156856897X00357
Dai, J., \& Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. Journal of Experimental Psychology: General,

143(4), 1489-1514. http://doi.org/10.1037/a0035976
Doyle, J. R. (2013). Survey of Time Preference, Delay Discounting Models. Judgment and Decision Making, 8(2), 116-135. http://doi.org/10.2139/ssrn. 1685861
Frederick, S., Loewenstein, G., \& O’Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. Journal of Economic Literature, 40(2), 351401.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spatial Vision, 10(4), 437-442. http://doi.org/10.1163/156856897X00366
Rodriguez, C. A., Turner, B. M., \& McClure, S. M. (2014). Intertemporal choice as discounted value accumulation. PLoS ONE, 9(2). http://doi.org/10.1371/journal.pone. 0090138
Scherbaum, S., Dshemuchadse, M., \& Goschke, T. (2012). Building a bridge into the future: Dynamic connectionist modeling as an integrative tool for research on intertemporal choice. Frontiers in Psychology, 3(NOV), 1-14. http://doi.org/10.3389/fpsyg.2012.00514
Scherbaum, S., Dshemuchadse, M., \& Kalis, A. (2008). Making decisions with a continuous mind. Cognitive, Affective, \& Behavioral Neuroscience, 8(4), 454-474. http://doi.org/10.3758/CABN.8.4.454
Scherbaum, S., Dshemuchadse, M., Leiberg, S., \& Goschke, T. (2013). Harder than Expected: Increased Conflict in Clearly Disadvantageous Delayed Choices in a Computer Game. PLoS ONE, 8(11), e79310. http://doi.org/10.1371/journal.pone. 0079310
Scherbaum, S., Frisch, S., Leiberg, S., Lade, S. J., Goschke, T., \& Dshemuchadse, M. (2016). Process dynamics in delay discounting decisions: An attractor dynamics approach. Judgement and Decision Making, 11(5), 472-495.
Townsend, J. T., \& Busemeyer, J. R. (1989). Approachavoidance: Return to dynamic decision behavior. In C. Izawa (Ed.), Current Issues in Cognitive Processes: The Tulane Flowerree Symposion in Cognition. NJ: Lawrence Erlbaum.
Tuller, B., Case, P., Ding, M., \& Kelso, J. A. S. (1994). The nonlinear dynamics of speech categorization. Journal of Experimental Psychology. Human Perception and Performance, 20(1), 3-16.
http://doi.org/10.1037/0096-1523.20.1.3

# Strategic exploration in human adaptive control 

Eric Schulz ${ }^{1}$, Edgar D. Klenske ${ }^{2,3}$, Neil R. Bramley ${ }^{4}$ \& Maarten Speekenbrink ${ }^{1}$<br>${ }^{1}$ Department of Experimental Psychology, University College London, London, WC1H0AP, UK<br>${ }^{2}$ Max Planck Institute for Intelligent Systems, Spemannstraße 38, 72076 Tübingen, Germany<br>${ }^{3}$ Max Planck ETH Center for Learning Systems, ETH Zurich, 8092 Zurich, Switzerland<br>${ }^{4}$ Department of Psychology, New York University, New York, NY, 10003, USA


#### Abstract

How do people explore in order to gain rewards in uncertain dynamical systems? Within a reinforcement learning paradigm, control normally involves trading off between exploration (i.e. trying out actions in order to gain more knowledge about the system) and exploitation (i.e. using current knowledge of the system to maximize reward). We study a novel control task in which participants must steer a boat on a grid, aiming to follow a path of high reward whilst learning how their actions affect the boat's position. We find that participants explore strategically yet conservatively, exploring more when mistakes are less costly and practicing actions that will be required later on.


Keywords: Reinforcement Learning, Strategic Exploration, Control, Exploration-Exploitation

## Introduction

Deciding how to act under uncertainty is a core problem for cognition. Cognitive agents must be able to navigate a world whose dynamics are initially unknown and generally uncertain, learning to generate rewards as they go along. In the context of reinforcement learning, we can think of control as a trade-off between exploration (i.e. trying out actions to gain more knowledge about the underlying system) and exploitation (i.e. using current knowledge of the system to maximize reward). However, whether and how human explorative control reflects future goals and current uncertainty is still unclear (Wilson, Geana, White, Ludvig, \& Cohen, 2014). Are human explorative actions strategic and goal-directed? Or are they rather passive, for instance involving a simple "exploration bonus" that treats uncertainty equally across all actions?

Traditionally, reinforcement learning models have addressed exploration rather implicitly, letting the agent learn about the underlying system en passant via outcomes produced while high rewards are chased (Rescorla, Wagner et al., 1972). Exploration, according to this definition, is what happens when an agent optimizes noisily. We will refer to this kind of exploration as passive exploration.

More recently, exploration has been incorporated into reinforcement learning models more explicitly via an exploration bonus (Schulz, Konstantinidis, \& Speekenbrink, 2016; Wu, Schulz, Speekenbrink, Nelson, \& Meder, 2017). An exploration bonus assigns additional utility to less explored actions and thereby assumes that the agent values uncertainty equally across all actions. Exploration, according to this definition, is what happens when expectations are inflated by their attached uncertainties. We will refer to this kind of exploration as $a g$ nostic exploration.

Another line of research tries to redefine exploration as goal-directed behavior (e.g., Thrun, 1992). The idea behind
this approach is that not all uncertainty should be treated equally but rather that exploration should be driven by both the current knowledge of the system and the agent's overall goal. Exploration, according to this definition, is a strategic action. We will refer to this kind of exploration as strategic exploration.

Many real-world control scenarios are non-episodic, such that there are no "second chances" and one may be unable to return to known states. In such situations, one must treat exploration strategically and with great caution to avoid accidents (Klenske \& Hennig, 2015). Imagine visiting a country with left-hand traffic from a country with right-hand traffic. Strategically exploring in order to learn how to drive on the left side could allow you to make your mistakes on the quiet roads first before hitting the highway. Moreover, as turning right will be more difficult than you are used to, practicing how to turn right is more important than practicing how to turn left and therefore should be exercised more frequently.

In machine learning, problems of planning under uncertainty have been approached via Bayesian reinforcement learning (BRL; Poupart, 2010), which assigns probabilistic beliefs over the dynamics of a system and the costs of states and actions in order to reason about potential changes to beliefs from future observations, and their influence on future decisions (Duff, 2002). BRL provides a useful framework for assessing strategic exploration behavior as we do here. More specifically, we will make use of the duality between reinforcement learning and control, that is tasks in which an agent has to keep a system at a certain state in order to generate rewards (Feldbaum, 1960; Klenske \& Hennig, 2015).

In what follows, we will assess how participants exert control within a novel control paradigm. Therein, strategic exploration allows them to produce greater long-term rewards formally, within a non-episodic, finite-horizon system with initially unknown dynamics. We will build on recent work by Klenske \& Hennig (2015) and assess behavior in two tasks: one in which, due to time-varying state costs, exploration can be delayed until it is more opportune; and one in which the learning agent can distinguish between more and less important exploration of directional actions. We first discuss in more detail the task and the three perspectives on exploration in control theory passive,agnostic and strategic exploration. We then assess qualitative predictions derived from these in two experiments. We find participants' behavior to be more in line with predictions derived by strategic exploration.

## Control task

In a simple computer game, participants have to navigate a boat as it crosses a sea towards regions in which they can earn a higher bonus. The boat moves incrementally from left to right and by changing its current angle of direction (see Figure 1), participants could attempt to steer the boat up or down, so as to remain in calm waters (blue) and avoid perilous rough seas (red). The overall goal of the game was to min-


Figure 1: Example path in Experiment 1, Free Late condition. Star $=$ starting position at $t=0$, circles $=$ subsequent positions. On each trial, the gray arrow shows contribution of underlying current and black arrow contribution of control angle. At $t=0$ the participant takes a control angle of 0 and drifts upward. On the 10 subsequent trials they attempt to counteract this upward drift by setting a negative angle. During the free exploration stage they use wider angles to explore the variation in the strength of the current at different $y$ positions. This allows them to discover that strength of the current is strongest in the center, approaching zero toward the top, constantly pulling the boat upwards.
imize the "cost" of the voyage while simultaneously learning both how to control their boat and about an underlying position-dependent "current" that drags the boat off course. In some periods, the area of low cost was very narrow, while in other periods, the area was very wide. Analogous to real sailing, participants had to learn to control the boat through experience, by trying different angles and observing the effect on the boat's position. This exploration is costly when the low-cost region is narrow, whilst exploration is almost "free" when the low-cost region is very wide.

Our control task is adapted from Klenske \& Hennig (2015). Therein, the boat is influenced by two factors, its current position $x_{1}$ and an underlying current $x_{2}$. This means that where the boat will end up on the next trial is influenced by the chosen angle, its current position, and the underlying current which is determined by an unknown nonlinear function. Within our experiments, the underlying current decreased from its full strength in the middle of the sea to zero at the upper and lower edges, and constantly pulled the boat upwards. For example, if participants entered the angle of 0 in the center of the sea, the boat would be pulled upwards
more than if they entered the same angle at another position further up. Formally, at each time $t$, the (vertical) position of the boat, $y_{t}$, depends on a two-dimensional latent state variable $x_{t}$ and independent random noise $\gamma_{t}$ as

$$
\begin{equation*}
y_{t}=\mathbf{C x}_{t}+\gamma_{t} \quad \gamma_{t} \sim \mathcal{N}\left(0, \sigma_{\gamma}\right) \tag{1}
\end{equation*}
$$

The latent state depends through a nonlinear function on the previous latent state, the controller input (i.e., the chosen angle) $u_{t}$, and additional noise $\xi_{t}$, as:

$$
\begin{equation*}
\mathrm{x}_{t+1}=\mathrm{A} \phi\left(\mathrm{x}_{t}\right)+\mathrm{B} u_{t}+\xi_{t} \quad \xi_{t} \sim \mathcal{N}\left(0, \Sigma_{\xi}\right) \tag{2}
\end{equation*}
$$

where

$$
\mathbf{A} \phi\left(\mathbf{x}_{t}\right)=\left[\begin{array}{cc}
1 & 0.4  \tag{3}\\
0 & 1
\end{array}\right]\left[\begin{array}{l}
x_{1, t} \\
x_{2, t}
\end{array}\right]+\left[\begin{array}{cc}
0 & 0 \\
\theta_{1} & \theta_{2}
\end{array}\right]\left[\begin{array}{c}
\frac{1}{1+e_{1, t}} \\
\frac{1}{1+e^{-x_{1, t}+5}}
\end{array}\right]
$$

$\theta=[0.8,0.4]^{\top}$, and $\mathbf{B}=[0,1]^{\top}$. The underlying drift is determined by the shifted sigmoid functions on the right-hand side of Equation 3. Given a finite-time horizon with terminal time $T$, the following quadratic cost function was used:

$$
\begin{equation*}
\mathcal{L}(\mathbf{x}, \mathbf{u})=\sum_{t=0}^{T}\left(\mathbf{x}_{t}-\mathbf{r}_{t}\right)^{\top} \mathbf{W}_{t}\left(\mathbf{x}_{t}-\mathbf{r}_{t}\right) \tag{4}
\end{equation*}
$$

where $\mathbf{r}=\left[\mathbf{r}_{0}, \ldots, \mathbf{r}_{T}\right]$ is the target trajectory and $\mathbf{W}_{t}$ the timevarying state cost. The goal of the controller is to find the action sequence $\mathbf{u}=\left[u_{0}, \ldots, u_{T}\right]$ that minimizes the expected cost (and thereby maximizes the expected reward) to the horizon $T$.

## Control strategies

Controlling a system as defined in Eqs. (1) - (3) is difficult, as the state dynamics are nonlinear with an unknown function $\phi$ and parameters $\mathbf{A}$ and $\mathbf{B}$. The controller then not only needs to control the states in accordance to the reference path $\mathbf{r}$, but also learn the parameters (and functions) in order to derive a good control strategy $\mathbf{u}$. Thus, the controller not only needs to control the states, but also control her knowledge about the model, hence the term dual control.

We will now provide a description of the three different forms of exploration mentioned earlier, and their qualitative predictions in the present control task. The predictions are shown graphically in Figure 2 for the variants of the task used in Experiment 1, which tests whether participants will postpone exploration until it is most opportune, and Experiment 2, which tests whether participants perform strategic (directional) exploration.

## Passive exploration by certainty equivalence

A certainty equivalence controller completely ignores uncertainty about the dynamics and derives a control strategy as if the current (mean) estimates of the system are accurate and knowledge about the system is perfect. Effectively, any learning about the system happens passively, as the control strategy does not focus on minimizing uncertainty. As no active

A


B


C


Figure 2: Control environments and qualitative model predictions for Experiment 1 and 2. The agent moves one step right each time point (trials are delimited by white vertical lines) and can control the angle upwards/downwards in which a boat is steering. The background color represent the cost function; the more red, the lower the score; dark blue areas mark free exploration trials. Qualitative model predictions taken from Klenske \& Hennig (2015) are represented by horizontal line called 'predictive regions'. Black lines represent the predictive region for passive exploration, white lines for agnostic exploration, and purple lines for strategic exploration. The more space between the horizontal lines at a trial, the wider the region and the higher the expected variance of the controller's actions.
A: Strategic exploration holds off exploration until it comes at lower cost (broader trust region during the dark blue patch) and consequently performs better than passive exploration later on (narrower trust region).
B: If free exploration phase is moved to the end, strategic exploration explores less overall and expedites exploration to earlier, more costly stages, thereby reducing performance early on in order to achieve the best performance later on.
C: Instead of agnostically exploring both directions in the same way, strategic exploration uses the free exploration phase to try to move in a trajectory which is rewarding in the future, thereby performing better later on.
exploration is encoded into this model, it might miss out on important information that could be beneficial to produce better rewards later on. This form of control predicts no exploration, even when exploration is 'free' and beneficial to future rewards.

## Agnostic exploration by exploration bonus

To promote exploration, a straightforward adaptation of the certainty equivalent controller is to introduce a Bayesian exploration bonus. Effectively, this means adapting the cost function so that the costs of actions which reduce uncertainty in the model of the control dynamics-as measured by the standard deviation of the posterior distribution over the parameters at each observation point-is temporarily reduced (cf. Srinivas, Krause, Kakade, \& Seeger, 2009). This model is still myopic as it only considers uncertainty at the current control step. Moreover, exploration is not strategic, as all uncertainty is treated equally and it does not take into account what knowledge might be most important later on. Under agnostic exploration, the expected behavior would be the attempt to identify all uncertain components, irrespective of their future usefulness.

## Strategic exploration as dual control

BRL involves reasoning about the effect of actions on future rewards and beliefs. Where an exploration bonus renders reducing uncertainty rewarding in itself, in BRL, reducing uncertainty is only attractive insofar as it is expected to result in an increase in future rewards. Optimal BRL requires determining the consequences of strengthening beliefs on future rewards, thereby finding the optimal balance between exploration and exploitation. Unfortunately, the optimal solution to the dual control problem of simultaneously controlling the
system as well as possible given current knowledge (exploitation) and learning about the system through experimentation in order to control it better later on (exploration), is known to be intractable.

Approximate dual control involves three conceptual steps which together yield what, from a contemporary perspective, amounts to an approximate solution to Bayesian RL: First, determine the optimal trajectory under the current mean model of the system (as in certainty equivalent control). Second, construct a local quadratic expansion around the nominal trajectory that approximates the effects of future observations. Third, within the current time step $t$, perform the prediction for an arbitrary control input $u_{t}$ and optimize $u_{t}$ numerically by repeated computation of steps 1 and 2 at varying $u_{k}$ to minimize the approximate cost (see Klenske \& Hennig, 2015, for implementation). Approximate dual control does not treat all exploration equally but rather explores strategically by, for example, holding off exploration until it is less costly or by exploring actions that will become important later on.

## Experiment 1: Holding off exploration

Our first experiment was designed to test passive exploration against both agnostic and strategic exploration by including a low-cost period which was either introduced relatively early ("Free Early" condition) or at the end of the task ("Free Late" condition). When a low-cost period is introduced early, controllers can make good use of it to explore and better their performance in later periods, while exploring in a low-cost period at the end of the task is not beneficial as there are no later rewards to reap.

Both conditions experienced an initial stage of medium state costs (see Figure 2). However, whereas for the Free

Early condition that stage is followed by a stage of free exploration (no costs of errors) which then leads to a stage of very high cost, the Free Late condition experiences the stage with high state costs first before then experiencing the stage with no costs (the two stages are swapped).

We expected participants to behave as strategic controllers and to initially hold off exploration in the Free Early condition until it comes at no cost in the low-cost period, allowing them to be prepared for the most difficult final stages. In contrast, participants in the Free Late condition were expected to explore more in the initial period, in order to be prepared for the second, most difficult stage. In addition, we expected participants in the Free Late condition to explore less in the low-cost period compared to those in the Free Early condition, as late exploration no longer brings benefits if the task is nearly over. Finally, we expected participants in the Free Early condition to generally perform better than participants in the Free Late group, as early exploration would enhance their knowledge of the system for the remainder of the task.

## Design

The manipulation involved changing the order of the reference trajectory (the state values that would produce the highest rewards) and state weightings.

In the Free Early-condition the reference trajectory and state weightings were:

$$
\begin{aligned}
\mathbf{r}_{1: 22} & =\left[\begin{array}{l}
0 \\
0
\end{array}\right] \quad \mathbf{r}_{23: 28}=\left[\begin{array}{l}
7 \\
0
\end{array}\right] \mathbf{r}_{29: 30}=\left[\begin{array}{l}
0 \\
0
\end{array}\right] \\
\mathbf{r}_{31: 36} & =\left[\begin{array}{c}
-7 \\
0
\end{array}\right] \mathbf{r}_{37: 40}=\left[\begin{array}{l}
0 \\
0
\end{array}\right]
\end{aligned}
$$

and

$$
\mathbf{W}_{1: 10}=\left[\begin{array}{ll}
1 & 0 \\
0 & 0
\end{array}\right] \quad \mathbf{W}_{11: 20}=\left[\begin{array}{ll}
0 & 0 \\
0 & 0
\end{array}\right] \quad \mathbf{W}_{21: 40}=\left[\begin{array}{cc}
10 & 0 \\
0 & 0
\end{array}\right]
$$

In the Free Late condition, these were:

$$
\begin{aligned}
& \mathbf{r}_{1: 12}=\left[\begin{array}{l}
0 \\
0
\end{array}\right] \mathbf{r}_{13: 18}=\left[\begin{array}{l}
7 \\
0
\end{array}\right] \mathbf{r}_{19: 20}=\left[\begin{array}{l}
0 \\
0
\end{array}\right] \\
& \mathbf{r}_{21: 26}=\left[\begin{array}{c}
-7 \\
0
\end{array}\right] \mathbf{r}_{27: 40}=\left[\begin{array}{l}
0 \\
0
\end{array}\right]
\end{aligned}
$$

and

$$
\mathbf{W}_{1: 10}=\left[\begin{array}{ll}
1 & 0 \\
0 & 0
\end{array}\right] \quad \mathbf{W}_{11: 30}=\left[\begin{array}{cc}
10 & 0 \\
0 & 0
\end{array}\right] \quad \mathbf{W}_{31: 40}=\left[\begin{array}{ll}
0 & 0 \\
0 & 0
\end{array}\right]
$$

## Materials

Participants were told that they had to navigate a boat through the sea in a sailing competition. On every trial, their boat was at a current position $\left(y_{t}\right)$ and they had to determine an angle ( $u_{t}$, between $-180^{\circ}$ and $180^{\circ}$ ) in which they wanted to sail.

Additionally, they had different target areas $\left(r_{t}\right)$ on each trial marked by dark blue colors and how far they were off from the target area was penalized differently (based on $\mathbf{W}_{t}$ ). An example trial from the task (for the Free Early condition) is depicted in Figure 1.

The cost function was shown to participants through the color of each position in the sea. Participants could earn between 0 (positions with a red background) and 100 points (positions with a blue background) per trial.

## Participants

Sixty-one participants were recruited via Amazon Mechanical Turk and received $\$ 1$ and a bonus of up to $\$ 1$. Thirty-nine participants were male and the mean age was $31.3 \pm 8.4$.

## Results

The distribution of boat position, as well as average chosen angles, are depicted in Figure 3. We can see that, overall,


Figure 3: Boat positions by condition in Experiment 1. The heat map reflects number of participants who were at that position on a given trial. Error bars represent the standard error of the average position per trial. Arrows indicate the average chosen angle. Black dots mark the target trajectory and periods with different state weights are delimited by vertical lines.
participants managed to steer the boat reasonably well. A linear regression of condition, cost function weights (coded as dummy variables), and trial number onto participants' scores (see Table 1) showed that, unsurprisingly, cost function
weights had the largest effect on participants' scores. Moreover, performance increased significantly over trials. Importantly, condition affected overall performance, such that participants in the Free Early condition performed better than participants in the Free Late condition. This confirms the hypothesis that participants would benefit from early free exploration.

Table 1: Regression estimates for Experiment 1. $r^{2}=0.38$.

|  | Estimate | Std. Err. | t value | $\operatorname{Pr}(>\|\mathrm{t}\|)$ |
| ---: | ---: | ---: | ---: | ---: |
| Intercept | 99.9 | 2.76 | 36.2 | 0.000 |
| Condition | -3.04 | 1.07 | -2.84 | 0.004 |
| Med. cost: 1 | -15.7 | 2.13 | -7.39 | 0.000 |
| High cost: 10 | -47.3 | 1.30 | -36.3 | 0.000 |
| Trial | 0.18 | 0.07 | 2.61 | 0.009 |

Another hypothesis was that participants in the Free Early condition would explore more during the free exploration stage than participants in the Free Late condition. Confirming this hypothesis, the participant-wise variance of chosen angles during the free exploration stage was significantly larger for the Free Early condition than for the Free Late condition $(t(59)=2.62, p<0.01)$. As such, participants indeed seemed to strategically adapt their exploration behavior to the underlying cost function.

While we expected participants in the Free Late condition to explore more in the initial stage of medium difficulty than those in the Free Early condition, a similar test to the one above did not confirm this $(t(59)=0.63, p>0.5)$. As such, there is no clear evidence that participants in the Free Late condition used the medium difficulty period to explore in order to perform better in the high-difficulty period.

Overall, participants in Experiment 1 showed hallmarks of strategic exploration. However, they did not explore as vigorously as approximate dual control predicted, often only doing so during completely free exploration periods. As soon as exploration is somewhat costly, participants seem to shift focus back to normal (perhaps certainty-equivalence based) control, thereby more conservatively trading off between exploration and exploitation.

## Experiment 2: Directional exploration

The second experiment was designed to distinguish between agnostic and strategic exploration, involving the explicit exploration of directional actions. The design was again based on ideas put forward by Klenske \& Hennig (2015). In both conditions, a free exploration phase was followed by a high difficulty period, in which controllers either had to move the boat first up then down again (Up-Down condition) or first down and then up again (Down-Up condition).

If exploration is indeed strategic rather than agnostic and simply based on an exploration bonus, then participants in the Up-Down condition should focus exploration in the free exploration phase on first learning to travel precise increments
upwards and then precise increments downwards, whereas participants in the Down-Up condition should explore to do the opposite, as knowledge about these actions will be useful later on. Note that mimicking the later target trajectory during the free exploration phase is better then trying upwards and downwards movements at one position as the current, and with that the effect of a chosen angle on the position, varies nonlinearly depending on the boat's position.

## Design

The underlying dynamics were exactly the same as in Experiment 1 . The manipulation solely concerned the reference trajectory, which for the Up-Down condition was:

$$
\begin{aligned}
\mathbf{r}_{1: 23} & =\left[\begin{array}{l}
0 \\
0
\end{array}\right] \mathbf{r}_{24: 26}=\left[\begin{array}{l}
3 \\
0
\end{array}\right] \mathbf{r}_{27: 29}=\left[\begin{array}{l}
5 \\
0
\end{array}\right] \\
\mathbf{r}_{30: 32} & =\left[\begin{array}{l}
7 \\
0
\end{array}\right] \mathbf{r}_{33: 35}=\left[\begin{array}{l}
5 \\
0
\end{array}\right] \mathbf{r}_{36: 38}=\left[\begin{array}{l}
3 \\
0
\end{array}\right] \mathbf{r}_{39: 40}=\left[\begin{array}{l}
0 \\
0
\end{array}\right]
\end{aligned}
$$

And for the Down-Up condition, the reference trajectory was:

$$
\mathbf{r}=-\mathbf{r}
$$

The state weighting was the same for both groups:

$$
\mathbf{W}_{1: 2}=\left[\begin{array}{ll}
1 & 0 \\
0 & 0
\end{array}\right] \quad \mathbf{W}_{3: 21}=\left[\begin{array}{ll}
0 & 0 \\
0 & 0
\end{array}\right] \quad \mathbf{W}_{22: 40}=\left[\begin{array}{cc}
10 & 0 \\
0 & 0
\end{array}\right]
$$

## Materials

Participants were again told that they were taking part in a sailing contest. Participants in the Up-Down condition were then shown the control environment sketched out in Figure 2 (right panel), whereas participants in the Down-Up condition experienced the same control environment but flipped around the center horizontal axis.

## Participants

Forty-six participants were recruited via Amazon Mechanical Turk and received $\$ 1$ and a bonus of up to $\$ 1.16$ participants were female and the mean age was $34.32 \pm 11.17$.

## Results

Figure 4 shows participants' boat position by group. Again, participants seemed to be able to learn how to steer the boat towards its targets in both groups. As before, we performed a linear regression of the weights, trials and condition onto participants' score (see Table 2).

The weights had again the largest effect on participants' scores and participants' scores improved over time. There was no significant difference between the scores of the two conditions.

Strategic exploration is visible in the Up-Down condition as participants' mean position goes up and then down again, thereby showing clear signs of practicing the route to come. This can also be found by testing the difference between condition's average position during times of free exploration,


Figure 4: Boat positions by condition for Experiment 2. See legend of Figure 3 for further details.

Table 2: Regression estimates for Experiment 2. $r^{2}=0.45$.

|  | Estimate | Std. Err. | t -value | $\operatorname{Pr}(>\|\mathrm{t}\|)$ |
| ---: | ---: | ---: | ---: | ---: |
| Intercept | 97.8 | 2.72 | 35.9 | 0.000 |
| Condition | 0.23 | 1.75 | 0.13 | 0.89 |
| Med. cost: 1 | -11.3 | 1.69 | -6.56 | 0.000 |
| High cost: 10 | -26.6 | 1.37 | -19.4 | 0.000 |
| Trial | 0.16 | 0.06 | 2.41 | 0.01 |

which was significantly higher for the Up-Down condition $(t(44)=3.21, p<0.01)$. Strategic exploration was not as pronounced in the Down-Up condition, as the mean position seems closer to a straight line than the later target trajectory. Since the prevailing current would nudge any passive participants who aimed straight ahead upward, a bias toward the upper half is to be expected in both conditions. There is no evidence that participants in the Down-Up condition explored less, as there was no difference in the variance of chosen inputs during the free exploration phase between the conditions $(t(44)=-0.32, p>0.75)$. Participants in the Down-Up condition chose angles which were on average more downwards during the first 10 trials $(t(44)=-3.17, p<0.01)$. Thus, perhaps participants in the Down-Up condition also explored strategically, but were less able to steer the boat in a clear and consistent "practice run" of the desired future route.

## Discussion and Conclusion

Scenarios in which we have to explore to effectively exploit dynamical systems are ubiquitous in daily life. We introduced a novel control task and assessed to what extent people's exploration can be seen as a strategic, opportunistic, and goaldirected behavior.

We found that participants displayed hallmarks of strategic exploration, exploring differently depending on the cost function and, in some cases, practicing part trajectories which would become important later on. However, strategic exploration seemed more conservative than that of an idealized approximate dual control strategy. During periods of medium cost, participants seemed reluctant to explore in order to benefit their performance during a following high-cost period in Experiment 1. For controllers who learn and choose actions more noisily than statistical algorithms, perhaps the future benefits of this costly exploration did not outweigh the immediate costs. Participants also did not always play out strategies of future importance during free exploration trials as indicated by Experiment 2. As participants in the Up-Down condition could easily follow the underlying upward-current, participants in the Down-Up condition had to go against the current in order to explore strategically. Therefore, the difference in exploration behavior could imply that, for humans, serendipity still plays a part in discovery of effective exploration strategies.

As strategic exploration requires considerable planning, even when dual control is approximate, it is likely to require considerable mental effort. Future research could look into possible heuristics which approximate strategic exploration whilst further reducing computational costs.

## References

Duff, M. O. (2002). Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes. Ph.D. thesis, University of Massachusetts Amherst.
Feldbaum, A. (1960). Dual control theory. Avtomatika i Telemekhanika, 21(9), 1240-1249.
Klenske, E. D., \& Hennig, P. (2015). Dual control for approximate bayesian reinforcement learning. arXiv preprint arXiv:1510.03591.
Poupart, P. (2010). Bayesian reinforcement learning. In Encyclopedia of Machine Learning, (pp. 90-93). Springer.
Rescorla, R. A., Wagner, A. R., et al. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory, 2, 64-99.
Schulz, E., Konstantinidis, E., \& Speekenbrink, M. (2016). Putting bandits into context: How function learning supports decision making. bioRxiv, (p. 081091).
Srinivas, N., Krause, A., Kakade, S. M., \& Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995.
Thrun, S. B. (1992). Efficient exploration in reinforcement learning. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., \& Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. Journal of Experimental Psychology: General, 143(6), 2074.
Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., \& Meder, B. (2017). Mapping the unknown: The spatially correlated multi-armed bandit. In Proceedings of the 39th Annual Conference of the Cognitive Science Society.

# How Does Instance-Based Inference About Event Frequencies Develop? An Analysis with a Computational Process Model 

Christin Schulze (cschulze@mpib-berlin.mpg.de)<br>Thorsten Pachur (pachur@mpib-berlin.mpg.de)<br>Ralph Hertwig (hertwig@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development, Lentzeallee 94<br>Berlin, 14195 Germany


#### Abstract

To make inferences about the frequency of events in the world (e.g., the prevalence of diseases or the popularity of consumer products), people often exploit observations of relevant instances sampled from their personal social network. How does this ability to infer event frequencies by searching and relying on personal instance knowledge develop from childhood to adulthood? To address this question, we conducted a study in which children (age 8-11 years) and adults (age 19-34 years) judged the relative frequencies of first names in Germany. Based on the recalled instances of the names in participants' social networks, we modeled their frequency judgments and the underlying search process with a Bayesian hierarchical latent-mixture approach encompassing different computational models. We found developmental differences in the inference strategies that children and adults used. Whereas the judgments of most adults were best described by a noncompensatory strategy that assumes limited and sequentially ordered search (social-circle model), the judgments of most children were best described by a compensatory strategy that assumes exhaustive search and information aggregation (availability-by-recall). Our results highlight that already children use instance knowledge to infer event frequencies but they appear to search more exhaustively for instances than adults. One interpretation of these results is that the ability to conduct ordered and focused search is a central aspect in the development of noncompensatory instance-based inference.


Keywords: child development; sampling; probabilistic inference; heuristics; availability

## Introduction

The relative frequency of events in the world is an important ecological characteristic that impacts many actions and decisions. For instance, the relative frequency of other people's behaviors hints at social norms that should be followed; the number of people having bought different products may indicate differences in product quality that influence consumer choice; and the prevalence of diseases hints at health risks that may guide precautionary actions. Decision makers commonly do not have access to summary tables of these frequency statistics but need to infer them. An easily accessible but informative indicator for event frequencies in the population is their occurrence among the people one knows personally. That is, by searching for relevant instances in their personal social network people can collect a variety of information about the frequency of events in the world, and use this information to form
subjective frequency judgments. In this paper, we examine how this ability to search proximate social spaces to judge the relative frequency of events develops from childhood to adulthood.

Previous work has garnered much insight into how adults make instance-based inferences. Most prominently, according to Tversky and Kahneman's (1973) availability heuristic, adults judge the frequency of events by assessing the ease with which instances of the events can be brought to mind. More recent research has elaborated the specific mechanisms guiding this search in and retrieval from mnemonic sample spaces. For instance, it has been shown that adults often restrict search to directly experienced instances in their social circles and that these social circles are searched sequentially (e.g., Hertwig, Pachur, \& Kurzenhäuser, 2005; Pachur, Hertwig, \& Rieskamp, 2013).

Yet, currently only little is known about how search for information in proximate social spaces develops ontogenetically. Do already children exploit their social memories to draw inferences about the frequency of events in the world? And if so, how much do they sample, in which order do they consult social circles, and how do they integrate the information to draw inferences? Existing developmental work on judgment and decision making is consistent with opposing predictions. On the one hand, working memory limitations may confine young children to using information-frugal strategies because processing and integrating large amounts of evidence may be difficult (e.g., Bereby-Meyer, Assor, \& Katz, 2004). On the other hand, limitations in the ability to selectively focus attention on relevant information may lead young children to use more exhaustive but unsystematic search strategies (e.g., Davidson, 1991; Mata, von Helversen, \& Rieskamp, 2011).

To disentangle these opposing predictions, we first introduce the social-circle model, a cognitive process model that parameterizes key components of the inference process-including search order, evidence threshold, and response noise. Second, we take a Bayesian hierarchical mixture approach to modeling the inferences of children and adults in a task in which they made judgments about the relative frequency of common first names in Germany.

## The Social-Circle Model

To model people's inferences based on recalled instances, Pachur et al. (2013) proposed that people search
sequentially through the circles of their social networkdefined as self, family, friends, and acquaintances-and stop search as soon as the instance evidence in a circle allows them to make an inference. It is thus assumed that people's search for relevant instances is guided by the welldocumented hierarchical structure in the ordering of discrete social groups that make up a person's social network (e.g., Hill \& Dunbar, 2003; Milardo, 1992), which has also been shown to be important for search in social memory (e.g., Hills \& Pachur, 2012). Adults' frequency judgments have been found to be equally well described by a model that assumes such a noncompensatory strategy and by a more exhaustive, compensatory search strategy (Pachur et al., 2013). Here, we formalize and extend the assumptions in Pachur et al.'s (2013) analysis and propose a generalized social-circle model (SCM) that allows for variability in the order in which circles are inspected and for probabilistic aspects in the search, stopping, and decision stages of inference.

The SCM assumes that in order to judge which of two events, A or B , is more frequent in the population, decision makers search distinct social circles, defined as self, family, friends, and acquaintances. At each inspected circle $i$ the evidence, $e_{i}$, is represented as the difference in the number of instances recalled for each event, expressed as a proportion:

$$
\begin{equation*}
e_{i}=\frac{n_{i A}}{n_{i A}+n_{i B}}-\frac{n_{i B}}{n_{i A}+n_{i B}} \tag{1}
\end{equation*}
$$

## Search Rule

The order in which the circles are inspected is represented by circle-weight parameters, one for each circle ( $w_{i}$, constrained by $\sum w_{i}=1$; see Bergert \& Nosofsky, 2007), that can be estimated from the data. These weights represent the probability that a circle is inspected as

$$
\begin{equation*}
p\left(\text { inspect } \operatorname{circle}_{i}\right)=\frac{w_{i}}{\sum_{i}^{N} w} \tag{2}
\end{equation*}
$$

Once a circle has been inspected, it is not considered further (i.e., the denominator is calculated only over circles that have not yet been inspected). Note that search is thus assumed to be probabilistic. The probability of following a particular search order $p\left(\right.$ order $\left._{j}\right)$ is given by the product of the individual probabilities of circle inspection,

$$
\begin{equation*}
p\left(\text { order }_{j}\right)=\Pi_{j} p\left(\text { inspect circle } e_{i}\right) . \tag{3}
\end{equation*}
$$

## Stopping Rule

In the SCM it is assumed that the proportional evidence obtained from each circle is compared against a decision threshold, $d$. If the evidence from the recalled instances reaches or exceeds the threshold, a choice is made; if it is lower than the threshold, the next circle is inspected. The SCM implements a probabilistic version of this stopping rule by assuming normally distributed error for each circle, denoted as $\varepsilon_{i}$, generated from a normal distribution with mean zero and standard deviation $\sigma$. Specifically, it is
assumed that, if the evidence in a given circle (with added error) meets or exceeds $d$, then the decision maker selects option A (i.e., $\left|e_{i}+\varepsilon_{i}\right| \geq d$ ); if the evidence meets $-d$, then the decision maker selects option B (i.e., $\left|e_{i}+\varepsilon_{i}\right| \leq-d$ ). Thus, the probability of making a choice after inspection of circle $i$ is given by

$$
\begin{align*}
p_{i}(\text { choice }) & =p\left(\left|e_{i}+\varepsilon_{i}\right| \geq d_{i}\right) \\
& =p\left(e_{i}+\varepsilon_{i} \geq d_{i}\right)+p\left(e_{i}+\varepsilon_{i} \leq-d_{i}\right) \\
& =\Phi\left(\frac{e_{i}-d_{i}}{\sigma}\right)+\Phi\left(\frac{-e_{i}-d_{i}}{\sigma}\right) \tag{4}
\end{align*}
$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

## Decision Rule

The probability of selecting option A based on a particular order, $p_{j}(A \mid A B)$, follows from combining the choice probabilities resulting from circle inspection in that order (cf. Rieskamp, 2008). For example, for the order $j=1,2,3,4$ :

$$
\begin{align*}
& p_{j=1,2,3,4}(A \mid A B)=p_{1}(A \mid A B)+\left[1-p_{1}(\text { choice })\right] \times \\
& \quad p_{2}(A \mid A B)+\left[1-p_{1}(\text { choice })\right] \times\left[1-p_{2}(\text { choice })\right] \times \\
& \quad p_{3}(A \mid A B)+\left[1-p_{1}(\text { choice })\right] \times\left[1-p_{2}(\text { choice })\right] \times \\
&  \tag{5}\\
& {\left[1-p_{3}(\text { choice })\right] \times p_{4}(A \mid A B) .}
\end{align*}
$$

The total probability of selecting option A is defined as the sum of all $p_{j}(A \mid A B)$, each weighted by the probability of the decision maker following the order (see Equation 3):

$$
\begin{equation*}
p(A \mid A B)=\sum_{j=1}^{N!} p_{j}(A \mid A B) \times p\left(\text { order }_{j}\right) \tag{6}
\end{equation*}
$$

In sum, the SCM parameterizes three key components of instance-based inference: the decision maker's preferred search order (circle weight parameters, $w_{i}$ ), evidence threshold $(d)$, and response noise $(\sigma)$. Thus, depending on its parametrization, the model can capture various noncompensatory inference processes. In what follows, we apply the SCM to inference data from an experiment in which children and adults were asked to judge the relative frequency of common first names in Germany, and examine how well it accounts for participants' inferences compared to a compensatory strategy and a guessing strategy.

## Experiment

## Method

Participants Forty children (age 8-11 years; 18 female) and 40 adults (age 19-34 years; 19 female) who were recruited via the subject pool of the Max Planck Institute for Human Development participated in the experiment. The data of five additional children were excluded from the analysis because the children showed insufficient readingcomprehension (two children aged 7 years); did not recall any or only one instance for each name in the same social circle, yielding a guessing prediction for instance-based models on every trial (two children); or terminated the experiment prematurely (one child). Participants received a performance-based payment (earning 0.04 EUR for each
correct inference but losing the same amount for each incorrect inference; $1 \mathrm{EUR} \approx 1.12$ USD at the time of the experiment), and an additional flat fee of 10 EUR.

Materials Table 1 lists the 22 first names ( 11 female) that were used in the experiment. Because no census data about the frequency and distribution of first names in Germany was available, we approximated a frequency ranking by weighting popular baby names between 1911 and 2010 (Bielefeld, 2016) with each cohort's proportion in the population to date (Statistisches Bundesamt, 2014). ${ }^{1}$ We constructed a set of all possible 231 paired comparisons of the names, and informed participants that the accuracy of their inferences was judged on the basis of the available data. Participants were instructed to ignore the particular spelling of each name and to judge the relative frequency of names by taking possible variants of a name into account.

Procedure The experiment consisted of two tasks, an inference and a retrieval task, that were completed by all participants in this order. In the inference task, participants were asked to judge which of two first names is more frequent in Germany for each of the 231 name pairs. The pairs were presented sequentially on a computer screen in blocks of 23 pairs ( 24 pairs in the final block). The order in which name pairs were presented was randomized across participants; the order of names in each pair was predetermined so that correct and incorrect inferences (according to our statistics) were distributed equally across the two response alternatives. Each trial started with the display of a fixation cross at the center of the screen, followed by the presentation of two black silhouettes (either male or female) which were labeled with the respective names in the comparison (see Figure 1A). Participants made a selection by pressing one of two designated keys on the keyboard. After each choice, the selected name's silhouette was shown on a podium at the center of the screen to confirm the selection to the participant. There was no trial-by-trial feedback about the accuracy of decisions. Participants were encouraged to make as many correct judgments as possible. There was a self-paced pause after each block and participants completed two training trials with fictitious names before the start of the inference task. In the retrieval task, participants were asked to recall how

[^197]Table 1: The 22 first names used in the experiment, their approximated frequency rank in Germany, and the total number of instances children and adults recalled from their own social networks.

|  |  |  | Total number of recalled |  |
| :--- | :---: | :---: | :---: | :---: |
| Name | Gender | Rank | instances |  |
|  |  |  | Children | Adults |
| Michael | m | 1 | 35 | 66 |
| Thomas | m | 2 | 34 | 72 |
| Peter | m | 3 | 29 | 45 |
| Andreas | m | 4 | 34 | 65 |
| Jan | m | 5 | 40 | 67 |
| Hans | m | 6 | 22 | 26 |
| Christian | m | 7 | 29 | 76 |
| Karin | f | 8 | 14 | 24 |
| Ursula | f | 9 | 4 | 15 |
| Julia | f | 10 | 34 | 78 |
| Anna | f | 11 | 41 | 70 |
| Sabine | f | 12 | 29 | 44 |
| Stefanie | f | 13 | 24 | 58 |
| Renate | f | 14 | 19 | 20 |
| Helga | f | 15 | 18 | 17 |
| Günter | m | 16 | 11 | 16 |
| Tim | m | 17 | 40 | 43 |
| Horst | m | 18 | 11 | 17 |
| Angelika | f | 19 | 16 | 27 |
| Lukas | m | 20 | 39 | 46 |
| Hannah | f | 21 | 42 | 44 |
| Gertrud | f | 22 | 6 | 9 |

many people with each of the 22 names shown in Table 1 they knew personally. For each name, participants counted each person among their family, friends, and acquaintances with that name by dragging and dropping pictorial representations of family members, friends, and acquaintances on a black silhouette labeled with the respective name (see Figure 1B). Following the retrieval of a person, participants were also asked to indicate their contact frequency with that person on a scale from one (less than once every six months) to five (multiple times per week). Additionally, participants could allocate a pictorial person labeled "self" to indicate the shown name was their own. Each recalled person was listed on the screen and counted toward an overall tally of persons with a particular name also shown on the screen. Before the start of the retrieval task, a training trial familiarized participants with the controls of this task. At the end of the experiment, participants were informed about their overall accuracy on the inference task and paid in cash by the experimenter.

Bayesian Hierarchical Mixture Modeling Based on the instances of names that each participant recalled from their social network in the retrieval task, we modeled each participant's decisions in the inference task with a Bayesian latent-mixture approach (see, e.g., Bartlema, Lee, Wetzels, \& Vanpaemel, 2014). Hierarchical mixture modeling allows

A


B


Figure 1: Illustration of the task screen and controls used during the inference task (A) and retrieval task (B).
us to simultaneously estimate discrete classes of participants who use categorically different inference strategies and to robustly model variation within each group of strategyusers, thus combining the advantages of pooling continuous individual differences hierarchically and assuming discrete differences among groups of individuals. We assumed three latent subgroups of participants, each using a different inference strategy: (a) the social-circle model, (b) availability-by-recall, which assumes a compensatory process (Hertwig et al., 2005; Pachur, Hertwig, \& Steinmann, 2012), and (c) a random guessing strategy.

According to availability-by-recall, all instances of an event are tallied across the entire social network and the option with the larger summed instance-evidence is chosen. For comparability, we applied the same response noise mechanism as for the SCM, which gives the probability of choosing option A as

$$
\begin{equation*}
p_{A b R}(A \mid A B)=\Phi\left(\frac{n_{A}-n_{B}}{\sigma}\right), \tag{7}
\end{equation*}
$$

where $n_{A}$ denotes the number of instances recalled for event A across all circles and $\sigma$ is a response noise parameter. For the guessing strategy, we assumed that participants randomly selected one of the two names in each pair with probability .50 . With this approach, we can estimate the proportion of participants using each strategy based on inference and recall data while taking into account the uncertainty surrounding such a classification. We modeled participants' inferences for all paired comparisons on which
a participant's instance knowledge allowed each strategy to make an unambiguous prediction. The two instance-based strategies did not make a prediction, if a participant recalled no or equal numbers of instances for both names in a comparison. The posterior distributions of model parameters were estimated via Gibbs sampling methods implemented in JAGS (Plummer, 2003). We used reasonably uninformative priors: For the $w_{i}$ and $d$ parameters of the SCM we assumed uniform priors on the group-level mean (beta distributions with shape parameters of 1 ) and gamma priors (with a shape parameter of 1.1051 and a scale parameter of 0.01051 ; see Bartlema et al., 2014) on the group-level precision. For the $\sigma$ parameters of the SCM and availability-by-recall we assumed uniform distributions constrained between 0.01-40 on the group-level mode and standard deviation. For the latent-mixture indicator variable we assumed a categorical prior that assigned equal weight to each strategy. ${ }^{2}$ To ensure efficient mixing, we used pseudo-priors that approximate the posterior density for the individual-level parameters. These pseudo-priors were obtained from an initial Bayesian hierarchical estimation procedure that was performed separately for each model (without a mixture component). In the model estimation, 16 chains each with 50,000 samples drawn from the posterior distributions were run after an initial burn-in period of 2000 samples. GelmanRubin statistics and visual inspections of the four chains indicated adequate chain convergence.

## Results

Behavioral Data We found differences between the age groups in inferential accuracy, $t(78)=5.17, p<.001$, $d=1.16, B F_{10}=8362$, and in reported instance knowledge, $t(60.00)=4.68, \quad p<.001, \quad d=1.05, \quad B F_{10}=1456$. On average, adults picked the more frequent first name more often than children ( $M=.64$ vs. $M=.57$ ) and recalled more people with any of the 22 first names in their social network ( $M=23.63$ vs. $M=14.28$; see also Table 1 ). One possible reason for children's lower inferential accuracy is that the instances they reported were less valid indicators of the actual frequency distribution of first names in the population (possibly because they recalled fewer instances overall). That is, for adults, there was a significant rank correlation between reported instances and actual frequency ranks, $r_{s}(20)=.524, p=.012, B F_{10}=4.99$. For children, however, no such correlation was found, $r_{s}(20)=.203, p=.364$, $B F_{10}=0.39 .{ }^{3}$

[^198]

Figure 2: Allocation of adult and children participants to three latent subgroups of strategy users.

Computational Modeling Figure 2 shows the membership probability of each adult (left panel) and each child (right panel) in each group of strategy-users, as derived from the posterior distribution of the latent-mixture variable. The figure shows that the judgments of most adults were best described by the SCM (55\% of adults compared to $38 \%$ of children). By contrast, the judgments of most children were best described by availability-by-recall ( $48 \%$ of children compared to $40 \%$ of adults). Only few participants were best described by the guessing strategy. Overall, there was greater uncertainty in the classification of children to latent groups than in the classification of adults. This was partly due to the lower number of instances children recalled resulting in poorer discriminability between the models.

Next, we compared children's and adults' search and decision processes by evaluating their group-level SCM parameter estimates. As shown in Figure 3, children and adults weighted the different circles in their social network similarly (although adults showed greater inter-individual variability in the weighting of different circles), applied similar decision thresholds, and did not differ on the response noise parameter (for all parameters, 95\% HDIs overlapped). Children's lower inferential accuracy was thus not due to a more error prone execution of an instance-based inference strategy. This also held for inferences described by availability-by-recall.

## Discussion

Our results suggest that already children systematically exploit their instance knowledge to make inferences about the frequency of events in the world. However, they do so differently than adults. Whereas the judgments of most adults were best described by a strategy that assumes limited information search, the judgments of most children were best accounted for by a strategy that assumes exhaustive search. This finding echoes previous research on multiattribute choice and cue-based inference which has found young children to use more exhaustive but unsystematic
search strategies (e.g., Davidson, 1991; Mata et al., 2011). A possible explanation for why children use more informationintensive strategies is that they have difficulties to selectively attend to relevant and diagnostic information (cf. Betsch, Lehmann, Lindow, Lang, \& Schoemann, 2016). In young children, this inability to effectively focus search may be driven by the required executive control functions being not yet fully developed (see Best \& Miller, 2010). In light of


Posterior Group-Level Parameter Estimates of the SCM

Figure 3: Posterior distributions of the group-level parameters of the SCM. Small circles and diamonds below the density plots show the posterior means for adults and children, respectively; lines show $95 \%$ HDIs.
children's more limited and less ecologically valid instance knowledge, their greater tendency to adopt exhaustive sampling strategies might represent an adaptive response to these limiting factors. However, it should also be noted that, due to children's lower instance knowledge, the discriminability between models was lower, which might have contributed to the more balanced strategy classification in children as well.

Our results extend previous research that has found children to use availability as a cue for judging the relative frequency of and their own memory for names (Davies \& White, 1994; Geurten, Willems, Germain, \& Meulemans, 2015). This prior work, however, did not use cognitive modeling to formalize and quantitatively analyze the development and use of instance-base inference strategies. By taking a formal computational modeling-based approach, our analysis enabled us to simultaneously detect developmental differences in the use of discrete strategies and parameterize the specific mechanisms underlying search for instances in memory. This approach highlighted that children search for instances more exhaustively but weight the subgroups in their social network similarly as do adults. The analysis also revealed substantial individual differences in the process of search for instances in memory among both age groups. In this respect, the social-circle model that we applied provides an advantage over previously proposed models of instance-based inference (e.g., Tversky \& Kahneman, 1973), which are silent regarding the specific mechanisms and order of instance sampling.

We conclude that the social-circle model provides an effective tool for capturing and illuminating individual and group differences in the cognitive processes that underlie instance-based inference. The insights gained with this model are consistent with the finding that search in social memory is guided by factors such as social proximity (Hills \& Pachur, 2012) and suggest that one important factor in the development of information-frugal strategies for judging frequencies is the ability to limit and selectively focus search on relevant instance knowledge.

## References

Bartlema, A., Lee, M., Wetzels, R., \& Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. Journal of Mathematical Psychology, 59, 132-150.
Bereby-Meyer, Y., Assor, A., \& Katz, I. (2004). Children's choice strategies: The effects of age and task demands. Cognitive Development, 19, 127-146.
Bergert, F. B., \& Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-thebest models of decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 107129.

Best, J. R., \& Miller, P. H. (2010). A developmental perspective on executive function. Child Development, 81, 1641-1660.

Betsch, T., Lehmann, A., Lindow, S., Lang, A., \& Schoemann, M. (2016). Lost in search: (Mal-)adaptation to probabilistic decision environments in children and adults. Developmental Psychology, 52, 311-325.
Bielefeld, K. (2016). Statistiken der häufigsten Vornamen für jeden Geburtsjahrgang [Statistics of the most common first names for every cohort; data set]. Retrieved from http://www.beliebte-vornamen.de/jahrgang
Davidson, D. (1991). Children's decision-making examined with an information-board procedure. Cognitive Development, 6, 77-90.
Davies, M., \& White, P. A. (1994). Use of the availability heuristic by children. British Journal of Developmental Psychology, 12, 503-505.
Geurten, M., Willems, S., Germain, S., \& Meulemans, T. (2015). Less is more: The availability heuristic in early childhood. British Journal of Developmental Psychology, 33, 405-410.
Hertwig, R., Pachur, T., \& Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 621-642.
Hill, R. A., \& Dunbar, R. I. M. (2003). Social network size in humans. Human Nature, 14, 53-72.
Hills, T. T., \& Pachur, T. (2012). Dynamic search and working memory in social recall. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 218228.

Mata, R., von Helversen, B., \& Rieskamp, J. (2011). When easy comes hard: The development of adaptive strategy selection. Child Development, 82, 687-700.
Milardo, R. M. (1992). Comparative methods for delineating social networks. Journal of Social and Personal Relationships, 9, 447-461.
Pachur, T., Hertwig, R., \& Rieskamp, J. (2013). Intuitive judgments of social statistics: How exhaustive does sampling need to be? Journal of Experimental Social Psychology, 49, 1059-1077.
Pachur, T., Hertwig, R., \& Steinmann, F. (2012). How do people judge risks: Availability heuristic, affect heuristic, or both? Journal of Experimental Psychology: Applied, 18, 314-330.
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, \& A. Zeileis (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing.
Rieskamp, J. (2008). The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 1446-1465.
Statistisches Bundesamt. (2014). Bevölkerung nach Geburtsjahren und Geschlecht für Deutschland [German population by year of birth and gender; data set]. Retrieved from http://www.zensus2011.de
Tversky, A., \& Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207-232.

# Prior Expectations in Linguistic Learning: A Stochastic Model of Individual Differences 

R. Alexander Schumacher<br>Northwestern University

Janet Pierrehumbert
University of Oxford


#### Abstract

When learners are exposed to inconsistent input, do they reproduce the probabilities in the input (probability matching), or produce some variants disproportionately often (regularization)? Laboratory results and computational models of artificial language learning both argue that the learning mechanism is basically probability matching, with regularization arising from additional factors. However, these models were fit to aggregated experimental data, which can exhibit probability matching even if all individuals regularize. To assess whether learning can be accurately characterized as basically probability matching or systematizing at the individual level, we ran a large-scale experiment. We found substantial individual variation. The structure of this variation is not predicted by recent beta-binomial models. We introduce a new model, the Double Scaling Sigmoid (DSS) model, fit its parameters on a by-participant basis, and show that it captures the patterns in the data. Prior expectations in the DSS are abstract, and do not entirely represent previous experience.


# Effects of Grammatical Gender on Object Description 

Arturs Semenuks ${ }^{1}$ (asemenuk@ucsd.edu) Webb Philips ${ }^{1}$ (webb.philips@gmail.com) Ioana Dalca ${ }^{2}$ (ioanadalca@gmail.com) Cora Kim ${ }^{3}$ (cora.kim@fu-berlin.de) Lera Boroditsky ${ }^{1}$ (lera@ucsd.edu)<br>${ }^{1}$ Department of Cognitive Science, University of California, San Diego<br>${ }^{2}$ School of Electronic Engineering and Computer Science, Queen Mary University of London<br>${ }^{3}$ Brain Language Laboratory, Department of Philosophy and Humanities, Freie Universität Berlin


#### Abstract

Can grammatical gender influence how people conceptualize the referents of nouns? Using an implicit measure, we investigated whether such an effect could be found in a task where neither grammatical nor biological gender is highlighted. In the current study, conducted in English, speakers of French, German and Romanian with knowledge of English were asked to generate adjectives they associate with referents of nouns. Afterwards, the gender valence of the adjectives was measured. The results showed that participants generated more feminine adjectives for nouns with majority feminine translations compared to nouns with majority masculine translations. We found a stronger effect of grammatical gender for some semantic categories than for others. Significant effects of grammatical gender were present starting with the $2^{\text {nd }}$ adjective generated by participants (effects were stronger for adjectives generated $2^{\text {nd }}$ and $3^{\text {rd }}$ by participants, as opposed to the $1^{\text {st }}$ adjective).


Keywords: grammatical gender; conceptualization; categorization; semantic features; representation

## Introduction

Can the structure of the language you speak affect how you perceive and conceptualize the world around you? Although the question has a long history (see Koerner (1992) for an overview), the last two decades have seen an especially large increase in the amount of empirical work examining and testing the predictions that the different answers to it imply. This has lead to an accumulation of results showing that there are cases where differences in seemingly nonlinguistic aspects of cognition between speakers of different languages arise due to specific differences between the structures of those languages (see Wolff \& Holmes (2011) for a survey). Nonetheless, the jury is still out on a number of questions aiming to flesh out the details of how, when and why linguistic differences can lead to differences in the operation of cognitive processes.

One question that is not conclusively resolved is what types of effects language could have on cognition. Research supports, for example, the existence of effects of language that play a large role in the operation of some specific cognitive domains. Frank et al. (2012) shows that number words are a cognitive tool that, when suppressed, impairs Western college educated people's ability to encode exact numerical values. Thus number vocabulary plays an
important role in numerical cognition. Other work indicates that language also has some early effects on cognition. A body of research has shown that color vocabulary has an effect on categorical color perception (see Regier et al. (2010) for a review). There are also other types of effects and research indicates that cognitive differences between speakers of different languages arise due to the metaphors they use for talking about time, their frequency of use of a particular spatial frame of reference and knowledge of constructions for talking about mental states, among other differences (Wolff \& Holmes, 2011).

One unresolved issue is whether language could have broad or pervasive effects. Are effects of language only restricted to specific domains (e.g. specific color distinctions, numbers, orientation in space)? Or can differences in the structure of languages lead to broadly distributed effects across a wide range of contexts? For this question, it is worth examining grammatical features that apply to say all nouns (e.g., gender) or all verbs (e.g., tense). In this paper we consider the role of grammatical gender in shaping the way people think about things that are named by nouns - a very broad potential scope of influence.

## Grammatical Gender

Grammatical gender is a system of categorizing nouns into distinct classes, i.e. genders, which manifests itself via morphologically marking what gender the noun belongs to on some of the words that that noun is morphosyntactically related to (Corbett, 1991).

The forms that gender systems take vary. The ones that have garnered the interest in the research on the effects of language on cognition, however, usually share several properties. First, these systems have masculine and feminine genders, i.e. for nouns with human referents the assignment is fully semantic and dependent on biological sex, bar a small number of isolated exceptions in some languages. Second, the majority of nouns with non-human referents fall into the semantic residue, i.e. the category for which the assignment is based on the word's phonological or morphological, but not semantic properties. Third, a number of words in the semantic residue is assigned to either masculine or feminine gender. Languages with such systems provide the opportunity to ask: can the conceptualization of
a noun's referent accrue gendered properties because of the noun's grammatical gender assignment? If so, grammatical gender could exhibit a broad and pervasive effect on concept formation, since, by having a gender system, a language requires all nouns to be assigned to one of the gender categories. This could result in a broad range of concepts acquiring gendered properties that would be absent otherwise.

Why might we expect entities without biological sex to acquire stereotypically masculine or feminine properties as a result of being assigned to a grammatical gender category? As we mentioned, grammatical gender can be considered to be a system of noun categorization. Biological sex, in that case, is a highly diagnostic feature for determining the category assignment of nouns for humans. As the research on categorization indicates, diagnostic features better suited for differentiating between members and non-members of a category draw more attention to themselves (Tversky, 1977). In addition, category features could develop or become accentuated in response to the requirements that tasks involving categorization impose (Schyns \& Rodet, 1997; Medin et al., 1993). This could mean that nouns belonging to a grammatical gender strongly associated with humans of a particular sex might acquire gendered properties just by virtue of being in those genders. Intuitions of native speakers of languages with grammatical gender for why nouns get assigned to their respective genders in their language supports this view (Basetti, 2011).

## Prior Research and the Goals of the Current Study

A body of prior work supports the idea that grammatical gender influences conceptualization of objects and animals (Basetti \& Nicoladis, 2016). For example, Konishi (1993) asked native speakers of German and Spanish to rate nouns on a variety of semantic differential scales in their native language. The words were either feminine in German and masculine in Spanish or vice versa. Subjects tended to give higher ratings on scales associated with potency for words that were masculine in their native language. In another set of experiments by Sera et al. (1994), Spanish and English speaking adults and children were asked to either give human names or assign male or female voices to pictures of objects. The results showed that older Spanish speakers tended to assign names and voices in these tasks in a grammatical gender-congruent fashion, which was not the case for English speakers or Spanish preschoolers. Phillips and Boroditsky (2003) investigated the effects of grammatical gender by teaching native speakers of English ${ }^{1}$ an artificial language with grammatical gender and analyzed whether it influenced their performance on a similarityrating task. The results indicated that participants rated pictures of humans and objects as more similar when they were in the same grammatical gender.

However, other studies suggest that there are limitations

[^199]on when and how grammatical gender could have an effect on conceptualization. For example, Vigliocco et al. (2005) found an effect of animal name grammatical gender on similarity in a triad similarity judgment task when the stimuli were names of animals, but not when the stimuli were pictures of animals. Furthermore, the effect was not found for either names or pictures of artifacts. In another experiment, Kousta et al. (2008) asked participants to name pictures that appeared at a random location on the screen, and afterwards analyzed cases where participants' produced an incorrect word. The subjects were either Italian or English monolinguals, or Italian-English bilinguals. The monolinguals responded in their native language, whereas the bilinguals participated in the experiment in both Italian and English. The results showed that the grammatical gender of the Italian noun describing the picture tended to affect the responses of Italian monolinguals - they preserved the grammatical gender of the target word almost twice as often as English monolinguals. The responses of ItalianEnglish bilinguals followed a pattern similar to Italian monolingual pattern when they were tested in Italian. When they were tested in English, however, their pattern of errors was similar to the one exhibited by English monolinguals. Ramos and Roberson (2010) examined how speakers of Portuguese conceptualize inanimate objects in a series of experiments using different methodologies and compared their responses to the responses of an English-speaking group. The authors found that the grammatical gender of an object's name strongly influenced whether the Portuguese speakers assigned a male or a female voice to it, but had a smaller influence on the participant's responses in word similarity rating task, or in a triad similarity judgment task, especially when the stimuli were pictures, and not words.
These findings raise several issues. One alternative interpretation of the previously obtained effects is that they are by-products of task demands or of the particular stimuli that were used in the study. For example, when people are asked to give a name to an object, they might consciously or unconsciously understand that gendered information is important for this task. This seems especially likely when participants come from cultures where sets of male and female names have very little overlap, or when they are asked to give the name to something that can be very easily anthropomorphized, e.g. an animal. This then could lead to a pattern of responses that suggests an effect of grammatical gender on ingrained features of concepts, but that actually arose due to the demands of a particular context. Similar arguments could be made for other methodologies.

Another possibility is that some of the results arise because of the influence of information contained in the experimental stimuli. Given a set of words, the ones that are more similar phonologically or morphologically are more likely to prime each other (Kinoshita \& Lupker, 2004), and words in the semantic residue are assigned to a grammatical gender specifically based on their phonology or morphology. When a speaker of Spanish is participating in a semantic differential scale task and reads a masculine noun
ending in " 0 ", this weakly activates other words of similar morphology and phonology. That, supported by the experimental context, could lead to higher activation of frequently used words for human males and their properties. It, in turn, could lead to participants giving higher ratings to masculine words on scales associated with masculinity.

The experiment described below aims to address these issues. Three groups of non-native English speakers whose native languages have grammatical gender systems were asked to generate adjectives for English nouns. The adjectives were then scored on how feminine or masculine they were, and adjectives generated for nouns differing in grammatical gender between the languages were examined. This design addresses the issues discussed above in the following ways. The manipulation is implicit - there is no way for participants to guess that gender is of interest in the study - they are simply given a list of nouns and asked to generate adjectives that describe the referent. The adjectives are generated for English words, which eliminates the possibility of phonetic or morphological similarity between words driving the effect.

## Methods

## Participants

A total of 273 participants completed the study. Out of them, 99 were Romanian-English bilinguals ( 69 female; mean age $=24.03, \mathrm{SD}=6.83$; mean English proficiency $=$ $4.38 / 5, \mathrm{SD}=0.62$ ), 90 were German-English bilinguals ( 53 female; mean age $=24.37, \mathrm{SD}=5.91$; mean English proficiency $=4.08 / 5, \mathrm{SD}=0.80$ ) and 84 were FrenchEnglish bilinguals ( 51 female; mean age $=28.96$, $\mathrm{SD}=$ 11.52; mean English proficiency $=4.38 / 5, \mathrm{SD}=0.64$ ).

## Materials

A total of 225 nouns served as stimuli. Out of them, 200 were the most frequently used nouns in the English language based on the data from the Corpus of Contemporary American English (Davies, 2008). These words were used for several reasons. First, it decreases the possibility of an effect arising due to an unconscious experimenter bias in the choice of items. Second, these words are often encountered in written and spoken language, so their meaning is unlikely to be misunderstood by the participants. Lastly, the procedure allows for an easier selection of items in future experiments using a similar methodology, but examining speakers of languages other than the ones used in this experiment.

We were also interested in examining whether the semantic category of the noun modulates what adjectives are generated for it, as previous work in the field has shown that the semantic category of the noun affects participants' responses (e.g. Vigliocco et al., 2005). Because the list described above lacked nouns with animal referents, 25 such words were added to the stimuli.

## Procedure

All of the instructions in the experiment were given in English. At the very beginning participants were asked to verify their native language proficiency by translating an English sentence.

After this, the participants were informed that in the next part of the study they would see a list of nouns one by one, and that they would need to list the properties that they associate with their referents. For each noun the participants had to list three adjectives. Half of the participants were instructed to personify the nouns, i.e. imagine them as humans, and generate adjectives that would best suit those personifications, whereas the other half was not given personification instructions (and were instructed simply to produce adjectives). This was done in order to establish whether explicit invitations to personify (as when assigning names or voices) are necessary to induce effects of grammatical gender.

Next, the participants continued onto the adjective generation portion of the experiment. This part of the study consisted of 225 trials. During each trial the participants saw an English noun from the stimulus set and were asked to generate three adjectives for it. The participants were presented with all of the items shown in a random order.

After this, the participants were asked to translate all of the nouns they had encountered previously into their native language. At the end of the experiment, the participants filled in a questionnaire about their language background, education and residency.

## Post-Processing

Grammatical Gender of Translations Some of the nouns were translated by the participants in multiple ways. For example, German participants translated English "difference" both as "(der) Unterschied" (masculine) and "(die) Differenz" (feminine).

Due to this, we first established the grammatical gender for each of the translations generated by the participants. Afterwards, the number of translations belonging to each of the grammatical genders was calculated for each noun in each language. We call the most common grammatical gender among the noun's translations in a particular language its most common gender (MCG) in that language.

Noun Semantic Categories In order to analyze the effect of noun's semantic category, each noun was categorized as being either an abstract noun, a noun denoting an animal, a body part or a concrete object by two coders.

Noun Selection After the most common grammatical genders of the nouns were established, only the nouns that had most common masculine gender in one of the experimental languages, and feminine in one of the others were left for analysis, leaving a total of 68 nouns. In cases where the noun's MCG in the third language was neuter, only adjectives generated by participants speaking the other
two languages were analyzed. 39 denoted abstract concepts, 13 - animals, 4 - body parts and 12 - other concrete objects.

Adjective Gender Valence Our next task was to operationalize how masculine or feminine the generated adjectives were. The participants generated a total of 45972 adjective tokens for the nouns remaining in the sample. Out of them, 36426 were also generated for nouns for humans (i) whose referent was clearly biologically male or female (e.g. "man", "woman") or (ii) that differed in translation depending on the sex of the referent (e.g. English "friend" = French "ami" (masc) or "amie" (fem)). For each of those adjectives, we calculated its token frequency among adjectives generated for males (a total of 12564) and adjectives generated for females (a total of 3681 tokens). After that, we subtracted the frequency of each adjective among adjectives generated for biological females from its frequency among those generated for biological males. We consider this to reflect adjectives' relative gender valence: relatively more positive values indicate more masculine connotations, relatively more negative values indicate more feminine connotations.

## Results

The data were analyzed with linear mixed-effects models using the lme4 and multcomp packages for R (Bates et al., 2015; Hothorn et al., 2016). The most common gender (MCG) of the noun in the participant's native language, the noun's semantic category, personification condition, i.e. whether the participant was asked to personify the nouns, and the order of the adjective were modeled as fixed effects, whereas subject IDs and items were modeled as random effects.

As a manipulation check, we examined whether the biological gender of the referent for nouns denoting humans predicted what adjectives would be generated for it. As can be seen on figure 1 , nouns for males on average received more masculine adjectives $(M=0.001)$ compared to nouns for females ( $\mathrm{M}=-0.005$ ). We compared mixed-effect models with gender valence of the adjective as the dependent variable, random intercept for subject and bysubject random slope for MCG of the noun, and a random intercept for noun. The model with MCG of the noun fit the data significantly better ( $\chi 2=21.871, \mathrm{p}<0.001$ ).

Following that, we analyzed the data from nouns for nonhumans. Only comparisons between adjectives generated for nouns that had either masculine or feminine MCG are reported. Besides the specified fixed effects, all of the models below contain random intercept for subject and bysubject random slope for MCG, as well as random intercept for noun and by-noun random slope for personification.

Adjectives generated by the participants for nouns with majority masculine translations had higher gender valence ratings $(M=0.0007)$ than adjectives generated for nouns with majority feminine translations ( $\mathrm{M}=0.0002$ ), as figure 2 shows. To investigate whether the difference is significant, we compared mixed-effect models with
personification, noun's semantic category, order of the adjective, as well as all of their interactions as fixed effects. Comparison of the models with and without MCG as a fixed effect revealed that it significantly improves the fit of the model ( $\chi 2=14.988, \mathrm{p}<0.01$ ).

The effect, however, could potentially arise due only to the adjectives generated by participants who were asked to personify the noun's referent. To examine that possibility, we compared mixed-effect models with and without an interaction between personification and MCG. The models contained all other possible main effects and interactions between them. Model comparison revealed no significant differences between them $(\chi 2=2.3089, p=0.13)$.


Figure 1: Gender valence of adjectives for humans.


Figure 2: Gender valence of adjectives for non-humans.
Comparison of models with and without an interaction between noun category and MCG revealed a significant difference between them ( $\chi 2=9.1268, \mathrm{p}<0.05$ ). Further multiple comparisons analysis showed a significant difference of the gender valences of adjectives for animal nouns ( z -value $=3.713, \mathrm{p}<0.01$ ) and a marginally significant difference for abstract nouns ( z -value $=2.929, \mathrm{p}$ $=0.078$ ) in the expected gender-congruent directions, but no significant differences in gender valence for nouns for body parts or concrete objects (see fig. 3). We note however that the categories of body parts and concrete objects contained very low numbers of items (4 and 12 respectively). Further,
our design does not license us to generalize the results found for these specific items to their respective semantic categories as a whole.


Figure 3: Gender valence of adjectives belonging to different semantic categories.


Figure 4: Gender valence of first, second and third adjectives generated by participants for non-human nouns.

Last, we examined when the effects of grammatical gender emerge by looking separately at adjectives generated first, second, and third for each noun. Comparison of models with fixed effects of MCG, personification and noun category, as well as all of their interactions, but differing in having or not having a fixed effect of adjective order, revealed a highly significant difference in their fit ( $\chi 2=$ 23.465 , $\mathrm{p}<0.001$ ). Additionally, a comparison of models with and without an interaction between MCG and adjective order was conducted. The models included all other possible main effects and interactions between them. The comparison revealed that the model with the interaction fit the data significantly better $(\chi 2=10.413, \mathrm{p}<0.01)$. Multiple comparisons analysis showed that adjectives generated first did not differ significantly depending on the MCG of the noun they were generated for $(z$-value $=-0.163$, $\mathrm{p}=1.00$ ). However, the difference in average gender valence was significant for second (z-value $=3.643, \mathrm{p}<$ 0.01 ) and third ( z -value $=4.054, \mathrm{p}<0.001$ ) adjectives generated for nouns with different MCG.

## Discussion

What do the results tell us? We see that the grammatical
gender of a noun in a particular language influences what adjectives are generated for it: feminine nouns tend to elicit relatively more feminine adjectives compared to masculine nouns.

The effect appeared even though the task was conducted in English and did not invite participants to think about biological or grammatical gender. This removes two possibilities for why it arose. The first possibility that can be ruled out is that phonological or morphological properties of the nouns used in the experiment made the grammatical gender more salient to the participants, since English nouns do not contain in themselves any information related to grammatical gender. Thus operations situated solely at the lexical level of processing could not explain the effect. Second, gender was not highlighted in the experimental context for the participants who received no instructions to personify. Additionally, the selection procedure for the items minimized the possibility that the stimuli set would implicitly push the participants towards thinking about gender when participating in the experiment. This suggests that the results could be taken as evidence for the existence of effects of grammatical gender on how referents of nouns are conceptualized. The exact mechanism through which this effect takes place is a question for future work, but the current study provides some suggestions. Similarly to some prior work (e.g. Vigliocco et al., 2005) grammatical gender had the most effect on adjectives for nouns denoting animals. The interaction between the noun's semantic category and the gender of the noun found in the current study provides some support for the hypothesis that anthropomorphization of the noun referent is the mechanism through which the effect comes into being.

Cross-linguistically nouns often constitute the most frequently occurring word class in a language (Liang \& Liu, 2013), and they are also used in everyday language for reference to humans, animals, objects, relations, categories and other types of entities. Because of this, the effect we found could have a broad and pervasive influence, affecting a wide range of processes relying on how referents of nouns are conceptualized.

The data also indicate that the effect did not emerge solely due to an invitation to personify the nouns that half of the participants received. Both participants who received instructions to personify and those who did not, showed effects of grammatical gender. It appears that an explicit invitation to personify (as when assigning names or voices in prior studies) is not necessary to induce effects of grammatical gender.

Of course, it is possible that participants personified the nouns even without being given the instruction to do so. For those who did not receive this instruction, the experiment did not make any suggestions for the participants to anthropomorphize the nouns' referents. If participants did engage in such unprompted personification, it seems likely that the effect would be observed outside of the context of this experimental task as well.

Last, we found that significant grammatical-gender
effects emerged starting with the second adjective participants generated for a given noun. This suggests that gender information is quite central in people's mental representations. There was however, no significant effect observed at the very first adjective.

This has the possibility of explaining why some studies using tasks where participants need to respond rapidly fail to observe effects of grammatical gender (e.g. Vigliocco et al., 2005; Kousta et al., 2008). The effect of grammatical gender is small in comparison to some other effects, for example the effect of cultural associations (Beller et al., 2015). It has also been suggested that more abstract pictures induce more schematic ways of conceptualizing what they depict and that verbal description of the less abstract pictures induces a similar effect (Holmes \& Wolff, 2010). This allows one to hypothesize that the strength of the gender effect in combination with its temporal development could leave it unnoticeable in conditions where the participants need to respond rapidly and where other perceptual or conceptual features of the stimulus are highly activated due to task demands or stimuli properties.

## Conclusion

The results obtained in this study support the view that grammatical gender affects object conceptualization. The effect was obtained in absence of any phonological or morphological aspects of the word carrying information about grammatical gender. Additionally, the effect was obtained without participants being invited to think about gender by any experimental instructions or demands (and participants could not have guessed that gender was of interest in the study). Finally, effects of grammatical gender emerged starting with the second adjective participants generated for a given noun. This suggests that gender information is quite central in people's mental representations (but did not emerge on the very first adjective). Furthermore, it is possible that the effect is quite pervasive, as it has the potential to affect anything that could be named by a noun.

## References

Bassetti, B. (2011) The grammatical and conceptual gender of animals in second language users. In Cook, V.J. and Bassetti, B. (Eds.) Language and Bilingual Cognition (pp. 357-384). Hove, UK: Psychology Press.
Bassetti, B., \& Nicoladis, E. (2016). Research on grammatical gender and thought in early and emergent bilinguals. Int. journal of bilingualism: interdisciplinary studies of multilingual behaviour, 20(1), 3-16.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Beller, S., Brattebø, K. F., Lavik, K. O., Reigstad, R. D., \& Bender, A. (2015). Culture or language: what drives effects of grammatical gender?. Cognitive Linguistics, 26(2), 331-359.
Corbett, G. (1991). Gender. Cambridge textbooks in
linguistics. Cambridge: Cambridge.
Davies, M. (2008). The Corpus of Contemporary American English: 425 million words, 1990 - present. Available online at http://corpus.byu.edu/coca/.
Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., \& Gibson, E. (2012). Verbal interference suppresses exact numerical representation. Cognitive psychology, 64(1), 74-92.
Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., \& Hothorn, M. T. (2016). Package 'multcomp'. Simultaneous inference in general parametric models. Project for Statistical Computing, Vienna, Austria.
Holmes, K. J., \& Wolff, P. (2010). Simulation from schematics: Dorsal stream processing and the perception of implied motion. In Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (pp. 2704-2709).
Kinoshita, S., \& Lupker, S. J. (2004). Masked priming: The state of the art. Psychology Press.
Koerner, E. F. (1992). The Sapir-Whorf Hypothesis: A Preliminary History and a Bibliographical Essay. Journal of Linguistic Anthropology, 2(2), 173-198.
Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. Journal of psycholinguistic research, 22(5), 519-534.
Kousta, S. T., Vinson, D. P., \& Vigliocco, G. (2008). Investigating linguistic relativity through bilingualism: The case of grammatical gender. JEP: Learning, Memory, and Cognition, 34(4), 843.
Liang, J., \& Liu, H. (2013). Noun Distribution in Natural Language. Poznań Studies in Contemporary Linguistics 49(4), 509-529
Medin, D. L., Goldstone, R. L., \& Gentner, D. (1993). Respects for Similarity. Psych. Review, 100(2), 254-278.
Phillips, W., \& Boroditsky, L. (2003) Can quirks of grammar affect the way you think? Grammatical gender and object concepts. Proceedings of the 25th annual meeting of the Cognitive Science Society (pp. 928-933).
Regier, T., Kay, P., Gilbert, A. L., \& Ivry, R. B. (2010). Which Side Are You on, Anyway? In B. Malt \& P. Wolff (Eds.), Words and the mind: How words capture human experience. Oxford: Oxford University Press.
Schyns, P. G., \& Rodet, L. (1997). Categorization creates functional features. Journal of Experimental Psychology Learning Memory and Cognition, 23, 681-696.
Sera, M. D., Berge, C. A., \& del Castillo Pintado, J. (1994). Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. Cognitive Development, 9(3), 261-292.
Tversky, A. (1977). Features of similarity. Psychological review, 84(4), 327.
Vigliocco, G., Vinson, D. P., Paganelli, F., \& Dworzynski, K. (2005). Grammatical gender effects on cognition: implications for language learning and language use. JEP: General, 134(4), 501.
Wolff, P., \& Holmes, K. J. (2011). Linguistic relativity. Wiley Interdisciplinary Reviews: Cognitive Science, 2(3), 253-265.

# Inferring Human Interaction from Motion Trajectories in Aerial Videos 

Tianmin Shu * Yujia Peng * Lifeng Fan Hongjing Lu Song-Chun Zhu<br>\{tianmin.shu, yjpeng, lfan\}@ucla.edu hongjing@ucla.edu sczhu@stat.ucla.edu Department of Psychology and Statistics, University of California, Los Angeles, USA


#### Abstract

People are adept at perceiving interactions from movements of simple shapes but the underlying mechanism remains unknown. Previous studies have often used object movements defined by experimenters. The present study used aerial videos recorded by drones in a real-life environment to generate decontextualized motion stimuli. Motion trajectories of displayed elements were the only visual input. We measured human judgments of interactiveness between two moving elements, and the dynamic change of such judgments over time. A hierarchical model was developed to account for human performance in this task, which represents interactivity using latent variables, and learns the distribution of critical movement features that signal potential interactivity. The model provides a good fit to human judgments and can also be generalized to the original Heider-Simmel animations (1944). The model can also synthesize decontextualized animations with controlled degree of interactiveness, providing a viable tool for studying animacy and social perception.


Keywords: social interaction; motion; decontextualized animation; hierarchical model; action understanding

## Introduction

People are adept at perceiving goal-directed action and inferring social interaction from movements of simple objects. In their pioneering work, Heider and Simmel (1944) presented video clips showing three simple geometrical shapes moving around, and asked human observers to describe what they saw. Almost all observers described the object movements in an anthropomorphic way, reporting a reliable impression of animacy and meaningful social interaction among the geometric shapes displayed in the decontextualized animation.

Later studies (Dittrich \& Lea, 1994; Scholl \& Tremoulet, 2000; Tremoulet \& Feldman, 2000, 2006; Gao, Newman, \& Scholl, 2009; Gao, McCarthy, \& Scholl, 2010) used more controlled stimuli and systematically examined what factors can impact the perception of goal-directed actions in a decontextualized animation. The results provided converging evidence that the perception of human-like interactions relies on some critical low-level motion cues, such as speed and motion direction. However, it remains unclear how the human visual system combines motion cues from different objects to infer interpersonal interactivity in the absence of any context cues.

To address this fundamental question, Baker, Saxe, and Tenenbaum (2009) developed a Bayesian model to reason about the intentions of an agent when moving in maze-like environments of the sort used by Heider and Simmel (1944). Other studies (Baker, Goodman, \& Tenenbaum, 2008; Ullman et al., 2009; Baker, 2012) developed similar models that could be generalized to situations with multiple agents and

[^200]

Figure 1: Stimulus illustration. (Left) An example frame of an aerial video recorded by a drone. Two people were being tracked (framed by red and green boxes). (Right) A sample frame of an experimental trial. The two people being tracked in the aerial video are presented as two dots, one in red and one in green, in a black background. A video demonstration can be viewed on the project website: http: / / www. stat.ucla.edu/~tianmin.shu/HeiderSimmel/CogSci17
different contexts. These modeling studies illustrate the potential fruitfulness of using a Bayesian approach as a principled framework for modeling human interaction shown in decontextualized animations. However, these models have been limited to experimenter-defined movements, and by computational constraints imposed by the modelers for particular application domains.

The present study aims to generate Heider-Simmel-type decontextualized animations using real-life videos of visual scenes. As a naturalistic example, imagine that you are watching a surveillance video recorded by a drone from a bird's eye view, as shown in Fig. 1. In such aerial videos, changes in human body postures can barely be seen, and the primary visual cues are the noisy movement trajectories of each person in the scene. This situation is analogous to the experimental stimuli used in Heider and Simmel's studies, but the trajectories of each entity are directly based on real-life human movements.

In the present study, we first used real-life aerial videos to generate decontextualized animations and to assess how human judgments of interactivity emerge over time. We developed a hierarchical model to account for human performance. One advantage of using aerial videos to generate decontextualized animations is that the technique provides sufficient training stimuli to enable the learning of a hierarchical model with hidden layers, which could illuminate the representations of critical movement patterns that signal potential interactivity between agents. Furthermore, we assessed whether the learning component in the model can be generalized to the original animations by Heider and Simmel (1944).

## Computational Model

We designed a hierarchical model with three layers. As shown in Fig. 2, the first layer (the $X$ layer) estimates spatiotemporal motion patterns within a short period of time.


Figure 2: Illustration of the hierarchical generative model. The solid nodes are observations of motion trajectories of two agents, and the remaining nodes are latent variables constituting the symbolic representation of an interaction, i.e., the original trajectories are coded as a sequence of sub-interactions $S$ and interaction labels $Y$.

The second layer (the $S$ layer) captures the involvement of various motion fields at different stages of interactivity over a long period by temporally decomposing interactivity with latent sub-interactions. The last layer (the $Y$ layer) indicates the presence or absence of interactiveness between two agents.

The inputs to the model are motion trajectories of two agents, denoted as $\Gamma_{a}=\left\{\mathbf{x}_{a}^{t}\right\}_{t=0, \cdots, T}, a=1,2$. The position of agent $a(a=1,2)$ at time $t$ is $\mathbf{x}_{a}^{t}=(x, y)$. The total length of the trajectory is $T$. Using the input of motion trajectories, we can readily compute the velocity sequence of agent $a$ $(a=1,2)$, i.e., $V_{a}=\left\{\mathbf{v}_{a}^{t}\right\}_{t=1, \cdots, T}$, where $\mathbf{v}_{a}^{t}=\mathbf{x}_{a}^{t}-\mathbf{x}_{a}^{t-1}$.

To capture the interactivity between two agents based on the observed trajectories of movements, the model builds on two basic components. (1) Interactivity between two agents can be represented by a sequence of latent motion fields, each capturing the relative motion between the two agents who perform meaningful social interactions. (2) Latent motion fields can vary over time, capturing the behavioral change of the agents over a long period of time. The details for quantifying the two key components are presented in the next two subsections.

## Conditional Interactive Fields

As illustrated in Fig. 3, we use conditional interactive fields (CIFs) to model how an agent moves with respect to a reference agent. We randomly select an agent to be the reference agent, and then model the partner agent's movement by estimating a vector field of the relative motion conditioned on a specific distribution of the reference agent's motion.

To ensure that the fields are orientation invariant, we perform a coordinate transformation as Fig. 3 illustrates. At each time point $t$, the transformed position of the reference agent is always located at $(0,0)$, and its transformed velocity direction is always pointed to the norm of the upward vertical direction. Consequently, the position and velocity of the second agent after the transformation, i.e., $\tilde{\Gamma}=\left\{\tilde{\mathbf{x}}^{t}\right\}_{t=0, \cdots, T}$ and $\tilde{V}=\left\{\tilde{\mathbf{v}}^{t}\right\}_{t=1, \cdots, T}$, can be used to model the relative motion.

For a sub-interaction $s$ (interactivity in a relatively short time sharing consistent motion patterns, e.g., approaching, walking together, standing together), we define its CIF as a


Figure 3: Illustration of a conditional interactive field (CIF): after a coordinate transformation w.r.t. the reference agent, we model the expected relative motion pattern $\tilde{\mathbf{x}}^{t}$ and $\tilde{\mathbf{v}}^{t}$ conditioned on the reference agent's motion.


Figure 4: Temporal parsing by $S$ (middle). The top demonstrates the change of CIFs in sub-interactions as the interaction proceeds. The bottom indicates the change of interactive behaviors in terms of motion trajectories. The colored bars in the middle depict the types of the sub-interactions.
linear dynamic system:

$$
\begin{equation*}
\tilde{\mathbf{v}}^{t} \sim \mathcal{N}\left(A_{S} \tilde{\mathbf{x}}^{t}+B_{s}, \Sigma_{s}\right) \tag{1}
\end{equation*}
$$

where $A_{s}, B_{s}$, and $\Sigma_{s}=\operatorname{diag}\left(\sigma_{s 1}^{2}, \sigma_{s 2}^{2}\right)$ are the parameters of the Gaussian distribution to be learned for each subinteraction $s . \quad A_{s} \tilde{\mathbf{x}}^{t}+B_{s}$ can be interpreted as the expected motion at location $\tilde{\mathbf{x}}$ in the field.

## Temporal Parsing by Latent Sub-Interactions

We assume that a long interactive sequence can be decomposed into several distinct sub-interactions each with a different CIF. For example, when observing that two people walk towards each other, shake hands and walk together, we can decompose this interactive sequence into three subinteractions. We represent meaningful interactivity as a sequence of latent sub-interactions $S=\left\{s_{k}\right\}_{k=1, \ldots, K}$, where a latent sub-interaction determines the category of the CIF involved in a time interval $\mathcal{I}_{k}=\left\{t: t_{k}^{1} \leq t \leq t_{k}^{2}\right\}$, such that $s^{t}=s_{k}, \forall t \in \mathcal{T}_{k} . s_{k}$ is the sub-interaction label in the $k$-th interval representing the consistent interactivity of two agents in the interval. Fig. 4 illustrates the temporal parsing.

In each interval $k$, we define an interaction label $y_{k} \in\{0,1\}$ to indicate the absence or presence of interactivity between the two agents. The interaction labels also constitute a sequence $Y=\left\{y^{t}\right\}_{t=1, \cdots, T}$. We have $y^{t}=y_{k}, \forall t \in \mathcal{T}_{k}$, where $y_{k}$ is the interaction label in interval $\mathcal{T}_{k}$.

## Model Formulation

Given the input of motion trajectories $\Gamma$, the model infers the posterior distribution of the latent variables $S$ and $Y$,

$$
\begin{equation*}
p(S, Y \mid \Gamma) \propto \underbrace{P(\Gamma \mid S, Y)}_{\text {likelihood }} \cdot \underbrace{P(S \mid Y)}_{\text {sub int. prior }} \cdot \underbrace{P(Y)}_{\text {int. prior }} . \tag{2}
\end{equation*}
$$

The likelihood assesses how well the motion fields under corresponding CIFs of sub-interactions can account for relative motion observed in the video input, the spatial density of the relative position and the observed motion of the reference agent:

$$
\begin{equation*}
p(\Gamma \mid S, Y)=\prod_{k=1}^{K} \prod_{t \in \mathcal{T}_{k}} p\left(\tilde{\mathbf{v}}^{t}, \tilde{\mathbf{x}}^{t}, \mathbf{v}_{1}^{t} \mid s^{t}=s_{k}, y^{t}=y_{k}\right) \tag{3}
\end{equation*}
$$

where

$$
\begin{align*}
& p\left(\tilde{\mathbf{v}}^{t}, \tilde{\mathbf{x}}^{t}, \mathbf{v}_{1}^{t} \mid s^{t}=s_{k}, y^{t}=y_{k}\right)  \tag{4}\\
& =\underbrace{p\left(\tilde{\mathbf{v}}^{t} \mid \tilde{\mathbf{x}}^{t}, s_{k}, y_{k}\right)}_{\text {rel. motion }} \cdot \underbrace{p\left(\tilde{\mathbf{x}}^{t} \mid s_{k}, y_{k}\right)}_{\text {rel. spatial density }} \cdot \underbrace{p\left(\| \mathbf{v}_{1}^{t}| | s_{k}, y_{k}\right)}_{\text {ref. motion }} .
\end{align*}
$$

Note that $\mathbf{v}_{1}^{t}$ is the reference agent's velocity. When $y_{k}=1$, the first term is defined in equation (1), the second term is learned by Gaussian kernel density estimation, and the third term is defined as a Weibull distribution, which is suitable for learning a long-tail distribution of a non-negative variable. When $y_{k}=0$, the first term is defined as a Gaussian distribution $\mathcal{N}\left([0,0]^{\top}, \Sigma_{0}=\operatorname{diag}\left(\sigma_{0}^{2}, \sigma_{0}^{2}\right)\right)$, and the remaining two terms are uniform distributions in quantized spaces.

We model the prior term of sub-interactions $P(S \mid Y)$ using two independent components, i) the duration of each subinteraction, and ii) the transition probability between two consecutive sub-interactions, as follows:

$$
\begin{equation*}
p(S \mid Y)=\prod_{k=1}^{K} \underbrace{p\left(\left|\mathcal{T}_{k}\right| \mid s_{k}, y_{k}\right)}_{\text {duration }} \prod_{k=2}^{K} \underbrace{p\left(s_{k} \mid s_{k-1}, y_{k}\right)}_{\text {transition }} \tag{5}
\end{equation*}
$$

When $y_{k}=1$, the two terms follow a log-normal distribution and a multinomial distribution respectively; when $y_{k}=0$, uniform distributions are used for the two terms instead.

Finally, we use a Bernoulli distribution to model the prior term of interactions $P(Y)$,

$$
\begin{equation*}
p(Y)=\prod_{k=1}^{K} \prod_{t \in \mathcal{T}_{k}} p\left(y^{t}=y_{k}\right)=\prod_{k=1}^{K} \prod_{t \in \mathcal{T}_{k}} \rho^{y^{t}}(1-\rho)^{1-y^{t}} \tag{6}
\end{equation*}
$$

## Inference and Prediction

The model infers the current status of latent variables and produces an online prediction of future trajectories. Inference and prediction are performed for each time point from 1 to $T$ sequentially (rather than offline prediction, which gives the labels after watching the entire video).

We denote trajectories from 0 to $t$ as $\Gamma_{0: t}$, and the subinteractions from 1 to $t-1$ as $S_{1: t-1}$. Without loss of generality, we assume there are $K$ sub-interactions in $S_{1: t-1}$ with $\mathcal{T}_{K}$ being the last interval and $s^{t-1}=s_{K}$. We first infer $s^{t}$ under the assumption of interaction (i.e., $y^{t}=1$ ) by maximizing

$$
\begin{equation*}
p\left(s^{t} \mid \Gamma_{0: t}, S_{1: t-1}, y^{t}\right) \propto p\left(\tilde{\mathbf{v}}^{t}, \tilde{\mathbf{x}}^{t}, v_{1}^{t} \mid s^{t}\right) p\left(s^{t} \mid S_{1: t-1}, y^{t}\right) \tag{7}
\end{equation*}
$$

where,

$$
\begin{align*}
& p\left(s^{t} \mid S_{1: t-1}, y^{t}\right) \\
& =\left\{\begin{array}{ll}
p\left(\tau \geq\left|\mathcal{T}_{k}\right|+1 \mid s^{t}=s^{t-1}, y^{t}\right) & \text { if } s^{t}=s^{t-1} \\
p\left(\tau \geq 1 \mid s^{t}, y^{t}\right) p\left(s^{t} \mid s^{t-1}\right) & \text { otherwise }
\end{array} .\right. \tag{8}
\end{align*}
$$

Then the posterior probability of $y^{t}=1$ given $s^{t} \in \mathcal{S}$ is defined as

$$
\begin{equation*}
p\left(y^{t} \mid s^{t}, \Gamma_{0: t}, S_{1: t-1}\right) \propto p\left(s^{t} \mid \Gamma_{0: t}, S_{1: t-1}, y^{t}\right) p\left(y^{t}\right) \tag{9}
\end{equation*}
$$

This computation makes it possible to perform the following inferences and online prediction: i) we maximize (7) to obtain the optimal $s^{t}$; ii) we use (9) to compute the posterior probability of two agents being interactive at $t$ under the CIF of $s^{t}$ as an approximation of the judgment of interaction/noninteraction provided by human observers; iii) the model can synthesize new trajectories using the following computation,

$$
\begin{align*}
& s^{t+1} \sim p\left(s^{t+1} \mid S_{1: t}, y^{t+1}\right)  \tag{10}\\
\mathbf{x}_{1}^{t+1}, \mathbf{x}_{2}^{t+1} & \sim p\left(\mathbf{x}_{1}^{t+1}, \mathbf{x}_{2}^{t+1} \mid \mathbf{x}_{1}^{t}, \mathbf{x}_{2}^{t}, s^{t+1}, y^{t+1}\right)  \tag{11}\\
& =p\left(\tilde{\mathbf{v}}^{t+1}, \tilde{\mathbf{x}}^{t+1}, v_{1}^{t+1} \mid s^{t+1}, y^{t+1}\right)
\end{align*}
$$

where $\tilde{\mathbf{v}}^{t+1}, \tilde{\mathbf{x}}^{t+1}$, and $v_{1}^{t+1}$ are given by $\mathbf{x}_{1}^{t}, \mathbf{x}_{1}^{t+1}, \mathbf{x}_{2}^{t}$ and $\mathbf{x}_{2}^{t+1}$, and the last term is defined in (4). By setting $y^{t+1}=1$ or $y^{t+1}=0$ in (10) and (11).

## Learning

## Algorithm

To train the model, we used Gibbs sampling to find the $S$ that maximizes the joint probability $P(Y, S, \Gamma)$. The implementation details are summarized below:

- Step 0: To initialize $S$, we first construct a feature vector for each time $t$, i.e., $\left[\left\|v_{1}^{t}\right\|, \tilde{\mathbf{x}}^{t}, \tilde{\mathbf{v}}^{t}\right]^{\top}$. A K-means clustering is then conducted to obtain the initial $\left\{s^{t}\right\}$, which also gives us the sub-interaction parsing $S$ after merging the same consecutive $s^{t}$.
- Step 1: At each time point $t$ of every training video, we update its sub-interaction label $s^{t}$ by

$$
\begin{equation*}
s^{t} \sim p\left(\Gamma \mid S_{-t} \cup\left\{s^{t}\right\}, Y\right) p\left(S_{-t} \cup\left\{s^{t}\right\} \mid Y\right) \tag{12}
\end{equation*}
$$

where $S_{-t}$ is the sub-interaction temporal parsing excluding time $t$, and $S_{-t} \cup\left\{s^{t}\right\}$ is a new sub-interaction sequence after adding the sub-interaction at $t$. Note that $Y$ is always fixed in the procedure; thus we do not need $p(Y)$ term for sampling purpose.

- Step 2: If $S$ does not change anymore, go to next step; otherwise, repeat step 1.
- Step 3: Since we do not include the non-interactive videos in the training set, we selected 22 videos in the first human experiment (a mixture of interactive and non-interactive videos) as a validation set to estimate $\rho$ and $\sigma_{0}$ by maximizing the correlation between the model prediction of (9) and the average human responses in the validation set.


## Model Simulation Results

We tested the model using two sets of training data. The first dataset is a UCLA aerial event dataset collected by Shu et al. (2015), in which about 20 people performed some group activities in two scenes (a park or a parking lot), such as group touring, queuing in front of a vending machine or playing frisbee. People's trajectories and their activities are manually annotated. The dataset is available at http://www. stat.ucla .edu/~tianmin.shu/AerialVideo/AerialVideo.html

We selected training videos including interactivity from the database, so that the two agents always interact with each other in all training stimuli. Thus, for any training video, $y^{t}=1, \forall t=1, \cdots, T$. During the training phase, we excluded the examples used in human experiments. In total, there were 131 training instances.

In the implementation, we manually define the maximum number of sub-interaction categories to be 15 in our full model (i.e., $|S|=15$ ), which is over-complete for our training data according to learning (low frequency in the tail of Fig. 6). With simulated annealing (Kirkpatrick, Gelatt, \& Vecchi, 1983), Gibbs sampling converges within 20 sweeps (where a sweep is defined as all the latent sub-interaction labels have been updated once). The frequencies of the top 15 CIFs are highly unbalanced. In fact, the top 10 CIFs account for $83.8 \%$ of the sub-interactions in the training data. The first row of Fig. 5 provides a visualization of the top 5 CIFs.

The second dataset was created from the original HeiderSimmel animation (i.e., two triangles and one circle). We extracted the trajectories of the three shapes, and thus obtained 3 pairs of two-agent interactions. We truncated the movie into short clips (about 10 seconds) to generate a total of 27 videos. The same algorithm was used to train the model with 15 types of CIFs. The most frequent five CIFs are visualized in the second row of Fig. 5. Clearly, the richer behavior in the Heider-Simmel animation yielded a variety of CIFs with distinct patterns compared to the CIFs learned from aerial videos. The frequencies of CIFs are also more distributed in this dataset, as shown in Fig. 6.

We observed a few critical CIFs that signal common interactions from the two simulation results. For instance, in aerial videos, we observed i) approaching, e.g., CIF 1 and ii) walking in parallel, or following, e.g., the lower part of CIF 2; the Heider-Simmel animation revealed additional patterns such as i) orbiting, e.g., CIF 1, ii) walking-by, e.g., CIF 5, and iii) leaving, e.g., CIF 4.

## Experiment

## Stimuli

24 interactive stimuli were generated from different pairs of human interactions in aerial videos. We selected two people interacting with each other in each aerial video. We then generated the decontextualized animations by depicting the two people as dots with different colors. The dots' coordinates were first extracted from the aerial videos by human annotators. Note that the two dots were first re-centered to localize
the midpoint at the center of the screen in the first frame. The coordinates were temporally smoothed by averaging across the adjacent 5 frames.

24 non-interactive stimuli were generated by interchanging motion trajectories of two people selected from two irrelevant interactive videos (e.g., the motion of one dot in video 1 recombined with the motion of a dot in video 2). The starting distances between two dots in non-interactive stimuli were kept the same as in the corresponding interactive stimuli.

The duration of stimuli varied from 239 frames to 500 frames (mean frame $=404$ ), corresponding to 15.9 to 33.3 seconds, with a recording refresh rate of 15 frames per second. The diameters of dots were $1^{\circ}$ of visual angle. One dot was displayed in red ( $1.8 \mathrm{~cd} / \mathrm{m}^{2}$ ) and the other in green $\left(30 \mathrm{~cd} / \mathrm{m}^{2}\right)$ on a black background $\left(0 \mathrm{~cd} / \mathrm{m}^{2}\right)$. Among the 48 pairs of stimuli, four pairs of actions (two interactive and two non-interactive) were used as practice.

## Participants

33 participants (mean age $=20.4 ; 18$ female) were enrolled from the subject pool at the University of California, Los Angeles (UCLA) Department of Psychology. They were compensated with course credit. All participants had normal or corrected-to-normal vision.

## Procedures

Participants were seated 35 cm in front of a screen, which had a resolution of $1024 \times 768$ and a 60 Hz refresh rate. First, participants were given a cover story: "Imagine that you are working for a company to infer whether two people carry out a social interaction based on their body locations measured by GPS signals. Based on the GPS signal, we generated two dots to indicate the location of the two people being tracked." The task was to determine when the two dots were interacting with each other and when they were not. Participants were asked to make continuous responses across the entire duration of the stimuli. They were to press and hold the left-arrow or right-arrow button for interactive or non-interactive moments respectively, and to press and hold the down-arrow button if they were unsure. If no button was pressed for more than one second, participants received a 500 Hz beep as a warning.

Participants were presented with four trials of practice at the beginning of the session to familiarize them with the task. Next, 44 trials of test stimuli were presented. The order of trials was randomized for each participant. No feedback was presented on any of the trials. The experiment lasted for about 30 minutes in total.

## Results

Interactive, unsure and non-interactive responses were coded as $1,0.5$, and 0 , respectively. Frames with no responses were removed from the comparison. Human responses were shown in Fig. 8 (left). A paired-sample t-test revealed that the average ratings of non-interactive actions $(\mathrm{M}=0.34, \mathrm{SD}=0.13)$ were significantly lower than interactive actions $(M=0.75$,


Figure 5: Interactive fields of the top five frequent CIFs learned from aerial videos (top) and Heider-Simmel movie (bottom) respectively. In each field, the reference agent (red dot) is at the center of a field i.e., $(0,0)$, moving towards north; the arrows represent the mean relative motion at different locations and the intensities of the arrows indicate the relative spatial density which increases from light to dark.


Figure 6: The frequencies of learned CIFs with the training data generated from aerial videos (top) and the Heider-Simmel movie (bottom). The numbers on the x axis indicate the IDs of CIFs, ranked according to the occurrence frequency in the training data.

| Method | HMM | One-Interaction | Hierarchical Model |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $\|\mathcal{S}\|=5$ | $\|\mathcal{S}\|=10$ | $\|\mathcal{S}\|=15$ |
| $r$ | 0.739 | 0.855 | 0.882 | 0.911 | 0.921 |
| RMSE | 0.277 | 0.165 | 0.158 | 0.139 | 0.134 |

Table 1: The quantitative results of all methods in experiment 1 using aerial videos as training data.
$\mathrm{SD}=0.13), t(32)=13.29, p<0.001$. This finding indicates that human observers are able to discriminate interactivity based on decontextualized animations generated from the real-life aerial videos.

To compare the model predictions with human continuous judgments, we computed the average human ratings, and ran the model to simulate online predictions of sub-interaction and interaction labels on the testing videos (excluding the ones in the validation set). Specifically, we used (9) to compute the probability of two agents being interactive with each other at any time point $t$. The model simulation used the hyper-parameters $\rho=10^{-11}$ and $\sigma_{0}=1.26$.

Table 1 summarizes the Pearson correlation coefficient $r$ and root-mean-square error (RMSE) between the model predictions and the human ratings using aerial videos as training data. We compare our hierarchical model with two baseline models: i) Hidden Markov Model (HMM), where the latent variables $s^{t}$ and $y^{t}$ only depend on their preceding variables $s^{t-1}$ and $y^{t-1}$; ii) a model with only one type of subinteraction. Both models yielded poorer fits to human judgments (i.e., lower correlation and higher RMSE) than the hierarchical model. In addition, we changed the number of subinteraction categories to examine how sensitive our model is
to this parameter. The results clearly show that i) only using one type of sub-interaction provides reasonably good results, $r=.855$, and ii) by increasing the number of sub-interactions $|S|$, the fits to human ratings were further improved until reaching a plateau with a sufficiently large number of subinteractions.

Fig. 7 shows results for a few videos, with both model predictions and human ratings. The model predictions accounted for human ratings quite well in most cases. However, the model predictions were slightly higher than the average human ratings, which may be due to the lack of negative examples in the training phase. We also observed high standard deviations in human responses, indicating the large variability of the online prediction task for every single frame in a dynamic animation. In general, the difference between our model's predictions and human responses are seldom larger than one standard deviation of human responses.

We also tested the model trained from the Heider-Simmel movie on the same testing set (generated from the aerial videos), yielding a correlation of 0.640 and RMSE of 0.227 . The reduced fitting result indicates the discrepancy between two types of videos. The CIFs learned from one dataset may be limited in generalization to the other dataset.

One advantage of developing a generative model is that it enables the synthesis of new videos by (10) and (11), based on randomly sampled initial positions of the two agents ( $\mathbf{x}_{1}^{0}$, $\mathbf{x}_{2}^{0}$ ) and the first sub-interaction $s^{1}$. By setting the interaction labels to be 1 or 0 , the synthesized stimuli can be controlled to vary the degree of interactiveness. We ran a second experiment using model synthesized animations ( 10 interactive and 10 non-interactive clips). These synthesized videos were presented to human observers in random orders and the interactive ratings were recorded. The interactiveness between the two agents in the synthesized videos was judged accurately by human observers (mean rating of 0.85 for synthesized interactive clips, and 0.15 for non-interactive clips), suggesting that the model effectively captured the visual features that signal potential interactivity between agents.


Figure 7: Comparison of online predictions by our full model $(|\mathcal{S}|=15)$ (orange) and humans (blue) over time (in seconds) on testing videos. The shaded areas show the standard deviations of human responses at each moment.


Figure 8: Mean ratings of the interactive versus non-interactive actions in the experiment. Error bars indicate $+/-1$ SEM.

## Conclusion

In this paper, we examined human perception of social interactions using decontextualized animations based on movement trajectories recorded in aerial videos of a real-life environment, as well as Heider-Simmel-type animations. The proposed hierarchical model built on two key components: conditional interactive fields of sub-interactions, and temporal parsing of interactivity. The model fit human judgments of interactiveness well, and suggests potential mechanisms underlying our understanding of meaningful human interactions. Human interactions can be decomposed into sub-interactions such as approaching, walking in parallel, or standing still in close proximity. Based on the transition probabilities and the duration of sub-components, humans are able to make inferences about how likely the two people are interacting.

The model could be extended to be applied to the field of behavioral recognition. While previous work has focused on actions of individuals based on detecting local spatialtemporal features embedded in videos (Dollár, Rabaud, Cottrell, \& Belongie, 2005), the current work can deal with multiagent interaction. Understanding of the relation between agents could facilitate the recognition of individual behaviors by putting single actions into meaningful social contexts. In addition, the current model is only based on visual motion cues. The model could be enhanced by incorporating a cognitive mechanism (e.g., a theory-of-mind framework) to enable explicit inference of intentions.

## Acknowledgement

This research was funded by a NSF grant BCS-1353391 to HL and DARPA MSEE project FA 8650-11-1-7149 and ONR MURI project N00014-16-1-2007 for SZ.

## References

Baker, C. L. (2012). Bayesian theory of mind: modeling hu-
man reasoning about beliefs, desires, goals, and social relations. Unpublished doctoral dissertation, Massachusetts Institute of Technology.
Baker, C. L., Goodman, N. D., \& Tenenbaum, J. B. (2008). Theory-based social goal inference. In Proceedings of the thirtieth annual conference of the cognitive science society (p. 1447-1452).

Baker, C. L., Saxe, R., \& Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3), 329349.

Dittrich, W. H., \& Lea, S. E. (1994). Visual perception of intentional motion. Perception, 23(3), 253-268.
Dollár, P., Rabaud, V., Cottrell, G., \& Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In In proceedings of ieee international conference on computer vision workshops (pp. 65-72).
Gao, T., McCarthy, G., \& Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. Psychological Science, 21, 1845-1853.
Gao, T., Newman, G. E., \& Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. Cognitive Psychology, 59(2), 154-179.
Heider, F., \& Simmel, M. (1944). An experimental study of apparent behavior. American Journal of Psychology, 57(2), 243-259.
Kirkpatrick, S., Gelatt, C. D., \& Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220(4598), 671680.

Scholl, B. J., \& Tremoulet, R. D. (2000). Perceptual causality and animacy. Trends in Cognitive Sciences, 4(8), 299-309.
Shu, T., Xie, D., Rothrock, B., Todorovic, S., \& Zhu, S.-C. (2015). Joint inference of groups, events and human roles in aerial videos. In Proceedings of ieee conference on computer vision and pattern recognition.
Tremoulet, P. D., \& Feldman, J. (2000). Perception of animacy from the motion of a single object. Perception, 29(8), 943-951.
Tremoulet, P. D., \& Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. Perception \& Pyschophysics, 68(6), 1047-1058.
Ullman, T., Baker, C. L., Macindoe, O., Evans, O., Goodman, N., \& Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In Proceedings of advances in neural information processing systems (p. 18741882).

# Children's spontaneous comparisons from 26 to 58 months predict performance in verbal and non-verbal analogy tests in 6th grade 

Catriona Silvey (casilvey@uchicago.edu)<br>Department of Psychology, 5848 S. University Avenue<br>Chicago, IL 60637 USA<br>Dedre Gentner (gentner@northwestern.edu)<br>Department of Psychology, 2029 Sheridan Road<br>Evanston, IL 60208 USA<br>Lindsey Richland (Irichland@uchicago.edu)<br>Department of Comparative Human Development, 1126 E. 59th St.<br>Chicago, IL 60637 USA<br>Susan Goldin-Meadow (sgm@uchicago.edu)<br>Departments of Psychology and Comparative Human Development, 5848 S. University Avenue/1126 E. 59th St. Chicago, IL 60637 USA


#### Abstract

Comparison supports the development of children's analogical reasoning. The evidence for this claim comes from laboratory studies. We describe spontaneous comparisons produced by 24 typically developing children from 26 to 58 months. Children tend to express similarity before expressing difference. They compare objects from the same category before objects from different categories, make global comparisons before specific comparisons, and specify perceptual features of similarity/difference before non-perceptual features. We then investigate how a theoretically interesting subset of children's comparisons - those expressing a specific feature of similarity or difference - relates to analogical reasoning as measured by verbal and non-verbal tests in 6th grade. The number of specific comparisons children produce before 58 months predicts their scores on both tests, controlling for vocabulary at 54 months. The results provide naturalistic support for experimental findings on comparison development, and demonstrate a strong relationship between children's early comparisons and their later analogical reasoning.


Keywords: comparison; similarity; language development; analogy

## Introduction

Comparison - the process of jointly examining two objects or events and assessing their similarities and differences - is crucial in the development of children's word learning, categorization, and analogical reasoning skills (Namy \& Gentner, 2002; Gentner \& Namy, 2006; Gentner, Anggoro, \& Klibanoff, 2011; Richland \& Simms, 2015). Comparison is an effective learning tool because it promotes structural alignment: the mapping of two representations in a way that enables the recognition of relational commonalities and alignable differences. A large body of experimental work shows that inviting children to compare exemplars helps them to move beyond overall or global similarity to more specific kinds of similarity, including similarity based on relational commonalities, as in analogical reasoning (Loewenstein \& Gentner, 2001; Christie \& Gentner, 2014; Gentner et al.,
2016). However, to get a full picture of the role of comparison in the development of children's analogical reasoning skills, it is important to relate this experimental work to children's spontaneous behavior in a naturalistic environment. Previous work has shown that children spontaneously produce comparative utterances from early in their language development: for example, children spontaneously generate metaphors from the age of around 2 (Winner, 1979) and are able to explain them in terms of similarity (Billow, 1981). However, the nature of the comparisons children produce is not static over time, but follows a developmental trajectory. Özçalışkan, Goldin-Meadow, Gentner, and Mylander (2009) found that while children's earliest comparisons tended to be between objects that were similar to each other in many features, the acquisition of the word 'like' was associated with an increase in the number of comparisons between objects that only shared a single feature. These specific comparisons are argued to be a more sophisticated stage in the development of children's understanding of similarity than are global comparisons (Smith, 1989; Gentner \& Rattermann, 1991). As such, the prevalence of specific comparisons in children's early speech could potentially be an index of their later analogical reasoning skill.

The current work has two aims: 1) a descriptive aim, to characterize common patterns in the development of children's spontaneous comparisons produced in naturalistic contexts in the home; 2) an inferential aim, to test the hypothesis that variation in children's production of specific, singlefeature comparisons predicts variation in their scores on tests of analogical ability given much later, in 6th grade.

## Methods

## Participants

24 children and their primary caregivers were drawn from a larger sample of 64 families who participated in a longitudi-
nal study of language development (the same sample drawn on by Özçalışkan et al., 2009). Families were recruited via direct mailings to targeted zip codes and an advertisement in a free monthly parenting magazine. Parents who responded were interviewed regarding background characteristics, and the final sample was selected to be representative of the greater Chicago area in terms of race, ethnicity and income. The sub-sample of 24 families in the current study was selected randomly, within the constraints of preserving the demographic spread of the original sample. Of the 24 children, 11 were male and 13 female; 18 were white (of whom 3 were Hispanic), 3 were Black or African-American, and 3 were of two or more races. The distribution of socio-economic status across the 24 families was similar to that of the original sample, ranging from families with an income of under $\$ 15,000$ where the primary caregiver had some high school education, to families with an income of over $\$ 100,000$ where the primary caregiver had an advanced degree.

## Procedure

Parents and children were visited in their homes and videotaped engaging in their normal daily activities for 90 minutes. Home visits began when the children were 14 months old and continued at 4-month intervals, ending when the children were 58 months old ( 12 sessions in total). ${ }^{1}$ All child speech, and all parent speech directed to the child, was transcribed. Transcription reliability was established by having a second individual transcribe $20 \%$ of each transcriber's tapes. Reliability was at or above $95 \%$.

## Coding

Comparisons were coded from the transcripts of child speech during the 12 sessions. The criterion for a comparison was that the child expressed a similarity or difference between an identifiable source and target. Sources and targets could be objects or events. In cases where the source and target of the comparison were unclear from the transcript alone, the original video was consulted. For each identified comparison, we coded the following:
Word. The word that made the utterance a comparison; e.g. 'I'm a funny one like you' would be coded as 'like'.
Word category. Comparative words were classified into six categories: like (the words 'like' and 'alike'), same/different (the words 'same' and 'different'), comparative/superlative (any comparative or superlative adjective, e.g., 'bigger', 'best'), too (used either in contexts like 'too big' or contexts like 'I'm dancing too'), match (e.g., 'these match each other'), and other.
Object or event. Comparisons were coded for whether the Source and Target were objects (e.g., 'this [rug] look like a skirt') or events (e.g., 'I win too').

[^201]Expressing similarity or difference. Comparisons were coded for whether they expressed similarity (e.g. 'go like a elephant') or difference (e.g. 'I'm bigger than everybody!').
Global or specific comparison. Comparisons were coded for whether they expressed global similarity/difference (e.g., for Objects, 'I have toys just like yours'; for Events, 'they both win'), or specific similarity/difference (e.g., for Objects, 'red like the ladybug'; for Events, 'I go a lot faster than when I was three'). Comparisons could be specific even if the objects compared were overall similar, e.g., 'this [tree] is the tallest [tree]'. We expect global comparisons to appear earlier than specific comparisons (Smith, 1989; Gentner \& Rattermann, 1991).

Feature specified. Where a feature of similarity or difference was specified, this feature was coded. Features were classified into 6 categories: Spatial (e.g., size, shape, distance, speed), Sensory (e.g., color, weight, taste, smell), Evaluative (e.g., goodness, prettiness, badness), Emotion (e.g., being tired, mad, scared), Preference (e.g., liking one thing better than another thing), ${ }^{2}$ and Other. Features were also classified as Perceptual (based on a readily perceptible attribute, e.g. color, size) or Non-Perceptual (based on a more abstract, not directly perceptible feature, e.g., preference, goodness).
Within or between-category comparison. Comparisons were coded for whether the objects compared were from the same or different superordinate categories. Superordinate categories were taken from Özçalışkan et al. (2009), with three additions to accommodate new data (in italics): people, animals, body parts, vehicles, clothing, furniture, appliances, kitchen utensils, tools, musical instruments, food, plants, activity toys, places, decorations/crafts, words/letters, and shapes.

In the case of events, the objects of interest were those with corresponding roles in the two events. For example, if the parent said she was going to use some yellow paint, and the child said 'think I'll do yellow too', the objects in corresponding roles (parent/child, and yellow paint/yellow paint) are in the same superordinate categories (people and decorations/crafts, respectively). This would therefore be coded as a within-category comparison. If the child said 'I'm going to act like a bee', the objects in corresponding roles (child and bee) are in different superordinate categories (people and animals); this would therefore be coded as a between-category comparison. If children initially rely on overall similarity, then within-category comparisons should emerge earlier than between-category comparisons.

A total of 532 comparisons were codable under these guidelines.

## Later outcomes

The same children were followed longitudinally as part of an ongoing language development project. When the chil-

[^202]

## Age in months

Figure 1: Frequencies of word categories across sessions.
dren were in 6th grade (aged around 13 years), we administered two tests of analogical reasoning: the Verbal Analogies subtest of the Woodcock-Johnson Tests of Cognitive Abilities (Woodcock, McGrew, \& Mather, 2001), and a nonverbal test, Raven's Progressive Matrices (Raven, 1938). The Woodcock-Johnson Verbal Analogies is an orally administered test that consists of sets of paired items. The participant has to fill in the missing item by abstracting the relation that holds between the first pair. For example, the participant is given the prompt 'mother is to father, as sister is to...', and expected to fill in the missing term 'brother'. Raven's Progressive Matrices consists of a series of geometric analogy problems. The participant is presented with a matrix that has one entry missing and must select the correct entry from an array of $6-8$ choices. These two measures were taken as outcomes in our analyses.

## Results

## Onset and prevalence of comparisons

Children varied in the age at which they produced their first comparison. For the purpose of this analysis, age of onset was defined as children's age during the session where they produced at least one comparison and also produced at least one comparison during the immediately following session. Under this criterion, the earliest onset was at 26 months, and the latest was at 50 months. The average age of onset was 36 months, with a standard deviation of 6 months. Comparisons were relatively infrequent: they ranged from $0 \%$ to $2.2 \%$ of a child's utterances in a given session. However, the fact that we reliably find comparisons even in short 90-minute sessions suggest they are a robust feature of children's early talk.


Figure 2: Frequencies of comparisons expressing similarity and difference across sessions.

## Comparison words

The most commonly used comparison word was 'like', followed by 'too', 'bigger', 'same', and 'better'. Together, these words accounted for $73 \%$ of the comparisons the children expressed. Table 1 shows counts and percentages for the word categories detailed in the Methods.

Table 1: Word categories.

| Word category | Number of uses | Percent |
| :--- | :--- | :--- |
| like | 219 | $41 \%$ |
| comparative/superlative | 142 | $27 \%$ |
| too | 76 | $14 \%$ |
| same/different | 45 | $8 \%$ |
| other | 34 | $6 \%$ |
| match | 16 | $3 \%$ |

Figure 1 shows the frequencies of the 4 most prevalent word categories over sessions. 'Like' is the first word category to reliably emerge. While 'like' and comparatives/superlatives are overall more frequent, all word categories generally show an increase in use across sessions.

## Expressing similarity and difference

Figure 2 shows the trend over sessions for expressing similarity versus difference. Similarity comparisons were more numerous overall ( 346 to 186). The general trend was for similarity comparisons to emerge earlier than difference comparisons, and to remain more numerous until the final session. On a by-individual level, 20 out of 24 children produced a similarity comparison before they produced a difference comparison; 1 produced a difference comparison before produc-


Figure 3: Frequencies of global and specific comparisons across sessions.
ing a similarity comparison; and 3 produced examples of both simultaneously. This trend for similarities to precede differences was significant, $\chi^{2}=27.25, p<.001$.

## Objects and events

While object comparisons were more numerous in general (358 compared to 174 event comparisons), the overall trend was for object and event comparisons to emerge at around the same time. 11 out of 24 children produced an object comparison before they produced an event comparison; 8 produced an event comparison before they produced an object comparison; and 5 produced examples of both simultaneously. The trend in ordering was not significant, $\chi^{2}=2.25, p=.32$. Thus it appears that from comparison onset, children are capable of expressing comparisons between events as well as comparisons between objects.

## Global and specific comparisons

The numbers of global and specific comparisons were broadly equivalent: 249 global to 283 specific. Figure 3 shows the trend over sessions. Global comparisons appear to be more numerous than specific comparisons in the first two sessions; in subsequent sessions they are at equivalent levels, until the final two sessions when specific comparisons are higher. By individuals, as predicted, global comparisons tended to precede specific comparisons: 14 of 24 children produced a global comparison before they produced a specific comparison, while 5 produced a specific comparison before they produced a global comparison, and 5 produced both in the same session. While not as strong as the tendency for similarity to precede difference, this trend in ordering was significant, $\chi^{2}=6.75, p=.034$.

## Features specified

The most frequently specified features were spatial or sensory; together, these accounted for $70 \%$ of the specific comparisons the children expressed. Table 2 shows overall counts and percentages.

Table 2: Feature categories.

| Feature category | Number of uses | Percent |
| :--- | :--- | :--- |
| Spatial | 136 | $48 \%$ |
| Sensory | 62 | $22 \%$ |
| Evaluative | 49 | $17 \%$ |
| Other | 30 | $11 \%$ |
| Emotion | 4 | $1 \%$ |
| Preference | 3 | $1 \%$ |

More perceptual features (202) were specified than nonperceptual features (80). The overall trend was for perceptual features to be specified earlier: by individual, 16 children specified perceptual features before they specified nonperceptual features, 4 specified non-perceptual features before they specified perceptual features, and 4 did both in one session. The trend for perceptual features to be specified first was significant, $\chi^{2}=12, p=.002$.

## Within- and between-category comparisons

Comparisons between objects in the same superordinate category (or between events involving objects in the same superordinate categories) were more numerous than comparisons between different superordinate categories (421 compared to 133). As predicted, comparisons between objects in the same category generally tended to precede comparisons between objects in different categories. 14 of 24 children produced a within-category comparison before a between-category comparison. 5 produced a between-category comparison first, and 5 children did both in one session. This trend in ordering was significantly different from chance, $\chi^{2}=6.75, p=.034$.

## Comparison type interactions

We also examined interactions between comparison types.
Firstly, we asked whether the children's comparisons expressing similarity were more likely to specify a feature than their comparisons expressing difference, or vice versa. 118 (34\%) of similarity comparisons specified a feature of similarity, while 165 ( $89 \%$ ) of difference comparisons specified a feature of difference. Given their marginal totals, similarity comparisons were less likely than expected to specify features, and difference comparisons were more likely than expected to specify features. This difference was significant, $\chi^{2}=145.38, p<.001$.

We then asked whether comparisons involving objects in the same superordinate category were more likely to express similarity or difference, as opposed to comparisons involving objects in different superordinate categories. Comparisons of within-category objects, or events involving within-category


Figure 4: Scatterplot showing number of specific comparisons produced from 26-58 months (x axis) and score in Verbal Analogies test in 6th grade (y axis).
objects, were broadly as likely to express similarity as difference: $240(60 \%)$ of these expressed similarity. On the other hand, comparisons of between-category objects, or events involving between-category objects, were more likely to express similarity ( 105 , or $80 \%$ ) than difference. This trend was significant, $\chi^{2}=15.32, p<.001$.

## Relation to later outcomes

We then tested the hypothesis motivated in the Introduction, that the number of specific comparisons (expressing a single feature of similarity or difference) that children made during the 12 observational sessions would predict their performance on tests of analogical reasoning in 6th grade.

Our outcome measures were the two analogy tests described in the Methods: the Woodcock-Johnson Verbal Analogies test, and Raven's Progressive Matrices. Both a verbal and a non-verbal test were administered in order to address the potential confound of language skill, which could influence both children's comparison production and their verbal analogy test scores. To further account for language proficiency, we controlled for the child's score on the Peabody Picture Vocabulary Test (PPVT-III; Dunn \& Dunn, 1997) at 54 months (the penultimate session of the 12 during which comparisons were collected).

Figures 4 and 5 show scatterplots of the relationship between the number of specific comparisons the children produced during the pre-school observation sessions and their 6th grade scores on the Verbal Analogies and Raven's Progressive Matrices tests, respectively.

Table 3 shows the results of the statistical model predicting Verbal Analogies score from specific comparison count and


Figure 5: Scatterplot showing number of specific comparisons produced from 26-58 months (x axis) and score in Raven's Progressive Matrices test in 6th grade (y axis).

PPVT at 54 months. Specific comparisons remained a significant predictor after controlling for PPVT, although PPVT had a larger effect. The adjusted $R^{2}$ for the model was .64, indicating that these two variables together explain around two-thirds of the variance in Verbal Analogies score.

Table 3: Verbal Analogies model

| Predictor | Standardized $\beta$ | $t$ | $p$ |
| :--- | :--- | :--- | :--- |
| \# specific comparisons | 0.37 | 2.44 | .024 |
| PPVT at 54 months | 0.55 | 3.60 | .002 |

Table 4: Raven's Progressive Matrices model

| Predictor | Standardized $\beta$ | $t$ | $p$ |
| :--- | :--- | :--- | :--- |
| \# specific comparisons | 0.67 | 4.27 | $<.001$ |

Table 4 shows the results of the model predicting Raven's Progressive Matrices score from specific comparison count. In this case, a likelihood ratio test showed that adding PPVT did not improve the model, $F(1)=1.05, p=.318$. The adjusted $R^{2}$ for the model was .43 , indicating that specific comparison count alone explains around $40 \%$ of the variance in Raven's Progressive Matrices scores.

## Discussion

Children's earliest comparisons tend to express global similarity between objects or events within the same superordi-
nate category. Later in development, children begin to express difference, to specify features of comparison, and to compare objects and events from different superordinate categories. Turning to the content of these comparisons, children are particularly motivated to comment first on similarities and differences in perceptual features such as size, color, and speed, and later on evaluative features such as goodness, prettiness, and their opposites.

While children are more likely to express global similarity than specific similarity, most difference comparisons are specific rather than global. This finding suggests that children are less motivated to comment on overall dissimilarity than on overall similarity: differences are only interesting insofar as they are specific. We also find that comparisons involving objects in different superordinate categories tend to disproportionately express similarity, rather than difference, despite these objects being a priori less similar to each other. This seemingly counter-intuitive result backs up existing theory: more similar objects are more likely to have salient, alignable differences than objects which are dissimilar (Markman \& Gentner, 1993; Gelman, Raman, \& Gentner, 2009).

The relationship we find between children's early comparisons and their later analogical reasoning skill can potentially be interpreted in a number of ways. One possibility is that children who make more specific comparisons gain more practice in identifying dimensions of similarity or difference: thus, making these comparisons directly helps build their analogical skills in ways that persist through later development. Another possibility is that both our predictor variable (the prevalence of specific comparisons in the pre-school years) and our outcome variable (performance on verbal and nonverbal analogy tests in 6th grade) can be traced back to an underlying variable such as intelligence. The current work cannot tease these explanations apart. However, in future work, we aim to code the comparisons parents produce during the sessions before their children start producing comparisons themselves. It will then be possible to use causal modeling to investigate the extent to which parent comparison input predicts child comparison production, controlling for parent IQ. If parent comparison input influences child production of comparisons beyond a heritable IQ effect, this outcome could potentially open the door for interventions aimed at boosting children's comparison production in the home by providing them with particularly helpful kinds of input.

## Acknowledgments

Thanks to Anjali Murthy for help with coding scheme refinement. This research was funded by NIH P01 HD040605 to Goldin-Meadow.

## References

Billow, R. M. (1981). Observing spontaneous metaphor in children. Journal of Experimental Child Psychology, 31, 430-445.

Christie, S., \& Gentner, D. (2014). Language helps children succeed on a classic analogy task. Cognitive Science, 38, 383-397.
Dunn, L. M., \& Dunn, L. M. (1997). PPVT-III: Peabody Picture Vocabulary Test. Circle Pines, MN: American Guidance Service.
Gelman, S. A., Raman, L., \& Gentner, D. (2009). Effects of language and similarity on comparison processing. Language Learning and Development, 5(3), 147-171.
Gentner, D., Anggoro, F. K., \& Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. Child Development, 82(4), 1173-88.
Gentner, D., Levine, S. C., Ping, R., Isaia, A., Dhillon, S., Bradley, C., \& Honke, G. (2016). Rapid learning in a children's museum via analogical comparison. Cognitive Science, 40(1), 224-240.
Gentner, D., \& Namy, L. L. (2006). Analogical processes in learning. Current Directions in Psychological Science, 15(6), 335 - 361.
Gentner, D., \& Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman \& J. P. Brynes (Eds.), Perspectives on thought and language: Interrelations in development (pp. 225-277). Cambridge: Cambridge University Press.
Loewenstein, J., \& Gentner, D. (2001). Spatial mapping in preschoolers: Close comparisons facilitate far mappings. Journal of Cognition and Development, 2(2), 189-219.
Markman, A. B., \& Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. Journal of Memory and Language, 32(4), 517-535.
Namy, L. L., \& Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. Journal of Experimental Psychology: General, 131(1), 5-15.
Özçalışkan, Ş., Goldin-Meadow, S., Gentner, D., \& Mylander, C. (2009). Does language about similarity play a role in fostering similarity comparison in children? Cognition, 112(2), 217-228.
Raven, J. C. (1938). Progressive matrices: A perceptual test of intelligence. London: HK Lewis.
Richland, L. E., \& Simms, N. (2015). Analogy, higher order thinking, and education. Wiley Interdisciplinary Reviews: Cognitive Science, 6(2), 177-192.
Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou \& A. Ortony (Eds.), Similarity and analogical reasoning (pp. 225-277). New York: Cambridge University Press.
Winner, E. (1979). New names for old things: The emergence of metaphoric language. Journal of Child Language, 6(03), 469-491.
Woodcock, R. W., McGrew, K. S., \& Mather, N. (2001). Woodcock-Johnson III tests of cognitive abilities. Itasca, IL: Riverside Publishing.

# Learning About Causal Systems Through Play 

Zi L. Sim* (zi@berkeley.edu)<br>Kuldeep K. Mahal* (kmahal@berkeley.edu)<br>Fei Xu (fei_xu@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94720 USA


#### Abstract

It is commonly believed that children are able to learn through play. Recent studies have found that children are able to learn causal rules through free play (Sim \& Xu, in press). One such study found that children learned how to correctly activate machines, using either a block that was the same shape or the same color as the machine, when given five minutes to play with them. However, would children be able to learn a more complex causal rule through free play as well and would their performance be comparable to children who were didactically taught the same causal rule? In the current study, we show that children are able to learn more complex causal rules through free play. We also show that children perform significantly better when learning these rules through free play or by first engaging in free play and then observing, as opposed to solely through observation.


Keywords: free play; causal learning; generalization

## Introduction

It is widely accepted that play is important for young children. Most elementary schools have a designated play time, where children are free to play and socialize with their peers. Studies have shown that on average children spend 51 hours a week, or $30 \%$ of their week, engaging in free play (Hofferth \& Sandberg, 2001). Play is also commonly encouraged by parents and educators, and access to play has been recognized by the United Nations Convention on the Rights of the Child (UNCRC) as a fundamental human right (Davey \& Lundy, 2011). It is evident that children spend a lot of time playing and are encouraged to do so, but it remains unclear why this is the case. Why do children spend such a large portion of their time playing, and why do adults actively encourage play behavior?

One explanation is that play has the potential to result in better learning than direction instruction because play provides learners with the opportunity to choose what they want to do (Whitebread, Coltman, Jameson, \& Lander, 2009; Weisberg, Hirsh-Pasek, \& Golinkoff, 2013; Weisberg, HirshPasek, Golinkoff, \& McCandliss, 2014). By doing so, learners may better encode new information in their memory (Metcalfe \& Kornell, 2005); process the problem structure more deeply (Sobel \& Kushnir, 2006); they may pay more attention and are more motivated (Corno \& Mandinach, 1983); or they may be able to focus on acquiring data that can address gaps in their knowledge (Markant \& Gureckis, 2013). There is some empirical evidence that adults benefit from the opportunity to select the information that they want to learn. In fact, research has shown that adults learn better when they
engage in active hypothesis testing, where they are able to select the data they observe to test their hypotheses, as compared to those who engaged in reception learning, where they observed data generated by another adult (Castro et al., 2009; Markant \& Gureckis, 2014; Sobel \& Kushnir, 2006).

For example, in one study, adults were shown a spectrum of sixteen "alien eggs" on a computer that went from spiky to smooth. Students were told that spiky eggs most likely hatched into alien snakes while smooth eggs most likely hatched into alien birds. They were asked to determine the boundary between the two types of eggs, so they had to determine the point at which the eggs shifted from hatching one species to the other species. To do this, subjects either selected a sequence of eggs to see which animal hatched from them (active learning condition) or observed randomly selected eggs being hatched (random condition). The study found that participants in the active learning condition generally performed better than those in the random condition (Castro et. al, 2009), which suggests that adults experience benefits in learning when they play an active role in gathering information.

Within the developmental literature, several studies have established that children learn successfully when they have the opportunity to choose what they want to do as well. For example, researchers have found that children as young as five years were able to use self-generated evidence to learn about an ambiguous causal system (McCormack, Bramley, Frosch, Patrick, \& Lagnado, 2016; Schulz, Gopnik, \& Glymour, 2007). In another study, Sim and Xu (in press) found that three-year-olds were capable of forming higherorder generalizations about a causal system after a short play period. In this study, children were presented with a causal learning task in which blocks would activate machines either based on a shape-rule (a block that matched the machine in shape activated the machine), or a color-rule (a block that matched the machine in color activated the machine). Children were randomly assigned to a didactic condition, such that an experimenter showed the children how to activate the machines, or a free play condition, such that the children were given the opportunity to play freely with these machines and blocks. The children were then tested using a first-order and a second-order generalization task. In the firstorder generalization task, children were asked to activate a familiar machine, and in the second-order generalization task, children were asked to activate a novel machine. The study found that both groups performed at levels well above chance, and there was no significant difference between the accuracy between the two conditions ( $\mathrm{Sim} \& \mathrm{Xu}$, in press).

Similarly, Smith and Dutton (1979) compared the performance of children in a play condition to both a training condition, where children were taught to use the materials by an experimenter, as well as a control condition, where children were neither taught by an experimenter nor played with the materials. Although children in the play and training condition performed significantly better in a problem-solving task compared to the control condition, there was no significant difference between the play and training conditions.

Although all of these studies suggest that children are able to learn through free play, there is as yet little evidence that young children's learning and generalizations under free play conditions would actually differ from those in didactic conditions. Learning in the free play condition has repeatedly been found to be comparable to a didactic condition, but has not been found to be different from a didactic condition.

Are there conditions under which young children may benefit more from the free play that they engage in independently, as compared to a training that is directed by an adult? In the current study, we examined this question by presenting children with a causal learning task in which the generalization to be acquired was more complicated than that examined in Sim and Xu (in press). To do so, machines in our study were activated when either two blocks that were the same shape as the machine, or two blocks that were the same color as the machine, were placed on the machine. Prior research have indicated that young children can learn causal rules of a similar form: Walker and Gopnik (2014) showed that after a short demonstration by an experimenter, 18- to 24-month-olds were able to learn a "same" or "different" causal rule, i.e. that a machine was activated only by placing two identical blocks on the machine, or that a machine was activated only by placing two dissimilar blocks on the machine. Given that the rule used in the current study is of a more complex form - the two blocks had to match each other and the machine on a specific dimension (i.e., shape or color) - we chose to test 3- and 4-year-olds in the current study.

Children were randomly assigned to a free play condition, where they were presented with six machines (three categories of machines, with two identical machines within each category) and twelve blocks to play with for approximately 10 minutes, or the didactic condition, where they observed an experimenter activate each machine once. To further examine any potential benefits of play, children who did not activate the machines at least once in the free play condition were placed in a third condition, the free-playfirst condition, where they observed the experimenter showing them how to activate the machines after they played by themselves for approximately 10 minutes. Similar to the study by Sim and Xu (in press), the children's ability to learn the correct rule was measured using a first- and second-order generalization test, where they were asked to activate both a familiar and a novel machine respectively.

## Method

## Participants

Sixty-one three- to five-year-old English-speaking children ( 29 boys and 32 girls) with the mean age of 48.4 months (range $=36.3$ months to 59.1 months) were tested. All were recruited from Berkeley, California and its surrounding communities. Children were tested either in a small testing room at our lab or in a small quiet room in a preschool. Each child was randomly assigned to the didactic $(\mathrm{N}=24)$ or free play condition ( $\mathrm{N}=22$ ). Children assigned to the free play condition but did not activate the machines at least once during the free play phase were placed in the free play first condition ( $\mathrm{N}=15$ ). The mean ages in the didactic, free play, and free play first condition were 48.1 months, 49.7 months, and 47.1 months respectively. An additional 9 children were tested but were excluded due to parent interference $(\mathrm{N}=5)$ and experimenter error $(\mathrm{N}=4)$.

## Materials

Five different types of machines were constructed for this experiment. Each type of machine made a distinct sound when activated. This activation was completed using a foot pedal connected to a remote that activated a doorbell that was placed inside the machines. There were two blue rectangle machines, two red triangle machines, two green circle machines, and one orange L-shaped machine. In addition, there was a colorful felt-covered plus-shaped machine (demonstration machine) that looked considerably different from the other four types of machines. Each machine was approximately $20 \mathrm{~cm} \times 12 \mathrm{~cm} \times 10 \mathrm{~cm}$.

A variety of small blocks (approximately $7 \mathrm{~cm} \times 5 \mathrm{~cm} \times$ 1 cm ) were used to activate the machines. The activator blocks were of different shapes and colors. Some matched the machines in shape but not in color, or in color but not in shape, while other distractor blocks matched in neither shape nor color. In total, twenty-two blocks were used.

## Procedure

Each child was tested individually. For children tested in our lab, parents sat next to the child during the procedure and were asked not to interact with their child. For children tested at preschools, an observer watched the procedure through a one-way mirror.

Both the didactic and free play conditions consisted of three phases: a demonstration phase, a training/free play phase, and a testing phase. For half the children, each machine was activated by placing two blocks on the machine that matched the machine in shape (shape rule). For the other half, each machine was activated by placing two blocks on the machine that matched the machine in color (color rule).

Children in the free play condition who did not activate the machines at least once during the free play phase formed a separate group: the free-play-first condition. This condition consisted four phases: a demonstration phase, a free play phase, a training phase, and a testing phase. In other words, children who did not activate any machines during free play
were then trained by the experimenter using the procedure of the didactic condition. For four of these children, machines were activated by the shape rule, and for eleven children machines were activated by the color rule.
Training and Free Play Phase

| Machines <br> (2 of each type) | Activators <br> (Shape Rule) | Activators <br> (Color Rule) |
| :---: | :---: | :---: |
| $\square$ | $\square$ | $\square$ |

First-Order Generalization Test


Figure 1: Schematic of materials and procedure.
Didactic Condition For the didactic condition, the children sat at a small table across from an experimenter. The demonstration phase began with the experimenter showing the child 12 blocks and pointing out that the blocks had different shapes and different colors. The blocks were then removed and the demonstration machine was placed on the table along with one block. The experimenter then showed the child how to make the machine go by placing the block on the machine and pressing down. After the machine made a sound, the experimenter noted that the blocks made the machine go and then allowed the child to try. Two new blocks were then placed in front of the child. The child activated the machine by pressing down with these two blocks and the experimenter stated that the blocks made the machine go. This was repeated next with three new blocks. The blocks and machine were then removed, ending the demonstration phase. The duration of this phase was around four minutes.

The experimenter then told the child that she had some new machines to show them. She also emphasized that these new machines were much pickier than the demonstration machine, so only some blocks would make them go.

The training phase that followed began with the experimenter presenting the first machine (e.g., blue
rectangle machine). The experimenter placed four activator blocks on the table next to the machine (e.g., the red, green, yellow, and purple rectangle blocks if the machine was activated by the shape rule; the blue triangle, blue circle, blue heart, and blue star blocks if the machine was activated by the color rule). The experimenter then stated, "Let me show you how to make this machine go," and placed two of the four activator blocks (e.g., purple rectangle block and yellow rectangle block, or blue circle block and blue star block) on the machine and pressed both blocks down, activating the machine. She then exclaimed, "The machine made a sound! It played music!" The experimenter then told the child that she had another machine that was identical to the one in front of them. This machine was placed on the table, and the other two activator blocks were now placed on the new machine, activating it. The machines and blocks were then removed and the process was repeated with the remaining two sets of machines. Once the child had seen all six machines activated one time each, the training phase was complete. The order of the presentation for the types of machines was counterbalanced. The duration of the training phase was approximately five minutes.

The testing phase consisted of both a first-order generalization test and a second-order generalization test. The order of the tests was counterbalanced. For the first-order test, the children were first presented with six separate blocks: two blocks that matched the machine in shape, two blocks that matched the machine in color, and two distractor blocks that did not match the machine in shape or color (see Figure 1). The experimenter presented the blue rectangle machine from the training phase, and said, "Remember this machine? Remember that I made this machine go just now? Can you show me how to make this machine go?" If the child placed the correct blocks on to the machine, the machine activated and the experimenter neutrally stated that the machine made a sound. If the child did not place the correct blocks, the machine did not activate and the experimenter neutrally stated that the machine did not make a sound. For the second-order generalization, the child was presented once more with six new blocks (see Figure 1). The child was then shown a novel machine (the orange L-shaped machine) and told that the machine "was a picky machine too". The experimenter then asked the child, "Can you show me how to make this machine go?" Once again, the experimenter neutrally stated that the machine made a sound if the child was correct, or did not make a sound if the child was incorrect.

Free Play Condition For the free play condition, the children sat on a blanket on the floor. The demonstration phase in the free play condition was identical to that in the didactic condition.

The free play phase began with the experimenter saying that she needed to check that all the machines worked (pilot testing suggested that this step was necessary in order to encourage children to keep playing even if they did not active any machines after a few attempts). The machines were taken behind a table and activated so that the child could hear them
activate by the sound they made, but could not see how they were activated. The experimenter then placed all six machines (two blue rectangle machines, two red triangle machines, and two green circle machines) in front of the child and noted that these machines were pickier and that only some blocks made them go. The twelve activator blocks shown in the demonstration phase were placed in front of the child as well. The experimenter then told the child, "I just remembered I have to do some work now, but while I work, you can play with these machines and these blocks." The children were given approximately ten minutes to play with the machines and blocks. If the child did not make any attempt to activate the machines for one minute, the experimenter prompted the child by saying, "Why don't you try to make the machines go?" After ten minutes, the blocks and machines were removed, ending the free play phase.

The testing phase was identical to that of the didactic condition.

Free-Play-First Condition For the free-play-first condition, the children sat on a blanket on the floor. The demonstration phase and free play phase were identical to that in the free play condition. Children were then moved to a table and the training phase was identical to the training phase in the didactic condition. The testing phase was identical to that of the didactic and free play conditions.

## Coding

For children exposed to the shape rule, selecting the two blocks that matched the machine in shape was scored as one point. On the other hand, for children exposed to the color rule, selecting the two blocks that matched the machine in color was scored as one point. Since each child completed two tests, the maximum score that a child could receive was two points.

## Results

We analyzed effects on children's responses with generalized linear mixed effects models in R, using an alpha level of 0.05 for all analyses. Children's responses were coded as a binary variable where correct responses were coded as 1 and incorrect responses were coded as 0 . In the model, subjects were specified as a random factor since this was a repeated measures task; each subject gave two responses. Although children in the didactic condition and the free play first condition all saw a total of six activations, the same cannot be said for children in the free play condition. In the free play condition, the total number of activations ( $M=17.6, S D=$ 12.6) each child saw as well as the amount of time that each child played for ( $M=7.91, S D=2.34$ ) varied. Although the didactic condition did not receive any negative evidence (i.e., observing unsuccessful activations), this was not the case for both the free play condition and the free-play-first condition. The amount of negative evidence generated by each child in the free play condition varied ( $\mathrm{M}=29.5, \mathrm{SD}=21.8$ ). Children in the free-play-first condition also generated a varied amount of negative evidence $(\mathrm{M}=28.1, \mathrm{SD}=16.1)$. Preliminary analysis showed no significant effect of age, sex, or
presentation order of the machines and testing phases. Additionally, there was no significant difference in the children's performance between first-order and second-order generalization tests in each of the three conditions.


Figure 2: Percent accuracy for the conditions. Dashed line represents a conservative calculation of chance. Error bars represent standard error.

There was, however, a statistically significant difference between the performance of children who were in the free play condition ( $M=0.4318, S D=0.4168$ ) and children who were in the didactic condition ( $M=0.10, S D=0.21$ ), as shown in Figure 2. More specifically, our analysis showed that the free play condition performed significantly better than the didactic condition ( $\beta=2.259, S E=0.787, p=0.004$ ). Analysis of the exponentiated coefficients revealed that being in the free play condition increased the children's odds of being correct by $856 \%$. The free-play-first condition ( $M=0.4333$, $S D=0.4169$ ) also performed significantly better than the didactic condition ( $\beta=2.267, S E=0.840, p=0.007$ ), and being in the free play first condition increased the children's odds of being correct by $865 \%$. However, there was no significant difference between the performance of children in the free-play-first condition and the free play condition. In addition, we analyzed whether the amount of negative evidence received was predictive of performance in the free play and free-play-first conditions. It was found that the amount of negative evidence did not have a significant effect on performance ( $\beta=-0.005, S E=0.018, p=0.787$ ), indicating that the extent to which children received negative evidence during play did not influence their performance during the generalization test trials.

The best fit model was also found by comparing various models that included potential predictors of performance such as condition, sex, age, rule, and amount of negative and positive feedback. Through model comparisons, it was found that the best fit model predicted accuracy from condition ( $\chi^{2}=$ 13.72, $d f=2, p=0.001$ ). This model outperformed all other models, including the null model.

A conservative value for chance was also calculated by considering all possible two block combinations that could be placed on the machine and then calculating the probability of
placing the correct blocks. This resulted in a value of 0.067 for chance performance. Children in the free play condition were significantly more likely to choose the correct blocks compared to chance, $t(43)=4.834, p=0.00002$, as were children in the free-play-first condition, $t(29)=3.985, p=$ 0.0004 . Children in the didactic condition, however, were not significantly more likely to choose the correct blocks over chance, $t(47)=0.842, p=0.404$.

## Discussion

In the current study, we demonstrate that 3- and 4-year-old children can successfully acquire fairly complex causal generalizations through free play. They independently generated evidence that allowed them to understand the causal system that they were presented with, and they formed higher-order generalizations at a level above chance. More strikingly, children's learning in the free play condition and the free-play-first condition was superior to that of children in the didactic condition. Just as in the study by Sim and Xu (in press), children were equally successful in learning firstorder and second-order generalizations during the course of free play, and there was no difference in their performance when it came to learning the shape or the color rule. It is interesting to note that there were more children in the free-play-first condition who had been exposed to the color rule rather than the shape rule, which may indicate the potential influence of a shape bias. However, we did not find an overall difference in performance between children who were exposed to the color rule vs. the shape rule.

We also sought to determine if there were scenarios in which children benefited more from learning through free play than through direction instruction by an experimenter. We found that children who engaged in free play performed significantly better at test than children in the didactic condition. Even children who were unable to activate the machines during play but who later observed an experimenter doing so performed significantly better than those assigned to the didactic condition, suggesting that the former group also benefited from engaging in free play. To the best of our knowledge, this study presents the first evidence that children can learn about a causal system more effectively through play than through training. So why was learning more effective under free play?

One possible reason for this difference between the two conditions is that children in the free play condition were able to engage actively with the materials, whereas children in the didactic condition played a passive role in learning about the machines, observing an experimenter activate them but never activating the machines themselves. Sobel and Sommerville (2010) found that four-year-old children learned a causal structure more accurately when they were given some time to engage with a causal system. Likewise, McCormack et al. (2016) showed that children who acted out interventions on a causal system following specific directions from an experimenter performed better than children who witnessed the same interventions but watched as the interventions were performed by an experimenter. Together, these studies
suggest that children may benefit more from intervening on a causal system, rather than observing an experimenter do so. This may explain why the free-play-first condition performed significantly better than the didactic condition, even though the children never successfully activated the machines while playing.

It is also possible that the children in the didactic condition struggled because they had different hypotheses about how the machines worked which were not contradicted by the evidence they witnessed (e.g., they may have thought that each of the two blocks would activate the machine by itself). Children in the free play condition and in the free-play-first condition, in contrast, had the opportunity to carefully test their own hypotheses, particularly through the generation of negative evidence. In other words, children in these conditions were able to see which combinations of blocks would make the machine go, as well as which combinations of blocks would not make the machine go. However, we did not find in our additional analyses that the amount of negative evidence that children generated was predictive of their performance at test, suggesting that there was something more to the evidence that the children saw during free play that assisted them in forming the correct generalizations. Further research is still necessary to understand the differences we found between the free play and didactic condition in our study. One worthwhile direction is to conduct an additional "yoked" didactic condition, where the experimenter presents children with evidence that was generated by children from the free play condition. This additional condition will clarify whether the differences found in the present study can be attributed solely to the difference in the quality of evidence between conditions.

We note that the current findings also appear to differ from those of other studies comparing the performance of children in free play and training/didactic conditions for other kinds of tasks. For example, Klahr and Nigam (2004) compared discovery learning, which they defined as learning that children engaged in by themselves without the assistance or feedback from a teacher, to direct instruction in third- and fourth-grade children for designing unconfounded experiments, which are experiments that clearly reveal the effect of a particular variable. The researchers found that children in the direct instruction condition performed significantly better than children in the discovery learning condition. However, in this particular study, children in the direct instruction condition were engaged in designing and manipulating variables during training as well, while children in the didactic condition in the current study engaged with the materials more passively, observing as an experimenter taught them how the machines worked. Another difference between the two studies is that in the study by Klahr and Nigam (2004), children in the discovery learning condition were given time to explore the ramp and marbles and design experiments, however they were not provided with any negative or positive feedback. In our study, on the other hand, children in the free play condition were provided with both negative and positive evidence, since the machine only made
a sound when activated correctly. Although the finding from Klahr and Nigam (2004) is sometimes used as evidence for the benefits of direct instruction for teaching children science, our study suggests that children have the potential to learn causal systems effectively through play when provided with useful feedback, even if the feedback does not come from an instructor.

Previous work has demonstrated that children attending child-centered preschools, where free play and child initiative are highly encouraged, were more motivated to learn, showed more pride in their accomplishments, and claimed to be less worried than children attending didactic, highly academic preschools (Stipek, Feiler, Daniels, \& Milburn, 1995). The current study extends these results by showing that children can learn effectively through free play, and under some conditions, learning within a free play context may be better than learning in a didactic context. However, it is important to note that free play is just one aspect of the child-centered instructional approach in preschools, and our results cannot speak directly to any potential learning differences between the two instructional approaches.

In summary, the present study provides evidence that young children can learn effectively through free play, and this learning might be better than the learning achieved through direct demonstration. Our results provide one source of empirical evidence on why play is important for children and why unstructured play should be incorporated into school curriculums. This study also suggests that there is merit to child-centered learning in preschools, as it appears that children are able to learn through play and that they are able to successfully engage in active learning.

## Acknowledgments

We thank the Berkeley Early Learning Lab (BELL) for their help in testing and recruitment, as well as the parents and children for their participation. Mahal and Sim are joint first authors.

## References

Chalnick, A., \& Billman, D. (1988). Unsupervised learning of correlational structure. Proceedings of the tenth annual conference of the cognitive science society (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.
Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., \& Zhu, X. (2009). Human active learning. In Advances in neural information processing systems, 241-248.
Corno, L., \& Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. Educational Psychologist, 18(2), 88-108.
Davey, C., \& Lundy, L. (2011). Towards Greater Recognition of the Right to Play: An Analysis of Article 31 of the UNCRC. Children \& Society, 25, 3-14.
Hofferth, S. L., \& Sandberg, J. F. (2001). How American Children Spend Their Time. Journal of Marriage and Family, 63, 295-308.
Klahr, D., \& Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct
instruction and discovery learning. Psychological Science, 15(10), 661-667.
Markant, D. B., \& Gureckis, T. M. (2014). Is it better to select or to receive? learning via active and passive hypothesis testing. Journal of Experimental Psychology: General, 143(1), 94-122.
McCormack, T., Bramley, N., Frosch, C., Patrick, F., \& Lagnado, D. (2016). Children's use of interventions to learn causal structure. Journal of Experimental Child Psychology, 141, 1-22.
Metcalfe, J., \& Kornell, N. (2005). A region of proximal learning model of study time allocation. Journal of Memory and Language, 52(4), 463-477.
Pepler, D. J., \& Ross, H. S. (1981). The Effects of Play on Convergent and Divergent Problem Solving. Child Development, 52(4), 1202-1210.
Schulz, L. E., Gopnik, A., \& Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. Developmental Science, 10(3), 322-332.
Sim, Z., \& Xu, F. (in press). Learning Higher-Order Generalizations through Free Play: Evidence from Twoand Three-Year-Old Children. Developmental Psychology.
Smith, P. K., \& Dutton, S. (1979). Play and training in direct and innovative problem solving. Child Development, 50(3), 830-836.
Sobel, D. M., \& Kushnir, T. (2006). The importance of decision making in causal learning from interventions. Memory \& Cognition, 34(2), 411-419.
Sobel, D. M., \& Sommerville, J. A. (2010). The importance of discovery in children's causal learning from interventions. Front. Psychol., 1,176.
Stipek, D., Feiler, R., Daniels, D., \& Milburn, S. (1995). Effects of different instructional approaches on young children's achievement and motivation. Child Development, 66(1), 209-223.
Walker, C. M., \& Gopnik, A. (2014). Toddlers infer higherorder relational principles in causal learning. Psychological science, 25(1), 161-169.
Weisberg, D.S., Hirsh-Pasek, K., \& Golinkoff, R. M. (2013). Guided play: Where curricular goals meet a playful pedagogy. Mind, Brain, and Education, 7(2), 104 - 112.
Weisberg, D.S., Hirsh-Pasek, K., Golinkoff, R. M., \& McCandliss, B. D. (2014). Mise en place: Setting the stage for thought and action. Trends in Cognitive Sciences, 18(6), 276 - 278.
Whitebread, D., Coltman, P., Jameson, H., \& Lander, R. (2009). Play, cognition and self-regulation: What exactly are children learning when they learn through play? Educational and Child Psychology, 26(2), 40-52.

# Conditionals, Individual Variation, and the Scorekeeping Task 

Niels Skovgaard-Olsen (niels.skovgaard.olsen@psychologie.uni-freiburg.de)
Department of Psychology, Engelbergerstraße 41, 79085 Freiburg Germany

David Kellen (dkellen@syr.edu)<br>Department of Psychology, Huntington Hall 516<br>Syracuse, NY 13244 USA

Ulrike Hahn (u.hahn@bbk.ac.ukk)<br>Department of Psychological Sciences, Malet Street<br>London, WC1E 7HX UK

Karl Christoph Klauer (christoph.klauer@psychologie.uni-freiburg.de)<br>Department of Psychology, Engelbergerstraße 41, 79085 Freiburg Germany


#### Abstract

In this manuscript we study individual variation in the interpretation of conditionals by establishing individual profiles of the participants based on their behavioral responses and reflective attitudes. To investigate the participants’ reflective attitudes we introduce a new experimental paradigm called the Scorekeeping Task, and a Bayesian mixture model tailored to analyze the data. The goal is thereby to identify the participants who follow the Suppositional Theory of conditionals and Inferentialism and to investigate their performance on the uncertain and-to-if inference task.


Keywords: conditionals; individual variation; and-to-if; norms; the Equation; inferentialism

## Introduction

According to a popular theory in the psychology of reasonning (the Suppositional Theory, or 'ST'), the probability of an indicative conditional (e.g. 'If I forget to pay the rent, then my landlord will complaint') is evaluated by a mental algorithm known as the Ramsey test (Evans \& Over, 2004; Oaksford \& Chater, 2007; Baratgin, Over, and Politzer, 2013).

The Ramsey Test: to evaluate P (if A , then C ) the participants add the antecedent to their background beliefs, make minimal adjustments to secure consistency, and evaluate the probability of the consequent on the basis of this temporarily augmented set of beliefs.
Quantitatively, this introduces the following prediction, which is known as "the Equation":

$$
\mathrm{PRED}_{1}: \mathrm{P}(\text { if } \mathrm{A} \text {, then } \mathrm{C})=\mathrm{P}(\mathrm{C} \mid \mathrm{A})
$$

Given that $\mathrm{P}(\mathrm{C} \mid \mathrm{A}) \geq \mathrm{P}(\mathrm{A}, \mathrm{C})$ follows from the axioms of probability theory (an inequality referred to as probabilistic coherence; PCh ), ST also predicts that:
PRED $_{2}$ : $\mathrm{P}($ if A, then C$) \geq \mathrm{P}(\mathrm{A}, \mathrm{C})$

Accordingly, the participants are predicted to conform to the following inequality in the so-called uncertain and-to-if inference (UAI), where they are presented with 'A and C' as a premise and 'if A, then C' as a conclusion and asked to assign probabilities to each:
Pred $_{2 \mathrm{~A}}: \mathrm{P}($ Conclusion $) \geq \mathrm{P}($ Premise $)$
Cruz, Baratgin, Oaksford, and Over (2015) found that the participants conformed to $\mathrm{PRED}_{2 \mathrm{~A}}$ at above chance levels. This has been taken as indirect evidence in favor of ST.
There is presently a considerable interest in and-to-if inferences, because recently a theory known as 'inferentialism' made its appearance into the psychology of reasoning, which posits that indicate conditionals express inferential relations. In the truth-conditional version of inferentialism, it rejects the validity of the and-to-if inference ' $\mathrm{A} \wedge \mathrm{C} \vDash$ if A , then C' (Douven, 2015). Truth-conditional inferentialism rejects the validity of this argument scheme, because the indicative conditional is viewed as expressing a reason relation and the mere truth of A and C does not ensure that they are inferentially connected. Rejecting the validity of the and-to-inference is a distinguishing feature of this approach that separates it from other popular semantics of conditionals like Stalnaker's possible worlds semantics or the de Finetti truth table endorsed by proponents of ST.
In Skovgaard-Olsen, Singmann, and Klauer (2016a) a weaker probabilistic implementation of inferentialism was given in the form of the Default and Penalty Hypothesis (DP), which employs the following explication of the reason relation:
PO: A is positively relevant for C (and a reason for C ) iff $\mathrm{P}(\mathrm{C} \mid \mathrm{A})>\mathrm{P}(\mathrm{C} \mid \sim \mathrm{A})$
NE: A is negatively relevant for C (and a reason against C$)$ iff $\mathrm{P}(\mathrm{C} \mid \mathrm{A})<\mathrm{P}(\mathrm{C} \mid \sim \mathrm{A})$
IR: A is irrelevant for $C$ iff $\mathrm{P}(\mathrm{C} \mid \mathrm{A})=\mathrm{P}(\mathrm{C} \mid \sim \mathrm{A})$
DP posits that the participants have the goal of evaluating whether a sufficient reason relation obtains when evaluating

P(if A, then C). According to Spohn's (2012: ch. 6) explication of the reason relation given above, this requires at least two things: (a) assessing whether A is positively relevant for C , and (b) assessing the sufficiency of A as a reason for C by means of $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$. DP moreover postulates that the participants follow the heuristic, when processing natural language conditionals, of making the default assumption that (a) is satisfied, which reduces their task of assessing P (if A , then C ) to assessing $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$. However, once the participants are negatively surprised by a violation of this default assumption, such as when they are presented with stimulus materials implementing the NE or IR category, they apply a penalty to $\mathrm{P}($ if A , then C ) to express the conditional's failure to express that A is a reason for C. An example would be the conditional 'If Oxford is in England, then Napoleon is dead' which sounds defective to the extent that the antecedent is obviously irrelevant for the consequent.
In support of DP, it was found in Skovgaard-Olsen et al. (2016a) that $\mathrm{PRED}_{1}$ only holds when $A$ is positively relevant for C in virtue of raising its probability. When A is negatively relevant by lowering C's probability, and when A is irrelevant for C by leaving its probability unchanged, violations of PRED $_{1}$ occur. Consistent with these findings, it was found in Skovgaard-Olsen et al. (2016b) that the abovechance level of conformity to $\mathrm{PRED}_{2 A}$ reported in Cruz et al. (2015) only holds for PO. In NE and IR the participants are performing below chance levels. Further-more, this is a pattern that is not reflected in their conformity to the theorem $\mathrm{P}(\mathrm{C} \mid \mathrm{A}) \geq \mathrm{P}(\mathrm{A}, \mathrm{C})$ across relevance levels, in spite of the fact the participants are supposed to conform to P(if A, then C$)=\mathrm{P}(\mathrm{C} \mid \mathrm{A})$, according to ST .
It is presently unclear whether this finding of lack of conformity to $\mathrm{PRED}_{2 \mathrm{~A}}$ in the NE and IR conditions indicates that the participants are making a reasoning error (by following ST) or whether they are not making a reasoning error but simply basing their performance on a different interpretation of conditionals (by following DP). The goal of the present study is to address this question.
In the present experiment, we seek to establish individual profiles of the participants based on their behavioral responses and reflective attitudes. In order to study their reflective attitudes we implemented a novel experimental paradigm - the Scorekeeping Task - suggested in Skovgaard-Olsen (2015), as well as a Bayesian mixture model tailored to classify the data coming from it (both are discussed in detail below). Based on this novel task and the associated data-analytic method, we were able to investigate two key questions: First, whether participants classified as ST accord with ST's PRED $_{2 \text { A }}$ prediction for the UAI across a relevance manipulation. Second, whether participants classified as DP accord with DP's prediction that PRED $_{2 \mathrm{~A}}$ only holds in the PO condition. In the IR condition, DP participants are expected to apply a penalty to conditionals in the conclusion of the UAI, such that P(if A, then C) < $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$ can occur, effectively dismissing $\mathrm{PRED}_{2 \mathrm{~A}}$.

## Experiment

## Method

## Participants

A total of 354 people from the USA, UK, Canada, and Australia completed the experiment, which was launched over the Internet (via Mechanical Turk) to obtain a large and demographically diverse sample. Participants were paid a small amount of money for their participation.
The following exclusion criteria were used: not having English as native language (6 participants), completing the experiment in less than 300 seconds (2 participants), failing to answer two simple SAT comprehension questions correctly in a warm-up phase (89 participants), and answerring 'not serious at all' to the question how serious they would take their participation at the beginning of the study (zero participants). Since some of these exclusion criteria were overlapping, the final sample consisted of 261 participants. Mean age was 36.53 years, ranging from 20 to $75,66 \%$ were female, $66 \%$ indicated that the highest level of education that they had completed was an undergraduate degree or higher.

## Design

The experiment implemented a within-subject design with two factors varied within participants: relevance (with two levels: PO, IR) and priors (with four levels: HH, HL, LH, LL, meaning, for example, that $\mathrm{P}(\mathrm{A})=$ low and $\mathrm{P}(\mathrm{C})=$ high for LH).

## Materials and Procedure

We used a slightly modified version of 12 of the scenarios presented in Skovgaard-Olsen et al. (2016b). For each scenario we had 8 conditions according to our design (i.e., 4 conditions for PO [i.e., HH, HL, LH, LL], 4 conditions for IR). Each participant worked on one randomly selected (without replacement) scenario for each of the 8 withinsubjects conditions such that each participant saw a different scenario for each condition. Following the recommendations of Reips (2002), to reduce dropout rates, we presented two SAT comprehension questions as an initial high hurdle in a warm-up phase (in addition to using them for excluding participants). The experiment was split into four phases and on average took ca. 23 minutes to complete. Here we focus on conveying the underlying conceptual ideas.

## Phase 1, Behavioral Responses

The first phase contained eight blocks, one for each withinsubjects condition. The order of the blocks was randomized anew for each participant and there were no breaks. Within each block, the participants were presented with four pages. On the first page, the participants were shown a scenario text like the following:
Scott was just out playing with his friends in the snow. He has now gone inside but is still freezing and takes a bath. As both he and his clothes are very dirty, he is likely to
make a mess in the process, which he knows his mother dislikes.

The idea was to use brief scenario texts concerning basic causal, functional, or behavioral information that uniformly activates stereotypical assumptions about the relevance and prior probabilities of the antecedent and the consequent of 8 conditionals that implement our experimental conditions for each scenario. So to introduce the 8 within-subjects conditions for the scenario above we, inter alia, exploited the fact that the participants would assume that Scott's turning on the warm water would raise the probability of Scott being warm soon ( PO ) and that Scott's friends being roughly the same age as Scott would be irrelevant for whether Scott will turn on the warm water (IR).
This scenario text was repeated on each of the following three pages, which measured $\mathrm{P}(\mathrm{A}$ and C$), \mathrm{P}(\mathrm{C} \mid \mathrm{A})$, and $\mathrm{P}($ if A , then C ) in random order. Throughout the experiment, the participants gave their probability assignments using sliders with values between 0 and $100 \%$. To measure $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$, the participants might thus be presented with the following question in an IR condition:
Suppose Scott's friends are roughly the same age as Scott. Under this assumption, how probable is it that the following sentence is true on a scale from 0 to 100\%:
Scott will turn on the warm water.

## Phase 2, the Scorekeeping Task

In this phase the participants were first presented with a new IRHH item to be rated in the same way as the items in phase one. Then the participants were presented with the following instruction:
When given the task you just completed, John and Robert responded very differently to some of the scenarios as outlined below.
And it was explained that John and Robert responded in the following way to the "if-then sentence" and the "supposesentence" (where the "suppose-sentence" had been identified for the participants as the type of question quoted above for measuring $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$ ):
John assigned 99\% to the suppose-sentence and 1\% to the if_then sentence.
Robert assigned $\mathbf{9 0 \%}$ to the suppose-sentence and $\mathbf{9 0 \%}$ to the if_then sentence.
Note that although John and Robert are fictive participants, these values were based on actual data provided by other participants in response to the IRHH item in previous experiments. In order to reduce the processing demands, these values were repeated on each of the following four pages along with the IRHH item, which John and Robert allegedly had responded to. The conditional took the following form, and it was evaluated in the context of a dating scenario describing Stephen's preparations for a date with Sara: 'If Stephen's neighbour prefers to put milk on his cornflakes, then Stephen will wear some of his best clothes on the date'.

As part of the scorekeeping task, the participants were instructed to apply a sanction to John or Robert's response based on its adequacy. Given their large divergence, the participants were instructed that at most one of John or Robert's responses could be approved as adequate.
Since the experiment was run on Mechanical Turk we exploited the fact that an ecologically valid sanction for the participants would be not to have a task (a "HIT") approved. Since the approval of HITs on Mechanical Turk determines whether the participants are paid for a completed task (and moreover counts towards their reputation on Mechanical Turk, which determines whether they can participate in future HITs) it is our experience that the participants care a lot about the approval of their HITs. We therefore expected that applying the sanction of not approving either John or Robert's HIT based on its adequacy would be a contextually salient sanction, which the participants would be highly motivated to reason with.
Next the participants were asked to state the reasons that they could think of which could be given for or against John and Robert's responses in an open entry question, which was included in the experiment for exploratory purposes.
On the two pages that followed, the participants were presented with John's criticism of Robert and Robert's criticism of John in random order. Robert made the following complaint about John's response:
Robert's no difference justification: "There is no difference between the two questions. So why do you give a lower probability to:
'IF Stephen's neighbour prefers to put milk on his cornflakes, THEN Stephen will wear some of his best clothes on the date'
than you gave to: 'Stephen will wear some of his best clothes on the date' under the assumption that 'Stephen's neighbour prefers to put milk on his cornflakes'?
This makes no sense!"
John in turn made the following complaint about Robert's response:
John's irrelevance justification: "Whether 'Stephen's neighbour prefers to put milk on his cornflakes' or not is irrelevant for whether 'Stephen will wear some of his best clothes on the date'.
So why do you give such a high probability to: 'IF Stephen's neighbour prefers to put milk on his cornflakes, THEN Stephen will wear some of his best clothes on the date'? This makes no sense!"
In each case, the participants were asked to indicate (yes/no) whether they agreed with the following statements:
John's irrelevance justification [/Robert's no difference justification] shows that Robert's [/John's] response is wrong.
Robert [/John] needs to come up with a very good response to John's [/Robert's] criticism, if his HIT is to be approved.
Finally, after having seen the justifications from both sides, the participants were asked which justification they found
most convincing by choosing between the following options, presented in random order:

## The two justifications are equally convincing John's irrelevance justification Robert's no difference justification

The participants then had to indicate who's HIT deserved to be approved based on their justifications by selecting one of the options below, presented in random order:
None of their HITs should be approved
Robert's HIT should be approved
John's HIT should be approved

## Phase 3, the Uncertain And-to-If Inference

This phase tested the participants' performance on the UAI under relevance manipulations. Phase 3 was used to measure whether the participants displayed a consistent behavior on the UAI with the interpretation of the conditional that they had been classified according based on their responses in phase 1 and phase 2.
Phase 3 contained 8 blocks implementing the same withinsubjects conditions as phase 1. For each participant, the same permutations of scenarios and within-subject conditions that had been randomly generated in phase 1 was displayed again in random order. First the participants were instructed that they would be presented with a scenario text as earlier and a short argument based on the scenario text. They were told that the premise and the conclusion of this argument could be uncertain and that it was their task to evaluate the probabilities of the premise and conclusion. Each block contained one page. On the top of the page the scenario text was placed as a reminder. Below the participants were instructed to read an argument containing the conjunction as a premise and the conditional as a conclusion, employing sentences that they assigned probabilities to in phase 1. Furthermore, the actual value of the probability that they had assigned to the premise in phase 1 was displayed to the participants in a salient blue color. We here illustrate it using the example from above from phase 1 of a POHH item:
Premise: Scott's turns on the warm water AND Scott will be warm soon.
Conclusion: IF Scott's turns on the warm water, THEN Scott will be warm soon.
You have estimated the probability of the premise as: $\mathbf{9 0 \%}$. Please rate the probability of the statement in the conclusion on a scale from 0 to $100 \%$.
In Phase 4, we tested the participants’ interpretation of the probabilities (Hertwig \& Gigerenzer, 1999). These results are beyond the scope of the present manuscript and therefore not reported here.

## Bayesian Mixture Modeling

In order to investigate the participants' interpretation of the conditional, the probability judgments they produced in Phase 1 were classified as coming from one of two latent classes using an indicator variable $w$. This classification was
achieved by means of a Bayesian Mixture model (for a similar approach, see Lee, 2016). In the PO condition, where both ST and DP make the same predictions (see the left panel of Figure 1), the mixture model assumed that responses from an individual $i$ were generated by ST/DP ( $w_{i}^{P O}=1$ ), or by an unclassifiable response-generation mechanism ( $w_{i}^{P O}=0$ ), for an item-pair $j$ :

$$
P(\text { if } A \text {, then } C)_{i, j}=\left\{\begin{array}{c}
\beta_{i, j}+\varepsilon_{i, j}, \quad w_{i}^{P O}=0 \\
P(C \mid A)_{i, j}+\varepsilon_{i, j}, \quad w_{i}^{P O}=1
\end{array}\right.
$$

where $0 \leq \beta_{i, j} \leq 100$.
When an individual follows ST/DP, the generated P (if A , then C ) are expected to follow $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$ along with some truncated Gaussian noise term $\varepsilon_{i, j}$ with mean 0 and variance $\sigma^{2}$ (see the left panel of Figure 1). This noise captures the variability that is commonly observed in probability judgments across the [0\%, 100\%] interval (see Costello \& Watts, 2016). When an individual follows an unclassified pattern, their responses were captured by a saturated model, which established a $\beta$ parameter per data point (predicting the latter perfectly). 1

In the IR condition, the model only considered participants that were classified as ST/DP in the PO condition (i.e., the PO condition served as a filter for the IR condition). Here, both ST $\left(w_{i}^{I R}=0\right)$ and $\mathrm{DP}\left(w_{i}^{I R}=1\right)$ make distinct predictions:

$$
P(\text { if } A, \text { then } C)_{i, j}=\left\{\begin{array}{cc}
P(C \mid A)_{i, j}+\varepsilon_{i, j}, & w_{i}^{I R}=0 \\
\theta_{i} P(C \mid A)_{i, j}+\varepsilon_{i, j}, & w_{i}^{I R}=1
\end{array}\right.
$$

with $0 \leq \theta_{i} \leq 1$.
When individuals follow ST, the generated P (if A , then C) are again expected to follow $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$. In contrast, when individuals follow DP, P (if A , then C ) follows a penalized version of $\mathrm{P}(\mathrm{C} \mid \mathrm{A})$ (with the penalty being determined by $\theta$ ).

Note that when $\theta=1$, the ST and DP models coincide, although the implied predictions are not really in accordance with the gist of DP. However, this point turns out not to be of practical import, because since ST is more parsimonious it will be preferred when $\theta=1$ (see Lee, 2016).


Figure 1. Predictions from both theoretical accounts (including some moderate degree of truncated noise).

[^203]

Figure 2. Individual associated to the different Phase 1 classifications, and their respective posterior individual-level classifications (note that in the IR condition, only participants classified as ST/DP in the PO condition were considered).

The key parameters of interest in this analysis are the posterior probabilities of $w_{\mathrm{i}}=1$ obtained in the PO and IR conditions. In the PO condition, when the mean of this posterior probability was estimated to be below or equal to .50, the individual was classified as following the saturated model. When the mean is estimated to be larger than .50 , the individual was classified as following ST/DP. In the IR condition, these same ranges of values led to the ST and DP classifications, respectively.

The individual classifications jointly obtained for PO and IR were used to characterize the conformity of individuals’ responses to theoretically-meaningful inequalities, namely UAI and PCh. For participant $i$, the probability that her response to a given item-pair $j$ conformed to a given inequality is given by $\Phi\left(\Delta_{i}+K_{i, j}\right)$, with $\Phi()$ being the probability function of the standard Normal distribution. Parameter $K_{\mathrm{i}, \mathrm{j}}$ is a correction term for participant $i$ and itempair $j$ such that $\Phi\left(K_{i, j}\right)$ corresponds to the probability that the responses to a given item-pair were inequality-conforming by chance alone (Singmann, Klauer, \& Over, 2014). Parameter $\Delta_{i}$ corresponds to that individual's displacement from chance (i.e., when $\Delta_{i}$ is positive, that individual produces inequality-conforming responses at an abovechance rate). Using a hierarchical framework, these individual parameters were assumed to come from a Normal group-level distribution, with mean $\mu_{\Delta}$ and standard deviation $\sigma_{\Delta}$. If individuals in general conform to the UAI or PCh, then their respective $\mu_{\Delta}$ should be consistently above 0 (i.e., the probability of $\mu_{\Delta}$ being below 0 should be very small). These parameters were estimated separately for individuals classified as ST and DP in the IR condition.

A very similar hierarchical approach was used to model the relative probability of an individual judging the nodifference justification (in line with ST) as most convincing after having seen both sides, as well as the relative probability attributing the HIT to such justification.

## Results

The posterior-parameter distributions of mixture model were estimated via Gibbs sampling using the generalpurpose software JAGS (Plummer, 2003). Chain convergence was confirmed via the R-hat statistic and visual inspection. The individual-level classifications shown in Figure 2 show that the probabilities generated by the majority (225 out of 261) of individuals in the PO condition were in line with ST/DP. In contrast, only a very small group of individuals were in line with ST in the IR condition (39 out of 225); most followed the predictions of DP. The individual data shown in Figure 1 shows that the data classified as ST/DP in the PO condition as well as ST and DP in the IR condition were in line with the model predictions. To address the worry that participants belonging to ST were misclassified as DP, we visually inspected the responses of every participant individually.
The classifications lead to clear differences in both UAI and PCh, as well as in the probability of judging the nodifference justification as most convincing. As shown in Table 1, for UAI the posterior $\mu_{\Delta}$ estimates in the IR condition for individuals classified as ST are systematically above 0 , but systematically below 0 for individuals classified as DP. In the case of PCh, the posterior $\mu_{\Delta}$ estimates were systematically above 0 , as expected. The latter result was less clear for ST, but this is expected given the small number of participants classified as being in line with ST.
Finally, the relative probabilities of judging the nodifference justification (consistent with ST) as most convincing and attributing the HIT were drastically different for individuals classified as following ST and DP. These posterior probabilities were considerably larger for ST (see Table 1). Note that these were conditional probabilities of finding the ST justification most convincing, and accepting the ST HIT, given the participants expressed preferences for either ST or DP in phase 2.

|  | ST Followers $(\mathrm{N}=39)$ | DP Followers $(\mathrm{N}=186)$ |
| :--- | :---: | :---: |
|  | $\mu_{\Delta}$ | $\mu_{\Delta}$ |
| UAI | $0.61[0.16,1.11](72 \%)$ | $-0.46[-0.65-0.28](47 \%)$ |
| PCh | $0.21[-0.07,0.51](68 \%)$ | $0.14[0.02,0.27](66 \%)$ |
| P(ST mc) | $.94[.78,1]$ | $.15[.09, .22]$ |
| P(ST HIT) | $.92[.77,1]$ | $.21[.15, .28]$ |

Table 1. Median group-level posterior parameter estimates (and their respective $95 \%$ credibility intervals) obtained in the IR condition. Percentages of responses conforming to UAI and PCh are given in parentheses. The estimates associated to $\mu_{\Delta}$ in the PO condition (where participants were classified as ST/DP) were 1.66 [1.14, 2.24] and 1.19 [ $0.82,1.61]$ for UAI and PCh, respectively. ' $\mathrm{P}(\mathrm{ST} \mathrm{mc})$ ' $=\mathrm{P}($ ST most convincing | ST or DP most convincing $)$. $' P(S T$ HIT $) '=P(S T$ receive HIT | ST or DP receive HIT).

## Discussion

In this paper we have presented a novel experimental design to study the reflective attitudes of the participants and an accompanying Bayesian mixture model to study individual variation. We have seen that it is possible to classify the participants according to whether they follow the Suppositional Theory of Conditionals or the Default and Penalty Hypothesis. We then used these classifications to study the participants' performance on the uncertain and-toif inference task to examine whether the participants consistently followed the assigned interpretation of the conditional in an inference task.
This experimental design gives us a very rich data set that we have not exhausted in this brief note. Nevertheless, the data we did analyze show a very clear pattern. In the PO condition of phase 1, $86 \%$ of the participants followed the Equation ( $\mathrm{PRED}_{1}$ ), whereas only 39 of these participants followed the Equation in the IR condition. The remaining 186 participants showed a clear tendency in the IR condition to assign lower probabilities than if they had treated the P (if A, then C) as a conditional probability. For the 39 ST participants from phase 1 there was a .94 probability that they find the ST character to be most convincing one, conditional on the fact that they had a preference. Of the 186 DP participants in phase 1, this conditional probability was .85 , this time in favor of the DP character.
Finally, the participants' performance on the uncertain and-to-if inference task in phase 3 indicated that the participants acted consistently with their assigned interpretation of the conditional. As a theorem of probability theory, the PCh inequality $(\mathrm{P}(\mathrm{C} \mid \mathrm{A}) \geq \mathrm{P}(\mathrm{A}, \mathrm{C})$ ) remains valid for both groups, so they should conform to it at above chance levels irrespectively of the relevance condition. In contrast, whether the participants should conform to the UAI inequality ( $\mathrm{P}($ Conclusion $) \geq \mathrm{P}($ Premise $)$ ) in the IR condition, depends on whether they interpret the conditional in the conclusion as a conditional probability.
In the PO condition both groups were above chance levels for conformity to both the UAI and PCh inequalities. For the ST participants, a tendency was found to continue to conform to the UAI and PCh inequalities in the IR condition at above chance levels. (However, the estimates were
connected with uncertainty given the modest size of the ST group.) In contrast, for the DP participants an interaction was revealed between relevance and type of inequality in that these participants continued to display conformity to PCh at above chance levels in the IR condition while ceasing to conform to the UAI inequality at above chance levels. The results thus indicate that it was possible to separate two individual profiles in the participants' interpretation of the conditional. For each profile, the participants were shown to behave consistently with their interpretation of the conditional in the uncertain and-to-if inference.
In Skovgaard-Olsen et al. (2016b), it was found that the above-chance level conformity to UAI, which Cruz et al. (2015) did not generalize to the IR condition. However, since these results were analyzed at the group level, it was hard to tell whether they indicated that the participants were incoherent or whether they followed DP instead. With the present results we have a first indicator that two groups can be identified at the individual level that consistently follow their assigned interpretation of the conditional in the uncertain and-to-if inference.

## References

Baratgin, J., Over, D. E., \& Politzer, G. (2013). Uncertainty and the de Finetti tables. Thinking \& Reasoning, 19, 308-28.
Costello, F., \& Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). Cognitive Psychology, 89, 106-133.
Cruz, N., Baratgin, J., Oaksford, M. and Over, D. E. (2015). Bayesian reasoning with 'if's and 'and's and 'or's. Frontiers in Psychology, 6, 192.
Douven, I. (2015). The Epistemology of Indicative Conditionals. Formal and Empirical Approaches. Cambridge, UK: Cambridge University Press.
Evans, J. St. B. T. and Over, D. (2004). If. Oxford: Oxford University Press.
Hertwig, R., \& Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. Journal of Behavioral Decision Making, 12, 275-305.
Lee, M. D. (2016). Bayesian outcome-based strategy classification. Behavior Research Methods, 48, 29-41.
Oaksford, M. and Chater, N. (2007). Bayesian Rationality: The Probabilistic Approach to Human Reasoning. Oxford: Oxford University Press.
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, \& A. Zeileis (Eds.), Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003).
Singmann, H., Klauer, K. C., \& Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. Frontiers in psychology, 5, 316.
Skovgaard-Olsen, N. (2015). The problem of logical omniscience, the preface paradox, and doxastic commitments. Synthese.
Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016a). The Relevance Effect and Conditionals. Cognition, 150, 26-36.
Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016b). Relevance and Reason Relations. Cognitive Science.
Spohn, W. (2012). The Laws of Beliefs. Oxford: Oxford University press.

# A case for systematic sound symbolism in pragmatics: The role of the first phoneme in question prediction in context 

Anita Slonimska (anita.slonimska@istc.enr.it)<br>CLS, Radboud University, Nijmegen, The Netherlands<br>Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy

Seán G. Roberts (sean.roberts@bristol.ac.uk)
Department of Archaeology and Anthropology, University of Bristol, UK
Language and Cognition Department, Max Planck Institute for Psycholinguistics, The Netherlands


#### Abstract

Turn-taking in conversation is a cognitively demanding process that proceeds rapidly due to interlocutors utilizing a range of cues to aid prediction. In the present study we set out to test recent claims that content question words (also called wh-words) sound similar within languages as an adaptation to help listeners predict that a question is about to be asked. We test whether upcoming questions can be predicted based on the first phoneme of a turn and the prior context. We analyze the Switchboard corpus of English by means of a decision tree to test whether $/ \mathrm{w} /$ and $/ \mathrm{h} /$ are good statistical cues of upcoming questions in conversation. Based on the results, we perform a controlled experiment to test whether people really use these cues to recognize questions. In both studies we show that both the initial phoneme and the sequential context help predict questions. This contributes converging evidence that elements of languages adapt to pragmatic pressures applied during conversation.


Keywords: questions; wh-words; question words; turntaking; speech-act recognition; question prediction

## Introduction

People spend an average of 2-3 hours every day in conversation, producing around 1200 turns (Levinson, 2016). The structure of conversation, far from being chaotic, places specific constraints on speakers (Sacks, Schegloff \& Jefferson, 1974). Recently, it has been recognized that these constraints have implications for processing and therefore for the way languages evolve (see Levinson, 2016). In this paper we explore a phenomenon at the interface of conversation, processing and cultural evolution.
Conversation progresses through exchanging bursts of information - mostly through use of language - that are orchestrated in consecutive turns produced by the speakers (Sacks et al., 1974). The surprising aspect of turn-taking is that it is orchestrated in a remarkably tight manner. Speakers strive to minimize gaps and overlaps between turns (Sacks et al., 1974), with the average gap length being only 200 ms cross-culturally (Stivers et al., 2009; Kendrick \& Torreira, 2015; Levinson \& Torreira, 2015). Thus, while languages themselves differ, the pressure for rapid turntaking is the same.

The surprising fact that turns are produced in such a tight window of time becomes even more puzzling if we take into account that it takes a minimum of 600 ms to plan and begin uttering a single word (Schriefers, Meyer, Levelt, 1990;

Levelt, 1993). In this context, one has to ask a question how is it possible that the gap between turns is shorter than the planning of the response? The obvious answer is prediction (Sacks, Schegloff \& Jefferson, 1974; Levinson, 2013). Listeners project what the current speaker will say and when their turn will end (Holler and Kendrick, 2015; Bögels \& Torreira, 2015). Thus, the next speaker can start preparing their turn in advance so that it can be delivered on time.

Predicting the specific type of a speech act is extremely important as different speech acts have different social and cognitive pressures on speakers. For example, when we are greeted, the greeter expects a greeting in response. Or when we are asked a question, we are socially obliged to give an answer, and hesitations can lead to inferences about the intent of the responder (Kendrick \& Torreira, 2015). Thus, social constraints put pressure on cognition to respond rapidly in interactive conversation. We suggest that languages should evolve to provide listeners with early cues that facilitate this process. Perhaps the context in which this would be most evident is in recognizing questions, to which we now turn.

Answering questions is a complex process involving understanding the question, retrieving or calculating the relevant answer and planning the response. Previous research suggests that the planning of the response starts as soon as an answer can be retrieved (Bögels, Magyari \& Levinson, 2015; Bögels, Casillas, \& Levinson, 2016; Barthel, Meyer \& Levinson, 2017). However, even before planning their answers, speakers first have to recognize that they are being asked a question.

Gisladottir, Chwilla, \& Levinson (2015) show that people can recognize the type of a speech act at an early stage if the preceding turns sufficiently constrain the context. For example, if I have just produced an initiating turn (like a greeting or asking a question), my interlocutor is most likely to produce a responding action (like an answer), rather than ask a question of their own. Therefore, one early cue as to whether a question will appear is the prior context.

Beyond that, there are also early cues in the question itself, before the turn can be identified as a question syntactically or semantically. Levinson (2013) suggests that question recognition is possible due to front-loading of the cues at the beginning of a turn. For example, questions can
be recognized by early cues in intonation (Levinson, 2013), pitch (Sicoli et al. 2014) and eye-gaze (Rossano, Brown \& Levinson, 2009; Rossano, 2012). Moreover, shifting question words to the initial position of the utterance (e.g., wh-movement in English) appears to be one of the most evident examples of front-loading (Levinson, 2013). Even when wh-movement is not permitted in the formal grammar of many languages, it is often evident in colloquial interactions (e.g. in Japanese, Levinson, 2013). Surprisingly, though, there is no quantitative research investigating whether this feature actually helps in question recognition.

Slonimska \& Roberts (accepted) were the first to quantitatively assess whether question words, also called wh-words, are plausible candidates as a cue to content question recognition. They suggest that a systematic phonetic similarity between question words within a language could provide a cue for that. In other words, if question words tend to sound similar, it would be easier for the addressee to predict that a question is about to be asked, and they can prepare themselves accordingly. For example, in English many question words begin with /w/ (what, why, where, when), and in Latvian many begin with /k/ (kas, kad, kur, kurš, kas, kāpēc).

Even though there is some qualitative research arguing that there is no systematicity in question words (Cysouw, 2004), Slonimska \& Roberts (accepted) show that there is a statistical tendency for question words to sound similar within languages. When they analyzed 266 languages the authors found that there is a higher similarity between the first phoneme of question words (within languages) than would be expected by chance, than other sets of words and also when controlling for historical factors. Accordingly, Slonimska \& Roberts argue that this phenomenon constitutes a product of cultural evolution that is selected for due to its benefit in interaction - i.e., rapid question recognition. Their study, however, is based purely on observational data of word forms. This leaves several issues to be addressed before their claim can be supported. First, are phonological regularities in question words actually statistically good predictors of questions in conversation? Secondly, do people actually use these cues to recognize questions? Finally, what is the relationship between the use of these cues and the prior conversational context?

We address these issues by means of two studies. First, we explore a large corpus of natural conversations and subsequently use the insights from the corpus study to design an experiment in which we test the hypotheses in a controlled setting by using stimuli from the same corpus.

As such, the present project not only informs the theoretical field in regard to question recognition, but it also makes a case for a new approach to research - namely, by creating a synergy between ecologically valid corpus analysis and experimentally controlled quantitative insights into the phenomenon.

## Corpus study

## Method

To assess whether we can gain support for our hypotheses, we first carried out an exploratory corpus analysis of naturalistic data - i.e., spoken conversations. We addressed this by means of the method of binary decision trees, also known as recursive partitioning (Strobl, Malley, and Tutz, 2009). A binary decision tree represents the optimal series of yes-no questions that a rational agent would ask about predictor variables in order to estimate an outcome variable (see Roberts et al., 2015).

In the current study we are interested in whether the first phoneme of the turn (first predictor) and context of the previous turn (second predictor) would help in recognizing an incoming turn as a content question (outcome variable). Namely, we predicted that the data would be clustered in such way that specific first phoneme (/w/, /h/ versus other phonemes in English) of the current turn and specific type of previous turn (non-initiating turn versus initiating turn) would help identify whether the current turn was a question. Unlike regression frameworks, the predictor variables that a binary decision tree uses are not set by the researchers, but chosen by an algorithm in order to maximize performance and parsimony. It could pick any combination of phonemes as identifying factors if suggested by the data. Therefore, our prediction of the form of the tree is a strong one.

Materials and design. We used the Switchboard corpus (Calhoun et al., 2010) that consists of telephone conversations in American English. This corpus is transcribed and annotated in detail, including a division of utterances into sequential turns by Roberts et al. (2015). The data was prepared for the analysis in R and later analyzed by means of the package "party" (Hothorn, Hornik \& Zeileis, 2006).

Each observation consisted of a transition between two turns between speaker A and speaker B. We specified the outcome variable - question - according to whether B's turn (i.e., current turn) was a question (content/open question) or not, according to the dialogue act annotation. We used the last speech act of A's turn (i.e., previous turn) for the second predictor variable specifying whether this turn was initiating or non-initiating (see Roberts et al., 2015). For example, B's turn was "What kind do you like to watch" - this was a turn that was a question and that started with $/ \mathrm{w} /$. The turn that preceded this question (i.e., A's turn) was "and uh you know there so there only a few that i that i like to watch routinely" - this was a statement (i.e., non-initiating turn).

We excluded the following fillers from the B's turn: ahm, er, ah, hmm, oh, uh, aa, um, ow. Then, the first phoneme from B's turns was extracted to create the predictor variable phoneme. This variable consisted of 34 unique phonemes (coded according to the transcription convention of Switchboard). Finally, we excluded all turns for which B's turn was a backchannel, considering that backchannel serves a monitoring rather than an informing function.

The final data included 9185 turns in total out of which 226 turns were content or open questions. Out of all turns, 1456 were initiating and 7729 were non-initiating turns. 1562 current urns ( $17 \%$ ) started with $/ \mathrm{w} /$ or $/ \mathrm{h} /$.

For the analysis we had 2 predictor variables: context from the A's turn (initiating or non-initiating) and first phoneme of the B's turn (34 unique phonemes). The outcome variable was whether the current turn (i.e., B's turn) was a content/open question.

## Results

The decision tree divides the data at each node of the tree starting from the top of the figure. Leaves of the tree at the bottom of the figure show the proportion of turns that are questions (see Fig.1).

The decision tree splits the data first based on the first phoneme of the turn. The exact division of the phonemes is as follows: $/ \mathrm{w} /$ and $/ \mathrm{h} /$ versus all the other phonemes, with the proportion of questions being higher for turns starting with $/ \mathrm{w} /$ and $/ \mathrm{h} /$. Thus, the decision tree, which is blind to our predictions, splits the data exactly in line with our predictions.

Following the branch that clusters the data on the right (/w/, /h/), the data is further clustered according to the type of the previous turn. If the previous turn was an initiating turn the proportion of question turns is considerably lower than if previous turn was not an initiating turn. If the

Figure 1: The decision tree of question turns split according to the sequential type of the previous turn and the first phoneme of the current turn. Non-IN: non-initiating turn, IN: initiating turn, phoneme transcription conventions come from the Switchboard corpus.

previous turn is not initiating, the data is further split into whether the phoneme of the current turn is $/ \mathrm{h} /$ or $/ \mathrm{w} /$. Note that proportion of questions is higher in $/ \mathrm{h} /(22 \%)$ leaf than in $/ \mathrm{w} /(13 \%)$. This may be because "well ...", is often used as a filler at the beginning of a turn and thus decreases the overall proportion of questions in $/ \mathrm{w} /$ leaf. Moreover, there are more turns overall that start with $/ \mathrm{w} /$ than with $/ \mathrm{h} /$, therefore the proportion in /w/ leaf is also lower.
In regard to the data clusters on the left (turns starting with phonemes other than $/ \mathrm{w} /$ and $/ \mathrm{h} /$ ), it is evident that the proportion of question turns is extremely low in all leaves of the tree.

Overall, the analysis confirmed our initial hypotheses. Furthermore, based on the analysis we can also expect that the probability of a turn being a question will be additionally boosted if both cues are present - namely, if an incoming turn starts with $/ \mathrm{w} /$ or $/ \mathrm{h} /$ and the previous turn is non-initiating.

## Experimental study

The corpus study suggested that the prior context and the initial phoneme of a turn helps identify questions statistically. The experimental study tests whether real people actually make use of these cues.

## Method

Participants. For the experiment 25 participants ( 14 male, 11 female) were recruited. Participants' age ranged from 21 -70 years $(M=32, S D=11)$. All participants were native speakers of English but had various nationalities (e.g., American, British, Canadian, Australian, Indian, Latvian).

Materials and design. In this experiment participants listened to series of audio samples extracted from the Switchboard corpus. Each sample consisted of a context turn (initiating or non-initiating) produced by the first speaker and a response produced by the second speaker.

The context turn type could be either initiating (yes/no questions and wh-questions) or non-initiating (statements). The response turn type could be either content questions or non-questions. Each response turn was clipped to contain only the first phoneme, which could either be a wh phoneme (/w/ or $/ \mathrm{h} /$ ) or another phoneme. We therefore had the following fully crossed $2 \times 2 \times 2$ design: context type (initiating/non-initiating) x response type (content question/other) x response phoneme (wh/other). In addition, the response turn could be blank (no audio, with context being initiating or non-initiating). This resulted in 10 conditions.

Table 1: Example of a 10 conditions consisting of 2 types of context turn (initiating/non - initiating) and 5 types of response turn.

| Context turn |  | Response turn |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | /w/ |  | Other |  | Blank <br> no 2nd <br> turn |
|  |  | /w/ ques. | /w/ not quest. | not/w/ quest. | not/w/ non-quest. |  |
| Not initial | I do enjoy playing | Wh [at is your ...] | $\boldsymbol{W}$ [ell I wish...] | D [o you have...] | $\boldsymbol{Q}$ [uite a while...] | - |
| Initial | And how did it go | $\boldsymbol{W h}$ [at is your...] | $\boldsymbol{W}$ [ell I wish...] | D [o you have...] | $\boldsymbol{Q}$ [uite a while...] | - |

Context type was manipulated to test the effect of context and response phoneme was manipulated to test the effect of the first phoneme. Response type was manipulated so that we could assess whether the other question cues (e.g., raised pitch at the beginning of the question word) contribute in question prediction. The blank turn was added to establish a baseline for predicting an upcoming question without an initial phoneme.

We used the software Praat (Boersma \& Weenink, 2014) to cut and concatenate each first turn with each second turn (e.g., (first turn: statement) + (second turn: /w/ from whquestion)). Subsequently, each turn pair was processed in the software Audacity (Mazzoni \& Dannenberg, 2000) by adjusting a gap between the turns, so that the gap between first and second turn was 250 ms .

We created 25 samples for each of the 10 conditions, resulting in 250 unique audio samples. These were split into 5 groups of 50 samples so that each context sample or response sample only appeared once inside each group.

Procedure. The experiment was presented via the online software Qualtrics (Snow \& Mann, 2010). In each trial, a participant clicked a button to listen to a sample through headphones. Then they were asked to determine whether the second person would ask a question or not by means of completing a sentence "The Second turn is $\qquad$ " on the screen by pressing one of two buttons: "not a question" or "a question". The experiment began with 2 practice trials ensuring that participants understood the task. Participants were assigned to an audio sample group and heard the samples from that group in a random order.

## Results

We excluded 1 participant from the analysis due to the fact that this participant took 3 times longer than other participants to complete the experiment ( 38 minutes compared to an average of 12 minutes).

A logistic mixed model was used to predict whether the participant thought the response turn was a question (binary decision, yes or no, using the R package lme4, Bates et al, 2015). The predictor variables were context (initiating/noninitiating) and phoneme (wh, other, none). These predictors were coded as fixed effects and compared to a baseline model which included fixed effect of trial, random effect of context sample and phoneme sample, random effect of
participant and random slopes for context and phoneme by participant.

There was a significant main effect of context $\left(\chi^{2}(1)=\right.$ $45.74, p<.001)$. Participants were more likely to rate the turn as a question when preceded by a non-initiating context than an initiating context (see Fig.2).

There was a significant main effect of phoneme $\left(\chi^{2}(2)=\right.$ $13.83, p<.001)$. Turns that started with wh phonemes were more likely to be rated as questions in comparison to turns starting with other phonemes or without the response from the second speaker. The model estimated that the probability of considering a turn a question was $90 \%$ for wh phonemes compared to $71 \%$ for other and $70 \%$ for none in noninitiating context. In initiating context this was $9 \%$ compared to $4 \%$ for other and $2 \%$ for none. There was no significant difference in question prediction between other phoneme and no response. Considering that there was only one variant of $/ \mathrm{h} /$ responses present in our stimuli, we ran analysis with these trials removed. There was no difference in the results with or without these trials.

Importantly, we also assessed whether participants could differentiate between the type of the response sample (a question or not) from which the phoneme was extracted. We found no effect of the response type $\left(\chi^{2}(1)=0.11, p=.75\right)$.

Figure 2: Raw proportions of participants answering that an incoming turn is a question based on the previous context and the first phoneme of the incoming turn.
Error bars indicate 95\% CI of observations grouped within participants.


## Context



Thus, participants answered comparably to the phoneme samples that actually were questions and samples that were not questions. Most importantly, there was no interaction between response phoneme and the type of the response $\left(\chi^{2}=0.008, p=0.93\right)$. Thus, participants treated wh phonemes from real questions comparably to wh phonemes from other speech acts. These results suggest that participants are responding to the phoneme, not any other acoustic cue in the sample.

There was no significant interaction between context and phoneme $\left(\chi^{2}(2)=1.34, p=.51\right)$, although the trend was in the predicted direction.

## Discussion and conclusions

In the present paper we set out to explore whether the first phoneme of a turn and the prior context can serve as a cue to question recognition. We found that both of these features contribute to this process. Although an effect of context was clearly expected, it was less certain whether there would be an effect of the first phoneme. This is the first experimental study supporting the claim of Slonimska \& Roberts (accepted) that the first phoneme of question words can be used to predict an upcoming question.

We approached this topic from two different but mutually enhancing perspectives. We first assessed the hypothesis by analyzing natural conversations. Thus, we could look for patterns in ecologically valid data. The fact that the decision tree generated the same predictions as our hypothesis served as a sound basis for an experimental testing. Indeed, the samples from the corpus were used as experimental materials and the design was partly informed by the interaction between the two factors that the corpus study suggested. The hypotheses were also confirmed in the experiment, but there were two minor differences. First, the initial phoneme had a stronger effect than context in corpus study and vice versa in the experiment. Secondly, the corpus study predicted an interaction between initial phoneme and context, which was not found in the experiment. This may be because the probability of occurrence of various combinations is different in the corpus compared to the experiment, the experiment did not have enough statistical power, or more generally there is a difference between cues that are present in the data and ones that are actually used by people.

Another obvious difference between the two studies is that the speakers in the corpus had more prior context information than participants in the experimental study. Future experimental studies could include more extensive contextual information for the participants to be able to make predictions about the incoming turn.

Furthermore, the experimental participants were only passive listeners of the audio samples and their responses were not on-line. Future studies could take advantage of new paradigms to make it possible to combine interactive conversation with the use of controlled audio samples (e.g. Bögels, Magyari \& Levinson, 2015).

Slonimska \& Roberts (accepted) argue that question words tend to sound similar at the beginning of the word within a language to trigger question recognition. This leads to a prediction that $/ \mathrm{w} /$ should be a better cue than $/ \mathrm{h} /$, considering that there are more question words starting with $/ \mathrm{w} /$ than $/ \mathrm{h} /$. We found support for this in the corpus study. However, there was only one instance of $/ \mathrm{h} /$ phoneme in the experimental samples. We ran analyses with /h/ samples excluded and found no difference in the results. Therefore, although $/ \mathrm{w} /$ appears to boost question recognition, generalization to wh phonemes in English may not be warranted. Future studies could consider the differences between hearing $/ \mathrm{w} /$ and $/ \mathrm{h} /$ at the beginning of a turn in regard to question recognition.

It could be argued that the effect sizes in either study are too small to cause an evolutionary change in the language. However, we point out that even a small pressure would exert itself many times even in one conversation, and across cultural evolutionary time, small changes can accumulate to cause substantial changes.

Importantly, we advocate the virtuous cycle of looking for the phenomena in natural data, testing it in a controlled way and referring back to the real world. It can raise new questions and, most importantly, research can proceed in a more valid way than by using a single approach. This is clearly evident in our study - two approaches used in our study revealed differences that are important to account for, and which a single approach would have missed.

The findings in this paper are limited to English language and future research should continue exploring this cue in other languages, as well as diachronically. Only in this way can we be certain that this is not a single-language phenomenon or based on some idiosyncrasy of English but is actually a universal tendency. However, the puzzle remains - why else would question words sound so similar within so many languages (given that Slonimska \& Roberts account for historical factors in their study and still find significant similarities)?

To summarize, by using different approaches in exploring the same topic we now have converging evidence for the question word similarity hypothesis: first, question words tend to sound similar within languages (Slonimska \& Roberts, accepted); also, this phonetic cue can help in predicting questions in real conversations as shown in the corpus analysis; and finally people actually use this cue to predict questions when presented in a semi-natural setting. Thus, we suggest that the tendency for question words to sound similar is not a random occurrence, but might have evolved under a selective pressure to act as one of the early cues for question recognition in interactive conversation.

## Acknowledgements

This research was supported by an European Research Council Advanced Grant No. 269484 INTERACT to Stephen Levinson and a Leverhulme early career fellowship to Seán Roberts (ECF-2016-435). We thank the Max Planck Institute for additional support.

## References

Barthel, M., Meyer, A. S., \& Levinson, S. C. (2017). Next Speakers Plan Their Turn Early and Speak after TurnFinal "Go-Signals". Frontiers in Psychology, 8, 393.
Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48
Boersma, P., \& Weenink, D. (2014). Praat: doing phonetics by computer. 7 http://www. fon. hum. uva. nl/praat/. Zugegriffen: 17.
Bögels, S., \& Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. Journal of Phonetics,52, 46-57.
Bögels, S., Casillas, M., \& Levinson, S. C. (2016). To plan or to listen? The trade-off between comprehension and production in conversation. In the Eighth Annual Meeting of the Society for the Neurobiology of Language.
Bögels, S., Kendrick, K. H., \& Levinson, S. C. (2015). Never Say No... How the Brain Interprets the Pregnant Pause in Conversation. PloS one, 10(12), e0145474.
Bögels, S., Magyari, L., \& Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. Scientific reports, 5, 12881.

Cysouw, M. (2004, February). Interrogative words: an exercise in lexical typology. In Presentation presented at the Bantu grammar: description and theory workshop, February (Vol. 13).
Enfield, N. J., Stivers, T., \& Levinson, S. C. (2010). Question-response sequences in conversation across ten languages: An introduction. Journal of Pragmatics, 42(10), 2615-2619.
Gisladottir, R. S., Chwilla, D. J., \& Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. PloS one, 10(3), e0120068.
Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., \& Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation, 44(4), 387-419.
Holler, J., \& Kendrick, K. H. (2015). Unaddressed participants’ gaze in multi-person interaction: optimizing recipiency. Front. Psychol, 6(98), 10-3389.
Holler, J., Kendrick, K. H., Casillas, M., \& Levinson, S. C. (2016). Turn-Taking in Human Communicative Interaction. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-825-2
Hothorn, T., Hornik, K., \& Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.
Kendrick, K. H., \& Torreira, F. (2015). The timing and construction of preference: a quantitative study. Discourse Processes, 52(4), 255-289.
Levelt, W. J. (1993). Speaking: From intention to articulation (Vol. 1). MIT press.

Levinson, S. C., \& Torreira, F. (2015). Timing in turntaking and its implications for processing models of language. Frontiers in Psychology, 6, 731.
Levinson, S. C. (2013). Action formation and ascription. In The handbook of conversation analysis (pp. 103-130). Wiley-Blackwell.
Levinson, S. C. (2016). Speech acts. In Y. Huang (Ed.), Pragmatics. Advance online publication. Oxford: Oxford University Press.
Mazzoni, D., \& Dannenberg, R. (2000). Audacity (software). The Audacity Team, Pittsburg, PA, USA.
Roberts, S. G., Torreira, F., \& Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. Frontiers in Psychology, 6, 509.
Rossano, F. (2012). Gaze in conversation. In J. Sidnell, \& T. Stivers (Eds.), The handbook of conversation analysis (pp. 308-329). Malden, MA: Wiley-Blackwell.
Rossano, F., Brown, P., \& Levinson, S. C. (2009). Gaze, questioning and culture. Conversation analysis: Comparative perspectives, 187-249.
Sacks, H., Schegloff, E. A., \& Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 696-735.
Schriefers, H., Meyer, A. S., \& Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies.Journal of Memory and Language, 29(1), 86-102.
Sicoli, M. A., Stivers, T., Enfield, N. J., \& Levinson, S. C. (2014). Marked initial pitch in questions signals marked communicative function. Language and Speech, 58(2), 204-223.
Snow, J., \& Mann, M. (2013). Qualtrics survey software: handbook for research professionals.
Slonimska \& Roberts (accepted). A case for systematic sound symbolism in pragmatics: universals in wh-words. Journal of Pragmatics.
Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... \& Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, 106(26), 10587-10592.
Strobl, C., Malley, J., \& Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods,14(4), 323.

# Learning to See People Like People: Predicting Social Impressions of Faces 

Amanda Song (mas065@ucsd.edu) ${ }^{\dagger}$<br>Department of Cognitive Science<br>University of California, San Diego, CA 92093

Li Linjie (li2477@purdue.edu) ${ }^{\dagger}$<br>Department of Computer Science, Purdue University 610 Purdue Mall, West Lafayette, IN 47907

Chad Atalla (catalla@ucsd.edu)<br>Department of Computer Science and Engineering<br>University of California, San Diego, CA 92093

Garrison Cottrell (gary@ucsd.edu)<br>Department of Computer Science and Engineering<br>University of California, San Diego, CA 92093


#### Abstract

Humans make complex inferences on faces, ranging from objective properties (gender, ethnicity, expression, age, identity, etc) to subjective judgments (facial attractiveness, trustworthiness, sociability, friendliness, etc). While the objective aspects of face perception have been extensively studied, relatively fewer computational models have been developed for the social impressions of faces. Bridging this gap, we develop a method to predict human impressions of faces in 40 subjective social dimensions, using deep representations from state-of-the-art neural networks. We find that model performance grows as the human consensus on a face trait increases, and that model predictions outperform human groups in correlation with human averages. This illustrates the learnability of subjective social perception of faces, especially when there is high human consensus. Our system can be used to decide which photographs from a personal collection will make the best impression. The results are significant for the field of social robotics, demonstrating that robots can learn the subjective judgments defining the underlying fabric of human interaction. Keywords: social impression; deep learning; face perception


## Introduction

With the huge success of deep learning techniques, current state-of-the-art computer vision algorithms have approached or exceeded human ability in recognizing a face (Taigman, Yang, Ranzato, \& Wolf, 2014; Stewart, Andriluka, \& Ng, 2016) and identifying the objective properties of a face, such as age and gender estimation, (Guo, Fu, Dyer, \& Huang, 2008). However, humans not only read objective properties from a face, like expression, age, and identity, but also form subjective impressions of social aspects of a face (Todorov, Olivola, Dotsch, \& Mende-Siedlecki, 2015) at first sight, such as facial attractiveness (Thornhill \& Gangestad, 1999), friendliness, trustworthiness (Todorov, Baron, \& Oosterhof, 2008), sociability, dominance (Mignault \& Chaudhuri, 2003), and typicality. In spite of the subjective nature of social perceptions, there is often a consensus among human in how they perceive attractiveness, trustworthiness, and dominance

[^204]in faces (Falvello, Vinson, Ferrari, \& Todorov, 2015; Eisenthal, Dror, \& Ruppin, 2006). This indicates that faces contain high-level visual cues for social inferences, therefore making it possible to model the inference process computationally. Social judgments, as an important part of people's daily interactions, have a significant impact on social outcomes, ranging from electoral success to sentencing decisions (Oosterhof \& Todorov, 2008; Willis \& Todorov, 2006).

Are deep learning models, which are successful in various visual tasks, also capable of predicting subjective social impressions of faces? Even before the advent of deep learning, there have been models using traditional computer vision algorithms and simulated faces to model the perception of facial attractiveness (Thornhill \& Gangestad, 1999; Eisenthal et al., 2006; Kagian et al., 2008; Gray, Yu, Xu, \& Gong, 2010), trustworthiness (Falvello et al., 2015; Todorov, Baron, \& Oosterhof, 2008), sociability, aggressiveness (Mignault \& Chaudhuri, 2003), familiarity (Peskin \& Newell, 2004), and memorability (Bainbridge, Isola, \& Oliva, 2013; Khosla, Bainbridge, Torralba, \& Oliva, 2013). Recently, there has been work on modeling the "big five" personality traits perceived by humans when viewing another person in video clips (Escalera et al., 2016).

In this paper, we examine human social perceptions of faces in 40 dimensions extensively and systematically. We evaluate the human consistency and correlation in 40 social features ( 20 relevant pairs) that are typically studied by social psychologists (Todorov, Said, Engell, \& Oosterhof, 2008), and relevant to social interactions (Todorov et al., 2015; Oosterhof \& Todorov, 2008), and use state-of-the-art deep learning algorithms to model all 40 of them. Using the internal representations learned from the deep learning models, our model can successfully predict human social perception whenever human have a consensus. We further visualize the key features defining different social attributes to facilitate a understanding of what makes a face salient in a certain social dimension.

## Methods

## Dataset

To predict human social impressions of faces, we use a public dataset (Bainbridge et al., 2013) consisting of 2,222 face images and annotations for 40 social attributes. Each attribute is rated on a scale of $1-9$ by 15 subjects. We take the average rating from all raters as a collective estimation of human judgment for the social features of each face.

The 40 social attributes consist of 20 pairs of related traits: (attractive, unattractive), (happy, unhappy), (friendly, unfriendly), etc. Some of these traits are highly correlated and predictable from others, especially within the trait pairs. To understand the human-perceived correlations between these traits, we compute the Spearman's rank correlation between the average human ratings of every pair of social features and show their correlations in a heatmap (Figure 1(a)). We order traits in the map based on similarity and positive or negative connotation. From the figure, we see that negative social features such as untrustworthy, aggressive, cold, introverted, and irresponsible form a correlated block. Likewise, the most positive features such as attractive, sociable, caring, friendly, happy, intelligent, interesting, and confident are highly correlated with each other. Although we choose 20 pairs of opposite features, they are not completely complementary and redundant. Principal Component Analysis of the covariance matrix shows that it takes 24 principal components to cover $95 \%$ of the variance.

## Regression Model for Social Attributes

After averaging human ratings, each face receives a continuous score from 1 to 9 in all social dimensions. We model these social scores with a regression model. We propose a ridge regression model on either features from deep convolutional neural networks (CNN) or traditional face geometry based features, and present results from both feature sets. Such visual features are usually high-dimensional, so we first perform Principal Component Analysis (PCA) on the extracted features of the training set to reduce dimensionality. The PCA dimensionality is chosen by cross-validation on a validation set, separately for each trait. The PCA weights are saved and further used in fine-tuning our CNN-regression model.

## Regression on Geometric Features

Past studies have found that facial attractiveness can be inferred from the geometric ratios and configurations of a face (Eisenthal et al., 2006; Kagian et al., 2008). We suggest that other social attributes can also be inferred from geometric features. We compute 29 geometric features based on definitions described in (Ma, Correll, \& Wittenbrink, 2015), and further extract a 'smoothness' feature and 'skin color' feature according to the procedure in (Eisenthal et al., 2006; Kagian et al., 2008). The smoothness of a face was evaluated by applying a Canny edge detector to regions from the cheek and forehead areas (Eisenthal et al., 2006). The more edges detected, the less smooth the skin is. The regions we chose
to compute smoothness and skin color are highlighted in the right subplot of Figure 2. The skin color feature is extracted from the same region as smoothness, converted from RGB to HSV. However, regressing on these handcrafted features alone is not enough to capture the richness of geometric details in a face. We therefore use a computer vision library (dlib, $\mathrm{C}++$ ) to automatically label 68 face landmarks (see Figure 2) for each face, and then compute distances and slopes between any two landmarks. Combining 29 handcrafted geometric features, smoothness, color and the distance-slope features, we obtain 4592 features in total. Since the features are highly correlated, we apply PCA to reduce dimensionality. Again, the PCA dimensionality is chosen by cross-validating on the hold out set separately for each facial attribute. Then a ridge regression model is applied to predict social attribute ratings of a face. The hyper-parameter of ridge regression is selected by leave-one-out validation within the training set.

## Regression on CNN Features

Previous studies have shown that pretrained deep learning models can provide feature representations versatile for related tasks. We therefore extract image features from pretrained neural networks, choosing from six architectures with different original training goals: (1) VGG16, trained for object recognition (Simonyan \& Zisserman, 2014), (2) VGGFace, trained for face identification (Simonyan \& Zisserman, 2014), (3) AlexNet, trained for object classification (Krizhevsky, Sutskever, \& Hinton, 2012), (4) Inception from Google, trained for object recognition (Szegedy et al., 2015), (5) a shallow Siamese neural network that we train from scratch to cluster faces by identity, (6) a state of the art VGG-derived network (Face-LandmarkNN) trained for the face landmark localization task.

To find the best CNN features among the six networks, we first find the best-performing feature layers of each network in the ridge regression prediction task. Before the ridge regression, we perform PCA and pick the PCA dimensionality that gives best results on the validation set. Then, we compare the results among networks to select the best features overall.

## Results

After comparing all 6 networks, we find that the conv5_2 layer of VGG16 (trained for object classification) lead to the best results. This set of features significantly outperforms the three networks trained solely on faces, while also slightly outperforming AlexNet and Inception networks. These bestperforming CNN features also exceed the prediction correlation of the geometric features in most attributes. Figure 3 compares prediction performance of the CNN model and the geometric feature model.

We speculate that the poor performance from the face recognition networks can be attributed to their optimization for specific facial tasks. Learning face landmark configurations and differences between faces that define identity may not correlate well with the task at hand, which looks for commonalities behind certain social features beyond identity. The


Figure 1: Correlation heatmaps among social features. (a): human; (b): CNN-based model.


Figure 2: 68 face landmarks labeled by dlib software automatically. The gray regions are used for computing smoothness and skin color.
landmark networks should presumably give results similar to the geometric features, but did not learn features corresponding to all of the features we manually extracted.

We also try fine-tuning the best performing CNN model with back propagation but do not observe further improvement in performance. Hence our reported results are without fine-tuning.

To evaluate model performance, we did a random train/validation/test split 50 times, with a ratio of 64/16/20 respectively. The prediction performance of our model is evaluated using Pearson's correlation with the average human ratings on the test set. For each social attribute, we also compute human group consistency as an index of the strength of learning signal.

Among the social attributes, human subjects agree most about 'happy' and disagree most about 'unfamiliar.' For both regression models (CNN based regression and geometric fea-
ture based regression), model performance grows as the consensus on a social trait increases.

Since a change in expression would produce a change in landmark locations, it is not surprising that landmark-based geometric features achieve comparable or slightly higher correlation with the CNN model when predicting social attributes which are highly related to expressions (such as 'happy', 'unhappy', 'cold' and 'friendly' etc). For other social attributes, the CNN model performs better, by about 0.04 higher in correlation on average. This implies that CNN features encode much more information than landmark-based features. It is useful to visualize such features to understand what aspects make them powerful enough to predict social attributes.

## Evaluating Against Human Consensus

An important gauge of model success is quantitative comparison between the subjective social features predicted by our best performing model and those perceived by humans. We take our model predictions, compute the Spearman correlation between every pair of traits, and display them in a heatmap (see Figure 1 (b)). The resulting heatmap shares similar patterns with the figure generated from average human ratings (see the left panel in Figure 1). Pearson Correlation between the upper triangle of the two similarity matrices (human and model prediction) is 0.9836 . This suggests that our model successfully preserves human-perceived relationships between traits.

Since these social impressions are subjective ratings, it is informative to examine the extent with which people agree


Figure 3: Model comparison on 40 social features.
with each other on these judgments. To calculate human group consistency, we perform the following procedure 50 times for each attribute and then average the results: (1) For each face, we randomly split the 15 raters into two groups of 7 and 8. (Note: The raters assigned to each face are generally different sets). (2) We calculate the two groups' average ratings for each face, obtaining two vectors of length 2,222 (the number of faces in the dataset). (3) Finally, we calculate the Pearson correlation between the two vectors. We find that human agreements covary with model performance and observe an extremely high correlation, as illustrated in Figure 4.


Figure 4: Human within group consistency vs. model's correlation with human average. Pearson correlation $\rho=0.98$, $p<10^{-5}$

## Feature Visualization

Here, we visualize features from our model which are important for social perceptions. We choose facial attractiveness as
an example, but the same method can be applied to the other social features.

To identify visual features that ignite attractiveness perception, we find the top 9 units of highest influence on attractiveness at conv5_2 as follows. First, we compute a product of three terms: (1) A unit's activation from conv5_2, (2) that unit's weight to the following fc_PCA layer, (3) the fc_PCA unit's weight to the output unit. We then sort all conv5_2 units' average products of these three terms and identify the top 9 neurons that contribute to the output neuron for the corresponding social feature. Then we employ the method described in (Yosinski, Clune, Nguyen, Fuchs, \& Lipson, 2015; Zeiler \& Fergus, 2014) to find top-9 input images that cause high activations in each of the top-9 conv5_2 neurons. Also we use deconvolution to create an image of the features activating that unit for each face, with varying levels of success.

Figure 5 captures the features that are important for predicting the attractiveness of a face. The feature importance descends from left to right and top to bottom. The important features identified by our model are related to eyes, hair with bangs, high nose-bridges, high cheeks, dark eyebrows, strong commanding jawlines, chins, and red lips. Note that among the 9 cropped input image patches, not all the faces are perceived as attractive overall; despite having a feature that contributes to attractiveness. An attractive face needs to activate more than one of these features in order to be considered attractive. This observation agrees with our intuition that attractiveness is a holistic judgment, requiring a combination of multiple features.

It also seems that several attractiveness features include relationships between different facial features. For example, while the first feature in the upper left of the figure emphasizes the eye, it also includes the nose. This is also true of the upper right feature. Additionally, smiling is important in perceived attractiveness, as emphasized by the feature in the lower left of the figure.


Figure 5: Visualization of features in the pretrained-VGG16 regression network. For conv5_2 layer, we show the top 9 activations of the top 9 neurons that maximally activate the attractiveness neuron across the training data, projected down to pixel space using the deconvolutional network approach (Zeiler \& Fergus, 2014) and their corresponding cropped image patches. Best viewed in electronic form, and zoomed in.

## Conclusion

We have shown that a deep network can be used to predict human social perception of faces, achieving high correlation with the average human ratings. As far as we know, this is the widest exploration of social judgment predictions, showing human-like perceptions on 40 social dimensions. Reflecting previous work in recognizing facial expressions, where happiness is the easiest to recognize, our highest correlation is on the happy feature. However, previous work in this area tends to classify a face as happy or not, rather than the degree of rated happiness. By predicting this as a continuous value, rather than categorical data, the subjective nature of human judgment is modeled smoothly, along with the subjective face trait landscape.

We find that, for attributes which are recognized via facial actions, such as happy, unhappy, or aggressive (probably associated with anger) or lack of facial action, such as cold or unemotional, a simple regression model based on the placement of facial landmarks works well, although the deep network performs nearly as well.

Of greater significance is our model's correlations with human judgments for traits such as trustworthiness, responsibleness, confidence, and intelligence, which correspond to more static features of the face. In this area, the deep network, which responds to facial textures and shape, has superior performance. While these judgments do not correspond to the traditional notion of "ground truth", they are descriptions for which humans have a fair amount of agreement, suggesting
the presence of a signal to be recognized.
Furthermore, we have shown, yet again, that a machine can recognize attractiveness. For this dataset, our deep network correlates with average human ratings at 0.75 . This provides a new benchmark for this dataset. This is one of a few areas where the deep network significantly outperforms the geometric features, as skin texture is likely to matter.

Many of these features are redundant. For example, friendly and happy are highly correlated (see Figure 1, and the red block indexed by happy and friendly). Similarly, aggressive and mean are highly correlated, which presumably requires not smiling. Meanwhile, it is also noteworthy that some traits considered to be "opposite" in this list are not simply the inverse of one another. For example, there is a large difference in human agreements on "sociable" $(0.74)$ versus "introverted" $(0.50)$, suggesting they are not opposites.

We also examined some of the features from the deep network. It is notable that these are difficult to verbalize, which is quite different from geometric features.

These results are significant for the field of social robotics. While a robot should not purely judge a human on appearance, much of human interaction is dictated by the underlying fabric of social impressions. Thus, it is important for a robot to be aware of this subjective social fabric, opening the door to useful knowledge such as whether humans might judge a person to be trustworthy. These judgments may happen subconsciously for humans, while a robot can be more objective, predicting these judgments and objectively choosing when to
consider them in a decision. A robot need not treat an attractive or unattractive person differently for its own purposes, but this knowledge could affect how interactions are made for the sake of the human, knowing in advance how that person may feel that they fit into the social landscape.

Expansions on this work may include investigating the image properties that determine high level social features, beyond the attractiveness features we display in Figure 5. Additionally, social trait prediction may benefit from a single model with a shared representation, while this paper approaches each attribute as a separate regression task.

For future work, we aim to develop a generative model which can automatically modify a face's attributes (either objective or subjective) while preserving its realism and identity. Practically speaking, such a model could improve a face's perceived social features in positive ways (e.g. make a face look more sociable, trustworthy). More importantly, it would enable psychologists to quantify human biases during the formation of social impression in a precise and systematic manner. Psychologists could generate variants of a real face differing in age, gender, race, and explore how various factors separately and jointly affect the social impressions of faces.

## References

Bainbridge, W. A., Isola, P., \& Oliva, A. (2013). The intrinsic memorability of face photographs. Journal of Experimental Psychology: General, 142(4), 1323.
Eisenthal, Y., Dror, G., \& Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. Neural Computation, 18(1), 119-142.
Escalera, S., Torres Torres, M., Martinez, B., Baró, X., Jair Escalante, H., Guyon, I., ... others (2016). Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-8).
Falvello, V., Vinson, M., Ferrari, C., \& Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. Social Cognition, 33(5), 368.
Gray, D., Yu, K., Xu, W., \& Gong, Y. (2010). Predicting facial beauty without landmarks. In The European Conference on Computer Vision (ECCV-2010) (pp. 434-447). Springer.
Guo, G., Fu, Y., Dyer, C. R., \& Huang, T. S. (2008). Imagebased human age estimation by manifold learning and locally adjusted robust regression. IEEE Transactions on Image Processing, 17(7), 1178-1188.
Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., \& Ruppin, E. (2008). A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. Vision Research, 48(2), 235-243.
Khosla, A., Bainbridge, W. A., Torralba, A., \& Oliva, A. (2013). Modifying the memorability of face photographs. In International Conference on Computer Vision (ICCV2013) (pp. 3200-3207).

Krizhevsky, A., Sutskever, I., \& Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1097-1105).
Ma, D. S., Correll, J., \& Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. Behavior Research Methods, 47(4), 1122-1135.
Mignault, A., \& Chaudhuri, A. (2003). The many faces of a neutral face: Head tilt and perception of dominance and emotion. Journal of Nonverbal Behavior, 27(2), 111-132.
Oosterhof, N. N., \& Todorov, A. (2008). The functional basis of face evaluation. Proceedings of the National Academy of Sciences, 105(32), 11087-11092.
Peskin, M., \& Newell, F. N. (2004). Familiarity breeds attraction: Effects of exposure on the attractiveness of typical and distinctive faces. Perception, 33(2), 147-158.
Simonyan, K., \& Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Stewart, R., Andriluka, M., \& Ng, A. Y. (2016, June). End-to-end people detection in crowded scenes. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2016).
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2015) (pp. 1-9).
Taigman, Y., Yang, M., Ranzato, M., \& Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2014) (pp. 1701-1708).
Thornhill, R., \& Gangestad, S. W. (1999). Facial attractiveness. Trends in Cognitive Sciences, 3(12), 452-460.
Todorov, A., Baron, S. G., \& Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. Social Cognitive and Affective Neuroscience, 3(2), 119-127.
Todorov, A., Olivola, C. Y., Dotsch, R., \& Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. Annual Reviews of Psychology, 66(1), 519.
Todorov, A., Said, C. P., Engell, A. D., \& Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. Trends in Cognitive Sciences, 12(12), 455-460.
Willis, J., \& Todorov, A. (2006). First impressions making up your mind after a $100-\mathrm{ms}$ exposure to a face. Psychological science, 17(7), 592-598.
Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., \& Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.
Zeiler, M. D., \& Fergus, R. (2014). Visualizing and understanding convolutional networks. In European Conference on Computer Vision (ECCV-2014) (pp. 818-833).

# A Rational Approach to Stereotype Change 

Jake Spicer (J.Spicer@ warwick.ac.uk)<br>Department of Psychology, Humanities Building, University of Warwick, Coventry, UK

Adam Sanborn (A.N.Sanborn@warwick.ac.uk)<br>Department of Psychology, Humanities Building, University of Warwick, Coventry, UK


#### Abstract

Existing theories of stereotype change have often made use of categorisation principles in order to provide qualitative explanations for both the revision and maintenance of stereotypical beliefs. The present paper examines the quantitative methods underlying these explanations, contrasting both rational and heuristic models of stereotype change using participant data and model fits. In a comparison of three models each simulating existing descriptions of stereotype change, both empirical data and model fits suggest that stereotypes are updated using rational categorisation processes. This presents stereotype use as a more rational behaviour than may commonly be assumed, and provides new avenues of encouraging stereotype change according to rational principles.


Keywords: Stereotypes; Categorisation; Rational Behaviour

## Introduction

Stereotypes have often been found to be resistant to change, with beliefs and expectations regarding a group often persisting even when faced with directly contradictory information (Hilton \& von Hippel, 1996). This presents a problem when trying to combat stereotypes underlying prejudice or discrimination through out-group exposure as has often been suggested by theories such as the Contact Hypothesis (Allport, 1954), as there is no assurance that simply demonstrating the inaccuracy of these beliefs will be effective in encouraging revision. It is therefore necessary to examine the processes by which stereotypes are updated with experience, and, in cases of stereotype persistence, determine how counter-stereotypical information can be disregarded in order to develop better methods to encourage change.

Past research into this field has offered three possible processes of stereotype revision: book-keeping, in which the stereotype is slowly adjusted with each relevant observation; conversion, in which the stereotype can undergo sudden and drastic changes in response to particularly notable contradictory exemplars; and subtyping, in which counter-stereotypical evidence is isolated from the rest of the category in a distinct subgroup, ignored when making category judgements. This presents three potential explanations for stereotype persistence: stereotype-incongruent exemplars may be noted via book-keeping but remain out-weighed by prior stereotypical beliefs; these exemplars may not have been sufficiently significant to evoke change via conversion; or these exemplars may have been excluded entirely via subtyping.

This distinction was examined by Weber and Crocker (1983) by manipulating the presentation format of counterstereotypical evidence in summaries of lawyers: equal
amounts of stereotype-incongruent evidence were either concentrated into only a few exemplars, or dispersed across many exemplars. This generates three competing expectations between the three models: conversion suggests that these concentrated exemplars would act as extreme disconfirmers, encouraging significant revision to the stereotype. Conversely, subtyping would suggest that concentrating incongruent evidence should make it easier to isolate, thereby preserving existing stereotypical beliefs. Book-keeping, meanwhile, focuses only on the amount of data rather than the presentation format, and so suggests no difference between these conditions. Measures of the strength of stereotypical beliefs following exposure to these exemplars were found to be stronger in the concentrated condition, supporting the subtyping model, an effect that has since been replicated in a number of studies (Bott \& Murphy, 2007; Johnston \& Hewstone, 1992).

This depicts stereotype persistence as an issue of categorisation, occurring where counter-stereotypical group members are placed in a distinct subgroup rather than integrated into existing structures. The mechanisms underlying stereotype revision could then be well described by existing models of categorisation, particularly those which perform a similar process of partitioning a category into lower-order subgroups. One key example of such a model is the Rational Model of Categorisation (RMC) developed by Anderson (1991), which organises a category into 'clusters' of exemplars based on similarities in observed features. The organisation of these clusters then determines the impact of stored data on subsequent judgements, with larger clusters tending to have more influence on expectations of given traits appearing in the category.

The subtyping effect could therefore be seen as the result of standard categorisation processes creating partitions of the category based on observed data patterns which determine the impact of incongruent information on later judgements: isolating this information leads to subtyping, diminishing its impact, while integration of congruent and incongruent information leads to book-keeping, and so greater stereotype revision. If so, subtyping may not be the result of a bias towards ignoring counter-stereotypical data in order to preserve stereotypical beliefs, but a rational incorporation of all available data under certain data patterns which happens to mitigate the influence of stereotype-incongruent information. As such, subtyping could be considered a more rational process
than it may initially seem, and so could be fought using similarly rational mechanisms to encourage stereotype change.

The present study therefore presents a rational approach to stereotype use; in the following paper, we develop several candidate models to approximate the existing depictions of stereotype revision, contrast the predictions of these models with participant data to assess their accuracy, and use these findings to offer some insight into the process of stereotype change.

## Model Details

We began by developing an edited version of the RMC in order to examine the categorical explanation proposed for the subtyping effect. This made use of the standard version of the RMC using discrete exemplar dimensions as defined by Anderson (1991), chosen for its reasonable level of simplicity and ease of application to the design of Weber and Crocker (1983). The RMC assigns exemplars sequentially to a cluster based on similarities in observed features using a Bayesian model to approximate the ideal partition:

$$
\begin{equation*}
p(k \mid f)=\frac{p(k) p(f \mid k)}{\sum_{k} p(k) p(f \mid k)} \tag{1}
\end{equation*}
$$

where $k$ is the cluster and $f$ is the feature set of the exemplar under consideration. This posterior probability is calculated for all existing clusters as well as a new potential cluster, with the highest probability determining assignment. Following Anderson (1991), the prior probability was defined as:

$$
p(k)= \begin{cases}\frac{c n_{k}}{(1-c)+c n} & \text { if } k \text { is old }  \tag{2}\\ \frac{(1-c)}{(1-c)+c n} & \text { if } k \text { is new }\end{cases}
$$

where $n_{k}$ is the number of exemplars in cluster $k, n$ is the total number of members assigned to the partition, and $c$ is a coupling parameter describing the probability of two exemplars being grouped together independent of any observations.

The likelihood also followed the format of Anderson (1991):

$$
\begin{equation*}
p(f \mid k)=\prod_{i} p_{i}(j \mid k) \tag{3}
\end{equation*}
$$

where the exemplar's features are divided into dimensions $i$ holding values $j$. As stated above, the discrete form of this probability was used, in keeping with the exemplar structure used by Weber and Crocker (1983):

$$
\begin{equation*}
p_{i}(j \mid k)=\frac{n_{j}+\alpha_{j}}{n_{k}+\alpha_{0}} \tag{4}
\end{equation*}
$$

where $n_{j}$ is the number of exemplars in cluster $k$ showing trait value $j$ on dimension $i, n_{k}$ is the number of cluster members showing a value on that dimension and $\alpha_{j}$ is a parameter reflecting the prior expectation of the occurrence of that value, while $\alpha_{0}$ is the sum of these alpha values.

Once a partition has been generated, the model is then able to calculate a probability value measuring the likelihood of a
new group member exhibiting a congruent trait value on any dimension. This is done by taking an average of the rate of congruent traits in each cluster weighted by the probability of that cluster:

$$
\begin{equation*}
p(\text { con })=\sum_{k} p(k) p(\text { con } \mid k) \tag{5}
\end{equation*}
$$

where $p($ con $\mid k)$ follows the format of Equation 4, focussing on congruent trait values. This explains how isolating incongruent data in a distinct subtype mitigates its impact: smaller clusters provide less evidence to outweigh prior expectations, here represented by the $\alpha$ parameter. As such, there is less confidence that future members of the incongruent cluster will demonstrate similar trait values, while the larger congruent cluster carries more certainty. To illustrate, consider a case in which 30 exemplars, 20 congruent and 10 incongruent, are either integrated or segregated. For the purposes of this illustration, $\alpha=1$ for both congruent and incongruent traits, and $c=1$, meaning no new cluster is considered:


As this shows, stereotype-congruency is estimated to be more probable in the segregated case because the $\alpha$ values are more impactful in the smaller cluster, offsetting the actual ratio of traits to a greater degree.

## Stereotypes as Prior Clusters

As the category in question is a familiar social group with which participants are likely to have previous experience, the model included a cluster of exemplars added to the partition before exposure to the main exemplar set in order to simulate this prior knowledge. This also provided a more valid depiction of the origins of the group stereotype by making the members of this prior cluster stereotype-congruent, as well as allowing for potential interactions between prior knowledge and new information, as have been observed in other categorical modelling studies (Heit, Briggs, \& Bott, 2004). Exemplars in the prior cluster therefore displayed stereotypecongruent values on all stereotypical dimensions, as well as group membership on a separate dimension, while the number of cluster members was added as an additional model parameter.

## Alternative Models

While the above demonstration does show that the RMC is able to predict a subtyping effect, in order to determine whether this is truly the result of a rational process, the RMC
must be compared with a more heuristic depiction of the subtyping mechanism. A second model was therefore developed in which the segregated incongruent data is ignored rather than simply mitigated, essentially redrawing category boundaries to exclude counter-stereotypical information. This was achieved by restricting the clusters considered when making probability estimates to only the cluster with the highest posterior probability. This restriction was based on the findings of Murphy and Ross (1994), which suggested that participants often only considered the most likely cluster when making probability estimates rather than all generated clusters. This creates an additional candidate model for assessment, labelled the Restricted Rational Model of Categorisation (RRMC).

In addition, as a counterpoint to this 'extreme subtyping' model, a third model was developed in which book-keeping was enforced; this was achieved by limiting the RMC to using only a single cluster by fixing the $c$ parameter at 1 , forcing all exemplars to be grouped together despite any differences in features. This 'Single Cluster Model' (SCM) therefore removes the possibility of segregating incongruent data, eliminating any influence of data format and focussing entirely on the ratio of traits in the partition.

These three models therefore present three different mechanisms of stereotype revision: while both the RMC and RRMC use a partition that flexibly adapts to observed data patterns, the RRMC subsequently simplifies this partition by focussing on only one cluster, heightening any effects this representation may have generated, while the RMC remains more moderate. Conversely, the SCM focuses on trait ratios rather than data pattern, thereby dismissing any effects that may be predicted by the other candidate models.

There is, however, a key distinction between these rational and heuristic models which can be used to determine their validity: in the RMC, the subtyping effect is dependent on the smaller size of the subtype cluster, meaning that increasing the size of the subtype by adding more incongruent members should reduce and ultimately eliminate this effect. In contrast, the RRMC will continue to ignore the subtype regardless of its size unless the subtype becomes so large that it is selected as the most likely cluster, at which point estimates will change drastically to reflect the subtype's much lower rate of congruency. This could essentially reverse the subtyping effect at higher volumes of incongruent evidence, focussing on counter-stereotypical rather than stereotypical clusters, and so bearing a closer resemblance to the conversion-effect described above. The SCM, meanwhile, is unable to exclude incongruent data at all, and therefore predicts no subtyping effect at any volume of incongruent information.

The accuracy of these models can therefore be contrasted according to the change in the subtyping effect with further exposure to stereotype-incongruent evidence: the RMC predicts a reduction in subtyping at higher volumes of counterstereotypical data; the RRMC predicts a stable subtyping effect until a sudden reversal; and the SCM predicts no sub-
typing effect at any point. The following experiment therefore set out to compare these model predictions by extending the concentration design of Weber and Crocker (1983) across a higher total volume of evidence and taking measures of stereotypical beliefs throughout exposure. This also provided direct behavioural data for use in assessing the fit of the candidate models for a more complete test of these predictions.

## Experimental Data

## Method

Participants One-hundred-and-sixteen participants were selected from a University of Warwick undergraduate psychology class as part of a course requirement. The sample included 102 females and 14 males, while age ranged between 18 and 27 years, with a mean of 19 .

Design and Materials The experiment followed the concentration design of Weber and Crocker (1983) with an additional within-subjects manipulation of data volume: measures of stereotypical beliefs were taken at fixed intervals during the observation of a set of exemplar descriptions where stereotype-incongruent information was either concentrated in a subset of exemplars or dispersed across all exemplars. Two exemplar sets were therefore created for use in the experiment, each containing 90 total exemplars displaying four trait dimensions: the first dimension described the occupational label, and so was identical for all exemplars, while the remaining three dimensions described personality traits with three possible values (stereotype-congruent, stereotypeincongruent or neutral). In both sets, two-thirds of the 270 total traits were incongruent, one-sixth were congruent and one-sixth were neutral; incongruent traits made up the majority in order to allow for a potential incongruent cluster to be larger than any other in the category. In the concentrated exemplar set, these incongruent traits were concentrated such that 60 exemplars each displayed incongruent traits on all three personality dimensions, with the congruent and neutral traits being distributed equally between the remaining exemplars. In the dispersed exemplar set, all traits were distributed as equally as possible.

As in Weber and Crocker (1983), exemplars were said to come from the category of lawyers; exemplars were therefore transformed into member summaries for use in the experiment by assigning each value on the three personality dimensions a unique trait label. Sixteen total labels were used: 5 congruent (Intelligent, Industrious, Neat, Out-going and Well-dressed), 5 incongruent (Incompetent, Lazy, Messy, Shy and Slovenly) and 6 neutral (Warm, Religious, Jovial, Obnoxious, Reserved and Meditative). These labels were taken from Weber and Crocker (1983), being based on pilot tests determining stereotypical and counter-stereotypical traits for the target category of lawyers. Three labels of each trait type were randomly selected at the start of each run of the experiment for use in exemplar summaries. Summaries were also assigned randomly selected names to assist in individuation.

Procedure Upon arriving at the lab, participants were first randomly assigned to one of the two concentration conditions, determining which set of exemplars would be viewed; this was balanced to provide equal numbers, meaning 58 participants were allocated to each condition. Participants were told the experiment tested how perceptions of a group changed with experience, involving both viewing summaries of group members and answering questions about the traits of the group in general.

The experiment began by asking participants to estimate the likelihood of certain traits appearing in the category of lawyers according to the number of members in a sample of 100 lawyers displaying that trait. Estimates were requested for all 16 possible personality traits, though only 9 were used in the subsequent member summaries. This first question block therefore provided a measure of baseline beliefs before any experimental exemplars were viewed.
After providing estimates for all traits, participants began a presentation block in which member summaries were shown on screen for the participants to examine. In order to maintain attention on this information, participants were asked to rate the pleasantness of each group member on a scale of 1-10, though this measure was not used during analysis.

At set intervals of presentation, the test block was repeated, and participants were again asked to estimate the likelihood of each of the 16 traits appearing in the category to measure any changes in expectation. This occurred after viewing 6, $18,36,60$ and 90 total exemplars, with the ratio of traits within each interval being consistent with that of the complete exemplar set. At the start of each test block, participants were informed that though some of the questions had been asked before, they should answer based on how they felt at that point in time.

After viewing all 90 lawyer summaries and completing the final test block, the experiment ended, and participants were debriefed as to the aims and expectations of the study.

## Results

## Data Analysis

The results of the experiment were analysed using a mixed linear regression model including the factors of evidence volume, concentration condition and trait type. As the first test block was intended to provide a baseline, being unaffected by either volume or concentration, ratings from this round were not included in the regression model. This was confirmed using independent $t$-tests, finding no significant difference between conditions in either congruent ratings $(\mathrm{t}(114)=.190$, p $=.850)$ or incongruent ratings $(\mathrm{t}(114)=.296, \mathrm{p}=.768)$ in the first test block.

The regression model showed significant effects for volume in both congruent $(\beta=-2.23, \mathrm{t}(5086)=6.41, \mathrm{p}<.001)$ and incongruent ratings $(\beta=6.49, \mathrm{t}(5086)=15.6, \mathrm{p}<.001)$, with congruent ratings decreasing and incongruent ratings increasing over the task. Similarly, condition is shown to be a significant predictor for both congruent ( $\beta=-6.36, \mathrm{t}(114)$
$=2.50, \mathrm{p}=.014$ ) and incongruent ratings ( $\beta=12.6$, $\mathrm{t}(114)$ $=3.19, \mathrm{p}=.001$ ), with congruent ratings being higher and incongruent ratings lower in the concentrated condition. Finally, the interaction between concentration and volume was found to significantly differ between congruent and incongruent ratings ( $\beta=-2.21, \mathrm{t}(5086)=3.76, \mathrm{p}<.001$ ), potentially indicating differences in the level of the subtyping effect over the task.

This was investigated further using two additional mixed linear regression models for each trait type, both including the factors of condition and evidence volume. Coefficient estimates from the congruent ratings model suggested evidence volume to be a significant predictor $(\beta=-3.11, \mathrm{t}(1620)=6.63$, $\mathrm{p}<.001$ ), but concentration condition to be non-significant ( $\beta=-3.63, \mathrm{t}(114)=1.24, \mathrm{p}=.218$ ), with no significant interaction between these factors $(\beta=.58, \mathrm{t}(1620)=.88, \mathrm{p}=$ .379). Conversely, the incongruent ratings model suggested a significant effect of volume $(\beta=4.48, \mathrm{t}(1620)=7.79, \mathrm{p}<$ $.001)$ and condition $(\beta=7.40, \mathrm{t}(114)=2.40, \mathrm{p}=.018)$, with a near-significant interaction $(\beta=-1.52, \mathrm{t}(1620)=1.87, \mathrm{p}=$ .062 ). The findings of the general model are therefore most evident in the incongruent ratings when the two trait types are separated, while congruent ratings do not display such strong effects.

## Model Comparison

Participant data was compared with model predictions made by the three candidate models to determine which provided the most accurate depiction of behaviour in the task. This used a grid search function across the three parameters, with the considered values being: for $c, 0.01$ to 0.99 in steps of 0.01 ; for $\alpha, 0.1,0.2,0.3,0.5,1,2,3,5,10,15,20,25$ and 30 ; and for membership frequency of the prior cluster, 0 to 50 in steps of 1 . The models were run through the same exemplar sets given to participants at each combination of parameter


Figure 1: Trait ratings for both trait types in both concentration conditions across the 6 test blocks. Error bars show 95\% CI.
values to generate estimates of the probability of both congruency and incongruency in new category members at each of the six exemplar intervals. These values were then used to calculate model likelihoods assuming identical parameter values for all participants in order to allow the model to fit both conditions simultaneously. Likelihoods were calculated in each of the six test blocks according to the product of the probability of all participant ratings for that trait type in that test block; these probabilities were defined according to a normal distribution using the model probability estimate as a mean and variance fit to maximise the final product. These values were then transformed into log likelihoods before being summed across test blocks and concentration conditions to create a single model log likelihood for all participants at that set of parameter values. Maximum log likelihoods from each model were then used to calculate BIC values for comparison. The RMC was found to have the lowest BIC score (11926, $\alpha=0.5, \mathrm{c}=0.01$, prior membership $=21$ ), indicating this model had a better fit to the experimental data than either the RRMC (11937, $\alpha=10, \mathrm{c}=0.09$, prior membership $=50$ ) or the $\operatorname{SCM}(11929, \alpha=10$, prior membership $=50)$.

Interestingly, when the predictions for this best fit for the RMC are examined, probability estimates for both measures are in fact identical between conditions; this is because all experimental exemplars were assigned to separate, singlemember clusters despite any similarities in features, mitigating all exemplars equally in both sets. This suggests that the differences observed in participant ratings between concentration conditions were sufficiently small such that the data could be best fit by identical behaviour in both conditions. As previously described, this is in fact a tenet of the SCM, which ignores the concentration of data via full integration of all exemplars; however, the SCM shows a steeper curve in both measures compared to the RMC, therefore predicting greater stereotype revision. As such, the scattering behaviour of the RMC better corresponds with the greater degree of maintenance observed in the data.

It is also notable that the best fit of the RRMC matches that of the SCM, as the maximum likelihood of the RRMC was found when all exemplars were grouped in a single cluster.


Figure 2: Trait probability estimates from the best fits of the three candidate models. Due to equality in estimates between conditions for all models, only one line is used for each measure.

As such, the RRMC also predicts greater revision than was observed in the experiment; however, because the SCM does not use a coupling parameter, the SCM holds a lower BIC value than the RRMC despite equal log likelihoods. It should be noted however that this comparison reveals only the best fit of the three candidate models rather than an absolute description of behaviour in the task; more complex models may therefore be needed to reflect the subtle differences observed in the participant data.

## Discussion

The results of the experiment provide three key findings: firstly, ratings of trait likelihood for both congruent and incongruent traits became less stereotypical over the course of the experiment, indicating that higher volumes of incongruent evidence were effective in evoking greater revision of stereotypical beliefs. Secondly, ratings were more stereotypical in the concentrated condition compared to the dispersed condition, as would be expected by subtyping. Thirdly, this concentration effect differed somewhat in size across the task, showing smaller differences between groups at higher volumes of evidence. When the trait types are separated, these findings are seen to be stronger in the incongruent ratings, while congruent ratings did not demonstrate the condition or interaction effects.

In general, these results appear to partially correspond with previous depictions of the subtyping effect: beliefs are more stereotypical where incongruent information is more easily segregated from existing category structures, whereas data patterns aiding integration demonstrate greater stereotype revision. This also matches with the categorical explanations for subtyping offered by both the RMC and RRMC, as both suggest that category partitions which place incongruent data in a separate cluster diminish the impact of this data on subsequent probability estimates, thereby leading to more stereotypical expectations.
There is, however, an additional aspect to the subtyping effect observed in this task which distinguishes between these models: the interaction between volume and concentration, while not quite reaching a significant effect in the separated regression models, does suggest that the subtyping effect did not remain entirely consistent across the task, but in fact dropped off in later test blocks, with incongruent trait ratings in particular appearing to converge between the two conditions. This finding corresponds with the predictions of the RMC made in the introduction to this study: because the subtyping effect in the RMC is the result of greater uncertainty in the data pattern of the subtype cluster due to its smaller size, increasing the size of this cluster attenuates the subtyping effect by providing more confidence in this pattern. This is in contrast to the RRMC's hypothesised cross-over from subtyping to conversion at higher volumes of incongruent evidence where the subtype becomes the most likely cluster due to its size, an effect which was not observed in the data.
The RMC therefore appears to provide the most accurate
theoretical account of the observed results, suggesting participants were most likely using a rational categorisation process to guide their judgements in this task. This suggestion is further supported by the RMC having the best fit to the behavioural data in the above model comparison, though it is notable that this best fit did not accurately capture the observed differences between concentration conditions. Even so, the RMC does still provide a better fit to the experimental results than either the RRMC or SCM, potentially suggesting the present findings are more likely to be due to standard rational processes than these more extreme depictions.

## Implications

The present study therefore provides evidence from both behavioural data and model fits that the maintenance of stereotypical beliefs generated by subtyping does appear to be the result of a rational incorporation of all available data rather than a heuristic strategy of stereotype preservation: isolating incongruent data in a distinct subtype does not completely exclude this information from consideration during category judgements, but instead mediates its impact according to the size of the subtype.

As such, the subtyping effect could be considered to be a normal aspect of standard categorisation processes operating on social groups, occurring where a particular data pattern inadvertently diminishes the impact of counter-stereotypical data. If so, stereotype change could be encouraged by using similarly rational techniques to circumvent subtyping, primarily by aiding the integration of incongruent data into preexisting clusters. More broadly, this finding provides a basis for a rational system underlying stereotyping, allowing for the generation of further predictions regarding stereotype use based on the principles of such rational models to be tested in future studies; such tests would be valuable in further developing the current model to provide a more complete depiction of rational stereotype use.

The current data also demonstrates that stereotype change can be drawn from even slight encounters with incongruent evidence in sufficient volume, with the effects observed in the experiment being based solely on the observation of member summaries rather than any significant interaction with actual counter-stereotypical group members. This is in contrast to past theories such as the Contact Hypothesis which often require intensive, long-term interaction with out-group members to generate a reduction in stereotypical beliefs (Allport, 1954). This is not to say that prior expectations have been completely overcome: revision still does not reach the level suggested by conversion (or indeed the actual ratio of evidence in the experimental data sets); even so, this does still provide limited evidence that stereotype maintenance can be counteracted through increased exposure to incongruent data.

The current design may therefore present a more economic path to combating prejudice, requiring less time and effort than some existing methodologies. What is more, the effects observed in this study could in fact be greater at more significant forms of encounter, potentially counteracting subtyping
at even lower volumes of incongruent data. It is not clear how the significance of an encounter should be represented within the current version of the model, but one basic option would be to represent a significant encounter as multiple observations in the partition, essentially viewing that individual as providing more data than a single exemplar. This suggestion should, however, be pilot tested to determine the validity of this representation before being incorporated into the model.

## Conclusion

The present study provides the starting point for a rational approach to stereotype use, providing both theoretical and empirical evidence that a rational model of stereotype change, while not universally accurate, does provide a reasonable account of behaviour both in this experiment and previous studies into stereotype maintenance. We therefore hope that this study can act as a foundation for continued work in this field, allowing subsequent research to further refine the presented models to provide a more accurate depiction of behaviour. This will serve to provide greater clarity regarding the operations underlying stereotype maintenance, and so aid in finding more potential methods for encouraging stereotype change.

## References

Allport, G. W. (1954). The nature of prejudice. Cambridge, Mass: Addison-Wesley.
Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98, 409-429.
Bott, L., \& Murphy, G. L. (2007). Subtyping as a knowledge preservation strategy in category learning. Memory \& Cognition, 35, 432-443.
Heit, E., Briggs, J., \& Bott, L. (2004). Modeling the effects of prior knowledge on learning incongruent features of category members. Journal of Experimental Psychology: Learning, Memory and Cognition, 30, 1065-1081.
Hilton, J. L., \& von Hippel, W. (1996). Stereotypes. Annual Review of Psychology, 47, 237-271.
Johnston, L., \& Hewstone, M. (1992). Cognitive models of stereotype change: 3 . subtpying and the perceived typicality of disconfirming group members. Journal of Experimental Social Psychology, 28, 360-386.
Murphy, G. L., \& Ross, B. H. (1994). Predictions from uncertain categorizations. Cognitive Psychology, 27, 148-193.
Weber, R., \& Crocker, J. (1983). Cognitive processes in the revsion of stereotypic beliefs. Journal of Personality and Social Psychology, 45, 961-977.

# Population size, learning, and innovation determine linguistic complexity 

Matthew Spike (matthew.spike@anu.edu.au)<br>Centre of Excellence for the Dynamics of Language<br>Coombs Building, The Australian National University, ACT 2601, Australia


#### Abstract

There are a number of claims regarding why linguistic complexity varies, for example: i) different types of societal structure (e.g. Wray \& Grace, 2007), ii) population size (e.g. Lupyan \& Dale, 2010), and iii) the proportion of child vs. adult learners (e.g. Trudgill, 2011). This simple model of interacting agents, capable of learning and innovation, partially supports all these accounts. However, several subtle points arise. Firstly, differences in the capacity or opportunity to learn determine how much complexity can remain stable. Secondly, small populations are susceptible to large amounts of drift and subsequent loss, unless innovation is frequent. Conversely, large populations remain resilient to change unless there is too much innovation, which leads to a collapse in complexity. Next, if adult learners are prevalent, we can instead expect less sustained complexity in large populations. Finally, creolisation does not imply simplification in smaller populations.


Keywords: linguistic complexity; language variation; innovation; social networks; agent-based models; cultural evolution.

## Introduction

Languages vary in complexity. This was a controversial idea for much of the last century, but a growing body of empirical evidence has led to a new consensus in its favour (see Joseph \& Newmeyer, 2012). More intriguingly, the most complex languages in the world are often the ones with the least speakers, spoken by remote, inaccessible, and sometimes non-literate societies. The Archi language, for example, "spoken by a thousand people in one village 2,300 metres above sea level in the Caucasus" (Nichols, 2009, p.3), features verbs with around 1.5 million inflected forms. At the other end of the spectrum, some languages are notable for their apparent simplicity; often creoles (e.g. McWhorter, 2001), but not exclusively so (e.g. Gil, 2001).

There are several lines of thought regarding the origin of this variation in complexity. Trudgill (2011) proposes that when a language community includes a large proportion of adult second-language learners, it leads to a corresponding reduction in that language's complexity, but that, when 'left alone', languages tend towards greater complexity: i.e. there is a directionality to such language change. In a somewhat related idea, Wray \& Grace (2007) argue that esoteric societies (where intra-group communication dominates) lead to further complexification, while simplification occurs in exoteric societies (where people frequently interact with strangers). Nettle (2012) indicates a link between population size and grammatical complexity, citing empirical support from Lupyan \& Dale (2010), who found a (negative) correlation between population size and morphological complexity: similarly to Trudgill, they argue that complex features of language undergo negative selection in large populations with many second language learners, but further conjecture that the high morpho-
logical complexity found in languages spoken by small communities assists in child language acquisition. Finally, authors such as McWhorter (2001) point to the youthfulness of creole languages as the explanation for their simplicity: they haven't been around long enough to build up the diachronic "ornamentation" found in older, more complex languages.

These claims require empirical validation. However, it is notable that despite the increasing availability of crosslinguistic documentation (e.g. WALS, Dryer \& Haspelmath, 2013), no uncontroversial, universally applicable measure of linguistic complexity has arisen. Information-theoretic measures of complexity (e.g. Juola, 2008) can be hard to interpret (the various dimensions of complexity, such as the size of the lexicon and segmental inventory, and paradigmatic vs. syntagmatic complexity are conflated in such measures, and furthermore do not distinguish between descriptive complexity and structural complexity, see Crutchfield, 1994). The alternative would be to employ traditional linguistic analysis, but as pointed out by Nichols (2009, p.111), "measuring the total complexity of a language in cross-linguistically comparable and quantifiable terms would be a massive task and unreasonably costly in time and effort", and moreover any such result would be theory-dependent, and as such subject to accusations of false equivalence (e.g. Haspelmath, 2010) and subjectivity (e.g. Martin, 2011).

As an alternative to empirical analysis, formal tools would seem a good way of - at the very least - assessing the internal consistency of the claims in question. Indeed, two such models have been produced, the first by Lupyan \& Dale (2010) and the second by Reali et al. (2014). Lupyan \& Dale argue that population size correlates with the proportion of L2 learners, and their model suggests that it is this which reduces complexity; Reali et al. show a more direct effect of population size. The model presented here represents an attempt to synthesize and extend these results in a more general format. Results suggest that three factors determine the complexity of a language. Firstly, a population's effective size. Secondly, the amount of linguistic regularisation: this can be determined by a number of factors, including the number of learning experiences, the memory limitations of individual agents, and any cognitive bias for regularity. Finally, linguistic innovation is crucial, as the same amount of innovation can sometimes support greater complexity, while at other times leads to a collapse in complexity, depending on the size and nature of the population.

## Previous Models

Lupyan \& Dale (2010) present a mathematical model in their supplementary materials which is analysed in terms of the
evolutionary fitness of languages depending on the proportion of L1/L2 learners. They find that, under various assumptions, a high proportion of L2 learners implies that simple languages are maximally fit. However, neither interaction nor social structure are taken into consideration.

Reali et al. (2014) explicitly investigate population size in a model where agent interactions are governed by Gilbert random graphs. Agents produce token-like conventions which can be either easy or hard. Crucially, easy tokens can be reproduced by another agent after a single exposure, while hard tokens require two exposures. Finally, new conventions are occasionally produced according to a Chinese restaurant process, and agents have a hard limit on the number of tokens they can store, i.e. a limited memory. The finding is that, in smaller groups, significantly more hard tokens are able to establish themselves across the entire population than is the case with larger groups. Reali et al. suggest that language, and indeed all culture, might become preferentially simpler as societies increase in size and social connectivity.

These models support two of the hypotheses found in the literature: both the type of language learner (child or adult) and the population size are arguably factors behind the variation found in linguistic complexity. However, we are left with a number of questions: 1) Can we reconcile these two predictions; 2) How do we incorporate ideas such as that of Wray \& Grace (2007) about esoteric and exoteric cultures, and McWhorter (2001) regarding creoles; 3) How robust are the previous models across different parameter settings and instantiations? To investigate these questions, we need to systematically vary not just the population size, but also i) the type of social network, ii) the amount of linguistic regularisation, iii) the amount of linguistic innovation, iv) the initial state of the population, and $v$ ) whether intergenerational language acquisition is included or not. This is the target of the model presented here.

## Model description

Agents produce and store tokens representing conventions, but there is no distinction between different types of token, e.g. easy vs. hard. Instead, the complexity of a conventional system is assessed by counting the number of populationwide shared types. This casts the complexity of a given population's language in terms of the total amount of information required to acquire that shared system, abstracting away from the details of how that system is stored, used, or acquired. A complex language, then, is when all agents share a large number of conventional types, while a very simple language is when almost no conventional types are shared throughout the population. Note that this does not imply that individual agents do not store a large number of types, or even that many conventions are not shared by sub-populations. Another way this might be conceptualised in terms of Reali et al.'s 2014 model is that this model deals only in hard-to-learn conventions, while easily-learned conventions are simply assumed to be learnt independently, in a way which does not interfere
with hard ones.
Conventions $c_{j}$ are drawn from an infinite set $C=$ $\left\{c_{1}, \ldots, c_{n}, \ldots\right\}$. There is no distinction between different types of convention, for example easy or difficult, and all are equally weighted as $W_{c}=1$. There are $n$ agents $a_{i} \in$ $A=\left\{a_{1}, \ldots, a_{n}\right\}$, which are modelled as variants on Hoppe urns (Hoppe, 1984), after the models of innovative signalling found in Skyrms (2010, see p.124), and similar to the learning agents described by Reali et al. (2014). Depending on the starting condition, an agent is initially composed of $t$ convention tokens, where $t \geq 0$, and a single 'innovation token' with weight $W_{v} \geq 0$. The number of tokens of convention type $c_{j}$ possessed by agent $a_{i}$ is denoted $N_{i j}$.

The initial state of the population is either homogeneous, sampled, or heterogeneous. Homogeneous populations consist of agents with exactly the same 50 types of token. Sampled populations initiate by sampling from a set of 100 initial tokens, meaning that initially no type is likely to be found in every individual in larger populations. Finally, heterogeneous populations consist of agents with entirely different initial sets of tokens.

When an agent $a_{i}$ 'speaks', it selects a particular convention type $c_{j}$ with probability $P_{i j}$ given by:

$$
\begin{equation*}
P_{i j}=\frac{N_{i j}}{W_{v}+\sum_{k \in C} N_{i k}} \tag{1}
\end{equation*}
$$

Alternatively, the agent may produce an entirely new convention, with probability $P_{i v}=1-\sum_{k \in C} P_{i k}=\frac{W_{v}}{W_{v}+\sum_{k \in C} N_{i k}}$. If $x$ conventions have been created by the population to date, the new convention is denoted $c_{x+1}$.

An interaction between two agents is simple: one is denoted 'sender', and another 'receiver'. The speaker chooses a convention according to the distributions given above, and the receiver adds exactly one new token of that type. When the learning capacity is cast as a memory limit, each agent has a hard limit of $m$ tokens: if the number of stored tokens exceeds $m$, then one of the tokens is selected for deletion with a probability proportional to $N_{i j}$, but excluding the innovation token (which is never selected for deletion). Put another way, conventional types which are more strongly represented via their association with more memory tokens are correspondingly more likely to be selected for deletion, and vice versa.

Population structure is defined by the non-directed graph $G$. Three types of graph structure are investigated: 1) Fullyconnected graphs, in which every agent node connects with every other, 2) Erdős-Rényi random graphs $G(n, p)$, generated by assigning a probability $p=0.4$ that any agent node connects with another, and finally 3) Newman-Watts-Strogatz small-world graphs $G(n, k=2, p=0.4)$ : agents are first connected in a ring-structure, then to each neighbour two nodes away, and then to another randomly-selected node with probability $p$. Small-world networks capture the property of reallife social networks in that while any one person may not be connected to many others, the number of nodes which must be traversed between any two people is typically small, e.g.


Figure 1: Results are robust across many individual simulations. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different populations sizes (5, 25, and 100: line colours) and network structures (fully-connected, random and small-world: line dashes), with 'sampled' initial states, learning capacity of 100 , and an innovation rate of 0.1 . Error bars represent 1 standard deviation from the mean for individual simulations. Also note that network structure has no apparent long-term effect.
the concept of 'six degrees of separation'.
Interactions proceed by selecting, with uniform probability, an agent to be sender. The receiver is then chosen from the set of agents which connect to the sender, also with uniform probability. The agents interact, and the simulation continues by reiterating the process.

Population turnover, when instantiated, is 'gradual': it proceeds by choosing an agent at random and replacing them with a new agent, who then is exposed to a given number of tokens from connected agents, representing the number of learning experiences. In this way, fewer learning experiences are taken to represent more adult-like learners, and more experiences to be child-like.

As outlined before, the method of analysis is to count the number of population-wide shared types.

## Results

The parameters adjusted in relation to each other were i) population size: 5, 25 or 100 agents; ii) population structure: fully-connected vs. random vs. small-world; iii) population dynamic: static vs. gradual turnover; iv) initial composition: homogeneous vs. sampled vs. heterogeneous. v) learning capacity: 100, 500, or 1000 tokens; vi) innovation rate: $W_{V}=0.1,1,10$, or 100 . The main results are as follows:

## 1. Long-term complexity is robustly determined.



Figure 2: Both population size and learning capacity determine stability. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different population sizes (5, 25 and 100: line colours) and learning capacities (100, 500, and 1000: line dashes), with small-world networks, homogeneous initial conditions and an innovation rate of 0.1 . Note that a small learning capacity always leads to a collapse in complexity, while even a large learning capacity is unable to prevent drift and loss in small populations.

Although simulations were stochastic, results were robust as regards long-term complexity. That is to say, the type of population (as determined by the parameters above) reliably determines a stable level of complexity which is robust across i) individual simulations and ii) time: see Figure 1 . This level of complexity is determined by multiple factors (which are outlined shortly), but the existence of a 'steady state' (which may take some time to reach) is important. Differently understood, this means that (given our assumptions) complexity will not remain in constant flux unless some new factor comes into play, e.g. a change in population size.

## 2. Learning capacity and population size determine stability.

Across all conditions, the learning capacity of individual agents determines how complex the population-wide language can be. When memory or learning experiences are limited in number, the effect of linguistic drift increases: see Figure 2. This leads to certain variants being lost and a decrease in complexity. Population size plays a similar role, for as the number of individuals increases, the less of an effect drift can play. In essence, either the individual or the population must act as a 'reservoir' to avoid loss. In the case of individuals, this requires a large memory and/or many instances of learning; for populations, a


Figure 3: Innovation can maintain, increase or decrease complexity. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different population sizes (5,25 and 100: line colours) and innovation rates (1, 10 and 100: line dashes) with fully-connected networks, homogeneous initial state and an learning capacity of 1000 . Note that high levels of innovation lead to very high levels of complexity in small populations, but to a collapse in complexity in larger populations.
smaller learning capacity is required because individual tokens will likely be shared across many individuals and are thus robust to loss in any one individual. However, when learning is not sufficient, complexity will collapse even in large populations.
3. Innovation can maintain, increase, and decrease complexity depending on population size.
For smaller populations, only high rates of innovation can counteract linguistic drift. When they do, however, this can push levels of complexity much higher than would be possible for adult learners with similar learning capacities: see Figure 3. Low levels of innovation lead to catastrophic collapses in complexity for small populations, even when learning capacities are high. Contrasting with this, large populations - which easily maintain a given level of complexity - are overwhelmed by large amounts of innovation: in this case, too much innovation leads to less overall complexity.

## 4. Adult learners reduce complexity

When we include gradual population turnover, decreasing the number of learning exposure leads to decreased complexity: see Figure 4. The rate of innovation is less important, as we see different rates of innovation pattern together. However, learning capacity is more important than


Figure 4: Intergenerational learning and innovation in large populations. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 5000 replacements with 1000 learning interactions each for populations of 100 agents with gradual turnover and different numbers of learning exposures ( 100,500 , and 1000: line colours) and rates of innovation ( 0.1 , and 1 : line dashes), with small-world networks and a heterogeneous initial state of 50 tokens. This shows that complexity is less stable in large populations of learners than is the case with interacting populations.
in static populations: when learning exposures are anything else than quite high, we can expect a decrease in complexity. As such, the maintenance of high levels of complexity requires child-like learners.

## 5. Creoles: complexity in small populations, simplicity in large populations

When a common language already exists, the level of complexity will either remain stable, or will be affected by the factors mentioned above: see Figure 5. On the other hand, when there is no common language, such as with the extreme state of interpersonal variation modelled by the 'heterogeneous' parameter, we see an interesting effect. When initial populations are large, these mixed societies never develop systems of any complexity. However, small groups with a similar composition lead to very high levels of complexity.

## 6. Social network structure has little effect:

Social network structure has a relatively small role to play in the development and maintenance of linguistic complexity. As long as networks have a small-world property, i.e. as long as the average path-length between any two people remains small (which is the case in all of the network types surveyed here), diffusion across the network is sufficiently


Figure 5: Creolisation does not necessarily imply simplicity. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different population sizes (5, 25, and 100: line colours) and innovation rates ( 0.1 and 1 : line dashes), with a heterogeneous initial state, small-world networks, and a learning capacity of 1000 tokens. When populations sizes are large, no complexity develops, but when population sizes are smaller then complexity is able to fixate.
large to ensure that the other results presented here remain valid.

## Analysis

Long-term complexity is reasonably deterministic given a set of assumptions about population size and structure, the rate of innovation and so on. All things remaining equal, then, population size and the nature of learning and innovation should have a predictable impact on linguistic complexity. On the other hand, it is worth noting that real-world populations are unlikely to remain static in regards to many of these assumptions: population sizes will rise and fall, societal pressures driving innovation will vary, and the nature of cultural integration between different social and linguistic groupings can drastically change over short periods of time. In the absence of more detailed case-specific analysis, however, these results should add weight to the theories discussed in the introduction.

Next, we can consider these findings in the light of wellestablished results from population genetics (e.g. the WrightFisher and Moran models of genetic drift) which show that i) small populations are highly susceptible to loss via drift while large populations are conservative, and that ii) fixation of new variants is much more likely in small populations than large ones. Taking these in turn:

1. The susceptibility of small populations to drift is in line
with the results which predict that maintaining high levels of complexity in small populations requires large amounts of innovation. Bromham et al. (2015), also citing the parallels between language change and evolutionary models, show that there is significantly more frequent word loss in smaller populations, so it seems reasonable to expect a similar process to occur at other levels of linguistic structure besides the lexicon. Perhaps a more pressing concern is that the model presented here is equivalent to a 'neutral model' of evolution. This runs against assumptions which are sometimes made in the literature regarding the directionality of linguistic complexification. Trudgill (2011) challenges previous assumptions that simplification is the natural direction of language change, arguing instead that when "left alone"(p.325), languages will gradually complexify, and that only external pressures such as a large proportion of second-language learners will lead to reduced complexity. This can be analysed in two ways: either that humans have something akin to a cognitive anti-regularisation bias which prevents drift-like processes from occurring, or that Trudgill simply perceives the natural state of linguistic development to take place in small groups with child learners. If the former, then recent work suggests that the opposite is the case: Ferdinand et al. (2013) identify a linguistic domain-specific bias in favour of regularisation. If the latter, then the model here corroborates with Trudgill's theories only if we can assume that the rate of innovation is very high.
2. Large populations are resistant to fixation or new variants, just as they are to the establishment of complexity. There are two factors behind this: firstly, when innovation rates are low, the probability of any new variant fixating within the population becomes very small. On the other hand, when there is too much innovation we see a collapse in overall complexity. This is in line with empirical results such as Lupyan \& Dale (2010), but the explanation differs. They argue that adult learners reduce complexity and child learners foster it: on the contrary, it appears that any more than an extremely sparse sampling by adult learners suffices to preserve population-wide complexity, due to the 'reservoir' like effect that large populations have. This, then, supports Trudgill (2011), but acts to constrain his theory: not just adult learners are necessary, but adult learners with extremely restricted exposure or learning capacities. The other condition in which we can expect adult learners to drive simplification is when they also contribute large amounts of innovation: this is an unexpected result, and is in need of empirical validation.

The results for large populations which tend towards either stability (when learning capacity is medium or high), or simplification (when learning capacity is very low), assume a static population where most change and innovation takes place in individual interactions. However, change and innovation also occur intergenerationally. Whether one or both of
these factors predominate had been a subject of perennial debate, but the results here make a solid prediction about what to expect if either is the case. That is, if interaction is at least one of the main factors, we should expect very little in the way of increasing complexification. If intergenerational change is the main factor, however, we should expect large populations of anything else than child learners to lead to dramatic simplification; if not, then we should expect simplification only when most learners have extremely sparse input. Whether this is or is not the case is a target for future empirical work.

Finally, the results indicate that creoles can attain complexity given reasonably small population sizes. In fact, this stands to reason given the previous results: given an initial pool of extremely wide variation, many variants are able to fixate in small populations, but very few to none in large populations. The take-home message from this is not that we should expect complexity in small mixed populations - as the assumptions made by this configuration of the model are particularly unlikely - but rather that we cannot assume that creolisation should automatically entail simplicity: we can expect it to appear under some circumstances.

## Conclusion

The relationship between linguistic complexity and social determinants is more nuanced than has been sometimes been assumed. At the very least, we need to consider not just the effective size of the population in question, but also give some thought to how learning proceeds - whether this is in terms of memory or learning exposures - and the nature of linguistic innovation. However, as previously observed, all of these factors can be difficult to accurately observe and/or measure, and undergo constant flux. In particular, linguistic innovation can be subject to a myriad of intrapersonal, interpersonal or larger cultural pressures and variations. Furthermore, the results presented here are from a highly idealised model of cultural learning and transmission: it may well be the case that including more detailed and realistic mechanisms, particularly as pertains to human language, will impact on some of the conclusions presented here. Even if this is the case, the model allows us to both draw several disparate theoretical claims together, while at the same time sharpening the predictions we can make regarding how social structure, population size, and the details of learning and innovation should impact linguistic complexity.

## References

Bromham, L., Hua, X., Fitzpatrick, T. G., \& Greenhill, S. J. (2015). Rate of language evolution is affected by population size. Proceedings of the National Academy of Sciences, 112(7), 201419704.
Crutchfield, J. P. (1994). The calculi of emergence: computation, dynamics and induction (Vol. 75) (No. 1-3).
Dryer, M., \& Haspelmath, M. (Eds.). (2013). The World Atlas of Language Structures. Leipzig: Max Planck Institute for Evolutionary Anthropology..

Ferdinand, V., Thompson, B., Kirby, S., \& Smith, K. (2013). Regularization behavior in a non-linguistic domain. University Proceedings of the 35th Annual Cognitive Science Society, 436-441.
Gil, D. (2001). Creoles, Complexity and Riau Indonesian. Linguistic Typology, 5, 325-371.
Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. Language, 86(3), 663-687.
Hoppe, F. (1984). Pólya-like urns and the Ewens' sampling formula. Journal of Mathematical Biology(20), 91-94.
Joseph, J. E., \& Newmeyer, F. J. (2012). All Languages Are Equally Complex' The rise and fall of a consensus*. Historiographica Linguistica, 39(2-3), 341-368.
Juola, P. (2008). Assessing linguistic complexity. In Language complexity: Typology, contact, change (pp. 89108).

Lupyan, G., \& Dale, R. (2010, jan). Language structure is partly determined by social structure. PloS one, 5(1), e8559.
Martin, F. d. P. (2011). The Mirage of Morphological Complexity. Proceedings of Cognitive Science Society Conference., 3524-3529.
McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. Linguistic Typology, 5(2), 125-166.
Nettle, D. (2012). Social scale and structural complexity in human languages. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1597), 1829-1836.
Nichols, J. (2009). Linguistic complexity: a comprehensive definition and survey. In G. Sampson, D. Gil, \& P. Trudgill (Eds.), Language complexity as an evolving variable (pp. 110-125). Oxford: Oxford University Press.
Reali, F., Chater, N., \& Christiansen, M. H. (2014). The paradox of linguistic complexity and community size. In E. Cartmill, S. Roberts, H. Lyn, \& H. Cornish (Eds.), The evolution of language - proceedings of the 10th international conference (pp. 270-277). Singapore.
Skyrms, B. (2010). Signals: Evolution, Learning \& Information. Oxford: Oxford University Press.
Trudgill, P. (2011). Sociolinguistic typology: Social determinants of linguistic complexity. Oxford University Press.
Wray, A., \& Grace, G. (2007, mar). The consequences of talking to strangers: Evolutionary corollaries of sociocultural influences on linguistic form. Lingua, 117(3), 543-578.

# A rational analysis of marketing strategies 

Nisheeth Srivastava (nsrivast@cse.iitk.ac.in)<br>Department of CSE, IIT Kanpur<br>Kalyanpur, UP 208016 India

Edward Vul (evul@ucsd.edu)<br>Department of Psychology, 9500 Gilman Drive<br>La Jolla, CA 92093 USA


#### Abstract

Rational accounts of decision-making are incompatible with the prevalence and success of ubiquitous marketing strategies. In this paper, we demonstrate, using computational experiments, how an ideal Bayesian observer model of preference learning is compatible with the manipulation of purchasing decisions via a number of well-known marketing techniques. The ability of this model to predict the effects of both familiar and novel marketing interventions suggests it as a plausible candidate theory of consumer marketing. Simultaneously, by clarifying the logic underneath the interplay between environmental exposure and preference distortions seen in economic decisions, this model rationalizes the seemingly irrational susceptibility of consumers to marketing.


Keywords: decision-making; preference learning; advertrising; marketing; rational analysis

## Introduction

Marketing constitutes a genre of economic activity that is mysterious to existing formal accounts of consumers' decision process. While such formal theories require consumers to be economically rational, doing so would make them impervious to marketing techniques. In fact, the very existence of marketing as a viable genre of activity violates the predictions of current formal accounts of consumer behavior. What possible new information can the $43^{r d}$ viewing of an insurance company's ad give a consumer? Clearly, consumers receive a lot more information about products than just their 'utility' through such repeated interactions. Such associative influences have been difficult to document and incorporate into formal theorizing - hence have historically been ignored in marketing research - save as unspecified exogenous influences to be parametrized in econometric analyses." Now that online activity can be meticulously logged across content platforms, we argue that the sort of side information that was treated as noise in earlier generations of marketing theories can be incorporated to construct computational models that can make testable predictions about the efficacy of marketing interventions. This is the goal we pursue in this paper.

We do so by developing a psychological model of preference formation that can quantitatively relate manipulations of marketing variables to consumer demand. Our approach diverges from existing accounts of consumer/buyer behavior in several key aspects. First, unlike classic (Belk, 1975) and modern (Malhotra, 1988) integrated models of consumer behavior, our model can offer constrained quantitative predictions by virtue of relying on only observable variables (such as price distributions, exposure frequency, and transaction history), rather than relying on unobservable and immeasurable consumer valuations and beliefs. Second, unlike current quantitative models of consumer psychology behavior that consider choice mechanisms in very narrow settings
(e.g., reference price models; (Winer, 1986)), we attempt to provide a general account that can capture the effects of many marketing interventions. Finally, unlike classical approaches designed for brick-and-mortar retail that have relied on population-level market and consumption variables, our approach considers frequency distributions of individual observers' transactions, which are increasingly more measurable and relevant in internet commerce.

## Existing models of consumer behavior

In the absence of formal theory, existing quantitative models of marketing are primarily econometric - they regress multiple available variables against outcomes of interest, use focus groups or deductive arguments to suggest that such variables can be changed by particular marketing interventions, then extrapolate these changes to the consumer base to predict how much the underlying outcomes will change.

Econometric models of marketing interventions are fundamentally data analytic models that impose microeconomic constraints on estimated parameters. Thus while they are good at retrospectively estimating the effects of marketing interventions on demand curves, they can only make predictions about such effects by extrapolating parameters. Perhaps the most rigorous models of consumer behavior are in the domain of pricing. Price sensitivity has been shown to follow a Weber law, such that consumers are sensitive to proportional price changes (Monroe, 1973). Moreover, consumers seem to evaluate prices relative to a "reference"" price range that varies across products and categories (Kalyanaram \& Winer, 1995), appears to be learned from transaction history (Emery, 1969), and may be influenced by brand strength (Biswas, 1992). These models are typically used to explain and motivate narrow experimental manipulations, and while they hold promise for predicting changes in aggregate demand curves from transaction history, they have not been applied in this way; perhaps largely due to the fact that they do not integrate the effects of long-term marketing strategies. Although these models can capture the effects of long-term marketing strategies on demand curves via free parameters to account for changes in reference price with branding, advertising, etc., they do not offer a predictive account of how marketing actions will influence the reference price, and thus can only retrospectively describe their effects on demand curves.

In contrast, theories of consumer behavior that aim to explain the psychological mechanisms of a broad range of marketing interventions rely on qualitative, verbal accounts of psychological processes and invoke unobservable, and immeasurable, latent traits and beliefs of consumers (e.g., (Bettman, 1979)). While these theories offer pithy qual-
itative summaries of marketing researchers' intuitions about the psychology of consumers, they are neither designed to, nor capable of, offering quantitative predictions.

The account we present in this paper aims to capitalize on the strengths of these different approaches: First, by constructing a model at the level of individual choices, we capture intuitions about the psychology underlying consumer behavior. Second, by basing the individual choice on historically observed price distributions and transactions, we accommodate known relative price-range effects. Finally, by relying on only externally observable quantities as the inputs to the individual choice model, we make our theory empirically identifiable in the same manner as standard econometric models.

## A cognitive model of consumer psychology

Ultimately, population demand curves are created from aggregating individuals' buy/don't buy decisions, therefore any formal analysis of the efficacy of marketing techniques must model how they influence individual purchasing choices. Framed this way, the question such an analysis must ask is 'how do prices and other market signals influence purchase decision?'

The standard way of addressing this question is to treat choices as the outcomes of utility maximization. On this view, whatever choices an observer makes can be attributed to some underlying hedonic calculation which shows a higher evaluation for the chosen option. While this is a mathematically elegant way of describing the choice outcome, it is a very poor description of the process underpinning these choices. Prior research has demonstrated that consumers' price estimates of products tend to be drastically altered by presentation formats (Tversky \& Kahneman, 1981), the set of available options (Huber \& Puto, 1983), as well as a variety of seemingly irrelevant psychological primes (Ariely, Loewenstein, \& Prelec, 2006).

The success of various marketing strategies in increasing consumer preference for the same underlying product (Kirmani \& Rao, 2000) lends credence to a less optimistic view of consumer preferences: choices are based on dynamic, context-dependent comparisons between options, rather than reliable hedonic value judgments (Ariely et al., 2006). Consumers are likely to make any particular decision by drawing upon past experiences with choices among similar options (Gilboa \& Schmeidler, 1995). Given variability in experiences, variability in recall, and variability in the comparison process used to generate preferences, the resulting preferences will be considerably uncertain.Our theory is that marketing strategies capitalize on this uncertainty by manipulating the information available to observers at intermediate steps of the preference-construction process to influence preferences.

The principal contribution of our work is demonstrating how Bayes-optimal combination of prior choice-relevant observations yields an interpretable, simple, testable, and parsi-
monious account of marketing psychology. In particular this account predicts the efficacy of a number of interesting marketing strategies on several important consumer choice outcomes by virtue of their influence on a small, factored representation of consumer price history and knowledge.

## Consumer representation

What are the observable building blocks of a theory of marketing psychology? An intuitive simplification of a typical economic transaction is that a buyer decides that the price for a particular product is fair in a particular context. Thus the observable units of individual transactions are consumers, prices, products, choices, and auxiliary contextual information (e.g. physical location, web portal, company brand, etc). Of these units marketers cannot directly influence consumers' choices ( $b$ ), but they can affect prices ( $m$ ), products $(x)$, contexts $(c)$, and critically, the frequency and cooccurrence statistics with which consumers encounter each.

Although the full set of experiences of an observer can be described as a joint distribution of $p(b, c, x, m)$, there are several reasons to consider the agent's representation not as this complete joint distribution, but instead a factored set of several conditional distributions. First, it seems implausible for humans to keep track of the full joint distribution given the extreme sparsity of observations therein. Second, an argument from introspection suggests that not all conditional probabilities are equally easy to access as we would expect if they were all calculated form the same joint distribution: e.g., $p(m \mid x, c)$ (how much does yogurt cost at Safeway?) seems intuitive while $p(x \mid m, c)$ (what costs $\$ 5$ at Safeway?) seems to require an awkward explicit search. Third, by factoring the joint distribution, a consumer can learn about the distribution of goods and prices from observing the transactions of others independently of tracking her own choices. Finally, a fourth, practical, reason to factor the joint distribution in a consumer choice model is that it makes it usable for predicting consumption behavior; whereas a model based on the full joint distribution would be inestimable to marketers who do not have access to the full set of experiences of a particular consumer.

Thus, to retain psychological plausibility, and practical usability, we assume that individuals represent the important elements of only some conditional and marginal probabilities from the joint distribution of purchasing decisions, products, prices, and contexts. Specifically, we assume consumers learn the following distributions from observations of the world around them:

- $p(c)$ - what contexts populate a consumer's daily life?
- $p(x \mid c)$ - what products are available in this context?
- $p(m \mid x, c)$ - how much does this product cost in this context?

And from their own experience, they keep track of:

- $p(b \mid m, x, c)$ - how often do I purchase a particular good in a particular context, at a particular price?


## Consumer choices

These tracked conditional distributions can be combined via the rules of probability to estimate the joint distribution over consumption choices, products, prices and contexts; and thus any conditional distribution of interest. Of particular interest in our case are the conditional distributions that observers must use to make consumer decisions:

- $p(b \mid x)$ - do I want to buy product $x$ ? (preference)
- $p(m \mid b, x)$ - what price am I willing to pay to buy $x$ ? (valuation)
- $p(b \mid m, x)$ - how does willingness to buy change with price? (demand curve)
- $p(b \mid c)$ - will I make a purchase in a given context?
- $p(c \mid b, x)$ - if I am going to buy $x$, in which context will I do so? (brand/retailer selection)

Each of these distributions capturing key aspects of consumer behavior can be predicted by marginalizing and conditioning the joint distribution obtained via $p(b, m, x, c)=$ $p(b \mid m, x, c) p(m \mid x, c) p(x \mid c) p(c)$.

On our account, consumers determine their propensity for buying particular goods using accumulated evidence of previous purchases:

$$
\begin{equation*}
p(b \mid \hat{x})=\frac{\sum_{c, m} p(b \mid m, \hat{x}, c) p(m \mid \hat{x}, c) p(\hat{x} \mid c) p(c)}{\sum_{c, m} p(m \mid \hat{x}, c) p(\hat{x} \mid c) p(c)} \tag{1}
\end{equation*}
$$

What is more interesting to a firm, though, is finding the greatest price a consumer would be willing to pay to purchase a product. Prior research has suggested that people typically generate a range of prices that they would be willing to pay for a product (Mazumdar, Raj, \& Sinha, 2005). We formalize this intuition by casting this as a distribution over possible prices,

$$
\begin{equation*}
p(m \mid b=1, \hat{x})=\frac{\sum_{c} p(b=1 \mid m, \hat{x}, c) p(m \mid \hat{x}, c) p(\hat{x} \mid c) p(c)}{\sum_{c, m} p(b=1 \mid m, \hat{x}, c) p(m \mid \hat{x}, c) p(\hat{x} \mid c) p(c)} \tag{2}
\end{equation*}
$$

which directly gives us the distribution of prices at which consumers are willing to purchase a good.

With only a slight reformulation, this yields the relationship needed to obtain classical demand curves: purchase propensity as a function of price:
$p\left(b \mid m=m_{a}, \hat{x}\right)=\frac{\sum_{c} p\left(b \mid m=m_{a}, \hat{x}, c\right) p\left(m=m_{a} \mid \hat{x}, c\right) p(\hat{x} \mid c) p(c)}{\sum_{c} p\left(m=m_{a} \mid \hat{x}, c\right) p(\hat{x} \mid c) p(c)}$,
Of particular interest to a retailer, is the propensity of consumers to purchase while in their store,

$$
\begin{equation*}
p(b \mid \hat{c})=\frac{\sum_{x, m} p(b \mid m, x, \hat{c}) p(m \mid x, \hat{c}) p(x \mid \hat{c}) p(\hat{c})}{\sum_{x, m} p(m \mid x, \hat{c}) p(x \mid \hat{c}) p(\hat{c})} \tag{4}
\end{equation*}
$$

Finally, brands and retailers alike are interested in the likelihood that a consumer will choose their store or brand when making a purchase of a particular product:

$$
\begin{equation*}
p(c \mid b=1, \hat{x})=\frac{\sum_{m} p(b=1 \mid m, \hat{x}, c) p(m \mid \hat{x}, c) p(\hat{x} \mid c) p(c)}{\sum_{c, m} p(b=1 \mid m, \hat{x}, c) p(m \mid \hat{x}, c) p(\hat{x} \mid c) p(c)} \tag{5}
\end{equation*}
$$

Critically, each of these key facets of consumer choice and behavior will change in predictable ways under various marketing interventions designed to alter the conditional distributions that consumers keep track of. Thus, this formal setup, while sparse, allows us to test the influence of manipulating prices and context information on consumer demand curves.

## Model predictions

To substantiate our intuitions about marketing-based distortions of consumer preferences, we simulated a small test market, containing three purchase contexts, two goods, and five price labels where a consumer's purchases were generated via the following generative model. A purchase context was sampled from a random seed distribution $p(c)$, a product was sampled from a discrete random seed probability $p(x \mid c)$ for this context, a price label was sampled from a random seed probability $p(m \mid x, c)$ for the already sampled tuple $\{x, c\}$. Finally, this observation was flagged as a purchase decision with a small probability $(p=0.2)$, and within the samples thus flagged, purchase decisions were randomly generated while maintaining an inverse relationship with price.

Using this generative procedure, we sampled 10000 events to obtain baseline empirical estimates for each of the conditional distributions implicated in our account. The experimental results we report in succeeding sections were constructed by appending this baseline event history with manipulated event sequences corresponding to various marketing interventions.

## Rationalizing product-brand associations

The most obvious form of marketing is advertising by displaying the product and its associated brand. This form of advertising could be rationalized as providing information to potential consumers. It is harder to make a similar argument for event sponsorships and brand awareness campaigns, wherein companies advertise only brands, not products. What rational purpose is served by simply presenting the company's logo to a consumer, disconnected from product information? Also, why belabor people with redundant and uninformative visuals over and over again? Surely once or a couple of occasions would be enough to convey any information? Why are "tip of tongue" (Mowen \& Gaeth, 1992) and brand recognition metrics (Munoz \& Kumar, 2004) so popular, influential, and desirable? The answer, of course, is that firms aim to increase the rate at which consumers think of their brand. But why would increasing the ease with which consumers think about the brand change consumer purchasing decisions?


Figure 1: (Left) The effects of increasing the baserate of a particular brand context $(p(c))$ via advertising without aiming to associate the brand with particular products: Increasing the baserate of $c$ increases how often observers would choose brand $c$ when they are buying something $(p(c \mid b)$ ), but does not increase their propensity to purchase given exposure to the brand $(p(b \mid c))$. (Middle) In contrast, if increased brand exposure coincides with increasing the association of that brand with desirable goods $(p(x \mid c))$, consumers will also be more likely to purchase goods given that brand ( $p(b \mid c)$ increases). (Right) This increase in propensity to purchase goods by brand $c$ coincides with an increment in the marginal demand curve for brand $c: p(b \mid m, c)$ is elevated after such targeted promotions.

On our account, changing brand recall and recognition amounts to changing the context probability $p(c)$ for that brand (Figure 1 left). The immediate effect of increasing brand recognition and recall is an increase in $p(c \mid b, x)$ : given that a consumer has decided to buy a product, which brand will she choose? So long as the brand is associated strongly with products a particular product $x(p(x \mid c))$ an increase in $p(c)$ will yield an increase in $p(c \mid b, x)$; in other words consumers will be more likely to choose brand $c$ when asking themselves "I want to buy an $x$, which brand/retailer should I choose?".

However, our model also predicts that simply increasing $p(c)$ will have no effect on the consumers' eagerness to buy its specific products $p(b \mid x)$ or increase their eagerness to buy the brand $p(b \mid c)$. Our account suggests one immediate strategy for increasing consumers' eagerness to buy the brand: selectively increasing $p(x \mid c)$ for $x$ s with high $p(b \mid x)$ - in other words, strategically associating the brand with desirable goods. If the advertising that increases $p(c)$ also strategically increases $p(x \mid c)$ in this manner, then not only are consumers more likely to choose brand $c$ when making a purchase $(p(c \mid b))$, but they will overall be more likely to purchase the brand $(p(b \mid c))$. Moreover, this increase in propensity to buy the brand yields a uniform increase to the demand curve for the brand ( $p(b \mid m, c)$; Figure 1 right), showing just how effective a carefully selected increase in brand-product association can be.

Another interesting theoretical prediction from our model concerns the overuse of promotions presenting that brand without an associated product; this may be counterproductive as it might result in product-brand delinking. This could occur if, for instance, a company overemphasizes event sponsorships over product ads, such that the linking probability $p(x \mid c)$ is diluted by frequent observations of brand $c$ without
associated products $x$. Since such dilution will be accompanied by $p(c)$ gains, this will be a risk primarily for already familiar brands, for which $p(c)$ improvements are showing diminishing returns. In such situations consumers will show high brand awareness $p(c)$, but this will not translate into changes in consumption behavior $p(b \mid c)$.

This account also reaffirms other important elements of brand competition. In particular, it emphasizes product differentiation (Dickson \& Ginter, 1987), frequently cited as one of the major causes of ad campaign failures. If the product ( $x$ ) that a brand is associated with is considered to be a unique entity (e.g., "a Diet Coke") rather than a generic category (e.g., "a diet cola beverage"), then the gains of increased brand recognition will translate directly to increased demand for that brand's product. However, when a market is over-crowded, product differentiation becomes harder and costlier, thus gains in $p(c)$ will be lost because $p(x \mid c)$ does not adequately pick out the product of that particular brand, thereby reducing the potential gains from a higher $p(c)$. Furthermore, this account emphasizes the arms race nature of branding campaigns - the advantage is determined by relative frequency, rather than absolute frequency of brand exposure, which naturally imposes barriers to entry in existing competitive markets, as suggested previously by (Schmalensee, 1982) using empirical data.

## Rationalizing loss-leader strategies

Classically, the economic tension between the retailer and consumers' incentives maintains a price equilibrium. One potential advantage for the retailer is the relatively high costs of searching for low prices for every product, which motivates consumers to generalize about price (dis)advantages of retailers in aggregate, rather than for isolated products. Thus, insofar as consumers use aggregate price advantages to predict

Table 1：Predicted direction of effect of various marketing strategies on distributions stored by consumers．Direction of arrow shows whether effect is predicted to be positive $(\Uparrow)$ or negative $(\Downarrow)$ from the marketer＇s perspective．Citations given present evidence favoring the predicted direction．Arrows in both directions represent ambiguity about the effect of the intervention．

| Strategies | $\mathbf{p}(\mathbf{b} \mid \mathbf{x}, \mathbf{m}, \mathbf{c})$ | $\mathbf{p}(\mathbf{m} \mid \mathbf{x}, \mathbf{c})$ | $\mathbf{p}(\mathbf{x} \mid \mathbf{c})$ | $\mathbf{p}(\mathbf{c})$ |
| :---: | :---: | :---: | :---: | :---: |
| Event sponsorships | －（Dean，1999） | － | $\Downarrow$ | 介（Mazodier \＆Rezaee，2013） |
| Advertisements | 介ね | － | 介（Simon \＆Sullivan，1993） | $\Uparrow$（Simon \＆Sullivan，1993） |
| Endorsements | $\Uparrow$（Dean，1999） | － | 介（Dean，1999） | 介（Dean，1999） |
| Product placement | － | － | 介（Morton \＆Friedman，2002） | $\Uparrow$（Morton \＆Friedman，2002） |
| Sales／discounts | － | 介（Kalwani \＆Yim，1992） | － | － |
| Cash back | － | － | 介 | 介 |
| Promotions | － | － | $\Uparrow \Downarrow$ | $\Uparrow$ |

prices for new products，there is an opportunity to offer some carefully chosen products at a discount，and thus distort the aggregate inference：$p(m \mid c)=\sum_{x} p(m \mid x, c) p(x \mid c)$ ．

Specifically，by choosing to offer price reductions on salient products（high $p(x \mid c)$ ），promoters can skew the con－ sumers＇estimates of the overall priciness of brand $c$ ．Figure 2A－B shows this strategy in action．Reducing prices on highly observed items reduces observers＇estimates of how pricy a particular retail outlet might be $p(m \mid c)$ which informs their propensity to make purchases in such contexts $p(b \mid c)$ ．

Although exploiting salient products to distort consumers＇ estimates of overall price tendencies may seem exotic，it is not original to this paper．Amazon adopts this very policy by un－ dercutting competitors on their most popular products，while increasing prices on less salient goods（Del Rey，2015）．This strategy makes perfect sense under our account：with a keen enough understanding of $p(x \mid c)$（which we expect Amazon＇s Big Data provides）it may even be possible to increase overall prices while simultaneously decreasing consumers＇estimates of the prices offered by the brand．

## Rationalizing money－losing brand extensions

The principal way in which manufacturers can benefit from brand＇equity＇is by extending the repertoire of products as－ sociated with it．The problem lies in the possibility of dilution of the brand＇s association with individual products by virtue of exposure alongside multiple products．In the simple ac－ count of brand extensions，manufacturers bring new products to market to increase profits at the expense of brand equity． Our analysis，however，reveals the possibility of an inversion of this basic process－a manufacturer could potentially im－ prove brand equity by bringing a low－priced new product to market－trading off profits（or even losses）for brand equity．

In this situation，a company would manufacture a new product that sells for low prices at high volumes，and is in－ delibly associated with the company＇s brand．Such a sce－ nario would most likely play out for companies whose pri－ mary products are big－ticket，low volume items，e．g．cars， vacations etc，and that are looking to improve their visibility． Availabilty of the product at sufficiently low prices will raise
$p(b \mid x, m, c)$ ，which will in turn increase not just $p(b \mid x)$ for this low－price and likely low margin product，but also $p(b \mid c)$ and $p(c \mid b)$ ，thus increasing brand equity at fairly low cost．

To test this possibility，we added exposure to a new good specific to a particular context to the baseline event history in our simulation，available at the bottom two price labels in a＇cheap＇condition and at the top two price labels in a ＇pricey＇condition．We measured gain in brand equity as rel－ ative change in $p(r \mid c)$ from that measured in the baseline condition for this context．Figure 2C，which plots the rel－ ative gain in $p(r \mid c)$ for 100 model simulations from differ－ ent initializations，shows how brand equity improves through adding a loss－leader，and drops through adding a relatively expensive product to the product line．The latter is more prof－ itable，so this simulation demonstrates the existence of a com－ petitive tension between brand equity and capital－companies could potentially trade one off against the other sequentially， modulo diminishing returns from product－line overcrowding．

## Conclusion

Beginning with the intuition that marketing strategies influ－ ence consumers＇preference formation processes via associa－ tive influences within the preference construction process，we have created a theory of consumer preference formation that is grounded strongly in observable correlates for marketing variables．With a series of computational experiments，we have substantiated various predictions that this model makes about the impact of both existing and novel marketing strate－ gies，thus rationalizing several lines of consumer research findings via a simple inductive explanation of how consump－ tion preferences are formed．The model opens up a large space of possible experiments testing the effect of each of the variables we have defined on consumer behavior．Table 1 sug－ gests a number of directional hypotheses derived within our framework．We expect the strong observability of our model， in combination with its novel predictions，will benefit both theory and practice of marketing and consumer research，par－ ticularly in online retail settings，where the conditional distri－ butions implicated in our account are easy to access．


Figure 2: Predictions for advanced and speculative marketing strategies (Left) Flooding retail displays with cheap or discounted goods reduces observers' internal estimates of the price distribution $p(m \mid c)$, (middle) which promotes their propensity to make purchases in the retailer's chosen context. (Right) Similarly, the introduction of a cheap brand extension to the market can result in an increase in $p(b \mid c)$ - a measure of brand equity. All changes are measured from baselines estimated on the initial event history. Histograms show results for 100 simulations each.

## References

Ariely, D., Loewenstein, G., \& Prelec, D. (2006). Tom sawyer and the construction of value. Journal of Economic Behavior \& Organization, 60(1), 1-10.
Belk, R. W. (1975). Situational variables and consumer behavior. Journal of Consumer research, 157-164.
Bettman, J. R. (1979). An information processing theory of consumer choice. Journal of Marketing.
Biswas, A. (1992). The moderating role of brand familiarity in reference price perceptions. Journal of Business Research, 25(3), 251-262.
Dean, D. H. (1999). Brand endorsement, popularity, and event sponsorship as advertising cues affecting consumer pre-purchase attitudes. Journal of Advertising, 28(3), 112.

Del Rey, J. (2015). How amazon tricks you into thinking it always has the lowest prices. ReCode. Retrieved from http://recode.net/2015/01/13/
Dickson, P. R., \& Ginter, J. L. (1987). Market segmentation, product differentiation, and marketing strategy. The Journal of Marketing, 1-10.
Emery, F. (1969). Some psychological aspects of price. Pricing strategy, 98-111.
Gilboa, I., \& Schmeidler, D. (1995). Case-based decision theory. The Quarterly Journal of Economics, 605-639.
Huber, J., \& Puto, C. (1983). Market boundaries and product choice: Illustrating attraction and substitution effects. Journal of Consumer Research, 31-44.
Kalwani, M. U., \& Yim, C. K. (1992). Consumer price and promotion expectations: An experimental study. Journal of marketing Research, 90-100.
Kalyanaram, G., \& Winer, R. S. (1995). Empirical generalizations from reference price research. Marketing Science, 14(3_supplement), G161-G169.
Kirmani, A., \& Rao, A. R. (2000). No pain, no gain: A
critical review of the literature on signaling unobservable product quality. Journal of marketing, 64(2), 66-79.
Malhotra, N. K. (1988). Self concept and product choice: An integrated perspective. Journal of Economic Psychology, 9(1), 1-28.
Mazodier, M., \& Rezaee, A. (2013). Are sponsorship announcements good news for the shareholders? evidence from international stock exchanges. Journal of the Academy of Marketing Science, 41(5), 586-600.
Mazumdar, T., Raj, S., \& Sinha, I. (2005). Reference price research: Review and propositions. Journal of marketing, 69(4), 84-102.
Monroe, K. B. (1973). Buyers' subjective perceptions of price. Journal of marketing research, 70-80.
Morton, C. R., \& Friedman, M. (2002). i saw it in the movies: Exploring the link between product placement beliefs and reported usage behavior. Journal of Current Issues \& Research in Advertising, 24(2), 33-40.
Mowen, J. C., \& Gaeth, G. J. (1992). The evaluation stage in marketing decision making. Journal of the Academy of Marketing Science, 20(2), 177-187.
Munoz, T., \& Kumar, S. (2004). Brand metrics: Gauging and linking brands with business performance. The Journal of Brand Management, 11(5), 381-387.
Schmalensee, R. (1982). Product differentiation advantages of pioneering brands. The American Economic Review, 349-365.
Simon, C. J., \& Sullivan, M. W. (1993). The measurement and determinants of brand equity: A financial approach. Marketing science, 12(1), 28-52.
Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453-458.
Winer, R. S. (1986). A reference price model of brand choice for frequently purchased products. Journal of consumer research, 250-256.

# Spatial language promotes cross-domain associations in early childhood 

Ariel Starr (arielstarr@berkeley.edu) \& Mahesh Srinivasan (srinivasan@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94720


#### Abstract

Spatial language is often used metaphorically to describe other domains, including time (a long sound) and pitch (a high sound). How does experience with these metaphors shape the associations we make across disparate domains? Here, we tested 3- to 6 -year-old English-speaking children and adults with a cross-domain matching task that assessed space-time and space-pitch mappings. We tested spatial relations that are expressed in English metaphors for time and pitch, as well as metaphors that are unfamiliar to English speakers, but expressed in other languages. Participants performed a perceptual matching task, in which they matched pictures and sounds, and a linguistic matching task, in which they matched pictures or sounds to verbal labels. Adults readily matched between space and time and between space and pitch, using relations expressed by both familiar and unfamiliar metaphors. Children showed an advantage for linguistic matching compared to perceptual matching, but their performance was similarly unaffected by metaphor familiarity. Together, these results suggest that spatial language promotes the development of cross-domain associations, and that experience with particular spatial metaphors is not required to produce this benefit.


Keywords: metaphor theory; linguistic relativity; crossmodal matching

## Introduction

Across languages and cultures, spatial language is frequently co-opted to describe other domains. In English, for example, we describe temporal duration as long or short, numbers as big or small, and auditory pitch as high or low. What can the prevalence of these spatial metaphors tell us about how we represent and reason about these other nonspatial domains? Previous work has demonstrated that spatial metaphors are not simply a communicative tool. Instead, they reflect our mental representations of nonspatial domains (Boroditsky, 2001; Casasanto \& Boroditsky, 2008; Dolscheid, Shayan, Majid, \& Casasanto, 2013). Here, we test how experience with spatial metaphors over development influences children's cross-domain associations.

Although spatial metaphors are common across languages, there is also variety in the exact spatial relations invoked. This cross-linguistic variation can be used to test hypotheses about the role of linguistic experience in the development of cross-domain associations. For example, in English, temporal durations are described in terms of twodimensional length, whereas languages including Greek and Spanish use three-dimensional spatial terms. Likewise, though English describes pitch in terms of height, languages including Turkish and Farsi use terms related to thickness (i.e., such that thicker sounds are lower in pitch).

Linguistic experience is not required to recognize crossdomain associations. Many studies have demonstrated that prelinguistic infants are already sensitive to many types of these correspondences (de Hevia \& Spelke, 2010; Lourenco \& Longo, 2010; Mondloch \& Maurer, 2004; Srinivasan \& Carey, 2010; Walker et al., 2010). Even neonates, for example, associate longer spatial lengths with longer temporal durations and larger numerical magnitudes (de Hevia, Izard, Coubart, Spelke, \& Streri, 2014). With regards to associations between pitch and space, infants appear to recognize both height-pitch and thickness-pitch mappings, even when only one of these associations is encoded in the language they are learning (Dolscheid, Hunnius, Casasanto, \& Majid, 2014; Walker et al., 2010; but see Lewkowicz \& Minar, 2014).

However, according to one recent study, young children may be less flexible than infants (Shayan, Ozturk, \& Bowerman, 2014). This study investigated thickness-pitch mappings in 2- to 5 -year-old children who spoke either German, Farsi, or Turkish, and found that while Turkish and Farsi speaking children (who speak languages that employ a thickness metaphor for pitch) can reliably map thickness to pitch, German-speaking children (who speak a language that does not employ a thickness-pitch metaphor) cannot. This finding suggests that maintenance of cross-domain associations between space and pitch may be dependent on the type of metaphorical mappings reinforced by language, at least during childhood.

In explicit matching tasks, adults appear to be more flexible than children. Adults can match pitch to both thickness and height, for example, and more generally can form mappings across innumerable domains, regardless of whether their language employs the relevant metaphors (Marks, 1978; Shayan et al., 2014) However, experience with linguistic metaphors does appear to influence the automaticity with which cross-domain associations are processed in adults. In one study, adult speakers of Dutch readily matched pitch to thickness in an explicit task, but their representations of pitch were only biased by irrelevant spatial height information and not by irrelevant thickness information (Dolscheid et al., 2013). This suggests that Dutch-speaking adults automatically process a height-pitch mapping but not a thickness-pitch mapping. The reverse pattern of results - a biasing effect of thickness but not height on pitch representations - occurs for Farsi speakers, consistent with the idea that experience with languagespecific spatial metaphors influences the automaticity with which cross-domain associations are accessed. A parallel pattern of results for the case of space-time mappings has been reported from English and Greek-speaking adults. In
this study, temporal duration judgments in English-speaking adults were biased by only irrelevant variation in spatial length (and not irrelevant volumetric information), and duration judgments in Greek-speaking adults were biased only by irrelevant volumetric information (and not irrelevant spatial length; Casasanto, 2008).

In sum, prior research suggests that experience with one's own language's spatial metaphors affects children's explicit cross-domain associations, as well as adults' implicit crossdomain associations. This suggests that experience with spatial metaphors may have a persistent influence on crossdomain associations throughout the lifespan. Here, we expand on previous studies of the development of crossdomain associations by testing both space-time and spacepitch associations, and by directly comparing the effects of familiar versus novel metaphors in the same population. In addition, we test a large number of children between the ages of three and six years to gain insight into the finegrained developmental trajectory of cross-domain associations. We test how children's experience with spatial metaphors influences their cross-domain associations by focusing on three critical factors.

First, we contrast cross-domain associations between space and time compared to associations between space and pitch. We focus on associations with time and pitch because while both time and pitch are frequently described using spatial language, they have different dimensional structures. Time, like spatial extent, is a prothetic dimension, meaning that it can be represented by an ordered continuum of increasing amount (Stevens, 1957). Differences in temporal duration and spatial extent are therefore quantitative. Pitch, on the other hand, is a metathetic dimension, meaning that differences in pitch are qualitative - pitches vary in frequency rather than amount. Therefore, it is possible that the shared ordinal structure of space and time may provide a specific advantage for cross-domain space-time mappings that does not extend to space-pitch mappings.

Second, we investigate the effect of experience with specific spatial metaphors by testing whether Englishspeaking children are better able to match domains in ways that reflect familiar spatial metaphors compared to unfamiliar metaphors. For time, we compared children's ability to map between sounds that varied in temporal durations and images that varied in either length (long vs. short; familiar relation) or in overall size (big vs. small; unfamiliar relation). For pitch, we compared children's ability to map between sounds that varied in auditory pitch and images that varied in either height (high vs. low; familiar relation) or thickness (thin vs. thick; unfamiliar relation). If experience with specific spatial metaphors constrains children's cross-domain associations, then matching performance should be higher for the familiar relations compared to the unfamiliar relations. In addition, the difference in performance between familiar and unfamiliar relations may increase with age, as children gain more experience with their language-specific spatial metaphors.

Finally, we probe whether children are better at matching across domains when the spatial metaphors are verbally labeled - even when the metaphors are unfamiliar compared to when the task is purely perceptual. Previous research suggests that verbal labels may help children organize their representations of perceptual domains by providing cues as to how to align the endpoints of disparate domains (Smith \& Sera, 1992). For example, if children understand that the word long in a spatial context refers to greater spatial extent, this may provide a cue for understanding that long in a temporal context refers to a greater temporal duration. Therefore, children's nonlinguistic ability to successfully map across domains (as measured by a perceptual matching task) may be enhanced by the presence of verbal spatial labels (as measured by a linguistic matching task).

## Methods

## Participants

80 children aged 3 to 6 years (mean age: 4.89 years, range: 3.13-6.98 years) and 16 adults (mean age: 21.25 years, range: 18.63-27.59 years) participated in this study. All participants were native English language speakers who were not regularly exposed to or fluent in a second language. Data from an additional 17 children and 6 adults were excluded from analyses due to proficiency with another language ( 9 children, 5 adults), failure to complete the experiment ( 5 children), inattention ( 3 children), or performance more than 3 standard deviations below the group mean (1 adult).

## Materials

Spatial stimuli consisted of pictures of cartoon aliens that varied in length (familiar space-time metaphor), overall size (unfamiliar space-time metaphor), vertical position (familiar space-pitch metaphor), or thickness (unfamiliar space-pitch metaphor). Temporal stimuli consisted of monotonic tones that varied in either duration or auditory pitch. Tones that varied in duration had a constant pitch of 384 Hz and were either 1 second or 3 seconds in length. Tones that varied in pitch had a constant duration of 2 seconds and a pitch of either 256 Hz or 512 Hz . All stimuli were presented using a laptop computer.

## Procedure

Participants' cross-domain matching ability was tested for both space-time and space-pitch pairings. Adult participants completed both the familiar (long/short for space-time and high/low for space-pitch) and unfamiliar (big/small for space-time and thin/thick for space-pitch) pairings. The order of the space-time and space-pitch blocks was counterbalanced across subjects, and the unfamiliar pairing always preceded the familiar pairing within each block. Child participants completed one block of space-time pairings and one block of space-pitch pairings, with the
order of the blocks and assignment to the familiar versus unfamiliar pairings counterbalanced across participants.

All experimental sessions began with the familiarization trials. At the beginning of each block, participants were first shown pictures of both relevant aliens and listened to both types of sounds. Within each block, the perceptual matching task always preceded the linguistic matching task so as to not bias participants' responses. Adults performed 8 trials of each task type, and children performed 4 trials of each task type.

## Familiarization Trials

The familiarization trials were designed to have the same structure as the test trials and involved matching pictures of animals to the appropriate sounds. In the first two trials, two animals were displayed on the screen. An animal sound was then played, and the participant was instructed to point to the animal that makes that sound. In the second two trials, one animal appeared centrally on the screen in front of two trees. The participant was told that the animal was looking for another animal just like it that was hiding in the jungle. The experimenter said that the animal was hiding behind one of the trees, and played a sound from each tree. The participant was instructed to point to the side where the animal sound that matched the visible animal was heard.

## Test Trials

Perceptual Matching Task In the perceptual matching task, participants matched pictures of aliens to the types of sounds they make. Critically, in these trials verbal labels were never used to describe the stimuli. There were two trial types: space as source (the referent is a spatial dimension and participants chose the sound that matched in pitch or duration) or space as target (the referent is a sound that varies in pitch or duration and participants chose the alien with a matching spatial attribute). For the space as source trials, a single alien was presented in front of two trees and the experimenter said that the alien was looking for another alien just like it. The experimenter said that the alien could be hiding behind either tree, and pointed to each tree as the sound of the alien hiding behind it was played. The participant was asked to point to the tree that had an alien behind it that was just like the visible alien. For the space as target trials, two aliens were presented on the screen and a single sound was played. The participant was instructed to choose which of the aliens is the one that makes that sound.

Linguistic Matching Task In the linguistic matching task, a verbal label was used to describe either the appearance of an alien or the type of sound made by an alien, and the participant matched this label to one of two exemplars in the opposite dimension. In the space as source trials, the spatial dimension was described and participants chose one of two auditory matches (e.g., the participant is told that a long alien is looking for another long alien just like it, and the long alien is hiding behind one of two trees; the two sounds are played and the participant chooses the sound that a long
alien makes). In the space as target trials, the type of sound was described and participants chose which of two visually presented aliens makes that type of sound (e.g., the experimenter asks the participant which of two aliens makes a long sound).

## Results

Accuracy for all matches was scored as correct if the match was in the direction reflected by spatial metaphors in language (e.g., matching the long or big alien to the long tone and the short or small alien to the short tone). Children's performance was analyzed using a repeated measures ANOVA with match type (perceptual or linguistic) and match direction (space as source or space as target) as within-subjects factors and dimension (space-time or space-pitch), familiarity (familiar or unfamiliar), and age as between-subjects factors. This model yielded a significant main effect of age $(F(3,144)=5.35, p<.005$; Figure 1A) and a significant main effect of match type ( $F(1$, $144)=17.14, p<.001$; Figure 1B). Interestingly, neither the main effect of dimension nor familiarity was significant, indicating that overall children were equally proficient at matching across space and time compared to space and pitch, and they performed just as well for spatial relations employed by English-language metaphors compared to unfamiliar space relations.


Figure 1: Children's cross-domain matching performance by age group (A) and by match type (B). Error bars indicate SEM.

Planned post-hoc comparisons revealed that 3-year-olds performed worse than the older age groups, all of whom performed at similar levels (3-year-olds: $\mathrm{M}=54.08$, $\mathrm{SEM}=$ 2.67; 4-year-olds: $\mathrm{M}=70.24$, $\mathrm{SEM}=2.68$; 5 -year-olds: $\mathrm{M}=$ $72.70, \mathrm{SEM}=3.16,6$-year-olds $=73.9, \mathrm{SEM}=3.42$; all $t \mathrm{~s}$ for 3-year-olds vs. older children $>3.9, p s<.001$; all $t$ s between 4-, 5-, and 6-year-olds < .7, ps > .5). In addition, performance was better for the linguistic matching trials compared to the perceptual matching trials (linguistic matching: 72.34, SEM: 2.09; perceptual matching: 61.56, SEM: 2.13, $t=3.12, p=.002$ ).

This analysis also revealed a significant interaction between match type and match direction $(F(1,144)=5.17$,


Figure 2: Children's matching performance on the linguistic and perceptual matching tasks, grouped by age and match direction. Error bars indicate SEM.


Figure 3. Children's matching performance for familiar and unfamiliar metaphors, grouped by age and match direction. Error bars indicate SEM.
$p<.05)$ and significant three-way interactions between age, match direction, and match type $(F(3,144)=3.62, p<.05)$, and between age, match direction, and familiarity $(F(3,144)$ $=3.12, p<.05)$. We analyzed the three-way interaction between age, match direction, and match type with separate ANOVAs for the linguistic and perceptual match trials (Figure 2). For the linguistic matching trials, the effects of both age and match direction were significant ( $F s>5.6, p$ s $<.05$ ). Overall, three-year-olds performed worse than older children, and performance was higher when space was the target dimension compared to the when space was the source dimension. For the perceptual matching trials, the main effect of age and the interaction between age and match direction were significant ( $F s>2.8, p s<.05$ ), indicating that older children outperformed younger children, and particularly so when space was the target dimension.

We analyzed the three-way interaction between age, match direction, and familiarity with separate ANOVAs for familiar versus unfamiliar metaphors (Figure 3). For familiar metaphors, there was a significant main effect of age and a significant interaction between age and match direction ( $F s>3.6, p s<.05$ ), indicating that older children
performed better than younger children, and the age effect was particularly pronounced when space was the source dimensions. For unfamiliar metaphor trials, there were no significant main effects, but the interaction between age and match direction was significant $(F(3,54)=3.11, p<.05)$, suggesting that the effect of age was more pronounced when space was the target dimension.

Adults' performance across all conditions was near ceiling (mean $=96.19$, $\mathrm{SEM}=.51$ ). Performance was analyzed using a repeated measures ANOVA with dimension (space-time or space-pitch), familiarity (familiar or unfamiliar), match type (perceptual or linguistic) and match direction (space as source or space as target) as within-subject factors. This analysis revealed no significant effects ( $F s<2.2, p s>.16$ ), nor any significant interactions ( $F s<3.1, p s>.1$ ). Therefore, adults performed equally well in all conditions.

## General Discussion

The present work explored cross-domain associations between space and time and between space and pitch, and the role that experience with spatial metaphors may play in shaping these representations. English-speaking children
and adults performed perceptual and linguistic cross-domain matching tasks for pairs of stimuli that varied in spatial extent and either temporal duration or auditory pitch. For both types of cross-domain pairings, we assessed matching performance for pairs that reflected familiar Englishlanguage spatial metaphors (length for time, height for pitch) or that reflected novel spatial metaphors not used in English (size for time, thickness for pitch). Consistent with previous work (e.g., Marks, 1978; Shayan et al., 2014), we found that adults readily matched time and pitch to both familiar and unfamiliar spatial attributes. Children's matching performance, however, revealed a nuanced developmental trajectory for these cross-domain associations, which we describe below.

Most notably, children's cross-domain matching performance was better for the linguistic matching task compared to the perceptual matching task. In the linguistic task, participants were provided with a verbal label that described a stimulus in terms of its spatial attributes, auditory pitch, or temporal duration, and needed to choose an exemplar from the other domain to match it. By comparison, in the perceptual matching task, participants matched the exemplars in the absence of a verbal label. This suggests that the presence of a verbal label conferred an advantage for cross-domain matching above and beyond matching the exemplars themselves.

Strikingly, we found that the verbal label provided an advantage even when it labeled a spatial relation not employed by English-language metaphors. In contrast to previous work (Shayan et al., 2014; but see Dolscheid, Hunnius, \& Majid, 2015), we found no significant effect of metaphor familiarity on children's matching performance. Children were equally proficient at matching spatial relations across domains that reflected both familiar and novel metaphors. Although English-speaking children presumably have little experience with thickness metaphors for pitch, they matched thickness onto pitch just as readily as they matched vertical height onto pitch. Likewise, they matched overall size onto duration just as readily as they matched spatial length onto duration. Given that previous findings have provided mixed results as to whether familiarity with specific spatial metaphors is required for children's success in space-pitch matching tasks, it is unclear whether these contrasting outcomes should be ascribed to differences in procedure or population. In addition, familiarity in the present experiment was defined based on the presence or absence of specific spatial metaphors in the English language. Therefore, it is possible that individual differences in children's experience with these metaphors may influence their matching performance. Regardless, the present results suggest that young children, like infants and adults, can flexibly map abstract domains onto multiple spatial reference frames. We also found no significant effect of dimension on children's mapping performance, which indicates that the shared ordinal structure between space and time did not confer an
advantage for mapping between these dimensions compared to mapping between space and pitch.

We found that verbal labels were most likely to improve performance for the youngest children when space was the target domain. On these trials, children were presented with two aliens and were asked to choose which one made a labeled sound (e.g., which alien makes a thick sound). Therefore, to match correctly, children could simply choose the alien whose visual appearance matched the label (e.g., a thick alien), without needing to represent the labeled sound (e.g., a thick sound). These were the trials on which three-year-olds performed the best, suggesting that understanding the spatial meaning of these words precedes understanding of the metaphorical meaning. Indeed, with regards to spatial metaphors for time, children typically produce the spatial meaning of the word earlier than the temporal meaning
(Clark, 1973). However, for older children the performance benefit for linguistic trials held both when space was the target domain and when space the source domain. Therefore, it was not solely the trials on which children could match a label to a spatial attribute that drive this effect because there was also improved performance when children mapped the label to a sound. Instead, it seems that the presence of labels themselves improves children's crossdomain matching performance.

The finding that children perform better on the linguistic matching task compared to the perceptual matching task is consistent with previous work suggesting that language is a facilitating factor in the development of cross-domain associations. This work suggests that children may initially form mappings between the labels for two domains, such that the association between the labels then drives the perceptual mapping. For example, when forming an association between size and auditory volume, children may initially map the word big onto the word loud and the word small onto the word quiet, and this linguistic association may lead children to think of loud as being more than quiet and lead to an association between the perceptions of size and volume (Smith \& Sera, 1992). This explanation can be logically extended to associations between space and time as well, with the common labels of long and short providing ordinal cues to children as to how to align and map these domains. However, it is less clear how this explanation applies to pitch, as the spatial metaphors used to describe pitch seemingly ascribe opposite ordinal anchors to the spectrum of pitch: both thick and low refer to low-frequency pitches, yet thick typically corresponds to more whereas low typically corresponds to less. Therefore, it seems that in the present task, labels must be providing an additional cue beyond an ordinal reference frame.

Another advantage that labels may provide is by clarifying what is otherwise an ambiguous task. When children are initially mapping between the pictures and the sounds, they may not spontaneously focus on the spatial attributes that are varying. However, labeling a particular dimension likely makes that dimension more salient, thus clarifying the goal of the task. For example, when shown a
thick alien and asked which of two sounds that type of alien would make, children may not immediately recognize that they could consider the width of the alien when making their choice. However, when asked which of two sounds a thick alien makes, children may perceive thickness as a relevant attribute. Therefore, the act of labeling itself may provide an additional cue for mapping that is not present in the perceptual matching task.

Although our results suggest that children's matching performance increases with age, the largest change in performance occurred between ages three and four. Overall, three-year-olds performed at chance, whereas four-, five-, and six-year-olds all performed at similar levels above chance. From the present data it is difficult to determine whether this jump in performance reflects improvements in cross-domain mapping ability, or whether the demands of our matching task may be too taxing for three-year-olds. Given that cross-domain associations have been demonstrated in infants using more implicit tasks, additional work is needed to trace the development of these associations between infancy and early childhood. Further, the present study involved making explicit matches between domains. Although adults can form explicit mappings across a multitude of domains (Stevens, 1957), there are constraints on the types of cross-domain associations that occur implicitly (e.g., Casasanto, 2008; Dolscheid et al., 2013; Srinivasan \& Carey, 2010). Therefore, it remains an open question whether children spontaneously associate space and time and space and pitch, and whether experience with particular spatial metaphors may influence the automaticity with which these associations are accessed. Reaction time measures may be useful for addressing this question, because implicit matching processes should proceed more rapidly than explicit matching processes.

Taken together, these findings suggest that spatial language promotes cross-domain associations in early childhood. Critically, this process appears to be equally accessible for spatial metaphors that are both familiar and novel, suggesting that experience with specific spatial metaphors is not necessary for forming these associations. Instead, spatial language may promote the perceptual organization of other domains by providing a reference frame for aligning these domains, as well as by highlighting relevant spatial attributes.

## Acknowledgments

We would like to thank Gillian Schwartz, Alagia Cirolia, and Leslie Wang for their assistance with data collection. This research is supported by NIH award F32HD085736 to AS and NSF award SBE-16302040 to MS.

## References

Boroditsky, L. (2001). Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time. Cognitive Psychology, 43(1), 1-22.
Casasanto, D. (2008). Who's afraid of the big bad Whorf? Crosslinguistic differences in temporal language and
thought. Language Learning, 58, 63-79.
Casasanto, D., \& Boroditsky, L. (2008). Time in the mind: Using space to think about time. Cognition, 106(2), 579593.

Clark, H. H. (1973). Space, time, semantics, and the child. In Cognitive Development and the Acquisition of Language. New York: Academic Press.
de Hevia, M. D., \& Spelke, E. S. (2010). Number-space mapping in human infants. Psychological Science, 21(5), 653-660.
de Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., \& Streri, A. (2014). Representations of space, time, and number in neonates. Proceedings of the National Academy of Sciences, 111(13), 4809-4813.
Dolscheid, S., Hunnius, S., \& Majid, A. (2015). When high pitches sound low: Children's acquisition of space-pitch metaphors. Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Dolscheid, S., Hunnius, S., Casasanto, D., \& Majid, A. (2014). Prelinguistic Infants Are Sensitive to Space-Pitch Associations Found Across Cultures. Psychological Science, 25(6), 1256-1261.
Dolscheid, S., Shayan, S., Majid, A., \& Casasanto, D. (2013). The Thickness of Musical Pitch: Psychophysical Evidence for Linguistic Relativity. Psychological Science, 24(5), 613-621.
Lewkowicz, D. J., \& Minar, N. J. (2014). Infants Are Not Sensitive to Synesthetic Cross-Modality Correspondences: A Comment on Walker et al. (2010). Psychological Science, 25(3), 832-834.
Lourenco, S. F., \& Longo, M. R. (2010). General magnitude representation in human infants. Psychological Science, 21(6), 873-881.
Marks, L. E. (1978). The unity of the senses: Interrelations among the modalities. Academic Press.
Mondloch, C., \& Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. Cognitive, Affective, \& Behavioral Neuroscience, 4(2), 133.

Shayan, S., Ozturk, O., \& Bowerman, M. (2014). Spatial metaphor in language can promote the development of cross-modal mappings in children. Developmental Science, 17(4), 636-643.
Smith, L. B., \& Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. Cognitive Psychology, 24(1), 99-142.
Srinivasan, M., \& Carey, S. (2010). The long and the short of it: On the nature and origin of functional overlap between representations of space and time. Cognition, 116(2), 217-241.
Stevens, S. S. (1957). On the psychophysical law. Psychological Review, 64(3), 153.
Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., \& Johnson, S. P. (2010). Preverbal Infants' Sensitivity to Synaesthetic Cross-Modality Correspondences. Psychological Science, 21(1), 21-25.

# Preemption in Singular Causation Judgments: A Computational Model 

Simon Stephan (simon.stephan@psych.uni-goettingen.de) Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)<br>Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany


#### Abstract

Causal queries about singular cases are ubiquitous, yet the question of how we assess whether a particular outcome was actually caused by a specific potential cause turns out to be difficult to answer. Relying on the causal power approach, Cheng and Novick (2005) proposed a model of causal attribution intended to help answering this question. We challenge this model, both conceptually and empirically. The central problem of this model is that it treats the presence of sufficient causes as necessarily causal in singular causation, and thus neglects that causes can be preempted in their efficacy. Also, the model does not take into account that reasoners incorporate uncertainty about the underlying causal structure and strength of causes when making causal inferences. We propose a new measure of causal attribution and embed it into our structure induction model of singular causation (SISC). Two experiments support the model.


Keywords: singular causation; causal attribution; preemption; causal reasoning; Bayesian modeling; computational modeling

## Introduction

Most people hold the belief that smoking causes lung cancer. Now, imagine that you learn that Peter, a passionate smoker, has contracted lung cancer. How strongly would you be willing to say that it was Peter's smoking that was causally responsible for his disease?

This example illustrates a scenario in which we seek an answer to a causal query about a singular case. Queries about singular causation are prevalent in everyday life and professional contexts, such as the law or medicine. How do people derive causal judgments about singular cases? Of course, the mere fact that two factors $C$ and $E$ are generally causally connected (e.g., smoking often causing lung cancer) does not necessarily imply that a singular or token co-occurrence of these events (e.g., Peter's smoking and his lung cancer) manifests a causal relationship - a singular co-occurrence might be a mere coincidence. On the other hand, as causality is not directly observable in the world, to what else than our general causal knowledge could we turn to obtain answers?

We are going to present a theory that builds on the idea, first formalized by Cheng and Novick (2005), that the notion of unobservable causal powers (Cheng, 1997) plays an essential role in singular causation judgments. Yet, we will demonstrate that Cheng and Novick's (2005) power PC model of causal attribution ( CN model) makes assumptions that are not always plausible. Generally, the CN model is intended to provide a normative answer to the question how we can determine whether an observed outcome was actually caused by a potential cause factor. For example, for cases like the one above about Peter in which a potential cause $c$ and an effect $e$ have been observed, the CN model delivers the probability
$P(c \rightarrow e \mid c, e)$ with the arrow denoting a causal relation. We argue that the key problem of the model is that it treats target causes as singular causes whenever they are sufficient for the effect in a specific situation. This appears to be at first sight a reasonable assumption, yet it ignores that the exact points in time at which different causes exert their powers play an important role in singular causation judgments (see Danks, 2017): crucially, sufficient causal powers can be preempted by others, and in such cases they should not be held causally responsible for the occurrence of the outcome. We will argue that preemption of causes by background factors frequently occurs in singular causation scenarios, and therefore presents a problem for the CN model.

Another problem of the CN model is that it does not take into account uncertainty about both the underlying causal structure and the causal parameters (e.g., the size of the causal powers). To incorporate uncertainty about the causal parameters, Holyoak, Lee, and Lu (2010) have proposed a Bayesian version of the CN model that uses probability distributions over the parameters instead of point estimates. However, their model also neglects uncertainty about the underlying causal structure. Both sources of uncertainty have been demonstrated to influence causal learning and reasoning (Griffiths \& Tenenbaum, 2005; Meder, Mayrhofer, \& Waldmann, 2014). For this reason, Stephan and Waldmann (2016) proposed the structure induction model of singular causation (SISC) that incorporates both types of uncertainty. Although three experiments (Stephan \& Waldmann, 2016) showed that SISC better accounted for the results than the standard power PC model of causal attribution, one shortcoming of the initial version of SISC was that it used the CN conceptualization of causal attribution that we are going to criticize in the present paper.

We will start with a theoretical section in which we defend a new measure of causal attribution as a component of SISC that is sensitive to preemption. We then present the results of two experiments. Experiment 1a confirmed that singular causation judgments deviate systematically from the predictions of the CN model in line with our revised causal attribution equation. Experiment 1 b assessed participants' notion of preemption. In Experiment 2 we used a larger set of contingencies to compare the revised SISC with the CN and other models. The results of this experiment showed that both a revision of the causal attribution equation and the consideration of statistical uncertainty are crucial to explain the findings.

## The Power PC Model of Causal Attribution

According to Cheng's (1997) power PC theory, causal power (or causal strength) is a hypothetical, unobservable en-
tity that represents the strength of causes. Mathematically, causal power is (in the generative case) expressed as the probability with which a target cause brings about its effect in a hypothetical world in which all alternative observed and unobserved causes of the effect are absent. Because of the possibility of unobserved alternative causes, causal power cannot be assessed directly but must be inferred based on the observed covariation and background assumptions. For generative causes, the following equation can be used to estimate the causal power $w_{c}$ of a target cause $C$ :

$$
\begin{equation*}
w_{c}=\frac{P(e \mid c)-P(e \mid \neg c)}{1-P(e \mid \neg c)}=\frac{\Delta P}{1-w_{a}} \tag{1}
\end{equation*}
$$

In this equation, $w_{a}$ represents the aggregate causal power of all alternative causes $A$ of the effect, which are assumed to exert their influence independently of $C$.

Under the causal Bayes net framework, the causal power of $C, w_{c}$, corresponds to the probabilistic weight of the causal arrow that connects $C$ with its effect $E$ in a common effect structure in which the target cause $C$ and the alternative causes $A$ combine in a noisy- $O R$ gate (see $S_{1}$ in Figure 2). Likewise, $w_{a}$ corresponds to the weight of node $A$.

## The CN Measure of Causal Attribution

Cheng and Novick (2005) proposed several measures of causal attribution that apply to different cases. The measure of causal attribution for cases in which both $c$ and $e$ are present, as in the example above about Peter, utilizes the concept of causal power in the following way to deliver the conditional probability $P(c \rightarrow e \mid c, e)$ :

$$
\begin{equation*}
P(c \rightarrow e \mid c, e)=\frac{w_{c}}{P(e \mid c)}=\frac{w_{c}}{w_{c}+w_{a}-w_{c} \cdot w_{a}} \tag{2}
\end{equation*}
$$

Equation 2 shows that the CN model defines the probability with which $c$ is causally responsible for $e$ given that both have co-occurred by the fraction of the causal power of $C$ and the conditional probability of the effect in the presence of $C$. Since the power PC theory assumes that $C$ and $A$ exert their causal powers independently of each other, $P(e \mid c)$ can be rewritten as the sum of both causal powers minus their intersection (see second step of Equation 2). Hence, what the CN model delivers is an estimation of the relative frequency of cases among all co-occurrences of $C$ and $E$ in which $C$ 's causal power is sufficient for the production of the effect. Our key criticism is that this relative frequency, because it neglects the possibility of preemption, frequently overestimates the true proportion of cases in which we should actually causally attribute $E$ 's occurrence to $C$.

To illustrate the problem, let us consider the results of the fictitious experiment shown in Figure 1 in which the influence of a chemical substance on the expression of a gene was investigated. As it is the case that all mice in the test group ( $P[e \mid c]=1$ ) but only one half in the control group $(P[e \mid \neg c]=.5$, the base rate) exhibit the gene, the results provide strong evidence for the existence of a strong effect of the chemical. In fact, by applying Equation 1 one can see that


Figure 1: Illustration of a hypothetical study testing the effect of a chemical on the expression of a gene. The control group is shown on the left and the test group treated with the chemical substance on the right. Mice having the gene expressed are depicted in blue.
the causal power of the substance equals 1. Crucially, so does $P(c \rightarrow e \mid c, e)$. What the CN model therefore prescribes is that we should attribute causal responsibility to the chemical whenever the chemical and the gene are both present. But should we really be maximally confident that the expression of the gene in, for instance, Mouse \# 25 must be causally attributed to the causal power of the chemical? If you have doubts, you were probably led by a prior assumption about the point in space and time at which the causal background factors $A$ produced the observed base rate of fifty percent: in the given scenario it seems likely that these factors (e.g., transcription factors) already produced their effects prior to the introduction of the chemical. Under this assumption, however, it seems likely that not only fifty percent of the mice in the control condition but also in the test group already possessed the gene prior to the study. Consequently, in those fifty percent of the mice it cannot be the chemical that is causally responsible for the effect because its causal efficacy has been preempted by the background factors.

## A New Measure of Causal Attribution

In the example it seems appropriate to say that the expression of the gene is caused by the chemical in only about half of the observed cases in which $C$ and $E$ have co-occurred. This conclusion is based on the assumption that in roughly half of the cases the causal power of the chemical has been preempted by the causal power of the background factors $A$. We propose a new measure of causal attribution that captures this intuition by refining Equation 2 so that all cases among the joint occurrences of $C$ and $E$ for which the effect of $C$ is assumed to be preempted by $A$ are partialed out. This refined measure is given by:

$$
\begin{equation*}
P(c \xrightarrow{\text { singular }} e \mid c, e)=\frac{w_{c} \cdot\left(1-\alpha \cdot w_{a}\right)}{w_{c}+w_{a}-w_{c} \cdot w_{a}}, \tag{3}
\end{equation*}
$$

in which we introduce $\alpha$ as a discounting parameter that represents the assumed probability with which $A$ is a preemptive cause of the effect. For illustration, if we assume that $A$ has caused the observed base rate of 0.5 prior to the application of the chemical, and if we assume further that $A$ 's causal power has produced roughly equal proportions of the effect in both groups, $\alpha$ takes on a value of 1 . In this case, the point estimate for $P(c \xrightarrow{\text { sing. }} e \mid c, e)$ is about .5 instead of 1.0 that is predicted


Figure 2: The two causal structures considered by the structure induction model of singular causation as mutually exclusive explanations for observed patterns of covariation. $C$ is a general cause of $E$ under $S_{1}$ whereas all co-occurrences of $C$ and $E$ are coincidental under $S_{0}$. The parameters $w_{c}$ and $w_{a}$ denote the causal powers of the observed cause $C$ and the unobserved background causes $A$, respectively; $b_{c}$ denotes the base rate of $C$.
by Equation 2. The difference between the two measures is that under the CN model the presence of two sufficient causal powers ( $w_{c}$ and $w_{a}$ ) is invariably conceptualized as a case of "symmetric overdetermination", whereas the possibility that causes can preempt each other is neglected. Equation 3 takes into account the possibility of preemption and thus delivers an estimation of the relative frequency of singular cases in which the target cause has actually been successful in generating the effect. In our view, preemption of $C$ by a previously present background factor $A$ seems to be prevalent in the cover stories typically reported in the literature (see e.g., Griffiths \& Tenenbaum, 2005). However, the discounting parameter $\alpha$ can be set to capture other cases. For example, cases of overdetermination or cases in which $A$ is preempted by $C$ could be modeled by setting $\alpha$ to 0 . In these cases, our model and the CN model make identical predictions because Equation 3 reduces to Equation 2.

## SISC: The Structure Induction Model of Singular Causation

Apart from the fact that the CN model does not take into account the possibility that preempted causes should not be classified as singular causes, a further problem of the CN model is that it is insensitive to statistical uncertainty about both the underlying causal structure and the size of the causal parameters. SISC (Stephan \& Waldmann, 2016) is sensitive to both types of uncertainty.

SISC was developed in the framework of causal Bayesian inference models; it takes observed data as evidence to update prior probabilities of mutually exclusive hypotheses. Under SISC, these competing hypotheses represent two causal structures that can account for a particular observed pattern of covariation. The two causal structures, $S_{0}$ and $S_{1}$, are depicted graphically in Figure 2. While there exists a causal arrow from $C$ to $E$ in $S_{1}$, which indicates that $C$ is a general cause of $E$, there is no causal arrow between $C$ and $E$ in $S_{0}$. Both models assume a background cause $A$.

The core principle of SISC can be illustrated with Figure 1. Assume someone suggests that $S_{0}$ is the causal structure that underlies the results. Under this hypothesis all observed cooccurrences of $C$ and $E$ would be mere coincidences. Yet, since the observed distribution of the events appears very unlikely to be coincidental, $S_{0}$ is weakened as an explanation while the alternative hypothesis, $S_{1}$, is proportionally strengthened. In fact, the probability computed by SISC for
$S_{1}$ for the data shown in Figure 1 (i.e., the posterior probability of $S_{1}$ ) is almost 1 . Now, imagine the same study had been conducted with a sample of merely eight mice but that $P(e \mid c)$ and $P(e \mid \neg c)$ remain the same. In this case, it seems less certain that $S_{1}$ underlies these results. Smaller samples not only increase uncertainty about the underlying causal structure, they also impede the reliable estimation of the size of parameters. SISC is sensitive to both types of uncertainty when estimating $P(c \xrightarrow{\text { sing. }} e \mid c, e)$.

SISC implements different steps. First, it derives the posterior probabilities for each causal structure illustrated in Figure 2. Applying Bayes' rule, the posterior probability for a causal structure is proportional to the likelihood of the data given the causal structure, weighted by the structure's prior probability:

$$
\begin{equation*}
P\left(S_{i} \mid D\right) \propto P\left(D \mid S_{i}\right) \cdot P\left(S_{i}\right) \tag{4}
\end{equation*}
$$

$P\left(D \mid S_{i}\right)$ is the likelihood of the data given a particular structure, which is the integral over the likelihood function of the parameter values under the particular structure. $P\left(S_{i}\right)$ represents a structure's prior probability. The model initially assumes that both structures are equally likely, that is, $P\left(S_{i}\right)$ $=1 / 2$. When data become available, the posterior for each causal structure varies systematically with the observed contingency: the higher the contingency, the more likely $S_{1}$ becomes.

Next, the model estimates the parameters $b_{c}, w_{c}$, and $w_{a}$, for each causal structure. To express parameter uncertainty, distributions rather than point estimates are inferred. The posterior probability distributions for the parameters, $P(w \mid D)$, are proportional to the likelihood of the data given the set of parameters $w$, weighted by the prior probability distributions of the parameters:

$$
\begin{equation*}
P(w \mid D) \propto P(D \mid w) \cdot P(w) . \tag{5}
\end{equation*}
$$

$P(D \mid w)$ is the likelihood of the data given the parameter values for $b_{c}, w_{c}$, and $w_{a} . P(w)$ is the prior joint probability of the parameters. The prior distributions of the parameters are independently set to flat, uninformative $\operatorname{beta}(1,1)$ distributions. Since $C$ does not cause $E$ under $S_{0}, w_{c}$ is held fixed at 0 for this causal structure.

In the last step, SISC computes $P(c \xrightarrow{\text { sing. }} e \mid c, e)$ for each parameterized structure. The new discounting parameter alpha is set based on background assumptions about the target scenario. For the scenarios we used in the present experiments it is set to 1 because preemption seems to be highly probable. As all co-occurrences of $c$ and $e$ are coincidences under $S_{0}, P(c \xrightarrow{\text { sing. }} e \mid c, e)$ is set to 0 for $S_{0}$. For $S_{1}$, Equation 3 is applied. The final output of SISC is a single estimate for $P(c \xrightarrow{\text { sing. }} e \mid c, e)$, which is obtained through integrating out the two causal structures by summing over the derived values of $P(c \xrightarrow{\text { sing. }} e \mid c, e)$ for each structure weighted by its posterior probability:

$$
\begin{equation*}
P(c \xrightarrow{\text { sing. }} e \mid c, e ; D)=\sum_{i} P\left(c \xrightarrow{\text { sing. }} e \mid c, e ; S_{i}\right) \cdot P\left(S_{i} \mid D\right) . \tag{6}
\end{equation*}
$$

## Experiment 1a

The goal of Experiment 1a was to test SISC against the CN model of causal attribution for data sets with a sufficient cause, i.e., $P(e \mid c)=w_{c}=1$, but varying base rates of the effect. Whereas the CN model predicts maximal confidence in singular causation assessments for any observed co-occurrence of $C$ and $E$ in this case, SISC predicts an interaction with the base rate under the assumption that $A$ 's causal power generally preempts the effect of $C$. The goal of Experiment 1a was to demonstrate that this predicted deviation from the CN model is expected for the conceptual reasons discussed above. To rule out uncertainty as an explanation, we used sample sizes in our data sets for which the posterior probabilities of $S_{1}$ computed by SISC are close to 1 . The predictions of the models are shown in Figure 3. We set $\alpha$ in Equation 3 to 1, which represents complete preemption of $C$ by $A$. We also considered a Bayesian variant of the CN model that has been proposed by Holyoak et al. (2010). This model is sensitive to parameter uncertainty; it uses probability distributions over the parameters instead of point estimates. As Figure 3 shows, the predictions of both variants of the CN model converge for large sample sizes because the influence of parameter uncertainty decreases.

## Methods

Participants 90 participants $\left(M_{\text {age }}=33.24, S D_{\text {age }}=12.50\right.$, 35 female) were recruited via Prolific Academic (www .prolific.ac) and received a monetary compensation of £0.60.
Design, Materials, and Procedure Three contingencies (see Figure 3) were manipulated between subjects with each participant responding to two causal test queries (general causation vs. singular causation). We included the general causation query to establish that uncertainty cannot account for the predicted pattern of singular causation ratings. The task was a standard elemental causal induction task. As cover story we used the gene expression scenario (cf. Griffiths \& Tenenbaum, 2005) mentioned above: subjects were asked to assume that they were biologists who are interested in whether a particular chemical causes the expression of a particular gene in mice. Subjects read that they will be asked to conduct an experiment on the computer screen in which they will treat a random sample of mice with the substance while a control sample will remain untreated. It was mentioned that the control sample is important as some individuals may show the gene expression for other reasons.

Participants were presented with an interactive animation showing the two samples arranged as in Figure 1, and a pipette containing a reddish chemical substance. All mice had gray color in this animation. Participants then dropped the substance into the test group area, whereupon the background color changed to a light red. On the next screen, subjects checked the results of the experimental manipulation by dragging a small magnifying glass over all the mice. Mice with the gene then became blue and those without became yellow. The final state of the animation looked like Figure 1.


Figure 3: Model predictions and results of Experiments 1a and b. The results show mean ratings and $95 \%$ bootstrapped CIs. Dark bars show general causation judgments; light bars singular judgments.

Subsequently, participants responded to two test questions. The general causation query referred to the causal structure. Participants were asked to indicate on a slider how confident they were that the chemical has an effect on the expression of the gene (from "very certain that the chemical has no effect" to "very certain that the chemical has an effect"). The singular causation query asked subjects about Mouse \# 25 from the test group. Participants were asked to indicate on a slider how confident they were that it was the chemical substance that caused the expression of the gene in this single case (from "very certain that it was not the chemical" to "very certain that it was the chemical").

## Results and Discussion

Figure 3 shows the results. The prediction for general causation responses corresponds to the posterior probability of $S_{1}$ computed by SISC. As predicted by the posterior probability of $S_{1}$, all general causation ratings were high, indicating very little uncertainty about the general causal structure. The singular causation ratings, by contrast, decreased with an increasing base rate of the effect, as predicted by SISC but not by the two CN models. The results of a multilevel model analysis revealed significant main effects for type of causal query, $\chi^{2}(1)=32.45, p<.001$, as well as contingency, $\chi^{2}(1)=12.63, p<.01$. General causation ratings were, on average, higher than singular causation ratings. Figure 3 shows that the main effect of contingency is driven by the decrease in singular causation ratings. Planned contrasts revealed that the general causation ratings neither differed between the first and second contingency, $t(80)=0.60$, nor between the second and third contingency, $t(80)=0.13$. Consequently, the interaction effect of query $\times$ contingency was also significant $\chi^{2}(1)=13.10$, $p<.01$. Planned contrasts breaking down this interaction effect showed that the difference between general and singular causation ratings was higher for the second than for the first contingency, $t(80)=2.10, p<.05, r=.23$, and also higher for the third compared to the second contingency, $t(80)=3.70$, $p<.001, r=.38$. In sum, both the trends for general as well as for singular causation ratings are captured well by SISC.

The observed trend for the singular causation judgments is, however, neither predicted by the CN model using point estimates nor by the Bayesian extension incorporating parameter uncertainty.

## Experiment 1b

The goal of Experiment 1 b was to assess how likely participants think it is that a particular individual from the test group already exhibited the effect caused by the background factors prior to the occurrence of the cause. Thus, instead of singular causation judgments for a particular individual, we asked subjects to provide a probability judgment. Crucially, responses to this query provide us with an estimate of the $\alpha$ value in Equation 3 that participants assumed.

## Methods

Participants 88 participants $\left(M_{\text {age }}=31.22, S D_{\text {age }}=10.84\right.$, 42 female) participated in this only study and received a monetary compensation of $£ 0.60$.
Design, Materials, and Procedure The study design and the materials were the same as in Experiment 1a. The only difference was that, instead of a singular causation judgment for Mouse \#25, we asked participants how likely they think it is that this individual already had the gene expressed prior to the experiment. The general causation query remained the same.

## Results and Discussion

Figure 3 shows that we replicated the pattern for general causation judgments found in Experiment 1a. Planned contrasts revealed that these ratings did not differ (all $t$ values $<1$ ). However, the probability judgments about the presence of the effect prior to the application of the chemical in the single case showed the opposite trend as the singular causation judgments in Experiment 1a. This finding supports our hypothesis that assumptions about preemption influence singular causation judgments, as predicted by Equation 3. Planned contrasts confirmed that ratings increased from the first to the second, $t(71)=2.67, p<.01, r=.30$, and also from the second to the third contingency, $t(71)=3.16, p<.01, r=.35$. Furthermore, the results indicate that participants indeed assumed high $\alpha$ values.

## Experiment 2

Experiment 1a showed that singular causation ratings for sufficient causes deviate systematically from the predictions of the CN models. This deviation is predicted as a consequence of assumptions about preemption relations between $C$ and $A$. Experiment 2 pursued two main goals: first, we aimed to test SISC using a larger set of contingencies with a combination of different levels of $P(e \mid c)$ and $P(e \mid \neg c)$. Second, we wanted to demonstrate that parameter and structure uncertainty indeed influence general and singular cause judgments. We used the set of contingencies studied in Buehner, Cheng, and Clifford (2003) but excluded the one contingency from the set in which the effect never occurs. It does not make sense to ask for singular causation if the effect is absent. The data sets and model predictions are shown in Figure 4. We set


Figure 4: Predictions of different models and results (means and within-subjects adjusted $95 \%$ CIs) of Experiment 2. Graphs (a) and (b) refer to general causation assessments. All other graphs refer to singular causation assessments.
the discount parameter $\alpha$ to 1 again.

## Methods

Participants 82 participants $\left(M_{\text {age }}=34.41, S D_{\text {age }}=10.42\right.$, 31 female) participated in this online study and were paid $£ 1.00$ for their participation.
Design, Materials, and Procedure The causal query (general causation vs. singular causation) was manipulated between subjects, whereas contingency was varied within subject. The fourteen contingency data sets were presented in random order. We used the same cover story as in Experiment 1 , except that subjects read that they will investigate the effects of fourteen different chemicals on fourteen different genes in fourteen different samples. We pointed out that the results of the studies are independent of each other. The assignment of mice to the cells of the contingency tables was randomly determined. Also the test mouse for the singular query showing both $c$ and $e$ was randomly chosen prior to the experiment.

## Results and Discussion

Figure 4 shows the results and the predictions of the different models: (a) and (b) display the predictions of SISC for general causation and the mean general causation responses. Panels (c) and (d) show the predictions of SISC and the results regarding the singular causation queries. Predictions of the standard CN model and its Bayesian variant are displayed in (e) and (f). Graph (g) shows predictions of SISC when $\alpha$ is set to zero but both structure and parameter uncertainty are incorporated. Finally, (h) shows point estimates of Equation 3 while neglecting statistical uncertainty.

Table 1: Model comparisons for singular causation judgments in Experiment 2. $\Delta P$ refers to the different contingency levels (.00, $.25, .75,1.00$ ) within the whole data set; $r_{\Delta P}$ expresses the model fits for these levels. $N / A$ represents undefined values.

| Fit measure | SISC | CN Model | Bayesian CN Model | SISC CN Model | Point Est. Eq. 3 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $r_{\Delta P=.00}$ | $\mathbf{. 7 2}$ | N/A | -.78 | -.68 | N/A |
| $r_{\Delta P=.25}$ | $\mathbf{1 . 0 0}$ | .21 | .38 | -.61 | .98 |
| $r_{\Delta P=.50}$ | $\mathbf{8 8}$ | .68 | .79 | .44 | .85 |
| $r_{\Delta P=.75}$ | 1.00 | N/A | 1.00 | -1.00 | 1.00 |
| $M_{r \Delta P}$ | .90 | .44 | .35 | -.46 | .94 |
| $r_{\text {overall }}$ | $\mathbf{. 9 4}$ | .88 | .93 | .90 | .90 |
| $R^{2}$ | $\mathbf{. 8 8}$ | .77 | .87 | .82 | .82 |
| $R M S E$ | .11 | .27 | $\mathbf{. 0 9}$ | .14 | .22 |

The overall pattern for both general and singular causation ratings was captured best by the revised version of SISC. As in our previous research, the results show that participants differentiated between general and singular queries. Moreover, the responses to the general causation query replicate those found in Griffiths and Tenenbaum (2005). Most importantly, the singular causation assessments were captured best by the revised SISC. The other models, by contrast, struggled to account for the local trends observed within the subsections of the contingency set in which $\Delta P$ is constant. The difference between the singular causation ratings and the point estimates for the revised causal attribution measure (Equation 3) implies that participants were sensitive to structure and parameter uncertainty.

A multilevel model analysis confirmed the main effect of contingency, $\chi^{2}(13)=1010.04, p<.001$, as well as the interaction between contingency $\times$ query, $\chi^{2}(13)=49.39$, $p<.001$, that is shown in Figure 4. To test the different models, we computed different fit measures shown in Table 1. As can be seen there, SISC achieved a good fit in the overall fit measures (bottom part of the table). It explained most variance, with $R^{2}=.88$, and yielded the second smallest $R M S E$ of .11 . Yet, all models obtained relatively high values on the global measures. Even the CN model with the lowest overall fit accounted for 77 percent of the variance. The similarity between the models is not unexpected, however, as all models are sensitive to $\Delta P$. More interesting are the fit measures for the subsections of the contingency set in which $\Delta P$ is kept constant. The upper part of Table 1 shows that SISC yielded high fit values there, too, and hence accounted well for these local trends, whereas the Bayesian CN model, which yielded the smallest RSME, even showed negative correlations here.

## General Discussion

We addressed two different problems that the power PC framework of causal attribution (Cheng \& Novick, 2005) faces: first, the CN model attributes causal responsibility for the occurrence of a particular effect $e$ to a present singular event $c$ whenever its causal power is sufficient to bring about the effect. We have argued that this conceptualization fails to take into account that people make assumptions about the point in time at which different causal powers exert their influences. Not every manifestation of a sufficient cause $c$ needs to be causally responsible for an observed outcome; it might be the case that a competing cause (e.g., a) preempts it. This problem of redundant causation, which occurs whenever two causes are individually sufficient for the effect, is widely acknowledged in the philosophical literature as a challenge for models of causation (see, e.g., Paul \& Hall, 2013). To account for the possibility of preemption we have modified the equation developed by Cheng and Novick (2005) as an account of causal attribution. The revised equation includes the discount parameter $\alpha$ that can be set to express domain-related assumptions about the temporal relations between the alternative causal factors. A second shortcoming of the standard causal attribution model (Cheng \& Novick, 2005) is that it
does not take into account statistical uncertainty about structure and causal parameters (cf. Griffiths \& Tenenbaum, 2005; Meder et al., 2014). Our model SISC remedies both shortcomings. It is sensitive to both the temporal relations between the alternative causes and to statistical uncertainty. Our experiments showed that both aspects are important to account for subjects' judgments about singular causation.

We have set the discount parameter $\alpha$ to 1 in Equation 3 which implies a complete preemption relation between $A$ and $C$ whenever A's causal power is sufficient in a situation. Better fits might be possible by estimating the size of $\alpha$ for each individual subject separately. We avoided this strategy to demonstrate that model improvements can already be achieved with very general assumptions. The goal of future experiments will be to manipulate the size of $\alpha$ by manipulating domain assumptions about the temporal relations between $C$ and $A$. Cases in which $\alpha$ is 1 are situations in which $A$ always preempts $C$. The cover stories used in the present experiments are an example in which it is plausible to assume that $A$ represents a temporally stable factor that has already been efficacious prior to the manipulation of $C$. Although preemption seems to be the default situation in most singular causation scenarios, there might be rare cases in which other assumptions need to be made. Consider cases of symmetric overdetermination that have also been discussed in the literature (see Paul \& Hall, 2013): in the famous firing squad scenario, for example, in which each shooter is a sufficient cause for the death of the target, a possible intuition is that each shooter should be counted as a singular cause of the death of the victim. In this case, alpha would have to be set to zero. Similarly, alpha would have to be set to zero if $C$ preempts $A$ so that $A$ cannot manifest its potential causal power. Cases of temporal variability between $C$ and $A$ might also be an interesting topic for future studies.
Acknowledgments We thank Jonas Nagel and Ralf Mayrhofer for helpful discussions.

## References

Buehner, M. J., Cheng, P. W., \& Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(6), 1119.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104(2), 367-405.
Cheng, P. W., \& Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). Psychological Review, 112, 694-706.
Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), The Oxford handbook of causal reasoning (pp. 201-215). New York: Oxford University Press.
Griffiths, T. L., \& Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive Psychology, 51, 334-384.
Holyoak, K. J., Lee, H. S., \& Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. Journal of Experimental Psychology: General, 139, 702-727.
Meder, B., Mayrhofer, R., \& Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. Psychological Review, 121, 277-301.
Paul, L. A., \& Hall, E. J. (2013). Causation: A user's guide. Oxford University Press.
Stephan, S., \& Waldmann, M. R. (2016). Answering causal queries about singular cases. In A. Papafragou, D. Grodner, D. Mirman, \& J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 2795-2801). Austin, TX: Cognitive Science Society.

# Marbles in Inaction: Counterfactual Simulation and Causation by Omission 

Simon Stephan ${ }^{1}$ (sstepha1@gwdg.de) Pascale Willemsen ${ }^{2}$ (pascale.willemsen@rub.de)<br>Tobias Gerstenberg ${ }^{3}$ (tger@mit.edu)<br>${ }^{1}$ Department of Psychology, Georg-August-University Göttingen and Leibniz ScienceCampus Primate Cognition<br>${ }^{2}$ Institute for Philosophy II, Ruhr-University Bochum<br>${ }^{3}$ Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology


#### Abstract

Consider the following causal explanation: The ball went through the goal because the defender didn't block it. There are at least two problems with citing omissions as causal explanations. First, how do we choose the relevant candidate omission (e.g. why the defender and not the goalkeeper). Second, how do we determine what would have happened in the relevant counterfactual situation (i.e. maybe the shot would still have gone through the goal even if it had been blocked). In this paper, we extend the counterfactual simulation model (CSM) of causal judgment (Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2014) to handle the second problem. In two experiments, we show how people's causal model of the situation affects their causal judgments via influencing what counterfactuals they consider. Omissions are considered causes to the extent that the outcome in the relevant counterfactual situation would have been different from what it actually was.


Keywords: causality; counterfactuals; causation by omission; causal attribution; mental simulation.

## Introduction

Billy is on his way home. He is driving on a lonely country road, when he notices a damaged car next to the road. The car seems to have collided with a tree, and the driver appears unconscious. Billy decides not to stop and keeps driving. A few days later, Billy reads in the newspaper that the driver died because he had not received any medical attention.

Many people would concur that Billy's not having stopped was causally relevant for the driver's death. However, there are two fundamental problems with citing omissions (i.e., events that did not happen) as causes. First, there is the problem of causal selection. Why cite Billy's not stopping as causally relevant for the driver's death? Why not cite the Queen of England? Second, there is the problem of underspecification. Assuming that Billy would have stopped to check on the driver, what would he have done? Would Billy's acting have prevented the driver's death, or would she have died anyway?

In this paper, we show how the counterfactual simulation model (CSM) of causal judgment developed in Gerstenberg, Goodman, Lagnado, and Tenenbaum (2012) (see also Gerstenberg et al., 2014; Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2015) provides a natural solution to the underspecification problem. The CSM predicts that an omission is a cause when the positive event that is chosen as its replacement would have changed the outcome of interest. More specifically, we show how people's causal model of a situation guides their selection of the relevant counterfactual which subsequently determines their judgment about whether the omission made a difference to the outcome.

The paper is organized as follows: We first describe the causal selection and the underspecification problem in more detail. We then propose an extension to the CSM as a solution to the underspecification problem. Thereafter, we present and discuss the results of two experiments which test the CSM.

## The Causal Selection Problem

Many philosophers argue that counterfactual approaches to causation are too inclusive when it comes to omissions (e.g. McGrath, 2005). If Billy had stopped and checked on the unconscious driver, the driver would not have died. Consequently, the driver died because Billy did not stop. However, following this logic, the same counterfactual seems to be true for the Queen of England. If the Queen of England had stopped, the driver would not have died either. However, intuitively, it is Billy's omission that was causally relevant, and not the Queen's. The problem of causal selection has been intensively discussed in both philosophy and empirical studies (e.g. Hesslow, 1988). Interestingly, while the causal selection problem presents a challenge to certain philosophical theories of causation, laypeople do not have any difficulty in selecting the cause of the driver's death. Based on evidence from research on causal cognition, it has been suggested that the concept of causation is not a purely descriptive one, but that it depends on reasoners' expectations (Willemsen, 2016). While we would have expected Billy to stop and help, we didn't entertain any such expectation for the Queen.

## The Underspecification Problem

When it comes to omissive causation a fundamental problem is how to define the relevant counterfactual contrast (cf. Schaffer, 2005). For positive events ("something happened"), the counterfactual contrast ("it didn't happen") is often welldefined. However, replacing a negative event with a positive event seems more problematic because there are a infinitely many ways in which events can come about. If Billy actually helped the driver, it seems to be pretty clear what would have happened if he had not helped (he would just have continued to drive on). However, if Billy did not help, it is unclear what would have happened if he had helped (would he have helped in a competent manner to prevent the driver's death, or would he have been too nervous and screwed things up?).

While the causal selection problem has received much attention in the literature (e.g., Henne, Pinillos, \& De Brigard, 2015; Livengood \& Machery, 2007), the underspecification problem has not. One exception is the account by Wolff, Barbey, and Hausknecht (2010) that addresses both problems. The general idea proposed by Wolff et al. (2010) is that cau-
sation by omission is linked to the removal of an actual (or anticipated) force that previously prevented a certain outcome from occurring. One problem of this account, however, is that it appears too restrictive in that it cannot account for cases in which no (apparent) force is removed. Imagine, for instance, sentences like "The lack of rain caused the drought in Somalia". Here, it would be a stretch to think of a the lack of rain as the removal of a force.

The extension of the CSM that we propose in this paper provides a different solution to the underspecification problem. Previous research has suggested that the extent to which a certain counterfactual is relevant is a function of both how likely we are to consider it, and how likely it would have changed the outcome of interest (Petrocelli, Percy, Sherman, \& Tormala, 2011). However, while this research has shown that these counterfactual probabilities affect people's causal judgments, it doesn't explain how we come up with the relevant probabilities in the first place. Here, we will show how the CSM provides a natural solution to determine whether an omission made a difference to the outcome.

## Counterfactual Simulation and Omission

The CSM predicts that people make causal judgments by comparing what actually happened with the outcome of a counterfactual simulation. So far, the model has been applied to capturing participants' judgments about events that actually happened (Gerstenberg et al., 2012, 2014, 2015). Consider the situation shown in Figure 1b (bottom) illustrated as the ideal path. Here, A collides with B and B subsequently goes through the gate. The CSM says that ball A's colliding with ball B caused ball B to go through the gate in this case, because it is obvious that ball B would have missed the gate but for the collision with A. More generally, the CSM predicts that causal judgments are a function of the reasoner's subjective degree of belief that the candidate cause made a difference to the outcome. More formally, we can express the degree of belief that $x$ caused $y$ as

$$
\begin{equation*}
P(x \triangleright y)=P\left(y^{\prime} \neq y \mid S, \operatorname{do}\left(x^{\prime}\right)\right), \tag{1}
\end{equation*}
$$

in which $x$ denotes the event of ball A hitting ball B , and the outcome $y$ captures the event of ball B going through the gate. We first condition on what actually happened $\mathcal{S}$ (i.e., the motion paths of each ball, the position of the walls, etc.). We then intervene to set the candidate cause event $x$ to be different from what it was in the actual situation, do $\left(x^{\prime}\right)$. Finally, we evaluate the probability that the outcome in this counterfactual situation $y^{\prime}$ would have been different from the outcome $y$ that actually happened. The results of several experiments (cf. Gerstenberg et al., 2012, 2014, 2015) have revealed that there exists a tight relationship between the counterfactual judgments of one group of participants (about what would have happened if the candidate cause had been absent), and the causal judgments of another group of participants.

To model causal judgments about positive events, the CSM considers counterfactuals in which the positive event (ball A colliding with $B$ ) is simply removed from the scene (indicated


Figure 1: Illustration of what actually happened (top) and the counterfactual simulation model (bottom). The diagrams illustrate the actual path that ball B took, as well as an ideal path for (a) A preventing $B$ from going through the gate, or (b) A causing $B$ to go through the gate. The sampled paths show example simulations that result from applying implementation noise to the ideal path. Note: In (a), A would have prevented B from going through the gate for both sampled paths. In (b), A would have caused B go through the gate in one sample but not so in the other in which $B$ would still have missed even though A hit B.
by $\operatorname{do}\left(x^{\prime}\right)$ in Equation1). Things become more intricate, however, when we want to model omissions as causes. As discussed above, it is often straightforward to replace an event with a non-event (e.g., a collision with no collision), but it is less clear how to replace a non-event with an event. Consider the situation shown in Figure 1a. Did ball B go through the gate because ball A did not hit it? The problem is that there are infinitely many ways for ball A to collide with ball B. Which of these events are we to consider? The collision event is severely underspecified. We will now show how the CSM can be extended to yield predictions about omissions as causes, and thereby provide a solution to the underspecification problem.

## Modeling Omissions

We assume that people solve the underspecification problem by sampling counterfactual possibilities based on their intuitive understanding of the situation (cf. Kahneman \& Tversky, 1982). The extent to which the omission is viewed as a cause of the outcome is assumed to be a function of the proportion of samples in which the outcome would have been different from what actually happened, assuming that the type of counterfactual event of interest was realized. Let us illustrate how the model works by example of the situation depicted in Figure 1a. In the actual situation, ball A did not move and ball B went right through the middle of the gate. We want to determine to what extent A's not hitting ball B was a cause of B's going through the gate. To do so, we simulate what would have happened if ball A had collided with B. More specifically, we need to determine the time $t$ at which A would have started to move, the direction $d$ in which ball

A would have moved, and the velocity $v$. Once we have determined these quantities, we can simulate what would have happened. For many combinations of values for $t, d$, and $v$ ball A would not have collided with ball B. We can discard all such situations since we are interested in evaluating what would have happened if ball A had hit ball B. For each situation in which the two balls collide, we record what the outcome would have been - would B have missed the gate, or would it still have gone through the gate? We can now obtain the probability that ball A's not hitting ball B was a cause of ball B's going through the gate (cf. Equation 1) by looking at the proportion of samples in which B would have missed the gate instead of going through.

But how do we determine what values to take for $t, d$, and $v$ which jointly determine what counterfactual situation we consider? We predict that prior expectations guide the counterfactuals we consider. In Experiment 1 below, we contrast situations in which participants don't have any expectations about what normally happens, with situations in which participants have statistical, or social expectations. We will now discuss how the model incorporates these expectations.

## Expectations Shape Counterfactual Simulations

No Expectations Let us first assume a situation in which an observer does not have any strong expectations concerning how the balls typically move in the given context. When asked whether A's not hitting B caused B to go through the gate, we have to generate situations in which A would have hit B. This already considerably constrains what kinds of situations we consider. For example, it would be futile to consider situations in which A only starts moving after B already went through the gate, or in which A moved toward the right.

We generated counterfactual samples in the following way: We first discretized the space for the time at which A starts moving $t$, the direction in which it moves $d$, and its velocity $v$. For $t$, we considered all values from 0 to $t_{\text {outcome }}$ where 0 corresponds to the time at which B starts moving and $t_{\text {outcome }}$ to the time at which ball B went through the gate (or hit the wall). For $d$, we considered the full range from A going straight to the left to going straight up. For $v$, we considered a reasonable range from A moving slowly to A moving fast. For each generated world, we noted whether A and B collided, and whether B went through the gate or missed the gate. We then discarded all situations in which the two balls did not collide, and recorded the proportion of situations in which B would have gone through the gate if the balls had collided.

The model makes the following predictions: For the situation in which $B$ is on a path toward the gate (Figure1a), there is a good chance that B would have missed the gate if ball A had hit it. The model predictions are shown in Figure2. As can be seen in the left panel, the CSM concludes that the probability that B would have missed the gate had A hit it is just as high as the probability that B would have passed the gate. By contrast, when B is on a path away from the gate ("missed" in Figure 2, cf. Figure 1 b top right) there is only a
relatively small chance that ball B would have gone through the gate if ball A had hit it. Thus, the CSM predicts that people will be more likely to agree that ball B went through the gate because ball A did not hit it than they will be to agree that ball B missed the gate because ball A did not hit it.
Social Expectations When nothing particular is known about how A and B typically move, the space of counterfactuals from which the CSM samples is relatively wide. It seems plausible, however, that what counterfactual possibilities are considered will be affected by different forms of prior expectations. Imagine, for example, that you learn that two players play a marble game. Player B wants to get her marble into the goal, while Player A wants to make sure that this does not happen. On a particular trial, Player A did not pay attention and forgot to flick his marble. Did Player B's marble go through the gate because Player A's marble did not hit it? When knowing that it is a player's job to prevent a marble from going through the gate, people may expect that this player would not have just flicked her marble randomly. Instead, she can be expected to try her best to make sure that the other marble does not go through the gate. Similarly, consider a situation in which Player A also wants that Player B's marble goes through the gate. In that case, it seems likely that Player A will try to flick his marble so that it makes sure that B's marble will go through the gate.

Figure 1 illustrates how the CSM incorporates how prior expectations constrain the space of counterfactual situations. We assume that the player would first determine a time $t$ at which to flick her marble. For any given point $t$, the player then determines an optimal $d$ and $v$ conditional on the player's goals. For a player who wants to prevent ball B from going through the gate, the player's goal is to maximize the distance between B's position and the middle of the gate. For a player who wants to cause B to go through the gate, the player's goal is to minimize the distance between B's position and the middle of the gate (i.e., she wants B to go right through the middle of the gate). For simplicity, we assume that players can plan their action optimally, but that they have some implementation noise. The CSM models this implementation noise by introducing a small perturbation to the ideal path on which A moves. As is illustrated in Figure1, the CSM incorporates implementation noise by slightly perturbing the "ideal path" vector.

Figure 1 shows the actual path that ball B took, the ideal paths that player A "wanted" the marbles to take, and two examples for paths that ball B actually took after subjecting A's ideal plan to some implementation noise. Notice that the implementation noise has a larger effect in Figure 1b where it leads to a situation in which ball B would have missed the gate even though ball A hit it. In contrast, in Figure 1a the implementation noise has less of an effect. Here, ball B would reliably miss the gate even if we apply some implementation noise to player A's intended plan. Accordingly, the CSM predicts that it is more likely that A's hitting B would have resulted in B missing the gate (when B actually went through,

Figure 1a) than it would have resulted in $B$ going through the gate (when B actually missed, Figure 1b). Since the sample of considered situations is biased toward optimal actions, the CSM predicts that judgments will overall be higher than when an observer does not have any prior expectations. The predictions for this situation are shown in the middle panel in Figure 2.
Statistical Expectations Now imagine that instead of learning anything about agents playing a game you get to see a few situations first that shape your expectations about what tends to happen. We incorporate such "statistical" expectations into the model in the same way in which we handled social expectations. However, we allow for the implementation noise to be different between these situations. Specifically, the size of the implementation noise parameter will depend on the kind of evidence that participants have seen. For example, if one has witnessed a series of trials in which A always hit B in such a way that B went straight through the gate, this would suggest a smaller implementation noise compared to one that is suggested by trials in which A hit B in such a way that B went through the gate in, for example, merely two third of the cases. The predictions for this situation are shown in the right panel in Figure 2.

## Experiment 1

Experiment 1 tests whether the CSM accurately predicts people's causal judgments for omissions in dynamic physical scenes. We look at causal judgments about situations in which ball A failed to hit ball B, and ball B either went through or missed the gate (see Figure 1). In line with the CSM, we predict that the degree to which people judge ball A's not hitting ball B as causally relevant to the outcome would be tightly coupled with the results of a mental simulation about what would have happened if a collision had occurred. Furthermore, we test the hypothesis that different types of expectations (social or statistical) influence people's causal judgments by affecting what counterfactual situations people consider.

## Methods

Participants and Materials 476 participants ( 239 female, $M_{\text {Age }}=33.83$ years, $S D_{\text {Age }}=12.03$ years) were recruited via Prolific Academic (www.prolific.ac) and participated in this experiment for a monetary compensation of $£ 0.25$. The clips were created in Adobe Flash CS5 using the physics engine Box2D.
Design and Procedure All factors were manipulated between subjects. We manipulated what actually happened (actual outcome: missed vs. went through), and the expectations of participants about what will happen (expectation: no expectations, statistical expectation, social expectation). Finally, we varied whether participants answered a causal question, or a (counterfactual) probability question (question: causation vs. probability).

In the "no expectations" condition, subjects simply read that they will see an animation in which a stage with solid


Figure 2: Experiment 1. Mean causal and probability judgments together with the predictions of the CSM. Note: Error bars indicate $95 \%$ bootstrapped CIs.
walls, two balls A and B, and a gate will be displayed. All subjects were shown a graphical illustration of the stimuli. Participants in the "statistical expectation" condition were presented four primer clips in which ball B actually collided with A . One group of subjects saw that the collision always caused B to go through the gate, while the other half always saw that A prevented B from going through the gate (see Figure 1). In the "social expectation" condition, subjects were instructed that the video clip (which was the same as in the "no expectations") shows what happened during a game of marbles played by two agents, Andy and Ben. We manipulated whether subjects believed that Andy wants to help Ben to flip his marble through the gate or whether he wants to hinder Ben from doing so.

Participants in the "causation" condition indicated how much they agreed with the claim that B missed the gate because A did not hit it, or that B went through the gate because A did not hit it, depending on the outcome. Participants in the "probability" condition gave a corresponding probability judgment: they indicated what they believed the chances were that B would have gone through / missed the gate if ball A had hit ball B. Participants indicated their ratings on a sliding scale.

Which outcome participants saw depended on the expectation condition: In the "social expectation" condition, participants who expected the agent to help saw that B actually missed the gate, and participants who expected the agent to hinder saw that B went through the gate. In the "statistical expectations" condition, participants who had seen the causation clips saw that B missed the gate, whereas those who had seen the prevention clips saw that B went through the gate.

## Results and Discussion

Figure 2 shows participants' mean causal ratings (white bars), probability ratings (gray bars), as well as the predictions of the CSM (black bars). The CSM correctly predicts a difference in agreement ratings for both the causal and probability condition as a function of the outcome (went through vs. missed). A global 2 (question) $\times 6$ (combination of expectation and outcome) factorial ANOVA shows a main effect of outcome, $F(5,464)=14.51, p<.001, \eta_{G}^{2}=.61$ but no main effect of question, $F(1,464)<1$. The interaction between question and expectation was significant,
$F(5,464)=2.74, p<.05$ but the effect is small, $\eta_{G}^{2}=.03$.
Importantly, participants saw A's not hitting ball B as more causal when B went through the gate compared to when it missed. This pattern was predicted by the CSM and indicates that participants' counterfactual simulations and their causal inferences were sensitive to the constraints imposed by the virtual physical environment. Because the displayed gate was relatively small, the probability that a collision would change the outcome is higher if B actually went though, than when it missed. Planned contrasts confirmed that the observed differences between "went through" and "missed" were significant in all conditions, with $t(464)=3.21, p<.01, r=.15$ in the "no expectations" condition, $t(464)=2.13, p<.05, r=.10$ in the "statistical expectation" condition, and $t(464)=3.53$, $p<.001, r=.16$ in the "social expectation" condition.

Besides the asymmetry between "went through" and "missed", we also expected to see higher causality ratings in the "statistical" and "social expectation" conditions than in the "no expectation" condition. This difference was predicted because we incorporated an ideal path in these situation that was then perturbed by imposing some implementation noise. As Figure 2 shows, we did indeed observe this pattern. A planned contrast confirmed that this difference was significant, $t(464)=5.98, p<.001, r=.27$.

Concerning the probability ratings, planned contrasts showed that the difference between "went through" and "missed" was significant in the "no expectations condition", $t(464)=2.33, p<.05, r=.11$, and the "statistical expectation" condition, $t(464)=1.73, p<.05, r=.08$, but not in the "social expectation" condition, $t(464)<1$. Concerning the predicted difference between the "no expectations" condition and the other two expectation conditions, Figure 2 shows that we obtained a similar pattern as for the causality judgments. In line with our expectations, the probability ratings for the "statistical expectation" and the "social expectation" condition were higher than the ratings for the "no expectation" condition, $t(464)=2.82, p<.01$, though this effect was smaller than the effect for the causality judgments, $r=.13$.

The results of Experiment 1 show that participants' causal judgments are qualitatively well accounted for by the CSM. The CSM also does a good job in accounting for the pattern quantitatively, as evidenced by a high correlation between model predictions and counterfactual probability judgments ( $r=.97, R M S E=14.00$ ), as well as between model predictions and causal judgments ( $r=.97, R M S E=6.06$ ). The fact that the model accounts slightly less well for the counterfactual probability judgments is mainly due to the relatively large difference between model predictions and probability judgments in the "no expectations" condition.

A key finding in Experiment 1 is the asymmetry in participants' causal judgments as a function of whether ball $B$ went through or missed the gate. The CSM predicts this pattern because it is more likely that A's hitting B would prevent B from going through the gate (cf. Figure 1a) than that it would cause


Figure 3: Illustration of the materials used in Experiment 2. Solid arrows indicate the actual path of the ball; dashed arrows show the hypothetical path of the wall. Graph (a) shows the "went through" and (b) the "missed" condition. The results for both conditions are included in brackets.

B to go through (cf. Figure 1b). One possibility, however, that Experiment 1 cannot rule out is that people are in general more likely to regard omissions as causes when the relevant counterfactual involves preventing compared to causing. In Experiment 2, we investigate whether there is such a general asymmetry between omissive causation and prevention.

## Experiment 2

The goal of Experiment 2 was to rule out that the observed difference between "went through" and "missed" in Experiment 1 came about because people generally treat omissive causation and omissive prevention differently. The CSM only predicts an asymmetry between two situations when the positive event of interest was more likely to make a difference in one situation compared to the other. Hence, our strategy in Experiment 2 was to hold this probability constant. To achieve this goal, we simply replaced ball A with a wall that had exactly the size of the gate. To model "missed" and "went through", we varied whether the wall blocked the gate or not, while a displayed ball always headed toward the gate (see Figure 3). Participants rated how much they agree that "the ball" missed the gate (or went through the gate) because the wall did not move. There is no ambiguity about the relevant counterfactual in this case - it is clear that the outcome would have been different, had the wall moved. Accordingly, the CSM predicts that participants' judgments should be high for both cases, no matter whether the ball went through the gate or missed the gate because of the omission.

## Method

Participants 65 participants ( 40 female, $M_{\text {age }}=32.86$, $S D_{\text {age }}=12.84$ ) who were again recruited via Prolific Academic completed this online experiment and received a monetary compensation of $£ 0.25$.
Design, Materials, and Procedure The final outcome, that is, whether the ball went through or missed the gate (see Figure 3) was manipulated between subjects. The instructions were similar those used in the "no expectations" condition in Experiment 1. Further, participants were presented an illustration showing the materials in which it was made clear that the wall can only be in two different positions, either right
in front of the gate or in the upper left corner of the stage (see Figure 3). Having read the instructions, participants were shown the respective video clip and provided the causal rating after the clip was finished.

## Results and Discussion

As expected, participants gave very high causal ratings for "went through" ( $M=87.51, S D=21.62$ ) and "missed" ( $M=89.00, S D=23.21$ ). As predicted by the CSM, the ratings were not different from each other, $t(63)<1$. The probability that the outcome would have been different in the relevant counterfactual, is close to maximal in both conditions.

The results of Experiment 2 are in line with the CSM. Further, the fact that the causality ratings were both very high and not different from each other rules out the potential alternative explanation that people might generally treat omissive causation and omissive prevention differently.

## General Discussion

We developed an extension of the Counterfactual Simulation Model to account for causation by omission. Based on previous research by Gerstenberg et al. (2014), we reasoned that people's causal judgments are closely linked to their subjective degree of belief that the outcome would have been different had the candidate cause been replaced. We argued that this replacement by a counterfactual contrast is particularly difficult in cases of omissions. The counterfactual contrast to "did not hit" is clearly "had hit", but it remains unclear what would have happened if "hitting" had taken place.

In two experiments we shed light on how to tackle the underspecification problem. We predicted that prior expectations would constrain what counterfactual contrasts people consider relevant to the scenario. Experiment 1 revealed an asymmetry: A's not hitting $B$ was judged less causal when $B$ missed the gate compared to when $B$ went through the gate. This is what the CSM predicts, and the results thus lend additional support to the hypothesis that causal judgments are grounded in counterfactually simulated probabilities. Adding expectations increased both people's causal judgments as well as their subjective degree of belief that a counterfactual collision would changed the outcome. This effect was particularly strong for social expectations, which the CSM explains by assuming that knowledge about intentions of agents limits the range of counterfactuals that are considered. Our results thus add to previous research indicating that intentional actions signal higher causal stability compared to unintentional ones (Lombrozo, 2010), and that causal stability is indeed a relevant dimension that affects causal reasoning (Nagel \& Stephan, 2016).

It might be objected that the asymmetry in causal attribution for "went through" and "missed" in Experiment 1 is not due to a difference in what would have happened in the relevant counterfactual simulations, but rather due to an inherent asymmetry between omissions that prevent and omissions that cause. Experiment 2 addressed this possible con-
found by looking at situations in which the relevant counterfactual event was clear (a wall that could only move in one direction), as well as what would have happened in case that event had happened. Just as predicted, we found that causal ratings were equally high irrespective of whether the ball "went through" and "missed" in this case. Instead of a general asymmetry between prevention and causation, participants judge omissions to be causal the more certain they are that the omission made a difference to the outcome.

As our introductory example demonstrates, omissions are particularly relevant in human interaction, especially so in morally or legally charged situations when we had clear expectations about what a person should have done. In this paper, we have shown how the CSM accounts for people's causal judgments of omissions in situations in a physical domain in which the relevant counterfactuals are relatively well constrained. However, we believe that the CSM has the potential to capture causal judgments about omissions of social agents as well. For example, the extent to which we blame someone for not having helped depends on how easy it would have been for the agent to help (cf. Jara-Ettinger, Tenenbaum, \& Schulz, 2015). In future research, we will explore the CSM in a richer social setup.
Acknowledgments This project was supported by the Leibniz Association through funding for the Leibniz ScienceCampus Primate Cognition. TG was supported by the Center for Brains, Minds \& Machines (CBMM), funded by NSF STC award CCF-1231216.

## References

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 378-383). Austin, TX: Cognitive Science Society.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, \& B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 523-528). Austin, TX: Cognitive Science Society.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 782-787). Austin, TX: Cognitive Science Society.
Henne, P., Pinillos, Á., \& De Brigard, F. (2015). Cause by omission and norm: Not watering plants. Australasian Journal of Philosophy, 1-14.
Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), Contemporary science and natural explanation: Commonsense conceptions of causality (pp. 11-32). Brighton, UK: Harvester Press.
Jara-Ettinger, J., Tenenbaum, J. B., \& Schulz, L. E. (2015). Not so innocent: Toddlers’ inferences about costs and culpability. Psychological Science, 26(5), 633-640. Retrieved from http://dx.doi.org/10.1177/0956797615572806 doi: 10.1177/0956797615572806
Kahneman, D., \& Tversky, A. (1982). The simulation heuristic. In D. Kahneman \& A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 201-208). New York: Cambridge University Press.
Livengood, J., \& Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. Midwest Studies in Philosophy, 31(1), 107-127.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. Cognitive Psychology, 61(4), 303-332.
McGrath, S. (2005). Causation by omission: A dilemma. Philosophical Studies, 123(1), 125-148.
Nagel, J., \& Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In A. Papafragou, D. Grodner, D. Miram, \& J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 806-812). Austin, TX: Cognitive Science Society.
Petrocelli, J. V., Percy, E. J., Sherman, S. J., \& Tormala, Z. L. (2011). Counterfactual potency. Journal of Personality and Social Psychology, 100(1), 30-46.
Schaffer, J. (2005). Contrastive causation. The Philosophical Review, 114(3), 327-358.
Willemsen, P. (2016). Omissions and expectations: A new approach to the things we failed to do. Synthese. Advance online publication. doi: 10.1007/s11229-016-1284-9
Wolff, P., Barbey, A. K., \& Hausknecht, M. (2010). For want of a nail: How absences cause events. Journal of Experimental Psychology: General, 139(2), 191-221.

# A Two-Step Signal Detection Model of Belief Bias 

Rachel G. Stephens (r.stephens@unsw.edu.au)<br>School of Psychology, University of New South Wales<br>Sydney, NSW 2052 Australia

John C. Dunn (john.dunn@uwa.edu.au)<br>School of Psychological Science, University of Western Australia<br>Perth, WA 6009 Australia

Brett K. Hayes (b.hayes@unsw.edu.au)<br>School of Psychology, University of New South Wales<br>Sydney, NSW 2052 Australia


#### Abstract

When asked to assess the deductive validity of an argument, people are influenced by their prior knowledge of the content. Recently, two competing explanations for this belief bias effect have been proposed, each based on signal detection theory. Under a response bias explanation, people set more lenient decision criteria for believable than for unbelievable arguments. Alternatively, believable and unbelievable arguments may differ in subjective argument strength for both valid and invalid items. Two experiments tested these accounts by asking participants to assess the validity of categorical syllogisms and rate their confidence. Conclusionbelievability was manipulated either within- or betweengroups. A two-step signal detection model was applied to examine the effects on the relative location of the decision threshold and the distributions of argument strength. Equivalent belief bias effects were found when believability was manipulated within- and between-groups, supporting the view that the belief bias effect is due to response bias.


Keywords: belief bias; deductive reasoning; signal detection theory; response bias

## Introduction

An important phenomenon for theories of reasoning is that people show a belief bias when asked to assess the logical validity of arguments. The tendency to accept or reject a conclusion as valid is not based purely on logical structure but is also swayed by its compatibility with prior knowledge (e.g., Evans, Newstead, \& Byrne, 1993; Markovits \& Nantel, 1989; Shynkaruk \& Thompson, 2006). Table 1 shows typical stimuli - categorical syllogisms - in which the validity of the argument is crossed with the believability of the conclusion. In the validity discrimination task, participants are asked to judge whether the conclusion below the line necessarily follows from the premises above the line. Key findings based on arguments like these are that people are more likely to endorse valid than invalid arguments, but they are also more likely to endorse arguments with believable than with unbelievable conclusions. In many cases these factors also interact; for example, the difference between the acceptance rates of valid and invalid arguments is often greater for unbelievable than for believable arguments (e.g., Dube, Rotello, \& Heit,

2010; Evans, Barston, \& Pollard, 1983; Newstead, Pollard, Evans, \& Allen, 1992; Roberts \& Sykes, 2003).

Table 1: Sample syllogisms.

|  | Believable | Unbelievable |
| :---: | :---: | :---: |
| Valid | No beers are krabbers. Some krabbers are drinks. | No drinks are krabbers. Some krabbers are beers. |
|  | Some drinks are not beers. | Some beers are not drinks. |
| Invalid | No drinks are krabbers. Some krabbers are beers. | No beers are krabbers. Some krabbers are drinks. |
|  | Some drinks are not beers. | Some beers are not drinks. |

Such effects are often seen as evidence that believability affects the quality of deductive reasoning - people's ability to distinguish valid from invalid arguments (see Dube et al., 2010 for a review). Theoretical accounts such as the selective scrutiny model (Evans et al., 1983), misinterpreted necessity model (e.g., Markovits \& Nantel, 1989; Newstead et al., 1992) or the mental models approach (e.g., Oakhill, Johnson-Laird, \& Garnham, 1989) propose explanations in which believability affects how validity is evaluated.

However, deciding whether an argument is valid also involves response bias - the willingness to endorse the argument, regardless of one's ability to discriminate valid and invalid forms. Controversially, recent work has used confidence ratings and signal detection theory to show that belief bias only reflects changes in response bias. That is, people are more willing to respond "valid" for believable arguments (Dube et al., 2010; Trippas et al., 2014). In this view, believability does not change one's subjective evaluation of argument validity.

In reaction to this response bias account, it has been suggested that data patterns consistent with changes in response bias can also be explained by believability affecting the subjective strength of both valid and invalid arguments (Klauer \& Kellen, 2011; Singmann \& Kellen, 2014). Under this alternative argument strength account, if an argument has a believable conclusion (whether valid or
invalid) then it will be viewed as more logically valid and thus garner more endorsements.

In adjudicating between these accounts, a key consideration is that believability is usually manipulated within a single experimental session. The argument strength account is consistent with evidence that response bias is unlikely to change from trial to trial (e.g., Stretch \& Wixted, 1998). However, to our knowledge it is currently unknown how believability affects performance if instead it is manipulated between different groups of participants, where response bias is free to differ. As we explain below, we hypothesized that if the response bias account is correct then the same belief bias effects on model parameters should appear when believability is manipulated within groups and between groups (i.e., equivalent ordinal effects on response bias and no effects on discriminability).

Given the important implications for theories of reasoning, we aimed to extend the investigation of response bias in deductive reasoning. We took three key steps. First, we sought to replicate the within-group findings of Dube et al. (2010) - including confidence ratings - that they used to support the response bias account. Second, we applied an extended signal detection model that was specifically tailored to the two-step task in which participants first make a binary valid/invalid decision, then rate their confidence. Our goal was to confirm whether such a model would still suggest that believability does not affect accuracy, but that the response bias or argument strength accounts are required. Third, to avoid the issue of whether response bias can change trial-by-trial, in a second experiment we manipulated believability between groups. Our goal was to examine whether the key effects generalized to this design, which would support the response bias account.

To this end, in the following sections, we outline (a) how signal detection theory can be applied to deductive reasoning, (b) the novel two-step signal detection model, and (c) two experiments that manipulate believability within or between groups, to which we apply the model.

## Signal Detection Theory and Belief Bias

Signal detection theory (SDT) is a useful framework to examine belief bias because it allows us to separate changes in discriminability (i.e., differentiating valid and invalid arguments) versus response bias (i.e., the "decision stage"; cf. Dube et al., 2010; Rotello \& Heit, 2009). In this framework, arguments fall along a continuum of subjective argument strength, with distinct Gaussian distributions for valid and invalid arguments, as shown in Figure 1. The distance between the means of these distributions reflects how well people can distinguish valid and invalid arguments. People also set a response threshold along the continuum, endorsing any argument that exceeds it in strength (i.e., the tallest "Invalid"/"Valid" threshold in the figure). Thus the hit rate (endorsement rate for valid arguments) is given by the area under the valid distribution to the right of the threshold, and the false alarm rate (endorsement rate for invalid arguments) is given by the
area under the invalid distribution to the right of the threshold. Two important ways that performance can change is by the threshold shifting (i.e., changes in response bias), and/or the valid distribution shifting relative to the invalid distribution (i.e., changes in discriminability or sensitivity).


Figure 1: Standard signal detection model.
Adding confidence judgments to the validity discrimination task allows for a more fine-grained analysis of changes in signal detection parameters. It is assumed that people set a response threshold for $n-1$ response options on the confidence scale - five are shown in Figure 1 for a sixpoint confidence scale. Performance can then be examined using receiver operating characteristic (ROC) curves, which plot hit rates against false alarm rates at different confidence levels (see examples in Figure 2). Evidence for a difference in the discriminability of valid and invalid arguments would be suggested by points from two conditions falling on different curves. Better discrimination is suggested by ROC curves that fall further from the diagonal, towards the upper left - hit rates are higher relative to false alarm rates. In contrast, conventional evidence for a difference in response bias is suggested by points from two conditions falling on different positions along the same curve. A more lenient threshold is suggested by points sitting further towards the right, corresponding to both higher hit rates and higher false alarm rates. Signal detection models can be fit to ROC curves to test for changes in argument discrimination or response bias, which would be supported by reductions in fit due to constraining either the relative location of the valid distribution or the criteria, respectively.


Figure 2: ROC curves from Dube et al. (2010).

An important but controversial result was reported by Dube et al. (2010), who compared and fit ROC curves for believable and unbelievable syllogisms like those in Table 1. Their ROC model fitting showed that argument believability affected response bias but did not affect discriminability. Participants were simply more willing to endorse believable arguments (see Figure 2). This response bias account of belief bias is illustrated in the top panel of Figure 3. Here there are two distributions - one for invalid and one for valid arguments - but two sets of decision thresholds - a more lenient set for believable arguments, and a more conservative set for unbelievable arguments (only three criteria per set are shown, to avoid clutter). A similar account has been proposed for belief bias in causal conditional arguments such as modus ponens (Trippas et al., 2014).

However, this is not the only way to interpret overlapping ROC curves. The response bias interpretation has been contested because an alternative argument strength account is possible, as illustrated in the bottom panel of Figure 3 (Klauer \& Kellen, 2011; Singmann \& Kellen, 2014). This approach assumes a single fixed set of decision thresholds, but four different distributions - distinct invalid and valid distributions for both unbelievable and believable arguments. Discriminability is assumed to be the same for believable and unbelievable arguments, but the believablevalid AND believable-invalid distributions are shifted to the right (i.e., they are stronger on average).
a


Argument Strength


Argument Strength

Figure 3: (a) The response bias account. There are fixed invalid (I) and valid (V) distributions. Criteria are shifted to the left for believable arguments (black lines) relative to unbelievable arguments (grey lines). (b) The argument strength account. There are fixed criteria. Invalid-believable (IB) and valid-believable (VB) distributions are shifted to the right, relative to the invalid-unbelievable (IU) and validunbelievable (VU) distributions.

Resolving this debate has been difficult because in many of the key studies (e.g., Dube et al., 2010; Trippas et al., 2014), believability has been manipulated within a block of arguments. The response bias account assumes that people will shift their criteria on a trial-by-trial basis, depending on whether an argument is believable or unbelievable. However, this assumption is controversial. In the recognition memory literature, although trial-by-trial shifts in criteria are possible, it appears that often this does not occur (Rotello \& Macmillan, 2007; Starns \& Olchowski, 2015; Stretch \& Wixted, 1998). One way to address this issue is to manipulate believability between participants. Uncontroversially, different groups are then free to set different response criteria.

In order to resolve whether belief bias is driven by changes in response bias or argument strength, we carried out two experiments and tested a new signal detection model of reasoning. Experiment 1 confirmed that we could replicate the ROC shifts found by Dube et al. (2010), with believability manipulated within-participants. In Experiment 2, we investigated whether the same effects appeared when believability was manipulated between groups. If the response bias account is correct, then the same distributions of response strength for valid and invalid arguments should apply to those seeing only believable or unbelievable arguments (because there are only two distributions), but the groups will differ in response criteria. Therefore, we would see different hit rates and false alarm rates for the believable and unbelievable argument groups, replicating the Dube et al. (2010) ROC shifts and differences in the response criterion parameter based on model fitting.

Alternatively, if the argument strength account is correct, then the pair of invalid and valid distributions would be in different locations for believable and unbelievable groups. However, each group would be free to set criteria relative to the locations of their invalid and valid distributions - each group has no reason to adopt criteria that are in different locations relative to their distributions. Therefore, we would see the same hit and false alarm rates for both groups, with no ROC shifts nor differences in the criterion parameter.

Accurately testing the competing accounts of belief bias requires model fitting with a model that properly captures the task. Therefore, we extended the signal detection model developed by Dube et al (2010), to treat the valid/invalid decision and confidence judgments as two separate steps. As outlined below, this kind of model is more appropriate for the two-step task than a traditional signal detection model (Moran, Teodorescu, \& Usher, 2015). We first present the model. We then report experiments using within- and between-participant manipulations of conclusion believability and fit the model to these data.

## Two-Step Signal Detection Model

In the two-step validity discrimination task that we use, participants make a "valid"/"invalid" decision, and then rate their confidence. Despite the sequential nature of these judgments, in the standard procedure for generating
empirical ROC curves, data from the response categories are recoded to form a single scale with judgments that range from high-confidence "valid" to low-confidence "valid", then low-confidence "invalid" to high-confidence "invalid" (e.g., Dube et al., 2010; Trippas et al., 2014). Typically, these ROC curves are then fit using the standard single-step SDT model that we outlined above, with a criterion parameter separating each adjacent pair of recoded confidence levels. However, visual inspection of these empirical ROC curves suggests that they differ from the smooth concave curve typically found - they instead exhibit a "hinge" or "elbow" where valid and invalid response categories join, as apparent in Figure 2, particularly for the unbelievable-ROC. In order to successfully model this feature, the standard SDT model was extended to incorporate changes in evidence accumulation and variability in the period between the initial validity judgment and the subsequent confidence judgment.

The two-step SDT model is similar to a standard SDT model with the exception that confidence judgments are based on a noisy version of the evidence value on which the validity judgment was made. Let $x \sim N(\mu, \sigma)$ be the strength of given argument. Let $c$ be a decision criterion such that if $x>c$, respond "valid", else respond "invalid". We propose that a confidence judgment is based on $x^{*}$, a noisy memory trace of argument strength, $x$. That is, $x^{*}=x+x^{\prime}$, for $x^{\prime} \sim N(v, \eta)$. If $v>0$ then additional argument strength is accumulated in the interval between the two decisions (cf., Moran et al., 2015). Suppose, there are $k$ confidence categories labeled, in sequence, from most confident to least confident. Then, associated with these category labels is set of points on the strength continuum, $U=\left\{u_{0}, u_{1}, \ldots, u_{k}\right\}$, such that, $u_{0}<u_{1}<\ldots<u_{k}, u_{0}=-\infty$, $u_{k}=\infty$, and a set of points, $V=\left\{v_{0}, v_{1}, \ldots, v_{k}\right\}$, such that, $v_{0}>v_{1}>\ldots>v_{k}, \quad v_{0}=\infty, \quad v_{k}=-\infty$. Then, if the response is "invalid" and $u_{i}<x^{*} \leq u_{i+1}$ or, if the response is "valid" and $v_{i} \leq x^{*}<v_{i-1}$ then respond with the $i$ th category label.

The hypotheses of interest were primarily tested by comparing the fits of nested versions of this model using the likelihood ratio test. Although the response bias and argument strength accounts are formally identical for a traditional signal detection model, this is not strictly true for the two-step model. Therefore, both accounts can be tested when believability is manipulated within-participants.

## Experiments

In two experiments, participants evaluated the validity of categorical syllogisms, which included logically valid and invalid arguments with believable or unbelievable conclusions in a $2 \times 2$ design. Experiments 1 and 2 manipulated believability within- and between-groups, respectively.

## Method

Participants. One-hundred-and-seventeen students (30 males) at the University of New South Wales, Sydney, participated for course credit. Mean age was 18.8 years (SD $=2.3$ ). Participants were randomly allocated to Experiment $1(\mathrm{~N}=38)$ or one of the groups in Experiment 2 (believable $\mathrm{N}=40$, unbelievable $\mathrm{N}=39$ ).

Stimuli. In Experiment 1, participants evaluated 64 arguments across two blocks of 32 trials, with 16 believable and 16 unbelievable arguments per block - half of which were valid in each case. In Experiment 2, participants evaluated either 32 believable or 32 unbelievable arguments (half valid).

Example stimuli are shown in Table 1. The arguments were based on those of Experiment 2 by Dube et al. (2010), and were constructed using their 16 syllogistic problem frames (e.g., All X are Y; Some Z are not Y; Therefore some $Z$ are not $X$ ). Half were valid and half were invalid. Each problem frame had the conclusion structure, Some $Z$ are not $X$ (or Some $X$ are not $Z$ ), and was assigned content involving a category-exemplar relationship (e.g., drinks-beers, dogspoodles, plants-weeds).

Conclusion believability was manipulated by simply reversing the order of the category and exemplar (e.g., Some drinks are not beers vs. Some beers are not drinks). We verified the believability of the conclusion statements in a separate study by 34 people drawn from a similar population to the main experiments. Based on ratings on a 5-point scale ( $1=$ unbelievable, $3=$ neutral, $5=$ believable), the 32 statement pairs with the most extreme average ratings were selected from a set of 38 pairs (Believable: $\mathrm{M}=4.95$, $\mathrm{SD}=$ 0.09 ; Unbelievable: $\mathrm{M}=1.59, \mathrm{SD}=0.35$ ). To minimize the effects of premise believability, the premises included a nonsense term (e.g., krabbers, junids).

The semantic content was split into four subsets of eight category-exemplar pairs, so the content could be assigned to all four believability-by-validity conditions, counterbalanced across participants. Experiment 1 participants (believable and unbelievable withinparticipants) saw the category-exemplar content once per block and the 16 problem frames twice per block (once as believable and once as unbelievable versions), forming the 64 arguments over two blocks. Content assignment was controlled for this group so that in the second block, each participant saw the same content in the same problem structures as in their first block, but with conclusion believability reversed. At the start of the second block, these participants were warned that there would be similar content but the specific arguments would be different. Experiment 2 participants (believable-only and unbelievable-only groups) saw each category-exemplar content once and the 16 problem frames twice, forming the 32 arguments.

Before beginning the experiment, all participants received two valid and two invalid practice problems with abstract content (e.g., "All M are P...") and different structures that were not included in the main task.

Procedure. Participants were shown the set of arguments in random order, presented one-by-one on a computer, with a line separating the conclusion from the premises. The instructions asked participants to assume that the premises were true and assess whether the conclusion logically followed from them. Valid arguments were defined as those for which the sentence below the line was necessarily true, given that the information above the line was true (and invalid $=$ not necessarily true $)$. Participants were told that the arguments would contain a nonsense word. A trial counter was presented at the top left corner of the screen. Participants clicked on either the "Valid" or "Invalid" button presented underneath a given argument, then rated their confidence on a scale that appeared, ranging from 50 (Guessing) to 100 (Certain) in increments of ten.

## Results

Both experiments replicated previously observed argument endorsement patterns and belief bias effects (see Table 2 ; e.g., Dube et al., 2010; Evans et al., 1983; Newstead et al., 1992). Analysis of variance (ANOVA) revealed that participants endorsed (i.e., responded "valid") valid arguments more often than invalid arguments: Experiment $1, F(1,37)=64.28, p<.001, \eta^{2}=.35$; Experiment $2, F(1$, 77) $=127.24, p<.001, \eta^{2}=.40$. Participants endorsed believable arguments more often than unbelievable arguments: Experiment $1, F(1,37)=38.59, p<.001, \eta^{2}=$ .12; Experiment 2, $F(1,77)=19.92, p<.001, \eta^{2}=.13$. Notably, as shown in the Table, there was a larger difference between the acceptance rates of valid and invalid arguments for unbelievable than for believable arguments: Experiment 1, $F(1,37)=5.50, p=.02, \eta^{2}=.01$; Experiment $2, F(1,77)=9.81, p=.002, \eta^{2}=.05$.

Table 2: Performance in Experiments 1 and 2. Hit rate is p("Valid"|Valid); False alarm rate is p("Valid"|Invalid).

| Experiment | Condition | Hit rate | False alarm rate |
| :---: | :--- | :---: | :---: |
| 1 | Believeable | 0.83 | 0.56 |
|  | Unbelievable | 0.71 | 0.37 |
| 2 | Believeable | 0.82 | 0.58 |
|  | Unbelievable | 0.75 | 0.34 |

The ROC curves for each experiment are presented in Figure 4 (unfilled points). Both show effects that are consistent with shifts in response criteria and comparable to Dube et al. (2010; cf. Figure 2), although we used more confidence response options. In each experiment, the points for believable and unbelievable arguments fall on similar curves, though the believable points are shifted further to the top-right corner than the unbelievable points.

We first fit an unconstrained two-step signal detection model to each experiment. As shown by the filled points in Figure 4, the predicted ROC points correspond reasonably well with the empirical results for both experiments, though there are some small departures for Experiment 1: Experiment 1, $G^{2}(12)=22.54, p=.03$; Experiment 2, $G^{2}(12)=15.83, p=.20$.


Figure 4: Observed ROC curves (Obs) and expected scores from the unconstrained model (Exp), for Experiments 1 and 2 (panels a and b, respectively).

We compared this unconstrained model against two nested models: a constant discriminability model and a constant criterion model in which (respectively) discriminability or the "valid"/"invalid" decision criterion for the initial binary judgment was constrained across believable and unbelievable conditions. For both experiments, the fit of the constant discriminability model did not significantly differ from that of the unconstrained model: Experiment 1, $G^{2}(1)=0.23, p=.63$; Experiment 2, $G^{2}(1)=0.001, p=.97$. This shows that, in line with Dube et al. (2010), discriminability did not differ between believability conditions.

The constant criterion model led to a reduction in fit compared to the unconstrained model: Experiment 1, $G^{2}(1)$ $=47.89, p<.001$; Experiment $2, G^{2}(1)=77.75, p<.001$. This indicates that, in line with the response bias account, the "valid"/"invalid" decision threshold differed between believability conditions. Importantly, this was true both when believability was manipulated within-groups (Experiment 1) and between-groups (Experiment 2).

When a (non-nested) variant of the two-step model was applied to Experiment 1 that allowed the believable distributions to shift (i.e., the argument strength account), we found that it also provided a satisfactory fit to the data: $G^{2}(20)=30.18, p=0.07$. In other words, an argument
strength account of belief bias could also explain the Experiment 1 data. Such a model cannot sensibly be applied to Experiment 2. Nevertheless, as we argued above, the response bias account can more readily explain belief bias effects that occur between-groups.

## Discussion

We investigated whether belief bias effects in deductive reasoning could be explained as a response bias effect. Experiment 1 replicated the belief bias effects of Dube et al. (2010), with conclusion believability manipulated withinblock. We applied a new two-step signal detection model to better suit the two-step task, and confirmed that belief bias effects are consistent with a shift in response bias, rather than discriminability. Experiment 2 extended the same results to an equivalent task with believability manipulated between-groups.

Under the response bias account (Dube et al., 2010; Trippas et al., 2014), this pattern is explained by a shift in decision threshold, such that there is a more lenient criterion for believable conclusions. Under the argument strength account (Klauer \& Kellen, 2011; Singmann \& Kellen, 2014), the belief bias effect reflects higher mean strength for believable-valid and believable-invalid arguments than for unbelievable-valid and unbelievable-invalid arguments.

It could be argued that participants in Experiment 1 were unlikely to change their criteria trial-to-trial for different levels of believability, favoring the argument strength account. However, this account would have difficulty with Experiment 2, where participants saw only believable or only unbelievable arguments. There, the two groups had no reason to position their criteria in different locations relative to their distributions. Thus if belief bias primarily reflects a change in argument strength, the belief bias effects should have disappeared. The fact that they did not suggests that the most plausible explanation of belief bias in the current data sets is a change in response bias.

Therefore, addressing the debate between response bias and argument strength accounts of belief bias, we agree that believable conclusions are most likely to affect the decision stage, lowering the decision threshold rather than appearing more logically valid. Just as people may require stronger evidence to endorse that an unusual event occurred (Starns \& Olchowski, 2015), it seems that people also require stronger evidence to endorse a syllogism with an unbelievable conclusion. As Dube et al. (2010) concluded, this is problematic for theories of reasoning that propose that believability affects the process of evaluating validity (e.g., Evans et al., 1983; Markovits \& Nantel, 1989; Newstead et al., 1992; Oakhill et al., 1989). Future work should address whether the same findings generalize to other reasoning problems such as causal conditionals.

## References

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 941-947.

Dube, C., Rotello, C. M., \& Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. Psychological Review, 117(3), 831-863.
Evans, J. St. B. T., Barston, J. L., \& Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. Memory \& Cognition, 11, 295-306.
Evans, J. St. B. T., Newstead, S. E., \& Byrne, R. M. J. (1993). Human reasoning: The psychology of deduction. Hove, UK: Lawrence Erlbaum Associates Ltd.
Klauer, K. C., \& Kellen, D. (2011). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (2010). Psychological Review, 118(1), 164-173.

Markovits, H., \& Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. Memory \& Cognition, 17, 11-17.
Moran, R., Teodorescu, A. R., \& Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. Cognitive Psychology, 78, 99-147.
Newstead, S. E., Pollard, P., Evans, J. St. B. T., \& Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. Cognition, 45, 257-284.
Oakhill, J., Johnson-Laird, P. N., \& Garnham, A. (1989). Believability and syllogistic reasoning. Cognition, 31, 117-140.
Roberts, M. J., \& Sykes, E. D. A. (2003). Belief bias and relational reasoning. The Quarterly Journal of Experimental Psychology: Section A, 56, 131-154.
Rotello, C. M. \& Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 1317-1330.
Rotello, C. M., \& Macmillan, N. A. (2007). Response bias in recognition memory. The Psychology of Learning and Motivation, 48, 61-94.
Singmann, H., \& Kellen, D. (2014). Concerns with the SDT approach to causal conditional reasoning: A comment on Trippas, Handley, Verde, Roser, McNair, and Evans (2014). Frontiers in Psychology, 5, 402.

Shynkaruk, J. M., \& Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. Memory \& Cognition, 34, 619-632.
Starns, J. J., \& Olchowski, J. E. (2015). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. Memory \& Cognition, 43, 49-59.
Stretch, V., \& Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 13791396.

Trippas, D., Verde, M. F., Handley, S. J., Roser, M. E., McNair, N. A., \& Evans, J. St. B. T. (2014). Modeling causal conditional reasoning data using SDT: Caveats and new insights. Frontiers in Psychology, 5, 217.

# Rational use of prosody predicts projection in manner adverb utterances 

Jon Scott Stevens, Marie-Catherine de Marneffe, Shari R. Speer, Judith Tonhauser<br>\{stevens.400,demarneffe.1,speer.21,tonhauser.1\}@osu.edu<br>Department of Linguistics, The Ohio State University, Columbus, OH 43210 USA


#### Abstract

Speakers can be taken to be committed to utterance content even when that content is contributed in the scope of an entailment-canceling operator, like negation (e.g., Chierchia \& McConnell-Ginet, 1990). We develop a probabilistic model of this phenomenon, called 'projection', that relies on the prosodic realization of utterances. We synthesize existing theoretical claims about prosody, information structure and projection into a model that assumes a rational speaker (Frank \& Goodman, 2012) who produces utterances with prosodic melodies that can signal which utterance content she is committed to. Predictions of the probabilistic model are compared to the responses of an experiment designed to test the effect of prosody on projection in manner adverb utterances. Key behaviors of the model are borne out empirically, and the quantitative fit is surprisingly good given that the model has only one free parameter. Our findings lend support to analyses of projection that are sensitive to the information structure of utterances (e.g., Simons, Beaver, Roberts, \& Tonhauser, 2017).


Keywords: Projection; prosody; information structure; probabilistic pragmatics; rational speech acts; manner adverbs

## Introduction

Projective content is utterance content that the speaker may be taken to be committed to even when the content is introduced by an expression in the scope of an entailment-canceling operator (e.g., Chierchia \& McConnell-Ginet, 1990). A speaker who utters (1a) is taken to be committed to the content $\phi$ that Sam is ill and to the content that Jo discovered $\phi$. A speaker who utters the negated variant in (1b) or the polar question in (1c) is not taken to be committed to Jo having discovered $\phi$ (rather, Jo's discovery is negated or asked about), but the speaker may still be taken to be committed to $\phi$, that Sam is ill. Hence, $\phi$ is projective content.
(1) a. Jo discovered that Sam is ill. b. Jo didn't discover that Sam is ill.
c. Did Jo discover that Sam is ill?

Simons et al. (2017) develop a question-based analysis of the projection of the content of the complement of predicates like discover according to which the speaker can be taken to be committed to the content of the complement if it is entailed by the Question Under Discussion (Roberts, 2012) that the utterance is taken to address. This analysis correctly predicts that whether the speaker is taken to be committed to the content of the complement depends on the prosodic realization of the utterance (see Tonhauser, 2016, for empirical evidence). For instance, if (1c) is uttered with prosodic prominence on discover, as in (2a), where capital letters indicate prosodic prominence, then the speaker is more likely to be taken to be committed to the content of the complement than if (1c) is uttered with prosodic prominence on ill, as in (2b).
a. Did Jo DISCOVER that Sam is ill?

## b. Did Jo discover that Sam is ILL?

This paper provides conceptual and empirical support for the question-based analysis of projection from a novel empirical domain: utterances with manner adverbs, like Masha didn't run quickly. To formalize the link between prosody, information structure and projection hypothesized in Simons et al. 2017, we develop a Rational Speech Act (RSA) model (Bergen \& Goodman, 2015; Frank \& Goodman, 2012, and many others) that predicts the projectivity of the so-called prejacent (that Masha ran) from the prosodic realization of manner adverb utterances. The model is evaluated based on empirical observations about the projectivity of the prejacent.

## Projective content in manner adverb utterances

In a manner adverb sentence, e.g., (3), a manner adverb like quickly, beautifully or easily modifies the activity-denoting verb. (3) entails that Masha ran (the prejacent). It has been long observed that the prejacent can project and that the projectivity of the prejacent depends on the prosody of the utterance (e.g., Abrusán, 2013; Simons, 2001). For instance, the speaker of (4a), with quickly prosodically prominent, may be taken to be committed to the prejacent even though the sentence is negated. On the other hand, the speaker of (4b), with Masha prosodically prominent, is not typically taken to be committed to the prejacent, i.e., the prejacent does not project. The speaker of (4b) may be committed to a different content, namely that somebody ran quickly. For reasons of space we only consider the prejacent here, though the formal proposal extends to the projectivity of other content.
(3) Masha ran quickly.

## a. Masha didn't run QUICKLY. <br> b. MASHA didn't run quickly.

Simons et al.'s (2017) question-based analysis of projection accounts for the dependence of the projectivity of the prejacent on prosody based on two independently made empirical observations. The first observation is that informationstructural focus can be prosodically indicated in American English (e.g., Eady \& Cooper, 1986). Compared to nonfocused expressions in the same intonational phrase, focused expressions are more likely to be realized with pitch accents, a longer duration, an expanded pitch range and/or greater intensity. Thus, the two manner adverb utterances in (4) are compatible with different expressions being focused, as shown in (5), where focus is marked by angle brackets subscripted with ' $F$ ':

[^205]The focused expression is used to calculate the focus alternatives set of an utterance: following Rooth 1992, a focus alternatives set of $(5 a)$ is the set of propositions of the form 'Masha didn't run $x$ ', with $x$ a modifier, and a focus alternatives set of (5b) is the set of propositions of the form ' $x$ didn't run quickly', with $x$ an entity. Following Beaver and Clark 2008, we assume that alternatives can also be calculated under negation: thus, another focus alternatives set of (5a) is the set of propositions of the form 'Masha ran $x$ ', with $x$ a modifier, and another focus alternatives set of (5b) is the set of propositions of the form ' $x$ ran quickly', with $x$ an entity.

Importantly, the prosody of utterances merely constrains what is focused but does not determine it. Thus, although the prosodic realization of (4a) is not typically compatible with the focus marking indicated in (5b), it is compatible with a larger expression of the sentence being focused, as in (6).

## (6) Masha didn't [run QUICKLY] ${ }_{F}$.

The second empirical observation is that answer utterances are congruent with the interrogative utterances they address (e.g., Paul, 1880; Rooth, 1992; von Stechow, 1990). Consider B's utterance in (7a), and assume that B's utterance is realized with the so-called rise-fall-rise contour, which consists of a rising pitch accent on quickly $\left(\mathrm{L}+\mathrm{H}^{*}\right.$ in the ToBI annotation scheme, Beckman \& Ayers, 1997) and a rising end contour (L-H\% in ToBI notation). When B's utterance in (7a) is realized with the rise-fall-rise contour, it is congruent with, i.e., judged to be acceptable in response to, A's interrogative utterance in (7a). However, because of the placement of the pitch accent on quickly, it is not congruent with A's interrogative utterance in (7b). Likewise, when B's utterance in (7b) is realized with the rise-fall-rise contour, it is congruent with A's utterance in (7b), but not with A's utterance in (7a).

> a. A: How did Masha run?
> B: Masha didn't run $[\mathrm{QUICKLY}]_{\mathrm{F}}$.
> b. A: Who ran quickly?
> B: [MASHA $]_{\mathrm{F}}$ didn't run quickly.

When B's utterances are realized with the rise-fall-rise contour, B indicates that their utterance does not provide a complete answer to the question (see e.g., Wagner, 2012); instead, B's utterance only eliminates a possible true answer to the question. A prosodically motivated model of projection must take these contributions of contours into account.

In alternative semantics, question-answer congruence is accounted for by assuming that a focus alternatives set of a congruent answer includes the denotation of the question, which is a contextually restricted set of propositions. For instance, B's answer in (7a) is congruent with A's interrogative utterance since the focus alternatives set 'Masha ran $x$ ' includes the set of propositions denoted by A's utterance (a subset of propositions of the form 'Masha ran $x$ '), but B's answer is not congruent with A's interrogative utterance in (7b) since the focus alternatives set 'Masha ran $x$ ' does not include the set of propositions denoted by A's utterance (a subset of propositions of the form ' $x$ ran quickly').

Importantly, in naturally occurring discourse, many utterances are not made in response to an interrogative utterance (an explicit question). Simons et al. (2017) assume that such utterances address an implicit question: given questionanswer congruence, the focus marking of an utterance that addresses an implicit question provides a cue to the question that the utterance addresses (e.g., Halliday, 1967; Most \& Saltz, 1979; Roberts, 2012). The potentially implicit question that is addressed by an utterance is called the Question Under Discussion (QUD). It follows that the prosodic realization of a manner adverb utterance provides listeners with a cue to the focus of the utterance, which in turn provides a cue to the QUD that the speaker was intending to address. Thus, (5a) can be taken to address the QUD 'How did Masha run?' and (5b) can be taken to address the QUD 'Who ran quickly?':
(8) a. Possible QUD of (5a): \{Masha ran quickly, Masha ran slowly, Masha ran clumsily,...\}
b. Possible QUD of (5b): \{Masha ran quickly, Jack ran quickly, Sue ran quickly,...\}
To predict projection, Simons et al. (2017) make the following assumption:
(9) Projection under the question-based analysis: Content $\phi$ of utterance $U$ projects if $\phi$ is entailed by each alternative in the QUD addressed by $U$.
The question-based analysis predicts that the prejacent is more likely to project from (4a) than from (4b). If a speaker utters (4a), with prosodic prominence on the manner adverb, and the utterance is taken to have the focus shown in (5a), then she can be taken by the listener to intend her utterance to address the QUD 'How did Masha run?' in (8a). Since each alternative in the QUD has the form 'Masha ran $x$ ', each entails that Masha ran and so the prejacent is predicted by (9) to project. On the other hand, if a speaker utters (4b), with prosodic prominence on the subject, and the utterance is taken to have the focus shown in (5b), she can be taken by the listener to intend her utterance to address the QUD 'Who ran quickly?' in (8b). Since each alternative in the QUD has the form ' $x$ ran quickly', the QUD entails that somebody ran quickly, but not the prejacent, that Masha ran. Thus, the prejacent is not predicted by (9) to project from (4b). Importantly, since prosody does not determine focus, but merely provides listeners with a cue, and since focus does not determine the QUD, but merely provides listeners with a cue, the questionbased analysis does not predict categorical (non-)projection of the prejacent from (4a) and (4b), but merely that the prejacent is more likely to project from (4a) than (4b).

## Modeling projection

For the purpose of modeling the link between prosody and projection, we consider an utterance to be a sentence with a melody. For the utterances considered here, a melody is the combination of a single pitch accent $\left(\mathrm{L}+\mathrm{H}^{*}, \mathrm{H}^{*}, \mathrm{~L}^{*}\right)$, aligned with the stressed syllable of the accented word, and an end contour (L-H\%, H-L\%, L-L\%). Our model considers pitch
accent positions either on the adverb, the verb, the negated auxiliary or the subject of the sentence. For example, $\mathrm{L}+\mathrm{H}^{*}$ $\mathrm{L}-\mathrm{H} \%$, with the $\mathrm{L}+\mathrm{H}^{*}$ realized on the first syllable of quickly may be the melody of the manner adverb utterance in (4a). We develop a probabilistic model of projection based on the assumption of a 'rational speaker', who chooses a melody for a given sentence to most effectively signal the identity of the QUD that an utterance of the sentence is intended to address. Following the basic framework of the RSA model, we begin by developing a notion of speaker utility, which is taken to be the 'usefulness' of a melody for a given sentence as a signal of which QUD is being addressed by the speaker's utterance. More concretely, the utility of a melody $M$ for a given sentence and a target QUD $Q$, is the probability that a hearer would randomly select $Q$, given that the sentence was uttered with melody $M$, and given what the hearer knows about the compatibility between $M$ and possible QUDs. We define this utility as follows, where $Q_{M}$ is the set of possible QUDs addressed by the given sentence that are compatible with $M$ :
(10) $U(Q, M)=\frac{1}{\left|Q_{M}\right|}$ if $Q \in Q_{M}$, else 0

To define what it means for a melody for a given sentence to be compatible with a QUD, we draw on the following independently motivated sets of assumptions from the prosodypragmatics literature:

- Assumption A - Compatibility of pitch accent with focus: A pitch accent on expression X is compatible with focus on $X$ or a constituent that contains $X$ and an immediately adjacent expression (e.g., Selkirk, 1996).
- Assumption B - Compatibility of focus with QUD:

The QUD that is addressed by the utterance must be congruent with the focus marking of the utterance (Beaver \& Clark, 2008, 45).

- Set of assumptions C - Compatibility of pitch contour with QUD:

1. Pitch contours that lack a final fall ( $\mathrm{L}-\mathrm{L} \%$ ) and contain either an $\mathrm{L}+\mathrm{H}^{*}$ pitch accent or a continuation rise ( $\mathrm{L}-\mathrm{H} \%$ ) — $\mathrm{H}^{*} \mathrm{~L}-\mathrm{H} \%, \mathrm{~L}+\mathrm{H}^{*} \mathrm{~L}-\mathrm{H} \%, \mathrm{~L} * \mathrm{~L}-\mathrm{H} \%$ and $\mathrm{L}+\mathrm{H}^{*} \mathrm{H}-\mathrm{L} \%$-are only compatible with incomplete answers to the QUD, i.e., answers that do not pick out a single true alternative (see e.g. Lai, 2012; Wagner, 2012, for evidence that $\mathrm{L}+\mathrm{H}^{*}$ and $\mathrm{L}-\mathrm{H} \%$ signal that the QUD has not been completely answered).
2. Pitch contours that either (i) have a final fall, or (ii) have neither an $\mathrm{L}+\mathrm{H}^{*}$ pitch accent nor continuation rise to suggest incompleteness- $\mathrm{H}^{*} \mathrm{~L}-\mathrm{L} \%, \mathrm{~L}+\mathrm{H}^{*} \mathrm{~L}-\mathrm{L} \%$, L* L-L\%, $\mathrm{H}^{*} \mathrm{H}-\mathrm{L} \%$ and $\mathrm{L} * \mathrm{H}-\mathrm{L} \%$-are only compatible with complete answers to the QUD.
For any melody-QUD pair $\langle M, Q\rangle$ for a given sentence, these assumptions can be used to generate the set $Q_{M}$ and therefore determine the utility of using $M$ to signal $Q$.

Adopting an RSA-based view, we posit a rational speaker who chooses melodies to maximize utility, i.e., maximize the chance that listeners retrieve the QUD intended by the
speaker, though the maximization is approximate, i.e., there still remains some probability of choosing a non-rational (non-utility-maximizing) melody. This is accomplished by setting the probability $P_{S}$ of the speaker producing a melody $M$ given a QUD $Q$ equal to a soft max function of $U(Q, M)$. The soft max function approximates utility maximization using a rationality parameter, $\lambda$, where higher values of $\lambda$ result in lower probability of a non-rational melody being chosen. Thus, in cases where there is a single utility-maximizing melody, the probability of selecting that melody will approach 1 as $\lambda$ increases. The formula for $P_{S}$ is given below, where $M^{\prime}$ is any member of the set of possible melodies the speaker could use:

$$
\begin{equation*}
P_{S}(M \mid Q)=\frac{e^{\lambda U(Q, M)}}{\sum_{M^{\prime}} e^{\lambda U\left(Q, M^{\prime}\right)}} \tag{11}
\end{equation*}
$$

We use Bayes' rule to determine the probability $P_{H}$ of the hearer deciding that the QUD is $Q$ given that she has heard the sentence uttered with melody $M$ :

$$
\begin{equation*}
P_{H}(Q \mid M)=\frac{P_{S}(M \mid Q) \times P(Q)}{\sum_{Q^{\prime}} P_{S}\left(M \mid Q^{\prime}\right) \times P\left(Q^{\prime}\right)} \tag{12}
\end{equation*}
$$

The denominator in this equation is a sum of probabilities over all possible QUDs $Q^{\prime}$, i.e., the set of QUDs that are compatible with any of the melodies we assume could have been used to utter the sentence. For instance, given the sentence Masha didn't run quickly, the set of all QUDs compatible with some melody for that sentence includes 'Who ran quickly?', 'Did Masha run quickly?', 'What did Masha do?', 'What did Masha do quickly?' and 'How did Masha run?', as well as the corresponding QUDs with negation ('Who didn't run quickly', etc.). For current purposes we assume a uniform prior probability distribution over QUDs.

We use $P_{H}$ to calculate the probability that the prejacent of a manner adverb sentence uttered with melody $M$ projects, i.e., that the speaker is taken to be committed to the prejacent. Recall that under assumption (9) from Simons et al. 2017, content projects if it is entailed by the QUD. For manner adverb sentences, the only QUD that entails the prejacent is the set of alternatives obtained by abstracting over the manner adverb (e.g., \{Masha ran $x \mid x$ is a modifier $\}$ entails that Masha ran, as discussed above). We call the prejacent $\phi$, and the prejacent-entailing QUD $Q_{\phi}$. The probability of $\phi$ projecting, given melody $M$, is the probability of the hearer assuming $Q_{\phi}$ given $M$ :
(13) $\quad P(\operatorname{PROJECT}(\phi) \mid M)=P_{H}\left(Q_{\phi} \mid M\right)$

The link between $Q_{\phi}$ and projection is not probabilistic-it is a categorical consequence of the theory set forth in Simons et al. 2017 (see (9)). The probabilistic character of the model results from the fuzzy link between prosodic melodies and the implicit questions that utterances with those melodies are taken to address. The hearer must determine how likely it is that $Q_{\phi}$ is the intended QUD, based on how the speaker selects melodies to convey the QUD she intends to address.

This model operates over the possible melodies that the speaker could use and considers those possibilities when calculating the probability of the projection of the prejacent. To test the model, we examine the model's predictions for two particular melodies and compare those predictions to experimental results. The next section provides information on the two melodies and how their effect on the projection of the prejacent was assessed experimentally.

## Experiment methodology

Using the method of Tonhauser 2016, participants listened to audio recordings of manner adverb utterances and judged whether the speaker was certain of the prejacent.

Participants. We recruited 100 self-reported native speakers of American English on Amazon's Mechanical Turk platform.

Stimuli. Each participant listened to 16 utterances- 10 target utterances and 6 fillers. The target sentences were all of the form, "subject didn't verb adverb", where each adverb was a manner adverb, and where each subject was a proper name. The target sentences were:
(14) a. Amanda didn't clap loudly.
b. Jennifer didn't drive carelessly.
c. Elizabeth didn't leave silently.
d. Linda didn't write neatly.
e. Susan didn't sing beautifully.
f. Jerry didn't knock frantically.
g. Justin didn't smile cheerfully.
h. Alexander didn't sneeze softly.
i. Tyler didn't lie deliberately.
j. Dennis didn't win easily.

The filler items were:
(15) a. Sandy wasn't invited to the party.
b. Did Mario bring a chocolate cake?
c. Who knows if Maggie is at the party?
d. Mike forgot to bring the ketchup.
e. Paul loves that pie!
f. Mandy was out gardening in the yard.

Each target sentence was uttered with one of two possible melodies, $\mathrm{L}+\mathrm{H}^{*} \mathrm{~L}-\mathrm{H} \%$ with the pitch accent on the adverb (the LH-Adverb condition), or $\mathrm{L}+\mathrm{H}^{*} \mathrm{~L}-\mathrm{H} \%$ with the pitch accent on the proper name subject (the LH-Name condition). Fillers were pronounced with pitch accents on words other than the subject noun, and contained a variety of pitch accent and end tone types not used in the target utterances.

In addition to exposing participants to melodies other than the ones used for the target items, filler utterance-question pairs tested whether participants comprehended direct consequences of an action described by the speaker. For example, the utterance, Mike forgot to bring the ketchup was followed by the question, 'Is Debby certain that Mike brought the ketchup?'. Given that Debby's utterance implies that Mike forgot the ketchup, participants were expected to rate the speaker as "not certain", but to reliably respond in this


Figure 1: A screenshot of one experimental item.
way requires attention to the meaning of the utterance as well as its component words.

Each participant was assigned to one of two lists, where the two lists contained the same sentences but were counterbalanced for prosodic condition. The same 6 fillers occurred on both lists.

Procedure. Participants were instructed to imagine themselves at a party, where they overhear Debby, the host, utter various sentences to somebody else. For each of the 16 utterances, the participant was asked to rate on a 7-point Likert scale labeled at 4 points ( $1 /$ "No, not certain", $3 /$ "Possibly not certain", 5/"Possibly certain", 7/"Yes, certain") whether Debby was certain of some content based on what she said and how she said it. On each trial, participants were presented with a display as in Fig. 1, clicked the audio icon, heard an utterance, read the related question, clicked on the radio button that corresponded to their chosen response, and clicked the 'continue' button to proceed to the next trial. For the target item shown in Fig. 1, the utterance was Amanda didn't clap $L O U D L Y$, and the participant was asked to rate the speaker's certainty about the prejacent, i.e., 'Amanda clapped'.
Data exclusion. If participants answered more than one filler incorrectly (an answer greater than 3 on the Likert scale for something that Debby would be uncertain about or an answer smaller than 5 for something that she would be certain about), their responses were excluded from analysis. We excluded 28 participants on these grounds, leaving 72 participants whose responses we analyzed. Whether these participants are excluded does not change the main effect of condition on response.

## Model predictions

If participants take Debby to be committed to the truth of $\phi$, we expect them to respond that Debby is certain that $\phi$. If Debby is not taken to be committed to the truth of $\phi$, we expect participants to respond that Debby is not certain that $\phi$. However, we do not expect mean responses at the extreme ends of the 7-point Likert scale, because it is possible for participants to exhibit uncertainty about whether Debby is committed to $\phi$. Participants can therefore give a response in the mid-range of the scale. To directly compare our model's predictions to the experimental results, we use the model to pre-


Figure 2: Model predictions about the extent to which listeners take the speaker to be certain of the prejacent, given the two melodies. The x -axis represents the $\lambda$-parameter, which encodes the degree to which the predictions reflect utilitymaximizing reasoning. The y-axis is mapped to a 7 -point scale to parallel the experimental task.
dict participants' probabilistic evaluation of whether Debby is certain that $\phi$. The probability $P(\phi \mid M)$ is the probability that Debby is certain that $\phi$, given $M$. This is expected to be 1 when the participant takes $\phi$ to project based on $M$-which occurs with probability $P(\operatorname{PROJECT}(\phi) \mid M)$. When the participant does not take $\phi$ to project based on $M$, we expect some baseline uncertainty about whether Debby is committed to $\phi$, which we encode as a prior probability $P(\phi)$. We thus define $P(\phi \mid M)$ as follows:

$$
\begin{align*}
P(\phi \mid M) & =P(\operatorname{PROJECT}(\phi) \mid M) \times 1  \tag{16}\\
& +P(\neg \operatorname{PROJECT}(\phi) \mid M) \times P(\phi)
\end{align*}
$$

To account for the fact that our stimuli do not provide any prior evidence (i.e., evidence apart from the manipulated prosody) for whether the speaker is committed to $\phi$, we take $P(\phi)$ to be uniform, i.e., equal to 0.5 . Assuming a uniform prior over $\phi$ maintains a model with only one free parameter, the rationality parameter $\lambda$, and makes the model more informative by limiting the range of predictions that it can make.

Fig. 2 shows the model predictions ${ }^{1}$ as $\lambda$ increases to 10 (a relatively high value given ones used in the literature). We see that the modeled participant responses in the two prosodic conditions LH-Adverb and LH-Name diverge rather shallowly, predicting significantly higher certainty in the LHAdverb condition, but not by a huge margin. The model predicts that projection in the LH-Adverb condition, though higher, will not be at ceiling. This is because utterances in the LH-Adverb condition are not only compatible with the prejacent-entailing QUDs but also with QUDs that do not entail the prejacent, and thus the probability of projection never exceeds fifty percent ( 4 on the Likert scale), even for high values of $\lambda .^{2}$

[^206]

Figure 3: Responses by prosodic condition. Violin plots show frequency of participant means. Bar plots show overall means with $95 \%$ bootstrap confidence intervals.

## Experiment results

The experimental results are shown in Fig. 3. Mean Likert scale response in the LH-Adverb condition was 5.7 , compared to 4.8 in the LH-Name condition. This difference is in the expected direction: participants rated Debby as being less certain of the prejacent when the pitch accent was on the subject than when it was on the manner adverb. A mixedeffects ordinal regression model with random intercepts for participant and item and random slope for participant shows responses to be significantly lower in the LH-Name condition than in the LH-Adverb condition ( $\beta=-1.13, S E=0.24$, $z=-4.68, p<0.0001$ ).

The model predictions in Fig. 2 are in line with the experimental results shown in Fig. 3 in three key ways:

1. The model correctly predicts a significant difference in mean responses between the two conditions, with the LHAdverb items showing higher certainty ratings.
2. The model correctly predicts the magnitude of this difference to be rather small (within about one point on the Likert scale).
3. The model correctly predicts that even in the LH-Adverb condition, where projection is expected, the ratings are not at ceiling.
Thus, three qualitative experimental behaviors are accounted for by our RSA model, which builds on existing theoretical assumptions about the links between prosody, the QUD and projection, and incorporates those assumptions into a probabilistic pragmatic model.

The quantitative match with the model's predictions is not exact-certainty is a bit higher across the board than predicted—but as we see in Fig. 4, it is not far off, either. We would expect a more exact match if we experimentally obtained priors over the hearer's evaluation of the speaker's certainty for the various sentences used (instead of assuming a uniform prior of 0.5 ), a possible task for future research.

## Discussion

This paper showed that the question-based analysis of projection developed in Simons et al. 2017 can be extended to manner adverb utterances and formalized in an RSA model.


Figure 4: Model predictions by $\lambda$ value (blue), along with the experimentally observed means (black).

The experimental findings empirically support the predictions of the model and, hence, the question-based analysis of projection. They also add to the growing empirical evidence that formal analyses of projection, including conventional triggering analyses (e.g., Heim, 1983; van der Sandt, 1992), need to be sensitive to information structure (e.g., Beaver, Roberts, Simons, \& Tonhauser, 2017; Tonhauser, 2016). Finally, the RSA model demonstrates the feasibility of formal pragmatic analyses of projection.

Future research needs to investigate the predictions of the model for other projective contents of manner adverb utterances, other prosodic realizations of such utterances, and the projective contents of other utterances. We also observed that the influence of prosody on the projectivity of the prejacent was heterogeneous across items. This observation suggests enriching the model with information about listeners' prior expectations about the prejacent, e.g., about how likely somebody is to smile given that they didn't smile cheerfully.

## Acknowledgments

This work was funded by the American Council of Learned Societies, the OSU Humanities and Arts Discovery Theme and NSF grant BCS-1452674 (Roberts \& Tonhauser). Special thanks to members of the OSU pragmatics working group for their helpful feedback on this work.

## References

Abrusán, M. (2013). A note on quasi-presuppositions and focus. Journal of Semantics, 30, 257-265.
Beaver, D., \& Clark, B. (2008). Sense and sensitivity: How focus determines meaning. Oxford: Wiley-Blackwell.
Beaver, D., Roberts, C., Simons, M., \& Tonhauser, J. (2017). Questions Under Discussion: Where information structure meets projective content. Annual Review of Linguistics, 3 .
Beckman, M. E., \& Ayers, G. (1997). Guidelines for ToBI labelling, version 3.0 [Computer software manual].
Bergen, L., \& Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. Topics in Cognitive Science, 7(2), 336-350.
Chierchia, G., \& McConnell-Ginet, S. (1990). Meaning and grammar. Cambridge, MA: MIT Press.

Eady, S., \& Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. JASA, 80, 402-415.
Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084), 998. doi: 10.1126/science. 1218633
Halliday, M. A. K. (1967). Notes on transitivity and theme in English (part 2). Journal of Linguistics, 3, 177-274.
Heim, I. (1983). On the projection problem for presuppositions. In M. Barlow, D. Flickinger, \& M. Westcoat (Eds.), WCCFL 2 (p. 114-125).
Lai, C. (2012). Rises all the way up: The interpretation of prosody, discourse attitudes and dialogue structure (Unpublished doctoral dissertation). University of Pennsylvania.
Most, R., \& Saltz, E. (1979). Information structure in sentences: New information. Language and Speech, 22, 89-95.
Paul, H. (1880). Prinzipien der Sprachgeschichte. Tübingen: Niemeyer.
Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. Semantics \& Pragmatics, 5, 1-69.
Rooth, M. (1992). A theory of focus interpretation. Natural Language Semantics, 1, 75-116.
Selkirk, E. O. (1996). Sentence prosody: Intonation, stress, and phrasing. In J. Goldsmith (Ed.), The handbook of phonological theory (p. 550-569). Blackwell.
Simons, M. (2001). On the conversational basis of some presuppositions. In Proceedings of Semantics and Linguistics Theory (SALT) XI (p. 431-448). Ithaca, NY: CLC Publications.
Simons, M., Beaver, D., Roberts, C., \& Tonhauser, J. (2017). The Best Question: Explaining the projection behavior of factive verbs. Discourse Processes, 54, 187-206.
Tonhauser, J. (2016). Prosodic cues to speaker commitment. In Proceedings of Semantics and Linguistic Theory (SALT) XXVI. Ithaca, NY: CLC Publications.
van der Sandt, R. (1992). Presupposition projection as anaphora resolution. Journal of Semantics, 9, 333-377.
von Stechow, A. (1990). Focusing and backgrounding operators. In W. Abraham (Ed.), Discourse particles (p. 3784). Amsterdam: John Benjamins.

Wagner, M. (2012). Contrastive topics decomposed. Semantics and Pragmatics, 5(8), 1-54.

# A Common Neural Component for Finger Gnosis and Magnitude Comparison 

Terrence C. Stewart (tcstewar@uwaterloo.ca)<br>Centre for Theoretical Neuroscience, University of Waterloo, 200 University Avenue W, Waterloo, ON, N2L 3G1, Canada<br>Marcie Penner-Wilger (mpennerw@uwo.ca)<br>Department of Psychology, King's University College at Western University<br>266 Epworth Avenue, London, ON, Canada N6A 2M3a<br>Rylan J. Waring (waringr@uoguelph.ca)<br>Department of Psychology, University of Guelph, Guelph, Canada<br>Michael L. Anderson (michael.anderson@fandm.edu)<br>Department of Psychology, Franklin \& Marshall College, Lancaster, PA 17604 USA<br>Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA


#### Abstract

Finger gnosis (the ability to identify which finger has been touched) and magnitude comparison (the ability to determine which of two numbers is larger) are surprisingly correlated. We present a spiking neuron model of a common component that could be used in both tasks: an array of pointers. We show that if the model's single tuned parameter is set to match human accuracy performance in one task, then it also matches on the other task (with the exception of one data point). This provides a novel explanation of the relation, and proposes a common component that could be used across cognitive tasks.


Keywords: finger gnosis; magnitude comparison; spiking neurons; neural engineering framework, numerical cognition

## Introduction

Finger gnosis, the ability to differentiate which finger has been touched, in absence of visual feedback, is related to math performance (Fayol, Barrouillet \& Marinthe, 1998; Noël, 2005; Penner-Wilger et al., 2007, 2009, 2014, 2015). Finger gnosis is commonly measured using a finger localization task (Baron, 2004; Noël, 2005), wherein the participant's hand is occluded from their view while a finger, or two fingers, are touched. The participant is then asked to indicate the touched finger(s). Performance is measured in terms of number of fingers correctly identified.

Finger gnosis ability predicts performance on a variety of math measures in children, both concurrently and longitudinally ( $\beta$ 's range from . 22 to .36 ; Fayol et al., 1998; Noël, 2005; Penner-Wilger et al., 2007, 2009). Finger gnosis ability also predicts performance on a variety of math measures in adults ( $\beta$ 's range from .21 to .30 ; PennerWilger et al., 2014, 2015). The relation between finger gnosis and math skill is reproducible across labs, different samples, age groups, and measures of math skill, despite controlling for many other variables (e.g., visuo-spatial working memory, finger agility, processing speed, and nonverbal IQ).

The relation between finger gnosis and math skill is partially mediated by symbolic number comparison performance (Penner-Wilger et al., 2009, in prep.). In symbolic number comparison tasks, participants are shown two digits (e.g., 2 3) and asked to indicate which number is
more (or in some variants asked to compare a target digit to a standard). One robust finding in number comparison is the distance effect - performance is faster and more accurate when numbers are father apart in magnitude (e.g., 2 7) than when they are closer together (e.g., 2 3; Moyer \& Landauer, 1967). The distance effect is proposed to reflect mapping between numerals and their associated magnitude, with greater distance effects reflecting noisier mappings (Dehaene, Dehaene-Lambertz \& Cohen, 1998; cf. Lyons, Nuerk \& Ansari, 2015). Children who perform better in finger gnosis, reflecting a more precise finger representation, also demonstrate smaller distance effects in number comparison, reflecting a more precise number representation (Penner-Wilger et al., 2009).

Why are finger gnosis and math performance, specifically a task indexing the precision/strength of number representations, related? On the redeployment view (PennerWilger \& Anderson, 2008, 2013), the relation between finger gnosis and number representation arises because the two tasks use overlapping neural substrates. On this view, the relation is an example of neural reuse, the use of local regions of the brain to support multiple tasks across domains (Anderson, 2010, 2014). Neural reuse is a dynamic process, impacting the functional organization of the brain across both evolutionary and developmental time, whereby individual regions of the brain contribute to multiple highlevel uses (e.g., finger representation and number representation). There are two forms of neural reuse: redeployment and neuromodulation. In redeployment, the same brain region supports multiple uses, across evolutionary and/or developmental time, while maintaining the same operation (Anderson, 2014). In neuromodulation, the same brain region supports multiple uses, at any given point in developmental time, without maintaining the same operation - its operation is modulated as a result of internal or external variables (Anderson, 2014; Bargmann, 2012; Marder, 2012). The redeployment view posits that the behavioural link between finger and number representations is at least partially explained by neural reuse, and that the specific type of neural reuse involved is redeployment. Thus, one (or more) local brain regions, over evolutionary
and/or developmental time, has come to perform the same operation in support of both uses.

In support of the redeployment view, regions associated with finger gnosis are activated during tasks requiring the representation of number (Andres, Michaux \& Pesenti, 2012; Dehaene et al., 1996; Zago et al., 2001), rTMS and direct cortical stimulation disrupt both finger gnosis and tasks requiring the representation of number (Rusconi, Walsh, \& Butterworth, 2005; Roux et al., 2003), and there is interference between tasks involving finger gnosis and tasks requiring the representation of number (Brozzoli et al., 2008). Zago et al. (2001) pinpointed a region of overlap between finger and number representation in the leftprecentral gyrus (-42, 0, 38). Penner-Wilger and Anderson (2011) conducted a meta-analysis of imaging data to determine the full complement of tasks, across domains, that this ROI was implicated in, with the goal of identifying common requirements across tasks/uses to guide structurefunction mapping. In addition to number and finger representation tasks, the ROI was implicated in generation, inhibition and order tasks. Common requirements across these uses were identified, including ordered storage and mapping, and a candidate working that could implement both these requirements was proposed - an array of pointers. An array is an ordered group, meeting the requirements for ordered storage, and a pointer is a data structure that designates a memory location and can indicate different data types. Thus, an array of pointers allows for storage and access of ordered elements, which are able to point to-or index-representations or locations in memory, allowing for mapping between different representational forms.

The neural overlap between finger and number representation could reflect redeployment, wherein the brain region is reused in both tasks while retaining the same operation. Alternatively, the overlap could reflect neuromodulation, wherein the operation of the region is modulated. In the current paper, we use computational modelling as a means of demonstrating whether the same proposed working -an array of pointers- could contribute to both number and finger representation. The goals of the current research are to evaluate the redeployment view and proposed shared working by (1) providing an in-principal demonstration that the same working could contribute to both uses, (2) determining the psychological plausibility of the model by comparing it to human performance on finger gnosis and number comparison tasks, and (3) differentiating between support for redeployment (same ROI, same working) over neuromodulation (same ROI, but different working).

## Common Component: A Cognitive Pointer

The core theoretical claim here is that both finger gnosis and magnitude comparison could plausibly make use of a neural system that is able to store a list of items, and each of those items can be used to indicate other information. For example, these items could mean a particular number (e.g.

ONE or THREE) or they could mean any other known concept. For the purposes of this paper, we choose these vectors randomly, but we could use other vector-based representation methods such as LSA or word2vec.

To be explicit about what we mean by such a system, let us define it mathematically. First, we need a (small) set of numerical values which are our "pointers": $p_{1}, p_{2}, p_{3}, p_{4}$, and $p_{5}$. For the purposes of this paper, we keep the size of this set to 5 (the number of fingers on a hand). Each of these pointers is a numerical vector, and different values can have different meanings. For example, there could be one value that means the number ONE, with other values meaning other concepts like DOG.

In the absence of input, these pointers should not change their value. However, we also need some way of changing their value when needed. For this, we need two things: a new input value $x$ and a way to indicate which pointer should be set to the new value. This input control we call a mask m and it is a list of values indicating which pointer should be set. For example, if $m=[0,1,0,0,0]$, then the input $x$ will be set to the second pointer $p_{2}$.

Mathematically, we can write this as follows, where $i$ indexes the different pointers:

$$
p_{i} \Leftarrow \begin{cases}p_{i} & \text { if } m_{i}=0  \tag{1}\\ x & \text { if } m_{i}=1\end{cases}
$$

We postulate that the two tasks use this component as follows. For the finger gnosis task, consider what happens if two fingers are touched, the index finger and the ring finger. We can treat each pointer as a separate finger, and load in a vector that means TOUCHED into the correct pointers by setting $x=$ TOUCHED and $m=[0,1,0,1,0]$.

For the magnitude comparison task, we load the first value into the first pointer and the second value in the second pointer. For the case of comparing 5 and 7 this means setting $x=$ FIVE and $m=[1,0,0,0,0]$, and afterwards setting $x=$ SEVEN and $m=[0,1,0,0,0]$. Over time, this process proceeds stepwise as follows, and maintains its state as shown:

## Finger Gnosis Task

| $\boldsymbol{x}$ | $\boldsymbol{m}$ | $\boldsymbol{p}_{1}$ | $\boldsymbol{p}_{2}$ | $\boldsymbol{p}_{3}$ | $\boldsymbol{p}_{4}$ | $\boldsymbol{p}_{5}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| -- | 00000 | -- | - | -- | - | -- |
| TOUCHED | 01010 | -- | TOUCHED | -- | TOUCHED | -- |
| -- | 00000 | -- | TOUCHED | -- | TOUCHED | -- |
| Magnitude Comparison Task |  |  |  |  |  |  |


| $\boldsymbol{x}$ | $\boldsymbol{m}$ | $\boldsymbol{p}_{1}$ | $\boldsymbol{p}_{2}$ | $\boldsymbol{p}_{3}$ | $\boldsymbol{p}_{4}$ | $\boldsymbol{p}_{5}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| -- | 00000 | - | -- | -- | -- | -- |
| FIVE | 10000 FIVE | -- | -- | -- | -- |  |
| SEVEN | 01000 FIVE | SEVEN | -- | -- | -- |  |
| -- | 00000 FIVE | SEVEN | -- | -- | -- |  |

## Neural Implementation

While the above algorithm gives us a conceptual understanding of this array of cognitive pointers, we also want to determine how neurons could implement such an algorithm. By examining this neural mechanism, we can
gain insights into how accurate it would be in different conditions, and hopefully gain insight into individual differences and cognitive deficits.

For our neurons, we use standard leaky-integrate-and-fire (LIF) neurons. These increase in voltage given their input, and emit a spike and reset when the voltage reaches a threshold. These spikes are transmitted to all neurons that the spiking neuron is connected to, with synaptic weights controlling how much current is added to (or subtracted from) the target neuron each time a spike occurs. Each connection also has a post-synaptic time constant that controls the time it takes for a spike's effect to decay away.


Figure 1: Neural implementation of an array of pointers. Only two pointers are shown.

Figure 1 shows the basic approach used to implement this functionality. The groups of neurons on the right store the individual pointer values $p_{1}, p_{2}$, etc. They are recurrently connected such that they will have stable firing patterns over time (i.e. whatever pattern of firing is present right now will cause a similar firing pattern in the near future).

The connections from $\boldsymbol{x}$ to the channels and from the channels to the pointers are all set such that the neurons simply pass along the value without altering it. That is, if we input a particular value $\boldsymbol{x}$, this will cause a particular (and unique) firing pattern in channel 1 and channel 2. These in turn will cause particular firing patterns in pointer 1 and pointer 2. If $\boldsymbol{x}$ is removed (i.e. set to zero), then the pointer patterns will stay as they were. Thus, they implement a memory of previously presented patterns.

However, we also want to be able to selectively set one or the other pointer. For this reason, we also include the $\boldsymbol{m}$ (mask) input. This can selectively inhibit the channel neurons. If these are inhibited, then they do not fire, and so do not affect the pointer neurons. So, if we want to set the value in pointer 2 only (and not change whatever is stored in pointer 1 ), then we inhibit channel 1 when inputting $\boldsymbol{x}$.

To actually create this network, we use the Neural Engineering Framework (Eliasmith \& Anderson, 2003) and the software toolkit Nengo (Bekolay et al., 2014). In this approach, we assume that $\boldsymbol{x}$ is a vector of some dimensionality that is smaller than the number of neurons in a group. This means that there is redundancy in the neural code, and the value $\boldsymbol{x}$ is distributed across the neural population. Here, for simplicity, we assume $\boldsymbol{x}$ is an 8 dimensional vector. Previous work (Crawford, Gingerich \& Eliasmith, 2013) has shown that 512-dimensions should be sufficient for high-level reasoning applications, but that is not needed for the tasks considered here.

Within each group of 400 neurons, each individual neuron has a randomly chosen preferred vector. That is, each neuron will have some particular $\boldsymbol{x}$ value for which it fires the fastest. This is a generalization of the standard preferred direction vectors observed throughout cortex (e.g. Georgopoulos et al., 1986).

To generate the actual connections between neurons, the NEF uses least-squares minimization to directly solve for the optimal synaptic connection weights that will do the best job of transferring a value $\boldsymbol{x}$ from one group to the next. This same process is used to generate the recurrent connections for the pointers.

The synaptic time constants were set to 10 ms for the feedforward connections (based on the fast AMPA synapses found in cortex) and 100 ms for the recurrent connections (based on the slower NMDA synapses found in recurrent connections in cortex).

It should be noted that there is nothing in the model so far that is fit to a particular task. The optimization of the connection weights is over all possible $\boldsymbol{x}$ values, not the particular $\boldsymbol{x}$ values that mean ONE or TWO or TOUCHED in the magnitude comparison and finger gnosis tasks themselves. This is meant to be a generic component, not one that is specialized for exactly these tasks.


Figure 2: Spiking activity for an example magnitude comparison task. Top row shows input to the model. Other rows show spiking neuron activity over time. The text indicates which vector $\boldsymbol{x}$ is represented by the pattern of activity. Note that pointer 1 and pointer 2 maintain their pattern (approximately) after the input has been removed.

Figure 2 shows the neural activity when this array of pointers is used to store two numbers. Initially, both pointer 1 and pointer 2 are firing with some random background firing rate. At $\mathrm{t}=0.2 \mathrm{~s}$, we set the input $\boldsymbol{x}$ to be the vector for FIVE (randomly chosen) and set the mask such that channel 1 is the only group not being inhibited. This drives the neurons in pointer 1 to also fire with the pattern for FIVE. At $\mathrm{t}=0.4 \mathrm{~s}$, we change the input to SEVEN and change the mask so that channel 2 is not inhibited. This drives pointer 2 to represent SEVEN. Importantly, after $\boldsymbol{x}$ is removed, the neurons in pointer 1 and pointer 2 retain their firing pattern.

While the system described above behaves as desired, note that it is not perfect. The neurons in channel 1 and channel 2 are not perfectly inhibited. Also, the neurons in pointer 1 and pointer 2 do not perfectly maintain exactly the desired firing pattern. This is as expected, as neurons only approximate the desired functions. We can now test whether the resulting model can still perform the two tasks, and whether the errors made by the model due to these imperfections are comparable to the errors made by people.

## Task 1: Magnitude Comparison

The first task believed to make use of this component is magnitude comparison. Two single-digit numbers are presented, and the system must decide which is larger.

To implement this task neurally, we add two new neural groups. First, a comparison group, which takes as input the vectors in the first two pointer populations. This means that the comparison group has a 16-dimensional input, with the value from $p_{l}$ as the first 8 dimensions and the value from $p_{2}$ as the second 8 dimensions. Second, we have an accumulator. This takes as input a single number which should be positive if the first number is larger, and should be negative if the second number is larger. This is recurrently connected to itself, so that even for small inputs, it will eventually build up until it reaches a threshold, making it a standard accumulate-to-threshold decisionmaking system.


Figure 3: The magnitude comparison model. The dotted area indicates the array of pointers.

The connection between the comparison and accumulator neurons needs to convert from 16 dimensions (the two numbers being presented) to 1 dimension (which number to choose). We implement this function by generating 2000 training examples of randomly chosen digits, along with the correct answer of +1 if the first number is larger, and -1 if the second number is larger. We then used Nengo to find the optimal connection weights between these neurons to best approximate this mapping.

To evaluate this model, we collected human participant data from 88 undergraduate students at King's University College who received course-credit for their participation (age: $M=21.28$ years, $S D=3.8$ years; 64 female). Two single digit numbers (ranging from 1 to 9 ) were presented simultaneously on an iPad screen. Participants were asked to choose the numerically larger number as fast as they could without making any errors. Stimuli remained on the screen for 7800 ms or until the participant made a choice, and the time between trials was 1000 ms . Participants performed a total of 72 trials. Dependent measures were reaction time and percent error.


Figure 4: Results from participants and from the model. Standard errors of the mean are shown.

The participant data (Figure 4) displays the expected distance effect: as the difference between the digits increases, accuracies improve and reaction times are quicker. For confirmation, a repeated measures ANOVA with a Greenhouse-Geisser correction revealed that mean RTs differed significantly between distances, $F(4.41$, $383.91)=46.96, p<.01$. A second repeated-measures ANOVA revealed that mean percent error also differed significantly between distances, $F(7,609)=21.37, p<.01$.

Importantly, the model data shows the same effects. To achieve the quantitative fit for the accuracy measure (Figure 4, left side), we only fit one parameter: the strength of the inhibition $\boldsymbol{m}$. That is, rather than having it always be strong enough to completely stop all neurons in the channel from firing ( $s_{\text {inhibition }}=1.0$ ), we allowed this value to be reduced. This causes some "leakage", where values meant for one pointer slightly affect the other pointers, since the other channels are not perfectly inhibited. If there are more than 2 pointers, we assume that the inhibition gets proportionally stronger for pointers farther from the target pointer. For the data shown above, $s_{\text {inhibition }}=0.875$. Surprisingly, this distance effect occurs even though the neural activity pattern for each digit is randomly chosen. See (Stewart \& Penner-Wilger, 2017) for further analysis.

For the reaction time data (Figure 4, right side), two additional parameters were fit. First, we added a fixed reaction-time value (i.e. the amount of time needed for perception and the motor action) $T_{\text {fixed }}$. Second, we allowed a scaling factor on the connection from the comparison neurons to the accumulator neurons. This controls the rate of evidence accumulation $S_{\text {evidence. }}$. This is a common feature of decision-making models. After fitting, $T_{\text {fixed }}=290 \mathrm{~ms}$ and $S_{\text {evidence }}=5.9$. All other parameters in the model were left at their default values.

## Task 2: Finger Gnosis

In the finger gnosis task, two fingers are touched on the participant's hand while that hand is occluded from their view. They must then report which fingers were touched.

To implement this task, we use the same array of pointers, but connect it to a different set of neurons, as depicted in Figure 5. The first group of neurons takes the input from all the pointers and combines them together as one vector. The second group stores the reported answer. As with the previous task, we use Nengo to find the connection weights
that best approximate the function between the combination neurons and the answer neurons. In this case, however, rather than determining which value is larger, here we do not need to perform any complex operation as we just need to extract the information that is already encoded in the neurons. Thus, here we use Nengo to approximate the identity function, where the output is the same as the input.


Figure 5: The finger gnosis model. Only 2 pointers are shown, but the full model uses 5 pointers.

Importantly, if this same array of pointers is to be used in two different tasks, a flexible neural routing system would be needed, so that the output of the pointer array can be sent to this combination system when doing the finger gnosis task, and sent to the comparison system when doing magnitude comparison. We have previously shown how to implement such a routing system using a model of the cortex-basal ganglia-thalamus loop (Stewart, Choo, \& Eliasmith, 2010), and so do not consider that here.

To evaluate this model, we used the same 88 undergraduates as for the first task. Participants first performed the magnitude comparison task, followed by the finger gnosis task as part of a larger study. As shown in Figure 6, a repeated-measures ANOVA revealed that mean percent error differed significantly between distances, $\mathrm{F}(3$, $261)=6.88, p<.01$.

Figure 6 also shows the model performance. Importantly, no parameters were tuned to achieve this result. We used $s_{\text {inhibition }}=0.875$, as that was the best fit value in the first task, and all other parameters were left as they were. The model is statistically significantly different at a distance of 1 , but does not statistically differ for distances 2,3 , and 4 .


Figure 6: Results from participants and model for the finger gnosis task. Standard errors of the mean are shown.

Since the only tuned parameter in the model is $s_{\text {inhibition }}$, we also examined how the model's performance changes on the two tasks as this parameter is varied (Figure 7). From this, we note that the error rates on these two tasks change drastically, given small changes in this parameter. This indicates a strong connection between the model's
performance on one task and on the other. The fact that a similar parameter value is needed in each task in order to fit the human data lends support to the idea that there is a shared working that is redeployed for these two tasks.


Figure 7: Effects of changing $s_{\text {inhibition }}$ in both tasks.

## Conclusions

On the redeployment view (Penner-Wilger \& Anderson, 2008, 2013), finger gnosis and math ability are linked because at least one local brain region, over evolutionary and/or developmental time, has come to perform the same operation in support of both finger and number representation. The goal of the current research was to evaluate the redeployment view and the proposed shared operation - an array of pointers (Penner-Wilger \& Anderson, 2011). To this end, we built a computational model to perform both the standard finger gnosis and number comparison tasks. We then compared the performance of this model to human performance data (RT and accuracy) and showed a close match on both tasks with one parameter ( $s_{\text {inhibition }}$ ) tuned to a common value.

First, our work provides an in-principal demonstration that the same working - an array of pointers - could contribute to multiple uses, as the same system successfully performed two different tasks. Our previous meta-analysis (Penner-Wilger \& Anderson, 2011) also indicates this region may be involved in a variety of other tasks, which we intend to include in future research.

Second, given that the model could successfully perform both tasks using the same operation, and that the model performance mirrored that of human participants, it is a psychologically plausible explanation, which lends support for the view that the observed neural overlap between finger and number representation reflects redeployment (same ROI, same working) rather than neuromodulation (same ROI, different working). It follows that damage to the ROI should impact performance on both finger gnosis and number comparison tasks. We are currently testing this in our computational model and it could be tested in human participants using rTMS applied to our ROI in the left precentral gyrus. Previous work using rTMS applied to the left angular gyrus has already been shown to disrupt performance on both tasks (Rusconi et al., 2005).

Third, by offering another concrete instance of the reuse of a basic operation in a high-level, abstract cognitive task, the model does not just bolster the neural reuse framework, but also serves the goal of enhancing our understanding of the nature of and processes involved in numerical cognition.

Finally, modeling efforts like this potentially enhance our efforts to map the functional structure of the brain. We currently lack the capacity to determine in vivo when neuromodulation has changed the underlying configuration of a local neural network, which hinders our ability to attribute function to structure. This approach offers some first steps toward developing reliable methods for detecting changes to the underlying operation a given local region supports, thereby refining our efforts to describe what the brain is actually doing at any given time.

## Acknowledgements

This research was supported by a grant from King's University College at Western University to MPW. TCS and MPW contributed equally to the research.

## References

Anderson, M.L. (2010). Neural reuse: A fundamental organizational principle of the brain. (Target article) Behavioral and Brain Sciences, 33(4): 245-66.
Anderson, M.L. (2014). After Phrenology: Neural Reuse and the Interactive Brain. Cambridge, MA: MIT Press.
Andres, M., Michaux, N., \& Pesenti, M. (2012). Common substrate for mental arithmetic and finger representation in the parietal cortex. Neuroimage, 62(3), 1520-1528.
Bargmann, C. I. (2012). Beyond the connectome. Bioessays, 34(6), 458-465.
Baron, I. S. (2004). Neuropsychological Evaluation of the Child. New York, NY: Oxford University Press.
Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T.C., Rasmussen, D., Choo, X., Voelker, A., \& Eliasmith, C.. (2014). Nengo: a python tool for building large-scale functional brain models. Frontiers in Neuroinformatics.
Brozzoli, C., Ishihara, M., Göbel, S. M., Salemme, R., Rossetti, Y., and Farnè, A. (2008). Touch perception reveals the dominance of spatial over digital representation of numbers. Proc. Natl. Acad. Sci. U.S.A. 105, 5644-5648.
Crawford, E., Gingerich, M., \& Eliasmith, C. (2013). Biologically plausible, human-scale knowledge representation. $35^{\text {th }}$ Cognitive Science Society, 412-417.
Dehaene, S., Dehaene-Lambertz, G., \& Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. Trends in Neuroscience, 21, 355-361.
Dehaene, S., Tzourio, N., Frak, V., Raynaud, L., Cohen, L., Mehler, J., \& Mazoyer, B. (1996). Cerebral activations during number multiplication and comparison: A PET study. Neuropsychologia, 34, 1097-1106.
Eliasmith, C. \& Anderson, C.. (2003). Neural engineering. MIT Press, Cambridge, MA.
Fayol, M., Barrouillet, P., \& Marinthe, C. (1998). Predicting arithmetical achievement from neuro-psychological performance. Cognition, 68, B63-B70.
Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. Science 233, 1416-1419.

Lyons, I. M., Nuerk, H.C. \& Ansari, D. (2015). Rethinking the implications of numerical ratio effects for understanding the development of representational precision and numerical processing across formats. JEP: General, 144 (5):1021-1035.
Marder, E. (2012). Neuromodulation of neuronal circuits: back to the future. Neuron, 76(1), 1-11.
Moyer, R. S., \& Landauer, T. K. (1967). Time required for judgments of numerical inequality. Nature 15,1519-1520.
Noël, M.-P. (2005). Finger gnosia: A predictor of numerical abilities in children? Child Neuropsychology, 11, 413430.

Penner-Wilger, M., \& Anderson, M. L. (2013). The relation between finger gnosis and mathematical ability: Why redeployment of neural circuits best explains the finding. Frontiers in Psychology, 4, 877.
Penner-Wilger, M., \& Anderson, M. L. (2011). The relation between finger gnosis and mathematical ability: Can we attribute function to cortical structure with cross-domain modeling? 33 ${ }^{\text {rd }}$ Cognitive Science Society, 2445-2450.
Penner-Wilger, M., \& Anderson, M. L. (2008). An alternative view of the relation between finger gnosis and math ability. $30^{\text {th }}$ Cognitive Science Society. 1647-1652.
Penner-Wilger, M., Fast, L., LeFevre, J., Smith-Chant, B. L., Skwarchuk, S., Kamawar, D., \& Bisanz, J. (2009). Subitizing, finger gnosis, and the representation of number. $31^{\text {st }}$ Annual Cognitive Science Society, 520-525.
Penner-Wilger, M., Fast, L., LeFevre, J., Smith-Chant, B. L., Skwarchuk, S., Kamawar, D., \& Bisanz, J. (2007). The foundations of numeracy: Subitizing, finger gnosia, and fine-motor ability. $29^{\text {th }}$ Cog. Science Society, 1385-1390.
Penner-Wilger, M., Waring, R. J., \& Newton, A. T. (2014). Subitizing and finger gnosis predict calculation fluency in adults. $36^{\text {th }}$ Cognitive Science Society, 1150-1155.
Penner-Wilger, M., Waring, R. J., Newton, A. T., \& White, C. (2015). Finger gnosis and symbolic number comparison as robust predictors of adult numeracy. $37^{\text {th }}$ Conference of the Cognitive Science Society, 2963.
Roux, F.-E., Boetto, S., Sacko, O., Chollet, F., \& Tremoulet, M. (2003). Writing, calculating, and finger recognition in the region of the angular gyrus: a cortical study of Gerstmann syndrome. J. Neurosurgery, 99, 716-727.
Rusconi, E., Walsh, V., \& Butterworth, B. (2005). Dexterity with numbers: rTMS over left angular gyrus disrupts finger gnosis and number processing. Neuropsychologia, 43(11), 1609-1624.
Stewart, T.C. \& Penner-Wilger, M. (2017). Analysis of a common neural component for finger gnosis and magnitude comparison. Int. Conf. on Cognitive Modelling.
Stewart, T.C., Choo, X., \& Eliasmith, C.. (2010). Symbolic reasoning in spiking neurons: a model of the cortex/basal ganglia/thalamus loop. $32^{\text {nd }}$ Annual Meeting of the Cognitive Science Society, 1100-1105.
Zago, L., Pesenti, M., Mellet, E., Crivello, F., Mazoyer, B., \& Tzourio-Mazoyer, N. (2001). Neural correlates of simple and complex mental calculation. NeuroImage, 13, 314-327.

# Flexible integration of a navigable, clustered environment 

Marianne Strickrodt<br>Max Planck Institute for Biological Cybernetics<br>Heinrich H. Bülthoff<br>Max Planck Institute for Biological Cybernetics<br>Tobias Meilinger<br>Max Planck Institute for Biological Cybernetics


#### Abstract

The representation of navigable space, consisting of multiple interconnected spaces, yet is not well understood. We examined different levels of integration within memory (local, regional, global). Participants learned two distinctive regions of a virtual environment that converged at a common transition-point. Subsequently, we tested their memory with a pointing task, varying body alignment during pointing, corridor distance to and regional belonging of the target. Pointing latency increased with increasing distance to the target and when pointing into the other region. Further, alignment with local, regional and global reference frames were found to facilitate pointing latency. These findings suggest that participants memorized local corridors, clustered corridors into regions, and also formed global reference frames, thus, represented the environment on multiple levels of integration. They are inconsistent with conceptions of spatial memory for navigable environments based either on exclusive representation within a single reference frame or exclusive reliance on local reference frames.


# The Semantic Spaces of Child-Directed Speech, Child Speech and Adult-directed Speech: a Manifold Perspective 

Hao Sun (hsun7@buffalo.edu)<br>Department of Linguistics, University at Buffalo Buffalo, NY 14260 USA<br>John Pate (johnpate@buffalo.edu)<br>Department of Linguistics, University at Buffalo<br>Buffalo, NY 14260 USA


#### Abstract

Child-directed speech (CDS) is a talking style adopted by caregivers when they talk to toddlers (Snow, 1995). We consider the role of distributional semantic features of CDS in language acquisition. We view semantic structure as a manifold on which words lie. We compare the semantic structure of verbs in CDS to the semantic structure of child speech (CS) and adult-directed speech (ADS) by measuring how easy it is to align the manifolds. We find that it is easier to align verbs in CS to CDS than to align CS to ADS, suggesting that the semantic structure of CDS is reflected in child productions. We also find, by measuring verbs vertex degrees in a semantic graph, that a mixed initialized set of verbs with high degrees and medium degrees has the best performance among all alignments, suggesting that both semantic generality and diversity may be important for developing semantic representations.


Keywords: child-directed speech; lexical development; manifold learning; distributional semantics; graph theory

## Introduction

One of the biggest puzzles in cognitive science is how children learn language from language input, namely childdirected speech. Child-directed speech is characterized by simplified sentence structures, restricted vocabulary, exaggerated intonation, and hyperarticulation, and previous work has proposed that these features facilitate language acquisition (Golinkoff and Alioto, 1995; Snow, 1995; Thiessen, Hill, and Saffran, 2005). Here, we compare the semantic spaces of child speech, child-directed speech and adult-directed speech, spanned by verbs, using state-of-theart computational tools.

The contributions of this paper are both theoretical and methodological. Theoretically, we explore various proposals about roles of verbs meanings in CDS, represented using a state-of-the-art distributional semantics approach. Distributional methods map each word to a point in highdimensional space so that words with similar meanings are near each other. We view the semantic structure of the vocabulary as a high-dimensional surface in this space, called a manifold, and compare manifolds estimated from CDS to manifolds estimated from child speech (CS) and adultdirected speech (ADS). Young children often broaden the use
of nouns and verbs and we model such differences in word meaning as a mismatch of data points in a semantic space.

Methodologically, we adapt a novel semi-supervised manifold alignment algorithm to compare semantic spaces (Ham et al, 2005), which maps two manifolds into a common subspace to measure the similarity of these manifolds. This algorithm takes as input a subset of initial points that must be aligned (i.e., pairs of points, one on each manifold, that correspond to the same verb), and produces an alignment for the rest of the verbs. We then measure the similarity of the manifolds in terms of the accuracy of the alignment: how often a verb is mapped to the same region of the common subspace.

We find that alignment between the CS and CDS is more accurate than the alignment between CS and ADS. Additionally, we obtain more accurate alignments when using verbs with many nearest neighbors (which have broader meanings) as the initial points than verbs with few near neighbors. Together, these results indicate that the semantic structure of CS reflects the semantic structure of CDS, and verbs with broad meanings may provide useful cues to children in acquiring the overall semantic structure of verbs. On the one hand, what children can learn from CDS deviates semantically from unfamiliar conversations in ADS, which suggests that further learning is required. On the other hand, caregivers might align their semantic spaces to children's semantic spaces, which lies within the general framework of conversational alignment (Pickering \& Garod, 2004).

## Model Setting

We combine models from two different traditions into a general framework of semantic representation. To compare the semantic spaces of CS, CDS and ADS, we use a manifoldbased algorithm. The similarities between semantic spaces are measured by how easy it is to map one semantic space to another. We represent the meaning of each verb by using the global vector model (Pennington, Socher \& Manning, 2014) to embed words into a 50-dimensional space, which we call a semantic space. . Following the associationist tradition in psychology (Anderson, 1973), we represent the meaning structure of the verbal lexicon as a whole by considering how a collection of verbs is situated in this space, as expressed by a neighborhood graph (Steyvers \& Tenenbaum, 2005).

Estimating verb meanings from different datasets produces different semantic spaces, and we compare the spaces using a semisupervised manifold alignment algorithm (Ham et al., 2005). This algorithm maps verbal semantic graphs into a common semantic space and discovers the data point correspondences by finding pairs of points with the smallest Euclidean distances.

## Lexical Semantic Representation

The past three decades saw efforts to model the mental representation of concepts (Launder \& Dumais, 1997). The inspiration for recent computational work on lexical semantics dates back to Harris's (1954) hypothesis that synonymous words appear in similar contexts.

One of the most successful semantic representation models is proposed by Launder \& Dumais (1997), known as Latent Semantic Analysis (LSA), which uses word-context cooccurrence matrices to produce a low-dimensional representation by singular value decomposition. The lexical semantic representation model used in this paper is based on a state-of-the-art algorithm, GloVe (Pennington, Socher \& Manning, 2014), which is an extension of LSA. Instead of explicitly decomposing a word-context co-occurrence matrix, GloVe implicitly decomposes a word-context logfrequency matrix. GloVe uses a weighted regression objective function to reconstruct a log word-context count matrix $\log (X)$ with bias terms, as shown in Equation (1), where $w$ and $b$ are bias vectors, $X$ is the co-occurrence matrix and $f$ is a heuristic weighting function. The optimization problem is iteratively solved using AdaGrad (Duchi, Hazan \& Singer, 2011).

$$
\begin{equation*}
J=\sum_{i, j=1}^{V} f\left(X_{i j}\right)\left(w_{i}^{T} \widetilde{w}_{j}+b_{i}+\tilde{b}_{j}-\log X_{i j}\right)^{2} \tag{1}
\end{equation*}
$$

Even though GloVe has better performance than traditional singular-value-decomposition-based LSA, careful analysis of the objective function suggests that GloVe is fundamentally probabilistic matrix factorizations (Levy \& Goldberg, 2014).

## Semantic Graphs

The manifold alignment algorithm we use approximates the underlying manifold by constructing a similarity graph $G=$ ( $V, E$ ), where the vertex set $V$ is the set of verbs and the edge set $E$ is a set of pairs of verbs that are near to each other. The weight of an edge is set to the cosine similarity between the verbs associated by the edge. The degree of a vertex is the sum of weights of all the edges linking to the vertex. In semantic networks, vertex degrees can be interpreted as contextual diversity. There are several ways to build such a similarity graph. Ozaki et al (2011) found that undirected mutual $k$ nearest neighbor ( $m k N N$ ) graphs give good performance for alignment of natural language data, so we use $m k N N$ graphs. An $m k N N$ graph has an edge ( $v_{1}, v_{2}$ ) if either $v_{1}$ or $v_{2}$ is within the $k$ nearest neighbors of the other. We set $k$ to 15 for the first experiment. In the second experiment, we increase $k$ to 20 to better investigate the degree effects. The unnormalized graph Laplacian $(L)$ of
graph $W$ is defined in Equation (2). D is the degree matrix, a diagonal matrix with vertex degrees on the diagonal.

$$
\begin{equation*}
L=W-D \tag{2}
\end{equation*}
$$

We use a symmetric graph Laplacian normalized by vertex degree (Shi \& Malik, 2000), as

$$
\begin{equation*}
L_{\text {sym }}=D^{-1 / 2} L D^{-1 / 2}=I-D^{-1 / 2} W D^{1 / 2} \tag{3}
\end{equation*}
$$

## Aligning Semantic Spaces

We compare the semantic spaces of CS, CDS and ADS using the semisupervised manifold alignment algorithm. A manifold is defined as a topological structure with every local point with a neighborhood similar to a Euclidean space. The goal of the manifold alignment algorithm is to pair up data points from two high-dimensional data sets. For example, the algorithm aims to match give in CS to give in CDS. A semisupervised algorithm, using both labeled and unlabeled data as input, combines the strength of supervised and unsupervised learning. The general goal of manifold alignment is to map two high-dimensional data sets to a common low-dimensional space simultaneously (Ham et al., 2005), which essentially is an extension of manifold-based nonlinear dimensionality reduction (Belkin \& Niyogi, 2003). Manifold-based methods are based on the geometric assumption that data in high dimensional space lie in lowdimension manifolds.
Ham et al.'s algorithm defines a function $f$ that maps the first manifold to a common space, and a function $g$ that maps the second manifold to a common subspace. These functions strike a tradeoff between mapping labeled pairs to the same point in the common space, and respecting local structure on the original manifolds as expressed by the graph Laplacian $L^{x}$ for the first space and $L^{y}$ for the second space. As we have both labeled ( $l$ ) and unlabeled (u) points, $L^{x}$ and $L^{y}$ are block matrices:

$$
L^{x}=\left[\begin{array}{ll}
L_{l u}^{x} & L_{u l}^{x}  \tag{4}\\
L_{l u}^{x} & L_{u u}^{x}
\end{array}\right]
$$

The cost of the mapping is then:

$$
\begin{equation*}
\tilde{C}(\boldsymbol{f}, \boldsymbol{g})=\frac{C(\boldsymbol{f}, \boldsymbol{g})}{\boldsymbol{f}^{T} \boldsymbol{f}+\boldsymbol{g}^{T} \boldsymbol{g}} \tag{5}
\end{equation*}
$$

where $\mu$ expresses the tradeoff between mapping points exactly and preserving local structure on the original manifolds. The first term is the sum of distances between paired data points in the common space, and the second two terms represent faithfulness to the graph Laplacian. Ham et al. point out that Equation 4 is unsuitable for optimization, since it ignores simultaneous scaling of $\boldsymbol{f}$ and $\boldsymbol{g}$, and so instead minimize the Rayleigh quotient:

$$
\begin{equation*}
C(\boldsymbol{f}, \boldsymbol{g})=\mu \sum_{i}\left|f_{i}-g_{i}\right|^{2}+\boldsymbol{f}^{T} L^{x} \boldsymbol{f}+\boldsymbol{g}^{T} L^{y} \boldsymbol{g} \tag{6}
\end{equation*}
$$

We set $\mu$ to positive infinity to impose a hard constraint for labeled pairs to be mapped directly on top of each other.

The analytic solution to the optimization is then given by the generalized graph Laplacian $L^{z}$ in Equation 7.

$$
L^{z}=\left[\begin{array}{ccc}
L_{l l}^{x}+L_{l l}^{y} & L_{l u}^{x} & L_{l u}^{y}  \tag{7}\\
L_{u l}^{x} & L_{u u}^{x} & 0 \\
L_{u l}^{y} & 0 & L_{u u}^{y}
\end{array}\right]
$$

The semisupervised manifold alignment algorithm adopted from Ham et al. 2005 is described in Algorithm 1.

## Algorithm 1: Semisupervised Manifold Alignment Algorithm (Ham et al., 2005)

Input: data points from two data sets, with N initially aligned data point pairs
Output: a matching of data points

1. Construct similarity graphs $G_{1}, G_{2}$, for both data sets respectively, using $\mathrm{m} k N N$
2. Compute the symmetric graph Laplacians of $G_{l}$ and $G_{2}, L^{x}$ and $L^{y}$, using Equation (3)
3. Compute a graph Laplacian for a joint graph $L^{z}$ using Equations (6) and (7)
4. Compute the eigenvectors of $L^{z}$ and take eigenvectors corresponding to the smallest non-zero eigenvalues, the results of which are the vectors in a lower-dimensional space
5. Find the data points with smallest Euclidean distance weighted by the inverse of their respective eigenvalues

## Experiment Setup

## Corpora

The training set for CDS and CS is a combined data set from CHILDES (MacWhinney, 2000), which consists of all the data on American English-speaking monolingual 3 to 7 yearold children with typical language and cognitive development, excluding diary studies. To simplify data collection, only utterances annotated as child are considered child speech and only utterances annotated as mother and father are considered as child-directed speech. The CS and CDS corpora contain 5 million and 9 million word tokens, respectively. To prevent the CS from being similar to CDS purely due to priming effects, we divided the data into two halves so that the CDS and CS data were not drawn from the same contexts

Our ADS data is drawn from the spoken portion of the Corpus of Contemporary American English (COCA, Davies, 2008). Although this data may differ from more casual conversations, it provides a large amount of spontaneous speech in the form of unscripted conversations from 150 television and radio programs.

## Materials

The target words used in this model are all verbs, which are understudied in the literature. We included the first 100 English verbs acquired by infants (Fenson et al., 1994), the
most frequent 200 English verbs in adult language productions (Davies, 2008) and verbs that appear in three common constructions (Levin, 1993).

The classes of verbs are the ones that appear in 3 constructions: the ditransitive (John gave Mary a book), the locative (The man loaded hay onto a truck) and the conative (The police shot at the criminal). Since CHILDES suffers from data sparcity, verbs missing in either CS or CDS were excluded from analysis. We end up with 811 data points for CS, CDS and spoken COCA respectively.

## Data Preprocessing

The adult-directed speech data from spoken COCA and the child speech and child-directed speech data from CHILDES data were preprocessed using regular expressions. Verbs in different inflectional forms were treated as separate verb types.

## Model Training

Global Vector Training We used the implementation of GloVe from the Stanford NLP website to train 50dimensional vectors for each of our three datasets (Pennington, Socher \& Manning, 2014). We trained each set of vectors for 50 epochs with a context window size of 10 , used a frequency cut-off of 2 for the CS and CDS datasets and a cut-off of 10 for the ADS dataset.
Similarity Graph Construction We construct mkNN graphs consistently throughout this paper. In the first simulations, we fix the number of mutual nearest neighbors to 15 . In the second simulation, we test the effect of vertex degrees and we set the number of mutual nearest neighbors to 20 to increase the range of vertex degree.
Manifold Alignment The parameters that we need to specify in the manifold alignment module include the initial labeled alignments and the dimensionality of the manifold. In addition to the number of labeled data, the identity of the labeled data can also influence the quality of alignment. The dimensionality of the manifold controls the abstraction of semantic information contained in the word vectors. The lower the dimension, the more abstract the representation.

## Evaluation

Because the alignment algorithm pairs up labeled data points exactly, we only evaluate alignments on unlabeled data. We use a random alignment averaged over 5 times as the baseline condition. Ideally, corresponding data points from two data sets should be mutual nearest neighbor in the lower dimensional space. We relax the evaluation requirements by giving every alignment a k-nearest neighborhood evaluation radius. If one data point is one of the k -nearest neighbors of the corresponding point, we take it as a hit. When the evaluation neighborhood radius equals 1, the measures quantify the exact alignment.

## Simulation 1: Mapping CS to CDS and COCA

In this section, we demonstrate that CS-CDS alignment is a less demanding task than CS-COCA alignment even when potential priming effects from linguistic and non-linguistic contexts are removed. We also predict that with the increase of labeled data, the alignment accuracy also increases.

## Method

We performed verb semantic graph alignments of CS to CDS and to ADS for alignment spaces of dimensionality from 5 to 30. The unlabeled precisions are evaluated by the windowsize at 1 and at 20, as demonstrated in the contour heat maps in Figure 1. The colors of different areas in the contours indicate different levels of unlabeled accuracy and the data points with the same unlabeled accuracy are connected by the isolines in the maps.


Figure 1 Accuracies of mapping CS to CDS and COCA

## Results

The general trend is that the highest unlabeled precisions are found in the upper right corners of the contour maps whereas the lowest unlabeled precisions tend to lie close to the x -axis. The dimensionality of the embedding space can be interpreted as the granularity of children's representations.

The result of the alignments is demonstrated graphically in Figures 1 and 2. In the alignments from CS to CDS and CS to COCA, the CS-COCA alignment achieves only $50 \%$ to $60 \%$ of the unlabeled precision of the CS-CDS alignment. The unlabeled precision of the CS-CDS alignment is consistently higher than the unlabeled precision of the CSCOCA alignment across all conditions. Both alignments have much larger unlabeled accuracy than the random baseline.

The CS data are aligned to both the spoken COCA and CDS corpora. The CS-CDS alignment precision wins over the CS-COCA precision across all conditions. In other words, child speech is much easier to map to child-directed speech than to spoken COCA. This easier alignment can be interpreted as similarity in semantic spaces across corpora.

Since the CS and the CDS word vectors are trained on speech data from different experiments, the relative similarity between CS and CDS lexical semantics, this similarity does not reflect mere priming effects. There are two possible
interpretations for this result. First, the result can be viewed as an imitation effect in which children mirror child-directed speech semantically. Second, adult caregivers might adapt their mental representations to children's when they talk to children, which sits well with the conversational alignment theory (Pickering \& Garrod, 2004). The big semantic gap between initial language input and adult-to-adult conversations on TV shows or radios suggests that learning from CDS alone is not sufficient for real world language processing. Adapting to TV or radio conversations constitute one part of further learning, which supports a continuous theory of language development.

## Simulation 2: Semantic generality

In Simulation 2, we use a fixed list of labeled data to investigate the effect of initialization in alignment, instead of random initialization. The motivation is that language scientists argue for the importance of a few important "pathbreaking" word exemplars in language learning (Ninio, 1999; Goldberg, Casenhiser \& Sethuraman, 2004). Some words attract more vertices than others, which is known as preferential attachment in network growth (Steyvers \& Tenenbaum, 2005). We evaluate the proposal that semantically general verbs are better starting points for language learning than semantically specific verbs, by measuring the vertex degrees.


Figure 2 Unlabeled accuracies of CS-CDS and CS-COCA alignments with a random alignment as the baseline

The degree of a vertex measures the association between a vertex and its neighboring vertices. The prediction is that vertices with large degree are better labeled data than vertices with small degree. Cognitively, the verbs with high degree are semantically general verbs whereas the verbs with low degree are the ones with less general meanings.

## Method

Verbs are ranked based on their vertex degree in a semantic network. As shown in Table 1, what we use as labeled data is 100 verbs with the largest degrees, 100 with the smallest degrees, and medium-degree verbs with degree rank of 201
to 300 . We also mixed half of high degree verbs with half of medium degree verbs in the mixed condition. The baseline condition is averaged over 5 random initializations. We set the number of mutual nearest neighbors, the evaluation radius and the dimensionality all to 20 .

## Results

The alignment precisions shown in Figure 3 show a clear advantage of high-degree and medium degree conditions over the low degree condition, but both high-degree and lowdegree have below random performances. We can also see an advantage of medium degree initialization, which is parallel to the basic level categorization theories. When we use a mixed set of high-degree and medium-degree verbs, we get the best results on all the conditions, which suggests that a diverse-degree initialization facilitates semantic space alignment.

Table 1: Verbs with the largest, medium and smallest vertex degrees in ADS

| largest | medium | smallest |
| :---: | :---: | :---: |
| get | giving | tickles |
| go | tearing | points |
| want | taken | shooting |
| put | poured | design |
| think | tipping | tapping |

## General Discussion

In Simulation 1, we demonstrate that CS has semantic properties very similar to CDS in comparison to ADS. This result supports a usage-based approach to language acquisition: children imitate their caregivers. The results can be interpreted in multiple perspectives. First, the result suggests that child speech is built upon restricted linguistic contexts. One of the biggest characteristics of human memory is context-dependency. Early language experience is built upon restricted contexts and usages requires further learning to achieve the adult form. Second, child-directed speech is used in young children's living environments. Children seem to use words highly consistent with their caregivers. Third, talking to children in child-directed speech is a double-edged sword. On the one hand, children might have an easier time initializing their language capacities at an early language development stage because their hypothesis space is restricted by child-directed speech. On the other hand, the mismatch between child-directed speech and adult-directed speech requires children to shift their semantic representations at later development stages.

In Simulation 2, we show empirically that semantically moderately general verbs are better starting points for language development. Our simulations show mixed results for the "path-breaking" argument that semantically generic verbs are important for language learning (Ninio, 1999). Our results suggest that both semantic generality and semantic diversity play a role here. Although semantically general
verbs help in general, verbs that are semantically too general may not be that helpful.


Figure 3 Unlabeled accuracies of alignments with highdegree, medium-degree, low degree, mixed-degree and random initializations

## Speaker Normalization by Manifold Alignment

In speech recognition and perception, speaker normalization is the task of automatically adjusting to acoustic differences between different speakers. Our work is inspired by Plummer et al. (2010), who proposed manifold alignment as an account for how young children learn to handle phonetic variability in vowel production during language acquisition.

Aside from working with semantic, rather than acoustic, representations, our work differs from theirs in two respects. First, they used synthesized data as input, while we used naturalistic corpus data. Second, since two token pronunciations of vowels will never be the same, they imposed only a soft alignment constraint that labeled pairs be aligned, while we imposed a hard constraint.

## Crosslinguistic Alignment of Polysemous Words

Youn et al. (2016) investigated semantic universals by constructing networks of corresponding polysemous nouns from 81 languages sampled from different language families. Using an approach reminiscent of thesaurus-based synonym induction, they established semantic correspondences between nouns using bilingual dictionaries. The target polysemous words were selected from the Swadesh 200 basic vocabulary list. The procedure described in this paper is automatic and takes into consideration the matching of semantic spaces in one language, whereas Youn and colleagues manually establishes semantic correspondences for a few basic words in bilingual data.

## Conclusions

The contribution of this paper is a novel integrated framework that compares semantic spaces of children and their caregivers based on naturalistic language productions. We combined methods from three traditions, distributed semantic representations, graph theory, and manifold
alignment, into one framework for approaching the semantic structure of the lexicon. We used naturalistic language productions from CHILDES to compare the semantic spaces spanned by verbs and demonstrated that (i) that CDS is more similar to ADS than CS in terms the semantic spaces spanned by verbs and that (ii) verbs with relatively large and diverse degrees are especially useful for aligning semantic structures.

While the general computational framework proposed in this paper does not provide an account of how children might exploit this manifold-based and graph-theoretic information, it does suggest that useful information about the structure of the adult lexicon is available to children. Even though our framework is on the computational level, using Marr's terminology (1982), it is very likely that semantic manifold alignment plays a role in children's semantic development. Additionally, this framework may be of use to other fields that are interested in the semantic structure of different lexicons. For example, this approach may be useful for performing semantic comparisons between languages or across time over the course of language change, and understanding the semantic organization of bilingual lexicons.

## References

Anderson, J. R., \& Bower, G. H. (1973). Human associative memory. Psychology press.
Belkin, M., Niyogi, P., \& Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of machine learning research, 7(Nov), 2399-2434.
Davies, Mark. (2008-) The Corpus of Contemporary American English: 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.
Diaz, F., \& Metzler, D. (2007, January). Pseudo-Aligned Multilingual Corpora. In IJCAI (pp. 2727-2732).
Duchi, J., Hazan, E., \& Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul), 2121-2159.
Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48(1), 71-99.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... \& Stiles, J. (1994). Variability in early communicative development. Monographs of the society for research in child development, i-185.
Goldberg, A. E., Casenhiser, D. M., \& Sethuraman, N. (2004). Learning argument structure generalizations. Cognitive linguistics, 15(3), 289-316.
Goldfield, B. A., \& Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. Journal of child language, 17(01), 171-183.
Golinkoff, R. M., \& Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: Implications for language acquisition. Journal of Child Language, 22(3), 703-726.

Ham, J., Lee, D. D., \& Saul, L. K. (2005, January). Semisupervised alignment of manifolds. In AISTATS (pp. 120-127).
Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211.
Levy, O., \& Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems (pp. 2177-2185).
MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.
Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. The MIT Press.
Ninio, A. (1999). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. Journal of child language, 26(03), 619-653.
Pennington, J., Socher, R., \& Manning, C. D. (2014, October). Glove: Global Vectors for Word Representation. In EMNLP (Vol. 14, pp. 1532-1543).
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and brain sciences, 27(02), 169-190.
Plummer, A. R., Beckman, M. E., Belkin, M., Fosler-Lussier, E., \& Munson, B. (2010). Learning speaker normalization using semisupervised manifold alignment. In INTERSPEECH (pp. 2918-2921).
Thiessen, E. D., Hill, E. A., \& Saffran, J. R. (2005). Infantdirected speech facilitates word segmentation. Infancy, 7(1), 53-71.
Shi, J., \& Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), 888-905.
Snow, C. (1995). Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes. The handbook of child language, ed. by Paul Fletcher and Brian MacWhinney. Oxford: Blackwell.
Steyvers, M., \& Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science, 29(1), 4178.

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., ... \& Bhattacharya, T. (2016). On the universal structure of human lexical semantics. Proceedings of the National Academy of Sciences, 113(7), 1766-1771.

# ¿From Abstract to Concrete? <br> Evidence for designing learning platforms that adapt to user's proficiencies. 

Michael I. Swart (MIS2125@tc.columbia.edu) Sorachai Kornkasem<br>Nirmaliz Colon-Acosta Amy Hachigan John B. Black<br>Department of Human Development, Teachers College Columbia University, $525 \mathrm{~W} .120^{\text {th }} \mathrm{St}$. New York, NY 10027 USA<br>Jon M. Vitale<br>University of California Berkeley, Graduate School of Education 1607 Tolman Hall<br>Berkeley, CA 94720 USA


#### Abstract

Digital-tablets distribute cognition through visual, auditory and haptic interactivity. We designed a tutor-game that explored how narratives ((S)trong/(W)eak) and gestures ((I)conic/(D)eictic) could be combined to situate embodied learning. Students played seven levels of a fractions game designed to teach them how to create and compare fractions. One hundred thirty-one students ( $\mathrm{N}=131$, age $\overline{\mathrm{x}}=8.78$ yrs, $52.6 \%$ Female) were randomly assigned to one of four groups (SI, SD, WI, WD) in a $2 \times 2$ factorial experiment. Students completed pre/post direct and transfer assessments and tutorgame $\log$ data was mined to explore characteristics of students learning. Results revealed a significant interaction between narrative and gesture moderated by student proficiency. In effect, students new to fractions performed better in an abstract environment using deictic (pointing) gestures. However, as students' proficiencies improved, they learned better using iconically enactive gestures in strong narrative with setting, characters and a plot. This has important implications for designing adaptive learning platforms and curricula for teaching fractions.


Keywords: embodied, situated, grounded cognition; narrative, gestures; design-based research; DBR; data-mining; adaptive learning.

## Introduction

Tutor-games provide learners with dynamic experiences that channel their visual (sight), aural (sound) and haptic (touch) perceptions into their cognitions (Baddeley, 1986; Ricker, AuBuschon \& Cowan, 2010). As virtual portals, digital tablets allow educators to situate learning in various contexts that scaffold the processes that connect concepts (Barab et al., 2007; Brown, Collins \& Duguid, 1989; Saxe, 1988; Lave, 1988; Schwartz \& Bransford, 1999). The touch-based gestural interface of digital tablets accesses the haptic channel as a means for embodying concepts (Varela, Thompson, \& Rosch, 1990; Barsalou, 1999; Glenberg \& Kaschak, 2002; Lakoff \& Johnson, 1980). The multi-modal ecology of digital tablets allows researchers to scaffold experiences that afford (Gibson, 1979) students freedom to explore with feedback that guides their learning (Dewey, 1938/1963).

## Theoretical Background

Developing Narrative. Developing an effective narrative invests the audience in the continuity of the characters,
locations, objects, actions and themes and invests them into the plot's trajectory (Graesser, Singer \& Trabasso, 1994). These details (microstructure) are the access points to a larger interactive narrative (macrostructure) that situates the concepts (van Dijk \& Kintsch, 1983). Thus, designers must create assets that engage players in problem spaces through the processes that foster correct mental model constructions (Johnson-Laird, 1980). Black and Bower (1980) found that the structure of stories, with actors, settings, problems and solutions, aided in participants inference making and recall. In effect, the coherence of narrative schemas helps participants chunk details into mental models (Black, Turner \& Bower, 1979) and ideally, the audiences' investment in the narrative can motivate player's explorations of the processes for creating and comparing fractions in a problem space conducive for discovery (Brown, Collins \& Duguid, 1989).

Developing Gestural Mechanics. Goldin-Meadow, Cook and Mitchell (2009) demonstrated that a pairing gesture (i.e., two fingers to identify two numbers as a pairing) facilitated elementary students strategies for arithmetic problems and demonstrates how gestures as abstractions are still rooted in relation to the body. In the cognitive science literature, gestures have been typically defined as spontaneous co-articulations with speech (Kendon, 1972; McNeill, 1992), but in the digital age, the physicality of gestures has been co-opted into gestural mechanics as an interface with touch and motion based digital technologies. Educators can leverage the mechanics of gestures as communications of concepts and strategies by simulating perceptual states to activate learners' understandings (Goldin-Meadow, 1999).

Exploratory studies (Swart et al., 2014) revealed types of gestures learners used when explaining fractions. Echoing Hostetter's and Alibali's (2008) Gestures as Simulated Action, students majoritively used either iconic gestures (I) (metaphorical, enactive, symbolic) that enact their understandings or deictic (D) gestures (pointing) that identify them (Fig. 1).


Fig. 1: Iconic \& Deictic Gestures

## The Tutor-Game: Mobile Movement Mathematics (M3).

The human ability to think mathematically manifests from the endowments of our perceptual systems. It includes our abilities to estimate the magnitudes of spaces and durations of time as well as enumerate objects by differentiating the intensities of stimuli in our surroundings (Dehaene, 1997). These experiences ground the embodied metaphors of mathematical thinking (Lakoff \& Núñez, 2001; see Fig 2) and we recognize that fractions originate in the processes of fracturing wholes into parts. Thus, we chose to use object fracturing as the metaphor for developing a situatively embodied curriculum.


Fig. 2: Embodied Experiences of Mathematical Fractions
The tutor-game consisted of 7 levels of 5 fractions that were situated in either a strong $(S)$ or weak $(W)$ narrative. The strong narrative had a setting, characters and plot based on the PBS series Cyberchase, and was compared to a weak, non-descript environment without narrative elements (see Fig. 3). We characterized it as "weak" in lieu of "no" narrative to account for researchers inability to control for any internal narratives that students might devise.


Fig. 3. Strong Narrative (L) \& Weak Narrative (R)
To play, students used either iconic or deictic gestures in a 2-part tutor-game: [Part 1] Players estimated, denominated, numerated and re-estimated using the fractivator (a hybrid of a rectangular area model and a number line (Siegler \& Opfer, 2003); [Part 2] Players determined equivalency between fractions by ordering them, magnifying their height and delineating each onto a vertical number line (Fig. 4).


Fig. 4. Part 1 (Obj. Fracturing) \& Part 2: Obj. Equivalency

## The Experiment

In order to isolate for the impact of gesture (I vs. D) and narrative (S vs. W) on learning, we devised, designed and developed 4 versions of the digital tablet tutor-game (M3) that resulted in the following experimental conditions: $S I$, $S D, W I$, and $W D$, and all other factors (curriculum, assets, instructions, feedback, and scaffolding) were held constant.

Under the gesture hypothesis (embodied), iconic gestures with richer perceptual affordances (Black, Segal, Vitale and Fadjo, 2012) should help learners embody mathematical concepts better than deictic gestures. We predicted that iconic gestures, by grounding concepts in real-world actions, connect internal processes of our cognition and affect better than deictic gestures.

For the narrative hypothesis (situated), contextualizing problem spaces (via setting, characters and plot) helps learners engage in the construction of their own conceptual models. By situating learning, we predicted that the strong (S) narrative will produce higher levels of engagement and motivation and higher levels of learning compared to a weak (W) narrative.

The third hypothesis arises from the interplay of design and how independent factors will interact. The interaction hypothesis suggests that combinations of narrative types (S vs. W) and gesture types (I vs. D) will create learning environments that vary in their efficiency for the learner. In favor of the situated and embodied condition, we predicted that the SI condition would perform better than SD and or WI conditions, while the WD condition would perform better than SD and or WI.

The fourth hypothesis stems from our classroom observations of students' play and the prospect for differential efficiencies between SI and WD. The proficiency hypothesis suggests that learners' existing proficiencies at fractions will moderate how they play and learn. In favor of the situated and embodied condition, we predicted that students with lower proficiencies would benefit more from the situated embodied experience of the SI condition while students with higher proficiencies would benefit from the abstractions of the WI condition.

## Methods

Participants. One hundred thirty-one participants from grades $3\left(\mathrm{~N}=131 ; \overline{\mathrm{x}}_{\text {age }}=8.78\right.$ years [1.36], $52.6 \%$ female $)$ at afterschool programs in New York City obtained parental consent to participate in the program.

Procedure. Researchers formally tested a total of 131 students in specially designated classrooms where researchers and monitors proctored over the sessions, administered assessments, collected observational and video record the sessions. In a $2 \times 2$ randomized factorial, students were assigned to play one of four game-based environments (Strong-Iconic (SI, $\mathrm{n}_{\mathrm{si}}=35$ ), Strong-Deictic (SD, $\mathrm{n}_{\mathrm{sd}}=27$ ), Weak-Iconic (WI, $\mathrm{n}_{\mathrm{wi}}=34$ ), Weak-Diectic (WD, $\mathrm{n}_{\mathrm{wd}}=$ 35)). Each student completed 3 one-hour sessions that in total included pre-tests, game play, post-tests and exit-


Fig. 5. $2 \times 2$ Randomize Factorial w Repeated Measures
surveys (see Fig. 5). Students' sessions were run in separate groups of 10 (5/condition) with a total of 2 sessions per day (total of 20 students/day; 5 per condition) for 3 days each week, extended over multiple weeks and some students participated in an optional $4^{\text {th }}$-day clinical interview.

## Materials

## Assessments.

Direct Pre/Post Test: Parallel Forms A \& B of fraction problems directly from the game curriculum. Representations of fractions were similar to static versions of what students saw in the game, including estimation, denomination, numeration and determining equivalency between fractions ( 40 items).

General Pre/Post Test: Parallel Forms C \& D of general fraction assessment that included problems using objects, collections of objects, number lines, numerical fractions, arithmetic, and word problems. Questions included items asking students to estimate, denominate, numerate and determine equivalency between fractions ( 43 items).

## M3: Digital Tablet Tutor-Game.

Log Data: The backend of the game was designed to deliver user $\log$ data (i.e., telemetry data) to helps researchers create profiles of students' learning by tracking players' time, accuracy/error, attempts and strategies during tutor-game play.

## Equipment.

iPad Air \& Sony MDR-ZX100 Headphones: A class set of 10 each; Flip Video UltraHD Camcorder: 2 camcorders w/ Tripods for Video.

## Results

## Formal Assessments

Direct Assessments. ANOVA revealed a significant interaction between gesture and narrative on Direct Assessment Total Difference scores (post - pre), $\mathrm{F}_{(1,126)}=$ $7.324, \mathrm{p}<.008, \mathrm{~d}=.482,(1-\beta)=.766$ (Figure 59). The significance of this interaction supports the both the narrative and gesture hypotheses that each can impact learning. Since the interaction is significant, the main effects of gesture or narrative are unclear. However, Fig. 6 clearly depicts the interaction and illustrates how students in the SI and WD groups show significantly higher rates of learning across amongst all the M3 groups.

T-tests for independence revealed differences between conditions for Direct Assessment Total Difference scores, with students in the SI group ( $\overline{\mathrm{x}}_{\text {pre }}=.208, \mathrm{SD}=0.143$ ) scoring higher than students in the SD group ( $\overline{\mathrm{x}}_{\mathrm{D}}=.143$, SD


Fig. 6. ANOVA revealed a significant interaction between gesture and narrative on Direct Assess total scores.
$=0.138), \mathrm{t}_{(60)}=1.79, \mathrm{p}<.079, \mathrm{~d}=.451$ and significantly higher than students in the WI group ( $\overline{\mathrm{x}}_{\mathrm{D}}=.129$, $\mathrm{SD}=$ $0.147), \mathrm{t}_{(67)}=2.25, \mathrm{p}<.028, \mathrm{~d}=.526$ ) while the WD group $\left(\overline{\mathrm{x}}_{\mathrm{D}}=.215, \mathrm{SD}=0.194\right)$ scored higher than $\mathrm{SD}, \mathrm{t}_{(60)}=1.79$, $\mathrm{p}<.107$ and significantly higher than WI, $\mathrm{t}_{(67)}=2.069, \mathrm{p}<$ $.041, \mathrm{~d}=.486$.

Preliminarily, this suggests that the strong narrative combined with iconic gestures as well as the deictic gestures combined with weak narrative both provide a learning experience significantly more efficient than either the strong-deictic or weak-iconic pairings.

Transfer Assessment. ANOVA revealed no significant main effects of gesture or interaction between gesture and narrative for Transfer Assessment Numeration Difference scores $F_{(1,128)}=1.70, p<.195, d=.229,(1-\beta)=.254$. Though $t$-tests for independence of the difference scores (post - pre) were not significant between groups, the pattern

DIFFERENCE SCORES: TRANSFER ASSESSMENT TOTAL SCORES


Fig. 7. ANOVA revealed a significant interaction between gesture and narrative on Direct Assess total score.s
of results in Fig. 7 show that students in the SI group ( $\overline{\mathrm{x}}_{\text {pre }}=$ $.147, \mathrm{SD}=.192)$ scored higher than students in $S D\left(\overline{\mathrm{x}}_{\mathrm{D}}=\right.$ $.084, \mathrm{SD}=0.180), \mathrm{t}_{(60)}=1.296, p<.20, d=.330$, higher than students in $W I\left(\overline{\mathrm{x}}_{\mathrm{D}}=0.07, \mathrm{SD}=0.246\right), \mathrm{t}_{(67)}=1.272, p$ $<.305, d=.394$, and higher than students in the WD condition ( $\overline{\mathrm{x}}_{\mathrm{D}}=.061, \mathrm{SD}=0.194$ ), $\mathrm{t}_{(68)}=1.857, p<.068, d$ $=.443$.

A one-way contrast showed that the $S I$ group performed
significantly better than the other three groups t (127) $=$ 1.763, $\mathrm{SE}=.122, \mathrm{p}<.080$. Unlike the direct assessment interaction, results from the transfer assessment suggested that the situated and embodied condition (SI) contributed to better transfer. Simply, enacting the processes of fracturing objects while situated in a narratively contextualized problem space seems to contribute to better transfer.


Fig. 8. HLR model regressing PreTest, Telemetry Data and Condition on PostTest scores. Direct Effect of X on Y ; Indirect Effect of X on Y via MEi $=\left(a_{i}\right)\left(b_{i}\right) ; \mathrm{COV}$ on $\mathrm{Y}=c_{i}$

## Tutor-game Log Data

Mediation with a Covariate Models. The next series of analyses looked principally at how condition and tutor-game play account for the variance in students' post-test scores while controlling for pre-test scores. Fig. 8 depicts the conceptual path model used for the stepwise construction of the Hierarchical Linear Regressions (HLR) predicting the variance in the assessment scores.

The path model depicts how the variance in dependent variable ( Y , post-test assessment score) is accounted for by the independent variable ( X , condition - SI, SD, WI, WD), while controlling for a covariate (COV, pre-test assessment score) and mediated by students' tutor-game play (ME, telemetry data).

Direct Assessment Total Post-Test. The first HLR regresses condition, pre-test scores and tutor-game play on direct assessment total scores. The complete meditational covariate model significantly predicted the outcome of students Direct Assessment Post-Test scores $R=.645, \mathrm{~F}_{(7 \text {, }}$ 4577) $=543.80, p<.001$. With the covariance of pre-test controlled, tutor-game play predicted a significant amount of the variance in Direct Post-Test Assessment scores ( $\mathrm{B}=$ $.623, \mathrm{SEB}=0.012, \beta=.607, p<.001,95 \% \mathrm{CI}[.599, .646])$.

Transfer Assessment: Total Score. The complete model significantly predicted the outcome of students Transfer Assessment Total Post-Test scores $R=.632, \mathrm{~F}_{(8,4576)}=$ $379.80, p<.001$. With the covariance of pre-test controlled, tutor-game play predicted a significant amount of the variance in direct post-test assessment scores $R=.626$, $\mathrm{F}_{(3,4580)}=35.47, \mathrm{p}<.001$..

Moderated Mediation Models. With solid evidence that both the SI and WD conditions were efficient environments for learning, it was important to clarify the nature of the
interaction between narrative and gesture and determine if the situated embodied approach (SI) was better for low proficiency students (i.e., early learning is situationally embodied) or those with higher proficiencies. The second path model determines if students' initial proficiencies $\left(\mathrm{MO}_{\mathrm{i}}\right.$, pre-test score) moderated how students played $\left(\mathrm{ME}_{\mathrm{i}}\right.$, telemetry data) and improved on formal assessments $\left(\mathrm{Y}_{\mathrm{i}}\right)$.


Fig. 9. HLR model of PreTest, Telemetry Data and Condition on PostTest scores. Direct Effect of X on Y; Indirect Effect of X on Y via $M E_{i}=\left(a_{i}\right)\left(b_{i}\right) ; \mathrm{MO}_{\mathrm{i}}$ on $\mathrm{X} \rightarrow \mathrm{Y}=c_{i}$ and $M E_{i} \rightarrow \mathrm{Y}$

In Fig. 10, we can see that there are two distinct slopes for the $S I$ ( $R^{2}=.474$ ) and $W D \quad\left(R^{2}=.183\right)$ conditions,
indicating two distinct trajectories of improvement from pretest (xaxis) to post-test (y-axis) scores. The dashed red boxes indicate the


Fig. 10. Scatterplot of Pre-Test (X axis) and Post-Test (Y Axis) scores by groups (SI; WD). median split between low and high initial proficiencies. Visual inspection suggests that the WD group shows better learning when their initial proficiencies are lower while the SI group seems to show better learning when their initial proficiencies are higher.

The moderated meditational model of the proficiency hypothesis confirmed that student performances in the game on formal assessments were significantly moderated by their existing proficiencies with fractions. Fig. 11 (top) shows the moderated mediation of direct assessment scores by condition and proficiency $R=.630, \mathrm{MSE}=122.36, \mathrm{~F}_{(5,2444)}$ $=353.72$, $p<.0001$. Students with lowest proficiencies $\left(10^{\text {th }}\right.$ percentile $\left(\mathrm{x}_{\text {pre }}=11.50 ; \mathrm{B}=-9.32, \mathrm{SE}_{\mathrm{B}}=.832, \mathrm{t}_{(2443)}=-\right.$ $11.20, p<.0001,95 \%$ CI [-10.95, -7.68]), benefitted the most if they were in the WD condition $(\beta<0)$ condition compared to the SI $(\beta>0)$, but as proficiency improved, students began to benefit more in the SI condition ( $90^{\text {th }}$ percentile ( $\mathrm{x}_{\mathrm{pre}}=46.00 ; \mathrm{B}=5.29, \mathrm{SE}_{\mathrm{B}}=.645, \mathrm{t}_{(2443)}=8.21, p$ $<.0001,95 \%$ CI $[4.03,6.56])$. We see a similar transition for low to high proficiencies from WD to SI for the transfer assessment (see Fig. 11, bottom). In this case, the
transition from the WD to the SI condition takes place at lower initial proficiencies for transfer of learning.


Fig. 11a \& 11b. Moderated Mediation of formative assessment scores by the interaction between condition and existing fractions proficiency. Scores on the pre-test are stratified by percentiles along the x -axis $\left(10^{\text {th }}\right.$, $\left.25^{\text {th }}, 50^{\text {th }}, 75^{\text {th }}, 90^{\text {th }} \%\right)$, and values on the $y$-axis are the weights of the $B$ coefficients for changes in Direct Assessment Post-Test scores. Coefficient values below the zero line on the $y$-axis indicate that the WD improved more on post-test at that percentile and coefficient values above the zero line indicate that students in the SI group improved more.

## Discussion

The Gesture, Narrative \& Interactions Hypotheses. The significant interaction between gesture and narrative on the direct assessment of the M3 curriculum shows that types of gestures may be conceptualized differently depending on the contexts in which they are embedded. It calls into question our original theoretical assumptions that situating cognition through narrative and embodying procedural learning through iconic gestures would produce better learning.

The HLRs on students direct and transfer assessment total scores showed that students tutor-game play, including their accuracy denominating, numerating and estimating significantly predicted learning, supporting the position that the act of splitting objects is central to learning fractions (Steffe, 2004; Norton \& Wilkins, 2009). Improvement on transfer assessment seems to suggest that the procedural and conceptual knowledge that players are developing is robust enough that the curriculum prepared
them for future learning (Schwartz \& Bransford, 1998) of near transfer representations and new domains for fraction.

The Efficiency Principle. Although our initial hypotheses predicted the superlative performances by the SI conditions for both assessments, the significant interaction between gesture and narrative suggests that both the SI and WD conditions are both efficient platforms for learning. Schwartz, Bransford and Sears (2005) note that efficiency often means rapid retrieval with accurate appropriation and application of knowledge and skills for understanding, solving and explaining a problem. Though the situated embodied SI environment provided a perceptually rich experience (Black et al., 2012) that promoted better transfer, students using deictic gestures in the weak narrative (i.e., without seductive details, Harp \& Mayer, 1998; Adams et al., 2012) also showed significantly better learning. Might the minimal and abstracted environment of the WD condition make procedures and concepts easily salient?

The Proficiency Principle. Students with low initial proficiencies benefitted more from playing in the WD version of the game, while students with higher initial proficiencies benefitted more in the SI environment. This finding was contrary to our hypothesis and the principle of concreteness fading (i.e., start concrete and fade to abstract; Fyfe, McNeil, Son \& Goldstone, 2014). Still to be determined is how these results fit with The Expertise Reversal Effect (i.e., experts require reduced guidance; Sweller, Ayres, Kalyuga, \& Chandler, 2003). Does the presence of the strong narrative make instruction and guidance invasive (i.e., reduced)? Nonetheless, the current results support findings from a study by Kaminski, Sloutsky and Heckler $(2006 ; 2008)$ that found that students learned division with remainders better using abstract symbols rather than concrete real world depictions.

## Significance

The current research demonstrated that combinations of different narratives and gestures produced differential learning. Ribbons and Malliet (2010) advocate for simulational realism in gaming. They argue that there must be balance between the rules that govern gaming experiences (e.g., gestures) and their relevance to the situated environment (e.g., the interactive narrative). This research suggests that when educators are designing pedagogy and curricula for mathematical fractions, students should begin working with abstractions and as their proficiency improves the learning platform should adapt to concrete experiences.

## Acknowledgments

Supported by NSF Cyberlearning Grant 1217093. Thank you: Sandra Sheppard and Kristin DiQuallo at WNET-13; Jan \& Nic at Curious Media.

## References

Adams, D. M., Mayer, R. E., MacNamara, A, Koenig, A, \& Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. Journal of Educational Psychology, 104(1) 235-249.
Baddeley, A. (1986). Working Memory. New York: Oxford University Press.
Barab, S., Zuiker, S., Warren, S., Hickey, D., Ingram-Goble, A., Eun-Ju Kwon, E-J., Kouper, I., \& Herring, S. (2007). Situationally embodied curriculum: Relating formalisms and contexts. Science Education, 91(5), 750-782.
Barsalou, L.W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 557-660.
Black, J.B. \& Bower, G.H. (1980) Story understanding as problem-solving. Poetics, 9, 223-250
Black, J. B., Turner, T. J., \& Bower, G. H. (1979). Point of view in narrative comprehension, memory, and production. Journal of Verbal Learning and Verbal Behavior, 18(2), 187-198.
Black, J.B., Segal, A., Vitale, J. and Fadjo, C.L. (2012). Embodied Cognition and learning environment design. In D. Jonassen and S. Lamb (Eds.), Theoretical Foundations of StudentCentered Learning Environments. New York: Routledge.
Bransford, J. D., \& Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad \& P. D. Pearson (Eds.), Review of research in education (Vol. 24, pp. 61-100). Washington, DC: AERA.
Brown, J.S. Collins, A. \& Duguid, P. (1989). Situated Cognition and the Culture of Learning. Ed. Researcher, 18, 33-42.
Dehaene, S. (1997/2011). The number sense: How the mind creates mathematics. New York: Oxford University Press.
Dewey, J. (1938/1963). Experience and Education. New York: Collier Books.
Gibson, J.J. (1979) The ecological approach to visual perception. Boston, MA: Houghton Mifflin.
Glenberg, A.M. \& Kaschak, M.P. (2002) Grounding Language in Action. Psychononmic Bulletin \& Review. 9(3), 558-565.
Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. Trends in Cognitive Science, 3, 419-429.
Goldin-Meadow, S., Cook, S.W., and Mitchell, Z.A. (2009). Gesturing Gives Children New Ideas About Math, Psychological Science, 20(3), p. 267-272.
Goldstone, R. L., \& Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. The Journal of the Learning Sciences, 14, 69-110.
Graesser, A. C., Singer, M., \& Trabasso, T. (1994). Constructing inferences during narrative text comprehension. Psychological Review, 101(3), 371-395.
Harp, S. F. \& Mayer, R.E. (1998). A theory of cognitive interest in science learning. Journal of Educational Psychology, 90, (3), 414-434.
Hostetter, A. B. \& Alibali, M. W. (2008). Visible embodiment: Gestures as simulatedaction. Psychonomic Bulletin and Review, 15, 495-514.
Johnson-Laird, P.N. (1983). Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge: Cambridge University Press.
Kaminski, J., Sloutsky, V. M., \& Heckler, A. F. (2006). Do children need concrete instantiations to learn an abstract concept? Proceedings of the $28^{\text {th }}$ Annual Conference of the Cognitive Science Society (pp. 411-416)
Kaminski, J. A., Sloutsky, V. M., \& Heckler, A. F. (2008). The advantage of abstract examples in learning math. Science, 230, 454-455.

Kendon, A. (1972). Some relationships between body motion and speech. In A. Siegman \& B. Pope (Eds.), Studies in dyadic communication (pp. 177-210). New York: Pergamon Press.
Lakoff, G., \& Johnson, M. (1980). The metaphorical structure of the human conceptual system. Cognitive Science, 4(2), 195208.

Lakoff, G., \& Núñez, R. E. (2000). Where mathematics comes from: How the embodied mind brings mathematics into being. Basic books.
Lave, J. (1988). Cognition in Practice: Mind, Mathematics and Culture in Everyday Life (Learning in Doing). Cambridge: Cambridge University Press.
McNeill, D. (1992). Hand and Mind: What Gestures Reveal About Thought. Chicago:Chicago University Press.
Norton, A., \& Wilkins, J.L.M. (2009). A quantitative analysis of children's splitting operations and fraction schemes. Journal of Mathematical Behavior 28, 150-161.
Ricker, T.J., AuBuchon., A.M. \& Cowan, N. (2010). Working Memory. Wiley Interdisciplinary Reviews: Cognitive Science, 1(4), 573-585.
Ribbens, W., \& Malliet, S. (2010). Perceived digital game realism: A quantitative exploration of its structure. Presence: Teleoperators and Virtual Environments, 19(6), 585-600.
Saxe, G. (1988). The Mathematics of Street Vendors. Child Development, 59, 1415-1425.
Schwartz, D. L. \& Bransford, J.D. (1998). A time for telling. Cognition and Instruction, 16(4), 475-522.
Schwartz, D. L., Bransford, J. D., \& Sears, D. (2005). Efficiency and innovation in transfer. Transfer of learning from a modern multidisciplinary perspective, 1-51.
Segal, A., Tversky, B. and Black, J.B. (2014). Conceptually congruent actions can promote thought. Journal of Applied Research in Memory and Cognition, http://dx.doi.org/10.1016/j.jarmac.2014.06.004
Siegler, R.S. \& Ramani, G.B. (2008). Playing linear numerical board games promotes low-income children's numerical development Developmental Science, 11(5) 655-661.
Steffe, L. P. (2004). On the construction of learning trajectories of children: The case of commensurate fractions. Mathematical Thinking and Learning, 6(2), 129-162.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12 (1988), pp. 257-285.
Swart, M. I, Friedman, B., Kornkasem, S., Hollenburg, S., Lowes, S., Black, J.B., Vitale, J.M., Sheppard, S., \& Nankin, F. (2014). Mobile Movement Mathematics: Exploring the gestures students make while explaining Fractions. Presented at 2014 AERA National Conference, Philadelphia, PA.
Swart, M.I., Friedman, B., Kornkasem, S., Lee, A., Lyashevsky, I., Vitale, J.M., Sheppard, S., Black, J.B., (2016). A DesignBased approach to Situating Embodied learning of Mathematical fractions using Narratives and Gestures in a tablet-based game. 2016 AERA National Conference, Washington, DC.
Varela, F., Thompson, E., \& Rosch, E. (1991). The embodied mind: Cognitive science and human experience. Cambridge, MA: MIT Press.
Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin \& Review, 9, 625-636.
van Dijk, T.A. \& Kintsch, W. (1983) Strategies of Discourse Comprehension. New York: Academic Press.

# Supporting Low-Performing Students by Manipulating Self-efficacy in Digital Tutees 

Betty Tärning (betty.tarning @lucs.lu.se)<br>Lund University Cognitive Science, Lund University, 22222 Lund, Sweden<br>Magnus Haake (magnus.haake@lucs.lu.se)<br>Lund University Cognitive Science, Lund University, 22222 Lund, Sweden<br>Agneta Gulz (agneta.gulz@lucs.lu.se)<br>Lund University Cognitive Science, Lund University, 22222 Lund, Sweden


#### Abstract

Educational software based on teachable agents has repeatedly proven to have positive effects on students' learning outcomes. The strongest effects have been shown for low-performers. A number of mechanisms have been proposed to explore this outcome, in particular mechanisms that involve attributions of social agency to teachable agents. Our study examined whether an expression of high versus low self-efficacy in a teachable agent would affect lowperforming students with respect to their learning outcomes and with respect to a potential change in their own selfefficacy. The learning domain was mathematics, specifically the base-ten system. Results were that the learning outcomes of low-performers who taught a low self-efficacy agent were significantly better than the learning outcomes of lowperformers who taught a high self-efficacy agent. There were no effects from the manipulation of self-efficacy expressed by the teachable agent on changes of the low-performing students' own self-efficacy.


Keywords: social agency; educational software; teachable agent; math self-efficacy; math performance

## Introduction

A teachable agent (TA) is a graphical computer character in a tutee role. The basic idea is that the student instructs and guides the TA (Brophy, Biswas, Katzlberger, Bransford, \& Schwartz, 1999). In essence, TA-based educational software implements the pedagogical approach learning by teaching, (Bargh \& Schul, 1980).

To date a set of TA-based learning games targeting the STEM areas have been developed and evaluated, and repeatedly proven to have positive effects on students' learning outcomes. Some studies have compared effects of TA-based software with ordinary teaching (regular classroom practice) (Pareto, Haake, Lindström, Sjödén, \& Gulz, 2012; Chin, Dohmen, \& Schwartz, 2013). Others have compared educational software versions with and without a teachable agent included (Chase, Chin, Oppezzo, \& Schwartz, 2009; Pareto, Schwartz, \& Svensson, 2009).

An observation from several of the studies is how readily the metaphor of the computer figure as a tutee (digital tutee) is accepted by students. They express engagement for the task of teaching the character, although it is in fact nothing but a computer artifact (Chase et al., 2009; Lindström, Gulz, Haake, \& Sjödén, 2011.) They also make more effort to
learn in order to teach their digital tutee than to learn for themselves (Chase et al., 2009). In effect, students attribute mental states and responsibility to the digital tutee as if it were a social agent (Chase et al., 2009; Lindström et al., 2011). They see the agent as a socio-cognitive actor that can learn (respond to being taught by them) and that can be ascribed traits such as 'brave', 'slow', 'smart', 'forgetful' etc.

## TA-systems and Low-Performing Students

Several studies show that the students who benefit most by educational software with teachable agents - whether compared to equivalent software without TA or compared to ordinary classroom teaching - are the low-performing students. When comparing eleven year olds who used an educational game in biology with or without TA, the former spent more time on learning activities and also learned more, with the effects most pronounced for lower performing students (Chase et al., 2009). In a study by Sjödén and Gulz (2015), 9-10 year-olds used a TA-based educational math game in school over a period of eight weeks. Thereafter, the students were divided into two groups, matched according to their pretest scores, and randomly assigned to a post-test with or without the TA present (the TA did not act in order to influence the test but was merely present). Results showed that low-performers (according to the pretest) improved significantly more than high-performers but only when tested with the TA. Pareto et al. (2009), likewise found a considerably stronger improvement for low ability students than for high ability students when they used a math game with a TA feature compared to the math game without the TA.

## Mechanisms in TA-Systems that may Support Low-Performing Students

A number of explanations for the pedagogical power of TAbased games have been proposed, including some that also provide possible rationales for why the effect is often larger for low-performers.

First, in a TA-based game, the student is positioned as the one that is most able, the one who can teach someone else that knows less. This experience - being someone who is
capable, who knows more than someone else - can potentially affect a student's view on her own competence in a positive way. This will likely benefit low-performers more than high-performers, since the latter are more likely to already have experienced the role of 'teaching someone else' and 'knowing more'. High-performers are more likely than low-performers to spontaneously take a teacher role (or be assigned this role in class). Acting teacher can potentially strengthen the student's belief in her own capability in the domain in question, and this may in turn have effects on performance.

Second, a teachable agent can be a model of learning behaviors (Blair, Schwartz, Biswas, \& Leelawong, 2007). A TA is often designed to model fruitful and productive student behaviors, such as being curious, asking questions, reasoning, being explicit about parts of 'knowledge'. It is, however, more likely that high-performing students already have such behaviors on their repertoire compared to lowperforming students, and that the latter therefore are more helped by being inspired by productive learning behavior in a TA.

Thirdly and crucially a TA is teachable. More specifically a TA models someone who from the beginning has little or no knowledge but learns incrementally or step-by-step. In other words, a teachable agent (re)presents or models an incrementalist theory of competence in contrast to an entity theory of competence according to which some individuals are held to be gifted and others non-gifted. This latter view is quite common among students (Dweck, 2006). Specifically it holds in the domain of mathematics, where it has also been shown that teachers to a larger extent than for other subjects used terms such as 'talented' and 'not talented' (Rattan, Good, \& Dweck, 2012). In principle both high- and low-performing students can have an entity view of competence, and potentially benefit from viewing competence (in this study competence in math) as something that can be changed with effort. However, it is more likely that low-achievers with an entity view of competence are trapped in a circle, where they don't think they are talented and see no meaning in making an effort; therefore make little effort; therefore don't achieve and thus confirm they are not talented. In other words, they create a self-fulfilling prophecy.

Fourthly, Chase et al. (2009), propose a mechanism named ego-protective buffer. In TA-system it is the TA that is tested for its knowledge. When the TA fails at a test, the failure or non-success does not come as close onto the student as when she takes a test herself. Even if students are aware that the TA's knowledge reflects how the TA has been taught by themselves, the responsibility for failing is not only theirs. Instead of bearing the full burden of a failure, the responsibility of failure can be shared between the TA and student. Even though this may benefit highperformers as well, low-performers are more used at failing at school and thus the ego-protective buffer mechanism may
explain why in particular low-achieving students perform better when working with a TA.

In sum, there is a set of proposed mechanisms that may explain why low-performers benefit more than highperformers from using teachable agents. All mechanisms involve the tendency of students to attribute social characteristics and agency to the agent, and interact intellectually and socially with it. For instance, to view the TA as someone that it is possible to share a failure with; to view the TA as someone who can accomplish a task (or not), as someone whose knowledge is different from mine and that I can influence by teaching it; to view that TA as someone that can learn - and as learner be slow, quick, smart, forgetful, etc.

In view of the above, we found it plausible that students would also tend to attribute high or low self-efficacy to an agent, if designed in an adequate manner. Spelled out, they would tend to attribute to an agent high or low belief in its own capability to learn and be successful - in our case with respect to math and base ten problems. The present study thus approaches the trait of self-efficacy, which to our knowledge has not been studied before in teachable agents.

## Does TA Self-Efficacy Matter for Student Progress

Having an ability to learn, i.e. being teachable, is the very essence of a digital tutee or teachable agent. However, whether other kinds of properties are attributed to a TA depends in the first place on how the TA is designed and implemented, and also on the student interacting with the TA. For instance, depending on how it is implemented, a TA can be (perceived as) a quick learner or a learner that needs many rehearsals. A TA can be (perceived as) more or less challenging or questioning (Kirkegaard, 2016).

In our study the TA was designed to express either high or low belief in its own capacity to learn and perform in a math game. We will soon present our predictions but first discuss the phenomenon of self-efficacy in real human students. For human learners we know that there is a relation between self-efficacy and actual performance (Bandura, 1997) in that self-efficacy predicts subsequent performance. Low self-efficacy predicts low performance, and high self-efficacy predicts high performance. Proposed mechanisms are that student's self-efficacy influences how much effort she puts into a task, her tendency to persist, how high she sets her aspirations and her tendency to persevere when being challenged by the task. Individuals with high self-efficacy often achieve more in intellectual terms (Bandura, 1997). Importantly, however, the relations are correlational and on a group level. There are no causal or absolute relations between individual's self-efficacy and her performance; students may over-estimate as well as underestimate their own capacity.

We now return to self-efficacy in teachable agents. The central research question in the present study was whether a teachable agent expressing low or high self-efficacy,
respectively, would have different impact on lowperforming students in terms of their learning and progress. In addition we explored whether there would be any effects on students' own self-efficacy in either of the conditions.

## Research Questions and Predictions

Research Question 1 (RQ1) Will learning and progress differ between low-performing students who teach a TA expressing low self-efficacy (lowSE-TA) and lowperforming students who teach a TA expressing high selfefficacy (highSE-TA)?

As a basis for our predictions we used two different theories: (i) role-modeling theory by Bandura (1977) and (ii) the theory of the TA protégée effect by Chase et al. (2009). This resulted in two alternative predictions that point in opposite directions. As such this is not surprising since the predictions are generated from theories not related to one another.

The first, alternative, prediction in line with Bandura's idea of role modeling focuses on teachable agents as behavioral models, as discussed in the introduction. A highSE-TA models a learner with a strong belief in her own abilities to learn, a willingness to persist and not give up, etc. Together with the TA:s incremental progression (given that it is reasonably taught by the student) this is likely to be a positive model for low-performers, that often themselves have low self-efficacy. Thus we predict that low-performers will make larger progress if they teach a highSE-TA than if they teach a lowSE-TA.

The second, alternative, prediction is based on the protégée-effect mentioned above: in general, students seem to take responsibility for a TA and make an effort to teach it. Now, a lowSE-TA expresses uncertainty in its own capacity, and seems in considerable need for support and engagement from the teacher (i.e. student), whereas a highSE-TA expresses confidence in its own capability to learn and manage and seems in less need for help from the teacher. Therefore low-performers may be more motivated to take responsibility and make an effort to teach a lowSETA compared to a highSE-TA. Consequently they will also themselves make more progress. Thus we predict that lowperformers will make larger progress if they teach a lowSETA than if they teach a highSE-TA.

There is also third possible result, namely that whether the TA expresses low or high self-efficacy will not matter for low-performers progress.

Research Question 2 (RQ2) Will a potential change in selfefficacy in low-performing students differ between those students who teach a TA expressing low self-efficacy and those who teach a TA expressing high self-efficacy?

If the TA functions as a behavioral model with respect to self-efficacy, low-performers are more likely to increase their own self-efficacy if they teach a highSE-TA than if
they teach a lowSE-TA. The reason is that they may be inspired to model the TA along the line "If this character, my digital tutee, believes strongly in its capability, why shouldn't its teacher, that is me, do so too?"'
From the protégée effect no straightforward prediction can be derived on potential self-efficacy change in students, depending on TA self-efficacy. As discussed under RQ1, if the protégée effect is at work, participants will put particularly large effort into teaching a lowSE-TA, since such a TA signals a greater need of help and support than a highSE-TA that signals that can learn on its own. But whether students that take more responsibility and make a larger effort to teach their TA also change their belief in their own capacity to learn is not obvious. On the one hand, an interplay between performance and self-efficacy is likely but such influences may take time.

Again there is a third possible result, namely that whether the TA expresses low or high self-efficacy does not matter with respect to low-performers potential self-efficacy change.

To sum up, the present study made use of a learning game in math including a TA, where we manipulated the TA:s expressed belief in its own capability to perform and learn math as expected in the game. Our two research questions were: RQ1: Would the manipulation of TA self-efficacy have an effect on low-performing students' progress in the game (i.e. their learning math)? RQ2: Would the manipulation of TA self-efficacy have an effect on potential change in self-efficacy in the low-performing students?

## Method

## Participants

Participants were 166 students ( 83 girls and 83 boys) aged 10-11 years from 4 schools and 9 classes in Southern Sweden from areas with relatively low socio-economic status and school performance below average. Students were randomly assigned one of the conditions: teaching a digital tutee that expressed high self-efficacy (highSE-TA) or teaching a digital tutee that expressed low self-efficacy (lowSE-TA). Out of the initial set of participants, 24 were excluded due to missing data points or low attendance. Next, out of the 142 remaining students, the 62 students who performed below the median on a math performance test were selected for further analysis. The math test was based on a representative part of the national tests in mathematics and consisted of 21 problems relating to place value. Thus, in the final data set, there were 28 students in the lowSE-TA condition and 34 in the highSE-TA condition.

## The Educational Game

The TA math game, developed by Lena Pareto (Pareto, 2014), targets basic arithmetic skills related to the place value system, where the student teaches a digital tutee
named Lo, so that Lo can compete against other students' digital tutees or against a computer actor in different digital board games. Lo's knowledge - based on the system's knowledge domain (Pareto, 2014) - develops entirely on the basis of what the student teaches her (and if taught wrong, Lo will learn wrong).

A central part of the student's teaching consists of answering questions from the digital tutee about the math content, specifically regarding place value, via multiplechoice for answering (see figure 1). The other main interaction between student and digital tutee takes place via a free text chat (Silvervarg \& Jönsson, 2011). This is also where Lo, the TA, expresses her self-efficacy (see figure 1).


Figure 1: The math game with multiple choice conversation and 'free text chat' conversation (overlay).

## Self-Efficacy in the Teachable Agent

High or low self-efficacy in or study was defined as high or low belief in ones capability to make progress and perform well in the math game. In turn, this requires making adequate moves and answering questions regarding the place value system correctly. The definition can be compared to a more general definition of self-efficacy in mathematics as the belief in ones capability to successfully learn mathematics (Bandura, 1997).

After each round of the game where Lo (the TA) has been active - observing and posing questions to the student or being guided by student - the chat conversation starts. The chat begins with Lo commenting on the previous round saying for example: "Awesome! We won! I have a good grip now of tens and hundreds and all that you teach me." (reflecting high self-efficacy), "Oh I won, did I? Nice. But I feel very uncertain about how to play well." (reflecting low self-efficacy).

The chat conversation also contains other comments and reflections from Lo on her own learning, for instance: "I'm learning the rules in the math game slowly. I'm not a very brilliant student." (reflecting low self-efficacy), "It's going to get better and better. I have so quickly learned so many
things about how to play the game." (expressing high selfefficacy), and "I am not sure I can learn these things." (expressing low self-efficacy).

The chat always ended with a sentence from Lo regarding her thoughts about the upcoming round, for example: "I have a feeling that the next round will go really well. Let's play!" (expressing high self-efficacy) or "It doesn't seem like I understand much really, but let's play another round." (expressing low self-efficacy).

Lo's utterances had previously been evaluated with regard to whether they sounded as uttered by someone who was confident, not confident, or neither nor in her ability to learn and perform. The evaluators were 22 fourth graders from a school not participating in the study. The evaluation resulted in the removal of a few sentences and slight modifications of others, resulting in a set of 136 sentences, 68 reflecting a digital tutee with high self-efficacy and 68 reflecting a digital tutee with low self-efficacy.

In addition the manipulation - low and high self-efficacy in the TA - was validated within the present study by participating students. At the end of the last study session they were asked to evaluate Lo's belief in her/his own capability to play the math game on a Likert scale. A MannWhitney test showed a significant difference ( $Z=-4.85, p<$ $.001, r=.39$ ) between the low SE-TA and the high SE-TA, confirming that the manipulation had intended effects on the perception of the TAs self-efficacy.

## Procedure

All study sessions took place in ordinary classrooms and lasted about 30 minutes. At the pre-test session, students completed a math pre-test targeting the place value system, and a pre-questionnaire targeting their self-efficacy in math with respect to the place value system. The students' math pre-test scores were used to identify the target group for this study's research questions, i.e. low-performers (in math).

Thereafter students participated during seven gameplaying sessions, once a week. At the post-session, students again filled out the questionnaire targeting their self-efficacy in math and the place value system and were debriefed about the two different types of digital tutees and the purpose of the study.

## Measurements

Performance During Game Play Students' performance while teaching the digital tutee is a reflection on how well they perform themselves. In line with this we calculated a performance score for each student on the basis of the datalogging. Through the game the digital tutee poses questions to the student that concerns the conceptual model and principles of the place value system. For instance: "How many orange square boxes are there in the 2 yellow square boxes on the game board?" and "How many red square boxes are needed to fill a yellow square box?" The tutee
posed three such questions during each game session, and the student had to choose one out of four alternative answers (one correct, two incorrect and the alternative "I don't know."). The performance score was calculated as the percentage of correct answers minus the percentage of incorrect answers. Additionally, a study by Pareto (2014) showed that in-game performance in this math game correlated with standard paper-and-pencil tests on the placevalue system.
Self-Efficacy Change To measure this we used a selfefficacy pre- and post-questionnaire based on Bandura, Barbaranelli, Caprara, and Pastorelli (1996); for this study translated into Swedish

The seven items targeted the students' self-efficacy with regard to the place value system and the question "How good are you at solving this type of task?" Item one to five regarded calculation tasks such as " $1136+346$ ", and item six and seven targeted place value concepts, such as: "Which digit has the highest place value in the number 6275 ?" All items were graded in five steps from "Not good at all" to "Very good at".

## Results

Statistical analyses were conducted in R v3.2.4 (R Core Team, 2016). Of the 142 participants with complete data, the 62 performing below the median on the pre-test in math were included in the analysis.

## Effects TA Self-Efficacy on Low-Performing Students' Performance During Game Play

An unmatched two sample $t$-test showed a significant difference $(t(60)=3.40, p=.0012$, Cohen's $d=0.87$ ) of TA self-efficacy on student performance with the students in the lowSE-TA condition ( $M=54.8, S D=13.7$ ) outperforming the students in the highSE-TA condition $(M=43.7, S D=$ 12.0).

## Effects of TA Self-Efficacy on Low-Performing Students' Self-Efficacy Change

An unmatched two sample $t$-test showed no significant difference $(t(60)=0.35, p=.73)$ of TA self-efficacy on student self-efficacy change between the students in the lowSE-TA condition $(M=1.18, S D=3.81)$ and the students in the highSE-TA condition $(M=1.53, S D=4.00)$.

## Discussion

Teaching a lowSE-TA compared to teaching a highSE-TA made the participants perform significantly better, as measured by their in-game performance scores. But the two conditions did not differ with respect to whether the participants changed their own self-efficacy. Changes were small and did not differ between the conditions.

These results contribute to our knowledge about mechanisms in a TA-based educational game with respect to
why low-performers tend to benefit more than highperformers from these games. First, we showed that a manipulation of expressed self-efficacy in a TA can influence performance for low-performers: a TA that expressed low self-efficacy was more beneficial than a TA that expressed high self-efficacy. The effect as such, regardless of direction, confirms that at least some of the pedagogical power in a TA-based game derives from attributions of social agency to TA:s, in this case attributing to the TA a weak or strong belief in its own capability. Consequently this is one of the traits that a TA designer ought to be aware of; a trait that can explain why lowperformers benefit more than high-performers from TAbased games.

With respect to student performance, we based our predictions on two different theoretical models: role modeling according to which a highSE-TA should have the most positive influence on the performance of lowperformers, and the protégée effect according to which a lowSE-TA should have the most positive influence on the low-performers performance. The latter theory was supported and can be further elaborated on by means of the results of our study. According to the protégée-effect students tend to make more effort and take more responsibility for the task of teaching a TA than for the task of learning for themselves (Chase et al., 2009). In our study the outcome was better when low-performers taught a lowSE-TA compared to a highSE-TA. It is near at hand that they made an even larger effort and took even more responsibility for a TA with low self-efficacy since this TA expresses a low trust in her own ability to learn, and likely comes across as someone who is more in need of help than a TA with high self-efficacy. A highSE-TA, on the other hand, indicates that s/he is capable to learn and perform, and is in less need of help.

The lacking effect on students self-efficacy change, depending on high or low self-efficacy in the TA, means that the role-modeling hypothesis proposed above was not supported. Students were not inspired by a highSE-TA as a model to increase their own self-efficacy. Neither did teaching a lowSE-TA lead to an increase in the students' self-efficacy. However, it did lead to an increase in their performance, and we can thus conclude that the increased performance was not caused by an increased self-efficacy, at least not as measured in our study. It should also be pointed out that an increase in self-efficacy is not always desirable, in particular not for students who overestimate their capabilities. At the same time, given the interactions between self-efficacy and performance, it is often a good thing when students with low self-efficacy in a domain gain more confidence in their abilities to make progress. What is desirable in general is that as many students as possible have an incrementalist rather than an entity view of intellectual capabilities - something that the use of TA-
based educational games may contribute to (Chase et al., 2009).

## Limitations of the Study and Future Research

The study should be seen as a first examination about how the manipulation of self-efficacy in a digital tutee can influence student performance. Some limitations should be kept in mind when interpreting the results. One is that there was no group of students who taught a digital tutee that expressed a neutral mode of self-efficacy. In future research such a condition should be included. Furthermore, rather than aiming to be conclusive, the present study opens up for associated studies. For instance, one relevant question is whether the results will replicate or not with other age groups than 10-11 year olds. Another interesting line of research could be to explore a TA with adaptive selfefficacy that reflects the rate at which it actually learns, which in turn reflects the proficiency of the student that is teaching it.

## Acknowledgments

This research was funded in part by Marcus and Amalia Wallenberg Foundation and the research environment Cognition, Communication and Learning. The authors thank all students and teachers who participated in the study.

## References

Bandura, A. (1997). Self-Efficacy: The exercise of control. New York, NY: W.H. Freeman.
Bandura, A. (1977). Social learning theory. Englewood Cliffs, NJ: Prentice-Hall.
Bandura, A., Barbaranelli, C., Caprara, G. V., \& Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. Child Development, 67(3), 12061222.

Bargh, J. A., \& Schul, Y. (1980). On the cognitive benefits of teaching. Journal of Educational Psychology, 72(5), 593-604.
Blair, K., Schwartz, D., Biswas, G., \& Leelawong, K. (2007). Pedagogical agents for learning by teaching: Teachable agents. Educational Technology, 47(1), 56-61.
Brophy, S., Biswas, G., Katzlberger, T., Bransford, J., \& Schwartz, D. (1999). Teachable agents: Combining insights from learning theory and computer science. In S.P. Lajoie \& M. Vivet (Eds.), Frontiers in Artificial Intelligence and Applications, Vol 50. Proc. of AIED 1999. Amsterdam, The Netherlands: IOS Press.

Chase, C., Chin, D., Oppezzo, M., \& Schwartz, D. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. Journal of Science Education and Technology, 18, 334-352.

Chin, D. B., Dohmen, I. M., \& Schwartz, D. L. (2013). Young children can learn scientific reasoning with teachable agents. IEEE Transactions on Learning Technologies, 6(3), 248-257.
Dweck, C. (2006). Mindset: The new psychology of success. Random House.
Kirkegaard, C. (2016). Adding challenge to a teachable agent in a virtual learning environment. Licentiate Thesis in Cognitive Science, Linköping University. Linköping, Sweden: Linköping University Electronic Press.
Lindström, P., Gulz, A., Haake, M., \& Sjödén, B. (2011). Matching and mismatching between the pedagogical design principles of a maths game and the actual practices of play. Journal of Computer Assisted Learning, 27, 90102.

Pareto, L. (2014). A teachable agent game engaging primary school children to learn arithmetic concepts and reasoning. International Journal of Artificial Intelligence in Education, 24(3), 251-283.
Pareto, L., Schwartz, D. L., \& Svensson, L. (2009, July). Learning by guiding a teachable agent to play an educational game. In V. Dimitrova, R. Mizoguchi, B. du Boulay, A. C. Graesser (Eds.), Frontiers in Artificial Intelligence and Applications, Vol 200. Proc. of AIED 2009 (pp. 662-664). Amsterdam, The Netherlands: IOS Press.
Pareto, L., Haake, M., Lindström, P., Sjödén, B., \& Gulz, A. (2012). A teachable agent based game affording collaboration and competition - Evaluating math comprehension and motivation. Educational Technology Research and Development, 60, 723-751.
Rattan, A., Good, C., \& Dweck, C. (2012). "It's ok - Not everyone can be good at math": Instructors with an entity theory comfort (and demotivate) students. Journal of Experimental Social Psychology, 48(3), 731-737.
Silvervarg, A., \& Jönsson, A. (2011). Subjective and Objective Evaluation of Conversational Agents. In Proceedings of the 7th Workshop on Knowledge and Reasoning in Practical Dialogue Systems (pp. 65-72). Barcelona, Spain.
Sjödén, B., \& Gulz, A. (2015). From Learning Companions to Testing Companions. In C. Conati, N. Heffernan, A. Mitrovic, \& M.F. Verdejo (Eds.), LNAI/LNCS: Vol. 9112. Proc. of AIED 2015 (pp. 459-469). Berlin/Heidelberg, Germany: Springer-Verlag.
R Core Team (2016). $R$ : A language and environment for statistical computing [Software]. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/.

# The impact of practice frequency on learning and retention 

Caitlin Tenison (ctenison@ andrew.cmu.edu)<br>Department of Psychology, Carnegie Mellon University

John R. Anderson (ja+@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University


#### Abstract

The current study manipulated how frequently different problems were practiced during a first day of practice, with the more frequent items being more closely spaced. Fitting the data to a skill acquisition model, we find that greater spacing between items is associated with an increased probability of transitioning to more efficient phases of performance, but with a shallower speedup within each phase. Three days after training, we find that performance is predicted not by the practice frequency during training, but rather by the phase of skill acquisition attained during training. Thus, it is type of processing achieved not the amount and spacing of practice, that determines retention. Spacing, however, promotes learning by driving changes in cognitive processing.


Keywords: Skill acquisition; Practice frequency; Spacing effect; Learning; Retention

## Introduction

A widely held observation, labeled the 'Spacing Effect', shows that increasing the time between practice opportunities improves retention. In contrast, massing practice opportunities together improves performance during training but negatively impacts retention. However, when the amount of time available for practice is limited spacing practice results in fewer total practice opportunities which may negatively impact learning (Anderson, 1982). Given the importance of the trade-off between spacing and amount of practice in understanding how humans learn, along with the clear educational benefits to improving retention, these effects have been studied extensively over the past 100 years. Explanation of both the effects of spacing and practice is critical for understanding skill acquisition. In this paper, we apply a model of skill acquisition (Tenison \& Anderson, 2016) to a spaced-practiced data set to explore how spacing and the amount practice impacts the type of cognitive processes and affects the speed at which those processes are used to complete a task.

To capture the spacing effect, theories must explain both the short-term effects of practice frequency on skill acquisition and the long-term effects of spacing on retention. While several theories explain the locus of the spacing effect, we consider two qualitatively distinct sets of theories that study the processes occurring during practice that contribute to the spacing effect (Bjork \& Allen, 1970). The set of deficitprocessing theories focuses on the role of attention in the encoding of each practice opportunity. These theories hypothesize that, when practice opportunities are massed together, information used to encode and recall items is stored in working memory and little attention is needed to maintain these features. When practiced opportunities are spaced apart, features relevant to encoding are not stored in working memory
and greater attentional resources must be applied to the maintenance and retrieval of memorized associations (Cuddy \& Jacoby, 1982).

The second set of theories attributes the spacing effect to variation in encoding. These theories hypothesize that spacing between practice opportunities increases variability in contextual information used in the encoding of the item. When the item is seen again, the degree to which the item is encoded and associated with prior exposures is mediated by the time and number of items viewed between exposures (Landauer, 1969; Raaijmakers, 2003). This work explains why the benefit of spaced practice becomes apparent when retention is tested after a long period. An interesting phenomenon discussed and modeled by many encoding variability models investigates the increase in probability of recalling two different items, which have been seen only once, when those items were spaced further apart during training (Raaijmakers, 2003; Lohnas, Polyn, \& Kahana, 2011). In this case, as with spacing, the probability of recalling both items at a much later time increases the more separated those items were when first seen.

The two types of theories view learning in terms of strengthening a memory rather than acquiring a skill. While this view may be sufficient for paired-associates task, it potentially over-simplifies the learning processes of more complex procedures. Given that the spacing effect has been observed across domains, including the learning of complex procedures (Shea, Lai, Black, \& Park, 2000) and motor tasks (Shea et al., 2000), we stand to benefit from a closer consideration of how spacing and practice frequency influence procedural skill acquisition. Recent work by Tenison and Anderson (2016) suggests that skill acquisition is best described by both distinct changes in the cognitive processes used to perform a task, as well as quantitative improvements in the speed with which those cognitive processes are executed. With evidence from both behavioral and neuroimaging data, Tenison, Fincham, and Anderson (2016) suggest that these shifts follow the three phases of skill acquisition proposed by Fitts and Posner (1967). In the first phase, the Cognitive Phase, people must execute a series of procedures to perform a task. By the second phase, the Associative Phase, the response has been memorized and the task involves a single retrieval from memory. The third phase, the Autonomous Phase, the retrieval is dropped and the skill becomes a stimulus-response process. While this model has been operationalized in the ACT-R cognitive architecture (Anderson, 1982), prior work modeling the spacing effect within ACT-R predominantly accounts for the
role of forgetting and memory rather than tracking changes in cognitive processing (Pavlik \& Anderson, 2005). While this computational model has successfully modeled the impact of spacing on learning and retention, it is unclear how forgetting may interfere with skill acquisition.

In the current study, we apply an unsupervised learning approach developed to detect the phases of skill acquisition (Tenison \& Anderson, 2016) to the problem solving latency generated in a spaced-practice task. In this task, participants are introduced to a novel mathematics operation and practice a set of problems. Different items receive different amounts of practice within the same learning period, resulting in greater spacing for the less frequently practiced items and less spacing between more frequently practiced items. Using a within-subjects design, participants were each exposed to 4 practice frequency conditions. We look at the effects of this different practice frequency manipulations on both learning during the task and retention three days later. We hypothesize that our manipulation will affect both the rate of achieving more advanced phases and the rate at which problem solving speeds up within a phase. We were interested in determining whether the phase of skill acquisition achieved during study and amount of practice within a phase would predict retention.

## Methods

We ran our experiment on Amazon Mechanical Turk (MTurk). Participants signed up for two sessions, separated by a 66-72 hour period. On Day 1, participants were introduced to a novel math operator and practiced solving a fixed set of repeated problems. On Day 2 of the study, participants returned to complete the test including the items seen on Day 1 and answer questions about the strategies they used to solve the problems.

## Participants

43 participants ( 15 female) completed both days of our experiment. All participants were from the United States. Participants represented a diverse age range, 20 to 53 years ( $\mathrm{M}=30.9, \mathrm{SD}=6.3$ ), and education levels (highest level: 5 high school, 30 college, 8 graduate school). Reviewing the problem solving strategies participants had reported on Day 2, we excluded 8 participants who had used external aids on either day to solve problems. Our final sample included 35 participants. Participants were paid 2 cents for correctly solved problems (up to $\$ 8.60$ on Day 1 and $\$ 3.60$ on Day 2), as well as a $\$ 2$ base pay for Day 1 and a $\$ 10$ bonus for completing both days. This study was approved by the university internal review board and all participants gave informed consent for participation.

## Materials and procedure

Participants learned a novel type of mathematics called a Pyramid problem. Pyramid problems follow the form of Base\$Height, where the base indicates the first term in the additive sequence, and the height determines the number of
terms to be added together (e.g., $8 \$ 4=8+7+6+5$ ). Problem sets included heights of $3,4,5$, and 6 . The bases of our problem set varied from 4 to 11 with the restriction that the minimum base for a given height was height plus one. Because height determines the number of terms to sum, we use this as a means of manipulating problem difficulty. A total of 36 unique problems were used in the experiment. This study, however, will focus on only the 16 unique problems that were practiced on Day 1. For each problem, we recorded the accuracy and problem solving latency.

Day 1 After giving consent and completing a demographic information questionnaire, participants were introduced to the pyramid operation and given two blocks of 36 unique items to solve. Each item was presented on the screen in the form ' $8 \$ 4=$ _'. Following each input, we displayed corrective feedback by showing the expanded calculation and correct answer in the form ' $8+7+6+5=26$ '. After these pre-test blocks, participants then completed 10 practice blocks, which included 40 items each. During the practice period, participants practiced 16 unique problems. Practiced problems were divided into four Practice Frequency (PF) groups. Items in Practice Frequency-1 (PF-1) were seen once per block; in PF-2, twice per block; in PF-3, three times per block; and, in PF-4, four times per block. We included problems of four different heights (3-6) in each PF group, so that each block included 40 items total and 16 unique problems. By the end of Day 1, PF-1 items were seen 10 times, whereas PF-4 items were seen 40 times, as a result, our analyses are sensitive to both the effects of spacing and of general practice. Participants were given 4 hours to complete the tasks in Day 1.

Day 2 Participants were emailed a link to the retention test 66 hours after the initial completion of their Day 1 session. After the link was sent, participants had a total of 12 hours to begin the experiment (once started participants were limited to 2 hours to complete the experiment). Similar to the pre-test collected during the first two-blocks of Day 1, the retention test consisted of 10 blocks of 36 unique problems. This included 20 items that were only seen during the pretest on Day 1 , and 16 problems from the four PF groups.

## Results

The aim of this study is to explore the impact of spacing on the acquisition and retention of procedural skills. We divide our results into three sections. First, we report the general impact of spacing and practice frequency on learning and retention. We next fit the Tenison and Anderson (2016) to the data to generate parameter estimates and phase labels. Finally, we use mixed-effects modeling to explore the relationship between experimental condition and phase of skill acquisition.

## Descriptive statistics

Before fitting a model to the data to identify learning phases, we examined the effect of spacing on the speed and accuracy of problem solving. A repeated measures analysis of


Figure 1: Problem solving latency for items averaged within each experimental block. Separate means are calculated for items of each practice frequency group. (a) indicates performance on Day 1 (b) indicates performance on Day 2. Error bars represent standard error
variance (ANOVA) run on mean, log transformed latency data revealed a significant main effect of practice frequency group $(\mathrm{F}(9,306)=68.6, \mathrm{p}<.001)$ and block $(\mathrm{F}(3,102)=34.2$, $\mathrm{p}<.001$ ), and a significant interaction between PF and block $(\mathrm{F}(27,918)=2.3, \mathrm{p}<.01)$. Figure 1 shows that response latency decreased across blocks and appears lower for higherPF groups than lower-PF groups. The significant interaction suggests that the impact of block on speedup differs between PF groups. The average accuracy of items within the different PF groups also increases from PF-1 ( $\mathrm{M}=92 \%, \mathrm{SD}=.7 \%$ ) to PF4 ( $\mathrm{M}=97 \%, \mathrm{SD}=.2 \%$ ). A repeated measures ANOVA on accuracy data finds a significant main effect of PF group ( $\mathrm{F}(3,102)=8.7, \mathrm{p}<.001)$ such that PF-4 items were more accurate than lower practice frequencies, but the impact of block $(\mathrm{F}(9,306)=.9, \mathrm{p}=.5)$ and the interaction between block and PF group $(\mathrm{F}(27,918)=1.2, \mathrm{p}=.3)$ are not significant.

From the last items practiced on Day 1 to the first item practiced on Day 2, we see average decreases in accuracy from $95.6 \%(.5 \%)$ to $94 \%(1.0 \%)$, and decreases in reaction time from $2.8(.01 \mathrm{~s})$ to $5.4(.02 \mathrm{~s})$. We will focus on latency for Day 2 , which shows the large effect. A repeated measures ANOVA showed a significant main effect of practice frequency group $(\mathrm{F}(3,102)=3.4, \mathrm{p}<.05)$ and problem difficulty ( $\mathrm{F}(3,102)=5.4, \mathrm{p}<.005$ ) but no interaction $(\mathrm{F}(9,306)=1.2, \mathrm{p}=.3)$. The effect of problem difficulty present in all groups suggested that on Day 2 many of these problems were solved rather than retrieved. Furthermore, the mean response times for these items indicated that items in the higher practice frequency groups were solved more quickly. These analyses show that while the effects present in Day 1 remain on Day 2, they are quite attenuated. However, as we will see this is because of the mixing of items that have reached different phases of learning on Day 1.

## Model fitting

We fit the Tenison and Anderson (2016) power-law skill acquisition model to the response latencies for the items solved during the 10 practice blocks completed on Day 1. This
model (refer to Tenison and Anderson (2016) for a detailed description) uses a Hidden Markov model (HMM) to track both the participants learning phase for any given problem and the number of practice opportunities a participant has had within a given phase. Using the within-phase tracking, we estimated parameters for a power-law function to describe speedup in the execution of the cognitive processes specific to each of the phases. However, according to the model, larger, abrupt changes are caused by transitioning to a more advanced phase of processing. We fit our model to each PF group separately. The model was fit separately for each item solved by a participant, but used trends across all participants solving items within a PF group in order to generate parameter estimates. We considered the number of phases that best fit the data by fitting HMMs with 1 through 5 possible phase transitions. Thus, we fit a total of 20 models ( 1 to 5 phases fit for each PF group). We used two measures to evaluate which model best fit the data: Bayesian Information Criterion, which penalizes models for added parameters, and log likelihood generated from a leave-one-subject-out cross validation. We best fit a 3-phase model for all 4 PF groups, replicating the result from our earlier studies. Once we determined the number of phases best fit by a model, we refit it to all the data and labeled each item with the phase the model identifies as most probable.

Because models are fit separately for each practice group, we first needed to establish that the 3 Phases identified by each model were in fact the same cognitive processes. In prior work, we found evidence that participants used calculation strategies in the first learning Phase and retrieval strategies in Phase 2 and 3 (Tenison \& Anderson, 2016; Tenison et al., 2016). Because height determines the number of items participants sum together, we used the four different heights present in each PF group as a stand-in for problem difficulty. We would expect this effect of difficulty to disappear when participants use the retrieval strategies of Phase 2 and Phase 3. In Figure 2, each quadrant illustrates for each PF-group the effect of problem difficulty for the items in each Phase.


Figure 2: Mean problem solving latency for problems of different heights (i.e. difficulty levels) Each quadrant represents a different practice frequency group. Error bars represent standard error

We fit a mixed effects model to participants log latency data (Kliegl, Masson, \& Richter, 2010) with random intercepts for each participant. Our model included an interaction between height and learning phase, along with a fixed-effect for training. We initially fit a maximal model in which we considered a random-effect for each fixed-effect in our model, and then used BIC to test whether or not removing those random effects improved the model fit (Barr, Levy, Scheepers, \& Tily, 2013). Using this approach, we found our model was improved by inclusion of a random-effect for Phase, suggesting some variation across individuals in the impact of Phase on latency. We used the Kenward-Rodger (Kenward \& Roger, 1997) approximation for degrees of freedom . In our final model, we could account for a significant amount of the variance in response time, with a fixed effect for the PF group of the item $(\mathrm{F}(3,1194.8)=22.2, \mathrm{p}<.001)$, problem difficulty $(\mathrm{F}(3,1203.7)=43.9, \mathrm{p}<.001)$ and Phase $(\mathrm{F}(2,44.9)=487.5$, $\mathrm{p}<.001$ ), and a significant interaction between problem difficulty and Phase $(\mathrm{F}(6,1200.3)=34.5, \mathrm{p}<.001)$. This analysis, as evident in Figure 2, shows that for all 4 practice frequency groups there is an effect of problem difficulty present in Phase 1 , but not in Phase 2 or 3 . The main effect of PF group, while difficult to discern from Figure 2, suggests a slight tendency for the more frequent items to be faster.

## Interpreting model parameters

We gain insight into these effects by looking at the parameters of the 3-stage models that were estimated separately for each practice level. In this model, we fit a power function to each stage to reflect the within-stage practice. This is a 3-parameter function:

$$
\begin{equation*}
\mu_{r e t}=I+\beta n^{-\alpha} \tag{1}
\end{equation*}
$$

Where $\mu_{\text {ret }}$ is the time it takes to retrieve the answer, $I$ is
the asymptotic latency (i.e., the fastest possible time), $\beta$ is the amount of latency that can be reduced with practice, $n$ is the number of practice opportunities, and $\alpha$ is the learning rate. Asymptotic latency, $I$, and learning rate, $\alpha$, parameters were estimated across all three Phases, and $\beta$ was estimated separately for each Phase. These parameters capture the speed up within a Phase while our transition parameters, $T_{12}$ and $T_{23}$, describe the probability of transitioning from one Phase to the next. Table 1 shows these parameters. Across models of the different practice-frequency groups the intercepts are essentially zero, implying that practice will always reduce the latency to some degree. The learning rate is controlled by the parameter a, which is small, indicating relatively small within-phase speedup. The parameters are remarkably similar across practice groups. There is a tendency for the speed up parameter to be greater for the higher-practice groups partially accounting for the faster within-phase times in Figure 1.

Table 1: Parameters for the three phase model of skill acquisition for each practice frequency group.

|  | $I$ | $\alpha$ | $\beta_{p 1}$ | $\beta_{p 2}$ | $\beta_{p 3}$ | $T_{12}$ | $T_{23}$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| PF-1 | $3.7 \mathrm{e}-8$ | -.07 | 7.8 | 3.2 | 1.6 | .17 | .15 |
| PF-2 | $9.5 \mathrm{e}-8$ | -.10 | 8.3 | 3.5 | 1.7 | .13 | .09 |
| PF-3 | $3.5 \mathrm{e}-7$ | -.11 | 8.2 | 3.3 | 1.6 | .07 | .06 |
| PF-4 | $5.9 \mathrm{e}-9$ | -.12 | 8.1 | 3.3 | 1.6 | .06 | .06 |

Because the number of practice opportunities vary between PF groups, the number of problems that reach Phase 3 is significantly different between the four groups $(\mathrm{F}(3,102)=6.6$, $\mathrm{p}<.001$ ), with $55.7 \%$ (6.7) of PF-4 problems reaching phase $3,47.1 \%$ of PF-3 (6.1), 50\% of PF-2 (5.7), and 35.6\% of PF1(6.7). However, the transition parameters, $T_{12}$ (Phase 1 to 2) and $T_{23}$ (Phase 2 to 3), are smaller for the high frequency group indicating that high frequency items spend more time in a phase before transitioning into the next Phase. On average, PF-4 items are seen 14.3 (1.1) times before transitioning to Phase 2 and 12.9 (1.0) before Phase 3, whereas PF-1 items are seen 4.7 (.29) times before Phase 2 and 3.6 (.25) times before Phase 3.

## Retention after three days

To understand how the performance on Day 2 was impacted by practice on Day 1, we fit a mixed-effects model to the log latency of the first observation of each problem on Day 2. We explored fixed effects for practice frequency, Phase reached on Day 1, time spent within that Phase on Day 1, and problem difficulty. Fitting a maximal model with all factors and random effects, we systematically removed factors that did not account for variance within the model (see Barr et al. (2013) for method). In fitting our model, we found no improvement justifying the inclusion of random effects; and, we found that neither the practice frequency group nor the amount of practice within the last phase reached accounted


Figure 3: (a) To the left of the vertical line, mean problem solving latency for items in experimental blocks on Day 1; line types distinguish the mean latency for items in each Phase. To the right of the vertical line, mean problem solving latencies for items grouped by last Phase reached on Day 1 (b) Mean latency of first block of Day 2. Line type indicates Phase item reached on Day 1. Error bars represent standard error
for enough variance in response latency to justify either fixed effect. Our final model indicated a significant effect of problem difficulty $(\mathrm{F}(3,515.8)=5.5, \mathrm{p}<.001)$, Phase reached on Day $1(\mathrm{~F}(2,542.5)=20.2, \mathrm{p}<.001)$, and a significant interaction between Difficulty and Phase $(\mathrm{F}(6,519.1)=2.4, \mathrm{p}<.05)$. Figure 3a shows the mean effect of phase. It is striking that the items that were still in Phase 1 show little change in speed, while the items in Phases 2 and 3 slow down from Day 1 to Day 2. Figure 3b shows the mean response latencies for the first items solved on Day 2. The effect of problem difficulty appears to be present for both items that reached Phase 1 and 2 on Day 1, but less of an observable effect for items that reached Phase 3 on Day 1. Our interpretation of these results is that items in Phase 1 on Day 1 stay in Phase 1 and therefore show no changes in their latency patterns. However, some of the items in Phases 2 and 3 slip back a phase over the retention interval and therefore slow down. Items that slip back to Phase 1 will show a problem difficulty effect, possibly explaining the presence of a problem difficulty effect for items in Phase 2 (Figure 3b).

## Discussion

Anderson and Milson (1989) suggest that memory phenomenon represent a joint function between general properties of memory and the strategies individuals use to process information. Our findings are aligned with prior studies that consider the impact of the spacing and practice frequency performance during training. Higher frequency in practice contributes to greater improvements in accuracy and response latency during training (e.g. Cepeda, Pashler, Vul, Wixted, \& Rohrer, 2006). The application of the skill acquisition model gives us insight into how problem solving strategies change in response to practice. We find the problem-solving latency advantage of massed trials is concentrated in the speedup within a phase rather than the transition between phases. This could be envisioned as 'rich get richer' process (Simon, 1955), in which learning strengthens both the probability of applying
the previously used strategy and the speed with which the sub-procedures of that strategy are executed. Items that are spaced further apart exhibit shallower learning rates, which may make the search for a more efficient strategy more rewarding than the learning of problem solving sub-procedures. While theories of deficit processing or contextual variability could provide a mechanism for the differences in within phase speed up, the shift between phases of skill acquisition may be driven by a different mechanism.

Unlike prior work, which largely uses accuracy to measure the impact of spacing on retention, in our study we use response latency and the effect of problem difficulty. When we include information about what phase each item reached on the first day of training, we find that practice frequency no longer accounts for significant variance in problem solving latency at the retention test. Analyzing the speed of problem solving on Day 2, it appears that problems that items that reach Phase 2 and 3 on Day 1 are solved more quickly than items that remain in Phase 1. Additionally, the significant interaction between Phase and problem difficulty suggests that Phase 3 items may still be retrieved on Day 2, while Phase 1 and 2 items are calculated. This work is consistent with findings of Sisti, Glass, and Shors (2007) who found that the survival of neurons in the dentate gyrus and the strength of memory in an animal model was predicted not by whether or not practice was spaced or massed, but by how well the animals learned the task. While this study provides a biological mechanism for memory preservation, it is unclear computationally what memory process would explain the impact of phase on retention, but not on spacing nor general practice. In future work, we will explore how phase impacts retention by incorporating forgetting into our computational model of skill acquisition. In incorporating this capability, we will consider forgetting both in terms of regressing to a prior phase, and as regression within a phase. Additionally, including regression into the model will allow us to explore how spacing and skill acquisition on Day 1 impacts relearning on Day 2. In this
future work we will limit the practice of more frequent items so to dissociate the effects of spacing from those of practice frequency. This work, while in an early stage, suggests that without considering the impact of skill acquisition on problem solving strategies, our models of the spacing effect, and memory more generally, are incomplete.

## Acknowledgments

This work was supported by the National Science Foundation grant DRL-1420008, a James S. McDonnell Scholar Award, and by Carnegie Mellon University's Program in Interdisciplinary Education Research funded by the US Department of Education grant R305B090023. We would like to thank Shawn Betts for his help running the Mechanical Turk study.

## References

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological review, 89(4), 369.
Anderson, J. R., \& Milson, R. (1989). Human memory: An adaptive perspective. Psychological Review, 96(4), 703.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.
Bjork, R. A., \& Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? Journal of Verbal Learning and Verbal Behavior, 9(5), 567-572.
Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., \& Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological bulletin, 132(3), 354.
Cuddy, L. J., \& Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. Journal of Verbal Learning and Verbal Behavior, 21(4), 451-467.
Fitts, P. M., \& Posner, M. I. (1967). Human performance. Oxford: Brooks/Cole.
Kenward, M. G., \& Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 983-997.
Kliegl, R., Masson, M. E., \& Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. Visual Cognition, 18(5), 655-681.
Landauer, T. K. (1969). Reinforcement as consolidation. Psychological Review, 76(1), 82.
Lohnas, L. J., Polyn, S. M., \& Kahana, M. J. (2011). Contextual variability in free recall. Journal of memory and language, 64(3), 249-255.
Pavlik, P. I., \& Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cognitive Science, 29(4), 559586.

Raaijmakers, J. G. (2003). Spacing and repetition effects in human memory: Application of the sam model. Cognitive Science, 27(3), 431-452.

Shea, C. H., Lai, Q., Black, C., \& Park, J.-H. (2000). Spacing practice sessions across days benefits the learning of motor skills. Human movement science, 19(5), 737-760.
Simon, H. A. (1955). On a class of skew distribution functions. Biometrika, 42(3/4), 425-440.
Sisti, H. M., Glass, A. L., \& Shors, T. J. (2007). Neurogenesis and the spacing effect: learning over time enhances memory and the survival of new neurons. Learning \& memory, 14(5), 368-375.
Tenison, C., \& Anderson, J. R. (2016). Modeling the distinct phases of skill acquisition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(5), 749.
Tenison, C., Fincham, J. M., \& Anderson, J. R. (2016). Phases of learning: How skill acquisition impacts cognitive processing. Cognitive psychology, 87, 1-28.

# Warm (for winter): Comparison class understanding in vague language 

Michael Henry Tessler ${ }^{1}$, Michael Lopez-Brau ${ }^{2}$, and Noah D. Goodman ${ }^{1}$<br>mtessler@stanford.edu, lopezmic@knights.ucf.edu, ngoodman@stanford.edu<br>${ }^{1}$ Dept. of Psychology, Stanford University, ${ }^{2}$ Dept. of Electrical \& Computer Engineering, University of Central Florida


#### Abstract

Speakers often refer to context only implicitly when using language. The utterance "it's warm outside" could signal it's warm relative to other days of the year or just relative to the current season (e.g., it's warm for winter). Warm vaguely conveys that the temperature is high relative to some contextual comparison class, but little is known about how a listener decides upon such a standard of comparison. Here, we formalize how world knowledge and listeners' internal models of speech production can drive the resolution of a comparison class in context. We introduce a Rational Speech Act model and derive two novel predictions from it, which we validate using a paraphrase experiment to measure listeners' beliefs about the likely comparison class used by a speaker. Our model makes quantitative predictions given prior world knowledge for the domains in question. We triangulate this knowledge with a follow-up language task in the same domains, using Bayesian data analysis to infer priors from both data sets. Keywords: comparison class; pragmatics; Rational Speech Act; Bayesian cognitive model; Bayesian data analysis


If it's $75^{\circ} \mathrm{F}\left(24^{\circ} \mathrm{C}\right)$ outside, you could say "it's warm." If it's $60^{\circ} \mathrm{F}\left(16^{\circ} \mathrm{C}\right)$, you might not consider it warm. Unless it's January; it could be warm for January. Warm is relative, and its felicity depends upon what the speaker uses as a basis of comparison-the comparison class (e.g., other days of the year or other days in January). Comparison classes are necessary for understanding adjectives and, in fact, any part of language whose meaning must be pragmatically reconstructed from context, including vague quantifiers (e.g., "He ate a lot of burgers."; Scholler \& Franke, 2015) and generic language (e.g., "Dogs are friendly"; Tessler \& Goodman, 2016a). The challenge for listeners is that the comparison class often goes unsaid (e.g., in "It's warm outside.").

The existence of comparison classes for understanding vague language is uncontroversial (Bale, 2011; Solt, 2009). Four-year-olds categorize novel creatures (pimwits) as either "tall" or "short" depending on the distribution of heights of pimwits and not the heights of creatures that are not called pimwits, suggesting the comparison class in that context is other pimwits (Barner \& Snedeker, 2008). Adult judgments of the felicity for adjectives like "dark" or "tall" similarly depend upon fine-grained details of the statistics of the comparison class (Qing \& Franke, 2014b; Schmidt, Goodman, Barner, \& Tenenbaum, 2009; Solt \& Gotzner, 2012).

Any particular object of discourse, however, can be conceptualized or categorized in multiple ways, giving rise to multiple possible comparison classes. A day in January is also a day of the year; if it's warm, it could be warm for winter or warm for the year. Why should one comparison class be preferred over another? To our knowledge, this question has not been addressed formally or empirically ${ }^{1}$ We pro-

[^207]pose that listeners actively combine category knowledge with pragmatic considerations to infer the comparison class implicitly used by the speaker. We introduce a minimal extension to the Rational Speech Act (RSA) model for gradable adjectives (Lassiter \& Goodman, 2013) to allow it to flexibly reason about the implicit comparison class.

We derive two novel qualitative predictions from this model. Saying "it's warm" in winter should signal it's warm for winter (as opposed to for the year) more so than saying "it's cold". The opposite relationship should hold in summer, where "it's cold" should signal it's cold for summer more so than "it's warm". This prediction is driven by the a priori probability that the adjective could apply to the class (e.g., the probability that a given day in winter is warm; Prediction 1). In addition, regardless of the season and the adjective form (e.g., "warm" or "cold"), listeners who expect speakers to be informative will prefer classes that are relatively specific (e.g., relative to the current season as opposed to the whole year), as they carry more information content (Prediction 2). We test these predictions by eliciting the comparison class using a paraphrase dependent measure (Expt. 1).

As with any Bayesian cognitive model, explicitly specifying relevant prior knowledge (e.g., beliefs about temperatures) is necessary for the model to make quantitative predictions. The current methodological standard is to measure beliefs by having participants estimate quantities or give likelihood judgments (Franke et al., 2016). We pursue a different methodology. The RSA model captures a productive fragment of natural language; thus, it makes predictions about a related natural language task (Expt. 2). Critically, we can use the model to predict natural language judgments that require the same prior knowledge as in Expt. 1 and use Bayesian data analysis to jointly infer the shared priors. This approach harnesses the productivity of language into experiment design and allows us to reconstruct priors without having participants engage in challenging numerical estimation tasks.

## Understanding comparison classes

Adjectives like warm and cold are vague descriptions of an underlying quantitative scale (e.g., temperature). The vagueness and context-sensitivity of these adjectival utterances can be modeled using threshold semantics $([\llbracket u \rrbracket]=x>\theta$, for utterance $u$, scalar degree $x$, and threshold $\theta$ ), where the threshold is probabilistically set with respect to a comparison class $c$ via pragmatic reasoning (Lassiter \& Goodman, 2013; see also Qing \& Franke, 2014a):

[^208]\[

$$
\begin{align*}
L_{1}(x, \theta \mid u) & \propto S_{1}(u \mid x, \theta) \cdot P_{c}(x) \cdot P(\theta)  \tag{1}\\
S_{1}(u \mid x, \theta) & \propto \exp \left(\alpha_{1} \cdot \ln L_{0}(x \mid u, \theta)\right)  \tag{2}\\
L_{0}(x \mid u, \theta) & \propto \delta_{\llbracket u\rfloor(x, \theta)} \cdot P_{c}(x) \tag{3}
\end{align*}
$$
\]

This is a Rational Speech Act (RSA) model, a recursive Bayesian model where speaker $S$ and listener $L$ coordinate on an intended meaning (for a review, see Goodman \& Frank, 2016). In this framework, the pragmatic listener $L_{1}$ tries to resolve the state of the world $x$ (e.g., the temperature) from the utterance she heard $u$ (e.g., "it's warm"). She imagines the utterance came from an approximately rational Bayesian speaker $S_{1}$ trying to inform a naive listener $L_{0}$, who in turn updates her prior beliefs $P_{c}(x)$ via an utterance's literal meaning $\llbracket u \rrbracket](x)$. Lassiter \& Goodman (2013) introduced into RSA uncertainty over a semantic variable: the truth-functional threshold $\theta$ (Eq. 11). $\theta$ comes from an uninformed prior and is resolved by the listener by reasoning about the likely states of the world $P_{c}(x)$ (e.g., possible temperatures) and the likelihood that a speaker would say the adjective given a state and a threshold $S(u \mid x, \theta)$. The prior distribution over world-states $P_{c}(x)$ is always relative to some comparison class $c$ (Eqs. $1 \&$ (3) but where does the comparison class come from?

When a listener hears only that "it's warm outside" without an explicit comparison class (e.g., "... for the season"), we posit the listener infers the comparison class using her world knowledge of what worlds are plausible given different comparison classes $P(x \mid c)$, what comparison classes are likely to be talked about $P(c)$, and how a rational speaker would behave in a given world and comparison class $S_{1}(u \mid x, c, \theta)$ (Eq. 4.) As a first test of this idea, we consider an idealized case where the comparison class can be either a relatively specific (subordinate) or relatively general (superordinate) categorization (e.g., warm relative to days in winter or relative to days of the year). Crucially in this situation, the listener is aware that the target entity is a member of the subordinate class (e.g., aware that it is winter) and draws likely values of the degree (e.g., temperature) from the subordinate class prior $P\left(x \mid c_{\text {sub }}\right)$. With these assumptions, the model becomes:

$$
\begin{align*}
& L_{1}(x, c, \theta \mid u) \propto S_{1}(u \mid x, c, \theta) \cdot P\left(x \mid c_{s u b}\right) \cdot P(c) \cdot P(\theta)  \tag{4}\\
& S_{1}(u \mid x, c, \theta) \propto \exp \left(\alpha_{1} \cdot \ln L_{0}(x \mid u, c, \theta)\right)  \tag{5}\\
& L_{0}(x \mid u, c, \theta) \propto \delta_{\llbracket u \rrbracket(x, \theta)} \cdot P(x \mid c) \tag{6}
\end{align*}
$$

We are interested in the behavior of the model with the underspecified utterance (e.g., "It's warm"), and we assume the speaker has two alternative utterances in which the comparison class is explicit (e.g., "It's warm relative to other days in winter." and "It's warm relative to other days of the year."). The predictions of this model depend on the details of the listener's knowledge of the subordinate and superordinate categories: $P\left(x \mid c_{\text {sub }}\right)$ and $P\left(x \mid c_{\text {super }}\right)$, as well as the prior distribution on comparison classes $P(c)$ in Eq. 4

Comparison class prior $P(c)$ reflects listeners' expectations of what classes are likely to be discussed. As a proxy
for comparison class usage frequency, we use empirical frequency $\hat{f}$ estimated from the Google WebGram corpus ${ }^{2}$, and scale it by a free parameter $\beta$ such that $P(c) \propto \exp (\beta \cdot \log \hat{f})$.
Degree priors (World knowledge) Only the relative values for $P\left(x \mid c_{\text {sub }}\right)$ and $P\left(x \mid c_{\text {super }}\right)$ affect model predictions. Hence we fix each superordinate distribution to be a standard normal distribution $P\left(x \mid c_{\text {super }}\right)=\mathcal{N}(0,1)$ and the subordinate priors to also be Gaussian distributions $P\left(x \mid c_{\text {sub }}\right)=$ $\mathcal{N}\left(\mu_{\text {sub }}, \sigma_{\text {sub }}\right)$; the subordinate priors thus have standardized units. We will eventually infer the parameters of the subordinate priors from experimental data.





Figure 1: Left: Three hypothetical subordinate class prior distributions over a degree (fixing the superordinate class to be a unit-normal distribution, in grey). Right: Predicted listener inferences for an intended subordinate class interpretation given positive and negative form adjectives with different subordinate degree priors.
Qualitative model predictions Figure 1 (left) shows schematic superordinate and subordinate priors; e.g., temperatures over the whole year (super), in winter (low), fall (medium), and summer (high). The subordinate distributions have lower variance than the superordinate, and the "low" and "high" distributions have different means (e.g., temperatures in winter are expected to be lower and have lower variance than temperatures over the whole year).

Two intuitions explain the inferences of the pragmatic listener model (shown in Figure 1 right). First, certain classes are more or less likely to have an adjective felicitously apply. For example, any given day in winter is less likely to be warm than cold. Thus, hearing "it's warm" (a positive-form adjective) in winter (low prior) will signal it's warm for winter (the subordinate class) more so than hearing "it's cold" (negativeform), because it's more likely to be true (Prediction 1).

[^209]| Scale (adjectives) | Subordinate classes | Superordinate |
| :--- | :--- | :--- |
| Height (tall, short) | (professional) gymnast, soccer player, basketball player | people |
| Price (expensive, cheap) | bottle opener, toaster, dishwasher | kitchen appliances |
| Temperature (warm, cold) | winter, fall, summer (day in Maryland) | days in the year |
| Time (long, short) | video of a cute animal, music video, movie | things you watch online |
| Weight (heavy, light) | grape, apple, watermelon | produce |

Table 1: Items used in Experiments 1 and 2. Subordinate categories were designed to fall near the low end, high end, and somewhere in the middle of the degree scale

Second, the amount of information conveyed by a vague utterance depends upon the variability in the comparison class. Comparison classes that have higher variance will result in relatively less information gain by the listener. All else being equal, listeners will prefer lower variance (e.g., subordinate) comparison classes because they are more informative (Prediction 2). Figure 1 (right) shows that subordinate class interpretations are above baseline regardless of the adjective polarity (positive or negative) or the mean of the subordinate prior (low, medium, high).

In sum, we see two predictions: The pragmatic listener overall prefers subordinate comparison classes, though the extent of this preference is modulated by the a priori probability that the adjective is true of the subordinate category. We test these two predictions in our first experiment.

Overview of data analytic approach As described above, specifying the relevant prior knowledge yields two free parameters per subordinate class. We will put priors over these parameters and infer their likely values using Bayesian data analysis. The data from the comparison class experiment (Expt. 1) would be insufficient, however, to reliably estimate all of the parameters of this data analytic model. To alleviate this, we use the same RSA model to predict additional data about related language use in the same domains (Expt. 2). Specifically, we gather judgments about adjectives when the comparison class is explicit: whether or not an adjective would apply to a subordinate member explicitly relative to the superordinate category (e.g., Is a day in winter warm relative to other days of the year?).

To model Expt. 2 data, we remove comparison class uncertainty by setting $P\left(c_{\text {super }}\right)=1$, since the sentences provide an explicit comparison to the superordinate class. We model sentence endorsement using a pragmatic speaker (following Qing \& Franke, 2014a; Tessler \& Goodman, 2016a, 2016b):

$$
\begin{equation*}
S_{2}\left(u \mid c_{\text {sub }}\right) \propto \exp \left(\alpha_{2} \cdot \mathbb{E}_{x \sim P_{c_{\text {sub }}}} \ln L_{1}(x \mid u)\right) \tag{7}
\end{equation*}
$$

Note that $L_{1}(x \mid u)$ is defined from Eq. 4 by marginalization.
Eqs. 4 and 7 define models for the data we will gather from Expts. 1 and 2, and depend on the same background knowledge $P(x \mid c)$. We can thus use data from both experiments to jointly reconstruct the shared prior knowledge and generate predictions for the two data sets. Experimental paradigms, computational models, preregistration report, and data for this paper can be found at https://mhtess.github.io

## Behavioral experiments

Experiment 1 tests the qualitative predictions of the model. Experiment 2 collects further data about adjective usage in order to constrain the quantitative predictions of the RSA model, which will be used to predict data from both experiments. The materials and much of the design of the two experiments are shared. Participants were recruited from Amazon's Mechanical Turk and were restricted to those with U.S. IP addresses with at least a $95 \%$ work approval rating. Each experiment took about 5 minutes and participants were compensated $\$ 0.50$ for their work.

Materials We used positive- and negative-form gradable adjectives describing five scales (Table 11. Each scale was paired with a superordinate category, and for each superordinate category, we used three subordinate categories that aimed to be situated near the high-end, low-end, and intermediate part of the degree scale (as in Figure 1 left). This resulted in 30 unique items ( $\{3$ subordinate categories $\} \times\{5$ scales $\}$ x $\{2$ adjective forms $\}$ ). Each participant saw 15 trials: one for each subordinate category paired with either the positive or negative form of its corresponding adjective. Participants never judged the same subordinate category for both adjective forms (e.g., cold and warm winter days) and back-to-back trials involved different scales to avoid fatigue.

## Experiment 1: Comparison class inference

In this experiment, we gather human judgments of comparison classes in ambiguous contexts, testing the two predictions described in Qualitative Model Predictions.

Participants and procedure We recruited 264 participants and 2 were excluded for failing an attention check. On each trial, participants were given a context sentence to introduce the subordinate category (e.g., Tanya lives in Maryland and steps outside in winter.). This was followed by an adjective sentence, which predicated either a positive- or negativeform gradable adjective over the item (e.g., Tanya says to her friend, "It's warm."). Participants were asked "What do you think Tanya meant?" and given a two-alternative forcedchoice to rephrase the adjective sentence with either an explicit subordinate or superordinate comparison class:
\{She / He / It \} is ADJECTIVE (e.g., warm) relative to other SUBORDINATES (e.g., days in winter) or SUPERORDINATES (e.g., days of the year)


Figure 2: Empirical comparison class data, inferred world priors, and empirically derived comparison class priors. Top: Experiment 1 results. Comparison class judgments in terms of proportion judgments in favor of subordinate comparison class. Middle: Inferred prior distributions of world knowledge used to model Experiment 1 and 2 data. Bottom: Inferred prior probability of the subordinate comparison classes based on Google WebGram frequencies. Error bars correspond to $95 \%$ Bayesian credible intervals (for bottom plot, derived from the posterior on the $\beta$ scale parameter).

In addition to all of the above design parameters, half of our participants completed trials where an additional sentence introduced the superordinate category at the beginning (e.g., Tanya lives in Maryland and checks the weather every day.), with the intention of making the superordinate paraphrase more salient.

Results We observed no systematic differences between participants' responses when the superordinate category was previously mentioned in the context and those when it was not; thus, we collapse across these two conditions for all analyses. Figure 2 (top) shows the proportion of participants choosing the subordinate paraphrase for each item, revealing considerable variability both within- and across- scales. The predicted effects are visually apparent within each scale (compare with Figure 1 right).

Our qualitative predictions are confirmed using a generalized linear mixed effects model with main effects of adjective form (positive vs. negative) and the a priori judgment by the first author of whether the sub-category was expected to be
low or high on the degree scale, and of critical theoretical interest, the interaction between these two variables. In addition, we included by-participant random effects of intercept and by-subordinate category random effects of intercept and iteraction between form and strength ${ }^{3}$. Confirming our two qualitative model predictions, there was an interaction between form and strength ( $\beta=-3.75$; $S E=0.58 ; z=-6.49$ ) and there was an overall preference for subordinate category paraphrases ( $\beta=1.21 ; S E=0.37 ; z=3.27$ ). The main effects of form and strength were not significant.

We then test the simple effects. For items low on the degree scale (e.g., temperatures in winter), positive form adjectives were significantly more likely to imply subordinate comparison classes ( $\beta=1.41$; SE $=0.15 ; z=9.43$ ), while the opposite is true for items high on the scale (e.g., summer days; $\beta=-2.5 ; S E=0.19 ; z=-13.15)$. Participants reason pragmatically to resolve the comparison class, combining world knowledge with informativity as predicted by our model.

[^210]

Figure 3: Human endorsement of subordinate comparison class paraphrases (middle; Expt. 1) and adjective sentences (left; Expt. 2) as a function of listener model $L_{1}$ and speaker model $S_{2}$ predictions, respectively. The right facet displays a subset of the paraphrase data (Expt. 1) to reveal good quantitative fit even in a small dynamic range. Error bars correspond to $95 \%$ Bayesian credible intervals.

## Experiment 2: Adjective endorsement

In this experiment, we collected data about adjective endorsement that would require the same prior knowledge relevant for Expt. 1. We use this data to further constrain the RSA model's quantitative predictions.
Participants and procedure We recruited 100 participants and 5 were excluded for failing an attention check. On each trial, participants were given a sentence introducing the subordinate category (e.g., Alicia lives in Maryland and steps outside in winter.). This was followed by a question asking if the participant would endorse an adjective explicitly relative to the superordinate category (e.g., Do you think the day in winter would be warm relative to other days of the year?).

Results The judgments in this experiment were consistent with the a priori ordering of the subordinate categories on the degree scale. On the y-axis of Figure 3 (left), we see that the endorsement of adjectival phrases in these domains is markedly more categorical than the comparison class inference task (compare vertical spread of left and middle facets).

## Full model analysis and results

The RSA listener (Eq. 4) and speaker (Eq. 7) models make quantitative predictions about comparison class interpretation and adjective endorsement, respectively. We construct a single data-analytic model with each of these RSA components as sub-models in order to make quantitative predictions about the data from both of our experiments.

The listener and speaker sub-models share their prior world knowledge $P(x \mid c)$ (e.g., temperatures in winter), described in the Degree Priors section. We put the same priors over the parameters of each subordinate distribution: $\mu \sim$ Uniform $(-3,3), \sigma \sim \operatorname{Uniform}(0,5)$, since they have standardized units. The comparison class prior $P(c)$ in Eq. 4
scales the empirical frequency $\hat{f}$ by a free parameter, which we give the following prior: $\beta \sim \operatorname{Uniform}(0,3)$.

The full model has three additional parameters not of direct theoretical interest: the speaker optimality parameters $\alpha_{i}^{\text {expt }}$, which can vary across the two tasks. The pragmatic listener $L_{1}$ model (Eq. 4) has one speaker optimality: $\boldsymbol{\alpha}_{1}^{1}$. The pragmatic speaker $S_{2}$ model (Eq. 7) has two speaker optimality parameters: $\left\{\alpha_{1}^{2}, \alpha_{2}^{2}\right\}$. We use priors consistent with the previous literature: $\alpha_{1} \sim \operatorname{Uniform}(0,20), \alpha_{2} \sim \operatorname{Uniform}(0,5)$

We implemented the RSA and Bayesian data analysis models in the probabilistic programming language WebPPL (Goodman \& Stuhlmuller, 2014). To learn about the credible values of the parameters, we collecting 2 chains of 50 k iterations (after 25k burn-in) using an incrementalized version of MCMC (Ritchie, Stuhlmuller, \& Goodman, 2016).

Results The full model's posterior over the RSA and dataanalytic parameters were consistent with prior literature and intuition. The maximum a-posteriori (MAP) estimate and 95\% highest probability density (HPD) intervals for model parameters specific to the $L_{1}$ model used for Expt. 1 were $\alpha_{1}^{1}=1.6[1.1,2.5], \beta=0.13[0.11,0.19]$. Model parameters specific to the $S_{2}$ model used for Expt. 2: $\alpha_{1}^{2}=3.5[0.6,13.2]$, $\alpha_{2}^{2}=3.2[2.6,3.8]$. The inferred distributions corresponding to subordinate class priors were consistent with the a priori ordering of these subordinate classes (low, medium, high) used in these tasks (Figure 2 middle).

Finally, the full model's posterior predictive distribution does an excellent job at capturing the quantitative variability in responses for Expt. 1: $r^{2}(30)=0.965$, and Expt. 2: $r^{2}(30)=0.985$ (Figure 3). Because of the overall preference for the subordinate comparison class, many of the data points are distributed above 0.5 . Even for these fine-grained differences, the model does a good job at explaining the quantitative variability in participants' data (Figure 3 right).

## Discussion

The words we say are often too vague to have a single, precise meaning and only make sense in context. Context, however, can also be underspecified, as there are many possible dimensions or categories that a speaker might be implicitly referring to or comparing against. Here, we investigate the flexibility in the class against which an entity can be implicitly compared.

We introduced a minimal extension to an adjective interpretation Rational Speech Act model to allow it to flexibly reason about the comparison class. This model made two novel predictions about how listeners should prioritize one class over another. It also made quantitative predictions about how background knowledge about the degree scale should inform this inference in a graded fashion. Both qualitative predictions of the model were borne out in our first experiment, and the quantitative predictions were confirmed using a novel data analytic technique. To our knowledge, this is the first experiment to demonstrate how reference classes for adjective interpretation can adjust based on world knowledge.

We observe in our modeling results for Expt. 1 that a uniform prior distribution over the experimentally supplied comparison class alternatives is unlikely (Figure 2 bottom). For example, the comparison class of "people" for heights of individuals is relatively more salient than the class of "produce" for the weights of fruits and vegetables. We used the frequency of the class in a corpus as a proxy for their prior probability $P(c)$, which was sufficient to account for differences in baseline class probability both between- and within-scales.

Corpus frequency is a composite measurement of factors relevant for speech production. Its utility in this model suggests that utterances without an explicit comparison class (e.g., "It's warm outside") may in fact be incomplete sentences, in a way analogous to sentence fragments studied in noisy-channel models of production and comprehension (Bergen \& Goodman, 2015). Another (non-mutually exclusive) possibility is that the comparison class prior reflects basic-level effects in categorization (Rosch \& Mervis, 1975). Future work should attempt to understand these factors to construct a more complete theory of the comparison class prior.

The second contribution of this paper is a novel dataanalytic approach, where prior knowledge used in the Bayesian language model is reconstructed from converging evidence gathered from related language experiments. In previous work, we have attempted to measure prior knowledge by decomposing what would be a single, implicitly multilayered, numerical estimation question into multiple simpler questions. Then, we construct a Bayesian data analytic model to back out the prior knowledge (Tessler \& Goodman, 2016a, 2016b). We extend this approach by using the same core RSA model to model behavior across two language experiments. The major feature of this method is that participants respond only to simple, natural language questions rather than estimating numerical quantities for which complicated linking functions must be designed (e.g., Franke et al., 2016). The
fully Bayesian language approach we pioneer here also provides a further constraint on the language model, which must predict data from two similar but distinct language experiments. The productivity of natural language can thus be harnessed to productively design experiments that further constrain and test computational models of language and cognition.

## Acknowledgements

The authors would like to thank Ali Horowitz for help in stimuli design. This work was supported in part by NSF Graduate Research Fellowship DGE-114747 to MHT, a Stanford CSLI Summer Internship for MLB, and a Sloan Research Fellowship, ONR grant N00014-13-1-0788, and DARPA grant FA8750-14-2-0009 to NDG.

## References

Bale, A. C. (2011). Scales and comparison classes. Natural Language Semantics, 19(2), 169-190.
Barner, D., \& Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. Child Development, 79(3), 594-608.
Bergen, L., \& Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. Topics in Cognitive Science, 7(2), 336-350.
Franke, M., Dablander, F., Scholler, A., Bennett, E., Degen, J., Tessler, M. H., ... Goodman, N. D. (2016). What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data. In Proceedings of the 38th annual meeting of the cognitive science society.
Goodman, N. D., \& Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences, 20(11), 818-829.
Goodman, N. D., \& Stuhlmuller, A. (2014). The Design and Implementation of Probabilistic Programming Languages. http: //dippl.org
Lassiter, D., \& Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Semantics and linguistic theory (Vol. 23, pp. 587-610).
Qing, C., \& Franke, M. (2014a). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In Semantics and linguistic theory (Vol. 24, pp. 23-41).
Qing, C., \& Franke, M. (2014b). Meaning and Use of Gradable Adjectives: Formal Modeling Meets Empirical Data. In Proceedings of the 36th annual conference of the cognitive science society.
Ritchie, D., Stuhlmuller, A., \& Goodman, N. D. (2016). C3: Lightweight incrementalized mcmc for probabilistic programs using continuations and callsite caching. In AISTATS 2016.
Rosch, E., \& Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7(4), 573-605.
Schmidt, L. A., Goodman, N. D., Barner, D., \& Tenenbaum, J. B. (2009). How Tall Is Tall? Compositionality, Statistics, and Gradable Adjectives. In Proceedings of the 31st annual conference of the cognitive science society.
Scholler, A., \& Franke, M. (2015). Semantic values as latent parameters: Surprising few \& many. In Semantics and linguistic theory (Vol. 25, pp. 143-162).
Solt, S. (2009). Notes on the Comparison Class. In International workshop on vagueness in communication.
Solt, S., \& Gotzner, N. (2012). Experimenting with degree. In Semantics and linguistic theory (Vol. 22, pp. 166-187).
Tessler, M. H., \& Goodman, N. D. (2016a). A pragmatic theory of generic language. ArXiv Preprint ArXiv:1608.02926.
Tessler, M. H., \& Goodman, N. D. (2016b). Communicating generalizations about events. In Proceedings of the 38th annual meeting of the cognitive science society.

# Neural and computational arguments for memory as a compressed supported timeline 

Zoran Tiganj (zoran.tiganj@gmail.com)<br>Karthik H. Shankar (kshankar79@gmail.com)<br>Marc W. Howard (marcwhoward777@gmail.com)<br>Center for Memory and Brain<br>Department of Psychological and Brain Sciences<br>Boston University, Boston, MA 02215


#### Abstract

It is well known that, all things being equal, the accuracy of mammalian timing and memory decays gradually with the passage of time. The gradual decay of temporal accuracy is also observed in single-unit neural recordings. Here we review recent modeling work describing a specific mechanism for timing and memory and relevant neural data. The model describes a neural mechanism that can give rise to a logarithmically compressed representation of the recent past. We examine the specific predictions of the model, in particular that the elapse of time is represented by sequentially activated cells which fire for a circumscribed period of time. Such cells, called time cells, have been observed in neural recordings from several brain regions in multiple species. As predicted by the model, the cells show accuracy that decreases with time.


Keywords: scale-invariance, memory, interval timing, time cells.

## Introduction

Behavioral experiments on humans and other animals have demonstrated that the accuracy in estimating the duration of a time interval decays gradually with the interval duration itself. More specifically, the variability of the response is proportional to the interval duration (Rakitin et al., 1998; Ivry \& Hazeltine, 1995). For instance, in interval timing the response distributions appear to be scale-invariant: distributions corresponding to different interval durations overlap when linearly scaled (Roberts, 1981; Smith, 1968). Furthermore, animal literature suggests that in conditioning paradigms, the number of trials needed for an animal to learn the association between conditioned and unconditioned stimuli scales with a ratio of the reinforcement latency and intertrial interval (Gallistel \& Gibbon, 2000), indicating again the scale-invariance in the animals’ behavior (Balsam \& Gallistel, 2009; Shankar \& Howard, 2012; Gibbon, 1977).

In addition to timing, scale-invariance has been argued to be one of the key properties of memory. Gradual decay of memory without a characteristic scale has been observed in a number of behavioral experiments (Anderson \& Schooler, 1991; Chater \& Brown, 2008; Wixted \& Ebbesen, 1991) and it was often refereed to as power-law of forgetting. For example, Donkin and Nosofsky (2012) reported that in item recognition task the strength of the memory was decaying as a power law function of the lag between studied items and a test probe. It has been argued that scale-invariance should be thought of as universal law of cognition (Chater \& Brown, 1999). A number of cognitive models have been
constructed to account for these properties (Brown, Neath, \& Chater, 2007; Howard, Shankar, Aue, \& Criss, 2015; Donkin \& Nosofsky, 2012).

Neural mechanisms that could support the scale-invariance of time and memory are still unclear. It has been argued that working memory is represented with persistent neural activity observed in different areas of the prefrontal cortex (PFC) during for instance a delayed match to sample task (P. S. Goldman-Rakic, 1991; P. Goldman-Rakic, 1995). Even though such persistent activity can account well for the demands of a particular task, it is not clear how it could account for a gradual decay of the memory representation. More recent studies have found that in some behavioral tasks a subset of neurons activates sequentially, tiling the task relevant interval, typically lasting for several seconds (see e.g. Pastalkova, Itskov, Amarasingham, and Buzsaki (2008); MacDonald, Lepage, Eden, and Eichenbaum (2011) for the first reports of such activity). These neurons that fire sequentially, each during a circumscribed period of time, are called time cells (Howard \& Eichenbaum, 2015; Eichenbaum, 2013). It has been argued that time cells play important role in timing and memory (MacDonald, Fortin, Sakata, \& Meck, 2014; Howard et al., 2014; Eichenbaum, 2014).

Time cells provide a direct readout of when the delay interval has started: there is no need for population decoding in a classical sense (Murray et al., 2016; Stokes, 2015). This is because time cells activate sequentially effectively providing temporal basis functions and constituting an internal timeline. As we will discuss later, this timeline is compressed, such that the temporal resolution gradually decays with the elapsed time, just as expected given the behavioral findings on memory and timing we mentioned above. It is unclear, however, what are the neural mechanisms that could give rise to such a compressed timeline.

Here we utilize a computational model for compressed scale-invariant dynamical memory representation introduced in Shankar and Howard (2012). We compare specific predictions of the model with the neural recordings of sequentially activated time cells. The model provides a unique solution to constructing a scale-invariant memory representation (Shankar, 2015). The model has been used to account for results of various timing and memory experiments including judgment of recency and serial scanning in long and short term memory (Howard, Shankar, Aue, \& Criss, 2015). Here


Figure 1: Examples of sequentially activated neurons from tetrode recordings in rat hippocampus (plot a.) and PFC (plot b.). On both plots each row on the heatplot corresponds to a single cell and displays normalized firing rate averaged across trials during a delay interval of a behavioral experiment. Red corresponds to high firing rate, while blue corresponds to low firing rate. The cells are sorted with respect to their peak time. Despite the fact that both recordings are done during a rather different behavioral experiment, they both show similar qualitative properties. In particular we point to two features related to the temporal accuracy: 1) time fields later in the delay are more broad than time fields earlier in the delay (the central ridge is widening as the peak moves to the right); 2) peak times of the time cells are not evenly distributed across the delay, with later time periods represented by fewer cells than early time periods (this is apparent from the curvature of the central ridge; a uniform distribution of time fields would manifest as a straight line). a. Hippocampal CA1 neurons recorded during object-delay-odor sequence task (reprint from MacDonald et al. (2011)). In order to obtain a reward the animals had to memorize the identity of the stimulus during the delay interval and match it to the appropriate odor. b. PFC neurons recorded during a temporal discrimination task (reprint from Tiganj et al. (2016), original data first reported in (Kim et al., 2013)). In order to obtain a reward the animals had to estimate whether the presented delay interval was larger than some baseline duration and make a left or right turn accordingly following the delay.
we focus on the neural side, looking into specific predictions about individual neural activity that can be derived from the model.

## Sequential activation as a neural correlate of timing and memory

Starting with a report from Pastalkova et al. (2008), a number of studies have reported sequential neural activation in different timing and memory tasks from different brain regions: hippocampus (MacDonald et al., 2011; Salz et al., 2016; Gill, Mizumori, \& Smith, 2011; Kraus, Robinson, White, Eichenbaum, \& Hasselmo, 2013; MacDonald, Carrow, Place, \& Eichenbaum, 2013; Modi, Ashesh, \& Bhalla, 2014; Naya \& Suzuki, 2011), PFC (Tiganj et al., 2016) and the striatum (Mello, Soares, \& Paton, 2015; Adler et al., 2012) in a variety of behavioral tasks. This activity has been hypothesized to be a neural basis for representation of memory and elapsed time in a gradually decaying fashion (Howard, Shankar, Aue, \& Criss, 2015; Howard \& Eichenbaum, 2015; Eichenbaum, 2014, 2013). The studies were done on different animals, including rats (MacDonald et al., 2011; Salz et al., 2016; Gill et al., 2011; Kraus et al., 2013; MacDonald et al., 2013; Pastalkova et al., 2008; Mello et al., 2015; Tiganj et al., 2016), mice (Modi et al., 2014) and monkeys (Naya \& Suzuki, 2011; Adler et al., 2012). Even though the majority of studies used tetrode recordings, Modi et al. (2014) used two-photon calcium imaging minimizing the probability that the results were observed due to some sort of recording artifact. Most of the studies were done on animals that were allowed to move, but
some were done on head-fixed animals (MacDonald et al., 2013; Modi et al., 2014; Naya \& Suzuki, 2011; Adler et al., 2012) confirming that the results were not coming from position-related artifacts.

It is worth noting that sequentially activated time cells were observed in these studies despite the different cognitive demands on the animals, which included temporal discrimination (e.g. Tiganj et al. (2016)) or memory demands (e.g. Salz et al. (2016); MacDonald et al. (2011)). The duration of the intervals where such cells were measured was ranging from a couple of seconds up to 60 s (Mello et al., 2015).

Several studies have observed decreasing temporal accuracy as a function of delay, due to spread in time field width (Howard et al., 2014; MacDonald et al., 2011; Mello et al., 2015; Adler et al., 2012; Kraus et al., 2013; Salz et al., 2016; Tiganj et al., 2016) and/or due to a non-uniform distribution of time fields (Kraus et al., 2013; Salz et al., 2016; Mello et al., 2015; Tiganj et al., 2016). Two examples of neural representation with decreasing temporal accuracy are provided in Figure 1.

## Computational model for compressed scale-invariant dynamical memory representation

The computational model reviewed here was initially introduced in (Shankar \& Howard, 2012). It consists of a twolayer feedforward neural network with analytically derived weights. Here we briefly describe the model and then focus on its predictions regarding neural activity. Notice that below
we define the model as a model of memory, as it was initially introduced in (Shankar \& Howard, 2012). Its application in timing is restricted to the stimulus that initiates the delay interval, as the only stimulus that needs to be remembered.

We first define an input vector $\mathbf{f}$ consisting of $N$ elements such that each of its elements corresponds to a unique stimulus. Thus observing for example stimulus $A$ makes an element in $\mathbf{f}$ that corresponds to stimulus $A, f_{A}$, equal to one for the time $A$ is presented and zero otherwise. Each element of the input vector $\mathbf{f}$ has a two-layer dynamical compressed memory representation. The first layer of the network implements an approximation of an integral transform of the input (Laplace transform, but as a function of a real rather than a complex variable). This means that nodes in the first layer, $\mathbf{F}\left(t,{ }^{*}\right)$, act as leaky integrators (first order low-pass filters) with a spectrum of time constant defined with $k /{ }^{*}$, where $k$ is positive integer (Figure 2):

$$
\begin{equation*}
\frac{\mathbf{F}(t, \stackrel{*}{t})}{d t}=-\stackrel{{ }_{*}^{*}}{t} \mathbf{F}(t, \stackrel{*}{t})+\mathbf{f}(t) \tag{1}
\end{equation*}
$$

Leaky integrators project to the second layer, $\tilde{\mathbf{f}}$, through fixed weights that implement an approximation of the inverse of the transform by applying a $k^{\text {th }}$ order derivative with respect to $k /^{*}$, denoted as $\mathbf{F}^{(k)}(t, \stackrel{*}{t})$ (the inverse is derived based on Post's inversion formula (Post, 1930), see Shankar and Howard $(2012,2013)$ for further details on the derivation):

$$
\begin{equation*}
\tilde{\mathbf{f}} t, \stackrel{*}{t})=C_{k}\left(\frac{k}{\stackrel{k}{t}}\right)^{k+1} \mathbf{F}^{(k)}(t, \stackrel{*}{t}) \tag{2}
\end{equation*}
$$

where $C_{k}$ is a constant that depends only on $k$. The cells in the second layer constitute a dynamical memory representation of the input signal. To understand the properties of the memory representation we consider an impulse response of a cell in $\tilde{\mathbf{f}}$. For $f_{A}(\tau)=\delta(\tau=0)$ the corresponding activation of the cells in the second layer is:

$$
\begin{equation*}
\tilde{f}_{A}(t, \stackrel{*}{t})=C_{k} \frac{1}{\underset{*}{*}}\binom{t}{\frac{\tilde{*}}{t}}^{k} e^{-k_{\dot{*}}^{t}}, \tag{3}
\end{equation*}
$$

where $C_{k}$ here is a different constant that depends only on $k$. The activity of each node in $\tilde{f}_{A}\left(t,{ }_{t}^{*}\right)$ is the product of an increasing power term $\left(\frac{t}{t^{*}}\right)^{k}$ and a decreasing exponential term $e^{-k_{\text {仡 }}^{t}}$. Consequently, each node in $\tilde{f}_{A}(t, \stackrel{*}{t})$ has a peak that corresponds to the ${ }_{t}^{*}$ value of that node: $\frac{d \tilde{f}_{A}(t, t)}{d t}=0 \Rightarrow t=\stackrel{*}{\tilde{f}}$. Thus, following a transient input, cells in $\tilde{f}_{A}$ activate sequentially in time constituting a dynamical memory representation of the input $A$ (Figure 3).

This memory representation has perfect accuracy in the limit when $k \rightarrow \infty$. In a realistic biological or artificial neural network, where $k$ is finite and $\stackrel{*}{t}$ is a discrete variable sup-


Figure 2: Constructing a scale-invariant compressed memory representation through an integral transform and its inverse. A transient input stimulus $\mathbf{f}(t)$ (top row) is presented twice and feeds into a layer of leaky integrators $\mathbf{F}\left(t,{ }_{t}^{*}\right)$ with a spectrum of time constants $\stackrel{*}{t}$ constituting a discrete approximation of an integral transform (middle row). The transform is denoted as $\mathbf{L}$ since it is equivalent to the real part of the Laplace transform. Only three nodes in $\mathbf{F}(t, \stackrel{*}{t})$ are shown. Each leaky integrator is characterized with its time constant, ${ }_{t}^{*}$. F projects onto $\tilde{\mathbf{f}}(t, \stackrel{*}{t})$ through a set of weights defined with the operator denoted as $\mathbf{L}_{k}^{-1}$ which implements an approximation of the inverse of the Laplace transform. Nodes in $\tilde{\mathbf{f}}(t, \stackrel{*}{t})$ activate sequentially following the stimulus presentation creating a memory representation. The width of the activation of each node scales with the peak time determined by the corresponding ${ }_{t}^{*}$, making the memory scale-invariant. Logarithmic spacing of the $\stackrel{*}{t}$ assures that the memory representation is compressed.
ported with a limited number of nodes, the memory representation becomes an approximation of the past. The approximation is scale-invariant (Figure 4) since the width of the activation of each node scales with the peak time (this is scale-invariant since rescaling the temporal axis rescales the width of the activation by the same amount). In other words, the accuracy of the memory representation decreases with the elapse of time since the stimulus presentation. With appropriately distributed ${ }^{*}$ the representation can be made logarithmically compressed.

To establish biological plausibility of the model we have shown that leaky integrators with a spectrum of time constants are biologically realistic (Tiganj, Hasselmo, \& Howard, 2015; Tiganj, Shankar, \& Howard, 2013). In addition, taking derivatives with respect to $k / t^{*}$ amounts to lateral inhibition, making it biologically plausible as well (Howard et al., 2014). To implement the derivative it is required that each neurons
a

b


Figure 3: Activity of the cells in the compressed memory representation generated by the model. Analogous to the heatmaps in Figure 1, each row corresponds to a single cell and displays its normalized activity across time. The cells are sorted with respect to the peak time defined by their value of $t$. The two features observed in Figure 1 are fully captured by the model: the time fields later in the delay were more broad than the time fields earlier in the delay and the density of time fields decreased as a function of time ( $\stackrel{*}{t}$ was logartihmically spaced). This illustrates that the model can indeed account for the firing dynamics of the sequentially activated time cells that form a compressed representation of time. The two plots, $\mathbf{a}$ and $\mathbf{b}$, show the activity of the cell ensemble for two different values of parameter $k$. Increasing $k$ makes the firing fields more narrow and the memory representation more precise. Notice that, from the biological perspective, larger $k$ is more difficult to obtain, since it requires higher order of derivative with respect to $k /{ }^{*}$. This requires broader connectivity between the two layers.
of the first layer only projects to the $k$ neighboring neurons of the second layer. The connectivity pattern is the same across the entire projection, since it always implements a derivative with respect to $k /{ }^{*}$. In addition, qualitative alignment of the model with the sequential neural activity as shown in Figure 3 further supports its biological plausibility.

## Discussion

We reviewed the predictions from a computational model for compressed scale-invariant memory representation and compared them to the results from recently-published neural recordings. The model maintains a dynamical representation of the recent past through a set of sequentially activated neurons. Such sequential activation appears qualitatively similar to the data published in multiple studies over the past several years including different regions of the brain including the hippocampus, PFC and striatum.

Several of the studies align with the model exhibiting compressed memory representation. In particular, the width of the time fields increased with the peak time and more cells had time fields earlier than later in the delay interval (notice the common trend in the plots in Figure 1 and Figure 3). These findings suggest that the model can indeed account for the neural representation of the elapsed time.

The model makes specific prediction on the scale-
invariance of the memory representation which was inspired by the behavioral experiments on timing and memory. Existing neural data were thus far not sufficient to explicitly test that prediction. However, the qualitative observations made here are consistent with the scale-invaraince prediction, but they are not sufficient to quantitatively verify it.

In addition to the model described here, several other computational models predict the qualitative properties found in the data. The common aspect of most of such models is the functional form that gives rise to time fields: as in the model described here, the activity increase is governed by a powerlaw and then later attenuated by a damping exponential. In particular, Grossberg and Schmajuk (1989); De Vries and Principe (1992); Machado (1997) propose different mechanistic solutions for achieving such form. However, unlike in the model described here, rescaling the time axis in these models would change the functional form of the representation. Others (for instance Tank and Hopfield (1987); Ludvig, Sutton, and Kehoe (2012)) directly used the functional form that provides spreading temporal basis functions as seen here.

Experimental data allowed us thus far to verify some of the predictions computational models make regarding the compressed representation of time. However, the model described here makes specific predictions regarding how memory is maintained in general. Here we assumed that the stimulus that marks the onset of the delay interval is the only one that has the memory representation. The model is designed to capture a variety of stimuli and maintain an independent compressed memory representation for each of them. In fact, associations between the independent representations allowed us to test the model on a variety of memory tasks (Howard, Shankar, Aue, \& Criss, 2015). It is to be tested whether the neural representation indeed supports such independent, stimulus specific compressed memory representations (see Tiganj, Cromer, Roy, Miller, and Howard (2017) for recent evidence of this).

Maintaining temporal information through sequential activation has a critical computational property in that it provides a direct readout of the elapsed time. Notice that cells in the first layer of the model (leaky integrators) contain the same amount of temporal information as the cells in the second layer (sequentially activated neurons). Thus one could apply population decoding techniques and extract the temporal information from the first layer directly. In fact, this is exactly what the inverse transform is doing. However, instead of training a classifier, which would be a common decoding procedure, it provides a simple form of linear readout using a mechanism analogous to lateral inhibition, which is known to exist in the nervous system. An additional advantage of having such a mechanism is that it provides access to the realvalue Laplace domain, where computations that are otherwise hard to achieve in a neural network become straightforward. These in particular include addition and subtraction of probability distributions as well as temporal translation (Howard, Shankar, \& Tiganj, 2015; Shankar, Singh, \& Howard, 2016).
a

b

c

d


Figure 4: Illustration of scale-invariance in the compressed memory representation generated by the model. Scaling the number of cells and the temporal duration by the same factor results in identical memory representation (plots a. to d. appear identical despite the fact that both x and y axes are rescaled on each plot). This property follows from Equations (3) since ${ }^{*}$ and $t$ appear only as a ratio (except for the scaling factor in front that does not influence the functional form). Scale-invariance is consistent with behavioral experiments, but it remains unclear whether neural data exhibits this property as well, even though the results shown in Figure 1 are consistent with scale-invariance.

## Conclusion

We showed that a computational model for constructing compressed dynamical representations of the recent past aligns well with recent neural data showing sequential neural activation. The sequential activation constitutes a compressed supported timeline, providing a mechanism for representing the elapse of time and potentially a mechanism for maintaining a dynamical memory representation.

## Acknowledgments

This work was supported by NIBIB R01EB022864, NIMH R01MH112169 and MURI N00014-16-1-2832.

## References

Adler, A., Katabi, S., Finkes, I., Israel, Z., Prut, Y., \& Bergman, H. (2012). Temporal convergence of dynamic cell assemblies in the striato-pallidal network. Journal of Neuroscience, 32(7), 2473-84. doi: 10.1523/JNEUROSCI.4830-11.2012

Anderson, J., \& Schooler, L. (1991). Reflections of the environment in memory. Psychological science, 2(6), 396408.

Balsam, P. D., \& Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. Trends in Neuroscience, 32(2), 73-78.
Brown, G. D. A., Neath, I., \& Chater, N. (2007). A temporal ratio model of memory. Psychological Review, 114(3), 539-76.
Chater, N., \& Brown, G. D. (1999). Scale-invariance as a unifying psychological principle. Cognition, 69(3), B17B24.
Chater, N., \& Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. Cognitive Science, 32(1), 36-67. doi: 10.1080/03640210701801941
De Vries, B., \& Principe, J. C. (1992). The gamma modela new neural model for temporal processing. Neural networks, 5(4), 565-576.
Donkin, C., \& Nosofsky, R. M. (2012). A powerlaw model of psychological memory strength in short-
and long-term recognition. Psychological Science. doi: 10.1177/0956797611430961

Eichenbaum, H. (2013). Memory on time. Trends in Cognitive Sciences, 17(2), 81-8. doi: 10.1016/j.tics.2012.12.007
Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. Nature Reviews Neuroscience, 15(11), 732-744.
Gallistel, C. R., \& Gibbon, J. (2000). Time, rate, and conditioning. Psychological Review, 107(2), 289-344.
Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. Psychological Review, 84(3), 279-325.
Gill, P. R., Mizumori, S. J. Y., \& Smith, D. M. (2011). Hippocampal episode fields develop with learning. Hippocampus, 21(11), 1240-9. doi: 10.1002/hipo. 20832
Goldman-Rakic, P. (1995). Cellular basis of working memory. Neuron, 14, 477-85.
Goldman-Rakic, P. S. (1991). Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. Progress in brain research, 85, 325-336.
Grossberg, S., \& Schmajuk, N. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. Neural Networks, 2(2), 79-102.
Howard, M. W., \& Eichenbaum, H. (2015). Time and space in the hippocampus. Brain research, 1621, 345-354.
Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., \& Eichenbaum, H. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. Journal of Neuroscience, 34(13), 4692-707. doi: 10.1523/JNEUROSCI.5808-12.2014

Howard, M. W., Shankar, K. H., Aue, W., \& Criss, A. H. (2015). A distributed representation of internal time. Psychological Review, 122(1), 24-53. doi: 10.1037/a0037840
Howard, M. W., Shankar, K. H., \& Tiganj, Z. (2015). Efficient neural computation in the laplace domain. In Proceedings of the 2015th international conference on cognitive computation: Integrating neural and symbolic approaches-volume 1583 (pp. 61-68).
Ivry, R. B., \& Hazeltine, R. E. (1995, Feb). Perception and production of temporal intervals across a range of dura-
tions: evidence for a common timing mechanism. J Exp Psychol Hum Percept Perform, 21(1), 3-18.
Kim, J., Ghim, J.-W., Lee, J. H., \& Jung, M. W. (2013). Neural correlates of interval timing in rodent prefrontal cortex. Journal of Neuroscience, 33(34), 13834-47. doi: 10.1523/JNEUROSCI.1443-13.2013

Kraus, B. J., Robinson, R. J., 2nd, White, J. A., Eichenbaum, H., \& Hasselmo, M. E. (2013). Hippocampal "time cells": time versus path integration. Neuron, 78(6), 1090-101. doi: 10.1016/j.neuron.2013.04.015

Ludvig, E. A., Sutton, R. S., \& Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. Learning \& Behavior, 40(3), 305-319.
MacDonald, C. J., Carrow, S., Place, R., \& Eichenbaum, H. (2013). Distinct hippocampal time cell sequences represent odor memories immobilized rats. Journal of Neuroscience, 33(36), 14607-14616.
MacDonald, C. J., Fortin, N. J., Sakata, S., \& Meck, W. H. (2014). Retrospective and prospective views on the role of the hippocampus in interval timing and memory for elapsed time. Timing \& Time Perception, 2(1), 51-61.
MacDonald, C. J., Lepage, K. Q., Eden, U. T., \& Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. Neuron, 71, 737-749.
Machado, A. (1997). Learning the temporal dynamics of behavior. Psychological review, 104(2), 241.
Mello, G. B. M., Soares, S., \& Paton, J. J. (2015). A Scalable Population Code for Time in the Striatum. Current Biology, 25(9), 1113-1122.
Modi, N. M., Ashesh, D. K., \& Bhalla, S. U. (2014). CA1 cell activity sequences emerge after reorganization of network correlation structure during associative learning. eLife, 3(0). Retrieved from http://dx.doi.org/10.7554/eLife. 01982 doi: 10.7554/eLife. 01982

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., \& Wang, X.-J. (2016). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proceedings of the National Academy of Sciences, 201619449.
Naya, Y., \& Suzuki, W. (2011). Integrating what and when across the primate medial temporal lobe. Science, 333(6043), 773-776.
Pastalkova, E., Itskov, V., Amarasingham, A., \& Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. Science, 321(5894), 1322-7.
Post, E. (1930). Generalized differentiation. Transactions of the American Mathematical Society, 32, 723-781.
Rakitin, B. C., Gibbon, J., Penny, T. B., Malapani, C., Hinton, S. C., \& Meck, W. H. (1998). Scalar expectancy theory and peak-interval timing in humans. Journal of Experimental Psychology: Animal Behavior Processes, 24, 15-33.
Roberts, S. (1981). Isolation of an internal clock. Journal of Experimental Psychology: Animal Behavior Processes, 7, 242-268.

Salz, D. M., Tiganj, Z., Khasnabish, S., Kohley, A., Sheehan, D., Howard, M. W., \& Eichenbaum, H. (2016). Time cells in hippocampal area ca3. The Journal of Neuroscience, 36(28), 7476-7484.
Shankar, K. H. (2015). Generic construction of scaleinvariantly coarse grained memory. In Australasian conference on artificial life and computational intelligence (pp. 175-184).
Shankar, K. H., \& Howard, M. W. (2012). A scale-invariant representation of time. Neural Computation, 24, 134-193.
Shankar, K. H., \& Howard, M. W. (2013). Optimally fuzzy scale-free memory. Journal of Machine Learning Research, 14, 3753-3780.
Shankar, K. H., Singh, I., \& Howard, M. W. (2016). Neural mechanism to simulate a scale-invariant future. Neural Computation, 28(12).
Smith, M. C. (1968). CS-US interval and US intensity in classical conditioning of rabbit's nictitating membrane response. Journal of Comparative and Physiological Psychology, 3, 679-687.
Stokes, M. G. (2015). 'activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends in Cognitive Sciences, 19(7), 394-405.
Tank, D., \& Hopfield, J. (1987). Neural computation by concentrating information in time. Proceedings of the National Academy of Sciences, 84(7), 1896-1900.
Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., \& Howard, M. W. (2017). Compressed timeline of recent experience in monkey lpfc. bioRxiv, 126219.
Tiganj, Z., Hasselmo, M. E., \& Howard, M. W. (2015). A simple biophysically plausible model for long time constants in single neurons. Hippocampus, 25(1), 27-37.
Tiganj, Z., Kim, J., Jung, M. W., \& Howard, M. W. (2016). Sequential firing codes for time in rodent mPFC. Cerebral Cortex(1-9). doi: 10.1093/cercor/bhw336
Tiganj, Z., Shankar, K. H., \& Howard, M. W. (2013). Encoding the laplace transform of stimulus history using mechanisms for persistent firing. BMC Neuroscience, 14(Suppl 1), P356.

Wixted, J. T., \& Ebbesen, E. B. (1991). On the form of forgetting. Psychological Science, 2, 409-415.

# Picturing time: Children's preferences for visual representations of events 

Katharine A. Tillman ${ }^{1}$, Eren Fukuda ${ }^{1}$, \& David Barner ${ }^{1,2}$<br>University of California, San Diego<br>Departments of Psychology ${ }^{1}$ and Linguistics ${ }^{2}$<br>${ }^{1}$ katillman@ucsd.edu; ${ }^{2}$ efukuda@ucsd.edu; ${ }^{3}$ barner@ucsd.edu


#### Abstract

English-speaking adults recruit a left-to-right "mental timeline" (MTL) when thinking about time. The origins of the MTL are debated, with some arguing that it is a cultural construct and others arguing that it is rooted in innate associations between time and space. Here we ask whether preschoolers, with limited experience with cultural practices thought to shape the MTL, prefer conventional linear representations of temporal events. English-speaking preschoolers and adults were told stories and asked to choose which of two visual representations best illustrated the story. As expected, adults overwhelmingly preferred images that were linearly ordered from left-to-right. Five-year-olds also preferred left-to-right to right-to-left series, but were equally likely choose left-to-right and top-to-bottom. By contrast, 3-year-olds chose at random, apparently insensitive to the spatial ordering of event-denoting images. These results suggest that attention to the ordinal structure of visual representations of time increases across early childhood, and that adults' preference for horizontal space-time mappings results from increased cultural conditioning.


Keywords: time; space; mental timeline; events; abstract concepts

## 1. Introduction

Time and space are deeply interwoven in human experience and culture. For example, diverse societies use spatial tools to depict, measure, and track time; languages often use the same words to refer to both time and space (e.g., long and short); and readers repeatedly experience temporal narratives unfolding in a particular spatial direction across the page. Behavioral and neuroscientific studies suggest that adults have implicit linear associations between specific locations in time and positions in space (for a review, Bonato et al., 2012). The nature of the relationship between this "mental timeline" (MTL) and cultural practices that link time and space is debated. On the one hand, systematic cross-cultural differences in the direction of the MTL (e.g., Boroditsky, 2011; Bergen \& Lau, 2012) suggest that it is learned. On the other hand, evidence of space-time mappings in infants (e.g., de Hevia et al., 2014; Lourenco \& Longo, 2010; Srinivasan \& Carey, 2010), and the ubiquity of spatial artifacts and metaphors across cultures (Haspelmath, 1997) suggest that some form of MTL may be intrinsic to human cognition. Do cultural tools linking time and space create mental associations across domains, or do they simply capitalize on a low-level,
biological predisposition to think about time spatially? Understanding the development of space-time associations in children who cannot yet read or use spatial artifacts for time could shed light on this question. Here, we test whether 3- to 5-year-old preschoolers show adult-like preferences for linear representations of events.

Cross-cultural comparisons involving adults and school-aged children have revealed reliable differences in the orientation and direction of ordinal space-time mappings (e.g., Tversky, Kugelmass, \& Winter, 1991). The left-to-right (LR) mental timeline is robust in speakers of English and many other languages using an LR orthography, but speakers of languages that are written from right-to-left (RL), often construe of time in an RL line (e.g., Ouellet et al. 2010; Tversky, Kugelmass, \& Winter, 1991). Vertical associations between time and space have also been found in speakers of Chinese, which can be written top-to-bottom (TB) and also contains vertical time-space metaphors (e.g., Boroditsky, 2011). Many cultural and environmental sources of the MTL (and the analogous "mental number-line," MNL) have been posited. These include: reading/writing direction, space-time metaphor in language, exposure to artifacts such as calendars, counting-related practices, early visual experiences, and simply growing up in a community with existing space-time associations.

In contrast to purely cultural accounts, some theories contend that we have an innate predisposition to associate space and time. One such theory posits that space, time, and number rely on a single system for magnitude representation (Walsh, 2003). Consistent with the idea that language and social cues are not the sole sources of the MTL, infants and even neonates appear to make implicit associations between duration and spatial length (e.g., de Hevia et al., 2014; Srinivasan \& Carey, 2010). Going beyond a general magnitude account, others have argued that the ordinal structure of the MTL/MNL also has a neurophysiological and evolutionary basis, and may be LR by default (Chatterjee, 2001; Rugani et al., 2015).

Importantly, cross-cultural differences in the direction of the adult MTL indicate that, even if innate ordinal space-time mappings exist, they can be modulated by reading-writing behavior or other types of cultural conditioning. It is therefore difficult to pinpoint the developmental origins of the MTL, or to disentangle its potential biological or environmental causes, in adults populations with many relevant types of cultural knowledge.

Here, we explore when and how time-space mappings develop in a population whose exposure to cultural input is more limited: children. Because formal instruction in reading/writing and spatial tools for time often begins in the early school years, evidence of linear space-time associations in younger children might suggest these abilities are not critical to the formation of the MTL. The purpose of the current study is to test whether preschoolers already have a preference for visual representations of events depicted in conventional ordered lines. If so, this might suggest that the tendency to form mental mappings between time and space is not entirely culturally constructed.

Several prior studies argue that directional space-number associations are present in preliterate preschoolers (see Nuerk et al., 2015). For instance, English-speaking preschoolers spontaneously count objects from LR, while Hebrew speakers count from RL. Biases on purely spatial tasks such as line-bisection have also been observed in preschoolers. These effects are generally stronger in older children and adults. To the extent that both the MNL and MTL draw on similar spatial representations, we might expect to observe similarly early biases toward LR representations of time in English-speaking preschoolers.

Relatively few studies have investigated space-time mappings in preschoolers. Timeline tasks indicate that 4-year-olds can place events on an LR line more accurately than chance, but that this ability improves considerably over the next $3+$ years (e.g., Hudson \& Mayhew, 2011; Tillman et al., 2017). Importantly, tasks in which a single type of timeline is provided for children to use cannot address whether they privilege particular spatial orientations or directions. However, even without a template, a majority of school-aged children place stickers representing events in ordered lines with a culture-specific direction (e.g., LR for English-speakers; Tversky, Kugelmass, \& Winter, 1991).

In contrast to older children, preschoolers rarely place event-denoting stickers in lines spontaneously, and those who do so show a much more modest, if any, bias toward LR lines (Tillman, Tulagan, \& Barner, 2015). Similarly, older children, but not preschoolers, produce spatial representations of single events in which the agent, object, and recipient are linearly ordered in a culturally-conventional direction (Dobel, Diesendrunk, \& Bolte, 2007). Together, these studies suggest that the automatic deployment and the direction-specificity of the MTL develop slowly in early childhood, and may rely on literacy and/or formal schooling to become fully engrained.

Critically, tasks like those discussed above either require children to use sophisticated artifacts or to create visual representations of time, and therefore may require
significant visuospatial, motor, and working-memory skills. For instance, the sticker-placement task requires an ability to use non-iconic stickers symbolically, sufficient motor control to put them in specific spatial locations, and memory of what previously-used stickers represent. It is therefore possible that the difficulty of these tasks could have masked existing associations between time and space in preschoolers. To address this concern, the present study employs a forced-choice task with minimal response demands to test whether English-speaking preschoolers prefer conventional linear representations of time. Preschoolers were told brief stories describing three-step event sequences, given a choice between two spatial depictions of each story, and asked which of the two was better. In Experiment 1, to test whether children have direction preferences, they chose between (conventional) LR, RL, TB, and bottom-to-top (BT) representations of events. In Experiment 2, to test whether children were sensitive to the ordinality of the images, they chose between ordered and unordered sequences.

## 2. Experiment 1

### 2.1 Methods

2.1.1 Participants. Participants included 62 3-year-old children $(M$ age $=3 ; 6)$, 605 -year-old children $(M$ age $=$ $5 ; 5)$, and 85 adult controls. They were pseudo-randomly assigned to one of 3 conditions: Horizontal ( $\mathrm{n}=213 \mathrm{YO} ; 21$ $5 \mathrm{YO} ; 29$ adults), Vertical ( $\mathrm{n}=203 \mathrm{YO} ; 205 \mathrm{YO} ; 29$ adults), and Mixed ( $\mathrm{n}=213 \mathrm{YO} ; 205 \mathrm{YO} ; 27$ adults). Children were recruited from museums and daycares in the San Diego, CA, area, and adults were workers on Amazon Mechanical Turk. All participants spoke English as their primary language, and none spoke a secondary language with non-LR orthography. Adults and parents of children gave informed consent to participate. Children were awarded a small prize, and adults were compensated $\$ 1$. An additional 9 children were tested but excluded from analysis because either English was not their primary language ( $\mathrm{n}=3$ ), they spoke a second language with a non-LR orthography ( $\mathrm{n}=2$ ), they failed to complete the task $(\mathrm{n}=2)$, developmental delay ( $\mathrm{n}=$ $1)$, or clerical error $(\mathrm{n}=1)$. Five adults were excluded from analysis due to speaking a language with non-LR orthography $(\mathrm{n}=2)$ and lack of attention to the task, as indexed by failing a "catch" trial $(\mathrm{n}=3)$.
2.1.2 Procedure. On each of 8 trials, children heard a story involving 3 steps (see Table 1). The experimenter placed two cards on the table in front of the child, and asked: "Which card shows that story? Which one is better?"

Table 1. Event stories.

| Event | First. | Then.. | And.. |
| :--- | :--- | :--- | :--- |
| Egg | there was an egg | the egg cracked | a baby chick came out! |
| Ice Cream | there was an ice-cream | it started melting | and it was all gone! |
| Drawing | someone started drawing a stem | they added blue petals | it was a flower! |
| Caterpillar | there was a caterpillar | it made a cocoon | it turned into a butterfly! |
| Baby | a baby was born | he started growing | he was a big kid! |
| Apple | there was an apple | someone took a bite | they ate it all! |
| Rose | a rose started growing | it opened | it got big and pink! |
| Watermelon there was a watermelon | we cut it all up | everyone ate it! |  |

After the child pointed to their choice, the cards were removed, and the next trial began.

Participants in the Horizontal condition always chose between one card with three pictures depicting the story in order from left-to-right (LR; see Table 1 and Fig 1A) and another with the same 3 pictures ordered from right-to-left (Fig. 1B). Participants in the Vertical condition chose between cards with images arranged from top-to-bottom (Fig 1C) vs. bottom-to-top (Fig 1D), and, in the Mixed condition, between LR and TB lines (Fig. 1A vs. 1C). The two cards were placed side-by-side in the Vertical and Mixed conditions, but were positioned one above the other in the Horizontal condition. Every child heard the Egg story first. Half the children heard the remaining stories in the order listed in Table 1, and half heard them in the reverse order. The positioning of the two cards was counterbalanced across subjects and items. Adults read the stories on a computer, and clicked the image they thought was better. Data analysis was done using R and the lme 4 package.

### 2.2 Results

2.2.1 Horizontal condition. Participants in the Horizontal condition chose between LR and RL sets of images (e.g., Fig. 1A vs. 1B). To test for direction preferences, we calculated the percentage of trials on which each subject chose the LR card. As expected, virtually all adults ( $\mathrm{n}=28$ of 29) chose the LR card on every trial (Fig 2A). In contrast, the median percentage of LR choices for 5-year-olds was lower, at $75 \%$, and these children were less consistent across trials than were adults (see Fig 2A). The median percentage of LR picks by 3 -year-olds was $50 \%$. Exact Wilcoxon signed-rank tests confirmed that 3-year-olds' performance was consistent with random guessing ( $\mathrm{V}=24.5, p=0.5$ ), but five-year-olds selected the LR card significantly more often than chance $(\mathrm{V}=126.5, p=0.02)$.


Figure 1: Example picture cards. The three images on each card depict the three stages in the Egg story (see Table 1). Cards used in Experiment 1: (A) LR, left-to-right, (B) RL, right-to-left, (C) TB, top-to-bottom, and (D) BT, bottom-to-top. Additional cards used in Experiment 2: (E) Scrambled Horizontal and (F) Scrambled Vertical.
2.2.2 Vertical condition. Participants in the Vertical condition chose between TB and BT images (Fig 1C vs. 1D). As shown in Figure 2B, $90 \%$ of adults ( $\mathrm{n}=26$ of 29) chose the TB card on every trial. A subset of 5-year-olds

## A. Horizontal: LR vs. RL



Figure 2: Direction and orientation preferences. Histograms showing the number of subjects who picked the more conventional representation of time at each degree of consistency.
( $\mathrm{n}=9$ of 20) also showed a strong preference for TB cards, bringing the group median to $63.5 \%$, significantly higher than chance (Exact Wilcoxon signed-rank test, $\mathrm{V}=111, p=$ 0.02 ). The median response among 3 -year-olds was $50 \%$ TB, consistent with random guessing (Exact Wilcoxon signed-rank test, $\mathrm{V}=54, p=0.3$ ).

We next asked whether children's directional biases were stronger along one spatial axis than the other. In other words, did children have a significantly stronger preference for LR in the Horizontal condition than they had for TB in the Vertical condition? We used mixed-effects logistic regression to model the likelihood of a "conventional" choice (i.e., LR in the Horizontal condition; TB in Vertical) as a function of Age Group (3-year-olds vs. 5-year-olds) and Condition (Horizontal vs. Vertical). The model included the interaction of fixed effects as well as a random effect of subjects ${ }^{1}$. Examining this model, we found only a main effect of Age Group ( $\beta=0.65, p=0.01 ; \chi 2(1)=10.1, p=$ 0.002 ). The effect of Condition did not reach significance ( $\beta$ $=0.15, p=0.6 ; \chi 2(1)=0.4, p=0.5)$. Thus, the results

[^211]indicate that children have equally strong (or weak) directional preferences within the horizontal and vertical axes.
2.2.3 Mixed condition. Participants in the Mixed condition chose between (horizontal) LR- and (vertical) TB-ordered images (Fig. 1A vs. 1C). In contrast to their near-perfect consistency in the other conditions, only about half the adult sample ( $\mathrm{n}=12$ of 27 ) chose the LR card on every trial ( $\mathrm{n}=$ 12), resulting in a median response of $87.5 \%$ LR (Fig. 2C). In contrast, the median percentages of LR picks for both 3and 5 -year-olds were $50 \%$, consistent with random guessing (Exact Wilcoxon signed rank tests, $p$ 's $>0.05$ ).

Next, we asked whether children's likelihood of choosing the LR card was impacted by the orientation of the comparison set of images, by fitting a mixed-effects logistic model to data from the Horizontal and Mixed conditions. As predictors, we entered Age Group and Condition, their interaction, and a random effect of subjects. Examining the model, we found significant main effects of both Age Group ( $\beta=0.69, p=0.01 ; \chi 2(1)=9.9, p=0.01)$ and Condition $(\beta$ $=-0.51, p=0.02 ; \chi 2(1)=5.1, p=0.02)$, with no interaction. In other words, when given a choice, children chose LR more often than RL, but not more often than TB.

Together, the results of Experiment 1 suggest that directional linear associations of time emerge between 3 and 5 years of age, and that children's biases within spatial axes develop earlier than biases across axes.

## 3. Experiment 2.

When given choices between two ordinal representations of a story that had different directions, the majority of 3-year-olds in Experiment 1 did not demonstrate a preference. One explanation for this behavior is that, for 3-year-olds, all ordered series of images are equally compelling illustrations of stories. An alternative explanation is that 3 -year-olds simply did not attend to the relative ordering of the images on the cards. Experiment 2 tests this hypothesis.

Rather than choosing between two ordered sets varying in direction, children in Experiment 2 chose between one ordered set (either LR or TB) and one unordered set with the same orientation (horizontal or vertical). If children are sensitive to the ordinal relations among images, we would expect them to choose cards showing ordered temporal sequences (e.g., caterpillar-cocoon-butterfly) more often than cards showing scrambled sequences (e.g., caterpillar-butterfly-cocoon). On the other hand, if 3 -year-olds do not attend to the order of the pictures (in relation to the order of events in the story), we would expect the same pattern of results found in Experiment 1.

### 3.1 Methods.

3.1.1 Participants. Thirty-eight 3-year-olds ( $M$ age $=3 ; 7$ ) were recruited from daycares and museums in the Comox
valley, BC, and San Diego, CA, areas. Nineteen were assigned to the Scrambled Horizontal condition and 19 to the Scrambled Vertical condition. An additional 4 children were excluded because English was not their primary language $(\mathrm{n}=1)$, they spoke a second language with a non-LR orthography $(\mathrm{n}=2)$, and experimenter error $(\mathrm{n}=1)$.
3.1.2 Materials and procedures were identical to those used in the Horizontal and Vertical conditions of Experiment 1, except that each RL card was replaced with a Scrambled Horizontal card (Fig. 1E), and each BT card was replaced with a Scrambled Vertical card (Fig. 1F).

### 3.2 Results and Discussion.

3.2.1 Horizontal Scrambled condition. Figure 3A plots the distribution of children who chose the ordered (LR) card with each level of consistency across trials. The median percentage of LR choices was $50 \%$, again consistent with random guessing (Exact Wilcoxon signed-rank test, $\mathrm{V}=$ $52.5, p=0.3$ ).


Figure 3. Spatial ordinality preferences. Histograms showing the number of 3 -year-olds who picked the ordinal representation of time over an unordered one, at each degree of consistency.

To compare 3-year-olds' performance on the Horizontal Scrambled (Exp. 2) and unscrambled Horizontal (Exp. 1) conditions, we used a mixed-effects logistic model predicting the likelihood of an LR choice as a function of Condition (Horizontal vs. Horizontal Scrambled), with a random effect of subjects. The Condition factor did not improve the fit of the model over a null model $(\beta=0.23, p=$ $0.3 ; \chi 2(1)=1.0, p=0.3)$. Children were no better at choosing the LR card over an unordered sequence than they were at choosing LR over RL or TB in Experiment 1.
3.2.2 Vertical Scrambled condition. Results from the Vertical Scrambled condition are shown in Fig 3B. As in the Horizontal Scrambled condition, most 3-year-olds picked the TB card on $50 \%$ of trials, consistent with chance (Exact Wilcoxon signed-rank test, $\mathrm{V}=44, p=0.4$ ) and the addition of Condition (Vertical vs. Vertical Scrambled) as a factor did not significantly improve the fit of a model predicting children's likelihood of choosing the TB card $(\beta=0.36, p=$ $0.1 ; \chi 2(1)=2.5, p=0.1)$.

Together, the results of Experiment 2 suggest that 3-year-olds are insensitive to the ordinal relationships among images depicting temporal events.

## 4. General Discussion

We explored the development of mental associations between time and space, by asking whether preschoolers prefer visual representations of events that have a conventional linear structure (i.e., left-to-right for English speakers). Consistent with conventions in their culture, we found that 5-year-olds prefered depictions of events ordered from left-to-right to those ordered right-to-left. Furthermore, even though vertical artifacts for time are rare in their culture, 5 -year-olds prefered top-to-bottom representations of events to bottom-to-top ones. However, unlike adults, 5-year-olds showed no preference for horizontal (LR) over vertical (TB) depictions of events. Furthermore, younger preschoolers, 3-year-olds, not only appeared to lack direction or orientation preferences for ordered sequences, but also did not prefer ordered sets of pictures to unordered ones. Together, these findings suggest that children may not initially attend to the ordinal structure of event-depicting images, and that the "mental timeline" is constructed gradually in early childhood.

A substantial body of cross-cultural evidence indicates that the direction of mature linear mappings between time and space varies according to factors such as writing direction (e.g., Ouellet et al., 2015). A smaller number of studies indicate that these differences may emerge in childhood (Dobel, Diesendrunk, \& Bolte, 2007; Tillman, Tulagan, \& Barner, 2015; Tversky et al., 1991). The present study adds to this existing literature, by providing new evidence that cultural factors shape the direction of the mental timeline during childhood. Specifically, we found that preliterate 3-year-olds did not privilege conventional LR representations of time, and that LR biases appeared around age 5 , when literacy often begins to emerge ${ }^{2}$. Going beyond prior work, the current study also suggests that preliterate 3-year-olds may not map sequential temporal events to ordinal lines at all, regardless of the direction of those lines. If so, this suggests that both the directionality and the ordinal structure of the "mental timeline" are constructed during childhood, in response to increased environmental input.

The task used here was designed to give children more scaffolding for the formation of space-time mappings than previous studies have provided, while also making fewer response demands. In contrast to the classic sticker-placement task (Tversky, Kugelmass, \& Winter, 1991), for example, the present task did not require children to produce a spatial representation, or to recruit an implicit mental timeline "from scratch." Our task provided both the

[^212]temporal stimulus (a verbal story) and the spatial stimulus (images in lines) to be associated. The child simply needed to compare the two alternative mappings afforded by the two cards, and to pick the best match of temporal structure to spatial structure. However, given that 3-year-olds' performance in both experiments did not differ from chance, we cannot rule out the possibility that their failure stemmed from some less theoretically interesting incomprehension of the task. For example, it is possible that these children may have failed to recognize the images, or to remember the ordering of the three parts of the story. We are currently conducting a new experiment to test these possibilities.

Our findings are inconsistent with theories suggesting the LR direction of the MTL is a biological default that must be over-ridden to achieve an RL or TB mental timeline (Chatterjee, 2001; Rugani, 2015). Our findings also suggest that perceptual mappings between duration and length observed in infants cannot account for the ordinal MTL observed in adults and older children, in which positions in space (e.g., on the left) represent locations in time (e.g., in the past, see Winter, Marghetis, \& Matlock, 2015, for discussion). Several studies indicate that, if presented with a stimulus that is spatially "long" (e.g., a visual line) and temporally "long" (e.g., an auditory tone), prelinguistic infants associate these two dimensions automatically, and can detect mismatches between duration and length (de Hevia et al, 2014; Srinivasan and Carey). In contrast, preschoolers in the present study did not appear to align 3-part temporal sequences and analogous 3-part spatial representations. It is therefore possible that space-time associations in infancy apply only to temporal properties of single events, not to event sequences.

Our findings in 5-year-olds may also provide a hint into the process by which linear space-time mappings are shaped. In particular, we observed a developmental trajectory in which within-axis direction preferences (LR > $R L ; T B>B T$ ) emerged prior to a preference for one axis over the other. Can a literacy-based theory of MTL-acquisition account for this? In considering this question, it is interesting to note that English orthography has both a horizontal and a vertical component, with text progressing rightward across lines and downward through the page. Indeed, the vertical component of text may be more salient in children's books, which have fewer words per line than books for adults. Additional research will be needed to directly test whether children with more print exposure are more likely to make linear mappings between time and space - whether horizontal, vertical, or both.

## References

Bergen, B. K., \& Lau, T. T. C. (2012). Writing direction affects how people map space onto time. Frontiers in psychology, 3.

Bonato, M., Zorzi, M., \& Umiltà, C. (2012). When time is space: evidence for a mental timeline. Neuroscience \& Biobehavioral Reviews, 36(10), 2257-2273.
Boroditsky, L. (2011). How languages construct time. In Dehaene \& Brannon (Eds.), Space, time and number in the brain: Searching for the foundations of mathematical thought. 978-0-12-385948-8. Elsevier
Chatterjee, A. (2001). Language and space: Some interactions. Trends in cognitive sciences, 5(2), 55-61.
de Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., \& Streri, A. (2014). Representations of space, time, and number in neonates. Proceedings of the National Academy of Sciences, 111(13), 4809-4813.
Dobel, C., Diesendruck, G., \& Bölte, J. (2007). How writing system and age influence spatial representations of actions a developmental, cross-linguistic study. Psychological Science, 18(6), 487-491.
Haspelmath, M. (1997). From space to time: Temporal adverbials in the world's languages. Lincom Europa.
Hudson, J. A., \& Mayhew, E. M. (2011). Children's temporal judgments for autobiographical past and future events. Cognitive Development, 26(4), 331-342.
Nuerk, H. C., Patro, K., Cress, U., Schild, U., Friedrich, C. K., \& Göbel, S. M. (2015). How space-number associations may be created in preliterate children: six distinct mechanisms. Frontiers in psychology, 6, 215.
Ouellet, M., Santiago, J., Israeli, Z., \& Gabay, S. (2015). Is the future the right time?. Experimental psychology.
Rugani, R., Vallortigara, G., Priftis, K., \& Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. Science, 347(6221), 534-536.
Srinivasan, M., \& Carey, S. (2010). The long and the short of it: on the nature and origin of functional overlap between representations of space and time. Cognition, 116(2), 217-241.
Tillman, K. A., Marghetis, T, Barner, D., \& Srinivasan, M. (2017). Today is tomorrow's yesterday: Children's acquisition of deictic time words. Cognitive Psychology, 92, 87-100.
Tillman, K., Tulagan, N., \& Barner, D. (2015). Building the mental timeline: Spatial representations of time in preschoolers. Proc. CogSci.
Tversky, B., Kugelmass, S., \& Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. Cognitive Psychology, 23(4), 515-557.
Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. Trends in cognitive sciences, 7(11), 483-488.
Winter, B., Marghetis, T., \& Matlock, T. (2015). Of magnitudes and metaphors: Explaining cognitive interactions between space, time, and number. Cortex, 64, 209-224.

# Computational Exploration of Lexical Development in Down Syndrome 

Angel Eugenio Tovar (eugeniot@comunidad.unam.mx)<br>Department of Psychology, UNAM<br>Avenida Universidad 3004, Ciudad Universitaria, Mexico

Gert Westermann (g.westermann@lancaster.ac.uk)<br>Department of Psychology, Lancaster University<br>Lancaster LA1 4YF, United Kingdom


#### Abstract

Research on lexical development in Down syndrome (DS) has emphasized a dissociation between language comprehension and production abilities, with production of words being relatively more impaired than comprehension. Current theories stress the role of associative learning on lexical development. However, there have been no attempts to explain the atypical lexical development in DS based on atypical associative learning. The long-term potentiation (LTP) and long-term depression (LTD) of synapses, underlying associative learning, are altered in DS. Here we present a neural network model that instantiates notions from neurophysiological studies to account for the disparities between lexical comprehension and production in DS. Our simulations show that an atypical LTP/LTD balance affects comprehension and production differently in an associative model of lexical development.


Keywords: Down syndrome; lexical development; associative learning; comprehension/production asymmetries; neurocomputational model.

Down syndrome (DS) is the most common genetic cause of intellectual disability. There has been extensive research in behavioral and neurophysiological sciences to understand how DS affects cognitive development.

One of the behavioral domains that has attracted particular attention in DS is language development, and specifically, lexical development. This is interesting because lexical development has been argued to be based on associative learning mechanisms (McMurray, Horst, \& Samuelson, 2012), while studies on the neurophysiology of DS have consistently described an altered mechanism for synaptic adaptation (Begenisic et al., 2014; Scott-McKean \& Costa, 2011) which lies at the core of associative learning. Nevertheless, the role of atypical associative learning in lexical development in DS has not been explored.

In this paper, we address this gap by describing a neurobiologically informed computational model that implements an altered associative learning mechanism described in DS to account for the atypical lexical development in DS. Our focus is on explaining an apparent dissociation between lexical comprehension and production in DS. We want to address to what extent this observed dissociation is based on general atypical associative learning mechanisms. Our hypothesis is that interactions between experience and the neurophysiological constraints of DS are
sufficient to account for the differences in performance between lexical comprehension and production in this population.

This hypothesis is in accordance with a domain-general view of cognitive development, where the process of associative learning is affected overall, but depending on the demands of the task (i.e., comprehension or production) the observed outcomes are qualitatively different. We test this hypothesis in our computational model of lexical development. Therefore, a second aim of this paper is to provide a computational model of atypical lexical acquisition, biologically informed.

## Lexical development in Down Syndrome (The Process)

Language development in DS, as in other developmental disorders, has attracted considerable attention for both theoretical and practical reasons. On the one hand, descriptions emphasizing a relatively greater impairment in language abilities in DS (Chapman \& Hesketh, 2000; Rice, Warren, \& Betz, 2005; Vicari et al., 2004) have motivated theoretical debate on the nature of language as a process resulting from a cognitive system with domain specific vs. domain general components (Marcus \& Rabagliati, 2006; Stojanovik, 2014; Thomas \& Karmiloff-Smith, 2005). On the other hand, there is interest in understanding atypical language trajectories in DS to develop better interventions and minimize dysfunction in these patients. Among the different domains of language development, in this review we focus on lexical development.

Lexical acquisition is traditionally studied through the number of words produced and number of words comprehended in a certain age range. These numbers are lower in DS when compared to typically developing children (TD) of the same chronological age, but the discrepancy between DS and TD diminishes when DS individuals are compared with TD children of the same mental age (i.e., level of non-verbal cognitive ability) (Galeote, Soto, Sebastián, Rey, \& Checa, 2012).

In DS it is commonly reported that language comprehension abilities exceed language production abilities (Galeote et al., 2012; Kay-Raining Bird, Chapman, \& Schwartz, 2004; Vicari et al., 2004). This pattern replicates a canonical finding in research of lexical development in TD: the number of words comprehended initially exceeds the number of words produced (McMurray
et al., 2012). However, critically, a number of studies have found that the discrepancy between the comprehension and production of words in DS is greater than expected on the basis of mental age, with comprehension at or near mental-age-typical levels, but production selectively impaired (Kay-Raining Bird et al., 2004; Vicari et al., 2004).

Some studies also suggest that production and comprehension of words in DS follow qualitatively different developmental trajectories (Chapman, Hesketh, \& Kistler, 2002; Galeote et al., 2012), with one study reporting that comprehension of words in DS even exceeded the level of non-verbal mental age (Glenn \& Cunningham, 2005). However, in contrast to these results, other evidence has suggested that in DS both expressive and receptive language are significantly more impaired than what is expected on the basis of mental age (Bello, Onofrio, \& Caselli, 2014).

Due to conflicting results it has been difficult to characterize a unique profile of cognitive and linguistic abilities in DS. High inter-individual variability in the DS population (Karmiloff-Smith et al., 2016), along with methodological constraints including small sample size and the use of different measures and procedures, may explain some disparities between studies. In an effort to analyze a larger sample of DS individuals in verbal skills, Næss and colleagues (2011) meta-analyzed data reported by different research groups between 1988 and 2009, and found that performance of children with DS is in line with TD children, matched by mental age, in receptive vocabulary but is significantly impaired in measures of expressive vocabulary.

A number of questions arise from this apparently uneven profile between lexical comprehension and production, and its failed predictability from the overall level of cognitive development: is lexical development in DS only delayed or deviated from the TD pattern? Is there a dissociation between lexical comprehension and production in DS? Is it possible to account for these results with a domain general approach?

## Lexical Acquisition and Associative Learning (The Theory)

There is a vast literature on lexical acquisition and the study of word learning is at the core of this field. Word learning is viewed as the process by which we learn to link a phonological representation with a category of objects. Word learning involves a sequence of complex processes; the learner faces the challenge of selecting discrete phonological representations, picking a specific object in a cluttered visual scene, and creating meaningful representations linking the sounds and the visual objects.

Attempts to explain how the cognitive system deals with such a complexity have been based on three theoretical accounts. First, under the lexical constraints account, word learning is guided by a set of default assumptions (i.e., constraints) on hypotheses (Woodward \& Markman, 1998). For example, the mutual exclusivity constraint describes the process of inferring which word corresponds with which
object on the basis of knowing already the names of the other objects present in the visual scene.

Second, the social-pragmatic account argues that children use cues such as the speaker's (e.g., caregiver) gaze or intention to learn the correspondences between sounds and objects (see Ambridge \& Lieven, 2011). Third, the associative learning account explains word learning as a process governed by the domain-general rules of learning. The focus is on the linkages created between sounds and objects without appealing to any other prerequisites such as lexical constraints or social cues, even when these can exert a modulatory role on word learning. In recent years, this account has been formalized and tested through computational models (Mayor \& Plunkett, 2010; McMurray et al., 2012, Westermann \& Mareschal, 2014). Computational simulations have provided precise descriptions on how the qualitative properties of lexical development, empirically observed, as is the initial asymmetry between comprehension and production, the vocabulary spurt, and mutual exclusivity, emerge in a system that operates by establishing associations with language-like inputs (e.g., McMurray et al., 2012).

In this paper, we focus on analyzing the disruptive effects that atypical mechanisms of associative learning have on word learning for the DS population. For this reason, our approach is based on the model proposed by McMurray, Horst and Samuelson (2012), we call this the MHS model from here on. We selected this model for the following reasons: first, the theoretical account underlying this model distils the process to its basic computational components and develops an approach focused on the role of associative learning, and this is convenient for our purpose of analyzing atypical forms of associative learning on lexical development. Second, the architecture of this model is well suited to incorporating our computational formalization of biological descriptions of atypical learning in DS. Third, by building on previous work, we extend this previous and well accepted model to account for atypical behavior and in this extension (in terms of behavior, and populations) additional evidence is provided for the associative account of word learning.

## Associative Learning in Down Syndrome (The Underlying Mechanism)

From a neurobiological perspective, associative learning results from the adaptation of synaptic connections between neurons. Such adaptations are activity dependent; following Hebbian descriptions high co-activation between pre- and post-synaptic neurons lead to a strengthening of the synaptic connection. Complementary to the Hebbian account, empirical research has shown that decays in the efficacy of synaptic connections are also triggered by the co-activation between the pre- and post-synaptic units. A co-activation threshold is assumed to exist (Bienenstock, Cooper, \& Munro, 1982) so that below-threshold co-activation values produce decays in the synaptic efficacy (i.e. long-term depression or LTD) and above-threshold co-activation
values lead to increase the synaptic efficacy (i.e., long-term potentiation or LTP).

A vast literature on the biological bases of associative learning in DS has described an atypical balance between LTP and LTD in different mouse models of this syndrome. When compared with euploid control mice, LTP is limited and LTD is increased in DS (Begenisic et al., 2014; ScottMcKean \& Costa, 2011; Siarey, Villar, Epstein, \& Galdzicki, 2005). This pattern of synaptic adaptation functionally corresponds with an increased co-activation threshold, where the same level of stimulation produces limited gains and increased decays in the connection strengths in DS relative to TD. An increased co-activation threshold has been proposed for other populations that show cognitive impairment (Meredith \& Mansvelder, 2010).

While considerable progress has been made in the study of LTP/LTD in DS, with an emphasis on the design of pharmacological interventions (e.g., Begenisic et al., 2014), building the bridge from the basic level of altered neurophysiology to the high level of cognitive function has seen less progress. For example, it is not clear what is the role of the altered LTP/LTD balance on language development in DS. Descriptions of the exact way by which biological differences contribute to language impairments in different populations (e.g., TD, Williams syndrome, fragile X syndrome) will inform us on what is common across populations, the nature of language impairments, and how the language capacity is vulnerable (Rice et al., 2005).

Given the evidence from two fields of research, one informed by behavioral studies suggesting a preserved and marked asymmetry between comprehension and production of words, and another informed by neurophysiological studies describing an altered mechanism for associative learning, and in the context of an associative learning account to word learning, in this paper our focus is on exploring, the role of atypical associative learning mechanisms in word learning in DS.

## Computational Model

Overview and Architecture The present model is based on the MHS model. It is designed to analyze the role of associative learning in the establishment of correspondences between auditory word forms and visual objects. In the following, we describe our model and we indicate the differences between the present model and the MHS model.

The present model is composed of a neural network with three layers of units. Two of these layers represent processing in the auditory and visual systems. These layers are used to present input patterns to the network and to collect responses. These layers are not directly connected with each other; instead they are indirectly connected through a third layer of "lexical units" (see Figure 1).

One assumption of this approach is that the auditory and visual systems can already categorize objects and select discrete elements from the environment. The units in the visual and auditory layers are localist; each unit represent only one stimulus.


Figure 1: Architecture of the neural network with the visual, auditory and lexical layers. Only a few connections are shown to represent connectivity from auditory and visual units to lexical units.

The auditory and visual layers have 40 units each. Thus, 40 is the total number of words that the network is able to learn. The lexical layer contains 100 units. There are more lexical units than would be needed to learn 40 words -this allows for better learning (McMurray et al., 2012). Since the model could initially randomly associate two different inputs with the same lexical unit, increasing the number of lexical units prevents mismappings and increases discrimination of words (McMurray et al., 2012).

The architecture of the model is similar to the one presented by McMurray and colleagues (2012), but a key difference is in the number of units. The MHS model has 35 input units in the auditory and visual layers, and 500 lexical units. Our model incorporates more input units and fewer lexical units; thus our model requires less computational power to simulate the learning of a higher number of words.

Each unit in the input layers is connected to all the units in the lexical layer. These connections are bidirectional and their weights are initially randomized. In the MHS model, connections are not functionally bidirectional, since they use a different temperature parameter for feed-forward and feed-back connections.

Activation values of units range between 0 and 1. The activation values of the lexical units are initially normalized, such that the sum of all activation values equals 1 . When an auditory or visual stimulus is presented to the input layers, the unit that represents this stimulus is activated with a value of 1 , and all remaining inputs are set to 0 . The activation flows through the connections and reaches the lexical layer, which then computes the net input as the sum of activations coming from the auditory and visual inputs weighted by the corresponding connection values. The activation values in the lexical layer then go through a process of normalization (Equation 1), during 7 cycles. In our model 7 cycles are optimal to stabilize 100 lexical units. It is not clear how many cycles the MHS model requires.

$$
\begin{equation*}
a_{i}^{(t+1)}=\frac{\left(a_{i}^{(t)}\right)^{2}}{\sum_{i}^{N}\left(a_{i}^{(t)}\right)^{2}} \tag{1}
\end{equation*}
$$

The activation of the lexical units then feeds back to the auditory and visual layers; these units then sum the net input coming from the lexical layer with the activation from direct stimulation. This process allows integration of bottom-up with top-down information. Then, the connection weights are updated according to the rule described below.

Learning The MHS model incorporates a Hebbian learning algorithm that strengthens connections between co-acvtive units. The decay terms in the MHS model weaken the connections when either the lexical-, or the input units are inactive. In our model, the learning algorithm is designed to capture the functional differences in synaptic adaptation between TD and DS, as informed by studies with mouse models. Thus, both strengthening and weakening of connections result from the co-activation of units. Our algorithm incorporates a co-activation threshold ( $\theta$ ). Those co-activation values that surpass $\theta$ lead to gains in the connection weights, and co-activation values below $\theta$ lead to decays in connection weights. The simulations of DS use a relatively higher value for $\theta$ than simulations of TD (i.e., $\theta$ $=0.9$ for DS and 0.7 for TD). Higher values of $\theta$ restrict connection strengthening and increase connection decay; in this way we simulate the atypical pattern of increased LTD and limited LTP that has been described in DS.

To stabilize changes in connection weights we also include a self-adjusting parameter called lambda ( $\lambda$ ). It keeps weights between 0 and 1, by reducing changes as weights approach 1. As shown in Equation 2, for above- $\theta$ values, $\lambda$ depends on the difference between the coactivation and the current connection weight. It is computed by subtracting the value of the current weight from the current co-activation. For below- $\theta$ values, lambda acquires a negative value proportional to the current weight.

$$
\begin{gather*}
\text { If }\left(a_{\mathrm{i}}^{*} * \mathrm{a}_{\mathrm{j}}\right)>\theta, \text { Then } \lambda=\left(\mathrm{a}_{\mathrm{i}} * \mathrm{a}_{\mathrm{j}}\right)-\mathrm{W}_{\mathrm{ij}}  \tag{2}\\
\text { Else } \quad \lambda=-\mathrm{W}_{\mathrm{ij}}
\end{gather*}
$$

Lambda is a multiplicative parameter in the final learning algorithm (Equation 3).

$$
\begin{equation*}
\mathrm{W}_{\mathrm{ij}}(\mathrm{t}+1)=\mathrm{W}_{\mathrm{ij}}(\mathrm{t})+\lambda \beta\left(\mathrm{a}_{\mathrm{i}} * \mathrm{a}_{\mathrm{j}}\right) \tag{3}
\end{equation*}
$$

Changes in weights $\left(\mathrm{W}_{\mathrm{ij}}\right)$ then depend on the co-activation value ( $\mathrm{a}_{\mathrm{i}} * \mathrm{a}_{\mathrm{j}}$ ) modulated by the interaction between the current state of the connection and the co-activation computed by $\lambda$, and a learning rate ( $\beta$ ). We ran two sets of simulations for DS. In the first set (DS-1) we used a relatively lower $\beta$ in DS compared to TD simulations to capture additional neurophysiological abnormalities in DS with impact on computing power, namely, a reduction of synapse density and inhibitory predominance (Dierssen,
2012). In the second set of simulations of DS (DS-2) we kept the same value $\beta$ as the one used in TD. We did this to be able to compare and explore the effects of an increased $\theta$ alone vs. increased $\theta$ and lower $\beta$. $(\beta=0.001$ for TD and DS-2; and $\beta=0.0005$ for DS-1).

## Simulations

Training One auditory object was presented during each training trial along with many visual objects (usually five). These presentations simulate natural scenes where, in a discrete moment, one auditory word form is presented (spoken) to the child in the presence of a cluttered visual scene. For example, the first time a child hears the word /cat/, she can observe a visual scene that contains a cat, but also contains a dog, a container with milk, a ball of yarn, etc. Thus, the word /cat/ could initially refer to any of these visual objects. This problem of referential ambiguity needs to be solved by the child across many trials. Let's consider a second trial when the word /cat/ is presented again, but now the visual scene contains the cat, the container with milk, a pillow, and a table. If the child is sensitive to the environmental regularities, across many trials she will learn the correct correspondences between auditory words and visual objects (Smith \& Yu, 2008). But this is a slow process that requires numerous trials. To capture this process, in our simulations, each time that an auditory word was presented, the correct visual object was presented with another 4 different visual objects. The additional visual objects changed for every trial. We simulated the learning of 40 words, by presenting each auditory-visual pairing a total of 20000 times.

Testing We presented trials to evaluate comprehension and production of words. Tests for comprehension were designed, as in the MHS model, to simulate a traditional test of lexical comprehension, The N -alternative forced choice, where a number of different visual objects are presented to the child and she is asked to point or select one in particular (e.g., where is the pencil? which one is the pencil?). In our simulations one auditory stimulus (e.g., pencil) was active, as well as 4 visual objects (e.g., pencil, cat, table, glass) in the visual layer. Activation flowed from inputs to the lexical layer and back. Then the unit in the visual layer with the highest activation (e.g., pencil) was taken as the response of the model. In this way, comprehension was conceptualized as the correct activation of the visual object in the presence of one particular auditory word form.

Following again simulations in the MHS model, tests for production of words were designed to simulate the "child says" measures of the MacArthur-Bates Communicative Development Inventory. In these trials one single visual object was active and all possible auditory word forms were active. Activation flowed from inputs to the lexical units and back, then the auditory unit with the highest activation value was taken as the response of the model. Production then corresponded to evaluating the activation of auditory word forms in the presence of a particular visual object.

The comprehension and production test trials were run after every 50 training epochs (each epoch was composed of the presentation of the 40 training trials). A total of 400 measures of comprehension and production were obtained for each simulation. We ran 20 simulations of TD, 20 of DS-1 and 20 of DS-2.

## Results and Discussion

Figure 2 shows the mean values of words comprehended and produced for TD, DS-1, and DS-2. The standard deviation values are shown in the error bars (gray areas).

Our simulation of TD (Fig. 2A) shows that comprehension surpassed production in the early stages of learning; then, from the test trial 51 until the end of the simulation, comprehension and production were matched, and show complete learning of vocabulary.

The simulations of DS-1 and DS-2 (Fig. 2 B and C) show a qualitatively different trajectory of lexical acquisition. Some aspects shown by these simulations are of particular interest in the context of our theoretical and empirical review. First, performance in the comprehension task is always above the performance in the production task. Moreover, production of words never reaches the maximum possible value of 40 words. Second, DS-1 is more affected than DS-2. DS-1 used a higher co-activation threshold with a lower learning rate, while DS-2 used the higher coactivation threshold with a high learning rate. Data from DS-2 suggests that the atypical synaptic learning process in DS has a direct consequence on lexical development on its own, and the difference between DS-1 and DS-2 suggests that the learning rate has an additional effect. Third, the standard deviations show that the performance in the DS

groups was more variable than the performance in TD. DS-1 showed the highest variability. These patterns replicate the high inter-individual variability usually observed in DS compared with TD (see Karmiloff-Smith et al., 2016).

Comprehension and production tests were different tasks in our simulations. Comprehension required the selection of a visual stimulus from a sample of a few objects, while production, a more demanding task, required the selection of an auditory stimulus from the total number of auditory word forms. These tasks were designed to reproduce the top-down and bottom-up interactions that a child processes when she produces names vs. when she comprehends auditory words. Then, in our model, the asymmetries between comprehension and production are (partially) explained by the properties of the tasks. Remarkably, the disparity between comprehension and production in TD was overcome as training continued, but this disparity persisted for the DS simulations, thus pointing to the atypical associative learning mechanism as an explanation for the persistence and more marked disparity between comprehension and production of words in DS.

Other factors may as well contribute to the lexical comprehension/production asymmetry in DS, such as an atypical physical development that affects correct articulation of words and therefore restricts experience with lexical production. Our model, however, shows that the atypical pattern of synaptic strengthening directly affects lexical development.

Our approach supports a domain-general view of cognitive development, and we argue that it also strengthens the associative learning account to lexical development, since it explains a pattern of uneven development of lexical abilities in Down syndrome as a result of an altered domain-

Number of Test Trials
Figure 2: Mean values of comprehension and production across the 400 test trials for TD (Panel A), DS-1 (Panel B) and DS-2 (Panel C). The values from the three populations appear for comparison purposes in Panel D. Gray areas in Panels A, B and C show the standard deviation.
general mechanism in combination with the properties of the behavioral task.

## Acknowledgments

AET was funded by Secretaría de Ciencia Tecnología e Innovación-Ciudad de México. GW is funded by the ESRC International Centre for Language and Communicative Development (LuCiD) at Lancaster University (ES/L008955/1) and by a British Academy/Leverhulme Trust Senior Researcher Fellowship.

## References

Ambridge, B., \& Lieven, E. V. (2011). Child language acquisition: Contrasting theoretical approaches. UK: Cambridge University Press.
Begenisic, T., Baroncelli, L., Sansevero, G., Milanese, M., Bonifacino, T., Bonanno, G.,... Sale, A. (2014). Fluoxetine in adulthood normalizes GABA release and rescues hippocampal synaptic plasticity and spatial memory in a mouse model of Down Syndrome. Neurobiology of Disease, 63, 12-19.
Bello, A., Onofrio, D., \& Caselli, M. C. (2014). Nouns and predicates comprehension and production in children with Down syndrome. Research in Developmental Disabilities, 35(4), 761-775.
Bienenstock, E. L., Cooper, L. N., \& Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 2(1), 32-48.
Chapman, R. S., \& Hesketh, L. J. (2000). Behavioral phenotype of individuals with Down syndrome. Mental Retardation and Developmental Disabilities Research Reviews, 6(2), 84-95.
Chapman, R. S., Hesketh, L. J., \& Kistler, D. J. (2002). Predicting longitudinal change in language production and comprehension in individuals with Down syndrome: hierarchical linear modeling. Journal of Speech, Language, and Hearing Research, 45(5), 902-915.
Dierssen, M. (2012). Down syndrome: the brain in trisomic mode. Nature Reviews. Neuroscience, 13(12), 844-858.
Galeote, M., Soto, P., Sebastián, E., Rey, R., \& Checa, E. (2012). Vocabulary acquisition in children with Down syndrome: Normative data and developmental trends. Infancia y Aprendizaje, 35(1), 111-122.
Glenn, S., \& Cunningham, C. (2005). Performance of young people with Down syndrome on the Leiter-R and British picture vocabulary scales. Journal of Intellectual Disability Research, 49(4), 239-244.
Karmiloff-Smith, A., Al-Janabi, T., D'Souza, H., Groet, J., Massand, E., Mok, K., ... Strydom, A. (2016). The importance of understanding individual differences in Down syndrome. F1000Research, 5.
Kay-Raining Bird, E., Chapman, R. S., \& Schwartz, S. E. (2004). Fast mapping of words and story recall by individuals with Down syndrome. Journal of Speech, Language, and Hearing Research, 47(6), 1286-1300.

Marcus, G., \& Rabagliati, H. (2006). What developmental disorders can tell us about the nature and origins of language. Nature Neuroscience, 9(10), 1226-1229.
Mayor, J., \& Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. Psychological Review, 117(1), 1-31.
McMurray, B., Horst, J. S., \& Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological Review, 119(4), 831-877.
Meredith, R. M., \& Mansvelder, H. D. (2010). STDP and mental retardation: dysregulation of dendritic excitability in Fragile X syndrome. Frontiers in Synaptic Neuroscience, 2.
Næss, K.-A. B., Lyster, S.-A. H., Hulme, C., \& MelbyLervåg, M. (2011). Language and verbal short-term memory skills in children with Down syndrome: A metaanalytic review. Research in Developmental Disabilities, 32(6), 2225-2234.
Rice, M. L., Warren, S. F., \& Betz, S. K. (2005). Language symptoms of developmental language disorders: An overview of autism, Down syndrome, fragile X, specific language impairment, and Williams syndrome. Applied Psycholinguistics, 26(1), 7-27.
Scott-McKean, J. J., \& Costa, A. C. S. (2011). Exaggerated NMDA mediated LTD in a mouse model of Down syndrome and pharmacological rescuing by memantine. Learning \& Memory, 18(12), 774-778.
Siarey, R. J., Villar, A. J., Epstein, C. J., \& Galdzicki, Z. (2005). Abnormal synaptic plasticity in the Ts1Cje segmental trisomy 16 mouse model of Down syndrome. Neuropharmacology, 49(1), 122-128.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Stojanovik, V. (2014). Language in genetic syndromes and cognitive modularity. En L. Cummings (Ed.), The Cambridge handbook of communication disorders (pp. 541-558). Cambridge: Cambridge University Press.
Thomas, M., \& Karmiloff-Smith, A. (2005). Can Developmental Disorders Reveal the Component Parts of the Human Language Faculty? Language Learning and Development, 1(1), 65-92.
Vicari, S., Bates, E., Caselli, M. C., Pasqualetti, P., Gagliardi, C., Tonucci, F., \& Volterra, V. (2004). Neuropsychological profile of Italians with Williams syndrome: an example of a dissociation between language and cognition? Journal of the International Neuropsychological Society, 10(6), 862-876.
Westermann, G. and Mareschal, D. (2014) From perceptual to language-mediated categorization. Philosophical Transactions of the Royal Society B, 369, 20120391
Woodward, A. L., \& Markman, E. M. (1998). Early word learning. In Handbook of child psychology: Vol. 2: Cognition, perception, and language ( $\mathrm{pp} .371-420$ ). Hoboken, NJ: John Wiley \& Sons Inc.

# The temporal dynamics of base rate neglect: People may not be intuitive statisticians after all 

Eoin Travers<br>Institute of Philosophy, University of London<br>Jonathan J. Rolison<br>Department of Psychology, University of Essex<br>Aidan Feeney<br>School of Psychology, Queen's University Belfast


#### Abstract

According to a classic view of reasoning, intuition is fast but fallible, while reflection is slow but reliable. Biases, therefore, emerge when a reasoner's intuitions are wrong and they fail to notice. Recent evidence, however, suggests that people may be aware when their intuitions are incorrect. A possible explanation reason for this is that both correct and incorrect responses are cued in parallel, but the strongly-cued incorrect response is given unless people can inhibit it. We tested this explanation using base rate neglect problems, and recorded participants' mouse cursor movements as they chose between possible answers under time pressure. Descriptions affected both participants' early movements and ultimate responses, and interfered with their use of the base rates, while base rates rarely interfered with participants' use of descriptions, and then only at a later point in time. Thus, despite suggestive findings elsewhere, our results support the classic of view reasoning.


# Promoting Spontaneous Analogical Transfer by Idealizing Target Representations 

Máximo Trench (maxtrench@gmail.com)<br>Department of Psychology, Universidad Nacional del Comahue, IPEHCS/CONICET-UNCo. Quintral 1250, Bariloche, Rio Negro, Argentina<br>Lucía Micaela Tavernini (micaelatavernini@gmail.com)<br>Department of Psychology, Universidad Nacional del Comahue<br>IPEHCS/CONICET-UNCo, Irigoyen 2000, Cipolletti, Río Negro, Argentina

Robert L. Goldstone (rgoldsto@indiana.edu)<br>Department of Psychological and Brain Sciences, Indiana University<br>1101 E. 10th St., Bloomington, Indiana


#### Abstract

Recent results demonstrate that inducing an abstract representation of target analogs at retrieval time aids access to analogous situations with mismatching surface features (i.e., the late abstraction principle). A limitation of current implementations of this principle is that they either require the external provision of target-specific information or demand very high intellectual engagement. Experiment 1 demonstrated that constructing an idealized situation model of a target problem increases the rate of correct solutions compared to constructing either concrete simulations or no simulations. Experiment 2 confirmed that these results were based on an advantage for accessing the base analog, and not merely on an advantage of idealized simulations for understanding the target problem in its own terms. This target idealization strategy has broader applicability than prior interventions based on the late abstraction principle, because it can be achieved by a greater proportion of participants and without the need to receive target-specific information.


Key words: analogy, transfer, idealization, retrieval

## Introduction

Analogical reasoning represents a powerful heuristic for creative problem solving. By matching an unsolved situation (the target analog) to a stored exemplar whose solution is known (the base analog), the base solution can be transferred to the target problem. One of the most robust findings in the experimental literature on analogical transfer is that people often fail to spontaneously retrieve analogous situations when they do not share surface features with the target situation being processed (Gick \& Holyoak, 1980; Keane, 1987; Trench, Oberholzer, \& Minervino, 2009, for a discussion of naturalistic findings, see Trench \& Minervino, 2015).

A considerable body of research has sought to enhance spontaneous analogical retrieval by means of promoting a more abstract encoding of the base analogs, so as to render them more accessible during later encounters with analogous situations lacking surface similarities with the base analogs. Two successful interventions have consisted in presenting the base analog together with its abstract schema (Goldstone \& Wilensky, 2008) or with a second analogous situation
(Catrambone \& Holyoak, 1989), and asking participants to compare them. More stripped-down interventions include asking participants to discuss the base analog with another student (Schwartz, 1995), to explain the problem to themselves (Ahn, Brewer, \& Mooney, 1992) or to construct a structurally equivalent problem (Bernardo, 2001). Even if participants are not asked to elaborate on the base situations, transfer advantages can still be obtained by means of removing irrelevant information in the base analog (Goldstone \& Sakamoto, 2003), and even by replacing domain-specific terms of the base situation with domaingeneral ones (e.g., replacing "typing" by "writing", Clement, Mawby, \& Giles, 1994). What all of these interventions have in common is the highlighting of the abstract structure of the base analogs. As future, relationally similar examples will have a stronger match with such stripped-down representations than they will with specific examples having surface features that mismatch, the future retrievability of relationally encoded base analogs increases. Despite the relative success of these interventions, they cannot be applied to already learned situations or procedures that had not been originally encoded in ways that highlighted their abstract structure.

## The late abstraction principle

Kurtz and Loewenstein (2007) reasoned that as retrieval depends on the degree of match between the stored items and the memory probe, the beneficial effect of relational schemas should also apply when elaborating on the target analog at retrieval time. The removal of lower-level information was hypothesized to increase distant retrieval (1) by granting more weight to structural predicates due to the normalization of content vectors, and (2) by reducing the unwanted activation of competing situations that maintain only superficial similarity with the target. To gather behavioral evidence for this theoryladen prediction, Kurtz and Loewenstein (2007, Experiment 1) assessed the effectiveness of an intervention that consisted in providing participants with a second (unsolved) problem that was isomorphic to the target problem to be solved, and asking them to compare both problems prior to attempting their solution.

As was the case with the "base comparison" interventions, the abstraction process induced by this "target comparison" procedure resulted in enhanced transfer of the base solution as compared to the standard base-target paradigm. In subsequent work, Gentner, Loewenstein, Thompson and Forbus (2009) generalized the benefits of the target comparison strategy to autobiographical memories that were acquired several years prior to the experimental session, and also simulated the process of backward transfer using a retrieval algorithm and a set of stories that were developed before the late analogical encoding hypothesis had been proposed. To carry out these simulations, Gentner et al. (2009) fed MAC/FAC (Forbus, Gentner, \& Law, 1995), with either the original stories from the Karla the Hawk series of studies (Gentner, Rattermann, \& Forbus, 1993) or with their respective abstract schemas, and had it run on a long-term memory comprising analogical matches, mere appearance matches, and several filler stories. In line with the target-comparison studies, MAC/FAC retrieved more analogical matches when using the schemas rather than detailed stories as working memory cues.

As suggested by the results of the target-comparison studies, the process of late analogical abstraction opens a promising avenue for retrieving base situations whose initial encoding was not especially engineered to highlight their abstract properties, and which represent the vast majority of the situations we learn within and outside instructional settings. In contrast to the widespread potential applicability of the late analogical abstraction principle, however, the specific targetcomparison intervention falls short of representing a truly portable cognitive strategy because participants will depend on the external provision of a second analogous problem for every new target problem they are to solve.

With the aim of helping learners capitalize on late analogical abstraction without needing to be provided with additional information about the target, Minervino, Olguín and Trench (2017) demonstrated that analogical transfer from a distant source analog can be enhanced by asking participants to invent a new unsolved problem analogous to the target. Even though successful problem constructors were much more likely than unsuccessful constructors to transfer the base solution to the target problem, only a small proportion of participants succeeded at fabricating an analogous problem, an activity that seems to require a great deal of worldknowledge and above-average intellectual engagement.

In order to devise more widely applicable ways of capitalizing on the late abstraction principle, in the present study we identified an easily executed strategy credited with having enhanced the retrievability of base analogs during their initial encoding, and assessed whether its application to the target analog proves advantageous for retrieving analogous problems lacking surface similarities.

## Concrete vs. idealized representations

Goldstone and Sakamoto (2003) examined whether there was an effect of training with concrete or idealized graphics on spontaneous transfer of a general principle called "competitive specialization." Participants were trained with
an Ants and Food simulation with concrete graphics (black ants and small fruit) or idealized elements (black dots and green blobs) as shown in Figure 1. Afterwards they were asked to explore another instance of the competitive specialization principle in which initially undifferentiated matrices progressively learn to respond to a predefined set of letter inputs. Results revealed that participants in the idealized condition showed better transfer to the Sensors and Inputs quiz than in the concrete condition.


Figure 1. Snapshots of the concrete and idealized simulations of base analogs employed by Goldstone and Sakamoto (2003).

In order to assess whether a comparable transfer advantage can be obtained by inducing a more idealized representation of the target analog at retrieval time, we had three groups of participants learn how to solve a "collision" problem in which a plane and a helicopter travelled towards each other at different speeds. After a distracting task, participants were presented with a problem pertaining to a different family of algebra problems (i.e., "work problems"), but whose abstract structure was similar to that of the learned problem. In this problem, participants had to calculate the time that two painters would need to jointly paint a wall, given the times that each of them would have needed to paint it on his own. Before being asked to actually solve the problem, two of the groups were presented with a set of manipulatives and were tasked with carrying out an approximate representation of the situation described by the target problem as it unfolded from the initial moment until the moment when the wall got completed. While participants in the concrete condition received a realistic illustration of a horizontally laden wall and two smaller rectangles printed with drawings of painters, participants in the idealized condition received similarly sized paper rectangles without any figurative illustrations.

## Experiment 1

## Method

Participants and design A total of 90 participants were recruited from the Department of Psychological and Brain Sciences participant pool at Indiana University-Bloomington. All participants signed an informed consent for participation in the study, and were compensated with course credit. An equal number of participants $(N=30)$ were randomly assigned to the idealized, the concrete, and the no simulation conditions.

Procedure and materials The experimental session was introduced to participants as dealing with the effectiveness of instructions for solving different kinds of algebra problems. Participants were told that for most of the problem types to be covered during the session, they would begin by trying to solve a problem of such type on their own, follow by reading instructions on how to solve such problem, and finish by applying the learned strategy to a subsequent problem of the same type. Unbeknownst to participants, the first block served to encode a base analog and its solution, the third block was used as a test of whether participants spontaneously applied the base solution to a seemingly unrelated problem that admitted a similar solution strategy, and the middle block served to contextually separate the first and third blocks. Upon receiving a booklet containing the materials, they were told that they would be informed in advance how much time they would have for carrying out each of the tasks, and that they could only proceed to the following page of the booklet once the experimenter had notified them that the allotted time for the current activity had elapsed. Participants were also provided with a pencil, an eraser, and an electronic calculator. The session was administered in small groups ranging from one to ten, with each participant working individually.
During the first block of problems (i.e., the encoding phase), participants of all groups were presented with a typical "collision" problem in which a plane and a helicopter initially located at two cities 2000 miles apart started travelling towards each other at different speeds (See Table 1). Participants were allotted 5 min to calculate the time the aircrafts would need to pass next to each other. Once the allotted time had elapsed, they were given 3 min to read a worked solution to such problem that included a standard illustration in which the plane and the helicopter were located at their respective cities A and B , which were in turn connected by a straight horizontal line. Participants were given 4 more min to apply the learned strategy to a similar problem in which a helium balloon and an elevator located at the top vs. bottom of a tall building begin travelling toward each other at different speeds (see Table 1). Given that achieving a basic understanding of the base problem and its solution represents a necessary prerequisite for subsequent transfer to occur, participants who failed to apply the base solution to this second problem were withdrawn from further analysis.

Table 1: Base and target problems used in Experiment 1

Base problem: A plane flies at 600 mph , while a helicopter flies at 100 mph . Imagine that the plane starts flying from City A to City B at the same time that the helicopter departs from city B to City A. How long will it take them to pass each other, if the cities are 2000 miles apart?
Base problem 2: While a helium balloon goes up at a speed of 2 feet per second, an external elevator travels at a rate of 4.5 feet per second. Suppose that the elevator starts descending from an altitude of 100 feet at the same moment that the balloon is freed from street level. How long will it take them to pass each other?
Target problem: Fred can paint an 18 -feet wall in 8 hours, while Bob can paint such wall in 5 hours. How long will it take
them to paint such wall in case they painted it together?

The second block of problems had the same structure and time allowances as the encoding phase, with the difference that it involved learning and applying a simple procedure for solving combinatorics problems that were unrelated to the prior problems. It thus served to contextually separate the encoding and transfer phases.
The third section (i.e., the transfer phase) was presented to participants of all groups as dealing with "work" problems, and had a different structure than the two previous phases. For brevity, we begin by describing the procedure followed by the concrete simulation group, and proceed by describing how the other conditions differed from such condition.

Participants of the concrete simulation condition received a typical work problem in which they had to calculate the time that two painters would need to jointly paint a wall, given the times that each of them would have needed to paint it on his own (see Table 1). They were given 2 min to read the problem very carefully, but they were asked to refrain from attempting a solution until explicitly indicated by the experimenter. Right below the problem text, the page displayed a 6.37 in $x$ 1.84 in sized illustration of a brick wall printed in greyscale. Upon receiving two small paper rectangles each one illustrated with a figurative drawing of a painter (one grey and one black, see Figure 2), participants were asked to take advantage of these manipulatives to carry out an approximate representation of how the painting of the wall unfolds over time, from the moment the painters start their job until the moment when it gets completed. In order to get a record of the specifics of each participant's simulation, the next page included three similar walls meant to represent three different snapshots of the dynamic simulation they had just performed. Upon receiving four additional paper painters (two grey and two black) and a glue stick, they were allotted 2 min to produce a record of the simulation they had just performed by means of sticking two painters onto each wall in a manner faithful to the locations of each of the painters at three different moments: (1) at the exact moment when they started painting [top wall], (2) at an intermediate stage of the process [center wall], and (3) at the exact moment when the painting job was completed [bottom wall]. Once the time allotted to this activity had elapsed, participants were given 5 min to solve the problem by any means.

The procedure followed by the idealized simulation group was identical to that of the concrete simulation condition, with the difference that the manipulatives used during the simulation were relatively more abstract. While the wall consisted of a white 6.37 in $x 1.84$ in sized rectangle, the two painters were represented by 1.6 in $x 0.75$ in sized grey/black paper rectangles.
The procedure followed by the no simulation group was identical to that of the simulation conditions, with the difference that participants were not asked to simulate the situation models of the target problem prior to attempting its solution.


Figure 2: Manipulatives employed for the idealized (top panel) and concrete (lower panel) representations of the target problem.

Data analysis Two independent judges sorted the solutions to the posttest (the "collision" problem featuring a balloon and a helicopter) as either correct or incorrect. Solutions were scored as "correct" whenever (1) the collision time obtained was expressed with at least one decimal position and coincided with the exact solution, and (2) the participant showed how such result was derived. Eight participants (five from the concrete simulation condition, three from the no Simulation condition and one from the idealized Simulation condition) were not able to apply the base solution to the balloon problem, and were thus removed from further analyses. Two additional judges blind to the purposes of the experiment followed the above criteria to score participants' solutions to the target problem. Judges reached $94 \%$ agreement regarding solutions to the balloon problem and $96 \%$ agreement regarding solutions to the target problem. Cases of disagreement were resolved by discussion.

## Results and Discussion

The rates of correct solutions to the target problem were $36 \%, 79 \%$, and $51 \%$ for the concrete, idealized and no simulation conditions, respectively. The spontaneous transfer rate in the idealized condition was reliably greater than those obtained in the concrete, $\chi^{2}(1, N=54)=10.43, p=.0012$, and in the no simulation conditions, $\chi^{2}(1, N=56)=4.7, p=.0302$.

The rates of spontaneous transfer did not differ between the concrete and the no simulation conditions, $\chi^{2}(1, N=52)=1.32$, $p=.2506$ These results indicate that idealized representations were more advantageous than concrete representations for eliciting correct answers to the work problem. On the other hand, the fact that the idealized simulation condition also outperformed the no simulation condition suggests that there are genuine benefits of idealization as opposed to disadvantages due to concrete representations.

In a manner similar to the transfer advantage of comparing two analogous targets (Gentner et al. 2009), the observed advantage of the Idealized Simulation Group in generating correct solutions suggests that there is a general advantage of lean representations for accessing analogous situations lacking superficial similarities with the target. However, an alternative explanation could be that the concrete representations of the painters might have invited a dynamical representation that was inconsistent with the "convergent" representation that characterized the base problems. If the concrete simulation of the painters' activity recruited a "socially laden" representation in which the painters advance in parallel fashion-e.g., to talk to each other-rather than in the more transfer-appropriate "converging" motion, this idiosyncratic accidental feature could have contributed to their inferior transfer performance. In order to assess this possibility, we sorted participants' representations as "convergent" vs. "non convergent" according to the way in which they had glued the painters onto the three walls that were meant to record three informative snapshots of how participants intuitively imaged the process as it unfolded over time. This analysis revealed a nonsignificant trend towards a greater use of the convergent representation in the concrete simulation condition (96\%) than in the idealized simulation condition (76\%), $p=.056$ (Fisher exact test). Given that the opposite trend would have been expected under the socially-laden interpretation account, the relative advantage of idealized simulations appears not to be due to an intrinsic advantage this kind of representations for prompting a convergent motion simulation.

Another alternative explanation for the superiority of target representations for eliciting correct solutions to the work problem could be that such advantage was originated, not in the benefits of our idealized materials for analogical transfer (as posited here), but rather in their potential to promote a better understanding of the target problem in its own terms, thus leading to a higher probability of solving such problem by first principles. According to various authors (see Belenky \& Schalk, 2014 for a discussion) learning is facilitated when representations convey the minimum detail that is necessary to grasp the quantitative structure of a problem. As an example, the removal of potentially distracting irrelevant features like the quasi-regular pattern of the bricks or the left vs. right handedness of the painters could have helped participants build a more accurate representation of the temporal dynamics of the problem (e.g., the different speeds of each painter), which may in turn serve as a secure foundation from which to control the accuracy and soundness of algebraic manipulations (Minervino, Trench, \& Oberholzer, 2009).

In order to assess how the concrete and idealized simulations enforced in Experiment 1 impacted the raw probabilities of solving the target problem in a non-analogical fashion, in Experiment 2 the transfer phases of the idealized, concrete and no simulation conditions were not preceded by the presentation of a structurally equivalent base analog.

## Experiment 2

## Method

Participants and design A total of 90 participants ( $N=30$ per condition) were recruited from the Department of Psychological and Brain Sciences participant pool at Indiana UniversityBloomington, and were compensated with course credit.

Procedure and materials The experimental session was introduced to participants as dealing with the effectiveness of different instructional aids for solving algebra problems, and took place after participants completed an unrelated experiment whose length was roughly equivalent to the time taken by participants of Experiment 1 to complete the encoding plus distracter phases. Upon receiving a booklet containing the materials, they were told that they would be informed in advance how much time they would have for carrying out each of the tasks, and that they could only proceed to the following page of the booklet once the experimenter had notified them that the time allotted to the current activity had elapsed. Participants were also provided with a pencil, an eraser, and an electronic calculator. The session was administered in small groups ranging from one to ten, with each participant working individually.
Participants of the simulation conditions received the painters' problem coupled with the same manipulatives and the same simulation tasks as in the corresponding groups of Experiment 1. After completing the simulation tasks, they were given 5 min to try solving the problem by whatever means. The procedure followed by the No Simulation Group was identical to that of the simulation conditions, with the difference that participants were neither provided with manipulatives nor invited to simulate the situation model of the problem prior to attempting its solution. Coding of correct solutions followed the same criteria as in Experiment 1, with judges reaching total agreement.

## Results and Discussion

The rates of correct solutions to the target problem were $37 \%$, $30 \%$, and $33 \%$ for the concrete, idealized and no simulation conditions, respectively. The rate of correct solutions in the idealized condition did not differ from that obtained in the concrete condition, $\chi^{2}(1, N=60)=0.3, p=.5839$. Similarly, differences were neither found between the no simulation and the idealized simulation conditions, $\chi^{2}(1, N=60)=0.08, p=.7773$, nor between the no simulation and the concrete simulation conditions, $\chi^{2}(1, N=60)=0.07, p=.7913$. The fact that the rate of correct solutions obtained by the Idealized Simulation Group was not even numerically higher than those of the concrete and the no Simulation conditions (in fact it was slightly lower) confirms that the advantage of idealized simulations over the
other conditions of Experiment 1 did not originate in their ability to promote a better comprehension of the target problem, but rather in an advantage for transferring a previously learned solution to a superficially dissimilar target.

## General Discussion

The present results are compatible with Gentner et al.'s (2009) late abstraction principle, which postulates that just as source abstractions can be beneficial for later analogical retrieval (i.e. forward transfer), manipulations aimed at highlighting the structure of the target can enhance the retrieval of superficially similar base analogs whose encoding was not intended to emphasize their structural features. It should be noted, however, that the perceptual nature of our concrete vs. idealized manipulation is very different from the "conceptual" abstraction induced by Kurtz and Loewenstein (2007) or Minervino et al. (2017), and computationally simulated by Gentner et al. (2009). In the above studies (see Trench \& Minervino, 2017 for a review), the domain-specific elements of the original problems (e.g., "destroy a tumor") are allegedly replaced by more domain-general expressions (e.g., "neutralize a central target"), which could promote distant retrieval in at least two different ways: (1) by granting more relative weight to the relational predicates of target representations, and (2) by decreasing the retrieval of mere appearance matches that could outcompete useful base situations with dissimilar surface features but similar structure. The fact that we obtained similar results by means of removing perceptual detail from the target representations suggests a subtle parallelism between the abstraction process that takes place in tasks like problem comparison or problem construction and the kind of idealization induced by our manipulation of the target. Akin to the advantage of abstract retrieval cues in the MAC/FAC simulations of the late abstraction principle, the observed advantage of idealized simulations of the target analog might have originated in their tendency to be, on average, perceptually more similar to the superficially dissimilar base analogs compared to their alternative concrete representations, as well as in their being less likely to evoke superficially matching situations that could outcompete the base analog. The present results thus contribute to enlarging the empirical basis of the late abstraction principle, while at the same time broadening its scope so as to include a perceptual dimension that has not been thus far discussed in the existing literature.

Much of the excitement over target elaborations stems from the possibility of retrieving base analogs learned under conditions that were not especially engineered to highlight their abstract features. If the encoding specificity hypothesis applied, however, any advantage of distilling abstract or idealized representations of the target would be limited to maximizing the retrieval of stored representations whose initial encoding had already emphasized those same features (Tulving \& Thompson, 1973). As discussed in more detail elsewhere (Trench \& Minervino, 2017), there are several ways in which a base analog can be suboptimally encoded, and yet benefit from a more structural representation of the target.

Beyond their relevance for theoretical models of analogical retrieval, the present results bear implications for the design of interventions aimed at fostering a flexible use of learned contents. On the one hand, the fact that asking participants to carry out idealized simulations led to higher solution rates than not requiring them to perform any kind of simulation indicates that the superior performance of the idealized condition was not based on an intrinsically detrimental effect of concrete simulations. More importantly, the activity of constructing idealized representations of the target overcomes important limitations of previous instantiations of the late abstraction principle. With regards to Kurtz and Loewenstein's (2007) target-comparison intervention, an important shortcoming had to do with the need to provide participants with a second analogous target for every problem to be solved by analogy. Even though Minervino et al.'s (2017) target-construction intervention was not subject to this crucial limitation, only a small proportion of participants were able to generate an isomorphic problem. In contrast to the above instantiations of the late abstraction principle, the cognitive strategy assessed in the present study can be easily implemented by a great majority of participants, and without needing to be provided with additional information about the target. Future research should assess whether the advantages of target idealization can be combined with the benefits of strategic search (see e.g., Trench, Olguín, \& Minervino, 2016), as well as whether they generalize to other educationally relevant activities such as generating explanatory hypotheses for poorly understood phenomena or communicating complex ideas to others.

## Acknowledgments

This work was supported by Grants PICT 2352 and 2650 from the ANPCyT of Argentina, Grants B213, C123 and C134 from the National University of Comahue, and by a Fulbright research fellowship to the first author of this paper.

## References

Ahn, W. K., Brewer, W. F., \& Mooney, R. J. (1992). Schema acquisition from a single example. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 391-412.
Belenky, D. M., \& Schalk, L. (2014). The effects of idealized and grounded materials on learning, transfer, and interest: An organizing framework for categorizing external knowledge representations. Educational Psychology Review, 26, 27-50.
Bernardo, A. B. I. (2001). Analogical problem construction and transfer in mathematical problem solving. Educational Psychology, 21, 137-150.
Catrambone, R., \& Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1147-1156.
Clement, C., Mawby, R., \& Giles, D. (1994). The effects of manifest relational similarity on analog retrieval. Journal of Memory and Language, 33, 396-420.
Forbus, K., Gentner, D., \& Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. Cognitive Science, 19, 141-204.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Gentner, D., Loewenstein, J., Thompson, L., \& Forbus, K. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. Cognitive Science, 3, 1343-1382.
Gentner, D., Rattermann, M. J. \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive Psychology, 25, 431-467.
Gick, M. L., \& Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychology, 12, 306-355.
Goldstone, R. L., \& Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. Cognitive Psychology, 46, 414-466.
Goldstone, R., L., \& Wilensky, U. (2008). Promoting transfer by grounding complex systems principles. Journal of the Learning Sciences, 17, 465-516
Keane, M.T. (1987). On retrieving analogues when solving problems. Quarterly Journal of Experimental Psychology, 39, 29-41.
Kurtz, K., \& Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. Memory \& Cognition, 35, 334-341.
Minervino, R., Olguín, V., \& Trench, M. (2017). Promoting interdomain analogical transfer: When creating a problem helps to solve a problem. Memory \& Cognition, 45, 221-232.
Minervino, R., Trench, M., \& Oberholzer, N., (2009). Concrete and imagined simulations of situation models enhance transfer of solutions to structurally different algebra word problems. In N. A. Taatgen \& H. Van Rijn (Eds.), Proceedings of the $31^{\text {st }}$ Annual Conference of the Cognitive Science Society (pp. 1394-1399). Austin, TX: Cognitive Science Society.
Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. The Journal of the Learning Sciences, 4, 321-354.
Trench, M., \& Minervino, R. (2015). The role of surface similarity in analogical retrieval: Bridging the gap between the naturalistic and the experimental traditions. Cognitive Science, 39, 1292-1319.
Trench, M., \& Minervino, R. (2017). Cracking the problem of inert knowledge: Portable interventions to access distant analogs from memory. In B. H. Ross (Ed.), The Psychology of Learning and Motivation, Vol. 66 (pp. 1-42). San Diego, CA, Academic Press.
Trench, M., Oberholzer, N., \& Minervino, R. (2009). Dissolving the analogical paradox: Retrieval under a production paradigm is highly constrained by superficial similarity. In B. Kokinov, K. Holyoak \& D. Gentner (Eds.), New frontiers in analogy research (pp. 443-452). Sofia: NBU Press.
Trench, M., Olguín, V., \& Minervino, R. (2016). Seek, and Ye shall find: Differences between spontaneous and voluntary analogical retrieval. Quarterly Journal of Experimental Psychology, 69, 698-712.
Tulving, E., \& Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80, 352-373.

# It's all in your head: Effects of expertise on real-time access to knowledge during written sentence processing 

Melissa Troyer ${ }^{1}$ (mtroyer@ucsd.edu) \& Marta Kutas ${ }^{1,2}$ (mkutas@ucsd.edu)

Department of Cognitive Science ${ }^{1}$, Neurosciences ${ }^{2}$, UCSD
9500 Gilman Dr., La Jolla, CA 92093


#### Abstract

Real-time sentence processing involves connecting linguistic input with knowledge. Here, we ask how variability in semantic memory (specific domain knowledge) may influence semantic access in realtime sentence processing. We recorded EEG while participants more/less knowledgeable about the narrative world of Harry Potter (HP) read sentences. In Experiment 1, all participants showed N400 predictability effects for general-knowledge sentences, but only those with high HP knowledge showed predictability effects for sentences about Harry Potter. This effect was driven by graded brain responses to predictable endings as a function of knowledge. Experiment 2 revealed greater semantic activation (inferred from N400 effects) for HP items participants reported knowing. Highknowledge participants also showed greater semantic activation for items they reported not knowing/remembering. These findings suggest that amount and/or functional organization of knowledge has real-time consequences on written sentence processing and implicate implicit/partial access to domain knowledge for experts when information is not explicitly recalled.


Keywords: sentence processing; knowledge; ERPs; individual differences

## Introduction

Depending on your background, the question, "What's your patronus?" might leave you bewildered. But if you've spent a sizable chunk of your life obsessing over the fictional world of Harry Potter created by J.K. Rowling, you might have a response quickly at hand (e.g., a dolphin or a cat).

Variability in individual experiences helps determine an individual's knowledge, whether the domain is a fictional narrative world like Harry Potter, a game like chess, or an academic discipline, like physics. Moreover, knowledge differences in many domains have been shown to systematically influence various aspects of the organization of knowledge, including depth, breadth, and hierarchical information structure (Chi, 2006). Such differences in knowledge seem likely to impact real-time semantic access, including perceiving an utterance (or text), relating it to prior knowledge, and forming expectations about upcoming content. Yet despite the inevitable link between an individual's knowledge and semantic access, the specific role(s) of knowledge variability has received relatively little attention in models of real-time language processing.

Decades of psycholinguistic research have revealed that language processing is incremental: we update our mental representations word-by-word (e.g., Tanenhaus et al., 1995; Kamide, Altmann, \& Haywood, 2003). Upon encountering an incoming word, world knowledge is used as soon as possible (e.g., Kutas \& Hillyard, 1980). Real-time access to such knowledge is influenced by a host of contextual factors,
both linguistic and nonlinguistic, including sentence and discourse context (Kutas \& Hillyard, 1980), discourse context (Nieuwland \& Van Berkum, 2006), and who the speaker is (Van Berkum et al., 2008).

These (and many other) studies have used event-related brain potentials to investigate the activation and organization of the semantic system during real-time language processing. A well-known ERP signature called N400 (a broad centroparietally distributed, negative-going potential peaking approximately 400 ms after stimulus onset) shows finegrained sensitivity to semantic relationships, with stronger relations between context and input yielding less negative- or more positive-going potentials between 200-500 ms postinput onset (Kutas \& Federmeier, 2000).

The content and organization of long-term memory (i.e., knowledge) influence semantic access as reflected in N400 modulation both within sentences and simpler (or even no) context. For words presented in isolation, N400 amplitude is reduced for high-, compared to low-frequency, words (Kutas \& Federmeier, 2000). Moreover, N400 amplitude is sensitive to category membership. Following a category label (e.g., 'a type of bird'), typical category exemplars ('robin') yield reduced N400 amplitude compared to atypical exemplars ('turkey'), and both are reduced compared to unrelated words ('broom') (Federmeier, Kutas, \& Schul, 2010).

Such effects rely on long-term knowledge likely available due to years of experience with concepts like birds. N400 studies, however, have also revealed sensitivities to culturally-specific information (e.g., the fact that Dutch trains are yellow, not white; Hagoort et al., 2004) and fictional information (Nieuwland \& Van Berkum, 2006; Filik \& Leuthold, 2013). Taken together, these N400 findings offer a window into the relationship between language input and structured, flexible knowledge use. Moreover, N400 amplitude provides an excellent proxy measure of the ease of access to semantic information.

Here, we use the N400 to explore the notion that systematic variability in the content and organization of individuals' knowledge, as a function of their expertise, will have systematic influences on real-time semantic access. To that end, we conducted two ERP studies with individuals varying in their knowledge of the narrative world of Harry Potter. We first asked whether domain knowledge will systematically influence N400 effects, possibly reflecting ease of semantic access and/or availability of information in long-term memory (Experiment 1). Next, we dissociated knowledge of individual facts from domain knowledge, allowing us to ask whether or not, and if so how, domain knowledge influences ease of semantic access when individuals think they know, or don't know/remember, the information (Experiment 2).

## Experiment 1

In Experiment 1, participants ranging in their knowledge of Harry Potter (based on an objective offline measure) read Control sentences about general, real-world topics as well as sentences about the narrative world of Harry Potter (HP) while we recorded EEG. Sentences of both types ended with either a Predictable or an Unpredictable word.

Based on the literature (Kutas \& Federmeier, 2000), we expected predictability effects for Control sentences, with Predictable items eliciting reduced N400 amplitudes compared to Unpredictable items. Moreover, we expected that for HP sentences, specific knowledge of Harry Potter would have its effect during this N400 time window, with knowledgeable individuals showing a reliable predictability effect and less knowledgeable individuals showing a smaller (or no) difference between Predictable and Unpredictable words.

Though our study focused on predictions during the N400 time window, we also anticipated later positive effects. Late positive complexes, often occurring post-N400, have been related to attention-driven processing, including integration, revision, or updating of ongoing interpretations in the presence of unexpected items (e.g., Van Petten \& Luka, 2012; Brouwer et al., 2012). We suspected that we might observe effects of Predictability on late positivities to words ending Control sentences, and possibly (for knowledgeable individuals) to words ending HP sentences, to the extent that individuals revise their interpretations.

## Methods

Participants 40 right-handed students at UCSD participated for partial course credit and some monetary compensation.

Sentence Materials During the ERP portion of the study, participants read sentence pairs of two types. Control sentences described commonplace scenarios and ended in a Predictable (Offline Cloze > 87\%) or an Unpredictable (Offline Cloze $=0 \%$ ), albeit plausible, word, determined by offline norming studies. Harry Potter (HP) sentences described situations and entities from the Harry Potter book series and ended in a Predictable (book-consistent) or an Unpredictable (book-inconsistent) word. Unpredictable words were matched to Predictable words for the broad classes of words they belonged to (common noun, proper noun, Harry Potter-specific noun) and in many cases belonged to the same, more specific category. For example:
(1) Control: We had been watching the blue jay for days. The bird laid her eggs in the nest. (Predictable) yard. (Unpredictable)
HP: The character Peter Pettigrew changes his shape at times. He takes the form of a rat. (Predictable)
dog. (Unpredictable)

[^213]The Predictable Harry Potter sentence endings were only predictable assuming perfect knowledge of the books. Unpredictable endings were inconsistent with the books but were designed to be similarly plausible endings, assuming no knowledge of the books. A total of 216 sentence frames (108 Control, 108 HP) were constructed, each with two ending types (Predictable, Unpredictable). Two lists were constructed such that each sentence frame appeared with only one ending type. Participants therefore saw a total of 216 sentences ( 54 sentences of each type).

Experimental Procedures Participants were told they would be reading sentences for comprehension and that they would be asked questions about the materials at the end of the study. After a practice session, blocks of Control sentences were presented first, followed by blocks of HP sentences. For each sentence pair, the first sentence appeared all at once in the center of the screen. When ready, participants pressed a button to move on to the second sentence, presented one word at a time in the center of the screen with a 500 ms SOA (200 ms on, 300 ms off). Immediately following the ERP study, participants were given a Control memory quiz followed by an HP memory quiz. For each, participants saw a list of 90 words, 60 of which had appeared in sentence-final position (half Predictable, half Unpredictable). They were instructed to circle the words they remembered as an ending to the second sentence of each pair in the study. After clean-up, participants completed 10 multiple-choice questions about the Harry Potter books. Raw scores are henceforth referred to as HP Knowledge. A median split on these scores determined High- and Low-Knowledge Groups.

ERP Recording and Data Analysis The electroencephalogram (EEG) was recorded from 26 tin electrodes geodesically arranged in an ElectroCap, with impedances kept below $5 \mathrm{~K} \Omega$. Recordings were referenced online to the left mastoid and re-referenced offline to an average of the left and right mastoids. EEG was recorded by Grass bioamplifiers with a bandpass of $.01-100 \mathrm{~Hz}$ at a sampling rate of 250 Hz . Trials contaminated by artifacts (e.g., eye movements or blinks) were not included in analyses.

Grand average ERPs to sentence-final words were computed across all 26 recording sites by Sentence Type (Control/HP) and Ending Type (Predictable/Unpredictable). We performed statistical analyses on mean amplitudes of these waveforms in two time periods: a canonical N400 time period ( $250-500 \mathrm{~ms}$ ) and a post-N400 period ( $500-750 \mathrm{~ms}$ ) relative to a 500 ms pre-stimulus baseline. For each time period, we subjected data to an omnibus ANOVA including Channel ${ }^{1}$ ( 26 levels), Sentence Type (Control, HP), Ending Type (Predictable, Unpredictable) as within-subjects factors and Knowledge Group as a between-subjects factor. Subsequently we focused on a region of interest (ROI) including 8 centro-parietally distributed channels (MiCe, LMCe, RMCe, MiPa, LDPa, RDPa, LMOc, and RMOc).

Channel in the N400 region reflect the fact that N400 amplitude (and N400 effects) are largest over the middle and back of the head.

## ALL PARTICIPANTS HIGH HARRY POTTER LOW HARRY POTTER KNOWLEDGE <br> KNOWLEDGE

GENERAL KNOWLEDGE SENTENCES (CONTROL)


Figure 1: ERPs from a central-parietal ROI to sentence-final critical words from Experiment 1 are plotted for Predictable (black lines) and Unpredictable (red lines) endings relative to a 200 ms baseline for illustrative purposes. Shaded regions depict the area between 250 and 500 ms (N400 time window). All participants showed Predictability effects for Control sentences while Predictability effects for HP sentences were driven by the High-Knowledge group.

## Results

Memory task Participants correctly recognized an average of 15 out of 60 Control words ( $25 \%$ ) and false alarmed to an average of 2 words (7\%). On the HP recognition test, participants correctly recognized an average of 30 out of 60 HP words ( $50 \%$ ) and false alarmed to an average of 2 words (7\%). Participants were therefore able to discriminate between words they had and had not seen for both the Control and HP memory tests.

To control for false alarms, we subtracted the number of false alarms for each memory test (Control, HP) from the number of items correctly recognized. We subjected these to a repeated measures ANOVA with Sentence Type (Control, HP) and Ending Type (Predictable, Unpredictable) as factors. There was a main effect of Sentence Type ( $p<.0001$ ), with higher accuracy for HP compared to Control sentences. There was also an interaction between Sentence Type and Ending Type ( $\mathrm{p}<.001$ ); while memory for HP words was similar irrespective of the Ending Type (corrected accuracy for Predictable $=44 \%$; corrected accuracy for Unpredictable $=$ $40 \%$ ), memory for Control words was better for Unpredictable words (22\%) compared to Predictable words (15\%).

As predicted, HP knowledge was not correlated with accuracy for Control words, but HP knowledge was correlated with accuracy for HP words (Predictable: $r=.471$, $\mathrm{p}<.005$; Unpredictable: $r=.478, \mathrm{p}<.005$ ).

ERPs ERPs from our centro-parietal ROI are shown in Fig. 1. ERPs for both Control and HP sentences are characterized by two early sensory components, a negative-going peak around 100 ms (N1) and a positive-going peak around 200 ms (P2). Across all participants, for Predictable endings, the P 2 is followed by a positivity in the N 400 time window
( $\sim 250-500 \mathrm{~ms}$ ). For Unpredictable endings, the P 2 is followed by a relative negativity in this window.

Effects of knowledge during the N400 time window. Our primary hypothesis was that specific domain knowledge would influence semantic access, reflected by interactions between Knowledge Group, Sentence Type, and Ending Type during the N400 time window ( 250 to 500 ms ). In the omnibus ANOVA, we observed a main effect of Ending Type ( $\mathrm{p}<.005$ ) but no effect of Sentence Type, reflecting the pattern observed in Fig. 1: for both sentence types, N400 amplitude is reduced for Predictable items. Of note, Knowledge Group interacted with Sentence Type ( $p<.005$ ), and a three-way interaction was observed between Knowledge Group, Sentence Type, and Ending Type (p < .05).

To follow up on the effects of Knowledge Group, Sentence Type, and Ending Type on N400 amplitude, we examined Control and HP sentences separately at our centro-parietal ROI. Within Control sentences, there was an effect of Ending Type on N400 amplitude ( $\mathrm{p}<.0001$ ) but no main effect of Knowledge Group or interaction between Knowledge Group and Ending Type. Conversely, within HP sentences, there was a main effect of Ending Type ( $p<.0001$ ), and an interaction between Knowledge Group and Ending Type (p $<$ .005). Follow-up t-tests revealed a larger reduction in N400 amplitude for Predictable versus Unpredictable endings for the High-Knowledge Group compared to the LowKnowledge Group ( $\mathrm{p}<.01$ ), supporting the notion that specific knowledge at the level of the individual reduces N400 amplitude (see Fig. 2). Furthermore, the differential knowledge had its effect primarily for Predictable endings, as High-Knowledge and Low-Knowledge individuals showed differences in N400 activity to Predictable endings ( $\mathrm{p}<.05$ ) but similar N400 activity to Unpredictable endings ( $\mathrm{p}=.630$ ).

GENERAL KNOWLEDGE SENTENCES (CONTROL)

HARRY POTTER
SENTENCES



Figure 2. Difference ERPs for Unpredictable minus Predictable endings from Experiment 1 are plotted for the High HP Knowledge Group (solid lines) and the Low HP Knowledge Group (dashed lines) relative to a 200 ms baseline. Predictability effects for Control sentences were similar for both groups but only the High HP Knowledge Group showed sizable Predictability effects for HP sentences.

Analysis involving our continuous measure of HP Knowledge coincided with this pattern of results. We observed a graded relationship between the N400 effect (mean amplitude to Unpredictable minus Predictable endings) and knowledge scores, $r=.457, \mathrm{p}<.005$; this relationship was driven by the correlation between knowledge and mean amplitude to Predictable endings ( $r=$ $.473, \mathrm{p}<.005$ ) whereas no correlation obtained between knowledge and mean amplitude to Unpredictable endings ( $r$ $=.171, \mathrm{p}=.293$ ).

Effects of knowledge post-N400 (500-700 ms). In the omnibus ANOVA, we observed an interaction effect of Sentence Type and Ending Type ( $p<.01$ ), which resulted from a significant difference between Predictable and Unpredictable endings to Control sentences ( $\mathrm{p}<.0001$ ), with Unpredictable endings associated with greater positivities, but only a marginal difference between Predictable and Unpredictable endings for HP sentences, with the reverse pattern ( $\mathrm{p}=.08$ ). Apart from effects involving Channel, there were no other main effects or interactions in this analysis. ${ }^{2}$

## Experiment 2

Experiment 1 demonstrated that specific domain knowledge influences real-time semantic access, inferred from N400 predictability effects. This pattern could obtain for multiple reasons. By definition, experts know more information, but expert knowledge also may be functionally organized differently, with greater structure and/or depth than that of less-knowledgeable individuals (Chi, 2006). We expect that this organization, in whichever form it may take, may influence semantic access above and beyond the successful retrieval of any given known item.

[^214]To tease apart contributions of (1) knowledge of individual items and (2) knowledge of the domain (Harry Potter) to semantic access, we asked participants to read sentences, all of which were consistent with the world of Harry Potter, and to respond with judgments of their knowledge along with their confidence in them. We were particularly interested in whether domain knowledge might have independent effect on semantic activation (inferred from N400 amplitude) for (a) items people say they know, (b) items people say they don't know, or (c) both.

## Methods

Participants 41 right-handed students at UCSD participated for partial course credit and some monetary compensation.
Sentence Materials Materials consisted of 172 sentence pairs describing the world of Harry Potter, including the 108 from Experiment 1 plus an additional 64. All sentences ended in a word consistent with the Harry Potter books (i.e., the Predictable endings from Experiment 1).

Experimental procedure Sentence presentation was as in Experiment 1. After each sentence pair was presented, participants were first asked to make a non-speeded judgment about whether they knew the information in the sentences ahead of time, followed by a judgment of their certainty (we report only on responses to the first question in this paper).

After clean-up, participants completed 40 multiple-choice questions about the Harry Potter books (including the 10 questions used in Experiment 1). Raw scores are henceforth referred to as HP Knowledge. A median split on these scores was used to determine High- and Low-knowledge groups.

ERP recording and data analysis ERPs were recorded and processed as in Experiment 1. Because the design of Experiment 2 involves binning data based on subject responses, we used mixed-effects models (Baayen et al., 2008), which allow for the analysis of unbalanced data (e.g., Tibon \& Levy, 2015). For both N400 and post-N400 time windows, we start by employing models that include fixed effects of (1) Judgment of Knowledge (two levels: "Yes," "No") and (2) HP Knowledge (continuous measure) along with random by-items and by-subjects intercepts. To unpack interactions, we follow up with similar mixed-effects models designed to isolate the $\operatorname{root}(\mathrm{s})$ of the interactions. These models were applied to data from our centro-parietal ROI (see Experiment 1). As in Experiment 1, we examined a window centered around the N400 $(250-500 \mathrm{~ms})$ and a postN400 window ( $500-750 \mathrm{~ms}$ ). For illustrative purposes, when plotting these data, we weight trials equally (rather than plotting grand averages as is typical, where each subject is weighted equally).

[^215]

Figure 3: (A) ERPs from a central-parietal ROI to sentence-final critical words from Experiment 2 are plotted for words judged as known (black lines) and unknown (red lines) relative to a 200 ms baseline. Shaded regions depict the area between 250 and 500 ms (N400 time window). Across all participants, words judged as known led to more positive-going waves during this time. (B) During the N400 time window, HP Knowledge influenced mean amplitude only for Unknown (red), but not known (black) items.

## Results

Behavior On average, participants responded that they knew 102 out of 172 items ( $60 \%$ ). As expected, high-knowledge participants reported knowing more items (80\%) than lowknowledge participants (38\%), with a strong correlation between HP Knowledge and number of items judged as known, $r=.85, \mathrm{p}<.0001$.

We trimmed response times three standard deviations greater than the mean for all responses. Response times for judgments of knowledge were overall slower for "No" responses ( 1015 ms ) than "Yes" responses ( 851 ms ), p < .0001. Moreover, HP knowledge interacted with Judgment of Knowledge ( $\mathrm{p}<.0001$ ): high-knowledge individuals responded faster for "Yes" responses ( 809 ms ) than "No" responses ( 1193 ms ) ( $\mathrm{p}<.0001$ ), but Low-Knowledge individuals showed only slightly faster RTs for "Yes" (943 ms ) than "No" ( 964 ms ) (n.s.). Pair-wise differences between High- and Low-Knowledge Groups were significant for "No" responses ( $\mathrm{p}<.005$ ) but not for "Yes" responses. Individuals therefore responded with similar speed when they judged items as known, but those with greater HP knowledge took longer to judge an item as unknown.

ERPs Grand average ERPs to sentence-final words were computed across all 26 recording sites grouped by participants' responses. See Fig. 3 for plots from the centroparietal ROI.

Effects of knowledge during the N400 time window. As expected, we observed overall more positive-going waveforms during the N400 time window for highknowledge compared to low-knowledge individuals (p $<$
.005). In addition, positive Judgments of Knowledge (i.e., "Yes" responses) resulted in reduced N400 amplitudes (p $<$ .0001; see Fig. 3). We also observed an interaction of Judgment and HP Knowledge ( $\mathrm{p}<.005$ ). Follow-up comparisons revealed that this interaction was driven by effects of HP Knowledge on N400 amplitude for "No" responses ( p .05 ) but not for "Yes" Responses, demonstrating that specific domain knowledge has its primary influence on items which participants say they did not know (recollect) at the time.

Effects of knowledge post-N400 (500-750 ms). Mean amplitude during the post-N400 window was influenced both by HP Knowledge ( $\mathrm{p}<.05$ ) and Judgments of Knowledge ( p $<.0001$ ), with the two terms also interacting ( $\mathrm{p}<.0001$ ). Overall, "Yes" responses yielded more positive-going waves than "No" responses, and greater HP Knowledge was also related to more positive-going potentials. Follow-up analyses revealed that this interaction was driven by an effect of HP Knowledge on "No" responses ( $\mathrm{p}<.05$ ), with no relationship between HP Knowledge and positive-going potentials for "Yes" responses $(p=.869)$. For "No" responses, individuals with more knowledge had more post-N400 positive-going activity.

## General Discussion

We set out to investigate the relationship between specific domain knowledge and semantic access during real-time written sentence processing. Experiment 1 provided a strong indication that knowing about a domain (in our case, the narrative world of Harry Potter) influences semantic access, but only within that domain. As predicted, we observed no effects of Harry Potter-specific knowledge on processing of sentences about general topics. However, Harry Potterspecific knowledge did mediate N400 effects for Harry Potter sentences. We found that the size of the N400 predictability effect was correlated with HP knowledge score, with the correlation being driven by a graded relationship between knowledge and the neural response to predictable words.

In many ERP studies of sentence processing, predictability is defined using offline Cloze norming measures (that is, how likely an individual is to provide a word given a sentence context). It is worth noting that offline Cloze measures for our Harry Potter sentences provide a different type of metric than for our Control sentences. That is, predictable endings for Harry Potter sentences are predictable by virtue of being factual (within the narrative world); predictable endings for Control sentences are predictable based on world knowledge, but have no "correct" ending. Even so, our analyses revealed no main effect of sentence type (general/control vs. Harry Potter sentences) in Experiment 1. That is, across the whole group, we observed similar N400 effects of predictability for both Control and Harry Potter sentences (see Fig. 1). Our findings concur with many reports that N400 amplitude is sensitive to a word's predictability and/or contextual fit, in factual and non-factual scenarios (e.g., Hagoort et al., 2004).

Our finding that semantic access (inferred from N400 effects) is driven by knowledge is not surprising. In order to
access information, the information must exist in the first place. However, there are at least two reasons why HP knowledge might relate to N400 predictability effects in Experiment 1: (1) low-knowledge individuals know fewer facts than high-knowledge individuals on average; and (2) there are potentially additional contributions of domain knowledge on semantic processing when individuals know (or don't know) items, respectively. In Experiment 2, we examined these possibilities by asking participants whether they knew each item (i.e., each Harry Potter fact) ahead of time. While both high- and low-knowledge groups showed large differences in N400 activity based on their own (metacognitive) judgments of whether they knew specific items, high-knowledge (compared to low-knowledge) individuals showed greater positivities during the N 400 window even for information they reported not knowing at the time.

There are multiple reasons why domain knowledge might modulate N400 amplitude for items that are not immediately recognized. We cannot currently rule out the possibility that high- vs. low-knowledge individuals perceive different task demands or use different criteria when making judgments.

We believe a more likely explanation is that enhanced N400 reduction for high- compared to low-knowledge individuals suggests some level of implicit activation of information outside of conscious awareness. This activation may be restricted to a specific word and its semantic features or it may extend to related words / concepts. The precise nature of such implicitly activated information, and precisely how it is modulated by variation in level or amount of knowledge, have yet to be determined. Some possibilities include information that is taxonomically / categorically related to a predictable word (e.g., Federmeier \& Kutas, 1999) and information related to the scenario / event being described (e.g., Metusalem et al., 2012).

As for post-N400 activity, we observed systematic effects of both HP Knowledge and judgments of knowledge on late positivities in Experiment 2, with high-knowledge individuals showing greater positivities than low-knowledge individuals for items they did not know. One way of interpreting this interaction is that when high-knowledge individuals do not know an item, they continue to search for it (perhaps because they believe they may know, but not currently be able to retrieve, the knowledge).

Our findings build on work showing that the functional organization of long-term memory plays an important role in the real-time construction of meaning (e.g., Federmeier \& Kutas, 1999). We have demonstrated that variability among individuals in their knowledge of a domain is an important contributor to real-time access to meaning. More specifically, our data suggest that the amount and/or organization of domain knowledge appear to influence access to knowledge above and beyond explicit knowledge of individual items: expert-like knowledge organization in a domain may lead to implicit or partial activation of domain-related information, even when individuals do not explicitly recall a given piece of information.

## Acknowledgments

This work was supported by NICHD grant HD22614 to Marta Kutas and UCSD Kroner Fellowship and FISP grant to Melissa Troyer. The authors would like to thank Seana Coulson, Katherine DeLong, Jeff Elman, Vic Ferreira, Zhuowen Tu, and Thomas Urbach for helpful feedback.

## References

Baayen, R.H., Davidson, D.J., \& Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Brouwer, H., Fitz, H., Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. Brain Research, 1446, 127-143.
Chi, M.T.H. (2006) Laboratory methods for assessing experts' and novices' knowledge. In Ericsson et al. (Eds.), The Cambridge Handbook of Expertise and Expert Performance. Cambridge University Press.
Federmeier, K.D. \& Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. Journal of Memory and Language, 41, 469-495.
Federmeier, K.D., Kutas, M., \& Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. Brain \& Language, 115, 149-161.
Filik, R. \& Leuthold, H. (2013). The role of character-based knowledge in online narrative comprehension: Evidence from eye movements and ERPs. Brain Research, 1506, 94-104.
Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M. (2004). Integration of word meaning and world knowledge in language comprehension. Science, 304, 438-441.
Kamide, Y., Altmann, G.T.M., \& Haywood, S.L. (2003). The timecourse of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language, 49, 133-156.
Kutas, M. \& Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences, 4(12), 463-470.
Kutas, M. \& Hillyard, S.A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. Science, 207, 203205.

Metusalem, R., Kutas, M., Urbach, T.P., Hare, M., McRae, K., \& Elman, J.L. (2012). Generalized event knowledge activation during online sentence comprehension. Journal of Memory and Language, 66(4), 545-567.
Nieuwland, M.S. \& Van Berkum, J.J.A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. Journal of Cognitive Neuroscience, 18(7), 1098-111.
Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., \& Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.
Tibon, R. \& Levy, D.A. (2015). Striking a balance: analyzing unbalanced event-related potential data. Frontiers in Psychology, 6, 555.
Van Berkum, J.J.A, van den Brink, D., Tesink, C.M.J.Y., Kos, M., \& Hagoort, P. (2008). The neural integration of speaker and message. Journal of Cognitive Neuroscience, 20(4), 580-591.
Van Petten, C. \& Luka, B.J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. International Journal of Psychophysiology, 83,176-190.

# Inhibitory Control Supports Referential Context Use in Language Production and Comprehension 

Alison M. Trude (atrude2@jhmi.edu)<br>Department of Neurology, Johns Hopkins University<br>1629 Thames St., Suite 350, Baltimore, MD, 21231 USA<br>Nazbanou Nozari (nozari@jhu.edu)<br>Department of Neurology, Department of Cognitive Science, Johns Hopkins University 1629 Thames St., Suite 350, Baltimore, MD, 21231 USA


#### Abstract

Using referential context in language (e.g., saying "blue pen" when two different-colored pens are visible) makes communication efficient. But it is still unclear which general cognitive processes support the use of context in conversation. Research on pragmatic use in language implicates working memory and inhibitory control; however, no studies have shown evidence of a shared cognitive mechanism in both production and comprehension within an individual. The current study asked a) whether referential context use is supported by the same cognitive mechanisms in production and comprehension, b) which processes are implicated, and c) whether the nature of the context itself affects processing. Participants completed a referential communication eye-tracking task in which a disambiguating adjective was either necessary or over-informative, as well as a cognitive test battery. The results implicated inhibitory control in both production and comprehension (although the comprehension results were more variable), suggesting a shared underlying cognitive mechanism across domains.


Keywords: language production; language comprehension; discourse; pragmatics; inhibitory control; working memory

## Introduction

The ability to take context into account often facilitates communication in interactive settings. Imagine you are cooking with a friend. There are two identical spoons on the table: one next to a big bowl, and one next to a small bowl. You may know that you would like you friend to hand you the former; however, if you just say, "hand me the spoon next to the bowl," he will likely not understand which one you mean. Thus, in order to effectively communicate your intent, you would need to use a disambiguating adjective (i.e., "hand me the spoon next to the big bowl"). On the other hand, if only one spoon and one bowl were visible, saying "hand me the spoon next to the big bowl" would be confusing, as it implies to your listener that there is more than one option to choose from. This paper investigates how speakers and listeners behave in situations when adding or subtracting an adjective is most appropriate for clear communication given the referential context.

Related to the idea that listeners and speakers will tailor their language to the referential context is the Gricean Maxim of Quantity (Grice, 1975), which specifies that speakers should make their utterances only as informative as is required. Thus, enough information should be provided to
distinguish the intended referent from its potential competitors (e.g., the big bowl when two bowls are visible), and providing information that is not necessary (e.g., the big bowl, when no other bowls are in view), should be avoided. Although, ideally, following this maxim would help to make communication maximally efficient, in reality, speakers and listeners and often fail to behave in a completely Gricean manner (e.g., Deutsch \& Pechmann, 1982; Sedivy, 2005).

Speakers' and listeners' ability to make their utterances optimally informative, or to appropriately interpret the utterance they are hearing, within the current referential context may depend upon the cognitive demands that this process places upon them. Referential context adaptation (RCA) is a complex process, involving not only language production or comprehension, but also selectively attending to certain objects in one's surroundings, remembering what information has already been introduced into the discourse, or refraining from mentioning irrelevant or confusing information, to name a few. Therefore, general cognitive functions, such as working memory (WM) and inhibitory control (IC), could play an important role in RCA.

The current work investigates whether the ability to take referential context into account is supported by the same cognitive processes in language production and comprehension, and under different linguistic demands. If so, we would expect (a) individuals with better RCA to demonstrate this ability in both production and comprehension, and (b) the same general cognitive operations to drive referential context consideration in comprehension and production, and perhaps in situations with different linguistic demands.

## Pragmatic Language Use and Cognitive Abilities

The fact that speakers often fail to observe the Gricean Maxim of Quantity provides evidence that taking referential context into account may be a cognitively demanding process. While research in this field has not previously addressed the relationship between RCA and general cognitive abilities, some work on a related linguistic process, perspective-taking, has attempted to identify the cognitive processes underlying pragmatic language use. Perspective-taking and RCA are similar in that both involve the on-line incorporation of referential context in interactive conversation. Thus, it is possible that a similar set of
cognitive mechanisms supports both processes.
In one study of perspective-taking, speakers were tested on their ability to refrain from producing descriptions that were over-informative from the perspective of their listener (e.g., "the big star" when two stars are visible to the speaker, but only one is visible to the listener; Wardlow, 2013). The results showed that this ability was correlated with both WM and IC. In comprehension, Brown-Schmidt (2009) found that participants with greater IC were more likely to take the experimenter's perspective when interpreting their questions about a display containing some pictures that were visible to only the listener. Lin, Keysar, and Epley (2010) found that comprehenders with higher WM capacity performed better on a similar task.

While these studies all point to a role for general cognitive resources in perspective-taking, an important question remains: Do the same abilities underlie perspective-taking in production and comprehension? The two studies that have investigated perspective-taking in both domains within the same individuals found contradicting results: Ryskin, Benjamin, Tullis, \& Brown-Schmidt (2015) found WM to be correlated with perspective-taking in production, but none of their executive control measures were correlated with comprehension. On the other hand, Nilsen and Graham (2009) found that IC negatively correlated with egocentric behaviors in a comprehension task in 3-5 year olds; however, none of their executive control measures correlated with production. The question, thus, remains: is the ability to take into account the referential context the same in both domains?

## The Current Study

The current study investigates how speakers and listeners adapt their language processing to referential context when observing the same visual display. Specifically, we asked: is the ability to take referential context into account related in production and comprehension, and under different linguistic and contextual demands? To this end, each participant took turns as both speaker and listener in a referential communication task with a visual world design, and their eye fixations, as well as their utterances (when acting as speakers) were recorded. Two conditions with different contextual demands were created: in the Adj+ condition, the most felicitous utterance required the inclusion of an adjective (e.g., "Click on the heart under the green gorilla," Figure 1a). In the Adj- condition, the most felicitous utterance was one without an adjective (e.g., "Click on the heart under the gorilla," Figure 1b). To test which cognitive processes underlie RCA in production and comprehension, each participant also completed a battery of tasks that included three WM and three IC measures. The WM measures were selected so as to include a mixture of both linguistic and non-linguistic tasks. The IC measures were selected to probe both competitive inhibition (the ability to inhibit a strongly competing response, e.g., not saying red gorilla in Figure 1a.) and global inhibition (the ability to inhibit a prepotent response, e.g., refraining from
using an adjective in Figure 1b.) (Munakata et al., 2011). Due to the rich set of independent and dependent variables in the current design, we adopted a statistical approach that is well-suited to handling this type of data structure: partial least squares path modeling (PLS-PM). This method allows us to look at multiple dependent variables simultaneously. In addition, this method allows us to similarly group WM and IC tasks into latent constructs in order to minimize taskspecific effects and avoid the issue of collinearity.


Figure 1: Example Adj+ (a) and Adj- (b) displays used in the referential communication task.

## Methods

## Participants

Twenty-eight native English speakers, ages 18-30, participated for $\$ 40$.

## Materials

Referential Communication Task Each display consisted of four black card suit shapes (club, diamond, heart, or spade). Above each shape was a drawing of an animal. The animal stimuli were chosen to be cohort competitors with either a size or color adjective (e.g., big buffalo, green gorilla). On each trial, the target and competitor shapes were the same (e.g., both spades), ensuring that the target shape always needed to be disambiguated from the competitor by describing the animal above the target. Trials were presented in one of four fixed random orders.

There were two critical trial types. On Adj+ trials ( $n=96$ ), the same cue animal appeared above the target and competitor shapes in two different colors or sizes. Critically, the animal above the target shape was rendered in its cohortcompetitor adjective (e.g., green gorilla). These trials were designed such that, if participants used RCA, in production, they would include the adjective necessary to disambiguate the target cue from the competitor cue, and in comprehension, upon hearing the initial phoneme of the adjective (e.g., the /g/ in green), they would interpret this sound as the beginning of an adjective, and not a noun, and as a result fixate more on the cue above the target (green gorilla) than the cue above the competitor (red gorilla) (Table 1).

On Adj- trials ( $\mathrm{n}=96$ ), different cue animals appeared
above the target and competitor shapes. The cue animal above the target shape was always a cohort competitor with the adjective corresponding to the cue above the competitor shape (e.g., red gorilla vs. green raccoon). These trials were designed such that, if participants used RCA, no adjective would be produced in production, as it would be overinformative, and in comprehension, upon hearing the initial phoneme of the target noun (e.g., the $/ \mathrm{g} / \mathrm{in}$ gorilla), the participant would interpret this phoneme as the onset of a noun (gorilla), and not of the adjective preceding the competitor cue (green raccoon) (Table 1).

Table 1: Felicitous utterances on Adj+ and Adj- trials.

|  | Adj+ | Adj- |
| :---: | :---: | :---: |
| Production | Say "the heart under <br> the green gorilla." | Say "the heart under <br> the gorilla." |
| Comprehension | Upon hearing "heart <br> under g..." look for <br> "green," not "gorilla" | Upon hearing "heart <br> under g.." look for <br> "gorilla," not "green" |

## Cognitive Battery

A spatial WM task (Corsi Block, Kessels et al., 2000) and two linguistic WM tasks (Category and Rhyme Probe, Freedman \& Martin, 2001), measuring semantic and phonological WM respectively, were administered. Three IC measures were also employed: Fish Flanker along with embedded NoGo (Nozari, Trueswell, \& Thompson-Schill, 2016), as well as semantic blocking elicited through cyclic naming (e.g., Schnur et al., 2006), in which participants named twelve sets of six images, either in semanticallyhomogenous or heterogeneous blocks. Semantic blocking was determined as the difference in response latencies between the two block types.

## Procedure

Participants completed two sessions 3-7 days apart. Each session began with the eye-tracking task, followed by half of the cognitive measures. Stimuli were displayed using MATLAB and Psychophysics Toolbox (Brainard, 1997), and the participant's eye movements during the referential communication tasks were recorded using an Eyelink 1000 Plus desktop-mounted eye-tracker (SR Research).

Referential Communication Task, Session 1. On each trial, a fixation cross was presented for 500 ms before the stimuli appeared. After a 1500 ms preview period, a tone sounded and the target shape began to flash, cuing the participant to begin speaking. After 2.5 seconds, a lower tone sounded, indicating the end of the trial. The participant and experimenter viewed separate monitors containing the same stimuli. The participant instructed the experimenter to click on the target shape that was cued on the participant's screen, using sentences with this structure: "Click on the
[target shape] under the [adjective, if needed] [target cue]." They were told that the goal of the task was to provide instructions as quickly as possible, and to avoid unnecessary words to meet the temporal deadline, thus motivating them to drop the adjective when not necessary.

Referential Communication Task, Session 2. Participants followed instructions like the ones they had given during Session 1. Each trial began with a 500 ms fixation point. The stimuli then appeared onscreen, and after a 1000 ms preview period, a pre-recorded instruction played, instructing the participant to click on one of the shapes. Instructions were always pragmatically appropriate (i.e., an adjective was always provided on Adj+ trials, and no adjective was provided on Adj- trials).

## Analyses

## Behavioral

Participants made a total of 345 infelicitous utterances out of 5376 critical trials ( $6.4 \%$ error rate, $M=12.3$ errors, $S D=$ 8.5). Of these infelicitous utterances, 187 were made on Adj+ trials (i.e., adjective underuse), while 157 were made on Adj- trials (i.e., adjective overuse). Participants were expected to perform near ceiling on the comprehension task, so no accuracy measures were collected.

## Eye-Tracking

In production, eye-tracking analyses combined each card suit and its animal cue into a single region. Data was analyzed from $400-1350 \mathrm{~ms}$ post-target cuing, comprising the time period between the minimum and maximum proportion of fixations to the competitor after attention was initially drawn to the flashing target (see Figure 2a). Competitor fixations were considered critical because, in order to establish the referential context of the display, participants needed to divert their attention from the extrinsically-cued target to its competitor. Thus, our dependent measure of analysis was a competitor advantage score, or the proportion of fixations to the competitor minus the proportion of fixations to the target.

In comprehension, eye-tracking analyses were completed on the target advantage score (proportion of fixations to the target and its cue minus the proportion of fixations to the competitor and its cue) over a time window beginning 300 ms before the onset of the critical word (cue noun in the Adj- and cue adjective in the Adj+ condition) and ending 200 ms after the onset of the critical word (see Figure 2b). The time window was chosen in order to encompass coarticulatory cues from the word preceding the critical word, as well as processing of the initial cohort phoneme of the critical word. This target preference measure should indicate how well participants took referential context into account (i.e., disregarded the competitor). By using target advantage as a DV in comprehension and competitor advantage in production, we ensured that in both cases, a more positive eye-tracking score would index better RCA.


Figure 2: Time course of eye-tracking data for all critical trials in Production (a) and Comprehension (b). Highlighted regions indicate windows of analysis.

## Partial Least Squares Path Modeling (PLS-PM)

Analyses were conducted using partial least squares path modeling, implemented in the plspm package in $R$ (Sanchez, 2013). PLS-PM is a partial least square approach to Structural Equation Modeling suitable for analyzing the relationship between latent variables (psychological constructs such as WM) and manifest variables (observed data from tasks assumed to index these variables) as a network of multiple interconnected linear regressions.

Figure 3a shows our initial theoretical model. The model has three latent variables: RCA, Working Memory (WM), and Inhibitory Control (IC), each measured through a number of manifest variables. The direction of the arrows indicates the direction of causal influence.

Model quality assessment takes three general steps. The first is to verify the relationship between manifest variables and the latent variables hypothesized to underlie them. This is done by first assessing unidimensionality, or the extent to which a change in the latent variable affects all manifest variables in the same direction. Unidimensionality is indexed by Dillon-Goldstein's rho (DG rho). A DG rho above 0.7 is favorable. Second, the relative contribution of the latent variable (vs. noise/task-specific factors) to each manifest variable is calculated. Indicators with a loading of


Figure 3: Structure of initial path model (a) and revised model for production data only (b).
less than 0.6 are not good indices of the latent construct. Third, the cross-loadings of the manifest variables are checked to ensure that the loading of a manifest variable is indeed highest on the latent variable it is assumed to represent, and not on another latent variable in the model.

The first three steps are used to revise the model by dropping manifest variables or re-partitioning the latent variable constructs. The revised model is then re-checked.

Once a viable model is obtained, the overall fit is assessed, which also includes the relationship between the latent variables. $\mathrm{R}^{2}$ is reported for the latent variable of interest, and similar to simple regression models, indicates the amount of variance explained by the independent latent variables (WM and IC).

The part of the model that answers questions about the contribution of general cognitive functions on RCA is the inner model, or the links between latent variables. Its output is similar to that of any generalized linear model. Significance levels of path coefficients are estimated via bootstrapping with 1000 iterations.

## Results

## Model 1: General Model

Model 1 is based on the following theoretical assumptions: (a) the three WM tests measure a unified WM construct, (b) the three IC tests measure a unified inhibition construct, and (c) all the RCA scores measure a unified RCA construct.

Examination of this general model resulted in three main revisions: (a) Rhyme Probe was dropped because it did not contribute substantially to the latent WM construct (loading
$=0.12$ ) (b) The loadings and cross-loadings of the latent variables revealed that the IC variable had low unidimensionality ( DG rho $=.11$ ). NoGo errors had a large negative loading on the IC latent variable ( -0.93 ), while the other two IC manifest variables had positive loadings, indicating that the NoGo errors were measuring a fundamentally different construct. Thus, the IC latent variable was broken down into two variables: indirect competitive (CI; Semantic Blocking and Flanker) and global (GI; NoGo) to reflect two types of IC that have been shown to have different cognitive and neural underpinnings (Munakata et al., 2011). (c) While all measures of RCA in production had factor loadings above 0.6 on the RCA latent variable, RCA measures in comprehension had low factor loadings and low unidimensionality across Adj+ and Adjconditions. This finding points to dissociation of the RCA construct in production and comprehension and calls for separate examination of the two tasks.

## Model 2: Production

This model is shown in Figure 3b, and includes only manifest variables indexing RCA in production. RCA had a DG rho of 0.86 , and high factor loadings for all four manifest variables. This finding supports defining RCA in language production as a unified construct, regardless of linguistic demands. DG rho's for WM and CI were 0.7 and 0.63 respectively, with $\mathrm{R}^{2}$ of 0.51 .

Table 2 shows the results of the bootstrapping on Model 2. The only latent variable that significantly predicted RCA in production was GI. Post-hoc modeling of Adj+ and Adjmanifest variables separately also revealed a reliable contribution of this variable to RCA in both trial types.

Table 2: Results of Bootstrapping for Model 2. $\mathrm{CI}=$ Competitive Inhibition; GI = Global Inhibition; WM = Working Memory.

| Latent <br> IV | Path <br> Coeff. | Bootstrapping <br> Means | SE | Lower <br> $95 \%$ <br> CI | Upper <br> $95 \%$ <br> CI |
| :--- | :---: | :---: | :---: | :--- | :--- |
| WM | 0.18 | 0.20 | 0.20 | -0.24 | 0.54 |
| CI | 0.26 | 0.18 | 0.28 | -0.42 | 0.57 |
| GI | -0.56 | -0.54 | 0.18 | -0.80 | -0.21 |

Analysis of eye-tracking data in production showed that participants' average competitor advantage scores were positively correlated with RCA accuracy in both the Adj+ and Adj- condition ( $r=.45$ and .58 , respectively; $p$ 's $<.05$ ), and that eye-tracking performance itself was highly correlated across the two conditions ( $r=.96, p<.001$ ).

## Analysis of Comprehension Data

In the path model, the two measures of RCA in comprehension had opposite loadings on the RCA latent variable, thus forcing us to examine them separately. To understand why the eye-tracking measures in comprehension had opposite effects on the RCA variable,
we first examined the correlation between participants' target advantage scores in the Adj+ and Adj- conditions. The two measures showed a significant negative correlation ( $r=-.47, p=.01$ ), in contrast to the production results. Nearly half of the participants (13 of 28) showed a negative target advantage score in one condition and a positive score in the other, indicating that in one condition, participants systematically interpreted the critical cohort phoneme as the onset of the competitor, instead of the target. Recall that in the Adj+ condition, the target word was always an adjective and its cohort competitor was always a noun, while in the Adj- condition, the opposite was true. Thus, it appears that these participants adopted a strategy of always interpreting the cohort phoneme as being from the same part of speech.

Because our main question of interest was whether RCA in production and comprehension rely upon the same cognitive processes within an individual, we directly compared performance in these two domains within the subset of participants who were clearly engaging in RCA during comprehension (Adapters; $\mathrm{n}=11$ ). While no relationship was found between production and comprehension on Adj+ trials, on Adj- trials, these participants' eye-tracking performance in comprehension was significantly positively correlated with their production accuracy ( $r=.63, p<.05$ ) and marginally positively correlated with their eye-tracking performance in production ( $r=.49, p=.12$ ).

We also tested whether the same cognitive abilities were responsible for performance across conditions in Adapters. Due to the univariate nature of the dependent measures in these analyses, we ran multilevel models separately on Adj+ and Adj- eye-tracking data, with all six cognitive tests included as predictor variables. While none of the cognitive measures predicted performance on Adj+ trials, RCA on Adj- trials was significantly predicted by the Flanker ( $t=$ $6.2, p<.01$ ) and NoGo tasks ( $t=5.1, p<.05$ ).

In summary, comprehension results differentiated two groups of individuals: those who flexibly adapted their processing to referential context and those who did not. In those who did, RCA abilities were correlated in production and comprehension, at least on Adj- trials, and in both cases they were well predicted by a measure of IC.

## Discussion

This experiment tested three central questions: 1) Which cognitive processes underlie RCA? 2) Are these processes consistent across production and comprehension? 3) Do situations in which better RCA is marked by addition of an adjective differ from those in which it is marked by omission of an adjective? In answer to question 1, we observed a clear effect of IC across the domains of production and comprehension. In production, NoGo scores were predictive of better RCA for both Adj+ and Adj- trials. In comprehension, amongst Adapters, scores on both NoGo and Flanker tasks predicted performance on Adj- trials. These findings suggest a definite role for IC in RCA. Since common to both production and comprehension, this finding
most likely points to the role of IC in capturing the critical contrast by preventing fixed attention to the target (hence ignoring the critical competitor).

Two lines of evidence can be used to answer question 2. First, the fact that IC played a role in RCA for both production and comprehension is evidence that the RCA abilities are related in the two domains. Second, performance on Adj- trials was correlated between production and comprehension amongst Adapters. Together, these findings provide the first piece of evidence for a common basis of RCA in production and comprehension.

In answer to question 3, production analyses strongly suggested RCA ability was independent of trial type: Eyetracking measures across Adj+ and Adj- trials were highly correlated, PLS-PM revealed the RCA latent variable in production to be a highly coherent construct, and the same underlying cognitive mechanism, global inhibition, was implicated in both Adj+ and Adj- trials. In comprehension, more variability across conditions and participants was observed. PLS-PM showed a lack of unidimensionality across comprehension conditions, and further inspection of the data revealed that nearly half of the participants did not use referential context. Even within Adapters, only RCA on Adj- trials correlated with production and was predicted by measures of IC. This is in part due to low internal reliability of RCA in comprehension ( $\rho^{*}=.34$ for $\operatorname{Adj}-$ and $\rho^{*}=-.16$ for $\mathrm{Adj}+)^{1}$, as also reported by Ryskin et al. (2015), which stands in sharp contrast to the high split-half reliability of RCA in production ( $\rho^{*}=.97$ in both conditions). However, our results suggested that apart from consistency issues, listeners did often default to a fixed strategy, as opposed to flexibly adapting to context, as they did in production.

In summary, these results represent the first evidence for shared underlying cognitive mechanisms of pragmatic processing in production and comprehension. In production, this mechanism, global inhibition, was recruited regardless of the particulars of the referential context, while in comprehension, results were less uniform across conditions, pointing to specific strategies adopted by listeners in locating the referent. These findings provide insights into the cognitive processes that drive pragmatic use during spoken language comprehension and production, and help to situate pragmatic processing within a larger and more general cognitive framework.

## Acknowledgments

This research was supported by a Ruth L. Kirschstein NRSA Postdoctoral Fellowship (F32DC015390) to AMT \& the Therapeutic Cognitive Neuroscience Fund.

## References

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433-436.
Brown-Schmidt, S. (2009). The role of executive function in

[^216]perspective taking in online language comprehension. Psychonomic Bulletin \& Review, 16, 893900.

Deutsch, W., \& Pechmann, T. (1982). Social interaction and the development of definite descriptions. Cognition, 11, 159-184.
Freedman, M. L., \& Martin, R. C. (2001). Dissociable components of short-term memory and their relation to long-term learning. Cognitive Neuropsychology, 18, 193226.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan (eds.) Syntax and Semantics Volume 3: Speech Acts. New York: Academic Press.
Kessels, R. P., Van Zandvoort, M. J., Postma, A., Kappelle, L. J., \& De Haan, E. H. (2000). The Corsi block-tapping task: standardization and normative data. Applied Neuropsychology, 7, 252-258.
Lin, S., Keysar, B., \& Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. Journal of Experimental Social Psychology, 46, 551-556.
Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., \& O’Reilly, R. C. (2011). A unified framework for inhibitory control. Trends in Cognitive Sciences, 15, 453-459.
Nilsen, E. S., \& Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. Cognitive Psychology, 58, 220249.

Nozari, N., Trueswell, J. C., \& Thompson-Schill, S. L. (in press). The interplay of local attraction, context and domain-general cognitive control in activation and suppression of semantic distractors during sentence comprehension. Psychonomic Bulletin \& Review.
Ryskin, R. A., Benjamin, A. S., Tullis, J., \& BrownSchmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. Journal of Experimental Psychology: General, 144, 898-915.
Sanchez, G., Trinchera, L., \& Russolillo, G. (2013). plspm: tools for partial least squares path modeling (PLS-PM). R package version $0.4,1$.
Schnur, T. T., Schwartz, M. F., Brecher, A., \& Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. Journal of Memory and Language, 54, 199-227.
Sedivy, J. C. (2005). Evaluating Explanations for Referential Context Effects: Evidence for Cricean Mechanisms in Online Language Interpretation. In Trueswell, J. C., \& Tanenhaus, M. K. (eds). Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions. MIT Press.
Wardlow, L. (2013). Individual differences in speakers' perspective taking: The roles of executive control and working memory. Psychonomic Bulletin \& Review, 20, 766-772.

# A Dynamic Process Model for Predicting Workload in an Air Traffic Controller Task 

Martina Truschzinski (mtru@hrz.tu-chemnitz.de)<br>Department of Automation Technology<br>Technische Universität Chemnitz<br>Chemnitz, Germany<br>Maria Wirzberger (wirm@hrz.tu-chemnitz.de)<br>Department of E-Learning and New Media<br>Technische Universität Chemnitz<br>Chemnitz, Germany


#### Abstract

We present a dynamic process model for workload, developed according to a conducted experiment, which recorded the pupil dilation during an air traffic controller simulation. We describe how we built such a dynamic system based on the collected data. Logged events that happened in our simulation were used as system input and the recorded pupil dilation as output. Afterwards, we used the MATLAB system identification toolbox to identify the transfer function between input and output. The identified model is validated with a validation data set that has been excluded from the identification process. Results show that we are able to explain nearly $50 \%$ of the variance of the recorded pupil dilation data in the air traffic controller simulation. Moreover, the model explains some contrary results of the statistical analysis from our experiment.


Keywords: Dynamic process model; System theory; Workload; Pupillometry; Air traffic controllers

## Introduction

According to current statistics, the amount of airline passengers will continue its positive development over the next years, with expected annual growth rates of up to five percent (IATA, 2017; Boeing, 2017). To maintain the resulting needs and ensure smooth and safe traveling, the duty of air traffic controllers (ATCs) is of high importance. However, tasks like this are rather complex and put high demands on the available resources of such job holders (Mogford, Guttman, Morrow, \& Kopardekar, 1995). Beyond this, it is proven that predefined factors like traffic volume or frequency congestion influence ATCs' mental workload (Mogford et al., 1995).

As discussed by Gopher and Donchin (1986), the concept of mental workload enfolds various dimensions and facets. Although it has been broadly inspected, deriving a clear definition forms a rather difficult matter. Nevertheless, there are two constituting aspects that build a common ground in most cases. While task difficulty results from the demands required to successfully solve a task in a given time (Galy, Cariou, \& Mélan, 2012), resource supply refers to the information processing capacity available for this purpose. In this vein, mental workload comprises the difference between required capacities of the information processing system to achieve satisfying task performance and available capacity at a given time (Gopher \& Donchin, 1986; Wickens, 2008). Based on the assumption that tasks with increased difficulty require additional resources, a significant decrease in performance due
to the lack of resources should appear as soon as resource demands exceed resource supply (Wickens, G., Banbury, \& Parasuraman, 2013).

There are different possibilities to estimate human workload (Prewett, Johnson, Saboe, Elliott, \& Coovert, 2010; Beatty \& Lucero-Wagoner, 2000; Reiner \& Gelfeld, 2014) and build workload models (Gopher \& Braune, 1984; Wickens, 2008). Beatty and Lucero-Wagoner (2000) described nonreflexive phasic pupillary movements as indicators for brain processes that underlie dynamic, intensive aspects of human cognition. In several research investigating the cognitive functions, task-evoked responses of the pupil (TERPs) (Beatty, 1982; Beatty \& Lucero-Wagoner, 2000) were used to measure cognitive effort, workload and cognitive load (Haapalainen, Kim, Forlizzi, \& Dey, 2010; Wierwille, 1979; Paas, Tuovinen, Tabbers, \& Van Gerven, 2003). Therefore, in our approach we used the measured TERPs to model and validate a dynamic workload model to investigate and simulate workload in ATC tasks.

## Experiment

We collected data from 25 volunteers located at the campus of University Pompeu Fabra ( $M_{\text {Age }}=28.12$; SD $=5.67,64 \%$ male). The majority of $84 \%$ participants had no prior experience in ATC tasks (including, but not limited to, video games).


Figure 1: Experimental procedure with preparation phase, practice session and two different conditions. Measures on personality (BFI) and mood (MDMQ) are not reported in this paper.

## Experimental Design

In our 2 (workload condition) x 3 (events) within-subjects design, each participant completed a simulated ATC task. The simulation was divided into a practice session and two conditions of 4 min , a low workload condition (LWC) with a lower degree of difficulty and a high workload condition (HWC) with a higher degree of difficulty. In the LWC, participants had to manage and control less airplanes appearing at a lower frequency (airplanes appeared every $4-8 \mathrm{~s}$ ). This results in an easy task difficulty with lower time pressure, since participants had more time to handle each airplane. In the HWC, participants had to manage a greater number of airplanes appearing in a higher frequency (airplanes appeared every $1-5$ s), which resulted in higher task difficulty and time pressure to avoid collisions between airplanes.

Workload Measurement Since Beatty and LuceroWagoner (2000) reported a significant increase in pupil dilation due to an increase in workload, we recorded the pupillary response during each condition. We used the mobile eye tracking headset from Pupil Labs with a sample rate of 60 Hz and analyzed the TERPs by calculating the mean pupil dilatation during an 1.5 s window for three classes of events that were assumed to trigger an increase in workload. The time window of 1.5 s was chosen in line with Beatty and LuceroWagoner (2000), who identified that TERPs are recognized due to an increase in mental workload between 1-2 s after the presentation of a stimulus.

One class of events included all collisions caused by the participant, another one included participants actions of changing height or direction of an airplane. The third class of events included system-induced occurrences of a new couple of airplanes. All events were logged by a self-programmed protocol system, which was part of the simulated scenario. The obtained log files included timestamps for each pupillary response, which were sent to the system via a wlanconnection, as well as simultaneously recorded timestamps for each event occurrence.

## Procedure

The study was conducted in a virtual reality room, called XIM, and participants were recruited directly from the campus plaza. After completing the consent forms, they were invited to enter the XIM. For each participant, the experiment started with a preparation phase, where the eye tracking glasses were put on, the Big Five Inventory (BFI) was completed and some instructions regarding the virtual reality room were given (see Figure 1). In addition, there was a calibration phase that also ensured stable light conditions with and without planes presented on the screen. Afterwards, the practice session started, in which participants received an instruction on how to structure their commands to change airplane routes (see Fig. 2) and how to avoid collisions between airplanes. Following this instructions, participants had to manage the airplanes appearing at the screen on their own.


Figure 2: Experimental setup with eye-tracking device.

This section was finished as soon as participants were able to manage the scenario, measured by 10 correct answers in a row. After the practice section, participants were exposed to two trial sections presented in static order, the LWC followed by the HWC.
In each section, the airspace was divided in several airspace areas, whereof the subjects were responsible for the middle airspace (green rectangle). At a predefined frequency (HWC: $1-5 \mathrm{~s}$; LWC: 4-8 s) two airplanes with a given number and a random height appeared from both sides or from top and down heading to the same randomized point in the responsible airspace. During the whole experiment, participants had to keep in mind that airplanes, which do not collide in a 3D space, could appear at the same screen location due to the 2D display. This indirect 3D perception demands information processing resources as well (Wickens et al., 2013), but since all participants were exposed to comparable requirements, we did not expect additional effects on the measured level of workload. For avoiding collisions, participants had to use control commands with a similar structure compared to real ATC commands (see Figure 3). In detail, they had to provide to the number of the chosen airplane plus the information about what they want to change, for instance the direction or height of the airplane. The experimenter in the back adopted the role of the pilot, controlling the airplanes by sending special keyboard sequences to the simulation. Each session comprised a break as well. During this time span, participants reported their mood state by completing the Multidimensional Mood State Questionnaire (MDMQ). After finishing the LWC and HWC, the experimenter removed the eye tracker glasses.

## Data Analysis and Model Preparation

After conducting the experiment and preparing the pupil dilation data, we analyzed the data statistically and computed a dynamic model of workload over the task with an identified and fitted dynamic system (Isermann \& Münchhof, 2011). With reference to the latter, we developed some hypothetical assumptions based on the curve progressions of figure 6 in Beatty (1982). We assumed that dealing with appearing airplanes and setting a command will increase participants' workload (see Fig. 4). If a collision happened, we expected participants to immediately recognize their mistake and think about. However, at the same time the complexity of the air space should be reduced due to the reduced number of air


Figure 3: Schematic representation how participants had to change the height of an airplane.
planes on the screen. On this account, we assumed an initial increase in workload after collisions, directly followed by a decrease caused by the reduced amount of airplanes.

## Data preparation

To calculate the mean pupil dilation, the recorded data had to be cleaned from artifacts, blinks and other undesired patterns in the data stream (Beatty \& Lucero-Wagoner, 2000). Therefore, we used MATLAB-functions to implement standard methods for cleaning and analyzing pupil dilation data. First, we deleted all blinks in the signal, which are characterized by zero values in the data stream. Then, we interpolated the missing values and used a median filter in order to clean the signal from outliers. Participants with more than $18 \%$ blinks or zeros in the data stream were excluded from the statistical analysis, as the filtering functions and the evaluation could be falsified by very noisy signals. For the statistical approach, we calculated the respective level of workload for the events "collision", "disappear" and "appear" as mean of all occurred TERPs after the system had logged the collision or the appearance of airplanes. Due to the fact that our simulation only recognized if an airplane changed its direction or height, we measured the level of workload for the event "action" from TERPs calculated during an 1.5 s window before the change happened. Within our statistical analyses, we calculated the mean of the pupil dilation of LWC and HWC as measure of workload for the particular condition.

## Statistical Results

We conducted a repeated measures analysis of variance (ANOVA), to validate TERPs as predictor for workload. Event ("collision" vs. "appear" vs. "disappear") and workload condition (LWC vs. HWC) were regarded as independent variables and the recorded TERPs as indicator of workload were defined as dependent variable. Mauchly's test indicated a violation in the assumption of sphericity for the main effect of event, $\chi^{2}(5)=73.049, p<.001$, as well as the interaction between condition and event, $\chi^{2}(5)=$ $42.331, p<.001$, thus degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity for event, $\varepsilon=.388$, and the interaction, $\varepsilon=.573$. We found a significant main effect for event, $F(1.17,19.79)=12.394, p<.05, \eta_{p}^{2}=$ 0.42 , but no significant main effect of workload condition. Post-hoc pairwise comparisons with Bonferroni correction


Figure 4: Assumed schematic curve progression of TERPs during different events.
pointed out that the workload after an action $(p<.001)$ and after a collision ( $p<.05$ ) was significantly higher than the average workload in the whole condition. However, the workload after the appearance of an airplane was significantly lower ( $p<.05$ ). Moreover, a significant interaction effect between condition and event showed up, $F(1.72,29.24)=$ $3.701, p<.05, \eta_{p}^{2}=.18$, indicating a difference in the workload between events in both conditions. Since we had a static order, the level of workload in the HWC could have been influenced by the LWC. To control for this effect, we computed pupil dilation means during a 1 s window at the beginning of the LWC and HWC and conducted a paired-samples t-test. It did not show significant results, $t(19)=-1.927, p>.05$, thus we can assume that there was no influence on workload evoked by the static order.

## Workload Model

We assume that different levels of workload in both conditions result from task difficulty and the different events corresponding to the behavior of simulation and participants. Therefore, each simulated event as well as the spoken commands should have a direct influence on the level of workload in each condition, resulting in different TERPs. Thus, in the dynamic approach, the pupil dilation as level of workload is described as output that is dependent on the events, which are described as inputs. If there was a stable unique relation between input and output, we should be able to find a mathematical model for the temporal behavior of the workload (TERPs) from the measured input of the events.

## System Description

In system theory, such model can be described as multiple input and single output model (MISO), at which appearance, disappearance and collisions of airplanes as well as actions commanded by the participants are inputs, whereas the obtained TERPs are regarded as output. In detail, we measured a continuous time signal, with pupillary response and the related events as impulse responses appearing in each condition. An impulse response can be defined as the output of a process being excited by an impulse $(\delta(s)$ ) (Isermann \& Münchhof, 2011).

$$
\delta(t)=\left\{\begin{array}{lll}
\infty & \text { for } \quad t=0  \tag{1}\\
0 & \text { for } \quad t \neq 0
\end{array}\right.
$$

For a better understanding, while the step response or impulse response can be measured easily in many cases, we modeled our input events as step functions, whereas a step $(\sigma(s))$ can be obtained by integrating the impulse with respect to time $t$ (Isermann \& Münchhof, 2011).

$$
\sigma(t)= \begin{cases}1 & \text { for } \quad t \geq 0  \tag{2}\\ 0 & \text { for } \quad t<0\end{cases}
$$

To estimate important system parameters, such as settling time or the the damping coefficient and other characteristic values, we can use the following generic transfer function

$$
\begin{equation*}
G(s)=\frac{y(s)}{u(s)}=\frac{b_{0}+b_{1}(s)+b_{m-1}(s)^{m-1}+b_{m}(s)^{m}}{a_{0}+a_{1}(s)+a_{m-1}(s)^{m-1}+a_{m}(s)^{m}} \tag{3}
\end{equation*}
$$

which is the Laplace transformation of an ordinary differential equation (ODE) for a lumped parameter system (for further details see Isermann and Münchhof). Since we model our input data with step responses, we can directly take some individual characteristic values from the calculated step response of the system, which might be used to determine coefficients of special transfer functions by means of simple calculations (Isermann \& Münchhof, 2011). With the system identification toolbox, MATLAB offers a great database of identification methods to solve such process identification problems. Therefore, we used MATLAB to identify our workload model based on the data we collected during the experiment.

Modeling the input For modeling the system input, we used the timestamps of appearances and disappearances of airplanes as well as collisions and actions within the tasks from participants' $\log$ files. Based on this, we created several time series for each event class, which contained a step response at each event timestamp recorded by the simulation. Since airplanes stayed on the screen till they disappeared, we had to take into account that appearance and disappearance of airplanes have a different influence on the resulting workload compared to commands and collisions. Thus, the $\sigma$-function of appear is increased by two if an airplane couple appeared on the screen and the $\sigma$-function of disappear is increased by the number of airplanes which left the air space unharmed. By contrast, the influence of actions and collisions lasted only a limited time (an action during the time the participant spoke and the collision as long as the collision sound was played and the airplanes disappeared). Therefore, the $\sigma$-function of action was set to 1 for the time frame of 2 s before an action happened (see Fig. 5(a)). We chose these time window because the middle duration of commands was 2 s . Due to the fact that an airplane collision reported relatively short by a collision sound, we modeled the $\sigma$-function of collision by setting it to 1 for the time frame of 1 s after a collision happened (see Figure 5(b)).

Identify the types of transfer functions for events After modeling the input, we analyzed the behavior of the pupil


Figure 5: Example step response function of all events.
dilation over 4 s time windows during the recorded events. The chosen time windows doubled the recommended 2 s time window (Beatty \& Lucero-Wagoner, 2000), as for identifying the dynamic of the system we had to ensure that the eventrelated response was included even with potential reaction time differences between participants. Moreover, the additional information of the signal behavior could help to find the right time constants.

In system theory, there exist several LTI(Linear-Time-Invariant)-systems, which describe different patterns of behavior in signals with linear ODEs. In the mathematical view, this behavior is described with the transfer function $G(s)$, which describes how a step response(s) influences the output signal. For example, a transfer function for an $P T_{1}$-system can be described by

$$
\begin{equation*}
G(s)=\frac{y(s)}{u(s)}=\frac{b_{0}}{1+a_{1} s}=K \frac{1}{1+T_{1} s} \tag{4}
\end{equation*}
$$

in which $P T_{1}$-system results depend on an step response in an increase of K during the time $T_{1} / \mathrm{T}$, whereas T is the sampling rate of the signal. Thus, the time constant $T_{1}$ describes how fast the signal reached the value K. In terms of the workload description such a system would describe how the workload will be influenced over time. Whereas $T_{1}$ describes how fast the workload is increased and $K$ describes the absolute increase or decrease of the workload after an event is recorded. A more detailed view and the explanation of all types of LTIsystems are described in Isermann and Münchhof (2011).

Since we focused on developing a general model for each event and there might be some disturbing influences within recorded TERPs, we calculated the mean of all TERPs and regarded this as a baseline within our identification process. Such disturbing influences could be seen in miscalculated workload within overlapping, unrecognized or overwhelming events. By calculating the mean progression of the TERPs


Figure 6: Curve progression of TERPs at different events.
during the events, we assume that disturbing influences might distracted from the characteristic behavior. In Figure 6, we show the behavior of the TERPs based on this means (black lines in each figure). Moreover, it outlines that we identified transfer functions for the each step function of several events, therefore the mathematical description of our transfer function could be seen as the mathematical description of our TERPs depending on event inputs (colored lines). We identified a $P T_{2}$-system for the action, which shows a short initial decrease in workload, followed by a steep increase. The under-dumped $P T_{2}$-system with a death time for the behavior of collision, shows that there was no significant increase in workload after a collision but a decrease after $0.5-1.0 \mathrm{~s}$. The identified system of appear comprises a $D T_{2}$-system with a death time, which shows a significant increase in workload during an spoken command. The signal of disappearance reveals that the reaction of this event is very small (signal range is between 0.08 and -0.05 ). Potential reasons might be the lack of reaction in pupil dilation to this event or an ineptly small size of the chosen time window for identifying a significant change. Thus, we have to handle the identified $P T_{2} Z$ system with death time carefully, since it might be incorrect. Of course, these identified models are "ideal" models to the mean behavior of the TERPs, but they can provide a hint on the type of underlying system and a clue for the range of the used time constants. Such applies in particular for collision and action, since these events are most likely to trigger direct and fast input-response behavior.
Identify the overall system behavior Based on the identified dynamic system for the TERPs, we aimed to identify the

Table 1: Identified parameters of the mean curve progression of all event-based TERPs.

| System | K | $T_{1}$ | $T_{2}$ | $T_{z}$ | $T_{d}$ | $T_{w}$ | $\zeta$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Action | 1.25 | 1.84 | 1.59 |  |  |  |  |
| Appear | .51 | .52 | .52 | -1.4 | .59 |  |  |
| Collision | -10 | 1.41 |  |  |  | .96 | .89 |
| Disapp | -.2 | 4 | 4 | -10 | .66 |  |  |


(a) LWC of participant 8.

(b) LWC of participant 25 .

Figure 7: Comparison of recorded and modeled pupil dilation for participants 8 (training data) and 25 (validation data).
underlying dynamic system behavior of pupil dilation for the whole conditions. We assumed that the measured pupil dilation reflects the sum of responses to the input. For validating our model, we divided the data in a training data set, containing 16 participants ( $80 \%$ of the sample), and a validation data set, containing four participants ( $20 \%$ of the sample). Data of LWC and HWC were represented as independent experiments and contained the time series of the events (see Figure 5) and the corresponding recorded measurements of the pupil dilation. We defined the types of systems and their possible range of parameters detected in the TERP-analysis as system structure, to identify the complete model of pupil dilation behavior. Afterwards, we used the system identification toolbox to identify the best model describing change of pupil dilation over time depending on the event inputs. Figure 7(a) shows the simulated and the experimental output of the dynamical pupil dilatation system of two examples of the training and validation data set. We see the typical increase of the pupil dilation during the conditions of participants 8 and 25 (black lines) and the corresponding model outputs (red lines) with their exponential curve progression. The goodness-of-fit is calculated by the normalized root mean square error $\left(P_{8}=46.06 \%, P_{25}=45.68\right)$ and shows that the peaks in the pupil dilation are the result of TERPs from actions and the dips are the result of TERPs from happened collisions. The fitted model parameters for different transfer functions are displayed in table 2. As expected, in the fitted time constants of appear and disappear event-related influences are very slow ( $T_{1 a p}=104.2 ; T_{1 d i}=211.2$ ), compared to action and collision ( $T_{1 a c}=1.72 ; T_{1 c o}=1.5$ ). Furthermore, the absolute influence of an action and a collision ( $K_{a c}=2.77$; $K_{c o}=-9.58$ ) to the workload is greater than the appearance and disappearance of airplanes ( $K_{a p}=0.93 ; K_{d i}=-1.0$ ).

Table 2: Identified parameters of event-based TERPs based on pupil dilation data of the whole conditions.

| System | K | $T_{1}$ | $T_{2}$ | $T_{z}$ | $T_{d}$ |
| :--- | :---: | :---: | :---: | :---: | :--- |
| Action | 2.77 | 1.72 | 1 |  |  |
| Appear | .93 | 104.2 | .61 | -.62 | .6 |
| Collision | -9.58 | 1.5 | .96 |  | 1.14 |
| Disappear | -1 | 211.2 | 55.65 |  | .74 |



Figure 8: Comparison of experiment and fitted model data

Validation of the model The model described above was identified by data from the LWC and HWC of the first 16 participants. To measure how good our model can represent the recorded data of participants the model never seen, we calculated the deviation between model and experiment data over all participants and conditions by the normalized root mean square error (NRMSE) and the normalized coefficient of determination $\left(N R^{2}\right)$. In Figure 8 shows the goodness-of-fit-results of our training data set and the validation data set. For the NRMSE, we reached mean values of $28.87 \%$ for the validation data set and $32.99 \%$ for the training data set. If we look at the $N R^{2}$, the model is able to explain $49.42 \%$ of variance in the validation data and $55.10 \%$ in the training data.

## Discussion

We developed and identified a dynamic model for the TERPs within a simulated ATC scenario. Corresponding to our expectations, statistical analyses show a significant increase in participants' TERPs due to collisions and actions, indicating metacognitive reflections about commands or mistakes. Contradictory results show up with a significant decrease in TERPs after the appearance of a couple of airplanes that afterwards increases again (see Figure 6(b)). These effects are very slow and the sole calculation of state based statistics is prone to loose this information. On this account, we built and validated a dynamic model to predict workload of ATC-Tasks based on the experimental results. We used different models for each event logged in the session, and thus can conclude that not each visual input provides the same TERPs (Beatty \& Lucero-Wagoner, 2000). Furthermore, we show that the resulting workload in our condition is the sum of the responses of our system to the events. However, the increase is not a straight line, but rather an exponential increase, which might occur as well in similar experiments that investigate workload. Moreover, we can conclude that there is a stable unique relation between events in the simulation and the resulting TERPs, as we were able to find a mathematical model for the temporal behavior of the pupil dilation. Still, this model is just an approximation of the dynamic processes of workload that might be limited by the underlying linear process model. Nevertheless, we were already able to explain and predict nearly $50 \%$ of the variance in the resulting workload.

## Further steps

In the next instance, we will conduct another experiment with a duration of 7 min and two conditions, an emotional and a neutral session. In this vein, we can validate our identified workload progression for the extended time frame and furthermore investigate how the emotional influence in the second condition changes the dynamics of our model. Based on these results, we will extend our model by an emotional component, simulating and predicting the influence of emotions to the workload and TERPs.

## References

Beatty, J. (1982, March). Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, 91(2), 276-292.
Beatty, J., \& Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, \& G. G. Berntson (Eds.), Handbool of Psychophysiology (2nd ed.). Cambridge: Cambridge University Press.
Boeing. (2017). Estimated annual growth rates for passenger and cargo air traffic from 2016 to 2035, by region. in statista - the statistics portal. Retrieved from https://www.statista.com/statistics/269919/ growth-rates-for-passenger-and-cargo-air- traffic
Galy, E., Cariou, M., \& Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? International Journal of Psychophysiology, 83, 269-275.
Gopher, D., \& Braune, R. (1984). On the Psychophysics of Workload: Why Bother with Subjective Measures? Human Factors: The Journal of the Human Factors and Ergonomics Society, 26, 519-532.
Gopher, D., \& Donchin, E. (1986). Workload - An examination of the concept. In K. R. Boff, L. Kaufmann, \& J. P. Thomas (Eds.), Handbook of Perception and Human Performance. Vol. II. Cognitive Processes and Performance. New York: Wiley \& Sons.
Haapalainen, E., Kim, S., Forlizzi, J. F., \& Dey, A. K. (2010). Psycho-Physiological Measures for Assessing Cognitive Load. Copenhagen.
IATA. (2017). Annual growth in global air traffic passenger demand from 2005 to 2017. in statista - the statistics portal. Retrieved from https://www.statista.com/statistics/193533/ growth-of-global-air-traffic-passenger-demand
Isermann, R., \& Münchhof, M. (2011). Identification of Dynamic Systems. Berlin, Heidelberg: Springer Berlin Heidelberg.
Mogford, R. H., Guttman, J. A., Morrow, S. L., \& Kopardekar, P. (1995). The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature. (Tech. Rep.). DTIC Document.
Paas, F., Tuovinen, J. E., Tabbers, H., \& Van Gerven, P. W. M. (2003, March). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. Educational Psychologist, 38(1), 63-71.
Prewett, M. S., Johnson, R. C., Saboe, K. N., Elliott, L. R., \& Coovert, M. D. (2010). Managing workload in humanrobot interaction: A review of empirical studies. Computers in Human Behavior, 26, 840-856.
Reiner, M., \& Gelfeld, T. M. (2014). Estimating mental workload through event-related fluctuations of pupil area during a task in a virtual world. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology, 93, 38-44.
Wickens, C. D. (2008). Multiple Resources and Mental Workload. Human Factors: The Journal of the Human Factors and Ergonomics Society, 50, 449-455.
Wickens, C. D., G., H. J., Banbury, S., \& Parasuraman, R. (2013). Engineering psychology and human performance. Upper Saddle River, New Jersey: Pearson Education.
Wierwille, W. W. (1979). Physiological measures of aircrew mental workload. Human Factors: The Journal of the Human Factors and Ergonomics Society, 21(5), 575-593.

# The Dynamics of Selective Integration during Rapid Experiential Decisions 

Konstantinos Tsetsos (k.tsetsos@uke.de)<br>Department of Neurophysiology \& Pathophysiology, University Medical Center Hamburg D-20246 Hamburg, Germany


#### Abstract

When making decisions humans often violate the principles of rational choice theory. Recent experiments, involving rapid experiential decisions, uncovered a mechanism that is responsible for various rationality violations. According to this selective gating mechanism, incoming value samples are accumulated across time, but prior to their accumulation they are weighted in proportion to their momentary rank-order. Here, using a data-driven approach, I present a dynamic extension of this mechanism, which involves potentially asymmetric inhibition between the inputs. As a result, and contrary to the previous selective gating implementation, the vigour of gating is modulated by the difference between two value samples (a distance effect) as well as by the absolute magnitude of the samples (a magnitude effect). This extension offers a superior explanation to existing and new data; and links high-level decision phenomena with computational principles previously described in theories of selective attention and visual search.


Keywords: Selective integration; Experiential decisions; Risk-seeking; Intransitivity

## Introduction

Behaving organisms update their preferences in response to changes in their internal status and the state of the world; but they also do so in the absence of such changes. For instance, people are found to prefer $A$ (e.g., fresh fish) over $B$ (e.g., steak), when these two alternatives are offered, but $B$ over $A$ when a third - inferior and unchosen- alternative, $C$ (e.g., frozen fish), becomes available (Huber, Payne, \& Puto, 1982). Reversals of preference, as in the example above violate the axioms of rational choice theory (Von Neumann \& Morgenstern, 2007) and indicate that the valuation of an alternative is context-sensitive, not only dependent on the agreement between the goals of the decision-maker and the properties of the judged alternative but also dependent on the properties of other alternatives in the choice set.

Context-sensitive valuation (hereafter CSV) phenomena are reported as preference reversals (Tsetsos, Usher, \& Chater, 2010) or transitivity violations (Tversky, 1969), elicited in multiattribute choice experiments. Recently, analogues of these CSV phenomena were obtained in a rapid experiential decision task, labelled value psychophysics (Tsetsos, Chater, \& Usher, 2012; Tsetsos et al., 2016). Obtaining CSV phenomena in a psychophysical task enabled the detailed computational modelling of the involved decision processes, pointing to a selective integration model (hereafter SI) that underlies several rationality violations.

According to SI, value information is accumulated across time but, prior to accumulation, higher value samples suppress lower samples via a selective gating mechanism. In the extant implementation of the model the vigour of this suppression is constant. For instance, a value sample of 60 will suppress a competing value sample of 59 in the same fashion
that it would suppress a competing value sample of 49 . Here, I focus exclusively on binary choices and show that this invariant mechanism fails to capture some qualitative patterns in past data. To overcome this limitation, I propose a dynamical mechanism in which the inputs compete against each other via potentially asymmetric inhibition.

In what follows, I describe the extant implementation of the SI model and its trademark behavioural signatures. Then, I use previously published data and outline qualitative patterns that SI in its current form cannot explain. These patterns are replicated in a new experiment. Third, I propose three extensions of the selective integration model and fit these extensions to the new data. Finally, I present results from a new experiment that decisively disentangles these three selective integration extensions.

## Selective Integration: description and behavioural signatures

In the SI model for binary choices, value samples in support of the two alternatives arrive simultaneously ${ }^{1}$ and are accumulated over time, as in sequential sampling models of perceptual discrimination and categorisation (Bogacz, Brown, Moehlis, Holmes, \& Cohen, 2006). Importantly, the values that are momentarily higher are passed onto the accumulation layer unaffected but the relatively lower values are truncated, akin to an attentional process that selectively prioritises the accumulation of local winners over local losers. In this section I outline the mathematical details of the extant implementation of selective integration for binary choices and point to key behavioural phenomena predicted by the model.

## Model description

The model described here applies to decisions based on two sequences of inputs, presented simultaneously. The two sequences have thus the same number of samples and each pair of samples is presented at a discrete time-step, for a fixed time interval. Here, based on the findings in Tsetsos et al. (2016), I assume that the incoming samples are not corrupted by noise prior to accumulation. The two sequences are labelled $S_{A}$ and $S_{B}$, with $S_{A}(t)$ indicating the value of sequence $A$ at the (discrete) sample $t$. Two accumulators ( $Y_{A}$ and $Y_{B}$ ) integrate the values of the sequences across time according to the following difference equations:

[^217]\[

$$
\begin{align*}
& Y_{A}(t)=(1-\lambda) \cdot Y_{A}(t-1)+I_{A}(t)+\xi \cdot \zeta_{A}(t)  \tag{1}\\
& Y_{B}(t)=(1-\lambda) \cdot Y_{B}(t-1)+I_{B}(t)+\xi \cdot \zeta_{B}(t) \tag{2}
\end{align*}
$$
\]

In the above $t$ indicates the current discrete time-step (or sample), $\lambda$ is accumulation leakage, $I_{A, B}(t)$ is the input to the two accumulators on a given time-step, $\xi$ is the standard deviation of the noise at the accumulation level and $\zeta_{A, B}(t)$ are standard Gaussian samples, independent from each other and across time-steps. The accumulators are initialised at 0 : $Y_{A}(t)=Y_{B}(t)=0$. At the end of the accumulation period (at $t=T$, with $T$ being the total number of samples presented in each sequence) a decision is made in favour of the accumulator with the higher tally. If both accumulators end up with equal tallies, a decision is made randomly.

The inputs to the two accumulators $\left(I_{A, B}(t)\right)$ reflect the modified sequence values after the selective integration filter is applied. I refer to this filter as selective gating. Selective gating is implemented as follows.

$$
\begin{align*}
& I_{A}(t)=\theta\left(S_{A}(t), S_{B}(t)\right) \cdot S_{A}(t)  \tag{3}\\
& I_{B}(t)=\theta\left(S_{B}(t), S_{A}(t)\right) \cdot S_{B}(t) \tag{4}
\end{align*}
$$

Function $\theta$ returns a value of 1 if the first argument is equal or larger than the second and a value $w$ (selective gating parameter) otherwise:

$$
\theta(x, y)=\left\{\begin{align*}
1 & \text { if } x>=y  \tag{5}\\
w & \text { if } x<y
\end{align*}\right.
$$

## Behavioural signatures of selective integration

Pro variance (PV) effect. Consider two sequences, $A$ and $B$, with values sampled from normal distributions with means $\mu_{A}, \mu_{B}$ and standard deviations $\sigma_{A}, \sigma_{B}$. The pro variance effect (PV) occurs when participants choose more often sequence $A$ when $\mu_{A}=\mu_{B}$ and $\sigma_{A}>\sigma_{B}$. Equivalently, the PV effect is present when accuracy is higher in trials where $\mu_{A}>\mu_{B}$ and $\sigma_{A}>\sigma_{B}$ (correct answer is $A$ ) relative to the accuracy in trials where the means are swapped ( $\mu_{A}<\mu_{B}$ and $\sigma_{A}>\sigma_{B}$; correct answer is $B$ ). The PV effect was originally demonstrated in Tsetsos et al. (2012) and it was robustly replicated in Tsetsos et al. (2016). SI explains the PV effect as follows: a losing sample from the high variance distribution will more likely have low value. In the low variance distribution, a losing sample will more likely have a mediocre value. Multiplicatively downweighting a low value results in a smaller loss relative to downweighting a mediocre value (for a value of 30 and for $w=0.5$ the loss is 15 ; for a value of 50 the respective loss is 25 )(Tsetsos et al., 2012).

Frequent-winner (FW) effect. Consider two sequences $A$ and $B$, consisting of the same three low $(L)$, medium $(M)$ and high $(H)$ value samples such that $H-M=M-L$. The order of appearance of these three samples differs in the two sequences: $A \rightarrow L M H$ and $B \rightarrow H L M$. Thus, when presented
simultaneously in that order, sequence $A$ wins locally twice by a small margin ( $M$ vs. $L$ and $H$ vs. $M$ ) and loses once by a larger margin ( $L$ vs. $H$ ). According to the SI model ( $w<1$ ) a choice bias in favour of sequence $A$ is expected since $M+H+L \cdot w>L \cdot w+M \cdot w+H$ or $M>M \cdot w$. This frequent-winner effect (hereafter FW) can also appear when accuracy is higher in trials in which all values in $A$ are augmented by a small constant $c$ (such that $A$ has a higher total value and dominates $B$ in 2 out of 3 samples), relative to the accuracy in trials in which all values of $B$ are augmented by a small constant $c$ (in that case $B$ has a higher value but $A$ still dominates $B$ in 2 out of 3 samples). Importantly, SI can lead to intransitive preference cycles when a third sequence $C$ with values $M H L$ is considered. In that case, in the respective binary choices, $A$ will be preferred over $B, B$ will be preferred over $C$ and $C$ will be preferred over $A$. The FW effect, and the corresponding weak stochastic transitivity violations, were robustly obtained across 4 experiments, in which participants had to choose between bars of different length, presented sequentially (Tsetsos et al., 2016).

## Challenges for Selective Integration

Under the selective integration framework the PV and FW effects occur due to the same mechanism, controlled by the selective gating parameter. It is therefore expected that the two effects will be strongly correlated across participants. However, a re-examination of the 4 experiments reported in Tsetsos et al. (2016) $(N=93)$ reveals no correlation between the two effects $(r=0.000)$. On the contrary, the effects predicted by fitting the SI model show indeed significant positive correlation ( $r=0.323, p=0.002$ ). Additionally, for the same parametrisation, the SI model predicts a much stronger FW than PV effect (difference in the predicted effects: $M=$ $0.128, S E=0.014, t(92)=9.150, p<0.001)$. However, in the observed data this difference does not occur (difference in the observed effects: $M=0.006, S E=0.023, t(92)=$ $0.240, p=0.811$ ). The model predicts well the magnitude of the FW effect but, although the predicted PV effect is significant $(M=0.036, S E=0.004, t(92)=9.177, p<0.001)$, it is much smaller than the observed one (observed and predicted difference for the PV effect: $M=0.102, S E=0.014, t(92)=$ $7.274, p<0.001$ ).

The lack of correlation between the PV and FW effects and the underestimation of the PV effect challenge the existing implementation of SI. It is conceivable, however, that these patterns are specific to the design and stimuli used in Tsetsos et al. (2016) (i.e. accumulation of lengths) and do not reflect generalisable limitations of the model. I examine next whether this is the case by characterising the PV and FW effects in an experiment that involves the accumulation of numerical values (c.f. Tsetsos et al. (2012)).

## Experiment 1

Participants. 25 participants $\left(M_{\text {age }}=28.1, S D_{\text {age }}=6.4,14\right.$ female) with normal or corrected-to-normal vision and no
history of neurological or psychiatric impairment were recruited from Birkbeck's (University of London) participants pool. All participants gave informed consent to participated and all procedures were approved by the local ethics committee.

Task \& Procedure. On each trial, participants observed pairs of black 2-digit numerical values presented rapidly and sequentially, to the left and right of a central fixation point and against gray background. The viewing distance was 60 cm and each numerical character was $0.93^{\circ}$ wide and $1.5^{\circ}$ long. After the presentation of 9 pairs of numbers, the central fixation point turned blue and participants were asked to choose which stream had on average the higher value. After giving a response the blue dot turned green (red) to indicate a correct (incorrect) response. The presentation rate of the numbers was 800 ms and 1 second gap was left between trials. Overall there were 4 blocks with 65 trials each. At the end of each block participants could see their accuracy so far. At the end of the experiment participants received $£ 7$ and a $£ 2$ bonus if their accuracy exceeded $75 \%$.

Design. There were 4 types of trials ( 65 trials per type) in the experiment, presented in random order. In all trials there was a correct answer, with the sum of the higher sequence differing from the sum of the lower sequence by 72 units. Two types of trials were associated with the PV effect while the other two with the FW effect. In the PV trials the sequences were generated from Gaussian distributions, with the mean of the higher sequence $\left(\mu_{H}\right)$ sampled from $\mu_{H} \sim U(45,65)$. The mean of the lower sequence was $\mu_{L}=m u_{h}-8$. In one type of PV trials, referred to as $\mathrm{PV}_{1}$ trials, the standard deviation of the higher sequence was $\sigma_{H}=20$ while the standard deviation of the lower sequence was $\sigma_{L}=10$. In $\mathrm{PV}_{2}$ trials the standard deviations changed with $\sigma_{H}=10$ and $\sigma_{L}=20$. The accuracy difference between $\mathrm{PV}_{1}$ and $\mathrm{PV}_{2}$ trials quantifies the PV effect. In the FW trials, the mean values of the higher and lower sequences were matched to those in the PV trials. However, the temporal distribution of the sequences was manipulated such that one alternative always dominated the other in 6 out of 9 samples (see also Tsetsos et al. (2016)). In $\mathrm{FW}_{1}$ trials the higher sequence dominated the lower sequence more often. When the higher sequence dominated, it did so by $U(23,28)$ units and in the less often cases when the lower sequence dominated it did so by $U(22,32)$ units. In $\mathrm{FW}_{1}$ trials the higher sequence dominated the lower sequence in 3 out of 9 samples only. When the higher sequence dominated it did so by $U(38,48)$ units and when the lower sequence dominated it did so by $U(7,13)$ units. The accuracy difference between $\mathrm{FW}_{1}$ and $\mathrm{FW}_{2}$ trials quantifies the FW effect. In all trials the generated sequences were constrained so as to involve only 2-digit numbers ranging from 10 to 90 .

Results. Participants performed above chance in all trials (accuracy: $M=0.774, S E=0.023$ ). Accuracy in $\mathrm{PV}_{1}$ trials was higher than in $\mathrm{PV}_{2}$ trials $(M=0.167, S E=$ $0.032, t(24)=5.170, p<0.001, d=1.034)$ replicating thus
the PV effect obtained elsewhere (Tsetsos et al., 2012, 2016). Accuracy in $\mathrm{FW}_{1}$ trials was higher than in $\mathrm{FW}_{2}$ trials $(M=0.050, S E=0.017, t(24)=2.897, p=0.008, d=$ 0.579 ). This finding replicates with different stimuli (i.e. numbers) the effect reported in Tsetsos et al. (2016). Contrary to what SI predicts, there was no correlation between the two effects ( $r=-0.014, p=0.949$ ) and the FW effect was weaker than the PV effect.

## Extensions of selective integration

The challenges that were identified for SI using the datasets from Tsetsos et al. (2016) persist in Experiment 1: the two critical effects were not correlated and the PV effect was larger than the FW effect. To address these challenges I here propose 3 extensions of the SI model. One extension is static, with selective gating invariance as in the original model, while the other two are dynamic and biologically inspired. The primary aim of these extensions is to decorrelate the PV and FW effects. Hereafter, the baseline SI model described earlier will be referred to as $\mathcal{M}_{S 0}$.

## A static extension

The first SI extension involves a transducer function that transforms objective values into their subjective counterparts. The model is thus identical to the one described in Eq. 1-5, with the exception that Eq. 3-4 take the form:

$$
\begin{align*}
& I_{A}(t)=\theta\left(S_{A}(t)^{\alpha}, S_{B}(t)^{\alpha}\right) \cdot S_{A}(t)^{\alpha}  \tag{6}\\
& I_{B}(t)=\theta\left(S_{B}(t)^{\alpha}, S_{A}(t)^{\alpha}\right) \cdot S_{B}(t)^{\alpha} \tag{7}
\end{align*}
$$

Exponentiating the inputs allows the PV effect to occur independent of the selective gating parameter when $\alpha>1$. In such cases the value function is convex resulting in a riskseeking bias. Although convex low-level representation is undocumented in numbers (Feigenson, Dehaene, \& Spelke, 2004) or lengths (Stevens, 1957)), it is possible that, at a higher processing level, large quantities stand out. This model will be referred to as $\mathcal{M}_{S 1}$.

## Dynamic extensions

Implementations $\mathscr{M}_{S 0,1}$ are static in nature. Here I explore dynamical implementations, in which selective gating falls out from continuous competition between the input units. The competition is mediated by inhibition as in models of visual attention (Lee, Itti, Koch, \& Braun, 1999). To illustrate the basic idea, the incoming values compete against each other and the activation states of the input units feed continuously, in cascade (McClelland, 1979), to the accumulation level (variables $Y$ in Eq. 2). Almost equivalently, and to maintain comparability between static and dynamic SI implementations, I assume that a given accumulator receives discrete updates. These updates are equal to the temporal average of the activation in the corresponding input unit, for the period during which the stimulus was presented. The first dynamic model, labelled $\mathscr{M}_{D 1}$, involves mutual inhibition between the
input units (Usher \& McClelland, 2001). Eq. 1-2 remain intact but Eq. 3-5, which implement the selective gating filtering, are replaced as follows:

$$
\begin{align*}
I_{A}(t) & =\frac{\int_{0}^{P} X_{A}(x) d x}{P}  \tag{8}\\
I_{B}(t) & =\frac{\int_{0}^{P} X_{B}(x) d x}{P} \tag{9}
\end{align*}
$$

Variables $X_{A, B}$ reflect the input units. Variable $P$ is the duration (in units of time) that a given pair of samples is presented for and $d x$ is a small time interval (set in simulations to $d x=0.001$ seconds). The input units are initiated at 0 and their dynamics are governed by the following coupled differential equations:

$$
\begin{align*}
d X_{A} & =\left(-\kappa X_{A}-\beta f\left(X_{B}\right)+S_{A}(t)\right) d x  \tag{10}\\
d X_{B} & =\left(-\kappa X_{B}-\beta f\left(X_{A}\right)+S_{B}(t)\right) d x \tag{11}
\end{align*}
$$

In the above, $\kappa$ is a leak parameter (set to 1 throughout this paper), $f$ is the identity function $(f(x)=x)$ and $\beta$ is the strength of mutual inhibition. The two input units thus reflect sustained input, corresponding to the presented values, $S_{A, B}(t)$ at time-step $t$. For simplicity, the above equations are deterministic, consistent with the finding that during value psychophysics noise at the representation level is negligible (Tsetsos et al., 2016). The input units are subject to a reflecting boundary at 0 that prevents activation states from being negative:

$$
\begin{align*}
& X_{A}=\max \left(X_{A}, 0\right)  \tag{12}\\
& X_{B}=\max \left(X_{B}, 0\right) \tag{13}
\end{align*}
$$

When $\beta=0$, the two input units quickly converge to their nominal values. When inhibition is present, the larger value suppresses the smaller value, implementing that way a form of selective gating. Figure 1a shows examples of the evolution of activation states in the input units.

The second dynamic extension $\left(\mathcal{M}_{D 2}\right)$ is identical to $\mathcal{M}_{D 1}$ with the exception that, following Brown and Holmes (2001), each unit inhibits the other via a sigmoid activation function.Thus, $f$, which was the identity function in $\mathcal{M}_{D 1}$, now becomes:

$$
\begin{equation*}
f(x)=\frac{1}{1+e^{(-g(x-b))}} \tag{14}
\end{equation*}
$$

In the above, $g$ is the slope of the activation function (here set to 1 ) and $b$ the inflection point of the sigmoid. In other words, $b$ controls when selective gating will kick in while the inhibition strength $(\beta)$ controls the strength of selective gating. The difference between $\left(\mathcal{M}_{D 1}\right)$ and $\left(\mathcal{M}_{D 2}\right)$ is that, in the former, inhibition is mutual while in the latter inhibition can be non-reciprocal (Figure 1b) and inactive from inputs of low value.


Figure 1: (a) Input activation trajectories for $\mathscr{M}_{D 1}$ for different value samples and for $\kappa=1$. Leftmost panel depicts the special case without inhibition. (b) Same as (a) but for $\mathcal{M}_{D 2}$ and $b=38, g=1$ and $\kappa=1$.

## Comparison of selective integration extensions

## Quantitative comparison of the models

Here I fit the models to the choice data of each participant from Experiment 1. Model predictions for each trial are derived numerically. The negative log likelihood of each parametrisation is calculated on a trial-by-trial basis and summed across trials. $\mathcal{M}_{S 0}$ has three free parameters $(w, \xi, \lambda)$ and $\mathcal{M}_{S 1}$ has one extra free parameter ( $\alpha$ ). For the dynamic extensions, some parameters are set to fixed values ( $\kappa=1, g=1, P=500 \mathrm{~ms}, d x=1 \mathrm{~ms}$ ), which leaves $\mathcal{M}_{D 1}$ with 3 free parameters $(\sigma, \lambda, \beta)$ and $\mathscr{M}_{D 2}$ with a fourth parameter (b). The models are compared based on their BIC values, aggregated across participants, and also on a participant-byparticipant basis (i.e. the proportion of participants for which a model has the lowest BIC score). Additionally, the correlation between the PV and FW predicted effects is examined. The results are summarised in Table 1.

Table 1: Model comparison in Experiment 1.

| Model | Total BIC | \% BIC lowest | $r$ |
| :--- | :--- | :--- | :---: |
| $\mathcal{M}_{S 0}$ | 8,458 | $28 \%$ | 0.736 |
| $\mathcal{M}_{S 1}$ | 8,554 | $0 \%$ | 0.254 |
| $\mathcal{M}_{D 1}$ | 8,441 | $24 \%$ | 0.779 |
| $\mathcal{M}_{D 2}$ | 8,398 | $48 \%$ | 0.087 |

$\mathcal{M}_{D 2}$ explains the data best and succeeds to decorrelate the PV and FW effects (last column in Table 1). It is also the only model that does not underestimate dramatically the PV effect and predicts that it will have higher magnitude than the FW effect (Figure 2). In this model, the PV effect is partly driven by parameter $b$. If this parameter is set above the middle of the value range, selective gating will be inactive for comparisons between mediocre and low values, further exaggerating


Figure 2: Data and model fits in Experiment 1. Error bars correspond to 2 SE .
the choice bias for a high variance sequence. The FW effect is independent of the $b$ value (as long as it is not too high, deactivating altogether selective gating). $\mathcal{M}_{S 1}$ also succeeds to decorrelate the effects but provides a poor fit. Importantly, $\mathcal{M}_{D 1}$ appears to suffer from the same limitation as $\mathcal{M}_{S 0}$ : selective gating is controlled by one parameter $(\beta)$ and the PV and FW are strongly correlated.

## Qualitative differences among models

The distinction between the four selective integration implementations is the way selective gating is implemented. In the two static implementations $\left(\mathcal{M}_{S 0,1}\right)$ the weight applied on the local loser is invariant to the difference between the winning and losing value samples. On the contrary, in the dynamic extensions, this difference matters. In Figure 3, I show the effective weight applied on the losing sample for several combinations of pairs of values (c.f. Figure 1). Static implementations predict, by definition, invariance of weighting (here for $w=0.5$ ). $\mathcal{M}_{D 1,2}$ both predict that as the difference between the two inputs increases (c.f. leftmost and middle bars), suppression of the loser increases too. This is reminiscent of the distance effect encountered in numerical cognition (Moyer \& Landauer, 1967).

Adding a constant to both input samples (c.f. leftmost and rightmost bars) results in opposing predictions in the two dynamic models. $\mathcal{M}_{D 1}$ predicts weaker suppression for increased values, since the competition between the two inputs takes longer to resolve. $\mathscr{M}_{D 2}$ predicts a stronger suppression of the loser, since the winning unit will breach first the $b$ barrier and will start inhibiting the other unit strongly, resulting in enhanced winner-take-all dynamics (Figure 1, rightmost panels). All models thus make distinctive qualitative predictions regarding a magnitude effect, which I exploit in the next experiment.

## Experiment 2

In this experiment I examine how the PV effect changes when all sequence values increase or decrease by the same constant amount (Figure 4a). According to static models no change


Figure 3: Effective weighting is the ratio between the filtered lower sample and the corresponding unfiltered values. The unfiltered values in static models are nominal values (without selective gating, $w=1$ ) and in dynamic models values, as per Eq. $8-9$, in the $\beta=0$ instantiation. The values of the different pairs are given in the x -axis.
is expected, since selective gating is invariant to the magnitude of the values. $\mathcal{M}_{D 1}$ predicts weaker selective gating for high values, while $\mathscr{M}_{D 2}$ predicts a stronger PV effect for high values.

Participants. 18 participants $\left(M_{\text {age }}=26.4, S D_{\text {age }}=5.2,11\right.$ female) took part. The rest of details are as per Experiment 1.

Task \& Procedure. The task and procedure was identical to Experiment 1 except that there were 12 samples in each stream and that the presentation rate was 0.5 seconds. Overall participants did 6 blocks with 50 trials each.

Design. The PV effect was elicited as in Experiment 1, using two trial types ( $\mathrm{PV}_{1,2}$ ) and examining the accuracy difference between them. Here, there were 3 conditions giving rise to 6 trial types ( 50 trials for each type, randomly presented). In the baseline condition $\left(\mathrm{PV}_{B_{1,2}}\right)$, the correct sequence had always a mean of 50 and the incorrect a mean of 42 . In the negative offset condition $\left(\mathrm{PV}_{-1,2}\right)$, for a given trial, 6 pairs of values were created as per the baseline condition. The remaining 6 pairs were created by subtracting from the mean of both Gaussians a constant $(c=15)$. The regular and lower pairs were presented in random order in a given trial. Equivalently, in the positive offset condition ( $\mathrm{PV}_{+_{1,2}}$ ) a constant $(c=15)$ was added to the values of 6 pairs.

Results. The PV effect increased as both sequences increased in absolute values as indicated by a repeated measures ANOVA $\left(F(2,34)=18.74, p<0.001, n^{2}=0.524\right)$. Tukey post-hoc tests revealed that the PV effect was lower in the negative offset condition relative to the baseline ( $p<$ 0.001 ) and the positive offset ( $p<0.001$ ) conditions. The difference between the baseline and the positive offset condition was not significant ( $p=0.171$ ). As predicted, this PV increasing pattern was solely captured by $\mathscr{M}_{D 2}$. The advan-


Figure 4: (a) Outline of experimental design. (b) PV effect as a function of negative offset (N), baseline (B) and positive offset ( P ) conditions for data and models (colour code for models as per Figures 2-3). Error bars correspond to $95 \%$ CI.
tage of $\mathscr{M}_{D 2}$ in this experiment was clear also in quantitative comparisons (data fitted for each participant separately) using BIC (Table 2).

Table 2: Model comparison in Experiment 2.

| Model | Total BIC | \% BIC lowest |
| :--- | :---: | :---: |
| $\mathcal{M}_{S 0}$ | 5,976 | $6 \%$ |
| $\mathcal{M}_{S 1}$ | 5,466 | $0 \%$ |
| $\mathcal{M}_{D 1}$ | 5,448 | $6 \%$ |
| $\mathcal{M}_{D 2}$ | 5,253 | $88 \%$ |

## Conclusion

Selective integration is a decision making model that has successfully explained several rationality violations. One potential criticism against this model is that its applicability is limited to rapid decisions from experience, in which attentional demands are increased. However, it has been previously shown that selective gating increases under lower attentional demands (Tsetsos et al., 2016). Additionally, behavioural signatures that are routinely obtained in high-level decisions are also obtained in rapid experiential decisions, implying that the latter can offer a window to more complex choice mechanisms.

Here I presented challenges for the extant implementation of selective integration. These challenges were successfully addressed by a dynamic extension of the model, in which the inputs compete for accumulation via inhibiting each other, as in models of selective attention and visual search. This dynamic extension predicts that the vigour of selective integration increases both when the distance and the absolute magnitudes of the two inputs increase. This prediction was experimentally confirmed. Oveall, this dynamic and biologically inspired (Usher \& McClelland, 2001) extension presented here, significantly improves the descriptive adequacy of the selective integration framework and facilitates its validation at the neurophysiological level.

## Acknowledgments

I thank Artemis Maipa for assistance with data collection. This work was funded by a Wellcome Career Development fellowship, a British Academy/ Leverhulme Trust awarded and a Marie Curie Individual Fellowship awarded to the author.

## References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., \& Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review, 113(4), 700.
Brown, E., \& Holmes, P. (2001). Modeling a simple choice task: stochastic dynamics of mutually inhibitory neural groups. Stochastics and dynamics, 1(02), 159-191.
Feigenson, L., Dehaene, S., \& Spelke, E. (2004). Core systems of number. Trends in cognitive sciences, 8(7), 307314.

Huber, J., Payne, J. W., \& Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. Journal of consumer research, 9(1), 90-98.
Lee, D. K., Itti, L., Koch, C., \& Braun, J. (1999). Attention activates winner-take-all competition among visual filters. Nature neuroscience, 2(4), 375-381.
McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. Psychological review, 86(4), 287.
Moyer, R. S., \& Landauer, T. K. (1967). Time required for judgements of numerical inequality. Nature, 215(5109), 1519-1520.
Stevens, S. S. (1957). On the psychophysical law. Psychological review, 64(3), 153.
Tsetsos, K., Chater, N., \& Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. Proceedings of the National Academy of Sciences, 109(24), 9659-9664.
Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., \& Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. Proceedings of the National Academy of Sciences, 113(11), 3102-3107. doi: 10.1073/pnas. 1519157113

Tsetsos, K., Usher, M., \& Chater, N. (2010). Preference reversal in multiattribute choice. Psychological review, 117(4), 1275.
Tversky, A. (1969). Intransitivity of preferences. Psychological review, 76(1), 31.
Usher, M., \& McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. Psychological review, 108(3), 550.
Von Neumann, J., \& Morgenstern, O. (2007). Theory of games and economic behavior. Princeton university press.

# Cake or Hat? Words Change How Young Children Process Visual Objects 

Catarina Vales (cvales@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue<br>Pittsburgh, PA 15214 USA

Linda B. Smith (smith4@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E $10^{\text {th }} \mathrm{St}$
Bloomington, IN 47405 USA


#### Abstract

A large literature shows that language influences cognition. Yet, we know very little about when and how linguistic influences on cognition become important in development. Here we test the proposal that one pathway by which language affects cognition is by activating category information which influences visual processing, and that this influence starts early. Across two experiments, we show that category information affects visual processing and that words can activate category information in young children.


Keywords: language; attention; cognitive development; vision.

## Introduction

A large literature has documented that linguistic information changes other cognitive processes. Evidence for this comes from laboratory tasks in which people perform differently if they experience the same event associated with different kinds of linguistic information (Feist \& Gentner, 2007; Loftus \& Palmer, 1974) or associated with linguistic information vs. information presented in another modality (Lupyan \& Spivey, 2010; Lupyan \& Thompson-Schill, 2012), and from cross-linguistic research that shows influences of language on presumably non-linguistic processes (Fausey \& Boroditsky, 2011; Levinson \& Haviland, 1994). Taken together, these results show that language - and words, in particular - changes how adults perform on a wide variety of tasks, and that these cognitive processes are permeable to linguistic information.

Despite this wealth of evidence, we know very little about when and how linguistic influences on cognition become important in development. Understanding the development of linguistic effects on cognition is essential to understand the development of human cognition and the nature of individual differences in cognitive abilities - differences that start early and have downstream consequences into later development (Morgan, et al., 2015; Stanovich, 1986). One possibility is that language influences cognition by activating information about the objects or events to which it refers, and this information changes how visual information is processed. This hypothesis is plausible for three reasons. First, there is evidence supporting the link between language and visual processing. For example, adults listening to spoken sentences look at possible visual referents even when the visual array is irrelevant to the task (see Huettig, Rommers \& Meyer, 2011), and when adults
hear a word (e.g. "snake") they are likely to look at objects that share aspects with the referent of the word (e.g. a rope, similar shape, Huettig \& Altmann, 2007). Similarly, adults’ ability to detect a visual item is boosted by labeling the item (Lupyan \& Spivey, 2008), and children's ability to find a target in a cluttered display is boosted by hearing the spoken name of the target object (Vales \& Smith, 2015).

Second, word learning and object recognition are two related developmental achievements. Children's ability to identify visually degraded objects (Pereira \& Smith, 2009), to attend to the configuration among the parts of a novel object (Augustine, et al., 2011), and to recognize sparse versions of known object categories (Smith, 2003), all are positively related to the number of words a child knows. Object recognition continues to be coordinated with word comprehension into adulthood (Huettig et al., 2011).

Third, there is evidence suggesting that words activate knowledge about the categories to which they refer. Words - object names in particular - do not refer to a specific item but rather to more abstract knowledge. Empirical results have shown that, relative to other cues (e.g. environmental sounds), words activate more decontextualized, categorical knowledge (Edmiston \& Lupyan, 2015; Lupyan, 2008).

Taken together, this evidence supports our proposal that one pathway by which language affects human cognition is by activating category information which then influences visual processing, and that this pathway likely starts in early childhood. To directly test this proposal, in Experiment 1 we asked whether visual processing can be influenced by visual category information, and in Experiment 2 we examined whether category information activated by words can also influence visual processing. We tested 3-year-old children, who know several hundreds of object names and are at the start of the long developmental course in visual object recognition and in language development.

## Rationale for the experiments

A large literature on categorical perception suggests that categorical information activated through visual means can change how adults process visual stimuli (Beale \& Keil, 1995; Daoutis, Pilling \& Davies, 2006; Goldstone, 1995; Goldstone, Lippa \& Shiffrin, 2001; Livingston, Andrews \& Harnad, 1998). By hypothesis, having learned that items belong to the same category changes in-task perceived similarity, making within-category discriminations harder than between-category discriminations. This idea that
within-category comparisons are more difficult than between-category comparisons has been conceptually replicated with multiple kinds of tasks and stimuli (Jonides \& Gleitman, 1972; MacKain, Best \& Strange, 1981; Newell \& Bülthoff, 2002), including some with infants and children (Eimas, Siqueland, Jusczyk \& Vigorito, 1971; Jusczyk, Rosner, Cutting, Foard \& Smith, 1977; Massaro, 1984).

To test if categorical information influences visual processing in young children, the present experiments tested children's ability to find a target in a cluttered array. The visual arrays were composed of items of the same category as the target (Within-Category search) or items of a different category than the target (Between-Category search). In Experiment 1, the category information was instantiated via visual information, and in Experiment 2, the category information was instantiated via linguistic information. In both experiments, we used two categories that share visual similarity but minimal conceptual similarity, and that children are likely to be familiar with: cakes and hats.

## Experiment 1

If categorical information instantiated by visual information changes children's ability to visually process a visual array, then searching for a target amidst items of the same visual category should be more difficult than searching for that target amidst items of a distinct visual category.

## Methods

Participants. Thirty-two children ( 15 females, $\mathrm{M}_{\text {age }}=36$ months, $\mathrm{SD}=1.92$ ) were randomly assigned to either the Within- or the Between-Category condition. Children had no developmental disorders, and English was the main or only language spoken by all families. Two additional children were recruited but not included due to experimenter error and being unable to follow task instructions during the familiarization phase. Parental consent was obtained for all participants, and all children received a toy for participating.
Apparatus and Stimuli. Stimuli were presented on a 17 " touchscreen monitor. E-Prime (PST, Pittsburgh, PA) was used to present the stimuli and to record participants' responses. The stimuli were placed in 16 possible locations. The stimulus set is depicted in Figure 1A; four hats and four cakes were selected in pairs and recolored in red scale, such that a hat and a cake in the same pair were similar to each other in color appearance, overall shape, and details (e.g. in pair 1, both items have stripes and a smaller component at the top). The Within- vs. Between-Category manipulation was realized by changing target/distractor assignments; for instance, for pair 2, the Within-Category search array was composed of hat 2 as the target and hats 1,3 , and 4 as distractors; for the Between-Category search, cake 2 served as the target amidst the same (i.e. 1, 3, and 4) hat distractors.

To ensure that young children could recognize the items used, 12 children who did not participate in the main experiments ( 7 females, $\mathrm{M}_{\text {age }}=36$ months, $\mathrm{SD}=2.67$ ) were tested in a 4-alternative forced choice recognition task; on
each trial, children were asked to select the picture that matched the heard word (e.g. "Where is the cake?"). Each child was asked to recognize all the items in the stimulus set twice, with target category (hat vs. cake) blocked, order of block presentation counterbalanced across children, and items presented in random order. On average, children selected the correct item on $81 \%$ of the trials ( $\mathrm{SD}=0.22$ ). No differences in accuracy were found across the two category of items $(t(11)=1.11, p=0.29)$, or time taken to respond to cakes vs. hats items $(t(11)=1.92, p=0.08)$.


Figure 1, A: Full stimulus set. The hat and cake stimuli were used as targets and distractors in Experiment 1. The ambiguous items were used as targets in Experiment 2.

B: Experiment 1, Trial structure.
Because the visual search task requires participants to discriminate the items in the search array from each other, 8 additional children ( 2 females, $\mathrm{M}_{\text {age }}=36$ months, $\mathrm{SD}=1.04$ ) were tested in an immediate match-to-sample task that probed their ability to discriminate pairs of stimuli. On each trial, children were presented with a sample object at the top center of the screen and then asked to indicate which of two options at the bottom matched the sample; if children could discriminate the two stimulus pictures, then they should be able to correctly select the option that matched the sample. All possible combinations of items that would be presented as targets and distractors in the visual search were tested. Each child was asked to discriminate one hat and one cake from the remaining items; this was done so that, for each child, a foil never became a target and vice-versa. The target category (hat vs. cake) was blocked, order of block presentation was counterbalanced across children, and items were presented in random order. Each child was tested on a given contrast (e.g. hat 1 vs. hat 2) twice. On average, children selected the correct option on $89 \%$ of the trials ( $\mathrm{SD}=0.19$ ). No differences were found in children's ability to discriminate Between- vs. Within-category items $(t(7)=1.07, p=0.32)$ or time taken to respond to cakes vs. hats items $(t(7)=0.41, p=0.69)$.
Design and Procedure. Each trial started with a "fixation" slide that encouraged children to rest their hands on the table. The experimenter made sure the child was looking at the screen before displaying a preview of the target. After 1 s , the search array was automatically displayed and the child was asked to find the target picture and touch it; the trial ended once a manual response was detected (see Figure

1B). Children had up to 15 secs to make a response, and were encouraged to find the picture as fast as possible. Across test trials, the target was displayed equally often on the left and right side of the screen. Prior to the test phase, children were familiarized with using the touch screen and with the idea of searching for the object that matched the visual preview as fast as possible. Each child was assigned to one target, and searched for that target for 24 trials. None of the objects were labeled. The experimenter gave general encouragement (e.g. "thanks for your help finding the pictures") but no feedback was provided. Children received stickers to maintain their interest in the task.

## Results and Discussion

Initial inspection of the data suggested that participants were, on average, both faster and more accurate in the Between-Category condition than in the Within-Category condition. Traditional analysis of variance would require analyzing response time and accuracy separately, which implicitly assumes that these two variables are independent (e.g. Davidson \& Martin, 2013). Instead, we analyzed RT and accuracy together by comparing the relationship between RT and accuracy across the two conditions. Figure 2A depicts this relation and shows that the two conditions differ in how time taken to respond (plotted as quantiles) influences the likelihood of correctly identifying the target; accuracy in the Between-Category condition was overall higher and less influenced by response time. A generalized linear mixed effects analysis was performed using the geepack package (Højsgaard, Halekoh \& Yan, 2006) in the R environment. The model was fit with a logit link function, a binomial variance function, a scale parameter fixed at 1 , and an independent correlation structure. The variables Condition (Between-Category vs. Within-Category) and RT (as a continuous variable) were included as fixed effects with the interaction term, and participant was included as a random effect; RT was centered to decrease the differences in the scales of the model parameters. Odds Ratios were calculated by exponentiating the model estimates.

The model showed that Condition (Odds Ratio, OR=0.2, $p<0.001$ ) and the interaction between condition and RT ( $\mathrm{OR}=0.43, p<0.05$ ) were significant predictors of accuracy. RT was marginally predictive of accuracy, $\mathrm{OR}=1.8, p=0.08$. This suggests that processing the visual information in the within-category condition is challenging, and when children are asked to find a target amidst distractors of the same visual category they either cannot maintain the target active for more than a few seconds, or they disengage with the task if they fail to find the target within a few seconds. Notice that visual category information affected visual processing even though children could discriminate pairwise presentations of within- and between-category items equally well in the calibration study, even though children searched for the same visual target across all trials and thus should be able to remember the particular visual target, and even though the visual properties of the items were equated as best as possible across the two conditions.

## A. Experiment 1


B. Experiment 2


Figure 2, Mean proportion of correct responses per Reaction Time quantiles. Dashed line: Between-Category, Solid line: Within-Category.

These results show that category information presented through visual means influences visual processing in young children; this is, to the best of our knowledge, the first demonstration that visual categories directly influence visual processing in young children. This result adds to past research showing that infants, children, and adults (e.g. Eimas, Siqueland, Jusczyk \& Vigorito, 1971; Goldstone, Lippa \& Shiffrin, 2001; Massaro, 1984) are sensitive to category information, showing that category information matters for how children visually process a scene. When children encoded, for example, a hat target and saw the other hats in the within-category search array, that samecategory information seems to have disrupted their ability to find the target hat. In sum, Experiment 1 shows that categorical information perceived in a visual scene can directly influence visual processing, as predicted by our proposal. In Experiment 2 we test the hypothesis that words, through activating categorical information, should also change the perceived target-distractor similarity, and therefore change visual processing.

## Experiment 2

The goal of Experiment 2 was to test the hypothesis that hearing the spoken name of an object activates information about that object's category, which influences visual processing. This hypothesis predicts that, for example, if an ambiguous target is labeled as a hat and is placed amidst other hat distractors, it should be more difficult to find than when that target is labeled as a cake and is placed amidst the same hat distractors. In other words, when presented with the same visual information in the search array, children should be better able to find a target if it was labeled as a category other than the distractors.

## Methods

Participants. Thirty-two children ( 18 females, $\mathrm{M}_{\mathrm{age}}=36$ months, $\mathrm{SD}=2.12$ ) were randomly assigned to either the Within-Category or the Between-Category condition. These children did not participate in Experiment 1. Four additional children were recruited but not included due to refusal to participate $(\mathrm{N}=3)$ and failure to follow task instructions.
Stimuli and Procedure. Four ambiguous items were created by blending together the two items (cake and hat) of
each pair (see Figure 7, rightmost column); the ambiguous items included aspects of both the cake and the hat of the pair (e.g. the frosting and the ribbon), and were edited to look like a plausible visual object (e.g. smooth surface, even coloring, even edges). The Within- vs. Between-Category manipulation was realized by changing how the target object was labeled during the visual preview. For instance, in the Within-Category condition, children would preview the ambiguous item \#2 and hear it labeled as hat, and then be asked to find it amidst hats 1,3 , and 4 . In the BetweenCategory condition, children would preview the same ambiguous item \#2, but hear it labeled as cake and then be asked to find it amidst hats 1,3 , and 4 . Notice that the visual information presented in the preview and in the search array is exactly the same in both conditions - the difference between the two conditions is whether the ambiguous target is labeled a member of the same vs. a different category as the distractors while it is previewed prior to search. All other aspects of the procedure for the visual search were the same as Experiment 1.

To ensure that the ambiguous stimulus pictures were equally likely to be recognized as hats and cakes, 8 additional children ( 6 females, $\mathrm{M}_{\text {age }}=36$ months, $\mathrm{SD}=2.62$ ) were tested in a 4-alternative forced choice recognition task similar to the one used in Experiment 1. Each child was tested with all the ambiguous items, two of them as "cake" and the other two as "hat"; across children, each ambiguous item was equally likely to be tested as "hat" and as "cake". The target category (hat vs. cake) was blocked for each child, the order of block presentation was counterbalanced across children, and the presentation order of the items was randomized within each block. On average, children selected the ambiguous items as "cake" on $81 \%$ ( $\mathrm{SD}=0.40$ ) of the trials, and as "hat" on $84 \%(\mathrm{SD}=0.35)$ of the trials (paired $t(7)=0.16, p=0.88$ ) - suggesting that the ambiguous items were equally likely to be recognized as hats and cakes. Children took a similar amount of time to respond to cakes vs. hats trials $(t(7)=0.45, \mathrm{p}=0.67)$. In addition, to ensure that children were not selecting the ambiguous items merely because they looked more unfamiliar or novel than any of the foils, children were presented with 4 "catch" trials (2 after each block) where they were asked to find a balloon; on these "catch" trials, one of the foils was a novel-shaped object and the other two foils were known objects. Children correctly identified the balloon on $91 \% ~(\mathrm{SD}=0.18)$ of the trials, suggesting that they were not relying on novelty to respond in this task.

## Results and Discussion

Similar to Experiment 1, initial data inspection suggested that participants were both faster and more accurate in the Between-Category condition than in the Within-Category condition. Figure 2B shows that the relationship between RT and accuracy is the same across conditions, in that participants' accuracy does not depend on time taken to respond, but participants in the Between-Category condition were more accurate than participants in the Within-Category
condition. A generalized linear mixed effects model (fit in the same way as in Experiment 1) showed that Condition was the only significant predictor of accuracy, $\mathrm{OR}=0.12$, $p<0.001$. Time taken to respond $[\mathrm{OR}=1.1, p=0.8]$ and the interaction between Condition and RT [OR=0.8, $p=0.4]$ were not predictive of accuracy. This suggests that hearing the name of an object activates visual information about that objects' category, which affects visual processing. When presented with the same visual information, children's ability to find a visual target in a cluttered array depended on how that visual target had been labeled while it was being previewed. This is a robust demonstration of the effect of language on visual processing - encoding an ambiguous object as a hat or as a cake changed children's ability to find that object amidst the same set of distractors.

## General Discussion

The experiments presented here support our proposal that language affects human cognition by activating category information, which in turn influences visual processing. In Experiment 1, children's ability to find a visual target was hindered by the presence of same-category distractors; this influence of categorical information on visual processing was instantiated through visual means. Experiment 2 extended those results by showing that words can also activate categorical information which influences visual processing. Together, these results show that visual processing is influenced by categorical information, and that heard words can instantiate categorical information.

How does categorical information - through visual or linguistic means - influence visual processing? Past research on categorical perception suggests that categorical information changes the perceived similarity among the items, with items that belong to the same category being perceived as more similar to one another than items that belong to different categories (Goldstone \& Hendrickson, 2010; see also Sloutsky \& Fisher, 2004 for a related developmental model). This perceived similarity could influence visual processing in multiple ways. One possibility is that perceiving an item as a member of the same category as the distractors - and consequently as more similar to the distractors - lowers the threshold for accepting an item in the array as the target (e.g. Elman, 1979). Another possibility is that perceiving all the items as items of the same category influences children's ability to bind all the features of the target object together (Treisman \& Schmidt, 1982); there is evidence that object feature binding is still developing in late childhood (Lorsbach \& Reimer, 2005) and that children are prone to making conjunction errors (Dessalegn \& Landau, 2008). Through increasing the perceived similarity between the target and the distractors, category information might lead children to incorrectly bind features of the target and the distractors, increasing the likelihood of making an incorrect selection. Interestingly, language has been suggested to play a role in the binding of visual features in young children (Dessalegn
\& Landau, 2008; 2013), perhaps through the activation of categorical information.

These are empirically testable possibilities that merit future research. But notwithstanding the specific process by which children's ability to find a target was impaired by the presence of distractor items of the same category as the target, the point is that it was impaired - both when the categorical information was presented through visual means (Experiment 1) and through language (Experiment 2). The current results support the idea that words influence visual processing by highlighting information about the objects' category. This idea has important consequences to conceptualize the pathway by which words change visual object processing in young children, proposing that words activate categorical information that may change how objects are perceived and processed. Future research should examine what specific aspects of the objects' category are being activated when a word is heard. The ability to recognize the components of an object, and how those components relate to each other, is one critical aspect of visual object recognition (Tarr \& Bülthoff, 1998), and the developmental literature on visual object recognition suggests a long and protracted development on the ability to use configural information (Augustine et al., 2011: Jüttner et al., 2013). Given the strong links between word learning and visual object recognition in early childhood (Augustine, et al., 2011; Pereira \& Smith, 2009; Smith, 2003), it is possible that language comes to change what aspects of the objects children attend to.

These results also highlight the importance of understanding the nature of visual processing in young children. Contemporary accounts of visual processing in adults propose a reciprocal interaction between the shortterm encoding of visual information and long-term visual representations (e.g. Brady, Konkle, \& Alvarez, 2011). Research across levels of analysis suggest that both processes might be permeable to top-down influences (e.g. Hemmer \& Steyvers, 2009; Olsson \& Poom, 2005), but we have very little understanding of how these develop. That is, what information do children use to visually process objects in the moment, what do those visual representations include, and what factors influence the long-term encoding and fidelity of those visual representations? All these processes are likely to mature and improve with age (e.g. Burnett Heyes et al., 2012; Simmering \& Perone, 2012) and might be weak in children (Riggs, McTaggart, Simpson \& Freeman, 2006; Zhang, Shen, Tang, Zhao \& Gao, 2013). Importantly, visual processing and visual working memory have been shown to be immature in children with language impairments (Collisson et al., 2015), further underscoring the importance of understanding the development of these processes.

In sum, we documented that words influence visual processing, likely by highlighting information about the objects' category. This fits with our proposal that one pathway by which language influences human cognition is by activating category information, which influences visual
processing, and that this pathway likely starts in early childhood.

## Acknowledgments

The authors thank Brianna Le, Lizzie Hart, and Adriana Valtierra for their help with stimuli creation, recruitment, and data collection; and the parents and children who participated in these studies. This work was supported by a Graduate Fellowship from the Portuguese Foundation for Science and Technology (SFRH/BD/68553/2010) awarded to CV and a grant from the National Institute of Child Health and Development (HD28675) awarded to LBS.

## References

Augustine, E., Smith, L. B., \& Jones, S. S. (2011). Parts and relations in young children's shape-based object recognition. Journal of Cognition and Development, 12(4), 556-572.
Beale J.M., \& Keil F.C. (1995) Categorical effects in the perception of faces. Cognition 57: 217-39.
Brady, T. F., Konkle, T., \& Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. Journal of Vision, 11(5), 4.

Burnett Heyes, S., Zokaei, N., van der Staaij, I., Bays, P. M., \& Husain, M. (2012). Development of visual working memory precision in childhood. Developmental Science, 15(4), 528-539.
Collisson, B. A., Grela, B., Spaulding, T., Rueck1, J. G., \& Magnuson, J. S. (2015). Individual differences in the shape bias in preschool children with specific language impairment and typical language development: theoretical and clinical implications. Developmental Science, 18(3), 373-388.
Davidson, D. J., \& Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. Acta Psychologica, 144(1), 83-96.
Daoutis, C. A., Pilling, M., \& Davies, I. R. (2006). Categorical effects in visual search for colour. Visual Cognition, 14(2), 217-240.
Dessalegn, B., \& Landau, B. (2008). More than meets the eye: The role of language in binding and maintaining feature conjunctions. Psychological Science, 19(2), 189195.

Dessalegn, B., \& Landau, B. (2013). Interaction between language and vision: It's momentary, abstract, and it develops. Cognition, 127(3), 331-344.
Edmiston, P., \& Lupyan, G. (2015). What makes words special? Words as unmotivated cues. Cognition, 143, 93100.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., \& Vigorito, J. (1971). Speech perception in infants. Science, 171(3968), 303-306.
Elman, J. L. (1979). Perceptual origins of the phoneme boundary effect and selective adaptation to speech: A signal detection theory analysis. The Journal of the Acoustical Society of America, 65(1), 190-207.

Fausey, C. M., \& Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. Psychonomic bulletin \& review, 18(1), 150-157.
Feist, M. I., \& Gentner, D. (2007). Spatial language influences memory for spatial scenes. Memory \& Cognition, 35(2), 283-296.
Goldstone, R. L. (1995). Effects of categorization on color perception. Psychological Science, 298-304.
Goldstone, R. L., \& Hendrickson, A. T. (2010). Categorical perception. Wiley Interdisciplinary Reviews: Cognitive Science, 1(1), 69-78.
Goldstone, R. L., Lippa, Y., \& Shiffrin, R. M. (2001). Altering object representations through category learning. Cognition, 78(1), 27-43.
Hemmer, P., \& Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. Psychonomic Bulletin \& Review, 16(1), 80-87.
Højsgaard, S., Halekoh, U. \& Yan J. (2006) The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software, 15, 2, 1-11
Huettig, F., \& Altmann, G.T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. Visual Cognition, 15(8), 985-1018.
Huettig, F., Rommers, J., \& Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. Acta Psychologica, 137(2), 151-171.
Jonides, J., \& Gleitman, H. (1972). A conceptual category effect in visual search: O as letter or as digit. Perception \& Psychophysics, 12(6), 457-460.
Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F., \& Smith, L. B. (1977). Categorical perception of nonspeech sounds by 2-month-old infants. Perception \& Psychophysics, 21(1), 50-54.
Jüttner, M., Wakui, E., Petters, D., Kaur, S. \& Davidoff, J. (2013). Developmental trajectories of part-based and configural object recognition in adolescence. Developmental Psychology 49, 161-176.
Levinson, S. C., \& Haviland, J. B. (1994). Introduction: Spatial conceptualization in Mayan languages. Linguistics, 32(4-5), 613-622.
Livingston, K. R., Andrews, J. K., \& Harnad, S. (1998). Categorical perception effects induced by category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(3), 732.
Loftus, E. F., \& Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior, 13(5), 585-589.
Lorsbach, T. C., \& Reimer, J. F. (2005). Feature binding in children and young adults. The Journal of Genetic Psychology, 166(3), 313-328.
Lupyan, G. (2008). From chair to" chair": A representational shift account of object labeling effects on memory. Journal of Experimental Psychology: General, 137(2), 348.

Lupyan, G. \& Spivey, M.J. (2008). Ascribing meaning to unfamiliar items facilitates visual processing. Current Biology, 18: R410-R412
Lupyan, G. \& Spivey, M.J. (2010). Making the invisible visible: Verbal but not visual cues enhance visual detection. PLoS One 5(7).
Lupyan, G., \& Thompson-Schill, S.L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. Journal of Experimental Psychology: General. 141(1), 170-186.
MacKain, K. S., Best, C. T., \& Strange, W. (1981). Categorical perception of English/r/and/1/by Japanese bilinguals. Applied Psycholinguistics, 2(04), 369-390.
Massaro, D. W. (1984). Children's perception of visual and auditory speech. Child Development, 1777-1788.
Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., \& Maczuga, S. (2015). 24-Month-Old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. Child Development, 86(5), 1351-1370.
Newell, F. N., \& Bülthoff, H. H. (2002). Categorical perception of familiar objects. Cognition, 85(2), 113-143.
Olsson, H., \& Poom, L. (2005). Visual memory needs categories. Proceedings of the National Academy of Sciences of the United States of America, 102, 8776-8780.
Pereira, A. F., \& Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. Developmental science, 12(1), 67-80.
Simmering, V. R., \& Perone, S. (2012). Working memory capacity as a dynamic process. Frontiers in Psychology, 3.
Sloutsky, V. M., \& Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General, 133(2), 166.
Smith, L. B. (2003). Learning to recognize objects. Psychological Science, 14(3), 244-250.
Riggs, K. J., McTaggart, J., Simpson, A., \& Freeman, R. P. (2006). Changes in the capacity of visual working memory in 5-to 10-year-olds. Journal of Experimental Child Psychology, 95(1), 18-26.
Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly, 360-407.
Tarr, M. J., \& Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. Cognition, 67(1), 1-20.
Treisman, A., \& Schmidt, H. (1982). Illusory conjunctions in the perception of objects. Cognitive Psychology, 14(1), 107-141.
Vales, C., \& Smith, L. B. (2015). Words, shape, visual search and visual working memory in 3-year-old children. Developmental Science, 18(1), 65-79.
Zhang, Q., Shen, M., Tang, N., Zhao, G., \& Gao, Z. (2013). Object-based encoding in visual working memory: A life span study. Journal of Vision, 13(10), 11.

# Looking for the Cat and Seeing the Dog: Using Visual Search to Study Semantic Knowledge in Children 

Catarina Vales (cvales@andrew.cmu.edu)<br>Layla Unger (lunger@andrew.cmu.edu)<br>Anna V. Fisher (fisher49@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue<br>Pittsburgh, PA 15214 USA


#### Abstract

Semantic knowledge influences various higher-order cognitive processes; therefore, it is important to understand how it changes with development. The Match-to-Sample task is perhaps the most common paradigm for studying changes in semantic knowledge over development, yet this paradigm has a number of limitations. Here we provide initial evidence validating a Visual Search paradigm as a measure of semantic knowledge in preschoolers, and discuss the potential of this paradigm to address the limitations posed by the Match-toSample task to study semantic knowledge development.


Keywords: semantic knowledge; visual attention; visual search; match-to-sample; language; children.

## Introduction

Knowledge about the world supports efficient behavior. For example, knowing that cats are often playful and have sharp claws makes one careful when playing with a cat, and knowing that light bulbs generate light makes one likely to check the light bulb if a lamp stops working. This knowledge about objects, facts, and concepts (Clark, 1973) is thought to be represented in a semantic network that links entities by multiple meaningful relations (McClelland \& Rogers, 2003). Structured semantic knowledge influences multiple cognitive processes, including memory, reasoning, word learning, and visual attention (Bower et al., 1969; Chi et al., 1981; Moores et al., 2003; Roediger \& McDermott, 1995; Xu \& Tenenbaum, 2007), and individual differences in semantic knowledge have been putatively related to the ability to make inferences about novel instances (Coley, et al., 2004; Gobbo \& Chi, 1986; Fisher, 2015).

## Developmental changes in semantic knowledge

As semantic knowledge plays an important role in supporting efficient behavior, there is a large literature investigating what aspects of semantic knowledge change over development to give rise to mature, adult-like behavior. One of the most widely used tasks to study the development of semantic knowledge is the Match-to-Sample task; in this task, participants are shown a target object (e.g. chicken) and asked to match it with one of two options - often a thematic match (an item that is likely to co-occur with the target item, such as pig) and a taxonomic match (an item that belongs to the same stable category of items that share intrinsic properties, such as eagle). Research using this task has documented marked age-related changes starting in the
preschool years in preferences for matching items on the basis of different types of relations (Smiley \& Brown, 1979; Walsh, et al., 1993). However, the Match-to-Sample task presents two main limitations to study developmental changes in semantic knowledge. First, this task cannot be used with young children who are unable to follow verbal instructions and indicate their choices. Prior research examining semantic knowledge development in infants and toddlers has used other tasks (e.g. Arias-Trejo \& Plunkett, 2009; Chow et al., 2017), which may result in confounding developmental changes and task demands. Second, because the Match-to-Sample task requires participants to make explicit judgments about the items, performance in this task might stem not only from knowledge of semantic relatedness but also from other deliberative processes. Indeed, past research with children has shown that performance in the Match-to-Sample task is modulated by the wording of instructions (e.g. Waxman \& Namy, 1997), suggesting that interpreting the pragmatics of the task plays a role in which objects children select. In sum, this task is not ideally suited to study changes in semantic knowledge.

Priming procedures have been used to bypass the limitations of the Match-to-Sample task in adults; however, traditional priming paradigms are difficult to implement with young children. Although several studies have used semantic priming paradigms in infants using looking behavior measures (vs. manual response times) (e.g., AriasTrejo \& Plunkett, 2009), paradigms developed for infants are often not suitable for older children who are not content to inspect visual displays in the absence of an overt task. Below we suggest that a measure of visual attention has the potential to address the limitations outlined above and be used to study semantic knowledge over the lifespan.

## The Visual Search paradigm

Research with adults has used visual attention measures to study knowledge associated with concepts (e.g. Huettig et al., 2011). In these studies, participants are cued about an upcoming target (e.g. by hearing a word or a sentence) and asked to locate the target in a cluttered display. Participants' response times to detect the target (Moores et al., 2003) or their gaze while scanning the array (Huettig \& Altmann, 2005; Mirman \& Magnuson, 2009) are taken as a proxy for the co-activation of concepts related to the target.
A Visual Search paradigm has two main advantages for studying the development of semantic knowledge. First,
visual search tasks have been successfully implemented across the lifespan (Gerhardstein, \& Rovee-Collier, 2002; Vales \& Smith, 2015), in children with developmental disorders (Kaldy et al., 2011), and with varying degrees of language knowledge (Vales \& Smith, 2017). Thus, a Visual Search paradigm is well suited to studying developmental changes in semantic knowledge, reducing differences in task demands from using different tasks with different populations. Second, the Visual Search task allows semantic knowledge to be measured by manipulating the distractors present in the array; because participants do not make explicit judgements about these related distractors, deliberative processes are greatly reduced (see Chun \& Jiang, 1998 for evidence that people are often unaware of experimental manipulations in the visual array). In sum, the Visual Search task is a good candidate to address the limitations of the Match-to-Sample task outlined above.

## The current study

In this study, we seek initial evidence that a Visual Search paradigm can provide estimates of semantic relatedness that are broadly consistent with the estimates from tasks used in prior research. If this is the case, then items judged as more strongly related in a Match-to-Sample task should also more strongly influence performance in a Visual Search task. To this end, we used the Match-to-Sample task to select pairs of target-distractor items. Each potential distractor was tested against a foil that we identified as more distantly related to the target. Measuring the rate at which children chose each distractor versus the foil allowed us to calibrate two distractors for each target, one strongly- and one weakly-related. While the Match-to-Sample procedure suffers from low resolution to detect graded responses, as it allows only binary judgements on each trial, it should still provide a coarse measure of semantic relatedness. We then tested the effect of target-distractor strength in a Visual Search task by asking children to indicate if a target was present in an array of distractors. Across trials, we manipulated the presence of the related distractors; performance in the critical trials in which a related distractor was present was compared with performance in baseline trials in which the related distractors were replaced by items unrelated to the target object. The degree to which children performed more poorly in the presence versus absence of a related distractor was used as the measure of the strength of the perceived relation. As in previous studies with adults and infants (e.g. Chow et al., 2017; Moores et al., 2003), we focused our analyses on target-absent trials, because participants' attention to the target on target-present trials leaves little room for related distractors to influence performance. Target-present trials were included to ensure that children were completing the target identification task.

Prior work using the Match-to-Sample task suggests that children under the age of four are unlikely to consistently select an item related to the target if the foil is a strong competitor (e.g. a visually similar item; Godwin \& Fisher, 2015). As such, we recruited 4 and 5-year-old children, as
this is the youngest age group that we can confidently expect to complete the Match-to-Sample task and thus provide reliable relatedness judgements.

## Methods

We first describe the Match-to-Sample task used to select pairs of items, and report the items selected. Next, we describe the Visual Search paradigm used to test the hypothesis that items judged as more similar in the Match-to-Sample task should also more strongly influence performance in the Visual Search task.

## Stimuli Selection: Match-to-sample task

To select pairs of items with varying strength, we conducted a calibration study with 16 children $\left(M_{\text {age }}=4.9\right.$ years, range $=4.0-5.9,6$ females); children were recruited from local preschools and from a university-affiliated laboratory school in Pittsburgh, PA and tested in a quiet location.

We selected 10 target objects that were likely to be recognized by young children from a prior study investigating the role of semantic relations in a Visual Search task (Moores et al., 2003). For each target, we selected four related items to be tested in the Match-toSample task with the goal of selecting two related items (one strong relation and one weaker relation); the relation strength between each target and each related item was tested in the presence of the same foil, judged by the authors to be a plausible competitor. For example, to test the strength between cat and the items bear, bird, dog, and mouse, participants were presented with the following triads: cat-bear-butterfly, cat-bird-butterfly, cat-dogbutterfly, cat-mouse-butterfly. Two testing sets were created by randomly selecting two of the four triads for each target. Each participant completed one of the sets, for a total of 20 test trials; the order of the trials was randomly determined for each subject, with the constraint that the same target was not presented on consecutive trials. The target was displayed at the top center of a computer screen, and followed by the presentation of the two options (related item and foil); these were presented on the left and right bottom of the screen, with side counterbalanced across trials.

To ensure that children understood the task and were not arbitrarily selecting items, five "catch" trials were randomly placed amid the test trials. These catch triads were intended to introduce no conflict and included items not used in the test trials (e.g. cherries-apple-stapler). On average, children selected the related object on $95 \%(S D=0.12)$ of the catch trials, suggesting that they understood the task (one additional child failed to complete at least 3 out of the 5 catch trials correctly and was not included in the sample). Additionally, because the Match-to-Sample task presented the same foil for each target (and thus children saw the same target- $X$-foil triad more than once), we checked that children were not learning to reject the foil for each target by presenting a block of 10 "control" trials after the experimental trials. In these control triads, the foil used on the test trials was presented against an item judged to be
unrelated to the target (e.g. cat-butterfly-watch). If children were learning to reject the foil over the course of the experimental trials, they should select the non-foil item (the watch in this example); on the other hand, if children were responding to each triad by considering how the items are related, they should select the item that is more strongly related to the target within each triad, even if previously they have not selected that item (the butterfly in the example above). On average, children selected the related item (the butterfly in this example) on $81 \%(S D=0.24)$ of the control trials; this suggests that children were considering how the items were related within each trial.

For each target, we selected two related items that varied in the degree of semantic relatedness, a strongly-related item (selected by most children) and a weakly-related item (selected at a lower rate, at or above chance). Two targets items (banana and cow) failed to produce relations that satisfied these conditions and were not used in the Visual Search task. Table 1 presents the eight sets of items consisting of a target, strongly-related item, and weaklyrelated item to be used in the Visual Search task and the proportion of trials in which each related item was selected in the Match-to-Sample task. On average, strongly-related items were selected on $94 \%(S D=0.09)$ of the trials and weakly-related items were selected on $61 \%(S D=0.10)$ of the trials in the Match-to-Sample task, $t(14)=6.97, p<0.001$. These sets of items were used to create 16 target-related match pairs for the Visual Search task.

Table 1: Proportion of trials each related item was selected in the Match-to-Sample task.

| Target | Strongly-related |  | Weakly-related |  |
| :--- | :--- | :---: | :--- | :--- |
| bike | skateboard | 0.86 | train | 0.71 |
| carrots | rabbit | 0.78 | horse | 0.58 |
| cat | dog | 1.00 | mouse | 0.67 |
| chair | table | 0.89 | bed | 0.57 |
| chicken | turkey | 1.00 | eagle | 0.55 |
| drum | guitar | 1.00 | piano | 0.57 |
| foot | shoe | 1.00 | glove | 0.77 |
| lamp | flashlight | 1.00 | candle | 0.44 |

## Visual search task

Participants. Twenty-four children $\left(M_{\text {age }}=4.8\right.$ years, range $=4.0-5.8,12$ females) were recruited from a universityaffiliated laboratory school in Pittsburgh, PA and tested in a quiet location; these children had not participated in the calibration study. One additional child was recruited but not included in the final sample due to computer malfunction. Children had no known developmental or visual impairments, and English was their only or main language.
Apparatus and Stimuli. Stimuli were presented on a 15.6" touchscreen laptop and responses (accuracy and latency) were recorded using E-Prime (PST, Pittsburgh, PA). To prevent color information from guiding participants' search,
each image was recolored in sepia; recolored images were rendered in a $200 \times 200$ pixel area on a white background. The audio files used to present the spoken names of the targets were recorded by a female native speaker of English.

Design and Procedure. There were six trial types, resulting from all combinations of target presence (present/absent) and related distractor presence (strong present, weak present, related absent), with equal occurrence of each trial type. On each trial, children saw four objects, one on each quadrant of the screen and all equally distant from the center of the screen. Depending on the trial, the four objects were combinations of target, related distractor, and random distractor objects. The order of the test trials was randomly determined for each child, provided that the same target did not appear on consecutive trials. Across trials, the target and related distractors appeared equally often on the left and right side of the screen. A unique token of each concept was used on each trial; for example, each trial probing cat used a different token (see Figure 1 for examples).


Figure 1: Tokens used to instantiate the concept "cat".
Figure 2 shows the temporal order of events on each trial. A "fixation" slide encouraged the child to rest their hands on the table before the trial started (Fig. 2a); the experimenter ensured that the child had their hands down and was looking at the screen before starting a trial. The spoken name of the target (Fig. 2b) was then presented and followed by the search array (Fig. 2c); upon viewing the search array, children had to indicate if the target was present or absent by touching one of two buttons.


Figure 2: Visual Search, trial structure. The target (carrots) is absent and the strong distractor (rabbit) is present.

Children sat in front of the laptop and were told that the goal of the game was to look for pictures on the screen. They were first shown which buttons to touch ("Touch this button if you see the picture on the screen and this button if you do not see the picture on the screen"); the location of the two buttons was counterbalanced across participants. Children were asked to repeat the instructions ("Can you show me which button you touch if you do not see the picture on the screen?") and all children correctly repeated the instructions. Next, children completed 4 "warm-up" trials in which they were familiarized with putting their
hands down during the "fixation" slide, listening to the audio cue, and touching the appropriate button to indicate the target's presence or absence; feedback was provided and children were reminded of the instructions if necessary. Children then completed 48 test trials. The experimenter gave general encouragement throughout the task (e.g. "You are doing great") but did not provide explicit feedback. A short break was introduced every 16 trials during which children could stamp a progress chart.

## Results

To confirm that children were performing the task, we start by analyzing performance on target present trials. Next, we focus on target absent trials to test the hypothesis that items judged as more similar to the target in the Match-to-Sample task also more strongly influence performance in the Visual Search task. We used linear mixed models to analyze the effect of target-distractor relatedness on the time taken to indicate the target's absence (RT). Differently from a traditional analysis of variance, which requires data to be aggregated and incorrect trials to be excluded, a mixed model can include all data and take accuracy into account by modeling the data at the trial level. We included both subject and target item as random factors, that is, varying around a group mean; modeling both subject and target item as random effects is particularly important in experimental designs in which the two factors are fully crossed (Baayen, Davidson, \& Bates, 2008; Jude, Westfall, \& Kenny, 2012).

## Target present trials

Children correctly indicated the target's presence on $83 \%$ ( $S D=0.37$ ) of these trials, suggesting they were trying to locate the target. No main effect of distractor strength on accuracy was found, $F(2,46)=1.78, p=0.18$.

## Target absent trials

Children correctly indicated the target's absence on $91 \%$ ( $S D=0.28$ ) of these trials. This supports the conclusion from target present trials that children were searching for the target. However, there was a main effect of distractor strength on accuracy, $F(2,46)=10.49, \quad p<0.001, \quad$ as participants were less accurate when the strongly-related distractor was present in the array ( $M_{\mathrm{acc}}=0.84, S D=0.36$ ) than when the weakly-related distractor was present in the array ( $M_{\text {acc }}=0.95, S D=0.22$ ) or when all items in the array were unrelated to the target $\left(M_{\mathrm{acc}}=0.95, S D=0.21\right)$. Analyzing RT for correct trials only would exclude different amounts of data from each condition; instead, we include accuracy as a factor in the analyses below. The same pattern of results is found when we consider only correct trials.

Figure 3 depicts mean RT across the three types of trials. Relative to baseline trials ( $M=3.6 \mathrm{~s}$ ), children took over a second longer to judge the target's absence in the presence of a strongly-related distractor ( $M=4.7 \mathrm{~s}$ ), but only slightly longer in the presence of a weakly-related distractor ( $M=3.8 \mathrm{~s}$ ). To assess how the strength of the distractors influenced the ability to correctly indicate the target's
absence, we implemented a linear mixed model using the lme4 package (Bates et al., 2015) in the R environment. We specified accuracy (correct, incorrect) and strength of the distractors (unrelated, weak, strong) as fixed effects, and subject and target item as random effects. The RT outcome variable was log-transformed. Wald $F$ tests and respective $p$-values were calculated using Kenward-Roger's approximation. The model showed only a significant main effect of distractor strength, $F(2,16.19)=3.64, p<0.05$. The main effect of accuracy $[F(1,13.45)=2.40, p=0.14]$ and the interaction between accuracy and distractor strength $[F(2,415.42)=0.14, p=0.87]$ were not significant predictors of RT. Planned contrasts (adjusted using a Bonferroni correction) showed that participants were significantly slower when a strongly-related distractor was present in the array compared to baseline trials in which no related distractor was present, $F(1,16.70)=7.64, p=0.003$. The difference between weakly-related distractor trials and baseline trials, $F(1,16.76)=0.60, p=1.00$, and the difference between strongly-related and weakly-related distractors, $F(1,17.03)=2.77, p=0.34$, were not significant.


Figure 3: Mean RT per trial type in the Visual Search task. Error bars display standard errors of the mean.

## Discussion

The goal of this experiment was to examine if estimates of semantic relatedness as measured by a Match-to-Sample task converged with performance in a Visual Search task. Specifically, for each target item we selected two related items that varied in how strongly they were judged to be related to the target (one strongly-related and one weaklyrelated distractor) in a Match-to-Sample task, and tested the effect of each type of distractor on children's ability to search for that target. Our finding that strongly- but not weakly-related items influenced children's ability to indicate the absence of a target provides initial evidence that Visual Search task performance is influenced by semantic relation strength in children. As such, the Visual Search task is a promising alternative to the Match-to-Sample task, addressing the limitations of this task as outlined above.

## Using Visual Search to study the development of semantic knowledge

Semantic knowledge exerts a pervasive influence on cognitive processes. This knowledge about objects, facts,
and concepts deeply influences how people search for information in the environment (Moores et. al, 2003), retrieve information from memory (Bower et al., 1969), make predictions about objects (Coley et al., 2004), or make sense of events (McNamara \& Kintsch, 1996). Despite the important role that semantic knowledge plays in organizing efficient behavior, we still have a limited understanding of how this knowledge is acquired and how its structure changes with experience. Currently, one obstacle to study the development of semantic knowledge and how it changes with experience is the lack of measures that can be used across the lifespan; as outlined in the Introduction, the most commonly used measure, the Match-to-Sample task, is not well-suited to do so. Visual search paradigms may be a viable alternative as they have been extensively used with adults to study many facets of knowledge (Huettig et al., 2005), and some recent work with toddlers shows similar evidence in younger populations (Chow et al., 2017).

This paradigm also shows promise in capturing individual variation among children. Figure 4 shows participants RT (subtracted from baseline trials RT) to indicate the target's absence from the visual array when the strongly-related and the weakly-related distractors were present. Each bar depicts the relative response time of a single participant, indicating how much that participant was affected by the presence of the related distractor. The range of variability suggests that this task may be a promising tool to study how individual differences in semantic knowledge contribute to individual differences in processes theorized to rely on semantic knowledge (e.g. Fisher, 2015).

Together, the present data suggest that a Visual Search paradigm both complements the Match-to-Sample paradigm and potentially addresses many of its limitations. Below we discuss some of the unresolved questions and important future directions of this work.


Figure 4: RT difference scores, calculated as: related-absent trials (i.e. baseline) minus related present trials.

## Unresolved questions and future steps

The present study revealed no evidence that the presence of a weakly-related distractor influenced search performance. One possibility is that our RT measure was insufficient to detect subtler effects of semantic relatedness. More finegrained moment-to-moment measures taken while children
are looking for the target may detect these effects. Indeed, prior research that found graded effects of knowledge associated with a target concept made use of higher resolution measures, such as eye-tracking or mouse-tracking (e.g. Mirman \& Magnuson, 2009). Another possibility is that children's performance in the Match-to-Sample task is idiosyncratic, and the findings of our calibration study do not generalize across children. To address this possibility, we retested in the Match-to-Sample task all children who participated in the Visual Search task; the estimates were comparable across the two samples, lending some confidence to the estimates from the stimuli selection study.

As this was the first attempt at using the Visual Search task to measure semantic knowledge in young children, we were agnostic as to which relations to probe and thus imposed no constraint when selecting the target-distractor pairs. In the current set of items, some items are linked by multiple relations (e.g. chair and table are both furniture items and often co-occur in the environment, and thus share two types of relations), while other items share only one type of relation (e.g. carrots and rabbit). Previous research using a spatial arrangement task showed that young children seem to consider items that share multiple relations as being more strongly related than items that share only one relation (Unger et al., 2016), and thus it is possible that items that are related in more than one way more strongly influence performance in a Visual Search. Future work can more closely test this prediction by systematically selecting stimuli that vary in the types of relations depicted.

We also did not control for visual similarity between targets and distractors. It is well known that visual similarity influences visual search (e.g. Vales \& Smith, 2015); when selecting the visual tokens, we selected items that were easily discriminable both within- and between-categories, but we did not empirically measure visual similarity. Although it is not trivial to obtain a pure measure of similarity (see Medin et al., 1993; Chow et al., 2017), some recent work has tried to address these issues (e.g. De Groot et al., 2016), and as such it will be important to try to more systematically measure visual similarity in future studies.

## Conclusions

The present study demonstrates the Visual Search paradigm as a feasible approach to investigate the development of semantic knowledge. In contrast with paradigms commonly used in prior research, this paradigm is age-appropriate across a wide developmental range, and greatly reduces the influence of deliberative processes on performance. Thus, the Visual Search task has the potential to shed new light on the development of semantic knowledge and its role in a variety of cognitive processes.

## Acknowledgments

We thank R. Almasi, K. Boyle, H. Do, S. Jahanian, and A. Vande Velde for help with stimuli creation, participant recruitment and testing; and the CMU Children's School, Sacred Hearth Elementary School, and Beth Shalom Early

Learning for making this research possible. This research was supported by the James S. McDonnell Foundation $21^{\text {st }}$ Century Science Initiative in Understanding Human Cognition - Scholar Award (220020401) to A.V.F.

## References

Arias-Trejo, N., \& Plunkett, K. (2009). Lexical-semantic priming effects during infancy. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364(1536), 3633-3647.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Bower, G. H., Clark, M. C., Lesgold, A. M., \& Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. Journal of Verbal Learning and Verbal Behavior, 8(3), 323-343.
Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), Cognitive development and the acquisition of language (pp. 65-110). New York, NY: Academic Press.
Chi, M. T., Feltovich, P. J., \& Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive science, 5(2), 121-152.
Chow, J., Davies, A. A., \& Plunkett, K. (2017). Spokenword recognition in 2-year-olds: The tug of war between phonological and semantic activation. Journal of Memory and Language, 93, 104-134.
Chun, M. M., \& Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive psychology, 36(1), 28-71.
Coley, J. D., Hayes, B., Lawson, C. A., \& Maloney, M. (2004). Knowledge, expectations, and inductive reasoning within conceptual hierarchies. Cognition, 90, 217-253.
De Groot, F., Koelewijn, T., Huettig, F., \& Olivers, C. N. (2016). A stimulus set of words and pictures matched for visual and semantic similarity. Journal of Cognitive Psychology, 28(1), 1-15.
Fisher, A.V. (2015). Development of inductive reasoning. Child Development Perspectives, 9(3), 172-177.
Gerhardstein, P., \& Rovee-Collier, C. (2002). The development of visual search in infants and very young children. Journal of Experimental Child Psychology, 81(2), 194-215.
Gobbo, C., \& Chi, M. (1986). How knowledge is structured and used by expert and novice children. Cognitive Development, 1(3), 221-237.
Godwin, K. E., \& Fisher, A. V. (2015). Inductive generalization with familiar categories: developmental changes in children's reliance on perceptual similarity and kind information. Frontiers in psychology, 6.

Huettig, F., \& Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. Cognition, 96(1), B23-B32. Huettig, F., Rommers, J., \& Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. Acta Psychologica, 137(2), 151-171.
Kaldy, Z., Kraper, C., Carter, A. S., \& Blaser, E. (2011). Toddlers with autism spectrum disorder are more successful at visual search than typically developing toddlers. Developmental Science, 14(5), 980-988.
McClelland, J. L., \& Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. Nature Reviews Neuroscience, 4, 310-322.
McNamara, D. S., \& Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. Discourse Processes, 22(3), 247-288.
Medin, D.L., Goldstone, R.L., \& Gentner, D. (1993). Respects for similarity. Psychological Review, 100, 254278.

Mirman, D., \& Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. Memory \& Cognition, 37(7), 1026-1039.
Moores, E., Laiti, L., \& Chelazzi, L. (2003). Associative Knowledge Controls Deployment of Visual Selective Attention. Nature Neuroscience, 6(2), 182-189.
Roediger, H. L., \& McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 803-814.
Smiley, S. S., \& Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations: A nonmonotonic age trend from preschool to old age. Journal of Experimental Child Psychology, 28, 249-257.
Unger, L., Fisher, A. V., Nugent, R., Ventura, S. L., \& MacLellan, C. J. (2016). Developmental changes in semantic knowledge organization. Journal of Experimental Child Psychology, 146, 202-222.
Vales, C., \& Smith, L. B. (2015). Words, shape, visual search and visual working memory in 3-year-old children. Developmental science, 18(1), 65-79.
Vales, C., \& Smith, L. B. (2017). The words they know: Individual differences in how language affects visual processing. Talk presented at the Biennial Meeting of the Society for Research in Child Development. Austin, Texas.
Walsh, M., Richardson, K., \& Faulkner, D. (1993). Perceptual, thematic, and taxonomic relations in children's mental representations: Responses to triads. European Journal of Psychology of Education, 8, 85-102.
Waxman, S. R., \& Namy, L. L. (1997). Challenging the notion of a thematic preference in young children. Developmental Psychology, 33, 555-567.
Xu, F., \& Tenenbaum, J. B. (2007). Word learning as Bayesian inference. Psychological Review, 114(2), 245.

# Interactivity and Ego Depletion in Insight Problem Solving 

Frédéric Vallée-Tourangeau<br>Department of Psychology, Kingston University<br>Kingston-upon-Thames UNITED KINGDOM KT1 2EE<br>f.vallee-tourangeau@kingston.ac.uk


#### Abstract

In the triangle of coins problem coins are arranged to create a triangle pointing down and the solution involves moving a few coins to change its orientation. The task ecology can be designed such that participants can work on it in a low interactivity environment, maintaining a mental representation of simulated moves, or in a high interactivity environment, thinking with and through a physical model of the problem. These task ecologies involve working memory to a different degree: Problem solving draws more on working memory the lower the degree of physical interaction. Participants first engaged in a writing task that required vigilance to inhibit common word choices, a degree of self regulation designed to induce a so-called ego depletion; participants then worked on the ToC problem in either a low or high interactivity environment. Solution rates were determined by level of interactivity; the preceding depletion experience did not impact performance.


Keywords: Interactivity, ego depletion, insight problem solving, working memory

## Introduction

The relative contribution of working memory in analytic and insight problem solving has been explored using a broad range of methodologies. Prototypical analytic problems such as those requiring arithmetic operations would involve the maintenance and updating of interim results, the strategic allocation of attentional resources, and retrieval of long term memory knowledge, processes that draw heavily on working memory resources; high working memory capacity (WMC) is a reliable predictor of analytic problem solving performance (Wiley \& Jarosz, 2012). In turn, insight problems are designed to resist initial analytic efforts or the direct transfer of long term memory knowledge. They are presented in a manner that trigger prepotent responses, but these lead to an impasse. Solving these problems then requires letting go of the incorrect interpretation. Less rather than more focus and attentional control might facilitate abandoning the incorrect interpretation, and hence a lower or transiently lowered WMC may better predict insight problem solving success (DeCaro \& Beilock, 2010).

However, there is also evidence that working memory capacity is involved in solving insight problems (Gilhooly \& Fioratou, 2009). For example, verbal working memory scores predict solution rates for compound remote associates (Chein \& Weisberg, 2014) and spatial working memory scores correlate significantly with solution rates for the nine-dot problem (Chein, Weisberg, Streeter, \& Kwok, 2010). Weisberg's (2015) integrated framework stresses the central role of analytic processes in solving insight problems, processes that draw on WMC.
A temporary reduction in deliberate executive function skills using a self-control exercise prior to engaging in an insight problem solving task might offer an interesting means to throw light on these somewhat conflicting findings. Hoffman, Schmeichel and Baddeley (2012) proposed that key features of executive functioning subserve self-regulation. So-called ego depletion tasks involve participants actively inhibiting a response set; these inhibitory efforts have negative aftereffects in subsequent tasks that require executive functions. For example, in Schmeichel (2007) participants' backward digit span was significantly lower after they engaged in a writing task that prohibited the use of the letters $a$ and $n$ than after an unconstrained writing task. Such findings support a fuel metaphor: executive control processes rely on a limited resource that fluctuates as a function of effort, rest and, more controversially, glucose level. The transient depletion of executive functions may impact analytic and insight problem solving differently. For example, performance on a mental arithmetic problem that requires temporary storage, retrieval of long term arithmetic knowledge and the strategic allocation of attentional resources, might be influenced by prior exposure to an ego-depletion manipulation. The prediction for insight problem solving however depends on the purported involvement of WMC. If a looser focus and 'leaky' attention (Wiley \& Jarosz, 2012) are important, then ego depletion might actually enhance insight problem solving. In contrast, if, as Weisberg (2015) contends, analytic processes are implicated in insight problem solving, then an ego-depletion manipulation should have a negative aftereffect on performance.

## Interactivity

The debate concerning the mental processes and capacities involved in problem solving reflects a commitment to methodological individualism (ValléeTourangeau \& Vallée-Tourangeau, 2017): that the locus of cognition is person, or more specifically, skull bound. Problem solving outside the laboratory naturally involves interacting with the world, recruiting artefacts, building models of proto solutions, assembling non-mental resources that scaffold creativity and problem solving. The world is there to see and act upon, its dynamic configuration triggers different actions and guides attention. As such, then, the working memory burden of keeping a detailed representation of the problem is not the same when participants can interact with a physical model of the problem. In some sense, interactivity might result in a functional increase in WMC (Vallée-Tourangeau, Sirota, Vallée-Tourangeau, 2016), and profiling participants in terms of their WMC might not be as informative as profiling the working memory resources of the agent-environment system created through interactivity.


Figure 1: The triangle points down: Which three coins should be moved to make the triangle point up? Solving the triangle of coins problem involves moving the coins that mark the three vertices.

Real-world interactivity can be scaled down under laboratory conditions if the problem solving environment, or the cognitive ecosystem (Hutchins, 2010) affords interacting with a physical model of the problem (ValleeeTourangeau, Sirota, Vallée-Tourangeau, \& Makri, 2015). For example, Fleck and Weisberg (2013) used the triangle of coins problem (ToC see Fig. 1), among other insight problems, to explore problem solving strategies (based on verbal protocols). For this problem, they supplied participants actual coins to manipulate to determine the solution. They quote, at length (see p. 452) the protocol of one participant who works on the problem by initially moving coins in a trial and error fashion. The changes in the problem configuration guide and constrain the problem solving trajectory. Fleck and Weisberg also
describe what they term 'data-driven' restructuring, or how a productive interpretation of the problem is triggered by manipulating the physical model; for example, an exploratory rotation of the entire model helped a participant notice that only the three vertices should be moved and the hexagonal core should be untouched (see pp. 452-453). Thus a productive strategy to solve the problem was triggered by changes in the physical model of the problem, not through the mental manipulation of a problem representation.

## The Present Experiment

The present experiment explored the impact of an ego depletion task and level of interactivity on insight problem solving using the ToC problem. The experiment employed a $2 \times 2$ between subjects design. The first independent variable was the nature of a six-minute writing task before participants tackled the ToC problem: Half the participants had to closely monitor and inhibit certain responses to ensure that words containing the letters $a$ or $n$ were not used, while the other half wrote freely. The second independent variable was the level of interactivity. After the writing task, half of the participants worked on the ToC in a low interactivity condition, that is by looking at a static visual display of the problem and dictating possible moves to an experimenter. They did so by keeping their hands palm down in front of them and could not point at the coins or simulate movements with their fingers. The other half of the participants worked on the problem in a high interactivity condition: Participants were presented with a physical model of the problem and they could touch, point to or move the tokens in determining which three could be moved to change the triangle's orientation. The low interactivity condition draws more heavily on WMC since participants must mentally simulate move, keeping track of the simulated movements of certain tokens, while evaluating the movement of which other token(s) would mentally change the orientation of the triangle. None of this mental activity could be supported with complementary actions (Kirsh, 1995). To the extent that the constrained writing exercise prior to working on the ToC problem depletes executive functions and solving the ToC problem requires analytic processes that draw on WMC, then performance should be better after the free writing session than after the constrained writing session. On the other hand, if solving the ToC does not proceed from focused deliberate analytic efforts, then participants might actually perform better with prior exposure to the ego depletion task, that is a transient reduction in executive functions might be beneficial. Participants working on the ToC in the high interactivity condition are not confronted with the same kind of WMC taxing environment, and as such we predict a much higher rate of problem solving success in the high than in the low interactivity condition. In the high interactivity condition,
participants think with and through the physical model of the problem. Transient executive functions depletion might have little or no impact on problem solving performance.

## Method

## Participants

Eighty undergraduate and postgraduate students (60 females) received course credits for their participation ( $M_{\text {age }}=24.2, S D=7.3$ ).

## Procedure

Participants were tested individually in a quiet cubicle. The experimental session was composed of three parts. Once an information sheet was read and understood and a consent form signed, participants first engaged for six minutes in a writing exercise modelled after the one reported in Schmeichel (2007). They were instructed "to write a short story describing a trip you have taken recently". Half of the participants experienced the depletion version of the task, where they were further instructed that the letters ' A ' or ' N ' could not be used: "you must pay close attention to the words you are using and aim to describe the trip with words that don't use these two letters". The other participants wrote freely without having to inhibit word choice responses.
During the second part of the experiment, all the participants were shown a sheet of paper (size A4) on which 17 digits were randomly printed. They were instructed to add these numbers as quickly and accurately as possible; however, participants had to keep their hands palm down on the table top and hence the mental arithmetic could not be supplemented or supported by complementary actions, such as touching the printed numbers of pointing at them.
In the third and final part of the experiment participants worked on the triangle of coins problem for five minutes; half of the participants were allocated to the low interactivity condition, half to the high interactivity condition. The problem was illustrated on a 9 x 9 grid printed on a sheet of paper: Columns were labelled with letters (A-I) and rows with numbers (1-9). Ten tokens were arrayed on that grid, each token labelled with an individual letter (see Fig. 2); the solution to the problem involved moving tokens R, W, A to cells E4, B7, and H7 to reorient the triangle such that it pointed up rather than down. Participants in the low interactivity condition worked on the problem with hands palm down on the table top, and voiced their proposed moves to the experimenter in groups of three moves: They would name the token and then the cell coordinate where it should be moved. The experimenter noted the moves on a record sheet hidden from the participants' view, and provided feedback. It's important to note that the problem configuration always
remained the same, that is participants had to mentally project moves on the grid and verbalise these moves, while looking at the instruction sheet show in Figure 2; the experimenter never modified the triangle as such, and only provided feedback. In addition, feedback provided was all or none, that is participants were not informed if the projected position of one or two tokens was right on a given trial. In the high interactivity condition, participants worked with a laminated 9 x 9 grid (measuring 21 cm x 29 cm ) with rows and columns labelled as in the low interactivity condition. Ten tokens ( 2.2 cm in diameter) were arrayed as in the low interactivity condition, creating a triangle pointing down; each token had a letter printed on it, just as it did in the low interactivity condition. The token and grid coordinates helped the experimenter record the participants' moves, but the participants were not required to verbalise moves by identifying tokens and cell destinations. Rather participants were invited "to touch the tokens" and "trace their movement with your finger". If after moving the three tokens, the pattern created did not result in the correct answer, the experimenter put the tokens back to their original place, and participants could try to move a new set of three tokens

The triangla points to the bottorn of the board.


Which set of 3 tokens can be moved to make the triangle peint to the top of the boand? To solve the problem, think of a set of three, announce their movement to the experimenter. The experimerter will note your proposed set and will give you foedback. if the set you propose is incorrect, you will be asked to propose another set of three. You have five minutes to propose as many sets of three as you want.

To announce a set of moves, you need to say which token you wart to move. and idersty the coortinates on the grid of where you think the token should be placed. For example, imagine you worked on a different probiem and warted to propose that token $Z$ move from $A 1$ to $A 2$, then you would say $Z$ from $A 1$ to $A Z^{\prime}$.


As with the riertal arithmetic task, you must place your handa in front of you flat on the tabietop. You are not allowed to touch the lokens or use your fingers to indicate where a token should be placed. Aather, you till the experimenter which set of three lokene should be moved and where.
Figure 2: The problem instructions in the low interactivity condition.

The experimental design was thus a 2 (Depletion: constrained writing or free writing) x 2 (Interactivity: low, high) between subjects design. Participants were allocated randomly to each of the four experimental conditions.

## Results

The mean number of words written in the constrained-ego depletion version of the story writing task ( $M_{\text {depletion }}=28.7$, $S D=12.26$ ) was significantly lower than in the unconstrained free version of the task ( $M_{\text {no depletion }}=122.7$, $S D=25.48), t(56)=-20.8, p<.001$ (the degrees of freedom were adjusted to account for heterogeneity of variance). Thus participants complied with the task instructions and struggled to write a story when they had to suppress words containing the letters $a$ or $n$. However, performance on the mental arithmetic task, immediately following the story writing task, was unaffected by the prior exposure to the ego depletion manipulation in terms of the magnitude of absolute calculation errors $\left(M_{\text {depletion }}=\right.$ $\left.3.45, S D=5.16, M_{\text {no depletion }}=5.23, S D=8.18\right) t(66)=-$ $1.16, p=.250$, or latency $(\mathrm{s})$ to solution $\left(M_{\text {depletion }}=71.1\right.$, $\left.S D=42.64, M_{\text {no depletion }}=79.4, S D=57.81\right), t(72)=-0.729$, $p=.469$ (degrees of freedom were adjusted to account for heterogeneity of variance).
Performance on the triangle of coin problem is illustrated in Figure 3. With low interactivity, 4 participants (or 20\%) solved the problem following the constrained writing task, while 5 (or $25 \%$ ) did so after the unconstrained writing task. With high interactivity, 14 participants (or $70 \%$ ) and 10 (or $50 \%$ ) solved the problem following the constrained and unconstrained writing task, respectively. Summing across depletion levels, more participants ( $60 \%$ ) solved the ToC problem in the high interactivity condition, than in the low interactivity condition ( $23 \%$ ), $\chi^{2}(1, N=80)=11.61, p$ $=.001$; solution rates between the two high interactivity conditions did not differ significantly, $\chi^{2}(1, N=40)=$ $1.67, p=.196$.


Figure 3: Percentage of correct solutions for the triangle of coins problem in the low and high interactivity task environment after engaging in a self-regulation task (depletion) or not (no depletion).

The solution rate data were analysed using a binary logistic regression. The outcome variable was the probability of solving the triangle of coin problem. Three models were tested: the first included only depletion as a predictor variable, the second included both depletion and interactivity as predictors, and the third model included depletion, interactivity, and their interaction as predictors. The first model was not significant, $\chi^{2}(1)=0.465, p=$ .495; adding interactivity produced a significant model, $\chi^{2}$ (2) $=12.026, p=.001$, Nagelkerke $R^{2}=.195$; however, adding an interaction term did not increase the significance of the model, $\chi^{2}(1)=1.280, p=.258$. The only significant predictor of success using the Wald criterion was level of interactivity ( $p=.001$ ) with an odds ratio of 5.231 (see Table 1).

Table 1: Summary of Binary Logistic Regression with the Model Involving Depletion and Interactivity as Predictors.

|  |  | $95 \%$ C.I. for OR |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Variable | $B$ | $S E$ | OR | Lower | Upper | Wald | $p$ |
|  |  |  |  |  |  |  |  |
| Depletion | 0.363 | 0.494 | 1.438 | 0.546 | 3.789 | 0.541 | .462 |
| Interactivity | 1.655 | 0.500 | 5.231 | 1.962 | 13.947 | 10.937 | .001 |

## Discussion

The triangle of coins problem was difficult to solve within the time allocated for participants in the low interactivity conditions; the success rate was very similar between those who had undergone the ego-depletion manipulation and those who had engaged in the free writing exercise. In turn, the success rates were substantially higher for participants in the high interactivity conditions. The interaction between ego-depletion and level of interactivity was not a significant predictor of the solution rate, however.
The level of interactivity afforded by the thinking environment substantially influenced problem solving performance, and corroborates recent findings with other types of insight problems (e.g., matchstick algebra, the 17 animals problem, see Weller, Villejoubert, \& ValléeTourangeau, 2011; Vallée-Tourangeau, Steffensen, ValléeTourangeau, \& Sirota 2016). The overall solution rate in the high interactivity condition in this experiment is similar to the one reported in Fleck and Weisberg (2013) who presented the ToC problem in a high interactivity environment, even if their work did not explicitly explore and contrast levels of interactivity in insight problem solving. A physical and modifiable model of the problem reduces working memory demands because changes in the problem configuration cue new actions and guide attention. Moves need not be premeditated, and need not be mentally simulated; participants observe the results of their action, and changes in the world, that is changes in the physical model of the problem, convey new information. Participants can more readily see how to solve the problem, they need not mentally represent possible changes, imagine their outcome, and while maintaining
this simulated modification to the problem in working memory, project and simulate the next move.
As for the ego-depletion manipulation, whether the constrained writing task achieved its purpose of temporarily reducing executive functions, remains uncertain. On the one hand, word production was much lower in the constrained writing condition than with the free writing task. This suggests that participants complied with the task instructions and struggled to write. Assuming that the sustained inhibition of common word choices led to a depletion in executive functions, the predicted impact of such transient depletions on insight problem solving depends on the prominence attributed to deliberate analytical processes. What is less controversial is how such an ego-depletion manipulation should have influenced performance on the mental arithmetic task immediately following the writing exercise, but it did not; if anything absolute calculation errors were marginally lower following ego depletion. As for the ToC problem, performance in the low interactivity conditions was very poor, and such a floor effect might have masked any influence of ego depletion. Thus, on the one hand, the potential window on the importance of WMC in insight problem solving that a purported depletion in executive functions might offer was undermined by the very low rates of success in both low interactivity conditions. On the other hand, the mental arithmetic data suggest that the depletion manipulation did not work. In the final analysis, the controversy surrounding the very existence of the egodepletion phenomenon (Hagger, Chatzisarantis, Alberts, Anggono, Batailler, et al., 2016) suggests that such a manipulation does not offer an interesting tool to gauge the importance of WMC and executive functions in problem solving. There is also the possibility that the mental effort invested in the mental arithmetic task depleted executive functions more so than the constrained writing exercise, and the low solution rate in the low interactivity conditions reflects this depletion. Thus, possible avenues for future research involve eliminating the intervening mental arithmetic task (and devising an alternative ego depletion manipulation check) or employing a more exacting depletion task such as a computation span test of the kind used to gauge WMC.
Fleck and Weisberg (2013) reported that some participants solved the ToC problem through an analytic and incremental strategy while for others the solution appeared to reflect a non-incremental insight. On the basis of concurrent verbal protocols or post-participation interviews future research could thus better determine the strategies and processes employed by a given participant, and make more specific predictions as to the degree of WMC involvement in solving that particular problem (I thank Robert Weisberg for this point).
In light of the controversy surrounding the concept of depletion and its potential negative aftereffects on performance, perhaps an altogether more productive
research programme could look at burdening working memory with a secondary task (e.g., Lavric, Forstmeier, \& Rippon, 2000) to determine how it would affect insight problem solving as a function of the level of interactivity afforded by the task environment. Recent work suggests that the impact of articulatory suppression on mental arithmetic was much greater in a low than in a high interactivity environment (Vallée-Tourangeau et al., 2016). Such a paradigm might be usefully employed with insight problem solving, not only as means to adjudicate the different proposals concerning the involvement of working memory but also to assess how working memory capacity is functionally enhanced in a high interactivity environment.

## Concluding Remarks

Individual differences in cognitive capacities and thinking dispositions are often measured to throw light on thinking processes (e.g., Stanovich \& West, 1998). Correlational and latent variable analyses are conducted to determine the underlying factors that best account for thinking performance. This strategy is employed in problem solving research as reviewed earlier (see also Chuderski, 2014). The tests designed to measure cognitive capacities-such as working memory-and the problem solving tasks typically involve little or no interactivity with physical problems. Thus, the commitment to methodological individualism is implicitly reinforced rather than challenged. The substantial improvement in problem solving performance in the high interactivity environment observed in this and other experiments (e.g., ValléeTourangeau, Abadie, \& Vallée-Tourangeau, 2015; ValléeTourangeau et al., 2016) suggests that researchers should be mindful of the importance of interacting with a physical model when solving a problem.

## Acknowledgements

I would like to thank Juan Aracena and Paige Edwards for recruiting and running the participants, as well as Emma Henderson, Gaëlle Vallé-Tourangeau and Robert Weisberg for thoughtful comments on a previous version of this manuscript.

## References

Chein, J. M., \& Weisberg, R. W. (2014). Working memory and insight in verbal problems: Analysis of compound remote associates. Memory \& Cognition, 42, 67-83.
Chein, J. M., Weisberg, R. W., Streeter, N. L., \& Kwok, S. (2010). Working memory and insight in the nine-dot problem. Memory and Cognition, 38, 883-892.
Chuderski, A. (2014). How well can storage capacity, executive control, and fluid reasoning explain insight problem solving. Intelligence, 46,258-270.
DeCaro, M. S., \& Beilock, S. L. (2010). The benefits and perils of attentional control. In B. Bruya (Ed.),

Effortless attention: A new perspective in the cognitive science of attention and action (pp. 51-73). Cambridge, MA: MIT Press.
Fleck, J. I., \& Weisberg, R. W. (2013). Insight versus analysis: Evidence for diverse methods in problem solving. Journal of Cognitive Psychology, 25, 436-463.
Gilhooly, K. J., \& Fioratou, E. (2009). Executive functions in insight versus non-insight problem solving: An individual differences approach. Thinking and Reasoning, 15, 355-376.
Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., \& Zwienenberg, M. (2016). A multi-lab pre-registered replication of the ego-depletion effect. Perspectives on Psychological Science, 11, 546-573.
Hofmann, W., Schmeichel, B. J., \& Baddeley, A. D. (2012). Executive functions and self-regulation. Trends in Cognitive Science, 16, 174-180.
Hutchins, E. (2010). Cognitive ecology. Topics in Cognitive Science, 2, 705-715.
Kirsh, D. (1995). Complementary strategies: Why we use our hands when we think. In J. M. Moore \& J. L. Lehman (Eds.), Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society (pp. 212217). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
Lavric, A., Forstmeier, S., \& Rippon, G. (2000). Differences in working memory involvement in analytical and creative tasks: An ERP study. NeuroReport, 11, 1613-1618.
Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. Journal of Experimental Psychology: General, 136, 241-255.
Stanovich, K. E., \& West, R. F. (1998). Individual differences in rational thought. Journal of Experimental Psychology: General, 127, 161-188.
Vallée-Tourangeau, F. (2014). Insight, materiality and interactivity. Pragmatics \& Cognition, 22, 27-44.
Vallée-Tourangeau, F., Sirota, M., \& Vallée-Tourangeau, G. (2016). Interactivity mitigates the impact of working
memory depletion on mental arithmetic performance. Cognitive Research: Principles and Implications, 1, 26.
Vallée-Tourangeau, F., Steffensen, S. V., ValléeTourangeau, G., \& Makri, A. (2015). Insight and cognitive ecosystems. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, \& P. P. Maglio (Eds.), Proceedings of the Thirtyseventh Annual Conference of the Cognitive Science Society (pp. 2457-2462). Austin, TX: Cognitive Science Society.
Vallée-Tourangeau, F., Steffensen, S. V., ValléeTourangeau, G., \& Sirota, M. (2016). Insight with hands and things. Acta Psychologica, 170, 195-205.
Vallée-Tourangeau, G., \& Vallée-Tourangeau, F. (2017). Cognition beyond the classical information processing model: Cognitive interactivity and the Systemic Thinking Model (SysTM). In S. J. Cowley, \& F. Vallée-Tourangeau (Eds.), Cognition beyond the brain: Interactivity, cognition and human artifice ( $2^{\text {nd }}$ Edition, pp. 133-154). Dordrecht: Springer.
Vallée-Tourangeau, G., Abadie, M., \& Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. Journal of Experimental Psychology: General, 144, 581-603.
Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In R. J. Sternberg \& J. E. Davidson (Eds.), The Nature of Insight (pp. 157-196). Cambridge MA: MIT Press.
Weisberg, R.W. (2015). Toward an integrated theory of insight in problem solving. Thinking \& Reasoning, 22, 5-39.
Weller, A., Villejoubert, G., Vallee-Tourangeau, F. (2011). Interactive insight problem solving. Thinking \& Reasoning, 17, 429-439.
Wiley, J., \& Jarosz, A. F. (2012). Working memory capacity, attentional focus, and problem solving. Current Directions in Psychological Science, 21, 258262.

# A computational model for decision tree search 

Bas van Opheusden, Gianni Galbiati, Zahy Bnaya, Yunqi Li \& Wei Ji Ma<br>Center for Neural Science \& Department of Psychology, New York University<br>\{basvanopheusden,gianni.galbiati,zahy.bnaya,yunqi.li,weijima\} @ nyu.edu


#### Abstract

How do people plan ahead in sequential decision-making tasks? In this article, we compare computational models of human behavior in a challenging variant of tic-tac-toe, to investigate the cognitive processes underlying sequential planning. We validate the most successful model by predicting choices during games, two-alternative forced choices and board evaluations. We then use this model to study individual skill differences, the effects of time pressure and the nature of expertise. Our findings suggest that people perform less tree search under time pressure, and that players search more as they improve during learning.


Keywords: Sequential decision-making, Behavioral modeling, Expertise

## Introduction

Imagine you are deciding if you should run for President of the United States in 2020. To make that choice, you have to consider a sequence of future decisions. Will you run as a Republican, Democrat or Independent? If Democrat, will you run as a moderate or progressive candidate? What positions will you take on abortion or gun control? How will you distinguish yourself during the primaries? What line of attack will you choose in the Presidential Debates? You face a sequence of decisions, which together determine your electoral success. In short, you have to explore a decision tree.

Although the computations underlying human decisionmaking are extensively studied, the process by which people explore decision trees is less understood. Most work focuses on the neural implementation of learning and decisionmaking in small decision trees (Solway \& Botvinick, 2015; Simon \& Daw, 2011). However, with more choices and more available options, the decision tree grows exponentially, and people need to prune the tree (Huys et al., 2012).

There exists a large literature exploring human decisionmaking in chess, starting with de Groot's seminal article (A. D. de Groot, 1946). One central question in this literature is whether the superior performance of experts relies primarily on enhanced pattern recognition (Chase \& Simon, 1973), increased tree search (Holding, 1985), or both. The relation between tree search and expertise is especially controversial, with both positive (Campitelli \& Gobet, 2004) and negative (A. D. de Groot, 1946) results.

In this article, we investigate sequential decision-making in a two-player board game, which is much simpler than chess, but much more complex than traditional decision-making tasks. We develop a computational model that predicts people's choices on individual trials, and fit this model to data from individual participants. We then ask whether the computations performed by our model mimic the process by which people arrive at their decisions. Finally, we use our model to investigate the nature of expertise in our game.

## Experiments

Task. To investigate the computations underlying sequential decision-making, we collected data from people playing a variant of tic-tac-toe, in which players need to make 4-in-arow on a 4-by-9 board (figure 1A). Despite these simple rules, the game is surprisingly challenging and fun to play. Because the game is deterministic without hidden information, it is theoretically solvable. Using alpha-beta pruning and threat tree search (Allis et al., 1994), we were able to derive a weak solution: the first player can force a win by opening on the central square. However, with perfect defense, the second player can delay the win for 17 moves.


Figure 1: Task. A. Two players take turns placing black or white pieces on a 4-by-9 board, and the first to achieve 4-in-a-row (horizontal, diagonal or vertical), wins the game. B. In the 2AFC task, participants see a board and two candidate moves, and indicate their preferences. C. In the evaluation task, participants see a board position and report their estimated winning chances on a 7 -point scale.

Participants. We conducted four experiments: human-vshuman ( $N=40$ participants), generalization ( $N=40$ ), time pressure $(N=30)$, and learning $(N=30)$. We recruited participants through the NYU psychology research participant system, flyers, a sign-up link on our lab webpage or personal communication. We did not collect demographic data. We compensated participants 12 per hour, but did not incentivize task performance.
Procedure. In the human-vs-human experiment, we divided participants into pairs. Participants in each pair played games against each other without time constraints for 50 minutes, switching colors every game. In the generalization experiment, participants performed three tasks: playing the game against a computer opponent for 30 minutes, 82 trials of a two-alternative forced-choice (2AFC) between moves in a given board position (figure 1B), and 82 board evaluation trials, in which they rated their winning chances in given board positions on a 7 -point scale (figure 1C). The time pressure experiment was identical to the human-vs-computer component of the generalization experiment, except that for each game, we added a time limit randomly selected between 5, 10 or 20 seconds per move. If participants exceeded the time
limit, they lost the game. The learning experiment consisted of 5 sessions, no more than 3 days apart. In sessions $1,3 \&$ 5 , participants played against computers for 30 minutes, then completed 60 trials each of the 2 AFC and evaluation tasks. In session $2 \& 4$, they played against computers for the entire 50-minute session.

In all human-vs-computer games, the computer opponents implemented an early version of our computational model for people's decision-making process, with parameters adapted from fits on human-vs-human games. We created 30 AI agents, grouped by playing strength into 6 groups of 5 agents each, and matched participants with AI opponents through a one-up, one-down staircase procedure.

In the 2 AFC and evaluation task, each participant completed the same trials in shuffled order. We selected board positions and move options that maximize an approximation to mutual information between model parameters and move choice, in order to present participants with interesting and informative choices.

## Model

Value function. The core component of our model is an evaluation function $V(s)$ which assigns values to board states $s$. We use a weighted linear sum of 5 features: center, connected 2-in-a-row, unconnected 2-in-a-row, 3-in-a-row and 4-in-arow. The center feature assigns a value to each square, and sums up the values of all squares occupied by the player's pieces. This value of each square is inversely proportional to its Euclidean distance from the board center. The other features count how often particular patterns occur on the board (horizontally, vertically, or diagonally):
Connected 2-in-a-row: two adjacent pieces with enough empty squares around them to complete 4-in-a-row.
Unconnected 2-in-a-row: two non-adjacent pieces which lie on a line of four contiguous squares, with the remaining two squares empty.
3-in-a-row: three pieces which lie on a line of four contiguous squares, with the remaining square empty. This pattern represents an immediate winning threat.
4-in-a-row: four pieces in a row. This pattern appears only in board states where a player has already won the game.

We associate weights $w_{i}$ to these features, and write

$$
V(s)=c_{\text {self }} \sum_{i=0}^{4} w_{i} f_{i}(s, \text { self })-c_{\text {opp }} \sum_{i=0}^{4} w_{i} f_{i}(s, \text { opponent })
$$

where $c_{\text {self }}=C$ and $c_{\text {opp }}=1$ whenever the player is to move in state $s$, and $c_{\text {self }}=1$ and $c_{\mathrm{opp}}=C$ when it is the opponent's move. The scaling constant $C$ captures value differences between "active" and "passive" features. For example, a three-in-a-row feature signals an immediate win on the player's own move, but not the opponent's.

Tree search. The evaluation function guides the construction of a decision tree with an iterative best-first search algorithm. Each iteration, the algorithm chooses a board position to explore further, evaluates the positions resulting from each
legal move, and prunes all moves with value below that of the best move minus a threshold. After each iteration, the algorithm stops with a probability $\gamma$, resulting in a geometric distribution over the total number of iterations.

Noise. To account for variability in people's choices, we add three sources of noise. Before constructing the decision tree, we randomly drop features (at specific locations and orientations), which are omitted during the calculation of $V(s)$ anywhere in the tree. During tree search, we add Gaussian noise to $V(s)$ in each node. Finally, we include a lapse rate $\lambda$.

The components of our computational model are inspired by behavioral studies of human decision-making. Tree search, as a mechanism whereby people mentally simulate the consequences of available actions, is similar to "levelK reasoning" (Arad \& Rubinstein, 2012) in behavioral economics. In other decision-making tasks, people have been shown to prune away options leading to immediate losses but long-term gains (Huys et al., 2012). Feature dropping reflects shift in endogenous attention (to spatial locations, orientation or feature types), corresponding to participants overlooking relevant features on the board. Finally, feature-based evaluation functions, value noise and lapse rates are all common in reinforcement learning.

There also exists neural evidence consistent with our model. In rats, dynamic search and exploration of possible paths at junctions in a T-maze have been linked to preplay sequences in hippocampal place cells (Johnson \& Redish, 2007). In humans, tree search is associated with neural activity in the ventral striatum (Simon \& Daw, 2011) and ventromedial prefrontal cortex (Lee, Shimojo, \& ODoherty, 2014).

## Methods

Estimating task performance. To quantify task performance in human-vs-computer games, we use the Elo rating system (Elo, 1978), which estimates playing strength from game results, independent of the moves played. We append the results of games from all 4 experiments to a computer-vs-computer tournament, and estimate ratings jointly for all humans and computers with a Bayesian optimization algorithm (Hunter, 2004). To calculate performance in the 2AFC task, we calculate the agreement between a participant's choices and those of an optimal agent with random tie-breaking. In the evaluation task, we define performance as the correlation between a participant's choices and the optimal rankings.

Estimating model parameters The model has 10 parameters: the 5 feature weights, the active-passive scaling constant $C$, the pruning threshold, stopping probability $\gamma$, feature drop rate $\delta$ and the lapse rate $\lambda$. We infer these parameters for individual participants and individual learning sessions or time limit conditions with maximum-likelihood estimation. We estimate the log probability of a participant's move in a given board position with inverse binomial sampling (M. H. de Groot, 1959), and optimize the log-likelihood function with multilevel coordinate search (Huyer \& Neu-
maier, 1999). We account for potential overfitting by reporting 5 -fold cross-validated log-likelihoods, with the same testing-training splits for all models.

## Model comparison

To test how well our model predict participants' choices, we compare its log-likelihood on human-vs-human games to that of 25 alternative models (figure 2). We test four categories of alternative models: lesions, generated by removing model components; extensions, generated by adding new model components; modifications, generated by replacing a model component with a similar implementation; and controls, which are structurally different from the main model.


Figure 2: Cross-validated log-likelihood/move for our main model and 25 alternatives on the human-vs-human data. The bars show mean and s.e.m. across participants $(N=40)$. The main model fits better than lesions, most controls and some modifications, and approximately equally good as extensions or some other modifications.

Lesions. We create lesion models by forcing either one of the feature weights to zero, or removing the feature dropping, pruning, value noise, active-passive scaling or the entire decision tree. The no-tree model evaluates the positions after each possible move, and chooses the one with maximum value. It contains feature dropping and value noise but no pruning.

Extensions. We consider extending the model with a feature that recognizes to a three-piece pattern arranged in a triangle, or multiplying the weights for diagonally and vertically oriented features by scaling constants $c_{\text {diag }}$ or $c_{\mathrm{vert}}$, respectively. Alternatively, we extend the main model by allowing feature drop rates to differ between features of different types (2-in-a-row, 3-in-a-row, etc) or orientations. Finally, we test a model in which all weights for the opponent's features are scaled by a factor $c_{\text {opp }}$, which thereby controls the balance between attack and defense.

Modifications. We modify the model by fixing the number of iterations of the search algorithm to a constant instead of the geometric distribution prescribed by the main model. Alternatively, we amend the search process to explore each branch of the tree up to fixed depth, or the pruning rule to keep only the $K$ best moves (according to the evaluation function), where the branching factor $K$ is again fixed. For a more drastic modification, Monte Carlo Tree Search (MCTS) estimates state values not by calling the evaluation function $V(s)$, but by aggregating outcomes of simulated games between notree agents. It also extends the best-first search algorithm by adding a term that favors exploration (investigating unexplored moves) over exploitation (further investigating already explored moves). We consider fixing the feature weights to the optimal solution, i.e. those weights that maximize the correlation between $V(s)$ and the game-theoretic value of the position $s$. Finally, we modify the attention mechanism from dropping random features from the evaluation function to dropping random branches from the decision tree.

Controls. We consider MCTS with completely random playouts, or a mixture model between optimal and random play. The optimal agent enumerates all candidate moves that preserve the game-theoretic value of the position, and chooses randomly between them. Another control model, labeled soft-max, assigns a value to each square on the board (enforced to obey reflection/rotation symmetry), and chooses a move with a softmax decision rule, constrained to unoccupied squares.

All lesioned models fit worse than the full model. The most impactful lesions are specific features (3-in-a-row, connected 2-in-a-row and center) and sources of variability (value noise and feature dropping). Lesioning the pruning mechanism or the entire tree search algorithm has a less dramatic effect, which can be partially explained by parameter trade-offs. Finally, some lesions (active-passive scaling, unconnected 2-in-a-row and 4-in-a-row) cause only small reductions in loglikelihood. Most modifications also worsen the main model, but the Monte Carlo Tree Search model is equally good and the "fixed iterations" model slightly outperforms it. The model extensions also slightly increase the main model's performance. Finally, all control models fit much worse than the main model.

Unfortunately, the model comparison does not reveal a unique best-fitting model, meaning that we did not collect enough data to determine precise details of people's thought process. For example, we cannot distinguish between tree search algorithms (best-first search or MCTS) or determine specifics of the best-first search algorithm (pruning and number of iterations). Alternatively, different participants may use different strategies. However, the model comparison does suggest that any model that can predict human choices needs to contain a feature-based evaluation function, and mechanisms for attentional oversights and tree search.

## Generalization to 2AFC and Evaluation

Next, we show that the model can generalize by estimating parameters from subjects' choices in games against computers and predicting their choices in the 2AFC or evaluation tasks with minimal additional assumptions. To select an option on a 2 AFC trial, the only change we make is to initialize the tree search algorithm with a three-node decision tree with the current board position as the initial node and the two available candidate moves as children. On an evaluation trial, we execute the tree search algorithm as usual, then measure the value of the root node. We then convert this value to a seven-point scale by transforming $v \rightarrow 3+4 \tanh (v / 20)$ and rounding to the nearest integer.


Figure 3: A. Histogram of the percentage of correctly predicted 2AFC choices by our model across $N=40$ participants. We fit parameters for each participant on their choices in games against computers. The dashed line indicates the accuracy of a random prediction. B. Same for the evaluation task, where we quantify goodness-of-fit as the correlation across trials between rankings predicted by the model and reported by a participant.

The average accuracy of the model prediction on 2AFC data is $58.6 \pm 1.0 \%$ (figure 3A), the average correlation between predicted and observed evaluations is $\rho=0.38 \pm 0.04$ (figure 3B). The prediction is better than chance for $36 / 40$ participants in the 2AFC task, and 37/40 for evaluation.

Even though our model predicts participants' choices in these additional tasks well on average, the goodness-of-fit is highly variable across participants. This variability in goodness-of-fit is correlated across subjects between the three tasks (2AFC-evaluation: $\rho=0.54, p<0.001 ; 2 A F C$-games: $\rho=0.35, p<0.05$; evaluation-games: $\rho=0.24, p=0.12$ ). Moreover, on the 2AFC and evaluation task, goodness-offit correlates with participants' objective task performance (2AFC: $\rho=0.56, p<0.001$; evaluation: $\rho=0.96, p<$ 0.001 ). This suggests that the variability in goodness-of-fit can at least partially be explained by differences in intrinsic variability across participants.

## How experimental manipulations affect model parameters

To further support the model, we investigate whether its parameters respond in predictable ways to experimental manip-
ulations. As our first manipulation, we introduce time constraints of 5,10 or 20 seconds per move. Second, we conduct an experiment in which participants play the game for 5 sessions.

Given a set of parameters for an individual participant in a time limit condition or learning session, we simulate moves made by the model in a database of pre-determined positions and measure 3 statistics of its process: the percentage of dropped features, the value quality (correlation between $V(s)$ and the game-theoretic value $\left.V^{*}(s)\right)$ and the mean tree size (number of nodes in its decision tree). Note that tree size incorporates both the width and depth of the decision tree.

Based on the literature on expertise and time pressure in chess, we expected that time constraints would reduce tree size but not affect value function quality. In the learning experiment, we expected the value function quality to increase across sessions and the tree size to remain constant or increase only slightly. Since chess algorithms often do not explicitly include feature dropping or similar mechanisms, we made no predictions for its trajectory. Finally, we predict that experience increases participants' task performance while time pressure reduces it.
Time pressure To test the effectiveness of time constraints to manipulate participants' behavior, we first plot the distribution of response times in the three conditions, as well as the response times from the unconstrained (generalization) experiment (figure 4A). Adding time pressure causes an overall shift in the response time distribution regardless of the time limit. Additionally, participants play faster with shorter time constraints. Surprisingly, there is no consistent effect of time constraints on participants' performance (figure 4B).


Figure 4: A. Empirical cdf of response times in the three conditions of the time pressure experiment (red), and the generalization experiment (blue). In the latter experiment, players could take arbitrary amounts of time, which we denote as an infinite time limit. People play faster with shorter time limits. B. Task performance, quantified by Elo rating, for the same experiments and conditions. Error bars indicate mean and s.e.m. across participants $(N=30)$. The effect of time limits on performance is unclear.

In figure 5 (top), we show the feature drop rate, value function quality and tree size in different time limit conditions. Compared to the unconstrained experiment, participants build
smaller trees and drop more features, while the value function quality is similar. The impact of the time constraint on tree size becomes larger with shorter time limits, but the feature drop rate shows the opposite trend and is at its highest in the 20 -second condition. We speculate that the stress of potentially losing on time causes participants to pay more attention with shorter time limits, whereas with 20 seconds, they are more relaxed and make more attentional lapses.


Figure 5: Top row. Estimated model parameters in the time pressure and generalization experiments. Error bars denote mean and s.e.m. across participants. The model infers a relation between time limit and tree size, but unclear effects on feature dropping and the value function quality. Bottom row. Model parameters and Elo rating for each participant in each time limit condition. The tree size and feature drop rate correlate with Elo rating, but value function quality does not.

To understand the surprising negative result of figure 4, we investigate how Elo rating and parameter estimates correlate across both individuals and time limit conditions (figure 5, bottom). Stronger players (in all time limit conditions) are estimated to build larger decision trees and drop fewer features. Therefore, the increased tree size with longer time limit predicts a performance increase, but the increased feature drop rate predicts decreased performance. These opposite effects happen to be approximately equal, which explains the lack of correlation between time limit and Elo rating.
Learning We first validate that experience affects participants' behavior by plotting Elo rating as a function of session number (figure 6). Next, we investigate changes in parameters across sessions (figure 7, top). Tree size increases across sessions, feature drop rate decreases and value function quality remains constant. As in the time pressure experiment, tree size and feature drop rate correlate with Elo rating on an individual level (figure 7, bottom), and the change in parameter estimates across sessions explains changes in task performance. Experienced players build larger decision trees and drop fewer features, both of which predict increased playing strength, which matches the data.


Figure 6: Elo rating of $N=30$ participants in the learning experiment (mean and s.e.m. across participants). As participants gain expertise, they play stronger.


Figure 7: Top: Model parameters as a function of sessions completed in the learning experiment. Over the course of learning, tree size is estimated to increase while feature dropping decreases. The value function quality decreases, but only slightly. Bottom: Model parameters and Elo ratings for each participant in each session of the learning experiment. Both tree size and feature dropping correlate with Elo, but value function quality does not.

## Discussion

Limitations. Our model has three conceptual limitations. First, although its parameters shift as participants acquire expertise, the model does not describe how these shifts arise from their experience (their specific move choices and rewards). Instead, model parameters are stationary within each session. Moreover, because model parameters are constant while participants play against multiple AI opponents per session, the model cannot capture strategic adaptations based on an opponent's game play. Finally, the model assumes that people make decisions independently on every move, ignoring potential long-term planning or caching of partial game trees between moves. We make these assumptions out of necessity, because parameter inference is already challenging.
Relation with chess literature. Contrary to the chess literature, in which the superior pattern recognition of chess experts is evident from board reconstruction experi-
ments (Chase \& Simon, 1973) and eye movements (Reingold, Charness, Pomplun, \& Stampe, 2001), we find no changes in value function quality with expertise or individual skill differences. Stronger players might use features outside our model space, and the lack of correlation could be a false negative. Alternatively, perhaps chess and 4-in-a-row are qualitatively different domains of expertise. Chess contains many non-obvious features (pawn structure, the bishop pair) or non-obvious feature weights (bishops and knights are equally strong). By contrast, in our task, people's intuitive priors (three-in-a-row is good) happen to be correct.

Our finding of increased tree search with longer time controls is consistent with chess studies that conceptualize pattern recognition and tree search as fast and slow processes, respectively (Chabris \& Hearst, 2003). However, the strong dependence between expertise and tree search is unexpected. We first investigate whether this effect could have arisen from incorrect model assumptions. Specifically, players may use unmodeled features, stronger players may assign those features higher weights, and those feature weights may trade off with additional tree search in our model. However, by analyzing parameter estimates in lesion models, we find no such trade-offs. Therefore, our results reflect differences between 4-in-a-row and chess, or a methodological improvement. Conclusions about tree search in chess derive almost solely from verbal reports, whereas we use the more principled method of parameter inference in a behavioral model.

## Conclusion

We built a computational model that predicts people's choices in a two-player board game. The model posits three computational principles for sequential decision-making: a featurebased evaluation function, attentional oversights and tree search. All three components are necessary to explain participants' behavior, but the data does not constrain details of their implementation such as the order by which nodes are visited during search, or how long the search process continues before players finalize their decision.

The model generalizes to predict choices in a twoalternative forced-choice task and a board evaluation task. This suggests that the model doesn't just fit a mapping from boards to moves, but that it captures aspects of the computational process that underlies decision-making in all three tasks. Furthermore, the feature drop rate and tree size change in predictable ways when we expose participants to manipulations in time pressure and experience. These changes account for participants' task performance, suggesting that these specific parameters reflect some task-relevant characteristic of participants' cognitive process. Furthermore, these two behavioral characteristics are dissociable, since in the time pressure experiment, both tree size and feature dropping increase across conditions, whereas in the learning experiment, tree size increases while feature dropping decreases. In the future, we aim to further support our model as a description of the computational process underlying people's move choices
by using it to predict response times and eye movements.

## Acknowledgments

This work was supported by grant IIS-1344256 from the National Science Foundation

## References

Allis, L. V., et al. (1994). Searching for solutions in games and artificial intelligence. Ponsen \& Looijen.
Arad, A., \& Rubinstein, A. (2012). The 11-20 money request game: a level-k reasoning study. The American Economic Review, 102(7), 3561-3573.
Campitelli, G., \& Gobet, F. (2004). Adaptive expert decision making: Skilled chess players search more and deeper.
Chabris, C. F., \& Hearst, E. S. (2003). Visualization, pattern recognition, and forward search: Effects of playing speed and sight of the position on grandmaster chess errors. Cognitive Science, 27(4), 637-648.
Chase, W. G., \& Simon, H. A. (1973). Perception in chess. Cognitive psychology, 4(1), 55-81.
de Groot, A. D. (1946). Het denken van den schaker: een experimenteel-psychologische studie. Noord-Hollandsche Uitgevers Maatschappij.
de Groot, M. H. (1959). Unbiased sequential estimation for binomial populations. The Annals of Mathematical Statistics, 80-101.
Elo, A. E. (1978). The rating of chessplayers, past and present. Arco Pub.
Holding, D. H. (1985). The psychology of chess skill. Lawrence Erlbaum.
Hunter, D. R. (2004). Mm algorithms for generalized bradley-terry models. Annals of Statistics, 384-406.
Huyer, W., \& Neumaier, A. (1999). Global optimization by multilevel coordinate search. Journal of Global Optimization, 14(4), 331-355.
Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., \& Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol, 8(3), e1002410.
Johnson, A., \& Redish, A. D. (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. Journal of Neuroscience, 27(45), 1217612189.

Lee, S. W., Shimojo, S., \& ODoherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. Neuron, 81(3), 687-699.
Reingold, E. M., Charness, N., Pomplun, M., \& Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. Psychological Science, 12(1), 48-55.
Simon, D. A., \& Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. Journal of Neuroscience, 31(14), 5526-5539.
Solway, A., \& Botvinick, M. M. (2015). Evidence integration in model-based tree search. Proceedings of the National Academy of Sciences, 112(37), 11708-11713.

# Approximations of Predictive Entropy Correlate with Reading Times 

Marten van Schijndel (van-schijndel.1@osu.edu)<br>Department of Linguistics<br>The Ohio State University

William Schuler (schuler.77@osu.edu)
Department of Linguistics
The Ohio State University


#### Abstract

The lexical frequency of an upcoming word affects reading times even when the upcoming word is masked from readers (Angele et al., 2015). One explanation for this observation is that readers may slow down if there is high uncertainty about upcoming material. In line with this hypothesis, this study finds a positive correlation between predictive entropy and self-paced reading times. This study also demonstrates that such predictive entropy can be effectively approximated by the surprisal of upcoming observations and that this future surprisal estimate is more predictive of reading times when the grammar is more granular, which would be prohibitively expensive for predictive entropy. These results suggest readers engage in fine-grained predictive estimations of certainty about upcoming lexical and syntactic material, that such predictions influence reading times, and that estimating that uncertainty can be done less expensively and more robustly with information-theoretic surprisal. Keywords: Self-Paced Reading; Information Theory; Language Modeling; Corpus Studies


## Introduction

The lexical frequencies of upcoming words affects reading times even when the upcoming word is masked from readers (Angele et al., 2015). Angele et al. suggest that the driving factor behind their result may be anticipation of upcoming difficulty. For example, a less constraining context (i.e. less predictable upcoming words) may produce slower reading. This study uses informationtheoretic entropy to test their hypothesis and to investigate the level of linguistic detail predicted by readers.

This work is scientifically important because it uses a large self-paced reading corpus to show that reading times are influenced both by uncertainty over upcoming syntactic constructions and by uncertainty over upcoming lexical items, which supports the hypothesis of Angele et al. (2015) that anticipation of upcoming difficulty influences reading times. While previous work has found evidence of prediction during language processing through responses to violated predictions (Wicha, Moreno, \& Kutas, 2004; Van Berkum, Brown, Zwitserlood, Kooijman, \& Hagoort, 2005; Fine, Jaeger, Farmer, \& Qian, 2013; DeLong, Troyer, \& Kutas, 2014), the present work demonstrates that the influence of prediction can be reliably detected in reading times prior to any violation of that prediction. Other work, for example using a visual world paradigm (Altmann \& Kamide, 1999; Kamide, Altmann, \& Haywood, 2003; Ito \& Speer,
2008), has also demonstrated predictive processing absent a prediction violation, but the present work demonstrates that such an effect is also observable in a broadcoverage self-paced reading corpus such as can be collected via Mechanical Turk. Finally, Roark, Bachrach, Cardenas, and Pallier (2009) have previously shown that the entropy of upcoming syntactic categories influences self-paced reading times, but their entropy measure is extremely expensive to compute, they used a much smaller corpus, ${ }^{1}$ and they did not find an influence of upcoming lexical uncertainty on reading times, unlike the present work.

In addition, this work demonstrates that surprisal (Hale, 2001; Levy, 2008), typically only used to estimate responses to observed stimuli, can be used to quantify predictive influences as well. From a computational perspective, this work provides an inexpensive way to estimate the uncertainty experienced by readers, which will allow future studies to test the cognitive plausibility of various grammars and parsing algorithms, providing a tool with which to probe predictive human sentence processing outside of highly constraining experimental stimuli.

## Background

Angele et al. (2015) wanted to test whether lexical successor effects (influences of upcoming material) could be elicited even when readers were unable to view the upcoming words. They used a moving mask to hide upcoming words from readers but still found that the trigram predictability of the next hidden word was a significant predictor of reading times. Angele et al. (2015) hypothesized that readers may anticipate upcoming difficulty and slow down. That is, an unconstrained context with several plausible continuations might produce slower reading (due to each continuation's low predictability) than a highly constraining context with a smaller number of plausible continuations. To test this hypothesis, we use information-theoretic entropy to predict reading times.

Under information theory (Shannon, 1948), the entropy $(H)$ of a random variable $(X)$ is defined by the component probabilities of each possible value $(x)$ of that

[^218]variable:
\[

$$
\begin{equation*}
H(X)=-\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \tag{1}
\end{equation*}
$$

\]

In the case of language processing, the possible values are words that have yet to be observed, and entropy is typically computed from the conditional probability of each possible value given the observations that have already been made.

Linzen and Jaeger (2015) distinguished single-step predictive entropy (uncertainty about the next processing step) from full entropy (uncertainty about the rest of the sentence). Since Angele et al. (2015) found that lexical frequency successor effects were only dependent on the word following a fixation, the present work is concerned with single-step predictive entropy. Linzen and Jaeger (2015) found that when single-step predictive entropy was computed over upcoming syntactic constituents based on verb subcategorization biases, it was not predictive of self-paced reading times. However, they hypothesize that the fit of entropy may improve when computed over finer-grained categories (they only computed probabilities for 6 subcategorization classes). The results in Analysis 4 of this paper support their hypothesis.

Roark et al. (2009) defined two variants of single-step predictive entropy to distinguish syntactic uncertainty from lexical uncertainty. Syntactic entropy is computed over the conditional probability of each preterminal $(p)$ in the grammar $(G)$ given the previously observed lexical sequence $\left(w_{1 . . i-1}\right)$ :

$$
\begin{align*}
& \operatorname{Syn} H_{G}^{1}\left(w_{1 . . i-1}\right) \stackrel{\text { def }}{=} \\
& \quad-\sum_{p_{i} \in G} \mathrm{P}_{G}\left(p_{i} \mid w_{1 . . i-1}\right) \log \mathrm{P}_{G}\left(p_{i} \mid w_{1 . . i-1}\right) \tag{2}
\end{align*}
$$

Syntactic entropy is computed in practice by generating all possible syntactic derivations ${ }^{2}$ that can generate each possible upcoming word $\left(w_{i}\right)$ in the vocabulary $(V)$ and then subtracting from each derivation's probability the emission probability of generating $w_{i}$ from the chosen preterminal $\left(p_{i}\right)$.

Lexical entropy is computed over the conditional probability of each possible upcoming lexeme, given the previously observed lexical sequence:

$$
\begin{align*}
& \operatorname{Lex} H_{G}^{1}\left(w_{1 . . i-1}\right) \stackrel{\text { def }}{=} \\
& \quad-\sum_{w_{i} \in V} \mathrm{P}_{G}\left(w_{i} \mid w_{1 . . i-1}\right) \log \mathrm{P}_{G}\left(w_{i} \mid w_{1 . . i-1}\right) \tag{3}
\end{align*}
$$

Roark et al. (2009) found that syntactic entropy was predictive of self-paced reading times but that lexical entropy was not, which we were able to replicate on the corpus in this study as well. Roark et al. suggested that the

[^219]failure of lexical entropy to predict reading times may be due to the fact that their grammar was trained on the relatively small Brown portion of the Penn Treebank (Marcus, Santorini, \& Marcinkiewicz, 1993), so their lexical probabilities may not have been robust enough.

It is interesting to note that 'single-step prediction' was defined slightly differently for these two sets of authors. Roark et al. (2009) define it as a prediction over the next word in a lexical sequence, while Linzen and Jaeger (2015) define it as a prediction over the next syntactic category (e.g., noun phrase) that will branch from a partial derivation ending in a verb phrase. To avoid making a commitment as to the particular parsing strategy adopted by readers, this paper will use the definition of 'single-step prediction' from Roark et al. (2009) to mean uncertainty about the next lexical observation.

## Data

This study makes use of the Natural Stories self-paced reading corpus (Futrell et al., in prep). The corpus is a set of 10 texts ( 485 sentences) written to sound fluent but still containing many low-frequency and marked syntactic constructions. The sentences within each text were presented in order, and self-paced reading time data was collected from 181 native English speakers. Reading times were excluded if they occurred at the beginning or end of a sentence, or if they were less than 100 ms or greater than 3000 ms . Approximately one third of the sentences (255,554 events) were used for exploration and two thirds of the sentences ( 512,469 events) were used as a confirmatory partition for significance testing to reduce the risk of false positives due to multiple comparisons. All significance results reported in this paper are from the confirmatory partition.

## Models

This study fits reading times using linear mixed effects models computed with the lme4 (version 1.1-7) R package (Bates, Maechler, Bolker, \& Walker, 2014). All models include a baseline of fixed effect predictors for word length, sentence position, and 5 -gram surprisal. ${ }^{3}$ The models also include random intercepts for each word, each subject, and each subject/sentence pair. The last random intercept corrects for the fact that multiple nonindependent observations are drawn from each sentence. Finally, each model includes by-subject random slopes for all the fixed effects. All predictors were z-transformed prior to fitting. Significance values for each predictor were obtained using a likelihood ratio test between two

[^220]mixed models: one of which contained both a by-subject random slope and a fixed effect for the predictor of interest, and the other of which omitted the fixed effect for that predictor.

## Analyses

## Analysis 1: Single-Step Predictive Entropy

First, we test whether the original finding of Roark et al. (2009) that syntactic predictive entropy positively correlates with reading times holds up on the Natural Stories corpus (Futrell et al., in prep). We compute single-step predictive syntactic and lexical entropy using the Roark (2001) top-down incremental parser. Our findings are consistent with those of Roark et al. (2009): syntactic entropy has a significant positive effect on selfpaced reading times in the Natural Stories confirmatory partition over the baseline model $(\hat{\beta}=4.53, \hat{\sigma}=0.54$, p -value $<0.001$ ), and lexical entropy is not a significant predictor of reading times.

As Roark et al. (2009) point out, the lack of predictivity of lexical entropy may stem from the sparseness of the training data. Unfortunately, computing predictive entropy is very expensive since it requires predictively running the parser over a large set of hallucinated observations whose cardinality is the size of the vocabulary for for each actual observation. Therefore, meaningfully increasing the vocabulary is not generally practical. ${ }^{4}$

## Analysis 2: Surprisal as Entropy Approximation

Angele et al. (2015) found that the trigram surprisal of an upcoming word is predictive of reading times and speculated that such an effect could be driven by uncertainty over future events, so this section tests whether the predictive entropy effect observed in Analysis 1 can be approximated by the PCFG surprisal of the upcoming word.

Roark (2011) showed that single-step predictive lexical entropy is mathematically equivalent to the expected value of total surprisal $S$ :

$$
\begin{align*}
& \quad S_{G}\left(w_{i}, w_{1 . . i-1}\right) \stackrel{\text { def }}{=}-\log \mathrm{P}_{G}\left(w_{i} \mid w_{1 . . i-1}\right)  \tag{4}\\
& \operatorname{Lex} H_{G}^{1}\left(w_{1 . . i-1}\right) \\
& \stackrel{\text { def }}{=} \sum_{w_{i} \in V}-\mathrm{P}_{G}\left(w_{i} \mid w_{1 . . i-1}\right) \log \mathrm{P}_{G}\left(w_{i} \mid w_{1 . . i-1}\right)  \tag{5}\\
& =\sum_{w_{i} \in V} \mathrm{P}_{G}\left(w_{i} \mid w_{1 . . i-1}\right) S_{G}\left(w_{i}, w_{1 . . i-1}\right)  \tag{6}\\
& =\mathrm{E}\left[S_{G}\left(w_{i}, w_{1 . . i-1}\right)\right] \tag{7}
\end{align*}
$$

[^221]where $w_{i}$ is the current lexical item, $w_{1 . . i-1}$ is the sequence of previously observed lexical items and $V$ is the vocabulary of the language.

Therefore, surprisal is a single sample from the conditional probability distribution over which single-step lexical entropy is computed, where the sampled observation is the occurrence that ultimately is observed. Over several trials, then, future surprisal should approximate entropy since each observed occurrence should happen proportionately to its expected occurrence frequency. As a moving window self-paced reading corpus, participants were physically unable to see upcoming words, similar to the masked condition used by Angele et al. (2015).

To test surprisal as an approximation of entropy, we use the Roark (2001) parser's estimate of surprisal of each observation to predict the reading time of the preceding observation. This measure (future surprisal) also has a significant positive effect on reading times ( $\hat{\beta}=4.96, \hat{\sigma}=0.63, \mathrm{p}$-value $<0.001$ ). This measure may be thought of as an aggregate approximation to entropy, whereas the lexical entropy output by the Roark (2001) parser may be thought of as a point-wise approximation to entropy. That is, Roark lexical entropy approximates the true lexical entropy for each new observation as the weighted average of the conditional probability distribution at that point according to the parser's grammar, while future surprisal approximates the true lexical entropy over the entire corpus (aggregated over all observations) by sampling from the conditional probability distribution for each observation. The fact that future surprisal is able to fit reading times more consistently than point-wise lexical entropy gives hope that this less expensive aggregate approximation of entropy is a more robust means of computing entropy than a point-wise approximation.

## Analysis 3: $N$-grams as Better Entropy Approximation

Since the Roark (2001) parser computes surprisal based on a relatively small and coarse-grained Penn Treebank grammar, the previous results may be skewed by the small amount of training data. In order to obtain conditional probabilities based on more data, we use a 5 gram back-off model computed with the KenLM toolkit (Heafield, Pouzyrevsky, Clark, \& Koehn, 2013) on the Gigaword 4.0 corpus (Graff \& Cieri, 2003), which consists of 2.96 billion words from English newswire text. Again, the 5-gram surprisal of each word was used to predict the reading time of the preceding word. Similar to future Roark surprisal, future 5 -gram surprisal has a significant positive correlation to reading times $(\hat{\beta}=4.49$, $\hat{\sigma}=0.57$, p-value $<0.001$ ), and when future 5 -gram surprisal is in the model, future Roark surprisal ceases to be a significant predictor of reading times.

This result aligns with work by van Schijndel and Schuler (2016) who found that future PCFG surprisal,
computed with a Penn Treebank PCFG, is an effective predictor of reading times in eye-tracking, but that it ceased to be predictive when future $n$-gram surprisal was included in their model. They also found that future $n$ gram surprisal was only predictive for one or two words following a fixation, similar to the finding of Angele et al. (2015) that only the frequency of the word following a fixation was predictive of reading times.

## Analysis 4: Fine-Grained Syntactic Prediction

Although future $n$-gram surprisal seems to account for a lexical entropy effect, it is unable to account theoretically for the effect of Roark syntactic entropy, since $n$-gram surprisal reflects lexical probabilities and syntactic entropy reflects syntactic probabilities (without lexical emission probabilities). However, future Roark PCFG surprisal using the default set of Penn Treebank syntactic categories was unable to predict reading times when future $n$-gram surprisal was in the model. Previous work on predictive processing has suggested that predictions can be relatively fine-grained (Luke \& Christiansen, 2015; Kim \& Lai, 2012), so this section explores whether humans predict upcoming material with finegrained syntactic specificity.

Whereas the above experiments used the Roark (2001) parser with the default Penn Treebank tag set, this section uses the van Schijndel, Exley, and Schuler (2013) parser, which computes surprisal using the Petrov, Barrett, Thibaux, and Klein (2006) latent-variable grammar computed from sections 2-21 of the Wall Street Journal portion of the Penn Treebank and thereby achieves higher parsing accuracy than the Roark parser (van Schijndel et al., 2013). The latent-variable grammar is derived from a split-merge algorithm that creates finegrained subcategory tags from the basic Penn Treebank category tags. For this experiment, the grammar underwent 5 split-merge operations to obtain optimally tuned tags, following the recommendations of Petrov et al.

When future surprisal is computed with a finergrained tag set, it is able to obtain a significant positive correlation with reading times, even in the presence of future 5 -gram surprisal and syntactic entropy ( $\hat{\beta}=4.10$, $\hat{\sigma}=0.74$, p-value $<0.001$ ).

## Discussion

Much previous psycholinguistic and neurolinguistic work has shown that prediction plays a role in language processing (DeLong et al., 2014; Kuperberg \& Jaeger, 2015). Angele et al. (2015) observed that even when upcoming material is masked, its predictability can affect reading times. They suggest that their observation is likely driven by readers predicting difficult material and slowing in anticipation of it. The findings in this paper of a positive correlation between self-paced reading times

|  | $\hat{\beta}$ | $\hat{\sigma}$ | t |
| :--- | ---: | ---: | ---: |
| Syntactic Entropy | 4.53 | 0.54 | 8.36 |
| Future Roark Surprisal | 4.96 | 0.63 | 7.85 |
| Future 5-gram Surprisal | 4.49 | 0.57 | 7.89 |
| Future Fine PCFG Surprisal | 4.10 | 0.74 | 5.58 |

Table 1: Effect sizes for each predictor of interest over the baseline described in the Models section. Each predictor was tested over the baseline factors and all predictors listed above it in the table. Future Roark Surprisal is not significant once Future 5 -gram surprisal is added.
and predictive entropy are consistent with that hypothesis and suggest that, in particular, readers slow due to increased probabilistic uncertainty over upcoming material.

Previous studies have claimed that a positive correlation between entropy and reading times would indicate that there is a competition cost between multiple parse hypotheses (Linzen \& Jaeger, 2015), but this is not the only possible explanation for such a correlation. For example, similar reasoning to the Uniform Information Density hypothesis (UID; Jaeger, 2010) might apply to readers. That is, if readers have more uncertainty about upcoming material, they may anticipatorily slow their reading in order to better process the less expected information (reducing their expected per-millisecond surprise to channel capacity). If, instead, readers are reasonably confident about what words they are about to encounter, they may speed up in order to maximize the per-millisecond informativity of their observations. This sort of tuning may be exaggerated in the moving window self-paced reading paradigm, where readers will be unable to regress if they speed past an unexpected observation, which could be why previous work using eye tracking has only been able to find an effect of future $n$ gram surprisal on reading times (Angele et al., 2015; van Schijndel \& Schuler, 2016), while the present self-paced reading study also found an effect for future PCFG surprisal.

The fact that both future 5 -grams and future PCFG surprisal are predictive of reading times suggests that predictions of upcoming difficulty are being made both about lexical items and syntactic constructions. Surprisal is computationally much less expensive than entropy, and therefore it can provide samples from a much finer-grained conditional probability distribution over possible analyses than would be practical for entropy calculation.

The present results show that future latent-variable PCFG surprisal can fit reading times even when the coarser Roark et al. (2009) surprisal and lexical entropy cannot, which suggests that humans predict upcoming material at a relatively fine-grained level (both syntactic and lexical) as suggested by previous work (Luke
\& Christiansen, 2015; Kim \& Lai, 2012). These results further indicate that the fit of entropy to reading times improves as the granularity of the grammar becomes finer, which supports the hypothesis of Linzen and Jaeger (2015) that their subcategorization entropy was likely too coarse-grained to reveal entropy's influence.

The finding that Roark syntactic entropy retains its reading time predictivity in the presence of future 5gram surprisal and future latent-variable surprisal suggests that humans estimate certainty about upcoming parses based on multiple samples from the distribution over upcoming observations. Such a finding is consistent with parallel models of sentence processing but may be problematic for serial processing models. Another interpretation of this finding is that a point-wise entropy approximation is more stable and so can serve as a back-off for the less stable but more nuanced aggregate approximations provided by both the $n$-gram and latent-variable surprisal models. It is left to future work to differentiate between these two possibilities.

It may seem strange that total latent-variable surprisal was used in this study instead of syntactic latent-variable surprisal (without lexical probabilities) since the goal of moving beyond future $n$-gram surprisal was to capture something of syntactic entropy, which omits lexical emission probabilities; however, explorations on the development partition revealed that total surprisal generally provides better fits to reading times than syntactic surprisal even in the presence of future 5 -gram surprisal. In any case, the goal was not necessarily to approximate Roark syntactic entropy but to capture an aspect of the uncertainty experienced by readers, of which Roark lexical entropy and Roark syntactic entropy are themselves approximations. In fact, the consistent correlation between future surprisal (both $n$-gram and latent-variable) and reading times compared to Roark lexical entropy suggests that fine-grained aggregate entropy approximation via future surprisal is more robust than the coarser but more intuitive point-wise lexical entropy approximation output by the Roark (2001) parser.

The entropy findings in this paper are distinct from those in the entropy reduction literature. The Entropy Reduction Hypothesis states that readers slow according to the informativity of the words they encounter (as measured by a decrease in entropy; Hale, 2006). It is possible that the two effects are independent and that people slow down before areas of greater uncertainty, while also slowing down due to larger information gains. These effects are not necessarily mutually exclusive because entropy reduction deals with changes in entropy while predictive entropy deals with the overall level of uncertainty in a text. That is, an entropy reduction of $k$ may predict the same $k \cdot \beta_{\Delta H} \mathrm{~ms}$ effect on reading times whether the resulting entropy is low or high. In contrast, the experiments in this paper highlight a broad-
coverage correlation of fine-grained predictive entropy to self-paced reading times.

## Conclusion

This paper has replicated previous findings that singlestep predictive entropy is positively correlated with selfpaced reading times and presented new results that show this correlation can be inexpensively approximated using both future $n$-gram surprisal and future latent-variable PCFG surprisal. The present results also demonstrate that such approximations improve as the granularity of the approximation increases. By showing that greater uncertainty over upcoming words and syntactic constructions slows reading times, these results support the hypothesis of Angele et al. (2015) that anticipation of upcoming difficulty affects reading.

## Acknowledgements

Thanks to the computational linguistics and cognitive modeling discussion groups at OSU (Clippers and CaCL) for helpful feedback on this work. This work was supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012.

## References

Altmann, G. T. M., \& Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73 (3), 247-264.
Angele, B., Schotter, E. R., Slattery, T. J., Tenenbaum, T. L., Bicknell, K., \& Rayner, K. (2015). Do successor effects in reading reflect lexical parafoveal processing? evidence from corpus-based and experimental eye movement data. Journal of Memory and Language, 79-80, 76-96.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4 [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 1.1-7)
DeLong, K. A., Troyer, M., \& Kutas, M. (2014). Preprocessing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. Language and Linguistics Compass, 8(12), 631-645.
Fine, A. B., Jaeger, T. F., Farmer, T. A., \& Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. PloS ONE, 8(10), 1-18.
Futrell, R., Gibson, E., Tily, H., Vishnevetsky, A., Piantadosi, S., \& Fedorenko, E. (in prep). Natural stories corpus.
Graff, D., \& Cieri, C. (2003). English Gigaword LDC2003T05 [Computer software manual]. Linguistic Data Consortium.
Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In Proceedings of the second meeting of the north american chapter of the as-
sociation for computational linguistics (pp. 159166). Pittsburgh, PA.

Hale, J. (2006). Uncertainty about the rest of the sentence. Cognitive Science, 30(4), 609-642.
Heafield, K., Pouzyrevsky, I., Clark, J. H., \& Koehn, P. (2013, August). Scalable modified Kneser-Ney language model estimation. In Proceedings of the $51 s t$ annual meeting of the association for computational linguistics (pp. 690-696). Sofia, Bulgaria.
Ito, K., \& Speer, S. R. (2008). Anticipatory effect of intonation: Eye movements during instructed visual search. Journal of Memory and Language, 58, 541-573.
Jaeger, T. F. (2010, August). Redundancy and reduction: Speakers manage information density. Cognitive Psychology, 61(1), 23-62. Retrieved from http://dx.doi.org/10.1016/j .cogpsych.2010.02.002
Kamide, Y., Altmann, G. T. M., \& Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language, $49(1), 133-156$.
Kim, A., \& Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from erps. Journal of Cognitive Neuroscience, 24(5), 11041112.

Kuperberg, G. R., \& Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? Language, Cognition and Neuroscience, 1-29.
Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, $106(3), 1126-1177$.
Linzen, T., \& Jaeger, T. F. (2015). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. Cognitive Science, 1-30.
Luke, S. G., \& Christiansen, K. (2015). Predicting inflectional morphology from context. Language, Cognition and Neuroscience, 1-14.
Marcus, M. P., Santorini, B., \& Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, $19(2), 313-330$.
Petrov, S., Barrett, L., Thibaux, R., \& Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In Proceedings of the 44 th annual meeting of the association for computational linguistics (COLING/ACL'06).
Roark, B. (2001). Probabilistic top-down parsing and language modeling. Computational Linguistics, 27(2), 249-276.
Roark, B. (2011). Expected surprisal and entropy (Tech. Rep. No. CSLU-11-004). Portland, OR: Center for Spoken Language Processing, Oregon Health and

Science University.
Roark, B., Bachrach, A., Cardenas, C., \& Pallier, C. (2009). Deriving lexical and syntactic expectationbased measures for psycholinguistic modeling via incremental top-down parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Langauge Processing, 324-333.
Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379-423, 623-656.
Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., \& Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31 (3), 443.
van Schijndel, M., Exley, A., \& Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. Topics in Cognitive Science, 5(3), 522-540.
van Schijndel, M., \& Schuler, W. (2016). Addressing surprisal deficiencies in reading time models. In Proceedings of the computational linguistics for linguistic complexity workshop. Association for Computational Linguistics.
Wicha, N. Y. Y., Moreno, E. M., \& Kutas, M. (2004). Anticipating words and their gender: an eventrelated brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. Journal of Cognitive Neuroscience, 16(7), 1272-1288.

# The Development of Structural Thinking about Social Categories 

Nadya Vasilyeva (vasilyeva@berkeley.edu), Alison Gopnik (gopnik@berkeley.edu), Tania Lombrozo (lombrozo@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720 USA


#### Abstract

Representations of social categories help us make sense of the social world, supporting predictions and explanations about groups and individuals. Here we explore whether children and adults are able to understanding category-property associations in structural terms, locating an object of explanation within a larger structure and identifying structural constraints that act on elements of the structure. We show that children as young 3-4 years of age show signs of structural thinking, but that this capacity does not fully develop until after 7 years of age. These findings introduce a viable alternative to internalist accounts of social categories, such as psychological essentialism.


Keywords: structural explanation, structural factors, social categories, essentialism, category representation

Imagine that a school introduces a dress code stating that all items of a student's clothing must match in color. When school begins, most boys show up wearing blue, and most girls show up wearing pink. What explains the correlation between gender and color? One explanation is that boys naturally prefer blue, and girls pink. But a quick glance at history shows that in the $19^{\text {th }}$ century, pink was considered the vigorous, masculine color, whereas girls wore "delicate and dainty" blue (Fausto-Sterling, 2012). If an explanation that appeals to intrinsic preferences is inadequate, an alternative might be to appeal to a structural feature of the environment: department stores reliably stock more pink options for girls than for boys. In this case, availability could be a sufficient explanation for the observed correlation.
This example illustrates what we call "structural thinking." A hallmark of structural thinking is locating an object of explanation within a larger structure and identifying structural constraints that act on components of the structure to shape the distribution of outcomes for each component. In our example, girls occupy a position within larger social and institutional structures that make them more likely than boys to choose pink over blue. A structural approach to social categories differs from internalist approaches, which focus on essential or inherent properties of the category itself. In the current paper, we ask whether and when children develop the ability to think about social categories in structural terms.
Internalist approaches to category representation. One prominent approach to theorizing about the representation of social categories (such as "girl") is based on the notion of psychological essentialism, which refers to the tendency to represent (some) categories in terms of an underlying essence that is constitutive of category membership and/or causally responsible for key category features (Gelman, 2003). Psychological essentialism can support efficient generalizations about natural kinds, but can also lead to unwarranted normative expectations about categories, stereotypical generalizations, and prejudice (Leslie, 2015).

A related internalist approach comes from Cimpian and Salomon (2014), who argue for the inherence heuristic, defined as the tendency to explain observed patterns in terms of the inherent properties of the objects that instantiate them. If girls wear pink, people might infer that it must be due to something inherent about pink ("it is delicate") and/or girls ("they are attracted to delicate colors"), rather than considering a broader range of external, historical factors. Cimpian and Salomon argue that the inherence heuristic is distinct from, but potentially a precursor to, essentialized representations of social categories.
A final approach, the aspect hypothesis, comes from Prasada and Dilllingham (2006, 2009), who offer a nonessentialist account of categorical representation. On this view, some features of a category are viewed as aspects of the kind. For example, "fighting crime" is an aspect of being a police officer (in contrast to merely statistical associations, such as between police officers and "eating donuts").
While psychological essentialism, the inherence heuristic, and the aspect hypothesis are importantly distinct in their commitments regarding categorical representations, they all support internalist explanations for associations between a category and a feature (e.g., "she chose pink because girls like warm colors"), as well as formal explanations that appeal to category membership (e.g., "she chose pink because she is a girl"). By contrast, they lack mechanisms for differentiating kinds (i.e., "girls") from the structures in which they are embedded (i.e., the social position occupied by girls). As a result, they cannot readily accommodate the kind of structural thinking supported by a structural approach.
A structural approach to category representation. Our study explores an alternative to internalist accounts. According to a structural view of categorical representation, reliable connections between properties and categories can be represented as a consequence of stable structural constraints acting on categories from the outside.
This approach is based on the notion of structural explanation developed in philosophy of social sciences, where it is defined by situating the object of explanation in a network of relationships within a larger, organized whole (a structure), and identifying how relationships to other parts of the whole modify the probability distribution over possible states of the part whose behavior is explained (relative to a hypothetical case outside a structure, relative to other nodes within the structure, or relative to different structures; Haslanger, 2015). For example, an internalist explanation for why many women in heterosexual relationships leave their jobs after having a child might appeal to women's priorities or abilities, whereas a structural explanation would identify constraints that affect women in virtue of their position within the social structure (e.g., lack of paid parental leave, a
gender wage gap, etc.). These structural constraints shift the probability distribution across different outcomes for women versus men. For another position subjected to different structural constraints (e.g., "men," "women in a different culture"), the same event (having a child) need not trigger the same outcomes. Rather than pinpoint triggering causes (e.g., the baby's arrival), structural explanations identify constraints that shape the causal relationships between triggering causes and their effects (Dretske, 1988).
The structural view capitalizes on the distinction between nodes (positions within social structures) and node-occupiers (categories that occupy those positions, and come to possess particular properties by virtue of their location within the structure; Haslanger, 2015). This distinction brings to light a potential ambiguity in formal explanations (e.g. "Smith quit her job after having a baby because she is a woman"), given that the term "woman" can refer to either the node or the node-occupier. Such explanations could attribute properties directly to the node (i.e., women's location in a structure), without necessarily tying them to its inherent nature (i.e., to women themselves). In other words, a formal explanation could support both structural and internalist interpretations, a prediction that our experiments test.
Structural vs. other externalist approaches. One way to appreciate what constitutes a structural explanation is to consider what it is not. Structural explanations are not merely "situation" explanations from the traditional person-situation dichotomy (Ross \& Nisbett, 2011), nor "causal history of reasons" explanations from Malle's (2004) taxonomy, which are narrower in their restriction to intentional behavior, yet broader in allowing for non-structural antecedents to reasons. Structural explanations are a sub-type of externalist explanations that invoke stable constraints acting on a category in virtue of its position in a structure.
It's useful to think of structural explanations in terms of the ANOVA or "cube model" (Kelley, 1973), in which a behavior is attributed to co-varying factors (person, situation, or stimulus). However, the cube model assumes that the data (behaviors) come from an "unconfounded" factorial design, where person and external factors vary independently. Structural thinking is instead sensitive to confounds between people and situations; within a social structure, categories are often constrained by their nodes. The category "women" can only occupy the "women" node, which constrains the range of properties the occupier can display.
The notion of a confound between a category and its social location also helps to position the structural view of categories relative to role-based categories, such as guest, which specify a role in a relational structure (Asmuth \& Gentner, 2016; Markman \& Stilwell, 2001). Role-based categories involve relational structure, but structural thinking about social categories critically applies to cases in which a relational position is confounded with membership in a (perceived) taxonomic category.
Cross-cultural research on independent vs. interdependent (object vs. field) construals (Nisbett, 2003) suggests that the reasoning style associated with structural thinking is not as "unnatural" as it may seem. Research on analogy (Gentner,

1983; 2005) and recent work on role-based concepts (Goldwater, Bainbridge, \& Murphy, 2016) offer additional indications that people have the representational capacities to reason about structures. If people possess the requisite resources for engaging in structural reasoning, the question is: do they? And if so, when does this capacity develop? These are the questions our study addresses.
The development of structural thinking. Our study evaluates two competing hypotheses. Hypothesis 1 is that young children lack the conceptual prerequisites and/or knowledge to engage in structural thinking. Hypothesis 2 is that young children can successfully engage in structural thinking from an early age.
Each hypothesis receives some support from existing research. In favor of the first hypothesis, prior work demonstrates that children view some social categories (such as gender) as essentialized natural kinds from an early age (Rhodes \& Gelman, 2009; Taylor, 1996), even when cultural input suggests otherwise (Astuti et al., 2004). There is also evidence that young children have trouble endorsing environmental mechanisms that could produce category features (Rhodes \& Taylor, 2009), although the "environmental factors" that were examined were primarily non-structural in nature. Finally, as young as 4-5 years of age, children tend to generate and endorse "inherent" explanations of categorical patterns over "extrinsic" ones (Cimpian \& Markman, 2011; Cimpian \& Steinberg, 2014).
Beyond evidence of early essentialist and inherence-based reasoning, there is evidence that children lack capacities involved in structural thinking. Structural explanation could rely on structure-wide counterfactual alternatives, which do not fully emerge until age 7-8 (Beck et al., 2006; Rafetseder, Cristi-Vargas, \& Perner, 2010). Structural reasoning also relies on representing relations, and research on relational reasoning suggests a developmental shift in relevant capacities throughout and beyond the preschool years (e.g., Gentner, 1988; Richland, Morrison, \& Holyoak, 2006).
On the other hand, there is evidence that potentially favors Hypothesis 2. Several findings suggest that young children appreciate external constraints on social categories. Seiver, Gopnik, and Goodman (2013) demonstrated that children as young as 4 can use situational information in explanation and prediction when appropriate covariation evidence is available. Four-year-olds also recognize moral constraints on their own behavior (Chernyak \& Kushnir, 2014) and acknowledge that the behavior of members of a social category can be driven by common norms (Kalish, 2011; see also Kalish \& Shiverick, 2004; Rakoczy, Warneken, \& Tomasello, 2008; Smetana, 1981; Turiel, 1983). ${ }^{1}$

[^222]A final and more intriguing possibility is that young children may be more open to structural reasoning than older children and adults. Rhodes and Gelman (2009) showed that young children are more flexible than older children about some social categories, such as race. In Seiver et al. (2013), it was older children, not younger children, who showed an overall bias for person over situation explanations. Moreover, young children may be less biased by prior assumptions than adults, and thus open to learning a broad range of causal relationships (Lucas, Bridgers, Griffiths, \& Gopnik, 2013). This body of work suggests that relative to older children and adults, young children could be more open to integrating external constraints in their representations of social categories and relying on structural relations in reasoning.

## Experiment

This study had three goals: to determine whether and when children can successfully engage in structural thinking in explaining the association between a category and a property; to determine whether a structural construal can be experimentally induced; and to evaluate the prediction that structural thinking can support formal explanations under a structural interpretation of the category.

To address these three goals, we adopted an approach mirroring Prasada and Dillingham (2006, 2009), who developed a set of tasks that can be used to identify whether people construe the connection between a feature and a category as principled (such as between "fighting crime" and being a police officer) or statistical (such as between "eating donuts" and being a police officer). They showed that only the principled connections between kinds and features supported judgments of feature immutability (a person who does not fight crime is not really a police officer), partial definitions (a police officer is a person who fights crime), and formal explanations ("this person fights crime because she is a police officer"). With the aim of detecting structural thinking and differentiating it from internalist thinking, we modified these three measures (described below). Vasilyeva and Lombrozo (in prep) found that with adults, responses across these judgments can successfully be used as a "profiling tool" to detect structural thinking, which generates a unique signature: relatively high mutability ratings, low partial definition ratings, and high formal explanation ratings. In contrast, the pattern for an internalist construal should be low - high - high. To further validate the profiling tool, we additionally included an open-ended explanation prompt and close-ended causal explanation evaluations.

## Method

Participants We recruited 41 3-4-year-olds (mean age 4.3 years, range 3.0-4.9; 23 females, 18 males) and 48 5-6-yearolds (mean age 5.6 years, range 5.0-6.9; 23 females, 25 males). Additionally, 67 adults (mean age 33 years, range 1971; 33 females, 64 males) were recruited via Amazon Mechanical Turk; participation was restricted to users with an IP address within the US and an approval rating of at least $95 \%$ based on at least 50 previous tasks. Children were tested in person using an illustrated storybook presented on a laptop; adults were tested online.

Materials, Design, and Procedure Participants were first introduced to a school where girls and boys study in separate classrooms, and presented with fictitious data about students playing different games during recess: girls predominantly played Yellow-Ball while boys predominantly played GreenBall. Participants were told that the game each child played was determined by tossing a pebble towards two buckets standing side by side: if the pebble fell into the yellow bucket, the child played Yellow-Ball that day, and if the pebble fell into the green bucket, that child played GreenBall that day (Figure 1a). The critical manipulation concerned the sizes of the buckets. In the internalist condition, both buckets were of the same size (Figure 1b), inviting participants to infer that the correlation between category membership and game choice was the product of inherent preferences (see Kushnir, Xu, \& Wellman, 2010, for evidence that even younger children can use statistical evidence to infer a preference). In the structural condition, one bucket was instead much larger than the other: in the girls' classroom the yellow bucket was larger than the green bucket, with the reverse in the boys' classroom (Figure 1c). The size difference imposed a stable structural constraint on the probability distribution over options available to members of each category, inviting a structural interpretation of the category-property connection.

After comprehension checks, all participants completed a series of measures designed to differentiate an internalist from a structural construal of the property. First, in the openended explanation task, participants were asked why girls in the girls' classroom play Yellow-Ball a lot at their school. Second, participants completed a causal explanation evaluation task and the three profiling tools measures: a mutability judgment, a partial definition, and formal explanation ratings.

In the causal explanation evaluation task, children evaluated three kinds of causal explanations offered by puppets that "sometimes say things that are smart, and sometimes say things that are silly." The puppets explained that girls tend to play Yellow-Ball "because girls like playing Yellow-Ball" (internalist); "because in the girls' classroom, it's easier to throw a pebble in the yellow bucket" (structural); or "because they got sprinkled with water" (an incidental explanation invoking an irrelevant fact from the cover story). Participants evaluated each explanation using a two-step, four-


Figure 1: Illustrations of the procedure determining which game each student played in the story (a) and of the different constraints on the probability of outcomes in the internalist (b) and structural conditions (c).
point thumb scale: they first chose one of two thumbs representing "good explanation" (up) and "bad explanation" (down), and they then chose between two subsequent options based on their choice: "kind of good/bad" (small thumb) or "really good/bad" (big thumb) - a scale previously shown to work well to measure children's agreement with explanations (Cimpian \& Steinberg, 2014; Hussak \& Cimpian, 2015).

For the mutability judgment, participants were told that after a change in the school's rules allowing children to attend any classroom, Suzy's parents transferred her to the boys' classroom. Participants were asked to guess which game Suzy would play after transferring, responding on a two-step, four-point scale ranging from "for sure YellowBall" to "for sure Green-Ball." This mutability judgment mirrors more familiar "switched at birth" tasks in the essentialism literature (Gelman \& Wellman, 1991), in which children are asked, e.g., whether a cow raised by pigs will moo or go oink. Similarly, our mutability judgment involves a change in environment (structural constraints), and participants are asked to infer whether a property will match the exemplar's category (the node occupier) or the new environment (the node). A shift in predictions from YellowBall to Green-Ball should track the causal influence of the node, and indicate structural thinking (as well as show that structural positions are seen as influencing behavior, rather than merely reflecting existing internal preferences).

For the partial definition task, participants rated whether an alien did a good job telling what a girl is to another alien who had never heard about girls: "A girl is a person who plays Yellow-Ball a lot." Participants used a two-step, four-point scale ("really bad job" - "really good job").

In the formal explanation task, participants met Suzy who "plays Yellow-Ball a lot at her school" and were asked to evaluate a formal explanation offered by a puppet - "Because Suzy is a girl" - using the two-step, four-point thumb scale ranging from "really bad" to "really good."

## Results and Discussion

Due to differing test formats and sample sizes, data from children and adults were analyzed separately. For the openended explanation task (see Figure 2), participants' explanations were coded as internalist ("maybe the girls just like it better, so they always aim to get their pebbles into the yellow ball bucket"), structural ("because the pebble went into the yellow bin, because the yellow one is bigger"), or miscellaneous, comprised of "I don't know," question restatements, and unclassifiable responses ("the yellow ball is brighter"). The distribution of explanations was affected by condition for each age group (3-4-year-olds: $\chi^{2}(N=41)=6.19$, $p=.045$; 5-6-year-olds: $\chi^{2}(N=48)=16.80, \quad p<.001$; adults: $\left.\chi^{2}(N=67)=42.86, \quad p<.001\right)$. Critically, in the structural condition some proportion of participants in each age group produced structural explanations (Figure 2, right panel, black bars). There was also evidence of developmental change: in the structural condition, the percentage of internalist explanations dropped as the percentage of structural explanations increased, so that the overall preference for internalist explanations in the younger age group flipped to a preference for structural explanations for older children.


Figure 2: Distribution of internalist and structural explanations generated in response to question about why girls play Yellow-Ball, as a function of condition and age group.

Children's evaluations of causal explanations (see Figure 3) were analyzed as a function of explanation type (internalist, structural, incidental), condition (internalist, structural), and age group (3-4, 5-6 year-olds) in a mixed ANOVA, with the key prediction concerning an interaction between explanation type and condition. The analysis revealed a main effect of explanation type, $F(2,170)=9.87, p<.001, \eta_{p}{ }^{2}=.104$, which was qualified by a significant interaction between explanation type and condition, $F(2,170)=6.00, p=.003, \eta_{p}^{2}=.066$ : only the structural explanation ratings were boosted by the structural framing. Most importantly, we observed the target three-way interaction: $F(2,170)=3.73, p=.026, \quad \eta_{p}{ }^{2}=.042$, driven by the selective effect of condition on 5-6-year-olds' evaluations of the structural explanation: older children, but not younger children, rated structural explanations higher in the structural condition than in the internalist condition ( $p_{\text {older }}<.001, p_{\text {younger }}=.390$ ). The interaction remained significant when restricting the analysis to internalist and structural explanations, $p=.012$. For adults, an explanation type (essentialist, structural, incidental) by condition (essentialist, structural) mixed ANOVA revealed a significant effect of explanation type, $F(2,126)=171.15, p<.001, \eta_{p}^{2}=.731$, and a marginal effect of condition, $F(1,63)=3.74, p=.058, \eta_{p}{ }^{2}=.056$, qualified by a significant interaction, $F(2,126)=117.83$, $p<.001, \eta_{p}{ }^{2}=.652$ : adults favored the internalist explanation over the structural in the internalist condition, with the reverse in the structural condition ( $p$ ' $s<.001$, see Figure 3).

Having succeeded in finding evidence of structural thinking in our open- and close-ended causal explanation tasks, we next turn to the profiling tool measures to see whether they reveal developmental differences mirroring these patterns. For adults, who exhibited high levels of structural thinking, we would predict the following for the structural condition relative to the internalist condition: more frequent predictions


Figure 3: Explanation evaluation as a function of explanation type, framing condition, and age group.


Figure 4: Mutability (a), partial definition (b), and formal explanation ratings (c) as a function of framing condition and age group.
that Suzy will play Green-Ball when switched to the boys' classroom, lower endorsement of the partial definition, and no difference in endorsement of the formal explanation. For children, we would predict the same patterns, with smaller effects for the younger children. This is what we found.

For the mutability judgment task (see Figure 4a), the predicted main effect of condition was marginal for the youngest group, $t(39)=1.96, p=.057, d=.42$, but significant for the older children, $t(46)=2.29, p=.027, d=.63$, and for adults, $t(65)=8.04, p<.001, d=2.00$. The age by condition interaction for children was not significant, $F(1.85)<.01$, $\mathrm{p}=.984$, but it appears that the property (playing Yellow-Ball for girls) was seen as more mutable in the structural condition than in the essentialist condition by age 5-6.

For the partial definition task (Figure 4b), we predicted that properties construed as internalist should support definitions better than properties construed as structural. Neither younger nor older children displayed such a pattern ( $p$ ' $\mathrm{s} \geq .687$ ), but adults did, $t(65)=2.11, p=.039, d=.52$.

Finally, as predicted, formal explanation ratings did not significantly differ across the essentialist and structural conditions for any age group, all $p$ 's $\geq .915$ (see Figure 4c), suggesting that these explanations support both internalist and structural construals.

These results show that even young children are capable of structural thinking, as reflected in their openended explanations. They also provide the first demonstration that across all age groups formal (categorical) explanations support two interpretations: essentialist and structural. Beyond these age-general effects, they reveal developmental changes in structural thinking, with older children and adults more readily engaged in structural thinking. (Notably, we have reasons to believe that the observed pattern of developmental change is not due to younger children simply not understanding the task or explanations: in the explanation generation task younger children produced predominantly internalist explanations regardless of the framing, and when asked to break ties in the explanation evaluation task, they ranked internalist explanations higher under the internalist framing.) Moreover, these results suggest that internalist versus structural construals can be effectively induced, though in reality, they likely coexist, and are triggered by different cues. Finally, our results show that the profiling tool can effectively track internalist versus structural thinking across development.

## General Discussion

Using novel tasks designed to assess structural thinking, we find evidence that even young children are able to reason about social categories in structural terms. By age 5-6, children preferentially generated and accepted structural explanations for a category-property association when a structural constraint was presented, with hints of an emerging sensitivity by ages 3-4.
Recognizing structural reasoning as a distinct cognitive phenomenon invites us to rethink some of the findings in the literature on psychological essentialism. For example, many discussions of essentialism emphasize its capacity to support predictions and promote generalizations across category members (Gelman, 2003). In fact, generalization tasks are often used to measure the extent to which a category representation is essence-based. However, a structural representation of a category can likewise support such generalizations: structural forces shape properties of the nodes within the structure, and the occupiers of the nodes, being subject to these forces qua occupiers, are likely to obtain the properties in spite of idiosyncrasies in their individual histories and predispositions. Haslanger (2015) correspondingly praises structural explanations for their stability and identification of broad patterns that hold across "inessential perturbations," suggesting that such explanations may be particularly good in supporting generalizations within stable structures. These features of structural thinking challenge the widespread assumption that the stability and generalizability of category properties imply internalist (essentialist) representations. More generally, our findings lay the groundwork for refining internalist claims and the evidence that is taken to support them, and for making more finegrained distinctions when it comes to externalist alternatives.
We have also demonstrated that formal explanations support both structural and internalist interpretations. Introducing structural connections as a new type of nonaccidental relationship between a property and a category raises new questions about generics (e.g., "Girls prefer pink"), which are implicated in perpetuating stereotypes. On most accounts, generics are interpreted as expressing something about the underlying nature of the category, reinforcing essentialist beliefs (Cimpian \& Markman, 2011; Leslie, 2014; Prasada \& Dillingham, 2009). For example, Leslie argues that generics are by default interpreted as expressing "generalizations that hold because of common, inherent features of the members of the kind" (p. 217), where the only alternative available to people is interpreting generics as describing statistical connections, along the lines of "police officers eat donuts," on the basis of "specific worldly knowledge." But if people can interpret generics structurally, by construing features of category members as products of structural constraints rather than inherent attributes of the kind, this potentially opens up a new way to mitigate harmful side-effects of generic language without purging it from everyday speech (or, equally implausibly, convincing people that many associations between properties and social categories are merely "accidental").

By introducing a structural alternative into the dichotomy of internal vs. vaguely and variably defined external (situational) factors in explanations of behavior, we have unmasked a gap in our understanding of categorical reasoning, and opened up new directions of study that may help account for some of the mixed evidence in research on the development of relational reasoning, essentialist beliefs about social categories, and reasoning about moral and conventional norms. The reported work already calls for revision of current accounts of generic language and formal explanation, and highlights the need to study categories embedded in relational structures. But of course, a lot more remains to be done.

## Acknowledgments

This work was supported by the Varieties of Understanding Project, funded by the John Templeton Foundation, and by the McDonnell Foundation.

## References

Asmuth, J., \& Gentner, D. (2016). Relational categories are more mutable than entity categories. Quarterly Journal of Experimental Psychology, 0, 1-19.
Astuti, R., Solomon, G. E., Carey, S., Ingold, T., \& Miller, P. H. (2004). Constraints on conceptual development: A case study of the acquisition of folkbiological and folksociological knowledge in Madagascar. Monographs of the society for Research in Child Development, i-161.
Beck, S. R., Robinson, E. J., Carroll, D. J.,\& Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. Child Development, 77(2), 413-426.
Chernyak, N., \& Kushnir, T. (2014). The self as a moral agent: Preschoolers behave morally but believe in the freedom to do otherwise. Journal of Cognition and Development, 15(3), 453464.

Cimpian, A.,\& Markman, E.M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. Child Development, 82(2), 471-492.
Cimpian, A., \& Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. Behavioral and Brain Sciences, 37(5), 461-480.
Cimpian, A., \& Steinberg, O. D. (2014). The inherence heuristic across development: Systematic differences between children's and adults' explanations for everyday facts. Cognitive Psychology, 75, 130-154.
Dretske, F. (1988). Explaining behavior: Reasons in a world of causes. Cambridge, MA: MIT Press.
Fausto-Sterling, A. (2012). Sex/Gender: Biology in a social world. Routledge.
Gelman, S.A. (2003). The essential child: Origins of essentialism in everyday thought. Oxford University Press.
Gelman, S.A., \& Wellman, H.M. (1991). Insides and essences: Early understandings of the nonobvious. Cognition, 23, 183-209.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive science, 7(2), 155-170.
Gentner, D. (1988). Metaphor as structure-mapping: The relational shift. Child Development, 59, 47-59.
Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe \& D. Rakison (Eds.), Building object categories in developmental time (pp. 245-275). Mahwah, NJ: Erlbaum.
Goldwater, M. B., Bainbridge, R., \& Murphy, G. L. (2016). Learning of role-governed and thematic categories. Acta Psychologica, 164, 112-126.
Haslanger, S. (2015). What is a (social) structural explanation? Philosophical Studies, 1-18.

Hussak, L. J., \& Cimpian, A. (2015). An early-emerging explanatory heuristic promotes support for the status quo. Journal of Personality and Socisal Psychology, 109(5), 739-752.
Kalish, C.W. (2011). Generalizing norms and preferences within social categories and individuals. Developmental Psychology, 48(4), 1133-1143.
Kalish, C.W., \& Shiverick, S.M. (2004). Children's reasoning about norms and traits as motives for behavior. Cognitive Development, 19(3), 401-416.
Kelley, H. H. (1973). The processes of causal attribution. American Psychologists, 28(2), 107-128.
Kushnir, T., Xu, F., \& Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. Psychological Science, 21(8), 1134-1140.
Leslie, S.-J. (2015). The original sin of cognition: Fear, prejudice and generalization. The Journal of Philosophy, 3.
Leslie, S.-J. (2014). Carving up the social world with generics. In T. Lombrozo, J. Knobe, \& S. Nichols (Eds.), Oxford Studies in Exp. Philosophy, (pp. 208-232).
Lucas, C. G., Bridgers, S., Griffiths, T. L., \& Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. Cognition, 131(2), 284-299.
Malle, B. (2004). How the mind explains behavior: Folk explanations, meaning, and social interaction. Cambridge, MA: The MIT Press.
Markman, A.B., \& Stilwell, C.H. (2001). Role-governed categories. Journal of Experimental \& Theoretical Artificial Intelligence, 13(4), 329-358.
Nisbett, R.E. (2003). The geography of thought: How Asians and Westerners think differently... And why. New York, NY: The Free Press.
Prasada, S., \& Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. Cognition, 99, 73-112.
Prasada, S.,\& Dillingham, E. M.(2009). Representation of principled connections: a window onto the formal aspect of common sense conception. Cogitive Science, 33(3), 401-48.
Rafetseder, E., Cristi-Vargas, R., \& Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". Child development, 81(1), 376-389.
Rakoczy, H., Warneken, F., \& Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. Developmental Psychology, 44(3), 875-881.
Richland, L. E., Morrison, R. G., \& Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. Journal of Experimental Child Psychology, 94, 249-273.
Rhodes, M., \& Gelman, S. (2009). A developmental examination of the conceptual structure. Cognitive Psychology, 59(3), 244-274.
Ross, L., \& Nisbett, R. E. (2011). The person and the situation: Perspectives of social psychology. Pinter \& Martin Publishers.
Seiver, E., Gopnik, A., \& Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. Child Development, 84, 443-454.
Smetana, J.G. (1981). Preschool children's conceptions of moral and social rules. Child Development, 52, 1333-1336.
Taylor, M.G. (1996). The development of children's beliefs about social and biological aspects of gender differences. Child Development, 67, 1555-1571.
Thibaut, J. P., French, R., \& Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive function. Journal of Experimental Child Psychology, 106(1), 1-19.
Turiel, E. (1983). The development of social knowledge: Morality and convention. Cambridge, UK: Cambridge University Press. Vasilyeva, N. \& Lombrozo, T. (in prep.) Structural thinking.

# The statistical significance filter leads to overconfident expectations of replicability 

Shravan Vasishth (vasishth@uni-potsdam.de)<br>Department of Linguistics, University of Potsdam, Potsdam 14476, Germany.<br>Andrew Gelman (gelman@stat.columbia.edu)<br>Department of Statistics, Columbia University, New York, NY 10027, USA.


#### Abstract

We show that publishing results using the statistical significance filter-publishing only when the p-value is less than 0.05 -leads to a vicious cycle of overoptimistic expectation of the replicability of results. First, we show analytically that when true statistical power is relatively low, computing power based on statistically significant results will lead to overestimates of power. Then, we present a case study using 10 experimental comparisons drawn from a recently published metaanalysis in psycholinguistics (Jäger et al., 2017). We show that the statistically significant results yield an illusion of replicability. This illusion holds even if the researcher doesn't conduct any formal power analysis but just uses statistical significance to informally assess robustness (i.e., replicability) of results.


Keywords: Statistical significance; p-values; replicability
"‘. .. in [an]. . . academic environment that only publishes positive findings and rewards publication, an efficient way to succeed is to conduct low power studies. Why? Such studies are cheap and can be farmed for significant results, especially when hypotheses only predict differences from the null, rather than precise quantitative differences and trends." (Smaldino \& McElreath, 2016, p. 5)

## Introduction

The statistical significance filter tells us that significant results-those findings in which the p -value is less than 0.05 -are positively biased. The statistically significant estimate is, by definition, more than $t$ standard errors away from zero, where $t$ is some critical value determined by a statistical test (such as the t-test) and the pre-specified Type I error (the probability, under repeated sampling, of incorrectly rejecting the null hypothesis).

Statistical power is the probability, under repeated sampling, of correctly rejecting the null hypothesis assuming that the parameter of interest has some true point value $\mu .{ }^{1}$ It is well-known that when statistical power is low, the effect (the sample mean) will tend to be exaggerated. These are referred to as Type M errors by Gelman and Carlin (2014) (also see Gelman \& Tuerlinckx, 2000). This exaggeration of effects has been noticed in previous work (Hedges, 1984; Lane \& Dunlap, 1978), and most recently in neuroscience and epidemiology, where Button et al. (2013) refer to the exaggeration of effects in neuroscience as the "winner's curse" and "the vibration of effects." In related work, Ioannidis (2008)

[^223]discusses this exaggeration of effects in epidemiological studies in terms of the vibration ratio: the ratio of largest to smallest observed effects.

These overestimates get published and fill the literature. Now consider what happens when researchers design a new study. They read the literature and see all these big effects, then plan their next study. They do a power calculation based on these big effects and get an exaggerated estimate of power, and can easily convince themselves that they have a high powered study. Alternatively-and this is probably the more common route in many fields, such as psychology-they don't do a formal power analysis, but just rely on the informal observation that most of the previously published results had a significant effect and so the effect must be present.

A related observation about overestimation comes from the replication attempts reported by the Open Science Collaboration (2015). The authors report that the magnitude of the published p-values from the original studies were predictive of replication success. As they put it (p. 943): "...correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original P value) was ... predictive of replication success ..." From this, researchers might erroneously conclude that lower p-values are generally more predictive of replication success. In other words, an erroneous conclusion would be that a lower p-value suggests a higher probability that the effect can be detected in future repeated studies.

We show that if statistical significance is used as a filter for publishing a result, and the observed effect (or p-value) is used to determine replicability, this will lead the researcher to overestimate replicability. We demonstrate this point analytically, and then present a case study involving 10 reading studies in psycholinguistics that illustrates this illusion.

## The relationship between $p$-values and estimated power

Assume for simplicity the case that we carry out a one-sided statistical test where the null hypothesis is that the null hypothesis mean is $\mu_{0}=0$ and the alternative is that $\mu>0 .{ }^{2}$ Given some continuous data $x_{1}, \ldots, x_{n}$, we can compute the t -statistic and derive the p -value from it. For a large sample size $n$, a normal approximation allows us to use the $z$-statistic, $Z=\frac{\bar{X}-\mu_{0}}{\sigma_{X} / \sqrt{n}}$, to compute the p-value. Here, $\bar{X}$ is the mean, $\sigma_{X}$ the standard deviation, and $n$ the sample size.

[^224]The p-value is the probability of observing the z -statistic or a value more extreme assuming that the null hypothesis is true. The p-value is a random variable $P$ with the probability density function (Hung, O'Neill, Bauer, \& Kohne, 1997):

$$
\begin{equation*}
g_{\delta}(p)=\frac{\phi\left(Z_{p}-\delta\right)}{\phi\left(Z_{p}\right)}, \quad 0<p<1 \tag{1}
\end{equation*}
$$

where

- $\phi(\cdot)$ is the pdf of the standard normal distribution, Nor$\operatorname{mal}(0,1)$.
- $Z_{p}$, a random variable, is the (1-p)th percentile of the standard normal distribution.
- $\delta=\frac{\mu-\mu_{0}}{\sigma_{X} / \sqrt{n}}$ is the true point value expressed as a z-score. Here, $\mu$ is the true (unknown) point value of the parameter of interest.

Hung et al. (1997) further observe that the cumulative distribution function (cdf) of $P$ is:

$$
\begin{equation*}
G_{\delta}(p)=\int_{0}^{p} g_{\delta}(x) d x=1-\Phi\left(Z_{p}-\delta\right), \quad 0<p<1 \tag{2}
\end{equation*}
$$

where $\Phi(\cdot)$ is the cdf of the standard normal.
Once we have observed a particular z-statistic $z_{p}$, the cdf $G_{\delta}(p)$ allows us to estimate power based on the z -statistic (Hoenig \& Heisey, 2001). To estimate the p-value given that the null hypothesis is true, let the true value be $\mu=0$. It follows that $\delta=0$. Then:

$$
\begin{equation*}
p=1-\Phi\left(z_{p}\right) \tag{3}
\end{equation*}
$$

To estimate power from the observed $z_{p}$, set $\delta$ to be the observed statistic $z_{p}$, and let the critical z-score be $z_{\alpha}$, where $\alpha$ is the Type I error (typically 0.05 ). The power is therefore:

$$
\begin{equation*}
G_{z_{p}}(\alpha)=1-\Phi\left(z_{\alpha}-z_{p}\right) \tag{4}
\end{equation*}
$$

In other words, power estimated from the observed statistic is a monotonically increasing function of the observed z statistic: the larger the statistic, the higher the power estimate based on this statistic (Figure 1). Together with the common practice that only statistically significant results get published, and especially results with a large z-statistic, this leads to overestimates of power. As mentioned above, one doesn't need to actually estimate power in order to fall prey to the illusion; merely scanning the statistically significant z-scores gives an impression of consistency and invites the inference that the effect is replicable and robust. The word "reliable" is frequently used in psychology, presumably with the meaning that the result is replicable and represents the reality.

A direct consequence of Equation 4 is that overestimates of the z-statistic will lead to overestimates of power. For example, if we have 36 data points and the true effect is 0.1 on


Figure 1: The relationship between power and the observed zscore. The larger z-scores are easier to publish due to the statistical significance filter, and these published studies therefore give a mistaken impression of higher power.
some scale and standard deviation is 1 , then statistical power is $15 \%$. ${ }^{3}$

If we now re-run the same study, collecting 36 data points each time, and impose the condition that only statistically significant results with Type I error $\alpha=0.05$ are published, then only observed z-scores larger than 1.64 (for a one-sided test) would be published and the power estimate based on these z-scores must have a lower bound of

$$
\begin{equation*}
G_{Z_{\alpha}}(\alpha)=1-\Phi(1.64-1.64)=0.5 \tag{5}
\end{equation*}
$$

Thus, in a scenario where the real power is $15 \%$, and only z scores greater than or equal to $z_{\alpha}$ are published, power based on the z-score will be overestimated by at least a factor of $0.5 / 0.15=3.33$. Call this ratio the Power Inflation Index (PII).

Now, lower p-values are widely regarded as more "reliable" than $p$-values near the Type I error probability of 0.05. ${ }^{4}$ This incorrect belief, widely shared by editors, reviewers, and authors in areas like psychology and linguistics, has the effect that studies with lower p-values are more likely to be reported

[^225]and published, with the consequence that the PII will tend to be even higher than the lower bound discussed here.

We turn next to a case study involving psycholinguistic data that illustrates the illusion of replicability.

## Case study: Interference effects in reading studies

To illustrate the illusion of replicability, we consider the 10 experiments that were reviewed in the literature review and meta-analysis presented in Jäger, Engelmann, and Vasishth (2017). These were psycholinguistic studies in which the dependent measure was reading time in milliseconds of words. The experimental manipulation involved pairs of sentence types where one type was easier to read than the other; the empirical phenomenon of interest here is interference in working memory. Here, an appropriate statistical test is the two-sided paired t-test (one could do a one-sided t-test, although this is less common in psycholinguistics).

We had the raw data from these 10 studies and so were able to carry out the pairwise comparison. As discussed in detail in Jäger et al. (2017), theory predicts an effect with a negative sign. The original results as published were analyzed on the raw milliseconds scale, but here we analyze the data on the log milliseconds scale because the reading time data were log-normally distributed.

A summary of the pairwise t-test is shown in Table 1. From the table, it is clear that the studies consistently found negative values for the coefficient; this consistent result raises our confidence in the reproducibility of the result. A formal power analysis based on these studies, also shown in the last column of the table, leads to estimates of power ranging from 17 to $60 \%$.

|  |  |  | t | d | n | se | s |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| pval | power |  |  |  |  |  |  |
| 1 | -1.9 | -0.1 | 40 | 0.0 | 0.2 | 0.1 | 0.3 |
| 2 | -3.1 | -0.1 | 32 | 0.0 | 0.1 | 0.0 | 0.6 |
| 3 | -1.5 | -0.0 | 32 | 0.0 | 0.2 | 0.2 | 0.2 |
| 4 | -2.1 | -0.0 | 32 | 0.0 | 0.1 | 0.0 | 0.3 |
| 5 | -1.7 | -0.0 | 32 | 0.0 | 0.1 | 0.1 | 0.2 |
| 6 | -2.6 | -0.1 | 28 | 0.0 | 0.2 | 0.0 | 0.4 |
| 7 | -1.6 | -0.0 | 60 | 0.0 | 0.2 | 0.1 | 0.2 |
| 8 | -3.2 | -0.1 | 44 | 0.0 | 0.2 | 0.0 | 0.6 |
| 9 | -1.9 | -0.1 | 60 | 0.0 | 0.2 | 0.1 | 0.3 |
| 10 | -2.6 | -0.0 | 114 | 0.0 | 0.2 | 0.0 | 0.5 |

Table 1: Results from the paired $t$-tests for the 10 experimental comparisons. Shown are the $t$-score, the effect $d$ in $\log$ ms , the sample size n , the standard error se, the standard deviation $s$, and the $p$-value. The $t$-tests were done on the raw data from the original studies (the $t$-values reported here may deviate slightly from the published t -values). Also shown is the power estimated from each study.

## Using a Bayesian random-effects meta-analysis to estimate the power function

In Table 1, we calculated power based on the individual studies. As discussed above, these will tend to be overestimates because there is a preference to publish effects with low pvalues. How can we check this for the 10 studies? True power is unknown so we have no basis for comparing the power estimates from individual studies with a true value for power.

One way to arrive at a conservative estimate of the true power given these 10 studies is to carry out a Bayesian random-effects meta-analysis (Gelman et al., 2014). This hierarchical modelling approach allows us to determine the posterior distribution of the effect, which can then be used for computing an estimate of power. As discussed in Button et al. (2013), using estimates from a meta-analysis yields a more conservative estimate of power. In the random-effects meta-analysis, this conservativity arises due to the shrinkage property of hierarchical models: Larger sample studies receive a greater weighting in determining the posterior than smaller sample studies. Note, however, that even here the power may be an overestimate due to the fact that the studies that go into the meta-analysis are likely to have publication bias. But as we show below, the estimates of power from individual studies tend to be ever larger.

The random-effects meta-analysis model was set up as follows. Let $y_{i}$ be the effect size in log milliseconds in the $i$-th study, where $i$ ranges from 1 to $n$. Let $\mu$ be the true (unknown) effect in log ms, to be estimated by the model, and $\mu_{i}$ the true (unknown) effect in each study. Let $\sigma_{i} \log \mathrm{~ms}$ be the true standard deviation of the sampling distribution; each $\sigma_{i}$ is estimated from the sample standard error from study $i$. The standard deviation parameter $\tau$ represents between-study variability.

Then, our model for $n$ studies is as follows. The model assumes the i-th data point (the effect observed on the $\log \mathrm{ms}$ scale) $y_{i}$ is generated from a normal distribution with mean $\mu_{i}$ and some standard error $\sigma$, estimated from the sample's standard error. Each of the true underlying means $\mu_{i}$ are assumed to be generated from a normal distribution with true mean $\mu$ and between-study standard deviation $\tau$. We assign Cauchy $(0,2.5)$ priors to the parameters $\mu$ and $\mu_{i}$, and a truncated Cauchy $(0,2.5)$ prior for the between-study standard deviation $\tau$, truncated so that $\tau$ is greater than 0 . The model can be stated mathematically as follows:

## Likelihoods:

$$
\begin{align*}
& y_{i} \mid \mu_{i}, \sigma_{i}^{2} \sim \operatorname{Normal}\left(\mu_{i}, \sigma_{i}^{2}\right) \quad i=1, \ldots, n \\
& \mu_{i} \mid \theta, \tau^{2} \sim \operatorname{Normal}\left(\mu, \tau^{2}\right), \\
& \text { Priors: }  \tag{6}\\
& \mu \sim \operatorname{Cauchy}(0,2.5), \\
& \mu_{i} \sim \operatorname{Cauchy}(0,2.5), \\
& \tau \sim \operatorname{Cauchy}(0,2.5), \tau>0
\end{align*}
$$



Figure 2: Posterior distributions of the estimated effect ( $\hat{\mu}$ ), and the standard deviation of estimate of the between-study variability $(\hat{\tau})$ in the random-effects meta-analysis.

We fit the model using Stan 2.14.2 (Stan Development Team, 2016), running four chains with 4000 iterations (half of which were warm-ups). Convergence was successful, as diagnosed using the $\hat{R}$ diagnostic (Gelman et al., 2014). The posterior distributions of $\hat{\mu}$ and of the between-study standard deviation $\hat{\tau}$ are shown in Figure 2. The posterior mean of the effect is $-0.05 \log \mathrm{~ms}$, with $95 \%$ credible interval [-0.08,$0.03]$. Next, we use this estimate of the posterior distribution to compute a power distribution.
Computing the power distribution using the posterior distribution of the effect An analysis of reading studies, including the ones considered here, showed that the precisions (the inverse of the variance) in reading time studies have mean values 16.3 and standard deviation 7.07 (the unit for precision is $1 / \log \mathrm{ms}^{2}$ ). Since precision can be modelled as a Gamma distribution, we assumed that precisions are distributed as $\operatorname{Gamma}(\alpha=5.3, \beta=0.3)$. These parameters of the Gamma distribution were computed by taking the mean $\bar{x}$ and standard deviation $s$ of the precisions, and then deriving the parameters of the Gamma distribution by solving for $\alpha$ and $\beta$. We use the fact that for a random variable generated from a Gamma distribution with parameters $\alpha$ and $\beta$, the expectation $\mu$ and variance $\sigma^{2}$ are:

$$
\begin{equation*}
E(X)=\frac{\alpha}{\beta}=\mu \quad \text { and } \operatorname{Var}(X)=\frac{\alpha}{\beta^{2}}=\sigma^{2} \tag{7}
\end{equation*}
$$

Having obtained the estimate of the effect (through the meta-analysis) and the distribution of the precisions, we used these estimates to carry out 100,000 Monte Carlo simulations to derive a power distribution for different sample sizes ( $n=20, \ldots, 50$ ) in the following manner. For each sample size, we repeatedly computed power after obtaining:

- one sample for the effect by sampling from the distribution $\operatorname{Normal}(-0.05,0.01)$; this is the posterior distribution of the effect derived from the random-effects meta-analysis;
- one sample for the precision by sampling from the Gamma $(5.3,0.3)$, and then converting this to a standard deviation.

Such a Monte Carlo sampling procedure gives a probability distribution of power values and allows us to quantify our uncertainty about the estimated power by taking all sources of uncertainty into account-the uncertainty regarding the effect, and the uncertainty regarding the standard deviation.

Figure 3 shows the resulting power distributions for power given different sample sizes. These power distributions are of course only estimates, not the true power; and as Button et al. (2013) point out, are probably slight overestimates if the studies themselves have publication bias.

The power distributions illustrate two important points. First, the range of most likely power values is remarkably low for typical sample sizes used in psycholinguistic reading experiments relating to interference effects (see Table 1). As an aside, we note that our estimates are similar to those from a recent review of 44 meta-analyses of research in social and behavioural sciences published between 1960-2011; they report a mean power of 0.24 with most studies suggesting power to be below 0.4 (Smaldino \& McElreath, 2016, p. 6, Fig. 1). The second observation is that the power values computed from individual studies (the red dots) tend to be overestimates relative to the mean of each power distribution shown. The power from each study tends to be higher than the mean of each power distribution. Of course, if the statistical power of the original studies were very high (approximately $80 \%$ or higher), then the overestimation problem would disappear or at least be negligible.

We can quantify the overestimation of power by computing the Power Inflation Index: the ratio of the power computed from individual studies to the power distribution computed using Monte Carlo simulations. If power is overestimated, then the distribution of the PII will be such that the mean ratio will be greater than 1 . These distributions of PIIs are computed for a typical sample size used in psycholinguistic studies ( $\mathrm{n}=20,30,40$ ) in Table 2. Here, we can see that the PII can be as high as 12 .

## Discussion

We have shown that if statistical significance is used to decide whether to publish a result, overestimates of the effect will tend to be published, leading to an over-enthusiastic belief in the replicability of the effect.

Recently, the replication project reported by Open Science Collaboration (2015) showed that only $47 \%$ of the studies they investigated could be replicated. One factor causing these failures to replicate could have been low power in the original studies. Even before the replication project, Cohen $(1962,1988)$ and others have repeatedly warned against running low-powered studies. Despite these injunctions, many researchers do not believe that there is a problem of low power. For example, Gilbert, King, Pettigrew, and Wilson (2016) contested the $47 \%$ replication rate and argued that


Figure 3: Power distributions for different sample sizes (log reading times). The histogram shows the power distribution (generated through Monte Carlo sampling; see text for details). The red dots show power estimates from the 10 individual experimental comparisons considered in this case study. The white dot shows the mean of each power distribution.

| $\mathrm{n}=20$ |  |  |  | $\mathrm{n}=30$ |  | $\mathrm{n}=40$ |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
| Study | $2.5 \%$ | $97.5 \%$ | $2.5 \%$ | $97.5 \%$ | $2.5 \%$ | $97.5 \%$ |  |
| 1 | 1.37 | 4.64 | 0.98 | 3.95 | 0.76 | 3.47 |  |
| 2 | 3.67 | 12.45 | 2.63 | 10.61 | 2.04 | 9.30 |  |
| 3 | 1.08 | 3.67 | 0.78 | 3.13 | 0.60 | 2.74 |  |
| 4 | 1.95 | 6.61 | 1.40 | 5.64 | 1.08 | 4.94 |  |
| 5 | 1.41 | 4.76 | 1.01 | 4.06 | 0.78 | 3.56 |  |
| 6 | 3.06 | 10.37 | 2.19 | 8.83 | 1.70 | 7.75 |  |
| 7 | 0.76 | 2.59 | 0.55 | 2.20 | 0.42 | 1.93 |  |
| 8 | 2.99 | 10.12 | 2.14 | 8.62 | 1.65 | 7.56 |  |
| 9 | 0.98 | 3.34 | 0.71 | 2.84 | 0.55 | 2.49 |  |
| 10 | 1.02 | 3.47 | 0.73 | 2.96 | 0.57 | 2.59 |  |

Table 2: The 95\% credible intervals of the Power Inflation Index for each of the 10 experimental comparisons, for different sample sizes. The Power Inflation Index can be as large as 12.
the replication rate may be much higher, perhaps even "statistically indistinguishable from $100 \%$." The objections of Gilbert et al. (2016) were largely based on arguments about the lack of fidelity to the original design, but it is possible that, in addition to concerns about fidelity, Gilbert et al. are, like many researchers, generally overconfident about the replicability and robustness of their results. This overconfidence is also evident in reading research in psycholinguistics, where it is routine to run experiments with sample sizes ranging from 20 to 40 participants. Recent work has argued that sample sizes of 20-40 partipants may be too low for reading studies on interference (Jäger et al., 2017). We are hopeful that future work will take this finding into account when planning studies.

Currently, the replication problems in psycholinguistics are serious. For example, in recent work (Mertzen, Jäger, \& Vasishth, 2017) we carried out six replication attempts of two eyetracking experiments published in the Journal of Memory and Language. We were unable to replicate any of the claims in the paper. There is thus an urgent need to attempt to replicate published results, and not just in psycholinguistics. For example, Makel, Plucker, and Hegarty (2012) present a quantitative analysis of the low rate of successful replications in psychology ( $1 \%$ ). Other fields are also affected. For example, Button et al. (2013) have shown that in neuroscience studies, power may also be quite low, ranging from 8 to $31 \%$. Smaldino and McElreath (2016) have shown through a 50year meta-analysis in behavioural science that power has not improved (mean power: $24 \%$ ). In biomedical sciences, approximately $50 \%$ of studies have power in the $0-10 \%^{5}$ or 11$20 \%$ range (Dumas-Mallet, Button, Boraud, Gonon, \& Munafò, 2017).

Despite these indications, many researchers remain overconfident about the robustness of their results. This overconfidence is in part due to the statistical significance filter.

## Concluding remarks

We have shown that the statistical significance filter directly leads to over-optimistic expectations of replicability of published research. Even if the researcher doesn't conduct any formal power analyses, they can fall prey to this illusion because of the informal assessment of replicability afforded by the statistical significance filter. We illustrated the illusion of replicability through a case-study involving 10 published experimental comparisons.

Many psychology journals are beginning to require that power analyses be included in submitted manuscripts. But our results, echoing those of others who have studied this problem, suggest that such analyses, which invariably are based on previously published work, will tend to provide overestimates of power.

To resolve or at least reduce this problem, we offer two pieces of advice. First, we recommend entirely abandon-

[^226]ing the concept of power, which is based on the idea that " $p<.05$ " is a win, an attitude that fails miserably when effect sizes are small and measurements are noisy. Second, when performing design analysis, consider possible effect sizes based on subject-matter understanding; see Gelman and Carlin (2014) for further discussion of this point. It can make sense to consider a range of reasonable effect sizes.

## Appendix

Here, we review the well-known proof that for a point null hypothesis and a continuous dependent variable, the distribution of the p-value under the null is Uniform $(0,1)$.

When a random variable $Z$ comes from a $\operatorname{Uniform}(0,1)$ distribution, then the probability that $Z$ is less than (or equal to) some value $z$ is exactly $z: P(Z \leq z)=z$.

The p -value is a random variable, call it $Z$. The p -value is computed by calculating the probability of seeing a $t$-statistic or something more extreme under the null hypothesis. The t -statistic comes from a random variable $T$ that is a transformation of the random variable $\bar{X}: T=(\bar{X}-\mu) /(\sigma / \sqrt{n})$. This random variable T has a $\mathrm{CDF} F$.

We can establish that if a random variable $Z=F(T)$, then $Z \sim \operatorname{Uniform}(0,1)$, i.e., that the p-value's distribution under the null hypothesis is $\operatorname{Uniform}(0,1)$. This is proved next.

Let $Z=F(T)$. Then: $\quad P(Z \leq z)=P(F(T) \leq z)=$ $P\left(F^{-1} F(T) \leq F^{-1}(z)\right)=P\left(T \leq F^{-1}(z)\right)=F\left(F^{-1}(z)\right)=z$.

Since $P(Z \leq z)=z, \mathrm{Z}$ is uniformly distributed, that is, $\operatorname{Uniform}(0,1)$.

## Acknowledgements

We thank Lena Jäger, Reinhold Kliegl, Christian Robert, and Bruno Nicenboim for helpful discussions. For partial support of this research, we thank the Volkswagen Foundation through grant 89 953, and the U.S. Office of Naval Research through grant N00014-15-1-2541.

## References

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., \& Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365376.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. The Journal of Abnormal and Social Psychology, 65(3), 145.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., \& Munafò, M. R. (2017). Low statistical power in biomedical science: A review of three human research domains. Royal Society Open Science, 160254. doi: 10.1098/rsos. 160254
Gelman, A., \& Carlin, J. (2014). Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. Perspectives on Psychological Science, 9(6), 641-651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., \& Rubin, D. B. (2014). Bayesian data analysis (Third ed.). Chapman and Hall/CRC.
Gelman, A., \& Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics, 15(3), 373-390.
Gilbert, D. T., King, G., Pettigrew, S., \& Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". Science, 351(6277), 1037-1037. doi: 10.1126/science.aad7243
Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. Journal of Educational Statistics, 9(1), 61-85.
Hoenig, J. M., \& Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician, 55(1), 19-24.
Hung, H. J., O'Neill, R. T., Bauer, P., \& Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. Biometrics, 11-22.
Ioannidis, J. P. (2008). Why most discovered true associations are inflated. Epidemiology, 19(5), 640-648.
Jäger, L. A., Engelmann, F., \& Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. Journal of Memory and Language, 94, 316-339. doi: 10.1016/j.jml.2017.01.004

Lane, D. M., \& Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. British Journal of Mathematical and Statistical Psychology, 31(2), 107-112.
Makel, M. C., Plucker, J. A., \& Hegarty, B. (2012). Replications in psychology research: How often do they really occur? Perspectives on Psychological Science, 7(6), 537542.

Mertzen, D., Jäger, L. A., \& Vasishth, S. (2017). The importance of replication in psycholinguistics. In Proceedings of the 30th Annual CUNY Conference on Sentence Processing. Boston, USA. Retrieved from https://osf.io/j66z5/
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.
R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org
Smaldino, P. E., \& McElreath, R. (2016). The natural selection of bad science. Royal Society Open Science, 3(9), 160384.

Stan Development Team. (2016). Stan modeling language users guide and reference manual, version 2.12 [Computer software manual]. Retrieved from http://mc-stan.org/

# Modelling dependency completion in sentence comprehension as a Bayesian hierarchical mixture process: A case study involving Chinese relative clauses 

Shravan Vasishth (vasishth@uni-potsdam.de)<br>Department of Linguistics, University of Potsdam, Germany.<br>Nicolas Chopin (nicolas.chopin@ensae.fr)<br>École Nationale de la Statistique et de l'administration économique, Malakoff, France.<br>Robin Ryder (ryder@ceremade.dauphine.fr)<br>Centre de Recherche en Mathématiques de la Décision, CNRS, UMR 7534, Université Paris-Dauphine, PSL Research University, Paris, France.<br>Bruno Nicenboim (bruno.nicenboim@uni-potsdam.de)<br>Department of Linguistics, University of Potsdam, Germany.


#### Abstract

We present a case-study demonstrating the usefulness of Bayesian hierarchical mixture modelling for investigating cognitive processes. In sentence comprehension, it is widely assumed that the distance between linguistic co-dependents affects the latency of dependency resolution: the longer the distance, the longer the retrieval time (the distance-based account). An alternative theory, direct-access, assumes that retrieval times are a mixture of two distributions: one distribution represents successful retrievals (these are independent of dependency distance) and the other represents an initial failure to retrieve the correct dependent, followed by a reanalysis that leads to successful retrieval. We implement both models as Bayesian hierarchical models and show that the direct-access model explains Chinese relative clause reading time data better than the distance account.


Keywords: Bayesian Hierarchical Finite Mixture Models; Psycholinguistics; Sentence Comprehension; Chinese Relative Clauses; Direct-Access Model; K-fold Cross-Validation

## Introduction

Bayesian cognitive modelling (Lee \& Wagenmakers, 2014), using probabilistic programming languages like JAGS (Plummer, 2012), is an important tool in cognitive science. We present a case study from sentence processing research showing how hierarchical mixture models can be profitably used to develop probabilistic models of cognitive processes. Although the case study concerns a specialized topic in psycholinguistics, the approach developed here will be of general interest to the cognitive science community.

In sentence comprehension research, dependency completion is assumed by many theories to be a key event. For example, consider a sentence such as (1):
(1) a. The man (on the bench) was sleeping

In order to understand who was doing what, the noun The man must be recognized to be the subject of the verb phrase was sleeping; this dependency is represented here as a directed arrow. One well-known proposal (Just \& Carpenter, 1992), which we will call the distance account, is that dependency distance between linguistically related elements partly
determines comprehension difficulty as measured by reading times or question-response accuracy. For example, the Dependency Locality Theory (DLT) by Gibson (2000) and the cue-based retrieval account of Lewis and Vasishth (2005) both assume that the longer the distance between two codependents such as a subject and a verb, the greater the retrieval difficulty at the moment of dependency completion. As shown in (1), the distance between co-dependents can increase if a phrase intervenes.

As another example, consider the self-paced reading study in Gibson and Wu (2013) in Chinese subject and object relative clauses. The dependent variable here was the reading time at the head noun (official). As shown in (2), the distance between the head noun and the gap it is coindexed with is larger in subject relatives compared to object relatives. ${ }^{1}$ Thus, the distance account predicts an object relative advantage. For simplicity, we operationalize distance here as the number of words intervening between the gap inside the relative clause and the head noun. In the DLT, distance is operationalized as the number of (new) discourse referents intervening between two co-dependents; and in the cue-based retrieval model, distance is operationalized in terms of decay in working memory (i.e., time passing by).

## a. Subject relative

[GAP $_{i}$ yaoqing fuhao de] guanyuan
GAP invite tycoon DE official
xinhuaibugui
have bad intentions
'The official who invited the tycoon has bad intentions.
b. Object relative

[^227][fuhao yaoqing $\mathrm{GAP}_{i}$ de] guanyuan ${ }_{i}$
tycoon invite
 xinhuaibugui
have bad intentions
'The official who the tycoon invited has bad intentions.

In the Gibson and Wu study, reading times were recorded using self-paced reading in the two conditions, with 37 subjects and 15 items, presented in a standard Latin square design. The experiment originally had 16 items, but one item was removed in the published analysis due to a mistake in the item. We coded subject relatives as $-1 / 2$, and object relatives as $+1 / 2$; this implies that an overall object relative advantage would show a negative coefficient. In other words, an object relative advantage corresponds to a negative sign on the estimate.

The distance account's predictions can be evaluated by fitting the hierarchical linear model shown in (1). Assume that (i) $i$ indexes participants, $i=1, \ldots, I$ and $j$ indexes items, $j=1, \ldots, J$; (ii) $y_{i j}$ is the reading time in milliseconds for the $i$-th participant reading the $j$-th item; and (iii) the predictor $X$ is sum-coded $( \pm 1 / 2)$, as explained above. Then, the data $y_{i j}$ (reading times in milliseconds) are defined to be generated by the following model:

$$
\begin{equation*}
y_{i j}=\beta_{0}+\beta_{1} X_{i j}+u_{i}+w_{j}+\varepsilon_{i j} \tag{1}
\end{equation*}
$$

where $u_{i} \sim \operatorname{Normal}\left(0, \sigma_{u}^{2}\right), w_{j} \sim \operatorname{Normal}\left(0, \sigma_{w}^{2}\right)$ and $\varepsilon_{i j} \sim$ $\operatorname{Normal}\left(0, \sigma_{e}^{2}\right)$; all three sources of variance are assumed to be independent. The terms $u_{i}$ and $w_{j}$ are called varying intercepts for participants and items respectively; they represent by-subject and by-item adjustments to the fixed-effect intercept $\beta_{0}$. Their variances, $\sigma_{u}^{2}$ and $\sigma_{w}^{2}$ represent betweenparticipant (respectively item) variance.

This model is effectively a statement about the generative process that produced the data. If the distance account is correct, we would expect to find evidence that the slope $\beta_{1}$ is negative; specifically, reading times for object relatives are expected to be shorter than those for subject relatives. As shown in Table 1, this prediction appears, at first sight, to be borne out. Subject relatives are estimated to be read 120 ms slower than object relatives, apparently consistent with the predictions of the distance account.

|  | Estimate | Std. Error | t value |
| :--- | ---: | ---: | ---: |
| $\hat{\beta}_{0}$ | 548.43 | 51.56 | $10.64^{*}$ |
| $\hat{\beta}_{1}$ | -120.39 | 48.01 | $-2.51^{*}$ |

Table 1: A linear mixed model using raw reading times in milliseconds as dependent variable, corresponding to the reported results in Gibson and Wu 2013. Statistical significance is shown by an asterisk.

The object relative advantage shown in Table 1 was origi-
nally presented in Gibson and Wu (2013) as a repeated measures ANOVA.

To summarize, the conclusion from the above result would be that in Chinese, subject relatives are harder to process than object relatives because the gap inside the relative clause is more distant from the head noun in subject vs. object relatives. This makes it more difficult to complete the gap-head noun dependency in subject relatives. This distance-based explanation of processing difficulty is plausible given the considerable independent evidence from languages such as English, German, Hindi, Persian and Russian that dependency distance can affect reading time (see review in Safavi, Husain, and Vasishth (2016)).


Figure 1: Boxplots showing the distribution of reading times by condition of the Gibson and Wu (2013) data.

However, the distributions of the reading times for the two conditions show an interesting asymmetry that cannot be straightforwardly explained by the distance account. At the head noun, the reading times in subject relatives are much more spread out than in object relatives. This is shown in Figure 1, where reading times are shown on the log scale. Although this spread was ignored in the original analysis, a standard response to heterogeneous variances (heteroscedasticity) is to delete "outliers" based on some criterion; a common criterion is to delete all data lying beyond $\pm 2.5 S D$ in each condition. ${ }^{2}$ This procedure assumes that the data points identified as extreme are irrelevant to the question being investigated. An alternative approach is to not delete data but to downweight the extreme values by applying a variance stabilizing transform (Box \& Cox, 1964). Taking a log-transform

[^228]of the reading time data, or a reciprocal transform, can reduce the heterogeneity in variance; see Vasishth, Chen, Li, and Guo (2013) for analyses of the Gibson and Wu data using a transformation.

One might think that if subject and object relatives are generated by LogNormal distributions with different means, then modelling the data as being generated by LogNormals would adequately explain the data. Table 2 shows that if we assume such a model, there is no longer a statistically significant object relative advantage: the absolute $t$-value for the estimate of the $\beta_{1}$ parameter is smaller than the critical value of 2 (Bates, Maechler, Bolker, \& Walker, 2015). Thus, assuming that the data are generated by LogNormal distributions with different means for the subject and object relatives leads to the conclusion that there isn't much evidence for the distance account.

|  | Estimate | Std. Error | t value |
| :--- | ---: | ---: | ---: |
| $\hat{\beta}_{0}$ | 6.06 | 0.07 | $92.64^{*}$ |
| $\hat{\beta}_{1}$ | -0.07 | 0.04 | -1.61 |

Table 2: A linear mixed model using log reading times in milliseconds as dependent variable in the Gibson and Wu, 2013, data.

Consider next the possibility that the heteroscedasticity in subject and object relatives in the Gibson and Wu data reflects a systematic difference in the underlying generative processes of reading times in the two relative clause types. We investigate this question by modelling the extreme values as being generated from a mixture distribution.

Using the probabilistic programming language Stan (Stan Development Team, 2016), we show that a hierarchical mixture model provides a better fit to the data (in terms of predictive accuracy) than several simpler hierarchical models. As Nicenboim and Vasishth (2017) pointed out, the underlying generative process implied by a mixture model is consistent with the direct-access model of McElree, Foraker, and Dyer (2003). We therefore suggest that, at least for the Chinese relative clause data considered here, the direct-access model may be a better way to characterize the dependency resolution process than the distance account.

We can implement the direct-access model as a hierarchical mixture model with retrieval time assumed to be generated from one of two distributions, where the proportion of trials in which a retrieval failure occurs (the mixing proportion) is $p_{s r}$ in subject relatives, and $p_{o r}$ in object relatives. The expectation here is the extreme values that are seen in subject relatives are due to $p_{s r}$ being larger than $p_{o r}$.

Subject relatives

$$
\begin{align*}
y_{i j} & \sim p_{s r} \cdot \operatorname{LogNormal}\left(\beta+\delta+u_{i}+w_{j}, \sigma_{e^{\prime}}^{2}\right) \\
& +\left(1-p_{s r}\right) \cdot \operatorname{LogNormal}\left(\beta+u_{i}+w_{j}, \sigma_{e}^{2}\right) \tag{2}
\end{align*}
$$

Object relatives

$$
\begin{aligned}
y_{i j} & \sim p_{o r} \cdot \log N o r m a l \\
& \left(\beta+\delta+u_{i}+w_{j}, \sigma_{e^{\prime}}^{2}\right) \\
& +\left(1-p_{\text {or }}\right) \cdot \operatorname{LogNormal}\left(\beta+u_{i}+w_{j}, \sigma_{e}^{2}\right)
\end{aligned}
$$

Here, the terms $u_{i}$ and $w_{j}$ have the same interpretation as in equation 1.

## Model comparison

Bayesian model comparison can be carried out using different methods. Here, we use Bayesian k-fold cross-validation as discussed in Vehtari, Gelman, and Gabry (2016). This method evaluates the predictive performance of alternative models, and models with different numbers of parameters can be compared (Vehtari, Ojanen, et al., 2012; Gelman, Hwang, \& Vehtari, 2014).

The k-fold cross-validation algorithm is as follows:

1. Split data pseudo-randomly into $K$ held-out $\operatorname{sets} \mathbf{y}_{(k)}$, where $k=1, \ldots, K$ that are a fraction of the original data, and $K$ training sets, $\mathbf{y}_{(-k)}$. Here, we use $K=10$, and the length of the held-out data-vector $\mathbf{y}_{(k)}$ is approximately $1 / K$-th the size of the full data-set. We ensure that each participant's data appears in the training set and contains an approximately balanced number of data points for each condition.
2. Sample from the model using each of the $K$ training sets, and obtain posterior distributions $p_{\operatorname{post}(-\mathrm{k})}(\theta)=p(\theta \mid$ $\left.\mathbf{y}_{(-k)}\right)$, where $\theta$ is the vector of model parameters.
3. Each posterior distribution $p\left(\theta \mid \mathbf{y}_{(-k)}\right)$ is used to compute predictive accuracy for each held-out data-point $y_{i}$ :

$$
\begin{equation*}
\log p\left(y_{i} \mid \mathbf{y}_{(-k)}\right)=\log \int p\left(y_{i} \mid \theta\right) p\left(\theta \mid \mathbf{y}_{(-k)}\right) d \theta \tag{3}
\end{equation*}
$$

4. Given that the posterior distribution $p\left(\theta \mid \mathbf{y}_{(-k)}\right)$ is summarized by $s=1, \ldots, S$ simulations, i.e., $\theta^{k, s}$, $\log$ predictive density for each data point $y_{i}$ in subset $k$ is computed as

$$
\begin{equation*}
\widehat{e l p d}_{i}=\log \left(\frac{1}{S} \sum_{s=1}^{S} p\left(y_{i} \mid \theta^{k, s}\right)\right) \tag{4}
\end{equation*}
$$

5. Given that all the held-out data in the $K$ subsets are $y_{i}$, where $i=1, \ldots, n$, we obtain the $\widehat{\text { elpd }}$ for all the held-out data points by summing up the $\widehat{\text { elpd }}_{i}$ :

$$
\begin{equation*}
\widehat{e l p d}=\sum_{i=1}^{n} \widehat{e l p d}_{i} \tag{5}
\end{equation*}
$$

The difference between the $\widehat{\text { elpd}}$ 's of two competing models is a measure of relative predictive performance. We can also compute the standard deviation of the sampling distribution (the standard error) of the difference in $\widehat{\text { elpd }}$ using the formula discussed in Vehtari et al. (2016). Letting $\widehat{E L P D}$ be the vector $\widehat{e l p d}_{1}, \ldots, \widehat{e l p d}_{n}$, we can write:

$$
\begin{equation*}
\operatorname{se}\left(\widehat{e l p d}_{m 0}-\widehat{e l p d}_{m 1}\right)=\sqrt{n \operatorname{Var}(\widehat{E L P D})} \tag{6}
\end{equation*}
$$

When we compare the model (1) with (2), if (2) has a higher $\widehat{e l p d}$, then it has a better predictive performance compared to (1).

The quantity $\widehat{\text { elpd }}$ is a Bayesian alternative to the Akaike Information Criterion (Akaike, 1974). Note that the relative complexity of the models to be compared is not relevant: the sole criterion here is out-of-sample predictive performance. As we discuss below (Results section), increasing complexity will not automatically lead to better predictive performance. See Vehtari et al. (2012); Gelman, Hwang, and Vehtari (2014) for further details. ${ }^{3}$

## The data

The evaluation of these models was carried out using two separate data-sets. The first was the original study from Gibson and Wu (2013) that was discussed in the introduction. The second study was a replication of the Gibson and Wu study that was published in Vasishth et al. (2013). This second study served the purpose of validating whether independent evidence can be found for the mixture model selected using the original Gibson and Wu data.

## Results

In the models presented below, the dependent variable is reading time in milliseconds. Priors are defined for the model parameters as follows. All standard deviations are constrained to be greater than 0 and have priors Cauchy $(0,2.5)$ (Gelman, Carlin, et al., 2014); probabilities have priors $\operatorname{Beta}(1,1)$; and all coefficients ( $\beta$ parameters) have priors Cauchy $(0,2.5)$.

Fake-data simulation for validating model Before evaluating relative model fit, we first simulated data from a mixture distribution with known parameter values, and then sampled from the models representing the distance account and the direct-access model. The goal of fake-data simulation was to validate the models and the model comparison method: with reference to the simulated data, we asked (a) whether the $95 \%$ credible intervals of the posterior distributions of the parameters in the mixture model contain the true parameter values used to generate the data; and (b) whether k-fold cross validation can identify the mixture model as the correct one when the underlying generative process matches the mixture model.

[^229]The answer to both questions was "yes". This raises our confidence that the models can identify the underlying parameters with real data. The fake-data simulation also showed that when the true underlying generative process was consistent with the distance account but not the direct access model, the hierarchical linear model and the mixture model had comparable predictive performance. In other words, the mixture model furnished a superior fit only when the true underlying generative process for the data was in fact a mixture process. Further details are omitted here due to lack of space.
The original Gibson and Wu study The estimates from the hierarchical linear model (equation 1) and the mixture model (equation 2) are shown in Tables 3 and 4. Note that in Bayesian modelling we are not interested in "statistical significance" here; rather, the goal is inference and comparing predictive performance of two competing models.

|  | mean | lower | upper |
| :--- | ---: | ---: | ---: |
| $\hat{\beta}_{1}$ | 6.06 | 5.91 | 6.20 |
| $\hat{\beta}_{2}$ | -0.07 | -0.16 | 0.02 |
| $\hat{\sigma}_{e}$ | 0.52 | 0.49 | 0.55 |
| $\hat{\sigma}_{u}$ | 0.25 | 0.18 | 0.34 |
| $\hat{\sigma}_{w}$ | 0.20 | 0.12 | 0.33 |

Table 3: Posterior parameter estimates from the hierarchical linear model (equation 1) corresponding to the distance account. The data are from Gibson and $\mathrm{Wu}, 2013$. Shown are the mean and $95 \%$ credible intervals for each parameter.

|  | mean | lower | upper |
| ---: | ---: | ---: | ---: |
| $\hat{\boldsymbol{\beta}}_{0}$ | 5.85 | 5.76 | 5.95 |
| $\hat{\delta}$ | 0.93 | 0.73 | 1.14 |
| $\hat{p}_{s r}-\hat{p}_{o r}$ | 0.04 | -0.04 | 0.13 |
| $\hat{p}_{s r}$ | 0.25 | 0.17 | 0.34 |
| $\hat{p}_{o r}$ | 0.21 | 0.14 | 0.29 |
| $\hat{\sigma}_{e^{\prime}}$ | 0.64 | 0.54 | 0.74 |
| $\hat{\boldsymbol{\sigma}}_{e}$ | 0.22 | 0.20 | 0.25 |
| $\hat{\boldsymbol{\sigma}}_{u}$ | 0.24 | 0.18 | 0.31 |
| $\hat{\sigma}_{w}$ | 0.09 | 0.05 | 0.16 |

Table 4: Posterior parameter estimates from the hierarchical mixture model (equation 2) corresponding to the directaccess model. The data are from Gibson and Wu, 2013. Shown are the mean and $95 \%$ credible intervals for each parameter.

Table 4 shows that the mean difference between the probability $p_{s r}$ and $p_{o r}$ is $4 \%$; the posterior probability of this difference being greater than zero is $82 \%$. K-fold crossvalidation shows that $\widehat{e l p d}$ for the hierarchical model is -3761 (SE: 38) and for the mixture model is -3614 (35). The difference between the two $\widehat{\text { elpds }}$ is 148 (18). The larger $\widehat{e l p d}$ in the hierarchical mixture model suggests that it has better predictive performance than the hierarchical lin-
ear model. In other words, the direct-access model has better predictive performance than the distance model.
The replication of the Gibson and Wu study This dataset, originally reported by Vasishth et al. (2013), had 40 participants and the same 15 items as in Gibson and Wu's data. Figure 2 shows the distribution of the data by condition; there seems to a similar skew as in the original study, although the spread is not as dramatic as in the original study.


Figure 2: Boxplots showing the distribution of reading times by condition of the replication of the Gibson and Wu data.

Tables 5 and 6 show the estimates of the posterior distributions from the two models. Table 4 shows that the mean difference between the probability $p_{s r}$ and $p_{o r}$ is $7 \%$; the posterior probability of this difference being greater than zero is $96 \%$.

The $\widehat{\text { elpd }}$ for the hierarchical model is -3959 (53), and for the hierarchical mixture model, -3801 (38). The difference in $\widehat{e l p d}$ is 158 (29). Thus, in the replication data as well, the predictive performance of the mixture model is better than the hierarchical linear model.

|  | mean | lower | upper |
| :--- | ---: | ---: | ---: |
| $\hat{\beta}_{0}$ | 6.00 | 5.88 | 6.12 |
| $\hat{\beta}_{1}$ | -0.09 | -0.16 | -0.01 |
| $\hat{\sigma}_{e}$ | 0.44 | 0.41 | 0.47 |
| $\hat{\sigma}_{u}$ | 0.25 | 0.19 | 0.33 |
| $\hat{\sigma}_{w}$ | 0.16 | 0.10 | 0.26 |

Table 5: Posterior parameter estimates from the hierarchical linear model (equation 1) corresponding to the distance account. The data are from the replication of Gibson and Wu, 2013 reported in Vasishth et al., 2013. Shown are the mean and $95 \%$ credible intervals for each parameter.

|  | mean | lower | upper |
| ---: | ---: | ---: | ---: |
| $\hat{\boldsymbol{\beta}}_{0}$ | 5.86 | 5.78 | 5.95 |
| $\hat{\delta}^{\delta}$ | 0.75 | 0.56 | 0.97 |
| $\hat{p}_{s r}-\hat{p}_{o r}$ | 0.07 | -0.01 | 0.15 |
| $\hat{p}_{s r}$ | 0.23 | 0.15 | 0.33 |
| $\hat{p}_{o r}$ | 0.16 | 0.09 | 0.25 |
| $\hat{\sigma}_{e^{\prime}}$ | 0.69 | 0.59 | 0.81 |
| $\hat{\sigma}_{e}$ | 0.21 | 0.18 | 0.23 |
| $\hat{\sigma}_{u}$ | 0.22 | 0.17 | 0.29 |
| $\hat{\sigma}_{w}$ | 0.07 | 0.04 | 0.12 |

Table 6: Posterior parameter estimates from the hierarchical linear model (equation 2) corresponding to the direct-access model. The data are from the replication of Gibson and Wu, 2013 reported in Vasishth et al., 2013. Shown are the mean and $95 \%$ credible intervals for each parameter.

## Discussion

The model comparison and parameter estimates presented above suggest that, at least as far as the Chinese relative clause data are concerned, a better way to characterize the dependency completion process is in terms of the direct-access model and not the distance account implied by Gibson and Wu (2013) and Lewis and Vasishth (2005). Specifically, there is suggestive evidence in the Gibson and Wu (2013) data that a higher proportion of retrieval failures occurred in subject relatives compared to object relatives. In other words, increased dependency distance may have the effect that it increases the proportion of retrieval failures (followed by reanalysis). ${ }^{4}$

There is one potential objection to the conclusion above. It would be important to obtain independent evidence as to which dependency was eventually created in each trial. This could be achieved by asking participants multiple-choice questions to find out which dependency they built in each trial. Although such data is not available for the present study, in other work (on number interference) (Nicenboim, Engelmann, Suckow, \& Vasishth, 2016) did collect this information. There, too, we found that the direct-access model best explains the data (Nicenboim \& Vasishth, 2017). In future work on Chinese relatives, it would be helpful to carry out a similar study to determine which dependency was completed in each trial. In the present work, the modelling at least shows how the extreme values in subject relatives can be accounted for by assuming a two-mixture process.

## Conclusion

The mixture models suggest that, in the specific case of Chinese relative clauses, increased processing difficulty in subject relatives is not due to dependency distance leading to longer reading times, as suggested by Gibson and Wu (2013).

[^230]Rather, a more plausible explanation for these data is in terms of the direct-access model of McElree et al. (2003). Under this view, retrieval times are not affected by the distance between co-dependents, but a higher proportion of retrieval failures occur in subject relatives compared to object relatives. This leads to a mixture distribution in both subject and object relatives, but the proportion of the failure distribution is higher in subject relatives.

In conclusion, this paper serves as a case study demonstrating the flexibility of Bayesian cognitive modelling using finite mixture models. This kind of modelling approach can be used flexibly in many different research problems in cognitive science. One example is the above-mentioned work by Nicenboim and Vasishth (2017). Another example, also from sentence comprehension, is the evidence for feature overwriting (Nairne, 1990) in parsing (Vasishth, Jäger, \& Nicenboim, 2017).

## Acknowledgments

We are very grateful to Ted Gibson for generously providing the raw data and the experimental items from Gibson and Wu (2013). Thanks also go to Lena Jäger for many insightful comments. Helpful observations by Aki Vehtari are also gratefully acknowledged. For partial support of this research, we thank the Volkswagen Foundation through grant 89953.

## References

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716-723.
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1-48. doi: 10.18637/jss.v067.i01
Box, G. E., \& Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), 211-252.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., \& Rubin, D. B. (2014). Bayesian data analysis (Third ed.). Chapman and Hall/CRC.
Gelman, A., Hwang, J., \& Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. Statistics and Computing, 24(6), 997-1016.
Gibson, E. (2000). Dependency locality theory: A distancebased theory of linguistic complexity. In A. Marantz, Y. Miyashita, \& W. O'Neil (Eds.), Image, language, brain: Papers from the first mind articulation project symposium. Cambridge, MA: MIT Press.
Gibson, E., \& Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. Language and Cognitive Processes, 28(1-2), 125-155.
Just, M. A., \& Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99(1), 122-149.
Lee, M. D., \& Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: A practical course. Cambridge University Press.

Lewis, R. L., \& Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. Cognitive Science, 29, 1-45.
McElree, B., Foraker, S., \& Dyer, L. (2003). Memory structures that subserve sentence comprehension. Journal of Memory and Language, 48, 67-91.
Nairne, J. S. (1990). A feature model of immediate memory. Memory \& Cognition, 18(3), 251-269.
Nicenboim, B., Engelmann, F., Suckow, K., \& Vasishth, S. (2016). Number interference in German: Evidence for cuebased retrieval. Retrieved from https://osf.io/mmr7s/ (submitted to Cognitive Science)
Nicenboim, B., \& Vasishth, S. (2017). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. Retrieved from https://arxiv.org/abs/1612.04174 (Under revision following review, Journal of Memory and Language)
Plummer, M. (2012). JAGS version 3.3.0 manual. International Agency for Research on Cancer. Lyon, France.
Safavi, M. S., Husain, S., \& Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates: Implications for expectation and memory-based accounts. Frontiers in Psychology, 7. doi: 10.3389/fpsyg.2016.00403
Stan Development Team. (2016). Stan modeling language users guide and reference manual, version 2.12 [Computer software manual]. Retrieved from http://mc-stan.org/
Vasishth, S., Chen, Z., Li, Q., \& Guo, G. (2013, 10). Processing Chinese relative clauses: Evidence for the subjectrelative advantage. PLoS ONE, 8(10), 1-14.
Vasishth, S., Jäger, L. A., \& Nicenboim, B. (2017). Feature overwriting as a finite mixture process: Evidence from comprehension data. In Proceedings of MathPsych/ICCM. Warwick, UK. Retrieved from https://arxiv.org/abs/1703.04081
Vehtari, A., Gelman, A., \& Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out crossvalidation and WAIC. Statistics and Computing.
Vehtari, A., Ojanen, J., et al. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys, 6, 142-228.

# Interpreting actions by attributing compositional desires 

Joey Velez-Ginorio ${ }^{1}$ (joeyv@mit.edu), Max Siegel ${ }^{1}$ (maxs@mit.edu), Joshua B. Tenenbaum ${ }^{1}$ (jbt@mit.edu), \& Julian Jara-Ettinger ${ }^{2}$ (julian.jara-ettinger@ yale.edu)<br>${ }^{1}$ Department of Brain and Cognitive Sciences, MIT. Cambridge, MA. 02139<br>${ }^{2}$ Department of Psychology, Yale University. New Haven, CT. 06520


#### Abstract

We cannot see others' mental states, so we infer them by watching how people behave. Bayesian inference in a model of rational action - called inverse planning - captures how humans infer desires from observable actions. These models represent desires as simple associations between agents and world states. In this paper we show that by representing desires as probabilistic programs, an inverse planning model can infer complex desires underlying complex behaviors-desires with temporal and logical structure, which can be fulfilled in different ways. Our model, which combines basic desires via logical primitives, is inspired by recent probabilistic grammarbased models of concept learning. Through an experiment where we vary behaviors parametrically, we show that our model predicts with high accuracy how people infer complex desires. Our work sheds light on the representations underlying mental states, and paves the way towards algorithms that can reason about others' minds as we do.


Keywords: social cognition; theory of mind; computational modeling; Bayesian inference.

## Introduction

As social creatures, humans routinely have to make sense of what other people are doing, and we do so by appealing to mental states such as beliefs, desires, and intentions. Because we cannot see these internal mental states we need to infer them by watching how people act.

Research into this capacity, called a Theory of Mind (Gopnik \& Meltzoff, 1997; Dennett, 1989), suggests that mental state inferences are driven by the assumption that agents act efficiently, subject to constraints imposed by their environment (Gergely \& Csibra, 2003). If, for instance, an agent takes a straight path towards a cookie jar, we can guess that her goal is to get a cookie, even before she has reached it. By contrast, if she gets there after wandering around for a while, we may infer that she found it without having deliberately searched.

In such scenarios, it makes sense to equate goals with desires. But in more complex scenarios it is important to distinguish between the two: a one-to-one correspondence between desires and goals is rare. Consider, for instance, if Bob wants to have breakfast. He can do this in several different ways that each require a different plan: he can stay home and prepare breakfast; he could go to the local café near his house; or he could go to a coffee shop that is out of the way. If he chooses to eat at the local café, he can show up and request food. By contrast, if he chooses to cook, he may have to go to the grocery store first and then go to his kitchen, in that order. While Bob is at the grocery store, he may need to
buy coffee and milk, but the order in which he buys them does not matter. Finally, before Bob has had breakfast, many states of the world are rewarding (eating at the café or having a scone at home, for example), but once he eats something, all rewards associated with breakfast disappear.

These examples reveal three key properties of desires. First, desires can often be fulfilled in more than one way. So from an observer's standpoint, goals cannot be equated with desires. Second, desires can have logical and temporal structure: they can be fulfilled in different ways (get tea or coffee), they can break into subgoals (get coffee and milk), and they can have temporal structure (go to the café and then buy a scone). Finally, the logical and temporal structure of desires interacts with the underlying rewards. If Bob is thirsty, then both soda and water are rewarding. But once he's had one of them, the other loses its immediate appeal. If Bob wants to exercise and then bathe before work, he has to do them in that specific order; doing them in the wrong order does not suffice. In other cases, the order does not matter, but the reward is only achieved once all the necessary prerequisites are fulfilled. If Bob likes his coffee with milk, then having coffee and milk together is rewarding, but having only one of them is not.

Computational models of mental-state attribution that successfully explain human mental-state inferences assume a relatively simplistic representation of desires: each desire can only be fulfilled in one way, and it is fulfilled by reaching one and only one physical state of the world (e.g. Baker et al., 2017; Baker et al., 2012). This assumption implicitly blurs together desires, intentions, goals, and physical states of the world. As our examples show, this is overly limiting; people may require conjunctions ( A and B ) or disjunctions of goals (A or B), with temporal properties (A then B).

In this paper we develop a richer representation of desires, and clarify the multiple computational levels that transform desires into actions. To solve the representational challenges, we draw on advances in concept learning that support concepts of unbounded complexity (Piantadosi et al., 2012; Goodman et al., 2008, 2014). To solve the inferential challenges that arise with more sophisticated representations, we draw on advances in mental-state attribution beyond goal inference (Lucas et al., 2014; Jara-Ettinger et al., 2016, under review). In the remainder of the paper, we sketch out the computational framework and we present an experiment testing quantitative predictions of our model.


Figure 1. (a) schematic of the generative model. (b) example of how an expression combines primitives and objects to determine how to satisfy a desire. This expression corresponds to an agent who first wants either coffee and milk, or just tea, and then a scone afterwards. The tree below shows the space of possible intentions that can fulfill the desire.

## Computational model

We take as a starting point the idea that social cognition is supported by a probabilistic generative model that determines how mental states lead to actions (Baker, et al., 2017). We expand on this approach by building a more powerful representation of desires, and how they relate to behavior.

Figure 1a shows the overall schematic of our model. We argued that a realistic model of commonsense psychology should distinguish between desires, goals, intentions, and actions, and our model attempts to do so.

At the top level we place desires, which combine logical (and/or) and temporal (then) primitives with simple goals (such as arriving to certain physical locations). This approach enables us to represent desires that directly map onto a single goal (e.g. "go to get coffee") as well as desires that can be fulfilled in different ways (e.g. "eat breakfast first, and then either get coffee and milk, or alternatively get tea"). This representation is inspired and based on computational models that combine logical primitives with unitary concepts to explain the productivity and compositionality of conceptual knowledge (Piantadosi et al., 2012; Goodman et al., 2008, 2014).

Following Goodman et al. (2008), we model the space of desires with a probabilistic grammar, which builds arbitrarily complex desires by composing simple ones. The grammar implements production rules that recursively conjoin primitives and units to yield desire expressions. We endow the grammar with several primitives - And, Or, and Then but the framework is general. These primitives are motivated by common-sense intuitions, but our primary goal is to develop a framework for compositional desires, not to identify the exact primitives that underlie goal-directed behavior.

To connect desires to actions, we rely on an intermediate representation of intentions (see Jara-Ettinger et al., under review). Given a composite desire, our model derives the space of intentions as the set of all ordered sequences of subgoals that satisfy it. For instance, if an agent desires to get either coffee and milk, or just tea, and then a scone afterwards
(Fig 1b), her space of intentions is \{get tea and then a scone; get milk, coffee, and then a scone; and get coffee, milk, and then a scone $\}$.

To model how the agent selects an intention and transforms it into an action plan, we rely on advances in commonsense psychology that suggest that we interpret other people's behavior through the assumption that they act to maximize their subjective utilities - the difference between the rewards they obtain and the costs they incur (Jara-Ettinger et al., 2016, under review; Lucas et al., 2014). This assumption operates at two levels: given a space of intentions, the agent will choose the one that maximizes her subjective utilities, and given an intention, the agent will attempt to complete it as efficiently as possible (for an agent to maximize utilities, they must also minimize costs).


Figure 2: Examples of the experimental stimuli. (a-b) examples of stimuli that consist of a single event. (c) example of stimuli that consists of two events.

To compute each intention's utility, we rely on planning algorithms developed in the robotics literature (Puterman, 2014) that have been successfully applied to model mentalstate attribution (Baker et al., 2009; 2017): Markov Decision Processes (MDPs). Given a set of states, a set of actions, and an underlying reward function, MDPs allow us to determine the sequence of actions that an agent should take to fulfill her goal as efficiently as possible. By using MDPs, we can compute the expected cost of achieving each goal, and define an intention's utility as the reward gained by fulfilling the desire minus the sum of the costs for achieving each goal in the intention. Given each intention's utility, we assume that agents probabilistically select an intention:

$$
\begin{equation*}
p(I) \propto \exp \left(\frac{U(I)}{\tau}\right) \tag{1}
\end{equation*}
$$



Hypothesis
Figure 3: Detailed results from the experiment. Each plot represents one trial from the experiment. The x-axis shows the model's top three hypotheses and the y-axis shows the z -scored prediction with participant judgments. Blue lines and dots show model predictions and red lines and dots show participant judgments. Vertical bars show $95 \%$ confidence intervals. In each plot, the schematic represents the paths the agent took in the event (see Figure 2 for examples of the actual stimuli).
where $\tau$ is a parameter that captures expectations about the agent's rationality. When $\tau$ is low, the agent invariably selects the intention with the highest utility; as $\tau$ increases, the agent is more likely to choose a suboptimal intention.

Finally, once the agent has selected an intention, we define the action plan as the ordered sequence of goals along with the motor programs that complete each goal (computed through MDPs).

## Inference in the generative model

We have specified a generative model for compositional desires, intentions, and action plans. To recover a desire
given some observed actions, we use Bayesian inference to invert the generative model. Given an observable set of actions $A$, the posterior belief for each underlying desire $D$ is given by:

$$
\begin{equation*}
p(D \mid A) \propto l(A \mid D) p(D) \tag{2}
\end{equation*}
$$

where the prior $p(D)$ is set to favor simpler explanations using a simple penalization for the length of the expression (as in Goodman et al., 2008).

To compute the likelihood, $l(A \mid D)$, we integrate over the space of all possible intentions the agent could have:

$$
\begin{equation*}
l(A \mid D)=\sum_{I \in \text { Intentions }} p(A \mid I) p(I \mid D) \tag{3}
\end{equation*}
$$

Both the probability of the intention given the desire $(p(I \mid D))$, and the probability of the action, given the intention $(p(A \mid I))$ are computed through the assumption that agents act to maximize their utilities-the difference between the subjective reward for fulfilling their desires minus the cost for fulfilling it. This expectation implies that agents are more likely to act efficiently given their intention, but that they are also more likely to select the intention that can fulfill the desires with the overall lowest cost. We enumerate a set of desires using breadth-first-search over the grammar, and then approximate the posterior over that space using Bayesian inference.

## Simplicity prior alternative model

To better understand our model, we developed a simple alternative that uses a deterministic likelihood function, where the probability of a desire generating an action $(p(A \mid D))$ is 1 if the action satisfies the desire and 0 otherwise. This model continues to have much of the power of the full model: it has access to rich representations of desires and the prior over hypotheses creates a preference for simpler explanations. Unlike the main model, this model is insensitive to the intermediate representations of intentions, as it does not account for how the agent chooses the intention that will fulfill their desires.

## Experiment

## Design

To evaluate our model, we designed a simple task where participants watched an agent's behavior across one or two days and were asked to determine their belief that the agent had certain desires (see Figure 2).

## Methods

Participants 33 participants, mean age $(S D)=32.13$ years ( 9.38 years), range $=20-61$ years from the US (as determined by their IP address) were recruited using Amazon's Mechanical Turk Framework.

## Stimuli

Figure 2 shows an example of the stimuli. Stimuli consisted of 19 two-dimensional images of an agent traveling to one or more of three potential static locations. Eight of these trials consisted of a single event and the remaining 11 consisted of two events. The one event trials were built by designing all possible efficient paths agents could take to reach between 1 and 3 of the locations and removing equivalent paths (i.e. identical under a rotation or reflection of the map).

Trials with two events were built by first creating a set including possible efficient paths between 1 and 2 of the


Figure 4. Comparison between our model (simplicity \& efficiency) and the alternative model (prior only). Each dot represents a judgment of a hypothesis for a given trial. The x -axis shows the model's prediction and the y -axis shows participant judgments.
locations, omitting paths between 3 locations to prevent the stimuli set from growing too large. In contrast to the single event set, we keep the equivalent paths, as they become necessary to construct the most primitive desires occurring over two events e.g. (A or B). This creates a base set of 9 paths. To generate the trials with two events, we first split the 9 paths into two classes, one for paths that go to only one location (3 paths) and another for paths that go to two locations ( 6 paths). For each class we compute the cartesian product of itself, and after removing duplicate pairs of stimuli in each class, (e.g. $A, B=B, A$ ), this provided a set of 27 two event trials. From that set, events that violated the principle of rational action were removed ( 10 trials). Additionally, if a trial with repeated events was the reflection or rotation of another trial with two events, it was removed ( 5 trials); e.g. between (A,A) and (C,C), we kept (A,A). Last, trials with two events were removed if only one possible hypothesis could explain the trials ( 2 trials), these trials trials impact our ability to get graded responses on alternative plausible hypotheses (an ideal trial would have more than one plausible explanation, to determine if the model captures the same graded measure humans have for alternatives). For example, if the agent only goes to the farthest location on event 1 and 2 , it's clear the only compatible hypothesis is that the agent wants to go that location. As an exception, we included one of these cases in the final set, just to show that the model was capable of inferring the only plausible hypothesis. After filtering the original 27 two event stimuli, 11 remained. These 11 plus the 8 one event trials result in the 19 stimuli used in the experiment.

## Procedure

Participants first read a tutorial that explained the logic of the task. Participants then completed a short survey that ensured they had read the instructions, and the test phase followed immediately after.

During the test phase, participants completed 19 trials. In each trial participants saw the stimuli on the left side, and they were asked to rate their belief that the agent had each of three different desires. Each desire was rated on a scale from $0-10$ for each, with 0 indicating "Definitely not"; 5 "Maybe";
and 10 "Definitely." The three desires were obtained by selecting the three hypotheses with the highest posterior distribution according to the model. In order to present these hypotheses to participants, we translated the description from the model into descriptions in English. To ensure their accuracy, two coders blind to the original hypotheses backtranslated the descriptions into the model's original representations. The two coders showed $100 \%$ agreement and recovered the correct model hypothesis in all trials.


Figure 5. Detailed results one of the trials. The top left plot shows the schematic of the stimuli we used. The top right plot shows participant judgments (z-scored); the bottom two plots show the predictions of the full model and the alternative model (z-scored). This example illustrates how, by removing the probabilistic nature of the likelihood function, the model loses sensitivity to variability in participant judgments.

## Results

Figure 3 shows the results from the experiment. Qualitatively, our model fit participant judgments well. Our model predictions showed a correlation of $\mathrm{r}=0.92$ with participant judgments ( $95 \%$ CI: 0.86-0.95). See Figure 4. By contrast, the alternative model (prior only) showed a weaker correlation ( $\mathrm{r}=0.80 ; 95 \% \mathrm{CI}: 0.69-0.88$ ). A bootstrap over the correlation difference showed that the full model performed reliably better than the alternative model (correlation difference $=0.11 ; 95 \% \mathrm{CI}: 0.009-0.18$ ).

Figure 5 shows the detailed results of a single trial that illustrates how the alternative model with a deterministic likelihood function fails to capture participant judgments. In this trial the agent begins by going to the top left location (which is one of the closest ones, together with the bottom
right location), and then travels diagonally to the bottom right location. Our full model gives a high probability to the desire that the agent wanted to visit those two locations in that specific order (A then C ), an average probability to the desire that she could have wanted to visit the locations in any order ( A and C ), and a low probability to the desire that the agent wanted to visit either A or B first, then C ((A or B) then C). Although all hypotheses explain the actions, our model is sensitive to the probability that each desire would generate the observed actions relative to competing ways to fulfill the same desire (driving the difference between the first and second hypotheses) and to the baseline complexity of the desires (driving the difference between the second and third hypotheses). That is, our model recognizes that there are two equally good intentions that fulfill the desire "A and C" (A and then C , or C and then A ), but only one that fulfills the ordered desire "A then C " (A and then C ). This makes our model favor the ordered explanation, as participants do (see Figure 5). This is not captured in the prior only model, as it is only sensitive to expression complexity. These results show how people are both sensitive to the likelihood that a desire would generate the observed actions, and to the complexity of the ascribed desire. Figure 6 shows how this failure becomes even stronger in the case where participants watch the agent behave identically across two events.

## Discussion

Here we presented a formal model of action understanding that represents desires as composite entities sampled from a probabilistic context free grammar. Desires get transformed into intentions and then into action plans by the assumption that agents act to maximize their utilities. By performing Bayesian inference over this generative model, we showed how we can capture desires that have rich logical and temporal structure, as well as enabling us to represent desires that can be fulfilled in more than one way. We tested our model by comparing its inferences with those made by human participants, finding that it closely mirrors their judgments, and that an alternative model is less successful.

Our model shows that combinations of primitives and objects using a probabilistic context free grammar supports rich representations of desires in Theory of Mind. The primitives, composing over objects, generate structured desires that capture temporal and logical structure.

Our goal was to develop a more nuanced representation of desires, and the framework we propose works for any arbitrary set of primitives and objects. To test our model, we focused on three specific primitives: And, Or, and Then. Our results do not imply that these are the only primitives people use when they reason about others' desires, or even that they are central in action-understanding. Other primitives such as If, Any, and Not, are likely also at play when we reason about other people's behavior. More research is needed to characterize the primitives we use in action-understanding, and their developmental origins.

To characterize desire complexity, we used a simple prior
that penalized the length of the expression (based on Goodman et al., 2008). Although this is a useful approximation, different primitives may have different priors which capture both their conceptual complexity and the extent to which they are useful in explaining behavior. Future work may attempt to uncover primitive-specific priors and the forces that shape these priors.


Figure 6. Results from the trial where participants watch two repeated events. While the prior only model continues to make the same predictions, both participants and our model have a stronger belief that the order mattered, in comparison to the trial with a single event (Figure 5)

In our current work, we focused specifically on desires and we assumed that the agents had full knowledge about the environment. In more realistic cases, agents can be uncertain, ignorant, or wrong about the world, and people's reasoning about others is sensitive to this fact (Baker et al., 2017; Kovács, Téglás, \& Endress, 2010). Our grammatical approach to desires may also support more structured representations about beliefs. Intuitively, people's beliefs are often structured logically (e.g. my laptop is in my backpack or at home; she thinks he is hungry and tired). In future work we will investigate the power and limitations of applying this approach to the representations of beliefs, and to the interaction of beliefs and desires.

Although in our work we focused on these representations as applying to desires, these desires often inherit their structure from how the world works. If Bob wants to shoot a water gun, he needs to pour water into the tank first, then pump air into valve, and then press the trigger, in that order. The fact that Bob's desire takes this structure is a reflection of how water guns work. This opens the possibility that, through the ability to reason about other people's desires, we may simultaneously learn procedural knowledge about how to make changes to the world. As such, our model may shed
light on how we learn about the world by watching more competent agents (see also Jara-Ettinger, Baker \& Tenenbaum, 2012).

## Acknowledgments

This work was supported by the Simons Center for the Social Brain. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

## References

Baker, C. L., Saxe, R., \& Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3).
Baker, C. L., Jara-Ettinger J., Saxe, R., \& Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing. Nature Human Behavior
Dennett, D. C. (1989). The intentional stance. MIT press.
Gergely, G., \& Csibra, G. (2003). Teleological reasoning in infancy: The nave theory of rational action. Trends in cognitive sciences, 7(7), 287-292.
Goodman, N. D., Tenenbaum, J. B., Feldman, J., \& Griffiths, T. L. (2008). A Rational Analysis of Rule-Based Concept Learning. Cognitive Science, 32(1), 108-154.
Goodman, N. D., Tenenbaum, J. B., \& Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. Center for Brains, Minds and Machines (CBMM).
Gopnik, A., Meltzoff, A. N., \& Bryant, P. (1997). Words, thoughts, and theories (Vol. 1). Cambridge, MA: Mit Press.
Jara-Ettinger, J., Schulz, L. E., \& Tenenbaum, J. B. (under review). The naïve utility calculus as a rational, quantitative foundation of action understanding.
Jara-Ettinger, J., Baker, C. L., \& Tenenbaum, J. B. (2012). Learning What is Where from Social Observations. In CogSci.
Jara-Ettinger, J., Gweon, H., Schulz, L. E., \& Tenenbaum, J. B. (2016). The naïve utility calculus: computational principles underlying commonsense psychology. Trends in cognitive sciences, 20(8), 589-604.
Kovács, Á. M., Téglás, E., \& Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. Science, 330(6012), 1830-1834.
Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... \& Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. PloS one, 9(3), e92160.
Piantadosi, S. T., Tenenbaum, J. B., \& Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. Cognition, 123(2).
Puterman, M. L. (2014). Markov decision processes: discrete stochastic dynamic programming. John Wiley \& Sons.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279-1285.

# Non-Symbolic Exact Quantity Representation in a Language Impaired Population 

John Verbos (verbosj@duq.edu)<br>Department of Psychology, 600 Forbes Avenue<br>Pittsburgh, PA 15282

Sarah E. Wallace (wallaces@duq.edu)<br>Department of Speech-Language Pathology, 600 Forbes Avenue<br>Pittsburgh, PA 15282

Alexander Kranjec (kranjeca@duq.edu)<br>Department of Psychology, 600 Forbes Avenue Pittsburgh, PA 15282<br>Center for the Neural Basis of Cognition, 4400 Fifth Avenue, Suite 115<br>Pittsburgh, PA, 15213


#### Abstract

English-speakers whose access to number language is artificially compromised by verbal interference and the Pirahã (an Amazonian tribe without exact number words) appear to rely on analog magnitude estimation for representing nonsymbolic exact quantities greater than 3 . Here, 16 participants with aphasia performed the 5 counting tasks from these previous studies. Performance was poorest when targets were not visible during response ( $70 \%$ correct, task $4 ; 71 \%$ correct, task 5) and best when targets were presented as subitizable groups of 2 and 3 ( $98 \%$ correct, task 2). Western Aphasia Battery-Revised subtest scores correlated with task performance, suggesting diverse forms of language impairment may contribute to errors. Coefficients of variation for tasks and significant correlations of target magnitude with error rate ( $r^{2}=.88$ ) and error size ( $r^{2}=.87$ ) across tasks suggest participant use of analog magnitude estimation. Experiments involving people with aphasia may further refine our understanding of how language and thought interact.


Keywords: aphasia, language, number

## Introduction

"Linguistic relativity" occupies the broad theoretical middle ground where language and cognition interact, where the grammatical structures and lexicons of a language are believed to influence thought to a greater or lesser degree. While the idea that language can influence thought, perception, and action has a long history in Western philosophy, Whorf (1956) provided the first and clearest articulation of a strong version of this position. According to linguistic relativity, words aren't just names for pre-existing concepts; thought is influenced by the way particular languages are structured, what languages have words for, and what they don't. When a language is transmitted from one generation to the next, so are particular ways of "cutting up" the world that come with speaking that language.

Everett (2013) compiles a diverse array of recent research that explores domains like space, time, quantity, gender, and color and draws positive conclusions about the effects of language on thought. Similarly, Frank, Fedorenko, Lai,

Saxe, and Gibson (2012), review several studies that find "meaningful cognitive differences" ( p .75 ) between speakers of languages that have words for particular concepts and those that don't. Such cognitive differences appear to exist both across cultures and across development. At the same time, experimentally manipulated verbal interference can temporarily remove differences otherwise present.

The domain of number is a good entry point for testing the linguistic relativity hypothesis. Numeracy develops alongside language in humans, and there are clear differences between the ways adult speakers of different languages perform number-related tasks. The Pirahã, an indigenous Amazonian tribe, are of particular interest here, as their language lacks words for exact number. Gordon (2004) engaged seven Pirahã tribe members in a series of nonverbal matching tasks where participants were asked to reproduce a visual array that matched a model. The Pirahã struggled to accurately reproduce any set of objects containing more than three items, even when the model was visible to copy. Gordon (2004) also noted Pirahã responses produced a coefficient of variation (CoV) of approximately 0.15 , congruent with evidence that without access to number language and counting, people use less accurate but inborn abilities to estimate quantities larger than three ${ }^{1}$.

Frank, Everett, Fedorenko, and Gibson (2008) replicated the tasks from Gordon (2004) with fourteen participants in a different Pirahã village. The authors found similar results for each task with the exception of the one-to-one matching task, where results were near ceiling. Consequently, Frank et al. (2008) concluded that some of the startling results of Gordon (2004) might be the product of participants not understanding the task or inconsistencies in the experiment.

Everett and Madora (2012) sought to resolve the conflicting results of Gordon (2004) and Frank et al. (2008). The authors recreated the three tasks from Frank et al.

[^231](2008) with fourteen participants in a third Pirahã village. With one exception, Everett and Madora (2012) found no significant differences when making intra- or inter-study task comparisons. The exception was the one-to-one matching task from Frank et al. (2008), which was significantly different from control tasks and the Everett and Madora (2012) one-to-one matching task. The CoVs for all tasks in Everett and Madora (2012) were 0.15, consistent with Gordon (2004) and the hypothesis that the Pirahã were employing analog estimation strategies.

Frank et al. (2012) extends the experimental tasks performed with the Pirahã to a numerate population by using verbal interference in an attempt to force participants to resort to analog magnitude estimation (Whalen et al., 1999). The authors hypothesized that if language is not crucial to establishing exact number, then participants should successfully perform non-verbal number tasks under verbal interference. Should language be necessary for exact numeracy, however, these same participants should fall back on analog magnitude estimation under verbal interference revealing a constant CoV , as seen in other studies. To test this, Frank et al. (2012) had thirty-five MIT students attempt the matching tasks performed with the Pirahã while simultaneously repeating radio news broadcasts aloud. The results of these experiments were then compared to each other and to the results of the same experiments with the Pirahã from Frank et al. (2008).

While the English-speakers were found to be more accurate than the Pirahã, both groups made "significant and systematic errors" (p. 79) on the "nuts-in-a-can" task (see Figure 1 below), where participants have no access to a direct or remembered visual representation of the array. Here, college students under verbal interference, like the Pirahã, produced a flat CoV of 0.15 across targets, suggesting the use of analog magnitude estimation. Frank et al. (2012) drew the conclusion that the concept of "exact match" does not require language, but that language is crucial to storing and manipulating exact quantities greater than three. This conclusion is in line with the language as a technology or tool-kit version of the linguistic relativity hypothesis, wherein language allows us to transcend our pre-linguistic cognitive capacities (Gentner \& GoldinMeadow, 2003).

The evidence to date strongly suggests that language for number has a significant influence on how quickly and accurately we comprehend and process quantities larger than three. At the same time, there is room for debate as to how fundamental number language is to the correct apprehension of exact quantity. One largely unexplored route to an understanding of the relationship between language and counting (and more generally, questions regarding linguistic relativity) involves studying people with organic language impairments. People with focal brain lesions-either as a result of infarcts, tumor resections or other restricted lesions-may acquire aphasia, an impairment of a person's ability to comprehend and formulate language across multiple modalities, including
speaking, reading, writing, and listening (Rosenbek, LaPointe, \& Wertz, 1989). Consequently, people with aphasia may experience difficulty in the use of language for number and calculation (Dragoy, Akinina, \& Dronkers, 2016). McNeil and Pratt (2001) specify that aphasia is a processing or performance disorder-that is, a problem in using language for a known concept. By this reasoning, if aphasia were to affect a person's ability to represent exact quantity on a non-symbolic task such as the one employed in the current study, it may work in a similar fashion to verbal interference-by disrupting access to a number concept and consequently impairing comprehension or speech in relation to that concept. However, it is conceivable that aphasia may impair some individuals' ability to represent exact quantity in a manner more like the Pirahã, who have no exact number language to employ. In such a scenario, a person with aphasia may be impaired because they have no stored verbal label for exact quantity available for access. While the current study cannot adjudicate between these possibilities, we hope the diversity of impairment within the present aphasia population may provide a window into qualitative differences that account for errors across the kinds of tasks used with the Pirahã. We also hope to suggest ways that aphasia populations may generally contribute to investigations of the linguistic relativity hypothesis.

While several case studies have examined the impact of aphasia on calculation-e.g., Dragoy et al. (2016), where 7 of 10 participants with aphasia struggled with basic arithmetic and when comparing Arabic representations of quantities-little research to date has examined the impact of language impairment on non-symbolic representation of quantity. Lemer, Dehaene, Spelke, and Cohen (2003) examined a person with acalculia due to a focal lesion of the left parietal lobe and another person with semantic dementia from predominantly left temporal hypometabolism to demonstrate dissociations between tasks associated with counting and those associated with innate quantity systems of number processing. As predicted by a lesion in the parietal lobe, the patient with acalculia showed a severe slowness in approximation, and exhibited impairments in subitizing and numerical comparison tasks. Meanwhile, the patient with semantic dementia had intact approximation abilities and showed preserved processing of non-symbolic small numbers-that is, her "quantity processing" systems were functioning as expected-but struggled with tasks that required intact verbal processing and counting. Given these findings and related results with other populations, language impairment in the form of aphasia may be predicted to negatively affect the individual's ability to produce nonverbal and non-symbolic representations of exact quantity.

In the current study, participants with aphasia performed the same set of five, increasingly complex matching tasks used with the Pirahã and English-speakers whose access to language was artificially compromised by verbal interference (Frank et al., 2012). It bears noting that unlike the previously studied groups, a clinical aphasia population
consists of individuals with a diversity of verbal and nonverbal impairments. Regardless, we hypothesize that participants will make more frequent and larger errors (1) in proportion to target size; (2) on each subsequent, more difficult, task; and (3) produce a flat coefficient of variation (CoV) on each task and across target quantities, suggesting reliance on the analog magnitude system to estimate quantity. Such results would lend further support to the hypothesis that access to language for exact number is necessary for the recognition and representation of exact quantities. While general severity of language impairment is predicted to correlate with performance across tasks, we are also interested in whether particular aspects of language impairment point to specific qualities of language involved in counting and exact quantity representation.

Results suggesting that aphasia limits a person's ability to represent non-symbolic exact quantities would complement the body of evidence demonstrating a relationship between exact number language and the ability to perform nonsymbolic exact quantity tasks. When taken alongside similar evidence from previous studies with different human populations-i.e. children raised in numerate cultures but who have yet to develop number-language skills (e.g., Condry \& Spelke, 2008), adults in numerate cultures under verbal interference, and adults in an anumeric culture-it would seem difficult not to conclude that access to exact number language has an effect on the way that humans think about numbers. More broadly, these findings may refine hypotheses generated by linguistic relativity with regard to the necessity and/or effective use of language in representing basic number concepts. The linguistic diversity present within the present clinical aphasia population may provide deeper insight into relations between particular aspects of language function and the representation of exact quantity.

## Methods

Sixteen participants (3 female) completed aphasia assessments and the set of five non-verbal and non-symbolic exact quantity representation tasks from Everett and Madora (2012) and Frank et al. (2012). Thirteen participants also completed a numeral elicitation task, confrontation naming task, and free counting task. Eight completed tests of nonverbal semantic processing and short-term memory-the Semantic Category Probe (Freedman \& Martin, 2001), and Pyramids and Palm Trees tests (Howard \& Patterson, 1992). All participants had aphasia resulting from a lefthemisphere stroke as determined by their score on the Aphasia Quotient (AQ) portion of the Western Aphasia Battery-Revised (WAB-R) (Kertesz, 2006) and a speechlanguage pathologist. Within this framework, 5 participants are considered to have Broca's aphasia, 6 Anomic aphasia, 2 Wernicke's aphasia, 2 conduction aphasia, and 1 global aphasia. Eligible participants were a minimum of six months post onset of aphasia ( $M=73$ months, $R=9-159$ months), between the ages of eighteen and eighty-five years ( $\mathrm{M}=61, \mathrm{R}=43-75$ ) and native English speakers.

Aphasia assessment. Participants completed the AQ portion of the WAB-R (Kertesz, 2006). This formal assessment includes tasks such as answering simple questions, describing pictures, manipulating and naming common objects, following directions, repeating words, and matching pictures to printed words and sentences.

Matching tasks (Everett \& Madora, 2012; Frank et al., 2012). Participants completed five non-verbal and nonsymbolic exact quantity representation tasks in the following order: a one-to-one matching task, an uneven matching task, an orthogonal matching task, a hidden matching task, and a "nuts-in-a-can" task (see Figure 1). In every task, the experimenter presented a quantity of spools of thread (approximately 1 " tall, $3 / 4$ " in diameter) and asked the participant to construct a row of un-inflated balloons (approximately $4 "$ long and $2 "$ wide) that matches the number of spools of thread. In the one-to-one task, the experimenter placed the spools one at a time in an evenly spaced line from left to right. In the uneven task, the spools were presented in the same manner as in the one-to-one task, but broken randomly into smaller groups of two and three. The orthogonal task is identical to the one-to-one task except that the row of spools is presented in a line perpendicular to the participant. The hidden matching task is identical to the one-to-one task except that the row of spools is hidden from the participant after being presented. In the "nuts-in-a-can" task, the experimenter places spools one by one into an opaque cup. Participants were tested once per task on each quantity from four to twelve in one of two random orders, totaling forty-five trials per participant.


Figure 1: Schematic of each matching task. From left to right: one-to-one match (task 1), uneven match (task 2), orthogonal match (task 3), hidden match (task 4), "nuts-in-a-can" (task 5). Image is from Frank et al. (2012).

Numeral elicitation task. Participants were asked to name the number of spools of thread presented, increasing from one to twelve and then decreasing from twelve to one. In each case, participants were asked, "How many spools of thread are there?" by the researcher. Divergence between performance on this task and on the matching tasks might illuminate whether the participant is having difficulty recognizing, articulating, or representing the target quantity.

Confrontation naming task. Participants were asked to name the Arabic numerals one through twenty as presented individually on flashcards. In each case, participants were asked, "What number is this?" This task assessed the participant's ability to recognize and name Arabic numerals. Confluent or divergent performance on this task when
compared to the matching and counting tasks might help differentiate the participant's ability to recognize and name symbolic and non-symbolic numbers.

Free counting task. Participants were asked to count up from one to twenty and down from twenty to one. The researcher says, "Please count from one up to twenty" and "Please count from twenty down to one." Participants were allowed five minutes to recite each count list. Performance on this task indicates the participant's capacity to access and articulate counting numbers in order, a factor in the participant's performance on the matching tasks.

Semantic Category Probe Test (Freedman \& Martin, 2001). Participants listened to a list of three or more words and determined whether the final word is from the same category as any of the preceding words by saying or pointing to "Yes" or "No." This task assesses the participant's capacity to retain semantic information in their short-term memory, where impairment might be a potential reason for poorer performance on the matching tasks.

Pyramids and Palm Trees Test (Howard \& Patterson, 1992). Participants matched a pictured item to the closest associate among a set of two pictured choices (e.g., fish matched to: cat, table). This task assessed the participant's capacity to process non-verbal semantic information. Distinguishing between semantic and verbal impairments may help explain performance on the matching tasks.

## Results

There was notable variation across participants and tasks. Percent correct scores for all tasks ranged from $53 \%$ to $98 \%$ (Table 1). Participants responded correctly on $83 \%$ of task 1 trials, $98 \%$ of task 2 trials, $90 \%$ of task 3 trials, $70 \%$ of task 4 trials, and $71 \%$ of task 5 trials (Fig. 2, far left).

Participants' accuracy descreased as the target quantity increased across all tasks ( $r^{2}=0.87$ ) (Fig. 2, center left) and for each individual task (Fig. 3, top row). Similarly, error magnitude increased as target quantity increased ( $r^{2}=0.88$ ) (Fig. 2, center right). CoV was similar across target
quantities and tasks (Fig. 2, far right), but higher on task 4 (0.10) and task 5 (0.11) (Fig. 3, bottom row). Across analyses, aphasia participants' performance was remarkably similar to the performance of English speakers under verbal interference from Frank et al. (2012) (Figs. 2 and 3). Compared to the Pirahã (Figs. 2 and 3, aggregated from Everett \& Madora, 2012; Frank et al., 2008; and Gordon, 2004), participants with aphasia and English speakers under verbal interference were generally more accurate and made smaller errors, but all three groups showed similar patterns of responding across tasks.

|  | Task |  |  |  |  | Total | $\begin{gathered} \% \\ \text { Correct } \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1 | 2 | 3 | 4 | 5 |  |  |
| 2 | 0 | 0 | 0 | 1 | 0 | 1 | 97.8 |
| 16 | 0 | 0 | 0 | 1 | 1 | 2 | 95.6 |
| 7 | 0 | 0 | 0 | 1 | 1 | 2 | 95.6 |
| 1 | 0 | 0 | 0 | 1 | 1 | 2 | 95.6 |
| 10 | 0 | 0 | 0 | 1 | 1 | 2 | 95.6 |
| 12 | 1 | 0 | 0 | 1 | 1 | 3 | 93.3 |
| 光 13 | 0 | 1 | 3 | 1 | 2 | 7 | 84.4 |
| . 11 | 0 | 0 | 0 | 5 | 2 | 7 | 84.4 |
| 15 | 0 | 0 | 0 | 4 | 5 | 9 | 80.0 |
| $\approx 3$ | 2 | 0 | 1 | 3 | 4 | 10 | 77.8 |
| 9 | 1 | 0 | 2 | 2 | 5 | 10 | 77.8 |
| 6 | 4 | 0 | 2 | 2 | 2 | 10 | 77.8 |
| 14 | 2 | 0 | 2 | 4 | 4 | 12 | 73.3 |
| 4 | 5 | 1 | 0 | 3 | 4 | 13 | 71.1 |
| 8 | 5 | 0 | 2 | 6 | 3 | 16 | 64.4 |
| 5 | 4 | 1 | 3 | 7 | 6 | 21 | 53.3 |

Table 1: Participant errors across tasks. The maximum number of errors on each task is nine. Darker colors indicate more errors.

WAB-R AQ and subtest scores were reliably correlated with task performance on tasks 4 and 5. AQ and subtest scores were most predictive of performance on task 5, the "nuts-in-a-can" task (Table 2).

Thirteen participants completed additional number tasks. While, generally speaking, participants with higher AQ scores who had made fewer errors on the nonverbal matching tasks also performed better on the additional



Figure 2: Matching task summary data for participants with aphasia, Pirahã, and adults under verbal interference. Far left: For participants with aphasia, performance was poorest when targets were not visible during response ( $70 \%$ correct, task $4 ; 71 \%$ correct, task 5 ) and best when targets were presented as subitizable groups of 2 and 3 ( $98 \%$ correct, task 2 ). Center left: Significant correlations were found between target magnitude with both error rate ( $r^{2}=.87$ ) and error size ( $r^{2}=.88$ ) (Center right) across tasks. Far right: Coefficients of variation for participants with aphasia mirrored those of adults under verbal interference. "Pirahã" data is from Everett and Madora (2012); Frank et al. (2008); and Gordon (2004). "Verbal Interference" data is from Frank et al. (2012).
counting tasks, there were exceptions. Participant 13, who has a high AQ score and made no errors on the additional number tasks made seven errors across matching tasks. Participant 11 made as many matching task errors as Participant 13 (refer to Table 1), but scored only 4 of 12 on the numeral elicitation task. Additionally, Participant 11, despite correctly reciting 18 of 20 numbers on the ascending free counting task, could not count backwards from 20 to 1 , receiving a score of zero on the descending free counting task. Across all 8 participants who completed the nonverbal semantic processing and short-term memory tasks, higher AQ scores predicted better performance on the Pyramids and Palm Trees and Semantic Category Probe tests.

|  |  | Task |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 1 | 2 | 3 | 4 | 5 |
|  | AQ | 0.41 | -0.02 | 0.23 | 0.61 | 0.77 |
|  | Speech | 0.30 | -0.04 | 0.18 | 0.62 | 0.74 |
|  | Comprehension | 0.42 | 0.04 | 0.10 | 0.38 | 0.69 |
|  | Repetition | 0.38 | -0.01 | 0.16 | 0.50 | 0.68 |
|  | Naming | 0.49 | -0.10 | 0.33 | 0.66 | 0.72 |

Table 2: Correlations between task performance and WAB-R subtest scores. AQ = Aphasia Quotient, Speech = Spontaneous Speech, Comprehension $=$ Auditory Comprehension, Naming $=$ Naming and Word Finding. Darker colors indicate larger $r$-values.

## Discussion

Generally, participants (1) made more errors for larger target quantities, (2) made errors of greater magnitude for larger target quantities, and (3) had more difficulty with tasks where targets were not visible during response. There was consistency among those participants with the greatest overall task impairments. Participants who made ten or more incorrect responses also made errors across tasks 1, 3, 4, and 5. Eight different participants responded incorrectly
to at least one trial of task 1, where the target remained visible and did not require conservation in space or time, nor, presumably, counting: correct responding only required participants to match one object to another. The results of task 1 stand in stark contrast to near-ceiling results on task 2. In task 2, targets were presented in groups of 2 and 3. This is the only difference between tasks 1 and 2 , suggesting that many participants were able to subitize the visible targets on task 2 in order to answer accurately, but were unable to do so consistently on task 1 . Near-ceiling performance on task 2 also suggests that perceptual and/or attentional impairments (e.g., field cuts, neglect) do not explain poor performance on tasks $1,3,4$, and 5 ; this represents an important control condition in a stroke population with expected neurological and behavioral heterogeneity. Surprisingly, performance on task 3 was superior to performance on task 1, despite the required spatial translation between the perpendicular target array and horizontal response. Participants responded incorrectly on $10 \%$ of task 1,2 , and 3 trials, where the target remained visible for comparison, matching, and recounting. Performance on tasks 4 and 5 was poorer, as expected: both involve responding without the target array still visible.

These results mirror those of previous studies with the Pirahã and adults under verbal interference, although the Pirahã made more frequent and larger errors, more clearly suggesting a reliance on analog magnitude estimation in attempting to represent target quantities. Of all the research of this kind conducted with the Pirahã, only the one-to-one matching task in Frank et al. (2008) produced a CoV markedly different from 0.15 . Everett and Madora (2012) offered a speculative explanation: unlike the others, the village tested in Frank et al. (2008) had been exposed to math tutoring that included neologisms for number words. It is the neologisms for number words that are exceptionalall the villages had been exposed to the one-to-one matching
Accuracy: one-to-one

Figure 3: Task accuracy and CoV for participants with aphasia, Pirahã, and adults under verbal interference. Accuracy (Top row) and CoV (Bottom row) for participants with aphasia mirrored those of adults under verbal interference. "Pirahã" data is from Everett and Madora (2012); Frank et al. (2008); and Gordon (2004). "Verbal Interference" data is from Frank et al. (2012).
task and other attempts at basic math training by the Brazilian government, but only the site of Frank et al. (2008) had been exposed to number word neologisms. The authors are clear that this is speculation on their part, but it dovetails with a possible explanation as to the task performance differences between the Pirahã on the one hand and the verbal interference and aphasia participants on the other. In attempting to account for the lower CoVs and greater accuracy of the verbal interference participants, Frank et al. (2012) suggests that participants' "differential cultural experience with mathematics and other uses of exact numerosity led to their relatively more precise representation of analog magnitude" (p. 82). The same could be suggested of the aphasia participants in this study.

Certainly there are differences between the current population of people with aphasia, people of an anumeric culture, and English-speakers under verbal interference. What separates the Pirahã from other populations under discussion here is that they exist in a world without exactquantity language and may not have a concept of number to access. English speakers under verbal interference, meanwhile, are members of a numeric culture who have had their ability to use language temporarily disrupted, and people who have aphasia are members of the same culture with a more permanent disruption. Also, an aphasia population consists of individuals with distinct lesions, resulting in a range of verbal and nonverbal impairments and significant heterogeneity is to be expected, compared to a population of English speakers undergoing experimental manipulation via verbal interference. While diversity within the current aphasia population is viewed as a potentially rich source for identifying particular aspects of language (e.g., comprehension, speech) that may uniquely affect particular aspects of number use (e.g., mental representation of exact quantity, counting), it also suggests caution before drawing definitive conclusions based on group performance.

That several studies have repeatedly found similar results despite population differences lends support to established ways of thinking about number, thought, and language. According to the model put forth by Feigenson, Dehaene, and Spelke (2004), we are born with two systems for the cognitive representation of number-a parallelindividuation system that can track up to three or four discrete objects and an analog magnitude estimation system we use to approximate large quantities. While these cognitive systems are also found in other animals, humans appear to use exact number words as tools that enhance our capacity to do things with quantities by bridging these systems. The results of the present and previous studies fit this model: language impairment, like verbal interference and living in a culture without exact number words, makes it difficult, if not impossible, for individuals to bridge the two systems for cognitively representing quantities. The present study also suggests that experiments involving people with aphasia may serve to further refine our understanding of how language and thought interact.

## Acknowledgments

The authors would like to thank Michael C. Frank and Peter Gordon for sharing raw data from Frank et al. (2008), Frank et al. (2012), and Gordon (2004), respectively.

## References

Condry, K., \& Spelke, E. (2008). The development of language and abstract concepts: The case of natural number. Journal of Experimental Psychology: General, 137(1), 22-38.
Dragoy, O., Akinina, Y., \& Dronkers, N. (2016). Toward a functional neuroanatomy of semantic aphasia: A history and ten new cases, Cortex, 1-19.
Everett, C. (2013). Linguistic relativity: Evidence across languages and cognitive domains. Berlin: De Gruyter Mouton.
Everett, C., \& Madora, K. (2012). Quantity recognition among speakers of an anumeric language. Cognitive Science, 36(1), 130-141.
Feigenson, L., Dehaene, S., \& Spelke, E. (2004). Core systems of number. Trends in Cognitive Sciences, 8(7), 307-314.
Frank, M., Fedorenko, E., Lai, P., Saxe, R., \& Gibson, E. (2012). Verbal interference suppresses exact numerical representation. Cognitive Psychology, 64(1-2), 74-92.
Frank, M., Everett, D., Fedorenko, E., \& Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. Cognition, 108(3), 819-824.
Freedman, M., \& Martin, R. (2001). Dissociable components of short-term memory and their relation to long-term learning. Cognitive Neuropsychology 18(3), 193-226.
Gentner, D., \& Goldin-Meadow, S. (Eds.). (2003). Language in mind: Advances in the study of language and thought. Cambridge, MA: MIT Press.
Gordon, P. (2004). Numerical Cognition Without Words: Evidence from Amazonia. Science, 306(5695), 496-499.
Howard, D., Patterson, K. (1992). Pyramids and palm trees: A test of semantic access from pictures and words. Bury St. Edmunds, UK: Thames Valley Test Company.
Kertesz, A. (2006). Western Aphasia Battery-Revised. New York, NY: Pearson.
Lemer, C., Dehaene, S., Spelke, E., \& Cohen, L. (2003). Approximate quantities and exact number words: Dissociable systems. Neuropsychologia, 41(14), 19421958.

McNeil, M. \& Pratt, S. (2001). Defining aphasia: Some theoretical and clinical implications of operating from a formal definition. Aphasiology, 15(10/11), 901-911.
Rosenbek, J., LaPointe, L., \& Wertz, R. (1989). Aphasia: A clinical approach. Boston, MA: Little, Brown \& Co.
Whalen, J., Gallistel, C., \& Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. Psychological Science, 10(2), 130-137.
Whorf, B. (1956). Language, thought and reality: Selected writings of Benjamin Lee Whorf (J.B. Carroll, Ed.). Cambridge, MA: MIT Press.

# Audiovisual integration is affected by performing a task jointly 

Basil Wahn (bwahn@uos.de)<br>Institute of Cognitive Science - Neurobiopsychology, University of Osnabrück, Wachsbleiche 27, 49090 Osnabrück, Germany<br>Ashima Keshava (akeshava@uos.de)<br>Institute of Cognitive Science - Neurobiopsychology, University of Osnabrück, Wachsbleiche 27, 49090 Osnabrück, Germany<br>Scott Sinnett (ssinnett@hawaii.edu)<br>Department of Psychology, University of Hawai'i at Mānoa, 2530 Dole Street, Honolulu, HI 96822-2294, USA

Alan Kingstone (alan.kingstone@ubc.ca)
Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia, Canada V6T1Z4

Peter König (pkoenig@uos.de)
Institute of Cognitive Science - Neurobiopsychology, University of Osnabrück, Wachsbleiche 27, 49090 Osnabrück, Germany
Institut für Neurophysiologie und Pathophysiologie, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany


#### Abstract

Humans constantly receive sensory input from several sensory modalities. Via the process of multisensory integration, this input is often integrated into a unitary percept. Researchers have investigated several factors that could affect the process of multisensory integration. However, in this field of research, social factors (i.e., whether a task is performed alone or jointly) have been widely neglected. Using an audiovisual crossmodal congruency task we investigated whether social factors affect audiovisual integration. Pairs of participants received congruent or incongruent audiovisual stimuli and were required to indicate the elevation of these stimuli. We found that the reaction time cost of responding to incongruent stimuli (relative to congruent stimuli) was reduced significantly when participants performed the task jointly compared to when they performed the task alone. These results extend earlier findings on visuotactile integration by showing that audiovisual integration is also affected by social factors.


Keywords: multisensory integration; joint action; task distribution; social cognition.

## Introduction

In everyday life, humans constantly process sensory input from several sensory modalities. If sensory input from multiple sensory modalities coincides in space and/or time, it is frequently integrated into a unitary percept (Alais \& Burr, 2004; Ernst \& Banks, 2002; Körding et al. 2007; Rohe \& Noppeney, 2015; for a review, see: Spence, 2007) - a process referred to as "multisensory integration". Multisensory integration can result in perceptual benefits as
well as costs. In particular, if the multisensory inputs contain redundant information (e.g., visual and auditory stimuli originate from the same spatial location), human localization performance is faster and more accurate (e.g., Körding et al. 2007; Rohe \& Noppeney, 2015; Wahn \& König 2015a,b; 2016). Yet, if the sensory inputs provide conflicting information (e.g., visual and auditory stimuli originate from different spatial locations but still coincide in time), human localization performance is slowed down and less accurate (Heed, Boukje, Sebanz, \& Knoblich, 2010; Plöchl et al., 2016; Rohe \& Noppeney, 2015; Spence, Pavani, \& Driver, 2004). In the past, researchers have explored how attentional processes influence the multisensory integration process, and more generally, how attentional processing is distributed across the sensory modalities (e.g., Alais, Morrone, \& Burr, 2006; Alsius, Navarra, Campbell, \& Soto-Faraco, 2005; Helbig \& Ernst, 2008; Wahn \& König, 2015a,b, 2016; Wahn, Murali, Sinnett, \& König, 2017; for recent reviews, see Talsma, 2015; Wahn \& König, 2017). However, to date, researchers have largely neglected how social factors could affect the integration process. Thus we know relatively little about how the social presence of another person, and/or how performing a task with another person, influences the process of multisensory integration.

To date, to the best of our knowledge, only two studies (Heed et al., 2010; Teneggi, Canzoneri, di Pellegrino, \& Serino, 2013) have addressed the extent to which social factors can modulate multisensory integration. In Heed et al.'s experiment participants performed a visuotactile congruency task that was either performed alone, or jointly
with another person. When the task was performed alone, participants were required to hold two foam cubes, one in each hand, and indicate with foot pedal presses the spatial elevation of a tactile stimulus that could either appear at the top of the cube (i.e., felt at the index finger) or at the bottom of the cube (i.e., felt at the thumb). The participants also simultaneously received irrelevant visual stimuli that either appeared at the same spatial location as the tactile stimulus or not (i.e., stimuli were presented either in congruent or incongruent positions). Thus, the visual stimuli provided either conflicting or redundant spatial information, resulting in costs or benefits of multisensory integration, respectively. Heed et al. (2010) replicated earlier results (Spence, Pavani, \& Driver, 1998; 2004) by finding that reaction times were faster when indicating the location of the tactile target if the visual stimulus appeared in a congruent position compared to an incongruent position. This effect is referred to as the "crossmodal congruency effect" (CCE). That is, when the tactile and visual stimuli provide redundant information (i.e., are presented in the same (congruent) position), localization performance is faster compared to when conflicting information is provided (i.e., stimuli are presented in different, i.e. incongruent positions). When participants performed the task in pairs, one of them indicated the elevation of the tactile stimuli (as before) while the second participant indicated the elevation of the visual stimuli. But note, the person detecting tactile stimuli was still exposed to the congruent or incongruent visual stimuli. Heed et al. found that the magnitude of the CCE was reduced when performing the crossmodal congruency task jointly compared to performing it alone. In particular, when participants performed the task jointly, incongruent presentations had less of an effect on reaction times when compared to performing the task alone. This observation suggests that the cost of incongruent presentations on multisensory integration is reduced when the task is performed jointly.

To date, the modulation of the CCE by social factors as found by Heed et al. (2010) has not been investigated with other sensory modalities. In particular, it is an open question whether audiovisual integration is similarly affected by social factors. Given that the tactile sensory modality processes events in close proximity while the auditory sensory modality is also able to sense more distal events, it is not clear whether audiovisual integration would be similarly affected by social factors as visuotactile integration. Thus rather than the visuotactile congruency task as used in Heed et al. the present study required participants to perform an audiovisual congruency task, either alone or jointly. If social factors modulate the CCE for audiovisual stimuli, we predict that the CCE will be reduced when performing the audiovisual congruency task jointly as compared to performing the task alone. Conversely, if social factors do not affect the CCE, then the CCE should not be modulated regardless of whether the task is performed jointly or alone.

## Methods

## Participants

Twelve pairs of individuals ( 15 female, $M=21.92$ years, $S D=3.35$ years) participated in the study at the University of Osnabrück. Prior to the experiment, participants signed informed written consent. The study was approved by the ethics committee of the University of Osnabrück. After the experiment had been completed, participants were debriefed and received monetary compensation or participation hours.

## Experimental setup

Participants sat in a dark room in front of a computer screen (Apple 30" LCD screen, resolution $2560 \times 1600$ pixels, $77.53 \times 48.46$ visual degrees) at a distance of 50 cm . Four USB speakers (Mini HiFi USB 2.0 mini speaker), which were connected via a USB hub (Orico HF9US-2P USB 9-Port HUB) were arranged in a $2 \times 2$ grid above and below the monitor (vertically and horizontally 1600 pixels, equivalent to 48.45 visual degrees, apart) in front of the participants (Figure 1). The positions of the visual flashes ( 80 x 80 pixels, 2.42 visual degrees wide, 100 ms ) were arranged in the same $2 \times 2$ grid, such that the visual flashes were observed from approximately the same spatial locations as the auditory stimuli (sine wave tone, 4800 Hz , 100 ms ) - they were vertically displaced by 2.4 cm .

Participants sat in two chairs placed in front of the computer screen (left and right of the fixation cross, respectively) with keyboards on their laps.


Figure 1: Experimental Setup.

## Experimental conditions and procedure

In the experiment, participants performed an audiovisual congruency task either alone or jointly. In this task, participants received visual flashes and auditory tones, originating either from the same (i.e., congruent) or a different (i.e., incongruent) spatial elevation. In addition, stimuli could originate either from the same or opposite side. For example, either both stimuli could originate from the left side or one could originate from the left and one from the right side. The task was to indicate the elevation of one of these stimuli using the keyboard with the mapping of keys $\mathrm{F} / \mathrm{up} \& \mathrm{C} /$ down for the visual stimuli; keys $\mathrm{K} / \mathrm{up}$ \&
$\mathrm{M} /$ down for the auditory stimuli. We set the time limit for responses to 2 seconds (see Figure 2, for a trial overview).

When participants performed the task jointly, they sat next to each other in front of the computer screen in close proximity ( $\sim 10 \mathrm{~cm}$ ) to ensure that they shared peripersonal space (Heed et al., 2010). In this condition, one participant would indicate the elevation of the auditory stimuli while the other participant would indicate the elevation of the visual stimuli. When participants performed the task alone, one participant was asked to wait outside the experiment room while the other participant performed the task, indicating the stimulus elevation for their assigned modality. Note, regardless of whether participants performed the task alone or jointly, the seating positions of participants remained constant in all conditions within a pair and were counterbalanced across pairs (i.e., in half of the pairs, the participant responding to the auditory stimuli was sitting on the right side).


100 ms

Figure 2: Trial overview. (A) Participants simultaneously received a visual and auditory stimulus. (B) Participants were required to indicate the elevation of one of the stimuli using the keyboard. In this example trial, the auditory stimulus would be in the upper location on the right side, the visual stimulus in the bottom location on the left side (i.e., an incongruent opposite side trial). After two seconds passed, the next trial started automatically.

In sum, the experiment consisted of a $2 \times 2 \times 2$ factorial design with Congruency (Congruent, Incongruent), Side (Same, Opposite), and Condition (Individual, Joint) as factors.

The experiment consisted of six blocks, each composed of 144 trials. In these trials, each combination of the factor levels for the factors Congruency and Side occurred equally often in a randomized order. The factor Condition was varied across blocks. That is, there were three types of blocks: 1) The participant responding to the visual stimuli performing the task alone, 2) the participant responding to the auditory stimuli performing the task alone, 3) both participants performing the task jointly. Participants performed a pseudorandomized sequence of these three types of blocks twice. We avoided repetitions of the same block type in consecutive blocks.
The experiment took approximately 40 minutes. It was programmed in Python 2.7.3.

## Data preparation and analysis

In line with Heed et al. (2010), we restricted our analysis to the participant in a pair responding to the auditory stimuli. That is, given that visual stimuli are considerably
easier to localize than auditory stimuli, CCE effects are only observed for the participants responding to the auditory stimuli. Prior to performing inferential statistical tests, we tested whether the normality assumption was given with a Shapiro-Wilk test. In the case of a violation, we transformed the data using a log transformation.

## Results

On a descriptive level (see Figure 3A \& B), when examining the reaction times of correctly localized auditory cues, participants were slower to localize the cues in the incongruent condition compared to the congruent condition. This observation establishes the well-known CCE effect. Furthermore, in line with earlier studies (Heed et al., 2010), the CCE was more pronounced for stimuli that were shown on the same side compared to the opposite side. Importantly, for same side stimuli, the CCE was reduced profoundly in the joint condition relative to the individual condition.


Figure 3: Mean reaction time (in seconds) as a function of the factors Condition (Individual, Joint) and Congruency (congruent, incongruent), separately for same side (A) and opposite side stimuli (B). Error bars in both panels are standard error of the mean.

We tested whether these observations were statistically reliable by performing a $2 \times 2 \times 2$ repeated measures ANOVA with the factors Congruency (Congruent, Incongruent), Side (Same, Opposite), and Condition (Individual, Joint). As the assumption of normality was violated, we applied a log transformation to the reaction times prior to entering them to the ANOVA.

We found a significant main effect for the factor Congruency $(F(1,11)=19.73, p<.001)$. We found significant two-way interactions between the factors Side and Congruency $(F(1,11)=38.42, p<.001)$ and the factors Condition and Congruency $(F(1,11)=6.00, p=.032)$. The former interaction effect suggests that the magnitude of the CCE is reduced for opposite side stimuli compared to same side stimuli. Importantly, the latter interaction effect suggests that the CCE is reduced for the joint condition compared to the individual condition. In addition, we also observed a three-way interaction $(F(1,11)=6.61, p=.026)$, suggesting that the reduced CCE for the joint condition compared to the individual condition depends on whether stimuli appear on the same side or opposite sides. To further investigate the three-way interaction effect, we performed two $2 \times 2$ repeated measures ANOVAs (Condition $x$ Congruency), restricting the data either to only same side or opposite side stimuli. For same side stimuli, we found a significant main effect of Congruency $(F(1,11)=27.29, p<$ .001 ) and a significant interaction between the factors Condition and Congruency $(F(1,11)=9.62, p=.01)$. This demonstrates that for same side stimuli, performing a task jointly indeed reduced the CCE. However, for opposite side stimuli, we only found a significant main effect of Congruency $(F(1,11)=5.58, p=.038)$ but no interaction effect between the factors Condition and Congruency $(F(1,11)=0.07, p=.801)$. Both of these results are in line with the findings by Heed et al. (2010). That is, when investigating visuotactile integration, Heed et al. (2010) similarly found that the CCE effect was reduced in the joint condition relative to the individual condition for same side stimuli but not for opposite side stimuli.

We also tested an alternative explanation of these results by a speed-accuracy tradeoff. That is, in the joint condition, participants potentially could have localized the incongruent cues faster at the expense of being less accurate in their responses. To investigate this, we repeated the $2 \times 2 \times 2$ repeated measures ANOVA with the dependent variable fraction correct (for a descriptive overview, see Figure 4A \& B). We found significant main effects for the factors Side $(F(1,11)=73.32, p<.001)$ and Congruency $(F(1,11)=$ 29.87, $p<.001$ ) and a significant interaction effect between these two factors $(F(1,11)=66.14, p<.001)$. Importantly, we did not find a significant main effect or interaction
involving the factor Condition (Condition: $F(1,11)=0.31, p$ $=.588$; Condition x Congruency: $F(1,11)=0.02, p=.881$; Condition x Congruency x Side: $F(1,11)=0.003, p=.954$ ). These results indicate that a speed-accuracy tradeoff does not explain the reduced CCE for the joint condition relative to the individual condition reported above because the accuracy did not vary as a function of whether the task was performed in pairs or alone. Thus the latency benefit of the joint condition relative to the alone condition was not acquired at the expense of committing more errors.

In sum, the results for same side stimuli indicate that the CCE is reduced significantly when participants perform an audiovisual crossmodal congruency task jointly compared to when they perform it alone.


Figure 4: Mean fraction correct as a function of the factors Condition (individual, joint) and Congruency (Congruent, Incongruent), separately for same side (A) and opposite side stimuli (B). Error bars in both panels are standard error of the mean.

## Discussion

The present study investigated whether the modulation of the CCE by social factors found in earlier studies investigating visuotactile integration (Heed et al. 2010) can also be observed for audiovisual presentations. In line with Heed et al., we found that the CCE is indeed reduced for same side stimuli when participants perform an audiovisual crossmodal congruency task jointly compared to performing it alone. Furthermore, we found that the data are not explained by a speed-accuracy tradeoff. Collectively, the present results extend Heed et al.'s earlier findings of a modulation of the CCE for a visuotactile crossmodal congruency task, and indicate that this social effect generalizes to audiovisual integration.

A possible "mechanism" for our present social effect could be a co-representation process (Sebanz, Knoblich, \& Prinz, 2003; for reviews see: Sebanz, Bekkering, \& Knoblich, 2006; Vesper et al., 2017). That is, when participants perform the task jointly, participants corepresent the task of their partner (e.g., that the partner responds to the visual stimuli) which could lead to a reduced processing of the stimuli relevant for the partner but irrelevant for the own task. As a consequence, the irrelevant stimuli could be perceived as less distracting for incongruent stimulus presentation but still sufficiently processed for congruent presentations, yielding faster reaction times. Alternatively, the effects in the present study could be explained by a dynamic modulation of the coactor's peripersonal space as found in an earlier study (Teneggi et al., 2013) or by a general withdrawal of attention to the stimuli to which the co-actor responds (Szpak et al., 2015).

Future studies could discern further how social factors contribute to the modulation of the CCE. In the present study, pairs of participants performed the crossmodal congruency task in the same peripersonal space and both participants performed the task. Earlier findings (Heed et al., 2010) showed that the CCE for visuotactile stimuli is only affected by social factors if both participants perform the task and are located in their respective peripersonal spaces. It is an open question whether a reduction of the CCE for audiovisual stimuli would be observed when only one of these factors is manipulated. For instance, when participants are in the same peripersonal space but only one of them performs the task, or when both of them perform the task but from separate peripersonal spaces. In contrast to the tactile modality, both the visual and the auditory modality investigated here sample distant events. Thus, it is quite conceivable that visuotactile integration is dependent on jointly executing the task in peripersonal space while this might not be the case for audiovisual integration.

As another point of note, our finding that performing the crossmodal congruency task jointly affects the CCE for same side stimuli but not for opposite side stimuli could be explained by the observation that for opposite side stimuli the CCE was already greatly reduced in the individual
condition. That is, an already lower CCE may not allow for any additional modulations by social factors.

Future studies could also test whether the social effects found in this study can alternatively be explained by other factors (Stenzel \& Liepelt, 2016). For instance, it could be investigated whether a non-human co-actor (e.g., a robot) responding to the distractors is sufficient to find the effects in the present study (Stenzel et al., 2012).

More generally, the present findings are relevant to, and may benefit, real-world situations in which humans perform tasks jointly while processing multisensory information. That is, our data and the earlier findings of Heed et al., (2010) suggest that the benefits of multisensory integration are preserved when performing a task jointly (i.e., participants respond faster to congruent multisensory stimuli) while the costs of multisensory integration are reduced (i.e., participants are slowed down less by incongruent stimuli). Future studies could investigate further how the benefits of multisensory processing (e.g., due to multisensory integration (Alais \& Burr, 2004; Ernst \& Banks, 2002; Körding et al. 2007; Rohe \& Noppeney, 2015), sensory augmentation (König et al., 2016; Goeke, Planera, Finger, \& König, 2016), or circumventing limited attentional resources (Alais \& Burr, 2004; Arrighi, Lunardi, \& Burr, 2011; Wahn, et al. 2016; for a review, see: Wahn \& König, 2017)) may facilitate human performance in other joint settings.

## Acknowledgments

This research was supported by H2020-H2020-FETPROACT-2014641321—socSMCs (for BW \& PK).

## References

Alais, D., \& Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Current Biology, 14, 257-262.

Alais, D., Morrone, C., \& Burr, D. (2006). Separate attentional resources for vision and audition. Proceedings of the Royal Society of London B: Biological Sciences, 273, 1339-1345.

Alsius, A., Navarra, J., Campbell, R., \& Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. Current Biology, 15(9), 839-843.

Arrighi, R., Lunardi, R., and Burr, D. (2011). Vision and audition do not share attentional resources in sustained tasks. Frontiers in Psychology. 2:56.
doi:10.3389/fpsyg. 2011.00056
Ernst, M. O., \& Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415, 429-433.

Goeke, C. M., Planera, S., Finger, H., \& König, P. (2016). Bayesian alternation during tactile augmentation. Frontiers in Behavioral Neuroscience, 10 .

Heed, T., Habets, B., Sebanz, N., Knoblich, G. (2010). Others' actions reduce crossmodal integration in peripersonal space. Current Biology, 20, 1345-1349.

Helbig, H. B., \& Ernst, M. O. (2008). Visual-haptic cue weighting is independent of modality-specific attention. Journal of Vision, 8, 1-16.

König, S. U., Schumann, F., Keyser, J., Goeke, C., Krause, C., Wache, S., ... \& König, P. (2016). Learning new sensorimotor contingencies: Effects of long-term use of sensory augmentation on the brain and conscious perception. PLoS ONE, 11(12), e0166647. doi:
10.1371/journal.pone. 0166647

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., \& Shams, L. (2007). Causal inference in multisensory perception. PLoS ONE, 2(9), e943.

Plöchl, M., Gaston, J., Mermagen, T., König, P., \& Hairston, W. D. (2016). Oscillatory activity in auditory cortex reflects the perceptual level of audio-tactile integration. Scientific Reports, 6:33693.

Rohe, T., \& Noppeney, U. (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception. PLoS Biology, 13(2), e1002073.

Sebanz, N., Knoblich, G., \& Prinz, W . (2003). Representing others' actions: just like one's own? Cognition, 88, 11-21.

Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: bodies and minds moving together. Trends in Cognitive Sciences, 10, 70-76.

Spence, C., Pavani, F., \& Driver, J. (1998). What crossing the hands can reveal about crossmodal links in spatial attention. Abstracts of the Psychonomic Society, 3, 13.

Spence, C., Pavani, F., \& Driver, J. (2004). Spatial constraints on visual-tactile cross-modal distractor congruency effects. Cognitive, Affective, \& Behavioral Neuroscience, 4, 148-169.

Spence, C. (2007). Audiovisual multisensory integration. Acoustical Science and Technology, 28(2), 61-70.

Stenzel, A., \& Liepelt, R. (2016). Joint Simon effects for non-human co-actors. Attention, Perception, \& Psychophysics, 78(1), 143-158.

Stenzel, A., Chinellato, E., Bou, M. A. T., del Pobil, Á. P., Lappe, M., \& Liepelt, R. (2012). When humanoid robots
become human-like interaction partners: Corepresentation of robotic actions. Journal of Experimental Psychology: Human Perception and Performance, 38(5), 1073.

Szpak, A., Loetscher, T., Churches, O., Thomas, N. A., Spence, C. J., \& Nicholls, M. E. (2015). Keeping your distance: Attentional withdrawal in individuals who show physiological signs of social discomfort. Neuropsychologia, 70, 462-467.

Talsma, D. (2015). Predictive coding and multisensory integration: An attentional account of the multisensory mind. Frontiers in Integrative Neuroscience, 9:19. doi:10.3389/fnint.2015.00019

Teneggi, C., Canzoneri, E., di Pellegrino, G., \& Serino, A. (2013). Social modulation of peripersonal space boundaries. Current Biology, 23(5), 406-411.

Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., ... , \& Wahn, B. (2017). Joint action: Mental representations, shared information and general mechanisms for coordinating with others. Frontiers in Psychology, 7, 2039.

Wahn B., \& König P. (2015a) Audition and vision share spatial attentional resources, yet attentional load does not disrupt audiovisual integration. Frontiers in Psychology, 6:1084. doi:10.3389/fpsyg. 2015.01084

Wahn B., \& König P. (2015b) Vision and haptics share spatial attentional resources and visuotactile integration is not affected by high attentional load. Multisensory Research 28, 371-392. doi:10.1163/22134808-00002482

Wahn, B., Schwandt, J., Krüger, M., Crafa, D., Nunnendorf, V., \& König, P. (2016). Multisensory teamwork: using a tactile or an auditory display to exchange gaze information improves performance in joint visual search. Ergonomics, 59, 781-795. doi: 10.1080/00140139.2015.1099742

Wahn B., \& König P. (2016) Attentional resource allocation in visuotactile processing depends on the task, but optimal visuotactile integration does not depend on attentional resources. Frontiers in Integrative Neuroscience, 10:13.

Wahn B., Murali, S., Sinnett, S., \& König P. (2017) Auditory stimulus detection partially depends on visuospatial attentional resources. i-Perception, 1-17. doi: 10.1177/2041669516688026

Wahn B., \& König P. (2017) Is attentional resource allocation across sensory modalities task-dependent? Advances in Cognitive Psychology, 13(1), 83-96. doi: 10.5709/acp-0209-2

# More than meets the eye: <br> Early relational reasoning cannot be reduced to perceptual heuristics 

Caren M. Walker (carenwalker@ucsd.edu)<br>Department of Psychology, McGill Hall (\#0109), University of California San Diego<br>9500 Gilman Dr., La Jolla, CA 92093 USA

Alison Gopnik (gopnik@berkeley.edu)<br>Department of Psychology, 3210 Tolman Hall, University of California, Berkeley<br>Berkeley, CA 94720 USA


#### Abstract

The ability to represent same-different relations is a condition for abstract thought. However, there is mixed evidence for when this ability develops, both ontogenetically and phylogenetically. Apparent success in relational reasoning may be evidence for conceptual understanding or may be due to low-level, perceptual strategies. We introduce a method to discriminate these possibilities by pitting two conditions that are perceptually matched but conceptually different: in a "fused" condition, same and different objects are joined, creating single objects that have the same perceptual features as the pairs in the "relational" condition. However, the "fused" objects do not provide evidence for the relation. Using this method in a causal task provides evidence for genuine conceptual understanding. This novel technique offers a simple manipulation that may be applied to a variety of existing match-to-sample procedures used to assess same-different reasoning to include in future research with non-human animals, as well as human infants.


Keywords: cognitive development; causal inference; relational reasoning; perceptual processes

The ability to represent relations between objects and events is an essential condition for abstract thought; some have suggested that relational abilities may be the key to the cognitive differences between humans and other animals (Penn, Holyoak, \& Povinelli, 2008). However, there is mixed evidence about when this ability develops, both ontogenetically and phylogenetically. Traditionally, there was little evidence for relational reasoning in either young children or non-human animals. More recent results, particularly involving the foundational relations "same" and "different" challenge that conclusion. Ducklings can generalize these relations in an imprinting paradigm (Martinho \& Kacelnik, 2016). Human infants are able to generalize these relations in looking-time experiments. In particular, pre-verbal infants can be habituated to pairs of same and different objects (Addyman \& Mareschal, 2010; Ferry, Hespos, \& Gentner, 2015; Hochman, Mody, \& Carey, 2016; Tyrell et al., 1991), discriminate and generalize patterns of repeated visual or auditory elements (ABA/AAB/ABB) (Dawson \& Gerken, 2009; Johnson et al., 2009; Marcus et al., 1999; Saffran et al., 2007), and provide
a conditioned response to pairs of identical stimuli (Kovács, 2014; Hochmann, 2010). Moreover, very young toddlers can apparently use same-different relations in an active causal learning paradigm (Walker \& Gopnik, 2014), although this ability declines in the preschool period (Walker, Bridgers, \& Gopnik, 2016). In these studies, toddlers, aged 18-30-months, were able to infer same-different relations in a causal version of a match to sample task (i.e., matching AA' with $\mathrm{BB}^{\prime}$, not CD, and matching EF with CD, not BB').

On the other hand, it is possible that these successes may be mediated by perceptual factors that are quite separate from the abstract same-different concepts that these tasks are intended to assess (see Addyman \& Mareschal, 2010 for a review). It is clear that both human and non-human animals are able to perceive the similarity of objects, agents, and events in their environment; these abilities are necessary for basic cognitive functions (Martinho \& Kacelnik, 2016; Hochman, Mody, \& Carey, 2016). However, noticing similarity does not necessarily imply the existence of the conceptual representation, same. This distinction is difficult to make, and this point has been widely debated in the comparative literature (Penn, Holyoak, \& Povinelli, 2008; Thompson \& Oden, 1996).

For example, non-human primates (Wasserman, Fagot, \& Young, 2001) and several species of birds (Smirnova et al., 2015; Pepperberg, 1987) have succeeded in solving similar relational problems, in the context of multiple trials in reinforcement learning paradigms (Wasserman, Fagot, \& Young, 2001; Smirnova et al., 2015; Pepperberg, 1987), suggesting that these species, like humans, may possess the ability to learn abstract relational properties (Cook \& Wasserman, 2007). However, there is also growing evidence indicating that these trained abilities may be grounded in perceptual expertise, reflecting learned sensitivity to surface cues, rather than higher-order reasoning, per se (Thompson \& Oden, 2000).

This suggests that the match to sample tasks that have historically served as the standard for assessing samedifferent understanding across species may be passed in the absence of genuine conceptual representations. In particular, lower-level, perceptual strategies, like attention to the symmetry, contrast, and the variance of the stimuli could contribute to success (Young \& Wasserman, 2001; Smith et
al. 2008; Blaisdell \& Cook, 2005). Might infants, toddlers, and non-human animals in an imprinting paradigm, like nonhuman animals in reinforcement training, be responding to a perceptual analysis of the stimuli pairs rather than a samedifferent strategy?

One candidate for such a strategy is a low-level heuristic, called "perceptual entropy," that has been proposed to facilitate relational recognition in non-human animals (Penn, Holyoak, \& Povinelli, 2008; Wasserman, Fagot, \& Young, 2001; Fagot, Wasserman, \& Young, 2001; Young \& Wasserman, 1997; Wasserman, Young, \& Cook, 2004; Wasserman \& Young, 2010; Zentall et al., 2008). In particular, any visual display can be reduced to "a continuous analog estimate of the degree of perceptual variability between the elements" (Penn, Holyoak, \& Povinelli, 2008, pg. 112), a strategy similar to a process of conceptual chunking (Halford, Wilson, \& Phillips, 1998). In other words, because there is a lower amount of perceptual variation among the elements for 'same' displays (AA') than for 'different' displays (AB), toddlers (as well as human infants and non-human animals) may succeed by learning and applying the following rule: If the variability of the effective training sample is low, select the test pair that also has low variability.

This attention to variance would also subsume a range of other perceptual cues including symmetry, oddity, and spatial orientation, among others (Cook \& Wasserman, 2007). Adult humans show some sensitivity to the amount of perceptual variance in a display, but this evidence is not sufficient to prove that it is responsible for their performance. In fact, previous findings suggest that additional processes of categorization likely play $a$ role in the human conceptualization of "same-different" relations (Smith et al., 2008; Fagot, Wasserman, \& Young, 2001). Interestingly, similar findings have been recently found with baboons (Flemming, Thompson, \& Fagot, 2013).

Discriminating between conceptual and perceptual learning strategies in non-verbal relational reasoning tasks is a notoriously difficult problem to solve in both developmental and comparative contexts. In the current study, we introduce a novel method designed to directly pit the perceptual and conceptual accounts against one another. The method involves a contrast between one condition relying upon a traditional match to sample task involving same-different relations (i.e., matching AA' with BB', not CD, and matching EF with CD, not BB') and a "fused" object condition. Exactly the same objects are used in the two conditions, but in the "fused" condition the objects are physically joined to create a single compound object. Importantly, the amount of perceptual entropy, or variance, as well as other perceptual features such as symmetry is matched between the two conditions. However, only the unfused/relational condition also provides evidence for the higher-order relation 'same.' In the fused/single object case, there is no relation between objects to learn - there is only one object present.

As a proof of concept, we applied this method to assess human toddlers in a causal match to sample task originally developed by Walker and colleagues (Walker \& Gopnik, 2014; Walker, Bridgers, \& Gopnik, 2016). In the current study, children observed two trials in which a pair of "same" objects, or a fusion of those objects, activated a machine, but a pair or fusion of two "different" objects did not. Then, children had to select a novel pair of objects or a novel fused object to activate the machine (see Figure 1). If children are indeed relying upon a low-level perceptual heuristic, they should select the lower entropy pair (i.e., the pair with less variance among its features) consistently across both conditions, whether they are fused or not. On the other hand, if children learn the abstract relation 'same' during the training trials, they should privilege this test pair only in the unfused/relational condition, where there is a relation between objects to learn.

Although the current study applies this method to assess human reasoning in a previously published causal reasoning paradigm, this same technique is intended to be used for discriminating perceptual strategies from genuine relational reasoning in a variety of existing paradigms, across species.

## Method

## Participants

A total of 80 18-30-month-olds participated $(M=24.3$ months; $S D=3.6$ months; range $=17.9-31.1$ months; 40 girls), with 40 toddlers randomly assigned to one of two conditions (fused/single object or unfused/relational). There was no difference in age between conditions, $t(1)=1.21, p=$ .23, and approximately equal numbers of males and females were assigned to each. Sixteen additional children were tested but excluded for failure to complete the study (11) or due to experimenter error (5). Children were recruited from a local museum.

## Materials

The toy was a 10 " $\times 6$ " $\times 4$ " opaque cardboard box containing a wireless doorbell. When a block or pair of blocks "activated" the toy, the doorbell played a novel melody. In fact, the toy was surrepticiously activated by a remote control. Eight painted wooden blocks in assorted colors and shapes (2 pairs of 'same' blocks and 2 pairs of 'different' blocks) were placed on the toy in pairs during the unfused/relational condition training. The 'same/low entropy' blocks were identical in color and shape, and the 'different/high entropy' blocks were distinct in color and shape. An identical set of these eight painted blocks were used to create the "fused" objects to be placed on the toy as single objects in the fused/single object condition training. In this condition, each pair of training blocks were glued together to create a single, larger block. Four additional blocks were used during the test phase of each condition, including 1 novel pair of 'same' and 1 novel pair of 'different' blocks. The test blocks either appeared as two pairs of blocks or as two fused, single objects, depending
upon condition (see Figure 1). The pairs of test blocks in each condition were placed on 4 " $\times 4$ " plastic trays.

Two different complete sets of blocks were constructed for each condition. In the simple set, all blocks were composed of simple, symmetrical geometric shapes (e.g., cubes, spheres, cylinders) with a single color and no pattern. In the complex set, all blocks were composed of asymmetrical, irregular polygons. Half of the children in each condition were randomly assigned to receive each stimuli set.

## Procedure

All children were tested one-on-one, seated at a table across from the experimenter. Following a brief warm-up, the experiementer introduced a toy that was placed on the table. The experimenter said, "This is my toy. Some things make my toy play music and some things do not. Let's try some things on my toy and find out how it works."

In the unfused/relational condition, children observed as the experimenter placed a pair of 'same' blocks (AA') on the toy, causing it to activate and play music (twice). They then observed that a pair of 'different' blocks (BC) failed to activate the toy (twice). This procedure was repeated for two additional pairs, one pair of 'same' (DD') and one pair of 'different' blocks (EF) (see Figure 1). The 'same' pairs (AA', DD') were composed of individual blocks that were identical in both color and shape, and the 'different' pairs (BC, EF) were composed of individual blocks distinct in both color and shape. In the fused/single object condition, children observed an identical presentation with one critical exception: each pair of blocks were glued together to form single objects (Alow entropy, B-high entropy, C-low entropy, D-high entropy) (see Figure 1).

In detail, the experimenter selected the first pair [block], saying, "Let's try!" and placed them [it] on the toy. Children in both conditions observed the 'same' pair ['low entropy' block'] activate the toy. The experimenter said, "Music! Let's try again!", picked up the pair [block], and placed them [it] back on the toy a second time, and children observed the outcome. The experimenter said, "Music! These ones [this one] made my toy play music." After this second demonstration, the experimenter removed the pair [block], selected another - a 'different' pair or a 'high entropy' block - and placed it on the toy. This time, children in both conditions observed no effect. The experimenter said, "No music. Let's try again!" As with the first pair [block], this was demonstrated a second time. The experimenter concluded, "No music. These ones [this one] did not make my toy play music."

This procedure was repeated for all 4 pairs [blocks]: 2 pairs [blocks] of 'same' ['low entropy'] objects and 2 pairs [blocks] of 'different' ['high entropy'] objects. All pairs were placed on the toy twice. Therefore, children observed a total of 8 outcomes (4 positive and 4 negative). The order that the individual pairs [blocks] were presented was randomized, however, the order of the presentation pairs was fixed, beginning with a causal pair, and alternating between causal and inert pairs. In all cases, the experimenter placed all pairs
of objects on the toy in the same orientation as the objects that formed the fused blocks, so that they were perceptually identical. Except for the particular objects used in the training trials (fused or unfused), there were no other differences in procedure between conditions.


Figure 1: Schematic of study design (simple set). On training trials, pairs of blocks were placed on the toy. In the fused/single object condition, fused, identical/low entropy objects activated the toy, while fused, distinct/high entropy objects did not. In the unfused/relational condition, pairs of identical/low entropy objects activated the toy while pairs of distinct/high entropy objects did not. Participants observed 4 pairs (2 causal, 2 inert). On each test trial, the child selected between 2 novel pairs ("low entropy [same]" or "high entropy [different]").

Following the training phrase in both conditions, the experimenter said, "Now it is your turn. Can you help me pick the thing[s] that will make my toy play music?" The experimenter produced 2 pairs of test blocks (1 novel 'same' pair ['low entropy' block], 1 novel 'different' pair ['high entropy' block]). In order to avoid a novelty preference, both test pairs were composed of novel objects. The pairs were presented to the child on trays. The experimenter held up the two trays, saying, "I have these [this] and I have these [this]. Only one of these trays has the thing[s] that will make my toy
play music." She then lowered the trays and placed them on opposite sides of the table in front of the child, saying, "Can you point to the one[s] that will make my toy play music?" The side on which the correct pair was placed was randomized between subjects.

Coding The first tray that the child selected (pointing, reaching, picking up objects) was recorded. Children received 1 point for selecting the low entropy pair/object that was consistent with their training and 0 points for selecting the high entropy pair/object. Children's responses were recorded by a second researcher during the testing session, and all sessions were video recorded for independent coding by a third researcher who was naïve to the the hypotheses of the experiment. Interrater reliability was very high; the two coders agreed on $99 \%$ of the children's responses to the test questions.

## Results

Results show no difference between the complex objects and simple objects, in either condition, $\chi^{2}(1)=0, \mathrm{p}=1, \varphi=0$ (fused/single object); $\chi^{2}(1)=.13, \mathrm{p}=.72, \varphi=-.06$ (unfused/relational). We therefore combined data from the two stimuli sets within each condition for all subsequent analyses. Children in the unfused/relational condition selected the 'same' test pair more often than chance (73\%), $p$ $=.006$ (two-tailed, exact binomial). These results replicate previous findings with 18-30-month-olds (Walker \& Gopnik, 2014; Walker et al., 2016). However, in contrast with the perceptual account, children of the same age in the fused/single object condition selected at chance (40\%), $p=$ .27 (two-tailed, exact binomial). There was a significant difference between conditions, $\chi^{2}(1)=8.58, \quad p=.004, \varphi=$ . 33.

## Discussion

Results demonstrate that when perceptual cues are matched, but no relation is present, toddlers do not appear to learn and generalize an abstract concept of 'same' to a novel set of objects. These findings therefore suggest that early relational competence in humans found here and elsewhere is unlikely to be the result of reliance on a perceptual heuristic, and provide evidence for genuine conceptual understanding of 'same' at this young age.

This novel method offers a simple, non-verbal manipulation that may be applied to a variety of existing match-to-sample procedures used to assess same-different reasoning to include in future research with non-human animals across species, as well as human infants. If infants or animals show the discriminative pattern of the toddlers in this experiment - generalizing the unfused/relational but not the fused/single objects - that suggests that they genuinely understand the relations. On the other hand, if they respond in the same manner to both conditions, the perceptual hypothesis would gain more weight. The latter pattern would not eliminate the possibility that relational reasoning was in play - perhaps children or animals are using different kinds
of reasoning in the two conditions. But it would place the burden of proof on the relational claim.

Whatever the results of non-human animals or infants might turn out to be, the present results are consistent with previous claims that, from a very early age, as young as 18 months, humans posess cognitive tools for genuine conceptual understanding of same-different relations. These findings are also consistent with the idea that humans may possess a qualitatively different system for abstracting relations.

## Acknowledgments

Research for this project was funded by the American Psychological Foundation to C. Walker (Elizabeth Munsterberg Koppitz Child Psychology Graduate Fellowship) and the National Science Foundation to A. Gopnik. We thank all of the parents and children who participated in this research, as well as the Childhood Creativity Center at the Bay Area Discovery Museum for facilitating recruitment. Finally, we thank Rosie Aboody, Hannah Broidy, and Gillian Rush for their efforts towards data collection.

## References

Addyman, C., \& Mareschal, D. (2010). The perceptual origins of the abstract Same/Different concept in human infants. Animal Cognition, 13(6), 817-833.
Blaisdell, A. P., \& Cook, R. G. (2005). Two-item same different concept learning in pigeons. Animal Learning \& Behavior, 33(1), 67-77.
Cook, R. G., \& Wasserman, E. A. (2007). Learning and transfer of relational matching-to-sample by pigeons. Psychonomic Bulletin \& Review, 14(6), 11071114.

Dawson, C., \& Gerken, L. A. (2009). From domaingenerality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. Cognition, 111, 378-382.
Fagot, J., Wasserman, E. A., \& Young, M. E. (2001). Discriminating the relation between relations: the role of entropy in abstract conceptualization by baboons (Papio papio) and humans (Homo sapiens). Journal of Experimental Psychology: Animal Behavior Processes, 27(4), 316.
Ferry, A., Hespos, S. J., \& Gentner, D. (2015). Prelinguistic relational concepts: Investigating the origin of analogy in infants. Child Development, 86(5), 1386-1405.
Flemming, T. M., Thompson, R. K., \& Fagot, J. (2013). Baboons, like humans, solve analogy by categorical abstraction of relations. Animal cognition, 16(3), 519-524.
Halford, G. S., Wilson, W. H., \& Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences, 21(06), 803-831.
Hochmann, J. R. (2010). Categories, words and rules in language acquisition (Doctoral Dissertation).

Hochmann, J. R., Mody, S., \& Carey, S. (2016). Infants’ representations of same and different in match-and non-match-to-sample. Cognitive psychology, 86, 87-111.
Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N. Z., Marcus, G. F., Rabagliati, H., \& Slemmer, J. A. (2009). Abstract rule learning for visual sequences in 8- and 11-month-olds. Infancy, 14, 2-18.
Kovács, Á. M. (2014). Extracting regularities from noise: Do infants encode patterns based on same and different relations? Language Learning. http://dx.doi.org/10.1111/lang. 12056.
Marcus, G. F., Vijayan, S., Bandi Rao, S., \& Vishton, P. M. (1999). Rule-learning in seven-month-old infants. Science, 283, 77-80.
Martinho, A., \& Kacelnik, A. (2016). Ducklings imprint on the relational concept of "same or different". Science, 353 (6296), 286-288.

Penn, D.C., Holyoak, K.J., Povinelli, D.J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behavioral and Brain Sciences, 31, 109178.

Pepperberg, I. M. (1987). Acquisition of the same/different concept by an African Grey parrot (Psittacus erithacus): Learning with respect to categories of color, shape, and material. Animal Learning \& Behavior, 15(4), 423-432.
Saffran, J. R., Pollak, S. D., Seibel, R. L., \& Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. Cognition, 105(3), 669-680.
Smirnova, A., Zorina, Z., Obozova, T., \& Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. Current Biology, 25(2), 256-260.
Smith, J. D., Redford, J. S., Haas, S. M., Coutinho, M. V., \& Couchman, J. J. (2008). The comparative psychology of same-different judgments by humans (Homo sapiens) and monkeys (Macaca mulatta). Journal of Experimental Psychology: Animal Behavior Processes, 34(3), 361.
Thompson, R. K. R., \& Oden, D. L. (1996). A profound disparity revisited: Perception and judgment of abstract identity relations by chimpanzees, human infants, and monkeys. Behavioral Processes, 35, 149-161.

Thompson, R. K. R., \& Oden, D. L. (2000). Categorical perception and conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape. Cognitive Science, 24(3), 363-396.
Tyrell, D. J., Stauffer, L. B., \& Snowman, L. G. (1991). Perception of abstract identity/difference relationship by infants. Infant Behavior and Development, 14, 125-129.
Walker, C.M., Bridgers, S., \& Gopnik, A. (2016). The early emergence and puzzling decline of relational reasoning: Effects of knowledge and search on inferring abstract concepts. Cognition, 156, 30-40.
Walker, C. M., \& Gopnik, A. (2014). Toddlers infer higherorder relational principles in causal learning. Psychological Science, 25(1), 161-169.
Wasserman, E.A., Fagot, F., \& Young, M.E. (2001). Samedifferent conceptualizations by baboons (Papio papio): The role of entropy. Journal of Comparative Psychology, 115(1): 42-52.
Wasserman, E. A., Young, M. E., \& Cook, R. G. (2004). Variability discrimination in humans and animals: implications for adaptive action. American Psychologist, 59(9), 879.
Wasserman, E. A., \& Young, M. E. (2010). Same-different discrimination: The keel and backbone of thought and reasoning. Journal of Experimental Psychology: Animal Behavior Processes, 36(1), 3.
Young, M. E., \& Wasserman, E. A. (1997). Entropy detection by pigeons: Response to mixed visual displays after same-different discrimination training. Journal of Experimental Psychology: Animal Behavior Processes, 23(2), 157.
Young, M. E., \& Wasserman, E. A. (2001). Entropy and variability discrimination. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(1), 278.

Zentall, T. R., Wasserman, E. A., Lazareva, O. F., Thompson, R. K., \& Rattermann, M. J. (2008). Concept learning in animals. Comp Cogn Behav Rev, 3, 13-45.

# Simultaneous acquisition of vocabulary and grammar in an artificial language learning task 

Neil Walker (n.walker@lancaster.ac.uk)<br>Department of Linguistics and English Language<br>Lancaster University, Lancaster LA1 4YL, UK

Christine Schoetensack (s1649700@sms.ed.ac.uk)<br>School of Philosophy, Psychology and Language Sciences<br>Edinburgh University, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD

Padraic Monaghan (p.monaghan@lancaster.ac.uk)<br>Department of Psychology<br>Lancaster University, Lancaster LA1 4YF, UK<br>Patrick Rebuschat (p.rebuschat@lancaster.ac.uk)<br>Department of Linguistics and English Language<br>Lancaster University, Lancaster LA1 4YL, UK


#### Abstract

Learning syntax requires determining relations between the grammatical categories of words in the language, but learning those categories requires understanding the role of words in the syntax. In this study, we examined how this chicken and egg problem is resolved by learners of an artificial language comprising nouns, verbs, adjectives and case markers following syntactic rules. We found that the language could be acquired through cross-situational statistical correspondences with complex scenes and without explicit feedback, and that knowledge was maintained after 24 hours. Results also showed that verbs and word order were the first to be acquired, followed by nouns, adjectives and finally case markers. Interdependencies in learning were found for word order and verbs, and also for nouns, adjectives and case markers. Grammar and vocabulary can be acquired simultaneously, but with distinctive patterns of acquisition grammar and the role of verbs first, then the vocabulary of other lexical categories.


Keywords: language acquisition; grammar; vocabulary; artificial language learning; statistical learning.

## Introduction

The early stages of language learning involve a great deal of ambiguity as learners must make sense of the stream of input they hear by noticing words boundaries, decoding the meanings of words, identifying lexical categories and understanding the relations between categories defined by the syntax. How this is achieved and the order in which vocabulary and grammatical knowledge is acquired has been a critical question in the cognitive sciences (Marcus, 1996).

## Cross-situational learning

Recent research has shown that it is possible for children and adults to learn vocabulary within basic categories of words when they are presented across numerous ambiguous learning situations without any feedback, a mechanism
known as cross-situational learning. Smith and Yu (2008) showed that 12 to 14 -month old infants could learn the meanings of novel nouns by keeping track of cross-trial statistics. Scott and Fisher (2012) also demonstrated that it is possible for 2.5 year-old toddlers to learn novel verbs, utilising syntactic cues, knowledge of nouns and other situational referents. Monaghan et al. (2015) found that nouns and verbs could be learned simultaneously without any syntactic cues, although nouns were learned slightly more quickly. They suggested that this prioritisation of nouns could be explained by the greater saliency and stability of object versus action referents.

However, these studies on the cross-situational learning of nouns and verbs are a substantial abstraction from the complexity of natural language acquisition. In child-directed speech, children are generally exposed to multi-word utterances containing many word categories (Mintz, 2006). With every new word category or syntactic phrase added, the number of possible referents for any given word increases, making the tracking of statistical probabilities more complex.

On the other hand, with greater complexity comes greater interdependency between content words, function words and word order. And so conversely, this extra complexity may also provide additional cues from which to constrain learning. Monaghan and Mattock (2012), for example, found that function words could aid the learning of nouns in a cross-situational learning paradigm.

A key question is, therefore, how learners can break into this complex stream, given the difficulty of attempting to acquire the syntax to indicate the role of grammatical categories, and learning the vocabulary to populate those categories. This chicken and egg problem has proven difficult to resolve, and has led to proposals either for independence of learning grammar and vocabulary (e.g., Marcus, 1996), or their inter-relatedness (Bates \& Goodman, 1997). Under these latter accounts, learning a
few words can give rise to syntactic bootstrapping (Gleitman, 1990), which can then be used to promote further vocabulary acquisition, predicting correlations in children's performance for vocabulary and grammatical processing tasks. However, evidence from actual language learning tasks in which both vocabulary and grammar must be acquired has not been extensively explored. Many previous studies of artificial language learning have trained participants on vocabulary before testing them on a language structure.

In the current study, we investigated whether it is possible to learn more complex artificial languages that combine content words, function words and syntactic structures under cross-situational learning conditions without any feedback.

## The acquisition order of linguistic categories

The second question this study addressed was that of acquisition order: When are nouns, verbs, adjectives, case markers, and syntactic constraints on word order acquired and are some aspects learned before others? The vast majority of studies into early childhood language learning support Gentner's (1982) claim that across all languages children learn nouns before verbs and adjectives (e.g, Bornstein et al., 2004). One reason might be that predicates are more semantically complex as they modify and depend on nouns (Dixon, 1982), whether that be adjectives (the black dog) or verbs (the cat pounced on the mouse). Therefore, in order to learn the verb or adjective, learners need to also encode information about the noun (Gleitman et al., 2005).

However, there is some evidence that in languages such as Korean and Japanese, where the verb is found in a highly-salient sentence-final position and subjects and objects are often left out of utterances by caregivers, verbs may be learned earlier than in SVO languages such as English (Choi \& Gopnik, 1995). There is, however, a lack of consensus as to whether verbs in these languages are learned at the same time as nouns (see Bornstein et al., 2004).

Regarding adjectives, Booth and Waxman (2003) demonstrated that 14 -month-old children could extract the meaning from nouns but not adjectives when presented with basic syntactic and visual frames. In a large corpus-based study, Behrens (2006) found that German children aged 1 year 11 months produced more verbs than adjectives.

Finally, case markers which indicate the agent and patient of a sentence have been shown to be understood by children as young as two years old (Göksun et al., 2008). However, in this and other studies, a small vocabulary of nouns and verbs had been acquired before comprehension of case markers was demonstrated.

The participants in the current study were adults who have already mastered their first language, and so it is arguable that the acquisition order observed in child language development may not apply to these learners. An alternative strand of research comes from first exposure
studies of adults learning a second language (L2; for an overview, see Indefrey \& Gullberg, 2010). In a study into the initial stages of learning an L2 by adults in a classroom setting, Shoemaker and Rast (2013) found that it was the words in sentence-initial and sentence-final positions that were most easily recognised in a stream of speech. They argued that this was due to not only silence bordering the initial and final words, but that working memory is less burdened for the final word of the utterance.

Another factor that influences whether a word can be picked out of a stream of speech is the number of syllables it contains (Gullberg et al., 2012). With many function words monosyllabic, this could render them less easily noticed than highly salient content words, despite the frequency with which they occur in utterances.

Overall, if the learning of nouns and verbs follows the findings of child language research, then we can expect nouns to be learned before verbs. Alternatively, if sentence position is a more important factor, we could expect verbs, which in our current study occupy the sentence-final position, to be learned before nouns, which are mostly in medial position. We then predict that adjectives will be learned next, followed by case markers, although given the short duration of the learning paradigm, it is possible that the latter may not be learned at all (e.g., DeKeyser, 2005).

## The learning mechanisms of vocabulary and syntax

A final aim of this study was to investigate how the different types of language structure cohere. Is the meaning of vocabulary items (nouns, verbs, adjectives) learned in the same way as grammatical items (word order, case markers) or do they depend on different mechanisms? Research from models of learning data (Frost \& Monaghan, 2016), neuropsychology patient studies (Alario \& Cohen, 2004), theoretical models (Bock \& Levelt, 1994), and memory models (Ullman, 2004) treat vocabulary and syntax as distinct. If this were the case, we might expect word order and case markers to be interdependent, with nouns, verbs and adjectives also grouped together. Alternatively, if syntax and vocabulary share the same learning mechanism, as is postulated in single-system models (MacWhinney, 1987), we might expect to see no interdependency of word order and case markers. Instead, as word order is determined by the position of the verb, it is possible that learning which word is the verb and the word order will be linked.

## Method

## Participants

Sixty-four native speakers of English (47 women) were randomly assigned to two conditions (massed vs distributed, each $n=32$ ) which varied in terms of whether there were pauses between blocks of training on an artificial language learning task. Participants were students or graduates of universities in the North West of England. The mean age was 26.0 years $(S D=7.1)$. None of the participants had previously studied any verb-final languages. Participants in
the massed group received 20 GBP and participants in the distributed group received 28 GBP . The difference was due to the extra time involved in the distributed condition.


Figure 1: Screenshot of the cross-situational learning task. Participants see two dynamic scenes and hear a sentence and decide which scene the sentence refers to.

## Materials

A novel artificial language was created for this experiment. The lexicon consists of 16 pseudowords, taken from Monaghan and Mattock (2012). Fourteen bisyllabic pseudowords were content words: Eight nouns, four verbs, and two adjectives. Two monosyllabic pseudowords served as function words that reliably indicated if the preceding noun referred to the subject or the object of the sentence. The words were recorded by a female native speaker of British English who was instructed to produce the words in a monotone.

In terms of syntax, the artificial language was based on Japanese. Sentences could either be SOV or OSV, i.e. verbs had to be placed in final position but the order of subject and object noun phrases (NPs) was free. NPs had to contain a noun as its head and a post-nominal case marker that indicated if the preceding noun was the agent or the patient of the action. Adjectives were optional and only occurred in half the NPs. Adjectives occurred pre-nominally.

Eight alien cartoon characters served as referents for the language (see Figure 1). The aliens could either appear in red or blue and were depicted performing one of four actions (hiding, jumping, lifting, pushing) in dynamic scenes generated by E-Prime (version 2.0). Figure 1 shows a sample screen shot, containing the target scene and a distractor scene. Each noun referred to one alien, the adjectives referred to the colours of aliens, and the verbs referred to the actions. Word-referent mappings were randomly generated for each participant to control for preferences in associating certain sounds to objects, motions or colours.
For training, there were 12 blocks of 16 trials each. In each trial, two scenes were presented and an artificial language sentence played. The sentence described only one of the scenes and the participants had to match the sentence to the correct scene. Within each block, each alien and action occurred an equal number of times; half the
utterances in each block were SOV, the other half OSV. In the distractor scene, no actions were the same and the aliens and their colours were randomly selected. The locations of the target scene were counterbalanced.
For testing, each type of information in the language was assessed by presenting an utterance and varying the target and distractor scenes by one piece of information: For testing nouns, target and distractor scenes were identical except for one of the aliens; for testing verbs, only the scenes' actions differed; for testing adjectives, one of the colours of an alien was changed; and for testing marker words, the two scenes depicted the same aliens performing the same actions but with opposite agent-patient assignment. Testing trials were intermingled with every third training block. The purpose of this was to make it less likely that participants would know they were being tested. For testing word order, grammatical and ungrammatical sentences were presented: Half the trials followed the licensed SOV or OSV order in sentences that had not been presented in the crosssituational learning trials, whereas the other half contained syntactic violations (*VSO, *VOS, *OVS, *SVO).

## Procedure

Participants were trained and tested on the artificial language on two days. Participants first completed 16 training and testing blocks. Twenty-four hours later, they returned to the lab to complete a delayed post-test. There were eight pure training blocks, four mixed training and testing blocks, and four grammaticality judgment test (GJT) blocks. In the cross-situational learning task, participants were instructed to observe the two scenes on the screen and listen to the sentence played over headphones. Their task was to decide, as quickly and accurately as possible, which scene the sentence referred to. Participants received no feedback regarding the accuracy of their choice. For the word order trials, participants were told that they would see only one scene and hear a sentence spoken by another alien from a very different planet who was also learning the new language. Their task was to listen carefully and decide if the sentence sounded "good" or "funny".
Presentation order of trials within each block was randomized but all participants completed blocks in the same sequence. There were two training blocks, then one mixed training and vocabulary testing block, then a word order test block. This sequence was then repeated four times.
The massed group completed the first 16 blocks consecutively while the distributed group had three 20minute breaks after every four blocks, in which they watched a natural history documentary on mute. Training and testing on day 1 took between 70 and 90 minutes for the massed group and between 130 and 150 minutes for the distributed group. The delayed test on day 2 comprised a final block of vocabulary testing trials and then a block of word order testing, and five cognitive tests (not reported here), and lasted approximately 90 minutes.

## Results

## Performance on training trials

We first performed a mixed analysis of variance on accuracy within each block, with training block as within subjects factor and the two training conditions (massed and distributed) as between subjects factor. There was a significant main effect for block, using the GreenhouseGeisser correction, $F(4.55,282)=42.0, p<.001, \eta_{\mathrm{p}}{ }^{2}=.40$. This indicates that subjects improved with more training. However, there was no significant main effect for group, nor was there a significant interaction between time and group, both $F<1$. We therefore pooled the data from the two training conditions for the remaining analyses.

In order to ascertain when learning had taken place during the training blocks, a one-sample t-test was conducted to compare the mean scores for each block to a chance score of .5. Participants performed significantly above chance from block two ( $M=.57, S D=.18$ ) onwards, $95 \%$ CI $[.028$ to .12], $t(63)=3.27, p=.002$. In other words, 32 trials of exposure (without feedback) were enough to lead to abovechance performance in the cross-situational learning task.

## Performance on test trials

In order to determine performance for each type of information in the language, we performed one sample ttests to establish the first test block at which accuracy was above chance (at .5). We then carried out a series of repeated measures ANOVAs in order to determine the effects of test block on the scores for word order, nouns, verbs, adjectives and markers. Finally, we conducted further repeated measures ANOVAs for test blocks 4 (immediate post-test) and 5 (delayed post-test) for each word type and word order to assess the role of the 24-hour delay. The results are displayed in Figure 2 and Table 1 and Table 2.

Table 1: Summary of repeated measures ANOVA over test blocks 1 to 4 showing effect for block.

| Test | F | p | $\mathrm{h}_{\mathrm{p}}{ }^{2}$ |
| :--- | :--- | :--- | :--- |
| Word order | 7.82 | $<.001$ | 11 |
| Noun | 14.5 | $<.001$ | .19 |
| Verb | 2.46 | .064 | .038 |
| Adjective | 2.76 | .043 | .043 |
| Case marker | .63 | .60 | .010 |

Table 2: Summary of repeated measures ANOVA over blocks 4 to delayed test block 5 showing effect for block.

| Test | F | P | $\mathrm{h}_{\mathrm{p}}{ }^{2}$ |
| :--- | :--- | :--- | :--- |
| Word order | .025 | .88 | .00 |
| Noun | 2.90 | .59 | .005 |
| Verb | 4.61 | .036 | .069 |
| Adjective | 2.50 | .12 | .040 |
| Case marker | 2.63 | .11 | .041 |



Figure 2: Proportion of correct trials across the five test blocks. Test blocks 1 to 4 were completed on day 1 . Test block 5 was administered with a 24 -hr delay.

Participants performed significantly above chance from test block 1 onwards for both the word order tests, ( $M=.76$, $S D=.19), 95 \%$ CI $[.22$ to .31$], t(63)=10.9, p=<.001$ and also verb tests, $(M=.70, S D=.25,95 \%$ CI [. 13 to .26$]$, $t(63)=6.16, p<.001)$. For noun tests, participants performed significantly better than chance from test block 2 onwards, $(M=.60, S D=.19,95 \%$ CI [.052 to .15$], t(63)=$ 4.16, $p<.001$ ). Adjective test results were significantly above chance from test block 4 onwards, ( $M=.64, S D=$ .27), $95 \%$ CI [.076 to .21$], t(63)=4.22, p<.001)$. Finally, case markers only reached significantly above chance on test block 5, ( $M=.54, S D=.16$ ), $95 \%$ CI [.001 to .079$]$, $t(63)=2.06, p=.043$.

## Determining relations between learning different information types

In order to determine the factors driving performance in the task - whether learning was independent or interdependent for different types of information, we conducted a principal components analysis on test performance for the final test block for word order, nouns, verbs, adjectives, and marker words. There were two components with eigenvalues greater than 1 , and the loadings of the individual tests on these components, with varimax rotation, showed a simple solution (i.e., each test loaded $>0.4$ on only one component). The components and their loadings are shown in Table 3.
The first component related to learning nouns, adjectives, and marker words, and the second component related to learning word order and verbs. This indicated that performance across the five information types was effectively explained by two aspects of the data: The first relates to learning the vocabulary items of nouns and

Table 3: Loadings of the five delayed tests on the two principal components.

| Test | First component | Second component |
| :--- | :--- | :--- |
| Noun | $\mathbf{. 7 7 8}$ | .104 |
| Adjective | .769 | .034 |
| Marker words | $\mathbf{. 6 0 4}$ | .081 |
| Verb | .322 | .718 |
| Word order | -.090 | $\mathbf{. 8 7 3}$ |

adjectives and how the marker words affected the role of the adjective-noun phrases, and the second indicated a close relation between learning the identities of verbs and learning that the word order of sentences was verb-final.

## Discussion

In this study we investigated whether adult learners could acquire the syntax and vocabulary of a novel language by keeping track of cross-trial statistics, without feedback and without any explicit instruction about the structure of the language or its vocabulary. We also provided a delayed post-test after 24 hours to determine whether any acquired knowledge had been maintained. Furthermore, we examined the order of acquisition and investigated how learning of syntax and of vocabulary cohered.

## Simultaneous learning of words and syntax

Our results indicated that adult learners can rapidly acquire both syntax and vocabulary of the language simultaneously. Previous cross-situational learning studies only investigated nouns (Smith \& Yu, 2008) or verbs (Scott \& Fisher, 2012) or nouns and verbs simultaneously (Monaghan et al., 2015). Our results extend these findings to demonstrate that it is possible for adults to acquire a wider range of information, including adjectives and case markers. We cannot rule out, however, that this occurred because of the nature of the lexical test design, in which the two scenes presented differed only in terms of the lexical item being tested, artificially making these word categories more salient. It also remains to be seen whether children can also learn such a complex system via cross-situational learning, and this is an important question that we are currently addressing. The results also show that the learning effects can be retained overnight. This is an important methodological observation as the majority of studies in statistical learning do not have a delayed post-test, which means that it is unclear whether the learning is robust. By including a $24-\mathrm{hr}$ delayed post-test, we show that learning is indeed robust and that this applies to words and syntax.

For the case markers, it was only after 24 hours that test scores were significantly above chance. This corresponds with Grey, Williams and Rebuschat's (2015) study that found no learning effect for Japanese morphology on an immediate post-test, but a significant effect after a two-week delay. These findings suggest that consolidation may be valuable, particularly for the function words' role in the language. Indeed, there is evidence that sleep aids in the
generalization of grammatical rules (Walker \& Stickgold, 2010). The case marker results also raise another important methodological consideration. Without the delayed test, we would have underestimated the amount of learning that had taken place and would have concluded that case markers had not been learned at all. Whereas, with the delayed test, there is evidence, albeit a small effect, that learning of case markers does in fact take place. It is recommended, therefore, that future studies into cross-situational learning include delayed post-tests to show that learning is robust and to catch any learning effects brought on through consolidation.

## Order of acquisition

Although learners were exposed to both vocabulary and syntax simultaneously, they performed above chance on different aspects of the language at distinct stages: First, verbs and word order were acquired, then nouns, then adjectives, and finally case markers (see Figure 2). It is interesting to note that verbs were learned before nouns in this artificial language and thus differed from the majority of first language acquisition studies. One possible reason for this is the saliency of the final-position verb compared to the mostly medial-position nouns (Shoemaker \& Rast, 2013). Another possibility is that adult learners already possess syntactic and lexical knowledge of word categories in their L1 and so can transfer them to their L2. This would then allow the learner to concentrate on deciding which words map onto the different lexical categories, rather than also working out the lexical categories as infants do.

## The coherence of vocabulary and syntax

Regarding the coherence of learning of syntax and vocabulary, we found that acquisition of word order and verb learning were interdependent. Upon learning that the final word in the sentence was a verb, participants were able to gain an understanding of the basic word order of the sentence. It is conceivable that such an understanding could be gained by breaking into the stream of input through any word category, with the greater salience of verbs due to final utterance position promoting this acquisition. In addition, we found that nouns, adjectives and case markers were also interdependent but acquired somewhat independently of verbs and word order. This result supports an emergentist view that syntactic knowledge associated with case markers develops only after a core vocabulary of content words has been learned (Bannard et al., 2009).

These results demonstrate that the chicken and egg problem of acquiring grammar and vocabulary can be resolved by the learner through using cross-situational statistics with events in the environment. An alternative explanation is that once verbs are learned, this knowledge is then bootstrapped to aid the acquisition of the other lexical categories. The patterns of results we found for this verbfinal language in our experimental paradigm did not neatly correspond with a distinction between grammar and vocabulary learning (e.g., Ullman, 2004), with word order
being related to verb acquisition, and case marking being related to noun and adjective learning. Complex interactions between grammatical categories and grammar do not appear to lend themselves to a neat distinction in acquisition of these sources of linguistic knowledge.

## References

Alario, F., \& Cohen, L. (2004). Closed-class words in sentence production: Evidence from a modality-specific dissociation. Cognitive Neuropsychology, 21(8), 787-819.
Bannard, C., Lieven, E., \& Tomasello, M. (2009). Modeling children's early grammatical knowledge. Proceedings of the National Academy of Sciences, 106(41), 17284-17289.
Bates, E., \& Goodman, J. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia, and real-time processing. Language and Cognitive Processes, 12, 507-584.
Behrens, H. (2006). The input-output relationship in first language acquisition. Language and Cognitive Processes, 21(1-3), 2-24.
Bock, K., \& Levelt, W. (1994). Language production: Grammatical encoding. Handbook of psycholinguistics, ed. by Morton Ann Gernsbacher, 945-84.
Booth, A. E., \& Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. Child Development, 80(1), 15-22.
Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S., Pascual, L., \& Vyt, A. (2004). Cross-Linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. Child Development, 75(4), 1115-1139.
Choi, S., \& Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. Journal of child language, 22(03), 497-529.
DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. Language Learning, 55(S1), 1-25.
Dixon, R. M. (1982). Where have all the adjectives gone?: and other essays in semantics and syntax (Vol. 107). Walter de Gruyter.
Frost, R. L., \& Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. Cognition, 147, 70-74.
Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. Technical report no. 257.
Gleitman, L. (1990). The structural sources of verb meanings. Language Acquisition, 1(1), 3-55.
Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., \& Trueswell, J. C. (2005). Hard words. Language Learning and Development, 1(1), 23-64.
Göksun, T., Küntay, A. C., \& Naigles, L. R. (2008). Turkish children use morphosyntactic bootstrapping in interpreting verb meaning. Journal of Child Language, 35(02), 291-323.

Grey, S., Williams, J. N., \& Rebuschat, P. (2015). Individual differences in incidental language learning: Phonological working memory, learning styles, and personality. Learning and Individual Differences, 38, 4453.

Gullberg, M., Roberts, L., \& Dimroth, C. (2012). What word-level knowledge can adult learners acquire after minimal exposure to a new language? International Review of Applied Linguistics in Language Teaching, 50(4), pp. 239-276
Indefrey, P., \& Gullberg, M. (2010). The earliest stages of language learning: Introduction. Language Learning, 60(s2), 1-4.
MacWhinney, B. (1987). The competition model. Mechanisms of Language Acquisition, 249-308.
Marcus, G. (1996). Why do children say "breaked"? Current Directions in Psychological Science, 5, 81-85.
Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. Action Meets Word: How Children Learn Verbs, 31-63.
Monaghan, P., \& Mattock, K. (2012). Integrating constraints for learning word-referent mappings. Cognition, 123(1), 133-143.
Monaghan, P., Mattock, K., Davies, R. A., \& Smith, A. C. (2015). Gavagai is as gavagai does: Learning nouns and verbs from Cross-Situational statistics. Cognitive Science, 39(5), 1099-1112.
Scott, R. M., \& Fisher, C. (2012). 2.5-year-olds use crosssituational consistency to learn verbs under referential uncertainty. Cognition, 122(2), 163-180.
Shoemaker, E., \& Rast, R. (2013). Extracting words from the speech stream at first exposure. Second Language Research, 29(2), 165-183.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. Cognition, 92(1), 231-270.
Walker, M. P., \& Stickgold, R. (2010). Overnight alchemy: sleep-dependent memory evolution. Nature Reviews Neuroscience, 11(3), 218-218.

# Please Explain: Radical Enactivism and its Explanatory Debt 

Lachlan Douglas Walmsley (lachlan.walmsley@anu.edu.au)<br>School of Philosophy, 9 Fellows Road<br>Canberra, ACT 2602 Australia


#### Abstract

Radical Enactivism is a position in the philosophy of cognitive science that aims to displace representationalism, the dominant position in cognitive science for the last 50-60 years. To accomplish this aim, radical enactivism must provide an alternative explanation of cognition. Radical enactivism offers two alternative explanations of cognition. The first I call the dynamical explanation and the second I call the historical explanation. The mechanists have given us reasons for doubting that the first alternative makes for a good explanation. The historical explanation does not hit the right explanatory target without the introduction of a proximate mechanism, but the proximate mechanisms suggested by radical enactivism are associationist mechanisms, the limitations of which led to the initial widespread endorsement of representationalism. Therefore, radical enactivism cannot displace representationalism in cognitive science.


Keywords: radical enactivism, representation, dynamical explanation, computationalism, explanation

## 1 Introduction

Radically enactive cognition (REC) is a position in the philosophy of cognitive science aiming to displace representationalism (Hutto \& Myin 2013), the dominant position in cognitive science for the last $50-60$ years. ${ }^{1}$ To accomplish this aim, proponents of REC-or RECersmust settle their explanatory debt by providing us with an alternative explanation of cognition. Cognition, here, is understood in a non-question begging biological sense as that system functioning to coordinate behaviour intelligently (Godfrey-Smith 1996). ${ }^{2}$ It's a familiar phenomenon. It's what happened when you wrote your last paper or organised your last workshop. It's probably what happened when you performed some less sophisticated tasks too. The RECers owes us an explanation of that. For half a century, the representational explanation of cognition has been the defining explanation of cognitive science-"the only game in town" (Fodor 1975). But the RECers argue that the representational explanation is bad. If we can't play the representation game anymore what game shall we play?

REC is committed to two alternative explanations of cognition. The first, I call the dynamical explanation. Here, RECers presuppose that all dynamical models make for good explanations, and David Kaplan and William Bechtel

[^232]have given us good reasons to think otherwise (Kaplan 2015; Kaplan \& Bechtel 2011). Although many of the dynamical models appealed to by the RECers provide elegant and predictive descriptions of phenomena, they do not explain those phenomena. Their second alternative, I call the historical explanation. Although this is a good explanation, it is, by itself, not the right kind of explanation to compete with the representational explanation. Even with the addition of associationist mechanisms, the explanation still fails to explain certain intelligent behaviour, a limitation that led to the initial widespread endorsement of the representational explanation in cognitive science. Therefore, RECers owe us an explanation of cognition that can displace the representational explanation. ${ }^{3}$

## 2 The Representational Explanation

According to the representational explanation, intelligent behaviour is coordinated through the manipulation and transformation of information-bearing structures called representations. There are a number of versions of the explanation, however, owing to the different ways in which the term "representation" has been used. William Ramsey identifies four ways in which the term has been used (Ramsey 2007). Here, I defend only the representational explanation of the classical computational theory of cognition (CCTC), so only two of the four notions of representation Ramsey identifies are relevant: the $I O$-notion and $S$-representation. For brevity's sake, I will discuss only the IO-notion (Ramsey 2007: 68-77). ${ }^{4}$

The IO-notion of representation is used to describe those situations in which some structure standing in for another is taken as a system's input (hence the "I") and transformed into another structure (again standing in for yet another structure), which is its output (hence the "O"). If this is a little abstract, imagine a calculator taking as input some structures standing in for numbers and mathematical operators-say, " 2 ," " 3 ," " + ," and " $=$." All going well, the calculator transforms its input into another structure, "5," which stands in for the number five, and which the

[^233]calculator gives as its output. Here's another example. Imagine a face recognising computer. Again, its input is a structure standing in for another structure, a face in this case. Let's say its input is a portrait of Charles Darwin. All going well, the computer transforms its input and gives as output a different structure that stands in for the individual to whom the face belongs. Something like "Charles Darwin." CCTC models the transformation of a structure from input to output as a series of operations carried out by component sub-systems. Each sub-system also takes as its input structures standing in for other structures and gives as its output yet more such structures. It is due to this on-going replacement and its role in the transformation of system input to system output that CCTC takes such systems to manipulate and transform structures standing in for other structures-that is, representations. ${ }^{5}$

If you are sceptical of CCTC's representational explanation then consider the digital computer. The digital computer can perform calculation and facial recognition as in the examples above and does so in the way I've described. The digital computer is a paradigm case of a physical system for which the representational explanation is good. It is, in effect, an existence proof for good representational explanations. No matter what position you occupy in the philosophy of cognitive science, if it follows from your argument that the representational explanation of a digital computer is bad, then that is a reductio against your argument. Whether or not the representational explanation is good for human cognition is an empirical question. But RECers argues that the representational explanation of human cognition is bad on theoretical grounds.

### 2.1 The Cost of the Representational Explanation

RECers claim that we should reject the representational explanation because it is too metaphysically demanding to be naturalised. ${ }^{6}$ A naturalistic explanation, here, is understood as one that can be squared with our current scientific knowledge-no spooky stuff. The representational explanation is too metaphysically demanding, REC claims, because it posits the existence of content, which determines what a structure stands for: the structure " 5 " stands in for the number five because of its content. According to REC, no naturalistic account of content has succeeded in explaining the "special properties," such as "truth, reference, implication," attributed to content (Hutto \& Myin 2013: 67). These properties make content, and hence the

[^234]representational explanation, "too metaphysically extravagant to be accepted by hard- nosed naturalists" (21).

As a naturalist, I shy away from metaphysical extravagance, and I agree that no naturalistic account of representation has explained content as REC understands it. But I resist the assumption that content must have such metaphysically demanding properties, such as truth and reference. One reason why REC might make this assumption is because their emphasis is partly on mentality and the mind: "Enactivism is inspired by the insight that the embedded and embodied activity of living beings provides the right model of understanding minds" (Hutto \& Myin 2013: 4, my emphasis). The focus of the representational explanation radical enactivists hope to displace, however, is not on the mind but on cognition. It may prove difficult to give good naturalistic explanations of the mind's features because the mind simply doesn't lend itself to good naturalistic explanation. But I will leave that to the philosophers of mind.

Can we make do with less metaphysically demanding accounts of representational in cognitive science? I think we can. Furthermore, I think REC must also make do with less metaphysically demanding accounts of representation because they are committed to the existence of public representational systems, such as public language. How does " 5 " come to stand in for (at least in most instances) the number five in our public language? Radical enactivists cannot answer that " 5 " stands for the number five because of representations internal to the language users. The meaning of public representations can't be due to something metaphysically extravagant in your head or in mine. Such an answer is anathema to RECers. Instead, public structures like " 5 " come to stand in for what they do in virtue of the interactions between public language users (Hutto 2008). This was, roughly, Wittgenstein's view of language in his Philosophical Investigations. According to this view, "5" stands in for the number five because we use that structure, either as a written symbol or as an utterance, in those situations involving five-type things, such as when we ask someone to fetch five stones or to wait five days. Over time, language users become expert at recognising situations like these and can use " 5 " in situations involving much more complicated or abstract entities, like dollars and electrons. For Wittgenstein, structures stand in for other structures (in most cases) in virtue of their functional role in a system of public language users. In slogan form: "the meaning of a word is its use in the language" (Wittgenstein 2009/1953: $\S 43)$. So " 5 " or "five" means (has as its content) the number five because of the way in which " 5 " is used in a community of language users. ${ }^{7}$

REC is committed to something like Wittgenstein's meaning-as-use account. A metaphysically undemanding explanation of representation such as this can generalise to cognitive systems. Just as utterances can be said to stand in for other structures in virtue of their functional role in a

[^235]public system, so can structures in a computational system. These structures can be said to stand in for others in virtue of how they are passed between the different sub-systems of the computational system and how they are transformed by each sub-system-that is, how they are used within and by the system. You might be sceptical of a Wittgensteinflavoured meaning-as-use account of representation. And it's fine if you are. All that's important here is that REC is committed to an explanation of interactions between organisms-that is, human public practices-that involves representation and avoids metaphysical extravagance. If public practice can be given a metaphysically undemanding representational explanation then so can cognition. There is no reason to suppose that representational explanations of intraorganism processes will be any more metaphysically extravagant than those for interorganism processes.

Advancing the representational explanation, at least as it is understood in CCTC, does not entail positing anything as problematic as REC suggests it does. As I outlined above, within CCTC we can posit representations and explain representational content in virtue of a structure's function within the computational system. A structure stands in for what it does because of the role that structure plays in the computational system. Furthermore, this way of grounding content is analogous to REC's strategy for grounding the content of public representations. So, if REC remains sceptical of CCTC's representational explanation, as I have outlined it in this section, then they must rethink their own commitments to the existence of public representational systems.

Furthermore, according to CCTC's representational explanation, structures in a computational system are not causally efficacious because of their content. They are causally efficacious because of how their formal properties map onto physical parts of the target system, such as transistors or neurons (Fodor 1981; Pylyshyn 1984; Gallistel and King 2009). If the RECers suppose the representational explanation requires that structures be causally efficacious because of their content then they are setting up a straw person. Hence, I take it this is not their position. If the RECers accept CCTC's representational explanation but reject other versions of the representational explanation, such as those according to which structures are causally efficacious in virtue of their content, then they are conservatives, classical computationalists rather than radicals. Hence, I take it this is not their position. Instead, I take it that REC is a position according to which CCTC's representational explanation is unnecessary for explaining any intelligent behaviour. This is an empirical question. And we have good reasons to answer it in favour of the representational explanation (see $\S 4.1$ especially). From here on, I argue that REC's candidate replacement explanations are not genuine alternatives to the representational explanation of CCTC or, where they are genuine alternatives, they fail to explain some intelligent behaviour that the representational explanation can.

## 3 The Dynamical Explanation

The first alternative to the representational explanation offered by REC is the dynamical explanation: "the vast sea of what humans do and experience is best understood by appealing to dynamically unfolding, situated embodied interactions and engagements with worldly offerings" (Hutto \& Myin 2013: ix). Dynamical explanations are constructed with the language of dynamical systems theory, which models how physical systems change over time with differential and difference equations. These equations quickly become analytically intractable as structures or details are added but their solutions can be satisfactorily approximated using numerical methods and computer simulation. From the approximate solutions, modellers create geometric visualisations of the different ways in which the system can change over time as transitions through a state space. The explanation's language is complex, but it need not concern us here. The problem with dynamical explanations is not the language in which they are described.

A paradigm case of a dynamical model is the Haken-Kelso-Bunz (HKB) model of human hand movements (Haken et al. 1985). The model captures "voluntary oscillatory motions of the two index fingers"-that is, the movement of your index fingers when you move them back and forth in a coordinated fashion, either symmetrically or asymmetrically. In particular, it captures the abrupt change from asymmetrical and symmetrical coordination when the oscillations reach a certain frequency. HKB is a phenomenological model, built to have a close qualitative fit with the system's behaviour: the "first step in the development of the model is to provide a mathematically accurate description of the main qualitative features of the data" (349).

HKB has received much attention because it is a minimal model of a "relatively simple two-component system" (Bressler \& Kelso 2001: 28) with predictive power, capturing the dynamics of a wide range of interactions including those between an agent and their environment (Kelso 1994) and between two agents (Schmidt et al. 1990). For RECers and similarly inclined anti-representationalists, models like this provide good alternative explanations because they make simple and generalisable predictions (without positing content): "If models are accurate enough to describe observed phenomena and to predict what would have happened had circumstances been different, they are sufficient as explanations" (Chemero \& Silberstein 2008: 12). Although accurately describing the behaviour of a large class of systems is a virtue of these models, good descriptions and predictions are not sufficient for explanation.

Dynamical explanations of the intelligent coordination of behaviour can only be genuine alternatives to the representational explanation if they are genuine explanations. Dynamical explanations are only genuine explanations if the predictions and descriptions of behaviour offered by models like HKB are also explanatory.

Description and prediction are certainly similar to explanations. The covering law account of explanation, for example, treats them as having the "same logical character" as each other (Hempel 1958: 37; 1965). But they are importantly different from explanation (Kaplan 2015; Kaplan \& Bechtel 2011). The difference between description and explanation is obvious: a description of a phenomenon is simply a statement of the explanandum. The difference between prediction and explanation is not so obvious but just as real.

To see the difference between prediction and explanation, imagine a flagpole, which casts a shadow as the sun rises and sets. As the position of the sun changes, so does the shape and size of the shadow. The two change together with law-like regularity. Hence, we can use the height of the flagpole along with the position of the sun and some mathematics to predict the shape and size of the shadow. We can also use the shape and size of the shadow along with the position of the sun and some mathematics to predict the height of the flagpole, but we cannot explain the height of the flagpole in virtue of the shadow's shape and size. Although predictions can run either way, from flagpole to shadow and from shadow to flagpole, explanations run in only one direction - in this case, from flagpole to shadow (Bromberg 1966; also Kaplan \& Bechtel 2011: 440-441). Explanations must inform us of that which gives rise to a phenomenon, so an explanation of the height of the flagpole would appeal to the factory in which it was made, but not to its shadow.

Precise mathematical models of behaviour like HKB are not the explanans. ${ }^{8}$ They are the explanandum. This is not especially controversial; dynamical modellers themselves are aware of this. Haken et al., for example, admit of their model that it describes the coupling between the two hands but says nothing about what gave rise to that coupling and leave this for "further theoretical and experimental research" (Haken et al. 1985: 355). Short an actual explanation of the phenomena described by HKB, Haken et al. provide a howpossible explanation, describing a mechanism that might be responsible for causing the phenomena: "one coupling might be established via the corpus callosum, the wellknown band of fibres that joins the two hemispheres of the brain" (ibid.). Another explanation of the regularities described by dynamical models like HKB may involve the manipulation and transformation of information-bearing structures. The behaviour of digital computers, for example, can be modelled using the tools of dynamical systems theory, but, as I said above, it is a paradigm case of a system for which the representational explanation is good. Hence, dynamical models of cognition-explanatory or not-are compatible with representational explanations. Even if

[^236]REC's dynamical explanation were genuinely explanatory it would be compatible with representational explanation.

## 4 The Historical Explanation

REC's second alternative to the representational explanation is the historical explanation. In this case, cognition is explained in virtue of an agent's "history of previous engagements and not in some set of internally stored mental rules and representations" (Hutto \& Myin 2013: 9). To make this concrete, imagine some behaviour:
"Someone is living in a house with a kitchen in the hallway, such that she has to walk around a sideboard to get to the other side. Suppose that at some point the sideboard gets removed, but that the person still takes the same curve to get to the other side of the hall." (Degenaar \& Myin 2014: 3642)

Here is the historical explanation of that behaviour:
"In the new situation, the person is going through the same old motions in absence of the environmental basis for these motions. Over the years, a behavioural pattern has emerged: the person tends to take a particular trajectory when walking through the hallway." (Degenaar \& Myin 2014: 3642)

The historical explanation is neither mere prediction nor description. Unlike REC's first alternative explanation, its second is genuinely explanatory. However, it is still not a genuine alternative to the representational explanation. Rather, it is compatible with the representational explanation. As Jan Degenaar and Erik Myin say of the above example, "This might involve representations or it might not" (Degenaar \& Myin 2014: 3642).

The historical explanation is not the right kind of explanation to be an alternative to the representational explanation because the historical explanation is an ultimate explanation, while CCTC's representational explanation is a proximate explanation. Niko Tinbergen (1963) first made the distinction between proximate and ultimate explanations. An example will help illustrate the distinction: humans regularly help needy others at a cost to themselves. One explanation of this behaviour is that empathising with needy others motivates us to help them (Batson 2011). This is a proximate explanation. It tells you about the mechanism here and now-empathy-that produces the helping behaviour. But why this sort of mechanism? Why are we empathetic? This question calls for an ultimate explanation, which might explain the helping behaviour as the result of selection for a particular behavioural disposition in terms of benefits to an organism or group's fitness. For example, perhaps our empathetic ancestors were better carers for their and their kin's young, so our empathetic ancestors did better than our nonempathetic ones and empathy spread through the population (De Waal 2008). Importantly, ultimate explanations need not refer to evolution. They can also refer to an agent's developmental history (Baum 1994). For example, an ultimate explanation of an agent without the
disposition to help may be that their helping was rewarded materially early in their developmental history such that the agent came to expect material rewards to follow from helping (Warnaken \& Tomasello 2014). Hence, when there are no material rewards on the horizon, the agent doesn't help. Here, our ultimate explanation refers to a learning process rather than an evolutionary one.

Proximate and ultimate explanations are natural partners, with one explaining the mechanism producing the behaviour here and now and the other explaining why that kind of mechanism exists instead of another. Since REC's historical explanation is compatible with the representational explanation it is no real alternative at all.

### 4.1 The Proximate Explanation

RECers might respond that their historical explanation does involve a proximate mechanism, which is something akin to what Kim Sterelny (2003) calls a detection system. Detection systems link certain environmental stimuli with certain behavioural responses and do not involve the manipulation and transformation of information-bearing structures. ${ }^{9}$ In the example above, entering the hallway has been linked with the response of taking a particular trajectory. This connection has been wired up through something like simple association-based learning-the hallway becomes associated with taking the trajectory in virtue of certain rewards, such as not bumping into the sideboard. But detection systems can also be wired up through evolutionary processes. Organisms can be born responding to particular stimuli with particular responses because such organisms have had greater reproductive success. The infamous male Photuris firefly, for example, is born with such a detection system, which links a certain series of flashes with the response of flying toward the source of the flashes. Female Photuris fireflies produce these flashes and males find them, so the two can mate.

Although often effective, detection systems are fragile. The Photinus firefly's detection system is exploited by the Photuris firefly. They produce the flashing just like the Photinus females, catching and eating a fair number of unfortunate males (Lloyd 1965). Sterelny's robust systems are, as the name suggests, less fragile than detection systems. These link a number of environmental stimuli with a particular response. But even these have their limits. Once the causal chain through which a relevant aspect of an environment and an appropriate behavioural response are linked becomes sufficiently complex and rare, it becomes invisible to whatever processes build detection and robust systems, such as associationist learning or evolution by selection. Sterelny argues that this is the case in complex social environments in which deception is common and multi-place relations between group members matter, and from which language can emerge as it has in the case of human lineage. Explaining human cognition, then, will

[^237]require a proximate explanation appealing to more than detection systems and associationist learning.

If the only good proximate explanation RECers have up their sleeves is one involving associationist mechanisms linking a stimulus with response in virtue of a history of interaction then that's a problem. As Sterelny argues, the complexities of social life are such that stimulus-response systems like Photinus's just won't do. But REC needs more than a good proximate explanation for human social behaviour (here, RECers will argue is richly scaffolded by shared practices and hence not as computationally demanding as it seems). REC also needs one for the behaviour of much simpler organisms, such as insects (Gallistel 1990, Gallistel \& King 2009). Desert ants, among other insects, often trace winding paths away from their nests as they forage. Upon returning to their nests, they don't retrace their steps, but take an almost direct route. This is known as path integration or dead reckoning. It requires integrating information both about the distance and direction travelled from the nest. This ability has been experimentally demonstrated (Wehner \& Srinivasan 1981). In Wehner and Srinivasan's experiment, ants forage from their nest to a feeder station 20 m away. Upon reaching the feeder station, the ants are transferred to a test area several hundred meters away with a replica feeder station. From this replica, the ants take a direct path to where their nest should be. When they reach this point they begin searching for their nest in different directions. In these experiments, the ants are clearly not using environmental cues. If they were, they would find their actual nest, not where their nest should be. Their destination is a novel location, so they cannot be navigating by anything like habit. In this instance, there is no history of on-going interactions to appeal to and no stable environmental stimuli with which behaviour can be associated. A good explanation is one positing a computational process involving representations of the ants' location relative to the nest and feeder station.

## 5 Revolution?

The radical enactivists owe me an explanation. They owe you one too. They owe us all an explanation of how biological systems like you and I behave in the complicated ways we do. They owe us an explanation of cognition. Importantly, the explanation cannot one of those advanced before the cognitive revolution. You cannot displace the representational explanation with stimulus-response mechanisms because the representational explanation initially gained traction in virtue of the limitations of such mechanisms.

So what new explanations are on offer? There is the dynamical explanation, according to which our actions are the results of dynamically unfolding interactions with our environments. But what explains why these dynamics obtain instead of others? When I become reciprocally coupled with my environment, what initiates and maintains that coupling such that my behaviour can be predicted with a set of elegant differential equations? Although the models of the
dynamical explanation can offer mathematically precise descriptions of behaviour, they don't explain why those descriptions hold. If those in the radical camp want an antirepresentational revolution, they must fill the explanatory gap left by the representational explanation. So far, they have failed to do this.

## Acknowledgments

Thank you to Ronald Planer and David Kalkman for their comments on drafts of this paper. Thank you also to three anonymous reviewers whose comments were both fair and insightful. Funding was provided by Australian Research Council (Grant No. ARC FL13).

## References

Batson, C. D. (2011). Altruism in humans. Oxford: Oxford University Press.
Baum, W. M. (1994). Understanding behaviorism: Science, behavior, and culture. New York: Harpercollins College Division.
Bressler, S. L. \& Kelso, J. A. S. (2001). Cortical coordination dynamics and cognition. Trends in cognitive science, 5, 26-36.
Bromberger, S. (1966). Why questions. In R. G. Colodney (ed.). Mind and Cosmos. Pittsburgh, PA: University of Pittsburgh Press.
Chemero, A. \& Silberstein, M. (2008). After the philosophy of mind: replacing scholasticism with science. Philosophy of Science, 75, 1-27.
De Waal, F. B. M. (2008). Putting the altruism back into altruism: the evolution of empathy. Annual Review of Psychology, 59, 279-300.
Degenaar, J. \& Myin, E. (2014). Representation-hunger reconsidered. Synthese, 191, 3639-3648.
Fodor, J. A. (1975). The Language of Thought, Cambridge, MA: Harvard University Press.
Fodor, J. A. (1981). Representations. Cambridge, MA: MIT Press.
Gallistel, C. R. \& King, A. (2009). Memory and the Computational Brain. Malden: Wiley-Blackwell.
Godfrey-Smith, P. (1996). Complexity and the Function of Mind in Nature. Cambridge: Cambridge University Press.
Haken, H., Kelso, J. A. S., \& Bunz, H. (1985). "A theoretical model of phase transitions in human hand movements," Biological Cybernetics, 51, 347-356.
Hempel, C. G. (1958). The theoretician's dilemma: A study in the logic of theory construction. Minnesota Studies in the Philosophy of Science, 2, 173-226.
Hempel, C. G. (1965). Aspects of scientific explanation. Aspects of Scientific Explanation; And other Essays in the Philosophy of Science. New York: The Free Press.
Hutto, D. D. (2008). Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons. Cambridge, MA: MIT Press.
Hutto, D. D. \& Myin, M. (2013). Radicalizing Enactivism: Basic Minds Without Content. Cambridge, MA: MIT Press.

Kaplan, D. M. (2015). Moving parts: the natural alliance between dynamical and mechanistic modeling approaches. Biology \& Philosophy, 30, 757-786.
Kaplan, D. M. \& Bechtel, W. (2011). Dynamical models: an alternative or complement to mechanistic explanations? Topics in Cognitive Science, 3, 438-444.
Kelso, J. A. S. (1994). Elementary coordination dynamics. In S. P. Swinnen, H. Heuer, J. Massion, \& P. Casaer (eds.). Interlimb Coordination: Neural, Dynamical, and Cognitive Constraints. San Diego: Academic Press: 301318.

Lloyd, J. E. (1965). Aggressive mimicry in Photuris: firefly femmes fatales. Science, 149, 653-654.
Pylyshyn, Z. (1984). Computation and Cognition. Cambridge, MA: MIT Press.
Ramsey, W. M. (2007). Representation Reconsidered. Cambridge: Cambridge University Press.
Schmidt, R. C., Carello, C., \& Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. Journal of Experimental Psychology: Human Perception and Performance, 16, 227.
Sterelny, K. (2003). Thought in a Hostile World: The Evolution of Human Cognition. Malden: Blackwell Publishing.
Stich, S. P. (1983). From Folk Psychology to Cognitive Science: The Case Against Belief. Cambridge, MA: MIT Press.
Tinbergen, N. (1963). On aims and methods of ethology. Zeitschrift für Tierpsychologie, 20, 410-433.
Warneken, F. \& Tomasello, M. (2014). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. Motivation Science, 1, 43-48.
behaviour of desert ants, genus Cataglyphis (Formicidae, Hymenoptera). Journal of Comparative Physiology, 142, 315-338.
Wittgenstein, L. (2009/1953). Philosophical Investigations. Trans. G. E. M. Anscombe, P. M. S. Hacker, \& J. Schulte. 4th ed. Malden: Wiley-Blackwell.

# Learning to reinforcement learn 

Jane Wang<br>DeepMind<br>Zeb Kurth-Nelson<br>DeepMind, Max Planck-UCL Centre for Computational Psychiatry<br>Hubert Soyer<br>DeepMind<br>Joel Leibo<br>DeepMind<br>Dhruva Tirumala<br>DeepMind<br>Remi Munos<br>DeepMind<br>\section*{Charles Blundell}<br>DeepMind<br>Dharshan Kumaran<br>DeepMind, Institute of Cognitive Neuroscience, UCL<br>Matt Botvinick<br>DeepMind, Gatsby Computational Neuroscience Unit, UCL


#### Abstract

In recent years deep reinforcement learning (RL) systems have attained superhuman performance in a number of challenging task domains, but are constrained by a demand for large training sets. A critical present objective is thus to develop deep RL methods that can adapt rapidly to new tasks. In the present work we introduce a novel approach to this challenge, which we refer to as deep meta-reinforcement learning. Previous work has shown that recurrent networks can support metalearning in a fully supervised context. We extend this approach to the RL setting. What emerges is a system that is trained using one RL algorithm, but whose recurrent dynamics implement a second, quite separate RL procedure. This second, learned RL algorithm can differ from the original one in arbitrary ways and exploit structure in the training domain. We unpack these points in five proof-of-concept experiments to examine key aspects of deep meta-RL.


# The Learning of Subordinate Word Meanings 

Hao Wang<br>University of Pennsylvania<br>Lila Gleitman<br>University of Pennsylvania<br>John Trueswell<br>University of Pennsylvania


#### Abstract

In three experiments, adults attempted to learn words with subordinate-level meanings (dalmatian) by sampling the referent world cross-situationally. Xu \& Tenenbaum, 2007 predicted that encountering three uses of a word, each referring to a dalmatian would evoke "suspicious coincidence" inferencing, leading to the subordinate meaning (dalmatian). Exp. 1 found little evidence for this; cross-situational exposure led to a basic-level bias. This bias was unchanged even when the sample was increased to five subordinate exemplars (Exp. 2). Exp. 3 encouraged semantic contrast by simultaneously teaching each subject a word for the subordinate-level and the basic-level category within the same semantic domain (dap=dalmatian; blit=dog). Participants now showed non-basic level learning, but more in line with mutual exclusivity: they may think "dap" means dalmatian but "blit" means all-dogs-except-dalmatians. We conclude that the basic-level interpretation is powerful and cannot be removed by the mere observation of exemplar items over multiple word instances.


# Anticipatory Synchronization in Artificial Agents 

Auriel Washburn (washbual@mail.uc.edu)<br>Center for Computer Research in Music and Acoustics, Department of Music, Stanford University Stanford, CA 94305 USA

Rachel W. Kallen (rachel.kallen@uc.edu), Maurice Lamb (lambmi@ucmail.uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati Cincinnati, OH 45221 USA

Nigel Stepp (ndstepp@hrl.com)
HRL Laboratories
Malibu, CA 90265 USA

Kevin Shockley (kevin.shockley@uc.edu) and Michael J. Richardson (michael.richardson@uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati Cincinnati, OH 45221 USA


#### Abstract

By integrating theories and methodologies from a diverse range of scientific disciplines (e.g., physics, neuroscience, cognitive science, psychology and robotics engineering) the present work is aimed at harnessing self-organized anticipatory synchronization in order to advance humanrobotic interaction (HRI). This phenomenon is characterized by the emergence of anticipatory behavior by one system coupled to the chaotic behavior of another, following the introduction of short self-referential delays in the coordinating system. The current set of studies involved the creation of an artificial agent based on a time-delayed, low-dimensional dynamical model capable of behaving prospectively during an interaction with a human actor performing complex, unpredictable behaviors. By achieving characteristics similar to those observed during natural human interaction and coordination, the time-delayed modeling approached advocated here provides the potential for considerable future advancements in HRI.


Key words: human-robotic interaction; artificial agents; dynamical modeling; virtual reality; anticipatory synchronization; interpersonal coordination; chaos

Rapid advances in cyber-technologies and robotics present increasing opportunities for the implementation of interactive, artificial agents within contexts of human behavior. This includes, but is not limited to, assistance during the performance of everyday tasks and the development of new skills. Work has already been done, for example, on the development of virtual agents able to assist elderly individuals with the organization of their daily activities (Yaghoubzadeh et al., 2013), and to create a robot whose structured interaction may help to improve interpersonal coordination in children with autism spectrum disorders (Palatinus, 2014). However, Lorenz and Hirche (2014) have recently drawn attention to the fact that engineers working to design virtual and robotic agents do not always prioritize those aspects which will allow for
smooth, effortless human interaction, while psychologists studying interpersonal or joint-action do not always take into account technical realizability in describing what they see as the fundamental elements of successful multi-agent coordination.

One potential solution to this issue is to identify and model the behavioral dynamics (Warren, 2006) of natural human-human interaction using low-dimensional differential equations that can be easily implemented within interactive robotic or machine systems. Recent work by Dumas et al. (2014) and Zhai et al. (2014) has already provided support for the idea that relatively simple selfsustaining, nonlinear dynamical systems can be used to construct virtual interaction partners capable of successful, flexible coordination with human actors. Both groups of researchers used long-standing oscillator models of biological coordination to develop virtual agent systems capable of synchronizing with a selection of behaviors exhibited by a human actor. For instance, Dumas et al. (2014) have developed variations of their Human Dynamic Clamp (HDC) system that can coordinate with continuous and discrete finger movements of a human actor. Zhai et al. (2014) have designed a similarly adaptive virtual agent that is capable of coordinating with an individual during a continuous, one-dimensional movement-mirroring task.

The development of these dynamical, artificial agents has primarily focused on their ability to exhibit coordination with periodic behaviors, or synchronize with fluctuating movement speeds using a velocity estimation algorithm. However, one only has to consider a pedestrian navigating a busy city sidewalk to be reminded that people are often capable of prospectively coordinating their behavior with highly variable, seemingly unforeseeable events in an effortless manner. Recent research in human motor control and joint-action has demonstrated that small perceptualmotor feedback delays, such as those known to exist within the human nervous system, may actually facilitate the ability
to achieve anticipation of such continuous chaotic events (Stepp, 2009; Washburn et al., 2015). This phenomenon, referred to as strong anticipation or self-organized anticipatory synchronization, has been found to emerge when a unidirectional coupling exists between a "slave" system and a chaotically behaving "master" system (e.g., Masoller, 2001; Stepp \& Turvey, 2015; Voss, 2000). Surprisingly, as the slave system begins to synchronize with the chaotic behavior of the master system, the introduction of small temporal feedback delays results in the slave system anticipating the ongoing behavior exhibited by the chaotic master system.

Of particular significance here, is that the dynamics of chaotic anticipation during interpersonal coordination can be captured using a low-dimensional dynamical model and can be easily implemented in artificial agents. Such models of self-organized anticipatory synchronization could therefore provide an opportunity for significant advancement in HCI and HRI through the development of artificial systems capable of anticipating chaotic human behavior during real-time interaction. In the current study, two experiments were conducted to examine whether a virtual, artificial agent, whose arm movements were controlled by a time-delayed dynamical model, could not only coordinate with the chaotic movements of human actors in real time, but could do so in a self-organized anticipatory manner akin to human-human perceptual-motor coordination.

## Method

## Participants

Twelve students were recruited from the University of Cincinnati to take part in Experiment 1 along with four individuals from the greater Cincinnati area, for a total of 16 participants. Participants ranged in age from 19 to 31 years.

Seventeen University of Cincinnati undergraduate students participated in Experiment 2 (eight in the 1.5 coupling strength condition and nine in the 2.0 coupling strength condition). Participants ranged in age from 18 to 31 years.

## Procedure and Design

A virtual reality (VR) interface was employed in both experiments as it afforded the opportunity to examine the phenomenon of human-human and human-machine anticipatory synchronization within a realistic, yet highly controllable setting. A seated participant interacted with a simple virtual environment created using Unity 3D and viewed via a head-mounted Oculus Rift. Within the virtual environment participants saw a robot avatar sitting directly in front of them, and an additional avatar arm that moved along with their own right arm movements. The movements of this virtual participant arm were generated through the inverse kinematics function available within Unity 3D by coupling the pointer finger of the virtual arm to the real time
position of a wired motion sensor attached to the first two fingers of a participant's right hand. A Polhemus Liberty electro-magnetic motion capture system $(\sim 0.1 \mathrm{~mm}$ accuracy) (Polhemus Liberty, Polhemus Corporation, Colchester, VT) was used to record and track participants' movements at 120 Hz . The horizontal and vertical coordinates of participant movement were also recorded from the magnetic tracking system at a sampling rate of 75 Hz for later analysis. The receiver for this system was positioned approximately 10 cm in front of the fingers of a participant's right arm outstretched directly in front of their body.

Experiment 1: Human (slave) - Avatar (master) Experiment 1 was designed to establish the coordinative dynamics exhibited by human actors coordinating with an artificial agent via a novel VR setup. That is, we examined whether small perceptual-motor feedback delays could enhance a human actor's ability to anticipate the chaotic movements of the artificial agent system. Experiment 1 was also conducted to assess the degree to which bidirectional coupling (from master to slave) might influence the emergence of anticipatory synchronization. At the beginning of each experimental trial, the robot avatar began to move its left arm with the index finger pointed in a continuous trajectory. The participants' task was to synchronize their own arm movements with those of the moving stimulus (in this case the robot avatar's arm). The movements of the robot avatar (master system) were defined online by means of a chaotic spring system,

$$
\begin{aligned}
& \dot{x}_{1}=x_{2}+C\left(p_{1}-x_{1}\right) \\
& \dot{x}_{2}=-\left(\omega \pi\left(\frac{x_{3}}{\alpha}+\beta\right)\right)^{2} x_{1}+C\left(p_{2}-x_{2}\right) \\
& \dot{x}_{3}=-x_{4}-x_{5} \\
& \dot{x}_{4}=x_{3}+\alpha x_{4} \\
& \dot{x}_{5}=b+x_{5}\left(x_{3}-c\right)
\end{aligned}
$$

with the $x_{3}, x_{4}$ and $x_{5}$ dimensions defining a standard Rössler attractor (Stepp, 2009). This attractor generates the chaotic dynamics used to define position of the ' $x$ ' and ' $y$ ' dimensions for a simple harmonic oscillator specified in $x_{I}$ and $x_{2}$. The resulting system maintains an elliptical trajectory over time while exhibiting chaotic fluctuations in amplitude and frequency. Nine sets of system parameters $a$, $b, c, \alpha, \beta, \omega$ and initial conditions $x_{1}, x_{2}, x_{3}, x_{4}$, and $x_{5}$ were selected for use based on support of the evolution of bounded chaotic behavior.

Generating this behavior online allowed us to introduce a coupling term, $C$, between the virtual robot avatar and the behavior of the human participant. This system included an influence of the ' $x$ ' coordinates of a participant's arm
movements, $p_{1}$, on the ' $x$ ' coordinates of robot avatar arm movements, $x_{I}$, as well as a symmetrical influence of the ' $y$ ' coordinates of participant arm movements, $p_{2}$, on the ' $y$ ' coordinates of the robot avatar arm movements, $x_{2}$. The weight of avatar-participant coupling was manipulated to allow for more or less influence of the movement of the participant on that of the robot avatar, resulting in three total coupling strength conditions ( $0, .025$, and .05 ). Feedback delays of $26.67^{1}, 200$, and 400 ms were introduced between the participant's movements and the movement of their virtual arm. The average movement frequency exhibited by the robot avatar for a given trial in this study was between .23 and $.30 \mathrm{~Hz}^{2}$. Trials lasted 60 s . The first 10 s and last 5 s of each time series were discarded to remove transients.

Experiment 2: Human (master) - Avatar (slave) Experiment 2 examined whether an artificial agent, as a slave system, could anticipate the chaotic movements of a human master system. Participants were initially asked to complete two training trials in which they were to synchronize with robot avatar movement defined by fully chaotic, 2-D movement sequences generated ahead of time (i.e., there was no influence of participant movements on robot avatar master system behavior). The same two chaotic robot avatar movement sequences ${ }^{3}$ were provided to all individuals. During these trials participants saw their own virtual arm within the environment at the minimum delay possible (i.e., 26.67 ms ). Each sequence lasted 100 s . For the remainder of the experiment participants were asked to continue making the same kinds of movements they had been making during the training period: "generally circular and always in the same direction, but somewhat unpredictable in terms of the speed and size of movements". They were also informed that they would be switching roles with the robot avatar, so that they were now the leader and the avatar would be coordinating with their movements. For these test trials the system of equations specifying the baseline slave behavior of the robot avatar consisted of a harmonic spring oscillator ${ }^{4}$

[^238]\[

$$
\begin{aligned}
& \dot{x}_{1}=x_{3}+C\left(m_{1}-x_{1 d}\right) \\
& \dot{x}_{2}=x_{4}+C\left(m_{2}-x_{2 d}\right) \\
& \dot{x}_{3}=-\omega x_{1} \\
& \dot{x}_{4}=-\omega x_{2}
\end{aligned}
$$
\]

As in the harmonic spring system used in the previous experiment, this system includes a coupling term, $C$, here to modulate the strength of coupling between the robot avatar and the ' x ', $m_{1}$, and ' y ', $m_{2}$, dimensions of a 2-D master system (i.e., human participant) behavior. This method of delay-coupling results in a function that incorporates the ' x ' and ' y ' dimensions of its' past behavior, $x_{I d}$ and $x_{2 d}$, into the terms that reference the velocity of movement in each of the ' $x$ ' and ' $y$ ' dimensions, $x_{3}$ and $x_{4}$, effectively constituting a feedback delay within the system (see Stepp \& Turvey, 2015; Voss, 2000). Here the past behavior being referenced, $x_{d}$, is always that which occurred at a constant, set length of time, $\tau$, prior to the current time point, $t$,

$$
x_{d}=x(t-\tau)
$$

The remaining terms in the system of equations responsible for robot avatar movement include the variable specifying spring stiffness, $\omega$, through interaction with the ' $x$ ' and ' $y$ ' position variables, $x_{1}$ and $x_{2}$. Two different values for the slave-master coupling term, $C$, were introduced within this system (1.5 and 2), and were treated as a between subjects variable such that participants either interacted with the avatar system coupled to them with the lower or higher strength. Five different delay latencies were also introduced within the robot avatar system as $\tau(26.67,106.64,199.95$, 306.59 , and 399.90 ms ). These coupling strengths and delay latencies were chosen based on preliminary simulations using a chaotic spring master system and the current harmonic spring oscillator slave system. Each delay latency was instituted once per participant, with the order of presentation randomized over the five test trials experienced by each participant. Each trial lasted a total of 60 s . As in Experiment 1, the first 10 s and last 5 s of each time series were discarded for analysis.

## Data Analysis \& Results

## Largest Lyapunov Exponent

Calculation of the largest Lyapunov exponent (LLE) provided an initial measure of the chaotic dynamics within master system movement time series (see Washburn et al., 2015 for details). Average LLE values of robot avatar movement sequences from Experiment 1 were all positive ( $\mathrm{M}=0.024, \mathrm{SD}=0.008$ ), indicating that the robot avatar exhibited consistent chaotic movement dynamics even when it was coupled to the coordinating behavior of the human participant. LLE values associated with human participant
behavior in Experiment 2 were also positive for all combinations of feedback delay latency and slave-master coupling conditions except one (feedback delay: 26.67 ms , avatar-actor coupling: 2.0 ) (overall $\mathrm{M}=0.034, \mathrm{SD}=0.046$ ), indicating that the participants produced reasonably consistent chaotic movement dynamics when acting as the master system.

## Cross-Correlation and Phase Lead

To evaluate whether anticipatory synchronization occurred between the slave and master systems in Experiments 1 and 2, we first performed a cross-correlation analysis. This analysis indexes the degree of synchrony between two behavioral time series across a range of possible temporal relationships (Stepp, 2009). Of relevance for identifying anticipatory synchronization is the maximum degree of synchrony that occurred (indexed by the maximum observed cross-correlation coefficient) and the corresponding time lag (or lead) at which the synchrony occurred.


Figure 1: Average maximum cross-correlation (left) and temporal lead/lag (right) between artificial agent and human participant movements for Exp. 1 (top) and 2 (bottom). Line graphs in this figure are presented as means $\pm$ SEM. $* p<$ .05; two-way analysis of variance (ANOVA), using Bonferroni post hoc comparisons.

The results of this analysis for Experiment 1 were very similar to those found in previous studies of human anticipatory synchronization (Stepp, 2009; Washburn et al., 2015). Namely, that although overall coordination decreased slightly with increases in perceptual-motor feedback delay, anticipatory synchronization was observed for delays
between 200-400 ms (Fig. 1, top). Interestingly, no significant differences in anticipation were observed for the different coupling strengths employed. This is also consistent with existing studies in agent-environment and interpersonal human coordination, indicating that the VR paradigm employed here is suitable for the continued investigation of human anticipatory synchronization during uni-directional and bi-directional slave-to-master coupling situations.

In Exp. 2, maximum cross-correlation analysis also revealed a decrease in coordination with increases in timedelay, here implemented within the artificial agent slave system (Fig. 1, bottom left). More importantly, increases in time-delay were associated with a progressive decrease in lag latency between the artificial agent and human participant, with the artificial agent achieving temporal synchrony with the human participant for the 399.90 ms delay latency (Fig. 1, bottom right).

## Instantaneous Relative Phase

To gain further information about the anticipatory coordination that occurred between the human and artificial agent, an analysis of the relative phase between the movements of the slave and master systems in each experiment was conducted. Relative phase captures the spatial-temporal patterning of the coordination that occurs between two movement time-series. Of particular relevance for the current study was the distribution of relative phase angles that occurred for each feedback delay condition (i.e., how often a particular relative phase relationship was observed between the coordinator and producer over the course of a behavioral trial), with peaks in the distribution indicative of the stability of the coordination (higher peaks $=$ higher stability) and the degree to which the slave system led or lagged behind the movements of the master system (Schmidt \& O'Brien, 1997).

IRP distributions for participant with respect to avatar movements in Experiment 1 consistently indicated the occurrence of intermittent leading and lagging behavior, with more frequent leading than lagging in all combinations of coupling strength and feedback delay conditions (see Fig. 2). This kind of intermittent, or relative, coordination is consistent with the coordinative dynamics exhibited during interpersonal anticipatory synchronization (Washburn et al., 2015), and characterizes weakly coupled physical or biological limit-cycle oscillators (see Kelso \& Ding, 1993), including visually coupled rhythmic limb movements of coacting individuals (Schmidt \& O’Brien, 1997). These distributions look similar across conditions with some decreased stability apparent in the 400 ms delay condition, especially when there was no coupling from robot avatar to participant. There also seemed to be less relative difference in the frequency of leading to lagging in both of the bidirectional coupling conditions as compared to the no coupling condition at the 26.67 ms feedback delay. There
were very few differences in these distributions between the low and high coupling strengths conditions examined.


Figure 2: Distribution of average instantaneous relative phase (IRP) values between artificial agent and human actor as a function of the coupling strengths and delay conditions examined in Experiment 1.

Consistent with the maximum cross-correlation results above, when the artificial agent slave system was coupled to the live human actor master system in Experiment 2, most combinations of feedback delays and coupling strengths were associated with the artificial agent lagging behind the human actor (see Fig. 3). There was in fact relatively more anticipation than lagging at the longest feedback delay in Experiment 2 (i.e., 399.90 ms ), but the overall stability the phase relationships at this delay was reduced in comparison to the shorter delays. It is important to keep in mind that both the IRP frequency distributions and the maximum cross-correlation analysis represent average phase and temporal relationships between the artificial agent and the master system to which it is coupled. Furthermore, a participant-wise examination revealed that the artificial agent achieved anticipation for three of the eight participants in the 1.5 coupling strength condition, and five of the nine participants in the 2.0 coupling strength condition. This provides strong support for the idea that the kind of artificial agent developed and tested here can produce adaptive, prospectively coordinated behavior during ongoing, bi-directionally coupled interaction with a human actor.

## Discussion

The current project extends a rapidly emerging line of work investigating the process of coordination and self-organized
anticipatory synchronization during human-human and human-machine interaction. The findings of Experiment 1, demonstrated that anticipation similar to that observed during interpersonal interaction is also exhibited by human actors with respect to a chaotically behaving virtual coactor. Experiment 2 used the same novel VR paradigm to evaluate the anticipatory abilities of time-delayed artificial agent during interaction with a human co-actor. The movements of this artificial agent were defined by a low dimensional, harmonic oscillator system, coupled to the real-time behavior of the human co-actor. The results of this experiment revealed that the addition of feedback delays reduced the degree to which the avatar lagged behind the human actor.


Figure 3: Distribution of average instantaneous relative phase (IRP) values between human participant and robot avatar for coupling strengths of 1.5 (left) and 2.0 (right) and in each feedback delay condition examined in Experiment 2.

It is important to appreciate that while the addition of feedback delays in the artificial agent only, on average, reduced the lag between artificial agent and the human coactor, this should not be taken to indicate that the current agent is ill-suited to achieving self-organized anticipatory synchronization during human-machine interaction. The fact that human actors are intentional agents means they likely exhibited some adaptation to the artificial agent during interaction even though they were instructed to focus on producing their own movements and simply allow the avatar to follow them. This could account for the finding that the artificial agent only consistently achieved more anticipation than lagging of the human co-actor in the context of the longest time-delay. Furthermore, the patterns of intermittent anticipatory coordination observed in Experiment 2 were still quite similar to those seen in instances of interpersonal anticipatory synchronization, suggesting that small feedback delays in artificial agents induce a coordinative dynamic analogous to natural to human-human interaction.

Indeed, overall the current findings present a potentially transformative advance in the development of artificial agents and HRI. An agent defined by a low-dimensional dynamical model was able to display adaptive, anticipatory coordination during real time interaction with a human actor performing complex, seemingly unpredictable movements. The coordinative patterns exhibited by this agent were analogous to those observed during the occurrence of visual-motor agent-environment and interpersonal anticipatory synchronization in humans. This supports the idea that the dynamical models employed in the current research capture universal properties intrinsic to many physical systems, including complex biological behaviors like the human neural and movement processes that exhibit the kind of unpredictable determinism characteristic of chaos (e.g., Mitra et al., 1997). In displaying behavior that is qualitatively similar to human individuals the artificial agent developed here is likely capable of not only participating in the kind of interpersonal coordination known to support the successful completion of many everyday human tasks, but also engendering some of the associated increases interpersonal rapport and the facilitation of social awareness found following behavioral coordination between individuals (e.g., Miles et al., 2011). The current outcomes therefore suggest that engaging in coordinated interaction with such agents in the process of some higher order task goal will not only allow for more successful and efficient interactions during a wide variety of tasks, but may also result in the kinds of positive social outcomes associated with naturally occurring human interaction.

## Acknowledgments

This research was supported by the National Institutes of Health (R01GM105045).

## References

Dumas, G., de Guzman, G. C., Tognoli, E., \& Kelso, J. S. (2014). The human dynamic clamp as a paradigm for social interaction. Proceedings of the National Academy of Sciences, 111(35), E3726-E3734.
Kelso, J. A. S., \& Ding, M. (1993). Fluctuations, intermittency, and controllable chaos in biological coordination. Variability and motor control, 291-316.
Lopresti-Goodman, S. M., Richardson, M. J., Silva, P. L., \& Schmidt, R. C. (2008). Period basin of entrainment for unintentional visual coordination. Journal of Motor Behavior, 40(1), 3-10.
Lorenz, T., \& Hirche, S. (2014). Movement coordination in repetitive joint action tasks: Considerations on humanhuman and human-robot interaction. From 2014 IEEE International Conference.
Masoller, C. (2001). Anticipation in the synchronization of chaotic semiconductor lasers with optical feedback. Physical Review Letters, 86(13), 2782.
Miles, L. K., Lumsden, J., Richardson, M. J., \& Macrae, C.N. (2011). Do birds of a feather move together? Group membership and behavioral synchrony. Experimental brain research, 211(3), 495-503.
Mitra, S. S., Riley, M. A., \& Turvey, M. T. (1997). Chaos in human rhythmic movement. Journal Of Motor Behavior, 29(3), 195-198.
Palatinus, K. (2014). The Effects of Robot-Child Interactions on Interpersonal and Intrapersonal Coordination. Master's thesis, University of Connecticut, Storrs, Connecticut.
Schmidt, R. C., \& O'Brien, B. (1997). Evaluating the dynamics of unintended interpersonal coordination. Ecological Psychology, 9(3), 189-206.
Stepp, N. (2009). Anticipation in feedback-delayed manual tracking of a chaotic oscillator. Experimental brain research, 198(4), 521-525.
Stepp, N., \& Turvey, M. T. (2015). The Muddle of Anticipation. Ecological Psychology, 27(2), 103-126.
Voss, H. U. (2000). Anticipating chaotic synchronization. Physical review E, 61(5), 5115.
Warren, W. H. (2006). The dynamics of perception and action. Psychological review, 113(2), 358-389.
Washburn, A., Kallen, R. W., Coey, C. A., Shockley, K., \& Richardson, M. J. (2015). Harmony from Chaos? Perceptual-Motor Delays Enhance Social Anticipation. Journal of Experimental Psychology: Human Perception and Performance.
Yaghoubzadeh, R., Kramer, M., Pitsch, K., \& Kopp, S. (2013, January). Virtual agents as daily assistants for elderly or cognitively impaired people. In Intelligent Virtual Agents (pp. 79-91). Springer Berlin Heidelberg.
Zhai, C., Alderisio, F., Tsaneva-Atanasova, K., \& di Bernardo, M. (2014, December). Adaptive tracking control of a virtual player in the mirror game. In Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on (pp. 7005-7010). IEEE.

# Structure Learning in Motor Control: A Deep Reinforcement Learning Model 

Ari Weinstein (ariweinstein@google.com)<br>Deepmind, London, UK<br>Matthew M. Botvinick (botvinick@google.com)<br>Deepmind, London, UK<br>Gatsby Computational Neuroscience Unit, University College London


#### Abstract

Motor adaptation displays a structure-learning effect: adaptation to a new perturbation occurs more quickly when the subject has prior exposure to perturbations with related structure. Although this 'learning-to-learn' effect is well documented, its underlying computational mechanisms are poorly understood. We present a new model of motor structure learning, approaching it from the point of view of deep reinforcement learning. Previous work outside of motor control has shown how recurrent neural networks can account for learning-to-learn effects. We leverage this insight to address motor learning, by importing it into the setting of model-based reinforcement learning. We apply the resulting processing architecture to empirical findings from a landmark study of structure learning in targetdirected reaching (Braun et al., 2009), and discuss its implications for a wider range of learning-to-learn phenomena.


Keywords: motor adaptation; reinforcement learning; learning to learn; structure learning; system identification

## Introduction

Learning can be defined as a process that improves performance as exposure to a task increases. However, research on human and animal learning makes clear that this simple definition is not quite enough to explain the observed relationship between experience and performance. The full picture must also include 'learning-to-learn,' a process whereby growing experience causes learning itself to become more efficient (Harlow, 1949). More specifically, learning-to-learn (also referred to as meta-learning and structure learning) occurs in settings where the learner encounters a series of tasks that share some underlying structure, and gains from these an ability to quickly adapt to a new task that displays the same general form (Thrun \& Pratt, 1998).

A vivid example of learning-to-learn, which provides a concrete focus for the present research, comes from research on motor adaptation. Many studies have documented the ability of human subjects to adapt to perturbations of motor dynamics or kinematics, as for example in prism adaptation (Harris, 1963). However, a series of studies by Braun and colleagues (Braun et al., 2009, 2010; Braun \& Wolpert, 2012) went beyond this to show that adaptation can occur faster when the subject has prior exposure to perturbations that share structure with the final test conditions. In one specific experiment, upon which we will continue to focus, Braun and colleagues (2009) studied reaching under visuomotor rotation. They examined the speed with which targetdirected reaching adapted to a 60-degree rotation, manipulating between subjects the content of a preceding set of training
trials. In one condition, which we will refer to as Rot, subjects dealt with a series of rotations (though never the one presented at test). In a comparison condition Rot+, subjects dealt with a more diverse set of transformations, each made up of a rotation along with shear and scale components. Results showed that subjects in the Rot group adapted faster to the probe rotation problem (Figure 1). Braun et al. (2009) interpreted this as learning-to-learn effect, which they referred to as "motor structure learning": Subjects in the Rot group evidently learned that the transformations being presented were restricted to a particular structurally coherent set (rotations), and this allowed them to infer and adapt rapidly to the probe transformation. This structure learning was less feasible in the Rot+ condition because the structure underlying the training set was more complex, thus offering weaker constraint on inference when facing a new transformation.

In the present study, we consider the computational mechanisms underlying motor structure learning, treating it as a case study in learning-to-learn. Despite widespread agreement that learning-to-learn effects are both real and important, the precise computational processes underlying such effects are poorly understood. The most widely proposed idea comes from a Bayesian perspective, and proposes that learning-to-learn involves refining the structure and hyperparameters of a generative model of the relevant task domain (Lake et al., 2015). Braun and colleagues initially proposed, and later investigated (Genewein et al., 2015) a model of this sort to account for their structure learning results.

A different computational proposal, which has been less widely considered in cognitive science, comes from neural network or deep learning research. In classic work, Hochreiter and colleagues (2001) showed how a recurrent neural network (RNN) can learn to learn, by integrating information about past outcomes into predictions concerning new observations. Recent applications of this idea (Wang et al., 2016; Duan et al., 2016) have treated the RNN in Hochreiter's scheme as a mechanism for directly selecting actions. In the present work, we leverage Hochreiter's (2001) insight in a different way, using an RNN as an adaptive model of the task domain, which is leveraged by a separate actionselection mechanism. In this sense, the aim of our work is to bridge Bayesian and deep learning perspectives on learning-to-learn. On a more immediate level, we show how the resulting approach can be used to account for the findings of Braun


Figure 1: Results redrawn from Braun et al.(2009), showing mean cumulative error on a series of five reaches under a 60degree visuomotor rotation.
and colleagues (2009) in motor adaptation.
The motor control literature suggests that actions such as reaching are based, at least in part, on an internal predictive or forward model of reaching dynamics (Wolpert et al., 1995; Miall \& Wolpert, 1996), and analyses of motor adaptation have portrayed adaptation as reflecting a progressive adjustment of this internal model to fit with current observations (Berniker \& Kording, 2008; Haith \& Krakauer, 2013). Following this idea, and a great deal of previous work in computational motor control, we construe action selection as a form of model-based reinforcement learning (RL).

Formulating the problem in this way begins by casting it as a finite-time Markov decision process (MDP) $M$, which is made of a set of states $S$, a set of possible actions $A$, a transition function $\mathcal{T}$, and reward function $R$ (in many settings a discount factor $\gamma$ is included, but since we formulate the task as a finite-time problem this is unnecessary). The goal is to select actions that maximize the cumulative reward up to some time $T$ : $\sum_{t=0}^{T} r_{t+1}$, where $t$ indexes discrete time steps up to some maximum $T$, and $r_{t}$ is the reward received on each step. Focusing on target-directed reaching, the task studied by Braun and colleagues (2009), the problem is defined as follows: $M$ is the entire reaching task; $S$ are possible arm configurations; $A$ are possible motor inputs; $\mathcal{T}$ defines the dynamics of the arm based on motor inputs; $R$ is the negative distance from the cursor to the target. In order to be consistent with the literature on structure learning in motor control, we will use the terms reward maximization and penalty (or error) minimization interchangeably.

In model-based RL, a model $\hat{M}$ of the environment $M$ is built, and then used by a planner $P$ in order to construct an action-selection policy. The general form of a model-based learning architecture is diagrammed in Figure 2, left. Here a planner $P$ is informed of a current state $s$ by the true MDP $M$. Based on the particular policy of $P$, the planner queries the model $\hat{M}$ with a series of state-action pairs $\left(s_{t}, a_{t}\right)$, and in turn receives an estimated next state $s_{t+1}$ and reward $r_{t+1}$. After the planner completes querying $\hat{M}$, either because it has


Figure 2: Model-based reinforcement learning with a fixed model (left) and an adaptive model (right).
taken as much data as it needs or due to some outside pressure such as a time limit, it returns an action $a$ which is executed in $M$, which results in a new state and reward and the process repeats.

In the architecture shown in Figure 2 (left), one way of implementing the forward model $M$ is as a feed-forward neural network. This approach has been explored in a number of previous studies (Jordan \& Rumelhart, 1992; Hamrick et al., 2016). However, a feed-forward neural network will not suffice to address the learning-to-learn phenomena we are concerned with here. Indeed, the overall architecture must be fundamentally changed in order to address the learning-tolearn problem.

As introduced earlier, learning-to-learn arises in a setting where the learner encounters a series of interrelated problems or tasks, and must adapt to each one in turn. Using our terminology, each task $M_{n}, n=1 \ldots N$ becomes a sample from a task distribution $\mathbb{M}$. As such, the properties of each $M_{n}$ must be inferred based on observed action-outcome pairs (a process referred to in the engineering literature as system identification). On a formal level, this demand changes the MDP we have been considering into a partially observable Markov decision process (POMDP). By definition, a POMDP is an MDP which additionally has an observation space $O$ and observation function $\Omega$ which takes its internal state and outputs an observation $o$ to the agent. Instead of the true state, an agent only has access to observations, which unlike state, is generally insufficient to act optimally when considered in isolation. In order for $\hat{M}$ to adjust to each $M_{n}$, it must have some form of memory $\alpha$ to keep a relevant summary of interactions with the environment, allowing for integration over previous timesteps in order to accurately estimate problem dynamics.

These requirements yield the interaction and planning structure diagrammed in Figure 2 (right). Instead of states $s$, information presented is in terms of observations $o$. The model $\hat{M}$ must now directly consume $o$ and $r$ from $M_{n}$ at every time step, which causes it to update its memory $\alpha . P$ now only passes a sequence of actions along trajectories to the model. This is because it does not have access to the
true state, and because observations alone are not sufficient for planning or modelling ${ }^{1}$. Additionally, during the planning trajectories, the $P$ must signal to $\hat{M}$ when its evaluation of a simulated trajectory is complete so that $\alpha$ can be reset (omitted from figure for clarity).

Note that, unlike in the simpler MDP, in the POMDP setting the internal model $\hat{M}$ cannot be accurately implemented as a feed-forward neural network, because such networks do not have memory or persistent internal state. The key move in the present work is to substitute for the feed-forward network a RNN, whose recurrent connectivity endows it with the memory needed to support system identification and, as we will show, learning-to-learn.

## Simulation Study

We predicted that the proposed architecture, if trained in an appropriate multi-task setting, would display learning-tolearn, leveraging experience with past tasks to adapt rapidly to a new task sharing in the same structure. In order to test this idea, we applied the architecture to the task paradigm employed in they study of motor structure learning by Braun and colleagues (2009).

## Model implementation and task design

We implement the architecture shown in Figure 2 (right), instantiating the forward model in the form of a recurrent neural network (which is naturally deep as it is unrolled over time). More specifically, this involves one LSTM layer (Hochreiter \& Schmidhuber, 1997) followed by two more fully connected layers containing rectified linear units (Nair \& Hinton, 2010), where each layer contains 100 units. The planner is an open-loop planner based on cross-entropy optimization, as described in Weinstein \& Littman (2013), with the addition of "warm starting." In warm starting, planning is done from scratch on the first step of a trajectory, but all subsequent steps in the actual domain initiate planning with the result from the previous step. At each time step only the first action in the current plan is executed in the true domain before partial replanning in this manner occurs. For simplicity, we assume (without loss of generality) that $\hat{M}$ has access to the reachtarget coordinates and can compute the reward function.

In order to model target-directed reaching, we implemented a simple arm model. While not intended to offer a detailed model of biomechanics, this was intended to capture the most important aspects in terms of possible arm geometry, velocity, and acceleration (Nagasaki, 1989). As simulated, the underlying state space of the problem has four dimensions: horizontal shoulder angle, elbow angle, and corresponding angular velocities. Observations emitted are the Euclidian position of the cursor controlled by the arm's tip as seen in the experiment (meaning that $\hat{M}$ must also learn to

[^239]estimate velocities), and the goal. The two dimensional action space sets the angular accelerations of the joints, and the reward is the negative Euclidean distance of the cursor from the center of the goal region.

In the reaching task, the cursor is always initialized at the origin and is controlled by the transformation of the underlying position of the simulated hand. Before each trial a goal location is selected which is set to be 8 cm from the origin at a uniformly distributed angle.

## Training and testing procedure

The simulation study, like the experiment by Braun et al. (2009) was divided into training and testing phases. During training, the RNN model was trained to predict each sequential outcome observation exactly along a trajectory consisting of observations and randomly selected actions, that is, following a random walk. Again as in the empirical study, two versions of the model were trained in different environments. One model, which we label (in a minor abuse of terminology) Rot, was trained on a series of visuomotor rotations, simulated by appropriately transforming the observed cursor coordinates. The second model instance, Rot + , was trained on a combination of rotations, shears, and scales (following the design described in Braun et al., 2009, Supplemental Data). Following the design imposed by Braun and colleagues, when the rotation to be presented to Rot + fell between $\pm 50^{\circ}$ and $70^{\circ}$, a rotation of $\pm 60^{\circ}$ was substituted and no linear transform was applied. As a result, both Rot and Rot+ had roughly equal exposure to the transformation used during test trials. In both conditions, the model was trained by backpropagation through time on 2,000 trajectories of random-walk data, with each trajectory containing three seconds of simulation time, and training starting from the initial observation of each trajectory.

In the testing phase the RNN weight parameters were frozen and reaches were elicited only under only pure rotations of $\pm 60^{\circ}$, as in the testing phase of the experiment by Braun and colleagues. Goal locations were placed at a randomly selected angle 8 cm from the start location of the cursor. The radius of the goal region is 1.6 cm . In order to simulate a series of reaches, the angle of the imposed visuomotor rotation was held constant while the position of the goal varied between reaches. Test reach trajectories ran for a maximum of two seconds, terminating early if the cursor was brought within the goal region for 500 ms .

## Results

Training of models for both the Rot and Rot + conditions were successful, but the model trained on Rot was able to achieve an average error of about 0.002 cm per time step for trajectories in the training set, while Rot + an error of about 0.03 cm by the same metric. In the Rot condition, the RNN model learned to act as an adaptive forward model, adjusting its predictions to fit with accumulating action-outcome observations. Figure 3 shows the average observation-prediction errors of both Rot and Rot+ models during an initial random-


Figure 3: Model errors by step in initial trial.
walk trajectory which was not part of the training set. The initial data-point is the error of the model prior to any experience in the test MDP. In interpreting the values on the $y$-axis of the plot, it should be taken into account that in our simulation two seconds of time takes 28 discrete time steps, and error compounds over these steps. In contrast to the Rot model, the Rot+ model adapts much less successfully, despite having been trained on an identical amount of data.

Figure 4 shows the mean cumulative penalty when the model is coupled with a planner, for each reach at test for both models. This is intended for comparison with the empirical data from humans in preceeding work shown in Figure 1. As predicted, Rot is better able to conduct structure learning, by adapting more rapidly and completely to the test rotation (the manipulation both models were exposed to during training) than Rot+. This qualitatively replicates the experimental findings from Braun et al. (2009).

Figure 5a shows average trajectories for five successive reaches (normalized by rotation and goal angles), for both Rot and Rot+ models. Both models adapted across reaches (starting with smaller initial angular errors after the initial reach), but the effects were stronger in the Rot model. Quite striking is the standard deviation of the final position of the first trial of Rot, and Rot+ in cyan and magenta, respectively. Although on average Rot+ tracked toward target, there is a tremendous amount of variability in its trajectories, and was not able to consistently reach the goal region, whereas Rot usually terminated within the target.

We also consider other indirect metrics of performance which are presented in the human studies such as initial angular error, velocity, and minimum distance to goal region, which are presented in Figures 5b through 5d, respectively. In general the results with these metrics are similar to the previous plots, with Rot improving quickly and performing better than Rot+. We also note the higher variance of Rot+, which manifests itself in wider confidence intervals across all Figures, especially Figure 5d. These results are qualitatively aligned with those reported in the experimental study.

In fact, of these metrics, the only one which shows improvement by Rot+ is the initial angular error. Even with


Figure 4: Average cumulative penalty by trial.
this improvement, the agent frequently falls short of reaching goal region (which would allow for an early termination of distance penalties). This is most likely due to the fact that on average testing data in Rot + has a scaling amount of roughly 1.3 (this design is part of the original human study), and indeed Rot+ almost uniformly tells the planner that actions will result in greater changes in location than actually occur.

Although Rot+ was less effective at structure learning than Rot, it is not the case that it failed entirely. The average penalty of a trajectory for agent using a uniform random policy is approximately 220 units which is significantly poorer than what Rot+ was able to achieve.

We note that our goal was not to fit the results from the human data quantitatively, but rather to demonstrate the same phenomenon which is that structure learning becomes more difficult as the the amount of variability in the problem increases. And although Rot + was not able to perform well, the overall architecture does have the capacity to do effective structure learning; expanding the data corpus size by a factor of five produces models that have statistically equal, high quality performance on test tasks for both Rot and Rot+.

## Discussion

Learning-to-learn is a fundamental aspect of human behavior, but its computational basis is not yet well understood. We have presented a new model of learning-to-learn in the setting of motor adaptation. This task, defined by Braun and colleagues (2009) involves learning to learn in the sense that the subject must gather data on a current situation in order to infer the hidden parameters of the dynamics, and indeed Braun and colleagues state that learning to learn can be recast as structure learning. On the other hand, a stronger definition of learning to learn could require learning to adapt to a situation it has not experienced in the past, perhaps in terms of new objects to interact with that follow some prelearned rules (Harlow, 1949). This has been considered in a different simulated setting in RL where the agent learns policies (Wang et al., 2016), as opposed to models of the environment as is done here.

Adopting the standard approach, we assume that motor


Figure 5a: Average trajectories by trial. Standard deviation of final position of first trial in shaded region. Goal region in black.


Figure 5c: Average velocities by trial
adaptation involves updating an internal forward model of reaching dynamics. Our novel contribution is to instantiate this internal model as a recurrent neural network. Through simulations of a key experimental study, we have shown that the resulting system not only learns to adapt to changing perturbations, but also that its adaptation becomes more effective when there is prior exposure to structurally related conditions, as seen empirically in motor structure learning. Importantly, no special measures were required in order to secure this learning-to-learn effect. Through error-correcting learning, the parameters of the RNN are, perforce, fit to the structure of the pre-training data. That same structure is thus naturally - indeed inevitably - expressed in its later inferences at test.

We consider learning to learn as refining a (potentially implicit) hypothesis set based on experience. If the problem has a large underlying dimension, then the hypothesis set learned by the model must be of corresponding size. This is in turn fundamentally linked to the amount of data required to both train the model, as well as do inference, accurately. For these reasons, it is to be expected that when comparing the data requirements of doing both in Rot versus Rot + , Rot leads to lower data requirements. Just as is the case with Braun and colleagues (2009), we do not attempt to disentangle these issues, although a more detailed investigation warrants future


Figure 5b: Average angle error from goal after 200 ms of simulated time.


Figure 5d: Average minimum distance to goal by trial
attention.
As noted earlier, our use of RNN dynamics to capture learning-to-learn effects builds directly on pioneering work by Hochreiter and colleagues (2001), in which an RNN model was applied to the problem of function induction (see also Wang et al., 2016; Santoro et al., 2016). In contrast to that work, we deployed our RNN as a forward model situated within a larger model-based RL system. In this sense, our implementation bridges between Hochreiter's original proposal and models of motor adaptation that have embedded an adaptive Bayesian model of limb dynamics (e.g. Berniker \& Kording, 2008; Genewein et al., 2015). The approach we have introduced also relates to other work in which RNNs have been used as forward models in support of motor adaptation, but where multiple fixed models are assumed (Haruno et al., 2001; Pitti et al., 2013), rather than a single adaptive model used here. These fixed models lack memory, meaning that reweighing fixed models aside, adaptation is only possible by retraining the system. Implicitly, our work implements a sort of Kalman filter which has also been considered previously in recurrent networks (Wolpert et al., 1995). Undertaking a careful comparison between these related approaches and the one we have introduced here offers an important objective for next-step research.

Our implementation of the reaching task was deliberately minimal, simplifying both the underlying biomechanics and the motor planning process, in order to foreground our central computational proposal. Naturally, a more detailed evaluation of the approach, incorporating a higher degree of empirical constraint, will be desirable in further evaluating the viability of our approach as a theory of motor adaptation. A related opportunity is to consider the potential parallel between the recurrent connectivity underlying the function of our adaptive model and the recurrent connectivity inherent in biological neural circuits underlying motor control and adaptation, including circuits running through the basal ganglia and cerebellum.

At the same time, however, we feel it may also be fruitful to apply the model-based framework we have introduced here in domains beyond motor control, in particular other domains that display the characteristics of a POMDP and where learning-to-learn effects have been observed. Such tasks are indeed ubiquitous, ranging from structured bandit tasks to video-game play (Wang et al., 2016; Lake et al., 2015). To the extent that the framework we have presented here can be adapted and (more challenging) effectively scaled to these other settings, it offers to provide a more general new perspective on the problem of learning-to-learn.

## Acknowledgements

We would like to thank Daniel Braun and Konrad Kording for valuable feedback.

## References

Berniker, M., \& Kording, K. (2008). Estimating the sources of motor errors for adaptation and generalization. Nature Neuroscience, 11, 1454-1461.
Braun, D. A., Aersten, A., \& Wolpert, D. M. (2009). Motor task variation induces structural learning. Current Biology, 19, 352-357.
Braun, D. A., Mehring, C., \& Wolpert, D. M. (2010). Structure learning in action. Behavioural Brain Research, 206, 157-165.
Braun, D. A., \& Wolpert, D. M. (2012). Structural learning in sensorimotor control. In Encyclopedia of the sciences of learning (pp. 3208-3211).
Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., \& Abbeel, P. (2016). RL ${ }^{2}$ : Fast reinforcement learning via slow reinforcement learning. CoRR, abs/1611.02779.
Genewein, T., Hez, E., Razzaghpanah, Z., \& Braun, D. A. (2015). Structure learning in bayesian sensorimotor integration. PLoS Computational Biology, 11(8).
Haith, A. M., \& Krakauer, J. W. (2013). Model-based and model-free mechanisms of human motor learning. In Progress in motor control: Neural, computational and dynamic approaches (pp. 1-21).
Hamrick, J. B., Pascanu, R., Vinyals, O., Ballard, A., Heess, N., \& Battaglia, P. W. (2016). Imagination-based decision
making with physical models in deep neural networks. In NIPS 2016 workshop on intuitive physics.
Harlow, H. F. (1949). The formation of learning sets. Psychological Review, 56(1), 51-65.
Harris, C. S. (1963). Adaptation to displaced vision: Visual, motor, or proprioceptive change? Science, 140(3568), 812-813.
Haruno, M., Wolpert, D. M., \& Kawato, M. (2001). Mosaic model for sensorimotor learning and control. Neural Computation, 13, 2201-2220.
Hochreiter, S., \& Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8).
Hochreiter, S., Younger, A. S., \& Conwell, P. R. (2001). Learning to learn using gradient descent. In Artificial neural networks ICANN (Vol. 2130, pp. 87-94).
Jordan, M. I., \& Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. Cognitive Science, 3, 307-354.
Kaelbling, L. P., Littman, M. L., \& Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. Aritficial Intelligence, 1-2.
Lake, B. M., Salakhutdinov, R., \& Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science, 350(6266), 1332-1338.
Littman, M. L., Sutton, R. S., \& Singh, S. (2001). Predictive representations of state. In Neural information processing systems (Vol. 14).
Miall, R., \& Wolpert, D. (1996). Forward models for physiological motor control. Neural Networks, 9(8), 1265-1279.
Nagasaki, H. (1989). Asymmetric velocity and acceleration profiles of human arm movements. Experimental Brain Research, 74(2), 319-326.
Nair, V., \& Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In International conference on machine learning (pp. 807-814).
Pitti, A., Braud, R., Mah, S., Quoy, M., \& Gaussier, P. (2013). Neural model for learning-to-learn of novel task sets in the motor domain. Frontiers in Psychology, 22.
Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., \& Lillicrap, T. (2016). One-shot learning with memory-augmented neural networks. arXiv preprint arXiv:1605.06065.
Thrun, S., \& Pratt, L. (Eds.). (1998). Learning to learn. Kluwer Academic Publishers.
Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2016). Learning to reinforcement learn. arXiv preprint arXiv:1611.05763v2.
Weinstein, A., \& Littman, M. L. (2013). Open-loop planning in large-scale stochastic domains. In AAAI conference on artificial intelligence (Vol. 27, pp. 1436-1442).
Wolpert, D., Ghahramani, Z., \& Jordan, M. (1995). An internal model for sensorimotor integration. Science, 269(5232), 1880-1882.

# Children's intuitions about the structure of mental life 

Kara Weisman<br>(kweisman@stanford.edu)<br>Department of Psychology<br>Stanford University

Carol S. Dweck<br>(dweck@stanford.edu)<br>Department of Psychology<br>Stanford University

Ellen M. Markman<br>(markman@stanford.edu)<br>Department of Psychology<br>Stanford University


#### Abstract

We investigated children's understanding of mental life by analyzing attributions of perceptual, cognitive, affective, and other capacities. 200 children ( $7-9 \mathrm{y}$ ) and 200 adults evaluated the mental capacities of beetles or robots. By assessing which capacities traveled together when participants disagreed about these controversial "edge cases," we reconstructed the latent structure underlying mental capacity judgments from the bottom up-a novel approach to elucidating conceptual structure among children. For both children and adults, factor analyses revealed a distinction between social-emotional, physiological, and perceptual-cognitive capacities, hinting at three fundamental ways of explaining and predicting others' actions: as social partners, biological creatures, and goaldirected agents (each involving related forms of both "experience" and "agency"; Gray et al., 2007). Relative to adults, children attributed greater social-emotional capacities to beetles and robots, suggesting that intuitive ontologies of mental life could be critical for making sense of children's developing understanding of the social world.


Keywords: mind perception; sentience; animate-inanimate distinction; cognitive development.

## Introduction

Questions about the nature of mental life extend back to antiquity, but it is only recently that cognitive scientists have begun to explore lay people's conceptions of the mind.

One particularly exciting approach was pioneered by Gray, Gray, and Wegner (2007) in their work on mind perception. From participants' responses to simple questions about the mental capacities of various characters (e.g., "Which is more capable of experiencing joy: a frog or an infant?"), Gray et al. extracted a conceptual space characterized by two dimensions: "experience," the extent to which a character is capable of hunger, fear, pride, and other inner experiences; and "agency," the extent to which a character is capable of self-control, morality, memory, and other capacities central to acting in the world.

This bottom-up approach has tremendous potential in elucidating the kinds of deep conceptual structures that are difficult for participants to report on directly (and for experimenters to anticipate a priori). Rather than imposing theory-driven categories onto participants' responses, Gray et al. (2007) let the data speak for themselves.

However, Gray et al.'s (2007) study focused participants' attention on the similarities and differences between characters, thus illuminating the dimensions along which social beings are thought to differ from each other-an important part of social reasoning, but not equivalent to intuitions about the structure of mental life itself.

Inspired by their approach, we recently conducted a series of studies designed to assess intuitive ontologies of mental life directly (Weisman, Dweck, \& Markman, 2016). We focused participants' attention on the connections and divisions between different aspects of mental life by asking them to evaluate a wide variety of mental capacities for a single character (e.g., a robot or a beetle). By analyzing patterns of attributions across participants, we uncovered a 3-part conceptual structure that emerged reliably across several studies: Physiological sensations and self-initiated behaviors hung together to form a suite of capacities related to the body; social-emotional experiences and moral agency formed a suite of capacities related to the "soul"; and perceptual-cognitive abilities and goal pursuit formed a suite of capacities related to the mind. Interestingly, each of these three factors encompassed aspects of both "experience" and "agency." Instead of the broad distinction that seems to characterize adults' understanding of social beings (Gray et al., 2007), adults' understanding of the structure of mental life itself seems to hinge on distinctions among varieties of experience and agency, and connections among related kinds of experience and agency.

Intuitions about mental life are at the core of many of the oldest and richest lines of research in developmental psychology, including animism (Piaget, 1929), lay biology and psychology (Carey, 1985), and theory of mind (Wellman \& Woolley, 1990). But most of this work has relied on a priori distinctions between perception, desires, emotions, intentions, beliefs, knowledge, etc. (Flavell, 1999), leaving the actual conceptual structure underlying children's reasoning and behavior unknown; to our knowledge, there have been no attempts to map out the ontology of mental life from the ground up with children. This may be due in part to the challenges of implementing bottom-up approaches, which generally require hundreds of participants to answer dozens of questions-not the typical design for studies with young children. On the other hand, studies like Gray et al. (2007) and Weisman et al. (2016) are built on the premise that these complex conceptual structures can be uncovered from participants' answers to relatively simple questions, suggesting that this approach might lend itself to adaptation for younger participants.

Thus, in the current study we developed a bottom-up approach for uncovering children's intuitions about the structure of mental life. We believe these intuitions are critical for making sense of children's social and moral reasoning about the people, animals, and other social partners in their lives.

## Study

We based our experimental paradigm on our previous work with adults, in which participants evaluated a target character on 40 mental capacities using a 7 -point scale from not at all capable to highly capable. Pilot testing suggested two necessary modifications for children: rewording some of the mental capacity items, and using a 3-point response scale (no, kinda, yes). Although a 3-point scale is not optimal for factor analyses, it allowed children to move fast enough through the study to answer all 40 questions, and maintaining this within-subjects design was our top priority for the planned factor analysis.

As in our previous work, we focused on judgments of the mental capacities of two "edge cases" in social reasoning: a beetle and a robot. Because beetles are animals and robots are artifacts, this pair provides insight into the role of biological life in attributions of mental life-an issue of particular interest from a developmental perspective, given the long history of work on the development of the animateinanimate distinction and its relation to folk psychology. Most critically for our bottom-up approach to uncovering intuitive ontological structures, the "mental lives" of these entities are controversial: People differ in their assessments of the mental capacities of beetles and robots (Weisman et al., 2016). This allowed us to address the following question: When children disagree about the mental capacities of some entity, which capacities "go together"?

Pilot testing suggested that children as young as 7 y found the paradigm easy and enjoyable, and work on the development of lay biology and psychology has suggested that these concepts may continue to develop well into middle childhood (e.g., Carey, 1985; Hatano \& Inagaki, 1997; Piaget, 1929; cf. S. Gelman \& Opfer, 2002). Thus, we targeted 7- to $9-y$-old children for our child sample.

We also recruited a group of adults to validate our childfriendly paradigm, i.e., to evaluate whether it replicated our earlier work with adults (Weisman et al., 2016).

## Methods

Participants. 400 people participated in this study.
Children ( $n=200$ ) participated at one of several Bay Area museums or at their younger sibling's preschool (median study duration: 5.18 min ). Children ranged in age from $7.0-$ 10.0 y (median: 8.3 y ). An additional 12 children participated but were excluded for being outside the target age range ( $n=7$ ), being of unknown age ( $n=4$ ), or being shown a target character other than a beetle or a robot $(n=1)$.

Adults ( $n=200$ ) participated via MTurk. Adult participants had gained approval for $\geq 95 \%$ of previous work on MTurk; had verified accounts based in the US; and indicated that they were $\geq 18 y$ old. Adults were paid $\$ 0.30$ (median duration: 2.48 min ). Repeat participation was prevented.

Materials and procedure. Participants were randomly assigned to evaluate one of two target characters: a beetle, accompanied by a photograph of a black beetle on a leaf
( $n=98$ adults, 104 children), or a robot, accompanied by a photograph of a humanoid robot (Sony Qrio; $n=102$ adults, 96 children). The picture and label (a beetle or a robot) were present throughout the survey.

Instructions focused on the idea that we wanted to know what participants thought "[beetles/robots] can do and can not do." Participants rated the target character on 40 mental capacities, presented in a random order for each participant. On each trial, participants responded no, kinda, or yes to the question "Do you think a [beetle/robot] can...?"

The 40 mental capacities were designed to be as close as possible to those in our previous studies (Weisman et al., 2016) while being comprehensible to children in early elementary school. This set of items included physiological sensations related to biological needs (e.g., get hungry); emotional experiences (e.g., feel happy); perceptual abilities (e.g., hear sounds); cognitive abilities (e.g., remember things); capacities related to autonomy or agency (e.g., decide what to do); social abilities (e.g., feel guilty); and several additional items (e.g., be aware of itself). Each of these a priori categories included at least five items of varying valence, complexity, and phrasing (see Table 1).

We also prepared a short definition for each item, so as to be consistent in our responses to participants if they asked for clarification. Children were encouraged at the beginning of the study to ask questions if they did not know what a word meant, in which case they given these definitions; adults were told that they could access these definitions by hovering over the text. Pilot testing suggested that 7 items required clarification for most children, so these items were always accompanied by their definitions from the beginning of the trial (for both children and adults), as follows: have a personality, like when someone is shy and somebody else is silly; have beliefs, like when you think something is true; feel pleasure, like when something feels really good; have desires, like when you really want something; have selfcontrol, like when you stop yourself from doing something you shouldn't do; have goals, like when you're trying hard to do something or make something happen; and feel sick, like when you feel like you might throw up.

Data preparation. We scored responses of no as 0 , kinda as 0.5 , and yes as 1 . We dropped trials with response times that were faster than a preset criterion of 250 ms ( $n=3$ child trials, 97 adult trials) and retained participants regardless of skipped trials ( $n=55$ child trials, 1 adult trial). Overall, only $1 \%$ of adult trials and $1 \%$ of child trials were missing data.

Analysis plan. Our primary goal was to determine which mental capacities go together: e.g., if a participant indicated that a character was capable of hunger, what other capacities did she endorse? To do this, we used exploratory factor analyses (EFA) to reveal the covariance structure underlying participants' responses, collapsing across characters and using Pearson correlations to find minimum residual solutions. We first examined maximal (13-factor) unrotated solutions to determine how many factors to extract, using
the following preset retention criteria: Each factor must have an eigenvalue $>1.0$; individually account for $>5 \%$ of the total variance; and be the "dominant" factor (the factor with the highest factor loading) for $\geq 1$ mental capacity item. We focus our interpretation on varimax-rotated solutions, extracting the number of factors that met these criteria. (Using polychoric correlations and/or oblimin rotation yielded similar latent structures.)

## Results and Discussion

We first assess the validity of our child-friendly paradigm relative to our previous work with adults by examining an EFA of adults' responses. We then address our primary question-children's intuitions about the structure of mental life-via EFA of children's responses. Finally, we analyze differences in factor scores between children and adults.

EFA: Adults. EFA revealed 3 factors that met our criteria.
After rotation, the first factor corresponded primarily to physiological sensations related to biological needs. It was the dominant factor for such items as get hungry, do math (negative loading), feel pain, feel scared, and feel tired. Factor 1 accounted for $25 \%$ of the variance in the rotated maximal solution.

The second factor corresponded primarily to capacities for self- and other-relevant emotions. It was the dominant factor for such items as feel joy, feel proud, feel sad, feel happy, and feel love. Factor 2 accounted for $21 \%$ of the variance in the rotated maximal solution.

Finally, the third factor corresponded primarily to perceptual-cognitive abilities to detect and use information about the environment. It was the dominant factor for such items as recognize somebody else, figure out how to do things, remember things, sense whether something is close by or far away, and communicate with somebody else. Factor 3 accounted for $10 \%$ of the variance in the rotated maximal solution. (See Table 1 for all factor loadings.)

In sum, as in our original studies (Weisman et al., 2016), a three-factor structure emerged from adults' mental capacity attributions, characterized by a distinction between physiological, social-emotional, and perceptual-cognitive abilities. This suggests that our child-friendly paradigm was valid: Using reworded items and a 3 -point response scale elicited the same intuitive ontology of mental life, among adults, as revealed by our "adult-friendly" paradigm.

EFA: Children. Again, 3 factors met our retention criteria.
After rotation, the first factor corresponded primarily to social-emotional abilities. It was the dominant factor for such items as feel proud, feel happy, feel joy, get hurt feelings, and feel sad. Factor 1 accounted for $25 \%$ of the variance in the rotated maximal solution.

The second factor corresponded primarily to physiological sensations. It was the dominant factor for such items as get hungry, feel pain, do math (negative loading), smell things, and feel scared. Factor 2 accounted for $18 \%$ of the variance in the rotated maximal solution.

The third factor corresponded primarily to perceptualcognitive abilities. It was the dominant factor for such items as be aware of itself, figure out how to do things, be aware of things, sense whether something is close by or far away, and sense temperatures. Factor 3 accounted for $7 \%$ of the variance in the rotated maximal solution.

In sum, like adults, children's mental capacity attributions were dominated by a 3-way distinction between socialemotional, physiological, and perceptual-cognitive abilities.

Note that a number of additional or alternative latent factors could have emerged from this analysis. For example, children might have distinguished primarily between internal experience and external action (Gray et al., 2007), or they might have demonstrated finer-grained groupings of mental capacities based on phrasing, rote knowledge, etc. Instead, the latent conceptual structure underlying children's responses appears to be very similar to that of adults.

Children vs. adults. To formally compare responses from children and adults, we considered the full, combined dataset and examined factor scores by age group.

EFA using the combined dataset revealed three factors that met our retention criteria. Unsurprisingly, these three factors were very similar to those revealed for adults and children analyzed independently: They corresponded to social-emotional abilities, physiological sensations, and perceptual-cognitive abilities (see Table 1).

The purpose of this combined EFA was to examine differences in adults' and children's evaluations of beetles and robots within this 3-part structure. To do so, we derived factor scores (via the ten Berge method) using the rotated 3factor solution. This yielded 3 scores for each participant, corresponding, in principle, to holistic judgments of the social-emotional, physiological, and perceptual-cognitive abilities of the target character the participant evaluated. (Note that each of these 3 scores takes into account factor loadings for all 40 mental capacities, as listed in Table 1.)

This allowed us to examine the effects of age group (adult, child), character (beetle, robot), and factor (socialemotional, physiological, perceptual-cognitive) on these scores via mixed effects linear regression. See Table 2 for the results of a maximal model and Fig. 1 for mean scores.

Collapsing across age groups and domains (physiological, social-emotional, and perceptual-cognitive), factor scores suggest that participants generally attributed fewer mental capacities to the robot than the beetle $(b=-0.25)$. However, this appears to be entirely due to the huge discrepancy between characters in the physiological domain; the difference between characters was reduced to nothing in the social-emotional domain ( $b=0.26$ ), and reversed in the perceptual-cognitive domain ( $b=0.39$ ). Collapsing across entities (beetle, robot), children tended to attribute more mental capacities adults ( $b=0.19$ ), but this was driven primarily by the social-emotional domain ( $b=0.46$ ), and was reversed in the perceptual-cognitive domain ( $b=-0.30$ ).

Scores in the physiological and perceptual-cognitive domains were very similar for children and adults: Both

Table 1: Factor loadings from exploratory factor analyses for adults alone ( $n=200$ ), children alone ( $n=200$ ), and the combined dataset. Loadings are from 3-factor varimax-rotated minimum residual solutions. Items are grouped according to their dominant factor (the factor with the strongest factor loading) in the combined analysis; loadings $>0.60$ or $<-0.60$ are in bold. Items marked with an asterisk were accompanied by a brief definition (see main text).

| Item | Social-emotional |  |  | Physiological |  |  | Perceptual-cog. |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Do you think a [target] can...? | Ad. | Ch. | ALL | Ad. | Ch. | ALL | Ad. | Ch. | ALL |
| feel proud | 0.81 | 0.78 | 0.86 | 0.13 | -0.03 | 0.03 | 0.08 | -0.02 | -0.05 |
| feel happy | 0.77 | 0.76 | 0.83 | 0.33 | 0.07 | 0.18 | 0.05 | 0.04 | 0.02 |
| feel joy | 0.81 | 0.75 | 0.82 | 0.30 | 0.12 | 0.18 | 0.02 | -0.04 | -0.03 |
| feel sad | 0.80 | 0.66 | 0.77 | 0.26 | 0.27 | 0.23 | 0.04 | 0.02 | 0.00 |
| get hurt feelings | 0.70 | 0.66 | 0.77 | 0.21 | 0.19 | 0.16 | 0.04 | 0.10 | 0.00 |
| feel love | 0.76 | 0.63 | 0.74 | 0.26 | 0.11 | 0.16 | 0.14 | 0.00 | 0.03 |
| feel guilty | 0.69 | 0.59 | 0.71 | 0.14 | 0.06 | 0.07 | 0.06 | 0.06 | 0.00 |
| get angry | 0.51 | 0.50 | 0.67 | 0.38 | 0.31 | 0.30 | 0.15 | 0.05 | 0.04 |
| have beliefs* | 0.51 | 0.53 | 0.65 | -0.03 | -0.04 | -0.04 | 0.33 | 0.22 | 0.18 |
| feel embarrassed | 0.60 | 0.57 | 0.65 | 0.09 | 0.04 | 0.05 | 0.03 | -0.06 | -0.06 |
| have a personality ${ }^{*}$ | 0.50 | 0.51 | 0.64 | -0.05 | -0.06 | -0.06 | 0.26 | 0.30 | 0.20 |
| feel pleasure* | 0.47 | 0.62 | 0.64 | 0.55 | 0.09 | 0.30 | 0.08 | 0.02 | 0.04 |
| feel calm | 0.43 | 0.48 | 0.60 | 0.53 | 0.22 | 0.36 | 0.16 | 0.12 | 0.11 |
| have thoughts | 0.36 | 0.46 | 0.55 | 0.24 | 0.24 | 0.22 | 0.37 | 0.32 | 0.30 |
| know what's nice and what's mean | 0.42 | 0.47 | 0.54 | -0.20 | -0.18 | -0.19 | 0.34 | 0.20 | 0.22 |
| have desires* | 0.36 | 0.43 | 0.53 | 0.53 | 0.33 | 0.39 | 0.19 | 0.03 | 0.09 |
| understand how somebody else is feeling | 0.42 | 0.40 | 0.51 | -0.09 | -0.31 | -0.21 | 0.31 | 0.28 | 0.24 |
| have self-control* | 0.42 | 0.26 | 0.47 | 0.00 | 0.02 | 0.00 | 0.34 | 0.28 | 0.25 |
| have goals* | 0.21 | 0.37 | 0.42 | 0.16 | -0.17 | -0.01 | 0.42 | 0.22 | 0.29 |
| get hungry | 0.04 | 0.12 | 0.14 | 0.94 | 0.87 | 0.90 | -0.08 | -0.07 | -0.04 |
| do math | 0.05 | 0.14 | 0.05 | -0.83 | -0.71 | -0.79 | 0.36 | 0.34 | 0.31 |
| feel pain | 0.17 | 0.21 | 0.26 | 0.82 | 0.79 | 0.79 | 0.06 | 0.01 | 0.06 |
| smell things | 0.01 | -0.10 | -0.08 | 0.67 | 0.64 | 0.64 | 0.21 | 0.11 | 0.22 |
| feel scared | 0.32 | 0.39 | 0.46 | 0.75 | 0.53 | 0.62 | 0.13 | 0.06 | 0.10 |
| feel sick ${ }^{*}$ | 0.29 | 0.16 | 0.21 | 0.66 | 0.51 | 0.58 | 0.14 | -0.06 | 0.09 |
| feel tired | 0.24 | 0.27 | 0.41 | 0.72 | 0.46 | 0.58 | 0.22 | -0.01 | 0.10 |
| feel safe | 0.28 | 0.42 | 0.47 | 0.71 | 0.33 | 0.50 | 0.23 | 0.31 | 0.25 |
| figure out how to do things | 0.16 | 0.12 | 0.18 | 0.00 | -0.04 | -0.04 | 0.59 | 0.49 | 0.55 |
| be aware of things | 0.06 | 0.17 | 0.08 | 0.32 | 0.20 | 0.23 | 0.50 | 0.49 | 0.50 |
| sense whether something is close by or far away | -0.03 | 0.02 | -0.16 | 0.10 | 0.01 | 0.00 | 0.57 | 0.44 | 0.49 |
| remember things | 0.19 | 0.10 | 0.16 | -0.33 | -0.40 | -0.39 | 0.57 | 0.39 | 0.47 |
| sense temperatures | 0.00 | -0.12 | -0.26 | 0.19 | -0.13 | -0.03 | 0.51 | 0.42 | 0.46 |
| make choices | 0.14 | 0.28 | 0.23 | 0.08 | 0.18 | 0.09 | 0.57 | 0.36 | 0.46 |
| recognize somebody else | 0.21 | 0.18 | 0.14 | -0.45 | -0.16 | -0.34 | 0.61 | 0.32 | 0.46 |
| decide what to do | 0.09 | 0.31 | 0.20 | 0.09 | 0.28 | 0.14 | 0.48 | 0.40 | 0.45 |
| be aware of itself | 0.21 | 0.11 | 0.31 | 0.23 | 0.06 | 0.14 | 0.41 | 0.52 | 0.42 |
| hear sounds | 0.01 | -0.18 | -0.11 | 0.13 | 0.01 | 0.05 | 0.50 | 0.33 | 0.42 |
| see things | -0.03 | -0.13 | 0.03 | 0.24 | -0.05 | 0.11 | 0.55 | 0.23 | 0.40 |
| communicate with somebody else | 0.14 | 0.08 | 0.17 | -0.32 | -0.18 | -0.26 | 0.57 | 0.24 | 0.40 |
| make plans | 0.28 | 0.32 | 0.33 | -0.31 | -0.18 | -0.27 | 0.46 | 0.41 | 0.40 |
| \% variance ...3-factor solution: | 37\% | 50\% | 53\% | 37\% | 30\% | 28\% | 26\% | 20\% | 19\% |
| explained ...maximal (13-factor) solution: | 21\% | 25\% | 37\% | 25\% | 18\% | 20\% | 10\% | 7\% | 8\% |

Table 2: Results of a mixed effects linear regression of factor scores on target character, factor, and age group, with random intercepts by participant. Categorical predictors were effect-coded and compared to the grand mean (GM). "Significant" predictors $(|t|>2)$ are in bold.

| Predictor | $b$ | $s e$ | $t$ |
| :--- | ---: | :--- | ---: |
| (Intercept) | -0.01 | 0.02 | -0.25 |
| character (robot vs. GM) | $\mathbf{- 0 . 2 5}$ | $\mathbf{0 . 0 2}$ | $\mathbf{- 1 0 . 0 0}$ |
| factor 1 (vs. GM) | 0.00 | 0.03 | -0.06 |
| factor 3 (vs. GM) | 0.01 | 0.03 | 0.41 |
| age group (children vs. GM) | $\mathbf{0 . 1 9}$ | $\mathbf{0 . 0 2}$ | $\mathbf{7 . 6 5}$ |
| character * factor 1 | $\mathbf{0 . 2 6}$ | $\mathbf{0 . 0 3}$ | $\mathbf{8 . 6 4}$ |
| character * factor 3 | $\mathbf{0 . 3 9}$ | $\mathbf{0 . 0 3}$ | $\mathbf{1 3 . 0 0}$ |
| character * age group | 0.05 | 0.02 | 1.87 |
| factor 1 * age group | $\mathbf{0 . 4 6}$ | $\mathbf{0 . 0 3}$ | $\mathbf{1 5 . 1 0}$ |
| factor 3 * age group | $\mathbf{- 0 . 3 0}$ | $\mathbf{0 . 0 3}$ | $\mathbf{- 9 . 8 8}$ |
| character * factor 1 * age group | 0.00 | 0.03 | 0.12 |
| character * factor 3 * age group | -0.04 | 0.03 | -1.35 |

children and adults marked a clear difference between the robot and the beetle in the physiological domain (Fig. 1, center), in line with the animate-inanimate distinction ${ }^{1}$; and both age groups credited the robot with slightly greater perceptual-cognitive skills than the beetle (right). In contrast, in the social-emotional domain (left) both the beetle and the robot received rather low scores among adults, but very high scores among children. See Fig. 2 for raw counts of no, kinda, and yes responses for all items, grouped by character, age group, and dominant factor.

In sum, we see only minor differences between children and adults in their attributions of physiological and perceptual-cognitive abilities to beetles and robots-but a major difference in the social-emotional domain: Relative to adults, children tended to credit both beetles and robots with much greater social-emotional abilities.

## General Discussion

A bottom-up approach designed to shed light on children's intuitions about the ontology of mental life revealed an adult-like conceptual structure in place among 7 - to 9 -y-old children. Patterns of mental capacity attributions revealed a shared fundamental distinction between social-emotional, physiological, and perceptual-cognitive abilities. To our knowledge, this is the first bottom-up exploration of children's intuitions about the structure of mental life.

In a close parallel to adults (Weisman et al., 2016), the distinction that loomed the largest in children's responses

[^240]

Fig. 1: Mean factor scores for the beetle and the robot for each of the three factors (social-emotional, physiological, perceptual-cognitive), among adults ( $n=200$ ) and children ( $n=200$ ). Error bars are non-parametric bootstrap 95\% CIs.
was not between experience and agency (Gray et al., 2007), but between three varieties of experience: emotional, physiological, and perceptual. Echoing this previous work, different aspects of agency were distributed across these factors: The social-emotional factor included several items related to moral agency (e.g., understand how somebody else is feeling, know what's nice and what's mean), while items related to goal pursuit tended to pattern with perceptual-cognitive abilities (decide what to do, make plans). ${ }^{2}$ For both children and adults, connections between related varieties of experience and agency seemed to play a particularly important role in intuitive ontologies of mental life-perhaps because they allow us to explain and predict others' actions in several fundamental domains (interactions among social partners, the bodily needs of animals, and the goal-directed actions of agents).

Although the conceptual structure underlying children's mental capacity attributions was quite similar to that of adults', there was one striking difference in their evaluation of entities within that structure: Children were far more generous in their assessment of the social-emotional abilities of both beetles and robots. The specificity of this age difference-which emerged dramatically in one domain, but not others-suggests that this is unlikely to be due either to a general tendency toward "mentalizing" these characters (or a simple "yes" bias). But its extension to both beetles and robots raises many questions. With regard to robots, children growing up in the $21^{\text {st }}$ century might be converging on a new understanding of technological "beings" as inanimate objects with some degree of social-emotional life (see Kahn, Gary, \& Shen, 2013)—but this kind of historical conceptual change would not predict the high rates of social-emotional attributions to beetles that we observed. Our findings are perhaps more consistent with a general openness to untraditional social partners that extends into middle childhood (but not adulthood)-or with a difference in construals of what it means to feel proud, happy, guilty,

[^241]
## Responses by mental capacity item



Fig. 2: Raw counts of responses to each item, by character (beetle, robot), age group (children, adults), and dominant factor (social-emotional, physiological, perceptual-cognitive), listed in descending order of absolute factor loading (see Table 1). Opacity indicates responses of yes (dark), kinda (medium), or no (light).
etc. To what kinds of entities would children of this age deny social-emotional abilities, and how do they draw this line? What aspects of attributing pride, happiness, or guilt might change between 7-9y and adulthood?

Our findings point to the importance of distinguishing between different aspects of mental life in building theories of how social cognitive reasoning might evolve-both over the lifespan and across history and cultures. The current studies offer the major advantage of making these distinctions on the basis of children's own conceptual structure, rather than a priori categories generated by experimenters-an approach that could prove particularly powerful in making sense of children's beliefs about and behaviors toward the many kinds of human, animal, and technological "beings" in the modern social world.

## Acknowledgments

This material is based upon work supported by the NSF GRFP under Grant No. DGE-114747, and by a William R. \& Sara Hart Kimball Stanford Graduate Fellowship. Thanks to the Tech Museum of Innovation, Palo Alto Junior Museum \& Zoo, Bay Area Discovery Museum, Bing Nursery School, research assistants, and families.

## References

Carey (1985). Conceptual Change in Childhood. MIT Press. Flavell (1999). Cognitive development: Children's knowledge about the mind. Annu Rev Psychol, 50, 21-45.
Gelman \& Opfer (2002). Development of the animateinanimate distinction. In Goswami (Ed.), Blackwell Handbook of Childhood Cognitive Development (pp. 151-66). Blackwell Publishers Ltd.
Gray, H.M., Gray, K., \& Wegner (2007). Dimensions of mind perception. Science, 315(5812), 619.
Hatano \& Inagaki (1997). Qualitative changes in intuitive biology. Eur J Psychol Edu, 12(2), 111-30.
Kahn, Gary, \& Shen (2013). Children's social relationships with current and near-future robots. Child Dev Perspec, $7(1), 32-7$.
Piaget (1929). The Child's Conception of the World. Routledge \& Kegan Paul Ltd.
Weisman, Dweck, \& Markman (2016). Varieties of experience: A new look at folk philosophy of mind. In Proc Annu Meet CogSci Soc (pp. 2741-6).
Wellman \& Woolley (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. Cog, 35(3), 245-75.

# The Effect of expertise and biscriptalism on letter perception: The complexity benefit 

Robert W. Wiley (wiley@cogsci.jhu.edu)<br>Department of Cognitive Science, Johns Hopkins University, Krieger Hall<br>3400 N. Charles Street, Baltimore, MD 21218, USA<br>Brenda Rapp (rapp@cogsci.jhu.edu)<br>Department of Cognitive Science, Johns Hopkins University, Krieger Hall<br>3400 N. Charles Street, Baltimore, MD 21218, USA


#### Abstract

Previous work has demonstrated that the visual complexity of letter-shapes is processed differently by naïve and expert observers. Specifically, fluent readers of the Arabic alphabet were found to discriminate complex letters more readily than less complex letters, whereas naïve observers exhibited the opposite effect. This "complexity benefit", wherein complex letters confer a processing advantage to expert observers, is not yet well understood. In a new study, we investigate whether this effect generalizes across scripts, and whether it is unique to individuals with biscriptal experience (knowledge of reading two different scripts). The results of the three experiments confirm that the complexity benefit is characteristic of expert monoscriptal and biscriptal readers, and that, furthermore, there may be a biscriptal advantage in processing visual complexity.


Keywords: biscriptal; orthography; visual complexity; perceptual expertise

## Background

Letter perception and identification require detection and processing of a letter's component visual features (Grainger, Rey, \& Dufau, 2008). For example, Pelli and colleagues (Pelli, Burns, Farell, \& Moore-Page, 2006) determined that letters are identified by detecting $7 \pm 2$ visual features. While core properties of the human visual system certainly determine how and which visual features are detected for in letter identification, there is increasing evidence that the extent and type of experience with letters influences how the visual system processes them.

Wiley, Wilson, \& Rapp (2016) examined the effects of both alphabet and expertise on Arabic letter perception by comparing same/different letter judgments of expert, biscriptal Arabic-English readers, and naïve, monoscriptal English-only readers. Among the findings was that letter complexity, defined as the number of visual features in a letter ${ }^{1}$, was associated with slower/less accurate responses for naïve observers, but faster/more accurate responses for expert, biscriptal readers. This finding suggests that

[^242]extensive experience leads to more efficient visual processing of complex shapes. In other words, whereas for naïve observers, complex letter-shapes are more difficult to discriminate than are simple ones, for expert observers the reverse is true. This effect was referred to as the "complexity benefit". As a first step to furthering our understanding of the complexity benefit, the current study seeks to determine (a) whether the complexity effect is specific to Arabic, and (b) whether the magnitude of the effect is related to the amount of experience with a specific script or if extends across scripts.

## The Current Study

Whereas Wiley, Wilson, \& Rapp (2016) focused only on comparing the effects in letter perception of the amount of expertise (naïve or expert observers), it is also the case that those participants can be divided along another dimension: monoscriptal and biscriptal. Here, we make use of the biscriptal experience to better understand the nature of the complexity benefit. Specifically, we address two questions:

Question 1: Is the complexity benefit limited to Arabic letters?

Question 2: Is the complexity benefit affected by the amount of expertise with a script?

Question 3: Does biscriptalism affect the perception of Roman letters?

The answers to these questions have implications for our understanding of whether and how the visual system is affected by extensive reading experience. There are at least two relevant hypotheses that are evaluated: (1) the complexity benefit is a consequence of extensive experience with letter identification within a specific set (e.g. the Roman alphabet). In that case, expertise with one script should have no bearing on the visual processing of another. (2) The complexity benefit may be related to the manner in which visual features are processed, regardless of the letter in which they appear; in this case, expertise with one script may influence the processing of another, depending on the extent to which they make use of similar sets of visual
features. This latter possibility would support a "biscriptal advantage", such that biscriptal Arabic-English readers should show a greater complexity benefit than monoscriptal readers. In addressing these questions, we also determine whether or not our original finding of a "complexity benefit" is replicable, whether or not it is an artifact of the Arabic alphabet, and whether or not it is true of monoscriptal as well as biscriptal individuals.

The questions are addressed in three experiments. In Experiment 1, the experimental protocol from Wiley, Wilson, \& Rapp (2016) was used with a considerably larger sample size of monoscriptal Roman-only readers, with implementation in Amazon's Mechanical Turk, with both Arabic and Roman letters. This experiment directly tests whether the complexity benefit is unique to the Arabic alphabet, or whether it is also present in monoscriptal participants viewing the Roman alphabet.

Experiment 2 is a re-analysis of the data from Wiley, Wilson, \& Rapp (2016), specifically the reaction time measurements from the same-different judgment task with pairs of Arabic letters. We separate the expert, biscriptal participants into two groups, one low-proficiency and the other high-proficiency, to shed light on whether the amount of expertise with reading a script affects the magnitude of the complexity benefit.

Finally, in Experiment 3 we use the same protocol as in Experiment 2 with new samples of both monoscriptal (Roman-only) and biscriptal (Arabic \& Roman) participants, viewing both Arabic and Roman letters. This experiment allows us to address whether expertise with reading one script affects the perception of a second script, specifically evaluating whether or not being biscriptal provides an advantage in terms of the complexity benefit.

## Experiment 1: Is the complexity benefit limited to Arabic letters?

Following Wiley, Wilson, \& Rapp (2016): we used a samedifferent judgment task with pairs of letters, using lettershapes from both the Arabic and Roman alphabet. For all experiments, the questions of interest are addressed on the basis of reaction times (RT, on correct trials), analyzed using linear mixed-effects modeling (LMEM; including random intercepts and slopes by both participants and items).

## Participants

167 participants were recruited online via Amazon's Mechanical Turk (MTurk), receiving payment of \$7.50/hour for their participation. 86 participants completed the task with Arabic letters and 81 with Roman letters. All participants reported no knowledge of any language written in a non-Roman script, and thus all are considered monoscriptal (MS).

## Stimuli

A set of 23 letter-shapes from the Arabic alphabet was presented in Adobe Arabic, in font size 24 (stimuli subtended $0.17^{\circ}-0.31^{\circ}$ and $0.05^{\circ}-0.35^{\circ}$ of visual angle, respectively in the vertical and horizontal dimensions). A set of 23 letter-shapes from the Roman alphabet was also presented in Arial, font size 16, thereby equating the size range of the two alphabets.
Both sets of stimuli included 8 pairs of allographs (i.e. 8 letters were presented with two letter-shapes, such as "A" and "a"; see Table 2). The stimuli are listed in Tables 1 and 2.

## Procedure

Each trial began with a central fixation cross ( 250 ms ), which disappeared and was replaced by a pair of letters simultaneously on either side of fixation, 48 pixels apart. Each pair of letters was presented for 2000 ms or until a response of "same" or "different" (by pressing either the "a" or "l" key on the keyboard was recorded. After a response or two-second timeout there was a 500 ms intertrial blank screen. Participants completed either the task in Arabic or in Roman letters but not both; the ratio of same to different trials was $40 / 60$, for a total of 437 trials.

## Analysis

Using only correct responses, a single LMEM was fit to the "same" pairs ${ }^{2}$ data to determine the effect of complexity (number of visual features from a list of 14) on reaction time, and whether this effect differed across groups. The regression model was fit using R ( R Core Team, 2015), package lme4 (Bates, Mächler, Bolker, \& Walker, 2015), and confidence intervals for the parameters of interest were determined using parametric bootstrapping; plots are provided based on the R package effects (Fox, 2003).

Regression predictors: For the fixed effects, two predictors of interest were included: the categorical variable Alphabet (Arabic or Roman, with sum-coding) and the continuous variable Complexity (total number of visual features, ranging from 4-12). Two additional predictors were included as control variables: Trial Order and Previous RT (reaction time on the preceding trial), to control for trends in RT across the duration of the experiment. Finally, we included the interaction Alphabet X Complexity.

The following crossed random effects were included: random intercepts were included both by participants and by items, as was a random slope for the effect of Complexity by participant.

[^243]
## Results

The results are reported in Figure 1. Confidence intervals are based on 1,000 bootstrap simulations.

The estimated complexity benefit is significant for Roman letters, beta $=-0.014[-0.017,-0.010]$, whereas there is a significant increased RT as the number of features increases for the Arabic alphabet, beta $=0.005$ [0.002, 0.009]. The interaction between alphabets is significant, a beta difference $=0.019$ [0.015, 0.024].

Summary: The finding that the complexity benefit also exists for monoscriptal participants in the Roman alphabet (in which they are experts) is replicated in a large MTurk sample, as is the finding that the opposite effect (slower RT on more complex letters) for monoscriptal observers with no experience in reading Arabic.


Figure 1: Experiment 1, predicted RT (ms) as a function of complexity (\# of visual features) in the Arabic (green) and Roman (red) alphabets, measured in response to "same" pairs.

## Experiment 2: Is the complexity benefit affected by amount of expertise?

Experiment 2 is a reanalysis of data originally presented in Wiley, Wilson, \& Rapp (2016). The procedure was the same as that described for Experiment 1, with the following differences in participants and stimuli.

## Participants

There were 34 participants, all from the Johns Hopkins University community, who took part in two one-hour sessions, receiving either course credit or $\$ 20$ for their participation. The participants were organized into three groups:

Low-proficiency biscriptal (L-BS, $\mathrm{n}=11$ ): individuals whose first written language is English and who have had 23 years of studying Arabic.

High-proficiency biscriptal (H-BS, $\mathrm{n}=11$ ): individuals who learned to read and write Arabic simultaneously with English, or as a second language with at least 4 years of study.

Monoscriptal (MS, $\mathrm{n}=12$ ): consists of participants whose first language is English, and who have had no exposure to reading or writing in non-Roman scripts ${ }^{3}$.

## Stimuli

The stimuli were a superset of the Arabic letters used in Experiment 1, for a total of 45 shapes. However only the 23 stimuli used for Experiment 1 are analyzed here in order to better compare results across experiments.

## Procedure

Stimuli were presented using E-Prime 2.0 (Psychology Software Tools, Pittsburg, PA). Participants completed the experiment over two sessions, with each session consisting of 990 trials with a $50 / 50$ ratio of $50 / 50$ same to different trials. For this analysis, a total of 506 trials were used.

## Analysis

The same analysis was used as in Experiment 1.
Regression predictors: The model structure was the same as outlined in Experiment 1 except that the predictor Alphabet replaced by the predictor Group (MS, L-BS, or HBS, with sum-coding).

Table 1: Arabic letter-shapes and their complexity, the mean RTs across Experiments 2 and 3, for each group for each letter, and the correlation between complexity and mean RT (bottom row).

| Letter | Complexity | MS | L-BS | H-BS |
| :---: | :---: | :---: | :---: | :---: |
| b | 10 | 563 | 549 | 622 |
| $\tau$ | 7 | 583 | 567 | 666 |
| , | 4 | 567 | 536 | 625 |
| $\varepsilon$ | 10 | 575 | 557 | 639 |
| - | 10 | 573 | 560 | 648 |
| $\varepsilon$ | 10 | 591 | 593 | 702 |
| ب | 7 | 556 | 569 | 605 |
| j | 6 | 579 | 582 | 668 |
| ظ | 11 | 609 | 586 | 617 |
| غ | 11 | 604 | 579 | 608 |
| $\otimes$ | 6 | 557 | 559 | 614 |
| - | 8 | 572 | 558 | 619 |
| ج | 8 | 607 | 578 | 631 |
| ك | 6 | 606 | 573 | 616 |
| 5 | 6 | 567 | 552 | 611 |
| $\checkmark$ | 5 | 552 | 554 | 588 |
| $\lrcorner$ | 4 | 603 | 568 | 668 |
| ن | 7 | 587 | 571 | 610 |
| - | 4 | 573 | 613 | 709 |
| , | 4 | 562 | 571 | 605 |
| $\omega$ | 12 | 579 | 566 | 593 |
| - | 9 | 576 | 581 | 634 |
| j | 5 | 577 | 570 | 645 |
|  | $r=$ | 0.298 | 0.071 | -0.148 |

[^244]
## Results

The re-analysis of data from Wiley, Wilson, and Rapp (2016) is reported in Figure 2 based on the LMEM as previously described; confidence intervals are based on 1000 bootstrap simulations.

The estimated beta-weight for the effect of Complexity is for the MS group $=0.015,95 \%$ CI [0.010, 0.020] Thus, we again find that among the naïve (monoscriptal) participants more complex letters lead to significantly slower reaction times.

For the biscriptal groups, for the L-BS the effect is estimated $=-0.002[-0.008,0.002]$; and for the $\mathrm{H}-\mathrm{BS}=-$ 0.007 [-0.013, -0.001$]$. Thus, only the H-BS show a significant complexity benefit, while the L-BS show only a trend toward faster RT on more complex letters.

The estimated difference between the MS and L-BS is $=$ 0.017 [0.011, 0.024], and between the MS and H-BS $=$ 0.022 [0.014, 0.029]. Both biscriptal groups show significantly more negative (hence, more of a complexity benefit) than the monoscriptal group. The estimated difference between the two biscriptal groups $=0.005[-$ $0.003,0.011]$, with a nonsignificant trend toward a greater complexity benefit for the H-BS relative to the L-BS.

Summary: Both biscriptal groups show a numerically larger complexity benefit than the monoscriptal group; although only for the H-BS group is the complexity benefit statistically significant.


Figure 2: Experiment 2, Predicted RT (ms) as a function of complexity (\# of visual features) in the Arabic alphabet, measured for each group of participants in response to "same" pairs.

## Experiment 3: Does biscriptalism affect the perception of Roman letters?

The same procedure as outlined in Experiment 1 was used, with a few differences noted as follows.

## Participants

29 students from Johns Hopkins University (ages 18-22), all different from those in Experiment 2, took part in the onehour experiment, receiving either course credit or $\$ 10$ for their participation. The participants were divided into L-BS
$(\mathrm{n}=7), \mathrm{H}-\mathrm{BS}(\mathrm{n}=5)$, and MS $(\mathrm{n}=17)$ for a total of 29 participants.

## Stimuli

The stimuli were identical to those used in Experiment 1.

## Procedure

The same procedure as Experiment 1 was used, except participants completed the task for both alphabets separately across two sessions, with the order (Arabic-Roman or Roman-Arabic) counterbalanced across participants.

## Analysis

The same analysis as described for Experiment 1 was conducted, plus the addition of the variable Group (MS, LBS, or H-BS, sum-coded) and the 3-way interactions of Alphabet X Group X Complexity and Alphabet X Group X Previous RT. The random effects structure was the same as in Experiment 1, with the addition of (correlated) random slopes for the effect of Alphabet by participants.

Table 2: Roman letter-shapes and their complexity, the mean RT from Experiment 3, for each group for each letter, and the correlation between complexity and mean RT

| Letter | (bottom row). |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Complexity | MS | L-BS | H-BS |
| a | 10 | 532 | 506 | 605 |
| A | 10 | 507 | 497 | 563 |
| b | 8 | 528 | 553 | 613 |
| B | 14 | 518 | 527 | 595 |
| C | 6 | 545 | 530 | 627 |
| d | 8 | 519 | 541 | 658 |
| D | 7 | 512 | 510 | 593 |
| E | 12 | 520 | 542 | 584 |
| g | 9 | 516 | 536 | 626 |
| G | 7 | 530 | 537 | 612 |
| I | 5 | 547 | 552 | 648 |
| j | 5 | 529 | 539 | 631 |
| J | 4 | 526 | 519 | 646 |
| 0 | 5 | 522 | 545 | 628 |
| q | 9 | 523 | 538 | 657 |
| Q | 9 | 517 | 531 | 593 |
| r | 6 | 529 | 540 | 658 |
| R | 10 | 519 | 519 | 623 |
| S | 7 | 528 | 520 | 600 |
| t | 9 | 538 | 543 | 643 |
| T | 8 | 523 | 522 | 585 |
| W | 11 | 533 | 528 | 642 |
| X | 8 | 546 | 544 | 619 |
|  | $r=$ | -0.353 | -0.231 | -0.424 |

## Results

The results from participants completing the same-different task with both alphabets are reported in Figure 3 (Arabic) and Figure 4 (Roman). Confidence intervals are based on 1,000 bootstrap simulations.

For the Arabic alphabet, the MS show significantly slower RTs on more complex letters, beta estimated $=0.013$
[0.007, 0.017$]$. The L-BS show a nonsignificant trend in the same direction, beta $=0.007[-0.001,0.014]$, whereas the HBS show a nonsignificant trend toward a complexity benefit, beta $=-0.007[-0.018,0.001]$.

While the complexity benefit is not significant within either biscriptal group, the difference between the H-BS both of the other two groups is significant: the H-BS betaweight is significantly different than that for the MS, by an estimated $0.019[0.009,0.031]$ and than the L-BS by beta $=$ 0.014 [0.003, 0.025]. The difference between the MS and LBS is not significant $($ beta $=0.006[-0.003,0.014])$.


Figure 3: Experiment 3, predicted RT (ms) as a function of complexity (\# of visual features) in the Arabic alphabet, measured for each group of participants in response to "same" pairs.

For the Roman alphabet, the MS show a significant complexity benefit $=-0.005[-0.010,-0.0002]$. The L-BS show a marginally significant complexity benefit, beta $=-$ 0.008 [-0.016, 0.0001]. The H-BS show a significant complexity benefit, beta $=-0.020[-0.029,-0.013]$.

Finally, the $\mathrm{H}-\mathrm{BS}$ show a significantly greater effect than both the MS (beta $=0.015[0.005,0.026])$ and the L-BS (beta $=0.013[0.002,0.025])$. There is no difference between the MS and the L-BS (beta $=0.002[-0.006$, 0.011]).


Figure 4: Experiment 3, Predicted RT (ms) as a function of complexity (\# of visual features) in the Roman alphabet, measured for each group of participants in response to "same" pairs.

Summary: We find a similar pattern of results for the Arabic alphabet as in Experiment 2 with the H-BS group showing a significantly greater complexity benefit than either then MS or the L-BS groups. Critically, a complexity benefit is found for the Roman alphabet for all groups, including the monoscriptal (English-only) participants, indicating that it is not an artifact of the Arabic alphabet or of being biscriptal. The magnitude of the effect is significantly greater in the H-BS than in either the MS or LBS groups, suggesting a possible biscriptal advantage.

## Discussion

We investigated the role that expertise and biscriptalism play in the visual processing of letter-shapes. Specifically, we sought to determine whether: (1) the complexity benefit, wherein expert readers of a script identify complex letters significantly more quickly than simpler letters, occurs for scripts other than Arabic where it was first reported, (2) the complexity benefit is limited to biscriptal individuals or is present also in monoscriptals, and (3) there is a biscriptal advantage for visual processing of letters, such that biscriptals show a greater complexity advantage or if, instead, the magnitude of the complexity benefit is simply tied to the amount of experience with a script. There were three participant groups: monoscriptal, English-only readers (MS), and two biscriptal Arabic-English reader groups, one with four or more years of experience (H-BS) and one with two or three years (L-BS). We used LMEM to determine the direction and strength of the relationship between letter complexity (as defined by the number of visual features), and whether this relationship differs across groups of participants and across alphabets.

The results of Experiments 1 and 3 both reveal that the complexity benefit is not an artifact of the Arabic alphabet. Monoscriptal participants who participated in the laboratory or the MTurk experiments all exhibited a complexity benefit when performing the same-different task with Roman letter stimuli. Thus, it would seem that the complexity benefit is not only a general trait of reading expertise, but also is not unique to individuals with biscriptal experience.

Additionally, the results of Experiments 2 and 3 provide further details regarding the complexity benefit phenomenon. While a significant complexity benefit was not limited to biscriptal individuals, the effect was greatest in the high-proficiency biscriptal individuals. This group showed a larger complexity benefit than the other two groups in both Arabic and Roman scripts. This is particularly interesting, given that the monoscriptal and biscriptal participants presumably had comparable expertise with the Roman alphabet. In fact, if anything the monoscriptal participants are likely to have had more experience with the Roman alphabet, as the biscriptal participants would have spent some of their time reading in Arabic instead of Roman letters. It is possible that this division of reading time between the two scripts may
underlie the overall slower reaction times exhibited by this group, analogous to the rationale provided for some of the findings in the literature on spoken word production with bilinguals (Gollan et al., 2008). This possibility will require more targeted experimental work. Nonetheless, the larger complexity benefit observed for the high expertise biscriptal participants indicates that there may be a biscriptal advantage for processing visual complexity, at least for letters.

The mechanism underlying the complexity benefit itself is not yet well understood. There are multiple possible explanations for why expert observers learn to identify more complex letters more quickly or accurately. One possibility is that expertise leads to the creation of new visual features- such that features are "bundled" together, making a complex letter no longer complex. For example, the letter " $w$ " may not be processed as four slanted lines, three intersections, two terminations, with symmetry and cyclicity, but instead as fewer features or even a single feature, "w". This type of expertise effect is consistent with findings in perceptual learning research (e.g. Goldstone, 1998; Kellman \& Garrigan, 2009; Sireteanu \& Rettenbach, 2000).

Another explanation for the complexity benefit is that it is related to the distinctiveness of letter-shapes within the set of shapes being processed. Under such an account, a complex letter like "w" may be easier to identify because its greater number of features provide more possible ways to distinguish it from other letters. This is compatible with findings from visual crowding effects, indicating that a target is easier to identify within an array of distractors if it is relatively more complex than those distractors (Bernard \& Chung, 2011; Chanceaux, Mathôt, \& Grainger, 2014). Accordingly, with increasing expertise, one learns not only the visual properties of each of the letters, but the distribution of features across the set of letters.

Relatedly, it may be that experts learn a greater number of ways to identify complex letters relative to simpler letters, allowing the identification process to terminate sooner. For example, whereas an observer with minimal experience may identify " $w$ " only after considering all of its features, an expert may identify it as soon as some distinct combination of features (a subset of the total number of features) are recognized. In this case, a "simple" letter such as ' 1 " may be more difficult to distinguish from other letters, because while a complex letter like " $w$ " can be identified without full consideration of all of its features and without searching for the absence of certain features, an " 1 " does not afford these opportunities.

Of these possibilities, perhaps the one most consistent with a biscriptal advantage would be the creation of new complex features from simpler features-biscriptal individuals' expertise with a wider range of letter-shapes may result in a larger feature 'vocabulary' that allows relatively more complex shapes to be more readily processed. In future research, it will be important to examine if the biscriptal complexity advantage extends to
other types of visual stimuli, and to identify evidence to adjudicate between possible mechanisms that support the complexity benefit.

## Acknowledgments

The research reported here was supported by the Science of Learning Institute at Johns Hopkins University, Distinguished Science of Learning Fellowship to R.W.W. and by NIH grant DC006740 to B.R. We acknowledge Melissa Greenberger for her contributions to this project.

## References

Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1-48.
Bernard, J.-B., \& Chung, S. T. L. (2011). The dependence of crowding on flanker complexity and target-flanker similarity. Journal of Vision, 11(8), 1-1.
Chanceaux, M., Mathot, S., \& Grainger, J. (2014). Effects of number, complexity, and familiarity of flankers on crowded letter identification. Journal of Vision, 14(6), 77.

Fox, J., \& Hong, J. (2009). Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package. Journal of Statistical Software, 32(1), 1-24.
Goldstone, R. L. (1998). Perceptual learning. Annual Review of Psychology, 49(1), 585-612.
Gollan, T., Montoya, R., Cera, C., \& Sandoval, T. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. Journal of Memory and Language, 58(3), 787814.

Grainger, J., Rey, A., \& Dufau, S. (2008). Letter perception: from pixels to pandemonium. Trends in Cognitive Sciences, 12(10), 381-387.
Kellman, P. J., \& Garrigan, P. (2009). Perceptual learning and human expertise. Physics of Life Reviews, 6(2), 5384.

Palmer, S. E. (1999). Vision science: Photons to phenomenology (Vol. 1). Cambridge, MA: MIT Press.
Pelli, D. G., Burns, C. W., Farell, B., \& Moore-Page, D. C. (2006). Feature detection and letter identification. Vision Research, 46(28), 4646-4674.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Sireteanu, R., \& Rettenbach, R. (2000). Perceptual learning in visual search generalizes over tasks, locations, and eyes. Vision Research, 40, 2925-2949.
Wiley, R. W., Wilson, C., \& Rapp, B. (2016). The Effects of Alphabet and Expertise on Letter Perception. Journal of Experimental Psychology: Human Perception and Performance, 42(8), 1186-1203.

# Examining Multiscale Movement Coordination in Collaborative Problem Solving 

Travis J. Wiltshire (wiltshire@sdu.dk)<br>Centre for Human Interactivity, Department of Language \& Communication, Campusvej 55 DK-5230 Odense M, Denmark<br>Sune Vork Steffensen (s.v.steffensen@sdu.dk)<br>Centre for Human Interactivity, Department of Language \& Communication, Campusvej 55<br>DK-5230 Odense M, Denmark


#### Abstract

During collaborative problem solving (CPS), coordination occurs at different spatial and temporal scales. This multiscale coordination should, at least on some scales, play a functional role in facilitating effective collaboration outcomes. To evaluate this, we conducted a study of computer-based CPS with 42 dyads. We used cross-wavelet coherence as a way to examine the degree to which movement coordination is evident at a variety of scales and tested whether the observed coordination was greater than both the amount expected due to chance and due to task demands. We found that coordination at scales less than 2 s was greater than expected due to chance and at most scales (except 16s, 1m, and 2m) was greater than expected due to task demands. Lastly, we evaluated whether the degree of coherence at scales less than 2 s , and the form of coordination (in terms of relative phase), were predictive of CPS performance. We found that .25 s and 1 s scales were predictive of performance. When including relative phase, our results suggest that higher in-phase movement coordination at the 1 s scale was the strongest predictor of CPS performance. We discuss these findings and detail their relevance for expanding our knowledge on how coordination facilitates CPS.


Keywords: coordination; collaboration; problem solving; team performance; dynamical systems; synchrony.

## Introduction

Collaborative problem solving (CPS) is a cognitive skill pervasive in many human interaction contexts ranging from everyday life to highly complex work environments. CPS is defined as "a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution" (OECD, 2015, p. 6). Given the increasing complexity of problems in contemporary societal practices, and the need for multiple disciplines to solve them, CPS has been recognized as an essential $21^{\text {st }}$ century skill. However, while some research has examined CPS in a variety of laboratory (e.g., Berg, Johnson, Meegan, \& Strough, 2003; Roschelle \& Teasley, 1995) and naturalistic contexts (e.g., Fiore, Wiltshire, Oglesby, O’Keefe, \& Salas, 2014; Jordan \& McDaniel Jr, 2014), the state of the science is still limited. In this paper, we explore the implications of the notion that CPS is a multiscale phenomenon, and investigate the degree to which movement coordination, at various scales, plays a functional role in effective CPS.

Human interaction in any context is a dynamic, multiscale phenomenon (e.g., Dale, Fusaroli, Duran, \& Richardson,

2013; Steffensen \& Pedersen, 2014). For example, during a conversation, neural events transpire on the order of milliseconds, speech production and gestures over seconds, and the conversation itself on the order of minutes (Hasson, Ghazanfar, Galantucci, Garrod, \& Keysers, 2012). This point, while oversimplified, illustrates the fact that human interaction involves a variety of temporal and spatial scales (e.g., neural, physiological, bodily). Recognizing human interaction as multiscale, especially during CPS, implies that coordination, both intra- and inter-personally, must span a variety of these spatial and temporal scales in order to effectively accomplish joint goals (Eiler, Kallen, Harrison, \& Richardson, 2013).

Indeed, a common question, particularly in the movement sciences, has been to understand how systems with high degrees of freedom, are able to functionally coordinate (e.g., Mitra, Amazeen, \& Turvey, 1998). Coordination in this context is simply the ways in which components and processes of a system change together over time (Butner, Berg, Baucom, \& Wiebe, 2014). In an interpersonal context, a wide variety of terms have been used to describe different forms of coordination (Butler, 2011) such as synchronization, co-regulation, entrainment, and coupling. Evidence for many forms of interpersonal coordination are quite pervasive amongst differing modalities (Fusaroli \& Tylén, 2016; Louwerse, Dale, Bard, \& Jeuniaux, 2012) and contexts (Palumbo et al., 2016). But, while many different forms of coordination have been discovered, the ways in which they facilitate effective interaction outcomes is less studied (cf., Timmons, Margolin, \& Saxbe, 2015), particularly in collaborative contexts.

Prior research has suggested that coordination is required for the accomplishment of joint goals (Mills, 2014) and that stronger coordination should contribute to better collaborative results (Barron, 2000). Findings so far have been mixed, though, with regard to how coordination, albeit in different modalities and scales, relates to optimal performance on joint tasks (Gallotti, Fairhurst, \& Frith, 2017). In one example, performance on a dyadic movement task was predicted by a measure of coordination that reflected interaction across multiple time scales (Davis, Brooks, \& Dixon, 2016). In another example, Abney, Paxton, Dale, and Kello (2015) found that stronger coordination in bodily movements were associated with poorer performance on a movement-based dyadic problem solving task. Louwerse et al. (2012) found that as task difficulty increased, so too did coordination. Further,
coordination of bodily movements in psychotherapy, a highly collaborative endeavor (Tryon \& Winograd, 2011), were shown to link to effective treatment outcomes (Ramseyer \& Tschacher, 2011). While not related to performance specifically, a number of interactional benefits have been observed following periods of interpersonal coordination such as increased affiliation (Hove \& Risen, 2009) and cooperative behavior (Valdesolo, Ouyang, \& DeSteno, 2010; Wiltermuth \& Heath, 2009).

Given the extant research, briefly reviewed here, we expect that multiscale coordination of bodily movements should have a functional relationship with performance in complex, CPS. Relatedly, a recent theoretical account of dialog proposed that high coordination in lower level behaviors (e.g., posture) may provide a necessary foundation for more variability and complementarity at higher levels of dialog (Fusaroli, Rączaszek-Leonardi, \& Tylén, 2014). So, a key aspect that distinguishes the present work from prior research is that we focus on movement coordination in a challenging computer-based CPS context.

## The Current Study

The present work is part of a larger study examining team interaction dynamics during a dyadic CPS task (Wiltshire, 2015; Wiltshire, Butner, \& Fiore, 2017). Whereas in prior work, we have examined how transitions in communication structures and their complexity relate to CPS performance (Wiltshire et al., 2017), here we focus on coordination of bodily movements at various time scales and how that coordination relates to CPS performance.

We utilize cross-wavelet coherence as a way of examining coordination of human interaction that is largely unstructured, at least when compared to rhythmic movement tasks (Fujiwara \& Daibo, 2016; Issartel, Bardainne, Gaillot, \& Marin, 2015). This method allows for evaluation of the degree of coordination of two continuous time series and whether that coordination is in-phase or anti-phase. One of its key strengths is that it retains a high level of precision in both the time and frequency domains (Issartel et al., 2015). This method has been used previously to examine movement coordination in a variety of interactive contexts such as the exchange of jokes (Schmidt, Morr, Fitzpatrick, \& Richardson, 2012), dialog (Fujiwara \& Daibo, 2016), the coordination of jazz musicians (Walton, Richardson, Langland-Hassan, \& Chemero, 2015), and dancers (Washburn et al., 2014).

We expect that the coordination of bodily movements will serve a functional role in facilitating effective CPS performance. However, we also expect that this functional role will vary based on time scales. In other words, movement coordination at some time scales should be more relevant to CPS than others. We thus adopt an exploratory approach to determine what scales are important in predicting effective CPS performance. Given the nature of the task, it is likely that smaller time scale movements will matter such as those that occur while controlling the computer-based task as well as during speech. When
examining interpersonal coordination dynamics, it is essential to demonstrate that the observed coordination is greater than can be expected due to chance alone (Ramseyer \& Tschacher, 2010), and that it is not solely due to task constraints (Strang, Funke, Russell, Dukes, \& Middendorf, 2014). Thus, we advance the following research hypotheses $(\mathrm{H})$ and research questions (RQ):

- H1: Movement coordination will be greater than chance, at least at lower scales.
- H2: Movement coordination will be greater than can be expected due to task constraints.
- RQ1: At what scales does movement coordination predict CPS performance?
- RQ2: Does the form of coordination (e.g., in-phase, anti-phase) at these scales relate to performance?


## Method

## Participants

84 undergraduate students (31 female, $M_{\text {age }}=19.2$ years, range 18-28 years; ~ 67\% White, $8 \%$ Black, 10\% Hispanic, $10 \%$ Asian, and 5\% other) from a large United States university voluntarily participated in this experiment comprising 42 dyadic teams. There were five female-only teams, 17 male-only teams, and 20 mixed-gender teams. Participants must have had general video game experience using a mouse and keyboard for third-person video games, no prior history of seizures, no experience using the Moonbase Alpha simulation, and no prior acquaintance.

## Materials

Participants sat face-to-face with each other with two desktop computers offset to one side. This setup allowed them to view the other's face and torso. The computer screens were placed back-to-back. A Logitech HD webcam model C615 was used to record the participants from a profile view. All videos were collected in 720p resolution.

## Task

NASA's Moonbase Alpha is a complex, CPS task (NASA, 2011) that places team members in a simulated scenario where a meteor strike damages critical life support systems of a moonbase. The goal of the Moonbase Alpha task is for participants to fully restore oxygen to the settlement in 25 minutes or less. Both team members must work together to solve the problem by figuring out how to fix and/or replace damaged components of the life support system such as solar panels, power cables, couplers, and a power distributor. A variety of tools and coordination strategies must be employed to complete the task; however, there are no predefined guidelines for how to completely repair the settlement in the given timeframe.

## Procedure

Participants were briefed about the nature of the experiment and asked to introduce themselves to each other by
providing a greeting and sharing their name with the other participant. Participants were then given an informed consent document to review and asked to complete a biographical questionnaire.

Participants were then provided a PowerPoint tutorial that covered the basics of the Moonbase Alpha simulation, which were derived from the simulation's instruction manual (NASA, 2011). Further, participants were told that they would be tested on the content. After completing the PowerPoint, they received a 10 -item multiple-choice knowledge assessment (see Wiltshire, 2015).

After completion of the knowledge assessment, the necessity for communication to complete the task was reiterated. Participants were then instructed to begin the simulation. A short video introduced the problem (i.e., the moonbase was damaged by a meteorite and life support functions need to be restored) before participants began the 25-minute task. The task was considered complete either when time ran out or once participants fully restored oxygen, whichever came first.

## CPS Performance

Problem solving performance was determined by a rescaled combination of three variables: (a) the total time taken to restore life support ( $0-25$ minutes), (b) the total percentage of oxygen restored $(0-100 \%)$, and (c) a ratio of completed object repairs to the total possible repairs ( $0-25$; only for teams that restored zero oxygen). The rescale function in R (R Core R Core Team, 2016) was used to place teams whose performance restored no oxygen at all into a range of $0-33$ as a function of their ratio of object repairs/total possible object repairs. Those teams that restored some, but not all, oxygen were rescaled to fit a range of 34-66. Lastly, for those teams that restored all oxygen, the time to complete the task was inversely rescaled to fit the range of 67-100 (with lower times leading to higher scores).

## Analytic Strategy

Frame-Differencing We used Paxton and Dale's (2013) video frame-differencing technique to extract a time series representing the level of bodily movement for each participant from all videos at an 8 Hz sampling rate. For the current task, this measure of bodily movement captures behaviors such as speech, postural sway, gestures, adjustment of position, hand movements controlling the mouse and keyboard, and shifting of the legs and/or feet. In general, this technique provides an objective measure of the amount of movement a given participant is exhibiting moment-by-moment over the duration of the task with higher values corresponding to more movement. The tradeoff when using this type of method is that there is a loss of specificity with regard to the types of movements that are coordinated, but movement can be extracted with relatively little effort and time compared to more specific movement coding systems (cf. Louwerse et al., 2012; Paxton \& Dale, 2013).

Cross-Wavelet Coherence We examined dyadic movement coordination with the cross-wavelet transformation method by using the wtc function from the biwavelet package (Gouhier, Grinsted, \& Simko, 2016) in R. This is a spectral decomposition method that allows for examination of time localized oscillations in a variety of frequencies and how the spectrum changes in those frequencies over time (Issartel, Marin, Gaillot, Bardainne, \& Cadopi, 2006). This method is known to be robust to nonstationary time series (Issartel et al., 2015). We extracted the average coherence and average relative phase values from the following frequency ranges: $.25 \mathrm{~s}, .5 \mathrm{~s}, 1 \mathrm{~s}, 2 \mathrm{~s}, 4 \mathrm{~s}, 8 \mathrm{~s}, 16 \mathrm{~s}, 32 \mathrm{~s}, \sim 1 \mathrm{~m}, \sim 2 \mathrm{~m}$, and $\sim 4.5 \mathrm{~m}$ within $+/-.5$ scales (frequency scales are converted to time domain by multiplying them by the 8 Hz sampling rate and dividing by 60 for minutes). Coherence is the spectral equivalent to a cross-correlation. Values of 0 convey no coordination and a value of 1 conveys absolute coordination (Schmidt, Nie, Franco, \& Richardson, 2014). Relative phase indicates whether the oscillations are in-phase $\left(0^{\circ}\right)$, antiphase $\left(180^{\circ}\right)$, or exhibiting a lag (between $0^{\circ}$ and $180^{\circ}$ ).

Surrogate and Virtual Pairs Analyses Surrogate analysis was conducted by computing a shuffled transformation of each observed movement time series and repeated the crosswavelet analyses for each dyad. This effectively destroys the temporal pattern in the data while preserving the distributional properties (Louwerse et al., 2012). Any measures of coordination applied to these are widely interpreted as the degree of coordination expected due to chance (Ramseyer \& Tschacher, 2010).

For the virtual pairs analysis, 42 randomized combinations of individuals who did not interact together were created. Because these individuals did not interact with each other, but were performing the same task, coordination measures calculated from virtual pairs have been interpreted as the coordination that can be expected due to the task demands (Strang et al., 2014). Thus, cross-wavelet analyses were conducted on these virtual pairs. Where time series were of unequal length, the longer time series was truncated to the length of the shorter series. Separate paired-sample ttests were used to compare between observed coherence and surrogate coherence as well as between observed coherence and virtual pairs coherence for each time scale.

Examining Relationship Between Coordination and Performance Our approach to answering RQs 1 and 2 was exploratory based on the results from H1. Specifically, we first conducted a linear multiple regression model with the observed coherence values at scales that were significantly greater than chance as predictors of performance. Then, we took those values that were significant predictors of performance and included them in a second multiple regression model with the relative phase values for those respective scales.

## Results

In order to examine H 1 , that the observed coordination would be greater than chance at some scales, the coherence
values for the surrogate data were compared to the coherence of the observed data. Results (see Table 1) suggested that coherence was significantly greater than chance at the $.25 \mathrm{~s}, .5 \mathrm{~s}, 1 \mathrm{~s}$, and 2 s frequency scales.

Table 1: Paired sample t-tests comparing observed to surrogate coherence and to virtual pairs coherence.

| Freq. <br> Scale | Observed <br> Coherence | Surrogate <br> Coherence | Virtual Pairs <br> Coherence |
| :--- | :--- | :--- | :--- |
| .25 s | $0.84(.04)$ | $0.30(.02)^{* * *}$ | $0.72(.05)^{* * *}$ |
| .5 s | $0.93(.03)$ | $0.54(.09)^{* * *}$ | $0.79(.17)^{* * *}$ |
| 1 s | $0.42(.06)$ | $0.29(.02)^{* * *}$ | $0.30(.01)^{* * *}$ |
| 2 s | $0.37(.05)$ | $0.33(.03)^{* * *}$ | $0.30(.03)^{* * *}$ |
| 4 s | $0.28(.02)$ | $0.29(.01)$ | $0.26(.02)^{* * *}$ |
| 8 s | $0.28(.03)$ | $0.29(.04)$ | $0.26(.02)^{* *}$ |
| 16 s | $0.30(.04)$ | $0.31(.04)$ | $0.29(.04)$ |
| 32 s | $0.32(.07)$ | $0.31(.07)$ | $0.29(.06)^{*}$ |
| 1 m | $0.33(.08)$ | $0.34(.08)$ | $0.31(.07)$ |
| 2 m | $0.38(.13)$ | $0.36(.11)$ | $0.33(.11)$ |
| 4.5 m | $0.48(.18)$ | $0.43(.21)$ | $0.36(.17)^{*}$ |
| Note. Values are mean and standard deviation. ${ }^{*} p<.05 ;$ |  |  |  |
| $* * p<.01 ;{ }^{* * *} ; p<.001$ |  |  |  |
| Wavelet Coherence: T9 High Performing |  |  |  |



Wavelet Coherence: T19 Low Peforming


Figure 1: Cross-wavelet coherence plots for a high performing team (top) and low performing team (bottom).
Likewise, in order to examine H2, that the observed interpersonal movement coordination would be greater than due to task demands and environment, the coherence values for the virtual pairs data were compared to the coherence of
the observed data. Results (see Table 1) suggested that coherence was significantly greater than could be expected due to task demands and environment alone for all frequency scales except $16 \mathrm{~s}, 1 \mathrm{~m}$, and 2 m .

To better understand the relationship between coherence and performance, we present two examples of cross-wavelet coherence plots in Figure 1. The top example is derived from the top performing team and the bottom example is derived from the lowest performing team. The $y$-axis corresponds to the frequency scale (which when divided by 8 can be related to time in seconds). The x-axis corresponds to the time on task with each point corresponding to $1 / 8$ of a second (or one video frame). The colors correspond to the amount of coherence with warmer colors indicating high coherence. Arrows indicate phase relationships with right arrows conveying in-phase and left arrows conveying antiphase. Arrows shifted up or down convey a lag in the oscillations between participants.

Next, we turn to RQs 1 and 2. In our first model, we included the four scales that were significantly more coordinated than expected due to chance alone as predictors of CPS performance (. $25 \mathrm{~s}, .5 \mathrm{~s}, 1 \mathrm{~s}$, and 2 s ). Overall, this model accounted for a significant $30.2 \%\left(R_{\text {adj }}^{2}=.226\right.$; $F(4,37)=4.00, p=.009)$ of the variability in CPS performance with coherence at the .25s ( $\beta=-.584, p=.017$ ) and 1s scales ( $\beta=.789, p=.003$ ) as significant predictors of performance. The .5 s and 2 s scales were not significant ( $p$ s $>.05$ ). Regarding RQ1, these results suggest that whereas stronger movement coordination at the 1 s scale is a strong predictor of better CPS performance, stronger coordination at the .25 s scale is associated with poorer performance.

Next, we sought to better understand the form of coordination at these scales. Thus, we conducted a second model that included coherence as well as relative phase at .25 s and 1 s scales. This model accounted for a significant $34.6 \% ~\left(R_{\text {adj }}^{2}=.276 ; F(4,37)=4.90, p=.003\right)$ of the variability in CPS performance with coherence ( $\beta=.614$, $p$ $=.007$ ) and relative phase $\left(M=3^{\circ}, S D=2^{\circ} ; \beta=-.294, p=\right.$ .038) at the 1 s scale as significant predictors of performance. Now, however, coherence at .25 s was not significant ( $\beta=-.412, p=.06$ ) nor was relative phase at .25 s ( $p=.24$ ). Thus, these results suggest that movement coordination at the 1 s scale is a primary predictor of performance and further, that relative phase values at the 1 s scale closer to $0^{\circ}$ (more in-phase) are associated with better CPS performance.

## Discussion

In this work, we investigated the multiscale, movement coordination dynamics that emerge in computer-based CPS. We found that movements in $.25 \mathrm{~s}-2 \mathrm{~s}$ scales were significantly more coordinated than chance and that all but the $16 \mathrm{~s}, 1 \mathrm{~m}$, and 2 m scales were more coordinated than expected due to task demands. We also observed that where coordination was greater than chance, both .25 s and 1 s were associated with CPS performance. However, when also accounting for relative phase, it appeared that higher in-
phase coordination at the 1s scale was the best predictor of CPS performance. Thus, some significant variability in CPS performance, in this context, appears to be explained by specific, low-scale patterns of coherence.

Given the low specificity of the movement data extracted from video, the question remains as to what is coordinated at these low scales and why they matter. In general, interactional phenomena that play out on (and below) a . 25 s timescale differ qualitatively from phenomena at a .5s timescale and beyond. For example, how interlocutors orient to each other's behavior as meaningful for the interaction depends on timing. Short pauses in interaction (typically $<.25$ s) are treated as idiosyncratic variation in speech; pauses around .5 s mark a transition space where the next speaker can take the word; and longer pauses (> 1s) "are often treated as flagging something unusual or troublesome about the interaction" (Mushin \& Gardner, 2009, p. 2035). Although, in addition to capturing these aspects of dialog, the observed coordination also captures mouse and keyboard movements, which likely unfold at these low scales as well. In general, many of the modalities captured by our movement measure can be argued to be task relevant as they capture dialogical events and computer input required for collaboration, but future work could consider more specific modalities such as how mouse movements are coordinated.

As far as future work is concerned, it is important to note some observable differences in coherence between the high and low performing teams in Fig. 1. There appear to be differences at higher scales, although average coherence was not generally above chance at these scales. However, we can speculate that participants’ performance may reflect their ability to create functional coherence across scales (e.g., between bodily ability and task demands), which could be assessed with fractal analyses (Davis et al., 2016). Further, it may be that successful CPS performance relies on higher-order transitions (Wiltshire et al., 2017) in coordination at one or more slow scales, as could be tenuously suggested by the pattern of high-low-high coherence near the 2 m scale across the duration of the task. Thus, future work should also consider extracting not only specific scales, but also time ranges that could be theoretically important to CPS.

More generally, research of this nature is important because it advances an efficient means of unobtrusively examining coordination processes during collaboration with a goal of working toward systems that can elicit forms of coordination that enable effective collaboration (Fiore \& Wiltshire, 2016; Kim, Chang, Holland, \& Pentland, 2008; Wiltshire \& Fiore, 2014). However, more work is necessary to understand if movement coordination is related to CPS performance in larger teams (de Montjoye, Stopczynski, Shmueli, Pentland, \& Lehmann, 2014), with different roles and disciplinary expertise (Bergmann, Dale, Sattari, Heit, \& Bhat, 2016), and when the teams are not co-located. Of course, such pursuits may require considering alternative modalities in which multiscale coordination might also
occur. We expect that such endeavors are essential to advancing our knowledge of the way that coordination during human interaction relates to collaborative cognition.

## Acknowledgments

This research was supported by Grant CF16-0455 from the Carlsberg Foundation awarded to Travis J. Wiltshire. We greatly appreciate the efforts of Stephen M. Fiore who was instrumental in the design and collection of data used for the present paper.

## References

Abney, D. H., Paxton, A., Dale, R., \& Kello, C. T. (2015). Movement dynamics reflect a functional role for weak coupling and role structure in dyadic problem solving. Cognitive processing, 16(4), 325-332.
Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. The journal of the learning sciences, 9(4), 403-436.
Berg, C. A., Johnson, M. M., Meegan, S. P., \& Strough, J. (2003). Collaborative problem-solving interactions in young and old married couples. Discourse Processes, 35(1), 33-58.
Bergmann, T., Dale, R., Sattari, N., Heit, E., \& Bhat, H. S. (2016). The interdisciplinarity of collaborations in Cognitive Science. Cognitive Science.
Butler, E. A. (2011). Temporal interpersonal emotion systems: The "TIES" that form relationships. Personality and Social Psychology Review, 15(4), 367-393.
Butner, J. E., Berg, C. A., Baucom, B. R., \& Wiebe, D. J. (2014). Modeling coordination in multiple simultaneous latent change scores. Multivariate behavioral research, 49(6), 554-570.
Dale, R., Fusaroli, R., Duran, N., \& Richardson, D. C. (2013). The self-organization of human interaction. Psychology of learning and motivation, 59, 43-95.
Davis, T. J., Brooks, T. R., \& Dixon, J. A. (2016). Multi-scale interactions in interpersonal coordination. Journal of Sport and Health Science, 5(1), 25-34.
de Montjoye, Y.-A., Stopczynski, A., Shmueli, E., Pentland, A., \& Lehmann, S. (2014). The strength of the strongest ties in collaborative problem solving. Scientific reports, 4.
Eiler, B. A., Kallen, R. W., Harrison, S. J., \& Richardson, M. J. (2013). Origins of order in joint activity and social behavior. Ecological Psychology, 25(3), 316-326.
Fiore, S. M., \& Wiltshire, T. J. (2016). Technology as Teammate: Examining the Role of External Cognition in Support of Team Cognitive Processes. Frontiers in psychology, 7.
Fiore, S. M., Wiltshire, T. J., Oglesby, J. M., O’Keefe, W. S., \& Salas, E. (2014). Complex Collaborative Problem-Solving Processes in Mission Control. Aviation, space, and environmental medicine, 85(4), 456-461.
Fujiwara, K., \& Daibo, I. (2016). Evaluating Interpersonal Synchrony: Wavelet Transform Toward an Unstructured Conversation. Frontiers in psychology, 7(7), 1-9.
Fusaroli, R., Rączaszek-Leonardi, J., \& Tylén, K. (2014). Dialog as interpersonal synergy. New Ideas in Psychology, 32, 147157.

Fusaroli, R., \& Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. Cognitive Science, 40(1), 145171.

Gallotti, M., Fairhurst, M., \& Frith, C. (2017). Alignment in social interactions. Consciousness and Cognition, 48, 253-261.
Gouhier, T. C., Grinsted, A., \& Simko, V. (2016). biwavelet: Conduct univariate and bivariate wavelet analyses (Version 0.20.10). Retrieved from http://github.com/tgouhier/biwavelet

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., \& Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends in cognitive sciences, 16(2), 114-121.
Hove, M. J., \& Risen, J. L. (2009). It's all in the timing: Interpersonal synchrony increases affiliation. Social Cognition, 27(6), 949-960.
Issartel, J., Bardainne, T., Gaillot, P., \& Marin, L. (2015). The relevance of the cross-wavelet transform in the analysis of human interaction-a tutorial. Frontiers in psychology, 5(1566), 1-18.
Issartel, J., Marin, L., Gaillot, P., Bardainne, T., \& Cadopi, M. (2006). A practical guide to time-frequency analysis in the study of human motor behavior: the contribution of wavelet transform. Journal of motor behavior, 38(2), 139-159.
Jordan, M. E., \& McDaniel Jr, R. R. (2014). Managing uncertainty during collaborative problem solving in elementary school teams: The role of peer influence in robotics engineering activity. Journal of the Learning Sciences, 23(4), 490-536.
Kim, T., Chang, A., Holland, L., \& Pentland, A. S. (2008). Meeting mediator: enhancing group collaborationusing sociometric feedback. Paper presented at the Proceedings of the 2008 ACM conference on Computer supported cooperative work.
Louwerse, M. M., Dale, R., Bard, E. G., \& Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. Cognitive Science, 36(8), 1404-1426.
Mills, G. J. (2014). Dialogue in joint activity: complementarity, convergence and conventionalization. New Ideas in Psychology, 32, 158-173.
Mitra, S., Amazeen, P. G., \& Turvey, M. T. (1998). Intermediate motor learning as decreasing active (dynamical) degrees of freedom. Human Movement Science, 17(1), 17-65.
Mushin, I., \& Gardner, R. (2009). Silence is talk: Conversational silence in Australian Aboriginal talk-in-interaction. Journal of Pragmatics, 41(10), 2033-2052. doi:http://dx.doi.org/10.1016/j.pragma.2008.11.004
NASA. (2011). Moonbase Alpha. Retrieved from http://www.nasa.gov/offices/education/programs/national/ltp/ games/moonbasealpha/index.html
OECD. (2015). PISA 2015 collaborative problem solving framework. Retrieved from https://www.oecd.org/pisa/pisaproducts/Draft\ PISA\ 2 015\%20Collaborative\%20Problem\%20Solving\%20Framewor k\%20.pdf.
Palumbo, R. V., Marraccini, M. E., Weyandt, L. L., Wilder-Smith, O., McGee, H. A., Liu, S., \& Goodwin, M. S. (2016). Interpersonal Autonomic Physiology A Systematic Review of the Literature. Personality and Social Psychology Review, 1088868316628405.

Paxton, A., \& Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. Behavior Research Methods, 45(2), 329-343.
R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria.: R Foundation for Statistical Computing. Retrieved from https://www.Rproject.org/

Ramseyer, F., \& Tschacher, W. (2010). Nonverbal synchrony or random coincidence? How to tell the difference Development of multimodal interfaces: active listening and synchrony (pp. 182-196): Springer.
Ramseyer, F., \& Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. Journal of consulting and clinical psychology, 79(3), 284.
Roschelle, J., \& Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. Paper presented at the Computer supported collaborative learning.
Schmidt, R., Morr, S., Fitzpatrick, P., \& Richardson, M. J. (2012). Measuring the dynamics of interactional synchrony. Journal of Nonverbal Behavior, 36(4), 263-279.
Schmidt, R., Nie, L., Franco, A., \& Richardson, M. J. (2014). Bodily synchronization underlying joke telling. Frontiers in human neuroscience, 8, 633.
Steffensen, S., \& Pedersen, S. B. (2014). Temporal dynamics in human interaction. Cybernetics \& Human Knowing, 21(1-2), 80-97.
Strang, A. J., Funke, G. J., Russell, S. M., Dukes, A. W., \& Middendorf, M. S. (2014). Physio-behavioral coupling in a cooperative team task: Contributors and relations. Journal of Experimental Psychology: Human Perception and Performance, 40(1), 145.
Timmons, A. C., Margolin, G., \& Saxbe, D. E. (2015). Physiological linkage in couples and its implications for individual and interpersonal functioning: A literature review. J. Fam. Psychol, 29(720), 10.1037.

Tryon, G. S., \& Winograd, G. (2011). Goal Consensus and Collaboration. Psychotherapy, 48(1), 50-57. doi:10.1037/a0022061
Valdesolo, P., Ouyang, J., \& DeSteno, D. (2010). The rhythm of joint action: Synchrony promotes cooperative ability. Journal of Experimental Social Psychology, 46(4), 693-695.
Walton, A. E., Richardson, M. J., Langland-Hassan, P., \& Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. Frontiers in psychology, 6, 313.
Washburn, A., DeMarco, M., de Vries, S., Ariyabuddhiphongs, K., Schmidt, R., Richardson, M. J., \& Riley, M. A. (2014). Dancers entrain more effectively than non-dancers to another actor's movements. Frontiers in human neuroscience, 8, 800.
Wiltermuth, S. S., \& Heath, C. (2009). Synchrony and cooperation. Psychological science, 20(1), 1-5.
Wiltshire, T. J. (2015). Team Interaction Dynamics during Collaborative Problem Solving. University of Central Florida Orlando, Florida.
Wiltshire, T. J., Butner, J., \& Fiore, S. M. (2017). Problem solving phase transitions during team collaboration. Cognitive Science.
Wiltshire, T. J., \& Fiore, S. M. (2014). Social Cognitive and Affective Neuroscience in Human-Machine Systems: A Roadmap for Improving Training, Human-Robot Interaction, and Team Performance. IEEE Transactions on HumanMachine Systems, 44(6), 779-787.

# A Computational Model for Constructing Preferences for Multiple Choice Options 

Lena M. Wollschlaeger (l.wollschlaeger@jacobs-university.de)<br>Jacobs University Bremen, Psychology \& Methods<br>Campus Ring 1, 28759 Bremen, Germany

Adele Diederich (a.diederich@jacobs-university.de)<br>Jacobs University Bremen, Life Sciences \& Chemistry<br>Campus Ring 1, 28759 Bremen, Germany


#### Abstract

When choosing between multiple alternatives, people usually do not have ready-made preferences in their mind but rather construct them on the go. The 2 N -ary Choice Tree Model (Wollschlaeger \& Diederich, 2012) proposes a preference construction process for N choice options from description, which is based on attribute weights, differences between attribute values, and noise. It is able to produce similarity, attraction, and compromise effects, which have become a benchmark for multi-alternative choice models, but also several other context and reference point effects. Here, we present a new and mathematically tractable version of the model - the Simple Choice Tree Model - which also explains the above mentioned effects and additionally accounts for the positive correlation between the attraction and compromise effect, and the negative correlation between these two and the similarity effect as observed by Berkowitsch, Scheibehenne, and Rieskamp (2014).


Keywords: computational model; multi-alternative choice; choice from description; preference construction; context effects

## Introduction

The decision making process involves various steps such as setting and prioritizing objectives, identifying choice alternatives, searching for information, developing preferences, and eventually taking a course of action. Here, we focus on developing preferences in multi-alternative choice situations and use in the following decision making from description as basic paradigm. Given a set of at least three choice alternatives that are described by at least two attributes, which they have in common, how do people choose one of these options? Simon (1955) argues that preferences in this kind of situation are dynamically constructed over time due to limited processing capacities. The decision maker experiences preference uncertainty (cf. Simonson, 1989) and tries to overcome it by gradually integrating the given information (see Payne, Bettman, \& Johnson, 1992, for a review on constructive processing in decision making). The resulting preferences are stochastic and highly dependent on the context, i.e., on the alternatives in the choice set and on any external reference points. Naturally, a model describing multi-alternative decision making from description should be a context-sensitive cognitive process model. The recently proposed 2 N -ary Choice Tree Model for preference construction for N choice options (2NCT; Wollschlaeger \& Diederich, 2012) assumes that the decision maker compares attribute values within attributes and between alternatives in a pairwise manner. Attributes are selected for examination based on attribute weights that reflect salience. Within attributes,
pairs of attribute values are selected for comparison based on so-called comparison values. In the 2NCT Model, the comparison values have a "global" component that remains constant over time during preference construction, a "local" component that depends on the outcomes of previous comparisons (reflecting leakage and inhibition, cf. Roe, Busemeyer, \& Townsend, 2001; Usher \& McClelland, 2004), and a random component. Advantageous and disadvantageous comparison outcomes for each alternative are counted separately and the difference of these counters is compared to two thresholds: a positive choice criterion and a negative elimination criterion. Implementation of an asymmetric value function (emphasizing disadvantageous comparison outcomes, cf. Usher \& McClelland, 2004) into the 2NCT Model is possible. Here, we present a revised and simpler version of the 2 N -ary Choice Tree Model, the Simple Choice Tree (SCT) Model. Therein, the local component is omitted from the definition of comparison values, making the model mathematically tractable while maintaining its ability to account for similarity, attraction and compromise effects. Furthermore, a new parameter, the focus weight $\lambda$, is introduced. It replaces the asymmetric value function and allows the SCT Model to account for correlations between the effects.

## Benchmark: Context Effects

Three context effects, demonstrating the influence of choice set composition on preferences, have played a prominent role in the multi-alternative preference construction modeling literature: The similarity effect, the compromise effect, and the attraction effect. All three effects occur when adding a third alternative to a set of two equally attractive yet clearly distinguishable options described by two attributes. Let $A_{1}$ and $A_{2}$ be two choice alternatives with two common attributes, $D_{1}$ and $D_{2}$, describing them. We assume that $D_{1}$ is the unique strongest attribute for $A_{1}$ and $D_{2}$ is the unique strongest attribute for $A_{2}$, that is, $A_{1}$ scores high on $D_{1}$ but low on $D_{2}$ and vice versa for $A_{2}$. One can think of the alternatives as placed in a two-dimensional space with dimensions $D_{1}$ and $D_{2}$. We further assume that the probability for choosing alternative $A_{1}$ from the binary choice set is equal to the probability for choosing alternative $A_{2}, P\left(A_{1} \mid A_{1}, A_{2}\right)=P\left(A_{2} \mid A_{1}, A_{2}\right)$.

Similarity Effect The similarity effect was named and first studied systematically by Tversky (1972). He observed the effect when comparing the binary choice set $\left\{A_{1}, A_{2}\right\}$ to the ternary choice set $\left\{A_{1}, A_{2}, A_{3}\right\}$ where $A_{3}$ is similar
to one of the original alternatives, say $A_{1}$, in scoring high on attribute $D_{1}$ and low on attribute $D_{2}$ while overall being similarly attractive (i.e. $P\left(A_{1} \mid A_{1}, A_{3}\right)=P\left(A_{3} \mid A_{1}, A_{3}\right)$ ). The probability of choosing $A_{1}$ over $A_{2}$ decreases when the decision maker chooses from the ternary choice set as compared to the binary set: $P\left(A_{1} \mid A_{1}, A_{2}\right) / P\left(A_{2} \mid A_{1}, A_{2}\right)>$ $P\left(A_{1} \mid A_{1}, A_{2}, A_{3}\right) / P\left(A_{2} \mid A_{1}, A_{2}, A_{3}\right)$.

Attraction Effect The attraction effect (or decoy effect or asymmetric dominance effect) was introduced by Huber, Payne, and Puto (1982) as consistent violation of the regularity principle. This principle, as presumed for example by the theory of Elimination by Aspects (Tversky, 1972), states that additional alternatives cannot increase the choice probabilities of the original options. However, Huber et al. (1982) claim that the relative probability for choosing alternative, say, $A_{1}$ can be increased by adding a third alternative $A_{3}$ to the choice set that is similar to but dominated by $A_{1}$ (and symmetrically for alternative $A_{2}$ ). $A_{3}$ then serves as a decoy for alternative $A_{1}$, drawing attention to it and therewith improving its evaluation and increasing its choice probability.

Compromise Effect Originally intended to explain the attraction effect, the theory of Reason-based Choice (Simonson, 1989) predicts an additional context effect, the compromise effect. It occurs when a third alternative $A_{3}$, equally attractive as the original alternatives $A_{1}$ and $A_{2}$, but more extreme with respect to the attribute values, is added to the choice set. If $A_{3}$ is more extreme than alternative $A_{1}$, that is, if it scores higher than $A_{1}$ on attribute $D_{1}$ but lower on attribute $D_{2}$, then it increases the choice share of $A_{1}$ as compared to the binary situation (and vice versa for alternative $\left.A_{2}\right)$ : $\quad P\left(A_{1} \mid A_{1}, A_{2}, A_{3}\right) / P\left(A_{2} \mid A_{1}, A_{2}, A_{3}\right)>$ $P\left(A_{1} \mid A_{1}, A_{2}\right) / P\left(A_{2} \mid A_{1}, A_{2}\right)$. However, note that the more similar the additional extreme alternative $A_{3}$ is to its adjacent alternative $A_{1}$, the more shares it takes away from $A_{1}$ via the similarity effect.

Interrelations of the Effects Recently, several studies have explored similarity, attraction and compromise effects and their interrelations in different choice scenarios. In a withinsubject consumer choice design, Berkowitsch et al. (2014) find that the similarity effect is negatively correlated with both the attraction and the compromise effect while the latter two are positively correlated. In a similar vein, Liew, Howe, and Little (2016) criticize that most of the results regarding context effects are based on averages over participants, not taking into account individual differences. Before analyzing the data from their inference and consumer choice experiments, they cluster it according to the observed choice patterns. The differences between clusters are remarkable, some even show negative (reverse) context effects while positive effects are observed in the averaged data. Before explaining how the Simple Choice Tree (SCT) Model accounts for the similarity, attraction and compromise effects and their interrelations, we introduce the basic mechanisms of the model.

## The Simple Choice Tree Model

Let $n_{a}$ be the number of alternatives under consideration, $\left\{A_{1}, A_{2}, \ldots, A_{n_{a}}\right\}$, and $n_{d}$ the number of attributes, $\left\{D_{1}, \ldots, D_{n_{d}}\right\}$, that characterize them. The decision maker is provided with one attribute value per alternative per attribute, that is, $n_{a} \cdot n_{d}$ attribute values in total. Let $m_{i j}$ be the attribute value for alternative $A_{i}$ with respect to attribute $D_{j}$. Attribute values within attributes and between alternatives are repeatedly compared and the resulting evidence is accumulated in two counters $S_{i}^{+}$and $S_{i}^{-}$for each alternative $A_{i}, i \in\left\{1, \ldots, n_{a}\right\}$. The positive counter $S_{i}^{+}$accumulates evidence for choosing alternative $A_{i}$ and the negative counter $S_{i}^{-}$accumulates evidence for rejecting it. Here, the initial counter states are set to zero, $S_{i}^{+}(0)=0=S_{i}^{-}(0)$. Definition of non-zero initial counter states accounting for prior knowledge about the choice alternatives is possible. However, these additional free parameters make the model less parsimonious and complicate parameter estimation. The counter states at time $t, S_{i}^{+}(t)$ and $S_{i}^{-}(t)$, are the initial counter states increased by the respective evidence accumulated until $t$. Their difference defines the momentary preference state for alternative $A_{i}$ at time $t: \operatorname{Pref}\left(A_{i}, t\right)=S_{i}^{+}(t)-S_{i}^{-}(t)$. We will now answer the following questions: (1) How is attention allocated between choice alternatives and attribute values? (2) How are alternatives evaluated and how is evidence accumulated? (3) When does evidence accumulation stop and which alternative is chosen?

## Attention Allocation

At the beginning of the process, when information about the alternatives and attributes is made available to the decision maker, each attribute $D_{j}, j \in\left\{1, \ldots, n_{d}\right\}$, is assigned a weight $\omega_{j}, 0 \leq \omega_{j} \leq 1$, reflecting its salience. The attribute weights determine how much attention the decision maker gives to the respective attributes during the preference construction process. Attributes with higher weights get more attention than attributes with lower weights. To allow for at least some of the attention to be allocated randomly between attributes, we define a random component (see below) for which an additional weight $\omega_{0}, 0 \leq \omega_{0} \leq 1$ is designated. Assuming that the weights sum up to one, $\sum_{j=0}^{n_{d}} \omega_{j}=1$, they can be interpreted as attention probabilities for the attributes: At each points of the preference construction process, the decision maker concentrates on attribute $D_{j}, j \in\left\{1, \ldots, n_{d}\right\}$ with probability $\omega_{j}$.

Having selected an attribute $D_{j}$, the decision maker concentrates on the specific attribute values of two alternatives and compares them. Pairs of attribute values are selected for comparison according to their importance for the decision. The more diagnostic the attribute values are, i.e., the more they discriminate between the alternatives, the more important they become for the decision. Pair selection probabilities within attribute $D_{j}$ are therefore defined to be proportional to the absolute differences $d_{i k j}=\left|m_{i j}-m_{k j}\right|, i \neq k \in\left\{1, \ldots, n_{a}\right\}$. In order to obtain probabilities, we normalize these differences to sum up to one: The probability for selecting the pair
$\left\{m_{i j}, m_{k j}\right\}$ for comparison is $p_{i k j}=d_{i k j} / \sum_{\{l, m\}} d_{l m j}, l \neq m \in$ $\left\{1, \ldots, n_{a}\right\}$. Note that the normalization of absolute differences balances out inequalities between attributes with - on average - bigger or smaller differences. Higher salience of an attribute $D_{j}, j \in\left\{1, \ldots, n_{d}\right\}$, with, for example, higher absolute differences, is thus not hard-wired into the model but is reflected in a higher attribute weight $\omega_{j}$ instead.

## Preference Sampling

The actual comparison of the two selected attribute values $m_{i j}$ and $m_{k j}$ is ordinal and directional: Let $m_{i j}>m_{k j}$, then the comparison can be either positively phrased, e.g. " $m_{i j}$ is greater than $m_{k j}$ ", or it can be negatively phrased, e.g. " $m_{k j}$ is smaller than $m_{i j}$ ". For the positive phrasing, $m_{i j}$ is called focus value and $m_{k j}$ is called reference value. The focus value determines the counter whose state is increased by +1 , here $S_{i}^{+}$, since the comparison is advantageous for the associated alternative $A_{i}$. For the negative phrasing, $m_{k j}$ is the focus value and $m_{i j}$ is the reference value, leading to an increase by +1 of counter $S_{k}^{-}$, since the comparison is disadvantageous for alternative $A_{k}$. Which phrasing the decision maker uses for the comparison and therewith which counter is updated might, for example, depend on the wording of the task or the decision maker's attitude (cf. Choplin \& Hummel, 2002). It is implemented into the model via the focus weight $\lambda, 0 \leq \lambda \leq 1$. If $\lambda=1-\lambda=0.5$, the decision maker uses the positive and negative phrasing both about equally often. If $\lambda>0.5$, the decision maker has a tendency towards the negative phrasing and towards updating negative counters. If $\lambda<0.5$, the decision maker has a tendency towards the positive phrasing and towards updating positive counters. The focus weight $\lambda$ replaces the asymmetric value function that was applied to the absolute differences between attribute values in the original 2NCT Model (Wollschlaeger \& Diederich, 2012). While the asymmetric value function hard-wired a tendency towards updating negative counters into the 2 NCT Model, weighting with $\lambda$ allows for flexible balancing of attention to positive versus negative aspects of the alternatives in the SCT Model. It is therefore especially useful in situations without a loss/gain-framing, e.g., in perceptual or preferential choice. Note that $\lambda$ is a global weight and independent from the attributes and attribute values. However, it allows us to define counter updating probabilities for the positive and negative counter of alternative $A_{i}, i \in\left\{1, \ldots, n_{a}\right\}$ with respect to attribute $D_{j}, j \in\left\{1, \ldots, n_{d}\right\}: p_{i j}^{+}=\sum_{k:\left(m_{i j}>m_{k j}\right)}(1-\lambda) \cdot p_{i k j}$ for updating $S_{i}^{+}$and $p_{i j}^{-}=\sum_{k:\left(m_{i j}<m_{k j}\right)} \lambda \cdot p_{i k j}$ for updating $S_{i}^{-}$.

Finally, the random component accounts for times where counter states are updated at random and without any connection to the actual attribute values (for instance due to inattention or misperception, cf. Busemeyer \& Townsend, 1993). Technically, it is treated as an additional (phantom) attribute $D_{0}$. The counter updating probabilities $p_{i 0}^{+}=p_{i 0}^{-}=1 /\left(2 \cdot n_{a}\right)$, $i \in\left\{1, \ldots, n_{a}\right\}$ with respect to $D_{0}$ depend on the number of available choice alternatives and therefore sum up to one: $\sum_{i=1}^{n_{a}}\left(p_{i 0}^{+}+p_{i 0}^{-}\right)=1$.

Combining attribute-wise counter updating probabilities $p_{i j}^{ \pm}$with attribute weights $\omega_{j}$, we can now define weighted counter updating probabilities for the positive and negative counter of alternative $A_{i}$ :

$$
\begin{equation*}
p_{i}^{+}=\sum_{j=0}^{n_{d}} p_{i j}^{+} \cdot w_{j} \quad \text { and } \quad p_{i}^{-}=\sum_{j=0}^{n_{d}} p_{i j}^{-} \cdot w_{j} \tag{1}
\end{equation*}
$$

## Choice Tree and Stopping Rules

Starting with the presentation of the choice alternatives and their attribute values, the preference construction process consists of a sequence of counter updates. In principle, every possible sequence of counter updates may occur and it is therefore of interest to have them conveniently summarized. For this purpose, we introduce the $\left(2 \cdot n_{a}\right)$-ary choice tree $T=(V, E, r)$ with vertices $V$, edges $E \subseteq V \times V$ and root $r \in V$, where all vertices are directed away from $r$ and each internal vertex $v \in V$ has $2 \cdot n_{a}$ children that are associated with the $2 \cdot n_{a}$ counters. Figure 1 shows an example with three choice alternatives and six counters. The preference construction process is represented by a random walk on $T$, beginning at the root and passing from there through an edge to another vertex, triggering the update (increase by +1 ) of the associated counter, moving on through another edge and so forth. The next edge to pass through is chosen according to the updating probability of the counter associated with its endpoint. Note that for each vertex the transition probabilities of all outgoing edges sum up to one. An example path of this random walk is pictured in bold in Figure 1.

The preference construction process stops when enough evidence has been accumulated to make the required choice. To this end, the preference states $\operatorname{Pref}\left(A_{i}, t\right)=S_{i}^{+}(t)-$ $S_{i}^{-}(t), i \in\left\{1, \ldots, n_{a}\right\}$ are constantly compared to two thresholds, a positive threshold $\theta^{+}$and a negative threshold $\theta^{-}=$ $-\theta^{+}$. If the preference state for alternative $A_{i}$ hits the positive threshold, the process stops and $A_{i}$ is chosen. If, on the other hand, the preference state for alternative $A_{k}$ hits the negative threshold, $A_{k}$ is eliminated from the choice set and the process continues with the remaining alternatives until one of them is chosen or until all but one of them have been eliminated. Consider a simple example with three choice alternatives $\left\{A_{1}, A_{2}, A_{3}\right\}$ and thresholds $\theta^{+}=2$ and $\theta^{-}=-2$. The sample path in Figure 1 with its associated sequence of counter updates $S_{2}^{+}, S_{1}^{-}, S_{1}^{-}, S_{2}^{+}$, leads to elimination of alternative $A_{1}$ after three steps and choice of alternative $A_{2}$ after four steps. Other possible sequences resulting in choice of alternative $A_{2}$ include $S_{3}^{+}, S_{1}^{-}, S_{2}^{+}, S_{2}^{+}$with direct choice of $A_{2}$ after four steps, and $S_{1}^{-}, S_{3}^{-}, S_{3}^{-}, S_{1}^{-}$with elimination of alternatives $A_{3}$ after three steps and $A_{1}$ after four steps and therewith choice of the only remaining alternative $A_{2}$.

## Choice Probabilities and Expected Response Times

The probability for walking along a specific path as, for example, shown in Figure 1, is the product of the transition probabilities along the respective edges. The choice probability for alternative $A_{i}, i \in\left\{1, \ldots, n_{a}\right\}$ is equal to the sum


Figure 1: A random walk on the choice tree for three alternatives. The associated sequence of counter updates is $S_{2}^{+}, S_{1}^{-}, S_{1}^{-}, S_{2}^{+}$and the probability for walking along this specific path is $p_{2}^{+} \cdot p_{1}^{-} \cdot p_{1}^{-} \cdot p_{2}^{+}$. Supposing that the rejection threshold $\theta^{-}$is equal to -2 and the choice threshold $\theta^{+}$is equal to 2 , this sequence implicates first rejection of alternative $A_{1}$ and then choice of alternative $A_{2}$. When $A_{1}$ is eliminated from the choice set, the vertices associated with its counters no longer appear in the choice tree, as can be seen in the bottom row of vertices here.
of the probabilities for walking along all the specific paths that lead to choice of alternative $A_{i}$. Since it is not feasible to calculate probabilities separately for each path and sum them up, we will analyze preference states, choice probabilities and response times instead by interpreting them as independent birth-death Markov chains with absorbing boundaries $\theta^{+}$and $\theta^{-}$. The state space of these birth-death chains $\operatorname{Pref}\left(A_{i}, t\right)=S_{i}^{+}(t)-S_{i}^{-}(t)=: S_{i}(t), i \in\left\{1, \ldots, n_{a}\right\}$ is $\mathcal{S}:=$ $\left\{\theta^{-}, \ldots,-1,0,1, \ldots, \theta^{+}\right\}$, with $|\mathcal{S}|=\theta^{+}-\theta^{-}+1$. The transition probabilities are

$$
\left.\begin{array}{ll}
p_{i}(x, x+1) & =p_{i}^{+}>0 \\
p_{i}(x, x-1) & =p_{i}^{-}>0 \\
p_{i}(x, x)=1-p_{i}^{+}-p_{i}^{-} & =p_{i}^{0}>0
\end{array}\right\} \text { for } x \in \mathcal{S}-\left\{-\theta^{-}, \theta^{+}\right\},
$$

where $p_{i}^{ \pm}$is defined in Eq. 1 above; $p_{i}(x, x+1)=p_{i}(x, x-$ $1)=0, p_{i}(x, x)=1$, for $x \in\left\{-\theta^{-}, \theta^{+}\right\}$; and zero otherwise. They form a $|\mathcal{S}| \times|\mathcal{S}|$ transition probability matrix $P_{i}^{\prime}=\left(p_{r s}^{\prime}\right)_{r, s=1, \ldots, \mid S}$, where $p_{r s}^{\prime}$ is the probability for the birthdeath chain to transition from state $x_{r}$ to state $x_{s}$ in one step. $P_{i}^{\prime}$ can be written in its canonical form $P_{i}$ by rearranging the rows and columns (changing the indices of the states such that the absorbing states $-\theta^{-}$and $\theta$ come first). $P_{i}$ can be de-
composed into a $2 \times 2$ identity matrix $I_{2}$, a $2 \times n_{t}$ matrix 0 of zeros with $n_{t}=|\mathcal{S}|-2$ (the number of transient states in $\mathcal{S}$ ), a $n_{t} \times 2$ matrix $R_{i}$, containing the probabilities for entering the absorbing states $\theta^{+}$and $\theta^{-}$, that is, for hitting the elimination or choice threshold, and a $n_{t} \times n_{t}$ matrix $Q_{i}$, containing the transition probabilities between transient states (cf. Diederich, 1997): $P_{i}=\left(\begin{array}{cc}I_{2} & 0 \\ R_{i} & Q_{i}\end{array}\right)$.

Given a row vector $Z_{i}$ of length $n_{t}$ which represents the initial preference state (e.g., ( $\left.\begin{array}{lllll}0 & 0 & 1 & 0 & 0\end{array}\right)$ ) or the initial distribution of preference over the transient states (e.g., $\left(\begin{array}{lllll}0.05 & 0.10 & 0.70 & 0.10 & 0.05\end{array}\right)$, cf. Diederich \& Busemeyer, 2003) for alternative $A_{i}$, the probability that the process is absorbed during the first step can be obtained by multiplying $Z_{i}$ and $R_{i}$, yielding a vector of length $2: Z_{i} \cdot R_{i}=$ $\left[P\left(S_{i}(1)=\theta^{+}\right), P\left(S_{i}(1)=-\theta^{-}\right)\right]$. In the case that the process was not absorbed during the first step, the distribution of preference over the transient states after the first step is given by $Z_{i} \cdot Q_{i}$, a vector of length $n_{t}$. Multiplying the result with the matrix $R_{i}$ yields the probabilities of absorption in the second step: $Z_{i} \cdot Q_{i} \cdot R_{i}=\left[P\left(S_{i}(2)=\theta^{+}\right), P\left(S_{i}(2)=-\theta^{-}\right)\right]$. The distribution of preference over the transient states is given by $\left(Z_{i} \cdot Q_{i}\right) \cdot Q_{i}=Z_{i} \cdot\left(Q_{i} \cdot Q_{i}\right)=Z_{i} \cdot\left(Q_{i}\right)^{2}$. The entries of the $n_{t} \times n_{t}$ matrix $\left(Q_{i}\right)^{2}$ are 2-step transition probabilities between the transient states, allowing for calculation of absorption in the third step: $Z_{i} \cdot\left(Q_{i}\right)^{2} \cdot R_{i}=\left[P\left(S_{i}(3)=\right.\right.$ $\left.\left.\theta^{+}\right), P\left(S_{i}(3)=-\theta^{-}\right)\right]$. Iterating these results indicates that all the relevant probabilities can be obtained from the vector $Z_{i}$, the matrix $R_{i}$ and powers of the matrix $Q_{i}$. Since $Q_{i}$ is a tridiagonal Toeplitz matrix (the entries on the main diagonal are all equal to $p_{i}^{0}$, the entries on the diagonal above the main diagonal are equal to $p_{i}^{+}$and the entries on the diagonal below the main diagonal are equal to $p_{i}^{-}$), its eigenvalues, eigenvectors and its powers are known and given in closed form (Salkuyeh, 2006), making it easy to compute all the relevant quantities.

We are interested in the conditional probabilities and expected hitting times for each alternative $A_{i}, i \in\left\{1, \ldots, n_{a}\right\}$, given that $A_{i}$ is the first alternative to be chosen/eliminated. Therefore, we have to determine the probability that alternative $A_{k}, k \in\left\{1, \ldots, n_{a}\right\}$ with $k \neq i$, has not been chosen/eliminated until time $t$. It is given by

$$
\begin{aligned}
& P\left(-\theta^{-}<S_{k}(T)<\theta\right)=1-\sum_{t=1}^{T} Z_{k} \cdot\left(Q_{k}\right)^{t-1} \cdot R_{k} \cdot\binom{1}{1} \\
& =1-Z_{k} \cdot\left(\sum_{t=1}^{T}\left(Q_{k}\right)^{t-1}\right) \cdot R_{k} \cdot\binom{1}{1} .
\end{aligned}
$$

The choice and elimination probability for alternative $A_{i}$ at time T is then equal to

$$
\begin{aligned}
& {\left[P\left(S_{i}(T)=-\theta^{-}\right), P\left(S_{i}(T)=\theta\right)\right]} \\
& =\left(Z_{i} \cdot \sum_{t=1}^{T}\left(Q_{i}\right)^{t-1} \cdot R_{i}\right) \cdot \prod_{k \neq i}\left(P\left(-\theta^{-}<S_{k}(T)<\theta\right)\right) .
\end{aligned}
$$

Overall, this yields probabilities

$$
\begin{aligned}
& {\left[P\left(\text { choose }_{i}\right), P\left(\text { eliminate }_{i}\right)\right]} \\
& =\sum_{T=1}^{\infty}\left(\left[P\left(S_{i}(T)=-\theta^{-}\right), P\left(S_{i}(T)=\theta\right)\right]\right)
\end{aligned}
$$

and expected response times

$$
\begin{aligned}
& {\left[E\left(T_{i} \mid \text { choose }_{i}\right), E\left(T_{i} \mid \text { eliminate }_{i}\right)\right]} \\
& =\sum_{T=1}^{\infty} T \cdot\left(\left[P\left(S_{i}(T)=-\theta^{-}\right), P\left(S_{i}(T)=\theta\right)\right]\right)
\end{aligned}
$$

Note that the infinite sums over T have only a finite number of nonzero addends, since $P\left(N_{i}<\infty\right)=1$ for all $i \in\left\{1, \ldots, n_{a}\right\}$, thus the choice/elimination probabilities and expected response times can be easily computed.

## Context Effects Explained

Three interacting mechanisms produce similarity, attraction, and compromise effects in the Simple Choice Tree Model: (1) selection of pairs of attribute values for comparison based on normalized differences, (2) the possibility to eliminate unwanted alternatives from the choice set, and (3) weighting of attributes based on salience. The first mechanism leads to a higher impact of dissimilar alternatives on the updating probabilities and thus faster evidence accumulation for alternatives with more distant competitors. In the similarity and attraction settings, this applies to the dissimilar alternative $A_{2}$, and in the compromise situation to the extreme alternatives $A_{2}$ and $A_{3}$. The second mechanism and the related focus weight $\lambda$ determine whether choices are more likely to be based on eliminations or to be made directly. The greater $\lambda$, the more likely are the choices based on eliminations. In the similarity situation, greater $\lambda$ leads to faster elimination of the dissimilar alternative $A_{2}$ and subsequent choice or elimination of either alternative $A_{1}$ or $A_{3}$, that is, a small or even negative similarity effect. On the other hand, smaller $\lambda$ leads to more direct choices of alternative $A_{2}$ and thus a higher similarity effect. Regarding the dissimilar alternative $A_{2}$, the same is true in the attraction situation. Greater $\lambda$ leads to faster elimination of $A_{2}$ while smaller $\lambda$ leads to more direct choices of alternative $A_{2}$. However, the attraction effect is higher for greater $\lambda$, since after elimination of alternative $A_{2}$, either the dominating option $A_{1}$ is chosen directly or the dominated option $A_{3}$ is eliminated first. In the compromise setting, greater $\lambda$ increases the probability for the extreme options to be eliminated from the choice set, leaving the decision maker with the compromise option. Smaller $\lambda$ on the other hand more likely leads to choice of an extreme option and thus a smaller or even negative compromise effect. Attribute weights further moderate the strengths of the context effects, but as long as they are more or less balanced, they play a minor role in the explanation of the similarity, attraction, and compromise effects. However, a high attribute weight is able to bias choice towards the alternative that scores highest on that attribute, covering any context effect.


Figure 2: Simulations of choice probabilities for changing focus weight $\lambda$ in the similarity, attraction, and compromise situation. There is a positive similarity effect for smaller $\lambda$ and a negative similarity effect for larger $\lambda$ (upper left) and vice versa for the attraction effect (upper right). The compromise effect (lower left and right) shows for larger $\lambda$ and is reversed for smaller $\lambda$.

We ran several simulations to illustrate these mechanisms. The available choice alternatives were $A_{1}=(70,30), A_{2}=$ $(30,70)$ and $A_{3}=(70,30)$ for the similarity effect, $A_{3}=$ $(65,25)$ for the attraction effect, $A_{3}=(90,10)$ for the asymmetric compromise effect, or $A_{3}=(50,50)$ for the symmetric compromise effect. The attribute weights were $\omega_{0}=0.1$ and $\omega_{1}=\omega_{2}=0.45$, and the focus weight $\lambda$ varied between 0 and 1 in steps of 0.1 . For each data point we ran 10000 simulations and the resulting choice probabilities are presented in figure 2. According to the simulations, the similarity effect is opposed to the attraction and the compromise effect. The similarity effect is strongest for low $\lambda$, whereas the attraction and the compromise effect are strongest for high $\lambda$. This prediction is consistent with the finding that the attraction and the compromise effect are positively correlated with each other and negatively correlated with the similarity effect (Berkowitsch et al., 2014). Note that $\lambda$ is assumed to be a global weight that does not change between trials but may vary between participants.

## Conclusion

We propose a revised and simpler version of the $2 N$-ary Choice Tree Model (Wollschlaeger \& Diederich, 2012), the Simple Choice Tree (SCT) Model. It predicts choice probabilities and response times in multi-alternative multi-attribute preferential choice from description and accounts for several
effects observed in these situations, including the similarity, attraction, and compromise effect. The SCT Model shares several aspects with existing models: Like Decision by Sampling (DbS; Stewart, Chater, \& Brown, 2006), it proposes binary ordinal comparisons and frequency accumulation as basic mechanisms. In DbS , however, pairs of attribute values are chosen at random and reference values may be sampled from long-term memory as well as from the given context. Only advantageous comparisons are counted and the model is not able to account for the above mentioned context effects, nor does it provide solutions for choice probabilities or choice response times. Multi-alternative Decision Field Theory (MDFT; Roe et al., 2001) and the Leaky Competing Accumulator (LCA) Model (Usher \& McClelland, 2001, 2004) provide such solutions only for fixed stopping times. Both models, like the SCT Model, are based on pairwise differences of attribute values. To account for the similarity, attraction, and compromise effect simultaneously, however, additional non-linear mechanisms (among others leakage and inhibition, cf. the original 2 NCT Model) are required, preventing the models from providing mathematically tractable solutions for optional stopping times. Elimination by Aspects (EBA; Tversky, 1972) proposes "a covert elimination process based on sequential selection of aspects" (p. 296). As an early example for a cognitive process model, it does not make any predictions about choice response times and accounts only for the similarity effect. The SCT model mimics EBA for high values of the focus weight $\lambda$, where mostly disadvantageous comparison outcomes are considered and decisions are based on the elimination of choice alternatives. The Multiattribute Linear Ballistic Accumulator Model (MLBA; Trueblood, Brown, \& Heathcote, 2014), basically a deterministic version of MDFT, provides analytic solutions for expected response times and choice probabilities like the SCT Model. However, it is unclear if and how the response times are related to the actual integration of information. Furthermore, the model has mostly been applied with fixed stopping times until now. Additional mechanisms allow the MLBA model to account for the compromise effect (a curved subjective value function) and the similarity effect (a higher weight on supportive information as compared to disconfirmatory evidence). The latter is comparable to low values of the focus weight $\lambda$ in the SCT Model. To summarize, the SCT Model combines aspects of competing models in a new way, yielding qualitatively new explanations for the context effects and additionally predicting correlation patterns amongst the effects. It provides mathematically tractable solutions for both choice probabilities and expected choice response times for optional stopping times, by that outperforming existing models.

## Acknowledgement

This research was supported by grant DFG DI506/16-1.

## References

Berkowitsch, N. A. J., Scheibehenne, B., \& Rieskamp, J. (2014). Rigorously testing multialternative decision field
theory against random utility models. Journal of Experimental Psychology: General, 143(3), 1331-1348.
Busemeyer, J. R., \& Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review, 100(3), 432-459.
Choplin, J. M., \& Hummel, J. E. (2002). Magnitude comparisons distort mental representations of magnitude. Journal of Experimental Psychology: General, 131(2), 270-286.
Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. Journal of Mathematical Psychology, 41(3), 260-274.
Diederich, A., \& Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. Journal of Mathematical Psychology, 47(3), 304-322.
Huber, J., Payne, J. W., \& Puto, C. P. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. Journal of Consumer Research, 9(1), 90-98.
Liew, S. X., Howe, P. D. L., \& Little, D. R. (2016). The appropriacy of averaging in the study of context effects. Psychonomic Bulletin \& Review, 23(5), 1639-1646.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1992). Behavioral Decision Research - a Constructive Processing Perspective. Annual Review of Psychology, 43, 87-131.
Roe, R. M., Busemeyer, J. R., \& Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. Psychological Review, 108(2), 370-392.
Salkuyeh, D. K. (2006). Positive integer powers of the tridiagonal Toeplitz matrices. International Mathematical Forum, l(22), 1061-1065.
Simon, H. A. (1955). A behavioral model of rational choice. The Quarterly Journal of Economics, 69(1), 99-118.
Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. Journal of Consumer Research, 16(2), 158-174.
Stewart, N., Chater, N., \& Brown, G. D. A. (2006). Decision by sampling. Cognitive Psychology, 53(1), 1-26.
Trueblood, J. S., Brown, S. D., \& Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. Psychological Review, 121(2), 179-205.
Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological Review, 79(4), 281-299.
Usher, M., \& McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review, 108(3), 550-592.
Usher, M., \& McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. Psychological Review, 111(3), 757-769.
Wollschlaeger, L. M., \& Diederich, A. (2012). The 2Nary choice tree model for N -alternative preferential choice. Frontiers in psychology, 3(189).

# Mapping the unknown: The spatially correlated multi-armed bandit 

Charley M. Wu ${ }^{1}$, Eric Schulz ${ }^{2}$, Maarten Speekenbrink ${ }^{2}$, Jonathan D. Nelson ${ }^{1}$ \& Björn Meder ${ }^{1}$<br>${ }^{1}$ Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany<br>${ }^{2}$ Department of Experimental Psychology, University College London, London, WC1H0AP


#### Abstract

We introduce the spatially correlated multi-armed bandit as a task coupling function learning with the explorationexploitation trade-off. Participants interacted with bi-variate reward functions on a two-dimensional grid, with the goal of either gaining the largest average score or finding the largest payoff. By providing an opportunity to learn the underlying reward function through spatial correlations, we model to what extent people form beliefs about unexplored payoffs and how that guides search behavior. Participants adapted to assigned payoff conditions, performed better in smooth than in rough environments, and-surprisingly-sometimes performed equally well in short as in long search horizons. Our modeling results indicate a preference for local search options, which when accounted for, still suggests participants were best-described as forming local inferences about unexplored regions, combined with a search strategy that directly traded off between exploiting high expected rewards and exploring to reduce uncertainty about the spatial structure of rewards.


Keywords: Exploration-exploitation; Multi-armed bandits; Active Learning; Gaussian Processes;

## Introduction

Modern humans descend from capable foragers and hunters, who have migrated and survived in almost every environment on Earth. Our ancestors were able to adaptively learn the distribution of resources in new environments and make good decisions about where to search, balancing the dual goals of exploring to acquire new information and exploiting existing knowledge for immediate gains. What strategies do humans use to search for resources in unknown environments?

We present a new framework for studying human search behavior using a spatially correlated multi-armed bandit task, where nearby arms (i.e., search options) have correlated rewards. Spatial correlations provide an opportunity to learn about the underlying reward function, extending the traditional reinforcement learning paradigm (Sutton \& Barto, 1998) to allow for generalization of learned rewards to unobserved actions using spatial context. We compare search behavior across different payoff conditions, search horizons, and types of environments, finding that participants adapt to their environment, tend to perform very local inferences about unexplored regions and choose arms based on a trade-off between expectations and their attached uncertainties.

## Spatially Correlated Multi-Armed Bandits

We adapt the multi-armed bandit (MAB) setting by adding spatial correlation to rewards and placing the arms in a twodimensional grid (Fig. 1). Each tile represents a playable arm of the bandit, which are initially blank and display the numerical reward value (along with a color aid) after an arm has been chosen. Traditionally, the goal in an MAB task is to
Smooth Environment

| 7 | 5 | 10 | 22 | 32 | 32 | 28 | 24 | 22 | 26 | 33 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 6 | 11 | 19 | 29 | 38 | 41 | 42 | 40 | 37 | 36 | 40 |
| 22 | 27 | 30 | 35 | 43 | 50 | 53 | 53 | 51 | 49 | 46 |
| 45 | 44 | 38 | 36 | 40 | 46 | 47 | 49 | 54 | 55 | 48 |
| 61 | 55 | 46 | 40 | 37 | 32 | 27 | 31 | 44 | 52 | 44 |
| 62 | 59 | 57 | 54 | 44 | 27 | 14 | 17 | 33 | 46 | 45 |
| 53 | 59 | 68 | 71 | 59 | 36 | 17 | 15 | 28 | 45 | 51 |
| 46 | 57 | 71 | 77 | 67 | 47 | 26 | 18 | 27 | 45 | 56 |
| 45 | 56 | 65 | 67 | 60 | 46 | 29 | 20 | 27 | 42 | 55 |
| 51 | 57 | 58 | 53 | 47 | 40 | 30 | 23 | 28 | 40 | 49 |
| 60 | 62 | 58 | 47 | 39 | 38 | 35 | 31 | 35 | 41 | 46 |

Figure 1: Examples of the underlying reward functions for the two classes of environments.
maximize cumulative payoffs by sequentially choosing one of the N -arms of the bandit that stochastically generate rewards (Steyvers, Lee, \& Wagenmakers, 2009), with learning happening independently for each arm (i.e., reinforcement learning). In our case, because proximate arms generate similar rewards, there is the opportunity to form inductive beliefs about unobserved rewards (i.e., function learning). This allows us to study how people generate beliefs about unobserved rewards and how this influences their search behavior.

The spatially correlated MAB is related to the optimal foraging context (Krebs, Kacelnik, \& Taylor, 1978), whereby a forager is not only guided by the search for resources, but also by the need to acquire information about the distribution of resources in the environment (Schulz, Huys, Bach, Speekenbrink, \& Krause, 2016). This creates a natural trade-off between exploration and exploitation (March, 1991), where an effective search policy needs to adequately balance exploring areas with higher uncertainty, while also exploiting existing information to obtain rewards. One key difference in our task is that the decision-maker must determine where to search, and not only whether to stay or to leave a patch.

## Modeling Adaptive Search Behavior

We consider various computational models for describing human behavior, which all make sequential predictions about where people are likely to search. We present both simple strategies without an explicit representation of the environment, along with more complex function generalization models representing the task as a combination of (i) a function learning model and (ii) a decision strategy. We use a form of Gaussian Process regression as a flexible and universal function learning model, which forms inferential beliefs about the underlying reward function, conditioned on previous observations of rewards. Decision strategies are used to transform beliefs into predictions about where to search next. The recovered parameter estimates of our models describe the extent to which people make spatial inferences and how they trade off between exploration and exploitation.

## Simple Strategies

Local search. While simple, a tendency to stay local to the previous search decision-regardless of outcome-has been observed in many different contexts, such as semantic foraging (Hills, Jones, \& Todd, 2012), causal learning (Bramley, Dayan, Griffiths, \& Lagnado, 2017), and eye movements (Hoppe \& Rothkopf, 2016). We use inverse Manhattan distance (IMD) to quantify locality:

$$
\begin{equation*}
\operatorname{IMD}\left(\mathbf{x}, \mathbf{x}^{\prime}\right)=\frac{1}{\left|x_{1}-x_{1}^{\prime}\right|+\left|x_{2}-x_{2}^{\prime}\right|} \tag{1}
\end{equation*}
$$

which compares the location of two arms $\mathbf{x}$ and $\mathbf{x}^{\prime}$, where $x_{1}$ and $x_{2}$ are the grid coordinates. For the special case where $\mathbf{x}=\mathbf{x}^{\prime}$, we set $\operatorname{IMD}\left(\mathbf{x}, \mathbf{x}^{\prime}\right)=1$. At each time $t$, we compute the IMD for each arm based on the choice at $\mathbf{x}_{t-1}$, and then use a softmax function (Eq. 11) to transform locality into choice probabilities, such that arms closer to the previous search decision have a higher probability of being chosen.

Win-stay lose-shift. We also consider a form of the winstay lose-shift (WSLS) heuristic (Herbert, 1952), where a win is defined as finding a payoff with a higher or equal value than the previous best. When the decision-maker "wins", we assume that any tile with a Manhattan distance $\leq 1$ is chosen (i.e., a repeat or any of the four cardinal neighbors) with equal probability. Losing is defined as the failure to improve, and results in choosing any unrevealed tile with equal probability.

## Function Generalization Models

We use a combination of (i) Gaussian Process ( $\mathcal{G} \mathcal{P}$ ) regression as a model of how people form beliefs about the underlying reward function conditioned on previous observations (Lucas, Griffiths, Williams, \& Kalish, 2015), and (ii) a decision strategy that transforms beliefs into predictions about where a participant will sample next. This approach has recently been applied to human behavior in contextual multi-armed bandits (Schulz, Konstantinidis, \& Speekenbrink, 2016) and is the only known computational algorithm to have any guarantees in a bandit setting (i.e., bounded regret; Srinivas, Krause, Kakade, \& Seeger, 2010).

Gaussian process learning. A $\mathcal{G} \mathcal{P}$ defines a distribution $P(f)$ over possible functions $f(\mathbf{x})$ that map inputs $\mathbf{x}$ to output $y$, in our case, grid location to reward. A $\mathcal{G} P$ is completely defined by a mean $\mu(\mathbf{x})$ and a kernel function, $k\left(\mathbf{x}, \mathbf{x}^{\prime}\right)$ :

$$
\begin{align*}
\mu(\mathbf{x}) & =\mathbb{E}[f(\mathbf{x})]  \tag{2}\\
k\left(\mathbf{x}, \mathbf{x}^{\prime}\right) & =\mathbb{E}\left[(f(\mathbf{x})-\mu(\mathbf{x}))\left(f\left(\mathbf{x}^{\prime}\right)-\mu\left(\mathbf{x}^{\prime}\right)\right)\right] \tag{3}
\end{align*}
$$

Here, we fix the prior mean to the median value of payoffs, $\mu(\mathbf{x})=50$ and use a radial basis function kernel (Eq. 7).

Suppose we have collected observations $\mathbf{y}_{T}=$ $\left[y_{1}, y_{2}, \ldots, y_{T}\right]^{\top}$ at inputs $\mathbf{X}_{T}=\left\{\mathbf{x}_{1}, \ldots, \mathbf{x}_{T}\right\}$, and assume

$$
\begin{equation*}
y_{t}=f\left(\mathbf{x}_{t}\right)+\varepsilon_{t} \quad \varepsilon_{t} \sim \mathcal{N}(0,1) \tag{4}
\end{equation*}
$$

Given a $\mathcal{G} \mathcal{P}$ prior on functions $f(\mathbf{x}) \sim \mathcal{G} \mathcal{P}\left(\mu(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}^{\prime}\right)\right)$, the posterior distribution over $f\left(\mathbf{x}_{T}\right)$ given inputs $\mathbf{X}_{T}$ is also
a $\mathcal{G} \mathcal{P}$ with the following mean and covariance:

$$
\begin{align*}
\mu_{T}(\mathbf{x}) & =\mathbf{k}_{T}(\mathbf{x})^{\top}\left(\mathbf{K}_{T}+\sigma^{2} \mathbf{I}\right) \mathbf{y}_{T}  \tag{5}\\
k_{T}\left(\mathbf{x}, \mathbf{x}^{\prime}\right) & =k\left(\mathbf{x}, \mathbf{x}^{\prime}\right)-\mathbf{k}_{T}(\mathbf{x})^{\top}\left(\mathbf{K}_{T}+\sigma^{2} \mathbf{I}\right)^{-1} \mathbf{k}_{T}\left(\mathbf{x}^{\prime}\right) \tag{6}
\end{align*}
$$

where $\mathbf{k}_{T}(\mathbf{x})=\left[k\left(\mathbf{x}_{1}, \mathbf{x}\right), \ldots, k\left(\mathbf{x}_{T}, \mathbf{x}\right)\right]^{\top}$ and $\mathbf{K}_{T}$ is the positive definite kernel matrix $\left[k\left(\mathbf{x}_{i}, \mathbf{x}_{j}\right)\right]_{i, j=1, \ldots, T}$. This posterior distribution is used to derive normally distributed predictions about the rewards for each arm of the bandit (Fig. 2).

The kernel function $k\left(\mathbf{x}, \mathbf{x}^{\prime}\right)$ encodes prior assumptions about the underlying function. We use the radial basis function (RBF) kernel

$$
\begin{equation*}
k_{\mathrm{RBF}}\left(\mathbf{x}, \mathbf{x}^{\prime}\right)=\exp \left(-\frac{\left\|\mathbf{x}-\mathbf{x}^{\prime}\right\|^{2}}{2 \lambda^{2}}\right) \tag{7}
\end{equation*}
$$

which is a universal function learner and assumes infinitely smooth functions (i.e., correlations between two points $\mathbf{x}$ and $\mathbf{x}^{\prime}$ slowly decay as an exponential function of their distance). The RBF kernel uses $\lambda$ (length-scale) as a free parameter, which determines how far correlations extend: larger values of $\lambda$ result in longer spatial correlations, whereas $\lambda \rightarrow 0^{+}$assumes complete independence of spatial information. We use recovered parameter estimates of $\lambda$ to learn about the extent to which humans make inferences about unobserved rewards.

Decision strategies. The $\mathcal{G} P$ learning model generates normally distributed predictions about the expectation $\mu(\mathbf{x})$ and the uncertainty $\sigma(\mathbf{x})$ for each arm, which are available to the decision strategies ${ }^{1}$ for evaluating the quality, $q(\mathbf{x})$, and ultimately making a prediction about where to sample next.

The Variance Greedy (VG) strategy values an arm using only the estimated uncertainty

$$
\begin{equation*}
q_{V G}(\mathbf{x})=\sigma(\mathbf{x}) \tag{8}
\end{equation*}
$$

and is an efficient step-wise (greedy) approximation of information gain (Srinivas et al., 2010), which seeks to learn the global reward function as rapidly as possible. VG achieves at least a constant fraction of the optimal information gain value (Krause \& Guestrin, 2005); however, it fails to adequately trade-off between exploration and exploitation, because effort is wasted exploring the function where $f(\mathbf{x})$ is small.

The Mean Greedy (MG) strategy is also step-wise greedy, valuing arms using only the estimated mean reward

$$
\begin{equation*}
q_{M G}(\mathbf{x})=\mu(\mathbf{x}) \tag{9}
\end{equation*}
$$

although this strategy carries no known guarantees and is prone to getting stuck in local optima.

Upper confidence bound sampling (UCB) combines the VG and MG strategies

$$
\begin{equation*}
q_{U C B}(\mathbf{x})=\mu(\mathbf{x})+\beta \sigma(\mathbf{x}) \tag{10}
\end{equation*}
$$

[^245]where the exploration factor $\beta$ determines how the reduction of uncertainty trades off against exploiting high expected rewards. This is sometimes referred to as optimistic "sampling with confidence" as it inflates expectations with respect to the upper confidence bounds (Srinivas et al., 2010), creating a natural balance between exploration and exploitation.


Figure 2: Modeling human performance. Column left represents the initial state of the task and column right is after 10 clicks. Top row: screenshots from the experiment. 2nd row: posterior predictions of expected reward $\mu(\mathbf{x})$, from a $\mathcal{G} \mathcal{P}$ with an RBF kernel (not shown: the estimated variance). 3rd row: the values of each tile $q(\mathbf{x})$ using the UCB acquisition function. Bottom row: the softmax prediction surface transforming the UCB values into choice probabilities.

## Choice Probabilities

For all models, we use a softmax function (Fig. 2 bottom row) to convert the value of an option $q(\mathbf{x})$ into a choice probability

$$
\begin{equation*}
P(\mathbf{x})=\frac{\exp (q(\mathbf{x}) / \tau)}{\sum_{j=1}^{N} \exp \left(q\left(\mathbf{x}_{j}\right) / \tau\right)} \tag{11}
\end{equation*}
$$

where $\tau$ is the temperature parameter. As $\tau \rightarrow 0$ the highestvalue arm is chosen with a probability of 1 (i.e., argmax), and when $\tau \rightarrow \infty$, all options are equally likely, with predictions converging to random choice. We use $\tau$ as a free parameter, where lower estimates can be interpreted as more precise predictions about choice behavior.

## Experiment

We present a bi-variate MAB problem with spatially correlated rewards. The problem space was represented by a twodimensional grid, measuring $11 \times 11$, resulting in 121 unique tiles in total. Participants could click to reveal unexplored tiles or re-click previously uncovered tiles to exploit known rewards (see Fig. 2 top row for screenshots).

## Methods

Participants. We recruited 80 participants from Amazon Mechanical Turk ( 25 Female; mean age $\pm$ SD $32 \pm 9$ ). Each participant was paid a participation fee of $\$ 0.50$ and a performance contingent bonus up to $\$ 1.50$. Subjects earned on average $\$ 1.64 \pm 0.20$ and spent $8 \pm 4$ minutes on the task.

Design. We used a $2 \times 2$ between subject design, where participants were randomly assigned to one of two different pay-off structures (Average Reward vs. Maximum Reward) and one of two different classes of environments (Smooth vs. Rough). Each grid represented a bi-variate function, with each observation including normally distributed noise, $\varepsilon \sim \mathcal{N}(0,1)$. The task was presented over 8 blocks on different grid worlds drawn from the same class of environments. In each block, participants had either a Short ( 20 clicks) or Long ( 40 clicks) search horizon to interact with the grid. The search horizon alternated between blocks (within subject), with initial horizon length counterbalanced between subjects. Per block, observations were scaled to a uniformly sampled maximum value in the range of 65 to 85 , so that the value of the global optima could not be easily guessed (e.g., a value of 100).

Materials and procedure. Before starting, participants were shown four fully revealed grids in order to familiarize themselves with the task. Example environments were drawn from the same class of environments assigned to the participant (Smooth or Rough) and underwent the same random scaling of observations. Additionally, three comprehension questions were used to ensure full understanding of the task.

At the beginning of each of the 8 blocks, one random tile was revealed and participants could use their mouse to click any of the 121 tiles in the grid until the search horizon was exhausted, including re-clicking previously revealed tiles. Clicking an unrevealed tile displayed the numerical value of the reward along with a corresponding color aid, where darker colors indicated higher point values (Fig. 1). Previously revealed tiles could also be re-clicked, although there were variations in the observed value due to noise. For repeat clicks, the most recent observation was displayed numerically, while hovering over the tile would display the entire history of observations. The color of the tile corresponded to the mean of all previous observations.

Payoff conditions. We compared performance under two different payoff conditions, requiring either a balance between exploration and exploitation (Average Reward) or a pure exploration context (Maximum Reward). Previous work
has shown that people can adapt (sometimes with great difficulty) to different payoff conditions in information acquisition tasks (Meder \& Nelson, 2012).

In each payoff condition, participants received a performance contingent bonus of up to $\$ 1.50$. Average Reward participants were told to "gain as many points as possible across all 8 grids" and were given a bonus based on the average value of all clicks as a fraction of the global optima, $\frac{1}{T} \sum\left(\frac{y_{t}}{y^{*}}\right)$, where $y^{*}$ is the global optimum. Maximum Reward participants were told to "learn where the largest reward is" and were giving a bonus using the ratio of the highest observed reward to the global optimum, $\left(\frac{\max y_{t}}{y^{*}}\right)^{4}$, taken to the power of 4 to exaggerate differences in the upper range of performance and for parity in expected earnings across payoff conditions. All 8 blocks were weighted equally, using noisy but unscaled observations to assign a bonus of up to $\$ 1.50$. Subjects were informed in dollars about the bonus earned at the end of each block.

Smoothness of the environment. We used two different classes of environments, corresponding to different levels of smoothness (Fig. 1). All environments were sampled from a $\mathcal{G} \mathcal{P}$ prior parameterized with a RBF kernel, where the lengthscale parameter $(\lambda)$ determines the rate at which the correlations of rewards decay over distance. We sampled 20 Smooth environments using $\lambda=2$ and 20 Rough environments using $\lambda=1$. Subjects performed the task on 8 grids randomly drawn (without replacement) from their assigned class of environments, while the four fully revealed environments used to familiarize subjects with the task were drawn (without replacement) from the remaining 12 environments.
Search horizons. The length of the search horizon influences the value of information learned about the environment, with respect to the assigned payoff condition. Longer horizons provide more opportunities for exploiting acquired information, thereby making early exploration more valuable. We chose two horizon lengths (Short $=20$ and Long $=40$ ) that were fewer than the total number of tiles on the grid (121), and varied within subject (alternating between blocks).

## Results

Figure 3 shows task performance. In all conditions, performance improved as a function of the trial number (i.e., with each additional click), as measured by both the overall correlation between average reward and trial number ( $r=.32$, $p=.04$ ) and between the maximum observed reward and trial number ( $r=.83, p<.001$ ). There were no learning effects across blocks (i.e., over successive grids), indicated by a lack of correlation between average reward and block number ( $r=.19, p=.65$ ), or between maximum reward and block number ( $r=-.37, p=.36$ ). Performance improved as more information was revealed (i.e., over trials), but not over additional blocks of identically parameterized environments.

Payoff conditions. Payoff conditions influenced search behavior, with participants in the Maximum Reward condition
displaying more variance in the locations sampled $(t) 78)=$ $-2.48, p=.02$ ). There were some differences in the number of unique tiles revealed (Fig. 3C) and the number of repeat clicks across the payoff conditions (Fig. 3D), although the effect size is largest for smooth environments given long search horizons. However, these behavioral differences did not manifest in terms of performance, with no systematic differences across payoff conditions in terms of the average reward obtained $t(78)=1.32, p=.2$ ) or in the maximum revealed reward $(t(78)=.001, p=.99)$.
Environment and horizon. Independent of the payoff condition, participants assigned to Smooth environments achieved higher average rewards $(t(78)=6.55, p<.001)$ and higher maximum rewards $(t(78)=5.45, p<.001)$, than those assigned to the Rough environments (Fig. 3E), suggesting that stronger correlations of payoffs make the task easier. Interestingly, longer horizons did not lead to better overall performance in the Average Reward condition $(t) 80)=.34$, $p=.73$ ), although participants given longer horizons found larger maximum rewards for all payoffs and environment conditions $(t(158)=7.62, p<.001)$. There may be a less-is-more-effect, with longer horizons leading to over-exploration, given the goal of maximizing average rewards.


Figure 3: Overview of task performance. (A) Average reward earned and (B) maximum reward revealed, where colors correspond to payoff condition and line-types to horizon length. Black lines show simulated performance of a random model averaged over 10,000 randomly sampled grids. (C) The average number of unique tiles clicked in each block and (D) the average number of repeat clicks in each block. (E) The distribution of rewards earned during each block, grouped first by environment type and then by horizon length.

## Model Comparison

We describe each model's ability to predict participant behavior using leave-one-block-out cross validation. For each participant, we analyzed the four short and the four long horizon blocks separately. Cross-validation was performed by holding out a single block as a test set, and fitting the model parameters using a maximum likelihood estimate (MLE) on the remaining three blocks. Iterating through each of the four hold-out blocks, for both short and long horizons, we calculated a model's out-of-sample log loss (i.e., test set prediction accuracy) and then summed up the results over all blocks. We use McFadden's $R^{2}$ values (McFadden, 1974) to compare the out-of-sample log loss for each model to that of a random model (Fig. 4), where $R^{2}=0$ indicates chance performance and $R^{2}=1$ is a perfect model.


Reward Condition $\square$ Average Reward $\square$ Maximum Reward
Figure 4: Model Comparison. The height of the bars show the group mean and error bars indicate standard error. McFadden's $R^{2}$ is a goodness of fit measure comparing each model $\mathcal{M}_{k}$ to a random model $\mathcal{M}_{\text {rand }}$. Using the out-of-sample log loss for each model, $R_{\mathrm{McF}}^{2}=1-\log \mathcal{L}\left(\mathscr{M}_{k}\right) / \log \mathcal{L}\left(\mathcal{M}_{\text {rand }}\right)$.

A large amount of the variance in participant behavior is explained by local search $\left(R^{2}=.28\right.$; all conditions); however, locality alone fails to achieve similar task performance as humans, with performance almost identical to random in terms of average reward and worse than random in maximum reward (Fig. 5). WSLS by comparison, was a poor approximation of search behavior ( $R^{2}=.05$ ), and was excluded from the model performance comparison.

Among the GP models, UCB performed best ( $R^{2}=.23$ ), with MG showing comparable results $\left(R^{2}=.17\right)$ and VG performing poorly ( $R^{2}=.01$ ). Interestingly, the performance of the GP-UCB model was remarkably similar to human subjects in terms of both average and maximum reward (Fig. 5). Both humans and the GP-UCB model explore beyond what is adaptive in the average reward context as evidenced by the peak around $t=15$, continuing to explore after most highvalue rewards have been revealed and thus failing to consistently improve average rewards ${ }^{2}$.

To harmonize the different aspects of human behavior captured by local search and by the GP-UCB model, we added a

[^246]local variant of each GP model (Local GP), which weighs the $q(\mathbf{x})$ for each arm by the inverse Manhattan distance to the previous choice, $q_{\text {Local }}\left(\mathbf{x}_{t}\right)=q\left(\mathbf{x}_{t}\right) \cdot \operatorname{IMD}\left(\mathbf{x}_{t}, \mathbf{x}_{t-1}\right)$. Adding locality to the GP models only improved prediction accuracy (Fig. 4 right), with the Local GP-UCB model having the highest overall out-of-sample prediction accuracy ( $R^{2}=.38$ ).

Overall, the modeling results show that humans display a preference for local search, but that locality alone fails to achieve comparable performance levels. The best model (Local GP-UCB) incorporated this tendency for local search into a computational model that combines function learning with a decision strategy explicitly trading off between both high expected rewards and high uncertainty.


Figure 5: Comparison of simulated model performance over 10,000 replications, where parameters were sampled from the crossvalidated MLEs of the subject population. Human results are averaged across payoff conditions and horizon length.

## Parameter Estimation

Figure 6 shows the cross-validated parameter estimates of the best predicting Local GP-UCB model. The estimates indicate subjects systematically under-estimated the smoothness of the underlying environments, with $\lambda$ values lower than the true underlying function $\left(\lambda_{\text {Smooth }}=2, \lambda_{\text {Rough }}=1\right)$, for both Rough environments $(t(36)=-4.80, p<.001)$ and Smooth environments $(t(42)=-18.33, p<.001)$, using the median parameter estimate for each subject. Participants not only had a tendency towards selecting local search options, but also made local inferences about the correlation of rewards.

All participants valued the reduction of uncertainty ( $\beta>$ 0 ), with long horizons often yielding larger $\beta$ estimates than short horizons ( 51 out of 80 subjects; $t(79)=-2.02, p=$ $.047)^{3}$. There were no differences between payoff conditions $(t(78)=-1.65, p=.1)$ or environments $(t(78)=.5, p>.1)$.

Subjects in the average reward condition yielded smaller estimates of the softmax temperature parameter ( $\tau$ ) than those in the maximum reward condition $(t(78)=-2.66, p=.009)$, This is consistent with almost all models making better predictions for average reward than for maximum reward subjects (Fig. 4), since smaller values of $\tau$ indicate more precise predictions. The larger number of unique tiles searched in the maximum reward condition (Fig. 3C) may indicate a more difficult prediction problem.

[^247]

Figure 6: Cross-validated parameter estimates for the Local GPUCB model, showing the median estimate for each participant.

## General Discussion

The results presented here can be seen as a first step towards uncovering how people search to acquire rewards in the presence of spatial correlations. We have re-cast the multi-armed bandit problem as a framework for studying both functionlearning and the exploration-exploitation trade-off by adding spatial correlations to rewards. Within a simple experiment about searching for rewards on a two-dimensional grid, we found that participants adapt to the underlying payoff condition, perform better in smooth than in rough environments, and-surprisingly-sometimes seem to perform as well in short as in long horizon settings.

Our modeling results show a tendency to prioritize local search options, which may indicate the presence of innate search costs (e.g., mouse movements or some additional cognitive processing). Even accounting for this local search behavior, our best predicting model (Local GP-UCB) indicates that people still systematically underestimate the extent of spatial correlation of rewards, preferring instead to make very local inferences about unexplored rewards. Additionally, we also found that search behavior was best predicted by a combination of both high expected reward and high uncertainty, embodied in the UCB decision strategy, which implicitly negotiates the exploration-exploitation trade-off.

Future studies could expand on this work by assessing a more diverse and perhaps combinatorial set of kernel functions (Schulz, Tenenbaum, Duvenaud, Speekenbrink, \& Gershman, 2016) or by speeding up GP-inference using approximation methods such as sparse inference (Lawrence, Seeger, \& Herbrich, 2003) or more parsimonious neural network representations (Neal, 2012). Indeed, the result that participants formed only very local beliefs about spatial correlations could be used to find heuristic approximations to GP models in the future, which could effectively trade-off a small loss in accuracy for reduced computational complexity.

## Conclusion

We compared both simple strategies and more complex function generalization models in their ability to make out-ofsample predictions about participant sampling behavior. Our modeling results indicate that there may be innate search costs, creating a tendency to prioritize local search options. Furthermore, even accounting for this local search behavior, our best performing model (Local GP-UCB) indicates that people also have a systematic tendency to underestimate the
extent of spatial correlation of rewards, preferring instead to make very local inferences about unexplored rewards.

## Acknowledgments

CMW is supported by the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World. ES is supported by the UK Centre for Financial Computing and Analytics. BM was supported by Grant ME 3717/2-2, JDN was supported by Grant NE 1713/1-2 from the Deutsche Forschungsgemeinschaft (DFG; SPP 1516). Code is available at https:// github.com/charleywu/gridsearch

## References

Bramley, N. R., Dayan, P., Griffiths, T. L., \& Lagnado, D. A. (2017). Formalizing neuraths ship: Approximate algorithms for online causal learning. Psychological Review, 124(3), 301.
Herbert, R. (1952). Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58, 527-535.
Hills, T. T., Jones, M. N., \& Todd, P. M. (2012). Optimal foraging in semantic memory. Psychological Review, 119, 431-40.
Hoppe, D., \& Rothkopf, C. A. (2016). Learning rational temporal eye movement strategies. Proceedings of the National Academy of Sciences, 113, 8332-8337.
Krause, A., \& Guestrin, C. (2005). Near-optimal nonmyopic value of information in graphical models. In Proceedings of the TwentyFirst Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05) (pp. 324-331).
Krebs, J. R., Kacelnik, A., \& Taylor, P. (1978). Test of optimal sampling by foraging great tits. Nature, 275, 27-31.
Lawrence, N., Seeger, M., \& Herbrich, R. (2003). Fast sparse Gaussian process methods: The informative vector machine. In Proceedings of the 16th Annual Conference on Neural Information Processing Systems (pp. 609-616).
Lucas, C. G., Griffiths, T. L., Williams, J. J., \& Kalish, M. L. (2015). A rational model of function learning. Psychonomic Bulletin \& Review, 22, 1193-1215.
March, J. G. (1991). Exploration and exploitation in organizational learning. Organization Science, 2, 71-87.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics, 1(2), 105-142.
Meder, B., \& Nelson, J. D. (2012). Information search with situation-specific reward functions. Judgment and Decision Making, 7, 119-148.
Neal, R. M. (2012). Bayesian learning for neural networks (Vol. 118). Springer Science \& Business Media.
Schulz, E., Huys, Q. J., Bach, D. R., Speekenbrink, M., \& Krause, A. (2016). Better safe than sorry: Risky function exploitation through safe optimization. In Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 1140-1145).
Schulz, E., Konstantinidis, E., \& Speekenbrink, M. (2016). Putting bandits into context: How function learning supports decision making. bioRxiv. Retrieved from http://www.biorxiv.org/ content/early/2016/10/14/081091
Schulz, E., Tenenbaum, J., Duvenaud, D. K., Speekenbrink, M., \& Gershman, S. J. (2016). Probing the compositionality of intuitive functions. In Advances in neural information processing systems (pp. 3729-3737).
Speekenbrink, M., \& Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. Topics in Cognitive Science, 7, 351-367.
Srinivas, N., Krause, A., Kakade, S. M., \& Seeger, M. (2010). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. Proceedings of the 27th International Conference on Machine Learning (ICML 2010), 1015-1022.
Steyvers, M., Lee, M. D., \& Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology, 53, 168-179.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction (Vol. 1) (No. 1). Cambridge, MA: MIT Press.

# What do you really think? Children's ability to infer others' desires when emotional expressions change between social and nonsocial contexts 

Yang Wu (yangwu@mit.edu), Laura E. Schulz (Ischulz@mit.edu)<br>Department of Brain and Cognitive Sciences, MIT<br>77 Massachusetts Avenue, Cambridge, MA 02139 USA


#### Abstract

We investigate children's ability to use social display rules to infer agents' otherwise under-determined desires. In Experiment 1, seven-to-ten-year-olds saw a protagonist express one emotional reaction to an event in front of her social partner (the Social Context), and a different expression behind her social partner's back (the Nonsocial Context). Children were able to use the expression in the Social Context to infer the social partner's desire and the expression in the Nonsocial Context to infer the protagonist's desire. This ability increased between ages seven and ten (Experiment 1). When task demands were reduced (Experiment 2), seven-to-eight-year-olds, but not five-to-six-year-olds, succeeded. These results suggest that although it is not easy for observers to infer emotions masked by social display rules, changing emotional expressions between social and non-social contexts allow even children to recover not only the desire of the person displaying the emotions, but also that of the audience.


Keywords: emotional expression; social display rule; mental state inference

## Introduction

Young children can use emotional expressions to draw inferences about both external events in the world (e.g., Berman, Chambers, \& Graham, 2010; Feinman, Roberts, Hsieh, Sawyer, \& Swanson, 1992; Wu, Muentener, \& Schulz, 2015), and others' internal mental states (e.g., Repacholi \& Gopnik, 1997; Rieffe, Terwogt, \& Cowan, 2005; Wellman, Phillips \& Rodriguez, 2000; Wu \& Schulz, 2017). However, because people sometimes go to great lengths to disguise their true feelings, emotional expressions can be misleading. When speaking in front of a large audience, an adult will pretend to be calm, even if she is nervous. When receiving an undesirable gift, a polite child will pretend to be happy even if she is disappointed. As we will review, a relatively large body of work has looked at children's understanding of social display rules and masked emotions. Here however, we consider a feature of social display rules that has been largely overlooked in prior work: they may disguise an individual's feelings while being informative about the feelings of her social partner's. When someone congratulates a friend in public but fumes in private, we learn not only that this person's true feelings about the event are negative, but also that her friend's true feelings are probably positive. Thus, masked emotions may reveal (about social partner's) as much as they conceal (about the individual herself). Given evidence about someone's feelings in both social and non-social contexts, an observer might therefore recover information both about
the individual's mental states, and those of the society she keeps.

This kind of inference is non-trivial: it requires tracking someone's emotional expressions across social and nonsocial contexts, reasoning recursively about the mental states of at least two parties. To our knowledge, despite abundant work on emotion understanding and theory of mind in early childhood (see Wellman, 2014 for review), no one has yet looked at whether children can use real and apparent emotions to infer not only the true feelings of the person expressing the emotions but also of their intended audience. That is our goal here.

First however, we note that there is a long line of work on children's ability to understand others' real and apparent emotions and their ability to respect display rules in their own behavior. Research suggests that young children modulate both their verbal and nonverbal responses in social contexts (Cole, 1986; Saarni, 1984; Talwar \& Lee, 2002; Talwar, Murphy, \& Lee, 2007; Xu, Bao, Fu, Talwar, \& Lee, 2010). If for instance, an experimenter has lipstick on her nose and asks a child how she looks, children as young as three lie and tell her that she looks okay (Talwar \& Lee, 2002). By three and four, children (in the laboratory anyhow) inhibit their negative emotional responses to an undesirable gift in front of a gift giver (Cole, 1986). As children get older, they are more likely to lie for pro-social purposes than for self-protective purposes (Xu, Bao, Fu, Talwar, \& Lee, 2010), and some evidence suggests that girls are better than boys at regulating their verbal and nonverbal behaviors (Cole, 1986; Davis, 1995; Saarni, 1984).

Between ages three and ten, children also show an increasing ability to understand others' masked emotions in social contexts. When predicting a recipient's response to an undesirable gift, children invoke both verbal display rules (e.g., judging that the recipient will tell a white lie) and facial display rules (e.g., judging that she will express happiness rather than disappointment; Broomfield, Robinson, \& Robinson, 2002; see also Gnepp \& Hess, 1986). Children appear to understand verbal display rules earlier than facial display rules (Broomfield, Robinson, \& Robinson, 2002), and are better at understanding display rules for pro-social purposes than for self-protective purposes (Gnepp \& Hess, 1986; but see Misailidi, 2006). The latter may be influenced by family emotional climates. For example, negative expressiveness in a family environment correlates positively with children's understanding of self-protective display rules and negatively with their understanding of pro-social display rules (Jones, Abbey, \& Cumberland, 1998). Additionally, some
researchers (Banerjee, 2002; Banerjee \& Yuill, 1999a, 1999b; Naito \& Seki, 2009) argue that the understanding of social display rules relies on an ability to represent secondorder mental state information. In support of this, children's performance on a second-order false belief task predicts their understanding of self-protective display rules (Banerjee \& Yuill, 1999b) and a more recent study suggests that it predicts both their understanding of self-protective and pro-social display rules (Naito \& Seki, 2009).

Although fruitful, much of this literature has used tasks with very rich contextual information (Banerjee, 1997; Harris, Donnelly, Guz, Pitt-Watson, 1986; Misailidi, 2006; Josephs, 1994; Wellman \& Liu, 2004; Naito \& Seki, 2009; Gross \& Harris, 1988). This is especially true for studies involving very young children. For example, in Banerjee's study (1997), preschoolers were read stories including an eliciting event (e.g., "Michelle is sleeping over at her cousin's house but she forgot her favorite teddy bear at home"), an agent's mental state (i.e., "Michelle is really sad that she forgot her teddy bear"), an intention to hide the agent's true feeling (i.e., "Michelle doesn't want her cousin to see how sad she is"), and a reason for hiding that feeling (i.e., "because her cousin will call her a baby"). Children were then asked about what the agent really feels and what she will try to look on her face. In such contexts, children may succeed without going much beyond the information available in the stories.

Consistent with this concern, studies using less informative contexts have found that an understanding of masked emotion and social display rules emerges much later in development (Broomfield, Robinson, \& Robinson, 2002; Gnepp \& Hess, 1986; Jones, Abbey, \& Cumberland, 1998). For instance, Gnepp \& Hess (1986) provided children (first, third, fifth, and tenth graders) with an eliciting event and an agent's mental state but did not explicitly mention the agent's intention to hide her feelings or any reason for her doing so. Children failed to predict the use of verbal display rules until third grade. Even adolescents (who successfully predicted the use of verbal display rules) frequently failed to predict that the agents would try to regulate their facial expressions. However, with less information in the stories, there is more uncertainty about whether the protagonist intended to be polite or not; children may have preferred to predict the emotional expression that directly mapped onto the protagonist's true mental state.

Thus, there remains some ambiguity about what children understand, and when, about masked emotions. Rich detailed scenarios may overestimate children's ability to understand social display rules, while less informative scenarios may be open to interpretations that do not involve social display rules at all.

More critically for the present purposes, previous work does not ask whether children can recover information, not only about the person displaying the emotion, but also about the person who is the intended audience of the emotion. To test this, we introduce children to a simple context where one of two teams wins a game. An observer of the game
displays one of two emotional reactions (happy or sad) in front of a social partner and the contrasting emotional expression (sad or happy) behind the social partner's back. We ask children both the desire of the person expressing the emotion, and that of his social partner. Since abundant work suggests that even infants and toddlers understand that someone whose desires are fulfilled will be happy and that someone whose desires are thwarted will be sad (see e.g., Skerry \& Spelke, 2014; Stein \& Levine, 1989; Wellman \& Woolley, 1990; Yuill, 1984), we took it for granted that by middle childhood, children could make this inference. The critical question was whether children could recover each participant's true desires given that one person (henceforth the Protagonist) displayed contradictory emotions in the social and non-social contexts, and the other person (henceforth the Social Partner) never displayed any emotion at all. (Not only do children not see the social partner's face, they have no other source of information about his emotions or desires. Thus the only way they can infer the social partner's desires is by using the protagonist's display of a false, misleading emotion in his presence. Given that without considerable scaffolding, children only appear to understand masked emotion relatively late in development (e.g., Broomfield et al., 2002; Gnepp \& Hess, 1986; Jones et al., 1998), in Experiment 1 we test seven- to ten-year-olds. In Experiment 2, we reduce the task demands and test fiveto eight-year-olds. In both cases, we look at whether children can use the emotional expression in the nonsocial context to infer the protagonist's desire and the emotional expression in the social context to infer the social partner's desire.

## Experiment 1

## Method

Participants Thirty-two children $(\mathrm{M}=8.8$ years; range: 7.2-10.8; 56\% girls) were recruited from an urban children's museum. To ensure a balanced distribution across ages, children were recruited in age bins consisting of 16 sevenand eight-year-olds ( $\mathrm{M}=7.9$ years; range: 7.2-8.8; $63 \%$ girls) and 16 nine- and ten-year-olds ( $\mathrm{M}=9.8$ years; range: $9.0-10.8 ; 50 \%$ girls). While most of the children were white and middle class, a range of ethnicities and socioeconomic backgrounds reflecting the diversity of the local population (47\% European American, 24\% African American, 9\% Asian, $17 \%$ Latino, $4 \%$ two or more races) and the museum population ( $29 \%$ of museum attendees receive free or discounted admission) were represented throughout.
Materials Each child saw two illustrated stories, one presenting the Happy-Sad condition (e.g., Tom was happy in front of Bryan but sad behind Bryan's back) and the other presenting the Sad-Happy condition. The facial expressions were from istock photos (http://www.istockphoto.com/) and have been used by previous research (Wu \& Schulz, 2017). The mapping between stories and conditions, and the order of conditions were counterbalanced across participants, resulting in a total of 4 storybooks. Different agents and
games were used in each storybook (Tom, Bryan, and basketball in one story and Sally, Diana, and volleyball in the other).
Procedure Children were tested individually; all sessions were videotaped. Children were asked check questions to encourage them to follow along. Incorrect responses were corrected throughout. Children had little difficulty with the check questions. Check questions were used only to maintain children's attention; they were not analyzed or used as inclusion criteria.

Each story was read consecutively, as follows (using the basketball-game story as an example). The experimenter placed the first picture on the table and said, "There is a basketball game today. It's the Tiger team against the Lion team." She introduced the next picture and said, "This is Tom. Tom is a basketball fan. He loves watching basketball games. He goes to watch the game. He is either a fan of the Tiger team, or the Lion team, but we don't know which one." Children were asked (Check question 1): "Do we know which team Tom is a fan of?" The experimenter introduced the third picture and said, "This is Bryan. Bryan was Tom's friend when they were little, but now they don't get to see each other very much. Bryan becomes a basketball player. He plays in the game. He either plays for the Tiger team or the Lion team, but we don't know which one." Children were asked (Check question 2): "Do we know which team Bryan plays for?" The experimenter introduced the fourth picture and said, "The result of the game was that the Tiger team won, and the Lion team lost." Then the experimenter introduced the fifth picture and said, "After the game, Bryan ran back to the locker room. Tom was passing by and saw Bryan. It was a very noisy and crowded room and they didn't have a chance to talk. However, in front of Bryan, when Tom came passing by, Tom made a face like this." Children were asked (Check question 3): "Did Tom look happy or sad?" The experimenter introduced the sixth picture and said, "However, behind Bryan's back, as soon as Bryan passed by and couldn't see Tom, Tom made another face." Children were asked (Check question 4): "Did Tom look happy or sad?" We controlled for the complexity between the social and nonsocial contexts by having two people in both contexts; the difference was only that in the social context, they were facing towards each other, and in the nonsocial context, they were facing away from each other. (See Figure 1.)

Finally, the experimenter asked two test questions. The first question was about the protagonist (Protagonist Question): "Now I am going to ask you some questions. In front of Bryan, Tom looked [happy/sad] but behind Bryan's back, Tom looked [sad/happy]. Do you think Tom is a fan of the Tiger team or Lion team?" The experimenter then asked the other test question (Social Partner Question): "Does Bryan play for the Tiger team or the Lion team?" ${ }^{1}$

[^248]

Figure 1 Example of the materials used in Experiments 1 and 2 (corresponding to the fourth to sixth pictures described in Procedure).

Coding We scored children's responses separately for the Protagonist and the Social Partner. Children received one point for answering a question correctly and none for answering it incorrectly.

## Results and discussion

Participants performed equally well in the Happy-Sad and Sad-Happy conditions (protagonist: $X^{2}=.59, p=.442$; social partner: $X^{2}=.00, p=1.00$ ). Additionally, there was no order effect between the first and second stories (protagonist: $X^{2}=.07, p=.798$; social partner: $X^{2}=.00, p=$ 1.00). Thus, we collapsed children's scores across the two conditions. This resulted in a score of $0-2$ for the Protagonist and a score of 0-2 for the Social Partner.

Using age as a continuous variable, we found that children between ages seven and ten showed an increasing ability to use the emotional expression in the nonsocial context to reason about the protagonist's desire $(\beta=.75, S E$ $=.36, z=2.11, p=0.035$; Ordinal Logistic Regression), and the emotional expression in the social context to recover the

[^249]social partner's desire ( $\beta=.69, S E=.34, z=2.02$, $p=0.044$ ). See Figure 2.

Overall, children did not recover the protagonist's desire above chance ( $z=$ 1.40, $p=.230$ ) and showed a non-significant trend towards recovering the social partner's desire ( $z=$ 2.13, $p=.052$ ) but there was no significant difference between the two ( $z=-1.13, p$ $=.453$ ). Because of the age effect, we did a median split on age. Seven- and eight-year-olds did not perform above chance on either


Figure 2 Results of Experiments 1 and 2. The top row shows individual children's performance on the questions about each agent, collapsing across the Happy-Sad and Sad-Happy conditions. The bottom row shows children's performance by age bin. Error bars indicate $95 \%$ confidence intervals. question (protagonist: $z=-$
$.58, p=.774$; social partner: $z=.00, p=1.00$ ); however, nine- and ten-year-olds performed above chance on both (protagonist: $z=2.50, p=.022$; social partner: $z=2.67, p$ $=.013$ ). See Figure 2.

These results suggest that nine- and ten-year-olds can use changing emotional expressions between social and nonsocial contexts to recover the desires of both the protagonist and the social partner in a masked emotion context. However, why did younger children fail? As noted, many previous studies suggest that by seven and eight, children can predict an agent's real and apparent emotions given relatively rich contextual information (Banerjee, 1997; Harris, Donnelly, Guz, Pitt-Watson, 1986; Misailidi, 2006; Josephs, 1994; Wellman \& Liu, 2004; Naito \& Seki, 2009; Gross \& Harris, 1988; Gnepp \& Hess, 1986; Broomfield, Robinson, \& Robinson, 2002; Jones, Abbey, \& Cumberland, 1998). They can also represent second-order mental state information (Perner \& Wimmer, 1985; Sullivan, Zaitchik, \& Tager-Flusberg, 1994), which supports the understanding of social display rules. Thus, it is possible that children's chance performance here was due to task demands. In particular, children may have tripped up by the fact that the first expression they saw was an apparent, misleading emotional expression. Only when children saw the second expression, did they have the information to tell that the first expression was a fake one.

In the next experiment, we reduce these task demands by flipping the order of the social and nonsocial contexts. Thus, children first see the agent's emotional expression in the nonsocial context and then a different one in the social context. This order does not require children to re-interpret the first emotional expression; additionally, the first expression may provide a basis for children to understand the expression displayed in the social context. To see if even younger children might succeed given these reduced task demands, we test both seven- and eight-year-olds and five- and six-year-olds.

## Experiment 2

## Method

Participants Thirty-two children ( $\mathrm{M}=7.0$ years; range: 5.3-8.8; $66 \%$ girls) were recruited from the children's museum. Half of them were seven- and eight-year-olds ( $n=$ $16 ; M=8.0$ years; range: $7.1-8.8 ; 75 \%$ girls) and the other half were five- and six-year-olds ( $n=16 ; M=6.0$ years; range: 5.3-6.8; $56 \%$ girls).
Materials, procedure and coding The materials, procedure and coding were identical to Experiment 1 except that we flipped the order of the social and nonsocial contexts. See Figure 1. For example, instead of first showing Tom's emotional expression in front of Bryan, the experimenter presented Tom's expression behind Bryan's back: "After the game, Tom made a face like this. At this moment, Bryan was nearby but Tom didn't see him." Children were asked a check question: "Did Tom look happy or sad?" The experimenter then introduced the next picture and said, "However, Tom turned around and saw Bryan. Tom made another face." Children were asked another check question: "Did Tom look happy or sad?"

## Results and discussion

As in Experiment 1, participants performed equally well in the Happy-Sad and Sad-Happy conditions (protagonist: $X^{2}=$ $2.82, p=.093$; social partner: $X^{2}=.67, p=.412$ ). There was no order effect between the first and second stories (protagonist: $X^{2}=.93, p=.335$; social partner: $X^{2}=.67, p$ $=.412$ ). Thus, children's scores were collapsed across the two conditions.

We used the same analyses as in Experiment 1. Taking age as a continuous variable, we found that children between ages five and eight showed an increasing ability to recover both the protagonist's $(\beta=.89, S E=.35, z=2.53, p$ $=.011$ ) and the social partner's desires $(\beta=.81, S E=.36, z$ $=2.25, p=0.024$ ). See Figure 2.

Overall, there was a non-significant trend for children to recover the protagonist's desire $(z=1.96, p=.078)$ and children successfully recovered the social partner's desire ( $z$ $=3.15, p=.002$ ); there was no significant difference between the two $(z=-1.13, p=.453)$. Given the age effect, we performed a planned median split on age. Five- and six-year-olds did not perform above chance in either question (protagonist: $z=.00, p=1.00$; social partner: $z=1.34, p$ $=.375$ ) but seven- and eight-year-olds succeeded in both (protagonist: $z=2.50, p=.022$; social partner: $z=2.89, p$ $=.006$ ). See Figure 2. Thus, we found that at least by ages seven and eight, children can use changing emotional expressions between social and nonsocial contexts to recover the desires of both participants in a social exchange, even when one participant is masking her emotions and the only cue to the other participant's desires is that misleading, masked emotional reaction.

## General Discussion

In two experiments, we investigated children's ability to use the information embedded in social display rules to recover others' otherwise under-determined mental states. Children saw an emotional expression when a protagonist was in front of a social partner, and a different expression when the protagonist was behind the social partner's back. Children successfully used the expression in the nonsocial context to infer the protagonist's desire, and the expression in the social context to infer the social partner's desire. Children's ability to draw these inferences improved between ages five and eight.

Our study builds on many previous studies that have looked at children's ability to predict an agent's real and apparent emotions given rich mental state information (e.g., the agent's desires, true feelings, her intentions, and a motivation to hide her true feelings; Banerjee, 1997; Harris, Donnelly, Guz, Pitt-Watson, 1986; Misailidi, 2006; Josephs, 1994; Wellman \& Liu, 2004; Naito \& Seki, 2009; Gross \& Harris, 1988; Gnepp \& Hess, 1986; Broomfield, Robinson, \& Robinson, 2002; Jones, Abbey, \& Cumberland, 1998). In contrast, here we provided children with very minimal background information, and no direct information about the agent's mental states. Children's ability to use the social context to recover the desires of an agent who provided two contradictory emotional reactions to an event, and also the desire of a social partner, whose emotional expressions were never observed at all, is consistent with other studies finding that children can recover rich unobserved information from observed emotional cues (e.g., Berman, Chambers, \& Graham, 2010; Feinman, Roberts, Hsieh, Sawyer, \& Swanson, 1992; Wu, Muentener, \& Schulz, 2015; Repacholi \& Gopnik, 1997; Rieffe, Terwogt, \& Cowan, 2005; Wellman, Philips \& Rodriguez, 2000; Wu \& Schulz, 2017). However, our study goes beyond those studies in suggesting that children can also detect and understand the conditions in which real emotions are masked.

Although emotional expressions are misleading when people mask their true feelings, our results indicate that the
masking behavior itself (if detected) can be richly informative. Note that feigning an emotional expression in front of others reflects one's beliefs and desires about others' beliefs or desires. Thus when a feigned emotional expression is detected, it contains recursive mental state information about what one agent thinks about what another agent thinks. Although there has been debate on the extent to which reasoning about pro-social display rules requires second-order mental state representation (Banerjee, 2002; Banerjee \& Yuill, 1999a, 1999b; Naito \& Seki, 2009), in our task, the social partner's beliefs, desires, and emotions were unknown throughout. To recover information about the social partner, children had to refer to the protagonist and selectively use the protagonist's emotional expressions to gain insight into the mind of his audience. We suggest that this kind of inference does require recursive mental state reasoning, and the current results suggest that the ability to make these inferences develops over middle childhood.

Critically, children succeeded here in a very tightly constrained context: there were only two possible outcomes (one of two teams won a game), two possible emotional responses (happy or sad) and two social partners. Moreover, the task design virtually eliminated any memory demands: children did not need to track the changing emotional expressions over time; they were all concurrently displayed in the storybook card format, together with the social context. Future work might look at children's ability to draw comparable inferences when they must track changing emotional dynamics over time and in more complex, multi-participant scenarios. Note however, that although more realistic scenarios may add processing demands and complexity, they may also provide children with richer cues to agents' mental states.

The current results however, suggest that by age seven, children can recover underlying mental states from changes between real and apparent emotional expressions. Intriguingly, the current results also suggest that there is a limit to how much we can hide when we hide our feelings: in disguising our true feelings, we may reveal what we think about what other people want.

## Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. Warm thanks to the Boston Children's Museum and participating parents and children. Thanks to Rachel Magid, Michelle Wang, and Kary Richardson for helpful feedback, and to Catherine Wu and Elizabeth Rizzoni for help with data collection and coding.

## References

Banerjee, M. (1997). Hidden emotions: Preschoolers' knowledge of appearance-reality and emotion display rules. Social Cognition, 15(2), 107-132.
Banerjee, R. (2002). Children's understanding of selfpresentational behavior: Links with mental-state
reasoning and the attribution of embarrassment. MerrillPalmer Quarterly, 48(4), 378-404.
Banerjee, R., \& Yuill, N. (1999a). Children's explanations for self-presentational behaviour. European Journal of Social Psychology, 29(1), 105-111.
Banerjee, R., \& Yuill, N. (1999b). Children's understanding of self-presentational display rules: Associations with mental-state understanding. British Journal of Developmental Psychology, 17(1), 111-124.
Berman, J. M., Chambers, C. G., \& Graham, S. A. (2010). Preschoolers' appreciation of speaker vocal affect as a cue to referential intent. Journal of Experimental Child Psychology, 107(2), 87-99.
Broomfield, K. A., Robinson, E. J., \& Robinson, W. P. (2002). Children's understanding about white lies. British Journal of Developmental Psychology, 20(1), 47-65.
Cole, P. M. (1986). Children's spontaneous control of facial expression. Child Development, 57, 1309-1321.
Davis, T. L. (1995). Gender differences in masking negative emotions: Ability or motivation?. Developmental Psychology, 31(4), 660.
Feinman, S., Roberts, D., Hsieh, K. F., Sawyer, D., \& Swanson, D. (1992). A critical review of social referencing in infancy. In Social referencing and the social construction of reality in infancy (pp. 15-54). Springer US.
Gnepp, J., \& Hess, D. L. (1986). Children's understanding of verbal and facial display rules. Developmental Psychology, 22(1), 103.
Gross, D., \& Harris, P. L. (1988). False beliefs about emotion: Children's understanding of misleading emotional displays. International Journal of Behavioral Development, 11(4), 475-488.
Harris, P. L., Donnelly, K., Guz, G. R., \& Pitt-Watson, R. (1986). Children's understanding of the distinction between real and apparent emotion. Child Development, 895-909.
Jones, D. C., Abbey, B. B., \& Cumberland, A. (1998). The development of display rule knowledge: Linkages with family expressiveness and social competence. Child Development, 69(4), 1209-1222.
Josephs, I. E. (1994). Display rule behavior and understanding in preschool children. Journal of Nonverbal Behavior, 18(4), 301-326.
Misailidi, P. (2006). Young children's display rule knowledge: Understanding the distinction between apparent and real emotions and the motives underlying the use of display rules. Social Behavior and Personality: an international journal, 34(10), 1285-1296.
Naito, M., \& Seki, Y. (2009). The relationship between second-order false belief and display rules reasoning: the integration of cognitive and affective social understanding. Developmental Science, 12(1), 150-164.
Perner, J., \& Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. Journal of Experimental Child Psychology, 39(3), 437-471.

Repacholi, B. M., \& Gopnik, A. (1997). Early reasoning about desires: evidence from 14 -and 18 -month-olds. Developmental Psychology, 33(1), 12.
Rieffe, C., Terwogt, M. M., \& Cowan, R. (2005). Children's understanding of mental states as causes of emotions. Infant and Child Development, 14(3), 259-272.
Saarni, C. (1984). An observational study of children's attempts to monitor their expressive behavior. Child Development, 1504-1513.
Skerry, A. E., \& Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. Cognition, 130(2), 204-216.
Stein, N. L., \& Levine, L. J. (1989). The causal organisation of emotional knowledge: A developmental study. Cognition \& Emotion, 3(4), 343-378.
Sullivan, K., Zaitchik, D., \& Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. Developmental Psychology, 30(3), 395.
Talwar, V., \& Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. International Journal of Behavioral Development, 26(5), 436-444.
Talwar, V., Murphy, S. M., \& Lee, K. (2007). White lietelling in children for politeness purposes. International Journal of Behavioral Development, 31(1), 1-11.
Wellman, H. M. (2014). Making minds: How theory of mind develops. Oxford University Press.
Wellman, H. M., \& Liu, D. (2004). Scaling of theory-of-mind tasks. Child Development, 75(2), 523541.

Wellman, H. M., Phillips, A. T., \& Rodriguez, T. (2000). Young children's understanding of perception, desire, and emotion. Child Development, 71(4), 895-912.
Wellman, H. M., \& Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. Cognition, 35(3), 245-275.
Wu, Y., Muentener, P., \& Schulz, L. (2015). A fine-grained understanding of emotions: Young children match withinvalence emotional expressions to their causes. In Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 2785-2690).
Wu, Y. \& Schulz, L. E. (2017). Inferring beliefs and desires from emotional reactions to anticipated and observed events. Child Development. DOI: 10.1111/cdev. 12759 First Published online on 8 March 2017.
Xu, F., Bao, X., Fu, G., Talwar, V., \& Lee, K. (2010). Lying and truth-telling in children: From concept to action. Child Development, 81(2), 581-596.
Yuill, N. (1984). Young children's coordination of motive and outcome in judgements of satisfaction and morality. British Journal of Developmental Psychology, 2, 73-81.

# Discovering Multicausality in the Development of Coordinated Behavior 

Tian Linger Xu (txu@indiana.edu)<br>Computer Science and Cognitive Science Department, 1101 East 10th Street, Indiana University<br>Bloomington, IN 47405 USA<br>Drew H. Abney (dhabney@indiana.edu)<br>Psychological \& Brain Sciences Department, 1101 East 10th Street, Indiana University<br>Bloomington, IN, 47405

Chen Yu (chenyu@indiana.edu)<br>Psychological \& Brain Sciences, Cognitive Sciences, and Informatics Department, 1101 East 10th Street, Indiana University Bloomington, IN, 47405


#### Abstract

Human interaction involves the organization of a collection of sensorimotor systems across space and time. The study of how coordination develops in child-parent interaction has primarily focused on understanding the development of specific coordination patterns from individual modalities. However, less work has taken a systems view and investigated the development of coordination among multiple interdependent behaviors. In the present work, we used Granger causality as a mathematical model to construct dyadic causal networks of multimodal data collected from a longitudinal study of child-parent interaction. At a grouplevel, we observed increases in the number of causal links and in the strength of such links in dyadic interaction from 9months to 12 -months. At an individual-level, we observed high variability in the types of causal links that emerged across developmental ages. We discuss these results in terms of a multicausality hypothesis for the development of human coordination.


Keywords: Interpersonal Coordination; Social Interaction; Child-Parent Interaction; Granger Causality; Multimodal Social Interaction; Multivariate Autoregressive Model

## Introduction

Human interaction entails the organization of a vast array of sensorimotor systems across space and time (Kendon, 1970). We imitate, align and synchronize over a spectrum of social behaviors with our social partners during communication and studies have shown fine-grained temporal structures across modalities in interpersonal coordination (Fusaroli \& Tylén, 2016; Garrod \& Pickering, 2009; Louwerse, Dale, Bard, \& Jeuniaux, 2012). How we are able to organize behaviors across multiple modalities and achieve seamless coordination in only fractions of a second is one of the most important questions about human cognition (Marsh, Richardson, \& Schmidt, 2009).
One effective approach to answering this question is to examine how such smooth coordination evolves during development. In developmental science, past research have shown that within specific behavioral modalities, coordinated behaviors emerge early in life and develop incrementally with age (Yale, Messinger, Cobo-Lewis, \& Delgado, 2003). For example, infants start to follow and
coordinate the gaze direction of their social partner (Scaife \& Bruner, 1975) and form vocal and facial expression feedback loops with their parents early in their first year of life (Cohn \& Tronick, 1988). Such social contingencies are suggested to be indicative of later language development (Goldstein, King, \& West, 2003; Mundy \& Newell, 2007; Warlaumont, Richards, Gilkerson, \& Oller, 2014).

Development is about change. The multicausality assumption in dynamical systems theory (Smith \& Thelen, 2003) indicates that change and growth in the system emerge through the relationships between different interdependent components, without an executive preprogrammed and unified path. Certain patterns and behavioral influences emerge or diminish at different developmental ages, and through different developmental pathways. In light of this, the aim of the present study is to examine the change in the organization of coordination among multiple interdependent behaviors. More specifically, we want to investigate the connectivity and directional influences from one modality to another in the course of development.

Towards this goal, in this paper, we proposed a novel approach to modeling multimodal coordinated behaviors between children and parents as a directed graph network with Granger causality (Bressler \& Seth, 2011; Granger, 1969). A longitudinal study was conducted in which we invited children at 9 months and their parents to participate in a toy play experiment, and again at 12 months. During the toy play sessions, we recorded the dyad's momentary eye gaze and manual action data with eye-trackers and multiview video recording. With this study and our analytic approach, we can investigate the development of human coordination through directional causal relations among a network of interdependent behavioral variables.

This framework of modeling child-parent interaction as causal networks allowed us to determine changes in the amount of causal links and the strength of causal links across 9- and 12-months. We tested two specific hypotheses about the development of coordination. First, the developmental hypothesis: on a group level, we expected that the number and strength of causal links in the childparent coordination network would increase from 9- to 12-
months. Second, the multicausality hypothesis: we expected the increased coordination to be achieved by the emergence of new causal influences in the network, among multiple different behavior variables. One key assumption of this hypothesis is that no causal link has developmental priority. If dyads show individual differences in their coordination development pattern, it would be an indication that they each follow distinct pathways to achieve increased levels of sensorimotor coordination.

## Granger Causality for Point Process Data

Coordination patterns change throughout the course of an interaction and require real-time adjustment of actions and predictions in accordance with their sensorimotor input (Clark \& Brennan, 1991; Yu \& Smith, 2016). When we study interpersonal coordination and development from a dynamical approach, one challenge is quantifying directional influence and connectivity between two specific variables. This is due, in part, to the interconnectivity and complexity of information exchange among behavioral variables (Fusaroli, Konvalinka, \& Wallot, 2014; Hidaka \& Yu, 2010).

Granger causality, or G-causality, is a well-established and effective method for the investigation of directional relationships among a set of interdependent variables in many domains (Bressler \& Seth, 2011). Granger (1969) formalized the basic idea of causality between signals introduced by Wiener (1956) based on multivariate autoregressive (MVAR) models: if past values of $\boldsymbol{Y}$ contain information that help predict $\boldsymbol{X}$ above and beyond the information contained in the past values of $\boldsymbol{X}$ alone, then $\boldsymbol{Y}$ is said to Granger-cause $\boldsymbol{X}$.
Kim et al. (2011) proposed a point process framework to enable G-causality to be applied to point process data with a discrete nature. A temporal point process is a stochastic time series of binary events that occurs in continuous time. It can only take on two values at any point in time, indicating whether or not an event has occurred. With a time series dataset of an ensemble of variables, the occurring likelihood of the event variable $\boldsymbol{X}$ can be modeled by the generalized linear model (GLM): a linear combination of time series $\boldsymbol{X}$ 's dependency to the history of each individual element in the ensemble. Given a set of multivariate temporal streams, the causal relationships from variable $\boldsymbol{Y}$ to $\boldsymbol{X}$ is assessed by calculating the relative reduction in the likelihood of producing this particular history of time series of $X$ when the history of $\boldsymbol{Y}$ is excluded, compared with the likelihood if all the available covariates are used in the prediction calculation. If the prediction likelihood is reduced when the history of variable $\boldsymbol{Y}$ is excluded from calculation, then there exists a Granger causal relationship from $\boldsymbol{Y}$ to $\boldsymbol{X}$. In addition, Kim et al. (2011) proposed that the sign of averaged influence of the occurring history of variable $\boldsymbol{Y}$ on $\boldsymbol{X}$ can be used to distinguish excitatory (positive estimate) and inhibitory (negative estimate) influences: whether the event history of $\boldsymbol{Y}$ is more or less likely to lead to the event occurring for variable $\boldsymbol{X}$. Finally, the point process
framework also affords researchers to identify the statistical significance of a causal link based on the likelihood ratio test statistic. The goodness-of-fit statistics were applied by comparing the deviance between the estimated model with trigger variable $\boldsymbol{Y}$ excluded and the estimated full model in the GLM framework. Then, a multiple hypothesis testing error measure, FDR, proposed in (Benjamini \& Hochberg, 1995; Storey, 2002) was used to control the expected proportion of false discovery rate when the number of hypothesis tests is large and the number of rejected null hypotheses is consequentially large.

Calculating G-causality with GLM model fitting makes very general assumptions about the data (Barnett, Barrett, \& Seth, 2009) and with the point process framework, we are able to apply G-cause to categorical behavioral data. In the present paper, we used this framework to construct quantitative causal networks among different behavioral modalities in child-parent interaction and study child's coordination development.

## Methods

## Participants

21 parent-child dyads participated in this study. Dyads came into the lab when the children were 9 -months-old and 12-months-old.


Figure 1: (a) A dual eye-tracking child-parent interaction paradigm. (b) Eye movement and object manipulation events from both the child and parent were coded into categorical data streams. The data streams were then divided into three different ROI groups, preserving only the onset of events. Finally, per subject, the three groups were concatenated as input for subsequent calculations of Granger causality.

## Procedure

Figure 1(a) shows the experimental setup of our dual eyetracking child-parent interaction experimental paradigm (Yu \& Smith, 2013, 2016). Parents and their children were seated across from each other at a plain white table $(61 \mathrm{~cm} \times$ $91 \mathrm{~cm} \times 64 \mathrm{~cm}$ ). Head-mounted Positive Science eye trackers (Franchak, Kretch, Soska, \& Adolph, 2011) were put on both the child and parent to capture their gaze data in real time. Each eye-tracking system includes an infrared camera that records eye images (mounted on the head and pointed to the right eye, see Figure 1a), and a scene camera capturing the first-person view from the participant's perspective. The scene camera's visual field is 108 degrees providing a broad view. Each eye-tracking system recorded both the first-person view video and precise gaze allocation in that view, with a sampling rate of 30 frames per second. Another high-resolution camera was mounted above the table and provided a bird's-eye view at a recording rate of 30 frames per second.

For each trial of the experiment, there were two sets of toys. Each set consisted of three toy objects with three different colors (blue, green, red). The toys were of similar size and weight. Parents were told that the goal of the experiment was to study how parents and toddlers interacted with objects during free play and they were asked to engage their children with the toys as what they would naturally do in daily life. Each of the two sets of toys was played with twice for 90 seconds, resulting in approximately six minutes of play over four trials from each dyad. Toy set order (ABAB or BABA) was counterbalanced across dyads.

## Data Processing

Human coders went through the videos from multiple viewpoints and manually annotated frame-by-frame about which object was gazed at and held by the child and the parent with both of their hands. In this study, we coded four Region-Of-Interest (ROI)s for the eye movement data: blue, green and red object categories (1-3) and other (0). Each value represents where the child or the parent was looking at in every frame. The participants could be looking at each other's face, but our analysis didn't include face looking events in this paper. The same object and empty ROIs (0-3) were also the coding categories for hand action data streams, indicating the target object was held by either the left or the right hand of the child and the parent. For each trial, after data processing, four coded categorical data event streams (child gaze events, child holding events, parent gaze events, and parent holding events) were obtained.

The next step was to convert our behavioral temporal data streams into multivariate point processes. All behavioral data streams were divided into three groups by different ROIs and then only the onsets of object ROI events were preserved to fit the point process framework for calculating G-causality. Figure 1(b) shows the point process data streams from one experimental trial. After point process conversion, for each dyad, three groups were concatenated as input data for calculating G-causality. In each group, all
streams contained the onset of the same category of events. With this point process data transformation, we extracted Granger causality among different behavioral variables acting on the same object. For example, we estimated Gcausality from the event of child looking at the red object to the occurrence of the parent looking at the same object.

## Analysis

For each dyad, we constructed a dyadic causal network among four behavioral variables (child eye movement, child hand action, parent eye movement and parent hand action) at 9 months and 12 months. Figure 2 shows the G-cause network constructed with two dyads' interaction data. In each network, there are 4 behavioral variables (child eye, child hand, parent eye and parent hand) and 12 different types of directional links between every pair of variables. The different types of directional links are illustrated in Figure 2.

Significance tests based on the likelihood ratio test statistic with FDR controlling false positive causal interactions (Storey, 2002; Kim et al., 2011) was performed to determine the statistical significance of every causal link with regard to the entire network. In Figure 2, red colored links indicate the significantly positive links with number at the end of each link representing the G-cause value from one behavioral variable to the other. For example, at 12 months, Dyad\#1 had a significantly positive causal link from child's gaze to child's holding behavior. This means that the child was looking at a certain object and the occurrence of this event significantly increased the likelihood of the child holding the same object. In addition, to best comprehend the magnitude of G-cause values for our


Figure 2: (a) The G-cause coordination network among child eye, child hand, parent eye and parent hand time series for Dyad\#1 at 9 months (left) and 12 months (right); red links are significantly positive G-cause links and the number indicates the G-cause value of that causal relation. (b) The G-cause coordination network for Dyad\#2 at 9 months (left) and 12 months (right).
multimodal coordination data, we also calculated the baseline G-cause network for every interaction. This was done by randomizing the order of event streams (with all ROIs and their event durations) for the behavioral variables. Then, the randomized onsets of object ROI events were preserved to convert the data to fit point process model for baseline G-cause network calculation.
The source code, a more detailed explanation of the Granger causality calculation process and more supplementary materials of this study are available at: https://github.com/lingerxu/Granger_causality_coordination

## Results

To examine our developmental hypothesis - increased coordination from 9 -months to 12 -months - we first looked at two group-level measures: the number of significantly positive G-cause links and the average G-cause value per link in each interaction network. For example, in Figure 2a, Dyad\#1 had 3 significantly positive links at 9 months and 5 links at 12 months and the average G-cause value per link was 2.96 at 9 months (baseline value 0.19 ) and -2.27 at 12 months (baseline value -0.04). Average baseline G-cause values obtained with the randomized event streams were close to 0 for both age groups. In the present paper, we focused on examining the significantly positive G-cause links, which have much higher values than baseline and entail a strong causal link from one behavior variable to another.
As shown in Figure 3, we observed more significantly positive G-cause links at 12 months ( $M=3.95, S D=0.23$ ) compared to each dyad's network at 9 months ( $M=2.38$, $S D=0.20), t(20)=3.27, p=.004$. We also observed that the Gcause network for 12 month olds ( $M=5.50, S D=0.39$ ) had significantly higher average G-cause values per link than 9 months ( $M=2.52, S D=0.26$ ), $t(20)=3.85, p<.0001$. Overall, the multimodal coordination between child and parent showed increased developmental changes from 9 months to 12 months. The observation of increased positive causal links in the network and higher G-cause values on average from 9 - to 12 -months, suggests that the coordinative patterns of the child-parent dyadic system are becoming more dense and stronger.

## Multicausality and Individual Differences

The main proposal of the multicausality hypothesis is that increased coordination is achieved by the emergence of multiple new causal influences between different pairs of behavioral variables and that no causal link has developmental priority. The results observed in the last section provided clear evidence that child-parent dyadic systems become more coordinated from 9 months to 12 months. Next, we want to look at how this increased level of coordination was achieved and whether we will observe individual differences in the developmental pattern in the dyadic causal network.


Figure 3: (a) Amount of significantly positive G-cause links and (b) average G-cause values of child-parent eye hand coordination networks at 9 months and 12 months.

When we take a closer look at the individual development between the two networks of each dyad, and how each causal link in the network changed from 9 months to 12 months, there are multiple types of change. Here we will mainly focus on examining the emergence of new significantly positive link, which means that this positive causal link did not exist in the 9 -month coordination network, and only appeared in the 12 -month network.

With 12 different types of G-cause links in total, the development of the coordination network can be described by a vector of developmental changes in each type of causal relations. The developmental coordination row vector for each dyad is visualized in Figure 4a. Three causal relation links, child hand $\rightarrow$ child eye, parent hand $\rightarrow$ child hand and child hand $\rightarrow$ parent hand, are omitted in the illustration because we did not observe any emergence of new positive links in these three link types. For example, the two dyads in Figure 2 can be mapped to the first two vector representations in Figure 4a. For Dyad\#1, two new positive links emerged in their G-cause network at 12 months. This emergence is depicted in the developmental coordination vector: two red cells in parent eye $\rightarrow$ child eye and child eye $\rightarrow$ child hand categories (see Figure 4a, row 1). In another example, for Dyad\#2 (see Figure 4a, row 2), five new links emerged from 9 months to 12 months. And we can see that, between the two dyads, four out of five emergent links from Dyad\#2 were completely different from the G-cause relation types in which Dyad\#1's emergent links belonged to.

Finally, if increased coordination from 9 months to 12 months was achieved through one type of causal link with causal priority, then the hypothesized frequency distribution of emergent links will be similar to Figure 4b. We can observe that the majority of emergent links belong to the same causal relation type. Alternatively, the multicausality hypothesis entails that increased coordination is achieved via multiple different causal relations. In an ideal situation, we would observe a uniform frequency distribution of emergent causal relations. This possibility is depicted in Figure 4 c . Figure 4 d shows the empirical frequency distribution of emergent links. The empirical distribution provides evidence for a diffuse collection of emergent causal relations, supporting the multicausality hypothesis


RANK ORDERED DIStribution of New Links among 12 types of g-Cause link
Figure 4: (a) The development coordination vector for each dyad's G-cause network. Red cells indicate the emergence of significantly positive G-cause links from 9 months to 12 months between different pairs of behavior variables. Each row represents the developmental change in coordination network for one dyad. Each column represents the developmental change for a particular type of causal relation link. Three causal relation links, child hand $\rightarrow$ child eye, parent hand $\rightarrow$ child hand and child hand $\rightarrow$ parent hand, are omitted here because we did not observe the emergence of significantly positive links. (b) The hypothesized frequency distribution of emergent causal links if increased coordination was achieved by only one link with causal priority. (c) Illustration of the frequency distribution of emergent links for the ideal uniform distribution under the multicausality hypothesis. (d) The empirical frequency distribution of emergent links in our results.
that child-parent dyads are utilizing multiple coordination patterns to achieve increased coordination.

## General Discussion

The goal of the present paper was to investigate the development of multimodal organization in naturalistic child-parent interactions. We used a novel causal network modeling approach to better understand how multimodal dyadic systems change across developmental age. The observed results provide preliminary evidence for the developmental and multicausality hypotheses that we proposed at the outset of the paper.
At a group-level, we observed an increase in the amount of causal links and an increase in the strength of causal links from 9 months to 12 months. These results provided support for the developmental hypothesis, suggesting that the multimodal coordination patterns across the child-parent dyadic system became stronger with more components being coordinated within the dyadic system. This is an important observation because it provides novel evidence for an important property of the developing child-parent dyadic system: development includes adding redundancy to the social interaction by creating new pathways for coordination to occur (Yu \& Smith, 2016). Redundancy is an important property for any complex system because it affords adaptability in the face of intrinsic and extrinsic
perturbations (Kugler \& Turvey, 1987; Thelen \& Smith, 1998).

At an individual level, we observed that the causal relation links were distributed among all types of G-cause relations between two behavioral variables both within and between agents. Furthermore, the frequency distribution of emergent causal links was approximately uniform suggesting that there was no single behavioral link taking developmental causal priority in the network. These results add preliminary support for the multicausality hypothesis. These observations provide important conceptual and empirical contributions. Multicausality has been proposed to be an important property of a complex system (Smith \& Thelen, 2003), however there has been little work to extend the proposal of multicausality to a dyadic model of childparent interactions. This framework quantifies the directional causal influences between different behavioral variables to model the complex system of interpersonal coordination at sensorimotor level. Thus, it can provide heuristics towards understanding the individual differences in the establishment of joint attention and possibly the reasons underlying the correlations between joint attention and many developmental outcomes (Mundy et al., 2007; Tomasello \& Farrar, 1986; Yu \& Smith, 2016). Finally, to our knowledge, this is the first study to use MVAR-based Granger causality to model multimodal coordination as directed causal networks. Our results provide evidence for
the promise of this analysis method as a novel dynamic modeling method for many domains, such as developmental science, behavioral science, etc.

## Acknowledgments

We thank Melissa Elston, Steven Elmlinger, Charlotte Wozniak, Melissa Hall, Charlene Tay, and Seth Foster for collection of the data; Seth Foster and Thomas Smith for developing data management and processing software; and Linda Smith, Lei Yuan and Lauren Stone for fruitful discussions. This work was funded by NIH grants R01 HD074601, R21 EY017843, T32 HD07475

## References

Barnett, L., Barrett, A. B., \& Seth, A. K. (2009). Granger causality and transfer entropy Are equivalent for gaussian variables. Physical Review Letters, 103(23), 1-10.
Benjamini, Y., \& Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B.

Bressler, S. L., \& Seth, A. K. (2011). Wiener-Granger Causality: A well established methodology. NeuroImage, 58(2), 323-329.
Clark, H. H., \& Brennan, S. E. (1991). Grounding in communication. Perspectives on Socially Shared Cognition.
Cohn, J. F., \& Tronick, E. Z. (1988). Mother-infant face-toface interaction: Influence is bidirectional and unrelated to periodic cycles in either partner's behavior. Developmental Psychology, 24(3), 386-392.
Franchak, J. M., Kretch, K. S., Soska, K. C., \& Adolph, K. E. (2011). Head-mounted eye-tracking: A new method to describe infant looking. Child Development, 82(6), 17381750.

Fusaroli, R., Konvalinka, I., \& Wallot, S. (2014). Analyzing Social Interactions: The Promises and Challenges of Using Cross Recurrence Quantification Analysis. Translational Recurrences, 137-155.
Fusaroli, R., \& Tylén, K. (2016). Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. Cognitive Science, 40(1), 145-171.
Garrod, S., \& Pickering, M. J. (2009). Joint Action, Interactive Alignment, and Dialog. Topics in Cognitive Science, 1(2), 292-304.
Goldstein, M. H., King, A. P., \& West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. Proceedings of the National Academy of Sciences of the United States of America, 100(13), 8030-5.
Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica, 37(3), 424-438.
Hidaka, S., \& Yu, C. (2010). Analyzing multimodal time series as dynamical systems. In International Conference on Multimodal Interfaces and the Workshop on Machine

Learning for Multimodal Interaction (p. 53: 1-8). ACM.
Kendon, A. (1970). Movement coordination in social interaction: Some examples described. Acta Psychologica, 32(C), 101-125.
Kim, S., Putrino, D., Ghosh, S., \& Brown, E. N. (2011). A Granger causality measure for point process models of ensemble neural spiking activity. PLoS Computational Biology, 7(3), 1-13.
Kugler, P. N., \& Turvey, M. T. (1987). Information, natural law, and the self-assembly of rhythmic movement. Resources for Ecological Psychology, 481.
Louwerse, M. M., Dale, R., Bard, E. G., \& Jeuniaux, P. (2012). Behavior matching in multimodal communication Is synchronized. Cognitive Science, 36(8), 1404-1426.
Marsh, K. L., Richardson, M. J., \& Schmidt, R. C. (2009). Social Connection Through Joint Action and Interpersonal Coordination. Topics in Cognitive Science, 1(2), 320-339.
Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., \& Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. Child Development, 78(3), 938-954.
Mundy, P., \& Newell, L. (2007). Attention, joint attention, and social cognition. Current Directions in Psychological Science, 16(5), 269-274.
Scaife, M., \& Bruner, J. S. (1975). The capacity for joint visual attention in the infant. Nature, 253(5489), 265266.

Smith, L. B., \& Thelen, E. (2003). Development as a dynamic system. Trends in Cognitive Sciences, 7(8), 343348.

Storey, J. D. (2002). A Direct Approach to False Discovery Rates on JSTOR. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 479-498.
Thelen, E., \& Smith, L. B. (1998). Dynamic systems theories. In Handbook of child psychology.
Tomasello, M., \& Farrar, M. J. (1986). Joint attention and early language. Child Development, 1454-1463.
Warlaumont, A. S., Richards, J. a, Gilkerson, J., \& Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. Psychological Science, 25, 1314-1324.
Wiener, N. (1956). The theory of prediction. In E. F. Beckenbach (Ed.), Modern mathematics for engineers (Vol. 1, pp. 165-190). New York: McGraw-Hill.
Yale, M. E., Messinger, D. S., Cobo-Lewis, A. B., \& Delgado, C. F. (2003). The temporal coordination of early infant communication. Developmental Psychology, 39(5), 815-824.
Yu, C., \& Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. PLoS ONE, 8(11).
Yu, C., \& Smith, L. B. (2016). Multiple Sensory-Motor Pathways Lead to Coordinated Visual Attention. Cognitive Science, 1-27.

# Unifying recommendation and active learning for human-algorithm interactions 

Scott Cheng-Hsin Yang ${ }^{1}$ (scott.cheng.hsin.yang@ gmail.com), Jake Alden Whritner ${ }^{1}$ (jake.whritner@ rutgers.edu), Olfa Nasraoui ${ }^{2}$ (olfa.nasraoui@ gmail.com) \& Patrick Shafto ${ }^{1}$ (patrick.shafto@gmail.com)<br>${ }^{1}$ Department of Mathematics \& Computer Science, Rutgers University-Newark<br>${ }^{2}$ Department of Computer Engineering and Computer Science, University of Louisville


#### Abstract

The enormous scale of the available information and products on the Internet has necessitated the development of algorithms that intermediate between options and human users. These algorithms do not select information at random, but attempt to provide the user with relevant information. In doing so, the algorithms may incur potential negative consequences related to, for example, "filter bubbles." Building from existing algorithms, we introduce a parametrized model that unifies and interpolates between recommending relevant information and active learning. In a concept learning paradigm, we illustrate the trade-offs of optimizing prediction and recommendation, show that there is a broad parameter region of stable performance that optimizes for both, identify a specific regime that is most robust to human variability, and identify the cause of this optimized performance. We conclude by discussing implications for the cognitive science of concept learning and the practice of machine learning in the real world.


Keywords: Recommender systems, active learning, concept learning, filter bubble

Historically, the information each individual had access to was defined by one's local environment: what one could directly observe , who one had to talk to or do business with, and available texts or catalogs one could access. With the advent of the Internet, information and products became available at a global scale. This vast potential resource creates a problem: how to choose-from billions or trillions of options-which information or products to present to an individual at a given time. Solutions to these problems form the foundation that supports major players in the online business world-from search engines and e-commerce to social network services-such as Google, Amazon, and Facebook. These algorithmic solutions radically affect not only what information and products we are exposed to, but also which information and products we have the chance to be exposed to. Thus, these algorithms mediate between us and reality, not by providing a random sample from what is possible, but by carefully selecting a sample which optimizes some underlying goals and metrics. The consequences of these humanalgorithm interactions have been insufficiently explored despite recent interest in cases such as filter bubbles (Pariser, 2011), algorithmic bias (Baeza-Yates, 2016), and humanalgorithm interaction biases (Nasraoui \& Shafto, 2016).

A well-established doctrine in cognitive science asserts that a driving factor of our beliefs is the information we are exposed to. However, the situations investigated in the most typical concept learning experiments (Bruner, Goodnow, \& Austin, 1956; Shepard, Hovland, \& Jenkins, 1961) differ sharply from the kinds of situations we encounter with recommender systems. In concept learning experiments, examples are typically sampled either exhaustively or randomly, neither
of which is feasible in the context of Internet-scale problems. Obviously, enumeration is not feasible. Random sampling is also not feasible because if the quality of the algorithm's suggestions were too poor, human users could simply choose to go elsewhere. This yields a thorny problem: how to select information and products to maximize relevance, while also accurately estimating what users want.

Two classes of algorithms-information filters (Sparck Jones, 1970; Van Rijsbergen, 1979; Salton, Fox, \& Wu, 1983) and recommender systems (Goldberg, Nichols, Oki, \& Terry, 1992; Maes et al., 1994; Adomavicius \& Tuzhilin, 2005)—have been developed to facilitate selection of information for users. Although different in some ways, they share a core assumption that the goal is to deliver humans relevant information or products. Given that these sorts of algorithms have raised concerns about not exposing people to the breadth of potentially relevant information, basic questions one might raise are whether the data they obtain allow them to accurately estimate human users' preferences, and whether there are small adjustments that could be made to optimize for recommendation and learning about the users' preferences.

One approach for obtaining optimally informative data is active learning, well known in both cognitive science (Nelson, 2005) and computer science (MacKay, 1992). Active learning has been proposed both as a model for how humans search for information and as an algorithm for how machines learn about the world. In the current context, active learning is a method for learning what information is relevant to the human user (Elahi, Ricci, \& Rubens, 2016), whichwhile having advantages for estimating the user's beliefs-is unlikely to produce quality recommendations.

Drawing inspiration from real-world problems of information filtering and recommendation, we seat these problems in a concept learning framework that allows for experimental control of which examples are relevant. This framework allows for an exploration of algorithms that perform better or worse on the joint problems of optimizing recommendations and inferring relevance. We introduce a novel approach to investigating algorithm performance that merges aspects of computational simulation and user testing: people are trained on the true concept, then they interact with the algorithm to test performance. Unlike in computational simulations and user testing, there is both defined ground truth and naturalistic human variability in behavior.

Our approach uses the simple concept learning task previously used by Markant and Gureckis (2014). We present two experiments. The first validates the method by demon-
strating expected limitations of recommendation and active learning. The second investigates a simple one-parameter generalization-active recommendation-that unifies both approaches. We show that while the extreme cases of pure active learning and pure recommendation yield poor performance, all intermediate values converge to near optimal recommendation and prediction. We also observe that due to human variability, parameter values that are closer to pure recommendation yield the best performance. We conclude by discussing implications for cognitive science in the lab and machine learning in the real world. Overall, the main contributions of our work are: (i) the use of an experiment to gauge how real human users interact with a system that spans graded shades between recommendation and active learning and (ii) how a unified, yet simple and generic model is beneficial in the design and interpreting of real user experiments.

## Unifying recommending and active learning

Given a dataset, $D=\left\{x_{i}, y_{i}\right\}_{i}^{N}$, the goal of a probabilistic classification algorithm is to predict the probability that a new data point $x^{*}$ belongs to class $y, P\left(y \mid x^{*}, D\right)$. We will be concerned with learning two classes corresponding to irrelevant and relevant information, $y \in\{0,1\}$. These predictions form the basis of the recommendation and active learning algorithms we will consider. Intuitively, the goal of recommendation is to provide the user with examples that are relevant. This intuition can be formalized directly,

$$
\begin{equation*}
x_{\text {rec }}=\underset{x^{*}}{\arg \max } P\left(y=1 \mid x^{*}, D\right) \tag{1}
\end{equation*}
$$

At any point-given previously observed data-this defines which examples are optimal for recommendation: those that maximize the probability of being relevant. Within the closely related problem of probabilistic retrieval (ranking relevant information), this coincides with optimal probabilistic retrieval for the most relevant item (Robertson, 1977).

One intuitive formalization of active learning is to select examples that reduce our uncertainty about which examples are relevant. This can also be formalized directly,

$$
\begin{equation*}
x_{a c t}=\underset{x^{*}}{\arg \min }\left|0.5-P\left(y=1 \mid x^{*}, D\right)\right| . \tag{2}
\end{equation*}
$$

Given previously observed data, the optimal example to observe is the one about which we have the greatest predictive uncertainty.

It is worth noting that this is not the only formalization of active learning that one may consider. Other well known strategies include optimizing information gain (K-L divergence), diagnosticity, and probability gain (Nelson, 2005). While each differs in formal detail, in many practically relevant situations, their predictions are quite similar. We formalize active learning as the selection of maximally uncertain data to facilitate integration with the recommendation criterion in Eq. 1, as will be seen below.

We propose a unified model of active recommendation that exploits the parallel structure in previous models. Our single parameter generalization includes filtering and active learn-
ing as extreme cases and thus unifies the two approaches and interpolates between them. Formally,

$$
\begin{equation*}
x_{\alpha}=\underset{x^{*}}{\arg \min }\left|\alpha-P\left(y=1 \mid x^{*}, D\right)\right|, \tag{3}
\end{equation*}
$$

where $\alpha \in[0.5,1]$. When $\alpha=0.5$ we recover active learning, as is obvious from inspection of Eq. 2. When $\alpha=1$, we recover recommendation. In our context, subtracting from 1 and taking the min is equivalent to taking the max in Eq. 1.

Of interest is what happens between the extremes that correspond to recommendation and active learning. Are there parameterizations of active recommendation that optimize accuracy in terms of recommendation and prediction? Are there parameterizations that are more robust to the kinds of variability that are characteristic of human behavior?

## Experiments

In what follows, we empirically investigate these questions using a novel approach. Human subjects were first trained on the underlying conceptual structure that defines which examples are relevant and which are not. The classes of relevant examples are defined by axis-aligned logistic function with data standardization in two dimensions. Next, people are randomly assigned to an algorithm, and are presented with a series of examples, which they label as being in the relevant class or not. The algorithm updates upon receiving each example-label pair, and then selects a new example. This method combines aspects of computational simulation and user testing by providing a ground truth, yet allowing human variability in responses. It thus provides information about when we would expect algorithms to perform well-both absolutely and in the presence of human variability.

The questions of interest are: which algorithms perform well in terms of recommendation and prediction and which ones perform well in terms of robustness to human variability? An algorithm's trial-by-trial recommendation accuracy is the fraction of examples labeled as relevant, at each trial, by a population of participants. Its predictive accuracy for trial $i$ is the fraction of correct predictions-made by the classification algorithm trained with data up to trial $i-$ where predictions are tested on a grid of predetermined, held-out test examples. The correctness is judged against the optimal decision boundary that was set in the beginning of the experiment.

Two experiments follow. Experiment 1 investigates the performance of pure recommendation and active learning, and compares them with random sampling. This experiment allows us to validate that recommendation and active learning fail to predict and recommend well, respectively, and provides a random sampling baseline. Experiment 2 investigates the unified active recommendation model, which interpolates between pure recommendation and pure active learning. This experiment characterizes recommendation and prediction accuracies of the algorithm in the context of concept learning.

## Experiment 1

Participants. The experiment was run on Amazon's Mechanical Turk (MTurk) with 30 participants in each of the
three conditions: recommend, active learning, and random.
Stimuli. Following Markant and Gureckis (2014), the stimuli were circles with a central diameter. The stimuli varied along two dimensions- the size of the circle's radius in pixels and the orientation of the central diameter in degrees (for an example, see Figure 5 B in Markant and Gureckis (2014)). The ranges of the size and orientation were fixed to 110 pixels and 140 degrees, respectively. The minimum radius and minimum orientation for the classes were sampled independently and uniformly from 10 to 30 units and fixed for the whole experiment. This procedure determined a pair of minimum and maximum values $\{\min , \max \}$ for each dimension.

For each experiment, one of the dimensions (size or orientation) was randomly selected as the separable dimension. Let the $\left\{\min _{s}, \max _{s}\right\}$ be the minimum and maximum values of the separable dimension, and $\left\{\min _{t}, \max _{t}\right\}$ be the values for the other dimension. Two classes were defined by two two-dimensional normal distributions. Along the separable dimension, the variances of the two classes were both 75 , and their means were set at $\left(\min _{s}+\max _{s}\right) / 2 \pm 30$. Along the other dimension, the variances were both 2250 , and the means were both $\left(\min _{t}+\max _{t}\right) / 2$. Stimuli were sampled from the twodimensional Gaussian described above. Those that happened to be outside the determined range were resampled.

The experiment consisted of three phases: training, interaction, and testing. In each trial of the training phase, a class was randomly sampled, and a stimulus was sampled according to its class distribution. In the interaction phase, there were several sampling algorithms. Random sampling used the procedure as in the training phase. Recommendation and active-learning sampling followed Eqs. 1 and 2 respectively. The choice was made from a fixed pool of 400 randomly sampled stimuli for each experiment. In the test phase, the stimuli were no longer sampled from the classes but from a test set. The test set consisted of $16 \times 16$ samples that lied on a regular grid covering the area of feature space defined by $[10,140]$ pixels $\times[10,170]$ degrees. Five stimuli were randomly selected from each of the four quadrants in that area to form the 20 test stimuli used in the test phase.

Procedures. Before the training phase, participants were instructed that throughout the experiment, they would see a series of "loop antennas" that receive signals from music stations called "Beat" and "Sonic" (the two classes of stimuli described above). They were instructed that the station received depends upon the antenna's radius and the orientation of its diameter. The goal of the training phase, as described to the participants, was for them to learn which station was received by a given class of antennas (e.g., Beat antennas have large diameters and Sonic have small). Participants provided input by clicking on one of two buttons (labeled Beat and Sonic respectively). After responding, participants received feedback on whether or not their input was correct. The participants moved on to the interaction phase once they had 19 correct answers in the past 20 trials.

The interaction phase was comprised of two parts. Participants were first instructed to pretend that they preferred either Beat or Sonic. Given this preference, participants were told that they would teach an algorithm to recommend the station that they preferred by indicating-by clicking on a button-that the antenna it chose was one that they "like" or "dislike." Participants were instructed to pay attention to whether the algorithm was improving or not. This part of the interaction phase continued for 20 trials. Next, participants rated the algorithm's improvement-that is, how well the participants thought the algorithm learned to recommend their preferred station-using a slide bar from "very poor" to "excellent."

The final phase of the experiment, the test phase, entailed a classification test to confirm whether participants still remembered the categories correctly. This phase followed the same procedure as the initial training phase, but did not provide participants with feedback (i.e., they were not told whether or not their categorization was correct). Afterwards, participants were asked to provide feedback about the experiment and identify the rule behind the classification they were trained on. Participants who successfully completed all phases of the experiment were compensated via MTurk.
Analysis. We quantify the behavior of the sampling algorithms by their trial-by-trial recommendation and predictive accuracies, as described previously. The test examples for computing the predictive accuracy consisted of a grid of 10-by-10 examples covering the area spanned by two pairs of \{min, max\} sampled for each experiment (see the Stimuli section under Experiment 1).

We report the first trial index by which an algorithm's recommendation accuracy becomes statistically different from $50 \%$ as well as the first trial index at which its recommendation accuracy becomes statistically no different from $95 \%$. For these we use the binomial test and claim statistical significance when p-value is less than 0.05 . We also report the trial index at which an algorithm's predictive accuracy converges. We formalize this as the first trial at which the predictive accuracy is not statistically different from the prediction accuracy at the last trial, using a one-sample t-test. The accuracy at the final trial is reported as the converged value.

We omit subjects whose test accuracy is below 18 out of 20 (below 90\%). For the included subjects, we compute a consistency score, which is the fraction of their responses to the recommended examples that matched the expected response. For subjects whose consistency score is below $50 \%$, we computed the predictive accuracies after flipping all their responses in the recommendation phase. This allowed us to correct for the responses from subjects who misremembered the preference during the interaction phase. A $0 \%$ consistency would flip the classification algorithm's prediction on every test example. We assume that the fraction of properly predicted examples is proportional to consistency. Thus, to maximize the fraction of proper predictions, we flip responses when consistency is $<50 \%$. The number of included subjects are 26/30 (3 flipped) for random, 27/30 (4 flipped) for active


Figure 1: Predictive accuracy and recommendation accuracy for all six conditions over time (trial index indicated on x -axis).
learning, and 27/30 (3 flipped) for recommendation.
Results. Figure 1 shows the recommendation and predictive accuracies of the different sampling algorithms. As expected, examples chosen under the recommendation objective result in high recommendation accuracy, but low predictive accuracy. As a function of the number of examples seen (trial index), recommendation accuracy rises above chance level and reaches $95 \%$ after 4 examples, ${ }^{1}$ while predictive accuracy converges after 5 examples to $81 \%$, which is low compared to the active learning or random algorithms.

Conversely, recommendation accuracy under the activelearning algorithm results in low recommendation accuracy, but high predictive accuracy. As a function of trial index, recommendation accuracy remains at chance level, while predictive accuracy converges after 5 examples to $92 \%$.

For reference, results of random sampling are also presented. These show a pattern similar to that observed for active learning. There is a rapid increase in predictive accuracy, converging after 8 examples to $95 \%$. Recommendation accuracy remains at chance level throughout.

## Experiment 2: Exploring active recommendation

An ideal algorithm would combine both high recommendation and high predictive accuracy. As a function of the number of examples given, one hopes that the recommendation accuracy will approach 1 after a few examples, and the predictive accuracy will steadily increase to 1 . Given the sharp dichotomy between the performance on recommendation and active learning, it is not obvious how best to achieve this.

We explore a simple, one parameter generalization of recommendation and active learning that we call, active recommendation. We investigate its trace of accuracy under a range of $\alpha=(0.55,0.75,0.95)$. We look at how the predictive and recommendation accuracies interpolate between the active learning and recommendation sampling as a function of $\alpha$. The new sampling algorithm is as described in Eq. 3. The stimuli and procedure are the same as Experiment 1.

[^250]

Figure 2: Predictive and recommendation accuracies for all included subjects broken down by condition. Red traces corresponds to individual subjects, and the blue curve is the average. (a) Active training; (b) $\alpha=0.55$; (c) $\alpha=0.75$; (d) $\alpha=0.95$; (e) Recommend. Note that a few mislabeled examples in the early trials can lead to unstable behavior, such as those curves that dip below chance level.

Participants. The experiment was run on MTurk with 30 subjects for each of the 3 conditions: $\alpha=(0.55,0.75,0.95)$. Following the criteria described above, the number of subjects included in the analysis is 27/30 (4 flipped) for $\alpha=0.55$, $26 / 30$ (5 flipped) for $\alpha=0.75$, and 23/30 (3 flipped) for $\alpha=0.95$.
Results. Figure 1 shows the plot of predictive and recommendation accuracies for all conditions. The predictive accuracies in the active-recommendation conditions converge to $93 \%, 90 \%$, and $93 \%$ for $\alpha=(0.55,0.75,0.95)$, respectively. These are similar to the $92 \%$ in the active condition and bet-


Figure 3: The distributions of like/dislike examples for each condition. The dotted lines in the distributions indicate the distributions' quartiles.
ter than the $81 \%$ in the recommend condition. The predictive accuracies of the active-recommendation conditions converged after 5 , 5 , and 8 examples for $\alpha=(0.55,0.75,0.95)$, respectively. The recommendation accuracies in the activerecommendation conditions reached $95 \%$ after 12, 16, and 4 examples for $\alpha=(0.55,0.75,0.95)$, respectively. These are similar to the recommendation condition in that all reached $95 \%$, whereas recommendation accuracies in the active and random conditions remain at chance level.

Importantly, if we move slightly away from the active condition (sampling from slightly farther away from the boundary than active; i.e., from $\alpha=0.5$ to 0.55 ), we can achieve much higher recommendation accuracy (it rises above chance after 8 examples and reaches $95 \%$ after 12 examples vs. at chance level throughout), while also achieving similar predictive accuracy. Similarly, if we move slightly away from the recommendation condition (from $\alpha=1$ to 0.95 ), we can maintain the recommendation accuracy while improving the predictive accuracy. Thus, these intermediate conditions (in terms of $\alpha$ ) appear to allow the algorithms to uncover more of the space that is relevant.

Figure 2 employs the same measures as Figure 1, but displays each of the six conditions separately, with individual participant performance (red lines) and the results averaged over all participants (blue lines). Figures 2a-2e allow for a closer look at individual variability during the experiment, and in particular highlight the difference in recommendation accuracy from the active learning and $\alpha=0.55$ conditions. If we compare Figures 2 a and 2 b , we can see that variation in recommendation accuracy across individuals persists for all trials in the active condition, while reducing greatly after 8 to 12 trials in the $\alpha=0.55$ condition. Comparing Figure 2 d and 2 e , we can see the predictive accuracy across individuals varies much less in the $\alpha=0.95$ condition than in the recommendation condition, resulting in the better average predictive accuracy for $\alpha=0.95$.

The cause of the improved performance of intermediate $\alpha$ values can be traced back to the examples they select. The distributions of "likes" and "dislikes" are plotted in Figure 3 alongside random sampling, active learning, and recommen-
dation. At the top, random sampling replicates the true distribution (up to some small number of inconsistent responses). Active learning selects examples that are evenly distributed across likes and dislikes but shifted toward the boundary between the two categories. At the bottom, recommendation selects examples that are skewed away from the boundary and the balance of examples is strongly tilted toward likes, consistent with the goal of recommending relevant examples. Of particular interest are the three alpha conditions. There are minor differences focused on the distribution of disliked items. What is most notable are the similarities among them and the active learning distribution for likes. Unlike the recommend condition, all three intermediate conditions disproportionately select "liked" examples that are close to the boundary. They all also select relatively few "disliked" examples. Cross-referencing against Figure 2, these disliked items happen only in the early trials. To summarize, the advantage of the active recommendation approach is a bias to select uncertain items within the relevant category. This allows them to achieve both high recommendation and predictive accuracy.

Interestingly, if we include only the fully consistent subjects, the $\alpha$ values dictate a strict ordering in both the predictive and recommendation accuracy. Increasing $\alpha$ from 0.5 to 1 , one sees a monotonic decrease in the converged predictive accuracy and a monotonic increase in the rate at which recommendation accuracy reaches 1 . The stochasticity in the subjects' responses can break the ordering in two ways. First, algorithms that provide examples closer to the boundary will receive more noisily labeled examples. Second, randomness in responses slows down the convergence of the classification algorithm. These effects cause the converged prediction accuracies, in small $\alpha$ conditions, to be lower than what they could be with less variable responses.

## Discussion

Information filters and recommender systems mediate between humans and the vast information and product stores on the Internet. Naturally, these algorithms aim to provide relevant information, but this goal may also lead to negative consequences by overly restricting experience. Embedding recommendation into a concept learning framework, we investigate the conditions under which we may observe high recommendation and predictive accuracy, in the presence of naturalistic human variability. We introduced a unified model of recommendation and active learning which we call active recommendation. In well-controlled experiments, we show that-across a wide range of parameterizations-active recommendation converges toward optimal predictive and recommendation accuracy. We also observe that parameterizations closer to pure recommendation yield better performance in terms of faster convergence and greater robustness to human variability. We trace the success of active recommendation to the fact that all parameterizations automatically combine rapid convergence toward selecting only relevant items and actively exploring informative examples from within that
set. Parameterizations close to pure recommendation perform best because they minimize exploration of regions of the space where human actions are most variable-near the boundary and in the non-focal category.

Our approach is unusual in that the goal is to use humans to investigate the behavior of algorithms. This makes sense because the algorithms are meant for recommending options to humans. In contexts where recommendation is typically applied, however, there is no known ground truth, which makes assessing the performance of algorithms difficult. One could assume a ground truth and perform computational simulations, but these assume that your simulation is robust to human-like variability, which is rarely known or checked. In our experiments, humans were taught very simple concepts that governed relevance. They then labeled data for the algorithm, which captures the kinds of uncertainty associated with cognition-stochasticity across time, in response to recent input, and features of the concepts. The results bear the fruits of the approach. If one considers only the people who labeled correctly in their interactions, active recommendation performs comparably well across a wide range of parameterizations. However, human variability is concentrated at the boundary and toward the non-focal concept, which gives parameterizations closer to pure recommendation a distinct advantage in recommendation and predictive accuracy.

Our proposed unified model of active recommendation takes pure recommendation and active learning as a starting point. However, across a wide range of parameterizations, the unified model exhibits behavior that is qualitatively different from either. That is, it achieves good performance on both goals of recommendation and active learning simultaneously. It is useful to consider this behavior in contrast with more explicit alternative approaches, namely, managing the exploitation-exploration trade-off in reinforcement learning. Formalizing and training a policy about when to apply recommendation (exploitation) or active learning (exploration) would certainly be more involved than the simple model we presented; it would also arguably miss the point. The active recommendation approach, in denying the existence of the dichotomy, allows simultaneous optimization of recommendation and prediction.

Active recommendation can be recast as a social active learning model where an agent asks questions to learn from another agent who may not answer because of disinterest, ignorance, or some other factors. In these social scenarios, good questions should depend on the answerer's preference, knowledge state, etc.. In the cognitive development literature, empirical studies have shown that children select questions based on the answerer's expertise (Kushnir, Vredenburgh, \& Schneider, 2013). This exemplifies an interesting connection between our model and human social learning.

Although active recommendation has demonstrated excellent performance, the problems considered here are vastly simpler than those more typical of real-world recommendation or information filtering. In light of this, one may rea-
sonably ask whether the results are likely to generalize to more complex, high-dimensional problems. Of course, active learning becomes decreasingly tractable as the space grows. This is why active recommendation may be expected to perform well. Instead of exploring the space of possibilities, active recommendation focuses on exploring the space of relevant possibilities. An important direction for future work is to formalize and test this question.

Often experimental control and real-world relevance are seen in competition. However, there are ways in which they can and should be complementary. Real-world applications of machine learning are especially amenable to this due to their algorithmic nature. In addition to the user studies that are typical of the applied computer science, we propose that more controlled experimental and modeling approaches in cognitive science can shed light on the core strengths and limitations of these algorithms.

## Acknowledgments

This research was supported in part by NSF grant NSF-1549981 to P.S. and O.N. through IIS and SBE.

## References

Adomavicius, G., \& Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734-749.
Baeza-Yates, R. (2016). Data and algorithmic bias in the web. In Proceedings of the 8th ACM Conference on Web Science.
Bruner, J. S., Goodnow, J., \& Austin, G. (1956). A study of thinking. New Brunswick, New Jersey: Transaction Publishers.
Elahi, M., Ricci, F., \& Rubens, N. (2016). A survey of active learning in collaborative filtering recommender systems. Computer Science Review, 20, 29-50.
Goldberg, D., Nichols, D., Oki, B. M., \& Terry, D. (1992, December). Using collaborative filtering to weave an information tapestry. Commun. ACM, 35(12), 61-70.
Kushnir, T., Vredenburgh, C., \& Schneider, L. A. (2013). who can help me fix this toy? the distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. Developmental psychology, 49(3), 446.
MacKay, D. J. (1992). Information-based objective functions for active data selection. Neural computation, 4(4), 590-604.
Maes, P., et al. (1994). Agents that reduce work and information overload. Communications of the ACM, 37(7), 30-40.
Markant, D. B., \& Gureckis, T. M. (2014). Is it better to select or to receive? learning via active and passive hypothesis testing. Journal of Experimental Psychology: General, 143(1), 94.
Nasraoui, O., \& Shafto, P. (2016). Human-algorithm interaction biases in the big data cycle: A markov chain iterated learning framework. arXiv preprint arXiv:1608.07895.
Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. Psychological Review, 112(4), 979.
Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin Press.
Robertson, S. (1977). The probability ranking principle in ir. Journal of Documentation, 33(4), 294-304.
Salton, G., Fox, E. A., \& Wu, H. (1983). Extended boolean information retrieval. Communications of the ACM, 26(11), 1022-1036.
Shepard, R. N., Hovland, C. I., \& Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs: General and Applied, 75(13), 1.
Sparck Jones, K. (1970). Some thoughts on classification for retrieval. Journal of Documentation, 26(2), 89-101.
Van Rijsbergen, C. (1979). Information retrieval. London, UK: Butterworths.

# A non-parametric Bayesian prior for causal inference of auditory streaming 

Tim Yates ${ }^{1}$ (t.a.yates@bham.ac.uk)<br>Nathanael Larigaldie ${ }^{2}$ (nathanael.c.larigaldie@durham.ac.uk) Ulrik R. Beierholm ${ }^{1,2}$ (ulrik.beierholm@durham.ac.uk)<br>${ }^{1}$ Psychology, University of Birmingham, Edgbaston, B15 2TT, UK<br>${ }^{2}$ Psychology, Durham University, Durham, DH1 3LE, UK


#### Abstract

Human perceptual grouping of sequential auditory cues has traditionally been modeled using a mechanistic approach. The problem however is essentially one of source inference - a problem that has recently been tackled using statistical Bayesian models in visual and auditory-visual modalities. Usually the models are restricted to performing inference over just one or two possible sources, but human perceptual systems have to deal with much more complex scenarios. To characterize human perception we have developed a Bayesian inference model that allows an unlimited number of signal sources to be considered: it is general enough to allow any discrete sequential cues, from any modality. The model uses a non-parametric prior, hence increased complexity of the signal does not necessitate more parameters. The model not only determines the most likely number of sources, but also specifies the source that each signal is associated with. The model gives an excellent fit to data from an auditory stream segregation experiment in which the pitch and presentation rate of pure tones determined the perceived number of sources.


Keywords: Bayesian modeling; Cognitive model; Causal reasoning; Computational neuroscience; Audition.

## Introduction

Ambiguity in perceptual systems is a blight for inference. When we hear two sounds sequentially, we may infer that they came from two different sources, $A$ and $B$, or the same source repeated. A third sound is heard - are the sources $\mathrm{AAA}, \mathrm{AAB}, \mathrm{ABA}, \mathrm{ABB}$ or ABC ? By the time four, five and six sounds are heard the number of combinations reaches 15, 52, 858. The ambiguity breeds to generate $a$ combinatorial explosion, and yet the human auditory system is able to reliably allocate multiple sources of sound in complex, real world situations. Features of the signal are consistently associated with different sources, allowing us to keep track of a speaker's voice and the wail of an ambulance siren, separate from the noise of background traffic and falling rain.

For several decades, the human ability to segregate sequential sounds into streams corresponding to sources has been investigated using simple sequences of either pure tones or more complex sounds (reviewed in (B. C. J. Moore \& Gockel, 2012)). The time interval between tones, their pitch difference and the duration of a sequence are among the factors that play an important role (Anstis \& Saida,

1985; Bregman \& Campbell, 1971; van Noorden, 1975): explanations of how the factors are used based on principles such as Gestalt laws and Occam's razor have been incorporated into the sophisticated conceptual model of Bregman (Bregman, 1994). Descriptive models based on peripheral excitation (Beauvois \& Meddis, 1997), coherence of coupled oscillators (Wang, 1996) and cortical streaming modules (McCabe \& Denham, 1997) provide mechanisms to estimate the number of streams, but do not specify which sound is associated with which source. While some of the models are expandable to allow more sources to be inferred, it is not known if they would cope with the combinatorial explosion. Furthermore, Moore \& Gockel (B. Moore \& Gockel, 2002) conclude from an extensive review of the literature that any sufficiently salient factor can induce stream segregation. This indicates that a more general model of inference is needed, that can incorporate any auditory perceptual cue and multiple sounds with different sources.
If ambiguity is a blight for inference, regularities in natural signals are the cure. Not all combinations of signal sources are equally likely - when perceptual systems generate a model of the world, we assume that they infer the most likely interpretation because the perceptual systems are optimized to the statistics of natural signals (Barlow, 1961; McDermott \& Simoncelli, 2011). Bayesian inference has had considerable success in modeling many visual and multi-sensory percepts as a generative, probabilistic process (Shams, et al. 2005; Weiss et al. 2002). Despite these successes, and the increasing evidence for the importance of predictability for auditory perception (for a review see Bendixen, 2014), we still have no general, principled model of how the auditory system solves the source inference problem.

A Bayesian approach to auditory stream segregation has been used to model the dynamics of perceptual bistability (Lee \& Habibi, 2009) but assumes that only two percepts are possible. Turner (2010) has developed methods of analyzing statistics of sounds based on Bayesian inference, and constructed a model to synthesize realistic auditory textures. Promisingly, inference in the model can qualitatively replicate many known auditory grouping rules.

In our model the probability of many alternative stream configurations (given the input signal) are calculated and the percept generated corresponds to the most probable configuration. The probabilities are calculated using Bayes'
rule to combine the likelihood of generating a signal given a postulated stream configuration, with the prior probability of sounds being associated with different sources. The likelihood and prior probability distributions are iteratively updated in a principled manner as information accumulates. The forms of the distributions are presumably optimized to natural signal statistics: the likelihood distribution we use is based on considerations of the physical limitations of oscillators. However, the framework of the model allows formulations of multiple explanatory factors, such as those determined by Bregman (1994) from psychophysics experiments, to be simply incorporated in the distributions. Furthermore, while the current study uses simple pure tones (replicating work by Bregman), the framework allows more complex cues from audition and other modalities to be used as long as their perceptual difference can be quantified.

## Human inference model

Pure tones are the indivisible atoms of input to the model each being assigned to just one sound source, or stream. Inspired by work done on non-parametric priors (Froyen, Feldman, \& Singh, 2015; Orbanz \& Teh, 2010; Wood, Goldwater, \& Black, 2006) we assume the existence of an infinite number of potential sources, leading to a sequence of tones with pitch $f_{1}, f_{2} \ldots$, onset time $t_{1}^{o n}, t_{2}^{o n} \ldots$ and an offset time, $t_{1}^{\text {off }}, t_{2}$ off $\ldots$ and the sound sources/streams that generated the tones are denoted by positive integers $S_{l}, S_{2} \ldots$ We rename the sources when necessary so that the first tone heard will always be generated by source 1 (i.e. $S_{l}=1$ ), and a subsequent tone, $S_{n}$ can be associated with source $1: \max \left(S_{1} \ldots S_{n-1}\right)+1$.

## Generative model

Given a source $S_{i}$ we assume that the frequency of tone $i$ is governed by physical constraints and statistical regularities of the source. If two sounds $f_{l}$ and $f_{2}$ with frequencies $F_{I}$ and $F_{2}$ are produced by the same source, the pitch cannot change at an infinitely fast rate: to make an oscillator change its frequency discontinuously would require an infinite impulse of energy. We assume that, all things being equal, a pure tone sound source is most likely to continue oscillating at the same frequency as it has in the past, and the probability of it changing at a rate $\Delta F / \Delta t$ will decrease as $\Delta F / \Delta t$ increases. More specifically we assume a normal probability distribution:

$$
\begin{equation*}
p\left(f_{i} \mid S_{i}, f_{i-t}, S_{\mathrm{i}}=S_{i-t}\right)=\frac{1}{\sqrt{2 \pi \sigma^{2}}} e^{-\frac{\left(\frac{\Delta F}{\Delta t}\right)^{2}}{2 \sigma^{2}}} \tag{1}
\end{equation*}
$$

where $\sigma$ is a constant. We here assume that the observer has a perfect noise free access to the generated frequency.

## Inference

The task of the observer is to infer the sources generating each of the tones, i.e. to find the $S_{1} S_{2} S_{3} \ldots$ that maximize $p\left(S_{1} S_{2} S_{3} \ldots \mid f_{1} f_{2} f_{3} \ldots\right)$, as illustrated in figure 1. As an example we use a sequence of three tones $f_{1} f_{2} f_{3}$, for which the observer wishes to infer the likely sources $S_{1} S_{2} S_{3}$. Thus the probability $p\left(S_{I} S_{2} S_{3} \mid f_{1} f_{2} f_{3}\right)$ that a sequence of three


Figure 1: a) Example of the integration or segregation of tones, either as 1 stream or 2 streams. b) Example of the condition [ $\begin{array}{lllll}3 & 1 & 9 & 1 & 3\end{array} 9$ ] from Exp. 2 (top) and the model's sequential maximum a posteriori assignment of tones within a stream (bottom). As each tone arrives the model reassigns the entire set of tones to streams (1->12->123 etc.).
tones was generated by sources $S_{1} S_{2} S_{3}$, has to be calculated over the five combinations: [ $\left.S_{l}=1, S_{2}=1, S_{3=1}\right]$, $\left[S_{l}=1, S_{2}=1\right.$, $\left.S_{3=2}\right],\left[S_{I=1}, S_{2}=2, S_{3=1}\right], \quad\left[S_{I=1}, S_{2}=2, S_{3=2}\right], \quad\left[S_{l=1}, S_{2=2}\right.$, $S_{3}=3$ ] corresponding to the five unique configurations of sources generating three sounds. Note that the first source is always assigned the value 1 , the next different source is assigned 2, etc.. Bayes' rule relates each conditional probability (the posterior distribution) to the likelihood $p\left(f_{1} f_{2} f_{3} \mid S_{1} S_{2} S_{3}\right)$ of each configuration of sound sources generating the sequence of tones, by

$$
\begin{equation*}
p\left(S_{1} S_{2} S_{3} \mid f_{1} f_{2} f_{3}\right)=p\left(f_{1} f_{2} f_{3} \mid S_{1} S_{2} S_{3}\right) p\left(S_{1} S_{2} S_{3}\right) / Z \tag{2}
\end{equation*}
$$

where $Z$ is a normalization constant, and $p\left(S_{1} S_{2} S_{3}\right)$ is the prior probability of the particular configuration of sound sources, regardless of the frequency, etc. of the tones

Assuming conditional independence of the tones and tone-source causality, this can be rewritten as

$$
\begin{align*}
& p\left(S_{1} S_{2} S_{3} \mid f_{1} f_{2} f_{3}\right)  \tag{3}\\
& =p\left(f_{3} \mid S_{1} S_{2} S_{3}\right) / p\left(f_{3}\right) \times p\left(S_{3} \mid S_{1} S_{2}\right) \times p\left(S_{1} S_{2} \mid f_{1} f_{2}\right)
\end{align*}
$$

The final term is the posterior generated from the first two tones. The latter two terms can be considered together as the prior for the third source, allowing us to use an iterative approach to the inference. After each tone we grow the tree of possible source sequence (e.g. $11 \rightarrow 111$ and 112), by multiplying the previous posterior $p\left(S_{1} S_{2} \mid f_{1} f_{2}\right)$ with two terms; the likelihood $p\left(f_{3} \mid S_{1} S_{2} S_{3}\right)$ and a prior for how likely the next 'branch' is, $p\left(S_{3} \mid S_{1} S_{2}\right)$.

We now consider how to determine the likelihood and prior probabilities. The first source can only be associated with one source, so $p\left(S_{l}=1\right)=1$. The principle of Occam's razor would suggest that $p\left(S_{I}=1, S_{2}=1\right)>p\left(S_{I}=1, S_{2}=2\right)$, i.e. if we haven't heard any of the sounds, the most probable acoustic scene is the simplest one: all sounds come from the same source. The value of $p\left(S_{1}=1, S_{2}=1\right)$ for an individual can be determined from fitting their data, and the value $p\left(S_{1}=1, S_{2}=2\right)$ is simply $1-p\left(S_{1}=1, S_{2}=1\right)$. The values may depend on factors such as the environment, which are not considered in the model: natural signal statistics may
provide guidance for how the prior probabilities are assigned. For successive sources, we use the probability given by a Chinese restaurant process (CRP) (Aldous, 1985), which can be considered as an extension of Occam's rule:

$$
\begin{aligned}
& p\left(S_{N}=\right.\left.i \mid S_{l} \ldots S_{N-I}\right)=n_{i} /(N-1+\alpha) \\
& \text { when } \mathrm{n}_{\mathrm{i}} \text { of the previous sources } S_{l} \ldots S_{N-I} \text { is equal to } i \\
& p\left(S_{N}=i \mid S_{l \ldots} . \ldots S_{N-I}\right)=\alpha /(N-1+\alpha) \\
& \text { when none of the previous sources is equal to } i
\end{aligned}
$$

where $N$ is the total number of sounds heard.
Regarding the likelihood function, the observer assumes the generative probability $p\left(f_{i} \mid S_{i}, f_{i-t,}, S_{i}=S_{i-t}\right)$. Note that this applies even when the sounds generated by the same source are separated by one or more sounds associated with different sources. The only transition that matters is that between the most recent tone and the last tone in the same stream, so if three tones $f_{1} f_{2}$ and $f_{3}$ had all been associated with the same stream, we would only consider the transition from $f_{2}$ to $f_{3}$, whereas if $f_{2}$ was associated with a different stream, we would only consider the transition from $f_{1}$ to $f_{3}$.

If a sound comes from a new source, then we assume that the likelihood is independent of previous tones:

$$
\begin{equation*}
p\left(f_{n} \mid S_{l}, \ldots S_{n-l}, \quad S_{n} \notin S_{l} \ldots S_{n-l}\right)=\frac{1}{\sqrt{2 \pi \sigma^{2}}} e^{-\frac{\left(f_{n}-\widehat{f}\right)^{2}}{2 \sigma^{2}}} \tag{5}
\end{equation*}
$$

where $\hat{f}$ is the midpoint of the range of auditory frequencies presented for the trial. The final model has two parameters, $\alpha$ and $\sigma$.

## Posterior approximation

Using the iterative scheme above we can calculate analytically the possible combinations of tones, but as the tone sequence progresses the number of possible source combinations - and hence the size of the posterior distribution - increases exponentially. To prevent combinatorial explosion two methods are used to generate an approximation of the full posterior distribution. The first limits the number of tones that are retained when using the previous posterior as the next prior, i.e. the algorithm only retains e.g. the last 10 tones and their potential allocations to sources

Limiting the number of tones eases the computational load, and can also be seen as a crude model of a limited memory capacity. Although the iteratively constructed prior retains some stream information of all previous tones, when a very short memory is used this may not be sufficient to generate stable stream allocation as the CRP prior probabilities fluctuate greatly when the number of previous tones is small. Furthermore, if the structure of the sequence is an important cue for streaming, a larger memory may be necessary to determine regularities in the sequence.

Even when the memory is limited to the previous six tones, allocating a stream to the seventh tone requires a posterior distribution taking 858 values, most of which must necessarily have very small probabilities. A second method to limit the size of the posterior is simply to select only the most probable stream combinations by imposing a probability threshold, hence we only used stream
combinations with $p>0.001$. Together these approximation methods allow a reasonable memory length of 10 tones (to avoid instability), while avoiding combinatorial explosion.

## Experiment 1

To compare the model to human performance we conducted a psychophysics experiment, in which six participants with normal hearing listened to simple auditory sequences and performed a subjective judgment task (a variant of experiments by van Noorden (1975)). Subjects were undergraduate students and received course credits for their participation. Each subject was fully briefed, provided informed consent and was given brief training on the task through exposure to 5 trial stimuli.

## Experimental setup

Figure 1a shows a schematic of the stimuli used - each sequence comprised 30 tones in repeated LHL- triplets, where the dash represents a silent gap. Each tone was 50 ms in duration, including 10 ms raised cosine onset and offset ramps. A $2 \times 2$ factorial design was used: the pitch of the high tones taking values of $3,6,9,12$ and 15 semitones above the low tone, which had a fixed frequency of 1000 Hz , and the offset to onset interval taking values $17,33,50$ and 67 ms . The duration of the silent gap was equal to the tone duration plus the offset-onset interval. Conditions were ordered randomly - each condition was tested 20 times over 5 runs, each run lasting approximately 7 minutes. Stimuli were presented through Sennheiser 280 headphones at a comfortable supra-threshold level. At the end of the sequence participants pressed a key to report whether the percept at the end of the sequence was most like a single stream (a galloping rhythm) or two separate streams of notes. The fraction of 2 -stream responses per condition is shown in figure $2 b$ for all six participants.

## Model response

To determine the response of the model to a tone sequence, the posterior for each possible sequence, $C$, is calculated tone-by-tone until all 30 tones have been presented. To relate the final posterior over sequences to subject responses, $s_{r}$ ('1 or 2 streams') $P_{\text {model }}\left(s_{r} \mid t o n e s, C\right.$ ), we defined a metric between two sequences. While the simple Hamming distance was considered we found it did not capture the similarities and differences between sequences. As an example, the Hamming distance between the sequence [11111] and [12222], $H(11111,12222)=4$, does not capture the intuition that a change of labels ( $2->1$ ) implies a distance of 1 . Instead we define a transition matrix, $M_{C}$ with elements $m_{i, j}=C_{i}-C_{j}$ i.e. the difference in the stream number for entry $i$ and $j$ of sequence $C$.

A transition matrix $M_{p C}$ is calculated for each posterior stream combination $C$, and also for the 'ideal' one or two stream response percepts (i.e. $M_{l}$ corresponding to 111 $111 \ldots$ and $M_{2}$ corresponding to $121121 \ldots$ ). The sum of the absolute difference between elements of $M_{p C}$ and both $M_{I}$ and $M_{2}, d_{C l}=\left|M_{p C}-M_{1}\right|$ and $d_{C 2}=\left|M_{p C}-M_{2}\right|$ give measures of the distances $d_{C 1}$ and $d_{C 2}$ from $C$ to the ideal response


Figure 2: a) Model prediction, based on fitted parameters from subject KC, giving the fraction of trials in which participant responded ' 2 ' for the number of streams perceived. Axes give the pitch difference for the middle tone and the inter stimulus interval (ISI): the time between the offset of one tone and the onset of the next. b) The results from 6 subjects.
percepts. This method can also give the fixed distance $\mathrm{d}_{12}$ between the ideal responses, $d_{12}=\left|M_{1^{-}}-M_{2}\right|$, thus streams $C$, $111111 \ldots$ and $121121 \ldots$ are represented by a triangle with sides of length $d_{C 1}, d_{C 2}$ and $d_{12}$. The vertex corresponding to stream $C$ can be projected onto the side $d_{12}$ giving $D_{l}$, the relative difference between $C$ and the two response percepts:

$$
D_{1}=\left(d_{12}^{2}+d_{C 1}^{2}-d_{C 2}^{2}\right) / 2 d_{12}^{2}
$$

$D_{l}$ is restricted to the range $[0,1]$, and each projected point is weighted by its posterior probability to give the marginal distribution of the posterior projected onto the axis joining the two responses. The distance $D_{l}$ gives the probability of subjects response, $s_{r}, 1$ or 2 , given $C$, i.e. $P\left(s_{r}\right.$ $=2 \mid C)=D_{l}$ and $P\left(s_{r}=1 \mid C\right)=1-P\left(s_{r}=2 \mid C\right)$. Lastly we marginalize over the possible sequences, and assume that participants draw a sample from the posterior when responding, giving

$$
P_{\text {model }}\left(s_{r} \mid \text { tones }\right)=\sum_{\mathrm{C}} P\left(s_{r} \mid \text { tones }, C\right) P(C \mid \text { tones })
$$

The parameters of the model (as well as for the alternative models below) were optimised using the MATLAB fminsearch routine to maximise the log-likelihood of the data, $\Sigma \ln \left(P_{\text {model }}\left(s_{r} \mid\right.\right.$ tones $\left.)\right)$ independently for each subject. During each iteration of the search, a sequence of 30 tones was presented to the model for each condition, and the probability of response ' 1 ' was calculated per condition.

## Model performance and comparison

The model was compared against three alternatives that used different priors to constrain the number of possible streams to two:
A. When the stream combination comprised only one stream (repeated), the prior probability of the next stream being 1 or 2 was allocated according to the CRP, but if the combination already contained two streams, the prior probability of allocating stream 1 or 2 was simply the fraction of previous tones that were allocated to stream 1 or 2 respectively.
B. The prior probabilities of a new tone being allocated to stream 1 or stream 2 was given by $P_{1}$, and $1-P_{1}$ respectively, where $P_{1}$ is a free parameter.
C. The prior probabilities of a new tone being allocated to stream 1 or 2 were fixed at 0.5 .
As mentioned earlier, an alternative response measure based on the Hamming distance was also tested: in this case we used the original, unconstrained CRP prior model. In the results, this is referred to as alternative D.

Because alternative model C has only one free parameter (all others have two), we use the Bayesian information criterion ( $\mathrm{BIC}=-2 \log P($ resp $\mid$ tones $)+k^{*} \log (n)$, where $k$ is the number of parameters and $n$ is the number of data points fitted over) to compare model performance in table 1. With the exception of participant LHH, the unconstrained model gives a better fit (smaller BIC) than all the alternatives considered. The mean $\pm$ SEM of the optimised parameters for the unconstrained model are $\alpha=0.81 \pm 0.12$ (equivalently $P(11)=0.56 \pm 0.04$ ) and $\sigma=105 \pm 7$ [semitones/sec]. Data from all subjects and the unconstrained model output for participant KC is shown in figure 2.

| Partici- <br> pant | Uncon- <br> strained | Alternative <br> A | B | C | D |
| :---: | :---: | :---: | :---: | :---: | :---: |
| SAG | $\mathbf{2 4 6 . 3}$ | 328.9 | 276.6 | 395.2 | 286.1 |
| TAY | $\mathbf{2 1 9 . 5}$ | 279.8 | 225.3 | 313.3 | 302.7 |
| KC | $\mathbf{2 8 7 . 4}$ | 345.4 | 314.1 | 414.1 | 489.9 |
| LHH | 288.5 | 335.9 | $\mathbf{2 5 7 . 8}$ | 424.1 | 284.5 |
| GM | $\mathbf{2 0 0 . 7}$ | 303.7 | 207.8 | 323.9 | 227.3 |
| MLP | $\mathbf{3 0 8 . 7}$ | 338.7 | 322.1 | 431.3 | 566.3 |

Table 1. BIC per participant. A smaller number indicates better relative performance (best model for each subject indicated in bold).


Figure 3: Averaged subject responses as a function of the auditory tone condition (see example in Fig. 1b). The horizontal labels indicate the tone-sequence of the condition, ordered by increasing step sizes. Error bars are standard errors.

## Experiment 2

While the model above theoretically allows an unlimited number of tones to be segregated into an unrestricted number of streams, the classical experiment (presented above) only allows a sequence of 3 tones to be separated into 1 or 2 streams. However, the model predicts that subjects should generally segregate based on frequency and temporal distances between tones. To test this further we performed a novel follow-up experiment where subjects were presented with seven tones and had to indicate the number of streams perceived. Nine conditions were created with sequentially larger discrepancy in frequency between tones and thus a larger probability of being assigned to different streams according to the model. The temporal gap between tones (ISI) were kept constant at 33.3 ms , unlike experiment 1 . For each condition, of the seven tones (see fig 1 b for one condition) three tones were unique. Five further subjects (see above) performed this new task. Results showed that subjects perceived an increasing number of streams (fig. 3), in accordance with predictions from the model, rising from 1 to close to 3 ( $\mathrm{p}<0.0001, \mathrm{~F}=20.39$, oneway anova, $\mathrm{df}=8$ ). None of the subjects perceived more than 3 streams for any of the conditions.

## Discussion

We have presented a simple Bayesian statistical model for grouping of discrete sequential stimuli. Utilizing a nonparametric Bayesian prior the model iteratively updates the posterior distribution over the assigned group of each stimuli and provides an excellent description of the perceptual interpretation of simple auditory sequences in human observers.

With just two parameters, the model gives a good account of the basic characteristics of auditory stream segregation the variation in the probability of perceiving a single sound source as a function of the repetition rate and pitch difference of the sounds. Although the ultimate goal is to characterize complex problems such as human speech
segregation, for experimental simplicity we tested a well known paradigm from auditory psychophysics. The proposed model gave a better fit to the data than alternative models that were constrained to interpret the sounds as being produced from just one or two streams. Predictions from the model were also in accordance with results from a novel experiment with larger number of tones (exp. 2).

Importantly the model goes beyond giving just the number of sources, but says which sounds are produced by each source. While the combinatorial space of the posterior distribution in experiment 1 was collapsed to give a marginal distribution in a continuous 1-d response space (leading to an estimate of response probability), the maximum a posterior (MAP) for all participants was always located at either 111-111... or 121-121..., depending on the stimulus condition (figure 2 b ). This is reassuring as it is consistent with the anecdotal evidence that participants always perceive either a galloping rhythm (streams 111111...) or a high-pitch and a low pitch stream (121-121...), i.e. the percept is always at the MAP. Indeed, the percept cannot in general be at the mean because the space of possible percepts is discrete: there is no percept between, say, 111 and 121.

One consequence of the inference model that is not addressed by mechanistic models of stream segregation is that when a percept changes from say 111-111 to 121-121, the source allocation of previous sounds is changed. Ironically, this 'non-causal' effect is essentially a feature of causal inference - when an observer decides that the percept has changed to 121-121, this is based on previous evidence, and yet at the time that the previous tones were heard, they were all associated with one source. A similar effect is commonly encountered when mis-interpreted speech (perhaps mis-heard due to background noise) suddenly makes sense when an essential word is heard - the previous words are reinterpreted, similar to the letters in predictive text message systems.

The framework of the model is very general, and allows for the incorporation of other factors into the likelihood to describe other aspects of auditory stream segregation. Adding terms in the likelihood function may be able to explain other effects seen in the literature, such as segregation based on bandwidth (Cusack \& Roberts, 2000), or build-up and resetting of segregation (Roberts, Glasberg, \& Moore, 2008). Furthermore, in the current study we assume that there is no ambiguity in the percept of the pure tones, the uncertainty arises from lack of knowledge about the underlying generative structure of the data. In a realistic situation perceptual ambiguity would have to be taken into account using an approach such as suggested by Turner and Sahani (Turner \& Sahani, 2011). Nevertheless, we should emphasize that even though we are dealing with a Markov property (each tone within a stream only depends on the previous tone), the mixture of streams makes the problem very different from work on e.g. Hidden Markov Models (or even Infinite Hidden Markov Models) for which the goal would be to infer underlying states despite perceptual ambiguity. Note also that while there are algorithms
developed to separate audio signals (e.g. Roweis, 2001), these are not meant to mimic human perception, although a future comparison would be very interesting.

In the current implementation we used numerical approximations in order to handle the complexity of the model. As an alternative to calculating our results analytically we could use Monte Carlo techniques (e.g. Markov Chain Monte Carlo sampling, a different type of approximation), which have become a standard tool for solving complex statistical models.

The proposed model of auditory stream segregation is a specific instantiation of an iterative probabilistic approach towards inference of perceptual information. A major issue for this approach is the problem of dealing with multiple sources, as represented by the work done on causal inference (Shams \& Beierholm, 2010). Until now models of causal inference have been unable to handle more than two sources, due to the escalating number of parameters needed for parametric priors. The use of a non-parametric prior allows a complex of many stimuli to be interpreted without running into this problem, potentially allowing for an arbitrary number of causes in the world. This approach is very general - it can be applied to any set of discrete sequential cues involving multiple sources - and it gives a simple, principled way to incorporate natural signal constraints into the generative model.

## References

Aldous, D. J. (1985). Exchangeability and related topics. Lecture Notes in Mathematics, 1117, 1-198.
Anstis, S. M., \& Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. Journal of Experimental Psychology: Human Perception and Performance,, 11(3), 257-271.
Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. Sensory Communication, 217 - 234.
Beauvois, M. W., \& Meddis, R. (1997). Time decay of auditory stream biasing. Perception \& Psychophysics, 59(1), 81-6.
Bendixen, A. (2014). Predictability effects in auditory scene analysis: A review. Front. in Neuroscience, 8, 60, 1-16
Bregman, A. S. (1994). Auditory Scene Analysis: The Perceptual Organization of Sound (Bradford Book). MIT Press.
Bregman, A. S., \& Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. Journal of Experimental Psychology, 89(2), 244-9.
Froyen, V., Feldman, J., \& Singh, M. (2015). Bayesian Hierarchical Grouping: Perceptual Grouping as Mixture Estimation. Psychological Review, 122(4), 575-597.
Cusack, R., \& Roberts, B. (2000). Effects of differences in timbre on sequential grouping. Perception \& Psychophysics, 62, 1112-1120.
Hedges, J. H., Stocker, A. A., \& Simoncelli, E. P. (2011). Optimal inference explains the perceptual coherence of visual motion stimuli. Journal of Vision, 11, 1-16.

Knill, D. C. (2007). Learning Bayesian priors for depth perception. J Vis, 7(8), 13.
Knill, D. C., \& Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci, 27, 712-719.
Lee, M. D., \& Habibi, A. (2009). A Cyclic Sequential Sampling Model of Bistable Auditory Perception. In Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 2669-2674).
McCabe, S. L., \& Denham, M. J. (1997). A model of auditory streaming. The Journal of the Acoustical Society of America, 101(3), 1611.
McDermott, J. H., \& Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron, 71(5), 926-40.
Moore, B. C. J., \& Gockel, H. E. (2012). Properties of auditory stream formation. Philos Trans $R$ Soc Lond B Biol Sci, 367(1591), 919-31.
Moore, B., \& Gockel, H. (2002). Factors Influencing Sequential Stream Segregation. Acta Acustica, 88(3), 320-333.
Olshausen, B. A., \& Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381(6583), 607-9.
Orbanz, P., \& Teh, Y. W. (2010). Bayesian Nonpar. Models. In Encyclopedia of Machine Learning. Springer.
Roberts, B., Glasberg, B. R., \& Moore, B. C. J. (2008). Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. Journal of Experimental Psychology. Human Perception and Performance, 34(4), 992-1006.
Roweis, S. T. (2001). One Microphone Source Separation. Advances in Neural Information Processing Systems, 13,793-799.
Shams, L., \& Beierholm, U. R. (2010). Causal inference in perception. Trends in Cognitive Sciences, 1-8
Shams, L., Ma, W. J. W. J., \& Beierholm, U. (2005). Soundinduced flash illusion as an optimal percept. Neuroreport, 16(17), 1923-7.
Turner, R. E. (2010). Statistical models for natural sounds. PhD Thesis. University College London.
Turner, R. E., \& Sahani, M. (2011). Probabilistic amplitude and frequency demodulation, Advances in Neural Information Processing (pp. 981-989). New York: Red Hook.
Van Noorden, L. (1975). Temporal coherence in the perception of tone sequences. PhD Thesis. University of Technology, Eindhoven, The Netherlands.
Wang, D. (1996). Primitive auditory segregation based on oscillatory correlation. Cogn Science, 20(3), 409-456.
Weiss, Y., Simoncelli, E. P., \& Adelson, E. H. (2002). Motion illusions as optimal percepts. Nat Neurosci, 5, 598-604.
Wood, F., Goldwater, S., \& Black, M. J. (2006). A nonparametric Bayesian approach to spike sorting. International Conference of the IEEE Engineering in Medicine and Biology Society, 1, 1165-8.

# Perceived similarity mediates violations of independence in probabilistic judgments 

James M. Yearsley (james.m.yearsley@ vanderbilt.edu)<br>Daier Yuan (daier.yuan@ vanderbilt.edu)<br>Jennifer S. Trueblood (jennifer.s.trueblood@ vanderbilt.edu)<br>Department of Psychology, Vanderbilt University, Nashville TN, USA


#### Abstract

We outline a simple way of representing sets of non-normative judgements that makes them look as similar as possible to normative ones. This representation allows us to view certain types of non-normative judgments, such as conjunction fallacies, as arising from a misestimation of the correlation between events, that might arise when decision-makers have no prior information about the frequency of co-occurrence. We suggest that decision-makers use the perceived similarity between events to make inferences about correlation, and we describe the results of an experiment showing that judged correlation and violations of independence in probabilistic judgments are strongly influenced by the perceived similarity between events.


Keywords: Conjunction fallacy, similarity, normative reasoning

## Introduction

Being able to make unbiased and reasonably accurate likelihood judgments about simple events is a foundation for more complicated tasks such as inference or causal reasoning. Human reasoners are often very competent at providing such judgments, in the sense that their judgments align well with normative prescription. However there are classic results, such as those associated with the famous Tversky and Kahneman research program, which show that in some cases human reasoners may provide likelihood judgments for simple combinations of events that show systematic biases. In causal reasoning tasks for example, this can lead to violations of the predictions of models based on classical probability theory, with a resulting need to supplement classical models with extra unobserved relationships (Rehder, 2014), or even to reject classical probability theory entirely and attempt to construct models based on other theories of probability, such as quantum probability theory (Pothos \& Busemeyer, 2013).

Several decades of experience have taught researchers where to expect violations of normative rules when making likelihood judgments, but there has been less success in determining why such violations occur. Tversky and Kahneman (1983) argued that conjunction fallacies occur because of a representativeness heuristic, while explanations for violations of normative rules such as the Markov Condition in causal reasoning often involve the presence of additional enabling or disabling causes (Rehder, 2014). Meanwhile models based on non-classical probability theory posit that violations of classical probability rules occur because of a somewhat mysterious property known as 'incompatiblity' (Busemeyer \& Bruza, 2012). Perhaps one or more of these explanations is correct, perhaps none are, but regardless they all
suffer from the same problem that it is difficult to predict in advance whether a given set of events will be 'representative' or 'incompatible' etc.

An additional problem faced when attempting to understand why some judgments are normative and some are not is that different frameworks are used to model the different types of judgments. Partly this is due to the fact that non-normative judgements are defined by what they are not, viz. those that can be explained by an underlying classical (Bayesian) belief state. One way of modeling non-normative judgments is via heuristics (see e.g., Gigerenzer et al, 2015) which may bear no relationship to classical probability computations. Another is to replace classical sample spaces with quantum vector (Hilbert) spaces, as done in quantum models of cognition (Busemeyer \& Bruza, 2012), which again appear to have little relation to classical probability theory.

There would appear, therefore, to be a disconnect between the way cognitive states, and computations on them, are represented depending on whether one is dealing with normative or non-normative reasoning. This poses a challenge if we wish to understand the reasons why we might sometimes give normative judgments and sometimes not, or, for example, if we wish to understand how corrective feedback may improve performance.

What we want to do in this contribution is to introduce a way of thinking about non-normative judgments that makes them look as similar as possible to normative ones. This representation can be used in a variety of settings to study transitions between non-normative and normative behaviors. We will show that for non-normative judgments the notion that an underlying probability distribution 'does not exist' can be formalized by considering quasi-distributions, which are similar to standard probability distributions except that some elements may be negative. This gives us a way to think about smoothly transitioning from non-normative judgments, represented by quasi-distributions, to normative ones, where all elements of the distribution are positive and it may be interpreted as a classical probability distribution.

By itself this achieves little beyond expressing the problem of non-normative judgments in a different language, however we will argue that this representation provides a new way to understand the origin of non-normative judgments, and even a way of visualizing how learning can cause a transition to normative behavior. We will see that non-normative judgments can arise because of an misestimation of the correlation be-
tween events, and we will argue that this may occur when events are perceived to be highly similar. We test this in an experiment, looking at judgments about the joint occurrence of events with different degrees of similarity.

The rest of this contribution is structured as follows, first we provide a brief introduction to quasi-probability distributions, our aim being to show how they may be used to encode non-normative judgments. Next, we use this framework to develop a novel empirical prediction that a high degree of perceived similarity between features of an object can give rise to violations of independence and conjunction fallacies in probabilistic judgments. Then, we describe an experiment to test this prediction. We conclude with some possible avenues for further study.

## Probability distributions and quasi-distributions

A common theme in experiments on probabilistic judgment is that participants are asked to make a set of judgments about the likelihood of some events, e.g. $p(A), p(B), p(A \cap C)$ etc, and the normative status of these judgments is assessed by proving that there either does or does not exist a probability distribution $p\left(A_{i}, B_{j}, C_{k}\right)$ such that all the measured judgments can be thought of as marginals of this joint distribution. For example, the conjunction fallacy, wherein participants judge $p(A \cap B)>p(A)$, is non-normative because it is impossible for participants to have a single joint probability distribution for $A$ and $B$ with this property.

## Joint probability distributions

Establishing whether a set of probability or likelihood judgments is normative is therefore equivalent to the following:

Given some set of probabilities, $S=$ $\{p(A), \ldots p(A, B), \ldots, p(A, B, C), \ldots\}$ etc does there exist a joint distribution $p\left(A_{i}, B_{j}, C_{k} \ldots\right)$ of which all elements of the set $\mathcal{S}$ may be considered as marginals?

If such a probability distribution exists, then the set of judgments $\mathcal{S}$ are normative, otherwise they are non-normative.

## Some Examples

- This definition includes trivial cases, e.g. where $\sum_{i} p\left(A_{i}\right) \neq$ 1. Such cases are obviously non-normative.
- A simple example is provided by the set $S=$ $\left\{p\left(A_{i}\right), p\left(B_{j}\right), p\left(C_{k}\right), \ldots\right\}$ where participants are only asked to make judgments about a single event. In this case we can easily find a joint distribution that has the single event probabilities as marginals, e.g. $p\left(A_{i}, B_{j}, C_{k}, \ldots\right)=$ $p\left(A_{i}\right) p\left(B_{j}\right) p\left(C_{k}\right) \ldots$ will work (there are many choices).

A Less Trivial Example Suppose we have three binary events $A, B, C$ and we are given the joint probabilities, $\mathcal{S}=$ $\left\{p\left(A_{i}, B_{j}\right), p\left(B_{j}, C_{k}\right), p\left(A_{i}, C_{k}\right)\right\}$. An important result is that it is not always possible to find a joint distribution $p\left(A_{i}, B_{j}, C_{k}\right)$ with these marginals. The conditions under which this is possible are when the Bell inequalities are satisfied (Fine, 1982).

## Joint quasi-probability distributions

If a set of judgments $S$ is not normative, then a probability distribution capturing these judgments does not exist. However there may still exist some function $q\left(A_{i}, B_{j}, C_{k} \ldots\right)$ such that all the elements in the set $S$ can be obtained by summing out the other variables in $q(\ldots)$. This function $q\left(A_{i}, B_{j}, C_{k} \ldots\right)$ will generally fail to be a probability distribution because it will not be non-negative.
Example Suppose $\mathcal{S}=\{p(A), p(B), p(A \cap B)\}$ for some binary valued features $A, B$ and $p(A \cap B)>p(A)$. Clearly there is no probability distribution which can have $\mathcal{S}$ as its marginals. However a quasi-distribution with these properties may be given as:

$$
\begin{aligned}
& q(A, B)=p(A \cap B) \\
& q(A, \bar{B})=p(A)-p(A \cap B) \\
& q(\bar{A}, B)=p(B)-p(A \cap B) \\
& q(\bar{A}, \bar{B})=1-p(A)-p(B)+p(A \cap B)
\end{aligned}
$$

Note that this has the desired marginals, e.g. $q(\bar{A}, B)+$ $q(\bar{A}, \bar{B})=1-p(A)$, but that $q(A, \bar{B})<0$.

The use of quasi-distributions in psychology to understand inconsistent judgments has been advocated before, most notably by de Barros, (e.g. de Barros, 2013). In physics there is a long history of trying to apply 'extended' probabilities to understand aspects of quantum theory (see e.g. Muckenheim, 1986). Their interpretation can be challenging (Halliwell \& Yearsley, 2013) but here we shall avoid assigning any meaning to them and regard them simply as a computational tool.

So far all we have done is to express some classes of nonnormative judgments in terms an object which is superficially similar to a joint probability distribution, but which fails to be one in some (rather drastic) way. Why is this useful? Well the usefulness of quasi-probability distributions lies in part in the fact that they smoothly capture the idea of transitioning between non-normative behavior (where one or more of the elements of the distribution is negative) to normative behavior, where all elements are non-negative. Suppose for example that we are performing an experiment where participants have to bet on the outcome of some gamble involving a conjunction. If they commit a conjunction fallacy in their reasoning, they may initially perform badly, but with corrective feedback they may revise their estimate of the probabilities of the outcomes. At some point their beliefs will change from non-normative to normative discontinuously, but in terms of quasi-distributions their their belief state may change smoothly as they learn.

Quasi-distributions also allow us to define a notion of distance from normative behavior, for example one could define the degree of non-normativity as,

$$
\begin{equation*}
\Delta=\sum_{i, j, k \ldots}\left|q\left(A_{i}, B_{j}, C_{k} \ldots\right)\right|-1 \tag{1}
\end{equation*}
$$

which is zero if $q(\ldots)$ is a probability distribution and nonzero otherwise. For example, for the case above of a conjunction fallacy, $\Delta=2|p(A, B)-p(A)|$, which is an appealing measure of the non-normativity.

One reasonable proposal would be to look at cases where a set of judgements $\mathcal{S}$ only just fails to be normative by this measure. One could then try to define a genuine probability distribution $p(\ldots)$ and a new set of judgments $S^{\prime}$ which are 'close' to the real judgments $\mathcal{S}$ in the sense that $p(\ldots)$ is close to $q(\ldots)$. If there is a sense in which this is possible then one might regard the judgments $\mathcal{S}$ as almost normative, and perhaps attribute the discrepancy to some sort of response noise.

We will not pursue this further here. Instead we will make use of another advantage of quasi-distributions, which is that by expressing normative and non-normative behaviors in a similar language, they suggest ways to understand how nonnormative behaviors may come about. We shall explore one such idea in the next section.

## A Simple Proposal

In a typical conjunction fallacy type experiment, participants might be expected to have some information about the rate of occurrence of the features $A$ and $B$, and be asked to guess the likelihood of the conjunction $p(A \cap B)$. It is important to realize that even given $p(A), p(B)$ there is no 'correct' answer to this question. Rather there are a range of possible allowed values, in other words, the marginal probabilities $p(A), p(B)$ under-specify the joint distribution. One extra piece of information is needed, one of the joint probabilities would do, as would some linear combination of these. One possibility is to consider the quantity,

$$
\begin{equation*}
S_{A B}=p(A \cap B)+p(\bar{A} \cap \bar{B}) \tag{2}
\end{equation*}
$$

which is closely related to the correlation. The difference between a normative probability distribution and a nonnormative quasi-distribution can be thought of, perhaps simplistically, as the difference between choosing a value for $S_{A B}$ within or outside of the allowed range. This is important because decision makers armed only with $p(A), p(B)$ have no information about $S_{A B}$, and there is therefore a significant possibility that they may chose incorrectly. To put it another way, if decision makers make an incorrect guess for the correlation between $A$ and $B$, this can lead to a non-normative set of judgments.

Now $S_{A B}$ is a number which varies between 1 if the events always happen together, to 0 if the presence of one event implies the absence of the other and vice versa. One possibility is that decision makers simply pick a value for $S_{A B}$ based on a uniform prior. Another possibility is that decision makers equate $S_{A B}$ with a more primitive quantity such as the similarity between $A$ and $B$. (We note in passing that the idea of similarity as essentially joint probability appears in accounts of similarity judgment based on quantum cognitive models (Pothos et al, 2015).)

This leads to an important prediction, which we will test below: In the absence of any information about joint occurrence, human decision makers will use features of events such as their similarity to construct a joint distribution. Manipulating these relationships in an experimental setting should lead
to greater or lesser degrees of violations of independence for these events, and more generally to changes in the correlation between events. We describe an experiment to test these ideas in the next section.

## An Experiment

## Methods

58 undergraduate students from Vanderbilt University participated in the experiment online at a time of their choosing for course credit. Participants answered questions about three different novel categories, an animal, a natural object and a human made object, adapted from previous work on causal reasoning (Rehder, 2014). Each object had three binary features $(A, B$, and $C)$. For each feature participants were told that 'most' members of the category had a high value for that feature, while 'a few' members of the category had a low value for the feature. Participants were not told about any relationships between the features. For example, in the Kehoe Ant category, $A=$ Blood iron level (high or low amount), $B=$ Immune system activation level (hyperactive or suppressed), and $C=$ Blood thickness (thick or thin).

After this, participants answered a number of questions where they were told that a new member of the category had been discovered, and were asked to indicate how likely they thought it was that the new object had various features. There were three question types: (1) how likely it was that the object had a particular feature, e.g. blood high in iron sulphate, (2) how likely it was that the object had a combination of features, e.g. blood high in iron sulphate and an immune system that is hyperactive, and (3) a conditional, e.g. a hyperactive immune system given that a previous test had established a high level of iron sulphate in the blood. Participants were asked about all possible conjunctions of events, but only about conditionals where one feature was conditioned on the presence of a low value for another feature. The reason for this was to reduce the overall number of questions in the experiment, particularly since our expectation was that features would be positively correlated, which would be likely to lead to floor or ceiling effects for the other possible conditionals.

The responses were either requested as whole numbers between 0 and 100 , or as points on a 9 point Likert scale. The response format for all questions concerning a given category were the same, and participants were randomly assigned either the whole number or the Likert response options for each category. After completing the likelihood judgment questions for each category, participants were asked to rate the similarity between the feature types on a 7 point Likert scale. The order in which the features appeared in the similarity question (e.g. how similar is feature 1 to feature 2 ) was randomized between participants for every judgment, however there were no significant order effects in the similarity judgments.

After finishing the main part of the experiment, participants completed an extended version of the Cognitive Reflection Test (CRT, Frederick, 2005), but there was no significant effect of CRT and it will not be discussed further here.

## Results

We first wanted to examine whether we had conjunction fallacies and violations of independence in this data set. For each pair of likelihood judgments, e.g. $\{A, A \cap B\}$ or $\{A, A \mid B\}$ we can perform a paired samples t-test to assess the presence of these effects, however this procedure would generate a substantial volume of test statistics without giving much insight. Instead we will plot the relevant likelihoods, and quote some representative statistics. In this contribution, we only report Bayesian statistical tests that were performed using JASP (JASP team, 2016). In particular we report Bayes factors for the alternative versus the null hypothesis, so that values $>1$ indicate evidence for the alternative hypothesis.

We begin by assessing the conjunctions. For each pair $\{A, A \cap B\}$ we plotted the average values across participants of the single event and the conjunction. It is useful to split these pairs up into four different types, depending on whether each of the events has high or low individual probability. The results are shown in Fig 1, and we have separated out data that comes from responses using whole numbers and from the Likert scale. Points that lie above the diagonal correspond to conjunction fallacies. The first thing to note is that there are three obvious clusterings of data points. We see straight away that pairs of the form $\{A, A \cap \bar{B}\}$ behave as expected, there are no conjunction fallacies. Equally the pairs $\{A, A \cap B\}$ do not display robust conjunction fallacies (All Bayes factors for t-tests $<1$ ), although these data points are slightly odd in another way, which we will return to shortly. The pairs which do display conjunction fallacies are $\{\bar{A}, \bar{A} \cap B\}$ and $\{\bar{A}, \bar{A} \cap \bar{B}\}$.

The pairs of the form $\{\bar{A}, \bar{A} \cap B\}$ display robust conjunction fallacies. For the whole number responses 14 out of 18 of the Bayes factors for t -tests are $>3$, and 10 out of 18 are $>10$. For the Likert responses 16 out of 18 of the Bayes factors for t-tests are $>3$, and 11 out of 18 are $>10$. These are the pairs for which conjunction fallacies are typically expected, with one likely and one unlikely event. In contrast the presence of conjunction fallacies in the pairs of the form $\{\bar{A}, \bar{A} \cap \bar{B}\}$ are less expected. For the whole number responses none of the Bayes factors for t-tests are $>3$, but for the Likert responses 7 out of 18 of the Bayes factors for t -tests are $>3$, and 5 out of 18 are $>10$. We will return to why this may be so later.

Overall then, we have good evidence for conjunction fallacies in some of these judgments, for both response types. Note also that there do not appear to be large systematic differences between the data obtained from different response modes, which is reassuring.

Next we check for violations of independence. Note that these violations are not necessarily non-normative, since no information about the relationship between features was given to participants. However systematic violations of independence would still be a surprising finding. We proceed as for the conjunctions, plotting the pairs $\{A, A \mid \bar{B}\}$ and $\{\bar{A}, \bar{A} \mid \bar{B}\}$ separately and also separating out whole number and Likert responses. The results are shown in Fig 2. Independence would be indicated by data points lying on the diagonal.


Figure 1: Plots of likelihood judgments for conjunctions against single constituent events. Data points above the diagonal indicate conjunction fallacies. a) Likert scale responses. b) Whole number responses.

For the pairs of the form $\{A, A \mid \bar{B}\}$ all points appear to lie below the diagonal, and this is confirmed by Bayesian t-tests. For the whole number responses and for the Likert responses all Bayes Factors are $>10$.

For the pairs of the form $\{\bar{A}, \bar{A} \mid \bar{B}\}$ all points appear to lie above the diagonal, and this is confirmed by Bayesian t-tests. For the whole number responses 13 out of 18 Bayes Factors are $>3$, and 7 out of 18 are $>10$. For the Likert responses we also have 13 out of 18 Bayes Factors $>3$, and 7 out of 18 $>10$. Again overall there is good evidence for violations of independence in this data.

More specifically, the conjunctions and the conditionals point to similar behavior - namely participants appear to believe that there is strong correlation between the different features, such that "high" or "low" values of these features are likely to occur together.

Now we turn to the question of whether the perceived similarity between features mediates the correlations and violations of independence.

We begin with the correlation, defined for each pair of features $A, B$ as $p(A \cap B)+p(\bar{A} \cap \bar{B})-p(A \cap \bar{B})-p(\bar{A} \cap B)$. We ran a Bayesian ANOVA with the perceived similarity as the


Figure 2: Plots of likelihood judgments for conditionals against the conditioned events. Data points on the diagonal indicate independence. Data points off the diagonal indicate violations of independence. a) Likert scale responses. b) Whole number responses.
independent variable. Note that we collapse across features and scenarios here, since we have similarity data for each individual feature pair.

For the whole number responses the results of the Bayesian ANOVA show that a model including perceived similarity is preferred over the null model $\left(\mathrm{BF}_{10}>10^{7}\right)$. For the Likert responses the results of the Bayesian ANOVA also show that a model including perceived similarity is preferred over the null model $\left(\mathrm{BF}_{10}>10^{3}\right)$. Analysis of effects for both ANOVAs are given in Table 1.

We plot in Fig 3a below the correlation as a function of the similarity. Since we saw in the analysis of the conjunctions and conditionals there were no obvious differences between the response types we have converted the Likert scale responses to numbers in the range $0-100$ and plotted them on the same axis. This lets us establish that the same qualitative pattern holds for both response types, namely there is some apparent decrease in correlation for very small similarity ratings, but then a robust increase in correlation with increasing perceived similarity. The reason for the decrease in correlation for small similarity ratings is unclear, although it is worth noting that the number of participants who gave similarity ratings from 1-3 is small (7-11) compared with the number of
participants who gave higher similarity ratings (40-66). In addition, for the lowest similarity rating a higher than expected proportion of participants giving this rating ( 5 out of 8 for the Likert scale and 6 out of 9 for the whole numbers) had the highest possible CRT score. This is significant because if two events, each of which has an individual probability of 0.8 , are independent, then the expected correlation is 0.32 , which is in fact close to the observed value for a similarity rating of 1 in the whole numbers condition.

Next we analyze the violations of independence. We compute a violation 'score' which is just the sum of the absolute value of the difference between the conditional and the single event, e.g. $|p(A \mid B)-p(A)|$ for all the conditionals we measured. Again we ran a Bayesian ANOVA, with perceived similarity as the independent variable.

For the whole number responses, the results of the Bayesian ANOVA show that a model including perceived similarity is preferred over the null model ( $\mathrm{BF}_{10} \sim$ " $\infty$ "). For the Likert responses the results of the Bayesian ANOVA show that a model including perceived similarity is again preferred over the null model $\left(\mathrm{BF}_{10}>10^{10}\right)$. Analysis of effects for both ANOVAs are given in Table 2.

In Fig 3b we plot the violation of independence score as a function of perceived similarity. Again we transform the values for the Likert scale responses allowing us to plot them on the same axes. The pattern is qualitatively similar to that for the correlation; a general trend towards larger violations of independence for higher values of the perceived similarity.

Table 1: Analysis of effects for Bayesian ANOVA of Correlation

| Whole Numbers |  |  |  |
| :--- | :--- | :--- | :--- |
| Effect | $p($ incl $)$ | $p($ incl $\mid$ data $)$ | $\mathrm{BF}_{\text {Inclusion }}$ |
| Similarity | 0.500 | 1.000 | $3.01 \times 10^{7}$ |
|  |  |  |  |
| Likert Scale |  |  |  |
| Effect | $p($ incl $)$ | $p($ incl ldata $)$ | $\mathrm{BF}_{\text {Inclusion }}$ |
| Similarity | 0.500 | 0.999 | $1.58 \times 10^{3}$ |

Table 2: Analysis of effects for Bayesian ANOVA of Violations of Independence

| Whole Numbers |  |  |  |
| :---: | :---: | :---: | :---: |
| Effect | $p$ (incl) | $p$ (incl ${ }_{\text {data }}$ ) | $\mathrm{BF}_{\text {Inclusion }}$ |
| Similarity | 0.500 | 1.000 | " ${ }^{\prime \prime}$ |
| Likert Scale |  |  |  |
| Effect | $p$ (incl) | $p$ (incl data) | $\mathrm{BF}_{\text {Inclusion }}$ |
| Similarity | 0.500 | 1.000 | $3.08 \times 10^{10}$ |

Overall the data provide strong support for our proposal that perceived similarity mediates perceptions of correlation and violations of independence in probabilistic judgments. It is also worth noting that there are strong positive correlations between the correlation function and violations of independence, (Pearson's rho $=0.481, \mathrm{BF}_{10}>10^{14}$ for whole number responses, Kendall's tau $=0.345, \mathrm{BF}_{10}>10^{13}$ for the Likert responses.)


Figure 3: Plots of the correlation and degree of violation of independence against the perceived similarity. a) Correlation as a function of perceived similarity for the whole number responses (blue line) and the Likert scale responses (red line.) b) Violations of independence as a function of perceived similarity for the whole number responses (blue line) and the Likert scale responses (red line.)

## Conclusions and Future Directions

We have shown that quasi-probability distributions can be used to encode certain sets of probabilistic judgments which are non-normative, in the sense that they cannot be regarded as marginals of a joint probability distribution. Quasidistributions generalize regular probability distributions in that they can have negative elements. By themselves these do not provide any great insight into non-normative behavior, but the fact that one can define analogues of properties such as correlations for quasi-distibutions lets us examine the ways in which sets of judgments fail to be normative, and perhaps suggest some possible reasons why. We proposed that in the absence of information about joint occurrence, human decision makers might use properties such as the similarity between features to set the correlation, which we demonstrated experimentally. Similarity does seem to mediate correlation and violations of independence. We also showed that these results are largely independent of the response format.

These findings are particularly significant for attempts to assess the normative status of human causal inference using stimuli of this nature (e.g. Rehder, 2014). In these experi-
ments, participants are given extra information about the features in the form of causal relationships between them. Judgments about correlations in this case could then reasonably be interpreted as meaning participants believe the presence of one feature caused another. This work suggests that care is needed when interpreting these studies - participants may believe that features are correlated even in the absence of causal relationships, which may lead to overestimation of perceived causality. Future work should explore this possibility.

Finally, the results of this study suggest quasi-distributions may be a valuable way of thinking about non-normative reasoning, and we are hopeful that this approach may be used fruitfully in other areas. One important task is the development of a learning model which works directly with quasidistributions. This could help us understand how people learn to avoid committing probabilistic fallacies (Nilsson, 2008).

## Acknowledgments

JMY and JST were supported by NSF grant SES-1556415.

## References

de Barros, JA (2013). Decision making for inconsistent expert judgments using negative probabilities. In, Atmanspacher, H et al (Eds.) Proceedings of the Sixth Quantum Interaction Symposium. LNCS, 8369, 257269.

Busemeyer, JR \& Bruza, P (2011). Quantum models of cognition and decision. CUP: Cambridge, UK.
Fine, A. (1982). Joint distributions, quantum correlations, and commuting observables. Journal of Mathematical Physics, 23, 1306-1310.
Frederick, S (2005). Cognitive reflection and decision making. Journal of Economic Perspectives. 19, 25-42.
Gigerenzer, G, Hertwig, R \& Pachur, T (eds.) (2015). Heuristics: The foundations of adaptive behavior. (OUP).
Halliwell, JJ \& Yearsley, JM. (2013). Negative probabilities, Fine's theorem and linear positivity. Phys. Rev. A 87, 022114.

JASP Team. (2016). Jasp. https://jasp-stats.org
Muckenheim, G. (1986). A review of extended probabilities. Phys. Rep. 113(6), 337-401.
Nilsson, H. (2008). Exploring the conjunction fallacy within a category learning framework. Journal of Behavioral Decision Making. 21, 471-490.
Pothos, E. M. \& Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? Behavioral \& Brain Sciences, 36, 255-327.
Pothos, EM, Barque-Duran, A, Yearsley, JM, Trueblood, JS, Busemeyer, J, \& Hampton, JA. (2015). Progress and current challenges with the quantum similarity model. Frontiers in Psychology, 6, 205.
Rehder, B (2014). Independence and dependence in human causal reasoning. Cognitive Psychology, 72, 54-107.
Tversky, A., \& Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjuctive fallacy in probability judgment. Psychological Review, 90, 293-315.

# Causal and compositional generative models in online perception 

Ilker Yildirim<br>MIT<br>Michael Janner<br>MIT<br>Mario Belledonne<br>MIT<br>Christian Wallraven<br>Korea University<br>\section*{Winrich Freiwald}<br>Rockefeller University<br>\section*{Josh Tenenbaum}<br>MIT


#### Abstract

From a quick glance or the touch of an object, our brains map sensory signals to scenes composed of rich and detailed shapes and surfaces. Unlike the standard approaches to perception, we argue that this mapping draws on internal causal and compositional models of the physical world and these internal models underlie the generalization capacity of human perception. Here, we present a generative model of visual and multisensory perception in which the latent variables encode intrinsic (e.g., shape) and extrinsic (e.g., occlusion) object properties. Latent variables are inputs to causal models that output sense-specific signals. We present a recognition network that performs efficient inference in the generative model, computing at a speed similar to online perception. We show that our model, but not alternatives, can account for human performance in an occluded face matching task and in a visual-to-haptic face matching task.


# Individual Differences in Gaze Dynamics in Risky Decision-making 

Siyuan Yin (siyuan.yin@ vanderbilt.edu)<br>Jennifer S. Trueblood (jennifer.s.trueblood@ vanderbilt.edu)<br>Department of Psychology, Vanderbilt University, $11121^{\text {st }}$ Ave S<br>Nashville, TN 37240 USA


#### Abstract

In risky decision-making, expected utility (EU) theory is widely used to examine people's risk attitude and choice behavior. However, it is unknown how risk attitude relates to attention and information search. In this paper, we explore the relationship between risk attitude (as measured by a variant of EU) and eye movement patterns (which serve as a proxy for attention and information search). Participants made choices between gambles presented perceptually as flickering grids in which monetary values were indicated by colors and probabilities by color proportions. To explore attention and information search patterns, we investigated eye movement patterns when faced with different gambles and correlated these patterns with the parameters of EU. We observed that people who are more risk-seeking (as determined by modeling) tend to look at risky options more often. These results bridge choice behaviors conceptualized by EU and information search strategies under risky decision-making revealed by eye movements.


Keywords: risky decision-making; eye movements; cumulative prospect theory; hierarchical Bayesian parameter estimation; individual differences

## Introduction

We face decision-making under risk every day in our lives, from financial investment decisions, choosing a new job, to voting for a presidential candidate. Lotteries, or gambles, consisting of well-defined sets of options, are widely used in psychological research to explore how people make decisions under risk. Expected Utility (EU) theory (Neumann \& Morgenstern, 1947) has been widely used to predict choices with well-defined sets of gambles in different forms, but its connection to information search and attention (e.g., as measured by eye movements) has not been clearly revealed. For example, do people who are more riskseeking look more often at risky options?

Eye movements, which have been studied as a process tracing methodology (Glaholt \& Reingold, 2011), have been shown to be related to decisions under different tasks (Krajbich, Armel, \& Rangel, 2010; Krajbich \& Rangel, 2011; Shimojo et al., 2003; Stewart, Hermens, \& Matthews, 2016) and tested by different models (Brandstätter \& Körner, 2014; Fiedler \& Glökner, 2014). Nevertheless, the link between risk attitude revealed by EU and attention and information search strategy is still missing.

In the present study, participants were asked to make risky decisions between two gambles presented perceptually
while their eye movements were monitored. Gambles were represented as a grid of colored pixels where the colors were associated with monetary amounts and the proportion of pixels with the probability of winning that amount. One gamble was risky (higher possible monetary payout, but lower probability) and one was safe (lower possible monetary payout, but higher probability). We used a variant of EU to model participants' choice behavior and then examined the relationship between model parameters and eye movement characteristics. This experimental paradigm enables us to bridge the gap between decision processes (as modeled by EU) and attention and information search patterns (as measured by eye movements).

## Experiment: Speeded Risky Gambling

## Participants

39 undergraduate students (31 female) from Vanderbilt University participated in the experiment for course credit. Their age ranged from 18 to 22 years old (mean $=19.3$ ). We tested 28 participants with a right dominant eye and 11 with a left dominant eye.

## Methods

All stimuli were presented on a 23.5 -inch ViewSonic screen with a 60 Hz refresh rate at $1980 \times 1020$-pixel resolution. The viewing distance was 68 cm and each gamble had an overall size of $4.5^{\circ} \times 4.5^{\circ}$ of visual angle. In this experiment, each trial began with a fixation cross displayed for 0.5 second. Following the fixation cross, two square grids were always presented diagonally at two of the corners on the screen for maximal 2 seconds. These two grids consisted of $20 \times 2010$-pixel squares, which were filled in with grey indicating a zero payout and one of three colors (blue, rose, and yellow) indicating different positive monetary payouts that participants learned from instructions and practice. The proportion of color to grey (i.e., positive payout to zero payout) was randomly selected from 15 pairs of gambles (Table 1). Thus, participants were faced with a choice between two nonnegative gambles that offered different probabilities of winning different amounts of money. The configuration of colored elements in the grids was randomly rearranged every four frames to avoid potential perceptual pattern biases (thus the grids


Figure 1. Speeded risky gambling experiment procedure. There were three practice blocks before the main task. In this case, rose, yellow, and blue represent $\$ 2, \$ 4.5$, and $\$ 7$. In the second and the third practice blocks as well as the main task, participants pressed ' $z$ ' or ' $m$ ' to indicate choosing the left or the right grid, respectively. In practice block 2, participants were instructed to select the gamble with a higher expected value. The feedback indicated whether their choice was correct or not. In practice block 3, participants were allowed to select whichever gamble they preferred. They received feedback about their choices after every trial. In the main experiment, the participants received feedback at the end of blocks. There was no trial-by-trial feedback during the main task.
appeared to flicker). Participants were instructed to select the gamble that they wanted to play. The results of the chosen gambles were provided to participants during block breaks as their cumulative payouts.

Table 1: Gambles used in the Speeded Risky Gambling Task

| $v(\mathrm{R})$ | $p(\mathrm{R})$ | $v(\mathrm{~S})$ | $p(\mathrm{R})$ | $\Delta \mathrm{EV}(\mathrm{R}-\mathrm{S})$ |
| :---: | :---: | :---: | :---: | :---: |
| $\$ 4.5$ | 0.46 | $\$ 2$ | 0.54 | 0.99 |
| $\$ 7$ | 0.33 | $\$ 2$ | 0.67 | 0.97 |
| $\$ 7$ | 0.48 | $\$ 4.5$ | 0.52 | 1.02 |
| $\$ 4.5$ | 0.38 | $\$ 2$ | 0.62 | 0.47 |
| $\$ 7$ | 0.28 | $\$ 2$ | 0.72 | 0.52 |
| $\$ 7$ | 0.43 | $\$ 4.5$ | 0.57 | 0.45 |
| $\$ 4.5$ | 0.31 | $\$ 2$ | 0.69 | 0.02 |
| $\$ 7$ | 0.22 | $\$ 2$ | 0.78 | -0.02 |
| $\$ 7$ | 0.39 | $\$ 4.5$ | 0.61 | -0.02 |
| $\$ 4.5$ | 0.23 | $\$ 2$ | 0.76 | -0.49 |
| $\$ 7$ | 0.17 | $\$ 2$ | 0.83 | -0.47 |
| $\$ 7$ | 0.35 | $\$ 4.5$ | 0.65 | -0.48 |
| $\$ 4.5$ | 0.15 | $\$ 2$ | 0.85 | -1.03 |
| $\$ 7$ | 0.11 | $\$ 2$ | 0.89 | -1.01 |
| $\$ 7$ | 0.3 | $\$ 4.5$ | 0.70 | -1.05 |

$R$ denotes the risky gamble, $S$ denotes the safe gamble
The experiment had three practice blocks before the main experiment, which consisted of 16 blocks (Figure 1). The first practice had 20 trials. In the first practice block, we asked participants to select the monetary value associated with a particular color and provided feedback based on their responses. In the second practice block, two grids were presented that had two different colors (i.e., different monetary amounts), but in the same proportions. Participants needed to choose the grid with the greater expected value (i.e., the grid with the color associated with the higher value). The third practice block was similar to the main experiment, except that feedback on the gambling results was provided for each trial. Both the second and the
third practice had 15 trials. In the main experiment, each block had 66 trials of which $10 \%$ are catch trials. During block breaks, the payout for the current block and cumulative payout were provided. Each gamble pair had one risky gamble and one safe gamble. Risky gambles were defined as gambles where positive payouts were greater than that of safe gambles, but where probabilities of winning were less than that of safe gambles. The difference in expected values ( $\Delta \mathrm{EV}$, defined as EV (risky gamble) EV(safe gamble)) of 15 gamble pairs ranged from about -1 to 1 (Table 1). We had three gamble pairs for each $\Delta \mathrm{EV}$ condition: $\$ 4.5$ and $\$ 2, \$ 7$ and $\$ 2$, and $\$ 7$ and $\$ 4.5$. Catch trials were gamble pairs where the risky gamble had a higher probability and greater monetary payout than the safe gamble.


Figure 2. Panel A: Areas of interest diagram. Panel B: Representative eye trajectory (red lines) within one trial. Two gambles were presented diagonally in randomized locations. Solid black square is the center of fixation. Solid rose and yellow squares are the center of two gambles. Black squares are AOIs (solid: the chosen gamble; dotted: fixation area and the unchosen gamble).

Eye movements were monitored by an EyeLink 1000 desk-mounted eye tracker (SR Research, Ontario, Canada). We tracked participants' dominant eye movement with both pupil and corneal reflection settings at a sampling rate of 1000 Hz . Area of interest (AOI) was defined around the centers of the grids and fixation cross with the size of $400 \times$ 400 pixels. We used these AOIs to determine when participants were looking at each gamble and to explore the gaze dynamics during their deliberation (Figure 2).

After the speeded risky gambling task, every participant completed a set of surveys, which included the Cognitive Reflection Test (CRT, Frederick, 2005), DOSPERT scale (Blais \& Weber, 2006), and Holt and Laury gambles (Holt \& Laury, 2002). In this paper, we did not include the results from the CRT, DOSPERT scale, and Holt and Laury gambles. Those results will be reported elsewhere.

## Results

Two participants, one of whom had less than $30 \%$ correct on the catch trials, one who did not move his or her eyes at all throughout the experiment, and another five participants’ data were not successfully recorded due to technical issues, were excluded. We first analyzed the effect of $\Delta E V$ and different gamble pairs on risky choices and response time. We observed that the probability of risky choice increased as the $\Delta \mathrm{EV}$ increased (Figure 3A). The probabilities of risky choice under the three gamble pairs were different, with the probability of risky choice under the $\$ 7-\$ 4.5$ pair being the highest and the $\$ 7-\$ 2$ pair being the lowest. When $\Delta \mathrm{EV}$ is greater than zero, which indicates that the risky gamble had a greater EV compared to the safe gamble, the risky gamble was more likely to be chosen. We used Bayesian methods to analyze the data and report the resulting Bayes Factors (BF), Based on a Bayesian two-way ANOVA, we found that the model with $\triangle \mathrm{EV}$ and gamble pair without their interaction was preferred to all other models $\left(\mathrm{BF}_{\text {Model }}=636.06\right)$ as well as to the null model $\left(\mathrm{BF}_{10}=2.03 \times 10^{37}\right)$. The Bayes Factors for including the variables $\Delta \mathrm{EV}$, gamble pairings, and their interaction were $\mathrm{BF}_{\text {Inclusion }} \sim \infty$, and $\mathrm{BF}_{\text {Inclusion }}=731.74$, respectively. Regarding response time, risky decisions in general took longer as the $\Delta \mathrm{EV}$ increased. Response time increased as $\Delta E V$ increased under the $\$ 7-\$ 2$ and $\$ 4.5-\$ 2$ pairs, but did not change much under the $\$ 7-\$ 4.5$ pair. Based on a Bayesian two-way ANOVA, we found that a model with both $\Delta \mathrm{EV}$ and gamble pairings and no interaction was preferred to all other models $\left(\mathrm{BF}_{\text {Model }}=\right.$ $580.70)$ as well as the null model $\left(\mathrm{BF}_{10}=2.36 \times 10^{13}\right)$. The Bayes Factors for including $\triangle \mathrm{EV}$ and gamble pair were $\mathrm{BF}_{\text {Inclusion }}=2.17 \times 10^{2}$ and $\mathrm{BF}_{\text {Inclusion }}=2.98 \times 10^{2}$, respectively.

Next, we used number of fixations to investigate information search patterns under the five $\triangle \mathrm{EV}$ conditions with the three gamble pairs (Figure4A). The number of fixations is the average fixation count in non-catch trials prior to the decision. We observed that the number of fixations increased with increasing $\Delta \mathrm{EV}$, which was consistent with response time patterns. Based on a Bayesian two-way ANOVA, we found that the model with both $\triangle \mathrm{EV}$
and gamble pair and no interaction was preferred to all other models $\left(\mathrm{BF}_{\text {Model }}=133.90\right)$ as well as the null model $\left(\mathrm{BF}_{10}=\right.$ $5.32 \times 10^{5}$ ). The Bayes Factors for including $\Delta \mathrm{EV}$ and gamble pair were $\mathrm{BF}_{\text {Inclusion }}=1.48 \times 10^{3}$ and $\mathrm{BF}_{\text {Inclusion }}=$ 286.17, respectively.


Figure 3. Psychometrics in speeded risky gambling experiment. Panel A: probability of risky choices under different $\triangle E V$ conditions. Panel B: the effect of different $\Delta \mathrm{EVs}$ and gamble pairs on response time. Error bars are the standard error of the mean. Dark, medium, and light green lines represent the three gamble pairings.


Figure 4. Eye movement results. Panel A: Fixation numbers under different $\triangle E V$ conditions. Panel B: Probability of choosing the last seen gamble and the other gamble. The error bars are the standard error of the mean.

We also observed the same gaze biases reported in previous studies showing that eye movements made during a choice have a strong relationship with the final choice (Krajbich, Armel, \& Rangel, 2010; Krajbich \& Rangel, 2011; Stewart, Hermens, \& Matthews, 2016). We compared choice proportions when the last gaze was on the chosen gamble with that of the unchosen gamble, and found that the last seen gamble was more likely to be chosen as compared to the other gamble for the three gamble pairs separately $\left(\$ 7-\$ 4.5: \mathrm{BF}_{10}=37.93 ; \$ 7-\$ 2: \mathrm{BF}_{10}=4.58 \times 10^{4} ; \$ 4.5-\$ 2\right.$ : $\mathrm{BF}_{10}=79.17$ ) (Figure 4B). To investigate the influence of different gamble pairs on the relationship of last gaze and risky choices, we further examined the difference in proportion of risky choice given the first or the last gaze was in the AOI of the risky gambles. The first gaze had less influence on final choice of the risky gambles compared to the last gaze. For the three gamble pairs, the proportion of choices for the risky gambles was greater when the last gaze
was on the risky gamble than when the last gaze was on the safe gamble (Table 2 and Figure 5).


Figure 5. Proportion of risky choices given the first or last gaze was on the risky versus safe option. Solid lines: proportion of risky choices when the first or last gaze was on the risky gamble. Dotted lines: proportion of risky choices with the first or last gaze was on the safe gamble. The error bars are the standard error of the mean.

Table 2. Bayesian ANOVA for first and last gaze effects.

|  | Gamble pairs | Best model | $\mathrm{BF}_{\text {model }}$ | $\mathrm{BF}_{10}$ |
| :---: | :---: | :---: | :---: | :---: |
| First Gaze (FG) effect | \$7-\$4.5 | $\Delta \mathrm{EV}$ | 6.3 | $1.4 \times 10^{24}$ |
|  | \$7-\$2 | $\Delta \mathrm{EV}$ | 26.9 | $1.4 \times 10^{17}$ |
|  | \$4.5-\$2 | $\Delta \mathrm{EV}$ | 21.2 | $1.9 \times 10^{55}$ |
| Last Gaze (LG) effect | \$7-\$4.5 | LG $+\Delta \mathrm{EV}$ | 146.9 | $2.8 \times 10^{24}$ |
|  | \$7-\$2 | $\mathrm{LG}+\Delta \mathrm{EV}$ | 34.9 | $1.3 \times 10^{17}$ |
|  | \$4.5-\$2 | $\begin{gathered} \mathrm{LG}+\Delta \mathrm{EV}+ \\ \mathrm{LG}^{*} \Delta \mathrm{EV} \end{gathered}$ | 4.4 | $6.0 \times 10^{53}$ |

The first (last) gaze effect is the difference in the probability of selecting the risky option when the first (last) gaze is on the risky option as compared to the safe option.

## Hierarchical Bayesian Parameter Estimation of "Perceptual" Expected Utility Theory

To account for choice behavior, we developed a "perceptual" variant of EU theory. The parameters of this model were estimated using a hierarchical Bayesian parameter estimation approach. There are two components of our variant of EU: (1) the subjective utility function and
(2) the perception of probabilities. The subjective utility function is governed by alpha, which indicates an individual's risk attitude. If alpha is less than 1 , it indicates that the person is risk-aversive. If alpha is greater than 1 , it indicates that the person is risk-seeking. If alpha is equal to 1 , that person is risk-neutral.

In EU theory, the subjective value of a two-outcome gamble $G$ is determined by

$$
\begin{equation*}
E U(G)=p_{1} u\left(x_{1}\right)+p_{2} u\left(x_{2}\right) \tag{1}
\end{equation*}
$$

where $u(\cdot)$ is the utility of the outcomes defined as

$$
\begin{equation*}
u\left(x_{i}\right)=x_{i}^{\alpha} \tag{2}
\end{equation*}
$$

where $\alpha$ is a free parameter that is greater than 0 and quantifies the curvature of the utility function.

In EU theory, the objective probabilities are used to compute the expected utilities of the gambles. In our experiment, however, the probabilities of both safe and risky gambles are presented as proportions of colors in the flickering grids. It is possible that people's perception or estimation of the actual proportion does not match the exact probabilities shown in the grids. That is, there may be perceptual distortion of the probability estimation. In order to capture this feature, we assume that perceived probability is given by the following function:
where $\varphi$ is the adaptation level for the proportion of color in the grid and $\beta$ is a shape parameter. The probability of choosing option A over option B is modeled as:

$$
\begin{equation*}
p(A, B)=\frac{E U(A)}{E U(A)+E U(B)} \tag{4}
\end{equation*}
$$

Hierarchical modeling serves as a compromise between a no-individual-differences model and a full-individualdifferences model. In a hierarchical model, individual parameters are drawn from group-level distributions, usually normal distributions with estimated mean and standard deviation. The estimated means quantify different cognitive processes. The standard deviations quantify the similarity among individual participants' behavior.

We used the prior distributions for these three parameters as following. Individual $\alpha_{i}$ is drawn from the normal distribution with two group-level parameters $\mu^{\alpha} \sim U(0,5)$ and $\sigma^{\alpha} \sim U(0,10)$. Individual $\varphi_{i}$ is drawn from the normal distribution with two group-level parameters $\mu^{\varphi} \sim U(0,100)$ and $\sigma^{\varphi} \sim U(0,10)$. Individual $\beta_{i}$ is from the normal distribution with its group-level parameter $\mu^{\beta} \sim U(0,100)$ and $\sigma^{\beta} \sim U(0,10)$. We implemented the hierarchical EU model in JAGS. Posterior distributions were approximated by 3 MCMC chains with 5000 samples from each chain, after a burn-in of 1000 samples. Convergence of chains was evaluated by computing the $\hat{R}$ statistic. Figure 6 shows the posterior distributions of three group-level means. The means of $\mu^{\alpha}, \mu^{\varphi}$, and $\mu^{\beta}$ are $0.92,0.41$, and 13.13, respectively.

By comparing observed choices with the predicted choices, we are able to assess how accurately the model captures people's choice behavior. We plot the observed risky choice proportions from the data with the posterior predictive of the model using the individual-level parameters (Figure 7). The model predictions are reasonably close to the actual proportion of risky choices indicating that the model accounts for the group-level data.


Figure 6. Posterior distributions of EU group-level parameters.


Figure 7. Comparison of the risky choice proportions from the data and predicted by EU theory. Light grey bins represent risky choice proportion from data, and dark grey bins represent risky choice proportion predicted by EU theory.


Figure 8. Comparison of risky choice proportion from the data and that predicted by EU. The line is $\mathrm{y}=\mathrm{x}$.

To examine the performance of EU theory at the individual level, we plot the individual data and predictions from EU theory using the individual-level parameters (Figure 8). Most of the data points fall on the diagonal,
meaning that the performance of EU is reasonable at the individual level as well.

## Individual differences in gaze dynamics

To bridge the gap between the conceptualized EU model and actual information search dynamics, we examined the correlation of EU parameters and eye movement statistics. We examined the correlations with four eye movement measures: \% of trials with first fixation on risky, \% of trials with last fixation on risky, proportion of gaze duration on risk, proportion of gaze duration on chosen gambles, as well as response time. See Table 3 for definitions of these five measures.

In EU, the parameter $\alpha$ captures participants' risk preferences. Note that EU assumes that the subjective utility function is concave if $0<\alpha<1$, implying that people are risk-averse, while the subjective utility function is convex if $\alpha>1$, implying that people are risk-seeking. A larger value of $\alpha$ implies less risk-aversion (or relatively greater riskseeking behavior). We found that $\alpha$ was positively correlated with \% of trials with last fixation on risky and proportion of gaze duration on risky gamble (see Table 3). For the two measures, the $\mathrm{BF}_{10}$ was greater than 100 , indicting extremely strong support for the correlations. These correlations suggest that people who are more riskseeking tend to look more at risky options.

Table 3. Correlations and Bayes Factors (BF) between eye movement statistics, response time, and EU parameters

| Eye movement measures | EU parameters |  |  |
| :---: | :---: | :---: | :---: |
|  | $\alpha$ | $\varphi$ | $\beta$ |
| \% of trials with first | 0.31 | -0.47 | -0.1 |
| fixation on risky | (0.942) | (7.21) | (0.29) |
| \% of trials with last | 0.65*** | -0.60*** | -0.14 |
| fixation on risky | (479.29) | (133.72) | (0.30) |
| proportion of gaze | 0.70*** | $-0.73 * * *$ | -0.23 |
| duration on risky | (2787.37) | (11149.51) | (0.46) |
| proportion of gaze | $\begin{gathered} -0.14 \\ (0.29) \end{gathered}$ | $\begin{aligned} & -0.17 \\ & (0.33) \end{aligned}$ | $\begin{gathered} 0.05 \\ (0.23) \end{gathered}$ |
|  | (0.29) | (0.33) |  |
| response time | $\begin{gathered} 0.38 \\ (1.87) \\ \hline \end{gathered}$ | $\begin{array}{r} -0.29 \\ (0.76) \\ \hline \end{array}$ | $\begin{aligned} & 0.004 \\ & (0.22) \\ & \hline \end{aligned}$ |
| $\mathrm{BF}_{10}$ enclosed in parentheses. * $\mathrm{BF}_{10}>10, * * \mathrm{BF}_{10}>30$, *** $\mathrm{BF}_{10}>100$ |  |  |  |
| - \% of trials with first fixation on risky: proportion of trials i which the first fixation after gambling presentation was on the |  |  |  |
| which the last fixation before decisions were made was on the risky. |  |  |  |
| - proportion of gaze duration on risk: ratio of gaze duration on th risky to the response time of each trial. |  |  |  |
| - proportion of gaze duration on chosen: ratio of gaze duration on the chosen gamble to the response time of each trial. |  |  |  |
| - response time: from stimuli onset to responses by pressing keys. |  |  |  |

The parameter $\varphi$ determines the adaptation level of the perceived probability transform function. When $\varphi<0.5$,
individuals are more adapted to small probabilities, which also correspond to riskier options in our task (i.e., risky gambles in our experiment have high values and small probabilities). Thus, smaller values of $\varphi$ indicate increased adaptation for risky options. We found that the parameter $\varphi$ was negatively correlated with $\%$ of trials with the last fixation on risky and the proportion of gaze duration on risky.

The $\beta$ parameter is the shape parameter for the probability transform function. This parameter was not correlated with any of the eye movement measures we calculated.

## Discussion

In this study, we investigated the relationship between gaze dynamics and "perceptual" EU parameters in risky decisionmaking. First, we corroborated previous findings that the last fixation was closely related to actual choices (Krajbich, Armel, \& Rangel, 2010; Krajbich \& Rangel, 2011; Shimojo et al., 2003; Stewart, Hermens, \& Matthews, 2016). Going beyond this, we observed that people with different risk attitudes have different patterns of eye movements, which serve as a proxy for information search and attention. In particular, we found that (i) the utility shape parameter of EU was positively correlated with measures related to gaze duration on the risky option and to the proportion of last gaze on the risky option, and (ii) the adaptation level in perceived probability was negatively correlated with the proportion of the last gaze on the risky option and with the gaze duration on the risky option. These results establish the connection between risky choice behavior conceptualized by EU and information search strategies under risky decision-making revealed by gaze dynamics.

Given the fact that eye movements are only considered as a proxy of internal processes of attention, we cannot rule out the possibility that participants held the two gambles in a mental comparison while moving their eyes. Thus, we cannot conclude that the decision processes conceptualized by the EU model caused specific information search strategies or vice versa. Future studies are needed to explore the causal relationship between gaze dynamics and choice. For example, future studies could examine if changing information search strategies by manipulating the salience of risky gambles might influence people's risky choices. Also, manipulating exposure time of options might influence choices.

We conclude by addressing some of the limitations of the present study. First, all thirty-nine students were granted course credit regardless of their performance. It is possible that their behavior might change if there was actual monetary reward rather than a hypothetical situation. The three estimated parameters of EU might be different when participants are more engaged to maximize their final payouts (Holt \& Laury, 2002). Second, in this study we did not include gambles with pure losses or mixtures of both gains and losses. People may adopt different strategies in
this speeded risky gambling task when losses are introduced. Addressing these issues would be suitable for future studies. Nevertheless, we did observe individual differences in risk preferences as measured by EU and these differences where related to differences in gaze dynamics. This suggests that information search and attention is related to underlying decision processes.

## Acknowledgement

JST was supported by NSF grant SES-1556325.

## References

Blais, A. R., \& Weber, E. U. (2006). A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. Judgment and Decision Making, 1(1), 33-47.
Brandstätter, E., \& Körner, C. (2014). Attention in risky choice. Acta Psychologica, 152, 166-176.
Fiedler, S. \& Glökner, A. (2012). The dynamics of decision making in risky choice: an eye-tracking analysis. Frontiers in Psychology, 3(335), 1-18.
Frederick, S (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19, 25-42.
Glaholt, M. G. \& Reingold, E. M. (2011). Eye movement monitoring as a process methodology in decision making research. Journal of Neuroscience, Psychology, and Economics, 4(2), 125-146.
Holt, C. A. \& Laury, S. K. (2002). Risk aversion and incentive effects. The American Economic Review, 92(5), 1644-1655.
Krajbich, I., Armel, C., Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. Nature Neuroscience, 13(10), 1292-1298.
Krajbich, I. \& Rangel, A. (2011). Multialternative driftdiffusion model predicts the relationship between visual fixations and choice in value-based decisions. PNAS, 108(33), 13852-13857.
Neumann, von, J., \& Morgenstern, O. (1947). Theory of Games and Economic Behavior (60 ed.). Princeton University Press.
Shimojo, S., Simion, C., Shimojo, E., Scheier, C. (2003). Gaze bias both reflects and influences preference. Nature Neuroscience, 6(12), 1317-1322.
Stewart, N., Hermens, F., Matthews, W. J. (2016). Eye movements in risky choice. Journal of Behavioral Decision Making, 29, 116-136.

# Novel Evidence for the Bilingual Advantage: Effects of Language Control on Executive Function in Balanced and Unbalanced Dual-Language Users 

W. Quin Yow (quin@sutd.edu.sg), Adam C. Oei (adam_oei@sutd.edu.sg), \& Xiaoqian Li<br>(xiaoqian_li@mymail.sutd.edu.sg)<br>Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372, Singapore


#### Abstract

Bilinguals' need to monitor and inhibit non-relevant languages over a relevant one confers advantage in cognitive control. No studies have demonstrated that the dual-language control process directly contributes to the bilingual cognitive advantage. We utilized a novel language control manipulation paradigm where 83 English-Chinese bilingual adults completed a reading and comprehension task in either singlelanguage (low-language-control) or dual-language (high-language-control) prior to performing nonverbal executive control tasks (Stroop, task-switching, and n-back). Results showed that language control had significant effects on subsequent cognitive performance, depending on whether the participants were regular dual language users or not. In the dual-language condition, but not the single-language condition, participants who used both languages regularly demonstrated a smaller mixing cost in task-switching and a greater sensitivity in $n$-back detection compared to participants who did not. This suggests that dual language control utilizes similar resources as executive function and frequent dual language use enhances this resource.


Keywords: bilingualism; cognitive resources; task mixing; working memory; adults

## Introduction

Research evidence suggests that the ability to speak more than one language confers an advantage in cognition, specifically in executive control. Executive control refers to a set of top-down mental processes such as the ability to inhibit impulses, monitor and update working memory representations and task switching (Miyake et al., 2000). The evidence for the bilingual advantage in executive control (henceforth bilingual advantage) is robust and is seen across the lifespan from young children to the elderly across several executive control tasks (e.g., Bialystok, Craik, \& Luk, 2008; Macnamara \& Conway, 2013).

Why would being bilingual have positive consequences to executive control? The transfer from bilingualism to enhanced executive control is likely a two-step process.

First, for a bilingual, different languages are activated in parallel during language processing (e.g., Crinion et al., 2006). Increased executive control resources are therefore needed to monitor this parallel activation and prevent crosslinguistic interference. Various executive control processes are hypothesized to be necessary for language control in different interactional contexts (Green \& Abutalebi, 2013). For example, a bilingual speaker must maintain a task goal and inhibit the non-relevant language when speaking in one language rather than another. These processes are akin to updating and inhibition components in the Miyake et al.
(2000) model of executive control. The cognitive demands on such processes are likely to increase in a dual-language context whereby both languages known to a bilingual are used and switched within a conversation. Such codeswitching acts (i.e., switching between languages) add another dimension of control necessary for disengagement from a prior language and engagement of the language in use. Neuroimaging studies have shown overlapping neural substrates between language control and nonverbal executive control (e.g., De Baene, Duyck, Brass, \& Carreiras, 2015; Rodríguez-Pujadas et al., 2014).

Second, the control processes associated with language control in bilingual speakers are hypothesized to adapt to the demands imposed on them, in turn enhancing domaingeneral cognitive control (e.g., Bialystok \& Craik, 2010; Green \& Abutalebi, 2013). Behavioral evidence indicates that executive control processes implicated in language control such as updating, interference suppression and shifting are enhanced in bilinguals (e.g., Morales, Calvo, \& Bialystok, 2013; Wiseheart, Viswanathan, \& Bialystok, 2014). Neuroimaging work suggests that bilingual language control exerts neuroplastic effects both structurally and functionally in brain areas of importance for executive control (e.g., Klein, Mok, Chen, \& Watkins, 2014). Bilinguals show reduced activation relative to monolinguals when performing conflict and inhibitory control tasks in the anterior cingulate (Rodríguez-Pujadas et al., 2014), indicating greater neural efficiency. These findings suggest that bilingual speakers may hold larger cognitive resources because of the adaptive language control processes.

A more specific account has been put forward recently, that these general "spill-over" positive effects of bilingualism in executive control skills may be due to the dual language switching behavior that bilinguals regularly engage in, which represents a skillful control of language use (e.g., Prior \& Gollan, 2011; Yim \& Bialystok, 2012). For example, Soveri, Rodriguez-Fornells, and Laine (2011) found that higher self-reported daily language switching frequency is associated with reduced task-mixing cost in adults from 30 to 75 years old. In addition, Yow and Li (2015) demonstrated that bilinguals who used both languages regularly have lower Stroop interference effects and task mixing costs than those who used one language significantly more than the other language. This bilingual advantage based on how often a bilingual uses both languages is also apparent across the lifespan from childhood to the elderly. Specifically, Thomas-Sunesson, Hakuta, and Bialystok (2016) showed that the more
balanced Spanish-English bilingual young children were in their language proficiency between two languages, the better they performed in working memory and conflict resolution.

Bilinguals often differ from each other in language proficiency and usage frequency of each of their languages, as well as language switching behavior. These differences in language usage behavior might lead to different neural and cognitive consequences due to the executive control processes implicated in language control. However, it is currently unknown whether active engagement in language control activities would have a positive or negative impact on bilingual speakers' performance in nonverbal tasks tapping executive control processes. No studies have experimentally manipulated language control to determine its effects on executive control. This present study aims to explore this question and to further understand the mechanisms underlying the effects of language control on executive control.

In the current study, we manipulated active engagement in language control by introducing a novel reading and comprehension task. In the single-language condition, participants read articles and answered questions in only English. In the dual-language condition, participants read and answered questions intermixed in English and Chinese, which would induce higher cognitive demands on language control compared to those in the single-language condition. Following the reading comprehension task, participants completed three executive control tasks (i.e., Stroop, number-letter task-switching, and n-back) to assess different executive control components (i.e., inhibition, shifting, and updating, respectively) (e.g., Miyake et al., 2000). If the engagement in language switching behavior indeed caused a higher cognitive demand (in the dual-language condition), then it could be expected that less cognitive resources would be available for integrating the text information, resulting in lower performance in the comprehension test. We also hypothesized that participants' performance in the subsequent executive control tasks would be negatively affected if they were in the dual-language condition, as cognitive resources would be depleted for these participants who had to engage in greater language control.

Since previous studies indicate that bilinguals who regularly use both languages may have a protective advantage in cognitive control, current participants who are balanced bilinguals (defined as either a balanced use or a balanced level of proficiency in two language systems) would be less likely to be affected by the language control manipulation than those who are unbalanced bilinguals.

## Method

## Participants

Eighty-three undergraduates ( 56 females, $M_{\text {age }}=22.60, S D$ $=1.83$, range $=19-25$ ) were recruited from the authors' university. All participants are Chinese Singaporeans and have been living in Singapore since they were born.

Singapore is a multilingual country with English as the main official language. The bilingual policy in Singapore encourages citizens to be proficient in both English and a mother tongue, which is Chinese for the participants of this study. All participants provided written informed consent prior to their participation and received credit points or reimbursement for their time of participation.

## Materials and Measures

Language Background Questionnaire (LBQ) This questionnaire asked participants to name all the languages that they know and to provide details about each of the listed languages (e.g., age of language acquisition, language proficiency, usage frequency, and language switching habits). For language proficiency, participants rated their proficiency in listening, reading, speaking and writing for each language on a 10 -point scale $(1=$ not proficient to $10=$ very proficient). We defined most and second most proficient language based on the two languages that have the highest average rating across these four domains. Usage frequency for each of the languages was assessed by asking participants to approximate the percentage they use each language when communicating with different groups of people (e.g., family members, colleagues, friends) in different contexts in a typical week. The usage of all different languages would add up to $100 \%$. In addition, participants rated on a 5 -point scale $(1=$ never to $5=$ always) for nine questions relating to how frequent they switch languages during discourse. A higher score therefore indicates more frequent language switching.

We followed the procedures in Yow and Li (2015) to estimate the individual differences in the degree of bilingualism:
(I) Balanced Proficiency: Most proficient language rating minus second most proficient language rating; a metric of balanced bilingualism of relative competency between participants' most and second most proficient languages.
(II) Balanced Usage: Frequency of most used language minus frequency of second most used language; a metric of balanced bilingualism of relative usage frequency between participants' most used and second most used languages.

For both measures, a score closer to 0 indicates more balance between two languages. Conversely, a higher score indicates more dominant proficiency or use in one language over the other. See Table 1 for all key language variables, including details between the balanced and unbalanced users related to the current study. Differences between the balanced and unbalanced users are reported as this is of interest in the current study.

Reading Comprehension Task This task was administered as a manipulation of language control. Participants were randomly assigned to a single-language condition (i.e., $\boldsymbol{S L}$; $n=38$ ) or a dual-language condition (i.e., $\boldsymbol{D L} ; n=45$ ). For both conditions, participants read four passages pertaining to current events happening at the authors' university (two before executive control tasks and two after).

Table 1: Language background measures.

| Language variable | Overall | Balanced users |  |
| :--- | :--- | :--- | :--- |
| Mean (SD) | Unbalanced users <br> Mean (SD) <br> $n=40$ | Mean (SD) <br> $n=39^{*}$ |  |
| Age of first acquired language | $n=83$ | $2.23(1.51)$ | $2.35(1.29)$ |
| Age second acquired language | $2.79(1.75)$ | $2.40(1.26)$ | $3.00(1.69)$ |
| Most proficient language proficiency rating | $8.49(1.29)$ | $8.47(1.10)$ | $8.47(1.48)$ |
| Second most proficient language proficiency rating | $6.54(1.76)$ | $7.36(1.34)$ | $5.63(1.80)$ |
| Frequency of most used language | $0.76(0.19)$ | $0.66(0.93)$ | $0.91(0.58)$ |
| Frequency of second most used language | $0.21(0.17)$ | $0.32(0.09)$ | $0.09(0.06)$ |
| Balanced proficiency (Most proficient minus $2^{\text {nd }}$ most proficient) | $1.95(1.84)$ | $1.11(1.00)$ | $2.84(2.16)$ |
| Balanced usage (Most used minus $2^{\text {nd }}$ most used) | $0.55(0.34)$ | $0.35(0.17)$ | $0.82(0.11)$ |
| Language switching | $25.96(4.26)$ | $26.60(4.60)$ | $25.11(3.91)$ |

*Note: Four participants did not provide their usage frequency for their languages. Hence, we were not able to categorize them as balanced or unbalanced users.

Participants assigned to the SL condition read passages presented in English only. The English reading passages are suitable for students in grade 12 (between grade 11 to 13), based on Flesch-Kincaid readability tests performed online (https://readability-score.com). The English passages contained on average 332.75 words (range $=305$ to 366 words). In the DL condition, participants were presented with passages intermixed in English and Chinese. The passages contained 438.5 words on average (range $=406$ to 458 words). Only inter-sentential switches (i.e., switching from English to Chinese sentence and vice versa) were used ( $M=13.5$ switches, range $=11$ to 15 ).

After reading each passage, participants were required to answer eight questions in oral and then written form. For each passage, four questions required an oral response and four required a written response. However, four were filler questions (two oral and two written) that were not scored. Participants were given one point for each correctly answered question (not including the filler questions). A maximum score a participant can receive is therefore four. Participants could answer in either English or Chinese for any of the questions, regardless of whether they are in the SL or DL condition. The questions were all presented on the computer screen. The passages remained on screen for the participants to refer to if needed.

In the SL condition, all questions were posed in English only. In the DL condition, participants were presented the questions in English and then Chinese in alternating fashion. For each of the four questions posed (oral and written), the answers could only be found in the passage printed in the opposite language. In other words, for a question posed in English, its corresponding answer is printed in Chinese in the passage and vice versa. This therefore necessitated participants in the DL condition to engage in codeswitching to answer the questions.

The sum of all correct responses was calculated for each form of questions (oral or written) and for each two passages before or after the executive control tasks as DVs
(dependent variables) from this task.
Executive Control Tasks We selected three tasks, colorword Stroop, number-letter switching, and n-back, to measure executive control components of inhibition, mental-set shifting, and information updating and monitoring, respectively (Miyake et al., 2000; see Yow \& Li, 2015 for details about stimuli design and procedure for each of the tasks).

In the Stroop task, participants were required to indicate the color that the stimuli were printed in and ignore the color names. The dependent measure in this task is the Stroop effect, taken as the difference in response time (RT) between the incongruent and neutral trials. Greater Stroop effects reflect the poorer inhibition.

In the letter-number switching task, participants saw number-letter pairs and either determined if the number was even or odd or if the letter was a vowel or consonant. The DVs for this task are switch cost and mixing cost. Switch cost reflects more transient control processes to updating goals or task demands, while mixing cost reflects cognitive control in actively maintaining representations of multiple task demands.

In the n-back task, participants completed two blocks of 2-back and two blocks of 3-back test trials. The dependent measures in this task are $d^{\prime}$ calculated separately for 2 and 3-back, reflecting detection sensitivity according to the signal detection theory.

## General Procedure

All tasks were administered individually in a quiet room at the authors' university. Participants first completed the LBQ. Participants then performed the two reading comprehension tasks and subsequently completed three executive control tasks: Stroop task, task-switching and nback task, programmed in MATLAB (Version 7.10). The order of the Stroop and task-switching task was counterbalanced between participants but the n-back task
was always performed last. When all the executive control tasks were completed, participants completed two other reading comprehension tasks. Visual stimuli were presented to participants from about 70 cm via a 23 -inch monitor with a refresh rate of about 60 Hz . For the executive control tasks, participants were instructed to respond as quickly and as accurately as possible. Each experimental session took about 1.5 to 2 hours.

## Results

Prior to data analyses, we screened the data from all executive control tasks to remove incorrect trials as well as trials with those RTs shorter than 200 ms or longer than 3000 ms . These discarded trials amounted to less than $6 \%$ ( $3 \%-5 \%$ ) of the total number of trials for each task. We also discarded DVs (Stroop interference RT, switch cost RT, mixing cost RT, 2-back $d^{\prime}$ and 3-back $d^{\prime}$ ') of interest that were greater or less than 2.5 SDs from the group mean. These outliers were rare and amounted to about $2 \%$ of the total number of data points. These data trimming procedures are typical in studies using similar experimental tasks.

We performed a median split on balanced proficiency and balanced usage to categorize participants into balanced and unbalanced proficient groups as well as balanced and unbalanced dual-language users. This is to directly compare these groups and subsequently perform analyses to determine the effect of proficiency and usage interactions on executive control. As participants' language switching could potentially confound the relationship between language usage and proficiency on our executive control measures, we controlled for its effects by entering language switching as a covariate in separate analyses of covariance.

## Reading Comprehension Performance

We first performed 2 (time: reading comprehension before and after executive control tasks) x 2 (response type: oral and written) x 2 (condition: SL and DL) x 2 (balanced usage: balanced and unbalanced) x 2 (balanced proficiency: balanced and unbalanced) mixed ANCOVA with language switching as a covariate to determine whether performance in the reading comprehension task differed as a function of time, bilingualism, response type and condition. Importantly, this also allowed us to perform a manipulation check to determine if the DL condition was indeed more cognitively demanding than the SL condition.

There was a Condition effect, $F(1,63)=21.71, p<.001$, $\eta_{p}^{2}=0.26$. This was qualified by a Condition x Balanced Usage interaction, $F(1,63)=4.14, p=.05, \eta_{\mathrm{p}}{ }^{2}=0.06$. Specifically, unbalanced language users' reading comprehension performance was poorer when required to codeswitch in the DL condition $(M=11.80)$ compared to the SL condition where codeswitching was not required ( $M$ $=13.56), t(36)=2.34, p=.03$. Similarly, balanced users performed poorer in reading comprehension in the DL ( $M=$ 10.22) than in the SL condition $(M=14.12)$.

There was also a Time effect, $F(1,63)=8.10, p=.01, \eta_{\mathrm{p}}{ }^{2}$ $=0.11$, indicating that participants performed better in the
reading comprehension task following $(M=3.18)$ the executive control tasks than before $(M=3.06)$. This was however qualified by a Time x Language Switching interaction. We followed up this significant interaction with a median split of language switching and compared the reading comprehension scores for frequent and infrequent language switchers. While the infrequent language switchers performed better $(M=6.57)$ after the executive control tasks than before $(M=5.87), t(36)=-3.03, p=.01$, the frequent language switchers performed equally well before $(M=$ 6.13) and after $(M=6.16)$ the executive control tasks, $t(37)$ $=-0.1, p=.93$. This suggests infrequent switchers but not the frequent switchers gained from a prior practice in the reading comprehension task. No other main effects or interactions approached or were statistically significant.

Overall, the significant main effect of condition showed that the dual-language condition was more demanding and indicated that our manipulation was successful in increasing executive control load.

## Executive Control Performance

Inhibition A 2 (balanced proficiency) x 2 (balanced usage) x 2 (condition) ANCOVA with language switching as a covariate to account for unintentional switching in language use as a potential confound was conducted. Language switching was not significantly related to Stroop effect ( $p=$ .53). There was a significant Balanced Proficiency x Condition interaction, $F(1,65)=4.22, p=.04$. Follow up comparisons however showed no significant difference in Stroop interference effects between the two conditions amongst balanced proficient bilinguals ( $p=.095$ ). This was similar for the unbalanced proficient bilinguals $(p=.22)$. No other significant effects were found (all $p \mathrm{~s}>.43$ ).

Shifting The switch cost and mixing cost in task-switching were evaluated in a similar way. There were no significant main effects or interactions on the switch cost (all $p \mathrm{~s}>.23$ ). In contrast, for the mixing cost, the Balanced Usage x Condition interaction was significant, $F(1,67)=4.87, p=$ $.03, \eta_{\mathrm{p}}{ }^{2}=0.07$ (see Figure 1). In the SL condition, no significant differences were found between balanced and unbalanced dual language users in mixing cost ( $p=.26$ ). However, in the DL condition, the balanced users had a significantly smaller mixing cost than the unbalanced users ( $p=.04$ ). These results suggest that codeswitching in the DL condition negatively affected unbalanced participants in task mixing but not on balanced usage participants.

Updating On the 2-back trials, there was a significant Balanced Usage x Condition interaction on the sensitivity to targets, $F(1,67)=4.48, p=.04, \eta_{\mathrm{p}}^{2}=0.06$ (see Figure 2). Follow up comparisons revealed that although there was no difference in 2-back $d^{\prime}$ between balanced and unbalanced users in the SL condition ( $p=.31$ ), 2-back $d^{\prime}$ differed between the two groups in the DL condition $(p=.047)$. This demonstrated that codeswitching adversely affected
information updating (in 2-back) for the unbalanced users, but had no effect on the balanced users.

On the 3-back trials, a significant main effect of balanced usage was evident, $F(1,68)=6.17, p=.02, \eta_{\mathrm{p}}{ }^{2}=0.08$. This main effect was qualified by a Balanced Usage x Condition interaction, $F(1,68)=4.80, p=.03, \eta_{\mathrm{p}}^{2}=0.07$. Similar as the 2-back task, no significant differences in performance were found between balanced and unbalanced duallanguage users in the SL condition ( $p=.84$ ). However, balanced users had significantly higher 3-back $d^{\prime}$ scores compared to unbalanced dual language users in the DL condition ( $p=.001$ ). This indicates that balanced dual language users were less affected by codeswitching than did their unbalanced counterparts (see Figure 3). Additionally, a significant balanced proficiency main effect was found for the 3-back $d$ ' scores, $F(1,68)=5.72, p=.02, \eta_{\mathrm{p}}{ }^{2}=0.08$. The Balanced Usage x Balanced Proficiency interaction was also significant, $F(1,68)=6.91, p=.01, \eta_{\mathrm{p}}^{2}=0.09$.

## Discussion

This study investigated the effects of language control on domain-general executive control in bilingual adults who differed in the degree of bilingualism. By implementing a novel language control manipulation and adopting an individual differences approach to study the bilingual advantage, we found that (1) engaging bilingual speakers in a dual-language context (involving codeswitching) increased cognitive load and interfered with information organization and integration, resulting in poorer comprehension performance, (2) increased cognitive demands on language control depleted general executive control resources and negatively impacted executive control components such as maintenance of mental representations and sensitivity to targets, and (3) effects of language control on executive control were modulated by factors that influenced the exposure and opportunity for bilingual speakers to practice language control in daily life.

When required to codeswitch prior to performing executive control tasks, participants who were balanced dual language users demonstrated smaller mixing cost and better target discrimination compared to those unbalanced users. These results suggest that a more balanced use of two languages may function as a cognitive reserve that would mitigate the effects of language control on executive control, i.e., balanced bilinguals may have larger cognitive resources for executive control than less balanced bilinguals. This advantage is likely due to balanced dual language users having more opportunity to be involved in interactional contexts where both languages are used with different speakers, and the constant need to monitor and control attention to the target language system over the competing other language, which in turn lead to the development of larger cognitive resources for adaptive language control processes.

The interaction effects of balanced usage and language control condition were significant for only mixing cost and working memory updating, but not for Stroop interference


Figure 1. Mixing cost for balanced and unbalanced usage participants in single- and DL conditions. Error bars denote $95 \%$ confidence intervals.


Figure 2. Performance in 2-back ( $d^{\prime}$ ) by balanced usage and condition. Error bars denote $95 \%$ confidence intervals.


Figure 3. Performance in 3-back ( $d^{\prime}$ ) by balanced usage and condition Error bars denote 95\% confidence intervals.
effect or switch cost, indicating that only some cognitive control processes were used in the current dual-language context. This is consistent with the adaptive control hypothesis proposing that the demand on different control processes varies as a function of the interactional context. For instance, it is hypothesized that, for bilingual speakers, both single-language and dual-language contexts increase the demands on interference control that was assessed by the Stroop task in this study. Thus, resources for inhibition may not be depleted in the dual-language context, even amongst unbalanced users. Our failure to find bilingual advantages in switch cost has been documented previously (e.g., Yow \& Li, 2015). Being more balanced in dual language use may not have conferred an advantage to the transient switching cost because task switching is also akin to frequent topic changes during discourse, in which both balanced and unbalanced users are equally likely to engage in (Wiseheart et al., 2014). In contrast, mixing cost reflects global
sustained control necessary for maintaining competing mental representations, which is akin to using different languages in dual-language contexts. Balanced dual language users may have developed a larger executive control resource to hold multiple language rules "on-line", preventing resource depletion following our codeswitching manipulation. Lastly, updating and monitoring is postulated to be necessary for maintenance of task goals during dual language discourse. Unbalanced dual language users who hold smaller working memory resources are therefore more affected in updating after having to codeswitch. Taken together, these results indicate that bilingual advantage is limited to certain but not all components within the Miyake et al. (2000) model.

One limitation inherent in our study is the use of selfreport for language proficiency. Although this approach is consistent with many other studies (e.g., Wiseheart et al., 2014), we acknowledge that using self- report measures may result in less accurate proficiency ratings than objective measures. However, given that many participants may know more than two languages, objective measures of their language proficiency may not be feasible without substantially prolonging the experimental session.
In conclusion, we provide novel evidence showing that language control and executive control depend on shared resources by experimentally manipulating language control via codeswitching. Crucially, we provide evidence of the bilingual advantage in showing that bilinguals who are more balanced in dual language use have larger working memory and task mixing resources that buffered against performance decline following language control. Theoretically, our results add to the current understanding of the mechanism of the bilingual advantage in executive control: language and executive control share similar resources, and this shared resource can be enhanced by using more than one language equally frequently.

## Acknowledgments

The authors wish to thank the participants who participated in the study, and Stephanie Dietz and Tony Lim for their help in experiment design and data collection. This work was partially supported by the AcRF (SUTDT12015007) and DIRP grant (IGDSP1502032) to the first author.

## References

Bialystok, E., Craik, F., \& Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(4), 859-873.
Bialystok, E., \& Craik, F. I. M. (2010). Cognitive and linguistic processing in the bilingual mind. Current Directions in Psychological Science, 19(1), 19-23.
Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J. T., . . . Price, C. J. (2006). Language control in the bilingual brain. Science, 312(5779), 1537-1540.

De Baene, W., Duyck, W., Brass, M., \& Carreiras, M. (2015). Brain circuit for cognitive control is shared by task and language switching. Journal of Cognitive Neuroscience, 27(9), 1752-1765.
Green, D., \& Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. Journal of Cognitive Psychology, 25(5), 515-530.
Klein, D., Mok, K., Chen, J.-K., \& Watkins, K. E. (2014). Age of language learning shapes brain structure: A cortical thickness study of bilingual and monolingual individuals. Brain and Language, 131, 20-24.
Macnamara, B., \& Conway, A. (2013). Novel evidence in support of the bilingual advantage: Influences of task demands and experience on cognitive control and working memory. Psychonomic Bulletin \& Review, 21(2), 520-525.
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., \& Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive Psychology, 41(1), 49-100.
Morales, J., Calvo, A., \& Bialystok, E. (2013). Working memory development in monolingual and bilingual children. Journal of Experimental Child Psychology, 114(2), 187-202.
Prior, A., \& Gollan, T. H. (2011). Good language-switchers are good task-switchers: Evidence from Spanish-English and Mandarin-English bilinguals. Journal of the International Neuropsychological Society: JINS, 17(4), 682.

Rodríguez-Pujadas, A., Sanjuán, A., Fuentes, P., VenturaCampos, N., Barrós-Loscertales, A., \& Ávila, C. (2014). Differential neural control in early bilinguals and monolinguals during response inhibition. Brain and Language, 132, 43-51.
Soveri, A., Rodriguez-Fornells, A., \& Laine, M. (2011). Is there a relationship between language switching and executive functions in bilingualism? Introducing a within group analysis approach. Frontiers in Psychology, 2, 183. doi:10.3389/fpsyg.2011.00183
Thomas-Sunesson, D., Hakuta, K., \& Bialystok, E. (2016). Degree of bilingualism modifies executive control in Hispanic children in the USA. International Journal of Bilingual Education and Bilingualism, 1-10.
Wiseheart, M., Viswanathan, M., \& Bialystok, E. (2014). Flexibility in task switching by monolinguals and bilinguals. Bilingualism: Language and Cognition, 19(1), 1-6.
Yim, O., \& Bialystok, E. (2012). Degree of conversational code-switching enhances verbal task switching in Cantonese-English bilinguals. Bilingualism: Language and Cognition, 15(04), 873-883.
Yow, W. Q., \& Li, X. (2015). Balanced bilingualism and early age of second language acquisition as the underlying mechanisms of a bilingual executive control advantage: Why variations in bilingual experiences matter. Frontiers in Psychology, 6, 164. doi:10.3389/fpsyg. 2015.00164

# Inconvenient samples: Modeling the effects of non-consent by coupling observational and experimental results 

Yue Yu ${ }^{12}$ (yue.yu@ rutgers.edu), Elizabeth Bonawitz ${ }^{2}$ (elizabeth.bonawitz@rutgers.edu), Patrick Shafto ${ }^{1}$ (patrick.shafto@gmail.com)<br>Department of ${ }^{1}$ Math and Computer Science \& ${ }^{2}$ Psychology, Rutgers University - Newark Newark, NJ 07102 USA


#### Abstract

Biased sampling of participants presents a major limiting factor for the generalizability of findings from behavioral studies. This effect may be especially pronounced in developmental studies, where parents serve as both the primary environmental input and decide whether their child participates in a study. To estimate the effects of parental non-consent, we coupled naturalistic observations of parent-child interactions with a behavioral test. Results showed that one particular parenting practice, the tendency to use questions to teach, associated with both children's behavior in the test and parents' tendencies to participate. Exploiting these associations with a model-based multiple imputation, we estimated that the means of the consented and not-consented groups could differ as much as 0.2 standard deviations for five of the seven test measurements we used, and standard deviations are likely underestimated. These results suggest that ignoring the role of consent may lead to systematic biases when generalizing beyond lab samples.


Keywords: sampling; generalization; parent-child interaction; learning; exploration; multiple imputation.

## Introduction

Sampling and generalizability are the methodological bedrocks of science. Researchers often rely on measurements taken from a small group of volunteers to draw conclusions for a much broader population, so knowing whether the sample is representative of the population becomes critical to the validity and generalizability of research findings. Among the many factors that may bias the sampling process, one prevalent but under-studied factor is the refusal to participate in research. Because ethical treatment of research participants requires informed consent prior to their participation, we know very little about what characteristics are associated with non-consent, and what those who did not consent would have done if they had participated.

The problem is potentially more acute for fields in which behavior tends to be heterogeneous along factors that may associate with non-consent. One such field is experimental research with young children: On the one hand, before the start of schooling, children's experiences are heavily influenced by the values and practices of their parents, which are known to be heterogeneous both within and between social groups (Bornstein, 1991; Hoff, Laursen, Tardif, \& Bornstein, 2002). At the same time, parents are also the ones who decide whether their children could participate in research, and the same values and practices may play a role in their decision. Given that parents' decisions are crucial for the composition of research
samples for most of the prevalent recruitment methods used in developmental experiments with young children (e.g., direct phone calls, recruitment from day care centers, preschools, and public spaces like museums), non-consent can present a major hurdle when evaluating the generalizability of findings from the field.

However, to date little is known about the factors associated with parents' non-consent to have their young children participate in experiments; consequently, it is difficult to speculate on the behavior of children who did not participate or on the implications for generalizability. This is in sharp contrast with the field of survey-based research with school-aged children and adolescence, where an extensive literature has associated parental consent and nonconsent with both parents' characteristics and children's behavioral outcomes. For example, studies from 1970s-80s have shown that U.S. parents who are female, white and well-educated are more likely to return written consents for their children to participate in research (Kearney, Hopkins, Mauss, \& Weisheit, 1983; Lueptow, Mueller, Hammes, \& Master, 1977), and children who have better school performance and fewer behavioral problems are more likely to receive parental consents (Kearney et al., 1983; Severson \& Ary, 1983). Moreover, the method of consent also matters. Compared to passive consent which requires a reply to opt out of a study, active consent which requires a reply to optin can bias the sample towards parents who are white and well-educated, and towards children who live in two-parent households, who have better school performance and satisfaction, who involve in more extracurricular activities and less risk-seeking behavior, and who are higher on selfesteem and assertiveness (Anderman et al., 1995; Dent et al., 1993). Critical to estimating these effects is the availability of relevant correlates, such as school records that contain demographic information and students' performance.

This study takes a first step to investigate whether parents' non-consent is also associated with preschoolers' behavior in standard experimental settings. Given that methods used to discover factors associated with non-consent in schoolaged children (such as passive consent procedures and school records) are not usually applicable in the research with preschoolers, we developed a new method that is wellsuited to many settings in which developmental psychologists collect data: coupling naturalistic observations of parent-child interactions with behavioral experiments.

By conducting the observation in public spaces without the awareness of the dyad, we aim to start with a relatively representative population that is unaffected by the consent
process. We then invite participation in a behavioral test to those who were observed. By analyzing the correlations between the observational and test data, and between the observational data and participation, we look for predictors that may associate test data from the participated group with participation itself, which would indicate the potential effects of non-consent on estimates of test measurements. If such association can be established, we would then use a model-based multiple imputation to simulate the behavior of children who did not participate, and compare the participated sample with the initial population on the test measurements.

The domain we chose to examine is one where there is known heterogeneity in parenting practices: the use of questions to teach. This line of research is grounded in a rich literature about informal pedagogy (Bonawitz et al., 2011; Csibra \& Gergely, 2009), which suggests that the format in which parents and educators chose to present evidence to children influence how children infer and learn. Specifically, recent experiments (Yu, Bonawitz, \& Shafto, under review) have shown that pedagogical questions asked by knowledgeable teachers are particularly effective in facilitating children's learning and exploration of a novel artifact. Given that the tendency to ask pedagogical questions has been shown to vary across parents (Yu, Bonawitz, \& Shafto, 2016), we explore the effects of children's experiences with pedagogical questions on their responses in the experiment. We did so by replicating one condition of the previous experiment with an added observation phase, in which parents' pedagogical questions towards children were measured along with other parentchild interaction measurements. This allows us to look for associations between parents' pedagogical questions and children's responses to these questions. And because the observational data is available for children who did not participate, these associations could then be used to estimate the test data for the whole population.

## Method

## Testing sites

We set up the study in two sites: an indoor reptile exhibit in a zoo, and an indoor playground. The zoo is in Essex County, NJ, which is one of the nation's most racially diverse counties, and has one of the nation's most unequal economies measured by the Gini Index (U.S. Census Bureau, 2016). The playground is in Middlesex County, NJ, which is overall more affluent, but also has a highly racially diverse population, of which $30 \%$ were born outside the U.S. (U.S. Census Bureau, 2016). We chose these two sites to ensure diversity in the population we initially observed. Consents from these two sites and from the internal review board was obtained before conducting the study.

These two sites differ in the expected level of supervision and involvement from parents. The zoo is an open environment that requires parents to constantly supervise their young children. Exhibits in the zoo also feature many
textual materials which require parents' explanations for young children to understand. In comparison, the playground features spaces and activities which young children can navigate on their own, and the closed environment allows minimal supervision from their parents. This contrast allows us to test whether the characteristics associated with parental non-consent differ by the type of facilities from which they were recruited.

## Participants

Between the two sites, we observed a total of 109 parentchild dyads. Of these 109 dyads, 31 were not invited for the test because of one of the following: the dyad left before the observation was finished (18), parent interrupted researchers during observation (1), researchers did not get a chance to invite parent (4), adult accompanying child was not the child's parent (4), child was out of the target age range (2), or child did not speak English (2). The remaining 78 parentchild dyads comprised our "population", which is unaffected by the consent process. Among them 41 were recruited from the zoo, and 37 were recruited from the playground.

## Procedure

During each trip to the testing sites, three researchers collected data from parent-child dyads in three phases: Two coders first observed and coded the interactions between the parent and the child (observation phase). Then a third researcher invited the dyad to a test (recruitment phase). She and one of the coders conducted the test if the dyad agreed to participate (testing phase).

Observation phase. Two coders pretended to be visitors of the zoo or the playground, so that they could code parentchild interactions without the dyad's awareness. The coders first looked for a child who was estimated to be between 3 and 6 years of age, and determined who were the adults accompanying the child. If at least one adult looked like the child's parent, the coders would record the members in the group (e.g., father, mother, and two daughters), and agree on a target dyad for observation (e.g., mother and younger daughter). To reduce potential selection biases from the coders, they always observed the first dyad they saw that fitted the requirements, and the observation always started immediately once the target dyad was determined. Each dyad was observed for 5 minutes, during which the coders independently coded both the quantity and the quality of parent-child interactions. Quantity of interaction was measured by the length of time period of dyadic activities (parent and child engaging in the same activity), supervised activities (parent watching, following, or taking pictures of child while child is engaging in his or her own activities), and unsupervised activities (parent and child engaging in different activities). We coded these as mutually exclusive categories, and they added up to the total time length (5 minutes). Quality of interaction was coded as a set of frequency measurements, adapted from the Dyadic ParentChild Interaction Coding System (Eyberg, Nelson, Ginn,

Bhuiyan, \& Boggs, 2013): The coders recorded the numbers of parents' questions, statements, and commands towards children.

Critical to our interest, parents' questions were further differentiated based on their functions (Yu et al., 2016): Those used to help children learn were coded as "pedagogical questions", whereas those used to request information from children were coded as "informationseeking questions". All coders were trained for approximately 5 hours and practiced the coding scheme with at least 5 parent-child dyads before formal data collection. Inter-rater reliabilities were high for both the quantity and quality of parent-child interactions: Inter-rater correlation $r$ $=.78 \sim .84$ for quantity codes, and $r=.79 \sim .86$ for quality codes. The average of the two coders' codes were used for data analysis.

Recruitment phase. After the 5-minute observation, a third researcher who was blind to the observation phase approached the parent and invited the parent-child dyad to participate in a test. The recruitment procedure followed a script which resembled that of a typical developmental experiment: The research started with a brief selfintroduction, then described the research as a study of how children learn and explore a novel toy, then briefly explained the consent form, and finally asked if the parent would be interested to have his or her child participate in the test. For parents who had multiple children with them, we specifically asked for the child who had been observed. Among the population of 78 parent-child dyads that were observed, 59 agreed to participate (the "consented" group) and 19 refused (the "not-consented" group). Of the 59 parents who agreed, 11 children did not participate in the test ( 2 were busy playing and did not get a chance to come, 8 refused to come, and 1 did not understand English), and the video was missing for one additional child, so data from the testing phase were available for 47 children. According to parental report, children who participated in the testing phase were diverse regarding race ( $51 \%$ white, $4 \%$ black, $15 \%$ Hispanic-Latino, $13 \%$ Asian, $17 \%$ multi-racial), but most came from middle- or upper-class families $(91 \%$ of the main caregivers have college diploma or above, $84 \%$ of the families have annual house hold income of $\$ 50 \mathrm{~K}$ or above).

Testing phase. Parents and children who agreed to participate were led to a corner of the zoo exhibit or a separate room in the indoor playground, where the test was conducted by the recruiter (acting as an experimenter) and one of the coders (acting as a confederate). The materials and procedure of the test was adapted from Bonawitz, et al. (2011), and was identical to the pedagogical question condition in our recent experiment ( Yu et al., under review).

A novel toy of approximately $14 " \times 7.5^{\prime \prime} \times 14.5^{\prime \prime}$ was used in the test. In addition to several inert properties, the toy had five functional parts: a tower that lit up when a button was pushed, a knob that produced a squeaking sound when squeezed, a lady bug pin light that flashed in three
different patterns when pushed, a flower magnet that moved between three different places on the toy, and a turtle hidden in a pipe that was visible through a magnifying window.

During the test, the child sat at a table opposite the experimenter and the confederate. The toy was initially hidden out of sight. The experimenter first said that she knows about the toy and the confederate does not, and asked the confederate to bring out the toy. After the confederate brought out the toy and handed it over to the her, the experimenter then asked a pedagogical question to the child, "I'm asking you to think about: What does this button do?", while pointing to the button on the tower without activating it. Then she told the child it is his or her turn to play with the toy, and to let the researchers know when he or she is done. The test ended when the child stopped playing and signaled the researchers, and a sticker was presented as a reward. The whole phase was video recorded.

## Video coding

After data collection, the videos from the testing phase were coded by another research assistant who was blind to the observation phase and to the hypotheses of the study. She first determined the total time children spent playing with the toy, and then coded three measurements regarding both the whole playing period, and the first minute after children started playing: whether children activated the target function (the tower with the button), the number of unique actions they performed with the toy, and the number of nontarget functions (out of 4) they activated. A second coder coded $14(30 \%)$ of the videos, and the inter-coder reliability was high for all measurements: total time playing: $r=.98$; activating target function: Cohen's $\kappa=1$ for both total time and first minute; number of non-target functions activated: Cohen's $\kappa=81$ (total time) and $\kappa=75$ (first minute); number of unique actions performed: $r=.79$ (total time) and $r=.92$ (first minute).

## Results

Our population consisted 32 mother-son dyads, 16 motherdaughter dyads, 17 father-son dyads, and 13 father-daughter dyads. Parent-child interactions varied both across sites and within sites: Compared to dyads in the playground, dyads in the zoo spent more time on dyadic activities, and less time on supervised (but not dyadic) activities or unsupervised activities, $t \mathrm{~s}>2.6, p \mathrm{~s}<.01$. Parents also asked more pedagogical and information seeking questions, and said more statements in the zoo than in the playground, $t \mathrm{~s}>3.4$, $p \mathrm{~s}<.001$. The difference in parents' commands toward children was marginally significant, $t(67.7)=1.75, p=.09$. These results suggest that the testing site needs to be taken into account when interpreting parent-child interactions. Therefore, testing site had been entered as a control variable for all further analyses. We also observed large within-site variations: For all measurements, standard deviations were higher than $1 / 3$ of the mean for both the zoo and the playground. These variations suggest that the population we observed was diverse with regard to parent-child


Figure 1. The estimated effects of non-consent on one of the test measurements: the number of unique actions children performed during the first minute of play. We tested whether the mean and the standard deviation measured from the consented group (1c, blue bars) are unbiased estimates of the population. To do so, we separated the consented group according to the number of PQs parents asked children during the observation prior to the test (1a). Parents who asked more PQs tended to have children who performed more unique actions during the test. On the other hand, as shown in (1b), parents who asked more PQs also tended to consent their children to participate (blue) rather than not consent (red). These two associations resulted in the simulated not-consented group to have a lower mean than the consented group (1c), by an estimated effect size (Cohen's $d$ ) of 0.2. Compared to the population (purple bars), focusing on the consented group may result in an overestimation of the mean and an underestimation of the standard deviation. $P Q=$ pedagogical questions. Error bars denote standard errors across children (1a) or across simulations (1c).
interactions, which serves as a basis for further correlational analyses. Though parent-child interactions differed by site, the proportion of parents who agreed to participate did not differ significantly, playground: 25 agreed, 12 refused ( $68 \%$ vs. $32 \%$ ); zoo: 34 agreed, 7 refused ( $83 \%$ vs. $17 \%$ ); Fisher's exact $p=.19$.

## Are parent-children interactions associated with children's behavior in the test?

Test data was available for 47 children ranging from 3.0 y to $6.3 y$, of which 27 were recruited from the zoo and 20 from the playground. Children from the two sites did not differ with regard to the activation of target and other functions, or the number of unique actions they performed on the toy, $t$ < $1.4, p \mathrm{~s}>.1$. However, there was a trend of children playing longer with the toy in the playground than in the zoo, $M_{\mathrm{zoo}}=$ $189 \mathrm{~s}, M_{\text {playground }}=132 \mathrm{~s}, t(29.4)=1.83, p=.08, d=0.58$. When comparing these results with previous experiments we conducted in preschools ( $n=30$, age range $=4.0 \mathrm{y}$ to 6.0 y ) using the exact same protocol (Yu et al., under review), none of children's response measurements differed significantly across the three sites, $F \mathrm{~s}<2.2, p \mathrm{~s}>.1$.

Next we looked at the relation between children's responses during the test and parent-child interaction measurements during the observation. After controlling for testing site and age, measurements regarding the composition of the group being observed (parent' and child's gender, and whether they were accompanied by other adults or children) did not correlate with any of children's responses, $p \mathrm{~s}>.1$. However, measurements of parent-child interaction did correlate with children's responses: Children of parents who spent more time
watching and following them were less likely to discover the target function during the first minute of play, $r(42)$ $=.33, p=.02$. At the same time, children whose parents asked more pedagogical questions discovered more other functions of the toy, $r(42)=.32, p=.03$, and also performed more unique actions during first minute of play (Figure 1a), $r(42)=.29, p=.05$. These results suggest that patterns observed in parent-child interactions were indeed associated with children's learning and exploration during the test.

## Are parent-child interactions associated with participation?

We then examined whether patterns observed in parentchild interactions also predicted parents' responses to the invitation for research. We fitted a logistic regression model with participation as the dependent variable and the observational measurements as the predictors. Overall the model predicted actual participation with $80 \%$ accuracy. With regard to individual predictors, parents were more likely to have their boys participate than girls, $B=1.47, p$ $=.03$; and those parents who asked more pedagogical questions during the observation were more likely to participate, $B=1.49, p=.05$ (Figure 1b).

## What can be predicted for children who did not participate?

Results so far have shown that the number of pedagogical questions parents asked children predicted both children's participation in a test and their behavior during the test. This indicates that children's participation and behavior may be related as well-that is, if we have tested children whose
parents did not consent them to participate, they may have responded differently than children who did participate.

To test this hypothesis, we applied model-based multiple imputation to our data (Rubin, 2004). ${ }^{1}$ The model we used for multiple imputation was a stochastic regression model, implemented with IBM SPSS 22. The seven observational measurements were used to model the seven test measurements, based on data from the consented group. The resulting models were then used to predict behavior of the not-consented group stochastically (with random noise) for 100 independent runs of simulations. ${ }^{2}$

Results showed that across the 100 runs of simulations, the means of the not-consented group were consistently different from that of the consented group for five out of seven test measurements including activating target function (total time and first minute), number of non-target functions activated (first minute), and number of unique actions performed (total time and first minute). The departure was towards the same direction-the participated children learned and explored more with the toy (Figure 1c shows one example). The differences between the means of the consented and not-consented group were estimated to be between 0.09 and 0.20 standard deviations for these five measurements. In addition, compared to the population, focusing on the consented group alone would lead to consistent underestimation of the standard deviations across children, and this is true for all test measurements we examined.

## Discussion

This study takes a first step towards evaluating whether

[^251]results from children who participated in an experiment could generalize to children whose parents did not consent for them to participate. We attempted to estimate these potential biases with a novel approach by pairing a behavioral test with naturalistic observations of parent-child interactions prior to parental consent. Results have shown that a specific parenting practice-asking questions to help children learn-correlated with both parents' tendencies to have their children participate in the test, and children's learning and exploration during the test. And since the observational data was available for both those who participated and did not participate, we were able to exploit these associations to impute behavior for children who did not participate. Results from the imputation showed differences in group means between the consented and notconsented group for five out of the seven test measurements, with estimated effect sizes (Cohen's $d$ s) between 0.09 and 0.20 . Furthermore, the consented group showed a lower standard deviation than the population for all test measurements.

Before discussing the implications of the results, it is worth noting that several assumptions underlie these simulated estimates. First, we assumed no direct causal relation between parents' decisions to have their children participate and children's potential behavior in the test. This assumption is plausible in our case: Because parents were not given much detail about the testing procedures, their decision to participate is unlikely to be based on what they expect their children to do. However, in other situations this assumption could be violated, which could render the imputation analysis invalid. For example, in a study that measures children's executive functions, if children drop out from the study exactly because of low executive functions, then it would be invalid to impute executive functions for the dropout group even when all relevant correlates have been observed and entered into the model. Second, our approach is valid because we saw variations in parent-child interactions for both the consented and not-consented groups, as well as significant overlap between the two groups. This allows imputation to be done as interpolations within the ranges of empirical support. In cases where the consented and not-consented groups do not overlap, our approach could be invalid, as the relations found in the consented group may not extend to the not-consented group. In sum, our methods to generalize experimental results are themselves subject to usual conditions for generalization.

How much this new approach could be and should be implemented in developmental experiments would also depend on various factors. The first factor is the recruitment method. Our approach could be beneficial for research settings that provide opportunities to observe and recruit from a relatively diverse population, such as in public spaces. On the other hand, for studies recruiting from places with a preselected population, such as preschools, the demographics of the preselected population may present a stronger sampling bias than parents' consent. The second factor is the research topic. Our approach could be more
valuable for domains in which parent-child interactions have been, or are expected to be, associated with children's behavior. The last factor is the research ethics. Pre-consent observations are ethically viable only for public actions, and needs to be performed with caution.

In cases where our approach can be applied, it could benefit the interpretation and generalization of experimental findings in several ways: First, it could reveal correlations between parent-child interactions and children's behavior, which may help explain the cognitive mechanisms and environmental inputs associated with the observed behavior. Second, it could inform the generalizability of experimental findings to children whose parents did not consent them to participate. Third, it can serve as an empirical base for future research to recruit a more representative sample. By knowing the associations between parental consent and patterns in parent-child interaction, it may be possible to intentionally focus recruitment on parent-child dyads who are likely underrepresented in typical recruitment procedures.

Our results may also have implications for developmental theories. Many developmental theories are built upon findings from experiments, as experimental design has advantages in addressing a range of developmental questions: These include depicting developmental trajectories ("Children do X at age Y "), disentangling causal mechanisms underlying children's behavior ("Children do X because of Z'), and testing causal effects of interventions ("T helps children do X "). In typical cases, random assignment of participants across groups removes unwanted systematic differences between groups, so that the effects of age, condition, or treatment can be detected by comparing between-group differences with within-group differences. Our results have shown that parental non-consent may have biased this comparison in two ways that random assignment cannot solve: First, it could lead to an underestimation of within-group variations, and thus Type I errors may be underestimated and effect sizes may be overestimated. Second, compared to the general population, children who received consent may be more susceptible or insusceptible to certain manipulations or treatments, therefore biasing the estimation of the between-group differences. Because findings from developmental experiments often guide realworld practices which apply to the general population, understanding factors and biases associated with nonconsent is essential when interpreting and applying these findings.

To conclude, this study provided a first empirical demonstration that children with and without parental consent to participate in research may have differed in behavior measured in an experiment. Therefore, parental non-consent should be considered an important factor when evaluating the generalizability of experimental findings, and the theories built upon them. In addition, we provided a method that, in certain contexts, could be used to estimate the effect of parental non-consent and generalizability of experimental results.

## Acknowledgments

This study is supported by NSF CAREER, DRL-1149116 to P.S., NSF Awards SMA-1640816 and ECR-1660885 to E.B. and P.S., and Jacob Foundation fellowship to E.B. We thank Reham Bader, Merna Botros, Milagros Grados, Anishka Jean, and Natasha Patel for help in conducting the study.

## References

Anderman, C., Cheadle, A., Curry, S., Diehr, P., Shultz, L., \& Wagner, E. (1995). Selection bias related to parental consent in school-based survey research. Evaluation Review, 19(6), 663-674.
Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., \& Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. Cognition, 120(3), 322-330.
Bornstein, M. H. (1991). Cultural approaches to parenting. Hillsdale, NJ: Lawrence Erlbaum.
Csibra, G., \& Gergely, G. (2009). Natural pedagogy. Trends in Cognitive Sciences, 13(4), 148-153.
Dent, C. W., Galaif, J., Sussman, S., Stacy, A., Burtun, D., \& Flay, B. R. (1993). Demographic, psychosocial and behavioral differences in samples of actively and passively consented adolescents. Addictive Behaviors, 18(1), 51-56.
Eyberg, S., Nelson, M., Ginn, N., Bhuiyan, N., \& Boggs, S. (2013). Dyadic parent-child interaction coding system (DPICS): Comprehensive manual for research and training (4th ed.). Gainesville, FL: PCIT International.
Hoff, E., Laursen, B., Tardif, T., \& Bornstein, M. (2002). Socioeconomic status and parenting. Handbook of parenting Volume 2: Biology and ecology of parenting, 8(2), 231-252.
Kearney, K. A., Hopkins, R. H., Mauss, A. L., \& Weisheit, R. A. (1983). Sample bias resulting from a requirement for written parental consent. Public Opinion Quarterly, 47(1), 96-102.
Lueptow, L., Mueller, S. A., Hammes, R. R., \& Master, L. S. (1977). The impact of informed consent regulations on response rate and response bias. Sociological methods \& research, 6(2), 183-204.
Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys. Hoboken, NJ: John Wiley \& Sons.
Severson, H. H., \& Ary, D. V. (1983). Sampling bias due to consent procedures with adolescents. Addictive Behaviors, 8(4), 433-437.
Sinharay, S., Stern, H. S., \& Russell, D. (2001). The use of multiple imputation for the analysis of missing data. Psychological Methods, 6(4), 317.
U.S. Census Bureau. (2016). 2011-2015 American community survey. Retrieved from factfinder.census.gov Yu, Y., Bonawitz, E., \& Shafto, P. (in press). Pedagogical questions in parent-child conversations. Child Development.
Yu, Y., Bonawitz, E., \& Shafto, P. (under review). Questioning supports transmission of knowledge and exploratory learning in pre-kindergarden children.

# Seeing Is Not Enough for Sustained Visual Attention 

Lei Yuan, Tian Linger Xu, Chen Yu and Linda Smith<br>Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University Bloomington, IN, 47405 USA


#### Abstract

Sustained visual attention is crucial to many developmental outcomes. We demonstrate that, consistent with the developmental systems view, sustained visual attention emerges from and is tightly tied to sensory motor coordination. We examined whether changes in manual behavior alter toddlers' eye gaze by giving one group of children heavy toys that were hard to pick up, while giving another group of children perceptually identical toys that were lighter, easy to pick up and hold. We found a tight temporal coupling between the dynamics of visual attention and the dynamics of manual activities on objects, a relation that cannot be explained by interest alone. In the Heavy condition, toddlers looked at objects just as much as did toddlers in the Light condition but did so through many brief glances, whereas in Light condition looks to the objects were longer and sustained. We discuss the implication of hand-eye coordination in the development of visual attention.


Keywords: Sustained visual attention; hand-eye coordination; multimodal; perception action; manual behavior; developmental systems

## Introduction

The ability to focus attention on an individual object or event for a period of time, often in the face of distractions, is predictive of learning and general cognitive capacities (Lansink, Mintz, Richards, 2000; Ruff \& Lawson, 1990). The ability to sustain visual attention undergoes substantial developmental change from infancy to early childhood with a steady increase in both total duration and the ability to resist distractions (Ruff \& Lawson, 1990; Kannass, Oakes \& Shaddy, 2006). Prior research on the development of visual attention has focused on both the effect of low-level stimulusdriven properties (exogenous) and the emergence of topdown internal control of attention (endogenous) (Colombo, 2001). However, like the development of many other cognitive capacities, visual attention interacts with and is influenced by other sensory modalities within the developmental system (Thelen \& Smith, 1994). The ability to sustain attention may not emerge directly from the development of internal controls but rather externally-from the coupling of vision with physical action.

Within this view, visual attentional skills are not built solely on the development of vision; but rather are influenced, altered, and coordinated with other sensory modalities (Yu, Smith L, Shen, Pereira, \& Smith T, 2009). One apt example is the demonstration that deaf children performed worse on a non-auditory visual attention task than their age-matched controls; but, deaf children who had cochlear implant for at least one year performed similarly to hearing children (Quittner, Smith, Osberger, Mitchell, \& Katz, 1994). Because the visual attention task did not rely on
auditory process at all, the deficit shown by deaf children without the implant was solely attributable to an impoverished capacity of visual attention. A history of having auditory experience with the aid of cochlear implant helped to build visual attention, which was then successively recruited to perform a task that did not rely on auditory information. Thus, visual information alone is not enough for building visual attention; the interaction of multiple sensory modalities may be critically involved in the pathways to internal control of attention.

We focus here on the role that manual behavior plays in the control of visual attention. It has long been recognized that the development of perception is driven by the development of motor behaviors (Gibson, 1979). For example, as infants achieve motor milestones (e.g., sitting, crawling and walking), they are able to receive different perceptual experiences (e.g., stably held objects, optical flow), leading to the development of various perceptual abilities such as object recognition and depth perception. Research has also shown that changes on the affordance of objects (or how they can be held) alters the visual input infants receive, which in turn alters the outcome of object recognition (Pereira, James, Jones, \& Smith, 2010). Thus, changes in manual behavior may alter infants' visual attention on objects through the coordination between hands and eyes.

Recent research suggests that infants' hands and eyes are dynamically coupled during toy play (Pereira, Smith, \& Yu, 2014; Yu \& Smith, 2014, 2016) and this coupling may play a causative role in sustained attention. The natural learning environment is complex, often presenting multiple visually interesting objects in a cluttered setting. In these visually complex contexts, infants may rely on manual behaviors to externally select and maintain attention on a target of interest. For example, Pereira, Smith, \& Yu (2014) have shown that infants own manual actions on objects help them to select target, reduce visual clutter, and create larger input images in the visual field, leading to sustained visual attention on objects, better object recognition and early word learning. In this sense, manual action helps to regulate and sustain visual attention. Conversely, it has been found that irregular attentional patterns in atypical development co-occur with perturbations between the visual and manual modalities (Koterba, Leezenbaum, \& Iverson, 2014).

Both between- and within-person hand and eye coordination may contribute to a more mature control of visual attention in social contexts. Yu and Smith (2016) demonstrated that parent's visual attention often follows infant's hands to the object to which the infant was directed; this between person hand and eye coordination substantially prolonged infant's sustained attention on the same object during toy play. Thus, visual attention is not a sole product of
vision or perhaps even the individual but also influenced by cross-person sensory-motor coordination. Consistent with this idea, recent findings suggest that joint attention is also dependent on the child's hand-eye coordination. One- to two-year-olds who showed more tightly coordinated hands and eyes were also better able to coordinate their attention with their parents and better able to sustain joint attention with a parent (Yu \& Smith, 2013, 2014).

All previous findings linking hand-eye coordination in toddlers to sustained visual attention were correlational. Here we attempt to show a causal link. We manipulated manual behavior by giving one group of children heavy toys that were hard to pick up but that could be poked and touched in various interesting ways. We gave another group of children perceptually identical toys that were lighter and easy to pick up and hold. The expectation is that the duration of each individual hand contact will be less for the heavy toys than the light toys that can be picked up and held. However, if the toys are equally engaging-which we designed them to bethe total amount of hand contact may not differ between the two conditions. The key expectation then is on the dynamics of individual contacting events: more briefer touches (pokes and touches) in the heavy case and fewer but longer touches (poking and touching while holding) in the light case.

We illustrate the expectations under the two hypotheses in Figure 1. First, if infants' hands and eyes are dynamically coupled-when hands are on an object, eyes are more likely to be on the same object-then the different dynamic properties of the manual behaviors caused by the weights of object should lead to different dynamic patterns of visual attention, with less sustained attention in the heavy condition (Fig. 1, H1). Second, and in opposition, if visual attention is independent of hand actions (if visual properties of objects solely determine gaze) then, when presented with novel and interesting toys, infants would visually look at them and for similar durations at each looking event, irrespective to whether the object can be held or not (Fig. 1, H2). We expect that the results will support Hypothesis 1: children from both conditions will manually handle and visually attend to the objects for the same total amount of time over the whole play session, but the dynamic properties of gaze will differ considerably and aligned with the different dynamics of the hands.


Figure 1. Two hypotheses for the looking patterns in the heavy condition

## Methods

## Participants

The final sample consisted of thirty-one parent - toddler (mean age $=21$ months old, range $=18-25$ ) dyads. Roughly half (16) of the dyads were assigned to play with light weight toys, while the other half (15) played with heavy weight toys. Children were recruited from a population of working and middle class families in a Midwestern town.

## Stimuli

Two sets of six novel toys (12 in total) were developed from extensive pilot work to be engaging for manual play with moveable elements, openings, and possible actions. They were made of hardened clay, painted in red, blue or green, and were roughly the same size $(9.5 \times 6.5 \times 5 \mathrm{~cm})$. The two sets were identical in terms of shape, size and color, with the only difference being their weights. The heavy set of toys was on average 1.4 lbs , seven times heavier than the average weight of the light set, which was 0.21 lbs .

## Apparatus

Parent and child sat across a small table ( $61 \mathrm{~cm} \times 91 \mathrm{~cm} x$ 64 cm ) (see Fig. 2). The child was strapped loosely into a small chair and the parent sat cross-legged on a pillow. Both participants wore head-mounted eye trackers with a sampling rate of 30 hz (positive science, LLC; also see Franchak et al., 2011). The eye tracker consists of a scene camera that captures the egocentric view of the participant, and an infrared camera that is mounted on the head, points to the right eye of the participant, and records the eye-in-head position ( x and y ) in the captured scene. Another highresolution camera (recording rate 30 frames per sec) was mounted above the table and provided a bird's eye view that was independent of participants' movements.


Figure 2. The experimental setup.

## Procedures

To place the eye tracker on the child head, one experimenter attracted the child's attention with an interesting toy, while another experimenter put the eyetracking gear low on the child's forehead. To calibrate the eye tracker, the experimenter directed the child's eyes toward an interesting toy, which were repeated 15 times while the toy was placed at various locations on the table. Parents were instructed to place the eye tracker on their heads. Parents' eye


Figure 3. Frequency of manual activity events that last for different durations
tracker was calibrated in a similar way. After this initial set up, parents were told that the goal of the experiment is to study how parents and their toddlers interact during toy play, and were instructed to play with their toddlers as naturally as possible.

The free play session lasted for a total of 6 minutes that was composed of four trials with each lasted 1.5 minutes. The six novel toys were grouped into two sets (A and B) with each set having three different colored objects (red, blue and green). The sets were interleaved with the order of the sets counterbalanced across dyads (ABAB or BABA). At the end of each trial, the experimenter signaled parent with a clicking sound, and quickly replaced the old set of toys with a new set.

## Coding

Three regions-of-interest (ROI) were defined for both the eye tracking data and the manual action data: the green, blue and red object. These ROIs were coded manually by coders who annotated frame-by-frame when the cross-hairs overlapped with any of the three ROIs. Another coder independently coded $10 \%$ of the frames with $95 \%$ agreement between coders. The final dataset consisted of a total of 203,316 frames.

## Results

Because our manipulation was on the weight of objects, we first analyzed toddlers' manual activity. We then turned to visual attention as measured by gaze patterns. Finally, we examined the hand-eye coordination as a possible mechanism that drives the observed effects.

## Manual activity

We defined manual activity event as any event during which the toddles' hands were in contact with any of the three objects (data from two hands were coded individually and then combined with a manual contact defined as either or both hands). Results showed that children in the heavy condition handled the objects for a comparable amount of total time as those in the light condition (Fig. 4), suggesting that overall the Heavy and Light versions of the toys were both manually engaging. There was no significant difference in the proportion of total time children in the light ( $M=84 \%$, $S D=7 \%$ ) and heavy condition ( $M=87 \%, S D=6 \%$ ) were in manual contact with the objects, $t(29)=.24, p=.8$. This is important to rule out the possibility that due to object weight,
children in one of the conditions were more interested in the objects and played with them more than the other condition.

Children in the heavy condition $(M=21.31, S D=7.26)$ produced manual activities at a higher frequency (count of events per minute) than those in the light condition ( $M=$ $15.48, S D=4.42$ ), $t(29)=2.67, p=.01$ (Fig. $3 \& 4$ ). But, children in the heavy condition ( $M=2.63 \mathrm{~s}, S D=.99 \mathrm{~s}$ ) spend less time in each manual activity event than those the light condition $(M=3.46 \mathrm{~s}, S D=1.38 \mathrm{~s}), t(29)=1.91, p=.06$. Thus, it appears that children in the light condition would pick up and hold objects, resulting in many long manual activity events. In contrast, children in the heavy condition generated more short manual activity events because they can't hold the objects for a long time if at all, and would probably more often touch the object that sat on the table. This prediction was confirmed by the data: during manual activity events, compared to the heavy condition, children in the light condition had on average a larger visual image size (the size of the object in proportion to the entire visual field captured by the ego-centric view recording in the eye tracker), Light: $M=5.84 \%, S D=.99 \%$, Heavy: $M=4.21 \%$, $S D=1.17 \%, t(29)=4.24, p<.001$.


Figure 4. The proportion of time, mean duration and frequency of manual activity events for both conditions.

Despite the similarity in the total duration of manual activity, the way children handled the objects were different between the two conditions. Because previous studies have used 3 seconds as the threshold of sustained attention (Ruff \& Lawson, 1990; Yu \& Smith, 2016), here we defined sustained manual activity as any manual action that lasted for more than 3 s . Consistent with our prediction, children in the heavy condition had significantly more short (less than 3 seconds) manual activity events per six minutes (session length) than did children in the light condition (Heavy $=1553$, Light $=994$ ); in contrast, the number of sustained manual action events per six minutes were comparable between conditions (Heavy $=493$, Light $=399$ ). Chi-square test of independence indicated that there was a significant


Figure 5. Frequency of looking events that last for different durations
relationship between the number of sustained manual activity events and the weight of objects, $\chi^{2}(1, \mathrm{~N}=3439)=8.92, p=$ .002.
These results set the stage for answering the key question: given that hand dynamics differ, do eye dynamics-and sustained attention episodes-differ as well?

## Visual attention

To analyze children's visual attention, we first examined all looking events during which the child had fixated on any of the objects (the ROIs). There was no significant difference in the proportion of total time children in the Light ( $M=67 \%$, $S D=2 \%$ ) and Heavy conditions ( $M=65 \%, S D=2 \%$ ) looked at the objects, $t(29)=.51, p=.61$. Thus, children from both conditions were visually interested in the objects by this measure.

The mean duration of looking events was significantly lower in the heavy condition ( $M=2 \mathrm{~s}, S D=0.44 \mathrm{~s}$ ) than the light condition $(M=2.43 \mathrm{~s}, S D=0.63 \mathrm{~s}), t(29)=2.21, p=.03$. However, the looking events in the heavy condition ( $M=$ 20.55, $S D=5$ ) had a slightly higher frequency (count per minute) than those in the light condition ( $M=17.64, S D=$ 4.6), although this difference was not statistically significant, $t(29)=1.67, p=.1$. Similar to the manual activity analysis and to previous research (Ruff \& Lawson, 1990; Yu \& Smith, 2015), we defined sustained looking as any looking event that lasted for more than 3 seconds. As shown in Fig. 5, children in the heavy condition had significantly more short (less than 3 seconds) looking events per six minutes (session length) than did children in the light condition (Heavy $=1789$, Light $=1625$ ); in contrast, the number of sustained looking events per six minutes were comparable between conditions (Heavy $=350$, Light $=425$ ). Chi-square test of independence indicated a significant relationship between the number of sustained looking events and the weight of objects, $\chi^{2}(1, \mathrm{~N}=$ $4189)=13.25, \mathrm{p}=.0003$.

Overall, the results of the looking patterns mirror the results from the manual activity: children in the heavy condition produced more rapid but frequent manual activity events, as well as more rapid but frequent looking events. By our hypothesis, the dynamic hand-eye coordination is responsible for the corresponding differences in the hand and eye patterns in the two conditions.

## Hand-eye coordination

We propose that the result-that heavy condition had more short and rapid manual activity events, as well as more short and rapid looking events than the light condition-is driven by the hand-eye coordination of the child. In other words, because child's hands and eyes are closely coupled such that when hands are on the object, the eyes are also more likely to be on the same object-sustained hand actions create and support sustained visual attention. To demonstrate this link, we measured the durations of joint hand-eye to the same object. If this is the case, then we would expect to see more short but rapid hand-eye coordination events-the hands and eyes of the child were on the same object-in the heavy than the light condition.

As predicted, the mean duration of hand-eye coordination events was significantly lower in the heavy condition ( $M=$ $1.04 \mathrm{~s}, S D=0.25 \mathrm{~s})$ than the light condition $(M=1.33 \mathrm{~s}, S D=$ $0.44 \mathrm{~s}), t(29)=2.26, p=.03$. However, the hand-eye coordination events in the heavy condition $(M=17.83, S D=$ 3.93) had a significantly higher frequency (count per minute) than those in the light condition $(M=14.19, S D=3.26)$, $t$ $(29)=2.55, p=.01$. Again, we used 3 seconds as the threshold to define sustained hand-eye coordination event and found that children in the heavy condition had significantly more short hand-eye coordination events per six minutes (session length) than did children in the light condition $($ Heavy $=1577$, Light $=1116)$; in contrast, the number of sustained hand-eye coordination events per six minutes were comparable between conditions (Heavy $=134$, Light $=151$ ). Chi-square test of independence showed a significant relationship between the number of sustained hand-eye coordination events and the weight of objects, $\chi^{2}(1$, $\mathrm{N}=2978)=14.05, p=.0002$.

## General Discussion

When actively engaged with objects-the context for much real-world learning and problem solving-infants' visual attention is dynamically tied to their hand actions. The implications of this for the development of visual attention and for the underlying brain mechanisms are profound: this sensory motor coordination could be a core driving force for visual development, setting up the behavioral and neural networks for the mature control of visual attention (Byrge, Smith, \& Sporns, 2014).

The direct connection between bodily movement, gaze direction and internal cognitive processing has been supported in many studies of adults' cognition. For example, it has been shown that bodily movement or direction of eye gaze serves as the basis for establishing deictic (pointing) reference to objects as well as the spatial relations between objects, suggesting that visual attention and action may share overlapping spatial referent frames (Ballard et al, 1997; Yuan, Uttal, \& Franconeri, 2016). Manual actions can also directly guide or bias visual attention. The position of hands elicits unique neural responses in several brain areas and serves to prioritize visual attention (Makin, Holmes, \& Zohary, 2007). Using a visual covert-orienting paradigm, for example, Reed, Grubb and Steele (2006) have shown that placing a hand on the side of the screen where a target would appear facilitated target detection, but the presence of visual anchors did not produce the same effect. This result suggests that adults have a hand-centered representation within peripersonal space (i.e., space that is close to a person's body), raising the possibility that children may have a similar or even stronger hand-centered representation in near space as they had shorter arms than adults and often hold objects very close to their body.

Manual action is a crucial way through which infants select and learn about the visual properties of objects in the world. Despite the complex and often cluttered real-world learning environment, the ego-centric view of infants suggests that they often attend to one dominant object at one time (Yu et al., 2009), which is crucial for developing visual attention to the detailed properties of objects. Importantly, although social partner occasionally brings an object in front of an infant's face, the predominant pathway through which infants create this optimal learning moments is through his or her own hand actions - it is hand actions that bring objects closer to the body and eyes, allowing for close examination of the various properties of the object, multiple sampling of the dynamic views of objects, leading to sustained visual attention and helping to build representations of the threedimensional structure of objects (Bambach, Crandall, Smith, \& Yu, 2016; Soska, Adolph, and Johnson, 2010).

The current study offers another pathway through which manual action exerts influence on visual attention-by changing the frequency and duration of looking events. Because hands and eyes are closely synchronized during play, the temporal characteristics of manual actions can influence those of vision. An analogous example is the demonstration that auditory input, particularly the rhythm of sounds, can facilitate visual learning for both adults (Iordanescu, Guzman-Martinez, Grabowecky, \& Suzuki, 2008) and children (Bahrick \& Lickliter, 2000). This multimodal learning not only provides redundant information to recruit sustained attention, but also capitalizes on the interconnection among sensory modalities-activities in one domain can influence and promote that of another domain. In this sense, the sensory motor coordination is the core driving force for the development of cognitive capacities.

Hand-eye coordination can help to build and integrate multiple neural networks that underpin cognitive development. Time-locked signals from perception and action not only afford the direct mapping between the physical properties of the object to the neuronal activity of the visual network, between the physical properties of the object to the neuronal activity of the haptic system, but also allow for cross-modality integration and enrichment: activity of the visual system and the activity of the haptic system are directly mapped to each other (Edelman, 1987; Smith \& Gasser, 2005). For example, as one holds and manipulates an object, the neuronal activity of the visual system is timelocked to the activity of the haptic system-each different hold is linked to each unique visual representation of the object. As a result, a particular sight of an object may elicit its corresponding neuronally mapped action. For instance, in one visual recognition task, adults were shown a picture of a pitcher and answered the question "Is this a pitcher" by pressing either a left or a right button. Adults responded faster when the "yes" button was on the same side of the pitcher's handle, suggesting that the sight of the object may have elicited corresponding motor activity, facilitating the motor execution of button press on the same side (Ellis \& Tucker, 2000).

This multimodal learning mechanism has important implications for development and learning. For example, manual actions can be leveraged to train a mature control of visual attention. One classic demonstration of this idea is the A-not-B error (Piaget, 1954). After repeating the sequence of seeing one object being hidden at location A and retrieving the object at location A several times, infants were shown the object being hidden at location $B$. Despite seeing the object being hidden in the new location B , infants continued searching at location A. However, changes in the motor actions or giving infants more motor experiences led to improved performance (Bertenthal, Campos, \& Barrett, 1984). For instance, changing the manual behavior that children need to perform to approach the object through changes in posture (sitting vs. standing) led children to search at the correct location much often (Smith, Thelen, Titzer, \& McLin, 1999).

In conclusion, the current study supports the developmental systems view of visual attention: visual attention emerges from the interaction among multiple sensory modalities, which are dynamically coordinated during moment-by-moment perception and action events to support cognitive development. In particular, the current study showed that changes in manual behavior alter the patterns of toddlers' visual attention during toy play. Further, we provided evidence that the hand-eye coordination is the underlying mechanism: toddlers' hands and eyes were dynamically coupled, such that when hands were on an object, the eyes were also likely to be on the same object. These results have implications for the research and development of visual attention, as well as the possibility to leverage on manual action as a way for training the control of visual attention.

## Acknowledgments

We thank Melissa Elston, Charlotte, Wozniak, Melissa Hall, Charlene Tay, Madison Philips and Seth Foster for data collection. This research was supported by National Institutes of Health Grants R01 HD074601 and R21 EY017843.

## References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., \& Rao, R. P. (1997). Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences, 20(04), 723-742.
Bahrick, L. E., \& Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. Developmental psychology, 36(2), 190.

Bambach, S., Crandall, D. J., Smith, L. B., \& Yu, C. (2016). Active Viewing in Toddlers Facilitates Visual Object Learning: An Egocentric Vision Approach.
Bertenthal, B. I., Campos, J. J., \& Barrett, K. C. (1984). Selfproduced locomotion. In Continuities and discontinuities in development (pp. 175-210). Springer US.
Byrge, L., Sporns, O., \& Smith, L. B. (2014). Developmental process emerges from extended brain-body-behavior networks. Trends in cognitive sciences, 18(8), 395-403.
Colombo, J. (2001). The development of visual attention in infancy. Annual Review of Psychology, 52, 337-367.
Edelman, G. M. (1987). Neural Darwinism: The theory of neuronal group selection. Basic Books.
Ellis, R., \& Tucker, M. (2000). Micro- affordance: The potentiation of components of action by seen objects. British journal of psychology, 91(4), 451-471.
Franchak, J. M., Kretch, K. S., Soska, K. C., \& Adolph, K. E. (2011). Head-Mounted Eye Tracking: A New Method to Describe Infant Looking. Child development, 82(6), 17381750.

Gibson, J. J. (1979). The ecological approach to visual perception. Boston, MA: Houghton Mifflin.
Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., \& Suzuki, S. (2008). Characteristic sounds facilitate visual search. Psychonomic Bulletin \& Review, 15(3), 548-554.
Kannass, K. N., Oakes, L. M., \& Shaddy, D. J. (2006). A longitudinal investigation of the development of attention and distractibility. Journal of Cognition and Development, 7(3), 381-409.
Koterba, E., Leezenbaum, N. B., \& Iverson, J. M. (2014). Object exploration at 6 and 9 months in infants with and without risk for autism. Autism, 18(2), 97-105.
Lansink, J. M., Mintz, S., \& Richards, J. E. (2000). The distribution of infant attention during object examination. Developmental Science, 3(2), 163-170.
Makin, T. R., Holmes, N. P., \& Zohary, E. (2007). Is that near my hand? Multisensory representation of peripersonal space in human intraparietal sulcus. Journal of Neuroscience, 27(4), 731-740.
Pereira, A. F., James, K. H., Jones, S. S., \& Smith, L. B. (2010). Early biases and developmental changes in selfgenerated object views. Journal of Vision, $10(11), 1-13$.

Pereira, A. F., Smith, L. B., \& Yu, C. (2014). A bottom-up view of toddler word learning. Psychonomic Bulletin \& Review, 21(1), 178-85.
Piaget, J. (1954). The construction of reality in the child. New York: Basic Books.
Quittner, A. L., Smith, L. B., Osberger, M. J., Mitchell, T. V, \& Katz, D. B. (1994). The impact of audition on the development of visual attention. Psychological Science, 5(6), 347-353.
Reed, C. L., Grubb, J. D., \& Steele, C. (2006). Hands up: attentional prioritization of space near the hand. Journal of Experimental Psychology: Human Perception and Performance, 32(1), 166.
Ruff, H. A., \& Lawson, K. R. (1990). Development of sustained, focused attention in young children during free play. Developmental Psychology, 26(1), 85-93.
Smith, L., \& Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. Artificial life, 11(1-2), 13-29.
Smith, L. B., Thelen, E., Titzer, R., \& McLin, D. (1999). Knowing in the context of acting: the task dynamics of the A-not-B error. Psychological review, 106(2), 235.
Soska, K. C., Adolph, K. E., \& Johnson, S. P. (2010). Systems in development: motor skill acquisition facilitates three-dimensional object completion. Developmental psychology, 46(1), 129.Mullette-Gillman, Cohen, \& Groh, 2005.

Thelen, E., \& Smith, L. B. (1994). A dynamic systems approach to the development of cognition and action. Cambridge, MA: MIT Press.
Yu, C., \& Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. PloS one, 8(11), e79659.
Yu, C., \& Smith, L. B. (2014). Linking Joint Attention with Hand-Eye Coordination - A Sensorimotor Approach to Understanding Child-Parent Social Interaction, 27632768.

Yu, C., \& Smith, L. B. (2016). The Social Origins of Sustained Attention in One-Year-Old Human Infants. Current Biology, 1-6.
Yu, C., \& Smith, L. B. (2016). Multiple Sensory-Motor Pathways Lead to Coordinated Visual Attention. Cognitive Science, 1-27.
Yu, C., Smith.L.B., Shen, H., Pereira, A. F.,\& Smith,T.(2009). Active information selection:Visual attention through the hands. Autonomous Mental Development, IEEE Transactions on, 1(2), 141-151.
Yuan, L., Uttal, D., \& Franconeri, S. (2016). Are Categorical Spatial Relations Encoded by Shifting Visual Attention between Objects? PloS one, 11(10), e0163141.

# Mindshaping the world can make mindreading tractable: Bridging the gap between philosophy and computational complexity analysis 

Andrea Zeppi (azeppi@unime.it)<br>Department of Cognitive Science, University of Messina Messina, Italy<br>Mark Blokpoel (m.blokpoel@psych.ru.nl)<br>Department of Artificial Intelligence, Radboud University<br>Nijmegen, The Netherlands


#### Abstract

It is often assumed that the socio-cultural context positively influences mindreading performances. Among the available theories, mindshaping is proposed to consist of cultural mechanisms that make the social domain homogeneous and, hence, easier to interpret. Proponents of the mindshaping hypothesis claim that homogeneity is responsible for the computational tractability of mindreading, which is otherwise intractable. In this paper, we examine this core claim of mindshaping and investigate how homogeneity influences mindreading tractability. By taking action understanding as a case-study for mindreading, we formally operationalize mindshaping homogeneity in different ways with the goal of bridging the gap between informal claims and formal (in)tractability results. The analysis shows that only specific combinations of homogeneity may lead to tractable mindreading, whilst others do not. Additionally, the analysis reveals the possibility of a yet undiscovered mindshaping mechanism.


Keywords: mindshaping; mindreading; computational intractability; culture; goal inference; conceptual/philosophical analysis; computational modeling

## Introduction

The ability to understand what motivates other people's behavior is often considered a defining capacity of human cognition. Theories of this mindreading capacity, however, are challenged with explaining how humans can interpret behaviors in a timely manner, because the available theories are often computationally intractable (Alechina \& Logan, 2010; Apperly, 2010; Zawidzki, 2013). As Gigerenzer and colleagues proposed:

> "The computations postulated by a model of cognition need to be tractable in the real world in which people live, not only in the small world of an experiment with only a few cues. This eliminates NP-hard models that lead to computational explosion, ..." Gigerenzer (2008)

Given that mindreading is performed in a complex, realworld socio-cultural environment (Adams et al., 2010; PerezZapata, Slaughter, \& Henry, 2016; Tomasello, Carpenter, Call, Behne, \& Moll, 2005), it stands to reason that intractable (NP-hard) theories of mindreading cannot explain how humans can perform the computations postulated by the theory quickly.

In an attempt to address this theoretical paradox, Zawidzki has proposed that the socio-cultural environment plays a key role. Zawidzki proposes that this environment is shaped by agents themselves so that mindreading can be tractable (Mameli, 2001; Zawidzki, 2008, 2013). Introduced as the mindshaping hypothesis, this claim entails a collection of (social) cognitive and evolutionary mechanisms that bring structure to the environment.

In this paper, we assess the potential that the mindshaping hypothesis has to solve the intractability paradox of mindreading. Given that computational (in)tractability is a well-defined mathematical property of computational-level theories (Marr, 1982; van Rooij, 2008), a bridge will have to be built between Zawidzki's informal theoretical contributions and formal complexity-theoretic results. We propose that such a bridge can be built by taking action understanding as a special-case proxy for evaluating how mindshaping mechanisms may pave the way for tractable mindreading.

In order to do this, we take two steps. First, we analyze the mindshaping hypothesis and extract specific claims about the effects that mindshaping mechanisms may have on the structure of the socio-cultural environment and consequently the (in)tractability of mindreading. Second, we assess the plausibility of the claims identified in the first step by operationalizing the mindshaping structuring effects in a computational-level model of action understanding (viz., Bayesian inverse planning; Baker, Saxe, \& Tenenbaum, 2009; Baker, Tenenbaum, \& Saxe, 2008). This allows us to relate mindshaping effects to formal (in)tractability results (Blokpoel, Kwisthout, van der Weide, Wareham, \& van Rooij, 2013). The analysis will show that only certain combinations of mindshaping effects lead to tractable mindreading, whilst other effects do not. Furthermore, the analysis also suggests the possibility of a novel effect that is necessary for tractability, which may lead to the discovery of new mindshaping mechanisms.

## Mindreading as abductive inference

Several theoretical accounts of mindreading have been proposed. Fast and frugal heuristics theories (Chater, Oaksford, Nakisa, \& Redington, 2003) conjecture that humans can understand what motivates other people's behavior through simple cue-based rules. Simulation theory
(Goldman, 2006) conjectures that we understand external behaviors by the means of mental simulations. In this paper, we focus on a third hypothesis proposed by Zawidzki. By rejecting modularity, acknowledging domain-generality and the relevance problem (Fodor, 1983; Heal, 1996) for social reasoning, Zawidzki implicitly accepts isotropy and with it a mindreading account that is inferential in nature (Zawidzki, 2013). ${ }^{1}$ Under this view, mindreading can be construed as a mapping from an observed socio-cultural environment (consisting of observed behaviors, actions and context) and social knowledge to the intentional attributions that best explain the observed social environment. Unfortunately, with the notion that mindreading is inferential also comes intractability.

Zawidzki attributes intractability of mindreading to the problem of holism/isotropy. Because in principle any information that a person has might be relevant for any inference that is made, every possibility must be considered (Fodor, 2001). The intractability of inferential mindreading is corroborated by the fact that many of our best theories of inferential cognitive capacities are computationally intractable (NP-hard or worse) (Cherniak, 1986; Frixione, 2001; Levesque, 1988; Thagard \& Verbeurgt, 1998; Tsotos, 1990; van Rooij, 2008; van Rooij \& Wareham, 2012).

An intractable theory makes unrealistic (exponential or worse) demands on computational resources (van Rooij, 2008). Hence, such a theory cannot satisfactorily explain why people can 'mindread' as quick as they do (see Table 1). This leads to the paradox mentioned in the introduction. However, rather than rejecting the inferential mindreading account altogether, the paradox may be resolved if the effects of mindshaping mechanisms are adequately fleshed out. In the next section, we discuss the different possible structuring effects that mindshaping mechanisms may have on the environment.

## Deconstructing mindshaping

First proposed by Mameli (2001) and later developed by Zawidzki (2009; 2008; 2013), the mindshaping hypothesis proposes that the success of human social cognition is explained by the (evolutionary) development of behavioral mechanisms that "shape our socio-cultural environment in ways that make coordination exponentially more tractable" (Zawidzki, 2013). Examples of mindshaping mechanisms are: imitation, over-imitation, the chameleon effect (Chartrand \& Bargh, 1999), pedagogy, norm following and self-constituting narratives (Zawidzki 2013). Although these mechanisms are individually quite different, they all implement a form of social expectancy and conformity mechanism that 'mindshape' both the socio-cultural environment and social knowledge through biased transmission and selection of behaviors. Under the

[^252]assumption that mindreading is a capacity that operates on the social environment and social knowledge, mindshaping may potentially have a positive effect on the tractability of mindreading. To assess this claim, however, it is necessary to characterize in more detail the different kinds of effects that mindshaping may have on the input of mindreading.

Table 1: An illustration of polynomial and exponential time requirements. The input size corresponds to the size of the representation of the input (e.g., the observed social environment and social knowledge encoded in a Bayesian network). The other columns illustrate the difference between tractable (i.e., polynomial time) intractable (i.e., exponential time or worse). The time required to compute intractable theories quickly outgrows the age of our universe.

| Input <br> size $\boldsymbol{n}$ | Polynomial time <br> required $\boldsymbol{n}^{\mathbf{2}}$ | Exponential time <br> required $\boldsymbol{2}^{\boldsymbol{n}}$ |
| ---: | ---: | ---: |
| $\mathbf{5}$ | $0,25 \mathrm{msec}$. | $0,32 \mathrm{msec}$. |
| $\mathbf{1 0}$ | 1 msec. | 10 msec. |
| $\mathbf{2 0}$ | 4 msec. | $10,5 \mathrm{sec}$. |
| $\mathbf{5 0}$ | 25 msec. | 130312 days |
| $\mathbf{1 0 0}$ | $0,10 \mathrm{sec}$. | $4 \times 10^{17}$ years |
| $\mathbf{2 5 0}$ | $0,63 \mathrm{sec}$. | $5,7 \times 10^{62}$ years |
| $\mathbf{5 0 0}$ | $2,50 \mathrm{sec}$. | $1,0 \times 10^{138}$ years |

Unfortunately, the literature only vaguely provides such a characterization of the effects of mindshaping, which is thought to 'homogenize' the social environment and knowledge to make mindreading easier. Such a characterization is insufficient as it states only the ultimate effect of mindshaping (i.e., tractability of mindreading), which is exactly that which needs to be explained. In order to unravel if and how mindshaping can render mindreading tractable, we need to understand two things. First, we need to understand that by putting constraints on the input of a mindreading, those constraints may render mindreading tractable. Second, we need to understand how mindshaping can implement such constraints through homogeneity.

## How homogeneity should affect mindreading

The main claim put forth by mindshaping is that mindshaping mechanisms positively affect the reliability of mindreading (Zawidzki, 2013). The term reliability, however, conflates two different meanings: accuracy and tractability. In order for mindreading to be reliable, inferred propositional attitudes need to be good and the computations need to be performed in a short amount of time (tractability). Perhaps counterintuitively, accuracy does not always cause intractability. An intractable function can be extremely inaccurate and it is also possible for a tractable function to,
bathwater. If, as originally proposed, we can show how Mindshaping mechanisms can render inferential mindreading tractable, it would strengthen the plausibility of the mindshaping account whether it is separate of mindreading or not.
instead, be accurate. Even approximate accuracy (compared to e.g., optimality) does not necessarily grant tractability (van Rooij \& Wareham, 2012). It seems, therefore, that the reason mindreading is tractable for humans lies not in mindreading trading off accuracy, but in the homogeneity effect that mindshaping has.

Research has focused on characterizing the phylogenetic, ontogenetic and cultural evolution of human social cognition (Mameli, 2001; Zawidzki, 2008, 2013). Although the computational details are underspecified, the mindshaping hypothesis clearly states that the tractability of mindreading is not obtained by altering the mindreading capacity, but by changing the socio-cultural environment and social knowledge on which mindreading operates. Mindshaping mechanisms are hence not modules (Zawidzki, 2009, 2013). Therefore, for the purpose of assessing the claim about mindreading tractability, they can be considered complementary to mindreading. This allows us to abstract away from the evolutionary mechanisms that underlie mindshaping and focus on their homogeneity effect.

The solution for tractable mindreading lies "not within human mind readers, but, rather, outside of them" (Zawidzki, 2009). The idea is promising, since it is known that some intractable functions $f: I \rightarrow O$ can be tractable when their input domain is constrained $f^{\prime}: I^{\prime} \rightarrow O$, where $I^{\prime} \subset I$ (Downey \& Fellows, 1999; van Rooij, 2008). These constraints can be the result of naturally occurring or 'mindshaped' structure in the world. They are formally defined as restrictions on properties of the input of computational-level models, called parameters (e.g., see Figure 1). When the tractability of a function is obtained through such restrictions it is said to be fixedparameter tractable for that subset of the input. Fixedparameter tractability, however, is a formal, mathematical property of computational-level models. In order to assess if homogeneity can render mindreading tractable, we need a formal computational-level model of mindreading. Although such an account does not yet exist for mindreading in general, we can investigate the tractability claim using a special-case capacity for mindreading. In the next section, we take (inferential) action understanding as a special case of (inferential) mindreading and present possible ways of operationalizing homogeneity in Bayesian inverse planning (Baker, Saxe, Tenenbaum 2009). We then show that only certain combinations of homogeneity effects render Bayesian inverse planning tractable, whilst other do not.

## Bridging homogeneity to tractability of Bayesian inverse planning

The ability to understand what goals underlie the actions of others is a prime example of the human capacity to mindread. In the Bayesian inverse planning model (Baker et al. 2009), action understanding is characterized as inferring the most probable goal given observed social behavior. In other words, a mapping from an observed socio-cultural environment (consisting of observed behaviors, actions and context) and social knowledge (about planning) to the intentional attributions (goals) that best explain the
observations. Table 2 compares the input and output domains of mindreading and Bayesian inverse planning and Figure 1 illustrates the Bayesian network that underlies Bayesian inverse planning.

Figure 1. In Bayesian inverse planning knowledge about planning is represented by state, action and goal variables (circles) and the probabilistic dependencies between them (arrows). Each variable has a domain (boxes). The input of the model consists of such a network and observed states and actions (gray variables). The output is the most likely value assignment to the goal variables. Several parameters
are: The number of goals $|G|$, the number of observed actions $|A|$, the maximum number of values a goal variable can have $g$ and the maximum number of values an action variable can have $a$.


Table 2. Comparing the input and output of the mindreading capacity with those of the special case Bayesian inverse planning model.

|  | Mindreading | Bayesian inverse <br> planning |
| :--- | :--- | :--- |
| Input | observed socio- <br> cultural environment <br> and | observed states, <br> actions <br> and |
| social knowledge | knowledge about <br> planning encoded in a <br> Bayesian network |  |
| Output | intentional <br> attributions | most probable goals, <br> given the input |

Like Bayesian inference (Chater, Tenenbaum, \& Yuille, 2006; Martignon \& Hoffrage, 2002), Bayesian inverse planning is computationally intractable in general (Blokpoel et al., 2013). This is consistent with the idea that (inferential) mindreading is computationally intractable too. Blokpoel et al. (2013), however, used formal analysis to prove that when certain constraining assumptions are made on the input of Bayesian inverse planning, it becomes tractable. To investigate whether or not a more homogeneous sociocultural environment and more homogeneous social knowledge may lead to tractable mindreading, we have to build a bridge all the way to Bayesian inverse planning. We
first start by illustrating possible interpretations of homogeneity. We then operationalize these interpretations in Bayesian inverse planning, and finally relate the operationalized homogeneity effects with known computational tractability results.

## Interpreting homogeneity effects

There are two types of homogeneity that are consistent with the mindshaping literature: Cognitive homogeneity and mindshaping homogeneity. This analysis is a first attempt and by no means exhaustive, i.e., more homogeneity effects may be postulated/discovered in the future. For example, our analysis will point to possible novel homogeneity effect that is not yet covered by a mindshaping mechanism.

## Cognitive homogeneity

Since all humans have similar biological and cognitive systems, one could argue that humans also share the majority of their propositional attitudes. For example, if all humans behave rationally, they all have the same (rational) bias when deciding how to act to achieve a goal.

Cognitive homogeneity mechanisms may result in a population where sets of intentions overlap a lot (i.e., most of people's possible intentions are shared; CH-S). This, however, does not necessarily restrict the number of possible intentions, as the shared set can still be very large.

Furthermore, Zawidzki argues that tractability of mindreading cannot be achieved only by cognitive homogeneity. This seems to make sense from a computational perspective as well. Even if, for example, all humans mindread 'rationally' this does not explain why mindreading is tractable. And even if all humans share most of their intentions, that set can still be extremely big. Other restrictions are needed to provide any computational benefits.

## Mindshaping homogeneity

Mindshaping homogeneity is effected by a set of mechanisms either cognitive, cultural or evolutionary that decrease the heterogeneity of the socio-cultural environment and social knowledge. For example, by having a culture that keeps reinforcing the same knowledge and behaviors through pedagogy, norms enforcement and imitation (Zawidzki 2013), the knowledge in that population can become more and more homogeneous over generations.

Mindshaping homogeneity can be conjectured to result in the following restrictions as a consequence of biased transmission of knowledge over generations:

- Biased transmission can restrict on the number of available behaviors in a population (MH-B);
- Biased transmission can make social knowledge (i.e., the relations between behaviors and intentions) less ambiguous (MH-A).
- Mindshaping mechanisms resulting in ritualization phenomena may limit, regardless of the total number of possible actions available, the number of executed and observed actions (MH-R).
- Habitualization and culture codification may further limit the complexity of what people can achieve to facilitate social understanding (MH-C).
In the mindshaping literature, the focus has been on identifying the nature of the mindshaping mechanisms that lead to homogeneity. Here instead, we focus on the actual contribution of mindshaping mechanisms and the relative homogeneity. Hence, we assume the validity of mindshaping as a starting point, together with homogeneity, and investigate if their effect can render an intractable model of a mindreading capacity tractable.


## Operationalizing homogeneity effects in Bayesian inverse planning

Taking Bayesian inverse planning as our case-study we can now investigate the possible input-restrictions that can result from the mindshaping hypothesis for action understanding, and show which of the homogeneity-based restrictions may lead to tractability of action understanding. To this end, we build the final part of our bridge by linking the homogeneitybased restrictions to input restrictions of the Bayesian inverse planning model. The effect of these restrictions on the (in)tractability has been investigated by Blokpoel et al. (2013). Table 3 and Figure 1 provide an overview of the five parameters they analyzed.

The above-mentioned parameterizations of Bayesian inverse planning and the previously given interpretations of homogeneity make it possible to operationalize homogeneity. Our contribution is to provide these operationalizations as restrictions on input parameters for Bayesian inverse planning. We go beyond what is currently in the literature by fleshing out more detailed effects that mindshaping might have.

Table 3. Possible parameters for the Bayesian inverse planning model (taken from Blokpoel et al. 2013) and their associated homogeneity hypothesis.

|  | Homogeneity | Description |
| :--- | :--- | :--- |
| $\|\mathbf{A}\|$ | MH-R (partly) | The number of actions that are <br> observed by an interpreter. |
| $\boldsymbol{a} \mid$ | MH-C | The number of goal variables <br> that are inferred by an <br> interpreter. |
| $\boldsymbol{g}$ | unknown | The number of available actions <br> values per action variable. <br> The number of available values <br> per goal variable. |
| $\mathbf{1 - p}$ | MH-A | The probability of the most <br> likely goal inference, dependent <br> on the probabilistic knowledge <br> encoded in the Bayesian <br> network. |

## Restricting the number of observed actions $|A|$

Parameter $|A|$ defines the number of actions that an interpreter observes in order to infer the underlying goal.

Mindshaping mechanisms at work in phenomena like ritualization may limit the number of executed and observed actions. However, ritualized behavior may only explain why $|A|$ may be small in those codified situations. Action understanding transcends those cases. If MH-R proponents are committed to small $|A|$ in general, the account needs to be strengthened. Regardless, restricting $|A|$ does not lead to any known tractability result.

## Restricting the number of inferred goals $|G|$

If action understanding is to be tractable, then one option is for $|G|$, the number of possible goals that an interpreter actually pursues, to be small (together with $g$ ). If, within a social community, an actor would like his/her actions to be timely interpretable to others, then this actor might pursue few goals at a given time so as to make $|G|$ small. This behavior might be the result of mindshaping mechanisms such as habitualization, culture and phenomena like ritualization (MH-C).

## Restricting the domain of actions $a$

Parameter $a$ can be seen as the maximum number of possible actions that are available at any point in time. This number is upper-bounded by the total number of possible actions that are available to a person (MH-B).

## Restricting the domain of goals $g$

Parameter $g$ can be seen as the maximum number of possible goal attributions available for the given inference. This number is upper-bounded by the total number of intentions available to an agent. One might argue that sharing the same intentions (CH-S) may lead to a restriction on $g$, but it does not as humans may in principle share even an infinitely large set. If anything restricts $g$, it seems there must be some not yet discovered mindshaping process that does so, or another cognitive process that selects the relevant intentions from the set of all possible intentions. The latter, however, would imply solving the relevance problem (Fodor, 1983, 2001; Pylyshyn, 1989) which is notoriously hard but perhaps the solution lies in a combination of mindshaping and cognitive relevance selection. Due to the ubiquity of $g$ in tractability results discovering these processes would be paramount for having a complete picture of the relation between homogeneity and tractability of mindreading.

## Restricting the ambiguity of behavior to make 1-p low

The relational probabilities between variables can be seen as encoding the social knowledge that is brought to bear when inferring the most probable goal. The prior probabilities of variables can be seen as the disposition a person has towards particular unobserved variables (such as goals) at the time of the inference. Together, these probabilistic relations between variables and the prior probability of variables may be shaped

[^253]by pedagogy, norm following and imitation such that, 1-p is low (MH-A).

## Computational complexity of Bayesian inverse planning in 'mindshaped' worlds

Blokpoel et al. (2013) proved several computational complexity results for Bayesian inverse planning. ${ }^{2}$ These results show that tractability is not easily achieved. Even restricting multiple parameters simultaneously does not necessarily render the model computationally tractable. The following two intractability results prove that either by themselves or in combination, these restrictions do not make Bayesian inverse planning tractable:

1. Restricting $|A|, a$, and $|G|$ simultaneously, or
2. Restricting $|A|, a$ and $g$ simultaneously

Importantly, none of the parameters by themselves render Bayesian inverse planning tractable. ${ }^{3}$ Only when the right combination of parameters is restricted, i.e., only when the world is mindshaped in the right way, Bayesian inverse planning does become tractable. The following two results show that if either (3) or (4) or both conditions hold, then action understanding is tractable.
3. Restricting $|G|$ and $g$, and/or
4. Restricting $1-p$ and $g$

These two tractability results show that in principle, under the correctly (mind)shaped conditions, Bayesian inverse planning can be tractable. However, the results also reveal (at least for the restricted case of action understanding) a gap in the mindshaping theory. While one of the main claims of mindshaping is the importance of homogeneity for the tractability of mindreading, homogeneity alone cannot (yet) fully explain tractability. All known tractability results show that a restriction on $g$ is always necessary for tractability, but no known mindshaping process leads to that restriction.

## Discussion

Explaining why people can understand what motivates other people's behavior quickly is, at least from a computational perspective, not trivial. Often, culture or evolution are used to trivialize the paradox of mindreading intractability and explain the speed at which people understand the social world around them. These ideas are embodied by Zawidzki's mindshaping hypothesis. In our analysis, we have shown that, a bridge can be built between mindshaping and a special-case capacity for mindreading, a lot of ground still needs to be covered if these ideas are to fully deal with the intractability paradox.

By detailing possible interpretations of mindshaping effects and relating those to known (in)tractability results for a computational model of action understanding, we have

[^254]shown that only very specific combinations of mindshaping effects have the potential to explain the performance of human mindreading.

The analysis has also revealed that a restriction on the number of available intentions (specifically, the maximum number of possible goal attributions) is a necessary condition for tractability. At the same time, no clear homogeneity effect leads to this restriction. Theoreticians interested in computationally explaining the speed of human mindreading through mindshaping may look for mindshaping mechanisms that specifically lead to this constraint.

Even for a restricted case of mindreading such as action understanding, some of these restrictions have an effect on the tractability of this capacity, while others do not. It stands to reason that caution is in order when claims about tractability are concerned. While not exhaustive, our analysis can be seen as a structured attempt at capturing philosophical and psychological claims about the influence of culture on mindreading into a systematic computational framework.

## References

Adams, R., Rule, N., Franklin, R., Wang, E., Stevenson, M., Yoshikawa, S., ... Ambady, N. (2010). Cross-cultural reading the mind in the eyes: an fMRI investigation. Journal of Cognitive Neuroscience, 22(1), 97-108.
Alechina, N., \& Logan, B. (2010). Belief ascription under bounded resources. Synthese, 173(2), 179-197.
Apperly, I. (2010). Mindreaders: the cognitive basis of* theory of mind'. Psychology Press.
Baker, C. L., Saxe, R., \& Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3), 329-349.
Baker, C. L., Tenenbaum, J. B., \& Saxe, R. R. (2008). Bayesian models of human action understanding. Consciousness and Cognition, 17(1), 136-144.
Blokpoel, M., Kwisthout, J., van der Weide, T. P., Wareham, T., \& van Rooij, I. (2013). A computational-level explanation of the speed of goal inference. Journal of Mathematical Psychology, 57(3-4), 117-133.
Chartrand, T. L., \& Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology, 76(6), 893.
Chater, N., Oaksford, M., Nakisa, R., \& Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. Organizational Behavior and Human Decision Processes, 90(1), 63-86.
Chater, N., Tenenbaum, J. B., \& Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. Trends in Cognitive Sciences, 10(7), 287291.

Cherniak, C. (1986). Limits for knowledge. Philosophical Studies, 49(1), 1-18.
Downey, R. G., \& Fellows, M. R. (1999). Parameterized Complexity. New York, NY: Springer New York.
Fodor, J. (1983). The Modularity of Mind. (Z. W. Pylyshyn, W. Demopoulos, Z. W. E. Pylyshyn, \& W. E.

Demopoulos, Eds.)Philosophical Review (Vol. 94). MIT Press.
Fodor, J. (2001). The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology. Representation and mind (Vol. 10). MIT Press.
Frixione, M. (2001). Tractable competence. Minds and Machines, 11(3), 379-397.
Gigerenzer, G. (2008). Why heuristics work. Perspectives on Psychological Science, 3(1), 20-29.
Goldman, A. I. (2006). Simulating Minds. Philosophical Books (Vol. 49).
Heal, J. (1996). Simulation, theory, and content. Theories of Theories of Mind, 75-89.
Levesque, H. J. (1988). Logic and the complexity of reasoning. Journal of Philosophical Logic, 17(4), 355389.

Mameli, M. (2001). Mindreading, Mindshaping, and Evolution, 597-628.
Marr, D. (1982). Vision. San Francisco: W.H. Freeman and Company.
Martignon, L., \& Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. Theory and Decision, 52(1), 29-71.
Perez-Zapata, D., Slaughter, V., \& Henry, J. D. (2016). Cultural effects on mindreading. Cognition, 146, 410414.

Pylyshyn, Z. W. (1989). The Robot's Dilemma: The Frame Problem in Artificial Intelligence. Norwood: Ablec Publishing Corporation.
Thagard, P., \& Verbeurgt, K. (1998). Coherence as constraint satisfaction. Cognitive Science, 22(1), 1-24.
Tomasello, M., Carpenter, M., Call, J., Behne, T., \& Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28(5), 675-691.
Tsotos, J. K. (1990). Analyzing vision at the complexity level. Behavioral and Brain Sciences, Behavioral(13), 423-469.
van Rooij, I. (2008). The Tractable Cognition Thesis. Cognitive Science: A Multidisciplinary Journal, 32(6), 939-984.
van Rooij, I., \& Wareham, T. (2012). Intractability and approximation of optimization theories of cognition. Journal of Mathematical Psychology, 56(4), 232-247.
Zawidzki, T. (forthcoming). Mindshaping. In A. Newen, L. de Bruin, \& S. Gallagher (Eds.), Oxford Handbook of $4 E$ Cognition. Oxford University Press.
Zawidzki, T. (2008). The function of folk psychology: mind reading or mind shaping? Philosophical Explorations, 11(3).
Zawidzki, T. (2009). Theory of mind, computational tractability, and mind shaping: 2009 Performance Metrics for Intelligent Systems Workshop. In Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems (pp. 149-154). ACM.
Zawidzki, T. (2013). Mindshaping: A New framework for understanding human social cognition. MIT Press.

# Using mouse-tracking data to visualise decision landscapes 

Arkady Zgonnikov (arkady.zgonnikov@nuigalway.ie)<br>School of Psychology, National University of Ireland, Galway

Andrea Aleni (andrea.aleni@gmail.com)<br>Department of Electrical Engineering and Information Technology University of Naples Federico II, Italy

Petri Piiroinen (petri.piiroinen@nuigalway.ie)
School of Mathematics, Statistics \& Applied Mathematics National University of Ireland, Galway

Denis O’Hora (denis.ohora@nuigalway.ie)
School of Psychology, National University of Ireland, Galway


#### Abstract

Computerised paradigms have enabled decision making researchers to gather rich data on human behaviour, including information on motor execution of a decision, e.g., by tracking mouse cursor trajectories. As the number and complexity of mouse-tracking studies rapidly increase, more sophisticated methodology is needed to analyse the decision trajectories. Here we present a new computational approach to generating decision landscape visualisations based on mouse-tracking data. Decision landscape is an analogue of energy potential field mathematically derived from velocity of mouse movement during a decision. Visualised as a 3D surface, it provides a comprehensive overview of motor evolution of decisions. Employing the dynamical systems theory framework, we develop a new method for generating decision landscapes based on arbitrary number of trajectories. The decision landscape visualisation have potential to become a novel tool for analysing mouse trajectories during decision execution, which can provide new insights into the dynamics of decision making.


Keywords: decision making; mouse tracking; dynamical systems; visualisation

## Introduction

Traditionally, decision making studies have been focused on what people choose and explaining the mechanisms leading to the observed choice outcome distributions. Sequential sampling models (Busemeyer \& Townsend, 1993; Ratcliff \& Rouder, 1998) make predictions on the time required for the decision maker to arrive to a decision, thereby stimulating empirical research to measure response times in addition to choice outcomes. However, in the past decade, decision making researchers have been employing more advanced experimental paradigms, measuring behavioural activity during decision-making to investigate whether how we choose is meaningfully related to what we choose.

A variety of experimental methods have been used to study the cognitive processes underlying decision making. One class of paradigms, including eye tracking (Orquin \& Loose, 2013) and different variations of information search paradigm (Payne, 1976), taps attentional processes, trying to answer the question of what information is attended to in the course of decision. Another strand of research, focused on
hand or mouse tracking, examines how decisions are executed through motor system. These studies interpret motor output of a decision as a continuous trace of decisional processes. In a typical experiment on mouse tracking, the participant chooses between the two options presented in the corners of a computer screen (Fig. 1). The dynamics of the response, as expressed in recorded mouse cursor trajectories, can then reveal the degree of competition between the two options during choice.

Mouse (or hand) tracking have been employed to investigate decision making dynamics in a variety of different domains, e.g., speech processing (Spivey, Grosjean, \& Knoblich, 2005), social categorisation (Freeman, Ambady, Rule, \& Johnson, 2008; Freeman \& Ambady, 2011), and intertemporal choice (Dshemuchadse, Scherbaum, \& Goschke, 2013; O'Hora, Carey, Kervick, Crowley, \& Dabrowski, 2016). The recent development of "off-the-shelf" solutions for capturing mouse cursor data (Freeman \& Ambady, 2010; Kieslich, Wulff, Henninger, \& Haslbeck, 2017) has further increased the amount and complexity of the data generated by mouse-tracking studies. However, the vast majority of available studies utilise only few basic measures derived from rich mouse-tracking data. At the same time, much potentially important information conveyed by mouse trajectories is still very often ignored. More advanced analysis methods can potentially enable us to get deeper insights from the rich data provided by the mouse-tracking paradigm.

Here we present a new computational approach for illustrating mouse-tracking data via three-dimensional visualisations of decision landscape. Recently, a method was suggested to infer such visualisations directly from the data (O’Hora, Dale, Piiroinen, \& Connolly, 2013). Here, we build up on this work, employing an alternative, modelbased approach. We assume that the decision process, as reflected by a mouse trajectory, is governed by a 'potential energy' landscape. The parameters defining the form of the landscape for each decision can then be fitted to capture specific mouse motion trajectory during that decision. Visualised as a 3D surface, this decision landscape provides a com-


Figure 1: Typical setup of a mouse-tracking experiment. Plotted in green is an actual mouse trajectory representing binary choice in a learning task (O’Hora et al., 2013)


Figure 2: Mouse trajectory from Fig. 1 and a hypothetical decision landscape driving the decision process. The grey line is a projection of the trajectory on the decision landscape. The figure is based in part on visualisation of hypothetical decision attractor manifold by Spivey and Dale (2006).
prehensive overview of motor evolution of decisions. The suggested method can generate illustrations of decision landscapes based on arbitrary number of trajectories. Using previously collected data on a learning task (O'Hora et al., 2013), we demonstrate how decision landscape visualisations can be used to compare sets of mouse trajectories between experimental conditions or individual decision makers in a comprehensive, visually appealing way.

## Visualising decision landscape

The proposed method is aimed at reconstructing a 3D decision landscape based on a mouse trajectory of a decision (or a set of trajectories). We assume that each trajectory can be described by a dynamical system of a specific form, which incorporates a parametrised function describing the shape of the two-well landscape. By fitting this dynamical system to a set of trajectories, we obtain specific values of the parameters characterising these particular trajectories. We can then use these parameters to generate the 3D visualisation of the decision landscape characterising all of the given decisions. The source code implementing all the procedures of the method in Python is available via Open Science Framework (Zgonnikov, Aleni, Piiroinen, O’Hora, \& di Bernardo, 2017).

## Data requirements and preprocessing

To visualise decision landscapes, we use two-dimensional trajectories obtained in a typical mouse-tracking experiment (Fig. 1). We assume that each decision trajectory starts in the bottom centre part of the screen and ends in either top left or top right corner. The method can be generalised to the case of more than two choice options; with minor adjustments (e.g., using two-dimensional projections of 3D trajectories) it can as well be used with any other experimental paradigm generating simple enough continuous trajectories, for instance, arm reaching (Song \& Nakayama, 2008; Gallivan \& Chapman, 2014).

Importantly, the current version of the method assumes continuity and smoothness of the trajectories, which is not always the case. In a fraction of experimental trials, participants change their mind in the course of a trial, which is indicated by major shifts in the $x$-direction of a decision trajectory (Resulaj, Kiani, Wolpert, \& Shadlen, 2009; Freeman, 2014). The deterministic dynamical model of a decision trajectory utilised in the present method does not account for such changes-of-mind, so we recommend using the method only with the trajectories without abrupt shifts.

As screen size and proportions can differ between experiments, we illustrate the method for spatially normalised trajectories. In particular, the screen coordinates are assumed to be rescaled such that each trajectory originates near $(x, y)=$ $(0,0)$ and ends in the vicinity of $(x, y)=(-1,1)$ (left target) or $(x, y)=(1,1)$ (right target).

Each trajectory is supposed to be described by the time series of $x$ - and $y$-coordinates of the mouse cursor on the screen. In addition to this data, the method also requires $x$ - and $y$ velocities of the mouse cursor at each time step, which are computed numerically.

## Model of trajectory dynamics

Without aiming at developing a model explaining the dynamics of a decision, we use a simple dynamical system to describe the decision trajectory, capturing the high-level features of motion of the mouse cursor. We describe the time dependent $x$ - and $y$-components of a decision trajectory by a system of ordinary differential equations

$$
\begin{align*}
\tau \dot{x} & =-\frac{\partial V}{\partial x} \\
\tau \dot{y} & =-\frac{\partial V}{\partial y} \tag{1}
\end{align*}
$$

where $\dot{x}=\frac{\mathrm{d} x}{\mathrm{~d} t}$ and $\dot{y}=\frac{d y}{d t}$ are the time derivatives of $x$ and $y, \tau>0$ is the time scale parameter expressed in seconds, $V(x, y)$ is a function describing the decision landscape, which defines the dynamics of the system, and $\frac{\partial V}{\partial x}$ and $\frac{\partial V}{\partial y}$ are its partial derivatives with respect to $x$ and $y$.

Our method is not constrained by some particular function $V(x, y)$; here we use one of the simplest possible variants. We assume that $V(x, y)$ comprises a fixed baseline component
$V_{x}(x)+V_{y}(y)$ and a parametrised component $V_{x y}(x, y)$ so that

$$
\begin{equation*}
V(x, y)=V_{x}(x)+V_{y}(y)+V_{x y}(x, y) \tag{2}
\end{equation*}
$$

where $V_{x}(x)$ and $V_{y}(y)$ are polynomials chosen in a way that the two target locations, $(-1,1)$ and $(1,1)$, are attracting steady states ("attractors"), and the starting location $(0,0)$ is a repelling state of the system (1) given that $V_{x y}(x, y) \equiv 0$. Thus,

$$
\begin{gather*}
V_{x}(x)=\int \frac{\partial V_{x}}{\partial x} \mathrm{~d} x=\int x(x+1)(x-1) \mathrm{d} x=\frac{x^{4}}{4}-\frac{x^{2}}{2}  \tag{3}\\
V_{y}(y)=\int \frac{\partial V_{y}}{\partial y} \mathrm{~d} y=\int y(y-1) \mathrm{d} y=\frac{y^{3}}{3}-\frac{y^{2}}{2} \tag{4}
\end{gather*}
$$

Having a two-attractor decision landscape as a baseline, we introduce the parametrised polynomial component $V_{x y}(x, y)$ to be able to account for asymmetry in the landscape and other, more intricate properties of experimental trajectories. Here, also for the reason of simplicity, for $V_{x y}(x, y)$ we use a polynomial function of $x$ and $y$

$$
\begin{equation*}
V_{x y}(x, y)=\sum_{k=2}^{\alpha} \sum_{\substack{i, j>0 \\ i+j=k}} c_{i j} x^{i} y^{j} /(k-1), \tag{5}
\end{equation*}
$$

where the parameter $\alpha \geq 2$ determines the number of terms in the polynomial, which in turn regulates flexibility of the model, and the coefficients $c_{i j}$ are fitted to the data. Note that with increasing $\alpha$ the number of free parameters increases, thus the fitted values of these parameters may be difficult to interpret for large $\alpha$. We recommend using the method with $\alpha=2,3$, or 4 , depending on the complexity of the trajectories, and to take into account the trade-off between approximation accuracy and interpretability of the parameters.

The effect of the model parameters $\tau$ and $c_{i j}$ on the shape of the decision landscape can be analysed independently of the experimental data (Fig. 3). For any $\alpha$, two parameters always enter the model, $\tau$ and $c_{11}$. The parameter $\tau$ affects the characteristic time scale of the system motion: the larger the value of $\tau$, the slower the motion of the mouse generated by the model (in both directions). In what follows we use the baseline value $\tau=0.05 \mathrm{~s}$.

The parameter $c_{11}$, corresponding to the only second-order polynomial term of the model, is the primary determinant of the asymmetry of the decision landscape. Such asymmetry may be caused, for instance, by strong prevalence of one decision outcome over the other. Another possible example of asymmetry would be a situation when the trajectories towards one option are consistently faster compared to the trajectories pointing to the other option, with the two options being chosen equally likely.

When $\alpha \geq 3$, additional polynomial terms enter $V_{x, y}(x, y)$. The effects of the parameters $c_{i j}$ in front of these higher-order terms are somewhat similar to those of $\tau$ and $c_{11}$, but allow for much finer tuning of the decision landscape to the experimental trajectories.

## Fitting the model to the single trajectory

For a single experimental trajectory, we aim to find the parameters allowing the model (1)-(5) to reproduce this trajectory as closely as possible. We can define the fitting error in two ways: as a function of the positional difference between the data and the modelled trajectory, or based on the difference in mouse velocities between the data and the model. The first approach would result in a more accurate approximation of trajectories, but requires substantially more computational time, as on each step of the fitting algorithm the system of differential equations (1) has to be integrated numerically. The second approach, employing the velocity-based fitting error, is much more efficient in terms of computational resources, sometimes at the expense of approximation accuracy. Here we focus on the latter; the supplied source code implements both approaches.

Given an experimental mouse trajectory sampled at $m$ time steps and the numerically derived mouse velocities $\left\{v_{x}^{\text {data }}\right.$, $\left.v_{y}^{\text {data }}\right\}$, we define the fitting error

$$
\begin{align*}
& H\left(\tau, c_{i j},\left\{v_{x}^{\mathrm{data}}, v_{y}^{\mathrm{data}}\right\}\right)= \\
& \frac{1}{m} \sum_{i=0}^{m}\left(v_{x}^{\mathrm{model}}\left(x_{i}, y_{i}\right)-v_{x}^{\mathrm{data}}\left(t_{i}\right)\right)^{2}+\left(v_{y}^{\operatorname{model}}\left(x_{i}, y_{i}\right)-v_{y}^{\mathrm{data}}\left(t_{i}\right)\right) \tag{6}
\end{align*}
$$

where $v_{x, y}^{\text {model }}\left(x_{i}, y_{i}\right)$ are the values of the right-hand side of the system (1) computed at each point $\left(x_{i}, y_{i}\right)$ along the experimental trajectory. These values depend on the current parametrisation of the model, so the defined error function depends both on the model parameter values and the experimental trajectory.

Using numerical optimisation routines (available, e.g., in the Python package scipy.optimize), one can find the values of the model parameters $\tau, c_{i j}$ minimising function (6) for a given mouse trajectory. These parameters are substituted in Eq. (5), which, along with Eqs. (2)-(4), fully specifies $V(x, y)$. The 3D plot of the function $V(x, y) / \tau$ then visualises the decision landscape representing the original trajectory.

## Fitting to multiple trajectories

Visualising a decision landscape that would integrate the properties of multiple trials (within a single experimental condition, individual participant, or a group of participants) is where the method can prove most useful. To be able to do this, we use the same approach as in the case of a single trial, and minimise the average error across individual trials in a set of trials. Given the set of $N$ trajectories and their velocities $\left\{v_{x}^{\text {data }}, v_{y}^{\text {data }}\right\}_{n=1}^{N}$, the fitting error for multiple trajectories is defined by

$$
\begin{equation*}
\hat{H}\left(\tau, c_{i j},\left\{v_{x}^{\mathrm{data}}, v_{y}^{\mathrm{data}}\right\}_{n=1}^{N}\right)=\frac{1}{N} \sum_{n=1}^{N} H\left(\tau, c_{i j},\left\{v_{x}^{\mathrm{data}}, v_{y}^{\mathrm{data}}\right\}_{n}\right), \tag{7}
\end{equation*}
$$

where $H$ is defined in (6).


Figure 3: Changes in the baseline decision landscape depending on the parameters of the model for $\alpha=2$. In each panel, all parameters except for the one in the panel title are fixed at the baseline levels $\tau=0.05, c_{i, j}=0$; the baseline landscape is shown in grey colour. In this and the following figures, the black marble marks the starting location of a trial.

## Example scenarios

We illustrate several potential scenarios of using our method to visualise mouse tracking data by applying it to the previously obtained experimental data on a simple learning task (O'Hora et al., 2013). The task consisted of a series of binary choices between abstract symbols, with each symbol yielding either low or high reward (e.g., 5 or 20 points). The goal of the participants was to get as many points as possible throughout a set of 36 trials, which included low vs. low, high vs. low, and high vs. high choices. By the end of the experiment, most of the participants successfully learned to choose only the symbols associated with high reward.

Here we only consider part of the dataset corresponding to high vs. low choices, so that there is always a "correct" choice. The data are preprocessed so that the correct ("high") option is mapped to right-hand corner of the screen, and incorrect option is located in the left-hand corner. To fit the experimental data, we used the version of the model with $\alpha=4$, which has seven free parameters. The baseline values of the parameters were set to $\tau=0.05, c_{i j}=0$.

Fig. 4 shows the example decision landscapes obtained for three trajectories generated by Participant 444 . The shape of the fitted landscape changes depending on the dynamics of the decision. Two key properties of a mouse trajectory reflected by the fitted decision surface are: motion time, i.e., the how long it takes for the cursor to reach the response area once it leaves the starting location, and maximum deviation of the trajectory from the ideal, straight-line trajectory (termed "max-d").

Importantly, the decision landscape is supposed to capture dynamics rather than geometry of the mouse trajectory. However, with increasing deviation of the trajectory towards unchosen option, the strength of the attractor corresponding to that option increases. Moreover, in extreme cases, when deviation towards competitor option is very large, the attractor representing the unchosen option can be even stronger than the attractor corresponding to the eventually chosen option (yellow surface in Fig. 4). This situation, paradoxically indicating that the attraction towards the unchosen option during


Figure 4: Decision landscapes of three representative trials of Participant 444. In all three trials, the subject had chosen the option on the right ("correct choice"). Trial 2 (yellow surface) was slow (motion time 0.72 s ) and mildly conflicted (max-d 0.23 ). In trial 15 (purple surface), the mouse trajectory was fast (motion time 0.15 s ) and was close to straight line (maxd 0.02). In trial 28 (green surface), the trajectory was still relatively fast (motion time 0.37 s ), but substantially curved towards the unchosen option (max-d 0.77).


Figure 5: Decision landscapes of the two representative participants. Participant 2 (blue surface) had chosen right- and left-hand-side options 7 times each (mean motion times 0.56 and 0.45 s). Red surface illustrates Participant 3 with the right-hand-side option chosen 3 times (mean motion time 0.5 s) and the left-hand-side option chosen 10 times (mean motion time 0.4 s ).
the trial was stronger than towards the chosen one, is very rare and happens only for landscapes of individual trajectories with high max-d.

One of the potential applications of the present method is highlighting individual differences between two participants performing the same task. To do this, one needs to obtain decision landscapes individually for each participant by fitting the model to all trials of that individual simultaneously. In the case of multiple-trajectory fitting, the fitting error is defined as the average error across all trajectories of a given participant, so the resulting decision landscape will integrate the information on how often and how fast each option was chosen, providing a comprehensive overview of the participant's decisions throughout the experiment.

We illustrate this by comparing the decision landscapes of the two participants (Fig. 5). Participant 2 was equally likely to chose either option, with faster trajectories reaching towards the incorrect (left-hand-side) option. Participant 3 had chosen the incorrect (left) response more often, and it was chosen on average faster than the correct option.

If the experimental task involves adaptation, the current method can be used to highlight learning patterns within subjects. Fig. 6 represents three decision landscapes separately fitted to the trajectories of three consecutive blocks of trials of Participant 4 (with each block containing six to eight trials). The decision surface gradually changes from the twoattractor landscape slightly favouring the right attractor to the single-attractor configuration, thereby tracking the learninginduced evolution of preference across blocks.

## Discussion

The decision landscape visualisations provide comprehensive overview of a mouse-tracking data. Each visualisation integrates the information on 1) the likelihood of each option to


Figure 6: Evolution of decisions of Participant 4 throughout three consecutive blocks of trials. Each surface corresponds to all trials of a block.
be chosen, 2) the duration of the response, and 3) the degree of competition between the options.

The first attempt to develop an algorithm to visualise decision landscapes based on mouse tracking data has been recently made by O'Hora et al. (2013). Their approach, however, is purely data-driven, and thus requires hundreds of trajectories to generate a reliable visualisation of the decision surface. By incorporating prior assumptions about decision landscape $V(x, y)$ into a parametrised model, we dramatically reduce the data requirements of our method. As demonstrated above, the method proposed here can be used with as few trajectories as one, but can also incorporate arbitrary number of trajectories.

One of the main limitations of the proposed method is its focus on trajectories without changes-of-mind. An important (although relatively rare) class of decisions are those involving preference reversals during choice execution. These happen even in simple perceptual discrimination tasks (Resulaj et al., 2009); in more cognitively demanding tasks the frequency of changes-of-mind can increase up to $20 \%$ (Freeman, 2014). Developing a way of visualising decision landscapes of the change-of-mind trajectories is one of our foremost future research directions.

Attractor models have proved useful in understanding outcomes of cognitive processes such as categorisation (Tuller, Case, Ding, \& Kelso, 1994), risky decision making (van Rooij, Favela, Malone, \& Richardson, 2013), and binary decision making in intermittent motor control (Zgonnikov \& Lubashevsky, 2015). This work is among the first attempts to apply the concepts of dynamical systems theory to process data characterising decisions. We hope that the proposed decision landscape visualisation approach will eventually grow in a new tool for analysing decision trajectories, which will be able to provide new insights into dynamics of decision making.

## Acknowledgements

This work was supported in part by Government of Ireland Postdoctoral Fellowship GOIPD/2015/481.

## References

Busemeyer, J. R., \& Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological review, 100(3), 432.

Dshemuchadse, M., Scherbaum, S., \& Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. Journal of Experimental Psychology: General, 142(1), 93.
Freeman, J. B. (2014). Abrupt category shifts during realtime person perception. Psychonomic bulletin \& review, 21(1), 85-92.
Freeman, J. B., \& Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behavior Research Methods, 42(1), 226-241.
Freeman, J. B., \& Ambady, N. (2011). A dynamic interactive theory of person construal. Psychological review, 118(2), 247.

Freeman, J. B., Ambady, N., Rule, N. O., \& Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. Journal of Experimental Psychology: General, 137(4), 673.
Gallivan, J. P., \& Chapman, C. S. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. Frontiers in Neuroscience, 8 .
Kieslich, P. J., Wulff, D. U., Henninger, F., \& Haslbeck, J. M. B. (2017). Mousetrap: An R package for processing and analyzing mouse-tracking data. Retrieved from https://doi.org/10.5281/zenodo. 290283 doi: 10.5281/zenodo. 290283

O’Hora, D., Carey, R., Kervick, A., Crowley, D., \& Dabrowski, M. (2016). Decisions in Motion: Decision Dynamics during Intertemporal Choice reflect Subjective Evaluation of Delayed Rewards. Scientific Reports, 6, 20740.

O'Hora, D., Dale, R., Piiroinen, P. T., \& Connolly, F. (2013). Local dynamics in decision making: The evolution of preference within and across decisions. Scientific reports, 3, 2210.

Orquin, J. L., \& Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. Acta psychologica, 144(1), 190-206.
Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. Organizational behavior and human performance, 16(2), 366-387.
Ratcliff, R., \& Rouder, J. N. (1998). Modeling response times for two-choice decisions. Psychological Science, 9(5), 347-356.
Resulaj, A., Kiani, R., Wolpert, D. M., \& Shadlen, M. N. (2009). Changes of mind in decision-making. Nature, 461(7261), 263-266.

Song, J.-H., \& Nakayama, K. (2008). Numeric comparison in a visually-guided manual reaching task. Cognition, 106(2), 994-1003.
Spivey, M. J., \& Dale, R. (2006). Continuous dynamics in real-time cognition. Current Directions in Psychological Science, 15(5), 207-211.
Spivey, M. J., Grosjean, M., \& Knoblich, G. (2005). Continuous attraction toward phonological competitors. Proceedings of the National Academy of Sciences of the United States of America, 102(29), 10393-10398.
Tuller, B., Case, P., Ding, M., \& Kelso, J. (1994). The nonlinear dynamics of speech categorization. Journal of Experimental Psychology: Human Perception and Performance, 20(1), 3.
van Rooij, M. M., Favela, L. H., Malone, M., \& Richardson, M. J. (2013). Modeling the Dynamics of Risky Choice. Ecological Psychology, 25(3), 293-303.
Zgonnikov, A., Aleni, A., Piiroinen, P. T., O’Hora, D., \& di Bernardo, M. (2017, Mar). Visualising decision landscapes using mouse-tracking data. Open Science Framework. Retrieved from https://osf.io/5q364/
Zgonnikov, A., \& Lubashevsky, I. (2015). Double-well dynamics of noise-driven control activation in human intermittent control: the case of stick balancing. Cognitive Processing, 16(4), 351-358.

# Insomniacs Misidentify Angry Faces as Fearful Faces Because of Missing the Eyes: An Eye-Tracking Study 

Jinxiao Zhang (tepzhang@hku.hk)<br>Department of Psychology, The University of Hong Kong, Pok Fu Lam, Hong Kong<br>Antoni B. Chan (abchan@cityu.edu.hk)<br>Department of Computer Science, The City University of Hong Kong, Kowloon Tong, Hong Kong<br>Esther Y.Y. Lau (laueyy@eduhk.hk)<br>Department of Psychology, The Education University of Hong Kong, Tai Po, Hong Kong<br>Janet H. Hsiao (jhsiao@hku.hk)<br>Department of Psychology, The University of Hong Kong, Pok Fu Lam, Hong Kong


#### Abstract

Insomniacs were found to have compromised perception of facial expressions. Through eye movement examinations, here we test the hypothesis that this effect is due to impaired visual attention functions for retrieving diagnostic features in facial expression judgments. 23 individuals with insomnia symptoms and 23 non-insomniac controls completed a task to categorize happy, sad, fearful, and angry faces. The insomniacs were less accurate to recognize angry faces and made more "fearful" mistakes than controls. A hidden Markov modeling approach for eye movement data analysis revealed that when recognizing angry faces, more insomniacs adopted an eye movement pattern focusing on the mouth while more controls adopted a pattern attending to both the eyes and the mouth. This result is consistent with previous findings that the primary diagnostic feature for recognizing angry faces is the eyes suggesting that impaired information selection through visual attention control may account for the compromised emotion perception in insomniac individuals.


Keywords: insomnia; eye-tracking; hidden Markov model; facial expression

## Introduction

Insomnia is closely related to emotional disorders such as anxiety and depression (see Baglioni et al., 2010 for a review). In particular, compromised perception of emotional facial expressions, which has an important role in one's socioemotional functioning, has been frequently found among sleep-deprived individuals or those with insomnia symptoms. For example, individuals with physiological insomnia were found to perceive fearful and sad faces as less emotional compared with good sleepers (Kyle et al., 2014). Another study found that 31.5 -hour sleep deprivation led to less accurate recognition of sad faces (Cote et al., 2014). In addition, an fMRI study found that sleep deprivation made participants more likely to classify facial expressions as angry, and this effect was coupled with their diminished neural discrimination between threatening and non-threatening stimuli (in the anterior cingulate and anterior insula; Goldstein-Piekarski et al., 2015). Nevertheless, the underlying mechanism for the disturbed perception of emotional facial expressions among insomniacs remains unclear (Kyle et al., 2014).

In addition to emotional functioning, insomniacs and sleep-deprived individuals are commonly found to have impaired performance in visuospatial attention tasks
(Marchetti et al., 2006), and this behavioral impairment is associated with attenuated activation in the attention neural network comprising the prefrontal, parietal, and cingulate cortex (Tomasi et al., 2009; Mander et al., 2008). The impairment in the attention network may have a profound impact on cognitive performance in general, as it can significantly influence how task relevant information is selected. Indeed, attenuated activations in the attention network were reported to be associated with less explorative eye movement patterns and worse performance in face recognition (Chan et al., 2016). It is thus possible that insomniac individuals adopt different eye movement patterns from good sleepers in emotional facial expression judgments as a result of their impaired visual attention control, leading to compromised recognition performance.

Here we aim to investigate the role of visual attention functions in accounting for insomniacs' compromised identification of emotional facial expressions through eye tracking. Individuals with insomnia symptoms and noninsomniac controls completed an emotional facial expression judgement task in which they were required to recognize emotional facial expressions and rate the emotional intensity with eye tracking. Recent studies have suggested four basic facial expressions that are recognized across cultures: 'happy', 'sad', 'fear' and 'anger' (Jack et al., 2016). Accordingly, here we examine participants' perception of these four facial expressions. Previous research did not consistently find disturbed perception of a particular facial expression in sleep-deprived or insomniac individuals. However, most of the expressions reported to be affected were negative expressions (e.g., sadness in Cote et al., 2014; anger in Goldstein-Piekarski et al., 2015; sadness and fear in Kyle et al., 2014). Thus, here we hypothesize that insomniac individuals may be less accurate in recognizing the negative facial expressions than noninsomniac controls, and that this behavioral difference may be associated with differences in eye movement patterns adopted by the two groups.

While eye movements are important measures for visual attention functions, recent studies have reported substantial individual differences in eye movements in visual tasks (e.g. Kanan et al., 2015), which were not adequately reflected in most of the current analysis methods. In view of this, Chuk, Chan, and Hsiao (2014) have recently proposed a Hidden

Markov Model (HMM, a type of time-series probabilistic model in machine learning) based approach for analyzing eye movement data. This approach assumes that the current eye fixation during a task is conditioned on previous fixations. Thus, eye movements in the task can be considered a Markov process, which can be better understood using HMM. In this approach, each individual's eye movements are modeled with an HMM, including both person-specific regions of interests (ROIs) and transitions among the ROIs. Thus, it reflects individual differences in both spatial and temporal dimensions of eye movements. Individual HMMs can be clustered to discover common patterns among individuals (Coviello, Chan, \& Lanckriet, 2014), and similarities between individual eye movement patterns can be quantitatively assessed by calculating the likelihoods of the patterns being generated by a given HMM. Thus, this approach is especially suitable for examining the relationship between eye movement patterns and other outcome measures such as task performance (e.g., Chuk, Chan, \& Hsiao, in press). Here we aim to apply this method to examine the relationship among insomnia, eye movements, and performance in facial expression categorization. We hypothesize that there may be more insomniac individuals adopting an eye movement pattern that overlooks diagnostic features for the negative facial expressions than non-insomniac controls, and participants' likelihoods of adopting this pattern may be associated with their performance in recognizing the negative expressions.

## Methods

## Participants

23 individuals with insomnia symptoms and 23 noninsomniac controls classified by the Sleep Condition Indicator (SCI, Espie et al., 2014) were recruited (Table 1). The SCI consists of 8 items concerning an individual's sleep condition during the recent month in a 0-4 Likert-style scale. The Chinese SCI has been validated and recommended as a screening tool for clinical insomnia with an original cut-off at 21/22 (Wong et al., 2017). To increase the contrast between the two groups, individuals with SCI scores $\leq 19$ were classified as individuals with insomnia symptoms, and those with SCI scores $\geq 24$ were classified as non-insomnia controls. Participants in the two groups were individually matched in gender and age. They were ethnically Chinese from Hong Kong and right-handed (Edinburgh Handedness Inventory, EHI; Oldfield, 1971). They had normal or corrected-to-normal vision and no history of head trauma or psychiatric conditions.

Table 1. Participants' demographics and sleep conditions

|  | Controls <br> $(n=23)$ | INS <br> $(n=23)$ | Comparison test |
| :--- | :---: | :---: | :---: |
| Age $(M+S D)$ | $18.91 \pm 0.90$ | $18.74 \pm 0.81$ | $t(44)=.689$ |
| Gender | $30.43 \%$ | $30.43 \%$ | $\chi^{2}(1)=0$ |
| $(\%$ male $)$ |  | $15.70 \pm 2.94$ | $t(44)=16.54^{* *}$ |
| SCI $(M \pm S D)$ | $27.61 \pm 1.80$ | 1 |  |

SCI: Sleep Condition Indicator; INS: insomnia group.

* $p<.05 ; * * p<.01$


## Design \& Procedures

Participants completed an emotional facial expression judgment task adapted from Kyle et al. (2014), which required participants to categorize and rate emotional intensity of 4 facial expressions (i.e. happiness, sadness, fear, and anger; Figure 1A). The task consisted of 2 blocks with 40 trials in each block ( 10 trials for each expression). In each trial, a solid dot first appeared at the screen center for drift correction, and it was replaced by a fixation cross for 500 ms . Once a fixation was detected at the cross at the end of the 500 ms , a color picture of an Asian individual's face with an emotional facial expression (Figure 1B; 450 x 600 pixels) was presented either above or below the center of the screen until the participants categorized it as a happy, sad, fearful, or angry face by pressing corresponding buttons. Participants were asked to respond as quickly as possible. After a $250-\mathrm{ms}$ pause, they were asked to rate the emotional intensity of the facial expression on a 6-point scale, ranging from " 1 -not very intense" to " 6 -extremely intense". Half of the face images were male faces; the faces spanned around $8^{\circ}$ of visual angle at the viewing distance of 60 cm (Hsiao \& Cottrell, 2008). The participants had a practice of 24 trials at the beginning of the task.

Participants' eye movements were recorded by an EyeLink 1000 eye tracker. The standard 9-point calibration procedure was used at the beginning of each block and was repeated whenever the drift-correction error was larger than $1^{\circ}$ of visual angle. The tracking mode was pupil and corneal reflection and the sampling rate was 2000 Hz . A chinrest was used during the task to reduce head movements.


Figure 1. (A) A demonstration of a trial with an angry face. (B) The average pictures of the 20 stimuli used in each emotional expression. From left to right: happiness, sadness, fear, anger.

The eye movement data were analyzed using the EMHMM (Eye Movement analysis with Hidden Markov Models, http://visal.cs.cityu.edu.hk/research/emhmm/; Chuk et al., 2014) approach. Each participant's eye movements while viewing one type of facial expressions was summarized with an HMM. The optimal number of ROIs for each model was determined automatically through a variational Bayesian approach, by selecting the model with the highest marginal likelihood. For each facial expression, we clustered all individual HMMs into 2 groups to reveal common patterns. We then examined the distributions of the insomniacs and the controls in the 2 common patterns, and
the correlations between their likelihoods of adopting each pattern and behavioural performances.

## Results

## Behavioral results

A 2 (group: insomnia vs. control) by 4 (emotion: happy, sad, fearful and angry) repeated measures ANOVA revealed a significant interaction between group and emotion on the accuracy of the facial expression judgment task, $F(3,132)=$ 2.68, $p=.049, \eta^{2}=.057$. Independent t -tests between the insomnia and the control group in each emotion condition indicated that there was $8.4 \%$ higher accuracy on average to recognize angry faces in the control group than the insomnia group, $t(44)=2.12, p=.039, d=.63$ (Figure 2A). This group difference was not found in other emotion conditions, $p s>.05$. When we examined the responses participants made towards angry faces (Figure 2B), the 2 by 4 repeated measures ANOVA indicated a significant interaction between group and emotion, $F(3,132)=3.57, p=.016, \eta^{2}=$ .075. Post-hoc between-group t-tests showed a significantly higher percentage of "fearful" responses in the insomnia group than the control group, $t(44)=2.07, p=.045, d=.62$.


Figure 2. (A) The accuracy to categorize happy, sad, fearful, and angry facial emotions in the control and the insomnia group. (B) Reponses made while angry faces were presented. (C) Response time to accurately categorize emotional facial expressions. (D) Emotional intensity rating of the 4 facial emotions in the two groups. (* $p<.05 ; \dagger .05<p<.10$; error bars: 1 s.e.m.)

In the correct response time (RT) data (Figure 2C), a 2 (group) by 4 (emotion) repeated measures ANOVA indicated a main effect of emotion (adjusted post-hoc comparisons: happy < sad, fear, and anger), $F(3,132)=$ 40.72, $p<.001, \eta^{2}=.48$. This effect did not interact with group. When we examined the difference between the two groups in categorizing different expressions separately, the insomnia group responded marginally slower in identifying angry faces than the control group, $t(44)=1.768, p=.084, d$ $=.53$. The group by emotion repeated measures ANOVA on
emotional intensity rating showed a main effect of emotion (adjusted post-hoc comparisons: fearful > happy, angry > sad; Figure 2D), $F(3,132)=23.24, p<.001, \eta^{2}=.35$. However, this effect did not interact with group.

## Eye movement data

We modeled each participant's eye movements for viewing each type of facial expressions with an HMM. For each expression type, we clustered all participants' HMMs into 2 representative patterns and examined the distributions of the insomniacs and controls adopting the 2 patterns. We observed that the insomnia and control groups differed significantly in their frequencies of adopting the 2 representative patterns when viewing angry faces. Consistent with our behavioral data, this difference was not observed in viewing other expressions ${ }^{1}$. Figure 3A and 3B show the 2 representative patterns. The 3 ROIs were in red, green, and blue respectively and the table showed the priors (the probability of the first fixation being located at an ROI) of the ROIs and the transition probabilities among them. In the eye-mouth pattern (Pattern 1; $n=21$ ), the first fixation was most likely to be in eye region (red ROI, $47 \%$ ) or the mouth region (blue ROI, 42\%). The next fixation from the eye ROI had a high probability to stay in the same (eye) ROI, whereas the next fixation from the mouth ROI had a $26 \%$ probability to move to the eye region. Thus, participants adopting this pattern focused the most on the eye region, followed by the mouth region, while viewing angry faces. In the nose-mouth pattern (Pattern 2; $n=25$ ), the first fixation was most likely to be at the nose/face center (red ROI, $46 \%$ ) or the mouth and chin region (green ROI, $42 \%$ ). The next fixation following the first was most likely to stay in the same ROI as the first fixation (> 99\%), suggesting few transitions among the ROIs. Thus, participants adopting this pattern tended to focus on the lower part of the face (nose, mouth and chin) but neglect the eye region. Figure 3C shows the difference heat map between the two patterns: the eye-mouth pattern had more fixations on the right eye region (warm colors) while the nose-mouth pattern had more fixations on the mouth and chin regions (cold colors). This clear separation of the eye movement patterns demonstrated well the power of machine learning methods.

Importantly, significantly more insomniacs adopted the nose-mouth eye movement pattern, and more controls adopted the eye-mouth pattern (Figure 3D), $\chi^{2}(1)=4.29, p=$ .038. In addition, insomniacs' eye movement patterns had higher similarities to the nose-mouth pattern than those from controls, $t(44)=-2.37, p=.022, d=.713$, as measured in the log-likelihoods of their eye movement patterns being generated by the HMM of the nose-moth pattern. In contrast, there was no significant difference in the similarity of eye movement patterns to the eye-mouth pattern between insomniacs and controls, $p>.05$.

[^255]
Pattern 1: eyes + mouth

|  | To Red | To Green | To Blue |
| :--- | :---: | :---: | :---: |
| Priors | .47 | .12 | .41 |
| From Red | .94 | .06 | .00 |
| From Green | .26 | .01 | .73 |
| From Blue | .04 | .01 | .95 |



| D |  |  |  |
| :--- | :---: | :---: | :---: |
| No. of <br> individuals | Controls | Insomniacs | Total |
| Pattern 1 | 14 <br> $(60.9 \%$ <br> controls) | 7 <br> $(30.4 \%$ IIS $)$ | 21 |
| Pattern 2 | 9 <br> $(39.1 \%$ <br> controls) | 16 <br> $(69.6 \% ~ I I S)$ | 25 |
| Total | 23 | 23 | 46 |

Figure 3. (A and B) The eye-mouth and nose-mouth representative eye movement patterns for viewing angry faces as the result of clustering. Images from left to right: 3 ROIs, actual assignments of the fixations to the ROIs, and heat map of eye fixations. The tables contain the priors and transition probabilities of the ROIs. (C) Difference map of actual fixations between the two patterns. (D) Distribution of the eye-mouth (Pattern 1) and nose-mouth (Pattern 2) patterns in the insomnia group and the control group.

To demonstrate the advantage of using the EMHMM approach to reveal these differences, we plotted the heat maps of the fixations of the insomnia and control groups and the difference map between them (Caldara \& Miellet, 2011). As can be seen in Figure 4 difference map, the significantly different areas (circled in white) were scattered and not easily interpretable. This phenomenon was due to significant individual differences in eye movement patterns within each group. The EMHMM approach allows us to identify the eye-mouth and nose-mouth patterns in a data driven fashion and clearly reveal the difference between the two participant groups while accounting for individual differences in eye movement patterns.

When we examined the relationship between eye movement patterns and performances in the facial expression judgment task, we found that the log-likelihood of participants' patterns being classified as the nose-mouth pattern was positively correlated the percentage of "fearful"
responses (Figure 5A), $r=.423, p=.003^{2}$ : the more similar the pattern to the nose-mouth pattern when viewing angry faces, the more "fearful" mistakes made. In addition, in the control group, the log-likelihood of being classified as the eye-mouth pattern was negatively correlated with the correct RT of identifying angry faces (Figure 5B), $r=-.451, p$ $=.031$ : the more similar the pattern to the eye-mouth pattern, the faster the correct RT. This correlation was not significant in the insomnia group, $p>.05$.


Figure 4. Fixation heat maps of the control and insomnia group and the difference map between the two groups. The areas surrounded by white contours showed significant differences. Warm colors: control > insomnia; cold colors: insomnia > control.

## Discussion

In the current study, we aim to test the hypothesis that the compromised perception of emotional facial expressions in insomniacs is related to impaired visual attention functions for selecting diagnostic features as revealed in their eye movements. Our results showed that individuals with insomnia symptoms were less accurate and marginally slower to identify angry faces than non-insomniac controls. Furthermore, insomniacs tended to misidentify angry faces as fearful faces more often than controls. Through the EMHMM approach (Chuk et al., 2014), we discovered two common eye movement patterns among the participants when viewing angry faces: an eye-mouth pattern that looked at the eyes and mouth primarily, and a nose-mouth pattern that fixated at either the nose or the mouth/chin region. Significantly more controls adopted the eye-mouth pattern and more insomniacs adopted the nose-mouth pattern. Indeed, the eye-mouth pattern was associated with faster identification of angry faces in the control group, whereas the nose-mouth pattern was associated with more misidentification of angry faces as fearful faces. These results suggest that insomniacs misidentified angry faces as fearful faces because of missing the eyes. The EMHMM approach is a data-driven method that reflects individual differences in both spatial and temporal dimensions of eye movements and provides quantitative assessments of

[^256]similarities among individual eye movement patterns, making it possible to reveal these effects. These findings were not possible with traditional approaches to eye movement data analysis such as using predefined regions of interest (ROIs; Henderson et al., 2005) or fixation heat maps (iMap, Caldara \& Miellet, 2011) between the insomniacs and the controls (Figure 3 vs. Figure 4).


Figure 5. (A) A positive correlation between the log-likelihood of being the nose-mouth pattern and the "fearful" response rate among all participants. (B) A negative correlation between the loglikelihood of being the eye-mouth pattern and the RT to identify angry faces in the control group.

Our finding is consistent with previous studies suggesting that insomnia and sleep loss are associated with compromised recognition of emotional facial expressions (e.g. Kyle et al., 2014; Cote et al., 2014). In particular, in an fMRI study, Goldstein-Piekarski and colleagues (2015) found that experimental sleep deprivation impaired behavioral and neural discrimination of angry faces from neutral faces. Angry facial emotions signal social threats, and thus misidentification or slower identification of angry faces may elevate interpersonal conflicts of insomniac individuals. Interestingly, insomniacs were more likely to misidentify angry faces as fearful faces than controls, suggesting that they may misidentify social threats senders as social threats receivers.

The misidentification of facial anger as facial fear in insomniacs corresponds to their eye movement patterns. Most of the insomniacs adopted a pattern that focused on either the nose or the mouth while missing the eye region. In contrast, most controls adopted a pattern that looked at mainly the eye region or both the eye and the mouth region.

The finding that insomniacs missed the eye region may be related to their impaired perception of angry faces. Indeed, through the 'Bubbles' reverse-correlation technique, Smith, Cottrell, Gosselin, and Schyns (2005) showed that eyes are the most diagnostic feature for recognizing angry expressions, whereas the most diagnostic features for recognizing the other three expressions (i.e. 'happy', 'sad', and 'fearful') were either mainly on the mouth region or comprised both the mouth and the eyes (see also Schyns, Petro, \& Smith, 2009). Consistent with this finding, Eisenbarth and Alpers (2011) showed that participants looked at the eyes longer than the mouth in recognizing anger and sad expressions, the mouth longer than the eyes for happy expressions, and the mouth and the eyes equally for fear and neural expressions. The exclusive importance of the eye region for recognizing angry faces may explain why we only observed behavioral differences between insomniacs and controls in identifying angry faces, since identifying other expressions do not require specific attention to the eye region as much as identifying angry expressions.

In the literature, biased interpretation of emotional information after sleep loss has typically been attributed to impaired functioning of limbic structures such as amygdala and anterior cingulate cortex and the functional connectivity between the prefrontal cortex and these limbic structures towards emotional stimuli (e.g., Yoo et al., 2007; GoldsteinPiekarski et al., 2015). In addition to this emotional brain network, the current study suggests that impaired attentional functioning may also play an important role in accounting for the misinterpretation of emotional information after sleep loss. Indeed, sleep loss is shown to affect visual attention control and activation in the attention brain network (Tomasi et al., 2009; Mander et al., 2008). Decreased activations in the attention brain network (e.g., the frontal eye field and intraparietal sulcus) are associated with maladaptive eye movement patterns and impaired recognition performance during face viewing (Chan et al., 2016). Impaired visual attention functions may cause failure of selecting diagnostic information for emotional face perception, leading to biased interpretation of emotional information. Our finding is consistent with Cote et al.'s (2014) study, which showed that impaired facial expression identification in sleep-deprived individuals was reflected in early visual ERP components including P1 and N170.

While the current study showed impaired recognition of angry expressions but not other expressions in insomniacs, some previous studies have reported disturbed perception of sad and fearful faces in addition to angry faces (e.g., Kyle et al., 2014). Cote et al. (2014) showed that altered early visual ERP responses due to sleep deprivation were observed for all expressions, whereas difference in identification accuracy between sleep deprived individuals and controls was only observed in sad faces. This effect suggests that while the modulation of sleep loss in attentional functioning may apply to all expressions in general, whether it results in decreased identification accuracy may depend on how it affects selection of diagnostic features, since different facial expressions differ in their diagnostic features (Schyns et al.,
2009). There may be individual differences in how features are selected and used for identification; individual differences in emotional functioning may also play a role. Future work will examine these possibilities.

In conclusion, here we showed that insomniacs misrecognized angry expressions because of missing diagnostic features in the eye region. This effect suggests that the impaired perception of facial expressions after sleep loss may be due to diminished visual attention control in addition to impaired emotional functioning. To our knowledge, this is the first to report the role of eye movement in the biased perception of emotional information due to sleep loss. Future studies will examine eye movements in clinical insomnia samples and sleep-deprived individuals to further examine the role of visual attention control in emotional perception after sleep loss.

## Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (\#17402814 to Hsiao and CityU110513 to Chan). We thank Dr. Tim Chuk for his support in data analysis and Ms. Li-Cih Hsu for her assistance in data collection.

## References

Baglioni, C., Spiegelhalder, K., Lombardo, C., \& Riemann, D. (2010). Sleep and emotions: a focus on insomnia. Sleep medicine reviews, 14(4), 227-238.
Caldara, R. \& Miellet, S. (2011). iMap: a novel method for statistical fixation mapping of eye movement data. Behavior Research Methods, 43(3), 864-878.
Chan, C. Y. H., Wong, J. J., Chan, A. B., Lee, T. M. C., \& Hsiao, J. H. W. (2016). Analytic eye movement patterns in face recognition are associated with better performance and more top-down control of visual attention: an fMRI study. In Proceedings of the 38th Annual Conference of the Cognitive Science Society, 854-859.
Chuk, T., Chan, A. B., \& Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. Journal of vision, 14(11), 8-8.
Chuk, T., Chan, A. B., \& Hsiao, J. H. (in press). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. Vision Research.
Cote, K. A., Mondloch, C. J., Sergeeva, V., Taylor, M., \& Semplonius, T. (2014). Impact of total sleep deprivation on behavioural neural processing of emotionally expressive faces. Experimental brain research, 232(5), 1429-1442.
Coviello, E., Chan, A. B., \& Lanckriet, G. R. G. (2014). Clustering hidden Markov models with variational HEM. Journal of Machine Learning Research, 15(1), 697-747.
Eisenbarth, H., \& Alpers, G. W. (2011). Happy mouth and sad eyes: scanning emotional facial expressions. Emotion, 11(4), 860.
Goldstein-Piekarski, A. N., Greer, S. M., Saletin, J. M., \& Walker, M. P. (2015). Sleep deprivation impairs the human central and peripheral nervous system
discrimination of social threat. The Journal of Neuroscience, 35(28), 10135-10145.
Hsiao, J. H., \& Cottrell, G. W. (2008). Two fixations suffice in face recognition. Psychol. Sci., 9(10), 998-1006.
Henderson, J. M., Williams, C. C., \& Falk, R. J. (2005). Eye movements are functional during face learning. Memory \& Cognition, 33(1), 98-106.
Kanan, C., Bseiso, D. N., Ray, N. A., Hsiao, J. H., \& Cottrell, G. W. (2015). Humans have idiosyncratic and task-specific scanpaths for judging faces. Vision Research, 108, 67-76.
Kyle, S. D., Beattie, L., Spiegelhalder, K., Rogers, Z., \& Espie, C. A. (2014). Altered emotion perception in insomnia disorder. Sleep, 37(4), 775-783.
Mander, B. A., Reid, K. J., Davuluri, V. K., Small, D. M., Parrish, T. B., Mesulam, M. M., Zee, P.C., \& Gitelman, D. R. (2008). Sleep deprivation alters functioning within the neural network underlying the covert orienting of attention. Brain research, 1217, 148-156.
Marchetti, L. M., Biello, S. M., Broomfield, N. M., Macmahon, K., \& Espie, C. A. (2006). Who is preoccupied with sleep? A comparison of attention bias in people with psychophysiological insomnia, delayed sleep phase syndrome and good sleepers using the induced change blindness paradigm. Journal of Sleep Research, 15(2), 212-221.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. Neuropsychologia, 9(1), 97-113.
Schyns, P. G., Petro, L. S., \& Smith, M. L. (2009). Transmission of facial expressions of emotion co-evolved with their efficient decoding in the brain: behavioral and brain evidence. PLoS ONE, 4(5), e5625.
Smith, M. L., Cottrell, G. W., Gosselin, F., \& Schyns, P. G. (2005). Transmitting and decoding facial expressions. Psychological Science, 16(3), pp. 184-189.
Tomasi, D., Wang, R. L., Telang, F., Boronikolas, V., Jayne, M. C., Wang, G. J., Fowler, J.S., \& Volkow, N. D. (2009). Impairment of attentional networks after 1 night of sleep deprivation. Cerebral cortex, 19(1), 233-240.
Wong, M. L., Lau, K. N. T., Espie, C. A., Luik, A. I., Kyle, S. D., \& Lau, E. Y. Y. (2017). Psychometric properties of the Sleep Condition Indicator and Insomnia Severity Index in the evaluation of insomnia disorder. Sleep Medicine, 33, 76-81.
Yoo, S. S., Gujar, N., Hu, P., Jolesz, F. A., \& Walker, M. P. (2007). The human emotional brain without sleep-a prefrontal amygdala disconnect. Current Biology, 17(20), 877-878.

# Low Dimensional Representations in Multi-Cue Judgment 

Joyce Wenjia Zhao (zhaowenj@sas.upenn.edu)<br>Department of Psychology, University of Pennsylvania Philadelphia, PA Address

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania Philadelphia, PA Address

Clintin P. Davis-Stober (stoberc@missouri.edu)<br>Department of Psychological Sciences, University of Missouri<br>Columbia, MO


#### Abstract

The study of multi-cue judgment investigates how decision makers integrate cues to predict the value of a criterion variable. We consider a multi-cue judgment task in which decision makers have prior knowledge of inter-cue relationships but are ignorant of how the cues correlate with the criterion. In this setting, a naive judgment strategy prescribes an equal weight for each cue. However, we find that many participants appear to use a weighting scheme based on a low-dimensional representation of the cue space. The use of such a representation is consistent with core insights in semantic memory research and has important optimality properties concerning judgment accuracy.


Keywords: judgment and decision making; cue integration; improper linear models; dimensionality reduction; semantic memory

## Introduction

Effective judgment and decision making involves the aggregation of multiple cues, or pieces of information, to evaluate a criterion variable. For example, individuals may receive advice from two or more friends regarding a financial investment, and aggregate this advice to calculate the expected return on the investment. Alternatively, they may have to choose between job candidates with multiple attributes, and have to aggregate these attributes to determine the quality of the candidates.

Traditionally, many normative and descriptive models of judgment and decision making adopt a linear approach and propose that decision makers compute the value of the criterion using a weighted average of the cues, with the weights being proportional to the observed relationship between the cues and the criterion (Brunswik, 1952; Keeney \& Raiffa, 1993). Linear models are often criticized as they require large amounts of information and abundant cognitive resources in order to be accurate. Thus, many researchers have proposed that individuals use improper linear models, such as heuristics. These models involve a fixed weighting scheme that assigns a priori weights to the cues. For example, an equal weights model gives each cue
the same weight, and the lexicographic model assigns all the weight to a single cue (Dawes, 1979; Gigerenzer \& Todd 1996). These models have been shown to perform as well as, if not better than, proper linear models in many situations, ranging from graduate student admission to clinical predictions (Dawes, 1979).
In addition to being cognitively simpler, improper linear models can also be used in situations where proper linear models are inapplicable. Consider settings where the relationship between the cues and the criterion is completely unknown. For example, individuals using the advice of their friends to judge an investment may not have previously observed how well their friends predict the performance of such investments. Likewise, individuals evaluating job candidates for novel or unconventional jobs may have never observed the value of different candidate attributes in the context of these jobs. In these situations, decision makers may have detailed knowledge about the relationship between the cues (e.g. how often their friends agree with each other or how frequently job candidate attributes co-occur) but have no way to assign weights to the cues in accordance with the standard linear model (where weights depend on the cues' relationship with the criterion). However, an a priori weighting scheme, as proposed by improper linear models, can still be used to make an evaluation.
For multi-cue judgment with known inter-cue relationships, but unknown cue-criterion relationships, the key questions of interest are the following: Which improper weighting scheme should decision makers use and which schemes do decision makers use. The former question has been tackled by Davis-Stober, Dana \& Budescu (2010a, 2010b). Davis-Sober et al. propose that any possible weighting scheme, $\boldsymbol{\beta}$, can be assessed with regards to how far it deviates from the true weight vector, $\boldsymbol{\beta}^{*}$, by taking the sum of squared difference between the weights in $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{*}$, i.e. $\sum_{i}\left(\beta_{i}-\beta_{i}{ }^{*}\right)^{2}$. When the cuecriterion relationships are unknown, $\boldsymbol{\beta}^{*}$ is also unknown. In these settings optimizing $\boldsymbol{\beta}$ can be seen as involving minimizing the risk, defined as the expectation of sum of squared error of $\boldsymbol{\beta}, \Sigma_{i}\left(\beta_{i}-\beta_{i}{ }^{*}\right)^{2}$. By this standard, the best improper linear weighting scheme is the eigenvector
corresponding to the first (i.e. largest) eigenvalue of the inter-cue correlation matrix (see Davis-Stober et al. 2010 for details). We will refer to this weighting scheme as $\boldsymbol{\beta}_{\boldsymbol{E V 1} 1}$. Here $E V 1$ in the subscript refers to the use of the eigenvector corresponding to the first eigenvalue.
$\boldsymbol{\beta}_{\boldsymbol{E V} \boldsymbol{1}}$ depends on the relationship between the cues in the judgment task, and can be shown, in appropriate settings, to approximate other existing improper linear models. For example, if the cues are equally, and positively, correlated with each other, $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ assigns an equal weight to each cue, similar to the equal weights model. In contrast, if cue 1 is highly correlated with all the other cues, and all the other cues are moderately correlated or uncorrelated internally, $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ overweighs cue 1 relative to other cues. This can mimic a lexicographic judgment strategy.

Normative solutions aside, descriptively, which weighting scheme do decision makers actually use when integrating multiple cues with unknown cue-criterion relationships? A first guess involves an equal weights model: Without knowing which cues are more related to the criterion than others, it seems conceivable that decision makers assign the same weights to all the available cues. This corresponds to a type of ignorance prior. However, a more principled guess could rely on insights regarding semantic representation. Decision makers with prior experience with the cues may have learnt mental representations of the cues. These representations, in many settings, correspond to projections of the decision makers' experiences with the cues onto a low-dimensional space. Such projections can be approximated by a principle components analysis on the cue-correlation matrix, or equivalently, a singular value decomposition on the matrix of cue-context co-occurrence. Indeed, such a decomposition is a key component of numerous existing approaches to modelling semantic representation, including latent semantic analysis (Landauer \& Dumais, 1997), multi-dimensional scaling (Kruskal \& Wish, 1978), and neural network models of semantic memory (Saxe, McClelland \& Ganguli, 2013). Interestingly, such a decomposition also yields the normative $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ model when only the first latent dimension of the projection is used to evaluate the criterion.

The goal of this paper is to investigate the plausibility of the $\boldsymbol{\beta}_{\boldsymbol{E V 1}}$ weighting scheme, and to compare its ability to predict participant judgments with alternate improper linear models such as the equal weights rule and the lexicographic rule. To appropriately test these models, we examine settings where participants have prior knowledge of inter-cue correlations but do not know how the different cues correlate with the criterion in consideration. Additionally, we systematically vary the cue-correlation matrix, and subsequently $\boldsymbol{\beta}_{E V \mathbf{1}}$, in order to adequately differentiate the predictions of this weighting scheme from those of alternate weighting schemes in our studies. We demonstrate the applicability of $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ for describing participant behavior in two ways: 1) by examining the model fits for $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ relative to other improper linear models, and 2 ) by testing whether the weights assigned by
$\boldsymbol{\beta}_{\boldsymbol{E V 1}}$ predict decision makers' use of these other improper models.

## General Method

In our three studies, the multi-cue judgment task was presented as an advice integration task, with the cues in consideration corresponding to the judgments of four advisors (similar to Bröder, 2003). They were described as predicted stock prices in Studies 1 and 2 and restaurant ratings in Study 3. Correspondingly, the criterion was the true stock price in Studies 1 and 2 and the true restaurant quality in Study 3. The cue-criterion correlations were never revealed to the participants.
The studies consisted of three tasks. The first two tasks exposed the participants to the cues, so as to allow them to form mental representations of the cue space. The third task asked participants to predict the criterion value based on the cues. In addition to being stated numerically, cue values in the three tasks were also shaded based on their magnitude. Participants were told that the cue values ranged from $0-100$ and were all centered at 50 . They were also told that some cues (advisors) might be more similar to each other, and that it was useful to pay attention to how closely different cues agreed with each other.


Figure 1. Stimuli display for Task 1-3.

In task 1, participants saw the four cues in 25 trials (Figure 1: upper left) displayed in four boxes. Each trial presented a set of cue values, and participants were asked to merely observe the cue values, without providing a response. In task 2, participants continued to learn the cue values, this time with feedback. Particularly, only three of the four cues were shown to participants (Figure 1: lower left). Participants had to guess the value of the fourth cue based on their knowledge of the inter-cue relationships. After the participant's guess, the real cue value was revealed. To increase motivation, participants were provided with a summary of their performance accuracy after every 50 trials. The cue to be guessed was determined at random in each trial.
In task 3, participants were shown all four cue values, and were asked to make a guess regarding the value of the
criterion (real stock price for Studies 1 and 2 and actual restaurant quality for Study 3; Figure 1: right). The true value of the criterion was not revealed after participants’ guesses, so that participants stayed uninformed regarding the cue-criterion relationship. Task 3 was the most relevant to our research question, as it provided a direct test of how cue values were integrated to make a judgment of the criterion.

## Study 1

Study 1 examines the predictions of the $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ weighting rule by considering a setting in which inter-cue relationships lead to a larger weight on one cue and smaller weights on the remaining cues.

## Methods

44 participants ( 37 females; Mean Age $=19.7$, SD Age $=$ 1.2), recruited from a university experimental participation pool, completed this study in a behavioral laboratory.

The study involved a hypothetical stock prediction task. The cue values were stock prices predicted by four advisors. For each cue, the values were normally distributed, with a mean of 50 and a standard deviation of 25. The inter-cue correlation matrix of the advisors is shown in Figure 2a.


Figure 2 (a) Inter-cue correlation matrix for Study 1 and the treatment condition of Study 2. (b) Inter-cue correlation matrix for the control condition of Study 2. (c) Inter-cue correlation matrix for Study 3.

As can be seen in this matrix, cue 1 is highly correlated with all the three other cues, with a correlation coefficient of 0.6 . The internal correlation among the remaining cues is very weak, with a correlation coefficient of 0.05 . The eigenvector corresponding to the first (largest) eigenvalue of the cue correlation matrix is $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}=[0.35,0.22,0.22$, 0.22 ]. Using this weighting vector, leads to an overweighting of cue 1 , and a relative underweighting of the remaining cues.
We used the above distributions to generate a single set of stimuli for all participants, for tasks 1,2 and 3. For each participant, the display position for each of the four cues was randomly chosen at the beginning of the study and stayed unchanged for the entire session. In other words, the specific advisor (advisor A, B, C or D) associated with cue 1 , was counterbalanced.

## Results

We first examined participants' performance in task 2, where they used three cue values for guessing the remaining cue value. Our analysis of behavior in this task suggested that participants were able to successfully learn the underlying cue structure. Particularly they placed a higher weight on cue 1 relative to the other cues when predicting the remaining cues ( $p<0.001$ ). Due to space constraints we will not outline these results in more detail (they will be reported elsewhere).
We also investigated the weighting scheme used by participants when integrating cues to predict criterion values in task 3. For this purpose, we considered a number of candidate weighting schemes, including $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ (corresponding to the first eigenvector of the cuecorrelation matrix) and $\boldsymbol{\beta}_{\boldsymbol{E W}}$ (corresponding to the equal weighting rule). We also considered lexicographic rules. Here, we tested four models that put all the weights on a single cue. These were referred to as $\boldsymbol{\beta}_{\text {LEX1 }}, \boldsymbol{\beta}_{\boldsymbol{L E X 2}}$, $\boldsymbol{\beta}_{\boldsymbol{L E X 3}}$ and $\boldsymbol{\beta}_{\boldsymbol{L E X 4}}$, corresponding to the cue that was given the unit weight. In addition to $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$, we also considered the linear weighting schemes corresponding to the remaining three eigenvectors of the cue correlation matrix. These are referred to as $\boldsymbol{\beta}_{\boldsymbol{E V 2}}, \boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{3}}$ and $\boldsymbol{\beta}_{\boldsymbol{E V 4} \mathbf{4}}$. Therefore, we have in total nine improper linear weighting schemes to compare. Each linear weighting scheme defined a weighting vector for the four cues. E.g. $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}=$ $[0.35,0.22,0.22,0.22], \boldsymbol{\beta}_{E W}=[0.25,025,0.25,0.25]$, $\boldsymbol{\beta}_{L E X 1}=[1,0,0,0]$, etc. For comparability, the weights in each scheme were constrained to add up to one.

In order to scale the criterion estimate generated by these improper linear models to match the participants' guesses, we introduced two additional participant-level parameters, $\alpha_{0}$ and $\alpha_{1}$, so that the predicted guess for each weighting scheme was $\alpha_{0}+\alpha_{1} \boldsymbol{\beta} \cdot \boldsymbol{C}$. Here $\boldsymbol{\beta}$ corresponds to the weighting vector of the model in consideration, and $\boldsymbol{C}$ is the vector of cue values presented in the trial. We also assumed a normally distribute error, with standard deviation $\sigma$, and subsequently fit each of these nine models by maximizing log-likelihood. $\alpha_{0}, \alpha_{1}$ and $\sigma$ were allowed to vary across the nine models. The model fitting was done on the participant level. Because the linear weighting schemes were pre-determined, each model used same number of parameters ( 3 parameters: $\alpha_{0}, \alpha_{1}$ and $\sigma$ ) to predict each participant's 100 guesses in task 3 .
We compared participant level log likelihood values for the nine candidate models (Table 1). Since all models have the same number of parameters, our model comparison is equivalent to model selection by AIC. Among 44 participants, 10 participants' predictions were best described by $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}, 23$ by $\boldsymbol{\beta}_{\boldsymbol{E W}}, 8$ by $\boldsymbol{\beta}_{\boldsymbol{L E X} \mathbf{1}}, 1$ by $\boldsymbol{\beta}_{\boldsymbol{E V \mathbf { 2 }}}$, 1 by $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{3}}$, and 1 by $\boldsymbol{\beta}_{\boldsymbol{L E X 4}}$. When comparing only $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ and $\boldsymbol{\beta}_{\boldsymbol{E W}}, 19$ participants were better described by $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$, whereas 25 were better described by $\boldsymbol{\beta}_{\boldsymbol{E W}}$. According to a paired Wilcoxon test on participant level model fits, there was no significant difference between log likelihood values of the $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ model (Median $=-351.38$ ) and the $\boldsymbol{\beta}_{E W}$ model (Median $=-351.59$ ), $Z=1.24, p=0.216$. Paired Wilcoxon tests also indicated that fits for all of the
remaining models were significantly worse than those for the $\boldsymbol{\beta}_{\boldsymbol{E V} \boldsymbol{1}}$ model and $\boldsymbol{\beta}_{\boldsymbol{E W}}$ model $(p<0.001)$.

Table 1: Comparison of model fits for Study 1

|  | Parameter (Median) |  |  | Log Likelihood |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Model | $\alpha_{0}$ | $\alpha_{1}$ | $\sigma$ | Median | Mean | \# best |
| EV1 | 7.32 | 0.84 | 8.57 | -351.38 | -357.16 | 10 |
| EV2 | 49.92 | 0.02 | 18.66 | -424.01 | -420.56 | 1 |
| EV3 | 49.96 | 0 | 18.63 | -423.86 | -420.64 | 1 |
| EV4 | 45.37 | 0.32 | 18.63 | -424.05 | -420.6 | 0 |
| EW | 4.94 | 0.89 | 8.56 | -351.59 | -357.10 | 23 |
| LEX1 | 17.94 | 0.63 | 8.76 | -351.39 | -358.54 | 8 |
| LEX2 | 29.81 | 0.37 | 15.83 | -408.81 | -406.45 | 0 |
| LEX3 | 32.17 | 0.38 | 15.95 | -410.32 | -407.24 | 0 |
| LEX4 | 29.7 | 0.39 | 15.58 | -408.7 | -405.83 | 1 |

Although participants had no information regarding the validities of any cues, a substantial subgroup of participants did not simply assign equal weights to cues. Instead, they overweighed cue 1 as predicted by $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$. The fact that some participants were actually best fit by $\boldsymbol{\beta}_{\boldsymbol{L E X 1}}$ suggested that some participants overweighed cue 1 even more than $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ recommended. Only one participant was best described by the other three lexicographic rules, indicating that $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ can predict which single cue participants tend to overweigh.

## Study 2

In Study 1, we found that the EV1 and EW models described participant level data about equally well, in terms of average log likelihood and proportion best fit. However, it is possible that predictions made by EV1 and EW were similar enough to be practically indistinguishable (given the noise in the data). This could confound our interpretation of model fit. Study 2 addresses this alternative explanation by manipulating the inter-cue relationships between subjects.

## Methods

64 participants ( 35 females; Mean Age $=19.9$, SD Age $=$ 1.1), recruited from a university experimental participation pool, completed this study in a behavioral laboratory.

All aspects of the study design were kept identical to Study 1, except that the cue correlation matrix varied between a treatment condition and a control condition. Participants were randomly assigned to one of these two conditions at the start of the study.

For the treatment condition, the inter-cue correlation matrix was identical to that in Study 1 (Figure 2a), generating an optimal weighting scheme with $\boldsymbol{\beta}_{E V 1}^{T r e a t}=$ [ $0.35,0.22,0.22,0.22$ ] (here we use the superscript to distinguish the treatment vs. control condition). For the control condition, the cue correlation matrix kept the correlation between all the cues constant at 0.4 (Figure 2b). Therefore, the weighting vectors predicted by the optimal weighting scheme and the equal weights rule were both $\boldsymbol{\beta}_{E V 1}^{C o n t}=\boldsymbol{\beta}_{E W}=[0.25,0.25,0.25,0.25]$. Due to different
inter-cue relationships across conditions, $\boldsymbol{\beta}_{E V 1}^{T r e a t}$ should provide a better account of behavior in the treatment condition compared to the control condition. Likewise $\boldsymbol{\beta}_{E V \mathbf{1}}^{C o n t}=\boldsymbol{\beta}_{E W}$ should provide a better account of behavior in the control condition compared to the treatment condition (even if a large subgroup of participants in the treatment condition do place an equal weight on all cues).

## Results

31 participants were assigned to the treatment condition and 33 participants were assigned to the control condition. As in Study 1, we first looked at participant learning in task 2. In the treatment condition, participants did learn the special status of cue 1. Particularly, as in Study 1, they placed a higher weight on cue 1 relative to the other cues when predicting the remaining cues ( $p<0.001$ ). In the control condition, participants placed similar weights on cues 1-4 when predicting cue values, indicating that they learnt different inter-cue relationships for the two conditions. Manipulating the cue-correlation matrix thus had an effect on participant learning. This laid the basis for task 3, where participants integrated cues to predict criterion values (again, due to space constraints, we will not expand on these results here).

Next, we examined which weighting schemes were used by participants in task 3. For both conditions, we applied the model fitting procedures of Study 1, and nine linear weighting schemes were compared on the participant level (Tables 2 and 3 ). Out of 31 participants in the treatment condition, 6 were best described by $\boldsymbol{\beta}_{V V 1}^{\text {Treat }}$, 18 by $\boldsymbol{\beta}_{\boldsymbol{E W}}, 6$ by $\boldsymbol{\beta}_{\boldsymbol{L E X 1}}$ and 1 by $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{2}} \cdot \boldsymbol{\beta}_{\boldsymbol{E V 1}}^{T r e a t}$ outperformed $\boldsymbol{\beta}_{\boldsymbol{E W}}$ for a substantial subgroup of participants (13 out of 31). As in Study 1, some participants were best described by $\boldsymbol{\beta}_{\boldsymbol{L E X 1}}$, indicating that they overweighed cue 1 more than recommended by $\boldsymbol{\beta}_{E V 1}^{T r e a t}$. No participant was best described by the other three lexicographic rules, indicating that $\boldsymbol{\beta}_{E V \mathbf{1}}^{T r e a t}$ can predict decision makers' use of other improper linear models in the treatment condition.

We also compared the log likelihood values of the fits. Although the log likelihood values of the $\boldsymbol{\beta}_{E V 1}^{T r e a t}$ model were significantly smaller than those of the $\boldsymbol{\beta}_{\boldsymbol{E} \boldsymbol{W}}$ model ( $Z=-2.06, p=0.040$ ), the effect size was small ( Median $_{E V 1}=-350.56$, Median ${ }_{E W}=-350.17$ ). Additionally both the $\boldsymbol{\beta}_{\boldsymbol{E V 1}}^{\text {Treat }}$ model and $\boldsymbol{\beta}_{\boldsymbol{E W}}$ model predicted participant level data significantly better than all other models ( $p<0.001$ ). These results replicate findings of Study 1.
In the control condition, the inter-cue correlation matrix was balanced and the weighting schemes for $\boldsymbol{\beta}_{E V 1}^{C o n t}$ and $\boldsymbol{\beta}_{E W}$ were identical. Unsurprisingly, all 33 participants were better described by $\boldsymbol{\beta}_{E V 1}^{C o n t}=\boldsymbol{\beta}_{E W}$ than any other models (Table 3). The fact that no participants were best fit by lexicographic rules in the control condition but some were best fit by $\boldsymbol{\beta}_{\boldsymbol{L E X 1}}$ in the treatment condition again indicated that participants' cue weighting behavior can be predicted by the inter-cue correlation matrix.
Lastly, we examined the predictions of $\boldsymbol{\beta}_{E V 1}^{T r e a t}$ on the data from the control condition. For this purpose we fit a
tenth model in the control condition, with weights given by $\boldsymbol{\beta}_{E V 1}^{\text {Treat }}$ (and $\alpha_{0}, \alpha_{1}$ and $\sigma$ flexible). Unlike the treatment condition, this model outperformed the $\boldsymbol{\beta}_{E W}=$ $\boldsymbol{\beta}_{E V 1}^{C o n t}$ model for only 4 out of 33 participants in the control condition. A paired Wilcoxon test indicated that the log likelihoods of the $\boldsymbol{\beta}_{E V 1}^{T r e a t}$ model on the control-condition data (Median $=-336.73$ ) were significantly lower than those of the $\boldsymbol{\beta}_{E W}=\boldsymbol{\beta}_{E V 1}^{\text {Cont }}$ model (Median $=-335.76$ ), $Z=-4.24, p<0.001$.

Table 2: Comparison of model fits for Study 2 (treatment)

|  | Parameter (Median) |  |  | Log Likelihood |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Model | $\alpha_{0}$ | $\alpha_{1}$ | $\sigma$ | Median | Mean | \#best |
| EV1 | 7.70 | 0.90 | 10.07 | -350.56 | -363.55 | 6 |
| EV2 | 50.65 | 0.03 | 19.73 | -424.88 | -423.34 | 1 |
| EV3 | 50.72 | 0.01 | 19.73 | -425.08 | -423.60 | 0 |
| EV4 | 46.79 | 0.33 | 19.66 | -424.85 | -423.52 | 0 |
| EW | 5.10 | 0.95 | 10.07 | -350.17 | -363.35 | 18 |
| LEX1 | 19.28 | 0.67 | 10.19 | -353.77 | -365.33 | 6 |
| LEX2 | 31.88 | 0.40 | 16.28 | -407.65 | -408.75 | 0 |
| LEX3 | 33.68 | 0.39 | 17.28 | -412.35 | -411.83 | 0 |
| LEX4 | 30.22 | 0.44 | 16.59 | -411.66 | -409.48 | 0 |

$\begin{array}{cc}\text { Table } 3 \text { Comparison of model fits for Study } 2 \text { (control) } \\ \text { Parameter (Median) } & \text { Log Likelihood }\end{array}$

| Model | $\alpha_{0}$ | $\alpha_{1}$ | $\sigma$ | Median | Mean | \#best |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| EV1/EW | 10.59 | 0.83 | 7.49 | -335.76 | -336.13 | 29 |
| EV2 | 51.41 | 0.00 | 17.64 | -415.22 | -411.91 | 0 |
| EV3 | 51.41 | 0.00 | 17.64 | -415.22 | -411.92 | 0 |
| EV4 | 51.39 | 0.07 | 17.63 | -414.84 | -411.69 | 0 |
| LEX1 | 28.91 | 0.45 | 13.00 | -387.19 | -385.25 | 0 |
| LEX2 | 30.35 | 0.43 | 13.30 | -391.31 | -388.82 | 0 |
| LEX3 | 28.55 | 0.46 | 12.20 | -382.23 | -382.13 | 0 |
| LEX4 | 31.17 | 0.44 | 13.32 | -390.59 | -387.50 | 0 |
| EV1Treat | 11.04 | 0.81 | 7.75 | -336.73 | -340.25 | 4 |

Overall, the differences in the mean and median $\log$ likelihoods of the $\boldsymbol{\beta}_{E V 1}^{T r e a t}$ and the $\boldsymbol{\beta}_{E V \mathbf{1}}^{\text {Control }}=\boldsymbol{\beta}_{\boldsymbol{E W}}$ models in the control condition were 4.12 and 0.97 respectively. These were larger than the equivalent differences in the treatment condition, which were 0.20 and 0.39 (these differences were 0.06 and -0.21 in Study 1). These results indicate that the relatively good fits for the $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ model in the treatment condition of Study 2 and in Study 1 were not due to this model mimicking the equal weights rule.

## Study 3

Study 3 provides a more stringent test of the $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ model by considering a setting with more complex inter-cue relationships. It also examines judgments of restaurant quality rather than stock performance.

## Methods

46 participants ( 34 females; Age Mean = 19.3, SD Age $=$ 1.0 ) recruited from a university experimental participation pool, completed this study in a behavioral laboratory.

The study was framed as involving judgments of restaurant quality. Here the cue values were restaurant scores rated by four reviewers, and the criterion corresponded to the real restaurant quality. Other aspects of the study design were kept identical to Study 1 , except the inter-cue correlation matrix, which was changed to the matrix displayed in Figure 2c. Here cue 1 is highly correlated with cue 2 , cue 3 is moderately correlated with cue 4 , and cues 1 and 2 are weakly correlated with cues 3 and 4 . With this inter-cue correlation structure, $\boldsymbol{\beta}_{\boldsymbol{E V 1}}$ predicts a weighting vector of $[0.35,0.35,0.15,0.15]$, i.e. an overweighting of cues 1 and 2 , relative to 3 and 4 .

## Results

As in Studies 1 and 2, our analysis of behavior in task 2 suggested that participants were able to successfully learn the underlying cue structure. Particularly they relied more on cues 1 and 2 than on cue 3 and 4 when guessing for cues 1 and 2 ; they also relied more on cues 3 and 4 than on cues 1 and 2 when guessing for cues 3 and 4. Again, due to space constraints we will not outline these results in more detail.
Next, we investigated the linear weighting schemes used by participants in task 3. The nine candidate weighting schemes and the model fitting procedures were the same as in Study 1 (though, of course, $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}, \boldsymbol{\beta}_{\boldsymbol{E V 2}}$, $\boldsymbol{\beta}_{\boldsymbol{E V}}, \boldsymbol{\beta}_{\boldsymbol{E V} 4}$ assigned different weights to the cues in Study 3, compared to the corresponding models in Study 1).

Out of the 46 participants, 7 were best fit by $\boldsymbol{\beta}_{\boldsymbol{E V} 1}$ and 38 were best fit by $\boldsymbol{\beta}_{\boldsymbol{E W}}$. The remaining participant was best fit by $\boldsymbol{\beta}_{\boldsymbol{L E X 2}}$. When comparing only $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ and $\boldsymbol{\beta}_{\boldsymbol{E W}}$, we found that 8 participants were better described by $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ than $\boldsymbol{\beta}_{\boldsymbol{E W}}$. Additionally, $\boldsymbol{\beta}_{\boldsymbol{E W}}$ (Median $=-350.82$ ) was significantly better than $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ (Median $=-353.15$ ) according to a paired Wilcoxon test performed on participant level log likelihood values, $Z=3.98, p<$ 0.001 . Except for $\boldsymbol{\beta}_{\boldsymbol{E W}}, \boldsymbol{\beta}_{\boldsymbol{E V} \boldsymbol{1}}$ outperformed all the other candidate models (with $p<0.001$ ). As would be predicted by the $\boldsymbol{\beta}_{E V \mathbf{1}}$ model, the lexicographic models did not provide a good account of participant behavior in this study. Table 4 provides additional details regarding these fits.

Table 4 Comparison of model fits for Study 3

|  | Parameter (Median) |  |  | Log Likelihood |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Model | $\alpha_{0}$ | $\alpha_{1}$ | $\sigma$ | Median | Mean | \# best |
| EV1 | 14.04 | 0.76 | 9.28 | -353.15 | -358.26 | 7 |
| EV2 | 44.77 | 0.26 | 16.76 | -412.07 | -410.59 | 0 |
| EV3 | 51.04 | 0.10 | 17.29 | -415.25 | -413.63 | 0 |
| EV4 | 50.95 | 0.03 | 17.35 | -415.29 | -413.82 | 0 |
| EW | 6.32 | 0.88 | 8.75 | -350.82 | -349.09 | 38 |
| LEX1 | 30.22 | 0.44 | 12.58 | -383.14 | -384.67 | 0 |
| LEX2 | 28.31 | 0.49 | 11.94 | -379.23 | -381.29 | 1 |
| LEX3 | 32.89 | 0.34 | 15.32 | -403.08 | -401.97 | 0 |
| LEX4 | 32.01 | 0.37 | 14.63 | -400.57 | -398.18 | 0 |

As in previous studies, we found that a significant subgroup of participants overweighed some cues (as
suggested by $\boldsymbol{\beta}_{\boldsymbol{E V 1} 1}$ ), rather than simply averaging all the available cues (as suggested by $\boldsymbol{\beta}_{\boldsymbol{E W}}$ ). That said, the performance of $\boldsymbol{\beta}_{\boldsymbol{E V} \boldsymbol{1}}$ was relatively worse in this study compared to our previous studies. This could be due to the differences in the cue-correlation matrices, suggesting that decision makers are less likely to use the $\boldsymbol{\beta}_{\boldsymbol{E V 1} 1}$ scheme when the underlying cue structure is complex. These differences could also, however, be attributed to the change in the task frame. Restaurant quality is more subjective than stock performance, and decision makers may be less likely to rely on the cue-correlation structure in these subjective settings.

## Discussion

In three studies, we investigated how decision makers weigh cues when cue criterion relationships are unknown. The optimal improper linear model uses the eigenvector, $\boldsymbol{\beta}_{\boldsymbol{E V I}}$, corresponding to the largest eigenvalue of the cue correlation matrix (Davis-Stober et al., 2010a, 2010b). Low dimensional representations of the cue space, learnt by some common models of semantic memory (Kruskal \& Wish, 1978; Landauer \& Dumais, 1997; Saxe et al., 2013), can also produce this type of weighting scheme.

Our results suggest that $\boldsymbol{\beta}_{\boldsymbol{E V} \boldsymbol{1}}$ provides a good description of participants' behavior. This model outperformed all other improper linear models tested in this paper, except for the equal weights model (with weights $\boldsymbol{\beta}_{\boldsymbol{E W}}$ ). On the aggregate level, the log likelihoods for the $\boldsymbol{\beta}_{\boldsymbol{E V} \boldsymbol{1}}$ and $\boldsymbol{\beta}_{\boldsymbol{E W}}$ weighting scheme were relatively close, showing no meaningful differences in Study 1, very minor differences in the treatment condition of Study 2, and somewhat larger differences in Study 3. As for individual level fits, there existed a substantial group of participants for whom $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ outperformed $\boldsymbol{\beta}_{\boldsymbol{E W}}$. The size of this group ranged from $43 \%$ of the participant pool in Study $1,42 \%$ in the treatment condition of Study 2, and $17 \%$ in Study 3. Moreover, a comparison of the control and the treatment conditions of Study 2 showed that experimental manipulations that varied the inter-cue correlation matrix influenced relative model fits.
$\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ was also able to predict when and how participants used lexicographic weights. When $\boldsymbol{\beta}_{E V \mathbf{1}}$ prescribed equal weights (control condition of Study 2) or the overweighing two cues (Study 3), there were almost no participants who were best described by such lexicographic weighting schemes. In contrast, in Study 1 and the treatment condition of Study 2, $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ overweighed a single cue. In these conditions, a substantial group of participants ( $18 \%$ in Study 1 and $19 \%$ in the treatment condition of Study 2) behaved according to a lexicographic rule that placed all of the weight on this cue (in contrast lexicographic rules that prioritize other cues all performed very poorly).

That said, $\boldsymbol{\beta}_{E V 1}$ did not provide a good account of behavior in Study 3, which adopted a more complex intercue correlation matrix. The results of this study suggest that such a weighting scheme may not be used in all settings. Additionally, the equal weights rule was the majority model in all studies, indicating that most participants tend to use the simpler equal weights strategy (corresponding to an ignorance prior) in the absence of
cue-criterion knowledge. Further work should examine the effect of inter-cue correlation structure and individual differences on the use of the $\boldsymbol{\beta}_{\boldsymbol{E V} \boldsymbol{1}}$ weighting rule. This work may extend the insights of other cognitive models of multi-cue judgment, such as those relying on neural network representations (Glöckner, Hilbig \& Jekel, 2014) or exemplar memory-based predictions (Juslin, Karlsson \& Olsson, 2008). Such models have not been applied to settings in which cue-criterion relationships are unknown. However, they nonetheless provide formal predictions regarding the learning and representation of cue knowledge and its relationship with the statistical structure of the judgment environment. For this reason they may provide a more adequate framework for understanding the cognitive underpinnings of the $\boldsymbol{\beta}_{\boldsymbol{E V} \mathbf{1}}$ weighting model.

## References

Bröder, A. (2003). Decision making with the" adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(4), 611.
Brunswik, E. (1952). The conceptual framework of psychology. Psychological Bulletin, 49(6), 654-656.
Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34(7), 571.
Davis-Stober, C. P., Dana, J., \& Budescu, D. V. (2010). Why recognition is rational: Optimality results on single-variable decision rules. Judgment and Decision Making, 5(4), 216.
Davis-Stober, C. P., Dana, J., \& Budescu, D. V. (2010). A constrained linear estimator for multiple regression. Psychometrika, 75(3), 521-541.
Gigerenzer, G., \& Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 103, 650-669.
Gigerenzer, G., \& Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In Simple heuristics that make us smart (pp. 3-34). Oxford University Press.
Glöckner, A., Hilbig, B. E., \& Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. Cognition, 133(3), 641666.

Juslin, P., Karlsson, L., \& Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. Cognition, 106(1), 259-298.
Keeney, R. L., \& Raiffa, H. (1993). Decisions with multiple objectives: preferences and value trade-offs. Cambridge university press.
Kruskal, J. B., \& Wish, M. (1978). Multidimensional scaling (Vol. 11). Sage.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211.
Saxe, A. M., McClelland, J. L., \& Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks. In Proceedings of the 35th annual meeting of the Cognitive Science Society (pp. 1271-1276).

# Using single unit recordings in PDP and localist models to better understand how knowledge is coded in the cortex 

Jeffrey S. Bowers (j.bowers@bristol.ac.uk)<br>Department of Experimental Psychology, 12a Priory Road Bristol, BS8-1TU UK

Keywords: localist representation; distributed representation; grandmother cell; neural network

## Introduction

There is long history of studies documenting that some neurons respond to images of objects, faces, and scenes in a highly selective manner. This includes neurons in the human hippocampus (e.g., the famous example of a neuron responding to images of the actress Jennifer Aniston) and neurons in high-level visual cortex in monkey (for reviews see Bowers, 2009; Ison, Quian Quiroga, \& Fried, 2015). These findings have led to a growing interest in the claim that some neurons code for information in a localist ('grandmother cell') manner, as reflected in the many contributions to a recent special issue on this topic in the journal Language, Cognition, \& Neuroscience (Bowers, 2017).

By contrast, it is only recently that interest in characterizing the selectivity of single units in connectionist networks has gathered speed. Critically, these studies also show that networks learn highly selective representations under a number of conditions, as detailed below. In this talk I will summarize recent research in my lab that explores the conditions in which artificial networks learn selective codes, and research comparing the responses of selective neurons and localist representations used in cognitive models. These findings suggest when and why some neurons in cortex respond in a highly selective manner, and highlight the biological plausibility of localist models in psychology.

## Selective codes as a solution to the superposition catastrophe

In Bowers, Vankov, Damian, and Davis $(2014,2016)$ we carried out single-unit recordings on networks trained to coactivate multiple words at the same time in short-term memory (STM). We adapted models by Botvinick and Plaut (2006) who demonstrated that recurrent PDP networks can support human-like performance on STM tasks, and claimed that the models succeeded on the basis of co-activating learned distributed representations. This claim is important because it challenges the hypothesis that overlapping distributed representations result in blend patterns that are ambiguous, the so-called superposition catastrophe (Von Der Malsburg, 1986). The superposition catastrophe has been one of the key arguments in support of localist representations (Bowers, 2002; Page, 2000)

However, we showed that the Botvinick and Plaut (2006) and related models solved the superposition catastrophe by learning localist representations. Adapting an analytical tool developed by Berkeley, Dawson, Medler, Schopflocher,
and Hornsby (1995), we carried out single unit recordings of the hidden units of trained networks. We showed that the models learned more selective codes when the superposition constraint became more challenging. For example, Figure 1 depicts a hidden unit (unit 89 of 200 hidden units) that responded selectively to the trained word 'cot' (taken from Bowers et al., 2014).

Figure 1


Furthermore, we found that recurrent networks of STM were only able to recall lists of novel words when they learned localist representations (Bowers et al., 2016), contrary to the widespread assumption that distributed codes are better able to support generalization. These findings extend our understanding of when and why some neurons respond selectively: Just as neurons in the hippocampus are thought to code information in a selective manner in order to support fast learning without forgetting (Marr, 1971), our findings suggest some neurons in cortex learn selective codes for the sake of STM.

## Selective codes as a solution to some forms of arbitrary input-output mappings.

Recently there has been an explosion of interest in characterizing the selectivity of single hidden units in socalled 'deep' networks that achieve state-of-the-art performance on a range of tasks, including object and spoken word identification (for review, see Bowers, 2017). The striking finding is that these networks often learn highly selective representations even when trained on items one-at-a-time. This raises the question as to why we found that networks learned non-selective representations when trained on items one-at-a-time (Bowers et al., 2016).

Vankov and Bowers (2017) began to explore the conditions in which PDP networks learn selective and nonselective codes when trained on words one-at-a-time. Models learned non-selective distributed codes under a range of conditions, including when trained on many arbitrary input-output mappings. For example, a 3-layered model trained to map random patterns of binary inputs (with input units taking on an activation of 1 or 0 ) to another
random pattern of binary output units learned distributed codes.

However, we found one condition in which a 3-layered network learned localist codes: when trained on images of faces when input units took on continuous values and the model was trained on many-to-one mappings (with multiple different images of a given person mapping onto the same output representation). For example, Figure 2 depicts 16 localist units (units that are highlighted) out of 500 hidden units that selectively fire to a given face. To illustrate, the face images that activate units 312 and 404 are displayed.

Figure 2


We are currently carrying out more simulations to better understand the conditions in which networks learn distributed and localist codes when trained on items one-at-a-time. For example, is it many-to-one mappings that is critical, or the nature of the images themselves?

## Comparing the selectivity of single neurons to the selectivity of single units in localist models in psychology.

Even when neurons are identified that selectively respond to images of one person or object within an experiment, it is often claimed that the neuron would responds to other (untested) categories of images. For example, Waydo et al. (2006) estimated that the most selective neurons observed in Quian Quiroga et al. (2005) study would respond to between 50-150 different people or objects if researchers had more time to find the relevant images. This is taken as inconsistent with grandmother cells.

However, Gubian, Davis, Alderman, and Bowers (2017) showed that the analysis of Waydo et al. (2006) is consistent with localist models in psychology. We carried out singleunit recordings in the Spatial Coding Model of visual word identification that represents $\sim 30,000$ words in a localist manner (Davis, 2010). Under parameter conditions that allow the model to correctly identify words we found that that the localist representations responded to approximately to 50 different words (e.g., the word DOG responds most strongly to the input DOG, but also responds above baseline to LOG, FOG, JOG, etc.). Page (2017) also provides evidence that localist models can account for single-cell recording data taken to support distributed coding.

Together, these results highlight the computational reasons why some neurons in cortex respond in a highly selective manner, and show that localist (grandmother cell) representations are in fact consistent with single-cell recording data.

## References

Berkeley, I. S. N., et al. (1995). Density plots of hidden unit activations reveal interpretable bands. Connection Science, 7, 167-186

Bowers, J. S. (2002). Challenging the widespread assumption that connectionism and distributed representations go hand-in-hand. Cognitive Psychology, 45, 413-445.

Bowers, J.S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. Psychological Review, 116, 220-251.

Bowers, J. S. (2017). Grandmother cells and localist representations: A review of current thinking. Language, Cognition, \& Neuroscience, 32, 257-273.

Bowers, J. S., Vankov, I. I., Damian, M. F., \& Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. Psychological Review, 121, 248-261.

Bowers, J. S., Vankov, I. I., Damian, M. F., \& Davis, C. J. (2016). Why do some neurons in cortex respond to information in a selective manner? Insights from artificial neural networks. Cognition, 148, 47-63.

Davis, C. J. (2010). The spatial coding model of visual word identification. Psychological Review, 117, 713-758.

Gubian, M., Davis, C.I., Adelman, J.S., \& Bowers, J.S. (2017). Comparing single-unit recordings taken from a localist model to single-cell recording data: A good match. Language, Cognition, \& Neuroscience, 32, 380-391.

Ison, M. J., Quian Quiroga, R., \& Fried, I. (2015). Rapid encoding of new memories by individual neurons in the human brain. Neuron, 87(1), 220-230.

Marr, D. (1971). Simple memory: A theory for archicortex. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 23-81

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. Behavioral and Brain Sciences, 23, 443-512.

Page, M. (2017). Localist models are compatible with information measures, sparseness indices, and complementary-learning systems in the brain. Language, Cognition and Neuroscience, 32(3), 366-379.

Waydo, S et al. (2006). Sparse representation in the human medial temporal lobe. Journal of Neuroscience, 26, 10232-10234.

Vankov, I. I., Bowers, J. S. (2017). Do arbitrary inputoutput mappings in parallel distributed processing networks require localist coding? Language, Cognition, \& Neuroscience, 32, 392-399.

Von Der Malsburg, C. (1986). Am I thinking assemblies? In Brain theory (pp. 161-176). Springer Berlin Heidelberg.

# Interoception: The Forgotten Modality in Perceptual Grounding of Concepts 

Louise Connell (l.connell@lancaster.ac.uk)<br>Dermot Lynott (d.lynott@lancaster.ac.uk)<br>James Carney (j.carney@lancaster.ac.uk)<br>Department of Psychology, Fylde College, Lancaster University, Bailrigg, Lancaster LA1 4F, UK

Keywords: concepts; perceptual strength; grounded cognition; embodied cognition; word recognition

Concepts are the basis of the human cognitive system, and the question of what constitutes the content of these mental representations has long occupied the cognitive sciences. Work in psychology, linguistics and cognitive neuroscience has converged on the idea that we develop our conceptual representations through our perception of and interaction with our environment. To date, such research has typically restricted consideration to the perceptual modalities of vision, touch, sound, taste, and smell. However, there is another major modality of perceptual information that is distinct from these traditional five senses; that is, interoception, or sensations within the body. In this paper, we explore the role of interoception in the perceptual grounding of concepts.

Recently, modality-specific measures of the strength of perceptual experience (Lynott \& Connell 2009, 2013) have proven themselves important predictors of human behaviour in a range of conceptual tasks including word recognition and reading (Connell \& Lynott, 2010, 2012, 2014a, 2014b, $2015,2016)$. In a megastudy of over 32,000 words from across the abstract-concrete spectrum, we asked people to provide modality-specific ratings of perceptual strength for six modalities: the usual five (auditory, haptic, gustatory, olfactory, visual) plus the new category of interoceptive strength. We found that interoceptive information dominates the perceptual profile of a sizeable number of concepts ( $9 \%$; e.g., hangover, eternal, remorse), less than the proportion of concepts dominated by vision ( $74 \%$; e.g., book) or sound ( $12 \%$; e.g., melody), but more than are dominated by touch ( $3 \%$; e.g., silky), gustation ( $2 \%$; e.g., candy), or olfaction ( $<1 \%$; e.g., bleach). Using principal components analysis to examine how interoception relates to the other perceptual modalities, we found that it tends to be strongly loaded against visual and haptic strength (i.e., that which is sensed within the body can be neither seen nor touched) but is relatively distinct from sound, taste, and smell.

Finally, we tested whether interoceptive strength offers valuable information to conceptual content by examining its role in semantic facilitation of word recognition. Maximum perceptual strength (i.e., strength in the dominant modality) has previously been shown to predict word recognition performance better than concreteness or imageability (Connell \& Lynott, 2012). We therefore compared the predictive ability of two different versions of maximum
perceptual strength: the original measure based on five traditional modalities, and a new version based on six modalities including interoceptive strength. In a regression analysis of lexical decision and word naming performance, interoceptive information considerably improved the efficacy of maximum perceptual strength in predicting both response time and accuracy (Bayes Factors ranged from $\mathrm{BF}_{10}=3.303 \times 10^{7}$ to $\left.\mathrm{BF}_{10}=3.059 \times 10^{16}\right)$. That is, perceptually strong words were recognized more quickly and accurately than perceptually weak words, and interoceptive strength was a valuable component in this perceptual facilitation. Overall, these findings suggest that interoception has comparable status to other modalities in contributing to the perceptual grounding of concepts.

## References

Connell, L., \& Lynott, D. (2010). Look but don't touch: Tactile disadvantage in processing modality-specific words. Cognition, 115, 1-9.
Connell, L., \& Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition, 125, 452465.

Connell, L., \& Lynott, D. (2014a). I see/hear what you mean: semantic activation in visual word recognition depends on perceptual attention. Journal of Experimental Psychology: General, 143, 527-533.
Connell, L., \& Lynott, D. (2014b). Principles of representation: Why you can't represent the same concept twice. Topics in Cognitive Science, 6, 390-406.
Connell, L., \& Lynott, D. (2015). Embodied semantic effects in visual word recognition. In Y. Coello, \& M. Fischer (Eds.), Foundations of Embodied Cognition (Vol. 2): Conceptual and Interactive Embodiment. Hove, UK: Psychology Press.
Connell, L., \& Lynott, D. (2016). Do we know what we're simulating? Information loss on transferring unconscious perceptual simulation to conscious imagery. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42, 1218-1232.
Lynott, D., \& Connell, L. (2009). Modality exclusivity norms for 423 object properties. Behavior Research Methods, 41, 558-564.
Lynott, D., \& Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behavior Research Methods, 45, 516-526.

# Tracking Meaning Change Over Time: A Dynamic Field Theory Model 

Kenny R. Coventryı (k.coventry@uea.ac.uk)<br>School of Psychology, University of East Anglia, Norwich, UK

Harmen B. Gudde (h.gudde@uea.ac.uk)<br>School of Psychology, University of East Anglia, Norwich, UK<br>Holger Schultheis (schulth@informatik.uni-bremen.de)<br>Cognitive Systems Group, University of Bremen, Bremen, Germany

Keywords: meaning; spatial language; dynamic field theory

## Meaning Change

Words are often regarded as "slippery customers" (Labov, 1973). First, it is difficult to come up with a fixed content that seems to apply across all uses of a single word (Wittgensten, 1953). Second, the meaning of a word is subject to wider contextual constraints beyond the company it keeps with other words within a sentence. Third, the assumption that a word means a fixed thing is at odds with the fact that word meaning shifts over time, as do the objects a word refers to. Taken together, theories of word meaning need to be able to account for the intuition that words have content associated with them, while allowing for variation in how a word is used in context, and how word meaning can change over time.

We provide an approach to word meaning within the spirit of dynamic systems theory (DST) models of cognition that accounts for the slipperiness of words. First, we draw an analogy between models of (non-linguistic) spatial behavior and how the meaning of language changes over time while exhibiting regularity. In particular, we identify two key features of DST models - multicausality and the building of temporally-bound attractor states - that afford application to theories of meaning. Second, we take these features of DST models and test them in experiments examining the comprehension of spatial expressions over time using object placement behavior as a measure (e.g. "Place the oil paint tube over the toothbrush"). Building on earlier work examining the constraints in which spatial language is used (Carlson-Radvansky et al., 1999; Coventry, 2013, 2015; Coventry \& Garrod, 2004; Coventry et al., 2001, 2010, 2013, 2016; Gudde et al. 2016), we present a programme of studies mirroring early models of spatial memory, with experimental data showing striking similarity with results from other (non-linguistic) spatial tasks (namely the A not B error tasks and associated model produced by Smith \& Thelen; e.g. Smith \& Thelen, 2003; Thelen et al., 2001). Third, we take these data, and show that a previous DST model originally developed to account for infant
perseverative reaching behavior (Thelen et al., 2001) provides an elegant model of the changing meaning of spatial expressions over time.

Figure 1: Examples of scenes used (A) and movement manipulation


The programme of experiments involved using placement behavior as a proxy for situation-specific meaning (adapting a method from Carlson-Radvansky et al., 1999). Participants were given spatial expressions of the form PLACE OBJECT A 'PREPOSITION' OBJECT B, followed by a picture displayed on a computer screen. The task was to move OBJECT A so that the relation between objects matched the location denoted by the sentence (prepositions used were over/underlabovelbelow). OBJECT B was always an object with a functional part at one end (e.g. a toothbrush), and these objects were always displayed in sideways view (Figure 1). Critically the similarity between the probe and the prime trials was manipulated - analogous to the different object locations in the A not B error task (Thelen et al., 2001). The objects to be placed were either functionally related (e.g. a toothpaste tube and a toothbrush, hereafter $F$ ) or non-functionally related (e.g. a tube of paint and a toothbrush, hereafter NF). Previously it has been shown that placement behavior for F object pairs is different from placement behavior for NF object pairs when participants are given spatial sentences with the preposition above in them. Placements for an F object were nearer the

[^257]functional part of the other object than for NF objects; placements for NF objects were nearer the mid-point (centre of mass) of the other object than for F objects (CarlsonRadvansky et al., 1999).

The programme of studies varied the similarity between prime and probe trials (in terms of same/different spatial relations, functional relations between objects, and the way in which objects are moved) as well as the number of prime trials (i.e. the extent to which an attractor state is built prior to probe placements). Among the results in the series of studies we find evidence that object placements on a probe trial are affected by the number of prime trials presented first, and the nature of the similarity between prime and probe trials. For example, when a probe involves functionally related objects, placements are more functional (i.e. more over the bristles of the toothbrush) following previous functional prime trials with different objects than when the primes trials were non-functionally related objects, etc.

We also manipulated how participants placed objects (Figure 1B). Consistent with DST and the A not B error model, we postulated that the temporal binding of spatial language to objects might also involve interaction with those objects. Placing a toothpaste tube over a toothbrush may call up an attractor state involving an action component as the toothpaste tube and toothbrush are held in specific ways associated with a brushing routine when those objects are in that relation. Participants either moved the objects on the touch screen with their hand upright (palm pointing downwards, Figure 1B, left panel), in a manner affording normal interaction with that object, or they moved the object with hand rotated in a manner that did not afford interaction (Figure 1B, right panel). We predicted that placements would be nearer the functional part of the other object when the movement was one that afforded action. Critically, we wanted to test whether this effect, if present, occurs for both functionally related and non-functionally related objects. If the effect only occurs for functionally related objects, one can argue that it is the action at encoding that it is important rather than any affordance to do with how the objects are moved per se. This was indeed what we found.

Overall results mirror the results from the A not B error task. Following the building of an attractor state over four prime trials, placement behavior reflecting comprehension of spatial language on the critical probe trials is dragged in the direction of previous object placements for incongruent prime-probe combinations - analogous to an infant searching in the wrong location on the A not B task. Second, the (incidental) way in which an object was moved on the screen also affected placement behavior, but only reliably so for F objects. This is consistent with the view that what objects are, how they interact, and how we interact with them becomes temporally coupled during learning, and forms a multimodal attractor state for spatial language.

Taking this data, we present a working DST computational model that also makes predictions tested in further later experiments. Overall, the novel approach to
word meaning allows the appearance of stable underlying "senses' of words while accounting for changes in meaning on a moment-to-moment basis.

## References/Indicative Publications

Carlson-Radvansky, L. A., Covey, E. S., \& Lattanzi, K. M. (1999). "What" effects on "where": Functional influences on spatial relations. Psychological Science, 10, 516-521.
Coventry, K. R., Prat-Sala, M., \& Richards, L. V. (2001). The interplay between geometry and function in the comprehension of "over", "under", "above" and "below". Journal of Memory and Language, 44, 376-398.
Coventry, K. R. (2013). On the mapping between spatial language and the vision and action systems. In Y. Coello \& A. Bartolo (Eds.), Language and Action in Cognitive Neuroscience, pp.209-223. Psychology Press.
Coventry, K. R. (2015). Space. In E. Dabrowska \& D. Divjak (Eds.), Handbook of Cognitive Linguistics, pp.489-507. De Gruyter Mouton.
Coventry, K. R., Christophel, T., Fehr, T., Valdés-Conroy, B., \& Herrmann, M. (2013). Multiple routes to mental animation: Language and functional relations drive motion processing for static images. Psychological Science, 24(8), 1379-1388.
Coventry, K. R. \& Garrod, S. C. (2004). Seeing, saying and acting. The psychological semantics of spatial prepositions. Psychology press: Taylor Francis.
Coventry, K. R., Griffiths, D., \& Hamilton, C. J. (2014). Spatial demonstratives and perceptual space: Describing and remembering object location. Cognitive Psychology, 69, 46-70.
Coventry, K. R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., \& Richardson, D. C. (2010). Spatial language, visual attention, and perceptual simulation. Brain \& Language, 112(3), 202-213.
Gudde, H. B., Coventry, K. R., \& Engelhardt, P. E. (2016). Language and memory for object location. Cognition, 153, 199-207.
Labov, W. (1973). The boundaries of words and their meanings. In C. J. Bailey \& R. Shuy (Eds.), New ways of analyzing variation in English. Washington, DC: Georgetown University Press.
Smith, L. B. \& Thelen, E. (2003). Development as a dynamic system. Trends in Cognitive Science, 7(8), 343348.

Thelen, E. et al. (2001) The dynamics of embodiment: a field theory of infant perseverative reaching. Behavioural and Brain Sciences, 24, 1-86
Wittgenstein, L. (1997). Philosophical investigations. Oxford, England: Blackwell Publishers. (Original work published 1953)

# Converging Evidence for Abstract Phonological Knowledge in Speech Processing 

Anne Cutler (a.cutler@westernsydney.edu.au)<br>The MARCS Institute, and ARC Centre of Excellence for the Dynamics of Language, Western Sydney University, Locked Bag 1797, Penrith NSW 2751, Australia;<br>Max Planck Institute for Psycholinguistics, P.O. Box 310, Nijmegen 6500 AH, The Netherlands.

Keywords: speech perception; phonological knowledge; first and second language; talker perception; infancy; abstraction

## Introduction

The perceptual processing of speech is a constant interplay of multiple competing albeit convergent processes: acoustic input vs. higher-level representations, universal mechanisms vs. language-specific, veridical traces of speech experience vs. construction and activation of abstract representations.

The present summary concerns the third of these issues. The ability to generalise across experience and to deal with resulting abstractions is the hallmark of human cognition, visible even in early infancy. In speech processing, abstract representations play a necessary role in both production and perception. New sorts of evidence are now informing our understanding of the breadth of this role. Two earlier and more detailed reviews of the role of abstraction in speech processing (Cutler, 2008; 2010) also embrace, respectively, evidence on the lexical representation of form versus meaning, and evidence on prosodic processing.

## Evidence from the second-language lexicon

Learners of a second language (L2) endure persistent perceptual trouble (not fixable just by accruing experience) with L2 phonemic distinctions that their first language (L1) lacks. The classic explanation of this is that L1 phonology (here, abstract knowledge about which phonemic contrasts may be encountered in speech signals) captures the input.

This is not the full account of this difficulty, however. Abstract knowledge of a contrast's existence (from reading, or from teaching, e.g., that light and write are supposed to be different) influences the construction of phonological representations in the lexicon. In the L2 lexicon, these representations thus become distinct. Speech perception, however, still fails to deliver the discrimination this requires (Weber \& Cutler, 2004; Cutler, Weber \& Otake, 2006; Broersma \& Cutler, 2008); both L2 sounds are perceived as (more or less good) realisations of a single phoneme. (That will usually be the L2 phoneme acoustically closest to the single L1 sound; for Japanese listeners hearing English [r/l], this is [1]). In righteous or rightful, the initial syllable will then actually activate light, not right. The second syllable will be needed in order to produce the desired lexical entry as the closest match to the input as a whole. The result is that word recognition in L2 is slower than it should be, because more competitor words are activated, and the competition from the spuriously activated ones is also more persistent (Broersma \& Cutler, 2011; Cutler, 2015).

## Evidence from talker adaptation

We adapt so rapidly to talkers we have never before heard by using existing knowledge to resolve phonetic ambiguity, and in consequence adjusting phoneme category boundaries for that specific talker (Norris, McQueen \& Cutler, 2003; Eisner \& McQueen, 2005). The adjustment generalises to words and phonetic contexts in which the phonemes in question have not previously been heard from the new talker (McQueen, Cutler \& Norris, 2006). Thus the adaptation has concerned phonemic categories, not veridical traces of experience (Cutler, 2010). Episodic models of lexical storage and retrieval, in which stored traces of lexical experience are activated in proportion to their match to the current input, cannot cope with this generalisation result (Cutler, Eisner, McQueen \& Norris, 2010), because the models are unable to assign the novel pronunciation instances uniquely to the phonemic category they should represent.

## Evidence from cross-modal generalisation

Cross-modal priming is popular in psycholinguistics, not necessarily because it calls on representations abstracted across different modalities; it is just a robust and useful task. Interestingly, recourse to the supra-modal representation even informs priming across modalities when the target is the same articulatory event - hearing words facilitates later phonological processing from lipreading the same spoken words, compared with new words (van der Zande, Jesse \& Cutler, 2014a). Notably, the lipreading here was facilitated whether or not the talker was the same one who had been heard in the priming phase; there was always an advantage for old words over new, but no effect of talker familiarity. Talker adaptation (as outlined in the section above) was likewise unaffected by visual information indicating another talker (van der Zande, Jesse \& Cutler, 2014b). These results confirm that phonological representations in the lexicon are shared across auditory and visual processing, and also show that talker information is not transferred across modalities at the lexical level. The abstract representations are stronger than, or unaffected by, modality-specific experience.

## Evidence from talker recognition

One of the best-known effects in talker recognition is that listeners find it easier to recognise talkers (pick them out from a set, as in a forensic lineup) when they are talking the listeners' native language. This turns out not to be due to a need to understand what is being said, because this nativelanguage effect appears even with seven-month-old infants:

Dutch-learning infants at this age perk up when a new talker is added to a set of three female talkers uttering unrelated (adult-style) Dutch sentences, but do not notice when a new talker is added to a set speaking Italian, or a set speaking Japanese (Johnson, Westrek, Nazzi \& Cutler, 2011).

At seven months, infants are acquiring the phonological structure of the language around them, but do not yet have a functional vocabulary that would allow them to understand such input. Thus the effect is here based on familiarity with the phonology in the one set of input but not in the others.

Analogously, adult listeners show equivalent efficiency with two phonologically comparable dialects of a language as opposed to a language with a differing phonology, and this works each way - native speakers of one of the dialects recognise talkers equally well in either dialect (but worse in the phonologically different language), while non-native listeners perform the talker recognition task equally badly in either dialect (but better in their phonologically different own tongue; Johnson, Bruggeman \& Cutler, in press). Again, the phonological familiarity predicts the results.

## Evidence from a lost language

Children adopted into another country lose all conscious knowledge of their first language and become essentially native speakers of a new language. But traces remain of the first, as many studies, with many languages, have shown. A recurring finding is that adoptees (in comparison to controls) show an accelerated trajectory of learning phonological structures found in the birth language but not in the current native tongue. In the largest such adoptee study so far, we replicated this for speech perception (Choi, Broersma \& Cutler, 2017), and also found that the perceptual mastery transferred to speech production (Choi, Cutler \& Broersma, 2017). This transfer, and a further generalisation of training on one phoneme contrast to other places of articulation, indicate that the observed benefit is based on abstract phonological representations. Most strikingly, the adoptee benefit was independent of age at adoption; infants adopted under the age of six months (before vocabulary building, or phoneme repertoire mastery, or talking) showed as much evidence of phonological retention as those adopted over the age of one. Thus abstract phonological knowledge is compiled and laid down even before six months of age, in preparation for the later stages of language acquisition.

## Conclusion

Abstract phonological knowledge plays a role in all aspects of speech processing. This is true even of those processing realms which may seem to form natural sources of evidence for memory-based effects. Thus we can see that abstractions are involved in many kinds of processing where differences between talkers are at issue. Likewise, though phonological structures are language-specific and hence not inborn, whereby language acquisition needs speech input to set it going, it also appears that construction of abstract phonological generalisations across this input must form part of linguistic processing even in the earliest months of life.

## Acknowledgments

Thanks to all collaborators listed in the References below, as well as to the Max Planck Society, NWO Spinoza and the Australian Research Council for financial support.

## References

Broersma, M. \& Cutler, A. (2008). Phantom word recognition in L2. System: An International Journal of Education Technology and Applied Linguistics, 36, 22-34.
Broersma, M. \& Cutler, A. (2011). Competition dynamics of second-language listening. Quarterly Journal of Experimental Psychology, 64, 74-95.
Choi, J., Broersma, M. \& Cutler, A. (2017). Early phonology: Birth-language retention from the earliest months of international adoptees' life. PNAS, to appear.
Choi, J., Cutler, A. \& Broersma, M. (2017). Early development of abstract language knowledge: Evidence from perception-production transfer of birth-language memory. Royal Society Open Science, 4, 160660.
Cutler, A. (2008). The abstract representations in speech processing. Quarterly Journal of Experimental Psychology, 61, 1601-1619.
Cutler, A. (2010). Abstraction-based efficiency in the lexicon. Laboratory Phonology, 1, 301-318.
Cutler, A. (2015). Representation of second language phonology. Applied Psycholinguistics, 36, 115-128.
Cutler, A., Eisner, F., McQueen, J.M. \& Norris, D . (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M.P. d’Imperio \& N. Vallée (Eds.), Papers in Laboratory Phonology 10. Berlin: Mouton de Gruyter.
Cutler, A., Weber, A. \& Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in secondlanguage listening. Journal of Phonetics, 34, 269-284.
Eisner, F. \& McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. Perception \& Psychophysics, 67, 224-238.
Johnson, E.K., Bruggeman, L. \& Cutler, A. (in press). Abstraction and the (misnamed) Language Familiarity Effect. Cognitive Science.
Johnson, E.K., Westrek, E., Nazzi, T. \& Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. Developmental Science, 14, 1002-1011.
McQueen, J.M., Cutler, A. \& Norris, D. (2006). Phonological abstraction in the mental lexicon. Cognitive Science, 30, 1113-1126.
Norris, D., McQueen, J.M. \& Cutler, A. (2003). Perceptual learning in speech. Cognitive Psychology, 47, 204-238.
Weber, A. \& Cutler, A. (2004). Lexical competition in nonnative spoken-word recognition. Journal of Memory and Language, 50, 1-25.
Zande, P. van der, Jesse, A. \& Cutler, A. (2014a). Hearing words helps seeing words: A cross-modal word repetition effect. Speech Communication, 59, 31-43.
Zande, P. van der, Jesse, A. \& Cutler, A. (2014b). Crossspeaker generalisation in two phoneme-level perceptual adaptation processes. Journal of Phonetics, 43, 38-46.

# Analogy and Episodic Memory to Support Domain Learning in a Cognitive Architecture: An Exploration 

Kenneth D. Forbus (forbus@ northwestern.edu)<br>Computer Science, Northwestern University, 2133 Sheridan Road<br>Evanston, IL, 60208, USA

Keywords: Analogy, Cognitive Architecture, Cognitive Simulation

## Introduction

Organisms learn from experience in many ways. One component of learning from experience is recording what has happened in the world when actions are taken, a form of episodic memory, and distilling such experience over time to learn models of phenomena for generating expectations. As further actions are taken, the accuracy of such models can be monitored, to detect surprises and to help identify and prioritize learning goals. This publication-based talk will describe some recent results in exploring the use of analogical generalization over episodic memories in the Companion cognitive architecture to formulate models of the effects of actions in a complex dynamic world. Measures of novelty, surprise and for prioritization of learning goals will be discussed.

## Episodic Memory and Analogy

How human episodic memory is organized is still an open question. Given the centrality of analogy in human cognition (e.g. Gentner, 2003), it seems reasonable that a common way of structuring episodic memories could be as cases, so that they can be accessed via analogical retrieval (e.g. MAC/FAC, Forbus et al. 1995) with more transferable knowledge constructed incrementally via generalization (e.g. SAGE, McLure et al. 2015). The Companion cognitive architecture (Forbus et al. 2009; Forbus 2016; Forbus \& Hinrichs, in press) incorporates these analogical processing models, along with SME (Forbus et al 2016), which MAC/FAC and SAGE are built upon. The Companion architecture also includes facilities for language understanding, sketch understanding (Forbus et al. 2011), and integration with simulators. For example, Companions can interact with Freeciv ${ }^{1}$, an opensource version of Civilization 2, which is a popular strategy game. The attraction of such games to players is their complexity, e.g. building civilizations and transportation networks, exploration, technology research, military operations, over hundreds of turns. Such complexity makes Freeciv useful for exploring learning about complex dynamics (McFate et al. 2014; Hinrichs \& Forbus, 2016). For example, by storing cases of both positive outcomes and negative outcomes generated by experimentation, a

Companion has learned to perform city management (Hinrichs \& Forbus, 2007).

This talk goes beyond that work by focusing on how a Companion can distill models of actions via analogical generalization while observing human players. For each action the person takes, the Companion records information about the state of the world before and after the action, and uses some general-purpose heuristics to attempt to explain immediate events in terms of the action. For each occurrence of each action, a case consisting of this information is stored. Storage occurs via a SAGE generalization pool for each command (e.g., doMove, doIrrigate). The generalization pool for a command can be thought of as an analogy-derived model for what happens when that command is used. By letting the system watch replays from six different games, it builds up over 4,200 cases across 34 different commands.

Inspecting these generalization pools leads to some interesting insights. First, the number of generalizations and outliers in a pool provides an indication of how well the action is understood. If there are many cases all forming a single generalization, then that command has straightforward local consequences (e.g. doIrrigate). When there are multiple generalizations, comparing their structures can be illuminating: For example, in doResearch, the generalizations differ only in the number of requirements and opportunities, making them artifacts of the encoding strategy, which could be eliminated via re-representation. Thus properties of the generalization pools provide a signal about how encoding strategies might be improved.

Analogical generalization also provides a means of detecting and quantifying novelty and surprise. Novelty can be detected in two ways: Failure to retrieve a similar experience, and by analysis of candidate inferences indicating differences. When little is known, all is novel surprise, I argue, occurs when a novel situation is experienced for a type of situation that was considered to already be well understood. The degree of surprise can be estimated based on the number of cases in the pool and frequency information for relationships within them that are computed for the generalizations: When there are many cases and highly certain outcomes, a new outcome can be more surprising. doMove provides an excellent example: It occurs frequently, so a dominant analogical model is quickly built up. But when a unit moves into a hut, there are five different things that might happen, leading initially to surprises.

[^258]In addition to summarizing the results of these experiments, I will describe work in progress on making adaptable encoding strategies guided by the system's own analysis of its experience.

## Acknowledgments

This research is sponsored by the AFOSR Computational Cognition and Machine Intelligence Program.

## References

Forbus, K. (2016). Software Social Organisms: Implications for Measuring AI Progress. AI Magazine. 37(1):85-90.
Forbus, K., Ferguson, R., Lovett, A., \& Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. Cognitive Science, DOI:10.1111/cogs.12377, pp. 1-50.
Forbus, K., Gentner, D. \& Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. Cognitive Science, 19:141-205.
Forbus, K. \& Hinrichs, T., (in press). Analogy and Qualitative Representations in the Companion Cognitive Architecture. AI Magazine.
Forbus, K., Klenk, M., \& Hinrichs, T. (2009). Companion Cognitive Systems: Design Goals and Lessons Learned So Far. IEEE Intelligent Systems, 24(4):36-46
Forbus, K., Usher, J., Lovett, A., Lockwood, K., \& Wetzel, J. (2011). CogSketch: Sketch understanding for Cognitive Science Research and for Education. Topics in Cognitive Science, 3(4):648-666.
Gentner, D. (2003) Why we're so smart. In D. Gentner \& S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought. MIT Press
Hinrichs, T., \& Forbus, K. (2007). Analogical Learning in a Turn-based Strategy Game. Proceedings of IJCAI-2007.
Hinrichs, T. and Forbus, K. (2016). Qualitative Models for Strategic Planning. In Advances in Cognitive Systems, Volume 4, 2016, pages 75-92.
McFate, C., Forbus, K., \& Hinrichs, T. (2014). Using Narrative Function to Extract Qualitative Information from Natural Language Texts. Proceedings of AAAI-2014.
McLure, M.D., Friedman S.E. and Forbus, K.D. (2015). Extending Analogical Generalization with Near-Misses. Proceedings of AAAI-2015.
Tomai, E. and Forbus, K. (2009). EA NLU: Practical Language Understanding for Cognitive Modeling. Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference. Sanibel Island, Florida.

# Non-syntactic Processing Explains Cortical Entrainment During Speech Perception 

Stefan L. Frank (s.frank@let.ru.nl)<br>Centre for Language Studies, Radboud University<br>P.O. Box 9103,6500 HD Nijmegen, The Netherlands<br>Jin-Biao Yang (ray306@gmail.com)<br>Institute of Brain and Cognitive Science, NYU Shanghai<br>3663 North Zhongshan Road, Shanghai, China 200062

Keywords: Language processing; cortical entrainment; distributional semantics

## Introduction

There is considerable debate on the precise role of hierarchical syntactic structure during the comprehension of sentences, with some arguing that a full hierarchical analysis is required for comprehension (e.g., Ding, Melloni, Tian, \& Poeppel, 2017) and others claiming that non-hierarchical processing is more common (e.g., Frank, Bod, \& Christiansen, 2012). Ding, Melloni, Zhang, Tian, and Poeppel (2016) recently presented evidence that cortical entrainment during speech perception reflects the neural tracking of hierarchical syntactic structure of simple sentences, which would support the view that hierarchical processing is unavoidable. However, we show that the same entrainment effects appear in a computational model that does not incorporate syntax or any other linguistic knowledge or process beyond the word level. Hence, the cortical entrainment results do not need to be indicative of syntactic processing.

Ding et al. (2016) had participants listen to Chinese or English four-syllable sentences, with syllables presented at a fixed rate while cortical activity was recorded with MEG. A frequency analysis of the MEG signal revealed peaks in the power spectrum at exactly the occurrence frequencies of syllables, phrases, and sentences. For example, when sentences with [NP VP] structure (such as "dry fur rubs skin") were presented at a rate of 250 ms per monosyllablic word, peaks would appear at $4 \mathrm{~Hz}, 2 \mathrm{~Hz}$, and 1 Hz , corresponding to the syllable/word, phrase, and sentence rate, respectively (see Figure 1). Likewise, a sequence consisting of only NPs or only VPs resulted in peaks at 4 Hz and 2 Hz but not 1 Hz , while presenting a sequence of Chinese syllables without word or phrase structure resulted in only the 4 Hz peak.

Although these results can indeed be interpreted in term of the sentences' syntactic structures, we propose a simpler explanation: The power spectrum merely reflects responses to regularities in word-level properties, such as (approximate) syntactic or semantic category. For example, in the [NP VP] sentences, verbs occur at 1 Hz and nouns at 2 Hz . We implemented this alternative explanation in a simple computational model and show that it indeed predicts the MEG power spectra in different experimental conditions.

## The model

The only linguistic knowledge available to the model is encoded in word vector representations. These were generated by a distributional semantics model (Mikolov, Chen, Corrado, \& Dean, 2013) trained on large corpora of Chinese or English texts (the same model and English corpus were used to obtain word representations that Frank \& Willems, in press, and Frank, 2017, applied to account for N400 and reading time effects). Words that occur in similar contexts get similar vectors so that representations of words from the same syntactic/semantic category tend to be clustered together.

The stimuli from the (Ding et al., 2016) experiments were presented to the model at a simulated rate of 4 Hz per English word or Chinese syllable. Twelve different subjects were simulated by retraining the distributional semantic model and randomly varying stimuli presentation order. The sequence of vector representations, at a simulated time resolution of 5 ms , were analysed by applying a Discrete Fourier Transform to obtain a power spectrum, just like Ding et al. (2016) do in their analysis of the MEG signal.

## Results

Figure 2 shows that the model predicts the same peaks in the power spectrum as in the original MEG study. The minor peak at 3 Hz , which did not reach significance in the MEG data, is most likely merely the second subharmonic of the 1 Hz peak (Zhou, Melloni, Poeppel, \& Ding, 2016). The model further correctly accounts for the outcomes of experiments with two-syllable NP or VP sequences that lack full sentence structure, and predicts results very similar to those in the MEG data when syllable sequences are scrambled to remove any higher linguistic structure (see Frank \& Yang, 2017)

## Conclusion

The only linguistic knowledge in the model is encoded in the input vectors, so it remains at the lexical level. Furthermore, the model does not include any intergrative processing. The resulting power spectra can therefore not reflect any (hierarchical) syntactic processing. Consequently, the original MEG results may also be explained without syntax.



Figure 1: Left: MEG results for Chinese [NP VP] sentences. Right: MEG results for English [NP VP] sentences, reproduced from Ding et al. (2016, Figure 2e) with permission (the frequency scale was adapted to match simulated presentation rate). Shaded areas represents the standard error over subjects; lines are the average over subjects. Stars indicate significant peaks after multiple comparison correction.


Figure 2: Model results for Chinese (left) and English (right) [NP VP] sentences. Grey lines represent individual simulated subjects; coloured lines are the averages over simulated subjects. Stars indicate significant peaks after multiple comparison correction.

## References

Ding, N., Melloni, L., Tian, X., \& Poeppel, D. (2017). Rulebased and word-level statistics-based processing of language: insights from neuroscience. Language, Cognition and Neuroscience, 32, 570-575.
Ding, N., Melloni, L., Zhang, H., Tian, X., \& Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience, 19, 158164.

Frank, S. L. (2017). Word embedding distance does not predict word reading time. In Proceedings of the 39th Annual Conference of the Cognitive Science Society.
Frank, S. L., Bod, R., \& Christiansen, M. H. (2012). How hierarchical is language use? Proceedings of the Royal Society B: Biological Sciences, 279, 4522-4531.
Frank, S. L., \& Willems, R. M. (in press). Word predictability and semantic similarity show distinct patterns of brain activity in language comprehension. Language, Cognition and Neuroscience.
Frank, S. L., \& Yang, J. (2017). Lexical representation explains cortical entrainment during speech perception. Manuscript submitted for publications.
Mikolov, T., Chen, K., Corrado, G., \& Dean, J. (2013). Efficient estimation of word representations in vector space. In

## Proceedings of the ICLR Workshop.

Zhou, H., Melloni, L., Poeppel, D., \& Ding, N. (2016). Interpretations of frequency domain analyses of neural entrainment: periodicity, fundamental frequency, and harmonics. Frontiers in Human Neuroscience, 10, 274.

# Comparative analysis of visual category learning 

John H. Freeman (iohn-freeman@uiowa.edu)<br>Matthew B. Broschard (matthew-broschard@uiowa.edu)<br>Jangjin Kim (jangiin-kim@uiowa.edu)<br>Leyre Castro (leyre-castroruiz@uiowa.edu)<br>Edward A. Wasserman (ed-wasserman@uiowa.edu)<br>Department of Psychological and Brain Sciences, The University of Iowa 301 East Jefferson Street, Iowa City, IA 52242, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Department of Psychology, The Ohio State University
1835 Neil Avenue, Columbus, OH 43210, USA

Keywords: category learning; category density; supervision; comparative; pigeon; rat

## Introduction

Categorization is critical for our ability to organize information. A comparative analysis may provide important insights into the cognitive and neural mechanisms underlying category learning. We have examined category learning in rats and pigeons because of differences in brain organization between mammals and birds. Species differences in category learning and representation can indicate how the differences in brain organization lead to differences in cognition.

Category structure and supervision are factors importantly influencing category learning and representation in humans (Kloos \& Sloutsky, 2008; Love, 2002). Humans can learn categories with dense defining features with no supervision, but we need supervision to learn categories with sparse features. These finding have been interpreted as evidence for multiple category learning systems in the brain. The current study examined the roles of feature density and supervision in visual category learning in rats and pigeons.

## Experiment 1

Rats were trained on a discrimination task earlier used for category learning with photographic stimuli (Brooks et al., 2013). Rats were trained with two visual categories in which feature density could be precisely manipulated (see Figure 1). One category was associated with a left response, whereas the other category was associated with a right response.


Figure 1: Exemplars of one high-density (dense) category (top) and one low-density (sparse) category (bottom).

Each stimulus category had five features. In the dense condition, three of the features were category-relevant, whereas the sparse condition had only one relevant feature. High supervision was defined as delivery of a food reward only after a correct choice. In contrast, low supervision was defined as delivery of a food reward regardless of whether or not the choice was "correct." Rats were trained in a $2 \times 2$ design with density and supervision as factors: dense-high supervision, sparse-high supervision, dense-low supervision, and sparse-low supervision. The rats were trained until reaching a criterion of $75 \%$ correct responding for both categories for 2 consecutive days or for a maximum of 60 days. After meeting the training criterion, the rats were given testing sessions in which training stimuli were mixed with probe trials. Probe trials included novel exemplars (novel irrelevant features), rotated stimuli (in which the relevant features appeared in different locations), and singleton stimuli (only one relevant feature was presented, in the absence of any other features).


Figure 2: Mean accuracy of rats trained in the dense-high supervision (dense-high), sparse-high supervision (sparsehigh), dense-low supervision (dense-low), and sparse-low supervision (sparse-low) conditions.

All rats in the dense-high supervision condition (6/6) showed rapid learning of the two categories (see Figure 2). They also showed very high accuracy with novel stimuli. Accuracy
dropped significantly when rotated or singleton stimuli were presented, suggesting that the rats' representations of the categories included feature-location binding. Some of the rats in the sparse-high supervision condition (2/6) learned and showed substantial generalization to novel exemplars. Like rats in the dense-high supervision condition, the rats in the sparse-high supervision condition that learned showed a significant drop in accuracy during presentations of the rotated and singleton test stimuli. Rats trained in the low supervision conditions did not learn. Only one rat in the dense-low supervision condition reached criterion performance. Rats that did not learn showed a position bias and were slower to learn when switched to high supervision.

## Experiment 2

Pigeons were trained and tested under identical conditions as the rats. This experiment is in progress and, currently, training and testing has been conducted with a limited number of animals: dense-high supervision (2), sparse-high supervision (3), dense-low supervision (2), and sparse-low supervision (3). All of these pigeons learned relatively quickly, compared to the rats.


Figure 3: Mean accuracy of pigeons trained in the dense-high supervision (dense-high), sparse-high supervision (sparsehigh), dense-low supervision (dense-low), and sparse-low supervision (sparse-low) conditions.

As can be seen in Figure 3, pigeons in the high density conditions learned rapidly, in five or fewer sessions, regardless of the level of supervision. They also showed very high accuracy to novel stimuli (above 90\%), and to rotated and singleton stimuli as well (above 90\%). Pigeons in the sparse conditions took longer to learn but, just as in the dense conditions, the level of supervision minimally affected their rate of learning. In both sparse conditions, accuracy to novel stimuli was high, albeit lower than in the dense conditions (85\%). Accuracy to the singleton stimuli was high as well (85\%), but it dropped a bit more for the rotated stimuli (75\%). Pigeons' representations of the categories seemed to include feature-location binding as well, just as we observed in the rats; however, this factor played a much smaller role in the pigeons' performance.

## Conclusions

The results indicate clear differences in category learning between rats and pigeons. Pigeons learned rapidly in all four
conditions and their learning rate was not affected by the level of supervision. For pigeons, the most important factor was the density of category-relevant features-dense categories were learned faster than sparse categories. In contrast, rats showed robust learning only in the dense-high supervision condition. Statistical density is therefore a crucial factor for visual category learning in birds, rodents, and humans. The interaction of density and supervision is more complex, however, and may be related to whether the organism is remembering visual features, binding features and spatial locations, or learning category rules.

The differences in category learning between pigeons and rats may reflect differences in brain organization. Birds do not have a laminar cortex or a prefrontal cortex. Thus, the pigeons' insensitivity to the level of supervision might be related to the absence of prefrontal processing of differential reinforcement. The clear superiority in learning rate in the pigeons relative to the rats suggests, however, an advantage in memory for visual stimuli, which might be related to specializations within the visual areas of the avian brain.

## Acknowledgments

This research was supported by the Eunice Kennedy Shriver National Institute of Child Health \& Human Development of the NIH under Award Number P01HD080679.

## Cited and Extended References

Brooks, D.I., Ng, K.H., Buss, E.W., Marshall, A.T., Freeman, J.H., \& Wasserman, E.A. (2013). Categorization of photographic images by rats using shape-based image dimensions. Journal of Experimental Psychology: Animal Behavior Processes, 39, 85-92.
Farley, S.J., Radley, J.J., \& Freeman, J.H. (2016). Amygdala modulation of cerebellar learning. The Journal of Neuroscience, 36, 2190-2201.
Kim, J., Goldsberry, M.E., Harmon, T.C., \& Freeman, J.H. (2016). Developmental changes in hippocampal CA1 single neuron firing and theta activity during associative learning. PloS One, 11, 1-22.
Kim, J., Wasserman, E.A., Castro, L., \& Freeman, J.H. (2016). Anterior cingulate cortex inactivation impairs rodent visual selective attention and prospective memory. Behavioral Neuroscience, 130, 75-90.
Kloos, \& Sloutsky, V. M. (2008). What's behind different kinds of kinds: effects of statistical density on learning and representation of categories. Journal of Experimental Psychology: General, 137, 52-72.
Love, B. (2002). Comparing supervised and unsupervised category learning. Psychonomic Bulletin \& Review, 9, 829835.

Steinmetz, A.B., Harmon, T.C., \& Freeman, J.H. (2013). Visual cortical contributions to associative cerebellar learning. Neurobiology of Learning and Memory, 104, 103-109.
Wasserman, E.A., Castro, L., \& Freeman, J.H. (2012). Samedifferent categorization in rats. Learning \& Memory, 19, 142-145.

# Cognitive mechanisms for imitation and the detection of imitation in human dyadic interactions 

Antonia F. de C Hamilton (a.hamilton@ucl.ac.uk)<br>Institute of Cognitive Neuroscience, University College London<br>Alexandra House, 17 Queen Square, London, WC1N 3AR

Keywords: imitation; mimicry; social interaction; eye contact; gaze; signaling;

## Introduction

Imitation is a ubiquitous human behavior which has been linked to both social learning and social bonding (Uzgiris, 1981). Here, we examine how imitation is used in the context of social affiliation, with a particular focus on the unconscious mimicry of body postures or gestures which are sometimes referred to as the 'chameleon effect' (Chartrand \& Bargh, 1999). The 'social glue' hypothesis of mimicry claims that mimicry behavior has a key causal role in social affiliation (Lakin, Jefferis, Cheng, \& Chartrand, 2003; van Baaren, Janssen, Chartrand, \& Dijksterhuis, 2009). For example, if Anna mimics an action by Bert (without awareness in either), the theory claims that Anna sends a prosocial signal to Bert and Bert receives that information (Wang \& Hamilton, 2012).

A correlational relationship between increased bodily mimicry in dyadic interactions and positive ratings of the interaction has been repeatedly observed (Pentland, 2008). However, direct experimental evidence that mimicry has a social signaling role remains weak. Two major types of evidence can show if an action functions as a social signal first, does the sender's behavior change depending on who can see the signal? and second, does the receiver act on the signal? This talk will examine recent evidence for each of these, and will thus test the social glue hypothesis of mimicry.

## Is mimicry changed by who can see?

Several studies suggest that mimicry is enhanced when another person makes eye contact and can receive a signal from the mimicker (Bavelas, Black, Lemery, \& Mullett, 1986; Wang, Newport, \& Hamilton, 2011). Here I will focus on three recent studies which show how mimicry in children and adults is modulated by the gaze behaviour of an observer. First, we report that children performing an overimitation task (similar to Marsh, Ropar, \& Hamilton, 2014) show more imitation behaviour when observed by an adult than when the adult turns her back (Marsh \& Hamilton, n.d.). Second, we show that rapid hand action mimicry is enhanced when the participant is observed at the time of response, but not if the observer's gaze is occluded just before the response (Wang \& Hamilton, 2013). In a third study, we asked dyads to complete a leader-follower
task where the leader demonstrated a movement sequence and we measured how closely the follower copied the kinematics of the sequence despite not being instructed to do so. We find that followers imitate with higher fidelity when the leaders eyes are open than when they are closed (Krishnan-Barman \& Hamilton, n.d.), matching the predictions of the signaling hypothesis.

Together, this series of studies provides clear evidence that the production of mimicry behavior varies according to whether the mimicry can be seen by another person or not. This is true across children, adult reaction time studies and adult dyadic interactions. These results are compatible with the idea that senders are producing mimicry as a social signal, in order to convey information to another person.

## Is mimicry detected by receivers?

For mimicry to function effectively as a signal, the message must be send and also received. That is, Bert must (on some level) detect that Anna is mimicking his action and respond to that signal. It is hard to find strong evidence for this, partly because it is not an easy experiment to implement. Most approaches require that a confederate should mimic or not-mimic the actions of a participant in a well-controlled manner. While some studies report positive effects (Chartrand \& Bargh, 1999; Müller, Maaskant, van Baaren, \& Dijksterhuis, 2012), others report mixed results or null effects (van Swol, 2003; Verberne, Ham, Ponnada, \& Midden, 2013). A full review of these results is provided in (Hale \& Hamilton, 2016a).

We propose that the most rigorous way to test the hypothesis that being mimicked leads to a positive social effect is to use virtual reality. In virtual reality mimicry, the experimenter has full control of the interaction and can ensure that mimicry (and only mimicry) is the factor which differs between experimental conditions, and that all participants receive a consistent experience. An early virtual reality study reported positive effects of being mimicked in VR (Bailenson \& Yee, 2005). We recently extended this result and examined how participants respond to being mimicked or not by a virtual character from their own culture or a different culture. In a pre-registered study with a large sample size, we find that mimicry of head motion which is not detected by participants has no impact on rapport or trust (Hale \& Hamilton, 2016b).

Several factors could account for this null result. First, we examined only mimicry of head motion, and mimicry of other motion features (e.g. gesture or posture) might lead to larger effects. Second, imperfections in the VR itself might negate any positive social effects, through similar VR systems can replicate many other psychological phenomena. Finally, it is possible that being mimicked is not implicitly detected as a social signal, arguing against the social signaling interpretation of mimicry.

## Conclusions

The present data suggests that the production of mimicry depends on who can see, but it is not yet clear if receivers respond positively to mimicry. This means that the role of mimicry as a social signal is not yet firmly established. We suggest that acquiring high-resolution motion capture data to better establish how dyads use mimicry will also enable the creation of better VR mimicry. This can provide a more definitive test of the claim that mimicry is used as a signal of social affiliation.

## Acknowledgments

AH is supported by ERC grant 313398 INTERACT

## References

Bailenson, J. N., \& Yee, N. (2005). Digital chameleons: automatic assimilation of nonverbal gestures in immersive virtual environments. Psychol Sci, 16(10), 814-819.
Bavelas, J. B., Black, A., Lemery, C. R., \& Mullett, J. (1986). "I show how you feel": Motor mimicry as a communicative act. Journal of Personality and Social Psychology, 50(2), 322-329.
Chartrand, T. L., \& Bargh, J. a. (1999). The chameleon effect: the perception-behavior link and social interaction. Journal of Personality and Social Psychology, 76(6), 893-910.
Hale, J., \& Hamilton, A. F. de C. (2016a). Cognitive mechanisms for responding to mimicry from others. Neuroscience \& Biobehavioral Reviews, 63, 106-123.
Hale, J., \& Hamilton, A. F. de C. (2016b). Testing the relationship between mimicry, trust and rapport in virtual reality conversations. Scientific Reports, 6, 35295.

Krishnan-Barman, S., \& Hamilton, A. F. de C. (n.d.). Spontaneous mimicry in dyads is modulated by gaze. In Prep.
Lakin, J. L., Jefferis, V. E., Cheng, C. M., \& Chartrand, T. L. (2003). THE CHAMELEON EFFECT AS SOCIAL GLUE: EVIDENCE FOR THE EVOLUTIONARY SIGNIFICANCE OF NONCONSCIOUS MIMICRY. Journal of Nonverbal Behavior, 27(3), 145-162.
Marsh, L. E., \& Hamilton, A. F. de C. (n.d.). Modulation of
over-imitation by observation in childhood. In Prep.
Marsh, L. E., Ropar, D., \& Hamilton, A. F. de C. (2014). The Social Modulation of Imitation Fidelity in School-Age Children. PLoS ONE, 9(1), e86127.
Müller, B. C. N., Maaskant, A. J., van Baaren, R. B., \& Dijksterhuis, A. P. (2012). Prosocial consequences of imitation. Psychological Reports, 110(3), 891-898.
Pentland, A. (2008). Honest Signals. MIT Press.
Uzgiris, I. C. (1981). Two Functions of Imitation During Infancy. International Journal of Behavioral Development, 4(1), 1-12.
van Baaren, R. B., Janssen, L., Chartrand, T. L., \& Dijksterhuis, A. (2009). Where is the love? The social aspects of mimicry. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364(1528), 2381-9.
van Swol, L. M. (2003). The Effects of Nonverbal Mirroring on Perceived Persuasiveness, Agreement with an Imitator, and Reciprocity in a Group Discussion. Communication Research, 30(4), 461480.

Verberne, F. M. F., Ham, J., Ponnada, A., \& Midden, C. J. H. (2013). Trusting Digital Chameleons: The Effect of Mimicry by a Virtual Social Agent on User Trust (pp. 234-245).
Wang, Y., \& Hamilton, A. F. de C. (2012). Social top-down response modulation (STORM): a model of the control of mimicry in social interaction. Frontiers in Human Neuroscience, 6(June), 153.
Wang, Y., \& Hamilton, A. F. de C. (2013). Why does gaze enhance mimicry? Placing gaze-mimicry effects in relation to other gaze phenomena. Quarterly Journal of Experimental Psychology (2006), (August), 37-41.
Wang, Y., Newport, R., \& Hamilton, A. F. de C. (2011). Eye contact enhances mimicry of intransitive hand movements. Biology Letters, 7(1), 7-10.

# Deep Networks as Models of Human and Animal Categorization 

Bradley C. Love (b.love@ucl.ac.uk)<br>Olivia Guest (o.guest@ucl.ac.uk)<br>Department of Experimental Psychology, University College London<br>London, UK

Piotr Slomka (piotr.slomka@cshs.org)<br>Department of Imaging Cedars-Sinai Medical Center<br>David Geffen School of Medicine, University of California Los Angeles<br>Los Angeles, California, USA

Victor M. Navarro (victor-navarro@uiowa.edu)
Edward Wasserman (ed-wasserman@uiowa.edu)
Department of Psychological \& Brain Sciences, The University of Iowa Iowa City, Iowa, USA

Keywords: Categorization; Neural Network; Deep Learning; Convolutional Neural Network; Animal Learning; Object Recognition; Neural Computation.

## Introduction

Convolutional neural networks (CNNs) trained as classifiers learn by associating visual inputs (e.g., photographs of objects) with appropriate output labels (e.g., "crow", "dog", "car"). These complex models, which contain millions of weights, are the state-of-the art in machine vision, rivaling humans in object recognition tasks (LeCun, Bengio, \& Hinton. 2015, Krizhevsky, Sutskever, \& Hinton, 2012). What these networks learn displays some commonalities with human learning (Kubilius, Bracci, \& de Beeck, 2016, Lake, Zaremba, Fergus, \& Gureckis, 2015). Furthermore, the layers in these networks have been related to neural activity along the ventral stream (Khaligh-Razavi \& Kriegeskorte, 2014, Yamins \& DiCarlo, 2016)

The similarity spaces created by these models at various network layers allow us to draw parallels with the brain's neural coding schemes (Guest \& Love, 2017). At earlier layers, networks display similarity spaces that reflect the high-level categories found in the input space, e.g., lions and tigers are more similar to one another than to mopeds. At the more advanced layers, similarity structure tends to break down such that representations of different object categories become orthogonal.

Can these networks also shed light on how non-human animals categorize? CNNs can be used to determine at what level of representation (i.e., what network layer) animals are coding similarities between images. For example, are animals learning regularities at a very low level, close to the pixels in the image, or are they seizing upon more abstract shape features? In this contribution, we address this question by examining data from pigeons trained to categorize images of cardiograms as normal or abnormal.

Pigeons are excellent at classifying visual stimuli (Bhatt, Wasserman, Reynolds, \& Knauss, 1988). For example, pigeons trained to discriminate between medical images of nor-
a)

b)


Figure 1: Two examples of the stimuli that the pigeons and network are asked to classify: a) a normal cardiogram without any perfusion damage; and b) an abnormal cardiogram with total perfusion damage 20 (of a maximum of 51 ).
mal and cancerous breast tissue generalized to novel stimuli and attained human-level accuracy (Levenson, Krupinski, Navarro, \& Wasserman, 2015). Importantly, knowledge transfer was only true in certain circumstances. Pigeons only generalized within image magnification levels - they were not scale-invariant. Also, generalization was significantly compromised, although still above chance, when tested on grayscale images (perhaps to be expected given the loss of hue and brightness cues). However, the pigeons' performance improved with additional training on greyscale images.

Can CNNs explain such patterns of performance? At the most advanced layers of these networks, representations should be somewhat invariant to changes in size, luminance, translation, etc. However, at lower layers the network will be more sensitive to such changes and will not generalize as broadly. Which network layer best captures how pigeons categorize?

Here we consider data from an a yet unpublished study by Wasserman and colleagues in which pigeons are trained to classify cardiograms as normal or abnormal, see Figure 1. Pigeons can correctly determine whether a cardiogram is ab-
normal or normal in much the same way as a skilled human, and can correctly classify unseen cardiogram images.

To parallel the pigeons, we also show the same stimuli to a CNN, namely Inception-v3 GoogLeNet (Krizhevsky et al., 2012). In line with the pigeons, the network can also determine whether a stimulus is normal or abnormal. Also like the pigeons, Inception-v3 GoogLeNet is very sensitive to changes in color, having serious problems generalizing when trained on color images and tested on grayscale without additional training. Importantly, even though the model can differentiate between the two classes at the output layer it can also do so at much lower layers. The output layer is trained to represent very high-level conceptual categories ( 1000 mutually exclusive classes, e.g., sunglasses, moped, jellyfish, etc.). Although these output classes do not contain options for normal and abnormal cardiograms, the network provides a distributed answer across these categories thus solving the classification task. In other words, the output shows a similarity structure matching the normal/abnormal distinction in the inputs.

As mentioned, at lower layers including the input layer, the network can also differentiate the two types of stimuli into normal and abnormal. This means that basic stimulus properties, which are what the network and the pigeons are extracting and learning, are sufficient to separate the two classes of cardiograms shown in Figure 1. This is important because it implies that more complex and abstract features, or even representations of basic shapes, are not required for the type of learning problem the pigeons are solving. In addition, this predicts that generalization will be poor in both the animal and computational models we have considered. We consider the broader implications of these results for how humans and non-human animals categorize.

## Six Relevant Papers by BCL

Gigure, G. \& Love, B.C. (2013). Limits in decision making arise from limits in memory retrieval. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 110 (19), 7613-7618.

Guest, O., Love, B.C. (2017). What the Success of Brain Imaging Implies about the Neural Code. eLife,6:e21397.

Love, B.C. (2015). The Algorithmic Level is the Bridge Between Computation and Brain. Topics in Cognitive Science, 7, 230-242.

Mack, M.L., Love, B.C., \& Preston, A.R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. Proceedings of the National Academy of Sciences (PNAS), 113(46), 1320313208.

Mack, M.L., Preston, A.R. \& Love, B.C. (2013). Decoding the Brain's Algorithm for Categorization from its Neural Implementation. Current Biology, 23, 2023-2027.

Turner, B.M., Forstmann, B.U., Love, B.C., Palmeri, T.J. \& Van Maanen, L. (2016). Approaches to Analysis in Modelbased Cognitive Neuroscience. Journal of Mathematical Psychology. http://dx.doi.org/10.1016/j.jmp.2016.01.001.

## Acknowledgments

This work was supported by NIH (Grant 1P01HD080679) to BCL and EW. We thank all the members of the P01 team, led by Vladimir Sloutsky. Additional support from this work comes from the Leverhulme Trust (Grant RPG-2014-075) and a Wellcome Trust Investigator Award (Grant WT106931MA) to BCL, as well as The Alan Turing Institute under the EPSRC grant EP/N510129/1.

## References

Bhatt, R., Wasserman, E., Reynolds, W., \& Knauss, K. (1988). Conceptual behavior in pigeons: Categorization of both familiar and novel examples from four classes of natural and artificial stimuli. Journal of Experimental Psychology: Animal Behavior Processes, 14(3), 219.
Guest, O., \& Love, B. C. (2017, jan). What the success of brain imaging implies about the neural code. eLife, 6 , e21397. doi: 10.7554/eLife. 21397
Khaligh-Razavi, S.-M., \& Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol, 10(11), 129. doi: 10.1371/journal.pcbi. 1003915

Krizhevsky, A., Sutskever, I., \& Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
Kubilius, J., Bracci, S., \& de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. PLoS Comput Biol, 12(4), e1004896.
Lake, B. M., Zaremba, W., Fergus, R., \& Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In Proceedings of the annual meeting of the cognitive science society (p. 1-6).
LeCun, Y., Bengio, Y., \& Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. doi: 10.1038/nature14539
Levenson, R. M., Krupinski, E. A., Navarro, V. M., \& Wasserman, E. A. (2015, 11). Pigeons (columba livia) as trainable observers of pathology and radiology breast cancer images. PLOS ONE, $10(11), 1-21$. doi: 10.1371/ journal.pone. 0141357
Yamins, D. L., \& DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), 356-365. doi: 10.1038/nn. 4244

# A Unified Model of Entropy and the Value of Information 

Jonathan D Nelson (jonathan.d.nelson@gmail.com) ${ }^{1,2}$ Vincenzo Crupi (vincenzo.crupi@unito.it) ${ }^{3}$ Björn Meder (meder@mpib-berlin.mpg.de) ${ }^{1}$ Gustavo Cevolani (g.cevolani@gmail.com) ${ }^{2}$ Katya Tentori (katya.tentori@unitn.it) ${ }^{4}$<br>${ }^{1}$ Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany<br>${ }^{2}$ School of Psychological Sciences, University of Surrey, Guildford, UK<br>${ }^{3}$ Center for Logic, Language and Cognition, Department of Psychology and Education, University of Turin, Italy<br>${ }^{4}$ Center for Mind and Brain Sciences, University of Trento, Italy

Keywords: uncertainty; entropy; information; information gain, probability gain


#### Abstract

Notions of entropy and uncertainty are fundamental to many domains, ranging from the philosophy of science to physics. One important application is to quantify the expected usefulness of possible experiments (or questions or tests). Many different entropy models could be used; different models do not in general lead to the same conclusions about which tests (or experiments) are most valuable. It is often unclear whether this is due to different theoretical and practical goals or are merely due to historical accident. We introduce a unified two-parameter family of entropy models that incorporates a great deal of entropies as special cases. This family of models offers insight into heretofore perplexing psychological results, and generates predictions for future research.


## Uncertainty and Information

Notions of entropy and uncertainty are fundamental to many domains, ranging from the philosophy of science to physics. One important application of uncertainty is to quantify the expected usefulness of possible experiments (or questions or tests). Lindley (1956) suggested that an experiment's usefulness could be quantified in terms of how much it reduces expected Shannon (1948) uncertainty about the possible states of the world. This idea has proven useful in psychological models (Oaksford \& Chater, 1994) as well. In a psychological context, the possible states could be the different categories that an object might belong to, and "experiments" could be a child's queries to learn more about the category. Other entropy models, such as Quadratic entropy (Crupi \& Tentori, 2014) or Bayes's error (Baron, Beattie \& Hershey, 1998; Crupi, Tentori \& Lombardi, 2009) could also be used. Different models do not in general lead to the same conclusions about which tests (or experiments) are most valuable (Nelson, 2005, 2008, 2009).

What kind of entropy model best characterizes people's goals in searching for information? Some data suggest that reduction in Bayes's error (probability gain) is a more plausible intuitive model than reduction in Shannon entropy
(Nelson, McKenzie, Cottrell \& Sejnowski, 2010; Meder \& Nelson, 2012). Probability gain appears to have its own limitations, however, as it does not show a preference for questions with close to a $50: 50$ split in 20-questions games (Nelson, Divjak, Gudmundsdottir, Martignon \& Meder, 2014).

Many different ideas of important axioms for entropy measures have been proposed (Csiszár, 2008). Interestingly, particular entropy measures have been predominant in particular research areas, and it is often unclear whether this is due to different theoretical and practical goals or are merely due to historical accident.
Is there any possibility for a formal model of uncertainty that would be able to describe people's behavior across a wide variety of tasks? Could such a model also have theoretically desirable properties?

Entropy is often thought of as expected surprise. But (1) what constitutes surprise, and (2) what constitutes an expectation? Depending on how surprise and expectation are defined, different entropy measures result. Combining these two ideas, we show that many entropy measures, including Hartley (1928), Shannon (1948) and Quadratic entropy, and the families of Tsallis (1988), Rényi (1961), and Arimoto (1971) entropies, can all be derived as special cases in the Sharma-Mittal (1975) framework for entropy measures.
Figure 1 depicts the Sharma-Mittal space of entropy measures graphically, where the horizontal axis (the order $r$ ) specifies the type of averaging function, and the vertical axis (the degree $t$ ) specifies the surprise function. A number of heuristic ideas of uncertainty, for instance the number of possibilities, and whether or not you know for sure (analogous to a Popperian formulation, Popper, 1959), also arise as special cases in this framework.

Can psychological insight be derived from this formalism? We show that many heretofore disparateseeming empirical results and normative desiderata can be accommodated by specific entropy measures within this formalism. Importantly, this framework affords more than a post hoc story; novel predictions can be derived for future experiments, to better characterize the psychological bases of uncertainty and information.


Figure 1: The Sharma-Mittal framework

## Acknowledgments

This research was supported by grants CR 409/1-2, NE $1713 / 1-2$, and ME 3717/2-2 from the Deutsche Forschungsgemeinschaft as part of the "New Frameworks of Rationality" priority program (SPP 1516). We thank Nick Chater, Laura Martignon, Andrea Passerini, and Paul Pedersen for helpful comments and discussions.

## References

Arimoto S. (1971). Information-theoretical considerations on estimation problems. Information and Control, 19, 181-194.
Baron, J., Beattie, J., \& Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. Organizational Behavior and Human Decision Processes, 42, 88-110. doi: 10.1016/0749-5978(88)90021-0

Crupi, V., \& Tentori, K. (2014). State of the field: Measuring information and confirmation. Studies in History and Philosophy of Science, 47(C), 81-90. doi: 10.1016/j.shpsa.2014.05.002

Crupi, V., Tentori, K., \& Lombardi, L. (2009). Pseudodiagnosticity revisited. Psychological Review, 116, 971-985. doi: 10.1037/a0017050
Csiszár, I. (2008). Axiomatic characterizations of information measures. Entropy, 10, 261-273. doi: 10.3390/e10030261

Hartley R. (1928). Transmission of information. Bell System Technical Journal, 7, 535-563.
Lindley, D. V. (1956). On a measure of the information provided by an experiment. The Annals of Mathematical Statistics, 27, 986-1005. doi: 10.1214/aoms/1177728069

Meder, B., \& Nelson, J. D. (2012). Information search with situation-specific reward functions. Judgment and Decision Making, 7, 119-148.
Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. Psychological Review, 112, 979-999. doi: 10.1037/0033295X.112.4.979
Nelson, J. D. (2008). Towards a rational theory of human information acquisition. In N. Chater and M. Oaksford (Eds.), The Probabilistic Mind: Prospects for Rational Models of Cognition (pp. 143-163). Oxford: Oxford University Press.
Nelson, J. D. (2009). Naïve optimality: Subjects’ heuristics can be better-motivated than experimenters' optimal models. Behavioral and Brain Sciences, 32, 94-95. doi: 10.1017/S0140525X09000405

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., \& Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. Cognition, 130, 74-80. doi: 10.1016/j.cognition.2013.09.007

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., \& Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. Psychological Science, 21, 960-969. doi: 10.1177/0956797610372637
Oaksford, M., \& Chater, N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review, 101, 608-631. doi: 10.1037/0033295X.101.4.608
Popper, K. R. (1959). The logic of scientific discovery. London: Hutchinson.
Rényi, A. (1961). On measures of entropy and information. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 547-561.
Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423, 623-656.
Sharma, B. \& Mittal, D. (1975). New non-additive measures of entropy for discrete probability distributions. Journal of Mathematical Sciences (Delhi), 10, 28-40.
Tsallis, C. (1988). Possible generalization of BoltzmannGibbs statistics. Journal of Statistical Physics, 52, 479487. doi: 10.1007/BF01016429

# What The Shape and Material Biases Can Tell Us About Object Recognition 

Lynn K. Perry (lkperry@miami.edu)<br>Department of Psychology, 5665 Ponce de Leon Blvd<br>Coral Gables, FL, 33146 USA

Keywords: vocabulary, novel noun generalization, word learning, object recognition, drawing

## Motivation and Background

Imagine a child who is just beginning to produce words. To figure out what a word like CUP means, she has to not only identify the individual cup, but also understand what properties are relevant for belonging to the category of cups in order to remember the word later and apply it to new cups she encounters (e.g., its abstract shape rather than its purple plastic material).

At the same time this child is learning all about cups, she is also learning other words, many of which also name categories of objects similar in shape (e.g., BALL and CAR), from which she will learn a bias to attend to shape when generalizing novel names to novel objects (i.e., shape bias). A child's prior knowledge and experiences help her to not only learn individual words and categories, but also, more importantly, to learn how to learn words, making subsequent word learning easier.

In this talk I will discuss my research on children's word learning biases and the consequences these biases have on their future learning and generalization. Additionally, I will discuss how our understanding of the development of these biases can inform our understanding of children and adults' visual recognition abilities.

## Attention to Shape

At first glance, the "shape bias" may seem like an artificial laboratory phenomenon. Children are presented with three novel objects: an exemplar, one object that matches the exemplar in shape but differs in material and color, and one object that matches the exemplar in material but differs in shape and color. When the experimenter names the exemplar and asks the child to generalize it to one of the other two saying, by the time children are about 2-years-old they systematically select the object mapping in shape (Landau, Smith, \& Jones, 1988). Although this phenomenon may seem simple, it has been shown to have consequences for children's future word learning (e.g., Perry et al., 2010) and is window into children's developing object recognition (Yee, Jones, \& Smith, 2012) and memory (Vlach, 2016).

Additionally, evidence from atypical populations further suggests that the shape bias can tell us about developmental process. For example, children with autism (Tek, Jaffery, Fein, \& Naigles, 2008), children who are late talkers (Jones \& Smith, 2005), and children who are deaf or hard of hearing and wear cochlear implants (Quittner, Cejas, Wang, Niparko, \& Barker, 2016) all show delayed or atypical biases when generalizing novel nouns.

My own work reveals interesting individual differences even within typically developing populations. Children whose vocabularies differ from the norm show generalization biases that differ from the norm (Perry \& Samuelson, 2011). The more words a child knew naming solid objects in categories organized by similarity in shape (e.g., CUP), the more likely she was to generalize novel names by similarity in shape. However, the more words she knew naming solid objects in categories organized by similarity in material (e.g., CHALK), the more likely she was to generalize novel names by similarity in material.

I have extended this line of research to examine how vocabulary differences lead to differences in memory for objects' features (Perry, Axelsson, \& Horst, 2015) and recognition of familiar objects (Perry \& Saffran, 2016). For example, regardless of vocabulary size, children who knew relatively few names for categories organized by shape had more trouble recognizing objects in the wrong colors (e.g., pink cow) than children who know more categories organized by shape. The particular words children already know, bias their future word learning and recognition.

Importantly, longitudinal training studies suggest vocabulary regularities play a causal role in shape bias development. Teaching young children categories organized by similarity in shape leads them to develop a precocious shape bias and learn new words at an increased rate (e.g., Perry et al., 2010). Together, this work on the shape bias offers insights about developmental process: 1) the process of learning words has cascading consequences for future word learning; and 2) the structure of a child's vocabulary influences what information they attend to and remember.

## Attention to Material

In addition to learning about solid objects like CUP, children also learn about nonsolid substances like applesauce and JUICE, for which material is important. When generalizing the names of novel nonsolids, older children and adults attend to similarity in material ("material bias"). Compared to the shape bias, the material bias is later acquired (Samuelson \& Smith, 1999) and is sensitive to stimuli and task changes (Samuelson \& Horst, 2007).

One reason for this difference in development is that children learn about nonsolids in a relatively constrained context-all early-learned nonsolids are foods seen at mealtimes, while solids are seen across a variety of contexts. I found that putting children in a highchair allows them to explore stimuli as they would at mealtimes and led them to show a material bias several years earlier than they do in a standard lab context (Perry, Samuelson, \& Burdinie, 2014).

An additional difference is that materials might be difficult to recognize from static visual information and may
require tactile information. Indeed, it was the children who touched stimuli the most who showed the strongest material bias in my study (Perry et al., 2014). And those in the highchair were messiest because that setting increased context-dependent action patterns that proved necessary for recognizing similarity between materials (cf Perry, 2015). Children's developing attention to material similarity builds on what we already know about attention to shape, and also provides new insight into the importance of context and exploration in recognition and generalization.

How do children eventually learn to pay attention to substances' materials outside of this specific context? Adults don't need to sit in a highchair to distinguish whiskey from juice. How do children learn to visually recognize materials? What do adults even know about substances? These questions have important applications beyond understanding word learning: although we can teach artificial intelligence systems to recognize solid objects, it is nearly impossible to teach them to recognize nonsolid substances (Adelson, 2001). To begin answering these questions, I conducted a study in which adults and children drew familiar objects and substances from memory.

## New Insights From Drawing

In my recent work, I assessed children and adults' drawings of familiar objects and substances from memory. Examining these drawings allows us to assess what visual information is relevant to representations of different kinds of things and how this information changes over development. As such, this study is an important first step in understanding how we recognize objects and substances.

Amazon Mechanical Turk participants identified drawings. Critically, they were more accurate in identifying drawings of solid objects than nonsolid substances. Both children and adults tended to include container information for nonsolids rather than draw the substance itself. Drawings of nonsolids that depicted distinct, prototypical containers (e.g., milk carton, coffee mug) aided recognition.

Additionally, adult were quite consistent in color usee.g., all adults drew brown (i.e., chocolate) pudding and purple grapes, while children used a variety of colors. These results suggest 1) color might be more important to object representations than previously believed and 2) that as children develop, they become more systematic and prototypical in the colors they associate with objects.

Overall, these new findings build on my previous work examining children's attention to shape and material by demonstrating what information we use to remember and recognize solids and nonsolids. These results demonstrate that visual recognition of both solids and nonsolids-is aided by shape, suggesting we may conceptualize nonsolids as more object-like than was thought.

## Relevant publications

My publications most relevant to this presentation are: Perry et al., 2010; Perry \& Samuelson, 2011; Perry et al., 2014; Perry, 2015; Perry et al., 2015; and Perry \& Saffran, 2016.

## References

Adelson, E. H. (2001). On seeing stuff: the perception of materials by humans and machines (4299, p1-12).
Jones, S. S., \& Smith, L. B. (2005). Object name learning and object perception: a deficit in late talkers. Journal of Child Language, 32(1), 223-240.
Landau, B., Smith, L. B., \& Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive Development, 3(3), 299-321.
Perry, L. K. (2015). To have and to hold: looking vs. touching in the study of categorization. Frontiers in Psychology, 6.
Perry, L. K., Axelsson, E. L., \& Horst, J. S. (2015). Learning What to Remember: Vocabulary Knowledge and Children's Memory for Object Names and Features. Infant and Child Development.
Perry, L. K., \& Saffran, J. R. (2016). Is a Pink Cow Still a Cow? Individual Differences in Toddlers' Vocabulary Knowledge and Lexical Representations. Cognitive Science.
Perry, L. K., \& Samuelson, L. K. (2011). The Shape of the Vocabulary Predicts the Shape of the Bias. Frontiers in Psychology, 2.
Perry, L. K., Samuelson, L. K., \& Burdinie, J. B. (2014). Highchair philosophers: the impact of seating context-dependent exploration on children's naming biases. Developmental Science.
Perry, L. K., Samuelson, L. K., Malloy, L. M., \& Schiffer, R. N. (2010). Learn Locally, Think Globally Exemplar Variability Supports Higher-Order Generalization and Word Learning. Psychological Science, 21(12), 1894-1902.
Quittner, A. L., Cejas, I., Wang, N.-Y., Niparko, J. K., \& Barker, D. H. (2016). Symbolic Play and Novel Noun Learning in Deaf and Hearing Children: Longitudinal Effects of Access to Sound on Early Precursors of Language. PLOS ONE, 11(5).
Samuelson, L. K., \& Horst, J. S. (2007). Dynamic Noun Generalization: Moment-to-Moment Interactions Shape Children's Naming Biases. Infancy, 11(1), 97-110.
Samuelson, L. K., \& Smith, L. B. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond? Cognition, 73(1), 1-33.
Tek, S., Jaffery, G., Fein, D., \& Naigles, L. R. (2008). Do children with autism spectrum disorders show a shape bias in word learning? Autism Research, 1(4), 208-222.
Vlach, H. A. (2016). How we categorize objects is related to how we remember them: The shape bias as a memory bias. Journal of Experimental Child Psychology, 152, 12-30.
Yee, M., Jones, S. S., \& Smith, L. B. (2012). Changes in Visual Object Recognition Precede the Shape Bias in Early Noun Learning. Frontiers in Psychology 3.

# The role of learning mechanisms in understanding spoken words 

Jennifer M Rodd (j.rodd@ucl.ac.uk)<br>Rebecca A Gilbert (becky.gilbert@ucl.ac.uk)<br>Hannah N Betts (hannah.betts.10@ucl.ac.uk)<br>Department of Experimental Psychology, University College London<br>26 Bedford Way, London, WC1H 0AP, UK


#### Abstract

Word meaning priming has become a key method to study how listeners (and readers) retune their lexical semantic representations in response to their linguistic environment in order to facilitate access to word meanings. We present a summary of recent findings using this method that help to constrain our theories of how this important form of lexicalsemantic learning occurs.


Keywords: lexical ambiguity; semantic ambiguity; learning; speech; language

## Background: Lexical Ambiguity

Access to word meanings during natural language comprehension is made difficult by the ubiquity of lexical ambiguity: $80 \%$ of common English words have multiple dictionary definitions (Rodd, Gaskell \& Marlsen-Wilson, 2002). Take for example the first sentence of a reading comprehension text that was recently given to 11-year-old children in England: "Dawn was casting spun-gold threads across a rosy sky over Sawubona game reserve". The words in this sentence have on average 8.8 dictionary definitions. The reader must work out that "Dawn" does not refer to a girl's name and that "game" does not refer to a form of competitive sport. And they must realise that the words "casting" and "threads" are not referring to a physical action and object, but are instead being used in a somewhat metaphorical sense.
When a listener/reader encounters an ambiguous word, they usually rapidly retrieve the most appropriate meaning and ignore any other irrelevant meaning(s). A very large body of psycholinguistics experiments conducted over the last 40 years have provided important constraints on our understanding of how this disambiguation process operates. The literature has converged on the view, exemplified in the reordered access model (Duffy, Morris, and Rayner, 1988, see Vitello and Rodd, 2014 for review) that whenever a reader/listener encounters an ambiguous word, its multiple different meanings are activated in parallel, but this activation is modulated by the sentence context and the relative frequencies of the different meanings: meanings that are highly frequent or compatible with the preceding context are more readily available.

## Word Meaning Priming: Published Findings

Recent studies using a novel word-meaning priming paradigm (Rodd et al., 2013; 2016) have supplemented this view of lexical disambiguation with evidence that learning
mechanisms make a key contribution to disambiguation fluency, by allowing listeners to make use of the past experience to boost the availability of meanings that are more likely to occur in the future. For example, comprehension of a sentence such as "the sheep were put into the pen", is usually relatively difficult because the intended 'animal-enclosure' meaning of "pen" is far less frequent than the dominant 'writing-instrument' meaning. But learning mechanisms can make such sentences easier in conditions where the listener has increased prior experience with the lower frequency meaning.

Specifically, these word-meaning priming experiments have revealed the key role of recent experience in modulating the availability of word meanings. For example, if the lower-frequency meaning of "pen" is encountered as part of a sentence comprehension prime task, then this meaning will be more readily available after a 20-40 minute delay (compared with an unprimed control; Rodd et al., 2013; 2016). This form of word meaning priming does NOT reflect a general forms of semantic priming; a control condition in which participants were primed with different but synonymous words showed no priming at this relatively long delay (Rodd et al., 2013). Word-meaning priming only occurs when the specific ambiguous word (e.g., "pen") is encountered in both the prime and test phases. In natural listening situations this dynamic 'retuning' of lexicalsemantic representations will act to improve comprehension fluency for cases where an ambiguous word is encountered multiple times within the same conversation.

In addition to these lab-based experiments that have shown word-meaning priming at 20-40 minute delays (with little decay during this time window), experiments conducted with larger sets of participants outside the lab have shown that even larger priming effects occur as a consequence of naturalistic encounters with word meanings. For example when recreational rowers encounter the specific rowing-related meanings of common words like "catch" and "feather" during their training, a significant and numerically large priming effect was observed after a median delay of eight hours (Rodd et al., 2016). In addition, these relatively large effects of same-day experience with word meanings leave residual traces that accumulate incrementally over many years to alter a listener's overall preferences for the different meanings: the number of years rowing experience that an individual rower had was a strong predictor of meaning access (Rodd et al., 2016).

Taken together, these results indicate that adult lexicalsemantic representations are relatively fluid and are
constantly being retuned on the basis of experience to improve the fluency and accuracy of comprehension.

## Word Meaning Priming: Recent Developments

A set of seven (unpublished) word-meaning priming experiments have been conducted to help constrain our theories of how exactly the availability of word meanings is boosted as a consequence of our experience.

Effects of prime/target modality. Two experiments (using different tasks at test) show that word-meaning priming occurs when the ambiguous words are presented in different modalities (i.e. spoken and written) at prime and test, and that such cross-modal priming is not reduced compared with uni-modal priming. This indicates that learning may occur at a relatively abstract lexical-semantic level, and that knowledge about words learned in one modality influences comprehension in the other modality.

Effects of word position. Two experiments (using different tasks at test) show that word-meaning priming is NOT modulated by the position of the ambiguous word within the sentence: there is no significant difference in priming when the disambiguating context occurs before or after the ambiguity (e.g., "the sheep were enclosed in a PEN" vs "a PEN was used to enclose the sheep"). These results are incompatible with an account where learning is triggered by the detection of an error signal that indicates that the ambiguous word has been misinterpreted, as this would predict increased priming for late-disambiguation sentences. The results are also incompatible with an account in which the co-activation of the word form and the contextuallyappropriate drives learning: this would predict more priming for the early-disambiguation sentences. Instead the results indicate that lexical semantic representations are modulated on the basis of a word's final, comprehended meaning and do not seem to be influenced by partial, transient activation of irrelevant meaning during comprehension.

Effects of multiple encounters. Three experiments show that listeners keep track of the likelihood of different meanings across multiple encounters with the ambiguous word. If the word is used repeatedly with the same meaning, the priming effects accumulate to increase the availability of this meaning relative to a single presentation control condition. In contrast, if different meanings are encountered then the effects of these experiences cancel each other out. Importantly, the cumulative effects of repeated exposure are dependent on the spacing of the words - no benefit of repetition is observed if the word is encountered multiple times in adjacent sentences.

## Summary

Word meaning priming has become a key method to study how listeners (and readers) retune their lexical semantic representations in response to their linguist environment.

We present recent findings using this method that help to constrain our theories of how this learning occurs and will guide the development of our connectionist model of how words are represented and processed (Rodd et al., 2004).

## Acknowledgments

This work was supported by the Economic and Social Research Council [grant numbers ES/K013351/1; ES/L010224/1; ES/J500185/1]

## Key Relevant Publications by the Author

(See jennirodd.com for full list)
Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G. et al. (2012a). Learning new meanings for old words: effects of semantic relatedness. Memory and Cognition, 40, 1095-1108.
Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., \& Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. Journal of Memory and Language, 87, 16-37.
Rodd, J. M., Davis, M. H., \& Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cerebral Cortex, 15, 12611269.

Rodd, J. M., Gaskell, M. G., \& Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. Cognitive Science, 28, 89-104.
Rodd, J. M., Johnsrude, I. S., \& Davis, M. H. (2010). The role of domain-general frontal systems in language comprehension: Evidence from dual-task interference and semantic ambiguity. Brain and Language.
Rodd, J. M., Johnsrude, I. S., \& Davis, M. H. (2012). Dissociating frontotemporal contributions to semantic ambiguity resolution in spoken sentences. Cerebral Cortex, 22, 1761-1773.
Rodd, J. M., Lopez Cutrin, B., Kirsch, H., Millar, A., \& Davis, M. H. (2013). Long-term priming of the meanings of ambiguous words. Journal of Memory and Language, 68, 180-198.
Rodd, J. M. (2004). The effect of semantic ambiguity on reading aloud: A twist in the tale. Psychonomic Bulletin and Review, 11, 440-445.
Rodd, J. M., Gaskell, M. G., \& Marslen-Wilson, W. D. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. Journal of Memory and Language, 46, 245-266.
Vitello, S. \& Rodd, J. M. (2015). Resolving Semantic Ambiguities in Sentences: Cognitive Processes and Brain Mechanisms. Linguistics and Language Compass, 9, 391405.

## Other References

Duffy, S. A., Morris, R. K., \& Rayner, K. (1988). Lexical ambiguity and fixation times in reading. Journal of Memory and Language, 27, 429-446.

# Far Transfer: Does it Exist? 

Giovanni Sala (giovanni.sala@liv.ac.uk)<br>Department of Psychological Sciences, Bedford Street South, University of Liverpool, L69 7ZA, United Kingdom

Fernand Gobet (fernand.gobet@liv.ac.uk)<br>Department of Psychological Sciences, Bedford Street South, University of Liverpool, L69 7ZA, United Kingdom


#### Abstract

Implementing interventions that are supposed to enhance students' general learning skill and overall cognitive ability is still a common practice in education. The basic idea on which this approach relies is that improving domain-general skills provides benefits for a broad range of domain-specific areas, such as academic disciplines. Thus, it is assumed that there is far transfer i.e., the generalization of a set of skills between domains loosely related to each other. In recent years, chess instruction, music instruction, and working memory training have been claimed to be able to train domain-general abilities (e.g., fluid reasoning/intelligence) which, in turn, generalize to other cognitive and academic skills (e.g., mathematics). We tested these claims in the population of healthy children via meta-analysis. The results showed small to moderate overall far-transfer effects in all the outcome measures of the three meta-analyses. However, the effect sizes were inversely related to the design quality (e.g., presence of active control groups), which casts doubts on the effectiveness of the three activities. We discuss the theoretical and practical implications of these findings for education and expertise and extend the debate to another type of training, video games training.


Keywords: chess; education; learning; mathematics; music; transfer; working memory.

## Introduction

The question of transfer is central to cognitive science. Near transfer can be defined as the generalization of a set of trained skills across domains closely related to each other. Far transfer can be defined as the generalization of a set of trained skills across domains loosely related to each other. Ever since Thorndike and Woodworth's (1901) common elements theory, psychology has documented the difficulty of far transfer. As noted by these authors, transfer from one domain to another can only happen when the two domains share common elements. Thus, while near transfer is expected to occur fairly often (e.g., transfer is expected between geometry and calculus), far transfer is much less likely, as the source and target domains share few elements (e.g., no transfer is expected between Latin and calculus).

The field of education has been much more sanguine about the possibility of far transfer. For example, in a very influential book, Papert (1980) argued that the skills acquired in learning the programming language LOGO would transfer to mathematics and indeed would improve learning generally. (Considerable research has shown that this was unlikely to be the case; e.g., Gurtner et al., 1990.) More recently, very strong claims have been made in
educational quarters, on the web and in the popular press about the possible benefits of music, chess, and working memory training for improving academic achievements and a large variety of cognitive abilities (e.g., fluid intelligence, cognitive control, phonological processing, and spatial ability).

## Recent Evidence

In recent meta-analyses (Sala \& Gobet, 2016, 2017a, 2017b), we evaluated the evidence of transfer in three domains (chess, music, and working memory training). In all cases, we focused on healthy children and young adolescents.

The results showed null to medium overall effect sizes in all three meta-analyses. Moreover, the size of the effects was inversely related to the quality of the experimental design. Specifically, when the participants were randomly allocated to the groups and the experimental groups were compared to active control groups, the overall effect sizes were minimal or null.

Design quality thus accounts for the variability between the studies. The three treatments provide either minimal overall effects on academic achievement and overall cognitive ability (music and working memory training), or medium effects possibly due to placebo effects (chess) and/or statistical artefacts due to lack of randomization. Overall, these results support Thorndike and Woodworth's (1901) theory.

## Link with Expertise Research

The lack of far transfer in these domains might appear surprising, because there is considerable evidence in these domains for correlations with intelligence and other intelligence-related measures. In a survey, Schellenberg (2006) reported medium correlations between time spent for music lessons and several measures of cognitive ability and academic attainment in a sample of children and undergraduate students. Similarly, Burgoyne et al. (2016) found that with chess players, the correlation between skill and fluid intelligence was $\bar{r}=, 24$. Sala et al. (2017) found that chess players are more intelligent than individuals who do not play chess ( $\bar{d}=0.49$ ). Finally, the correlation between working memory and intelligence is about $r=.70$ (Kane, Hambrick, \& Conway, 2005).

Of course, this evidence is correlational, and drawing conclusions about causality is notoriously difficult. Nevertheless, the pattern of results (near absence of far transfer in training studies and correlations with intelligence) suggests that, in the domains we have reviewed, intelligence comes first and explains why some individuals perform better than others.
We also note that the difficulty of far transfer is consistent with several theories of expertise. In particular, both chunking theory (Simon \& Chase, 1973) and template theory (Gobet \& Simon, 1996) propose that the development of expertise is closely linked to the acquisition of a large number of perceptual patterns. These patterns allow experts to memorise domain-specific material better than non-experts, even when the global structure of the material is destroyed by randomization (Sala \& Gobet, 2017c). Accessing knowledge depends on productions matching these perceptual patterns. Thus, even if two domains share common elements at an abstract level, these two theories predict that far transfer is unlikely, as the perceptual patterns met in the two domains are different (Gobet, 2015, 2016; Gobet \& Campitelli, 2006).

We are currently carrying out a meta-analysis of the effects of video game playing to further test these hypotheses. We strongly predict that near transfer will be common, but that far transfer will be rare. The results of this meta-analysis will be ready for the conference.

## Implications for Education

The studies discussed above suggest that the widespread notion that practicing any cognitively demanding activity enhances one or more cognitive skills beyond the trained activity has little empirical support. In other words, the benefits of cognitive training seem to be, to a large extent, domain- and task-specific. Practically, the unlikely occurrence of far transfer suggests that the most effective way of improving a skill is to train that particular skill. This sobering conclusion should discourage educators and trainers from proposing curricula aimed at fostering domaingeneral skills. Rather, curricula and training programs with a considerable amount of domain-specific content should be preferred. Thus, if the aim is to teach mathematics, lessons focusing on mathematics are better than lessons containing material on music, chess or working memory training.

## References

Burgoyne, A. P., Sala, G., Gobet, F., Macnamara, B. N., Campitelli, G., \& Hambrick, D. Z. (2016). The relationship between cognitive ability and chess skill: A comprehensive meta-analysis. Intelligence, 59, 72-83.
Gobet, F. (2015). Cognitive aspects of learning in formal and non-formal contexts: Lessons from expertise research. British Journal of Educational Psychology, Monograph Series II: Number 11, Learning beyond the Classroom, 23-37.
Gobet, F. (2016). Understanding expertise: A multidisciplinary approach. London: Palgrave/Macmillan.

Gobet, F., \& Campitelli, G. (2006). Educational benefits of chess instruction. A critical review. In T. Redman (Ed.), Chess and education. Selected essays from the Koltanowski Conference (pp. 124-143). Dallas, TX: University of Texas at Dallas.
Gobet, F., \& Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. Cognitive Psychology, 31, 1-40.
Gurtner, J. L., Gex, C., Gobet, F., Núñez, R., \& Retschitzki, J. (1990). La récursivité rend-elle l'intelligence artificielle? Revue Suisse de Psychologie, 49, 17-26.
Kane, M. J., Hambrick, D. Z., \& Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). Psychological Bulletin, 131, 6671.

Papert, S. (1980). Mindstorms: Children, computers, and powerful ideas. New York: Basic Books.
Sala, G., Burgoyne, A. P., Macnamara, B. N., Hambrick, D. Z., Campitelli, G., \& Gobet, F. (2017). Checking the "academic selection" argument. Chess players outperform non-chess players in cognitive skills related to intelligence: A meta-analysis. Intelligence, 61, 130-139.
Sala, G., \& Gobet, F. (2016). Do the benefits of chess instruction transfer to academic and cognitive skills? A meta-analysis. Educational Research Review, 18, 46-57.
Sala, G., \& Gobet, F. (2017a). When the music's over. Does music skill transfer to children's and young adolescents' cognitive and academic skills? A meta-analysis. Educational Research Review, 20, 55-67.
Sala, G., \& Gobet, F. (2017b). Working memory training in typically developing children: A meta-analysis of the available evidence. Developmental Psychology, 53, 671685.

Sala, G., \& Gobet, F. (2017c). Experts' memory superiority for domain-specific random material generalizes across fields of expertise: A meta-analysis. Memory \& Cognition, 45, 183-193.
Sala, G., Gobet, F., Trinchero, R., \& Ventura, S. (2016). Does chess instruction enhance mathematical ability in children? A three-group design to control for placebo effects. Proceedings of the $38^{\text {th }}$ Annual Meeting of the Cognitive Science Society.
Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. Journal of Educational Psychology, 98, 457-468.
Simon, H. A., \& Chase, W. G. (1973). Skill in chess. American Scientist, 61, 393-403.
Thorndike, E. L., \& Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions (I). Psychological Review, 8, 247-261.

# Scientific Sensemaking: A Critical Resource for Science Learning in School 

Christian D. Schunn (schunn@pitt.edu)<br>814 LRDC, University of Pittsburgh, 3939 O'Hara St<br>Pittsburgh, PA 15260 USA

Keywords: Science Learning; cognitive resources

## Conceptualizing Integration \& Transfer

Science consists of both a body of knowledge and a process by which the knowledge is produced. Historically, these two aspects were often assessed separately (i.e., test items on knowledge and test items on skills) and taught relatively separately (e.g., with an introduction section on skills or via isolated projects or labs). The last decade has been marked by a substantial shift to an integrated view of both how science should be taught and how science learning should be assessed. Now, consensus reports (e.g., NRC, 2007, 2011) assert that scientific processes (renamed practices) should be used to learn science content (e.g., by designing, conducted, and interpreting experiments, or by arguing from existing sources). Further, new science standards (e.g., NGSS) strongly claim that science practices must be demonstrated in use with scientific content and that scientific content must be demonstrated through use with scientific practices.

While the central point about the importance of practice and content integration is well supported by existing data (for a summary, see NRC 2011), embedded within these new conceptions of teaching and learning science are some open cognitive foundations questions that bear further investigation. These questions have important implications for both assessment and instruction. The first open question is the about the generativity and transferability of practices across content. If students learn sciences practices in one science content area (e.g., in biology), are they able to apply those science practices in another domain (e.g., chemistry)? Expert scientists have some transferability of their skills (Schunn \& Anderson, 1998), but will students also show such transferability? If so, they will be better positioned to learn new content having mastered practices in a prior science content area. However, if practices are very tightly bound to science content given how they are taught and learned, students may struggle with using these practices in new content areas.

The second but related open question has to do with the coherence of practices. If science consists of independent practices, is it meaningful to report an overall mastery level of science practices? However, if science practices work together in overall cycles of inquiry, then students who master some practices will be better positioned to master other practices, and there will be meaningful overall mastery level of science practices which can be taught and assessed.

Taking on both of these open questions, here I present recent tests of the general hypothesis that there is a general overall mastery level of core scientific practices that drives
learning in new science content areas. If supported, instruction should be organized around developing these core practices early in instruction (to accelerate later learning). Also, support for this approach suggests that computational agents could be developed to systematically acquire science content through experimentation and reading using scientific sensemaking skills as a foundation.

Conceptualizing this general mastery level as scientific sensemaking, and efficiently measuring it using scenarios that invoke shared, intuitive understandings of the natural world, I will describe recently obtained evidence that 1 ) students tend to vary along coherently along this overall sensemaking dimension; 2) overall sensemaking levels are a strong predictor of science learning; and 5) this overall sensemaking dimension can improve with effective science instruction.

## Conceptualizing Scientific Sensemaking

Approaching learning of science-related content as a sensemaking activity means recognizing that science is not a series of facts, but rather an ongoing and iterative employment of a set of practices that used in the pursuit of an increasingly rich understanding of natural and physical phenomena. These practices include asking good questions, seeking mechanistic explanations for natural and physical phenomena, engaging in argumentation about scientific ideas, interpreting data tables, designing investigations, and understanding the changing nature of science (Apedoe \& Ford, 2009; Lehrer, Schauble, \& Petrosino, 2001). Each of these practices play an important and complementary role in science learning. In selecting practices for inclusion within the scientific sensemaking construct, several criteria needed to be met: 1) an existing research-base for its role in predicting science achievement; 2) uniqueness in its contribution to learning; and 3) the flexibility to be improved through targeted instruction.

## Measuring Scientific Sensemaking

To cleanly measure scientific sensemaking, it is important consider several critical issues. First, the measure had to be about some scientific content: one cannot engage in scientific sensemaking void of content-science has some logic to its processes, but most of the logic is context specific about which assumptions or inferences are merited. Consider the class Control-of-Variables strategy (Zimmerman, 2007). This strategy can only be applied when it is clear what variables are possible to vary (and plausibly causal), which involves thinking about content.

A second consideration is that the assessment be effortworthy. Engaging in scientific sensemaking requires effort, and there is often little incentive for students to put forth effort to perform on an assessment. If students are not putting effort into the assessment, then the scores obtained from the assessment are an underestimate of the abilities students have. In order to motivate students to put forth effort, the content we selected for the assessment scenarios were so-called "charismatic mega-fauna" (i.e., Dolphins, Monkeys, \& Eagles), in which a general interest in the topic motivates some basic level of effort (Bathgate, et al. 2013).

A third consideration is assessment length. Items that are cognitively demanding of students and require them to make sense of scientific information, take relatively longer amounts of time than items simply requiring content recall.

## Empirical Tests of Scientific Sensemaking

After consideration of design considerations listed above, a new measure was created, and its psychometric properties were verified. Then, the validity of scientific sensemaking as a predictor of future content learning was tested in a large-scale study of students learning diverse science content across middle school and early high school grades in diverse curricula (included more hands-on and more text-book-based). Finally, changes in scientific sensemaking were examined in relation to levels of student engagement in classroom learning, to show that is malleable, rather than a fix construct like IQ.

## Acknowledgments

This work was funded by a grant DRL-1348468 from the National Science Foundation.

## Recent Prior Publications in this Area

Bathgate, M.E., Crowell, A.J., Cannady, M., Dorph, R. \& Schunn, C.D. (2015). The learning benefits of being willing and able to engage in scientific argumentation. International Journal of Science Education, 37(10), 15901612.

Crowell, A. J., \& Schunn, C. D. (2014). The contextspecificity of scientifically literate action: Key barriers and facilitators across contexts and contents. Public Understanding of Science, 23(6), 718-733.
Crowell, A. J., \& Schunn, C. D. (2016). Unpacking the relationship between science education and applied scientific literacy. Research in Science Education, 46(1), 129-140.
Dorph, R., Cannady, M., \& Schunn, C. D. (2016). How science learning activation enables success for youth in science learning. Electronic Journal of Science Education, 20(8).
Lin, P.-Y., \& Schunn, C. D. (2016). The dimensions and impact of informal science learning experiences on middle schoolers' attitudes and abilities in science. International Journal of Science Education, 38(17), 25512572.

Paletz, S. B. F., Kim, K., Schunn, C. D., Tollinger, I., \& Vera, A. (2013). The development of adaptive expertise, routine expertise, and novelty in a large research team. Applied Cognitive Psychology, 27(4), 415-428.
Peffer, M. E., Beckler, M. L., Schunn, C. D., Renken, M., \& Revak, A. (2015). Science classroom inquiry (SCI) simulations: A novel method to scaffold science learning. PLOS ONE, 10(3), e0120638.
Schuchardt, A., \& Schunn, C. D. (2016). Modeling scientific processes with mathematics equations enhances student qualitative conceptual understanding and quantitative problem solving. Science Education, 100(2), 290-320.
Schunn, C. D. (In press). Building from in vivo research to the future of research on relational thinking and learning. Educational Psychology Review.
Schunn, C. D., \& Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. Cognitive Science, 23(3), 337-370.
Sha, L., Schunn, C. D. \& Bathgate, M. (2015). Measuring choice to participate in optional science learning experiences during early adolescence. Journal of Research in Science Teaching, 52(5), 686-709.
Silk, E. M., Schunn, C. D., \& Strand-Cary, M. (2009). The impact of an engineering design curriculum on science reasoning in an urban setting. Journal of Science Education and Technology, 18(3), 209-223.
Tekkumru-Kisa, M., Stein, M. K., \& Schunn, C. D. (2015). A framework for analyzing cognitive demand and content-practices integration: Task analysis guide in science. Journal of Research in Science Teaching, 52(5), 659-685.
Vincent Ruz, P., \& Schunn, C. D. (2017). The increasingly important role of science competency beliefs for science learning in girls. Journal of Research in Science Teaching.

## Other Cited References

Apedoe, X., \& Ford, M. (2009). The Empirical Attitude, Material Practice and Design Activities. Science \& Education, 19(2), 165-186.
Lehrer, R., Schauble, L., \& Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. In K. Crowley, C. Schunn, \& T. Okada (Eds.), Designing for Science: Implications from Everyday, Classroom, and Professional Settings, p. 251-278.
National Research Council. (2007). Taking science to school: learning and teaching science in grades K-8. Washington, DC: The National Academies Press.
National Research Council. (2011). A framework for $K-12$ science education; practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. Developmental Review, 27(2), 172-223.

# Unsupervised Learning in an Animal Model 

Edward A. Wasserman (ed-wasserman@uiowa.edu)<br>Leyre Castro (leyre-castroruiz@uiowa.edu)<br>Department of Psychological and Brain Sciences, The University of Iowa<br>301 East Jefferson Street, Iowa City, IA 52242, USA<br>Vladimir M. Sloutsky (sloutsky.1@osu.edu)<br>Department of Psychology, The Ohio State University<br>1835 Neil Avenue, Columbus, OH 43210, USA

Keywords: unsupervised learning; supervised learning; associative learning; animal model; pigeons

## Introduction

The learning of rich associative networks is foundational to language and other advanced cognitive competencies. Such learning is commonly encouraged by direct supervision entailing explicit feedback for correct and incorrect responses. The power of explicit feedback has recently been demonstrated in an animal model (Wasserman, Brooks, \& McMurray, 2015), in which pigeons successfully learned to categorize 128 stimuli into 16 human language categories via both associative strengthening and weakening processes (Roembke, Wasserman, \& McMurray, 2016).

The use of supervised training strongly implies "cajoling" an organism to respond correctly. Most supervised learning tasks for animals differentially reinforce behavior, so that reward is given when the animal's response is correct, but not when the animal's response is incorrect. Arranging unsupervised associative learning tasks in an animal model is decidedly more difficult, as doing so requires that experimenters nondifferentially reinforce behavior.

Nonetheless, unsupervised learning may materially participate in the acquisition of rich associative networks. Given that supervision can dramatically change how stimuli are learned and represented (Love, 2002), a key and as yet unmet challenge is to devise behavioral tasks that demonstrate associative learning in animals and that do not involve explicit supervision. Extending work in infant behavior (Sloutsky \& Robinson, 2013), we devised and deployed a promising new paradigm to assess unsupervised learning in a pigeon model.

## Experiment

Two groups of four pigeons each were shown eight object images and eight color patterns on a touch-sensitive computer screen. In the Consistent Pairings group, each object was paired with a particular pattern (e.g., Object A was always paired with Pattern 1, etc.), so that these birds could learn eight specific object-pattern pairs. In the Random Pairings group, each of the eight objects was presented an equal number of times with each of the eight color patterns, so that these birds could not learn any consistent object-pattern pairs. All birds were first trained
without supervision and later trained with supervision, yielding two different ways to assess associative learning.

## Unsupervised Phase

Daily sessions comprised 128 trials, in which each object image was shown 16 times: always followed by the same color pattern in the Consistent Pairings group or randomly followed by each of the color patterns in the Random Pairings group. Birds simply had to peck each of the images a fixed number of times (gradually increased to 10). After completing that requirement-first to the object and next to the color pattern-food was always given. There were no correct or incorrect responses; so, no differential feedback was ever provided. Under these unsupervised training conditions, were the Consistent Pairings birds learning the statistical relations between each object and color pattern?

To find out, after 5 sessions of training, we gave 1 testing session in which-in addition to the 128 training trials-we included 28 testing trials. On testing trials, after pecking the object, two color patterns (rather than one) appeared to the left and right of the object. In the Consistent Pairings group, one color pattern was the same one that had consistently followed the object in training sessions, whereas the other was one of the remaining color patterns, randomly chosen. In the Random Pairings group, the two color patterns were randomly chosen. We repeated this sequence of 5 sessions of training followed by 1 session of testing 8 times. Food always followed any choice response the bird made.


Figure 1: Mean percent correct for Consistent and Random Pairings groups on test trials during the unsupervised phase.

Results on testing trials are shown in Figure 1. The percent of correct responses reflects the tendency of birds in the Consistent Pairings group to choose the color pattern
that had been paired with the object. There was no correct choice possible for the birds in the Random Pairings group, so their choice was bound to lie near the $50 \%$ chance level ( $M=52 \%$ ). Surprisingly, birds in the Consistent Pairings group actually preferred to peck the other color pattern, not the one that had been paired with the object; their choice performance was significantly below chance ( $M=35 \%$ ).

Thus, pigeons in the Consistent Pairings group did learn which specific color pattern was associated with each specific object; however, in choice tests, they displayed a preference for the other color pattern. This type of choice appears to parallel the preference for novelty in the classical children's preferential looking paradigm. Critically, all of the stimuli in our experiment were equally novel in both the Consistent Pairings group and in the Random Pairings group. Thus, the difference in performance here clearly implicates associative learning.

## Supervised Phase

After eight cycles of unsupervised training and choice testing, we began the supervised training phase, which was the same for both groups of pigeons. Now, all trials presented one object followed by two color patterns. Choice of the correct color pattern was followed by food, whereas choice of the incorrect color pattern was not. Figure 2 shows the number of training sessions necessary to reach $65 \%$, $75 \%$, and $85 \%$ accuracy levels.


Figure 2: Mean number of days for Consistent and Random Pairings groups to reach $65 \%, 75 \%$, and $85 \%$ accuracy levels in the supervised phase.

Mean accuracy on the first session of supervised training was $38 \%$ for the Consistent Pairings group and $51 \%$ for the Random pairings group, agreeing with their earlier behavior. Yet, despite this initial disadvantage, the Consistent Pairings group was actually faster to reach the $65 \%, 75 \%$, and $85 \%$ correct levels than the Random Pairings group. Again, the Consistent Pairings group showed clear evidence of having learned the object-pattern pairs during the unsupervised training phase, here by learning faster than the Random Pairings group during the supervised learning phase.

## Conclusions

These results clearly demonstrate unsupervised associative learning in pigeons using the same general stimuli and response options used in our earlier work in pigeons'
supervised category learning. One must appreciate, however, that supervised learning tasks are not altogether free from the influence of unsupervised learning. When correct responses are made and positive feedback is given, a statistical regularity is enforced which can strengthen the stimulus-response bond on correct-choice trials. What most strikingly distinguishes supervised from unsupervised learning is therefore likely to be what happens on incorrectchoice trials; here, weakening or pruning of stimulusresponse bonds can further direct responses to the correct choice option. Just such independent evidence was provided in the research of Roembke et al. (2016).

What remains to be determined is why-during the supervised phase-pigeons in the Consistent Pairings group were reluctant to peck the color pattern that had earlier been paired with the object image. Further work is underway to make that determination.

## Acknowledgments

This research was supported by the Eunice Kennedy Shriver National Institute of Child Health \& Human Development of the NIH under Award Number P01HD080679.

## Cited and Extended References

Couto, K. C., Navarro, V. M., Smith, T. R., \& Wasserman, E. A. (2016). Concept learning without differential reinforcement in pigeons by means of contextual cueing. Journal of Experimental Psychology: Animal Learning and Cognition, 42, 221-227.
Love, B. (2002). Comparing supervised and unsupervised category learning. Psychonomic Bulletin \& Review, 9, 829-835.
Roembke, T. C., Wasserman, E. A., \& McMurray, B. (2016). Learning in rich networks involves both positive and negative associations. Journal of Experimental Psychology: General, 145, 1062-1074.
Sloutsky, V. M. \& Robinson, C. W. (2013). Redundancy matters: Flexible learning of multiple contingencies in infants. Cognition, 126, 156-164.
Wasserman, E. A. (2016). Conceptualization in pigeons: The evolution of a paradigm. Behavioural Processes, 123, 4-14.
Wasserman, E. A., Brooks, D. I., \& McMurray, B. (2015). Pigeons acquire multiple categories in parallel via associative learning: A parallel to human word learning? Cognition, 136, 99-122.
Wasserman, E. A., Castro, L., \& Freeman, J. H. (2012). Same-different categorization in rats. Learning \& Memory, 19, 142-145.
Wasserman, E. A., \& Young, M. E. (2010). Same-different discrimination: The keel and backbone of thought and reasoning. Journal of Experimental Psychology: Animal Behavior Processes, 36, 3-22.
Zentall, T. R., \& Wasserman, E. A. (2012). Oxford handbook of comparative cognition. New York: Oxford University Press.

# The effects of gesture restriction on spatial language in young and elderly adults 

Demet Özer (dozer@ku.edu.tr)<br>Merve Tansan (mtansan@ku.edu.tr)<br>Ege Ekin Özer (ekinozer@ku.edu.tr)<br>Department of Psychology, Koç University<br>Rumelifeneri Yolu Sariyer 34450, Istanbul - TURKEY

Katsiaryna Malykhina (kat.malykhina@gmail.com)<br>Anjan Chatterjee (anjan@mail.med.upenn.edu)<br>Department of Neurology, Pennsylvania Hospital, 330 South $9^{\text {th }}$ Street<br>Philadelphia, PA 19107 USA

Tilbe Göksun (tgoksun@ku.edu.tr)<br>Department of Psychology, Koç University<br>Rumelifeneri Yolu Sariyer 34450, Istanbul - TURKEY


#### Abstract

There is contradictory evidence on whether speech production gets impaired or enhanced when people are restrained from gesturing. There is also very little research on how this effect can change with aging. The present study sought evidence for these by asking young and elderly adults to describe two different routes on a map in spontaneous speech and when gestures were prohibited. We found that elderly adults produced more spatial language when they were restricted to use gestures compared to their spontaneous speech, whereas young adults produced comparable levels of spatial language in both conditions. Young and elderly adults used comparable levels of gestures in their spontaneous route descriptions. Yet, only young adults' gesture use correlated positively with their spatial language production. Thus, the results of gesture prohibition on speech production are different for young and elderly adults.


Keywords: gesture restriction; speech production; aging; spatial language

## Introduction

People produce spontaneous gestures as they speak. Gestures improve communication as listeners understand a spoken message better when it is accompanied by a visible gesture (Hostetter, 2011). This enhanced communication might arise because gestures provide an image that is particularly informative about the spatial or motor aspects of the message that are not easily encoded in speech. That is, gestures, especially the ones accompanying spatial or motor information, improve communication (Alibali, 2005). Apart from the effects on the listener, gestures also benefit speakers. Gestures can directly convey imagistic components of thought due to the isomorphism between spatial-motor images and representational gestures. This
might be helpful for speakers describing spatial-motor events (Church \& Goldin-Meadow, 1986). The current study investigates the role of gesturing on the use of spatial language in a spatial task in young and elderly adults. We examined the relation between gesture and speech to address two questions: (1) How does people's use of spatial language change when they are restricted from using gestures in a spatial task? (2) How does aging influence the link between gesture production and spatial language use?

## How does gesture restriction affect speech production?

There are multiple accounts suggesting that gesturing benefits speakers as they speak (e.g., Kita, 2000; Krauss, Chen, \& Gottesman, 2000; Melinger \& Levelt, 2004). However, these accounts differ in the proposed mechanism for this benefit in speech. The Information Packaging Hypothesis (Kita, 2000) states that gestures help speakers to organize and package visual-spatial information into the linear and segmented units of language. Forming an image of the referent by gesturing might induce attention on the specific properties of that image, thus, helping speakers to break their rich spatial representations into units that are codable in speech. According to this model, when people are restricted from gesturing while talking about spatial information, people have difficulty in organizing their rich spatial-motor ideas into the units of language. As a result, they produce less spatial information in their speech compared to the cases in which they can spontaneously produce gestures. Indeed, Rimé and colleagues (1984) found that when people are restricted from gesturing, their speech contained less vivid descriptions. In line with the Information Packaging Hypothesis, when producing motor descriptions (e.g. how to tie a shoe), people who were free
to produce gestures used more semantically rich verbs referring to key elements in motoric descriptions (Hostetter, Alibali, \& Kita, 2007).

Another account, the Lexical Access Hypothesis, suggests that gestures might facilitate speech to retrieve words especially when expressing spatial information (Hadar \& Butterworth, 1997; Krauss, Chen, \& Gottesman, 2000). One prediction of this account is that if speakers are restricted from gesturing, they will have problems in retrieving the correct words particularly when they describe spatial-motor events. Thus, in line with Information Packaging Hypothesis, this account argues that speech will be impaired in the absence of gesture use. Many studies support this account (Krauss, 1998; Rauscher, Krauss, \& Chen, 1996). For example, Hostetter, Alibali, and Kita (2007) found that speakers who were restricted from gesturing started their speech with more connectors such as "and" compared to speakers who were free to use gestures when describing motoric events. The disfluencies observed in the absence of gesture is more evident in spatial contents. When restricted from gesturing, speakers were found to be speaking slower when talking about spatial information compared to the nonspatial aspects of a cartoon (Krauss, 1998). Also, speakers produced higher proportion of filled pauses (e.g. um, uh) when they could not gesture, indicating that they had difficulty in accessing the lexical items.

Another account suggests that speakers use gestures to supplement and/or complement their speech (Melinger \& Levelt, 2004). It is easier to convey the imagistic properties of spatial-motor events with gestures in a more global manner compared to speech (see also Kita, 2000). Yet, according to this account, speakers use speech and gesture concurrently; but if a gesture expresses necessary information, then, that information can be omitted in speech. For instance, Melinger and Levelt (2004) found that speakers who produced iconic gestures representing the spatial relations omitted the required spatial information from their speech more compared to speakers who did not gesture. Also, when restricted from gesturing, speakers used more spatial language when describing the spatial relations between objects (Graham \& Heywood, 1975).

Although all three accounts state that gesturing not only benefit communication of the listener but also of the speaker, they diverge on the mechanism of this benefit. They also propose different predictions on how speech will be affected in the absence of gesture use. The Information Packaging Hypothesis and the Lexical Access Hypothesis predict that speech will be impaired when speakers are restricted from gesturing, whereas the third account by Melinger and Levelt (2004) suggests that speech and gesture work as two different channels of expression, mutually
compensating each other. Thus, speech can be enhanced in the absence of gesture use.

## How does aging affect gesture production?

Although the effect of aging on communication has been studied considerably in the literature, the impact of aging on gesture use has received less attention. Studies investigating the effects of aging on gesture show that production (Cohen \& Borsoi, 1996), imitation (Dimeck, Roy, \& Hall, 1998), and comprehension (e.g., Cocks, Morgan, \& Kita, 2000) of gestures are all impaired by aging.

Aging can either affect cognition globally or increase problems in specific components of the cognitive system, such as working memory and spatial integration in visual processing (Andersen \& Ni, 2008; Copeland \& Radvansky, 2007). People with poor visual-spatial working memory and spatial transformation ability used more representational gestures (Chu, Foulkes, Meyer, \& Kita, 2014). Thus, it is possible that elderly people use fewer representational gestures in spontaneous speech due to problems in their working memory system. Cohen and Borsoi (1996) found that in an object description task, elderly adults produced fewer descriptive (i.e., representational) gestures compared to young adults; however, the use of non-descriptive (i.e., beat) gestures did not differ between young and elderly adults. These findings were interpreted as a specific consequence of reduced use of visual imagery in elderly people. No difference in the description quality of speech was found as a function of age. This study, in contrary to the Information Packaging Hypothesis (Kita, 2000), suggests that the less frequent use of gestures did not impair speech in elderly. Feyereisen and Harvard (1999) also investigated the production of representational and beat gestures in different description contents varying the likelihood of generating mental images. Elderly adults used fewer representational gestures when they talked about visuospatial content generating more mental imagery, whereas the use of beats for low mental imagery events was comparable between young and elderly.

The findings on the gesture use of elderly adults are not conclusive. Although elderly adults less frequently use representational gestures in spontaneous speech in spatial contents, the effects of gesture restriction on their speech is unknown. To our knowledge, no study has investigated how the effects of gesture restriction on speech, particularly spatial information, differ between young and elderly adults.

## The Present Study

The purpose of the present study is to further understand the effects of gesture restriction on speech production as a function of aging. We asked young and elderly adults to
describe two different routes on a directional map task without mentioning gesture use (i.e., spontaneous gesture production) and by prohibiting them from gesturing.

First, in line with the previous studies (e.g., Feyereisen \& Harvard, 1999; Cohen \& Borsoi, 1996), we expect that elderly adults would use fewer gestures compared to young adults when spontaneously describing the routes on the map. If gestures facilitate speech production by either helping speakers to organize and package visual-spatial information into units of speech (as suggested by the Information Packaging Hypothesis) or helping them to retrieve words (as suggested by the Lexical Access Hypothesis), then we would expect that young and elderly adults should use more spatial language in the gesture unrestricted condition compared to the gesture restricted condition. However, when we consider the already sparse use of gestures in elderly, we might also observe no difference between spatial language use in spontaneous speech and gesture restricted conditions in elderly adults. If, on the other hand, gesture helps speakers by easily conveying information that is not necessarily available in speech (Melinger \& Levelt, 2014), then, young adults should use more spatial language when restricted from gesturing compared to spontaneous speech. However, if elderly adults use fewer gestures compared to young adults, then elderly adults would use more spatial language in spontaneous speech compared to young ones and use comparable levels of spatial language across spontaneous speech and gesture restriction conditions.

## Method

## Participants

Twenty young ( $M_{\text {age }}=20.3, S D=2.18$, range: $18-22,9$ females) and 19 elderly ( $M_{\text {age }}=62, S D=7.17$, range: 52-78, 15 females) adults agreed to participate in the experiment in exchange for course credit or $\$ 10$ for an hour. All participants were right-handed and native English-speakers. Before the sessions, they were provided with a written, informed consent in accordance with the policies of the University of Pennsylvania's Institutional Review Board.

## Stimuli and Procedure

Participants were asked to describe two different routes (Route 1 vs. Route 2) on a San Diego Zoo Map in two different counterbalanced conditions in a quiet room with an individual setting (see Figure 1). All participants were seated on an armless chair to promote gesturing. The map was printed on an A1 size cartoon ( $594 \times 841 \mathrm{~mm}$ ) so that the routes were visible to the participants. The large and unnecessary identification signs were erased and targets were circled in pink to make finding them easier on the
map. Two routes (Route 1 vs. Route 2) were created. Route 1 was from a landmark at the bottom left to another landmark at the top right. Route 2 was the other diagonal route; from a landmark at the top right to another landmark at the bottom left. The map was present throughout the session and the experimenter held the map for the participant to describe the routes.


Figure 1: San Diego Zoo Map
In the first condition (spontaneous gesture condition; SG), participants were asked to describe the path they would take to go from one pink-circled landmark to another. In this condition, they were given no specific information about the use of speech or gesture. In the second condition (gesture restricted condition; GR), however, they were asked to sit on their hands and explain how they would continue from a different marked landmark to another just with speech. Participants always completed the SG condition first and GR condition second not to make them aware of the gesture use in the SG condition. However, the order of routes (Route 1 vs. Route 2) on the map was counterbalanced across different conditions. Sessions were videotaped for further coding.

## Coding

Speech. Participants' speech was transcribed verbatim in both conditions by a native English speaker. For each condition, we first calculated the number of utterances coded as the units of speech bounded by silence. Next, we coded three different spatial information in speech: (1) direction describing the course of movement in relation to other objects (e.g., down, south), (2) street names (e.g., Hippo trail or Parkway), and (3) landmarks (e.g., restaurant). Participants' frequency of using each spatial information in speech was calculated. We also calculated a composite speech score that was the sum of all spatial information in speech.

Gesture. Participants' use of spontaneous co-speech gestures was coded in SG condition. A change in the path of hand movement determined a new gesture. We coded three different gestures: (1) pointing (e.g., pointing at a landmark on the map), (2) tracing (e.g., continuously moving the finger or hand on the map to show the route), and (3) iconic (e.g., moving the hand off the map to represent a direction).

We also calculated a composite gesture score that was the total of all 3 specific gesture types. Both speech and gesture were coded by the first author.

## Results

Young and elderly adults did not differ in the total number of utterances in $\mathrm{SG}(F(1,37)=.77, p>.05)$ and $\mathrm{GR}(F(1$, 37) $=.04, p>.05)$ conditions, and the total trial duration it took to describe the routes in $\operatorname{SG}(F(1,37)=.30, p>.05)$ and GR $(F(1,37)=2.99, p>.05)$ conditions. There was also no gender difference in the total number of utterances in $\operatorname{SG}(F(1,37)=.57, p>.05)$ and $\operatorname{GR}(F(1,37)=1.13, p$ $>.05)$ and, the total trial duration in $\mathrm{SG}(F(1,37)=.19, p>$ $.05)$ and $\operatorname{GR}(F(1,37)=.31, p>.05)$ conditions. Thus, we merged gender for further analyses.

To see if the order of the routes (Route 1 first vs. Route 2 first) influenced the number of utterances used and the duration of describing the routes, we conducted 2 separate mixed ANOVA with total number of utterances (SG vs. GR conditions) and total trial duration (SG vs. GR conditions) as within subject variables and the order of routes (R1 first vs. R2 first) as the between subject variable. There was an interaction between the number of utterances and the order of the routes, and between the trial duration and the order of the routes, $F(1,37)=7.59, p<.01$ and $F(1,37)=6.69, p<$ .02 , respectively. Participants produced more utterances and spent more time in GR ( $M=10.25$ and $M=78.35$ seconds, respectively) compared to $\mathrm{SG}(M=6.35$ and $M=39.80$ seconds, respectively) only when they completed the second route (R2) in GR condition. Thus, for further analyses, we used normalized scores, obtained by the total number of utterances (i.e. raw scores) divided by the number of utterances for each subject in the respective condition. For the next analyses, we used Bonferroni adjusted alpha levels in pairwise comparisons for multiple hypotheses testing and applied Greenhouse-Geisser correction when sphericity assumption was violated (see Tables 1 and 2 for the mean raw scores)

## Speech Analyses

First, we conducted a mixed ANOVA with composite speech score ( SG vs. GR conditions) as within subject variable and group (young vs. elderly) as the between subject variable to see if total spatial information used in speech from SG to GR conditions differed between young and elderly. There was a main effect of the condition on the composite speech scores, $F(1,37)=27.68, p<.001$. However, this main effect was qualified by interaction, $F$ (1, 37) $=7.07, p<.05$. Young participants $(M=2.18)$ used more spatial information compared to elderly ( $M=1.42$ ) in SG condition. However, no difference was found for the use
of spatial information between young ( $M=2.58$ ) and elderly ( $M=2.62$ ) participants in GR condition. In addition, elderly adults used more spatial information in GR compared to SG, whereas young participants produced comparable spatial information in speech in SG and GR conditions (see also Table 1 for raw scores). Thus, even though young adults used similar spatial information in both conditions, when gesture use was restricted, elderly adults' use of spatial information increased.

To see if specific spatial information used in speech (street name, landmark or direction) in SG to GR conditions differed between groups, we conducted a $2 \times 2 \times 3$ mixed ANOVA with group (young and elderly adults) as the between subject variable and condition (SG and GR) and specific spatial information (street name, landmark, and direction) as within subject variables. There was a main effect of condition, $F(1,37)=27.68, p<.001$ and an interaction of condition by group, $F(1,37)=7.07, p<.02$. As reported earlier, elderly, on average, used more spatial information in GR compared to SG conditions, whereas young produced comparable levels of spatial information in SG and GR conditions. There was also a main effect of the specific type of spatial information, $F(2,74)=16.35, p<$ .001. Regardless of the condition, all participants used more direction information $(M=.99)$ compared to street name ( $M$ $=.60)$ and landmark $(M=.61)$ information in speech. No other interactions among spatial information, group, or condition were found, $p s>.05$.

Table 1: Mean raw speech scores for each condition (SG and GR) and group. The values in parentheses are standard errors of mean.

|  | Young |  | Elderly |  |
| :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{S G}$ | $\boldsymbol{G R}$ | $\boldsymbol{S G}$ | $\boldsymbol{G R}$ |
| Number of Utterances | 9.6 | 9 | 7.3 | 9.5 |
|  | $(2.2)$ | $(1.5)$ | $(1.4)$ | $(1.7)$ |
| Composite Speech | 18.0 | 21.1 | 11.6 | 21.4 |
| Score | $(3.5)$ | $(2.6)$ | $(2.9)$ | $(2.6)$ |
|  |  |  |  |  |
| Direction in speech | 8.3 | 9.3 | 5.6 | 9.3 |
|  | $(1.9)$ | $(1.5)$ | $(1.2)$ | $(1.4)$ |
| Landmark in speech | 5 | 6.5 | 3.5 | 7.0 |
|  | $(1.3)$ | $(1.0)$ | $(1.1)$ | $(1.2)$ |
| Street Name in speech | 4.7 | 5.2 | 2.4 | 5.2 |
|  | $(0.6)$ | $(0.5)$ | $(0.8)$ | $(0.6)$ |

## Gesture Analyses

There was no difference in the total number of gestures produced by young and elderly participants, $F(1,39)=$ $1.15, p>.05$. We conducted a mixed ANOVA with different gesture types (pointing, iconic and tracing) as the within subject variable and group (elderly vs. young) as the between subject variable. There was a main effect of gesture
type, $F(1.20,44.34)=46.59, p<.001$. There was also a marginally significant interaction between gesture type and group, $F(1.20,44.34)=3.79, p=.051$. Both young and elderly adults used more tracing gestures ( $M=.80$ and $M=$ 1.23, respectively) compared to pointing ( $M=.29$ and $M=$ .16 , respectively) and iconic gestures ( $M=.06$ and $M=.06$, respectively). However, the difference between pointing and iconic gestures was only significant for young adults, $M_{\text {diff }}=$ $.23, p<.01$. Elderly adults produced comparable number of pointing and iconic gestures, $M_{d i f f}=.10, p>.05$.

Table 2: Mean raw gesture scores. The values in parentheses are standard errors of mean.

|  | Young | Elderly |
| :---: | :---: | :---: |
| Composite Gesture Score | 8.4 | 7.3 |
|  | $(1.6)$ | $(1.2)$ |
| Pointing Gestures | 2.5 | 1.2 |
|  | $(0.6)$ | $(0.3)$ |
| Tracing Gestures | 4.9 | 5.9 |
|  | $(0.9)$ | $(1.0)$ |
| Dynamic Gestures | 1 | 0.2 |
|  | $(0.5)$ | $(0.2)$ |

## Gesture \& Speech Analyses

Young adults' total number of gestures and spatial language positively correlated in both $\operatorname{SG}(r=.54, p<.05)$ and GR ( $r$ $=.70, p<.01$ ) conditions. There was also a positive correlation for spatial language use in SG and GR conditions ( $r=.58, p<.01$ ) in young adults. The total number of gestures did not correlate with spatial language use in SG ( $r=-.42, p>.05$ ) and GR ( $r=.01, p>.05$ ) conditions in elderly adults. Moreover, the spatial language use in SG and GR conditions did not correlate in elderly adults as well, $r=.08, p>.05$.

## Discussion

This is one of the first studies investigating how the effects of gesture restriction on speech production differ with aging. We asked whether gesture prohibition impair or enhance spatial speech production of young and elderly adults. Our results showed that the effects of gesture restriction on the production of spatial language differed between young and elderly. Even though elderly and young adults produced comparable number of gestures in spontaneous speech condition, gesture restriction increased the use of spatial information in speech only for elderly adults. Overall, younger individuals produced more spatial language and the use of gestures correlated with their use of spatial information.

In the first task when people were allowed to gesture, young and elderly individuals produced similar number of
gestures. This result contradicts with the previous findings that showed evidence for decreased amount of gestures in elderly adults (Cohen \& Borsoi, 1996; Feyereisen \& Harvard, 1999). However, the majority of the gestures used by young and elderly adults in our study were nonrepresentational (e.g. tracing and pointing gestures). Since the map was present throughout the experiment, this might trigger the frequent use of pointing to the map (i.e. pointing gestures) or continuously moving finger on map to trace the route (i.e. tracing gestures). When people are asked to talk about an object from memory, they use more representational gestures compared to a condition where the object of interest is present when they talk about it (Wesp et al., 2001). Thus, the sparse use of iconic gestures might lead us not to find any difference between young and elderly adults in the use of representational gestures.

Our results showed that elderly adults used more spatial language when they were restricted from gesturing compared to spontaneous speech, whereas young adults produced comparable levels of spatial language in both conditions. The higher use of spatial language when gesture use was restricted in elderly adults is compatible with the account suggesting that gestures are used to supplement and/or complement the speech. That is, when gestures convey the imagistic properties of spatial events, this information can be omitted from speech (Melinger \& Levelt, 2004; Graham \& Heywood, 1975). Our results, however, did not find support for the Information Packaging Hypothesis (Kita, 2000) or the Lexical Retrieval Hypothesis (Hadar \& Butterworth, 1997; Krauss, Chen, \& Gottesman, 2000). On the other hand, our findings from young adults did not support any of accounts regarding the relation between speech production and gesture restriction. For young adults, gesture restriction did not affect their spatial speech production. However, we found that the gesture and speech production were positively correlated in young adults. Thus, young adults, who produced more gestures (possibly high in spatial skills) could use more spatial information overall.

Why does gesture restriction influence only elderly adults' speech? We cannot answer this question with certainty, but state some possible explanations. First, people's general verbal skills could be related to their gesture use. Hostetter and Alibali (2007) found that people with low verbal skills produced gestures to facilitate their speech, yet people with high verbal skills only supplemented their speech with gestures. Thus, elderly individuals who had high verbal skills could produce more spatial information in a gesture prohibited context. This does not necessarily explain the difference between young and elderly adults, which require future studies to find
answers. Second, not only verbal abilities, but also spatial skills (e.g., Chu \& Kita, 2011; Chu et al., 2014) or working memory and spatial integration in visual processing (Andersen \& Ni, 2008; Copeland \& Radvansky, 2007) could play a role in how people benefit from gestures. Increased problems in working memory and visual-spatial problems might particularly be a problem in the elderly group. Again, we did not examine the participants' spatial skills or working memory in the current study, hence we cannot make a conclusion regarding this issue. Future studies should investigate skill differences between these age groups to draw stronger conclusions. Also, the composition of our two different age groups in terms of sex might create a problem in the interpretation of findings. Finally, even though we told everyone not to gesture, it is possible that some people might have moved other body areas such as lips, eyes or parts of the body (Rimé et al., 1984).

Taken together, the present study provided new evidence for the role of gesture restriction on spatial language use from young and elderly adults. Surprisingly, we did not find detrimental effects of gesture prohibition on spatial language use in either groups. On the contrary, elderly people benefited from not using gestures. These findings suggest that gestures might serve different purposes for young and elderly people in the context of spatial language use.

## References

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. Spatial Cognition and Computation, 5(4), 307-331.
Andersen, G. J., \& Ni, R. (2008). Aging and visual processing: Declines in spatial not temporal integration. Vision Research, 48(1), 109-118.
Chu, M., \& Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. Journal of Experimental Psychology: General, 140(1), 102-116.
Chu, M., Meyer, A., Foulkes, L., \& Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. Journal of Experimental Psychology: General, 143(2), 694-709.
Church, R. B., \& Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. Cognition, 23(1), 43-71.
Cocks, N., Morgan, G., \& Kita, S. (2011). Iconic gesture and speech integration in younger and older adults. Gesture, 11(1), 24-39.
Cohen, R. L., \& Borsoi, D. (1996). The role of gestures in description-communication: A cross-sectional study of aging. Journal of Nonverbal Behavior, 20(1), 45-63.

Copeland, D. E., \& Radvansky, G. A. (2007). Aging and integrating spatial mental models. Psychology and Aging, 22(3), 569-579.
Dimeck, P. T., Roy, E. A., \& Hall, C. R. (1998). Aging and working memory in gesture imitation. Brain and Cognition, 37(1), 124-127.
Feyereisen, P., \& Havard, I. (1999). Mental imagery and production of hand gestures while speaking in younger and older adults. Journal of Nonverbal Behavior, 23(2), 153-171.
Graham, J. A., \& Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. European Journal of Social Psychology, 5(2), 189-195.
Hadar, U., \& Butterworth, B. (1997). Iconic gestures, imagery, and word retrieval in speech. Semiotica, 115(12), 147-172.

Hostetter, A. B., \& Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. Gesture, 7(1), 73-95.
Hostetter, A. B., Alibali, M. W., \& Kita, S. (2007). Does sitting on your hands make you bite your tongue? The effects of gesture prohibition on speech during motor descriptions. In D.S. McNamara \& J. G. Trafton (Eds.), Proceedings of the $29^{\text {th }}$ Annual Cognitive Science Society (pp. 1097-1102). Austin, TX: Cognitive Science Society.
Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. Psychological Bulletin, 137(2), 297-315.
Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), Language and gesture (pp. 162-185). Cambridge, UK: Cambridge University Press.
Krauss, R. M. (1998). Why do we gesture when we speak? Current Directions in Psychological Science, 7(2), 54-54.
Krauss, R. M., Chen, Y., \& Gottesman, R. F. (2000). Lexical gestures and lexical access: a process model. In D. McNeill (Ed.), Language and gesture (pp. 261-284). Cambridge, UK: Cambridge University Press.
Melinger, A., \& Levelt, W. J. (2004). Gesture and the communicative intention of the speaker. Gesture, 4(2), 119-141.
Rauscher, F. H., Krauss, R. M., \& Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. Psychological Science, 7(4), 226231.

Rimé, B., Schiaratura, L., Hupert, M., \& Ghysselinckx, A. (1984). Effects of relative immobilization on the speaker's nonverbal behavior and on the dialogue imagery level. Motivation and Emotion, 8, 311-325.
Wesp, R., Hesse, J., Keutmann, D. (2001). Gestures maintain spatial imagery. The American Journal of Psychology, 114(4), 591-600.

# The Effects of Duration Words and Spatial-Temporal Metaphors on Perceived Duration 

Demet Özer (dozer@ku.edu.tr)<br>Department of Psychology, Koç University<br>Rumelifeneri Yolu Sariyer 34450, Istanbul - TURKEY

Alexander Kranjec (kranjeca@duq.edu)
Department of Psychology, Duquesne University
600 Forbes Avenue Pittsburgh, PA 15282

Fuat Balcı (fbalci@ku.edu.tr)<br>Tilbe Göksun (tgoksun@ku.edu.tr)<br>Department of Psychology, Koç University<br>Rumelifeneri Yolu Sariyer 34450, Istanbul - TURKEY


#### Abstract

Subjective duration estimates are positively related to the magnitude of various non-temporal stimuli (e.g. Xuan et al., 2007). Our study investigated whether temporal and spatial magnitude information conveyed by linguistic stimuli would affect perceived duration in a temporal reproduction task. We used time-related words referring to different exact durations (e.g. second; Experiment 1), and spatial-temporal metaphors (e.g. long), referring to indistinct temporal as well as spatial magnitudes (Experiment 2). In both experiments, participants over-reproduced the shorter target duration ( 2.4 s ) and underreproduced the longer target duration (4.8 s). In Experiment 1, participants under-reproduced the longer target duration more when they saw "week" in the training and "year" in the reproduction. Yet, we did not observe the same semantic magnitude effect in other word pairs either in Experiment 1 or 2. Overall, we did not find supporting evidence for magnitude information conveyed by language affecting subjective time estimates.


Keywords: time perception; language

## Introduction

The perception of time is a key feature of many biological and behavioral processes. Although accurate timing is essential to many daily tasks, substantial evidence shows that the subjective experience of time is not perfectly isomorphic to physical time (Zakay, 1993). Rather, perceived durations are contracted or dilated depending on many factors, including changes in non-temporal stimulus properties.

In this study, we investigated how perceived durations are modulated by temporal magnitude information provided in the medium of language.

## The Interaction Between Non-Temporal Stimulus Magnitude and Perceived Duration

Subjective duration estimates are positively related to the magnitude of various non-temporal stimuli presented in
different modalities. In visual domain, duration judgments were observed to be longer for larger numbers (Xuan et al., 2007; Oliveri et al., 2008; Vicario, 2011), stimulus size (Ono \& Kawahara, 2007), stimulus luminosity (Goldstone, Lhamon \& Sechzer, 1978), and complexity (Schiffman \& Bobko, 1974). For example, people were more accurate and faster when classifying the duration of smaller magnitude numbers presented for a shorter time (congruent trials) than smaller magnitude numbers presented for a longer duration (incongruent trials) in a Stroop-like paradigm (Xuan et al., 2007). The effect of stimulus size on perceived duration has also been documented in many studies (e.g.,Ono \& Kawahara, 2007; Xuan et al., 2007; Rammsayer \& Verner, 2014). For example; when categorizing the durations of stimuli of different sizes by pressing one of four keys (" 1 " for short and " 4 " for long) in a temporal categorization task, people perceived larger visual stimuli as lasting longer compared to smaller visual stimuli of an equivalent duration (Ono \& Kawahara, 2007).

Although the studies cited above each investigate the effects of non-temporal magnitude information on perceived duration, to our knowledge, no study so far has investigated the effects of the temporal magnitude (i.e. duration) implied by word stimuli on time perception. If there is an effect of magnitude on subjective time estimations, then we should be able to see the same effect of magnitude information derived from the semantic representations activated by linguistic stimuli. However, how semantic representations of duration and magnitude information encoded by individual words interact with the representation of duration is mostly unknown.

## Interaction Between Language Processing and Low-Level Sensory/Perceptual Processing

A growing body of research investigating the interaction between language and perceptual processing suggests that semantic representations activated as we process linguistic stimuli affect the content-specific domain of low-level
sensory and perceptual processing (Glenberg, Kaschak, 2002; Zwaan, 2004; Kaschak et al., 2005). According to theories of embodied language processing, comprehension involves the perceptual and motor simulation of the situation described in the linguistic input. Thus, the comprehension of words referring to a particular modal event should interact with low-level perceptual processing of that event (Barsalou, 1999; Glenberg \& Kaschak, 2002; Zwaan, 2004; Kaschak et al., 2005). Many behavioral studies have provided evidence for an interaction between comprehension and perceptual processing, suggesting that higher-level semantic knowledge influences low-level sensory processing in visual perception (e.g. Spivey et al., 2001; Stanfield \& Zwaan, 2001; Zwaan, Stanfield \& Yaxley, 2002).

While the effects of language-activated semantic information on cognitive processing across a range of domains has been investigated, the effects of temporal magnitude representations activated by duration words and metaphors on the content-specific area of perceptual processing, namely duration perception, has not been studied. The present study aimed to fill this gap, in order to provide evidence informing both the duration perception and language processing literatures. To this end, in two experiments, we investigated how participants' reproduced duration estimations of a target interval are modulated when presented as different word types: 1) distinct temporal magnitudes (i.e. duration words; e.g. week vs. year) or 2 ) indistinct magnitudes or durations (i.e. spatial-temporal metaphors; e.g. long vs. short).

## Experiment 1

In Experiment 1, we investigated how words referring to different exact durations (e.g. second, year) affect duration estimations. We hypothesized that when the word in training refers to a shorter duration compared to the word presented in the reproduction (e.g. seeing the word "second" in the training and "minute" in the reproduction), participants would under-estimate (i.e. over-reproduce) the target interval and vice versa. We did not expect any systematic difference in reproduced duration estimations when participants are presented with the same words in both reproduction and training.

## Method

## Participants

Twenty-five Koç University students (16 females, $\mathrm{M}_{\text {age }}=21.7$ ) agreed to participate in exchange for course credit. We discarded one female subject because her average coefficient of variation (CV) was high (average CV across conditions $=.51$ ). All experiments were approved by the Institutional Review Panel for Human Subjects of Koç University.

## Task and Stimuli

We used a temporal reproduction task. In this task, we asked participants to reproduce a given target duration by pressing a pre-designated response button to approximate the target duration as closely as possible.

At the beginning of a trial, a word ("training word") was visually presented for one of two different target intervals ( 2400 ms or 4800 ms ). At the end of the target interval, a blank screen was presented for 1 second, followed by a fixation cross presented for a random interval between 500 ms and 1500 ms . Participants were then instructed to initiate the reproduction interval by pressing the space bar. Upon pressing the spacebar, another word ("reproduction word") appeared at the center of the screen, remaining for the entirety of the reproduction interval. The interval ended when the reproduction was perceived as temporally equivalent to the target and the participant released the spacebar. Following the termination of the reproduction interval, the next trial was presented after a random interval between 1000 and 2000 ms .

In Experiment 1, we chose four words referring to different exact durations in Turkish: saniye ("second"), dakika ("minute"), hafta ("week") and sene ("year"). There were two conditions presented in two different sessions. In Condition 1, the words appearing in training and reproduction were different (different word pairs). In this condition, we created two-word pairs out of these four words: ("second") vs. ("minute") and ("week") vs. ("year"). The order of the words also changed. Thus, in some trials participants saw the word referring to the shorter duration during training (e.g. "week" in training and "year" in reproduction) and vice versa (e.g. "year" in training and "week" in reproduction), making four different training word - reproduction word pairs. In Condition 2, however, the same word appeared both in the training and the reproduction intervals (same word pairs). Thus, in this condition, four words appeared both in training and reproduction (e.g. "week" in training and "week" in reproduction, etc.).

## Procedure

All words were presented at the center of the screen, printed in white on a black background. There were 30 presentations for each training word-reproduction word pair at each of the target durations. Hence, in each session, for two target durations and four word pairs, there were 240 experimental trials. We also added 24 trials ( $10 \%$ of the experimental trials) in which the target words appeared for a random interval between 500 and 5000 ms . We added them in order to avoid participants to habituate the two target intervals and label them as "short" and "long" durations throughout the experiment. the selected target intervals as "short" and "long" durations. Thus, in each session, there were 264 trials in total, 240 of which were used in the analyses. All trials were presented randomly. Additionally, to verify that participants looked at the screen, we asked them to report the last word they saw on the screen on 12
randomly selected trials. Participants who could not correctly identify the words three or more times were discarded from analyses (Only one participant in Experiment 2 was discarded on this front). Each participant completed the two sessions and the order of the sessions was counterbalanced across participants. Each experimental session lasted $50-60$ minutes and was separated by a minimum of 1 and a maximum of 5 days.

## Results

For every participant, we calculated the normalized reproduced time (i.e. the reproduced duration divided by the target duration) and averaged those scores for each word pair-target duration combination. Also, for each participant, we calculated the coefficient of variation (CV; i.e. standard deviation of each condition divided by its mean) for each condition. Reproduced intervals that were greater than three times, or smaller than one third of the target duration were excluded from the analyses. Also, the mean normalized reproduction scores and CVs that were above and below three standard deviations of the sample mean for any of the word pairs for a specific target duration were treated as outliers and excluded from further analyses.

The mean normalized reproduction times across target durations for second vs. minute and week vs. year can be found in Table 1. This table presents the over-reproduction of the 2.4 s duration and an under-reproduction of 4.8 s duration regardless of the word pair type (same vs. different) or the specific word pairs used.

Table 1: Mean normalized reproduction scores for two word pairs across two target durations. The first word in the pair is the one that was presented in training, and the second one during reproduction. The values in parentheses are the standard errors of the mean.

|  | 2.4 s | 4.8 s |
| :---: | :---: | :---: |
| second - minute | $1.19(.05)$ | $.84(.03)$ |
| minute - second | $1.21(.06)$ | $.84(.03)$ |
| second - second | $1.22(.05)$ | $.89(.03)$ |
| minute - minute | $1.22(.06)$ | $.86(.03)$ |
| week - year |  |  |
| year - week | $1.20(.05)$ | $.62(.03)$ |
| week - week | $1.15(.04)$ | $.83(.04)$ |
| year - year | $1.20(.05)$ | $.88(.03)$ |
|  | $1.18(.05)$ | $.87(.04)$ |

There was no effect of sex (all $p \mathrm{~s}>.19$ ) and the order of the conditions (all $p s>.18$ ) in any of the word pair-target duration combinations. Also, there was no interaction between sex and the session order (all $p s>.22$ ).

In Condition 1, we conducted three-way repeated measures ANOVA. The results showed that word pairs, $F$
$(1,20)=116.56, p<.001$, target duration, $F(1,20)=$ $110.72, p<.001$, and the order of the word referring to the shorter duration, $F(1,20)=40.60, p<.001$, had significant main effects. However, these main effects were qualified by an interaction between all three repeated factors, $F(1,20)=$ 49.66, $p<.001$. Further comparisons showed that, for week vs. year, reproduced durations were greater when week was given in reproduction ( $M=.80$ ) compared to training ( $M=$ .60 ) only when they were presented for 4.8 s .
In Condition 2, we conducted a two-way repeated measures ANOVA and found a main effect of the target duration, $F(1,21)=73.35, p<.001$, and the same word pairs, $F(3,63)=4.38, p=.007$. Pairwise comparisons showed that the mean normalized reproduced durations were greater for 2.4 s for all same word pairs compared to $4.8 \mathrm{~s}\left(M_{\text {diff }}=.34, p<.001\right)$. However, there were no significant differences between any of the same word pairs when we consider Bonferroni adjusted alpha levels of .008 per test (.05/6) in the pairwise comparisons. There was also no interaction between target duration and same word pairs, $F(3,63)=1.50, p=.22$.
To investigate the difference between the same and different word pairs, we averaged the mean normalized scores for the same and different word pairs separately and conducted a two-way repeated measures ANOVA. The results showed only a main effect of target duration, $F$ (1, 19) $=92.18, p<.001$. There was no difference between same and different word pairs, $F(1,19)=3.09, p=.095$ or the interaction between word pair type and target duration, $F(1,19)=3.26, p=.087$.

A two one-way ANOVA with all word pairs regardless of the word pair type (same vs. different) and the order of the shorter duration as repeated measures and CV scores as dependent measure was conducted separately for each target duration. For 2.4 s , there was no significant effect of word pair on CV scores, $F(4.14,95.18)=1.30, p=.25$. However, for 4.8 s , there was a main effect of word pair, $\mathrm{F}(7,161)=$ $78.65, p<.001$. Pairwise comparisons revealed that CVs were greater when participants saw "week" in the training and "year" in the reproduction compared to all other word pairs in 4.8 s (all $M_{\text {diff }}>.173$, all $p \mathrm{~s}<.001$ ). To see whether variability in perceived durations differed between Condition 1 and 2, we computed grand total CVs for same and different word pairs separately for each target duration and conducted a two-way repeated measures ANOVA. Results revealed a main effect of the target duration, $F$ ( 1 , 19) $=92.18, p<.001$. The CVs were greater for $2.4 \mathrm{~s}(M=$ 1.172) compared to $4.8 \mathrm{~s}(M=.806)$. There was no main effect of the word pair type (same vs. different), $F(1,19)=$ $3.09, p=.095$ and no interaction between two, $F(1,19)=$ $3.26, p=.087$.

In sum, in Experiment 1, we found that regardless of the word pair type and specific order, participants overreproduced the target duration of 2.4 s and under-reproduced 4.8 s . We also found that participants under-reproduced 4.8 s more when they saw "week" in the training and "year" in the reproduction compared to all word pair conditions. The

CV was also greater for that word pair ("week-year") compared to all other.

## Experiment 2

In Experiment 2, we investigated how words implying both a temporal magnitude as well as a spatial magnitude modulated duration estimations. To this end, we employed quantifiers that can refer metaphorically to different indistinct durations as well as spatial magnitudes (i.e. the size of an object; e.g. long vs. short). Our hypotheses were same as with Experiment 1.

## Method

## Participants

Twenty-five Koç University students (14 females, $\mathrm{M}_{\mathrm{age}}=21$ ) agreed to participate in exchange for course credit. One male participant was discarded because he did not pay attention to the experiment and one female participant was discarded because her mean normalized reproduced scores were outliers in 10 out of the 16 conditions.

## Task, Procedure \& Stimuli

The task and the procedure were identical to Experiment 1, except for the word stimuli used in the task. In Experiment 2, we used spatial adjectives and adverbs that are used as spatial-temporal metaphors referring to indistinct durations. We chose four words: uzun ("long"), klsa ("short"), geniş ("wide") and dar ("narrow"). In Condition 1, participants were trained with a spatial-temporal adjective and presented with the antonym of that word in the reproduction phase. We created 2 word pairs: "long" vs. "short" and "wide" vs. "narrow". The order of the words was reversed in this condition. In Condition 2, participants saw the same spatialtemporal adjective both in training and reproduction.

## Results

We used the same exclusion criteria as in Experiment 1. The averaged mean normalized reproduced intervals for short vs. long and narrow vs. wide for each target interval can be found in Table 2. Visual inspection of Table 2 suggests the over-reproduction of 2.4 s and an under-reproduction of 4.8 s in both word pairs.

Table 2: Mean normalized reproduction scores for two word pairs across two target durations. The values in parentheses are the standard errors of the mean.

|  | $2.4 \boldsymbol{s}$ | $\mathbf{4 . 8 ~ s}$ |
| :---: | :---: | :---: |
| short - long | $1.30(.05)$ | $.85(.02)$ |
| long - short | $1.30(.05)$ | $.86(.02)$ |
| short - short | $1.18(.04)$ | $.83(.03)$ |
| long - long | $1.20(.04)$ | $.83(.03)$ |


|  | $2.4 \boldsymbol{s}$ | $\mathbf{4 . 8 ~ s}$ |
| :---: | :---: | :---: |
| narrow - wide | $1.30(.05)$ | $.88(.02)$ |
| wide - narrow | $1.27(.04)$ | $.86(.02)$ |
| narrow - narrow | $1.21(.04)$ | $.81(.03)$ |
| wide $\boldsymbol{-}$ wide | $1.23(.05)$ | $.82(.03)$ |

Neither sex (all ps > .12) nor the order in which participants attended the two conditions (all ps > .09) affected the normalized reproduction scores in any of the word pair-target duration combinations. Also, there was no interaction between sex and the session order in any of the conditions (all $p \mathrm{~s}>.15$ ).

For Condition 1, we conducted a three-way repeated measures ANOVA. The results revealed only a main effect of target duration, $F(1,19)=122.96, p<.001$. The mean normalized reproduction scores were greater in 2.4 s ( $M=$ 1.29) compared to $4.8 \mathrm{~s}(M=.86)$ for all different word pairs. There were no main effects of the specific word pair (short vs. long and narrow vs. wide), $F(1,19)=.286, p=$ .599 , or the order of the shorter duration, $F(1,19)=.147, p$ $=.706$. Also, there was no interaction between all three repeated factors, $F(1,19)=.013, p=.910$.

For Condition 2, our analysis revealed only a main effect of target duration, $F(1,20)=69.56, p<.001$. Pairwise comparisons showed that the mean normalized reproduced durations were greater in $2.4 \mathrm{~s}(M=1.18)$ compared to 4.8 s ( $M=.81$ ). There was no main effect of the same word pairs, $F(3,60)=2.095, p=.110$. There was also no significant interaction between same word pairs and target duration, $F(3,60)=2.39, p=.078$.

We conducted two separate two-way repeated measures ANOVA with the averaged mean normalized reproduced durations for the same and different word pairs for each target duration. The results showed only a main effect of target duration, $F(1,17)=87.54, p<.001$. There was no significant difference between the averaged normalized reproduced durations for same and different word pairs, $F$ $(1,17)=2.70, p=.119$. Also, there was no interaction between two repeated factors, $F(1,17)=.999, p=.334$.

With participants' CV scores, we conducted a two-way repeated measures ANOVA with all word pairs and the target duration as the two repeated factors and the CVs as the dependent measure. The results showed a significant effect of target duration, $F(1,22)=35.338, p<.001$. Pairwise comparisons revealed that CVs were greater in 2.4 s ( $M=.260$ ) compared to $4.8 \mathrm{~s}(M=.213)$. There was no difference between any word pair, $F(3.31,72.83)=.639, p$ $=.607$. However, these results were qualified by an interaction between two, $F(7,154)=2.674, p=.012$. The follow-up multiple t-tests show that, when we consider Bonferroni adjusted alpha levels (.05/8 =.0062), CV scores were greater in 2.4 s in word pairs "long-short" $(M=.271)$, "wide-narrow" ( $M=.276$ ), "short-long" $(M=.259)$, "narrow-wide" $(M=.265)$, "wide-wide" $(M=.252)$ and "narrow-narrow" ( $\mathrm{M}=.255$ ) compared to the target
duration of 4.8s ( $M=.207, .204, .201, .211, .208, .218$; respectively).

In sum, in Experiment 2, we found an over-reproduction of 2.4 s and an under-reproduction of 4.8 s regardless of the word pair conditions, as in Experiment 1. However, we did not find any difference in mean normalized reproduced duration between any of the word pairs. We also found that CVs were greater in 2.4 s compared to 4.8 s for all four different word pairs as well as two of the same word pairs ("short" and "wide").

## General Discussion

In this study, we asked how language affects time perception. Specifically, we investigated how the temporal magnitude (Experiment 1; duration words) and spatialtemporal magnitude (Experiment 2; spatial-temporal metaphors) implied by words influenced subjective time estimates as assessed in temporal reproduction task. We hypothesized that increasing the magnitude conveyed by words from training to reproduction would lead to the overreproduction of the target duration, and vice versa. We found that (1) in two experiments, participants overreproduced 2.4 s and under-reproduced 4.8 s , regardless of the implied temporal / spatial magnitude of words (Figure 1 and 2), (2) CVs were greater in 2.4 s compared to 4.8 s in both experiments, and (3) participants' reproduced durations were smaller and CVs greater when they saw "week" in the training and "year" in the reproduction in 4.8 s compared to all other conditions in Experiment 1. Last, (4) we did not find any systematic effect of the temporal/spatial magnitude implied by words on perceived duration in both experiments.

The over-reproduction of 2.4 s and the under-reproduction of 4.8 s in our current study are in line with Vierordt's Law (for a review see Lejeune \& Wearden, 2009) and found in many timing studies in the literature across multiple timing tasks (e.g., Karşılar \& Balcı, 2016). This migration effect, which is the regression of duration estimates toward the mid-range of the target duration series, is likely due to the fact that all word pair-target duration conditions were presented randomly (i.e. interleaved) rather than in blocks.

We also detected a trend that CVs were greater for 2.4 s compared to 4.8 s . According to Weber's Law, although the variation of the reproduced duration increases proportionally with the to-be-timed intervals, these results might be best explained by an additive source of variability due to experimental manipulations (other than duration) in addition to the proportional one due to timing mechanism itself (e.g. generalized form of Weber's Law).

In Experiment 1, we found that the word pair week-year was under-reproduced more when presented for 4.8 s compared to all other word pairs. It means that participants thought of the target duration of 4.8 s as shorter when "week" in the training was followed by a word implying a larger temporal magnitude, like "year". However, we did not see the same effect in other exact duration word pairs in

Experiment 1 and spatial-temporal metaphor pairs in Experiment 2. This might be due to the larger temporal magnitude difference between these two words compared to the other word pair. Furthermore, the opposite effect was not observed for the year-week pair suggesting an asymmetrical form of time warping (see also Karşılar \& Balc, 2016). Further investigation is needed to determine if this effect is reliable.

Overall, we could not find supporting evidence for the effect of language on time perception. Both temporal magnitude and temporal/spatial magnitude information conveyed by words did not affect perceived duration (other than the word pair of week-year in 4.8s). Yet, it should be noted that there is no hypothesized model for the interaction between time perception and language. Thus, the current study is an exploratory one. However, in a recent study, Bottini and Casasanto (2010) investigated the effects of implicit spatial length information encoded in different object nouns (e.g. cigarettes, clothesline, footpath) on perceived duration and found a positive effect of spatial magnitude information conveyed in linguistic medium on time perception. Object nouns with relatively shorter implicit spatial lengths (e.g. cigarette) were remembered as appearing for shorter durations compared to nouns with longer implicit spatial lengths (e.g. footpath) despite each being presented for the same amount of time. However, we did not find the same kinds of effects. It is interesting when we consider that we used direct spatial magnitude information in Experiment 2, rather than an implicit one as in Bottini and Casasanto (2010). One possibility for falling short to replicate the findings of this study might be that the previously documented effects of magnitude on time perception are only for spatial and numerical magnitude (i.e. non-temporal) and not for temporal magnitudes. In other words, those findings might be present only for crossdomain effects. In the current study, however, we tested the impact of duration magnitude on duration perception, which is a within-domain interaction. Yet, in Experiment 2, we used spatial-temporal metaphors that implied both temporal and spatial magnitudes. One reason for the null effect in this experiment concerns the everyday use of spatial -temporal metaphors. Space and time are so intertwined that spatial adjectives are commonly understood as temporal concepts, especially in the context of a time reproduction task (Lakoff \& Johnson, 1980).

Another possible explanation for not finding data to support our hypothesis in both experiments, concerns the nature of our to-be-timed stimuli. Larger, more complex, and intense stimuli expand perceived duration (Eagleman, 2008). One mechanism for this effect is the modulation of attention and arousal by the non-temporal properties of the to-be-timed stimulus. For example, intense negative sounds expand subjective duration since they heighten physiological arousal (Mella et al., 2011). Also, apart from emotional valence, attentional modulation by highly dynamic stimuli might affect duration perception. For example, Karşılar and Balcı (2016) found that higher motion
coherence in a highly dynamic moving dot array may capture more attention to the non-temporal properties of the stimulus at the expense of attention to the timing task itself. This may result in the over-reproduction of a target interval when the coherence level is increased from training to reproduction. However, our stimuli were not emotionally arousing nor attention capturing. Also, magnitude was not inherently perceptible in the to-be-timed stimuli, but implied by words. Concrete, visual magnitude information presented as an inherent property of the external stimuli might affect perceived duration by better directing attentional resources to stimulus properties.

Last, the task we used might not be the most sensitive for exploring the possible effects of language on perceived duration. Other tasks, like temporal bisection (Allan \& Gibbon, 1991) or categorical timing (Wearden, 1992) that force participants to decide on whether the perceived target duration is shorter or longer compared to a reference interval, might better detect differences between conditions due to its specificity to perceptual time in future studies.

In sum, the current study did not support the hypothesis that temporal and spatial magnitude information conveyed by linguistic stimuli influences subjective duration estimations. Limitations of the current study and the absence of an hypothesized model to be rejected prevent strong conclusions, but higher-order linguistic representations may not reliably interact with a low-level domain like interval timing across experimental paradigms.

## References

Allan, L. G., \& Gibbon, J. (1991). Human bisection at the geometric mean. Learning and Motivation, 22(1), 39-58.
Barsalou, L.W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-660.
Bottini, R., \& Casasanto, D. (2010). Implicit spatial length modulates time estimates, but not vice versa. In Spatial Cognition VII (pp. 152-162). Springer Berlin Heidelberg.
Eagleman, D. M. (2008). Human time perception and its illusions. Current Opinion in Neurobiology, 18(2), 131136.

Glenberg, A. M. \& Kaschak M. P. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Goldstone, S., Lhamon, W. T., \& Sechzer, J. (1978). Light intensity and judged duration. Bulletin of the Psychonomic Society 12.1: 83-84.
Karşılar, H., \& Balcı, F. (2016). Asymmetrical modulation of time perception by increase versus decrease in coherence of motion. Attention, Perception, \& Psychophysics, 78 (8), 2690-2707.
Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A. \& Zwaan, R. (2005). Perception of motion affects language processing. Cognition, 94, B79-B89.

Lakoff, G., \& Johnson, M. (1980). Conceptual metaphor in everyday language. The Journal of Philosophy, 77(8), 453-486.
Lejeune, H., \& Wearden, J. H. (2009). Vierordt's The Experimental Study of the Time Sense (1868) and its legacy. European Journal of Cognitive Psychology, 21(6), 941-960.
Mella, N., Conty, L., \& Pouthas, V. (2011). The role of physiological arousal in time perception: psychophysiological evidence from an emotion regulation paradigm. Brain and Cognition, 75(2), 182-187.
Oliveri, M., Vicario, C. M., Salerno, S., Koch, G., Turriziani, P., Mangano, R., ... \& Caltagirone, C. (2008). Perceiving numbers alters time perception. Neuroscience Letters, 438(3), 308-311.
Ono, F., \& Kawahara, J. I. (2007). The subjective size of visual stimuli affects the perceived duration of their presentation. Perception \& Psychophysics, 69(6), 952957.

Rammsayer, T. H., Verner, M. (2014). The effect of nontemporal stimulus size on perceived duration as assessed by the method of reproduction. Journal of Vision, 14(5):17, 1-10.
Schiffman, H. R. \& Bobko, D. J. (1974). Effects of stimulus complexity on the perception of brief temporal intervals. Journal of Experimental Psychology 103.1: 156.
Spivey, M. J., Tyler, M. J., Eberhard, K. M., \& Tanenhaus, M. K. (2001). Linguistically mediated visual search. Psychological Science, 12(4), 282-286.
Stanfield, R. A. \& Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. Psychological Science, 12, 153-156.
Vicario, C. M. (2011). Perceiving numbers affects the subjective temporal midpoint. Perception- London, 40 (1), 23.

Wearden, J. H. (1992). Temporal generalization in humans. Journal of Experimental Psychology: Animal Behavior Processes, 18(2), 134.
Xuan, B., Zhang, D., He, S., \& Chen, X. (2007). Larger stimuli are judged to last longer. Journal of Vision, 7(10), 2.

Zakay, D. (1993). Time estimation methods do they influence prospective duration estimates? Perception London, 22, 91-91.
Zwaan, R. A., Stanfield, R. A. \& Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. Psychological Science, 13(2), 168-171.
Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. In B.H. Ross (Ed.), Psychology of Learning and Motivation (Vol. 44, pp. 35-62). San Diego, CA: Academic Press.

# Action and actor gaze mismatch effects during spoken sentence processing 

Dato Abashidze (dato.abashidze@hu-berlin.de)<br>CITEC, Department of Linguistics, Bielefeld University, Germany<br>Psycholinguistics, Humboldt University of Berlin, Germany

Pia Knoeferle (pia.knoeferle@hu-berlin.de)<br>Psycholinguistics, Humboldt University of Berlin, Germany


#### Abstract

Eye tracking research on situated language comprehension has shown that participants rely more on a recent event than on a plausible future event during spoken sentence comprehension. When people saw a recent action event and then they listened to a German (NP1-Verb-Adv-NP2) past or futuric present tense sentence, they preferentially looked at the recent event target over another plausible target object (that might be involved in a future action) independent of tense. This preferential inspection persisted even when future events and futuric present sentences were much more frequent within the experiment, or when a gaze cue biased towards the future action target. The present experiments extend this line of research by introducing incongruence (in Experiment 1 a past tense verb mismatched the recently seen action and in Experiment 2 an actor gaze cue mismatched the past tense sentence condition). Can the verb-action and the gaze-sentence mismatches eliminate the recent-event inspection preference? Would participants recall information in post-experimental memory tests better for matches (the futuric present tense condition) than mismatches (the past tense condition)? Results revealed inspection of the recent event target as participants processed the verb-action mismatch (Exp 1) and actor gaze incongruence (Exp 2). However, the gaze (but not the verb-action) incongruence eliminated the overall recent event preference in the NP2 region. The memory tests also showed some evidence for a reversal of the recent-event preference.


Keywords: Eye-tracking; spoken sentence comprehension; visual world paradigm; recent-event preference; eventsentence incongruence; actor gaze mismatch

## Introduction

Every day people see or hear about events in the world and effortlessly integrate language with what they see. Although previous research has examined how people understand language referring to events, little is known about how we interpret reference to a preceding event context in relation to language about future events. Previous research has revealed that both visual and linguistic context can rapidly guide the listeners' visual attention (e.g., Chambers, Tanenhaus, Eberhard, Filip, \& Carlson, 2002; Tanenhaus, Spivey-Knowlton, Eberhard, \& Sedivy, 1995) and expectations about events. For example, cues to event tense provided by the utterance can help comprehenders in developing expectations about future events (e.g., Altmann \& Kamide, 1999; Kamide, Scheepers, \& Altmann, 2003). At the same time, listeners tend to prioritize recently inspected event depictions (Knoeferle \& Crocker, 2007), or real-world event portrayals (over expectations of future events) when both recent and future events could, temporarily, relate to an utterance (Abashidze, Knoeferle, Carminati, \& Essig, 2011; Knoeferle, Carminati, Abashidze, \& Essig, 2011). We present two eye-tracking ex-
periments that are situated in the context of these extant findings. The present experiments examine to which extent the priority accorded to recent (vs. future) events in interpreting an utterance holds up when we weaken the congruence between the recent event and the unfolding utterance. To this end, we created mismatches between the recent event context (the action) and (the verb in) an unfolding utterance (Experiment 1) / an actor's gaze behavior (Experiment 2).

## The robustness of the recent-event preference

A number of studies have provided evidence for an attentional behaviour that has been dubbed the 'recent event preference' (e.g., Abashidze et al., 2011; Knoeferle \& Crocker, 2007): In Experiment 2 by Abashidze et al. (2011) participants saw a person performing an action (e.g., sugaring strawberries) and then they listened to either a past tense sentence (Der Versuchsleiter zuckerte kürzlich die Erdbeeren, 'The experimenter recently sugared the strawberries') or a futuric present tense sentence (Der Versuchsleiter zuckert demnächst die Pfannkuchen, 'The experimenter will soon sugar the pancakes'). During the sentence they saw the person in a static position and two objects on the table in front of him (e.g., pancakes and strawberries, i.e., Fig 1-B). After the sentence presentation a second event showed again a sugaring action (the 'future' action) but this time on the other object (e.g., sugaring pancakes). While participants listened to the sentence, their eye gaze to the two potential targets (of the recently seen action, and of a potential future action) were monitored. Results showed that participants preferentially inspected the recent event target (i.e., the strawberries) over the other plausible future event target. This happened even during the futuric present tense sentence, and they shifted gaze to the pancakes (the plausible future event target) only as it was mentioned. Follow-up studies examined this issue by increasing the number of the future events and of futuric present tense sentences up to $88 \%$ (Abashidze, Carminati, \& Knoeferle, 2014), by having the actor gaze at the targets before their mention (Abashidze, Knoeferle, \& Carminati, 2015), and by moving linguistic cues that could bias against the recent-event preference to the sentence beginning (Abashidze \& Chambers, 2016). Despite these strong visual and linguistic cues in favor of the future event target, these experiments replicated the overall recent event preference.

## The impact of incongruence and gaze cues

In language processing research, many studies have employed picture-sentence incongruence and verification as a
method. Experiments using this method have found that participants are sensitive to the incongruence, so that they responded faster to congruent than incongruent picturesentence pairs (Carpenter \& Just, 1975; Underwood, Jebbett, \& Roberts, 2004). In an eye-tracking study by Knoeferle and Crocker (2005) participants were presented with depicted scenes and either matching or mismatching np1-verb-advnp 2 sentences. The authors found an incongruence effect in the verb and adverb regions. Participants were faster reading these sentence regions in the congruous than incongruous condition (see also related findings on gender stereotype effects in a picture-sentence verification task by Rodríguez, Burigo, \& Knoeferle, 2015). Further studies used the sentence-picture verification procedure while participants were presented with positive and negated sentences (Glenberg, Robertson, Jansen, \& Johnson-Glenberg, 1999). Results showed that pictures matching the presented sentences (even when the sentences were negated) elicited faster responses than pictures mismatching the sentences. In addition, another cue that has been shown to rapidly influence comprehension (and guide participants' visual attention even when it was incongruous with language) is a speaker's gaze (e.g., Hanna \& Brennan, 2007; Kreysa \& Knoeferle, 2013; Staudte, Crocker, Heloir, \& Kipp, 2014).

The present experiments relied on incongruence in verbaction relations and an actor's gaze (to the future target) as a way to stress-test listeners' preference of inspecting the target of a recent action. The causes underlying the preferential inspection of the recent event are unclear. Perhaps the preferential inspection is guided by the verb. The verb could be linked to representations of the recently inspected action and its location, prompting participants to shift gaze to the location of the action when they encounter the verb. If so, then a match between the recent event and the sentence referring to it could boost the attention towards the recent event. By contrast, a mismatch between the visual and linguistic information could reduce the recent event preference. Alternatively, what we see is a general recency effect (i.e., participants inspect the object that is the target of the recent action, independent of verb meaning). If this were the case, then a mismatch between the recent action and the verb should not interfere with the recent-event preference but a gaze cue (e.g., the actor shifting gaze during the verb to the future target object) might diminish a recency effect and direct the listener's attention to the future target object.

## The present experiments

Given that incongruence has been shown to influence participant' eye-movements and their reaction times during picturesentence verification tasks, two eye-tracking studies examined to which extent incongruence could bias against the recent-event preference. In Experiment 1, the verb of the past (but not futuric present) tense sentences mismatched the recent action. In Experiment 2, the actor began to inspect the future target object at verb onset in the past tense sentences. We tested to what extent these incongruences will reduce the
preferential inspection of the recent event target during the verb. At this point people could realize that the verb does not match the action they saw in Exp 1 and notice that the actor shifts his gaze towards the future target from the verb onset while the past tense sentence refers to the recent event target in Exp 2. In both experiments the experimental trials were incongruent in the past tense condition only. These experiments used the design from Abashidze et al. (2011, Exp 2), presented above, with one factor, viz. tense (past vs. futuric present). If the recent event preference is sensitive to the incongruence between the recent event and either the past tense sentence $(\operatorname{Exp} 1)$ or the actor's gaze $(\operatorname{Exp} 2)$, then we should see a decrease in looks to the recent event target starting from the verb region. Can the congruence in the futuric present tense increase the inspections towards the future event target and override the preferential inspection of the recent event? The incongruence in the past tense might strengthen the congruence in the futuric present tense condition (e.g., Glenberg et al., 1999).

After the eye-tracking session, participants took part in a gated-memory (Exp 1) and a memory (Exp 2) test. Previous studies reported a better recall of the future event (Abashidze et al., 2015, Exp 1) which was not in agreement with the gaze data; however, other findings revealed a better recall of the recent events (Abashidze et al., 2014, Exp 1 and 2), and a better recall of the past tense sentence (Abashidze et al., 2015, Exp 2) underscoring the recent-event preference in the gaze data. If the incongruence affects the recent-event preference and the incongruence effects are long-lasting, then we might see a reduced recall performance for recent compared with future events in the memory test. Alternatively, the incongruence does not affect the recent-event preference and / or its effects are short-lived, in which case we might see better recall of recent than future events.

## Experiments 1 and 2: Methods

## Participants

Thirty-two native German University students in each experiment (aged 18 to 32) with normal or corrected-to-normal vision gave informed consent and received 6 Euros each for their participation. The study was approved by an ethics vote (Experiment 1: Bielefeld University ethics committee, Experiment 2: DGfS).

## Materials and design

The current experiments used the experimental sentences from Abashidze et al. (2014, see Table 1). All sentences ( $N=24$ ) had the structure NP-VERB-ADV-NP and two native German speakers recorded them. The sentences were in two tense conditions and referred either to a recently seen event or a plausible future event. In one condition, the verb was in the present tense and a time adverb (demnächst, 'soon') indicated the futuric present tense condition (Table 1b). In the other condition, the verb was in the simple past, and a time adverb (kürzlich, 'recently') indicated the past tense condi-
tion (Table 1a). The critical sentences employed only regular German verbs in which the verb was tense ambiguous up to but excluding the word-final phoneme which disambiguated towards the simple past in the past tense condition. As we can see in Table 1, the experiments used two sentences for each tense condition. With this counterbalancing we ensured that each object was once the target of the recent and once the target of the future event. The critical words in a sentence were matched for spoken syllables and lemma frequency within an item (Baayen, Piepenbrock, \& Gulikers, 1995).

Experiment 1 also used the videos from Abashidze et al. (2014) for futuric present tense sentences. These videos ( $M_{\text {duration }}=5015$ ) showed an actor sitting at a table in front of two objects (e.g., strawberries and pancakes, one on the left and one on the right; both of the objects can be sweetened). We additionally recorded new videos (for the recent events). For example, instead of sugaring strawberries, a first video presented the actor tasting the strawberries (see Fig 1A), and the verb in the past tense sentence (translated literally: 'The experimenter sugared recently the strawberries') never matched the recent action. By contrast, the futuric present sentence always matched the future event (see Fig 1-C).

In Experiment 2, for each experimental trial, participants saw a short video before and after hearing a sentence about a person performing an action. For the incongruence, we used the gaze videos from Experiment 2 by Abashidze et al. (2015). For instance, when participants listened to a past tense sentence they saw a video of the actor shifting his gaze towards the future event target (i.e., pancakes) from the onset of the verb, where it remained until the end of the sentence (the gaze cue mismatched the referential past tense sentence). By contrast, in the futuric present tense condition no gaze cue was present. Participants saw a snapshot from the last frame of the first video showing the actor in a static position looking straight ahead (i.e., see Fig 1-B). In both experiments the incongruence biased against the recent event preference.

In addition to the experimental items we created 36 filler sentences. To balance the incongruence across/within experiments, 12 fillers in Experiment 1 featured an incongruence between the futuric present verb and future event. Similarly, 12 fillers in Experiment 2 showed the actor looking at the recent event target during a futuric present sentence. In both experiments, recent and future events appeared equally often.

Thus, both experiments manipulated 1 factor: sentence tense (past vs futuric present); in half of the trials the sentence was in the past tense and in the other half in the futuric present tense (see Table 1 for counterbalancing). The resulting four lists used a Latin square design. Each experimental list contained every critical item in only one condition and all fillers. Each subject saw an individually pseudo-randomized version of one of the four experimental lists.

## Procedure

An Eyelink 1000 eye-tracker recorded participants’ eye movements. After a successful 9-point calibration, the experiment began. Participants were asked to inspect the scene

Figure 1: Sequence of events of a typical experimental trial for Experiment 1

and to listen carefully to the sentences. As in the previous studies by Abashidze et al. (2011), on a given trial, a participant first saw a video of a person (the actor) performing one action before the sentence (e.g., tasting strawberries for Exp 1 and sugaring strawberries for $\operatorname{Exp} 2$ ); then participants saw a static photo (see Fig 1-B). 700 ms after the onset of the static photo, a sentence was presented via the loud speakers either in (a) the past tense or (b) the futuric present tense (see Table 1). In Experiment 1 (experimental items), the past tense verb did not match the recent event; but the verb of the futuric present matched the future event (shown after the sentence had ended). In Experiment 2, during the past tense sentence the actor directed his gaze towards the future event target from verb onset (the gaze cue mismatched the past tense sentence and its NP2 referent). However, during the futuric

Table 1: Example experimental sentences. The indices (') indicate counterbalancing versions

|  <br> counterbalancing | Sentences |
| :--- | :--- |
| 1a past tense | Der Versuchsleiter zuckerte kürzlich <br> die Erdbeeren <br> 'The experimenter recently sugared <br> the strawberries' |
| 1a' past tense | Der Versuchsleiter zuckerte kürzlich <br> die Pfannkuchen <br> 'The experimenter recently sugared <br> the pancakes' |
| 1b futuric present | Der Versuchsleiter zuckert demnächst <br> die Pfannkuchen |
| 'The experimenter will soon sugar |  |
| 1b' futuric present | Der Versuchsleiter zuckert demnächst <br> die Erdbeeren <br> 'The experimenter will soon sugar <br> the strawberries' |

present sentence participants saw the actor in a static position throughout the sentence (as in Experiment 1, see Fig 1-B). 700 ms after the sentence had ended, participants saw a video of the actor performing the second action event (e.g., sugaring pancakes, Fig 1-C, both experiments). Post-experiment, participants completed a gated memory test in Experiment 1 and a memory test in Experiment 2 (Fig. 2). At the end, they were debriefed. Each experiment lasted approximately 50-55 minutes.

## Memory tests

Experiment 1 tested participants' later memory of the linguistic information and Experiment 2 examined participants' later memory of the visual information. Experiment 1 shows an example sentence as presented in the gated memory test in a 3-stage procedure (Fig 2, Exp 1). At the first stage, participants saw only the first noun phrase and the verb stem and had to verbally complete the verb tense. The second stage added the temporal adverb, and they had to recall the second noun phrase. If they were unable to do so, they received a further prompt at the third stage and had to select the correct referent out of three objects. Two of these were from that sentence trial and the third was a distractor from another filler item.

Figure 2: An example of a sequence of stages in the gated memory test, $\operatorname{Exp} 1$ and display for the memory test, $\operatorname{Exp} 2$


For the memory test in Experiment 2, we created two snapshots of the first and second video of each experimental item, i.e., showing the experimenter performing one of the two actions (Fig 2, Exp 2). The two snapshots associated with each item were combined into one display and shown to participants. Two versions were created in which the respective location of the two pictures was counterbalanced and participants responded with a button press. Above the picture, one of two questions appeared:
(a) Welche Aktion wurde VOR dem Satz durchgeführt?
"Which action was performed before the sentence?"
(b) Welche Aktion wurde NACH dem Satz durchgeführt? "Which action was performed after the sentence?"

## Experiments 1 and 2: Analyses and results

## Eye tracking

For the eye-tracking data we divided each experimental sentence into three time regions, (the verb, the adverb and the NP2). Each word region lasted from its onset to the onset of the following word region and NP2 ended at sentence offset. The measure of interest was inspection of the recent and future target objects. Because looks to one of the objects implied fewer looks to the other target objects, we computed mean log gaze probability ratios for the recent relative to the future target $(\ln (P($ recent target $) / P$ (future target $))$ ). A score of zero indicates that both targets are inspected equally often; a positive value means more looks to the recent event target; a negative value means more looks go to the future event target (see Knoeferle et al., 2011).

For the inferential analyses, we performed separate ANOVAs on the mean log ratio averaged for each condition (past vs. futuric present) and word region by participants and by items respectively. The independent variable was tense, with two levels, past and future tense. We tested the significance of the intercept overall (a positive intercept represents a preference of inspection of the recent event target).

In Figure 3, the dotted lines indicate the past tense condition and the solid lines indicate the futuric present tense condition. As we can see, the incongruence influenced target inspection during sentence comprehension. In Experiment 1, participants decreased their attention towards the recent event target at the end of the verb region (i.e., following the mismatch); however both lines (in the mismatching past and matching futuric present tense) remain above zero, meaning that people continued to preferentially inspect the recent (vs. future) event target. In Experiment 2, the preferential looks towards the recent event target lasted until the middle of the adverb region in both tense conditions. Interestingly, despite the incongruent gaze in the past tense condition, participants' attention towards the recent event target persisted until sentence end but decreased as the target was mentioned.

A noticeable difference between Experiment 2 and Experiment 1 is thus that the gaze incongruence (but not the verbaction incongruence) seems to have prompted participants to decrease their attention to the recent target object during NP2. At the end of the NP2 (the name of the recent event target), participants inspected the recent and future event targets equally often in the past tense condition (unlike in Experiment 1).

ANOVAs revealed a tense effect in the NP2 region in Experiment 1 and in the Adverb and NP2 regions in Experiment 2, reflecting that tense modulated the listeners' looks to the recent (vs. future) event target. The grand mean (i.e., the mean of both conditions / the intercept) was positive in all regions in Experiment 1 and in the Verb and Adverb regions in Experiment 2, which indicates an overall recent-event preference (significant intercept in all the ANOVAs by region). Thus, Experiment 1 replicated the overall preference to look at the recent event target in all three word regions indepen-

Figure 3: Mean log gaze probability rations ( $\ln (P$ (recent target) $/ P$ (future target))) by condition from verb onset for Exps 1 and 2


dent of tense condition, whereas Experiment 2 did not reveal a significant intercept in the last word region (NP2), suggesting the gaze incongruence eliminated the overall recent event preference in the NP2 region. Furthermore, a post-hoc two-way ANOVA on the mean log gaze ratios of the NP2 region revealed a tense effect and a marginal experiment effect but no interaction. This experiment effect suggests that the gaze incongruence not only eliminated the overall recent event preference but also marginally decreased participants’ attention towards the recent event target in Experiment 2 compared with Experiment 1.

## Memory tests

Gated memory test We calculated the percentage of correct responses by conditions. Participants correctly answered questions from all three stages on average with $61 \%$. They correctly recalled $54 \%$ at stage one, which is more accurate than at stage two with $40 \%$. Importantly, the highest accuracy emerged at stage three ( $89 \%$ ). Subjects recalled the futuric present sentence (match) better than the past tense sentence (mismatch). Logistic linear mixed effect (LME) analyses showed a marginal tense effect at stage $1(p<.08)$ and a fully significant tense effect at stage $2(p<.02)$, indicating higher accuracy for the futuric present than past tense sentence condition. Memory test In Experiment 2 we calculated the percentage of correct answers in indicating the event targeted in the image. Participants correctly answered questions on average with $61 \%$. They were slightly more accurate in recognizing the future (matching) events ( $60 \%$ ) than the recent (mismatching) events (59\%). The LME analyses did not reveal any significant difference in recalling the recent versus future events.

## Discussion

Across two experiments, we examined the recent-event preference (e.g., Abashidze et al., 2014; Knoeferle et al., 2011) and stress-tested it with two types of incongruence (verbaction and the actor's gaze to the future target in the past tense condition).

We had predicted that if the recent event preference were guided by the verb, then participants' overall preference to inspect the recent event target should disappear when they realize that the recent action mismatched the verb in meaning (this did not happen, Exp 1). Analyses of the data from both eye-tracking experiments did not show an early preferential inspection (during the verb and adverb) of the future event target but rather corroborated participants' preference to gaze at the recent event target. It is possible that the effect of the verb-action incongruence was weak since the past tense sentence, while mismatching at the verb, did mention the correct event target. Perhaps for this reason, inspection of the recent event target persisted during the Adverb and increased during NP2 in the past tense sentences of Experiment 1.

Furthermore, if a recency effect underlies the recent-event inspection preference and more generally object inspection, then an incongruent actor's gaze to the future target (as the most recent cue) should have guided the listeners' attention to that target during the verb and adverb for the past tense sentences, which did not happen; rather participants preferentially inspected the target of the past event i.e., in line with the sentence tense and the recent-event preference.

While previous studies revealed an immediate gaze effect in a congruent environment (at around $300-500 \mathrm{~ms}$ after its onset, e.g., Kreysa \& Knoeferle, 2013), participants in Experiment 2 fully decreased their inspection of the not-gazedat recent event target only at sentence end (i.e., after mention of the target), thus ignoring that the actor gazed at the future target from verb onset. This suggests a strong reliance on the recently-seen event and a relatively slow effect of the actor's gaze when it had to compete with the preceding action event referenced by the verb. Although the gaze mismatch eliminated the overall recent event preference in the NP2 region of Experiment 2, a between-experiment comparison of the same word region did not reveal a fully significant experiment effect between Experiment 1 and 2. The two types of incongruence hence did not differ reliably in the extent to which they disrupted the recent event preference at sentence end. For the Adverb region, however, between-experiment analyses clarified that participants were more likely to inspect the recent event target in Experiment 2 than 1, corroborating that gaze did not immediately modulate this inspection preference. Tense effects in the Adverb and NP2 regions in Experiment 2 (compared with only the NP2 region in Experiment 1) revealed that the gaze incongruence (and the actor's attention to the future event target) did boost the integration of tense after the verb in Experiment 2 compared with 1.

The post-experiment memory tests results in Experiment 1 did not agree with the overall recent event preference in the
gaze data (recall was reliably better for the futuric present tense sentences, in conflict with the recent event inspection preference (see also Abashidze et al., 2015)). In Experiment 2 , no reliable difference in recall emerged for the recent versus future events, suggesting short-lived effects of the gaze mismatches. The better recall of the futuric present condition in Experiment 1 could be explained if we assume that the congruent recent events and linguistic information evoked more in-depth processing and increased attention to the stimuli that then also benefitted the later recall of event information.

In conclusion, the incongruences in the past tense sentences did not reduce the overall recent event preference immediately (during the verb and adverb in both experiments); but at least actor gaze incongruence did eliminate the overall preference eventually, in the NP2 region of Experiment 2 and it boosted the tense effects. What these results suggest is that the recent-event inspection preference it robust, and that it is not entirely dependent upon verb reference or cue recency. The recall accuracy in Experiment 1, by contrast, suggests that the verb-action mismatches affected short-term memory of the events. Gaze mismatches, by contrast, seem to have had immediate effects in the sense that they boosted tense effects but they neither reduced the overall inspection preference more than verb-action mismatches at NP2, nor did they modulate recall of the stimuli.

## Acknowledgments

This research was funded by the Cognitive Interaction Technology Excellence Center 277 (German Research Foundation, DFG). We thank Johanna Bokelmann, David Nothdurfter and Luise Henneberg for help in conducting Experiment 2.

## References

Abashidze, D., Carminati, M. N., \& Knoeferle, P. (2014). How robust is the recent event preference? In Proceedings of the 36th Annual Meeting of the Cognitive Science Society (pp. 92-97). Cognitive Science Society.
Abashidze, D., \& Chambers, C. G. (2016). The role of early linguistic cues in the recent event preference. In Proceedings of the Conference on Architectures and Mechanisms for Language Processing.
Abashidze, D., Knoeferle, P., \& Carminati, M. N. (2015). Eye-tracking situated language comprehension: Immediate actor gaze versus recent action events. In Proceedings of the 37th Aannual Meeting of the Cognitive Science Society (pp. 31-36). Cognitive Science Society.
Abashidze, D., Knoeferle, P., Carminati, M. N., \& Essig, K. (2011). The role of recent real-world versus future events in the comprehension of referentially ambiguous sentences: Evidence from eye tracking. In Proceedings of the EuroCogSci. New Bulgarian University Press.
Altmann, G., \& Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition, 73, 247-264.

Baayen, R., Piepenbrock, R., \& Gulikers, L. (1995). The celex lexical database philadelphia: University of Pennsylvania. Linguistic Data Consortium.
Carpenter, P. A., \& Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. Psychological Review, 82(1), 45.
Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., \& Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. Journal of Memory and Language, 47(1), 30-49.
Glenberg, A. M., Robertson, D. A., Jansen, J. L., \& JohnsonGlenberg, M. C. (1999). Not propositions. Cognitive Systems Research, 1(1), 19-33.
Hanna, J. E., \& Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. Journal of Memory and Language, 57(4), 596-615.
Kamide, Y., Scheepers, C., \& Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. Journal of Psycholinguistic Research, 32, 3755.

Knoeferle, P., Carminati, M. N., Abashidze, D., \& Essig, K. (2011). Preferential inspection of recent real-world events over future events: Evidence from eye tracking during spoken sentence comprehension. Frontiers in Psychology, 2.
Knoeferle, P., \& Crocker, M. (2005). Incremental effects of mismatch during picture-sentence integration: Evidence from eye-tracking. In Proceedings of the 26th Annual Meeting of the Cognitive Science Society (pp. 1166-1171).
Knoeferle, P., \& Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: evidence from eye-movements. Journal of Memory and Language, 75, 519-543.
Kreysa, H., \& Knoeferle, P. (2013). Reference-related speaker gaze as a cue in online sentence processing. In D. Kluge A. \& Söffker (Ed.), Mensch Teams Systeme und Automaten. Duisburg: DuEPublico.de.
Rodríguez, A., Burigo, M., \& Knoeferle, P. (2015). Visual gender cues elicit agent expectations: different mismatches in situated language comprehension. In Proceedings of the Euro Asian Pacific CogSci (pp. 234-239).
Staudte, M., Crocker, M. W., Heloir, A., \& Kipp, M. (2014). The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts. Cognition, 133(1), 317-328.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., \& Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 632-634.
Underwood, G., Jebbett, L., \& Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. Quarterly Journal of Experimental Psychology Section A, 57(1), 165-182.

# It's Time: Quantifying the Relevant Timescales for Joint Attention 

Drew H. Abney (dhabney@indiana.edu)<br>Linda B. Smith (smith4@indiana.edu)<br>Chen Yu (chenyu@indiana.edu)<br>Department of Psychological and Brain Sciences<br>Indiana University, Bloomington, IN 47405 USA


#### Abstract

The study of the coordination of attention, a term called joint attention (JA), has resulted in a better understanding of the dynamics and development of communication. Despite the important insights gained from studying JA, there is little consensus regarding the specific components that are included in operationalizing JA. The present work explored a parameter space of JA during a dyadic naturalistic toy play task between 9-monthold infants and their parents. We systematically measured the temporal properties of two components commonly used to operationalize JA: the duration of continuous alignment of parent and infant visual fixations and the flexibility of fluctuations of attention. The results show that very brief bouts of JA are important predictors for vocabulary development. The results from this work provide new insights into the specific properties used to operationalize JA and point to the importance of considering multiple timescales of behavior that make up JA.


Keywords: joint attention; communication; development; language development; methodology

## Introduction

Human interaction consists of behaviors that occur across multiple timescales. When an infant interacts with a parent, physiological rhythms are coordinated within a 1 s timescale (Feldman, Margoi-Cohen, Galli, Singer, \& Louzoun, 2011), vocalizations at a 3 s timescale (van Egeren, Barratt, \& Roach, 2001; Harder et al., 2015), and leader-follower dynamics of vocalizations fluctuate across a 10s-temporal window (Abney, Warlaumont, Oller, Wallot, \& Kello, 2016). Infants and their parents also coordinate their attention onto objects, a coordinative behavior called joint attention. The achievement of joint attention emerges early in the first year of life (Scaife \& Bruner, 1975), and has been shown to be a fundamental component of communicative skills ranging from the development of language to social competencies (Mundy \& Newell, 2007). The main goal of the current paper is to determine the relevant timescales for joint attention during infancy.

The empirical study of joint attention was initiated by the seminal work of Scaife and Bruner (1975) observing that infants could follow the direction of a partner's gaze within the first year, and that this behavior increased in frequency with age. Since Scaife and Bruner's original findings, decades of research have led to important theoretical and empirical contributions to areas of psychology ranging from
basic questions and connections about attentional processes (Corkum \& Moore, 1995; Mundy, Card, \& Fox, 2000), to whether or not joint attention is critical for language development (Baldwin, 1995; Tomasello, 1988; Akhtar \& Gernsbacher, 2007), and has led to proposals about the origins of theory of mind (Baron-Cohen, 1991) and communication (Tomasello, 2010).

Table 1: Summary of an abbreviated literature review of studies investigating the relationship between joint attention and language. Note: u.r. $=$ under review. ${ }^{n}=$ semi-naturalistic play paradigm. ${ }^{h t}=$ head turn paradigm. Age in is months. T $=$ timescale ( s )

| Author | Year | N | Age | T |
| :--- | :--- | :--- | :--- | :--- |
| Bakeman \& Adamson $^{n}$ | 1994 | 28 | $6-18$ | 3 |
| Brooks \& Meltzoff $^{\text {ftt }}$ | 2005 | 96 | $9-11$ | 6.5 |
| Brooks \& Meltzoff $^{h t}$ | 2008 | 32 | $10-11$ | 6.5 |
| Carpenter et al. $^{n}$ | 1998 | 24 | $9-15$ | 3 s |
| Morales et al. $^{n}$ | 1998 | 22 | $2-18$ | NA |
| Mundy, Sigman, \& Kasari $^{n}$ | 1990 | 45 | 33 | NA |
| Tomasello \& Farrar (Exp. 1) $^{n}$ | 1986 | 24 | $15-21$ | 3 |
| Tomasello \& Farrar (Exp. 2) $^{n}$ | 1986 | 10 | 17 | 3 |
| Tomasello \& Todd $^{n}$ | 1983 | 6 | $12-13$ | 3 |
| Yu, Suanda, \& Smith $^{n}$ | u.r. | 26 | 9 | .5 |

Although the overall consensus is that joint attention is an important ability, there is less agreement and consistency regarding how joint attention is defined and operationalized. For example, in Scaife and Bruner's (1975) original work, a positive joint attention behavior was coded if an infant (a) looked in the same direction as the experimenter without (b) intervening looks elsewhere within (c) 7s of the experimenter's look. In another example, Bakeman and Adamson (1984) coded behavior as a coordinated joint engagement state if the infant (a) actively coordinates his or her attention with another person and an object for (b) a particular duration with (c) only brief attention shifts to other objects for less than 3s. Finally, Tomasello and colleagues (Tomasello \& Todd, 1983; Tomasello \& Farrar, 1986) defined joint attention as when (a) infant and parent both visually attended to the same object for (b) at least 3 s with (c) only brief looks elsewhere. Table 1 provides an abbreviated review of the timescales used to operationalize
joint attention for studies focused on the relationship between joint attention and language.

In free-flowing parent-infant everyday interaction, joint attention is embedded in a stream of free-flowing activity in which parents both react to and attempt to control toddlers' behaviors and in which toddlers react to, direct, and sometimes ignore parents as they pursue their own goals. In those naturalistic contexts, we know that adults generate on average 3 eye fixations per second (Hayhoe, Shrivastava, Mruczek, \& Pelz, 2003) and we also know, from our recent head-mounted eye tracking studies, infants generate lots of short looks in toy play (Yu \& Smith, 2016). That means the exquisite real-time "dance" of social interactions require effective adjustments within the dyad and socially coordinated shifts in attention have to be resolved in fractions of a second. In this context, it is possible that joint attention spans multiple timescales, displaying important variability at short timescales and also at longer timescales.

The specific goal of the current paper is to investigate the relevance of two key parameters used to operationalize joint attention that have varied considerably across research groups: the duration of continuous alignment of parent and infant visual fixations and the flexibility of fluctuations of attention. We refer to the former parameter as minimum joint duration and the later parameter as minimum duration. For minimum joint duration, we varied the duration to estimate joint attention across micro-level (e.g., 500 ms ) and macro-level (e.g., 10s) timescales. We focused on two aspects of this manipulation. First, we investigated how properties of joint attention, like mean duration, frequency, and proportion, varied across the minimum joint attention dimension. Importantly, we were also interested in how minimum joint duration affected how many dyads in our sample exhibited joint attention at a particular timescale. For example, it is possible that some infant-parent dyads do not exhibit any macro-level joint attention bouts, which would require us to omit them from subsequent analyses. For minimum duration, we first kept the parameter fixed at 0 ms to simulate coding schemes that only identified a JA bout as continuous visual alignment on a target object with no looks elsewhere (Study 1) and then manipulated the parameter to allow for brief fluctuations of attention less than 300 ms (Study 2).

Second, we asked how different values of minimum joint duration impacted the predictive value of joint attention. Previous research has observed that joint attention correlates with concurrent - and predicts future - vocabulary size (Baldwin, 1995; Tomasello \& Farrar, 1986; Tomasello \& Todd, 1983; Smith et al., 1988; Carpenter et al., 1998; Morales et al., 2000; Mundy, Sigman, \& Kasari, 1990). Therefore, to determine how minimum joint duration affects the predictive value of joint attention for vocabulary size, in Study 3, we estimated joint attention for different values along the minimum joint duration dimension for 9-month infants and their parents, and tested whether joint attention across various timescales predicted future vocabulary size.

## Methods

## Participants

26 parent-infant dyads participated ( 15 female and 11 male). The mean age of infants was 9.21 months ( $S D=0.23$ ). Parent reports of vocabulary were collected three and six months later when the infants were 12 months and 15 months.

## Stimuli

Six toys (car, cup, and train; duck, plane, and boat), organized into two sets of three were used. Each toy in the two sets had a unique uniform color (red, blue, green).

## Stimuli

Parents and their infants sat across from each other at a table $(61 \mathrm{~cm} \times 91 \mathrm{~cm} \times 64 \mathrm{~cm})$. The infants sat in a custom highchair and the parents sat on the floor. Both infants and parents wore head-mounted eye trackers (positive science, LLC). The head-mounted eye-tracking system includes two cameras: (1) An infrared camera that is placed just below and is pointed to the right eye that records eye images, and (2) A scene camera that is placed low on the forehead and is pointed outwards captures the user's first-person view ( $90^{\circ}$ visual field). Each eye tracking system recorded egocentricview video and gaze direction ( $\mathrm{x}-$, y -coordinates) in that view, sampled at 30 Hz . Another camera ( 30 Hz ) was mounted above the table and provided a bird's eye view of the dyadic interaction (see Yu \& Smith (2013) for additional technical details).

## Procedure

Parents and infants were fitted with the eye-tracking gear (see Figure 1). Once the eye-tracking gear was securely affixed to the participants, a calibration phase was completed. To collect calibration points for each eyetracker, an experimenter directed the infant's attention toward a toy that was only used for calibration while another experimenter recorded the moment the child attended to the location of the toy. This procedure was repeated 15 times with the calibration toy played in various locations on the tabletop. A similar procedure was used to calibrate the parent's eye tracker. The calibration procedure took approximately five minutes.

Once the calibration phase was complete, an experimenter placed one of the object sets on the table and the first play trial began. During object play, parents were instructed to engage with their infant as they naturally would. After approximately 60 seconds of play, an experimenter swapped out the objects with the second set of objects, and the second trial began. This procedure was repeated and dyads completed up to four trials for a total of six minutes of play. Not all dyads completed the full play session. Twenty-four dyads completed all four trials and two dyads completed three trials, for a total average playtime of five minutes, eight seconds.


Figure 1: (A) A dual eye-tracking set-up. (B) Sample infant and parent ROI streams.

## Data Processing

Eye-tracking software yielded scene camera footage with crosshairs superimposed, this footage was then sampled at a rate of 30 frames per second. Using an in-house coding program, trained coders annotated frame-by-frame the target of gaze. Three regions of interest (ROIs) were defined for the three objects. ROIs were manually coded frame-byframe from a first-person view video. An ROI was annotated when a cross-hair overlapped on any portion of an object or face. To assess reliability, a second coder coded a randomly-selected $10 \%$ of the frames with $95 \%$ agreement.


Figure 2: Operationalizing joint attention bouts using minimum duration and minimum joint duration.

## Joint Attention Parameters

Two parameters were used to determine joint attention: minimum joint duration and minimum duration (see Figure 2). Minimum joint duration is the temporal duration of continuous alignment of parent and infant fixations on a particular object ROI. Minimum duration is the temporal duration of brief looks elsewhere other than the joint attention ROI, e.g., another object, partner's face. Previous research has incorporated this component into various coding schemes, sometimes allowing for brief looks elsewhere, and for other coding schemes, not allowing the flexibility of brief looks. For Study 1 , we keep this parameter fixed at 0 ms to not allow for brief looks
elsewhere. In other words, all joint attention bouts estimated for this study only included simultaneous and continuous fixations from infant and parent. For example, in Figure 2, Bout 1 (blue object) would not be considered a JA bout because there is a brief look from the parent to the infant's face (pink). Bout 2 (red object) would be considered a JA bout because (1) the infant's and parent's fixations were on the same object (red object) for longer than a particular duration set in our parameter exploration and (2) the fixations were continuous with no brief fixations elsewhere.
In Study 2, we manipulated the minimum duration parameter to equal either 0 ms or 300 ms . To return back to the example in Figure 2, when minimum duration equals 300 ms , Bout 1 (blue object) would now be considered a JA bout because the brief look from the parent to the infant's face (pink) is shorter in duration than 300 ms .

## Exploration of Joint Attention Parameter Space

To explore the parameter space of minimum joint duration, we created six different temporal durations for minimum joint duration $(500 \mathrm{~ms}, 1 \mathrm{~s}, 2 \mathrm{~s}, 3 \mathrm{~s}, 5 \mathrm{~s}$, and 10 s$)$. The other parameter we manipulated was minimum duration, and varied this parameter as either 0 ms or 300 ms . In Study 1, we fixed the minimum duration parameter to 0 ms , and explored the minimum joint duration parameter across all six duration. In Study 2, we explored a combination of a subset of the minimum joint duration parameter ( 500 ms and 1 s ) and the minimum duration parameter ( 0 ms and 300 ms ).

## Properties of Joint Attention Bouts

In Study 1, we estimated joint attention bouts across the minimum joint duration parameter space for each of the 26 infant-parent dyads. Thus, for each dyad, we had 6 joint attention streams. In Study 2, we estimated joint attention bouts across the limited minimum joint duration parameter and the minimum duration parameter, equating to 4 joint attention streams across the parameter space combinations. For each joint attention bout stream, we estimated three properties: proportion ( $\%$ of time in joint attention), frequency (rate $/ \mathrm{min}$ ), and average bout duration. To determine how many dyads with at least one joint attention bout in each parameter space value, we calculated a parameter-level measure as the percentage of the sample (26 dyads) that yielded at least one joint attention bout in a particular parameter value. In Study 3, we explored how joint attention proportion estimates across the minimum joint duration parameter space at 9 -months of age predicted vocabulary size at 12 and 15 months of age.

## Vocabulary Size

Infants and parents returned to the laboratory at ages 12and 15 -month to complete the MacArthur-Bates Communicative Development Inventory (Fenson et al., 1994). We used total receptive vocabulary as our measure of vocabulary scores at each age.

## Study 1

We first investigated the properties of joint attention bouts across parameter space. We conducted a linear mixedeffects model (Baayen, Davidson, \& Bates, 2008) to examine the effects of minimum joint duration on the joint attention bout properties (proportion, frequency, and average bout duration). We included dyad membership as a random slope with the maximally permitted random intercept. Because only one dyad had at least one joint attention bout longer than the minimum joint duration parameter at 10 s , we excluded the 10 s duration from the minimum joint duration parameter in subsequent analyses (see Figure 3).


Figure 3: Joint attention properties across minimum joint duration parameters. Error bars reflect 95\% CIs.

For joint attention rate, there was a significant effect of minimum joint duration ( $\beta=-.0009, S E=0.00007, p<.001$ ), suggesting that as minimum joint durations increased, the rate of joint attention bouts (per minute) decreased. When increasing the minimum joint duration parameter from the shortest duration reflecting the micro-level timescale, 500 ms ( $M_{\text {rate }}=4.99, S E_{\text {rate }}=.39$ ), to the frequently-used timescale in previous literature, 3000 ms ( $M_{\text {rate }}=1.07, S E_{\text {rate }}=.11$ ) (see Table 1; Bakeman \& Adamson, 1994; Carpenter et al., 1998; Tomasello \& Todd, 1983; Tomasello \& Farrar, 1986), we observed a $78 \%$ decrease in JA bout rate.

As expected, for mean joint attention duration, there was a significant effect of minimum joint duration ( $\beta=.001, S E=$ $0.00003, p<.001$ ), suggesting that as minimum joint durations increased, the mean duration of joint attention bouts increased.

Similar to what was observed for rate, for joint attention proportion, there was a significant effect of minimum joint
duration ( $\beta=-.0001, S E=0.00002, p<.001$ ), suggesting that as minimum joint durations increased, proportion decreased. When increasing the minimum joint duration parameter from $500 \mathrm{~ms}\left(M_{\text {proportion }}=.16, S E_{\text {proportion }}=.02\right)$ to 3000 ms $\left(M_{\text {proportion }}=.07, S E_{\text {proportion }}=.01\right)$, we observed a $53 \%$ decrease in JA proportion.

Calculating the percentage of dyads with at least one joint attention bout for a particular parameter space value provides a metric of how the sample size changes as a function of parameter value choices. Considering that investigations of joint attention utilize properties of joint attention bouts, we interpret this value below $100 \%$ for a particular combination to be suboptimal for the study of joint attention. Inspection of these estimates yielded some important observations. Estimates decreased as the duration of the minimum joint duration parameter increased. At 2 s , the amount of dyads with at least one JA bout dropped below $100 \%$ and at 5 s , only approximately $50 \%$ of the sample had at least one joint attention bout. This is an important observation because it points to a particular timescale, 2-3s, when the behavior of interest, joint attention, does not occur for some dyads in a sample.

We have established that after a particular timescale, 2-3s, the amount of dyads producing at least one bout of joint attention drops considerably with increases in the minimum joint duration parameter.

## Study 2

To investigate the potential combinatory effects of both minimum joint duration and minimum duration, we estimated JA bouts across the minimum duration parameter ( 0 ms and 300 ms ) and a subset of the minimum joint duration parameter dimension ( 500 ms and 1000 ms ). Our parameter space therefore consisted of 4 possible combinations of the two parameters. We chose these parameters because (1) in Study 1, we observed $100 \%$ of the dyads had at least one JA bout for the 500 ms and 1000 ms minimum joint duration values and (2) 300 ms as a value for the minimum duration parameter has been used in previous research to allow for flexibility in the fluctuations of attention (Yu \& Smith, 2013, 2016).

Consistent with Study 1, for joint attention rate, there was a significant effect of minimum joint duration ( $\beta=-.0004$, $S E=0.00003, p<.001$ ), suggesting that as minimum joint durations increased, rate of joint attention bouts increased. There was no significant effect of minimum duration ( $\beta=.0001, S E=0.0004, p=.76$ ) nor was the interaction significant $\quad(\beta=.0000001, \quad S E=0.0000006, \quad p=.78)$, suggesting, despite allowing for brief attentional flexibility (minimum duration $=300 \mathrm{~ms}$ ) compared to no flexibility (minimum duration $=0 \mathrm{~ms}$ ), joint attention rate remained the same (see Figure 5).

Mean joint attention duration increased as minimum joint duration increased ( $\beta=.002, S E=0.0001, p<.001$ ), but there was no significant effect of minimum duration ( $\beta=-.0002$,
$S E=0.0004, p=.75$ ) nor was the interaction significant ( $\beta=.00000003, S E=0.0000006, p=.56$ ). Mean joint attention proportion decreased as minimum joint duration increased ( $\mathrm{b}=-.0005, S E=0.00004, p<.001$ ), but there was no significant effect of minimum duration ( $\mathrm{b}=.00005, S E=$ $0.0001, p=.74$ ) nor was the interaction significant ( $\mathrm{b}=$ $.00000005, S E=0.0000002, p=.81)$.


Figure 5: Joint attention properties across minimum joint duration and minimum duration parameters. Error bars reflect 95\% CIs.

## Study 3

We analyzed the relationship between joint attention proportion estimates across the minimum joint duration parameter space at 9 -months of age and vocabulary size at 12 and 15 months of age. From the previous analyses, we know that increases in minimum joint duration lead to decreases in the amount of joint attention bouts and also the proportion of joint attention bouts. It is possible that even though joint attention proportion decreases, the overall variability still predicts vocabulary size. Alternatively, it is also possible that decreases in joint attention proportion also reduces the likelihood that the variability of joint attention proportion covaries with later vocabulary size. Table 2 shows the correlation results across parameter space values for joint attention proportion at 9 months and vocabulary size at 12 and 15 months. As observed in previous analyses, the amount of dyads with proportion estimates varies across the parameter space. Therefore, it is important to note the varying degrees of freedom of the correlations across the parameter space values.

There are a few important observations from this analysis. First, the pattern of correlations between joint attention proportion and vocabulary size does not vary across 12 months to 15 months. This suggests that the predictive value of joint attention at 9 months extends into the second year of life. Second, the results suggest that after exceeding a minimum joint duration of 3 s , joint attention is no longer predictive of language development. This is an important
finding because it strengthens the argument that the relevant timescales for joint attention, and subsequent predictive value for language development, include durations shorter than 3 seconds. We will discuss the implications of this result in the Discussion section.

Table 2: Summary of correlation coefficients (degrees of freedom in parentheses) between JA proportion and 12- and

15-month vocabulary size across the Minimum Joint
Duration parameter.

| Minimum Joint Duration | 12 months | 15 months |
| :--- | :--- | :--- |
| 500 ms | $.58(24)^{* * *}$ | $.54(24)^{* * *}$ |
| 1000 ms | $.61(24)^{* * *}$ | $.56(24)^{* * *}$ |
| 2000 ms | $.50(23)^{* * *}$ | $.45(23)^{* * *}$ |
| 3000 ms | $.53(21)^{* *}$ | $.53(21)^{*}$ |
| 5000 ms | $.18(14)$ | $.16(14)$ |
| 10000 ms | - | - |

Note. ${ }^{*} p<.05,{ }^{* *} p<.01,{ }^{* * *} p<.001$.

## Discussion

The present study investigated the relevant timescales for joint attention in infant-parent naturalistic free-play. To answer this question, we explored a parameter space consisting of two frequently used components implemented to operationalize joint attention: minimum joint duration and minimum duration. Across three studies, we observed a collection of important results that provide insight into the consequences of choosing specific parameters for operationalizing joint attention. First, the observation of joint attention behavior drops precipitously when the duration of continuous alignment of parent and infant visual fixations (minimum joint duration parameter) extends longer than $\sim 3 \mathrm{~s}$. Second, allowing for brief fluctuations of attention away from the target object (minimum duration parameter), does not appear to impact the overall properties of joint attention. Third, the predictive value of joint attention for language development reduces in strength when joint attention bouts shorter than 3 s are omitted from analysis.

Perhaps the most important observation from this study was that when we only included joint attention bouts exceeding 3 seconds, the properties of joint attention changed significantly: rate and proportion of joint attention bouts were reduced by $50 \%$ or more. Furthermore, only including joint attention bouts exceeding a 3 -second duration resulted in the loss of predictive value of joint attention for vocabulary size. Taken together, these results suggest that a purely macro-level approach to the study of joint attention can lead to a loss of important variability that captures the phenomenon of joint attention.

We also observed that the inclusion of a parameter that affords brief fluctuations of attention to parts of the visual environment other than the target object does not significantly affect the properties of joint attention. It is
important to point out that we limited this analysis to only a subset of the minimum joint duration parameter in order to include all joint attention bouts longer than 500 ms and 1000 ms . It is possible that the inclusion of macro-level values of minimum joint attention (e.g., 500 ms ) and longer values of minimum duration - extending the duration of fluctuations of attention - would affect the joint attention properties beyond nominal differences. We plan to attend to this question in more detail in subsequent research.

Investigations of joint attention provide unique insights into the development and dynamics of human communication. The present study focused on an important methodological and theoretical question: what are the relevant timescales for joint attention? Our results, generated from a deductive technique to explore different areas of the parameter space of joint attention, suggest that the inclusion of micro-level temporal specifications of joint attention (e.g., <3s) is important for capturing a more vibrant picture of joint attention.

## Acknowledgments

We thank Melissa Elston, Steven Elmlinger, Charlotte Wozniak, Charlene Tay, Seth Foster for collection of the data. This investigation was supported by the National Institutes of Heath, T32 Grant \# HD07475.

## References

Abney, D. H., Warlaumont, A. S., Oller, D. K., Wallot, S., \& Kello, C. T. (2016). Multiple Coordination Patterns in Infant and Adult Vocalizations. Infancy. 1-26.
Akhtar, N., \& Gernsbacher, M. A. (2007). Joint attention and vocabulary development: a critical look. Language and linguistics compass, 1(3), 195-207.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language, 59(4), 390-412.
Bakeman, R., \& Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peerinfant interaction. Child development, 55(4), 1278-1289.
Baldwin, D. A., Moore, C., \& Dunham, P. J. (1995). Understanding the link between joint attention and language. Joint attention: Its origins and role in development, 131-158.
Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others.
Brooks, R., \& Meltzoff, A. N. (2005). The development of gaze following and its relation to language. Developmental science, 8(6), 535-543.
Brooks, R., \& Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. Journal of child language, 35(01), 207220.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., \& Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the society for research in child development, i-174.
Corkum, V., \& Moore, C. (1995). Development of joint visual attention in infants. In Joint attention: Its origins and role in development. Lawrence Erlbaum Associates, Inc.
Feldman, R., Magori-Cohen, R., Galili, G., Singer, M., \& Louzoun, Y. (2011). Mother and infant coordinate heart rhythms through episodes of interaction synchrony. Infant Behavior and Development, 34(4), 569-577.
Hayhoe, M. M., Shrivastava, A., Mruczek, R., \& Pelz, J. B. (2003). Visual memory and motor planning in a natural task. Journal of vision, 3(1), 6-6.
Harder, S., Lange, T., Hansen, G. F., Væver, M., \& Køppe, S. (2015). A longitudinal study of coordination in mother-infant vocal interaction from age 4 to 10 months. Developmental psychology, 51(12), 1778.
Morales, M., Mundy, P., Delgado, C. E., Yale, M., Messinger, D., Neal, R., \& Schwartz, H. K. (2000). Responding to joint attention across the 6-through 24month age period and early language acquisition. Journal of applied developmental psychology, 21(3), 283-298.
Mundy, P., Card, J., \& Fox, N. (2000). EEG correlates of the development of infant joint attention skills. Developmental psychobiology, 36(4), 325.
Mundy, P., \& Newell, L. (2007). Attention, joint attention, and social cognition. Current directions in psychological science, 16(5), 269-274.
Scaife, M., \& Bruner, J. S. (1975). The capacity for joint visual attention in the infant. Nature. 253, 265-266.
Tomasello, M., \& Todd, J. (1983). Joint attention and lexical acquisition style. First language, 4(12), 197-211.
Tomasello, M., \& Farrar, M. J. (1986). Object permanence and relational words: A lexical training study. Journal of Child Language, 13(03), 495-505.
Tomasello, M. (1988). The role of joint attentional processes in early language development. Language Sciences, 10(1), 69-88.
Yu, C., \& Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. PloS one, 8(11), e79659.
Yu, C., \& Smith, L. B. (2016). The social origins of sustained attention in one-year-old human infants. Current Biology, 26(9), 1235-1240
Van Egeren, L. A., Barratt, M. S., \& Roach, M. A. (2001). Mother-infant responsiveness: Timing, mutual regulation, and interactional context. Developmental psychology, 37(5), 684.

# The Role of Letter Frequency on Eye Movements in Sentential Pseudoword Reading 

Cengiz Acartürk<br>(acarturk@metu.edu.tr)<br>Middle East Technical University<br>Cognitive Science Program<br>Ankara, Turkey

Özkan Kılıc<br>(ozkankilic@ybu.edu.tr)<br>Yildirim Beyazit University, Computer Engineering<br>Middle East Technical University, Cognitive Science Ankara, Turkey

Bilal Kırkıcı<br>(bkirkici@metu.edu.tr)<br>Middle East Technical University, Foreign Language Education<br>Ankara, Turkey

Burcu Can<br>(burcucan@cs.hacettepe.edu.tr)<br>Hacettepe University, Computer Engineering<br>Ankara, Turkey

Ayşegül Özkan<br>(oaysegul@metu.edu.tr)<br>Middle East Technical University, Cognitive Science Program Ankara, Turkey


#### Abstract

For a language learner, any new word is a pseudoword. A pseudoword is a string of of letters or phonemes that sounds like an existing word in a language, though it has no meaning in the lexicon. On the other hand, speakers are well aware of permissible phonemes, their frequencies and collocations in their language due to the phonotactics inherent in the language. For example, saktal is a pseudoword in Turkish, whereas szyan is not, due to Turkish phonotactics. This study investigates the relationship between pseudoword letter formation and eye movement characteristics in reading. In particular, we examine the role of Turkish vowel harmony, middle-word consonant collocation, and word-initial and word-final consonants on eye movements with adult native speakers reading sentences that involve predesigned Turkish pseudowords. The results of an experiment with 34 participants are indicative of the role of pseudoword formation on a set of eye movement parameters.


Keywords: Consonant collocations; Eye movements, Pseudowords; Reading; Turkish; Vowel harmony.

## Introduction

If native speakers of English are asked to judge the acceptability of the words, wug, toysion, or craphen, they are likely to rate them as acceptable. Similarly, Turkish speakers may judge talar as a possible Turkish word though they hear it for the first time. These are all pseudowords, which sound like existing words in a language, without semantic content.

Pseudowords are commonly used as experiment stimuli in research in psychology, linguistics, neuroscience, and cognitive science. They are especially useful when researchers aim at overcoming likely effects of semantics in the experiments. Phonological well-formedness of words (Hammond, 2004), morphological productivity (Anshen \& Aronoff, 1988), language development (Dabrowska, 2006), judgment of semantic similarity (MacDonald \& Ramscar, 2001), vowel harmony (Pycha, Novak, Shosted \& Shin, 2003), machine learning for orthography (Testolin,

Stoianov, Sperduti \& Zorfib, 2015), neuroimaging of reading (Mechelli, Gorno-Tempini \& Price, 2003), and dyslexia (Grainger, Bouttevin, Truc, Bastien \& Ziegler, 2003; Houpt, Sussman, Townsend \& Newman, 2015) have been among the major topics that have been studied through the use of pseudowords as experimental stimuli. Pseudowords have been employed to test models of word and letter perception, such as the interactive activation model of context effects in letter perception (McClelland, \& Rumelhart, 1981).

Previous studies have revealed that native speakers seem to make their judgments by using a probable combination of sounds (Hammond, 2004; Shademan, 2007), the cooccurrence of syllables or consonant collocations locally (Hay, Pierrehumbert \& Beckman, 2004), non-locally (Finley, 2012; Frisch \& Zawaydeh, 2001; Koo \& Callahan; 2011), or through nucleus-coda combination probabilities (Treiman, Kessler, Knewasser, Tincoff \& Bowman, 2000). Accordingly, the acceptability judgments of pseudowords are connected to orthographic neighborhood (e.g., Davis, 2010, see Rayner, Pollatsek, Ashby, Clifton Jr., 2012, for a review), which takes into account the relative positions of consonants and vowels (Perea \& Lupker, 2004), as well as letter position coding through adjacent bigrams (Seidenberg \& McClelland, 1989) and open bigrams of letters (Whitney, 2001).

For Turkish, a language with shallow orthography, it has been shown that some of the properties that characterize pseudowords can be captured from a written corpus in parallel to native speakers' judgment (Kilic, 2014). Against this background, the present study first employs corpusbased frequencies for the formation of pseudowords according to a set of criteria, including the relative positions of consonants and vowels in pseudowords. Then we investigate pseudoword processing by locating the pseudowords in sentential contexts and having native speakers read them in sentential contexts. The underlying approach is that we capture eye movement characteristics as
an indicator of word processing, which may enrich our knowledge about word processing by providing data that go beyond response time (cf. lexical decision tasks). In other words, we propose that sentential pseudoword reading, as we name it, has the potential to contribute to our understanding of word processing through the study of the relationship between subword characteristics of pseudowords and eye movement parameters.

Three dominant experimental factors that have been found to influence eye movement characteristics in reading are word length, frequency and sentential predictability (Rayner, Sereno \& Raney, 1996; Rayner, 1998; Rayner, Pollatsek, Ashby, Clifton Jr., 2012). In the present study, we controlled word length by forming pseudowords of six letters. Using pseudowords allowed us to ignore sentential predictability and word-level frequency as factors for experimental control. Instead, we focused on subword frequencies that are specified by the combinations of the letters that form six-letter pseudowords. We studied three specific types of combinations by designing high-frequency and low-frequency letter bigrams for experimental purposes:

- Vowel combinations that follow vowel harmony
- Middle-word consonant collocation
- Word-beginning and word-end consonants

Silent reading is under the influence of inner speech. If a pseudoword is hard to speak out loud, it is also hard to read silently. In this study, we hypothesize that these three variables will affect pseudoword reading in Turkish as high frequency ensures easy pronunciation, whereas low frequency of these variables lead to difficult pronunciation. It is also expected that adjacent dependencies will have a strong influence on pseudoword reading. In the following section, we present an overview of these characteristics of word structure in Turkish.

## The Turkish Language

Turkish is an agglutinating language, read from left-to-right, with a considerably shallow orthography using 8 vowels and 21 consonants derived from the Roman Alphabet (Göksel \& Kerslake, 2005; Lewis, 2000). The description of Turkish word structure depends on morphophonological constraints. The continuation of a morpheme and the selection of the corresponding morph are determined by the preceding morph. The final vowel in the preceding morph affects the form of the vowel in the incoming morpheme (Turkish vowel harmony). Similarly, the final consonant in the most recent morph causes some changes on the first consonant of the next morpheme (Assimilation).

While a morpheme with a vowel is concatenated to a string in Turkish, its vowel is modified with respect to the roundedness and backness properties of the most recent vowel in the string as in (1).
at-lar
horse-Plu
horses
kedi-ler
cat-Plu
cats
okul-lar school-Plu schools

This is not an immediate dependency because diphthongs and consecutive vowel collocations are usually not allowed in Turkish. When the corpus frequencies are investigated, it is seen that the frequencies of words containing $a \ldots a$ or $e \ldots i$ as substrings are higher than the frequencies for the words with $l \ldots \ddot{u}$ or $o \ldots e$. In other words, frequencies mimic vowel harmony. Similarly, some immediate consonant collocations are more frequent than others due to assimilation as shown in (2) for the ablative case marker -DAn.

$$
\begin{array}{lll}
\text { ev-den } & \text { et-ten } & \text { yatak-tan }  \tag{2}\\
\text { house-Abl } & \text { meat-Abl } & \text { bed-Abl } \\
\text { from house } & \text { from meat } & \text { from bed }
\end{array}
$$

Accordingly, some consonant collocations, such as $v d$, $t t$, and $k t$, are more frequent than $v t$, $t d$, and $k d$. These are infrequent but not zero frequencies because there are exceptions to assimilation as well as vowel harmony.

Another salient aspect of Turkish word structure is word boundary. Some letters, e.g., $k, g, d, z$ and $y$, are observed more frequently than, e.g., $c, g, r, v$, and $f$ in word-initial or word-final position. Even Turkish words in root forms mostly follow the regularities briefly exemplified above. In this study, word-initial and word-final boundaries are treated as a single variable (either low or high frequency letters at both boundaries) to keep the stimuli size manageable and the experiment duration reasonable.

The present study examines the effects of the aforementioned three major aspect of word formation in Turkish, namely the frequency of vowel combinations that follow vowel harmony (henceforth, vowel harmony collocation), the frequency of middle-word consonant collocations (henceforth, consonant collocation), and the frequency of word-beginning and word-end consonants (henceforth, word boundary collocation) on eye movement characteristics in sentential reading. These three variables are assumed to be binary variables with either high frequency or low frequency in a $2 \times 2 \times 2$ design. A total of 80 pseudowords were created as experimental stimuli. The letter bigram frequencies were obtained from the METU Turkish Corpus (Say, Zeyrek, Oflazer \& Ozge, 2002). All pseudowords had the same template of vowels (V) and consonants (C) shown in (3).

## (3) $\mathrm{C}_{1} \mathrm{~V}_{1} \mathrm{C}_{2} \mathrm{C}_{3} \mathrm{~V}_{2} \mathrm{C}_{4}$

Low-frequency or high-frequency vowel collocations, i.e. $\mathrm{V}_{1} \mathrm{~V}_{2}$, were selected, where high-frequency pairs of $\mathrm{V}_{1} \mathrm{~V}_{2}$ followed the rules of Turkish vowel harmony. Similarly, high-frequency and low-frequency pairs were selected for the consonant collocation $\mathrm{C}_{2} \mathrm{C}_{3}$, and for the word boundary
collocation $\mathrm{C}_{1} \mathrm{C}_{4}$. Below, we present the experiment, its methods, materials and the results obtained.

## Experiment

Thirty-four university students at Middle East Technical University (METU), Turkey, participated in the experiment (18 female and 16 male university students, mean age: 23.59, SD: 5.04). The participants were asked to read silently 80 single-line sentences displayed separately on a computer screen while their eye-movements were recorded by an EyeLink 1000 Hz desktop eye tracker. Each sentence included one pseudoword, which was located around the middle of the display. The stimuli were shown in random order. Simple true/false questions were asked randomly after some screens to ensure active engagement of the reader with the task. 20 random and simple true/false questions, which were about words other than the pseudowords in the sentences, were asked as well in order to ensure that the participants had been actively reading. The experiment was conducted in single sessions. Each session took approximately 40 minutes. The participants were paid 25 TL (approximately 6\$) as an incentive for participation.

## Material

A total of 80 pseudowords of the form $\mathrm{C}_{1} \mathrm{~V}_{1} \mathrm{C}_{2} \mathrm{C}_{3} \mathrm{~V}_{2} \mathrm{C}_{4}$ were formed for the eight categories (cf. $2 \times 2 \times 2$ design of highand low-frequency bigrams) by using orthographic frequencies obtained from the METU Turkish Corpus. Frequency was specified as a binary variable, with either high or low values. For example, if the frequency of a consonant bigram is above the mean of the frequencies of all possible bigrams, it was assumed to be a high consonant collocation frequency for selecting $\mathrm{C}_{2} \mathrm{C}_{3}$. Otherwise, it was assumed to be a low frequency consonant collocation. Similarly, high-frequency and low-frequency bigrams were selected for the word boundary collocations $\mathrm{C}_{1} \mathrm{C}_{4}$, and for the vowel harmony collocations $\mathrm{V}_{1} \mathrm{~V}_{2}$. We did not distinguish between onset and offset frequencies while studying the bigrams in the corpus to keep the experiment design simple. Below, we present more detail about the three types of collocations.

Turkish Vowel Harmony Collocation ( $\left.V_{1} \mathbf{V}_{2}\right)$. Since Turkish has eight vowels, $8 \times 8$ vowel bigrams were produced and their frequencies were calculated from the METU corpus. The vowels might have zero or more intervening characters in between. For example, while $a \ldots a$ is a very frequent vowel substring, $l \ldots \ddot{u}$ is an unlikely one. We calculated all available vowel bigrams. The highfrequency bigrams represented Turkish vowel harmony, whereas the ones with low frequency were due to the words that were exceptions to vowel harmony.

Consonant Collocation ( $\mathbf{C}_{2} \mathbf{C}_{3}$ ). Turkish has 21 consonants. Therefore, $21 \times 21$ bigrams from $b b$ to $z z$ were produced by calculating their bigram frequencies from the corpus. For example, $m l$ is much above the average of all possible
bigram frequencies while $f v$ is very rare in the corpus. Accordingly, high-frequency consonant collocations and low-frequency consonant collocations were identified for forming the pseudoword consonant collocations of the form $\mathrm{C}_{2} \mathrm{C}_{3}$. In contrast to vowel harmony collocations, only adjacent bigrams were calculated for consonant collocation since $\mathrm{C}_{2} \mathrm{C}_{3}$ is an adjacent-bigram collocation.

Word Boundary Collocation ( $\mathbf{C}_{1} \mathbf{C}_{4}$ ). The frequency values of the $21 \times 21$ bigrams, from $b \ldots b$ to $z \ldots z$, were produced from the 21 consonants in Turkish and their word boundary frequencies were calculated from the corpus. Since $C_{1} C_{4}$ is not an adjacent-bigram collocation, we allowed intervening characters between the first character of words and the last character when calculating the bigram frequencies. For example, $k$ is a frequent word-initial boundary while $z$ is a frequent word-final boundary in Turkish, making $k \ldots . z$ a frequent word boundary pattern, whereas $f \ldots b$ is a very rare word boundary co-occurrence in Turkish.

Since each of the three independent variables can be either high or low frequency, eight ( $2 \times 2 \times 2$ ) groups, each of which had ten representative pseudowords, were formed. Eight samples from the 80 -pseudoword set are shown in Table 1.

Table 1: Pseudoword groups and representative samples

|  | $\mathbf{V}_{\mathbf{1}} \mathbf{V}_{\mathbf{2}}:$ Low |  | $\mathbf{V}_{\mathbf{1}} \mathbf{V}_{\mathbf{2}}:$ High |  |
| :--- | :---: | :---: | :---: | :---: |
|  | $\mathbf{C}_{\mathbf{2}} \mathbf{C}_{3}:$ | $\mathbf{C}_{\mathbf{2}} \mathbf{C}_{\mathbf{3}}:$ | $\mathbf{C}_{\mathbf{2}} \mathbf{C}_{\mathbf{3}}:$ | $\mathbf{C}_{\mathbf{2}} \mathbf{C}_{\mathbf{3}}:$ |
|  | Low | $\mathbf{H i g h}$ | $\mathbf{L o w}$ | $\mathbf{H i g h}$ |
| $\mathbf{C}_{\mathbf{1}} \mathbf{C}_{\mathbf{4}}:$ Low | vöfvac | nindüd | lagşav | remliv |
| $\mathbf{C}_{\mathbf{1}} \mathbf{C}_{\mathbf{4}}:$ High | töjkur | köndız | hupşar | kadraz |

$\mathrm{V}_{1} \mathrm{~V}_{2}$ : Vowel harmony frequency
$\mathrm{C}_{2} \mathrm{C}_{3}$ : Consonant collocation frequency
$\mathrm{C}_{1} \mathrm{C}_{4}$ : Word boundary frequency
The pseudowords were located around the center of 80 meaningful and different sentences of similar length (63 to 65 characters). For avoiding likely effects of the orthographic features of the neighbor words, the pre-target word was fixed in the sentences: The word "aslinda" (English 'actually'), which is a very frequent Turkish adverb was consistently used as the single pre-target word. For preventing parafoveal information intake from the posttarget word, we used the frequent Turkish bigram bi- such that all the post-target words started with this same bigram (cf. Lima \& Inhoff, 1985). A sample sentence is shown in (4), where GEN is the genitive marker, PLU is the plural marker, NEG is the negative marker, PASS is the passive marker, and PAST is the past tense marker in Turkish. The target pseudoword is underlined.
(4) Siz-in mektuplar aslında köndız bitmez dendiği için Your-GEN letter-PLU indeed köndız end-NEG say-PASS for dağıtılmamış.
distribute-PAST-NEG.

In (4), the target word is the pseudoword köndız. It has a low-frequency vowel harmony collocation $\mathrm{V}_{1} \mathrm{~V}_{2}$, in this case $\ddot{o} l$; a high-frequency consonant collocation $\mathrm{C}_{2} \mathrm{C}_{3} n d$; and a high-frequency boundary collocation $\mathrm{C}_{1} \mathrm{C}_{4} \mathrm{kz}$. (sample sentences are provided in the Appendix).

## Results \& Discussion

We calculated a set of eye movement measures, as listed below.

- First fixation duration is the duration of the first fixation on the first-pass reading of the target pseudoword.
- First pass gaze duration is the sum of individual fixation durations in the first-pass reading of the target pseudoword. In other words, the entire word is an area-of-interest and this value represents the total time spent by a participant before his/her gaze left the word for the first time.
- First pass fixation count is the sum of individual fixations in the first-pass reading of the target pseudoword. The first-pass reading covers all the fixations on the word without leaving it. If the eye shifts from a pseudoword to another word on its right, the first pass is over. If the eye shifts from a pseudoword to another word on its left, the first pass is over, as well.
- Regression in count is the number of regression fixations that return to the target pseudoword after the first-pass reading. If the gaze re-fixates on a previously fixated word, but this time from its right, this is called Regression-in.

The mean values for those eye movement parameters, for the eight pseudoword categories are shown in Table 2.

Table 2: Mean values for eye movement parameters on target pseudowords

|  | $\mathrm{V}_{1} \mathrm{~V}_{2}$ : Low |  | $\mathrm{V}_{1} \mathrm{~V}_{2}$ : High |  |
| :---: | :---: | :---: | :---: | :---: |
|  | $\begin{aligned} & \mathrm{C}_{2} \mathrm{C}_{3} \text { : } \\ & \text { Low } \\ & \hline \end{aligned}$ | $\begin{aligned} & \mathrm{C}_{2} \mathrm{C}_{3}: \\ & \mathrm{High} \\ & \hline \end{aligned}$ | $\begin{aligned} & \mathrm{C}_{2} \mathrm{C}_{3}: \\ & \text { Low } \\ & \hline \end{aligned}$ | $\begin{aligned} & \mathrm{C}_{2} \mathrm{C}_{3}: \\ & \text { High } \\ & \hline \end{aligned}$ |
|  | First Fixation Duration (ms) |  |  |  |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : Low | 265.97 | 253.67 | 259.58 | 253.72 |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : High | 259.74 | 247.01 | 274.45 | 258.06 |
|  | First Pass Gaze Duration (ms) |  |  |  |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : Low | 467.97 | 435.13 | 454.60 | 390.80 |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : High | 426.58 | 360.19 | 431.32 | 331.75 |
|  | First Pass Fixation Count |  |  |  |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : Low | 1.86 | 1.86 | 1.84 | 1.66 |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : High | 1.76 | 1.57 | 1.71 | 1.40 |
|  | Regression in Count |  |  |  |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : Low | 1.27 | 1.29 | 1.18 | 1.22 |
| $\mathrm{C}_{1} \mathrm{C}_{4}$ : High | 1.23 | 1.28 | 1.28 | 1.34 |

We also calculated the averages for eye movement measures for the six-letter, legitimate words of the form CVCCVC in the sentences (excluding the sentence-initial word, the sentence-end words, and post-target words). We found nine words as such, in the stimuli sentence set. Table 3 shows the mean values, which can be taken as a baseline for comparison with the pseudoword values.

Table 3: Mean values for eye movement parameters on legitimate words

| First Fixation Duration (ms) | 201.65 |
| :--- | :--- |
| First Pass Gaze Duration (ms) | 237.33 |
| First Pass Fixation Count | 1.20 |
| Regression in Count | 1.20 |

The data in Table 3 are not sufficiently representative due to the limited number of words. On the other hand, the values in Table 3 are close to the average values obtained for sixletter words in another study with a richer data set (Acartürk, et al., in preparation). Therefore, we believe that these values can be conceived as close to representative mean values.

## Analyses

Four three-way ANOVAs were run on the data from 34 participants to examine the role of vowel harmony collocation, consonant collocation and word boundary collocation on the following eye movement measures in sentence reading: First Fixation Duration, First Pass Dwelling Time, First Pass Fixation Count, and Regression in Count. No significant three-way interaction was observed between the variables and these measures $(F(1,33)=.266$, $p=.61, \quad \eta^{2}=.001 ; \quad F(1,33)=.002, \quad p=.97, \quad \eta^{2}=.000$; $F(1,33)=.208, p=.65, \eta^{2}=.006 ;$ and $F(1,33)=.000, p=.99$, $\eta^{2}=.000$ respectively). However, there were significant twoway interactions and single effects, as presented below.

First Fixation Duration. Consonant collocation frequency had a significant effect on the first fixation duration, $F(1,33)=8.428, p<.05, \eta^{2}=.20$. As the frequency of the collocated consonant bigram decreased, the average duration of the first fixation on the pseudoword significantly increased, and vice versa.

First Pass Gaze Duration. A statistically significant interaction was found between consonant collocation frequency and word boundary collocation frequency, $F(1,33)=4.608, p<.05, \eta^{2}=.12$. In other words, the effect of consonant collocation is greater in the high-frequency wordboundary collocation condition than in the low-frequency word-boundary collocation condition, and the effect of word-boundary collocation is greater in the high-frequency consonant collocation condition than in the low-frequency consonant collocation condition.

First Pass Fixation Count. There was a statistically significant interaction between consonant collocation frequency and word boundary collocation frequency, $F(1,33)=4.341, \quad p<.05, \quad \eta^{2}=.12$, similar to the finding obtained for the gaze duration. This means that if the frequency of consonant collocation is high, then word boundary collocation frequency has an effect on the first pass fixation count, such that if the word boundary collocation frequency changes from high to low, first pass fixation count increases. If both the consonant collocation frequency and the word boundary collocation frequency change, the effect is stronger.

Regression in Count. A statistically significant interaction was obtained between vowel harmony collocation frequency and word boundary collocation frequency, $F(1,34)=5.002$, $p<.05, \eta 2=.13$, without an interaction with consonant collocation frequency. In particular, when the consonant collocation frequency is high, pseudowords with high word boundary collocation frequency gets more re-fixations from its right if the vowel harmony collocation frequency is also high. This effect disappears if the dominant variable, the consonant collocation, becomes infrequent.

## Discussion and Conclusion

Our findings revealed mixed results. This indicates the need for further research on pseudoword reading in Turkish. In particular, the findings for the first fixation durations showed that the consonant collocation frequency is the dominant aspect that influences eye movements in reading, since low frequency consonant collocations result in longer fixation durations and vice versa. The frequency of consonant collocations also significantly interacts with the frequency of word boundary collocations, according to the findings obtained for the first pass fixation count and first pass gaze duration, showing that when they both have low frequency, a higher number of first pass fixation count and a longer first pass gaze duration is observed. Finally, the regression-in-count findings suggest that when the consonant collocation frequency is already high, regressionin counts increase when the vowel harmony collocation frequency and the frequency of word boundary collocations are also high. It is likely that this is because pseudowords of this type look like real words. This might have caused the participants to assume misreading a known word, which was followed by re-fixations after the first pass.

In reading research, the first fixation duration on a word is usually conceived as the most valuable indicator of word recognition, since it can be seen as a reflection of the initialstage processes in word recognition (Rayner, 1998). The second fixation on a word (viz. refixation), further fixations on the same word and regression fixations reveal more complex processes than word recognition, such as syntactic and semantic processes at a sentential-level. Accordingly, at this stage, our findings indicate the frequency of the middleword consonant allocation as a dominant factor that
influences pseudoword reading in Turkish. We also believe that the sentential pseudoword reading paradigm has the potential to enrich our knowledge about word processing with higher ecological validity compared to alternative approaches, such as lexical decision tasks and naming tasks.

## Acknowledgments

This study has been partially supported by TUBITAK (The Scientific and Technological Research Council of Turkey) 113K723, EU H2020 ISCH COST Action IS1404 Evolution of reading in the age of digitisation (E-READ), and EU FP7 Marie Curie Actions IRIS, 610986, FP7-PEOPLE-2013IAPP. The authors thank Zuhal Ormanoğlu and Irmak Olcaysoy Ökten for their contribution.

## References

Acartürk, C., Kırkıcı, B., Beken, F. (in preparation). Eye movement characteristics in Turkish reading: A corpusanalytic study.
Anshen, F., \& Aronoff, M. (1988). Producing morphologically complex words. Linguistics, 26, 641655.

Dabrowska, E. (2006). Low-level schemas or general rules? The role of diminutives in the acquisition of Polish case inflections. Language Sciences, 28, 120-135.
Davis, C. J. (2010). The spatial coding model of visual word identification. Psychological Review, 117(3), 713.
Finley, S. (2012). Testing the limits of long-distance learning: learning beyond a three-segment window. Cognitive Science, 36, 740-756.
Frisch, S. A., \& Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. Language, 77, 91-106.
Göksel, A., \& Kerslake, C. (2005). Turkish: A Comprehensive Grammar. Routledge: London and New York.
Grainger, J., Bouttevin, S., Truc, C., Bastien, M., \& Ziegler, J. (2003), Word superiority, pseudoword superiority, and learning to read: A comparison of dyslexic and normal readers. Brain and Language, 87, 432-440.
Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. International Journal of English Studies, 4, 1-24.
Hay, J., Pierrehumbert, J., \& Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. In: J. Local, R. Ogden, and R. Temple (Eds.), Phonetic Interpretation: Papersbin Laboratory Phonology VI. Cambridge: Cambridge University Press.
Houpt, J. W., Sussman, B. L., Townsend, J. T., \& Newman, S. D. (2015). Dyslexia and configural perception of character sequences. Frontiers in Psychology, 6, 482.
Kilic, O. (2014). Using corpus statistics to evaluate nonce words. In M. Colinet, S. Katrenko, R. K. Rendsvig (Ed.s), Pristine Perspectives on Logic, Language, and Computation: ESSLLI 2012 and ESSLLI 2013 Student Sessions, Selected Papers (pp. 26-35). Berlin, Heidelberg: Springer Berlin Heidelberg.

Koo, H., \& Callahan, L. (2011). Tier-adjacency is not a necessary condition for learning phonotactic dependencies. Language and Cognitive Processes, 77, 18.

Lewis, G. (2000). Turkish Grammar, Second edition. Oxford: University Press (2000)
Lima, S. D., \& Inhoff, A. W. (1985). Lexical access during eye fixations in reading: effects of word-initial letter sequence. Journal of Experimental Psychology: Human Perception and Performance, 11(3), 272.
MacDonald, S., \& Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. Proceedings of the 23rd Annual Conference of the Cognitive Science Society (611-616), University of Edinburgh.
Machelli, A., Gorno-Tempini, M. L., \& Price, C. J. (2003). Neuroimaging studies of word and pseudoword reading: Consistencies, inconsistencies, and limitations. Journal of Cognitive Neuroscience, 15(2), 260-271.
McClelland, J. L., \& Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological Review, 88(5), 375.
Perea, M., \& Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. Journal of Memory and Language, 51(2), 231-246.
Pycha, A., Novak, P., Shosted, R., \& Shin, E. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In G. Garding and M. Tsujimura (Eds.), Proceedings of the WCCFL 22, (423435).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372-422.
Rayner, K., Pollatsek, A., Ashby, J., \& Clifton Jr, C. (2012). Psychology of reading ( $2^{\text {nd }}$ ed.) Psychology Press.
Rayner, K., Sereno, S. C., \& Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. Journal of Experimental Psychology: Human Perception and Performance, 22, 1188-1200.
Say, B., Zeyrek, D., Oflazer, K., \& Ozge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. Proceedings of the Eleventh International Conference of Turkish Linguistics.
Seidenberg, M. S., \& McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological Review, 96(4), 523.
Shademan, S. (2007). From clusters to words: Grammatical models of nonce word acceptability. Grammar and analogy in phonotactic well-formedness Judgments. $P h D$ Dissertation, University of California, Los Angeles (2007)

Testolin, A., Stoianov, I., Sperduti, A., \& Zorfib, M. (2015). Learning orthographic structure with sequential generative neural networks. Cognitive Science, 1-28.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., \& Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In: M. B. Broe and J. Pierrehumbert (Eds.), Papers in Laboratory Phonology V: Acquisition and the Lexicon (pp. 269-282). Cambridge: Cambridge University Press.
Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. Psychonomic Bulletin \& Review, 8(2), 221-243.

## Appendix

Below we present sample stimuli for each combination of vowel harmony collocation $\mathrm{V}_{1} \mathrm{~V}_{2}$, middle-word consonant collocation $\mathrm{C}_{2} \mathrm{C}_{3}$, and boundary word collocation $\mathrm{C}_{1} \mathrm{C}_{4}$, respectively. H is used for high frequency, and L is used for low frequency. Pseudowords are underlined for demonstration.

## $V_{1} V_{2}$ is high, $C_{2} C_{3}$ is high, $C_{1} C_{4}$ is high

- Kaldırmaya çabalasan aslında biylen birden yere düşüp kırılmazdı.
- Mirası harcamasaydım aslında sirden binası şimdi çoktan bitmişti.
$V_{1} V_{2}$ is high, $C_{2} C_{3}$ is high, $C_{1} C_{4}$ is low
- Hasan cebindeki parayı aslında niyled birimi olarak düşünüyordu.
- Son yılın modası aslında lirdev birası ile karides pişirmekmiş.
$\mathbf{V}_{1} V_{2}$ is low, $C_{2} C_{3}$ is high, $C_{1} C_{4}$ is high
- Sizin mektuplar aslında köndız bitmez dendiği için dağıtılmamış.
- Dün Emine'nin nişanlısı aslında bındül binayı bulamadı gerçekten.
$V_{1} V_{2}$ is low, $C_{2} C_{3}$ is high, $C_{1} C_{4}$ is low
- Radyonun sesini açarsan aslında nındüd bizden sonuçları öğrenir.
- Şu kediler sokakta aslında ü̈lrıv birini bulup sürtünüyorlarmış.
$\mathrm{V}_{1} \mathrm{~V}_{2}$ is high, $\mathrm{C}_{2} \mathrm{C}_{3}$ is low, $\mathrm{C}_{1} \mathrm{C}_{4}$ is high
- Yarın deniz kenarında aslında bıgvıl birini izleyerek eğlenecek.
- Baharda sigara içmek aslında sagșan binası çevresinde yasaktır.
$V_{1} V_{2}$ is high, $C_{2} C_{3}$ is low, $C_{1} C_{4}$ is low
- Bu bölgedeki kuşların aslında lagşav birini korkuttuğu söylenir.
- Fizik dersi haricinde aslında revşev bilimi konusundan bahsetti.
$V_{1} V_{2}$ is low, $C_{2} C_{3}$ is low, $C_{1} C_{4}$ is high
- İstatistik dersinde aslında söcșun birden bütün verileri bozmuş.
- Hastaneye sabah gelenler aslında töjkır birini görmek istiyorlar.
$V_{1} V_{2}$ is low, $C_{2} C_{3}$ is low, $C_{1} C_{4}$ is low
- Doğa resimlerinde aslında löcșuv binası betimlemelerini kullanır.
- Cuma gecesi televizyonda aslında şobçüc birini izleyerek uyudum.


# Children's familiarity preference in self-directed study improves recognition memory 

Katherine A. Adams*1, George Kachergis*2,3, Douglas Markant ${ }^{4}$<br>${ }^{1}$ Department of Applied Psychology, New York University<br>${ }^{2}$ Department of Artificial Intelligence, Radboud University<br>${ }^{3}$ Donders Institute for Brain, Cognition and Behavior<br>${ }^{4}$ Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin<br>*equal contribution


#### Abstract

In both adults and school-age children, volitional control over the presentation of stimuli during study leads to enhanced recognition memory. Yet little is known about how very young learners choose to allocate their time and attention during self-directed study. Using a recognition memory task, we investigate self-directed study in low-income preschoolers, who are at an age when attention, memory, and executive function skills rapidly develop and learning strategies emerge. By pre-exposing children to some items before self-directed study, we aimed to discover how familiarity modulates their study strategies. We found that children showed a preference for studying pre-exposed items. Overall, items studied longer led to increased recognition of those items at test. We also compared recognition task performance and strategies with measures of cognitive control skills, finding that children's selective attention skills support recognition performance. These findings may inform both theory and educational intervention.


Keywords: active learning; recognition memory; executive function; attention; cognitive development

## Introduction

Children learn through active exploration of their environments. They ask questions, test hypotheses, and probe novel or confounding objects that could shed new light on how the world works (Schulz \& Bonawitz, 2007). Recent research in cognitive science suggests that schoolage children learn better when allowed to control the content and timing of information flow compared to passively receiving information (Partridge, McGovern, Yung, \& Kidd, 2015; Sim, Tanner, Alpert, \& Xu, 2015). Little is known, however, about how self-directed information gathering develops during preschool ages, a time of great plasticity in the neural networks that support executive function, attention, and memory (Blair \& Raver, 2015). Understanding patterns in young children's active information gathering and examining the mechanisms through which self-directed control affects learning may inform cognitive science as well as educational initiatives, particularly for low-income preschoolers at higher risk of poor learning outcomes (Ursache, Blair, \& Raver, 2012).

Episodic memory is one ability that has been found to benefit from active learning. Memory is aided by topdown, meta-cognitive control processes, such as when learners prioritize study of items close to mastery and
avoid content that is already learned or that is too difficult to master (Markant, Ruggeri, Gureckis, \& Xu, 2016). Bottom-up influences of cognitive control can also support episodic memory. In adult recognition memory tasks with self-paced study, alignment of stimulus exposure with attentional resources improved later recognition (Markant, DuBrow, Davachi, \& Gureckis, 2014). These active control behaviors enhance representations and strengthen associative networks, both of which help to encode and retrieve experienced stimuli (Markant et al., 2016).

Voss and colleagues (Voss, Galvan, \& Gonsalves, 2011; Voss, Gonsalves, Federmeier, Tranel, \& Cohen, 2011) examined how adults' study patterns influence the benefits of active encoding for recognition memory. They tasked participants with memorizing a set of objects arranged in a $5 \times 5$ grid. A moving window allowed only one object to be visible at a time, with control over the window given to the participant during active blocks. During yoked blocks, participants watched the window move according to the recorded movements of a previous participant. Importantly, the yoked condition allowed the authors to distinguish the effects of active control over and above the visual stimulus information experienced during study. They found both an overall active study advantage, as well as benefits to particular study patterns. Recognition improved when objects were studied for longer duration and revisited within a short time frame, but the benefits of these study features were only found in active and not yoked conditions.

Ruggeri, Markant, Gureckis, and Xu (2016) adapted the Voss et al. paradigm to examine study patterns during active encoding with school-age children. They found that 6 -to 8 -year-olds had better recognition memory when given volitional control over the presentation of stimuli during study, as compared to being yoked to study sequences generated by other children. Moreover, the recognition memory advantages of self-directed study were present following a one-week delay. In contrast to Voss et al.'s findings in adults, school-age children showed generalized benefits of certain study patterns on their memory encoding: participants had improved recognition memory in both active and passive conditions for items visited often and studied longer (Ruggeri
et al., 2016). The authors suggest that children benefited in both conditions from attentional cues (red outlines that indicated which object would be presented for study next). For school-age children, attentional cueing appears to support benefits of longer study even in the yoked condition. This finding is consistent with previous research showing that even subtle opportunities to coordinate the learner's attentional state to incoming information (i.e., by giving learners control over when the next stimulus appears) can improve episodic memory (Markant et al., 2014).

These studies suggest that multiple levels of control may enhance memory from an early age, but the developmental course of these processes remains unclear. One possibility is that the effects of active encoding vary based on the maturation of the neural networks that support volitional control, working memory, and attention. These neural networks undergo tremendous growth during the preschool years, leading to meaningful individual differences in children's attention and self-regulatory control (Blair \& Raver, 2015). These cognitive control skills support school readiness, and are targeted for intervention to close income-based early achievement gaps (Ursache et al., 2012). Little is known about the effects of active encoding on recognition memory at preschool ages. Young children's variability in attentional control may make study duration and attentional coordination particularly critical factors for active encoding. Developing cognitive control skills may also affect children's metacognitive ability to strategically allocate study effort based on their current familiarity with the materials.

This study examines whether low-income preschoolers use active control to engage in strategic study during a recognition memory task. If so, what patterns of sampling emerge, and how do these patterns change with varied stimulus familiarity? Another gap in the research literature is whether individual differences in young children's control skills influence the effects of active sampling on encoding. This study addresses these questions by examining active memory performance in a large sample of low-income preschoolers. We use a novel extension of the Ruggeri et al. (2016) task design that varies stimuli pre-exposure, as well as a battery of well-validated executive function and attention measures.

## Experiment

## Methods

Participants Ninety-four 5-year-olds from low-income backgrounds were recruited and tested as part of a school readiness study run in collaboration with two Head Start preschool centers. An additional 16 participants were tested but excluded due to incomplete data due to experimenter error or connectivity problems, or because of difficulty understanding task instructions. Children were tested in their preschools by trained assessors. Administration of the tasks was divided over two testing
days scheduled within one week of each other. EF and attention tasks were administered on day 1 , and lasted about 5 minutes each for a total of 15 minutes. The recognition memory task was administered on day 2 and lasted about 10 minutes.

## Memory Task

Materials Stimuli were taken from Ruggeri et al. (2016), which included 149 color line drawings of animals and objects that are used frequently by children younger than 5 -years-old in everyday conversations (MacWhinney \& Snow, 1985). Items were randomly sampled from the stimulus set and presented in a series of 4 x 3 grids. Stimuli not presented during the practice or study phases were randomly sampled in the test phase and used as novel foils. The task was presented on a touchscreen laptop, with timing and choice data logged to a database via psiTurk (Gureckis et al., 2015).

Procedure The task was presented as a simple memory card game (see Figure 1). Children were instructed to study a grid of images on the touchscreen tablet, presented initially "face-down" as empty rectangles. Children could "turn cards over" by touching the empty rectangle to reveal the image underneath. Later, they were asked to recognize studied items presented among novel distracter images. The design and procedure closely followed described in Ruggeri et al. (2016), with a key design modification: this version experimentally manipulates the pre-exposure of items during the study phase in order to examine the role of exposure on children's active study behavior.

Practice phase. Children were presented with a 2 x 2 practice grid. Half the items on the grid were simultaneously revealed during a pre-exposure phase that lasted 6 s while the other half remained face down, and children were instructed to "Remember these pictures!" The practice study phase ( 30 s ) followed, with all cards presented face down. Children were told to tap the cards they wanted to see. Once a card was touched, the image underneath was revealed until the child "tapped" off by touching the image again, or touched another card. Only one item was revealed at a time. Next the untimed test phase presented a $3 \times 2$ grid showing all 4 items in the study phase as well as 2 additional novel distracter items, in random grid locations. Children were instructed to touch all the "old" pictures they saw before, and not touch the "new" pictures. A red box appeared around each item when selected, and was toggled off if tapped again. Children were not restricted in how many or few items could be selected during the test phase. Once children indicated that their selections were complete, the assessor praised correct answers and gave feedback on incorrect answers. The practice phase could be played 1 to 3 times with different stimuli. If the child was unable
to understand directions, the task was ended.
Study phase. The study phase consisted of 3 blocks, each presenting a $4 x 3$ grid of randomly sampled images. The procedure is similar to that described in the practice phase. Simultaneous pre-exposure of half the items lasted 2 s per item ( 12 s in the two half pre-exposed blocks, and 24 s in the all pre-exposed block) before cards were turned over and the child could then actively select and turn over cards to study for 36 s . This shorter duration of the active study phase (as compared to Ruggeri et al., 2016) was chosen to enhance the potential effect of pre-exposure on search behavior. The 3 study grids were presented consecutively before the test phase.

Test phase. The test phase consisted of 6 blocks. Each $4 x 3$ test grid was a random sample drawn without replacement from a pool of 72 stimuli, including the 36 included in the study phase along with 36 novel images. The number of old items in each grid ranged from 0 to 12 (randomly chosen) in order to minimize strategic responding based on the proportion of items selected within each block. All 36 studied stimuli and all 36 novel stimuli were presented only once at test. Children were instructed to "Touch the pictures you remember!" and not to select new pictures. Once the child indicated that they were done with selection, the assessor prompted. "Are you sure you touched only the pictures you saw before and not any new pictures?" If the child said yes, the assessor advanced to the next test grid. If the child answered no, the assessor reminded them to choose only "old" pictures seen before.

Both hit rate (proportion of studied items correctly selected as "old") and correct rejection rate (proportion of novel items correctly not selected) were calculated. In addition, total study time per item and study repetitions per item was computed for pre-exposed vs. non-preexposed items and conditions.
Executive Function and Attention Tasks Attention Network Test. The Attention Network Test (ANT; Rueda et al., 2004) is a well-known behavioral measure thought to map onto the neural networks supporting attentional control. The child version of the ANT presents either a single fish or a horizontal row of five fish. Children are instructed to feed the center fish by pressing a blue box in the lower corners of either side of the screen indicating in which direction the central fish is swimming. Children are asked to ignore the flanker fish pointing either in the same (congruent) or opposite direction (incongruent) as the target middle fish. Mean accuracy and reaction time are computed.

Visual Search Task. The Visual Search task (Steele, Karmiloff-Smith, Cornish, \& Scerif, 2012) measures the ability to select relevant stimuli (targets) while ignoring distracters (non-targets). Children are presented with a search display on the touch screen monitor. Each display contains 90 items, made up of 20 targets (animals) and


Figure 1: Each study phase of the experiment was preceded by pre-exposure of half (6) or all (12) of the items, for 2 s per item (i.e., 12 s in the two half-pre-exposed conditions and 24 s in the all-pre-exposed condition). All three cycles of pre-exposure and study were completed before the six screens of testing were performed.

70 non-targets (objects). Children are instructed to find animals, which are replaced with a star when successfully touched. The task ends when a total of 18 correct responses is reached, or 40 responses are made overall. Mean search speed (time between touches), and number of errors are recorded.

Continuous Performance Test. The Continuous Performance Test (CPT; (Steele et al., 2012) measures the ability to sustain attention for a prolonged period without distraction. In this version, the child is instructed to touch the screen as soon as an animal appears. One hundred pictures are randomly presented one at a time, including 20 presentations of the target stimuli (animals) and 80 presentations of nontarget stimuli (objects). Each stimulus appears on the screen for 300 ms followed by a blank screen for 1250 ms . In addition to response time, number of missed responses to targets (omission error) and incorrect touches to distracters (commission error) are recorded.

Digit Span. Digit Span is a widely used executive function task that assesses children's working memory. Children are instructed to repeat number sequences of sequentially longer length in forward and backward conditions. Total number of correct responses per condition is recorded. Children in this sample were largely unable to repeat sequences backwards, so only performance on the forward condition are used here.

## Results

Data from 94 participants were analyzed with respect to recognition (selection) of studied items (i.e., hit rate), correct rejection of unstudied items, and the number of
repetitions and total study time for studied items. Participants' mean hit rate (HR) was 0.65 , and the mean correct rejection (CR) rate was 0.56 .
Study Behavior Studied items were selected for study on average 1.78 times (median: 1 ; maximum: 10). The mean study time for old items was 3037 ms (median: 2067 ms ). Table 1 shows the distribution of how many times children repeated study items and cumulative study time per item (median, mean, and SD). Children most often studied items a single time (38.6\%), but it was not uncommon to study an item twice ( $22.6 \%$ ) or even three times ( $10.6 \%$ ). A surprising number of items ( $23.4 \%$ ) were not actively selected for study at all, and these were well-distributed among the participants, who left a median of 6 of the 36 items unstudied (mean 8.3, bootstrapped $95 \%$ confidence intervals: $(6.8,10.1)$ ).

| Reps | Median Time | Mean Time | SD | N |
| ---: | ---: | ---: | ---: | :--- |
| 0 | 0 ms | 0 ms | 0 ms | 783 |
| 1 | 1,452 | 2,443 | 3,137 | 1,307 |
| 2 | 2,305 | 3,371 | 3,331 | 766 |
| 3 | 2,860 | 3,528 | 2,413 | 360 |
| 4 | 3,645 | 4,491 | 2,891 | 121 |
| 5 | 5,069 | 6,622 | 5,028 | 37 |
| $>5$ | 5,467 | 6,984 | 5,552 | 8 |

Table 1: Statistics of study repetitions and time (ms).

Pre-exposure Effects To investigate the impact of pre-exposure on study time and repetitions, we fit mixedeffects regression models to separately predict triallevel study time and study repetitions (both scaled and centered to $[-1,1]$ ) for only the blocks with half preexposed items. Subject was included as a random factor, and item pre-exposure as a binary predictor ( R syntax: Study Time $\sim$ Preexp + (1|Subject) and Repetitions $\sim$ Preexp + (1|Subject). Shown in Table 2 , the regression predicting study time found a significant positive intercept ( $\beta=0.19, Z=5.02, p<.001$ ). Moreover, there was a significant positive effect of preexposure ( $\beta=0.11, Z=2.23, p<.05$ ), indicating that pre-exposure led to increased study time. ${ }^{1}$ On average, pre-exposed items in these conditions were studied for 3477 ms , whereas the hidden items were studied for 3001 ms . Shown in Table 3, the regression predicting study repetitions found a significant positive intercept ( $\beta=0.32, Z=5.58, p<.001$ ). There was a positivelytrending effect of pre-exposure $(\beta=0.07, Z=1.88$, $p=.06$ ), suggesting that pre-exposed items may be selected more often for study. On average, pre-exposed items in these conditions were selected 1.80 times, while the hidden items were selected 1.65 times.

[^259]|  | $\beta$ | SE | $Z$-score | $p$-value |
| :--- | ---: | ---: | ---: | ---: |
| Intercept | 0.185 | 0.037 | 5.016 | $p<.001^{* *}$ |
| Pre-exposed | 0.112 | 0.050 | 2.233 | $p<0.05^{*}$ |

Table 2: Regression predicting study time.

|  | $\beta$ | SE | $Z$-score | $p$-value |
| :--- | ---: | ---: | ---: | ---: |
| Intercept | 0.315 | 0.057 | 5.578 | $p<.001^{* * *}$ |
| Pre-exposed | 0.070 | 0.037 | 1.884 | $p=0.06$. |

Table 3: Regression predicting study repetitions.

Recognition Accuracy To investigate the impact of pre-exposure, repetitions, and study time on recognition performance, we fit two logistic mixed-effects regression models to the item-level accuracy data for old stimuli, separating study time and repetitions since they are correlated. Subject was included as a random factor, and study repetitions and study time (scaled and centered to $[-1,1])$ were included in their respective models as fixed, continuous predictors, allowed to interact with item preexposure, a binary predictor ( R syntax for study time model: Correct ~ Preexp * Time + (1|Subject); and substitute Reps for Time in the other model).

In the study repetitions model, there was a significant positive intercept, showing that participants were more likely to correctly recognize rather than miss the old items ( $\beta=0.95, Z=5.16, p<.001$ ). There was a significant positive effect of study repetitions $(\beta=0.52$, $Z=4.19, p<.001$ ), showing that studying items more often led to higher recognition of those items. There was also a significant positive effect of pre-exposure ( $\beta=0.27, Z=2.37, p=.02$ ), showing that pre-exposure increased the likelihood of correctly recognizing an old item. Finally, there was a significant negative interaction of pre-exposure and repetitions $(\beta=-0.31, Z=2.23$, $p=.03)$ : with pre-exposure, there was less accuracy benefit of more study repetitions. Figure 2 shows the mean hit rate as a function of pre-exposure and study repetitions, along with the relative frequency of each level of repetitions.

In the study time model, in addition to a significant positive intercept, $(\beta=1.05, Z=5.83, p<.001)$, there was a significant positive effect of study time ( $\beta=0.36, Z=3.19, p=.001$ ), showing that studying items longer led to increased recognition of those items. There was also a positively-trending effect of preexposure ( $\beta=0.21, Z=1.92, p=.06$ ), suggesting that pre-exposure may increase their chance of recognizing old items. Finally, there was a significant negative interaction of pre-exposure and repetitions $(\beta=-0.29$, $Z=2.32, p=.02$ ), showing that pre-exposure lessens the accuracy benefit of longer study time.

The AIC of the study repetition model was 2712.8, and the AIC of the study time model was 2722.6, making the


Figure 2: Items that were studied more often had higher hit rates, but most items were not studied more than one or two times. (Not pictured: participants were at chance for unstudied 'old' items.)
relative likelihood of the study time model 0.007 . Thus, although both models have similar interpretations, the repetitions model provides a better account of the data.

Self-directed Memory and Executive Function We next examined the link between behavior in the selfdirected memory task and the various attention and executive function (EF) measures using three mixed-effects regression models to predict item-level ( $\mathrm{N}=2,770$ ) 1) recognition accuracy, 2) study time, and 3) study repetitions for old items. All three models included subject as a random factor, and the following EF measures (scaled and centered to $[-1,1]$ ) as fixed predictors: working memory, visual search errors, visual search reaction time, commission errors, and omission errors, and ANT accuracy and RT (R syntax: Correct ~ + EFvar1 + EFvar2 + . $+(1 \mid$ Subject $)$ ).
For the logistic model predicting recognition accuracy, besides a significant positive intercept ( $\beta=0.77$, $Z=4.35, p<.001)$, there was a negatively-trending coefficient for visual search errors ( $\beta=-0.41, Z=-1.92$, $p=.05$ ). All other predictors were insignificant ( $p$ 's $>.1$ ). In summary, this suggests that fewer visual search errors, an index for selective attention skills, is associated with increased recognition.

The model predicting study time (log-transformed, scaled and centered to $[-1,1]$ ) found a positive coefficient for visual search time ( $\beta=0.05, Z=2.11, p=.03$ ), with all other predictors insignificant ( $p$ 's $>.1$ ). This indicates that participants with longer visual search times also spent longer studying items during the study phase.

The model predicting study repetitions (with a Poisson linking function) found no significant predictors in the EF measures.

## Discussion

The present study examined low-income preschool children's study behavior in a self-directed recognition memory task, and compared 5 -year-olds' active study behaviors to patterns found in older samples in previous literature. We next examined if stimuli pre-exposure affects active encoding. Finally, we explored how individual differences in executive function and attention skills may influence study strategies and recognition.

First, we found that children were above-chance at recognizing old items and correctly rejecting new items, indicating that they can meaningfully engage in a developmentally-complex paradigm requiring selfdirected study. We found increased recognition accuracy for items with greater repetitions, and for items with greater study time, replicating classic repetition effects from both traditional (experimenter-directed) recognition memory experiments, as well as self-directed versions (e.g., Voss et al., 2011a; Voss et al., 2011b).

Second, we found that pre-exposure significantly increased study time, suggesting a preference to allocate study effort to familiar material at the outset of study. Pre-exposed items were also more likely to be recognized, but this effect appeared to overlap with other helpful study behaviors. For pre-exposed items, both repetitions and study time showed less benefit to recognition compared to items without pre-exposure. Thus, although children use their familiarity with items to guide their study, they appeared to benefit more generally from stimulus exposure, be it through passive pre-exposure or active selection (i.e., increased study time or repetitions). Ruggeri et al. (2016), finding similar results for 6 - to 8 -year-olds who had better recognition memory in both active and yoked conditions for items visited often and studied longer, suggested that children in yoked conditions were able to benefit from attentional cueing, allowing them to coordinate their attention with the presentation of new information. Notably, we found that preschoolers benefited from passive pre-exposure, which provides no attentional cueing. These findings suggest that duration of stimuli exposure alone may be particularly important for memory encoding at preschool ages.

Third, we found that greater recognition accuracy was predicted by both fewer visual search errors and longer visual search response times in a developmental selective attention task. These data suggest that selective attention skills support children's active study during preschool, a period of neurocognitive plasticity in systems that support attention, executive function, and memory (Blair \& Raver, 2015). While it may be surprising that longer visual search time supports recognition memory, it is important to note that these behav-
ioral measures often exhibit a speed-accuracy trade off (Davidson, Amso, Anderson, \& Diamond, 2006). Young children who search more carefully may be slower to respond but more successful in encoding stimulus information. The relation between stimuli pre-exposure and increased study time suggests that one possible study strategy for children is to focus attention on familiar items. As attentional focus is a more effortful and limited resource at this young age, children may benefit from allocating study time to known items. Prioritizing study of items close to mastery is a learning strategy described in Metcalfe's zone of proximal development framework (Metcalfe, 2011). In this framework, optimal learning strategies should focus on the easiest possible as-yetunlearned items, as focus on items too difficult may be maladaptive and potentially disheartening. In this task, the difficulty of unexposed items to encode is unknown until they are "turned over" and revealed, whereas young children have time during pre-exposure to evaluate preexposed item difficulty and engage attentional resources. Continued experimental investigation is needed to better understand the role of attention skills and search strategies on young children's active encoding.

This study is a first step in examining the effects of executive function and attention on low-income preschool children's active learning. We found that selective attention supports recognition memory, but measures of inhibitory control and working memory were not significant unique predictors. One possibility is that demands of the recognition memory task were particularly dependent on visual search and attentional focus skills. Future experimental studies should aim to tease apart how various cognitive control skills might contribute to different types of active learning tasks. A limitation to this study is that the narrow range of socio-economic status (SES) for our sample may limit generalizability of the findings. Notably, long-term exposure to chronic stress associated with poverty has been found to have negative consequences on children's selective attention and memory (McEwen, 2000). Thus, examining mechanisms that support active encoding may be particularly important for understanding the effects of poverty on early learning. We are planning additional data collection with a higher income cohort to examine relations between SES, cognitive control skills, and active encoding. Future work may also seek not only to measure children's self-directed study strategies, but to improve them via intervention.

## Acknowledgments

We thank Todd Gureckis and C. Cybele Raver for their support and guidance on this study.

## References

Blair, C., \& Raver, C. C. (2015). School readiness and selfregulation: A developmental psychobiological approach. Annual Review of Psychology, 66, 711-731.
Davidson, M. C., Amso, D., Anderson, L. C., \& Diamond, A. (2006). Development of cognitive control and executive
functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia, 44 (11), 2037-2078.
Gureckis, T. M., Martin, J., McDonnell, J. V., Rich, A. S., Markant, D. B., Coenen, A., ... Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. Behavior Research Methods, 1-14.
MacWhinney, B., \& Snow, C. (1985). The child language data exchange system. Journal of child language, 12(02), 271-295.
Markant, D. B., DuBrow, S., Davachi, L., \& Gureckis, T. M. (2014). Deconstructing the effect of self-directed study on episodic memory. Memory and Cognition, 42, 1211-1224.
Markant, D. B., Ruggeri, A., Gureckis, T. M., \& Xu, F. (2016). Enhanced memory as a common effect of active learning. Mind, Brain, and Education, 10(3), 142-152.
McEwen, B. S. (2000). The neurobiology of stress: from serendipity to clinical relevance. Brain research, 886(1), 172-189.
Metcalfe, J. (2011). Desirable difficulties and studying in the region of proximal learning. Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork, 259-276.
Partridge, E., McGovern, M. G., Yung, A., \& Kidd, C. (2015). Young children's self-directed information gathering on touchscreens. In R. Dale et al. (Eds.), Proc. of CogSci 37. Austin, TX: Cog. Sci. Society.
Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., \& Posner, M. I. (2004). Development of attentional networks in childhood. Neuropsychologia, 42(8), 1029-1040.
Ruggeri, A., Markant, D. B., Gureckis, T. M., \& Xu, F. (2016). Active control of study leads to improved recognition memory in children. In A. Papafragou, D. Grodner, D. Mirman, \& J. Trueswell (Eds.), Proc. of CogSci 38. Austin, TX: Cognitive Science Society.
Schulz, L., \& Bonawitz, E. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. Developmental Psychology, 43(4), 1045-1050.
Sim, Z. L., Tanner, M., Alpert, N. Y., \& Xu, F. (2015). Children learn better when they select their own data. In R. Dale. et al. (Eds.), Proc. of CogSci 37. Austin, TX: Cognitive Science Society.
Steele, A., Karmiloff-Smith, A., Cornish, K., \& Scerif, G. (2012). The multiple subfunctions of attention: Differential developmental gateways to literacy and numeracy. Child Development, 83(6), 2028-2041.
Ursache, A., Blair, C., \& Raver, C. C. (2012). The promotion of self-regulation as a means of enhancing school readiness and early achievement in children at risk for school failure. Child Development Perspectives, 6(2), 122-128.
Voss, J., Galvan, A., \& Gonsalves, B. (2011). Cortical regions recruited for complex active-learning strategies and action planning exhibit rapid reactivation during memory retrieval. Neuropsychologia, 49, 3956-3966.
Voss, J., Gonsalves, B., Federmeier, K., Tranel, D., \& Cohen, N. (2011). Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. Nature Neuroscience, 14, 115-120.

# From Words to Sentences \& Back: Characterizing Context-dependent Meaning Representations in the Brain 

Nora Aguirre-Celis (naguirre@cs.utexas.edu)<br>ITESM, Monterrey, Mexico \&<br>Department of Computer Science, 2317 Speedway<br>Austin, TX 78712 USA

Manuel Valenzuela-Rendon (valenzuela@itesm.mx)<br>ITESM, Monterrey, Mexico

Risto Miikkulainen (risto@ cs.utexas.edu)
Department of Computer Science, 2317 Speedway
Austin, TX 78712 USA


#### Abstract

Recent Machine Learning systems in vision and language processing have drawn attention to single-word vector spaces, where concepts are represented by a set of basic features or attributes based on textual and perceptual input. However, such representations are still shallow and fall short from symbol grounding. In contrast, Grounded Cognition theories such as CAR (Concept Attribute Representation; Binder et al., 2009) provide an intrinsic analysis of word meaning in terms of sensory, motor, spatial, temporal, affective and social features, as well as a mapping to corresponding brain networks. Building on this theory, this research aims to understand an intriguing effect of grounding, i.e. how word meaning changes depending on context. CAR representations of words are mapped to fMRI images of subjects reading different sentences, and the contributions of each word determined through Multiple Linear Regression and the FGREP nonlinear neural network. As a result, the FGREP model in particular identifies significant changes on the CARs for the same word used in different sentences, thus supporting the hypothesis that context adapts the meaning of words in the brain. In future work, such context-modified word vectors could be used as representations for a natural language processing system, making it more effective and robust.


Keywords: Neural Networks; FGREP; Concept Attribute Representation theory; fMRI; Context; Meaning; Semantics; Embodied Cognition

## Introduction

Recently, Deep Learning systems of vision and natural language processing (NLP) have drawn special attention into single-word vector spaces. They are able to extract low level features in order to recognize concepts (e.g. cat), but they are incapable of forming an abstract notion of the concept (symbol). In general, these models build semantic representations from text corpora where words that appear in the same context are likely to have similar meanings (Harris, 1970; Landauer \& Dumais, 1997, Burgess, 1998; Baroni et. al., 2010). However, such representations lack intrinsic meaning, which means sometimes even different concepts may appear similar. This problem has driven
researchers to develop new componential approaches, where concepts are represented by a set of basic features, or attributes, based on textual and perceptual input. (Bruni, et al., 2012; Silberer \& Lapata, 2014, Vinyals et. al., 2015). However, even with their multimodal embedding space, such vector representations fall short from symbol grounding.

In contrast, embodiment theories of knowledge representation (Regier, 1996; Landau et al., 1998, Barsalou, 2008) provide a direct analysis in terms of sensory, motor, spatial, temporal, affective, and social phenomena. Further, these theories can be mapped to brain networks. Recent fMRI studies helped identify a distributed large-scale network of sensory association, multimodal and cognitive regulator systems linked with the storage and retrieval of conceptual information (Binder et al., 2009). This network was then used as a basis for Concept Attribute Representation (CAR) theory, an embodiment theory that enumerates semantic features of concepts and grounds them in brain networks (Binder et al., 2009, 2011 and 2016).

An intriguing challenge to such theories is that concepts are dynamic, i.e. word meaning depends on context and recent experience (Pecher, Zeelenberg, \& Barsalou, 2004). For example, a pianist would invoke different aspects of the word piano depending on whether he will be playing in a concert or moving the piano. When thinking about a coming performance, the emphasis will be on the piano's function, including sound and fine hand movements. When moving the piano, the emphasis will be on shape, size, weight and other larger limb movements.

This paper focuses on addressing these challenges based on the CAR theory. The main idea is that different attributes in CARs can be weighted differently depending on context, i.e. according to how important each attribute is in that context. More specifically, neutral CARs of words are first used to form an expected fMRI pattern of a subject reading a sentence. That pattern is compared to an actual fMRI image. Two techniques, multiple linear regression and a FGREP neural network, are then used to determine how the CARs would have to change to account for the actual fMRI
pattern. These changes represent the weighting in context; it is thus possible to track the dynamic meanings of words by tracking how the weighting changes across contexts.
Experiments with available fMRI data show that the approach is feasible, demonstrating meaningful differences for e.g. human communication vs. noise from a machine; dangerous storm vs. dangerous person; live mouse vs. dead mouse. These changes are principled and could be captured e.g. by a neural network. It might then be possible to create them dynamically, and form as a basis for a more robust and grounded natural language processing system.

The CAR theory is first reviewed below, and the sentence fMRI and word representation data described. The methods for determining semantic changes, i.e. multiple linear regression and FGREP, are then presented, followed by an analysis of the results.


Figure 1: Perceptual Grounding. CARs are composed of a list of known modalities that relate to specialized sensory, motor and affective brain processes, systems processing spatial, temporal, and casual information, and areas involved in social cognition. They capture aspects of experience central to the acquisition of abstract and concrete event as well as object concepts.

## Concept Attribute Representation Theory

CARs represent the basic components of meaning defined in terms of known neural processes and brain systems (Binder, 2016). They relate semantic content to systematic modulation in neuroimaging activity. And are therefore not limited to the classical sensory-motor dimensions of most embodied theories.

CARs are composed of a list of well-known modalities that correspond to specialized sensory, motor and affective brain processes, systems processing spatial, temporal, and casual information, and areas involved in social cognition. They capture aspects of experience central to the acquisition of event and object concepts (both abstract and concrete).

These attributes were selected after an extensive body of physiological evidence based on two assumptions: (1) all aspects of mental experience can contribute to concept acquisition and consequently concept composition; (2) experiential phenomena are grounded on neural processors representing a particular aspect of experience (Figure 1).
These aspects of mental experience model each word as a collection of a 66 -dimensional feature vector that captures the strength of association between each neural attribute and
the word meaning. An example is shown in Figure 2. For a more detailed account of the attribute selection and definition see Binder et al. (2009, 2011 and 2016).

## Data Collection and Preprocessing

Two existing data sets were used in this study: fMRI images of sentences and CARs obtained via Mechanical Turk.

## Neural Images

The stimuli shown to subjects consisted of a list of 240 every day written sentences prepared in the Knowledge Representation in Neural Systems (KRNS) project (Glasgow et al., 2016). The sentences are composed by three to nine words from a set of 242 words ( 141 nouns, 39 adjectives and 62 verbs). Eleven subjects took part in this experiment producing 12 repetitions each. Participants viewed the sentences word by word while in the scanner. The data was acquired by the Center for Imagining Research of the Medical College of Wisconsin (Anderson et al., 2016). The fMRI data was preprocessed and transformed into a single sentence fMRI representation per participant (by averaging all the repetitions), with a final selection of 396 voxels per sentence on a scale from 0.2-0.8, for further use in the computational models.


Figure 2: Bar plot for CAR 66 semantic features. The attribute ratings represent the basic features of chair. Given that this concept is an object, gets low weightings on human-related attributes: face, speech, social, and emotion and strong on visual, shape, touch, manipulation, and some others.

## Semantic Vectors

The semantic attribute ratings were collected thru Amazon Mechanical Turk for each of the 242 words (e.g. family, hospital, chair, small, green, laughed, listened, walked). In a scale of $0 . .6$, the participants were asked to assign the degree to which a given concept is associated to a specific type of neural component of experience (e.g. "To what degree do you think of a chair as having a fixed location, as on a map?"). Approximately 30 ratings (all attributes for each word) were collected. After averaging all the ratings and removing outliers, the final attributes were transformed to unit length yielding a collection of 66-dimensional feature vector that captures the weights of association between each neural attribute and the 242 words. Note that in this manner, the richness and complexity of representations is based on intrinsic meaning of each word, and not on word co-occurrence (Figure 2).

## Data Preparation

The data set did not include fMRI images for words in isolation, a technique developed by Anderson et al. (2016) was adopted to approximate them. The voxel values for words were obtained by averaging all fMRI images for the sentence where each word occurred. Thus, the vectors include a combination of examples of that word along with other words that appear in the same sentence (context). Because of the limited number of combinations, some of these became identical, and were excluded from the dataset.

Given the final set of 237 sentences and 236 words (138 nouns, 38 adjectives and 60 verbs), the next step was to identify pairs of sentences with differences on word meanings such as live mouse vs. dead mouse, good soldier vs. soldier fighting, built hospital vs. damaged hospital, and playing soccer vs. watching soccer. This collection will allow the computational models to evaluate distinctive attribute representations and consequently adjust the baseline meaning of a word to convey the effects of context and conceptual combination.

A collection of 77 such sentences, with different shades of meaning for verbs, nouns and adjectives, as well as different contexts for nouns and adjectives was assembled. This collection will be used as Words of Interest (WoI) for the analysis of context in the experiments (Table 1).

Table 1: Contrasting Sentences. Eight sentences from the collection of the 77 contrasting sentences. Here, for instance, the verb kicked is used in two different contexts, playing with a ball (as in Soccer) vs. breaking the door (as an aggressive behavior).

| SEMANTIC <br> CONTRAST | SENTENCES |
| :--- | :--- |
| GOOD | 94 The soldier delivered the medicine during the flood. |
| AGGRESSIVE | 112 The soldier kicked the door. |
| PLAY <br> (SOCCER) | 239 The artist kicked the football. <br> 62 The boy kicked the stone along the street. |
| BREAK | 112 The soldier kicked the door. |
| BAD PEOPLE | 119 The dangerous criminal stole the television. <br> 152 The mob was dangerous. |
| NATURE | 99 The flood was dangerous. |

## Computational Models

A new technique is proposed in this section for analyzing data imaging. It is grounded on the CAR theory and implemented using Multiple Linear Regression (LReg) and the FGREP neural network (Forming Global Representations with Extended BP; Miikkulainen \& Dyer, 1991). The main idea is to predict sentence fMRI by mapping CARWord to SynthWord (fMRI) (top of Figure 3). The SynthWord is then combined by averaging to form SyntSent for the predicted sentence. Next, the SynthSent is compared to the actual fMRISent (middle of Figure 3). The differences are included by modifying the SynthWord that map to fMRISent and by modifying the CARWord that map to the modified SynthWord (bottom of Figure 3). The resulting CARWord indicate how word meaning change across sentences.


Figure 3: General System framework and data flow. Mapping CARWord to SynthWord (top). Then SynthWord is combined by averaging to form SyntSent and to be compared to the actual fMRISent (middle). Invert the process to modify the CARWords via SynthWord revised (bottom). The Revised CARWord includes different word meaning across sentences.

## Multiple Linear Regression

At the word level, Multiple regression (LReg) is used to learn the mapping between CARWord and SynthWord voxels. The training set has attribute vectors of words as independent variables and the corresponding SynthWord vectors as the dependent variable, predicting one voxel at the time. Similarly, at the sentence level, the training contains assembled sentences (SynthSent) as independent and the corresponding Observed fMRISent as the dependent variable. Once the prediction error is calculated, LReg is inverted (which is possible because it is linear), to determine what the CARWord values should have been to make the error zero.

## Neural Network with FGREP

It is possible that the linear prediction based on LReg is not powerful enough to account for the context effects. Therefore, a nonlinear approach based on neural networks is tested as well. A neural network is trained to map CARWord to SynthWord, which are then averaged (as before) into a prediction of the sentence SynthSent (Figure 4). The prediction error is used (through backpropagation) to train the network.

After training, this network is used to determine how the CARWords should change to eliminate the error. That is, for each sentence, the CARWords are propagated and the error is formed as before, but during backpropagation, the network is no longer changed. Instead, the error is used to change the CARWords themselves (which is the FGREP method---Forming Global Representations through Extended backPropagation; Miikkulainen et al., 1991). This modification can be carried out until the error goes to zero, or no additional change is possible (because the CAR values are already at their max or min limits).


Figure 4: The FGREP model to account for context effects. Propagate CARWord to SynthWord. Compose SynthSent by averaging the words into a prediction of the sentence. Compare SynthSent against Observed fMRISent. Backpropagate the error with FGREP for each sentence, freezing network weights and changing only CARWord. Repeat until error reaches zero.

Training the neural network requires as input the 236 CARWord 66-dimensional vectors (W1, W2, W3) and as target, the equivalent corresponding 396-dimensional SynthWord vector (W'1, W'2, W'3). The network then learns a general mapping of words across all sentences. This mapping is then utilized in the FGREP phase to change the CARWord for each different sentence separately (Figure 4). As the last step, the changes in the semantic attributes are analyzed according to the CAR theory for each affected sentence. At this point, due to scarcity of data this is a manual process verifying that the changes made sense.

## Results

The two approaches LReg and FGREP were evaluated in a preliminary experiment of distinguishing between the different meanings of the verb listened. LReg was found to be inadequate in this task and therefore in two subsequent experiments, focusing on the the adjective dangerous and in the noun mouse only the FGREP approach was used. The analysis was performed on the individual subjects for which the fMRI data in general was most consistent.

## Different contexts for the verb "listened "

Both models were used in this experiment to compare the contrasting meanings of HUMAN COMMUNICATION vs. NOISE FROM A MACHINE for the word listened as expressed in 89: The mayor listened to the voter, 92: The lonely patient listened to the loud television. The left side of Figure 5 shows the results for LReg between the original and transformed CARs. Although the CARs adjusted in all sentences, the changes were small and unprincipled, unable to characterize the difference between human communication versus noise from a machine. In contrast, the outcome for FGREP resulted in context-dependent
changes as shown, for sentences 89 and 92 in the right side of Figure 5.

CARs in Sentence 89 presented salient activations in human-related attributes like Face, and Body, Audition, and Speech, as well as Human, Communication, and Cognition, presumably denoting human verbal interaction. For Sentence 92, high activations on Vision, Bright, Color, Pattern, Large, Shape, Complexity, Touch, Temperature, Weight, Scene, Near, Harm, Unpleasant, Happy, and Angry describe a loud and large object such as a television. These results suggest that the linear mapping that LReg performs is not powerful enough to capture context, but the nonlinear mapping of FGREP is. The following experiments therefore both used the FGREP method for this task.

## Different contexts for the adjective "dangerous"

This experiment compared the contrasting meanings of NATURE vs. BAD PEOPLE for the word "dangerous", as expressed in 98: The flood was dangerous, 118: The dangerous criminal stole the television. Figure 6 shows the differences resulting from the FGREP method. As with the verb listened, context-dependent changes did emerge.

CARs in Sentence 98 present changes on activation for Large, Motion, SOMS attributes Texture and Weight, and event attributes Time, Short, and Caused, reflecting moving water. The attributes Toward, Harm, Unpleasant, and the emotion of Angry, represent the experiential and personal nature of danger. Conversely, Sentence 118 shows high activation for Vision, Complexity, Face, and Speech, because they represent human types and roles such as a criminal. Motor attribute Lower Limb as well as evaluation attributes Benefit, Angry, Disgusted, and Fearful can be associated with a dangerous act by a criminal. The FGREP method, therefore, was largely able to differentiate between


Figure 5: Results for the word listened in two contrasting sentences. LReg (left) did not capture context. All changes were insignificant to characterizing the context-dependent representations. The green line shows the original CARs for comparison. FGREP (right) did grasp context. The CARs for Sentence 89 have increased activations in human-related attributes like Face and Body, Auditory attributes, as well as Human, Communication and Cognition. In contrast, Sentence 92 activations on Vision, Color, Large, Shape, Complexity, Touch Temperature, High sound, and Unpleasant, depict a loud object such as a television.
the contrasting relevant dimensions of dangerous act of nature and humans.

## Different contexts for the noun "mouse"

This experiment compared the contrasting meanings of DEAD vs. ALIVE for the word mouse as expressed in sentences 56: The mouse ran into the forest, 60: The man saw the dead mouse. Figure 7 shows the differences resulting from the FGREP method, which are again systematic and meaningful.


Figure 6: FGREP results for the adjective dangerous across two contrasting sentences. CARs in Sentence 98 changed activation for Large, Motion, Texture and Weight, Time, Short, and Caused, reflecting moving water. The attributes Toward, Harm, Unpleasant, and Angry, represent the experiential nature of danger. Sentence 118 shows high activation for Vision, Complexity, Face, and Speech, because they represent human types and roles. Lower Limb, Benefit, Angry, Disgusted and Fearful can be associated

CARs in Sentence 56 have increased activation for Vision, Motion, Complexity, High, and Sound, possibly suggesting animate properties of the live mouse. Upper Limb, spatial attributes Path and Away, and event attributes Time, Duration, Short, and Consequence, symbolize activity such as running. Emotions of Fearful and Surprised may well be associated with seeing a live mouse. In contrast, Sentence 60 shows increased activation for Temperature, Weight, and Smell, as well as emotions Sad, Angry, Disgusted and Fearful, which may be associated to the dead mouse. These changes indicate different aspects of mouse in two contrasting contexts.

## Discussion and Further Work

The experiments in this paper suggest that different aspects of word meaning are activated in different contexts, and it is possible to see those changes in the corresponding fMRI images. These changes are likely to be nonlinear: The linear mapping approach (regression) tends to muddle them, but a nonlinear mapping (FGREP neural network) can tease them apart.

This result is remarkable considering that the dataset was not originally designed to answer the question of dynamic meaning. In particular, having fMRI images for isolated words available, instead of having to synthesize them, should amplify the observed effects significantly. It should also be possible to include sentences with contrasting contexts systematically, thus increasing the number of possible observations, and making it possible to identify differences in a more comprehensive manner.

With such a larger dataset, it should be possible to characterize changes across multiple sentences. Different kinds of changes may occur in nouns, adjectives, and verbs, and there are likely to be interactions between them. Moreover, the semantic changes can vary from individual to individual. As the first step, only single subjects were analyzed in this paper. In the future, the analysis can be
extended to more subjects, identifying which changes are consistent across subjects, and which ones are more individualistic. For instance, the subject in experiment 3 was Sad that the mouse was dead; another subject could show a different emotion.

After formulating such principles, the next step would be to utilize them in building artificial natural language processing systems. It may be possible to train e.g. a neural network to predict how meaning changes in context. Such a network could be then used as a part of an engineered natural language processing system, dynamically modifying the vector representations for the words to fit the context. Such a system should be more effective and more robust in its inference, and match human behavior better.


Figure 7: FGREP results for the noun mouse across two contrasting sentences. CARs in Sentence 56 increased activation for Vision, Motion, Complexity, High, and Sound, presumably to indicate the animate properties of the live mouse. Upper Limb, Path, Away, Time, Duration, Short, and Consequence, suggest activity such as running. In contrast, Sentence 60 shows increased activation for Temperature, Weight, and Smell, as well as Sad, Angry, Disgusted and Fearful, which can be associated to the dead mouse. These changes indicate different aspects of mouse in two contrasting contexts.

## Conclusion

Concepts are dynamic; their meaning depends on context and recent experience. In this paper, word meaning was represented as a collection of attributes (CARs), grounded in observed brain networks. Multiple Linear Regression analysis and a nonlinear FGREP Neural Network were used to understand how the CARs could change to construct the actual sentence representations seen in fMRI images. Preliminary results suggest that there are indeed systematic changes in CARs, and they make sense in each sentence context. These changes could only be seen in the FGREP analysis, suggesting that they are likely to be nonlinear. In the future, such changes could be characterized more fully and used to make artificial natural language systems sensitive to context.

## Acknowledgments

We would like to thank Jeffery Binder (Medical College of Wisconsin), Rajeev Raizada and Andrew Anderson (University of Rochester), Mario Aguilar and Patrick Connolly (Teledyne Scientific Company) for their work and valuable help regarding this research. This work was supported in part by IARPA-FA8650-14-C-7357 grant.

## References

Anderson, A. J., Binder, J. R., Fernandino, L., Humpries C. J., Conant L. L., Aguilar M., Wang X., Doko, S., Raizada, R. D. (2016). Perdicting Neural activity patterns associated with sentences using neurobiologically motivated model of semantic representation. Cer. Cortex. 1-17. Doi:10.1093/cercor/bhw240.
Baroni, M., Murphi, B., Barbu, E., Poesio, M. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. Cognitive Science, 34(2):222-254.
Barsalou, L. W. (2008). Grounded Cognition. Annual Review of Psychology, 59:617-845.
Binder, J. R., Desai, R. H., Graves, W. W., Conant L. L. (2009). Where is the semantic system? A critical review and metaanalysis of 120 functional neuroimaging studies. Cerebral Cortex, 19:2767-2769.
Binder, J. R., Desai, R. H. (2011). The neurobiology of semantic memory. Trends Cognitive Sci, 15(11):527-536.
Binder, J. R., Conant L. L., Humpries C. J., Fernandino L., Simons S., Aguilar M., Desai R. (2016). Toward a brain-based componential semantic representation. Cog. Neuropsychology, 33:3-4, 130-174.
Binder, J. R. (2016). In defense of abstract conceptual representations. Psychonomic Bulletin \& Review, 23.
Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with HAL model. Behavior Research Methods, Inst. \& Com., 30, 188-198.
Burni, E., Tran, N., Baroni, M. (2014). Multimodal distributional semantics. J. Artif. Intell. R. (JAIR), 49:1-47
Glasgow, K., Roos, M., Haufler, A. J., Chevillet, M., A., Wolmetz, M. (2016). Evaluating semantic models with word-sentence relatedness. arXiv:1603.07253.
Harris, Z. (1970). Distributional Structure. In Papers in Structure and Transformational Linguistics, 775-794.
Landau, B., Smith, L., and Jones, S. (1998). Object Perception and Object Naming in Early Develop. Trends in CosSci 27:19-24.
Landauer, T.K., Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
Miikkulainen, R., Dyer, M., G. (1991). Natural Language Processing with Modular PDP Networks and Distributed Lexicon. Cognitive Science, 15, 343-399.
Pecher, D., Zeelenberg, R., Barsalou, L. W. (2004). Sensorimotors simulations underlie conceptual representations: Modalityspecific effects of prior activation. Psychonomic Bulletin \& Review, 11, 164-167.
Regier, T. (1996). The Human Semantic potential. MIT Press, Cambridge, Massachusetts.
Silberer, C., Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. Proceedings of the $52^{\text {nd }}$ Annual Meeting of the Association for Computational Linguistics, 721-732.
Vinyals, O., Toshev, A., Bengio, S., Erham, D. (2015). Show and Tell: A New Image Caption Generator. arXiv:1506.03134v2.

# The Influence of Pop-Culture on Misattribution of Memory 

Daljit Ahluwalia (Daljit.ahluwalia@rutgers.edu)<br>Pernille Hemmer (Pernille.hemmer@rutgers.edu)<br>Kimele Persaud (Kimele.persaud@rutgers.edu)<br>Department of Psychology, Rutgers University, New Brunswick<br>152 Frelinghuysen Road, Piscataway, NJ 08854


#### Abstract

Social media platforms provide a source for transmitting information that can become widely accepted. However, in this process of transmission, information becomes susceptible to distortion. In this study, we assessed people's semantic (i.e., prior expectations) and recognition memory for pop culture content, as a function of confidence and perceived information source. In Experiment 1, we investigated semantic memory for ubiquitous movie quotes (e.g., the famous Star Wars quote "Luke I am your father"). Notably this quote is incorrect, but we found that a majority of participants accepted these lure quotes as true with high confidence and indicated they had experienced the original source. In Experiment 2, participants viewed the original movie sources before a recognition test of the quotes. We found that while there was some improvement, people still preferred the lure quote with high confidence. We discuss the findings in terms of the strength of people's prior expectations when reconstructing events from memory.


Keywords: Semantic memory, recognition memory, source attribution, pop culture, confidence ratings.

## Introduction

Social media platforms such as Reddit, Facebook, and Twitter are popular outlets to discuss a broad range of topics including people's memories of favorite movies, movie quotes, and scenes. Recently a viral online forum topic focused on people who recalled watching a movie dating back to the '90s called 'Shazaam,' which starred the actor Sinbad playing the role of a genie ("People Claim to Have Seen", 2016). Interestingly, there is no record of a movie called 'Shazaam,' and Sinbad himself denies the movie ever existed. This raises the question: how could so many people feel so strongly about a shared false memory of watching a movie that never was? While misremembering a movie might be harmless, imagine being misinformed about a real-world event-such as falling prey to fake news about politics or terrorist attacks. In this paper, we assess people's semantic (i.e., prior expectations) and recognition memory for pop culture content, as a function of their confidence and perceived information source.

There are a number of cognitive mechanisms that might underpin this behavior. One explanation might be that people are falling victim to "the social contagion of memory", which occurs when one person's recollection of events influences and shapes another person's recollection (Roediger, Meade, \& Bergman, 2001). People integrate the false information with their true source representation. In this case, the contagion becomes exacerbated online as people's personal
recollections are being influenced by the recollections of people from around the world in an instant, thus allowing these false memories to be shared and spread.

Take the game of telephone as an example. It begins with one person whispering a message to the person next to them, and so on. Oftentimes, the last message is significantly different from the original. The transmission for any individual in the chain would become a question of reconstructing the noisy information from memory (Xu \& Griffiths, 2009).

These same factors also have implications for everyday memory such as singing the wrong song lyrics or misquoting common phrases and movie lines. As illustrated by the shared internet memory of 'Shazaam,' many of our memories for ubiquitous and pop culture concepts are not learned through first-hand experience with the original source-such as learning a quote directly from a movie-but rather are the result of information being passed from one person to another. In this case, the information transmitted takes on an abstracted nature.

Such distortion can happen in three different ways: assimilation, leveling, and sharpening (Bartlett, 1932). Assimilation occurs when details of the information are altered in memory to reflect one's own culture. Leveling occurs when information that is deemed non-essential is left out, reflecting a 'gist' representation of the event rather than the details. Sharpening occurs when the order of some details is changed. This illustrates how the reconstructive process can influence the transmission of information, leading to an event being slightly altered with each retelling to reflect the biases of the people participating in the transmission process.

While assimilation, leveling, and sharpening can result unintentionally from either the person transmitting the information or the person encoding the information, it is also possible to intentionally distort the shared information. This has been apparent in many experiments involving eyewitness testimony. Repeated exposure to suggestion, specifically, has been found to alter recall for a previous event (e.g., Loftus \& Palmer, 1974). When shown a video of a car stopped at a stop sign and asked questions containing misinformation, such as "did the car stop at the yield sign," most participants responded in the way suggested by the misleading interview question, rather than what they had witnessed in the video (Loftus, Miller, \& Burns, 1978; Loftus \& Palmer, 1974; Mitchell \& Zaragoza, 1996). It has been suggested that when misleading information is presented, it is introduced into the representation for the event and causes an alteration of that representation (Loftus, Miller, \& Burns, 1978). This
illustrates a reconstructive process similar to that of the telephone game leading to biases in memory.

One possible hypothesis to explain this willingness to accept misinformation as part of the original event is source misattribution, where people incorrectly attribute the source of the memory (e.g., Schacter, 2001). According to the source monitoring framework, participants fall prey to misinformation because they have confused the source of the original information with the source of the misinformation (e.g., Johnson et al., 1993; Lindsay, 1994; Lindsay \& Johnson, 1989).

In addition to confusing the source of information, people tend to express high levels of confidence in their memory, whether it is accurate or not (Bacon, 1979; Mitchell \& Zaragoza, 1996). People are equally confident when reporting false information about an event that they heard from a secondhand source, as when they are correctly recalling the event as they experienced it (Mitchell \& Zaragoza, 1996). Higher confidence has also been reported for more familiar statements (those that have been heard before), regardless of the accuracy of those statements (Hasher, Goldstein, \& Toppino, 1977). This suggests that confidence in memory is not directly related to the accuracy of that memory.

Each of these cases demonstrates how memory is a constructive process, prone to distortions from factors including intrusions from semantic memory, source misattribution, and misinformation. Although there is extensive literature on false memories and source misattribution across a number of domains, one domain that has not been widely studied is memory for ubiquitous concepts, specifically the type that would be prone to influences from popular culture. Therefore, we were interested in evaluating people's semantic memory (i.e. prior expectations) for pop-culture content and how their semantic memory influences recognition memory for this content. We tested semantic and recognition memory for well-known quotes that tend to be misrepresented in popular culture, specifically focusing on how pervasive the misremembering of popular incorrect information is, what source people attribute the information to, and how confident they are in their responses. We compared these ubiquitous quotes to performance on common graphics such as the Apple logo, which are less likely to undergo transmission distortion.

In Experiment 1 we assessed participant's semantic memory and attribution of the source of the information, but we did not test recognition memory given the true source information. In Experiment 2 we investigated recognition memory for the original source of the ubiquitous information, in order to compare it to the semantic memory in Experiment 1. We hypothesized that selections of the incorrect popular lure quote would be made with high confidence, and that the likelihood a selecting the lure quote even after studying the video would be similar to the likelihood of selecting the lure

[^260]quote in Experiment 1. If prior expectations exert a strong influence on memory, it is possible that even after viewing the original source material participants will still recall the misinformation that is pervasive in popular culture.

## Experiment 1

## Method

Participants Sixty-three Rutgers University undergraduate students participated in this study in exchange for course credit.
Materials The stimuli consisted of a brief demographics questionnaire (i.e. age, primary language, major, and media usage), and 273 -alternative forced choice (AFCs) questions: three movie quotes, three famous quotes, four logos (note: the Google logo had six individual test questions, one for each letter ${ }^{1}$, so the total number of logo questions was nine), and 12 distractor questions. See Figure 1 for the set of target stimuli. The 3-AFCs for each question included the correct response, a critical lure (i.e., the quote that has become common usage), and a non-critical lure. We selected the critical lures based on what has circulated on the Internet (e.g. In 2012's "Snow White and the Huntsman," Charlize Theron's character can be heard saying, "Mirror, Mirror, on the wall..."). For the graphics questions, the lures were selected based on one that closely resembled the target and one that did not. Both the demographics questionnaire and the

## Movie Quotes:

Star Wars: "No. I am your father."
Snow White and The Seven Dwarves: "Magic, mirror on the wall, who is the fairest of them all?"
Forrest Gump: "Life was like a box of chocolates."

## Graphics:



Figure 1. Stimuli (movie quotes and graphics) used in Experiments 1 and 2.


Figure 2. First panel: demographic questionnaire; Second panel: sample stimuli.

3-AFC questions were written in Matlab and presented on 23inch Dell monitors
Design Given the types of questions that were asked (e.g. recognition of the Apple logo), participants were instructed to put away all cellphones and other electronic devices before the experiment began. First, participants completed the demographic questionnaire (Figure 2, first panel). Next, participants answered 27 questions, one at a time, which consisted of three parts: 1) participants responded to each question by selecting one of the 3-AFCs from a drop-down menu 2) participants rated their level of confidence in their answer on a seven point Likert scale, with one being not confident at all and seven being very confident 3 ) participants indicated from a drop-down menu if and how they had previously been exposed to the information presented in the question. The options included: "I recently discussed this with a friend", "I have seen the TV show/movie, read the book/play, or heard the song/phrase before", "I have not seen the TV show/movie, read the book/play, or heard the song, but I have seen this referenced elsewhere", "I have never seen or heard of this before", or "Other", which then allowed participants to elaborate on their source of knowledge using the keyboard (Figure 2, second panel). The presentation order of the 27 questions was randomized across participants. On average, it took participants 20 minutes to complete the experiment.

## Results

Because the goal was to compare performance on the ubiquitous movie quotes and graphics between Experiments 1 and 2, only these target questions were used in the analysis. To evaluate whether participants could correctly identify the true quotes/graphics, we computed the response probability for the target, the critical lure, and the non-critical lure for each question (see Figure 3). We found that the preferred response for the movie questions was the critical lure (Star Wars: $95 \%$; Snow White: $95 \%$; Forrest Gump: 92\%). The preferred response for most of the graphic questions was the target (Apple logo: 82\%; American flag: 79\%). However, for the Microsoft logo question, participants were split, with $48 \%$ of participants choosing the target, and $48 \%$ choosing the more closely matching lure.

A series of binomial tests were conducted to assess whether the proportion of correct responses for each question was different from chance. The tests revealed that performance was significantly worse than chance for the movie quote

Table 1: Proportion Correct Compared to Chance

| Question | Exp1 | Exp2 | Chi Square <br> Exp1 \& Exp2 |
| :--- | :--- | :--- | :--- |
| Star Wars | $p<.001^{* *}$ | $p=.08$ | $\chi^{2}=25.04, p<.001^{* *}$ |
| Snow White | $p<.001^{* *}$ | $p=.01^{*}$ | $\chi^{2}=2.39, p=0.12$ |
| Forrest Gump | $p<.001^{* *}$ | $p=.004^{*}$ | $\chi^{2}=4.54, p=0.03^{*}$ |
| Apple Logo | $p<.001^{* *}$ | $p=.08$ | $\chi^{2}=11.22, p<.001^{* *}$ |
| Flag | $p<.001^{* *}$ | $p<.001^{* *}$ | $\chi^{2}=0.09, p=0.77$ |
| Microsoft Logo | $p<.001^{* *}$ | $p=.007^{*}$ | $\chi^{2}=1.34, p=0.25$ |
| ${ }^{*} p<.05 ;{ }^{* *} p<.001$ |  |  |  |



Figure 3: Response probabilities for the target and critical lure for each question in Experiments 1 and 2.
questions and significantly better than chance for the graphics questions (see Table 1).

We then assessed confidence for the preferred response for each question. For the movie related questions, a majority of those who chose the critical lure also responded with high confidence, defined as five or higher (Star Wars: 86\%; Snow White: $92 \%$; Forrest Gump: 77\%). For the graphics questions, a majority of those who chose the target responded with high confidence (Apple logo: 78\%, American flag: 94\%; Microsoft logo, target: $55 \%$, most closely matching lure: $45 \%$ ), see Figure 4. When comparing confidence ratings for targets to critical lures, we found that for most of the movierelated questions, confidence was greater for the critical lure than the target. We report the mean and standard deviation because there were not enough participants who chose the target to conduct a statistical test comparing confidence between target and critical lure (Star Wars: critical lure $M=6.14$, $S D=1.63$, target $M=3.50, S D=3.54$; Snow White, critical lure $M=6.55, S D=0.82$, target $M=3.50, S D=2.12$; Forrest Gump, critical lure $M=5.6, S D=1.6$, no participants chose the target for this question). There was no significant difference in confidence ratings between targets and more closely matching lures for the Apple logo, American flag, and Microsoft logo questions.

For source attribution, we computed the response probability for each possible source conditioned on target and critical lure responses. Here we report the most frequently selected source for the preferred response for each question. Given that a majority of participants responded with the critical lure for the movie related questions, we analyzed which source they attributed their response to and found that they responded "watched the movie" as their direct source of knowledge (Star Wars: 71\%; Snow White: 83\%; Forrest Gump: 84\%), see Figure 5.


Figure 4: Confidence ratings for target and critical lures for movie related and graphic questions in Experiments 1 (top row) and 2 (bottom row).


Figure 5: Source attribution for the movie related questions in Experiment 1 (left column) and Experiment 2 (right column).

## Experiment 2

## Method

Participants Thirty-six Rutgers University undergraduates participated in exchange for course credit.
Materials The demographic questionnaire was identical to Experiment 1. The study stimuli consisted of eight videos corresponding to the movie quotes and graphics, as well as one distractor clip. The video clips were scenes from movies that contained the quotes and commercials that contained the logos that were the basis of the questions in Experiment 1. The clips varied in length from 30 seconds to one minute. The test stimuli consisted of 123 -AFC questions identical to those in Experiment 1, and related to the video clip content (three movie quotes, three logos, and three Google logo related questions ${ }^{2}$ ). The demographics questionnaire and 3-AFC

[^261]questions were administered through Qualtrics Survey System on Dell computers in the lab.
Design After completing the demographic questionnaire, participants viewed seven of the eight video clips, one at a time, wearing a pair of provided headphones. Participants were instructed that they would receive a memory test on the content of the video clips. After viewing all seven target video clips, participants then watched an eighth clip which was unrelated to the memory test. This clip served as a distractor between study and test and lasted for roughly five minutes. After the distractor, participants answered 12 recognition memory questions related to the seven clips they had just viewed. Importantly, these questions were identical to those in Experiment 1 including the 3-AFC (target response, critical lure, and non-critical lure), confidence ratings, and source attribution questions. The recognition
questions were self-paced and took participants 15 minutes, on average, to complete.

## Results

To evaluate whether participants could correctly recognize the target quote and graphic after having viewed the original source, we computed the response probability for the target, critical lure, and non-critical lure for each question (see Figure 3). We found that participants were split for the Star Wars question with $44 \%$ choosing the lure response and $47 \%$ choosing the target response, indicating that performance was substantially better for participants after watching the original source (with $95 \%$ choosing the critical lure in Experiment 1). Interestingly, however, for the other movie related questions the preferred response was still the critical lure (Snow White: 75\%; Forrest Gump: 78\%). Participants were split for the Apple logo question, with $47 \%$ of participants choosing the target, and $33 \%$ choosing the more closely matching lure. It is important to note that more participants responded correctly with the target in Experiment 1 than in Experiment 2. For the other graphic questions, the preferred response remained the target (American flag: 84\%; Microsoft logo: 56\%).

A series of binomial tests were conducted to assess whether the proportion of correct responses for each question was different from chance (see Table 1). The tests revealed no significant difference for the Star Wars question and the Apple logo question. There was a significant difference in performance for the remaining questions, with participants doing significantly worse than chance for the Snow White and Forrest Gump questions, and significantly better than chance for the American flag and Microsoft Logo questions.

In order to further examine whether the introduction of the video clips improved performance, a chi-square test was conducted to compare the proportion of correct responses in Experiment 1 to the proportion of correct responses in Experiment 2 (see Table 1). The test revealed that performance on the Star Wars and Forrest Gump questions were slightly better for those participants who watched the video clips in Experiment 2. For the Star Wars question, there were significantly more people who answered correctly in Experiment $2(47.22 \%)$ compared to Experiment 1 $(3.28 \%)$. For the Forrest Gump question, there were significantly more people who answered correctly in Experiment 2 ( $11.11 \%$ ) compared to Experiment 1 ( $0 \%$ ), indicating that recognition accuracy was higher for those participants who had studied the direct source. However, there was no significant difference in performance between experiments for the Snow White, American flag, or Microsoft logo questions, indicating that for some questions, even for participants who had seen the video clips, recognition performance was no better than for participants in Experiment 1 who had not viewed the direct source. For the Apple Logo question, there were significantly more people who answered correctly in Experiment 1 (81.97\%) compared to Experiment 2 ( $47.22 \%$ ).

We then assessed confidence for the preferred response for each question in Experiment 2. For the movie related
questions, a majority of those choosing the critical lure also responded with high confidence, defined as five or higher (Star Wars: 88\%; Snow White: 81\%; Forrest Gump: 75\%, Apple logo, critical lure: $71 \%$, more closely matching lure: $58 \%$ ). It is important to note that for the Apple Logo question, more people answered with the critical lure in Experiment 2 than in Experiment 1. However, for the remaining graphic questions, a majority of those choosing the target responded with high confidence (American flag: 90\%; Microsoft logo: $55 \%$ ), see Figure 4 . When comparing confidence ratings for targets to critical lures, we found that there was no significant difference in confidence ratings between target and critical lures for any of the questions, indicating that participants who chose the critical lure were just as confident as those who chose the target.

For source attribution, in Experiment 2 we computed the response probability for each possible source conditioned on target and critical lure responses. Here we report the most frequently selected source for the preferred response for each question. We found that a majority of those who chose the critical lure also responded that they had "watched the movie" as their source of knowledge (Star Wars, critical lure: $50 \%$, target: $53 \%$; Snow White, critical lure: $93 \%$; Forrest Gump, critical lure: 82\%), see Figure 5.

## Discussion

The current study investigated people's semantic memory (i.e., prior expectations) and recognition memory for pop culture content, as a function of their confidence and perceived information source. We found that people chose the ubiquitous incorrect (critical) lure for the movie related questions. These results are consistent with the finding that people remember the gist of sources and not most of the details (Sachs, 1967). These responses were given with high confidence, with participants indicating that they had learned this information from the direct source, i.e., the movie. This suggests that people have strong prior expectations for pop culture content frequently circulated through the media even though these expectations are not correct.

We found that in Experiment 2, exposure to original source material did not overwhelm pop-cultural distortions or lead to higher recognition accuracy for all of the movie quotes. Participants were still prone to select the critical lure, and did so with high confidence. This suggests that when prior expectations are strong, even if they are misaligned to the truth, viewing the original source cannot always overcome this inaccuracy.

Although there has been some success in correcting false memories (Brewer, 1977; Fazio \& Marsh, 2010), we did not observe this in our study. Providing the original source material may have been ineffective in our task because we did not explicitly inform our participants to attend to the original source material (movie quotes or graphics). This was purposely done to simulate real world settings where people may passively encode information.

These two findings together present a somewhat dangerous picture of what misinformation can do to episodic memory.

For example, for the Apple logo where semantic memory (Experiment 1) was highly accurate, recognition memory (Experiment 2) faltered. This might be because the somewhat abstract noisy semantic representation is integrated with a noisy episodic representation in the reconstruction process. In contrast, for the pop culture movie quotes, recognition memory was only slightly better than semantic memory. This might be a result of semantic representations exerting a strong influence on the noisy episodic traces in the reconstruction process-consistent with a Bayesian interpretation of memory (e.g., Hemmer \& Steyvers, 2009). In other words, the false semantic representation provides a high baseline contribution to episodic memory that is too strong to overcome even with exposure to the true source.

The implications of these findings as they relate to real world events are far reaching. Much of the "fake news" that was circulated the Internet throughout the 2016 presidential election consisted of fabricated stories posing as professional journalism. These stories were spreading misinformation, and ultimately became a means to influence public opinion. The importance of this issue has grown over time, as more people have reported that they get their news from the Internet (Lee, 2016). If we examine contemporary popular culture and the focus on social network distribution, it is easy to see how information spreads very rapidly through re-posts, re-tweets, or sharing via word of mouth throughout the Internet population, often in ways that the original producers cannot determine or control (Burgess, 2008). So, the next time you share a post on Facebook, or quote a movie, make sure you check your sources.

## Acknowledgments

We would like to acknowledge that the first author of this paper is an undergraduate in the Rutgers University Honors in Psychology program. This research was supported by the National Science Foundation grant \#1453276.

## References

Bacon, F.T. (1979). Credibility of repeated statements: Memory for trivia. Journal of Experimental Psychology, 5, 241-252.
Bartlett, F. (1932). Remembering: A study in Experimental and Social Psychology. Cambridge: Cambridge University Press.
Brewer, M. (1977). Memory for the pragmatic implications of sentences. Memory \& Cognition, 5(6), 673-678.
Burgess, Jean (2008) 'All your chocolate rain are belong to us?' Viral Video, YouTube and the dynamics of participatory culture. In Video Vortex Reader: Responses to YouTube. Institute of Network Cultures, Amsterdam, pp. 101-109.
Fazio, L. K., \& Marsh, E. J. (2010). Correcting false memories. Psychological Science, 21, 801-803.
Hasher, L., Goldstein, D., \& Toppino, T. (1977). Frequency and the conference of referential validity. Journal of Verbal Learning and Verbal Behavior, 16, 107- 112.

Hemmer, P., Steyvers, M. (2009). Integrating Episodic Memories and Prior Knowledge at Multiple Levels of Abstraction. Psychonomic Bulletin \& Review, 16, 80-87.
Lee, T. B. (2016, November 16). Facebook's fake news problem, explained. Retrieved January 18, 2017, from http://www.vox.com/newmoney/2016/11/16/13637310/fa cebook-fake-news-explained.
Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., \& Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest, Supplement, 13(3), 106-131.
Lindsay, D. S. (1994). Memory source monitoring and eyewitness testimony. In D. F. Ross, J. D. Read, \& M. P. Toglia (Eds.), Adult eyewitness testimony: Current trends and developments (pp. 27-55). New York: Cambridge University Press.
Loftus, E. F., \& Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior, 13, 585-589.
Loftus, E. F., Miller, D. G. \& Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. Human Learning and Memory, 4, 19-31.
McKenzie, S. \& Eichenbaum, H. (2011). Consolidation and reconsolidation: two lives of memories? Neuron, 71 (2011), pp. 224-233.

Mitchell, K. J., \& Zaragoza, M. S. (1996). Repeated exposure to suggestion and false memory: The role of contextual variability. J Mem Lang, 35, 246-260.
People Claim to Have Seen a Mysterious Movie That Never Existed. (2016, December 31). Relevant Magazine.
Retrieved from http://www.relevantmagazine.com/slices/ people-claim-have-seen-mysterious-movie-never-existed.
Roediger, H. L., III, Meade, M. L., \& Bergman, E. T. (2001). Social contagion of memory. Psychonomic Bulletin and Review, 8, 365-371.
Roediger, H. L., \& DeSoto, K. A. (2016). The power of collective memory: What do large groups of people remember - and forget? Scientific American.
Sachs, J. (1967). Recopition memory for syntactic and semantic aspects of connected discourse. Attention, Perception, \& Psychophysics, 2(9), 437-442.
Schacter, D. L. (2001). The seven sins of memory: How the mind forgets and remembers. Boston, MA: Houghton Mifflin.
Xu, J. \& Griffiths, T.L. (2009). How memory biases affect information transmission: A rational analysis of serial reproduction. Advances in Neural Information Processing Systems, 21.
Zaragoza, M. S., \& Lane, S. M. (1994). Source misattributions and the suggestibility of eyewitness memory. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 20, 934-945.

# A Computational Model for Reasoning About the Paper Folding Task Using Visual Mental Images 

James Ainooson (james.ainooson@ vanderbilt.edu)<br>Maithilee Kunda (mkunda@vanderbilt.edu)<br>Department of Electrical Engineering and Computer Science, PMB 351679<br>2301 Vanderbilt Place, Nashville, TN 37235-1679, USA


#### Abstract

The paper folding task is commonly used for the evaluation of nonverbal, spatial reasoning skills. In this paper, we present a computational model that attempts to use visual-imagerybased representations and operations to solve this task. The model was tested against all problems from the standard paper folding task and achieved a perfect score, illustrating that visual-imagery-based representations and operations are sufficiently expressive to capture at least one successful solution strategy. Although the model does not closely resemble human cognitive processing, and thus should not be considered in its current form to be a plausible psychological model of human task performance, the assumptions made and their implications for our understanding of human cognition on the paper folding task point to fruitful lines of future work towards this goal.


Keywords: artificial intelligence; cognitive assessment; paper folding; spatial skills.

## Introduction

Paper folding tasks are cognitive assessment tools used in the evaluation of spatial, non-verbal reasoning skills. Visuospatial skills in general are thought to be critical to a variety of human endeavors, including scientific discovery (Miller, 1984), art (Arnheim, 1969), engineering (Ferguson, 1994), computer programming (Petre \& Blackwell, 1999), mathematics (Giaquinto, 2007), education (Silverman, 2002), and even feats of memory (Foer, 2011). Visuospatial skills also seem to be areas of intact or even superior performance for certain individuals with developmental conditions such as autism (Soulieres et al., 2011; Kunda \& Goel, 2011) and Prader-Willi syndrome (Verdine et al., 2008).

In research on Science, Technology, Engineering and Mathematics (STEM) education, visuospatial ability is viewed as a key contributor to math learning (National Research Council, 2009) and to pursuing degrees and careers in STEM disciplines (Wai et al., 2009). Studies suggest that visuospatial ability can improve with training (Uttal et al., 2013), and that such training can enhance math performance in children (Cheng \& Mix, 2014).

Thus, there is an urgent need for effective visuospatial assessments as well as training interventions to promote learning outcomes, creative discoveries, effective design work, and more. Understanding the specific cognitive mechanisms that underlie visuospatial ability is a critical step along this path.

Of course, studying visuospatial ability purely through observations of human behavior is challenging because many of the underlying cognitive processes are not directly observable. Even neuroimaging yields only a coarse view of the specific information processing steps that take place as someone is solving a task.


Figure 1: A sample task from the VZ-2 paper folding test. The images on the left of the vertical line depict the stages in a fold. The images on the right of the line are possible choices of how the paper may look when unfolded.

In this paper, we instead adopt the approach of implementing a computational cognitive model that simulates solving the task-the cognitive systems approach of artificial intelligence (AI) (Thagard, 2005). Cognitive systems model how intelligent agents combine different cognitive processes, like learning, reasoning, and memory, to perform a task. By implementing a cognitive system that simulates solving visuospatial tasks, we can look "under the hood" at specific types of information processing mechanisms that might drive visuospatial ability.

In previous work, we have implemented similar models that investigate aspects of visuospatial cognition on other cognitive assessment tasks. Previous work on the Raven's Progressive Matrices intelligence test has examined the role of visual mental representations in solving difficult test problems (Kunda et al., 2013), the contributions of different types of imagery operators (Kunda et al., 2013), the meaning of different patterns of errors (Kunda, Soulières, et al., 2016), and visual mechanisms for maintaining goal-subgoal hierarchies (Kunda, 2015). Previous work on the Block Design task has looked at relationships between internal mental representations and external deployments of visual attention (Kunda, El Banani, \& Rehg, 2016), and previous work on the Embedded Figures task has looked at capacity limits in visuospatial memory, in particular the effects on task performance of internal deployments of visual attention to different parts of a visual mental representation (Kunda \& Ting, 2015).

In this paper, we present initial results from a new computational model of the paper folding task. Although the model does not closely resemble human cognitive processing, and thus should not be considered in its current form to be a plausible psychological model of human task performance, the assumptions made and their implications for our understanding of human cognition on the paper folding task point to fruitful lines of future work towards this goal.

In particular, the model we present can be considered as
an experiment on the sufficiency of certain imagery-based representations and operations for solving paper folding, which is valuable for understanding how different cognitive mechanisms might in theory contribute to visuospatial ability in people, and especially how certain cognitive limitations might affect task performance. Ultimately, we hope that results from this line of work will serve as a basis to suggest routes for how such cognitive limitations might eventually be overcome, i.e., in developing new visuospatial training interventions for use in education.

## About Paper Folding Tasks

Paper folding tasks are usually presented as line-drawings of paper cut-outs or folded pieces of paper. People are then asked to imagine changes that happen when this paper is manipulated in different ways. Several forms of the test exist.

Shepard and Feng (1972) presented a form of the paper folding test which required test subjects to fold a cube out of six connected squares. Two of the squares have arrows that point to an edge and one square is greyed out to show that it is the base of the cube. People are then required to predict whether the arrows align and point to the same edge when the cube is re-constructed. This essentially requires them to mentally reconstruct the cube by imagining folding the paper as though the shape was cut-out of the paper. Lovett and Forbus (2013) developed a computational model developed to reason about this specific case of paper folding tasks. Their model takes the approach of simplifying the task by removing unnecessary details and essentially focusing primarily on the orientation of critical edges to solve the task.

Another form of the paper folding task (which is also called "the punched hole test"), developed by Ekstrom and colleagues (1976), is administered as a 6 minute pencil and paper test in two parts ( 3 minutes for each part). During the test, subjects are presented with a sequence of images showing the stages in folding a square piece of paper. A hole is then punched on this folded piece. Test subjects are also presented with five possible outcomes of how the paper looks when it is unfolded. See Figure 1 for an example of this type of problem. Designed as part of the "Kit of Factor-Referenced Cognitive Tests", this task appears as the second test under the group of tests that evaluate the visualisation cognitive factor (VZ-2).

A number of research studies have used paper folding to evaluate spatial reasoning skills. While testing a hypothesis on learning styles amongst individuals, Mayer and Massa (2003) used this test as part of their measure of spatial skills. Keehner et al. (2004) also used this test as one of their tests of spatial ability while investigating the correlation between spatial ability, experience and skill in laparoscopic surgery. Another example is the study by Silvia (2008) in which the paper folding test was used as part of a measure of fluid intelligence, while investigating relationships between creativity and intelligence.

There is much that is still unknown about the direct cogni-
tive mechanisms involved in paper folding tests. Mental rotations are believed to play a major role (Shepard \& Feng, 1972). In addition to the complexity of the mental rotations, people may have to deal with the additional component of mental folding (Glass et al., 2012) and in the case of a punched hole test, how the holes affect the final output. Wright and colleagues (2008) showed that training on mental rotation tasks improved performance on paper folding tasks, just as training on paper folding tasks improved mental rotations.

Next, we present a computational model that attempts to solve the paper folding task using simulated "mental rotations" in "three dimensions". The exact formulation of the paper folding task we intend to tackle with this model is "The Punched Hole" test (Ekstrom et al., 1976).


Figure 2: A sequence of images sent as input to the model (blue and white), and the corresponding bitmaps that are used by the model after inputs are processed (black and white). The first input row corresponds to the initial "problem" part of a paper folding item. The second input row contains the possible choices the model is presented with.

## The Model

We present a computational model that attempts to solve "The Punched Hole" paper folding task (Ekstrom et al., 1976) using only image based operations. The model is built in the Python programming language and relies extensively on the Pillow fork of the Python Imaging Library (PIL) to perform low level image manipulation.

The main task of the model is to analyze a sequence of images that depict the folding and punching in a problem
of the paper folding task to determine what the paper would look like when unfolded. It achieves this by maintaining a three dimensional representation of the paper which is stored as a stack of two dimensional images. Each image on the stack represents a single level of folding performed in a single time-slice. The actual fold operations are performed with image reflections that provide a simplified simulation of threedimensional rotations.

## Input

Inputs to the model are presented in three stages. The first stage consists of a sequence of images that represent the state of the folded paper in each time-slice. The second stage consists of a single image that represents the state of the folded paper after the hole has been punched. Finally, the third stage presents the possible solutions from which the model could select an answer after it is done predicting a solution.

All the inputs are presented as line drawings with sections that contain paper filled with pixels to ensure the model can properly differentiate between paper and empty space. Before any images are passed on for further processing, they are converted to single colour images. This makes it easy for the model to perform logical bitwise operations between images. Once converted to a single colour image, any pixel that is set to 1 in the image is considered to be an area containing paper, and pixels set to 0 are considered to be empty spaces. See Figure 2 for a sample sequence of input images and their corresponding bitmap representations.

## Strategy

The model's strategy for solving the tasks relies heavily on two stacks. The first stack (which we call the image stack) keeps track of images that represent the layers of folds. The second stack (which we call the operations stack) keeps track of the operations that are performed on the images as folds occur. In predicting the solution, the model utilizes four main operations: Initialize, Fold, Punch and Unfold.

The Initialize operation sets up the model before solving any task. It places an image which has all its bits set to 1 on the images stack. This image is meant to represent the initial piece of unfolded paper on which fold operations will be performed.

The Fold operation receives as input an image of the state of the paper after a given fold has occurred. The model will attempt to use this image and other images on the image stack to find the best estimation of the line along which the fold was made. To accomplish this, the following processing steps are performed on the bitmap representation of the fold input image for every image on the stack, starting from the bottom:

1. The current layer to be processed is retrieved from the image stack. In the case of the first fold operation, this image has all pixels set to 1 .
2. An intersection operation is performed between an inverse of the fold input image and the image retrieved from the stack. This operation is constrained by the bounding box
around the image that was retrieved from the stack. The resulting image is an image of the flap to be folded, as illustrated in Figure 3.


Figure 3: Stages the images go through to generate the folded flap image.
3. Another intersection operation is performed between the fold input and the image retrieved from the stack. This new image replaces the original image on the stack, as shown in Figure 4.


Figure 4: The intersection between the input image and the existing image on the stack to generate a replacement for the image on the stack.
4. In order to determine the line along which the fold would occur, a single pixel border is drawn around both the folded flap image, and the modified image on the stack. This is to make both images larger so they can slightly overlap. An intersection operation is computed between the two new overlapping images to generate an image containing the line along which the fold is to be made. A search for two extreme coordinates of this image is then performed on the pixels in this image. This search is biased towards the pixels that are from the folded flap image. Search results will now contain the coordinates of the fold line. These coordinates are then pushed onto the operations stack.


Figure 5: Intersecting the new overlapping images to determine the fold line.
5. Finally, the folded flap image is reflected across the fold line. This reflection operation is analogous to a $180^{\circ}$ rotation in three dimensional space about the fold line axis. The reflected image is then pushed to the top of the stack to act as one of the base images for any subsequent fold operations. After a series of fold operations is performed, each image on the stack will represent a folded layer.


Figure 6: The final reflection operation of the fold.

Once all fold operations are completed, the Punch operation is performed. This operation takes the punch input, and computes the intersection of this input with all the images on the stack, replacing all the contents on the stack with the results. See Figure 7 for an image depicting all the changes that take place on the stack for fold operations and the punch operation.


Figure 7: The various states of the image stack as inputs are fed to the model.

Unfold is the next operation after the punch has been performed. This operation runs through the operations stack and performs a reverse of all operations. It works by picking the image on top of the stack and the one below the stack. It performs the reverse operation on the image on top of the stack (which will be the folded flap).

It then performs an OR operation between the folded flap and the base image. The new image generated after the OR operation is placed onto a new image stack. The unfold operations are recursively called on the newly generated image stack until the stack contains a single image. This image will represent the model's predicted solution to the problem.

A final solution can be chosen by the model with the last image generated after the unfold operation. A pixel by pixel comparison is performed between the model's prediction and each of the possible solutions. The comparison that yields the largest number of matching pixels is selected as the solution.


Figure 8: The various states of the image stack during the unfold operations.

## Experiments and Observations

We tested the model against all twenty items from "The Punched Hole" test (Ekstrom et al., 1976). Input images (for both the fold stages and answer choices) for this experiment were taken from the original test but redrawn as "clean" versions using the Inkscape Vector Graphics editor. Redrawn vector images were converted to raster images before being passed to the model.

Results for the experiment were a count of the number of items on which the model was able to select the correct answer. When this experiment was performed, the model selected the correct answer on all items in the test-a score of 20 out of 20 .

Looking more carefully at the operation of the model, we observed that a constant number of operations are performed for each fold simulated. Also, the size of the stack grows exponentially with respect to the number of folds performed. For every $n$ folds, there are a total of $2^{n}$ items on the stack.

Interestingly, the set of problems had 1 to 3 fold levels. This meant that the maximum stack size required for the tasks varied from 2 (for single folds) to 8 (for triple folds). If we take the size of the image stack to be analogous to "working memory usage" in our model, the maximum number of items stored while solving any of the paper folding problems is consistent with what is known about visuospatial working memory capacity limits in people (Luck \& Vogel, 1997).

Also, working in the image domain gives the model the ability to operate on arbitrary folding tasks. To test this ability, we ran a "paper snowflake" simulation through the model to evaluate its output. From Figure 9, we can clearly observe that the model generated an output that corresponds to the snowflake folds passed through it.


Figure 9: The model's "solution" to a paper folding problem with arbitrary fold and punch shapes. The top row of images show the input to the model and the bottom row of images showt the output of the model at the various stages of unfolding the snowflake.

## Discussion and Future Work

One current assumption of this model has to do with the comparison technique used to match possible solution choices to the predicted solution. As stated earlier, this operation is performed using a pixel by pixel comparison technique, where pixels on the predicted solution must closely match pixels on a possible solution. For the case of our experiment, these possible solutions were carefully drawn such that holes in both the right and wrong choices were precisely placed.

However, as we have observed in other standardized cognitive assessments (Kunda et al., 2013; Kunda \& Ting, 2015), printed figures in test booklets are not always so precise at the pixel level. (We surmise that many of these tests must have been hand-drafted when they were first created.) On the actual paper folding test by Ekstrom et al. (1976), many of the positions of the punched holes are not necessarily aligned perfectly. However, people are still able to solve these tasks, which suggests that the model should have more robust pattern recognition and processing abilities.

Our model has shown one possible set of cognitive mechanisms, based on visual mental images, that are sufficient for solving the paper folding test. Other strategies undoubtedly exist. What is more interesting, perhaps, is a consideration of why people might fail to solve paper folding items. What mechanism or set of mechanisms might they lack?

One possibility is working memory capacity, simulated in our model as the size of the image stack. Clearly, limiting the size of the stack will immediately reduce the model's ability to successfully solve paper folding problems. However, there are other possibilities as well.

One idea is that people might "forget" where the fold is on a folded up piece of mental paper, and proceed to unfold the paper in the wrong direction. (For example, they might interchange the folded side and the open side of a folded page, at the moment when they are unfolding it.) This is a subtle error that does not have to do with raw capacity but more like a limit on attentional capability, or the accurate persistence
of information in working memory. We speculate that these types of fold-forgetting errors may lead participants to choose some of the distracter answer choices that are provided.

In continued work on the model, we will implement some of these cognitive limitations to see what might lead the model to make particular types of errors. Then, we could compare the errors made by different configurations of the model to the errors made by people, to see if there are suggestive connections between cognitive strategy variations and behavioral error patterns (Kunda, Soulières, et al., 2016).

## References

Arnheim, R. (1969). Visual thinking. University of California Press.
Cheng, Y.-L., \& Mix, K. S. (2014). Spatial training improves children's mathematics ability. Journal of Cognition and Development, 15(1), 2-11.
Ekstrom, R. B., French, J. W., Harman, H. H., \& Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational testing service.
Ferguson, E. S. (1994). Engineering and the mind's eye. The MIT press.
Foer, J. (2011). Moonwalking with Einstein: The art and science of remembering everything. Penguin.
Giaquinto, M. (2007). Visual thinking in mathematics. Oxford University Press.
Glass, L., Krueger, F., Solomon, J., Raymont, V., \& Grafman, J. (2012). Mental paper folding performance following penetrating traumatic brain injury in combat veterans: a lesion mapping study. Cerebral Cortex, bhs153.
Keehner, M. M., Tendick, F., Meng, M. V., Anwar, H. P., Hegarty, M., Stoller, M. L., \& Duh, Q.-Y. (2004). Spatial ability, experience, and skill in laparoscopic surgery. The American Journal of Surgery, 188(1), 71-75.
Kunda, M. (2015). Computational mental imagery, and visual mechanisms for maintaining a goal-subgoal hierarchy. In Proceedings of the third annual conference on advances in cognitive systems acs (p. 4).
Kunda, M., El Banani, M., \& Rehg, J. M. (2016). A computational exploration of problem-solving strategies and gaze behaviors on the block design task. In 38th annual conference of the cognitive science society, philadelphia, usa.
Kunda, M., \& Goel, A. K. (2011). Thinking in pictures as a cognitive account of autism. Journal of autism and developmental disorders, 41(9), 1157-1177.
Kunda, M., McGreggor, K., \& Goel, A. K. (2013). A computational model for solving problems from the Ravens Progressive Matrices intelligence test using iconic visual representations. Cognitive Systems Research, 22, 47-66.
Kunda, M., Soulières, I., Rozga, A., \& Goel, A. K. (2016). Error patterns on the Raven's Standard Progressive Matrices Test. Intelligence, 59, 181-198.
Kunda, M., \& Ting, J. (2015). Looking around the minds eye: How internal deployments of attention can affect vi-
sual search performance. In Proc. third annual conf. advances in cognitive systems (acs).
Lovett, A. M., \& Forbus, K. D. (2013). Modeling spatial ability in mental rotation and paper-folding. In Cogsci.
Luck, S. J., \& Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390(6657), 279-281.
Mayer, R. E., \& Massa, L. J. (2003). Three facets of visual and verbal learners: Cognitive ability, cognitive style, and learning preference. Journal of educational psychology, 95(4), 833.
Miller, A. I. (1984). Imagery in scientific thought: Creating 20th-century physics. Boston: Birkhauserm.
National Research Council. (2009). Mathematics learning in early childhood: Paths toward excellence and equity. Na tional Academies Press.
Petre, M., \& Blackwell, A. F. (1999). Mental imagery in program design and visual programming. International Journal of Human-Computer Studies, 51(1), 7-30.
Shepard, R. N., \& Feng, C. (1972). A chronometric study of mental paper folding. Cognitive psychology, 3(2), 228243.

Silverman, L. K. (2002). Upside-down brilliance: The visualspatial learner. DeLeon Publishing Denver, CO.
Silvia, P. J. (2008). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. Personality and Individual differences, 44(4), 1012-1021.
Soulieres, I., Zeffiro, T., Girard, M., \& Mottron, L. (2011). Enhanced mental image mapping in autism. Neuropsychologia, 49(5), 848-857.
Thagard, P. (2005). Mind: Introduction to cognitive science. MIT Press.
Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., \& Newcombe, N. S. (2013). The malleability of spatial skills: a meta-analysis of training studies. Psychological bulletin, 139(2), 352.
Verdine, B. N., Troseth, G. L., Hodapp, R. M., Dykens, E. M., \& Conners, F. (2008). Strategies and correlates of jigsaw puzzle and visuospatial performance by persons with prader-willi syndrome. American Journal on Mental Retardation, 113(5), 343-355.
Wai, J., Lubinski, D., \& Benbow, C. P. (2009). Spatial ability for stem domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. Journal of Educational Psychology, 101(4), 817.
Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., \& Kosslyn, S. M. (2008). Training generalized spatial skills. Psychonomic Bulletin \& Review, 15(4), 763-771.

# Legal HARKing: theoretical grounding in interaction research 

# Saul Albert (saul.albert@tufts.edu) 

J.P. de Ruiter (jp.deruiter@tufts.edu)<br>Department of Psychology, Tufts University<br>Medford, MA 02155 USA


#### Abstract

In psychology, we tend to follow the general logic of falsificationism: we separate the 'context of discovery' (how we come up with theories) from the 'context of justification' (how we test them). However, when studying human interaction, separating these contexts can lead to theories with low ecological validity that do not generalize well to life outside the lab. We propose borrowing research practices from formal inductive methodologies during the process of discovering new regularities and analyzing natural data without being led by theory. From the perspective of experimental psychology, this approach may appear similar to the 'questionable research practice' of HARKing (Hypothesizing After The Results are Known). We argue that a carefully constructed form of HARKing can be used systematically and transparently during exploratory research and can lead to more robust and ecologically valid theories. Keywords: HARKing; experimental psychology; conversation analysis; methodology; interaction


## Performance-enhancing questionable practices

Most discussions of the current 'replication crisis' in psychology and the social sciences (Pashler \& Harris, 2012; Pashler \& Wagenmakers, 2012) focus on identifying and mitigating the biases and incentives that lead researchers to adopt questionable research practices (QRPs): a range of methods for manipulating experimental results and processes that John, Loewenstein, and Prelec (2012) describe as "the steroids of scientific competition, artificially enhancing performance". But science-at least ideally-is not about competition, and the highest scientific achievements are of benefit to all. It makes sense, therefore, to look at some QRPs and their underlying rationales in more detail: why are they so tempting? What makes them 'safe' or 'unsafe' for science in specific contexts? For example, qualitative, inductive methods often used in cognitive science such as grounded theory (Glaser \& Strauss, 1967) are very useful for exploratory studies in many research areas, but may produce misleading inferences when used to code certain kinds of behavioral phenomena for confirmatory, quantitative research into language and human interaction (Stivers, 2015). However, rather than simply labeling all such methods as QRPs in the context of experimental, confirmatory research, we may be able to borrow from them to enhance our research results without compromising our methodological rigor. In many of the failed replications reported in Open Science Collaboration (2015), it seems that QRPs are used to increase the probability of 'finding' an effect predicted by the stated theory. Theorizing about an interactional phenomenon that has no grounding in interactional reality makes QRPs attractive, simply because they make it more likely that researchers will be able to report significant effects that support their theory. The issue underlying the use
of QRPs in the study of human interaction, then, may be intrinsically related to the broader problem of groundless theorizing, where theories are formulated without being familiar with the situations they theorize about. We suggest that this problem, in turn, stems from some uncritical assumptions about science and human interaction.

## The problem of groundless theorizing

A common assumption about falsificationism, still implicitly or explicitly a major philosophical underpinning of empirical science, is that as long as a theory can be falsified by testing a hypothesis, the scientist is free to theorize any conceivable causal relationship between any measurable variables. There is nothing inherently wrong with this approach if all plausible confounding variables can be controlled, and this theoretical freedom of movement is tremendously powerful. Popper was inspired by how the freedom to theorize raised the stakes for cosmologists such as Einstein, whose entire theory of general relativity could have been falsified if just one of his audacious predictions about electromagnetism and gravitational potential had turned out to be false. However, in the context of human interaction research, it is notoriously difficult to control for confounds because there are many human behaviors that are very difficult if not impossible to emulate in controlled conditions (De Ruiter, 2013; Schegloff, 2006), and just recording or observing people interacting may change the ways they interact in unpredictable ways (Labov, 1972). In this paper we describe a set of pre-theoretical research procedures that interaction researchers can use to constrain their theories to match observable facts within the domain of interest. While this may sound like the questionable research practice of HARKing (Hypothesizing After the Results are Known) (Kerr, 1998), we argue that systematic, inductive methods for analyzing social interaction can provide a principled and effective way to ground theorizing about human interaction, leading to more robust and relevant theories.

## Contexts of discovery and contexts of justification

The 'context of discovery' is the situation in which a phenomenon of interest is discovered. For example, when studying human interaction, a useful context of discovery would be an otherwise naturalistic conversation that happened to be recorded for analysis (Potter, 2002). 'Contexts of justification', in this example, might then include the laboratory, the conference paper, and the academic literature within which the empirical details are reported, analyzed and formulated as a scientific discovery (Bjelic \& Lynch, 1992). These

| Context | Participants' resources | Analysts' resources |
| :--- | :--- | :--- |
| Discovery | Observable actions <br> Observable settings | Observable actions <br> Observable settings |
|  | Knowledge \& experience <br> beyond current interaction | Ability to fast forward, <br> rewind and replay. |
| Justification | Sequential organization of <br> talk \& social actions <br> Self/other correction <br> (repair) | Sequential organization of <br> talk \& social actions <br> Self/other correction <br> (repair) |
|  | Introspection <br> Inductive reasoning | Quantification <br> Deductive analyses |

Figure 1: Interactional resources and methods within participants' or analysts' contexts of discovery and justification.
are important distinctions for what we think of as theoretical grounding: drawing together the contexts of discovery and contexts of justification in order to place principled limits on theorizing about interaction. However, drawing these contexts together in interaction research requires analysts to take account of critical distinctions between what kinds of evidence is available to analysts and the kinds of interactional resources available to participants in the situation itself (Garfinkel, 1964; Lynch, 2012). Figure 1 lists a few key resources participants and analysts can both use when discovering and justifying interactional phenomena, and (underneath, in red) some of the resources for making sense of interaction that are only available from one perspective or the other. Some of these resources are shared, for example, both participants and overlooking analysts can use observable features of the setting and the visible actions of the people within it to discover new phenomena. Both participants and analysts can also observe when these visible actions are contiguous and uninterrupted (Sacks, 1987), and can see if certain actions are routinely matched into patterns of paired or 'adjacent' initiations and responses (Heritage, 1984, p.256). Similarly, both analysts and participants can observe when contiguous flows of initiation and response seem to break down, falter or require repair to re-establish orderliness and ongoing interaction (Schegloff, Jefferson, \& Sacks, 1977). By contrast, many other resources and methods for making sense of the situation are exclusively available to one or another role. For example, analysts can repeatedly listen to a recording, slow it down, speed it up, and can precisely measure, quantify, and deduce cumulative facts that would be unimaginable to participants in the interaction. On the other hand, participants may draw on their store of tacit knowledge and use introspectionoptions which are not necessarily observable for overlooking analysts-to make sense the current state and consequences of the interaction. 'Discovery' for participants, then, is some action or phenomenon observably discovered and treated as mutually relevant with others in the situation. Justification is the interactional work participants do with others in the setting to display and uphold the mutual intelligibility and rationale of their actions: an imperative that Garfinkel (1967)
describes as 'mutual accountability'. For analysts, discovery and justification use as many of the same resources as possible, but are motivated by different concerns i.e. to provide causal explanations for the events and phenomena discovered for the purposes of scientific research, but without the urgent imperatives of mutual accountability. The challenge for analysts wishing to improve their theories by bringing together contexts of discovery and contexts of justification is to constrain themselves to testing theories that deal with resources and methods that are evidently available to both analysts and participants.

## Pre-experimental HARKing for better theories

When we advocate pre-experimental HARKing, it should be clear that this proviso about using interactional resources available to both participants and analysts excludes 'exploratory data analysis' (Jebb, Parrigon, \& Woo, 2016) or other uses of inferential statistics for pre-confirmatory theorizing since this is not something participants would be able to use as a resource within their contexts of discovery or justification. Rather, the research procedures recommended here are inspired by conversation analysis (CA): an approach to interaction research which exemplifies the use of empirical constraints on theorizing (Schegloff, 2007, pp. xii-xiii), and which has tended to avoid engagement with experimental studies that necessarily prioritize theorizing in order to arrive at causal explanations (Kendrick, 2017). The 'theoretical asceticism' (Levinson, 1983, p. 295) of CA's research practices makes them very useful for drawing together contexts of discovery and justification in a principled and coherent way (De Ruiter \& Albert, 2017). In relation to theorizing, we call these practices 'pre-experimental HARKing' to draw attention to the distinction between HARKing as a QRP (after having produced a theory and tested it with an experiment), and CA's "qualitative, inductive, and strictly empirical" (Haddington, Mondada, \& Nevile, 2013, p.7) research processes of systematic observation and ongoing informal peer review that takes place before any theorizing is allowed. One of the ironies of theories and experiments in interaction psychology that use corpora is that the 'results' (i.e. what actually happened in the interaction) usually are known before the hypotheses or research questions are formulated. It therefore makes sense to use these data to develop better theories and operationalizations before having to make key decisions about coding, quantifying and analyzing interactional phenomena. The risk otherwise is that what gets coded, quantified and tested may not turn out to be observably relevant to the participants in the interaction at all (Schegloff, 1993; Stivers, 2015). It should be clear by now that the term HARKing is not used perjoratively here. Since existing data, intuitions, and past results often provide the basis for theorizing at a pre-experimental stage in any case, we advocate using CA's systematic and transparent procedures to constrain and ground those theories empirically.

## Sharing contexts of discovery and justification

A 'result' in the participants' context of discovery can be thought of as the achievement of a reciprocal action in a social situation such as successfully ordering a beer in a bar. This is motivated quite differently from from the 'results' that might be discussed in the analyst's context if the researcher were, for example, designing an experiment to try to figure out what behaviors enable people to obtain beer in bars. Loth, Huth, and De Ruiter (2013) show that going to a bar and systematically observing how beer-ordering is achieved through interaction provides very informative and somewhat surprising results as the basis for formulating new theories. They found that all customers have to do to initiate a successful beerordering interaction is to stand at the bar looking towards the bartender and that any use of the stereotypical ordering-like actions they had anticipated in fact proved to be unnecessary and even potentially disruptive. The first step in drawing together contexts of discovery and justification, then, is to find a setting where participants do observable interactional work to achieve their results (getting a beer in a bar) in ways that are informative for the analyst's results (finding out how people get beer in bars). CA terms this kind of social situation that can be used as a starting point for analysis a 'perspicuous setting' from the Latin perspicio 'to see through', denoting a situation that functions like a microscope that analysts can use to examine the local organization of human affairs. Garfinkel (1992, pp. 184-186) emphasizes that in perspicuous settings participants' affairs are 'locally produced, locally occasioned and locally ordered" and that these function as contexts of discovery and justification of what is relevant for the participants, whose interactions in those contexts are conducted without reference to analyst's concerns. The bar is an obvious choice as a perspicuous setting for exploring beer-ordering, but even if there is no specific domain of inquiry, new questions can also emerge from repeated viewing and 'unmotivated' analysis of data. For example, a corpus of video recordings of guided walking tours has provided a perspicuous setting for discovering questions about how people organize themselves as mobile groups (De Stefani \& Mondada, 2013), about the roles and procedures involved in getting the group to examine something (De Stefani, 2010), and to then coordinate the process of walking away together interactionally (Broth \& Mondada, 2013). The starting point for Legal HARKing, then, is to find a perspicuous setting where participants work together to achieve a given outcome in ways that researchers can then observe and analyze as the basis for formulating more interactionally grounded theories.

## Transcribe interactionally relevant details

Conversational turn-taking is one of the most clearly observable systematic forms of organization in interaction (Sacks, Schegloff, \& Jefferson, 1974). In this sense conversation is a useful example of a context of discovery and justification that is shared between participants and analysts alike. In the context of conversation, participants discover things like
whose turn it is to talk next, and justify their discoveries using a clearly organized protocol for turn-allocation and turntransition. For conversation analysts, the turn-taking system became a foundational context for discovery and justification when Sacks et al. (1974) showed how it could explain systematic features of everyday interaction such as the tendency for minimal gaps and overlaps in natural talk (a discovery that has subsequently been tested experimentally and across multiple languages (Stivers et al., 2009)). CA's transcription system was devised by Gail Jefferson to highlight the systematic patterns of overlap and variations in prosody and intonation (Hepburn \& Bolden, 2012). Although phonetic transcription in IPA notation provides a much higher degree of accuracy than standard orthography, these objective levels of description are not necessarily available to participants themselves, and in any case people in everyday interactions do not usually make an issue of pronunciation. Jeffersonian transcription is relatively simple to read and use, and is optimised to spatialize and represent the features of talk such as speed-ups, stress, stretches, overlaps, and gaps that seem most relevant to participants' contexts of discovery and justification. Most importantly, the activity of hand-transcribing conversational data is a very useful pre-analytical activity in itself through which researchers can become intimately familiar with their data by watching repeatedly while trying to capture the fine details of whatever features are observably relevant to the participants themselves (Bolden, 2015). While all transcription systems introduce the analytic perspectives and assumptions of the analyst doing the transcription (Ochs, 1979), it makes sense to use a system designed specifically to capture the details of talk most demonstrably relevant to how participants maintain the smooth operation of the turn-taking system.

## Use intersubjective review of subjective judgments

Another way to draw together the participants' and analysts' contexts of discovery and justification is to use the interactional aptitudes of the analysts themselves as a heuristic device to explore what is going on in the interaction. This may sound like an overly subjective form of judgment, but since the object of inquiry for analysts is human interaction where we have no better measuring device than our own social intelligence, it makes sense to use our skills as interactants, even if we may not understand how these abilities work. This problems of reliance on subjective intuition can be mitigated through interaction itself. The conversation analytic 'data session' is a research practice where analysts present their data, describe what they see, and have their observations tested against the intuitions and reasoned arguments of other analysts. This is one of the least well-documented aspects of CA, and is barely mentioned in the research or training literature (Sidnell \& Stivers, 2012), although Ten Have (1999, pp. 140-141) provides a brief explanatory description.
"[The data session] often involves playing (a part of) a tape recording and distributing a transcript...The session starts with a period of seeing/hearing and/or read-
ing the data, sometimes preceded by the provision of some background information by the 'owner' of the data. Then the participants are invited to proffer some observations on the data, to select an episode which they find 'interesting' for whatever reason, and formulate their understanding, or puzzlement, regarding that episode. Then anyone can come in to react to these remarks, offering alternatives, raising doubts, or whatever."

The group often consists of both experienced and novice analysts, so there is an element of 'tradecraft' and apprenticeship built into the structure of the data session (Jordan \& Henderson, 1995; Harris, Theobald, Danby, Reynolds, \& Rintel, 2012). Ten Have (1999) in fact attributes his learning CA to having attended data sessions with Gail Jefferson and Emanuel Schegloff, so despite the lack of documentation, the data session has clearly been central to CA from the start. There is a scattering of advice about how to run such sessions in some textbooks, in a few short papers (Hindmarsh, 2012) and even within some reflexive studies that explore CA data sessions as interactional situations in themselves using CA (Antaki, 2008; Harris et al., 2012). These accounts also provide some useful technical advice, for example, Jordan and Henderson (1995) suggest that the 'owner' of the data plays back short clips of up to twenty seconds, then discusses each clip, but limits the discussion to 5 minutes before looking at more data, so that "no single participant can speculate for very long without being called upon to ground her or his argument in the empirical evidence, that is to say, in renewed recourse to the tape." Heath, Hindmarsh, and Luff (2010, pp. 156-157) have some similarly practical advice to limit the session to 20 or fewer people, and not to "cheat and look ahead, or rely on information exogenous to the clip itself": essentially following the rule of thumb to avoid using resources or methods unavailable to participants themselves. Of course the frequent fast forwarding and rewinding of recordings and many of the other analytical methods described here do, nonetheless, rely on resources not necessarily available within the participants' contexts of discovery and justification. However, since the data session is an interactional situation where peers are involved in grounding one another's assumptions about the interaction through interaction, there is also a degree of mutual accountability at work that may compensate for the loosening of CA's strict methodological constraints. Through the data session, theories and assumptions are subjected to open and self-critical debate. Ten Have (1999) sums up this analytic attitude neatly:
"What is most important in these discussions is that the participants are, on the one hand, free to bring in anything they like, but, on the other hand, required to ground their observations in the data at hand, although they may also support them with reference to their own data-based findings or those published in the literature. One often gets, then, a kind of mixture, or coming together,
of substantial observations, methodological discussions, and also theoretical points."

Even after the analyst's painstaking transcripts and observations have run the gauntlet of multiple data sessions where flaws in theory may be identified and discussed, the CA research cycle has just begun by finding candidate phenomena for analysis.

## The analytical phases of pre-experimental HARKing

Having attended multiple data sessions to explore candidate phenomena and findings, there are several further stages required to develop analytically grounded theories about interaction. Conversation analytic primers are now available (Schegloff, 2007; Sidnell \& Stivers, 2012; Ten Have, 1999), so only a summary of analytic procedures is provided here.

After a series of data sessions, analysts collect multiple instances of a target phenomenon each with minor variations in terms of their composition, sequential structure and their range of uses in interaction. Analysts often then work on 'single case analyses' involving an extended study of a few episodes of interaction featuring the target phenomenon in great detail. Over time, the analyst may build up hundreds of cases, organized into 'collections' (Schegloff, 1996), working towards a more complete characterization of the phenomenon and its specialized variations. For example, Schegloff (1968) describes collecting 499 cases of telephone call openings, and considering his collection almost complete and ready to be analyzed. It was the 500th case, however, which provided him with a single 'deviant case' that forced him to re-evaluate his findings about the sequential order of ringing and greeting exchanges in telephone call openings. This example is often cited to demonstrate the difference between these approaches and more conventional case studies. Each single case starts from first (interactional) principles in trying to explore the setting from a vantage point as close to the context of discovery and justification of the participants as possible. For this reason, Schegloff's (1968) example functions as a kind of applied falsificationism: the only way the 500th case could make sense from the analyst's context was to (quite radically) change the theory. Furthermore, long-standing collections of often-analyzed phenomena become theory-like over time, and can be subject to falsification and ongoing modification through contradiction by subsequent CA findings, or through changes in people's patterns of behavior over time. For example, since the mid-2000s the most common telephone call opening sequence has changed significantly due to the prevalence of caller-ID on mobile phones (Raudaskoski, 2009). This process of careful, qualitative analysis is required before CA researchers even consider developing a formal coding scheme (Dingemanse, Kendrick, \& Enfield, 2016; Stivers \& Enfield, 2010) with which to quantify their findings (Stivers, 2015) and run experiments-although these last few steps are still not widely accepted, and remain controversial within CA
(Kendrick, 2017). While this overall procedure is clearly extremely laborious, it does have the reassuring advantage that the phenomena described are guaranteed to have actually occurred in reality, not only in our theoretical imagination.

## Summary: Better theorizing after legal HARKing

This paper argues for researchers of human interaction to devote attention and resources to systematically exploring the context of discovery where their theories will be formulated by extending the falsificationist paradigm. Before we theorize and then test our predictions experimentally, we suggest researchers borrow methods from conversation analysis and other formal inductive methods to enhance the performance of our theories with a kind of pre-experimental 'legal HARKing'. This procedure involves using detailed Jeffersonianstyle transcription, holding data sessions and subjecting our qualitative findings to ongoing, critical analysis before developing theories. By proceeding with our analysis with a sensitivity to the kinds of resources that participants themselves have at hand, we can identify interactional practices that are psychologically relevant and consequential for participants (and not just researchers), and empirically grounded in natural interaction. We expect this kind of grounding to improve the relevance, robustness, and replicability of human interaction research by producing more theoretically grounded hypotheses that we can then test using traditional experimental methods. As long as-at this stage-we pre-register our experiments, we can harness the performance-enhancing benefits of legal HARKing while excluding the dangerous possibility of 'illegal' post-experimental HARKing. More generally, since research practices are seen as 'questionable' in relation to the conventions of a specific methodological framework, we suggest that if we reconsider them at a critical distance from any one methodology, these practices may have many potentially beneficial applications. From the perspective of the experimental research practices that predominate within cognitive science and psychology (Toomela, 2014), for example, the inductive categorizing and coding methods of grounded theory may be seen as 'questionable'. Similarly, from the perspective of generalization-oriented experimental studies, the scope of theories derived from micro-analytic methods such as CA (e.g. about turn-taking) can seem almost trivial (Heritage, 2008). However, taken together the body of work derived from CA's empirical studies constitutes a very broad set of findings about interaction against which generalized theories can be tested (De Ruiter \& Albert, 2017). While different research practices address different problems and questions at different scales, they may also have some useful practical and philosophical intersections. One scientists' 'questionable' research practice can be another's means of rigorous inquiry, and perhaps remaining 'questionable'in the sense of being open to critical review-is something more researchers could aim for in their research practices.

## References

Antaki, C. (2008). Accounting for moral judgments in academic talk : The case of a conversation analysis data session. Text \& Talk, 28(1), 1-30.
Bjelic, D., \& Lynch, M. (1992). The work of a (scientific) demonstration: Respecifying Newton's and goethe's theories of prismatic color. In G. Watson \& R. M. Seiler (Eds.), Text in context: Contributions to ethnomethodology (pp. 52-78). Sage Publications Newbury Park, CA.
Bolden, G. B. (2015, jul). Transcribing as research: "manual" transcription and conversation analysis. Research on Language and Social Interaction, 48(3), 276-280.
Broth, M., \& Mondada, L. (2013, \#feb\#). Walking away: The embodied achievement of activity closings in mobile interaction. Journal of Pragmatics, 47(1), 41-58.
De Ruiter, J. P. (2013). Methodological paradigms in interaction research. In I. Wachsmuth, J. P. De Ruiter, P. Jaecks, \& S. Kopp (Eds.), Alignment in communication: Towards a new theory of communication. John Benjamins Publishing Company.
De Ruiter, J. P., \& Albert, S. (2017, jan). An appeal for a methodological fusion of conversation analysis and experimental psychology. Research on Language and Social Interaction, 1-18.
De Stefani, E. (2010). Reference as an interactively and multimodally accomplished practice. organizing spatial reorientation in guided tours. In M. Pettorino, A. Giannini, I. Chiari, \& F. M. Dovetto (Eds.), Spoken communication (pp. 137-170). Newcastle: Cambridge Scholars Publishing.
De Stefani, E., \& Mondada, L. (2013, \#dec\#). Reorganizing mobile formations: When "guided" participants initiate reorientations in guided tours. Space and Culture, 17(2), 157-175.
Dingemanse, M., Kendrick, K. H., \& Enfield, N. J. (2016, jan). A coding scheme for other-initiated repair across languages. Open Linguistics, 2(1).
Garfinkel, H. (1964). Studies of the Routine Grounds of Everyday Activities. Social Problems, 11(3), 225-250.
Garfinkel, H. (1967). Studies in ethnomethodology. Englewood Cliffs, New Jersey: Prentice-Halll.
Garfinkel, H., \& Wieder, D. L. (1992). Two incommensurable, asymmetrically alternate technologies of social analysis. In G. Watson \& R. M. Seiler (Eds.), Text in context: Contributions to ethnomethodology (pp. 175-206). Newbury Park, CA: Sage New York.
Glaser, B. G., \& Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. New York, NY: Aldine de Gruyter.
Haddington, P., Mondada, L., \& Nevile, M. (2013). Interaction and mobility: Language and the body in motion (P. Haddington, L. Mondada, \& M. Nevile, Eds.). Berlin, Boston: De Gruyter.
Harris, J., Theobald, M. A., Danby, S. J., Reynolds, E., \& Rintel, S. (2012). "What's going on here?" The pedagogy
of a data analysis session. In A. Lee \& S. J. Danby (Eds.), Reshaping doctoral education: International approaches and pedagogies (pp. 83-96). London: Routledge.
Heath, C., Hindmarsh, J., \& Luff, P. (2010). Video in qualitative research: analysing social interaction in everyday life. London: Sage Publications.
Hepburn, A., \& Bolden, G. B. (2012). The Conversation Analytic Approach to Transcription. In J. Sidnell \& T. Stivers (Eds.), The Handbook of Conversation Analysis (pp. 5776). Oxford: John Wiley \& Sons.

Heritage, J. (1984). Garfinkel and ethnomethodology. Cambridge: Polity Press.
Heritage, J. (2008). Conversation analysis as social theory. In Bryan Turner (Ed.), The new Blackwell companion to social theory (pp. 300-320). London: Blackwell.
Hindmarsh, J. (2012). Heath's natural habitat: The data session. In P. Luff, J. Hindmarsh, D. vom Lehn, \& B. Schnettler (Eds.), Work, interaction and technology: A festschrift for christian heath. (pp. 21-23). London: Dept. of Management, Kings College London.
Jebb, A. T., Parrigon, S., \& Woo, S. (2016, aug). Exploratory data analysis as a foundation of inductive research. Human Resource Management Review.
John, L. K., Loewenstein, G., \& Prelec, D. (2012, may). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23(5), 524-532.
Jordan, B., \& Henderson, A. (1995). Interaction analysis: Foundations and practice. The journal of the learning sciences, 4(1), 39-103.
Kendrick, K. H. (2017, jan). Using conversation analysis in the lab. Research on Language and Social Interaction, 1-11.
Kerr, N. L. (1998, aug). HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review, 2(3), 196-217.
Labov, W. (1972). Sociolinguistic patterns. University of Pennsylvania Press, Incorporated.
Levinson, S. C. (1983). Pragmatics. Cambridge: Cambridge University Press.
Loth, S., Huth, K., \& De Ruiter, J. P. (2013). Automatic detection of service initiation signals used in bars. Frontiers in Psychology, 4.
Lynch, M. (2012). Revisiting the cultural dope. Human Studies, 35(2), 223-233.
Ochs, E. (1979). Transcription as theory. In E. Ochs \& B. B. Schieffelin (Eds.), Developmental pragmatics (pp. 43-72). New York: Academic Press.
Open Science Collaboration. (2015, August). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716-aac4716.
Pashler, H., \& Harris, C. R. (2012, \#nov\#). Is the replicability crisis overblown? three arguments examined. Perspectives on Psychological Science, 7(6), 531-536.
Pashler, H., \& Wagenmakers, E. (2012, nov). Editors' intro-
duction to the special section on replicability in psychological science. Perspectives on Psychological Science, 7(6), 528-530.
Potter, J. (2002, aug). Two kinds of natural. Discourse Studies, 4(4), 539-542.
Raudaskoski, S. (2009). Tool and machine: The affordances of the mobile phone (Unpublished doctoral dissertation).
Sacks, H. (1987). On the preferences for agreement and contiguity in sequences in conversation. In G. Button \& J. Lee (Eds.), Talk and social organization (pp. 54-69). Clevedon: Multilingual Matters.
Sacks, H., Schegloff, E. A., \& Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50(4), 696-735.
Schegloff, E. A. (1968, \#dec\#). Sequencing in Conversational Openings. American Anthropologist, 70(6), 1075-1095.
Schegloff, E. A. (1993). Reflections on Quantification in the Study of Conversation. Research on Language \& Social Interaction, 26(1), 99-128.
Schegloff, E. A. (1996). Confirming allusions: Toward an empirical account of action. American Journal of Sociology, 102(1), 161-216.
Schegloff, E. A. (2006). On possibles. Discourse Studies, 8(1), 141157.
Schegloff, E. A. (2007). Sequence organization in interaction: Volume 1: A primer in conversation analysis. Cambridge: Cambridge University Press.
Schegloff, E. A., Jefferson, G., \& Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. Language, 53(2), 361-382.
Sidnell, J., \& Stivers, T. (2012). The Handbook of Conversation Analysis. Oxford: John Wiley \& Sons.
Stivers, T. (2015, \#jan\#). Coding social interaction: A heretical approach in conversation analysis? Research on Language and Social Interaction, 48(1), 119.
Stivers, T., \& Enfield, N. J. (2010, \#oct\#). A coding scheme for question-response sequences in conversation. Journal of Pragmatics, 42(10), 2620-2626.
Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... Levinson, S. C. (2009, \#jun\#). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences of the United States of America, 106(26), 10587-92.
Ten Have, P. (1999). Doing conversation analysis: A Practical Guide (1st ed.). London: Sage Publications.
Toomela, A. (2014). Mainstream psychology. In T. Teo (Ed.), Encyclopedia of critical psychology (1st ed., pp. 11171125). New York: Springer-Verlag.

# Segmentation as Retention and Recognition: the R\&R model 

Raquel G. Alhama (rgalhama@uva.nl) Willem Zuidema (zuidema@uva.nl)<br>Institute for Logic, Language and Computation, Science Park 107<br>Amsterdam, 1098XG, The Netherlands


#### Abstract

We present the Retention and Recognition model (R\&R), a probabilistic exemplar model that accounts for segmentation in Artificial Language Learning experiments. We show that R\&R provides an excellent fit to human responses in three segmentation experiments with adults (Frank et al., 2010), outperforming existing models. Additionally, we analyze the results of the simulations and propose alternative explanations for the experimental findings.


Keywords: artificial language learning; segmentation; statistical learning; cognitive modelling

## Introduction

A crucial step in the acquisition of a spoken language is to discover what the building blocks of a speech stream are. Children perform such segmentation by exploiting a variety of statistical and prosodic cues in the input. Understanding the unique ability of humans to acquire speech requires an understanding of the nature of this learning mechanism.

Artificial Language Learning (ALL henceforth) has, over the last 20 years, become a key paradigm to study the nature of learning biases in speech segmentation and rule generalization. In experiments in this paradigm, participants are exposed to artificial stimuli designed to incorporate particular aspects of speech and language, and they are subsequently tested on whether and under which conditions they discover the regularities in such artificial language.

A key result in this tradition is the demonstration that 8 month old infants are sensitive to transition probabilities between syllables, and can segment a speech stream based on these probabilities alone (Saffran, Aslin, and Newport (1996), Aslin, Saffran, and Newport (1998)). This ability to track statistics over concrete fragments of the input, known in the literature as statistical learning, has also been demonstrated in adults (Saffran, Newport, \& Aslin, 1996).

However, these experiments do not reveal whether the underlying cognitive mechanism does operate over transitional probabilities or, instead, it performs computations of an entirely different nature but which can be described as transitional probabilities. In order to reveal the precise underpinnings of such cognitive mechanism, a useful methodology is computational modeling.

There exist several segmentation models in the literature, offering alternative accounts of the nature of this process. Thus, these models need to be compared and analyzed against empirical data to validate their predictions. Possibly the most comprehensive study for the evaluation of computational models in segmentation is presented in Frank et al.
(2010). In that study, the authors evaluate a range of models based on their goodness of fit to three segmentation experiments that involve a great number of different conditions -thus providing a rich dataset for comparing the models.

In this paper we present one model for to account for segmentation experiments in ALL. Our model, called the Retention \& Recognition model (henceforth R\&R), is a novel processing model that explains segmentation based on the retention and recognition of subsequences of the input. Following Frank et al., we test our model against the experimental data from their study, and compare the goodness of fit of our model with those reported in previous studies.

## The R\&R Model

The model we propose, which we call the RetentionRecognition Model (R\&R), takes a sequence of syllables $X=$ $\left\langle x_{o}, x_{1}, x_{2}, \ldots, x_{m}\right\rangle$ as input, and considers all subsequences of length $l=1,2, \ldots, l_{\max }$ as potential segments to be memorized.

The model maintains a memory $M$, which is a set of segment types and their associated counts. The memory is initially empty ( $M_{0}=\emptyset$ ) and it changes with update steps that either add an entry (with count 1) or increase the count of an existing entry:

ADD: $\quad M_{t+1} \leftarrow M_{t} \cup\left\{\left\langle\left\langle x_{j}, \ldots, x_{k}\right\rangle, 1\right\rangle\right\}$

## INCREMENT:

$M_{t+1} \leftarrow M_{t}-\left\{\left\langle\left\langle x_{j}, \ldots, x_{k}\right\rangle, c\right\rangle\right\} \cup\left\{\left\langle\left\langle x_{j}, \ldots, x_{k}\right\rangle, c+1\right\rangle\right\}$
For any candidate segment $s \in S$ (with segments processed in the order they are encountered in the stream), the model checks whether it is stored in memory and, if so, what the count of that segment in memory is (its 'subjective frequency'). The model may (with a probability $p_{1}$ that increases with that count) recognize it (i.e., match it with a segment in memory). If it succeeds, the count is incremented with 1. If it fails to recognize the segment, the model might (with a probability $p_{2}$ that decreases with the length of the segment) still retain it (i.e., add it to memory with initial count of 1 if it was not stored, or in the event that a previously stored segment was not recognized and is retained -very rare in practise- increase the count by 1 as a form of 'late recognition'). In this way, the model builds a memory of segments that have different degrees of familiarity depending on their distribution in the stream. R\&R's flowchart is given in Figure 2.

The key components of the model are the equations for
computing the recognition probability $\left(p_{1}\right)$ and retention probability $\left(p_{2}\right)$. Recognition should become more probable the more often a segment has been recognized, but decrease with the number of segment types in memory $(|M|)$. Hence, we define $p_{1}$ as follows, with $B$ and $D$ free parameters $(0 \leqslant B, D \leqslant 1)$ that can be fitted to the data:

$$
\begin{equation*}
p_{1}(s, M)=\left(1-B^{\operatorname{COUNT}(s, M)}\right) \cdot D^{|M|} \tag{1}
\end{equation*}
$$

If a segment is not recognized, the model considers retaining it with a probability that decreases with the length of the segment $(l(s))$, and which can be boosted if there are additional cues favoring this segment (e.g., a pause preceding it). Hence, we define $p_{2}$ as follows, with $A$ and $\mu$ free parameters $(0 \leqslant A \leqslant 1 ; 0 \leqslant \mu)$ that can be fitted to the data:

$$
p_{2}(s)=A^{\operatorname{length}(s) \cdot \mu} \quad, \text { where } \mu=\left\{\begin{array}{l}
\mu_{w p} \text { after a pause }  \tag{2}\\
\mu_{n p} \text { otherwise }
\end{array}\right.
$$

The $A$ parameter thus describes how quickly the retention probability decreases with the length of a segment. The probability is also affected by the presence of additional cues; in this paper, we consider only the pauses between sentences as additional cues. ${ }^{1}$

Putting everything together, the model can be described in pseudocode as in Figure 1. As can be seen, R\&R is a simple model, but it gives a surprisingly accurate match with empirical data, as we will explore in the next sections, without even taking processes such as forgetting, priming, interference and generalization into account.

## Related Models

There exist several models of segmentation in the literature. We do not have the space to address them all here, but we discuss how our model relates to those to which it has more similarities.

The recognition component of our model yields rich-getricher dynamics (and thus consistently produces very skewed count distributions over segments in memory) similar to that of non-parametric Bayesian models, such as the Bayesian Lexical Model (BLM henceforth) in Goldwater, Griffiths, and Johnson (2009) (adapted for ALL in Frank et al. (2010)). The BLM implements such dynamics with a Dirichlet process. The main assumptions of this process are: (i) the probability of a word in the $i^{\text {th }}$ position is proportional to the number of occurrences of this word in previous positions; (ii) the

[^262]```
Input: Stream \(X\), and empty memory \(M_{0} \leftarrow 0\).
Output: Memory \(M_{n+1}\).
/* Compute candidate segments: */
\(S \leftarrow\left\langle s_{0}, s_{1}, \ldots, s_{n}\right\rangle\)
/* Process each segment: */
for \(i=0\) to \(n\) :
    /* Compute the recognition probability: */
    \(p_{1}=p_{1}\left(s_{i}, M_{i}\right)\)
    /* Compute the retention probability: */
    \(p_{2}=p_{2}\left(s_{i}, M_{i}\right)\)
    /* Draw two random numbers */
    \(r_{1} \sim \mathcal{U}(0,1)\)
    \(r_{2} \sim \mathcal{U}(0,1)\)
    /* Recognize, retain or ignore: */
    IF \(\left(r_{1}<p_{1}\right)\)
        \(M_{i+1} \leftarrow \operatorname{increment}\left(s_{i}, M_{i}\right)\)
    ELSE IF \(\left(r_{2}<p_{2}\right)\)
        \(M_{i+1} \leftarrow \operatorname{add}\left(s_{i}, M_{i}\right)\)
    ELSE
        \(M_{i+1} \leftarrow M_{i}\)
```

Figure 1: Pseudocode describing the R\&R model.
relative probability for a new word type in the $i^{t h}$ position is inversely correlated with the total number of word tokens, and (iii) a new word type is more probable if it is shorter. Assumption (ii) does not allow for direct comparison, since $R \& R$ is not a generative model, and therefore it does not provide a probability for new types -rather, the incorporation of new types to the memory of the model depends on the retention probability, and it is based on a preference for shorter sequences (an intuition encoded also in assumption (iii) of the Bayesian model). As for assumption (i), the same principle is incorporated in the recognition process in $R \& R$; however, in our model, the counts of the number of occurrences of a word is based on the subjective frequencies resulting from memorization, while in the BLM, these counts are based on absolute frequencies of the current hypothesis. This reflects a fundamental difference between the two approaches, which concerns their level of analysis (Marr, 1982). The Bayesian model is framed at Marr's computational level, and thus, it operates over the whole stimuli, since it does not incorporate perceptual or memory constraints (although some of the extensions in Frank et al. (2010) experiment with limitations on memory capacity, leading to a somewhat hybrid model; we return to this point later). In other words, the BLM is not proposed as a mechanistic explanation of the cognitive processes involved in the experiment; on the contrary, $R \& R$ is a processing model, which postulates that cognitive processes of retention and recognition, and psychological representations of exemplar segments are responsible for segmentation.

An existing model that is also pitched at Marr's processing level is PARSER (Perruchet \& Vinter, 1998). PARSER is a symbolic model, built around basic principles of associative learning and chunking, that shares many similarities with R\&R. Both PARSER and R\&R are exemplar-based models that build a lexicon of segments (exemplars), and use this


Figure 2: R\&R:The Retention-Recognition Model
lexicon of already-memorized segments to decide on further segments to memorize. Each segment in the lexicon is stored together with a score that determines the impact of this segment in the next steps of the segmentation process. Thus, the models are similar in their procedure, but there are notable differences between them. One of them is the probabilistic nature of their components. For PARSER, the stochasticity is limited to the random selection of the size of the next segment to read from the stream. In contrast, $R \& R$ considers all possible subsequences of the stream (up to a maximum length), as inspired by research in Data-Oriented Parsing tradition (Scha (1990), Zuidema (2006)). Additionally, the model is inherently probabilistic in its basic processes of retention and recognition.

There exist other differences in the procedure of these approaches. To begin with, the process of retention in R\&R penalizes longest segments, on the basis that they would require more working memory. However, PARSER is a chunking model, so it implements the opposite principle: whenever several segment candidates are possible, it selects those that are built of the longest units, creating in this way a bias for larger units. As for the process of recognition, it is implicitly implemented in PARSER when it maps the next segment to be read against the units in memory. This process involves a binary threshold: only units with weight above the threshold can be recognized as components of the segment (but those below the threshold are retained). In contrast, the interaction between recognition and retention in $R \& R$ is based on a graded probabilistic choice. Finally, an important difference between the models is that $R \& R$ does not implement any form of forgetting. Although we do not claim that humans are endowed with perfect memory, our results suggest that forgetting does not seem to play a key role in the timecourse of the experiments.

On the other extreme, at Marr's implementational level, we find TRACX (French et al., 2011; French \& Cottrell, 2014), a connectionist proposal that is also based on the recogni-
tion of subsequences. TRACX is an autoencoder model that learns a representation for the input data. The error of the output layer is computed by comparing it with the input, and it serves as an indication of the degree of recognition of the input. The model processes the input stream sequentially, maintaining a context window. After successful recognition of a segment, the internal representation learned by the network is used as the context for the next segment to be presented. In this way, contiguous segments that are successfully recognized are gradually represented as a single chunk, and therefore can be recognized as a unit. This approach shares with R\&R the intuition that words are consolidated in memory after repeated recognition; however, like PARSER, TRACX is a chunking model, that is, it is oriented to the integration of syllables in order to build larger fragments. In contrast, in $R \& R$, words emerge in a process that actually penalizes larger fragments, as a consequence of consolidated memorization of statistically salient segments.

To sum up, R\&R constitutes a new approach to modelling segmentation that offers a processing level explanation of the identification of words in a speech stream, which emerges as a result of the interplay between probabilistic memory processes. We now proceed to validate this model against empirical data.

## Fitting R\&R to Experimental Data

## Experimental Results

Frank et al. investigate how distributional aspects of an artificial language have an effect on the performance of human adults in segmentation. Each of their three experiments involves a range of conditions that vary in one particular dimension: (i) sentence length, (ii) amount of exposure (number of tokens) and (iii) vocabulary size (number of word types).

The stimuli consists of an auditory sequence of sentences, each of which is created from a sample of artificial (unexisting) words. The sentences are separated with a silence gap
of 500 ms , while there is no acoustic nor prosodic cue indicating the separation between words within a sentence. After the participants have been exposed to a sample of sentences thus constructed, they participate in a 2-Alternative-ForcedChoice test (2AFC). The two alternatives in the test consist on one word from the artificial language (a correctly segmented sequence), and one "part-word" (a sequence resulting from incorrect segmentation).

To analyze the results, the mean number of correct choices is computed across participants in each condition. The curves formed by these datapoints (ordered by condition value) is taken as indication of how segmentation performance is affected by the varied dimension. These curves (which are shown in the continuous line in Figure 3) show that: (i) human adults have more difficulty in segmenting words when sentences are longer, presumably because they do not benefit from the extra cue provided by the silence gaps; (ii) when the amount of word tokens is varied, more occurrences of words facilitate the identification of such words, and (iii) the size of the vocabulary seems to cause lower performance in the experiment, with an almost-linear inverse relation.

## Goodness of fit

The study by Frank et al. evaluates a number of segmentation models in terms of their goodness of fit to the curve that describes the average performance of the human subjects. The evaluated models include the ones previously described (BLM, PARSER, and later, also TRACX, reported in French et al. (2011)), and four additional approaches, all of them consisting on normative models: Transitional Probabilities (TP), a Bayesian version of TP (by Frank et al.), Mutual Information (MI), and a version of MI model that identifies words when they exceed a threshold both on MI and raw frequency counts (MI Clustering, Swingley (2005)).

In order to compare the models, Frank and colleagues convert the output of each model to a metric that can be interpreted as behavioural predictions for the 2AFC task. To do so, they employ the Luce Rule (Luce, 1963). Given a pair of sequences $s_{1}$ and $s_{2}$ in test, the Luce Rule defines the probability of choosing $s_{1}$ as can be seen in Equation 3:

$$
\begin{equation*}
P\left(s_{1}\right)=\frac{\operatorname{SubjFreq}\left(s_{1}\right)}{\operatorname{SubjFreq}\left(s_{1}\right)+\operatorname{SubjFreq}\left(s_{2}\right)} \tag{3}
\end{equation*}
$$

Once the scores have been transformed to probabilities, the performance of the models is computed as the mean probability of choosing the correct item, averaged over participants and test trials. These datapoints are arranged in a curve in the same way as with human participants, and the correlation in the shape of these curves -measured with Pearson's r- is taken as an indication of good fit.

Likewise, we run simulations of the three experiments with $R \& R$, transforming its output (the subjective frequencies) into test trials with the Luce Rule. We run a search over the parameter space, in order to find which parameters yield


Figure 3: Curve of performance for all conditions in the experiments in Frank et al. (2010).
best correlation with human performance ${ }^{23}$. The best results are shown in Table 1. As it can be seen, our model outper-

[^263]Table 1: Comparison of model results to human performance. The reported metric is Pearson's r. ${ }^{*}$ Experiment 2 was not reported in French et al. (2011). Therefore, the mean can be taken to be 0.63 (for a Pearson's $r$ of 0.0 in experiment 2) or 0.945 (averaging only over experiments 1 and 3 ).

|  |  | Exp. 1: <br> Sentence Length | Exp. 2: <br> Amount of tokens | Exp. 3: <br> Word types | Mean |
| ---: | :--- | :---: | :---: | :---: | :---: |
| 1 | Transitional Probabilities | 0.84 | 0.43 | -0.99 | 0.09 |
| 2 | Mutual Information | 0.83 | -0.32 | -0.99 | -0.16 |
| 3 | MI Clustering | 0.11 | -0.81 | 0.29 | -0.13 |
| 4 | PARSER | 0.00 | 0.86 | 0.00 | 0.28 |
| 5 | TRACX | 0.92 | - | 0.97 | $-{ }^{*}$ |
| 6 | BLM | 0.94 | 0.89 | -0.98 | 0.28 |
| 7 | Bayesian TPs 4\% data | 0.82 | 0.92 | 0.96 | 0.90 |
| 8 | BLM 4\% data | 0.88 | 0.85 | 0.90 | 0.87 |
| 9 | BLM Uniform forgetting (types) | 0.95 | 0.92 | 0.73 | 0.86 |
| 10 | BLM Prop. forgetting (types) | 0.88 | 0.87 | 0.88 | 0.87 |
| 11 | BLM Uniform forgetting (tokens) | 0.86 | 0.82 | 0.97 | 0.88 |
| 12 | R\&R | $\mathbf{0 . 9 8}$ | $\mathbf{0 . 9 4}$ | $\mathbf{0 . 9 8}$ | $\mathbf{0 . 9 7}$ |

forms all the other models in the three experiments, with a parameter setting that is common to the three experiments ( $A=0.008, B=0.923, D=0.866, \mu_{n p}=1.0, \mu_{w p}=0.234$ ). The curves of the performance of both human adults and R\&R can be see in Figure 3.

When it comes to experiment 1 , one possible explanation for this result is that $R \& R$ is the only model that explicitly models the effect of the silence gaps. By increasing the length of sentences while keeping the number of types and tokens constant, the stimuli necessarily consists of fewer sentences when those are made longer; therefore, the number of silence gaps also decreases. For this reason, the performance of $R \& R$ declines with longer sentences, since it cannot obtain the same benefit from exploiting silence gaps. This explanation for the superior performance can be supported by looking at the values of the $\mu_{w p}$ parameter: the best fit of the model requires a low value for this parameter ( $\mu_{w p}=0.234$ ) , so in the presence of a pause it substantially boosts the otherwise very small $\left(A^{\mu_{n p}}=0.008\right)$ retention probability.

In the second experiment, normative models based on point estimates (those based on TP and MI) do not offer a good fit with the data, since those metrics do not benefit from the accumulation of evidence offered by the increased number of tokens (contrary to humans). Frank et al. suggest that humans may be forgetting much of what they hear, which would explain the increased performance with the number of tokens. However, the extended versions of the BLM that incorporate some form of evidence limitation (with input data restricted to a random $4 \%$ sample) or forgetting exhibit mixed results (rows $8,9,10,11$ on table 1). Moreover, these extensions appear unrealistic from a cognitive perspective (e.g. one of the extensions forgets a random token when the memory capacity is full), and additionally, the resulting models are somewhat difficult to interpret, since after incorporating memory limitations, they are not computational level approaches anymore.

PARSER offers a more intuitive account of forgetting, with modest correlation with human data; however, this model has zero correlation in the other experiments. So this pattern of results suggests that a rich-get-richer form of recognition combined with a process of retention as defined in R\&R seems a more compelling explanation than a process of recognition with forgetting.

Also on experiment 3, the $\mathrm{R} \& \mathrm{R}$ model exhibits the best correlation with human data, followed closely by TRACX. Again, normative models show the opposite trend from humans (rows 1, 2, 3, 6 on table 1), since they do not have any memory limitations, and thus the effect of increasing vocabulary size only has an effect in the distributional properties of the stream, which result in less statistically coherent partwords. This is the case also for PARSER and the BLM. Frank et al. attribute this failure to the lack of forgetting in the models, but the same issues we have discussed above apply to this experiment. Therefore, the more convincing approaches are TRACX and R\&R. But although TRACX naturally reproduces the human results without forgetting, it is difficult to interpret what is the component of the model that is responsible for its success in this experiment. Conversely, R\&R explicitly incorporates a parameter that penalizes recognition based on the number of memorized types. In line with our intuitions, the corresponding parameter value for the best fit amounts to $D=0.86$, which results in a relatively large penalization for recognition ${ }^{4}$. Therefore, in conditions of high number of types, humans have an increased difficulty in recognizing sequences, most likely originating from the process of matching the input segment to one of the many segments stored in memory.

[^264]
## Discussion

With our model, R\&R, we provide a theory of the process of segmentation based on the interaction of two cognitive mechanisms of memorization. We believe that one of the best features of our model is its transparency: pitched at the processing level, and with a very simple formalization that involves clearly identified components, $R \& R$ allows for straightforward interpretation of the results. Even though, for reasons of space, we have not been able to report a thorough analysis of the behaviour of the model under different parameter settings, we have shown a glimpse on how these parameters allow for the identification of the relative importance of each component.

This study shows that our model can fit 2AFC data on human adults with a correlation that is at least on par with that of other models. Even though we consider that the evaluation data and procedure initiated by Frank et al. is one of the most thorough in the ALL modelling literature, in Alhama et al. (2015) we argue that averaging the responses over stimuli classes is likely to mask important differences between otherwise seemingly equivalent models. The work reported in this paper is a necessary first step to confirm that $R \& R$ is comparable to other models, but for future work it is important to move to evaluating models based on response distributions over individual test items (albeit our first attempts to evaluate our model with this procedure are inconclusive), and replace the Luce choice rule and correlation metric with a more cognitively realistic response model.

Finally, segmentation is a fundamental ability for language learners, but any segmentation model must at some point be related to other cognitive mechanisms that operate in natural and artificial language learning. In Alhama and Zuidema (2016) we show that the subjective frequencies computed by R\&R have the necessary distributional properties to explain some of the main results in rule learning in ALL. Future work may explore how the model relates to other linguistic processes (e.g. word learning), so that we can eventually achieve a complete understanding of how segmentation relates to the complete picture of language learning.

## Acknowledgments

This work was developed with Remko Scha, who sadly passed away before the finalization of this paper. We are grateful to Andreea Geambasu, Clara Levelt, Michelle Spierings, Carel ten Cate and Padraic Monaghan for their feedback. This research was funded by a grant from the Netherlands Organisation for Scientific Research (NWO), Division of Humanities, to Levelt, ten Cate and Zuidema (360-70450).

## References

Alhama, R. G., Scha, R., \& Zuidema, W. (2015). How should we evaluate models of segmentation in artificial language learning? In Proceedings of 13th International Conference on Cognitive Modeling.

Alhama, R. G., Scha, R., \& Zuidema, W. (2016). Memorization of sequence-segments by humans and non-human animals: the Retention-Recognition model. ILLC Prepublications, PP-2016-08.
Alhama, R. G., \& Zuidema, W. (2016). Generalization in Artificial Language Learning: Modelling the propensity to generalize. In Proceedings of the 7th workshop on Cognitive Aspects of Computational Language Learning (pp. 64-72). Berlin: Association for Computational Linguistics.
Aslin, R. N., Saffran, J. R., \& Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. Psychological Science, 9(4), 321-324.
Frank, M. C., Goldwater, S., Griffiths, T. L., \& Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. Cognition, 117(2), 107-125.
French, R. M., Addyman, C., \& Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. Psychological Review, 118(4), 614.
French, R. M., \& Cottrell, G. W. (2014). TRACX 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction. In Proceedings of the 36th annual conference of the cognitive science society.
Goldwater, S., Griffiths, T. L., \& Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. Cognition, 112, 21-54.
Luce, R. D. (1963). Detection and recognition. In Handbook of mathematical psychology. New York: Wiley.
Marr, D. (1982). Vision. A computational investigation into the human representation and processing of visual information. New York: W. H. Freeman.
Perruchet, P., \& Vinter, A. (1998). PARSER: A model for word segmentation. Journal of Memory and Language, 39(2), 246-263.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8 -month-old infants. Science, 274(5294), 1926-1928.
Saffran, J. R., Newport, E. L., \& Aslin, R. N. (1996). Word segmentation: the role of distributional cues. Journal of Memory and Language, 35(4), 606-621.
Scha, R. (1990). Taaltheorie en taaltechnologie; competence en performance. In R. de Kort \& G. Leerdam (Eds.), Computertoepassingen in de neerlandistiek (pp. 722). Almere, the Netherlands: LVVN. (English translation at http://iaaa.nl/rs/LeerdamE.html.)
Swingley, D. (2005). Statistical clustering and the contents of infant vocabulary. Cognitive Psychology, 50(1), 86-132.
Zuidema, W. (2006). What are the productive units of natural language grammar?: a DOP approach to the automatic identification of constructions. In Proceedings of the Tenth Conference on Computational Natural Language Learning (pp. 29-36).

# Phonological features in the bilingual lexicon: Insights from tonal accent in Swedish 

Nadja Althaus<br>University of East Anglia, Norwich, United Kingdom

Allison Wetterlin<br>University of Agder, Kristiansand, Norway<br>Aditi Lahiri<br>University of Oxford, Oxford, United Kingdom


#### Abstract

Scandinavian languages like Swedish employ tonal accent as a lexical phonological feature, where suprasegmental information can be the sole factor differentiating between words. Using cross-modal semantic fragment priming we tested the following: (a) Do monolingual speakers of Swedish use tonal accent information during lexical access? (b) Do bilingual speakers, who grew up with one tonal (Swedish) and one non-tonal language, treat this feature the same way as monolinguals? Our results show that for monolinguals, accent mispronunciations eliminate priming effects, implying that tone is used during lexical access. For bilinguals, by contrast, mispronunciation sensitivity depends on both the accent type and its distribution across the linguistic input, as well as on the lexical neighbourhood.


# Semantic Networks Generated from Early Linguistic Input 

Andrei Amatuni, Elika Bergelson<br>Department of Psychology and Neuroscience, Duke University<br>417 Chapel Dr. Durham, NC 27708 USA<br>\{andrei.amatuni, elika.bergelson\} @duke.edu


#### Abstract

Semantic networks generated from different word corpora show common structural characteristics, including high degrees of clustering, short average path lengths, and scale free degree distributions. Previous research has disagreed about whether these features emerge from internally- or externallydriven properties (i.e. words already in the lexicon vs. regularities in the external world), mapping onto preferential attachment and preferential acquisition accounts, respectively (Steyvers \& Tenenbaum, 2005; Hills, Maouene, Maouene, Sheya, \& Smith, 2009). Such accounts suggest that inherent semantic structure shapes new lexical growth. Here we extend previous work by creating semantic networks using the SEEDLingS corpus, a newly collected corpus of linguistic input to infants. Using a recently developed LSA-like approach (GLoVe vectors), we confirm the presence of previously reported structural characteristics, but only in certain ranges of semantic similarity space. Our results confirm the robustness of certain aspects of network organization, and provide novel evidence in support of preferential acquisition accounts.


Keywords: semantic networks; word learning; preferential acquisition

## Introduction

A word functions as an atomic unit of meaning, in principle carrying independent semantic content. In practice though, it occurs with its fellow words, as humans produce language. From this word-stream, infants begin to understand words by 6-9 months (Bergelson \& Swingley, 2012), and to produce them soon thereafter. Here we aim to shed light on how these semantic atoms are organized in the mental lexicon, and the degree to which this representational structure is reflective of the conceptual order found "out there" in the world.

To explore this, we turn to semantic networks, an idea dating back nearly a century (Trier, 1931). Given that words are related along semantic dimensions, characterizing these relations is a first step towards understanding their representational structure. Previous research on semantic networks generated from word corpora have shown small-world connectivity (i.e. any given word node is not very many nodes away from any other), as well as scale free degree distributions (i.e. a few nodes serve as 'hubs', and node distribution follows a power law such that probability $(k) \approx k^{-\alpha}$, for a node with degree $k$, and scaling parameter $\alpha$ ) (Sigman \& Cecchi, 2002; Steyvers \& Tenenbaum, 2005; Hills et al., 2009). This suggests that semantic information may be inherently structured in nonrandom, clustered, and highly organized ways, which internal representations may mirror or exploit ${ }^{1}$ (Todd, Hills, \& Robbins, 2012). Scale invariance

[^265](here equivalent to scale-free distributions), has been found in many cognitive domains and diverse natural phenomena; it is argued to be a general unifying principle of cognitive organization (Kello et al., 2010).

Barabási and Albert (1999) suggest that graphs with degree distributions that follow power laws imply constraints on the processes which formed them. Their model for generating such networks relies on incremental growth and a process of "preferential attachment" (hereafter PAT), whereby existing nodes with many connections are preferentially "chosen" by new nodes. While their resulting graphs display power law degree distributions, they did not find small world connectivity of the kind found in semantic networks, such as those generated from WordNet and Roget's Thesaurus. Building on this, Steyvers and Tenenbaum (2005) proposed a model for incrementally growing semantic networks similar to Barabási and Albert (1999), which indeed resulted in both small world and scale free structures. Their growth process centered on semantic differentiation, i.e. new words that are more contrastive with existing words are preferentially incorporated into the graph; they include a frequency parameter as well. The resulting semantic graphs showed degree distributions which reflected the relative time at which a particular node was added to the network: age-of-acquisition (AoA) norms for words corresponded to the relative number of connections in these graphs. PAT-based graphs inherently bias nodes which are added earlier to have higher degree.

Steyvers and Tenenbaum (2005) suggest that the structure of internal representations guides the selection process of new words or concepts. In contrast, Hills et al. (2009) propose that the connectivity of words in the external environment plays a guiding role in the acquisition of new words. In this alternative, dubbed preferential acquisition (hereafter PAQ), the relative salience between unlearned words directs new node integration into the lexicon. Under PAQ, the structure of the external semantic ground is itself scale-free, clustered, and small world, leading internal representations to mirror this structure as lexical items are added. This contrasts with PAT, which suggests that the structuring is a consequence of incremental semantic network growth. Under PAQ, the higher a word's contextual variety, the more interactions it has with other elements in the external ground. This results in more neighbors in semantic network space, making it more linguistically salient to the learner. Indeed, evidence by Hills, Maouene, Riordan, and Smith (2010) suggests a role for contextual variety and associative density in noun lexical development in particular. In the present work, we build on these previous results, combining approaches that suggest network
properties arise from incremental generative processes with networks that are definitively non-generative, as a window into how external and internal semantic spaces may influence the growing noun lexicon.

One limitation of previous work concerns operationalizing semantic relatedness, generally achieved through handtagged features or word associations (Steyvers \& Tenenbaum, 2005; Hills et al., 2009). Hand-tagged features may not reflect the underlying semantic organization given that they stem from an overt metalinguistic task. Indeed such features do not produce scale free graphs or predict AoA. Word association data lead to directed networks, which may obscure inherent transitivity between word pairs (unless both words have the other as their associative target). It's not clear whether this directedness is inherent in the lexicon. While asymmetry is conspicuous in human similarity judgments (Tversky, 1977), this may be a function of task rather than underlying semantic representations. In the present work, we rely on neither hand-tagged features nor directed associations in building our semantic graphs.

Given the goal of explaining how semantic structure emerges, a further limitation of previous work lies in the constituent word nodes in the semantic networks, which used e.g. free associations, Roget's Thesaurus, and WordNet, rather than child-directed corpora. In this paper, we make use of a new corpus of words from infant-caretaker interactions. This allows us to examine whether scale free distributions, small world connectivity, and links to lexical development trajectories are limited by corpus origin, and thus whether using a full range of concrete nouns children are exposed to in naturalistic settings renders different results.

Here we extend previous work and begin to address these limitations by building ecologically valid semantic graphs of early linguistic input. We use modern vector space methods to calculate undirected semantic relations, resulting in a gradient of networks parameterized by degree of similarity. We limit network nodes to only those which infants' hear and embed them in a space which approximates a common semantic ground shared by infants and adults alike. We also investigate links between word frequency in the corpus, and connectivity rank in network space.

## Present Study

We generate networks using a new model of semantic relatedness: vectors trained with GloVe (Pennington, Socher, \& Manning, 2014). We first determine whether our networks reproduce previously reported small world structure, and scale invariance (i.e. power law distributions). Such structures are consistent with PAT or PAQ. However, only PAT proposes that such structures arise due to incremental growth mechanisms (Barabási \& Albert, 1999). PAT suggests that words already in the internal lexicon guide new word selection: early words have higher degree than later-added words, i.e. new additions "prefer" to attach to words with higher degree. In contrast, PAQ proposes that external network connectivity drives
node addition, suggesting that internal structures mirrors external structure, which may be scale-free, small-world, or not. Because our networks are built using the GloVe vectors, they are, by definition, non-generative and non-incremental: showing scale free and small world behavior in our networks would suggest this structuring might exist without PAT's assumed incremental generative growth processes.

As a proxy for AoA, we make use of parent-reported vocabulary norms from WordBank (Frank, Braginsky, Yurovsky, \& Marchman, 2016), a compilation of the MacArthur-Bates Communicative Developmental Inventory (CDI.) We assume words known by more infants at a given age have been in the lexicon longer. Here we attempt to replicate network structure and AoA correlations originally presented as evidence for PAT, while violating PATs assumption of incremental growth. If successful, it would imply that scale-free structure does not itself depend on PAT.

We also test for evidence of PAQ, by determining whether words that go from being poorly-known to well-known over time have more connections in the externally-based network than those that remain poorly known over time. That is, we test PAQ's proposal that high degree nodes in networks generated from external linguistic input are acquired earlier than lower degree nodes in those same networks. Notably, PAQ models do not depend on power law distributions or scale free behavior, but rather on children selectively integrating salient (more densely connected) words from all possible lexical items they're exposed to. If adult sampling is also inherently biased to those words which have high degree in semantic network space, then we expect too that highly frequent words have higher connectivity relative to all child-directed words. Because our corpus is generated from a large sample of child-directed speech, we can further compare word frequency statistics with degree distributions generated using the same set of words.

## Method

## Data

The SEEDLingS corpus (Bergelson, 2016a, 2016b) comes from home recordings of 44 infants from upstate New York, followed from 6 to 17 months. Each month, a daylong audio recording and hour-long video recording were collected. All videos and 3-10 hours of each audio recording were manually tagged for concrete nouns directed to and/or attended by the child, creating tags of several thousand hours of naturalistic interactions between infants and caregivers. We exclude utterances made by the child, resulting in a final dataset of 4359 unique noun-types (194204 tokens). Plurals and diminutives were consolidated into a "basic level" proxy for word lemmas for each recording. These nouns were used to generate the SEEDLingS-All graphs. We also generate graphs for 6 month recordings alone (1855 types, 29289 tokens; SEEDLingS-6mo) and a $16 / 17$ month combined set (1708 types, 26969 tokens; SEEDLingS-16+17mo), to contrast networks generated from speech to pre-verbal infants
and speech to newly verbal toddlers; the SEEDLingS networks are our model of the external linguistic environment. We generated an additional network (WordBank), using only the 369 nouns on the CDI; this serves as our internal semantic network, given that it only includes words that (some) 16-30-month-olds produce.

As our measure of relative AoA, we used by-word summary data from the online WordBank repositories (Frank et al., 2016). This data includes productive vocabularies for children aged 16 to 30 months (reported productive vocabulary is generally more reliable than reported receptive vocabulary. ${ }^{2}$ We make the assumption that words said by more children at a given age entered the lexicon earlier. Indeed, the age at which a word is produced by $50 \%$ of children, the AoA metric used by Hills et al. (2010), is significantly inversely correlated with the percentage of production at 16,23 , and 30 months ( $r=-0.78,-0.97$, and -0.88 respectively; all $p<0.005$ ). Furthermore, AoA measures correlate with children's elicited naming rates (Morrison, Chappell, \& Ellis, 1997). We use WordBank norms rather than the SEEDLingS infants' own productions, as an independent and extremely large-n ( $n=5450$, for English) estimate of children's knowledge for each word in our networks, removing potential dependencies in our analyses.

## GloVe Vectors

Since our dataset is not tagged with semantic features, and since results with hand-engineered features have been mixed, we chose to follow a method described by Steyvers and Tenenbaum (2005) and use semantic vector space models to generate edges between any nodes above a given similarity threshold. We build on their use of Latent Semantic Analysis (LSA) vectors. In that work, LSA vectors (which along with other geometric methods, are non-incremental) did not generate scale free networks; this result was used to suggest that such approaches are incompatible with incremental growth and PAT. To generate our graphs, we use pre-trained word vectors produced by GloVe, a recently developed algorithm for word embedding (Pennington et al., 2014). Using this algorithm, we can investigate whether we find scale free and small world graphs; if so, the original failure to do so might be LSA-specific, and not a necessary consequence of PAT, as the authors suggested.

GloVe has been demonstrated to have higher performance on many different word similarity tasks compared to word2vec and matrix factorization methods using SVD. Here, we opted to use vectors trained on the Common Crawl corpus with 42 billion tokens, resulting in 300 dimensional vectors for 1.9 million unique words. ${ }^{3}$ In some sense this 'full' dataset provides word similarity proxy based on the target (i.e. adult) meanings the child is acquiring. Further analyses using vectors trained on CHILDES (MacWhinney, 2000),

[^266]displayed analogous and in some cases even stronger patterns than the current results. ${ }^{4}$ This to us suggests consistency in the linguistic manifestation of word meaning (and perhaps their concomitant cognitive processes) at both largeand narrow-sampling scales.

Similar to LSA, GloVe learns vector representations of words from co-occurrence matrices built from large text corpora. It instantiates the distributional hypothesis of linguistics, famously articulated by Firth (1957): "you shall know a word by the company it keeps". Because the GloVe vectors encode co-occurrence statistics derived from natural language, our similarity measures also indicate the degree to which two words share contextual coherence. I.e., the more connections a word has in the semantic network, the more words it shares this coherence with. Given this high dimensional encoding space, we can use a continuous metric of similarity. Iterating through similarity thresholds, we create a gradient of networks to study.

## Generating Semantic Networks

We generate graphs across a range of similarity thresholds ( $\varepsilon$ ). Our similarity measure is the cosine between two GloVe vectors. The cosine function also normalizes for word frequency (to some degree) since dot products are divided by their vector norms. For each corpus, for each word, we calculate $\cos (\theta)$ between it and every other word in the set.

We give an undirected edge between two words if their cosine is above a threshold $\varepsilon$. Since generating each graph is a quadratic operation we normalize the vectors to unit length before calculating cosines. We iterate $\varepsilon$ from 0 to 0.99 (step size $=0.01$ ), generating a graph for each similarity threshold. Further methods of edge generation are left for future research. Our code and IPython notebooks are on Github ${ }^{5,6}$.

## Results and Discussion

## Correlations Between Node Degree and Production

We generated 100 graphs for each corpus, one for each value of $\varepsilon$. We calculated Spearman's rank correlation coefficients between each word's number of connections and productive vocabulary norm (for the 369 CDI nouns), for each network and similarity threshold, at 16,23 and 30 months. Under both PAT and PAQ, we would expect to see that words with more connectivity have higher CDI production rates. Indeed, we find robust and significant correlations between the degree of a word in the network, and the percent of toddlers who produced it, for a range of $\varepsilon$, across corpora and ages; Fig.1.

More specifically, we find similar behavior across all networks, with a global peak in correlation for $\varepsilon=0.12-0.19$. All peak correlation values had Spearman's $\rho=0.43-0.52$, with $p<10^{-5}$, showing consistent behavior across networks and ages. This suggests that both the parent's word choice given

[^267]a child's age, and the child's responsiveness to external semantic density across time are roughly constant. This range of $\varepsilon$ where the correlation is at a maximum is relatively low, allowing very loose semantic associations to result in edges. In Figure 2 we show a subgraph from the SEEDLingS-All network, centered around the node "baby."


Figure 1: Correlation coefficients ( $\rho$ ) between number of edges and CDI production rate across words, as a function of similarity threshold $\varepsilon$ (all $\rho$ are significant at $p<0.005$ ). Color indicates which corpus created the network, shape indicates which month of CDI norms was used to calculate $\rho$. Vertical lines indicate the range in which we find scale free degree distributions ( $0.6-0.73$ ). (For thresholds $\varepsilon>0.75$ there were very few (or no) edges being created, explaining the discontinuity and lack of points towards the end of the scale.)


Figure 2: "Baby" subgraph from the SEEDLingS-All network at the similarity midpoint $(\varepsilon=.5)$, where "baby" has 40 neighbors.

## Power Law Degree Distributions

Surprisingly, at the values of $\varepsilon$ where the correlation is maximal, we did not find power law distributions. We did however find them at higher thresholds: at $\varepsilon=0.68$ we can fit a power law function with $\alpha=3.2 \pm 0.1$, with a $\log$ likelihood ratio in favor of power law over exponential fit ( $R=$ 115.73, $p=2.337 \times 10^{-21}$ ). Indeed, at $\varepsilon=0.63-0.75$ we find power law distributions ( $\alpha=2.39-3.73$ ) characteristic of scale free
networks. At these higher thresholds a word's neighbors are semantically very close, similar to other semantic graphs which have shown scale free distributions (e.g. Roget's Thesaurus), suggesting this property might depend on connections' high semantic proximity. See Figure 3.


Figure 3: Sample networks showing degree distributions with power law behavior ( $\alpha=3.1-3.2$; $\mathrm{SE}=.1, p<0.001$; similar behavior found across networks for $0.60 \leq \varepsilon \leq 0.73$ ). Distributions are plotted on a log-log scale with logarithmic spacing between points, which represent the edges of bins. Power law distributions appear linear on this scale.

PAT models presuppose power law distributions (indeed, PAT was initially proposed after observing scale free distributions in semantic networks, and arguing that this limits the kinds of mechanism which could have created them). We thus further analyze the range of $\varepsilon$ where our networks display power law behavior. Again, networks showing this distribution are critical for PAT (and thus our CDI-based Wordbank networks' proxie of an internal network), but incidental for PAQ, which makes no claims about power-law distributions.

Limiting our focus to these ranges (between the vertical lines in Figure 1,) we see that the degree of a given word in the SEEDLingS-All network has uniformly higher correlation with productive vocabulary norms compared to that same node in the internal (i.e. WordBank) networks. (To be clear, we can only calculate $\rho$ for words we have CDI norms for, but the SEEDLingS networks contain all the nouns infants heard, while the WordBank networks contain only CDI word nodes). This pattern is consistent with PAQ, where more densely connected words in the environment are preferentially incorporated into the learner's lexicon. The correlation between AoA and node degree for the WordBank networks, along with their scale-free organization, suggest that PAT is not a necessary condition for this behavior, since these graphs were generated using GloVe. The presence of these same correlations for our other networks (which serve as a proxy of an externally-based network) in this same range of $\varepsilon$, are also scale-free, and provide new support for PAQ.

This pattern validates our method of generating graphs using GloVe vectors: both the WordBank and our SEEDLingS networks display behavior consistent with previous accounts (i.e. scale free distributions in internal lexical networks
and node degrees correlated with AoA). If anything, the SEEDLingS networks show the predicted structure more strongly, suggesting the nouns infants actually hear may form a better representation than limiting the space to lexically simple, early-learned nouns alone. Our current analysis suggests that scale-free and small world structure can be produced without an incremental growth process, since our graphs were generated using a vector space model (i.e. GloVe). If words are to be incorporated using PAQ, this external structure would necessarily be mirrored in the internal lexical network. However, it remains possible that PAT and PAQ could both be at play during infant lexical development, perhaps with PAQ supplementing PAT by providing a structured sampling space for new word selection.

## Clustering Coefficients and Path Lengths

We next examined whether our semantic networks, based on natural language to children, exhibited two key small-world properties found in previous research: low average pathlength (L), and high clustering coefficients (C). We found that the SEEDLingS networks generally had lower C and higher L than the WordBank networks. In Table 1 we list $C$ and L of each network at their respective peak from Figure 1. We also generated Erdős-Renyi and Watts-Strogratz graphs for comparison (Watts \& Strogatz, 1998; Erdős \& Rényi, 1960); Erdős-Renyi gives us a baseline measure of a comparably sized graph built using a random process, while WattsStrogatz provides a prototypical example of a small world graph, with low L and high C. The SEEDLingS networks clearly showed higher clustering coefficients and smaller average path lengths compared to the Erdős-Renyi graph, and comparable behavior to the Watts-Strogatz graph. This small world organization is indicative of hub structures in the network, where a few very densely connected nodes establish routes between a large proportion of the graph, keeping the average shortest path length low. This is also a defining feature of networks with power law distributions, even though the networks we've listed in Table 1 do not fit that criteria. This small world organization, even in the absence of power law distributions, ${ }^{7}$ supports previous findings in other semantic networks and suggest that even in child-directed natural language input we see these structures.

## AoA as a Function of Frequency and Connectivity

The SEEDLingS corpus contains word frequency counts, a particularly powerful predictor of word acquisition (Goodman, Dale, \& Li, 2008), allowing us to examine the relationship between word frequency and network connectivity. ${ }^{8}$ A positive relationship would suggest that densely connected words are preferentially sampled in adult speech directed towards children. As shown above, more highly connected words were said by more toddlers, across our networks (at peak $\varepsilon$, all $\rho>.51$, all $p<.0001$; see Fig. 1). Using

[^268]| Corpus | $\varepsilon$ | C | L |
| :--- | :--- | :--- | :--- |
| SEEDLingS All | 0.13 | 0.594 | 1.749 |
| SEEDLingS 6 mo. | 0.16 | 0.669 | 1.739 |
| SEEDLingS 16+17 mo. | 0.12 | 0.726 | 1.534 |
| WordBank | 0.13 | 0.895 | 1.202 |
| Erdős-Renyi | - | 0.049 | 1.950 |
| Watts-Strogatz | - | 0.634 | 3.013 |

Table 1: Clustering coefficients (C) and average shortest path lengths (L) of the largest connected subgraph at peak values from Figure 1. Generated Erdős-Renyi $(\mathrm{n}=6404, \mathrm{p}=0.05)$ and Watts-Strogratz graphs ( $\mathrm{n}=6404, \mathrm{k}=64, \mathrm{p}=0.05$ ) are listed for comparison.
model comparison of simple linear models, we find that including both word frequency and node degree as predictors of word production (at 16,23 , and 30 months) accounts for significantly more variance than either alone (all $p<.01$; interaction term significantly improved model fit for months 23 and 30 only, both $p<.01$ ). ${ }^{9}$

Finally, to better understand whether our semantic networks find support for PAT and PAQ models, we tested one specific prediction of each. For PAT, we tested whether words that had been known longer (i.e. by proxy, were said by more children) had more connections than those that had not. Indeed, conducting a median split on words' production rates at each age (16, 23, and 30 months), we find that better known words have higher degree than less well known words ( $p<0.005$ by Wilcoxon Test). For PAQ, we tested whether the words that went from less-well-known to betterknown over 16-30 months had higher degree than those that remained poorly known. Indeed, for words produced below the median rate at 16 months, those below the median at 30 months had significantly lower degree than those above it ( $p<.005$ by Wilcoxon Test.) This supports PAQ's proposal that high degree nodes in networks generated from external linguistic input are acquired earlier than lower degree nodes in those same networks.

## Conclusions

Our results suggest there is inherent semantic structure present in the early linguistic environment, and that both the caregivers and their children are likely sensitive to this nonuniform distribution of semantic information. Because the SEEDLingS corpus provides a uniquely rich dataset of early linguistic input, we were able to construct ecologically valid networks and study differences in their structure across time for a constant set of infants. Our present findings support previous work addressing semantic network structure. Using a modern semantic vector space model to generate our graphs, we were able to confirm the presence of scale free degree distributions in our networks, as well as high clustering

[^269]coefficients and low average path lengths. This method for generating semantic networks avoids the need for hand engineered features and sidesteps the limits of free-association data, providing a potentially more advanced measure of semantic relatedness compared to the original work on PAQ.

That said, the process that generated the GloVe vectors here is not the same as that generating any human's lexicon; further work is needed to strengthen and test links between these representations. Moreover, the GloVe model does not speak to the origin of token distributions in natural language. It does, however, encode a geometric projection of a meshwork of causal substructures present in the external world. Future research will explore the link between these structures and their grounding in cognitive processes. While we have taken a few steps towards examining network growth over time (finding little difference in our 6 mo . and $16+17 \mathrm{mo}$. SEEDLingS networks, or over 16,23 , and 30 mo . CDI norms), more work is needed to better understand not only whether PAT and/or PAQ-compatible processes are at play, but how the interplay between input and uptake changes as the learner grows.

In their original work, Hills et al. (2009) were not able to produce scale free graphs using their hand made features, but were able to do so using adult free association data. In our own graphs we saw that the scale free property only manifested at relatively high values of $\varepsilon$, where only very closely related words (often synonyms) were connecting to each other. Because our measure of similarity was parameterized, we were able to produce a gradient of networks and study their behavior across a range of thresholds, focusing at different ranges of the scale as needed. By generating scale free networks using a non-incremental procedure, we lend support to the hypothesis that this structuring may be an inherent feature in the external environment, rather than a consequence of how it's integrated into internal representations. Building on Firth: our results suggest that words may indeed become known by the company they keep, and that the relevant neighbors may be both those inside the lexicon, and in the as-yet unknown external world of words.

## Acknowledgments

We thank the SEEDLingS team, and NIH DP5-OD019812.

## References

Barabási, A.-L., \& Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.
Bergelson, E. (2016a). Bergelson seedlings homebank corpus.
doi: 10.21415/T5PK6D
Bergelson, E. (2016b). Seedlings corpus. Retrieved 2017-0129, from https://nyu.databrary.org/volume/228
Bergelson, E., \& Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences, 109(9), 3253-3258.

Cohen, R., \& Havlin, S. (2003). Scale-free networks are ultrasmall. Physical review letters, 90(5), 058701.
Erdős, P., \& Rényi, A. (1960). On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5(1), 17-60.
Firth, J. R. (1957). \{A synopsis of linguistic theory, 19301955\}.
Frank, M. C., Braginsky, M., Yurovsky, D., \& Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. Journal of child language.
Goodman, J. C., Dale, P. S., \& Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. Journal of child language, 35(03), 515-531.
Hills, T. T., Maouene, J., Riordan, B., \& Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. Journal of memory and language, 63(3), 259-273.
Hills, T. T., Maouene, M., Maouene, J., Sheya, A., \& Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? Psychological Science, 20(6), 729-739.
Kello, C. T., Brown, G. D., Ferrer-i Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., \& Van Orden, G. C. (2010). Scaling laws in cognitive sciences. Trends in cognitive sciences, 14(5), 223-232.
MacWhinney, B. (2000). The childes project: The database (Vol. 2). Psychology Press.
Morrison, C. M., Chappell, T. D., \& Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. The Quarterly Journal of Experimental Psychology: Section A, 50(3), 528-559.
Pennington, J., Socher, R., \& Manning, C. D. (2014). Glove: Global vectors for word representation. In Empirical methods in natural language processing (emnlp) (pp. 1532-1543). Retrieved from http://www.aclweb.org/ anthology/D14-1162
Sigman, M., \& Cecchi, G. A. (2002). Global organization of the wordnet lexicon. Proceedings of the National Academy of Sciences, 99(3), 1742-1747.
Steyvers, M., \& Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science, 29(1), 4178.

Todd, P. M., Hills, T. T., \& Robbins, T. W. (2012). Cognitive search: Evolution, algorithms, and the brain. MIT press.
Trier, J. (1931). Der deutsche wortschatz im sinnbezirk des verstandes: die geschichte eines sprachlichen feldes. 1. von den anfängen bis zum beginn des 13. jahrhunderts. Winter.
Tversky, A. (1977). Features of similarity. Psychological review, 84(4), 327.
Watts, D. J., \& Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. nature, 393(6684), 440-442.

# Analogical Abstraction in Three-Month-Olds 

Erin M. Anderson (erinanderson2014@u.northwestern.edu) Yin-Juei Chang (alixyj@gmail.com) Susan J. Hespos (hespos@northwestern.edu) Dedre Gentner (gentner@northwestern.edu)<br>Department of Psychology, 2029 Sheridan Rd.<br>Evanston, IL 60208 USA


#### Abstract

This research tests whether analogical processing ability is present in 3-month-old infants. Infants are habituated to a series of analogous pairs, instantiating either same (e.g., AA, BB, etc.) or different (e.g., $\mathrm{AB}, \mathrm{CD}$, etc.), and then tested with further exemplars of the relations. If they can distinguish the familiar relation from the novel relation, even with new objects, this is evidence that for analogical abstraction across the study pairs. In Experiment 1, we did not find evidence of analogical abstraction when 3-month-olds were habituated to six pairs instantiating the relation. However, in Experiment 2, infants showed evidence of analogical abstraction after habituation to two alternating pairs (e.g., AA, BB, AA, BB...). Further, as with older groups, rendering individual objects salient disrupted relational learning. These results demonstrate that 3-month-old infants are capable of analogical comparison and abstraction. Our findings also place limits on the conditions under which these processes are likely to occur. We discuss implications for theories of relational learning.


Keywords: cognitive development, relational processing, infants

## Introduction

Analogical ability - the ability to make relational comparisons between objects, events, or ideas, and to see common relational pattern across different sets of objects - is a cornerstone of higher reasoning abilities. Learning by analogy is a powerful way of acquiring and transferring new information. Equally important, analogical comparison facilitates the formation of abstract categories and rules (Doumas \& Hummel, 2013; Gentner \& Medina, 1998; Gick \& Holyoak, 1983). Indeed, recent theoretical perspectives have asserted that analogical ability is the key capacity supporting higher-order cognition and differentiating human cognitive capacity from that of other primates (Gentner, 2003; 2010; Penn, Holyoak, \& Povinelli, 2008).

The relational abilities of adult humans are astounding. But there are many contributors to the sophistication of adult cognition. Adults have had the benefit of cultural transmission of knowledge - skills, cultural technologies of
various sorts, and symbol systems such as language and mathematics. In addition, adults have broad domain knowledge-another contributor to understanding relations (Gentner, 1988). It is therefore impossible to disentangle the roots/sources of our cognitive power by studying adults. To gain understanding of the nature and origin of our extraordinary relational ability, we must investigate infants who have not yet acquired these resources.

Although little is known about the very early development of human analogical ability, there has been considerable research on the development of analogical ability from preschool to adulthood. Analogical processing shows a relational shift (Gentner, 1988; Gentner \& Toupin, 1986; Halford, 1992; Richland, Morrison, \& Holyoak, 2006) with young children focusing on object matches and older children focusing relational matches and capable of using relational similarity in problem-solving (Chen, 1996). This shift has been attributed to increases in relational knowledge (Gentner \& Rattermann, 1991), to maturational increases in processing capacity (Halford, 1992) and to increases in executive ability, including inhibitory control (Doumas, Hummel, \& Sandhofer, 2008; Richland et al., 2006; Thibaut, French, \& Vezneva, 2010), and it is possible that all three play a role.

This work has also revealed characteristic patterns of relational learning, including factors that support or hinder it. One signature component of relational learning is that the ability to perceive abstract relational matches can be enhanced by comparing different instances of a relation. For example, Gick and Holyoak (1983) found that comparing two stories that had the same causal structure enabled people to transfer that structure to a further situation. Preschool children have shown similar benefits from comparison (e.g., Christie \& Gentner, 2010; Honomichi \& Chen, 2006). These findings are consistent with other research suggesting that the act of comparison entails a structural alignment process that highlights the relational commonalities between the compared items (Markman \& Gentner, 1993). The influence of structural alignment is a defining characteristic of analogical reasoning in adults (Doumas \& Hummel, 2013; Gentner, Holyoak, \& Kokinov, 2001), and the evidence of its influence in children as young as 3 years of age suggests that there may be continuity in the signature components of
relational learning through human development.
A second signature component of relational learning is that attention to individual objects can interfere with relational processing. Preschool children perform far worse on relational matching tasks when competing object matches are present (Gentner \& Toupin, 1986; Richland et al., 2006), especially if the objects involved are rich and distinctive (Gentner \& Rattermann, 1991; Paik \& Mix, 2006). The finding that attention to objects can overshadow attention to relations extends to very young age groups (Casasola, 2005; Maguire, Hirsh-Pasek, Golinkoff, \& Brandone, 2008).

The following experiments aim to trace the development of relational learning processes in infants. We focused on the same-different relation because it is among the simplest and most basic relations in the human repertoire. Additionally, Ferry, Hespos, \& Gentner (2015) found that 7- and 9-month infants can learn same-different relations from four exemplars of same or different toy pairs (e.g., $\mathrm{AA}, \mathrm{BB}, \mathrm{CC}$, DD or $\mathrm{AB}, \mathrm{CD}, \mathrm{BC}, \mathrm{DA}$ ). The key finding was that infants discriminated between the relation they had experienced and the novel relation, even when both were instantiated with new objects. Further, infants failed to discriminate between the learned relations when the test pairs contained objects that have been rendered individually salient prior to habituation. This was consistent with the findings among older children, for whom object salience interferes with analogical comparison (Gentner \& Toupin, 1986; Richland et al., 2006). These findings suggest that by 7 months, infants show the basic characteristics of analogical learning. In the present research, we took this investigation to even younger infants.

## Experiment 1

To fully understand the ontogenetic development of relational processes, we need to test for relational abstraction at the earliest age possible. This will this provide evidence as to when in development relational processing becomes possible. Further, it will serve as a base for capturing developmental changes in the learning process across age groups.

The key dependent measure in this study is whether infants can differentiate the familiar relation (e.g., same if habituated to same) from the unfamiliar one (different) when they see test pairs composed of new objects. The specific predictions are that if infants are learning via structural alignment, then (a) relational encoding and abstraction should benefit from comparing a series of exemplars and (b) relational encoding should be hampered for pairs that contain a highly salient object (based on findings that object focus interferes with relational encoding (Gentner \& Toupin, 1986; Richland et al., 2006)).

Participants. The participants were 31 healthy, full-term 3-month-old infants ( 17 male and 14 female) with an average age of 3 months, 2 days. Sixteen infants were assigned to the same condition and 15 to the different condition. Seventeen additional infants were tested but eliminated from the final analysis for fussiness (defined as fussy or crying on 4 or more test trials by two independent coders), breaks longer than 8 minutes, or because they looked the maximum amount of time on 7 out of 8 test trials, making their data uninterpretable.
Materials and Procedure. Coding and analysis procedure was closely modeled on Ferry et al. (2015). In Experiment 1, infants received training on either same or different relations. During test trials, infants saw pairs of objects instantiating the same and different relations (See Figure 1). The key question was whether infants would differentiate the familiar relation from the novel relation at test. Each infant saw four types of test trials, composing a $2 \times 2$ within-subject design. The first type consisted of entirely new objects (New). These trials tested the main prediction: whether infants had abstracted the relation across the habituation pairs and applied this relation to new instances. The second test type consisted of objects that had been rendered individually salient in the waiting room prior to habituation, but not shown in habituation trials (Object Experience only). These trials investigated whether object salience would disrupt relational processing. The third type was made up of objects that had been rendered individually salient in the waiting room and had subsequently appeared as part of pairs during the habituation trials (Object Experience + Pair Habituation). These trials tested whether repeated alignment across pairs would overcome initial object salience. The fourth test trial type was made of objects that were not seen in the waiting room, but were viewed in pairs during habituation trials (Pair Habituation only). These trials provided a check on whether infants recognized identical pairs. If infants failed to discriminate between a pair that they had seen in habituation and a novel pair, this would suggest failure to learn the exemplars even at a concrete level.

A small camera captured video of the infant's face while they watched an experimenter raise, lower and tilt a pair of objects in tandem on the stage. Two research assistants in a separate room viewed the image, each pressing a button when the infant attended to events on stage and releasing the button when the infant looked away. A software program recorded the looking times. Each trial ended when the software signaled that the infant had looked away from the stage for more than two consecutive seconds. If coder agreement was less than $90 \%$, recordings of the trials were re-coded by two new coders.

## Methods



Figure 1. Schematic of events in Experiment 1. A) In the waiting room before the experiment, infants were shown a subset of individual objects used in the experiment. (B) During habituation trials, infants were either shown pairs of same objects or pairs of different objects. (C) during test trials, infants saw pair of objects presented sequentially. There were four types of test trials that systematically varied the infants' object experience with the objects to measure the influence on performance. To give a sense of the variation across the stimuli, three sets of same and different pairs are shown in Figure 2.


Figure 2 Examples of the same and different pairs.
Pair habituation trials. When the screen was raised at the start of every trial, a pair of objects rested on the cardboard tray on the stage. To engage infants' attention, in both habituation and test trials, the experimenter grasped one object in each hand and raised the objects, tilted them to the left and right, then paused on the tray. This 8-s cycle repeated continuously until the trial ended.

Test trials. Infants viewed eight test trials. In each test trial, infants viewed one pair of objects, presented in the same motion pattern as in the habituation trials, while their looking time was recorded. Each infant received test trials with both same and different pairs of objects, presented in alternation, with order counterbalanced across infants.

Figure 3. Test trial looking times for Experiment 1. Looking durations to novel and familiar pairs for each test type were collapsed across same and different conditions. The
diamonds represent the mean. The horizontal line inside the rectangle represents the median. The area above and below the median represents the 1 st and 3rd quartiles respectively.

## Results

The results depicted in Figure 3 show no evidence of generalization to new pairs: infants did not distinguish novel from familiar relations on the test pairs with new objects. An


ANOVA testing the between-subject factor of habituation condition (same or different), and the within-subject factor of relation (novel or familiar) failed to show a significant effect of relation across all test trials, $F(1,30)=.967, p=.333$.

Critically, there was no evidence that either group-same or different-had abstracted the relation, because they showed no difference in looking time between novel and familiar relations when the relations were composed of new objects, $\mathrm{t}(30)<1, \mathrm{p}=.628$. This pattern suggests that infants recognized pairs they had seen previously, but did not generalize the relation.

## Discussion

Given infants' failure to generalize the relation to the novel objects in Experiment 1, there are at least three possible interpretations. First, three months-old infants may not yet be able to engage in analogical learning. Second, they may not be able to form abstract relations like same and different. However, a third possibility is that these young infants do already possess the relational learning processes, but that the training set used in Experiment 1 (i.e., six unique pairs of exemplars) was not adequate. For example, the range of exemplars given in habituation may have been too limited. Perhaps these very young infants need more variation and more exemplars to abstract the relation. This would be consistent with the standard assumption in learning theories that high variability in training enhances transfer, and with evidence that generalization improves when the number and
range of examples increases (e.g. Gerken, 2006; Quinn \& Bhatt, 2005; Xu \& Tenenbaum, 2007). The six exemplars we showed in Experiment 1, though, is already a larger training set than the four exemplars 7 - and 9 -month-olds saw in Ferry, et al. (2015). This larger set did not appear to benefit the 3-month-olds.

A second route-the one we pursued-is to show infants fewer pairs during habituation. Although this choice may seem counterintuitive, there is evidence that in early learning, fewer exemplars of a relation can lead to better learning of the relation. For example, Casasola (2005a) found that infants were better able to learn and generalize the spatial category of support when they were given two alternating exemplars of the relation than when they were given six exemplars of the relation (see also Maguire, et al., 2008). This pattern can be understood in terms of the general finding that relational alignment can be impeded by attention to objects (Gentner \& Medina, 1998; Gentner \& Toupin, 1986; Paik \& Mix, 2006)-particularly when the objects are rich and distinctive (Casasola \& Park, 2013).

## Experiment 2

In Experiment 2, we presented infants with only two pairs during habituation---either two same pairs (alternating between $A A$ and $B B$ ) or two different pairs ( AB and CD ). The idea is that alternating between just two pairs could allow that infants to become familiar enough with the objects to be able to attend to the relation between them. As in our previous studies, prior to habituation we showed the infants some of the objects (singly, not in pairs) in order to render those objects individually salient. This serves as a test of whether object salience disrupts relational learning in 3-month-olds. Thus, this study tests whether 3-month-old infants can abstract the same-different relation and generalize it to new test pairs and whether their ability to do so will be impeded for pairs containing high-salient objects (see Figure 4).

## Methods

Participants. The participants were 32 healthy, full-term, 3-month-old infants ( 19 male and 13 female) average age 3 months and 16 days, ranging from 2 months 10 days to 4 months 15 days. Half of the infants were assigned to the same condition; the other half, to the different condition. Ten additional infants were tested but eliminated from the final analyses (using the same criteria as Experiment 1). Procedure. As in Experiment 1, there were three types of test trials, varied according to infants' experience with the objects. Because fewer objects were used in habituation, we reduced the number of test trial types from four to three, dropping the Pair Habituation only trials (see Figure 4). The remaining test trial types were as in Experiment 1.


Figure 4. Schematic of events in Experiment 1. A) In the waiting room before the experiment, infants were shown a subset of individual objects used. (B) During habituation trials, infants were shown either alternating pairs of either same or different objects. (C) During test trials, infants saw six pairs of objects, presented sequentially.

## Results

The results (Figure 5) fit the predictions of an analogical learning account. First, infants looked significantly longer at the novel relation than at the familiar relation during test. Critically, this novelty preference held for test pairs containing new objects, demonstrating that the infants had abstracted the relation and could apply it to objects they had not seen before, $t(29)=3.616, p<.001$. Second, as predicted, prior experience with individual objects interfered with noticing the relation: there was no significant difference in looking time between the novel and familiar relations for pairs containing objects seen in the waiting room. This was true whether these salient objects appeared only in test or in pair habituation as well as test.


Figure 5. Test trial looking times for Experiment 2. Looking durations to novel and familiar pairs for each test
type collapsed across same and different conditions. The diamonds represent the mean. The horizontal line inside the rectangle represents the median. The area above and below the median represents the 1 st and 3rd quartiles respectively. The * indicates $p<.01$.

## Discussion

The findings in Experiment 2 are evidence for early relational learning. Critically, infants were able to distinguish the familiar relation from the novel relation even on pairs composed of new objects-the gold standard for testing whether infants abstracted the same and different relations. Consistent with other findings on analogical learning, infants performed significantly worse on test trials containing objects that had been seen prior to habituation. These findings show that the key signatures of analogical learning are already present by 3 months of age.

At the same time, the knowledge of the specific relations same and different in 3-month-olds appears to be learned: Experiments $1 \& 2$ show strong limits to the situations in which they could generalize these relations. Infants performed best with fewer exemplars. Even then, the infants did not discriminate between novel and familiar relations when they saw pairs that contained salient objects. Unlike with 7-month-olds in Ferry et al. (2015), 3-month-olds failed to overcome their object focus even after seeing the waiting room objects in pairs during habituation.

## General Discussion

There are two key findings. First, the results show that analogical learning processes are present in 3-month old infants. In Experiment 2, the infants showed two key signatures of analogical learning: (a) the ability to abstract a common relation across a sequence of pairs and (b) the detrimental effects of individual object salience. These findings suggest that the ability to abstract relations is an innate mechanism in human infants. If so, then analogical processing would join association and other domain-general processes as part of the core cognitive apparatus of humans.

The second key finding is that these young infants showed more learning when given just two pairs during habituation than when given six distinct pairs.
This pattern runs counter to the general finding that increasing the variability within a set of training stimuli increases learners' level of abstraction and therefore the range of transfer (Gerken, 2006; Gómez, 2002; Quinn \& Bhatt, 2005).

However, there is precedent for this kind of "less is more" finding (Casasola, 2005; Casasola \& Park, 2013; Maguire et al., 2008). What these studies have in common is that the objects participating in the relations are of high salience. Under these conditions, a participant given a series of different exemplars may attend only to the novel objects in each pair, and fail to attend to the relations. In this case,
reducing the range of instances so that a small set of exemplars is seen repeatedly may lead to better relational learning. As Casasola and Park (2013) note, although increasing the range of exemplars can help learners to isolate the relevant structures, "... the need for fewer exemplars arises when the relevant features, such as a spatial relation, risk becoming obscured by [...] the objects depicting that relation."

The finding then raises the question of when this pattern holds. As discussed earlier, many developmental studies have found better learning with more exemplars than with fewer (e.g., Bulf \& Johnson, 2011; Casasola \& Park 2013; Gerken 2006). Further, in our previous studies we found that 7- and 9 -month-olds successfully abstracted same and different relations when given four repeated exemplars (Ferry et al., 2015). Clearly, a goal for future research will be to understand the range of exemplar variability that best supports early relational learning across development.
Implications for learning theories. As noted above, a surprising finding is that in order for 3-month-old infants to learn the relations, they needed comparison across two repeating pairs rather than comparison across a greater variety of pairs. How do we square this finding with the many findings that greater variability during training leads to greater abstraction and transfer? We think that the key is that the current studies focus on relational learning. When the desired abstraction is at the level of overall exemplar similarity (e.g., learning a basic-level category such as dog, or learning a distribution of line lengths), then increasing the range of exemplars in learning should increase the level of generalization. However, if the desired generalization is a relational pattern, then it is crucial that the learner be able to compare and align the exemplars (Christie \& Gentner, 2010). In this case, whether the learner can align the exemplars may matter more than the amount of information potentially available. This leads us to suggest an amended learning principle: in relational learning, breadth of alignable training predicts breadth of transfer.
Summary. Together, the evidence from our experiments points neither to core knowledge of same or different nor to a process that arises entirely from experience, but to structural alignment as an early learning mechanism that becomes elaborated over development and with increases in language and domain knowledge.

## Acknowledgements

This research was supported by NSF Grant BCS-1423917 to Susan J. Hespos, NSF SLC Grant SBE-0541957 awarded to the Spatial Intelligence and Learning Center (SILC), and by ONR Grant N00014-92-J-1098 to Dedre Gentner. We are indebted to parents whose infants participated in the research. We thank members of the Infant Cognition Lab for help with data collection, and Ruxue Shao, Christian de Hoyos and Christine Schlaug for their insightful comments.

## References

Bulf, H., Johnson, S. P., \& Valenza, E. (2011). Visual statistical learning in the newborn infant. Cognition, 121(1), 127-132.
Casasola, M. (2005). When Less Is More: How Infants Learn to Form an Abstract Categorical Representation of Support. Child Development, 76(1), 279-290.
Casasola, M., \& Park, Y. (2013). Developmental changes in infant spatial categorization: When more is better and less is enough. Child Development, 84, 10041019.

Chen, Z. (1996). Children's Analogical Problem Solving: The Effects of Superficial, Structural, and Procedural Similarity. Journal of Experimental Child Psychology, 62(3), 410-431.
Christie, S., \& Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. Journal of Cognition and Development, 11(3), 356-373.
Doumas, L. A. A. \& Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. PLOS One, 8 (6), e63889.
Doumas, L. A. A. \& Hummel, J. E., \& Sandhofer C.M. (2008). A theory of the discovery and predication of relational concepts. Psychological Review, 115, 1-43.
Ferry, A., Hespos, S. J., \& Gentner, D. (2015). Prelinguistic Relational Concepts: Investigating Analogical Processing in Infants. Child Development, 96(5), 1386-1405.
Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child Development, 59, 47-59.
Gentner, D. (2003). Why we're so smart. Language in mind: Advances in the study of language and thought. In D. Gentner \& S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought (pp. 195-235). Cambridge, MA: MIT Press.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34(5), 752-775.
Gentner, D., \& Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. American Psychologist, 52, 32-34.
Gentner, D., \& Medina, J. (1998). Similarity and the development of rules. Cognition, 65, 263-297.
Gentner, D., \& Rattermann, M. (1991). Language and the career of similarity. In S. A. Gelman \& J. P. Byrnes (Eds.), Perspectives on language and thought: Interrelations in development (pp. 225-277). London: Cambridge University Press.

Gentner, D., \& Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. Cognitive Science, 10, 277-300.
Gerken LA. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. Cognition, 98, B67-B74.
Honomichi, R. D, \& Chen, Z. (2006) Learning to align relations: The effects of feedback and selfexplanation. Journal of Cognition and Development, 7(4), 527-550.
Halford, G. (1992). Analogical reasoning and conceptual complexity in cognitive development. Human Development, 35, 193-217.
Maguire, M. J., Hirsh-Pasek, K., Golinkoff, R. M., \& Brandone, A. C. (2008). Focusing on the relation: Fewer exemplars facilitate children's initial verb learning and extension. Developmental Science, 11(4), 628-634.
Paik, J.H. \& Mix, K.S. (2006). Preschooler's Use of Surface Similarity in Object Comparisons: Taking Context into Account. JECP, 95(3), 194-214.
Penn, D. C., Holyoak, K. J., \& Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behavioral and Brain Sciences, 31(2), 109-130.
Quinn, P. C., \& Bhatt, R. S. (2005). Learning perceptual organization in infancy. Psychological Science, 16, 515-519.
Richland, L. E., Morrison, R. G., \& Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. JECP, 94, 249-271.
Thibaut, J.-P., French, R. M., \& Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive functions. JECP , 106(1), 1-19.
Xu, F. and Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. Developmental Science 10(3), 288-297.

# A Preliminary P-Curve Meta-Analysis of Learned Categorical Perception Research 

Janet Andrews (andrewsj@vassar.edu)<br>Department of Cognitive Science, 124 Raymond Ave. Poughkeepsie, NY 12604 USA<br>Joshua de Leeuw (jdeleeuw@vassar.edu)<br>Department of Cognitive Science, 124 Raymond Ave. Poughkeepsie, NY 12604 USA<br>Calais Larson (calarson@vassar.edu)<br>Department of Cognitive Science, 124 Raymond Ave. Poughkeepsie, NY 12604 USA<br>Xiaoqing Xu (xixu@vassar.edu)<br>Department of Cognitive Science, 124 Raymond Ave. Poughkeepsie, NY 12604 USA


#### Abstract

A preliminary meta-analysis using the p-curve method (Simonsohn, Nelson, \& Simmons, 2014) was performed on a subset of the learned categorical perception literature to explore the robustness of the phenomenon. Only studies using novel visual categories and behavioral measures were included. The results strongly suggest that the phenomenon is robust but that the studies are somewhat underpowered. We argue that this is problematic because it renders both statistically significant and nonsignificant results very difficult to interpret, which impedes progress in understanding the learned CP phenomenon, for example, why expansion vs. compression is observed, or boundary vs. dimensional effects. Fortunately, there is a clear solution: conduct studies with greater statistical power.


Keywords: categorical perception; categorization; learning; p-curve; statistical power; expansion; compression; dimensional modulation

## Introduction

Learned categorical perception (CP) is a phenomenon whereby learning to place objects into categories alters some aspect of the way those objects are judged (for a review, see Goldstone \& Hendrickson, 2009). The classic patterns of change in judgments are that items placed in different categories become more distinguishable, sometimes called expansion, and/or that items placed in the same category become less distinguishable, sometimes called compression. These are category boundary effects, but other versions of learned CP are increased sensitivity to dimensions relevant to the category distinction and/or decreased sensitivity to dimensions irrelevant to the category distinction.

There has been a great deal of research on learned CP, but this paper will focus on visual categories learned in a laboratory setting. This is motivated mainly by our own interest in learned visual CP and the recognition that there may be important differences between CP in different
modalities that would make it inappropriate to group those studies together for this meta-analysis. Thus the large body of research on auditory CP, in particular for speech sound distinctions that are acquired in the lengthy process of learning a natural language, will not be considered here. It is notable that laboratory-induced learned CP effects are obtained with very little training compared to the kind of exposure that is usually given in real category learning, e.g., learning color categories. This makes the phenomenon appear to be pervasive and basic, but there are several important questions that need to be addressed.

First, in light of the recent attention given to failures to replicate (Open Science Collaboration, 2015), p-hacking (Head, Holman, Lanfear, Kahn, \& Jennions, 2015), the file drawer problem (Rosenthal, 1979), lack of sufficient statistical power (Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, \& Munafò, 2013), and so forth, we believe that it is essential to assess whether the published learned CP literature demonstrates convincing evidence for the effect. Like many, perhaps most, areas of research in cognitive science, the phenomenon of learned CP allows researchers many degrees of freedom that may, unintentionally no doubt, inflate the appearance of real effects. For example, within a single study, there are often a variety of different ways in which the data can be analyzed to look for evidence of learned CP effects. Researchers can investigate accuracy and/or response time data - both have been used as evidence of CP in the literature - in a multitude of combinations due to the many different possible behavioral patterns that count as learned CP. Furthermore, different potential criteria for what counts as successful learning (a precondition for testing for learned CP effects) can be used, leading to additional choices that can unintentionally bias the analysis. In short, this is exactly the kind of situation where preregistration is important to avoid mistaking the noise in the data for signal. To our knowledge, very little if any learned CP research has been preregistered as of yet, for replication purposes or otherwise.

One reason this is important is the current controversy over whether there are really any genuine top-down effects of cognition on perception, of which learned CP could potentially be one type. Firestone and Scholl (2016) argue that none of the vast amount of research claiming to have demonstrated such effects has actually successfully done so. It seems to us that before we can effectively debate whether learned CP provides evidence for such effects that is immune from Firestone and Scholl's criticisms, we must first establish that there is credible evidence that the effects themselves actually exist.

The optimist might point to the dozens of studies on CP that report significant results as compelling evidence that these effects are real. The problem with this approach is that it ignores the existence of publication bias in favor of statistically significant results. If learned CP turned out to not be replicable, it would not be the first example of a widely reported phenomenon that did not reliably replicate (e.g., Lurquin et al., 2016; Papesh, 2015; Shanks et al., 2013; Simmons \& Simonsohn, in press).

The purpose of this paper was to do a preliminary evaluation of published learned CP effects by using a pcurve analysis (Simonsohn, Nelson, \& Simmons, 2014). A p-curve is a meta-analytic technique that looks at the distribution of statistically significant $p$-values in a set of related studies. An advantage of the p-curve approach is that it nicely handles the file-drawer problem by looking only at statistically significant $p$-values. If the results come from a collection of well-powered studies investigating a real
effect, most of the p-values should be very small (well below .05). However, if many of the results are due to falsepositives, either because of a lack of statistical power or because the phenomenon being studied is not real, then the distribution of significant $p$-values will be flat, in the case of a null effect, or close to flat, in the case of low statistical power. (In extreme cases, the presence of substantial phacking can generate a p-curve with left skew.) Furthermore, the p-curve can reliably estimate the average statistical power of a set of studies because the distribution of observed p -values is directly related to statistical power when the null hypothesis is false. This estimation of power is an average estimate assuming that all studies are investigating the same basic effect.

In the analyses presented below, we calculate p-curves for a set of studies from the visual learned CP literature, but we caution that this analysis is preliminary. We expect that enlarging the scope of the search for relevant sources would yield many additional studies of learned CP that could be included in the analysis, though our sample size is large enough to likely be informative.

## Method and Results

The articles used in the analysis were selected by conducting a Scopus search of all articles citing Goldstone (1994) that had "CP" or "categorical perception" in the title, abstract, or key words and were deemed relevant (i.e., reported original empirical results in detail; used visual

Table 1. Articles used in the p-curve meta-analysis of learned categorical perception research.

| Authors | Year of publication | \# experiments included | \# analyses <br> included | Learned CP measure | Type of measure ${ }^{\text {a }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Corneille \& Judd | 1999 | 3 | 2 | Typicality | S |
| Folstein, Palmeri, \& Gauthier | 2014 | 1 | 1 | Same-different | O |
| Goldstone | 1994 | 4 | 8 | Same-different | O |
| Goldstone, Lippa, \& Shiffrin | 2001 | 1 | 1 | Similarity | S |
| Goldstone, Steyvers, \& Larimer | 1996 | 1 | 1 | Same-different | O |
| Grandison, Sowden, Drivonikou, Notman, Alexander, \& Davies | 2016 | 1 | 1 | Target location RT | O |
| Gureckis \& Goldstone | 2008 | 1 | 1 | XAB | O |
| Holmes \& Wolff | 2012 | 1 | 1 | Discrimination RT | O |
| Levin \& Beale | 2000 | 3 | 2 | XAB variant | O |
| Livingston \& Andrews | 2005 | 2 | 2 | Similarity, same-different | S, O |
| Livingston, Andrews, \& Harnad | 1998 | 3 | 2 | Similarity | S |
| Notman, Sowden, \& Özgen | 2005 | 2 | 2 | Same-different | O |
| Op de Beeck, Wagemans, \& Vogels | 2003 | 2 | 2 | Same-different | O |
| Özgen \& Davies | 2002 | 2 | 2 | Same-different | O |
| Stevenage | 1998 | 2 | 4 | Similarity | S |
| Zhou, Mo, Kay, Kwok, Ip, \& Tan | 2010 | 1 | 1 | Target location RT | O |

${ }^{\mathrm{a}} \mathrm{O}=$ objective, $\mathrm{S}=$ subjective


Note: The observed p-curve includes 33 statistically significant ( $p<.05$ ) results, of which 26 are p<. 025 .

Figure 1: P-curve for all relevant results in the articles listed in Table 1. The solid blue line shows the observed distribution of significant p-values. The dotted red line shows what the expected distribution of p -values would be if the null hypothesis were true. The right skew tests, for both the full set of p-values (all p-values $<.05$ ) and half set of p-values (all p-values $<$ .025 ), indicate that the p-values are more right skewed than would be expected if the null were true. The dashed green line represents the expected distribution of p-values if the set of studies had $33 \%$ power. The flat tests test if the observed distribution is flatter than the distribution that would be observed if the studies had $33 \%$ power.
stimuli, unfamiliar categories, and behavioral measures; and focused on learned CP). This yielded 14 articles; in addition, we included two conference papers meeting the same criteria for a total of 16 sources (see Table 1). Within those 16 sources were a total of 30 experiments reporting 42 distinct relevant statistical results. (For example, statistical results pertaining to the learning of the categories per se were not relevant.) Of these 42 statistical results, 33 were statistically significant in the predicted direction and these results were input to the p-curve app version 4.05 (http://www.p-curve.com/app4/) to produce the p-curve shown in Figure 1. Note that the p-curve analysis only considers the distribution of p -values below the 0.05 threshold.

In addition, separate p-curves were generated for the subset of results obtained using objective measures of learned CP, such as accuracy of same-different judgments, and the subset of results obtained using subjective measures such as similarity judgments. These are shown in Figures 2 and 3.

The p-curves shown in Figures 1-3 display the distribution of $p$-values that fall into five bins. For a real
effect with high power samples, most of the p -values should be in the leftmost bin ( $\mathrm{p}<0.01$ ). Shown in the figures are what the distribution would look like with a set of studies powered at $33 \%$ (green dashed line) and a set of studies testing a null effect (red dotted line). Note that because the distribution of p -values is determined by statistical power when investigating a real effect, the average power of the studies can be estimated from the observed distribution of p values.

The results of the overall p-curve meta-analysis show a curve that is right-skewed, which strongly suggests that the research has evidential value and is not the result of worrisome p-hacking (which produces a left-skewed curve). ${ }^{1}$ This is welcome news. However, the estimated power is only $62 \%$ ( $90 \%$ confidence interval is $41-78 \%$ ) which is not very high. If this estimate were correct, it would mean that only 6 in 10 studies of learned CP will detect an effect. We explain why this is a significant

[^270]problem for advancing our theory of learned CP in the discussion.

We conducted a separate analysis for objective and subjective measures of CP for two reasons. The first is that some researchers have suggested based on neural evidence that learned CP is not a genuinely perceptual effect but occurs at a higher, post-perceptual level (e.g., Clifford, Franklin, Holmes, Drivonikou, Özgen, \& Davies, 2012; but see Zhong, Li, Li, Xu, \& Mo, 2015 for counterevidence). We reasoned that if this were the case, studies employing more subjective measures of learned CP , such as ratings of similarity or typicality, should show stronger effects. The second is that there is a worry that learned CP effects could be the result of demand effects (Goldstone, Lippa, \& Shiffrin, 2001). That is, participants in these experiments may indicate that two items are subjectively more similar (dissimilar) to each other precisely because the experiment trains them that the two objects belong to the same category (different categories), and not because of any perceived change in similarity of the visual objects. If this is a contributing factor, then we would expect studies with subjective measures to have higher power, because this demand effect will only contribute for subjective measures of CP.


Figure 2: P-curve for learned CP results based on objective measures.

These ideas received some support from the p-curve patterns based on studies using subjective vs. objective learned CP measures, with a generally stronger pattern and higher power estimate for the subjective measure studies (79\%) than objective measure studies (52\%). However, given the preliminary nature of this analysis and the small set of results included, particularly for those using subjective measures, it is premature to draw any conclusions about this yet (note that the confidence intervals for the power estimates overlap substantially). Furthermore, we
can't distinguish between the two possible explanations of this result without directly investigating the matter.


Figure 3: P-curve for learned CP results based on subjective measures.

## Discussion

The literature for learned visual categorical perception contains evidentiary value, according to this meta-analysis. We can be reasonably confident that the studies reported are in general not reporting on a null effect. However, the relatively low statistical power shown by this analysis for the overall set of findings has important implications for how our theoretical understanding of learned CP is informed by these studies, and future studies with similar statistical power. We argue here that the statistical power of learned CP research must be improved in order to make robust advances in theory.

Several debates in learned CP research (e.g., to what degree is CP a perceptual or decision-making process; what kinds of judgments are changed by learning categories; is CP the result of demand effects) currently hinge on the observation of CP in some experimental contexts but not others. However, the overall lack of statistical power makes the pattern of significant and non-significant results difficult to interpret. Low power may well explain the occurrence of non-significant results. Low power also increases the likelihood that significant results are actually false positives. It follows from these two facts that when studies of learned CP are underpowered, the noise in the data makes it very difficult to distinguish among specific theoretical variants of what learned CP is. For example, it is difficult to distinguish among the different types of learned CP , of which there are at least four, as noted in the introduction (the boundary effects of compression and/or expansion and dimensional sensitization and desensitization based on category relevance). Since null results are impossible to interpret when statistical power is low (and using traditional
statistical methods), and patterns in statistically significant results may be just noise, it is correspondingly impossible to use the data to figure out under what conditions each of the types of learned CP do and do not occur. Yet this is essential to do in order to determine the nature of learned CP mechanisms and their purpose.

To understand the problem that this causes for our theoretical understanding of learned CP , it is important to keep in mind that the articles included in our analysis are not a set of direct replications, but rather a body of scientific evidence. Experiments in this set of results aim to build on the contributions of prior work to refine our theory of learned CP. Thus, we often rely on the pattern of findings within individual studies, or between small sets of studies, to constrain theorizing. But, as noted, with low statistical power comes the increased probability of false negatives and the increased probability that significant results are false positives. This leaves the theorist in a tough position. Are we improving our theoretical understanding with a new set of data, or merely reading the tea leaves of statistical noise?

Recent work in our lab provides an illustration of this problem (de Leeuw, Andrews, Livingston, \& Chin, 2016). We were primarily interested in why some learned CP studies had shown compression while others showed expansion. There seemed to be a relationship between both the type of learned CP measure (similarity vs. samedifferent judgment accuracy) and stimulus discriminability, on the one hand, and the pattern of learned CP on the other. Initial studies in our lab seemed to confirm variations of this kind but the patterns were somewhat bewildering. Only when we conducted a large scale study ( $\mathrm{N}>550$ ) simultaneously incorporating multiple measures and levels of stimulus discriminability and used Bayesian data analysis did a clear picture emerge: learned CP effects occurred (in fact, three of the four possible patterns occurred), but this pattern of effects did not differ systematically according to either of those variables. ${ }^{2}$

It is important to note a related set of problems with learned CP research that also presented a challenge for conducting the p-curve analysis. (1) Learned CP studies often don't test for more than one or two of the four possible types of effect, and the statistical analysis used may test for different effects separately or lump them together. Furthermore, there are methodological ambiguities in many of the studies that make separating out which effects occurred impossible. ${ }^{3}$ We therefore could not classify the pvalues according to which aspect of learned CP they

[^271]corresponded to, even though we would have liked to be able to do this. (2) Predictions are often vague in regard to the nature of a two-way interaction in an ANOVA, for example. But different results should be used for the pcurve for attenuation and reversal interaction predictions (just the overall interaction for attenuation and just the simple effects for reversal). Since we were limited to the information available in the articles, most of which did not predict a specific pattern of interaction, power may be overestimated by the p-curve. (3) A final caveat regarding our analysis is that our p-curve results could potentially be somewhat misleading if in fact certain learned CP effects (e.g., dimensional effects) are much stronger than others (e.g., boundary effects), which would mean that they are not really the same kind of effect as assumed by a combined pcurve analysis. If this were the case, it would suggest that power could be higher than our estimate for some aspects of learned CP, but lower for others. This would only further exacerbate the problem of drawing theoretical conclusions about the nature of learned CP , as the studies of certain kinds of effects that are deeply relevant to the theory would have even lower power.

The preliminary meta-analysis we report here strongly suggests that learned CP effects are real but also that our current knowledge of them is highly ambiguous and destined to remain so if we do not change the way we do research. In our view, only by conducting future studies with sufficient statistical power will we make significant progress understanding the phenomenon of learned CP.

## References

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., \& Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365-376.
Clifford, A., Franklin, A., Holmes, A., Drivonikou, V. G., Özgen, E., \& Davies, I. R. L. (2012). Neural correlates of acquired color category effects. Brain and Cognition, 80(1), 126-143.
Corneille, O., \& Judd, C. M. (1999). Accentuation and sensitization effects in the categorization of multifaceted stimuli. Journal of Personality and Social Psychology, 77(5), 927-941.
de Leeuw, J. R., Andrews, J. K., Livingston, K. R., \& Chin, B. M. (2016). The effects of categorization on perceptual judgment are robust across different assessment tasks. Collabra, 2(1), 1-9.
Firestone, C., \& Scholl,, B. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. Behavioral and Brain Sciences, 39.
Folstein, J. R., Palmeri, T. J., \& Gauthier, I. (2014). Perceptual advantage for category-relevant perceptual dimensions: The case of shape and motion. Frontiers in Psychology, 5.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General, 123, 178-200.
Goldstone, R. L., \& Hendrickson, A. T. (2009). Categorical perception. Interdisciplinary Reviews: Cognitive Science, 1, 6978.

Goldstone, R. L., Lippa, Y., \& Shiffrin, R. M. (2001). Altering object representations through category learning. Cognition, 78(1), 27-43.
Goldstone, R. L., Steyvers, M., \& Larimer, K. (1996). Categorical perception of novel dimensions. Proceedings of the eighteenth annual conference of the cognitive science society (pp. 243-248). Hillsdale, NJ: Lawrence Erlbaum Associates.
Grandison, A., Sowden, P. T., Drivonikou, V. G., Notman, L. A., Alexander, I., \& Davies, I. R. L. (2016). Chromatic perceptual learning but no category effects without linguistic input. Frontiers in Psychology, 7.
Gureckis, T. M., \& Goldstone, R. L. (2008). The effect of internal structure of categories on perception. Proceedings of the thirtieth annual conference of the cognitive science society (pp. 1876-1881). Washington, D.C.: Cognitive Science Society.

Head, M. L., Holman L., Lanfear, R., Kahn A. T., \& Jennions M. D. (2015). The extent and consequences of phacking in science. PLoS Biology, 13(3), e1002106.
Holmes, K. J., \& Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? Journal of Experimental Psychology: General, 141(3), 439-443.
Levin, D. T., \& Beale, J. M. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. Perception and Psychophysics, 62(2), 386401.

Livingston, K. R., \& Andrews, J. K. (2005). Evidence for an age-independent process in category learning. Developmental Science, 8(4), 319-325.
Livingston, K. R., Andrews, J. K., \& Harnad, S. (1998). Categorical perception effects induced by category learning. Journal of Experimental Psychology: Learning Memory and Cognition, 24(3), 732-753.
Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., von Bastian, C. C., Carruth, N. P., \& Miyake, A. (2016). No evidence of the ego-depletion effect across task characteristics and individual differences: A preregistered study. PLoS ONE, 11(2), e0147770.
Notman, L. A., Sowden, P. T., \& Özgen, E. (2005). The nature of learned categorical perception effects: A psychophysical approach. Cognition, 95(2), B1-B14.
Op de Beeck, H., Wagemans, J., \& Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. Journal of Experimental Psychology: General, 132(4), 491-511.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251).

Özgen, E., \& Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. Journal of Experimental Psychology: General, 131(4), 477-493.
Papesh, M. (2015). Just out of reach: On the reliability of the action-sentence compatibility effect. Journal of Experimental Psychology: General, 144(6), e116-e141.
Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86(3), 638-641.
Shanks, D.R., Newell, B.R., Lee, E.H., Balakrishnan, D., Ekelund, L., Cenac, Z., et al. (2013). Priming intelligent behavior: an elusive phenomenon. PLoS ONE, 8(4), e56515.
Simmons, J., \& Simonsohn, U. (in press). Power posing: P-curving the evidence Psychological Science.
Simonsohn, U., Nelson, L., \& Simmons, J. (2014). P-curve: A key to the file drawer. Journal of Experimental Psychology: General, 143(2), 534-547.
Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. British Journal of Psychology, 89(1), 39-57.
Zhong, W., Li, Y., Li, P., Xu, G., \& Mo, L. (2015). Shortterm trained lexical categories produce preattentive categorical perception of color: Evidence from ERPs. Psychophysiology, 52(1), 98-106
Zhou, K., Mo, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., \& Tan, L. H. (2010). Newly trained lexical categories produce lateralized categorical perception of color. Proceedings of the National Academy of Sciences of the United States of America, 107(22), 9974-9978.

# Reading Skill Test to Diagnose Basic Language Skills in Comparison to Machines 

Noriko H. Arai (arai@nii.ac.jp)<br>National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan<br>Naoya Todo (ntodo@niad.ac.jp)<br>NIAD-QE, 1-29-1 Gakuen-nishimachi, Kodaira-shi, Tokyo 187-8587, Japan<br>Teiko Arai (arai-teiko@g.ecc.u-tokyo.ac.jp), Kyosuke Bunji (bunji@p.u-tokyo.ac.jp)<br>University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan<br>Shingo Sugawara (shingo.sugawara.77@hosei.ac.jp)<br>Faculty of lifelong Learning and Career studies, Hosei University, 2-17-1 Fujimi, Chiyoda-ku, Tokyo 102-8160, Japan<br>Miwa Inuzuka (m_inuzuka@mail.tais.ac.jp)<br>Faculty of Humanity, Taisho University, 3-12-4 Nishi-sugamo, Toshima-ku, Tokyo 170-8470, Japan<br>Takuya Matsuzaki (matuzaki@nuee.nagoya-u.ac.jp)<br>Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan<br>Koken Ozaki (koken@gssm.otsuka.tsukuba.ac.jp)<br>Graduate School of Business Sciences, University of Tsukuba, 3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan


#### Abstract

A reading skill test to diagnose basic language skills is introduced. The test is designed to measure six component skills relevant to reading in comparison with those of state-of-the-art natural language processing technologies. The results of the first large-scale experiments using the test are reported. Surprisingly, almost half of Japanese junior high school students do no better than machines in dependency analysis. More than half of $7^{\text {th }}$ grade students do no better than making random choices on questions involving inferences and definition understanding.


Keywords: Reading Skills, Language Comprehension, Test Theory

## 1. Introduction

Artificial intelligence (A.I.) armed with machine learning technologies often surprises us by demonstrating its power. Arai et al. developed A.I. systems that were capable of passing the entrance examinations of more than half the universities in Japan (Arai \& Matsuzaki, 2014). On the other hand, teachers are facing the problem that many students come into their classrooms without the requisite knowledge, skills, or disposition to read and comprehend the materials placed before them (RAND, 2002).
This situation raises a natural question. Will there be any economic returns to education when A.I. is smart enough to "learn" better than most of us? Do we have to set different goals for education in the age of A.I.?
Before jumping to any conclusions, we must carefully study the performance of human beings in comparison with those of machines, especially of the skills and expertise that are believed to be acquired only through education. Reading comprehension is, of course, one such example.
In this paper, we introduce a new reading skill test (RST) for assessing an examinee's basic language skills involved in the comprehension of texts consisting of sentences taken from junior high and high school textbooks and dictionaries. It is a major version-up from the prototype developed in (Fujita et al., 2016). A unique feature of the RST is that it is designed to analyze language skills of both human beings
and machines. Consequently, the test results will tell us not only an examinee's language skills relative to others, but also to machines. It will also reveal what kinds of sentences (i.e. lexical, structural, thematic) are harder than others to comprehend (process) for human beings (for machines).
The RST contains six different types of question. The first two types are designed to measure an examinee's ability to analyze intra- and inter-sentential relations among words: dependency analysis and anaphora resolution. Statistical algorithms often achieve precisions around $80 \%-90 \%$ in parsing sentences and $60 \%-70 \%$ in anaphora resolution (Nivre et al., 2007; Pradhan et al., 2012), which indicates that not only examinees but also A.I. may be able to perform syntactic analysis of a sentence without understanding its meaning. The second two are designed to measure an examinee's inferential skills based on appropriate amounts of vocabulary and common sense. They are closely related to tasks called textural entailment recognition or synonymy recognition in the field of natural language processing, and both of them are known to be very hard (Dagan et al., 2013). The last two are designed to measure how examinees can map texts into meanings. They require high-level symbol grounding and abstract thinking, and neither a practical algorithm to solve them nor a theory to formalize them has been proposed yet.
If an examinee does equally well on the six different types of question, we can assume that he/she reads differently from machines. On the other hand, we had better doubt that an examinee reads like a machine if he/she does well on the first two types of question: he/she appears to understand the meaning of the texts, but actually may not. In other words, human-machine comparison and error analysis of machines may allow us to diagnose why many readers read poorly.
The results of the first large-scale investigation involving 1758 students from six public junior high schools are reported. Surprisingly, in a country like Japan where education is compulsory up to the end of junior high school, and which is among the top countries in PISA tests (OECD, 2016), more than half of the $7^{\text {th }}$ grade students did no better
than random choice on the third two types of question. These results lend support to the concerns expressed in (RAND, 2002). The performance of a popular Japanese dependency structure analyzer on dependency analysis questions is also reported for comparison.

## 2. Design of RST

2.1 Six Component Skills and their Measurement We define six component skills relevant to reading. Each skill is measured separately in the RST. We do not claim that basic language skills consist exclusively of these six. We plan to add new types as necessary.

1. Dependency Analysis (DEP): The skill of recognizing the dependency relations between words and phrases in a given sentence.

2 Anaphora Resolution (ANA): The skill of anaphora resolution. ANA is comprised of two elements: Demonstrative Anaphora Recognition (DANA) and Zero Anaphora Restoration (ZANA).

DANA: The skill of recognizing the anaphoric relation between a demonstrative pronoun in a sentence and its antecedent.

ZANA: The skill of restoring and recognizing a noun phrase implicitly omitted in a context.
3. Paraphrasing (PARA): The skill of recognizing that a sentence is the same in meaning as another one. PARA is comprised of three elements which are Lexical Paraphrasing (LeP), Structural Paraphrasing (SP), and Logical Paraphrasing (LoP). The participant reads two sentences and judges whether they are synonymous. The examinees are asked to choose "Yes" or "No".

LeP: The skill of recognizing the synonymy between words or short phrases.

SP: The skill of recognizing the synonymy between two sentences written in different voices (active/passive).

LoP: The skill of recognizing logical equivalency of two sentences.
4. Logical inference (INF): The skill of reading a sentence and determining what can be inferred from a proposition in the sentence, what conflicts with it, and what does not relate to it. Here, two sentences are presented to the examinees. The instruction asks the examinees whether the proposition in the second sentence (task sentence) can be inferred from the proposition in the first sentence (presented sentence). The examinees are asked to choose "Yes" if the sentence can be inferred, "No" if the first and the second propositions cannot hold true at the same time, and "Not known" if the propositions are not related to each other.
5. Representation (REP): The skill to represent an image (figure or table) by comprehending a sentence of the textbook. The participant reads a sentence and chooses the images correctly representing the sentence out of four (multiple responses).
6. Instantiation (INST): The skill to understand how to use a term correctly according to a given definition of the term. The participant reads a definition sentence and
chooses correct usages from four sentences (multiple responses).
2.2 Test settings Each RST question requires a considerable amount of concentration. We designed the RST so that examinees would not get confused or become exhausted. As a result, each examinee randomly takes three of six types of questions in the current setting. After answering two sample questions of a type, examinees are asked to answer questions randomly chosen from an item pool as precisely and quickly as possible in four minutes.

We intend to change the design of the test so that he/she takes all six types when we are ready to calculate $b$, the difficulties of the questions, and $\theta$, the ability of the examinee in Item Response Theory (IRT; Lord \& Novick, 1968; Hambleton \& Swaminathan, 1985) with fewer questions.
2.3 Interface RST is conducted as a Computer Based Test (CBT) or Paper Based Test (PBT). Figure 1 shows a screenshot of an REP question. For the details of the design, the reader should refer to (Fujita et al., 2016).


Figure 1: Question REP 39 shown in CBT
2.4 Materials We created all of the questions, except for the INST questions, on the basis of textbooks that have been approved by the Ministry of Education, Culture, Sports, Science and Technology and are being used in Japanese junior high and high schools. The INST questions were created using terms and definitions appearing either in the textbooks or in Japanese dictionaries.

## 3. Psychometric Properties of RST

An examinee's score is usually assessed by the sum score of all items to which he/she responded. However, because in the setting of RST, each examinee responds to different items, the sum score is not appropriate for an examinee's assessment. That is, the sum score is "item dependent", which means that the assessment result depends on the difficulties of the items that the examinee responded to as well as the examinee's characteristics.

Therefore, in this project, IRT is used for each examinee's assessment. One of the distinctive features of IRT is that it
is not item dependent. The reason is that an item's difficulty and an examinee's characteristics are treated as different parameters. An item $j$ 's difficulty parameter is denoted as $b_{j}$. The higher $b_{j}$ is, the more difficult the item is. An examinees $i$ 's characteristic is denoted as $\theta_{i}$. The higher $\theta_{i}$ is, the better the examinee's characteristic is, which is reading skill in this study. For the details of IRT, the reader is referred to the above references.

In the near future, we will start computerized adaptive testing (CAT, van der Linden and Glas, 2010). In CAT, each examinee answers items shown on a PC display or tablet. If the examinee correctly answers an item, the next item is more difficult, whereas if he/she incorrectly answers an item, the next item is easier. Note that CAT requires an item pool, which is a set of items whose item parameters have already been estimated. In CAT, an appropriate item for each examinee is selected from the item pool. Therefore, IRT is suitable for the CAT framework. This is another reason why IRT is used in the analysis.

The R software (version 3.1.0) was used to fit the IRT model. Estimations were performed for each component. Therefore, if an examinee took all six different types of tests, he/she would have six $\theta$ values.

Before going to the next analysis where $\theta$ is used, inappropriate items were detected and deleted and the IRT analysis was done once more. Inappropriate items were detected using item analysis, in particular, a trace line plot.

Figure 2 shows trace line plots of appropriate (left) and inappropriate (right) items. The horizontal axis of this figure is $\theta$. All the examinees who responded to these items were divided into four groups in accordance with $\theta$. The vertical axis of the figure is the ratio of the examinees who selected options 1 to 4 for each $\theta$ group. For both items, option 2 (bold line) is the correct one. Note that " $s$ " in this figure means 'skipped the item'.

The left item is appropriate because the higher $\theta$ is, the higher is the rate of the examinees correctly answering the item. This item will be examined in detail in the Results section. On the other hand, the right item is inappropriate because the higher $\theta$ is, the lower is the rate of the examinees correctly answering the item. Therefore, the right item was deleted.


Figure 2: Two trace plots
The three deletion criteria described below were applied to items responded by more than one hundred examinees. Items applied to more than one criterion were deleted.

1. The rate of the selecting correct option is almost one hundred percent for all of the four $\theta$ groups.
2. The higher $\theta$ groups do not have higher rates of selecting the correct option (right of Figure 2).
3 . The highest $\theta$ group is most likely to select an incorrect option (right of Figure 2).

Table 1 shows the numbers of deleted items and the numbers of remaining items.

Table 1: The number of deleted and the number of remaining items

|  | DEP | ANA | PARA | INF | REP | INST |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
| Deleted | 13 | 0 | 0 | 1 | 10 | 28 |
| Remained | 121 | 37 | 36 | 34 | 35 | 85 |

To examine the validity, reliability, and onedimensionality of each test, correlations between the six $\theta \mathrm{s}$, $\omega$ coefficients (McDonald, 1999) and the factor loadings in categorical factor analysis were estimated. Table 2 shows the results. Most of the correlations between the six $\theta$ s are above 0.5 , which means that the six tests all measured different aspects the same trait (reading skill). This shows that the tests have enough validity. Moreover, all six $\omega$ coefficients are very high, which shows that the tests have enough reliability. Finally, the means of the factor loadings are not small, which shows the one-dimensionality of each test, which is required in IRT.

Table 2: Correlations, omega coefficients, and mean of the

|  | factor loadings |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
|  | DEP | ANA | PARA | INF | REP | INST |
| ANA | 0.657 |  |  |  |  |  |
| PARA | 0.652 | 0.455 |  |  |  |  |
| INF | 0.541 | 0.575 | 0.541 |  |  |  |
| REP | 0.515 | 0.534 | NA | 0.620 |  |  |
| INST | 0.178 | 0.354 | 0.607 | 0.126 | NA |  |
| $\omega$ | 0.990 | 0.963 | 0.962 | 0.967 | 0.927 | 0.964 |
| mean loadings | 0.644 | 0.620 | 0.610 | 0.421 | 0.489 | 0.589 |

Note: NA means that the correlation could not be calculated because there was no examinee who took the two tests.

## 4. Related work in cognitive science

To answer the RST items, examinees need to parse sentences with unfamiliar content. In this situation, the literature suggests that human parsers tend to make errors with ambiguous sentences (Frazier \& Rayner, 1982). On the other hand, readers can construct coherence between sentences through automatic inferences (McKoon \& Ratcliff, 1992). However, studies on the human parsing process are mainly based on data collected from adult readers. Some studies suggest that there are different characteristics in the sentence processing of younger children (Otsu, 1994) and older adults (Baota et al., 2001), but there seems to be no evidence on sentence processing of young students. Moreover, despite that some school teachers recognize the possibility that the difficulties in parsing and building coherence between sentences are larger
than expected, achievement tests remain mainly concerned with higher levels of discourse.

## 5. Results of junior high school students

5.1 The appropriateness of RST Response data from six public junior high schools' students were analyzed, to show the appropriateness of the RST. These schools are in City A, whose schools are known to perform well (the best in the prefecture in 2016) in national standardized achievement tests. The responded included 613 students in grade 7, 537 in grade 8 , and 608 in grade 9 . The students responded to questions (items) taken from sentences from junior high and high school textbooks and from Japanese dictionaries.

The analyses calculated two statistics: the Correct Answer Rate (CAR) and the Rate of Students who may respond by Guessing (RSG). CAR is the percentage of items that a student correctly answered, while RSG is the rate of students who were not statistically significant in a one-sided hypothesis test assessing whether each student's correct answer rate is greater than that by guessing (null hypothesis). For example, in the PARA test, whose items have two alternatives, the expected correct answer rate by guessing is 0.5 .

First, we calculated CARs for each student in the six component tests. Although each examinee responded to different items as noted above, because these items were selected randomly, the CARs can be assumed to be comparable. The mean CAR was calculated for each grade (Table 3).

Table 3: CAR means of each grade in the six component tests

| Grade | DEP | ANA | PARA | INF | REP | INST |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 7 | 0.613 | 0.611 | 0.728 | 0.548 | 0.278 | 0.247 |
| 8 | 0.646 | 0.653 | 0.746 | 0.576 | 0.303 | 0.281 |
| 9 | 0.703 | 0.739 | 0.798 | 0.621 | 0.384 | 0.383 |

Table 3 indicates that in all the component tests, as the grade goes up, the mean CAR also increases. Generally speaking, reading skills improve as the grade goes up.

Table 4: Means of $\theta$ and RSG of each component skill

| in each grade |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Skills | grade 7 | grade 8 | grade 9 | grade 7 | grade 8 | grade 9 |
| DEP | -0.595 | -0.502 | -0.295 | 0.376 | 0.302 | 0.188 |
| ANA | -0.558 | -0.425 | -0.106 | 0.365 | 0.260 | 0.110 |
| PARA | -0.551 | -0.440 | -0.228 | 0.107 | 0.069 | 0.020 |
| INF | -0.470 | -0.443 | -0.200 | 0.660 | 0.531 | 0.423 |
| REP | -0.450 | -0.436 | -0.103 | 0.522 | 0.339 | 0.255 |
| INST | -0.154 | -0.072 | 0.232 | 0.583 | 0.505 | 0.312 |

Next, to examine the relationships between the six component skills and grades, the means of six $\theta$ s in each grade and the RSGs for each grade were calculated (Table 4). Including city A's, we collected responses from more
than 13000 participants, which are elementary-school students to adults. The $\theta$ s were estimated using the responses of all the examinees and the mean of the $\theta$ s was set to 0 for all six components. The means of the six $\theta$ s for the junior high school students therefore tend to be negative in this table. The table shows that like CARs, for all component skills, as the grade goes up, the means of $\theta$ also increase and the RSGs decrease.

Finally, to determine whether the relationships between six component skills, RSGs, and grades differ among schools, we calculated the means of the six $\theta \mathrm{s}$ and the RSGs of each grade in the six schools. The results indicate that the six junior high schools showed almost all the same tendencies as Table 4. That is, in all schools, as the grade goes up, the means of the six $\theta$ s tended to increase and the RSGs of the tests decreased.

All these results are evidence of the validity of the test.

### 5.2 Assessment of students' reading skills

It is a good sign that RSGs decrease as the grade goes up. However, the RSGs of the $7^{\text {th }}$ grade students on INF, REP and IST exceeded $50 \%$. In other words, more than half of them failed to make inferences correctly based on the knowledge given in the textbooks, map the texts into the correct images, or understand the definitions. Our statistics show that at least one fourth of students graduate from junior high school without the ability to read and comprehend textbooks at a level better than guessing. As far as we know, this is the first large-scale investigation revealing this inconvenient fact.
Read the following sentence.
Buddhism spread mainly to Southeast Asia and East Asia, Christianity
to Europe, North and South America and Oceania, and Islam to North
Africa, West Asia, Central Asia and Southeast Asia.
Choose the most appropriate answer from the given choices that correctly fill
the blank in the following sentence.
( ) has spread to Oceania.
Hinduism
Islam

Figure 3: Question DEP 103
Now, let us examine three items as to whether or not the items were tricky or too difficult for them to answer (Table 5). In DEP 103, given in Figure 3, one can choose the correct answer, Christianity, without knowledge of the four religions. Figure 2 shows the trace plot of DEP103. It shows the item was neither tricky nor inappropriate. Still, about $40 \%$ of $7^{\text {th }}$ graders, $50 \%$ of $8^{\text {th }}$ graders, and $33 \%$ of $9^{\text {th }}$ graders were not able to choose the correct answer.

Table 5: Percentage of correct answers to the three questions for each grade

| Question | grade 7 | grade 8 | grade 9 |
| :--- | :--- | :--- | :--- |
| DEP103 | 0.609 | 0.516 | 0.676 |
| REP39 | 0.070 | 0.281 | 0.298 |
| REP38 | 0.250 | 0.419 | 0.492 |

Moreover, all the $8^{\text {th }}$ grade students had learned the words appearing in REP39, in Figure 1, (i.e., circle, origin, x-axis, tangent to) in the $7^{\text {th }}$ grade. The gap between the CARs of the $7^{\text {th }}$ and $8^{\text {th }}$ grades (0.070: 0.281 ) might be explained by the unfamiliarity of these words to the $7^{\text {th }}$ graders. Then, how can we explain that only 28.1 percent of the $8^{\text {th }}$ grade students were able to choose the correct image of the text?

One may explain that unskillful readers fail to monitor when they are checking more than one condition. Here, we asked the following simpler question as REP38: "The circle passes through the origin O". The gap between CARs of REP 38 and 39 of the $8^{\text {th }}$ and $9^{\text {th }}$ grades might be explained by monitoring failure. Still half of the $9^{\text {th }}$ grade students failed to answer correctly. We could not find relevant literature to explain this phenomenon.
5.3 Correlation with schools' characteristics We calculated correlations between these statistics and the schools' characteristics; Distances from the nearest station (Dis), the Number of Students (NS), and Rates of Students receiving Financial Help for school attendance (RSFH) in each grade (Table 6).

Table 6: Correlations of means of $\theta$ and RSG with school characteristics

| Means of $\theta$ |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Rkills | Dis | NS | RSFH | Dis | NS | RSFH |
| DEP | -0.534 | 0.302 | -0.540 | 0.434 | -0.324 | 0.491 |
| ANA | -0.315 | 0.104 | -0.451 | 0.098 | 0.018 | 0.359 |
| PARA | -0.294 | 0.101 | -0.313 | 0.236 | -0.105 | 0.412 |
| INF | -0.262 | 0.251 | -0.288 | -0.001 | -0.114 | -0.016 |
| REP | -0.235 | 0.143 | -0.291 | 0.347 | -0.209 | 0.408 |
| INST | -0.156 | 0.279 | -0.310 | 0.160 | -0.345 | 0.237 |

Table 6 shows that in all the tests, the correlations of the means of $\theta$ with Dis and RSFH are negative, but positive with NS, and that in almost all of the tests, the correlations of RSG with Dis and RSFH are positive, but negative with NS. These results imply that students whose schools are near a station, are large, and offer less financial support tend to have higher component skills and therefore may respond to items not by guessing. We will continue to investigate these findings.

We asked examinees to answer a questionnaire including items on their attitudes toward reading and likes and dislikes of school subjects. City A conducts standardized achievement tests every year. We are planning to assess the relationship between the results of the RST, the responses to questionnaires and the scores of the achievement tests.

## 6. Comparison of performances with automatic dependency structure analyzer

We processed the test sentences of the RST dependency analysis questions (DEP) with the CaboCha parser (Kudo \& Matsumoto, 2002) and analyzed the errors. We hoped that
the analysis of errors made by a machine would help us to understand the human errors. CaboCha is a dependency parser based on Support Vector Machine. It was trained only on a news corpus, and its accuracy on news text is around $90 \%$ at the dependency relation level and $50 \%$ at the sentence level. The comparison with the human responses provided here is hence preliminary in that we expect the parser's accuracy will improve by retraining it on textbook data.

We analyzed the items on which we collected the responses from more than 100 students. DEP is a set of multiple-choice questions that ask for a phrase that stands in a certain grammatical relation to a phrase in a test sentence. We chose the answer based on CaboCha's output. The rate of correct answers by CaboCha was $66 \%$. For example, CaboCha parsed DEP103 (Figure 3) correctly.


Figure 4: DEP $\theta$ value of humans and CaboCha
Figure 4 shows the distribution of human $\theta$ and the estimated $\theta$ of CaboCha on the DEP questions. It reveals the mode of human $\theta$ is only slightly above that of CaboCha.

The most common error types made by CaboCha were as follows (the numbers in parentheses are the fractions of the errors of these types).

1. When the test sentence includes a phrase inside parentheses (7\%)
2. When the sentence is long ( $11 \%$ )
3. Unusual use of a comma or no use of comma (2\%)
4. Choice of the attachment site of a subordinate or parallel verb phrase ( $60 \%$ ): CaboCha made mistakes most frequently on the sentences including more than one subordinate or parallel verb phrase (VP). It corresponds to a sentence in the form of "... Verb ... VP1 ... VP2 ..." in English, where VP2 has two possible attachment sites, Verb (matrix verb) and VP1 (another subordinate VP), as in:

Adaptive immunity [verb includes] humoral immunity in which B cells [vP1 form proteins called antibodies] [vP2 to remove extracellular pathogens], and ...(snip).

There is no syntactic clue to choose between the two possibilities. Thus, it should be judged by meaning, and hence, it is difficult for CaboCha.
5. Wrong word segmentation (5\%)

In Japanese, words are not separated by whitespaces as in English. CaboCha often fails to segment technical terms correctly.

The errors of type 1, 3, and 5 would be reduced by retraining the parser on textbook data. On the other hand, the errors of type 4 require context and meaning to fix them. Table 7 lists the rate of correct answers by the human examinees on the questions on which CaboCha made mistakes. It suggests the choice of subordinate or parallel VP attachment is also difficult for humans. While Table 3 indicates that students gradually acquire the skill and knowledge to do it, it would remain a hard problem for an automatic parser since it requires some understanding of the meaning and context of a sentence.

We would like to confirm and extend these findings by examining more diverse samples collected through RST. Of special interest is a further analysis of the errors of human and automatic parsers on the basis of the cognitive studies on sentence processing (Mitchel 1994) such as the gardenpath theory (Frazier \& Fodor, 1978; Frazier \& Rayner, 1982) and minimalist hypothesis (McKoon \& Ratcliff, 1992).

Table 7: Human CARs on the questions on which the automatic analyzer made mistakes

| Error type of CaboCha | Human CAR |
| :--- | :---: |
| Parenthesized phrase | 0.584 |
| Long sentence | 0.572 |
| Unusual use or no use of comma | 0.615 |
| Attachment of subordinate VP | 0.549 |
| Word segmentation | 0.786 |

## 7. Conclusion

We developed a new reading skill test (RST) to measure six component skills relevant to reading. By analyzing the responses to the RST, we confirmed that it has enough reliability and validity. In addition, we analyzed response data of Japanese junior high school students to the RST, and the results implied that, surprisingly, the six component skills might be lower than expected. Finally, we compared the performances of the students with those of a Japanese dependency parser. The results implied that students do no better than a machine in dependency analysis.

## References

Arai, H. N., \& Matsuzaki, T. (2014). The impact of A.I. on education - Can a robot get into the University of Tokyo?. Proceedings of the 22nd International Conference on Computers in Education (pp. 1034-1042).
Balota, D. A., Cortese, M. J., \& Wenke, D. (2001). Ambiguity resolution as a function of reading skill, age, dementia, and schizophrenia: The role of attentional control. In Gorfein, D. S. (Ed). On the consequences of meaning selection: Perspectives on resolving lexical
ambiguity (pp. 87-102). Washington, DC, US: American Psychological Association.
Dagan, I., Roth, D., Sammons, M., \& Zanzotto, F. M. 2013. Recognizing Textual Entailment: Models and Applications. Morgan \& Claypool.
Frazier, L., \& Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. Cognition, 6(4), 291-325.
Frazier, L., \& Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology, 14, 178-210.
Fujita, A., Todo, N., Sugawara, S., Kageura, K., \& Arai, N. H. (2016). Development of a Reading Skill Test to Measure Basic Language Skills. Proceedings of the 8th IEEE International Conference on Technology for Education (pp.156-159).
Hambleton, R. K., \& Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Nijhof.
Kudo, T., \& Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. Proceedings of the 6th conference on Natural language learning-Volume 20 (pp. 63-69). Association for Computational Linguistics.
Lord, F. M., \& Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
McDonald, R. P. (1999). Test theory: A unified treatment. L. Erlbaum Associates, Mahwah, NJ.
McKoon, G., \& Ratcliff, R. (1992). Inference during reading. Psychological Review, 99, 440-466.
Mitchell, D. C. (1994). Sentence parsing. Handbook of psycholinguistics, 375-409.
Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., \& Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. (pp. 915932).

OECD (2016). PISA 2015 Results in Focus. Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf
Otsu, Y. (1994). Early acquisition of scrambling in Japanese. In: Teun Hoekstra \& Bonnie D. Schwartz (eds.) Language Acquisition Studies in Generative Grammar, 253-264. Amsterdam: John Benjamins Publishing.
Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., \& Zhang, Y. (2012). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. Proceedings of the Joint Conference on EMNLP and CoNLL - Shared task. (pp. 1-40).
RAND Reading Study Group (2002). Reading for understanding: Toward an R\&D program in reading comprehension. Santa Monica, CA: RAND Education.
van der Linden, W. J., \& Glas, C. A. W. (eds.). (2010). Elements of Adaptive Testing, New York, NY: Springer.

# Perception Meets Examination: Studying Deceptive Behaviors in VR 

Carla Aravena ${ }^{1 *} \quad$ Mark Vo ${ }^{1 *} \quad$ Tao Gao ${ }^{2} \quad$ Takaaki Shiratori ${ }^{1} \quad$ Lap-Fai $\mathrm{Yu}^{1}$<br>${ }^{1}$ University of Massachusetts Boston ${ }^{2}$ General Electric<br>*Equal Contributors


#### Abstract

Students cheating on an exam in an academic setting creates an environment where one person (the student) must reason about the perception of another (the teacher). In exploring the student's mindset, trends concerning how humans make decisions based on their understanding of another human's intentions and knowledge can be uncovered. In this work, we study human cheating behavior through simulated examinations in virtual reality, showing that the teacher's animacy and orientation plays a large part in the student's reasoning of the teacher's awareness. By utilizing a virtual classroom setting and accurately tracking a users behavior (through head tracking, eye movement, etc.), we have also demonstrated how a novel virtual reality approach can be used for such experiments involving human behavioral observations, which can be further explored in other cognitive science research experiments.


Keywords: deceptive behavior; behavior modeling; virtual reality; game experimentation; human vision; Theory of Mind

## Introduction

Imagine you are a teacher surveying your classroom of students who are taking a final exam. You know that the exam is very difficult and you expect that some students may attempt to cheat. Perhaps you find yourself scanning the room for signs of cheating behavior. But, what do such signs look like? A student looking around the room could simply be in the process of managing their thoughts. If a student is looking straight at you for a while, are they waiting for you to look away or are they wondering if they should ask you a question? How can you determine which students are simply feigning innocence while planning a cheating attempt? What if you have a large classroom of students; how would you be able to keep an eye on all of them at once?

Cheating remains a common problem in examinations, which creates an interesting scenario for cognitive psychology research in its need to have one person relying upon their perceptions of the intentions of another person to behave. In a classroom, the student must rely upon the teacher's actions and movements in order to determine when the best window to cheat off of another student's exam is. How the student reasons about a teacher's intentions, formulates the right time to cheat, and performs certain behaviors to "trick" the teacher are areas that have yet to be explored in a way that does not simply rely on asking students how they cheat. Discovering how a cheater successfully accomplishes his or her task is related to the cognitive science concept of the Theory of Mind, defined as "a mechanism that helps [one] to make sense of the behaviour of others in specific contexts and to predict their next action" (Dias, Aylett, Paiva, \& Reis, 2013). In the situation of cheating during an examination, the student must predict what the teacher is perceiving in order to gauge the success of their cheating attempt.

In this work, we study the signs and visual behaviors of cheating through virtual reality. Having actual students in real life to participate in our experiments and produce the cheating data gives us a realistic dataset to visualize and analyze how university students may actually cheat on classroom examinations. We recreate the cheating situations captured in

(a) Cheating in a real classroom

(b) Our game

Figure 1: (a) Cheating in a real classroom during an exam. (b) A screenshot of our game that mimics the settings of a real exam. The player cheats during a simulated examination in virtual reality while avoiding getting caught by the teacher similar to a real-life situation.
virtual reality to establish ecological validity and track various factors of the participants' behavior with respect to the state of the teacher.

Figure 1(b) shows a screenshot of our game. Our game premise is inspired by Sunken Places' virtual reality game "Classroom Aquatic". In their game, the player is asked to complete their multiple-choice examination within a time limit by cheating off of other students' examinations, avoiding the teacher's and students' gazes, and using objects as distractions to answer the questions correctly. This premise takes an ordinary situation and recreates it in virtual reality, encouraging players to use deceptive behavior to complete a task. From a cognitive science standpoint, such a game setting is interesting as it provides a good scenario for studying how humans exhibit cheating behavior and for understanding how humans judge the actions of others to be deceptive.

The major contributions of our work include the following:

- Analyzing how students cheat on classroom exams based on their behavior within a virtual setting.
- Demonstrating that virtual reality environments can produce a way to objectively gather data on human behavior.


## Hypothesis

Through our experiments, we seek to answer the question of how students use information of a teacher's movements to perceive the teacher's intent (i.e. how does the teacher's movement and placement in the room affect the student's judgment on when it is "safe" to cheat?). This question contributes to the discussion of the Theory of Mind and how humans process information of others' behaviors to make judgments. Inspired by the previous research on animacy and how that influences a human's decision-making process, we hypothesize that (1) players will cheat most often when the teacher is completely turned away from them. For this hypothesis, we believe that players will assume that the teacher will keep walking in the same direction, away from them, for a long enough time for them to safely cheat. As for how the players will choose to cheat, we also make the hypothesis that (2) the players will not look directly at another students' exam, and will instead cheat while employing their peripheral
vision in order to appear innocent to the teacher and also to keep an eye on the teacher as much as they can.

## Related Work

Common Cheating Behavior in Exams. Prior methods used to prevent cheating and to catch potential cheaters are common in research literature on the topic of student cheating. The factors behind why students cheat is welldocumented and analyzes how certain variables affect the students' likeliness to cheat including gender, GPA, parental pressure, etc. (Batool, Abbas, \& Naeemi, 2011). Although such findings are useful, their applications in preventing cheating are difficult to place given that one may not be able to generalize which students are likely to cheat based on their personal life (such as parental pressure) as well as the fact that some of these variables may not be accessible by an educational institution. An examination of classrooms and learning environments that are best able to dissuade cheating has been performed in prior research (Cizek, 2003), which found that students tend to cheat less often when (a) classes are smaller; (b) classroom conditions (both physical and instructional) are established that are conducive to learning; (c) instruction, assignments, and examinations are clear, welldesigned, meaningful, and relevant; (d) teachers take reasonable steps to prevent cheating. While (a) is an issue that is generally assumed to be handled by the educational institution, and (b) and (c) are obvious goals that any class setting should aim to achieve, it is (d) that we are most interested in cultivating by finding out how students actually cheat.

Many of the methods used to discover how cheating is conducted is qualitative and frequently relies upon students' readily answering questions on the matter (e.g., (Batool et al., 2011; Shon, 2006; Yee \& MacKown, 2009)). Relying upon student responses alone is problematic for the obvious reason that students could easily lie and fail to realize their exact behavioral habits when trying to be deceptive. By taking data from "cheaters" in a quantitative way through automatically tracking their actions throughout a cheating session, this problem would be resolved. Previous attempts to automate detection have applied text-mining based approaches, which only evaluate cheating after the fact and is limited to open-ended exams (Cavalcanti, Pires, Cavalcanti, Pires, et al., 2012). We are more interested in studying the cheating behavior at the instant it occurs, which forces the student to rely upon his or her reasoning of the teacher's perception, as it is highly relevant to the Theory of Mind.
Prior Cognitive Science Studies on Deceptive Behavior. According to previous studies, students cheat by first "qualifying the professor": determining how likely they could get away with cheating based on evaluating the teacher's behavior (Batool et al., 2011). This is an area that we would like to explore further as it is an extension of a human's Theory of Mind, defined as the ability to infer the full range of epistemic mental states of others, i.e., beliefs, desires, intentions and knowledge (Dias et al., 2013). This is a mechanism that helps to make sense of the behavior of others in specific contexts and to predict their next action. Through our experiments and analysis, we investigate the Theory of Mind in that we are trying to find out how people figure out whether or not the teacher is catching on to them and, based on the teacher's behavior, when it is safe to cheat.

In a recent study, (Dias et al., 2013) established a game in which different artificial intelligence models with varying levels of Theory of Mind were set to perform deceptive behaviors in a game setting, finding that those with higher levels of Theory of Mind (and thus a greater ability to reason and make inferences on others' behaviors) were more successful in deceiving, showing how frequently deception is founded when making judgments upon the actions of others. In our experiments, we examine the behaviors of the human players against the actions of a virtual teacher to discover how humans may similarly reason about the teacher's behavior in order to be successfully deceitful.

Gao's et al. (Gao, Newman, \& Scholl, 2009) research concerning how animacy is involved in a person's reasoning on intent provides a possible explanation for what we are observing in our players' behavior. Their study found that humans rely on the direction and orientation of an object to perceive its animacy, movement, and intent. Subjects were more likely to reason that a wolf shape was "chasing" a sheep shape if orientation was present. Therefore, animacy and orientation of movement may be influential when humans apply the Theory of Mind. In our experiment, we will observe to what extent the player will use the teacher's animacy and movements in making their decision of what time is best to cheat.

Investigating how humans reason about their game opponents in game settings is well-documented and is helpful to those in game-development who are concerned with realistic AI mechanisms. Prior research in turn-taking games follows a similar setup in exploring the Theory of Mind and looks into "which rules govern human strategic thinking" (Halder, Sharma, Ghosh, \& Verbrugge, 2015). Such findings inspired us to determine which aspects of human behavior we should track and record, taking note to include timing and duration of actions (such as cheating) in our own experiments. However, unlike such studies, we are dealing with continuous gameplay and so we cannot delineate a player's choices in a clearcut manner as we would in a turn-taking game. Having a continuous gameplay setup serves our purpose in mirroring a situation that would take place in everyday life.
Using Games for Cognitive Science Research. The involvement of videogames into academic research has an established, albeit relatively new, presence in academic research. Previous branches of research involving videogames in the fields of Cognitive and Social Psychology have studied the effect of entertaining games on basic cognitive skills while another has researched the success of educational gaming (Killingsworth \& Clark, 2013). Instead of contributing to either of these branches, our experiment seeks to instead use a videogame format in order to observe human behavior, which is an ability afforded to us through the use of virtual environments that are now even more immersive with the aid of a suitable virtual reality headset.

Previous research studies have used virtual reality technology to recreate real-life situations and examine human behavior in such instances (Kozlov \& Johansen, 2010; Olivier, Bruneau, Cirio, \& Pettré, 2014; Li, Liang, Quigley, Zhao, \& Yu, 2017; Rovira, Swapp, Spanlang, \& Slater, 2009). These studies have found that virtual scenarios are a good fit in measuring human responses to real events due to the close correspondence in human behavior between the two environments


Figure 2: Experiment setup. The participants performed the examinations using (a) an Oculus Rift virtual reality headset and (b) a Microsoft Xbox 360 game controller.
as well as the fact that the virtual reality settings can be finetuned by the researchers (Rovira et al., 2009). If the virtual setting is close enough to pass as its material counterpart, it is safe to claim that observations made on participants are valid.

## Approach

Overview. To study how students cheat, we seek to gather information on how players cheat on exams within a game session that models the real-life situation. In our experiments, the player takes on the role of a student taking a multiplechoice paper exam. The player is immersed in a virtual classroom environment and is asked to achieve two objectives: (1) answer as many questions on the examination correctly as possible by cheating off of other students' exams and (2) avoid detection of the teacher. If the teacher is able to catch the student in the act of cheating, the player will fail the task and that round will end.
User Interaction. Figure 2 depicts the experiment setup for the game sessions. We decided to use an Oculus Rift DK2 as our virtual reality hardware setup due to its ability to conduct the sessions in a small space and immerse players within the game setting. A Microsoft Xbox 360 wireless controller is used as the primary controller for which the player cheats and answers questions with. The player can change his viewpoint and shift his gaze as he would in reality with the use of the virtual reality headset. Other primary modes of interaction will be the act of cheating, which is triggered when the Left/Right Bumper Button is held down by the player (the answer retrieved from the examination paper of the classmate sitting on the left or right will be displayed); and the abilities to switch through and answer examination questions, which are respectively achieved by pressing Left/Right on the directional pad and either $\mathrm{A} / \mathrm{B} / \mathrm{Y} / \mathrm{X}$.
Virtual Environment Design. We designed a 3D virtual environment in the form of a classroom in which the player is a student surrounded by other students at individual desks, much like how most classrooms are set up at colleges today. The classroom objects, the teacher, the player character, and the other 3D student models were found at the Unity Asset Store. The design of the classroom is not flashy nor distracting so that players can focus on the task at hand.
Gameplay. The exams of the students to the left and right sides of the player will contain a correct answer to one of the questions on the exams; these answers will change in a preset interval every 5 seconds so that all answers can be obtained by cheating. The teacher moves along a preset path to hit certain points along the classroom, rotates about every six seconds, and seems to be checking over the room for signs of potential cheating. The path of the teacher is shown in


Figure 3: The path of the teacher during the game. Walking through the path once takes about 80 seconds and the teacher cycles through the path until the player finishes his exam in the round, which usually takes about $100-200$ seconds. Arrow directions correspond to the orientations of the teacher.

Figure 3. The questions that show up on the player's exams are derived from the website "Trivia Country" (Trivia Country, 2016). We chose questions from this website due to their specificity and very low chance that participants would know the correct answers. For example, one question we ask participants on the exam is: The Philadelphia mint started putting a 'P' mint mark on quarters in which year? Answer Choices: a)1980; b)1960; c)1950; d)never. We give the participants an incentive to cheat by telling them that the number of correct responses they answer is important to getting the maximum amount of money. Since the multiple-choice questions are derived from random, factual knowledge that the participants will most likely not know the answer to, cheating becomes a necessity to gathering the correct answers. Furthermore, we give an incentive for the players to cheat wisely by telling them that, if they are caught cheating, their potential to gather the maximum amount of points per round will be cut short. The player is also told that the teacher is powered by an advanced artificial intelligence and a human vision model so that he can perceive the environment like a real human, so that the player treats the teacher's actions with severity. In actuality, the teacher has a limited, cone-shaped range of vision that functions as a spotlight to detect cheating if the player is within the cone of vision and holding down the cheat button at the same time.

Data Collection. In order to gather data through the game session playthroughs, we constantly track the position and orientation of the teacher model. The position and orientations of the the teacher's head and the player's head movements are recorded during gameplay for later analysis. We measure the user's visual attention through game scripts that track an approximation of the player's gaze by creating a ray at a point where the player is looking, as well as taking note of at what time the player is cheating and which paper the player is cheating off of (either the left or right exam). During the player sessions, we also record a video of the gameplay, from both the teacher's and the student's point of view. Refer to our supplementary video which showcases some of this footage.


Figure 4: Human's field of vision.

## Experiments

Subjects. 25 participants were recruited to conduct this study with 22 being college students and the rest in high school and middle school. The subjects were recruited through emails sent out to the Computer Science department at UMass Boston as well as through word of mouth. About $60 \%$ of the subjects were of a Computer Science background. None of the students had played the aforementioned Classroom Aquatic game. Of all the subjects, there were 14 males and 11 females. Each player session consisted of 3 rounds. The participants participated individually for approximately 10 minutes for each player session and so completed each round in about 3 minutes. Participants were paid $\$ 10$ for their time upon completion of their experiments.
Implementation. The virtual reality program was run on a desktop computer installed with an AMD FX 8350 EightCore Processor $(4.00 \mathrm{GHz}), 16 \mathrm{~GB}$ memory (RAM) and a MSI Radeon R9 390 8GB Graphics Card. Participants were given a Xbox 360 wireless controller and an Oculus Rift DK2 headset to play the game created using the Unity game engine.

Briefing. Participants were given instructions on how the controls of the game functioned through a short tutorial session. This tutorial session allowed players to understand how to answer exam questions, how to cheat, as well as to become accustomed to the game surroundings. The virtual teacher was not introduced in the tutorial session. The players were then told the following message before the start of the game session:

You are a student looking to get a decent grade on an inclass, multiple-choice exam. The answers to the questions correspond to the $A-B-Y-X$ buttons on your controller. To gain a good score, you can cheat off of the exams of the students to the left and right of you. Hold down the left or right bumper to cheat off of their exams, which will have a correct answer to a question on the exam and will change answers in a set time interval. There will be three rounds with five questions on an exam each. You gain a point for each question that you correctly answer. You can earn up to 10 dollars depending on how many points you have at the end of the three rounds. The teacher in this game is wearing a blue suit and is designed with advanced AI and vision capabilities such that they can perceive the world like real humans can. If the teacher catches you teaching, it's game over for that round and you will go on to the next round.


Figure 5: The angles by which players cheated off of the left or right exam.

## Results and Analysis

Using the data we have gathered during the game experiments, we plot several different variables against each other to draw conclusions on students' cheating behavior. We ignore the first round of data for each participant as the latter two rounds are far more likely to contain the players' intent and are not marred by first-time errors.A total of 255 cheating attempts were recorded (not counting the first round).
Peripheral Vision. We first examine the player's use of peripheral vision in order to cheat by inspecting the angle between the player's estimated gaze and the exam they are cheating off of (either the left or the right exam). The angles at which the player cheated off of the exams are shown in Figure 5. The number of cheating attempts made off of the left exam were 117 while the right exam had 138 in total. We refer to the human's field of vision (Bhise, 2011) depicted in Figure 4 to compare the angle between the player's central ray and the left or right exam to this diagram in order to determine which area of their vision the participants used to cheat.

For those who cheated on the left exam, a mere $2 \%$ cheated using their central vision. $24 \%$ of participants utilized their near-peripheral vision. The majority of the left attempts, $70 \%$, cheated within the angle range considered to be midperipheral vision. The remaining $4 \%$ of these cheating attempts were conducted at an angle greater than 60 degrees. Similar results were found for the cheating attempts done on the right exam. $0 \%$ used their central vision on this side while $38 \%$ relied upon their near-peripheral vision. The remaining $62 \%$ all took place under the player's mid-peripheral vision.

Instead of staring directly at the other students' exams, the subjects chose to keep the exam they were cheating off of to the side of their vision. The reason for this result could be the effect of the player exhibiting signs of "sneaky" behavior. The player's cheating status is turned off or on simply by holding down the cheating buttons, which was made clear to the participants during the tutorial. Therefore, looking directly at another exam or not has no bearing on how quickly one can "stop" cheating. Because of this, we can say that the reason for this behavior is due to the player constantly keeping in mind the teacher, as the player exhibits signs of feeling wary of the teacher. By only "sneaking" glances at a nearby exam they are cheating off of, the student is trying not to be too "obvious" in their cheating. This action reveals that the


Figure 6: Appearances of the teacher at different relative orientations with the player.
student is letting their perception of the teacher's perception shape their own actions. The students exemplify the Theory of Mind: they are reasoning and making inferences on another's perception based on that person's behavior and are changing their own behavior accordingly (Dias et al., 2013).
Teacher's Orientation. The teacher in our experiment, in his pre-figured path around the room, pauses about every six seconds to rotate back and forth (as if he is scanning the room for signs of cheating) and turns his body like a normal person walking around a room would. Figure 6 shows the teacher's appearances at different orientations. We examine when
 tween the teacher's teacher's orientation relative to the player by noting the angle $\theta$ by which the teacher sees the student (depicted in Figure 7).

According to Figure 8, students mostly cheated during the moments where the teacher was farther away from the student and when the student was on the outskirts of the teacher's perceived vision. Out of 255 recorded cheating attempts, only $7 \%$ of attempts were undergone when participants were within the teacher's central range of vision (and got caught by the teacher during these attempts). $12 \%$ of cheating attempts were done under the teacher's near-peripheral range of vision and $9 \%$ under the teacher's mid-peripheral range. The majority of cheating was split between when the student was within the teacher's far-peripheral range of vision (the dark purple points, taking up $34 \%$ of all cheating attempts) or when he was not facing the student at all (the black points, which make up $38 \%$ ). When the teacher can see the student only with his far-peripheral vision or not at all, the player was able to be more certain that the teacher was not paying attention to him and therefore cheated the most during those moments.
Teacher's Position. Figure 8 also shows that more than half of all cheating attempts took place when the teacher's distance exceeded 3 meters, attesting to the player's feeling that, the farther away the teacher is, the "safer" the cheater is. Figure 9 shows further that students chose to initiate cheating when the teacher was located in areas that were farther away from them and did not feel safe enough to cheat when the teacher was not in sight. The moments in the teacher's path with the most frequent cheating attempts, at around $32 \%$ of all cheating instances, were located close to Area A, where the teacher was within the student's sight, was not relatively close to the student, and was moving in a clear path towards the back of the room. At this point, the student was able to cheat off of the left exam and keep the teacher within their line of sight. The least amount of cheating occurred around Area B, where the teacher was completely out of the student's range of vision. We see a brief peak in cheating when the


Figure 8: Each dot refers to an occurrence of cheating, plotted against the distance between the player and the teacher, and the angle by which the teacher saw the student.
teacher was close by but had his back turned to the player, in Area C. The player might assume they were safe at this point, thinking that the teacher would continue to walk away from them. However, the teacher turned around shortly after and walked by them again, moving back towards the front of the room. Cheating instances only became more frequent again, with around $9 \%$ of all cheating attempts happening, when the teacher had traveled a farther distance away in Area D.

The players' cheating patterns here reveal that students rely upon their perception of the teacher's perception to make judgments on the teacher's ability to spot them cheating. This explains why they choose to cheat only when they are able to see the teacher. It is also clear that the players used the animacy of the teacher to make predictions of the teacher's perception and behavior as they were more likely to cheat when he had his back turned or was walking away.

## Discussion

Feedback and Observations. We asked the participants how they felt about the experiment after receiving their compensation. The general feedback from the players was that the game was "fun" and "interesting". They claimed that, during the game, they felt slightly afraid of the teacher model. From our observations, we noted that, during the start of the first round, the participants would be reluctant to look away from their paper at all once they saw the teacher. Only after we reiterated the rules of the game (that the student is not cheating unless the corresponding button is held down) did the players feel comfortable looking around the room. However, even still, very few players turned completely around to see where the teacher was when he was behind the student, even though they could have done so with no penalty. In reality, few students would turn all the way around to spot the teacher as this might appear suspicious. This attests to the participants taking the game seriously and that they performed as they would in an actual classroom.
Limitations. The participants' experience with virtual reality and computer games may affect their performance in our experiments. Most of our participants were students from the Computer Science major. Many CS majors are exposed to and regularly enjoy computer games, so they may have had a slight advantage in escaping cheating during the game sessions, compared to participants from a non-CS background. Furthermore, virtual reality headsets affected different partic-


Figure 9: Cheating frequencies when teacher was at different locations along his path. Redness corresponds to the cheating frequency. Most cheating occurs when the teacher was at the far left. Refer to text for detailed description of the observations.
ipants differently, especially when some were wearing glasses or were just using a virtual reality device for the first time. As a result, some might have experienced slight discomfort with the virtual reality headset, perhaps leading to less ease in playing the game. We are also limited by the teacher's lack of realistic head movement. The teacher's torso and head move in the same direction throughout the game, which the player may or may not have realized as they were playing. Because of this, the results may be slightly varied from a reallife classroom setting.
Conclusion. In this study, we have verified our hypothesis about deceptive human behavior during an exam as the players (1) cheated most frequently when the teacher was turned completely away from them and (2) used their peripheral vision frequently to cheat off of other exams, showing that the teacher's animacy and orientation play a significant role in the student's likeliness to cheat and that their own judgment about "appearing" suspicious affected how they decided to cheat. By utilizing a virtual classroom setting and accurately tracking a user's behavior, we have also demonstrated how a novel virtual reality approach can be used for such experiments involving human behavioral observations, which can be further explored in other cognitive science research experiments. An interesting venue for future work is to use the human behavior data collected from virtual environments to train a realistic, human-like AI that can exhibit human deceptive behaviors. The application of new virtual reality devices that can accurately measure eye-tracking to this experiment would lend itself to the peripheral vision analysis as well. Furthermore, because students often cheat in collaboration with other students, our extension of this work will include a cooperative mode that allows two players to help each other cheat. We are also developing a quantitative method analysis of the player's cheating by establishing a metric to estimate the amount of visual attention the student is receiving from the teacher at any given point in time.

## Acknowledgments

This research is supported by the Oracle Undergraduate Research Fellowship, the McNair Program, the UMass Boston StartUp Grant P20150000029280 and the Joseph P. Healey

Research Grant Program provided by the Office of the Vice Provost for Research and Strategic Initiatives \& Dean of Graduate Studies of UMass Boston. This research is supported by the National Science Foundation under award number 1565978. We also acknowledge NVIDIA Corporation for graphics card donation.

## References

Batool, S., Abbas, A., \& Naeemi, Z. (2011). Cheating behavior among undergraduate students. International Journal of Business and Social Science, 2(3), 246-254.
Bhise, V. D. (2011). Ergonomics in the automotive design process. CRC Press.
Cavalcanti, E. R., Pires, V. F., Cavalcanti, E. P., Pires, C. E., et al. (2012). Detection and evaluation of cheating on college exams using supervised classification. Informatics in Education-An International Journal(Vol11_2), 169-190.
Cizek, G. J. (2003). Detecting and preventing classroom cheating: Promoting integrity in assessment. Corwin Press.
Clark, K. (2008). Professors use technology to fight student cheating. US News \& World Report.
Dias, J., Aylett, R., Paiva, A., \& Reis, H. (2013). The great deceivers: Virtual agents and believable lies. In Cogsci.
Gao, T., Newman, G. E., \& Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. Cognitive psychology, 59(2), 154-179.
Halder, T., Sharma, K., Ghosh, S., \& Verbrugge, R. (2015). How do adults reason about their opponent? Typologies of players in turn-taking games. In Cogsci.
Killingsworth, S. S., \& Clark, D. (2013). Connecting learning goals and component cognitive skills in digital games. In Cogsci.
Kozlov, M. D., \& Johansen, M. K. (2010). Real behavior in virtual environments: Psychology experiments in a simple virtual-reality paradigm using video games. Cyberpsychology, behavior, and social networking, 13(6), 711-714.
Li, C., Liang, W., Quigley, C., Zhao, Y., \& Yu, L.-F. (2017). Earthquake safety training through virtual drills. IEEE Transactions on Visualization and Computer Graphics.
Liang, W., Zhao, Y., Zhu, Y., \& Zhu, S.-C. (2015). Evaluating human cognition of containing relations with physical simulation. In Cogsci.
Olivier, A.-H., Bruneau, J., Cirio, G., \& Pettré, J. (2014). A virtual reality platform to study crowd behaviors. Transportation Research Procedia, 2, 114-122.
Rovira, A., Swapp, D., Spanlang, B., \& Slater, M. (2009). The use of virtual reality in the study of people's responses to violent incidents. Frontiers in Behavioral Neuroscience, 3, 59.
Shon, P. C. (2006). How college students cheat on in-class examinations: creativity, strain, and techniques of innovation. Info: Ann Arbor, MI: MPublishing, University of Michigan Library, 1.
Trivia Country. (2016). Trivia country - multiple choice trivia quiz questions. http: //www.triviacountry.com/ M1-Multiple-Choice-Trivia-Questions.htm. (Accessed: 01-30-2016)
Yee, K., \& MacKown, P. (2009). Detecting and preventing cheating during exams. PEDAGOGY, NOT POLICING, 141.

# Against the group actor assumption in joint action research 

Ed Baggs (e.baggs@ucl.ac.uk)<br>Bartlett School of Architecture, University College London, 22 Gordon Street, London WC1H 0QB<br>Department of Psychology, University of Cincinnati, Cincinnati, Ohio 45221


#### Abstract

A central assumption in joint action research is that in order to explain how individuals act as part of a group, we must first explain how the group comes into existence. This assumption has led to an unnecessarily narrow research programme: research has focussed largely on interpersonal coordination mechanisms. I outline an alternative approach predicated on a dynamic conception of the ecosystem. On this view, there is no need to assume that actors must first constitute a group agent with their fellows before entering into coordinated action. Such coordination can be more efficiently explained by recognizing that all actions perturb the structure of the ecosystem itself in a manner that can alter the action possibilities available to neighbouring actors. This move allows us to overcome entrenched debates over the nature of shared intentionality, and to instead focus on practical interventions in multiactor settings.


Keywords: joint action; shared intentionality; ecosystems; ecological psychology

## Introduction: The group actor assumption

The group actor assumption is not a commonly-used term; it is a term I introduce here to characterize the way that researchers working in the joint action tradition typically understand their project (Sebanz, Bekkering, \& Knoblich, 2006). The assumption might be analysed into the following four claims:

## 1. the basic form of action is individual action

2. the group is to be understood as a set of constraints and structures placed on individual action (the tactic here might be to reduce joint action to structures in individual minds, or, alternatively, it might be claimed that the group itself is an emergent structure that constrains the individual components)
3. the existence of the group must be explained in terms of the nature of the individuals involved, and the coordination activities they engage in
4. the group must come into existence first before a joint action can be implemented

I do not, however, wish to make a fetish out of these four items. So let us immediately simplify this analysis by replacing it with the following statement:

## 'The group actor assumptions reifies the group'

That is to say, under the group actor assumption, it is understood that the study of activity involving multiple actors requires a special mode of analysis. The group actor assumption says that the group is a real entity, and, furthermore, it
asserts that this groupness will play a crucial role in explaining whatever it is that we wish to explain about some phenomenon of interest. The group actor assumption invites us to divide the world, a priori, into individual actions and joint actions, and says that the latter type requires an additional layer of explanation over and above what is required for the former.

I will provide evidence for all of these claims below. But first it will be useful to consider a specific example of the kind of joint action that we might want to study. I will use this to argue that the group actor assumption leads to an unnecessarily narrow research programme. In the second half of the paper, I outline an alternative approach which avoids the problems identified, appealing to the concept of the ecosystem.

## Case study 1: Children's soccer training

When young children, say around age five, are first corralled onto a soccer pitch they can easily enough be divided into two teams and encouraged to act out a soccer game. What one will notice, however, is that these teams exhibit a striking absence of structure. The ball will be propelled in some direction, whereupon the children will chase the ball en masse, and then different children will try to coordinate the necessary limb movements in order to make some decisive connection with the ball. At the end of this process, the ball is propelled in some new direction, and the cycle begins again. What is going on here? Here is a succinct explanation: 'A child's basic urge is to run and chase the ball' (Quinn \& Carr, 2006). The children's chasing-urge, coupled with the quasi-random trajectory of the ball at a given moment, produces a situation in which all of the action appears to be reactive to the current spatial configuration; indeed, the ball itself almost appears to be driving the action.

Contrast this with an accomplished team performance played to a high degree of skill. Take the goal scored by Esteban Cambiasso for Argentina against Serbia and Montenegro at the 2006 World Cup. This is a famous goal because it came after the team had completed a sequence of 24 uninterrupted passes, and it thus serves as an object lesson in dominant possession-based soccer. But let's consider just the last four passes here. For each of these passes the player receiving the ball is already on the move and the passing player, detecting this movement, plays the ball into the space just in front of where the recipient is going to be. The whole sequence of four passes goes off in a fluid, continuous fashion. In contrast to the five-year-olds' game, we are no longer tempted to claim that the ball is driving the action. The ball is still central, of course, but it has now come under the control of a
disciplined, structured team.
The question is: What has changed between these two soccer games? How does one get from the five-year-olds' skill level to something approaching that of the Argentinian players? What does the learning process look like? How can we understand the skilled version of the game in a way that allows us to do useful things like come up with effective training interventions?

It would be reasonable to expect that research on joint action should give us something to say about such matters. A joint action research programme worth the name ought to provide us with some guidance about how to go about formulating and answering appropriate questions. Does it?

## From the study of the individual to the study of the group: Where did the action go?

Joint action research has its origins in attempts to expand the traditional, individualistic cognitivist research programme to encompass the study of interpersonal phenomena, such as discourse-level activities in spoken language. For the sake of clarity, here is a definition of that individualistic research programme: 'Cognitivism in psychology and philosophy is roughly the position that intelligent behavior can (only) be explained by appeal to internal "cognitive processes," that is, rational thought in a very broad sense' (Haugeland, 1978).

This immediately raises a problem. Any attempt to deal with multi-actor activities within cognitivism runs straight into an apparent contradiction: how can we appeal exclusively to internal processes when the 'inside' is distributed across multiple individuals? The way that researchers have typically dealt with this is through accepting the group actor assumption.

In perhaps the most widely-cited paper on the topic in recent years, Sebanz et al. (2006) frame the problem thus: 'As a working definition, joint action can be regarded as any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment.' Gilbert (1990) discusses the case of two people going for a walk, and insists that 'in order to go for a walk together, each of the parties must express willingness to constitute with the other a plural subject of the goal that they walk along in one another's company.' Bratman (1992) identifies as one of the characteristic features of joint activity that each actor has an 'appropriate commitment' to the joint activity. Marsh, Johnston, Richardson, and Schmidt (2009), who are not themselves cognitivists, go as far as to describe the group as a hybrid organism or 'chimera', an entity 'that has an implausible wholeness, despite the disparateness of the parts that compose it.' Searle (1990) provides a clear-eyed analysis of the problems associated with appealing to intentionality in the group setting, but he also posits a version of the group actor assumption: 'Intuitively, in the collective case the individual intentionality, expressed by "I am doing act A," is derivative from the collective intentionality, "We are doing act A."'

What can be seen in all of these statements is an acceptance of the view that the group must be understood as a real entity-a set of constraints or structures which are imposed on individual action and which must come into existence first before the joint action can be implemented. Debate within the joint action tradition can largely be read as a series of disagreements about what we should expect these constraints and structures to look like. ${ }^{1}$

In practice, this has led to the most intense research effort being directed at the question of interpersonal coordination (Vesper et al., 2017): How is it that a dispersed set of individual actors becomes a group? The ancillary assumption driving such investigations is that we should be able to identify a single, general mechanism which applies in all cases and which allows multiple individuals to coordinate with one another, no matter the activity they are actually engaging in.
In this vein, Tomasello et al. (2005) attribute the set of behaviours they consider to be uniquely human, such as language and culture, to an entity they call 'shared intentionality'. Tomasello notes at the end of his 2014 book that a 'particularly big' open question remains concerning 'the nature of the jointness or collectivity or "we-ness" that characterizes all forms of shared intentionality.' He favours an appeal to recursive mind-reading of the he-thinks-that-she-think-that-he-thinks variety, a solution similarly favoured by other researchers (e.g., Clark, 1996). On this account, a joint action can proceed only when all participants to the action understand themselves to be participating in the action as part of the group. The group exists in the minds of its members.

The most common alternative to recursive mind-reading hypotheses are contagion-based theories which posit that the group comes into existence automatically, as a result of lowlevel synchronization phenomena (e.g., Pickering \& Garrod, 2004; Marsh, Richardson, Baron, \& Schmidt, 2006). On such accounts, joint action is made possible-the group is able to act-because the individuals have already, spontaneously, become organized into a coordinated unit. The group exists precisely in the coordination of its members.
But notice that, in its pursuit of general mechanisms driving interpersonal coordination, the joint action research programme has indefinitely postponed the study of any particular action phenomenon as an instance of what it says it is: action. Nowhere does the group actor assumption lead researchers to ask the kinds of questions suggested above, about how to understand the movements of skilled soccer players, or how to come up with useful training interventions. The research programme appears to be misnamed, because what is actually being studied-interpersonal coordination-is understood only as a prerequisite to the real action. Where did the action go?

[^272]
## Case study 2: Group hunting in wolves and chimpanzees

In the wild, wolves are observed to hunt in packs, chasing large prey such as buffalo. A buffalo is a dangerous animal for a wolf: it has horns and powerful legs, which the wolf must avoid. Rather than attack the buffalo directly, the wolves chase the prey animal until it tires and collapses, whereupon the wolves will surround the prey on all sides, cutting off potential escape routes. Muro, Escobedo, Spector, and Coppinger (2011) were able to reproduce this behaviour in a computer simulation, modeling the wolves as agents following two rules: 1) get as close as possible to the buffalo without putting yourself directly in its path (i.e., avoid being trampled on or skewered), and 2) maximize your distance from neighbouring wolves.

Now, we might be tempted to conclude from this that the wolves are literally following a couple of simple rules. But the rules here are written as they are only because this is a computer simulation and it therefore requires explicit, symbolic rules in order to produce any output. In reality, the wolves cannot be following rules; the action must be driven by perception. What an individual wolf sees is a vista containing a potential prey animal along with a number of other wolves, all of which are already moving. The whole scene is perceived as a set of threats and opportunities which are continually rearranged as the situation unfolds. The question is, what aspects of this whole structure must a wolf learn to attend to and to exploit in order to successfully act?

A somewhat more structured type of group hunting behaviour has been observed in chimpanzees, who hunt monkeys (Boesch \& Boesch, 1989). It is attested that one chimpanzee will instigate the hunt, chasing the monkey from some direction, whereupon further chimpanzees will join in from the sides, driving the monkey into a space where eventually an ambushing chimpanzee will be in a position to catch and kill the monkey.

Boesch (2002) posits that the chimpanzees are able to coordinate this behaviour because they each adopt a specific 'role'. He identifies four roles: chimpanzees can act as driver, chaser, blocker, or ambusher. The suggestion is that these roles have some normative significance: a chimpanzee's role determines the share of the meat that they are entitled to. If this is the case, then some form of mind-reading must be involved, as each chimp would need to know what role the other chimps are playing.

Tomasello (2014) disputes this description, arguing that, in fact, each individual chimp's behaviour can be explained by selfish motives: 'chimpanzees in a group hunt are engaged in a kind of co-action in which each individual is pursuing his own individual goal of capturing the monkey'.

For Boesch, then, the group is reified as a set of normative role assignations. For Tomasello, the group simply does not exist because for it to exist the chimpanzees would have to have a human-like understanding of themselves and their fellows-the very idea is anathema.

But seeking to explain these hunting behaviours in terms of the reified group (or the absence of a group) draws our attention away from the actual action. It draws our attention away from the physical structures, movements, events, and reconfigurations that must necessarily be invoked in a description of the hunting behaviours: the chase, the closing in from the sides that narrows the monkey's escape possibilities, the space where the ambush can occur. In other words, the group actor assumption causes us to overlook the world itself. So let us reformulate our earlier critique:

## 'The group actor assumptions reifies the group while neglecting the world'

Perhaps 'the world' is too imprecise a term here. Below, it will be replaced with the concept of the ecosystem.

## The late appearance of the individual in evolution and development: from public action to private action

So far, we have left untouched the special position of individual action, generally understood as action in its most basic form. This assumption, too, must be challenged.

When we take a historical perspective on cognition, we note that the individual, as self-aware, symbol-using actor, is in fact a rather late achievement, in evolutionary terms as well as in terms of development within the human lifespan. Indeed, the notion that the individual should be treated as a given is only a self-evident truth when the matter is considered from within the cognitivist research programme.

One approach to psychology which rejects this progression from individual to joint action is that developed by Vygotsky and his followers. According to Vygotsky (1978), the child develops into a competent language-using being not by silently contemplating the actions and utterances of others, but by actually engaging in action: 'Prior to mastering his own behavior, the child begins to master his surroundings with the help of speech. This produces new relations with the environment in addition to the new organization of behavior itself.' Through the repetition of actions that were initially directed outwards, the child eventually starts to direct his actions inwards, at himself. Lake (2012) gives the example of a child speaking to himself as he looks at paint: 'I need the green paint.' Here, the speech is performed 'out loud' but is directed at regulating the child's own behaviour, not directed outwards to an audience.

Through continued practice, the child will eventually learn to regulate his behaviour without needing to say the words out loud. His private, self-directed speech becomes inner speech. This process is what Vygotsky referred to as 'internalization'. It should be noted that what has been internalized here is not some arbitrary symbol, or some Platonic notion of greenness, but the self-regulatory action. What was an action when it was carried out publicly remains an action when it is directed inwards and carried out privately: 'internalization reflects not "content" poured into a person's psychological structure, it is
how that structure is formed' (Lake, 2012). In other words, it is only through the mastery of externally-oriented actions that the individual arises.

But here we are talking about the development of speech. What does this have to do with joint action?

The difficulty is that as adult humans we are inescabably also language-users. Think back to the two soccer teams described above. While the five-year-olds have a limited vocabulary for explaining to themselves what they are doing, this is not the case for skilled adult players. Soccer players are barked at repeatedly on the training pitch year after year: "keep the ball moving", "give your teammates options", "don't get caught ball-watching", etc. The public behaviour of skilled soccer players is modified and channeled through private self-regulating actions and instructions.

The point here is that private action need not be taken as the starting point for the activity of groups. If action in its basic form is a public activity, directed outwards at the world and having immediate effects in changing the structure of that world itself, then perhaps the central contradiction driving joint action research-the contradiction which drives researchers to accept the group actor assumption-need never arise. The problem of explaining group action in terms of internal processes can be dissolved by recognizing that those internal processes do not arise until late in evolution and development. Private, self-directed behaviour is the consequence of public action, not a prerequisite for it.

## An ecosystems view of action: Reinstating space and movement

At the level of biology, an ecosystem is thought of as a rather slow-moving thing: a stable configuration of species, niches, soil types, atmospheric conditions, and so on that exist in some sort of equilibrium (Sarkar, 2016). At the level of psychology, the ecosystem must be thought of as a dynamic system, one in which every public action made by an animal alters, or perturbs, the configuration of the ecosystem itself. Moreover, because the ecosystem contains multiple actors, all public actions are inherently social, in the sense that the action changes the whole system in a manner which may be relevant to other animals: it may create new opportunities, erase previously existing ones, reveal threats, etc.

Consider the wolves again. After the pursuit of the buffalo the wolves surround their prey. Now, at this point any movement that a given wolf makes-an anti-clockwise rotation around the buffalo, say-alters the layout of opportunities for the neighbouring wolves, just as it changes the shape of the potential escape gaps for the buffalo. The second wolf may respond to the first movement in any number of ways. In no sense, though, would this response require an 'understanding' of what the first wolf was doing. It is enough to point out that the spatial configuration has changed. If we wish to understand the wolf's behaviour we should investigate what is going on in the relationship between the animal and its surroundings-that is, at the level of public action.

Does this mean that the wolf has direct perceptual access to the ecosystem? It does not. An ecosystem is not the kind of thing that can be perceived. Here it is necessary to make a distinction between the ecosystem and the environment.

James Gibson (1979) developed a psychology of perception that has as a central claim that perception cannot be thought of as something that an animal does on it own, merely in response to an environment; perception must instead be thought of as a process enacted within an animal-environment system. The animal's environment is already structured. If it is to survive, the animal must learn to make use of this structure in adaptive ways. Moving forwards creates an optic flow pattern which can be used to direct locomotion. This pattern only exists when the action is implemented and useable structure is present in the environment-the pattern does not exist for an animal that has never moved, nor is it present for an animal whose environment is filled with thick fog. For the pattern to exist requires both the animal and the environment to be in a certain relation, hence the animal-environment system.

An under-appreciated implication of this view is that it requires us to adopt a view of the environment as animalspecific. That is, it requires us to conclude that there are precisely as many environments in the world as there are animals (for a related argument about the concept of information see van Dijk, Withagen, \& Bongers, 2015). At the very start of his 1979 book, Gibson says the following about the animal and its environment:
[I]t is often neglected that the words animal and environment make an inseparable pair. Each term implies the other. No animal could exist without an environment surrounding it. Equally, although not so obvious, an environment implies an animal (or at least an organism) to be surrounded. This means that the surface of the earth, millions of years ago before life developed on it, was not an environment, properly speaking. The earth was a physical reality, a part of the universe, and the subject matter of geology. [...] We might agree to call it a world, but it was not an environment.

So what can we say about what the wolf sees? What it sees is a perspective on the ecosystem. But this is only to say that what is seen is a partial view of the structure that exists in the world. If an environment is the complement of an animal, then an environment it is what is experienced from a first-person perspective. An ecosystem, by contrast, is an analyst's label. It is a tool for capturing some of the structure in the world in a manner that can guide our investigation of certain phenomena of interest.

In the case of joint action research, the ecosystems concept is a tool that allows us to understand how multiple actors are able to negotiate a space that is populated with other actors. It allows us to move beyond the interpersonal coordination paradigm by making clear that such coordination is not necessarily something the animals must do. A basic level of


Figure 1: A pedestrian, represented by the black circle, waits to cross the road. In (a) the crossing will take 15 s , and the pedestrian must be attentive to potential threats 200 m down the road. In (b) the task difficulty is drastically diminished: with the crossing reduced to 5 s , the pedestrian now only needs to be aware of potential vehicles up to 67 m away.
coordination is already present by virtue of the fact that the animals already exist within a single ecosystem, whose structure is instantiated in just so many first-person perspectives.

The real question, though is whether the ecosystems view of action actually has any value in terms of its ability to generate practical research. I believe that it does. One example is in its potential application to the design of spaces for multiple actors, such as urban streetscapes.

## Case study 3: Pedestrian road crossing

Navigating an urban environment entails continuously coordinating with other actors. An urban design approach that has been popular in Europe in recent years, the shared space approach, argues that interpersonal coordination can be not only necessary but sufficient for managing urban traffic flows. This reasoning has led the UK government to publish design guidelines encouraging the removal of formal infrastructure elements such as kerbs and traffic signs, with the aim of forcing drivers to be more attentive to their immediate surroundings (Department for Transport, 2011). This approach has been criticized because it often renders these spaces less accessible for elderly and visually impaired pedestrians (Moody \& Melia, 2014). To see why this is the case, it is useful to adopt an ecosystems perspective on the activities of the road users. Here we will consider a simple instance in which a pedestrian wishes to cross a road that carries vehicle traffic. We will focus only on the pedestrian's task. ${ }^{2}$

[^273]The first thing to notice is that while this is an interpersonal coordination phenomenon, it is an asymmetrical one. Should a collision occur here, it is likely to be more disastrous for the pedestrian than for the driver of the car. The pedestrian thus has an immediate interest in being especially cautious.

The pedestrian's task here involves prospective control: it requires the pedestrian to be attentive to ongoing, or unfolding, movements in her environment (Von Hofsten, 1993). This is illustrated in Fig. 1. In order to decide whether to begin crossing, the pedestrian needs to take into account the movements of vehicles in the road that are on course to intercept her path. This decision task is made easier or harder depending on a number of factors, which include: how wide the road is, how fast the pedestrian is able to cross it, and the speed of the traffic. If the pedestrian can only move slowly (or has far to cross), as in (a), then that pedestrian must be aware of vehicles that are much further away compared to a pedestrian who is able to cross quickly (b). The difficulty of the decision-making task (how far the pedestrian has to look down the road) is proportional to the pedestrian's time-to-cross. Prospective control refers to the organization of behaviour with reference to perceived future movement, i.e. with reference to where objects are going to be, as the action unfolds. Older, slower-moving pedestrians have particular difficulty in crossing roads (Langlois et al., 1997), and Fig. 1 makes clear why this is the case: the slower one is able to move, the further one has to be able to see in order to control one's movements. In practice, for very wide roads or for very slow pedestrians it becomes impossible to perceive safe crossing opportunities.

From the fact that the crossing time is proportional to the distance one has to be able to see down the street, we can immediately derive a general design principle: for the road to be crossable by a given pedestrian, the crossing width, in seconds, can be no longer than the time it will take for a car to appear from beyond the horizon of visible space to intersect the pedestrian's path. In practical terms, this means that in order to make a road accessible to the widest possible range of pedestrians the crossing width should be kept to a minimum, and visibility maximized. This gives theoretical grounding to recent traffic design manuals that recommend just these measures (e.g., World Health Organization, 2013).

While the activity of crossing a road can certainly be described as an interpersonal coordination phenomenon, this is not all that it is. The ecosystems perspective allows us to reach a deeper understanding of the space in terms of: 1) the first-person perspective of the road users (pedestrians and drivers); 2) the movement capabilities, and vulnerabilities, of these actors; 3) the time-extended quality of the action; 4) the actual layout of the street. Some promising potential areas

[^274]of application are in designing roads to be inclusive for people of diverse capabilities and in redesigning 'accident black spots', which are spaces that are inherently unsafe.

## Implications for joint action research

The group actor assumption has restricted research on multiactor activities to the study of an unnecessarily narrow range of interpersonal coordination phenomena. Rejecting it is a liberating move: it frees researchers from having to address unsolvable questions about shared intentionality. The adoption of an ecosystems perspective, meanwhile, enables an investigation of multi-actor activities for what they are: instances of action. It gives us tools for addressing real-world problems in practical ways. There are some pitfalls to be avoided. I have barely touched on issues of language-use here, but human actions are inherently language-involving. For now, let us restate that internal language use is a kind of private action, and it is derived from public action. This public-private distinction is more useful, and less misleading, than the traditional individual-joint distinction. And let us restate also the ecosystem-environment distinction. The latter is a first-person perspective on the former, which itself is the setting of all social activity.

## Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 706432

## References

Boesch, C. (2002). Cooperative hunting roles among Taï chimpanzees. Human Nature, 13(1), 27-46.
Boesch, C., \& Boesch, H. (1989). Hunting behavior of wild chimpanzees in the Taï National Park. American Journal of Physical Anthropology, 78(4), 547-573.
Bratman, M. E. (1992). Shared cooperative activity. The Philosophical Review, 101(2), 327-341.
Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.
Department for Transport. (2011). Local transport note 1/11: Shared space. Norwich, UK: The Stationary Office.
Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton-Mifflin.
Gilbert, M. (1990). Walking together: A paradigmatic social phenomenon. Midwest Studies in Philosophy, 15(1), 1-14.
Haugeland, J. (1978). The nature and plausibility of cognitivism. Behavioral and Brain Sciences, 1(2), 215-226.
Hutchins, E. (1995a). Cognition in the wild. Cambridge, Massachusetts: MIT Press.
Hutchins, E. (1995b). How a cockpit remembers its speeds. Cognitive science, 19(3), 265-288.
Lake, R. (2012). Vygotsky on education. New York: Peter Lang.
Langlois, J. A., Keyl, P. M., Guralnik, J. M., Foley, D. J., Marottoli, R. A., \& Wallace, R. B. (1997). Characteristics
of older pedestrians who have difficulty crossing the street. American Journal of Public Health, 87(3), 393-397.
Marsh, K. L., Johnston, L., Richardson, M. J., \& Schmidt, R. (2009). Toward a radically embodied, embedded social psychology. European Journal of Social Psychology, 39(7), 1217-1225.
Marsh, K. L., Richardson, M. J., Baron, R. M., \& Schmidt, R. (2006). Contrasting approaches to perceiving and acting with others. Ecological Psychology, 18(1), 1-38.
Moody, S., \& Melia, S. (2014). Shared space: Research, policy and problems. Proceedings of the Institution of Civil Engineers - Transport, 167(6), 384-392.
Muro, C., Escobedo, R., Spector, L., \& Coppinger, R. (2011). Wolf-pack (Canis lupus) hunting strategies emerge from simple rules in computational simulations. Behavioural Processes, 88(3), 192-197.
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-190.
Quinn, R., \& Carr, D. (2006). Developmentally appropriate soccer activities for elementary school children. Journal of Physical Education, Recreation \& Dance, 77(5), 13-17.
Sarkar, S. (2016). Ecology. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2016 ed.). Metaphysics Research Lab, Stanford University.
Searle, J. R. (1990). Collective intentions and actions. In P. R. Cohen, J. L. Morgan, \& M. E. Pollack (Eds.), Intentions in communication (pp. 401-415). Cambridge, MA: The MIT Press.
Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: bodies and minds moving together. Trends in Cognitive Sciences, 10(2), 70-76.
Tomasello, M. (2014). A natural history of human thinking. Cambridge, Massachusetts: Harvard University Press.
Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., et al. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28(5), 675-690.
van Dijk, L., Withagen, R., \& Bongers, R. M. (2015). Information without content: A gibsonian reply to enactivists' worries. Cognition, 134, 210-214.
Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., ... Wahn, B. (2017). Joint action: Mental representations, shared information and general mechanisms for coordinating with others. Frontiers in Psychology, 7, 2039.
Von Hofsten, C. (1993). Prospective control: A basic aspect of action development. Human Development, 36(5), 253270.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, Massachusetts: Harvard University Press.
World Health Organization. (2013). Pedestrian safety: a road safety manual for decision-makers and practitioners. Geneva: WHO Press.

# Towards Automated Classification of Emotional Facial Expressions 

Lewis J. Baker (lewis.j.baker@rutgers.edu) ${ }^{1}$, Vanessa LoBue (vlobue@rutgers.edu) ${ }^{2}$, Elizabeth Bonawitz (elizabeth.bonawitz@ rutgers.edu) ${ }^{2}$, \& Patrick Shafto (patrick.shafto@gmail.com) ${ }^{1}$<br>${ }^{1}$ Department of Mathematics and Computer Science, ${ }^{2}$ Department of Psychology Rutgers University - Newark, 101 Warren St., Newark, NJ, 07102 USA


#### Abstract

Emotional state influences nearly every aspect of human cognition. However, coding emotional state is a costly process that relies on proprietary software or the subjective judgments of trained raters, highlighting the need for a reliable, automatic method of recognizing and labeling emotional expression. We demonstrate that machine learning methods can approach nearhuman levels for categorization of facial expression in naturalistic experiments. Our results show relative success of models on highly controlled stimuli and relative failure on less controlled images, emphasizing the need for real-world data for application to real-world experiments. We then test the potential of combining multiple freely available datasets to broadly categorize faces that vary across age, race, gender and photographic quality.


Keywords: Classification, machine learning, computer vision, support vector machines, emotion and cognition, facial recognition

## Introduction

Emotions are widely assumed to play a causal role in nearly every aspect of cognition (e.g., Pessoa, 2008), and yet many studies in cognitive science (and developmental science in particular) neglect to measure emotion because current measures are either expensive, tedious, or inaccurate. Consequently, many standard practices in the field have turned to indirect measures of affect. One prevalent example is the association of infant looking time with vastly different emotions depending on the researchers' theoretical stance, including preference (e.g., to positive emotional expressions; LaBarbera, Izard, Vietze, and Parisi, 1976), interest (e.g., to animate stimuli; Csibra, 2008), or surprise (e.g., to violations of belief; Baillargeon, Scott, and He, 2010). Another example is the notable lack of emotional state measures in studies on attention, learning and memory, even though the field has acknowledged the impact of emotion on these functions for over 50 years (Easterbrook, 1959). Rather than inferring the role of emotions, future studies could measure it efficiently using facial recognition algorithms. The advent of elegant machinelearning algorithms offers a free, reliable, non-invasive and easily implemented method that may be able to measure affective state in real-world settings at levels that meet or exceed trained human raters.

Here, we demonstrate automatic classification of emotional faces using three different datasets. We concentrate on young populations, as developmental science is particularly interested and constrained by hand-coding, but also demonstrate that methods are easily extended to adult populations. We also concentrate on relatively simple machine learning algorithms that may be flexibly implemented for a variety of
psychological studies. In doing so, we highlight the need for real-world data to solve real-world problems, as models based on well-curated training images that are common in the field often fail to accurately categorize messy, uncontrolled images. We further show how a single, large dataset that leverages controlled and uncontrolled images can improve generalization to real-world stimuli.

Recognition of facial expressions is a useful, non-invasive method of reasoning about another's thoughts. The seeming universality of emotional expressions further underscores their importance (Ekman \& Friesen, 1971). However, to understand how emotion influences cognition, researchers must be able to categorize facial expressions in continuous time and no existing measure can do this without great expense of time or money. Participant surveys lack temporal resolution and fall prey to metacognitive errors. Physiological methods require expensive or invasive apparatus such as galvanic skin response monitors or cortisol measurements Picard, Vyzas, and Healey, 2001. Even the gold-standard method of the Facial Action Coding System (FACS; Ekman and Rosenberg, 1997) requires hours of effort by trained technicians or prohibitively expensive proprietary software. As the demand for ecological experimentation increases, so too does the volume of video data for researchers (or more often their students) to scrutinize and label, frame by frame. The relatively constrained problems of identifying, labeling and categorizing facial features over thousands of datapoints is a prime opportunity for a machine learning solution.

Advances in data science and machine learning offer an affordable and accurate measure of participant emotional state using only filmed recordings. Computer scientists have demonstrated the uncanny accuracy of basic algorithms to classify highly controlled emotional images (e.g., Cohn, Zlochower, Lien, and Kanade, 1999; Littlewort, Bartlett, Fasel, Susskind, and Movellan, 2006), and recent efforts to categorize emotion "in the wild" (Yao, Shao, Ma, \& Chen, 2015) have the problem to unsupervised learning for less controlled images. Here we are interested in applying machine-learning to the varied contexts typical of cognitive science experiments. Laboratory settings offer more control than streaming surveillance footage, but less control than posed photography. We approximate this by comparing human performance to an algorithm trained on three datasets with unique attributes, each of which could reasonably be applied to experimental settings. We demonstrate the need for large amounts of highly varied data to consistently and accurately categorize human facial expressions. Furthermore, we present a model
trained on images that vary by age, ethnicity, gender and photographic conditions that nonetheless approaches human rater performance. It is worth noting that our goal is not to model human performance or develop new machine-learning methods; rather, we wish to explore the kinds of data required to approximate human-level emotion coding for cognitive experiments.

We begin by introducing several datasets with unique attributes of in interest to different applications. We define the methods required to use open-source libraries to create a simple machine-learning classifier. Applying this model to the datasets reveals that highly controlled training stimuli are more easily categorized, and that noisier, real-world stimuli are unsurprisingly more difficult. We discuss the tradeoff between accuracy and generality by amalgamating three datasets into a comprehensive model that is more robust to noisy input.

## General Methods

## Databases

Machine learning requires a large number of samples for reliable classification. However, the type of input can greatly affect the generalizability of the model. For instance, a model trained on only children's faces might not perform well with adult faces. Likewise, a model trained on highly controlled images might not perform well on naturalistic stimuli. We drew from three sources for training, each with a particular strength that could improve performance in a given setting. A "face" was defined as a front-facing image containing two eyes and no obstructions to facial features.

The CAFE dataset. Most face databases for psychology and machine learning focus on adults. However, a recent effort by LoBue and Thrasher (2015) documented the facial expressions of young children for applications in developmental psychology. Although stimulus sets exist for older children (aged 8-17, Egger et al., 2011) and adults (Cohn et al., 1999), the Child Affective Facial Expression (CAFE) set is the only collection featuring young children. The set contains photographs of 154 racially and ethnically diverse 2 - to 8 -yearold children posing for six emotional facial expressions (angry, disgusted, fearful, happy, sad, and surprised) as well as a resting neutral expression. Facial expressions were further labeled for "open" or "closed" mouths for angry, fearful, happy, sad and neutral faces. Disgust expressions were uniquely coded as with or without a protruding tongue. The CAFE set features multiple emotional faces for each child, though not every child demonstrated every subcategory of emotion. Altogether, the set contains 1192 images. Children's facial features offer a great deal of variability, and the ethnic diversity of the participant sample approximates the demographics of the United States.

The CAFE set was validated by a group of 100 independent adult raters, who viewed each image and labeled it with one of the seven emotions. Importantly, the images are labeled by the expression the child was asked to give, and not by
the labels most often generated by the raters. This variability makes the CAFE useful to compare to computer models, as we can test the model's success on "difficult" or "easy" faces compared to human performance.

The CK+ dataset. One method of producing cleaner data for machine learning is to extract images with tightly controlled visual features. Although our focus is on developmental populations, the CAFE set is the only publicly available database of children's faces. We therefore included a dataset comprised only of extensively vetted faces: the CohnKanade AU-Coded Expression Database, Version 2 (Lucey et al., 2010). The Cohn-Kande dataset (CK+) consists of over 11,000 image sequences of 120 adult models as they changed from neutral resting faces to peak emotional expression from 7 categories (the same as the CAFE expressions, with the addition of contempt). It is currently unknown whether training images from an adult dataset would improve performance on child facial categorization. Given the abundance of adult datasets, any improvement on child facial classification would expand the available training data for future models.

Machine vision researchers often use the CK+ dataset as a benchmark for performance of an algorithm (e.g., Littlewort et al., 2006). For our purposes, training the algorithm on the CK+ dataset allows us to test the best case scenario of facial classification, as it contains only highly controlled images with little cross-category variability. This comes at a cost to ecological validity, as all faces are of adults aged 18 to 30 , and less than $18 \%$ were minorities. Additionally, whereas items in the CAFE set were validated using subjective judgments from adult raters, the peak faces from the CK+ database were validated using the Facial Action Coding System (FACS). Briefly, FACS categorizes faces into emotional categories using reliable expressions of specific facial motor groups, or action units (Ekman \& Rosenberg, 1997). Lucey et al. (2010) validated the emotional labels given to each peak face using a linear support vector classifier trained on action units. Selecting these initial and peak faces generated 308 emotional expressions from the 6 emotional categories with a corresponding neutral face for each. A single neutral face was randomly selected for each participant to prevent over-fitting of neutral faces, leaving 120 neutral faces and a total of 428 faces.

Google image search by category. The CAFE and CK+ sets feature images taken under ideal lighting and camera positions, with labels that have been rigorously validated. However, real-world use of a facial expression classifier would necessarily include less-than-ideal photographic circumstances. To approximate the noisiness of real-world stimuli, we extracted images from a Google image search with the search term "X child face", where X was an emotional category of interest. Images were selected by research assistants, with the criteria that each image featured an individual human child's face (approximately aged 3-10) without obstruction on the face area.

Research assistants terminated the collection of images if the total number of collected images exceeded 100 exemplars or if the search returned more than 20 images in a row without a viable exemplar. This produced only 2 neutral exemplars, so an additional search was conducted using "calm" and "serious" as additional terms for neutral. This produced a total of 609 faces from all seven categories

## Face Extraction

Images from all datasets contained extraneous information, including body parts (e.g., hair and shoulders) or photographic artifacts (e.g., serial numbers in the CAFE and CK+ datasets, miscellaneous objects in the Google dataset). All images were passed through a facial recognition algorithm ${ }^{1}$ and reduced to a $300 \times 300$ pixel rectangle centered on the identified face.

Facial recognition was conducted using Haar Featurebased Cascade Classifiers. Generally, the cascade classifier breaks an image into clusters of pixels and excludes clusters that do not resemble facial features from later analysis. The process is then repeated until only clusters that resemble facial features remain. For methodological details and validation, see Viola and Jones (2001). The end result is a computationally efficient method for identifying facial regions.

A trained cascade classifier was obtained from the OpenCV website (Itseez, 2016). All faces from all datasets were passed through the classifier and cropped. A member of the research team then examined each extracted face and discarded false positives on non-face objects. This method produced 1187 faces ( 5 removed) from the CAFE set, 427 faces ( 1 removed) from the CK+ dataset, and 477 faces (132 removed) from the Google dataset.

## Human Validation

Image category labels for the CAFE and CK+ datasets were validated using adult human raters. To ensure that all images were of equal quality when training the classifier, we validated the Google dataset using 87 adult human raters recruited via Amazon Mechanical Turk. This was necessary to compare classifier and human performance for images that more closely resemble the real world.

Raters (median age: 31; 61 females, 41 college graduates, 28 parents) labeled a representative subset of the Google faces (between 47 and 49 images, evenly distributed across categories) into one of seven emotional categories. Raters also labeled a subset of the CAFE dataset (42 images, 6 from each category). The CAFE faces were evenly distributed by difficulty according to the CAFE set's previous validation metrics. This was done to compare the performance of in-person (live) and online raters.

[^275]

Figure 1: Validation of the CAFE and Google datasets by "online" human raters via Mechanical Turk vs in-person, "live" raters.

We first confirmed that Mechanical Turk raters performed comparably to live human raters (Figure 1). Overall accuracy of ratings for CAFE set images between live and online raters was significantly positively correlated ( $r=.783, t_{412}=$ $25.53, p<.0001$ ). Furthermore, accuracy of online and live raters were further correlated for all emotional categories (lowest $r_{\text {surprise }}=.410, t_{55}=3.340, p=.001$; highest $r_{\text {sad }}=$ $.820, t_{57}=10.807, p<.0001$ ), with the exception of happy expressions ( $r=.22, t_{57}=1.742, p=.08$ ), which likely had a reduced correlation due to ceiling effects. These results support the use of Mechanical Turk raters to validate the Google dataset.

We then confirmed that the human categorization performance for the novel Google images were comparable to the CAFE images. A two-way ANOVA modeling mean online rater categorization performance by dataset and emotional label found significant differences in categorization accuracy by dataset $\left(F_{1,877}=8.753, p=.003, \eta^{2}=.086\right)$ and emotion $\left(F_{6,877}=89.504, p<.0001, \eta^{2}=.881\right)$. There was also a significant interaction $\left(F_{6,877}=2.366, p=.028, \eta^{2}=.023\right)$, indicating no significant difference between online and live raters for angry, happy, neutral, sad and surprised expressions (highest $t_{877}[\mathrm{sad}]=.9694, p=.167$ ), and significantly poorer performance by online raters for disgusted and fearful expressions (lowest $t_{877}$ [disgust] $=2.134, p<.017, d=$ .144). These results suggest that the Google dataset is comparable to the CAFE set for five of seven emotions and follows the same trends for sadness and disgust, making the Google set ideal for testing an algorithm on ecological images.

## The Machine Learning Algorithm

Whole research communities are dedicated to the application of machine learning to emotional recognition, using both supervised and unsupervised algorithms for still image, video, audio or multimedia data (e.g., the EmotiW challenge at the annual ACM ICMI conference). Although there have been re-
cent successes modeling dynamic features (Littlewort et al., 2006), We opted to analyze still images to simplify implementation for the target demographic of psychologists, and used a supervised learning approach due to the relatively few training images available. We therefore selected a Support Vector Machine (SVM) algorithm, as SVMs are ideal for use by non-computer scientists for their simple implementation and ease of interpretation. SVMs have a long and successful history in image recognition (Tong \& Chang, 2001), particularly with facial recognition (Osuna, Freund, \& Girosit, 1997). An SVM is a type of supervised learning in which the algorithm identifies the optimal boundary between labeled data points. The boundary is defined by the support vectors, the subset of the data that define the boundary between classes. This boundary can then be used to infer, based on observed features, which category a novel image (or novel images) best fit. We recommend Cristianini and Shawe-Taylor (2000) for an in-depth overview.

We trained an SVM for each dataset, as well as on a comprehensive dataset containing training images from all three datasets. We used an open-source SVM classifier available through the scikit-learn database (Pedregosa et al., 2011). SVMs require the user to choose a similarity function, called a kernel, that governs the complexity of the possible boundaries between classes. There are many standard options for kernels including linear, polynomial, and radial basis function (RBF; aka Gaussian). Each computes similarity somewhat differently and they consequently differ in the kinds of classification boundaries they admit; as one might expect, a linear kernel gives a linear boundary and polynomial and RBF kernels allow non-linear boundaries. While these non-linear methods offer increased expressiveness, they also increase the risk of overfitting.

Additionally, there are two parameters that must be set and affect outcomes: the regularization parameter $C$ and kernel coefficient $\gamma . C$ is a regularization parameter which, when set to higher values, allows more complex solutions. The kernel coefficients $\gamma$ affect the influence of specific specific supports. When $\gamma$ is small, a support has broad influence on classification decisions, whereas when $\gamma$ is large the influence of each support is localized to the area near the supporting data point. A grid search for kernels, \{linear, polynomial, radial basis function (RBF) \}, penalty parameters, $C=(.001, .01, .1,1,10,100)$, and kernel coefficients, $\gamma=(.0001, .001, .01, .1,1,10,100,1000)$, yielded the optimal combination of a polynomial kernel with a $C=1$ and a $\gamma=.0001$, as assessed via cross-validation.

Although SVM classifiers are often used for facial recognition, training a classifier for emotional features offers unique problems. The classifier might divide faces by other similarities; emotional expressions are but a subset of the considerable variability between faces. For example, say a classifier is trained on two stimuli: Child A with an angry expression and Child B with a happy expression. When presented with a test image of Child A making a happy expression, the SVM


Figure 2: Classification of test images improved as a function of training data. The top line denotes average human accuracy across live and online human validation; the lower line denotes chance.
may be more likely to categorize the test image by the stable facial similarities of Child A than to the desired similarities in emotional expression of Child B. One solution might be to randomly select only one face per child participant in the CAFE and CK+ sets. This is not ideal, as it would greatly reduce the training set. Instead, we trained the SVM on all faces for a proportion of participants and tested on all faces for the withheld subset of participants. This eliminated the possibility that a test image might be paired with a training image of the same child, while also maximizing the richness of the dataset.

Another issue unique to emotional classification is the breadth of expression. For instance, the CAFE set makes a distinction between faces with open and closed mouths, and both the CAFE and Google sets contain exemplars that were difficult to label by human raters. We opted to include all instances under the basic emotional category, regardless of subordinate labels or validation score, so as to maximize training data with the greatest possible variation between features.

## Results

The algorithm was trained on incrementally increasing sizes of training data from all three datasets individually and a comprehensive dataset trained from all sources. Each sample size by dataset was repeated 40 times with a new random selection of training and test data to approximate error.

The primary goal of these analyses was to demonstrate machine-learning categorization on different training data versus human raters. Figure 2 illustrates overall performance by sample size for each of the datasets. An ANCOVA modeling dataset by training sample size revealed a significant effect of dataset $\left(F_{3,1512}=1412.39, p<.0001, \eta^{2}=\right.$ $.540)$ and training sample size $\left(F_{1,1512}=1103.76, p<\right.$ $\left..0001, \eta^{2}=.422\right)$, with a significant interaction $\left(F_{3,1512}=\right.$ 96.91, $p=.0001, \eta^{2}=.037$ ). Paired comparisons revealed that performance on within-dataset models increased faster than the comprehensive dataset as a function of training size (CAFE: $t_{1512}=12.246, p<.0001, d=.629$; Google:


Figure 3: Classification by source of testing data. Accuracy on uncurated Google images improved with the comprehensive model.
$t_{1512}=3.705, p<.0001, d=.191 ; \mathrm{CK}+: t_{1512}=11.886, p<$ $.0001, d=.611)$. Altogether, performance on all datasets improved as a function of training data, but performance on within-dataset models increased faster than the comprehensive model.

The bold dotted line on Figure 2 denotes average human categorization performance for the CAFE and Google sets, although it should be noted that no item-wise validation metrics were available for the $\mathrm{CK}+$ set. T-tests revealed that maximum training sizes on the CK+ dataset exceeded human performance ( $t_{40}=6.00, p<.0001, d=1.90$ ). Classifier performance was significantly below human performance for the CAFE ( $t_{40}=-3.62, p=.0006, d=1.145$ ), Google ( $t_{40}=-38.65, p<.0001, d=9.92$ ), and comprehensive datasets ( $t_{40}=-39.51, p<.0001, d=10.68$ ). Interestingly, human performance was not significantly correlated with classifier performance at maximum training sample size for any dataset (CK+: $r=-.031, p=.837$; CAFE: $r=.260, p=.105$; Google: $r=-.030, p=.869$; Comprehensive: $r=.240, p=.135$ ), suggesting that the basis on which categorization decisions were made by the algorithm differed from human judgments.

It is crucial for future applications that a classifier not only categorizes within a training dataset, but can also generalize beyond that set. A common method of gauging generalizability is to train models for each dataset and test on the other datasets. However, all of the present datasets, particularly the CK+ dataset, have unequal numbers of exemplars for each emotional category. As classifier performance is directly related to the amount of training data, we would have to hold training data constant to the minimum possible value across all emotional categories and datasets, which in this case would be only 25 exemplars per category (the number of exemplars for "fear" in the CK+ dataset), for a training set of only 175 images. Instead, we tested how well a single comprehensive model performs against maximally trained mod-


Figure 4: The comprehensive model from all three datasets paralleled human performance.
els for each individual dataset (the within-set models). This comparison demonstrates how the addition of training images outside the dataset improves performance. An ANOVA comparing accuracy by model type (within-dataset or comprehensive) and source of test images revealed a no effect of model $\left(F_{1,234}=0.001, p=.973, \eta^{2}<.001\right)$ but a significant effect of test image source $\left(F_{2,234}=1175.71, p<.0001, \eta^{2}=\right.$ .961) as well as a significant interaction ( $F_{3,234}=47.13, p<$ $.0001, \eta^{2}=.039$ ). Accuracy for the comprehensive model was significantly greater than the within-set model for Google ( $t_{234}=6.03, p<.0001, d=.792$ ), not significantly different for CK+ test images $\left(t_{234}=1.09, p=.140, d=.142\right)$ and significantly less for CAFE test images ( $t_{234}=6.50, p<$ $.0001, d=.849$ ). Comparing Figure 2 and Figure 3, the overall performance deficit of the comprehensive model relative to the CK+ and CAFE sets in Figure 2 are due to the high proportion of training images in the comprehensive model that come from the CAFE set (57.9\%). Importantly, these results show that a comprehensive dataset from multiple curated sources improves classification of more realistic and uncontrolled Google set.

Finally, it is worthwhile to see how a comprehensive dataset compares to human raters. Figure 4 compares human and comprehensive model performance by emotional category. An ANOVA modeling emotion by rating type (human vs the comprehensive algorithm) revealed than human raters were significantly more accurate than the classifier ( $F_{1,546}=1450.24, p<.0001, \eta^{2}=.567$ ). There were significant differences by emotion ( $F_{6,546}=1021.81, p<$ $\left..0001, \eta^{2}=.400\right)$ as well as a significant interaction $\left(F_{6,546}=\right.$ $\left.84.51, p<.0001, \eta^{2}=.048\right)$.

Overall, these results suggests that the comprehensive dataset follows similar trends as human raters. Sampling from multiple datasets improves performance on the highly uncontrolled Google stimuli and the highly controlled CK+ stimuli; although performance drops for the CAFE stimuli, it
likely stems from the high proportion of images in the comprehensive dataset that are sampled from the CAFE. These results suggest that models comprised of more training data may approach human performance on varied images.

## Discussion

Psychological theories emphasize the causal role of emotions across a variety of phenomena including learning, memory, and attention. However, emotion is rarely measured in such studies due to the cost, inefficiency and tediousness of modern methods. Widely available and accessible methods for coding emotion would greatly reduce barriers to advancing theory by allowing dense measurement of emotion in continuous time. Using a standard machine learning method, we explored the types of training data one would need to approach human-level coding of the big 7 emotional categories. These included curated image sets developed by psychological researchers and uncontrolled images drawn from Google with crowdsourced labels. We find that comprehensive models generated from multiple datasets improve classification of uncurated images. Overall model performance follows the same trends as human performance, and the inclusion of additional datasets promises to further approach human accuracy.

Cognitive science, and developmental science in particular, are greatly limited by the methods of the day. A typical developmental experiment takes place with one child and one experimenter for fifteen minutes. Such tight controls have led to important insights at the cost of ecological validity. The past 20 years have seen incredible improvements to computational theory and processing power that permit a more flexible study of human behavior. With machine-learning methods, scientists are no longer bound to brief interventions or constrained to discrete conditions. Rather, we can now continuously monitor affect and behavior as a response to the real world. Instead of inferring surprise from an infant's looking times, these models provide a method to measure a reliable indicator of emotion. Instead of assuming a role of affect in student outcomes, we can incorporate emotional expression with an intervention in real time.

This paper represents an effort toward integrating computational methods with cognitive science with the goal of actively measuring all features that support cognition. For now, we have demonstrated the feasibility of using publicly available software and data to code images in minutes rather than days. We have not yet reached human-level performance, but we have shown that the curated datasets that have traditionally been collected improve performance over training on naturalistic uncontrolled images. This marks the first step towards building theories that explain how emotion interacts with cognition in real-world learning scenarios.

## Acknowledgments

This research was supported in part by NSF grant CISE-1623486 to L.B., V.L., and P.S.

## References

Baillargeon, R., Scott, R. M., \& He, Z. (2010). False-belief understanding in infants. Trends in cognitive sciences, 14(3), 110118.

Bradski, G. (2000). The open source computer vision library. Dr. Dobb's Journal of Software Tools.
Cohn, J. F., Zlochower, A. J., Lien, J., \& Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. Psychophysiology, 36(1), 35-43.
Cristianini, N. \& Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.
Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. Cognition, 107(2), 705-717.
Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. Psychological Review, 66(3), 183.
Egger, H. L., Pine, D. S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K. E., \& Angold, A. (2011). The NIMH child emotional faces picture set (NIMH-ChEFS): a new set of children's facial emotion stimuli. International Journal of Methods in Psychiatric Research, 20(3), 145-156.
Ekman, P. \& Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124.
Ekman, P. \& Rosenberg, E. L. (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press.
Itseez. (2016). Open Source Computer Vision Library. https : / / github.com/itseez/opencv.
LaBarbera, J., Izard, C., Vietze, P., \& Parisi, S. (1976). Four-and six-month-old infants' visual responses to joy, anger, and neutral expressions. Child Development, 47(2), 535-538.
Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., \& Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. Image and Vision Computing, 24(6), 615625.

LoBue, V. \& Thrasher, C. (2015). The child affective facial expression (CAFE) set: Validity and reliability from untrained adults. Frontiers in Psychology, 5, 1532.
Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., \& Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotionspecified expression. In IEEE conference on computer vision and pattern recognition (pp. 94-101).
Osuna, E., Freund, R., \& Girosit, F. (1997). Training support vector machines: An application to face detection. In IEEE computer society conference on computer vision and pattern recognition (pp. 130-136).
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
Pessoa, L. (2008). On the relationship between emotion and cognition. Nature Reviews Neuroscience, 9(2), 148-158.
Picard, R. W., Vyzas, E., \& Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state. IEEE transactions on pattern analysis and machine intelligence, 23(10), 1175-1191.
Tong, S. \& Chang, E. (2001). Support vector machine active learning for image retrieval. In The 9th ACM international conference on multimedia (pp. 107-118).
Viola, P. \& Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE conference on computer vision and pattern recognition (Vol. 1).
Yao, A., Shao, J., Ma, N., \& Chen, Y. (2015). Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In Proceedings of the 2015 acm on international conference on multimodal interaction (pp. 451-458). ACM.

# Reasons and the "Motivated Numeracy Effect" 

Cristina Ballarini (cristina_ballarini@brown.edu)<br>Department of Cognitive, Linguistic, and Psychological Sciences, Brown University<br>190 Thayer St, Providence, RI 02912 USA

Steven Sloman (steven_sloman@brown.edu)<br>Department of Cognitive, Linguistic, and Psychological Sciences, Brown University<br>190 Thayer St, Providence, RI 02912 USA


#### Abstract

Does the ability to reason well make one less likely to engage in motivated reasoning? Following a paradigm used by Kahan, Peters, Dawson, and Slovic (2013), this study aims to replicate, extend, and explain the surprising finding that those most likely to process politicized data in a biased manner are those who score highest on a measure of numerical proficiency. Although our study found general effects of motivated reasoning, we failed to replicate Kahan et al.'s "motivated numeracy effect". However, our study did find that, when forced to consider competing statistical interpretations of the data before responding, highly numerate participants were more likely than less numerate ones to choose a correct but belief-contradicting interpretation of data. These results suggest that while numerate participants were biased when generating responses, they were not when evaluating reasons to justify their responses.


Keywords: reasoning; motivated reasoning; decision making, science communication; inference; intelligence; rationality

## Introduction

When science becomes politicized, who can we trust to maintain objectivity? Conventional wisdom tells us that when it comes to assessing politically-charged data, those most capable of seeing past their biases and recognizing "the facts" are those most proficient in quantitative reasoning. If that's the case, then people high in numeracy-a measure of the disposition and capacity to engage with quantitative information-ought to process information more objectively and therefore exhibit less bias in assessing it. However, a body of research suggests not only that polarization increases with numeracy, but also that highly numerate people process politicized data in a more biased manner (Kahan, Peters, Wittlin, Slovic, Ouellette, Braman, \& Mandel, 2012; Kahan, Peters, Dawson, \& Slovic, 2013; Kahan, Jenkins-Smith, \& Braman, 2010).

In one of these studies (Kahan et al., 2013), participants were faced with a problem that tested their "ability to draw valid causal inferences from empirical data." Participants all saw the same two-by-two data table, but the data were framed either as the results of a pharmaceutical study of a new rash cream or the results of a study of the effects of gun control on crime rates. Correctly interpreting the results required participants to detect covariance between the relevant intervention and two outcomes, and the numbers in each cell of the table were chosen such that a conclusion drawn using
one of two known "heuristic strategies" (comparing either the absolute value of positive outcomes or the difference in positive and negative outcomes between the two groups) did not agree with a conclusion correctly drawn using the "covariance strategy" (comparing the ratio of positive to negative outcomes between the two groups). Unsurprisingly, the authors found that participants highest in numeracy were most likely to give the correct response in the politicallyneutral rash cream version of the problem. But in the politically-sensitive gun control version, highly numerate participants only performed better than their low numeracy counterparts in cases where using the covariance strategy would lead to a conclusion that aligned with their political beliefs.

This "motivated numeracy effect" fits with a large body of literature on motivated reasoning. In her comprehensive review, Kunda (1990) writes that while "people are more likely to arrive at conclusions that they want to arrive at... their ability to do so is constrained by their ability to construct seemingly reasonable justifications for these conclusions". In many cases, such a "reasonable justification" will take the form of a processing strategy biased towards a favorable outcome (Giner-Sorolla \& Chaiken, 1997). If the information is quantitative in nature, this processing strategy may be a statistical heuristic (Ginossar \& Trope, 1987; Petty \& Cacioppo, 1986). Taken together, these studies predict that if a sophisticated strategy like covariance detection is required to arrive at a favorable conclusion, then people low in numeracy will be limited in their ability to engage in motivated reasoning. On the flip side, just because highly numerate people possess both a heuristic and normative strategy does not mean that their application of these strategies will be unbiased. To this point, Stanovich and West (2008) found that studies in which correlations were found between cognitive ability and unbiased processing supplied cues which signaled that unbiased processing was requiredmost often, these studies employed a within-subjects design where the conflict between a normative and heuristic rule was transparent. When this conflict was obscured, Stanovich, West, and Toplak (2013) found the degree of "myside bias"-the tendency to evaluate evidence, generate evidence, and test hypotheses in a manner biased towards one's priors-to be uncorrelated with measures of cognitive ability.

If the motivated numeracy effect depends crucially on both the salience and availability of heuristic and normative statistical strategies, what would happen if participants were to respond to Kahan et al.'s problems with both strategies at
hand? If motivated participants are inexorably biased in their selection of statistical strategies (as Kahan et al. suggest in their discussion), then providing "reasons" that describe each strategy before participants make their final judgement should infect the responses of those low in numeracy with the same bias exhibited by those high in numeracy. Alternatively, if motivation only obscures the need for heuristic override, then providing reasons may reduce bias in one of two ways. Assuming both low and highly numerate participants are equally cued by the reasons, performance should improve for all participants, regardless of whether the data they encounter are belief-affirming or belief-contradicting. However, if those high in numeracy are better able to recognize and respond to a conflict between normative and heuristic rules, then providing all participants with reasons should only benefit those high in numeracy.

## Study

This study examined the effect of reasons on the motivated numeracy effect, and attempted to replicate and generalize Kahan et al.'s (2013) results.

## Method

Participants Seventy-six undergraduates at Brown University participated in the study for course credit. Each participant attended one of four identical group sessions (excluding one who completed the survey during an individual session). All participants completed the study on personal computers.

Procedure Each participant was randomly assigned either to the "no reasons" or "reasons" condition. In both conditions, nine fictional experiments were described and the observed results were presented in the form of either a table, a scatterplot, or a histogram (three of each in total). Six of these data analysis problems dealt with politicized issues and three of the problems dealt with non-politicized issues.

In the "no reasons" condition, participants saw the results of the experiment and, on the same page, were asked to indicate which conclusion the described experiment supported. To respond, participants could select either between one of two opposing conclusions or could select "Other" and fill in a response. In the "reasons" condition, participants saw the results and, on the same page, were asked to select "the interpretation that best explains the data". Then, on a separate page, participants saw the data again and were asked to respond just as participants in the "no reasons" condition did.

Immediately following the study, participants completed a nine-item numeracy scale and were asked to report their political outlook (5-point Likert scale), political affiliation (7point Likert scale), and prior beliefs about each of the politicized issues presented in the study (7-point Likert scales). In a final series of questions, participants shared their
suspicions, confusions, and any other thoughts concerning the study.

Stimuli The six motivated problems looked at (1) the effect of gun control measures on crime rates, (2) the effect of mandatory anti-bias training on the number of minority civilians shot by police, (3) the effect of affirmative action on company profitability, (4) the effect of undocumented immigrant populations on violent crime rates, (5) the effect of stop-and-frisk practices on crime rates, and (6) the effect of taxing coal on unemployment rates. The three neutral problems looked at (1) the effectiveness of a rash cream, (2) the effectiveness of a fertilizer, and (3) the relationship between a property's distance from a city and the real estate commission earned on that property. For each problem, the conclusion that the data supported was randomized for each participant by switching column or axis labels.

Of the nine problems, three presented results in table form (gun control, affirmative action, and rash cream), three in the form of a scatterplot (stop-and-frisk, immigration, and property), and three in the form of a histogram (anti-bias, coal, and fertilizer). A random sequence of three blocks of table-scatterplot-histogram was generated, and the order in which the nine problems appeared was then counterbalanced using a Latin square. Selecting the correct answer in case of a table or a histogram required participants to detect covariance, just as in Kahan et al.'s (2013) original study. Selecting the correct answer in case of a scatterplot required participants to notice an overall positive (about 0.40 ) or overall negative (about -0.40 ) correlation rather than extreme outliers.

For tables and histograms, the interpretations presented in the "reasons" condition appealed to one of the two known "heuristic" responses first described by Wasserman, Dorner, and Kao (1990) or to covariance (a total of three reasons plus a fill-in-the-blank, "Other" response option). For scatterplots, the reasons drew attention to either overall correlation or outliers (a total of two reasons plus a fill-in-the-blank, "Other" response option).

Table 1: Reasons Provided for Gun Control Problems

| Heuristic A-C | Heuristic A-B | Covariance |
| :---: | :---: | :---: |
| The group that did ban carrying concealed handguns in public has more cities who saw an [increase/decrease] in crime than does the group that did not ban carrying concealed handguns in public. | Comparing the number of cities that saw crime <br> [increase/decrease] to the number of cities that saw crime [decrease/increase], there is a greater difference for the group that did ban carrying concealed handguns in public. | The ratio of cities that saw crime <br> [decrease/increase] to cities that saw crime [increase/decrease] is larger for the group that did ban carrying concealed handguns in public than for the group that did not. |

The numeracy scale used included five questions adapted from Weller et al. (2013) and four CRT questions-three
from Toplak, West, and Stanovich (2014) and one from Frederick (2005).

## Results

Participants performed at ceiling on problems in which the results were presented in the form of a scatterplot ( $M_{\text {stop-and- }}$ frisk $=0.95, M_{\text {immigration }}=0.91, M_{\text {property }}=0.99$ ), so responses to these problems were excluded from the analysis. Responses to the coal problem were also excluded, as the mean extremity of priors reported at the end of the study indicated that, for this group of participants, the issue was not a motivated one. On the coal issue, participants reported priors that averaged 0.64 points from "No opinion", compared to $1.6,2.1$, and 2.3 points on affirmative action, gun control, and bias, respectively. That left responses to three motivated problems (gun control, anti-bias, affirmative action) and two neutral problems (rash cream, fertilizer) from each participant for analysis- 228 responses in total.

Bias and Priors Participants were predominantly liberaldemocratic. In terms of outlook, $1 \%$ identified as "conservative" or "very conservative on a five-point scale. In terms of affiliation, 5\% identified as Republican leaning, Republican, or strong Republican on a seven-point scale. On all three motivated issues, participants reported substantial bias consistent with their liberal leanings-on a seven-point scale, $5 \%$ did not support gun control, $1 \%$ did not support mandatory anti-bias training in police departments, and 5\% did not support affirmative action. No significant difference in reported priors was found between HN participants and LN participants on gun-control $(t(72.99)=-0.26, p=0.80)$, antibias training $(t(64.47)=0.49, p=0.62)$, or affirmative action $(t(72.39)=-0.14, p=0.89)$. To simplify the analysis, problems that supported the conclusion that gun control lead to a decrease in crime, that anti-bias training lead to a decrease in the number of minority civilians shot by police, or that affirmative action lead to an increase in company profit are labeled as "motivated affirming". Problems that supported the conclusion that gun control lead to an increase in crime, that anti-bias training lead to an increase in the number of minority civilians shot by police, or that affirmative action lead to a decrease in company profit are labeled as "motivated contradicting".

Numeracy Average numeracy was 6.25 out of nine. Numeracy classes were assigned by a median split, with "high numeracy" (HN) referring to participants with numeracy scores of 7 or above and "low numeracy" (LN) referring to participants with numeracy scores of 6 or below. The sizes of the resulting groups were 36 and 40 , respectively. Numeracy scores were higher in the "reasons" condition ( $M=6.58, S D=2.39$ ) than in the "no reasons" condition ( $M=5.92, S D=1.75$ ), though not significantly so; $t(67.74), p=0.175$. The difference can be attributed to CRT items: Though not significant, the two groups differed in performance on the four CRT items, $t(71.24)=-1.50, p=$
0.14 , while performance on the other five numeracy items was the same, $t(71.31)=-0.79, p=0.43$.

Motivated Reasoning and the Effect of Numeracy After excluding 21 participants who did not encounter at least one problem of each valence (i.e. neutral, motivated affirming, and motivated contradicting), a repeated measures analysis of variance (ANOVA) revealed a significant effect of problem valence on performance, $F(2,102)=4.99, p=0.008, \eta_{G}{ }^{2}=$ 0.041 . This analysis also revealed a significant effect of numeracy on performance, $F(1,51)=20.01, p<0.001, \eta_{G}{ }^{2}=$ 0.18 , with highly numerate participants more likely to respond correctly than their low numeracy counterparts.

Unlike Kahan et al. (2013), in our "no reasons" condition, we found no significant difference in performance between HN and LN participants on neutral $(t(17.75)=1.59, p=0.13)$ motivated affirming $(t(21.90)=1.50, p=0.15)$, and motivated contradicting problems $(t(22.36)=0.64, p=0.53)$. After excluding the 21 participants, we also failed to find a significant interaction between numeracy and problem valence, $F(2,50)=0.058, p=0.944, \eta_{G}^{2}<0.001$. While participants clearly exhibit motivated reasoning, we failed to find a significant effect of motivated numeracy.

Figure 1: Performance in "No Reasons" and "Reasons" Conditions


The Effect of Reasons The analysis did reveal a significant interaction between numeracy and reasons, $F(1,51)=6.20, p$ $=0.016, \eta_{G}{ }^{2}=0.064$. HN participants performed significantly better in the "reasons" condition than in the "no reasons" condition, $\left(M_{\text {reasons }}=0.68, M_{\text {no reasons }}=0.43, t(55.52)=2.34\right.$, $p=0.023$ ), while LN participants performed significantly worse with reasons ( $M_{\text {reasons }}=0.13, M_{\text {no reasons }}=0.28, t(88)=$ $2.15, p=0.034)$. Note, however, that for both groups of participants, the "reasons" manipulation served to reduce bias, as measured by the difference in performance on motivated affirming and motivated contradicting problems (Figure 1). For HN participants, Bias $_{\text {no reasons }}=0.25$, while Bias $_{\text {reasons }}=0.03$. For LN participants Bias $_{\text {no reasons }}=0.26$, while Bias $_{\text {reasons }}=0.11$.

## General Discussion

While reasons served to reduce bias in all participants, those high in numeracy were better able to make use of those reasons to improve their performance. These data suggest that, in the presence of motivation, reasoners are not inexorably biased in their selection of statistical strategiescomparatively evaluating reasons can serve to block the effects of motivation, but only if one is able to understand those reasons. Motivation may encourage less reflective reasoning, but this effect is not irreparable.

Though we failed to find a significant effect of motivated numeracy, it is important to note that Kahan et al.'s (2013) original study analyzed 1111 observations from 1111 participants, while our study analyzed 206 observations from 55 participants ( 55 after the 21 participants who didn't encounter problems of each valence were excluded). As Kahan et al. note, in this paradigm, the "strength of inferences drawn from 'null' findings depends heavily on statistical power", and our sample may have been too small to detect the effect of motivated numeracy those researchers found. If Kahan et al.'s original findings are valid, these results support the hypothesis that the motivated numeracy effect results from belief bias obscuring the need for heuristic override. While motivation biased HN participants in their selection of an appropriate statistical strategy, they could appreciate the correct strategy when it was presented (and when the conflict between normative and heuristic strategies was apparent). But in any case, whether polarization increases with numeracy or whether it remains constant, our results suggest that evaluating reasons can reduce the effect of this polarization on reasoning.

If reasons served to block the effects of motivated reasoning, in virtue of what did they do so? The data suggest that evaluating reasons may have encouraged reflectiveness. To this point, not only were CRT scores higher in the "reasons" condition, but considered together, our four CRT items were the best predictor of a correct response on motivated contradicting, motivated affirming, and neutral problems. Recall that numeracy scales were completed after responding to the data analysis problems, suggesting that
evaluating reasons may have elicited a more analytic frame of mind.

Table 2: Correlations Between Numeracy Scale Items and Task Performance

| Numeracy Scale Item | Performance on <br> Motivated Tasks | Performance on <br> Neutral Tasks |
| :--- | :--- | :--- |
|  |  |  |
| N1 | -0.03 | 0.17 |
| N2 | 0.11 | 0.11 |
| N3 | 0.07 | 0.20 |
| N4 | 0.07 | 0.17 |
| N5 | 0.22 | 0.10 |
| CRT1 | 0.43 | 0.31 |
| CRT2 | 0.29 | 0.33 |
| CRT3 | 0.32 | 0.32 |
| CRT4 | 0.44 | 0.39 |
|  |  |  |
| NUMERACY-CRT | 0.17 | 0.24 |
| NUMERACY (incl. CRT) | 0.42 | 0.43 |
| CRT | 0.52 | 0.47 |

These results additionally suggest that CRT is not just a measure of numeracy, a position debated in the literature (Liberali, Reyna, Furlan, Stein, \& Pardo, 2012); the CRT scale was always a better predictor of performance on the covariance detection task than the numeracy scale considered without the CRT items. While there was a ceiling effect for three of the numeracy items ( $\mathrm{N} 1, \mathrm{~N} 2, \mathrm{~N} 4$ ), the two items for which this effect was absent still showed a lower correlation with performance compared to the CRT. Out of the non-CRT items, only N5 (a Bayes's rule problem) showed a correlation comparable to any of the CRT items. However, unlike N1N4, N5 may be more of a measure of reflectiveness than quantitative ability, per se-even with a frequency chart, the majority of people fail to attend to base-rates in problems like N5 (Bar-Hillel, 1980; Gigerenzer \& Hoffrage, 1995), and this base-rate neglect is correlated with low CRT scores (Hoppe \& Dusterer, 2011).

That said, considering that those low in numeracy generally performed worse with reasons, what appears to be a decrease in bias may not be the result of increased reflectiveness. To this point, unlike their highly numerate counterparts, LN participants in the "reasons" condition performed worse on CRT items (average scores of 0.94 ) than they did in the "no reasons" condition (average score of 1.25). Why might this have been the case? One hypothesis is that those low in numeracy had trouble understanding the reasons provided. But the number of LN participants who reported experiencing some confusion during the study (12\%) was comparable to the number of HN participants who reported experiencing confusion ( $9 \%$ ). What's more likely is that those low in numeracy were unable to appreciate and make use of the covariance strategy when it was presented as a reason.

There are three alternative explanations for our results that are important to consider. First, the difference in CRT scores between the "reasons" and "no reasons" condition may
not have reflected an effect of reasons on reflectiveness, but only an unfortunate selection confound. However, it's not clear how the presence of such a confound would affect our conclusions. As our analyses conditioned on numeracy, selection bias would only affect sample sizes, not mean performance scores.

It might also be suggested that the effect of reasons resulted from a task demand. Because in the "reasons" condition, participants saw similar reasons presented with both neutral and motivated versions of the problem, they may have come to suspect that the study was testing their bias. This may have been the case with HN participants in the "reasons" condition, $32 \%$ of whom reported suspicions about the study (e.g. "I thought that this study was probably testing how our beliefs influence our abilities to analyze the data"). Fewer LN participants ( $19 \%$ ) reported suspicions about the experiment, suggesting that if such a task demand was present, HN participants were better at picking up on (as well as responding to) it. The crucial point, though, is that even if HN participants were responding to a task demand, they could only supply responses that they thought experimenters wanted to hear if they could determine what those responses were. The fact that their responses were so often correct is consistent with our conclusion that HN participants were better at recognizing a correct response.

Third and most importantly, it could be argued that our results support an alternative explanation of Kahan et al.'s motivated numeracy effect: namely, that HN participants were more motivated because they had stronger priors. Here, we found no difference in the extremity of priors between low and highly numerate participants, and we also failed to find a significant motivated numeracy effect. Ultimately, while it is not clear how the alternative explanation could explain the effect of reasons, this is a pressing question for future research.

The implications of these results for science communication complicate Kahan et al.'s (2013) conclusions. While Kahan et al. concluded that "improving public understanding of science and propagating critical reason skills... cannot be expected to dissipate persistent public conflict over decision-relevant science", our study indicates that understanding and being able to make use of normatively correct interpretational strategies can make people more responsive to debiasing efforts, at least when those efforts encourage reflective processing.

## Conclusion

These data suggest that providing reasons can block the effects of motivated reasoning, and that such intervention is most successful for those high in numeracy. Though highly numerate people are more able to recognize when a sophisticated statistical strategy is appropriate, this recognition is impaired when a more immediate, heuristic strategy points to their desired conclusion. Making the need for heuristic override salient improves performance for those
high in numeracy, but is not enough to affect those low in numeracy.

## Acknowledgements

We thank Babak Hemmatian and Marc Lluis for their assistance with statistical analysis, and Dan Kahan for his helpful comments. This work was supported by the Varieties of Understanding Project at Fordham University and the John Templeton Foundation. The opinions expressed are the authors' and do not necessarily reflect the views of the Varieties of Understanding Project, Fordham University, or the John Templeton Foundation.

## References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44(3), 211-233.

Frederick S. (2005). Cognitive reflection and decision making. J. Econ. Perspect. 19, 25-42

Gigerenzer, G., \& Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. Psychological review, 102(4), 684.

Giner-Sorolila, R., \& Chaiken, S. (1997). Selective use of heunrstic and systematic processing under defense motivation. Personality and Social Psychology Bulletin, 23(1), 84-97.

Ginossar, Z., \& Trope, Y. (1987). Problem solving in judgment under uncertainty. Journal of Personality and social Psychology, 52(3), 464.
Hoppe, E. I., \& Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. Economics Letters, 110(2), 97-100.

Kahneman, D., \& Tversky, A. (1973). On the psychology of prediction. Psychological review, 80(4), 237.

Kahan, D. M., Jenkins-Smith, H., \& Braman, D. (2011). Cultural cognition of scientific consensus. Journal of Risk Research, 14(2), 147-174.

Kahan, D. M., Peters, E., Dawson, E. C., \& Slovic, P. (2013). Motivated numeracy and enlightened selfgovernment. Yale Law School, Public Law Working Paper, (307).

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., \& Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. Nature climate change, 2(10), 732-735.

Kunda, Z. (1990). The case for motivated reasoning. Psychological bulletin, 108(3), 480.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., \& Pardo, S. T. (2012). Individual Differences in Numeracy and Cognitive Reflection, with Implications for Biases and Fallacies in Probability Judgment. Journal of Behavioral Decision Making, 25(4).

Petty, R. E., \& Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In Communication and persuasion (pp. 1-24). Springer New York.

Stanovich, K. E., \& West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. Journal of personality and social psychology, 94(4), 672.

Stanovich, K. E., West, R. F., \& Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. Current Directions in Psychological Science, 22(4), 259-264.

Toplak M. E., West R. F., Stanovich K. E. (2014). Assessing miserly information processing: an expansion of the Cognitive Reflection Test. Think. Reason. 20, 147-168.

Weller J. A., Dieckmann N. F., Tusler M., Mertz C. K., Burns W. J., Peters E. (2013). Development and testing of an abbreviated numeracy scale: a Rasch analysis approach. J. Behav. Decis. Mak. 26, 198-212.

# Belief Updating and Argument Evaluation 

Megan D Bardolph (mbardolph@ucsd.edu)<br>Department of Cognitive Science (0515), 9500 Gilman Drive<br>La Jolla, CA 92093 USA<br>Seana Coulson (scoulson@ucsd.edu)<br>Department of Cognitive Science (0515), 9500 Gilman Drive<br>La Jolla, CA 92093 USA


#### Abstract

Studies of how evidence affects beliefs sometimes show belief polarization in response to mixed evidence. However, the nature of the mental processes leading to change in opinion is up for debate. Different accounts of how people process evidence and then update their beliefs make different predictions, especially about one-sided evidence, which is rarely examined. We presented subjects with multiple text arguments regarding socio-political topics as one-sided or mixed evidence. Participants rated arguments differently according to their extant beliefs, which is consistent with accounts of motivated reasoning. They did not polarize afterward, instead showing evidence of belief updating according to Bayesian principles: belief change is sensitive to prior opinions and to the direction and quality of the evidence presented. These data support rethinking some of the mental processes underlying incorporation of evidence into a personal belief structure.


Keywords: cognitive science; decision making; reasoning; language and thought; psychology; motivated reasoning; rationality

## Introduction

As people navigate a world filled with information, they must make decisions in order to accomplish various goals. The beliefs that individuals hold provide a structure in which new information is evaluated and potentially integrated with existing beliefs. Different models explain how people evaluate information and use that information to update their beliefs, leading to potentially different implications about human rationality (Oaksford \& Chater, 2007; Klayman \& Ha, 1987).

Information from the world can be thought of as data, or evidence, supporting or disconfirming a hypothesis. The evidence may be accepted without examination or it may be judged before it is used to update one's beliefs, or a hypothesis (Nisbett \& Ross, 1980; Pyszczynski \& Greenberg, 1987). From a Bayesian perspective, evidence is judged according to existing hypotheses about the world along with data that has already been observed. Bayes' rule provides a general model for constructing a posterior probability, $P$ (hypothesis $\mid$ data), as a function of prior beliefs and observed evidence: $P($ hypothesis $) * P($ data $\mid$ hypothesis $)$.

People's prior hypotheses about the world may differ depending on the data they have observed. Furthermore, the nature of people's hypothesis space for a given topic is not always easy to define (Tenenbaum, Kemp, Griffiths, \& Goodman, 2011; Jern, Chang, \& Kemp, 2014). Accounts of Bayesian updating explicitly allow for evidence to be treated differently depending on whether it is in agreement with one's prior beliefs (Gerber \& Green, 1999). However, unless there
is reason to suspect the source or validity of the evidence, Bayesian normative accounts still require that people update their prior beliefs in the direction of the evidence. This updating can be small, but it cannot be in the opposite direction. If such a shift occurs, it should be viewed as a violation of normative updating under this model.

Alternatively, differential rating of information (evidence) may be due to motivated, or hot cognition processes (Kunda, 1990; Ditto \& Lopez, 1992). Under motivated accounts, evidence compatible with an extant opinion is accepted, while incompatible evidence creates negative emotions and is therefore critically examined and judged more negatively because of its incompatibility. This difference in judgement can lead to attitude polarization, or belief polarization.

Lord, Ross, and Lepper (1979) suggested that attitude polarization occurs because people with opposing views can come to opposite conclusions from the very same set of evidence. In a classic study, the authors queried participants about their views on capital punishment, and then revealed the results of two studies, one which suggested the death penalty deters crime, and one which suggested the opposite conclusion. Participants were asked to rate the quality of each study, and then to recharacterize their views on the death penalty. The authors found that proponents of capital punishment rated the study showing the deterrent effect of the death penalty to be superior to that showing that the death penalty did not affect crime levels, and subsequently adjusted their beliefs to more strongly favor capital punishment. By contrast, opponents of capital punishment favored the study that showed the death penalty had little effect on crime, and subsequently adjusted their beliefs to more strongly oppose capital punishment. So-called biased assimilation is the phenomenon by which participants' prior beliefs impact the way they evaluate novel evidence, and it would seem to undermine the possibility of achieving consensus (Lord, Ross, \& Lepper, 1979).

Taber and Lodge (2006) suggest that "primacy and automaticity of affect kick-start the processes that spark motivated biases when citizens encounter attitudinally contrary information." Taber and colleagues (2009) found evidence of an attitude congruency bias, where people evaluate arguments and evidence that supports their prior opinions as stronger than nonsupporting information; and attitude polarization, where this bias leads to polarization with exposure to the same set of information.

The present study aims to clarify the differences between processing compatible and incompatible evidence, separating the effects of the two types of evidence. To do this, we will examine evidence rating for both mixed evidence (as used in prior studies) and one-sided evidence (previously missing from much of the literature). Studies using mixed evidence imply that participants process congruent and incongruent information using different processes; for example, readily accepting compatible arguments while spending more time and mental resources to undermine incompatible arguments (Edwards \& Smith, 1996; Taber, Cann, \& Kucsova, 2009). It is not clear whether compatible and incompatible arguments must be presented together to activate these processes or whether they apply to congruent and incongruent arguments due to the nature of the evidence alone. The inclusion of mixed and non-mixed (one-sided) evidence allows for examination of potential differences.

This study further aims to examine whether belief updating behavior supports a motivated reasoning account or a Bayesian account of belief updating. This will be assessed by testing whether participants' beliefs change as a function of biased assimilation of the evidence, dependent on their prior beliefs, or whether belief change depends on the direction and/or merits of the evidence.

## Methods

## Participants

Participants were 124 students ( 75 female) enrolled in Psychology, Linguistics, or Cognitive Science courses at the University of California, San Diego (UCSD) participating as part of a course requirement. All participants provided informed consent, and procedures were approved by the Institutional Review Board (IRB) at UCSD. Participants were between 18 and 35 years old, with a mean age of 21 . An additional two participants completed the survey, but their results were not included, either because their responses suggested they did not understand the rating scale $(\mathrm{n}=1)$, or because their age was greater than 35 years $(\mathrm{n}=1)$.

## Materials

The study concerned six socio-political issues: abortion, animal testing, assisted suicide, climate change, the death penalty, and school uniforms. These issues were among the most popular topics covered on two debate websites, www.procon.org and idebate.org.

Attitude measurements: For each issue, a single policy statement was chosen for participants to rate in terms of how much they agree or disagree (e.g., "Animal testing should be banned."). This was followed by four position statements for each issue selected from two headings under "Points for" on the idebate.org archive, and two from "Points against." Participants responded to all five of these position statements, and these responses formed the initial attitude measurement. After the experimental treatment, participants again responded to five position statements per issue to form the subsequent
attitude measurement.
Strength measurements: For each issue, participants were given four questions with a 9-point Likert scale to indicate how much they cared about, and had thought about, that issue. These four questions were combined to form a measure of strength of conviction.

Arguments: Using text from the websites, 6 supporting (Pro) and 6 opposing (Con) arguments were selected for each issue. Arguments were generally matched for content (i.e., if a Pro and a Con argument addressed the same point, both arguments were usually selected), and for length (mean argument length $=120$ words, $s d=11$ ). To create arguments of similar length, portions of longer arguments were omitted.

## Procedure

The study included three phases: initial collection of attitude and conviction strength measurements, the presentation and rating of arguments, and the subsequent collection of attitude and strength measurements.

Initial collection of attitude and strength measurements proceeded one issue at a time, as participants first rated their attitude on the issue, and then responded to the questions regarding the strength of their convictions on that issue. The presentation order of the six issues was randomly determined.

Following the collection of attitude and strength measurements, each participant was asked to read and rate arguments for three randomly chosen issues from the original set of six. For these three issues, one was randomly designated as the Pro condition, such that the participant read and rated six arguments in support of the original position; one was randomly designated as the Con condition, such that the participant read and rated six arguments against the original position; and one was randomly designated as the Mix condition, such that the participant read and rated three arguments in support of the original position, and three arguments against. The order of the issues was randomized, as was the order of the arguments presented within each issue. Treatment thus included four treatment conditions: Pro, Con, Mix, and None, with the None condition comprising four issues for which participants were not presented any argument text.

After reading all arguments, participants were again asked to rate their positions on all six issues. Next, participants completed a brief political knowledge quiz to assess their political sophistication, and two questions to assess openmindedness. Finally, they read a debriefing page that explained the goal of the study and provided links to the websites used for the argument texts.

## Analysis

Opinions were scaled from -5 to 5 , with -5 representing the opinion most against the issue and 5 representing the opinion most in favor of the issue (each issue is framed as a statement, e.g. "The death penalty should be banned."). Items where participants spent too long reading the argument text (more than 153 seconds, 3 standard deviations from the mean) were removed from analysis (28 items out of 2232).

Participants' prior opinions and strength of conviction were analyzed to ensure uniform representation across conditions, since within each issue, experimental conditions (Pro, Con, Mix, or None) were varied between subjects. A linear model of prior opinion as a function of treatment condition and issue showed that although opinions varied by issue, there were no significant differences among conditions (Pro, Con, Mix, None), nor was there any interaction of issue and condition. Similarly, strength of conviction did not vary as a function of treatment condition.

Models of argument rating were analyzed with a linear mixed effects regression (LMER) model using the lme4 package in R (Bates, Maechler, Bolker, Walker, et al., 2014; R Core Team, 2015). All experimental factors were allowed to interact initially; more complex models were compared with more parsimonious models using model ANOVA in R. Models were fit with random intercepts for subjects and items (viz. arguments). The reported models are those that included statistically significant predictors of argument rating and are not statistically different from more complex models (using cutoff $\mathrm{p}<.01$ ).

Models of belief updating were analzed with a linear model in R. Again, all experimental factors were allowed to interact initially; more complex models were compared with more parsimonious models using model ANOVA in R. This is equivalent to selecting all predictors with a significant p value ( $\mathrm{p}<.01$ ) in the model ANOVA.

## Results

The present study was designed (i) to replicate patterns of argument evaluation shown in other studies (Lord et al., 1979; Edwards \& Smith, 1996; Taber et al., 2009) and (ii) to critically examine whether biased argument rating leads to belief updating, as suggested by a motivated account of reasoning, or whether belief change can be better explained by a Bayesian account in which participants are sensitive to the merits of the evidence.

The motivated cognition account explains attitude polarization as resulting from a biased assimilation of the evidence, such that evidence compatible with participants' initial positions is weighted more heavily than incompatible evidence, and consequently has a disproportionate impact on the way participants update their beliefs. We first assessed whether participants evaluated the arguments in a biased manner by analyzing whether their ratings of these arguments differed systematically as a function of their prior beliefs. Next, we assessed the factors that influenced belief change in response to these arguments.

## Argument Rating

As noted above, our first question was whether participants rated evidence differentially as a function of its compatibility with their initial attitudes about the relevant issue. To examine this question, we began by modeling participants' argument ratings with a linear mixed effects model with predictors of treatment condition (Pro, Con, or Mix), argument po-


Figure 1: Average argument rating as a function of prior opinion ( -5 most opposed, 5 most in favor of the issue). Green lines represent Pro arguments presented in the Pro and Mix conditions; Red lines represent Con arguments presented in the Con and Mix conditions.
larity (Pro or Con), prior opinion, strength of conviction, issue, and political sophistication. Argument polarity is coded separately from Condition and represents Pro and Con arguments irrespective of which experimental condition they were presented in. The goal of this variable coding procedure was to separate potential effects of experimental condition from effects of argument polarity.

Our model selection procedure revealed that experimental condition per se was irrelevant. The best model predicts argument rating as a function of prior opinion and argument polarity only. There was a trending further interaction with strength of conviction, with the slope of the rating x prior opinion line being steeper for participants with high strength of conviction ( $\mathrm{p}=.015$ for the 3 -way interaction). Other experimental variables did not show main effects or interact with experimental variables. The mixed effects linear model includes random subject intercepts and individual argument intercepts. See Equation 1 and Table 1 for model results.

Argument rating $\sim$ prior opinion $*$ argument polarity

Table 1: Model results for Equation 1.

| Factor | df | F value |
| :--- | :---: | :---: |
| Prior opinion | 1 | 1.5 |
| Argument polarity | 1 | 0.3 |
| Prior x Argument polarity | 1 | 125.1 |

Figure 1 shows how argument ratings differ as a function of participants' prior opinions, with separate green regression lines shown for supporting arguments presented in the Pro condition and in the Mix condition, and separate red regression lines for opposing arguments presented in the Con condition and in the Mix condition. The positive slope of both green lines reflects the fact that the more participants support the issue, the higher they rate the Pro arguments compatible with their position. The similarity in the slope of the Mix and the Non-Mix line indicates that participants' ratings of these arguments were similar, regardless of whether they were presented in the context of other Pro arguments, or with a mixture of Pro and Con arguments. Likewise, the negative slope of both red lines reflects systematic bias in the ratings of opposing arguments, with opponents ( -5 on the $x$-axis) rating those arguments higher than supporters ( +5 on the $x$-axis), irrespective of whether opposing arguments were presented in a Con or a Mix block.

## Belief Updating

We are interested in what factors lead to belief updating, or opinion change, after participants read and rate the arguments. Specifically, experimental condition might interact with participants' prior opinions, showing that belief updating due to different types of evidence (i.e., that presented in the Pro, Con, and Mixed conditions) differs as a function of their original position regarding that issue. Strength of conviction may also influence opinion change if participants whose beliefs are stronger are either more motivated to defend their position or rely on a greater body of knowledge to form their prior opinion. Because participants may change their opinions differently by issue, issue is also included as a predictor. Finally, we included a measure of political sophistication because previous studies have suggested that sophisticated individuals are more likely to engage in motivated reasoning (Taber et al., 2009).

Opinion change was modeled as a function of treatment condition (Pro/Con/Mix), prior opinion, strength of conviction, issue, and political sophistication. Linear models as described in the Analysis section were created to investigate the effects of these factors on opinion change. The best model to predict opinion change is shown in Equation 2.

$$
\text { Opinion change } \sim \text { condition }+ \text { prior opinion } * \text { strength }
$$

Table 2: Model results for Equation 2.

| Factor | df | Estimate | F value | p value |
| :--- | :---: | :---: | :---: | :---: |
| Condition | 2 |  | 19.4 | $<.001$ |
| Prior opinion | 1 | -.41 | 96.2 | $<.001$ |
| Strength | 1 | -0.02 | 0.78 | .38 |
| Prior x Strength | 1 | 0.03 | 7.55 | $<.01$ |

The effect of experimental condition on opinion change is shown in Figure 2. On average, independent of prior beliefs,


Figure 2: Average opinion change for each treatment condition. Lines represent standard error.
participants' opinions were more in favor of an issue after viewing and rating arguments in the Pro condition; more opposed to the issue after viewing and rating arguments in the Con condition; and unchanged after viewing arguments in the Mix condition.

Overall, participants shifted their opinion toward a more moderate point of view (and also in the direction of the evidence), with participants more in favor of an issue changing their opinion to be less in favor, and those opposed changing their opinion to be more in favor. This center-trending behavior is represented in the negative coefficient of prior opinion in the model. Prior opinion further interacts with strength of conviction such that participants with lower strength show more center-trending than do those with higher strength of conviction.

The prior opinion x strength interaction is shown in Figure 3. Values for opinion change were baseline corrected by subtracting prior opinion $*$ opinion change slope for the None condition to show how much opinion changed when participants viewed and rated arguments. This visually removes the overall center-trending pattern observed for all conditions. Participants with high strength of conviction did not show a difference in opinion change compared to baseline. Those with low strength of conviction show an additional centertrending pattern, with participants more in favor of an issue changing to be more opposed, and participants more opposed to an issue becoming more in favor.

Finally, we were interested in whether participants' argument ratings would influence their beliefs in addition to the other factors. Motivated cognition accounts would predict that participants who exhibit biased rating behavior will be more likely to polarize, updating their beliefs in the direction


Figure 3: Interaction of prior opinion and strength. The blue line represents participants with low strength of conviction for a given issue, and the pink line represents those with high strength of conviction. Values have been corrected to remove the center-trending slope of the None condition to show their difference from baseline.
of their initial opinion. By contrast, Bayesian updating predicts that participants will rely only on the evidence. Consequently, they will either move in the direction of the evidence (irrespective of their prior beliefs), or maintain their original point of view.

Model comparison revealed that when participants’ average argument rating was included as a predictor, the most parsimonious account of opinion change is given by the factors in Equation 3. As in Equation 2, the main effect of treatment condition and the interaction between prior opinion and strength of conviction were present. In addition to the previous predictors, opinion change is further predicted by an interaction of argument polarity and argument rating. As shown in Figure 4, this interaction term results because participants' opinions on average change to be more congruent with the position of those arguments that participants rated highly. The higher a given participant rated Pro arguments, the more their opinion changed in the positive direction. The higher they rated Con arguments, the more their opinion changed in the negative direction.

$$
\begin{array}{r}
\text { Opinion change } \sim \text { condition }+ \text { prior opinion } * \text { strength } \\
+ \text { argument polarity } * \text { argument rating } \tag{3}
\end{array}
$$

Figure 4 shows this interaction of argument rating x argument polarity (Pro/Con). The occurrence of prior opinion and argument ratings in separate, additive terms in Equation

Opinion change by argument rating


Figure 4: Interaction of argument rating and argument polarity. The red line represents average opinion change for Con arguments in the Con or Mix condition; the green line represents average opinion change for Pro arguments.

3 suggests that the relationship between argument rating and opinion change was independent of participants prior beliefs. That is, whether or not participants initially agreed with the policy embraced in a given argument, they changed their positions to be more congruent with the arguments, especially for highly rated arguments.

## Discussion

## Argument rating

Participants rated arguments that were compatible with their prior policy opinions as objectively better than arguments that were incompatible with those opinions. Moreover, this bias scaled linearly with participants' prior opinions, as those at either end of the scale showed the greatest bias in argument ratings. This argument rating bias is consistent with previous findings, potentially supporting the motivated reasoning account. However, these findings are also consistent with a Bayesian reasoning account in which participants at the ends of the scale are assumed to assign a high prior probability to their own position, and naturally assess the likelihood of congruent evidence to be higher than that of incongruent evidence. To dissociate motivated from Bayesian reasoning, it is necessary to examine the opinion change data.

## Belief updating

The belief updating data provide support for a Bayesian account and show that even in the presence of biased argument ratings, participants changed their beliefs in response to the evidence. The final model of opinion change suggested that
for any given issue, participants' beliefs at the end of the experiment depended on three independent factors: treatment condition, an interaction between prior opinion and strength of conviction, and an interaction between argument polarity and argument rating. Whereas a motivated reasoning account predicts that treatment condition will interact with prior opinion, we instead found that condition had an independent effect. Participants who read Pro arguments adjusted their beliefs in a positive direction, those who read Con arguments adjusted their beliefs in a negative direction, and those in the Mix condition made almost no adjustment to their beliefs.

Further, while prior opinion was highly relevant for belief change, we found no evidence for the polarization phenomenon predicted by motivated reasoning. In fact, participants with weaker convictions moved a small amount away from their original positions, while those with strong convictions tended to maintain their existing beliefs.

Finally, the relationship between argument ratings and belief change was more consistent with a Bayesian account than the biased assimilation process predicted by motivated reasoning. That is, with motivated reasoning we would expect both highly-rated congruent arguments and low-rated incongruent ones to lead to opinion change in the direction of participants' prior opinions. Instead, we saw that highly-rated arguments, regardless of their congruency with participants' prior beliefs, were associated with movement in the direction of the arguments themselves. This is strong evidence in favor of a Bayesian account and shows that even in the presence of biased argument rating, belief change seems to be based on the quality of the evidence itself.

## Acknowledgments

This research was supported by a grant from the Frontiers of Innovation Scholars Program (FISP) at UC San Diego.

## References

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). lme4: Linear mixed-effects models using eigen and s4. $R$ package version, 1(7).
Ditto, P. H., \& Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. Journal of Personality and Social Psychology, 63(4), 568.
Edwards, K., \& Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. Journal of Personality and Social Psychology, 71(1), 5.
Gerber, A., \& Green, D. (1999). Misperceptions about perceptual bias. Annual review of political science, 2(1), 189210.

Jern, A., Chang, K.-M. K., \& Kemp, C. (2014). Belief polarization is not always irrational. Psychological review, 121(2), 206.
Klayman, J., \& Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological review, 94(2), 211.

Kunda, Z. (1990). The case for motivated reasoning. Psychological bulletin, 108(3), 480.
Lord, C. G., Ross, L., \& Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of personality and social psychology, 37(11), 2098.
Nisbett, R. E., \& Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, NJ: Prentice-Hall.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. Oxford University Press.
Pyszczynski, T., \& Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. Advances in experimental social psychology, 20, 297-340.
R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.
Taber, C. S., Cann, D., \& Kucsova, S. (2009). The motivated processing of political arguments. Political Behavior, 31(2), 137-155.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279-1285.

# The impact of the Digital Age in Moral Judgments 

Albert Barque-Duran ${ }^{1}$, Emmanuel M. Pothos ${ }^{1}$, James M. Yearsley ${ }^{2}$ \& James A. Hampton ${ }^{1}$<br>${ }^{1}$ Department of Psychology, City, University of London, London EC1V 0HB, UK.<br>${ }^{2}$ Department of Psychology, Vanderbilt University, Nashville, TN 37240, USA.<br>albert.barque-duran@city.ac.uk


#### Abstract

Nowadays, several of the situations in which we have to make decisions are in digital form. In a first experiment ( $\mathrm{N}=1010$ ) we showed that people's moral judgments depend on the Digital Context (Smartphone vs. PC) in which a dilemma is presented, becoming more utilitarian (vs. deontological) when using Smartphones. To provide additional evidence, we ran a second $(\mathrm{N}=250)$ and a third experiment $(\mathrm{N}=300)$, where we introduced time constraints and we manipulated time instructions. Our results provide an extended perspective on Dual-Process Models of Moral Judgment, as we showed that the use of smartphones, often assumed to be hurried which would be consistent with gut-feeling decision-making, increased the likelihood of utilitarian responses and decreased deontological ones. This is the first study to look at the impact of the digital age on moral judgments and the results presented have consequences for understanding moral choice in our increasingly virtualized world.


Keywords: Moral Judgment; Behavioural Ethics; DecisionMaking, Human-Computer Interaction.

## Introduction

In this digital age, we spend a lot of time interacting with computer screens, smartphones and other digital gadgets. We buy online, work on the cloud, our social relationships are sometimes online-based, etc. Thus, the contexts where we typically face ethical decisions and are asked to engage in moral behaviour have changed. Nowadays, moral dilemmas are often presented digitally, that is, relevant information is presented through and decisions are made on a technological device.

A key distinction regarding moral judgments concerns deontological versus utilitarian decisions (Singer, 1991). Recent dual-process accounts of moral judgment contrast deontological judgments, which are generally driven by automatic/unreflective/intuitive responses, prompted by the emotional content of a given dilemma, with utilitarian responses, which are the result of unemotional/rational/controlled reflection, driven by conscious evaluation of the potential outcomes (Greene et al., 2001; Greene \& Haidt, 2002). In this account, an individual's ethical mind-set (rule-based vs. outcome-based, Barque-Duran et al., 2015) can play a central role. A deontological perspective evaluates an act based on its conformity to a moral norm (Kant, 1785/1959) or perhaps just a rule (such a law). By contrast a consequentialist/utilitarian perspective evaluates an act depending on its consequences (Mill, 1861/1998).

People often believe that judgments about "right" and "wrong" should be consistent and unaffected by irrelevant aspects of a moral dilemma or by its context. However, studies have shown, for example, that manipulations of the language (foreign vs. mother tongue) in which a moral scenario is presented can affect moral judgments through increasing psychological distance from the situation, and so inducing utilitarianism (Costa et al, 2014). The choice of deontological versus utilitarian judgments can vary depending on the emotional reactivity triggered by the dilemma (Valdesolo \& DeSteno, 2006). As such, establishing which conditions favor each of these two influences is fundamental to understanding the psychology of moral choice.

Construal Level Theory (CLT) provides a framework of considerable potential relevance by linking mental representations to moral judgment. Individuals' judgments, decisions, and behaviours can differ as a function of construal levels. CLT proposes that the same event or object can be represented at multiple levels of abstraction (see Trope \& Liberman, 2010, for a review). More weight is given to global, abstract features at high-level construal, whereas local, concrete features are more influential at lowlevel construal. According to CLT, psychological distance is a major determinant of what level of construal is activated. Distancing a target on any dimension of psychological distance (i.e., time, space, social, and hypotheticality) leads to greater activation of high-level construal (directing attention to end states) than low-level construal. Crucially, high-level construal is often assumed to align with more utilitarian decision-making (Gong, Iliev, \& Sachdeva, 2012; Aguilar, Brussino, \& Fernández-Dols, 2013).

The present study explores whether a Digital Context (i.e. using a different digital device such a Smartphone or a PC, as hundreds of millions of individuals do every day) can have a systematic impact on these processes. Could Digital Contexts induce different construal levels (through psychological distance)?

There is evidence that people experience a so called "narrowing effect" when using smartphones in decisionmaking, which means that they channel or tunnel their focus toward a main task and ignore or filter out certain cues (Ariely, 2016). A narrowing effect is consistent with the idea that devices such as smartphones would increase psychological distance giving rise to an abstract representation of actions. In other words, the narrowing effect would seem to be aligned with a more utilitarian/ outcome-based mind-set, instead of a more emotional/
deontological one. For this reason we asked ourselves whether Digital Context, smartphone vs. PC, might influence the relation between different levels of construal (psychological distance), thus affecting the likelihood of utilitarian vs. deontological judgments.

To summarize, we hypothesize that Smartphones (vs. PCs) have the effect of channeling or tunneling the focus toward a main task at the expense of certain cues. This should induce high construal, increase psychological distance and give rise to an abstract representation of actions, thus biasing towards more utilitarian judgments.
We first tested this prediction using three versions of the well-known Trolley Problem (Switch, Fat Man, Balanced; Thomson, 1985; see Methods sections). To provide additional support we also ran a second and a third experiment where we introduced a Time Constraint (10 seconds vs. Unlimited Time to respond) and where we manipulated Time Instruction, relating to how participants were given information about the time constraints for reaching a decision (Instructing Unlimited Time vs. No Time Instruction).

## Experiment 1

The objective was to explore whether a manipulation of the Digital Context (Smartphone vs. PC) can have an impact on moral judgment. Specifically, we wanted to test if making moral judgments using a Smartphone increased the number of utilitarian responses in comparison to when using a PC.

## Participants

A total of 1010 participants, all US residents, were recruited on-line and received $\$ 1$ for doing the task ( 482 women, 528 men; mean age $=31.7$ years, $S D=9.6$ ). Sample sizes were based on extant research (Suter \& Hertwig, 2011).

## Materials and Procedure

The study was designed in Qualtrics, run on Amazon Mechanical Turk and lasted approximately 10-15 minutes. Digital Context (Smartphone vs. PC) ${ }^{1}$ and Version of the Trolley Problem (Switch vs. Fat Man vs. Balanced) were manipulated between participants. We used the frequency of Utilitarian vs. Deontological Responses as the dependent measure.

Participants were randomly told to switch to a Smartphone or a PC after reading and agreeing the general instructions on Amazon Mechanical Turk. Having a smartphone was a pre-requisite to participate in the experiment. Participants in the Smartphone condition had to respond to all questions from their smartphone devices. As a manipulation check for this condition, we tracked and verified through Qualtrics that the responses were indeed

[^276]made from an iPhone, Android, Windows Phone or Blackberry. Participants were randomly allocated to one of these six conditions: (1) Smartphone/Switch; (2) Smartphone/Fat Man; (3) Smartphone/Balanced; (4) PC/Switch; (5) PC/Fat Man; (6) PC/Balanced. One third of the participants (327 Participants) on each Digital condition were presented with the Fat Man version of the Trolley dilemma, where one imagines standing on a footbridge overlooking a train track. A small incoming train is about to kill five people and the only way to stop it is to push a heavy man off the footbridge in front of the train. This will kill him, but save the five people. A utilitarian analysis dictates sacrificing one to save five; but this would violate the moral prohibition against killing. Imagining physically pushing the man is emotionally difficult and therefore people typically avoid this choice (Thomson, 1985). According to our hypotheses, participants would be more likely to opt for sacrificing one man to save five when dealing with such moral dilemma using a smartphone in comparison to a PC, since this would induce high construal, increase psychological distance and give rise to an abstract representation of actions, which is aligned with more utilitarian judgments under time pressure; or would induce psychological closeness due to the link between low-level construal and a focus on means, which is also hypothesized to align with more utilitarian judgments, under conditions of no time pressure. Another third of participants (313 Participants) were presented with the Switch dilemma, where the trolley is headed towards the five men, but you can switch it with a lever to another track, where it would kill only one man. People are more willing to sacrifice the one man by pulling the switch than by pushing him off the footbridge and the extensively supported explanation is that pulling the switch is less emotionally aversive. The last third of participants (314 Participants) were presented with the Balanced version of the Trolley Problem. The Balanced dilemma had a setting similar to that in the Fat Man version, but with a different number of people one could save (15 instead of 5), so that utilitarian choice would increase. All participants first completed a filler task (10 trivia questions) before responding to one of the versions of the Trolley Problem. A "catch question" was introduced in the experiment, to control for attention during the task (i.e. "If you are paying attention to this question please select answer ' 36 ' from the options below"). Then, participants were presented with one of the three moral scenarios (Switch, Fat Man or Balanced) where they had to choose between Choice A (utilitarian) or Choice B (deontological). In all cases the dilemma was presented with both text and an illustration. Subsequently, participants completed another filler task (10 trivia questions). Finally, participants were asked to complete The Big Five Inventory (John et al., 1991) questionnaire, which is considered a quick (44-items), reliable, and accurate measure of the five dimensions of personality. We considered that the impact of digital content on moral choice could also interact with personality characteristics (Ozer \& Benet-Martínez, 2006) but the
results did not lead to firm conclusions and therefore will not be reported further.

## Results Experiment 1

We excluded participants whose first language was not English, as Costa et al., (2014) showed that the use of a foreign language (instead of a mother tongue) in a moral scenario increases psychological distance and induces utilitarianism when making moral judgments. We also excluded those participants who did not answer the catch question correctly. A total of 56 participants out of 1010 were thus excluded.

We first compared the percentage of Utilitarian Responses for the two Digital Contexts (Smartphone vs. PC) on each of the three Versions of the Trolley Problem that were employed (Switch vs. Fat Man vs. Balanced; Figure 1).


Figure 1: A) The experimental paradigm used in the Smartphone condition in Experiment 1. B) The illustrations used in each of the three moral conditions. C) Percentage of Utilitarian Responses for both Digital Contexts (Smartphone vs. PC) on each of the three versions of the Trolley problem
(Switch vs. Fat Man vs. Balanced). Error bars represent standard errors.

As expected, in the Fat Man dilemma more participants avoided the act of pushing the heavy man off the footbridge in front of the train, presumably because of the emotional burden of this choice. More importantly, participants were more likely to opt for sacrificing the Fat Man (utilitarian response) to save five men when using a Smartphone (33.5\%) than when using a PC ( $22.3 \%$ ). A $2 \times 2$ chi-square test of independence was performed to examine the frequency of Utilitarian vs. Deontological Responses against Digital Context in the Fat Man condition and this
revealed a significant association between the variables, $\chi 2$ $(1, \mathrm{~N}=327)=5.15, \mathrm{p}=.023$. This result supports our hypothesis that moral judgments in Smartphones increase utilitarian decision-making, than when using a PC.

We then analyzed the frequency of Utilitarian vs. Deontological Responses, across the two Digital Contexts, in the Switch condition. Slightly more participants decided to sacrifice one man by pulling the switch than to do nothing and let five people die ( $80.9 \%$ for the Smartphone users; $76.9 \%$ for the PC users), but there was no evidence for an association between the two variables, $\chi 2$ ( $1, \mathrm{~N}=313$ ) $=.741, \mathrm{p}=.389$. This result supports our expectation that in less emotional scenarios, such as the Switch dilemma, there is a reduced effect of Digital Context. That is, there is no difference in participants' moral judgments when using a Smartphone or a PC if the moral scenario is already highly utilitarian.

Finally, we examined the frequency of Utilitarian vs. Deontological Responses in the Balanced condition. Note, this condition was designed so that, in the PC condition at least, there would be fairly equivalent utilitarian and deontological influences, and this was approximately the case. Regarding the manipulation of interest, $40.4 \%$ of participants decided to push the heavy man off the footbridge in the PC and $36.7 \%$ in the Smartphone conditions. Nevertheless, a chi-square test of independence showed that the relation between these variables was not significant, $\chi 2(1, \mathrm{~N}=314)=.448, \mathrm{p}=.503$. The (tentative) conclusion from this experiment is that using a Smartphone ${ }^{2}$ rather than a PC has a reliable impact on moral judgments only when dilemmas or scenarios have high emotional content.

## Experiment 2a and 2b

The objective of Experiment 2a was to provide additional evidence for the increased number of utilitarian responses using a Smartphone by manipulating the amount of time available to form a moral judgment. We wanted to explore Digital Context (Smartphone vs. PC) and Time Constraint (10 seconds vs. Unlimited time to respond) on moral judgments. It is possible that the effect of Digital Context is independent from that of Time Constraint, in which case we cannot explain the former in terms of (just) the latter. Alternatively, Time Constraint may provide a bias on moral decision making opposite to the effect of Digital Context (e.g., a decrease of utilitarian responses, in the fat man scenario, when participants are using a Smartphone), which will create a complex picture regarding how using Smartphones in everyday moral judgments biases for and against utilitarian responses. In Experiment 2b, we addressed the challenge to explain the difference in the Fat Man condition of Experiment 1 and in the Unlimited Time

[^277]condition in Experiment 2a (where the effect of Digital Context had disappeared) by manipulating directly the Time Instruction to either specify that there was unlimited time available for a moral judgment, or not mentioning time at all (Instructing Unlimited Time vs. No Time Instruction). The key difference between these two conditions was that in Experiment 1 participants were not told anything regarding time, while in Experiment 2a, in the equivalent conditions, participants were specifically told they had unlimited time. We also measured participants' affective reaction with the Self Assessment Manikin test (Bradley and Lang, 1994).

## Participants

A total of 550 participants ( $250 \operatorname{Exp} 2 \mathrm{a}$ and $300 \operatorname{Exp} 2 b$ ), all of whom were US residents, were recruited on-line and received $\$ 0.80$ for doing the task ( 234 women, 316 men; mean age $=32.5$ years, $S D=9$ ).

## Materials and Procedure

The studies were designed in Qualtrics, run on Amazon Mechanical Turk and lasted less than 10 minutes. Digital Context (Smartphone vs. PC), Version of the Trolley Problem (Switch vs. Fat Man) and Time Constraint (10 seconds vs. Unlimited Time to respond) were manipulated between participants in Experiment 2a. There were therefore eight conditions. We used the frequency of Utilitarian vs. Deontological Responses as the dependent measure.

All participants followed a similar procedure as in Experiment 1. They first completed a filler task (10 trivia questions) including a catch question, as in Experiment 1. Then, participants were presented with one of the two moral scenarios (Switch or Fat Man). In all cases the dilemma was presented with both text and an illustration. Participants were alerted of the available time for responding depending on their condition (i.e. "You will only have 10 seconds to answer the question in the next screen" vs. "You will have unlimited time to answer the question in the next screen"). After the presentation of the scenario, in the " 10 seconds" condition participants had to choose between Choice A (utilitarian) or Choice B (deontological), while a countdown timer appeared at the top of their screen (both Smartphone and PC). In contrast, in the "Unlimited Time" condition, participants were explicitly told that they had to make their judgment taking as much time as they wanted. Finally, participants were asked to complete the Self Assessment Manikin test (Bradley and Lang, 1994), which is a technique that directly measures the pleasure, arousal and dominance associated with a person's affective reaction.

In Experiment 2b, Digital Context (Smartphone vs. PC) and Time Instruction (Instructing Unlimited Time vs. No Time Instruction) were manipulated between participants, using the Fat Man scenario. Time Instruction was manipulated in the following way. Half the participants were given the instructions as in the Experiment 2 a Unlimited Time condition. The other half did not have any
indication of the time they had to spend making their judgment (same procedure as in Experiment 1).

## Results across all Experiments 1, 2a and 2b

In this section we report the results of Experiment 2a, $2 b$ and then bring together the results from all experiments focusing on the Fat Man scenario (Figure 2).

First, we summarize the results from Experiment 2 a . We excluded a total of 10 participants out of 250 following the same criteria as in Experiment 1. As a manipulation check, we first examined the amount of time that participants took to finish the experiment ( 5 min 10 s in the Unlimited Time condition; 4 min 32 s in the 10 s condition).

We examined the differences in the percentage of Utilitarian Responses for the two Digital Contexts (Smartphone vs. PC) on each of the two versions of the Trolley Problem (Switch vs. Fat Man) and with or without time pressure ( 10 s vs. Unlimited Time).

In the time pressure (10s), Switch condition, slightly more participants decided to sacrifice one man by pulling the switch than to do nothing and let five people die, when using a Smartphone (79.31\%) than when using a PC ( $66.67 \%$ ), but this difference was not reliable, $\chi 2(1, N=65)$ $=1.282, \mathrm{p}=.257$.

Regarding the Unlimited Time condition, in the Switch condition, Digital Context also did not appear to play a role in moral judgments ( $85.71 \%$ and $83.87 \%$ for Smartphone and PC, respectively); regardless of Digital Context, we observed highly utilitarian responses. Thus, as before, the results in the Switch dilemma indicate that Digital Context and (as it seems) Time Constraint have a reliable impact on moral judgments only when dilemmas or scenarios have high emotional content. This result also supports our assumption that in less emotional scenarios, such as the Switch dilemma, any effect of either Digital Context or Time Constraint does not result in a reliable increase in utilitarian responding.

In the time pressure (10s), Fat Man condition, participants were more likely to opt for sacrificing the Fat Man (utilitarian response) to save five when using a Smartphone $(45.7 \%)$ than when using a PC (20.0\%), $\chi^{2}(1, \mathrm{~N}=60)=$ $4.239, \mathrm{p}=.04$. At face value, these results challenge the assumption that hurried responses necessarily lead to deontological moral judgments.

Then, we examined participant's responses in the Unlimited Time, Fat Man condition. The results here appear to conflict with our conclusion from Experiment 1, in that there was no difference in Utilitarian vs. Deontological responses, between the Smartphone and PC conditions ( $27.58 \%$ and $29.63 \%$, respectively, $\chi 2(1, \mathrm{~N}=64)=2.224$, $\mathrm{p}=.136$ ). In other words, when participants were specifically told to spend unlimited time to resolve the dilemma (Unlimited Time condition), the Digital Context effect vanished. We return to this finding in Experiment 2b.

We also considered whether the impact of Digital Content on moral choice could interact with the perceived emotionality of the scenario/context or affective reactions,
but the results did not lead us to firm conclusions and therefore will not be reported further.

Second, we summarize the results from Experiment 2b. In this experiment we excluded a total of 141 participants out of 300 following the same criteria as in Experiment 1 and 2a. One participant was rejected because she/he answered incorrectly to the catch question and one because English was not his/her first language. Additionally, 139 participants were eliminated because they said they had come across a moral choice in the context of the Trolley Problem before. The pattern of results does not change qualitatively if these participants are included, but we decided not to do so.

In this experiment we measured Response Time for the particular moral judgment, though we note that, as the experiment was run over the internet, the accuracy of these measurements is lower than in the lab. Did participants in the Instructing Unlimited Time condition take longer to respond than ones in the No Time Instruction one? There was no evidence that this was the case ( $2 \times 2$ ANOVA with Digital Context and Time Instruction, $\mathrm{F}<1$ for all effects). We suggest that the effects from Time Constraint and Time Instruction seen in Experiments 2a, 2b could result in a change of the participants' mind-set and approach to the problems, without corresponding clear differences in Response Time.


Figure 2: Summary of the relevant results from Experiments $1,2 \mathrm{a}$ and 2 b for the Fat Man problem. The vertical axis shows percentage of utilitarian responses and the horizontal axis the conditions of interest. Error bars represent standard errors.

The two leftmost bar clusters in Figure 2 show the results of Experiment 2b. Interestingly, using the data from Experiment 2b, we replicated the finding from Experiment 2a, that the mere fact of "nudging" participants to use unlimited time resulted in utilitarian responses that were not influenced by Digital Context. A $2 \times 2$ chi-square test with frequency of Utilitarian vs. Deontological Responses against Time Instruction (Instructing Unlimited Time vs. No Time Instruction) confirmed this conclusion, $\chi 2(1)=5.509$, $\mathrm{p}=.018$.

We next considered whether the results from Experiments 2b replicated the effect from Experiments 1 and 2a regarding Digital Context. The pattern of results from the

No Time Instruction condition in Experiment 2b closely matched the corresponding results in Experiment 1. In Experiment 2b, as expected, participants were more likely to opt for sacrificing the Fat Man (utilitarian response) to save five when using a Smartphone (28.6\%) than when using a PC ( $19 \%$ ). Even though the trend was as expected, a 2 x 2 chi-square test with frequency of Utilitarian vs. Deontological Responses against Digital Context (Smartphone vs. PC) was not significant, $\chi 2(1, \mathrm{~N}=70)=$ $0.864, \mathrm{p}=.35$. However, after collapsing the data (for the identical Fat Man, No Time Instruction conditions) from Experiments 1 and 2b, we obtained a significant association between frequency of Utilitarian vs. Deontological Responses and Digital Context (Smartphone vs. PC), $\chi 2$ (1, $\mathrm{N}=397)=6.27, \mathrm{p}=.012$. This result supports our hypothesis that moral judgments in Smartphones increase utilitarian decision-making, compared to when using a PC, when no information about time is provided.

Importantly, the results from Experiments 1, 2a and 2b put together indicate that under conditions of no time information and time pressure there is indeed a utilitarian bias. The only Time Instruction in which the utilitarian bias was eliminated was the Unlimited Time condition, in which participants were specifically told to take as long as they needed to respond. This finding has a plausible interpretation that, in the Unlimited Time condition, participants took into account the information they have been ignoring so far (which would include emotional cues) and this made the utilitarian bias disappear. Thus, the results so far support the hypothesis that, under most conditions, smartphones (vs. PC) are associated with more utilitarian decision-making (vs. deontological). An additional interesting finding is that utilitarian judgments emerge in both the No Time Instruction condition and the Time Pressure condition.

## Discussion

This is the first study to look at the impact of digital context in moral judgments. We considered whether the increasing tendency for our judgments to be mediated through the use of technological gadgets might be changing our approach to moral dilemmas. We have shown that people's moral judgments become more utilitarian (vs. deontological) when using Smartphones as opposed to PCs, under a variety of time-related manipulations (but not all). The present work was motivated by the idea Digital Context might impact the relation between different levels of construal (psychological distance) thus affecting utilitarian vs. deontological judgments. While our results are consistent with such a view, clearly further research is needed.

We first consider the implications of these results for the Dual-Process Models of Moral Judgment (Greene et al., 2001; Greene \& Haidt, 2002). A standard assumption is that moral dilemmas resolved in fast, gut-feeling conditions engage a deontological mode of responding, while utilitarian responses are typically the result of longer consideration and involve cognitive control. Instead, we
showed that participants under time pressure were more likely to opt for sacrificing the "fat man" to "save five" (utilitarian response) when using a Smartphone than when using a PC. That is, some digital contexts (i.e. Smartphones) can trigger utilitarian decision-making under time pressure, even though time pressure has traditionally been associated with deontological responding in moral choice. Dual route models have received extensive support and no doubt they are valid under most circumstances. Our results indicate a need to perhaps augment the available routes for utilitarian biases in such models.

Other research has provided a more complex picture regarding the impact of time on deontological vs. utilitarian judgments. Specifically, Suter and Hertwig (2011) showed that participants in a time-pressure condition (associated with fast, gut-feeling conditions), relative to a no-timepressure condition (associated with longer consideration and higher cognitive control), were more likely to give deontological responses only in high-conflict dilemmas. By contrast, in low-conflict and in impersonal dilemmas, the proportion of deontological responses did not differ between conditions. The results from the present experiments partly support these differences between high-low conflict dilemmas. In less emotional scenarios (Switch), neither Digital Context nor Time Constraint resulted in a reliable increase in utilitarian responding. By contrast, in more emotional scenarios (Fat Man), our results question the well-established assumption (from Suter \& Hertwig, 2011, amongst others) that hurried decisions enhance deontology, since we showed that moral judgments under a time constraint and in a specific Digital context (Smartphones) seem to make utilitarian judgments more common.

Clearly, more work is required to disentangle possible explanations for the exact effect of the different instructions concerning timing, especially regarding the possibility that keeping track of time may result in reduced cognitive resources. But the crucial point regarding the present study is that our conclusion considering Digital Context and moral judgments appears mostly independent of such considerations.

Our hypotheses regarding Digital Context and moral decision-making was largely motivated from the effects and implications from Construal Level Theory. According to CLT, psychological distance can vary on at least four dimensions: temporal, spatial, social and hypotheticality (i.e. probability for a scenario to become reality; Trope \& Liberman, 2010). Can we localize the particular effect of distance in considering responding using a smartphone vs. a PC? In further studies we will attempt to measure psychological distance directly. More generally, our results were inconclusive regarding the idea that the psychological distance elicited by a smartphone decreased the intensity of people's affective reactions. It is possible that smartphones induce a greater distance in other respects. For example, it might be the case that the use of digital devices interacts with/mediates the hypotheticality dimension.

Overall, the present work reveals a need for the further systematic study of how Digital Context affects moral choice, all the more so given that, increasingly, governments, charities and other institutions engage in intense campaigns over digital media to encourage moral choices for important aspects of our way of life.

## Acknowledgements

This work was supported by grants AFOSR FA8655-13-1-
3044 and Leverhulme Trust grant RPG-2015-311 and H2020-MSCA-IF-2015 grant 696331.

## References

Aguilar, P., Brussino, S., Fernandez-Dols, JM. (2013). Psychological distance increases uncompromising consequentialism. Journal of Experimental Psychology. 449-452.
Ariely, D. (2016). Time Pressure: Behavioral Science Considerations for Mobile Marketing. Think with Google.
Bradley MM., Lang PJ. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. J Behav Ther Exp Psychiatry. 25(1): 49-59.
Costa A, Foucart A, Hayakawa S, Aparici M, Apesteguia J, et al. (2014) Your Morals Depend on Language. PLoS ONE 9(4): e94842.
Barque-Duran et al., (2015) Patterns and Evolution of Moral Behavior: Moral Dynamics in Everyday Life. Thinking and Reasoning. 22, 31-56.
Gong, H., Iliev, R., Sachdeva, S. (2012). Consequences are far away: Psychological distance affects modes of moral decision making. Cognition.
Greene, J. D., Sommerville, R., Nystrom, L., Darley, J., \& Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 2105-2108.
Greene J, Haidt J. (2002). How (and where) does moral judgment work? Trends Cognitive Science 6: 517523.
Kant, I. Foundation of the metaphysics of morals. Beck, LW., translator. Indianapolis: Bobbs-Merrill; 17851959.
Mill, JS. Utilitarianism. Crisp, R., editor. New York: Oxford University Press; 18611998.
Ozer, D. J., \& Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. Annual Review of Psychology, 57, 401-421.
Singer, P. (1991). A companion to ethics. Oxford, England: Blackwell Reference.
Suter, R. \& Hertwig, R. (2011). Time and moral judgment. Cognition. 119 (2011) 454-458.
Thomson, J. (1985). The trolley problem. Yale Law, 94.
Trope, Y., \& Liberman, N. (2010). Construal Level Theory of Psychological Distance, Psychological Review, 117 (April), 440-63.
Valdesolo, P. \& DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. Psychological Science, 17, 476-477.

# Exploring Functions of Working Memory Related to Fluid Intelligence: Coordination and Relational Integration 

Joel E. Bateman (jbat2136@uni.sydney.edu.au) ${ }^{1}$<br>Damian P. Birney (damian.birney@sydney.edu.au) ${ }^{1}$<br>Vanessa Loh (vanessa.loh@sydney.edu.au) ${ }^{1}$<br>${ }^{1}$ School of Psychology, University of Sydney<br>Sydney, NSW, 2006, Australia


#### Abstract

Two hypothesized functions of working memory coordination (ability to maintain unrelated storage loads during processing) and integration (ability to integrate multiple elements into a relation) - were explored and compared to fluid intelligence. In Experiment 1, 130 participants completed a modified Latin-Square Task (LST) which experimentally added or reduced storage load. Results suggested that pure integration (with no storage load) could predict Gf, but no difference was found between coordination and integration. Experiment 2 employed the Arithmetic Chain Task (ACT), again with modifications to storage load. Results support replication of LST findings, though a distinction was found between coordination and integration when storage material could not be easily rehearsed. Findings from both experiments support a distinction between coordination and integration tasks in understanding the WM-Gf association.


Keywords: working memory; fluid intelligence; relational integration

## Introduction

Working memory ( $W M$ ) has consistently been linked to fluid intelligence $(G f)$, yet the intricacies underlying this relationship are not fully understood. This is in part because neither $W M$ nor $G f$ reflect a single cognitive process. Rather, $W M$ is a complex system responsible for processing and maintaining information, attention, and multi-tasking. Gf is similarly multi-dimensional, variously reflecting reasoning and the capacity to deal with novelty. Many $W M$ tasks (such as complex-span tasks; CSPANs) draw on coordination, in that information from one aspect of the task must be maintained in storage while performing a simultaneous but unrelated processing task. Conversely, many prototypical Gf tasks require relational integration (henceforth 'integration'; Halford, Wilson, \& Philips, 1998). Integration entails the ability to combine multiple representations and is critical to reasoning. In this report, we argue that advances in understanding the $W M-G f$ link have been slowed by the overrepresentation of coordination in $W M$ tasks, and the failure to consider integration as a component of $W M$.

We aim to redress this using two experiments. In Experiment 1, we modify the integration-based Latin-Square Task (LST; Birney, Halford, \& Andrews, 2006) by adding or removing storage load. In Experiment 2, we investigate the same processes in an arithmetic task (Oberauer, Demmrich, Mayr, \& Kliegl, 2001). We begin with an exposition of coordination and integration.

## Two Functions of Working Memory

Coordination $W M$ tasks typically involve some combination of processing and storage, reflecting the two components of $W M$. Coordination can be defined as the ability to coordinate stored elements with unrelated processing. CSPANs, such as the operation span, are examples of coordination tasks, because storage capacity is the primary outcome (e.g., the number of words that can be recalled) and processing (e.g., verifying the veracity of a math operation) is included to fulfil the simultaneous processing-storage conceptualization of WM (Baddeley \& Hitch, 1974). While a failure of either component does not represent a failure of the other, if both components are not given equal priority, the extent that CSPAN measures $W M$ is brought into question. However, even when processing is ensured (e.g., with an $85 \%$ threshold for operation verifications), the only measure indexing $W M$ is the recall. Because the processing is somewhat trivialized, it tells us little about how processing ability influences performance on $W M$ tasks. This does not, however, take away from the fact that coordination tasks are excellent for linking $W M$ and $G f$ (Ackerman, Beier, \& Boyle, 2005).

Integration Process-oriented accounts of $W M$ have led to the development of tasks that measure the ability to integrate representations into higher-order relational structures (Oberauer Süß, Wilhelm, and Wittman, 2008). All processing subtasks typically require some form of integration (e.g., integrating two digits to derive a sum), though some researchers have attempted to provide formalized accounts of processing tasks. Oberauer et al. (2008) employed the finding squares task, where participants monitored a 10 x 10 grid filled with 10 dots. Every few seconds, some dots would change position. The task was to monitor the dots and respond if a collective set of dots formed a square. Although tasks such as these have no storage requirements, they are still good predictors of $G f$ (Oberauer et al., 2008). This has led to the suggestion that integration forms the core of $W M$ (Oberauer, Süß, Wilhelm, \& Sander, 2007); and that rather than a 'storage capacity' limiting $W M$, constraints are instead dictated by the strength of bindings between integrated representations.

Halford and colleagues provide an alternate processoriented account of $W M$ limitations in terms of relational complexity (Halford et al., 1998), that formalizes individual
capacity for integration. $W M$ is framed not as a limitation in the number of elements, but by the complexity of relations between elements to-be-integrated. Complexity metrics have been shown to capture constraints in processing capacity (Birney et al., 2006).

Aims There is evidence that both coordination and integration can be implicated in the $W M-G f$ association. However, it is difficult to directly compare these functions, as they are typically operationalised in different tasks. The current research aims to compare coordination and integration within single tasks, experimentally.

## Experiment 1

One way to explore coordination and integration within a task is to consider a variant of a typical processing task, with reduced or additional storage load requirements. If the reduced storage load condition still associates with $G f$, it would suggest a role of pure integration within the $W M-G f$ link, as only the processing remains. Conversely, if the additional storage load condition associates with $G f$, it would suggest a role of coordination. This is the approach we used in Experiment 1, employing the LST as a processing task.

The LST was designed following the principles of relational complexity theory (Birney et al., 2006). The LST presents participants with an incomplete matrix of $4 \times 4$ cells with the governing rule that each row and column may contain only one element from a set of 4 elements. Complexity is manipulated by the number of rows and columns that must be considered in order to deduce a target cell (see Figure 1). In some items, participants must also solve interim cells, using information from those cells to solve the target. Thus, although the task is processingfocused, there is some storage costs associated with holding interim cell information. Birney et al. (2006) found that complexity captures $64 \%$ of variability in item difficulty, while the number of interim cells captures $16 \%$. Thus, $80 \%$ of difficulty variance is from identified processing and storage demands.


Figure 1. Example LST items

Because we were employing a variance-partitioning approach, it was possible to more directly compare the coordination condition with integration by varying whether the additional storage was processing-contingent or not. By partitioning out variance associated with the baseline task, we could derive variance associated solely with a coordination load and compare it to variance associated with an integration load, that were equivalent in task format and (potentially) difficulty. Thus, our four conditions for the LST were: basic, reduced storage, additional storage (coordination), and additional storage (integration); which were crossed with the standard manipulations of complexity and steps.

We hypothesized that the reduced storage condition would reduce the difficulty of the task, but maintain the association with $G f$, because pure integration was still required of the baseline task. We also hypothesized that the additional storage conditions would increase the difficulty of the task to similar degrees, and each would represent a unique contribution to predicting $G f$, as they represent the two functions of coordination and integration.

## Method

Participants and Procedure In total 130 first-year students (83 females) at the University of Sydney participated in exchange for course credit. The mean age was 19.04 ( $\mathrm{SD}=$ 1.6) years. Participants were tested in groups in 60 m sessions.

Measures Three LST sets, each with 12 unique items equally distributed across complexity ( $2 / 3 / 4$ ) and steps ( $1 / 2$ ), were adapted from Birney and Bowman (2009). Thus, all sets included an equal distribution of complexity and steps. The basic set consisted of 12 standard items (as in Figure 1).

The dynamic-completion (DC) set consisted of 12 items which allowed participants to insert interim solutions. Instead of simply selecting an answer, participants could place shapes into empty cells of the matrix, before placing a shape into the target cell to indicate their overall response. In this way, participants were able to work through the problem, offloading storage demands associated with interim cells.

The final set was additional load, consisting of 12 items. Participants were randomly allocated to coordination or integration items for this set. The actual items were identical, but the procedures were different.

For coordination items, there was a 5 s memory phase where participants viewed the matrix without the target indicated. During this phase, two shape-filled clue cells were coloured to indicate that they must be remembered. After this phase and a 2000 ms interlude, the typical test phase began with the target indicated. After responding to the item, the recall phase began. In this phase, there was first a five-second downtime with a black screen stating to "recall the cells". After this, a blank probe matrix appeared and the participant had to indicate the shapes and locations associated with the two marked cells (thus requiring coordination of stored elements and unrelated processing).

For integration items, a similar procedure was employed for coordination items except that the two marked cells during the memory phase were removed during the test phase. In this way, if the participants forgot the shapes in these cells, they would not be able to solve the problem (thus requiring integration of stored elements with related processing).

After the LST, participants completed a 20 -item Raven's Advanced Progressive Matrices (Raven, 1941; APM; odd items + items 34 and 36). The use of a single task does define $G f$ narrowly, because we cannot be certain that correlations between the LST and APM are due to an overlap in WM functions or due to task-specific factors, such as modality. However, there is a large comparative literature base to draw on to understand implications of this limitation.

## Results

Difficulty Effects Descriptive results are presented in Table 1. The overall LST-APM correlation was $r=.47, p<.001$, replicating prior work (Birney et al., 2012). Using a repeated measures ANCOVA, a complexity effect was investigated with APM entered as a moderator (covariate). Consistent with prior research (Birney \& Bowman, 2009), complexity was a significant predictor of performance $\left(\mathrm{F}_{2,256}=94.73\right.$, $m s e=0.485, \mathrm{p}<.001$, partial $-\eta^{2}=.425$ ), but APM did not moderate the effect, suggesting increases in complexity does not result in increased demand on $G f$-like resources.

A set (basic/DC/load) by complexity ( $2 / 3 / 4 \mathrm{D}$ ) repeatedmeasures ANOVA was conducted to determine whether set affected performance. There was a significant main effect of set $\left(F_{2,242}=43.17\right.$, mse $=0.56, p<.001$, partial $\left.-\eta^{2}=.26\right)$. As hypothesized, DC items were significantly easier ( $F_{1,121}=$ 35.14, $m s e=0.76, p<.001$ ) and load items were significantly more difficult $\left(F_{1,121}=14.66\right.$, mse $\left.=1.48, p<.001\right)$. A significant set-complexity interaction $\left(F_{4,484}=13.28\right.$, $m s e=$ $0.36, p<.001$, partial- $\eta^{2}=.10$ ) suggests complexity moderates the set effects. Simple-effect analyses suggest the difference between conditions, particularly the DC condition ( DC vs basic x quadratic complexity effect: $F_{1,121}=4.69, p=$ .03), is more pronounced for more complex items. Finally, separate analyses suggest that integration $(M=3.12)$ and coordination ( $\mathrm{M}=3.31$ ) conditions were not significantly different, $F_{1,128}=2.81, p=.10$. Although this was as hypothesized, there was a trend towards integration items being more difficult.

The results of these tests indicate the LST sets were performing as expected. That is, the DC condition was aiding participants and the load conditions were burdening. The next set of analyses sought to test the hypotheses on the links of set to predicting $G f$.

LST-Gf A series of multiple regressions were performed, regressing APM on LST set performance. Our first hypothesis was that DC should maintain the association with $G f$, despite having reduced storage demands.

When basic and DC items were entered together, $14.2 \%$ of variability in APM performance was accounted for $\left(R^{2}=\right.$ $\left..142, \mathrm{~F}_{2,127}=10.49, \mathrm{p}<.001\right)$. DC items explained $8 \%$ unique

Table 1. LST and APM Descriptives

| Scale (Total Scores) | Mean | (SD) | Range |
| :--- | :---: | :---: | :---: |
| LST combined | 31.17 | $(3.42)$ | $17-36$ |
| 2D Items | 11.42 | $(1.02)$ | $7-12$ |
| 3D Items | 10.96 | $(1.14)$ | $8-12$ |
| 4D items | 8.78 | $(2.11)$ | $2-12$ |
| Basic Set | 10.35 | $(1.50)$ | $5-12$ |
| DC Set | 11.17 | $(1.19)$ | $6-12$ |
| Load: Integration (n1 = 65) | 9.37 | $(1.98)$ | $3-12$ |
| Load: Coordination (n2 = 65) | 9.92 | $(1.78)$ | $4-12$ |
| Recall Cell 1 (Coordination, n2 only) | 10.78 | $(1.60)$ | $4-12$ |
| Recall Cell 2 (Coordination, n2 only) | 10.69 | $(1.98)$ | $0-12$ |
| APM | 13.35 | $(3.88)$ | $2-20$ |

variance $\left(\beta=.30, s r^{2}=.084, p=.001\right)$, whereas basic items did not significantly account for any additional systematic variance $\left(\beta=.14, s r^{2}=.018, p>.05\right)$. As hypothesized, DC did sustain the link with $G f$; and in fact, captured a larger proportion of variance in APM than basic items.

The final regression aimed to test the hypothesis that additional coordination and integration conditions could provide unique contributions to APM. The basic set was entered first, followed by load (regardless of type), then a load interaction variable distinguishing coordination from integration. Load items did account for a significant proportion of variance in APM performance, $\beta=.39, s r^{2}=$ $.14, p=.001$, over and above basic items, $\beta=.16, s r^{2}=.02$, $p>.05$. However, contrary to hypotheses, the regression lines were not different, $\beta=.10, s r^{2}=.01, p>.05$.

## Discussion

In Experiment 1, we flipped the typical $W M$ operationalisation, which uses recall as a primary task, to have processing as the primary task. Under these conditions, a storage-loaded version of the LST, relative to basic items, predicted a greater proportion of differences in $G f$, providing support for the notion that $W M$ does not have to be restricted to recall as an outcome, or processing as a distractor. The inability to distinguish integration from coordination was unexpected, though clashed with the results of the DC condition, which implicated pure integration alone as the strongest link between $W M$ and $G f$. It is possible the impact of the additional load conditions was confounded by the use of a primary task already highly loaded on integration processes. The burden of performing novel integration may have attenuated differences between the load conditions.

To address this limitation, Experiment 2 employed a different experimental task, the Arithmetic Chain Task (ACT; Oberauer et al., 2001). The ACT requires participants to solve a series of simple equations using mental arithmetic under additional load conditions, while mitigating the potentially high integration present in the LST by having a constant level of complexity. Furthermore, because the ACT is nonvisuospatial, it helps quell criticism that the modality overlap between the LST and APM was the core determinant of correlation. Although arithmetic is a form of integration
$(3+5=$ ? entails establishing the relation, sums-to( $3,5, ?)$; Halford et al., 1998), we argue that completing a chain of simple arithmetic provides a cognitively simpler instantiation of integration than the LST, thus allowing stronger differences to emerge between additional load conditions.

## Experiment 2

Oberauer et al. (2001) provided evidence for a distinction between coordination and integration in the ACT. They asked participants to complete a mental arithmetic task in which participants were shown an equation involving a number of digits, three of which were replaced by symbols (e.g., X, Y, and $Z$ ). In the control condition, participants were given a key showing the numerical values of $X Y Z$ for use in the equation. In the coordination condition, participants were briefly shown three additional numeric values associated with other symbols (A, B, and C). These variables were to be memorized and recalled later, though they were not relevant to the arithmetic. In the integration condition (dubbed 'access') however, $X Y Z$ was equated to $A B C$, necessitating both storage and integration (see Figure 2). The authors found that although the number of stored value mappings had little effect on performance in coordination; in the integration condition, higher levels of storage load produced declines in speed and accuracy. These diverging outcomes indicated the manipulations may have indeed tapped different functions.

In the current study, the ACT entails equations of six operations and seven addends. The format for control, coordination, and access from Oberauer et al. (2001) was used. We also introduced an additional condition, which modified the access condition to include fixed (e.g., $A B C=X Y Z$ ) as well as random (e.g., $A B C=Y Z X)$ mappings. Our complexity analysis (not reported here) suggests that random access imposes constraints on conceptual chunking, increasing the integration load, relative to access-fixed. In summary, the convenience of the serially ordered fixed mappings cannot be applied to random mappings, forcing participants to deconstruct and reconstruct the bindings holding the relation together - a critical source of demand in Oberauer et al.'s (2007) architecture of WM.

In addition to the ACT and APM, we employed the symmetry span as an additional criterion measure. We also aimed to replicate Experiment 1 by including the LST. If the LST-DC is indeed a measure of pure integration, it would provide a useful criterion measure.

Our primary hypothesis was that access and coordination aspects of the ACT should provide independent contributions to predicting APM variance. Further, we hypothesized that access-random should provide the strongest unique contribution, over-and-above other conditions, as it places the highest theoretical demand on a binding-based relational processing system of $W M$.

## Method

Participants and Procedure The participants were 60 firstyear students (44 females) at the University of Sydney who participated for course credit. The mean age was 19.22 (SD $=2.77$ ). Participants were tested in groups in 90 m sessions .

Measures The ACT required participants to solve arithmetic problems of six operations (additions/subtractions). Four blocks of problems (control, coordination, access-fixed, access-random) were generated such that all digits were between 1 and 7 , and final answers, between -9 and +9 . There were six items per block. Participants had practice with all conditions, then received the blocks in random order.

Control items were basic problems that entailed substituting variable-value mappings (e.g., $\mathrm{X}=2, \mathrm{Y}=1, \mathrm{Z}=4$ ) provided in the top half of the screen into equations where each operand was displayed one-at-a-time at a pace controlled by participants. After all 7 operands had been displayed, a textbox would appear prompting the participant for an answer. Feedback was then displayed.

Coordination items were identical to control items, with the exception that participants were given 6 s to memorize three variable-value mappings (e.g., $A=6, B=3, C=1$ ) to be recalled at the end of the trial.

Access-fixed items were similar to coordination items, except the $X Y Z$ variable-value mappings were directly linked to the $A B C$ mappings (e.g., $\mathrm{A}=6, \mathrm{~B}=3, \mathrm{C}=1$; and always, $\mathrm{X}=\mathrm{A}, \mathrm{Y}=\mathrm{B}, \mathrm{Z}=\mathrm{C})$. Again, participants were asked to reproduce the digits corresponding to $A B C$ after the equation had been solved. Thus, unlike the coordination condition, the $A B C$ mappings were required for the arithmetic. Accessrandom items were similar but the $X Y Z$ mappings were randomly linked to the $A B C$ mappings (e.g., $\mathrm{A}=6, \mathrm{~B}=3, \mathrm{C}=1$; and say, $\mathrm{X}=\mathrm{B}, \mathrm{Y}=\mathrm{C}, \mathrm{Z}=\mathrm{A}$ ).

Participants also completed the symmetry span, as in Kane et al. (2004), with set sizes of two to five (two of each). The
Phase 1: Memory

Figure 2. Example of Access condition of the Arithmetic Chain Task (adapted from Oberauer et al., 2001)
score analyzed was total number of recalled squares (0-28). The LST and APM were administered as in Experiment 1.

## Results

Difficulty Effects A repeated-measures ANOVA indicated differences in performance across conditions were significant ( $F_{3,180}=23.99$, mse $=1.51, p<.001$, partial $-\eta^{2}=.29$ ). Control performance ( $M=5.20, S D=1.01$ ) was not significantly different to coordination performance ( $M=5.00, S D=1.34$ ), $t_{59}=1.07, p=.29$. However, control performance was significantly higher than the access conditions on average (Access-fixed: $M=3.98, S D=1.75$; Access-random: $M=$ $3.60, S D=1.89$ ), $t_{59}=7.58, p<.001$. Although in the expected direction, the difference between fixed and random did not reach statistical significance, $t_{59}=1.60, p=.12$.

In summary, the ordering of performance was as expected. Recall was high for all conditions (coordination: 86.67\%, access-fixed: $92.59 \%$ and access-random, $90.37 \%$ ), meeting the criterion of the secondary task in CSPANs, though there was some evidence to suggest recall under conditions where the information was critical (access) is better than when it was irrelevant (coordination).

ACT Correlates The ACT correlated well with the APM, sharing $22 \%$ of variance ( $r=.47$ ). The total ACT-recall component correlated with the CSPAN $(r=.46)$, but was not related to either LST-DC or APM.

In efforts to understand the relationships among the data, step-wise analyses regressing each criterion measure (CSPAN, LST-DC, APM) on ACT were conducted. Results suggest different sets of unique predictors for each criterion in ways as might be expected. For CSPAN, the only ACT predictor accounting for significant variance was coordination recall. For both LST-DC and APM, control and access-random performance were unique predictors. In second models, the criterion measures not being predicted were added, but the results remained unchanged.

In order to fully explicate the ACT-Gf model, a hierarchical regression was conducted, with each condition predicting APM. Model 1, with just control items, predicted $15.3 \%$ of
variance in APM. Contrary to expectations, the coordination predictor did not account for additional unique variance in the second model $\left(\Delta R^{2}=.009, F_{l, 57}=.63, p=.431\right)$. Model 3 with access-fixed also failed to result in a significant change $\left(\Delta R^{2}=.032, \mathrm{~F}_{1,56}=2.19, p=.144\right)$, with control items and shared variance taking the majority of the contribution. However, model 4 with access-random added $6.4 \%$ of unique APM variance predicted - a significant contribution over-and-above all other variables, $\left(\Delta R^{2}=.064, \mathrm{~F}_{1,55}=4.77, p=\right.$ .03).

## Discussion

The findings for LST-DC and APM support the notion that integration is a key component of each of these tasks, drawing both on the control arithmetic (which is basic arithmetical integration) and access-random (which has the highest theoretical integration demands). CSPAN, which we have argued as capturing coordination, was related to recall in the coordination aspect of the ACT.

## General Discussion

The extant literature makes a distinction between coordination and integration functions of $W M$. We adopt a conceptualisation of coordination as the $W M$ function underlying dual-task requirements, where a storage load must be maintained despite ongoing, unrelated processing. This paradigm remains by far the most common used in investigations of the $W M-G f$ link (Ackerman et al., 2005). Process-oriented accounts of $W M$ instead focus on the capacity for integration: combining multiple representations into higher-order relational structures. Integration as a concept has been linked conceptually and empirically to $G f$ (Oberauer et al., 2008). The current work contributes to this research by investigating coordination and integration functions of $W M$ and their relationship to $G f$. A feature of our approach has been to focus on measures where the primary task is processing, rather than recall.

The LST provided mixed results on a distinction between coordination and integration. While additional load overall was incrementally predictive of $G f$, the load effect did not

Table 2. Significant stepwise correlates of cognitive criterion variables.

|  | Model CSPAN |  |  | Model LST-DC |  |  |  | Model APM |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Predictor | $\mathbf{b}$ | $\mathbf{t}$ | $\mathbf{p}$ | $\mathbf{b}$ | $\mathbf{t}$ | $\mathbf{p}$ | $\mathbf{b}$ | $\mathbf{t}$ | $\mathbf{p}$ |
| Control | - | - | - | .39 | 3.23 | .002 | .27 | 2.22 | .030 |
| Coordination | - | - | - | - | - | - | - | - | - |
| Coordination Recall | .42 | 3.54 | .001 | - | - | - | - | - | - |
| Access-Fixed | - | - | - | - | - | - | - | - | - |
| Access-Fixed Recall |  |  |  | - | - | - | - | - | - |
| Access-Random | - | - | - | .24 | 2.03 | .047 | .34 | 2.73 | .008 |
| Access-Random Recall | - | - | - | - | - | - | - | - | - |
| Symmetry Span | $* * *$ | $* * *$ | $* * *$ | - | - | - | - | - | - |
| LST-DC | - | - | - | $* * *$ | $* * *$ | $* * *$ | - | - | - |
| APM | - | - | - | - | - | - | $* * *$ | $* * *$ | $* * *$ |

depend on whether the recalled items were unrelated or related to the LST solution. However, evidence for integration was found in the DC condition, which exceeded expectations as a predictor of $G f$, improving performance while also increasing the association with APM. We replicated this in Experiment 2. This could be explained as a means of 'purifying' the LST into an assessment of raw integration, minimizing the impact of obfuscating storage demands associated with holding interim processing outcomes. DC may be a valuable tool for future use of processing tasks, in order to amplify the effect of integration.

We argued that one issue with the LST was the high integration load present in all manipulations, potentially swamping our additional load conditions by task-specific characteristics. This is especially plausible given the power of DC. The ACT was selected for Experiment 2 because those characteristics are less apparent. The ACT conditions were predictive of the criterion measures consistent with an account for a distinct coordination and integration. However, the coordination link to CSPAN only became apparent when using the recall portion of the ACT, indicating the relationship may have more to do with the outcome measure (i.e., recall) rather than a coordination function per se.

The results of the experiments support a compelling case for differentiating a specific role of integration in $G f$ over-and-above conceptualisations of $W M$ defined by CSPANs. The absence of storage in the LST-DC and other integrationbased tasks (Oberauer et al., 2008) contributes to the notion that storage maintenance is not a pre-requisite for $W M$ to be associated with $G f$, and supports process-oriented accounts of WM (Halford et al., 1998; Oberauer et al., 2007). Further, specific processing limits were alluded to in the results of access-random. That is, consistent with a relational binding approach (Oberauer et al., 2007), the random ordering forced participants to quickly and flexibly deconstruct and reconstruct the variable-value mappings from the way they were first presented into an order consistent with the way they were presented on the screen at the time of the equation. Because only this single condition could indicate binding as an ability, further research is needed to determine what processes contribute to the capacity for relational binding.

One limitation with the current results was that the DC variance could have represented a general task navigation ability (i.e., to apply the advantages of DC), as opposed to pure integration per se. It seems unlikely that such a strong unique effect (equal to $8.4 \%$ of variance in APM) could be attributed solely to DC (as opposed to any other condition), though there is no way to disprove such an explanation with the current data. Because participants could fill as many cells as they wished, we could not distinguish which cells were filled through trial-and-error and which were used as actual planning steps. This task navigation component could be explored using a variant of DC where participants are allowed only a limited number of cells to fill.

Another limitation was the LST and ACT both being integration-based tasks. While we have attempted to reconcile this by holding processing load constant in the

ACT , it is worth considering alternatives for future work. For one, it would be helpful to consider both a storage-based and a processing-based primary task, each with coordination and integration conditions. For instance, an integration version of the operation span could use numbers for the storage component, and these numbers could then be used in the processing component. While this does remove the ability to keep comparisons within a single task, it may at least provide some evidence of a coordination-integration dichotomy not restricted to processing-based tasks.

In conclusion, the current results offer mixed support for a strict coordination-integration functional dichotomy within $W M$. They do, however, provide evidence of a relational integration ability implicated within $G f$ across multiple task formats, with the storage-stripped DC set offering perhaps the strongest support. Further work is needed to determine the extent of integration across tasks; and to determine if coordination can be distinguished from mere recall.

## References

Ackerman, P.L., Beier, M.E., \& Boyle, M.O. (2005). Working memory and intelligence: The same or different constructs? Psychological Bulletin, 131, 30-60.
Baddeley, A.D., \& Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 8, pp. 47-89). New York: Academic Press.
Birney, D.P., \& Bowman, D.B. (2009). An experimental-differential investigation of cognitive complexity. Psychology Science Quarterly, 51, 449-469.
Birney, D.P., Halford, G.S., \& Andrews, G. (2006). Measuring the influence of complexity on relational reasoning: The development of the Latin Square Task. Educational and Psychological Measurement, 66, 146-171.
Halford, G.S., Wilson, W.H., \& Philips,S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences, 21, 803-865.
Kane, M.J., Hambrick, D.Z., Tuholiski, S.W., Wilhelm,O., Payne,T., \& Engle, R.W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. Journal of Experimental Psychology: General, 133, 189-217.
Oberauer, K., Demmrich, A., Mayr, U., \& Kliegl, R. (2001). Dissociating retention and access in working memory: An agecomparative study on mental arithmetic. Memory \& Cognition, 29, 18-33.
Oberauer, K., Süß, H.M., Wilhelm, O., \& Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In R.A. Conway, C. Jarrold, M.J. Kane, A. Miyake, \& J.N. Yowse (Eds.), Variation in Working Memory. New York: Oxford University Press.
Oberauer, K. Süß, H.M., Wilhelm, O., \& Wittman, W.W. (2008). Which working memory functions predict intelligence? Intelligence, 36, 641-652.
Raven, J. C. (1941). Standardisation of progressive matrices. British Journal of Psychology, XIX, 137-150.

# How Order of Label Presentation Impacts Semantic Processing: an ERP Study 

Jelena Batinić (jelenabatinic1@gmail.com)<br>Laboratory for Neurocognition and Applied Cognition, Department of Psychology, Faculty of Philosophy, University of Belgrade, Čika Ljubina 18-20, 11000 Belgrade, Serbia

Andrej Savić (andrej_savic@etf.rs)<br>Tecnalia Serbia Ltd.<br>University of Belgrade, School of Electrical Engineering, Bulevar Kralja Aleksandra 73, 11000 Belgrade, Serbia

Vanja Ković (vanja.kovic@f.bg.ac.rs)<br>Laboratory for Neurocognition and Applied Cognition, Department of Psychology, Faculty of Philosophy, University of Belgrade, Čika Ljubina 18-20, 11000 Belgrade, Serbia


#### Abstract

In this study, we wanted to investigate whether the processing of semantic information is easier when mapping names to pictures or is it the other way around. In order to test this hypothesis, we ran a behavioural and an ERP (Event Related Potential) study, with specific interest in the N400 component as an indicator of semantic processing. We compared three groups of participants who did a match/mismatch task with the only difference being that the labels would appear before, after or simultaneously with the pictures. Not surprisingly, the hardest condition was the one where the two information were presented simultaneously. The amplitude of the N400 was more prominent in the condition where labels were presented after the pictures in comparison to the condition where labels preceded picture presentation, suggesting that this second experimental situation led to smaller violation of expectation for our participants (word to picture condition) in comparison to mapping pictures to words.


Keywords: semantic processing; Event Related Potentials; N400; mental representations; word processing; picture processing

## Introduction

We are in constant interaction with novel and familiar objects on a daily basis. When learning about an object for the first time, we examine its visual characteristics and associate them with its name.

In this study, we wanted to investigate whether the processing of semantic information is easier when mapping names (as more abstract representations) to pictures (as more specific representations) or is it the other way around. Given that the most informative component which is well known to be sensitive to semantic processing/integration is the N 400 component it will be of our primary interest to test whether these mappings elicit differences in the N400 amplitude.

The discovery of the N400 component came in the now classical study of Kutas and Hillyard (1980) in which participants were presented with sentences (one word at a time), that ended with either congruent or
incongruent words. There were two types of incongruent endings: possible but improbable (She drinks tea with salt) or completely semantically unrelated to the previous context (She drinks tea with house). Incongruent words elicited a negative response at around 400 ms from the stimulus onset. Authors concluded that the N 400 is sensitive to context and semantic anomalies (Kutas \& Hillyard, 1980).

Since that study, different authors have reported finding the N400 in a variety of experimental tasks which required semantic processing such as match/mismatch task, semantic priming, word or picture recognition (Anderson \& Holcomb, 1995; Boutonnet, \& Lupyan, 2015; Ganis, Kutas, \& Sereno, 1996; Holcomb \& Anderson, 1993). In general, N400 is most prominent in the central and parietal regions of the scalp (Anderson \& Holcomb, 1995; Kutas \& Federmeier, 2011), but the topography changes depending on the experimental condition. For example, anterior regions are particularly active when processing pictures (Anderson \& Holcomb,1995). The latency of the component is usually in the time window of 200-600 ms from stimulus onset (Kutas \& Federmeier, 2011). The most interesting characteristic of the N 400 is its amplitude, given that it is most responsive to experimental manipulations, whereby a more negative amplitude is elicited by unexpected stimuli which are in turn harder to process (Kutas \& Federmeier, 2009).

## Semantic processing and the N400

Anderson and Holcomb used a semantic priming task in order to investigate the differences in processing of auditory and visually presented words (Anderson \& Holcomb, 1995). Word pairs (prim and target) were presented in the same modality (visual or auditory) using different stimulus onset asynchronies SOA- 0 ms , $200 \mathrm{~ms}, 800 \mathrm{~ms}$. N400 component was found in all of the experimental conditions, but lasted longer when word pairs were presented simultaneously (SOA-0 ms), suggesting that the processing of the two stimuli was
parallel.Apart from that, the highest error rates where in this experimental condition, supporting the hypothesis that it is harder to process two pieces of information at the same time.

Following a different line of research, Boutonnet and Lupyan (2015) where investigating whether visual processing of objects would be easier when they where primed with names or with nonverbal cues. In a match/mismatch task pictures of familiar animals and artifacts were preceded by their names or equally informative nonverbal cues (sound of dog barking preceding a picture of a dog). Participants were more succesful when they were cued with words, and the authors suggest that this is because words denote categories and are better at evoking mental representations which facilitates responding both to match and mismatch trials.

If in fact words evoke more general and abstract mental representations, it would be interesting to see in what way do pictures, that always represent a specific exemplar, can influence the processing of an object's name.

As previously mentioned, words can be treated as more abstract representations and refer to entire categories of objects, while a picture is always representing a single instance of an object and therefore evokes a more narrow and specific mental representation (Ković, Plunkett, \& Westermann, 2009; Ković, Plunkett, \& Westermann, 2010). Given that, to our knowledge, there are no studies directly comparing word to picture versus picture to word processing it remains unclear whether these processes differ, and if they do, which one is easier.

In order to investigate this, we constructed an experiment in which we manipulated the order of label presentation, thereby contrasting three experimental conditions: words preceding pictures, pictures preceding words, and words and pictures presented together. This allowed us to compare the processes of mapping abstract (word) to specific (pictures) representations and specific to abstract represenations by examining the amplitude of the N400 across conditions. Our hypothesis is that the hardest condition for our participants would be simultaneous presentation of words and pictures, given that they have to process two pieces of information at the same time (Anderson \& Holcomb, 1995). Furthermore, we expect that the easiest condition, which would elicit the smallest negative response, would be the pictures to words condition. Since names evoke broad mental representations (Boutonnet \& Lupyan, 2015), any picture shown after the label, no matter how typical of an exemplar it is, would most likely be somewhat different from our evoked mental representation which makes the task harder for the participant to respond. On the other hand, a picture can evoke only one name for a given object which makes the name easier to process when displayed after the picture.

## Method

## Participants

We tested sixty participants, twenty per experimental condition. Participants were psychology students at the University of Belgrade, all native Serbian speakers. They gave informed consent and received course credit for their participation. All participants reported normal or corrected-to-normal vision.

## Stimuli

The study consisted of 120 familiar, everyday objects from different categories such as: mammals, fruits, furniture, tools, clothes, etc. These objects were represented by pictures (original stimuli list taken from Kovic et al., 2009) and their coresponding labels. Labels were presented visually in order to control the duration of stimuli presentation, which wouldn't be possible in the case of auditory presentation. All stimuli were pretested and qualified as highly typical and highly familiar objects. We also conducted a naming task in which 8 participants were asked to name the objects presented in the pictures in order to ensure that there was only one appropriate name for a given picture. Hence, only pictures that were named in the same way by every participant, were included in the study.

## Experimental Design and Procedure

Participants completed 240 trials of a simple match/mismatch task. They were instructed to judge if the picture and the label represented the same object, and indicate their response by pressing one of two keys ( $C$ or $N$ ) on a keyboard (which were counterbalanced across participants). The number of match and mismatch trials was equal and the order of trials was randomized across participants. Depending on the experimental condition, participants were responding to pictures - when they were preceded by words (WP condition); words - when they were preceded by pictures (PW condition) or words and pictures when they were presented together (TO condition). The labels and pictures in the mismatch trials were from different categories and paired in a way to avoid phonological similarities and phonological onset competition (catcow); rhyme (dog-frog) as well as semantic association (cat-dog). Trials would start with a fixation cross, followed by a 700 ms presentation of word, picture, or word and picture together (depending on the experimental condition, with the difference being that in TO condition the stimuli would last until response, not only 700 ms ) after which they would see a picture or a word on which they had to respond to. The time sequence of a single trial for each experimental group is presented in Figure 1. In order to avoid preparatory movement potentials during the task a jitter of $\pm 200 \mathrm{~ms}$ for the fixation cross was introduced (Luck, 2005). According to Luck (2005) expecting a stimulus that requires a response can cause preparatory movement
potentials that are known to appear as contingent negative variations (CNV), a low frequency negative wave preceding an expected stimulus.


C Together Condition


Figure 1: Time sequence of individual trials for all three experimental conditions.

The experiment was conducted in a Faraday Cage. The participants were sitting in front of a computer, at approximately one meter distance from the screen. The stimuli were presented on a grey background at the center of the screen at eye level. Participants were instructed to avoid frequent blinking and reduce muscle movement as much as possible.

## ERP recordings

EEG signals were recorded continuously throughout the experiment. The signals were recorded from 15 electrodes placed at: F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, PC5, PC6, T5, T6, O1 and O2 sites according to the international $10-20$ standard. Two electrodes were placed on the earlobes as a reference, and the ground electrode was positioned on the participant's forehead.

PSYLAB EEG8 biological amplifier in combination with PSYLAB SAM unit (Contact Precision Instruments, London, UK) were used for EEG measurements. Skin-electrode contact impedance was below $5 \mathrm{k} \Omega$ at the beginning of the trials. EEG signal amplification was 20 k and hardware band-pass filtering over the range $0.03-40 \mathrm{~Hz}$. Signals were sampled at 500 Hz using NI USB- 6212 (National Instruments, Austin TX) card for analog to digital signal conversion. For EEG signal acquisition and online display a custom software with graphical user interface developed in LabVIEW 2010 was used (National Instruments, Austin, TX, USA) (Savic,Maleševic, \& Popovic, 2013). For determining the exact moment of stimulus onset upon which we time-lock the ERPs a sensor for detecting changes in brightness was placed in the upper-left corner of the screen. The stimuli had a black square in the sensor area which was not visible to the participants. This allowed a precision of 1 ms for determining stimulus onset.

## ERP processing

Offline EEG processing was conducted using custom routines in MATLAB (version 2010a, The Mathworks, Natick, MA, U.S.A.). EEG signals from all channels were filtered using a zero-phase 4th order Butterworth bandpass filter with $0.1-25 \mathrm{~Hz}$ cut-off frequencies.
The high pass component of the filter removes near-DC drift and the low pass component filters out muscle artifacts and 50 Hz noise, along with related harmonics. Data were then segmented into epochs including 100 ms baseline prior to stimulus onset, 900 ms following stimulus onset. The baseline was corrected in all EEG channels by subtracting from each epoch the mean of a 100 ms interval prior to the stimuli onset. Epochs contaminated with ocular-movements and/or other artefacts were rejected from further analysis if absolute value of the signal from any of the channels exceeded a threshold manually determined for each subject within a range of $40-60 \mu \mathrm{~V}$ (mean value: $48 \pm 6.4 \mu \mathrm{~V}$ ). The individual event related potential was calculated for each electrode site in each of the three experimental conditions. In the case where an individual electrode contained substantial noise compared to the average signal for the participant, only that individual electrode was removed, resulting in a small number of exclusions. Only three participants (one in each experimental condition) were excluded from the study on the basis of poor EEG signal.

## Results

## Behavioral Data

Accuracy rates for all three experimental conditions were extremely high, on average participants were correct $97 \%$ percent of the time. Mixed ANOVA analysis showed no difference in accuracy across conditions, but revealed a significant effect of match $/$ mismatch $\left(\mathrm{F}(1,59)=133.33 ; \mathrm{p}<.01,{ }_{\mathrm{p}} \eta^{2}=0.72\right)$, with participants making more errors in the match (5\%) trials in comparison to mismatch trials (1\%). Regarding

RTs, we found a main effect of experimental condition $\left(\mathrm{F}(2,59)=478.86 ; \mathrm{p}<.01,{ }_{\mathrm{p}} \eta^{2}=0.95\right)$. Post hoc tests revealed that only the TO condition differed significantly from both WP and PW condition (See Figure 3.).


Figure 3. RTs for match and mismatch trials across three experimental conditions (WP-word to picture; PW-picture to word; TO-together condition)

## ERP Data

The model used for analysis consisted of 9 electrode sites: F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, divided into three bands of coronal orientation (frontal-centralparietal), and three lateral regions (left-central-right). All analysis, including the determination of time windows of interest were done with difference waves. Namely, we substracted ERP wave forms of match trials from the ERP wave forms of mismatch trials in order to isolate the N400 component more accurately. For determining time windows of interest we adopted an exploratory approach to data analysis. Following the analysis of Kovic et al. (2010), mean amplitude measurements were extracted from the continuous EEG signal into 20 ms bins for each participant across all experimental conditions. Successive ANOVAs were conducted on each time bin. Windows of interest were defined if at least 3 consecutive 20 ms bins were significant ( $\mathrm{p}<.05$ ). After identification of windows, mean amplitudes across the window were computed for each experimental condition, and further analysis conducted. Two windows of interest ( $260 \mathrm{~ms}-440 \mathrm{~ms}$; $440 \mathrm{~ms}-680 \mathrm{~ms}$ ) were analysed with a $3 \times 3 \times 2$ repeated measures ANOVA with within-subjects factors of Frontality (Frontal, Central, Parietal) and Laterality (Left, Midline, Right), and between-subjects factor Time-condition (WP condition, PW condition). Given that the latency of the component of interest (namely N400) and pattern of responding was completely different between TO in comparison to WP and PW conditions, we decided to exclude the TO condition from the amplitude analysis.

Time window $260-440 \mathrm{~ms}$ We found that WP and PW condition differed in the N400 amplitude, given that there was a significant effect of Time-condition in the first time window $\left(\mathrm{F}(1,38)=6.01, \mathrm{p}<.01,{ }_{\mathrm{p}} \eta^{2}=0.14\right)$. PW condition elicited a more negative response than the WP condition (See Figure 4.) . Apart from that, we also
found a main effect of Laterality $\mathrm{F}(2,30)=79.24 ; \mathrm{p}<$ $.01,{ }_{\mathrm{p}} \eta^{2}=0.58$ ) and an interaction Laterality x Frontality $\left.\mathrm{F}(4,30)=7.86 ; \mathrm{p}<.01,{ }_{\mathrm{p}} \eta^{2}=0.18\right)$. The same effect reported here can be easily recognised in Figure 5. whereby the dark blue colour indicates the more prominent N400 effect across scalp distribution.


Figure 4. Difference waves showing the N400 and P600 effects across experimental conditions on the Cz electrode (WP-word to picture; PW-picture to word; TO-together condition)


## B Picture to Word Condition



C Word and Picture Together Condition


Figure 5. Heat maps showing time course of the distribution of the N400 and P600 effects across the scalp in all three experimental conditions. The dark blue color indicates the more negative amplitudes

Time window $440-680 \mathrm{~ms}$ A repeated measures ANOVA was conducted in order to analyze the difference of the P600 amplitude between the WP condition and PW condition. Analysis revealed a main effect of Time-condition $\mathrm{F}(1,38)=4.99 ; \mathrm{p}<.01,{ }_{\mathrm{p}} \eta^{2}=$ 0.12 ) with WP condition eliciting a more positive response (See Figure 4.). Similarly to the first, earlier time window, we found a main effect of Laterality
$\left(\mathrm{F}(2,30)=4.34 ; \mathrm{p}<.05,{ }_{\mathrm{p}} \eta^{2}=0.11\right)$ and a Laterality x Frontality interaction $\left(\mathrm{F}(4,30)=8.95 ; \mathrm{p}<.01,{ }_{\mathrm{p}} \eta^{2}=\right.$ 0.19 ). Figure 5. shows the distribution of activity through time and across scalpe. Orange colour indicates the more positive responses in amplitude.

## Discussion

In this study we tested if and how the order of stimuli presentation impacts semantic processing. In particular, we tested the hypothesis that the mapping from picture to name would be easier for processing in comparison to name-to-picture mapping given that there are many instances of pictures and thus mapping a single name to multiple potential objects seemed as a harder experimental condition in comparison to mapping from a particular picture to the name (which we pretested to select the most adequate for the given object).

The results we obtained demonstrate that the hardest condition for the semantic processing is the one in which the labels and pictures were presented at the same time. This finding was in accordance with our expectations because participants needed to process both information (word and picture) in parallel, which opens possibilities of interference as well as competitive processes making the task harder. Additionally, in the together condition, there was no priming in a strict sense as in the other two conditions, which prevented participants from forming any expectations which would help them with the task. It is noteworthy to say that in this study latency of ERP response in TO condition corresponds to a time window commonly associated with the P600 component. However, unlike the P600 component, here its polarity is negative. This is why we believe it is in fact a late N400 effect, delayed because of the difficulty of the task, which was also represented through longer RTs. All of this could account for the different morphology of the N400 in the TO condition.

Regarding the other two conditions, we observed a larger N400 amplitudes in PW condition in comparison to WP condition. Thus, in accordance with our expectations we found that picture to label mapping (PW) was easier for participants to process given that the N400 amplitude was more prominent in this condition in comparison to label to picture (WP) condition.

A more parsimonious way of interpreting this data would be in terms of violation of expectation reflected through the amplitude of the N400 component. This is consistent with a hypothesis that one can predict a word from a picture with more precision than the opposite (which leads to a larger violation of expectations).

The observed pattern of results in picture to word mapping would potentially be different in the case of less typical or atypical pictures. Violation of expectation in that case would certainly be higher than observed in the current study. Similarly, when mapping from word to picture, we would also expect greater violation of expectation then the one reported in this study.

Another interesting stream of research would be to contrast WP and PW mapping in the situation of novel object formation, that is - during the process of category formation. Here, we would have better control of the variability of the objects used in the study, given that with familiar objects the variability is much higher for the pictures (then for words).

Another component that turned out to be sensitive to semantic processing in this study, was the P600 component which in relevant literature is commonly related to syntactic processing (Kotz, Frisch, von Cramon \& Friederici, 2003; Osterhout \& Holcomb, 1992). However, there are a few studies which also reported P600 to be sensitive to semantic processing and interpreted as additional processing of meaning (Frisch, Schlesewsky, Saddy \& Alpermann, 2002; Martín-Loeches, Nigbur, Casado, Hohlfeld \& Sommer, 2006).

In our study, WP condition elicited a more positive P600 response in comparison to PW condition, which would suggest the information in this condition required additional processing in the later stages. However, given that there is an ongoing debate over the meaning of P600 in semantic processing, and since this component wasn't of main interest in this study, we would reserve from making firm claims when interpreting these results.

Practical implications of this research would be that in a classical priming experiments the best way to design the experiment would be to consistently map from pictures to words (that is, from more specific to more general representations), at least in the situation when typicality, familiarity and frequency of the selected pictures are high.

## Acknowledgments

We would like to express our gratitude to prof. Suzy Styles from Nanyang Techological University in Singapore for the valuable comments she made on the study, as well as to anonimous reviewers whose suggestions were incorporated in the final version of this paper.

This research was partially supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, Belgrade, Project No. 179033 and Project No. 179006.

## References

Anderson, J. E., \& Holcomb, P. J. (1995). Auditory and visual semantic priming using different stimulus onset asynchronies: An event-related brain potential study. Psychophysiology, 32(2), 177-190.
Boutonnet, B., \& Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. The Journal of Neuroscience, 35(25), 9329-9335.
Frisch, S., Schlesewsky, M., Saddy, D. \& Alpermann, A. (2002). The P600 as an indicator of syntactic ambiguity. Cognition, 85, B83-B92.
Ganis, G., Kutas, M., \& Sereno, M. E. (1996). The search for "common sense": An electrophysiological
study of the comprehension of words and pictures in reading. Journal of Cognitive Neuroscience 8(2), 89106.

Holcomb, P. J., \& Anderson, J. E. (1993). Cross-modal semantic priming: A time-course analysis using event-related brain potentials. Language and Cognitive Processes, 8(4), 379-411.
Kotz, S. A., Frisch,S., von Cramon D.Y. \& Friederici, A. D. (2003). Syntactic language processing: ERP lesion data on the role of the basal ganglia. Journal of the International Neuropsychological Society, 9, 1053-1060.
Ković, V., Plunkett, K., \& Westermann, G. (2009). Shared and/or separate representations of animate/inanimate categories: An ERP study. Psihologija, 42(1), 5-26.
Ković, V., Plunkett, K., \& Westermann, G. (2010). A unitary account of conceptual representations of animate/inanimate categories. Psihologija,43(2), 155165.

Kutas M, \& Federmeier K. D. (2009). N400. Scholarpedia. 4(10):7790.
Kutas, M., \& Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP).Annual review of psychology, 62, 621.
Kutas, M., \& Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. Science, 207(4427), 203-205.
Luck, S. J. (2005). An introduction to the event-related potential technique. MIT press.
Martín-Loeches, M., Nigbur, R., Casado, P., Hohlfeld, A., \& Sommer, W. (2006). Semantics prevalence over syntax during sentence processing: A brain potential study of noun-adjective agreement in Spanish. Brain Research, 1093, 178-189.
Osterhout, L., \& Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. Journal of Memory and Language, 31, 785-806.
Savić, A., Malešević, N., \& Popović, M. (2013). Motor imagery driven BCI with cuebasedselection of FES induced grasps. In J. L. Pons, D. Torricelli, \& M. Pajaro (Eds.). Converging clinical and engineering research on neurorehabilitation (Vol. 1, pp. 513-516). Berlin, Heidelberg: Springer.

# The Relationship Between Executive Functions and Science Achievement 

Drake Bauer<br>University of Minnesota<br>Sashank Varma<br>University of Minnesota

Keisha Varma<br>University of Minnesota

Martin Van Boekel
University of Illinois at Urbana Champaign

Alyssa Worley<br>University of Minnesota<br>Jean-Baptiste Quillien<br>University of Minnesota<br>Tayler Loiselle<br>University of Minnesota<br>Purav Patel<br>University of Minnesota


#### Abstract

Executive function is a fundamental component of the human cognitive architecture. Here, we investigate the relationship between executive function and scientific reasoning. Eighth graders completed measures of three executive functions (EFs): shifting, inhibiting, and updating. They also completed a measure of cognitive flexibility, the Wisconsin Card Sort Task (WCST), that has predicted scientific reasoning in prior studies. Scientific reasoning was measured by a standardized test of science achievement. A principal components analysis found that the three EFs were separable. Different EFs predicted different aspects of cognitive flexibility; notably, participants with poor shifting ability made more perseverative errors. Both EF and WCST predicted science achievement. Of note was the finding that better updating (i.e., working memory) was associated with higher science scores. These findings illuminate the role of EF in cognitive flexibility and scientific reasoning, and point the way to future studies of the effect of training EF on science achievement.


# Contrasts in reasoning about omissions 

Paul Bello, Christina Wasylyshyn, Gordon Briggs, Sangeet Khemlani \{paul.bello, christina.wasylyshyn, gordon.briggs.ctr, sangeet.khemlani\}@nrl.navy.mil<br>Naval Research Laboratory, Overlook Ave. S.W.<br>Washington, DC 20375 USA


#### Abstract

Omissions figure prominently in causal reasoning from diagnosis to ascriptions of negligence. One philosophical proposal posits that omissions are accompanied by a contrasting alternative that describes a case of orthodox (nonomissive) causation (Schaffer, 2005; Bernstein, 2014). A psychological hypothesis can be drawn from this contrast view of omissions: by default, humans should interpret omissive causations as representing at least two possibilities, i.e., a possibility representing the omission and a possibility representing a contrast. The theory of mental models supposes that reasoners construct only one possibility (the omission) by default, and that they consider separate alternative possibilities in sequential order. Two experiments test the contrast hypothesis against the model theory, and find evidence in favor of the model-theoretic account.


Keywords: omissive causation; mental models; reasoning; contrasts

## Introduction

A mechanical failure in a car to start causes a missed meeting. A friend's broken promise causes hurt feelings. The lack of rainfall causes drought. Each is a case of omissive causation, where an omission or lack of some event brings about some effect. Take the following oft-cited example:

> You come home after a business trip to find your rosebushes desiccated and ruined. You learn from your neighbor that your gardener did not show up to water the plants.

Omissive causes feature in both prediction and explanation. Intuition suggests that the death of the roses is explained by the failure of the gardener to show up. Similarly, we would expect that future failures of this kind would yield the same result. And omissions are often invoked in moral judgment. If you had signed a contract with the gardener and were especially litigious, you would have grounds to sue for damages. While we take for granted the fact that omissions are causes, omissions pose deep puzzles for theorists who wish to treat them in much the same way that "orthodox" (i.e., non-omissive) causes are treated.

It seems reasonable to think that orthodox causation concerns relations between events. But omissions are nonevents, and it is unclear how a non-event can be an argument to a causal relation. One idea is that omissions could be nothing at all (Clarke, 2014; Beebee, 2004), but this notion fails to explain why omissions seem to serve as sensible causal agents, as in, e.g., the lack of medicine caused sickness. If omissions were nothing, then they
couldn't be thought of as causes. Another proposal is that omissions denote a non-actualized possibility (Bernstein, 2014), such as that of the gardener showing up to water the bushes. Bernstein invokes the machinery of possible worlds to argue that omissions involve "counterpart relations" between actual omitted events and non-actualized contrast events at close-by possible worlds. A related idea in Schaffer (2005) is that omissions represent actual events, e.g., the event that occurred instead of the gardener showing up. The shared assumption from these latter two proposals about the nature of omissions is that they are definable in terms of contrasts between events (Schaffer, 2005), or between the omitted event and a non-actualized possibility (Bernstein, 2014).

Bernstein and Schaffer's accounts, while distinct, share the surface similarity of basing omissions on contrasts. In both cases, each theorist argues that an adequate metaphysics of omissions ought to not run afoul of human intuitions. In this spirit of consilience between intuition and metaphysical theorizing, it is worth exploring whether or not contrasts are present in mental representations of omissive causes and the inferences reasoners draw from them. The present paper explores if and how statements about omissions automatically refer to representations of contrasting events. Evidence from the psychology of counterfactual reasoning suggests that reasoners are in principle capable of maintaining two separate possibilities (Byrne, 2005), but whether they do so for reasoning about omissions remains unknown. If a strict interpretation of the contrast view is correct, reasoners should interpret omissive causation as referring to two representations by default.

In what follows, we review a psychological account of omissive-causal reasoning with mental models (Bello \& Khemlani, 2015). The theory predicts that reasoners tend to interpret omissive causation as referring to a single possibility: one in which the omitted event happens (e.g., the gardener fails to water the flowers), and the result follows (e.g., the flowers die). The theory posits that reasoners can potentially think about other possibilities including contrasts, e.g., the situation in which the gardener waters the flowers and they don't die, or the situation in which the gardener waters the flowers and they die for some other reason. But these alternative possibilities demand additional effort, and so by initially considering only one (non-contrasting) possibility, reasoners reduce the load on their working memories.

We describe two studies that test between the different hypotheses. The experiments support the model-based account in which reasoners do not represent contrasting possibilities by default, but instead consider alternatives
sequentially and in a systematic order. The paper concludes by discussing other puzzling aspects of omissive causation and plans for future research.

## The model theory

The mental model theory of reasoning - or "model theory" for short - posits that reasoners draw conclusions by building and scanning mental models, or iconic representations of possibilities (Johnson-Laird, 2006; Johnson-Laird \& Byrne, 1991, Goldvarg \& Johnson-Laird, 2001; Goodwin \& Johnson-Laird 2005). The model theory makes three central assumptions:

1. The principle of iconic possibilities. The contents of perception, memory, language, or imagination yield models, i.e., sets of discrete possibilities. Models are iconic, i.e., they are isomorphic to the structure of what they represent (Peirce, 19311958, Vol. 4), but they can also contain abstract tokens, such as a symbol denoting negation (Khemlani, Orenes, \& JohnsonLaird, 2012). And they can represent temporal sequences of events as discrete possibilities that unfold in time the way events do (Khemlani \& Johnson-Laird, 2013).
2. The principle of parsimony. Models require maintenance in working memory, and so inferences that demand more models are more difficult and take longer than those that demand fewer models. Hence, the theory posits two primary systems for reasoning: a fast system builds and scans models without the use of working memory, and so it posits that reasoners tend to reason with a single mental model in most scenarios. A slower system revises and rebuilds models, and it searches for alternative models consistent with the premises. It can correct the errors and biases that the fast system yields, but it is subject to the limitations of working memory.
3. The principle of truth. Reasoners initially build models that represent only what is true in a compound clause, and not what is false. They can flesh out the initial mental models to yield a set of fully-explicit models, i.e., those possibilities that denote both true, false, possible, and impossible scenarios. Fullyexplicit models form a complete representation of the possibilities to which a statement refers.

To illustrate these three principles, we now turn to summarizing the theory of omissive causation presented in (Bello \& Khemlani, 2015).

## A model-based account of omissive causation

According to the model theory, different sorts of causal verbs refer to different sets of mental models (Goldvarg \& Johnson-Laird, 2001; Khemlani, Barbey, \& Johnson-Laird, 2014). For instance, the statement, acid causes flowers to die, refers to three separate possibilities that constitute a fully-explicit model, which can be depicted in the following diagram:

$$
\begin{array}{rr}
\begin{array}{r}
\text { acid } \\
\neg \text { acid } \\
\neg \text { acid }
\end{array} \quad \neg \begin{array}{r}
\text { death } \\
\text { death }
\end{array} \\
\text { death }
\end{array}
$$

Each row of the diagram represents a different temporallyordered possibility that renders the statement true. Hence, the first row denotes the possibility in which acid is introduced and the flowers die; and the latter two rows denote possibilities in which acid isn't introduced and the flowers do not die (row 2) or die anyway (row 3). The model does not represent situations inconsistent with the statement (e.g., the situation in which acid is introduced and the flowers do not die, or any situation in which death occurs before acid is introduced). Moreover, maintaining three separate possibilities is difficult for reasoners, and the principle of parsimony implies that most reasoners only construct the first possibility, i.e., the mental model:
acid death

The mental model can be scanned and combined with other premises to yield inferences rapidly, but reasoners who rely on the mental model alone are prone to make reasoning errors on certain inferences. Moreover, each additional model in the set of fully-explicit models above demands working memory resources, and reasoners should be progressively less likely to consider them.

Omissive causation operates similarly to orthodox causation under the model theory, with the proviso that omissions imply that the antecedent events are negated (Bello \& Khemlani, 2015) and negations increase difficulty (Khemlani et al., 2012). For instance, the statement, the lack of water causes flowers to die, refers to the following mental model:

$$
\neg \text { water death }
$$

which can be fleshed out into the following fully-explicit models:

$$
\begin{array}{lr}
\neg \text { water } \\
\text { water } \\
\text { water }
\end{array} \quad \neg \begin{aligned}
& \text { death } \\
& \text { death }
\end{aligned}
$$

And so, just as in the case of orthodox causation, the model theory predicts that reasoners should often build only the mental model (i.e., the possibility in which there is a lack of water and the flowers die). Those who consider additional possibilities should construct the second possibility less often than the first, and the third possibility less often than the second.

## Models and contrasts

Do omissive causes entail contrasts? If so, then a central assumption of the model theory would be incorrect. That is, a default representation of contrasts would imply that statements such as the lack of water causes flowers to die should refer to the following two models:

$$
\begin{array}{cr}
\neg \text { water } \\
\text { water } & \quad \begin{array}{l}
\text { death } \\
\text { death }
\end{array}
\end{array}
$$

instead of just one mental model (see above). Reasoners do appear to consider the contrasting possibility often. In recent studies by Briggs and colleagues, participants evaluated
omissive causal relations (of a structure akin to: the lack of $A$ causes $B$ ) by assessing whether four separate scenarios (not- $A$ and $B, A$ and not- $B, A$ and $B$, and not- $A$ and not- $B$ ) were possible given the truth of the relation. Participants in one study, for instance, selected not- $A$ and $B$ at ceiling, and they selected $A$ and not- $B$ close to ceiling ( $98 \%$ and $83 \%$, respectively; see Briggs et al., under review, Experiment 3). The preponderance of $A$ and not- $B$ responses lends some tentative support to the idea that omissions are understood in terms of contrasts, as some metaphysicians have suggested. But, the data are also consistent with the view that reasoners select the contrasting possibility ( $A$ and not$B$ ) only after considering the mental model (not- $A$ and $B$ ) first. No studies directly test between the contrast view and the model theory, and so we carried out two experiments in which reasoners made inferences about omissive causation. Certain inferences should be less error-prone if reasoners represent contrasting possibilities, but Experiment 1 showed no such improvement. Experiment 2 showed that reasoners spontaneously generate contrasting possibilities less often and after - they represent possibilities corresponding to mental models. Both studies support the predictions of the model theory.

## Experiment 1

Experiment 1 tests reasoners' inferences about the pattern of reasoning known as modus tollens, which is an inference in sentential logic of the following form:

## If A then C .

Not C.
Therefore, not A.
The inference is valid because it is true in every case that the premises are true (Jeffrey, 1981, p. 1). But, reasoners have difficulty with modus tollens inferences: they tend to respond that nothing follows from the premises instead of inferring that not $A$ follows (Nickerson, 2015, p. 41 et seq.). A causal version of the inference is as follows:

## Overexposure to UV light causes snowblindness. A particular mountaineer doesn't have snowblindness. What follows?

A valid conclusion from these premises is that the mountaineer isn't overexposed to UV light. But, even in the causal domain, many reasoners have difficulty generating the valid conclusion, and they instead respond that nothing follows (Cummins, Lubart, Alksnis, \& Rist, 1991). The model theory explains why: if reasoners represent only a single mental model of overexposure to UV light causes snowblindness, e.g.,
UV-overexposure snowblindness
then that single possibility does not correspond to the possibility referred to in the second premise: a mountaineer doesn't have snowblindness. And so, reasoners respond that nothing follows. Only those reasoners who construct the fully explicit models of the causal relation:

$$
\begin{array}{rr} 
& \text { UV-overexposure } \\
\neg \text { UV-overexposure } & \neg \text { snowblindness } \\
\neg \text { snowblindness } \\
\neg \text { UV-overexposure } & \text { snowblindness }
\end{array}
$$

can make the valid deductive inference, because the second premise corresponds to the second possibility above.

The same prediction, mutatis mutandis, holds for omissive causal relations. Consider the following inference:

A lack of vitamin C causes scurvy.
A particular sailor doesn't have scurvy.
What follows?
If, as the model theory predicts, reasoners represent only a single mental model, e.g.,

$$
\neg \text { vitamin-C scurvy }
$$

then they should have difficulty drawing a valid conclusion from the premises. If, however, reasoners construct both the mental model and its contrasting possibility, e.g.,

$$
\begin{array}{rr}
\neg \text { vitamin-C } & \text { scurvy } \\
\text { vitamin-C } & \neg \text { scurvy }
\end{array}
$$

then they should be more likely to make the valid conclusion that the sailor doesn't have a vitamin $C$ deficiency. Hence, the contrastive view of omissive causation predicts that reasoners should respond more accurately on modus tollens inferences when they concern omissions than when they concern orthodox causation.

To test this prediction, participants in Experiment 1 wrote out their natural responses to short vignettes concerning omissive and orthodox causal reasoning arguments.

## Method

Participants. Thirty participants volunteered through the Amazon Mechanical Turk online platform (see Paolacci, Chandler, \& Ipeirotis, 2010, for a review). Fourteen participants reported no formal logic or advanced mathematical training and the remaining reported introductory to advanced training in logic. All participants were native English speakers.

Design, procedure, and materials. Participants carried out the experiment on a computer screen. The study was designed in psiTurk (Gureckis et al., 2015). After reading instructions, participants completed eight experimental problems. Half the problems concerned omissive causation by making use of the word "absence" to establish an omission; and the other half concerned orthodox causation by using the word "presence". Each problem comprised two premises. The first premise always established the presence or absence of a causal relation (e.g., $A$ causes $B$ ). For half of the problems, the second premise asserted that the event $(B)$ occurred (and therefore allowed participants to draw an inference known as affirming the consequent), and for the other half, the premise asserted that the event did not occur (not-B), and so participants could draw a modus tollens inference. Participants wrote out responses to the question "What, if anything, follows?" An example problem is as follows:

```
Suppose the following statements are true:
    1. The [presence/absence] of a particular part causes a
        machine to fail.
    2. On a particular day, the machine [did/didn't] fail.
What, if anything, follows?
```

The information for each problem was presented simultaneously, and participants were prevented from continuing to the next problem until they typed in at least one possibility. Participants were informed that they should write out that "nothing followed" if they thought there was not enough information in the premises to make any conclusion with certainty. The materials were drawn from four domains: biology, nature, socioeconomics, and mechanics. The presentation order of the content and problem type of the vignettes was randomized.

## Results and discussion

Two coders blind to the predictions of the study judged whether participants' natural responses were accurate or inaccurate; they agreed on $99 \%$ of trials (Cohen's $\kappa=.99$ ). Table 1 shows the percentage of accurate responses in Experiment 1 as a function of whether the inference concerned omissive or orthodox causation. Across the study as a whole, participants produced more accurate responses for orthodox causation than for omissive causation ( $47 \%$ vs. $32 \%$ correct; Wilcoxon test, $z=2.49, p=.01$, Cliff's $\delta=$ .15), which is the opposite of the pattern predicted by the contrast view. As in previous studies (e.g., Cummins et al., 1991), participants produced more accurate responses for modus tollens inferences than for affirming the consequent inferences ( $63 \%$ vs. $15 \%$ correct; Wilcoxon test, $z=6.65, p$ $<.0001$, Cliff's $\delta=.48$ ). The interaction between the type of causation and the type of inference was not reliable (Wilcoxon test, $z=.64, p=.52$, Cliff's $\delta=.03$ ).

Participants in Experiment 1 violated the prediction of the contrast view: they were less accurate for inferences concerning omissive causation than for those concerning orthodox causation. Indeed, their patterns of inference corroborate the model theory, which predicts that inferences about omissive causes should be slightly more difficult because reasoners represent negated possibilities.

|  | Type of causation |  |
| :--- | :---: | :---: |
| Inference | Orthodox <br> The presence <br> of $A \ldots$ | Omissive <br> The absence of <br> A... |
| Affirming the consequent: |  |  |
| A causes B. B. |  |  |
| What, if anything, follows? | $22 \%$ | $8 \%$ |
| Modus tollens: <br> A causes B. Not B. <br> What, if anything, follows? | $72 \%$ |  |

Table 1. Proportion of correct responses in Experiment 1 as a function of the type of inference and the type of causation.

## Experiment 2

Experiment 1 concerned inferences, and its results suggest that reasoners do not make use of contrasting possibilities in their modus tollens or affirming the consequent inferences. But, the data do not conclusively establish whether or not reasoners represent the contrast by default. After all, people might initially represent such possibilities but fail to consider them when drawing inferences. Hence, Experiment 2 tested reasoners' interpretations of omissive causation directly. It elicited natural responses to the different possibilities for orthodox and omissive cause and enabling conditions. Participants read a single short premise and were asked to list the possibilities that correspond to each premise. We analyzed the order in which participants constructed each possibility, as well as the first possibility they constructed. The contrast view predicts that reasoners should construct the possibilities that correspond to not-A and $B$ and $A$ and not- $B$ equally often when they interpret omissive causation. The model theory predicts that reasoners should construct the possibility that corresponds to not- $A$ and $B$ first, then (if at all) the possibility that corresponds to $A$ and not- $B$, and finally (if at all) the possibility that corresponds to $A$ and $B$. And the theory predicts an analogous trend in latencies: reasoners should build not- $A$ and $B$ faster than $A$ and not- $B$, and they should build $A$ and not- $B$ faster than $A$ and $B$.

## Method

Participants. Thirty-one participants volunteered through the Amazon Mechanical Turk online platform. Twenty-two participants reported no formal logic or advanced mathematical training and the remaining reported introductory to advanced training in logic. All were native English speakers.

Design, procedure, and materials. Participants completed two practice problems and eight experimental problems, and they acted as their own controls. Each problem presented one premise that consisted of two events and a causal verb. The experiment manipulated whether the first event concerned orthodox or omissive causation: half the problems used the word "presence" and the other half used the word "absence." The experiment also manipulated the relevant causal relation: half the problems concerned causation and half concerned enabling conditions, though for brevity we analyze only those problems concerning causation below. An example problem is as follows:

> Suppose the following statement is true: The [presence/absence] of a particular preservative [causes/enables] a substance to decay.
> What is possible given the above statement?

Participants were then asked to construct a list of possibilities using pre-populated drop-down menus. Figure 1 shows an example of the interface used in Experiment 2. Participants could choose any combination of the possibilities from the drop-down menus, they could change


Figure 1. The interface used to elicit responses in Experiment 2. Participants completed sentences using drop-down menus and added possibilities using a button marked " + ".
their answer choices at will, and they could add additional sentences if they thought the statement was true in a number of possibilities; but, the interface allowed the construction of at most four different sentences. The presentation order of the trials was randomized. The order in which the participants endorsed possibilities was recorded, as was the latency between when the premises appeared and when participants pushed a button to finish the trial.

## Results and discussion

Table 2 shows the percentage of trials on which participants constructed the four possible sentences as a function of whether the premise in the trial concerned an orthodox or an omissive causal relation. The table also shows, in parentheses, the percentages of trials on which a given sentence appeared first in the set of sentences constructed by the participants.

For omissive causation trials, participants constructed not$A$ and $B$ more often than $A$ and not- $B(85 \%$ vs. $69 \%$, respectively; Wilcoxon test, $z=2.88, p=.003$, Cliff's $\delta=$ .03 ), in violation of the contrast view. Instead, the data corroborate the trend predicted by the model theory; participants constructed not-A and $B$ most often ( $85 \%$ of trials), then $A$ and not- $B(69 \%)$, then $A \& B(47 \%)$, and rarely not- $A$ and not- $B$ (19\%). A nonparametric trend test revealed a significant trend in their responses (Page's trend test, $z=5.16, p<.0001$ ).

One way of understanding participants' performance is to examine only the first sentence in the set of sentences they constructed: doing so allows for a coarse analysis of their online preferences for possibilities. Participants constructed not- $A$ and $B$ as a first sentence more often than $A$ and not- $B$

| Type of <br> causation | The four sentences (in abbreviated form) |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | $A \& B$ | $A \& \neg B$ | $\neg A \& B$ | $\neg A \& \neg B$ |
| Orthodox | $100(100)$ | $6(0)$ | $31(0)$ | $74(0)$ |
| Omissive | $47(26)$ | $69(26)$ | $85(48)$ | $19(0)$ |

Table 2. Percentages of trials on which participants in Experiment 2 constructed four separate sentences for trials that concerned omissive and orthodox causal relations. ' $\neg$ ' denotes negation. In parentheses: percentages of trials on which a sentence appeared first in the set constructed by participants. (Not shown: data from trials that concerned enabling conditions.)
( $48 \%$ vs. $26 \%$, respectively; Wilcoxon test, $z=2.06, p=$ .04; Cliff's $\delta=.22$ ). And they constructed $A$ and not- $B$ and $A$ and $B$ equally as often ( $26 \%$ vs. $26 \%$ ). We recorded the latency between when the premises appeared to when participants made a response. While those latencies are inflated to include the amount of time they read the premises, they nevertheless revealed that participants were faster to construct not- $A$ and $B(26 \mathrm{~s})$ as a first sentence compared to $A$ and not-B (38 s); a test on their overall selections corroborated the trend predicted by the model theory (Jonckheere's trend test, $z=2.95, p=.001$ ).

Participants' responses to orthodox causal relations likewise corroborated the predictions of the model theory. Every participant responded $A$ and $B, 74 \%$ of participants responded not- $A$ and not- $B$, and $31 \%$ of participants responded not- $A$ and $B$ (Page's trend test, $z=3.54, p<$ .001). And, every participant constructed $A$ and $B$ first.

The results largely corroborate the predictions of the model theory. The significant trends in the proportions of constructing the different sentences suggest that reasoners consider possibilities sequentially.

## General discussion

Two experiments showed that when people interpret and reason about omissive causal relations, they do not represent contrasting alternatives by default. Reasoners can, in principle, hold two separate possibilities in mind - they seem to do precisely that when reasoning about counterfactual assertions (Byrne, 2005). But, as the present studies show, they tend to interpret omissive causes as referring to a single model of a negated cause and its associated effect (Experiment 1). When they are asked to list possibilities, they list the mental model earlier, more often, and faster than contrasting possibilities (Experiment 2). These results corroborate the model theory of omissive causation (Bello \& Khemlani, 2015; Briggs et al., under review), and no alternative theory of omissive causation, whether psychological (Wolff et al., 2010) or philosophical, presently account for the results from the two studies.

The model theory provides a specific ordering on what kinds of possibilities reasoners consider. It is a process theory that explains why some possibilities (the mental models) are considered by default and why others (the alternative models) demand additional cognitive resources to construct. And it provides some constraints on the contents of alternative properties, though the specific contents depend on the semantics of the particular verbs used to describe omissive-causal premises.

How are contrast events identified in the first place? Both Bernstein and Schaffer highlight this open question, and they suggest that its answer is critical for shoring up their respective theories. Philosophers who interpret omissive causation using possible worlds are faced with developing a theory that specifies how to pick through the infinitude of possible worlds to find those that contain the most relevant contrast event. Actual-event theorists like Schaffer shoulder the same explanatory burden. To illustrate, it is perfectly
reasonable to think that the Queen of England might have shown up to water the rosebushes. This would have prevented their death, after all. Does her failure to do so qualify her as an omissive cause of their dying? What explains why the gardener's failure to do so is a better candidate for the actual cause of their death?

Recent work shows norms and pragmatics help establish relevant contrasts when reasoning about omissions (see Henne, Pinillos, \& De Brigard, 2016). Analogously, the model theory posits that background knowledge of the meanings of words and their contexts can introduce relations and block the construction of certain possibilities in a process known as modulation (Johnson-Laird \& Byrne, 2002). For instance, reasoners often infer a temporal relation from the conjunction, Mary studied and she passed her test, such that she studied before she passed her test (Juhos et al., 2012). In the case of omissive causes, modulation may rely on knowledge of norms to introduce relations or contents for the different contrasting possibilities, and future work will investigate the processes by which norms bias the representation of omissive causes.

## Acknowledgments

This work was supported by grants from the Office of Naval Research to PB and SK. We are grateful to Knexus Research Corporation for their help in conducting the experiments, and we are especially grateful to Ruth Byrne, whose research on counterfactual inferences inspired the design of Experiment 1. Finally, we thank Monica Bucciarelli, Felipe de Brigard, Todd Gureckis, Paul Henne, Laura Hiatt, Phil Johnson-Laird, Greg Murphy, L.A. Paul, Bob Rehder, and Greg Trafton for their advice and comments.

## References

Beebee, H. (2004). Causing and nothingness. In J. Collins, N. Hall, and L. A. Paul (Eds.), Causation and Counterfactuals, Cambridge, MA: The MIT Press.
Bello, P., \& Khemlani, S. (2015). A model-based theory of omissive causation. In Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Bernstein, S. (2014). Omissions as possibilities. Philosophical Studies, 167.

Briggs, G., Wasylyshyn, C., Bello, P., \& Khemlani, S. (under review). Mental models of omissive causation.
Byrne, R. (2005). The rational imagination: How people create alternatives to reality. Cambridge, MA: The MIT Press.
Clarke, R. (2014). Omissions: Agency, metaphysics, and responsibility. Oxford University Press.
Cummins, D. D., Lubart, T., Alksnis, O., \& Rist, R. (1991). Conditional reasoning and causation. Memory \& cognition, 19.
Gureckis, T. M. et al. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. Behavior Research Methods, 1-14.
Goldvarg, E. \& Johnson-Laird, P. (2001). Naïve causality: A mental model theory of causal meaning and reasoning. Cognitive Science, 25.
Goodwin, G.P., \& Johnson-Laird, P.N. (2005). Reasoning about relations. Psychological Review, 112.
Henne, P., Pinillos, A., \& De Brigard, F. (2016). Cause by omission and norm: Not watering plants. Australasian Journal of Philosophy.
Jeffrey, R. (1981). Formal logic: Its scope and limits (2nd Ed). New York: McGraw-Hill.
Johnson-Laird, P.N. (2006). How we reason. NY: OUP.
Johnson-Laird, P. N., \& Byrne, R.M.J. (1991). Deduction. Hillsdale, NJ: Erlbaum.
Johnson-Laird P.N., \& Byrne R. (2002). Conditionals: A theory of meaning, pragmatics, and inference. Psychological Review, 109.
Juhos et al. (2012)
Khemlani, S., Barbey, A., \& Johnson-Laird, P. N. (2014). Causal reasoning with mental models. Frontiers in Human Neuroscience, 8, 849.
Khemlani, S., \& Johnson-Laird, P. N. (2013). The processes of inference. Argument \& Computation, 4, 1-20.
Khemlani, S., Orenes, I., \& Johnson-Laird, P.N. (2012). Negation: a theory of its meaning, representation, and use. Journal of Cognitive Psychology, 24.
Nickerson, R. S. (2015). Conditional reasoning: The unruly syntactics, semantics, thematics, and pragmatics of "If". New York Oxford University Press.
Paolacci, G., Chandler, J., \& Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. Judgment and Decision Making, 5.
Peirce, C.S. (1931-1958). Collected papers of Charles Sanders Peirce. 8 vols. C. Hartshorne, P. Weiss, and A. Burks, (Eds.). Cambridge, MA: Harvard University Press.
Schaffer, J. (2005). Contrastive Causation. Philosophical Review, 114.

Wolff P., Barbey A., Hausknecht M. (2010). For want of a nail: how absences cause events. Journal of Experimental Psychology: General, 139.

# Representing time in terms of space: Directions of mental timelines in Norwegian 

Andrea Bender ${ }^{1}$ (andrea.bender@uib.no)<br>Kristin Sjåfjell ${ }^{1}$ (kristin.sjafjell@ub.uib.no)<br>Annelie Rothe-Wulf ${ }^{2}$ (annelie.rothe@ psychologie.uni-freiburg.de)<br>Sieghard Beller ${ }^{1}$ (sieghard.beller@uib.no)<br>${ }^{1}$ Department of Psychosocial Science, University of Bergen<br>N-5020 Bergen, Norway<br>${ }^{2}$ Department of Psychology, University of Freiburg<br>D-79085 Freiburg, Germany


#### Abstract

People often use spatial vocabulary to describe temporal relations, and this has increasingly motivated attempts to map spatial frames of reference (FoRs) onto time. How people assign FRONT to time and to temporal entities depends on cultural conventions, and is crucial for diagnosing which temporal FoR a person actually adopts. Here, we report findings from a survey with speakers of Norwegian that aimed at assessing the cultural conventions involved in FRONT assignment. Data on temporal movements of events, on the temporal order of events, and on explicit FRONT assignments to events, time units, and "time itself" suggest that participants use different principles for describing fixed relations (static time) versus moving events (dynamic time).


Keywords: space; time; space-time mapping; frames of reference; mental timeline.

## Introduction

When talking about time, people tend to do so with vocabulary and concepts borrowed from the domain of space. Yet, while research in the two domains and the acknowledgement of cross-domain transfers do have a venerable tradition (reviewed in Núñez \& Cooperrider, 2013), the challenge of mapping a taxonomy of spatial representations onto the domain of time has been taken up only recently, and respective attempts differ considerably in terms of theoretical conceptualization and subsequent interpretation of data. Based on a review of advances in this field, we outlined how such taxonomies may be transferred from space to time (Bender \& Beller, 2014), with a focus on accounts that deal with frames of reference (FoRs).
Taking Levinson's (2003) well-established taxonomy of spatial frames of reference as starting point, the $t$-FoR account (Bender et al., 2010, 2012; Rothe-Wulf et al., 2015) derives a set of temporal frames of reference ( t -FoRs) following general design principles as described below. Yet, while these design principles provide an abstract structure for distinct frames of reference, their concrete specification depends on cultural conventions involved in how people assign FRONT to temporal entities per se. Previous accounts drew on intuitions regarding such conventions for speakers of English and related languages (such as that FRONT of an event is considered to be at its beginning). Here, we report findings from a survey that empirically assessed such conventions.

## Spatial and Temporal Frames of Reference

A frame of reference (FoR) is a coordinate system required to localize a figure $F$ in reference to a ground $G$ from an observer's point of view $V$. Levinson's (2003) taxonomy distinguishes three basic types of spatial FoRs, absolute, intrinsic, and relative, as well as different variants of the latter. In line with the underlying design principles, these FoRs can be mapped from space onto time as follows (Bender \& Beller, 2014; and see Table 1):

The absolute FoR is anchored in a superordinate field outside F, G, and V. As space itself is the superordinate field in the spatial domain, so is time in the temporal domain. Assignment of orientation to the field follows cultural conventions and may recruit, for instance, cardinal points, mountain slopes, rivers, or the land-sea axis on small islands (in the case of space), and correspondingly the asymmetry inherent in the 'arrow of time' (in the case of time), which is (presumably) pointing towards the future.

The intrinsic FoR is anchored in the reference or ground entity G and can thus only be adopted if G is perceived as being oriented itself (this includes an observer if serving as ground). Assignment of orientation again follows cultural conventions and may recruit, for instance, moving directions of objects such as cars (in space) and the beginning versus end of events (in time).

Table 1: Frames of reference and forward movements (indicated by the tips of the arrows) according to the $t$-FoR account (for more details, see Bender \& Beller, 2014), based on assumed cultural conventions for assigning FRONT in English speakers.


The relative FoR, finally, is anchored in the viewpoint V of an observer (separate from G). V's position can be established both in space (as the observer's location) and in time (as the observer's subjective present). In order to still be able to localize F in reference to G , the coordinate system primarily anchored in V needs to be shifted into G. This can be done in several ways, two of which are relevant here: the reflection variant under which FRONT is assigned to a position or time between G and V , and the translation variant under which FRONT is assigned to a position or time beyond G. In either case, FRONT assignments are point-symmetrical to the present, but have diverging directions (reflection: towards V; translation: away from V).
Each of these FoRs hinges on cultural conventions: the absolute FoR on how orientation is assigned to the superordinate field (for variation, see Núñez \& Sweetser, 2006), the intrinsic FoR on how orientation is assigned to the ground entity, and the relative FoR on which variant is preferred for shifting the primary coordinate system. So far, assumptions on these conventions are based more on intuitions than on data, especially for the domain of time. In the following, we explicate these for three Germanic languages.

## Frames of Reference in Germanic Languages

Empirical research in the spatial domain on three Germanic languages-English, German, and Swedish, (e.g., Beller et al., 2015; Grabowski \& Weiß, 1996; Majid et al., 2004)indicates that speakers of these languages make use of all basic spatial FoRs for describing locations and movements in space, with a pronounced preference in small-scale space for the reflection variant of the relative FoR and, albeit to a lesser extent, the intrinsic FoR.
In the temporal domain, the metaphorical space-time mapping emerging in language suggests a set of conventions for FRONT assignment that appear similar across the three languages. With regard to time itself (as the superordinate field in the absolute FoR), FRONT seems to be assigned to the future, as reflected in the 'arrow of time' pointing towards the future or in expressions such as "the future ahead", and "olden days passed by". With regard to events (as the ground entities in the intrinsic FoR), FRONT seems to be assigned to that part of time pertinent to the beginning of events, as reflected in expressions such as "the quiet before the storm". When it comes to the subjective viewpoint V of an observer (as the central point in the relative FoR ), it might be ventured that none of its variants are frequent in Germanic languages, as the point-symmetric patterns arising from them have been observed only infrequently (with $2.5 \%$ or less in any of the languages under investigation; see Rothe-Wulf et al., 2015).
Based on these assumptions, an absolute FoR would be diagnosed when events "in front of" other events or "moved forward" from their previous position are localized as further in the future, while an intrinsic FoR would be diagnosed when they are localized as further in the past (Table 1). Interestingly, the latter pattern has been described as canonical for all three Germanic languages for describing
fixed relations such as "the quiet before the storm" (German: "die Ruhe vor dem Sturm", Swedish: "lugnet före stormen"), while patterns in the three languages differ fundamentally when it comes to movement: Moving a meeting "forward" results in a later date (futurewards movement) for the vast majority of Swedish speakers, in an earlier date (pastwards movement) for the vast majority of German speakers, and in dissent between these variants for English speakers (Rothe-Wulf et al., 2015; and see Boroditsky \& Ramscar, 2002; McGlone \& Harding, 1998).

Whether, however, these diverging patterns can be diagnosed as arising from an absolute or intrinsic FoR, respectively, depends on whether our assumptions regarding the cultural conventions for FRONT assignment are correct. Data on this question was collected in the current study for a fourth Germanic language, namely Norwegian, for which research on spatial FoRs revealed the same preferences for references in small scale space (Beller \& Bender, 2017) as in the other three Germanic languages.

## Study

The study aimed at assessing whether and how FRONT is assigned to time itself (relevant for the absolute FoR) and to temporal entities such as events (relevant for the intrinsic FoR). We also remained open to the possibility of pointsymmetric response patterns indicative of a relative FoR.

## Methods

Participants. 81 volunteers participated in the survey; three were excluded from further analyses because they indicated a language other than Norwegian as their mother tongue. The resulting sample therefore consisted of 78 participants ( 59 female; age $M=25.3$ years, $S D=7.6$, range 19-62, with 5 not indicating their age).

Materials. The tasks described in the following were part of a larger paper-and-pencil survey, provided in Norwegian (bokmål). Here, we focus only on those tasks that are relevant for the questions under scrutiny in this paper.

The Event-Moving Task consisted of four items, with an event to be moved forward (Norwegian: fram) or backward (bakover) in time. Two items used the time scale days:

- The concert scheduled for Thursday last week was moved $\{$ forward/backward $\}$ two days. On which day of the week did it actually take place?
- The meeting scheduled for Wednesday next week will be moved \{forward/backward\} two days. On which day of the week will it now take place?
The other two items used the time scale hours:
- The departure scheduled for 9 a.m. yesterday was moved \{backward/forward\} three hours. At what time did it actually take place?
- The power cut scheduled for 4 p.m. tomorrow will be moved \{backward/forward\} three hours. At what time will it take place now?
For each time scale, a past and a future event was included;
this is necessary to be able to distinguish linear from pointsymmetric t-FoRs (cf. Table 1) that participants might adopt (Bender et al., 2010). The original scheduling of the events and the moving span were chosen so as to remain within the respective time cycle (e.g., for weekdays between Monday and Saturday), and hence to prevent ambiguous responses.

The items were implemented in four arrangements, crossing between-subjects two orders of time scales with the two moving directions. The task started either with the time scale days (first meeting, then concert), followed by hours (first power cut, then departure), or vice versa, and either with "forward" as moving direction for the first two events, followed by "backward" for the other two, or vice versa.
The Order Task consisted of six items that asked for the order of events, that is, whether a target event (figure F) is "in front of" (Norwegian: foran) or "behind" (bak) a reference event G. Four items used a forced-choice format:

- Lunch is normally ...
$\square$ in front of / $\square$ behind $\ldots$.. breakfast.
- Good Friday is two days ...
$\square$ in front of / $\square$ behind ... Easter Sunday.
- New Year's Eve is one week ...
$\square$ in front of / $\square$ behind ... Christmas Eve.
- The Stone Age was ...
$\square$ in front of / $\square$ behind ... the Middle Ages.
Two further items used an open format:
- The exam is generally nine days
\{in front of/behind \} the $17^{\text {th }}$ of May.
So, at which date does it take place?
- This year, Peter's birthday is three months \{in front of/behind \} midsummer.
So, in which month is his birthday?
The items were implemented in four arrangements, crossing between-subjects two orders of items either with two orders of response options for the items in the forced-choice format ("in front of" as the first vs. the second option) or with the two phrasings for the items in the open format ("in front of" for the birthday item and "behind" for the exam item, or vice versa). One item order was determined randomly with the second order being the exact reversal ${ }^{1}$.

The Front Task consisted of eight items that asked for indicating whether or not a time segment has a front (Norwegian: forside) or back (bakside), and if so, in which direction FRONT or BACK is pointing. All items followed the same schema and had four response options, here exemplified for the item on time in general:
\{Front/Back\} of time in general ...
$\square$ is at the beginning of time.
$\square$ is at the end of time.
$\square$ Something like that does not exist.
$\square$ Something else, namely $\qquad$ —.

As the two last response options were the same for all items, we explicate only the item-specific options for the remain-

[^278]ing items. Three items referred to the units of time day, month and year:

- $\{$ Front/Back $\}$ of today ...
$\square$ was early in the morning / $\square$ will be late at night.
- \{Front/Back\} of August ...
$\square$ is the $1^{\text {st }}$ of August / $\square$ is the $31^{\text {st }}$ of August.
- $\{$ Front/Back $\}$ of the current year ...
$\square$ was in January / $\square$ will be in December.
Four other items referred to events:
- \{Front/Back $\}$ of a meeting ...
$\square$ is at the introduction $/ \square$ is at the summary.
- $\{$ Front/Back $\}$ of a dinner...
$\square$ is at the appetizer / $\square$ is at the dessert.
- $\{$ Front/Back $\}$ of Easter ...
$\square$ is on Maundy Thursday / $\square$ is on Easter Monday.
- $\{$ Front/Back $\}$ of your life ...
$\square$ is when you were born / $\square$ is when you will die.
The items were implemented in four arrangements, crossing between-subjects two phrasings (asking for all items either for "front of X ..." or "back of X ...") with two orders of items (one random order starting with time in general, and the exact reversal ${ }^{1}$ ).

Design and Procedure. Four versions of questionnaires were constructed. The various types of tasks were presented within-subject in a fixed order (i.e., event-moving task followed by order task followed by front task) in line with the increasingly explicit nature of the task (asking for the "front" of time highlights the topic of interest more strongly than asking for the date to which an event is moved). The four item arrangements of each task were randomly assigned to one of the four versions of questionnaires, and varied between-subjects as indicated in the Materials section. Participants were instructed to work on all tasks in the given order.

## Results

For each task, we first describe how FRONT assignments were coded and then report participants' preferences.

Event-Moving Task. In this task, participants had to move an event either forward or backward in time. The responses were coded as whether they indicated that FRONT of the moving direction pointed towards the future or towards the past. For the items with "forward"-phrasing, the coding is obvious: If, for example, Wednesday's meeting is moved "forward" to Monday, then the assignment of FRONT and the moving direction points pastwards. For items with "back-ward"-phrasing, coding is reversed: If Wednesday's meeting is moved "backward" to Monday, then the assignment of FRONT and the corresponding forward direction point futurewards.

Each single item was tested first for potential influences of the order of time scales (days first vs. hours first) and the requested moving direction (forward vs. backward) on the actually chosen direction (futurewards vs. pastwards). No

Table 2: FRONT assignments (\%) in the event-moving task.

|  | Single items |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| FRONT <br> pointing | Concert <br> $(N=77)$ | Meeting <br> $(N=77)$ | Departure <br> $(N=77)$ | Power cut <br> $(N=76)$ | $M$ |
| futurewards | 63.6 | 71.4 | 58.4 | 68.4 | 65.5 |
| pastwards | 36.4 | 28.6 | 41.6 | 31.6 | 34.5 |
| Future/past items of time scale |  |  |  |  |  |

significant effects were found (all $G^{2} \leq 5.88 ; d f=3$; $p \geq .118$ ). Across items, futurewards movements dominated ( $65.5 \%$ on average; Table 2, upper half).

Then, we determined for each of the two pairs of items with the same time scale the pattern of FRONT assignments that resulted from considering both the future and past event as pointing futurewards, pastwards, towards V (present), or away from V (cf. Table 1). We checked whether the distribution of the four patterns for the time scale days differed from the distribution for the time scale hours, which would be indicative of an influence of time scale on temporal movements. According to a marginal homogeneity test for paired tasks, this was not the case ( $s t d$. MH statistic $=.906$, $p=.365$ ), thus justifying an aggregation across the time scales. Overall, the two linear patterns prevailed by far (Table 2, lower half). The majority of participants ( $60.8 \%$ on average) made futurewards movements, about one third ( $30.1 \%$ ) made pastwards movements, while of the two point-symmetric patterns, only the one with moving directions away from V (present) was chosen (9.2\%).

Order Task. In this task, participants had to specify whether a target event is "in front of" or "behind" a reference event. The responses were coded as whether they indicated that FRONT of the reference event was assigned to the end of the event, and hence pointing towards the future, or to the beginning of the event, pointing towards the past. For instance, Good Friday is always earlier in the year than Easter Sunday. The response "Good Friday is in front of Easter Sunday" therefore implies that FRONT of Easter Sunday is assigned to its beginning and points pastwards. With lunch as target event in reference to breakfast, coding would be reversed: As lunch is the later event, the response "Lunch is in front of breakfast" implies that FRONT of breakfast is assigned to its end and points futurewards.

Each item was tested first for potential effects of two factors on the coded FRONT of the event: the order of items (for all 6 items) and either the order of response options (for the 4 items with forced-choice format) or the phrasing (for the 2 items with open format). For the forced-choice items, no
significant effects were found (all $G^{2} \leq 3.77 ; d f=1$; $p \geq .052$ ). With some variation between events, the majority of responses indicated that FRONT of an event was assigned to its beginning and pointed pastwards ( $83.6 \%$ on average; Table 3, upper half). For the items with open format, main effects of the phrasing were found $\left(G^{2} \geq 12.95 ; d f=1\right.$; $p<.001$ ). Responses indicating that FRONT pointed pastwards were more frequent when participants had to specify whether the target is "in front of" the reference event $(100 \%)$ than when they had to specify whether the target is "behind" the reference event ( $76.4 \%$ on average; Table 3, lower half) ${ }^{2}$.

Finally, we determined how consistently the two possible FRONT assignments were made across the whole set of items. To this end, we counted for each participant how often FRONT pointed futurewards and how often it pointed pastwards. FRONT assignments were highly consistent. Participants used the same type of assignment on 5.17 (86.1\%) of the 6 items. Overall, 65 participants $(83.3 \%)$ had a preference for a pastwards directed FRONT and four participants (5.1\%) for a futurewards directed FRONT; the remaining 9 participants (11.5\%) had no preference.

Taken together, the results from the order task support the idea that FRONT of a time segment is at its beginning and that it points towards the past, at least for most of the participants. While in this task FRONT assignments were assessed indirectly from the order of events, the next task explicitly asked participants to indicate the "front" or "back" of events, time units, and time in general.

Front Task. In this task, participants had to specify whether "front" (or "back" respectively) of a temporal entity is at its beginning, at its end, does not exist, or something else. FRONT assignments were coded in four categories, with FRONT pointing futurewards, or pointing pastwards, is nonexistent, or something else ("other"). For the items asking to indicate the "front" of an event, coding is again obvious:

Table 3: FRONT assignments (\%) in the order task.

|  | Items with forced-choice format |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| FRONT <br> pointing | Lunch <br> $(N=77)$ | Good Friday <br> $(N=78)$ | New Year <br> $(N=78)$ | Stone Age <br> $(N=78)$ | $M$ |  |
| futurewards | 14.3 | 16.7 | 9.0 | 25.6 | 16.4 |  |
| pastwards | 85.7 | 83.3 | 91.0 | 74.4 | 83.6 |  |
| Items with |  |  |  |  |  | open format |
| FRONT | Exam $(N=41,37)$ |  |  |  |  |  |
| pointing | In front of | Behind | In front of | Behind |  |  |
| futurewards | - | 21.6 | - | 25.7 | 11.6 |  |
| pastwards | 100.0 | 78.4 | 100.0 | 74.3 | 88.4 |  |

[^279]Choosing, for example, the beginning as the "front of the event" implies that FRONT points towards the past. For the items asking to indicate the event's "back", coding was reversed: Choosing the beginning as its "back" implies that Front is assigned to the end of the event and points towards the future. Therefore, the response "the back of a meeting is at the summary" implies that FRONT is assigned to the beginning of the meeting and hence points towards the past.

Each item was tested first for potential effects of the phrasing and the order of items on the coded FRONT. A main effect phrasing was found in all cases (all $G^{2}>8.03 ; d f=3$; $p<.046$ ), a main effect order of items in three cases (life, year, and time; $G^{2}>9.38 ; d f=3 ; p<.025$ ), and an interaction of the two factors in two cases (dinner and Easter; $\left.G^{2}>11.15 ; d f=3 ; p<.011\right)$.
A joint log-linear analysis of the four event items (meeting, dinner, Easter, and life) suggested that the model phrasing $\times$ order of items was the simplest model that fitted the data ( $G^{2}=39.69$; $d f=36 ; p=.309$ ), justifying the aggregation across these items. As in the order task, a pastwards directed FRONT occurred more frequently when participants had to specify the "front" of an event (as compared to the "back"), but this was the preferred response only when the task did not begin with the item on time in general (cf. Table 4). In the other cases, the majority of participants indicated that something like "front" does not exist. Among the two directions future- and pastwards, pastwards assignments clearly prevailed ( $87.2 \%$; 129 of 148 responses).

A joint log-linear analysis of the four time and units items (day, month, year, and time) suggested again that the model phrasing $\times$ order of items was the simplest model that fitted the data ( $G^{2} \geq 40.903$; $d f=36 ; p=.264$ ), justifying to aggregate the data across these items. The results were quite similar to those from the event items: The pastwards directed FRONT occurred more frequently when participants had to specify the "front" of an event (as compared to the "back"), but this was the preferred response only when the

Table 4: front assignments (\%) in the front task.

| Front (pointing) | Order of items |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Time item first |  | Time item last |  |
|  | Phrasing |  | Phrasing |  |
|  | Front | Back | Front | Back |
|  | Event items (meeting, dinner, Easter, life) |  |  |  |
| futurewards | - | 7.4 | 2.5 | 15.2 |
| pastwards | 39.8 | 30.9 | 73.8 | 20.3 |
| Nonexistent | 51.8 | 50.0 | 13.8 | 53.2 |
| Other | 8.4 | 11.8 | 10.0 | 11.4 |
|  | Unit items and time (day, month, year, time) |  |  |  |
| futurewards | 2.4 | 13.6 | - | 20.0 |
| pastwards | 35.7 | 18.2 | 81.3 | 12.5 |
| Nonexistent | 53.6 | 51.5 | 18.7 | 65.5 |
| Other | 8.3 | 16.7 | - | 5.0 |

task did not begin with the item on time in general. In the other cases, a majority of participants indicated that something like "front" does not exist. Among the two directions future- and pastwards, pastwards assignments again prevailed ( $81.3 \%$; 117 of 144 responses). This pattern includes the item on time in general. Futurewards directed FRONT assignment, which was the prevailing pattern in the eventmovement task ( $65.5 \%$ ) (cf. Table 2), occurred rarely when asked explicitly, and almost only when "back" had to be indicated ( 6 of 37 responses $=16.2 \%$ ).

Finally, we determined how consistently different responses were given across the whole set of items. To this end, we counted for each participant how often FRONT pointed futurewards, how often it pointed pastwards, and how often it was declared as nonexistent. Responses were fairly consistent. Participants gave the same type of response on 6.37 ( $79.6 \%$ ) of the 8 items. 29 participants ( $37.2 \%$ ) had a preference for a pastwards directed FRONT, and four participants (5.1\%) for a futurewards directed FRONT; 31 participants ( $39.7 \%$ ) were consistent in declaring that something like "front" or "back" does not exist; the remaining 14 participants ( $17.9 \%$ ) had no preference.

Taken together, the front task yields three results: First, the high number of participants indicating that something like FRONT or BACK does not exist for temporal entities is eye catching. Second, if FRONT was assigned to an entity at all, then it was assigned to its beginning and pointed towards the past. Finally, this tendency was also found for the item representing time in general.

## Discussion

Summarizing the findings across the three tasks presented here, the results indicate a preference among speakers of Norwegian for a futurewards orientation when "moving forward" an event (about 60\%), but a pastwards orientation when localizing earlier events as "before" later events (about $80 \%$ ). The latter is largely in line with the explicit assignment of FRONT to the beginning (rather than end) of events and time units-in fact, even to time itself-but assignments also depended on the order of items to some extent. These findings are surprising in at least three ways.

First, assignment of FRONT to temporal entities does not seem to follow the same principles across tasks, even though they were aimed at tapping the same underlying concepts. This is not unexpected per se, as people may have more than one timeline (Miles et al., 2011; and see Bender \& Beller, 2014, for a review of respective data). Here, the static versus dynamic nature of the tasks seems to make the difference: While the pastwards orientation prevails for fixed relations (revolving around the order or orientation of events), the futurewards orientation takes over when movement is involved. A similar pattern was observed for spatial referencing, in a task where participants had to pick the "front" token (from a set of several tokens) and move it "forward" by a given number of fields. In this case, FRONT was assigned to the token and the movement in diverging ways: closer to Ego for the former, and away from Ego for
the latter (Bender et al., 2012). Since movement itself provides orientation, it may serve as a direct source for FRONT assignment and thereby even override possibly conflicting orientations of the entities involved (Talmy, 2000). Interestingly, however, in the cases discussed here, the direction of movement is not specified beforehand, but is a consequence of FRONT assignment. This suggests that FRONT assignments follow different a priori preferences, but where these preferences are grounded in remains an open question.

Second, in the front task, time itself is treated similar to the smaller units year, month, and day, which themselves are treated similar to events in time. This appears at odds with the observation that for speakers of most languages (and especially English and related languages), FRONT is typically pointing towards the future (evidence summarized in Bender \& Beller, 2014). However, we hesitate to interpret our current data as strong evidence to the contrary for three reasons: The high proportion of "does not exist" responses observed for all items alike hints at the possibility that the phrasings (i.e., forside and bakside) have been infelicitous. Even if one were willing to assign a FRONT or BACK to a virtual, one-dimensional notion as time, assigning a whole front or back side may seem undue. In addition, the response options "beginning of time" and "end of time" may have evoked a notion of time that resembles an (excessively long) event rather than the superordinate field the item was meant to refer to. And finally, since the time question was embedded in questions on events and smaller time units, set effects may have led to an overgeneralization of assignment patterns that are applied to events.
The third way in which our findings are surprising concerns cross-linguistic patterns. What we found for speakers of Norwegian is more similar to previous findings on English than on Swedish-despite the fact that Norwegian and Swedish are much more closely related, and actually mutually understandable. Besides sharing almost identical proportions of the reflective versus translational variant of the relative FoR in the spatial domain (Beller et al., 2015; Beller \& Bender, 2017), speakers of English and Norwegian also exhibit a mix of preferences in the event-moving task, whereas speakers of Swedish strongly prefer the futurewards direction (Rothe-Wulf et al., 2015). Only in terms of relative preferences of the futurewards over the pastwards direction is Norwegian closer to Swedish.
Two conclusions may be drawn from these patterns. Most importantly, they indicate that cultural conventions are indeed crucial for establishing concrete frames of reference, as they determine how FRONT is assigned to temporal entities such as events or to time itself, both for fixed relations and for movement. In order to be able to identify which temporal FoRs people actually adopt, these conventions need to be assessed independently on an empirical basis. Furthermore, while FRONT assignment and FoR selection obviously differ across languages, it is not the languages themselves that are decisive here, but rather the agreement among their speakers, as attested to by the greater similarity of the Norwegian pattern with the English than the Swedish pattern.

This, we propose, renders the observed pattern a matter of negotiation and consensus, and hence a cultural phenomenon.

## Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft DFG through a grant for the project Spatial referencing across languages: Cultural preferences and cognitive implications to AB (Be 2451/13-1) and SB (Be 2178/7-1). For help with data collection, we thank Marleen Wilms.

## References

Beller, S., \& Bender, A. (2017). How relative is the relative frame of reference? Front and back in Norwegian, Farsi, German, and Japanese. Proceedings of the $39^{\text {th }}$ Annual CogSci Meeting.
Beller, S., Singmann, H., \& Hüther, L., \& Bender. A. (2015). Turn around to have a look? Spatial referencing in dorsal versus frontal settings in cross-linguistic comparison. Frontiers in Psychology, 6:1283, 1-17.
Bender, A., \& Beller, S. (2014). Mapping spatial frames of reference onto time: A review of theoretical accounts and empirical findings. Cognition, 132, 342-382.
Bender, A., Beller, S., \& Bennardo, G. (2010). Temporal frames of reference: Conceptual analysis and empirical evidence from German, English, Mandarin Chinese, and Tongan. Journal of Cognition and Culture, 10, 283-307.
Bender, A., Rothe-Wulf, A., Hüther, L., \& Beller, S. (2012). Moving forward in space and time: How strong is the conceptual link between spatial and temporal frames of reference (FoRs)? Frontiers in Psychology, 3:486, 1-11.
Boroditsky, L., \& Ramscar, M. (2002). The roles of mind and body in abstract thought. Psychological Science, 13, 185-188.
Grabowski, J., \& Weiß, P. (1996). Determinanten der Interpretation dimensionaler Lokalisationsäußerungen: Experimente in fünf Sprachen. Sprache \& Kognition, 15, 234-250.
Miles, L. K., Tan, L., Noble, G. D., Lumsden, J., \& Macrae, C. N. (2011). Can a mind have two time lines? Exploring space-time mapping in Mandarin and English speakers. Psychonomic Bulletin \& Review, 18, 598-604.
Núñez, R. E., \& Cooperrider, K. (2013). The tangle of space and time in human cognition. Trends in Cognitive Sciences, 17, 220229.

Núñez, R. E., \& Sweetser, E. (2006). With the future behind them. Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. Cognitive Science, 30, 401-450.
Levinson, S. C. (2003). Space in language and cognition. Cambridge: Cambridge University Press.
Majid, A., Bowerman, M., Kita, S., Haun, D., \& Levinson, S. C. (2004). Can language restructure cognition? The case for space. Trends in Cognitive Sciences, 8, 108-114.
McGlone, M. S., \& Harding, J. L. (1998). Back (or forward?) to the future: The role of perspective in temporal language comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 1211-1223.
Rothe-Wulf, A., Beller, S., \& Bender, A. (2015). Temporal frames of reference in three Germanic languages: Individual consistency, cultural consensus, and cross-linguistic variability. The Quarterly Journal of Experimental Psychology, 68, 917-939.
Talmy, L. (2000). Toward a cognitive semantics (Vol. 1): Conceptual structuring systems. Cambridge: MIT Press.

# A Bayesian model of knowledge and metacognitive control: Applications to opt-in tasks 

Stephen T. Bennett (stbennet @uci.edu)<br>Department of Cognitive Science<br>University of California, Irvine

Aaron S. Benjamin (asbenjam@illinois.edu)<br>Department of Psychology and Beckman Institute for Advanced Science and Technology University of Illinois at Urbana-Champaign

## Mark Steyvers (mark.steyvers@uci.edu)

Department of Cognitive Science
University of California, Irvine


#### Abstract

In many ecologically situated cognitive tasks, participants engage in self-selection of the particular stimuli they choose to evaluate or test themselves on. This contrasts with a traditional experimental approach in which an experimenter has complete control over the participant's experience. Considering these two situations jointly provides an opportunity to understand why participants opt in to some stimuli or tasks but not to others. We present here a Bayesian model of cognitive and metacognitive processes that uses latent contextual knowledge to model how learners use knowledge to make opt-in decisions. We leverage the model to describe how performance on selfselected stimuli relates to performance on true experimental tasks that deny learners the opportunity for self-selection. We illustrate the utility of the approach with an application to a general-knowledge answering task.


Keywords: metacognitive control; Bayesian cognitive model; wisdom of the crowd; opt-in; missing not at random

## Background

In traditional approaches to experimental psychology, an experimenter has unilateral control over which stimuli a participant experiences and the tasks that they complete. Yet in many real-world situations, such as providing ratings to videos on the Internet, the participant has some or even total control over the specific stimuli and tasks that they experience. The choice behavior underlying such self-selection is an important domain of study called metacognition (Nelson \& Narens, 1990), and the self-selection of activities or stimuli is specifically called metacognitive control (Fiechter, Benjamin, \& Unsworth, 2016; Finley, Tullis, \& Benjamin, 2010). Some work on monitoring and control processes in memory tasks focused on confidence judgments as an indicator of self-selection questions (Kelley \& Sahakyan, 2003; Koriat \& Goldsmith, 1996). It is unclear precisely how this self-selection is generated, however. To better understand metacognitive control behavior, a model is needed that accounts for performance on the task of interest as well as the choice behavior that leads participants to select only some stimuli for exposure, evaluation, or testing.

The major difficulty of such an endeavor is that participants select tasks according to their interests and expertise, and so
the data is missing in a nonrandom fashion (see Little \& Rubin, 2014, for a description of other missing data scenarios). Consequently, participants can only be compared and their performance fairly evaluated if a model is specified for the opt-in process. If a participant does not opt in to a particular question, then we simply do not see that participant's response to that question.

A starting point in explaining opt-in behavior is that participants have some meta-knowledge of what it is they already know, and use that knowledge effectively in service of ongoing learning. People provide higher assessments of their ability to answer inference questions in domains in which they have greater expertise (Bradley, 1981), and learners often choose to engage more effective study techniques for material that is more difficult for them (A. S. Benjamin \& Bird, 2006). Memory reports are also considerably more accurate when respondents have the option of withholding answers that they are unsure of or of titrating the grain size of their answers to their perceived accuracy (Goldsmith \& Koriat, 2007).

Self-regulated learning often has substantial benefits in educational contexts (Mezirow, 1981; Zimmerman, 1989; Boekaerts \& Minnaert, 1999; Paris \& Paris, 2001). Learners use meta-knowledge to allocate time, resources, and activities to an array of learning goals, and this application increases overall performance compared to learners who have their learning activities dictated by an instructor (Winne \& Hadwin, 1998; Finley et al., 2010).

The benefits of self-control extend beyond these constrained tasks, however. In causal reasoning experiments, participants can more quickly understand the causal structure of a network if they intervene in the learning process and design their own "experiments" (Steyvers, Tenenbaum, Wagenmakers, \& Blum, 2003; Sobel \& Kushnir, 2006; Lagnado \& Sloman, 2004). Human strategy selection can be explained in terms of rational metareasoning, wherein humans flexibly choose strategies in accordance with their environment (Lieder \& Griffiths, 2015; Lieder et al., 2014).

The core claim across each of these examples is that selfselection within a task aimed at measuring performance is
driven by metacognitive knowledge, which leads to a higher rate of success, expertise, or interest for the selected items. This process makes it difficult to evaluate the stimuli and the participants in an unbiased way. One test-taker may, for example, outperform another not because they have greater knowledge but rather because they make more a judicious selection of problems.


Figure 1: Outline of our modeling approach. Latent knowledge and design both explain the performance on the task. In the case of a subject-chosen design, latent knowledge also explains the design.

The aim of this project is to develop a cognitive model of the metacognitive aspect of item selection. In doing so, it also provides a framework to relate performance on selfselected materials with performance on an unconstrained set of items or stimuli. Here we apply this model to data collected from participants answering general knowledge questions, but the model is considerably more general: the same principles could apply in other metacognitive control tasks, such as study time allocation or selection of items for restudy. We are aware of one current model of metacognitive control, which takes as a given the state of the world, which then causes the observed behaviors (Fleming \& Daw, 2017). We take a different approach, which starts with the latent knowledge that the participant is coming to the experiment with and uses that in both the selection process behind opting-in and the observed responses to questions, as illustrated in Figure 1 . Performance on the task is explained by both the de-sign-that is, the particular experience of the participant in the task-and the latent knowledge of the participant. In the case of a participant who can opt in to certain questions but avoid others, the design is also partially informed by the latent knowledge. We are interested in estimating the latent knowledge of each participant and evaluating how it relates both to performance on the task and to opt-in behavior. In order to infer latent knowledge from the observed data, we apply Bayes' rule in the equation below, where $\theta$ is the latent knowledge, $c$ is the experimenter design, $d$ is the subject-chosen design, $x$ are the performance data from a true experimental design (where subjects respond to all or to a random subject of probes), and $y$ are the performance data from a subjectchosen design (in which subjects choose which probes to re-
spond to):

$$
\begin{equation*}
p(\theta \mid c, d, x, y) \propto p(x \mid \theta, c) p(y \mid \theta, d) p(d \mid \theta) p(\theta) \tag{1}
\end{equation*}
$$

In a traditional cognitive model, the important part of the model is the specification of $p(x \mid \theta, c)$ and $p(y \mid \theta, d)$, termed the likelihood functions. These functions directly explain the empirical effect of interest by relating latent knowledge to performance on the task given the experimental design. The novel part of the model relates to the specification of the metacognitive control process $p(d \mid \theta)$, which explains how the participant self-designs on the basis of their latent knowledge. If we would ignore this model component, we would likely, and incorrectly, conclude that participants who selfdesigned were more knowledgeable than participants subject to the experimenter's design because they outperformed their experimenter-designed counterparts. Such an error could be catastrophic if we were trying to compare across individuals or across tests. Because subjects are randomly assigned to conditions, it is highly unlikely that they differ widely. The process by which the participants who self-designed outperformed those who could not lies in the opportunity to selfdesign. Here we see the importance of jointly modeling the selection process and the task at hand in order to understand the interplay between latent knowledge, opt-in behavior, and performance.

Since this is a task in which many participants give judgments to many questions, we also expect to find that averaging across participants leads to higher accuracy-an effect termed the wisdom of the crowd (Surowiecki, 2004; Steyvers, Miller, Hemmer, \& Lee, 2009). Here we have the opportunity to evaluate whether the opportunity to opt in to a selfselected portion of the questions will enhance or attenuate such benefits associated with averaging. Certainly, many participants will gravitate towards the same questions when they can opt in, which would potentially decrease the benefits of averaging across a crowd by virtue of reducing input to the more difficult questions. However, based on what is known about metacognition, we expect that participants will opt in to questions for which they have relevant knowledge, which could lead to a more informed set of responses to average with the remaining crowd. Crowd behavior provides an additional benchmark against which we can evaluate the performance of the metacognitive model.

## Experiment

Stimuli The question set consisted of 100 general-knowledge binary choice questions. The questions were drawn from 12 topics: World Facts, World History, Sports, Earth Sciences, Physical Sciences, Life Sciences, Psychology, Space \& Universe, Math \& Logic, Climate Change, Physical Geography, and Vocabulary. The question set was created by collecting from multiple sources. Two example questions are shown in Table 1. Based on the empirically observed accuracy levels, the first is difficult and the second is easy.

Participants A total of 83 participants were recruited through Amazon Mechanical Turk (AMT). Each participant

Table 1: Example questions.

| Difficulty Example |  |
| :--- | :--- |
| Hard | The Sun and the planets in our Solar system all rotate in the same direction because: (a) they were all formed from <br> the same spinning nebular cloud, or (b) of the way the gravitational forces of the Sun and the planets interact <br> Greenhouse effect refers to: (a) gases in the atmosphere that trap heat, or (b) impact to the Earth's ozone layer |

was compensated $\$ 1$ for the 30 minutes the experiment was expected to take, and assigned to one condition.

Design Participants could view the survey description on AMT. If they selected the survey they were redirected to another website. They were first directed to a study information sheet which provided details of the survey and compensation. If they agreed to continue, they were instructed to answer some demographic questions. Participants were randomly assigned to either a random condition $(N=44)$ or a self-selection condition $(N=39)$, determining the subject's role in selecting which questions to answer. Participants were not aware of the existence of other conditions. Each participant saw the questions in 5 blocks of 20 questions each. In each block, they were instructed to rate the difficulty of each question and then, if they were assigned to the opt-in condition, instructed to choose 5 of those 20 questions to answer. The participants in the random assignment condition were randomly assigned 5 questions from that block to answer. After rating the difficulty of all 100 questions and answering 25 of them, participants were thanked for their time and given instructions on how to receive payment.

## Model

The model utilizes an IRT model to generate subjective latent knowledge (the belief of a participant that she can answer a question), which informs all aspects of participants' responses including the observed accuracy and difficulty ratings, as well as the metacognitive process of question selection. We describe participants as opting-in to questions for which they believe they have knowledge, answering with accuracy dependent on whether or not they believe they have knowledge, and giving lower difficulty ratings when they believe they have knowledge.

We use an IRT model to generate the subjective latent knowledge, $\delta_{i, j}$, for each participant $i$ (across both the optin and random condition) and question $j$,

$$
\begin{equation*}
\delta_{i, j} \sim{\operatorname{Bernoulli}\left(\operatorname{logit}^{-1}\left(\theta_{i}+\eta_{j}\right)\right), ~}_{\text {a }} \tag{2}
\end{equation*}
$$

where $\theta_{i}$ is the self-perceived skill of participant $i, \eta_{j}$ is the perceived familiarity of question $j$, and $\operatorname{logit}^{-1}(x)=\frac{e^{x}}{1+e^{x}}$. This latent knowledge is represented as a 0 or 1 , indicating whether or not that participant believes that she has knowledge for that question. We place a Normal prior on the selfperceived skill, $\theta_{i} \sim \operatorname{Normal}(0, \sigma)$, such that participants are expected to have the same skill (on average) for both the selfselection and random conditions.

For the self-selection condition, we assume that participants have a preference to select questions for which they believe they have knowledge. Let $c$ represent the observed question selections with $c_{i, j}=1$ if question $j$ was selected by participant $i$. For each participant and question block, we model question selection in the opt-in condition by a sampling process:

$$
\begin{equation*}
c_{i} \sim \operatorname{SampleWR}\left(\left(\delta_{i, 1}+\kappa, \ldots, \delta_{i, K}+\kappa\right), M\right) \tag{3}
\end{equation*}
$$

where $K$ is the total number of questions available for selection in each block ( $K=20$ in our experiment), $M$ is the number of questions that need to be selected ( $M=5$ in the experiment), SampleWR ( $\delta, M$ ) represents a sampling without replacement distribution where $M$ items are sampled with probability proportional to $\delta$, and $\kappa$ is a fixed parameter that controls the randomness in the selection process. Higher $\kappa$ values make it more likely that questions are selected for which the participant has no subjective knowledge. For participants in the random condition, we assume that the questions are randomly sampled by a process that is under control of the experimenter (where $M$ out of $K$ questions are randomly allocated).

Let $x_{i, j}$ represent the observed accuracy for participant $i$ on question $j$. We do not assume a fixed relationship between belief of knowledge and accuracy. For each question, we introduce guessing rate parameters $\rho_{j}$ and $\lambda_{j}$ that control the probability of correct responding if the participant does or does not have subjective knowledge about a question:

$$
\begin{equation*}
x_{i, j} \sim \operatorname{Bernoulli}\left(\delta_{i, j} \rho_{j}+\left(1-\delta_{i, j}\right) \lambda_{j}\right) \tag{4}
\end{equation*}
$$

For example, with $\rho=0.8$ and $\lambda=0.4$, the probability of a correct response is 0.8 if a participant has subjective knowledge, but 0.4 if the participant does not. The guessing parameters are given Beta priors, $\rho_{j} \sim \operatorname{Beta}(\alpha, \beta), \lambda_{j} \sim \operatorname{Beta}(\alpha, \beta)$ where $\alpha$ and $\beta$ are hyperparameters that control the variability in guessing rates across questions.

To model the difficulty ratings, we use an ordered logit model (Williams et al., 2006). We assume that subjective latent knowledge informs the perceived difficulty of questions. Questions for which the participant believes they have knowledge are perceived as easier. Let $\phi_{i, j}$ represent the perceived difficulty for participant $i$ on question $j$. We determine the perceived difficulty by:

$$
\begin{equation*}
\phi_{i, j}=-\delta_{i, j}-\eta_{j} \xi+\omega_{i}-\beta_{j}+\sigma_{i, j} \tag{5}
\end{equation*}
$$

where $\beta_{j}$ and $\omega_{i}$ capture participant and item level effects (e.g. some participants might find all items easy, some items
might be judged as easy) independent of subjective knowledge. In addition, we also allow the perceived familiarity of a question $\eta_{j}$ to affect the perceived difficulty weighted by a fixed scaling parameter $\xi$. Finally, $\sigma_{i, j}$ represent small perturbations centered around 0 to explain the random variability in difficulty ratings unrelated to any of the previous factors mentioned. These perceived difficulties feed into the ordered logit model to generate the difficulty ratings $r_{i, j}$,

$$
\begin{equation*}
r_{i, j} \sim \operatorname{OrderedLogit}\left(\phi_{i, j}, \tau_{i}\right) \tag{6}
\end{equation*}
$$

where $\tau_{i}$ is the set of criteria cutoffs for participant $i$.
We used JAGS to perform parameter inference. All parameters were inferred jointly from the opt-in and random condition. All model predictions were derived from posterior predictives where we simulate new participants from the distribution and assess how they self-select from a new set of questions.

## Results

We examine several empirical effects within the data and observe that the model captures the appropriate trend in most cases.

Item selection and latent knowledge. The model captures the expected relationship between opting-in behavior and knowledge (see Figure 2). Participants were more likely to select questions for which they had pre-existing knowledge. Each question was randomly assigned to at least four participants in the random assignment condition. However, in the opt-in condition, there were seven questions that no participant chose to answer. Question selection strongly corresponded with the inferred latent knowledge ( $\delta_{i, j}$ ) for the participant-question pair, with participants choosing questions for which they had latent knowledge. Across conditions, latent knowledge is distributed in a similar manner: most participants have knowledge for popular questions, few participants have knowledge for unpopular questions, and some participants are more knowledgeable than others. However, the model has substantially more certainty about the localization of this knowledge in the opt-in condition compared to the random condition because it can leverage the opt-in behavior. In Figure 2, this certainty is expressed as black or white squares, while uncertainty is represented in gray. We see the uncertainty about which participants have knowledge for which question as a "blurring" of the latent knowledge space.

Effect of opting-in on participant performance. Average performance across questions was higher in the self-selection condition ( $86.05 \%$ ) than in the random condition ( $67.27 \%$ ). We computed a Bayes Factor (BF) given a binomial distribution with a shared or different rate of correct responding and find a $\log _{10} \mathrm{BF}$ of 21.12 in favor of a higher rate of correct responding in the opt-in condition. This corresponds to decisive evidence that average accuracy is higher in the opt-in condition than the random assignment condition. This occurs even when taking into account the fact that people tend to opt


Figure 2: Latent knowledge is similar between conditions and corresponds to opting-in behavior. Plotted are the opt-in behaviors and average $\delta_{i, j}$ values across conditions, all sorted by the popularity of the question in the opt-in condition. White corresponds to questions that the participant opted in to or the inferred presence of knowledge.
in to easier questions. In order to perform this analysis, we took the product of the evidence that performance is higher in the opt-in condition than the random assignment condition for each question and find a $\log _{10} \mathrm{BF}$ of 9.02. So, even when comparing on an item-by-item basis, opting-in provides an advantage.

Effect of opt-in on model performance. For the model, the average accuracy for posterior predictive samples in the self-selection condition (mean $=79.03 \%$ ) is also significantly higher than in the random condition ( mean $=67.07 \%$ ), both across all questions ( $99.86 \%$ of samples) and even within questions ( $68.93 \%$ of sample-question pairs). We observe this benefit in accuracy despite the average inferred ability of individual subjects $\left(\theta_{i}\right)$ being equivalent across conditions: $\bar{\theta}_{i}=0.00, \mathrm{SD}=0.99$ in the opt-in condition versus $\bar{\theta}_{i}=-0.09, \mathrm{SD}=2.04$ in the random assignment condition. This means that the benefit to accuracy that the model predicts is due to downstream consequences of the metacognitive selection process and not an (inaccurate) inference that participants in one condition were more skillful than in the other.

Difficulty Ratings. Participants tended to give lower average difficulty ratings to questions that they opted in to ( $\log _{10}$ $\mathrm{BF}=91.89$ ) and higher average difficulty ratings to questions that they did not opt in to $\left(\log _{10} B F=64.09\right)$, relative to the random condition. The model captures, but understates, this trend (see Figure 3).

Wisdom of the crowd. The left panel of Figure 4 shows the relationship between crowd size and crowd accuracy for the two conditions in the experiment, as well as a hybrid condition in which the two groups are combined. The right side of the Figure shows that the model captures this effect qualitatively. Crowd responses were determined by taking the most


Figure 3: Distribution of difficulty ratings for participants and model for questions that were selected or not selected in the opt-in condition and the random condition. Lower ratings indicate lower perceived difficulty.
common response across the participants in the crowd. Since seven questions went unanswered in the opt-in condition, we had to consider how unanswered questions impacted crowd performance. To treat the self-selection condition maximally conservatively, we graded any question that went unanswered as incorrect for that crowd. Even with this penalty, the crowd composed of the participants from the self-selection condition (79\%) outperformed the crowd of subjects from the random condition (73\%).

We also considered the impact of crowd size on performance. To do this, we evaluated the average performance of crowds composed of random samples of participants from a condition and varied the number of participants drawn to form the sample. We plot average crowd performance as a function of the total number of judgments, where a judgment is a person's response to a question. The hybrid condition provides a means of improving upon both conditions. To create a hybrid crowd, we first sampled participants that answered the question from the opt-in condition. If a question had no responses, we added the answer from one participant in the random condition in order to guarantee that all questions received at least one answer. This hybrid crowd has high performance across all questions. The model captures the general trends in the data in that larger crowds result in higher crowd accuracy, opt-in crowds outperform random-assignment crowds, and the hybrid crowds perform well across all questions.

Additional simulations. Given our model, we investigated which circumstances would likely lead to changes in the relative performance of the self-selection and random conditions in terms of both average overall accuracy and crowd performance. We varied the heterogeneity of perceived question difficulty $\left(\eta_{j}\right)$ and latent ability $\left(\theta_{i}\right)$. We did this by simulating experiments in which we varied the underlying hyperparameter corresponding to the variability of $\theta_{i}$ and $\eta_{j}$ by factors of $0.25,1$, and 4 while keeping other parameters constant


Figure 4: Crowd performance when varying the number of participants (measured by the total number of judgments)
(see Figure 5). We find that increasing the heterogeneity of perceived question difficulty increases self-selection accuracy overall, but decreases it at the crowd level since participants tend to avoid answering the same difficult questions. Heterogeneity in question difficulty does not have an appreciable impact on performance in the random condition. In both conditions, higher heterogeneity of participant skill leads to higher crowd performance and gives resilience to heterogeneously difficult questions in the opt-in condition. However, it detracts from overall accuracy in the self-selection condition.


Figure 5: Simulated performance depending on variability of question difficulty $\left(\eta_{j}\right)$ and participant skill $\left(\theta_{i}\right)$.

## Conclusions

A comprehensive model of cognition must make allowance for the fact that cognitive behavior is driven by motivations. We choose what we attend to and attempt to encode, and what we attempt to remember. Metacognitive behavior is at the heart of most learning outside the laboratory, and a fair amount within it as well (A. Benjamin \& Ross, 2008). The joint modeling of metacognitive behavior-like self-selection
of items-along with cognitive performance has the potential to address a wider and more representative range of realworld learning and testing behaviors, and can serve as the basis for drawing comparisons across individuals or tests that would otherwise be hopelessly confounded. Additionally, the model could be extended to explain various incentives given to the participant, which would impact how latent knowledge interacts with the task to generate opt-in behaviors. The model presented here provides a starting point for such an enterprise. It leads to a relatively good description of performance across a variety of metrics. A single latent knowledge state for each participant-question pair permits an explicit representation of the metacognitive process that governs the relationship between opt-in, accuracy, and difficulty behaviors. The model is successful in describing the nonrandom missing nature of the data that we observed by relying on principled psychological theories about why someone might choose one question over another.

An additional lesson of the current research can be seen in the crowd data. Opting in is generally beneficial to crowd accuracy in both the observed data and our model. This result indicates that the metacognitive skill of the individuals in self-selection can be leveraged in order to create a smarter crowd. This effect is sufficiently robust that it appears to outweigh the cost associated with small crowd sizes for some questions or no volunteered responses at all for a small number of questions. Such a result is particularly important when considering the widespread availability of datasets in which responses are self-selected.

## References

Benjamin, A., \& Ross, B. H. (2008). The psychology of learning and motivation: Skill and strategy in memory use (Vol. 48). Academic Press.
Benjamin, A. S., \& Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. Journal of Memory and Language, 55(1), 126-137.
Boekaerts, M., \& Minnaert, A. (1999). Self-regulation with respect to informal learning. International journal of educational research, 31(6), 533-544.
Bradley, J. V. (1981). Overconfidence in ignorant experts. Bulletin of the Psychonomic Society, 17(2), 82-84.
Fiechter, J. L., Benjamin, A. S., \& Unsworth, N. (2016). 16 the metacognitive foundations of effective remembering. The Oxford Handbook of Metamemory, 307.
Finley, J. R., Tullis, J. G., \& Benjamin, A. S. (2010). Metacognitive control of learning and remembering. In New science of learning (pp. 109-131). Springer.
Fleming, S. M., \& Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. Psychological Review, 124(1), 91.

Goldsmith, M., \& Koriat, A. (2007). The strategic regulation of memory accuracy and informativeness. Psychology of learning and motivation, 48, 1-60.

Kelley, C. M., \& Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. Journal of Memory and Language, 48(4).
Koriat, A., \& Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. Psychological review, 103(3), 490.
Lagnado, D. A., \& Sloman, S. (2004). The advantage of timely intervention. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(4), 856.
Lieder, F., \& Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In Cogsci.
Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., \& Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In Advances in neural information processing systems (pp. 2870-2878).
Little, R. J., \& Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley \& Sons.
Mezirow, J. (1981). A critical theory of adult learning and education. Adult education quarterly, 32(1), 3-24.
Nelson, T. O., \& Narens, L. (1990). The psychology of learning and motivation. Metamemory: A theoretical framework and new findings.
Paris, S. G., \& Paris, A. H. (2001). Classroom applications of research on self-regulated learning. Educational psychologist, 36(2), 89-101.
Sobel, D. M., \& Kushnir, T. (2006). The importance of decision making in causal learning from interventions. Memory \& Cognition, 34(2), 411-419.
Steyvers, M., Miller, B., Hemmer, P., \& Lee, M. D. (2009). The wisdom of crowds in the recollection of order information. In Advances in neural information processing systems (pp. 1785-1793).
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., \& Blum, B. (2003). Inferring causal networks from observations and interventions. Cognitive science, 27(3), 453489.

Surowiecki, J. (2004). The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economics, society and nations. Little, Brown.
Williams, R., et al. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. Stata Journal, 6(1), 58.
Winne, P. H., \& Hadwin, A. F. (1998). Studying as selfregulated learning. Metacognition in educational theory and practice, 93, 27-30.
Zimmerman, B. J. (1989). A social cognitive view of selfregulated academic learning. Journal of educational psychology, 81(3), 329.

# Modality Switch Effects Emerge Early and Increase throughout Conceptual Processing: Evidence from ERPs 

Pablo Bernabeu (pcbernabeu @gmail.com)<br>Centre for Language Studies, Radboud University<br>Erasmusplein 1, 6525 HT - Nijmegen, Netherlands<br>Roel M. Willems (r.willems@let.ru.nl)<br>Neurobiology of Language dept., Max Planck Institute for Psycholinguistics<br>Wundtlaan 1, 6525 XD - Nijmegen, Netherlands<br>Max M. Louwerse (m.m.louwerse@tilburguniversity.edu)<br>Tilburg center for Cognition and Communication, Tilburg University<br>Warandelaan 2, 5037 AB - Tilburg, Netherlands


#### Abstract

We tested whether conceptual processing is modality-specific by tracking the time course of the Conceptual Modality Switch effect. Forty-six participants verified the relation between property words and concept words. The conceptual modality of consecutive trials was manipulated in order to produce an Auditory-to-visual switch condition, a Haptic-tovisual switch condition, and a Visual-to-visual, no-switch condition. Event-Related Potentials (ERPs) were time-locked to the onset of the first word (property) in the target trials so as to measure the effect online and to avoid a within-trial confound. A switch effect was found, characterized by more negative ERP amplitudes for modality switches than noswitches. It proved significant in four typical time windows from 160 to 750 milliseconds post word onset, with greater strength in posterior brain regions, and after 350 milliseconds. These results suggest that conceptual processing may be modality-specific in certain tasks, but also that the early stage of processing is relatively amodal.


Keywords: conceptual processing; time; modality switch; perceptual simulation; amodal; event-related potentials; ERP

## Introduction

Research in the cognitive sciences has extensively investigated whether conceptual processing is modalityspecific (Barsalou, 2016). In a commonly used paradigm known as the Conceptual Modality Switch (CMS), participants perform a property verification task in which they decide whether certain property words can reasonably describe certain concept words. For instance, Pecher, Zeelenberg, and Barsalou (2003) presented sentences such as Blenders can be loud. Covertly, the conceptual modality of consecutive trials was manipulated in order to produce specific switches. A sentence like Blenders can be loud, which is mainly related to the auditory modality, could either be followed by a sentence within the same modalitye.g., Leaves can be rustling-, or by a sentence in a different modality-e.g., Cranberries can be tart (gustatory). Pecher et al. found that when the modalities of consecutive trials did not match, participants took longer to respond. Such an effect suggested that perceptual features of concepts (operationalized in the modality shifts) are
accessed during conceptual processing. More recently, however, the CMS effect was reanalysed using a nonperceptual alternative, language statistics (i.e., how words co-occur in a language). Louwerse and Connell (2011) found that language statistics were able to approximately predict what modality a concept and property pair belonged to. Specifically, they could predict a visual/haptic modality, an olfactory/gustatory modality, and an auditory modality, but they could not predict the subtler differences between visual and haptic, and between olfactory and gustatory, which seemed to be reserved for perceptual simulations. Moreover, when a language statistics explanation and a perceptual explanation were compared against one another, faster response times (RTs) were best explained by language statistics, whereas slower RTs were best explained by perceptual simulations (for similar findings with switches in emotion, see Tillman, Hutchinson, Jordan, \& Louwerse, 2013). Louwerse and Hutchinson (2012) further replicated these findings in an Electroencephalography (EEG) experiment in which they showed that those cortical regions commonly associated with language processing are relatively more active in the beginning of processing, whereas those regions commonly associated with perceptual processing are relatively more active later on. These studies demonstrated that the time course of processing is important in the study of language statistics and perceptual simulation.

## Time Course of Effects in Word Processing

The time course of word processing may be relevant for an effect such as the CMS. Hauk (2016) zooms into the one second during which a word is processed, proposing the following timeline. A reader or listener starts to identify a word and to access part of its meaning within around 150 milliseconds (ms) from word onset. Building on that information, working memory processes emerge at around 170 ms post word onset, followed by response-related processes at around 250 ms . Mental imagery and episodic memory are the last-emerging processes, both around 400 ms post word onset. Once started, each of these processes extends further, gradually overlapping with each other. This
timeline suggests two important things about the earlier and the later stages of word processing. First, having an early emergence-e.g., at 250 ms -does not make an effect lexicosemantic per se because the meaning encoded could have gone through working memory before activating the actual system of interest, e.g., sensorimotor (Mahon \& Caramazza, 2008). Second, it suggests that effects emerging later face that same challenge and, in addition, the potential influence of response-related and forthcoming processes.

If even an effect measured online with high temporal resolution (EEG or Magnetoencephalography) may be subject to alternative causation, effects measured with lower temporal resolution (functional Magnetic Resonance Imaging) or off-line (RT) are arguably more challenged by this. With regard to the current topic specifically, sensory and motor effects may possibly be epiphenomenal to the representation of concepts online, independently of what is suggested by the measurements off-line or online lagged.

There is a technique especially apt for testing the causality of cognitive systems, namely, Transcranial Magnetic Stimulation (TMS). Willems et al. (2011) found that comprehension of hand-related verbs was improved as a consequence of stimulating the hand area of premotor cortex. It was particularly improved when the TMS was applied in the hemisphere controlling the dominant hand. More recently, Vukovic et al. found an impairment in the processing of action-related words, along with an improvement in the processing of abstract words, after TMS was applied over motor cortex, 200 ms after word onset. The latter finding suggests that the contribution of modalityspecific systems (in this case, motor ones) can emerge relatively early (see also Amsel, Urbach, \& Kutas, 2014; Van Dam, Brazil, Bekkering, \& Rueschemeyer, 2014).

In our view, two interlocked questions stand out in the current topic area. The causality question asks whether modality-specific effects reflect a functionally relevant simulation process or arise only after basic conceptual processing has been attained. The compatibility question asks whether different processing systems, amodal and modal, may compatibly operate in conceptual processing.

## Experiment

We addressed the causality and the compatibility questions by revisiting the CMS paradigm (see most recent previous study in Scerrati, Lugli, Nicoletti, \& Borghi, 2016).

## Tracking the Time Course of the CMS

We measured the CMS online by time-locking EventRelated brain Potentials (ERPs) to the onset of the first word in the target trials. We wanted to establish where exactly the effect-indexing access to perceptual informationemerged, how far it extended, and the relative strength over the time course. These measures would allow us to relatively assess how strongly the CMS may be influenced by response-related and other extra-semantic processes (see Hauk, 2016). Concerning the compatibility question, previous research would predict an increase in the CMS
effect over time because earlier processing is relatively amodal (Louwerse \& Hutchinson, 2012).

The three previous ERP studies on the CMS time-locked the measurement to property words placed last in the target trials (Hald, Marshall, Janssen, \& Garnham, 2011; Collins, Pecher, Zeelenberg, \& Coulson, 2011; Hald, Hocking, Vernon, Marshall, \& Garnham, 2013). A potential problem of those measurements is a lack of certainty on the emergence of the effect, because a switch might reasonably emerge already at the first content word in the target trial. Therefore, in our design we placed the property word first in the target trial, and time-locked ERPs to its onset. This had an important advantage, as it helped avoid a confound caused by the relation between the property and the concept in each target trial (see Hald et al., 2013). The possibility of those two confounds-the lagged measurement and the within-trial relationship-could explain why the CMS effect has sometimes failed to appear in RTs (Hald et al., 2011; 2013; Collins et al., 2011; Scerrati et al., 2016).

We did not have clear hypotheses on what we would find as the time course of the CMS because we were the first to time-lock ERPs to the first word. Nonetheless, the effects found in the three ERP studies cited above were generally characterized as N400-linked to semantic violation-, with more negative amplitudes for modality switches than noswitches. The earliest emerging effect appeared in Hald et al. (2011), in a time window from 270 to 370 ms .

## Different Switches and Processing Speeds

In order to further explore the compatibility question, we drew on Louwerse and Connell (2011). As reviewed in the Introduction, they found that quick processing was able to pick up most switch types but missed the subtler ones, for instance, between haptic and visual. By contrast, slow processing had the advantage of picking up even those subtler switches. Here we brought these findings to a group design. We distinguished a Quick group of participants and a Slow group of participants based on their average RT. Maintaining the CMS as a within-subjects factor, we predicted that the larger modality switches (e.g., auditory to visual) would be picked up equally by both groups, whereas the subtler switches (e.g., haptic to visual) would be picked up only-or more clearly-by the Slow group.

## Method

Accuracy Pretest The task was validated in a behavioural pretest ( $N=19$; Radboud U., Tilburg U.) revealing that all participants but one had an average response accuracy over $50 \%$, and the overall average was $63 \%$ ( $S D=48 \mathrm{pp}$.).

Participants Forty-nine participants-native speakers of Dutch with no relevant disorders-were recruited at the Max Planck Institute for Psycholinguistics. They were paid a small fee after participating. Participants were randomly assigned to one of three experimental groups: a Quick response group ( $n=22$ ), a Self-paced response group ( $n=$ 21), and a Null group who got the same experimental design
as the Self-paced group but no instructions on response speed ( $n=5$ ) (see Figure 1). One participant had to prematurely leave the experiment. Another participant had to be removed from the data due to too noisy ERPs (7 retained trials out of 108). Under visual inspection, all other participants' waveforms-preprocessed and averaged per CMS condition-approximately presented the typical peaks of word reading. Last, one participant, the only one with an accuracy below $50 \%$-i.e., $37 \%$-, was also removed from further analyses. Forty-six participants remained. Because
the Groups presented rather close, significantly equal RTs, we pooled them together and re-split them in two groups on the basis of each participant's average RT. The effects CMS and CMS by Group were equally significant with the old and new groups in both ERPs and RTs. New groups were: Quick ( $n=23$; mean RT $=568.40 \mathrm{~ms}, S D=104.83$; age 1931, mean $=23.3$; 19 females), and Slow ( $n=23$; mean RT $=$ $937.21 \mathrm{~ms}, S D=265.56$; age $18-25$, mean $=22.2 ; 18$ females). The different tests-stimulus norming, pretest, and main experiment - did not share any participants.


Figure 1: Schema illustrating materials, design, and procedure. Note that Groups were pooled and re-split (see Participants).

Materials and Design As in previous CMS studies, the stimuli consisted of pairs of property and concept words, but we had a small novelty in this combination. Whereas previous studies presented the concept and the property in declarative sentences (Pecher et al., 2003; Louwerse \& Connell, 2011; Hald et al., 2011), or with the concept followed by the property alone (Collins et al., 2011), the current experiment presented the property followed by the concept alone, e.g., Soundless Answer. In this design, as in most, the property word took the most relevant position for the measurement, because properties are generally more modality-specific than concepts (Lynott \& Connell, 2013).

The properties and concepts, all in Dutch, were partly based on Lynott and Connell's (2009, 2013) norms. We normed our items similarly too, by asking forty-two respondents to rate 0 to 5 the extent to which they experienced each property or concept with the senses of hearing, touch, and vision. Then we computed the dominant modality of each word (Bernabeu, Louwerse, \& Willems, in prep.). Next, we created 216 trials by joining properties and concepts within the same modalities. ${ }^{1}$ Half of the trials contained a fairly related property and concept, while the other half presented rather unrelated pairs. These

[^280]relationships served to engage participants in a semantic task, yet conveniently did not affect the ERP measurement because ERPs were measured before the concept word was presented in each trial (Figure 1). In spite of this, we wish to acknowledge that some trials came out rather unnaturalLukewarm Volume-or fuzzy-Solid Ideal-because they were created out of a fixed set of modality norms (Bernabeu et al., in prep.). In order to alleviate that problem, the instructions of the experiment stated that the accuracy feedback following every response was based on the answers of previous participants (in reality, it wasn't). Furthermore, the stimuli and the task were validated by the accuracy rates in the pretest and in the main experiment.

For the critical CMS manipulation, trials were covertly paired as context and target trials. This was done pseudorandomly within participants and CMS conditions by using the software Presentation ${ }^{\mathrm{TM}}$. Three conditions were created-Auditory-to-visual, Haptic-to-visual, and Visual-to-visual-, each with 36 context trials and 36 target trials. One auditory-to-visual switch for one participant was: Soundless Answer | Bumpy Wage (bold added to ERP-target word). For another participant, the latter target trial was instead preceded by the context trial Loud Welcome. The pseudo-randomization ensured that ERP-target words (properties) were matched across CMS conditions on the
essential criteria-word frequency and length (letters), and semantic class. Also, ERP-target words occurred only once.

Procedure The entire experiment was in Dutch. By means of written instructions, participants were asked to respond in each trial whether the first word, a property, could be used to describe the second word, a concept. Two buttons were used to respond. An example was provided based on the property 'grey' and the concept 'snow.' Snow is often white, but it can also be grey. By contrast, a property that would not match is 'pink.' Then, the instructions diverged for the different groups of participants: while the Quick group was asked to respond as quickly as possible in every trial, the Self-paced group was asked to respond self-paced, and the null group was altogether unconstrained (see design constraints for each group in Figure 1). Further, the instructions stated that feedback would be provided for each response, and that this was based on all preceding answers (although it was not), and therefore participants need not worry too much about mistakes. Last, they were asked to move or blink as little as possible, and do so only while the cross was on the screen. Twelve practice trials ensued, after which participants could ask questions. The experiment ran on Presentation ${ }^{\mathrm{TM}}$. The experiment proper lasted about 20 minutes, with a break in the middle. Taking into account EEG procedures, it lasted about 1 hour and 45 minutes.

ERP Recording and Preprocessing The EEG signal was recorded with BrainVision Recorder $1^{\mathrm{TM}}$, in differential mode, utilizing 65 active $\mathrm{Ag} / \mathrm{AgCI}$ electrodes. The ground electrode was positioned just above the nose, at the glabella. Three other electrodes were used to register eye movements, two placed at the outer canthi of each eye, and one placed below the left eye. The remaining 59 electrodes were mounted in a custom, equidistant ActiCap cap. Impedance was kept below $10 \mathrm{k} \Omega$ by applying electrolyte gel at the tip of each electrode. The signal was amplified through BrainAmp DC amplifiers with a bandpass filter of 0.016100 Hz , and an online sampling frequency of 500 Hz (i.e., every 2 ms ). Afterwards, the signal was preprocessed in BrainVision Analyzer $2^{\text {TM }}$, with the following steps: CMS condition segmentation, automatic ocular correction, 200 ms baseline correction, artefact rejection via semi-automatic segment selection. ${ }^{2}$ The proportion of segments (trials) retained from the 46 final participants was: $77.4 \%$ in the Visual-to-visual condition, $78.0 \%$ in the Haptic-to-visual condition, and $78.6 \%$ in the Auditory-to-visual condition.

[^281]ERP Analysis The ERPs, averaged per CMS condition, were downsampled to 125 Hz due to computational demands. Electrodes were divided into an anterior and a posterior area (also done in Hald et al., 2011). Albeit a superficial division, we found it sufficient for the research question. Time windows were selected as in Hald et al., except for the last window, which was extended up to 750 ms post word onset, instead of 700 ms , because the characteristic component of that latency tends to extend until then, as we confirmed by visual inspection of these results. Window 1 was meant to capture N1-P2 components, window 2 the pre-N400, window 3 the N 400 , and window 4 the LPC/P600. Analyses were performed in the software R.

## Results

All final participants responded correctly in over half of the trials. The average accuracy was $63 \%$ ( $S D=48 \mathrm{pp}$.), nearly identical in each participant Group and CMS condition. ${ }^{3}$

ERPs The ERP results revealed a CMS effect from time window 1 on, larger after 350 ms . It appeared with both switch conditions, and was characterized by a more negative amplitude for the switch conditions compared to the noswitch condition. In certain parts over the time course, the effect appeared in both anterior and posterior areas, and in both participant groups, but it was generally stronger in the posterior area and in the Slow group (Figure 2).

The ERPs per window were analyzed with Linear Mixed Effects models (lmer R package). Random intercepts and slopes, and fixed effects, were tested with the critical factors and interactions, as well as with potential confounds, e.g., handedness, sex, age. Each inclusion was tested in a stepwise fashion based on the significance of the Likelihood Ratio. The final models presented good fits, with $R^{2}$ ranging from .748 (time window 4) to 862 (time window 2). Table 1 sums up the results. First, the CMS effect in time window 1 was confirmed significant (see detailed waveforms in Figure 3). Such an early emergence is unprecedented in the CMS literature, and it may have been enabled by the timelocking of ERPs to the first word in target trials. In this time window, the only process not lexicosemantic is possibly working memory (Hauk, 2016), and therefore this early emergence lends support to the possibility that the CMS had a lexicosemantic basis (but see Mahon \& Caramazza, 2008).

Whereas in time window 1 ( $160-216 \mathrm{~ms}$ ), the CMS effect was circumscribed to an interaction with Brain Area (anterior/posterior), by time window $2(270-370 \mathrm{~ms})$ a main effect of CMS emerged. Finally, in window 3 ( $350-550 \mathrm{~ms}$ ) and window $4(500-750 \mathrm{~ms})$, the only critical effect was CMS. Window 3 presented the largest main effect of CMS. Planned ANOVA contrasts into CMS conditions, corrected for multiple comparisons, revealed that the no-switch condition differed significantly from the switch conditions.

[^282]

Figure 2: Data per Group and Area, with $95 \%$ Confidence Intervals every 2 ms , and time windows. Negative up.

Table 1: Effect of CMS and its interaction with Ant./Pos. brain Area and with Group. ${ }^{* * *} p<.001 ;{ }^{* *} p<.01 ;{ }^{*} p<.05$.

| Time window | Factors | Effect: $\chi^{2}$ |
| :--- | :--- | :--- |
| $1) 160-216 \mathrm{~ms}$ | CMS | 1.40 |
|  | CMS x Ant/Pos Area | $48.59^{* * *}$ |
|  | CMS x Ant/Pos Area x Group | $23.63^{* *}$ |
| $2)$ | CMS | $6.40^{*}$ |
|  | CMS x Ant/Pos Area | $10.89^{* *}$ |
|  | CMS x Ant/Pos Area x Group | $4.13^{* *}$ |
| 3) $350-550 \mathrm{~ms}$ | CMS | $9.47^{* *}$ |
| 4$) 500-750 \mathrm{~ms}$ | CMS | $7.58^{*}$ |

By contrast, the switch conditions hardly differed from each other-statistically equal in some sections of the data-, fitting the CMS effect. The fit of these follow-up ANOVAs was high in time windows 1 to 3 , and medium in window 4 .

Although the interaction of Group and CMS was only significant in time windows 1 and 2, the waveforms in windows 2 , 3 , and 4 presented a pattern that precisely fitted


Figure 3: Subset of electrodes from the Slow group at time window 1 (the Quick group presented a slightly smaller but also significant effect). Y-axis ranges from $-1 \mu \mathrm{~V}$ to $+4 \mu \mathrm{~V}$. Red labels signal the equivalents in the 10-20 montage.
our predictions based on Louwerse and Connell (2011). Whereas the Slow group picked up the switches across all modalities similarly, the Quick group picked up the Auditory-to-visual switch more clearly than the Haptic-tovisual switch, fitting with an amodal-modal compatibility.

RTs This design was tailored to measure ERPs. RTs were not reliable enough regarding the CMS because the last word in the target trials-critical for RTs-had not been matched across conditions on the essential criteria (see Materials section above). Nonetheless, we analysed RTs, statistically controlling for the confounds. No effects involving CMS were found, all $p \mathrm{~s}<.05$ (model $R^{2}=.552$ ).

## Discussion

CMS effects are a well-known, replicated demonstration of the relevance of modality-specific information for conceptual processing. In the current study, we tracked this effect online in order to ascertain at what stages perceptual information is processed, and in what degree (see Mahon \& Caramazza, 2008; Hauk, 2016). Time-locking ERPs to the onset of the first word in the target trials brought the added advantages of cancelling confounds within the target trial and measuring the effect at the onset, un-lagged. On the other hand, this design had the disadvantage of some unnatural stimuli. In spite of these novelties, though, our broad randomization of trials and the results found suggest that this experiment preserved the essence of the CMS paradigm. We found the CMS effect emerging at the start of lexicosemantic and working memory processing, then
increasing through the rest of word processing. The virtually immediate effect upon word recognition offers further support for the suggestion that sensory brain regions have a functional role in conceptual processing, at least in a fairly demanding semantic analysis as in the current task (see Louwerse \& Hutchinson, 2012). Solving the causality question, nonetheless, may require in the future more fundamental research on word processing, in addition to TMS-based work, in order to qualify the degree of semantic and post-semantic processing in an effect (see Hauk, 2016). The increase in the CMS effect over the time course converges with previous findings in suggesting that distributional processing-language statistics-may play a greater role earlier on (Louwerse \& Connell, 2011; Louwerse \& Hutchinson, 2012). This early-late distribution fits with Hauk's (2016) word processing timeline, where the early stage has a greater relative proportion of lexicosemantic processing, which would presumably support language statistics. Increasing evidence on the compatibility of amodal and modal/embodied processing invites further research. Concerning the CMS specifically, we still need to establish whether this effect can best be explained by language statistics or by perceptual simulations. The current work at least demonstrates that it emerges early and increases throughout word processing.

## Supplementary Material, Acknowledgments

All data, including the stimuli, set-up, raw files, and results in detail, are available through https://osf.io/97unm/

We received funding from the Neurobiology of Language department at Max Planck Institute for Psycholinguistics, from the Experimental Psychology Society, and from the Cognitive Science Society. We received valuable advice from Gwilym Lockwood, Johan Weustink, Monique Flecken, Ronald Fischer, Sean Roberts, Brain Products staff, and the reviewers. Also, thanks to the participants.

## References

Amsel, B. D., Urbach, T. P., \& Kutas, M. (2014). Empirically grounding grounded cognition: the case of color. Neuroimage, 99, 149-157.
Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. Psychonomic Bulletin \& Review, 23.
Bernabeu, P., Louwerse, M. M., \& Willems, R. M. (2017, April 11). Modality exclusivity norms for 747 properties and concepts in Dutch: a replication of English http://doi.org/10.17605/OSF.IO/BRKJW
Collins, J., Pecher, D., Zeelenberg, R., \& Coulson, S. (2011). Modality switching in a property verification task: an ERP study of what happens when candles flicker after high heels click. Frontiers in Psychology, 2.
Hald, L. A., Hocking, I., Vernon, D., Marshall, J.-A., \& Garnham, A. (2013). Exploring modality switching effects in negated sentences: further evidence for grounded representations. Frontiers in Psychology, 4, 93.
Hald, L. A., Marshall, J.-A., Janssen, D. P., \& Garnham, A. (2011). Switching modalities in a sentence verification
task: ERP evidence for embodied language processing. Frontiers in Psychology, 2.
Hauk, O. (2016). Only time will tell-Why temporal information is essential for our neuroscientific understanding of semantics. Psychonomic Bulletin \& Review, 23, 4, 1072-1079.
Louwerse, M., \& Connell, L. (2011). A taste of words: linguistic context and perceptual simulation predict the modality of words. Cognitive Science, 35, 2, 381-98.
Louwerse, M., \& Hutchinson, S. (2012). Neurological evidence linguistic processes precede perceptual simulation in conceptual processing. Frontiers in Psychology, 3, 385. doi:10.3389/ fpsyg. 2012.00385
Lynott, D., \& Connell, L. (2009). Modality exclusivity norms for 423 object properties. Behavior Research Methods, 41, 2, 558-564.
Lynott, D., \& Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behavior Research Methods, 45, 516-526.
Mahon, B. Z., \& Caramazza, A. (2008). A critical look at the Embodied Cognition Hypothesis and a new proposal for grounding conceptual content. Journal of Physiology Paris, 102, 59-70.
Pecher, D., Zeelenberg, R., \& Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. Psychological Science, 14, 2, 119-24.
Scerrati, E., Baroni, G., Borghi, A. M., Galatolo, R., Lugli, L., \& Nicoletti, R. (2015). The modality-switch effect: visually and aurally presented prime sentences activate our senses. Frontiers in Psychology, 6, 1668.
Scerrati, E., Lugli, L., Nicoletti, R., \& Borghi, A. M. (2016). The Multilevel Modality-Switch Effect: What Happens When We See the Bees Buzzing and Hear the Diamonds Glistening. Psychonomic Bulletin \& Review, doi:10.3758/s13423-016-1150-2.
Tillman, R., Hutchinson, S., Jordan, S., \& Louwerse, M. M. (2013). Verifying properties from different emotions produces switching costs: Evidence for coarse-grained language statistics and fine-grained perceptual simulation. Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 3551-3556). Austin, TX: Cognitive Science Society.
Van Dam, W. O., Brazil, I. A., Bekkering, H., \& Rueschemeyer, S.-A. (2014). Flexibility in embodied language processing: context effects in lexical access. Topics in Cognitive Science, 6, 407-424.
Vukovic, V., Feurra, M., Shpektor, A., Myachykov, A., \& Shtyrov, Y. (2017). Primary motor cortex functionally contributes to language comprehension: An online rTMS study. Neuropsychologia, 96, 222-229.
Willems, R. M., Labruna, L., D., Esposito, M., Ivry, R., \& Casasanto, D. (2011). A functional role for the motor system in language understanding: Evidence from ThetaBurst Transcranial Magnetic Stimulation. Psychological Science, 22, 849-854.

# Object Representation in Multiattribute Choice 

Sudeep Bhatia (bhatiasu@sas.upenn.edu)<br>University of Pennsylvania, Philadelphia, PA.

Neil Stewart (neil.stewart@warwick.ac.uk)<br>University of Warwick, Coventry, UK.


#### Abstract

We propose a theoretical framework for understanding how everyday choice objects are represented and how decisions involving these objects are made. Our framework combines insights regarding object and concept representation in semantic memory research with multiattribute choice rules proposed by scholars of decision making. We also outline computational techniques for using our framework to quantitatively predict naturalistic multiattribute choices. We test our approach in two-object and three-object forced choice experiments involving common books, movies, and foods. Despite using complex naturalistic stimuli, we find that our approach achieves high predictive accuracy rates, and is also able to provide a good account of decision time distributions.


Keywords: Multiattribute choice, Semantic memory, Naturalistic decision making, Judgment and decision making

## Introduction

Most choices that people make on a day-to-day basis, from the books they read to the foods they eat, involve trading off attributes, so as to select the object whose attributes are overall the most desirable (Keeney \& Raiffa, 1993). There is, however, a disconnect between the way in which multiattribute choices are currently studied, and the way in which these day-to-day choices are typically made. Most multiattribute choice experiments explicitly present choice objects and their attributes to participants in a matrix of numerical quantities (e.g. Figure 1a). Everyday decisions, in contrast, are not usually composed of objects with a small set of explicitly presented and quantified attributes. Rather the objects in these decisions are much richer and complex (e.g. Figure 1b). Decision makers do have knowledge about these objects and their attributes, but this knowledge is represented in the decision makers' minds after having been learnt through prior experience with the choice domain.


Figure $1 a$ and $b$. Stimuli presentation in standard multiattribute choice experiments (left) and in Study 1 (right).

The divergence between the stylized stimuli used in current research and the complex multiattribute choices made in real-world settings is problematic. Choice processes
and resulting behaviors depend greatly on the ways in which attributes and objects are presented (e.g. Kleinmuntz \& Schkade, 1993) suggesting that real-world decisions, which seldom involve actual attribute-by-object matrices, may be different to the types of decisions observed in current experimental work. More importantly, by using artificial designs in which the attributes of objects are directly presented to decision makers, existing theoretical work has largely ignored the role of object representation. Storing, retrieving, and processing attribute information about the objects in a given choice problem is a pivotal part of the decision process, and a complete account of choice requires an approach that is able to specify the mechanisms involved at this stage in the decision, well as the relationship between these mechanisms and the final outcomes of the decision (see Bhatia, 2013 for a discussion).

This paper provides a theoretical framework capable of addressing these issues. It relies on insights in semantic memory research which suggest that low-dimensional attribute spaces are used to represent objects and concepts. For example, multi-dimensional scaling (Shepard, 1962) passes similarity ratings through a matrix decomposition algorithm, resulting in the recovery of a small number of latent attributes that best describe the structure of similarity for a given domain. Likewise, distributional models of semantic memory typically learn low-dimensional word representations through natural language. Some approaches, like latent semantic analysis, use singular value decomposition to perform dimensionality reduction on word-context occurrence matrices (Landauer \& Dumais, 1997). Others use Bayesian statistics or convolution based associative memory, but also result in low-dimensional representations for words (see Jones et al., 2015).

We suggest that these insights extend to everyday multiattribute choice, so that decision makers can be seen as using the distribution of observable features across choice objects in the environment to uncover low-dimensional latent attributes for representing the objects. Furthermore, we propose that it is these latent attributes that are evaluated and aggregated during the decision process. For simplicity we suggest that the recovery of latent attributes can be approximated using singular value decomposition on the observable feature space (as in e.g. Landauer \& Dumais, 1997), and that the evaluation of the latent attributes can be approximated with a linear model with decision weights for each latent attribute (as in e.g. Keeney \& Raiffa, 1993).

We also propose computational techniques for uncovering the latent attribute representations of common choice
objects. Particularly, keywords, tags, and other natural language descriptors for choice objects on internet websites, can be considered suitable proxies for the observable features of these objects. For a sufficiently rich online dataset, it is possible to train semantic models and learn the latent attribute representations for the objects in a choice environment, and subsequently examine peoples' choices between these objects.

## Framework

Let us consider a choice domain with $N$ total objects. Each of these objects has a set of observable features, and can be written as a vector of these features. If there are $M$ total unique features in the environment, then each for object $i$ we have $\boldsymbol{x}_{i}=\left(x_{i 1}, x_{i 2}, \ldots x_{i M}\right)$, with $x_{i j}=1$ or $x_{i j}=0$ based on whether or not feature $j$ is present in object $i$. Singular value decomposition involves processing the matrix $\mathbf{X}=\left[\boldsymbol{x}_{1}, \boldsymbol{x}_{2}, \ldots \boldsymbol{x}_{N}\right]$ to obtain $L \ll M$ latent attributes, corresponding to the $L$ largest singular values of $\mathbf{X}$. Using these singular values, we can represent an object $i$ as $z_{i}=$ $\left(z_{i 1}, z_{i 2}, \ldots z_{i L}\right)$, with $z_{i j}$ corresponding to the association between the object and the $j^{\text {th }}$ latent attribute. Note that $M$ can be very large in many naturalistic choice domains, whereas $L$ is typically much smaller.

The use of latent attributes for representing objects implies that our approach retains the multiattribute structure assumed by theoretical decision making research. Thus we can take common multiattribute decision rules and apply them very easily to latent attributes. We use a simple linear rule, which specifies a decision weight for each attribute and aggregates weighted attributes into a measure of utility for an object (Keeney \& Raiffa, 1993). The object with the higher utility is the one that is most frequently chosen. In the context of the latent attribute structure outlined here, this involves specifying an $L$ dimensional vector of weights $\boldsymbol{w}=$ ( $w_{1}, w_{2}, \ldots w_{L}$ ), and multiplying the latent attributes for an object $i$ by these weights, so as to obtain the utility for the object $U_{i}=\boldsymbol{w} \cdot z_{i}$. In order to permit random noise in the choice process we embed our utilities in the logit choice rule (Luce, 1959). In a two-object choice this specifies the probability of choosing an object $i$ over another object $i^{\prime}$ as $\operatorname{Pr}[i$ chosen $\left.]=\mathrm{e}^{U i}\right) /\left(\mathrm{e}^{U i}+\mathrm{e}^{U i^{\prime}}\right)=\mathrm{e}^{w \cdot z i} /\left(\mathrm{e}^{w \cdot z i}+\mathrm{e}^{w \cdot z i}\right)$. For the general case with $N^{\prime}$ choice objects we have $\operatorname{Pr}[i$ chosen $]=$ $\mathrm{e}^{U i} /\left(\Sigma_{\mathrm{n}=1 \ldots N^{\prime}} \mathrm{e}^{U n}\right)$.

In order to test our approach and illustrate its applicability we first need to uncover the actual attribute representations that characterize common choice objects. In related domains, such representations are usually obtained by asking experimental participants to generate features that describe the meaning of a given word (e.g. McRae et al., 2005). However common choice domains are so vast (involving thousands of features for thousands of objects) that the experimental elicitation of these feature norms may not practical. Thus we suggest that user-generated keywords, tags, and other descriptors for common choice objects on online datasets can be seen capturing the observable features that best describe the various objects.

In this paper, we use three large online datasets: www.GoodReads.com, which contains user-generated bookshelves for thousands of books; www.IMDB.com, which contains user-generated keywords for thousands of popular movies; and www.AllRecipes.com which contains user-specified ingredients for thousands of dishes. We scrapped these websites in 2014, and for each website we attempted to obtain as much information (as many objects and associated features) as was technically feasible. We obtained a total of 372,186 unique shelves for 15,737 books for the www.GoodReads.com dataset, a total of 160,322 unique keywords for 44,971 movies for the www.IMDB.com dataset, and a total of 24,688 unique ingredients for 39,979 recipes for the www.AllRecipes.com dataset. Using these user-generated descriptors as our observable features, each of the $N$ objects in each of the three datasets can be written as an $M$-dimensional feature vector $\boldsymbol{x}_{i}=\left(x_{i l}, x_{i 2}, \ldots x_{i M}\right)$, with $x_{i j}=1$ if object $i$ (a book, a movie, or a food dish) has observable feature $j$ (a keyword, a shelf, or an ingredient). A singular value decomposition on $\mathbf{X}=\left[x_{1}, x_{2}, \ldots x_{N}\right]$ can be subsequently performed to obtain $L$ $\ll \mathrm{M}$ latent attributes for the datasets.

## Study 1

In Study 1 we tested whether our theoretical framework and the computational techniques for applying this framework, actually predict peoples' everyday multiattribute choices. This is the primary experiment in this paper: It involves incentivized choices in the laboratory with reaction time measures. In later studies we examine variants of this design using non-incentivized online samples.

Method. In this study, 73 participant made binary choices between pairs of popular books. Participants were recruited from a university subject pool, and performed the study in a behavioral laboratory on computer screens. Participants were also incentivized, and one of their chosen books was selected at random and given to them at the end of the study.

Unlike most existing multiattribute choice experiments, the choice objects were not presented alongside a set of quantifiable attributes (as in e.g. Figure 1a). Rather they were shown to participants using just the covers of the books and the accompanying titles (as in e.g. Figure 1b). Overall, each of the 73 participants made 220 choices involving 150 unique books. The books used in this study were obtained from 30 different popular genres on www.GoodReads.com.

Model Fitting. We fit participant choices using the latent attributes recovered from a singular value decomposition (SVD) on the www.GoodReads.com data. We allowed the number of underlying latent attributes, $L$, to vary across participants. For a given value of $L$, we used the $L$ latent attributes with the highest singular values from the SVD on the www.GoodReads.com dataset. In order to ensure sufficient degrees of freedom for estimating decision weights, we restricted $L$ to a maximum of $L=100$ (and a minimum of $L=2$ ). In essence this leads to a total of 99 unique models for each participant, corresponding to $L=2$,
$L=3, \ldots L=100$. with a separate set of best fitting participant-level attribute weights for each model. The values of the 150 books in our study on the two latent attributes with the largest singular values are shown in Figure 2. Figure 2 also shows the ten shelves with the largest absolute weights for these two latent attributes.


Attribute 1

| Latent Attribute 1 |  | Latent Attribute 2 |  |
| :---: | :---: | :---: | :---: |
| Shelf | Weight | Shelf | Weight |
| default | -0.11 | fantasy | 0.14 |
| literature | -0.12 | teen | 0.13 |
| 2010 | -0.12 | ya-fiction | 0.12 |
| my-library | -0.10 | young-adult-fiction | 0.12 |
| 2009 | -0.10 | ya | 0.11 |
| series | 0.10 | sci-fi-fantasy | 0.11 |
| romance | 0.10 | adventure | 0.11 |
| part-of-a-series | 0.09 | ya-books | 0.11 |
| 2011 | -0.09 | magic | 0.11 |
| book-club | -0.09 | fantasy-sci-fi | 0.11 |

Figure 2. The values of the 150 books in Study 1 on the two latent attributes with the largest singular values, alongside the ten shelves with the largest absolute weights for these two latent attributes.

In order to avoid overfitting, we used ten-fold crossvalidation to test predictive accuracy and find the best performing model (i.e. best performing value of $L$ ) for describing each participant's choices. For model training, we recovered the weighting vector $\boldsymbol{w}$ that provided the best fit to the training data, with the assumption of a linear choice rule embedded in a logistic link function. This vector (whose dimensionality depended on the dimensionality of the model (value of $L$ ) in consideration), was recovered using maximum likelihood estimation. For model testing we calculated the proportion of choices in the test data predicted accurately by the recovered $\boldsymbol{w}$ for each model. A choice is considered to be predicted accurately if the utility assigned to the chosen option by the model in consideration is higher than the utility assigned to its competitor. Ultimately, the value of $L$ and corresponding weight vector $\boldsymbol{w}$ with the highest accuracy on the test data was considered to be the overall best fitting model.

Results. The mean accuracy of our approach for predicting the test data is $83 \%(S D=0.08)$, significantly above a baseline accuracy of $50 \%$ ( $p<0.01$ ). Additionally, the average best fitting value of $L$ across our participants is 39.67 ( $S D=27.95$. Table 1 summarizes statistics regarding model accuracy.

|  | Latent <br> Attributes | Random <br> Vectors | Lex <br> Heuristic | Tally <br> Heuristic |
| :--- | :---: | :---: | :---: | :---: |
| Study 1 |  |  |  |  |
| Mean | 0.83 | 0.69 | 0.50 | 0.72 |
| Std. Dev. | 0.08 | 0.08 | 0.03 | 0.09 |
| Median | 0.86 | 0.68 | 0.50 | 0.73 |
| Best Fit | 0.85 | 0.16 | 0.00 | 0.19 |
| Significant | 0.93 | 0.44 | 0.00 | 0.63 |
|  |  |  |  |  |
| Study 2 |  |  |  |  |
| Mean | 0.78 | 0.67 | 0.50 | 0.69 |
| Std. Dev. | 0.10 | 0.08 | 0.03 | 0.12 |
| Median | 0.80 | 0.65 | 0.50 | 0.70 |
| Best Fit | 0.74 | 0.18 | 0.00 | 0.21 |
| Significant | 0.88 | 0.46 | 0.00 | 0.59 |
|  |  |  |  |  |
| Study 3 |  |  |  |  |
| Mean | 0.74 | 0.51 | 0.33 | 0.53 |
| Std. Dev. | 0.13 | 0.07 | 0.01 | 0.12 |
| Median | 0.75 | 0.50 | 0.33 | 0.53 |
| Best Fit | 0.86 | 0.07 | 0.00 | 0.07 |
| Significant | 0.95 | 0.80 | 0.00 | 0.70 |
|  |  |  |  |  |
| Study 4 |  |  |  |  |
| Mean | 0.79 | 0.51 | 0.33 | 0.63 |
| Significant | 0.98 | 0.68 | 0.00 | 0.89 |
| Std. Dev. | 0.14 | 0.07 | 0.01 | 0.14 |
| Median | 0.80 | 0.50 | 0.33 | 0.65 |
| Sest Fit | 0.86 | 0.09 | 0.00 | 0.08 |
| Significant | 0.96 | 0.73 | 0.00 | 0.88 |
| Sest Fit 5 |  |  |  |  |
| Mean | 0.80 | 0.51 | 0.33 | 0.62 |
| Std. Dev. | 0.12 | 0.07 | 0.01 | 0.13 |
| Median | 0.80 | 0.50 | 0.33 | 0.62 |
|  | 0.91 | 0.04 | 0.00 | 0.09 |
|  |  |  |  |  |

Table 1. Summary of model fits. Mean", "Std. Dev." and "Median" indicate the distribution of best-fitting model accuracy rates on test data across participants. "Best Fit" describes the proportion of participants for which the model has the highest accuracy (these proportions sum to greater than one as models are sometimes tied) and "Significant" indicates the proportion of participants that outperform the baseline model with $p<0.05$.

One possibility is that our technique achieves its high accuracy rates by allowing flexible weights across a large number of dimensions. In order to control for this, we attempted the above model-fits with randomly generated attribute vectors. Particularly, for each participant and each object offered to the participant, we artificially created a 100 -dimensional vector with each dimension randomly and uniformly distributed in the range [0,1]. We then performed a 10 -fold cross validation procedure that examined the fits of linear models with flexible weights for $L$ dimensions of the random vectors. With this approach we found the mean accuracy to be $69 \%$ ( $S D=0.08$ ) Additionally, $84 \%$ of
participants achieved a higher accuracy rate using the recovered latent attributes from www.GoodReads.com, compared to the randomly generated vectors (and $8 \%$ of participants had equal accuracy with both approaches). A participant-level paired t-test indicates shows that this difference is significant ( $p<0.01$ ). Table 1 provides further statistics involving the random vectors approach.

Another alternative to our SVD-based attributes involves the use of the raw observable features for the books. Of course it is impossible to actually recover separate decision weights for each of these observable features. However, we can use well-known decision heuristics applied to these observable features. For example, using the lexicographic heuristic (Tversky, 1969) would involve considering only a single feature, and choosing the object that is the most desirable on this feature. Likewise, applying the tallying heuristic (Russo \& Dosher, 1983) would involve counting up the positive and negative features of each choice object, and choosing the object with highest number of positive features relative to negative features. We applied these two heuristics to participant-level choice using 10 -fold cross validation. For the lexicographic heuristic we used the training sample to determine which of the object's features has the highest absolute correlation with choice. We then used this single feature to predict the choices on our test sample. For the tallying heuristic, we used the training sample to determine whether each of the features were positively or negatively correlated with choice. If they were positively correlated with choice, they received a weight of +1 , and if they were negatively correlated with choice they received a weight of -1 . These weights were then applied to the observable features in the test data to predict choices according to the tallying heuristic.

We found that the lexicographic heuristic achieved a mean accuracy rate of exactly $50 \%(S D=0.03)$, indicating that it is not a suitable way of making multiattribute choices with such large features spaces. In contrast, the tallying heuristic achieved a mean accuracy rate of $72 \%(S D=0.09)$. When comparing these heuristics with our latent attribute approach, we found that all participants were better fit by our approach compared to the lexicographic heuristic, and that $78 \%$ of participants were better fit by our approach relative to the tallying heuristic (with another $16 \%$ tied). The differences in accuracy rates shown here are statistically significant when evaluated with a paired t-test ( $p<0.01$ for both heuristics). Table 1 provides further statistics involving the lexicographic and tallying heuristics.

How well do our model fits predict decision time? We can perform this test by embedding our best fitting utilities into a drift diffusion model (Ratcliff \& Rouder, 1978). Our utilities are a measure of the desirability of the objects and, within the drift diffusion framework, are likely to determine the drift rate. We can formalize this by allowing the mean drift rate in the drift diffusion model to be a linear function of the best fitting utility difference. Thus, for trial $a$ for participant $b$, we can write this mean drift rate as $v_{a b}=\beta_{0}+$ $\beta_{l} \cdot\left(U_{a b}{ }^{L}-U_{a b}{ }^{R}\right)$. Here $U_{a b}{ }^{L}$ is the predicted utility for the left
option in the trial for the participant, based on the best fitting model for the participant. Likewise, $U_{a b}{ }^{R}$ is the predicted utility for the right option. $\beta_{l}$ is a multiplier mapping this utility difference on to a drift rate, and $\beta_{0}$ is an intercept term capturing an absolute bias in drift for the left option. In this model, hitting the upper boundary leads to the left option being selected, whereas hitting the lower boundary leads to the right option being selected.

We fit this modified drift diffusion model permitting trial-to-trial variability in starting points and trial-to-trial variability drift rates. For this purpose, we adopted a hierarchical model fitting approach, as implemented by the HDDM toolbox (Wiecki et al., 2013). This approach recovers group mean parameters for the decision threshold, non-decision time, drift rates, trial-to-trial variability in starting points, trial-to-trial variability, and trial-to-trial variability drift rates, while also permitting individual differences in these parameters. Importantly this toolbox makes it easy to fit linear functions for drift rates as we wish to do in this paper. The best fitting group mean parameters from our specification, as recovered by the diffusion analysis, are presented in Table 2. Again $\beta_{l}$ represents the weight on utility difference in the drift term. As can be seen, the bulk of the distribution of this parameter lies above 0 , indicating that the best fitting utility difference has a strong positive relationship with mean drift in the model. Table 2 also displays the deviance information criterion (DIC) value for this fits.

|  | Mean | SD | Median |
| :--- | :---: | :---: | :---: |
| Full model |  |  |  |
| Boundary | 3.26 | 0.09 | 3.26 |
| Non decision time | 0.46 | 0.02 | 0.46 |
| $\beta_{0}$ | 0.01 | 0.01 | 0.01 |
| $\beta_{1}$ | 0.26 | 0.01 | 0.26 |
| DIC: $60,694.76$ |  |  |  |
|  |  |  |  |
| Restricted Model |  |  |  |
| Boundary | 3.08 | 0.08 | 3.08 |
| Non decision time | 0.48 | 0.03 | 0.48 |
| $\beta_{0}$ | 0.01 | 0.01 | 0.01 |
| DIC: $68,571.64$ |  |  |  |

Table 2. Summary of best fitting group mean parameters for the drift diffusion model fits in Study 1. Here $\beta_{1}$ represents the weight on utility difference in the drift term, in the full model. The restricted model sets this to O. DIC indicates the deviance information criterion value for the fits.

In a related analysis, we fitted a simplified version of this model in which $\beta_{I}=0$, and drift is independent of the predicted utility difference. As shown in Table 2, the fits for this model, measured through the deviance information criterion (DIC), are much lower than those for the extended model, suggesting that the utility differences specified by our approach do improve reaction time predictions in naturalistic multiattribute choice tasks.

## Studies 2-5

As a secondary demonstration we applied our approach to two other domains: food choice and movie choice. We conducted a series of online studies offering participants two-object and three-object choices between various food dishes and between various movies, and we predicted these choices using latent attributes obtained from user-generated ingredients on www.AllRecipes.com and user-generated keywords on www.IMDB.com.

Method. In Study 2, 90 participants recruited from Amazon Mechanical Turk made 200 binary choices between various food dishes. The food dishes were obtained from www.AllRecipes.com, and there were a total of 100 unique food dishes used in the study (which were the most popular dishes on www.AllRecipies.com). Choices in this study were presented on the screen using just the names of the dishes. Participants had to click on the names in order to indicate their choices. In Study 3, 88 participants recruited from Amazon Mechanical Turk made 200 three-object choices between various food dishes. The dishes used were the same as those in Study 2, and their presentation was identical to that in Study 2 (except that each screen offered three different choices, instead of two). Participants in both Studies 2 and 3 were compensated with money.

In Study 4, 75 participants recruited from an undergraduate student participant pool made 200 threeobject choices between different movies. There were a total of 100 unique movies used. These were the 100 most popular movies on www.IMDB.com (Internet Movie Data Base). The choices were presented on the computer screen using just the names of the movies and their IMDB movie posters. Participants had to click on the movie name or poster in order to indicate their choices. Participants were compensated with course credit. Study 5 was identical to Study 4, except that participants were recruited from Amazon Mechanical Turk. There were 223 total participants in this study, and they were compensated with a monetary payment.

Model Fitting. The model fitting in Study 2 was identical to Study 1, except that the latent attributes were recovered from a singular value decomposition on the www.AllRecipes.com data. Study 3 used a very similar model fitting technique, except that instead of a binary logit choice rule, there was a three-object (multinomial) logit choice rule. Studies 4 and 5 also used this choice rule, applied using latent attributes recovered from a singular value decomposition on the www.IMDB.com data.

Results. The accuracy rates from our analysis for the Studies 2-5 are displayed in Table 1. The mean accuracy for Study 2 is $78 \%(S D=0.10)$, the mean accuracy for Study 3 is $74 \% ~(S D=0.13)$, the mean accuracy for Study 4 is $79 \%(S D=0.14)$ and the mean accuracy for Study 5 is $80 \%(S D=0.12)$. All of these are significantly ( $p<0.01$ ) higher than the baseline accuracy of $50 \%$ (for Study 2) and $33 \%$ (for Studies 3-5).

We also found that the best fitting latent attribute models have a relatively low dimensionality, for most participants.

Overall, the average best fitting value of $L$ (i.e. number of dimensions) across our participants is $31.95(S D=28.55)$ for Study 2, $56.02(S D=27.18)$ for Study 3, $50.05(S D=$ 28.12) for Study 4, and $52.64(S D=25.93)$ for Study 5.

Table 1 also displays the results of a random vector model for these studies. Again it shows that the majority of participants are better described by our approach relative to the random vector approach. Finally, Table 1 shows the fits of the lexicographic and tallying heuristics. For Study 2, these fits are performed similarly to Study 1. However, Studies 3-5 involve three object choice. Thus the weights for the individual features necessary for fitting these heuristics cannot be obtained through a simple correlation analysis between the relative presence or absence of a feature and the choice in a trial. Instead we calculated, for each feature in each trial, Relative Presence $=C-0.5\left[U C_{1}\right.$ $\left.+U C_{2}\right]$. Here $C=1$ if the feature is present in the chosen option and 0 otherwise. Likewise, $U C_{I}=1$ if the feature is present in the first unchosen option and 0 otherwise, and $U C_{2}=1$ if the feature is present in the second unchosen option and 0 otherwise. For each feature, we summed Relative Presence over all the observations in the training data for the participant in consideration. This gave us a measure of the Total Relative Presence of the feature in the chosen options for the participant. For the lexicographic heuristic, we then selected the single feature with the highest absolute Total Relative Presence for the participant in the training data, and used this feature to predict the participant's choices in the test data. For the tallying heuristic we recoded the Total Relative Presence for a feature to generate a weight of +1 if Total Relative Presence was positive and -1 if it was negative. These binary weights were then used to predict the participant's choices according to the tallying heuristic. Using this approach, we again found that the lexicographic and tallying heuristics were out performed by the latent attribute approach, as shown in Table 1.

## Discussion

In this paper we have proposed that decision makers use low-dimensional latent attributes in order to make decisions in naturalistic multiattribute choice settings. We have obtained latent attribute representations for various everyday choice objects using user-generated object descriptors in large online datasets, and in five experiments, have predicted participant choices between these objects by fitting linear models with our latent attributes. Our fits reveal that our approach provides high accuracy rates, which significantly outperform accuracy rates obtained through other sophisticated methods (such as linear models with random attribute vectors, and lexicographic and tallying heuristics). The best fitting models in our analysis often have small or moderate number of dimensions. Additionally, these models are able to quantitatively predict decision times, when their estimated utilities are embedded within a drift diffusion process.

Our primary theoretical contribution involves the formal characterization of the processes involved in choosing between everyday choice objects. In doing so we extend insights from semantic memory research to the field of multiattribute decision making. The resulting framework attempts to describe all key aspects of the decision process, from the learning of object representations for common choice objects, to the use of these representations for evaluation and decision making. This is in contrast to most theories of multiattribute choice, which specify the mechanisms involved in aggregating decision attributes but seldom attempt to describe what these attributes actually are (see Bhatia, 2013 for a discussion).

Our results suggest that dimensionality reduction is not only at play in representing words, concepts, and various non-choice objects (as in e.g. Landauer \& Dumais, 1997; Shepard, 1962) but is also a critical feature of multiattribute choice object representation in preferential decision making. There are many reasons why this would be the case. Firstly, common multiattribute choice objects involve a large number of observable features, as well as systematic relationships between the features. Good decision making involves understanding these feature relationships, and using these relationships to make inferences about the objects. Even though the inferences in preferential choice are primarily evaluative, knowledge is used in a very similar manner as in categorization, language comprehension, object recognition, and other related tasks. Additionally, the use of latent attributes also offers a number of distinct advantages relative to the use of raw observable features. There are fewer latent attributes than there are observable features, and for this reason, latent attributes simplify the decision process. These attributes also reduce redundancy in object representation, and do so in the most efficient manner possible. In fact, our approach is not unlike principle components regression, which possesses a very similar set of statistical benefits (see Draper \& Smith, 1981).

That said, the approach presented in this paper is fairly simplistic: It involves a linear technique for dimensionality reduction combined with a linear multiattribute utility model. Both of these assumptions should be tested, it wouldn't be surprising if more sophisticated and more realistic approaches to building semantic representations ( Jones et al., 2015) and making choices (Oppenheimer \& Kelso, 2015) outperform the current approach. It may also be the case that the representations of choice objects depend not only on feature co-occurrence, but also on the reward structure of the domain in consideration. Individuals may, for example, learn object representations that best predict rewards, rather those that best predict feature occurrence. If this is the case then it would be necessary to train models of object representation alongside models of evaluation and choice (rather than training the former separately, as is done in this paper). This could be accomplished using neural networks with backpropagation from a preference (reward) layer to an object representation layer. Supervised topic
models may also facilitate the learning of such representations.

Despite the need to test more sophisticated representation and choice models, the success of our current approach nonetheless opens up a new avenue for studying naturalistic multiattribute choice. It can be applied to examine whether existing multiattribute choice effects also emerge in more realistic choice settings, where attribute information is not presented numerically (as in Figure 1a). It can also be used to extend the psychological analysis of multiattribute choice beyond the laboratory and predict real world choice data. Ultimately, by combining existing theories of semantic representation and multiattribute choice with rigorous analysis of large-scale data, this paper has proposed tools to capture the large number of important decisions made in the real-world, that are not currently within the scope of decision making research. This has the potential to significantly expand the theoretical, descriptive, and practical scope of this area of study.

## References

Bhatia, S. (2013). Associations and the accumulation of preference. Psychological Review, 120(3), 522.
Draper, N., and Smith, H. (1981), Applied Regression Analysis, New York: Wiley.
Jones, M. N., Willits, J. A., Dennis, S., \& Jones, M. (2015). Models of semantic memory. Oxford Handbook of Mathematical and Computational Psychology, 232-254.
Keeney, R. L., \& Raiffa, H. (1993). Decisions with Multiple Objectives: Preferences and Value Trade-offs. Cambridge University Press.
Kleinmuntz, D. N., \& Schkade, D. A. (1993). Information displays and decision processes. Psychological Science, 4(4), 221-227.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211.
McRae, K., Cree, G. S., Seidenberg, M. S., \& McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods, 37(4), 547-559.
Oppenheimer, D. M., \& Kelso, E. (2015). Information processing as a paradigm for decision making. Annual Review of Psychology, 66, 2774.
Russo, J. E., \& Dosher, B. A. (1983). Strategies for multiattribute binary choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(4), 676.
Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Psychometrika, 27(2), 125-140.
Tversky, A. (1969). Intransitivity of preferences. Psychological Review, 76(1), 31.
Wiecki, T. V., Sofer, I., \& Frank, M. J. (2013). HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. Frontiers in Neuroinformatics, 7, 14.

# The Interactive Shaping of Social Learning in Transmission Chains 

Lucas M. Bietti (lucas@bietti.org), Adrian Bangerter (adrian.bangerter@unine.ch), Eric Mayor (eric.mayor@unine.ch)<br>Institute of Work and Organizational Psychology, University of Neuchâtel, Rue Emile-Argand 11, Neuchâtel, CH-2000, Switzerland


#### Abstract

This study investigated the social transmission of memories and skills collected from a collaborative cooking task (raviolimaking) and across transmission chains. The transmission over three generations of pairs of participants occurred under two conditions. In the interactive condition, transmissions over generations occurred in face-to-face conversations, whereas in the non-interactive condition, generations videorecorded their instructions to the next generations. We analyzed the effects of verbal and embodied features of informational transfer on task performance. Our results show that performances improved over generations regardless of interactivity. In the discussion we suggest that tools (like cooking utensils) may have operated as cultural affordances encapsulating and transmitting important cultural knowledge for the successful completion of the task.


Keywords: social transmission; embodied interaction; social learning; joint complex task; cultural affordances; cooking

## Social learning and the necessary ingredients

 for cumulative cultural evolutionSocial learning (e.g. Bandura, 1977) is learning by observing or interacting with another individual or a product. Social learning mechanisms enable individual improvements in the efficiency or productivity of cultural artefacts (e.g. Ramstead, Veissière, \& Kirmayer 2016) to accumulate from one generation to the next (e.g. Boyd \& Richerson, 1994; Tomasello et al., 1993;). Such mechanisms include teaching (Kline, 2015), imitation or emulation (reverse engineering) (Caldwell \& Millen, 2008). Teaching and imitation represent cases of high-fidelity transmission, allegedly allowing "complex behaviors to disseminate and be retained in populations until beneficial modifications occur" (Vale, Flynn \& Kendal, 2012, p. 223). Currently, however, it is unclear whether teaching is a necessary ingredient for cumulative culture to accrue (Zwirner \& Thornton, 2015). The aim of our study is to investigate how different variants of teaching (interactive vs. non-interactive) affect cumulative cultural transmission of skills in a
complex joint task (collaborative cooking) in a laboratory setting.

## Cultural transmission in the laboratory

Transmission chains are a method used to study cultural evolution in the laboratory. Bartlett's (1932) seminal serial reproduction design allowed studying how content changes when transmitted from individuals of one generation to the next. Although Bartlett's method is more focused on constructive remembering of information originally provided to the first generation of participants and not on the accumulation of such information, it has largely inspired modern laboratory research on cumulative cultural evolution (e.g., Caldwell, Atkinson, \& Renner, 2016; Mesoudi \& Whiten, 2008) or the evolution of language and other communication systems (e.g. Fay, Arbib, \& Garrod, 2013; Kirby, Cornish, \& Smith, 2008). However, the method in its standard form prescribes one-way transmission without receiver feedback and has thus been criticized for neglecting the interactive processes germane to conversational remembering (Edwards \& Middleton, 1987). Indeed, recent work has shown that giving participants the opportunity to freely interact during transmission improves transmission quality (Tan \& Fay, 2011).

## Cultural transmission of manual tasks

In a series of experimental studies using transmission chains, individuals built paper airplanes or towers made of spaghetti and modeling clay (Caldwell \& Millen, 2008). For the paper plane task, successive generations had access to different types of social information: (i) information about actions (new generations observed what previous generations did); (ii) information about results (new generations observed final products and their performance measured in flying distance); and (iii) information
generated through teaching (new and old generations interacted about the completed task). These three types of social information were designed to enable imitation (information about actions); emulation (information about results) and instructed learning (information generated through teaching). The results indicated that cumulative learning was found in all three conditions (imitation, emulation and teaching). Such findings seem to challenge widespread claims about the necessity of social transmission for cumulative cultural evolution (e.g. Boyd \& Richerson, 1994).

Building paper airplanes is arguably an artificial task likely to be confounded by prior experience. Zwirner and Thornton (2015) extended these findings using a more realistic basket construction task. They found that teaching increased the accumulation of improvements over generations, but that it was not necessary for them to occur, further supporting the hypothesis that imitation and teaching are not "fundamental prerequisites for cumulative culture" (Zwirner \& Thornton, 2015, p. 7).

## Cultural transmission and social interaction

A recent study examined the influence of social interaction in transmission chains (Tan \& Fay, 2011). In an interactive condition, chains of participants interacted freely with one another to transmit narrative information from one generation to the next. In a non-interactive condition, receivers of the information had to listen to audio-recordings of narrations produced by senders (previous generation) and then recorded their own accounts of what they had listened to, which were passed on to a new generation of receivers for the same procedure. Transmission was more accurate in the interactive condition than in the non-interactive condition, and was due to the effect of receivers' behavior, including backchannels or clarification questions. The authors suggested that the motivation or ability to interact during information transmission may contribute to the emergence of cumulative culture.

## Our experiment

In Tan and Fay (2011), it could be the case that benefits of the interactive transmission of information are related to the nature of the task (information transfer) rather than to general mechanisms of cultural transmission. Hence, it is unclear whether interactivity plays a role in transmitting manual skills. Our experiment thus investigated the interactive context in which the cultural transmission of manual skills occurs. That is, we studied whether the teaching behaviors of senders who have experience with a skill is affected by the presence or absence of receivers from a subsequent generation. The experiment consisted in the cultural transmission of memories and skills collected from a collaborative cooking task (ravioli-making) via transmission chains. Chains of three generations (G1-G2-G3) of pairs of participants made ravioli and transmitted their experience to a pair in the next generation. This occurred under two conditions (interactive condition vs. non-interactive condition). In the interactive condition, transmissions occurred in face-to-face conversations, whereas in the noninteractive condition they were video-recorded as instructions to the next generation. All transmissions were video-recorded in order to analyze both verbal and embodied features of information transfer (e.g., gestures that depict an action), which may be particularly important for the transmission of manual skills.
In line with Caldwell and Millen's (2008) studies using manual tasks, we expect that performance will improve over generations due to the accumulation of learned improvements. We further expect interactive transmissions to allow receivers of information to ask questions and request clarifications (Tan \& Fay, 2011), and thus, to stimulate senders to talk and gesture more. This in turn may lead to a better transmission of skills. As a result, we also expect interactive transmissions to lead to better performance than non-interactive transmissions. Finally, we expect longer transmissions from senders to lead to subsequent higher performance in receivers when compared to performances following shorter transmission.

Thus, the hypotheses we tested were:
H1) Performance improves over generations;
H2) During transmissions (G1-G2; G2-G3), senders gesture more (H2a) and speak more (H2b) in the interactive condition than in the noninteractive condition;
H3) Performance improves more in the interactive condition than in the non-interactive condition;
H4) Performance is predicted by the number of words and the duration of gestures in the preceding transmission session.

## Method

## Participants

Participants ( $n=246 ; 117$ men) were recruited from the student population of the University of Neuchâtel (Age $M=23.2$; $S D=4.07$ ). They were fluent speakers of French, and reported having limited previous cooking experience. They had previous practice of simple skills like combining and heating ingredients but did not master more complex skills (e.g. preparing pie from scratch). Participants received 25 CHF compensation each for half an hour of their time along with an incentive of 0.25 CHF in total for each produced ravioli of good quality. There were 41 chains ( 20 in the interactive condition and 21 in the noninteractive condition). Pairs of participants were randomly assigned to different conditions (interactive vs. non-interactive) and generations (G1-G3) in the chains.

## Task

The task consisted of two kinds of sessions, performance sessions and transmission sessions (Fig. 1). In performance sessions, participants from each generation prepared ravioli together in pairs. Their goal was to produce as many goodquality ravioli as possible in 10 minutes. Each pair had at their disposal a ball of 150 grams of dough; 200 grams of filling made of ricotta cheese, concentrated tomato paste and salt; a 24-hole ravioli mold with zigzag sealing for easy release; a pasta maker; a rolling pin; a cutting board; 2 pizza cutters; 2 knives; 4 teaspoons; 2 kitchen cloths and kitchen paper; 250 grams of flour; and a stopwatch. Immediately after the time was up, the ravioli were evaluated by the experimenter.

Transmission sessions occurred immediately after each performance session (except for the last one, see Fig. 1). Pairs who had just completed the task explained to next-generation pairs how to prepare the ravioli. These sessions were unstructured and did not have time constraints (they typically lasted 2-8 minutes).


Figure 1. Sequence of sessions in the experiment and groups involved in each session.

## Procedure

Participants signed consent forms upon their arrival. G1 pairs watched a 3 min 47 sec video tutorial that was recorded for the study (Fig. 1). It provided information about the steps to be followed to prepare ravioli in pairs. They then completed Performance session 1, followed by Transmission session 1 (together with G2 pairs). Then, G2 pairs completed Performance session 2. During this time, G1 pairs were paid, debriefed and allowed to leave. After having completed Performance session 2, G2 pairs participated in Transmission session 2 (together with G3 pairs). Then, G3 pairs completed Performance session 3. During this time, G2 pairs were paid, debriefed and allowed to leave. After performance session 3, G3 pairs were paid, debriefed and allowed to leave.

## Measures

All sessions were videotaped and transcribed. Information transmission (teaching) was measured by the total number of words uttered by senders as well as the total duration of their manual gestures (iconic and pointing gestures) in the transmissions. Performance was measured as the quantity of "good" ravioli each pair produced. The criteria that experimenters considered to count ravioli as good exemplars were that they should contain enough filling and they should be perfectly sealed.

## Results

Descriptive results appear in Tables 1 (performance) and 2 (transmission).

Table 1. Performance ( $M, S D$ ) by condition and generation

|  |  | Interactive |
| :--- | :--- | :--- |
| G1 | $8.70(10.70)$ | Non-interactive |
| G2 | $10.65(11.51)$ | $10.33(9.36)$ |
| G3 | $14.80(10.75)$ | $12.57(8.81)$ |

Table 2. Transmission variables ( $M, S D$ ) by condition and transmission

|  | Interactive | Non-interactive |  |  |
| ---: | :--- | :--- | :---: | :---: |
|  | Sender words |  |  |  |
| G1G2 | $872.45(290.22)$ | $545.14(222.655)$ |  |  |
| G2G3 | $924.15(447.53)$ |  |  | $480.19(196.16)$ |
|  | Sender gestures |  |  |  |
| G1G2 | $196.19(69.93)$ | $152.40(76.00)$ |  |  |
| G2G3 | $216.80(98.52)$ | $117.27(57.64)$ |  |  |
| Note. | G1G2: Transmission | session |  |  |
| Transmission session 2. |  |  |  |  |

We tested our hypotheses using random intercept mixed-model regression (in R 3.4, packages lme4 and lmerTest). We included chains as clustering variables.

## H1: Performance improves over generations.

To test H1, we included condition and the linear trend of generation as predictors of performance.

Performance improved marginally (linear trend: $B$ $=2.79, S E=1.41, t=1.98, p=0.051)$. Condition was not a significant predictor of performance ( $B$ $=-0.19, S E=2.13, t=-0.09, p=0.92$ ). Thus, H1 is marginally supported.

H2: During transmissions (G1-G2; G2-G3), senders gesture more (H2a) and speak more (H2b) in the interactive condition than in the non-interactive condition.
To test H2, we included condition as a predictor of sender's gestures (H2a) and words (H2b), controlling in each model for transmission session. The interactive condition was the reference category. H2a was supported ( $B=-$ 71.04, $S E=20.59, t=-3.48, p=0.001$ ). H2b was also supported $(B=-385.63, S E=81.16, t=-4.75$, $p<.001$ ). Transmission (reference category: first transmission) was not related to the dependent variables (for sender's gestures: $B=-7.94, S E$ $=12.90, t=-0.615, p=0.54$; for senders' words: $B=-8.05, S E=49.06 ; t=-0.16, p=0.87$ ). Overall, H2 is supported.

H3: Performance improves more in the interactive condition than in the noninteractive condition.
To test H3, we included the condition, the linear trend of generation and their interaction term as predictors of performance. The performance of groups in the interactive condition improved across generations (linear trend: $B=4.31, S E=$ $2.03, t=2.12, p=0.04)$. Groups in the noninteractive condition do not improve less than groups in the interactive condition, as shown by the non-significant condition $x$ linear trend of generation interaction ( $B=-2.97, S E=2.84, t=-$ $1.04, p=0.30$ ). Thus, H3 is not supported.
The marginal improvement in performance in the test of H1 is probably driven by the interactive condition, as the magnitude of the negative interaction term in the test of H 3 is close to the main effect.

H4: Performance is predicted by the number of words and the duration of gestures in the preceding transmission session.

To test H4, we simultaneously included the number of words uttered by senders and their gestures during the transmission preceding the task as predictors of performance (we did not predict the performance of G1), controlling for condition and generation. H4 was not supported (for senders' words: $B=0.01, S E=0.06, t=0.15$, $p=0.87$; for senders' gestures: $B=0.03, S E=$ $0.02, t=1.41, p=0.16$ ). Neither Condition ( $B=$ $1.43, S E=2.70, t=0.53, p=0.60$ ) nor Generation ( $B=3.44, S E=2.04, t=1.68, p=0.10$ ) were related to performance.

## Discussion

We investigated whether the interactive context in which cultural transmission occurred affected the transmission process and its outcomes. We observed that senders behaved differently depending on the presence or absence of listeners (longer transmissions in terms of talk and gesture). However, such differences did not affect the subsequent performance of the receivers (see Table 1). That is, performance improved over generations (albeit marginally) regardless of interactivity, further supporting previous findings (e.g. Caldwell \& Millen, 2008; Zwirner \& Thornton, 2015) and research on cumulative learning in humans (e.g. Boyd \& Richerson, 1994). However, in contrast to the manual tasks previously employed (e.g. paper airplanes, spaghetti towers, and baskets), our collaborative cooking task presented some particularities. Some materials that the participants had the possibility to use (e.g. ravioli mold; pasta maker and rolling pin) may have operated as cultural affordances (e.g. Ramstead, Veissière, \& Kirmayer 2016) already encapsulating relevant information for the successful completion of the task. Material culture, as transmitted by cooking utensils, has played a central role in the evolution of human cognition (Malafouris, 2013). Future studies on cultural transmission in the laboratory should begin to take into consideration the importance of such cultural affordances if they want to better understand the actual ecologies of teaching and learning.

Tan and Fay (2011) showed that "that interaction between senders an receivers promotes more accurate recall and transmission of cultural information" (p. 405). Our analyses did not deal with verbal protocols and the amount of information accurate recalled over generations. However, based on this previous evidence we expected to find more increased performance in the interactive condition compared to in the noninteractive condition. In other words, if more accurate information were produced in interactive chains (as previous evidence suggests), it could lead to an increase in performance over generations in the interactive condition. Against our expectations, in our study, this was not the case. We did not find an effect of the number of words and duration of gestures produced by senders on performance of receivers over transmission chains (see Table 2). Although, senders spoke and gestured more in the interactive condition than in the non-interactive condition (see Table 2), this did not affect receivers' performance. This result suggests that there is no clear correspondence between the quantity of information transmitted over generations and performance.
In our analyses we did not examine the content of the information transmitted over generations or how it was recalled during performance. A possible theme for further investigation related to our current results is whether better or worse performing generations transmit more useful information over chains. It may well be the case that worse performing generations communicate more useful information to next generations if they focus their accounts on the errors they committed during performance.
In contrast to Tan and Fay (2011), our findings showed that in complex joint tasks having the possibility of asking questions and requesting clarifications (interactive condition) did not bring benefits. This could be related to the nature of the task participants were asked to perform. Whereas in Tan and Fay (2011) participants simply relayed information received from previous generations, in our study participants had the opportunity to perform the task before transmitting information to the next generation. This involved access to the
tools provided as well as the opportunity to test different strategies and solve problems. Throughout human evolution, "the social environment, not just individual minds, has become increasingly organized to support the flow of information across the generations" (Sterelny, 2012, p. 27). Looking at the multiple ways in which interactive contexts, task specificity, and cultural affordances affect the transmission of everyday skills (e.g. cooking) is an important step towards better understanding the mechanisms of cultural transmission.

## Acknowledgments

We gratefully acknowledge the support of the Swiss National Science Foundation [Ambizione Grant PZ00P1-154968].

## References

Bandura, A. (1977). Social learning theory. New York: General Learning Press.
Bartlett, F. (1932). Remembering. Cambridge: Cambridge University Press.
Boyd, R., \& Richerson, P.J. (1994). Why does culture increase human adaptability? Ethology and Sociobiology, 16, 125-143.
Caldwell, C.A. \& Millen, A. E. (2008). Experimental models for testing hypotheses about cumulative cultural evolution. Evolution and Human Behavior, 29, 165-171.
Caldwell, C. A., Atkinson, M., \& Renner, E. (2016). Experimental approaches to studying cumulative cultural evolution. Current Directions in Psychological Science, 25, 191195.

Edwards, D., \& Middleton, D. (1987). Conversation and remembering: Bartlett revisited. Applied Cognitive Psychology, 1, 7792.

Fay, N., Arbib, M., \& Garrod, S. (2013). How to bootstrap a human communication system. Cognitive Science, 37, 1356-1367.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105, 1068110686.

Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. Behavioral and Brain Sciences, 1-70. doi:10.1017/S0140525X14000090, e31.
Malafouris, L. (2013). How things shape the mind: A theory of material engagement. Cambridge, MA: MIT Press.
Mesoudi, A., \& Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. Philosophical Transactions of the Royal Society B, 363, 3489-3501.
Ramstead M.J.D., Veissière S.P.L., \& Kirmayer L.J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. Frontiers in Psychology, 7:1090. doi: 10.3389/fpsyg.2016.01090
Sterelny, K. (2012). The evolved apprentice: How evolution made humans unique. Cambridge, MA: MIT Press.
Tan, R., \& Fay, N. (2011). Cultural transmission in the laboratory: Agent interaction improves the intergenerational transfer of information. Evolution and Human Behavior, 32, 399-406.
Tomasello, M., Kruger, A.C., \& Ratner, H. (1993). Cultural learning. Behavioral and Brain Sciences, 16, 495-552.
Vale, G.L., Flynn, E.G., \& Kendal, G.L. (2012). Cumulative culture and future thinking: Is mental time travel a prerequisite to cumulative cultural evolution? Learning and Motivation, 43, 220-230.
Zwirner, E., \& Thornton, A. (2015). Cognitive requirements of cumulative culture: teaching is useful but not essential. Scientific Reports, 5, 16781; doi: 10.1038/srep16781.

# Language Modality Affects Responses in Left IFG during Processing of Semantically Ambiguous Sentences 

Lena M. Blott<br>University College London<br>Jennifer M. Rodd<br>University College London<br>Jane E. Warren<br>University College London


#### Abstract

Ambiguity resolution requires high-level interpretation processes, at least some of which are subserved by the inferior frontal gyrus (IFG), a region that is susceptible to modulation by task demands. This fMRI study investigates the extent to which ambiguity-related activation in IFG is modulated by the specific cognitive-linguistic demands posed by the modality in which a sentence is presented. In the present study, ambiguous sentences and matched unambiguous sentences were presented in three conditions: listening, reading, and rapid serial visual presentation (RSVP). The RSVP modality elicited stronger ambiguity-related haemodynamic responses than the other two modalities, particularly in left anterior IFG. This indicates that the RSVP modality cannot be used as a simple substitute for natural reading without taking into account the additional processing resources it requires.


# Generalized Representation of Syntactic Structures 

Reihane Boghrati (boghrati@usc.edu)<br>Department of Computer Science, University of Southern California Los Angeles, CA 90089 USA

Kate M. Johnson (katejohn@usc.edu)<br>Department of Psychology, University of Southern California Los Angeles, CA 90089 USA

Morteza Dehghani (mdehghan@usc.edu)
Department of Psychology and Department of Computer Science, University of Southern California Los Angeles, CA 90089 USA


#### Abstract

Analysis of language provides important insights into the underlying psychological properties of individuals and groups. While the majority of language analysis work in psychology has focused on semantics, psychological information is encoded not just in what people say, but how they say it. In the current work, we propose Conversation Level Syntax Similarity Metric-Group Representations (CASSIM-GR). This tool builds generalized representations of syntactic structures of documents, thus allowing researchers to distinguish between people and groups based on syntactic differences. CASSIMGR builds off of Conversation Level Syntax Similarity Metric by applying spectral clustering to syntactic similarity matrices and calculating the center of each cluster of documents. This resulting cluster centroid then represents the syntactical structure of the group of documents. To examine the effectiveness of CASSIM-GR, we conduct three experiments across three unique corpora. In each experiment, we calculate the clustering accuracy and compare our proposed technique to a bag-of-words approach. Our results provide evidence for the effectiveness of CASSIM-GR and demonstrate that combining syntactic similarity and tf -idf semantic information improves the total accuracy of group classification.


Keywords: Syntax; Text Clustering; Syntactic Similarity; Text Classification; CASSIM.

## Introduction

Language lies at the heart of human communication, and analysis of language has been shown to be an essential lens for investigating and understanding many different psychological properties. Language analysis has provided insight into depression (Ramirez-Esparza, Chung, Kacewicz, \& Pennebaker, 2008), moral values (Graham, Haidt, \& Nosek, 2009; Dehghani et al., 2016), neuroticism and extraversion (Mehl, Robbins, \& Holleran, 2012), political orientations (Dehghani, Sagae, Sachdeva, \& Gratch, 2014), and cultural backgrounds (Maass, Karasawa, Politi, \& Suga, 2006; Dehghani, Bang, et al., 2013) among many others.

Most of these studies, however, focus on quantifying word choice or semantics. While semantics undoubtedly play an important role in capturing psychological properties, it is vital to also include analysis of syntax in this process. Prior research has shown that syntactic structures also capture individuals and group differences for various demographic and psychological factors such as educational or regional background (Bresnan \& Hay, 2008), gender (Vigliocco \& Franck,
1999), socio-economics (Jahr, 1992), and emotional states and personality (Gawda, 2010).

Recently, several tools have been developed for automated analysis of syntactic structures. For example, Lu's (Lu, 2010) system analyzes fourteen different measures including the ratio of verb phrases, number of dependent clauses, and Tunits to calculate documents' syntactic complexity. Similarly, TAALES relies on several features such as frequency, range, academic language, and psycholinguistic word information to measure lexical sophistication (Kyle \& Crossley, 2015). By comparison, Coh-Metrix is a tool which provides measurement for over 200 different facets of syntax (e.g. mean number of modifiers per noun phrase, mean number of high-level constituents per word, and the incidence of word classes that signal logical or analytical difficulty) (Graesser, McNamara, Louwerse, \& Cai, 2004).

While each of these tools provides different mechanisms for measuring various syntactic features, they all rely on previously identified features of interest. More recently, we introduced ConversAtion Level Syntax Similarity Metric (CASSIM) to incorporate constituency parse trees when calculating the syntactic similarity of documents (Boghrati, Hoover, Johnson, Garten, \& Dehghani, 2017). CASSIM compares groups of documents based on underlying syntactic differences between groups of documents.

There are some situations, however, where hypothesis testing about predefined features or groups may not be the only aim. Instead, researchers may wish to identify new groupings of documents and the features which tie them together. These group-level linguistic representations can lead to important, novel discoveries about how a group communicates. Clustering techniques are widely used for this type of analysis. There is an extensive literature studying various text clustering approaches and their applications (Song, Li, \& Park, 2009; Sasaki \& Shinnou, 2005; Lin, Jiang, \& Lee, 2014). This literature demonstrates that many linguistic features facilitate improvements in text clustering (T. Liu, Liu, Chen, \& Ma, 2003; L. Liu, Kang, Yu, \& Wang, 2005), some of which address the effect of synonymy, hypernymy, syntax, and part of speech tags on text clustering methods (Sedding \& Kazakov, 2004; Lewis \& Croft, 1989; Lewis, 1992; Zheng, Kang,


Figure 1: CASSIM-Group Representation Process.
\& Kim, 2009).
In the current paper, we introduce ConversAtion Level Syntax Similarity Metric-Group Representations (CASSIMGR), a tool that captures the generalized representation of syntactic structure used by individuals in a certain group. CASSIM-GR groups documents into separate clusters based on their syntactic similarity scores, and uses the centroid of a cluster as a generalized representation of the syntactic structures used in that cluster. These centroid syntax representation can then be used to understand within-group syntax similarities and between-group syntax variations. As we will show, these generalizations of syntactic structures can be useful when analyzing differences between documents written by different individuals or groups.

This paper is structured as follows: First, we describe our proposed approach, CASSIM-GR, in more detail. Next, we validate the approach with a corpus of syntactically similar documents. Then, we apply CASSIM-GR to two other corpora: documents marked as dogmatic and non-dogmatic (Fast \& Horvitz, 2016) and documents from conservative and liberal weblogs (Dehghani, Sagae, Sachdeva, \& Gratch, 2013) and evaluate the classification accuracy of CASSIM-GR compared to tf-idf approach and a combination of the two approaches. Finally, we discuss limitation and future directions of our work.

## CASSIM-GR

In this section we describe CASSIM-GR for clustering groups of documents with similar syntactic structures. CASSIM-GR includes four general steps: 1 . constructing the syntactic similarity matrix, 2 . applying spectral clustering, 3 . calculating the center of clusters, 4. classification. Figure 1 demonstrates the steps involved in CASSIM-GR to compute the generalized representation of syntactic structures.

First, we use CASSIM (Boghrati et al., 2017) to calculate the syntactic similarity between each pair of documents. CASSIM relies on edit distance difference of constituency parse trees. It first generates parse trees for the sentences in each document. Next, it calculates the edit distance between each two sentences' constituency parse trees and matches the most syntactically similar sentences using Hungarian algorithm. Finally, it provides a score between 0 and 1 where
higher numbers indicate higher similarity between two documents. Using the syntactic similarity scores measured by CASSIM, we build a syntactic similarity matrix. With $N$ documents in our corpus, the syntax similarity matrix is $A_{N \times N}$; where $A_{i, j}$ is the syntactic similarity of the two documents $i$ and $j$.

Next, spectral clustering (Shi \& Malik, 2000) is used to cluster documents into a pre-defined number of groups. It has been shown that spectral clustering often outperforms traditional clustering algorithms (Von Luxburg, 2007). The general idea behind spectral clustering is to apply $k$-means clustering on eigenvectors of Laplacian matrix of $A$. The syntactic similarity matrix $A$, which is constructed in the previous step, and the number of clusters are provided as inputs to the spectral clustering method.

Clustering documents leads us to an essential next step which is extracting general attributes or representation of clusters. One way to address this concern is to calculate a centroid for each cluster. Clusters' centers facilitate researchers to better understand and analyze the syntactic structures used by a group of people or under certain situations by only analyzing center documents and without going through hundreds of documents. Hence, the third step in CASSIM-GR is calculating a centroid for each cluster. We define a cluster's center as the document which has the highest syntactic similarity to other documents in its cluster. To identify a cluster's center, we calculate average syntactic similarity of each document to other documents in its cluster and return the document with the highest average similarity. Additionally, we may return the top $n$ documents with the highest average syntactic similarity to other documents in a cluster as representative samples of that cluster.

Finally, we use cross-validation to test the accuracy and representativeness of the clusters' centers. To cross-validate, our approach uses CASSIM to calculate the syntactic similarity of the left-out document to each centroid and assigns the document to a cluster with the highest similarity. This process is repeated $N$ times and an accuracy of classification is reported by the method. In the following sections, we evaluate CASSIM-GR by performing classification experiments on three different corpora.

## Experiments

We conducted three experiments to validate CASSIM-GR and to examine the representativeness of the cluster centroids. Additionally, we examined how well documents with similar syntactic structures cluster together and demonstrate the importance of syntactic similarity in classification. Further, we compare the accuracy of syntactic clustering to bag-of-words clustering. For this purpose, we use the tf-idf similarity matrix as input to spectral clustering. Lastly, we combined tf-idf and CASSIM-GR to see how including both sets of information affect the classification accuracy. Below, we discuss the three experiments in detail.

## Experiment One

Experiment one was conducted on a corpus of syntactically similar documents. The corpus was generated by Amazon Mechanical Turk participants and consists of four groups of documents; each has high within-group syntactic similarity and low between-group syntactic similarity.

We used CASSIM-GR along with tf-idf, to group documents into clusters. Further, we combined these two approaches and calculated the overall accuracy. We first introduce the dataset and then report the results.
Data 118 MTurk participants answered a set of four questions. In each question they were asked to generate sentences with similar grammar rules to the sentence prompts in the question. Each of the four prompts had a different syntactic structure. Later, two independent coders, coded whether a sentence generated by a participant was grammatically similar to its prompt. Sentences which were identified as dissimilar by both coders were excluded from the dataset. Finally, a total of 272 documents, 68 documents in each group, were collected. See Boghrati et al. (2017) for more details.

Since participants were asked to write sentences similar to four different sets of prompts, the corpus is therefore divided to four separate groups, each associated to a question and its responses. Documents which are in the same group are considered to have similar syntactic structures.

Analysis We performed leave-one-out cross-validation for both of the clustering techniques. Namely, we ran the analysis on all the documents except for document $i$. Next, we labeled the clusters with the name of the group to which most of the documents belong. Then, we calculated similarity of document $i$ to each cluster's center. Finally, document $i$ was assigned to the cluster with which it had the highest syntactic similarity. The classification was considered successful if the assigned cluster's label and the document's group were identical.

We used the following approach to combine tf-idf and CASSIM-GR: First, we used CASSIM-GR and tf-idf approach separately to cluster documents into $k$ clusters. Cluster $j, j \in[1, k]$ in tf-idf approach and cluster $j^{\prime}, j^{\prime} \in[1, k]$ in CASSIM-GR were labeled with the same name, that is, the majority of documents in cluster $j$ and the majority of docu-
ments in cluster $j \prime$ were from the same group (e.g. 'liberals'). We averaged the syntactic similarity of document $i$ to center of cluster $j$ and the syntactic similarity of document $i$ to center of cluster $j \prime$. We repeated this procedure $k$ times to measure the similarity of document $i$ to all $k$ clusters and assigned document $i$ to the cluster with highest similarity score. If the cluster's label and document $i$ 's label were the same, we would conclude that prediction was successful.

Results Our results demonstrate that CASSIM-GR is able to accurately cluster the corpus. Following the instructions discussed in above, we performed leave-one-out crossvalidation on 272 documents. In each step, 271 documents were clustered in four groups and later the left-out document was assigned to one of the four clusters based on its similarity to the center of clusters.

Following this mechanism, CASSIM-GR yielded 95\% accuracy while tf-idf approach was only $84.5 \%$ accurate. Running a chi-squared test demonstrates that CASSIM-GR results in significantly higher accuracy than tf-idf, $X^{2}(1)=$ $17.01, p<.001$. Since the dataset consists of groups of syntactically similar documents, it is not surprising that clustering based on syntactic structures surpasses the word-based approach and achieves a higher accuracy.

Next, we combined the two approaches and obtained an accuracy of $97.8 \%$. While this result is not significantly higher than CASSIM-GR accuracy, $X^{2}(1)=2.67, p=.10$, we may conclude that incorporating syntactic and semantic information together could potentially improve clustering accuracy.

## Experiment Two

In the second experiment, we used the Dogmatism Dataset collected by Fast and Horvitz (2016). This dataset includes comments from New York Times which are rated based on their level of dogmatism. As explained below, we first categorized the documents as dogmatic or non-dogmatic based on this ratings. Next, we followed the procedure which was explained in the first experiment and clustered the documents using CASSIM-GR and the tf-idf approach. In the following subsections, we first introduce the dataset and then report the results.

Data The Dogmatism Dataset includes comments from New York Times. Amazon Mechanical Turk participants were asked to rate the level of dogmatism of each of the collected comments on a 5-point Likert scale. More details on the dataset and the annotation process are available at Fast and Horvitz (2016).
Analysis Dogmatism is subjective, and consequently interannotator agreement is higher for comments in both extreme sides of the spectrum. In other words, human coders tend to agree more on posts rated as very high in dogmatism and posts rated as very low in dogmatism (Fast \& Horvitz, 2016). Following the method used by Fast and Horvitz (2016), to have a representative and balanced dataset, we selected the top 250 and the bottom 250 documents based on the dogma-

Table 1: Corpora Overview.

|  | Experiment One | Experiment Two | Experiment Three |
| :--- | :--- | :--- | :--- |
| Corpus | Syntactically Similar Sentences | Dogmatism in New York Times | Political Weblog Posts |
| Number of Groups | 4 | 2 | 2 |
| Number of Documents | 272 | 500 | 452 |

Table 2: Accuracy of approaches in three experiments.

|  | Experiment One | Experiment Two | Experiment Three |
| :--- | :--- | :--- | :--- |
| CASSIM-GR | $95 \%$ | $54.8 \%$ | $69.9 \%$ |
| TF-IDF Approach | $84.5 \%$ | $61 \%$ | $64.4 \%$ |
| Combined Approach | $97.8 \%$ | $66.6 \%$ | $71.9 \%$ |

Table 3: Comparison of approaches in three experiments.

|  | Experiment One | Experiment Two | Experiment Three |
| :--- | :--- | :--- | :--- |
| CASSIM-GR vs. TF-IDF Approach | $X^{2}(1)=17.01, p<.001$ | $X^{2}(1)=3.94, p<.05$ | $X^{2}(1)=3.13, p=.07$ |
| TF-IDF Approach vs. Combined Approach | $X^{2}(1)=29.61, p<.001$ | $X^{2}(1)=3.39, p=.06$ | $X^{2}(1)=5.89, p<.05$ |
| CASSIM-GR vs. Combined Approach | $X^{2}(1)=2.67, p=.10$ | $X^{2}(1)=14.59, p<.001$ | $X^{2}(1)=.43, p=.51$ |

tism rating. We labeled the top 250 posts as dogmatic and the bottom 250 as non-dogmatic, hence the final dataset contained 500 posts with 250 in each group.

Results Following the instruction in Experiment 1, we performed leave-one-out cross-validation; we ran the clustering algorithm with 499 documents and left document $i, i \in$ [ 1,500 ], out. Then, we predicted to which cluster document $i$ belonged. CASSIM-GR and tf-idf approach resulted in $55 \%$ and $61 \%$ accuracy respectively. Even though, the tf-idf approach outperformed our approach significantly, $X^{2}(1)=$ $3.94, p<.05$, combining these two approaches resulted in a higher accuracy of $66.6 \%$, which is a marginally significant improvement over the tf-idf accuracy, $X^{2}(1)=3.39, p=.06$.

This result provides evidence for the importance of syntactic structure similarity in clustering documents. It demonstrates that not only what different groups of people say, but also how they say what they say provide important information about the characteristics of the group. This is evident by the fact adding syntactic similarity to word-level similarity can improve the clustering accuracy.

## Experiment Three

In this experiment, we applied CASSIM-GR on a corpus of political discussions taken from a set of conservative and liberal weblogs, and focus on the discussion about the Ground Zero Mosque (Dehghani, Sagae, et al., 2013).

Data The top five popular conservative and liberal news blogs were selected according to www.blogs.com. Next, a dataset of these weblogs posts which contained word mosque
and were written in the time frame of the debate, were complied. For more details about the dataset and the data collection process please refer to Dehghani, Sagae, et al. (2013).

Analysis In this experiment, we randomly selected 250 posts from conservative weblogs posts and 250 posts from liberal weblogs posts, but due to encoding issues the final dataset included 226 posts from each group (total of 452 posts).
Results Similar to the previous experiments, we used the leave-one-out cross-validation procedure described above. Specifically, we trained the clustering algorithm on 451 documents and predicted to which cluster the left-out document belonged. This process was repeated 452 so that each document was tested once.

CASSIM-GR was able to successfully predict the correct cluster for a document with $70 \%$ accuracy, while tf-idf was $64.4 \%$ accurate. This difference is only marginally significant, $X^{2}(1)=3.134, p=.0767$. Next, we combined these two approaches as described in the Experiment section. The total accuracy was $72 \%$ which is significantly more accurate than tf-idf approach alone, $X^{2}(1)=5.8905, p=.0152$.

These results demonstrated that, in some cases, syntactic structures similarity may capture more crucial features needed for clustering compared to tf -idf approach. However, there are some features that only tf-idf approach can pick up. Thus, the combination of these two sets of features is needed for more accurate clustering.

## Discussion and Future Work

Across three studies, we presented and validated a new approach called CASSIM-GR. CASSIM-GR clusters documents into separate groups based on their syntactic similarity and calculates a generalized representation of group-level syntax usage by performing four general steps: First, it creates a syntactic structure similarity matrix of documents using CASSIM. Second, it uses spectral clustering to group the documents into a pre-defined number of clusters using the syntactic similarity matrix generated in the previous step. Next, the algorithm selects the document which has the highest syntactic similarity to the other documents within each cluster and identifies it as the centroid of that cluster. Finally, it can be used to classify unknown documents based on the document's syntactic similarity to the clusters' centers.

We applied CASSIM-GR to three unique corpora (Table 1) across three experiments to compare its accuracy to both a bag-of-words approach and a combined approach incorporating tf-idf semantic information and CASSIM-GR. As Table 2 demonstrates, tf-idf and CASSIM-GR varied in their relative strength for clustering accuracy across studies. The combined approach incorporating both syntactic (CASSIM-GR) and semantic (tf-idf) information resulted in the highest clustering accuracy across all three experiments. While not a significant improvement beyond both single approaches, the combination approach significantly outperformed tf-idf in two of the three experiments and CASSIM-GR in the second experiment. Therefore, we may conclude that word-level similarity and syntactic similarity capture different aspects of language, and consequently, combining the two features' similarities results in more accurate clusters.

Our results indicate that methods assessing syntactic similarity may more accurately cluster documents than methods which rely on semantics alone. While there may be situations in which groups use the same general words to discuss a topic, syntactic similarity differences could still allow researchers to distinguish between different subsets of individuals.

More importantly, CASSIM-GR gives researchers an opportunity to study syntactic differences between groups by analyzing the prototypical syntactic structures at the clusters' centers. The syntactic structures used by a cluster's center document is defined as a generalized representation of syntactic structures of the documents in that cluster. Assessing differences in these structures may help to capture underlying psychological differences between groups in the ways that they conceptualize a topic or how they communicate with each other.

A vital component of CASSIM-GR is measuring syntactic similarity among documents using CASSIM. As mentioned previously, CASSIM's general focus is on comparing constituency parse trees. Building on CASSIM, we intend to compare dependency parse trees among sentences and documents to add another syntactic similarity measurement to CASSIM. Unlike constituency parse trees which posit the connection between part of speech tags, dependency parse
trees reveals the relationship between the words in a sentence. By incorporating this feature into CASSIM, researchers may further use CASSIM-GR not only to generalize syntactic structure of a group of documents, but also their dependency structures. This extension will help researchers study human language in finer grained detail by looking at the relationship between words.

In summary, we introduced a new method for computing generalized representations of syntactic structures of documents, allowing researchers to distinguish between groups of documents based on syntactic differences. Further, In the three experiments, we demonstrated the benefits of including syntactic structure similarity scores in clustering documents. In each experiment, we repeated a clustering procedure, once using CASSIM-GR and once using tf-idf similarity matrix. Then, we calculated clustering accuracy of each approach using leave-one-out cross-validation mechanism. Finally, we combined the results of these two approaches and calculated the accuracy when both sets of features were present. Our results support our assumption and demonstrated that syntactic similarity scores capture different aspects of language compared to bag-of-words, and therefore help improve clustering accuracy.

## Acknowledgments

This research was supported in part by NSF IBSS Grant \#1520031.

## References

Boghrati, R., Hoover, J., Johnson, K. M., Garten, J., \& Dehghani, M. (2017). Conversation level syntax similarity metric. Journal of Behavior Research Methods.
Bresnan, J., \& Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of give in new zealand and american english. Lingua, 118(2), 245-259.
Chalnick, A., \& Billman, D. (1988). Unsupervised learning of correlational structure. In Proceedings of the tenth annual conference of the cognitive science society (pp. 510516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dehghani, M., Bang, M., Medin, D., Marin, A., Leddon, E., \& Waxman, S. (2013). Epistemologies in the text of children's books: Native-and non-native-authored books. International Journal of Science Education, 35(13), 21332151.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., ... Graham, J. (2016). Purity homophily in social networks. Journal of Experimental Psychology: General.
Dehghani, M., Sagae, K., Sachdeva, S., \& Gratch, J. (2013). Linguistic analysis of the debate over the construction of the ground zero mosque. Journal of Information Technology \& Politics. Advance online publication. doi, 10(19331681.2013), 826613.
Dehghani, M., Sagae, K., Sachdeva, S., \& Gratch, J. (2014). Analyzing political rhetoric in conservative and liberal
weblogs related to the construction of the ground zero mosque. Journal of Information Technology \& Politics, 11(1), 1-14.
Fast, E., \& Horvitz, E. (2016). Identifying dogmatism in social media: Signals and models. arXiv preprint arXiv:1609.00425.
Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum \& J. Feldman (Eds.), Computers and thought. New York: McGraw-Hill.
Gawda, B. (2010). Syntax of emotional narratives of persons diagnosed with antisocial personality. Journal of psycholinguistic research, 39(4), 273-283.
Graesser, A. C., McNamara, D. S., Louwerse, M. M., \& Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, \& computers, 36(2), 193-202.
Graham, J., Haidt, J., \& Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of personality and social psychology, 96(5), 1029.
Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. Cognition and Brain Theory, 6, 287-317.
Jahr, E. H. (1992). Middle-aged male syntax. International Journal of the Sociology of Language, 94(1), 123-134.
Kyle, K., \& Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Quarterly, 49(4), 757-786.
Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In Proceedings of the workshop on speech and natural language (pp. 212-217).
Lewis, D. D., \& Croft, W. B. (1989). Term clustering of syntactic phrases. In Proceedings of the 13th annual international acm sigir conference on research and development in information retrieval (pp. 385-404).
Lin, Y.-S., Jiang, J.-Y., \& Lee, S.-J. (2014). A similarity measure for text classification and clustering. IEEE transactions on knowledge and data engineering, 26(7), 15751590.

Liu, L., Kang, J., Yu, J., \& Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. In Natural language processing and knowledge engineering, 2005. ieee nlp-ke'05. proceedings of 2005 ieee international conference on (pp. 597-601).
Liu, T., Liu, S., Chen, Z., \& Ma, W.-Y. (2003). An evaluation on feature selection for text clustering. In Icml (Vol. 3, pp. 488-495).
Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474-496.
Maass, A., Karasawa, M., Politi, F., \& Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? a cross-linguistic analysis of person representation. Journal of personality and social psychology, 90(5), 734.
Matlock, T. (2001). How real is fictive motion? Doctoral dissertation, Psychology Department, University of Cali-
fornia, Santa Cruz.
Mehl, M. R., Robbins, M. L., \& Holleran, S. E. (2012). How taking a word for a word can be problematic: Contextdependent linguistic markers of extraversion and neuroticism. Journal of Methods and Measurement in the Social Sciences, 3(2), 30-50.
Newell, A., \& Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Ohlsson, S., \& Langley, P. (1985). Identifying solution paths in cognitive diagnosis (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.
Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., \& Pennebaker, J. W. (2008). The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In Icwsm.
Sasaki, M., \& Shinnou, H. (2005). Spam detection using text clustering. In Cyberworlds, 2005. international conference on (pp. 4-pp).
Sedding, J., \& Kazakov, D. (2004). Wordnet-based text document clustering. In proceedings of the 3rd workshop on robust methods in analysis of natural language data (pp. 104-113).
Shi, J., \& Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), 888-905.
Shrager, J., \& Langley, P. (Eds.). (1990). Computational models of scientific discovery and theory formation. San Mateo, CA: Morgan Kaufmann.
Song, W., Li, C. H., \& Park, S. C. (2009). Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. Expert Systems with Applications, 36(5), 9095-9104.
Vigliocco, G., \& Franck, J. (1999). When sex and syntax go hand in hand: Gender agreement in language production. Journal of Memory and Language, 40(4), 455-478.
Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17(4), 395-416.
Zheng, H.-T., Kang, B.-Y., \& Kim, H.-G. (2009). Exploiting noun phrases and semantic relationships for text document clustering. Information Sciences, 179(13), 2249-2262.

# Context reduces coercion costs - Evidence from eyetracking during reading 

Oliver Bott (oliver.bott@uni-tuebingen.de)<br>Project Composition in Context (CiC), DFG Priority Program XPrag.de<br>SFB 833, Nauklerstraße 35, 72074 Tübingen, Germany


#### Abstract

This paper presents an eyetracking during reading experiment that investigated the role of supportive context on processing aspectual coercion. Coercion sentences in need of aspectual enrichment were embedded in discourse contexts providing the necessary information for successful interpretation. The findings of the reported experiment show that context information can be used immediately without disrupting reading of coercion sentences. The lack of coercion costs in supportive discourse contexts provides experimental evidence for the proposed Composition in Context Hypothesis and against theories that view semantic composition as largely encapsulated from context. Furthermore, the present experiment investigated the role of inter-individual differences in verbal working memory capacity on the immediate use of contextual information in computing coerced interpretations.


Keywords: Aspectual coercion; Discourse context; Semantic processing; Eyetracking during reading; Working memory capacity

## Introduction

A central assumption in semantics is that interpretation is governed by the principle of compositionality (see, e.g., Pelletier, 1994). The prinholds that the meaning of a complex expression is entirely determined by the meaning of its parts and their syntactic combination. However, linguistic expressions are at the same time highly context dependent, and the language interpretation system is therefore not only dependent on the parts of complex expressions in a bottom-up fashion, but also has to be open to top-down influences of the context of utterance. The present study investigates the interplay between sentential and contextual information during the online composition of the event interpretation. In particular, I tested whether contextual information is immediately used to resolve compositional conflicts during online interpretation.

Cases of coercion have prominently figured in studies on the time course of compositional interpretation, see e.g. Piñango and Deo (2016) for an overview of studies on complement coercion, and Bott (2010) and Paczynski, Jackendoff, and Kuperberg (2014) for aspectual coercion. (1) displays a compositional conflict calling for aspectual coercion.

## (1) \# Yesterday, Peter jogged in only thirty minutes.

When uttered out of the blue, sentence (1) is hardly interpretable. This is because an in-adverbial requires a telic event predicate of the accomplishment type (Vendler, 1957), but Peter jog- expresses an atelic activity. Under a coercion analysis Peter jog- has therefore to be shifted into an accomplishment, i.e. the event representation of the activity has to be enriched by adding a culminating event. The required operation can be summarized as follows:
(2) $\quad[$ in thirty minutes $[$ Peter jog-] $]] \rightsquigarrow_{\text {coerce }}[$ in thirty
$\quad$ minutes[ADD CULMINATION $[$ Peter jog- $]]]$

The inserted coercion operator ADD CULMINATION is a function that takes as input an activity and outputs an accomplishment (see, e.g., Dölling, 2014, for semantic representations of various coercion operators). The coercion operation solves the compositional problem. After the inclusion of the appropriate type shifting operator the resulting representation can be interpreted fully compositionally. Interestingly, the semantic problem in (1) seems to completely disappear once the sentence is embedded in an appropriate discourse context. Consider (1) in the context of (3).
(3) Half a year ago, Peter started to jog four kilometers $_{\text {culmination }}$ every day. When he began, he was quite slow but now he is really fast.

Based on the pragmatic literature (Recanati, 2010) two theoretical alternatives can be contrasted on how compositional interpretation might make use of contextual information. The Composition in Context ( CiC ) hypothesis predicts immediate availability of contextual information (e.g., Nieuwland \& van Berkum, 2006). Accordingly, the bounded path four kilometers from the preceding context should be immediately available when composing the adverbial with the rest of the target sentence in (1). Alternatively, however, compositional interpretation may operate in strictly locally in a bottom-up fashion (Cappelen \& Lepore, 2005). According to this view, which may be characterized as Encapsulated Composition $(E C)$ hypothesis, contextual information is only considered when the sentence information is not sufficient: Either to resolve compositional conflicts or to interpret context dependent expressions that are context dependent. In fact, in coercion theories it is standardly assumed that coercion operations are locally triggered by temporary semantic mismatch, i.e. many coercion analyses employ the EC hypothesis (cf. de Swart, 1998, p. 8).

According to the EC hypothesis the initial interpretation of (1) in the context of (3) should result in an aspectual mismatch that is only resolved in a second processing step. This should lead to measurable disruption during online interpretation. The CiC hypothesis, by contrast, predicts no processing costs of coercion sentences relative to non-coercing controls because the contextually given culmination can go right into the composed meaning.

Context effects on coercion have only been investigated in a small number of online studies so far. Traxler, McElree, Williams, and Pickering (2005) report a series of self-paced reading and eyetracking during reading experiments in which
they presented target sentences involving complement coercion (the student began the book: [begin[the book]] $\rightsquigarrow$ coerce [begin[to do something involving[the book]]]) and manipulated the preceding discourse context. Here is a sample item from Traxler et al. (2005, Exp. 3/4); c1/2 correspond to coercion vs. control contexts, $\mathrm{t} 1 / 2$ are coercion vs. control targets, critical regions underlined.
(4) c1) The student started a book in his dorm room.
c2) The student read a book in his dorm room.
t1) Before he started it, he checked his e-mail.
t2) Before he read it, he checked his e-mail.
Contextual facilitation was found in cases in which the context sentence either itself required complement coercion (c1), or explicitly introduced the relevant event (c2). In their Exp. 3 and 4, Traxler et al. (2005) observed coercion costs for c1 relative to c2. Crucially, however, the anaphor it in the coercion target t 1 was read as fast as in the control condition t 2 showing that the anaphor could be immediately linked to the relevant event in the discourse representation. This is prima facie evidence in favour of the CiC hypothesis. However, there is another explanation consistent with the EC hypothesis too. It is also possible that it was not interpreted as anaphoric to the book but rather as an event anaphor, which would make coercion unnecessary in the first place. The following example adopted from Asher (1993, p. 233) illustrates the difference between individual and event anaphora.

$$
\begin{equation*}
\text { If Timmy }{ }_{i} \text { hits }_{j} \text { John }_{k}, \mathrm{it}_{* i / j / * k} \text { will cause a fight. } \tag{5}
\end{equation*}
$$

As for aspectual coercion, Bott (2010, Exp. 3) reports a selfpaced reading experiment investigating whether German contexts such as (3) facilitate aspectual enrichment in German versions of (1). The findings provide preliminary evidence that supportive context can eliminate coercion costs observed for this type of sentences when presented out of context. After telic contexts such as (3) the reading times of coercion sentences did not differ from control targets allowing for plain compositional interpretation. However, this interpretation of the results is complicated by the fact that the comparison crucially involved a direct comparison between different adverbials (German in x time vs. for $x$ time).

Processing along the lines of the EC or the CiC hypothesis does not have to be an either or choice but could be subject to inter-individual differences. In particular, verbal working memory capacity may be an important constraining factor for being able to employ a highly context dependent processing strategy as put forward in the CiC hypothesis. Note that the CiC haypothesis presupposes full accessibility of all the relevant contextual information. Processing along the lines of the CiC hypothesis can therefore be expected to require more working memory capacity than the strictly local interpretation of the EC hypothesis. Existing research on sentential context effects has provided evidence that low-span readers make even less use of the immediate sentential context than highspan readers (van Petten, Weckerly, McIsaac, \& Kutas, 1997).

Thus, smooth aspectual enrichment in line with the CiC hypothesis may be especially expected for high-span readers, whereas low-span readers should be more likely to exhibit coercion costs. Inter-individual differences in working memory capacity have not been taken into account in coercion studies so far.

The present experiment studied contextual facilitation effects in aspectual coercion with four major modifications relative to prior research. First of all, the present study employed eyetracking during reading - an online method that provides us with a richer picture about the time course of interpretation than self-paced reading. Secondly, a larger set of experimental items was tested, and these materials were set up in such a way that the critical region was kept identical across conditions. Thirdly, the materials were more carefully pretested concerning their offline interpretation than the ones used in Bott (2010, Exp. 3). Finally, contextual facilitation effects were related to participants' verbal working memory capacity as measured by the reading-span task (Daneman \& Carpenter, 1980).

The experimental design of the present study included a coercion and a control condition as well as a mismatch condition. Two kinds of contexts were constructed. Both, telic contexts such as (3), and atelic contexts such as (6) introduced a repetitive event (contexts translated from German). The only difference is that the telic context (3) establishes a series of telic, bounded events (e.g., jog four kilometers) while the atelic context (6) introduces a series of atelic, unbounded activities instead. Both types of contexts put emphasis on the actual duration of the respective events at reference time now.
(6) Half a year ago, Peter started to jog every day. When he began, he had to stop after a short time but now he can run for quite a long time.

Target sentences were of two types manipulating the adverbial: telic sentences (7-a) including German in-adverbials, and atelic sentences (7-b) with German for-adverbials.
(7) a. Als es ihm vorhin gelang, in nur dreißig When he just managed in only thirty Minuten zu joggen,... minutes to jog...
'When he managed to jog in only $30 \mathrm{~min} . .$. '
b. Als es ihm vorhin gelang, ganze dreißig When he just managed for thirty Minuten zu joggen,... minutes to jog... 'When he managed to jog for $30 \mathrm{~min} . .$. '
c. ... war er sehr stolz auf sich.
... he was very proud of himself.
Discourse conditions were as follows. The coercion condition was constructed by combining telic contexts (3) with telic targets (7-a). The control condition combined atelic contexts (6) with atelic targets (7-b). For the aspectual mismatch condition atelic contexts (6) were paired with telic targets (7-a).


Figure 1: Mean judgments in Pretest 1 (left-hand side) and Pretest 2 (right-hand side). Error bars represent 95\% confidence intervals (by-participants analysis).

## Pretests

Two pretests were conducted. The first pretest was a sentence acceptability judgment experiment testing the telic (7-a) and the atelic (7-b) target sentences out of context. The second pretest was a discourse acceptability rating experiment that queried the felicity of the discourses in the coercion, the control, and the mismatch condition, respectively. The predictions are straightforward. Due to their need of coercion, telic target sentences should be less acceptable than atelic target sentences when encountered out of context. Supportive context should increase the acceptability of the coercion condition, though. After a telic context, a telic target should become as acceptable as the control condition. The mismatch condition should be judged as infelicitous.

## Pretest 1

Method 20 native German speakers (mean age: 26.9 years; 17 female) participated in the pretest for a payment of $€ 5$. Participants rated the acceptability of the telic and atelic target sentences on a scale from 1-7 from completely unacceptable to fully acceptable.

Target sentences were taken from the set of 24 items created for the eyetracking study. All were constructed following the scheme exemplified in (7). Pronouns were replaced by the proper names from the contexts (3)/(6). Two lists were constructed using a Latin square design and 100 filler sentences were added to each list. 60 of the fillers were infelicitous ('bad fillers') while the others were fully acceptable ('good fillers'). Ten participants were randomly assigned to each list.

Paricipants were tested individually in a quiet computer pool. Sentence materials were presented in randomized order in a single block which was preceded by a short practice of five trials. An experimental session took less than 30 minutes.

Results and discussion The mean judgments are shown in Figure 1. As predicted, acceptability of telic target sentences was judged significantly worse than atelic target sentences
$\left(t_{1}(19)=-5.4, p<.01 ; t_{2}(23)=-9.52, p<.01\right)$. The latter were judged even slightly better than the well-formed fillers (atelic targets: 5.90; well-formed fillers: 5.70) suggesting that the target sentences in the control condition are in fact fully acceptable. The telic target sentences received mean ratings of 4.24 and were thus well above the nonsensical fillers with a mean rating of 2.34 . Even though the telic target sentences were perceived as not fully acceptable when presented out of the blue, participants seemed to be aware of the fact that these sentences are in fact well-formed if embedded in an appropriate discourse context.

## Pretest 2

Method 30 new participants (mean age: 26.2 years; 16 female), all native speakers of German, took part in the pretest for a payment of $€ 5$. Participants rated the acceptability of the discourses in the coercion, the control, and the mismatch condition on a scale from $1-7$. In addition to the 24 experimental items 66 filler discourses were included ( 33 acceptable and 33 incoherent discourses). 20 of the incoherent filler discourses were globally incoherent, e.g. Lisa is very bad in maths. [...] . So, she wasn't surprised when she got an A., and 13 were locally incoherent, e.g. ...the jockey sat in his horse .... The items plus the fillers were distributed to three lists in a Latin square design. Ten participants were randomly assigned to each list. The procedure was the same as in the previous pretest
Results and discussion The mean judgments are also shown in Figure 1. As predicted, telic targets preceded by a telic context made the coercion condition fully acceptable. Paired t-tests revealed that the coercion condition did not differ reliably from the control condition $\left(t_{2}(29)=-1.65, p=\right.$ $.11 ; t_{1}(23)=-.76, p=.46$ ). Both, coercion (mean rating: 5.10) and control (mean rating: 5.27) received ratings in the range of the good fillers (mean rating: 5.82). As expected, the mismatch condition was judged similar to the incoherent fillers. Repeated measures ANOVAs revealed that mismatch was judged significantly worse than coercion and control $\left(F_{1}(2,58)=81.77, p<.01 ; F_{2}(2,46)=122,56, p<.01\right) .{ }^{1}$

Taken together, the results of the pretests show that the materials tested in the eyetracking study fully meet the assumptions stated in the introduction. The coercion targets are not fully interpretable on their own but require contextual support. Embedded in a telic context, however, the telic targets become fully acceptable. After atelic contexts, however, telic target sentences result in aspectually incoherent discourses.

## Eyetracking Experiment

The EC and the CiC hypotheses make fundamentally different predictions regarding the online processing of the three discourse conditions.

[^283]According to the EC hypothesis, the coercion operation is triggered by a temporary semantic mismatch during the intial interpretation of atelic activity verbs modified by inadverbials. Therefore, during first-pass reading the coercion condition should pattern with the mismatch condition and lead to processing difficulty relative to the control condition.

By contrast, the CiC hypothesis predicts smooth interpretation of the coercion condition. Coercion should therefore pattern with the control condition, and it should only be the mismatch condition that causes processing difficulty.

In order to assess effects of inter-individual differences in working memory capacity, between group analyses were conducted for high- versus medium- vs. low-span readers.

## Methods

Participants 48 new participants (mean age: 24.1 y., range $20-32$ y.; 40 female) all native speakers of German with normal or corrected-to-normal vision took part in the experiment for a payment of $€ 8$. Based on their performance in Daneman and Carpenter's (1980) reading span task they were divided into three groups. The first group consisted of 10 participants with a reading span of 3.0 , the median value in the sample. This group was included as MEDIUM SPAN readers in ANOVA analyses including the between factor READING SPAN. The group of LOW SPAN readers consisted of 19 participants with a mean reading span of 2.4 (range $2.0-2.5$ ). The group of HIGH SPAN readers consisted of 19 participants with a mean reading span of 3.7 (range $3.5-4.5$ ).
Materials The sentence materials were identical to those used in the pretest and the items were always presented on three lines. The first line contained the first context sentence, the second line the second context sentence, and the third line the target sentence. The target sentences were split up in ten regions of interest (ROIS):
(8) Als es ihm $\mid$ heute $\mid$ gelang, $\left.\right|_{a d v 1}$ in nur $\left.\right|_{a d v 2}$ dreißig $\left.\right|_{a d v 3}$ Minuten $\left.\right|_{\text {verb }}$ zu joggen, $\left.\right|_{\text {coda } 1}$ war er $\left.\right|_{\text {coda } 2}$ sehr stolz $\left.\right|_{\text {coda } 3}$ auf sich. $\mid$

The critical region was the verb RoI. Note that any aspectual mismatch only becomes evident at this ROI. For instance, als es ihm heute gelang, in nur dreißig Minuten vier Kilometer weit zu joggen, ... (when he managed today to jog four kilometers in only thirty minutes ...) would be fully acceptable even after an atelic context.

Apparatus and procedure Eye movements of the dominant eye were recorded with an SR Research Ltd. Eyelink 1000 eyetracker. The trial began with the presentation of a screen which served as calibration check and for drift correction with a yellow dot in the position where the centre of the first word would appear. If no fixation on the dot was registered within five seconds, recalibration was enforced. Otherwise, texts were presented. After reading, participants had to move their eyes to a yellow dot in the right bottom corner of the screen which triggered the presentation of the judgment
screen. Judgments had to be provided by pressing the left or the right button of a gamepad.

The experiment started with five discourses for practice, followed by the individually randomized presentation of the experimental trials in three blocks. A typical experimental session lasted less than 45 minutes. Immediately after the eyetracking experiment each participant was subjected to an experimenter-administered version of the reading span task (Friedman \& Miyake, 2004). Reading span was scored as follows. The highest stage for which at least two out of a total of three sequences could be correctly recalled determined a participant's basic reading span. If she was able to correctly recall one sequence from an even higher stage, a value of 0.5 was added to this basic value.
Eyetracking analysis Prior to all analyses the eyetracking data were preprocessed. Two trials with major track loss were exluded, and all fixations immediately preceding or following a blink were eliminated. All fixations shorter than 80 ms and further than 0.5 degrees from the last or next fixation as well as fixations longer than 800 ms were eliminated. Preprocessing affected $2.7 \%$ of all fixations.

The analyzed measures of first-pass reading included first fixation durations, first-pass times, and first-pass regression ratios, i.e. the proportions of regressions made during firstpass reading.Measures related to rereading included secondpass time and the proportions of regressions in.

## Results

The coercion condition was accepted $83.3 \%$ of all trials, control was accepted $86.9 \%$, but mismatch was rejected $66.5 \%$. Thus, discourses in the coercion and in the control condition were generally accepted while the aspectual mismatch condition was generally rejected. The analysis of judment RTs, corrected for outliers by eliminating all RTs more than 2.5 standard deviations above a participant's mean RT, revealed no significant differences between the three discourse conditions ( $F_{1 / 2}<1$ ).

Table 1 presents the descriptive statistics for the eyetracking measures related to first-pass parsing. The findings for the measures related to rereading are shown in Table 2.

First-pass reading Immediately when readers encountered the verb region, first fixation durations were longer in the mismatch condition than in the control condition $\left(t_{1}(47)=\right.$ $\left.2.29, p<.05 ; t_{2}(23)=2.12, p<.05\right)$. By contrast, verbs in the coercion condition were read equally fast as verbs in the control condition $\left(\left|t_{1 / 2}\right|<1\right)$. Before or after the critical verb ROI there were no significant differences in first fixation durations for any of the ROIs in the target sentences.

The analyses of first-pass times further corroborated this finding. At the verb ROI a clear mismatch effect was found $\left(t_{1}(47)=3.85, p<.01 ; t_{2}(23)=3.72, p<.05\right)$, but coercion did not differ from control $\left(\left|t_{1 / 2}\right|<1\right)$. The mismatch effect was again limited to the verb ROI, and discourse conditions did not differ reliably from each other at any target ROI.

Table 1: Mean first fixation durations (FFD), mean first-pass times (FPT), and mean first-pass regression ratios (FPRR) of the target sentences in the eyetracking experiment. Note: ROI a3 corresponds to the final adverbial region, verb to the critical verb region, and $c 1-c 3$ to the three sentence final ROIs.

|  |  | a3 | verb | c1 | c2 | c3 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| FFD | contr. | 205 | 251 | 238 | 252 | 256 |
| (in ms) | coerc. | 210 | 247 | 247 | 247 | 264 |
|  | mism. | 213 | 266 | 240 | 243 | 257 |
| FPT | contr. | 217 | 311 | 290 | 333 | 403 |
| (in ms) | coerc. | 219 | 317 | 317 | 312 | 424 |
|  | mism. | 219 | 345 | 304 | 307 | 417 |
| FPRR | contr. | 14.6 | 18.5 | 6.4 | 25.4 | 45.8 |
| (in \%) | coerc. | 15.7 | 16.9 | 9.3 | 26.3 | 48.3 |
|  | mism. | 16.2 | 27.1 | 12.2 | 21.9 | 51.2 |

Mismatch not only slowed down reading speed during firstpass reading, it also gave rise to more regressions from the verb ROI. The analysis of first-pass regression ratios showed that readers launched more regressions from mismatching verbs than from verbs in the control condition $\left(t_{1}(47)=\right.$ $\left.2.23, p<.05 ; t_{2}(23)=3.08, p<.01\right)$. Again, coercion did not differ from control $\left(\left|t_{1 / 2}\right|<1\right)$.

Taken together, the findings from the three measures reflecting the first-pass reading of the critical verb ROI show that the initial interpretation of the coercion targets was as smooth as that of the control targets. Without a preceding telic context in the mismatch condition, however, first-pass reading was severely disrupted.
Rereading The analyses of the proportions of regressions in revealed that readers regressed back to the three adverbial RoIs in the mismatch condition. For all three RoIs reliable mismatch effects were observed (first ROI: $t_{1}(47)=2.52, p<$ . 05 ; $t_{2}(23)=2.82, p<.05$; second ROI: $t_{1}(47)=2.70, p<$ $.05 ; t_{2}(23)=2.84, p<.01$; third ROI: $t_{1}(47)=2.02, p<.05$; $\left.t_{2}(23)=2.16, p<.05\right)$. Also, on the verb ROI a mismatch effect was found that was marginally significant by subjects and significant by items $\left(t_{1}(47)=1.87, p=.07 ; t_{2}(23)=\right.$ $2.18, p<.05$ ). The coercion analyses showed that the adverbial ROIS did not receive more regressions in the coercion condition than in the control condition (all $\left|t_{1 / 2}\right|<1$ ). However, it turned out that readers regressed more often back into the verb ROI than they did in the control condition. This was reflected by a (by-items marginally) significant coercion effect $\left(t_{1}(47)=2.46, p<.05 ; t_{2}(23)=1.97, p=.06\right)$.

A similar pattern of effects was observed in the secondpass times too. The mismatch condition led to longer secondpass times than the control condition persisting from the adverbial ROIs (first ROI: $t_{1}(47)=2.83, p<.01 ; t_{2}(23)=$ $2.23, p<.05$; second ROI: $t_{1}(47)=2.40, p<.05 ; t_{2}(23)=$ $2.64, p<.05$; third ROI: $t_{1}(47)=1.70, p=.10 ; t_{2}(23)=$ $1.80, p=.08)$ to the verb ROI $\left(t_{1}(47)=4.71, p<.01\right.$;

Table 2: Mean second-pass times (SPT), and mean proportions of regressions in (RI) of the relevant ROIs of the target sentences.

|  |  | a1 | a2 | a3 | verb | c1 | c2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| SPT | contr. | 81 | 106 | 56 | 81 | 88 | 115 |
| (in ms) | coerc. | 85 | 104 | 59 | 143 | 115 | 132 |
|  | mism. | 129 | 160 | 89 | 167 | 87 | 108 |
| $R I$ | contr. | 35.7 | 22.8 | 19.7 | 4.6 | - | - |
| (in \%) | coerc. | 37.9 | 26.3 | 22.0 | 9.2 | - | - |
|  | mism. | 47.5 | 33.5 | 30.7 | 8.1 | - | - |

$\left.t_{2}(23)=3.90, p<.01\right)$. Also, a coercion effect was present, and, consistent with what was observed for the regressions in, this effect was limited to the verb ROI $\left(t_{1}(47)=3.25, p<.01\right.$; $\left.t_{2}(23)=2.46, p<.05\right)$.

Taken together, the analyses of late eyetracking measures besides substantial mismatch effects - show that participants were more likely to reread the verbs in the coercion condition than in the control condition.

Analyses contingent on reading span In order to investigate whether early and late effects were modulated by interindividual differences in working memory capacity ANOVAs with the within-factor DISCOURSE CONDITION and the between factor READING SPAN (three levels: HIGH SPAN vs. MEDIUM SPAN vs. LOW SPAN) were computed analyzing the first and second-pass times of the verb ROI. Table 3 presents the mean reading times of the verb ROI split up by groups.

The analysis of first-pass times only revealed a significant main effect of DISCOURSE CONDITION $\left(F_{1}(2,90)=\right.$ $4.61, p<.05$ ), i.e. the above reported mismatch effect. Neither the main effect of READING SPAN nor its interaction with DISCOURSE CONDITION reached significance (both $F_{1}<.05$ ). Thus, the three reading span groups did not differ with respect to their first-pass reading times of the verb.

The analysis of second-pass times also revealed no differences between the three groups. The main effect of DISCOURSE CONDITION was reliable $\left(F_{1}(2,90)=11.88, p<\right.$ .01), but neither the main effect of READING SPAN nor the interaction reached significance (both $F_{1}<1.3$ ).

To summarize, the three reading span groups had strikingly similar patterns of results. The three groups showed equally sized early effects of aspectual mismatch and late coercion effects that only started during rereading the sentence.

## Discussion

The present study investigated whether coercion sentences embedded in supportive discourse context lead to measurable processing costs during their initial interpretation. According to the EC hypothesis, the compositional system operates strictly bottom-up, and the coercion targets should therefore lead to temporary aspectual mismatch during the initial inter-

Table 3: Mean first- (FPT) and second-pass times (SPT) of the verb ROI split up by reading span groups.

|  |  | $F P T(\mathrm{~ms})$ | SPT (ms) |
| :--- | :--- | :--- | :--- |
| low span | contr. | 325 | 73 |
|  | coerc. | 324 | 106 |
|  | mism. | 353 | 126 |
| medium span | contr. | 313 | 72 |
|  | coerc. | 316 | 158 |
|  | mism. | 344 | 230 |
| high span | contr. | 296 | 95 |
|  | coerc. | 309 | 171 |
|  | mism. | 338 | 174 |

pretation and subsequent context-driven repair. During firstpass reading target sentences in the coercion and the mismatch condition should exhibit qualitatively similar processing effects. The CiC hypothesis, by contrast, predicts smooth interpretation of the coercion targets if embedded in a supportive context because the culminating event from the context should be immediately available.

The findings of the present eyetracking experiment unambiguously provide evidence against the EC hypothesis. While aspectual mismatch led to substantial processing difficulty during first-pass reading, none of the three analyzed early eyetracking measures indicated any difficulty in the coercion condition. The processing of aspectual coercion involved a qualitatively different time course than aspectual mismatch.

Do the findings support the CiC hypothesis, then? Above, it was stated that supportive context should completely eliminate all coercion costs. The coercion effects observed during rereading of the coercion targets may therefore be taken as prima facie evidence against the CiC hypothesis. However, the exposition of this hypothesis in the introduction was grossly oversimplistic. On closer consideration, the hypothesis is fully consistent with overall higher processing demands in the coercion than in the control condition as long as difficulty only emerges after a smooth first composition step. Two potential sources of difficulty come to mind. Under the CiC hypothesis comprehenders still have to be able to exactly recall the culmination introduced two sentences before. Note that in order to make proper sense of the coercion target it is important to know the exact length of the path. Obviously, to jog a distance of three miles in half an hour is plausible but thirty miles is not. Therefore, successful coercion crucially depends on a precise representation of the context in all its particulars. Another potential source of difficulty may be attributed to the metalinguistic evaluation demanded from the participants: since the target sentences were exactly the same in the coercion and the mismatch condition, the metalinguistic evaluation may have been more difficult in the coercion than in the control condition.

Both explanations, the (un-)availability of the culminating event in working memory as well as difficulty during metalin-
guistic evaluation, would be consistent with the CiC hypothesis. The above reported analyses taking into account participants' working memory capacity suggest that the second explanation is more likely than the first.

## Acknowledgments

This research was funded by the DFG in the project Composition in Context of the Priority Program SPP 1727 XPrag.de. The author would like to thank Salome Mauthe and Maximilian Müller for conducting the experiments. The author would also like to thank four anonymous CogSci 2017 reviewers for their valuable comments on a first draft of this paper. The usual disclaimer applies.

## References

Asher, N. (1993). Reference to abstract objects in discourse. Dordrecht: Springer.
Bott, O. (2010). The processing of events. Amsterdam: John Benjamins.
Cappelen, H., \& Lepore, E. (2005). Insensitive semantics. Oxford: Blackwell.
Daneman, M., \& Carpenter, P. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19, 450-466.
de Swart, H. (1998). Aspect shift and coercion. Natural Language \& Linguistic Theory, 16(2), 347-385.
Dölling, J. (2014). Aspectual coercion and eventuality structure. In K. Robering (Ed.), Events, arguments and aspects topics in the semantics of verbs (p. 189-226). Amsterdam: John Benjamin.
Friedman, N. P., \& Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. Journal of Memory and Language, 51, 1-25.
Nieuwland, M. S., \& van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. Journal of Cognitive Neuroscience, 18(7), 1098-1111.
Paczynski, M., Jackendoff, R., \& Kuperberg, G. (2014). When events change their nature. Journal of Cognitive Neuroscience, 26(9), 1905-1917.
Pelletier, F. J. (1994). The principle of semantic compositionality. Topoi, 13(1), 11-24.
Piñango, M. M., \& Deo, A. (2016). Reanalyzing the complement coercion effect through a generalized lexical semantics for aspectual verbs. Journal of Semantics, 33(2), 359-408.
Recanati, F. (2010). Truth-conditional pragmatics. Oxford: Oxford University Press.
Traxler, M., McElree, B., Williams, R., \& Pickering, M. (2005). Context effects in coercion. Journal of Memory and Language, 53, 1-25.
van Petten, C., Weckerly, J., McIsaac, H., \& Kutas, M. (1997). Working memory capacity dissociates lexical and sentential context effects. Psychological Science, 8(3), 238-242.
Vendler, Z. (1957). Verbs and times. The Philosophical Review, 66(2), 143-160.

# The Structure of Goal Systems Predicts Human Performance 

David D. Bourgin ${ }^{\text {1a }}$ (ddbourgin@berkeley.edu)<br>Falk Lieder ${ }^{\text {1a }}$ (falk.lieder@berkeley.edu)<br>Daniel Reichman ${ }^{1 \text { 1a }}$ (daniel.reichman@gmail.com)<br>Nimrod Talmon ${ }^{\text {1b }}$ (nimrodtalmon77@gmail.com)<br>Thomas L. Griffiths ${ }^{\text {a }}$ (tom_griffiths@berkeley.edu)<br>${ }^{\text {a }}$ Department of Psychology<br>University of California Berkeley<br>Berkeley, CA 94720 USA<br>${ }^{\mathrm{b}}$ Department of Computer Science and Applied Mathematics<br>Weizmann Institute of Science<br>Rehovot, Israel<br>${ }^{1}$ These authors contributed equally.


#### Abstract

Most psychological theories attribute people's failure to achieve their goals exclusively to insufficient motivation or lack of skill. Here, we offer a complementary explanation that emphasizes the inherent complexity of the computational problems that arise from the structure of people's goal systems. Concretely, we hypothesize that people's capacity to achieve their goals can be predicted from combinatorial parameters of the structure of the network connecting their goals to the means available to pursue them. To test this hypothesis, we expressed the relationship between goals and means as a bipartite graph where edges between means and goals indicate which means can be used to achieve which goals. This allowed us to map two computational challenges that arise in goal achievement onto two classic NP-hard problems: Set Cover and Maximum Coverage. The connection between goal pursuit and NP-hard problems led us to predict that people should perform better with goal systems that are tree-like. Three behavioral experiments confirmed this prediction. Our results imply that network parameters that are instrumental to algorithm design could also be useful for understanding when and why people struggle in their goal pursuits.


Keywords: decision-making; goals; rational analysis; graph theory; computational complexity

## Introduction

The ability to set and achieve high-level goals, such as creating a CogSci paper, is a critical feature of human intelligence and a key challenge for artificial intelligence systems (Newell \& Simon, 1972). Critically, everyday problem solving requires people to juggle multiple goals in parallel (Atkinson \& Birch, 1970; Miller, Galanter, \& Pribram, 1960). Concretely, when people are given ten minutes to list their current pursuits they will report about 15 goals on average and each of those goals typically entails multiple subgoals at several levels of abstraction (Little \& Gee, 2007).

It is generally agreed that there are many situations in which people fail to act on their goals (Baumeister, Heatherton, \& Tice, 1994). The predominant explanations of such failures are lack of motivation, lack of planning (Gollwitzer, 1999), failure to delay gratification (Mischel, Shoda, \& Rodriguez, 1989), or the depletion of the capacity for selfcontrol (Muraven \& Slessareva, 2003). Here, we explore an alternative explanation for people's failure to achieve their
goals: the inherent complexity of the underlying computational problem.

The relationship between means and goals can be formalized in a bipartite graph whose vertices are divided into a set of means $M=\left\{m_{1}, \cdots, m_{k}\right\}$ and a set of goals $G=$ $\left\{g_{1}, \cdots, g_{l}\right\}$. In this graph, a mean $m$ is connected to a goal $g$ if and only if selecting $m$ will achieve the goal $g$. Such networks are called goal systems in the psychological literature (Kruglanski et al., 2002). For instance, the vertices at the top of the goal system illustrated in Figure 1 might correspond to your goals to become a prolific scientist $\left(g_{1}\right)$, be a wonderful partner $\left(g_{2}\right)$, become a great parent $\left(g_{3}\right)$, get physically fit $\left(g_{4}\right)$, and enjoy life to the fullest $\left(g_{5}\right)$.

Finding the best configuration of means for achieving a set of goals can involve considerable computational challenges. It has been suggested that findings from theoretical computer science can shed light on how people cope with hard computational problems (Van Rooij, 2008). For example, Van Rooij (2008) advocated applying the theory of fixed parameter tractability to study how people cope with hard computational problems. Yet, while previous research on problem solving has investigated which strategies people use to solve NP-hard problems (MacGregor \& Ormerod, 1996; MacGregor, Ormerod, \& Chronicle, 2000), this literature has focused on the Traveling Salesman problem and other problems that are structurally distinct from those that arise in goal pursuit. ${ }^{1}$ Here, we will fill this gap by analyzing goal achievement through the lens of computational complexity theory.

In theoretical computer science, it is well known that the performance of many combinatorial optimization algorithms critically depends on certain graph-theoretic properties of the networks they are applied to (Kleinberg \& Tardos, 2006). For instance, a well-documented phenomenon in algorithm design, artificial intelligence, and operational research is that NP-hard optimization problems often become easier on trees and tree-like graphs. Indeed, when restricted to trees, many

[^284]

Figure 1: Example of goal system: means appear at the bottom marked by $m_{1} \ldots m_{4}$. Goals appear on top marked by $g_{1} \ldots g_{5}$. A goal can be attained if at least one mean connected to it is chosen. The set $\left\{m_{1}, m_{3}\right\}$ is a minimal set of means covering all goals

NP-hard problems can be solved by efficient polynomial-time algorithms, such as divide-and-conquer methods and greedy algorithms. Here, we show that human performance on two goal management tasks is also well predicted by graph theoretic measures of the tree-likeness of the underlying goal system. Our results offer a fresh computational perspective on why people fail to achieve their goals. Our experimental results align well with theoretical knowledge from computer science and highlight that findings from computational complexity are relevant to cognitive psychology.

The plan for the paper is as follows. We start by formalizing two common challenges of goal achievement in terms of two classic NP-hard problems: Set Cover and Maximum Coverage. Next we derive our theoretical prediction by arguing that people's performance on these problems should increase with graph theoretic metrics of how tree-like the goal system is. We then test this prediction in three behavioral experiments and conclude with a summary of our findings and directions for future work.

## Formal analysis and predictions

Here we formalize, as well-defined NP-hard problems, two computational challenges that arise in means selection problems where one seeks to choose a set of means that are instrumental to the ends one is trying to achieve.

The first problem we consider is trying to achieve as many goals as possible with a fixed budget that limits the number of activities that one can perform. We formalize this challenge in terms of the Maximum Coverage problem (MC). In it, we are given a bipartite graph $H=(C, D, F)$ where $C, D$ are the sides of the bipartition, and $F$ is the set of edges connecting vertices in $C$ to vertices in $D$. We are also given a nonnegative integer $k \leq|C|$. We seek to find a set $C^{\prime} \subseteq C$, of cardinality at most $k$, maximizing the number of vertices in $D$ covered by vertices in $C^{\prime}$ (a vertex $b \in D$ is covered by a vertex $a \in C$ if $(a, b) \in F)$. As a goal system, the set $C$ corresponds to means, the set $D$ corresponds to goals, and $F$ represents to interconnections between goal and means. Observe that we assume that once a goal is covered by a single mean then it will be achieved. This assumption is made in order to simplify the experimental task, allowing for a simple and clean
description.
The second problem that we study is trying to achieve a given set of goals as efficiently as possible by selecting the minimal number of means that will accomplish all goals. We formalize this challenge in terms of the Set Cover problem (SC). In the Set Cover problem, we are given a bipartite graph $G=(A, B, E)$, where $A, B$ are the sides of the bipartition and $E$ is the set of edges connecting vertices in $A$ to vertices in $B$. As a goal system, the set $A$ corresponds to means, the set $B$ corresponds to goals, and the set $E$ corresponds to interconnections between goals and means. In the SC problem, our goal is to cover all vertices in $B$ by a set $A^{\prime} \subseteq A$ of minimal cardinality. Both MC and SC are NP-hard not only to solve exactly but also to approximate (Feige, 1998).

These two computational problems (MC and SC) capture essential aspects of means selection problems that have been studied in the psychological literature.

For example, it is generally agreed that people try to select means in order to maximize the number of attained goals (Kruglanski et al., 2002; Zhang, Fishbach, \& Kruglanski, 2007). Furthermore, representing goals as graphs and assuming interconnectedness between goals appear either explicitly or implicitly in several papers (Kruglanski et al., 2002; Thagard \& Millgram, 1995).

We shall use the following two performance measures in quantifying how well people do in MC and SC. Our first measure simply quantifies the ability to find the optimal solution. This is a binary measure that equals 1 if the person finds the optimal solution and 0 otherwise. Given an algorithm $A$ for MC or SC, the second measure, referred to as the solution quality, ranges between 0 (worst) and 1 (best). For the MC task, the solution quality is the number of goals achieved by $A$ divided by maximum number of goals that could be achieved with $k$ means. For the SC task, the solution quality is the minimum number of means that achieves all goals divided by the number of means selected by $A$. Solution quality (which is also referred to as the approximation ratio of $A$ ) is widely used in quantifying the quality of approximation algorithms for NP-hard problems (Vazirani, 2013).

The hardness of MC and SC gives a first indication for why people might find it difficult to juggle multiple goals at the same time. Yet, not every instance of these problems is equally difficult. Next, we introduce graph-theoretic measures that might be useful for distinguishing harder instances from easier ones.

## Features and predictions

A tree is a connected graph without cycles. Many NP-hard problems on graphs with small treewidth (Robertson \& Seymour, 1986, see below), allow exact or approximate algorithms which are significantly better than what is known to be achievable on worst-case instances.

The idea that tree-like graphs might be easier for people to deal with guided our search for features that quantify how similar a given network is to a tree. Four such features are presented below.

- Treewidth. Treewidth (tw) is a combinatorial parameter that is associated with a graph. ${ }^{2}$ Low treewidth implies that the nodes and edges of the network can be arranged in a way that resembles a tree (e.g., Kloks, 1994; Robertson \& Seymour, 1986). For a precise definition of treewidth see Kloks (1994).
- Combinatorial expansion. Given a graph $G=(V, E)$ and a nonempty subset $S \subseteq V$, let $\partial(S)$ be the set of edges crossing the cut $(S, V \backslash S)$ and let $N(S)$ be the set of all vertices in $V \backslash S$ having a neighbor in $S$. The vertex-expansion of $G$ is defined as $\min _{S \subseteq V, 0<|S| \leq|V| / 2} \frac{|N(S)|}{|S|}$. The edge-expansion of $G$ is $\min _{S \subseteq V, 0<|S| \leq|V| / 2} \frac{|\partial(S)|}{|S|}$. Trees of bounded degree have expansion $O(1 /|V|)$, hence large expansion suggests that the graph is dissimilar to a tree. We computed both vertex and edge expansion by solving an integer linear programs (ILPs) using IBM's CPLEX.
- Spectral expansion. The adjacency matrices of the graphs we consider are symmetric, hence have $n$ real eigenvalues $\lambda_{1} \geq \lambda_{2} \geq \ldots \geq \lambda_{n}$. The classical (discrete) Cheeger's inequality (e.g., Alon and Milman, 1985) implies that the larger $d-\lambda_{2}$ is, the larger the edge expansion of the graph. ${ }^{3}$

MC and SC instances with low treewidth $w$ are known to have exact algorithms that run in time $O\left(2^{w} n\right)$ (Alber \& Niedermeier, 2002), hence instances with low treewidth are likely to be easier to deal with algorithmically. Current algorithms that compute tree-decompositions are quite complicated, therefore it seems unlikely that people would use them to solve SC and MC problems. Nevertheless, treewidth might affect the performance of people's heuristics: the similarity between low treewidth graphs and trees might make the kinds of algorithms that people might use, such as greedy and divide-and-conquer methods, much more effective. Conversely, as worst-case instances of MC and SC are hard even to approximate, and as hard instances often have large treewidth and expansion (Clementi \& Trevisan, 1999), it is likely that it will be hard not only to solve exactly, but also to find approximate solutions for instances of large treewidth. Similar reasoning applies to our expansion measures. We therefore hypothesize that treewidth, vertex-expansion, edgeexpansion, and the spectral gap of $G$ are negatively correlated with the quality of people's solution to SC and MC problems and frequency with which they find an optimal solution.

## Additional predictors

A feedback vertex set (FVS) is a subset of vertices whose removal from a given graph results in a forest. The size of a minimal feedback vertex set is an alternative measure to the similarity of a graph to a tree. Hence we used this

[^285]feature as well. We calculated FVS using the implementation based on (Iwata, 2016; Wahlström, 2014) available at https://github.com/wata-orz/fvs.

Previous empirical hardness models have found additional features of graphs to be useful: the diameter, average eccentricity, and average path length (Leyton-Brown, Nudelman, \& Shoham, 2009). We thus included these features. The diameter of a graph is the longest distance between two vertices in the graph (where the distance, $\operatorname{dist}(u, v)$, is the number of edges in a shortest path connecting $u$ and $v$; all graphs considered are connected). The eccentricity $\varepsilon(v)$ of a vertex $v$ in an undirected graph $G=(V, E)$ is the maximal distance of a vertex in $V \backslash v$ from $v$. The average eccentricity (AvgEcc) is $\frac{1}{n} \sum_{v \in V} \varepsilon(v)$ ( $n$ is the number of vertices). The average path length (AvgPath) is $\frac{2}{n(n-1)} \sum \operatorname{dist}(u, v)$ where summation is taken over all pairs of distinct vertices.

## Behavioral experiments

To test how our predictors relate to human performance in MC and SC, we conducted three crowdsourced behavioral experiments. We used a between-subjects design where each participant was assigned randomly to one of twenty graphs with treewidths varying from 4 to 13 . In the first experiment, participants were asked to solve the SC problem, and in the second experiment participants were asked to solve the MC problem. In each case, the problem was graphically represented as a bipartite graph with 48 vertices. The 24 vertices at the bottom represent the available means (activities A-Z) and 24 vertices at the top represent the goals. Each edge from a means vertex to a goal vertex implies that completing that activity is sufficient to achieve the goal. In the SC task, participants were asked to select a minimal number of activities to achieve all of the goals. In the MC task, participants were asked to choose five activities that achieve as many goals as possible. The third and final experiment asked participants to solve a SC problem where goals are given semantic content and real values, and a different visual display is used to eliminate possible visualization effects. In each experiment we restricted our analyses to goal systems in which every goal and mean vertex had exactly 4 neighbors (i.e., each graph was 4-regular). This restriction meant that each graph required the same amount of memory to enable processing, ensuring that any difference between conditions cannot be explained by working memory limitations.

## Experiment 1: Human performance on Set Cover

Methods We recruited 655 participants on Amazon Mechanical Turk. Participants were paid $\$ 1.25$ and could earn a performance-dependent bonus of up to $\$ 2$. Each participant was randomly assigned to one of 20 conditions that differed only in the graph structure of the SC problem participants were asked to solve. After consenting to participate, participants read a cover story about a person trying to choose which set of activities (e.g., volunteer to improve the company's website and work out at the gym) they should perform in order to achieve all their goals (e.g., earn more money, improve


Figure 2: Interface and instructions for Experiment 1.
relationship with boss, get fit, etc.) with as few activities as possible because their time is limited. The story highlighted that some activities achieve multiple goals at the same time. Next, participants completed a simple practice trial involving only two means and two goals. Once they had solved this task successfully, participants proceeded to the Set-Cover problem they had been assigned to. When participants moused over an activity, this interface highlighted the goals it would achieve and the corresponding edges of the graph in green. Goals that had already been achieved and activities that had already been selected were highlighted by checkmarks, and the number of selected activities was shown at the bottom of the screen. Participants were asked to help a person achieve their 24 goals with as few activities as possible, and they were motivated by the prospect of earning a financial bonus of $\$ 2$ for achieving all goals as efficiently as possible. After submitting their solution, participants completed an exit questionnaire asking them for basic demographic information (age, gender, and primary language), and four 9-point Likert scales (anchors: "not at all", "somewhat", and "extremely") measuring the perceived difficulty of the task, motivation to achieve all goals, motivation to find the optimal solution, and their motivation to finish the task as quickly as possible. We excluded 30 participants ( $4.6 \%$ ) because their responses did not achieve all goals suggesting that they did not follow the instructions.

Results We found that treewidth alone explained $44.90 \%$ of the variance in the frequency with which people found the optimal solution across the 20 graphs $(F(1,18)=14.20, p=$ 0.0012 ): as we increased the treewidth of the graph the percentage of participants who discovered the optimal solution decreased significantly ( $\rho=-0.59, p=0.0058$ ) from more than $90 \%$ on the graph with treewidth 5 to only about $30 \%$ on the graph with treewidth 14 . We found that the aver-
age solution quality was negatively correlated with treewidth ( $\rho=-0.44, p=0.0525$ ) suggesting that our participants achieved fewer goals for goal systems with higher treewidth. Treewidth explained $17.59 \%$ of the variance in the median response time across problems $(F(1,18)=3.86, p=0.0650)$ : the median amount of time people took to solve the problems tended to increase with treewidth $(\rho=0.3426)$ but this effect was not statistically significant ( $p=0.1393$ ), and when we restricted this analysis to correct solutions the correlation was $\rho=0.3825$ ( $p=0.1297$ ). Perceived difficulty also tended to increase with treewidth ( $\rho=0.37$ ) but this correlation was not statistically significant ( $p=0.1062$ ). Our participants were highly motivated to find the optimal solution (average rating $7.91 \pm 0.06$ out of 9 ). Thus it appears unlikely that their motivation was a bottleneck to their performance. Furthermore, motivation appeared to be unaffected by treewidth ( $\rho=-0.23, p=0.34$ ). Thus the observed differences in performance appear to result from the inherent difficulties of the means selection problems posed by different goal systems.

Of the additional predictors evaluated, we found that graph diameter, average shortest path, and average graph eccentricity all were significantly positively correlated with the frequency of optimal solutions identified by our participants (graph diameter: $\rho=0.5691, p=0.0088$, avg. shortest path: $\rho=0.6516, p=0.0019$, avg. eccentricty: $\rho=0.6265, p=$ 0.0031 ). In addition, the spectral expansion (measured as $d-\lambda_{2}$ ) and the size of the graph vertex and edge expansions showed significant negative correlations with the frequency of optimal solutions (vertex expansion: $\rho=-0.5836, p=$ 0.0069 , edge expansion: $\rho=-0.4552, p=0.0437$, spectral expansion: $\rho=-0.6280, p=0.0030$ ). We also found that the average shortest path, average graph eccentricity, spectral expansion, and the size of the graph vertex expansions were significantly correlated with the average participant solution qualities (avg. shortest path: $\rho=0.4505, p=0.0462$, avg. eccentricity: $\rho=0.4316, p=0.0574$, spectral expansion: $\rho=$ $-0.4496, p=0.0467$, vertex expansion: $\rho=-0.4636, p=$ 0.0395 ). Finally, only the cardinality of the graph edge expansions exhibited a significant correlation with the median response times on the SC task ( $\rho=0.4538, p=0.0445$ ).

## Experiment 2: Human performance on Maximum Coverage

Methods Experiment 2 was identical to Experiment 1 except for the task: participants were now instructed to achieve as many goals as possible subject to the constraint that the person's limited time does not permit them to complete more than five activities. The 20 graphs and financial incentives were the same as in Experiment 1. The interface of Experiment 1 was modified to prevent participants from selecting more than five activities at a time. When a participant attempted to add a sixth activity they were told they would first have to remove one or more of the activities they had already selected. The cover story and survey were modified slightly to match the change in the task. We recruited 545 participants on Amazon Mechanical Turk. Participants were paid $\$ 1.25$
and could earn a bonus of up to $\$ 2$. The consent form specified that participants must not have participated in the previous version of this experiment. We excluded 23 participants $(4.2 \%)$ because they had selected fewer than five means.

Results The frequency with which people found the optimal solution decreased significantly with treewidth ( $\rho=$ $-0.4828, p=0.0311$ ). We found that treewidth alone explained $20.41 \%$ of the variance in the frequency with which people found the optimal solution across the 20 graphs $(F(1,18)=4.62, p=0.0455)$. We found that treewidth explained $25.25 \%$ of the variance in solution quality $(F(1,18)=6.08, p=0.0240)$ which significantly deteriorated as treewidth increased ( $\rho=-0.4972, p=0.0257$ ). The median amount of time people took to solve the problems did not increase significantly with treewidth ( $\rho=0.25, p=0.28$ ) and treewidth explained only $0.6 \%$ of our participants' median response times $(F(1,18)=0.10, p=0.76)$. When we restricted the analysis to the time taken by optimal solutions, the relationship was still not statistically significant ( $\rho=0.2994, p=0.1998 ; F(1,18)=0.10, p=0.76$ ). Finally, treewidth explained only $8.8 \%$ of the variance in the perceived problem difficulty across the 20 graphs $(F(1,18)=$ $1.74, p=0.20$ ), and the correlation between treewidth and perceived difficulty was not statistically significant ( $\rho=$ $0.26, p=0.26$ ). Our participants were highly motivated to find the optimal solution (average rating $8.11 \pm 0.06$ out of 9 ). Thus it appears unlikely that their motivation was a bottleneck to their performance. Furthermore, motivation appeared to be unaffected by treewidth $(\rho=-0.03, p=0.91)$. Thus the observed differences in performance appear to result from the inherent difficulties of the means selection problems posed by different goal systems.

In addition, we found that both the size of the graph edge expansion and the graph spectral expansion (measured again as $d-\lambda_{2}$ ) were significantly negatively correlated with the frequency of optimal solutions (edge expansion: $\rho=$ $-0.4802, p=0.0321$, spectral expansion: $\rho=-0.4782, p=$ 0.0330 ), while the average shortest path and average graph eccentricity showed a significant positive relationship (avg shortest path: $\rho=0.4912, p=0.0279$, avg. eccentricity: $\rho=0.4391, p=0.0528$ ). In contrast, only the size of the graph vertex and edge expansions showed a significant correlation with the average solution quality (vertex expansion: $\rho=-0.4431, p=0.0504$, edge expansion: $\rho=-0.4832, p=$ 0.0309 ), suggesting that in general graph treewidth and combinatorial expansions may be more robust predictors of human performance on the MC problem. None of the metrics surveyed were significantly correlated with median participant response times.

## Experiment 3: A more realistic Set-Cover task

While Experiments 1 and 2 capture some of the computational challenges of goal achievement, the tasks were relatively abstract. Experiment 3 addresses this limitation by as-
signing semantic labels to the 24 goals. These labels were common new-years resolutions such as "get in shape" and "earn more money". Similar semantic goals were used in previous research in goal-system theory (Zhang et al., 2007). We also used a different interface to avoid possible visualization effects that arise from graph drawings in the first two experiments.

Methods We recruited 600 participants on Amazon Mechanical Turk. Participants were paid $\$ 0.38$ for about 5 min of work plus a performance-dependent bonus of $\$ 0.50$ if they found an optimal solution. Each participant was randomly assigned to one of the twenty graph structures used in Experiments 1 and 2. For each graph, the order in which the means were listed and the order in which the goals were listed was randomized between participants. The participants' task was to achieve all goals with as few means as possible. The graphical interface of the task was changed to reduce visual clutter. Instead of drawing edges between mean and goals, the goals achieved by each mean were listed next to it (see Figure 3). The cover story was similar to the one used in Experiment 1 but the training trial used the new task interface shown in Figure 3. The consent form required that participants had not participated in any of our previous goal management experiments. All participants were included in the subsequent analyses.

Results On a scale from 1 to 9 participants rated their motivation to find a solution that achieves all goals with the minimal number of means as 7.38 , their motivation to finish the task as quickly as possible and move on as 4.35 , and the difficulty of the task as 5.67. We found that treewidth, the magnitude of the graph spectral expansion, cardinalities of the graph edge and vertex expansions, average eccentricity, average shortest path, and graph diameter were all significantly correlated with the frequency with which human participants identified the optimal solution (treewidth: $\rho=$ $-0.756, p=0.0001$; avg. eccentricity: $\rho=0.583, p=0.007$; avg. shortest path: $\rho=0.651, p=0.002$; graph diameter: $\rho=$ $0.525, p=0.017$; spectral expansion: $\rho:-0.708, p=0.0005$; edge expansion: $\rho=-0.7, p=0.0006$; vertex expansion: $\rho=-0.7303, p=0.0003$ ). Similarly, treewidth, the magnitude of the graph spectral expansion, cardinalities of the graph edge and vertex expansions, average eccentricity, and average shortest path were all significantly correlated with the average solution quality of human responses (treewidth: $\rho=-0.60, p=0.005 ;$ avg. eccentricity: $\rho=0.4872, p=$ 0.0293 ; avg. shortest path: $\rho=0.5243, p=0.0176$; spectral expansion: $\rho=-0.5308, p=0.0160$; edge expansion: $\rho=$ $-0.5670, p=0.0091$; vertex expansion: $\rho=-0.6047, p=$ 0.0047 ). None of the features were significantly correlated with median participant response times.


Figure 3: Graphical Interface of Experiment 3.

## Conclusions

We demonstrated that people's performance in Maximum Coverage and Set Cover can be reliably predicted from graph theoretic measures for the tree-likeness of the goal system, such as treewidth and expansion. Our data support the conclusion that tree-like goal systems are easier for people to handle. More generally, our results imply that parameters that are used in theoretical computer science to differentiate between hard and easy instances can be leveraged to predict human performance in NP-hard tasks.

One limitation of our experiments is that their complete, explicit representation of goals, means, and the connections between them is a simplifying idealization. In real life, people are often unaware of some of their goals and means, as well as some of the connections between goals and means. For example, maintaining goal systems of moderate size in working memory when solving means selection problems is likely to be nontrivial. Hence real-life representations of goals are likely to make means selection problems as those discussed here even more challenging to solve. Although such memory problems are not directly related to how treelike the goal system is, they are nevertheless consistent with our hypothesis that the cognitive difficulty of means selection is an important limiting factor for people's ability to achieve their goals.

In conclusion, our results suggest that even highly motivated people will likely fall short of achieving all their goals when they have to consider many goals and means in parallel. Our analyses provide a novel approach to predicting how likely people are to succeed in these settings, with implications for the design of goal systems that make it easier for people to meet their objectives.

## References

Alber, J., \& Niedermeier, R. (2002). Improved tree decomposition based algorithms for domination-like problems. In Latin american symposium on theoretical informatics (pp. 613-627).
Atkinson, J. W., \& Birch, D. (1970). The dynamics of action. Oxford: John Wiley.

Baumeister, R. F., Heatherton, T. F., \& Tice, D. M. (1994). Losing control: How and why people fail at self-regulation. San Diego: Academic Press.
Carruthers, S., Masson, M. E., \& Stege, U. (2012). Human performance on hard non-euclidean graph problems: vertex cover. The Journal of Problem Solving, 5(1), 5.
Clementi, A. E., \& Trevisan, L. (1999). Improved nonapproximability results for minimum vertex cover with density constraints. Theoretical Computer Science, 225(1), 113-128.
Feige, U. (1998). A threshold of $\ln \mathrm{n}$ for approximating set cover. Journal of the ACM, 45(4), 634-652.
Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. American Psychologist, 54(7), 493-503.
Iwata, Y. (2016). Linear-time kernelization for feedback vertex set. arXiv preprint arXiv:1608.01463.
Kleinberg, J., \& Tardos, É. (2006). Algorithm design. Pearson Education India.
Kloks, T. (1994). Treewidth: computations and approximations (Vol. 842). Springer Science \& Business Media.
Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., \& Sleeth-Keppler, D. (2002). A theory of goal systems. Advances in Experimental Social Psychology, 34, 331-378.
Leyton-Brown, K., Nudelman, E., \& Shoham, Y. (2009). Empirical hardness models: Methodology and a case study on combinatorial auctions. Journal of the ACM, 56(4), 22.
Little, B. R., \& Gee, T. (2007). The methodology of personal projects analysis: Four modules and a funnel. In B. R. Little, K. Salmela-Aro, \& S. D. Philipps (Eds.), Personal project pursuit: Goals, action, and human flourishing. New York: Psychology Press.
MacGregor, J. N., \& Ormerod, T. (1996). Human performance on the traveling salesman problem. Perception \& Psychophysics, 58(4), 527-539.
MacGregor, J. N., Ormerod, T. C., \& Chronicle, E. (2000). A model of human performance on the traveling salesperson problem. Memory \& Cognition, 28(7), 1183-1190.
Miller, G. A., Galanter, E., \& Pribram, K. H. (1960). Plans and the structure of behavior. New York: Adams Bannister Cox.
Mischel, W., Shoda, Y., \& Rodriguez, M. L. (1989). Delay of gratification in children. Science, 244(4907), 933-938.
Muraven, M., \& Slessareva, E. (2003). Mechanisms of self-control failure: Motivation and limited resources. Personality and Social Psychology Bulletin, 29(7), 894-906.
Newell, A., \& Simon, H. A. (1972). Human problem solving. Englewood Cliffs: Prentice-Hall.
Robertson, N., \& Seymour, P. D. (1986). Graph minors. ii. algorithmic aspects of tree-width. Journal of algorithms, 7(3), 309-322.
Thagard, P., \& Millgram, E. (1995). Inference to the best plan: A coherence theory of decision. In A. Ram \& D. B. Leake (Eds.), Goal driven learning (pp. 439-454). MIT Press.
Van Rooij, I. (2008). The tractable cognition thesis. Cognitive Science, 32(6), 939-984.
Vazirani, V. V. (2013). Approximation algorithms. Springer Science \& Business Media.
Wahlström, M. (2014). Half-integrality, LP-branching and FPT algorithms. In Proceedings of the twenty-fifth annual acm-siam symposium on discrete algorithms (pp. 1762-1781).
Zhang, Y., Fishbach, A., \& Kruglanski, A. W. (2007). The dilution model: how additional goals undermine the perceived instrumentality of a shared path. Journal of Personality and Social Psychology, 92(3), 389-401.

# The Effects of Autonomy on Emotions and Learning in Game-Based Learning Environments 

Amanda E. Bradbury, Michelle Taub, \& Roger Azevedo<br>\{aebradbu@ncsu.edu, mtaub@ncsu.edu, razeved@ncsu.edu\}<br>North Carolina State University, Department of Psychology, 2310 Stinson Drive, 640 Poe Hall Raleigh, NC 27695 USA


#### Abstract

The current study examined the impact of agency on college students' emotions and learning during gameplay with CRYSTAL ISLAND, a game-based learning environment designed to foster microbiology learning. 96 undergraduate students ( $59 \%$ female) from a large North American university participated in the study. Participants were randomly assigned to one of three experimental conditions (i.e., full agency, partial agency, no agency), based on the level of control granted during gameplay, and were asked to uncover the source, identity, and best treatment for a mysterious illness. Results revealed participants in the partial agency condition achieved the highest (pre- to post-test) proportional learning gain (PLG), even when controlling for session duration. Additionally, there was a positive correlation between evidence scores of four emotions (anger, fear, confusion, and frustration) and PLG within the partial agency condition-meaning the higher the evidence of the above emotions, the higher the PLG. Further, a stepwise multiple regression showed anger as the sole predictor of PLG. Results from this study have important implications for understanding the role of autonomy and emotions during learning and problem solving with GBLEs designed to foster scientific thinking in STEM. The current study suggests that although GBLEs offer significant learning benefits, they also induce several emotions that can facilitate or inhibit learning gains, requiring further examination.


Keywords: human agency; emotions; learning; game-based learning environments; science

Autonomy is a critical determinant in human learning, problem solving, and performance (Bandura, 2001). Despite its importance in cognitive science, there is a paucity of research that experimentally manipulates autonomy and explores its impact on learning and emotions, in STEM gamebased learning environments (GBLEs). Various levels of autonomy likely affect learners' abilities to monitor and regulate their cognitive, affective, metacognitive, and motivational processes in dynamic, non-linear learning environments involving planning (e.g., coordinating multiple goals), learning activities (e.g., reading scientific texts), and scientific reasoning (e.g., collecting evidence and testing hypotheses) in different ways. Further, little is understood of how autonomy affects emotions in GBLEs, and in turn, how these emotions affect learning outcomes (Azevedo, Taub, Mudrick, Farnsworth, \& Martin, 2016; D’Mello \& Graesser, 2012). Our study focuses on the effects of autonomy on emotions and the impact of both on learning and problem solving within the GBLE, CRystal IsLand.

GBLEs offer powerful platforms to enhance student learning, problem solving, and performance. However, a majority of the research focuses on engagement and motivation and is often criticized for (1) a lack of theoretical framing, (2)
questionable operationalizations of key constructs (e.g., engagement, motivation), (3) overreliance on self-report measures, and (4) dubious empirical support, based on a lack of experimental rigor, methodological shortcomings, and inappropriate analytical techniques (see Mayer, 2014). Additionally, much of this research fails to assess learning gains, choosing to take an "everything but learning" approach, such as measuring engagement or motivation alone while ignoring educational outcomes (Mayer, 2014). Further, GBLEs have been criticized for overshadowing educational content with game elements that are superfluous and distracting to learning goals, drawing learner attention away from important educational content (Mayer \& Johnson, 2010). Interestingly, many of these distractors (e.g., game narratives, interesting characters) are the very elements thought to increase student motivation, engagement and positive emotions (Sabourin \& Lester, 2014). Further, research has indicted that while distractors may present opportunities for off-task behaviors, leading to decreased learning gains (Rowe, McQuiggan, Robison, \& Lester, 2009), off-task behaviors could in-fact be a strategy to alleviate frustration, allowing the student to reduce frustration and thereby increase learning gains (Sabourin, Rowe, Mott, \& Lester, 2014).

Students experience a diverse range of emotions when learning, which likely influence cognitive processes and academic performance (see Calvo, D'Mello, Gratch, \& Kappas, 2014). We address this issue by using online trace methods (e.g., facial expression detection software [FACET; Version 6.2], and logfiles), to assess the impact of autonomy on emotions and learning during gameplay (see Azevedo et al., 2016; Calvo et al., 2014), thereby increasing understanding of emotional monitoring and regulation in GBLEs (Rowe, Shores, Mott, \& Lester 2011). This research can inform the design of future intelligent, adaptive GBLEs that not only teach complex instructional material effectively but also train the skills necessary to successfully monitor and regulate emotions during learning, leading to improved learning outcomes.

## Theoretical Framework

D'Mello and Graesser's (2012) model of affective dynamics suggests certain emotional states arise as the result of an impasse during deep learning, creating cognitive disequilibrium. This model focuses on four learner-centered emotional states: flow/engagement, confusion, frustration and boredom. When learners reach a state of disequilibrium (e.g., during reading complex text), they are likely to experience confusion which if unresolved will likely transition to frustration, which if also left unresolved, will lead to boredom
and disengagement from the activity (e.g., reading, inspecting diagrams). This model posits that students systematically shift between learning-centered states during complex learning and that these shifts are predictive of learning, problem solving, and scientific reasoning. For instance, frustration is much more likely to transition to boredom than to engagement/flow, as learners have not yet transitioned to confusion, where through effortful reasoning and problem solving they can resolve an impasse and return to equilibrium. However, this model has some drawbacks. For instance, it ignores other emotional states such as the seven basic emotions (e.g., anger; Ekman, 1973), assuming that other basic emotions are unimportant to learning. Lastly, this model has not been used to examine autonomy and extended learning with GBLEs such as CRYSTAL ISLAND.

## Current Study

The goal of the current study was to examine the effects of autonomy on emotions and learning during gameplay with GBLEs such as CRYSTAL ISLAND. By experimentally manipulating autonomy, we could empirically observe how different levels of autonomy (e.g., agency conditions) affected learning gains as well as emotional states, and in turn, how these emotional states affected learning gains. Our research questions were as follows: 1) What are the effects of autonomy on proportional learning gains with CRYSTAL ISLAND, after controlling for session duration? 2) What are the effects of autonomy on learners' emotions throughout their interaction with Crystal IsLand? 3) Do evidence scores of emotional states predict PLG during gameplay with CRYSTAL ISLAND and are there differences in emotion evidence scores between high and low performers?

Our hypotheses were as follows. (H1): Participants in the partial agency condition will show significant PLG compared to the full agency and no agency conditions. (H2): The full agency condition will exhibit the highest evidence of positive emotions such as joy and the lowest evidence of negative emotions such as anger and frustration compared to the partial and no agency conditions. (H3): Higher evidence scores of negative emotions such as anger, confusion and frustration will lead to increased PLG in all conditions.

## Method

## Participants

96 undergraduate students ( $59 \%$ female) from a large North American university participated in the current study. Participants' ages ranged from 18 to $29(M=19.99, S D=1.79)$ and were randomly assigned to one of three experimental condition: full agency, partial agency or no agency (see Experimental Procedure). Additionally, they were compensated \$10/hour for participating.

## Materials

At the start of the experimental session, participants read and completed the informed consent, a demographics question-
naire and a series of self-report questionnaires. These questionnaires probed participants' emotions and motivation (e.g., Emotions and Values; Pekrun, Goetz, Frenzel, Barchfeld, \& Perry, 2011) as well as achievement goals (Elliot \& Murayama, 2008). Participants also completed a pretest ( $M=11.94, S D=2.79 ; 57 \%$ correct) and post-test ( $M=$ 13.92, $S D=2.86 ; 66 \%$ correct) on microbiology knowledge: a 21-item, four-choice multiple-choice test, with 12 factual and 9 procedural questions. Participants also completed the Perceived Interest Questionnaire (Schraw, Bruning, \& Svoboda, 1995), Intrinsic Motivation Inventory (Ryan, 1982), and Presence Questionnaire (Witmer \& Singer, 1998).

## Crystal ISLAND

CRyStal IsLand is a narrative-centered GBLE used to foster students' self-regulated learning, scientific reasoning, and problem-solving skills (Rowe et al., 2011). Participants experience the game in first person perspective, arriving on a tropical island where they discover a mysterious illness has infected the community. Taking a protagonist role, participants explore the island, seek clues by speaking to residents and patients, read content on microbiology and use lab equipment to scan for possible transmission sources, all to discover the source, identity, and best treatment for the infectious disease.

Buildings CRYSTAL ISLAND has five buildings, each embedded with a multitude of books, research papers, posters, food items, and non-player characters (NPCs). In the infirmary, participants interview sick patients and interact with the NPC, Kim the camp nurse, who provides the game narrative. Through this interaction, they gather pertinent information such as overall goals, background information, and clues pointing towards possible illness types and transmission sources. In the two living quarters (a dorm room and a microbiologist's home), participants converse with microbiology experts and another patient, and read books and posters on various microbiology topics. In the dining hall, participants meet Quentin the camp cook, who offers insight into what foods he had prepared and sick patients had eaten prior to the outbreak. Using information and clues gathered from these buildings, participants can infer which items are the likely transmission source and then test these hypotheses by scanning these food items in the laboratory.

Game Elements Participants complete concept matrices as they read about microbiology in books and research articles. For example, as they read about E. coli, they must fill in a diagram asking questions related to the reading (i.e., where E. coli is located, symptoms and common diagnostic tests). Additionally, by interacting with NPCs, participants receive valuable information (i.e., evidence), such as symptoms and food eaten. As participants collect evidence and begin making inferences, they can track and organize symptoms, test results, and make a final diagnosis via a diagnosis worksheet. This worksheet supports problem-solving processes by allowing participants to offload information as they interact with the game environment, later using this information to
make a final diagnosis, identify the transmission source, and propose a treatment plan. For instance, they may read about influenza then check the diagnosis worksheet to find the symptoms match the current epidemic. Additionally, participants generate hypotheses regarding which food items are the likely transmission source as well as the type of pathogen they might carry. These hypotheses are tested by collecting and scanning food items, and testing for a virus, bacterium, mutagen, or carcinogen. If a test comes back positive for a pathogenic substance, the participant can confirm the transmission source and add their finding to the diagnosis worksheet. Once participants correctly identify the illness type, transmission source, and treatment plan, the mystery is solved and the game concludes.

## Experimental Procedure

Conditions Participants were randomly assigned to one of three conditions (i.e., full agency, partial agency, no agency) prior to gameplay. These conditions varied in the level of autonomy assigned to each player, ranging from full autonomy (full agency), to some autonomy (partial agency), to no autonomy at all (no agency). In the full condition, participants were free to explore the game environment and its elements as much or as little as they wished, choosing what buildings to visit, what books to read, and with which NPCs to interact. Conversely, the partial condition contained strict game parameters with a pre-set order in which players visited buildings and a requirement that they interact with all game artifacts (e.g., read all books/posters, speak with all NPCs, etc.) before advancing to other buildings. In the no condition, participants did not play CRYSTAL ISLAND but instead watched a narrated video of an expert playing the game. This was an optimal instructional path designed to enhance learning without the opportunity to exercise autonomy as participants had no control over any aspect of the gameplay or content.

Experimental Procedure The experimental session lasted one to two and a half hours depending on condition ( $M=$ $89.64 \mathrm{~min}, S D=18.37 \mathrm{~min}$ ). Upon arrival, participants were greeted, directed to the workstation and asked to review and complete the informed consent. Next, they received an overview of the study, donned an electro dermal activity (EDA) bracelet (Empatica E4), and completed the microbiology pretest. Then, the SMI RED 250 eye tracker was calibrated using a 9-point calibration. Following successful calibration, a baseline for the facial recognition of emotion software (FACET) and EDA were established using Attention Tool (Version 6.2). Participants were then given instructions for the experimental session that included an overview of the game scenario covering their role as the protagonist, the importance of reading (i.e., books, articles, and posters), interacting with NPCs and scanning food items to solve the mystery. During gameplay, we collected logfiles, eye-tracking, facial expressions of emotions, and physiological data on all participants in the full and partial agency conditions only. Upon game conclusion, participants completed several self-
report measures and the microbiology post-test, after which they were debriefed, thanked, and paid for their time.

## Coding and Scoring

For the purposes of the current study, only logfiles and FACET data were used. Additionally, pre- and post-test scores (out of 21 possible points) of microbiology content knowledge were used to generate a PLG score (see below).

Logfiles Logfile data captured the sequence and timing of participants' movements and actions within the game (e.g., talking to NPCs, reading books). For this study, only session duration was analyzed. This variable was extracted from the trace data. Additionally, logfile data were only captured in the full and partial agency conditions as the no agency condition watched a video play-through ( 91 min ) of CRYSTAL IsLAND rather than play, thus not generating any log-file data.

Facial Expression Data Each experimental session included a video of the participant, which was later analyzed using FACET, facial expression recognition software included with Attention Tool. We used FACET (sampling rate of 30 Hz ) to analyze the following nine basic and learning-centered emotions: joy, anger, contempt, frustration, confusion, surprise, fear, sadness and disgust (see, Dente, Küster, Skora, \& Krumhuber, 2017, regarding the software's validity). Each emotion was given an evidence score automatically generated by FACET representing the likelihood of an expert human coder to similarly categorize the expression. This score was based on a logarithmic scale (base 10), meaning that a score of one indicated the likelihood of 10 human coders coding for that emotions while a score of two indicated the likelihood of 100 human coders coding for that emotion, and so forth. For the purposes of the current study, the mean evidence score for the entire session duration was used for each participant. The range of evidence scores for all emotions and across participants was 0 to 1.98 , excluding negative values. Negative scores indicated the emotion was not likely present, and since we were interested in emotions present, all negative values were replaced with zero.

Proportional Learning Gain (PLG) PLG scores were calculated from pre- and post-test ratios scores of microbiology content knowledge, using Witherspoon, Azevedo, and D'Mello's (2008) formula. For example, if a participant scored an 11 out of 21 on the pre-test and a 15 out of 21 on the post-test then their PLG score was .40 .

Median Split High versus low performers were determined through a median split of the PLG variable for the partial agency condition. The median for this condition was .40 (range: -0.17 to 0.70 ).

## Results

Research Question 1: What are the effects of autonomy on proportional learning gains with Crystal IsLAND, after controlling for session duration?
To investigate the effects of autonomy on PLG, we conducted an ANCOVA, using condition as the independent variable and session duration as a covariate, see Table 1 for mean session duration by condition. Results indicated a significant main effect for condition, $F(2,88)=3.35, p=.003, \eta_{p} 2=.13$. Post hoc LSD analyses indicated that the partial agency condition ( $M=.35, S D=.23$ ) showed significantly higher PLG than both the full ( $M=.18, S D=.27$ ) and no agency conditions ( $M=.11, S D=.28$ ); however, there was no difference between the full and no agency conditions.

Table 1. Mean session duration (min) by condition.

|  | Full Agency <br> $M(S D)$ | Partial Agency <br> $M(S D)$ | No Agency <br> $M(S D)$ |
| :--- | :---: | :---: | :---: |
| Session <br> Duration | $78.69(21.92)$ | $98.65(18.43)$ | $91.00(0)$ |

## Research Question 2: What are the effects of autonomy on learners' emotions throughout their interaction with CRYSTAL ISLAND?

A MANCOVA was conducted using mean evidence scores of the basic and learner-centered emotions as the nine dependent variables and condition as the one independent variable. No significant main effect was found by condition; Wilk's $\lambda=.78, F(16,164)=1.39, \eta_{\mathrm{p}}^{2}=.12$. Univariate results revealed that disgust, $F(2,89)=4.15, p=.02, \eta_{p} 2=.09$, anger, $F(2,89)=4.12, p=.02, \eta_{\mathrm{p}} 2=.02$, and joy $F(2,92)=$ $3.48, p=.04, \eta_{p} 2=.07$, showed statistically significant differences between conditions. No other emotions demonstrated significant differences. Post hoc LSD analyses indicated that those in the full agency condition exhibited higher levels of disgust ( $M=.22, S D=.34$ ) and anger $(M=.55, S D$ $=.62$ ) compared to those in the partial agency condition $(M$ $=.14, S D=.24 ; M=.37, S D=.49$, respectively). Additionally, those in the full agency condition exhibited higher levels of joy ( $M=.25, S D=.44$ ) compared to the partial agency condition ( $M=.06, S D=.13$; see Table 2).

Table 2. Mean emotion evidence scores by condition.

|  | Experimental <br> Conditions |  |  |  | F-test Results |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Emotional <br> State | Full <br> Agen | Part <br> Agen | No <br> Agen | $F$-Stat | Comparisons |
|  | $M(S D)$ | $M(S D)$ | $M(S D)$ | $F(p)$ |  |
| Disgust | .22 | .14 | .04 | 4.15 | $(\mathrm{P}=\mathrm{F}>\mathrm{N}=\mathrm{P})$ |
|  | $(.34)$ | $(.24)$ | $(.12)$ | $(.02)$ |  |
| Anger | .55 | .37 | .18 | 4.12 | $(\mathrm{P}=\mathrm{F}>\mathrm{N}=\mathrm{P})$ |


|  | .25 | .06 | .09 | 3.48 | $(\mathrm{~F}>\mathrm{P}=\mathrm{N})$ |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Joy | $(.44)$ | $(.13)$ | $(.27)$ | $(.04)$ |  |
| Frustra- | .38 | .20 | .16 | 2.88 | $(\mathrm{P}=\mathrm{F}>\mathrm{N}=\mathrm{P})$ |
| tion | $(.49)$ | $(.32)$ | $(.31)$ | $(.06)$ |  |
|  | .16 | .19 | .11 | .48 |  |
| Surprise | $. .36)$ | $(.32)$ | $(.27)$ | $(.62)$ | $(\mathrm{F}=\mathrm{P}=\mathrm{N})$ |
|  | .18 | .10 | .09 | 1.45 | $(\mathrm{~F}=\mathrm{P}=\mathrm{N})$ |
| Fear | $(.31)$ | $(.15)$ | $(.20)$ | $(.24)$ |  |
|  | .06 | .06 | .05 | .05 | $(\mathrm{~F}=\mathrm{P}=\mathrm{N})$ |
| Contempt | $(.13)$ | $(.12)$ | $(.14)$ | $(.95)$ |  |
|  | .23 | .23 | .18 | .32 | $(\mathrm{~F}=\mathrm{P}=\mathrm{N})$ |
| Sadness | $(.28)$ | $(.29)$ | $(.31)$ | $(.73)$ |  |
|  |  |  |  |  |  |
| Confusion | .45 | .33 | .26 | 1.30 | $(\mathrm{~F}=\mathrm{P}=\mathrm{N})$ |
|  | $(.52)$ | $(.40)$ | $(.46)$ | $(.28)$ |  |

Note: $\mathrm{F}=$ full agency, $\mathrm{P}=$ no agency, $\mathrm{N}=$ no agency conditions

## Research Question 3: Do evidence scores of emotional states predict PLG during gameplay with CRYSTAL ISLAND and are there differences in emotion evidence scores between high and low performers?

To assess the relationship between emotions and PLG while playing CRYSTAL ISLAND, four correlation matrices were created: overall (all conditions; $n=92$ ), full agency ( $n=$ $30)$, partial agency $(n=32)$, and no agency $(n=30)$. The full and no agency conditions as well as all conditions combined showed no correlations between emotions and PLG; however, for the partial condition, four emotions were significantly positively correlated with PLG, anger, $r(30)=.39$, $p=.03$, fear, $r(30)=.36, p=.04$, confusion, $r(30)=.39, p=$ .03 , and frustration, $r(30)=.39$, meaning the higher the evidence of the above emotions, the higher the PLG.

To determine the predictive power of anger, fear, confusion, and frustration on PLG within the partial agency condition, a stepwise multiple regression analysis was conducted. Results indicated that anger ( $\beta=.39, p=.03, R^{2}=.15$ ) was the sole predictor of PLG, meaning that more evidence of anger predicted better PLG, accounting for $15 \%$ of the variability in PLG.

Given the regression results for the partial agency condition, we performed a median split on these participants' PLG to examine whether there were differences between high- and low-performers' experienced emotions. Result of an independent samples $t$-test revealed that high performers exhibited significantly more evidence of facially expressed frustration, $t(18)=-3.75, p<.002, d=-1.78$, anger, $t(19)=-3.47$, $p<.003, d=-1.58$, and confusion, $t(21)=-2.97, p<.007, d$ $=-1.29$, compared to low performers.

## Discussion

Results of the current study revealed that students achieved the highest PLG in the partial agency condition compared to the full and no agency conditions, even after controlling for sessions duration. These results support H1, demonstrating the positive impacts of seceding partial agency to improve learning outcomes in GBLEs. Previous research explains that while offering a high degree of user control allows learners
to regulate their own learning, constructing knowledge based on the representations they find useful, this responsibility can lead to disorientation and negative learning outcomes when learners are unsure which path to follow (Greene, Bolick, \& Robertson, 2010), suggesting there may be on optimal level of autonomy to improve learning outcomes in GBLE. Future research should empirically test different parametrization of autonomy on GBLEs to assess the optimal level of autonomy to foster learning across domains.

For research question two, participants in the full agency condition were more emotionally expressive than those in no and partial agency conditions. For instance, those in the full agency condition showed significantly higher evidence of joy than those in the partial and no agency conditions, as well as significantly higher evidence of anger and disgust compared to the no agency condition. These results run contrary to our original hypothesis (H2), expecting the full agency condition to experience the least negative emotions; however, the full agency condition did experience the highest evidence of joy, partially supporting H2. A plausible explanation could be that those in the full agency had a greater potential to express autonomy which led to more emotional expressivity throughout task performance (Azevedo et al., 2016). A next step involves a micro-level analysis mapping specific game events (e.g., reading books, testing evidence, etc.) with emotional expressivity (e.g., higher evidence scores) and emotional states.

As for research question three, no correlations between emotional states and PLG were found with 1) all conditions combined, 2) the full agency condition or 3) the no agency condition; however, this was not the case within the partial agency condition. The partial agency condition found significant positive correlations between PLG and evidence scores of facially expressed anger, fear, confusion and frustration, meaning the higher evidence of the above emotions, the higher a participant's PLG. After imputing the aforementioned emotions into a stepwise multiple regression conducted within the partial agency condition, anger was the sole predictor of PLG. Further, high performers in the partial agency condition exhibited significantly higher evidence of anger, frustration and confusion compared to low performers, demonstrating that negative emotions, typically thought as unconducive to learning (Sabourin \& Lester, 2014), can have positive effects on learning outcomes. Previous work has reach similar conclusions, finding confusion, if appropriately regulated and resolved, as beneficial to learning (D'Mello, Lehman, Pekrun,\& Graesser, 2014).

In the current study, fear, anger, frustration and confusion had a positive effect on PLG, but only when the participant seceded partial control of the learning environment (i.e., partial agency condition). One explanation for these results could be explained using the model of affective dynamics (D'Mello \& Graesser, 2012). For instance, participants are likely to experience confusion and frustration when learning difficult subject matter and will hence experience cognitive disequilibrium (D'Mello \& Graesser, 2012). Equilibrium (e.g., engagement/flow state) is regained
through effortful reasoning, problem solving and reflection; however, when left unresolved, learners can digress from confusion to frustration and eventually disengage from the learning activity (D'Mello et al., 2014).

In the current study, participants were asked to learn new information in order to solve complex problems: what disease was infecting the community, what was the transmission source, and how to best treat patients. However, each condition offered different paths to learn this information (via varying levels of autonomy) and in turn affected emotions and learning differently. For instance, in the no agency condition, participants might have felt frustrated at not being able to play the game and this frustration may have led to boredom and disengagement, explaining poor PLG. In the full agency condition, participants could reduce confusion and frustration by simply avoiding books, research articles, or interactions with aspects of the game they found unappealing; however, even though they would return to equilibrium through these actions, they would have missed valuable educational content, thus reducing PLG. Conversely, the partial agency condition was forced to interact with all elements of the game before leaving a room. This stipulation may have forced participants to work through the confusion and frustration they experienced because they could not progress with the not step of the game until required actions (e.g., finishing a conversion with the NPC, filling in a concept matrix correctly) were completed. Therefore, these participants were more likely to engage in the effortful reasoning and problem solving necessary for both deep learning and a return to equilibrium.

## Limitations

There were a number of limitations with the current study. First, the operationalization of autonomy in the partial and full agency condition as this was a first attempt to parameterize key assumptions of autonomy in a GBLE. Also, because we were looking at autonomy, there may have been other metacognitive processes (e.g., motivation) affecting learning gains that we did not control for or measure, we only used log-files, FACET, and learning outcomes data. Converging these data along with EDA and eye-tracking data would further elucidate the role of autonomy and emotions during learning. For instance, eye-tracking data could be used to examine what activity a participant was engaged in prior, during and after the onset of a certain emotion. Additionally, EDA data could be used to validate the presence and relevance of emotions. For instance, spikes in EDA data could be mapped onto emotion evidence scores to determine when spikes and high emotion evidence scores co-occur revealing the quality of appraisals mechanisms (Gross, 2015).

## Implications and Future Directions

These results have important implications for understanding the role of autonomy and emotions during learning and problem solving with GBLEs designed to foster scientific thinking in STEM. The current study suggests GBLEs induce several basic and learning-centered emotions depending on
the level of autonomy granted to a learner and that autonomy and emotions can either facilitate or inhibit learning. However, further empirical examination is required. Future research should design and test additional experimental manipulations that operationalize key assumptions of autonomy (Bandura, 2001). Further, our results revealed a need to extend models and theories of affect to include basic emotions when considering transitions between emotional states in learning environments (D'Mello \& Graesser, 2012) and would benefit by including Gross's process model of emotional regulation along with emotion regulation strategies (Gross, 2015).

Methodologically, converging the multimodal multichannel data will allow researchers to examine the impact of autonomy on emotions and their impact on learning, problem solving, and reasoning. For example, how do emotions fluctuate during different activities during learning with GBLEs? What is their specific behavioral signature in terms of onset/trigger event, intensity, duration, evidence of emotion regulation strategy, and so forth? How do these emotions related to specific GBLE activities (e.g., reading books and posters, interviewing patients, interpreting results, deriving hypotheses)? Such questions can be addressed by traditional statistics as well as data mining and machine learning techniques and lead to the design of intelligent GBLEs capable of detecting, tracking, modeling, and fostering adaptive, realtime scaffolding to learners, depending on their individual needs, thus ensuring optimal learning.

## Acknowledgments

This research has been supported by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006) and members of the SMART Lab and the IntelliMedia Group.

## References

Attention Tool (Version 6.2) [Computer software]. Boston, MA: IMotions.
Azevedo, R., Taub, M., Mudrick, N., Farnsworth, J., \& Martin, S. (2016). Using research methods to investigate emotions in computer-based learning environments. In P. Schutz \& M. Zembylas (Eds.), Methodological advances in research on emotion and education. Amsterdam, The Netherlands: Springer.
Bandura, A. (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52, 1-26.
Calvo, R. A., D'Mello, S., Gratch, J., \& Kappas, A. (Eds.). (2014). The Oxford handbook of affective computing. Oxford, England: Oxford University Press.
Dente, P., Küster, D., Skora, L., \& Krumhuber, E. (2017). Measures and metrics for automatic emotion classification via FACET. Proceedings of the Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour.
D'Mello, S., \& Graesser, A. (2012). Dynamics of affective states during complex learning. Learning and Instruction, 22, 145-157.

D’Mello, S., Lehman, B., Pekrun, R., \& Graesser, A. (2014). Confusion can be beneficial for learning. Learning and Instruction, 29, 153-170.
Elliot, A. J., \& Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. Journal of Educational Psychology, 100, 613-628.
Ekman P. (ed) (1973). Darwin and facial expression. New York, NY: Academic Press.
Greene, J. A., Bolick, C. M., \& Robertson, J. (2010). Fostering historical knowledge and thinking skills using hypermedia learning environments: The role of self-regulated learning. Computers \& Education, 54, 230-243.
Gross, J. J. (2015). Emotion regulation: Current status and future prospects. Psychological Inquiry, 26, 1-26.
Mayer, R. E. (2014). Computer games for learning: An evi-dence-based approach. Cambridge, MA: MIT Press.
Mayer, R. E., \& Johnson, C. I. (2010). Adding instructional features that promote learning in a game-like environment. Journal of Educational Computing Research, 42, 241-265.
Pekrun, R., Goetz, T., Frenzel, A., Barchfeld, P., \& Perry, R. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). Contemporary Educational Psychology, 36, 3648.

Rowe, J., McQuiggan, S., Robison, J., \& Lester, J. (2009). Off-task behavior in narrative-centered learning environments. In S. Craig \& D. Dicheva (Eds.), Proceedings of the 14th International Conference on Artificial Intelligence in Education. Berlin, Germany: Springer-Verlag.
Rowe, J., Shores, L., Mott, B., \& Lester, J. (2011). Integrating learning, problem solving, and engagement in narra-tive-centered learning environments. International Journal of Artificial Intelligence in Education, 21, 115-133.
Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. Journal of Personality and Social Psychology, 43, 450.

Sabourin, J. L., \& Lester, J. C. (2014). Affect and engagement in game-based learning environments. IEEE Transactions on Affective Computing, 5, 45-56.
Sabourin, J. L., Rowe, J. P., Mott, B. W., \& Lester, J. C. (2013). Considering alternate features to classify off-task behavior as emotion self-regulation: A supervised learning approach. Journal of Educational Data Mining, 5, 9-38.
Schraw, G., Bruning, R., \& Svoboda, C. (1995). Sources of situational interest. Journal of Literacy Research, 27, 1-17.
Witmer, B. G., \& Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. Presence: Teleoperators and Virtual Environments, 7, 225-240.
Witherspoon, A., Azevedo, R., \& D'Mello, S. (2008). The dynamics of self-regulatory processes within self- and externally regulated learning episodes during complex science learning with hypermedia. In B. P. Woolf, E. Aïmeur, R. Nkambou, \& S. Lajoie (Eds.), Intelligent Tutoring Systems 2008. Berlin, Germany: Springer.

# A Computational Model of the Role of Attention in Subitizing and Enumeration 

Gordon Briggs (gordon.briggs.ctr@nrl.navy.mil) Will Bridewell (will.bridewell @ nrl.navy.mil) Paul F. Bello (paul.bello@nrl.navy.mil)<br>Naval Research Laboratory<br>Washington, DC 20375 USA


#### Abstract

Recent studies in the perception of numerosity have indicated that subitizing (the rapid and accurate enumeration of small quantities) requires attention. We present a novel computational model of enumeration in which attention unifies distinct processes of numerosity approximation, subitizing, and explicit counting. We demonstrate how this model accounts for both the reaction time results from the subitizing literature and the effects of attentional load on subitizing accuracy.


Keywords: attention; subitizing; enumeration; perception of numerosity; counting; inattentional blindness

## Introduction

The perception of numerosity is one of the core faculties underlying much of the human ability to represent, reason, and communicate about number (Dehaene, 2011), and the precision and accuracy of this "number sense" is highly dependent on the amount of attention devoted to it. Imagine that you are asked to report the size of a small crowd gathered around a performer in a public park. Glancing at the crowd might suffice to obtain a rough figure (e.g., about thirty), but obtaining an exact number with high confidence would require looking at individual people and counting them. Converging lines of evidence suggest that numerical cognition is supported by two types of representations of numerosity (Feigenson, Dehaene, \& Spelke, 2004). The first is an approximate representation that can be quickly perceived through estimation. The second is a precise representation, associated with linguistic representations of number (Gelman \& Butterworth, 2005) that are usually generated through a slower process of explicitly fixating on and counting individual objects.

Yet, a simple dual-system account of numerosity judgment is incomplete. Since the 19th century, scientists have observed that the enumeration of small quantities of objects (under five) was both rapid and accurate (Jevons, 1871). This phenomenon is called subitizing (Kaufman, Lord, Reese, \& Volkmann, 1949). Within the subitizing range, each additional object requires only $40-100 \mathrm{~ms}$ more time to enumerate on average, whereas in the post-subitizing range, each additional object requires $250-350 \mathrm{~ms}$ to enumerate (Trick \& Pylyshyn, 1994). Previous work has proposed that subitizing is the result of a parallel and pre-attentive process (Trick \& Pylyshyn, 1994; Mandler \& Shebo, 1982) that transitions into slower, serial counting.

In contrast with this account, recent studies provide a body of evidence demonstrating that attentional manipulations, such as attentional blink or increased attentional load, adversely affect subitizing performance (Railo, Koivisto,

Revonsuo, \& Hannula, 2008; Olivers \& Watson, 2008; Egeth, Leonard, \& Palomares, 2008; Vetter, Butterworth, \& Bahrami, 2008). Therefore, subitizing has an attentional component, and an account of the processes underlying subitizing must explain both (a) why subitizing is significantly more rapid than explicit counting and (b) why it requires attention.

We claim that numerosity judgments of all types are subject to constraints on attention and that attention flexibly integrates the results of multiple number-processing capacities. Both classical and newer results in the subitizing literature can be explained by combining a capacity-limited, objectbased view of attention with a capacity for approximation that operates independently of attention. In this paper, we present a computational model of enumeration in which attention unifies distinct processes of numerosity approximation, subitizing, and explicit counting. We begin by introducing a computational cognitive modeling system, ARCADIA (Bridewell \& Bello, 2016b), in which attention features centrally. We then proceed to present a computational account of enumeration, proposing a model of rapid, serial subitizing in which the numerosity of objects encoded in visual short-term memory can be quickly established. We then demonstrate how this model of subitizing accounts for both the reaction time and the effects of attentional load. Finally, we discuss the various processes involved in the perception of numerosity and the role of attention in enabling and integrating them.

## ARCADIA's Architectural Features

ARCADIA provides a system in which attention is the primary organizing mechanism for perception, cognition, and action. In this section, we provide a brief overview of the key concepts necessary to understand ARCADIA models and point to a more detailed introduction by Bridewell and Bello (2016b). Each ARCADIA model consists of a set of components, which carry out all the processing for the model, that can read from or write to accessible content, a temporary buffer where components share their output on each cycle. Whereas representations inside of individual components can take arbitrary form, every component is designed to read from and express its results in a common representation called the interlingua. This shared language pulls together data in different formats so that they can be exchanged among components.

ARCADIA operates in discrete, cognitive cycles, each corresponding to 25 ms intervals (Bridewell \& Bello, 2016a). During each cycle, an attentional strategy selects an element
out of accessible content to be the focus of attention. Some components can respond to the focus of attention, while others operate independently of the focus. In this way, ARCADIA has natural resources for distinguishing processes that require attention from those that do not.

## Model of Enumeration

In this section, we present an ARCADIA model of numerosity judgment that includes both estimation and counting processes. The core processes within this model are illustrated in Figure 1. We model counting such that differences in enumeration speed result from the attentional and temporal constraints necessitated by maintaining count information in different memory stores. The model proposes that a count of the first set of seen objects can be calculated from visual shortterm memory (vSTM) and represents the subitized count. After this, an explicit, symbolic representation of the current count is maintained that requires rehearsal via subvocalization (i.e., inner speech) supported by a phonological loop.

The connection between subvocalization and slow counting has been previously established (Dehaene, 2011), and evidence has shown that articulatory suppression adversely affects counting performance (Logie \& Baddeley, 1987). Likewise, vSTM capacity-limits have been proposed as an explanation of subitizing (Railo et al., 2008), and there is evidence that vSTM load affects subitizing performance (Cutini \& Bonato, 2012).

## Visual Processing

For the sake of brevity, our discussion of these components is abridged. A more detailed walkthrough of visual processing mechanisms in ARCADIA can be found in Bridewell and Bello (2016a).
Early Vision: At the first stage of visual processing, image segmentation produces a set of segments, corresponding to hypotheses about object locations that we call proto-objects (Rensink, O’Regan, \& Clark, 1997). [Step 1].
Fixation Generation: Highlighter components produce candidate fixations on either individual proto-objects or a group of proto-objects. When a candidate fixation on a single protoobject becomes the focus of attention, an object-binding process begins [Step 2a]. If a candidate fixation for a group of proto-objects becomes the focus of attention, an estimation process is triggered [Step 2b].
Object Binding and Storage (vSTM): When a candidate fixation containing a single proto-object is focused on, a (visual) feature-binding process occurs, creating an object representation containing size, shape, color, and location information. This object representation can in turn be focused on, which updates vSTM. The ARCADIA model of vSTM has a maximum capacity of four objects (Bridewell \& Bello, 2016b).
Task Knowledge (Visual Cues): Task knowledge includes object property information used as visual cues to identify and distinguish between different types of task-relevant objects. This includes size and shape information about objects
that should be enumerated and mask objects that represent the end of a trial.

## Enumeration Components

Various components are responsible for subitizing and subvocal counting, estimation, and then merging these results into a single numerosity report.
Subitized Numerosity: When vSTM is full, or a mask object is detected, or when there are no more uncounted objects, then the number of objects to be enumerated in vSTM is returned. Note, this may be based on a subset of the objects in vSTM. Evidence supports the notion that people are able to selectively enumerate objects in vSTM based on various object properties (Chesney \& Haladjian, 2011).
Lexical Count: For precise numerosity representations to be remembered, they first must be converted into a lexicalized form (e.g., "one," "four," "eight"). This happens when a subitized numerosity representation is focused on [Step 3]. When a lexicalized count representation already exists in working memory, and a new object is focused on, the next lexicalized number in the count sequence is returned [Step 5]. The presence of lexicalized number representations triggers subvocalization in the phonological loop.
Phonological Loop: The phonological loop is implemented with a component that generates a series of subvocalization actions over multiple cycles, ensuring that the number of sequential subvocalization actions corresponds to a model of subvocalization time. Currently, the number of ARCADIA cycles required to subvocalize a word is calculated based on the approximation from Huss and Byrne (2003): 150 ms per syllable and one syllable per three characters in a lexeme. For example, "seventeen" would have three estimated syllabes for a total subvocalization time of 450 ms .
Approximate Number System (ANS): The ANS responds when the focus is a fixation on a group of proto-objects, by producing a noisy number-sense representation [Step 4]. The result is a normal distribution with a mean of $n_{e}$ (the number of proto-objects in the group fixation region) and standard deviation of $w \cdot n_{e}$, where $w$ is the model's Weber fraction parameter (Halberda \& Feigenson, 2008).
Numerosity Reporter: The numerosity reporter is responsible for merging both the results from the ANS and the serial count into a single numerosity judgment. If time allows for an explicit count to be fully generated, the explicit count is recorded. Otherwise, an educated guess is made:

$$
\operatorname{Guess}\left(n_{c}, n_{e}, w\right)=n_{c}+\operatorname{sample}\left(\mathcal{N}\left(n_{e}-n_{c}, \sqrt{w \cdot\left(n_{e}-n_{c}\right)}\right)\right)
$$

where $n_{c}$ denotes the number of explicitly counted objects. This reflects the basic phenomenon that enumeration error decreases gradually in conjunction with the number of items able to be explicitly counted (e.g., Mandler \& Shebo, 1982; Railo et al., 2008). Whether this can be fully explained by approximation on uncounted fixation candidates or approximate mathematical operations on partial results (e.g., Gallistel \& Gelman, 2000) is a topic for further investigation.


Figure 1: Information flow within the ARCADIA model of enumeration. Red lines indicate information flow that requires focus of attention.

## Modeling Rapid vs. Slow Counting

In this section, we demonstrate the application of the ARCADIA numerosity perception model to a basic enumeration task (the enumeration of one to eight objects with no effective time limit). Our goals in examining this base case are to investigate and account for the origins of the bilinear reaction time curve found in the subitizing literature. To evaluate the model, we generated 40 videos, each containing a randomized, irregular pattern of one to eight (non-overlapping) circles. There were five videos for each number of circles, and the model was run 10 times for each video.

## Attentional Strategy

Attentional strategies in ARCADIA determine which element from accessible content becomes the focus of attention on a given cognitive cycle. The attentional strategy for the enumeration model embodies two constraints and prioritizes selection to jointly satisfy them. First, each task-relevant object must be counted. As such, after a new candidate proto-object is focused on, priority is given to processing that updates the count. Second, objects must not be counted more than once. Candidate proto-object fixations are ordered in a left-to-right manner, and the position of the last counted object is tracked.

## Results

Figure 2 provides a plot of the simulated reaction times from the ARCADIA model of counting ( 25 ms per cycle) with human subject results from Trick and colleagues (1996) . The simulated RTs from the model are consistent with human subjects from the 22 year-old group, $r^{2}=.990, p<.001$.

The model predicts that enumerating a single object requires at least 375 ms (or 15 ARCADIA cycles). The first seven cycles are needed to generate a lexicalized representation of numerosity for report. The first cycle is needed for
image processing to occur and fixation candidates to be generated [Step 1]. The first fixation candidate is focused on in the second cycle [Step 2a]. The third cycle is required for object binding, and the fourth cycle is required to encode the object representation into vSTM. The fifth cycle is required for the subitized numerosity process to determine that there are no more unvisited candidate fixations. Finally, the sixth and seventh cycles are required for the subitized numerosity process to produce an object count from vSTM [Step 3] and for this numerosity representation to be converted into lexicalized form. The remaining eight cycles are required for subvocalization and generation of the final numerosity report.

When vSTM is below capacity (not filled to its four item limit) only 50 additional milliseconds are required for each additional item to be enumerated, which is consistent with the $40-100 \mathrm{~ms}$ per additional item result from the subitizing literature (Trick \& Pylyshyn, 1994). During subvocalized counting each new enumerated object necessitates updating the last enumerated point in working memory (to keep track of which points were counted) and an explicit subvocalization of the updated count. This additional attentional requirement adds roughly $250-350 \mathrm{~ms}$ per item ( 75 ms for object binding and inhibition updating and $175-275 \mathrm{~ms}$ for subvocalization and number report). Simulated RTs from the ARCADIA model are more consistent with previous human studies that the Peterson and Simon (2000) ACT-R model of subitizing (SUBIT-R), which predicts enumerating one to two objects as taking roughly 200 ms and over 1000 ms for four objects.

## Modeling Enumeration: Attentional Effects

As a second test of the model, we replicate the results from Railo and colleagues (2008). In that study, the authors used a paradigm originally applied to study inattentional blindness (Rock, Linnett, Grant, \& Mack, 1992). Subjects had two
potential tasks: (1) report which line of a centrally located cross is longer and (2) report the number of dots clustered in a quadrant outside the central cross. Videos consisted of a series of trials in which the cross appeared, with a critical trial in which a peripheral dot cluster appeared for the first time (and subjects were unaware of the enumeration task).

There were three experimental conditions (for more detail see Bridewell and Bello, 2016a). First, the inattention condition consisted of the critical trial. Subjects were asked first to report the results from the length comparison task. After the third, critical trial, subjects were asked whether they noticed any dots and, if so, to report their quantity. Next, the dividedattention condition consisted of trials in which subjects were asked to perform both tasks. Finally, the full-attention condition consisted of trials in which subjects were told to ignore the length comparison task and focus only on enumeration. In the inattention condition, subjects who were not inattentionally blind to the dots ( $\sim 80 \%$ ) had enumeration accuracy close to $100 \%$ for up to two dots, after which accuracy dropped to under $25 \%$. In the divided attention condition, enumeration accuracy was at or near $100 \%$ for one and two dots, after which accuracy more gradually declined. In the full attention condition, enumeration accuracy was at or near $100 \%$ for up to three dots before beginning a decline. Because the attentional manipulation is the result of a dual task, additional components are necessary to allow the model to produce results for both, which we outline below.

## Model Configuration

We generated 36 videos to serve as stimuli for our simulation of the task environment with six videos for each number of peripheral dots. Each video consisted of a fixation cross presented for 1500 ms followed by a 200 ms stimulus interval with a centrally located cross and peripheral dot cluster. After the stimulus interval in each video, a mask was displayed for 500 ms . The model was run on each video 20 times.

Additional Components Bridewell and Bello (2016a) present a computational model of inattentional blindness applied to similar stimuli. The current stimuli use the same paradigm, and as such many of the components from that early model are reused. ${ }^{1}$

Attentional Strategies As in the previous enumeration model, ARCADIA follows a left-to-right prioritization of candidate fixations for enumeration. However, in the inattention and divided attention conditions, peripheral fixations are inhibited until a length comparison result is encoded in working memory (ensuring enumeration processing occurs after the primary task is complete). After this, in the divided attention condition, group fixations are given prece-

[^286]

Figure 2: Simulated RT for ARCADIA enumeration model compared with human data from Trick and colleagues (1996).
dence over individual fixations (until an ANS result is memorized), whereas in the inattention case, group fixations are not prioritized (reflecting lack of awareness of the dot clusters and the time limited enumeration task). In the full attention condition, peripheral inhibition is absent, length comparison task elements are not prioritized, and group fixations are prioritized over individual fixations.

## Results

Bridewell and Bello (2016a) analyzed focus traces generated by ARCADIA to demonstrate the precise effect attentional and temporal constraints had on the processing of inattentional blindness stimuli. In Figure 3, we present similar focus traces for each experimental condition in the dual task. Because the stimulus interval for the dual task was 200 ms , only eight ARCADIA cycles were available for completing both tasks. The primary (length comparison) task required four cycles to accomplish, leaving only four cycles for any enumeration or secondary processing. The amount of enumeration processing in these four cycles (or eight in the case of the full attention condition) influences how accurate the numerosity judgment could be.

In the inattention condition, after the primary task is accomplished, there is time to fixate on and encode two dots into vSTM. However, these fixations are incidental, and because the attentional strategy in the inattentional case does not prioritize focus on any group fixations, there are no resulting estimation results from the ANS. As such, there is no basis for an educated guess. If there were only one or two dots, then a correct numerosity judgment would be made. Otherwise, the system would generate an incorrect report. In constrast, the system generates an estimate when the attentional strategy is configured for the divided attention condition. However, there is time to generate and memorize the ANS estimate only and not to begin a serial estimation process. Therefore, accuracy begins to drop after one item, which corresponds to the increased noise associated with the ANS judgment (and subsequent guess inaccuracy). Finally, in the full attention condition, estimation can occur in the time that would have


Figure 3: Focus traces from ARCADIA model of Railo et al. (2008) dual-task in different attentional conditions.
otherwise been required for the primary task. This provides time for two dots to be fully encoded in vSTM and explicitly enumerated. As such, performance does not begin to significantly decline until there are four dots or higher. ${ }^{2}$

The resulting performance curve can be found in Figure 4. A Weber fraction $w$ of 0.13 was used, which is consistent with the range observed in normal adults (Halberda \& Feigenson, 2008). Fischer's exact tests showed that accuracy in the full attention condition was significantly higher than in the divided attention condition for dot numbers of two ( $p=.002$ ), three ( $p<.001$ ), four ( $p<.001$ ), five ( $p<.001$ ), and six ( $p<.001$ ). Qualitatively, this matches the previously described attentional effects from the human subjects, with exception of the two dot results. Railo and colleagues did not find significant differences in performance for three or five dots, but ascribed this to potential perceptual difficulties in their stimuli. Our model may underestimate accuracy for two dots in the absence of other sources of numerosity information (i.e., pattern recognition, which we discuss in the next section).

## Discussion

The boundary between the subitizing range and the postsubitizing range is commonly found to be four objects (Atkinson, Campbell, \& Francis, 1976). Trick and Pylyshyn (1994) proposed that the four-object subitizing range (and the rapid enumeration within this range) emerges out of the limited capacity of "pre-attentive" individuation mechanisms. This "pre-attentive" characterization of subitizing stands in contrast to the serial, attention-bound enumeration mechanisms of explicit counting. However, in light of studies showing attentional effects within the subitizing range (e.g., Railo et al., 2008), the pre-attentive characterization must be reevaluated.

The results from Railo and colleagues (2008) support two key points about the role of attention in enumeration. First, subitizing requires serial focus on individual objects. Other-

[^287]

Figure 4: Mean model accuracy in different attentional conditions.
wise, there would be no significant performance differences between the divided and full attention conditions. In other words, subitizing (at least for arbitrary, irregular patterns) is unusually rapid, serial counting. The results also support the notion that even enumeration processes like the ANS have an attentional dependence, as subject performance in the divided and full attention conditions declined more gradually than in the inattention condition. In the inattention case, subjects attended incidentally to one or two dots, but in the absence of an intention to enumerate, they likely did not engage in approximation. As such, performance did not gradually decline. Attention, therefore, is a necessary feature of any unified account of numerosity perception.

The contrast between parallel and serial processes in enumeration may better be characterized by a distinction between a weak and strong sense of attentional involvement (rather than "pre-attentive" vs. "attentive"). Enumeration, in general, requires an intention to report on the absolute or relative quantity of objects in a visual scene. Therefore, there is at least the need to attend to the results of a parallel mechanism of numerosity judgment (the weak sense). In contrast, precise enumeration via counting requires attentional focus on each individual object to be enumerated (the strong sense). ${ }^{3}$

Other Subitizing Processes Regular and common patterns of objects (e.g., such as patterns found on dice) enable rapid and accurate numerosity beyond four objects (Mandler \& Shebo, 1982), suggesting that pattern recognition may play a role in subitizing performance for certain spatial arrangements of objects. We view the reported model of subitizing as a complementary rather than competing account to pattern-recognition based ones (Peterson \& Simon, 2000). A pattern-recognition component could be subsumed into the ARCADIA model as an alternate (and attentionally prioritized) number sensor that responds to focus on groups of proto-objects. This addition would enable us to make and model the following prediction: regular patterns such as those

[^288]presented by Mandler and Shebo (1982) would show minimal performance differences between the divided-attention and full-attention conditions in the dual task from Railo and colleagues (as compared to the irregular patterns used in the original study and in this paper).

As such, we view the perception of numerosity as a potentially four-part phenomenon. Estimation provides a rapid, parallel, but imprecise source of numerosity information, whereas subvocal counting provides a serial, slow, but accurate enumeration procedure. Subitized counting and pattern recognition provide both a serial and parallel mechanism, respectively, to achieve rapid and accurate enumeration.

To summarize, we have presented a novel computational model of numerosity perception in which attention unifies processes of subitizing, subvocal counting, and estimation. Attention is the glue that enables and binds these separate numerosity faculties together. The limits of attention-bound processes such as object-binding and subvocalization determine how quickly subjects can report numerosity judgments. Likewise, serial attentional focus to individual objects and the need for explicit attention to estimation is necessary to account for the accuracy of enumeration in dual-task settings.

## Acknowledgments

This work was supported by a National Research Council Research Associateship awards to GB, and by grants N0001417WX00574 and N0001417WX00153 by the Office of Naval Research. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

## References

Atkinson, J., Campbell, F. W., \& Francis, M. R. (1976). The magic number $4 \pm 0$ : A new look at visual numerosity judgements. Perception, 5, 327-334.
Bridewell, W., \& Bello, P. F. (2016a). Inattentional blindness in a coupled perceptual-cognitive system. In Proceedings of the thirty-eighth annual conference of the cognitive science society (pp. 2573-2578). Philadelphia, USA.
Bridewell, W., \& Bello, P. F. (2016b). A theory of attention for cognitive systems. In Proceedings of the fourth annual conference on advances in cognitive systems (pp. 1-16). Evanston, USA.
Chesney, D. L., \& Haladjian, H. H. (2011). Evidence for a shared mechanism used in multiple-object tracking and subitizing. Attention, Perception, \& Psychophysics, 73, 2457-2480.
Cutini, S., \& Bonato, M. (2012). Subitizing and visual shortterm memory in human and non-human species: A common shared system? Frontiers in Psychology, 3, 129-133.
Dehaene, S. (2011). The number sense: How the mind creates mathematics. New York, NY: Oxford University Press.
Egeth, H. E., Leonard, C. J., \& Palomares, M. (2008). The role of attention in subitizing: Is the magical number 1 ? Visual Cognition, 16, 463-473.

Feigenson, L., Dehaene, S., \& Spelke, E. (2004). Core systems of number. Trends in Cognitive Sciences, 8, 307-314.
Gallistel, C. R., \& Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. Trends in Cognitive Sciences, 4, 59-65.
Gelman, R., \& Butterworth, B. (2005). Number and language: How are they related? Trends in Cognitive Sciences, 9, 6-10.
Halberda, J., \& Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. Developmental Psychology, 44, 1457.
Huss, D., \& Byrne, M. (2003). An ACT-R/PM model of the articulatory loop. In Proceedings of the fifth international conference on cognitive modeling (pp. 135-140). Boston, USA.
Jevons, W. S. (1871). The power of numerical discrimination. Nature, 3, 281-282.
Kaufman, E. L., Lord, M. W., Reese, T. W., \& Volkmann, J. (1949). The discrimination of visual number. The American Journal of Psychology, 62, 498-525.
Logie, R. H., \& Baddeley, A. D. (1987). Cognitive processes in counting. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 310-326.
Mandler, G., \& Shebo, B. J. (1982). Subitizing: An analysis of its component processes. Journal of Experimental Psychology: General, 111, 1-22.
Olivers, C. N., \& Watson, D. G. (2008). Subitizing requires attention. Visual Cognition, 16, 439-462.
Peterson, S. A., \& Simon, T. J. (2000). Computational evidence for the subitizing phenomenon as an emergent property of the human cognitive architecture. Cognitive Science, 24, 93-122.
Railo, H., Koivisto, M., Revonsuo, A., \& Hannula, M. M. (2008). The role of attention in subitizing. Cognition, 107, 82-104.
Rensink, R. A., O’Regan, J. K., \& Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. Psychological Science, 8, 368-373.
Rock, I., Linnett, C. M., Grant, P., \& Mack, A. (1992). Perception without attention: Results of a new method. Cognitive Psychology, 24, 502-534.
Trick, L. M., Enns, J. T., \& Brodeur, D. A. (1996). Life span changes in visual enumeration: The number discrimination task. Developmental Psychology, 32, 925-932.
Trick, L. M., \& Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. Psychological Review, 101, 80-102.
Vetter, P., Butterworth, B., \& Bahrami, B. (2008). Modulating attentional load affects numerosity estimation: Evidence against a pre-attentive subitizing mechanism. PLoS One, 3, 1-6.

# Effects of transmission perturbation in the cultural evolution of language 

Thomas Brochhagen (t.s.brochhagen@uva.nl)<br>Institute for Logic, Language \& Computation, University of Amsterdam

Michael Franke (mchfranke@gmail.com)<br>Department of Linguistics, University of Tübingen


#### Abstract

Two factors seem to play a major role in the cultural evolution of language. On the one hand, there is functional pressure towards efficient transfer of information. On the other hand, languages have to be learned repeatedly and will therefore show traces of systematic stochastic disturbances of the transmission of linguistic knowledge. While a lot of attention has been paid to the effects of cognitive learning biases on the transmission of language, there is reason to expect that the class of possibly relevant transmission perturbations is much larger. This paper therefore explores some potential effects of transmission noise due to errors in the observation of states of the world. We look at three case studies on (i) vagueness, (ii) meaning deflation, and (iii) underspecified lexical meaning. These case studies suggest that transmission perturbations other than learning biases might help explain attested patterns in the cultural evolution of language and that perturbations due to perceptual noise may even produce effects very similar to learning biases.


Keywords: cognitive biases; iterated learning; language evolution

## Introduction

Language is shaped by its use and transmission across generations. Linguistic properties are therefore not necessarily solely due to functional pressure, such as the selection of more communicatively efficient behavior. They may also be effected by a pressure for learnability. In the extreme, an unlearnable language will not make it to the next generation. The effects that (iterated) learning has on language are often seen as stemming from a combination of general learning mechanisms and inductive cognitive biases (e.g. Griffiths \& Kalish 2007, Kirby et al. 2014, Tamariz \& Kirby 2016). Proposals of biases that shape language acquisition abound, e.g.; mutual exclusivity (Merriman \& Bowman 1989, Clark 2009), simplicity (Kirby et al. 2015), regularization (Hudson Kam \& Newport 2005), and generalization (Smith 2011). But forces other than learning biases may also systematically perturb the transmission of linguistic knowledge and thereby contribute to the shaping of language by cultural evolution (cf. Perfors \& Navarro 2014). In the following we focus on one particular source of transmission noise: agents' imperfect perception of the world. Our overall goal is to give a formalism with which to study the possible effects of such perturbations and to apply it to three case studies on (i) vagueness, (ii) meaning deflation, and (iii) underspecified lexical meaning.

## Iterated Bayesian learning

We model the transmission of linguistic knowledge as a process of iterated learning (Kirby et al. 2014, Tamariz \& Kirby 2016). More specifically, we focus on iterated Bayesian
learning, in which a language learner must infer unobservables, such as the lexical meaning of a word, from the observable behavior of a single teacher, who is a proficient language user (e.g. Griffiths \& Kalish 2007, Kirby et al. 2007). Concretely, the learner observes instances $\langle s, m\rangle$ of overt language use in context, where $s$ is a world state and $m$ is the message that the teacher used in state $s$. The learner's task is to infer which latent type $\tau$ (e.g., which set of lexical meanings or which grammar) may have produced a sequence of such observations. To do so, the learner considers the posterior probability of $\tau$ given a data sequence $d$ of $\langle s, m\rangle$ pairs:

$$
P(\tau \mid d) \propto P(\tau) P(d \mid \tau)
$$

where $P(\tau)$ is the learner's prior for type $\tau$ and $P(d \mid \tau)=$ $\prod_{\langle s, m\rangle \in d} P(m \mid s, \tau)$ is the likelihood of type $\tau$ producing the observed data $d$, with $P(m \mid s, \tau)$ the probability that a type $\tau$ produces message $m$ when in world state $s$. It is usually assumed that learners exposed to $d$ adopt type $\tau$ with probability $F(\tau \mid d) \propto P(\tau \mid d)^{\gamma}$, where $\gamma \geq 1$ regulates whether learners probability match $(\gamma=1)$ or tend towards choosing a maximum of the posterior distribution $(\gamma>1)$. If the set $D_{k}$ of data a learner may be exposed to is the set of all sequences with $k$ pairs $\langle s, m\rangle$, the probability that a learner acquires type $\tau_{i}$ when learning from a teacher of type $\tau_{j}$ is:

$$
P\left(\tau_{j} \rightarrow \tau_{i}\right) \propto \sum_{d \in D_{k}} P\left(d \mid \tau_{j}\right) F\left(\tau_{i} \mid d\right)
$$

If a population is a distribution over types, then iterated Bayesian learning predicts the most likely path of change in the population due to learning from finite observations.

The prior $P(\tau)$ can be understood as encoding learning biases. For example, learners may have an a priori preference for simpler languages over ones with a more complex grammar, or over ones with larger or more marked lexical or phonemic inventories (cf. Kirby et al. 2015). Crucially, even weak biases can magnify and have striking effects on an evolving linguistic system, especially if learning is fueled by only limited input (small $k$ ). Experimental and mathematical explorations of iterated learning have consequently suggested that the linguistic structure evinced by the outcome of this process reflects learners' inductive biases (Kirby et al. 2007; 2014).

## Iterated Bayesian learning with state-noise

Other stochastic factors beyond learning biases in $P(\tau)$ can influence the adoption of a linguistic type $\tau$ based on the observation of $\langle s, m\rangle$ sequences. One further potential source


Figure 1: State-noise during observation of language use.
of "transmission noise" are regular stochastic errors in the perception of world states (see Figure 1). Imperfect perception may lead teachers to produce utterances that deviate from their production behavior had they witnessed the state correctly. Similarly, learners may mistake utterances as applying to different states than the ones witnessed by the teacher who produced them. For instance, when learning the meaning of a vague adjective such as tall from utterances like "Jean is tall", agents may have diverging representations of how tall Jean actually is, even if she is in a shared perceptual environment. The main idea to be explored here is that regularities in misperceptions of states may have striking and possibly explanatory effects on language evolution.

We denote the probability that the teacher (learner) observes state $s_{t}\left(s_{l}\right)$ when the actual state is $s_{a}$ as $P_{N}\left(s_{t} \mid s_{a}\right)$ $\left(P_{N}\left(s_{l} \mid s_{a}\right)\right)$. The probability that $s_{a}$ is the actual state when the learner observes $s_{l}$ is therefore:

$$
P_{N}\left(s_{a} \mid s_{l}\right) \propto P\left(s_{a}\right) P_{N}\left(s_{l} \mid s_{a}\right) .
$$

Assuming a finite state space for convenience, the probability that the teacher observes $s_{t}$ when the learner observes $s_{l}$ is:

$$
P_{N}\left(s_{t} \mid s_{l}\right)=\sum_{s_{a}} P_{N}\left(s_{a} \mid s_{l}\right) P_{N}\left(s_{t} \mid s_{a}\right) .
$$

The probability that a teacher of type $\tau$ produces data that is perceived by the learner as a sequence $d_{l}$ of $\left\langle s_{l}, m\right\rangle$ pairs is:

$$
P_{N}\left(d_{l} \mid \tau\right)=\prod_{\left\langle s_{l}, m\right\rangle \in d_{l}} \sum_{s_{t}} P_{N}\left(s_{t} \mid s_{l}\right) P\left(m \mid s_{t}, \tau\right) .
$$

It is natural to assume that learners, even if they (in tendency) perform rational Bayesian inference of the likely teacher type $\tau$ based on observation $\left\langle s_{l}, m\right\rangle$, do not also reason about statenoise perturbations. In contrast to, e.g., noisy-channel models that have agents reason over potential message corruption caused by noise (e.g. Bergen \& Goodman 2015), our learners are not proficient language users that could leverage knowledge about the world and its linguistic codification to infer
likely state misperception. ${ }^{1}$ In this case the posterior probability of $\tau$ given the learner's perceived data sequence $d_{l}$ is as before: $P\left(\tau \mid d_{l}\right) \propto P(\tau) P\left(d_{l} \mid \tau\right)$. Still, state-noise affects the probability $P_{N}\left(\tau_{j} \rightarrow \tau_{i}\right)$ that the learner adopts $\tau_{i}$ given a teacher of type $\tau_{j}$, because it influences the probability of observing a sequence $d_{l}$ (with $F\left(\tau_{i} \mid d\right)$ as before):

$$
P_{N}\left(\tau_{j} \rightarrow \tau_{i}\right) \propto \sum_{d \in D_{k}} P_{N}\left(d_{l} \mid \tau_{j}\right) F\left(\tau_{i} \mid d\right)
$$

In sum, it may be that learner and/or teacher do not perceive the actual state as what it is. If they are not aware of this, they produce/learn as if what they observed was the actual state. In particular, the learner does not reason about noise when she tries to infer the teacher's type. She takes what she observes as the actual state that the teacher has seen as well, and infers which linguistic type (e.g. which set of lexical meanings or grammar) would have most likely generated the message to this state. This can lead to biases of inferring the "wrong" teacher type if noise makes some types err in a way that resembles the noiseless behavior of other types. That is, such environmental factors can, in principle, induce transmission perturbations that look as if there was a cognitive bias in favor of a particular type, simply because that type better explains the noise.

## Case studies

In what follows we present three case studies that show how iterated learning under noisy perception can lead to the emergence of linguistic phenomena. The studies are ordered from more to less obvious examples in which state-noise may be influential and explanatory: (i) vagueness, (ii) meaning deflation, and (iii) underspecification in the lexicon. No case study is meant to suggest that state-noise is the definite and only explanation of the phenomenon in question. Instead, our aim is to elucidate the role that transmission perturbations beyond inductive biases may play in shaping the cultural evolution of language. We therefore present minimal settings that isolate potential effects of state-noise in iterated learning.

## Vagueness

Many natural language expressions are notoriously vague and pose a challenge to logical analysis of meaning (e.g. Williamson 1994). Vagueness also challenges models of language evolution since functional pressure towards maximal information transfer should, under fairly general conditions, weed out vagueness (Lipman 2009). Many have therefore argued that vagueness is intrinsically useful for communication (e.g. van Deemter 2009, de Jaegher \& van Rooij 2011, Blume \& Board 2014). Others hold that vagueness arises naturally due to limits of perception, memory, or information processing (e.g. Franke et al. 2011, O’Connor 2014, Lassiter

[^289]\& Goodman 2015). We follow the latter line of exploration here, showing that vagueness can naturally arise under imperfect observability of states (see Franke \& Correia (to appear) for a different evolutionary dynamic based on the same idea).

Setup. We analyze the effects of noisy perception on the transmission of a simple language with 100 states, $s \in[0,99]$, and two messages, $m \in\left\{m_{1}, m_{2}\right\}$. The probability that agents perceive actual state $s_{a}$ as $s_{t} / s_{l}$ is given by a (discretized) normal distribution, truncated to [0;99], with $s_{a}$ as mean and standard deviation $\sigma$. Linguistic behavior is fixed by a type $\tau \in[0 ; 99]$ which is the threshold of applicability of $m_{1}$ : $P\left(m_{1} \mid s, \tau\right)=\delta_{s \geq \tau}=\left(1-P\left(m_{2} \mid s, \tau\right)\right)$. In words, if a speaker observes a state that is as large or larger than its type, then message $m_{1}$ is used (tall), otherwise $m_{2}$ (small).

Results. The effects of a single generational turnover under noisy transmission of a population that initially consisted exclusively of type $\tau=50$ is depicted in Figure 2. As learners try to infer this type from observed language use, even small $\sigma$ will lead to the emergence of vagueness in the sense that there is no longer a crisp and determinate cut-off point for message use in the population. Instead, borderline regions in which $m_{1}$ and $m_{2}$ are used almost interchangeably emerge. For larger $\sigma$, larger borderline regions ensue. The size of such regions further increases over generations with growth inversely related to $\gamma$ and $k$. As is to be expected, if $k$ is too small to discern even strikingly different types, then iterated learning under noisy perception leads to heterogeneous populations with (almost) no state being (almost) exclusively associated with $m_{1}$ or $m_{2}$.


Figure 2: Noisy iterated learning ( $\gamma=1, \sigma=0.4, k=20$ ).

Discussion. Transmission perturbations caused by noisy state perception reliably give rise to vague language use even if the initial population had a perfectly crisp and uniform convention. Clearly, this is a specific picture of vagueness. As
modeled here for simplicity, each speaker has a fixed and nonvague cut-off point $\tau$ in her lexicon. Still, the production behavior of a type- $\tau$ speaker in actual state $s_{a}$ is probabilistic and "vague", because of noisy perception:

$$
P_{N}\left(m \mid s_{a}, \tau\right)=\sum_{s_{t}} P\left(s_{t} \mid s_{a}\right) P\left(m \mid s_{t}, \tau\right) .
$$

An extension towards types as distributions over thresholds is straightforward but the main point would remain: systematic state-noise perturbs a population towards vagueness.

Of course, convergence on any particular population state will also depend on the functional (dis)advantages of particular patterns of language use. Functional pressure may therefore well be necessary for borderline regions to be kept in check, so to speak. Which factor or combination thereof plays a more central role for the emergence of vagueness is an empirical question we do not address here. Instead, we see these results as adding strength to the argument that one way in which vagueness may arise is as a byproduct of interactions between agents that may occasionally err in their perception of the environment. If state perception is systematically noisy and learners are not aware of this, some amount of vagueness may be the natural result.

## Deflation

Meaning deflation is a diachronic process by which a form's once restricted range of applicability broadens. Perhaps the most prominent example is Jespersen's cycle (Dahl 1979), the process by which emphatic negation, such as French ne ... pas, broadens over time and becomes a marker for standard negation. As argued by Bolinger (1981), certain word classes are particularly prone to slight and unnoticed reinterpretation. When retrieving their meaning from contextual cues, learners may consequently continuously spread their meaning out. For instance, Bolinger discusses how the indefinite quantifier several has progressively shifted from meaning a respectable number to broader a few in American English. We follow this line of reasoning and show how state confusability may lead to meaning deflation. Other formal models of deflationary processes in language change have rather stressed the role of conflicting interests between interlocutors (Ahern \& Clark 2014) or asymmetries in production frequencies during learning (Schaden 2012, Deo 2015).

Setup. The setup is the same as that of the previous case study, except that we now trace the change of a single message $m$, e.g., emphatic negation, without a fixed antonym being sent whenever $m$ does not apply. This is a crude way of modeling use of markers of emphasis or high relevance for which no corresponding "irrelevance marker" exists. Learners accordingly observe positive examples of use $\langle s, m\rangle$ but do not positively observe situations in which $m$ did not apply to a particular state. This causes asymmetry in the learning data because some types will reserve their message only for a small subset of the state space and otherwise remain silent.

Learners take the absence of observations into account but cannot know what it is that they did not observe. We assume that learners are aware of $k$ so that: ${ }^{2}$

$$
\begin{array}{r}
P\left(\tau \mid d_{l}\right) \propto \operatorname{Binom}\left(\text { successes }=k-\left|d_{l}\right|, \text { trials }=k,\right. \\
\text { succ.prob } \left.=\sum_{i=0}^{\tau-1} P(s=i)\right) \prod_{s \in d_{l}} P(m \mid s, \tau) .
\end{array}
$$

As before, the second factor corresponds to the likelihood of a type producing the perceived data. The first is the probability of a type not reporting $k-|d|$ events for a total of $k$ events. $P \in \Delta(S)$ is assumed to be uniform. In words, a long sequence of data consisting of mostly silence gives stronger evidence for the type producing it having a high threshold of applicability even if the few state-message pairs observed may be equally likely to be produced by types with lower thresholds.

Results. The development of an initially monomorphic population consisting only of $\tau=80$ is shown in Figure 3. Even little noise causes a message to gradually be applied to larger portions of the state space. The speed of meaning deflation is regulated by $\sigma, k$, and to lesser degree $\gamma$. In general, more state confusion due to higher $\sigma$, shorter sequences, or less posterior maximization will lead to more learners inferring lower types than present in the previous generation.


Figure 3: Noisy iterated learning $(\gamma=1, \sigma=0.4, k=30)$.

Discussion. In contrast to the previous case study, we now considered the effects of noisy perception under asymmetric data generation where overt linguistic evidence is not always produced, i.e., acquisition in a world in which not every state is equally likely to lead to an observable utterance. The outcome is nevertheless similar to the previous one: Noisy perception can cause transmission perturbations that gradually

[^290]relax formerly strict linguistic conventions. In contrast to the case of vagueness, if there are no relevant competing forms, e.g., small vs. tall, asymmetry in production and noise will iteratively increase the state space that a form carves out.

## Scalar expressions

Scalar expressions have been at the center of many studies on pragmatic inference. Examples include quantifiers such as some and most, adjectives such as cold and big, and numerals such as four and ten. Commonly, their use is taken to pragmatically convey an upper-bound which is not present in their lexical semantics (Horn 1972, Gazdar 1979). For instance, while "Bo ate some of the cookies" is semantically compatible with a state in which Bo ate all of them, this utterance is often taken to convey that Bo ate some but not all, as otherwise the speaker would have said all. A semantically weak meaning is thus pragmatically strengthened by interlocutors’ mutual reasoning about rational language use (Grice 1975).

Why does such pragmatic strengthening not lead to widespread lexicalization of upper-bounded meanings? To address this question, Brochhagen et al. (2016) explore an evolutionary model that combines functional pressure and iterated learning. This account assumes a prior that favors a lack of upper-bounds. Here, we demonstrate that state-noise can mimic the effects of such a cognitive learning bias.

Setup. The simplest possible model distinguishes two kinds of lexica and two behavioral strategies to use them, a pair of which constitutes a type. Both lexica specify the truthconditions of two messages in either of two states. Let us mnemonically label them $m_{\text {some }}, m_{\text {all }}, s_{\exists \neg \forall}$ and $s \forall$, where the former state is one in which natural language some but not all holds, and the latter one where all holds. In lexicon $L_{\text {bound }}$, which lexicalizes an upper-bound for some-like expressions, message $m_{\text {some }}$ is only true of $s_{\exists \neg \forall}$ and $m_{\text {all }}$ only of $s_{\forall}$. In the English-like lexicon $L_{\text {lack }}$, message $m_{\text {all }}$ is also only true of $s_{\forall}$, but the meaning of $m_{\text {some }}$ is underspecified and lexically holds in both states. Speakers follow one of two strategies of language use: literal or pragmatic. The former select a random true message, whereas the latter prefer to send the most informative messages from those that are true in the observed state (Grice 1975). This gives rise to probabilistic speaker behavior $P(m \mid s, \tau=\langle$ lexicon, use $\rangle)$ which approximates the following choice probabilities: ${ }^{3}$

| Literal | $\underline{L_{\text {bound }}}$ |  |  | $\underline{L_{\text {lack }}}$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $m_{\text {some }}$ | $m_{\text {all }}$ |  | $m_{\text {some }}$ | $m_{\text {all }}$ |
|  | $\begin{aligned} & s_{\forall} \\ & s_{\exists \neg \forall} \end{aligned}$ | $\left(\begin{array}{l}0 \\ 1\end{array}\right.$ | $\left.\begin{array}{l}1 \\ 0\end{array}\right)$ | $s_{\forall}$ $s_{\exists \rightarrow \forall}$ | $\left(\begin{array}{c}0.5 \\ 1\end{array}\right.$ | $\left.\begin{array}{c}0.5 \\ 0\end{array}\right)$ |
|  |  | $m_{\text {some }}$ | $m_{\text {all }}$ |  | $m_{\text {some }}$ | $m_{\text {all }}$ |
| Pragmatic | $\begin{aligned} & s_{\forall} \\ & s_{\exists \exists \forall} \end{aligned}$ | $\left(\begin{array}{l}0 \\ 1\end{array}\right.$ | $\left.\begin{array}{l}1 \\ 0\end{array}\right)$ | $s_{\forall}$ $s_{\exists \neg \forall}$ | $\left(\begin{array}{l}0 \\ 1\end{array}\right.$ | $\left.\begin{array}{l}1 \\ 0\end{array}\right)$, |

[^291]where $P(m \mid s, \tau)=M_{[s, m]}$ with $M$ being type $\tau$ 's choice matrix. As pragmatic users of $L_{\text {lack }}$ are (almost) indistinguishable from types with $L_{\text {bound }}$, the emergence of a predominance of $L_{\text {lack }}$ in a repeatedly learning population must come from transmission biases. A learning bias in favor of $L_{\text {lack }}$ in the learners' priors will select for it (Brochhagen et al. 2016), but here we assume no such cognitive bias. Rather we assume state-noise in the form of parameters $\varepsilon$ and $\delta$. The former is the probability of perceiving actual state $s_{\exists \neg \forall}$ as $s_{\forall}$, $P\left(s_{\forall} \mid s_{\exists \rightarrow \forall}\right)=\varepsilon$, and $P\left(s_{\exists \rightarrow \forall} \mid s_{\forall}\right)=\delta$. For instance, states may be perceived differently because different numbers of objects must be perceived (e.g., quantifiers and numerals) or they may be more or less hard to accurately retrieve from sensory information (e.g., adjectives).

Results. To quantify the effects of the dynamics we ran a fine-grained parameter sweep over $\varepsilon$ and $\delta$ with 50 independent simulations per parameter configuration. Each simulation started with a random initial population distribution over types and applied iterated learning with state-noise for 20 generations, after which no noteworthy change was registered. Mean proportions of resulting pragmatic users of $L_{\text {lack }}$ under different noise signatures are shown in Figure 4. These results suggest that when $\delta$ is small and $\varepsilon$ high, iterated noisy transmission can lead to populations consisting of almost exclusively English-like lexica with pragmatic language use. Similar results are obtained for larger $k$ or $\gamma$.


Figure 4: Mean proportion of pragmatic $L_{\text {lack }}$ users after 20 generations $(\gamma=1, k=5)$.

Discussion. The main goal of this case study was to show that noisy perception may mimic effects of learning biases. In the case of Brochhagen et al. the assumed bias was one for simplicity; learners had an a priori preference for not codifying upper-bounds lexically, which increased their propensity to infer pragmatic $L_{\text {lack }}$ over $L_{\text {bound }}$ even if the witnessed data
could not tease them apart. We assumed no such bias but nevertheless arrived at evolutionary outcomes that comparable to those predicted if the bias were present. However, this result strongly depends on the types involved. Whether a type thrives under a particular noise signature depends on the proportion of types confused with it during transmission. The addition or extraction of a single type may therefore lead to different results.

At present, it is unclear what role noisy perception should play in the selection of underspecified meaning. These results should therefore be taken as suggestive but not indicative of a relationship between the two. In the case of quantifiers, a possible way to explore this relation may lie in their connection to empirical work on the verification of quantified statements (see Szymanik 2016 for a recent overview). The idea being that some states are easier to verify, e.g., $s_{\forall}$, and therefore less confusable with other states than others, e.g., $s_{\exists \neg \forall}$.

## General discussion

We proposed a general model of iterated Bayesian learning that integrates systematic noise in agents' perception of world states, giving rise to stochastic perturbations that may influence and (potentially, partially) explain language change. We investigated the model's predictions in three case studies that show that iterated noisy transmission can lead to outcomes akin to those found in natural language. As stressed before, these results are not meant to suggest noisy perception to be the sole or main determinant of these phenomena. Instead, our aim was mainly conceptual and technical in nature.

Beyond technical aspects, we foregrounded two intertwined issues in the cultural evolution of language. First, the fact that noise signatures may mimic the effects of cognitive biases has consequences for the interpretation of outcomes of acquisition processes. Care must therefore be exercised in reading off the influence of possible learning biases from data obtained "in the wild" or the laboratory. Noisy perception instead offers a neutral model of cultural evolution that appeals to neither functional competition nor differential learnability among types (Reali \& Griffiths 2009). Second, and more importantly, these results can be seen as complementing and stressing the pivotal role of systematic transmission perturbations as explanatory and predictive devices of language change - independent of the perturbation's source. They thereby strengthen and widen the scope of research on iterated learning by bringing attention to forces beyond inductive biases (cf. Perfors \& Navarro 2014).

## Conclusion

Acquisition is a central force shaping linguistic structure. The consideration of the (imperfect) means by which such knowledge is transmitted is therefore crucial to our understanding of the cultural evolution of language. Here, we focused on one factor that may give rise to systematic stochastic perturbation in learning -agents' noisy perception of the worldand analyzed its effects in three case studies on (i) vagueness,
(ii) meaning deflation, and (iii) underspecified lexical meaning. Our results suggest that the class of relevant perturbation sources reaches beyond the well-studied effects of inductive learning biases. In particular, that some linguistic properties, such as (i), (ii) and more tentatively (iii), may emerge as a byproduct of constraints on agents' perception of the world.

## Acknowledgments

TB acknowledges funding from the European Community's Seventh Framework Programme under grant agreement no. 567652 /ESSENCE/. MF gratefully acknowledges financial support by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63).

## References

Ahern, C., \& Clark, R. (2014). Diachronic processes in language as signaling under conflicting interests. In E. A. Cartmill, S. Roberts, H. Lyn, \& H. Cornish (Eds.), Proceedings of EVOLANG 11 (pp. 25-32). Singapore: World Scientific Press.

Bergen, L., \& Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. Topics in Cognitive Science, 7(2), 336-350.
Blume, A., \& Board, O. (2014). Intentional vagueness. Erkenntnis, 79(4), 855-899.
Bolinger, D. (1981). The deflation of several. Journal of English Linguistics, 15(1), 1-3.
Brochhagen, T., Franke, M., \& van Rooij, R. (2016). Learning biases may prevent lexicalization of pragmatic inferences: a case study combining iterated (bayesian) learning and functional selection. In Proceedings of the 38th annual conference of the cognitive science society (pp. 20812086). Austin, TX: Cognitive Science Society.

Clark, E. V. (2009). Lexical meaning. In E. L. Bavin (Ed.), The cambridge handbook of child language (pp. 283-300). Cambridge University Press.
Dahl, Ö. (1979). Typology of sentence negation. Linguistics, 17(1-2).

Deo, A. (2015). The semantic and pragmatic underpinnings of grammaticalization paths: The progressive to imperfective shift. Semantics \& Pragmatics, 8(1), 1-52.

Franke, M., \& Correia, J. P. (to appear). Vagueness and imprecise imitation in signalling games. British Journal for the Philosophy of Science.
Franke, M., Jäger, G., \& van Rooij, R. (2011). Vagueness, signaling \& bounded rationality. In T. Onoda, D. Bekki, \& E. McCready (Eds.), JSAI-isAI 2010 (pp. 4559). Springer.

Gazdar, G. (1979). Pragmatics, implicature, presuposition and logical form. New York: Academic Press.
Grice, P. (1975). Logic and conversation. In Studies in the ways of words (pp. 22-40). Cambridge, MA: Harvard University Press.

Griffiths, T. L., \& Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. Cognitive Science, 31(3), 441-480.
Horn, L. R. (1972). On the semantic properties of logical operators in english. Bloomington, IN: Indiana University Linguistics Club.
Hudson Kam, C. L., \& Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language Learning and Development, 1(2), 151-195.
de Jaegher, K., \& van Rooij, R. (2011). Strategic vagueness, and appropriate contexts. In Language, games, and evolution (pp. 40-59). Berlin, Heidelberg: Springer.
Kirby, S., Dowman, M., \& Griffiths, T. L. (2007). Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences, 104(12), 5241-5245.
Kirby, S., Griffiths, T., \& Smith, K. (2014). Iterated learning and the evolution of language. Current Opinion in Neurobiology, 28, 108-114.
Kirby, S., Tamariz, M., Cornish, H., \& Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. Cognition, 141, 87-102.
Lassiter, D., \& Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. Synthese.
Lipman, B. L. (2009). Why is language vague? (Manuscript, Boston University)
Merriman, W. E., \& Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. Monographs of the Society for Research in Child Development, 54(3/4).
O'Connor, C. (2014). The evolution of vagueness. Erkenntnis, 79(4), 707-727.
Perfors, A., \& Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. Cognitive Science, 38(4), 775-793.
Reali, F., \& Griffiths, T. L. (2009). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. Proceedings of the Royal Society B: Biological Sciences, 277(1680), 429-436.
Schaden, G. (2012). Modelling the "aoristic drift of the present perfect" as inflation: An essay in historical pragmatics. International Review of Pragmatics, 4, 261-292.
Smith, K. (2011). Learning bias, cultural evolution of language, and the biological evolution of the language faculty. Human Biology, 83(2), 261-278.
Szymanik, J. (2016). Quantifiers and cognition: Logical and computational perspectives. Springer International Publishing.
Tamariz, M., \& Kirby, S. (2016). The cultural evolution of language. Current Opinion in Psychology, 8, 37-43.
van Deemter, K. (2009). Utility and language generation: The case of vagueness. Journal of Philosophical Logic, 38(6), 607-632.
Williamson, T. (1994). Vagueness. London and New York: Routledge.

# Does banana spontaneously activate yellow color? Color-related concepts help with color discrimination 

Lionel Brunel (lionel.brunel@univ-montp3.fr) \&Philippe Servajean (philippe.servajean@univ-montp3.fr)<br>Department of Psychology, Epsylon Laboratory, University Paul Valéry<br>Montpellier, France<br>Loic Heurley (heurleyloic@yahoo.fr)<br>Department STAPS, CeRSM Laboratory, University Paris X, Nanterre, France

Nicolas Vermeulen (nicolas.vermeulen@uclouvain.be)
Research Institute for Psychological Sciences, Université catholique de Louvain, Louvain-la-Neuve, Belgium


#### Abstract

Color is a critical part of objects representation as well as critical cue for recognizing objects. However, it is less clear how people represent color in memory. The present study aimed at investigating this issue. We designed a procedure based on short-term sensory memory load procedure mixed with a color-priming paradigm. Participants learned three visual stimuli (either non-words - lexical load condition - or visual-shapes - visual-shape load condition). Then, they performed a color discrimination task on colored patch (e.g., a yellow patch). Each target was preceded by a color-related concept word either congruent (e.g., word "banana") or not (e.g., word "lettuce"). Finally, they performed a recognition task either on non-words or on visual-shapes depending on the memory load condition). We showed that color-priming effect was selectively disrupted in visual-shape load condition. We interpreted this finding as an evidence that automatic modal simulations occur during access to the meaning of color-related concept.


Keywords: Color knowledge, perceptual simulation, priming effect, visual memory

## Introduction

Color knowledge about object is an important part of conceptual representation. Indeed, color is often a critical cue for recognizing object (Tanaka \& Presnell, 1999) or natural scene (Oliva \& Schyns, 2000) in everyday life. Hansen, Olkkonen, Walter, and Gegenfurtner (2006) suggested that research about color should investigate how people represent color rather than how people perceive color because color of an environment is never stable over the time (i.e., dependent of the objects' illumination) and, as a consequence, dependent from color knowledge. The question is how do people represent color knowledge?

First, color knowledge could be considered as stored in an amodal format within conceptual representation of an object. In that view, meaning of an object is distributed across semantic features (see Masson, 1995) and relationships between concepts are explained in term of
overlapping between semantic features. As soon as two concepts share the same semantic color feature, one can prime the other. For instance, hearing the word "lips" could facilitate the detection of the picture of a strawberry (Huettig \& Altmann, 2011). However, several studies suggest that access to color knowledge depend on the nature of the stimuli (i.e., lexical vs. visual, see Nijboer, Zandvoort \& Haan, 2006). Indeed, because the word "banana" and "yellow" co-occurs frequently in everyday language, color knowledge could also be stored in a lexical format: a color label (see Landauer \& Dumais, 1997). In that case, as soon as two concepts share the same color label, one can prime the other. For instance, hearing the word "pea" could facilitate the detection of the picture of a green blouse (see Huettig \& Altamann, 2011) because both objects share the label "green". In the same vein, Roberson and Davidoff (2000) have showed a loss of the categorical perception (i.e., better discrimination across color categories than within the same color category) when participants had to simultaneous discriminate between colors and maintain words in memory. This effect was not observed when they had to follow a curved line with the eyes. Accordingly, it seems that lexical access interferes with categorical perception, while visual interference does not. This result suggests the implication of lexical units during access to color knowledge. Moreover Nijboer and collaborators (2006) demonstrated that color-priming effect (from objectword or object-picture to colored-patch) is dependent of the nature of the prime (either picture or word, see also Heurley, et al., 2013). This result suggests the existence of both a semantic-based and lexical-based representation of color. These two formats of color knowledge are not necessary incompatible and authors suggested a time-based access distinction for accounting existence of both level of representation (Heurley et al., 2013).

Alternatively, theories of embodied or grounded cognition (see Barsalou, 2008) assume that access to a representation is linked with perceptual sensory simulations. In other words, performing a conceptual task like verifying that a
"banana" is "yellow" is associated with the automatic simulation of a former visual experience associated with the banana (Pecher, Zeelenberg, \& Barsalou, 2003, 2004; Van Dantzig, Pecher, Zeelenberg, \& Barsalou, 2008; Vermeulen, Chang, Corneille, Pleyers, \& Mermillod, 2013; Vermeulen, Corneille, \& Niedenthal, 2008). Thus, access to the meaning of concept involves automatic perceptual simulation in several sensory modalities and, as a consequence, influences processing of stimuli presented in the same modality than the simulated one (see Brunel, Goldstone, Vallet, Riou, \& Versace, 2013; Brunel, Labeye, Lesourd, \& Versace, 2009; Brunel, Lesourd, Labeye, \& Versace, 2010; Vallet, Brunel, \& Versace, 2010; for a review see Versace, Labeye, Badard, \& Rose, 2009). In that case, color knowledge could be defined as perceptual or modal rather than semantic or lexical. Indeed, neuroimagery studies showed that either perceiving or conceiving color involves common neural substrates (Simmons et al., 2007). Moreover, Richter and Zwaan (2009) showed that processing color words (e.g., word "red") involves perceptual simulation of the color. They showed that participants were faster at discriminating between targets colored-square displayed in the same category color (e.g., light red vs. dark red) when the prime color word and the targets colored squares match on their color rather than mismatch. According to the authors, their results ruled out an explanation based on a lexical competition since the prime word and the target squares shared the same color label. Finally, Yee, Ahmed and Thompson-Schill (2012) found a contextual-based color priming effect using a semantic priming procedure. Indeed they found that color-priming effect is observed only when color was sensitized before the priming procedure. According to the authors, this result seemed to indicate that color knowledge is context dependent rather than stable over the time that is consistent with an embodied approach but not with a semantic approach (see also Connell, 2007; Connell \& Lynott, 2009 for a similar conclusion)

Given existence of empirical evidence for both approaches (embodied vs. semantic or lexical), this article aims at addressing the issue about the representational format of color knowledge in memory. In the present study, we designed a single paradigm in order to test simultaneously each assumption regarding the nature of color knowledge (i.e., perceptual/modal vs. lexical/amodal). To do so, we adapted the procedure of Vermeulen and coworkers (2008, see also Vermeulen, Chang, Mermillod, Pleyers \& Corneille, in press, Experiment 2). In their experiment, they combined a short-term memory task (i.e., memory load) with a property verification task. Consequently, they manipulated both the nature of the memory load (i.e., visual or auditory) and the nature of the property (i.e., visual or auditory) during the property
verification task. First, participants had to learn items visually or auditory displayed. Then they had to perform a property verification task like verifying that a banana could be yellow. After that, they had to recognize the previously learnt items from a new list displayed in the same modality. The main results of their study is that participants were significantly slower at verifying visual properties preceded by a visual-shape load than an auditory load and conversely for auditory properties. Authors concluded that sensory memory and conceptual memory share the same modal properties. In our procedure, we changed the nature of the items in the short-term memory task and replaced the property verification task by a color priming procedure (Heurley, et al., 2013). As a consequence, participants firstly maintained three stimuli in visual memory either meaningless lexical stimuli (i.e., non-words) or meaningless visual-shapes (i.e., Gaussian blobs) before performing a color discrimination task on colored patches. Each target colored patch (e.g., a yellow patch) was preceded by a word prime that could represent a color-related concept (e.g., the word "banana") or not (e.g., the word "lettuce"). Heurley and collaborators (2013, see also Nijboer et al., 2006) found that participant were faster at discriminating the color of the patch when prime was congruent rather than incongruent, attesting a color-priming effect from color-related concept primes to colored targets patches. Finally, participants completed a recognition task on the previously learnt elements.

Our procedure should let us to directly test different assumptions about the format of the color knowledge. First, if accessing to color knowledge from a color-related concept involves semantic amodal knowledge about color, we should find a significant interaction between Prime Type (yellow or green color related concept) and Color Target (yellow or green patches) irrespective the nature of the memory load. A given word activates its conceptual semantic representation in memory (see, Masson, 1995) which include diagnostic feature about the concept. Since this representation is activated at a semantic level, it should not interact with any of the item of the load conditions because each item for these conditions is meaningless. This result would be a direct replication of Heurley and coworkers (2013) Experiments. Then, if accessing color from a color-related concept involves lexical knowledge (such as verbal label), we should find an interaction between Prime Type and Color Target and the Nature of the Memory Load. In that case, the word prime activates the color representation at a lexical level and should interact specifically with the items of the lexical load condition. As a consequence, we might predict a diminution (or a lost) of color-priming effect consecutive to the lexical load condition while the color-priming effect should be observed


Figure 1: Trial overview for both training and test phase in the visual-shape load condition. Note: The colored patch is represented into grey scale but was displayed either in yellow or green during the experiment.
with the visual-shape load condition. Finally, if accessing color from a color-related concept involves visual simulation, we should observe the opposite pattern. In that case, processing the word automatically simulates former visual experiences associated with the concept and those simulations should interact with the items of the visualshape load condition. As a consequence, we might predict a diminution (or a lost) of color-priming effect consecutive to the visual-shape load condition while the color-priming effect should be observed with the lexical load condition.

## Experiment

Participant - Twenty-Four native French speakers (student from Université Paul-Valery, Montpellier, France) were recruited and received courses credits for their participation. All have a normal or corrected to the normal vision and none of them reported having atypical color perception (Daltonism or synesthesia)

Stimuli \& Material - We created 12 Gaussian blobs (6 for training and 6 for test) that were equilibrated in term of surface, number of angles and of peaks in both sides regarding a vertical symmetric axis. We created also 12 CVC-CVC non-words (6 for training and 6 for test) following Reinitz, Lammers, and Cochran (1992) methodology. These stimuli were used in the short-term memory task, respectively in the visual-shape load and lexical load conditions. For the priming phase, we used the same material than Heurley and co-workers (2013; see also Reilhac \& Jiménez, 2006). The 16 priming words (4 for training and 12 for test) depicted either animal or vegetable typically associated with the color green (e.g., lettuce) or yellow (e.g., banana). The target stimuli were yellow ( $\mathrm{R}=$ 255; $\mathrm{G}=255 ; \mathrm{B}=0)$ or green patches $(\mathrm{R}=34 ; \mathrm{G}=163$; B $=13$ ) according to the RGB color model. We also used a
mask that was a white screen with 17 lines of 60 black stars (i.e., *).

Procedure - After filling out a consent form, participants were tested individually in a computer room. Each trial started with a fixation-cross lasting on the screen during 500 ms followed by three successive visual stimuli (either non-words or visual-shapes depending on the memory load condition), each lasting 500 ms . Participants were informed that they have to learn these stimuli in order to perform a later recognition task. Then a prime word was prompted on the screen $(150 \mathrm{~ms})$ and was immediately replaced by a visual mask ( 100 ms ) itself replaced by a blank screen ( 100 ms ). A target colored patch followed and participants had to judge as quickly and accurately as possible its color. After 1500 ms blank screen, 3 stimuli (non-words or visualshapes) were successively displayed and participants have to judge for each stimuli if it corresponded or not to a previously learnt stimulus. We set the inter-trial interval at 1500 ms (see Figure 1). Participant indicated their responses by pressing different keyboard's keys for the color discrimination task and for the recognition task. The responses keys were counterbalanced between participants.

The experiment started with a training phase ( 16 trials) followed by a test phase composed by 48 trials randomly presented: 24 in the visual-shape-load and 24 in lexical-load condition. Each prime was seen followed by each target patch and for each load condition. For the short-term memory task, we have controlled that the number of "same" and "different" was identical for each position during test compared to the learning and for each conditions: visual-shape-load and lexical-load.

## Results

Table 1. Mean correct RT and correct response rates for each experimental condition. Note: Priming effect was calculated by subtracting incongruent experimental conditions (e.g., "yellow" prime / green target condition) to the congruent ones (e.g., "yellow" prime / yellow target condition). A negative value indicates facilitation (i.e., a gain toward the congruent condition) whereas a positive value indicates a cost.

|  |  | Lexical |  |  |  | Visual-shape |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | "Yellow" Prime |  | "Green" Prime |  | "Yellow" Prime |  | "Green" Prime |  |
|  |  | RT (ms) | CR | RT(ms) | CR | RT(ms) | CR | RT (ms) | CR |
| Target Color | Yellow | 493 (21) | 0.965 (.018) | 531 (26) | 0.971 (.013) | 518 (28) | 0.978 (.012) | 525 (27) | 0.942 (.021) |
|  | Green | 529 (24) | 0.965 (.017) | 490 (28) | 0.976 (.013) | 517 (27) | 0.962 (.013) | 528 (29) | 0.957 (.018) |
|  | Priming effect | -36 |  | -42 |  | 1 |  | 2 |  |

Color Discrimination Task - The mean correct response latencies and mean percentages of correct responses were calculated across subjects for each experimental condition. Latencies below 250 ms and above $1,250 \mathrm{~ms}$ were removed (this cut-off resulted in the exclusion of $2.95 \%$ of the data, see Brunel et al., 2009). The participants performed the test color categorization task accurately (overall correct response rate of $96.45 \%$, see Table 1). Practice trials were removed from the analysis.

A repeated analysis of variance was performed with subjects as random variable, Nature of Memory Load (Lexical vs. Visual-shape), Prime Type (yellow-related concept vs. a green-related concept) and Target Color (Yellow vs. Green) as within-subjects variables.

Analysis revealed neither significant main effects of the Nature of Memory Load, $F(1,23)=1.79, p=.19, \eta_{\mathrm{p}}^{2}=.07$, Prime Type, $\mathrm{F}<1$, Target Color, $\mathrm{F}<1$ nor interaction between Nature of Memory Load and Prime Type, $\mathrm{F}<1$, and between Nature of memory Load and Target Color, F < 1. Analysis showed a significant interaction between Prime Type and Target Color, $F(1,23)=7.14, p<.05, \eta_{p}^{2}=.24$. However, the Nature of Memory Load modulates this interaction. Indeed analysis revealed a significant three way interaction between Nature of the Memory Load, Prime Type and Target Color, $F(1,23)=7.33, p<.05, \eta_{p}^{2}=.24$. This interaction is depicted in Figure 2.


Figure 2: Mean correct RT for each experimental condition. Error bar represents standard error.

As can be appreciated in Figure 2, the interaction between Prime Type and Target Color was observed in the lexical load condition, but not in the visual-shape load condition. Regarding the lexical load condition, participants were significantly faster at judging the green patch preceded by a congruent prime (i.e., a green-related concept like a lettuce) than an incongruent prime (i.e., a yellow-related concept like a banana), $\mathrm{F}(1,23)=7.60, \mathrm{p}<.05$. The reverse was observed for the yellow patch, $\mathrm{F}(1,23)=4.61, \mathrm{p}<.0 .5$.

Regarding the visual-shape load condition. The Prime Type did not modulate color discrimination. The difference between the primes was not significant for the yellow patch, $\mathrm{F}<1$, as well for the green patch, $\mathrm{F}<1$.

Recognition Task - A t-test conducted between the correct recognition rates of the different memory load conditions revealed that participants were significantly worst for the memory test in the visual-shape load condition ( $\mathrm{M}=.672$, $\mathrm{SE}=.021)$ than in lexical condition $(\mathrm{M}=.729, \mathrm{SE}=.015)$, $\mathrm{t}(23)=3.25, \mathrm{p}<.05$.

Correlation Analysis - We also tested the correlation between the memory test accuracy and the priming effect size (RT congruent - RT incongruent) ${ }^{1}$ for both condition of memory load and each participant. Indeed, Vermeulen and collaborators (2008) showed that memory performances were selectively influenced by the relation between the nature of the load (i.e., visual or auditory) and the nature of the to be verified property (i.e., visual or auditory). Accordingly, we might expect a negative correlation between the priming effect size and the memory performance in the visual-shape load condition. In other words, less the visual-shape load is efficient (attesting by a high recognition rate) the higher is the probability to observe a priming effect. Conversely, we should not observe any correlation for the lexical load condition.

[^292]We found a significant negative correlation (Spearman Correlation) between the priming effect size and the memory performance for the visual-shape load condition, $r(22)=-.46, p<.05$, but not for lexical load condition, $r(22)$ $=+.02$.

## Discussion

The aim of this study was to propose a procedure for disentangling between several conceptions about nature of color knowledge in memory. To do so, we combined a color-priming paradigm (Heurley, et al., 2013) with a shortterm memory load procedure (Vermeulen, et al., 2008). First, participants had to learn three visual elements (either non-words or shapes). Then, they had to perform a color discrimination task where each colored-target (e.g., yellow patch) was preceded by a congruent color-related word concept (e.g., "banana") or not (e.g., "lettuce"). Finally, they performed a recognition task on either non-words or visualshapes depending on the memory load condition. Our results seem to indicate that access to color knowledge involves perceptual simulation rather than lexical or semantic activation. Indeed, we found that color-priming effect (i.e., shorter RTs when the color of the patch was congruent with color-related concept word rather than incongruent) was incurred in the visual-shape load condition while it was observed in the lexical load condition. This should be due to a competition for same visual resources between the short-term storage of shapes in memory and the simulation of the color-related concepts (see Vermeulen, Corneille et al., 2008; Vermeulen, Chang et al., in press, for a similar conclusion). Moreover, the fact that we found a significant negative correlation between priming size effect and the accuracy in short-term memory task in the visualshape load condition (while the same correlation was not significant in the lexical load condition) is consistent with our interpretation. Moreover, this result is in accordance with Yee and co-workers' (2012) experiment. Indeed, they found a positive correlation between Stroop interference and color-priming gain only when the Stroop task was presented before the priming procedure. This result attested that colorpriming effect was modulated by the Stroop task. Taken together, these results indicate that access to color information related to object is not only contextually dependent but also sensory-based. Finally, our results bring direct evidence that access to an object concept using words involved automatic modal simulation (see also Vermeulen et al., in press). Indeed, this paper showed that words representing modal concept spontaneously involve perceptual simulations (without engaging participant in a property verification task) so that a perceptual load (in the same modality than the modal concept) selectively incurs memory for these words.

In conclusion, our study provides a strong argument in favor of the idea that access to conceptual knowledge is linked to the simulation of the sensory dimension captured within the concept (see Barsalou, 2008) so that experiencing
a concept in a given modality involves perceptual simulation in the same sensory modality and in the other related sensory (Brunel, Lesourd, et al., 2010) or motor modalities (Brouillet, Heurley, Martin, \& Brouillet, 2010).

## References

Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645. doi: 10.1146/annurev.psych.59.103006.093639

Brouillet, T., Heurley, L., Martin, S., \& Brouillet, D. (2010). The embodied cognition theory and the motor component of "yes" and "no" verbal responses. Acta Psychologica, 134, 310-317. doi: 10.1016/j.actpsy.2010.03.003
Brunel, L., Goldstone, R. L., Vallet, G., Riou, B., \& Versace, R. (2013). When seeing a dog activates the bark: multisensory generalization and distinctiveness effects. Experimental Psychology, 60, 100-112. doi: 10.1027/1618-3169/a000176

Brunel, L., Labeye, E., Lesourd, M., \& Versace, R. (2009). The sensory nature of episodic memory: sensory priming effects due to memory trace activation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 1081-1088. doi: 10.1037/a0015537
Brunel, L., Lesourd, M., Labeye, E., \& Versace, R. (2010). The sensory nature of knowledge: Sensory priming effects in semantic categorization. Quartely Journal of Experimental Psychology, 63, 955-954. doi: 10.1080/17470210903134369

Connell, L. (2007). Representing object color in language comprehension. Cognition, 102, 476-485. doi: 10.1016/j.cognition.2006.02.009

Connell, L., \& Lynott, D. (2009). Is bear is white in the woods? Parallel representation of implied object color during language comprehension. Psychonomic, Bulletin \& Review, 16, 573-577. doi: 10.3758/PBR.16.3.573
Hansen, T., Olkkonen, M., Walter, S., \& Gegenfurtner, K. R. (2006). Memory modulates color appearance. Nature Neuroscience, 9, 1367-1368. doi: 10.1038/nn1794
Heurley, L. P., Brouillet, T., Chesnoy, G., \& Brouillet, D. (2013). Color perception involves color representations firstly at a semantic level and then at a lexical level. Cognitive Processes, 14, 19-29. doi: 10.1007/s10339-012-0527-z
Huettig, F., \& Altmann, G.T.M. (2011). Looking at anything that is green when hearing 'frog' - How object surface color and stored object color knowledge influence language- mediated overt attention. The Quarterly Journal of Experimental Psychology, 64, 122-145. doi: 10.1080/17470218.2010.481474

Landauer, I. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104, 211-240.

Masson, M.E.J. (1995). A distributed memory model of semantic priming. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 3-23.
Nijboer, T. C. W., van Zandvoort, M. J. E., \& de Haan, E. H. F. (2006b). Seeing red primes tomato: Evidence for comparable priming from color and color name primes to semantically related word targets. Cognitive Processes, 7, 269-274. doi: 10.1007/s10339-006-0153-8.
Oliva, A., \& Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. Cognitive Psychology, 41, 176-210
Pecher, D., Zeelenberg, R., \& Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. Psychological Science, 14, 119124.

Pecher, D., Zeelenberg, R., \& Barsalou, L. W. (2004). Sensorimotor simulations underlie conceptual representations: modality-specific effects of prior activation. Psychonomic Bulletin \& Review, 11, 164-167.
Reilhac, G., \& Jiménez, M. (2006). Amorçage de la couleur typique d'un objet lors d'une tâche de catégorisation. Canadian Journal of Experimental Psychology, 60, 285293
Reinitz, M. T., Lammers, W. J., \& Cochran, B. P. (1992). Memory-conjunction errors: miscombination of stored stimulus features can produce illusions of memory. Memory \& Cognition, 20, 1-11.
Richter, T., \& Zwaan, R. A. (2009). Processing of color words activates color representations. Cognition, 111, 383-389. doi: 10.1016/j.cognition.2009.02.011
Roberson, D., \& Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. Memory \& Cognition, 28, 977-986.
Simmons, W. K., Ramjee, V., Beauchamp, M. S., Mcrae, K., Martin, A., \& Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. Neuropsychologia, 45, 2802-2810. doi: 10.1016/j.neuropsychologia.2007.05.002

Tanaka, J. W., \& Presnell, L. M. (1999). Color diagnosticity in object recognition. Perception \& Psychophysics, 61, 1140-1153.
Vallet, G., Brunel, L., \& Versace, R. (2010). The perceptual nature of the cross-modal priming effect: arguments in favor of a sensory-based conception of memory. Experimental Psychology, 57, 376-382. doi: 10.1027/1618-3169/a000045

Van Dantzig, S., Pecher, D., Zeelenberg, R., \& Barsalou, L. (2008). Perceptual Processing Affects Conceptual Processing. Cognitive Science: A Multidisciplinary Journal, 32, 579-590. doi: 10.1080/03640210802035365
Vermeulen, N., Chang, B., Corneille, O., Pleyers, G., \& Mermillod, M. (2013). Verifying properties of concepts spontaneously requires sharing resources with samemodality percept. Cognitive Processes, 14, 81-87. doi: 10.1007/s10339-012-0533-1

Vermeulen, N., Chang, B., Mermillod, M., Pleyers, G., \& Corneille, O. (2013). Memory for words representing
modal concepts: resource sharing with same-modality percepts is spontaneously required. Experimental Psychology, 60, 293-301. doi: 10.1027/16183169/a000199
Vermeulen, N., Corneille, O., \& Niedenthal, P. M. (2008). Sensory load incurs conceptual processing costs. Cognition, 109, 287-294. doi: 10.1016/j.cognition.2008.09.004

Versace, R., Labeye, E., Badard, G., \& Rose, M. (2009). The contents of long-term memory and the emergence of knowledge. European journal of cognitive psychology, 21, 522-560.
Yee, E., Ahmed, S. Z., Thompson-Schill, S. L. (2012). Colorless green ideas (can) prime furiously. Psychological Science, 23, 364-369. doi: 10.1177/0956797611430691

# Recursion in Children's Comprehension and Formulation of Algorithms 

Monica Bucciarelli<br>Università di Torino, Italy<br>Robert Mackiewicz<br>University of Social Sciences and Humanities, Warsaw, Poland<br>Sangeet Khemlani<br>US Naval Research Laboratory, Washington, DC 20375 USA<br>Philip Johnson-Laird<br>Princeton University, Princeton, New Jersey, United States and New York University, New York, NY<br>10003, USA


#### Abstract

Recursive loops in informal algorithms are difficult to formulate, even for naïve adults (Khemlani et al., 2013). Children can formulate algorithms that do not require loops (Bucciarelli et al., 2016), and anecdotal evidence suggests that they can understand loops. As there were no previous studies, we examined how they made deductions of the consequences of loops, and how they abduced loops in creating informal algorithms in everyday language. We therefore tested fifth-grade children's ability carry out both these tasks in algorithms that rearrange the order of cars on a toy railway track with one siding. Experiment 1 showed that they could deduce rearrangements from algorithms containing loops, and Experiment 2 showed that they could formulate at least some algorithms that contained loops. These abilities are the likely precursors to the comprehension of recursion and to computer programming.


# Is Structural Priming in Children Facilitated by Interactions between Animacy and Syntax? 

Leone Buckle (leone.buckle@postgrad.manchester.ac.uk)<br>ESRC International Centre for Language and Communicative Development (LuCiD), Coupland 1, University of Manchester, Oxford Road, Manchester, M13 9PL<br>Anna Theakston (anna.theakston@manchester.ac.uk)<br>ESRC International Centre for Language and Communicative Development (LuCiD), Coupland 1, University of Manchester, Oxford Road, Manchester, M13 9PL<br>Elena Lieven (elena.lieven@manchester.ac.uk)<br>ESRC International Centre for Language and Communicative Development (LuCiD), Coupland 1, University of Manchester, Oxford Road, Manchester, M13 9PL


#### Abstract

Sentence production relies on the activation of both semantic information (e.g. noun animacy) and syntactic frames that specify an order for grammatical functions (e.g. subject before object; Levelt, Roelofs \& Meyers, 1999). However, it is unclear whether these semantic and syntactic processes interact (Gámez \& Vasilyeva, 2015), and if this changes developmentally. We thus examined the extent to which animacy-semantic role mappings in dative prime sentences and target scenes influenced choice of syntactic structure. 143 participants ( 47 three year olds, 48 five year olds and 48 adults) alternated with the experimenter in describing animations. Animacy mappings for themes and goals were either prototypical or non-prototypical and either matched or mismatched across the experimenter's prime scenes and participants' target elicitation scenes. Prime sentences were either double-object datives (e.g. the girl brought the monkey a ball) or prepositional datives (e.g. the girl brought the ball to the monkey). Participants' target sentences were coded for syntactic form. All age groups showed a main structural priming effect. For the youngest group, animacy-semantic role mappings facilitated prepositional dative priming. No animacy facilitation was found for the older groups. Our results demonstrate the changing influence of animacy cues on sentence production through interactions with syntactic structure over the course of development. The theoretical implications of our findings are discussed.


Keywords: structural priming; animacy; language production; semantics; syntax.

## Introduction

In order to communicate ideas, speakers must map concepts to syntactic structures. Where one idea can be expressed using multiple structures, speakers tend to use the most recently heard structure (Bock, 1986). For example, structural priming occurs where speakers are more likely to describe the transfer of a ball between a girl and a monkey using the double-object dative (DOD) sentence the girl brought the monkey a ball instead of the prepositional dative
(PD) structure the girl brought a ball to the monkey, following a DOD, rather than a PD prime. This occurs in children (Rowland, Chang, Ambridge, Pine \& Lieven, 2012) and adults (Bock, 1986).

In their residual activation theory, Pickering and Branigan (1998) argue that abstract representations of verbs, grammatical roles (e.g. direct object) and combinatorial notes are activated upon hearing a DOD sentence (i.e. NPNP). Structural priming occurs where speakers reuse the currently activated NP-NP node to produce another DOD construction rather than activating the alternative NPprepositional phrase ( PP ) node to produce a PD sentence.

The residual activation theory cannot, however, account for instances where structural priming is enhanced by animacy-syntax interactions. Gámez and Vasilyeva (2015) found that priming of passive sentences in children was greatest where primes and targets both contained animate patients and inanimate agents. For datives, animacy may interact with semantic role-grammatical function mappings (e.g. theme-direct object), before these mapped constituents are ordered, to determine syntactic structures (de Swart, Lamers \& Lestrade, 2008). Prototypical DOD sentences contain animate goals before animate themes, whereas prototypical PD sentences feature inanimate themes before animate goals (Bresnan, Cueni, Nikitina \& Baayen, 2007). Demuth, Machobane, Maloi and Odato (2005) found that children best understood double object applicatives in Sesotho where they contained human, rather than inanimate, benefactives before inanimate, as opposed to animate, themes. These studies suggest that structural and semantic information may be inseparable and represented at varying levels of granularity (Ambridge, Kidd, Rowland \& Theakston, 2015).

However, methodological problems with structural priming studies have made it unclear whether animacysyntax interactions could drive priming effects (Chang, Bock \& Goldberg, 2003). DOD sentences (e.g. the girl brought the monkey a ball) may prime participants to repeat the abstract syntactic frame and produce DOD targets. Alternatively, they may prime speakers to reuse the
animate-inanimate noun ordering, leading them to produce PD sentences where targets contain animate themes and inanimate goals (e.g. the boy brought the tiger to the zoo). Studies manipulating both prime and target animacy cues are needed to identify whether semantic processes influence structural priming (Goldwater, Tomlinson, Echols \& Love (2010).

Priming could be greater with semantically prototypical primes (e.g. Vasilyeva \& Waterfall, 2015). Alternatively, priming might be greater with non-prototypical primes because they are more salient, according to Chang, Dell and Bock's (2006) error-based learning theory. Error-based learning effects may decrease with age due to increased exposure to uncommon sentence types (Peter, Chang, Pine, Blything \& Rowland (2015). Priming is greater in children (Goldwater, et al., 2010) and adults (Cleland and Pickering, 2003) where primes and targets are semantically similar. Sensitivity to animacy and its effects on structural priming may also decrease with age (Corrigan, 1988).

Priming research may provide insight into how children extract representations of grammatical functions and animacy-semantic role mappings from caregiver speech to produce their own sentences (Bock, Dell, Chang \& Onishi, 2007; Pickering \& Ferreira, 2008). By investigating possible specification of semantic, and not just lexical, information in children's sentence representations, we can more accurately conclude whether or not representations are entirely abstract (Rowland \& Noble, 2010).

We assessed the extent to which structural priming in three year olds, five year olds and adults was influenced by interactions between animacy cues and syntax by manipulating prime structures (DOD/PD), prime animacysemantic role mappings (prototypical [AN goal \& IN theme]/non-prototypical [AN theme \& IN theme]), and prime-target match in animacy-semantic role mappings (match/mismatch).

Prior research implies relatively strong interactions between animacy and syntax and that these effects on sentence processing are greater in younger children than in older children and adults. Thus, we tested the following hypotheses: (i) structural priming effects will be greater where primes have prototypical animacy cues. Alternatively, error-based learning may entail greater priming with reversed cues, (ii) priming will be greater where primes and targets have matching animacy-semantic role mappings, (iii) the relative increase in priming where animacy-semantic role mappings are prototypical and matching across primes and target pairs will decrease with age.

## Method

## Design

We used a $3 \times 2 \times 2 \times 2$ mixed design. Age (3 years/5 years/adults) and prime structure (double-object dative [DOD]/prepositional dative [PD]) were between-subject
independent variables. Prime animacy-semantic role mappings (prototypical [AN goal \& IN theme]/ nonprototypical [AN theme \& IN goal] and prime-target match in animacy-semantic role mappings (match/mismatch) were within-subjects independent variables. The production of DOD target responses was our dependent variable.

## Participants

We tested 143 monolingual British English speakers; 47 three year olds ( 24 females), 48 five year olds ( 25 females), and 48 adults ( 35 females). One three year old was excluded for their failure to produce any dative sentences.

## Visual Stimuli

Sixty-eight 10 -second animations were created in Anime Studio Pro 10 and presented on a laptop using Microsoft PowerPoint. Forty-eight ( 24 for primes and 24 for targets) portrayed ditransitive events (e.g. a girl bringing a monkey a ball). Twenty depicted intransitive events featuring two characters simultaneously acting in the centre of the screen (e.g. a boy and girl jumping). Eight of these were used as practice scenes (four each for the experimenter and participant) and 12 were used as fillers (six each).

## Sentence Stimuli

Eighty- two sentences were created as descriptions for the 68 animations. These included:

- Practice Items (4): Intransitive sentences for the experimenter's turn in practice trials to introduce participants to the task.
- Fillers (6): Present-tense intransitive sentences for the experimenter's turn in filler trials to limit priming effects across prime-target pairs.
- Primes (48): Past tense dative sentences which included 24 DOD and 24 PD counterparts corresponding to the 24 prime scenes. Six different prime sentences were assigned to each of the four experimental conditions.
- Targets (24): Six different verbs were included in sentence initiations for target sentences (e.g. the boy brought). Primes and targets always contained the same verb and participants completed these sentence initiations to produce the full target sentence. See Table 1 for example prime sentences and target elicitation scenes.


## Procedure

The experimenter played the animations on a laptop, beginning with four practice-practice trials, followed by alternating prime-target and filler-filler trials. She described the first scene and produced the first sentence in each pair, producing all primes and participants described the second scene in each pair, including all targets. On target trials, the
experimenter produced initial sentence initiations (e.g. the girl brought...) to encourage participants' use of datives. Participants formed their own target structures as they finished the sentence (e.g. the monkey a ball or the ball to the monkey). Adult participants often produced entire target sentences including the initial subject and verb.

Table 1: Example prime sentences and target elicitation scenes for each condition

| Condition | DOD Prime | PD Prime | Target Elicitation Scene |
| :---: | :---: | :---: | :---: |
| Prototypical Prime (AN goal \& IN theme) / Matched Target | The girl brought the monkey a ball | The girl brought a ball to the monkey | Transfer of a flower from boy to a snail |
| Prototypical <br> Prime <br>  <br> IN theme) / <br> Mismatched <br> Target | The girl brought the bee a flower | The girl <br> brought a flower to the bee | Transfer of a monkey from a boy to a zoo |
| Non- <br> prototypical <br> Prime <br>  <br> IN goal) / <br> Matched <br> Target | The girl brought the zoo a tiger | The girl brought a tiger to the zoo | Transfer of a bee from a boy to a zoo |
| Non- <br> prototypical <br> Prime <br>  <br> IN goal) / <br> Mismatched <br> Target | The girl brought the garden a snail | The girl brought a snail to the garden | Transfer of a ball between a boy and a tiger |

## Coding

Target responses were coded for syntactic structure (doubleobject dative [DOD], prepositional dative [PD] and OTHER). Only DOD and PD target sentences were included in the analyses.

DOD: sentences with a goal - theme structure (e.g. the boy brought the tiger a ball).

PD: sentences with a theme - preposition - goal structure (e.g. the boy brought a tiger for the monkey). Both to and for were suitable prepositions.
OTHER: Such responses were excluded from the analyses and included:

1. Sentences without a DOD or PD structure (e.g. intransitive and/or incomplete sentences with only one noun such as the boy threw the whale, or locatives such as the boy threw the way into the sea).
2. Incomplete sentences with one object and a preposition but no second object (e.g. the boy threw the food to)
3. Sentences where nouns were assigned to the wrong semantic role (e.g. the boy brought the ball [goal] a tiger [theme], where the target scene actually showed the transfer of a ball [theme] between a boy and tiger [goal]. A misunderstanding of the target scene may influence target structures where animacy cues might interact with syntactic structures.
4. Sentences with incorrectly named nouns, indicating participant's misunderstanding of the event shown in the target scene (e.g. the boy brought the zoo/mouse a ball instead of the boy brought the tiger a ball).
The percentage of OTHER target responses was $38 \%$ in three year olds, $28 \%$ in five year olds and $27 \%$ in adults. This is to be expected because although our events involved three participants, it is perfectly acceptable to focus on only a subset of these in a linguistic description of the scenes.

## Results

The data were analysed using logistic mixed effects models in R, using the glmer function of the lme4 package (lme4 version 1.1-11: R Core Team 2012). Fixed effects for all final models included: age ( 3 years $=-1 ; 5$ years $=0$; adult $=$ 1), prime animacy-semantic role mappings (prototypical [AN theme -IN goal] $=1$; non-prototypical [IN theme AN goal] $=0$ ) and prime-target match in animacy-semantic role mappings (match $=1$; mismatch $=0$ ). All variables were centred to reduce multicollinearity (Neter, Wasserman \& Kuttner, 1985). Participant was always included as a random effect. Sentence item was excluded as a random effect and the analyses were separated by age since the model initially fitted to the full data set did not converge. For each individual age group, the Bonferroni method was used with a corrected alpha level of .025 for post-hoc analyses. The mean proportion of DOD target responses produced in each condition is shown in Figure 1.

## Age Three

The model initially contained only main effects of prime structure, prime animacy-semantic role mappings and prime-target match, but was significantly improved by adding a three-way interaction term and all the two-way interaction terms that are derived from it $(p=.03)$. We found a significant main effect of prime structure whereby
more DOD targets were produced following DOD ( $M=$ $0.27, S E=0.02$ ) as opposed to PD primes $(M=0.06, S E=$ 0.01 ) and a significant three-way interaction between prime structure, prime animacy-semantic role mappings and prime-target match.

To interpret the three-way interaction a model was fitted for each level of prime structure (DOD and PD). Analysis of DOD primes failed to reveal any significant effect for prime animacy-semantic role mappings, $\beta=0.20(S E=0.31), z=$ $0.65, p=.518$, prime-target match, $\beta=-0.12(S E=0.31), z$ $=-0.40, p=.688$, or the interaction between the variables, $\beta=0.61(S E=0.61), z=1.02, p=.31$. Analysis of PD primes, however, revealed a significant two-way interaction between prime animacy-semantic role mappings and primetarget match, $\beta=3.89$ ( $S E=1.39$ ), $z=2.81, p=.005$.

Two further models were run for PD primes, one for each level of animacy-semantic role mapping (prototypical [AN goal \& IN theme]/non-prototypical [AN theme \& IN goal]). Where PD primes featured non-prototypical animacysemantic role mappings, there was a marginally significant effect of prime-target match, $\beta=-2.33(S E=1.07), \mathrm{z}=-$ $2.19, \mathrm{p}=.029$. Fewer DOD responses were produced where targets contained matched (non-prototypical) animacysemantic role mappings ( $M=0.01, S E=0.03$ ) as opposed to mismatched (prototypical) animacy-semantic role mappings $(M=0.11, S E=0.03)$. However, where PD primes contained prototypical animacy-semantic role mappings there was no significant effect of prime-target match, $\beta=$ $1.47(S E=0.85), z=1.73, p=.08$.

## Age Five

The model originally featured only main effects but was significantly improved by adding two-way interaction terms between the variables $(p=.007)$. There was a significant main effect of prime structure whereby more DOD targets were produced following DOD ( $M=0.30, S E=0.02$ ) as opposed to PD primes $(M=0.02, S E=0.01)$ and a significant two-way interaction between prime animacysemantic role mapping and prime-target match $\beta=1.15$ (SE $=0.51), z=2.28, p=.002$.

To interpret the two-way interaction, a model was fitted for each level of prime animacy-semantic role mapping (prototypical [AN goal \& IN theme]/non-prototypical [AN theme \& IN goal]). For prototypical prime animacysemantic role mappings there was a significant effect of prime-target match, $\beta=0.83(S E=0.35), z=2.35, \mathrm{p}=.018$. DOD production was higher where targets featured matched (prototypical; $M=0.18, S E=0.02$ ) as opposed to mismatched (non-prototypical; $M=0.09, S E=0.02$ ) animacy-semantic role mappings. However, where primes contained non-prototypical animacy-semantic role mappings there was no effect of prime-target match, $\beta=-0.48$ ( $S E=$ $0.36), z=-1.35 p=.177$. There was no difference in the production of DOD targets where targets featured matched (non-prototypical; $M=0.14, S E=0.03$ ) as compared with
mismatched (prototypical; $(M=0.09, S E=0.03)$ animacysemantic role mappings.


Figure 1: The mean proportion of DOD responses following DOD and PD primes where primes contained either prototypical or non-prototypical animacy-semantic role mappings and these mappings were either matched or mismatched across primes and targets ( $S E$ in error bars).

## Adults

The model originally featured only main effects but was significantly improved by adding two-way interaction terms between the variables ( $p<.001$ ). We found a significant effect of prime structure with more DOD targets produced following DOD $(M=0.64, S E=0.02)$ than PD primes $(M=$ $0.07, S E=0.12$ ) and a significant two-way interaction between prime animacy-semantic role mapping and primetarget match.

To interpret the two-way interaction a model was fitted for each level of prime animacy-semantic role mapping (prototypical [AN goal \& IN theme]/non-prototypical [AN theme \& IN goal]). For primes with prototypical animacysematic role mappings we found a significant effect of prime-target match, $\beta=2.608(S E=0.51), z=5.09, p<$ .001. DOD production was higher where targets featured matched (prototypical; $M=0.43, S E=0.03$ ) as opposed to mismatched (non-prototypical; $(M=0.25, S E=0.03)$ animacy-semantic role mappings. Where primes contained non-prototypical animacy-semantic role mappings, there was also a significant effect of prime-target match $\beta=-1.33$ $(S E=0.43), z=-3.12, p<.001$. Fewer DOD responses were produced where targets contained matched (nonprototypical; $M=0.30, S E=0.03$ ) as opposed to mismatched (prototypical; $(M=0.40, S E=0.03)$ animacysemantic role mappings.

## Results Summary

All age groups showed an effect of structural priming, producing more DOD responses following DOD primes, as compared to PD primes. Three year olds also exhibited effects of animacy-semantic role mappings on the magnitude of structural priming, showing an increase in PD sentence priming effects where primes and targets contained matching non-prototypical mappings (AN theme \& IN goal) (although no effects were observed for DOD primes). Our hypothesis that priming would be greater with prime-target match was met, indicating that animacy mappings were represented to a relatively strong degree. However, priming increased with non-prototypical rather than prototypical primes providing support for error-based learning accounts. As expected, animacy effects decreased with age; they had no influence on structural priming in five year olds or adults. Nevertheless, animacy did influence DOD target production in five year olds and adults, independently of prime structure. They produced more DOD sentences where targets (and also primes in the case of five year olds) contained prototypical animacy-semantic role mappings (AN goal \& IN theme).

## Discussion

Our results support claims of structural priming effects in children (Rowland et al., 2012) and adults (Bock, 1986) and more importantly, they provide further clarification as to how structural priming works. Our results reveal that priming relies, first and foremost, on the repetition of abstract syntactic frames and not the repetition of animacy noun orders. This was previously unclear due to methodological issues with earlier research (Chang, Bock \& Goldberg, 2003). Non-prototypical (AN theme \& IN goal) DOD primes with an inanimate-animate noun order (e.g. the girl brought the zoo a monkey) were just as likely to yield DOD targets as prototypical (AN goal \& IN theme) DOD primes with an animate-inanimate noun order (e.g. the girl brought the monkey a ball). Mere repetition of animacy noun ordering would have resulted in more prototypical PD targets with an inanimate-animate noun order (e.g. the girl brought the flower to the snail) following non-prototypical DOD targets. All age groups showed a main structural priming effect, suggesting that children's linguistic representations do not need to specify animacy-semantic role mappings for priming to occur.

PD sentence priming was enhanced in three year olds where there was prime-target match in non-prototypical (AN theme \& IN goal) animacy-semantic role mappings. This is consistent with Gámez and Vasilyeva's (2015) finding that prime-target match increased priming in five and six year olds. Our results are thus at odds with Pickering and Branigan's (1998) residual activation theory as it cannot explain how semantic information could influence structural priming. We found that animacy-semantic role mappings were specified and represented to a relatively strong degree and could influence priming through error-based learning in
support of Chang, et al (2006). We suggest that error-based learning may have been less likely to occur following DOD primes as three year olds generally use fewer DOD than PD constructions (Rowland, et al., 2012). They might not have been sensitive enough to the typical animacy mappings in DOD sentences for surprisal priming effects to occur.

Five year olds and adults showed no evidence of increased priming where primes contained non-prototypical (AN theme \& IN goal), as opposed to prototypical (AN goal \& IN theme) animacy-semantic role mappings. This developmental decrease in error-based learning may be due to increased exposure to such lower frequency sentence types and is consistent with Rowland, et al., (2012) and Peter, et al.'s (2015) results. Our results also complement those of Corrigan (1988) who found animacy effects on children's sentence interpretations to decrease with age.

Nevertheless, five year olds and adults produced more DOD targets where targets (and also primes for five year olds) contained prototypical (AN goal \& IN theme) mappings, regardless of prime structure. This indicates a preference to use animate goals and inanimate themes in DOD as opposed to PD constructions. These data therefore support claims that animacy interacts with semantic rolegrammatical function mappings and can influence subsequent word order (Zorzi \& Vigliocco, 1999; Goldberg, 1995; Garrett, 1975). Animate goals tended to be realised as indirect objects more than as oblique objects and were often placed before inanimate theme-direct objects, resulting in DOD constructions (e.g. brought the monkey [animate goal] a ball [inanimate theme]).

Speakers' tendency to encode animate goals and inanimate themes in DOD constructions increased with age. Very small surprisal effects may have moderated five year olds' DOD production. Non-prototypical (AN theme \& IN goal) PD primes may have subtly increased PD sentence priming through error-based learning. Non-prototypical DOD primes may have sometimes primed participants to reuse noun animacy orders in prototypical PD constructions. E.g. the DOD prime the girl brought the zoo [inanimate] a monkey [animate] could have prompted the PD response the boy brought the ball [inanimate] to the tiger [animate]).

We should however, also seek to clarify whether animacy could influence word orders in sentence production independently of syntax and/or grammatical roles. Bock, Loebell and Morey (1992) provide evidence to suggest that this is possible. Following primes with animate subjects before inanimate objects, participants were more likely to produce targets with the same noun animacy order than an inanimate subject-animate object order. It did not matter whether subjects were agents or patients of active or passive sentences. Little research has been conducted to address this topic in adults and it is yet to be explored in children.

## Conclusion

In our study animacy-syntax interactions appeared to facilitate structural priming in young children but this effect was subject to a developmental decrease. The extent to
which animacy-semantic role mappings could influence speakers' choice of syntactic structure independent of structural priming, rather increased with age. Animacysyntax interactions can therefore influence sentence production. We consequently propose that theories of structural priming and sentence production in general should seek to consider the role of animacy-syntax interactions.

## Acknowledgements

This work was funded by the Economic and Social Research Council (ESRC) North West Doctoral Training Centre (NWDTC) $1+3 \mathrm{PhD}$ award, grant number: 1512180 (Buckle) and was additionally supported by The University of Manchester President's Doctoral Scholar Scheme (Buckle, PhD). Professors Elena Lieven and Anna Theakston are supported by the ESRC Centre for Language and Communicative Development (LuCiD, ESRC grant number: ES/L008955/1).

## References

Ambridge, B., Kidd, E., Rowland, C. \& Theakston, A. (2015). The ubiquity of frequency effects in first language. Journal of Child Language, 42(2), 239-273.
Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12(4), 575.

Bock, K., Dell, G. S., Chang, F., \& Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. Cognition, 104(3), 437-458.
Bock, K., Loebell, H., \& Morey, R. (1992). From conceptual roles to structural relations: bridging the syntactic cleft. Psychological review, 99(1), 150.
Bresnan, J., Cueni, A., Nikitina, T., \& Baayen, R. H. (2007). Predicting the dative alternation. Cognitive foundations of interpretation, 69-94.
Chang, F., Bock, K., \& Goldberg, A. E. (2003). Can thematic roles leave traces of their places?. Cognition, 90(1), 29-49
Chang, F., Dell, G. S., \& Bock, K. (2006). Becoming syntactic. Psychological review, 113(2), 234.
Cleland, A. A., \& Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. Journal of Memory and Language, 49(2), 214230.

Corrigan, R. (1988). Children's identification of actors and patients in prototypical and nonprototypical sentence types. Cognitive Development, 3, 285-297.
de Swart, P., Lamers, M., \& Lestrade, S. (2008). Animacy, argument structure, and argument encoding. Lingua, 118(2), 131-140.

Demuth, K., Machobane, M., Moloi, F., \& Odato, C. (2005). Learning animacy hierarchy effects in Sesotho double object applicatives. Language, 421-447.
Gámez, P. B., \& Vasilyeva, M. (2015). Exploring interactions between semantic and syntactic processes: The role of animacy in syntactic priming. Journal of Experimental Child Psychology, 138, 15-30.
Garrett, M.F. (1975). The analysis of sentence production. In G.H. Bower (Ed.), The psychology of learning motivation. New York: Academic Press.
Goldberg, A. E. (1995). Constructions: a construction grammar approach to argument structure. University of Chicago Press, Chicago, IL.
Goldwater, M. B., Tomlinson, M. T., Echols, C. H., \& Love, B. C. (2011). Structural priming as structure Mapping: Children use analogies from previous utterances to guide sentence production. Cognitive Science, 35(1), 156-170.
Levelt, W. J., Roelofs, A., \& Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and brain sciences, 22(01), 1-38.
Neter, J., Wasserman, W., \& Kutner, M. H. (1985). Applied linear statistical methods. Richard D. Irwin. Inc., Homewood, IL.
Peter, M., Chang, F., Pine, J. M., Blything, R., \& Rowland, C. F. (2015). When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. Journal of Memory and Language, 81, 1-15.
Pickering, M. J., \& Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. Journal of Memory and Language, 39(4), 633-651.
Pickering, M. J., \& Ferreira, V. S. (2008). Structural priming: a critical review. Psychological bulletin, 134(3), 427.

R Core Team (2012). R: A language and environment for statistical computing. 3-900051-07-0. Vienna, Austria: R Foundation for Statistical Computing
Rowland, C. F., \& Noble, C. L. (2010). The role of syntactic structure in children's sentence comprehension: Evidence from the dative. Language Learning and Development, 7(1), 55-75.
Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M., \& Lieven, E. V. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. Cognition, 125(1), 49-63.
Zorzi, M., \& Vigliocco, G. (1999). Compositional semantics and the lemma dilemma. Behavioral and Brain Sciences, 22(01), 60-6

# Equiprobability principle or "no change" principle? Examining reasoning in the Monty Hall Dilemma using unequal probabilities 

Bruce D. Burns (bruce.burns@sydney.edu.au)<br>University of Sydney, School of Psychology, Brennan MacCallum Building (A18)<br>Camperdown, NSW 2006, Australia


#### Abstract

The Monty Hall Dilemma (MHD) is a well-known cognitive illusion. It is often claimed that one reason for the incorrect answers is that people apply the equiprobability principle: they assume that the probability of the two remaining options must be equal. An alternative explanation for assigning the same probabilities to options is that they had the same prior probabilities and people perceive no significant change. Standard MHD versions do not distinguish these possibilities, but a version with unequal prior probabilities could. Participants were given an unequal probabilities version of MHD and told that either the high or low probability option had been eliminated. This affected participants' choices and their posterior probabilities. Only $14 \%$ of participants' responses were consistent with applying the equiprobability principle, but $51 \%$ were consistent with a "no change" principle. Participants were sensitive to the implications of the prior probabilities but did not appear to use Bayesian updating.


Keywords: Monty Hall Dilemma, probabilistic reasoning, cognitive illusions, cognitive reflection

## Introduction

In the Monty Hall Dilemma (MHD) participants are presented with a scenario in which there are three doors, one of which conceals a prize. The participant initially chooses one and then some mechanism (a game show host in the original version) opens one of the other two doors to show that it does not conceal the prize. This mechanism never opens a door with the prize behind it, knows where the prize is, and it always operates. The participant is then offered the choice of staying with the original choice or switching to the other unopened door. Almost all participants say they would stay with their first choice and that the probability of winning is $50 \%$ (Granberg \& Brown, 1995), however the correct choice is to switch and the probability of winning is $2 / 3$. When Marilyn vos Savant published the MHD in her magazine column and gave its correct answer she reports having received thousands of letters with $92 \%$ of the general public disagreed with her, but so did $65 \%$ of letters with university addresses (vos Savant, 1997). As such the MHD has proved to be one of the best examples of a cognitive illusion (Piattelli-Palmarini, 1994) and it has been the subject of a number of research studies (see Krauss \& Wang, 2003; Tubau, Aguilar-Lleyda \& Johnson, 2015, for reviews).

## Factors driving the illusion

Tubau et al. (2015) point out that it has been observed in many empirical studies that when participants are given the MHD they display a strong tendency to see the two remaining
doors as equally likely to conceal the prize. Stibel, Dror, and Ben-Zeev (2009) found that even in the 100-door version of the MHD in which 98 doors were opened, participants still tended to say that the last two remaining doors had a $50 \%$ chance each of concealing the prize. Although Tubau et al. point out that a number of factors have been argued to contribute to the illusion in MHD, one factor that it has been argued is a strong factor is misapplication of the equiprobability principle (Falk, 1992; Johnson-Laird, Legrenzi, Girotto, Legrenzi, \& Caverni, 1999; Falk \& Lann, 2008). Once participants see the options as equally probable they then choose to "stay" due to either illusion of control (Granberg \& Dorr, 1998) or anticipation of regret (Gilovich, Medvec, \& Chen, 1995).
The equiprobability principle suggests that in the absence of any apparent reason to differentiate options, all options will be assigned the same probability (Johnson-Laird, et al, 1999). So if there are just two apparently identical options they must each have an equal $50 \%$ chances of being correct. The equiprobability principle has often been observed when people are faced with uncertain options. For example, Fischhoff, Parker, Bruine de Bruin, Palmgren, Dawes and Manski (2000) found that a large number (over 20\%) of US 16 year olds estimated a $50 \%$ chance of dying in the next year. Tversky and Kahneman (1974) report that when participants were presented with a party made up of people with one of two professions, but given a description of an individual that the representativeness heuristic could not allocate to one of the groups, they said the probability of him being in either group was $50 \%$ regardless of base-rate. Even when Burns and Wieth (2004) presented a variation of the MHD that induced $51 \%$ participants to see that it was better to switch, most still said the probability of wining was $50 \%$. Johnson-Laird et al. present ample evidence of the equiprobability principle being applied to a number of situations, but it appears to be particularly strong in the MHD. However the frequency of the $50 \%$ answer may not be as strong as evidence of application of equiprobability as has been assumed because there is more than one path to this answer.

Burns and Wieth (2004) suggested that a major barrier to correct reasoning about the MHD is that participants fail to understand the causal structure underlying it. Therefore they see the host's action in opening a door as having not having changed the underlying probabilities. Thus given that the two unopened doors had equal probabilities before any door was opened they may still have equal probabilities. Occasionally participants have report that each of the two doors now has a $33 \%$ chance of concealing the prize, making this reasoning
transparent, although mathematically incoherent. However participants may be applying this reasoning to conclude that the ratio of the two probabilities of the remaining doors has not changed, and thus the probability of winning after switching is $50 \%$. Thus participants could come to the same answer (i.e., that each door has the same $50 \%$ chance of concealing the prize) in two different ways: 1) by misapplying the equiprobability principle; 2 ) by assuming no change. With the standard version of the MHD it is difficult to distinguish which form of reasoning led to the $50 \%$ answer, but a modification used by Granberg (1999) could.

## MHD with unequal probabilities

Granberg (1999) tried to probe how people use conditional probabilities in the MHD by giving them an alternative version of the MHD with four doors and unequal probabilities. The Bayesian analysis shows that whether staying or switching has the highest probability of winning depends on which door is initially selected and which door is opened. The optimal strategy for this version is to first select the least likely alternative then switch away from it after a door is opened. Granberg gave participants 60 trials of either equal or unequal probability 4-door versions of the MHD. On the first trial only $11 \%$ switched in the equal-probability condition but even fewer ( $7 \%$ ) switched in the unequalprobability condition, despite there being a $75 \%$ chance of winning (on average) if switching in the later condition. Over 60 trials the switch rates improved in both conditions but at similar rates. Participants in unequal-probability condition increasingly utilized the optimal strategy, but still only used it on an average of about $30 \%$ of trials in their last block of 10. Granberg saw this as evidence that participants were satisficing (Simon, 1955).
Granberg (1999) did not ask participants about what they thought were the probabilities that they would win by switching, but by not doing so an opportunity was lost. Few participants would be expected to get these probabilities correct, given how poor participants are at getting the percentage correct in standard version of the MHD (Burns \& Wieth, 2004, found only $2 \%$ did so). However their incorrect answers could be windows into their reasoning. In the 3-door equal-probabilities version of the MHD the same percentage answer could result from either form of the erroneous reasoning identified above. If a participant was applying the equiprobability principle then they would say there is a $50 \%$, but they would answer the same if they thought that Monty had changed nothing (except for cutting down the options). However in an unequal-probabilities version applying the equiprobability principle would still yield a $50 \%$ chance of winning, but if they applied the "no change" principle then the probability of winning would be a function of the unequal probabilities. Thus an unequal probability version of the MHD can be used to probe to what extent are participants’ errors due to either of these principles.

## The current experiment

In the experiment presented here participants were given a three-door version of the MHD but told that the probability that Door A concealed the prize was $30 \%$, the probability that Door B did was $60 \%$ and the probability that Door C did was $10 \%$. (These probabilities were chosen because they yielded different correct answers to staying or switching depending on which door was opened, and because they allowed easy calculation of the percentages corresponding to reasoning I expected from participants.) They were then told that they had initially selected Door A (fixing the first choice eliminated the strategic considerations that were not the focus of this experiment), then that one of the other doors had been opened. Figure 1a shows diagrammatically the scenario for the condition in which Door B was the one unopened, and in Figure 1 b for when Door C was unopened.


Figure 1a: Diagrammatic depiction of the scenario presented to participants in the Door-B unopened condition.


Figure 1b: Diagrammatic depiction of the scenario presented to participants in the Door-C unopened condition.

Participants were then asked whether they would stay with their first choice or switch to the unopened door, and to express their probability of winning if they switched. They were also asked if they had ever seen a question like this before. There were two conditions: Door- $B$ in which participants were told that Door B was unopened; and Door$C$ in which participants were told that Door C was unopened.

Bayes' Theorem yields the correct answer to the problems in the two conditions. In the Door-B condition the optimal choice is to switch because there is a $80 \%$ chance of winning if they switch, but in the Door-C condition participants
should stay because they would have only a $40 \%$ chance of winning if they switched (these answers were confirmed by simulations assuming that if the prize was behind Door A then there was an equal chance of opening either Door B or C). If participants apply the equiprobability principle then they should indicate a $50 \%$ chance of winning by switching in both conditions. If they are applying the "no change" principle then in the Door-C conditions they may say $25 \%$ (.1/(.1+.3)) and in the Door-B conditions $67 \%$ (.6/(.6+.3)). However they may apply a cruder version of the "no change" principle and just repeat the prior probabilities of $10 \%$ for the Door-C condition or $60 \%$ for the Door-B condition.

Tubau et al. (2015) suggest that use of the equiprobability principle may be due to lack of cognitive reflection, so it may be possible to analyses reasoning about the MHD in terms of dual-system models (see Stanovich, 2011). Applying "no change" and repeating prior probabilities may be the least reflective response. However "no change" but updating the probabilities to maintain the same ratio may be the most reflective. Thus responses to the unequal probability MHD may also be used to probe how reflective was the thinking employed.

The goal of the experiment was to analyze the type of reasoning participants used in an unequal-probabilities version of the MHD and thus to investigate what factors are behind the cognitive illusion. An improved understanding of the MHD can improve understanding of how people reason about probabilities.

## Method

## Participants

A total of 373 participants completed the experiment as part of a class at the University of Sydney. Of these, 105 indicated that they had seen a similar question before, so they were excluded from this analysis. The remaining sample of 268 consisted of 160 women and 108 men.

## Materials and procedure

During a class, participants were presented with the task as part of a set of tasks completed on a computer. They were told to read the following questions carefully and answer all the questions. If there were any they did not know the answer to, then they were instructed to guess.
"Pretend you are on a game show, where you are allowed to choose one of three closed doors. Behind one door is a prize (a car), and behind each of the other doors is a goat. After you have chosen a door, the door remains closed for the time being. The game show host, Monty Hall, who knows what is behind the doors, now has to open one of the two unchosen doors and reveal a goat. After he shows you a goat, he asks you to decide whether you want to stay with your first choice or switch to the remaining unopened door."
"By watching the show many times you have calculated that there is a pattern to where the prize is initially placed. Door A has the prize $30 \%$ of the time, Door B has the prize $60 \%$ of the time and Door C has the prize only $10 \%$ of the
time." Participants were then instructed to pretend they had first chosen Door A and that Monty Hall then opens Door B (if in the Door-C condition, otherwise Door C) and reveals a goat. Now he asks you whether you want to stick with your first choice (Door A) or switch to the Door C (or Door B if in the Door-B condition). Participants were also shown Figure 1a or Figure 1b, which ever was relevant to their condition.

They were then asked "Would you choose to switch doors or stay with your original door (Door A)?" and "What do you think is the chance of winning the prize if you switch doors (to Door B) [Door C in Door-C condition]?

Participants also completed the Cognitive Reflection Task (CRT) of Frederick (2005) which consists of three problems that require participants to reflect on the answers rather than give the obvious ones. It is used to assess the extent to which participants are reasoning reflective and thus using System 2 rather than relying on System 1 (in terms of Stanovich, 2011).

## Results

## Choice

After eliminating participants who reported having seen the question before and two with missing responses, there were 115 in the Door-B condition and 151 in the Door-C condition. Table 1 shows the number of participants choosing to stay or to switch depending on which door was left unopened. There was a large effect of condition, $\chi^{2}(1)=$ 74.35, $p<.001$.

This result shows that participants were sensitive to the implication of the prior probabilities of the unopened door, with $86 \%$ correctly switching when the unopened door had a high prior probability and $66 \%$ correctly staying when the unopened door had a low prior probability. This result already argues that the equiprobability principle is not being commonly applied.

Table 1: Number of participants in each condition (Door-B in which the high probability door is left unopened and Door-C in which the low probability door is left unopened) deciding to stay or switch.

|  |  | Choice |  |
| :---: | :--- | :--- | :--- |
|  |  | Stay | Switch |
| Door-B <br> unopened | $(60 \%)$ | 16 | 99 |
| Door-C <br> unopened | $(10 \%)$ | 101 | 50 |
|  |  |  |  |

A slightly surprising aspect of Table 1 was how high the switch rate was for the Door-C condition, given that so few participants switch in standard versions of the MHD despite it being the correct response. This may be due to the initial choice of Door A being allocated to participants rather than
being a true choice, which could reduce both the illusion of control (Granberg \& Dorr, 1998) and the anticipation of regret (Gilovich et al, 1995) factors that have been seen as driving the excessive number of "stay" decision in the standard MHD. Consistent with this is that previous studies have found that eliminating participants' first choice increased switch rates substantially (Tubau \& Alonso, 2003).

The large effect of the prior probabilities is also surprising in light of the lack of an effect of unequal prior probabilities observed by Granberg (1999) on his first trial. Possibly this was because the differences in probabilities for Granberg's four doors were quite small: .1, $2, .3$ and .4. It is also possible that because Granberg's first trial was the first of 60 participants were using it to explore rather than thinking deeply about the right choice to make.

## Percent

Participants in the Door-B condition gave higher mean percentage ( $M=.61, S D=.16$ ) chances of winning by switching than did those in the Door-C condition ( $M=.41$, $S D=.21), t(264)=8.85, p<.001$. Consistent with the choice data, participants thought they had a better chance of winning by switching if the higher probability door was left unopened.

Table 2 shows the number of participants giving each percentage response depending of their condition and their choice. Very few participants gave the correct percentages, only $4 / 266$. That these four actually calculated the Bayesian posterior probabilities is thrown into doubt by the observation that two participants generated these responses even when they were incorrect for their condition. Critically, Table 2 reveals that relatively few participant gave the equiprobability response of $50 \%$ when asked how likely they were to win if they switched. In total only 37/266 (14\%) of participants did so. Equiprobable responses were no more likely in the Door-C than the Door-B condition, $\chi^{2}(1)=$ $0.475, p=.47$. As expected based on results from the standard MHD, equiprobability was associated with more "stay" decisions, $\chi^{2}(1)=5.76, p=.016$.

There was strong evidence that participants followed the principle that nothing had changed. A total of 93/266 (35\%) of participants gave responses consistent with them calculating the ratio of the prior probabilities of the opened and unopened door ( $66 \%$ when Door B was unopened; $25 \%$ when Door C was unopened). As expected, the $66 \%$ response was much more common from participants in the Door-B than the Door-C condition, $\chi^{2}(1)=18.9, p<.001$, and the reverse was true for $25 \%$ responses, $\chi^{2}(1)=19.2, p<.001$. There is also evidence of participants using a cruder version of the "no change" principle and simply giving the prior probabilities as the posterior probabilities ( $10 \%$ in Door C; $60 \%$ for Door B). A total of 50/266 (19\%) did this. Again as expected, the $60 \%$ response was much more common from participants in the Door-B than the Door-C condition, $\chi^{2}(1)$ $=31.4, p<.001$, and the reverse was true for $10 \%$ responses, $\chi^{2}(1)=20.1, p<.001$.

Although only generated by 9 participants, the most common percent response that is not apparently associated
with applying equiprobability or a "no change" principles was $70 \%$. This answer could be due to participants adding together the prior probabilities of the opened and unopened doors. Such addition is sometimes proposed as a way to explain the correct answer to the standard version of the MHD. The unequal probability version of the MHD illustrates why this is a poor way to explain the MHD, but it is interesting that a small number of participants appeared to reasoning using it. If we see the 6 participants responding $30 \%$ as some sort of inversion of the $70 \%$ reasoning, this leaves only 10 out of 266 other responses which cannot in some way be linked to the "sum", equiprobability, or "no change" approaches.

Table 2: Frequencies of different values of reported percent chances of winning if participant switched, split by whether the participants decided to stay or switch and their condition
(Door-B or Door-C not opened). Two categories cover ranges of responses rather than precise responses.

| Percent <br> chance <br> of win if <br> switch | Door C <br> $(.10)$ not <br> opened | Door B <br> $(.60)$ not <br> opened | Door C <br> $(.10)$ not <br> opened | Door B <br> $(.60)$ not <br> opened |
| :--- | :--- | :--- | :--- | :--- |
|  | 0 | 0 | 0 | 2 |
| $10 \%$ | 19 | 0 | 5 | 0 |
| $11-24 \%$ | 1 | 0 | 0 | 1 |
| $25 \%$ | 22 | 0 | 1 | 0 |
| $30 \%$ | 2 | 0 | 2 | 1 |
| $33 \%$ | 25 | 2 | 5 | 2 |
| $40 \%$ | 3 | 0 | 0 | 1 |
| $50 \%$ | 16 | 7 | 7 | 7 |
| $60 \%$ | 1 | 3 | 1 | 23 |
| $65-67 \%$ | 8 | 4 | 26 | 52 |
| $70 \%$ | 3 | 0 | 1 | 5 |
| $75 \%$ | 0 | 0 | 0 | 1 |
| $70 \%$ | 0 | 0 | 1 | 1 |
| $700 \%$ | 0 | 0 | 1 | 2 |
| Totals | 101 | 16 | 50 | 99 |

A surprising aspect of the data in Table 2 was that a high number of participants in the Door C condition who switched gave $66 \%$ as the percentage chance of winning. This is consist with their choice but given that the prior probability
of this door concealing the prize was $10 \%$ and its posterior probability was $40 \%$, it is hard to infer the reasoning behind this answer. It may just be due to participants who were confused about the problem. Alternatively it could be that participants who had an intuition to switch, but were unable to calculate a probability fell back to using a plausible sounding percentage for a problem proposing three entities. Consistent with this is the similar number of participants giving $33 \%$ as the answer, particularly when deciding to stay. Another possibility arises from the fact that two-thirds is the correct answer to the standard MHD, so it is possible that some participants who said that they have not seen a similar problem actually had seen the standard MHD and then repeated its solution.

## Cognitive Reflection Task (CRT)

Participants' responses to the CRT were scored by counting how many of the three problems they gave the correct answer to. Overall the participants did quite poorly given that the maximum score is 3.0 but their mean score was 0.55 ( $\mathrm{SD}=0.86$ ).

There appeared to be an association between CRT scores and how participants responded. Those who generated the equiprobability percent of $50 \%$ did not differ from other participants on CRT scores ( $M=0.43, S D=0.83, n=37$ verse $M=0.57, S D=0.86, n=229), t(264)=0.92, p=.36$. However participants repeating the prior probability, either $10 \%$ or 60\% depending on their condition, had lower CRT scores than other participants, $(M=0.30, S D=0.61, n=50$ verse $M=0.61, S D=0.89, n=216), t(264)=2.34, p=.020$. Furthermore participants maintaining the ratio of prior probabilities, either $25 \%$ or $66 \%$ depending on their condition, had higher CRT scores than other participants ( $M=0.72, S D=0.96, n=116$ verse $M=0.42, S D=0.74, n=150$ ), $t(264)=2.92, p=.004$. So it appears that participants' responses to the unequal probabilities MHD reflect not only what principles they applied, but also how reflective was their thinking.

## Discussion

The equiprobability principle has often been presented as a strong driver of the illusion behind the MHD, yet there has been no attempt to test this hypothesis beyond the observation that most participants when given standard versions of the MHD report that that they would have a $50 \%$ chance of winning if they switched. To the extent that participants use similar reasoning when faced with the equal and unequal probability versions of the MHD, the finding here that only $14 \%$ of participants gave equal probability answers throws into doubt that the equiprobability principle is a strong driver of this cognitive illusion. In a way, it is also a powerful demonstration of just how strong the equiprobability principle can be for some individuals, given that it appeared to be followed by some participants even when it not only led to the wrong answer but it had to be invoked out of thin-air (i.e., there are no mentions of $50 \%$ and percentages presented
are unequal). So this principle may explain the illusion of the MHD for some people, but they appear to be a small minority.

The experiment's findings represent strong evidence that what leads participants to think that each remaining door is equally probable after Monty has revealed one is that there has be no change to these doors, so they maintain their equal status. Overall $54 \%$ of participants gave percentages in which they repeated the prior probability or they maintained the ratio of the prior probabilities. This is consistent with Burns and Wieth's (2004) argument that the main obstacle to correctly seeing the need for conditional reasoning about the MHD is failure to recognize that Monty's actions change things beyond just removing an option. Burns and Wieth argued that the MHD is difficult because it has a causal structure that people have difficulty recognizing and thus they fail to see that Monty's actions have causal consequences. We demonstrated this effect by showing higher switch rates when the MHD was presented as a competition. The finding here using the unequal probability MHD show directly that a belief that there has been "no change" could be driving many people's failures to reason correctly about the MHD.

The current results show almost no evidence of Bayesian updating by participants. They are sensitive to prior probabilities but they do not appear to recognize the problem as one involving conditional probabilities. The more reflective thinkers (as measured by CRT scores) recognize that the posterior probabilities are different from the prior probabilities, so have updated their probabilities after a door was opened. However they have only accounted for the missing door, not taken into account the new conditional probabilities.

The unequal probabilities versions of the MHD have the potential to be useful tools for examining the reasoning people use in the MHD because they have the potential to reveal this reasoning through differential answers. Almost all participants gave percent responses which could be interpreted as revealing their underlying reasoning. Questions concerning both the MHD and probabilistic reasoning more generally might be fruitful explored using appropriate versions of the unequal probabilities MHD.

## Does this experiment really address the MHD?

Reviewers of this paper raised some interesting issues regarding whether any experiment using an unequalprobabilities MHD can tell us anything about how people reason in the common equal probabilities version of the MHD.

One claim was that the current experiment had nothing to do with the MHD because participants were told their initial choice and told what door was opened, whereas the common description of the MHD allows people to make their own first choice. However assigning to participants their first choice has been the case in most empirical studies of the MHD going back to the first published study by Granberg and Brown (1995). Doing so has not eliminated the strong bias to decide to stay, so it does not seem to be a critical factor.

The question was also raised as to whether assigning unequal prior probabilities to the doors in itself invalidates the experiment as an examination of reasoning in the MHD. However, that the probabilities have to be equal has not been presented as a critical aspect of the MHD. If it is, then why it is critical needs to be explained. The MHD is a conditional probability problem and Bayes' Law can be as effectively applied to the MHD with unequal probabilities as to a MHD with equal probabilities. From the point of view of conditional probability, it is the equal probabilities presentation that is a special case of a more general problem. The reasoning described as being behind the MHD does not seem to rely on the prior probabilities being equal, although it may simplify the calculations.

An interesting claim made was that the unequal probabilities MHD cannot possibly address the equiprobability principle because it does not have equal prior probabilities. As pointed out earlier the equiprobability principle applies when there is no apparent reason to differentiate the options, thus the presentation of unequal prior probabilities may prevent it from being applied. The probabilities of the unopened doors are also not equal in the equal probabilities version, but people often don't perceive that to be the case. However this interpretation of the principle makes it somewhat circular, that is, it becomes a statement that people judge probabilities to be equal when they don't perceive them as unequal. Such an interpretation limits the explanatory value of the equiprobability principle. (Johnson-Laird et al [1999] also use the equiprobability principle to generating mental models, so perhaps the principle is better seen as a step in reasoning than as a result.) The finding in the current experiment that some participants thought the doors to be equally likely despite the prior probabilities being unequal suggests that a stronger version of the equiporbability principle is used by some people.

## References

Burns, B.D., \& Wieth, M. (2004). The collider principle in causal reasoning: Why the Monty Hall Dilemma is so hard. Journal of Experimental Psychology: General. 133, 434-449.
Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. Cognition, 43, 197-223.
Falk, R., \& Lann, A. (2008).The allure of equality: Uniformity in probabilistic and statistical judgment. Cognitive Psychology, 57, 293-334.
Fischhoff, B., Parker, A. M., Bruine de Bruin, W., Downs, J., Palmgren, C., Dawes, R., \& Manski, C. F. (2000). Teen expectations for significant life events. The public opinion quarterly, 64, 89-205.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19, 25-42.
Granberg, D. (1999). A new version of the Monty Hall Dilemma with unequal probabilities. Behavioural Processes, 48, 25-34

Granberg, D., \& Brown, T. A. (1995). The Monty Hall dilemma. Personality and Social Psychology Bulletin, 21, 711-723.
Granberg, D., \& Dorr, N. (1998). Further exploration of two-stage decision making in the Monty Hall dilemma. American Journal of Psychology, 111, 561-579.
Gilovich, T., Medvec, V. H., \& Chen, S. (1995). Commission, omission, and dissonance reduction: Coping with regret in the "Monty Hall" problem. Personality and Social Psychology Bulletin, 21, 185-190.
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., \& Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62-88.
Krauss, S., \& Wang, X. T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. Journal of Experimental Psychology: General, 132, 3-22.
Piattelli-Palmarini, M. (1994). Inevitable illusions: How mistakes of reason rule our minds (M. Piattelli-Palmarini \& K. Botsford, Trans.). New York: John Wiley \& Sons.
vos Savant, M. (1997). The power of logical thinking. New York: St Martin's Press.
Simon, H. A. (1955). A behavioral model of rational choice. The Quarterly Journal of Economics, 69, 99-118.
Stanovich, K. E. (2011). Rationality and the reflective mind. New York, NY: Oxford University Press.
Stibel, J. M., Dror, I. E., \& Ben-Zeev, T. (2009). The collapsing choice theory: Dissociating choice and judgment in decision making. Theory and Decision, 66, 149-179.
Tubau E., Aguilar-Lleyda, D., \& Johnson. E. D. (2015). Reasoning and choice in the Monty Hall Dilemma (MHD): Implications for improving Bayesian reasoning. Frontiers in Psychology, 6 (Article 353), 1-11.
Tubau, E., \& Alonso, D. (2003). Overcoming illusory inferences in a probabilistic counterintuitive problem: The role of explicit representations. Memory \& Cognition, 31, 596-607.
Tversky, A., \& Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. Science, 185, 11241130.

# Discovering simple heuristics from mental simulation 

Frederick Callaway (fredcallaway@berkeley.edu)<br>Jessica B. Hamrick (jhamrick@berkeley.edu)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology; University of California, Berkeley; Berkeley, CA, USA


#### Abstract

In the history of cognitive science, there have been two competing philosophies regarding how people reason about the world. In one, people rely on rich, generative models to make predictions about a wide range of scenarios; while in the other, people have a large "bag of tricks", idiosyncratic heuristics that tend to work well in practice. In this paper, we suggest that rather than being in opposition to one another, these two ideas complement each other. We argue that people's capacity for mental simulation may support their ability to learn new cuebased heuristics, and demonstrate this phenomenon in two experiments. However, our results also indicate that participants are far less likely to learn a heuristic when there is no logical or explicitly conveyed relationship between the cue and the relevant outcome. Furthermore, simulation-while a potentially useful tool-is no substitute for real world experience.


Keywords: mental simulation, heuristics, physical reasoning

## Introduction

The world is a complex place, yet people are able to navigate it effortlessly. How is the mind able to do so much? One answer is that the mind builds rich, generative models of the world (Tenenbaum, Kemp, Griffiths, \& Goodman, 2011), which it then uses to "mentally simulate" potential futures and make inferences about objects and scenes. Indeed, there is a vast literature on how mental simulation underlies our core reasoning and problem solving abilities, including spatial reasoning (Hegarty, 2004; Shepard \& Metzler, 1971), physical scene understanding (Battaglia, Hamrick, \& Tenenbaum, 2013; Smith \& Vul, 2013), counterfactual reasoning (Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2014), and language comprehension (Bergen, Lindsay, Matlock, \& Narayanan, 2007; Matlock, 2004). Yet, despite the power and flexibility of mental simulation, there is a cost associated with its use: running simulations and evaluating their results takes time and resources. An alternative is to rely instead on simple heuristics that usually point to a good answer (Gigerenzer \& Todd, 1999). But, where do such heuristics come from in the first place?

Previous research has explored the notion of "learning by thinking" (Lombrozo, in press), demonstrating that people have the ability to learn new knowledge or re-represent old knowledge through internal processes such as simulation. For example, Hamrick, Battaglia, Griffiths, and Tenenbaum (2016) showed how people can use their mental simulations to learn about unobservable properties of the world such as the mass of objects; Khemlani, Mackiewicz, Bucciarelli, and Johnson-Laird (2013) illustrated how mental simulations can give rise to algorithmic problem-solving procedures; and Schwartz and Black (1996) demonstrated that people can learn simple rules about a physical system on the basis
of mental simulation. Thus, it is clear that people can acquire new knowledge or heuristics from mental simulation; but, under what circumstances will they do so?

In this paper, we propose that generative models can bootstrap the discovery of heuristics for novel tasks, but that people's prior biases strongly influence how likely they are to discover such heuristics. We pose three key questions regarding this claim. First, to what extent are people able to learn new information from their mental simulations? Second, to what extent do people use this information to construct new heuristics? And third, is mental simulation as reliable as realworld experience in learning such heuristics?

To determine how people learn heuristics from mental simulation, we designed and ran two experiments adapted from Hamrick, Smith, Griffiths, and Vul (2015) in which participants predict whether or not a ball would go through a hole based on its initial trajectory (see Figure 1). Importantly, we also manipulated an environmental cue-the color of the box containing the ball-that perfectly predicted the correct response. In the first experiment, we primed participants with the knowledge that a simple rule existed (but did not tell them the rule itself); in the second, we primed them with either weak expectations or no expectations, and then allowed them to do the task and discover the rule independently. Our results show that people are capable of crystallizing new rules solely on the basis of their mental simulations, though they are significantly less likely to do so if they are not already entertaining the hypothesis that a rule exists. Moreover, we show that mental simulation, while an avenue for learning such rules, is no substitute for real world experience.

## Experiment 1: Learning about known cues

In our first experiment, we asked to what extent people are able to learn heuristics from mental simulation when they are aware such a heuristic might exist. The heuristic took the form of an associative cue (see Stimuli) that perfectly predicted the correct response and that did not require mental simulation.

## Methods

Participants We recruited 119 participants on Amazon's Mechanical Turk using the psiTurk experimental framework (Gureckis et al., 2015). Participants were paid $\$ 1.50$ for roughly 14 minutes of work. We excluded 9 participants who did not finish the experiment and 8 participants who answered incorrectly on more than one catch trial. This left a total of 102 participants in our analysis.


Figure 1: Example of a medium trial.

Design We used a $3 \times 3 \times 2$ mixed design. We manipulated two within-subject variables, CUE and DIFFICULTY. CUE could take on three values: honest (the cue perfectly predicts the correct response), neutral (the cue contains no information), and deceitful (the cue predicts the incorrect response). DIFFICULTY could take on three values as well: easy, medium, and hard (see Stimuli). We manipulated one between-subjects variable, FEEDBACK, which determined whether people were allowed to see the full path of the ball (and thus the correct answer) after making a judgment.

Stimuli The stimuli were animations of a ball moving at a $400 \mathrm{px} / \mathrm{s}$ in a box with dimensions $900 \times 650 \mathrm{px}$. As the ball moved, it traced a gray line to reduce uncertainty about its direction. The initial stimulus presentation consisted of the ball moving for 0.2 seconds, after which the ball would freeze, remaining on screen along with its trace. The feedback animation picked up where the initial stimulus presentation left off, and showed the ball bouncing some number of times and then either (1) passing through the hole (a hit); or (2) bouncing off the central wall (a miss). The properties of a stimulus depended on the trial's difficulty. Easy stimuli had one bounce, a path length of 560 px , and a hole size of 300 px . Medium stimuli had one or two bounces, a path length of 880 px , and a hole size of 200 px. Finally, hard stimuli had two bounces, a path length of 1280 px , and a hole size of 100 px . The color of the background could be blue, green, or yellow depending on both the correct response and the value of CUE for that trial. For each participant, the three colors were mapped to hit, miss, and neutral (this mapping was counterbalanced). Thus, on an honest trial, the background would take the hit color if the ball would go through the hole, and the miss color, otherwise. This mapping was reversed for deceitful trials. Finally, on a neutral trial, the background was always the neutral color.

Procedure Participants were first given instructions in which the task was described. We specifically informed participants that they would observe three people playing a game on three different courts: "Player B" was playing on a blue court, "Player G" was playing on a green court, and "Player Y" was playing on a yellow court. We additionally told par-


Figure 2: Trial structure. The experiment begins with a block of eight "instruction" trials, shown with feedback regardless of condition. No cue is present. This is followed by nine twelve-trial "standard" blocks of increasing difficulty. Fiftyfour unique stimuli are each shown twice (in separate blocks), once with an honest cue and once with a neutral cue. Feedback is displayed on all or no trials depending on condition. At the end of each standard block, all participants saw their accuracy from the preceding block and responded to the cue quiz (see text). The final, "critical" block contains fourteen trials, shown without feedback. Trials with deceitful cues are interspersed to minimize the chance of participants noticing the change in cue reliability.
ticipants that one of the players was playing a game in which they were trying to get the ball in the hole, one was playing a game in which they were trying to avoid the hole, and one was playing a game in which they didn't care whether or not it went in. This backstory was designed to increase participants' subjective prior probability of and attention to the hypothesis that the background color was predictive of the correct response. Crucially, however, the backstory only motivated the existence of such a predictive relationship; it does not indicate its direction.

On each trial, participants were shown the scene, including the initial position of the ball and the location of the hole. Participants pressed 'space' to begin the trial, after which an animation of the initial stimulus began. Participants were then asked, "will the ball go in the hole?", and were instructed to press ' $q$ ' if they thought it would and ' $p$ ' otherwise. Participants in the feedback condition then saw "Correct!" or "Incorrect" as well as an animation showing the full remaining trajectory of the ball.

The structure of the experiment is shown in Figure 2. The early trials were easy so that participants in the no feedback condition had the chance to learn the cue when their simulation-based judgments were more reliable. The later trials were hard so they would be discriminative of participants' strategies: participants using simulation should perform poorly, while participants using the cue should be insensitive to trial difficulty. To assess declarative knowledge of the cue, we asked participants three multiple choice questions (the cue quiz) after each standard block: "Which player


Figure 3: Accuracy on critical and standard trials in Experiment 1. Error bars in all figures denote $95 \%$ confidence intervals by bootstrapping. (a) Critical trials were displayed either with an honest cue or with a deceitful cue. Participants performed better when the cue is honest, suggesting that they were relying on the cue rather than using simulation. (b) On standard trials, participants tended to respond more accurately on honest-cue trials than on neutral-cue trials.
is trying to get the ball into the hole?","Which player is trying to avoid the hole?", and "How confident are you in your response to the previous two questions?".

The last (critical) block was designed to provide evidence that participants were using the cue, as those using the cue should answer incorrectly on trials with the deceitful cue. The stimuli used for the first four honest trials and the deceitful trials were counterbalanced, and we excluded all honest trials after the first deceitful trial from analysis.

## Results

All analyses were planned unless specifically stated otherwise, and all contrasts are adjusted for multiple comparisons. All data and analysis code can be viewed at https://osf.io/ut3xp/
Hypotheses Based on our experimental design, we hypothesized the following: (1) Participants in both the no feedback and feedback conditions will learn the cue, as determined by their responses to the cue quizzes and based on their responses on the critical trials. (2) Participants in both the no feedback and feedback conditions will use their knowledge of the cue to respond more accurately in the task. (3) Participants in the feedback condition will be more likely to learn and use the cue than participants in the no feedback condition.
Cue quizzes To gauge whether a participant had successfully learned the cue after seeing all standard trials, we restricted our analysis to the final cue quiz. We conducted three one-tailed proportion tests comparing the proportion of participants in each condition who answered both questions correctly on the quiz, with a chance probability of $\frac{1}{6}$. We found that $56 \%$ of participants in the no feedback condition $\left(\chi^{2}(1)=54.465, p<0.001\right)$ and $68 \%$ of participants in the feedback condition $\left(\chi^{2}(1)=91.204, p<0.001\right)$ correctly identified the cue. These results suggest that participants were able to use their simulations to learn about the cue, confirming our first hypothesis.
Critical trials According to our first hypothesis, we anticipated that participants who learned and used the cue strat-
egy would fail on the four critical trials in which the cue was misinformative. We constructed a logistic regression model over accuracy on critical trials with factors for FEEDBACK and CUE. The results suggest that people in both the feedback and no feedback conditions were more likely to answer incorrectly on deceitful trials than on honest trials (Figure 3). Specifically, we found a significant main effect of CUE, with participants responding more accurately on honest trials than on deceitful trials ( $\chi^{2}(1)=4.252, p<0.05$ ). We also found a significant main effect of FEEDBACK $\left(\chi^{2}(1)=17.124, p<\right.$ 0.001 ), as well as an interaction between FEEDBACK and CUE $\left(\chi^{2}(1)=40.019, p<0.001\right)$. In both feedback conditions people were more likely to answer incorrectly on deceitful trials than on honest trials, though this difference was only marginally significant in the no feedback condition (for feedback, $\mathrm{LLR}=-2.31 \pm 0.23, z=-9.85, p<0.001$; for no feedback, LLR $=-0.41 \pm 0.20, z=-2.06, p=0.08$; where LLR is the log likelihood ratio).

The weak effect of cue honesty for the no feedback condition could be due to either an inability to identify the cue, or an inability to use knowledge of the cue to make predictions. To test these explanations, we conducted a posthoc analysis identical to that above but restricting the data to those participants that passed the quiz. We found highly significant effects of $\operatorname{CUE}\left(\chi^{2}(1)=10.862, p<0.001\right)$, FEEDВАСК $\left(\chi^{2}(1)=30.657, p<0.001\right)\left(\chi^{2}(1)=17.124, p<\right.$ 0.001 ), and the interaction between FEEDBACK and CUE $\left(\chi^{2}(1)=59.828, p<0.001\right)$. Contrasts revealed a significant effect of honesty in both the feedback (LLR $=-4.33 \pm$ $0.40, z=-10.85, p<0.001$ ) and no feedback (LLR $=$ $-0.88 \pm 0.27, z=-3.26, p<0.01)$ conditions. These results suggest that participants in the no feedback condition who identified the cue were also able to use it to make predictions, but not as well as those in the feedback condition.

Standard trials We also looked at the accuracy across trials during the main part of the experiment. We constructed a logistic regression model over accuracy with factors for FEEDBACK, DIFFICULTY, and CUE. The results are


Figure 4: Example cue learners. Each subplot shows a different participant in the no feedback condition of Experiment 1 who was identified as learning the cue by our model. The blue lines indicate average accuracy on each block when using the honest cue, while the gray lines correspond to the neutral cue. The vertical red lines indicate the trial, $C$, when our model inferred they switched from using simulation to using the cue. The title of each subplot displays the log likelihood ratio of the change model to the no change model, as well as the trial when they changed strategies.
shown in Figure 3. We found a main effect of difficulty $\left(\chi^{2}(2)=122.679, p<0.001\right)$, as well as a three-way interaction between FEEDBACK, DIFFICULTY, and CUE $\left(\chi^{2}(2)=\right.$ 12.363, $p<0.01$ ).

We investigated differences in accuracy within feedback conditions and cue types and found that, overall, people were more accurate on honest trials when they had feedback than when they did not have feedback (LLR $=-0.51 \pm 0.07, z=$ $-7.40, p<0.001$ ). This supports our third hypothesis that real data is more reliable than simulated data. We did not detect a difference between feedback conditions on neutral trials, however, indicating that feedback did not affect people's accuracy when using simulation (LLR $=0.03 \pm 0.06, z=$ $0.50, p=1.00$ ). We also found that people were more accurate when the honest cue was present than when the neutral cue was present, both in the feedback condition (LLR $=$ $-0.83 \pm 0.07, z=-12.24, p<0.001)$ and the no feedback condition $(\operatorname{LLR}=-0.29 \pm 0.06, z=-4.57, p<0.001)$.

## Modeling individual differences in cue learning

While the group-level effects in the previous sections confirmed our first and third hypotheses, we wanted to additionally investigate the individual behavior of participants who learned the cue. To this effect, we constructed a simple Markov model that allowed us to identify who actually used the cue and who did not.

Model For each participant, we defined a Markov model with observed states $J_{t}$ representing the participants' judgment on trial $t$. For each strategy, we defined a probability of answering correctly. For the simulation probability, we
fit $p_{\text {sim }}^{(\text {easy })}, p_{\text {sim }}^{(\text {med })}$, and $p_{\text {sim }}^{(\text {hard })}$ empirically based on the participant's average accuracy on trials without the cue for each level of difficulty. For the cue probability, we set $p_{\text {cue }}=0.95$ to reflect a high probability of answering correctly, but not perfectly. Finally, we introduced a variable $C \in\{1, \ldots, T\}$ which indicated the "change point" at which participants switched from using simulation to using the cue heuristic.

The probability of a participant's judgment was then:

$$
p\left(J_{t}=1 \mid C\right)= \begin{cases}p_{\mathrm{sim}}^{\left(d_{t}\right)} & t \leq C \\ p_{\text {cue }} & t>C\end{cases}
$$

where $d_{t}$ is difficulty of trial $t$. So, the probability of all responses was $\max _{C} p\left(\mathbf{J}_{1: T} \mid C\right)=\max _{C} \prod_{t=1}^{T} p\left(J_{t} \mid C\right)$, which we will refer to as the change model. We fit $C$ in the change model to each participant separately.

We additionally computed the likelihood of participants' responses under a no change model, in which we computed $p\left(\mathbf{J}_{1: T} \mid C=\infty\right)=\prod_{t=1}^{T} p\left(J_{t} \mid C=\infty\right)$, where the infinite change point $C$ indicates that the participant used the simulation strategy throughout the whole experiment.

Results To determine whether an individual participant learned the cue, we computed the log-likelihood ratio (LLR) between the change model and the null hypothesis (the no change model), and tested whether $2 \cdot$ LLR was significantly greater than zero under the $\chi^{2}$ distribution, with a significance threshold of $p=0.001$. Using this analysis, we found that 29 participants in the feedback condition switched to a cue-based strategy while 8 participants in the no feedback condition switched. To ensure these numbers were more than we would expect due to random chance, we additionally performed proportion tests with a probability of chance at 0.001 (corresponding to the significance threshold above). Both proportions were significantly different from chance (for feedback, $\chi^{2}(1)=16204, p<0.001$; for no feedback, $\left.\chi^{2}(1)=1068, p<0.001\right)$. The difference in proportions was also significant $\left(\chi^{2}(2)=17272, p<0.001\right)$

Figure 4 shows the two participants in the nofeedback condition with the highest log-likelihood ratios, and illustrates the clear effect of the cue: on the honest trials, the participants have nearly perfect performance, while on the neutral trials, they are significantly worse.

We additionally looked at the overlap between those participants who correctly answered the cue quiz and those who were identified by our model. The results, shown in Table 1, indicate that those people who were identified by the model answered correctly on the quiz, but not necessarily the other way around. This suggests that, counter to our second hypothesis, not everybody who explicitly identifies the cue is able to apply that knowledge when performing the task.

## Experiment 2: Discovering new heuristics

Based on the results of Experiment 1, it is clear that some people are able to use mental simulation to learn a cue-based heuristic-as long as they know that such a cue exists. In

Table 1: Number of participants identified by the quiz and/or model as having learned the cue in Experiments 1 and 2. FB = "Feedback", No FB = "No Feedback".

|  | Condition | Neither | Quiz <br> Only | Model <br> Only | Both |
| :--- | :--- | :---: | :---: | :---: | :---: |
| 1 | No $F B(52)$ | 20 | 24 | 3 | 5 |
|  | $F B(50)$ | 15 | 6 | 1 | 28 |
| 2A | No $F B(48)$ | 38 | 10 | 0 | 0 |
|  | $F B(49)$ | 31 | 14 | 1 | 3 |
| 2B | No $F B(52)$ | 41 | 11 | 0 | 0 |
|  | $F B(49)$ | 38 | 9 | 0 | 2 |

Experiment 2, we asked whether participants could discover and learn the heuristic without being given this information explicitly. By making two small alterations to the backstory presented in Experiment 1, we modulated the degree to which participants would expect the cue. Experiment 2A did not inform participants that the colors were predictive; it only associated the cue (color) with players. We hypothesized that this would allow participants to frame hypotheses about cue predictiveness in terms of more familiar concepts: one player might be more talented or have a different goal. Additionally, describing the colors in the instructions might increase their salience. In Experiment 2B, we did not verbally draw attention to the cue, nor did we provide any semantic meaning for the cue. Thus we expected participants would be even less likely to learn the cue, perhaps because they would not even consider the hypothesis that the colors are predictive.

## Methods

Participants We recruited 224 participants on Amazon's Mechanical Turk using the psiTurk experimental framework (Gureckis et al., 2015). Participants were treated in accordance with UC Berkeley IRB standards and were paid $\$ 1.50$ for fourteen minutes of work. We excluded 15 participants who did not finish the experiment and 11 participants who answered incorrectly on more than one catch trial. This left a total of 198 participants in our analysis.

Design and Procedure The design and procedure were identical to Experiment 1, with the following exceptions. In Experiment 2A we told participants that there were three different players, corresponding to three different colors, but not that they were playing different games. In Experiment 2B we gave participants a minimal backstory that made no reference to players or colors. In both experiments we administered the cue quiz once, at the end of the experiment.

## Results

Experiment 2A We performed the same analyses as in Experiment 1, and found that $35 \%$ of people in the feedback condition were able to identify the cue in the quiz $\left(\chi^{2}(1)=10.204, p<0.001\right)$. Without feedback, $21 \%$ of par-
ticipants identified the cue, which was marginally significant $\left(\chi^{2}(1)=1.680, p=0.10\right)$. We did not find an affect of CUE in the critical trials $\left(\chi^{2}(1)=0.518, p=0.47\right)$, though there was a trend towards people being more accurate on honest trials. We found no significant effect of the cue on accuracy in standard trials either $\left(\chi^{2}(1)=0.160, p=0.69\right)$.

The Markov model identified 4 people in the feedback condition $\left(\chi^{2}(1)=243, p<0.001\right)$ and 0 in the no feedback condition as having adopted the cue strategy. Together, these results show that when people are primed with a cover story that makes the cue plausible, some of them will indeed learn the cue; however, the majority still will not.
Experiment 2B Whereas in Experiment 2A the cue was explained with a cover story about people playing a game, in Experiment 2B the cue was entirely unexplained. The results suggest that when the cue is unexplained, participants are unlikely to discover the informativeness of the cue. We again performed the same analyses as those in Experiment 1, and found that people were not significantly different from chance at identifying the cue in the survey, regardless of whether they saw feedback ( $21 \%$ of participants, $\chi^{2}(1)=0.800, p=0.19$ ) or not ( $22 \%$ of participants, $\chi^{2}(1)=0.465, p=0.25$ ). We also found no effect of CUE in the critical trials $\left(\chi^{2}(1)=\right.$ 1.626, $p=0.20$ ), though as in Experiment 2A there was a trend toward people being more accurate on the honest trials. Again, we found no significant effect of the cue on accuracy in standard trials $\left(\chi^{2}(1)=1.269, p=0.26\right)$.

The Markov model identified 2 people in the feedback condition $\left(\chi^{2}(1)=43, p<0.001\right)$ and 0 in the no feedback condition as having adopted the cue strategy. These results suggest that when people are not already entertaining the hypothesis that a heuristic might exist, it is unlikely that they will spontaneously realize it.
Comparing Experiments Summary results of the three experiments are shown in Table 1 and Figure 5. We consistently find more evidence for cue-learning when feedback is given. However, our results suggest that an unexplained cue that has no intuitive relationship with the outcome is quite difficult to learn, even when feedback is present.

## Conclusion

In this work, we asked three questions: (1) are people able to learn about auxiliary properties in the world through mental simulation; (2) do they use their knowledge to make more accurate predictions; and (3) is mental simulation as reliable as real-world experience? In Experiment 1, we showed that (1) people can indeed learn a correlated cue through the use of mental simulation; and (2) people can sometimes apply such knowledge as a heuristic prediction strategy. However, (3) both discovery and application of the cue was weaker when people had to learn from only simulated data. We speculate on two potential explanations for the advantage of external over simulated data. First, simulations are noisy; thus, simulated data may not accurately reflect the world. In this study,


Figure 5: Comparing quiz, accuracy, and model results across experiments. (a) Participants who correctly identified the cue during the cue quiz. The dashed line indicates chance performance. (b) The difference in accuracy on the honest trials versus the neutral trials. Positive values indicate that participants were more accurate on honest trials. (c) The proportion of participants identified as having learned the cue by the Markov model. There are no error bars due to the particulars of this analysis; all non-zero proportions are significantly different from chance.
the cue was perfectly predictive; however, if one predicted incorrectly on $25 \%$ of trials, the cue would only $75 \%$ predictive. Furthermore, if people are aware that their simulations are error-prone, they may place less faith in the simulated data and any patterns therein. Second, simulations are costly, and it is possible that increased attentional and working memory load may have decreased participants' ability to simultaneously perform the task and pick up on the cue.

In both experiments, there was considerable withincondition variance in cue learning and use. The alignment between accuracy on the quiz and the Markov model predictions (Table 1) suggests that this is partly due to individual differences. It appears that some participants learned and applied the cue, while others completely ignored the cue. This between-subject variance could be due to true individual differences: perhaps some people are better able to learn associative cues (in general or specifically from simulated data). Alternatively, these differences could be the result of a constant learning ability that is stochastic and only occasionally expressed. Similar to flashes of intuition that strike seemingly at random, identifying a pattern in simulated data may be a powerful but rare event in human cognition.

Together, our results suggest that mental simulation on its own is not sufficient for learning: prior expectations are hugely important. This result is consistent with the ideas behind the hypothesis of theory-based causal induction (Tenenbaum, Griffiths, \& Kemp, 2006), which posits that inductive reasoning requires highly structured and systematic systems of causal knowledge. While it was possible for participants in our experiments to learn a new piece of causal knowledge (a heuristic), it was very difficult for them to do so if the cue did not easily fit into an existing causal framework. Thus, we suggest that while mental simulation can be a powerful tool for re-representing knowledge, it does not operate in a vacuum, and must work in tandem with other cognitive processes to fully realize its potential.

Acknowledgements This work was supported by grant number ONR MURI N00014-13-1-0341.

## References

Battaglia, P. W., Hamrick, J. B., \& Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45), 18327-18332.
Bergen, B. K., Lindsay, S., Matlock, T., \& Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. Cognitive Science, 31(5), 733-64.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2014). From Counterfactual Simulation to Causal Judgment. In Proceedings of the 36th Annual Meeting of the Cognitive Science Society.
Gigerenzer, G., \& Todd, P. M. (1999). Simple heuristics that make us smart. Oxford University Press, USA.
Gureckis, T. M., Martin, J. B., McDonnell, J. V., Alexander, R. S., Markant, D. B., Coenen, A., et al. (2015). psiTurk: An opensource framework for conducting replicable behavioral experiments online. Behavioral Research Methods, 2-16.
Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., \& Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. Cognition, 157, 61-76.
Hamrick, J. B., Smith, K. A., Griffiths, T. L., \& Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In Proceedings of the 37th Annual Meeting of the Cognitive Science Society.
Hegarty, M. (2004). Mechanical reasoning by mental simulation. Trends in Cognitive Sciences, 8(6), 280-285.
Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., \& Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. Proceedings of the National Academy of Sciences of the United States of America, 110(42), 16766-71.
Lombrozo, T. (in press). "Learning by thinking" in science and everyday life. In P. Godfrey-Smith \& A. Levy (Eds.), The Scientific Imagination. Oxford University Press.
Matlock, T. (2004). Fictive motion as cognitive simulation. Memory \& Cognition, 32(8), 1389-1400.
Schwartz, D. L., \& Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. Cognitive Science, 20(4), 457-497.
Shepard, R. N., \& Metzler, J. (1971). Mental Rotation of ThreeDimensional Objects. Science, 171(3972), 701-703.
Smith, K. A., \& Vul, E. (2013). Sources of uncertainty in intuitive physics. Topics in Cognitive Science, 5(1), 185-199.
Tenenbaum, J. B., Griffiths, T. L., \& Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences, 10(7), 309-318.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279-1285.

# Fast and Easy: Approximating Uniform Information Density in Language Production 

Jesús Calvillo (jesusc@coli.uni-saarland.de)<br>Saarland University, Germany


#### Abstract

A model of sentence production is presented, which implements a strategy that produces sentences with more uniform surprisal profiles, as compared to other strategies, and in accordance to the Uniform Information Density Hypothesis (Jaeger, 2006; Levy \& Jaeger, 2007). The model operates at the algorithmic level combining information concerning word probabilities and sentence lengths, representing a first attempt to model UID as resulting from underlying factors during language production. The sentences produced by this model showed indeed the expected tendency, having more uniform surprisal profiles and lower average word surprisal, in comparison to other production strategies.


Keywords: information density; sentence production; rational analysis; connectionist; semantics

## Introduction

For a given semantics, humans are able to produce a large number of surface representations that express its meaning. However, some constructions are preferred over others, some sentences are easier to understand, while some others are more difficult, so people tend to avoid them.

Uniform Information Density Hypothesis (UID, Jaeger, 2010; Levy \& Jaeger, 2007) presents one way to rank sentences according to how uniform their surprisal profiles are; where a sentence is preferred if the surprisal of each of its words remains uniform. This is explained as a rational strategy of language production at the computational level of analysis, as such strategy maximizes the probability of successful communication in a bandwidth-limited noisy channel while maximizing information transmission. Alternatively, and without the assumption of a noisy channel, comprehension effort is also minimized utilizing a UID strategy (Levy \& Jaeger, 2007), provided that the effect of surprisal on comprehension effort is superlinear (Hale, 2001; Levy, 2008).

Empirical evidence supports this hypothesis (e.g., Aylett \& Turk, 2004; Bell et al., 2003), however, as far as one can tell, no modeling attempts explore this at the algorithmic or implementational levels. Here, a mechanistic account of sentence production is presented, which balances on the one hand speed of information transmission and on the other hand comprehension and production effort. The sentences produced by this strategy present more uniform surprisal profiles, compared to other strategies, and thus, represent a first approximation to UID.

In particular, the model assumes that speakers act under three different pressures: a first one, pushing speakers to be fast under time restrictions; a second one, related to production effort, pushing speakers to produce available content first (see Ferreira \& Dell, 2000); and a third one, related to comprehension effort, pushing speakers to avoid high information
density structures. Here I present a way to balance these pressures in order to obtain sentences with more uniform surprisal profiles, which could be later linked to a bandwidth-limited communication channel.

The language production model proposed here extends the one presented by Calvillo, Brouwer, and Crocker (2016), which produces sentences describing a given semantics by maximizing word probabilities. The semantic representations used are a variation of those defined by the Distributed Situation Space model (DSS, Frank, Koppen, Noordman, \& Vonk, 2003; Frank, Haselager, \& van Rooij, 2009). The rest of this section briefly presents the DSS model as well as the model described by Calvillo et al. (2016).

## Distributed Situation Space

The DSS model (Frank et al., 2003, 2009) defines a microworld in terms of a finite set of basic events (e.g., play(charlie,chess)) -the smallest meaning-discerning units of propositional meaning in that world. Basic events can be conjoined to form complex events (e.g., play(charlie, chess) $\wedge$ win(charlie)). However, the microworld poses both hard and probabilistic constraints on event co-occurrence; as a result, some complex events are very common, and some others impossible to happen.

A situation-state space is a large set of $m$ microworld observations defined in terms of $n$ basic events, yielding an $m \times n$ matrix (see Table 1). Each observation in this matrix is encoded by setting basic events that are the case in the given observation to 1 (True) and those that are not to 0 (False). This matrix is constructed by sampling $m$ observations such that no observation violates any hard world knowledge constraint, and such that the $m$ observations approximate the probabilistic nature of the microworld. The resulting matrix encodes then all knowledge about the microworld, where each column, also called situation vector, represents the meaning of each basic event in terms of the observations in which the basic event is true.

Frank et al. (2009) successfully used these DSS representations in a connectionist comprehension model. They defined a microworld consisting of 44 basic events centered around three people. Then they constructed a situation-state space by sampling 25,000 observations. As an example, in this space the situation vector for play(charlie, chess) would correspond to a column in the matrix, where each dimension corresponds to one observation, and its value would be 1 if Charlie is playing chess in that observation. Finally, they reduced the dimensionality of the resulting $25 k$-dimensional situation vectors to 150 dimensions using a competitive layer algorithm.

Table 1: Situation-state space.

|  |  |  |  | $\ldots$ | \# |
| :---: | :---: | :---: | :---: | :---: | :---: |
| observation $_{1}$ | 1 | 0 | 0 | $\ldots$ | 1 |
| observation $_{2}$ | 0 | 1 | 1 | $\ldots$ | 1 |
| observation $_{3}$ | 1 | 1 | 0 | $\ldots$ | 0 |
| ... | . | . | . | .. | . |
| observation $_{m}$ | 0 | 1 | 0 |  | 0 |

## DSS Language Production

DSS representations were also used by Calvillo et al. (2016) in a connectionist model of language production, showing that they are suitable for modeling production.

While Calvillo et al. (2016) used the same microworld as Frank et al. (2009), the DSS representations were modified in order to avoid the competitive layer dimensionality reduction. Instead, the original $25-\mathrm{k}$ dimensional situation vectors were converted to belief vectors. Each dimension of the latter is equal to the conditional probability of each basic event given the original 25 k -dimensional DSS representation that is associated to each sentence. ${ }^{1}$ The result is a 44-dimensional vector that avoids the loss of information associated to the competitive layer algorithm, and consequently renders a higher performance in a language production task.

The architecture of the model presented by Calvillo et al. (2016), represented by the dotted rectangle in Figure 1, implements an extension of a Simple Recurrent Network (Elman, 1990) with a 45 -unit input layer, a 120 -unit recurrent hidden (htan) layer, and a 43 unit (softmax) output layer. The input layer contains 44 units corresponding to the 44 basic events in the microworld, plus one binary unit indicating whether the model must output an active sentence (1), or a passive one (0). The output layer contains 43 units matching the number of available words in the vocabulary.

Time in the model is discrete. At each time step $t$, the recurrent hidden layer receives as input the DSS representation, its own activation at time step $t-1$ (zeros at $t=0$ ) and the identity of the word that was produced at time step $t-1$ (ze$\operatorname{ros}$ at $t=0$ ). Activation of the hidden layer is then propagated to the softmax output layer.

The activation of the output layer yields a probability distribution over the available words, where the word produced at time-step $t$ is defined as the one with highest probability (highest activation). Production stops after an end-ofsentence marker has been produced.

[^293]The identity of the word that was produced at time-step $t-1$ is forwarded to the hidden layer through monitoring units connecting the output layer to the hidden layer, where only the output unit of the word produced at time-step $t-1$ is activated (set to 1 ), while all other units are set to 0 .

Finally, the hidden and output layers also receive input from a bias unit with a constant activation of 1 .

## UID Model

The here proposed model architecture, shown in Figure 1, consists of two paths of processing: the first one (above, inside the dotted rectangle), computes word probabilities given the context, and is identical to the model of Calvillo et al. (2016); and the second one (below), receives the output of the former and computes derivation length estimations, i.e., how long a sentence can be if a particular word is produced. We call probabilities the layer containing the output of the first path, and der_lengths the layer containing the output of the second path.

The output of these two paths is then combined in a final layer (words) that receives unmodified copies of the activation of probabilities and der_lengths and whose activation is a combination of these two types of information. At this point the model produces the word with the highest activation in words, whose identity is then passed to the first hidden recurrent layer through monitoring units in order to process the next word production. Finally, production stops when an end-of-sentence marker is produced.

The rest of this section presents in more detail each of these parts, along with their justification.


Figure 1: UID Production Model.

## Semantic and Linguistic Information

The information content or surprisal of a sentence $s$ is defined as its negative $\log$ probability $-\log P(s)$. Moreover, sentences express events in the world, such that a sentence can be paired with one or more events, and vice versa. Therefore, we can decompose the probability of a sentence $s$ into:

$$
P(s)=\sum_{i} P\left(s \mid e_{i}\right) P\left(e_{i}\right)
$$

where $e_{i}$ is an event in the world that is paired with $s$.
From this, we can distinguish two kinds of information: $P\left(e_{i}\right)$, related to each event that can be paired with the sentence; and $P\left(s \mid e_{i}\right)$, related to the linguistic elements used in this particular sentence to express $e_{i}$.

We call the first one semantic surprisal, and the second one linguistic surprisal (cf. Frank \& Vigliocco, 2011). Semantic surprisal represents how unexpected the events conveyed by the sentence are. Linguistic surprisal can be seen as the information that the sentence conveys, given that the semantics is already known; thus, it is not information about the world, but about the sentence itself.

These two types of information cannot be easily disentangled because they are embedded in each sentence/event. Knowing the identity of an event gives information about the possible related sentences, and vice versa. Nonetheless, based on our definition, we can express total semantic surprisal of a sentence $s$ as:

$$
\operatorname{SemSurp}(s)=-\log \sum_{i} P\left(e_{i}\right)
$$

where $e_{i}$ is each event that can be expressed by $s$.
While one sentence can be paired with several events, normally when a speaker produces a sentence, he/she has one specific event in mind $e_{\alpha}$. Thus, while total semantic surprisal is as described above, the semantic information/surprisal that the speaker is trying to communicate is only:

$$
-\log P\left(e_{\alpha}\right)
$$

As a result, the relevant information associated with a specific sentence $s$ assuming that the speaker is trying to communicate the event $e_{\alpha}$ is given by:

$$
\begin{gathered}
\operatorname{Surp}_{e_{\alpha}}(s)=-\log P\left(s \mid e_{\alpha}\right) P\left(e_{\alpha}\right) \\
\quad=-\log P\left(s \mid e_{\alpha}\right)-\log P\left(e_{\alpha}\right)
\end{gathered}
$$

where the semantic information $-\log P\left(e_{\alpha}\right)$ remains constant across all different surface realizations that could convey it; in contrast to the linguistic information $-\log P\left(s \mid e_{\alpha}\right)$, which can vary widely depending on the specific syntactic structures or words that the speaker chooses.

## Being Easy to Produce

Surprisal Theory (Hale, 2001; Levy, 2008) states that the cognitive effort associated to the processing of a word is proportional to its surprisal. Evidence supporting this has been shown for comprehension (e.g., Hale, 2001; Levy, 2008), and production (e.g., Griffin \& Bock, 1998). Therefore, one can assume that a rational model of production would try to minimize effort for both interlocutors.

While comprehension effort is minimized following a UID strategy, production effort can be minimized by following an Availability Based Production strategy (ABP, Ferreira \& Dell, 2000), where items are produced as they are available.

In this respect, producing the most probable word, and therefore most available, at each time step minimizes (to some extent) production effort by locally minimizing linguistic surprisal:

$$
w_{t+1}=\underset{w}{\arg \min }-\log P\left(w \mid D S S, w_{0}, \ldots, w_{t}\right)
$$

where $w$ is a word in the vocabulary and DSS is the semantic representation related to $e_{\alpha}$. This is already implemented by the model described by Calvillo et al. (2016), where the word produced at each time step is the one with highest conditional probability given the semantics and the previously produced words. In our model these probabilities are obtained at the Probabilities layer in Figure 1.

## Being Fast

The information contained by a sentence results from the sum of the information contained by each of its words. Thus, knowing that the semantic surprisal related to $e_{\alpha}$ should sum up to $-\log P\left(e_{\alpha}\right)$, and that this information is distributed among the words in the sentence, we can calculate average word semantic information/surprisal with respect to $e_{\alpha}$ :

$$
E\left[\text { WordSemSurp }_{e_{\alpha}}\right]=\frac{-\log P\left(e_{\alpha}\right)}{n}
$$

where $n$ is the number of words in the sentence. Hence, if one wants to maximize average semantic information transmission of the desired event $e_{\alpha}$, it suffices to minimize $n$.

We hypothesize that in general speakers tend to maximize information transmission of the desired semantics $e_{\alpha}$ by minimizing $n$, and therefore by favoring shorter sentences.

The model presented minimizes sentence lengths by estimating at each time step a score that reflects the expected derivation length that would follow the production of a certain word. This is done by the second path shown in Figure 1, below. This path is constituted by a hidden recurrent layer followed by a softmax layer. The recurrent layer contains 30 sigmoid units and receives as input the DSS semantic representation, the output of probabilities, and its own activation at time step $t-1$ (zeros at $t=0)$. Activation of this layer is then propagated to a softmax layer (der_lengths) with dimensionality equal to the size of the vocabulary(43), and that calculates for each word a probability value $D L$, where values closer to 0 represent longer derivations and values closer to 1 represent shorter derivations, and where probability mass is distributed among all words that can be produced at the given time step. Finally, these layers receive also input from a bias unit with a constant activation of 1 .

A model that produces at each time step the word that maximizes this score would prefer words leading to shorter derivations, regardless of their information content:

$$
w_{t+1}=\underset{w}{\arg \max } D L\left(w \mid D S S, \text { probabilities }_{t+1}\right)
$$

## Being Easy to Comprehend

A model combining the previous two strategies would produce sentences with more uniform surprisal profiles, compared to a model that only applies one of them. However, these strategies do not take into account that world events with high surprisal represent higher comprehension effort.

Speakers know beforehand how unexpected the event they are trying to communicate is. Therefore, one can propose that they balance these two strategies according to this information. That is, when the speaker is trying to communicate an event $e_{\alpha}$ with low surprisal, the speaker would prefer to be faster; but, when the event represents high surprisal, the speaker would prefer sentences with lower linguistic surprisal and possibly longer. Thus, at each time step, the model would produce the word that maximizes the score:

$$
w_{t+1}=\underset{w}{\arg \max }\left\{\left(1-P\left(e_{\alpha}\right)\right) P(w \mid \ldots)+P\left(e_{\alpha}\right) D L(w \mid \ldots)\right\}
$$

This final model is expected to produce sentences with more uniform surprisal profiles, compared to strategies that only maximize one of these measures, or that do not take into account semantic surprisal.

In our model this is computed at the words layer (see Figure 1), which receives $P(w \mid \ldots)$ values from the probabilities layer and $D L(w \mid \ldots)$ scores from the der_lenghts layer. The value of $P\left(e_{\alpha}\right)$ is assumed to be known.

## Training and Evaluation

## Examples Set

We use the same examples set as Calvillo et al. (2016), which consist of a set of pairs $\left.\left\{\left(D S S_{1}, \varphi_{1}\right), \ldots,\left(D S S_{n}, \varphi_{n}\right)\right)\right\}$ where each $D S S_{i} \in[0,1]^{45}$ is formed by a DSS representation plus an extra bit that indicates whether the model must produce a a passive sentence (0) or an active one (1); and $\varphi_{i}$ is the set of all the sentences that encode the information contained in the corresponding $D S S_{i}$ and in the expected voice.

The sentences are those generated by the microlanguage defined by Frank et al. (2009) (see their Tables 5-8). This microlanguage consists of 40 words that can be combined into 13556 sentences according to its grammar. After adding determiners (a,the) and an end-of-sentence marker (.), there were 43 words, which were encoded at the output layer probabilities in the form of localist vectors. After ruling out sentences expressing situations that are not allowed by the microworld, there were a total of 8201 sentences related to 782 DSS representations.

This set was used because it pairs each semantic representation with several sentences, allowing to define different ranking functions. In future work a new set could be defined in order to assess more specific phenomena.

Derivation Length Scores. For each DSS representation, we know beforehand the sentences that can encode it according to the grammar. Furthermore, we know at each derivation point what words can be produced and how long the sentences would be if a particular word is produced. Using this
information, we compute a probability distribution over the vocabulary that reflects the length of the sentences that one can expect after producing a particular word.

Given a DSS representation and a derivation point, for each possible word production $w_{i}$, we get its minimum derivation length min_dl $\left(w_{i}\right)$, which is the length of the shortest sentence that can be produced if $w_{i}$ is produced. Afterwards we calculate a score $d l\left(w_{i}\right)$ :

$$
d l\left(w_{i}\right)=\max _{w}\left\{\min _{-} d l(w)\right\}-\min _{-} d l\left(w_{i}\right)+1
$$

which is equal to the difference between the greatest min_dl value among all the words that can be currently produced and the min_dl associated to each specific word $w_{i}$, plus 1. Finally, in order to have a proper distribution, we normalize by dividing by the sum over all the possible word continuations.

These scores are the values expected at the output layer of der_lengths. According to these, all possible word productions at a specific derivation point have some probability mass that is inversely proportional to the length of the shortest sentence that can be obtained by following that production.

Semantic Probability. For each DSS representation in the examples set, a semantic probability value $P\left(e_{\alpha}\right)$ was computed. Considering that the model is trained only on the pairs given in the examples set and that all sentences are presented an equal number of times during training, then the probability of a DSS representation is given by the number of sentences related to that representation divided by the total number of sentences in the examples set.

However, since $P\left(e_{\alpha}\right)$ is used to balance word probabilities and derivation lengths, less biased values are needed because as it is, $P\left(e_{\alpha}\right)$ is in general very low, and $1-P\left(e_{\alpha}\right)$ is very high. Therefore instead of normalizing by the total number of sentences, normalization is done with respect to the highest number of sentences that can be related to a DSS representation, which is 130 . Hence, for each DSS, its probability $P\left(e_{\alpha}\right)$, or henceforth $P(D S S)$, is given by the number of sentences paired with the representation, divided by 130.

## Training Procedure

Since the output layer receives unmodified copies from probabilities and der_lengths, the connections from the latter to the former are fixed one-to-one and do not need training. In other words, the $i^{\text {th }}$ unit of probabilities is only connected to the $i^{t h}$ unit of words with a connection weight fixed to 1 , and likewise for the connections between der_lenghts and words.

Prior to training, all weights on the projections between layers (with the exception of those mentioned in the last paragraph) were initialized with random values drawn from a normal distribution $\mathcal{N}(0,0.1)$. Weights on the bias projections were initially set to zero.

Training consists of setting the connection weights leading to the computation on the one hand of probabilities and on the other hand of der_lengths, corresponding to the two paths of processing. Accordingly, training is performed in
two phases, in both cases using cross-entropy backpropagation (Rumelhart, Hinton, \& Williams, 1986) with weight updates after each word in the sentence of each training item.
probabilities. The first phase corresponds to the training of the path leading to probabilities, which is performed as described by Calvillo et al. (2016), where the model is trained to predict the next word given the semantic representation and the previously produced words.

During this phase, the monitoring units were set at time $t$ to what the model was supposed to produce at time $t-1$ (zeros for $t=0$ ). This reflects the notion that during training the word contained in the training sentence at time-step $t-1$ should be the one informing the next time step, regardless of the previously produced (and possibly different) word. During production, the monitoring units are set to 1.0 for the word that was actually produced and 0.0 everywhere else.

This path was trained for a maximum of 200 epochs, each one consisting of a full presentation of the training set, which was randomized before each epoch. Note that each item of this set consisted of a $D S S_{i}$ paired with one of the possible sentence realizations describing the state of affairs represented in $D S S_{i}$. Hence, during each epoch, the model saw all the possible realizations of $D S S_{i}$. An initial learning rate of 0.124 was used, which was halved each time there was no improvement of performance during 15 epochs. No momentum was used. Training halted if the maximum number of epochs was reached or if there was no performance improvement over a 40-epoch interval.
der lengths. The second path can be trained after the training of the first one is completed. During this phase, the connection weights calculated during the first phase are fixed, so that only the second path weights are modified.

At each time step, the DSS is fed into the first path, which outputs a probability distribution over the vocabulary. This is fed into the second recurrence, as well as the DSS representation. Monitoring units are handled exactly as in the first training phase. The activation of the second recurrence is then propagated to der_lengths. Its output is compared to the derivation length values, as defined in the previous section, and finally the connection weights are updated.

Training of this path was performed for a maximum of 80 epochs, with the training items arranged in the same way as in the previous phase. An initial learning rate of 0.24 was used, which was halved each time there was no improvement of performance during 10 epochs. No momentum was used. Training halted if the maximum number of epochs was reached or if there was no performance improvement over a 20-epoch interval.

## Evaluation

The model presented defines a production strategy as an interaction between production goals. Thus, in order to assess the model, its productions were compared to those obtained by using the following alternative strategies, where at each
time step the model produces the word with:

- Min Linguistic Surprisal
- Min Derivation Length
- Max Word Probability +/* Derivation Length Score
- Complete Model

For each DSS representation in the examples set that was related to more than one sentence (968), the model generated a sentence according to each production strategy.

In order to measure surprisal, a language model was trained implementing a Simple Recurrent Network (Elman, 1990). This model was trained on the whole set of sentences for 200 epochs with a learning rate of 0.24 which was halved each time there was no improvement in performance. Using this language model, surprisal values were calculated for each one of the words of the produced sentences.

Uniformity of information density was measured in terms of standard deviation of word surprisal, assuming that complete uniformity would produce a standard deviation of 0 .

## Results and Discussion

The results can be seen in Table 2, where the columns denote respectively: production strategy, production accuracy (Acc) as defined by Calvillo et al. (2016) and denoting how precise the sentences convey the given semantics, average sentence length (AvDL), average word surprisal (AvS), and standard deviation of surprisal (Std).

Table 2: Results of each production strategy.

|  | Acc | AvDL | AvS | Std |
| :---: | :---: | :---: | :---: | :---: |
| Min LS | 99.67 | 9.01 | 1.0 | 0.89 |
| Min DL | 99.86 | 7.55 | 1.20 | 0.97 |
| Max P(+/*) DL | 99.82 | 7.77 | 1.16 | 0.95 |
| Max 3P-2DL | 98.23 | 10.15 | 0.89 | 0.84 |
| SemSurp | 97.67 | 10.17 | 0.89 | 0.83 |

As expected, minimizing linguistic surprisal (Min LS) led to lower surprisal values compared to minimizing derivation lengths (Min DL). Combining these two strategies by a sum or product led to results almost identical to each other, and very close to Min DL, suggesting that derivation length scores were mostly dominating production.

Given that linguistic surprisal and derivation lengths are different in nature, one can expect a more complex relation between them in order for the resulting score to be helpful. Consequently, grid search was performed in order to find linear factors that would minimize the standard deviation of surprisal. The resulting model corresponds to the fourth row in Table 2, where the model produces at each time step the word that maximizes:

$$
3 P\left(w \mid D S S, w 0, . ., w_{n}\right)-2 D L(w \mid D S S, \text { probabilities })
$$

where one can see that minimizing linguistic surprisal is favored, while minimizing derivation lengths is penalized. As a result the sentences produced are longer than only minimizing linguistic surprisal. However, uniformity of information density is higher than with the previous models and additionally average surprisal is lowest.

The final row in Table 2 presents the results of the model that incorporates semantic probabilities. For this case grid search was also used, which led to a model that at each time step produces the word that maximizes:

$$
(3.5-P(D S S)) P(w \mid \ldots)+(P(D S S)-2.5) D L(w \mid \ldots)
$$

which is very similar to the previous model, but with some influence from semantic probabilities. While the performance of this model is very similar to the previous one, its sentences present slightly higher uniformity of information density; and the influence of semantic surprisal is in the expected direction, where semantics with high surprisal produce longer sentences and vice versa.

The small difference between the last two strategies could be caused by the nature of the language model, which receives no semantic information during training, which means that rather than being a joint model of semantics and sentences, it only considers word sequences. Furthermore, the production model here proposed uses semantic surprisal at a sentence level, while speakers can be sensitive to this information incrementally at a word level. These issues will be addressed in future work.

In general the model outlined here shows: first, that as expected, shorter sentences are more dense in terms of information content. Second, that longer sentences present information in a more uniform way. Third, that sentences with more uniform information densities present in average lower word surprisal, therefore minimizing comprehension effort. And finally and most importantly, that sentences with higher uniformity of information density can be produced by balancing sentence lengths and word probabilities. In future work, this can help to address uniformity for a given channel capacity.

## Conclusion

This article presents a model of language production that takes into account word probabilities and sentence lengths in order to produce sentences with uniform surprisal profiles, and in order to model the Uniform Information Density Hypothesis. The sentences produced by this model were compared to those produced using other strategies, showing that the proposed model produces sentences with more uniform surprisal profiles and lower average word surprisal. This model represents a first attempt to model the Uniform Information Density Hypothesis at the algorithmic level, where uniformity arises by balancing word probabilities and sentence lengths in a mechanistic way.

## Acknowledgements

I would like to thank Matthew W. Crocker and Harm Brouwer for their helpful comments while writing this article.

## References

Aylett, M., \& Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and speech, 47(1), 31-56.
Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., \& Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. The Journal of the Acoustical Society of America, 113(2), 1001-1024.
Calvillo, J., Brouwer, H., \& Crocker, M. W. (2016). Connectionist semantic systematicity in language production. In Proceedings of the 38th annual conference of the cognitive science society.
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.
Ferreira, V. S., \& Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive psychology, 40(4), 296-340.
Frank, S., Haselager, W. F. G., \& van Rooij, I. (2009). Connectionist semantic systematicity. Cognition, 110(3), 358379.

Frank, S., Koppen, M., Noordman, L. G. M., \& Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. Cognitive Science, 27(6), 875-910.
Frank, S., \& Vigliocco, G. (2011). Sentence comprehension as mental simulation: an information-theoretic perspective. Information, 2(4), 672-696.
Griffin, Z. M., \& Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. Journal of Memory and Language, 38(3), 313-338.
Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies (pp. 18). Stroudsburg, PA, USA: Association for Computational Linguistics.
Jaeger, T. F. (2006). Redundancy and syntactic reduction in spontaneous speech. Unpublished doctoral dissertation, Stanford University.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. Cognitive psychology, 61(1), 23-62.
Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3), 1126-1177.
Levy, R., \& Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. Advances in neural information processing systems, 19, 849.
Rumelhart, D. E., Hinton, G. E., \& Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

# Learning to Learn Visual Object Categories by Integrating Deep Learning with Hierarchical Bayes 

Andres Campero (campero@mit.edu)<br>Andrew Francl (francl@mit.edu)<br>Joshua B. Tenenbaum (jbt@mit.edu)<br>Department of Brain and Cognitive Sciences<br>Massachusetts Institute of Technology, MA 02139 USA


#### Abstract

Humans are capable of generalizing and learning new concepts after very little experience. They have the ability to create semantic structures from concepts they acquire, they can learn appropriate inductive biases that are later used as priors for different tasks, and they can learn novel categories from very few examples. While recent advances in neural networks and other machine learning methods are beginning to approach humanlevel capabilities in several tasks, building computational models that replicate these abilities has proven difficult. We propose a model that combines powerful features extracted from a deep neural network with a semantic structure inferred using probabilistic Hierarchical Bayes. We test and demonstrate the capabilities of our model in three different tasks: learning a new concept from a single example of a novel category, learning new categories from few examples of different categories, and learning the semantic tree from an unlabeled set of novel objects.


Keywords: hierarchical bayes; one-shot learning; inductive bias; neural networks; unsupervised learning

## Introduction

Recent advances in neural networks and other machine learning methods have led to computer vision object-recognition systems that are beginning to approach human-level performance. Trained on thousands of object categories, with thousands of labeled examples for each, deep convolutional networks can tell if a new image contains a familiar category almost as well as human adults can in a brief glance. Yet, even young children have abilities to learn and generalize that go beyond what current machine vision systems can do. Here we focus on three such abilities:
(1) By age 3, children can learn new object categories from just a single example. Furthermore, children generalize in different ways as appropriate for different kinds of categories: labels for artifacts with functionally relevant shapes are preferentially generalized according to those shapes, while labels for non solid substances or arbitrarily shaped objects are more likely to be generalized according to material properties.
(2) Children can learn to learn appropriate inductive biases, such as the shape and material biases described above, from experience with just a few examples each of a small number of categories that exemplify these biases in a consistent way. The shape-bias training studies of Smith and colleagues are the best known examples (Smith, Jones, Landau, GershkoffStowe, \& Samuelson, 2002).
(3) Children can, in a completely unsupervised way, sort novel objects into categories and supercategories in a meaningful way, and then use these hierarchical category structures
as strong constraints to learn and generalize names for objects from just one or a few examples.

Previous attempts to capture these abilities in computational models have had some success, but not with models that are "image-computable" on the same stimuli that people see. These earlier models have used either adult similarity judgments (Xu \& Tenenbaum, 2007) or highly simplified, idealized feature representations (Kemp, Perfors, \& Tenenbaum, 2007) to build their category hierarchies. Here we show that a computational framework can come close to capturing abilities (1-3) by combining two powerful representation-learning techniques: deep learning for feature construction and Hierarchical Bayes for unsupervised taxonomy construction.

We build on work by Salakhutdinov, Tenenbaum, and Torralba (2012) who build a Hierarchical Bayesian model that "learns to learn" by incorporating information from past experience into a prior when inferring statistical properties of a novel category. In particular, when presented with a few image examples of a new category, the model infers a supercategory and uses the higher-order knowledge abstracted from previous categories to identify the relevant features and allow generalization (Figure 1).


Figure 1: Learning a similarity metric for a new category. The goal is to identify the correct supercategory and estimate an appropriate similarity metric.

That work was extended by the same authors, who harnessed a two layer Deep-Boltzmann Machine to generate low level feature representations of the images while learning a prior using a hierarchical Dirichlet process. (Salakhutdinov,

Tenenbaum, \& Torralba, 2013). Their experimental data showed that using this prior in combination with more powerful features gave them a distinct advantage over other methods of classification. This progression of work suggests that building a model that combines complex feature spaces with a hierarchical semantic structure may lead to further increases in performance.

Building on this line of work, we contribute a model that combines the two components: powerful image representations extracted from Deep Neural Networks (DNNs) and a Hierarchical semantic structure that works as a Bayesian prior. We show how the combination of these two components can "learn to learn" in ways that resemble some aspects of child cognition. Additionally, we explore how this model's performance is affected as we vary different aspects of the model architecture and the structure of the training data.

Other approaches to combine probabilistic graphical models and DNNs have recently been proposed that focus on building unsupervised clustering algorithms (Dilokthanakul et al., 2016; Johnson et al. 2016). Instead, the focus of our model is to capture certain aspects of human cognition. This leads to some notable differences. First, representations in our model are a fixed set of visual relevant features instead of being learned for the inference task at hand. In addition, our model's generative component is limited to a hierarchical structure that aims to recover the semantic relations between concepts in a useful and meaningful way while other models are fully generative but tend to have graphs with simpler semantic structures. We therefore propose a relatively simple model that is not intended for general unsupervised learning but that instead focuses on traits of human object and category learning.

More specifically, we test our model's capacity to capture the previously discussed human abilities (1-3) in an image recognition framework. First, we evaluate the ability of our model to learn novel categories from only one or a few examples. To address this we allow the model to construct a semantic structure from labeled examples in a data set and then judge the model's performance on a one-shot learning task. Second, we assess the models capability to construct inductive biases in low data environments. We test this ability by repeating the first task but limiting the training data available to the model when it constructs the semantic tree. Finally, in a third task, we test the model's ability to learn a hierarchical semantic structure of novel objects in a completely unsupervised manner. Results suggest that this approach may be suitable for modeling certain aspects of cognition.

## Model and Learning to Learn

Our model combines two Machine Learning approaches that have recently been successful at a range of differing tasks. On one hand, powerful deep networks construct feature spaces that enable rapid and accurate classification. On the other, Hierarchical Bayesian Models have proven successful in creating taxonomies of the different concepts learned from pre-
vious experience. These taxonomies can then be used as a prior to identify the relevant features for learning a new category from one or a few examples based on the distribution of other similar categories. We create various versions of our model to compare combinations of feature spaces extracted from different architectures with variants of the Hierarchical Bayesian component.

Learning begins by constructing a 2-level tree of categories and supercategories that best explains the training observations under a Bayesian framework. The model learns structure in the observations by first generating useful general features from a DNN and then developing hierarchical priors that allow previous similar experiences to bias the learning of new concepts and categories. The priors are constructed by inferring the means and variances that define the most relevant dimensions from the DNN feature representations for each category and supercategory (Figure 1).

## Deep Network Features

We use features extracted from DNNs pretrained for object classification on ImageNet. We obtain a representation from each image by passing it through a network and extracting the response from the penultimate layer consisting of 4096 realvalued dimensions. In the regular deep network classification scheme, this response is then passed through a linear weighting and a generalized logistic regression layer. This layer maps this representation onto probabilities for each class in the specific classification task for which the network was trained.

We compare the performance of the different versions of our model on features extracted from two different DNN architectures: Alexnet (Krizhevsky, Sutskever, \& Hinton, 2012), which was the first implemented Deep Learning Model that significantly improved object classification on images; and VGG-16 (Simonyan \& Zisserman, 2014), a more recent architecture with 16 layers that achieves above $90 \%$ top 5 classification performance on ImageNet.

## Generative Semantic Organization

After obtaining a useful general image representation from the DNN, the Hierarchical Bayesian Model's parameters are inferred by approximating the posterior via Markov Chain Monte Carlo methods in the following way.

Consider a two-level hierarchy where $N$ observed inputs are partitioned into $C$ basic-level categories, these categories are in turn partitioned into $K$ supercategories. In this hierarchy of observations, categories, and supercategories, the higher levels determine a prior over the distribution of the lower levels. In particular, the distribution over observations (feature vector representations of images in our case) of each of the different basic level categories are assumed to be multivariate Gaussian with a category specific mean $M_{c}$ and with precision terms $\tau_{c}^{d}$ that are assumed to be independent across the $D$ dimensions of the feature space. These precision terms constitute a similarity metric by determining the relative importance of each of the features. In turn, we place a conjugate

Table 1: Performance results using the area under the ROC curve (AUROC) on the MSR dataset in the one-shot learning task

|  | \# Examples from Withheld Class |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Alexnet |  |  |  | VGG |  |  |  |
|  | 1ex | 2ex | 4 ex | 20 ex | 1ex | 2ex | 4 ex | 20ex |
| Oracle | .99 | 1 | 1 | 1 |  |  |  |  |
| HB-Full | .91 | .96 | .98 | .99 | .92 | .97 | .98 | .99 |
| One Supercategory | .87 | .94 | .97 | .99 | .88 | .95 | .98 | .99 |
| NearestN | .84 | .86 | .87 | .90 | .89 | .90 | .92 | .95 |
| T of T* | .76 | .80 | .84 | .87 |  |  |  |  |

Normal-Gamma prior over $\left\{M_{c}, \tau_{c}\right\}$, this prior is determined by the supercategory specific level-2 parameters $M_{k}, \tau_{k}, \alpha_{k}$, where $M_{k}$ and $\tau_{k}$ constitute the expected values of the lower level parameters and $\alpha_{k}$ controls the variability of $\tau_{c}$ around its mean. Finally, for the conjugate priors over the level-2 parameters, we respectively assume Normal, Exponential and Inverse-Gamma distributions that are further shaped by parameters $\alpha_{0}$ and $\gamma_{0}$. The full generative model is given in Figure 2 (Salakhutdinov et al., 2012).


Figure 2: Hierarchical Model
Given a set of observations, the model iteratively performs Bayesian inference by alternating between sampling the parameters and inferring the category assignments. When learning the distributions at each step of the iteration, the supercategory membership is fixed and the parameters are sampled from posteriors that are analytically computed using the conjugate priors ${ }^{1}$. The supercategory membership for each category is learned in a similar way by fixing the currrent parameters and the rest of the hierarchical structure. Every category can be assigned to any of the existing supercategories or to a newly created one. The posterior probability of belonging to a supercategory is computed as a combination the likelihood that the parameters of the category come from the parameters of the supercategory and a Chinese Restaurant Process (CRP) prior (Griffiths \& Tenenbaum, 2004). This nonparametric prior is a distribution over a partition on integers in which the $n^{\text {th }}$ number is assigned to set $k$ with probability:

[^294]\[

P\left(z_{n}=k \mid z_{1}, z_{2} ···, z_{n-1}\right)= $$
\begin{cases}\frac{n^{k}}{n-1+\gamma} & \text { if } n^{k}>0 \\ \frac{\gamma}{n-1+\gamma} & \text { if } \mathrm{k} \text { is new }\end{cases}
$$
\]

Where $n^{k}$ is the number of previous integers assigned to set $k$ and $\gamma$ is a concentration parameter sampled from a $\operatorname{Gamma}(1,1)$ distribution.

In an unsupervised setting where the categories of the observations are also unknown, the model utilizes a similar strategy to assign observations to categories as is used when assigning categories to supercategories. The model iterates through the observations and assigns each either to an existing or to a newly created category based on the prior and likelihood. By utilizing the CRP prior, the model can create an unbounded number of categories and supercategories. This entire process constitutes a Gibbs sampling procedure where both the tree structure and all of the parameters are simultaneously learned.

## Tests and Results

We test the model in scenarios that attempt to capture aspects of human cognition related to learning from limited data. First we measure the model's ability to generalize previous knowledge to learn novel categories from only a few examples. Next, we assess the model on this task when the training data for all of the categories is also limited to only a few examples. Finally, we exploit the model's full hierarchy in a completely unsupervised setting by exploring how the model recovers the underlying semantic structure.

## One-Shot Learning on MSR

In the first task, we test the model's ability to learn new categories form one or a few examples. First, we select a category that will be held-out for testing. Labeled observations for all other basic-level categories are provided for training. The model learns the semantic structure of the training set by clustering the basic categories into supercategories and inferring the relevant parameters at all levels of the Bayesian Hierarchy. The challenge is then to generalize the learned structure to the held-out category from only one or a few examples.

To do this, the model first infers the best supercategory from one or a few examples of the withheld category by

Table 2: Performance results using the area under the ROC curve (AUROC) on the MSR dataset with limited training data.

|  | \# Examples from Withheld Class |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Alexnet |  |  |  | VGG |  |  |  |
|  | 1 ex | 2 ex | 4 ex | 20 ex | 1 ex | 2 ex | 4 ex | 20 ex |
| \# Training Examples |  |  |  |  |  |  |  |  |
| 1 ex | .87 | .87 | .88 | .89 | .90 | .90 | .90 | .92 |
| 4 ex | .92 | .96 | .99 | .99 | .93 | .97 | .98 | .99 |
| 10 ex | .92 | .96 | .99 | .99 | .92 | .96 | .98 | .99 |
| 18 ex | .92 | .95 | .98 | .99 | .91 | .96 | .98 | .99 |
| All examples | .91 | .96 | .98 | .99 | .92 | .97 | .98 | .99 |

marginalizing over the category level parameters. Next, the model uses the supercategory priors and training examples to estimate the category similarity metric and mean for each dimension in the feature space.

We evaluate different versions of our model on the MSR Cambridge dataset (Kohli et al., 2005), which consists of 24 categories with varying numbers of images in each category. In total this dataset contains roughly 800 images. Figure 3 shows a typical partition over all the categories discovered by the full model. To quantify the models accuracy, a testset with unlabeled data from all categories is classified.

We repeatedly trained the model withholding one of the categories at a time and then inferred the withheld category parameters and supercategory membership using one or a few images. Next, we calculated the posterior probability for each testset image belonging to each category and variated a threshold to classify images as belonging to the heldout category or to any of the other categories. This created true and false positive rates for each point along our threshold which traced out a Receiver Operating Characteristic curve (ROC) for classifying objects from the withheld vs. all the other categories. The reported results are calculated by averaging the Area Under the ROC curve (AUROC) for the model trained with each of the 24 categories withheld (Table 1).

Performance is compared for each combination of an Inference Model and a Network Architecture. HB-Full is the full version of the model described above. One Supercategory places all the categories in the same single supercategory. NearestN classifies new points with the label of the nearest neighbor of its feature vector in euclidean distance. Texture of Textures (T of T)* replaces our DNN features with the set of responses from a three layer convolutional neural network that uses precomputed weights that resemble Gabor filters ${ }^{2}$. Finally, the Oracle is the same than our full model, but uses the true empirical mean and variances from the whole population (including testset). Table 1 shows the results for the two different feature spaces used.

[^295]

Figure 3: MSR semantic tree discovered by the Full Model
The results show that the model performs best when using the full hierarchy in combination with the feature space extracted from VGG. HB-Full considerably outperforms alternatives under both feature spaces, particularly for trials with one example from the withheld dataset. As more examples become available, the performance difference decreases reflecting the importance of the prior when little data is available. The importance of the learned features is highlighted when comparing with the T of $\mathrm{T}^{*}$ feature space where performance is considerably lower. It is interesting to note that the VGG representation improves most over Alexnet when in combination with NearestN, but the effect is mitigated when the hierarchy is used.

## Limited training data regimes

In a second task, we test the capability of our model to extract inductive biases from experience with just a few examples. To evaluative this capability, our full model was limited to only $1,4,10$ or 18 examples of each category used for training. The number of examples from the withheld category was varied separately. Table 2 shows the average AUROC for the same "one vs. all" metric used in the previous task ${ }^{3}$. For comparison, the full model performance from the previous table is included and labeled as "All examples" ${ }^{4}$.

We can see that the largest jump in performance happens when moving from 1 to 4 training examples. This likely reflects the fact that a single example provides information about the mean of the category but not about the variance or similarity metric, which has to be inferred completely from

[^296]the prior. However, 4 examples provide adequate information about the variance to allow the model to appropriately infer the parameters for new categories. As the number of training examples continues to increase, there are no further gains in performance. This is consistent with literature showing that children need at least two examples to learn inductive biases in certain contexts (Smith et al., 2002).

## Unsupervised Learning on Gazoobian Objects

Humans and children can sort new objects into categories and supercategories in a semantically meaningful way. While our model is also able to of recover meaningful structure from labeled examples (Figure 3), real situations often demand learning where labels are completely absent. Schmidt (2009) explores this human capability with a dataset composed of 45 novel objects that were generated using a modeling software to simulate a specific taxonomic structure. The dataset consists of three supercategories supposed to be alien equivalents of plants, tools and snails from the planet "Gazoob". The objects in each supercategory are further organized into a structure that can be approximated by basic-level categories (gray box in Figure 4).

Our model has the ability to infer both categories and supercategories in an unsupervised manner from observations. Schmidt (2009) shows that a model based on agglomerative clustering that uses adult similarity judgments is able to recover the taxonomic tree (Figure 4). Here our model is tested with the harder task of recovering the taxonomic tree directly from the same images that people saw. The model accomplishes this task in a fully unsupervised manner using a single image of each object.

This "image-computable" model is able, although with some mistakes, to recover the three supercategories and most of the basic-level category structure (Figure 5). Other unsupervised clustering algorithms were also able to capture some of the semantic structure, but the hierarchy between categories and supercategories was not evident.

## Discussion

One can think of the task of concept learning as consisting of two elements. The first involves obtaining relevant features to represent the objects and categories commonly observed in the world. The second involves constructing a semantic hierarchical structure with links between categories that humans can use to navigate and perform tasks. While recent results demonstrate the capabilities of DNNs to classify categories provided a large number of training examples, they struggle to perform tasks that require understanding the semantic relationships between classes. The ability of Hierarchical Bayesian Models to build these semantic structures can further help with understanding and classifying new categories.

We demonstrate how these two approaches can complement one another by combining them in a computational model. We tested the model's abilities tasks designed to approximate human capabilities that are currently difficult for computer vision systems such as concept generalization,
learning inductive biases, and constructing semantic structures. We show results for three tasks involving limited data availability. The model is able to learn relevant semantic structures from just a few examples of novel objects and effectively transfer appropriate similarity metrics from learned categories in the form of a prior. In all tasks, the computational framework comes close to capturing human abilities that other, more complex, machine vision systems struggle to reproduce.

## References

Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., \& Shanahan, M. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648.
Griffiths, T., \& Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. Advances in neural information processing systems, 16, 17.
Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P., \& Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. In Advances in neural information processing systems (pp. 2946-2954).
Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. Developmental science, 10(3), 307-321.
Kohli, P., Sharp, T., Minka, T., Winn, J., Shotton, J., \& Criminisi, A. (2005). Microsoft research in cambridge image dataset.
Krizhevsky, A., Sutskever, I., \& Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
Salakhutdinov, R., Tenenbaum, J. B., \& Torralba, A. (2012). One-shot learning with a hierarchical nonparametric bayesian model. In Icml unsupervised and transfer learning (pp. 195-206).
Salakhutdinov, R., Tenenbaum, J. B., \& Torralba, A. (2013). Learning with hierarchical-deep models. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1958-1971.
Schmidt, L. A. (2009). Meaning and compositionality as statistical induction of categories and constraints. Unpublished doctoral dissertation, Citeseer.
Simonyan, K., \& Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., \& Samuelson, L. (2002). Object name learning provides on-the-job training for attention. Psychological Science, 13(1), 13-19.
Xu, F., \& Tenenbaum, J. B. (2007). Word learning as bayesian inference. Psychological review, 114(2), 245.
Yildirim, I. (2012). Bayesian inference: Metropolis-hastings sampling. University of Rochester, $N Y$.


Figure 4: Ground Truth Tree of Gazoobian Objects as Generated from Human Similarity Judgments. Each of the three branches at the top of the tree denotes a supercategoy. The gray box in the lower left hand of the figure denotes a basic-level category.


Figure 5: Model's Inferred Semantic Hierarchy of Gazoobian Objects. Outer boxes denote supercategories inferred by the model. Dashed lines separate model generated categories within each supercategory. Colored boxes around each object denote the ground truth supercategories as shown above.

# The paradox of relational development is not universal: Abstract reasoning develops differently across cultures 

Alexandra Carstensen (a.carstensen@let.ru.nl)<br>Centre for Language Studies, Erasmusplein 1, Radboud University<br>6525 HT, Nijmegen, NL

Caren M. Walker (carenwalker@ucsd.edu)
Department of Psychology, 9500 Gilman Drive, UC San Diego
La Jolla, CA 92093 USA


#### Abstract

Recent studies demonstrate a puzzling decline in relational reasoning during development. Specifically, 3-year-olds fail in a relational match-to-sample (RMTS) task, while younger children (18-30 months) succeed (Walker, Bridgers, \& Gopnik, 2016). Hoyos, Shao, and Gentner (2016) propose that older children fail because of a bias toward individual object properties induced by "avid noun learning." If this is the case, children learning a language with a stronger emphasis on verbs, like Mandarin Chinese, may show an attenuated decline in relational reasoning. We first test this possibility by reproducing the causal RMTS task in China, and find that Mandarin-speaking 3 -year-olds outperform their Englishspeaking peers in the U.S. In a second experiment, we show that Mandarin speakers exhibit a corresponding bias toward relational solutions while English speakers prefer objectbased solutions in an ambiguous context. We discuss possible mechanisms through which language and culture may promote (or hinder) the early development of relational reasoning.


Keywords: cognitive development, causal learning, relational reasoning, overhypotheses, language, culture.

## The puzzling decline of relational reasoning

Relational reasoning is often cited as a defining feature of human cognition (e.g., Gentner, 2003), and a source of the differences between the abilities of humans and other primates (Penn, Holyoak, \& Povinelli, 2008). The ability to recognize relational similarities appears surprisingly early in human development: 7- and 9-month-old infants distinguish the abstract relations "same" and "different," looking longer at novel pairs of objects that differ from a habituated relation (Ferry, Hespos, \& Gentner, 2015; Tyrrell, Stauffer, \& Snowman, 1991).

Toddlers (18-30 months) can also employ these concepts to infer abstract causal properties in a relational match-tosample task (Walker \& Gopnik, 2014). In this task, children observe as four pairs of blocks are placed on a toy that plays music when "activated." Two of the pairs contain identical blocks ("same") and the other two pairs contain mismatched blocks ("different"). For toddlers in the same condition, the toy activates and plays music only when the "same" pairs are placed on top, while those in the different condition observe the opposite pattern. When shown novel pairs of "same" or "different" blocks and asked to choose which pair
would activate the toy, toddlers succeed in picking the pair that is relationally consistent with their training.

However, this early success in relational reasoning is quickly followed by a puzzling decline: 3-year-olds (36-48 months) fail to select the relational solution in precisely the same task (Walker, Bridgers, \& Gopnik, 2016). Similar difficulties have also been observed in a variety of relational reasoning tasks (e.g., Christie \& Gentner, 2007; 2010; 2014; Gentner, 1988; 2010; Hoyos, Shao, \& Gentner, 2016). By 4 years of age (52-60 months), children once again succeed in a standard RMTS task (Christie \& Gentner, 2014), but continue to neglect relational similarities in other contexts even at 5-6 years of age (e.g., Gentner, 1988). This pattern of early success, decline, and reemergence suggests that the development of relational reasoning may follow a U-shaped trajectory, rather than a continuous process of gradual improvement, as previously suggested (e.g., Gentner \& Medina, 1998). What causes this curious dip in children's relational reasoning?

One possibility is that preschoolers retain an early competence to reason about relations, but that this competence is overshadowed by a failure to attend to relational structure. In particular, Walker et al. (2016) suggest that 3 -year-olds neglect relational information as a result of a learned bias to attend to individual object kinds and their properties. This claim is consistent with a large literature demonstrating that preschool-aged children attend to objects and attributes, and proposals that children must overcome an "entity-based view" in order to effectively process relations (Christie \& Gentner, 2010; also, e.g., Christie \& Gentner, 2007; 2014; Gentner, 1988; Gentner \& Rattermann, 1991; Hall \& Waxman, 1993).

Several proposals link this well-documented object bias to language development, which has been shown to both foster and impair relational reasoning (e.g., Christie \& Gentner, 2014; Hoyos, Shao, \& Gentner, 2016). These seemingly incongruous findings have led some to regard the contradictory effects of language on relational thinking as a developmental paradox (Hoyos, Shao, \& Gentner, 2016).

## Noun learning and relational development

In a recent paper, Hoyos, Shao, and Gentner (2016) suggest that the decline of relational reasoning may stem from an object bias induced by language learning. They reason that
"avid noun-learning" in early childhood likely leads to a "captivation with objects," which in turn helps children to learn additional nouns. In support of this view, they provide evidence that an experimentally induced noun bias interferes with relational reasoning. In this experiment, they replicate a previously published finding (Christie \& Gentner, 2014) that 4-year-olds succeed on a standard RMTS task in a baseline condition, but show that priming nouns in a picture-labeling activity significantly reduces subsequent RMTS performance. On its own, this outcome suggests that language learning-and an emphasis on nouns in particular-may negatively impact relational reasoning in toddlers.

However, earlier work by Christie and Gentner (2014) leads to the opposite conclusion, that linguistic conceptsand nouns in particular-facilitate relational reasoning in the standard RMTS task. They find that providing children with a novel noun ("truffet") for pairs of objects improves toddlers' subsequent RMTS performance. Here, and elsewhere, the authors argue that young children do not initially have access to a hypothesis space that is "sufficient to allow for the range of possible semantic categories" but instead form hypotheses about relational meanings by comparing co-labeled items to identify common structure (Christie \& Gentner, 2010).

Taken together, this account and the conflicting findings create an apparent paradox, in which a linguistic emphasis on nouns orients young learners away from relations, but language simultaneously provides the necessary scaffolding for relational learning by highlighting relational structure. In this way, language learning appears to solve the very problem it creates. Accordingly, this account implies that language learning may be interpreted as a double-edged sword, drawing attention to objects at the expense of relations, but in doing so, ultimately helping children to construct novel relational categories.

## Language as a driver of children's hypotheses

The current research further explores the hypothesis that language learning influences the types of concepts and categories that young children entertain. Under the paradox account presented above, language learning helps children develop new relational categories to further populate their hypothesis space. An alternative possibility is that children have access to both relational and object-based hypotheses throughout development, but that the probabilities assigned to each type of hypothesis change as a result of prior knowledge and past experience, including language learning. This account draws on probabilistic models of cognitive development in which children are seen as Bayesian learners (e.g., Gopnik \& Wellman, 2012), who weight the likelihood of a given hypothesis (the probability of the data given the hypothesis) by its prior probability (the general probability of the hypothesis, before any data are observed). Consequently, if a hypothesis has high prior probability, it will require stronger data to overturn it. This reasoning may also be applied to entire categories of
hypotheses in the form of an overhypothesis, a general principle by which the learner assigns higher prior probability to particular types of hypotheses (Kemp, Perfors, \& Tenenbaum, 2007). From this perspective, the "noun explosion" in early language learning could motivate an object bias-and temporary dip in relational reasoningin the form of an overhypothesis that privileges object-based hypotheses over relational ones (for a discussion of language-induced overhypotheses and their relevance beyond language, see Colunga \& Smith, 2005). By this account, language acts as one of many possible influences that affect a learner's hypothesis space, not by providing for new hypotheses (as the paradox view suggests), but by adjusting children's existing prior expectations.

Despite this distinction, both of these accounts leave room for an important role of language in driving children's relational reasoning, and both predict that a noun focus in word learning would (at least initially) bias children toward object properties and away from relations.

Previous demonstrations (e.g., Christie \& Gentner, 2010; Hoyos, Shao, \& Gentner, 2016) have tested this hypothesis indirectly, showing that immediate exposure to nouns modulates success on RMTS tasks, presumably by directing the learner's attention toward or away from relational information. However, these findings (which may reflect simple priming effects) do not necessarily demonstrate a relationship between noun focus in word learning and RMTS performance, as the experimental groups all involve English speakers, without any systematic between-group differences in degree of noun focus.

Conveniently, not all word learning follows the same trajectory. In particular, the "noun explosion" that has been documented in English-language learners is not universal across languages. In Korean, for instance, there is evidence for a comparable "verb spurt" (Choi \& Gopnik, 1995). Similarly, several studies have found that children learning Mandarin Chinese produce more verbs than nouns in their spontaneous speech (both types and tokens), in contrast with English speakers of the same age, who produce a greater proportion of nouns than verbs (Tardif, 1996; Tardif, Shatz, \& Naigles, 1997).

If an emphasis on noun learning (relative to other parts of speech) indeed drives the dip in relational reasoning by fostering an object bias, then children learning a more verbcentric language should show an attenuated or reversed bias. While nouns may direct focus to object properties by relying on these in picking out meanings, verbs often signal relational meanings across multiple entities, and might serve to redirect attention accordingly.

The difference in noun focus between English and Mandarin Chinese therefore presents two natural conditions in word learning, which we exploit as a test of the proposal that properties emphasized in word learning induce a bias in reasoning more generally.

## Experiment 1: Causal relational reasoning in Mandarin-speaking children

To test for a relationship between noun focus in word learning and relational reasoning, we first reproduced Walker et al.'s (2016) causal RMTS task (see Figure 1) with Mandarin-speaking children (36-48 months) in China.

## Methods

Participants. A total of 64 Mandarin-speaking 36-48-month-olds ( $M=42.1$ months; 28 female) took part in Experiment 1. This sample size was chosen based on previously published studies using the same paradigm. Participants were pseudo-randomly assigned to either the same or different condition. Five additional participants were excluded due to experimenter error or failure to complete the study. All participants were native speakers of Mandarin Chinese, and were recruited and tested at preschools in China.

Materials and procedure. The materials and procedure replicated those used in Experiment 1 of Walker et al. (2016), with the exception that instructions were given in Mandarin Chinese. The original English instructions (described here in English) were independently translated and backtranslated to ensure accuracy.

Children were tested individually, seated at a table across from the experimenter. The causal RMTS task began after a brief warmup to familiarize the child with the experimenter. During the task, the experimenter placed matching and mismatched pairs of painted wooden blocks on top of a box which appeared to play music in response to certain blocks. In reality, the experimenter activated a wireless doorbell inside the box by surreptitiously pushing a button.

The experimenter began by placing an opaque cardboard box on the table, saying "This is my toy! Sometimes it plays music when I put blocks on top and other times it does not. Should we try some and see how it works?" The experimenter then produced two blocks, said "Let's try!" and put both blocks on top of the toy simultaneously. The toy played music and the experimenter said "Music! My toy played music!" The experimenter picked up the blocks and set them back on the toy, which again played music, saying "Music! These ones made my toy play music!" She then repeated this procedure with a new pair of blocks in the opposite relation. The new pair did not make the toy play music, and the experimenter responded to the first try with "No music! Do you hear anything? I don't hear anything," and after the second try, said "No music. These ones did not make my toy play music." This pattern was repeated with two additional pairs of blocks. The experimenter always began with a causal pair (identical blocks in the same condition and blocks of differing colors and shapes in the different condition), and alternated inert, causal, inert, using novel blocks in each new pair, and randomizing the specific blocks between participants.


Figure 1: Schematic illustration of training and test trials in Experiment 1. Reprinted from Walker et al. (2016).

After the four training trials, the experimenter said "Now that you've seen how my toy works, I need your help finding the things that will make it play music. I have two choices for you." The experimenter presented the child with two new pairs made of novel blocks, one "same" pair and one "different." Each pair was supported by a tray, which the experimenter held up as she said "I have these...and I have these. Only one of these trays has things that will make my toy play music. Can you point to the tray that has the things that will make it play?" They trays were placed on either side of the toy, just out of reach of the child, with the side of the correct pair and order of presentation counterbalanced between participants. The experimenter recorded the child's first point or reach, and scored the answer as correct if the child chose the test pair (same or different) that corresponded to her training.

## Results and discussion

Mandarin-speaking preschoolers selected the test pair that was consistent with their training in both same ( $69 \%$; onetailed binomial $p=.025$ ) and different ( $72 \%$; one-tailed binomial $p=.010$ ) conditions (see Figure 2). ${ }^{1}$

[^297]

Figure 2: Proportion of correct relational matches selected by English- and Mandarin-speaking toddlers. English speaker data is reproduced from Experiment 1 of Walker et al. (2016). Error bars indicate $95 \%$ confidence intervals.

As predicted, Mandarin-speaking preschoolers succeed in the RMTS task at an age at which their English-speaking counterparts fail (English speakers in the Walker et al. study performed at chance in both same (46\%) and different ( $43 \%$ ) conditions). ${ }^{2}$ Although this outcome is consistent with an account in which verb-focused word learning biases children toward relational solutions, it is also possible that Mandarin speakers succeed at the task for more general reasons (attention, etc.), without having a relational bias.

We can discriminate these two possibilities by examining bias in an ambiguous task, with both relational and object matches available, and no definitive correct answer. If Mandarin-speaking toddlers succeeded in Experiment 1 because of a general aptitude for test-taking, and not a specific bias toward relations, then they should respond at chance in a modified RMTS with no correct answer. Additionally, if an object bias is responsible for the poor performance of English speakers (and not just random responding), then we should observe systematic preferences for object matches when there is no conflicting evidence for relations. In Experiment 2, we assess these possibilities.

## Experiment 2: Comparing relational and object focus across cultures

Experiment 2 tests for baseline differences in bias toward relational or object-based hypotheses across Mandarin and English speakers. To do this, we created an ambiguous paradigm, in which it is unclear whether a particular object or the relationship between objects is causal. Specifically,

[^298]we presented children with a "different is causal" condition, in which the same object appears in each of the causal pairs (see Figure 3). In this case, it is perfectly reasonable to infer that either the individual object (i.e., the blue square) or the relation (i.e., different) produced the effect. We pit these options against each other by presenting the same objects in the test pairs. The individual objects come together to create a "same" pair-which is correct with respect to the object hypothesis, but incorrect with respect to the relational hypothesis, and the other objects associated with the effect come together to create a "different" pair-which is correct with respect to the relational hypothesis and incorrect with respect to the individual object hypothesis.

If a focus on verbs in early language learning induces a bias toward relational hypotheses, we should observe a tendency toward relational solutions in Mandarin-speaking toddlers, and a converse bias toward objects in nounfocused English-speaking toddlers.

## Methods

Participants. A total of 112 3-year-olds participated in Experiment 2, 56 native Mandarin speakers ( $M=41.4$ months; 28 female) and 56 native English speakers ( $M=$ 41.4 months; 21 female). An additional 11 children were tested but excluded as a result of experimenter error or failure to complete the study. Mandarin-speaking children were recruited and tested at preschools in China, and English speakers at preschools and museums in the U.S. In all settings, children were tested individually with the experimenter in a private room.
Materials and procedure. Materials were identical to those in Experiment 1, and the procedure closely resembled that of the "different" condition, but with several modifications to create an ambiguous causal structure (see Figure 3).


Test Trial
Figure 3: Schematic illustration of ambiguous training and test trials in Experiment 2, in which the evidence was consistent with both object and relational solutions.

First, one of the blocks (represented by the blue square in Figure 3) appears in both different pairs. This reoccurring block provides the object-based hypothesis (i.e., the blue
square is causal). Second, the test trial included two pairs composed of blocks that were previously observed in the "different" training pairs. Finally, due to the constraints of the study design, it was not possible to present an ambiguous same condition. As a result, Experiment 2 only included the different condition. As in the previous study, the experimenter asked the child to choose the pair that would activate the machine. The child's first point or reach was scored as consistent with either an object selection or a relational selection.

## Results and discussion

Given an ambiguous choice between object and relational matches, English-speaking preschoolers selected the object match ( $64 \%$; two-tailed binomial $p=.044$ ) and Mandarinspeaking preschoolers chose the relational match ( $66 \%$; two-tailed binomial $p=.022$; see Figure 4).


Figure 4: Proportion of object and relational matches selected by English- and Mandarin-speaking toddlers in Experiment 2. Error bars indicate $95 \%$ confidence intervals.

## General discussion

In two experiments, we find that Mandarin-speaking children tend to privilege relations whereas Englishspeaking children tend to privilege individual objects, often missing the abstract relation.

In Experiment 1, we evaluated whether the noun focus in English word learning can account for the dip in relational reasoning observed in English-speaking preschoolers. To do so, we examined relational reasoning in Mandarin-speaking preschoolers, whose early language learning is more focused on verbs. Consistent with the noun-focus account, we found that Mandarin-speaking preschoolers substantially outperform their English-speaking peers in identifying shared relational structure in the RMTS.

In Experiment 2, we tested for the key factor predicted to mediate the relationship between language and RMTS performance. This study explored whether English- and Mandarin-speaking preschoolers exhibit differing biases toward relational and object-based solutions. Indeed, we found that in an ambiguous context with no correct answer, Mandarin speakers tend to favor solutions consistent with
relational hypotheses and English speakers show a contrasting object bias.

It is important to note that while English-speaking preschoolers have often exhibited poor performance in relational tasks of the same format, their consistent selection of object-based matches in this experiment is not trivial. Choosing the object match may indeed present a more challenging cognitive task. In order to select the object match at test, children must track and remember the relevant object (the blue square) throughout the training trials, which (perhaps counterintuitively) increases the cognitive load compared with learning the abstract relation, which does not require tracking of any particular objects. Accordingly, this outcome demonstrates a surprising competence on the part of English-speaking preschoolers, which may also be attributable to their noun-centric language learning.

Taken together, these findings inform potential sources of bias in early learning and the development of relational reasoning. In particular, they rule out the possibility that language learning in general produces an object bias. Instead, we show that preschoolers of the same age in different linguistic and cultural contexts may have varying degrees of relational and object focus, and that these differences correlate with robust population-level differences in relational reasoning.

Our findings stand in contrast to the suggestion that language plays a paradoxical role in relational development, by both hindering relational reasoning and facilitating it (Hoyos et al., 2016). Although this may be true in nounfocused languages, like English, it does not appear to be a general feature of language learning.

Furthermore, we suggest that language may well act to hinder and facilitate relational reasoning, without the need to view this phenomenon as a paradox. Instead, it is possible that the object bias and the associated dip in relational reasoning observed in English speakers result from general learning processes with no exceptional role for language. Instead, the structures and features of language may be interpreted as some of many sources of input informing the types of concepts that are privileged during early learning.

Of course, several questions remain regarding the source of the population differences observed here. For example, it is certainly possible that cultural factors (other than language) play a role in facilitating a relational focus in Mandarin speakers. Indeed, there are well-documented differences between collectivist and individualist cultures, which may similarly result in an emphasis on relationships between entities or on characteristics of individual entity kinds (e.g., Chiu, 1972; Choi, Nisbett, \& Norenzayan, 1999; Nisbett, Peng, Choi, \& Norenzayan, 2001; Oyserman \& Lee, 2008; Peng \& Knowles, 2003). Our ongoing research is aimed at further pulling these hypotheses apart.

That said, regardless of whether language, culture, or some combination of the two is ultimately responsible for these effects, the current findings demonstrate that preschoolers have the capacity to infer relational properties, providing additional evidence that the object bias is learned after early competence in relational reasoning is achieved
(Walker et al., 2016). More broadly, we have established population-level differences in relational focus that occur naturally across cultures early in development and predict the developmental trajectory of relational reasoning.

## Acknowledgments

We thank Gail Heyman, Kang Lee, Jing Zhang, Caitlyn Lanigan, and Amy Zou for invaluable assistance with data collection. This work was supported by the Ammodo KNAW Award and the NSF Graduate Research Fellowship Program under grant DGE 1106400 to AC.

## References

Chiu, L. H. (1972). A cross-cultural comparison of cognitive styles in Chinese and American children. International journal of psychology, 7(4), 235242.

Choi, I., Nisbett, R. E., \& Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. Psychological bulletin, 125(1), 47.
Choi, S., \& Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. Journal of child language, 22(03), 497-529.
Christie, S. \& Gentner, D. (2007). Relational similarity in identity relation: The role of language. In S. Vosniadou \& D. Kayser (Eds.), Proceedings of the second European cognitive science conference. London, UK: Taylor \& Francis.
Christie, S. \& Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. Journal of Cognition and Development, 11(3), 356-373.
Christie, S., \& Gentner, D. (2014). Language helps children succeed on a classic analogy task. Cognitive Science, 38, 383-397.
Colunga, E., \& Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. Psychological review, 112(2), 347-382.
Ferry, A., Hespos, S., \& Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. Child Development, 86(5), 1386-1405.
Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child Development, 59, 47-59.
Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought (pp.195-235). Cambridge, MA: MIT Press.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34(5), 752-775.

Gentner, D., \& Medina, J. (1998). Similarity and the development of rules. Cognition, 65, 263-297.
Gentner, D., \& Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman \& J. P. Byrnes (Eds.), Perspectives on thought and language: Interrelations in development (pp. 225-277). London: Cambridge University Press.
Gopnik, A., \& Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. Psychological Bulletin, 138(6), 1085.
Hall, D. G., \& Waxman, S. R. (1993). Assumptions about word meaning: Individuation and basic-level kinds. Child Development, 64, 1550-1570.
Hoyos, C., Shao, R., \& Gentner, D. (2016). The paradox of relational development: Could language learning be (temporarily) harmful? D. Grodner, D. Mirman, A. Papafragou, J. Trueswell, J. Novick, S. Arunachalam, S. Christie, \& C. Norris (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental science, 10(3), 307-321.
Nisbett, R. E., Peng, K., Choi, I., \& Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. Psychological review, 108(2), 291.
Oyserman, D., \& Lee, S. W. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. Psychological bulletin, 134(2), 311.
Peng, K., \& Knowles, E. D. (2003). Culture, education, and the attribution of physical causality. Personality and Social Psychology Bulletin, 29(10), 1272-1284.
Penn, D. C., Holyoak, K. J., \& Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behavioral and Brain Sciences, 31(2), 109-130.
Tardif, T., Shatz, M., \& Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. Journal of Child Language, 24(03), 535-565.
Tyrrell, D. J., Stauffer, L. B., \& Snowman, L. G. (1991). Perception of abstract identity/difference relationships by infants. Infant Behavior Development, 14 (1), 125-129.
Walker, C.M., Bridgers, S., \& Gopnik, A. (2016). The early emergence and puzzling decline of relational reasoning: Effects of knowledge search on inferring abstract concepts. Cognition, 156, 30-40.
Walker, C. M., \& Gopnik, A. (2014). Toddlers infer higherorder relational principles in causal learning. Psychological Science, 25, 161-169.

# Grammar-Based and Lexicon-Based Techniques to Extract Personality Traits from Text 

Maira B. Carvalho (m.brandaocarvalho@uvt.nl)<br>Tilburg Center for Cognition and Communication<br>Tilburg University<br>Tilburg, The Netherlands<br>Max M. Louwerse (mlouwerse@uvt.nl)<br>Tilburg Center for Cognition and Communication<br>Tilburg University<br>Tilburg, The Netherlands


#### Abstract

Language provides an important source of information to predict human personality. However, most studies that have predicted personality traits using computational linguistic methods have focused on lexicon-based information. We investigate to what extent the performance of lexicon-based and grammarbased methods compare when predicting personality traits. We analyzed a corpus of student essays and their personality traits using two lexicon-based approaches, one top-down (Linguistic Inquiry and Word Count (LIWC)), one bottom-up (topic models) and one grammar-driven approach (Biber model), as well as combinations of these models. Results showed that the performance of the models and their combinations demonstrated similar performance, showing that lexicon-based topdown models and bottom-up models do not differ, and neither do lexicon-based models and grammar-based models. Moreover, combination of models did not improve performance. These findings suggest that predicting personality traits from text remains difficult, but that the performance from lexiconbased and grammar-based models are on par.


Keywords: language; personality; traits; machine learning; computational linguistics; lexicon-based; grammar-based

## Introduction

In our daily interactions, we guide our behavior towards other people using information that is collected throughout these interactions, but also using knowledge about the world and social groups (Rich, 1979). These judgments are oftentimes made unconsciously.

Models of users' behavior, thinking and feeling typically rely on the personality traits that can be identified (McCrae \& John, 1992). Trait theory is an approach to the study of human personality in which it is believed that humans exhibit habitual patterns of behavior, thought and emotion. It is presumed that there is a relatively small number of dimensions that can be used to describe personality (O’Connor, 2002). Independent analyses have consistently yielded five broad dimensions, called the Big Five (or Five Factor Model): openness to experience, conscientiousness, extraversion, agreeableness and neuroticism (McCrae \& John, 1992).

Personality traits are generally identified on the basis of data collected from the users who fill out standardized questionnaires. However, such an approach has certain drawbacks. Firstly, it can be costly for the researcher and timeconsuming for the user (Gauch, Speretta, Chandramouli, \& Micarelli, 2007). Secondly, people are not reliable sources
of information about themselves: there is evidence to suggest that self-descriptions are heavily influenced by the social groups in which a person finds himself (McGuire \& PadawerSinger, 1976), and dissimulation can be a problem in selfreports (Wright, 2014).

Automatic inference of personality offers the advantages of being less intrusive and possibly more environmentally valid. Indeed, a range of studies have investigated the extent to which personality traits can be predicted from user behavior. The lion's share of these studies use linguistic data as sources of information, both language and speech (Beukeboom, Tanis, \& Vermeulen, 2012; Gawda, 2009; Mairesse, Walker, Mehl, \& Moore, 2007; Mehl, Robbins, \& Holleran, 2012; Oberlander \& Gill, 2006; Oberlander \& Nowson, 2006). Linguistic data has also shown to indirectly shed light on personality. For instance, linguistic cues have shown to be linked to deception (Louwerse, Lin, Drescher, \& Semin, 2010), and to different registers of communication (Louwerse, McCarthy, McNamara, \& Graesser, 2004). Furthermore, a person's emotional state is reflected in language use (Tausczik \& Pennebaker, 2009), not only by explicit lexical content but also by implicit semantic associations (Recchia \& Louwerse, 2014).

In considering the cognitive science literature that aims to extract behavioral information from linguistic data, two approaches can be distinguished. On the one hand, studies use lexical cues to extract information from text, for example emotional expression (Kahn, Tobin, Massey, \& Anderson, 2007), deception (Newman, Pennebaker, Berry, \& Richards, 2003), political orientation (Dehghani, Sagae, Sachdeva, \& Gratch, 2013), moral foundations (Graham, Haidt, \& Nosek, 2009), romantic relationship outcomes (Ireland et al., 2011), among others. On the other hand, extracting behavioral information from explicit lexical information can be problematic. First, in controlled experimental settings it is easy for participants to carefully monitor their semantic content. For instance, in deception studies participants might avoid using specific words. Second, sparsity issues may emerge if algorithms detect specific word use.

An alternative approach lies in using grammar-based cues. By performing a manual analysis of seven syntax markers, Gawda (2009) identified an increased use of certain features
in emotional narratives written by individuals with antisocial personality disorder. Current computational linguistic tools are able to extract many grammar-based linguistic features automatically and efficiently, and it is reasonable to assume that such features could also carry information about personality. Biber (1988) conducted a study on linguistic variation across speech and writing, and computed the frequency of 67 linguistic features (e.g. frequency of auxiliary verbs, pronouns, main verbs, adjectives); he was able to identify different writing genres using naturally occurring word patterns. Graesser, McNamara, Louwerse, and Cai (2004) proposed the tool Coh-Metrix, which allows the user to analyze texts on discourse, cohesion, and world knowledge. The Suite of Linguistic Analysis Tools (SALAT) calculates scores for aspects such as syntactic complexity (Kyle, 2016) and cohesion (Crossley, Kyle, \& McNamara, 2015). These tools open up a range of possibilities for the investigation of the relationships between personality and language use.

In conclusion, two approaches can be identified in extracting personality traits from language use: a lexicon-driven approach and a grammar-driven approach. As pointed above, the majority of cognitive science literature has focused on the lexicon-based approach. The question is to what extent the findings from a grammar-driven approach are comparable with the lexicon-driven approach that currently dominates the literature. We address this question in the current work.

## Extracting personality traits from text

Most studies on extracting personality from text focused on identifying words, collocations and general linguistic features that occur in texts produced by one group of people versus another group, aiming to uncover which features are informative when trying to differentiate the groups. Early attempts relied on word counting and predefined dictionaries that sort words in categories (Tausczik \& Pennebaker, 2009). This approach, albeit basic, has been used in many studies that show links between word usage and certain psychological processes and personalities, e.g. Beukeboom et al. (2012), and Mehl et al. (2012). Other researchers used bottom-up approaches to associate linguistic features with personality types. Oberlander and Gill (2006) collected large corpora of text labeled with the personality of the author and performed stratified corpus comparisons. Interesting findings included the fact that people who scored high in extraversion used more inclusive expressions and connectives, while those with low score were more tentative and used adjectives less frequently. The authors also noted that people with high neuroticism scores had preference for multiple punctuation.

Another approach is to treat the problem as a supervised classification task, employing machine learning techniques to identify the personality of the author of a given text. Oberlander and Nowson (2006) investigated a corpus of weblog posts from 71 participants, who completed a personality questionnaire online as part of the study. The authors used Support Vector Machine (SVM) classifiers and feature sets consist-
ing of n-grams extracted from the text and selected according to different levels of restriction. The same approach was later applied to a larger sample of bloggers (Iacobelli, Gill, Nowson, \& Oberlander, 2011). Argamon, Dhawle, Koppel, and Pennebaker (2005) used SVMs and four sets of lexical features to differentiate high and low extraversion and neuroticism, using a corpus of around 2400 student essays and personality assessments, collected by Pennebaker and King (1999). Mairesse et al. (2007) also worked on the same corpus, employing a series of classification and regression techniques and features from both the Linguistic Inquiry and Word Count (LIWC) and the Medical Research Council (MRC) Psycholinguistic Database. Their results confirmed previous findings and reveal new correlations between linguistic markers and personality, such as use of swear words and use of pronouns. As for the accuracy of automatic classification, the authors reported accuracies that are, according to their evaluation, significantly above chance; however, it is not clear whether these values are high enough to be useful in real applications (Mairesse et al., 2007).

Following the work by Mairesse et al. (2007), the task of automatic identification of personality from text gained a lot of attention from the research community, mostly due to the Workshop on Computational Personality Recognition (Celli, Pianesi, Stillwell, \& Kosinski, 2013). As part of a shared task, the organizers made available two datasets of text labeled with the personality traits of the authors - including the Essay Corpus by Pennebaker and King (1999). As a result, many researchers tackled the problem with different learning algorithms (e.g. Naive Bayes, SVM, kNN, ensemble methods, logistical regression) and using different features such as n-grams, LIWC, MRC, lexical nuances, part-of-speech tags, emotional values from the AFINN database, word intensity scale, sentiment analysis and word associations to emotions (Celli et al., 2013).

Although the results from these attempts are encouraging, it had been noted that top-down approaches based on lexical resources seem to perform better than bottom-up approaches based only on words or n-grams (Celli et al., 2013). Nevertheless, there are benefits in employing approaches that do not rely on pre-defined vocabularies, for example allowing exploration of topics not previously considered, easier application in different genres and languages, and saving the effort of creating the word lists (Schwartz et al., 2013).

Schwartz et al. (2013) used a large dataset of Facebook posts (over 15.4 million Facebook messages collected from 75 thousand volunteers) to perform an open-vocabulary analysis of correlations between personality types and vocabulary use. The goal of the work was to discover unexpected relationships that would not necessarily be evident from using pre-defined word categories. Although the focus of the work was mainly to explore and gain insights on the data, the authors also used the approach to predict personality from text, with results that are comparable to previous literature. Liu, Wang, and Jiang (2016) also attempted to predict personality
from text while avoiding to rely on predefined vocabularies. The authors proposed a model that expands latent Dirichlet allocation (LDA) to include the assumption that topic distribution depends not only on the characteristics of the corpus itself, but also on the five personality traits of writers. However, the topics identified by their model seem to be affected disproportionally by individuals with less common personality combinations, and for this reason the model must be trained with a massive, representative corpus for which the personality of the writers is known. Since obtaining such corpora is difficult, the applicability of their approach seems to be limited.

In this paper, we investigate how we can predict user personality from written text by using features that do not rely on closed-vocabularies, and compare the results to the state of the art. In the next section, we describe our approach.

## Procedure

## Dataset

We use the Essay Corpus (Pennebaker \& King, 1999), which consists of 2468 essays, written by introductory psychology students of the University of Texas as part of their course assignments. The students also completed the Five Factor Inventory personality questionnaire (John, Donahue, \& Kentle, 1991), so that all essays could be marked with five personality scores for each Big-5 trait (Openness to Experience, Conscientiousness, Extraversion, Agreableness, Neuroticism). In addition to the scores, the corpus contains binary values for each trait (high/low), which were obtained using a median split over the scores. The class distribution of the binary values is shown in Table 1.

Table 1: Class distribution in dataset.

|  | OPN | CON | EXT | AGR | NEU |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Low | 1196 | 1214 | 1191 | 1158 | 1235 |
|  | $(48.46 \%)$ | $(49.19 \%)$ | $(48.26 \%)$ | $(46.92 \%)$ | $(50.04 \%)$ |
| High | 1272 | 1254 | 1277 | 1310 | 1233 |
|  | $(51.54 \%)$ | $(50.81 \%)$ | $(51.74 \%)$ | $(53.08 \%)$ | $(49.96 \%)$ |

## Features

We employed four groups of feature sets, which were chosen to investigate to what extent the performance of lexiconbased and grammar-based computational linguistic methods are comparable.

For the lexicon-based features, we used two main approaches: top-down (LIWC and MRC) and bottom-up (topic modeling with latent Dirichlet allocation (LDA)). For the grammar-based features, we selected the original Biber features (Biber, 1988).

## Lexicon-based, top-down

1) A total of 80 LIWC features were extracted using the LIWC2007 software, which outputs relative frequencies of
words found in each pre-defined category, and a few structural features such as word count and words per sentence.
2) The 14 MRC features refer to word length, number of syllables or phonemes, and values for frequency of use, imageability, concreteness, meaning, age of acquisition, among others. The MRC features were calculated by averaging the scores of the essay words found in the database (as opposed to averaging over total word count).

Lexicon-based, bottom-up For topic modeling, we preprocessed the corpus by lemmatizing the words using NLTK's WordNet lemmatizer, and removing non-English words (i.e. words that were not found in Wordnet). Given the relatively small size of the corpus, we did not filter out words based on frequency. Then, we trained three LDA models with different number of topics ( 30,65 and 100). Each document in the dataset was converted to a vector that represents the proportion in which each topic appears in the document. To train the model, we used the library Gensim, with 10 passes and default hyperparameters.

To illustrate the topics found in the corpus, these are the ten most relevant words for each of the three most frequently appearing topics extracted by the 30 -topic LDA model: "think, go, get, really, like, write, minute, time, wonder, need"; "go, get, really, time, friend, home, want, much, like, miss"; and "life, people, thing, know, think, time, one, make, feel, way".
Grammar-based, top-down For this study, we use the 67 features selected by Biber (1988) to reflect the linguistic structure of the text. These features primarily operate at the word level, such as parts-of-speech, and fall into categories such as tense and aspect markers, adverbials, pronouns, questions, nominal forms, passives, subordination features, prepositional phrases, coordinations and negations, and so on. These features were extracted from the text using software developed in-house.

Combinations In addition to considering these models separately, we investigated the model combinations in order to determine their complementary value.

## Classifiers

We trained five Support Vector Machine (SVM) classifiers with linear kernel, one for each personality trait. SVMs were chosen due to previous reports of them performing better on this task than other algorithms (Mairesse et al., 2007), and linear kernels were employed to retain interpretability of the model. For the implementation, we used the machine learning library Scikit-learn, which in turn uses an implementation based on Libsvm. The classifiers were trained without parameter tuning (i.e. penalty parameter $\mathrm{C}=1.0$ ).

## Results

Reservations have been expressed by the scientific community on the application of null-hypothesis statistical testing for comparison of machine learning algorithms for many rea-

Table 2: Average accuracy, precision and recall for each classifier, with $95 \%$ confidence interval.

|  |  | Baseline | Lexicon, top-down |  | Lexicon, bottom-up |  |  | Grammar | Combinations |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | N |  | $\stackrel{\text { Even }}{\stackrel{U}{U}}$ | $\stackrel{\stackrel{\rightharpoonup}{\mathrm{e}}}{\stackrel{\varepsilon}{8}}$ | $\frac{\sqrt{6}}{8}$ | $\stackrel{\Im}{8}$ | © |  | 为雨 |
| Openness | Acc | . 515 | . $617 \pm .006$ | . $602 \pm .006$ | . $613 \pm .005$ | . $613 \pm .006$ | . $602 \pm .006$ | . $576 \pm .007$ | . $606 \pm .006$ | . $602 \pm .006$ |
|  | P | . 515 | . $634 \pm .006$ | . $617 \pm .006$ | . $633 \pm .006$ | . $630 \pm .006$ | . $603 \pm .005$ | . $586 \pm .007$ | . $619 \pm .006$ | . $615 \pm .006$ |
|  | R | 1. | . $611 \pm .010$ | . $605 \pm .010$ | . $593 \pm .010$ | . $603 \pm .010$ | . $669 \pm .009$ | . $607 \pm .010$ | . $614 \pm .009$ | . $612 \pm .010$ |
| Conscientiousness | Acc | . 508 | . $547 \pm .006$ | . $551 \pm .006$ | . 5556.007 | . $544 \pm .006$ | . $534 \pm .006$ | . $551 \pm .006$ | . $546 \pm .006$ | . $548 \pm .007$ |
|  | P | . 508 | . $550 \pm .005$ | . $553 \pm .006$ | . $553 \pm .006$ | . $546 \pm .005$ | . $536 \pm .006$ | . $554 \pm .006$ | . $550 \pm .006$ | . $551 \pm .006$ |
|  | R | 1. | . $607 \pm .009$ | . $611 \pm .010$ | . $653 \pm .011$ | . $610 \pm .010$ | . $615 \pm .011$ | . $599 \pm .010$ | . $588 \pm .010$ | . $598 \pm .009$ |
| Extraversion | Acc | . 517 | . $562 \pm .007$ | . $545 \pm .006$ | . $546 \pm .006$ | . $555 \pm \pm .006$ | . $551 \pm .006$ | . $546 \pm .007$ | . $553 \pm .007$ | . $557 \pm .007$ |
|  | P | . 517 | . $570 \pm .006$ | . $554 \pm .005$ | . $544 \pm .004$ | . $551 \pm .004$ | . $550 \pm .004$ | . $553 \pm .006$ | . $562 \pm .006$ | . $563 \pm .006$ |
|  | R | 1. | . $624 \pm .010$ | . $624 \pm .010$ | . $764 \pm .009$ | . $756 \pm .010$ | . $731 \pm .011$ | . $635 \pm .010$ | . $614 \pm .010$ | . $644 \pm .010$ |
| Agreeableness | Acc | . 531 | . $545 \pm .007$ | . $552 \pm .007$ | . $557 \pm .005$ | . $548 \pm .006$ | . $542 \pm .006$ | . $549 \pm .006$ | . $552 \pm .006$ | . $553 \pm .007$ |
|  | P | . 531 | . $561 \pm .006$ | . $567 \pm .005$ | . $554 \pm .003$ | . $555 \pm \pm .004$ | . $546 \pm .004$ | . $562 \pm .005$ | . $570 \pm .005$ | . $570 \pm .006$ |
|  | R | 1. | . $651 \pm .011$ | . $664 \pm .010$ | . $843 \pm .007$ | . $758 \pm .012$ | . $821 \pm .009$ | . $678 \pm .010$ | . $636 \pm .010$ | . $649 \pm .010$ |
| Neuroticism | Acc | . 500 | . $565 \pm .007$ | . $571 \pm .006$ | . $532 \pm .007$ | . $527 \pm .007$ | . $520 \pm .006$ | . $545 \pm .006$ | . $552 \pm .006$ | . $545 \pm .007$ |
|  | P | 0. | . $563 \pm .007$ | . $571 \pm .007$ | . $529 \pm .006$ | . $527 \pm .008$ | . $519 \pm .006$ | . $545 \pm .006$ | . $552 \pm .006$ | . $544 \pm .006$ |
|  | R | 0. | . $577 \pm .012$ | . $571 \pm .011$ | . $581 \pm .014$ | . $531 \pm .012$ | . $540 \pm .011$ | . $535 \pm .010$ | . $551 \pm .011$ | . $550 \pm .011$ |

sons, not the least of which the fact that any difference between two algorithms, no matter how small, can be shown to be statistically significant, provided that enough data are used (Japkowicz \& Shah, 2011). For this reason, instead of traditional hypothesis testing, we chose to adopt error-estimation techniques to obtain relatively robust estimates of the performance of the algorithms, which in turn allows us to compare the results considering their practical differences.

Table 2 shows the performance scores of the classifiers trained using the eight different sets of features discussed earlier. We focus our discussion around accuracy (Acc), but we also report precision $(\mathrm{P})$ and recall $(\mathrm{R})$ to give a better overall indication of the performance of the classifiers ${ }^{1}$. The performance of a simple majority classifier (i.e. it always predicts the class with the highest number of instances) is used as baseline. We report the estimated mean of the scores, calculated by running $10 \times 10$-fold cross-validation, using all 100 individual scores to estimate the mean and variance, and using 10 degrees of freedom to calculate the $95 \%$ confidence interval, as suggested by Bouckaert (2003).

As can be seen in Table 2, the performance scores of the classifiers vary for different traits, with the best accuracies ranging from approximately $56 \%$ for Agreeableness and Conscientiousness to $62 \%$ for Openness to experience (accuracies are highlighted in the table, and the highest accuracy scores for each trait are marked in bold). Nevertheless, we can make some general observations on the overall performance of the different sets of features, which we list below.

Confirming previous findings, top-down lexicon-based ap-

[^299]proaches generally provide the best accuracies. The top-down approach proposed in the literature, which uses MRC features in addition to LIWC, does provide a small added value for classifying Extraversion and Openness to experience. Conversely, for the other three traits, MRC features do not seem to provide any real improvement.

We note that bottom-up lexicon-based approaches can offer comparable accuracies to top-down approaches, with performance being basically equivalent among top-down and bottom-up over all traits but Neuroticism. Furthermore, the number of topics matters, as accuracy degrades with 100 topics (when the features are likely to become more sparse).

We can observe that a grammar-based approach on its own seems to give a slightly worse accuracy than lexicon-based approaches for three of the five traits: around $2 \%$ less accurate for Neuroticism and Extraversion, and 4\% less accurate for Openness to experience. Nevertheless, for the other two traits (Conscientiousness and Agreeableness), the accuracies are basically the same.

Finally, combining grammar and lexicon approaches does not lead to significant improvements in accuracy. In fact, it even seems to degrade the results of the top-down lexiconbased approach slightly.

In summary, Table 2 shows that lexicon-based top-down features and bottom-up features do not seem to differ in a practical way, and while grammar-based features seem to have slightly worse accuracies than lexicon-based features, the difference can be considered too small to be of practical significance. Furthermore, the accuracies of our proposed sets of features are on par with the results obtained by previous studies.

## General Discussion

The differences in performance between the different algorithms are very small. Using different feature sets yields similar results, and combining different features does not improve the performance in any meaningful way.

One possible reason could be a floor effect, in which the questionnaire used to assess personality traits in this corpus would not be able to distinguish reliably between subjects at the lower end of the scale. This is unlikely, since the test used in this corpus is a standardized questionnaire that has been validated and used in numerous studies (John et al., 1991).

It is also possible that the use of self-assessments of personality makes this task particularly difficult due to the potential unreliability of self-reports, as discussed in the introduction. Future investigation could incorporate personality assessments made by human observers, to evaluate to which extent self-assessment and observed scores differ, and whether the algorithms could match the performance of human judges. Furthermore, replicating the study with other corpora could also indicate whether different text types could be more suitable for detecting certain personality traits.

In this study, we used a relatively limited set of non-topdown features, namely the features proposed by Biber (1988) and topic modeling. Future work could investigate whether applying other grammar-based and bottom-up lexicon-based features (e.g. cohesion, syntactic complexity, n-grams, skip grams, Word2Vec, semantic similarities) would result in better performances. In addition, we could try to improve the models by using non-linear kernels, performing parameter tuning, and employing ensemble machine learning methods for combining different sets of features.

However, the difficulty of identifying personality traits from text could signal a more fundamental issue. Mischel and Shoda (1995) have argued that individual differences in social behaviors are actually variable across different situations (situationism), and not completely stable as it is proposed by trait theory. As such, if personality scales and textual analyses tap into different social situations, tasks that use questionnaire scores as gold standard will not be able to achieve acceptable performance. Further research is needed to investigate this hypothesis, and whether other stable patterns of behavior could be used as gold standard for automatic personality inferences.

The current study has used the most common personality traits classification, the Big Five, and the most commonly used corpus to identify personality traits, the Essay Corpus, in order to compare the difference between top-down and bottom-up lexicon-based and grammar-based computational linguistic techniques. Our findings show that no differences were obtained between lexicon-based and grammar-based or between top-down and bottom-up approaches, nor complementary advantages for combinations of models, despite the fact that all methods were on par with the performance previously reported.

## Acknowledgments

We would like to thank James W. Pennebaker for providing us access to the original personality scores of the Essay Corpus.

## References

Argamon, S., Dhawle, S., Koppel, M., \& Pennebaker, J. W. (2005). Lexical predictors of personality type. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America.
Beukeboom, C. J., Tanis, M., \& Vermeulen, I. E. (2012). The language of extraversion: Extraverted people talk more abstractly, introverts are more concrete. Journal of Language and Social Psychology, 32(2), 191-201. doi: 10.1177/0261927x12460844

Biber, D. (1988). Variation across speech and writing. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. In Proceedings of the 20th International Conference on Machine Learning (ICML-03) (pp. 51-58).
Celli, F., Pianesi, F., Stillwell, D., \& Kosinski, M. (2013). Workshop on computational personality recognition (shared task). In Proceedings of the Workshop on Computational Personality Recognition.
Crossley, S. A., Kyle, K., \& McNamara, D. S. (2015). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior Research Methods, 48(4), 1227-1237. doi: 10.3758/s13428-015-0651-7

Dehghani, M., Sagae, K., Sachdeva, S., \& Gratch, J. (2013). Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque". Journal of Information Technology \& Politics, 11(1), 1-14. doi: $10.1080 / 19331681.2013 .826613$
Gauch, S., Speretta, M., Chandramouli, A., \& Micarelli, A. (2007). User profiles for personalized information access. In The adaptive web (pp. 54-89). Springer. doi: 10.1007/978-3-540-72079-9_2

Gawda, B. (2009). Syntax of emotional narratives of persons diagnosed with antisocial personality. Journal of Psycholinguistic Research, 39(4), 273-283. doi: 10.1007/s10936-009-9140-4

Graesser, A. C., McNamara, D. S., Louwerse, M. M., \& Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, \& Computers, 36(2), 193-202. doi: 10.3758/bf03195564
Graham, J., Haidt, J., \& Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96(5), 1029-1046. doi: 10.1037/a0015141
Iacobelli, F., Gill, A. J., Nowson, S., \& Oberlander, J. (2011). Large scale personality classification of bloggers. In Affective computing and intelligent interaction (pp. 568-577). Springer. doi: 10.1007/978-3-642-24571-8_71

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., \& Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. Psychological Science, 22(1), 39-44. doi: 10.1177/0956797610392928

Japkowicz, N., \& Shah, M. (2011). Evaluating learning algorithms: a classification perspective. Cambridge University Press.
John, O. P., Donahue, E. M., \& Kentle, R. L. (1991). The Big Five inventory - versions $4 a$ and 54. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.
Kahn, J. H., Tobin, R. M., Massey, A. E., \& Anderson, J. A. (2007). Measuring emotional expression with the linguistic inquiry and word count. The American Journal of Psychology, 120(2), 263-286.
Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. Doctoral dissertation, Georgia State University.
Liu, Y., Wang, J., \& Jiang, Y. (2016). PT-LDA: A latent variable model to predict personality traits of social network users. Neurocomputing, 210, 155-163. doi: 10.1016/j.neucom.2015.10.144

Louwerse, M. M., Lin, K.-I., Drescher, A., \& Semin, G. (2010). Linguistic cues predict fraudulent events in a corporate social network. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 961-966). Austin, TX: Cognitive Science Society.
Louwerse, M. M., McCarthy, P. M., McNamara, D. S., \& Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, \& T. Regier (Eds.), Proceedings of the 26th Annual Conference of the Cognitive Science Society (pp. 843-848). Austin, TX: Cognitive Science Society.
Mairesse, F., Walker, M. A., Mehl, M. R., \& Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research, 30, 457-500. doi: 10.1613/jair. 2349

McCrae, R. R., \& John, O. P. (1992). An introduction to the five-factor model and its applications. Journal of Personality, 60(2), 175-215. doi: 10.1111/j.14676494.1992.tb00970.x

McGuire, W. J., \& Padawer-Singer, A. (1976). Trait salience in the spontaneous self-concept. Journal of Personality and Social Psychology, 33(6), 743-754.
Mehl, M. R., Robbins, M. L., \& Holleran, S. E. (2012). How taking a word for a word can be problem-
atic: Context-dependent linguistic markers of extraversion and neuroticism. Journal of Methods and Measurement in the Social Sciences, 3(2), 30-50. doi: 10.2458/azu_jmmss_v3i2_mehi

Mischel, W., \& Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychological Review, 102(2), 246-268. doi: 10.1037/0033-295X.102.2.246

Newman, M. L., Pennebaker, J. W., Berry, D. S., \& Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin, 29(5), 665-675. doi: 10.1177/0146167203029005010
Oberlander, J., \& Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. Discourse Processes, 42(3), 239270. doi: $10.1207 / \mathrm{s} 15326950$ dp4203_1

Oberlander, J., \& Nowson, S. (2006). Whose thumb is it anyway?: classifying author personality from weblog text. In Proceedings of COLING/ACL (pp. 627-634). Sidney, Australia.
O'Connor, B. P. (2002). A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. Assessment, 9(2), 188-203. doi: 10.1177/1073191102092010

Pennebaker, J. W., \& King, L. A. (1999). Linguistic styles: language use as an individual difference. Journal of Personality and Social Psychology, 77(6), 1296-1312. doi: 10.1037/0022-3514.77.6.1296

Recchia, G., \& Louwerse, M. M. (2014). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. The Quarterly Journal of Experimental Psychology, 68(8), 1584-1598. doi: 10.1080/17470218.2014.941296

Rich, E. (1979). User modeling via stereotypes. Cognitive Science, 3(4), 329-354. doi: 10.1207/s15516709cog0304_3

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS ONE, 8(9). doi: 10.1371/journal.pone. 0073791

Tausczik, Y. R., \& Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24-54. doi: 10.1177/0261927x09351676
Wright, W. R. (2014). Personality profiling from text and grammar. In User Modeling, Adaptation, and Personalization (pp. 502-507). Springer. doi: 10.1007/978-3-319-08786-3_47

# Decomposability and Frequency in the Hindi/Urdu Number System 

Chundra Aroor Cathcart (chundra.cathcart@ling.lu.se)<br>Department of Linguistics and Phonetics Lund University


#### Abstract

Hindi/Urdu (HU) numbers 10-99 are highly irregular, unlike the transparent systems of most languages. I investigate the morphological decomposability of HU numbers using a series of computational models. While these models classify most forms accurately, problems are encountered in high-frequency forms of low cardinality, suggesting that some HU numbers are more transparent (i.e., morphologically decomposable) than others. These results are compatible with a dual-route access model proposed for the processing of numeral forms.


Keywords: numerals; computational modeling; Bayesian learning; Hindi/Urdu; phonology; morphology

## Introduction

Hindi/Urdu (HU, officially considered separate languages but differing in little other than orthography and high-register vocabulary) and most other modern Indic languages are unusual in that the numbers through 99 are highly opaque and irregular, undoubtedly posing difficulties in production and processing for language users. This phenomenon is understudied, and raises interesting questions regarding the design principles of cross-linguistic number systems, as well as the processing of complex morphological forms by language users.

In this paper, I investigate the mechanisms by which HU users extract structure and meaning from HU number terms. I address this issue using a series of unsupervised computational models designed to approximate HU users' processing of the numerals $10-99$. While it is generally agreed that number terms are acquired as individual lexical items, there is good reason to hypothesize that many numbers above a certain threshold of frequency are not accessed as individual lexemes during processing, but rather via their component parts, in line with a dual-route model of lexical storage and access (cf. Baayen, 1993). I expect that despite the irregularity of the HU number system, users can find regular patterns in various cues to numerical identity in the input, particularly in lowfrequency numbers, thus facilitating easier comprehension.

My methodology investigates the extent to which HU numbers can be morphologically decomposed. Brysbaert (2005) tentatively proposes a dual-route access model for English numeral storage, hypothesizing that frequent, opaque items like twelve are accessed directly, while morphologically transparent numbers of lower frequency (e.g., eighty-nine) are processed through decomposition. In line with this view, I predict that less frequent HU numbers can be segmented and labeled more accurately by a computational model, indicating greater morphological transparency.

I find that a model using $n$-grams as phonological features successfully assigns most HU numeral forms to the proper TENS/DIGITS cohort, but that, rather unsurprisingly, some highly opaque forms are misclassified. Major errors occur
among numbers of lower magnitude. Since these forms are highly frequent, this state of affairs is compatible with a dualroute account of processing. I find that in general, the model faces difficulties in capturing relationships between simplex (i.e., monomorphemic) forms (e.g., /วssi/ '80') and their complex counterparts (e.g., /corasi/ '84'), where a more sophisticated model of phonology might succeed. These results provide an important baseline for future investigations into mental representations of HU numerals.

## Background

The full list of numerals (taken from Comrie, n.d.) is given in Table 1. When encountering a datum like /bəjalis/ ' 42 ', listeners must infer the value of the TENS and DIGITS place with the aid of cues in the input, and must be able to contend with highly noisy allomorphy: $\operatorname{TENS}\{40\}$ and DIGITs $\{2\}$ have multiple surface realizations. In some cases, this allomorphy is suppletive (i.e., variants bear no phonological resemblance to each other). Listeners may possess the knowledge that HU is head-final, and that higher-order numerical information generally occurs closer to the root (Hurford, 1987), i.e., to the right. For some numerals, it seems plausible that high frequency facilitates access; for instance, $\mathrm{HU} /$ sola/ ' 16 ' is quite unlike other numerals with the feature digits $\{6\}$, all of which are $/ \mathrm{c}^{\mathrm{h}} /$-initial. This is a diachronic artifact; /sola/ faithfully continues Sanskrit sodaśa-, while other forms with DIGITS $\{6\}$ contain reflexes of an unattested dialectal variant *ks(v)aṭ- of attested Sanskrit saṣ- '6' (Turner, 1962-1966). It is also the only member of the teens which shows $/ 1 /$ in its allomorph of /das/ 'ten'. All the same, it may be used frequently enough that this twofold suppletion does not pose problems to speakers and listeners.

A major attempt to explore synchronic regularities among HU numbers is that of Bright (1969), who concludes that despite a lack of economy, implicit rules governing the system are available to language users. Berger (1992) outlines the complex historical development of HU numbers; sporadic phonological reduction, analogy, and language contact, among other phenomena, have resulted in a highly irregular and opaque system compared to the relatively transparent numbers of Sanskrit, HU's ancestor. These works aside, many aspects of the HU numeral system remain untreated.

## Representational issues

## Abstract representation of $\mathbf{H U}$ numerals

Above, I adopt the canonical abstract numerical representation found in much of the literature, where each surface form comprises two underlying factors corresponding to the TENS and DIGITS place. I make the assumption that DIGITs $\{0\}$

Table 1: HU numbers 1-99; rows represent the tens place, columns the digits place

|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | - | ek | do | tin | car | pãc | $\mathrm{c}^{\mathrm{h}} \varepsilon$ | sat | $a t^{\text {h }}$ | no |
| 10 | dəs | gjarə | barə | terə | codə | pəndrə | solə | sətrə | ət ${ }^{\text {a }}$ arə | unnis |
| 20 | bis | Ikkis | bais | teis | cobis | paccis | $\mathrm{c}^{\mathrm{h}}$ วbbis | sattais | əttais | vntis |
| 30 | tis | Ikəttis | bettis | taxtis | cõtis | paxtis | $\mathrm{c}^{\mathrm{h}}$ วttis | saxtis | ərtis | vntalis |
| 40 | calis | ıktalis | bəjalis | tãtalis | cavalis | pãtalis | $\mathrm{c}^{\text {}}{ }^{\text {rjadis }}$ | sãtalis | ərtalis | vncas |
| 50 | pacas | ıkjavən | bavan | tirpen | cəuvən | рәсрәп | $\mathrm{c}^{\text {h }}$ әрpən | səttavən | $\partial t^{\text {h }}$ avən | unsət ${ }^{\text {h }}$ |
| 60 | sat ${ }^{\text {h }}$ | Iksət ${ }^{\text {h }}$ | basat ${ }^{\text {h }}$ | tirsət ${ }^{\text {h }}$ | cõsət ${ }^{\text {h }}$ | pazsət ${ }^{\text {h }}$ | $\mathrm{c}^{\mathrm{h}} \mathrm{Ijas}^{\text {a }}{ }^{\text {h }}$ | sorsat ${ }^{\text {h }}$ | ərsət ${ }^{\text {h }}$ | vnhəttər |
| 70 | sattor | Ikhəttər | bəhəttər | trhəttər | cohəttər | pəchəttər | $\mathrm{c}^{\text {h }}$ Ihəttər | səthəttər | $\partial t^{\text {h }}$ həttər | vnjasi |
| 80 | assi | Ikjasi | bojasi | tirasi | corasi | pəcasi | $\mathrm{c}^{\mathrm{h}}$ ıjasi | səttasi | $\partial t^{\text {h }}$ asi | nəuasi |
| 90 | nəve | ıkjanve | banve | tiranve | coranve | pəcanve | $\mathrm{c}^{\mathrm{h}}$ rjanve | sattanve | ətt ${ }^{\text {hanve }}$ | nınjanve |

does not map to any overt phonological information. Additionally, for forms such as /vntalis/, there is a mismatch between the abstract representation TENS $\{3\}$ DIGITS $\{9\}$ and the phonological form, since the morpheme representing the tens place closely resembles /calis/ ' 40 ', not /tis/ ' 30 '; this suggests an intermediate calculation TENS $\{4\}$ DIGITS $\{-1\}$. I assume that the representation Digits $\{-1\}$ is an integral part of HU numerical computation and is reflected explicitly in the morphology.

## Surface representation of $\mathbf{H U}$ numerals

Brysbaert hesitates to draw a categorical distinction between transparent and opaque English numerals, citing semitransparent forms like thirteen. Along these lines, I seek to situate HU numerals along a cline between mild and extreme opacity. I quantify a number's transparency or decomposability via the performance of a computational model designed to segment and label HU numbers, both in terms of (1) accuracy of the labeling and (2) low posterior uncertainty.

At the outset, I lack a principled means of separating suppletive and non-suppletive allomorphy found in the system. Numbers 11-18 exhibit three allomorphs for TENS $\{1\}$, /-də/, $/-\mathrm{rr} /$ and /-lə/, all from the diachronic source -daśa-, though synchronic $/ \mathrm{d} / \sim / \mathrm{r} / \sim / \mathrm{l} /$ alternations are not well known in HU. Numbers 49-58 show multiple bases for TENS $\{5\}$, all descended from Sanskrit pañcāśat- '50' but formally very dissimilar. I make no a priori assumptions about the status of suppletive allomorphy in the morphological system, and allow the model to simply group together forms according to the configuration it infers. I do, however, treat digits $\{-1\}$ as a separate morpheme, given its systematic occurrence. Nonparametric models (which assume an unbounded number of underlying morphological labels) may alleviate some of the problems that result from forcing suppletive allomorphs to be classified together, which I set aside for future work.

A model of HU numerical processing should characterize the morphological structure of the data encountered. HU numbers are highly fusional, exhibiting the effects of millennia of phonological and morphological change. As Bright reports, no economical set of rules helps to derive the surface representations from their morphological bases. There is often unpredictable allomorphy between simplex and complex forms of a given decade: for instance, /s/ alternates with /h/
in complex forms of /səttrr/ '70' < Sanskrit saptatí', but not in complex forms of $/ \mathrm{sat}^{\mathrm{h}} /$ ' 60 ' < Sanskrit sasți-; however, the latter decade's complex forms contain a reduced vowel $/ \partial /$, alternating with $/ \mathrm{a} /$. The short vowel and geminate consonant found in /assi/ ' 80 ' alternate with a long vowel and singleton consonant in derived /-asi/, but the short vowel found in /nəve/ does not appear in derived forms.

Despite these challenges, listeners should be able to form a probability distribution over possible morphemes contained in a complex input datum. HU's highly fusional phonology notwithstanding, listeners should be able to approximate the location of morpheme boundaries. This question is a key part of this paper's computational inference, and should be of broad interest to phonological theory, as it has the potential to incorporate a number of strategies for morphological boundary detection. The models introduced in this paper draw morpheme boundaries on the basis of what is most likely under the current parameters of the model, and are dependent on distributional information found in other numerals. This task is easier than that of many types of unsupervised segmentation in that at most one boundary must be located per input datum; however, the model must contend with a wider distribution of allomorphs which must be unified. This model does not use external distributional information for the purpose of segmentation (as do Harris, 1955; Saffran et al., 1996).

A question relevant to this paper concerns the types of morphological segmentation that should be permitted. Crosslinguistically, a morphological segmentation of the type [b][əjalis] might be permissible, but non-inflectional morphemes in HU tend to consist minimally of a unit with prosodic weight. As such, I restrict the proposal distribution for segmentations of HU numbers to exclude morpheme boundaries following the first and penultimate segments; this additionally speeds up inference and ensures that short forms like /bis/ ' 20 ' will be treated as monomorphemic.

## Phonological features

The methodology developed in this paper must capture allomorphy in the HU numeral system, inferring that different surface strings such as /-jalis/ and /calis/ correspond to the same underlying morpheme, TENS $\{4\}$. I cluster allomorphs together on the basis of phonological features using essentially the same likelihood formula used in unsuper-
vised Naïve Bayes/Dirichlet-Multinomial classifiers, popular in bag-of-words models of document classification. This is a model of convenience which I find fairly effective, though it is admittedly crude; it is insensitive to positional information and alternations, and does not strongly penalize the absence of potentially crucial morpheme-level information.

The model described in this paper lends itself to the use of different types of phonological features, and provides opportunities to investigate model performance under features with differing degrees of abstraction. In this paper, I limit myself to domain-general string-based features, namely $n$ grams. In many contexts, I expect segmental bigrams to fare well in assigning cohort membership. However, there are some cases where I fear that bigrams may fail to capture alternations caused by (among other things) deletion or insertion, as between the base /nəve/ ' 90 ' and /-nve/, found in complex forms. A unigram model will be sensitive to the co-occurrence of $/ \mathrm{n} /$ and $/ \mathrm{v} /$, whereas a representation consisting solely of bigrams will not (on the use of separate autosegmental tiers for consonants and vowels, see Goldsmith \& Riggle, 2012). I attempt to circumvent this problem with a phonological representation that uses both unigrams and bigrams (though this technically violates the independence assumption of Naïve Bayes). This allows the model to capture some similarities between paradigmatically related forms that would otherwise be lost in a strict bigram model.

## Model

Here, I introduce the core model employed in this paper, designed to approximate a HU speaker's recognition of numbers 10-99 (I assume that 1-9 are primitives). When encountering a numerical form, the listener must determine whether it is simplex or complex. If simplex, the value of the TENS place must be inferred; if complex, the DIGITS place must be as well. The model assumes that a complex form is generated by independent draws from two mixtures, a DIGITS mixture (the labels of which correspond to the values $\{-1,1, \ldots, 9\}$ ) and a TENS mixture (the labels of which correspond to the values $\{1, \ldots, 9\}$ ). Because HU morphology is generally concatenative, I make the simplifying assumption that phonological elements generated by a given mixture are adjacent to one another - i.e., that a morpheme boundary can be located somewhere in a complex form, however approximately. I make the assumption that the lefthand morpheme is generated by the digits mixture and the righthand morpheme is generated by the tens mixture; this convention essentially incorporates Hurford's insight that higher numerical elements occur closer to the root, which in turn can be interpreted as prior knowledge of a morphosyntactic headedness parameter. This system of numerical classification is schematized in Figure 1 .

## Inference

This paper's basic model of numeral classification assigns each form to one or two mixtures, given a $10 \times F$ matrix $\Omega^{D}$ and a $9 \times F$ (where $F$ is the number of feature types in


Figure 1: Schema of a proposed morphological segmentation, tens classification, and digits classification for form /bajalis/
the input) matrix $\Omega^{T}$ (specifying a prior over feature distributions associated with each label of the digits and TENS place, respectively), as well as a word-level vector $\mu$ representing a prior over morpheme boundary locations. I initialize these matrices with symmetric concentration parameters $\alpha^{T} \alpha^{D}, \alpha^{\mu}$, set to . 1 in order to encourage sparseness, such that unshared features from unrelated labels are not clumped together. The generative model draws probability simplices $\phi_{j}^{D} \sim \operatorname{Dirichlet}\left(\omega_{j}^{D}\right), \phi_{i}^{T} \sim \operatorname{Dirichlet}\left(\omega_{i}^{T}\right)$ representing the feature distributions associated with levels $j$ and $i$ of the DIGITS and TENS place, and assumes that for every word $w$,
$\varsigma \sim \operatorname{Dirichlet}(\mu)$ (a simplex of morpheme boundary probabilities is drawn, including the probability $p(m=\emptyset)$, i.e., the probability that there is no morpheme boundary)
$m \sim$ Categorical $(\varsigma)$ (a morpheme boundary is drawn from $\varsigma$ ) If $m=\emptyset$,
$z_{j}^{D}=0$
for each feature $f \in w$

$$
f \sim \operatorname{Categorical}\left(\phi_{i}^{T}\right), i \in\{1, \ldots, 9\}
$$

If $m \neq \emptyset$,
For each feature $f \in w_{1, \ldots, m}$ (through index $m$ )

$$
f \sim \operatorname{Categorical}\left(\phi_{j}^{D}\right), j \in\{-1,1, \ldots, 9\}
$$

For each feature $f \in w_{m+1, \ldots,|w|}$ (from index $m+1$ through the end of the word)

$$
f \sim \text { Categorical }\left(\phi_{i}^{T}\right), i \in\{1, \ldots, 9\}
$$

I marginalize out the parameters $\varsigma, \phi_{i}^{T}, \phi_{j}^{D}$ to obtain collapsed Dirichlet-Categorical updates for $p(m \mid \mu), p\left(z^{T} \mid \Omega^{T}\right), p\left(z^{D} \mid \Omega^{D}\right)$. For a given word, this yields the following conditional probability if $m=\emptyset$ (adopted from Yin \& Wang, 2014):

$$
\begin{align*}
P\left(m=\emptyset, z_{i}^{T}, z^{D}=\right. & \left.0 \mid z_{-i}^{T}, z_{-0}^{D}, \Omega^{T}, \Omega^{D}, \mu\right) \propto \\
& \frac{\prod_{f \in w} \prod_{n=1}^{c(f) w} c(f)_{z_{j}^{-w}}^{-w}+\alpha^{T}+n-1}{\prod_{k=1}^{w \mid c} c(\cdot)_{z_{j}^{T}}^{-w}+F \alpha^{T}+k-1} \tag{1}
\end{align*}
$$

If $m \neq \emptyset$ :

$$
\begin{array}{r}
P\left(m, z_{i}^{T}, z_{j}^{D} \mid z_{-i}^{T}, z_{-j}^{D}, \Omega^{T}, \Omega^{D}, \mu\right) \propto \\
\frac{\prod_{f \in \lambda_{m}^{\prime}}^{m} \prod_{n=1}^{c(f)} \lambda_{\lambda_{m}^{\prime}} c(f)_{z_{j}^{D}}^{-w}+\alpha^{D}+n-1}{\prod_{k=1}^{m} c(\cdot)_{z_{j}^{D}}^{-w}+F \alpha^{D}+k-1} \cdot \frac{\prod_{f \in \lambda_{m}^{r}} \prod_{n=1}^{c(f) \lambda_{m}^{r}} c(f)_{z_{i}^{T}}^{-w}+\alpha^{T}+n-1}{\prod_{k=1}^{|w|-m} c(\cdot)_{z_{i}^{T}}^{-w}+F \alpha^{T}+k-1} \tag{2}
\end{array}
$$

Above, $c(f)_{z_{i}^{T}}^{-w}$ denotes the number of instances of $f$ currently associated with label $z_{i}^{T}$, and $c(\cdot)_{z_{i}^{T}}^{-w}$ the number of instances of any item currently associated with label $z_{i}^{T}$ (both terms exclude any instances contributed by $w$ ); $c(f)_{\sigma}$ signifies the number of instances of $f$ in element $\sigma$. For simplicity, I write $\lambda_{m}^{l}$ for $w_{1, \ldots, m}$ and $\lambda_{m}^{r}$ for $w_{m+1, \ldots,|w|}$. Once new values of $m, z_{i}^{T}, z_{j}^{D}$ are chosen for word $w$, counts for the features in $\lambda_{m}^{l}$ (if $m \neq \emptyset$ ) and $\lambda_{m}^{r}$ can be allocated to $z_{j}^{D}$ and $z_{i}^{T}$, respectively.

## Priors on morphological segmentation

For this paper's most basic inference procedures, the prior over morpheme boundaries is symmetric, with equal probability allocated to all possible segmentations of word $w$. In certain inference regimes, I employ one of two priors on segmentation incorporating the principle of Minimum Description Length, popular in unsupervised morphological segmentation (Goldsmith, 2001; Creutz \& Lagus, 2007); these priors favor the insertion of morpheme boundaries which minimize the length of the code that generates the data. There are a number of ways to interpret this principle. Most intuitively, the "code" can be construed as the list of morph types, or alternatively, the sum of the lengths of morph types. Hence, an MDL or exponential prior on morphological segmentations disfavors analyses that add to the list, or the sum of (string) lengths of types in the list.

The first prior (MDL1), designed to keep the list of analyzed morphs short, assigns probability to a morphological segmentation for word $w$ proportional to the inverse of the number of morph types as currently analyzed, including the proposed segmentation for $w$; under the second approach (MDL2), the prior probability is inversely proportional to the sum of lengths of current morph types. I have employed these priors due to the importance of MDL in the literature on unsupervised segmentation, but remain somewhat skeptical as to whether HU numeral morphology can be rendered compact in the same manner as the morphology traditionally analyzed with MDL priors (e.g., of English, Finnish, Turkish, etc.), given the noisy allomorphy seen.

## Priors on cluster membership

Readers may note that the above formulae depart from traditional Dirichlet-Multinomial mixture models in that the Chinese Restaurant Process prior (a rich-get-richer scheme) over cluster membership is excluded. This prior, which makes it more likely for an item to be assigned to a cluster that already has many data points, seems inappropriate for this paper's model, which iterates over one token of each number, and should learn classes of roughly equal size. In one sampling regime, I place an exponential prior on TENS and DIGITS label membership, inversely proportional to the number of items currently assigned to the label in question (plus a concentration parameter). The intention here is to introduce a pressure toward clusters of uniform size.

## Inference procedure

Inference is carried out via Markov chain Monte Carlo. I run different versions of the model on three chains for 10000 iterations, discarding the first half of samples as burn-in. Each chain is initialized by randomly segmenting and assigning each item to a TENS and DIGITS label. Parameters are updated via Gibbs Sampling; for each number in 10-99, a morphological segmentation $m$, a TENS label $z^{T}$ and if relevant, a DIGITs label $z^{D}$ are drawn conditional on the labels currently assigned to all other data points (see eqq. 1-2). I use a simulated annealing procedure, raising each vector of update probabilities to the power of a constant $\frac{1}{\gamma}$, with $\gamma$ decreasing from 10 to 1 over the course of the burn-in. Code can be found at github.com/chundrac/HUnumerals.

I carry out an inference procedure using only bigrams as a phonological feature representation $(2 \mathrm{~g})$; this is followed by a regime using unigrams and bigrams $(1+2 \mathrm{~g})$. I modify the $1+2 \mathrm{~g}$ procedure to incorporate an MDL prior sensitive to the length of the current list of morph types (MDL1), followed by an MDL prior sensitive to the sum of their lengths (MDL2). I attempted to see how the MDL1 prior (which showed better performance) affected the bigram model. Additionally, I ran a simulation which augmented the $1+2 \mathrm{~g} /$ MDL1 model with a prior over component membership designed to keep clusters uniform (denoted by U). ${ }^{1}$

## Results

I use the overall F-measure (Fung et al., 2003) and the Vmeasure (Rosenberg \& Hirschberg, 2007), two evaluation metrics designed to quantify the similarity between two classifications, in order to monitor convergence and measure overall accuracy (convergence was also assessed via chain log-likelihoods). I compute pairwise F- and V-measures between the maximum a posteriori (MAP) configuration of each chain to assess the degree to which chains return the same classification, interpreting values greater than .9 as a token of convergence between two chains. I evaluate each chain's accuracy by computing the F- and V-measures between the chain's MAP configuration and the true classification of the numbers. These values are found in Table 2. In general, MDL priors do not appear to improve inference for bigrams, and do not significantly improve inference for $1+2$ grams.

Table 3 displays the MAP configuration for the top chain (2) in the regime with highest overall accuracy $(1+2 \mathrm{~g} / \mathrm{MDL} 1 / \mathrm{U})$. To measure the ACCURACY with which this regime decomposes individual numbers, I calculate the F-scores for each number's MAP TENS and DIGITS classifications with respect to its true TENS and DIGITS classifications, averaging these values. The resulting values are then averaged across chains. I calculate POSTERIOR UNCERTAINTY

[^300]by averaging the entropy of the posterior sample (comprising blocked draws of $m, z^{T}, z^{D}$ ) for each chain.

I extract numeral frequencies from the EMILLE Hindi Webnews corpus (Baker et al., 2002). For each number, accuracy and posterior uncertainty are plotted according to frequency in Figure 2, along with correlation coefficients and $p$-values. Both correlations are significant (albeit noisy), providing support for the idea that the HU numbers can be processed via a dual-route model. As seen in the lefthand plot, the majority of HU numbers occupy a quasi-Pareto frontier, indicating an efficient trade-off between decomposability and frequency. Several numbers in the teens (seen in the upper righthand corner of the plot) are both highly frequent and decomposable. These outliers in no way contradict the dualroute model, since a form's decomposability does not preclude the possibility that it is stored whole. However, a handful of numbers are found beneath the frontier (near the lower lefthand corner), meaning that they are both relatively infrequent and difficult to parse. These items can be viewed as vulnerable points in the grammar of HU numbers, and may be prone to "leakage" or analogical restructuring.

Table 2: F-/V-measures for different inference regimes

|  | TEN | DIG | TEN | DIG | TEN | DIG |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| convergence | chain 1-2 |  | chain 1-3 |  | chain 2-3 |  |
| 2 g | . $88 / .87$ | . $77 / .74$ | . $86 / .86$ | . $81 / .78$ | . $92 / .91$ | . 951.93 |
| 2g,/MDL1 | .77/.80 | .86/.82 | . $78 / .81$ | .85/.84 | . $87 / .84$ | . $90 / .88$ |
| $1+2 \mathrm{~g}$ | .88/.86 | .89/.88 | . $90 / .88$ | . $90 / .89$ | .99/.98 | .99/. 99 |
| $1+2 \mathrm{~g} / \mathrm{MDL} 1$ | .93/.89 | .95/.95 | .94/.91 | . $95 / .95$ | .99/.98 | 1/1 |
| $1+2 \mathrm{~g} / \mathrm{MDL} 2$ | .94/.92 | .95/.94 | .94/.92 | .95/.94 | 1/1 | 1/1 |
| $1+2 \mathrm{~g} / \mathrm{MDL} 1 / \mathrm{U}$ | .88/.87 | . $92 / .92$ | . $89 / .89$ | . $92 / .92$ | .97/.96 | 1/1 |
| over. accuracy | chain 1 |  | chain 2 |  | chain 3 |  |
| 2 g | . $81 / .81$ | . $76 / .74$ | . $82 / .84$ | .86/.84 | . $87 / .88$ | .92/.89 |
| 2g/MDL1 | . $78 / .79$ | . $82 / .80$ | . $80 / .82$ | .89/.88 | .80/.81 | .85/.83 |
| $1+2 \mathrm{~g}$ | .87/.86 | .89/.87 | . $91 / .91$ | . $90 / .88$ | . $91 / .91$ | .92/.89 |
| $1+2 \mathrm{~g} / \mathrm{MDL} 1$ | . $90 / .89$ | .88/.86 | . $91 / .91$ | .9/.88 | .91/.91 | .91.88 |
| $1+2 \mathrm{~g} / \mathrm{MDL} 2$ | . $88 / .87$ | . $90 / .88$ | . $91 / .91$ | .9/.87 | . $91 / .91$ | .9/.87 |
| $1+2 \mathrm{~g} / \mathrm{MDL} 1 / \mathrm{U}$ | .91/.89 | .9/. 9 | .93/.91 | .92/.89 | .91/.9 | .92/.89 |

Table 3: MAP configuration for $1+2 \mathrm{~g} / \mathrm{MDL} 1 / \mathrm{U}$, chain 2. Rows represent tens classification; columns represent digits classification. Numbers are represented by cardinality for readability. Asterisks $(*)$ mark numbers where the numerical representation TENS $\{i\}$, $\operatorname{DIGITS}\{9\}$ maps to the representation TENS $\{i+1\}$, DIGITS $\{-1\}$

| 35 |  | 34 | 31 |  | $29 *$ | 32 | 36 | 38 |  | 30 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :---: |
| 75 | 77, <br> 70 | 74 | 71 | 73 | $69 *$ | 72 | 76 | 78 |  |  |
| 15 | 17, <br> 16 | 14 |  | 13 |  | 12 |  | 18 | 11 | 10 |
| 65 | 67 | 64 | 61 | 63 | $59 *$ | 62 | 66 | 68 |  | 60 |
| 44, | 47, | 40 | 41 | 43 | $39 *$, <br> $49 *$ | 42 | 46 | 28, |  |  |
| 45 | 27 |  |  |  |  |  |  | 48 |  |  |
| 25 | 37 | 24 | 21 | 33, | $19 *$ | 22 | 26 |  |  | 20 |
| 95 | 97 | 94 | 91 | 93 |  | 92 | 96 | 98 | 90, |  |
| 55 | 57 | 54 | 51 | 53 |  | 52 | 56 | 58 |  |  |
| 50, | 87 | 84 | 81 | 83 | $79 *$ | 82 | 86 | 88, | 89 |  |
| 85 |  |  |  |  |  |  |  | 80 |  |  |

## Discussion

The models presented in this paper show that although HU numerals 10-99 are morphologically irregular, a large number can be classified according to their component parts. However, quite a few forms are difficult to decompose, most of them of low magnitude and high frequency. In general, the models handled some types of allomorphy well, and others


Figure 2: Log frequency as a predictor of model accuracy ( $\rho=-.55, p<.001$ ) and post. uncertainty $(\rho=.38, p<.001)$
poorly. Forms containing the TENS $\{5\}$ allomorphs /-pən/ $\sim$ /-vən/ are grouped together, due to their agreement in two out of three segments. Surprisingly, /solə/ '16' was recognized as a member of the teens, despite the unique allomorph /-la/; however, the model failed to properly classify it according to its digits place. Other forms with highly suppletive allomorphy (e.g., /vncas/ '49') were misclassified. Additionally, many simplex forms were not analyzed as monomorphemic, unless only a monomorphemic analysis was permitted under the proposal distribution.

As stated above, my results show that the HU numeral system's design is largely compatible with a dual-route model of access. In general, high-frequency items were more difficult for a computational model to decompose, indicating greater opacity. (Berger shows that many of these numbers were historically subject to erosion and evidently resistant to analogical changes that would otherwise make them more transparent and perceptually distinct.) At the same time, there are exceptions to this generalization: certain high-frequency items in the teens showed high accuracy, though this does not rule out the possibility that they are stored whole. Additionally, some problematic items are more opaque than would be expected, given their low frequency. It is likely that such vulnerable forms cause problems in planning and production.

The EMILLE Spoken Hindi corpus contains intriguing numeral variants (e.g., /it ${ }^{\mathrm{h}}$ janve/ ' 91 ' by speaker ehinsp041, /sintrjanve/ '97' by ehinsp035, /vnanve/ ' 89 ' by ehinsp044), though the data are too sparse to serve as the basis of a rigorous quantitative study. Many numbers are missing in the corpus; furthermore, the variation observed may stem from sources other than production difficulty, including transcriber error, multilingualism (with another Indic language; for example, speaker ehinsp017 utters the form /bavis/ '22', standard in Marathi but not HU), and stylistic factors. Studies of variability in the production of HU numerals - either in experimental contexts or naturalistic speech - will serve as a valuable research direction, particularly with an eye to whether vulnerable forms (i.e., sub-
optimal forms with higher opacity than expected relative to frequency) are subject to greater instability.

## Conclusion

In this paper, I have employed a simple and somewhat crude model of allomorphy, inspired in part by bag-of-words models used in document classification and intended to serve as a baseline for future work. A goal of this study was to test the limits of a simple mixture model in a HU numerical recognition task. A more sophisticated model of phonological processes may relate potential allomorphs to each other in terms of edits, as has been done in some MDL approaches (Virpioja et al., 2010). However, while such models can contend with or recover relatively regular allomorphy, no model has been designed, to my knowledge, to capture the highly noisy allomorphy found in the HU numeral system.

A true test of any computational model's value is in how well it agrees with human performance. A future direction for this work will involve carrying out experimental research to see how HU speakers process and produce numerical forms. It will serve us well to see how model inaccuracy fares as a predictor of greater response latency in psycholinguistic tasks. A joint approach which considers limitations in both experimental performance and computational simulation will help us identify weak points in this and other complex morphological systems that can potentially (though not obligatorily) undergo analogical change.

I have shown that frequency may facilitate the processing of more opaque HU numbers, but the question remains as to why most Indic number systems are on average more irregular than exact number systems found in other languages. Sociocultural factors may be partially responsible, ${ }^{2}$ and their role in shaping cross-lingustic number systems should be taken into account along with that of functional need (cf. Xu \& Regier, 2014).

## Acknowledgements

I am grateful to Stephan Meylan, Terry Regier, and three anonymous reviewers for helpful comments. All errors and infelicities are my own responsibility.

## References

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij \& J. van Marle (Eds.), Yearbook of morphology 1992 (p. 181-208). Dordrecht: Kluwer.
Baker, P., Hardie, A., McEnery, T., Cunningham, H., \& Gaizauskas, R. (2002). EMILLE, a 67-million word corpus of Indic languages: Data collection, mark-up and harmonisation. In Proc. LREC 2002 (p. 819-825).

[^301]Berger, H. (1992). Modern Indo-Aryan. In J. Gvozdanović (Ed.), Indo-European numerals (Vol. 57, p. 243287). Berlin: Walter de Gruyter.

Bright, W. (1969). Hindi numerals. Working Papers in Linguistics (University of Hawaii), 9, 29-47.
Brysbaert, M. (2005). Number recognition in different formats. In J. I. Campbell (Ed.), Handbook of mathematical cognition (p. 23-42). New York, Hove: Psychology Press.
Comrie, B. (n.d.). Typology of numeral systems. Available at https://mpi-lingweb.shh.mpg.de/numeral/ TypNumCuhk_11ho.pdf. Accessed 1 October 2016.
Creutz, M., \& Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing, 4, 134.

Fung, B. C. M., Wang, K., \& Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In Proceedings of the SIAM International Conference on Data Mining.
Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27(2), 153-198.
Goldsmith, J., \& Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. Natural Language and Linguistic Theory, 30, 859-896.
Harris, Z. S. (1955). From phoneme to morpheme. Language, 31(2), 190-222.
Hurford, J. R. (1987). Language and number: The emergence of a cognitive system. Oxford: Blackwell.
Rosenberg, A., \& Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (p. 410-20). Prague: Association for Computational Linguistics.
Saffran, J. R., Newport, E. L., \& Aslin, R. N. (1996). Word segmentation: The role of distributional cues. Journal of Memory and Language, 35, 606-21.
Turner, R. L. (1962-1966). A comparative dictionary of the Indo-Aryan languages. London: Oxford University Press.
Virpioja, S., Kohonen, O., \& Lagus, K. (2010). Unsupervised morpheme analysis with Allomorfessor. In C. Peters et al. (Ed.), CLEF 2009 Workshop, Part I (Vol. 6241, p. 609616). Heidelberg: Springer.

Xu, Y., \& Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, \& B. Scassellati (Eds.), Proceedings of the 36th annual meeting of the Cognitive Science Society.
Yin, J., \& Wang, J. (2014). A Dirichlet Multinomial Mixture Model-based approach for short text clustering. In KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (p. 233-242). New York: ACM.

# Is the relative distribution of verbs and nouns modulated by socio-cultural influences? Evidence from bilingual infants and toddlers in Malaysia. 

Jun Ho Chai<br>The University of Nottingham Malaysia Campus<br>Hui Min Low<br>University Sains Malaysia<br>Tze Peng Wong<br>The University of Nottingham Malaysia Campus<br>Julien Mayor<br>University of Oslo, Oslo, Norway


#### Abstract

Early vocabularies in most languages tend to contain more nouns than verbs. Yet, the strength of this noun bias has been observed to vary across languages and cultures. Two main hypotheses have aimed at explaining such variations; either that the relative importance of nouns vs. verbs is language- specific, or that socio-cultural influences shape early vocabulary structures. The present study compares the relative distribution of verbs and nouns, in English, between two groups of bilingual infants and toddlers; Malay-English and Mandarin- English. We found that early English lexicons of Mandarin- English bilinguals contained more verbs than in the English lexicon of Malay-English bilinguals, in both comprehension and production. We discuss the potential role of socio-cultural influences on the vocabulary structure in young users of a language.


# Exploring the relations between oral language and reading instruction in a computational model of reading 

Ya-Ning Chang (y.chang5@lancaster.ac.uk)<br>Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK<br>J. S. H. Taylor (joanne.taylor@rhul.ac.uk)<br>Kathleen Rastle (kathy.rastle@rhul.ac.uk)<br>Department of Psychology, Royal Holloway, University of London, TW20 0EX, UK<br>Padraic Monaghan (p.monaghan@lancaster.ac.uk)<br>Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK


#### Abstract

To become a proficient reader, children have to learn mappings between print, sound and meaning. There is debate over whether reading instruction should focus on the relations between print and sound, as in phonics, or on the relationship between print and meaning, as in sight word reading. In a study where participants learned a novel artificial orthography, Taylor, Davis and Rastle (2017) compared print to sound focused or print to meaning focused reading training, demonstrating that sound training was superior for learning to read. However, a benefit from sound focused training is likely dependent on prior acquisition of effective sound to meaning relations of words. To explore this issue, we developed a connectionist model of reading. We exposed the model to a sound or a meaning focused training, but varied the model's pre-acquired oral language skills. The simulation results showed that proficiency in oral language is a determinant of the advantage of print to sound focused reading training, suggesting that reading training should address both oral language skills and print to sound mappings.


Keywords: reading instruction; oral language; reading development; computational modelling; word learning.

## Introduction

Learning to read requires mastery of a set of complex skills involving encoding phonology ( P ), semantics ( S ), and learning to map orthographic (O) forms onto those representations of sound and meaning. Even for alphabetic orthographies, where a letter, or set of letters, corresponds approximately regularly to a phoneme in the word, learning to read is effortful and frequently fraught with difficulties (Seidenberg, 2017). Effective early reading instruction is therefore critical to help children become proficient readers. There has been a vigorous debate over whether reading instruction should focus on the relations between print and sound or on the relationship between print and meaning. The former is typically characterized by phonics-style training, where the phonemes associated with particular letters or letter clusters are trained intensively, enabling children to decode letter-by-letter. The latter is often referred to as meaning-focused or whole-word language instruction, where the meaning and pronunciation of the whole word is provided to the child during training.

Proponents of the phonics method argue that reading instruction should focus on learning spelling-to-sound mappings because exploiting the systematicity of alphabetic writing systems ought to be substantially easier than acquiring more arbitrary spelling-to-meaning mappings, where the arbitrariness of the sign is dominant and learning can only be accomplished word by word, without the benefit of generalising from one learned word to the next. Evidence for the strong predictive relation between phonological decoding skills and reading acquisition (see, e.g., Rayner et al., 2001, for a review) demonstrates that phonological skills are key to reading success.

Alternatively, researchers who advocate the meaningfocused method (see, e.g., Davis, 2013, for a review) argue that the primary goal of reading is to access the meanings of words and so this ought to be the priority of reading training approaches. Although spelling-to-meaning mappings are hard to learn, they may still be acquired early in reading development (Nation, 2009; Taylor et al., 2015). For example, Nation and Cocksey (2009) demonstrated that 7-year-old children could access semantic categories of words from orthography very quickly without evidence that the phonological form of the words mediated responses.

## Effectiveness of sound-focused and meaningfocused reading instruction

According to the Simple View of reading (Gough \& Tunmer, 1986), reading comprehension is the product of phonological decoding and oral vocabulary. During reading training, learners acquire mappings from print to sound, and access meaning based on their knowledge of sound-tomeaning mappings acquired pre-literacy. There is some evidence that both print to sound mapping skills (as indexed by pseudoword reading tasks) as well as sound to meaning mapping skills (as reflected in oral vocabulary tasks) are predictors of silent reading comprehension performance (e.g., Curtis, 1980; Nation \& Snowling, 2004; Ouellette \& Beers, 2010; Ricketts, Nation, \& Bishop, 2007). However, the Simple View of reading does not consider an alternative, which involves the role of accessing meaning directly from print (Taylor et al., 2015).

Within the connectionist view of reading (Seidenberg \& McClelland, 1989; Harm \& Seidenberg, 2004; Plaut et al.
1996), learning to acquire the meaning of written forms of words could be via developing direct orthographic to semantic mappings. Alternatively, acquisition could be indirect, through the learner developing orthographic to phonological mappings, which then map, via oral language knowledge, onto semantic representations. Computational modelling investigations have established that there is division of labour along these direct and indirect pathways from orthography to semantics over development (Harm \& Seidenberg, 2004; Plaut et al. 1996). However, comparisons between reading training that focuses on developing the direct orthographic-to-semantic, versus the indirect orthographic-to-phonological, pathways have not as yet been undertaken.

One exception to this is a recent study by Taylor, Davis and Rastle (2017). In a laboratory study using adults, Taylor et al. compared reading acquisition when training was biased toward orthography-to-semantics (OS) mappings versus orthography-to-phonology (OP) mappings. They trained literate adult participants to read two sets of 24 novel words which were written in two different unfamiliar alphabetic orthographies (in each orthography, one character related to one phoneme) - see Figure 1. Each novel word was assigned a familiar concrete noun meaning (e.g., /ged/ referred to camel, and $/ \mathrm{k} \varepsilon \mathrm{s} /$ referred to parsnip), and the mappings between novel words and their referents were counterbalanced across participants).

## NSM X XN

Figure 1. $/ \mathrm{g} \varepsilon \mathrm{d} /$ and $/ \mathrm{k} \varepsilon \mathrm{s} /$ in the artificial orthography from Taylor et al. (2017).

Prior to reading training, participants were exposed to the mappings between phonology and semantics for the novel words. Then, participants learned orthographic-tophonological and orthographic-to-semantic mappings for both orthographies. For one orthography, participants received OP focused training, which involved three times as many orthographic-to-phonological training trials as orthographic-to-semantic training trials, whereas for the other orthography they received OS focused training, which involved three times as many orthographic-to-semantic as orthographic-to-phonological training trials. The results demonstrated that OP focused training led to better accuracy and speed in reading aloud, and it also had a transferable benefit to reading comprehension. By contrast, OS focused training resulted in faster but not more accurate reading comprehension, and showed no transferable benefit for the reading aloud task.

Taylor et al. (2017) demonstrated that both reading aloud and reading comprehension accuracy could be promoted by focusing on OP mappings during reading training. However, unlike children learning to read for the first time, participants were acquiring an orthography which very likely piggy-backs on the reading system that the participants already have. Thus, an outstanding question is the extent to which prior language skills, particularly
between phonology and semantics, are critical to the OP versus OS focused reading training differences.
Furthermore, a key aspect of Taylor et al.'s (2017) study design was that participants were pre-trained on mappings between phonological and semantic forms for the novel words. This previously tuned phonology-semantics system is crucial to allow the transference of knowledge from training on OP mappings to access meaning from print, since this requires using not only the OP but also the PS routes within the reading system.

According to the connectionist view of reading, then, phonics instruction will be most successful if the participant has acquired an effective level of oral language knowledge. Thus, in relating the laboratory-based studies of reading acquisition to the child's task of learning to read, the relative contribution of training from OP and OS on reading acquisition needs to be considered alongside the contribution of pre-literate oral language skills.

## Computational models of reading

Computational models of reading have converged on an architecture involving two different pathways that are active during reading - a subword orthographic to phonological pathway and an orthographic whole word pathway, which may map onto a whole-word phonological representation and/or a semantic representation of the word (Coltheart et al. 2001; Plaut et al. 1996). There are also mappings between phonological and semantic representations, meaning that words can be comprehended both by direct OS mappings, and also indirectly via OP then PS mappings. In the connectionist tradition, the relative contribution for generating phonology or semantics via different reading pathways is flexible, and can be determined by properties of individual words, such as high-frequency words more likely to be read via direct OS mappings, or due to properties of the orthographic system itself, such as ideographic writing systems more likely to utilize the direct OS mappings than alphabetic writing systems (Chang, Welbourne \& Lee, 2016; Harm \& Seidenberg, 2004; Plaut et al. 1996).

In this study, we implemented the two reading schemes tested in Taylor et al.'s (2017) study, in order to determine whether the connectionist triangle model of reading is able to replicate the behavioural effects of an OP focused versus an OS focused training regime. Furthermore, we examined whether the advantage for the OP focused training demonstrated in Taylor et al.'s (2017) study was present even for the model with poor oral language skills, or only when well-established mappings between phonological and semantic representations were in place. Tracking the relative benefit of OP and OS focused training according to preliterate oral language skills enables greater clarity on how different reading training schemes may benefit readers with varying language abilities.

Following Harm and Seidenberg (2004), we developed a fully implemented connectionist model of learning to read, that mapped between representations of orthography, phonology, and semantics of words. The model was
pretrained to different degrees of proficiency in mapping between phonological and semantic representations of words, to simulate pre-literate oral language skills. We tested three different quantities of pre-training to reflect a model with moderate, medium, and high levels of oral language skills, in terms of the overall fidelity of phonological and semantic representations within the model, and the proportion of words in the language for which the model was able to generate the correct semantic and phonological representations. We then compared the effects of two reading training regimes with different focuses of reading instruction - orthography to phonology (OP) focused model or orthography to semantics (OS) focused model. Prior to learning to read, both models received three different amounts of pretraining (i.e. 500, 1000, or 2000 epochs) on mappings between semantics and phonology. The OP focused model then received three times as much training on the OP mappings, while the OS focused model received three times as much training on the OS mappings. We evaluated the model's performance under these different training regimes using tasks of reading aloud and reading comprehension.


Figure 2. The architecture of the developmental model of reading.

## Method

## Network Architecture

The architecture of the model is shown in Figure 2, which was the same as the developmental model of reading implemented in Monaghan, Chang, Welbourne, and Brysbaert et al. (2017) and Chang, Monaghan, and Welbourne (2016). The model consisted of three key processing layers representing orthographic, phonological and semantic representations respectively, and four hidden layers that learned to map between the processing layers. An attractor layer, which contained 50 hidden units, was connected to and from the phonological layers. Similarly, there was a set of 50 hidden units for the semantic layer. The use of attractors was to help the model to develop stable phonological and semantic representations of words. The semantic layer was connected to the phonological layer through a set of 300 hidden units, and the phonological layer was connected back to the semantic layer through another
set of 300 hidden units. The orthographic layer was connected to both the phonological and semantic layers through different sets of 500 hidden units.

## Training Corpus: Artificial Words

The training corpus comprised 24 artificial words, taken from the materials in Taylor et al. (2017). For the phonological forms, all items were monosyllabic consonant-vowel-consonant pseudowords. All items were constructed from 12 consonants (/m/, /t/, /g/, /b/, /k/, /d/, /n/, /s/, /z/, /v/, $/ \mathrm{p} /$, and, $/ \mathrm{f} /$ ) and four vowel phonemes ( $/ \varepsilon /, / \mathrm{I} /, / \rho /$, and, $/ \Lambda /$ ). For phonology, each word was represented in the $3^{\text {rd }}, 4^{\text {th }}$ and $5^{\text {th }}$ slots of a set of eight phoneme slots, with each slot consisting of 25 phonological features. Each word was thus positioned with its vowel at the fourth phoneme slot. The first three slots were for onset consonants, and the last four slots were for coda consonants, but because all words in the set had one onset and one coda consonant, only one of these slots was used during training (so for the word "tep" its phonology was represented as ${ }_{-} \mathrm{t} \varepsilon \mathrm{p} \ldots_{-}$, where indicates an empty slot). For orthographic forms, the correspondence between letters and phonemes was transparent (i.e., there was a one-to-one correspondence). For orthography, each word was represented across a layer containing 14 letter slots with each slot comprising 26 units, each of which could represent a distinct letter, so an alphabet up to 26 letters could be represented. Words were positioned with their vowel aligned on the fifth slot. Consonants preceding the vowel were positioned in slots right before the vowel and consonants following the vowel were positioned starting from the seventh slot. This representation was in alignment with Chang et al. (2016), which enabled words up to 14 letters to be represented. However, because all words were three letters in length, with one onset and one coda consonant, words occupied only the $4^{\text {th }}, 5^{\text {th }}$, and $7^{\text {th }}$ slots (so for the word "tep" its
 Note that we use here Roman alphabet as a short hand to reflect the alphabet used in the laboratory-based study. There is nothing particular in the representations used in the model regarding the particular alphabet used, only that the model is able to distinguish the letters from one another from the outset, but does not know the properties of the letters in other respects in advance of commencing training.
For semantics, a set of familiar objects consisting of six fruits and vegetables, six vehicles, six animals, and six tools were randomly assigned to the 24 artificial words. The semantic representation for each word was derived from Wordnet (Miller, 1990), following Harm and Seidenberg (2004). Each semantic representation was composed of 2446 semantic features. The presence of semantic features was encoded as 1 and the absence of semantic features was encoded as 0 in the respective slot.

## Training Procedure

The model was trained on the 24 artificial words. All the training parameters were exactly the same as those used in
our previous modelling work (Chang et al., 2016). The training process had two phases: pretraining and reading training. For the pretraining, the model learned to map from phonological to semantic (PS) representations in an oral vocabulary task and from semantic to phonological (SP) representations in a meaning naming task (e.g. picture naming). To investigate how oral language skills affected literacy development, three different amounts of pretraining were used $-500,1000$, or 2000 learning trials. For the oral vocabulary (PS) task, the phonological representation of the word was clamped at the phonological layer for eight time steps, and the model generated a semantic representation at the semantic layer. The difference between the actual and the target semantic representation was then calculated, and the weights on connections between all the layers were adjusted according to gradient descent backpropagation through time in order to reduce the error. The training rate was 0.1 . Similarly, for the meaning naming task (SP), the semantic representation was clamped at the semantic layer for eight time steps, and the model was required to produce a phonological representation. In pretraining, the model additionally learned to develop a stable phonological attractor (PP), and a stable semantic attractor (SS), by presenting the phonological or the semantic representation for two time steps, then allowing the model to cycle activation for a further six time steps to reproduce the initial representation. During pretraining, these four tasks (PS, SP, PP, and SS) were interleaved, with $40 \%$ of trials for the oral vocabulary task, $40 \%$ of trials for the meaning naming task, $10 \%$ of trials for the phonological attractor and $10 \%$ for the semantic attractor. For each trial, a word was randomly selected.

After pretraining, the weights on connections between the semantics and the phonology layers were frozen. The model was then trained to learn to read with different focuses of reading instruction, in two separate simulations as either the OP focused or OS focused model. For the OP focused model, there were three OP trials for every OS trial, and for the OS focused model the reverse was true. For an OP trial, the model's error at the phonological layer at the final time step was computed and then backpropagation with gradient descent adjusted the weights to reduce this error. For an OS trial, error was propagated from the semantic representation. Each model was trained for 1000 reading trials. For each reading learning trial, a word was randomly selected and presented at the orthographic layer for 12 time steps. Five versions of each model were trained with different random initial weights and different random samplings from the words.

## Testing Procedure

For testing the model's phonological output, we determined the number of words for which all phonemes were correctly produced. The closest phoneme representation from the set of all phonemes in the language was derived from the model's actual production and this was then compared against the target phoneme. If the actual and
target phonemes were the same, then the model was judged to have spoken the word correctly. For testing the model's semantic output, the activation of units at the semantic layer was recorded. Accuracy was measured by computing the Euclidean distance between the model's actual semantic representation and the semantic representation of each word in the training corpus. If the smallest distance was for the target representation then the model was judged to be correct. We examined how the different training focuses affected reading performance at various stages during training.

## Results

## Network Performance

For the pretraining tasks, the model that was trained with 500,1000 , and 2000 presentations achieved $74 \%, 89.6 \%$, and $100 \%$ accuracy on the meaning naming (PS) task and $41.7 \%, 80.2 \%$ and $97.9 \%$ accuracy on the oral vocabulary (PS) task, respectively. This pattern of results is in line with performance of the model when trained with a substantially larger vocabulary (Monaghan et al., 2017). The three training schedules thus reflect different levels of pre-literate oral language skills, from poorer through to near-perfect vocabulary knowledge.

Figure 3 shows the average performance of the OP and OS focused models with the different amounts of pretraining at different stages of reading training. We analysed the model's performance by using generalized linear mixed effects models with accuracies in reading aloud or reading comprehension as the dependent variable, depending on the task. Item and simulation (simulations one to five) were included as random factors, and training focus (OP or OS), reading time (epoch 100 to 1000 ) and pretraining $(500,1000$, or 2000) were included as fixed factors.

Overall, the model performed better on the tasks for which it had undergone intensive training. For reading aloud, the OP focused model performed better than the OS focused model. Adding training focus as a fixed factor resulted in a significant improvement in model fit compared to a model with random effects of item and simulation and with fixed effects of reading time and pretraining, $\chi(1)=398.86, p$ $<.001$. For reading comprehension, the OS focused model performed better than the OP focused models, as again indexed by the fact that adding training focus improved model fit, $\chi(1)=314.25, p<.001$.

However, the effect of pretraining had an asymmetric effect on the reading aloud and reading comprehension tasks, according to whether the model had been trained with OP or OS focus. For reading aloud, the effect of different levels of pretraining, reflecting oral language skills, had a null effect on performance for both the OP and the OS focused models. Adding pretraining as a fixed factor did not result in a significant improvement in model fit compared to a model with random effects of item and simulation and with fixed effects of reading time and training focus, $p>.05$. Note that the trajectories of the lines for the OS focused model for 500
and for 2000 pretraining trials are very close together, as they are for the OP focused model. In contrast for reading comprehension, the effect of pretraining had a substantial effect on the OP training focused model - adding pretraining as a fixed factor improved model fit compared to a model with random effects of item and simulation and with fixed effects of reading time and training focus, $\chi(2)=34.42$, $p$ $<.001$. Specifically, after substantial pretraining (2000 pretraining trials, producing close to $100 \%$ in oral vocabulary and meaning naming tasks), the performance of the OP focused model began to converge with that of the OS focused model. The beneficial effect of the OP training focus is strongest for the model with advanced oral skills prior to literacy onset. This observation was confirmed by the fact that adding the interaction between pretraining and training focus as a fixed factor improved model fit compared to the model containing random and fixed effects, $\chi(2)=$ 9.86, $p<.001$. Looking at each level of pretraining, the difference between the OP focused model and the OS focused model was the smallest for 2000 pretraining trials, $\beta$ $=1.47$, followed by 1000 pretraining trials, $\beta=1.81$, and then 500 pretraining trials, $\beta=2.01$.

These results for the skilled oral language model are in tune with the behavioural results from Taylor et al. (2017). Figure 3 (right) shows the performance of the participants trained with the OP versus OS focus languages on each day taken from Taylor et al.'s figures 3 and 4. Similar to the behavioural data, the performance of the OP and OS focused models converged after substantial training and this is likely due to the fact that the training sample was relatively small.


Figure 3. The performance of the OP and OS focused models with different amounts of pretraining over the time course of the reading training (Left). The performance of the participants trained with the OP and OS focus languages on each day from Taylor et al. (2017, right). The error bars indicate $\pm$ SEM.

## Discussion

We developed a fully implemented connectionist model of
reading that mapped between orthography, phonology, and semantics and explored the influence of oral language on the effectiveness of different types of reading instruction. The laboratory study on which this work was based indicated that focusing on learning mappings between print and sound also transferred to promote mapping between print and meaning, whereas focusing on learning print to meaning mappings resulted in deficiencies in learning print to sound and had little advantage for mapping from print to meaning. The consequences of this, if they extend to children's learning, are that, given limited instructional time, learning should focus on phonics, rather than on meaning-based strategies for reading acquisition.

Our model replicated these effects: a model which focused on print to meaning (i.e., OS training focused model) had deficiencies in learning to map from print to sound, whereas a model which focused on print to sound (i.e., OP training focused model) was better at learning reading aloud tasks, and converged in performance for reading comprehension tasks with the OS training focused model which had three times as much experience of comprehension trials during training.
However, importantly this convergence was dependent upon the model's preliteracy training. Only when the model had high accuracy in its mappings between phonology and semantics was it able to transfer performance from OP training trials to perform well on reading comprehension. This pattern of performance from the OP training focused model with high oral language skills was similar to the behavioural data reported in Taylor et al. (2017). Our computational results demonstrate that the advantage of OP focused training only pertains in cases where good oral language skills are present. This is because the transfer from OP training trials to OS task performance requires effective mappings from phonology to semantics. If these are not present then the effective learning of OP mappings in the model stops just there - any high fidelity representation of phonology cannot then accurately activate the target semantic representation. OP training, then, is only advantageous for reading comprehension when the learner has good oral language knowledge, consistent with the view that addresses the role of oral language in reading (Gough \& Tunmer, 1986; Harm \& Seidenberg, 2004; Plaut et al. 1996).
The results are thus far compatible with empirical evidence of the benefit of both print to sound decoding skills and oral language skills on reading ability (e.g. Curtis, 1980; Nation \& Snowling, 2004; Ouellette \& Beers, 2010; Ricketts, Nation, \& Bishop, 2007), which relate to the two segments of the indirect route from orthography to semantics via phonology. Further investigation of the model's performance will enable us to determine whether this is the way in which the model functions to solve the mapping tasks. We suggest that, for reading aloud, the direct OP pathway is likely to be most effective for performing the task regardless of the training focus, because the systematic mappings are easier to learn compared to the indirect OSP pathway which requires two arbitrary mappings. Thus, more
training of the direct OP pathway is likely to be beneficial. In contrast, for reading comprehension, the indirect OPS pathway may again be more effectively used, because it exploits a regular OP mapping and a previously learned arbitrary PS mapping, whereas the direct OS pathway is arbitrary and needs to be acquired. We might then expect the indirect pathway to have a substantial contribution to reading comprehension performance for both OP and OS focused training, in the context of highly accurate PS mappings.

Previous studies have showed that the division of labour between the phonological and semantic pathways in connectionist models of reading could be shaped by word properties or orthographic systems (Chang, Welbourne \& Lee, 2016; Harm \& Seidenberg, 2004; Plaut et al. 1996). In this work we show that reading instruction and prior oral language skill also seem to alter the division of labour. This is likely due to the broadly systematic versus arbitrary nature of OP versus OS mappings in English.

In summary, our simulation results have demonstrated that oral language skills mediate the effectiveness of reading instruction in early literacy development. In particular, the beneficial effects of print to sound instruction for reading comprehension depend on high levels of oral vocabulary knowledge. Thus, in line with the Simple View of reading, our modelling work suggests that teaching children about spelling-to-sound mappings needs to be accompanied by substantial training on oral vocabulary, in order to promote reading comprehension. Interventions based on promoting print to sound skills should also ensure effective oral language skills, in order to exploit the benefit of enhancing the regularities available in OP mappings in alphabetic writing systems.

## Acknowledgments

This research was supported by ESRC grant RES-000-224049 and ESRC grant ES/L002264/1.

## References

Chang, Y.-N., Welbourne, S., \& Lee, C.-Y. (2016). Exploring orthographic neighborhood size effects in a computational model of Chinese character naming. Cognitive Psychology, 91, 1-23. doi: http://dx.doi.org/10.1016/j.cogpsych.2016.09.001
Chang, Y.-N., Monaghan, P., \& Welbourne, S. (2016). Effects of experience in a developmental model of reading. Paper presented at the Proceedings of the $38^{\text {th }}$ Annual Conference of the Cognitive Science Society, Philadelphia.
Coltheart, M., Rastle, K., Perry, C., Langdon, R., \& Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological Review, 108(1), 204.
Curtis, M. E. (1980). Development of components of reading skill. Journal of Educational Psychology, 72(5), 656.

Davis, A. (2013). To read or not to read: decoding Synthetic Phonics. Impact, 2013(20), 1-38.
Gough, P. B., \& Tunmer, W. E. (1986). Decoding, reading, and reading disability. Remedial and Special Education, 7(1), 6-10.
Harm, M. W., \& Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. Psychological Review, 111(3), 662-720.
Miller, G. A. (1990). WordNet: An on-line lexical database. International Journal of Lexicography, 3, 235-312.
Monaghan, P., Chang, Y.N., Welbourne, S., \& Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. Journal of Memory and Language, 93, 1-21.
Nation, K. (2009). Form-meaning links in the development of visual word recognition. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1536), 3665-3674. doi: 10.1098/rstb.2009.0119
Nation, K., \& Cocksey, J. (2009). Beginning readers activate semantics from sub-word orthography. Cognition, 110(2), 273-278. doi: 10.1016/j.cognition.2008.11.004
Nation, K., \& Snowling, M. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. Journal of Research in Reading, 27, 342-356.
Ouellette, G., \& Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. Reading and Writing, 23, 189-208.
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., \& Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review, 103(1), 56-115.
Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., \& Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. Psychological science in the Public Interest, 2(2), 31-74.
Ricketts, J., Nation, K., \& Bishop, D. (2007). Vocabulary is important for some, but not all reading skills. Scientific Studies of Reading, 11, 235-257.
Seidenberg, M. (2017). Language at the speed of sight: How we read, why so many can't, and what can be done about it. New York: Basic Books.
Seidenberg, M. S., \& McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological Review, 96(4), 523-568.
Taylor, J. S. H., Davis, M. H., \& Rastle, K. (2017). Comparing and validating methods of reading instruction using behavioural and neural findings in an artificial orthography. Journal of Experimental Psychology: General. Advance online publication.
Taylor, J.S.H., Duff, F.J., Woollams, A., Monaghan, P., \& Ricketts, J. (2015). How word meaning influences word reading. Current Directions in Psychological Science, 24, 322-238.

# Evaluating vector-space models of analogy 

Dawn Chen (sdawnchen@gmail.com)<br>Joshua C. Peterson (peterson.c.joshua@gmail.com)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, CA 94720 USA


#### Abstract

Vector-space representations provide geometric tools for reasoning about the similarity of a set of objects and their relationships. Recent machine learning methods for deriving vectorspace embeddings of words (e.g., word2vec) have achieved considerable success in natural language processing. These vector spaces have also been shown to exhibit a surprising capacity to capture verbal analogies, with similar results for natural images, giving new life to a classic model of analogies as parallelograms that was first proposed by cognitive scientists. We evaluate the parallelogram model of analogy as applied to modern word embeddings, providing a detailed analysis of the extent to which this approach captures human relational similarity judgments in a large benchmark dataset. We find that that some semantic relationships are better captured than others. We then provide evidence for deeper limitations of the parallelogram model based on the intrinsic geometric constraints of vector spaces, paralleling classic results for first-order similarity.


Keywords: analogy; word2vec; GloVe; vector space models

## Introduction

Recognizing that two situations have similar patterns of relationships, even though they may be superficially dissimilar, is essential for intelligence. This ability allows a reasoner to transfer knowledge from familiar situations to unfamiliar but analogous situations, enabling analogy to become a powerful teaching tool in math, science, and other fields (Richland \& Simms, 2015). Computational modeling of analogy has primarily focused on comparing structured representations that contain labeled relationships between entities (Gentner \& Forbus, 2011). However, the questions of where these relations come from and how to determine that the relationship between one pair of entities is the same as that between another pair are an unsolved mystery in such models. Some models, such as DORA (Doumas, Hummel, \& Sandhofer, 2008) and BART (Lu, Chen, \& Holyoak, 2012), try to learn relations from examples, but have only demonstrated success on comparative relations such as larger.

Another possibility is that the representations of entities themselves contain the information necessary to infer relationships between entities and that relations do not need to be learned separately. An instantiation of this hypothesis is the parallelogram model of analogy (see Figure 1), first proposed by Rumelhart and Abrahamson (1973) over 40 years ago. In this model, entities are represented as points in a Euclidean space and relations between entities are represented as their difference vectors. Even though two pairs of points may be far apart in the space (i.e., they are featurally dissimilar), they are considered relationally similar as long as their difference vectors are similar. Although Rumelhart and Abrahamson found that this simple model worked well for a small
domain of animal words with vectors obtained using multidimensional scaling, little progress was made on the parallelogram model after the initial proposal, with the exception of a handful of reasonably successful applications (see Ehresman \& Wessel, 1978).

However, in the past few years, the parallelogram model was reincarnated in the machine learning literature through popular word embedding methods such as word2vec (Mikolov, Sutskever, Chen, Corrado, \& Dean, 2013) and GloVe (Pennington, Socher, \& Manning, 2014). These word representations enable verbal analogy problems such as king : queen :: man : ? to be solved through simple vector arithmetic, i.e., $v_{\text {queen }}-v_{\text {king }}+v_{\text {man }}$ results in a vector very close (in terms of cosine distance) to $v_{\text {woman }}$. Word embeddings like word2vec and GloVe have also been used successfully in a variety of other natural language processing tasks, suggesting that these representations may indeed contain enough information for relations to be inferred from them directly. Recently, researchers in computer vision have been successful in extracting feature spaces that exhibit similar properties in both explicit (supervised) (Radford, Metz, \& Chintala, 2015) and implicit (unsupervised) (Reed, Zhang, Zhang, \& Lee, 2015) ways, yielding linearized semantic image transformations such as object rotations and high-level human face interpolations. The potential for applying the parallelogram model of analogy to vector space models appears to be domainagnostic, broadly applicable to both semantic and perceptual domains. This suggests a promising cognitive model and provides the opportunity to evaluate a classic theory in largescale, ecologically valid contexts.


Figure 1: The parallelogram model of analogy completes the analogy king : queen :: man : ? by adding the difference vector between king and queen to man. This forms a parallelogram in the underlying vector space.

In this paper, we evaluate the parallelogram model of analogy as applied to modern vector-space representations of words. Focusing on the predictions that this approach makes about the relational similarity of words, we provide a new dataset of over 5,000 comparisons between word pairs that exemplify 10 different types of semantic relations. We find that the parallelogram model captures human relational sim-
ilarity judgments for some semantic relations, but not others. We then show that human relational similarity judgments for pairs of words violate the geometric constraints of symmetry and triangle inequality, just as Tversky (1977) showed for judgments of first-order similarity between words. This creates a challenge for any vector space model that aims to make predictions about relational similarity.

## Relational Similarity

One way to evaluate vector space models such as word2vec and GloVe as accounts of analogy is to compare their assessments of relational similarity - the similarity of the relation between one pair of words to that of another - with human judgments. A good foundation for this task is the SemEval-2012 Task 2 dataset (Jurgens, Turney, Mohammad, \& Holyoak, 2012), which contains prototypicality scores based on human data for word pairs that exemplify 79 different semantic relations. These relations were taken from a taxonomy of semantic relations (Bejar, Chaffin, \& Embretson, 1991) and are subtypes of 10 general types, such as CLASS-INCLUSION, SIMILAR, and CONTRAST. Participants were given three or four paradigmatic examples of a relation and asked to generate additional examples of the same relation. A total of 3,218 unique word pairs were generated for the 79 relations, with an average of 41 word pairs per relation. A prototypicality score for each participant-generated word pair was calculated based on how often a second group of participants chose the word pair as the best and worst example of the relation among a set of choices. Table 1 shows examples illustrating two representative subtypes of each of the ten general types of relations.

According to the parallelogram model, two pairs of words ( $A: B$ and $C: D$ ) are relationally similar to the extent that their difference vectors $\left(v_{B}-v_{A}\right.$ and $\left.v_{D}-v_{C}\right)$ are similar. How appropriate is this geometric relationship for the various semantic relations? As a preliminary investigation of this question, we projected the 300 -dimensional word2vec vectors into a two-dimensional space using principal components analysis separately for each relational subtype in the SemEval dataset, and visualized the difference vectors for the participant-generated word pairs from each relation. Figure 2 shows the difference vectors for the 20 relational subtypes that are shown in Table 1.

Examining the difference vectors for each relation shows that the parallelogram rule does not appear to capture all relations. CASE RELATIONS Agent:Instrument (e.g., farmer : tractor) shows a nearly perfect correspondence with what we would expect under the parallelogram model, with all difference vectors aligning. However, many of the relations appear to have no clear geometric pattern. Nevertheless, simply looking at projections of the difference vectors is not sufficient to evaluate the power of geometric models of relational similarity to capture various relations, because information is lost in the projections. What is required is a detailed evaluation on judgments of relational similarity between word pairs

Table 1: Examples of word pairs instantiating each of two representative subtypes from each general relation type in the SemEval-2012 Task 2 dataset

| Relation type | Subtype | Example |
| :--- | :--- | :--- |
| 1. CLASS- <br> INCLUSION | Taxonomic <br> Class:Individual | flower $:$ tulip <br> river $:$ Nile |
| 2. PART-WHOLE | Object:Component <br> Collection:Member | car $:$ engine <br> forest $:$ tree |
| 3. SIMILAR | Synonymy <br> Dimensional Simi- <br> larity | car $:$ auto <br> simmer $:$ boil |
| 4. CONTRAST | Contrary <br> Reverse | old $:$ young <br> buy : sell |
| 5. ATTRIBUTE | Item:Attribute <br> Object:State | beggar $:$ poor <br> coward $:$ fear |
| 6. NON-ATTRIBUTE | Item:Nonattribute <br> Object:Nonstate | fire $:$ cold <br> corpse $:$ life |
| 7. CASE | Agent:Instrument <br> RELATIONS | soldier $:$ gun <br> plow $:$ earth |
| 8. CAUSE-PURPOSE | Cause:Effect <br> Cause:Compensa- <br> tory action | joke $:$ laughter <br> hunger $:$ eat |
| 9. SPACE-TIME | Location:Item <br> Time:Associated <br> Item | library $:$ book <br> winter $:$ snow |
| 10. REFERENCE | Sign:Significant <br> Representation | siren $:$ danger <br> diary $:$ person |

within each relation.
Although the SemEval-2012 dataset contains prototypicality scores for the participant-generated word pairs within each relation, which have been interpreted as the relational similarities between the participant-generated pairs and the paradigmatic pairs, prototypicality is influenced by other factors such as the production frequencies of words (Uyeda \& Mandler, 1980). Moreover, because participants were encouraged to focus on the relation illustrated by the paradigmatic examples, the prototypicality scores may not have much to do with the particular word pairs chosen as paradigmatic examples. Experiment 1 aims to address these problems.

## Experiment 1: Benchmarking Relational Similarity

To overcome the limitations of the SemEval-2012 Task 2 dataset for our purposes, we collected a new large dataset that directly measures human judgments of relational similarity between word pairs, focusing on comparisons between word pairs with similar relations.

Participants We recruited 823 participants from Amazon Mechanical Turk. Participants were paid $\$ 2.00$ for the $20-$

1. CLASS-INCLUSION

Taxonomic

3. SIMILAR Dimensional Similarity

$\leftrightarrow \times$
6. NON-ATTRIBUTE Item:Nonattribute

8. CAUSE-PURPOSE Cause:Compensatory Action

4. CONTRAST

6. NON-ATTRIBUTE Object:Nonstate

9. SPACE-TIME Item:Location

4. CONTRAST
Reverse

$\stackrel{\downarrow}{\text { CASE RELATIONS }}$

9. SPACE-TIME Time Associated Item


5. ATTRIBUTE

ItemAttribute


Cause:Effect

10. REFERENCE

Representation
3. SIMILAR

Synonymity

5. ATTRIBUTE


Figure 2: Visualizations of difference vectors for 20 relational subtypes using two-dimensional projections of word2vec word vectors obtained separately for each relation using principal components analysis.
minute study. We excluded 158 participants from the data analysis because they failed two or more of the attention checks (see below).

Stimuli The stimuli for this study were taken from the SemEval-2012 Task 2 dataset. We were mainly interested in how people rate relational similarities between participantgenerated word pairs within each of the 79 relational subtypes. However, because the total number of such "withinsubtype" pairwise comparisons is still enormous, we selected the most representative subtype for each relation type out of the two shown in Table 1. The subtype we chose is the first of each pair of subtypes that appears in Table 1. We then randomly chose 30 word pairs out of the entire participantgenerated set for each of the 10 subtypes and formed all possible within-subtype comparisons between these word pairs. This created a set of 4,350 within-subtype comparisons. Finally, in order to encourage participants to use the entire rating scale, we added 925 "between-subtype" comparisons, which are comparisons between word pairs from different subtypes within a type (e.g., Object:Component and Collection:Member, both subtypes of PART-WHOLE), and 925 "between-type" comparisons, which are comparisons between word pairs from the representative subtypes of different relational types (e.g., Object:Component and Taxonomic CLASS-INCLUSION).

Procedure Participants were given instructions about relational similarity, which included an example of two word pairs that have similar relationships (kitten : cat and chick : chicken) and an example of word pairs with dissimilar relationships (chick: chicken and hen : rooster). Participants then viewed two pairs of words side-by-side on each page and were asked to rate the similarity of the relationships shown by the two word pairs on a scale from 1 (extremely different) to 7 (extremely similar). They rated 100 comparisons in a random order, 70 of which were within-subtype, 15 of which were between-subtype, and 15 of which were between-type. The left-right order of the two word pairs on the screen was chosen randomly (but order within pairs was of course maintained). After every 20 trials, there was an attention check question that asked participants to indicate whether two words are the same or different.

Results \& Discussion We obtained at least 10 good ratings for each comparison, with an average of 10.74 ratings per comparison. The mean rating across all comparisons was $4.52(S D=2.17)$. As expected, we obtained the highest relational similarity ratings for within-subtype comparisons ( $M$ $=5.01, S D=1.98$ ), mid-level ratings for between-subtype comparisons ( $M=4.02, S D=2.14$ ) and the lowest ratings for between-type comparisons ( $M=2.70, S D=1.93$ ).

We calculated relational similarity for each comparison us-
ing word2vec and GloVe word representations. We used the 300 -dimensional word2vec vectors trained on the Google News corpus that were provided by Google (Mikolov et al., 2013), and the 300 -dimensional GloVe vectors trained on a Common Crawl web crawl corpus that were provided by Pennington et al. (2014). We tested two measures of similarity between difference vectors, cosine similarity and Euclidean distance. Specifically, for a given comparison between two word pairs, $A: B$ and $C: D$, letting $\mathbf{r}_{1}=v_{B}-v_{A}$ and $\mathbf{r}_{2}=v_{D}-v_{C}$, we calculated the cosine similarity,

$$
\frac{\mathbf{r}_{1} \cdot \mathbf{r}_{2}}{\left\|\mathbf{r}_{1}\right\|\left\|\mathbf{r}_{2}\right\|}
$$

as well as a similarity measure based on Euclidean distance,

$$
1-\left\|\mathbf{r}_{1}-\mathbf{r}_{2}\right\|
$$

Cosine similarity is typically used to measure similarity in vector spaces such as word2vec and GloVe. However, using Euclidean distance corresponds more closely to the original parallelogram model, in which not only the directions but also the lengths of the difference vectors needed to be similar for two word pairs to be considered relationally similar.

Figure 3 shows Pearson's correlations between predicted relational similarity scores and average human relational similarity ratings on each relation type (including both withinsubtype and between-subtype comparisons) for each vector space and similarity measure. There is considerable variation in the performance of word 2 vec and GloVe in predicting human relational similarity ratings. As might be expected from examining Figure 2, cosine similarity performs the best on CASE RELATIONS (relation 7). However, cosine similarity completely fails on SIMILAR (relation 3), CONTRAST (relation 4), and NON-ATTRIBUTE (relation 6). Euclidean distance boosts performance on the latter two relations, but still underperforms overall compared to most other relations. Nevertheless, Euclidean distance does perform very well on SPACETIME (relation 9).

These results indicate that a single relational comparison strategy cannot capture all semantic relations in the spaces provided. It is unclear if such a result is a reflection of the word embeddings or actual variation in human analogical strategies. Next, we turn to the broader question of the appropriateness of the class of geometric models in general for representing human relational similarity behavior.

## Violations of Geometric Constraints

Distance metrics in vector spaces must obey certain geometric constraints, such as symmetry (the distance from $x$ to $y$ is the same as the distance from $y$ to $x$ ) and the triangle inequality (if the distance between $x$ and $y$ is small and the distance between $y$ and $z$ is small, then the distance between $x$ and $z$ cannot be very large). Cosine similarity, used to measure similarity between word2vec representations, also obeys symmetry and an analogue of the triangle inequality (Griffiths, Steyvers, \& Tenenbaum, 2007). However, psychological representations of similarity do not always obey these
constraints (Tversky, 1977). The famous example of this is that people judge North Korea to be more similar to China than the other way around, a violation of symmetry. Griffiths et al. (2007) examined the word representations derived by Latent Semantic Analysis (Landauer \& Dumais, 1997), another well-known vector space model, and found that these representations are unable to account for violations of symmetry and the triangle inequality in human word association data. Nevertheless, all prior work has focused on first-order similarity between words, and second-order (relational) similarity between word pairs might be expected to follow a different pattern. In this section, we show that human judgments of relational similarity also do not satisfy the geometric constraints of symmetry and the triangle inequality. Vector space models such as word2vec and GloVe cannot account for these violations.

## Experiment 2: Symmetry

In this experiment, we examined whether there were any pairs of word pairs for which participants' judgments of relational similarity changed when the presentation order was reversed. We might expect such asymmetry to occur when a word pair has multiple relations and shares ones of its less salient relations with another word pair. For example, when presented with angry : smile - exhausted : run, one might think, "an angry person doesn't want to smile" and "an exhausted person doesn't want to run," but when presented with exhausted : run - angry : smile, one might think,"running makes a person exhausted, but smiling doesn't make a person angry." Thus, participants might give high relational similarity ratings in the first presentation order and low ratings in the second order.

Participants We recruited 1,102 participants from Amazon Mechanical Turk, who gave informed consent and were paid $\$ 1.00$ for the 10 -minute study. We excluded 99 participants from the data analysis because they failed two or more of the attention checks (see below).
Stimuli We randomly selected 220 within-subtype, 220 between-subtype, and 60 between-type comparisons from all possible comparisons formed using the entire SemEval-2012 Task 2 dataset. We created two versions of each comparison, in which the order of the word pairs were switched.
Procedure Participants were given instructions about relational similarity and the two examples used in Experiment 1 illustrating similar and dissimilar relationships. They saw one word pair in each comparison first and were asked to think of the relationship between the words. Then after a 600 ms delay, the other word pair was shown and participants were asked to rate the similarity of the relationships on a 7-point scale. Participants rated 50 comparisons, including 22 withinsubtype, 22 between-subtype, and 6 between-type comparisons. Each participant viewed each comparison in only one of its presentation orders. After every 10 trials, there was


Figure 3: Pearons's $r$ between human relational similarity ratings and model predictions on different relation types for (a) word2vec and (b) GloVe. The name and examples of each numbered relation type are shown in Table 1.
an attention check question that asked participants to indicate whether two words are the same or different.
Results \& Discussion We obtained about 50 ratings for each comparison in each presentation order. We conducted a $t$-test for each comparison to see if the two presentation orders resulted in significantly different relational similarity ratings. 77 of these $t$-tests were statistically significant at the .05 level. The number of $t$-tests that we would expect to be significant at the $\alpha=0.05$ level if presentation order did not matter for any of the comparisons is 25 . Assuming that the $t$-tests are independent, a binomial test reveals that this deviation is statistically significant, $p<.001$.

Examining the comparisons for which different presentation orders resulted in significantly different relational similarity ratings confirms our guess as to when people's judgments of relational similarity might not obey symmetry. The previously mentioned example of angry: smile and exhausted : run indeed elicited higher ratings in the direction shown here ( 4.76 mean rating) than in the opposite direction ( 2.36 mean rating). As another example, people rated hairdresser : comb - pitcher : baseball as more relationally similar (6.10 mean rating) than pitcher : baseball - hairdresser : comb (4.84 mean rating). In the first presentation order, participants might be thinking that "a hairdresser handles a comb" and "a pitcher handles a baseball," whereas in the second presentation order, they might be thinking "a pitcher plays a specific role in baseball," which doesn't fit with hairdresser : comb.

## Experiment 3: Triangle Inequality

For this experiment, we created triads of word pairs for which we expected people's relational similarity judgments to violate the triangle inequality, such as nurse : patient, mother : baby, and frog : tadpole. This triad violates the triangle inequality because nurse : patient $::$ mother : baby is a good analogy (relationally similar), and so is mother: baby :: frog : tadpole, but nurse : patient $::$ frog : tadpole is not. In this example, the middle pair has multiple relations and shares one of them with the first pair and a different one with the last pair. We presented the two word pairs in each analogy together and asked participants to rate the quality of the analogy rather than relational similarity, because we wanted to encourage participants to consider the two relations together rather than using one relation as a reference.

Participants We recruited 71 participants from Amazon Mechanical Turk, who gave informed consent and were paid $\$ 0.50$ for the 5-minute study. This group of participants did not overlap with the participants in Experiment 2. We excluded 11 participants from the data analysis because they failed one of the attention checks (see below).

Stimuli We created twelve triads of word pairs for which analogy quality judgments are likely to violate the triangle inequality. For every triad, the analogy formed between the first and third word pairs was expected to be rated low and the other two analogies were expected to be rated highly.

Procedure Participants were given instructions about verbal analogies and the two examples used in Experiments 1 and 2 as examples of good and bad analogies, respectively. They were then asked to rate the quality of each analogy on a scale from 1 (very bad) to 7 (very good). For each of the twelve triads, each participant viewed one of the three analogies. Each participant received four analogies formed between the first and second word pairs of various triads (analogy type 1-2), four formed between the second and third word pairs (type 2-3), and four formed between the first and third word pairs (type 1-3). Because two thirds of these analogies are expected to be rated highly, participants also viewed four "filler" analogies expected to be given low ratings. Finally, there were two attention check questions that asked to participants to simply choose 1 (or 7) for a bad (or good) analogy.

Results \& Discussion We obtained 20 ratings for each analogy. We calculated the mean participant rating for each analogy and conducted a one-way between-subjects ANOVA to test if there was an effect of the analogy type (1-2, 2-3, or $1-3$ ) on the mean analogy quality rating. This revealed a significant effect of analogy type, $F(2,33)=45.57, p<0.001$. Post hoc comparisons using the Tukey HSD test indicated that the mean ratings for both type 1-2 analogies $(M=5.44$, $S D=.99$ ) and type 2-3 analogies $(M=5.43, S D=.63)$ were significantly higher than the mean rating for type 1-3 analogies $(M=2.99, S D=.46), p<.001$, whereas the mean ratings for type 1-2 and type 2-3 analogies did not differ significantly from each other. This is consistent with our expectation that types 1-2 and 2-3 analogies would both be rated highly, whereas type 1-3 analogies would be rated lowly. These results indicate that participants' analogy quality ratings violated the triangle inequality.


Figure 4: Mean human ratings and predicted relational similarities (scaled to the range 0-1) for the triad lawyer : books, chemist : beakers, and librarian : books. Error bars indicate 1 SEM for scaled human ratings.

We obtained the predicted relational similarity between the word pairs in each analogy by calculating the cosine similarity between difference vectors using word2vec and GloVe representations. Then we conducted separate ANOVAs for the two representational spaces to test whether there was an effect of analogy type on the predicted relational similarity in each space. Neither ANOVA indicated a significant effect of analogy type. For word2vec, $F(2,33)=1.20, p=.31$. For GloVe, $F(2,33)=.24, p=.79$. These results indicate that the predictions of relational similarity made by word2vec and GloVe do not violate the triangle inequality for our stimuli.

Because each participant contributed ratings to more than one analogy (although only one per triad), the observations are not entirely independent in the first overall ANOVA that we conducted on the participant data. Thus, we conducted a separate between-subjects ANOVA for each of the twelve triads to test if there was an effect of analogy type (1-2, 23 , or 1-3) on the analogy quality ratings for each triad. All twelve ANOVAs were significant, with every $p<.01$. We then conducted post hoc comparisons using the Tukey HSD test. The pattern we expected was that types 1-2 and 2-3 analogies would have significantly higher ratings than type 1-3 analogies, but would not differ significantly from each other. We observed this pattern for seven of the twelve triads. For every one of the remaining triads, the mean ratings for the type 1-2 and type 2-3 analogies were higher than the mean rating for the type 2-3 analogy, but which differences were statistically significant differed among the triads. Figure 4 shows an example of one of the seven triads with the expected pattern and compares the mean participant ratings and predicted relational similarities for the three analogies.

## Conclusions

Our results provide a clearer picture of the utility of vectorspace models of analogy. The parallelogram model makes good predictions of human relational similarity judgments for some relations, but is less appropriate for others. For example, consider the word pairs represented as vectors in Figure 2. As one would expect, relation SIMILAR seems to be best represented by a short difference vector rather than the direction of the difference vector. More generally, in more complex analogies with the source and target each consist-
ing of many points in the vector space, one could imagine many ways of describing relationships between the two sets of points.

More challenging are the constraints posed by the geometric axioms. In our datasets, we found considerable violations of two of these axioms, which cannot be overcome through better embedding methods. In light of this, it would be interesting to follow the history of models of first-order similarity in considering the use of featural representations (Tversky, 1977), exploring methods of measuring similarity in vector spaces that are no longer subject to the constraints imposed by the metric axioms (Krumhansl, 1978), or reformulating the problem as probabilistic inference (Griffiths et al., 2007).
Acknowledgments. Supported by grant number FA9550-13-1-0170 from the Air Force Office of Scientific Research.

## References

Bejar, I. I., Chaffin, R., \& Embretson, S. (1991). A taxonomy of semantic relations. In Cognitive and psychometric analysis of analogical problem solving (pp. 55-91). Springer.
Doumas, L. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological review, 115(1), 1.
Ehresman, D., \& Wessel, D. L. (1978). Perception of timbral analogies. Centre Georges Pompidou.
Gentner, D., \& Forbus, K. D. (2011). Computational models of analogy. Wiley interdisciplinary reviews: cognitive science, 2(3), 266-276.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114(2), 211.
Jurgens, D. A., Turney, P. D., Mohammad, S. M., \& Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In Proceedings of the first joint conference on lexical and computational semantics-volume 1 (pp. 356364).

Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review, 85, 450463.

Landauer, T. K., \& Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211.
Lu, H., Chen, D., \& Holyoak, K. J. (2012). Bayesian analogy with relational transformations. Psychological review, 119(3), 617.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., \& Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
Pennington, J., Socher, R., \& Manning, C. D. (2014). Glove: Global vectors for word representation. In Emnlp (Vol. 14, pp. 15321543).

Radford, A., Metz, L., \& Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
Reed, S. E., Zhang, Y., Zhang, Y., \& Lee, H. (2015). Deep visual analogy-making. In Advances in neural information processing systems (pp. 1252-1260).
Richland, L. E., \& Simms, N. (2015). Analogy, higher order thinking, and education. Wiley Interdisciplinary Reviews: Cognitive Science, 6(2), 177-192.
Rumelhart, D. E., \& Abrahamson, A. A. (1973). A model for analogical reasoning. Cognitive Psychology, 5(1), 1-28.
Tversky, A. (1977). Features of similarity. Psychological review, 84(4), 327.
Uyeda, K. M., \& Mandler, G. (1980). Prototypicality norms for 28 semantic categories. Behavior Research Methods \& Instrumentation, 12(6), 587-595.

# Anticipation Effect after Implicit Distributional Learning 

Danlei Chen (dchen37@u.rochester.edu)<br>Carol A. Jew (carol.jew@rochester.edu)<br>Benjamin Zinszer (bzinszer@ur.rochester.edu)<br>Rajeev D. S. Raizada (rajeev.raizada@gmail.com)<br>Rochester Center for Brain Imaging<br>Department of Brain and Cognitive Sciences<br>University of Rochester<br>Rochester, NY 14627 USA


#### Abstract

Distributional learning research has established that humans can track the frequencies of sequentially presented stimuli in order to infer the probabilities of upcoming events (e.g., Hasher \& Zacks, 1984). Here, we set out to explore anticipation of a stimulus after implicit distributional learning. We hypothesize that as people learn the category frequency information implicitly, response times will scale according to the relative frequency of the stimulus category. Twelve adult participants viewed photographs of faces, tools, and buildings while performing a simple classification task. We found that response times significantly decreased with greater frequencies in the distribution of stimulus categories. This result suggested that distributional information about the internal representations of the stimuli could be learned and indicated the possibility that participants anticipated the stimuli proportional to the probability of the category appearing and thereby reduced response times for the more frequent categories.


Keywords: statistical learning; implicit distributional learning; anticipation; classification

## Introduction

Although complicated and dynamic, our sensory environment contains regularities distributed both spatially and temporally. Previous studies have shown that humans can acquire information from a probabilistic structure, and they are able to predict the upcoming stimulus using distributional knowledge. In this study, we used behavioral methods to investigate anticipation prior to object classification after distributional learning of the object category frequencies.

People are known to be sensitive to the distributional information, and they are able to actively use this information to make complex inferences, such as identifying underlying structures in sequences. Explicit probabilistic information can aid human decision-making in many situations (Arkes, Dawes, \& Christensen, 1986; Wiggs, 1993; Lin, Kung, \& Lin, 1997). In addition to explicit distributional learning, in fact, it is well established in the fields of human development, language acquisition, attention, and perception that people are sensitive to implicit distributional information (e.g., Attneave, 1953; Fiser \& Aslin, 2002; Hasher \& Zacks, 1984; Saffran, Aslin, \& Newport, 1996; Tryk,1968; Turk-Browne, Scholl, Chun, \& Johnson, 2009; Pelucchi, Hay, \& Saffran,
2009). In these studies, distributional information was not explicitly provided to participants, but the results showed that participants could track the stimulus input to infer its underlying causal structure and therefore make accurate predictions or judgments about which stimuli potentially fit or violate this structure. Thus, even when these statistical relationships are not explicitly presented and the stimuli are too numerous to be explicitly counted, people can discover an accurate distributional model of the input.

Particularly, classification tasks have been used to test implicit distributional learning (Forster \& Chambers, 1973; Stanners, Forbach \& Headley, 1971; Stanners, Jastrzembski, \& Westbrook, 1975; Whaley, 1978). Classifying responses can reflect distributional learning processes, as learned items can be recognized and discriminated from other items faster than unfamiliar items can. For example, Whaley (1978) found that response times for word and non-word classification were substantially faster with high-frequency initial and final consonants than for words with lowfrequency consonants in initial or final position or both. Although in the context of language, this finding shed lights on the correlation between implicit distributional learning and response times, and it demonstrated the methodology of using a classification task to test this correlation.

Response time has been used to measure anticipation in many studies (Haith, Hazan, \& Goodman, 1988; Hinrichs \& Krainz, 1970; Todorovic, van Ede, Maris, \& de Lange, 2011; Turk-Browne, Scholl, Johnson, \& Chun, 2010, Poulton, 1950). Some of these studies have found that when participants were instructed to predict the upcoming stimulus, response times were faster for correct predictions than for incorrect predictions (Bernstein \& Reese, 1965; Hinrichs et al., 1970). This finding suggests that anticipation of an upcoming stimulus influences the response time in the subsequent trial.

However, the effect of adult observers' use of implicit distributional learning on anticipation of the category of an upcoming stimulus remains largely unexplored. Most studies have focused on effect of frequency information about stimulus-stimulus association (e.g. Conway \& Christiansen, 2005; Kirkham, Slemmer, Johnson, 2002; O’Brien \& Raymond, 2012; Olson \& Chun, 2001; Turk-Browne, Jungé, Scholl, 2005), and little research have looked into the effect of the overall distributional information about the internal
representation of the stimuli (e.g. the categorical representation of the object).

Here, we aimed to establish evidence for anticipatory representations of the category of the upcoming stimulus emerging from distributional learning. Turk-Browne et al. (2010) investigated implicit anticipation triggered by probabilistic information. Their behavioral result showed that when the participants observed and made classification responses to every trial, the participants reacted faster to the trials that can be predicted from their immediate preceded trials. Based on this finding, in the experiment, we measured participants' anticipation of an object category by measuring the response times in a sequential classification task. We examined the anticipatory effects of the underlying distribution and predicted that response times for classification would decrease with greater category frequencies, suggesting that as people learned the category distributional information implicitly, anticipation was scaled according to the probability of the category appearing.

## Methods

## Participants

Twelve participants were recruited ( mean age $=20$ years; $S D$ $=1.7$ years; 7 females, 5 males) and were compensated $\$ 10$ per hour. All participants were undergraduate students at the University of Rochester. All participants reported being right-hand dominant. The experiment took around 45 minutes to complete. The study procedures were approved by the Institutional Review Board of the University of Rochester, and participants received an informed consent document prior to the study.

## Materials

We chose three categories of stimuli: faces, buildings, and tools. Each category has specific brain areas that reliably respond to one of these categories but not the others (Epstein \& Kanwisher, 1998; Kanwisher, McDermott, \& Chun, 1997; Chao \& Martin, 2000). The images from each of these categories were grey-scaled and edited to be the same size ( $640 \times 480$ pixels) using Preview software in Mac OSX. The images appeared in the middle of a $27{ }^{\prime \prime}$ iMAC monitor with $1920 \times 1080$ resolution. The images appeared in the middle of the screen against a white background. Face images were acquired from the Chicago Face Database (Ma, Correl, \& Wittenbrink, 2015); Building images were downloaded by Google Image search with the keywords "building" and "house"; Tool images were obtained from the BOSS database (Brodeur, Dionne-Dostie, Montreuil, \& Lepage M, 2010).

The frequency of each category ( $60 \%, 30 \%$, or $10 \%$ ) was counterbalanced across six different distributional conditions using Latin Squares (Winer, 1962). This manipulation counterbalanced carryover effects between conditions and ensured that participants see each of the conditions in the study. We chose $60 \%, 30 \%$, and $10 \%$ as frequencies considering the number of trials in each block ( 30 trials) and condition ( 90 trials). These frequencies produce integer instead of the decimal number of trials in each condition. The
complete information about these six conditions is shown in Table 1.
Three adjacent buttons on the computer keyboard were marked as " F ", " H " and " T ". To exclude the motor-related confounds that were the interest of this study, the key mapping was counterbalanced across subjects. Subjects were asked to always use the same three fingers (index, middle, and ring fingers) for the same keys.

Table 1: Distribution of categories in each condition

|  | Faces | Buildings | Tools | Num. of <br> Trials |
| :---: | :---: | :---: | :---: | :---: |
| Condition 1 | $60 \%$ | $30 \%$ | $10 \%$ | 90 |
| Condition 2 | $60 \%$ | $10 \%$ | $30 \%$ | 90 |
| Condition 3 | $10 \%$ | $60 \%$ | $30 \%$ | 90 |
| Condition 4 | $30 \%$ | $60 \%$ | $10 \%$ | 90 |
| Condition 5 | $10 \%$ | $30 \%$ | $60 \%$ | 90 |
| Condition 6 | $30 \%$ | $10 \%$ | $60 \%$ | 90 |
| Num. of <br> Trials | 180 | 180 | 180 | 540 |



Figure 1: Illustration of the experiment experimental protocol of first three trials in one condition. After participants read the instructions, images appeared on the screen and participants responded accordingly by pressing corresponding buttons. A score would appear above the picture after the participant pressed a button.

## Procedure

Subjects were asked to perform a simple classification task. The presentation of stimuli was programmed using MATLAB Psychophysics toolbox (Brainard, 1997; Kleiner, Brainard, Pelli, Ingling, \& Murray, 2007; Pelli, 1997). The experiment took place in a behavioral testing cubicle. During each condition, the participants were instructed to press a designated key to indicate the category of each presented stimulus. A score was given based on the reaction time
immediately after each response (zero for inaccurate or missing trials), and a total score was presented after each condition. The scores were presented in order to provide feedback and motivate participants to give faster and more accurate responses, and they were not used in the analysis. Stimulus-onset asynchrony (SOA; 2000, 4000, and 6000 ms ) was varied to prevent participants from predicting when the next stimulus would appear on the screen. Each trial had a fixed image duration of 1000 ms , and the image would not disappear after a response was recorded. Inter-trial interval (ITI) varied based on the SOA of that trial. Each subject went through all 6 conditions. Each condition had 3 blocks, and each block displayed 30 images. In total 90 images were presented in each condition, and 540 images in the whole experiment. The breaks between blocks were 15000 ms , and the breaks between conditions were two minutes. Instructions between each condition were designed to cue the participants to the new distributional information of the next condition: "Thank you for finishing the task. Please take a short break. A new but similar task will start in two minutes." The illustration of the experiment experimental protocol of first three trials in one condition is shown in Figure 1.

## Results

In total, 6480 responses were recorded $(M=521 \mathrm{~ms}$; $S D=$ 107 ms ; Accuracy $=0.948$ ). For analysis, we excluded the trials with incorrect responses or no responses. And across all participants, 336 out of 6480 trials ( $0.05 \%$ ) were excluded for this reason. Paired two-tailed t-tests showed that for all twelve subjects, mean response times of block 2 for each condition were significantly less than those of block 1 , $t=2.255, p=0.038<0.05$; average response times of block 3 for each condition were significantly less than those of block 1 , $t=2.585, p=0.019<0.05$; but average response times of block 3 for each condition were not significantly less than average response times of block $2, t=-0.683, p=0.504$. It is possible that subjects needed several trials to acquire the distributional information of the current condition and to replace the carryover distributional information from previous conditions. And after obtaining the current distributional information, the participants were able to perform the task using this knowledge. Thus, we excluded all trials from block 1 of each condition. The analyses only contained correct responses from block 2 and block 3. Mean, standard
deviation, and accuracy of response times for each category in each frequency are shown in Table 2.

Using linear mixed-effects models in R , we compared response times for different stimuli in each distribution. We first used a 3-way interaction model (SOA $\times$ category $\times$ frequency), but did not find any significant 3-way interaction. Instead, a reduced 2 -way interaction model (category $\times$ frequency + SOA $\times$ category + SOA $\times$ frequency) was estimated and reported in Table 3. Face was chosen as the arbitrary baseline category by R to prevent multi-collinearity in the indicator variables for the stimulus category.

SOA did not significantly interact with Frequency, and only marginally varied across Category ( $p=0.07$ ). A main effect of SOA was also significant ( $p=0.006$ ). Therefore, although SOA was not a variable of particular interest, we retained it in the model to control for the possible effects of the pre-stimulus waiting period on the anticipatory representation and thus on the change in response times.

Similarly, no main effect for Category was found, but we retained this term due to its significant interaction with Frequency (see below) and to account for Category-specific differences in baseline response rate (e.g., participants responded to faces faster than tools and buildings).

The main effect of frequency was highly significant ( $p<$ 0.001 ), consistent with our hypothesis that response times would differ as a function of the stimulus distribution. The post-hoc paired $t$-tests on subject-level mean response times found that response times indeed significantly decrease as the frequency of the stimulus category increased ( $60 \%$ vs. $30 \%$ : $t(11)=-6.08, p<0.001 ; 30 \%$ vs. $10 \%: t(11)=-5.01, p<$ 0.001).

The interaction between frequency and condition was also significant, so we examined the frequency effects specific to each category of the stimuli. We found that within each of the categories, response times generally decreased as its frequency increased (see Figure 2 for the summary of tests), although many of these tests would not survive correction for multiple comparisons. With the results from the linear mixedeffect models and the $t$-tests, we can conclude that this effect is in line with our hypothesis that participants' response times reduced proportionally to the increasing frequency of the category.

We also looked at the estimated slopes for the frequency $\times$ category interaction in the model (Figure 3), which was

Table 3: ANOVA performed on linear mixed-effects model with 2-way interactions

|  | SS | MS | Num. DF |  | Den. DF | F | -value |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| SOA | 0.38 | 0.38 |  | 1 | 6119.5 | 7.644 | 0.006 ** |
| Category | 1.63 | 0.82 |  | 2 | 6119.3 | 0.676 | 0.509 |
| Frequency | 4.72 | 4.72 |  | 1 | 6119.5 | 40.628 | $<0.001$ *** |
| Category $\times$ Frequency | 0.34 | 0.17 |  | 2 | 6119.2 | 6.224 | 0.002 ** |
| SOA: Category | 0.14 | 0.07 |  | 2 | 6119.3 | 2.574 | 0.076 |
| SOA: Frequency | 0.05 | 0.05 |  | 1 | 6119.5 | 1.963 | 0.161 |
| Notes. . $\mathrm{p}<.1$; ${ }^{*} \mathrm{p}<.05 ; * * \mathrm{p}<.01 ;{ }^{* * *} \mathrm{p}<.001$ |  |  |  |  |  |  |  |

highly significant in the ANOVA. Steeper slopes indicate a stronger influence of frequency on that category. It is clear that frequency influenced face more than it influenced tool, and it influenced tool more than it influenced building (slope(face) $=-2.123$; slope $($ building $)=-1.156 ;$ slope $($ tool $)=$ -1.395).


Figure 2: Average response times for each category in each frequency. The error bars were indicated by one standard error of the mean.


Figure 3: Frequency $\times$ Category Effect

## Discussion

In this study, we showed behavioral evidence that response times decreased with the higher frequency of occurrence of the upcoming stimulus. The evidence consisted of lower response times for categories with the higher frequencies of occurrence in the input as opposed to the category with lower frequencies of occurrence. Higher probability could be reasonably more predictable and therefore facilitated
classification response and reduced response times. This result could be explained if the response to each category engaged anticipatory processes that completed with the overall probability information across all categories. The results were in line with our hypothesis that the anticipatory representation acquired through distributional learning affects responses in a classification task by allowing faster response times according to the frequency of a category appearing.

Previous studies have been focused on frequency information learning between specific stimulus-stimulus associations (e.g. Conway et al. 2005; Kirkham et al. 2002; O’Brien \& Raymond, 2012; Olson \& Chun, 2001; TurkBrowne et al. 2005). However, no study, to our knowledge, has looked into the effect on distributional information about the internal representations of the stimuli. And here we present robust evidence that implicit distributional information about the internal representations of the stimuli could be learned, subsequently facilitated responses to trials of the more frequent category, and therefore caused the anticipation effect.

We also found that participants could learn new distributional information and this new information could override the previously learned distribution relatively quickly (i.e., within 30 trials or one block of the experiment). Although our finding gives a rather coarse estimate of distributional learning efficiency due to the use of response times as an index of learning with relatively low resolution, this result provides strong evidence for on-line learning, because the participants were required to give a response on every trial. Some previous studies relied on off-line learning tests so that they have not been able to study the speed of the distributional learning. Other studies that looked into the online learning of probabilistic information also suggested that probabilistic information could be obtained quickly (Abla, Katahira, \& Okanoya, 2008; Turk-Browne et al., 2009), although these studies used probabilistic information about stimulus-stimulus associations instead of the overall distributional pattern of the stimuli.

Additionally, results showed that building and tool categories were less affected by frequency than faces were. Humans are highly experienced at recognizing faces for evolutionary purposes (Leopold \& Rhodes, 2010; Little, Jones, \& DeBruine, 2010; Sheehan \& Nachman, 2014), and therefore it is possible that human faces can be more quickly recognized and distinguished than other categories can. The perception of faces might have a lower minimal response time in the high-frequency condition, and thus participants' performance for faces was far faster than the other conditions at the $60 \%$ frequency.

At the beginning of the paper, we intended to measure the anticipation of the probabilistically distributed category using comparisons between response times. The above results, using response times as the indication of the anticipation effect, showed that the participants successfully learned and used the distributional information. However, although response time has been used in some studies as an indication
of anticipation (Haith et al., 1988; Hinrichs et al., 1970; Todorovic et al., 2011; Poulton, 1950), it does not directly measure the neural response to anticipatory effect after distributional learning. It is possible that the categories were also directly encoded or primed at relative magnitudes in the brain as a function of frequency, producing this response time effect. This promising result from this behavioral experiment points to the possibility of a future experiment using neuroimaging techniques (e.g., fMRI) to test the hypothesis that probabilistically weighted brain activity also corresponds to the category frequencies, and can be found in the neural activity immediately prior to each trial.

Further studies can combine our behavioral results with the ability to detect categorical specific activation using fMRI to explore the neural basis of anticipation after implicit distributional learning. The adaptive nature of human categorization assumes that categorization reflects the optimal estimates of the probability of unseen features of objects (Anderson \& Milson, 1989). Turk-Browne et al. (2010) identified a neural mediator of anticipation for stimuli as a consequence of implicit distributional learning of paired and unpaired images using fMRI. A region of interest analysis of this study found increased activation of the category-specific brain area from the anticipation of that category and suppressed activation of the area when the predictive stimulus was from another category. These findings suggest that category-specific cortical activation due to implicit perceptual anticipation after implicit probabilistic learning is detectable in the category-specific brain regions using fMRI.

In sum, our study gave behavioral evidence that anticipation for the category of the upcoming stimulus is proportional to the distribution over all the categories. In the future, we hope to see neuroimaging experiment that shows anticipation after distributional learning can be measured in brain activity, and the representation is proportional to the learned distribution.

## Acknowledgement

We thank Donias Doko for pilot experiment design and data collection, J. J. Christine Otto for participant recruitment and data collection, and Andrew J. Anderson for discussions. This research was supported by NSF Grant \#1228261 to Rajeev D. S. Raizada and the Discover Grant from the University of Rochester Undergraduate Office.

## Reference

Abla, D., Katahira, K., \& Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. Journal of Cognitive Neuroscience, 20(6), 952964.

Anderson, J. R., \& Milson, R. (1989). Human memory: An adaptive perspective. Psychological Review, 96(4), 703.
Arkes, H. R., Dawes, R. M., \& Christensen, C. (1986).
Factors influencing the use of a decision rule in a probabilistic task. Organizational Behavior and Human

Decision Processes, 37(1), 93-110.
Attneave, F. (1953). Psychological probability as a function of experienced frequency. Journal of Experimental Psychology, 46(2), 81.
Bernstein, I. H., \& Reese, C. (1965). Behavioral hypotheses and choice reaction time. Psychonomic Science, 3(1-12), 259-260.
Brainard, D. H. (1997). The psychophysics toolbox. Spatial vision, 10, 433-436.
Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., \& Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. PloS one, 5(5), e10773.
Chao, L. L., \& Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. Neuroimage, 12(4), 478-484.
Conway, C. M., \& Christiansen, M. H. (2005). Modalityconstrained statistical learning of tactile, visual, and auditory sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(1), 24.
Epstein, R., \& Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598-601.
Fiser, J., \& Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences, 99(24), 15822-15826.
Forster, K. I., \& Chambers, S. M. (1973). Lexical access and naming time. Journal of verbal learning and verbal behavior, 12(6), 627-635.
Haith, M. M., Hazan, C., \& Goodman, G. S. (1988). Expectation and anticipation of dynamic visual events by 3.5-month-old babies. Child development, 467-479.

Hasher, L., \& Zacks, R. T. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. American Psychologist, 39(12), 1372.
Hinrichs, J. V., \& Krainz, P. L. (1970). Expectancy in choice reaction time: Anticipation of stimulus or response? Journal of Experimental Psychology, 85(3), 330.
Kirkham, N. Z., Slemmer, J. A., \& Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. Cognition, 83(2), B35-B42.
Kanwisher, N., McDermott, J., \& Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience, 17(11), 4302-4311.
Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., \& Broussard, C. (2007). What's new in Psychtoolbox-3. Perception, 36(14). Retrieved from http://www.kyb.mpg.de/fileadmin/user_upload/files/publi cations/attachments/ECVP2007Kleinerslides_5490\%5b0 \%5d.pdf
Leopold, D. A., \& Rhodes, G. (2010). A comparative view of face perception. Journal of Comparative Psychology, 124(3), 233.
Lin, S. H., Kung, S. Y., \& Lin, L. J. (1997). Face
recognition/detection by probabilistic decision-based neural network. IEEE transactions on neural networks, 8(1), 114-132.
Little, A. C., Jones, B. C., \& DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1571), 1638-1659.
Ma, D. S., Correll, J., \& Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. Behavior research methods, 47(4), 1122-1135.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial vision, 10(4), 437-442.
Pelucchi, B., Hay, J. F., \& Saffran, J. R. (2009). Statistical Learning in a Natural Language by 8-Month-Old Infants. Child development, 80(3), 674-685.
Poulton, E. C. (1950). Perceptual anticipation and reaction time. Quarterly Journal of Experimental Psychology, 2(3), 99-112.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants.
Sheehan, M. J., \& Nachman, M. W. (2014). Morphological and population genomic evidence that human faces have evolved to signal individual identity. Nature communications, 5, 4800.
Stanners, R. F., Forbach, G. B., \& Headley, D. B. (1971). Decision and search processes in word-nonword classification. Journal of Experimental Psychology, 90(1), 45.

Stanners, R. F., Jastrzembski, J. E., \& Westbrook, A. (1975). Frequency and visual quality in a word-nonword classification task. Journal of Verbal Learning and Verbal Behavior, 14(3), 259-264.
Todorovic, A., van Ede, F., Maris, E., \& de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. Journal of Neuroscience, 31(25), 9118-9123.
Tryk, H. E. (1968). Subjective scaling of word frequency. The American Journal of Psychology, 81(2), 170-177.
Turk-Browne, N. B., Scholl, B. J., Chun, M. M., \& Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. Journal of cognitive neuroscience, 21(10), 1934-1945.
Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., \& Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. Journal of Neuroscience, 30(33), 11177-11187.
Turk-Browne, N. B., Jungé, J. A., \& Scholl, B. J. (2005). The automaticity of visual statistical learning. Journal of Experimental Psychology: General, 134(4), 552.
O’Brien, J. L., \& Raymond, J. E. (2012). Learned predictiveness speeds visual processing. Psychological Science, 0956797611429800.
Olson, I. R., \& Chun, M. M. (2001). Temporal contextual cuing of visual attention. Journal of Experimental Psychology Learning Memory and Cognition, 27(5), 1299-
1313.

Whaley, C. P. (1978). Word—nonword classification time. Journal of Verbal Learning and Verbal Behavior, 17(2), 143-154.
Wiggs, C. L. (1993). Aging and memory for frequency of occurrence of novel, visual stimuli: Direct and indirect measures. Psychology and Aging, 8(3), 400.
Winer, B. (1962). Latin squares and related designs. Statistical principles in experimental design, 514-5.

# Analytic Causal Knowledge for Constructing Useable Empirical Causal Knowledge: Two Experiments on Preschoolers 

Patricia W. Cheng (cheng@lifesci.ucla.edu)<br>Department of Psychology, University of California, Los Angeles, CA 90095 USA<br>Mimi Liljeholm (m.liljeholm@uci.edu)<br>Department of Cognitive Sciences, University of California, Irvine, CA 92697 USA<br>Catherine M. Sandhofer (sandhof@psych.ucla.edu)<br>Department of Psychology, University of California, Los Angeles, CA 90095 USA


#### Abstract

The present paper examines what domain-general causal knowledge reasoners need for at least some outcome-variable types to construct useable content-specific causal knowledge. In particular, it explains why it is essential to have analytic knowledge of causal-invariance integration functions: knowledge for predicting the expected outcome assuming that the empirical knowledge acquired regarding a causal relation holds across the learning context and an application context. The paper reports two studies that support the hypothesis that preschool children have such knowledge regarding binary causes and effects, enabling them to generalize across contexts rationally, favoring the causal-invariance hypothesis over alternative hypotheses, including interaction (e.g., linear) integration functions, heuristics, and biases.


Keywords: Causal induction; causal learning; causal invariance, rationality; cognitive development

## Introduction

How do we humans best represent the world so that we are able to achieve desired outcomes? A basic requirement is that the world knowledge we acquire be useable. Whenever we use our past knowledge to achieve a desired outcome (e.g., avoid a certain food to prevent a skin or intestinal reaction), we are inevitably generalizing from the learning context (e.g., items for meals at home preceding past allergic reactions) to a subsequent new context (lunch at work the next day, food during foreign travel). By contexts with respect to a cause in question, we mean occasions or settings where (known or unknown) enabling conditions and alternative causes of the target outcome may occur with different probabilities.

By adulthood, humans appear to make causal judgments that suggest they assume causal invariance - namely, that causes operate in an invariant manner across the learning and application contexts - as a default and as a criterion for revising causal knowledge (e.g., Cheng, 1997; Liljeholm \& Cheng, 2007; Lu, Rojas, Beckers \& Yuille, 2016; Lu, Yuille, Liljeholm, Cheng \& Holyoak, 2008). If the concept of causal invariance is essential to the construction of useable causal knowledge, we would expect young children to use it just as adults do. Alternatively, if the concept operates as an acquired strategy or heuristic in causalknowledge construction, young children would be less likely to use it, especially when its use requires
mathematical skills that are far beyond the children's general level of mathematical capability.

A large literature on children's causal reasoning shows that children are able to reason causally from a young age (e.g., Gopnik, 2009; Gweon \& Schulz, 2011; Legare, 2012; Rakison \& Krogh, 2012). For example, like adults, children can learn deterministic conjunctive or disjunctive causal relationships and generalize the relationship that better fits the evidence to other variables (Lucas, Bridgers, Griffiths \& Gopnik, 2014). However, there has been little work on the essentiality of the causal-invariance concept in the shaping of causal knowledge. In particular, it is not known whether children use that concept, rather than simple approximations or heuristics. Young children's use of a probabilistic form of causal invariance would provide especially strong support for its essentiality.

To see why knowledge of causal invariance is essential for constructing useable causal knowledge, consider situations in which a) there may be background causes present, b) these causes may vary from context to context, and c) the set of candidate causes under evaluation may not include one that generalizes well across contexts. Natural settings often hold these challenges. When we want to infer what cures an illness, for example, the illness must have some non-zero probability of occurring due to some background generative cause. The illness need not occur across all individuals, suggesting that background alternative preventive causes may be present. And the illness may be more or less prevalent in different contexts (e.g., countries). The fact that the "no confounding" condition is a standard principle in experimental design is an indication of the pervasive need for the influence of a target cause to be teased apart from that of background causes. A further challenge is that our initial parsing of events to isolate distinct candidate causes may not yield predictions that generalize to application contexts. Moreover, generalizability is a matter of degree (Woodward, 2000, 2003). We may encounter occasions on which a relation that we have assumed to be generalizable unexpectedly fails to hold (e.g., when on a trip up a tall mountain we find that eggs boiled the usual amount of time remain uncooked). The replicability crisis in medical research is a reminder of failures to generalize even in costly planned investigations, not to say in everyday inferences. The need to go beyond one's current set of candidate causes is ever present.

Given the goal of formulating useable causal knowledge, information about a failure to reach that goal - failure indicated by a notable deviation from the outcome expected assuming that the acquired knowledge generalizes when applied - would be useful for assessing whether to retain or revise that knowledge. Along with our colleagues, we have proposed that mathematical functions characterizing the sameness of influence of a cause across contexts - functions which differ depending on the form of the cause and effect variables (e.g., binary vs continuous) rather than their content (e.g., tobacco smoking causes lung cancer) - play a critical role in the construction of causal knowledge (e.g., Cheng, Liljeholm \& Sandhofer, 2013; Cheng \& Lu, in press). We term these causal-invariance functions. Whenever there are too many possible causal models to exhaustively evaluate, causal invariance is a helpful signal.

We have further noted that the vastness of the search space of possible causal representations renders the use of causal invariance not merely helpful but essential. A basic tenet of cognitive science -- that our perception and conception of reality are our representations -- implies that the search space of the representation of reality is infinite. In an infinite search space, an exhaustive evaluation of the possible causal models is not merely practically infeasible, but in principle impossible. Given the nature of the problem of causal knowledge construction, the need to go beyond one's current candidate causes becomes clear. Deviation from the outcome expected based on causal-invariance functions serves as an essential navigating device.

What if the need for revision is signaled instead by deviation from a causal-interaction (i.e., non-causalinvariance) criterion? In that case, that is, if candidate $c$ 's influence on target effect $e$ is expected to vary depending on the state of the background causes, there would be a deviation from expectation -- signaling a need to revise causal knowledge -- when the influence of $c$ in fact generalizes across contexts. Conversely, no deviation from expectation would confirm that $c$ interacts with background causes (its inferred influence therefore should not generalize across contexts). But no deviation from expectation means no signal to revise. Given an inverted signal to revise, in the infinite search space of possible representations of reality, the acquired causal knowledge is unlikely to hold when applied or to replicate when further tested.

If our thesis on the essentiality of the concept of causal invariance is correct, we would expect young children to use the concept, even when its use requires mathematical skills that are far beyond the children's general level of mathematical competence, and even though such usage contradicts an irrational but common practice in medical or business research. Our two studies on preschool-aged children tested their use of a causal-invariance versus a causal-interaction criterion.

## Analytic Knowledge of Causal Invariance

For all situations, every observed outcome is inherently the outcome due to the totality of its causes; the contributing
causal relations are not differentiable by observation. When background causes are present, the unobservability of causation requires that causal learners adopt an assumption (either tacitly or explicitly) regarding how the total causal influence that results in the observed outcome is decomposed into the influences by the candidate and the background causes. The functions characterizing the decomposition are often called integration functions. Causal invariance functions are integration functions that specify the sameness of causal influence across contexts. Different integration functions yield different causal conclusions (e.g., see Lu et al, 2008). Our Study 1 presents a situation where multiple integration functions yield qualitatively different causal recommendations.

One might argue, however, "Why would a particular integration function have a special status? Which integration function is appropriate depends on the domain. Although causal-invariance functions explain the results from many experiments (e.g., see Lu et al., 2008), perhaps due to reasoners' prior knowledge of how some causes combine their influences in certain scenarios, other integration functions may be more appropriate for describing how causal influences combine in other domains." Even if causal-invariance functions are the default integration functions, the argument may go, "whenever these functions do not fit the data from a domain, they would be - and should be - given up in favor of a better-fitting integration function. Causal-invariance functions may be a convenience, but the key factor is how well an integration function explains causation in a domain. Adherence to particular integration functions regardless of domain would be irrational." This argument may appear to have empirical support: Adults and even children have been shown to be able to learn various causal integration functions and generalize their learning to novel variables presented in the experiments (e.g., Lucas et al., 2014; Melchers et al., 2004).

To explain the relation between our work and work on integration-function learning, we make two distinctions: 1) a distinction between analytic and empirical knowledge (cf. Hume's, 1739, "truths of reason" and "matters of fact") and 2) a part-whole distinction, between a "whole" cause (elemental or complex) and an interactive component within a whole cause. Whereas empirical knowledge is contentspecific and justified by experience or data, analytic knowledge is content- and domain-general (i.e., formal) and is justified by reason, by what deductively follows based on the meaning of the concepts in question. Previous work has studied the generalization of acquired empirical (data-based) integration functions. In contrast, our work studies the role of a causal-invariance function as analytic knowledge, operating as a default and a revision criterion in causalknowledge construction, with both roles motivated by the (tacit) goal of formulating useable causal knowledge.

The combination of biological factors that lead to "healthy forest growth" is a whole cause of that outcome; adequate nitrates in the forest soil is an interactive component in that complex whole cause. Arsonists and the
lumber industry are two other whole causes that influence that outcome. Likewise, the gravitational force from a celestial body $y$ on a celestial $x$ is a whole cause of $x$ 's motion; the masses of bodies $x$ and $y$ and the distance between them are interactive components within that whole cause. The gravitational forces from other celestial bodies on $x$ are other contributing whole causes of $x$ 's motion, independently influencing that motion.

Note that within the same domain (e.g., gravitational force), an interaction function (i.e., Newton's law of universal gravitation) integrates the influences from specific component factors (e.g., the masses of the two celestial bodies in a pairwise gravitational force) and a causalinvariance function (vector addition) integrates the influences from multiple whole (presumably noninteracting) causes (e.g., the gravitational forces from multiple bodies on a target body simply sum up). To enable prediction, the aim of causal-knowledge construction is to formulate whole causes (elemental or complex) that are teased apart from, that do not interact with, other causes (e.g., whole causes in the background).

Because causal-invariance and causal-interaction functions exist within the same domain, empirical integration functions are content- or context-specific rather than domain-specific. Whether an acquired interactionintegration function generalizes to other candidate causes depends on the perceived similarity between the relevant causal mechanisms (e.g., Lucas \& Griffiths, 2010, Expt. 5; Wheeler, Miller, and Beckers, 2008, Expt. 3) as well as on situational variables (e.g., Wheeler et al., 2008, Expts $1 \& 2$; the demand characteristics of an experiment). In contrast, causal-invariance functions (e.g., vector addition, the noisy-AND-NOT function in Eq. 2) are formal, specific to variable types (vectors \& binary variables, respectively), but general across domains, contents, and contexts. As explained earlier, for the goal of constructing useable causal knowledge, only causal-invariance functions can serve as a default and a revision criterion for integrating the influence of ideally whole candidate causes with the influence of (potentially unknown) other causes.

## Causal-Invariance Functions for Binary Variables

The causal-invariance functions for two binary causes of a binary effect - a candidate cause of an outcome and the background causes as group - are as follows (e.g., Cheng, 1997; Pearl, 1988). There are different but logically consistent functions for potentially generative and potentially preventive candidate causes.

For a candidate cause $c$ that potentially generates effect $e$ and does so independently of alternative causes in the context, denoted $a$ as a group, the probability of observing $e$ is given by a "noisy-OR" integration function,

$$
\begin{equation*}
P\left(e=1 \mid c ; w_{a}, q_{c}\right)=q_{c} \cdot c+w_{a}-q_{c} \cdot c \cdot w_{a} \tag{1}
\end{equation*}
$$

where $c \in\{0,1\}$ denotes the absence and the presence of candidate cause $c, e \in\{0,1\}$ denotes the absence and the presence of effect $e, q_{c}$ represents the generative power of the candidate cause $c$, and $w_{a}$ represents the probability that
$e$ occurs due to all background causes, known and unknown. For a candidate cause $c$ that potentially prevents effect $e$, the probability of observing $e$ is given by a "noisy-AND-NOT" integration function:

$$
\begin{equation*}
P\left(e=1 \mid c ; w_{a}, p_{c}\right)=w_{a}\left(1-p_{c} \cdot c\right) \tag{2}
\end{equation*}
$$

where $p_{c}$ is the preventive causal power of $c$. These "noisylogical" integration functions (terminology due to Yuille \& $\mathrm{Lu}, 2008$ ), under the assumption that there is no confounding [i.e., when $\mathrm{P}(a=1 \mid c=1)=\mathrm{P}(a=1 \mid c=0)$ ], imply respectively equations for estimating $q_{c}$ and $p_{c}$. The equation for estimating preventive power $p_{c}$, for example, is:

$$
\begin{equation*}
p_{c}=\frac{P(e=1 \mid c=0)-P(e=1 \mid c=1)}{P(e=1 \mid c=0)} \tag{3}
\end{equation*}
$$

Our experiments test preschoolers' use of noisy-logical functions, the probabilistic version of disjunction, in their role as analytic knowledge of causal invariance for binary variables. Testing for knowledge of probabilistic causal invariance rather than deterministic disjunction provides a stronger test of our thesis.

## Preschooler Experiments

Our two studies with preschool children tested our causalinvariance hypothesis against alternative hypotheses, including ones in addition to the linear-integration rule tested in Liljeholm and Cheng (2007). The linear rule states that the observed value of the outcome is explained by the sum of the individual causal influences present. Our studies concern evaluating the effects of two treatments for removing (or preventing) an undesirable outcome, to decide which treatment best removes the outcome. Generalizing across contexts in the scenario involves generalizing from a farm context to a zoo context. Study 1 tested a situation in which the noisy-AND-NOT and linear integration rules yield opposite recommendations for action, and the divergence does not diminish with increased sample size.

Unlike the event frequencies in Liljeholm and Cheng's experiments, the event frequencies in Study 1 (see Table 1) were constructed so that logistic regression and the linear rule recommend the same action (see Cheng et al., 2003 for an explanation of the shared recommendation), contrary to that recommended by the noisy-AND-NOT rule. Logistic regression is a widely used statistical procedure in the medical sciences for evaluating the causal effects of treatments for binary outcome variables. Binary variables are common in medicine (e.g., whether or not a bone is fractured, a tumor is malignant, a woman is pregnant, a patient survives).

In both Studies 1 and 2, the children listened to an interactive story that concerns two brothers - a farmer and a zookeeper - who noticed that some of their animals had red dots on their faces. They were told, "The animals didn't seem sick at all, but the red dots made them look kind of funny." They heard that two "really tasty" and healthy treats, one a grain and the other leaves, might make the red dots go away. The brothers decided to figure out whether the treats work. First, they visited the farm, and fed the
grain treat to every farm animal; later they visited the zoo, and fed both treats to every zoo animal.
Table 1 displays the pattern of event frequencies at the farm and at the zoo for Study 1. The critical "transfer" question is: To relieve red dots on new farm and zoo animals that have red dots on their faces, if one has to choose one and only one treat, what is one's best bet on which treat to use, grain or leaves? Assume that neither treat has any bad effects.

Table 1: Event frequencies for Study 1

| Intervention | Farm <br> Grain <br> only | Zoo <br>  <br> leaves |
| :--- | :--- | :--- |
| Animals with dots: <br> Pre intervention | $9 / 10$ | $4 / 10$ |
| Animals with dots: <br> Post intervention | $6 / 10$ | $1 / 10$ |
| Number Cured <br> Fraction Cured | 3 | 3 |

Regardless of how "sameness of influence" is defined, the rationale underlying the choice is: Assuming the grain operates the same way across contexts (i.e., farm and zoo), then if the influence of the intervention (grain at farm vs. both treats at zoo) is observed to be the same across contexts, one's best guess would be that leaves had no influence - grain alone would already explain the outcome. But, if the influence of the intervention varied across contexts, one would attribute the difference to leaves.

Whereas the causal-invariance function predicts recommending leaves, models adopting a linear integration rule, frequentist or Bayesian, recommend using grain. Here we briefly sketch inferences according to the two rules (for prediction details see Cheng et al., 2013).

First, according to the noisy-AND-NOT integration rule (using Bayesian maximum-likelihood estimates of causal strengths as the predictor), the outcomes at the farm suggest that the grain removes red dots in a farm animal with a $1 / 3$ probability. Assuming the grain's efficacy remains the same for the zoo animals as for the farm animals, grain would be expected to remove red dots with a probability of $1 / 3$ in every zoo animal. It should be clear that the treat ingested by each animal does not "know" what the treat does in other animals (the independent-trials assumption). It follows that only $1 / 3$ of the 4 zoo animals with red dots would be expected to have their red dots removed by the grain. Because in fact 3 of these 4 animals had their red dots removed, considerably more than $1 / 3$ of the 4 , the leaves must be explaining the large difference between the expected and the observed outcome. The causal-invariance function therefore predicts recommending leaves for the new animals. Note the use here of deviation from invariance as a criterion for revising one's causal beliefs, from grain as a preventive cause to leaves as a preventive cause also.

In contrast, according to the linear integration rule, because 3 of 10 animals were "cured" both at the farm and
at the zoo, the addition of the leaves treat at the zoo does not result in any additional cured animals. This rule therefore predicts that the leaves treat is noncausal, and recommends giving the new animals grain.
The transfer question can be equivalently stated in terms of an interaction with something in the context. Both variants of the question address whether one's initial causal belief regarding relieving red dots requires revision.

## Study 1

## Method

Participants The participants were 29 children ( 13 male and 16 female) Children's mean age was 3.42 years (range 2.61 to 4.84 years, $\mathrm{SD}=.60$ years). One additional child was excluded for failure to complete the task. Children were recruited from preschools in Los Angeles, CA. All children were fluent speakers of English and were learning English as a primary language.

Procedure As mentioned, children first listened to the story about the farm and zoo animals with and without red dots on their faces. The farm animals received a grain treat intervention and the zoo animals received a simultaneous grain and leaves treat intervention. In the last part of the study, children were shown new farm and zoo animals and asked to choose between two potential interventions.

Storybook Task The task was presented in a child friendly format, as an interactive storybook. The "reader" of the book was blind to any hypotheses of the study. Children were read the following cover story:
"Once upon a time there were two brothers, one was a farmer and the other a zookeeper. The two brothers loved their animals very much and took very good care of them. One day, the brothers noticed that some of their animals had red dots on their faces."

After being reassured that the animals were not sick, the children were told about the two treats, and were asked to determine their efficacy. They were told that both tasty treats would be loved by the animals.
"The two brothers decided to figure out whether the treats work. First, they went together to the farm. Then, they went over to the zoo. Let's look at what happened and see if YOU can figure out if the grain makes the red dots go away and if the leaves makes the red dots go away."

The farm context and the zoo context were presented separately, and the change in context was highlighted and emphasized. Animals in the farm context received the grain intervention only, whereas those in the zoo context received the grain and the leaves intervention in combination.

Figure 1 depicts examples of the pre- and postintervention pictures that children saw. Because it was critical for children to attend to 1) the presence or absence of the red dots and 2) the administered intervention, those aspects of the story were interactive. For example, children were told "Here is a cow before it ate anything today" and then were asked "Does this cow have any red dots?" Children's responses were acknowledged (e.g., "You're
right he does have red dots"). Children were then handed a cut-out of the treat to feed to the animal. Next, children were asked to make a prediction (e.g., "Do you think the cow will have red dots on its face now that it ate the grain?") After the child replied, the experimenter said, "Let's see!" and showed the picture of the treat inside the animal's tummy, and the presence or absence of red dots was noted regardless of how the child answered (e.g., "Look no more red dots!"). This procedure was repeated with all twenty animals.


Figure 1: Examples of the pre- and post- intervention pictures.

Treat Selection The critical test was presented to children at the conclusion of the story. Children were shown two new animals (one farm and one zoo animal) with red dots on their faces and were asked to select only one of the treats, either the grain or the leaves, to make the animals' red dots go away.

Event Frequency Table 1 depicts the event frequencies for the Study 1. To control for primacy and recency effects, the first trial at the farm and at the zoo showed the same event type; likewise, the last trial at the two locations showed the same event type. (A replication of the study randomized trial order; see note at end of Study 2.) As explained earlier, the noisy-AND-NOT integration rule predicts choosing leaves, but the linear integration rule predicts choosing grain. Note that the linear prediction requires a subset of the arithmetic steps required by the noisy-AND-NOT prediction. The linear rule also predicts the outcome at the zoo perfectly assuming fewer causes than the noisy-AND-NOT rule, namely, a single cause rather than two causes.

## Results

Children were attentive during the storybook reading and rarely responded incorrectly about the presence or absence of red dots. Across all children and all questions there were 7 initial incorrect responses (out of 360 total queries). For these seven responses children were corrected (e.g., "Look here are red dots") and queried again.
The critical result concerned which treat children selected to make the animals' red dots go away. As Figure 2 shows, children overwhelmingly chose the leaves $X^{2}(1)=12.4, p$ $=.0004$, suggesting that children's responses fit with the
noisy-AND-NOT rule rather than with the linear rule. They did so despite the linear rule's relative arithmetic simplicity and its perfect accuracy predicting the outcome at the zoo using fewer causes.


Figure 2: Results from Study 1 depicting the number of children selecting the grain treat versus the leaves treat.

## Study 2

There are alternative explanations for why the children selected leaves in Study 1. The children's attention could be biased toward the newer second treat. The children might simply have a bias toward leaves. Or they might have used a heuristic: pick the treat uniquely associated with the fewest animals with red dots after the intervention. Previous related experiments have not ruled out analogous hypotheses. To rule out all three alternative explanations, Study 2 presented the same story but with the event frequencies in Table 2 to a separate group of preschoolers. As should be clear, the heuristics and biases still predict choosing leaves. For example, as before, fewer animals had red dots after the intervention at the zoo than at the farm (one and two, respectively). The noisy-AND-NOT rule predicts choosing grain this time; the "treatment" maintained the same preventive strength of $3 / 4$ at the farm and at the zoo. Along with the above heuristics and biases, the linear rule predicts no change from the recommended action in Study 1.
Table 2. Event frequencies for Study 2

| Intervention | Farm <br> Grain <br> only | Zoo <br>  <br> leaves |
| :--- | :--- | :--- |
| Animals with dots: <br> Pre intervention | $8 / 8$ | $4 / 8$ |
| Animals with dots: <br> Post intervention | $2 / 8$ | $1 / 8$ |
| Number Cured <br> Fraction Cured | 6 | $3 / 8$ |

## Method

Participants The participants were 28 preschool-aged children ( $M=4.38$ years, range 2.61 years -5.18 years, SD $=.66$ years). 14 were male and 14 were female. An additional two children were excluded for failure to
complete the task and/or attend to the story. Children were recruited similarly using the same criteria as for Study 1.
Procedure The procedure replicated that in Study 1 except that there were 16 trials in total, with the event frequencies for the farm and zoo animals as specified in Table 2.

Results As before, the critical result concerned which treat children selected to make the animals' red dots go away. Figure 3 shows that children's pattern of responses reversed in Study 2: children were now significantly more likely to select the grain treat, $X^{2}(1)=5.14, p=.02$.

We replicated the pattern of results in Studies 1 and 2 in a variant in which the children were randomly assigned to the two studies, and the order of trials in each context (farm and zoo) was randomized for each child.


Figure 3: Results from Study 2 depicting the number of children selecting the grain treat vs. the leaves treat.

## Discussion

Our results favor young children's use of a causalinvariance function over use of the simpler linear function, a preference for one of the candidate causes, or a heuristic to choose the candidate more frequently paired with the desired outcome. Only the noisy-AND-NOT rule representing causal invariance can explain the opposite predominant choices across both our studies. More complex alternative hypotheses, such as use of the linear function in combination with a bias toward the candidate with the more frequent pairing, await further study.

The goal of our present paper is to provide support for the essentiality of the concept of causal invariance, as a default and a criterion for belief revision, in the construction of useable causal knowledge, when the set of possible causal representations is too large to exhaustively evaluate. Our findings indicating the early use of a probabilistic causalinvariance function -- embodying the rather abstract concept of the unchanging nature of the forces of change -- suggest that the generalizability of causal knowledge, along with parsimony and logical consistency, is not a mere wish but a constraint in the rational construction of causal knowledge.

## Acknowledgments

The research reported in this article was supported by AFOSR FA 9550-08-1-0489. We thank Chris Carroll and

Reid Hastie for helpful discussions, and Maryann Francis and Natalie Peri for conducting the experiments.

## References

Cheng, P.W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104,367-405.
Cheng, P.W., Liljeholm, M. \& Sandhofer, C. (2013). Logical consistency and objectivity in causal learning. In Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 2034-2039). Austin, TX: Cognitive Science Society.
Cheng, P.W. \& Lu, H. (in press). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing the invariance of causal power. In M.R. Waldmann (Ed). The Oxford Handbook of Causal Reasoning. Oxford, England: Oxford Univ Press.
Gopnik, A. (2009). The philosophical baby. New York: Farrar, Straus and Giroux.
Gweon, H. and Schulz, L.E. (2011) 16-month-olds rationally infer causes of failed actions. Science, 332, 1524.

Hume D. (1739/1987) A treatise of human nature (2nd edition, Clarendon Press, Oxford).
Legare, C.H. (2012) Exploring exploration: explaining inconsistent information guides hypothesis-testing behavior in young children. Child Development, 83, 173185.

Liljeholm, M., \& Cheng, P.W. (2007). When is a cause the "same"? Coherent generalization across contexts. Psychological Science, 18, 1014-1021.
Lu, H., Yuille, A., Liljeholm, M., Cheng, P.W., \& Holyoak, K.J. (2008). Bayesian generic priors for causal learning. Psychological Review, 115, 955-984.
Lu, H., Rojas, R. R., Beckers, T., \& Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. Cognitive Science, 40, 404-439.
Lucas, C.G., \& Griffiths, T.L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. Cognitive Science, 34(1), 113-147.
Lucas, C.G., Bridgers, S., Griffiths, T.L., Gopnik, A. (2014). When children are better (or at least more openminded) learners than adults: Developmental differences in learning the forms of causal relationships. Cognition, 131, 284-299.
Melchers, K. G., Lachnit, H., \& Shanks, D. R. (2004). Past experience influences the processing of stimulus compounds in human Pavlovian conditioning. Learning and Motivation, 35(3), 167-188.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Morgan Kaufmann.
Rakison, D. H. \& Krogh, L. (2012). Does causal action facilitate causal perception in infants younger than 6 months of age? Developmental Science, 15, 43-54.
Wheeler, D.S., Beckers, T. \& Miller, R.R. (2008). The effect of subadditive pretraining on blocking: Limits on generalization. Learning \& Behavior, 36 (4), 341-351.

# The relationship between fairness, cognitive control, and numerical encoding 

Nadia Chernyak (chernyak@bu.edu) \& Peter R. Blake (pblake@bu.edu)<br>Department of Psychological \& Brain Sciences, Boston University<br>Boston, MA 02215 USA


#### Abstract

Fairness, or the ability to distribute resources in a manner that accords with societally recognized principles of justice, is a hallmark of human cooperation. Young children rapidly develop the ability to enact fairness, but the cognitive underpinnings of this ability remain unknown. The present study investigated 4-7-year-olds' acquisition of three principles of fairness -- equality (the principle that all parties should have the same), merit (the principle that those who work harder should get more), and starting opportunity (the principle that those who started with less should get more) -in relation to their emerging cognitive control and memory for numerical information (numerical accuracy). Cognitive control predicted children's equal sharing, whereas numerical accuracy predicted merit-based sharing. Children up through the oldest age we tested ignored starting opportunities. The results suggest that different principles of fairness may be underpinned by distinct cognitive processes.


Keywords: fairness; cognitive control; resource distribution; children; social and cognitive development

## Introduction

Fairness, or the ability to distribute resources in a manner that accords with societally recognized principles of justice, serves as a foundation for human cooperation and is a critical cognitive achievement of early childhood. In spite of the fact that the concept of fairness itself is ubiquitous, its specific manifestation varies across individuals, cultures, and social groups (Schafer, Haun, \& Tomasello, 2015). Even within a given cultural group, many possible principles of fairness exist. For example, people endorse the idea that principles of equality and merit are both fair. The key empirical question is how people shift between different potential principles of fairness and what accounts for the acquisition of these different forms of fairness. In this work, we explored the cognitive predictors of young children's fairness behavior in a third-party resource allocation task.

Recent work in developmental psychology finds that even infants possess rudimentary concepts of fairness (Sloane, Baillargeon, \& Premack, 2012; Schmidt \& Sommerville, 2012). Throughout preschool and middle childhood, children appreciate at least three distinct principles of fairness: equality (Rakoczy, Kaufmann, \& Lohse, 2016; Smith, Blake, \& Harris, 2013), merit (Baumard, Mascaro, \& Chevallier, 2012; Damon, 1975; Jara-Ettinger, Gibson, Kidd, \& Piantadosi, 2015; Kanngiesser \& Warneken, 2012), and starting opportunity (McCrink, Bloom, \& Santos, 2010; Ng, Heyman, \& Barner, 2011). However, children do not always use these principles consistently. For example, although 3-year-olds pay attention to merit-based information (Baumard et al., 2012) and are able to incorporate it into their resource allocation decisions even
when doing so is costly (Hamann et al., 2011; Kanngiesser \& Warneken, 2012), they often ignore information about merit and enact equal distributions instead (Baumard, Mascaro, \& Chevallier, 2012; Damon, 1975; Rizzo, Elenbaas, Cooley, \& Killen, 2016).

One possibility for these discrepant results may be that different principles of fairness are underpinned by unique cognitive processes. In this work, we investigated two potential cognitive processes underlying children's resource distribution: children's memory for numerical information and children's cognitive control. Accurately encoding quantitative information is essential for both equality-based and merit-based allocation. Numerical skills operate at two levels for resource distribution tasks. First, a general numerical ability is necessary for children to execute even a simple division into two equal subsets that are matched on cardinal equivalence (see Muldoon, Lewis, \& Freeman, 2009; Sarnecka \& Wright, 2013). Indeed, counting abilities have been proposed and also recently found to relate to children's abilities to share resources equally (Chernyak, Sandham, Harris, \& Cordes, 2016; Frydman \& Bryant, 1988; Squire \& Bryant, 2002).

Merit-based distribution also requires attending to quantitative information about relative effort (i.e., that one person worked twice as hard as another) and subsequently incorporating that information into decisions about resource allocation (i.e., that the harder worker must therefore receive twice as much). Similarly, information about starting opportunities must be encoded in order to be used. Thus, sharing between two recipients based on merit or starting opportunity requires trial specific numerical encoding. In this work, we looked at whether children encoded exact numerical information for each trial or whether the information was encoded only approximately.

Finally, we looked at children's emerging cognitive control. Distributing resources according to merit or starting opportunity requires holding in mind multiple -- and often conflicting -- pieces of information and "rules" regarding resource distribution (Zelazo, Helwig, \& Lau, 1996). For example, a child must keep in mind that one person worked harder, but also that that person had a greater starting opportunity to begin with (McCrink et al., 2010). Prior work has found relationships between children's inhibitory control and their abilities to execute the normatively appropriate resource allocation in costly first-party tasks (Blake, Piovesan, Montinari, Warneken, \& Gino, 2015; Steinbeis \& Over, 2017). In third-party tasks of the type that we investigated, cognitive control may serve as a behavioral tool through which they may control their behavioral responses and implement a target distribution.

In this study, we presented 4-7-year-old children with a series of trials in which children were presented with stories about animal characters that had expended either equal or unequal amounts of work in order to acquire resources that would be jointly sold. Each character also had either equal or unequal amounts of starting opportunities. After hearing each story, children were provided with a set of resources to split between these characters. We tested children's memory for numerical information by asking children to recall the amounts of work and starting opportunities for each character after each trial. We also tested children's cognitive control by administering the Happy/Sad Stroop task (Lagattuta, Sayfan, \& Monsour, 2011).

## Method

## Participants

Participants were 67 children ( 35 female, 32 male) between the ages of 4 and 7. This age-range included a younger age group of $334-5$-year-olds (Mean age $=5.00$; Range $=4.11$ 5.82 ) and an older age group of 34 6-7-year-olds (Mean age $=7.10$, Range $=6.00-8.03$ ). Children were tested either in the laboratory or at a local preschool or elementary school.

## Materials

Materials were 12 sets of storybook panels (described below), 8 sets of plastic cookies for the resource allocation tasks, and a set of 24 black and white pictures of smiley faces ( 12 representing sad faces and 12 representing happy faces) for the Happy/Sad Stroop Task.

## Procedure

Children completed 3 pretest trials, followed by 8 focal resource allocation trials, followed by the Happy/Sad Stroop Task. All children were tested in a quiet room in the laboratory or at their local school by an experimenter. All children were videotaped with the exception of 6 children whose parents did not provide video consent and 4 due to technical issues. The experimenter or a trained research assistant also transcribed answers as children during test.

Pretest Trials All children began were introduced to the structure of the storybook task via 3 pretest trials aimed at making children understand the relevant components of each story. In the first panel, the experimenter showed the child two dinosaurs and said that sometimes in the stories some characters will have different amounts. She then indicated that one dinosaur had way more candy than another, and asked the child to recount which dinosaur had more. In the second panel, the experimenter showed the child two dinosaurs and said one worked harder than another and asked the child to recount which one worked harder. Finally, in the last panel, the experimenter showed the child two dinosaurs and said both had the same and worked the same amounts. She then asked the child to recount whether either of the dinosaurs had more and also to
recount if one had worked harder. Incorrect responses were followed with corrective feedback and re-prompts.

Test Trials In each resource allocation trial, children were presented with two animal characters (e.g., two cats) who each acquired resources to achieve a shared goal (e.g., catching fish to sell at the market). The characters contributed either equal or unequal amounts of work towards the shared goal (e.g., one cat caught 4 fish the other caught 2). The characters also had either the same or different starting opportunities (e.g., one cat fished from a pond with 4 fish whereas the other cat fished from a pond with only 2 fish).
An example of the materials and wording of the task is shown in Figure 1 below.

We used a 2 (Starting Opportunity: Equal or Unequal) x 2 (Work Expended: Equal or Unequal) design in which we presented each child with 4 different trial types (2 of each type totaling 8 trials per child): (a) all equal trials (i.e., trials in which characters had exactly the same starting opportunity and expended the same amounts of effort); (b) equal opportunity, unequal work trial, (c) unequal opportunity, equal work trials (e.g., a trial in which two cats both obtain 2 fish, but one started with a pond that only had 2 and another started with a pond that had 4), and (d) unequal opportunity, unequal work trials. In these last trials, characters produced unequal amounts of work, but also had different starting opportunities. For example, one cat caught 2 out of 4 fish and another caught 1 out of 2 fish. The ratio of opportunity to work expended was thus equal.

Figure 1: Example of a Resource Allocation Trial


The types of trials and numbers used in each trial are summarized Table 1 below. As may be noted in this table, the work ratios between the two characters were $1: 1$ if equal and always $2: 1$ if unequal (i.e., the character who worked more obtained twice as much). Additionally, ratios between
a given character's starting opportunity and work expended were also either $1: 1$ or $2: 1$.

Each of these four trial types were presented in 2 blocks: a large number block in which we used relatively large numbers of starting opportunities and work expended (e.g., 8 and 4 fish), and a small number block in which we used relatively small numbers of starting opportunities and work expended (e.g., 4 and 2 fish). Within each block, the presentation of the four different trial types were counterbalanced with a Williams Latin Square design. We also counterbalanced which block type (large vs. small) was presented first as well as whether the larger vs. smaller numbers appeared on the child's right or left side. Each trial used one of four possible animal pairs: cats that fished fish, rabbits that grew carrots, bears that picked apples, and monkeys that picked bananas. Presentation of animal types and colors of animal characters were fixed.

Resource Allocation After being read each scenario (trial type), children were shown 6 plastic cookies that the characters earned from their joint effort. Children were told that they had to decide which characters should get which cookies. Cookies were arranged in a linear array in between two cardboard boxes that depicted the two animal characters. We note that we used 6 cookies specifically because they enabled either distribution according to equality (i.e., 3 cookies to each character) or distribution according to a $2: 1$ merit ratio (i.e., 4 cookies to the harder worker and 2 to the less hard worker). We recorded the amount children gave to each character.

Table 1: Numbers used in each trial type

| Trial Type | Starting <br> Opportunity | Work <br> Expended | Block <br> Type |
| :--- | :--- | :--- | :--- |
| All equal | $(8,8)$ | $(4,4)$ | large |
| Equal opportunity, <br> unequal work | $(8,8)$ | $(8,4)$ | large |
| Unequal opportunity, <br> equal work | $(8,4)$ | $(4,4)$ | large |
| Unequal opportunity, <br> unequal work | $(8,4)$ | $(4,2)$ | large |
| All equal |  |  |  |
| Equal opportunity, <br> unequal work | $(4,4)$ | $(4,4)$ | small |
| Unequal opportunity, <br> equal work | $(4,2)$ | small |  |
| Unequal opportunity, <br> unequal work | $(4,2)$ | $(2,2)$ | small |
| Numerical Accuracy <br> each trial, boxes and materials were closed and put away, |  |  |  |

and children were asked four questions to assess their memories for the numerical information presented to them in the trial: two questions asking them to recall the characters' starting opportunities (e.g. "How many fish did green Kitty have in green Kitty's pond? How about red Kitty?") and two questions asking them to recall the characters' work expended (e.g., "And how many fish did green Kitty get out of green Kitty's pond? How about red Kitty?"). For each of these, we computed a continuous accuracy measure reflecting the Percent Absolute Error that children's answers displayed (Siegler \& Booth, 2004). Percent Absolute Error (PAE) was calculated via the following formula:

$$
\text { PAE }=\frac{\mid \text { child's answer }- \text { correct answer } \mid}{\text { correct answer }}
$$

This score reflected the deviation of the child's answer from the correct answer. For example, a child who answered that there were 5 fish when the correct answer was 4 would receive a PAE score of $|5-4| / 4=25 \%$. Thus, high PAE scores indicated lower accuracy, and low PAE scores reflected higher accuracy (with a score of 0 indicating having correctly recalled the exact number). On each trial, each child was given two scores assessing trial specific numerical encoding: a Starting Opportunity PAE reflecting the average PAE of the two opportunity questions, as well as a Work Expended PAE reflecting the average PAE of the two work expended questions on that trial. These are labeled as trial specific because they reflected children's accuracy for numbers on that specific trial only.

Additionally, children received an Overall PAE reflecting numerical accuracy average across all 32 questions asked (4 per trial), which reflected general numerical memory. This was referred to as general numerical memory because it reflected children's tendency to correctly estimate numbers overall, not on any given trial.

Cognitive Control After completion of all 8 resource distribution trials, children were administered a version of the Happy/Sad Stroop Task. Following procedures used in (Lagattuta et al., 2011), children were introduced to a happy and a sad face and asked to label them. They were then told they would be playing an "opposite" game in which they had to label happy faces as "sad" and sad faces as "happy". After ensuring all children understood task instructions, children completed 4 practice trials (corrective feedback was provided) followed by 20 test trials (no corrective feedback). Children were given a Cognitive Control Score between 0-20 reflecting the number of correct test trials.

## Results

Preliminary results showed no effects of gender or block type, so we collapsed data across these variables. We first sought to characterize children's resource allocation decisions by coding the outcomes as equal, merit, or other (unequal for equal trials or against merit; see Figure 2).

We ran separate models predicting children's tendency to share resources equally. We ran a within-subjects mixed linear model using equal sharing allocation as the dependent variable and Age (entered continuously), Work Expended (equal or unequal), Initial Opportunity (equal or unequal) and all interaction effects as the predictors. There was a significant main effect of Work Expended, $F(1,528)=$ 9.102, $p=.003$ a significant Age x Work Expended interaction, $F(1,528)=18.07, p<.001$, and no other significant effects ( $p$ 's $>.15$ ). Thus, equal allocations were predicted by age and whether the characters had expended equal amounts of work.

Figure 2: Allocation Types Across Trials


Note. Merit-based sharing is impossible in Equal Work Expended Trials

We also looked at predictors of merit-based sharing. Children were coded as having given a merit-based allocation if they had given more resources to the harder working character (i.e., one who produced a greater amount of resources). Because merit-based sharing was not possible in the All Equal trial, we excluded this trial from that analysis. There was a significant Age x Work Expended interaction, $F(1,396)=4.81, p=0.03$, and no other significant effects (all $p$ 's > .09). Thus, children made more merit-based allocations with age.

To better characterize these interactions, we explored how age impacted merit-based and equality-based sharing separately in each trial type. We thus ran follow-up models for equal work expended and unequal work expended trials using Age as a predictor. The results of these analyses are summarized in Table 2 (Model 1). We ran separate models using equal sharing as a response and then merit-based sharing as a response.

As shown in Table 2, Age predicted equal sharing in the trials in which characters produced equal amounts of work (top panel of Table 2; All Equal and Unequal Opportunity, Equal Work trials), whereas Age predicted merit-based sharing in trials in which characters produced unequal amounts of work (bottom panel of Table 2).

Therefore, confirming the previous analyses, all children ignored starting opportunities. However, age predicted children's likelihood of selecting the "correct" allocation type in each trial - equal-based sharing in the trials in which
characters produced equal amounts of work, and meritbased sharing in the trials in which characters produced unequal amounts of work.

We next investigated whether these age-related changes were explained by numerical accuracy or cognitive control. In particular, we first looked at whether children's numerical accuracy predicted their resource allocation decisions. Recall that on each trial, each child was given two scores: a Starting Opportunity PAE and a Work Expended PAE. Preliminary analyses revealed no differences between the two PAE types, suggesting that children were equally adept at encoding both types of information.

We first looked at the predictors of each PAE type. For each model, we ran a mixed linear model using Age, Work Expended, and Initial Opportunity as predictors. We also included Cognitive Control Total Correct as a covariate to ensure that any potential age-related changes in encoding accuracy were not simply attributable to changes in cognitive control.

For Starting Opportunity PAE, there was a significant effect of Age, $F(1,464)=93.71, p<.001$, with older children showing lower PAE (higher accuracy) for initial opportunity and no other significant effects (all $p$ 's $>.25$ ).

For Work Expended PAE, there was a significant effect of Age, $F(1,464)=43.89, p<.001$, and a significant effect of Work Expended Trial Type, $F(1,464)=20.05, p<.001$, with children showing worse encoding of work expended when the characters put in unequal amounts of work. Therefore, both age and trial type also predicted children's recall of the work expended. Children were better at encoding numerical information when characters had expended equal amounts of work.

Finally, we looked at predictors of Overall PAE. Age significantly predicted Overall PAE, $F(1,55)=24.32, p<$ .001, and Cognitive Control did not (once accounting for age; $p=0.42$ ). Therefore, numerical accuracy and cognitive control were dissociable, despite both getting better with age. ${ }^{1}$

We next investigated whether these age-related differences in encoding accuracy might explain age-related changes in children's resource allocation decisions. Because children ignored starting opportunities, we do not further consider the Starting Opportunity PAE. Recall that each child could be coded as giving either an "equal split", "merit-based split", or neither split ("other"). For the separate analyses we coded these as equal split or not and merit-based split or not.

We first looked at predictors of equal sharing. Preliminary analyses revealed significant interactions between various predictors for the trial types. Therefore, we first considered the equal work trials only. We ran two mixed binary logistic regression models using equal sharing as a response variable and Age, Work Expended PAE, Overall PAE and Cognitive

[^302]Control as the predictors. ${ }^{2}$ The critical question was whether either Work Expended PAE or Cognitive Control might predict children's equal resource allocation.

As shown in Table 2, in the Equal Work Expended trials, Cognitive Control and Overall PAE (general numerical memory - the average error across the 32 questions asked) predicted children's equal sharing behavior. Once Cognitive Control and Overall PAE were accounted for, there was no longer any significant effect of age. Therefore, both cognitive control and general numerical memory explained age-related changes in sharing resources equally.

We then looked at predictors of merit-based allocation. As shown in Table 2, Work Expended PAE was related to the propensity to split resources meritoriously, but cognitive control and Overall PAE were not. Age continued to be related to the propensity to make merit-based splits.

Table 2: Beta Coefficients (and Standard Errors)

| Response: Equal Sharing | Equal Work Expended | Unequal Work Expended |
| :---: | :---: | :---: |
| Model 1 |  |  |
| Age | . 40 (.13)** | -. 34 (.11)** |
| Model 2 |  |  |
| Age | -. 22 (.18) | -. 50 (.14)** |
| Cognitive Control | 0.15 (.06)* | . 11 (.06)* |
| Work Expended PAE | -. 46 (.35) | -. 16 (.21) |
| Model 3 |  |  |
| Age | -. 24 (.18) | -. 50 (0.14)** |
| Cognitive Control | . 14 (.06)* | . 13 (.05)* |
| Overall PAE | -. 28 (.13)* | -. 08 (.12) |
| Response: Merit Sharing | Equal Work Expended | Unequal Work Expended |
| Model 1 |  |  |
| Age | - | -43 (.11)*** |
| Model 2 |  |  |
| Age | - | . 35 (.15)* |
| Cognitive Control | - | . 006 (.06) |
| Work Expended PAE | - | -. 89 (.40)* |
| Model 3 |  |  |
| Age | - | . 40 (.15)** |
| Cognitive Control | - | . 006 (.06) |
| Overall PAE | - | . 11 (.16) |

The Table 2 results suggest two things: first, cognitive control and general numerical accuracy predicted propensity to make equal splits. Second, trial specific numerical accuracy for work expended (ability to properly encode merit-based information specifically) predicted children's abilities to make merit-based splits on trials that called for merit-based splits (i.e., trials in which characters produced unequal splits). Both sets of results held when controlling for age, suggesting that age-related changes in

[^303]equal sharing may be explained by changes in cognitive control and numerical accuracy, and that age-related changes in merit-based sharing may be partly explained by changes in encoding of merit-based information.

## Discussion

Recent work has taken an interest in the cognitive predictors of fairness. Our findings are consistent with prior work showing a mostly equality-based principle during the preschool age shifting to a merit-based principle by middle childhood. We extend these findings by showing that 6 and 7 -year olds actively create merit-based distributions even when making equal allocations is a viable alternative. Most importantly, we point to two cognitive predictors of sharing behavior: children's numerical encoding ability, and their emerging cognitive control, each of which exerted a unique effect on children's abilities to make resource allocations.

One possibility for why young children often do not employ merit-based resource allocations may be that they fail to encode trial-specific numerical information to begin with. Our findings show that this may be the case: encoding accuracy for the amount of work that each character expended predicted merit-based resource distribution on trials that called for such distribution. Interestingly, numerical encoding accuracy for starting opportunity information did not predict allocations based on this information. Thus, 6- and 7-year olds specifically encoded information about starting opportunity but did not use this information when allocating resources.

The effect of numerical accuracy held even when controlling for changes in age and cognitive control. Prior work has found that, in third-party unequal work tasks, younger children can and do distribute resources according to merit but under simplified conditions such as when an unequal allocation is the only option (Baumard et al., 2012). Three- and 5-year olds are also capable of using merit in first person distributions but the strategies used vary widely (Kanngiesser \& Warneken, 2012).We propose that these individual differences may be explained by differences in the ability to accurately encode numerical information inherent in meritocratic situations.

In contrast to the cognitive processes for merit-based allocations, equal allocation decisions depended on general numerical encoding. This result falls in line with prior work finding that general numerical cognition (i.e., counting ability) predicts equal sharing among preschool-aged children (Chernyak et al., 2016). General numerical accuracy and counting ability may tap into the same underlying construct of numerical fluency and understanding of numbers, which may then help children with creating equal sets of resources.

We also found that children's emerging cognitive control predicted equal, but not merit-based, allocations across all trial types. This suggests that cognitive control serves as a general behavioral tool that allows children to choose equal outcomes in spite of inequalities present in the scenarios. One possibility for why this might be the case is that older
children have acquired and might therefore need to inhibit other potential principles (e.g., merit) in order to enact the equality. Alternatively, cognitive control may simply help children ensure that two equal sets have been created. Most importantly, cognitive control failed to predict merit-based resource distribution, suggesting a dissociation between the types of cognitive mechanisms required for equality and merit-based resource allocation.

Few children used the information about starting opportunities, despite encoding this information. Although work has found that children are able to make evaluations of others' work based on their starting opportunity (McCrink et al., 2010; Ng et al., 2011), to our knowledge, there is no current work that has shown that children then use those evaluations to make resource allocation decisions. We therefore propose that children may be well aware of existing inequalities, but do not actively use such information when making resource allocation decisions.

Overall, our work points to two important cognitive predictors for different fairness principles. We propose that searching for individual differences in children's cognitive abilities may help account for and ultimately shape their social preferences.

## Acknowledgments

We would like to acknowledge Elizabeth Asal, Claire Park, Javier Rivera, Eliana Roth, and Shaina Yoo for assistance with data collection and coding. This work was funded by a postdoctoral fellowship from the Greater Good Science Center and the John Templeton Foundation to NC.

## References

Baumard, N., Mascaro, O., \& Chevallier, C. (2012). Preschoolers are able to take merit into account when distributing goods. Developmental Psychology, 48, 492-498.
Blake, P. R., Piovesan, M., Montinari, N., Warneken, F., \& Gino, F. (2015). Prosocial norms in the classroom: The role of self-regulation in following norms of giving. Journal of Economic Behavior \& Organization, 115, 18-29.
Chernyak, N., Sandham, B., Harris, P. L., \& Cordes, S. (2016). Numerical cognition explains age-related changes in third-party fairness. Developmental Psychology, 52, 1555-1562.
Damon, W. (1975). Early conceptions of positive justice as related to the development of logical operations. Child Development, 46, 301-312.
Frydman, O., \& Bryant, P. (1988). Sharing and the understanding of number equivalence by young children. Cognitive Development, 3, 323-339.
Hamann, K., Warneken, F., Greenberg, J. R., \& Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. Nature, 476, 328331.

Jara-Ettinger, J., Gibson, E., Kidd, C., \& Pinatadosi, S. (2015). Native Amazonian children forego
egalitarianism in merit-based tasks when they learn to count. Developmental Science, 19, 1104-1110.
Kanngiesser, P., \& Warneken, F. (2012). Young children consider merit when sharing resources with others. PloS One, 7(8), e43979.
Lagattuta, K. H., Sayfan, L., \& Monsour, M. (2011). A new measures for assessing executive function across a wide age range: Children and adults find happy-sad more difficult than day-night. Developmental Science, 14, 481-489.
McCrink, K., Bloom, P., \& Santos, L. R. (2010). Children's and adults' judgments of equitable resource distributions. Developmental Science, 13, 37-45.
Muldoon, K., Lewis, C., \& Freeman, N. (2009). Why setcomparison is vital in early number learning. Trends in Cognitive Sciences, 13, 203-208.
Ng, R., Heyman, G. D., \& Barner, D. (2011). Collaboration promotes proportional reasoning about resource distribution in young children. Developmental Psychology, 47, 1230.
Rakoczy, H., Kaufmann, M., \& Lohse, K. (2016). Young children understand the normative force of standards of equal resource distribution. Journal of Experimental Child Psychology, 150, 396-403.
Rizzo, M. T., Elenbaas, L., Cooley, S., \& Killen, M. (2016). Children's recognition of fairness and others' welfare in a resource allocation task: Age related changes. Developmental Psychology, 52, 1307-1317.
Sarnecka, B. W., \& Wright, C. E. (2013). The idea of an exact number: Children's understanding of cardinality and equinumerosity. Cognitive Science, 37, 14931506.

Schafer, M., Haun, D. B., \& Tomasello, M. (2015). Fair is not fair everywhere. Psychological Science, 26, 12521260.

Siegler, R. S., \& Booth, J. L. (2004). Development of numerical estimation in young children. Child Development, 75, 428-444.
Sloane, S., Baillargeon, R., \& Premack, D. (2012). Do infants have a sense of fairness? Psychological Science, 23, 196-204.
Smith, C. E., Blake, P. R., \& Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. PLoS One, 8, e59510.
Schmidt, M. F., \& Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15 -month-old human infants. PloS ONE, 6, e23223.
Squire, S., \& Bryant, P. (2002). From sharing to dividing: Young children's understanding of division. Developmental Science, 5, 452-466.
Steinbeis, N., \& Over, H. (2017). Enhancing behavioral control increases sharing in children. Journal of Experimental Child Psychology, 159, 310-318.
Zelazo, P. D., Helwig, C. C., \& Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. Child Development, 67, 2478-2492.

# Risky Decision Making for Medications: Age and Social Influence Effects 

Amy W. Chong (amywlchong@gatech.edu)<br>School of Psychology, Georgia Institute of Technology 654 Cherry Street NW, Atlanta, GA 30332-0170 USA

Michael T. Bixter (mbixter3@gatech.edu)
School of Psychology, Georgia Institute of Technology
654 Cherry Street NW, Atlanta, GA 30332-0170 USA
Wendy A. Rogers (wendyr@illinois.edu)
Department of Kinesiology and Community Health, University of Illinois Urbana-Champaign 1206 South Fourth Street, Champaign, IL 61820 USA


#### Abstract

Prior studies on older adults' risk taking have paid little attention to the healthcare domain or social influences on decision making. This study examined age-related differences in medication risk taking and the effects of a collaborative decision-making experience on individuals' tendency to take risks. We recruited 24 younger (mean age $=19.50, S D=1.41$ ) and 24 older adults (mean age $=70.54, S D=2.30$ ), and asked them to choose between hypothetical medications that differed in probabilities and outcomes of treatment success. To investigate the effects of risk-neutral versus riskadvantageous trials, participants chose between a risky option and a sure option that had equal expected values (risk-neutral) or between a risky option and a sure option that had a lower expected value (risk-advantageous). Participants completed the decision task first individually (the pre-collaboration phase), then in dyads (the collaboration phase), and once again individually (the post-collaboration phase). During the pre-collaboration phase older adults showed a smaller increase in risk-taking tendency in response to riskadvantageous trials compared to younger adults. The pre-and post-collaboration data showed that older adults' risk preferences converged towards their partner's preference to a greater extent following collaboration relative to younger adults. These findings highlight the importance of designing decision aids to encourage older adults to take risks when risk taking is beneficial, and considering how social processes influence patients' medication decisions.


Keywords: risky decision making, health, aging, social influence

When choosing between options in health care, the degree of risk involved is an important consideration that younger and older individuals must make. For example, an individual may have to choose between painkillers that have different probabilities of treatment success. A National Health Interview Survey in 2012 showed that $86 \%$ of US older adults aged 65 or older have at least one chronic condition, and $61 \%$ have at least two chronic conditions, compared to $27 \%$ and $7 \%$ of US adults aged 18 to 44 (Ward, Schiller, \& Goodman, 2014). As older adults are more likely to have multiple chronic conditions, they may need to make more medical choices involving risks.

Individuals often discuss their health care decisions with family members, friends, or physicians. Despite the plentiful literature on shared decision making between patients and physicians, the emphasis is rarely on how a collaborative experience would affect subsequent health-related behavior of individuals. The lack of research on this area demands attention because it is common for individuals to make a number of choices on their own after their discussion with other people. There is evidence that family and friends influence the health-related attitudes and beliefs of patients, particularly those who are less educated and non-white (Thompson, 2013). Therefore, research should consider how discussions about health care decisions take place within patients' social networks and how to improve the resulting decisional outcomes. One of the most common health care decisions facing patients is medication risk taking.

## Age-related Differences in Risk Taking

## Risk-Neutral Decisions

Most studies on aging and risk taking asked participants to make risk-neutral choices, which involve a risky option and a sure option that have equal expected values. The expected value of an option is calculated by multiplying outcomes by their respective probabilities, and taking the sum of the products (Bernoulli, 1954). A higher expected value represents a higher average value in the long run assuming the same option is chosen repeatedly.

A recent meta-analysis of these studies found that older adults were more risk averse than younger adults in making positively framed decisions (Best \& Charness, 2015). Positively framed decisions refer to choices in which positive aspects of the scenarios are highlighted using wordings such as "keep" and "save". This finding can be explained by fuzzy-trace theory.

Fuzzy-trace theory postulates that people simultaneously store and access two types of representations (Reyna \& Brainerd, 2011). A verbatim representation reflects the precise information. In contrast, a gist representation captures the subjective interpretation of information based on emotion, experience, level of development, and is vague
and qualitative. In the context of the Asian disease problem (Tversky \& Kahneman, 1981), a gist representation of a sure option of "saving 200 people" would be "saving some people" whereas a gist representation of a risky option of "a one-third probability of saving 600 people and a two-thirds probability of saving no people" would be "some probability of saving some people and some probability of saving no people." Hence, fuzzy-trace theory suggests that people would choose the sure option when they represent the positively framed situation at the gist level. Older adults are more likely than younger adults to rely on gist processing because they may have learned that it is a more effective means of making decisions (Peters, Hess, Västfjäll, \& Auman, 2007). In addition, gist processing is relatively well preserved with normal aging although verbatim processing declines as people age (Reyna \& Brainerd, 2011). Older adults' decisions are more gist-based, which may account for their tendency to be more risk averse in the positive frame.

The meta-analysis revealed that the presence of the age effect depended on the amount and the scenario type (Best \& Charness, 2015). That is, the age effect was found in small-amount financial and large-amount mortality scenarios, but not in large-amount financial and smallamount mortality scenarios. Younger and older adults' levels of risk taking depended on the scenario. Owing to the primary use of either financial risk seeking scenarios or the Asian disease problem in the aging literature, past findings on age-related differences in risk taking may not generalize to medication decision making.

## Risk-Advantageous Decisions

Studies have also explored younger and older adults' risk taking tendencies in situations where risk seeking is advantageous and disadvantageous. From an economic perspective, an option with a higher expected value is better than an option with a lower expected value. Analyzing trials on which the expected value of the risky option was more favorable than that of the sure option, older adults were shown to be more risk averse than people of age 5 to 64 (Weller, Levin, \& Denburg, 2011). That is, older adults were less risk taking than younger adults when risk taking was beneficial. However, that study used very broad age ranges.

Based on Peters et al. (2007), and Reyna and Brainerd (2011), older adults have an increased tendency to use gist processing relative to younger adults. Thus, they may be less sensitive to the expected values of the sure and risky options and more likely to stick to their preferred options on risk-neutral trials than younger adults.

## Collaborative Decision Making

If we consider how common it is for people to exchange views with others in everyday situations of making health care decisions, it is necessary to understand medical decision making in a collaborative context. Collaboration in patient-physician relationships is not emphasized in the
traditional care model, which depicts patients as passive followers of the orders set by physicians. However, a new collaborative care model is replacing the traditional model (Mitzner, McBride, Barg-Walkow, \& Rogers, 2013). In the collaborative model, patients and physicians share the primary caregiving responsibility and make decisions together. Hence, investigating collaborative decision making and how it influences decision makers' subsequent decisions would help people make better use of others' opinions.

Collaborative decision making has been studied in social psychology. Group decision-making phenomena that have been observed include group polarization and group convergence. The former occurs when the decisions made by groups are more extreme than the initial position of its members (Sunstein, 2002). Group convergence was found in Bixter, Trimber, and Luhmann's (2017) study that focused on intertemporal monetary preferences. Individuals’ post-collaboration decisions converged towards their respective group decisions. The social comparison process was proposed to explain the findings. Participants might have changed their preferences in accordance with their group members' preferences because they viewed others' behavior as a source of information about normatively appropriate behavior. Using a risky decision task, another study demonstrated a similar behavioral change (Suzuki, Jensen, Bossaerts, O’Doherty, 2016). Participants' risk preferences shifted towards the observed person's preferences. Research is needed to better understand whether a group polarization or group convergence effect would be present in medication risky decision-making scenarios.

## Age Differences in Susceptibility to Social Effects

Given evidence suggesting age-related differences in decision making between younger and older adults, it is reasonable to ask whether younger and older adults' experience of making decisions in a group would influence their individual decisions differently. Age-related differences in the tendency to be influenced by others have been investigated for young age groups. In Gardner and Steinberg's (2005) study, participants made riskier decisions and exhibited more risky behavior when in peer groups, and the influence of peers on risky decision making and risk taking was stronger among adolescents and youths than adults. However, no research has assessed age-related differences between younger adults and older adults.

One finding which suggested that older adults might be more prone to social influence than younger adults in making decisions is the age-related difference in perceived decision-making competence. Older adults rated themselves as less competent decision makers than did younger adults (Bruine de Bruin, Parker, \& Fischhoff, 2012). Despite older adults' accumulation of experience, they may have rated their decision-making competence based on perceived declines in their fluid cognitive abilities. Owing to their lower perceived competence, older adults might change their decisions more easily when different views are
presented. This prediction is supported by the finding that participants who lacked confidence in their answers to health knowledge questions were significantly more likely than those who were confident to change their answer after receiving online social feedback (Lau \& Coiera, 2008). Furthermore, previous research has demonstrated that higher uncertainty strengthened social effects on memory reports (Walther et al., 2002). If older adults are less confident and thus more uncertain about their decisions, they might be more susceptible to social influence.

## Overview of Study

Although older adults often have multiple medical conditions and need to make health care choices involving risks, past research has not assessed age-related differences in risk taking for medication decision tasks. The goal of the current study was to study age differences in medication risk taking when risk taking was advantageous or neutral. Younger and older adults were asked to make choices between medications that involved varying probabilities and outcomes of treatment success. On risk-neutral trials, they chose between options that were equally favorable. On riskadvantageous trials, they chose between options that favored risk taking. To investigate the effect of collaboration on subsequent individual decisions, they were asked to complete the decision task first independently, then in dyads, and finally independently. Hypotheses were:

H 1 : older adults are less risk taking than younger adults.
H 2 : people are more risk taking on risk-advantageous trials than on risk-neutral trials.

H3: there is an age by trial type interaction such that older adults show a smaller increase in risk taking when risk taking is beneficial.

H 4 : older adults, compared to younger adults, are more likely to be influenced by others.

## Method

## Participants

Participants were 24 English speaking younger adults (14 females) between the ages of 18 and $23(M=19.50, S D=$ 1.41) and 24 English speaking older adults ( 14 females) between the ages of 67 and $74(M=70.54, S D=2.30)$. Participants in each age group formed 12 age-group matched dyads. All participants had at least $20 / 50$ visual acuity for near vision (corrected or uncorrected) to ensure that they could see the stimuli. The majority of older adults were highly educated, with $83 \%$ reporting having some college or higher. Other descriptive variables were demographics and health, numeracy (Lipkus, Samsa, \& Rimer, 2001), personality (Gosling, Rentfrow, \& Swann, 2003), social intelligence (Silvera, Martinussen, \& Dahl, 2001), perceived decision-making competence (Greene, Hibbard \& Tusler, 2005), processing speed (Wechsler, 1997), verbal working memory span (Wechsler, 1997), and verbal ability (Shipley, 1986). Due to limited space, results
involving some of these variables are not included in this paper. Table 1 provides descriptive data.

Table 1: Younger and older adults's scores on health and cognitive measures.

|  | Younger <br> Adults | Older Adults |  | $\mathrm{t}-$ <br> value |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $M$ | $S D$ | $M$ | $S D$ |  |  |
| Health $^{\mathrm{a}}$ | 4.05 | .54 | 3.79 | .60 | 1.59 |
| Numeracy $^{\mathrm{b}}$ | 10.55 | .67 | 7.04 | 3.29 | $5.10^{* * *}$ |
| Processing <br> speed $^{\mathrm{c}}$ | 72.45 | 9.63 | 45.83 | 13.84 | $7.51^{* * *}$ |
| Verbal <br> working <br> memory | 8.86 | 2.30 | 7.63 | 2.50 | 1.75 |
| Verbal <br> ability | 32.08 | 3.28 | 32.96 | 4.85 | -.726 |

${ }^{\text {a }}$ Self-reported health (1=poor, 5=excellent); ${ }^{\text {b }}$ Numeracy (number of correct items from 0 to 11 on the numeracy scale); ${ }^{\text {c Processing speed (number of correct items on the }}$ digit-symbol substitution task from 0 to 100); ${ }^{\text {d }}$ Verbal working memory (number of correct items from 0 to 14 on the digits backward task); eVerbal ability (number of correct items from 0 to 40 on the Shipley institute of living scale); $* * * p<.001$.

## Materials

The experiment had three phases: pre-collaboration, collaboration, and post-collaboration. Participants made decisions independently in the pre- and post- collaboration phases, but in dyads in the collaboration phase.

Participants were asked to choose a medication for a family member who is the same age as them. We asked them to give advice to a family member rather than choose one for themselves because this was more ecologically valid with respect to the collaboration phase in which they have to interact with each other and reach a consensus. Every trial of the decision task consisted of a choice between two medications which had different probabilities and outcomes of treatment success. The sure option had $100 \%$ chance of some treatment success whereas the risky option had a variable outcome of treatment success.

There were two trial types, 20 risk-neutral trials and 20 risk-advantageous trials in each phase. On risk-neutral trials, the medications had equivalent expected values. On riskadvantageous trials, the medication with a sure outcome had a lower expected value than the medication with a variable outcome. Figures 1 and 2 show an example of each trial type. For both trial types, the risk magnitudes were $20 \%$, $40 \%, 60 \%$, and $80 \%$ on different trials and the number of days of sickness were $20,30,40,50$, and 60 on different trials. Within each phase of the experiment, the decision trials were presented in a randomized order to minimize
order effects. The percentage of time that participants chose the riskier option indicated their level of risk taking.

| The family member is going to experience 30 days of sickness because of a disease. Which medication would you recommend? |  |
| :---: | :---: |
| Medication A | Medication B |
| 20\% chance of protecting him/her | 100\% chance of protecting him/her |
| from 30 days of sickness and | from 6 days of sickness |
| $80 \%$ chance of protecting him/her from 0 days of sickness |  |

Figure 1: An example of a risk-neutral trial.
The family member is going to experience 60 days of sickness because of a disease. Which medication would you recommend?

| Medication A | Medication B |
| :--- | :--- |
| $100 \%$ chance of protecting him/her |  |
| from 12 days of sickness | $40 \%$ chance of protecting him/her |
|  | from 60 days of sickness |
|  | and |
|  | $60 \%$ chance of protecting him/her <br> from 0 days of sickness |

Figure 2: An example of a risk-advantageous trial.

## Design

Age was a grouping variable. Trial type and decisionmaking phase were the independent variables. Level of risk taking and difference in risk taking between dyad members were the dependent variables.

## Procedure

Before the experiment, participants received a consent form explaining the research study. After consent, they completed a questionnaire regarding demographics and health, and four ability tests. After that, participants were given both oral and written instructions about the decision task. Collaboration with other participants was not mentioned at this stage. The first phase of the experiment was the pre-collaboration phase. Participants made medication decisions involving risks individually. When the pre-collaboration phase was completed, participants moved on to the collaboration phase. They were notified that each of them would have to collaborate with another participant to give one answer as a group. The process of collaborative decision making was videotaped (with permission from the participants) for analysis in a separate study. When the collaboration phase was completed, participants entered the post-collaboration phase. Once again, they made similar decisions individually. After all decision trials were completed, participants filled out the self-report items, followed by other questionnaires, and then they were debriefed.

It took younger adults approximately one hour and older adults approximately two hours to complete the entire experiment.

## Results

## Individual Medication Risk Taking

First, individual risk taking data in the pre-collaboration phase were analyzed. Mixed-design ANOVA was conducted with age as the between-participants variable, and trial type as the within-participants variable.

As expected, older and young adults were significantly more risk taking on risk-advantageous trials $(M=.77, S D=$ .26) than on risk-neutral trials $(M=.41, S D=.32), F(1,46)$ $=82.66, p<.001, \eta_{\mathrm{p}}{ }^{2}=.64$. Overall, older adults $(M=.58$, $S D=.30$ ) were not significantly less risk taking than younger adults $(M=.60, S D=.18), F(1,46)=.024, p=$ $0.877, \eta_{\mathrm{p}}{ }^{2}=.001$. However, there was an age by trial type interaction such that older adults showed a smaller increase in risk taking in response to risk-advantageous trials than did younger adults, $F(1,46)=8.52, p<.01, \eta_{\mathrm{p}}{ }^{2}=.16$. Figure 3 shows the results.


Figure 3: Younger and older adults' level of risk taking on risk-neutral and risk-advantageous trials (error bars represent the standard error).

## Social Influence Effects

Risk taking data in the pre- and post-collaboration phases were compared. Mixed-design ANOVA was conducted with age as between- and phase as within-participants variable. The absolute difference between dyad members' level of risk taking was the dependent variable.

The difference in risk taking between dyad members was smaller in the post-collaboration phase ( $M=.20, S D=.19$ ) than in the pre-collaboration phase ( $M=.33, S D=.22$ ), $F(1$, 22) $=7.80, p<.05, \eta_{p}^{2}=.26$. The overall within-dyad difference in risk taking was not significantly different between younger ( $M=.22, S D=.12$ ) and older adults ( $M=$ $.31, S D=.16), F(1,22)=2.54, p=0.125, \eta_{\mathrm{p}}{ }^{2}=.10$. By contrast, the age by phase interaction was significant, $F(1$, 22) $=10.41, p<.01, \eta_{\mathrm{p}}{ }^{2}=.32$. Older adults' risk preferences converged towards their partner's preferences to a greater extent following collaboration relative to younger adults'. Figure 4 shows the results.


Figure 4: Younger and older adults' within-dyad difference in risk taking in pre- and post-collaboration phases (error bars represent the standard error).

## Discussion

Findings from the present study provide insights into younger and older adults' individual risky decision making for medications, and the effects of collaborating with a partner on subsequent risk-taking tendency. Confirming our expectation, younger and older adults took more risks when risk taking was beneficial than when risk taking and risk aversion were equally favorable. However, this effect of trial type differed between the two age groups such that the increase in risk taking among older adults was smaller than the increase in risk taking among younger adults when the risky option was favored. Regarding social influence effects, dyad members' risk-taking tendency was more similar to each other's after the collaborative decision-making experience compared to their initial difference. Older adults demonstrated a convergence effect, whereas the younger adults did not.

## The ore tical Implications

The present study adds to the literature in that it investigated age differences in risk taking in the medical domain, which has heretofore been understudied. Based on fuzzy-trace theory, when people represent positively framed scenarios at the gist level, they tend to be risk averse. Therefore, we predicted that older adults would be less likely to take risks than younger adults when choosing between medications that had different probabilities and outcomes of treatment success, consistent with the recent meta-analytic findings on age differences in the risky-choice framing effect (Best \& Charness, 2015). In our study, which focused on decision making in the medical domain, older adults were not significantly less risk taking than younger adults in making medication decisions. Because prior research mainly focused on financial and mortality domains, the pattern of finding in the present study could be additional evidence that age differences in risk preferences are context dependent (Best \& Charness, 2015).

Additionally, an interaction was found in the present study between age and trial type. Younger adults exhibited a substantially larger increase in risk taking than did older adults when comparing risk-advantageous trials with riskneutral trials. This finding is consistent with our expectation that younger adults are more sensitive to the expected values of options. When presented with a risky option and a sure option with a lower expected value, younger adults were more likely to choose the risky option that maximized their expected value gain in terms of the number of days protected from sickness. Relative to younger adults, older adults showed a more similar risk-taking tendency on riskneutral and risk-advantageous trials, suggesting that they were not as sensitive as younger adults to the expected values of options. This is consistent with the idea that older adults are more likely to use gist processing whereas younger adults are more likely to use verbatim processing in making medication risky decisions.

The present study demonstrated that people's medical risk taking propensities were prone to social influence effects. In addition, it explored the effects of collaborative decisionmaking on subsequent individual decisions in two different age groups. Consistent with prior studies on intertemporal choices (Bixter et al., 2017) and financial risky decisions (Suzuki et al., 2016), we found a group convergence effect following collaboration in older adults' medication risky decisions. Importantly, older adults' convergence effect was larger than younger adults'. This might reflect their greater tendency to conform to others. Older adults might change their decisions more easily when different views are presented because of their lower perceived decision-making competence (Bruine de Bruin et al., 2012). However, it could also be due to the greater initial difference within older dyads observed in the present sample. Future research should attempt to better understand age differences in social influence effects in risky decision-making contexts.

## Limitations and Future Directions

Several limitations have to be noted. The decision task may not resemble an everyday medical context and thus makes the results less generalizable to ecological settings. Asking participants to make third-person medication decisions might introduce bias. Moreover, people might perceive avoiding sickness as categorically different than shortening the duration of sickness. Additionally, numeracy differed between the age groups, and could be an alternative explanation for the individual risk taking and social influence findings. Future research should address these issues.

## Practical Implications

Findings from the current study offer some insights into how age and collaboration influence medication risk taking. Examining age differences in medical risk seeking would enable us to devise appropriate decision aids for people of different ages. In particular, it is important to encourage older adults to take risks when risk taking is beneficial.

Examining age differences in choice shift due to social influence would inform the public how social interactions alter patients' subsequent decisions as a function of their age. Current findings suggest that other people might be able to play a significant role in influencing older patients and helping them make improved decisions.

## Acknowledgments

This research was supported in part by a pilot grant from the National Institutes of Health (National Institute on Aging) Grant P01 AG17211 under the auspices of the Center for Research and Education on Aging and Technology Enhancement (CREATE; www.create-center.org), and by the Ruth L. Kirschstein National Research Service Award (NRSA) Institutional Research Training Grant from the National Institutes of Health (National Institute on Aging) Grant 2T32AG000175-26A1.

## References

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. Econometrica: Journal of the Econometric Society, 23-36.
Best, R., \& Charness, N. (2015). Age differences in the effect of framing on risky choice: A metaanalysis. Psychology and Aging, 30(3), 688-698.
Bixter, M. T., Trimber, E. M., \& Luhmann, C. C. (2017). Are intertemporal preferences contagious? Evidence from collaborative decision making. Memory \& Cognition.
Bruine de Bruin, W., Parker, A. M., \& Fischhoff, B. (2012). Explaining adult age-related differences in decision-making competence. Journal of Behavioral Decision Making, 25(4), 352-360.
Gardner, M., \& Steinberg, L. (2005). Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study. Developmental Psychology, 41(4), 625.
Gosling, S. D., Rentfrow, P. J., \& Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37(6), 504528.

Greene, J., Hibbard, J. H., \& Tusler, M. (2005). How much do health literacy and patient activation contribute to older adults' ability to manage their health? (pp. 200505). Washington, DC: AARP Public Policy Institute.

Lau, A. Y., \& Coiera, E. W. (2008). Impact of web searching and social feedback on consumer decision making: a prospective online experiment. Journal of Medical Internet Research, 10(1).
Lipkus, I. M., Samsa, G., \& Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. Medical Decision Making,21(1), 37-44.
Mitzner, T. L., McBride, S. E., Barg-Walkow, L. H., \& Rogers, W. A. (2013). Self-management of wellness and illness in an aging population. Reviews of Human Factors and Ergonomics, 8(1), 277-333.
Peters, E., Hess, T. M., Västfjäll, D., \& Auman, C. (2007). Adult age differences in dual information processes:

Implications for the role of affective and deliberative processes in older adults' decision making. Perspectives on Psychological Science, 2(1), 1-23.
Reyna, V. F., \& Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. Developmental Review, 31(2), 180206.

Shipley, W. C. (1986). Shipley Institute of Living Scale. Los Angeles: Western Psychological Services.
Silvera, D., Martinussen, M., \& Dahl, T. I. (2001). The Tromsø Social Intelligence Scale, a self-report measure of social intelligence. Scandinavian Journal of Psychology, 42(4), 313-319.
Sunstein, C. R. (2002). The law of group polarization. Journal of Political Philosophy, 10(2), 175195.

Suzuki, S., Jensen, E. L., Bossaerts, P., \& O’Doherty, J. P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. PNAS, 113(14), 37553760.

Thompson, V. L. S. (2013). Making decisions in a complex information environment: evidential preference and information we trust. BMC Medical Informatics and Decision Making, 13(Suppl3), S7.
Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453-458.
Walther, E., Bless, H., Strack, F., Rackstraw, P., Wagner, D., \& Werth, L. (2002). Conformity effects in memory as a function of group size, dissenters and uncertainty. Applied Cognitive Psychology, 16(7), 793810.

Ward, B.W., Schiller, J.S., \& Goodman, R.A. (2014). Multiple chronic conditions among US adults: a 2012 update. Prev Chronic Dis, 11, 130389. doi:http://dx.doi.org/10.5888/pcd11.130389
Wechsler, D. (1997). WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale. Psychological Corporation.
Weller, J. A., Levin, I. P., \& Denburg, N. L. (2011). Trajectory of risky decision making for potential gains and losses from ages 5 to 85. Journal of Behavioral Decision Making, 24(4), 331-344.

# Finding Creative New Ideas: Human-Centric Mindset Overshadows Mind-Wandering 

Yung-Yi Juliet Chou (juliet.chou@columbia.edu)

Barbara Tversky (btversky@stanford.edu)<br>Teachers College, Columbia University<br>New York, NY 10027 USA


#### Abstract

Finding creative new ideas requires both release from fixation and a productive search mindset. Recent research has shown that messy desks, walking, and mind-wandering can lead to more new uses for old objects. Here we show that a humancentric mindset is superior to mind-wandering for generating more alternative uses and more creative uses because it provides both release from fixation and an effective search strategy. A human-centric mindset entails perspective-taking, and perspective-taking is likely to be an effective general strategy for enhancing creativity, problem-solving and innovation.


Keywords: creativity; design; mindset;

## Introduction

How do you get an original idea? One way to catalyze the creative process is to recombine or transform old ideas into new ones. But starting with established ideas can often be counter-productive, leading to fixation (e.g., Jansson \& Smith, 1991; Finke, Ward, \& Smith, 1992; Smith, Ward, \& Schumacher, 1993; Purcell \& Gero, 1996; Chrysikou \& Weisberg, 2005). Finding new associations is regarded as key to overcoming fixation (e. g., Finke, 1990; Finke, Ward, \& Smith, 1992; Jansson \& Smith, 1991; Mednick, 1962; Smith, Ward, \& Finke, 1995).

Recent studies have shown a variety of ways to stimulate new ideas for alternative uses of ordinary objects, a classic creativity task (Guilford, Christensen, Merrifield \& Wilson, 1978) that is also frequently used in design classes as a warm-up activity. Messy desks in contrast to tidy ones have enabled people to think of more new uses for ping-pong balls (Vohs, Redden, \& Rahinel, 2013). Messy desks create ambiguous configurations and ambiguous configurations are deliberately used by designers to generate new ideas and successful in doing so (Tversky \& Suwa, 2009). Taking a walk rather than sitting has helped people generate more novel uses for common objects (Opezzo \& Schwartz, 2014); taking a walk exposes people to new stimuli that might inspire new associations. Mind-wandering has facilitated creative incubation for finding new uses for common objects (Baird, Smallwood, Mrazek, Kam, Franklin, \& Schooler, 2012; Smallwood \& Schooler, 2006) though this strategy has not always been successful (Hao, Wu, Runco, \& Pina, 2015). The proponents of mind-wandering use neuroscience research on the default network to argue for mind-wandering (Baird, et al., 2012). The default network
is activated when the mind turns inward rather than responding to external stimuli (Mason, Norton, Van Horn, \& Wegner, 2007; Smallwood, Beach, Schooler, \& Handy, 2008). Wandering in the mind, like wandering in the world, can bring new stimuli, and consequently new responses.
Messy desks, taking a walk, and mind wandering succeed in releasing thinkers from fixation by bring in new stimuli. An even simpler manipulation, interleaving different design problems rather than blocking them, accomplishes the same (Tversky \& Chou, 2010)—remember the old adage: Take a break. But bringing in new stimuli doesn't by itself provide a productive way to search for new ideas. Innovators need effective search strategies as well as release from fixation. Designers in prominent design firms, notably IDEO, have developed a systematic approach, Human-Centric design, to do exactly that. They have instituted elaborate practices to enable their designers to put themselves in the shoes of potential users in order to design effective systems, procedures or products for the target community (Kelley \& Littman, 2006). Although widely adopted, the humancentric approach has not been systematically evaluated.
Here we evaluate the Human-Centric approach by using a design task that laypeople frequently need to do, finding new uses for everyday objects. In our daily lives we often find ourselves improvising, to grasp an object out of reach by twisting a coat hanger or to tie a shoe together with a paper clip when a shoelace has snapped. This improvised design requires finding new uses for familiar objects. The new uses task has been used in considerable previous research, including the studies that stimulated our own. It is also used as a warm-up exercise in design course typically asking students to come up with many ways to use a brick. We asked participants to find new uses for ordinary objects under three mindsets: Human-Centric, Mind-Wandering, and a control condition with no special mindset. For the Human-Centric mindset, for each object, we directed participants to think of how different human roles might use the object. We chose roles that participants would be familiar with in their everyday interactions, such as artist, chef, physician, mechanic, and athlete. We pretested the roles to make sure our intuitions were correct. We selected six objects, also after pretesting to make sure that laypeople could generate alternative uses for the objects.
Adopting the perspectives of many roles should fulfill both requirements for original ideas. Changing perspective should lead to release from fixation and taking new
perspectives should provide effective ways to search for new ideas. The Human-Centric group was asked "to imagine how different people in different roles might reuse the objects in their activities." The Mind-Wandering group was given the instructions of Baird et al. (2012) "to simply relax and let your mind wander." The control group was given no special mindset.

## Measuring Fluency and Originality of Ideas

Here ideas generated by participants were evaluated on quantity or fluency and on creativity, that is degree of originality. The primary interest is in fluency, as in design evaluating suitability can only come after ideas are generated. Evaluating originality or creativity has typically been based either on judgments of creativity or on statistical rarity in the larger group (e.g. Hennessey \& Amabile, 2010; Runco \& Jaeger, 2012; Runco, 2004). Judgments of creativity can be biased and unstable (e.g. Kaufman, Baer, Cole, \& Sexton, 2008). Here we use the Sample-Specific Percentage Score of Mouchiroud and Lubart (2001) derived from Torrance's classic paradigm (1968). One point is given to each idea given by $2-5 \%$ of the sample and two points to each idea given by less than $2 \%$ of the sample. This method has been criticized for failing to differentiate the quality of originality from the fluency of responses. For example, a participant who gave 10 common responses might get a higher total originality score than a participant who gave only 2 answers, even if the 2 answers were more unusual. However, this did not seem to be a problem in the present study as the people who gave original ideas also gave many ideas.

## Methods

## Participants

Participants $(\mathrm{N}=105)$ were recruited through Amazon Mechanical Turks Web service, receiving $\$ 5$ for approximately 40 minutes of time. Participants' ages ranged from 21-65, with a mean of 33.19 and came from a wide range of educational backgrounds. Participants were randomly assigned to the three mindset conditions. There were 18 women and 17 men in the Mind-Wandering Group, 15 women and 20 men in the Human-Centric Group, 18 women and 17 men in the Control Group.

## Stimuli

The objects were selected from a review of objects in previous research and from a pilot study to make sure that they could be decomposed and would stimulate new uses from ordinary people: broom, flashlight, chair, umbrella, shoe, and smartphone. A smartphone is representative of contemporary and future design challenges.

## Procedure

The first screen that greeted participants described everyday ingenuity, such as using a hanger to grab an out of reach
object or rolling up a magazine to swat flies. Then participants were invited to discover and generate uncommon uses for six ordinary objects. All three mindset conditions next read general instructions: "On each trial you will be presented with the name of the object. Your job is to produce as many different novel uses as you can, uses that are different from the normal use. You will type your ideas in a text box, using only a few words, one idea at a time. Please do not repeat ideas. Eventually, you may run out of new ideas and then you will have a chance to proceed to the next object and generate new ideas for it. There are SIX objects. You will have 5 minutes to generate novel uses for each object. Please do not use any resources besides your own creative mind in this task." Participants were also told: "It's OK to use more than one of the objects and it's OK to use parts of the object."

Participants in the Mind-Wandering group were told that "One proven way to generate new ideas is to simply relax and let your mind wander. Please use that mindset to generate as many new uses as you can think of." Participants in the Human-Centric group were told that "One proven way to generate new ideas is to imagine how different people in different roles might reuse the objects in their activities. Other roles might include various kinds of athletes, gardeners, artists, chefs, musicians, mechanics, craftspeople, dancers, teachers, police, firefighters, plumbers, tailors, architects, physicians, writers and more. Please imagine the mindset of a variety of roles to generate as many new uses as you can think of." The Control group was not given any specific strategy or exemplars. Each participant had a practice trial with clothes hanger for 3 minutes before starting the real experiment. The screenshot of the human-centric mindset condition are shown after participants entered responses in Figure 1. Each response was assigned a position number by the system.


Figure 1: Screenshot of the Human-Centric mindset condition after entering new responses.

Participants were then presented with the names of six objects, one at a time, for the unusual uses task: Broom, Flashlight, Chair, Umbrella, Shoe, and Smart Phone. Each of those 5 objects except for Smart Phone was randomly ordered for each participant. For each object, the common use was presented under the name of the object on the screen After generating ideas for those 5 objects, the Smart

Phone was presented along with a new instruction by adding a paragraph, "Now that you've warmed up generating new uses for old objects, try your mind at generating new uses, including new apps, for a smart phone."

After participants finished generating new uses, participants responded to a questionnaire asking whether they used the mindset strategy suggested and how easy, how helpful it was to follow. The control group was asked if they used a mindset strategy, and if so, what?

## Results

Coding Counting ideas was a two-step process. Responses were first put through a spreadsheet that (a) counted the total number of answers, and (b) identified the likely original answers by eliminating all duplicates (repeated identical answers). During the initial examination, all duplicate answers were removed; total numbers generated for each participant were accurate, and all the unusual / unique answers were identified. Generalized items (e.g., "a broom to clean off the cobweb on the ceiling") were counted toward a participant's total number of responses but were not coded as original. To measure the originality of ideas, the task was coded with Sample-Specific Percentage scoring method derived from the classic Torrance's (1968) paradigm. Examples of both original and ordinary examples are provided in Figure 2.

| Objects | More Original Examples | Most Ordinary Examples |
| :---: | :---: | :---: |
| Broom | Mini golf course with multiple brooms; Wood flavoring source by shaving the stick; Fragrance holder; | Clean cobwebs off the ceiling; Animal deterrent; Weapon; |
| Flashlight | Martini Shaker; Meat Tenderizer; Wrapping longhair with decorative lighting; | Weapon; Signal; Starting a fire; |
| Chair | Use folding chair hinges in extending arm array; Spokes for a ring loom; Water Strainer; | Step stool; Weapon; Firewood; |
| Shoe | Stack up to use as soundproofing; Voice disguiser; Bottling lace grommets as a baby rattle; | Weapon; Planter; Animal Deterrent; |
| Umbrella | Cutting the fabric into strips to tie things together; Use mental spokes as shish kabobs; | Weapon; Walking cane; Protection from the Sun; |
| Smartphone | Physical obj.: Wrist splint; Cutting board in wilderness; Digital/Physical obj.: Smoke detector; Stud finder; Steak doneness tester; | Paperweight; Flashlight; Camera; |

Figure 2: Examples of original and ordinary ideas
Fluency of Ideas The 3 mindsets differed substantially in fluency (i.e. total number of ideas); the Human-Centric group ( $M=54.06, S D=27.30$ ) generated far more ideas than the Mind-Wandering ( $M=38.77, S D=15.52$ ) and Control groups ( $M=36.54, S D=17.05$ ) in Figure 3. Because Levene's test was significant ( $p<.001$ ), revealing that variances in the Human-Centric group were differed, violating the assumption of homogeneity of variance, a more robust Games-Howell method (instead of Tukey HSD)
was applied to interpret the F statistics for the post hoc results.

Welch's Robust ANOVA showed significant differences in fluency of three mindset groups, $F(2,65.38)=5.407, p=$ .007. Continuing, the Games-Howell post hoc testing revealed that the Human-Centric mindset group generated more ideas than the mind-wandering mindset group a mean increase of $15.229,95 \%$ CI [ 2.18 to 28.27]. There was also a mean increase of $17.514,95 \%$ CI [ 4.42 to 30.61 ] between the Human-Centric and Control groups, but there was no significant mean difference between the Mind-Wandering and Control groups, 2.229, $95 \%$ CI [-7.11 to 11.57]. The mean number of ideas generated by each mindset condition for each object can be viewed in Figure 4.


Figure 3: The Human-Centric mindset group generated more uses than Mind-Wandering and Control groups.


Figure 4: The Human-Centric mindset group generated more uses than the other groups for each object.

Originality of Ideas For originality, Levene's test was significant, and the assumption of homogeneity of variance was violated. The Welch test table was applied. There was a significant effect for the three mindset conditions differed significantly in originality of ideas $F(2,67.21)=4.34, p=$ .017. The post-hoc comparison using the Games-Howell test indicated that the mean score of originality for the HumanCentric group ( $M=31.89, S D=21.67$ ) was significantly different from the Mind-Wandering group $(M=20.43, S D=$
16.61) and from the Control group ( $M=18.77, S D=17.12$ ) . There were no differences between the Mind-Wandering and Control conditions, as shown in Figure 5.

Fluency of Original Ideas Two Pearson's product-moment correlations were run to assess the relationship between the quantity and originality of ideas. The first correlation refers to the total number of ideas, a summation of the number of ideas that each participant generated across 6 objects ( 5 min per object) and the sum of the originality score for those ideas. There was a strong correlation between the quantity of ideas generated by a participant and the overall originality scores irrespective of mindset conditions in the study, $r(103)=.885, p<.001$. The overall originality score (i.e. 2 points for each idea given by less than $2 \%$ of the sample; 1 point for each idea with a frequency seen in $2 \%$ to $5 \%$ of the sample; 0 points for ideas given more than $5 \%$ of the sample) is a summation of the originality score for 6 objects. The average participant generated approximately 43 ideas in the 30 minutes of the idea generation task. The second correlation refers to the total number of ideas and the average originality of ideas for each participant. There was a moderate positive correlation between the quantity of ideas and the mean originality score (sum of originality score divided by total number of ideas), $r(103)=.434, p<.001$.

There were no differences in quantity of ideas and originality of ideas for the different objects.


Figure 5: The Human-Centric mindset group generated more original ideas than the Mind-Wandering and Control groups.


Figure 6: Participants who generated at least 10 ideas for any object were more likely to produce more original ideas.

Position of Ideas: Original Ideas Come Later Many studies have found that ideas generated later tend to be better than early ideas since Christensen, Guilford, \& Wilson (1957) first demonstrated the effect. This result aligns with those from prior studies (Beaty \& Silvia, 2012).

Two methods were used to confirm that original ideas do come later. The graph in Figure 6 shows the mean score of originality (from 0 to 2 ) for ideas that appear in the ith position, $i=1,2, \ldots 10$, regardless of conditions and objects for this study. It was reasonable to choose 10 positions, because about half of the sample size generated at least 10 ideas. It appears that participants came up with more original ideas at the later position. Another bar graph Figure 7 is to show the percentage of ideas that were original (less than $5 \%$ of the sample generated the idea) for each position.


Figure 7: Original ideas tended to come later.
Self Report / Manipulation Check Regarding whether participants used the suggested mindset strategies to generate ideas; it appears that more than $75 \%$ of participants in both Human-Centric and Mind-Wandering groups claimed that they did follow the instruction. $69 \%$ of participants in the Human-Centric group and $60 \%$ of participants in the Mind-Wandering group did think it was helpful with the suggested strategy. Regarding how easy participants used the suggested mindset strategies to generate ideas; it appears that more than $50 \%$ of participants in both Mind-Wandering and Human-Centric groups self-reported it was easy for them to use the mindset. $89 \%$ of participants in the Control group self-reported that they simply let things come to mind, using a mindwandering mindset strategy.

## Discussion

Designers and problem solvers--and we are all designers and problem solvers--often get stuck. They/we get fixated on one idea or a set of them and then thinking goes in circles. Breaking fixation, breaking that circle, finding new ideas requires new associations. Messy desks, walks, and mind-wandering have all proven helpful for finding new uses for familiar objects. They work because each leads to new stimuli and new stimuli can bring new associations and perhaps new ideas.

Although wandering eyes, wandering bodies, and wandering minds can expose us to new stimuli, the paths of
search are still wandering, not directed in any meaningful way. There is no guarantee that the meandering and the associations are in any way related to the design or problem. Designers and problem solvers also need productive ways to search for and generate new ideas that are relevant to the problem at hand. The human-centric mindset does just that. The human-centric approach entails taking the perspectives of others, here diverse roles that participants are familiar with. Participants could make use of their knowledge of the roles to generate relevant uses: what could a gardener do with an umbrella? An artist with a shoe? An athlete with a chair? Participants with the Human-Centric mindset did use the roles that we gave them, and invented new roles of their own. Using a high criterion for relevance, nearly half the ideas generated by the group using the Human-Centric approach were directly related to one of the roles provided and another $10 \%$ derived from roles they invented, presumably because they used the mindset to take the perspectives of various roles. The most productive role was artist, followed by gardener, athlete, policeperson, mechanic, chef, and musician.

Consistent with that analysis, the Human-Centric mindset yielded more ideas than either the Mind-Wandering mindset or the no-mindset control. In fact, the Mind-Wandering mindset was no more successful than the no-mindset control group at generating new uses. This turns out to be unsurprising; in response to a question about how they searched for new ideas, many in the control group reported that they just let their minds wander.

Participants using the Human-Centric mindset generated more new uses and also generated more original new uses than those who adopted the other mindsets. Original new uses tended to come later; it's as if participants have to first get the ordinary alternative uses out of their heads in order to free their minds to find unusual ones. Sadly, one of the most common uses suggested for most of the objects was weapon. The vast majority of original responses were not only reasonable and appropriate, but clever, even if unusual. Remember that the instructions allowed using more than one of the objects. For a shoe, sound-proofing; for a chair, a water strainer; for a smart phone, a wrist splint, for a flashlight, a martini shaker.

Because of the overall quality of the original ideas, it is apparent that participants were editing their own responses. That process, of generating ideas and evaluating them, is supported by neuroscience research (Beaty, Benedek, Silvia, \& Schacter, 2016; Chrysikou, in press; Ellamil, Dobson, Beeman, \& Christoff, 2012; Mason, Norton, Van Horn, Wegner, Grafton, \& Macrae, 2007). The neuroscience findings suggest that creative problem solving is characterized by alternating activation in the default network, indicative of internal processing, and the frontal system, indicative of executive control. This iterative process, of generating ideas and evaluating them coincides with the experience of designers and problem solvers. It remains to be seen whether the neuroscience tools are sensitive enough to detect the large differences in mindset
demonstrated here. In the meantime, it should be clear that Mind-Wandering is not to be recommended as a general mindset for finding innovative ideas. A Human-Centric mindset is far more productive.
The Human-Centric mindset clearly has wide applicability. Diplomats negotiating peace agreements take the perspectives of each party, as do lawyers. Writers of books and screenplays take the perspectives of their readers or viewers. Product designers think deeply about the ways different users will interact with their products. Using the Human-Centric mindset entails adopting relevant and varying human roles. Yet there are many problems that demand creative solutions but do not involve humans, except as thinkers. Problems in mathematics or physics. Design of machines or robots for tasks that do not involve humans, except as designers. Taking different human perspectives might help, but that is probably not be the best way for those problems. However, taking different human perspectives is at its foundation taking different perspectives, and that mindset might just work for everything. Or nearly everything. Mathematicians reframe problems algebraically or geometrically. Temple Grandin, in designing runways for cattle, famously adopts the perspective of the cattle (Grandin \& Deesing, 2008). Taking different perspectives might sound simple, but deciding which alternative perspectives are relevant and productive also requires creative thought.

## Acknowledgments

We are grateful to the following grants that provided direct or indirect support to the research described or to the preparation of the manuscript: NSF CHS-1513841, NSF HHC 0905417, NSF IIS-0725223, NSF IIS-0855995, and The John Templeton Foundation through The Varieties of Understanding Project at Fordham University. The opinions expressed here are those of the author and do not necessarily reflect the views of The Varieties of Understanding Project, Fordham University, or The John Templeton Foundation.

## References

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W. Y., Franklin, M. S., \& Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. Psychological Science, 23, 1117-1122.
Beaty, R. E., Benedek, M., Silvia, P. J., \& Schacter, D. L. (2016). Creative cognition and brain network dynamics. Trends in cognitive sciences, 20(2), 87-95.
Beaty, R. E., \& Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. Psychology of Aesthetics, Creativity, and the Arts, 6(4), 309-319.
Chrysikou, E. G. (in press). The costs and benefits of cognitive control for creativity. In O. Vartanian and R. E. Jung (Eds.), The Cambridge Handbook of the Neuroscience of Creativity. Cambridge University Press.
Chrysikou, E. G., \& Weisberg, R. W. (2005). Following the wrong footsteps: fixation effects of pictorial examples in
a design problem-solving task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 11341148.

Ellamil, M., Dobson, C., Beeman, M., \& Christoff, K. (2012). Evaluative and generative modes of thought during the creative process. Neuroimage, 59(2), 17831794.

Finke, R. A. (1990). Creative imagery: Discoveries and inventions in visualization. Hillsdale NJ: Erlbaum.
Finke, R. A., Ward, T. B., \& Smith, S. M. (1992). Creative Cognition: Theory, Research, and Applications. Cambridge, MA: MIT Press.
Grandin, T., \& Deesing, M. (2008). Human livestock handling: understanding livestock behavior and building facilities for healthier animals. Storey publ., North Adams.
Guilford, J. P., Christensen, P. R., Merrifield, P. R., \& Wilson, R. C. (1978). Alternate uses: Manual of instructions and interpretations. Orange, CA: Sheridan Psychological Services.
Hao, N., Wu, M., Runco, M. A., \& Pina, J. (2015). More mind wandering, fewer original ideas: Be not distracted during creative idea generation. Acta Psychologica, 161, 110-116. http://doi.org/10.1016/j.actpsy.2015.09.001
Hennessey, B. A., \& Amabile, T. M. (2010). Creativity. Annual Review of Psychology, 61(1), 569-598.
Jansson, D. G., \& Smith, S. M. (1991). Design fixation. Design Studies, 12(1), 3-11.
Kaufman, J. C., Baer, J., Cole, J. C., \& Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. Creativity Research Journal, 20, 171-178.
Kelley, T., \& Littman, J. (2006). The ten faces of innovation: IDEO's strategies for defeating the devil's advocate and driving creativity throughout your organization. Crown Business.
Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., \& Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. Science, 315(5810), 393-395.
Mednick, S. (1962). The associative basis of the creative process. Psychological Review, 69, 220-232.
Mouchiroud, C., \& Lubart, T. (2001). Children's original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. The Journal of Genetic Psychology, 162(4), 382-401.
Oppezzo, M., \& Schwartz, D. L. (2014). Give your ideas some legs: The positive effect of walking on creative thinking. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(4), 1142-1152.
Purcell, A. T., \& Gero, J. S. (1996). Design and other types of fixation. Design Studies, 17, 363-383.
Runco, M. A. (2004). Everyone has creative potential. In R. J. Sternberg, E. L. Grigorenko, \& JL. Singer (Eds.), Creativity: From Potential to Realization (pp. 21-30) Washington, DC: American Psychological Association.

Runco, M. A., \& Jaeger, G. J. (2012). The standard definition of creativity. Creativity Research Journal, 24(1), 92-96.
Silvia, P. J. (2008). Creativity and intelligence revisited: A latent variable analysis of Wallach and Kogan (1965). Creativity Research Journal, 20(1), 34-39
Smallwood, J., and Schooler, J. W. (2006). The restless mind. Psychological Bulletin, 132, 946-958.
Smallwood, J. Beach, E., Schooler, J. W. and Handy, T. C. (2008). Going AWOL in the brain: mindwandering reduces cortical analysis of external events. Journal of Cognitive Neuroscience, 20-3, 458-469.
Smith, S. M., Ward, T. B., \& Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. Memory \& Cognition, 21(6), 837-845.
Smith, S. M., Ward, T. B., \& Finke, R. A. (Eds.). (1995). The creative cognition approach. Cambridge: MIT.
Torrance, E. P. (1968). Torrance tests of creative thinking. Princeton, N.J: Personnel Press, Inc.
Tversky, B. \& Suwa, M. (2009). Thinking with sketches. In A.B. Markman \& K.L. Wood (Eds.) Tools for innovation. Pp. 75-84. Oxford: Oxford University Press.
Tversky, B. and Chou, J. Y. (2010). Creativity: Depth and breadth. In T. Taura and Y. Nagai (Editors). Design creativity. Pp. 209-214. Dordrecht, Netherlands: Springer.
Vohs, K. D., Redden, J. P., \& Rahinel, R. (2013). Physical order produces healthy choices, generosity, and conventionality, whereas disorder produces creativity. Psychological Science, 24, 1860-1867.
Wilson, R. C., Guilford, J. P., \& Christensen, P. R. (1953). The measurement of individual differences in originality. Psychological Bulletin, 50(5), 362-370.

# Transcranial Direct Current Stimulation (tDCS) and the Face Inversion Effect: Anodal stimulation at $\mathbf{F p} 3$ reduces recognition for upright faces 

Ciro Civile (ciroc@mcmaster.ca)<br>Department of Psychology, Neuroscience, \& Behaviour, McMaster University, Canada.

Sukhvinder Obhi (obhi@mcmaster.ca)<br>. Department of Psychology, Neuroscience, \& Behaviour, McMaster University, Canada.

I.P.L. McLaren (I.P.L.McLaren@exeter.ac.uk)<br>School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.


#### Abstract

Perceptual learning is a key perceptual skill that people possess, in particular, it contributes to their ability to distinguish between faces thus recognize individuals. Recently, we showed that anodal transcranial Direct Current Stimulation (tDCS) at Fp3 abolishes the inversion effect (that would otherwise exist) for familiar checkerboards created from a prototype. Because of the close analogy between the inversion effect obtained with checkerboards, which we use as a marker for perceptual learning, and the traditional face inversion effect (upright faces recognized better than inverted ones), we investigated the effects of anodal tDCS at Fp3 during an old/new recognition task for upright and inverted faces. Results showed that stimulation significantly reduced the face inversion effect compared to controls. The effect was strongest in reducing recognition performance to upright faces. This result supports our account of perceptual learning and its role as a key factor in face recognition.


Keywords: TDCS; Perceptual learning; Face inversion effect; Old/new recognition task; Face recognition

## Introduction

Perceptual learning refers to an enhanced ability to distinguish between similar stimuli as a consequence of experience with them or stimuli like them. It also plays a key role in learning to identify stimuli as specific exemplars of a category, and not confuse one stimulus with another similar one (e.g. wine experts and wines, or bird watchers and warblers; James, 1890; see Hall, 1980 for a review). We know that people (and other animals) can improve their perceptual skills as a result of experience with stimuli, and recent studies have shown this phenomenon to be responsible for some key perceptual skills that people possess. In particular, it contributes to our ability to distinguish between faces and recognize individuals. For example, if we pre-expose someone to a set of checkerboards, all of which are produced by imposing random variation on one original prototype checkerboard, then this will have the effect of making them better able to distinguish between exemplars generated in this way -a
basic perceptual learning effect. They will now be able to tell two otherwise similar checkerboards apart where once they might have found it difficult to do so, and such preexposure improves their ability to identify checkerboards they have been asked to memorize in a subsequent recognition test (McLaren, Leevers \& Mackintosh, 1994). McLaren (1997) extended this result to show that the same procedures could also produce an inversion effect, with upright exemplars discriminated better than inverted ones.

Civile et al. (2014) further developed the case for perceptual learning as a contributor to the face inversion effect (i.e. that upright faces are recognized much better than inverted ones), by showing that these results can be obtained with the kind of old/new recognition paradigm conventionally used in such studies (Yin, 1969; Diamond \& Carey, 1986; see Maurer, Le Grand, \& Mondloch, 2002 for a review). Participants were trained to categorize (categorization task) checkerboard exemplars from two prototype-defined categories (the pre-exposure phase), before being shown an equal number of checkerboard exemplars (which they had not previously encountered) drawn from either one of the now familiar categories or a novel category, half of which were upright and half inverted. Participants were then tested for recognition of these exemplars after this study phase. The results confirmed the inversion effect for checkerboard exemplars drawn from a familiar category, and its absence for exemplars drawn from a novel category, strengthening the case for perceptual learning contributing to the inversion effect found with faces.

In a recent study, Civile et al. (2016) demonstrated that tDCS to dorsolateral prefrontal cortex (DLPFC) at Fp3 site significantly affected perceptual learning and reduced the inversion effect that can otherwise be obtained with checkerboards. The authors adopted the same old/new recognition task as in Civile et al. (2014)'s study which uses a categorization task to pre-expose participants to the stimuli i.e. checkerboards. A previous study by Ambrus et al., (2011) had found that anodal tDCS (compared to sham)
applied to the Fp 3 during the training phase of a categorization task where participants had to identify prototype and low-distortion patterns as category members reduced classification accuracy for the prototype. Thus, as Civile et al. (2014)'s study used prototype-defined checkerboard categories and formation of a strong representation of the prototype is a prerequisite for perceptual learning (McLaren, Kaye, \& Mackintosh,1989), Civile et al. (2016)'s study adopted the same Fp3 montage as that adopted by Ambrus et al. (2011). Civile et al. (2016) showed that the control condition (sham tDCS stimulation over Fp 3 delivered during the pre-exposure phase, i.e. the checkerboard categorization task) replicated the usual inversion effect for checkerboards drawn from a familiar category, but, as expected, not for checkerboard exemplars drawn from a control (novel) category that had not been preexposed. Critically, anodal tDCS to the same brain region changed this pattern, as there was now no inversion effect for stimuli drawn from either familiar or unfamiliar category, and the upright exemplars drawn from a familiar category were less well recognized than those drawn from the novel category, an indication that perceptual learning may even have been reversed. This remarkable and informative result suggested that perceptual learning in humans could be turned 'on' and 'off".

Civile et al.'s (2016) study is the first evidence that anodal tDCS administered during the pre-exposure phase can affect perceptual learning later on when participants are asked to memorize and recognize exemplars of checkerboards drawn from the checkerboard categories seen in during the pre-exposure phase (categorization task). The next important question to address is whether or not the same tDCS procedure would also affect perceptual learning that has already taken place. Given the lifelong expertise we have for faces, and given the already established analogy between the inversion effect obtained with checkerboards (McLaren, 1997; McLaren \& Civile, 2011; Civile et al., 2014; Civile et al., 2016) and that usually obtained with faces (for a review see Maurer et al., 2002), in the current study we extended the tDCS paradigm used in Civile et al.' (2016) to the inversion effect for faces. We expected to obtain a strong inversion effect for familiar faces in the sham tDCS group, but a significantly reduced inversion effect for familiar faces in the anodal tDCS group because, as was the case for Civile et al.'s (2016) familiar upright checkerboards, we expected anodal tDCS over Fp3 to disrupt recognition performance for familiar upright faces.

[^304]tDCS stimulation could reduce our ability to recognize upright familiar faces.

## Method

We adopted the tDCS montage used in Civile et al. (2016). Each subject was randomly assigned to either sham or anodal tDCS conditions. In the sham condition, the tDCS stimulation was only delivered for 30 s, to evoke the sensation of being stimulated, without causing neurophysiological changes that may influence performance. In the anodal tDCS condition, the stimulation was delivered for 10 mins while the subjects were completing an old/new recognition computer task that used images of faces. In both sample groups, the sham and tDCS stimulation started when the computer task began. In the first part of the computer task, the study phase, subjects were asked to memorize a set of upright and inverted faces presented one at a time. Following this, subjects were given a recognition task where they pressed one key if they thought they had seen the face before, and another key if they thought they had not seen the face before. All the faces seen in the study phase were presented again intermixed with an equal number of new faces of each type (i.e. upright faces, and inverted faces). This old/new recognition task is a standard method of assessing face processing and the inversion effect (Yin, 1969; Diamond \& Carey, 1986; Civile, McLaren, \& McLaren, 2016; Civile, McLaren, \& McLaren, 2014). Our main measure was accuracy scores during recognition converted into signal-detection d-prime " $d$ '". We also examined reaction time responses to check for any speed-accuracy trade-off that could affect our interpretation of the results.

## Subjects

Forty-eight students ( 39 women; mean age $=18.9$, age range $=18-22$ years) from McMaster University participated in this experiment. Twenty-four subjects were randomly assigned to each of two groups (sham tDCS, anodal tDCS). All subjects were right-handed and were given course credits for their participation. The experiment was approved by the research ethics committee at McMaster University. Written informed consent was obtained after the nature and possible consequences of the study were explained. Sample size was determined in advance based on previous studies (Civile et al., 2014; McLaren 1997) that found the original inversion effect for checkerboards and that showed a clear effect of tDCS on perceptual learning (Civile et al., 2016; McLaren, Carpenter, Civile, McLaren, Zhao, Ku, Milton, Verbruggen, 2016), as well as previous studies that adopted the same old/new recognition task and face stimuli that we used here (Civile, McLaren, McLaren, 2014; and Civile, McLaren, McLaren, 2016 obtained a strong face inversion effect with group samples of 24 subjects). Additionally, we conducted a post-hoc power analysis using G*power software (Faul, Erdfelder, Lang, \& Buchner, 2007) that revealed a statistical power of 0.92 , in
line with the recommended 0.80 level of power (Cohen, 1988).

## Materials

The study used 128 images of male faces. Only male faces were used because they allowed the inclusion of ears in the images as well. Men tend to have shorter hair with ears visible whereas women often have longer hair covering the ears, making the visibility of these features rather variable. The faces were standardized in gray-scale format and cropped around the hairline in Adobe Photoshop. The same set of faces was previously used in studies that adopted the same old/new recognition task with upright and inverted faces that we used in the study here reported (Civile, McLaren, \& McLaren, 2014; Civile, McLaren, \& McLaren, 2016).

## Transcranial Direct Current Stimulation (tDCS)

All participants first completed a brain stimulation safety screening questionnaire. Stimulation was delivered by a battery driven, constant current stimulator (Neuroelectrics) via a pair of surface sponge electrodes ( $25 \mathrm{~cm}^{2}$ ), soaked in a saline solution $(0.9 \% \mathrm{NaCl})$, and applied to the scalp at the target areas of stimulation. Electrodes delivered a constant current of 1.2 mA (current density: $0.048 \mathrm{~mA} / \mathrm{cm}^{2}$ ); the choice of the intensity is in line with Civile et al. (2016)'s study (see Neuroelectics website for a review of clinical studies that suggest keeping the average current densities in electrodes below $0.06 \mathrm{~mA} / \mathrm{cm} 2$ ). As in Civile et al. (2016)'s study, we adopted a bilateral bipolar-non-balanced montage with one of the electrodes (anode/target) placed over the left PFC ( Fp 3 ) and the other (Ambrus et al., 2011; Kincses et al., 2003) was placed on the forehead, just above the right eyebrow. In the anodal tDCS condition, the current was applied for 10 mins (fade-in and fade-out of 5 s ) from when the subjects began the computer task and throughout the old/new recognition task. Sham received the same 5 s fadein and fade-out, but only 30 s stimulation between them, which terminated shortly after the computer task started. The electrodes were left on the participant throughout the experiment (see Figure 1, Panel A).

## Behavioral Task

The old/new recognition task consisted of two parts: a 'study phase' and an 'old/new recognition phase' (Civile, McLaren, \& McLaren, 2014; Civile, McLaren, \& McLaren, 2016). In the study phase, each subject was shown upright and inverted faces with 32 images for each type ( 64 images in total). Faces were presented one at a time in random order. In the old/new recognition phase, 64 novel faces split into the same stimulus types were added to the 64 faces seen in the study phase, and all 128 images were presented one at a time in random order. Each face never appeared in more than one condition during the experiment for the same participant.

## Trial Structure

Following the instructions, in each trial of the study phase subjects saw a fixation cross in the center of the screen presented for 1 second. After this, one of the faces was presented on screen for 4 seconds. The next trial started with the presentation of a fixation cross again. After all 64 faces had been presented, the program displayed another set of instructions, explaining the recognition task. In this task, subjects were asked to press the '. ' key if they recognized the stimulus as having been shown in the study phase on any given trial, or press ' $x$ ' if they did not (the keys were counterbalanced). During the recognition task, the faces were shown for 4 seconds during which time subjects had to respond. The experiment was implemented using SuperLab 4.5 installed on a PC (see Figure 1, Panel B).


Figure 1: Panel a shows the electrode configuration of the tDCS and the stimulation set up on the Neuroelectrics software (NIC). Panel b shows the structure of the trials presented during the old/new recognition task.

## Data Analysis

Our primary measure was performance accuracy in the two recognition tasks. The data from all the participants was used in the signal detection $d^{\prime}$ analysis of the recognition task (old and new stimuli for each stimulus type) where a $d^{\prime}=$ of 0.00 indicates chance-level performance (Stanislaw \& Todorov, 1999). Each p-value reported in this paper is two-tailed, and we also report the F or t value along with measures of variability (SE or $S E M$ ) and effect size (Cohen's $d$ followed by the $95 \%$ confidence interval [CI] for d). The study had a $2 \times 2$ mixed model design using as a within-subjects factor Face Orientation (upright, inverted) and the between-subjects factor $t D C S$ (sham, anodal). Follow up, paired $t$-tests analyses were conducted to compare performance on upright and inverted faces (the inversion effect) in each tDCS group (sham, anodal). We also assessed performance against chance ( $d^{\prime}$ of 0 ) to show that both upright and inverted faces in the tDCS sham and anodal groups were recognized (for all four conditions we found a $p<.001$ ).

## Results

The statistical analysis (ANOVA) using the factors Face Orientation (upright/inverted) x $t D C S$ (anodal/sham) revealed a significant interaction, $F(1,46)=7.45, M S E=$ $0.12, p=.009, d=0.78, \mathrm{CI}=0.98,0.58$. We decomposed the interaction by looking at the inversion effect (upright faces - inverted faces) in each tDCS group (sham, anodal) separately. Following Civile et al's (2016) study, we expected to find the usual inversion effect for faces in the tDCS sham group. As predicted, a planned comparison showed a significant inversion effect with upright faces ( $M$ $=1.09, S E=0.11$ ) being recognized significantly better than inverted faces $(M=0.35, S E=0.07), t(23)=7.48, S E=$ $0.09, p<.001, d=1.59, \mathrm{CI}=1.78$, 1.41. Critically, we found a reduced (but still significant) inversion effect in the tDCS anodal group, recognition of upright faces ( $M=0.78$, $S E=0.11)$ compared to inverted faces $(M=0.44, S E=$ $0.08), t(23)=3.19, S E=0.11, p=.004, d=0.69, \mathrm{CI}=0.89$, 0.49 (see Figure 2). Thus, the inversion effect in the tDCS sham group was significantly greater than that in the tDCS anodal group, a similar result to that previously found in Civile et al. (2016)'s study using prototype-defined categories of familiar checkerboards.

Importantly, in Civile et al. (2016)'s study (Experiment 1) statistical analysis showed recognition of upright familiar checkerboards in the tDCS anodal group was reduced compared to that for familiar checkerboards in the tDCS sham group. We computed an additional analysis in our study to directly compare the recognition performance for upright faces in the two tDCS groups (sham, anodal). The results were that recognition for upright faces in the tDCS anodal group was reduced compared to that in the tDCS sham group, $t(46)=1.95, S E=0.14, p=$ .028 (1-tail), $d=0.56, \mathrm{CI}=0.78,0.34$. Thus, in both Civile et al. (2016)'s study (Experiment 1) and in our current
study, we have some evidence that anodal tDCS may affect the recognition of upright familiar stimuli (checkerboards in Civile et al, 2016, and faces in the current study). We calculated the Bayes factor using the procedures outlined by Dienes (2011) for this effect with faces using the effect for checkerboards in Civile et al. (2016)'s study (Experiment 1) as the prior, setting the standard deviation of $p$ (population value |theory) to the mean for the difference between recognition for familiar upright checkerboards in the tDCS sham group vs that in the tDCS anodal group (0.359). We used the standard error and the mean difference for tDCS sham upright faces vs tDCS anodal upright faces effect found in our study and assumed a one-tailed distribution for our theory and a mean of 0 . This gave a Bayes factor ( $B$ ) of 3.65. This factor is greater than 3, providing good support for this component of the reduction in the inversion effect (for Bayes factor calculator see Dienes, 2011).

Statistical analysis (ANOVA) of the response latencies was also conducted. Simple comparisons showed a significant inversion effect for both Anodal (p <.001) and Sham ( $\mathrm{p}=.009$ ) groups, and the inversion effect was numericaly larger for the Anodal group, but no significant interaction ( $p=.63$ ) was found. For completeness, we report the mean latencies for each stimulus condition: Sham upright faces, 1.37 s ; Sham inverted faces, 1.47 s ; Anodal upright faces, 1.48 s ; Anodal inverted faces, 1.61 s .

Finally, we also report here the SDT Bias estimates for each of the four stimulus' conditions: Sham upright faces, $\beta=1.33$; Sham inverted faces, $\beta=1.12$; Anodal upright faces, $\beta=1.70$; Anodal inverted faces, $\beta=1.04$.


Figure 2: The $y$-axis gives d' means for the old/new recognition task (higher _ better, $0_{\sim}$ chance), and the different stimulus' conditions in the two tDCS groups (sham, anodal) are shown on the $x$-axis. The dimensions of the stimuli were $6.95 \mathrm{~cm} \times 5.80 \mathrm{~cm}$. Participants sat 1 m away from the screen on which the images were presented.

## Discussion

We adopted the same procedures used in Civile et al. (2016) employing the old/new recognition task for faces that is a standard in the literature. The results indicate that anodal tDCS impaired recognition performance for upright faces, and as a consequence, the inversion effect was significantly reduced compared to the usual inversion effect found with faces that can be seen in the sham condition.

The MKM model (McLaren, Kaye and Mackintosh, 1989) and its later development in McLaren and Mackintosh (2000) and McLaren, Forrest and McLaren (2014) can explain the inversion effects reported by McLaren (1997) and Civile et al (2014) by appealing to perceptual learning as a consequence of experience with the category. But if the salience modulation based on prediction error implemented by this model is disrupted (by anodal tDCS ), then the MKM model turns into one more akin to McClelland and Rumelhart's (M\&R) (1985) model of categorization, and enhanced generalization between exemplars as a consequence of familiarity with that category is predicted rather than the enhanced discriminability that is the hallmark of perceptual learning. The result is the elimination of the inversion effect seen with artificial stimuli (that we take to be entirely due to perceptual learning), and even some reversal of the perceptual learning effect, explaining the pattern observed by Civile et al (2016). This interpretation of the results from Civile et al. (2016)'s study also applies to Ambrus et al. (2011)'s finding that tDCS reduces learning to the prototype, and increases generalization to random patterns. This would result in the elimination of the prototype effect, which is what we would expect if the MKM model of perceptual learning were, in effect, to be turned into the M\&R model of categorization by turning off the error-based modulation of salience that is the hallmark of MKM.

Our present data imply that anodal tDCS to Fp3 not only affects perceptual learning for artificial stimuli (the checkerboards in Civile et al., 2016) that were novel until encountered in the experimental setting but can also affect the long established perceptual learning for faces that is a result of experience over many years. This is a truly striking result that suggests that perhaps anodal tDCS over Fp3 may prevent individuals from exploiting "expertise" when called on to discriminate between stimuli of a class they are very familiar with.

These data strengthen the analogy between our checkerboard experiments and those with faces. In both cases, anodal tDCS reduces the inversion effect and reduces performance on upright exemplars taken from a familiar category. This suggests that the inversion effect obtained with what were novel, artificial stimuli, and that we attribute to perceptual learning, is at least one component of the face inversion effect. True, the inversion effect was completely eliminated by anodal stimulation in Civile et al (2016) but is still present in our stimulation group when we use faces. This could mean that any disruption of perceptual learning (which might be expected to be stronger after many years of
experience) is not complete in the current experiment, or it might be that there is a component of the face inversion effect that is not due to perceptual learning. We cannot say at present. What we can say is that the theory we have of how anodal tDCS to Fp3 works predicted a reduced inversion effect, and our salience modulation via error account of perceptual learning is, to that extent, further validated. We have also shown that we can turn perceptual learning in humans on and off, which opens the door to future applications.

These data also contribute to a recent line of studies that tested that effects of tDCS stimulation delivered at occipital brain regions on face recognition tasks. In one study the authors tested tDCS stimulation on an orientation judgment task for faces while recording brain activity with EEG. Results showed that anodal tDCS compared to sham, significantly reduced the N170 for both upright and inverted faces, despite not affecting the size of the inversion effect (Yang et al., 2014, Experiment1). In the same study (Experiment 2) the authors also showed that the same tDCS paradigm applied before a composite face effect task (the effect refers to an impairment at recognizing the top half of a familiar face when matched with the bottom half of another face) can significantly reduce the composite effect by enhancing performance for incongruent faces (composite faces created by mismatched top and bottom halves). In a similar vein, another study found that off-line (stimulation delivered before the task) anodal tDCS enhances memory performance for both upright faces and objects (inversion was not tested). In contrast, no enhancement was found for online (stimulation delivered during task execution) and sham tDCS stimulation (Barbieri et al., 2016). Together, the results from these studies show that tDCS at occipital regions seems to be effective at enhancing recognition performance (at least when tDCS is delivered off-line). Thus, this suggests that tDCS at occipital brain regions could possibly enhance perceptual learning in our experimental paradigm (either with checkerboards or faces). Future studies should test this and directly compare the effect of tDCS at Fp3 with that of tDCS at occipital sites during (and off-line) using Civile et al. (2016)'s checkerboard paradigm and our face paradigm.

## References

Ambrus G. G., Zimmer M., Kincses Z. T., Harza I., Kovacs G., Paulus W., and Antal, A. (2011). The enhancement of cortical excitability over the DLPFC before and during training impairs categorization in the prototype distortion task. Neuropsychologia 49, 1974-1980.
Barbieri, M., Negrini, M., Nitsche, M., and Rivolta, D. (2016). AnodaltDCS over the human right occipital cortex enhances the perception and memory of both faces and objects. Neuropsychologia, 81, 238-244
Civile, C., Zhao, D., Ku, Y., Elchlepp, H., Lavric, A., and McLaren, I.P.L. (2014). Perceptual learning and inversion effects: Recognition of prototype-defined familiar checkerboards. Journal of Experimental

Psychology: Animal Behavior Processes, 40, 144-61.
Civile, C., Verbruggen, F., McLaren, R., Zhao, D., Ku, Y., and McLaren, I.P.L. (2016). Switching off perceptual learning: Anodal transcranial direct current stimulation (tDCS) at Fp3 eliminates perceptual learning in humans. Journal of Experimental Psychology: Animal Learning and Cognition, 42, 290-296.
Civile, C., McLaren, R., and McLaren, I.P.L. (2016). The face inversion effect: Roles of first and second-order relational information. The American Journal of Psychology, 129, 23-35.
Civile, C., McLaren, R., and McLaren, I. P. L. (2014b). The face inversion effect: Parts and wholes. The Quarterly Journal of Experimental Psychology, 67, 728-746.
Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2nd ed.). HillSEale, NJ: Lawrence Earlbaum Associates.
Diamond, R. \& Carey, S. (1986). Why faces are and are not special: An effect of expertise. Journal of Experimental Psychology: General, 115, 107-117.
Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? Perspectives on Psychological Science, 6, 274-290.
Faul, F., Erdfelder, E., Lang, A., and Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. Behaviour Research Methods, 39, 175-191.
Hall, G. (1980). Exposure learning in animals. Psychological Bulletin, 88, 535-550.
James, W. (1890). Principles of psychology. New York: Holt.
Kincses T. Z., Antal A., Nitsche M. A., Bártfai O., and Paulus W. (2003). Facilitation of probabilistic classification learning by transcranial direct current stimulation of the prefrontal cortex in the human. Neuropsychologia, 42, 113-117
Maurer, D., Le Grand, R., and Mondloch, C. (2002). The many faces of configural processing. Trends in Cognitive Science, 6, 255-260.
McClelland, J.L. \& Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114, 159-197.
McLaren, I.P.L. (1997). Categorization and perceptual learning: An analogue of the face inversion effect. The Quarterly Journal of Experimental Psychology 50A, 257-273.
McLaren, I.P.L., Carpenter, K., Civile, C., McLaren, R., Zhao, D., Ku, Y., Milton, F., and Verbruggen, F. (2016). Categorisation and Perceptual Learning: Why tDCS to Left DLPC enhances generalisation. Associative Learning and Cognition. Homage to Prof. N.J. Mackintosh. Trobalon, J.B., and Chamizo, V.D. (Eds.), University of Barcelona.
McLaren, I.P.L., and Civile, C. (2011). Perceptual learning for a familiar category under inversion: An analogue of face inversion? In L. Carlson, C. Hoelscher, \& T.F.

Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society, (pp. 33203325). Austin, TX: Cognitive Science Society.

McLaren, I. P. L., Leevers, H. L., \& Mackintosh, N. J. (1994). Recognition, categorisation and perceptual learning. In C. Umilta \& M. Moscovitch (Eds.), Attention \& performance XV (pp. 889-909). Cambridge, MA: MIT Press.
McLaren, I.P.L., Kaye, H. \& Mackintosh, N.J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In R.G.M. Morris (Ed.) Parallel Distributed Processing - Implications for Psychology and Neurobiology. Oxford, Oxford University Press.
McLaren, I.P.L. and Mackintosh, N.J. (2000). An elemental model of associative learning: Latent inhibition and perceptual learning. Animal Learning and Behavior, 38, 211-246.
McLaren, I.P.L., Forrest, C.L., McLaren, R.P. (2012). Elemental representation and configural mappings: combining elemental and configural theories of associative learning. Learning and Behavior, 40, 320333.

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. Behaviour Research Methods, Instruments, and Computers, 31, 137-149
Yang, L-Z., Zhang, W., Shi, B., Yang, Z., Wei, Z., Gu, F., Zhang, J., Cui, G., Liu, Y., Zhou, T., Zhang, X., and Rao, H. (2014). Electrical Stimulation over bilateral occipito-temporal regions reduces N 170 in the right hemisphere and the Composite Face Effect. PLoS ONE 9(12): e115772. pmid:25531112
Yin, R. K. (1969). Looking at upside-down faces. Journal of Experimental Psychology, 81, 141-145.

# Beliefs about sparsity affect causal experimentation 

Anna Coenen ${ }^{1}$, Neil R. Bramley ${ }^{1}$, Azzurra Ruggeri ${ }^{2}$, and Todd M. Gureckis ${ }^{1}$<br>${ }^{1}$ Department of Psychology, NYU, New York; ${ }^{2}$ MPRG iSearch, Max Planck Institute for Human Development, Berlin<br>\{anna.coenen, neil.bramley, todd.gureckis\}@nyu.edu, ruggeri@mpib-berlin.mpg.de


#### Abstract

What is the best way of figuring out the structure of a causal system composed of multiple variables? One prominent idea is that learners should manipulate each candidate variable in isolation to avoid confounds (known as the "Control of Variables" strategy). Here, we demonstrate that this strategy is not always the most efficient method for learning. Using an optimal learner model which aims to minimize the number of tests, we show that when a causal system is sparse, that is, when the outcome of interest has few or even just one actual cause among the candidate variables, it is more efficient to test multiple variables at once. In a series of behavioral experiments, we then show that people are sensitive to causal sparsity when planning causal experiments.


Keywords: information search; causal learning; hypothesis testing

## Introduction

To develop a causal understanding of the world, we often need to find out how multiple candidate variables affect an outcome of interest. This problem arises in everyday situations (e.g., "Which of these switches can turn on the bathroom fan?"), during scientific exploration ("Which of these chemicals affect reaction $x$ ?"), and plays a role in answering economic and social questions ("How do these policies affect GDP?"). Often, the quickest and most effective method of resolving the causal relationships between variables and outcomes is to conduct experiments that manipulate variables (e.g., turning switches on or off) and help to decouple causation and correlation (Pearl, 2009).

In this paper, we explore how people interact with a novel causal system to understand how it works by manipulating multiple independent variables over a series of trials. We start by introducing two strategies for causal experimentation, one of which is so well known that it is codified in contemporary STEM education standards. We then describe an analysis of an optimal experimenter (i.e., an ideal actor model) which shows how the most informative strategy for learning critically depends on a learner's knowledge about the number of causes among the variables. We then evaluate the predictions of this model in three behavioral experiments.

## Test one variable at a time

The problem of disambiguating the effects of multiple variables has a long history in developmental psychology and education. Starting with Piaget (Inhelder \& Piaget, 1958), many educators and psychologists have stressed the importance of controlling or isolating variables. One important procedural component of this approach is the experimental strategy of changing one variable at a time and observing its effect while holding all other variables constant. In the STEM education
literature, considerable emphasis has been placed on teaching children this controlling variables strategy (e.g., Chen \& Klahr, 1999; Kuhn \& Brannock, 1977). In fact, it even appears in national standards for science education (National Academy of Sciences, 2013). A common finding from empirical studies is that children require extensive training to acquire the CV principle (e.g., Kuhn et al., 1995; Klahr, Fay, \& Dunbar, 1993; Kuhn \& Phelps, 1982). Adults and adolescents, although more likely to use the strategy spontaneously, sometimes also have a tendency to test multiple features at once instead of testing them one-by-one (Kuhn et al., 1995). Interesting exceptions have been found in more complex tasks. For example, Bramley, Dayan, Griffiths, and Lagnado (2017) tested people's intervention strategies in completely unconstrained multivariate systems (with no distinction between potential causes and outcomes) and found that participants often focused on testing one causal relationship at a time by holding most variables at a constant value.

In sum, CV is a widely regarded epistemic principle for learning about causal systems composed of multiple variables. A key advantage of a CV strategy is that it results in unconfounded data that is easy to interpret. Empirical work suggests that acquiring the ability to use the CV principle can be challenging, but adults sometimes adopt it more in complex tasks.

## Test half or test multiple variables

Changing variables one-by-one has the benefit of isolating the effect of every variable without confounding influence of the others. It is therefore particularly helpful when one believes that many variables are causes of the outcome. However, consider the case in which a learner expects only very few, and perhaps just a single variable to have a causal relationship to the outcome, but is faced with a number of equally plausible candidate variables. In that case, an alternative strategy is to test multiple variables at once to see if any of them affect the outcome at all. For example, imagine trying to figure out which out of 20 switches in a poorly labeled basement fusebox controls the bedroom fan. An optimal strategy for finding this switch is to turn on exactly half (10) of the switches to find out which half contains the target switch and then continue halving the remaining possibilities until only one switch remains. Compared to testing switches one-by-one, this will dramatically reduce the number of trips needed for checking what effect the current switch setting has on the fan.

This Test Multiple or (more specifically) Test Half strategy has been studied by psychologists in a slightly different type of information-seeking task, often based on popular games as "Twenty questions" or "Guess who?". In these games, chil-
dren or adults have to identify a target object, person or cause among a given set by asking as few yes/no questions as possible. Here too, the optimal strategy (in terms of expected information gain, see next section) is to ask about features that apply to half the possibilities under consideration (e.g., "Is the person female?", if the hypotheses are people and half are each sex), since it can reduce the number of possibilities more rapidly than asking about specific identities directly (e.g., Navarro \& Perfors, 2011). Both children and adults have been shown to use this method successfully (Nelson, Divjak, Gudmundsdottir, Martignon, \& Meder, 2014; Ruggeri \& Lombrozo, 2015). Interestingly, any Test Multiple strategy would be considered an error from the perspective of the education literature (e.g., a student adopting this strategy might be coded as failing an STEM education assessment), because by changing many things at once it momentarily confounds the influence of individual variables.

The Test One and Test Multiple/Half strategies are typically studied in different kinds of psychological tasks. However, as the switch example from above illustrates, they can both be reasonable approaches for testing the causal impact of multiple variables. Next, we show how the effectiveness of each strategy depends on the structure of the task.

## Sparsity determines effectiveness of strategies

As the switch example shows, an important factor that determines the effectiveness of a Test One or a Test Multiple strategy is the sparsity of a causal system. We define sparsity as the proportion of causes among variables (for related definitions and discussions of the importance of sparsity for hypothesis testing, see e.g., Navarro \& Perfors, 2011; Langsford, Hendrickson, Perfors, \& Navarro, 2014). In sparse environments (e.g., when we know that only one in 20 switches controls the fan), a learner can quickly narrow in on an effective cause by trying many variables at once. In contrast, if there are (known to be) many causes, trying many things at once will tend to be uninformative as the effect will almost always be generated and little will be learned about which variable(s) were responsible. The choice of an effective testing strategy in a particular situation is thus a question of ecological rationality.

Modeling the effect of sparsity To formalize this intuition, assume that a learner is faced with a simple causal system with $N$ binary independent input variables, $I$, and a single binary outcome, o. Given the subset of input variables, $C \subseteq$ $I$ that, when active, can cause the outcome to happen, the probability of the outcome given the current setting of inputs is

$$
P(o=1)= \begin{cases}1, & \text { if } \exists c \in C(c=1) \\ 0, & \text { otherwise }\end{cases}
$$

In other words, the outcome occurs if and only if any of the input variables in $C$ are currently active.

The learner must now decide how to manipulate the input variables to best figure out which of them are causes (i.e.,


Figure 1: Effect of the number of causes, $|C|$, and the number of variables, $N$, on the number of variables tested.
which are members of $C$ ). We assume that the learner's optimal strategy lies in choosing a switch setting, $s \in S$, that maximizes the expected Information Gain with respect to the system. Information gain is a common metric for quantifying the value of information-seeking actions, including causal interventions. It is computed as the expected difference of a learner's current uncertainty over their hypotheses $H$, and their expected new uncertainty after having made an intervention on the system and observed an outcome. In this case, a learner's hypotheses $H$ are all possible compositions of the set of causes, $C$, and there are two possible outcomes ( $o=1$ or $o=0$ ). Thus a learner's expected information gain is

$$
\begin{equation*}
E I G(s \mid H)=S E(H)-\sum_{j=0}^{1} P(o=j \mid s) S E(H \mid s) \tag{1}
\end{equation*}
$$

where SE denotes the Shannon Entropy over a distribution of beliefs (Shannon \& Weaver, 1949). To investigate the impact of causal sparsity, we use this model to explore how a learner's belief about sparsity affects the optimal strategy.

Figure 1 shows model predictions for the number of variables an optimal learner should manipulate upon their first encounter with a system of variables, based on their knowledge about the number of causes, $|C|$, and the number of variables in the system, $N$, assuming a uniform belief over all remaining hypotheses. In line with the intuition outlined above, when a learner expects only one cause the model predicts a Test Half strategy. As the number of causes increases (that is, as causal sparsity decreases), the optimal number of manipulated variables decreases and quickly reaches the strategy of changing only a single variable. This relationship is modulated by the total number of variables, which increases the degree of causal sparsity and consequently the number of variables that should be manipulated.

These results show that the causal sparsity of an environment should affect a learner's strategy for manipulating binary variables to find out how they affect some outcome of interest. This means that, even in the same task, there can exist a continuum of optimal strategies with respect to the num-


Figure 2: Wooden box used in Experiment 1.
ber of variables changed, which ranges from a Test Half to a Test One method. This observation leads to the core prediction we test in this paper. We hypothesize that when learning a causal system, people will use different strategies depending on their belief about the sparsity of the system. This result would offer a further demonstration that human intervention strategies are ecologically rational, in the sense of being well matched to the environment within which they are needed.

## Experiments

We now present three experiments that investigate how knowledge about sparsity affects people's causal testing strategies. Sparsity is manipulated in two ways, both suggested by the model results shown in Figure 1. We first vary the number of causes (i.e., variables that affect the outcome) in a system (Exp. $1 \& 2$ ) and second the number of total variables available for testing (Exp. 2). We also investigate what strategies people select given no prior instructions about sparsity (Exp. 3).

## Experiment 1 - manipulating number of causes

Participants 30 participants were recruited via the subject pool of New York University's Department of Psychology. Participants were paid at a rate of $\$ 5$ per hour and could win an additional bonus of up to $\$ 3$ (see below).

Stimuli Participants were presented with the wooden box depicted in Figure 2. The box had six different switches (inputs), a yellow wheel (output) and a red activation toggle. Each switch could be turned to the left (off) or the right (on). The activation toggle controlled whether the entire box was turned on or off. Participants were randomly assigned to one of two experimental conditions. In the sparse condition, only one of the switches caused the wheel to spin, whereas the remaining five switches were broken. In the non-sparse condition, five switches caused the wheel to spin and one switch was broken. A single working switch was sufficient to activate the wheel, and the position of the broken switches had no effect whatsoever. The wheel could only be activated if the activation toggle was currently in its on-position. Otherwise, participants were told that the box was turned off. The working and broken switches were chosen randomly for each participant. At the beginning of the experiment, participants
were given six plastic tokens, each of which was worth $\$ 0.50$. Participants had to pay one token every time they wanted to turn on the box via the activation toggle (see below).

Procedure Participants were first familiarized with the components on the box. They were told about the the binary (on/off) nature of the switches, and the difference between broken and working switches. Depending on the condition, participants were then told that they had to to identify the one broken switch (non-sparse condition) or the one working switch (sparse condition). Before starting the task, participants in both conditions were shown the same two demonstration trials. First, while the activation toggle was turned off, the experimenter turned all six switches to their on position and subsequently turned on the activation toggle, causing the wheel to spin. Second, after turning the activation toggle off again, the experimenter set all switches to their off state and turned the activation toggle back on, which did not cause the wheel to spin. In the main part of the experiment, participants could repeatedly test different settings of the switches to find out which one was broken/working. On each trial, they could change the switches in any way they liked while the activation toggle was off. They could then test their chosen switch setting by turning the activation toggle on and observing the effect on the wheel. Before the start of each new trial, the activation toggle had to be turned off again.

To incentivize participants to use as few trials as possible, they had to pay one of their six plastic tokens (worth $\$ 0.50$ each) for each time they performed a test by inserting it into a coin slot on the box. Participants could test the box up to six times (hence the use of six tokens), but could stop whenever they thought they had identified the one broken/working switch. After their final test, they indicated to the experimenter which of the switches was broken/working. If their choice was correct, they could trade in any remaining tokens for their corresponding monetary value. If it was incorrect or they used up all their tokens, they received no bonus.

Results To characterize participants' trial-by-trial behavior at a strategy level, we used the following classification scheme. In the non-sparse condition, participants' strategies were classified as Test One if they turned on one switch on every trial, while leaving all other switches turned off. If a participant manipulated multiple switches or kept testing the same switch more than once, their strategy was classified as Other. In the sparse condition, participants' strategies were classified as Test One if participants turned on one new switch each trial, even when they left previously tested, but ineffective, switches turned on. This is because these past switches would have shown to be broken and therefore could not contribute confounding evidence on future tests. Participants' strategies were classified as Test Multiple if participants tested half or multiple of the switches. As a sequential strategy, Test Half does not have a meaningful definition for participants in the non-sparse group, who would al-


Figure 3: Strategy use in Experiment 1
ways encounter confounding evidence when changing multiple switches. Note that we also classified participants as Test One or Test Multiple if they had some interspersed trials with zero Information Gain (e.g., from repeating the same test twice), assuming that they were using a more noisy version of the respective strategy. ${ }^{1}$

Figure 3 shows the number of participants using each testing strategy in the two conditions. For the purposes of this figure, the classification was based on a participant's sequence of tests up to the point at which an optimal learner would have been able to correctly identify the working or broken switch (some participants made further unnecessary tests). Note that, among participants classified as "Test Multiple", everyone actually manipulated exactly half of the switches (i.e., they used the optimal strategy according to our optimal model). We kept the more general classification as Test Multiple, to stay consistent with the results presented in the next experiment. The number of participants using a Test One strategy was lower in the sparse condition (4 in 15 vs. 14 in 15 , Fisher's exact $p<0.001$ ). However, even in the sparse condition around a quarter of the participants decided to change one variable at a time.

In sum, as predicted by the optimal learner model presented above, Experiment 1 found that instructing participants to expect either a sparse (one cause) or a non-sparse (five causes) environment, had an effect on how they proceeded to manipulate a set of six variables. However, even in the sparse condition, we found that some use of the less effective Test One strategy persists. The following experiments explore possible explanations for this finding.

## Experiment 2 - manipulating number of variables

Experiment 2 explores whether increasing the amount of sparsity by adding more variables would lead to more participants to adopt a Test Half strategy. In Experiment 1, the benefit of testing multiple variables over testing variables one-byone was relatively modest. In fact, testing half of the variables in the sparse condition would save participants less than one

[^305]

Figure 4: Expected number of trials needed to find the working switch in the sparse condition, when using a Test One or Test Half strategy.
step ( $2 / 3$ of a step) on average, compared to testing variables individually (this difference translated to an average saving of $\sim \$ 0.33$ ). This may have not provided sufficient incentive for participants to realize that a Test Half strategy would be more advantageous. As discussed above, one way to amplify the sparsity manipulation is to add more variables (see Figure 1). To illustrate the effect on the expected payoff from the two strategies, Figure 4 shows the average number of tests needed to find the working switch for a learner in a sparse (one cause) environment employing either a Test One or a Test Half strategy. It shows that as the number of switches increases, so does the benefit of the Test Half strategy over the Test One strategy.

To test if people are sensitive to the degree of sparsity, Experiment 2 manipulated the number of variables (switches). Participants on Amazon Mechanical Turk completed the same task as in Experiment 1, but were presented with either $4,6,10$, or 20 switches (all manipulations were betweensubjects). As before, they were given either sparse (one switch working) or non-sparse (one broken) instructions. Although adding variables should have no effect on behavior in the non-sparse condition, we decided to keep the manipulation to ensure that adding variables does not encourage a general increase in the number of variables participants would test on each trial. By including the 6 switches condition again, this experiment also served to replicate the results from Experiment 1 with an online sample.
Participants 120 participants were recruited on Amazon Mechanical Turk. Recruitment was restricted to AMT workers within the United States aged 18 or above. Participants were paid $\$ 0.50$ for their participation, with the possibility of earning an additional bonus of up to $\$ 1$ (see below).
Stimuli The task from Experiment 1 was adapted as faithfully as possible to be run on the web with some minor changes. Instead of a wheel, the outcome of interest was a light bulb, which lit up when it was turned on, and re-


Figure 5: Strategy use in Experiment 2
mained gray otherwise. All switches were of the same kind and would turn green when on and red when off.
Procedure The experiment followed a $4 \times 2$ betweensubjects design. Participants received different versions of the task with either $4,6,10$, or 20 switches, and were given either the sparse or the non-sparse instructions. The procedure was the same as in Experiment 1. Participants received similar instructions and were also asked to perform two demonstration trials in which first all and then none of the switches were turned on, to show that the light bulb would turn on and stay off, respectively. The per-trial payment was adjusted depending on condition, such that participants had to pay either $\$ 0.25, \$ 0.16, \$ 0.1$, or $\$ 0.05$ per additional test in the $4,6,10$, or 20 switches conditions, respectively. These payments were chosen so that the total potential bonus (starting at \$1) would be zero if participants decided to test every single switch in isolation.
Results Figure 5 shows the frequency of the Test One, Test Multiple, and Other strategies by condition. In the non-sparse group a large majority of participants changed a single variable at a time, irrespective of the number of switches. In the sparse condition, however, the proportion of Test One users varied with the number of switches (Fisher's exact, $p<.05$ ), such that participants confronted with more switches (10 or 20) were less likely to test individual switches than those confronted fewer (4 or 6). This development was also accompanied by the expected increase in strategy efficiency, such that the number of trials participants saved on average in the sparse condition compared to the non-sparse condition increased from 0.15 trials in the 4 variable condition to 6.53 trials in the 20 variable condition, in line with the predictions in Figure 4.

These results provide further evidence that information about sparsity affects how people learn actively in multiplevariable settings. Again, participants in the sparse group were more likely to manipulate multiple variables at a time, whereas those with in the non-sparse group chose to manipulate variables one-by-one. Furthermore, participants in the sparse condition, in which there is only one cause, were sensitive to the total number of variables. The more switches were presented to participants (the more sparse the environment),
the more prominent was their use of a Test Multiple strategy. Nevertheless, this experiment also replicated the finding from Experiment 1 that in the absence of a strong incentive to do otherwise, people have a tendency to change single variables, rather than multiple. In fact, the web sample revealed, if anything, an even stronger tendency to use a Test One strategy in the sparse condition, particularly when the number of switches was small.

## Experiment 3 - no sparsity information

Experiments 1 and 2 suggest that testing one variable at a time might serve as a default strategy that is only overridden, to some degree, by knowledge about the number of causes. To explore this possibility further, Experiment 3 asked what strategy people would use to test a multi-variable system when they had no prior information about the number of causes to begin with. By giving participants vague instructions, we aimed to instill an approximately flat "prior" over all possible combinations of working and broken switches.

Participants 57 participants were recruited on Amazon Mechanical Turk. Recruitment was restricted to AMT workers within the United States aged 18 or above. Participants were paid $\$ 0.50$ for their participation, with the possibility of earning an additional bonus of up to $\$ 1$.

Stimuli Materials were the same as the 6-switch condition of Experiment 2. In a between-subject design participants were again randomly assigned to a switchboard that either had one broken or one working switch.

Procedure The procedure was the same as in the previous Experiment, with the exception that participants were given the same set of instructions about the number of causes in both conditions. Instead of being told to find the one broken or one working switch, they were instructed to "find out which switch(es) are working or broken". After the switch testing phase, participants were asked to indicate which switch(es) were working or broken, now being able to make multiple selections.

Results Figure 6 shows the proportion of participants that chose to turn on any possible number of switches on the very first trial. Data is collapsed over both conditions, since the initial instructions were the same and hence the first trial should not lead to different behaviors. The vast majority of participants (\%58) chose to manipulate a single switch, with only a small number (\%10) manipulating half.

This experiment verified that with no instructions about sparsity, the majority of participants chose to manipulate variables one-by-one. Note that an optimal learner initialized with a flat prior (which translates into $2^{6}$ hypotheses, given 6 switches, each with a prior probability of $\frac{1}{2^{6}}$ ) also assigns higher expected Information Gain to testing one over testing


Figure 6: Number of switches tested on the first trial of Experiment 3.
multiple variables. Therefore, the Test One "default" shown by some participants in earlier experiments, could stem from them ignoring the constrained prior that they were instructed on and instead acting as if they knew nothing about the sparsity of the system. This behavior would still be in line with the optimal learner analysis presented above.

## General Discussion

In a series of experiments, we found that, in line with an optimal learner model, people's strategies to manipulate multivariate causal systems take into account the causal sparsity of the system. In non-sparse environments (e.g., only one non-cause) the majority of participants adhered to a strategy of testing one variable at a time, in line with a "Controlling Variables" principle (Kuhn \& Brannock, 1977). When causes were sparse (e.g., only one cause) participants were more likely to manipulate multiple (often half) of the candidate variables. We also found that increasing the degree of sparsity, by increasing the total number of variables, amplified this effect on people's strategy choices.

These findings demonstrate that people adaptively change their causal experimentation strategies in response to knowledge about the environment. Our study thus offers an example of the importance of "ecological learning" that allows people to flexibly adapt their inquiry strategies to the information structure of the task (Ruggeri \& Lombrozo, 2015). This idea tallies with other recent work on causal interventions showing that people's strategy choices were made adaptively with respect to internal constraints, like cognitive load, and external factors like the match of a strategy and the task environment (Coenen, Rehder, \& Gureckis, 2015). In finding that sparsity affects behavior, the experiments above also add to recent evidence from other (spatial) information search tasks, in which hypothesis sparsity was shown to affect people's hypothesis testing strategies (Hendrickson, Navarro, \& Perfors, 2016).

Interestingly, we also found that even in sparse environments a proportion of participants chose to test variables individually, despite the fact that changing multiple variables would have been more efficient. This is somewhat surprising since prior work has often found that the Controlling Variables principle is difficult to teach and often violated even
by adults (Kuhn et al., 1995). It is thus intriguing to think about why we found such pervasive use of a CV strategy. One possibility is that a Test One strategy is less risky than changing multiple variables under a wide range of possible prior beliefs about the system. If the underlying system is not sparse, changing multiple variables can result in ambiguous evidence and often no information gain. However, changing variables one-by-one will be informative even in a sparse environment. With some degree of uncertainty about the current environment, a learner might therefore just be better off testing one variable at a time. Another contributing factor might be the that changing one variable at a time is explicitly taught in schools as a principle of scientific experimentation (National Academy of Sciences, 2013). It is interesting to consider whether this curriculum standard might actually in some cases hinder efficient experimentation by promoting a narrow focus on the idea of testing variables individually, irrespective of situation specifics.

Acknowledgments This research was supported by NSF grant BCS-1255538 and a John S. McDonnell Foundation Scholar Award to TMG.

## References

Bramley, N. R., Dayan, P., Griffiths, T. L., \& Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. Psychological Review, to appear.
Chen, Z., \& Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. Child development, 70(5), 1098-1120.
Coenen, A., Rehder, B., \& Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. Cognitive Psychology, 79, 102-133.
Hendrickson, A. T., Navarro, D. J., \& Perfors, A. (2016). Sensitivity to hypothesis size during information search. Decision, 3(1), 62 .
Inhelder, B., \& Piaget, J. (1958). The growth of logical thinking. New York: Basic Books.
Klahr, D., Fay, A. L., \& Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. Cognitive psychology, 25(1), 111-146.
Kuhn, D., \& Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. Developmental Psychology, 13(1), 9-14.
Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., \& Carver, S. M. (1995). Strategies of knowledge acquisition. Monographs of the society for research in child development, i-157.
Kuhn, D., \& Phelps, E. (1982). The development of problem-solving strategies. Advances in child development and behavior, 17, 144.

Langsford, S., Hendrickson, A., Perfors, A., \& Navarro, D. J. (2014). People are sensitive to hypothesis sparsity during category discrimination..
National Academy of Sciences. (2013). Next generation science standards: For states, by states. Washington, DC: The National Academies Press.
Navarro, D. J., \& Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. Psychological review, 118(1), 120-134.
Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., \& Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. Cognition, 130(1), 7480.

Pearl, J. (2009). Causality. Cambridge university press.
Ruggeri, A., \& Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. Cognition, 143, 203-216.
Shannon, C. E., \& Weaver, W. (1949). The mathematical theory of information.

# Path salience in motion events from verbal and visual languages 

Neil Cohn (neilcohn@visuallanguagelab.com), Vivian Wong, Kaitlin Pederson \& Ryan Taylor<br>Tilburg center for Cognition and Communication (TiCC)<br>P.O. Box 90153, 5000 LE Tilburg, The Netherlands


#### Abstract

Languages differ in the way they convey paths. S-languages conveying manner of motion directly in a main verb, while Vlanguages require a separate verb. This difference has been shown to influence the conceptualization and narration of motion events. We therefore asked: would this difference arise in the paths that people draw, particularly in visual narratives? We annotated the representations of path information (source, trajectory, goal) in a corpus of 35 comics from S- and Vlanguages. We found that panels from S-languages depicted the path of an action more often than those from V-languages, consistent with previous research on increased motion event salience for S -languages. These findings suggest that the conceptualization of paths from spoken language may influence the graphic depiction of paths.


Keywords: visual language; motion events; linguistic relativity; paths; attention

## Introduction

Expressing information about the paths of actions is a challenge to communication systems. Spoken language expresses spatial relations with symbolic units. Languages break down actions as moving from a source (starting point) to a goal (endpoint) with some sort of manner (characteristic of the path, i.e., bouncing, wiggling, sauntering, etc.) (Talmy, 1985). This can be encoded in a sentence using a satelliteframed construction, which places both manner and motion into one verb (go, run), while a "satellite" expresses the path in a preposition (out, in), as in She ran out of the room. Alternatively, a verb-framed construction includes both motion and path in a verb (Spanish: salir-"exit", entrar"enter"), and separating manner into an additional verb (corriendo-"running"), as in She exited the room running. Corpus analysis has suggested that languages differ in which of these constructions they primarily use (see Slobin, 2003 for summary). Satellite-framed "S-languages" (English, German, Dutch, Mandarin) allow a main verb to contain information about manner of motion, while verb-framed "Vlanguages" (Spanish, French, Japanese, Hebrew) express this in a separate verb.

This difference in typology thus varies the salience of paths in motion events between languages. For example, in a sentence like He ran out of house, across the street, into the bar, an S-language can easily covey one main verb, and extend the manner of the path (running) across several event segments using prepositions. In contrast, V-languages would require a separate verb for each path (He exited the house,
crossed the street, entered the bar) and would need to add a verb repeatedly to each clause to specify manner (running).

Because V-languages frame manner in a separate verb, they demand extra effort and attention in contrast to the increased salience of paths in motion events by S-languages (Slobin, 2000, 2003). This difference in salience has manifested in studies of mental imagery, translation, and narration-often elicited from reading wordless visual narratives. Narratives told by speakers of S-languages tend to create units of successive events and draw focus to the manner of motion, while those by speakers of V-language often frame the setting and environment where motion events happen, leaving both paths and their manner to inference (Slobin, 2000, 2003).

Given that much previous work examined participants' verbalized narration of wordless, drawn visual narratives, would path salience be reflected in the drawing systems that appear in visual narratives themselves? In drawings, motion needs to be converted from dynamic movement into a static depiction. Such information is implied by the postures of figures in actions, and further clarified by graphic devices, like motion lines, which overtly depict the path of an action by trailing a moving object (Cohn, 2013; McCloud, 1993). Motion lines differ in the "visual languages" used to draw visual narratives throughout the world (Cohn, 2013; McCloud, 1993) and their understanding is learned over time (Friedman \& Stevenson, 1975) and conditioned by experience with comics (Cohn \& Maher, 2015; Nakazawa, 2016). In addition, both behavioral and neurocognitive research suggests that actions are more easily understood when motion lines depict their paths than without motion lines (Cohn \& Maher, 2015; Ito, Seno, \& Yamanaka, 2010).

Given that S - or V-languages differ in how they encode properties of paths, might this salience of paths in visual depictions differ based on a drawer's spoken language? That is, might the conceptualizations from one domain (e.g., speaking) "permeate" those of another domain (e.g., drawing) as a reflection of shared conceptual resources (Cohn, 2016)?

A study by Tversky and Chow (2009) sampled panels out of one comic each from Japan, Italy, America, and China. These books crossed distinctions of Western cultures (America and Italy) and Asian cultures (China and Japan), in addition to crossing linguistic typology of S-languages


Figure 1. A path depicted by a motion line segmented into its component parts.
(Chinese and English) and V-languages (Japanese and Italian). Panels were then rated on a scale of "action" vs. "setting the scene" by American and Japanese participants. Participants rated panels from China and America (Slanguages) as more "active" than those from Japan and Italy (V-languages), but within these contrasts panels from Asian countries (China, Japan) were rated as more active than those from Western countries (America, Italy). These results provide preliminary evidence that path information encoded differently in S- and V-languages could influence the drawings of speakers of those languages. Nevertheless, this work was limited in that it included panels from only single comics per group, and did not directly examine the depiction of paths within those works, instead using "action" as a proxy for paths.

We therefore further investigated this issue by coding the component parts of paths directly in panels from various comics from around the world. Paths were analyzed for their component segments: a source (the start of the path), the goal (the endpoint of the path), and the trajectory (the path traversed). The manner is typically encoded in the trajectory. In Figure 1, the ball is bouncing, but without that middle segment, the ball would appear to move in a straight path. As in Figure 1, an image could depict all three of these path segments at once, or a panel could frame isolated segments of a path (imagine each dotted box as images on their own, or combining them for Source-Trajectory or Trajectory-Goal segments).

We coded the properties of path information in 35 comics from around the world, drawn by speakers of S-languages (English, Mandarin, German) and V-languages (Japanese, Korean, French). However, we also considered the possibility
that path information could vary as a function of the influence of comics traditions. For example, American and Japanese comics have been observed to use different types of motion lines to depict the paths of moving objects (Cohn, 2013; McCloud, 1993, 1996). Yet, motion lines were a highly borrowed element from the Japanese Visual Language into mainstream American comics during the influx of manga into America in the 1990s. In addition, Original English Language (OEL) manga are comics drawn in the "style" of Japanese manga, but are created by speakers of English (an Slanguage). This is contrasted by Korean "manhwa," which are also imitative of the visual language of Japan, but come from Asia and speakers of Korean-another V-language. Chinese manhua (including "wuxia" from Hong Kong) are also from Asia, but do not necessarily imitate the Japanese Visual Language.

Following previous work, we reasoned that representations from visual narratives produced by speakers of S-languages (and optimized for those readers) would depict more paths than those from V-languages. Specifically, they should depict more trajectories, since that path segment illustrates its manner.

## Methods

## Materials

We selected 35 books for analysis, with 5 from each of our primary groups which varied in their continent of origin

Table 1. Characteristics of comics included in our corpus study.

| Comic type | Continent | Original <br> Language | Language <br> path type | Total <br> pages | Total <br> panels | Average <br> panels/page |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| American Mainstream | America | English | S-language | 106 | 541 | 5.16 |
| OEL Manga | America | English | S-language | 137 | 768 | 5.62 |
| Chinese manhua | Asia | Mandarin | S-language | 131 | 772 | 5.92 |
| German comics | Europe | German | S-language | 136 | 772 | 5.78 |
| French bande desinée | Europe | French | V-language | 100 | 769 | 7.73 |
| Japanese manga | Asia | Japanese | V-language | 106 | 563 | 5.62 |
| Korean manhwa | Asia | Korean | V-language | 118 | 579 | 5.2 |
| TOTAL |  |  |  | $\mathbf{8 3 4}$ | $\mathbf{4 , 7 6 3}$ | $\mathbf{5 . 8 6}$ |

(America, Europe, Asia) and path language type (S-language, V-language), as listed in Table 1. A full listing of works analyzed appears in the Appendix. We attempted to retain the same general genre (action, adventure, superhero, fantasy, sci-fi) throughout all the books as best as possible. Coders analyzed either a single issue, chapter, or episode in each book, 25 pages, or 150 panels, whichever came first. This amounted to an average of 23.8 pages and 136.1 panels per book, across a total of 4,763 panels in 834 pages (see Table 1). All annotations were incorporated into the Visual Language Research Corpus (VLRC: http://www.visuallanguagelab.com/vlrc/index.html).

## Data Analysis

Trained coders independently analyzed each book panel-by-panel across the areas of analysis with $60 \%$ of books were annotated by two coders. Coders were trained through an extensive course on visual language linguistics and prior to coding scored above an $85 \%$ agreement on assessments of their coding ability.

Coders identified whether a represented path depicted its source (starting point), trajectory (midpoint and the path itself), and/or goal (endpoint). Panels could involve an isolated path segment (just source, trajectory, or goal) or multiple segments together (ex. source-trajectory, trajectorygoal). We also annotated the cues used to signal these paths, be they graphic devices like motion lines, or the postural cues of figures in motion. We recorded the total number of path segments in a given panel, and calculated the mean number of instances by dividing the sum of path segments divided by the total number of panels per book. Final analyses averaged the means for each book between coders' scores.

Our analyses looked at panels which included any path segments (e.g., a panel with a trajectory, and/or both a trajectory and goal), and those that panels depicting only an isolated path segment (source, trajectory, or goal). Path segments were analyzed using repeated-measures ANOVAs that set path segment (i.e., source, trajectory, goal) as the


Figure 2: Path segments isolated to individual panels averaged across language types. Standard error is depicted.
within-groups factor and comic type (Table 1), continent of origin (America, Asia, Europe) or language type (S- vs. Vlanguage) as the between-groups factor. Follow up analyses examined the differences of each area of analysis within and between groups.

## Results

Our initial $2 \times 3$ repeated-measures ANOVA examined the difference between path segments (source, trajectory, goal) between different language types (S- vs. V-). We found main effects of path segments and language type, as well as an interaction between them (see Table 2). These analyses were carried out both for panels with any path segment, and those with isolated path segments. The main effect of path segment arose because, across language types, trajectories were used more than goals, which were used more than sources (all $p \mathrm{~s}<.05$; see Figure 2). However, across both analyses comics originally produced by speakers of S-languages used significantly more trajectories than those by V-languages (all $F \mathrm{~s}>7.3$, all $p \mathrm{~s}<.01$ ), but neither sources nor goals differed on the basis of language types (all $p \mathrm{~s}>.134$ ).

In line with Tversky and Chow's (2009) contrasts between cultures, we also analyzed path segments collapsed across continents (America, Europe, Asia) with a $3 \times 3$ repeatedmeasures ANOVA. Consistent with our other analyses, we again found main effects of path segments, but found no significant main effects of continent or interaction between path segments and continents. This held for analyses of all path segments and isolated path segments (Table 2).


Figure 3: Path segments isolated to individual panels from various comics. Standard error is depicted.

Finally, to assess the differences between each comic type, we used a $3 \times 7$ repeated-measures ANOVA to compare path segments across all comic types. We found main effects of path segments and comic type, and an interaction between them (Table 2). As in Figure 3, in almost all types of comics, trajectories appeared more than goals, which appeared more than sources (all $p s<.05$ ). The exception to this was Japanese manga, which depicted near equal amounts of trajectories and goals. Follow up analyses showed comic types did not differ in their depiction of isolated sources or goals when isolated in panels (all $p s>.364$ ), but did differ for trajectories in both analyses (all $F \mathrm{~s}>7.3$, all $p \mathrm{~s}<.001$ ) and between goals in panels depicting any paths, $F(6,28)=2.6, \mathrm{p}<.05$. Specifically, the difference in isolated trajectories seemed to be motivated by Chinese manhua, which used more trajectories than all other comics (all $t \mathrm{~s}>4.1$, all $p \mathrm{~s}<.01$ ) except French and German comics (all $p \mathrm{~s}>.16$ ), while panels with any paths including trajectories were higher in Chinese manhua than all other comics (all $t \mathrm{~s}>3.5$, all $p \mathrm{~s}<.05$ ), except trended higher than German comics ( $p=.076$ ). In addition, isolated trajectories in Japanese manga and Korean manhwa were fewer than in French and German comics (all $t \mathrm{~s}>3.1$, all $p \mathrm{~s}$ <.05).

## Discussion

This corpus analysis examined comics from seven different types of comics from around the world to investigate whether the depiction of path information varied on the basis of culture and/or spoken language typology. Though we found that path segments differed between comics, such variation did not vary based on the comics' continent of origin.

However, they did vary based on a comics' original language type.

On the whole, the trajectories of paths-i.e., the path itself-were depicted more than the goals (endpoint) of the paths, which in turn appeared more than the sources (starting point). The prominence of goals over sources aligns with findings that the endpoints of paths are more salient than starting points in verbal language, and in perception and attention (Lakusta \& Landau, 2005; Regier, 1996, 1997). However, the fairly consistent depiction of trajectories beyond sources and goals suggests an importance for the visual depiction of paths themselves in motion events, more than their start or endpoint. This is consistent with the idea that motion lines disambiguate actions by depicting their paths (Cohn \& Maher, 2015).

In addition, we found that trajectories isolated to their own panels appeared more in the depictions of paths from visual narratives from S-languages than V-languages. These findings are also consistent with findings that the conceptualization and narration of motion events are more salient for speakers of S-languages than those of V-languages (Slobin, 2000, 2003), and with previous work (Tversky \& Chow, 2009) showing panels from S-languages (English, Chinese) are as rated more action-oriented than those from V-languages (Japan, Italy). These findings support the idea that the framing of path information in a spoken language may influence its depiction in a drawn visual language.

Despite our finding of differences between comic panels on the basis of spoken language typology, we found no significant differences based on the comics' continent of origin. This differed from Tversky and Chow's (2009) finding of a split between how "active" panels were rated in

Western and Asian comics. However, inspection of Tversky and Chow's reported data suggests that Chinese panels may have driven the effect of higher ratings of "action" compared to all others type. We too found that manhua exceeded the depiction of paths of all other types of comics, though here it may have offset differences between continents given the relational similarities otherwise between pairs of books from Asia (Korean, Japanese), America (US mainstream, OEL manga) and Europe (French, German). The difference between manhua and other types of comics-including those from other S-languages - may support the classification of Mandarin as variant from the binary split of S- and Vlanguage types (Chen \& Guo, 2009).

The absence of variation between depictions of paths on the basis of culture contrasts from findings across other aspects of structure from the visual languages used in comics. For example, previous corpus analyses have suggested differences between cultures on the basis of comic panels' attentional framing structure (Cohn, Taylor-Weiner, \& Grossman, 2012), narrative patterns across panels (Cohn, In press), semantic transitions between panels (McCloud, 1993), and visual morphology like speech balloons (Forceville, Veale, \& Feyaerts, 2010). Combined with prior findings, our results suggest that cross-cultural variation in visual narrative systems may involve a diverse number of factors including cultural specificity, visual language patterns, and possibly influence from spoken languages.

Finally, given the differences between depictions of paths on the basis of S- versus V-languages, this work hints at "permeability" between the conceptualization made in expressive domains, here between spoken languages and drawings. This initial work could thus be followed by more extensive corpus research, in addition to experimental methods further examining these preferences. For example, both behavioral and neurocognitive work has shown that comic panels containing motion lines are easier to process than those omitting such visualized paths (Cohn \& Maher, 2015; Ito et al., 2010). While comic reading expertise modulated these costs (Cohn \& Maher, 2015), our findings here might suggest cross-cultural variation in such processing. Given the greater salience of paths for S languages, would speakers of these languages be more sensitive to the absence of motion lines than speakers of Vlanguages, for whom depicted paths may be less salient? Might path information thus be a factor in translations or interpretations of comics across languages? This work can hopefully sponsor further research into the potential permeability between the conceptualizations in spoken and visual languages.

## Works Analyzed

## American Mainstream

Jenkins, Paul, Dale Keowen, \& Matt Milla. 2004. Darkness
Resurrection. Vol 4. Top Cow Comics.
Larsen, Erik. 2013. Savage Dragon: The End. Image Comics.

Love, Jeremy \& Robert Love. 2004. Fierce. Vol. 1. Dark Horse Comics
Mignola, Mike \& Guy Davis. 2005. B.P.R.D.: The Black Flame. Dark Horse Comics.
Morris, Steve. 2006. Blessed Thistle. Dark Horse Comics.

## OEL manga

Bair, Katie \& Robby Bevard. 2006. Ninja High School Hawai'i. Vol 1. Antarctica Press.
Clugston, Chynna. 2005. Blue Monday. Vol. 4. Oni Press.
Espinosa, Rod. 2001. Chronicles of the Universe. Vol. 1. Antarctica Press.
Gunstone, Kevin \& Benn Dunn. 2003. The Agents. Antarctica Press.
Reid, Christopher \& John Kantz. 2003. Legends from Darkwood. Vol. 1. Antarctica Press.

## Japanese

Kazue, Kato. 2010. Ao no Exorcist. Vol. 4, Chapter 28. Kazé. Mashima, Hiro. 2007. Fairy Tail. Vol. 4, Chapter 26. Del Rey.
Naoshi, Komi. 2012. Nisekoi. Vol 4, Chapter 32. Jump Comics.
Otaka, Shinobu. 2012. Magi: The labyrinth of magic. Viz Media.
Saito, Kenji \& Nao Akinari. 2013. Trinity Seven. Vol. 7. Chapter 34.

## Korean

Geuk-Jin, Jeon \& Jin-Hwan Park. 2010. The Breaker New Waves. Chapter 36.
Han Yu-Rang. 2007. Boy of the Female Wolf. Vol 11, Chapter 69. Samyang Publisher.
Hwa, Kim Dong. 2009. The Color of Earth. Vol 1, Chapter 1. First Second.

Jung, Jee-Yun. 2004. Kwaidan. Dark Horse Comics.
Wann. 2006. 9 Faces of love. Vol 1. Netcomics.

## Chinese

Cha, Louis \& Wing Shing Ma. 2002. Heaven Sword and Dragon Sabre. Vol. 1. Comics One.
Ma, Wing Shing. 2002. Storm Riders. Vol. 1. Comics One
Wong, Tony. 2002. Mega Dragon \& Tiger Future Kung Fu Action. Vol. 5. Comics One.
Seto, Tony \& Ying-Hsiang Lin. 2002. Saint Legend: The Prelude. Comics One.
Yan, Win, King Tung, Bryce Gunkel, \& Calvin Chai. 2004. The King of Fighters. Vol. 3. DGN Productions.

## French

Dieter, Viviane Nicaise, \& Dina Kathelyn. 2004. La Vie En Rose. Glénat.
Galandron, Laurent \& Viviane Nicaise. 2010. Le Cahier à Fleurs. Cycle I.
Godard, Christian \& Various. 2003. Une Folie Très Ordinaire. Glénat.

Lapière, Denis, Pierre-Paul Render, \& Mathieu Reynès. 2011. Alter Ego: Fouad. Dupuis.

Recht, Robin \& Jean Bastide. 2012. Notre Dame. Glénat.

## German

Fil, Bei. 2001. Larry Potter.
Greulich, Jonas. Moga Mobo. Vol. 2. Kostenlos.
Gronle, Thomas. Moga Mobo. Vol. 3. Kostenlos.
Mawil, Markus. 2004. Die Band. Repordukt.
Sobottke, Bela. 2009. König Kobra. Gringo Comics.

## References

Chen, L., \& Guo, J. (2009). Motion events in Chinese novels: Evidence for an equipollently-framed language. Journal of Pragmatics, 41(9), 1749-1766. doi:http://dx.doi.org/10.1016/j.pragma.2008.10.015
Cohn, N. (2013). The visual language of comics: Introduction to the structure and cognition of sequential images. London, UK: Bloomsbury.
Cohn, N. (2016). Linguistic Relativity and Conceptual Permeability in Visual Narratives: New Distinctions in the Relationship between Language(s) and Thought. In N. Cohn (Ed.), The Visual Narrative Reader (pp. 315-340). London: Bloomsbury.
Cohn, N. (In press). Structural complexity in visual narratives: Theory, brains, and cross-cultural diversity. In M. Grishakova \& M. Poulaki (Eds.), Narrative Complexity and Media: Experiential and Cognitive Interfaces. Lincoln: University of Nebraska Press.
Cohn, N., \& Maher, S. (2015). The notion of the motion: The neurocognition of motion lines in visual narratives. Brain Research, 1601, 73-84. doi:10.1016/j.brainres.2015.01.018
Cohn, N., Taylor-Weiner, A., \& Grossman, S. (2012). Framing Attention in Japanese and American Comics: Cross-cultural Differences in Attentional Structure. Frontiers in Psychology - Cultural Psychology, 3, 1-12. doi:10.3389/fpsyg. 2012.00349
Forceville, C., Veale, T., \& Feyaerts, K. (2010). Balloonics: The Visuals of Balloons in Comics. In J. Goggin \& D. Hassler-Forest (Eds.), The Rise and Reason of Comics and Graphic Literature: Critical Essays on the Form (pp. 5673). Jefferson: McFarland \& Company, Inc.

Friedman, S. L., \& Stevenson, M. B. (1975). Developmental Changes in the Understanding of Implied Motion in Twodimensional Pictures. Child Development, 46, 773-778.
Ito, H., Seno, T., \& Yamanaka, M. (2010). Motion impressions enhanced by converging motion lines. Perception, 39(11), 1555-1561.
Lakusta, L., \& Landau, B. (2005). Starting at the end: the importance of goals in spatial language. Cognition, 96, 133.

McCloud, S. (1993). Understanding Comics: The Invisible Art. New York, NY: Harper Collins.
McCloud, S. (1996). Understanding Manga. Wizard Magazine, 56, 44-48.

Nakazawa, J. (2016). Manga literacy and manga comprehension in Japanese children. In N. Cohn (Ed.), The Visual Narrative Reader (pp. 157-184). London: Bloomsbury.
Regier, T. (1996). The human semantic potential: Spatial language and constrained connectionism. Cambridge, MA: MIT Press.
Regier, T. (1997). Constraints on the learning of spatial terms: A computational investigation. In R. Goldstone, P. Schyns, \& D. Medin (Eds.), Psychology of learning and motivation: Mechanisms of perceptual learning (Vol. 36, pp. 171-217). San Diego, CA: Academic Press.
Slobin, D. I. (2000). Verbalized events: A dynamic approach to linguistic relativity and determinism. In S. Niemeier \& R. Dirven (Eds.), Evidence for linguistic relativity (pp. 107-138). Amsterdam: John Benjamins.
Slobin, D. I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. In D. Gentner \& S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought (pp. 157-192). Cambridge, MA: MIT Press.
Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), Language Typology and Syntactic Description: Vol. 3. Grammatical categories and the lexicon (pp. 36-149). Cambridge: Cambridge University Press.
Tversky, B., \& Chow, T. (2009, 11/21/09). Comics: Language and Culture Affect Action in Depictions. Paper presented at the Psychonomics, Boston, MA.

# Using Prior Data to Inform Initial Performance Predictions of Individual Students 

Michael G. Collins (michael.collins.74@ctr@us.af.mil)<br>Kevin A. Gluck (kevin.gluck@us.af.mil)<br>Matthew Walsh (mmw118@gmail.com)<br>Michael Krusmark (michael.krusmark@us.af.mil)

Air Force Research Laboratory, 2620 Q Street, Building 852
Wright-Patterson Air Force Base, OH 45433


#### Abstract

The predictive performance equation (PPE) is a mathematical model of learning and retention that uses regularities seen in human learning to predict future performance. Previous research (Collins, Gluck, Walsh Krusmark \& Gunzelmann,, 2016) found that prior data could be used to inform PPE's free parameters when generating predictions of a group's aggregate performance, allowing for more accurate initial performance predictions. Here we investigate an extension of this methodology to predict performance of individuals, rather than aggregate samples. This paper documents the results of that investigation, which is on the critical path to the use of this cognitive technology in education and training.


Keywords: Mathematical model; Performance predictions; Skill learning; Parameter generalization; Educational data mining, Individual predictions

## Introduction

It is typical in training and education for instructors to have little to no information about the people who are about to begin the curriculum. Rather, individuals must complete some portion of the curriculum before for their knowledge can be assessed. This assessment period can lead to an increase in the overall amount of time that training and education takes, and can lead to individuals practicing skills that have already mastered (Beck \& Chang, 2007). Ideally, instructors could be able to estimate the future performance of both the incoming cohort of students as a whole in addition to the specific individuals based on the past performance of those who learned the same curriculum. This would allow instructors to better adjust a given curriculum to fit the needs of the cohort and of specific students.

In cognitive science, models of learning and retention have been developed to account for particular regularities in human learning such as the power law of learning (Newell \& Rosenbloom, 1981) and power law of decay (Rubin \& Wenzel, 1996), and the spacing effect (Bahrick, Bahrick, Bahrick, \& Bahrick, 1993 Although many of these models were created based on basic laboratory phenomena, they can also be used to generate predictions of future human performance (Anderson \& Schunn, 2000; Jastrzembski, Gluck, \& Gunzelmann, 2006; Mozer, Pashler, Cepeda, Lindsey, \& Vul, 2009; Pavlik \& Anderson, 2008; Raaijmakers, 2003). These models hold promise in training and education to increase mastery and /or decrease instruction time.

## The Predictive Performance Equation

The model discussed in this paper is the Predictive Performance Equation (PPE; Walsh et al, submitted). PPE is a mathematical model of human learning and retention that can generate performance predictions on declarative (know-what) and procedural (know-how) tasks. Prior research has validated PPE across a variety of different laboratory tasks (Walsh et al., submitted) as well as complex human performance data from F-16 fighter pilot training research (Jastrzembski et al., 2006) and education and training data (Collins, Gluck, \& Jastrzembski, 2015).

PPE represents the effects of three factors on knowledge acquisition and retention: recency of practice, frequency of practice, and the distribution of practice over time (i.e., spacing). The first factor, recency ( $T_{n}$ ), captures the amount of elapsed time since training began. $\mathrm{T}_{\mathrm{n}}$ is calculated as a weighted sum of the elapsed time since each of each previous training opportunities $\left(t_{i}\right)$ (Equation 1). The weight $\left(w_{i}\right)$ applied to the amount of time that has passed since a particular event decreases exponentially with time (Equation 2). Although in principle a free parameter, prior model exploration has found that the exponent, $x$, can be set to 0.6 , which we do in the analyses presented here.

$$
\begin{align*}
& T=\sum_{i=1}^{n} w_{i} * t_{i}  \tag{1}\\
& w_{i}=-t_{i}^{-x} \sum_{j=1}^{n} \frac{1}{t_{j}^{-x}} \tag{2}
\end{align*}
$$

The second factor, frequency $\left(\mathrm{N}_{\mathrm{n}}\right)$, represents the number of times that a particular knowledge or skill has been rehearsed. These two factors, elapsed time and frequency of practice, have a multiplicative effect on activation $\left(M_{n}\right)$, which is the strength of a particular memory or skill (Equation 3). Amount of practice is scaled by the learning rate $c$, which is fixed to 0.1 . As the amount of practice increases, activation rises at a decreasing rate, producing the power law of learning. The second term, comprised of $T$ and $d$, captures the effects of the power law of decay. The decay rate, d , captures spacing effects (Equation 4).

$$
\begin{gather*}
M_{n}=N^{c} * T^{-d}  \tag{3}\\
d_{n}=b+m *\left(\frac{1}{n-1} * \sum_{j=1}^{n-1} \frac{1}{\log \left(\operatorname{lag}_{j}+e\right)}\right)
\end{gather*}
$$

The precise effect of spacing on performance is determined by the summation term within the decay parameter. When lags between successive training opportunities $\left(\right.$ lag $\left._{j}\right)$ are
short, the summation term in Eq. 4 approaches 1 and decay increases, leading to a greater amount of forgetting. When the lags between training opportunities are long, and the summation term approaches 0 , the decay term decreases, leading to less forgetting over time. The decay rate equation includes a decay intercept parameter (b) and a decay slope parameter $(m)$. The activation value from Eq. 3 is scaled to performance through a logistic function (Equation 5). The function contains two additional free parameters controlling its slope $(s)$ and the intercept $(\tau)$.

$$
\begin{equation*}
P_{n}=\frac{1}{1+\exp \left(\frac{\tau-M_{n}}{s}\right)} \tag{5}
\end{equation*}
$$

In summary, PPE has four free parameters (i.e., $b, m, \mathrm{~s}, \tau$ ). These parameters can be calibrated based on existing performance data. Once a set of best fitting parameters have been found, PPE can use these parameters to predict future performance.

## Motivation

Reliable and valid parameter estimates for PPE cannot be found with PPE when calibrating to fewer than three training opportunities. There are two reasons for this. First, when fewer than three data points are available, multiple combinations of free parameter values that can fit the available training data equally well. This makes it difficult to determine which set of parameter values should be used to generate out-of-sample predictions (Beck \& Cheng 2007). Second, when calibrating PPE to so little data, PPE will likely fit to both the performance of the individual as well as to noise in the data (Geman, Bienenstock, \& Doursat, 1992). This overfitting, in turn, will reduce the accuracy of out-of-sample predictions. The combination of these two factors are likely to lead to inaccurate and uncertain out-of-sample performance predictions. To overcome this limitation, Collins et al. (2016) developed a method for using prior data (i.e., records of performance data collected from previous classes) to inform a subset of PPE's free parameters (prior predictions), under cases where there were not enough data points for accurate calibration. By using prior data to inform a subset of PPE's free parameters, PPE fits the available training data with a constrained parameter set. In circumstances where there is little training data, this increases PPE's prediction accuracy for early performance events.

This prior-informed prediction method was based on work from the Educational Data Mining (EDM) literature. EDM research applies data mining and statistical learning methodologies to educational data to improve student learning outcomes (Romero, Ventura, \& Baker, 2010). EDM methods are primarily data driven, meaning they require large amounts of data to develop predictions within a specific domain (Webb, Pazzani, \& Billsus, 2001). In contrast PPE is primarily theory driven, meaning that its predictions are based on mechanisms that account for general characteristics of human learning and retention.

The development of PPE's prior-informed prediction method balances the data-driven and theory driven approach of these two methods.

Although Collins et al. (2016) found that prior data could be used to generate predictions of the aggregate performance of multiple students attempting a single skill, their results did not indicate how accurate the predictions are at an individual student level of analysis. Using prior data to predict the initial performance at a finer level of aggregation is more difficult for two reasons. First, the performance of a single individual is characterized by greater variability, as compared to learning curves aggregating across the performance of multiple students, making performance of a single student more difficult to predict. Second, students are likely to learn skills at different rates, meaning that best fitting parameters for an aggregate learning curve may not generalize to account for the performance of a specific student attempting a particular skill.

In spite of these additional complexities when predicting the performance of individual students, educational data mining research has shown that prior data can be used to inform valid model parameter estimates for models used to account for the performance of individual students on single skills (Cen, Koedinger \& Junker 2007; Beck \& Chang 2007; Ritter et al., 2009). These findings suggest that prior data can serve as a useful tool that can be used to inform predictions of individual students and not just aggregate samples. In summary, we sought to expand our previous research by examining the extent to which our method for predicting early performance of groups of students generalizes to the individual student level of analysis. To evaluate the prior-informed method, we compare it against predictions to PPE's standard non-prior predictions during an individual student's first 4 attempts on a new skill.

## Method

The data used in this report were obtained from Learnlab.com's DataShop (Koedinger, Baker, Cunningham, Skogsholm, Leber, \& Stamper 2010), which is an online educational data repository for student $\log$ data. DataShop contains a collection of publicly available datasets from different math, science, and English classroom and tutoring studies. The data used in our analyses, consisted of $\log$ files of performance metrics of students completing their homework for an introductory physics class during six different semesters. Students used the ANDES tutoring system to complete their homework (VanLehn et al, 2005) at the United States Naval Academy (USNA). We chose these datasets because they contain the largest collection of data from multiple semesters collected from the same domain currently available on DataShop, allowing us to better investigate the utility of using prior data to inform PPE's performance predictions.

A single semester's worth of data on DataShop is called a dataset, which is composed of a record of the
performance of individuals who attempted to solve problems in a specific domain within a specific period of time. Each dataset contains the record of all of the students' actions across the curriculum's content. A curriculum is made up of problems, defined as "a task [attempted by] a student usually involving several steps." An example of a problem would be calculating the difference in velocity between trains A and B. Successfully solving a problem involves completing a series of steps, which are "an observable part of a solution to a problem", such as finding the velocity of train A. We choose to examine the performance of students while completing particular steps for two reasons. First, steps were the smallest level of resolution of data available on Datashop. Second, each step isolates a particular knowledge component. Because learning occurs at the level of individual knowledge components (Anderson \& Schunn, 2000), comparing analogous steps across problems is the proper way to observe the change in performance over time.

## Prediction Procedure

We systematically selected one of the six datasets as the prediction sample, and used the remaining five datasets as prior data to inform predictions for an individual on a particular step. Then the performance data of a single student on a particular skill was selctecd, from the prediction sample. All of the students from the prior data who also attempted the same skill were selected (relevant sample) and used to inform PPE's predictions. Due to the fact that the data collected from the ANDES tutoring system are data from homework assignments, the students' first exposure to the curriculum was during class and was not their first attempt on a particular step within the tutoring system. For this reason, we assumed a six-hour lag between class and when a student began to complete their homework. This assumption of a lag between class and home time allowed for a better estimation of PPE's model time as calculated from PPE's time variables (Eq. 1and 4). For the relevant sample to be able to inform a prediction, the average performance and model time variables across each participant during each event was calculated. Based on aggregate performance and model time computed from the relevant samples, PPE model parameters were estimated, and then used to make individualized predictions of a student's performance on a particular skill on the $2^{\text {nd }}, 3^{\text {rd }}$, and $4^{\text {th }}$ event.

For the analysis in this paper, we used PPE to generate predictions for two metrics of the students' performance: time to complete a particular step (seconds) and the number of incorrect attempts made by a student during a particular event. To generate a prior prediction, PPE first calibrated to the performance (i.e., completion time in seconds or number of incorrect attempts) of the first two events from the aggregate performance of the relevant sample. This yields a set of best fitting parameters values. The best fitting $b\left(\mathrm{~b}_{\text {prior }}\right)$ and $m$ ( $\mathrm{m}_{\text {prior }}$ ) parameters are then generalized to inform PPE's prior informed prediction of
an individual student's performance on the $2^{\text {nd }}$ event given their performance on the $1^{\text {st }}$ event. This is done by setting PPE's $b$ and $m$ free parameters to the $\mathrm{b}_{\text {prior }}$ and $\mathrm{m}_{\text {prior }}$ values and fitting PPE's remaining two free parameters $s$ and $\tau$ to the student's performance during the first event. After PPE is fitted to the student's performance on the $1^{\text {st }}$ event, the model is used to generate a prediction of the student's performance on the second event. This procedure was then repeated to generate predictions of the $3^{\text {rd }}$ and $4^{\text {th }}$ event, by increasing the number of events that PPE is calibrated to with the prior sample and the predicted individual before generating a performance prediction of the next event.

In addition to generating prior predictions, we used PPE to generate predictions of each student's performance on the 2nd, 3rd, and 4th events without using data from past participants. This involved fitting the model with the sparse, individual-specific data, and using the model to predict performance for the following event.

Across all of the six datasets collected from Datashop, a total of 10,499 predictions were made across 430 students and 161 individual steps across the $2^{\text {nd }}, 3^{\text {rd }}$, and $4^{\text {th }}$ performance event.

## Results

To examine the accuracy of PPE's prior and non-prior predictions the average model predictions from the $2^{\text {nd }}, 3^{\text {rd }}$, and $4^{\text {th }}$ events were compared to the average observations from students whose performance was predicted (Figure 1).

In addition to the looking at the average performance, the students' performance and PPE's predictions were separated in to two groups (i.e., canonical and noncanonical learning). The students in the canonical learning groups were students whose performance either improved or remained the same over the four observed learning events (Figure 2-A, 2-C). Students in the non-canonical learning group were students whose performance decreased during at least one of the four learning events (Figure 2-B, 2-D). The students' performance was separated into canonical and non-canonical learning groups, due to the fact the variability in the students' performance effects PPE's performance predictions. Additionally, we wanted to observe to test if PPE's could account for the two types of learning profiles.

## Completion Time

As seen in Figure 1, when predicting a student's performance on the $2^{\text {nd }}$ event, given their performance on the $1^{\text {st }}$ event, there is a significant difference between the mean completion time between PPE's prior ( $M=45.50$, $S D=85.50$ ) and non-prior ( $M=192.199, S D=196.78$; $t(10497)=90.932, p<.01)$ predictions compared to the students' average completion time $(M=37.85, S D=$ 70.43). Examining the root mean squared deviation $(R M S D)$ between PPE's prior $(R M S D=98.49)$ and nonprior predictions $(R M S D=250.18)$, we see that PPE's prior-informed predictions were more accurate than nonprior predictions. These results show that informing PPE's


Figure 1. The average performance metric, completion time (seconds) (left plot) and number of incorrect attempts (right plot) on the $2^{\text {nd }}, 3^{\text {rd }}$, and $4^{\text {th }}$ event, across human data (solid black line), prior informed predictions (dashed blue line), and non-prior informed predictions (dashed red line).
predictions using prior data can improve prediction accuracy when prediction performance of the $2^{\text {nd }}$ event

When predicting the students' performance on the $3^{\text {rd }}$ event, given their performance on the first 2 events, again a significant difference between PPE's prior ( $M=36.79$, $S D=81.76$ ) and non-prior ( $M=66.85, S D=143.21$; $t(10497)=22.78, p<.01)$ predictions is observed, compared to the students' average performance ( $M=$ $73.96, S D=179.11$ ). As was seen when predicting the students' average performance on the $2^{\text {nd }}$ event, a similar pattern is seen when predicting the $3^{\text {rd }}$ event. A lower RMSD was found between the students' average performance and PPE's prior $(R M S D=99.66)$ compared to non-prior predictions $(R M S D=151.80)$.

Finally, when predicting the students' performance on the $4^{\text {th }}$ event, given their performance on the previous 3 events, again a difference between the PPE's prior ( $M=$ $35.76, S D=73.11$ ) and non-prior ( $M=71.63, S D=172.86$; $\mathrm{t}(10497)=22.63, p<.01)$ predictions are observed, compared to the students' average performance ( $M=$ 33.56, $\mathrm{SD}=68.26$ ). Again PPE's prior informed predictions had a lower RMSD $(R M S D=92.13)$ compared to the non-prior predictions $(R M S D=181.53)$ when predicting the students' performance on the $4^{\text {th }}$ event.

## Correct Attempts: Canonical and Non-Canonical Learning Profile

Separating the students' performances into those who displayed canonical and non-canonical learning profiles, reveals two different sets of completion times. The performance of students who displayed a canonical learning profile was found to be monotonically improve over the course of the three events (Figure 2-A). Students who the non-canonical learning profile, on average


Figure 2. The average performance metric, completion time (A, B) and number of incorrect attempts (C, D) for both students who fit the canonical (A, C) and non-canonical (B D) learning profile, for both the human data (solid black line), non-prior predictions (dashed red line) and prior predictions (dashed blue line) on the $2^{\text {nd }}, 3^{\text {rd }}$, and $4^{\text {th }}$ event.
displayed non-monotonic improvement in their performance over the four events (Figure 2-B). Additionally, it is seen that the accuracy of PPE's prior and non-prior predictions varied based on the performance of the students' learning profile. When predicting the performance of students' who showed a canonical learning profile, PPE's prior and non-prior predictions became more accurate as PPE was calibrated to additional events before generating a prediction, during the $2^{\text {nd }}$ (Prior: $R M S D=$ 117.97; Non-prior: $R M S D=310.42$ ), $3{ }^{\text {rd }}$ (Prior: $R M S D=$ 82.90; Non-prior: $R M S D=132.78)$, and $4{ }^{\text {th }}$ (Prior: $R M S D=$ 56.31;Non-prior: $R M S D=91.52$ ) event (Figure 2-A). However, PPE's accuracy decreased when it was calibrated to each additional event when predicting performance of students' whose performance was found to have a noncanonical learning profile. When predicting the performance of students' who showed a non-canonical learning profile, PPE's prior and non-prior prediction accuracy decreased as PPE calibrated to additional events, during the $2^{\text {nd }}$ (Prior: $R M S D=41.26$; Non-prior: $R M S D=$ 95.67), $3^{\text {rd }}$, (Prior: $R M S D=101.62$; Non-prior: $R M S D=$ 154.06 ) and $4{ }^{\text {th }}$ (Prior: $R M S D=95.80$; Non-prior: $R M S D$ $=190.80$ ) event (Figure 2-B). Although, PPE's prediction accuracy varied based on the students' learning profile, PPE's prior performance predictions were more accurate than PPE's non-prior predictions.

## Number of Incorrect Attempts

Examining the average students' number of incorrect attempts on the $2^{\text {nd }}$ event given a students' previous performance on the first event (Figure 2), a large difference is observed in the predicted average number of incorrect attempts in PPE's prior ( $M=.47, S D=1.25$ ) and non-prior
$(M=2.19, S D=4.58 ; \mathrm{t}(10496)=38.97, p<.01)$ predictions, compared to the students' average number of incorrect attempts ( $M=.38, S D=1.09$ ). Looking at the RMSD between PPE's predictions and the students' performance, PPE's prior $(R M S D=1.51)$ predictions had a lower RMSD than PPE's non-prior informed predictions ( $R M S D=4.87$ ).

When predicting the students' average number of incorrect attempts $(M=.39, S D=1.63)$ on the $3^{\text {rd }}$ event, again a significant difference between PPE's prior ( $M=$ .49, $S D=1.63$ ) and non-prior predictions is observed ( $M=$ $1.20, S D=3.58 ; t(10496)=20.92, p<.01)$. However, unlike when predicting performance on the $2^{\text {nd }}$ event, the RMSD of PPE's prior informed predictions increased $(R M S D=2.14)$. While as well as PPE's non-prior ( $R M S D$ $=3.91$ ) decreased slightly.

Finally, when predicting the students' number of incorrect attempts on their $4^{\text {th }}$ event, given their performance on the previous three events, a similar pattern of predictions is seen. A significant difference was observed between PPE's prior $(M=.52, S D=2.13)$ and non-prior predictions $(M=1.24, S D=3.79 ; t(10496)=$ 22.62, $p<.01$ ), compared to the students' average performance was observed $(M=.39, S D=1.63)$. Additionally, the RMSD between the PPE's prior (RMSD $=2.22)$ and non-prior predictions $($ RMSD $=3.87)$ were not seen to improve. However, the PPE's prior informed predictions were lower than PPE's non-prior informed predictions.

## Incorrect Attempts: Canonical and Non-Canonical Learning Profile

Separating the students' performance into those who displayed canonical and non-canonical learning profiles, two different sets of the students' number of incorrect attempts are seen. From students who displayed a canonical learning profile, number of incorrect responses decreased over the course of the four learning events (Figure 2-C). Conversely, students who displayed a noncanonical learning profile on average displayed a nonmonotonic performance over the four events (Figure 2-D). The accuracy of PPE's prior and non-prior predictions varied based on the type of learning displayed by the students. When predicting the performance of students who showed a canonical learning profile, PPE's prior and nonprior predictions became more accurate when PPE calibrated to additional events, during the $2^{\text {nd }}$ (Prior: $R M S D$ $=1.04$, Non-Prior: $R M S D=4.69$ ), $3^{\text {rd }}$, (Prior: $R M S D=.68$ Non-Prior: $R M S D=2.27$ ) and $4^{\text {th }}$ (Prior: $R M S D=.56$ NonPrior: $R M S D=1.29$ ) event (Figure 2-C). However, PPE's accuracy decreased when it calibrated to additional events of students with a non-canonical learning profile. When predicting the performance of students' who showed a noncanonical learning profile, PPE's prior and non-prior predictions became less accurate as PPE calibrated to additional events, during the $2^{\text {nd }}$ (Prior: $R M S D=2.03$; NonPrior: $R M S D=5.14$ ), $3^{\text {rd }}$ (Prior: $R M S D=3.30$; Non-Prior:
$R M S D=5.57$ ), and $4^{\text {th }}$ (Prior: $R M S D=3.34$; Non-Prior: $R M S D=5.96$ ) (Figure $2-\mathrm{D}$ ). Although, prediction accuracy varied based on the students' average performance based on the learning profile of the student, PPE's prior performance predictions were more accurate than PPE's non-prior predictions.

## Discussion

The primary goal of this paper was to describe our assessment of the accuracy of PPE predictions of performance in the tutoring data available on DataShop, both with and without the use of informative priors. We find evidence that incorporating prior data into PPE's predictions at a lower (individual student) level of aggregation, slightly improves prediction accuracy, depending on the performance measure, the event being predicted, and the student's learning profile.

When predicting a student's completion time on the $2^{\text {nd }}$, $3^{\text {rd }}$, and $4^{\text {th }}$ event, we found that PPE's prior informed predictions were more accurate than PPE's individualized predictions. Additionally, we found that PPE's predictions varied based on the student's learning profile. When predicting the performance of students' who were found to have a canonical learning profile, the accuracy of PPE's increased as PPE was calibrated to additional events. However, the opposite results were observed when predicting the performance of students' who were found to have a non-canonical learning profile. Here it was observed that PPE's ability to predict performance depended on the variability of the students performance history in their performance. When variability in a student's performance history was low and improved regularly (i.e., canonical learning profile), PPE was better able to predict their future learning. When variability was high and a student's performance history showed both improvement and forgetting (i.e, non-canonical learning), the increased uncertainty in performance hindered the PPE's predictions from accurately predicting future performance. Although, the benefit of using priors was observed in PPE's predictions in each of these cases.

These results are partially consistent with results from Collins et al. (2016), where we found an initial benefit of using prior predictions to generate initial performance predictions of the $2^{\text {nd }}$ event, as was found when predicting the student's completion time. Without information from prior data, PPE's parameters must be estimated with sparse data from the student's prior performance during the first event. Because the model is under constrained in this case, the parameter estimates are likely unreliable.

Additionally, when predicting the average completion time and the number of incorrect attempts, a benefit of using a priors was found. When predicting a student's future performance, PPE is able to utilize information from other students who have previously performed the skill before, allowing for a better estimate of the student's future performance will be. These findings are in line with our
previous findings that PPE's prior predictions benefit PPE's predictions beyond the $2^{\text {nd }}$ event.

## Conclusion

The benefits of using prior data are not new to cognitive science. However, within the context of the PPE line of investigation, little previous research has been conducted on how prior data can be used to inform predictions, especially within the context of early performance predictions of individual students. In summary, we find evidence that our previously proposed method of incorporating information from prior data into PPE's free parameters (Collins et al. 2016), can add some benefit to prediction accuracy when attempting to predict the performance of individual students on particular skills. The results suggest that prior data is a useful source of information about the performance of individual students when generating predictions with PPE. Future work should attempt to incorporate information from prior data to generate initial performance predictions in order to decrease overall training or education time.

## Acknowledgements

The authors would also like to thank the Oak Ridge Institute for Science and Education (ORISE) who supported this research by appointing Michael Collins, to the Student Research Participant Program at the U.S. Air Force Research Laboratory (USAFRL), 711th Human Performance Wing, Human Effectiveness Directorate, Warfighter Readiness Research Division, Cognitive Models and Agents Branch administered by the ORISE.

## References

Anderson, J. R., \& Schunn, C. (2000). Implications of the ACT-R learning theory: No magic bullets. Advances in instructional psychology, Educational design and cognitive science, 5. 1-33.
Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., \& Bahrick, P. E. (1993) Maintenance of foreign language vocabulary and the spacing effect. Psychological Science, 4, 316-321.
Beck, J. E., \& Chang, K. M. (2007, July). Identifiability: A fundamental problem of student modeling. In International Conference on User Modeling, 137-146, Springer Berlin Heidelberg.
Collins, M. G, Gluck, K. A., \& Jastrzembski, T. S. (2015). Datashopping for performance predictions. Proceedings of the Foundations of Augmented Cognition, Los Angeles, California, (pp. 12-23).
Collins, M.G., Gluck, K.A., Walsh, M., Krusmark, M., Gunzelmann, G., (2016, July) Using prior data to inform model parameters in the predictive performance equation. In Proceedings of the 38th Annual Conference of the Cognitive Science Society. Philadelphia, PA
Cen, H., Koedinger, K., \& Junker, B. (2006, June). Learning factors analysis-a general method for cognitive model evaluation and improvement. In International

Conference on Intelligent Tutoring Systems (pp. 164175). Springer Berlin Heidelberg.

Corbett, A. T., \& Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4(4), 253-278.
Geman, S., Bienenstock, E., \& Doursat, R. (1992). Neural networks and the bias'variance dilemma. Neural Computation, 4, 1-58.
Jastrzembski , T. S. , Gluck , K. A. , \& Gunzelmann, G. ( 2006 ). Knowledge tracing and prediction of future trainee performance . In: The proceedings of the interservice/industry training, simulation, and education conference. Orlando, FL :National Training Systems Association (pp. 1498-1508).
Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R., \& Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, \& A. Culotta (Eds.), Advances in Neural Information Processing Systems 22 (pp. 1321-1329). La Jolla, CA: NIPS Foundation.
Newell, A., \& Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. Cognitive Skills and Their Acquisition, 1, 1-55
Pavlik, P. I., \& Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology: Applied, 14(2), 101.
Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. Cognitive Science, 27, 431-452.
Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R. C., \& Towle, B. (2009). Reducing the Knowledge Tracing Space. International Working Group on Educational Data Mining.
Romero, C., Ventura, S., Pechenizkiy, M., \& Baker, R. S. (Eds.). (2010). Handbook of educational data mining. CRC Press.
Rubin, D. C., \& Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. Psychological Review, 103, 734-760.
VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons Learned. International Journal of Artificial Intelligence and Education, 15 (3).

# Determinants of judgments of explanatory power: Credibility, Generalizability, and Causal Framing 

Matteo Colombo (m.colombo@uvt.nl)<br>Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Warandelaan 2, 5037 AB Tilburg<br>Tilburg, The Netherlands<br>Leandra Bucher (bucher@uni-wuppertal.de)<br>Department of Biological and General Psychology, Bergische Universität Wuppertal, Max-Horkheimer-Str. 20, 42119 Wuppertal, Germany<br>Jan Sprenger (j.sprenger@uvt.nl)<br>Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Warandelaan 2, 5037 AB Tilburg<br>Tilburg, The Netherlands


#### Abstract

This study investigates how judgments of explanatory power are affected by (i) the prior credibility of a potential explanation, (ii) the causal framing used to describe the explanation, and (iii) the generalizability of the explanation. We found that the prior credibility of a causal explanation plays a central role in explanatory reasoning: first, because of the presence of strong main effects on judgments of explanatory power, and second, because of the gate-keeping role prior credibility has for other factors. Highly credible explanations were not susceptible to causal framing effects. Instead, highly credible hypotheses were sensitive to the generalizability of an explanation. While these results yield a more nuanced understanding of the determinants of judgments of explanatory power, they also illuminate the close relationship between prior beliefs and explanatory power and the relationship between abductive and probabilistic reasoning.


Keywords: Explanation; Prior credibility; Causal framing; Generalizability; Abduction

## Introduction

Explanation is a central concept in human psychology. It supports a wide array of cognitive functions, including reasoning, categorization, learning, inference, and decisionmaking (Lombrozo, 2006; Keil \& Wilson, 2000; Keil, 2006). When presented with an explanation of why a certain event occurred, how a certain mechanism works, or why people behave the way they do, both scientists and laypeople have strong intuitions about what counts as a good explanation. Yet, more than sixty years after philosophers of science began to elucidate the nature of explanation (Hempel \& Oppenheim, 1948; Hempel, 1965; Salmon, 1989), the determinants of judgments of explanatory power remain unclear.

In this paper, we present three experiments on factors that may affect judgments of explanatory power. Motivated by a large body of theoretical results in epistemology and philosophy of science, as well as by a growing amount of empirical work in cognitive psychology (for respective surveys see Woodward, 2014; Lombrozo, 2012), we
examined how judgments of explanatory power are affected by (i) the prior credibility of a potential explanation, (ii) the causal framing used to describe the explanation, and (iii) the generalizability of the explanation.

First we hypothesized that the prior credibility of a causal explanation predicts judgments of explanatory power. Thus, throughout all three experiments, we manipulated the prior credibility of different explanations, and examined the effects of this manipulation on explanatory judgments.

Our focus on the prior credibility of causal explanation was motivated by the fact that most philosophical and psychological analyses of explanatory power agree that powerful explanations provide information about credible causal relationships. Credible causal information facilitates the manipulation and control of nature (Pearl, 2000; Woodward, 2003; Strevens, 2008) and plays distinctive roles in human psychology (Lombrozo, 2011; Sloman \& Lagnado, 2015). For example, credible causal information guides categorization (Carey, 1985; Murphy \& Medin, 1985; Lombrozo, 2009), supports inductive inference and learning (Holyoak \& Cheng, 2011; Legare \& Lombrozo, 2014; Walker et al. 2014), and calibrates metacognitive strategies involved in problem-solving (Chi et al., 1994; Aleven \& Koedinger, 2002).

Our second, related hypothesis was that presenting an explanatory hypothesis in causal terms predicts judgments of its explanatory power. Thus, we wanted to find out whether people's explanatory judgments are sensitive to causal framing effects.

The importance of this issue should be clear in the light of the fact that magazines and newspapers very often, even when it's not warranted, describe scientific explanations in terms of causal language (e.g., 'Processed meat causes cancer' or 'Economic recession leads to xenophobic violence') with the aim of capturing readers' attention and boosting their sense of understanding (Entmann 1993; Scheufele \& Scheufele, 2010). By combining prior credibility and causal framing as predictors of judgments of explanatory power, Experiment 1 and 2 examined the
impact of causality on the explanatory power of scientific hypotheses.

With Experiment 3, we tested the hypothesis that the generalizability (or scope) of a hypothesis determines its explanatory power. While the generalizability of scientific results is an obvious epistemic virtue that figures in the evidential assessments made by scientists, its relation to explanatory power is less clear. Previous psychological findings about the role of generalizability in explanatory reasoning are mixed. Read \& Marcus-Newhall (1993) found that generalizability predicts explanatory judgments. Preston \& Epley (2005) showed that hypotheses that apply to a wide range of observations are judged as more valuable. However, these studies involved no uncertainty about whether or not a causal effect was actually observed (cf., Khemlani, Sussman, \& Oppenheimer, 2011). So, whether or not generalizability is a robust determinant of explanatory judgment remains unclear.

In summary, bringing together different strands of research from philosophy and psychology, our study asks: How do the credibility, causal framing, and generalizability of a hypothesis influence judgments of explanatory power?

The pattern of our experimental findings supports the hypothesis that the prior credibility of a causal explanation plays a central role in explanatory reasoning: first, because of the presence of strong main effects on judgments of explanatory power, and second, because of the gate-keeping role it has for other factors. Highly credible explanations were not susceptible to causal framing effects. Instead, highly credible hypotheses were sensitive to the effects of factors which are usually considered relevant from a normative point of view like the generalizability of an explanation.

## Overview of the experiments and pre-tests

To warrant the validity of the experimental material, we conducted a series of pre-studies, where participants evaluated different levels of causal framing, credibility, and generalizability. Materials which corresponded to high, low, and neutral levels of these three factors were implemented in the vignettes of our three experiments, either as independent variables or as control variables.

Material evaluation and main experiments were both conducted online on Amazon Mechanical Turk, utilizing the Qualtrics Survey Software. We only allowed workers with an approval rate $>95 \%$ and with a number of HITs approved > 5000 to submit responses. Instructions and material were presented in English.

## Causal Framing

A sample of $\mathrm{N}=44$ participants (mean age 30.5 years, $S D=$ $7.3,28$ male) from America ( $\mathrm{n}=27$ ) and other countries rated eight brief statements, expressing relations between X and Y of the type " X co-occurs with Y "; "X is associated with Y", and so on. Participants judged how strongly they agreed or disagreed that a certain statement expressed a causal relation between X and Y. Judgments were collected
on a 7-point scale with options: "I strongly disagree" (-3), "I disagree", "I slightly disagree", "I neither agree nor disagree" (0), "I slightly agree", "I agree", "I strongly agree" (3). Based on participants' ratings, we selected three types of statements for our main experiments: statements with a neutral causal framing ("X co-occurs with Y"), with a weak causal framing (" X is associated with Y "), and with a strong causal framing ("X leads to Y " and " X causes Y ").

## Prior Credibility

We identified the prior credibility of different hypotheses by asking a new sample of $\mathrm{N}=42$ participants (mean age 30.7 years, $S D=7.5,16$ male) from America $(\mathrm{n}=29)$ and other countries to rate a list of 24 statements. Participants judged how strongly they disagreed or agreed that a certain hypothesis was credible. For all hypotheses, we used the phrasing "... co-occurs with..." to avoid the influence of causal framing. Based on participants' ratings, we selected four statements to use in our main experiments: two were highly credible, two were highly incredible (Table 1).

Table 1: The four hypotheses rated as least credible and as most credible.

| Credibility | Hypothesis |
| :---: | :--- |
| Low | Eating pizza co-occurs with immunity to flu. |
| Low | Drinking apple juice co-occurs with <br> anorexia. |
| High | Well-being co-occurs with frequent smiling. |
| High | Consuming anabolic steroids co-occurs with <br> physical strength. |

## Generalizability

This pre-study included two questionnaires, which were administered to two different groups of participants. One questionnaire presented descriptions of the samples used in scientific studies, which varied with regard to the number of people involved. The other questionnaire presented sample descriptions that varied with regard to the type of people in the sample.

Forty-two participants (mean age 33.5 years, $S D=10.8$, 27 male) from America ( $\mathrm{n}=38$ ) and other countries were presented with a list of six statements about a sample of a certain number of participants, e.g. "The study investigates five people"; "The study investigates 500 people".We found that the perceived generalizability of a study increased with the number of people in the sample of the study.

A new group of $\mathrm{N}=41$ participants (mean age 33.0 years, $S D=9.7,26$ male $)$ from America $(\mathrm{n}=36)$ and other countries was presented with a list of nine statements about samples of particular types of people, e.g. "The study investigates a group of people who sit in a park"; "The study investigates a group of people who work at a university". However, focusing on the number instead of the type of people in the sample allowed for a neater distinction between narrowly and widely generalizable results. Therefore, we characterized generalizability as a function of the number of participants in the main vignettes of the experiment.

## Vignettes of the Main Experiment

All experiments were performed using a $2 \times 2$ (withinsubject) design with explanatory power as dependent variable and prior credibility of the hypothesis being one of the independent variables. The other independent variable was either causal framing, or generalizability.

Participants were presented with four short reports about fictitious research studies. Two of these reports involved highly credible hypotheses, the other two reports involved incredible hypotheses. Two reports showed a high level of the other independent variable, while the other two reports showed a low level of that variable.

Each vignette in our experiments followed the same format as in this sample vignette.

## Consuming anabolic steroids leads to physical strength

A recent study by university researchers investigated the link between consuming anabolic steroids and physical strength. The researchers studied 240 persons. The level of physical strength was higher among participants who regularly consumed anabolic steroids than among the participants who did not regularly consume anabolic steroids. Family health history, age, and sex, which were controlled by the researchers, could not explain these results. The study therefore supports the hypothesis that consuming anabolic steroids leads to physical strength.

In all experiments, we varied the level of prior credibility of a hypothesis. In Experiment 1 and 2, we also varied the causal framing and interchanged "leads to" with "causes" and "is associated with", while we kept generalizability at its control. In Experiment 3, we varied the sample size (=generalizability) and controlled for causal framing by using the predicate "co-occurs with" in the headline and the conclusion. Participants were asked to rate our dependent variable: the explanatory power of the stated hypothesis for the results of the study.

## Experiment 1 and 2. Credibility x Causal Framing

## Participants, Design, and Material

Two-hundred-three participants (mean age 34.7 years, $S D=$ 10.5; 121 male) from America ( $\mathrm{n}=130$ ), India $(\mathrm{n}=67$ ) and other countries completed Experiment 1 for a small monetary payment. A new sample of two-hundred-eight participants (mean age 34.56 years, $S D=9.97$; 124 male) from America $(\mathrm{n}=154)$, India $(\mathrm{n}=43)$, and other countries completed Experiment 2 for a small monetary payment.

In both experiments, participants were presented with four short reports about fictitious research studies along the lines of the above vignette. Across vignettes, we manipulated the causal framing of the relationship between hypothesis and evidence as well as the choice of the hypothesis (credible vs. incredible). Generalizability was controlled for by setting it to its medium value (i.e., 240 participants). Two of the four reports involved highly
credible hypotheses, the other two involved incredible hypotheses. Similarly, two of these reports used weak causal framing (Experiment 1 and 2: "X is associated with Y") while the other two used strong causal framing (Experiment 1: "X leads to Y", Experiment 2: "X causes Y"). In other words, Experiment 1 used implicit causal language and Experiment 2 used explicit causal language, while the experiments were identical with respect to design, materials, and procedure.
To account for the possible influence of the content of a particular report, we counterbalanced the allocation of weak and strong causal framing conditions to the credibility conditions across the items, and created two versions of the experiments. The order of reports was individually randomized for each participant.
Participants judged each report in terms of the explanatory power of the hypothesis it described. Specifically, participants considered the statement: "The researchers' hypothesis explains the results of the study", and expressed their judgments on a 7 -point scale with the extremes (-3) "I strongly disagree" and (3) "I strongly agree", and the center pole (0) "I neither disagree nor agree".

## Analysis and Results

Separate two-way ANOVAs were calculated with the factors Credibility (low, high) and Causal Framing (weak, strong). ANOVA of Experiment 1 (implicit causal language) revealed a main effect of Credibility, $F(1,202)=$ 84.5; $p<.001 ; \eta_{\text {part }}^{2}=0.30$. There was no main effect of Causal Framing ( $p=.37$ ), and no interaction ( $p=.08$ ). Pairwise comparisons showed that incredible hypotheses were rated significantly lower than credible hypotheses, independently of the value of Causal Framing (incredible hypotheses: $M=0.26 ; S E M=0.10$; credible hypotheses: $M$ $=1.14 ; S E M=0.09 ; t$-test: $t(202)=-9.2 ; p<0.001 ; d=$ 0.67 ). The results of Experiment 1 therefore indicate that the prior credibility of a hypothesis was a strong predictor of judgments of explanatory power (Figure 1). Instead, framing a hypothesis with implicit causal language did not have effects on explanatory judgment.


Figure 1: Explanatory power ratings for credible and incredible statements in Experiment 1. Error bars show standard errors of the mean, and are expressed numerically, in parentheses next to the mean value.

ANOVA of Experiment 2 (explicit causal language) revealed main effects of Credibility $(F(1,207)=286.9 ; p$ <.001; $\left.\eta_{\text {part }}^{2}=0.58\right)$ and Causal Framing, $F(1,207)=31.0$; $p<.001 ; \eta_{\text {part }}{ }^{2}=0.13$, as well as a significant interaction Credibility x Causal Framing, $F(1,207)=37.6 ; p<.001$; $\eta_{\text {part }}{ }^{2}=0.15$. Figure 2 shows the effect sizes and the interaction between both factors as well as the relevant descriptives.


Figure 2: How explanatory power ratings vary with regard to Credibility and Causal Framing (Experiment 2). Error bars show standard errors of the mean and are expressed numerically, in parentheses next to the mean value.

The results of Experiment 2 confirm that the prior credibility of a hypothesis is a strong predictor of judgments of the hypothesis' explanatory power. Incredible hypotheses received negative explanatory power ratings, credible hypotheses receive positive ratings. The results also showed that explicit causal framing can increase ratings of explanatory power, but only for incredible hypotheses. While this effect may lead explanatory judgment astray, in most practical cases of explanatory reasoning, people are interested in the explanatory power of hypotheses which they find, at least to a certain extent, credible. As Figure 2 shows, there was no effect of causal framing on explanatory power in this important case.
This pattern of results confirms that the prior credibility of a hypothesis plays a gate-keeping-role in explanatory reasoning: only credible causal hypotheses qualify as explanatorily valuable. By contrast, implicit or explicit causal framing plays a small to negligible role in influencing judgments of explanatory power.

## Experiment 3: Credibility x Generalizability

## Participants, Design, and Material

Two-hundred-seven participants (mean age 33.4 years, $S D=$ 9.1; 123 male) from America ( $\mathrm{n}=156$ ), India $(\mathrm{n}=37)$ and other countries completed Experiment 3 for a small monetary payment.

The experiment resembled Experiment 1 and 2. Four vignettes, each of which included a headline and five sentences, presented credible and incredible hypotheses. The relation between hypothesis and evidence was expressed by using the causally neutral wording "X co-
occurs with $\mathrm{Y}^{\prime \prime}$. The critical manipulation concerned the sample descriptions used in the vignettes, which expressed either narrow or wide generalizability of the study's result. For narrowly generalizable results, the second sentence of a report indicated that the sample of the study encompassed around 5 people (e.g. "The researchers studied 6 people"). For widely generalizable results, the sample included about 10,000 people (wide generalizability condition, e.g. "The researchers studied 9891 people").
To control for the possible influence of the content of a particular report, we counterbalanced the allocation of narrow and wide generalizability conditions to the credibility conditions across the items, and created two versions of the experiments. The order in which reports were presented to the participants was individually randomized for each participant.

Participants were asked to carefully assess each report with regard to Explanatory Power. Participants’ ratings were collected on 7-point scales, with the extreme poles (-3) "I strongly disagree" and (3) "I strongly agree", and the center pole (0) "I neither disagree nor agree".

## Analysis and Results

The ratings were analyzed with a two-way ANOVA with the factors Credibility (low, high) and Generalizability (narrow, wide). ANOVA revealed significant main effects of Credibility, $F(1,206)=83.830 ; p<.001 ; \eta_{\text {part }}{ }^{2}=0.289$; and Generalizability, $F(1,206)=29.593 ; p<.001 ; \eta_{\text {part }}{ }^{2}=$ 0.126 , and no interaction Credibility x Generalizability ( $p=$ .085, n.s.).

As with Experiment 1 and 2, credible hypotheses achieved significantly higher ratings than incredible hypotheses (incredible hypotheses: $M=-0.01$; $\mathrm{SEM}=0.10$; credible hypotheses: $M=0.95$; SEM $=0.08$; $t$-test: $\mathrm{t}(206)=$ $-9.2 ; \mathrm{p}<.001 ; \mathrm{d}=0.72$ ). Furthermore, reports with wide generalizability achieved significantly higher ratings compared to reports with narrow generalizability (narrow: $M=0.21$; SEM $=0.10$; credible hypotheses: $M=0.73$; SEM $=0.08$; $t$-test: $\mathrm{t}(206)=-5.4 ; \mathrm{p}<.001 ; \mathrm{d}=0.40)$. Figures 3 and 4 show the main effects for both variables.


Figure 3: Explanatory power ratings as a function of Credibility. Error bars show standard errors and are also expressed numerically, next to the mean value.


Figure 4: Explanatory power ratings as a function of Generalizability. Error bars show standard errors and are also expressed numerically, next to the mean value.

## Discussion

We examined the impact of three factors---prior credibility, causal framing, and generalizability---on judgments of explanatory power. In a series of three experiments, we varied both the subjective credibility of an explanation and one of the other factors: causal framing and generalizability. In Experiments 1 and 2 we found that the impact of causal language on judgments of explanatory power was small to negligible. Experiment 3 showed that generalizable explanations with wider scope positively affected judgments of explanatory power.

Across all experiments, we found that the prior subjective credibility of a hypothesis had a striking effect on how participants assessed explanatory power. In particular, the credibility of an explanatory hypothesis had an important gate-keeping function: the impact of generalizability on explanatory power was more significant when credibility was high. On the other hand, the high credibility of a hypothesis controlled for the potentially misleading effect of causal framing on explanatory judgment.

This pattern of findings is consistent with existing psychological research demonstrating that people resist endorsing explanatory hypotheses that appear unnatural and unintuitive, given their background common-sense understanding of the physical and of the social world (Bloom \& Weisberg 2007). Our findings are also consistent with the idea that stable background personal ideologies (often referred to as "worldview") can reliably predict whether people are likely to reject well-confirmed scientific hypotheses (Lewandowsky et al., 2013; Colombo, Bucher, \& Inbar, 2016).

So, scientific hypotheses that are inconsistent with our prior, background, common-sense beliefs or in tension with personal ideologies are likely to be judged as implausible, and may not be endorsed as good explanations unless they are supported by extra-ordinary evidence gathered by some trustworthy source. On the other hand, for hypotheses that fit our prior, background belief or ideology, we often focus on information that, if the candidate explanatory hypothesis is true, would boost its goodness (Klayman \& Ha 1987).

This kind of psychological process of biased evidence evaluation and retention might have led participants to give
the highest ratings of explanatory power, across different experiments, when, in addition to a credible hypothesis, the report was widely generalizable. In comparison, the impact of causal framing was negligible in these cases. This result confirms that a good explanation has to be credible and widely generalizable, and that credible, widely generalizable explanations are not subject to misleading causal framing effects.

The interplay we observed between prior credibility and explanatory power is also relevant to understanding the relationship between abductive and probabilistic reasoning. Highly credible hypotheses were sensitive to the effects of factors which are usually considered explanatory virtues like the generalizability of an explanation.

In abductive reasoning, explanatory considerations are taken to boost the credibility of a target hypothesis while inducing a sense of understanding (Lipton, 2004). Previous psychological studies investigated the effect on people's assessments of explanatory power of factors like simplicity (Lombrozo, 2007; Bonawitz \& Lombrozo, 2012) and coherence (Koslowski et al. 2008). Our results advance this body of literature by suggesting that the generalizability of a hypothesis will boost the acceptability of the hypothesis, when the hypothesis has a high prior subjective credibility.

High prior credibility may also insulate an explanation from causal framing effects, which may produce a deceptive sense of understanding leading to erroneous explanatory judgments (Rozenblit \& Keil, 2002; Trout, 2002).

Overall, our experiments show that explanatory power is a complex concept, affected by considerations of prior credibility of a (causal) hypothesis, and its generalizability. These factors also figure prominently in (normative) philosophical theories of explanation. For instance, the D-N model (Hempel, 1965) stresses the generality of the proposed explanation, and the causal-mechanical account (Woodward, 2003) requires a credible causal mechanism.

On the other hand, the multitude of relevant factors in explanatory judgment explains why it has been difficult to come up with a theory of abductive inference that is both normatively compelling and descriptively accurate: after all, it is difficult to fit diverse determinants of explanatory judgment into a single unifying framework. In that spirit, we hope that our results will promote an interdisciplinary conversation between empirical evidence and philosophical theorizing, and about the "prospects for a naturalized philosophy of explanation" in particular (Lombrozo 2011, 549; Schupbach, 2015; Colombo, 2016).

## Acknowledgements

The project was financially supported by the European Research Council (ERC) under Starting Investigator Grant \#640638 (J.S.) and by the Deutsche Forschungsgemeinschaft (DFG) in the Priority Program 1516 "New Frameworks of Rationality" (M.C., J.S.).

## References

Aleven, V. A., \& Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive science, 26(2), 147-179.
Bloom, P., \& Weisberg, D. S. (2007). Childhood origins of adult resistance to science. science, 316(5827), 996-997.
Bonawitz, E.B. \& Lombrozo, T. (2012). Occam's rattle: children's use of simplicity and probability to constrain inference. Developmental Psychology, 48, 1156-1164.
Carey, S. (1985). Conceptual Change in Childhood. Plenum, Cambridge, MA
Chi, M.T.H., de Leeuw, N., Chiu, M., \& Lavancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.
Colombo, M. (2016). Experimental Philosophy of Explanation Rising. The case for a plurality of concepts of explanation. Cognitive Science. DOI: 10.1111/cogs. 12340
Colombo, M., Postma, M., \& Sprenger, J. (2016). Explanatory Judgment, Probability, and Abductive Inference. In Papafragou, A., Grodner, D., Mirman, D., \& Trueswell, J.C. (Eds.). Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 432437) Austin, TX: Cognitive Science Society.

Colombo, M., Bucher, L., \& Inbar, Y. (2016). Explanatory Judgment, Moral Offense, and Value-Free Science. An Empirical Study. The Review of Philosophy and Psychology, 7, 743-763.
Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of communication, 43(4), 5158.

Hempel, C.G..(1965). Aspects of Scientific Explanation. In: Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. The Free Press, pp. 331-496.
Hempel, C.G. \& Oppenheim, P. (1948). Studies in the Logic of Explanation. Philosophy of Science 15: 135-75.
Holyoak, K. J., \& Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. Annual review of psychology, 62, 135-163.
Keil F.C., \& Wilson R.A.. (2000). Explanation and Cognition. Cambridge, MA: MIT Press.
Keil, F. C. (2006). Explanation and understanding. Annual Review of Psychology, 57, 227-254.
Khemlani, S. S., Sussman, A. B., \& Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. Memory \& Cognition, 39(3), 527-535.
Klayman, J., \& Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94(2), 211-228.
Koslowski, B., Marasia, J., Chelenza, M., \& Dublin, R., (2008) Information Becomes Evidence when an Explanation Can Incorporate it into a Causal Framework. Cognitive Development, 23: 472-487.
Legare, C. H. \& Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. Journal of Experimental Child Psychology, 126, 198-212

Lipton, P. (2004). Inference to the Best Explanation (second edition). London: Routledge.
Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak \& R. G. Morrison (eds.): Oxford Handbook of Thinking and Reasoning, 260-276. Oxford, UK: Oxford University.
Lombrozo, T. (2011). The instrumental value of explanations. Philosophy Compass, 6, 539-551.
Lombrozo, T. (2009). Explanation and Categorization: How 'Why?' Informs 'What?' Cognition, 110, 248-253.
Lombrozo, T. (2007). Simplicity and probability in causal explanation. Cognitive Psychology, 55, 232-257.
Lombrozo, T. (2006). The structure and function of explanations. Trends in Cognitive Sciences, 10, 464-470.
Murphy, G. L., \& Medin, D. L. (1985). The role of theories in conceptual coherence. Psychological review, 92(3), 289.

Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge: Cambridge University Press.
Preston, J. \& Epley, N. (2005). Explanations Versus Applications: The Explanatory Power of Valuable Beliefs. Psychological Science 10, 826-832.
Rozenblit, L., \& Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. Cognitive Science 26, 521-562.
Salmon, W. (1989). Four Decades of Scientific Explanation, Minneapolis: University of Minnesota Press.
Scheufele, B. T., \& Scheufele, D. A. (2010). Of spreading activation, applicability, and schemas. Doing News Framing Analysis: Empirical and Theoretical Perspectives, New York, Routledge, 110-134.
Schupbach, J. N. (2015). Experimental Explication. Philosophy and Phenomenological Research. doi: 10.1111/phpr. 12207

Strevens, M. (2008) Depth: An Account of Scientific Explanation. Cambridge, MA: Harvard University Press.
Trout, J. D. (2002). Scientific Explanation And The Sense Of Understanding. Philosophy of Science, 69(2), 212-233.
Walker, C.M., Lombrozo, T., Legare, C., \& Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. Cognition, 133, 343357.

Woodward, J. (2014). Scientific Explanation. In E.N. Zalta (Ed.) The Stanford Enclycopedia of Philosophy. URL = [https://plato.stanford.edu/entries/scientific-explanation/](https://plato.stanford.edu/entries/scientific-explanation/)

# PACKER: An Exemplar Model of Category Generation 

Nolan Conaway (nconaway@wisc.edu)<br>Joseph L. Austerweil (austerweil@wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street<br>Madison, WI 53706 USA


#### Abstract

Generating new concepts is an intriguing yet understudied topic in cognitive science. In this paper, we present a novel exemplar model of category generation: PACKER (Producing Alike and Contrasting Knowledge using Exemplar Representations). PACKER's core design assumptions are (1) categories are represented as exemplars in a multidimensional psychological space, (2) generated items should be similar to exemplars of the same category, and (3) generated categories should be dissimilar to existing categories. A behavioral study reveals strong effects of contrast- and target-class similarity. These effects are novel empirical phenomena, which are directly predicted by the PACKER model but are not explained by existing formal approaches.


Keywords: Categorization, exemplar models, category generation, creative cognition, computational modeling.

## Introduction

The creation of new concepts and ideas is among the most interesting - yet infrequently studied - capabilities of human cognition. This paper focuses on one topic within the broader field of creative cognition: category generation. Foundational work on this topic (e.g., Smith, Ward, \& Schumacher, 1993; Ward, 1994, 1995; Ward, Patterson, Sifonis, Dodds, \& Saunders, 2002) has focused on the role of prior knowledge in generating novel concepts. A core phenomenon is that people generate categories with similar distributional properties as existing categories. For example, Ward (1994) asked participants to generate species of plants and animals that might exist on other planets. Generation was strongly constrained by prior knowledge of Earth species: People generated species with the same features as those found on Earth (e.g., eyes, legs, wings) and possessing the same feature correlations observed on Earth (e.g., feathers co-occur with wings).

Recent work has proposed and tested formal models to explain these observations. Jern and Kemp (2013) trained participants on experimenter-defined categories composed of exemplars within an artificial three-dimensional domain. After a short training phase, participants were asked to generate exemplars from a new category. Participants were provided with a set of scales to adjust the feature values of each generated stimulus, and were given unlimited time to create each example. As in the classic Ward (1994) experiment, Jern and Kemp (2013) found that generated categories possessed the same feature variance and correlations as the experimenterdefined categories in the domain.

Jern and Kemp (2013) tested several different computational models on their data. Most relevant to the present investigation, they tested a 'copy-and-tweak' model that generates items by copying and changing previous observations, and a hierarchical Bayesian model that uses the structure of
known categories to infer the structure of new ones. They found that the hierarchical Bayesian model provided the strongest account of the behavioral results.

In this paper, we introduce a novel exemplar-based approach to category generation, PACKER (Producing Alike and Contrasting Knowledge using Exemplar Representations), which creates categories by balancing two constraints: (1) new categories should be different from known categories (minimizing between-class similarity), and (2) new categories should be internally coherent (maximizing withinclass similarity). As such, PACKER is a significant departure from previous accounts of generation - rather than proposing that people create categories by abstracting and re-using knowledge of related categories, PACKER first considers how the generated category should differ from related categories. Further, it does so using the well-studied mechanics of exemplar representations and therefore possesses a rich connection to the wider body of research on category learning.

In the sections below, we formally describe the PACKER model and explore its predictions in a behavioral experiment. We compare its performance to copy-and-tweak and hierarchical Bayesian models by examining their fits to aggregate results and individual differences.

## PACKER: An Exemplar Model

The PACKER model is an extension of the Generalized Context Model of category learning (GCM; Nosofsky, 1984). It assumes that each category is encoded by a set of exemplars within a $k$-dimensional psychological space, and that generation is constrained by both similarity to members of the target category (the category in which a stimulus is being generated) as well as similarity to members of other categories.

As in the GCM, the similarity between two examples, $s\left(x_{i}, x_{j}\right)$, is an inverse-exponential function of distance:

$$
\begin{equation*}
s\left(x_{i}, x_{j}\right)=\exp \left\{-c \sum_{k}\left|x_{i k}-x_{j k}\right| w_{k}\right\} \tag{1}
\end{equation*}
$$

where $w_{k}$ is the attention weighting of dimension $k\left(w_{k} \geq 0\right.$ and $\sum_{k} w_{k}=1$ ), accounting for the relative importance of each dimension in similarity calculations, and $c(c>0)$ is a specificity parameter controlling the spread of exemplar generalization. For simplicity, our simulations will use uniform attention weights, except when otherwise noted.

To generate a new example, the model considers both the similarity to examples from contrast categories as well as the similarity to examples (if any exist) in the target category.


Figure 1: PACKER generation of a category ' B ' example, following exposure to one member of category ' A ' and category ' B '. The panels differ in how the trade-off between within- and between-category similarity is managed (via the $\gamma$ parameter).

The aggregated similarity $a$ between generation candidate $y$ and stored exemplars $x$ is:

$$
\begin{equation*}
a(y, x)=\sum_{j} f\left(x_{j}\right) s\left(y, x_{j}\right) \tag{2}
\end{equation*}
$$

where $f\left(x_{j}\right)$ is a function specifying the extent to which each exemplar contributes to the generation. A negative value for $f\left(x_{j}\right)$ produces a 'repelling' effect (items are less likely to be generated nearby $x_{j}$ ), and a positive value produces an 'attracting' effect (items are more likely to be generated nearby $x_{j}$ ). When $f\left(x_{j}\right)=0$, the exemplar does not contribute to generation.

PACKER sets $f\left(x_{j}\right)$ depending on exemplar $j$ 's category membership: $f\left(x_{j}\right)=\gamma$ if $x_{j}$ is a member of the target category, and $f\left(x_{j}\right)=\gamma-1$ if $x_{j}$ is a member of a contrast category. $\gamma$ is thus a free parameter $(0 \leq \gamma \leq 1)$ controlling the trade-off between within- and between-category similarity. PACKER's core proposal is that new categories should be different from existing categories, and same-category exemplars should be similar to one another. This is realized when $\gamma$ assumes an intermediate value: For example, when $\gamma=0.5, f\left(x_{j}\right)=0.5$ for members of the target category and $f\left(x_{j}\right)=-0.5$ for members of other categories; thus, the model is likely to generate items that are similar to members of the target category but are not similar to members of other categories. However, more extreme values can be used to produce different behavior, see Figure 1.

The probability that a given candidate $y$ will be generated is evaluated using an Exponentiated Luce (1977) choice rule. Candidates with greater values of $a$ are more likely to be generated than candidates with smaller values:

$$
\begin{equation*}
p(y)=\frac{\exp (\theta \cdot a(y, x))}{\sum_{i} \exp \left(\theta \cdot a\left(y_{i}, x\right)\right)} \tag{3}
\end{equation*}
$$

where $\theta(\theta \geq 0)$ controls response determinism.

## Summary

The proposed PACKER model suggests people generate categories by minimizing between-category similarity and maximizing within-category similarity. The underlying processes assumed by PACKER are highly similar to those in the GCM.

The main difference is that PACKER aggregates positiveand negative-valued similarities, rather than only aggregating positive-valued similarities. In later sections, we will explore the unique predictions yielded by these design principles. First, however, we contrast PACKER with other category generation models.

## Previous Accounts of Category Generation

Previous models of category generation focus on capturing the tendency for people to produce new categories that have similar distributional properties to existing categories. To the best of our knowledge, Jern and Kemp (2013) were the first to evaluate computational models of generation. Based on their work, we describe two alternative models: a formalization of the Path of Least Resistance hypothesis (later termed copy-and-tweak, see Jern \& Kemp, 2013), and the hierarchical sampling hypothesis (Jern \& Kemp, 2013).

## Copy-and-Tweak

The copy-and-tweak model, based broadly on the earlier Path of Least Resistance view (Ward, 1994, 1995), proposes that participants generate categories by retrieving an observation of the target class from memory, and then tweaking it to make something new. Jern and Kemp (2013) interpreted this proposal in terms of an exemplar model using the GCM (Nosofsky, 1984). Formally, their model is equivalent to PACKER with $\gamma=1$ (see Figure 1). In this case, $f\left(x_{j}\right)=1$ for members of the target category and $f\left(x_{j}\right)=0$ for members of other categories; thus, the model considers only target-class similarity, and when no members of the target class are known, the model generates items at random.

In our work we provide simulations from the copy-andtweak account, realized as a variant of PACKER with a fixed $\gamma$ parameter. Formalizing a model family where PACKER and copy-and-tweak are different parameterizations within the same framework is useful because comparison between the models provides a test of the explanatory value of the contrast mechanism: The account provided by copy-and-tweak will only equal that of PACKER if the contrast mechanism does not offer an advantage (i.e., if $\gamma<1$ significantly improves model fits).

## Hierarchical Sampling

Based on several results inconsistent with the copy-and-tweak account, Jern and Kemp (2013) advocated a hierarchical Bayesian model. Exemplars of each category were generated from a multivariate Normal distribution over the dimensions of stimulus space. The mean of each category was independently generated, but the covariance matrix (encoding feature variances and correlations) was generated from a common prior distribution. New categories are produced by generating a new mean (uniform over stimulus space) and covariance matrix from the common prior distribution. Because the shared prior distribution's parameters were unobserved, a hierarchical Bayesian model uses information from the previous categories (their feature variances and correlations) to generate the covariance matrix of the new category.

Each category's exemplars are assumed to be a multivariate Normal distribution with parameters $(\mu, \Sigma)$. Each category's covariance matrix is assumed to be inverse-Wishart distributed with parameters $\left(v, \kappa\right.$, and $\left.\Sigma_{D}\right) .{ }^{1} \quad \Sigma_{D}$ is the covariance matrix shared between categories. We assume the shared covariance matrix $\Sigma_{D}$ is generated from a Wishart distribution (for conjugacy) with parameters $v_{0}, \kappa_{0}$, and $\Sigma_{0}$. We set $v_{0}=4$, and $\Sigma_{0}=\lambda \mathbf{I}$, where $\lambda$ is a free parameter controlling the expected variance of dimensions (dimensions of the shared covariance matrix are expected to be uncorrelated) and $\mathbf{I}$ is the identity matrix.

To simplify the model predictions, we used maximum a posteriori (MAP) estimates for the hidden parameters and then generated new categories based on those estimates. Due to conjugacy, the MAP estimate for the shared covariance matrix $\Sigma_{D}=\Sigma_{0}+\sum_{c} C_{c}$, where $C_{c}$ is the empirical covariance matrix of category $c$. The MAP estimate of the covariance matrix for the target category $B$ is

$$
\begin{equation*}
\Sigma_{B}=\left[\Sigma_{D} v+C_{B}+\frac{\kappa n_{B}}{\kappa+n_{B}}\left(\overline{x_{B}}-\mu_{B}\right)\left(\overline{x_{B}}-\mu_{B}\right)^{T}\right]\left(v+n_{B}\right)^{-1} \tag{4}
\end{equation*}
$$

where $v(v>k-1)$ is an additional free parameter (from the Inverse-Wishart prior on $\Sigma_{B}$ ) weighting the importance of $\Sigma_{D}$. When the target category has no members (i.e., $n_{B}=0$ ), items are generated at random.

Generated exemplars are drawn from a multivariate Normal distribution specified by $\left(\mu_{B}, \Sigma_{B}\right)$. Thus, $p(y)$ is

$$
\begin{equation*}
p(y)=\frac{\exp \left(\theta \cdot \operatorname{Normal}\left(y ; \mu_{B}, \Sigma_{B}\right)\right)}{\sum_{i} \exp \left(\theta \cdot \operatorname{Normal}\left(y_{i} ; \mu_{B}, \Sigma_{B}\right)\right)} \tag{5}
\end{equation*}
$$

where $\theta$ is a response determinism parameter and $\operatorname{Normal}(y ; \mu, \Sigma)$ denotes a multivariate Normal density evaluated at $y$.

## Behavioral Experiment

The copy-and-tweak and hierarchical sampling models were designed to explain effects of prior knowledge on the struc-

[^306]

Figure 2: Conditions tested in the behavioral experiment.
ture of categories, but they do not make any assumptions about the role of between-category contrast. Indeed, when there are no known examples of the target category, both models assume that generation is random. PACKER is thus unique in its prediction that contrast categories should influence both the structure and location of generated categories. The behavioral experiment described below was designed to test this key prediction.

The experiment follows the paradigm developed by Jern and Kemp (2013): first, participants learn members of a known category ('Alpha', or 'A'), and are then asked to generate exemplars belonging to a new category ('Beta', or 'B'). We developed two Alpha categories (see Figure 2): the 'Bottom' Alpha category is a tight cluster in the bottom-center of the space, and the 'Middle' Alpha category is identical except that it lies in the center of stimulus space.

Although our manipulation is minimal, the PACKER model predicts strong between-condition differences. According to PACKER, the nature of the space not occupied by the Alpha category should determine where members of the Beta category are likely to be generated. Thus, the lower areas of the stimulus space should be less frequently used for generation in the Bottom condition compared to the Middle (as these areas possess greater similarity to the Bottom Alpha category). Conversely, the upper areas of the stimulus space should be used for generation more frequently in the Bottom condition compared to Middle.

More generally, PACKER proposes that the probability a stimulus $y$ will be generated is a function of its similarity to contrast categories and to members of the target category. Two more general predictions (not specific to either condition) follow from this proposal: (1) the location of Beta examples should be positively related to distance from the Alpha category, and (2) Beta examples should be more similar to one another than they are to members of the Alpha category.
Participants \& Materials We recruited 122 participants from Amazon Mechanical Turk from the US equally assigned to each condition. Stimuli were squares varying in color (grayscale $9.8 \%-90.2 \%$ ) and size ( $3.0-5.8 \mathrm{~cm}$ ). The assignment of perceptual features (color, size) to axes of the domain space ( $x, y$ ) was counterbalanced across participants.
Procedure Participants began the experiment with a short training phase ( 3 blocks of 4 trials), where they observed ex-


Figure 3: Sample generated categories.
emplars belonging to the 'Alpha' category. Participants were instructed to learn as much as they can about the Alpha category, and that they would answer a series of test questions afterwards. On each trial, a single Alpha category exemplar was presented, and participants were given as much time as they desired before moving on. Exemplars were randomly ordered within each block. Participants were shown the range of possible colors and sizes prior to training.

Following the training phase, participants were asked to generate four examples belonging to another category called 'Beta'. Participants were instructed that members of the Beta category could be quite similar or different depending on what they think makes the most sense for the category, but that they were not allowed to make the same example twice. As in Jern and Kemp (2013), generation was completed using a sliding-scale interface. Two scales controlled the features of the generated example. An on-screen preview of the example updated whenever one of the features was changed. Participants could generate any example along an evenly-spaced $9 \times 9$ grid, except for any previously generated Beta exemplars. Neither the members of the Alpha category nor the previously generated Beta examples were visible during generation.

Results Several sample Beta categories are depicted in Figure 3. Because the conditions differ only in their location along the $y$-axis, we first focus on how Beta exemplars are generated above and below the contrast category. As is evident in Figure 3, we observed broad individual differences in generation strategy: Whereas some participants generated all four Beta examples within a narrow y-axis range, others generated Beta examples along a wide range.

To evaluate the key predictions of PACKER, we determined the number of participants in each condition who placed at least one Beta exemplar on the top and bottom 'rows' of the space (the maximum and minimum possible $y$ axis value, respectively). The resulting contingencies data are shown in Table 1. Fisher's Exact Tests reveal that more Middle participants generated a Beta exemplar in the bottom row , $p<0.001$, but the conditions did not differ in use of the top of the space, $p=0.16$. More Middle participants placed Beta exemplars in the top and bottom rows, $p=0.038$.

To evaluate PACKER's other predictions, we computed the

Table 1: Behavioral results.

| Middle | Used top row | No top row |
| :--- | ---: | ---: |
| Used bottom row | 28 | 18 |
| No bottom row | 11 | 4 |
|  |  |  |
| Bottom | Used top row | No top row |
| Used bottom row | 16 | 8 |
| No bottom row | 31 | 6 |

number of exemplars produced at different distances to the center of the Alpha category. These data (Figure 4 left) reveal a strong preference for stimuli that are dissimilar to the Alpha category members: maximally distant items were by far the most frequently generated.

Finally, we computed for each participant the average distance between exemplars belonging to the same and opposite categories. These data (Figure 4 right) show that, as observed by Ward (1994), most people generated Beta categories in which members are closer to one another than they are to members of the Alpha category (i.e., more between- than within-category distance). We did however, observe a notable subset of individuals with greater within-class distance. These individuals tended to adopt a 'corners' approach, in which Beta examples were placed almost exclusively in the corners of the space.

## Summary

Our results support PACKER's predictions: People tend to generate items that are dissimilar from the contrast category and similar to the target category. We observed considerable differences in generation between the Middle and Bottom conditions: Participants in the Bottom condition were less likely to use the bottom row of the stimulus space for generation, and participants in the Middle condition were more likely to create categories spanning the entire $y$-axis (utilizing the top and bottom row of the space). This latter result is especially interesting as it conflicts with previous results: Qualitatively different types of categories were generated, depending only on the location of the Alphas.

Some aspects of the results described above are somewhat commonsense: They demonstrate that the location of existing categories imposes constraints on generation because people tend to generate examples in areas not occupied by existing categories. This principle, however, is novel and not predicted by existing models of generation - these models were designed to explain distributional correspondences between generated and existing categories, not effects of contrast.

## Model Evaluation

To obtain an overall sense of each model's ability to explain our results, we fit each model by maximizing the loglikelihood of the model's predictions of the human results. The $c, \gamma$, and $\theta$ parameters were fitted for PACKER; $c$, and $\theta$


Figure 4: Behavioral results. Left: Frequency of exemplar generation as a function of distance from the Alpha category normalized by the maximum possible distance. Right: Within- vs. between-category distance for every participant.
were fitted for the copy-and-tweak model ( $\gamma$ was fixed at 1 ), and $\kappa, \lambda, v$, and $\theta$ were fitted for the hierarchical sampling model. Note that each model possess a $\theta$ parameter fulfilling the same role (response determinism). Attention in PACKER and copy-and-tweak was set uniformly. Parameters were not allowed to vary between participants or conditions - the goal was to obtain the best-fitting values to our entire dataset.

Each model's best-fitting parameterization is shown in Table 2. Overall, PACKER outperformed copy-and-tweak and the hierarchical sampling model by a considerable margin ( $\sim 11 \%$ improvement in log-likelihood). The parameter settings associated with PACKER's best fit are exactly as expected: a strong preference for items that are similar to members of the target category but are dissimilar to members of the contrast category. A similar pattern of results was obtained when we only considered the second to fourth exemplars generated by each participant.

Our model-fitting results make sense given the assumptions made by each model. As the copy-and-tweak and hierarchical sampling models are not influenced by the location of contrast categories within the space, they do not capture the broad tendency for generated items to be dissimilar to existing classes.

## Relation Between Category Structure \& Location

Generally, our behavioral results showed that members of generated categories are dissimilar to opposite categories, and similar to members of their own category. However, we also observed a great deal of individual differences in generation style. Manually inspecting the data reveals four typical patterns (see Figure 3): 'corners' categories with one Beta example in each corner of the space, tight clusters, 'column'-like categories, and 'row'-like categories. This informal inspection also reveals that each of these category types tended to be generated into distinct regions of the domain, suggesting a link between category location and distributional structure.

To more systematically evaluate this possibility, we computed, for each stimulus in the domain, the difference in range between the features (range (size) - range(color)) across every generated category that had the stimulus as member. Ag-

Table 2: Model-fitting results.

| PACKER | Copy \& Tweak | Hierarchical <br> Sampling |
| :--- | :--- | :--- |
| $A I C=3474$ | AIC $=3914$ | AIC $=3972$ |
| $c=0.565$ | $c=4.894$ | $\kappa<0.001$ |
| $\gamma=0.469$ | $\gamma=1$ (fixed) | $\nu=4.660$ |
| $\theta=6.632$ | $\theta=3.712$ | $\lambda=0.423$ |
|  |  | $\theta=2.771$ |

gregating over these range differences yields a gradient describing how categories tended to distributed for each stimulus. These data (Figure 5) reveal a systematic relationship between category structure and location. Whereas column-like categories more often include stimuli to the left or right of the Alpha class, row-like categories appear above and below the Alpha class. Thus, participants modify the distributional structure of new categories to the maximize distance from the contrast category.

To simulate this finding, we set the attention weight parameters in PACKER and copy-and-tweak per participant. The other free parameters were set as in Table 2. While there exist methods to find the optimal attention weights for a given classification (see Vanpaemel \& Lee, 2012), for simplicity we approximated the weights using proportionally to inverse of each feature's range: Thus, the Alpha and Beta categories are assumed to be distinct along dimensions that the Betas do not vary on. To simulate the hierarchical sampling model we set the domain covariance prior, $\Sigma_{0}$, proportional to the range (not inverted) of each feature: Thus new categories were distributed more widely along the features that each participant used more widely. We then simulated 50 Beta categories with each participant's weighting scheme to obtain a sense of how the relative importance of each dimension affects what types of categories are generated and where they are generated. The results of these simulations are depicted in Figure 5.

When the x -axis is weighted more, PACKER creates column categories to the sides of the Alphas. Conversely, when the $y$-axis is weighted more, PACKER creates row categories above and below the Alphas. This behavior falls out from the nature of selective attention: Dimensions weighted more have a sharper similarity gradient. For example, when the xaxis is weighted more, PACKER favors Beta categories with more within-class similarity (less range), and less betweenclass similarity along the x-axis, resulting in column-like categories that differ from the Alphas along the x -axis.

Although differentially weighting the features results in different types of categories from the hierarchical sampling and copy-and-tweak models, the location of the Alpha category does not affect where items will be generated by these models. Thus, row- and column-like categories are not systematically generated in different areas of the stimulus space, resulting in the uniform predictions shown in Figure 5.


Figure 5: Generated category structure as a function of location. Orange areas in each gradient correspond to stimuli that were commonly generated into category possessing greater y-axis range (columns). Purple areas correspond to categories possessing greater x -axis range. White areas correspond to equal range along both features (or infrequent generation).

## Discussion

The creative use of conceptual knowledge is a fascinating yet understudied topic in categorization. In this paper, we presented a novel exemplar-based approach to explaining category generation. The PACKER model proposes that categories are represented as a collection of exemplars stored in memory, and that members of generated categories should be similar to one another, yet dissimilar to members of opposing categories. Exemplar models can be viewed as ImportanceSampling approximations of Bayesian models (Shi, Griffiths, Feldman, \& Sanborn, 2010). So, PACKER can be viewed as a rational process model, approximating the expected density of a new category based on a contrast category.

In a behavioral study and subsequent formal modeling, we found broad support for the PACKER model. Participants in our study more frequently generated items that are distant from members of contrast categories, and they tended to generate categories with more within-class than between-class similarity. Likewise, we found that the location of contrast categories (as opposed to their structure) shapes generation by imposing constraints on the areas of space that remain available for a new category. Formal simulations reveal that existing models (see Jern \& Kemp, 2013), making no assumptions about category-contrast, do not account for these effects.

The PACKER model is, in general, highly expressive in its performance. Under different parameter settings it is capable of generating tightly clustered or highly distributed categories, and adjusting the distribution of categories along each feature. Future work will focus on exploring the broad degree of individual differences we observed in generation, and whether PACKER can explain previous results in the field (Jern \& Kemp, 2013; Ward, 1994).

## Acknowledgments

Support for this research was provided by the Office of the VCRGE at the UW - Madison with funding from the WARF. We thank Kenneth Kurtz for helpful comments, and Alan Jern and Charles Kemp for providing code and data.

## References

Jern, A., \& Kemp, C. (2013). A probabilistic account of exemplar and category generation. Cognitive Psychology, 66(1), 85-125.
Luce, R. D. (1977). The choice axiom after twenty years. Journal of mathematical psychology, 15(3), 215-233.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 10(1), 104.
Shi, L., Griffiths, T. L., Feldman, N. H., \& Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. Psychonomic Bulletin \& Review, 17(4), 443-464.
Smith, S. M., Ward, T. B., \& Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. Memory \& Cognition, 21(6), 837-845.
Vanpaemel, W., \& Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. Psychonomic Bulletin \& Review, 19(6), 1047-1056.
Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. Cognitive Psychology, 27(1), 1-40.
Ward, T. B. (1995). Whats old about new ideas. The creative cognition approach, 157-178.
Ward, T. B., Patterson, M. J., Sifonis, C. M., Dodds, R. A., \& Saunders, K. N. (2002). The role of graded category structure in imaginative thought. Memory \& Cognition, 30(2), 199-216.

# Opening Up and Closing Down Discussion: Experimenting with Epistemic Status in Conversation 

Shauna Concannon (shauna.concannon@qmul.ac.uk),<br>Patrick G. T. Healey (p.healey@qmul.ac.uk) \& Matthew Purver (m.purver@qmul.ac.uk)<br>Cognitive Science Research Group<br>Queen Mary University of London, United Kingdom


#### Abstract

Managing disagreement in conversation requires subtle linguistic and pragmatics skills. One key dimension is the degree of 'knowingness' with which people present their stance on an issue. It has been hypothesised that framing stances as 'knowing', i.e. with higher implied levels of speaker certainty limits the potential for challenge by others. We present the first experimental test of this hypothesis. Using a text based chat-tool paradigm and a debating task we are able to systematically manipulate how 'knowing' people's turns appear to one-another. The results show that 'knowing' stances tend to close off discussion leading to less carefully formulated, truncated turns, but do not reliably affect the range of solutions considered. Unknowing stances, by contrast, do not affect turn length or formulation but do encourage more deliberation and include more signals of certainty in the message contents.


Keywords: Dialogue; Interaction; Disagreement; Stance; Deliberation.

## Introduction

During a debate people have choices about how they present their contributions. Amongst other things they can simply assert their position, they can modify it with a propositional attitude verb such as 'know' or 'think' or they can turn an assertion into a question rephrasing "'I think X" as "Do you think X?". These choices of attitude and modality all help to establish what a person's stance is and, in combination with the choices made by their interlocutor, set the tone and direction of a debate. One of the most important hypotheses about the impact of different stance markers on dialogue relates to expressions of epistemic certainty; framing of a stance as 'knowing' or 'unknowing' appears to significantly alter the deliberative quality of a discussion Heritage (2012a).

Although the interactional dimensions of stance have been discussed in some detail (Du Bois, 2007; Englebretson, 2007; Kärkkäinen, 2003), this work is based on case studies and corpus analyses. The causal effects of adopting different stance markers on the subsequent trajectory of a dialogue has not, as far as we are aware, been directly tested. One key reason for this is the practical difficulty of manipulating stance markers in a live dialogue. Here we use a technique introduced by Healey et al. (2003) that takes advantage of the potential of text-chat for enabling selective manipulation of people's turns, including the addition of stance markers, without their awareness. We use this technique to to assess how the epistemic status of a stance, i.e. whether it is framed as either unknowing or knowing, impacts on the quality of the joint action and deliberation in discussion dialogues.

## Knowing vs unknowing epistemic status

Heritage (2012a) defines 'epistemic status' as the relative positioning in which "persons recognize one another to be more or less knowledgeable concerning some domain of knowledge". Knowing all $(\mathrm{K}+)$ is typically conveyed through declaratives, while interrogative grammatical format is the most explicit way that a speaker can embody an 'unknowing' (K-) epistemic status. For example, the question 'what time is your appointment' positions the speaker in request of information, where as 'your appointment is at 3 pm ' positions the speaker in a K+ position. However, as highlighted by Drew (2012), how much speakers know relative to one another is not only encoded in the grammatical format, but also in incongruities between epistemic status and grammatical format, for example in posing a question to which you already know the answer (e.g. 'Aren't you going to be late?'). Speakers' relative positioning can alter from moment to moment, and be "disassembled by persons who deploy epistemic stance to appear more, or less knowledgeable than they really are" (Heritage, 2012a). There are significant potential social and interactional implications of positioning ourselves or others as either knowing or unknowing (Levinson, 2012).

In issuing a question the requester assumes an unknowing epistemic status and positions the recipient in a knowing one (Heritage, 2012a), creating an obligation for the recipient to respond (Levinson, 2012). Levinson (2012) observes that people prefer polar questions to other forms that require more knowledge-rich responses and often disguise them as assertions, thus demonstrating an unwillingness to locate oneself in an unknowing position, nor to impose too greatly upon an interlocutor by demanding a response. However, in a discussion context, in which individual contributions on the topic under discussion are warranted and expected, the ways in which requests are made could be influential to the deliberative quality of the discussion.

Furthermore, between the most explicit formats of K+ and K- constructions (i.e. declaratives and interrogatives), there are a range of other ways that speakers can encode epistemic stance, such as modals, hedges and epistemic adverbs, which can convey levels of speaker certainty, e.g. 'It was definitely red', and commitment 'I absolutely think...' and evidential markings which convey the source of a knowledge claim (i.e. direct evidentials based on sensorial/ visual evidence and indirect evidentials, such as inference and hearsay). Particularly within a discussion context the management of imbal-
ances in epistemic status is particularly pertinent as participants' contributions must necessarily negotiate alternative stance positions.

## Stance and Disagreement

Work in the emergent field of interactional linguistics posits that stance-taking is a fundamentally intersubjective process, with stance positions being co-constructed through interaction (Englebretson, 2007; Du Bois, 2007). This process of stance co-construction relies upon the expression of oppositions and alternatives. Disagreement is one perspicuous social activity which denotes the negotiation of differing stances and a potential process by which a shift in stance can occur. Disagreement is generally minimised in conversation (Pomerantz, 1984; Concannon et al., 2015a), and tends to be problematic when issued without mitigation (Chiu, 2008; Concannon et al., 2015b). However, in certain contexts, such as problem solving and discussion tasks, it can be important for advancing the deliberative quality of a dialogue and encouraging novel contributions (Chiu, 2008).

There is thus a delicate balance between mitigating the socially problematic aspects of disagreement while still being able to identify and resolve differences of opinion. This balance can be achieved in many different ways. Resources such as 'well'-prefacing (Pomerantz, 1984), stance markers such as 'I think' (Kärkkäinen, 2003) and reported speech (Holt \& Clift, 2007; Concannon et al., 2015b) all provide less explicit ways of marking what follows as potentially incongruous or in opposition to what went before.

## Marking Stance in the Balloon Task

The task chosen for the experiment reported below is the Balloon Task. Participants are presented with a fictional scenario in which an hot air balloon is losing altitude and about to crash. The only way for any of three passengers to survive is for one of them to jump to a certain death. The three passengers are: Dr. Nick Riviera, a cancer scientist, Mrs. Susie Derkins, a pregnant primary school teacher, and Mr. Tom Derkins, the balloon pilot and Susie's husband. The advantages of this task are that it is effective at generating debates between subjects and there is good scope for deliberation. To ensure we chose a relatively natural manipulation of epistemic stance for this task an initial analysis was conducted using control condition transcripts from previous balloon task discussions. Twelve transcripts were analysed for markers that conveyed 'knowing' or 'unknowing' states in relation to stance marking. 'I think' was frequently used as a resource to mark a stance position. 'I think' has been attributed a dual function, and can can also act as a hedge (Holmes, 1990), however in the discussion context it was used most frequently to convey a knowing stance, particularly when at the beginning of a turn.
(1) a. I think Tom should definitely stay in the balloon
b. I think Nick should definitely be the one to go
c. I think because there's an element of risk with whether Nick will actually end up coming up with a cure for cancer ... There's no point taking two risks by then letting go of Tom
d. ithink we have a couple mins left
e. A: so tom has to jump?

B: i think so
In $1 \mathrm{a}, 1 \mathrm{~b}$ and 1 c the marker 'I think' serves to accentuate the propositional content and emphasise the speaker's commitment to their proposition and focuses on a substantive aspect, namely, who should be sacrificed. In $1 d$ and $1 e$, however, the marker performs the opposite effect and suggests a lack of speaker commitment and acts as a hedging marker. There were 44 instances of 'I think' in the transcripts, 34 instances $(77.27 \%)$ served to emphasise the propositional content it was associated with, eight instances (18.18\%) acted in a 'hedging' or unknowing capacity, and the two remaining instances made manifest the cognate processes (e.g."whenever i think that nick should go, i think 'Are susie and tom really that important?'"). Of the 44 instances, 25 were turn-initial ( $56.82 \%$ ), 19 of which served to emphasise the speaker's ownership and commitment to the content that followed. Four instances of turn-initial 'I think' ( $20 \%$ ) were constructed in such a way that 'I think' functioned as a hedging marker and two instances were not possible to classify due to insufficient context (e.g. 'I think overall'). Closer inspection of the use cases showed that all instances of turn-initial 'I think', in which the proceeding content featured a character from the scenario, conveyed a 'knowing' stance. As such, using 'I think' as a turn-initial insertion for turns which contain a mention of one of the scenario's characters, should increase the likelihood of a consistent effect of framing the utterance as 'knowing', rather than performing a hedging effect.

While looking for markers which served to downgrade the epistemic strength of assertions, 'do you think' was one such 'unknowing' device that was used in the transcripts.

A: do you think the married couple would gang up on the doctor and throw him out
B: maybe. he is their friend though
'Do you think' makes a minimised contrast pair with 'I think' and can be inserted at a turn-initial position without changing the content of the turn. Consequently, 'I think' and 'do you think' were selected as our turn-initial inserts, to perform the role of framing the proceeding content as more or less 'knowing'.

## Hypothesis

Following Heritage (2012b) our general hypothesis is that framing a proposition as unknowing invites elaboration, sequence expansion and further discussion of the topic at hand (Heritage, 2012b). Conversely, a more knowing epistemic
stance, creates pressure for confirmation and sequence closing. As such, we predict that inserting 'knowing' and 'unknowing' stance markers will have different impacts on the course of a conversation even where, counterfactually, nothing about the content of the modified assertions is changed.

The analysis of previous dialogues enables us to operationalise our general hypothesis about the level of knowingness with which opinions are presented and inform the following predictions:

1. Fewer possible solutions will be considered when contributions are framed as knowing and responses will be less considered; this should affect turn formulation, with shorter typing times and less editing of turns. Framing contributions as knowing will close down the dialogue, as indicated by shorter and fewer turns.
2. Framing contributions as unknowing will open up dialogues, leading to longer turns and more possible solutions considered.
3. More possible solutions will be considered and more care will be taken in the construction of turns, as evidenced by slower typing times and more edits when contributions are introduced with an unknowing preface ('do you think X').
4. Framing contributions as unknowing will lead to higher frequencies of certainty and uncertainty markers.

## Method

In this experiment, to see how the epistemic framing of a contribution affects levels of deliberation in dyadic text-based conversations, participant contributions were manipulated using the DiET chat tool.

## The DiET chat tool

The participants communicate via a specially programmed chat tool, similar to other instant messenger interfaces they may have used previously. The Dialogue Experimental Toolkit (DiET) chat tool is a text-based chat interface facilitating real time manipulations of the dialogue. It is possible to programme several different types of interventions using the chat tool: turns may be altered prior to transmission, turns may not be relayed, and additional turns may be added, (e.g. Healey et al. (2003), insertion of spoof clarification requests). These manipulations occur as the dialogue progresses, thus making them minimally disruptive to the sequence of dialogue.

## Design

The experiment was conducted in pairs, with 10 dyads per condition. Pairs of participants were presented with a discussion task and instructed to discuss for 30 minutes and attempt to come to an agreement. Each pair of participants was assigned at random to one of three conditions; i) Control ii) Knowing iii) Unknowing. In the Control condition there were no interventions performed by the server; participants
received the dialogue turns exactly as they were typed. In the Knowing condition turn-Initial 'I think' insertions were made and in the Unknowing condition 'Do you think' insertions were added turn-initially. Manipulations were carried out every four turns, if and only if the turn included a reference to one of the characters in the scenario (e.g. Doctor, Susie, etc.). Interventions are not visible to the individual whose turn has been manipulated, only the recipient, so that there is no awareness that turns are being intercepted before being relayed. A pilot study was conducted to establish the acceptable frequency of interventions.

## Subjects and materials

The experiment was carried out on thirty pairs of students (41 females and 19 males) from the University of London who each received $£ 7.50$ or course credits for providing an hour of their time. They were invited to attend with someone they already knew to increase the likelihood that inter-pair participants were acquainted. All subjects were native speakers of English. Pairs of participants were seated at separate computers, at opposite ends of shared office ${ }^{1}$ and given an instruction sheet detailing the balloon task (see above for a description). Participants were told to take as much time as they needed to read the summary of the situation and then discuss with their partners via a chat tool set up on the computer at which they were seated.

## Analysis

The DiET chat tool records all interventions and key presses, including edits made before participants press ENTER. For a simple measure of authorial commitment counting frequencies of epistemic adverbials, modals and hedges were collected. Epistemic adverbials are separated into two categories (adapted from Biber et al. (1999); Biber \& Finegan (1988)): those which express certainty (e.g., surely, obviously) and those which express anything less than certainty, such as possibility or probability (e.g., maybe, probably). Our separation between those that express certainty and possibility is to acknowledge that through probability there is less authorial commitment. Uncertainty Markers therefore include uncertainty adverbials as well as modals ('may', 'might' and 'could') and hedges ('quite', 'sort of', etc.), but certainty and uncertainty adverbials are also presented individually for comparison ${ }^{2}$ Obvious typographical errors were corrected to

[^307]increase the accuracy of the frequency counts (e.g. possibiliyt $->$ possibility). The inserted fragments were also removed from the transcripts before frequency counts were conducted, to ensure that the figures reflected only what the participants actively contributed.

The decision processes were hand labelled for each conversation to detect the decision patterns for each participant. The transcripts were hand coded for the solutions being considered (for example, Undecided, Kill Tom, Kill Susie, Kill Nick) and the number of shifts from one solution to another during the conversation (e.g. Kill Susie $->$ Undecided $->$ Kill Tom). Furthermore, turns were counted in which participant A and B had matching or opposing stance states. Time spent in an undecided state, even if both participant A and B both were undecided is not counted as a matching stance as it is unclear what their current stance is. A matching stance would only be A: Tom B: Tom; A: Nick B: Nick or A: Susie B: Susie.

## Results

Table 1 provides the means for total number of words typed, turns and average words per turn per participant, the mean typing time in milliseconds, the speed of typing and details of the various self-edits participants made during turn construction, such as deletion and insertion of characters before pressing send to relay the message to their partner. Standard Deviations are provided in parentheses.

Table 1: Message construction data, per participant

|  | Control | Unknowing | Knowing |
| :--- | :--- | :--- | :--- |
| Total Words | 618.30 | 641.95 | 649.05 |
|  | $(182.12)$ | $(160.28)$ | $(196.08)$ |
| Total Turns | $81.05(35.91)$ | $74.35(34.57)$ | 94.65 |
|  |  |  | $(37.50)$ |
| Words/Turn | $8.71(4.43)$ | $9.45(3.62)$ | $6.65(1.87)$ |
| Type Time | 16005.50 | 17279.88 | 12988.29 |
|  | $(8118.06)$ | $(10793.22)$ | $(5300.58)$ |
| Type Speed | $3.09(0.88)$ | $3.71(1.17)$ | $3.26(0.81)$ |
| Self- | $0.25(3.79)$ | $0.31(0.68)$ | $0.03(0.10)$ |
| edits(Ins) |  |  |  |
| Self- <br> edits(Del) | 53.32 | 73.38 | 20.41 |
|  | $(125.10)$ | $(169.25)$ | $(49.01)$ |

Word counts A nonparametric Kruskal Wallis independent samples test shows that there is a significant omnibus effect of condition on the number of words typed per turn $\left(H_{(2)}\right.$ $=7.475, \mathrm{p}=0.02$ ). A post hoc pairwise comparison using Dunn's test shows that there is a significant difference in the number of words per turn between the Knowing and Unknowing cconditionondition, with Knowing dialogues containing fewer words per turn and Unknowing dialogues containing more words per turn ( $\mathrm{p}=0.02$ ). There is no significant difference between the number of words typed per turn in the

Control and Unknowing conditions ( $\mathrm{p}=0.74$ ), nor Knowing and Control conditions $(\mathrm{p}=0.35)$. A nonparametric Kruskal Wallis independent samples test shows that the total number of words typed was not significantly affected by the condition $\left.H_{(2)}=0.283, \mathrm{p}=0.87\right)$ and there is no significant effect of condition on the number of turns per dialogue $\left(H_{(2)}=3.556\right.$, $\mathrm{p}=0.17$ ).
Typing time Typing time averaged by participant was analysed using a Generalised Linear Mixed Models analysis (GLMM) with a Gamma distribution because the timing data was positively skewed. Participants was included as a random factor and condition as a fixed factor. This shows a clear main effect of condition $(\mathrm{F}(2,59)=13.18, \mathrm{p}<0.00)$. The estimated marginal means are: Control: 12,139, Unknowing: 13,404 and Knowing: 8,813. Pairwise Contrasts show that the Knowing condition has shorter typing times than Control $(\mathrm{t}=-3.606, \mathrm{p}<0.00)$ and shorter than the Unknowing condition ( $\mathrm{t}=-4.87, \mathrm{p}<0.00$ ) but Unknowing and Control are not reliably different ( $\mathrm{t}=1.16, \mathrm{p}=0.25$ ).
Self-edits The mean number of self-edit insertions per turn is substantially lower in the Knowing condition than the Control and Unknowing conditions. A Kruskal Wallis test shows that there is a significant omnibus effect of condition on the number of Self-edits (Inserts) per participant $\left(H_{(2)}=7.761\right.$, $\mathrm{p}=0.02$ ). A post hoc pairwise comparison using the Dunn's method shows that there are significantly fewer Self-edits (Inserts) in the Knowing condition than the Unknowing condition ( $\mathrm{p}=0.04$ ), but no significant difference between Knowing and Control ( $\mathrm{p}=0.06$ ), nor Unknowing and Control condition ( $\mathrm{p}=1.0$ ). The mean number of self-edit deletions per turn is higher in the Unknowing condition than the Control and Knowing conditions. However, a non-parametric Kruskal Wallis test shows that there is no significant effect of condition on the number of Self-edits (Deletions) $\left(H_{(2)}=4.560\right.$, $\mathrm{p}=0.10$ ).
Epistemic Strength Table 2 provides mean frequencies of epistemic markers, adverbials of certainty, adverbials of uncertainty and combined uncertainty markers (adverbials,hedges,modals) per 100 words. A non-parametric

Table 2: Epistemic marker mean frequencies

| Condition | Certainty <br> Adverbials | Uncertainty <br> Adverbial | Uncertainty <br> Markers |
| :--- | :--- | :--- | :--- |
| Control | $0.28(0.25)$ | $0.54(0.32)$ | $4.69(1.12)$ |
| Knowing | $0.33(0.34)$ | $0.55(0.21)$ | $4.60(1.19)$ |
| Unknowing | $0.67(0.35)$ | $0.65(0.39)$ | $4.69(0.88)$ |
| Total | $0.43(0.35)$ | $0.58(0.31)$ | $4.66(1.04)$ |

Kruskal Wallis test shows that there is an omnibus effect of condition on the frequency of certainty adverbs $\left(H_{(2)}=\right.$ $7.501 \mathrm{p}=0.02$ ). A post-hoc pairwise comparison Dunn's test shows that there are significantly more certainty adverbs in
the Unknowing condition compared to the Control condition ( $\mathrm{p}=0.04$ ), but no significant difference in frequencies between the Control and Knowing ( $\mathrm{p}=1.00$ ), nor Knowing and Unknowing conditions ( $\mathrm{p}=0.08$ ). A non-parametric Kruskal Wallis test shows that there is no omnibus effect of condition on the mean frequencies of uncertainty adverbials $\left(H_{(2)}\right.$ $=0.742 \mathrm{p}=0.690)$ or combined uncertainty markers $\left(H_{(2)}=\right.$ $0.148 \mathrm{p}=0.93$ ).

Deliberation quality Table 3 details the mean number of changes from a given stance position to another per participant over the course of the dialogue for each condition, as well as the total number of possible alternatives considered.

Table 3: Mean stance shifts during dialogue and possible solutions considered per participant by condition

| Condition | Shifts in Stance | Solutions Considered |
| :--- | :--- | :--- |
| Control | $4.85(1.84)$ | $3.10(0.97)$ |
| Unknowing | $6.80(2.63)$ | $3.30(0.66)$ |
| Knowing | $4.55(1.61)$ | $2.75(0.55)$ |
| Total | $5.40(2.27)$ | $3.05(0.77)$ |

There are a third more stance shifts in the Unknowing condition than the Control and Knowing conditions. A Kruskal Wallis non-parametric test shows that there is a significant omnibus effect of condition on the number of stance shifts traversed by a participant $\left(H_{(2)}=9.559 \mathrm{p}=0.008\right)$. A planned pairwise post hoc comparison using the Dunn's test shows that there are significantly more stance shifts in the Unknowing condition than the Knowing condition ( $\mathrm{p}=0.01$ ) but no confirmed significant effect between Unknowing and Control ( $\mathrm{p}=0.06$ ). There is an omnibus effect of condition on number of possible solutions considered $\left(H_{(2)}=6.146 \mathrm{p}<0.05\right)$. There are more possible solutions considered in the Unknowing condition than the Knowing condition ( $\mathrm{p}=0.044$ ). There is no significant difference between Knowing and Control ( $\mathrm{p}=0.33$ ) and nor Control and Unknowing ( $\mathrm{p}=.1 .00$ ).

Table 4 provides details of the mean percentage of turns in which participant A and B had matching and opposing stance states across conditions.

Table 4: Mean percent of dialogue in which participant A and B had matching and opposing stances

| Condition | Turns: Matching | Turns: Opposing |
| :--- | :--- | :--- |
| Control | $39.42 \%$ | $60.58 \%$ |
| Unknowing | $48.27 \%$ | $51.73 \%$ |
| Knowing | $32.74 \%$ | $67.26 \%$ |
| Total Mean | $40.15 \%$ | $59.85 \%$ |

Although, the distributions show approximately $16 \%$ difference in the ratio of opposing and matching stances between Knowing and Unknowing conditions, with more turns covered with opposing stances in the Knowing condition and
more matching stances in the Unknowing condition. A nonparametric Kruskal Wallis test find no significant effect of condition on the distribution of oppositional and matching stance states amongst participants $\left(H_{(2)}=3.850 \mathrm{p}=0.15\right)$.

## Discussion

In line with our prediction, the results show that framing statements as unknowing led to more deliberation in the dialogues. Not only was there a higher numbers of shifts in stance, indicating a thorough deliberation going back and forth over the possible solutions, there was also a fuller exploration of the total possible solutions (i.e. participants in the Unknowing condition were more likely to consider all of the four possible outcomes, and consider each person to be ejected rather than just sticking to one or two).

The results show that the introduction of the knowing stance marker 'I think' leads to fewer words per turn, i.e. shorter, or more terse responses. In part this may be so to the declarative format, compared to the question format of 'do you think', which obligates a reply. The greater efficiency in the construction of dialogue turns suggests that the introduction of the knowing stance marker leads to more direct exchange of opinions, which is supported by the fewer edits during turn construction in this condition. Less care is taken in the Knowing condition to alter the message prior to relaying it to a conversational partner, perhaps leading to less delicately constructed or polite turns, but more direct and less guarded opinion exchange. The results show that prefacing statements with a knowing preface (i.e. 'I think') forecloses the conversation, while the framing of the contribution with do you think leads to more considered and extended responses.

Counter to our predictions there was no significant effect of condition on the frequency of expressions of uncertainty. However, significantly more certainty adverbials are employed by participants in the Unknowing condition compared to the Control condition. This suggests that framing contributions as unknowing creates an environment in which participants are more likely to make manifest their commitment to a stance by upgrading the epistemic strength of a statement through certainty adverbials; as solutions are discussed more and potentially co-constructed, once a stance is established it can be committed to with greater conviction by participants in the Unknowing condition. So, although the Knowing condition features less guarded and more direct messages as indicated in the manner in which they are constructed, it is in the Unknowing condition that participants commit more firmly to the substantive essence of their utterance.

Interpreting these results together suggests that the introduction of 'Do you think' opens up the dialogue, inviting further elaboration of the topic at hand, while introducing 'I think' closes down the dialogue and limits the deliberative quality of the discussion. 'Do you think' positions the speaker in a position of unknowing epistemic status, and also directly invokes the hearer to collaborate in the co-construction of a
joint stance. In the Unknowing condition stance positions are more explicitly emphasised through certainty adverbials, i.e. when something is important, participants take care to make clear the focus of their stance and their strength of commitment to a given proposition. In part this may be due to the fact that 'do you think' directly invites input and therefore greater care is taken to make clear exactly what the opinion to which they are attaching themselves is. The interactive negotiation of the stance is more exaggerated. Conversely, the introduction of 'I think' to the dialogue has the opposite effect: the presentation of a knowing stance, leads to less consideration and more conviction among participants, demonstrated through fewer edits when constructing responses and more terse and direct turns. Opinions are expressed plainly and without additional specification.

## Conclusion

In this paper the causal effects of epistemic status, as expressed through particular stance markers, on the deliberative quality of a dialogue were investigated using an experimental approach. Framing a statement as unknowing has a significant impact on the deliberative quality of a dialogue and increases the likelihood that participants will consider multiple possible solutions, shifting their opinion more times before reaching a concluding stance. Furthermore, participants in the Unknowing condition, spent a larger proportion of dialogues considering one another's stance. This suggests that, within a discussion dialogue, the framing of a statement in a unknowing way can lead to a more flexible deliberation process and a greater willingness to engage with alternative viewpoints. Furthermore, while being more considerate of one another's views, this was not to the detriment of expressing a position with conviction, and actually led to greater displays of speaker commitment to a stance through certainty adverbials.

Framing a statement as knowing affects the ways in which individuals produce messages; specifically, they construct shorter and less edited responses. This suggests that there is less care taken in the construction of messages, and less conscientious effort put into producing polite, or considered turns. Shorter messages are typically more direct and the lack of editing may reflect decreased guardedness. By prefacing statements with 'I think', the context is set for the exchange of opinions; by introducing a stance with a knowing marker, the appropriateness for a response which is equally direct is established. Overall it seems that marking stances with a knowing preface leads to more direct and unguarded exchanges, but does not improve the deliberative quality of the dialogues. Conversely, prefacing statements with the unknowing preface 'do you think' encourages a more collaborative deliberation, in which more possible solutions are considered in turn before a final decision is reached.

## Acknowledgments

This work was funded by EPSRC through the Media and Arts Technology Programme, an RCUK Doctoral Training Centre

EP/G03723X/1. The authors would like to thank the reviewers for their helpful comments.

## References

Biber, D., \& Finegan, E. (1988). Adverbial stance types in english. Discourse processes, 11(1), 1-34.
Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., \& Quirk, R. (1999). Longman grammar of spoken and written english (Vol. 2). MIT Press.
Chiu, M. M. (2008). Flowing toward correct contributions during group problem solving: A statistical discourse analysis. the journal of the learning sciences, 17(3), 415-463.
Concannon, S., Healey, P. G., \& Purver, M. (2015a). Shifting opinions: An experiment on agreement and disagreement in dialogue. In Proceedings of the 19th semdial workshop on the semantics and pragmatics of dialogue (godial).
Concannon, S., Healey, P. G., \& Purver, M. (2015b). Taking a stance: a corpus study of reported speech. In Proceedings of the 19th semdial workshop on the semantics and pragmatics of dialogue (godial).
Drew, P. (2012). What drives sequences? Research on Language \& Social Interaction, 45(1), 61-68.
Du Bois, J. W. (2007). The stance triangle. Stancetaking in discourse: Subjectivity, evaluation, interaction, 139-182.
Englebretson, R. (2007). Stancetaking in discourse: Subjectivity, evaluation, interaction (Vol. 164). John Benjamins Publishing.
Healey, P., Purver, M., King, J., Ginzburg, J., \& Mills, G. (2003). Experimenting with clarification in dialogue. In Proceedings of the 25th annual meeting of the cognitive science society (pp. 539-544).
Heritage, J. (2012a). The epistemic engine: Sequence organization and territories of knowledge. Research on Language \& Social Interaction, 45(1), 30-52.
Heritage, J. (2012b). Epistemics in action: Action formation and territories of knowledge. Research on Language \& Social Interaction, 45(1), 1-29.
Holmes, J. (1990). Hedges and boosters in women's and men's speech. Language \& Communication, 10(3), 185205.

Holt, E., \& Clift, R. (Eds.). (2007). Reporting talk: Reported speech in interaction. Cambridge University Press.
Kärkkäinen, E. (2003). Epistemic stance in english conversation: A description of its interactional functions, with a focus on i think. John Benjamins Publishing.
Levinson, S. C. (2012). Interrogative intimations: On a possible social economics of interrogatives. In J. P. de Ruiter (Ed.), Questions: Formal, functional and interactional perspectives (pp. 11-32). Cambridge University Press.
Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. H. Atkinson J.M. (Ed.), Structures of social action: Studies in conversation analysis. Cambridge University Press, Cambridge.

# But where's the evidence? The effect of explanatory corrections on inferences about false information 

Saoirse Connor Desai (saoirse.connor-desai@city.ac.uk) \& Stian Reimers (stian.reimers.1@city.ac.uk)<br>City, University of London, Department of Psychology, Whiskin Street<br>London, EC1R 0JD


#### Abstract

Research on the continued influence effect has consistently shown that people continue to rely on false causal information despite being corrected by more recent information. Corrections are most effective when paired with an alternative explanation that 'fills the causal gap' left by the correction. However, it may not always be possible provide an alternative explanation. Previous research suggests people more readily discount unreliable information. Two experiments examined whether corrections to false causal information in a news report are more effective when the correction explains why the source of the false information was unreliable. The results showed that a correction did not fully eliminate reliance on false information and that an explanatory correction was no more effective than a non-explanatory correction. People also continued to rely on false information when there was limited information to support its validity. Possible explanations for the ineffectiveness of explanatory corrections are discussed.


Keywords: False Information; Continued Influence; Corrections; Inference; Explanations; Reasoning; Memory

## Introduction

Many news organizations now report breaking news via social media platforms. Although social media helps to keep people up to date, breaking news can be based on mistaken, inaccurate, or incomplete information. When false information ${ }^{1}$ is reported news organizations would ordinarily issue a correction, revising their original account. Provided that the false information has not proliferated, being aware of a correction should normatively neutralize belief in the false information. In contrast, numerous experiments on the continued influence effect (CIE) have shown that causal false information continues to be influential beyond a correction (Ecker, Lewandowsky, \& Apai, 2011; Ecker, Lewandowsky, Swire, \& Chang, 2011; Ecker, Lewandowsky, \& Tang, 2010; Johnson \& Seifert, 1994; Wilkes \& Leatherbarrow, 1988).

The standard experimental paradigm for studying the CIE involves reading a series of messages describing a fictional news story over time. Explanatory target information is presented and subsequently corrected for one group of participants, but remains uncorrected for a control group. Inferences and memory for the news report are then assessed through a series of open-ended questions. In Johnson and Seifert (1994), participants read a story about a warehouse fire wherein target information implies that

[^308]carelessly stored flammable materials (oil paint and gas cylinders), were a likely cause of the fire. Later in the story, some participants learn that no such materials had actually been found. A comprehension test follows, which includes indirect inference questions (e.g., "what could have caused the explosions?"), and questions assessing recall of basic facts (e.g., "what was the cost of the damage done?"). Inference responses are then coded in order to measure the extent to which the target information (oil paint and gas cylinder) has been discounted. Responses are coded according to whether they are consistent with the explanatory theme implied by the target information (e.g., "exploding gas cylinders") or not (e.g., "electrical short circuit").

The key finding from CIE studies (for reviews see (Lewandowsky, Ecker, Seifert, Schwarz, \& Cook, 2012; Seifert, 2014) is that corrections do not fully eliminate reliance on false information. People continue to rely on false information despite recalling the correction, when given prior warnings about false information in news reports; whether corrections are repeated, or appear immediately after the false information (Ecker, Lewandowsky, \& Apai, 2011; Ecker et al., 2011; Ecker et al., 2010; Johnson \& Seifert, 1994). The CIE has been replicated using various types of news stories, types of false information (e.g., Ecker et al., 2011), and using direct or indirect measures of reliance on false information (Connor Desai \& Reimers, 2016; Rich \& Zaragoza, 2016). Identifying the cognitive mechanisms underlying the successful correction of false information has timely, realworld implications in a wide variety of domains (e.g., news stories, public health information, in the courtroom).
Filling the causal gap One explanation for the CIE is that corrections are ineffective because a correction alone leaves a causal gap in a person's mental model of the reported event (e.g., Johnson \& Seifert, 1994; Wilkes \& Leatherbarrow, 1988; Lewandowsky et al., 2012). In this view, people maintain the false information because they prefer an inconsistent to an incoherent mental event model. In the warehouse fire example, an individual might infer that a fire started by an electrical short circuit was a result of negligence, based on information suggesting that flammable liquids were carelessly stored. Correcting a key piece of causal information (i.e., no flammable liquids) results in an incoherent mental model. People might continue to draw causal inferences from the false information because it is the only explanation available to them. In line with the mental models account, combining a correction with an alternative
explanation to 'fill the causal gap' considerably reduces the degree to which people rely on false information (e.g., there was evidence the fire was caused by arson; Ecker et al., 2011; Johnson \& Seifert, 1994a; Rich \& Zaragoza, 2016; Tenney, Cleary, \& Spellman, 2009).

In the real world it is not always possible to provide a single, coherent, alternative explanation to replace corrected false information (e.g., the true cause the Flight MH370 disappearance still remains unknown). Due to the fact that alternative explanations are not always available, it is important to identify other means of increasing the effectiveness of corrections.
Explanatory corrections One way of increasing the impact of corrections is to explain why the original information is no longer relevant or useful. For example, Bush, Johnson, and Seifert (1994) found that explaining that target information had been of poor quality (the storeroom actually contained cans of coffee and soda canisters), or was no longer relevant (a delivery of paint and gas cylinders was expected but never arrived), enhanced the effectiveness of the correction statement compared to a correction alone, but an explanatory correction was still not as effective an offering an alternative explanation. Bush et al., also found that ruling out the involvement of the corrected information (there was clear evidence that no paint or gas were ever on the premises) without providing an explanation actually decreased the effectiveness of the correction. These findings can be understood by the pragmatic inferences people draw about the conversational implications of the original statement (cf. Seifert, 2014). The validity of corrected information might be reinforced because people assume that a speakers only offer true (maxim of quality) and relevant (maxim of relevance) information (Grice, 1975). Bush et al's findings suggest that the person issuing the correction must explain why the original information should no longer be believed in order ensure the correction is understood. Legal decision-making studies support the idea that explaining why initial information is unreliable can enhance the effectiveness of a correction. For example, Kassin and Sommers (1997) found that mock-jurors who learned a key piece of incriminating evidence was inadmissible because it was unreliable (a taped confession secured without a warrant) were more likely to convict a defendant than mock-jurors who were told that the evidence was unreliable (the tape was inaudible). Similarly, Fein, McCloskey and Tomlinson (1997) found mock-jurors discounted inadmissible incriminating testimony when its reliability was called into question. Finally, Lagnado and Harvey (2008) showed that people providing evidence that an eyewitness has a 'longstanding grudge' against the suspect resulted in participants discounting that testimony. These studies suggest that explanatory corrections could be as effective as combining a correction with an alternative explanation when the correction explains why the initial source of the false information is unreliable.

## Pilot study

The pilot study tested whether explanatory corrections are more effective than a correction alone when the correction explains why the original source of the false information is unreliable (i.e., mistaken or intentionally deceptive). There were two main predictions: 1) Explanatory correction groups would produce fewer target information consistent inferences than the correction alone group, and 2) Correction only group would produce fewer target information consistent inferences than a group who was never exposed to a correction.

## Methods

Participants Forty-five U.S. based participants were recruited from Amazon Mechanical Turk (17 female, age $36.3 \pm 9.70$ ). Participants were paid $\$ 1$ and took an average of 14 minutes to complete the experiment.
Design Participants were randomly assigned to either the no correction (11), correction only (10), explanatory correction error (10), or explanatory correction lie (14) correction groups. There were four main dependent measures: 1) references to target information on inference questions, 2) recall on filler items, and 3) awareness of the correction.
Materials and Procedure Participants read a news story describing a warehouse fire, displayed as a series of sequentially presented short messages. Materials were reconstructed from an experiment by (Johnson \& Seifert, 1994; Exp 3a). There were 12 discrete messages (1, target message, 1 critical message, 1 causal detail message, 9 additional messages), in the style of 'Tweets' from the social media platform Twitter, an approach inspired by Hardwicke, Manning and Shanks (2016). The 'Tweets' originated from the same fictional news outlet, called "news now" and each message was no longer than 140 characters. Messages appeared one a time for a minimum of 5 seconds each; there was no maximum time. Participants clicked a button to proceed to the next message; they were unable to return and view previous messages.

Participants completed an instructional attentional check (e.g., Oppenheimer, Meyvis, \& Davidenko, 2009) before starting the experiment. The explanatory theme implied by the target message was that flammable materials had been carelessly stored in a storeroom. The target message, containing information about a possible cause of the fire (there were cans of oil paint and gas cylinders present in a storeroom), was presented at Message 5. The causal detail containing information consistent with the explanatory theme implied in the target message (thick, oily smoke + sheets of flames hinder firefighters efforts, intense heat has made the fire difficult to bring under control) appeared at Message 8. The critical message varied depending on condition and appeared at Message 11. The remaining (filler) messages provided event information, which was neutral with respect to the explanatory theme implied by the target message (e.g., Three warehouse workers working overtime, have been taken to St Columbus Hospital, due to smoke inhalation).

Table 1 Example questions and responses from pilot study

| Question | Example response to <br> receive score of 1 on false <br> information measure |
| :--- | :--- |
| What aspect of the fire <br> should the police focus on <br> in their investigation? | They should focus on the <br> chemical aspects because <br> it seemed to have started <br> from paint or gas. |
| What was the most likely <br> overall cause of the fire? | The oil paint cans and <br> pressurized gas cylinders. |
| Is there any evidence of <br> careless management in <br> relation to this fire? | Yes the pressurized <br> cylinders should not have <br> been kept indoors next to <br> paint cans. |

In the three correction conditions Message 11 corrected earlier information about the contents of the storeroom; participants in the no correction condition learned instead that warehouse workers taken to hospital had been released. The explanatory correction groups either learned that the target information had been corrected because an employee confused the soda canisters and coffee cans for paint and gas (error) or that an employee lied that there were flammable materials in the storeroom (lie). There were four narrative versions in total.

After reading all of the 'Tweets' participants completed a questionnaire consisting of seven inference questions, seven filler questions and two questions assessing awareness and understanding of the correction. Inference and filler questions were presented in a random order. Inference questions asked participants about information not explicitly mentioned in the news report (e.g., "Is there any evidence of careless management in relation to this fire?"), and included a question querying participants about what they thought the most likely cause of the fire was. Filler questions enquired about the explicit details included in additional (filler) messages included in the news story (e.g., "Which hospital were the workers taken to?"). Two further questions assessed awareness and understanding of the correction message. Participants typed a response to each of 16 questions in a text box, were required to use a minimum of 25 characters, and encouraged to answer using full sentences.

## Pilot study: Results

## Coding of Responses

The main dependent variable extracted from inference question responses was 'references to target information'. References that explicitly stated, or strongly implied, that the fire was caused by gas and oil paints were scored a 1 on the target information measure, and were otherwise scored as 0 . The maximum individual score for inference questions was 7. Filler question responses were scored for accuracy. Correct or partially correct responses were scored 1 and a score 0 was given for an incorrect response. The maximum individual score for filler questions was also 7. Awareness


Figure 1: Mean target information inference scores (left panel), filler accuracy scores (top right panel), and awareness of correction scores (bottom right panel) as a function of correction. Error bars represent $95 \%$ confidence interval of the mean.
of correction scores were computed using the same criteria; the maximum individual awareness of correction score was 2. One-way ANOVA analyzed differences between the correction conditions for all three measures. ${ }^{2}$ Fig 1 shows mean inference, filler and awareness of correction as a function of correction condition.
Inference scores There was significant effect of correction on the number references to target information, $F(3,41)=$ 3.32, $p<.05, \eta^{2}=$.20. Planned contrasts showed a correction reduced references to target information compared to no correction, $t(19)=-2.98, p<.01, d=1.46$. However, neither an error explanatory correction, $t(19)=-$ $0.55, d=0.23$, nor a lie explanatory correction, $t(23)=-$ $1.45, d=0.53$, reduced references to target information compared to no correction.
Filler recall accuracy There was a significant effect of correction on filler recall accuracy, $F(3,41)=4.39, p<.01$, $\eta^{2}=.24$. Tukey's tests showed the no correction group recalled significantly more filler details than the lie explanatory correction group, $t(23)=-3.34, p=.009, d$ $=1.14$. None of the other differences were significant ( $p$ 's $>$ .05).
Awareness of correction There was a significant effect of correction on awareness of correction scores, $F(2,31)=$ $6.52, p<.01, \eta^{2}=.30$. Tukey's tests revealed that the correction only group showed more awareness of the correction than the error, $t(18)=-3.09, p=.01 d=1.62$, or lie, $t(22)=-3.24, p<.01, d=1.22$ explanatory correction groups. The two explanatory correction groups did not significantly differ, $p=.10$.

[^309]

Figure 2: Content of critical messages in main experiment. In contrast to previous studies, the critical message in each of the correction conditions explicitly stated that the message was a correction.

## Interim Discussion

A correction alone reduced, but did not fully eliminate, target information references when compared to no correction. Pilot results also showed that an explanatory correction did not reduce references to target information compared to no correction. On average both explanatory correction groups made more target information references than the correction only group, although these differences were not significant. These results are inconsistent with previous findings showing that an explanatory correction was more effective at reducing reliance on target information than a correction alone (cf. Bush, Johnson \& Seifert, 1994).

The most likely reason that an explanatory correction was less effective than a correction alone is that participants in the explanatory correction groups showed poorer awareness and understanding of the correction, than the correction only group. Only $40 \%$ of the error explanatory correction group, and $21 \%$ of the lie explanatory correction group understood and were aware of the correction, compared to $90 \%$ of the correction only group. Both explanatory correction groups also recalled fewer story details on average than the correction only group. Some participants' responses indicated doubts about the credibility of the correction message (e.g., questioning whether the employee really lied about the contents of the storeroom), and other responses suggested misunderstanding of the correction message (e.g., the employee thought there was soda and coffee but there was actually paint and gas). A lack of clarity of the explanatory correction messages could explain poorer awareness and understanding in explanatory correction conditions could also explain why the current results do not replicate previous findings (cf. Bush et al., 1994). The main experiment sought to rectify these issues by enhancing the clarity of the correction messages.


Figure 3: Mean target information inference score as a function of condition in main experiment. Error bars represent the $95 \%$ confidence interval of the mean.

## Main Experiment

The same general setup was employed in the main experiment except that a number of changes were made to rule out explanations identified in the interim discussion. The hypotheses and predictions were also the same as the pilot study.
Participants Three-hundred and twelve U.S. based participants were recruited from Amazon Mechanical Turk (146 female, age $39.67 \pm 12.31$ ). Participants were paid $\$ 1$ and took an average of 20 minutes to complete the experiment.
Design, materials and procedure Participants were randomly assigned to either the no correction (71), correction only (87), explanatory correction error (71), or explanatory correction lie (83) groups. Dependent measures were the same as in the pilot study.

Content of the critical messages was modified from the pilot study in order to make it unequivocally clear that the target information was being corrected (see Fig 2). Unlike previous studies, the critical message for the correction conditions explicitly stated that the target information was being corrected.

## Results

Additional coding and analysis was performed on one of the filler questions to the total number of references indicating flammable substances had been in the storeroom before the fire. The additional 'discounting' measure further assessed the extent to which the false information had been disregarded. Responses were scored 1 if the response indicated there were flammable substances in the storeroom before the fire and 0 otherwise. One-way ANOVA analyzed differences between conditions for all four dependent measures.
Inference scores There was a significant effect of correction on references to the target information, $F(3,308)$ $=23.23, p<.001, \eta^{2}=.18$. Planned contrasts revealed a correction significantly reduced the number of references to target information, $t(156)=-6.84, p<.001, d=0.98$.

Likewise, the error, $t(140)=-6.90, p<.001, d=1.02$, and lie, $t(152)=-6.99, p<.001, d=1.01$., explanatory correction groups, made significantly fewer references to the target information than the no correction group. Mean target information inference scores are shown in Fig 3.
Filler recall accuracy There was no effect of correction on filler recall accuracy, $F(3,308)=0.64, p=.90, \eta^{2}=.01$, so it was not necessary to perform contrast analysis. Mean filler recall scores ranged from 4.58 to 4.96 (out of 7 ).
Awareness of correction No correction group responses were excluded from analysis because their responses to awareness of correction questions were meaningless. There was a significant effect of correction condition on awareness of correction scores, $F(2,238)=3.76, p<.05, \eta^{2}=.03$. Tukey's tests showed a significant difference between the correction only and explanatory correction error group, $t$ $(168)=-2.39, p=.05, d=0.37$. There were non-significant differences between the explanatory correction error and correction only groups ( $p=.06$ ), and both the explanatory corrections groups ( $p=1$ ). Given the small effect size the difference is considered negligible.
Discounting false information The inference scores suggest that explanatory corrections are treated the same as a correction alone. If this is the case, then the number of references indicating the storeroom contained flammable substances before the fire should be equivalent to inference scores. There was a significant effect of correction on the number of responses indicating the storeroom contained flammable substances before the fire, $F(3,308)=57.25, p$ $<.001, \eta 2=.36$. Planned contrasts confirmed the same pattern of results as inference scores; there were significantly higher number of references stating that flammable substances had been in the storeroom before the fire in the no correction than the correction only group, $t$ $(156)=9.22, p<.001, d=1.49$, the error correction group, $t$ (140) $=-12.27, p<.001, d=2.81$, or the correction lie group, $t$ (152) $=-9.99, p<.001, d=1.68$. A closer inspection of the responses suggested that explanatory corrections were not treated the same as a correction alone. Fig 4 shows the mean number of responses indicating that flammable substances were in storeroom before the fire. Tukey's tests showed that the explanatory correction error group significantly differed to the correction only, $t(156)=$ $-3.66, p<.01, d=0.55$, and explanatory correction lie group, $t(152)=-2.75, d=-.45$. The difference between the correction only and explanatory correction lie groups was not significant, $p=.80$.

## Discussion

The results show a clear continued influence effect; a correction significantly reduced, but did not eliminate, references to target information. A correction appeared to have a similar impact on inferences whether accompanied by an explanation as to why the original source of the false information should not be trusted, or not. A closer inspection of responses suggested fewer people continued to think that flammable substances had been in the storeroom


Figure 4: Mean number of references to presence of flammable substances in the storeroom before the fire as a function of correction condition. Error bars represent 95\% confidence interval of the mean.
before the fire when the correction replaced the contents of the storeroom (i.e., there were soda cans and gas canisters in the storeroom) than when the correction left the storeroom empty before the fire (i.e., the employee had lied about flammable materials in the storeroom). In addition, the continued influence effect was still observed despite the fact that the correction to target information was explicitly stated in the correction message.

## General Discussion

The experiments reported in this paper examined the impact of explanatory corrections on inferences about false information in the context of breaking news reports on social media. The findings reported here are consistent with previous studies showing that corrections do not fully eliminate reliance on false information (Ecker, Lewandowsky, Swire, et al., 2011; Ecker et al., 2010; Johnson \& Seifert, 1994; Wilkes \& Leatherbarrow, 1988). These results also provide a novel contribution to the literature on the continued influence effect. Specifically, a correction that explained why the original source of the false information was unreliable was no more effective in reducing reliance on false information, than a correction alone. Participants made an equivalent number of references to target information whether the correction provided an explanation for why the target information should no longer be believed (i.e., the current inaccuracy of the target information was directly attributed to a mistaken or a deceptive individual), or not.

Despite this finding there was evidence to suggest that corrections were not treated equally. People were less likely to say that flammable substances (oil paint and gas canisters) were in the storeroom before the fire when the contents of the storeroom were replaced with other objects (soda canisters and coffee cans) than when the contents of the storeroom were not replaced (i.e., the employee lied that there were flammable items in the storeroom). One explanation for this inconsistency between inferences and
memory of the storeroom contents is that people who received the error correction updated their representation of the contents of the room whilst maintaining an inconsistent mental model of the event. In contrast, people did not update their representation of the storeroom when the contents were not replaced with alternative materials. These results do not support previous findings showing that explanatory corrections are more effective than a correction alone (Bush, Johnson, \& Seifert, 1994).

The main methodological difference between the current study and previous study is that the explanatory correction conditions in this study involved an additional source of information. In addition to making a judgment about whether the correction sufficiently negated the false information participants had to establish why or how the original source of the information (i.e., the employee) provided the information in the first place. Without knowing why the employee lied about the flammable materials or how the employee was able to confuse flammable for nonflammable substances, people might still assume the false information is relevant. Another possible reason for the inconsistent results could be that in at least one of Bush et al's conditions the correction made it logically impossible to continue to rely on the false information whereas this was not the case in the current study. These findings further demonstrate that pragmatic inferences play an important role in successfully correcting false information.

The current studies also showed evidence of the continued to rely on false information even though the report only contained one piece of information that reinforced the false information explanatory theme. This suggests that people construct a mental model of the incident on the basis of limited causal information. If there is no information to indicate an alternative explanation then people fall back on the only explanatory information available to them. It is also possible that people interpret (or re-interpret) information as supporting their leading hypothesis (e.g., Carlson \& Russo, 2001). Future studies are necessary to address whether people re-interpret neutral information to fit false causal information or whether people construct their mental event model based on limited information.

While the current study provides initial steps, there is a lot more left to explore. It will be necessary to further explore why explaining why the information was unreliable was no more effective than withdrawing the false information and why the current findings are discrepant with previous continued influence (Bush et al., 1994) and legal decision making studies (e.g., Lagnado \& Harvey). Future studies will further investigate the role of source reliability in correcting false information, and use a wider range of scenarios as well as types of false information.

## Acknowledgments

We thank Dr. Tom Hardwicke for his insights and expertise regarding the continued influence effect paradigm.

## References

Bush, J. G., Johnson, H. M., \& Seifert, C. M. (1994). The implications of corrections: Then why did you mention it? In Proceedings of the Sixteenth Annual Cognitive Science Society Conference, Atlanta, Georgia (Vol. 27, pp. 112-117). Atlanta.
Carlson, K. a, \& Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. Journal of Experimental Psychology. Applied, 7(2), 91-103. http://doi.org/10.1037/1076-898X.7.2.91
Connor Desai, S., \& Reimers, S. (2016). Direct and indirect measures of the Continued Influence Effect. Unpublished manuscript.
Ecker, U. K. H., Lewandowsky, S., \& Apai, J. (2011). Terrorists brought down the plane!--No, actually it was a technical fault: processing corrections of emotive information. Quarterly Journal of Experimental Psychology (2006), 64(2), 283-310. http://doi.org/10.1080/17470218.2010.497927

Ecker, U. K. H., Lewandowsky, S., Swire, B., \& Chang, D. (2011). Correcting false information in memory: manipulating the strength of misinformation encoding and its retraction. Psychonomic Bulletin \& Review, 18(3), 570-578. http://doi.org/10.3758/s13423-011-0065-1
Ecker, U. K. H., Lewandowsky, S., \& Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory \& Cognition, 38(8), 1087-1100. http://doi.org/10.3758/mc.38.8.1087
Fein, S., McCloskey, a. L., \& Tomlinson, T. M. (1997). Can the Jury Disregard that Information? The Use of Suspicion to Reduce the Prejudicial Effects of Pretrial Publicity and Inadmissible Testimony. Personality and Social Psychology Bulletin, 23(11), 1215-1226. http://doi.org/10.1177/01461672972311008
Grice, H. P. (1975). Logic and conversation. In J. L. Cole, P., \& Morgan (Ed.), Syntax and semantics, Vol 3: Speech acts (pp. 41-58). New York: NY: Academic Press.
Johnson, H. M., \& Seifert, C. M. (1994). Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(6), 1420-1436. http://doi.org/10.1037/0278-7393.20.6.1420
Kassin, S. M., \& Sommers, S. R. (1997). Inadmissible Testimony, Instructions to Disregard, and the Jury: Substantive Versus Procedural Considerations. Personality and Social Psychology Bulletin. http://doi.org/10.1177/01461672972310005
Lagnado, D. a, \& Harvey, N. (2008). The impact of discredited evidence. Psychonomic Bulletin \& Review, 15(6), 1166-1173. http://doi.org/10.3758/PBR.15.6.1166
Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., \& Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest, 13(3), 106-131. http://doi.org/10.1177/1529100612451018
Oppenheimer, D. M., Meyvis, T., \& Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45(4), 867-872. http://doi.org/10.1016/j.jesp.2009.03.009
Rich, P. R., \& Zaragoza, M. S. (2016). The continued influence of implied and explicitly stated misinformation in news reports. Journal of Experimental Psychology. Learning, Memory \& Cognition, 42(1), 6274. http://doi.org/http://dx.doi.org/10.1037/xlm0000155

Seifert, C. M. (2014). The Continued Influence Effect: The Persistence of Misinformation in Memory and Reasoning Following Correction. (Rapp, David N., Braasch, Jason L.G., Ed.). Cambridge, Massachusetts: MIT Press.
Tenney, E. R., Cleary, H. M. D., \& Spellman, B. A. (2009). Unpacking the doubt in "beyond a reasonable doubt": Plausible alternative stories increase not guilty verdicts. Basic and Applied Social Psychology, 31(1), $1-8$. http://doi.org/10.1080/01973530802659687
Wilkes, A. L., \& Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. The Quarterly Journal of Experimental Psychology Section A, 40(2), 361-387. http://doi.org/10.1080/02724988843000168

# Learning Relational Concepts through Unitary versus Compositional Representations 

Daniel Corral (daniel.corral@colorado.edu) \& Matt Jones (mcj@colorado.edu)<br>Department of Psychology and Neuroscience, University Colorado Boulder, Boulder, CO 80309 USA


#### Abstract

Current theories of relational learning on structure mapping emphasize the importance of compositional representations, based on the concept's components and the relations among them. We consider the possibility that relational concepts can also be represented unitarily, whereby the concept is a property of the stimulus as a whole. The distinction between compositional and unitary representations of relational concepts is a natural consequence of structure-mapping theory, but its psychological implications have not been explored. We report two experiments in which we examine how encouraging subjects to represent relational concepts compositionally versus unitarily affects learning on classification- and inference-based category learning tasks. Our findings show that unitary representations lead to better learning than compositional representations, especially for the inference task. We conclude that unitary representations incur less cognitive load than structural alignment of compositional representations, and thus may be the default for everyday relational reasoning.


Keywords: Relational Learning; Relational Structure; Concept Representation; Category Learning; Inference.

## Introduction

On a daily basis, people encounter many complex concepts that are defined by a relational structure - the specific pattern in which two or more objects are bound together by interconnected relations (Corral \& Jones, 2014). For instance, consider a simple scenario in which a dog chases a cat. In this example, the $\operatorname{dog}$ and the cat share a specific relationship with one another, such that it is the dog that fills the role of the chaser and the cat fills the role of being chased (chase (dog, cat)). Critically, this structure is different from a scenario in which a cat chases a dog (chase (cat, dog)). These types of concepts differ from those that are defined by features, which can be identified by the presence of a given set of attributes (Estes, 1986). For example, a bird might be identified by the presence of certain prototypical features, such as \{feathers, beak, wings $\ldots\}$. Although feature-based representations can provide extensive knowledge about a given scenario, they do not convey structural information (Markman, 1999). Thus, a feature-based representation does not allow one to readily distinguish a simple scenario in which a dog chases a cat from an instance in which a cat chases a dog (Markman \& Gentner, 2000), as both would be represented as an unstructured set: $\{d o g$, cat, chase $\}$.

The ability to recognize and reason about structured concepts has been posited to be one of the cornerstones of human cognition (Penn, Holyoak, \& Povinelli, 2008). According to structure-mapping theory, the dominant theory
of relational learning, structured concepts are acquired via structure mapping, wherein the elements of two analogous scenarios are put into alignment in a way that preserves their common roles. For example, in the hypothetical scenarios described below, the dog in the first scenario maps to the cat in the second scenario because both fill the role of the chaser. Alignment of two scenarios highlights their common structure and facilitates abstraction of new relational concepts (Gentner, 1983; Hummel \& Holyoak, 2003).

Importantly, structure-mapping theory ${ }^{1}$ makes the implicit assumption that a relational concept can be represented in two fundamentally different ways: (1) as a system of relations, with meaning derived both from the identities of those relations and from how they are interconnected by shared role-fillers (Corral \& Jones, 2014); or (2) as a primitive, atomic relation that is explicitly represented. We refer to these as compositional and unitary representations. Although this logical distinction has been noted (Gentner, 1983), its potential psychological implications have largely been neglected.

To elaborate further, the first of these representational assumptions is premised on the idea that representations are constructed from two basic types of building blocks: objects and relations. The second assumption is based on the idea that a relation operates on a set of $n$ objects, that is, for every ordered set of $n$ objects, the relation returns a truthvalue indicating whether the objects satisfy the relation. Equivalently, for every ordered set of $n$ objects $\left(o_{1} \ldots, o_{n}\right)$ for which the relation holds, there is an explicit token of that relation: $R\left(o_{1} \ldots, o_{n}\right)$. We refer to any relation of this sort as a unitary relation.

In recent work, Corral, Kurtz, and Jones (under revision) raise the possibility that subjects might indeed represent some relational concepts unitarily, such that the concept is a component or a property of the stimulus as a whole. This type of representation would lack explicit structure and could be recognized directly in a stimulus, similarly to a feature. This idea is perhaps best exemplified in language comprehension, where people appear to seamlessly understand a multitude of rich relational concepts, without explicitly representing their substructure. For example, consider the concept of investigation. An investigation consists of an agent, a given question, the approach the agent takes to answering that question, and the specific

[^310]pattern of interconnections among these components. Nevertheless, people can likely recognize this concept without explicitly representing its structure. Likewise, a $t$ test involves a complex structure of mathematical elements and relations (as many hapless introductory statistics students will attest), but for experienced scientists it is easily conceived of as a unitary event-one can hear the sentence "I ran a $t$-test" and immediately comprehend its meaning without needing to invoke the concept's substructure.

The literature on structure-mapping theory has focused on compositional representations, through its emphasis on the alignment process. Furthermore, it has been proposed that people must use compositional representations in order to learn relational concepts (Markman \& Gentner, 2000). Compositional representations are computationally expensive (Forbus, Gentner, \& Law, 1995) and can place a high strain on working memory (Kintsch \& Bowles, 2002). They are also unnecessary for learning feature-based concepts (Markman, 1999), which can be recognized (without regard to structure) by attending to a stimulus' defining attributes (e.g., Nosofsky, 1986). Similarly, relational concepts that are represented unitarily can be explicitly recognized as a global attribute of the given scenario, and thus can be learned in an unstructured manner. Such representations allow for computationally efficient processing (Forbus et al., 1995), and based on principles of cognitive economy, it follows that people should avoid compositional representations and structural alignment whenever a unitary representation and setwise (featurestyle) comparisons are adequate.

Evidence from related literatures suggests that people in fact do not use compositional representations as much as might be expected based on structure-mapping theory. One prediction that follows from compositional representations is that people should be able to report the structural elements of the relational concepts they are familiar with. However, despite subjects reporting high confidence in their comprehension of various types of common relational systems (e.g., how helicopters fly), they are often mostly unaware of their structural elements (Keil, 2003; Rozenblit \& Keil, 2002). Another prediction from compositional representations is that, because relational structure must be explicitly represented (Kintsch \& Bowles, 2002), it should take longer to comprehend and recognize structured information than information that is not structured. However, various studies have found no differences in the time it takes subjects to comprehend structured (metaphors) and non-structured statements (e.g., "the ball is blue") (Glucksberg, Gildea, Bookin, 1982). Related work has shown that subjects can often understand metaphors automatically, with minimal explicit processing (Glucksberg, 2003). Taken together, these findings suggest that many relational concepts may not typically be represented compositionally.

Due to the representational flexibility that humans possess (Chalmers, French, Hofstadter, 1992), it seems plausible that relational concepts can be represented both unitarily
and compositionally. For instance, a person might represent a concept such as investigation based on a global attribute (e.g., an inspection), but can also likely represent its relational substructure when necessary (explicitly representing the agent, question, line of inquiry, and their interrelations). This idea leads to the question of which type of representation people use by default when learning a relational concept. The main hypothesis of the present paper is that, because unitary representations should allow for more efficient processing, subjects will use such representations when they are available. We test this prediction by giving subjects relational category learning tasks and encouraging them to represent the stimuli either compositionally or unitarily. If people typically learn relational concepts from structural alignment, then encouraging subjects to use compositional representations should aid learning. However, if people instead learn more efficiently with unitary representations, than the opposite outcome should be expected.

Half the subjects in our experiments were given a classification task, in which they were shown a series of stimuli and asked to make categorization judgments. Unitary representations seem especially well-suited for such a task, because they should enable subjects to directly recognize the diagnostic property in a stimulus, just as with feature-based categories. The other subjects were given an inference task, in which they were asked on each trial to determine a missing property of a stimulus that was presented together with its category label. Research with feature-based categories has shown that classification and inference learning tend to yield different category representations, with inference tasks encouraging learning of internal category structure, such as correlations among features (Markman \& Ross, 2003; Yamauchi \& Markman, 2000). This finding suggests that compositional representations should be particularly well-suited for inference learning with relational categories, as such representations highlight the internal structure of stimuli. The inference conditions of our experiments thus provide a more stringent test of our hypothesis that people can learn relational concepts better through unitary representations.

## Experiment 1

Experiment 1 examines how providing unitary and compositional descriptions of relational concepts affects learning on classification and inference tasks (description and task type both manipulated between subjects). Subjects were provided either a unitary or compositional hint at the start of learning and again after every third error, in order to assess whether each type of hint can improve learning. Control groups who were given no hints were also included in order to assess baseline performance in both tasks.

The stimuli used in this study were taken from Corral, et al. (under revision), which were adopted and modeled after those used by Rehder and Ross (2001). A stimulus consisted of three sentences, each of which describes a different component of a machine that works to remove waste
material: (1) the location of where the machine operates, (2) the waste material the machine removes, and (3) the instrument the machine uses.

Stimuli were sampled from two categories: coherent and incoherent. Each category consisted of 18 exemplars. The categories were determined by how a machine's components were related to one another. For exemplars from the coherent category, the machine's instrument is suited for collecting the waste material that the machine works to remove, which can be found in the location where the machine operates. Consider the following example: "Operates on the seafloor, works to remove lost fishing nets, and has a hook." This exemplar is coherent because of the secondary relations among the machine's component parts (presumed to be known by subjects), such that lost fishing nets can be found on the seafloor and a hook can be used to retrieve lost fishing nets. In contrast, exemplars from the incoherent category do not satisfy either of these secondorder relations (i.e., the machine's tool cannot be used to collect the machine's target waste material and that material cannot be found where the machine operates). Non-Morkels were thus made to be as incoherent as possible so as to maximally differentiate the categories and better facilitate learning of the task. Figure 1 illustrates the abstract relational structure of the two categories.

Half of the subjects completed an $A / \neg A$ classification task (in which each stimulus was to be categorized as either a category member or a nonmember), and the other half completed an inference task. On each trial, the subject was presented a single stimulus and asked to make an inference or classification judgment (depending on the condition). After the response, the subject was shown whether the response was correct along with the correct answer.
Operates on the surface of the water
Operates on the surface of the water
Works to clean spilled oil
Works to clean spilled oil
Has a spongy material
Has a spongy material


Coherent Item

Operates on the surface of the water Works to collect dangerous gaseous ions Has a shovel


Incoherent Item

Figure 1. Illustration of the relational structure for items in the coherent and incoherent categories in Experiment 1. The structures differ in that coherent items satisfy the relations indicated by diagonal lines: the machine's implement can remove the target, and the target is found in the machine's location. Recreated from Corral et al. (under revision).

## Method

Two hundred eighteen undergraduates from the University of Colorado Boulder participated for course credit in an introductory psychology course. Subjects were randomly
assigned to six conditions. Type of hint (compositional vs. unitary vs. control) was crossed with task type (classification vs. inference).

Subjects were told that they would be shown short descriptions of various types of cleaning machines, some of which were made by the Morkel Company (coherent category) and some were not (incoherent category). Subjects were provided a positive example of a Morkel (randomly selected) and told that all Morkels share a certain commonality and it was their job to figure out what it was.

Subjects in the unitary condition were shown the following hint: "On each trial try to think about how "well suited" the machine is for performing its task. Keep in mind that consumers say machines from Morkels are built "intuitively" in a way that makes sense." This hint was intended to shift subjects' attention toward finding a global attribute of the stimulus and away from the explicit relationships among its components. Using this hint, it is possible for subjects to learn how to distinguish the categories without explicit knowledge of their relational structure. This hint can therefore be said to encourage subjects to represent each stimulus unitarily.

Subjects in the compositional condition where shown the following hint: "On each trial try to think about the specific manner in which the machine's $1^{\text {st }}$ property relates to its $2^{\text {nd }}$ and $3^{\text {rd }}$ properties, as well as how its $2^{\text {nd }}$ property relates to its $3^{\text {rd }}$ property." This hint was intended to focus subjects' attention on the relationships among the component parts of the stimulus, and thus to encourage them to represent the stimulus compositionally.

Subjects were presented the appropriate hint during the initial task instructions, after the first trial, during rest breaks, and following every third error the subject committed (on a blank screen after corrective feedback was shown). Subjects were asked to read the hint carefully and press the spacebar when they were ready to continue. Subjects in the control group were not shown a hint and were instead asked to continue to try their best; this reminder was presented on every third error the subject committed and on rest breaks.

Each subject completed 72 trials. The order in which the items were presented was randomized for all subjects. In each block of 18 trials, all 18 stimuli appeared in a random order. After each block, subjects were given a self-paced rest break and were shown the proportion of correct responses they answered correctly over those trials, along with the number of trials they had completed and the number that remained.

On each trial in the classification condition, a single, complete stimulus was presented and the subject was asked to type "A" if the machine was a Morkel or "L" if it was not. On each trial in the inference condition, the category label for a stimulus was shown (Morkel or non-Morkel) directly above an incomplete stimulus consisting of two of its three components (i.e., sentences). Below the stimulus were two response options, one of which was the missing component and the other was a lure. The component the
subject was asked to infer (i.e., implement, target material, or location) was randomly selected on each trial. Subjects were asked to select which was the missing component by typing "A" if the correct choice was the top option or "L" if it was the bottom option. The order in which the two options were presented was randomized on every trial. For items that were Morkels, the correct response was the option that shared secondary relations with the given stimulus components. The lure did not share secondary relations with either of the stimulus components. For items that were nonMorkels, the correct response was the component that did not share any secondary relations with either of the stimulus components. The accompanying lure shared at least one secondary relation with one of the stimulus components. Figure 2 shows an example trial from the inference condition.


Figure 2. Example of a stimulus display from the coherent category (Morkels) from the inference task in Experiment 1.

## Results \& Discussion

Figure 3 shows average learning curves for subjects in each group. An ANOVA was conducted to examine differences in performance among groups. The analysis showed a main effect of hint, $F(2,212)=42.14, p<.0001, M S E=.014$, and an interaction, $F(1,212)=8.90, p=.0002, M S E=$ .014 , indicating that the main effect of hint depends on the type of task that subjects completed. On the classification task, control subjects $(M=.61, S E=.017)$ were outperformed by subjects in the compositional ( $M=.775$, $S E=.014 ; p<.0001$ ) and unitary groups $(M=.83, S E=$ $.016 ; p<.0001$ ). In the inference condition, only subjects who received a unitary hint $(M=.716, S E=.012)$ performed better than control subjects $(M=.585, S E=.012$; $p<.0001$ ), as no differences were observed between subjects who were presented a compositional hint ( $M=$ $.587, S E=.011$ ) and subjects in the control group.
Planned $t$-tests were conducted to compare the unitary and compositional groups, separately for each task. On the classification task, subjects in the unitary condition ( $M=$ $.83, S E=.014$ ) outperformed subjects in the compositional condition $(M=.775, S E=.014), t(71)=1.85, p=.068, d=$ .45. This same pattern was observed in the inference condition (unitary $M=.716, S E=.012$; compositional $M=$ $.587, S E=.012), t(67)=5.28, p<.0001, d=1.29$. An
additional 2 (unitary vs. compositional) $\times 2$ (classification vs. inference) ANOVA was conducted, which excluded control subjects. This analysis revealed an interaction, $F(1$, $138)=4.01, p=.047, M S E=.013$, indicating that the unitary advantage was stronger in the inference task than in the classification task.

Taken together, the findings presented here suggest that unitary and compositional representations can both be used to acquire relational concepts. However, subjects who were encouraged to represent the stimuli unitarily showed more robust learning than subjects who were encouraged to represent the stimuli compositionally, especially in the inference task. These findings thus provide support for our main hypothesis that, when both types of representations are available, subjects learn better with unitary than with compositional representations.


Figure 3. Average learning curves and standard errors across blocks of nine trials for each condition in Experiment 1.

## Experiment 2

Experiment 2 builds on the findings from Experiment 1 and examines how category learning is affected when subjects represent a relational concept one way (either unitarily or compositionally) and are subsequently made aware of an alternative representation. Experiment 2 used the stimuli from Experiment 1, and all subjects performed the classification task. All subjects were either provided a unitary or compositional hint prior to the start of learning. For half of the subjects, the hint was changed after the $18^{\text {th }}$ trial (i.e., the unitary hint was replaced with the compositional one and vice versa). For the other half of subjects, the hint they were shown remained the same throughout the study. These latter conditions were identical to the unitary and compositional classification conditions in Experiment 1.

## Method

One hundred fifty-seven subjects were randomly assigned to four conditions: unitary/switch $(N=40)$, compositional/switch $(N=39)$, unitary/no-switch $(N=39)$, and compositional/no-switch $(N=39)$. After the $18^{\text {th }}$ trial (i.e., in the first rest break), the screen was cleared and subjects in the switch conditions were shown a prompt that
notified them that Morkels could be represented differently from the initial hint and were shown the other hint. Following the $19^{\text {th }}$ trial, this hint was presented once more and subjects were reminded to use it to try to figure out what constitutes a Morkel. Subjects in the switch conditions were shown this hint for the remainder of the study (i.e., on rest breaks and following every $3^{\text {rd }}$ error), whereas noswitch subjects continued to see the hint they had seen at the beginning. The rest of the procedure was identical to that of Experiment 1.

## Results \& Discussion

Figure 4 shows average learning curves for subjects in each condition. A $t$-test showed that subjects in the unitary/noswitch condition ( $M=.79, S E=.017$ ) outperformed subjects in the compositional/no-switch condition $(M=.716, S E=$ $.017), t(76)=2.23, p=.03, d=.50$. This finding directly replicates the results from the classification condition in Experiment 1, which showed a unitary learning advantage.


Figure 3. Average learning curves and standard errors across blocks of nine trials for each condition in Experiment 2.

In addition to this analysis, a series of planned comparisons were conducted to examine differences among groups from the point at which subjects were introduced to the other hint (trials 19-72). The first analysis showed that subjects in the compositional/switch condition ( $M=.82, S E$ $=.017$ ) outperformed subjects in the compositional/noswitch condition $(M=.743, S E=.017), t(76)=2.26, p=$ $.027, d=.51$. Additionally, subjects in the unitary/switch condition ( $M=.802, S E=.018$ ) marginally outperformed subjects in the compositional/no-switch condition, $t(77)=$ 1.77, $p=.08, d=.45$. However, no differences in performance were observed among any of the three groups that were presented a unitary hint at some point in the study. Thus, it seems that as long as a unitary hint is presented, regardless of whether it is the only hint that is shown or if it is presented before or after a compositional hint, subjects are able to benefit from it. Taken together, these findings support the conclusion from Experiment 1 and suggest that subjects indeed learn better when they rely on unitary representations.

## General Discussion

We report two experiments that test how encouraging subjects to represent relational stimuli unitarily or compositionally affects concept learning. The findings from Experiment 1 showed that both types of hints can aid learning on a classification task, but only the unitary hint was a useful learning aid on the inference task. These findings provide support for the idea that subjects can indeed use both types of representations to understand and learn relational concepts, but that unitary representations are as or more effective than compositional ones. This latter conclusion challenges the emphasis on compositional representations at the core of most research on analogical reasoning.

Experiment 2 used only a classification task and was able to replicate the findings from the classification condition in Experiment 1, as subjects who received only a unitary hint outperformed subjects who received only a compositional hint. Furthermore, the results from this study showed that subjects who received a unitary hint at any point in the study (with a compositional hint coming before, after, or not at all) outperformed subjects who did not receive a unitary hint at all. No differences in performance were found among subjects in the groups who received a unitary hint. These results lend more support to the dominance of unitary representations, in that subjects will abandon or ignore suggestions for compositional representations if they have discovered a unitary one.

One surprising finding from Experiment 1 was that the unitary advantage was stronger for the inference task than for classification. The effect size for the inference task was actually quite dramatic (Cohen's $d$ of 1.29 ). We had predicted that, if anything, the interaction would go in the opposite direction, given that inference tasks encourage learning the relationships among a concept's components (Markman \& Ross, 2003; Yamauchi \& Markman, 2000). One speculative possibility is that inference learning encourages a top-down approach, in that subjects must reason from the category label to the stimulus, whereas classification encourages a bottom-up approach of reasoning from the stimulus to the category label. Likewise, a unitary representation is top-down in that it embodies a global property of a stimulus that can be used to deduce its internal structure, whereas a compositional representation is bottomup in that the local structure is explicitly represented and the global property emerges only implicitly from the relational system. Under this view, there might be a congruency effect between the stimulus representation and the processes involved in carrying out the task. In particular, a unitary representation might be more congruent with an inference task, because it facilitates conceiving of a concept by a single attribute that can then be used to infer missing parts of a stimulus.

These speculations aside, the main conclusion of the present studies is that, although relational concepts are defined by the interconnections among their component parts, subjects seem to learn these concepts better when they
can be represented unitarily, which might facilitate a global understanding that is easier to discover and use than an explicitly structured one. Furthermore, although compositional-based instruction can help subjects classify a given concept, it might not be optimal for inference-based reasoning.

These findings seem particularly applicable to education and instruction, as they might provide insight into how different types of descriptions for a given relational concept can affect students' representations, as well as how such representations affect learning. Indeed, students are often required to learn various types of structured concepts, and must often engage in both classification and inference. For instance, in mathematics, students must recognize various instantiations of a given problem type, a process that relies on classification, and must also make inferences about how to apply a given solution. These findings thus hold the potential to improve how relational concepts are taught in the classroom.

Furthermore, the present findings have theoretical implications for relational concept learning and representation, and have the potential to affect current theories of analogical reasoning and learning. In particular, research within the theoretical framework of structure mapping (Doumas, Hummel, \& Sandhofer, 2008; Hummel \& Holyoak, 2003) has placed a heavy emphasis on alignment processes operating on compositional representations, but our findings suggest that subjects more naturally represent such concepts unitarily, and that such representations produce a greater and more robust benefit to learning. During comparison of two scenarios, if the critical information can be represented unitarily, then there is no need for structural alignment, because the two can be recognized through the same sort of processing that is possible with feature-based representations, that is, flat (setwise) comparison to identify which properties they have in common. To be clear, this proposal is not intended to argue against the idea that structural alignment of compositional representations plays a prominent role in the more impressive feats of human reasoning (e.g., creativity or scientific discovery), but rather to point out that in more mundane cases, simpler processes and representations may be involved. Nevertheless, further work is necessary to better understand which conditions facilitate unitary and compositional representations.

Lastly, we note one potential shortcoming of the present studies. Although subjects were encouraged to represent the stimuli unitarily or compositionally, we cannot know for certain whether subjects adopted either of these representations. This issue has historically plagued researchers in this domain of study and highlights the need for improved assessment on concept representation. We welcome suggestions in helping us to address this challenge.

## References

Chalmers, D. J., French, R. M., \& Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A
critique of artificial intelligence methodology. Journal of Experimental and Theoretical and Artificial Intelligence, 4, 185-211.
Corral, D., \& Jones, M (2014). The effects of relational structure on analogical learning. Cognition, 132, 280-300.
Corral, D., Kurtz, K. J. \& Jones, M. Learning relational concepts from within- vs. between-category comparisons. Under revision.
Doumas, L. A. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological Review, 115, 1-43.
Estes, W. K. (1986). Array models for category learning. Cognitive Psychology, 18, 500-549.
Forbus, K. D., Gentner, D., \& Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. Cognitive Science, 19, 141-205.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Glucksberg, S. (2003). The psycholinguistics of metaphor. Trends in Cognitive Sciences, 7, 92-96.
Glucksberg, S., Gildea, P., \& Bookin, H. (1982). On understanding nonliteral speech: Can people ignore metaphors? Journal of Verbal Learning and Verbal Behavior, 21, 85-98.
Hummel J. E., \& Holyoak K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological Review, 110, 220-264.
Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. Trends in Cognitive Science, 7, 368-373.
Kintsch, W., \& Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? Metaphor and Symbol, 17, 249-262.
Markman, A. B. (1999). Knowledge representation. Hillsdale, NJ: Erlbaum.
Markman, A. B., \& Gentner, D. (2000). Structure-mapping in the comparison process. American Journal of Psychology, 113, 501-538.
Markman, A. B., \& Ross, B. H. (2003). Category use and category learning. Psychological Bulletin, 129, 592-615.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.
Penn, D. C., Holyoak, K J., \& Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behavioral and Brain Sciences, 31, 109-178.
Rehder, B. \& Ross, B.H. (2001). Abstract coherent concepts. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1261-1275.
Rozenblit, L. R. \& Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26, 521-562.
Yamauchi, T., \& Markman, A. B. (2000). Inference using categories. Journal of Experimental Psychology: Learning, Memory and Cognition, 26, 776-795.

# They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners 

Benjamin R. Cowan (benjamin.cowan@ucd.ie )<br>School of Information \& Communication Studies, University College Dublin<br>Belfield, Dublin 4<br>Holly Branigan (holly.branigan@ed.ac.uk)<br>Department of Psychology, University of Edinburgh<br>7 George Square, Edinburgh, EH8 9JZ<br>Habiba Begum (HXB368@student.bham.ac.uk)<br>HCI Centre, University of Birmingham<br>Edgbaston Campus, B15 2TT<br>Lucy McKenna (lucy.mckenna@adaptcentre.ie)<br>ADAPT Centre, Trinity College Dublin, Dublin 2<br>Eva Szekely (szekely@kth.se)<br>KTH Royal Institute of Technology, Stockholm


#### Abstract

Conversation partners' assumptions about each other's knowledge (their partner models) on a subject are important in spoken interaction. However, little is known about what influences our partner models in spoken interactions with artificial partners. In our experiment we asked people to name 15 British landmarks, and estimate their identifiability to a person as well as an automated conversational agent of either British or American origin. Our results show that people's assumptions about what an artificial partner knows are related to their estimates of what other people are likely to know but they generally estimate artificial partners to have more knowledge in the task than human partners. These findings shed light on the way in which people build partner models of artificial partners. Importantly, they suggest that people use assumptions about what other humans know as a heuristic when assessing an artificial partner's knowledge.


Keywords: knowledge estimation, human-computer interaction, partner modelling, theory of mind, human-computer dialogue

## Introduction

Psycholinguistic research on human-human dialogue (HHD) has shown that our language choices are affected by the assumptions we make about our partners as communicative and social beings (i.e. our partner models) (Branigan, Pickering, Pearson, McLean, \& Brown, 2011): People tend to estimate their conversational partner's knowledge and communicative abilities, and formulate their utterances accordingly. This complex set of judgements is simplified by using a range of heuristics such as accent and social cues (Clark, 1996; Nickerson, 1999) as well as our beliefs about the social distribution of knowledge, i.e., assumptions about what information is likely to be known to whom (e.g., students, residents of Dublin, opticians, birdwatchers) (Fussell \& Krauss, 1992a).

Such perspective taking is critical to successful communication and is not solely the preserve of HHD. People consistently perceive the flexibility and ability of automated artificial (computer) partners as far lower than those of a human dialogue partner, leading us to categorise them as 'at risk' listeners in dialogue (Oviatt, MacEachern, \& Levow, 1998). Moreover, our initial expectations about artificial partner's abilities affect our language choices in Human-Computer Dialogue (HCD) (Branigan et al., 2011; Edlund, Gustafson, Heldner, \& Hjalmarsson, 2008). Yet we know little of how people come to have these expectations: What factors impact people's preconceptions and expectations about what an artificial partner is likely to know, before they have even begun to interact with it? In other words, what determines people's initial partner models for artificial partners?

Understanding what governs and impacts our partner models when interacting with artificial partners, especially in speech-based interactions, has important theoretical and applied implications (e.g., in developing robust and effective speech-based interfaces). In this paper, we investigate whether our initial assumptions about what an artificial speech-based interaction partner knows are related to the sense we have of the social distribution of knowledge. We also look at how our initial beliefs about partner knowledge are influenced by (1) partner type (humans vs. artificial) as well as (2) the partner's signalled nationality.

## Perspective-Taking in Dialogue

Imagine that a stranger asks for directions to a local landmark. How do we ensure that the information we include and the language we use to communicate the message is appropriate for them? Research suggests that we
use verbal and non-verbal cues to assess our conversational partner's characteristics, e.g. where they are from, their language proficiency, their age, profession etc., and use these cues to construct a partner model to guide our language choices (Nickerson, 1999).

This initial global partner model (Brennan, Galati, \& Kuhlen, 2010), which is consulted at the stage of initial interaction, is formed through relatively superficial cues (e.g., stereotypes and pre-conceived expectations and assumptions that are in place prior to the dialogue) and assumptions about the social distribution of knowledge (Fussell \& Krauss, 1992a). These initial inferences act to give a speaker an initial model of common ground between interlocutors, i.e., a representation of mutual knowledge, assumptions and beliefs shared between the interlocutors in a conversation, crucial to successful and effective communication (Bromme, Rambow, \& Nückles, 2001; Clark, 1996). Although our partner models may be subsequently updated by local experiences within the dialogue interaction (e.g. feedback about comprehension, via verbal and non-verbal cues) (Brennan et al., 2010), the global model acts as a guide for our initial interaction, especially before feedback has been gathered from within the dialogue (Fussell \& Krauss, 1992a).

Understanding how we develop and form these initial models is important, as research shows that they guide our language choices. We tend to adjust our language based on our assumptions about our addressees' knowledge. For instance, when people are asked to describe items for their friends, they adapt their descriptions to their friend's knowledge - and these adjustments lead to better communication, i.e., higher accuracy in identification (Fussell \& Krauss, 1989). Crucially, studies also show that we are very accurate at assessing others' knowledge and that these assessments guide how we construct our initial message in communication (Fussell \& Krauss, 1991, 1992a).

Similar effects of partner models on language choice are thought to drive our dialogue interactions with artificial dialogue partners. People tend to see artificial partners as poorer interlocutors and alter their language choices and speech behaviours as a result (Branigan et al., 2011; Oviatt, Bernard, \& Levow, 1998). For example, people are more likely to converge (or align) with their partner's choice of referring expression when they believe their partner to be a computer rather than a human. In addition, they adjust their behaviour more in this way when they are led to believe that the artificial partner is a 'basic' interlocutor with restricted capability than a partner with more advanced capability (Branigan et al., 2011). Similarly, people's linguistic choices in a telephone conversation concerning air-fares and timetables change depending on whether they believe their partner to be a human or a computer (Amalberti, Carbonell, \& Falzon, 1993). Similar findings have been reported in
other work (Bell \& Gustafson, 1999; Kennedy, Wilkes, Elder, \& Murray, 1988). Compared to HHD, users tend to use simpler grammatical structures, use more words in their descriptions, use fewer pronominal anaphors (e.g. her/him; $h e / s h e$ ), and use simpler lexical choices (Amalberti et al., 1993; Kennedy et al., 1988). Such research assumes that people's perceptions and beliefs about their partner's abilities affect their language choices in these contexts. Yet it is not clear what factors determine these beliefs in the first place, and thus what may be driving people's global partner model during their initial interaction with an artificial partner. Our work aims to shed light on this question.

Research on robotic agents has shown that the perceived nationality of the agent, and the content that it is being asked to process, both influence participants' judgements about its abilities (Lee, Lau, Kiesler, \& Chiu, 2005). Participants used these cues in a similar way to that which they are used in HHD: When they were asked to judge the likelihood that a robot 'from New York' or 'from Hong Kong' would know and recognize a set of New York and Hong Kong landmarks, they judged that the robot would be more likely to identify landmarks associated with its perceived nationality (Lee et al., 2005). In this context, accent can play an important role. It acts as a strong signal of identity and a speaker's linguistic background (Ikeno \& Hansen, 2007), and allows listeners to identify characteristics such as age, gender and geographic affiliation, as well as stimulating specific stereotypes (Ryan, Giles, \& Sebastian, 1982).

## Research Aims and Hypotheses

There is currently little understanding of what factors affect people's assumptions about partner knowledge and abilities in HCD contexts. The limited existing research on people's perceptions of artificial dialogue partners tends to focus on affective factors such as interface likeability rather than on assumptions about a computer's knowledge and abilities. Other work in tangential fields such as HRI cannot be assumed to hold more widely as the embodiment of robots tend to facilitate the mapping of human abilities to a robot partner (Kiesler, 2005).

We present a study using a similar method to previous work investigating how people initially estimate human partners' knowledge (Fussell \& Krauss, 1992a), in order to investigate how people estimate artificial partners' knowledge. People are asked to name landmarks and judge the identifiability of those landmarks' names to others. We hypothesise that people will use the same heuristics to estimate partner knowledge for artificial partners as they use for human partners. That is, people will rate both human and artificial partners as more likely to know the name of those landmarks that are generally more accurately identified by other people (H1). This would be evidence that people have a sense of the spread of knowledge about a topic in the population (i.e., the social distribution of knowledge) with
this being related to their assessment of a partners’ likely knowledge, including artificial partners. We also expect a strong positive correlation between judgements of humans' and artificial partners' knowledge (H2), giving support to the idea that our judgements of artificial agents are related to our judgements of humans in this context. Based on the intimated difference in partner models between humans and artificial partners in the literature we also hypothesise that there will be a statistically significant difference between people's judgements of how likely a person versus an artificial agent is to know the name of the stimuli (H3). We also hypothesise that people will make different judgements about partner knowledge based on the relation between the system's signalled nationality (UK or US) and the type of content being judged in the experiment (i.e., UK landmarks) (H4).

## METHOD

## Participants

32 ( 16 F, 16 M) Native British English speakers with a mean age of 32.0 years (S.D.=12.1) from a UK university community took part in the study. The majority ( $\mathrm{N}=26$ ) of participants had previously spoken to an automated system. Those who had used such systems were asked to rate how frequently they used them on a 7 point Likert scale (Very Infrequently-Very Frequently). The mean rating suggests that their level of experience with these types of interfaces was low ( $\mathrm{M}=2.73$, $\mathrm{SD}=1.43$ ).

## Items

Fifteen UK landmarks were used as the stimuli in the study, selected based on the frequency of accurate naming in a prestudy. This was to ensure that there was variation in the frequency of accurate naming across the items in the experiment.

## Conditions

Partner Type All participants were asked to judge both an artificial partner's (i.e. automated agent) and a human partner's (within participants) likely knowledge of the landmark names. The order in which participants were asked to judge the artificial and the human partner was randomised. Participants judged all 15 landmarks in each condition. The display of the 15 landmarks in each partner condition was randomised to reduce potential order effects.

Nationality Participants were asked to judge how likely either an American $(\mathrm{N}=14)$ or British $(\mathrm{N}=18)$ partner (between participants- randomly assigned) would be to know the landmarks. When in the human partner condition, participants were asked to rate how identifiable the landmarks' names would be to either a British or American person (participants were told that 'identifiable' referred to the likelihood of knowing the landmark name). When in the artificial partner condition, participants were told that the
researchers were developing a British-based (British nationality condition) or a US-based (US nationality condition) automated agent. They then listened to a sample audio clip taken from the system. Participants listened to a sample audio introduction from the agent (e.g. "Hello, my name is Laura. How can I help you?"), simulating the type of content that would guide people's initial partner models in these types of interactions. To further emphasise the nationality, the introductory message from the service was played in either a British or a US accent. This procedure was used to make sure that participants who lacked previous experience with agents had a frame of reference for their ratings.

## Measures

Participant's ability to name landmarks To identify the spread of knowledge within the sample, all participants were initially asked to name the 15 landmarks used in the study. A $300 \times 250$ pixel image of each landmark was displayed along with a textbox. Participants were asked to name the item. They were informed that if they did not know the name of the item they could leave this box blank. The lead author then marked the names given by the participants as either accurate or inaccurate.

Others' knowledge of the landmark names Based on scales used in previous research on perception of others' knowledge in HHD (Fussell \& Krauss, 1992b) and humanrobot interaction (HRI) (Lee et al., 2005), participants were asked to judge how identifiable they felt the name of each landmark would be to others. This was measured using a 7point Likert scale from Not Identifiable (1) to Very Identifiable (7).

## Procedure

Participants were recruited via email from a British university community. Upon responding to the email participants were sent a link to the online survey. Participants completed the demographic section of the survey. They were then asked to name the 15 landmarks, and subsequently asked to judge how identifiable the name of the landmarks would be to a human (either a British or US person), and then how identifiable the name of the landmark would be to a computer (either British or US accented automated agent). Again, the order of these was randomised. They were then debriefed as to the purpose of the experiment.

## RESULTS

## Social Distribution of Knowledge

Following previous work on knowledge estimation in HHD (Bromme et al., 2001; Fussell \& Krauss, 1992b) we ran analysis on the item level data to test H1 and 2. Using the item level data means we can see whether landmarks that were more accurately named across the sample were rated
as more likely to be known to both human and artificial partners. This would give us a sense of how people's assumptions of knowledge for each item relate to actual levels of knowledge in the group of participants for each item. This type of fine grained insight would not be possible using the participant level data as we would only have a measure of accuracy for each participant, giving us no sense of the spread of knowledge of each item in the sample as a whole.


Figure 1: Relationship between percentage accurate item naming and human partner identifiability rating

There was a strong positive correlation between the percentage of accurate responses for an item and participants' mean judgements of other people's [r (13)= .85, p $<.001$ ] (Figure 1) as well as an artificial partner's knowledge of its name [r (13)=.86, p<.001] (Figure 2). There was also a strong positive correlation between judgments of other people's knowledge of the names and an artificial partner's knowledge $[\mathrm{r}(13)=.78, \mathrm{p}<.001]$ (Figure $3)$.


Figure 2: Relationship between percentage accurate naming and artificial partner identifiability rating

These correlations support our hypotheses (H1 and 2). They suggest that people have relatively accurate awareness of the actual distribution of knowledge (with respect to which
knowledge is more or less likely to be known) and that this has a strong relationship to their judgements of how likely the name is to be known to a person and an artificial partner.


Figure 3: Relationship between human and artificial partner identifiability ratings

Moreover, people's assessment of how identifiable a landmark's name is to an artificial partner seems related to how identifiable they believe it is to a human partner. This supports the idea that people's initial model of an artificial partner's knowledge is related to their initial model of other people's knowledge, with both closely reflecting people's actual rates of accuracy in naming each item.

## The Effect of Partner Type \& Nationality

To test H3 and H4, we analysed the data at the participant level using a $2 \times 2$ Mixed ANOVA looking at the effects of partner type (Human vs. Artificial -within participants) and nationality (US vs. British- between participants) on people's knowledge estimation. We saw a statistically significant main effect of partner type on people's knowledge estimations $\left[\mathrm{F}(1,30)=6.43, \mathrm{p}=.016, \eta_{\mathrm{G}}^{2}=\right.$ $0.058]$. People rated item names in general to be more identifiable to an artificial partner ( $\mathrm{M}=4.60$, S.D. $=1.06$ ) than to a human partner $(M=4.19$, S.D. $=0.74)$, supporting our hypothesis but contradicting the direction intimated by previous HCD work. There was no statistically significant main effect of nationality $[\mathrm{F}(1,30)=0.31, \mathrm{p}=.58$, $\left.\eta^{2}{ }_{G}=0.007\right]$ or interaction effect between partner type and nationality $\left[\mathrm{F}(1,30)=2.94, \mathrm{p}=.097, \eta_{\mathrm{G}}{ }^{=}=0.028\right]$. Therefore a partner's nationality did not affect people's knowledge judgements of human or artificial partners in relation to the landmarks; H4 was therefore not supported.

## DISCUSSION

We found that people have a strong sense of the social distribution of knowledge and this relates to people's judgements about others' knowledge, irrespective of the other being an artificial agent or a human. The number of times each item was named correctly correlated strongly and
positively with people's estimations of both artificial and human partners' knowledge of landmark names. We also found that people in general judged the names of the landmarks in the experiment to be more identifiable to a computer than a person. Surprisingly, partner nationality did not have statistically significant effects on knowledge estimation.

Our research highlights that people are relatively accurate at estimating what other people are likely to know based on a sense of the general distribution of that knowledge, similar to previous research (Fussell \& Krauss, 1992b; Lau, Chiu, \& Hong, 2001). But importantly, these effects also apply to our estimates of artificial partners' knowledge. The actual percentages of correct responses for each item correlated highly and positively with the knowledge estimates for both artificial and human partners. We therefore seem to use our estimates of what other people will know to inform our judgements of what an artificial partner will likely know. That is, people seem to use their perceptions of the social distribution of knowledge among humans to anchor their perceptions of an artificial partner's knowledge.

We also see that people judged an artificial partner as being more likely to know the name of the landmark in the study than a human partner. It is important to note that our finding may reflect users' assumptions about one specific dimension of an artificial partner's abilities (i.e., their knowledge of proper names) rather than their communicative capabilities or knowledge as a whole. Participants were asked to judge how identifiable the name of a landmark (e.g., Stonehenge) would be. Proper names pick out unique entities in the world. As such, they do not require any complex inferencing, knowledge of ontologies, conceptual relations between categories. They can (usually) be captured by a simple association between the name and a unique object, the kind of data that are prototypically perceived as easy for computer systems to store, index, and retrieve. This may explain why a computer was judged more likely than a human to know the name of the landmarks that we used. Other types of knowledge that involve more complex conceptual relationships, or operations over elements might not show the same pattern. Note however that people did not attribute complete omniscience to the artificial partner; their judgements about its knowledge were strongly related to the social distribution of knowledge.

There is also likely to be a distinction between what we perceive artificial partners to know and what we believe they can do with this knowledge in dialogue, or even whether these names will be recognised effectively in the first place. For instance people may assume that artificial partners know the proper names of landmarks but may not be sufficiently confident that these names will be recognised during speech recognition. Although vast improvements on error rates have been made in speech technology research, there may still be a perception within people's partner
models that recognition is poor and inflexible. Hence rather than artificial partners being seen as 'at risk' dialogue actors, people's partner models are likely more nuanced and multi-dimensional, presumably encompassing assumptions about both underlying knowledge and processing abilities.

To be clear, this study focused on how people establish estimates of knowledge in their initial global partner models, in the absence of dialogue interaction with the system. Our findings are particularly relevant to how people form a priori partner knowledge assumptions in a dialogue context. Yet when in dialogue, our perspective taking is likely to be informed by both the global models we create of our partner (e.g. assumptions of their knowledge and abilities formed by stereotypes and expectations before interaction) and local experiences within the dialogue (e.g. feedback of comprehension via verbal and non verbal cues) (Brennan et al., 2010). Indeed these factors are likely to interact in dialogue interactions. Work on HHD interaction has shown that behaviours within a dialogue that do not match our expected partner models impact our speech (Kuhlen \& Brennan, 2010). Research suggests that these models should be considered as being dynamic and adaptable over time (Fussell \& Krauss, 1991; Nickerson, 1999). Investigating the dynamism of partner models across the course of an interaction is a critical issue for future research in HCD as it has been in HHD.

In addition, although partner models are assumed to be important in influencing people's language choices and linguistic processing in HCD (Edlund et al., 2008), more research is needed to fully explore the role that they play. This question has received considerable attention in research on HHD, with particular reference to the extent to which our partner models impact processing: Is their influence immediate and pervasive, or delayed and restricted? (see Brennan et al., (2010) for summary of the main theoretical positions). Within HCD research, partner models have been invoked to explain the differences in language use between HHD and HCD (Branigan et al., 2011; Edlund et al., 2008), but recent research has shown that this may not be true in all contexts (Cowan \& Branigan, 2015; Cowan, Branigan, Bugis, Obregon, \& Beale, 2015). Clearly, partner models affect language choice and processing in both HHD and HCD - but it is not yet clear whether they do so in the same ways and to the same extent. An interesting possibility for future research is that partner models may play a more pervasive and far-reaching role in HCD than in HHD.

## Implications \& Conclusions

Our research set out to investigate the factors that affect people's expectations about what an artificial partner is likely to know, before they have begun to interact with it. Our findings suggest that we come to interactions with an existing presumption of what an artificial partner is likely to know that is based on assumptions of how knowledge is socially distributed. Moreover we found that under some
circumstances they may have the preconception that an artificial partner knows more than a human partner. These results suggest that models of human-human communication are applicable in important ways to communication with artificial agents. They also have important applied implications for HCD, by casting light on factors that can lead users towards or away from an appropriate mental model of a partner's abilities and intentions, with implications for successful communication (Kiesler, 2005). When designing artificial systems, developers should be aware that people bring with them assumptions about the social distribution of knowledge, which could significantly affect their interaction.

## References

Amalberti, R., Carbonell, N., \& Falzon, P. (1993). User representation of computer systems in humancomputer speech interaction. International Journal of Man-Machine Studies, 38, 547-566.
Bell, L., \& Gustafson, J. (1999). Interaction with an animated agent in a spoken dialogue system. In Proceedings of the Sixth European Conference on Speech Communication and Technology (pp. 1143-1146). Budapest, Hungary: ISCA.
Branigan, H. P., Pickering, M. J., Pearson, J. M., McLean, J. F., \& Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. Cognition, 121(1), 41-57.
Brennan, S., Galati, A., \& Kuhlen, A. (2010). Two Minds, One Dialog: Coordinating Speaking and Understanding. In B. Ross (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory (Vol. 53, pp. 301-344). Elsevier.
Bromme, R., Rambow, R., \& Nückles, M. (2001). Expertise and estimating what other people know: The influence of professional experience and type of knowledge. Journal of Experimental Psychology: Applied, 7(4), 317-330.
Clark, H. H. (1996). Using Language. Cambridge University Press.
Cowan, B. R., \& Branigan, H. P. (2015). Does voice anthropomorphism affect lexical alignment in speech-based human- computer dialogue? In Proceedings of Interspeech 2015. Dresden, Germany: ISCA.
Cowan, B. R., Branigan, H. P., Bugis, E., Obregon, M., \& Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic alignment in human-computer dialogue. International Journal of Human-Computer Studies, 83, 27-42.
Edlund, J., Gustafson, J., Heldner, M., \& Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. Speech Communication, 50(8-9), 630645.

Fussell, S. R., \& Krauss, R. M. (1989). Understanding friends and strangers: The effects of audience
design on message comprehension. European Journal of Social Psychology, 19, 509-525.
Fussell, S. R., \& Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. European Journal of Social Psychology, 21, 445-454.
Fussell, S. R., \& Krauss, R. M. (1992a). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. Journal of Personality and Social Psychology, 62(3), 378391.

Fussell, S. R., \& Krauss, R. M. (1992b). Coordination of knowledge in communication: Effects of speakers’ assumptions about what others know. Journal of Personality and Social Psychology, 62(3), 378391.

Ikeno, A., \& Hansen, J. H. L. (2007). The Effect of Listener Accent Background on Accent Perception and Comprehension. EURASIP J. Audio Speech Music Process., 2007(3), 4:1-4:8.
Kennedy, A., Wilkes, A., Elder, L., \& Murray, W. S. (1988). Dialogue with machines. Cognition, 30(1), 37-72.
Kiesler, S. (2005). Fostering common ground in humanrobot interaction. In IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005 (pp. 729-734).

Kuhlen, A. K., \& Brennan, S. E. (2010). Anticipating Distracted Addressees: How Speakers' Expectations and Addressees' Feedback Influence Storytelling. Discourse Processes, 47(7), 567-587.
Lau, I. Y.-M., Chiu, C., \& Hong, Y. (2001). I Know What You Know: Assumptions About Others' Knowledge and Their Effects on Message Construction. Social Cognition, 19(6), 587-600.
Lee, S., Lau, I., Kiesler, S., \& Chiu, C.-Y. (2005). Human Mental Models of Humanoid Robots. Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on, 27672772.
https://doi.org/10.1109/ROBOT.2005.1570532
Nickerson, R. S. (1999). How we know-and sometimes misjudge-what others know: Imputing one's own knowledge to others. Psychological Bulletin, 125(6), 737-759.
Oviatt, S., Bernard, J., \& Levow, G. A. (1998). Linguistic adaptations during spoken and multimodal error resolution. Language and Speech, 41 ( Pt 3-4), 419-442.
Oviatt, S., MacEachern, M., \& Levow, G.-A. (1998). Predicting hyperarticulate speech during humancomputer error resolution. Speech Communication, 24(2), 87-110.
Ryan, E. B., Giles, H., \& Sebastian, R. J. (1982). An integrative perspective for the study of attitudes towards language variation. In Attitudes Towards Language: Social and Applied Contexts. London: Arnold.

# Folk Attributions of Control and Intentionality Over Mental States 

Corey Cusimano (cusimano@sas.upenn.edu)<br>Geoffrey P. Goodwin (ggoodwin@psych.upenn.edu)<br>Department of Psychology, University of Pennsylvania, 3720 Walnut St Philadelphia, PA 19104 USA


#### Abstract

Influential theories in social psychology, philosophy, and linguistics assume that ordinary people judge many mental states as outside voluntary control, yet few studies have directly investigated these claims. We report four studies suggesting that, contrary to several prominent models, ordinary people attribute at least moderate intentional control to others over a wide variety of mental states. Furthermore, it appears that perceived control may vary systematically according to mental state type (e.g. emotions vs. desires vs. beliefs). These results point to several important directions for future research in behavior explanation and moral judgment.


Keywords: mental states; control; intentionality; agency

## Introduction

Mental states are not just used to explain and predict others' observable behavior, they are also often treated much like behaviors themselves in that we talk about them, care about them, and try to influence them (Frankfurt, 2004). For instance, when we learn that someone doesn't like us or respect us, it hurts and we feel angry (Leary, Springer, Negel, Ansell, \& Evans, 1998). When we dislike certain attitudes, either our own or someone else's, we try to change them (DeMarree, Wheeler, Briñol, \& Petty, 2014). And when we're exposed to someone's highly immoral emotions, desires, and thoughts, we form negative impressions of that person and try to avoid them (Ames \& Johar, 2009; Gromet, Goodwin, \& Goodman, 2016; Cohen \& Rozin, 2001).

Despite the importance of mental state evaluation and regulation in social life, very little research has studied how ordinary people think about others' agency over their own minds. This omission is particularly striking in light of years of research demonstrating that perceptions of behavioral control predict judgments of blame and responsibility, feelings of anger or pity, and helping or punishing behavior (see, Alicke, 2000; Malle, Guglielmo, \& Monroe, 2014; Weiner, 1995 for reviews).

One possible reason for this omission is a long-held assumption, based on early work in linguistics, that mental states are perceived as involuntary. In one of the first investigations on this topic, Katz and Postal (1964) argued, based on the observation that mental state verbs seem to be ungrammatical in the imperative form (e.g. compare "Want this pear!" and "Pick up this pear!"), that "being in such psychological states as belief, understanding, wanting and hoping is not subject to a person's will" (p. 77; see also Miller \& Johnson-Laird, 1978). Despite other work arguing that many mental states can be used in the imperative form (e.g. Huddleston, 1970), many linguists continued to assume that
mental states are involuntary (e.g., Brown \& Fish, 1983; Corrigan, 1988).

This work also influenced social psychology. For instance, Gilovich and Regan (1986) assumed that, unlike actions, mental states "do not necessarily involve any choice on the part of the person from among alternatives; they just happen" (p. 349). Similarly, Malle and Knobe (1997a) took as given that "prototypical actions... are both intentional and observable, whereas prototypical experiences (e.g. 'Ben is excited') are both unintentional and unobservable" (p. 289; emphasis added). These claims about ordinary attributions of intentionality and voluntariness play an important role in psychological models of behavior explanation. For instance, Malle \& Knobe (1997a) argued that people will be less motivated to try and explain others' mental states because they are unintentional. Gilovich \& Regan (1986) argued that, because mental states are involuntary and uncontrollable, people offer more dispositional (as opposed to situational) explanations for them (see also Lock \& Pennington, 1982). Finally, Malle (2004)'s model of behavior explanation posits that people provide mechanistic cause explanations (as opposed to reason explanations) for emotions, desires, beliefs, and other mental states, on the premise that ordinary people judge these as unintentional.

Scholars in philosophy and anthropology also frequently make assumptions about ordinary people's judgments of mental state voluntarism, but they often differentiate between mental states types. In his 'folk model of mind', D'Andrade (1987) claimed, similarly to Katz and Postal (1964), that people view desires as entirely involuntary and uncontrollable. However, D'Andrade (1987) also claimed that people view emotions as somewhat controllable and beliefs as highly controllable. A similar pattern emerges in philosophical theories. For instance, the idea that ordinary people judge beliefs as voluntary is echoed in Alston (1988), who invoked this point to explain why people blame each other for unjustified beliefs, while the idea that people view desires and other attitudes as involuntary is common in moral philosophy (see, e.g. Adams, 1985; Smith, 2008).

Despite the ubiquity of claims about the perceived controllability of mental states, only a handful of studies have been conducted which directly ask people about how they perceive them, and the evidence from these studies conflicts. In one study, Malle and Knobe (1997b) asked participants to rate the intentionality of 20 behaviors, three of which were mental states (e.g. "Anne was in a great mood"). Each of those mental states was rated low in intentionality (e.g. $M=$ 2.54 on a 1-7 scale). In contrast, Schlesinger (1992) asked people to rate how much control (Studies 1-4, 6) or
intentionality (Study 5) the experiencer of a mental state had over that state and found that people attributed a moderately high degree of control and intentionality (e.g. $M=4.58$ on a 1-7 scale, Study 6). However, Schlesinger (1992) relied on scenarios that were interpersonal and highly abstract (e.g. "A fears B"), making it unclear whether the results reflect prototypical attributional processes.

In light of this, we sought to test how much control and intentionality people typically attribute to others over their mental states. We improved over prior studies by testing a wide range of mental states and, following the possibility raised by D'Andrade (1987) and others, testing for differences between mental state categories. In the studies below, we also include observable behaviors including intentional acts (e.g. talk, avoid), accidents (e.g. slip, fall), and uncontrollable behaviors (e.g. sneeze, shiver), as foils. These foils acted as benchmarks, allowing us to test how judgments of mental states compared with judgments of prototypical controllable and uncontrollable behaviors, while also ensuring that participants used the control measure concepts in a predictable way.

## Study 1

The purpose of Study 1 was to measure prototypical judgments of control for a variety of mental states, and to compare these judgments with those for clearly intentional, unintentional, and uncontrollable behaviors. To obtain ecologically valid materials, we solicited vignettes from one sample of our population (University of Pennsylvania undergraduates), selected frequent examples, and then used them in a rating task for a separate sample from the same population.

## Methods

Stimulus generation and selection. We solicited stimuli for 43 items in total. These items consisted of 28 mental states, including four beliefs (believe that, conclude that, feel that, think that), four desires (crave, desire, hope, want), four emotions (anger, anxiety, embarrassment, happiness), four intentions (goal, intend, plan, resolve), four deliberations (consider, deliberate, speculate, think about), four evaluations (value, love, hate, appreciate), two imaginations (imagine, visualize), and two memory events (forget, remember). In addition to these 28 mental states, we included five intentional acts (play with, eat, say, search for, avoid), five accidents (fall off of, trip over, slip on, run into, drop), and five uncontrollable behaviors (sneeze, yawn, sweat, shiver, faint) as our foils.

80 University of Pennsylvania students participated (57 female) in a sentence completion task for course credit. Participants were provided with sentence fragments containing an ambiguous subject and a mental (or behavioral) verb, but no object (e.g. "He believed that...", "She wanted...", "He intended to..."). They were instructed to complete each sentence fragment in a way that made sense given the words provided and to avoid humor. The mental states were split across five lists and combined with
observable behaviors and 12-13 filler trials. Participants were randomly assigned to one of these lists, yielding 13-17 contents per item.

As expected, many of the topics participants wrote about were relevant to their lives as undergraduate students, including concerns about school (e.g. "She felt anxious about her upcoming exam", "He planned to do better on the next test"), romantic relationships (e.g. "She felt angry with her boyfriend", "She thought that she wasn't good enough for him"), and food (e.g. "He craved chocolate", "She thought about the lunch she would be having soon"). Similar responses appeared across item categories.

We selected five completions for each of the 28 mental states and 15 behavior foils, yielding 215 scenarios total, based on the frequency of similar completions. For items that produced few or no duplicate responses, we selected responses so as to maximize the diversity of content.

Main rating task. 143 University of Pennsylvania students (94 female) were recruited for an experiment about "understanding others' behavior" and completed the task for course credit.

The 215 scenarios were distributed across five lists. Each list contained one scenario from each of the 28 mental states and 15 behaviors yielding 43 trials total. Scenarios were presented on separate pages in a random order. For each trial participants responded to eight questions which, to avoid possible order effects, were presented in a new random order for each trial.

Four questions assessed how much agency participants attributed to the agent for the ascribed mental state. Two of these questions assessed general control: (1) How much control the agent had over that behavior; and (2) Whether, if desired, the agent could have done otherwise. The other two probed intentionality: (3) Whether the agent acted intentionally; and (4) Whether the agent chose to act/think/feel (etc.) that way.

Two questions probed participants' evaluations of the mental state, including (5) How good or bad the agent's behavior was; and (6) Whether the agent should have behaved in the manner described. Two final questions probed judgments of agent themselves: (7) How responsible the agent was for the behavior or mental state; and (8) How revealing it was of the agent. All questions used a 7-point rating scale.

To minimize ambiguity, all questions contained explicit reference to the mental or physical content (e.g., "How much control did she have over believing that she did well on the exam?"). At the end of the experiment, participants reported demographic variables including age, sex, political orientation, religiosity, and religious affiliation.

## Results

We combined our two control ( $1-2, r \mathrm{~s}=0.74-0.81$ ), and two intentionality (3-4, $r s=0.83-0.89$ ) measures into single measures of control and intentionality (Table 1 shows means and standard deviations for each behavior and mental state


Figure 1: Mean ratings (and standard errors) for each of the 43 item categories (black) with mean rating for each of the five scenarios (color) from Study 1. Each shape represents the mean one of the five scenarios.
category). To test degrees of agency, we ran a series of mixed-effect linear models comparing mental state categories to uncontrolled, accidental, or intentional behaviors, on bysubject means for each category.

Participants used the control concepts expected: accidents were seen as more controlled ( $b=0.78, S E=0.11, t=7.26$, $p<0.001$ ) but not more intentional ( $b=-0.15, S E=0.11, t$ $=-1.39, p=0$. 165) than uncontrollable behaviors. With two exceptions, mental states were seen as more controlled and intentional than the uncontrollable and unintentional behaviors ( $p s<0.001$ ). Only intentions were not see as less controlled ( $p=0.338$ ) or intentional ( $p=0.832$ ) than intentional action foils (see Figure 1).

We also conducted a set of exploratory analyses to test whether any mental state categories were significantly different from one another. We conducted a regression comparing each mental state category to its adjacent category based on the overall control and intentionality means. Results showed that, on average, most mental state category ratings were different from their adjacent category: emotions were less intentional than desires $(b=0.79, S E=0.08, t=10.07, p$

Table 1: Means (and SD) for agency responses in Study 1

| Behavior | Control | Intentionality | Responsibility |
| :--- | :--- | :--- | :--- |
| Uncontrolled Act | $2.51(1.53)$ | $2.13(1.40)$ | $2.66(1.70)$ |
| Accident | $3.30(1.51)$ | $2.29(1.41)$ | $3.72(1.73)$ |
| Emotion | $3.71(1.51)$ | $3.32(1.58)$ | $3.89(1.68)$ |
| Memory | $3.53(1.52)$ | $2.94(1.52)$ | $4.13(1.67)$ |
| Desire | $4.03(1.63)$ | $4.11(1.71)$ | $4.28(1.70)$ |
| Evaluation | $4.59(1.63)$ | $4.62(1.64)$ | $4.70(1.65)$ |
| Belief | $4.54(1.56)$ | $4.50(1.61)$ | $4.70(1.60$ |
| Deliberation | $4.99(1.43)$ | $5.04(1.39)$ | $5.07(1.43)$ |
| Imagination | $5.05(1.35)$ | $5.16(1.37)$ | $5.03(1.41)$ |
| Intention | $5.88(1.22)$ | $5.97(1.15)$ | $5.89(1.19)$ |
| Intentional Act | $5.98(1.18)$ | $5.99(1.13)$ | $5.86(1.28)$ |

$<0.001$ ), desires were less intentional than beliefs ( $b=0.39$, $\mathrm{SE}=0.08, t=4.92, p<0.001)$. Beliefs and evaluations were not significantly different from each other ( $b=0.11, S E=$ $0.08, t=1.42, p=0.157$ ), however evaluations were different from deliberations ( $b=0.44, S E=0.08, t=5.58, p<0.001$ ). This pattern was replicated in participants' control ratings.
Finally, we examined item-level means for each of the 28 mental state concepts and found that judgments of control and intentionality were highly correlated with one another $(r(26)$ $=0.95)$, and with judgments of responsibility $(r(26)=0.98$ and $r(26)=0.92$, respectively).

## Discussion

Results from this study provide evidence that, contrary to the theories cited above, many ordinary mental states are perceived to be moderately controllable and intentional. It also suggests that perceived agency might differ as a function of the type of mental state: emotions were judged as less voluntary than desires, beliefs, and other states.
However, this study has several notable shortcomings. First, the mental state scenarios were presented without the immediate context in which they occurred. It is possible that when possible proximate or situational causes for mental states are made salient, perceived control and choice is diminished. Second, the nature of our design was such that the content of the mental states was not held constant across mental state type: desires tended to be "about" different things than beliefs, evaluations, and so on. It is therefore possible that differences in control were due to what the mental states were about. Study 2 was designed to address these limitations.

## Study 2

Study 2 used a set of experimenter-generated stimuli to investigate judgments of three different measures of
voluntariness with more detailed and comparable vignettes. Because control measures were divided between subjects, Study 2 comprised three separate experiments: Study 2a investigated judgments of intentional choice, Study 2b investigated general control, and Study 2c investigated the ability to choose to stop thinking, feeling, or wanting something once it has started.

We predicted that, despite the additional constraints imposed in this experiment (see below), participants would still view mental states as moderately voluntary - more controllable than passive behaviors such as coughing or sweating - but not as fully controllable as intentional actions and, second, that there would be a step-wise increase in perceived control between emotions, desires, beliefs, and thoughts.

## Methods

Participants. A total of 442 participants were recruited from Amazon's Mechanical Turk to participate in Study 2. 146 individuals ( 65 female, 81 male; mean age $=35$ ) participated in Study 2a ("Choice"), 149 (66 Female, 78 Male, 5 unreported; mean age $=34$ ) participated in 2 b ("Control"), and the remaining 147 ( 60 Female, 85 Male, 2 unreported; mean age $=35$ ) participated in 2c ("Choose to stop"). No participants were excluded.

Stimuli. To generate contexts, we constructed 30 scenarios describing someone in an ordinary or believable situation (such as repairing a bike, photographing a wedding, walking down the street, and so on). Unlike Study 1, each scenario provided a great deal of context about the person and situation leading up to the mental state or behavior. Our primary manipulation was the last element of each scenario, which was either (1) an observable uncontrollable reaction, (2) an emotion, (3) a desire, (4) a belief, (5) thinking or ruminating on some idea, or (6) an observable intentional action. Below is one of the 30 scenarios with each of the six conditions:

Katy is nearing the end of her third year in college. She's studying chemistry and biology in order to eventually apply to medical school. Any low grade will hurt her chances at getting into the top medical schools. Today, however, she struggled through the final exam in her chemistry class. She did not complete it in time and had to guess on the entire last page of questions.
Walking out of the exam, Katy...

1. begins shivering in the cold. (uncontrollable reaction)
2. feels angry at her professor. (emotion)
3. wants to leave her professor a poor course evaluation. (desire)
4. believes that her professor deserves a poor course evaluation. (belief)
5. thinks about leaving her professor a poor course evaluation. (thinking)
6. fills out a negative course evaluation on her phone. (intentional act)

As this example illustrates, the context prior to the manipulation was held constant, and the attitudinal content of each mental state (and the intentional behavior) was also held as constant as possible (e.g., in the item above, a negative and retaliatory attitude towards the professor is conveyed in each case). There was, of course, no such content for the uncontrollable foils (shivering, sneezing, coughing, etc.). We varied the kind of emotion experienced by the agent in the scenario: across the 30 sets, the emotion condition featured the agent feeling either angry, sad, afraid, excited, or pleased. Given 30 scenarios, each of which split into six behavior conditions, there were 180 items in the whole experiment.

Design. The items were distributed across six lists (of 30 items each) using a Latin-square design. Each list had one item category from each of the 30 scenarios, yielding a total of five trials within each list for each item category. We balanced the distribution of emotion trials so that each of the five different emotions appeared in each list.

Dependent measures. In Study 2a, the main dependent variable was whether the agent chose the particular mental state he/she had at the end of the story. Participants indicated their answer on a rating scale ranging from 1 (definitely did not choose) to 7 (definitely did choose). As in Study 1, the full content of each item was included in each question and was italicized (e.g. "Did Katy choose to feel angry at her professor?").

Study 2 b measured perceptions of how much control the agent had over whether he/she had the particular mental state (or over the behavior). For instance, in the Katy vignette above, participants were asked "How much control did Katy have over whether she felt angry at her professor?" on a scale from 1 (no control at all) to 7 (complete control).

Finally, 2c, measured perceptions of the degree to which the agent could stop the particular mental state (or behavior) once it had started. In the Katy vignette above, participants were asked "Can Katy choose to stop feeling angry at her professor?" on a scale from 1 (definitely can not choose) to 7 (definitely can choose).

Procedure. At the beginning of the experiment participants were randomly assigned to one of the six stimulus lists. Participants were provided brief instructions that they would read 30 stories about different characters and answer a question about a behavior that the character performed. Each trial was presented on a separate page in a new random order for each participant. At the end of the study, participants filled out a brief demographics questionnaire. No other data was collected.

## Results

All analyses were performed by running a linear mixed-effect model (LMEM) regressing ratings on the within-subject, within-scenario behavior manipulation. We included random intercepts for participant and scenario ratings, as well as


Figure 2: Mean (and SE) ratings for (A) Choice, (B) Control, and (C) Choose to stop measures of agency across behavior conditions in Study 2
random slopes for by-subject and by-scenario variation in the effect of condition. For each control measure, Intentionality, Control, and Stopping, we ran three sets of analyses. One analysis compared the means of the four mental state categories to the uncontrollable behavior foil (dummy coded as the reference level); one analysis compared the means of the four mental state categories to the intentional act behavior foil; and finally one analysis compared mental state categories to each other following our hypothesized stepwise increase in control through emotions, desires, beliefs, and thinking. See Figure 2 for condition means and standard error across control measures.

2a: Choice. Replicating results from Study 1, emotions ( $M=$ 3.95, $S D=2.08$ ), desires $(M=5.48, S D=2.08)$, beliefs $(M$ $=5.43, S D=1.73)$, and thinking $(M=5.29, S D=1.77)$ were all judged significantly more intentional than uncontrollable reactions $(M=1.67, S D=1.37 ; p \mathrm{~s}<0.001)$. These states were also all judged significantly less chosen than intentional acts ( $M=6.45, S D=1.13 ; p \mathrm{~s}<0.001$ ). Also replicating Study 1, emotions were rated as less chosen than desires $(b=1.534$, $S E=0.08, t=19.243, p<0.001$ ) but, contrary to expectations, there were no differences between desires and beliefs or beliefs and thinking ( $p \mathrm{~s}>0.07$ ).

2b: Control. Emotions ( $M=4.26, S D=1.44$ ), desires ( $M=$ 4.96, $S D=1.85$ ), beliefs ( $M=5.09, S D=1.72$ ), and thinking ( $M=5.00, S D=1.81$ ) were all judged significantly more controllable than uncontrollable reactions ( $M=2.09, S D=$ 1.44; $p \mathrm{~s}<0.001$ ) and significantly less controllable than intentional acts ( $M=6.32, S D=1.23 ; p \mathrm{~s}<0.001$ ). Similar to Study 2 a , we observed a significant difference between emotions and desires ( $b=0.692, S E=0.076, t=9.125, p<$ 0.001 ), but not between desires, beliefs, or thinking ( $p \mathrm{~s}>$ 0.07).

2c: Choosing to Stop. Again, Emotions $(M=4.44, S D=$ 1.99), desires $(M=4.74, S D=1.98)$, beliefs $(M=4.77, S D=$ 1.93), and thinking ( $M=5.08, S D=1.84$ ) were all judged significantly easier to stop than uncontrollable reactions ( $M$
$=2.03, S D=1.58 ; p \mathrm{~s}<0.001)$ and significantly harder to stop than intentional acts $(M=6.08, S D=1.54 ; p \mathrm{~s}<0.001)$. Participants judged emotions as more difficult to stop than desires $(b=0.306, S E=0.084, t=3.662, p<0.001)$, and beliefs more difficult to stop than thinking ( $b=0.305, S E=$ $0.084, t=3.638, p<0.001$ ) but did not distinguish between beliefs and desires ( $b=0.03, S E=0.084, t=0.359, p=0.72$ ).

## Discussion

Study 2 replicated the main findings from Study 1: people attribute moderate to high agency to others over their emotions, desires, beliefs, and deliberative thoughts, whether that agency is conceptualized as "choice", general "control", or an ability to "choose to stop". This finding replicated in spite of more explicit portrayals of relevant situational constraints. We also replicated the finding that this control is not perceived as complete: individuals were granted less agency over all mental states (even traditionally "active" processes such as thinking) compared to observable intentional acts.

We also found that emotions were perceived as less voluntary than desires. Unexpectedly, once holding mental state content constant, the other differences in perceived agency, namely, those between desires and beliefs, and between beliefs and thinking, did not replicate except in the "choose to stop" condition. This may reflect the improved design in this study (i.e., the fact that background context and focal content were held constant), but suggests that some variation in general control may come from the kinds of content different mental states are usually about.

## General Discussion

Our results pose a challenge to a common assumption in linguistics, anthropology, and social psychology, namely that people view others' mental states as largely uncontrollable. Contrary to this assumption, we report that people judge many mental states to be quite controllable: they clearly do not perceive mental states as just happening (cf. Gilovich \& Regan, 1986), completely outside voluntary control (cf. Katz
\& Postal, 1964), nor as uniformly unintentional (cf. Malle \& Knobe, 1997a).

In line with models like those proposed by D'Andrade (1987), our results also suggest that people attribute different degrees of voluntary control to different mental state categories (even holding context and content constant). However, as D'Andrade (1987) never empirically tested his model, his specific predictions were wrong: for instance, we found that people viewed desires as moderately controllable, more controllable, on average, than emotions, whereas D'Andrade (1987) posited that desires were uncontrollable (while emotions were partially controllable). Future work should investigate the sources of variation in control both between (e.g. why beliefs easier are to control than emotions), and within mental state categories (e.g., why particular beliefs differ in their perceived controllability).

Finally, a great deal of work has shown that people are held accountable for their moral wrongs (e.g. Alicke, 2000; Malle et al, 2014). To date, however, notwithstanding some related work inferring poor character from knowledge of noxious mental states (see, e.g. Ames \& Johar, 2009; Gromet et al., 2016), no one has investigated the possibility that people hold each other accountable (i.e., blameworthy) for their immoral beliefs, desires, or emotions. Given that people apparently do attribute agency to others over everyday mental states, and control predicted judgments of responsibility in Study 1, future work should investigate whether these results replicate for immoral mental states, and whether perceived agency predicts blame, anger, and punishment.

To conclude: assumptions about the perceived agency of mental states are common, yet direct empirical investigations are rare. Across several studies, we found that these assumptions fail to track ordinary judgments of the controllability of mental states. Accordingly, our results may have important implications for a range of debates in social and cognitive psychology, and open up new questions about the sources of variation in perceptions of mental control.

## Acknowledgments

The authors would like to thank Jon Baron, Bertram Malle, Paul Rozin, and anonymous reviewers for their valuable feedback on this work.

## References

Adams, R. M. (1985). Involuntary sins. The Philosophical Review, 94, 3-31.
Alicke, M. D. (2000). Culpable control and the psychology of blame. Psychological Bulletin, 126, 556-574.
Alston, W. P. (1988). The deontological conception of epistemic justification. Philosophical Perspectives, 2, 257-299.
Ames, D. R., \& Johar, G. V. (2009). I'll know what you're like when I see how you feel: How and when affective displays influence behavior-based impressions. Psychological Science, 20, 586-593.
Cohen, A. B., \& Rozin, P. (2001). Religion and the morality of mentality. Journal of Personality and Social.

Psychology, 81, 697-710.
Corrigan, R. (1988). Who dun it? The influence of actorpatient animacy and type of verb in the making of causal attributions. Journal of Memory \& Language, 27, 447-465.
D'Andrade, R. (1987). A folk model of the mind. In D. Holland \& N. Quinn (Eds.), Cultural models in language and thought (pp. 112-148). New York, NY, US: Cambridge University Press.
DeMarree, K. G., Wheeler, S. C., Briñol, P., \& Petty, R. E. (2014). Wanting other attitudes: Actual-desired attitude discrepancies predict feelings of ambivalence and ambivalence consequences. Journal of Experimental Social Psychology, 53, 5-18.
Frankfurt, H. G. (1988). The importance of what we care about: Philosophical essays. Cambridge, England: Cambridge University Press.
Gilovich, T., \& Regan, D. T. (1986). The actor and the experiencer: Divergent patterns of causal attribution. Social Cognition, 4, 342-352.
Gromet, D. M., Goodwin, G. P., \& Goodman, R. (2016). Pleasure from another's pain: The influence of a target's hedonic states on attributions of immorality and evil. Personality \& Social Psychology Bulletin, 42, 1077-1091.
Huddleston, R. (1970). Some remarks on case-grammar. Linguistic Inquiry, 1, 501-511.
Leary, M. R., Springer, C., Negel, L., Ansell, E., \& Evans, K. (1998). The causes, phenomenology, and consequences of hurt feelings. Journal of Personality and Social Psychology, 74, 1225-1237.
Malle, B. F. (2004). How the mind explains behavior: Folk explanations, meaning, and social interaction. Cambridge, MA: MIT Press.
Malle, B. F., Guglielmo, S., \& Monroe, A. E. (2014). A theory of blame. Psychological Inquiry, 25, 147-186.
Malle, B. F., \& Knobe, J. (1997a). Which behaviors do people explain? A basic actor-observer asymmetry. Journal of Personality and Social Psychology, 72, 288304.

Malle, B. F., \& Knobe, J. (1997b). The folk concept of intentionality. Journal of Experimental Social Psychology, 33, 101-121.
Miller, G. A., \& Johnson-Laird, P. N. (1976). Language and perception. Belknap Press.
Locke, D., \& Pennington, D. (1982). Reasons and other causes: Their role in attribution processes. Journal of Personality and Social Psychology, 42, 212-223.
Schlesinger, I. M. (1992). The experiencer as an agent. Journal of Memory and Language, 31, 315-332.
Smith, A. (2008). Control, responsibility, and moral assessment. Philosophical Studies, 138, 367-392.
Weiner, B. (1995). Judgments of responsibility: A foundation for a theory of social conduct. New York, NY: Guilford Press.

# Judgment Before Emotion: <br> People Access Moral Evaluations Faster than Affective States 

Corey Cusimano (cusimano@sas.upenn.edu)<br>Department of Psychology, University of Pennsylvania, 3720 Walnut St Philadelphia, PA 19104 USA

Stuti Thapa Magar (stuti_thapa_magar@brown.edu) and Bertram F. Malle (bfmalle@brown.edu)<br>Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, 190 Thayer Street, Providence, RI 02912 USA


#### Abstract

Theories about the role of emotions in moral cognition make different predictions about the relative speed of moral and affective judgments: those that argue that felt emotions are causal inputs to moral judgments predict that recognition of affective states should precede moral judgments; theories that posit emotional states as the output of moral judgment predict the opposite. Across four studies, using a speeded reaction time task, we found that self-reports of felt emotion were delayed relative to reports of event-directed moral judgments (e.g. badness) and were no faster than persondirected moral judgments (e.g. blame). These results pose a challenge to prominent theories arguing that moral judgments are made on the basis of reflecting on affective states.


Keywords: affect, emotion, moral judgment, reaction time

## Introduction

There is broad agreement that affective phenomena play an important role in moral cognition; there is widespread disagreement, however, over the particular role that affect plays. Many theories suggest that emotion acts as an input to moral judgment: that affective states help distinguish moral from non-moral events (Nichols, 2002), indicate the severity of the transgression (Haidt, 2001), or bias downstream cognitive processes (Alicke, 2000). In contrast to these "emotion-as-input" (em-in) models, "emotion-asoutput" (em-out) models argue that, very often, considerations of rules, norms, risk or caused harm, and causal and mental information, guide moral judgments without any necessary causal precedence of affect (Huebner, Dwyer, \& Hauser, 2009; Mikhail, 2011). According to these models, emotions are typically connected to moral judgments because as they motivate and scale our social responses.

The focus of our paper is on a prominent subset of the emin theories, which claim that emotions precede and influence moral judgment through felt affect. For example, Schnall Haidt, Clore, and Jordan (2008) claim, "When making evaluative judgments, people attend to their own feelings, as if asking themselves: How do I feel about it?" (p. 1097, emphasis added). Similarly, Miller et al (2012) argue that "the likelihood of judging an action wrong is determined... by how upsetting you consider the action itself to be" (p
574). That is, moral judgments of some event are formed by recognizing and reporting one's emotional response to that event. This is why, according to these theories, the experience of negative affect on its own (i.e. absent appraisals of harm or risk) can yield negative moral evaluations (e.g. Haidt, 2001). The primary source of evidence for this comes from affect misattribution experiments in which inducing feelings of disgust (unrelated to the stimulus) both amplified the perceived wrongness of target behaviors (Schnall, et al., 2008; Cheng, Ottati, \& Price, 2013) and appeared to cause ordinarily permissible behaviors to be judged as wrong (Wheatley \& Haidt, 2005).

Though initially promising, many of the findings in favor of the em-in model of moral judgment have been called into question. First, it appears as though many moral judgments can be made absent any affective experience (Niedenthal, Rohmann, \& Dalle, 2003) and, conversely, many strong emotional reactions occur without any corresponding moral judgment (Royzman, Goodwin, \& Leeman, 2011). Additionally, the primary source of evidence for the causal role of felt affect has been called into question: a recent meta-analysis reports that, across dozens of experiments, there is not reliable effect of incidental disgust on moral judgment (Landy \& Goodwin, 2015).

However, even though these findings are consistent with em-out models, a major challenge in assessing any of the theories regarding the role of emotion in moral judgment is the dearth of experimental paradigms that get at the heart of the causal primacy question-whether the routine causal sequence is, according to one set of theories, event $\rightarrow$ emotion $\rightarrow$ moral judgment or, according to the other set of theories, event $\rightarrow$ moral judgment $\rightarrow$ emotion. What is needed are independent and time-locked measurements of the relevant moral and affective processes as they emerge in response to a range of different moral violations.

Because causality implies temporal precedence, we reasoned that if attended emotions cause moral evaluations, then participants ought to experience (and be able to report) certain emotions (such as feeling angry or upset) before being able to judge the moral status of that behavior. In contrast, if moral judgments guide affect or emotions based on perceived norm violations, causal and mental information, and so on, then felt emotions should follow moral judgments.

## Experimental Paradigm

To examine questions of causal primary, we conducted a series of reaction time experiments to test the relative speed of moral, non-moral, and affective reactions to value-laden events. We relied on a variant of the simultaneous inference paradigm (SIP, Smith \& Miller, 1983; Malle \& Holbrook, 2012) to measure the speed at which people make different judgments in response to short descriptions of moral transgressions. In the SIP, participants learn to associate a question with a short cue (or, hereafter, "probe") which is then used to elicit responses in a speeded-judgment task. These probes minimize the latency between the presented question and the participant's comprehension, as well as differences in the length and complexity of full questions.

Prior research using the SIP trained participants on dichotomous Yes-No judgments (e.g. "Did the behavior reveal a certain goal the actor has?"), which required modification for two reasons: First, many moral and affective reactions are graded: stabbing someone is worse than keying their car, which is worse than stealing their pencil. A simple Yes-No judgment does not indicate that someone is sensitive to these differences. Second, and relatedly, a prediction of em-in models is that the extremity of the affective reaction predicts the perceived severity of the transgression (e.g. Miller et al, 2012), which makes the best test of these models one in which both moral and affective judgments require reporting this more specific, nuanced information. To do this, we presented each probe along with a 7 -point rating scale from which participants selected their response as quickly as they could.

We also varied the type of moral scenario participants would react to. Different kinds of events reliably lead to different moral and affective reactions (e.g. people blame transgressors more for intentional harms relative to unintentional ones), and these different outputs are thought to reflect different underlying cognitive processes (Cushman, 2013; Malle et al, 2014). Furthermore, variation in encountered behavior better reflects the experience of encountering random morally relevant behaviors in the real world and generates variation that requires participants' attention. To this end, studies 1-2 mixed intentional and unintentional violations, while studies 3 and 4 mixed intentional, unintentional, and non-agent caused events.

Across four experiments, we measured reaction times for four response types: (1) non-moral judgments (e.g "Intentional?"), (2) moral evaluations of the event (e.g., "Bad?" or "Good?"), (3) moral judgments of the person (e.g., "Blame?"), and (4) reports of one's own affective state (e.g., "Angry?"). As argued above, the em-in models predict that people's responses to the affective probes should be faster than the responses to the moral probes, whereas the em-out models predict the opposite. For the purpose of the current report, we will focus on these a priori contrasts of response times, setting aside the speed of other probes and the specific ratings people provided.

## Study 1

## Methods

Participants. 241 people ( 130 self-reported as female, mean age $=35$ ) recruited from Amazon's Mechanical Turk (AMT) participated in this experiment.

Stimuli. We constructed 24 short descriptions of an agent causing harm either intentionally or unintentionally (e.g. "When she walked by a homeless man asking for money, Lisa spit on the ground in front of him"). Intentionality was verified through pretesting (mean intentionality ratings for intentional and unintentional descriptions were 8.17 and 2.35 respectively on a $1-9$ scale). The 12 intentional and unintentional sentences were matched on length (15.3 and 15.8 words for intentional and unintentional conditions, respectively) and varied in moral severity (valence ratings 1.1 to $-4.0, M=-2.45$, for intentional transgressions, and 0.65 to $-3.88, M=-2.00$ for unintentional transgressions on a -4 to +4 scale).
Judgments. Our non-moral, social judgment probed intentionality (Cue: INTENTIONAL? Full: Was the main character's behavior INTENTIONAL?) on a [1] definitely not intentional to [7] definitely intentional scale. Our eventdirected moral evaluation probed "badness" (Cue: BAD? Full: How BAD was the thing that happened? Scale: [1] not at all bad - [7] the most bad possible), while our persondirected moral judgment probed judgments of blameworthiness (How much BLAME does the main character deserve? Scale: [1] no blame at all - [7] the most blame possible). Finally, to assess participants' affective states, we used a general feeling probe (Cue: FEEL? Full: How much did the story make you FEEL something? Scale: [1] no feeling at all - [7] the most feeling possible).
Design. The study crossed two within-subject factors: behavior type (intentional vs unintentional) and judgment type (INTENTIONAL, BAD, BLAME, FEEL). The 24 experimental stimuli and four judgments types were distributed over participants such that they were probed for each judgment type 6 times, (three for intentional behaviors, three for unintentional behaviors). We used a Latin-square design to pair each of the four judgments with each of the 24 stories across four lists. The order of stimuli and probes was randomized for each participant within each list.

Procedure. The entire experiment was conducted through the participant's web browser. At the beginning of the experiment, participants received instructions, including a description of the cues and their associated meanings, as well as the fact that they would be doing a speededjudgment task and so would have limited time to read and respond to the vignettes. They then completed a training session in which they were taught the single-word cues for the associated judgments (e.g. "BAD?" for "How bad was the thing that happened?").

During the experiment, for each trial, a one-sentence description of a transgression was displayed in the center of
the screen. It remained for 4.5 s and was replaced by the probe and a seven-point rating scale (participants did not know which probe would be displayed for any trial). Participants were instructed to place their fingers on the number row of the keyboard and press the corresponding number to respond. Once they indicated their response, the cue and scale disappeared and the next trial automatically started. At the end of the experiment, participants filled out a brief demographics questionnaire indicating their gender, age, and language background.

## Results

We removed all trials with reaction times greater than 10 seconds ( $0.7 \%$ data loss). Otherwise, no other trials or participants were removed from data analysis.

Following previous studies (Malle \& Holbrook, 2012), we conducted simple effects tests comparing RTs for the affect probe with RTs for other judgment types separately within intentional and within unintentional behaviors. Our primary question was whether reaction times to the affective judgment probe were slower than to other judgment probes. We used linear mixed-effect models (LMEM) to regress RTs on judgment type, which was dummy coded with affective judgment (here, FEEL) as the baseline. Finally, due to the within-subject design, each model included random intercepts and slopes for each participant as well as a random intercept for each scenario.

When judging intentional behaviors, reaction times for FEEL judgments ( $M=2563, S D=1246$ ) were significantly slower compared to INTENTIONAL $(M=2113, S D=985, \mathrm{~b}=$ $-452.52, \mathrm{SE}=62.9, t=-7.19, p<0.001)$, BAD $(M=2332$, $S D=1204, \mathrm{~b}=-233.04, \mathrm{SE}=57.49, t=-4.054, p<0.001)$, and BLAME $(M=2355, S D=1211, \mathrm{~b}=-214.44$, $\mathrm{SE}=$ 55.964, $t=-3.832, p<0.001$ ). However, when judging unintentional transgressions, we observed no significant difference between FEEL ( $M=2640, S D=1267$ ) and INTENTIONAL $(M=2540, S D=1240, \mathrm{~b}=-98.71$, $\mathrm{SE}=$ 59.936, $t=-1.647, p=0.1)$, BAD $(M=2545, S D=1244, \mathrm{~b}=$ -96.17, $\mathrm{SE}=59.09, t=-1.63, p=0.104$ ), or BLAME ( $M=$ 2647, $S D=1339, \mathrm{~b}=13.88, \mathrm{SE}=60.93, t=0.23, p=0.82$ ).

## Discussion

Study 1 provides preliminary support against em-in models in favor of em-out models. Reaction times for the emotion probe FEEL were slower than those for judgments of INTENTIONALITY, BAD, and BLAME, at least when considering intentional norm violations. There were no comparable RT differences between the affect probe and the remaining probes in response to unintentional violations, perhaps because relevant moral rules are more difficult to access, harm more difficult to calculate, or responsibility more complicated to assess (e.g. Malle, et al., 2014).

One possible reason for the slow unfolding of affective reactions is that participants found it difficult to respond to a vague probe such as "feel". To address this possibility, we conducted another experiment using a more concrete easily identifiable and morally relevant emotion probe: angry.


Figure 1: Reaction time (and standard error) for participants' responses to probes in study 1 (A) and 2 (B)

## Study 2

## Methods

Participants, Materials, \& Procedure. 237 people (134 self-reported as female, mean age $=35$ ) recruited from Amazon's Mechanical Turk (AMT) participated in this experiment. Study 2 was identical to Study 1 except that the affective judgment probe assessed anger (Cue: ANGRY? Full: How ANGRY are you at the main character? Scale: [1] not at all angry to [7] the most angry possible).

## Results

Reaction times from Study 2 were analyzed identically to Study 1. Before conducting analyses, we removed all trials with RTs greater than 10s ( $2.8 \%$ data loss). No other data were removed.

Replicating Study 1, we found that, in the intentional condition, INTENTIONAL $(M=2111, S D=1119)$ and BAD ( $M$ $=2133, S D=1091$ ) judgments were both significantly faster than ANGER judgments $(M=2313, S D=1150$; INTENTIONAL: $\mathrm{b}=-203.90, \mathrm{SE}=52.32, t=-3.90, p<0.001$; BAD: $-183.18, \mathrm{SE}=52.90, t=-3.46, p=0.001$ ). We did not observe significant differences between ANGER ( $M=2448$, $S D=1300$ ) and other judgments for unintentional violations
(INTENTIONAL: $M=2358, S D=1189,-88.43, \mathrm{SE}=56.65, t$ $=-1.561, p=0.119$; BAD: $M=2382, S D=1291, \mathrm{~b}=-61.22$, $\mathrm{SE}=59.91, t=-1.02, p=0.307)$. Finally, we detected no significant difference between ANGER and BLAME for either intentional $(M=2231, S D=1214, \mathrm{~b}=-80.41, \mathrm{SE}=52.94, t$ $=-1.52, p=0.129)$ or unintentional $(M=2426, S D=1229$, $\mathrm{b}=-18.10, \mathrm{SE}=55.28, t=-0.33, p=0.743$ ) behaviors (see Figure 1).

## Discussion

Study 2 largely replicated the findings from Study 1. However, both studies are limited in several respects. First, the behaviors were always negative, and the agent in every description always a causer of harm. It is possible under these conditions that moral norms become more salient and accessible or that expectations of high causal agency sped up moral evaluations. Additionally, anger is typically directed at persons, and a different emotion term may be more appropriate for affective reactions to events. Finally, the studies did not limit participants' response time. While most responses occurred within several seconds, it is nevertheless possible that judgments are consciously accessible before then without an incentive to reveal these judgments as soon as they are accessible. The next two studies were designed to address these shortcomings.

## Studies 3 and 4

## Methods

Participants. 111 people ( 58 self-reported as female, mean age $=37.4$ ) in Study 3 and 193 people ( 90 self-reported as female, mean age $=33.4$ ) in Study 4, recruited from AMT participated in this experiment.

Stimuli. We constructed 28 single-sentence descriptions, 14 featuring a good event and 14 featuring a bad event. For each valence condition, we constructed four stimuli that had an agent with no causal role (Non-causal) in the good or bad event, four in which an agent unintentionally did a good or bad thing (Unintentional), and a final six in which an agent intentionally did something good or bad (Intentional). Pretesting ensured that these items were matched on intentionality and length across valence, and that all agency conditions were comparable (see Table 1).

Judgments. We modified the full questions associated with
Table 1: Pretest ratings for stimuli in Studies $3 \& 4$

| Behavior Types |  | Pretest Values |  |  |
| :--- | :--- | :---: | :---: | :---: |
| Valence | Agency | Intentionality | Valence | Words |
| Negative | Intentional | 8.41 | -2.58 | 16.83 |
|  | Unintentional | 2.40 | -2.57 | 16.00 |
|  | Non-causal |  | -2.70 | 13.75 |
| Positive | Intentional | 8.22 | 2.57 | 15.17 |
|  | Unintentional | 2.03 | 2.64 | 18.75 |
|  | Non-causal |  | 2.79 | 12.50 |

each cue to accommodate the greater variety of stimulus events. Similar to Studies 1-2, we included measures of (1) intentionality (Cue: INTENTIONAL? Full: "Was it intentional (what the character did)?" Scale: [1] definitely not intentional - [7] definitely intentional), and (2) the badness of the event (Cue: BAD? Full: "How bad was it (what happened)?" Scale: [1] not at all bad - [7] extremely bad). We also included a measure of (3) the goodness of the event in order to accommodate the positive valence items (Cue: GOOD? Full: "How good was it (what happened)?" Scale: [1] not good at all - [7] extremely good).

Studies 3 and 4 were identical except for the affective judgment probe. Study 3 measured anger (Cue: ANGRY? Full: "How angry were you (about what happened)?" Scale: [1] not at all angry - [7] extremely angry), whereas Study 4 measured "upset" (Cue: UPSET? Full: "How upset were you (about what happened)?" Scale: [1] not upset at all - [7] extremely angry).

Design. Studies 3 and 4 crossed two within-subject factors: event type (non-causal, intentional, and unintentional) and judgment type (INTENTIONAL, BAD, GOOD, and UPSET or ANGRY). The 28 experimental stimuli and four judgment types were distributed over participants in the following pattern: across the 28 items, participants responded to eight INTENTIONAL probes, four for intentional behavior conditions, two for unintentional behavior conditions, and two in the non-caused behavior conditions. This distribution meant that roughly half the probes would results in low intentionality ratings and the other half would result in high intentionality ratings. Affect probes were also distributed this way: four for intentional behavior stimuli, two for unintentional behavior stimuli, and two for non-caused. These probes were evenly divided between valence conditions. Finally, participants saw six BAD probes and six GOOD probes, which were matched to valence. That is, participants made BAD judgments only following negatively-valenced stimuli and GOOD judgments only following positively-valenced stimuli. The twelve moral evaluation probes ( 6 BAD and 6 GOOD) were evenly divided between event type conditions.

Probes were distributed across four stimulus lists according to a Latin-square design. At the beginning of the experiment, participants were randomly assigned to one of the four lists and, during the experiment, the order of the stimulus sentences and probes was randomized.

Procedure. The training and overall experiment procedures were the same as in Studies 1 and 2, with one exceptions: For each trial, the judgment screen containing the cue (e.g. "BAD?") and the rating scale disappeared after five seconds after being displayed. If no response was offered before then, no response was recorded for that trial. Participants were informed of the time restriction in the instructions. Prior to the experiment, participants conducted five practice trials to get accustomed to the procedure.

## Results

We removed all trials in which the participant did not provide an answer within the time constraint ( $1.6 \%$ data loss in Study 3, 3\% data loss in Study 4). No other data were removed. Similar to Studies 1 and 2, we conducted separate analyses on the Intentional, Unintentional, and Non-Causal behaviors, using the same mixed-effect regression models and adding the Valence (positive vs. negative) term, which predicted changes in the dummy variable (affect) as a function of positive or negative behaviors. Across all models in both studies, valence was not significant and did not improve model fit, and so was removed as a predictor.


Figure 2: Reaction time (and standard error) for participants' responses to probes in study 3 (A) and 4 (B)

Intentional Behaviors. In Study 3, moral evaluations were faster than ANGER ( $M=1746, S D=739$; BAD: $M=1601$, $S D=647, \mathrm{~b}=-178.46, \mathrm{SE}=37.35, t=-4.78, p<0.001$; GOOD: $M=1438, S D=645, \mathrm{~b}=-294.32, \mathrm{SE}=38.32, t=-$ 7.68, $p<0.001$ ), while INTENTIONAL ratings were not ( $M=$ 1687, $S D=739, \mathrm{~b}=-57.65, \mathrm{SE}=36.01, t=-1.60, p=$ 0.109). Similarly, in Study 4, intentional ( $M=1656, S D=$ $675)$, BAD $(M=1560, S D=607)$, and GOOD $(M=1456, S D$ $=642$ ) judgments were significantly faster than UPSET ( $M=$ 1822, $S D=682$; INTENTIONAL: $\mathrm{b}=-163.51, \mathrm{SE}=42.7, t=-$ 3.83, $p<0.001$; BAD: $\mathrm{b}=-273.68, \mathrm{SE}=48.50, t=-5.64, p<$ 0.001 ; GOOD: $\mathrm{b}=-366.28, \mathrm{SE}=48.19, t=-7.60, p<0.001$ ).

Unintentional Behaviors. In Study 3, moral evaluations were significantly faster than the ANGER ratings ( $M=1837$,
$S D=714 ;$ BAD: $M=1682, S D=694, \mathrm{~b}=-174.2, \mathrm{SE}=$ 43.07, $t=-4.05, p<0.001$; GOOD: $M=1601, S D=647, \mathrm{~b}=$ -193.3, $\mathrm{SE}=43.86, t=-4.41, p<0.001$ ), while INTENTIONAL was slower ( $M=1932, S D=775, \mathrm{~b}=115.11$, $\mathrm{SE}=41.88, t=2.75, p=0.006$ ). In Study 4, UPSET ( $M=$ 1760, $S D=617$ ) was significantly slower than GOOD ( $M=$ 1603, $S D=624, \mathrm{~b}=-157.62, \mathrm{SE}=53.89, t=-2.93, p=$ 0.003 ), and significantly faster than INTENTIONAL ( $M=$ 1972, $S D=769, \mathrm{~b}=212.92, \mathrm{SE}=53.78, t=3.96, p<$ $0.001)$, but not reliably different from $\operatorname{BAD}(M=1673, S D=$ $686, \mathrm{~b}=-82.75, \mathrm{SE}=53.86, t=-1.54, p=0.124$ ).

Non-Caused Behaviors. Moral evaluations of uncaused good and bad events were significantly faster than the ANGER judgments $(M=1728, S D=653$; BAD: $M=1561$, $S D=611, \mathrm{~b}=-163.31, \mathrm{SE}=38.24, t=-4.27, p<0.001$; GOOD: $M=1335, S D=541, \mathrm{~b}=-390.61, \mathrm{SE}=37.89, t=-$ 10.31, $p<0.001$ ), while INTENTIONAL ratings were slower ( $M=1969, S D=749, \mathrm{~b}=241.79, \mathrm{SE}=39.55, t=6.11, p<$ 0.001). In Study 4, UPSET ( $M=1659, S D=616$ ) was significantly slower than GOOD $(M=1340, S D=505, \mathrm{~b}=-$ 309.67, $\mathrm{SE}=50.41, t=-6.14, p<0.001)$ and $\operatorname{BAD}(M=$ 1556, $S D=639, \mathrm{~b}=-103.24, \mathrm{SE}=50.39, t=-2.05, p=$ $0.04)$, but significantly faster than INTENTIONAL ( $M=1947$, $S D=723, \mathrm{~b}=292.20, \mathrm{SE}=50.56, t=5.78, p<0.001)$.

## Discussion

Results from Studies 3 and 4 replicated our previous findings (see Figure 2): Moral evaluations of intentional violations were reliably faster than reports of felt anger and upsetness. For unintentional violations, reporting feeling upset was not significantly slower than negative moral evaluations, but anger was. Perhaps upset feelings are more globally sensitive to any unfortunate outcome and therefore converge with (but on average do not precede) badness judgments. Lastly, intentionality judgments were slowed in response to unintentional and uncaused events-which is not entirely surprising given that those events are clearly not intentional; the detection of negation may take time.

## General Discussion

Across four experiments, participants were reliably slower at reporting their emotional states in response to norm violations compared to reporting their moral judgments. More specifically, the results from these studies showed a clear speed advantage for event-directed judgments of badness and, often, intentionality judgments. These results fit both with theoretical models of moral judgment arguing that moral appraisals precede emotion, as well as prior work showing that intentionality and norm violation detection can occur extremely quickly (Malle \& Holbrook, 2012; Van Berkum et al, 2009). These findings did not extend to person-directed moral judgments (blame), consistent with theories that blame is more complex than event-directed evaluations (e.g. Cushman, 2013; Malle et al, 2014).

One important limitation of these studies comes from the observation that, while we are interested in characterizing
the cognitive processes underlying moral judgment when people are exposed to a morally relevant stimulus, we measured people's reaction times to probes that were displayed after the stimulus had been shown. It is possible that participants attended to their affective reactions when they were first exposed to the stimulus, which resulted in a moral judgment, which they later more quickly retrieved during the post-stimulus probe. However, while we cannot rule this out, it is not clear why affect would be more difficult to retrieve post-stimulus as opposed to duringstimulus (when one's attention is presumably directed outward toward reading the stimulus). Additionally, this account does not explain why the speed of retrieving affective information would change as a function of the behavior (Study 4). Finally, prior work using a simultaneous inference paradigm found that post-stimulus reaction times directly recapitulated online measures (Malle \& Holbrook, 2012).

Second, affect may have been slower relative to moral and social judgments because of an attention switching cost: the non-affect judgments targeted the stimulus while the affect judgment targeted oneself. Because em-in models explicitly predict a shift in attention from the behavior to one's affective state, a delay in reporting due to switching attention is not, in principle, a confound for our test. That said, this cost may have been exacerbated by the relative balance of event-directed (75\%) versus self-directed (25\%) probes, and future studies should use an even balance of affect and moral judgments.

Finally, even if we accept that felt emotions do occur after explicit moral judgment, our data do not rule out the possibility that pre-conscious affective processes play a role in moral judgment formation (say, by interfering with cognitive processes, Alicke, 2000). It is also possible that conscious affect may play a causal role when judging more ambiguous situations, in which relevant harm or rule information is difficult to access. Consistent with this, badness and blame judgments were not reliably faster than emotion reports when judging accidental bad behavior. Thus, our results may only hold for relatively common or extreme, but not unusual or novel situations.

In summary, we found that people could report moral evaluations of norm-violating events more quickly than their emotional reactions to these events. These results pose a challenge to models claiming that felt affect plays a necessary role in forming moral judgment.

## Acknowledgments

The authors would like to thank Fiery Cushman, John Voiklis, members of the SCSRL and anonymous reviewers for valuable feedback on this project. This research was funded in part by a grant from the Office of Naval Research (ONR), N00014-13-1-0269.

## References

Alicke, M. D. (2000). Culpable control and the psychology of blame. Psychological Bulletin, 126, 556-574.

Cheng, J. S., Ottati, V. C., \& Price, E. D. (2013). The arousal model of moral condemnation. Journal of Experimental Social Psychology, 49, 1012-1018.
Cushman, F. (2013). Action, outcome, and value a dualsystem framework for morality. Personality and Social Psychology Review, 17, 273-292.
Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review, 108, 814.
Huebner, B., Dwyer, S., \& Hauser, M. (2009). The role of emotion in moral psychology. Trends in Cognitive Sciences, 13, 1-6.
Landy, J. F., \& Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. Perspectives on Psychological Science, 10, 518-536.
Malle, B. F., Guglielmo, S., \& Monroe, A. E. (2014). A theory of blame. Psychological Inquiry, 25, 147-186.
Malle, B. F., \& Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. Journal of Personality and Social Psychology, 102, 661-684.
Mikhail, J. M. (2011). Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment. New York, NY: Cambridge Uni. Press.
Miller, R. M., Hannikainen, I. A., \& Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. Emotion, 14, 573-587.
Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. Cognition, 84, 221-236.
Niedenthal, P. M., Rohmann, A., \& Dalle, N. (2003). What is primed by emotion concepts and emotion words. The psychology of evaluation: Affective processes in cognition and emotion, 307-333.
Paulhus, D. L., \& Lim, D. T. (1994). Arousal and evaluative extremity in social judgments: A dynamic complexity model. European J. of Social Psychology, 24, 89-99.
Royzman, E. B., Goodwin, G. P., \& Leeman, R. F. (2011). When sentimental rules collide: "Norms with feelings" in the dilemmatic context. Cognition, 121, 101-114.
Schnall, S., Haidt, J., Clore, G. L., \& Jordan, A. H. (2008). Disgust as embodied moral judgment. Personality and Social Psychology Bulletin, 34, 1096-1109.
Smith, E. R., \& Miller, F. D. (1983). Mediation among attributional inferences and comprehension processes: Initial findings and a general method. Journal of Personality and Social Psychology, 44, 492.
Van Berkum, J. J., Holleman, B., Nieuwland, M., Otten, M., \& Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. Psychological Science, 20, 1092-1099.
Wheatley, T., \& Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. Psychological Science, 16(, 780-784.

# The Lego hands: changing the affording location of graspable objects 

Da Silva Fabrice (fabrice.da-silva@univ-montp3.fr) \& Brunel Lionel (lionel.brunel@univ-montp3.fr)<br>Epsylon Laboratory, Paul Valéry University, Montpellier, 34190, France


#### Abstract

The present study examined throughout three experiments the nature of stimulus-response compatibility (SRC) effects related to affordance perception in situations wherein object affordances and response effectors are irrelevant to each other. In the first experiment, using a foot-press response dispositive, we found a SRC effect between the orientation of the graspable part of the presented object and the laterality of the response. In Experiment 2a, we showed that constraining the subject hands in a given position (i.e., a Lego hand shape) during the same task interfered with the SRC effect. In Experiment 2b, participants performed a short training phase with their hands constrained before performing the experiment. This resulted in an inversion of the direction of the SRC effect previously observed. We discuss these results and provide arguments in favor of a specific motor activation account.


Keywords: Visual Perception; affordances; categorization; motor constraints

## Introduction

In stimulus-response compatibility (SRC) paradigms related to affordance perception (see Michaels, 1988), participants are usually faster and more accurate to categorize stimuli when the response hand and the presented objects are located on the same side (i.e., compatible) rather than on the lateral opposite side (i.e., incompatible). The present study aimed at disentangling between two alternatives explanations of this specific SRC effect.

## The affordances

Stimulus-response compatibility paradigms were first designed to highlight affordance effects. These affordances were defined as what a given environment offers the animal, what it provides or furnishes in terms of action possibilities (Gibson, 1979). These action possibilities are properties of the subject-environment system and emerge from the relation between an object and a subject (Stoffregen, 2003). For instance, stairs can afford an action of climbing only when their size does not exceed a certain proportion of the riser leg height (Warren, 1984). In SRC paradigms, participants generally performe a perceptual categorization task using a specific motor response. The critical manipulation is the compatibility (or congruence) between the motor response setting and the perceptual configuration of this object. For instance, Tucker and Ellis (1998) showed that participants were faster and more accurate to categorize the orientation (i.e., upright or downright) of daily life graspable objects when object handles and motor responses referred to the same side (i.e., compatible) than when they referred to the opposite side (i.e., incompatible). These results were interpreted as evidences of the affordance
effects. This interpretation was further supported by electrophysiological recording such as analyses of lateralized readiness potentials (LRP) during categorical judgments (Goslin, Dixon, Fischer, Cangelosi, \& Ellis, 2012), or studies about the link between the affordance perception and the dorsal stream activation through transcranial mental stimulations (Buccino, Sato, Cattaneo, Rodà \& Riggio, 2009).

## Specific motor activation versus abstract space coding

Despite the multiplication of experimental works concerning affordance perception for about twenty years, the nature of affordance-related SRC effects is still debated. For Tucker and Ellis (1998), SRC effects were observed because interactions with an object involve a representation about the range of actions that we can perform with and thereby, potentiate them. The nature of these representations was discussed in later works by the same authors including through micro-affordances (e.g. Tucker \& Ellis, 2001). However, other authors provided evidences that an alternative explanation of SRC effects might be considered. (Anderson, Yamagishi, \& Karavia, 2002). These authors interpreted such effects as a consequence of an attentional bias induced by a stimuli perceptual asymmetry (i.e., attention might be oriented to the left or the right depending on the perceptual configuration of the stimulus). This attentional orientation could be responsible for spatialrelated motor activations without requiring the potentiation of action-related properties of an object. This hypothesis is also in line with location-coding theories (e.g., Cho \& Proctor, 2011). In order to disentangle between specific motor activations and abstract location coding as mechanisms responsible for affordance related SRC effects, Phillips and Ward (2002) developed a method wherein a prime graspable object and the orientation of its handle were irrelevant to the participants' task. In this study, the handle of the priming object was oriented to the left or to the right side. Furthermore, the object handle was presented with an apparent depth towards or away from the participant. Participants had to respond to a target that appeared in the center of a computer screen. Authors found a main effect of the handle orientation congruent with common SRC studies but no significant effects involving its apparent depth. Nevertheless, this result could be attributed to methodological issues. Indeed, such proximity-related effects have been reported since (see Fischer \& Dahl, 2007). A more interesting point is that Phillips and Ward (2002) proposed another experiment in which participants had to respond to the same task pressing using foot switches with their left or their right foot. Like in the first experiment,
authors found a significant main effect of correspondence. Therefore, participants were faster to categorize stimuli when both the foot and the handle of the object were localized on the same side. The authors concluded that response facilitation effects arise from an abstract location coding.

However, if observing SRC effects with feet during the presentation of graspable objects seem to be inconsistent with the motor specific activations account, this is not sufficient to reject it. Indeed, grasping an object does not imply a single hand gesture but a more global engagement of the body. It is unclear in what extend the body is engaged during the perception of graspable objects and further investigations are necessary to question the implications of such generalized activations. Therefore, finding SRC effects while the response effector (i.e., the foot) and the action associated with the presented object (i.e., hand use) are seemly not directly related is not sufficient to conclude that no specific motor activations occurred.

## The current study

In the present study, we aimed at showing that motor activations could constitute the core mechanism of SRC effects. For this purpose, we conducted three experiments in which subjects had to categorize a common graspable object with foot-press responses. The experimental design was similar to the one used by Tucker and Ellis (1998) and consisted in a classical SRC paradigm. In the three experiments, participants had to categorize with their feet the orientation (i.e., upright or inverted) of a common mug displayed on a computer screen. In the first experiment, they responded while keeping their hands placed on the table in front of them. In the second experiment, participants were wearing gloves during the task, constraining their hands in an opened position (i.e., such as the Lego hand shape). Finally, in a third experiment, a last group responded while wearing the same gloves but after performing a short training phase.

## Experiment 1

The aim of this experiment was to replicate the SRC effects observed in Tucker and Ellis (1998) tasks using Phillips and Ward (2002) response setting.

## Method

Participants Twenty undergraduates students (17 women) from Paul Valéry Montpellier University aged from 18 to 36 years old ( $M=23.1, S D=4.72$ ) took part to this experiment and received course credits. All had normal or corrected to normal vision and were naïve to the purpose of the study. The experiment was realized in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

Apparatus and materials The experiment was performed using E-Prime 2 software (Schneider, Eschman, \&

Zuccolotto, 2002). The visual material consisted of pictures of a common mug disposed sideway on a white background and centered on the display. The mug dimensions were 340 x 320 pixels. All pictures were realized using the software Maya 16.0 (Palamar, 2014). The orientation of the initial picture was manipulated to produce two horizontal and two vertical orientations of the mug. Additionally, a second filler picture was presented in order to increase the difficulty of the task. The two stimuli are depicted in Figure 1.


Figure 1: The two visual stimuli used in all our experiments (in one orientation condition). The right panel object acted as a filler object.

As with the realistic mug, we reoriented the original picture of the filler object to produce four final pictures. The depicted objects were presented with an apparent size similar to their real size which is about 11 cm high and 7.5 cm wide and the participants performed the task at a distance of 45 cm from the screen and a visual angle of about $15^{\circ}$. All results associated with the filler object were not included into the analyses. Each of the eight pictures was presented ten times in the experiment so that a participant responded to an overall of 80 trials presented randomly. Trials wherein the object was presented with its handle oriented on the same side that the response effector were considered as compatible and incompatible when it was the reverse situation.

Procedure After filling out a consent form, participants sat in front of a computer and were asked to rest their hands on the table in front of them and their foots above the pedals. Then, they were asked to perform a forced choice categorization task on a computer by pressing a left or a right switch of a pedalboard. They had to determine as fast as possible if the displayed objects were disposed upright or inverted. Each pedal was attributed to a response category. This attribution was counterbalanced for the half of the sample. Each picture was displayed until the response and preceded by a fixation point which remained on the screen during 200 milliseconds.

## Results

First, we observed that participants have accurately performed the categorization task (less than $5 \%$ of error rate). Thus, we only considered latencies for analysis. We
excluded RTs above 1250 milliseconds (This cut-off led to the exclusion of $11.37 \%$ of trials). Thus, We performed a RT analysis between Compatible and Incompatible situations for responses exclusively related to the realistic mug pictures. Mean response times for the Compatible situations ( $M=713 \mathrm{~ms}, S D=76.77 \mathrm{~ms}$ ) were faster than mean response times for the Incompatible situations ( $M=$ $750 \mathrm{~ms}, S D=93 \mathrm{~ms}$ ). Using a bilateral paired t-test, we found that this difference was significant, $t(19)=2.55, p<$ $.05, d=.43$.

## Discussion

The results clearly demonstrated a SRC effect. Indeed the participants were faster to categorize objects when their graspable part was oriented on the same side as the response effector rather than the opposite side. With the same pattern of results, Phillips and Ward (2002) concluded that such facilitation effects had to be consecutive to an attentional shift and could not be associated with premotor activations. This conclusion is congruent with the fact that the graspable part of objects like the mug we used constitutes a visual protrusion that could capture the subject attention. Furthermore, a foot-press response seems to be not related to the perception of the affordance of the mug. However, to conclude that facilitation effects emerging from unsuited limb responses are not related to specific motor activations, this experimental design is insufficient. Indeed, contrary to the experimental context, the tool use in daily life implies a more global generation of movements and might involve that perception of graspable objects potentiate a higher range of muscles that the hand or even the arm. If the affordance theory is unclear to specify such implications, it is necessary to control the hand disposition during this kind of tasks for rejecting this motor hypothesis. If such of a hand constraint results in an alteration of facilitation effects as those observed in the experiment 1 , this would constitute evidence that such response facilitations are dependent on the subject action possibilities.

## Experiment 2a

In this experiment, we aimed at constraining subject hands to a certain position which is incompatible with the usual grasp of the object used in Experiment 1. The gloves used in this experiment were conceived to induce a large grasping position incompatible with the handle of the mug. If SRC effects such as the ones revealed by Tucker and Ellis (1998, 2001) or Phillips and Ward (2002) arise well from an abstract representational coding, this manipulation should not impact the effect. On the contrary, if the previous effect is a consequence of perceived affordances traduced by more general potentiations, wearing gloves inducing a particular grasp should invert the stimulus-response compatibility effect in the direction of the tank of the mug. This is precisely our hypothesis: considering that a large hand position is fitter with the manipulation of the tank of the mug and not anymore with its handle, the effects will be
inverted regarding Experiment 1 and the incompatible situation should be facilitating for the subjects.

## Method

Participants Twenty undergraduate students (19 women) from Paul Valéry Montpellier University aged from 18 to 41 years old $(M=22.45, S D=6.15)$ took part to this experiment and received course credits. All had normal or corrected to normal vision and were naïve to the purpose of the study. Among them, four left-handed were distributed into the two groups.

Apparatus and Materials The experimental setup and the materials remained the same as Experiment 1. The only change in the experimental design was that participants had this time to wear specific gloves. We constructed these gloves with Plaster bands in such a way that the participant's hands adopt the form of a large grasping position (i.e., a necessary position to grasp a mug by its tank).

Procedure The instructions were the same as Experiment 1. Nevertheless, participants were asked to wear the gloves and to place their hands on the table in front of them before beginning the experiment. Their hands were placed shoulder-width apart and with palms facing inward. Once the participants felt comfortable with the gloves and the pedalboard, they could begin the experiment.

## Results

Response errors accounted for $4.62 \%$ of the total of trials. There were no significant differences between error rates in compatible and incompatible situations. As in experiment 1, all response times exceeding 1250 milliseconds were removed from the analysis. Theses exclusions represented $17 \%$ of trials. The response time analysis between Compatible and Incompatible situations showed not statistically significant differences ( $p>.05$ ).

## Discussion

Apart from the fact that participants were wearing constraining gloves inducing a large grasp shape during the task, the second experiment was identical to experiment 1. This single difference has seemly altered the results in the way that there was no significant SRC effect anymore. However, contrary to our hypothesis, the hand constraint did not reverse the result's pattern but seemed to have interfered with the previous effect. This interference could represent an argument which is not in favor of the abstract coding. Indeed, if facilitation effects arise from an abstract spatial coding, there is no rational that a change applied to subject hands during foot-press responses impacts the facilitation effect. Nevertheless, this null result cannot be fully interpreted as it stands and there is a doubt as to whether the gloves acted on subjects. While the purpose of the gloves was to potentiate a large grasping position and thus promote
a facilitation effect directed toward the location of the mug tank, they may have been perceived by the subjects as a simple immobilization.

## Experiment 2b

Experiment 2a showed that a constraint applied to subject hands during a categorization task with foot-press responses seems to have significantly altered the facilitation effect previously observed in Experiment 1. It is unclear that this was due to a simple interference or to a conflict between a specific motor activation of the feet on one side and of another activation of hands on the other side. This could stem for an insufficient subject integration of the gloves possibilities. Indeed, in Experiment 2a, no rationales about the glove shaped were given to participants. Thus, in the present experiment, we proposed to new participants to perform a five minutes training in which they had to move a real mug along several drawn points on a sheet while wearing the gloves. The purpose of this training was to strengthen the potential disposition to grasp the mug tank with hands during the task. Our hypothesis is that a better integration of the grasping possibility of the gloves will produce a facilitation effect in favor of the side of the mug tank and thus, will facilitate incompatible responses regarding the handle location.

## Method

Participants A new sample of twenty undergraduate students (12 women) from Paul Valéry Montpellier University aged from 17 to 29 years old $(M=21.5, S D=$ 4.32) took part to this experiment and received course credits. All had normal or corrected to normal vision and were naïve to the purpose of the study. Among them, two left-handed were distributed into the two groups.

Apparatus and materials The experiment itself remained unchanged. The participants were preforming the same task while wearing the same gloves that in Experiment 2a. The difference was that they had to perform a training phase during which they had to move a real mug (with the same appearance that the one modelised for the experiment) on a sheet plotted course. The plotted course consisted in nine drawn circles. These circles were numbered from 1 to 9 . To ensure that participants remain focused during this phase, the mug was filled with water.

Procedure To perform the training phase, participants manipulated the mug while wearing the gloves (see Figure 2 ). They were told to move the filled mug circles by circles in the ascending and then in the descending order on the sheet with the left hand and after that, with the right hand. They had to put down the mug on each circle before moving on the next one. This course was repeated two times. Regarding the shape of the gloves, the participants were obliged to grasp the mug by its tank. No instructions were given concerning the better way to grasp the mug in this situation. The participants spontaneously grasped it by the
tank and performed the training without dropping the mug. After this training phase, they performed the same categorization task than the one proposed in the two previous experiments.


Figure 2: Plotted course for the training phase

## Results

Response errors accounted for $6.25 \%$ of the total of trials. There were no significant differences between error rates in compatible and incompatible situations. Response times exceeding 1250 milliseconds were excluded and represented also $6.25 \%$ of trials. Regarding the compatibility between response feet and handle locations, mean responses for the Incompatible situations ( $M=686 \mathrm{~ms} .14, S D=111.44 \mathrm{~ms}$ ) were faster than mean responses for the Compatible situations ( $M=713.66 \mathrm{~ms}, S D=126.76 \mathrm{~ms}$ ). A dependent t test revealed that the difference was statistically significant, $t(19)=2.64, p<.02, d=.23$. This pattern was therefore the reverse than the one observed in Experiment 1 (see Table 1).

Table 1: Summary of the chronometric results in milliseconds (with the standard deviations in parentheses) and associated $p$-values for the three experiments

|  | Compatible | Incompatible | $p$ |
| :--- | :--- | :--- | ---: |
| Experiment 1 | $713(76.77)$ | $749.96(93)$ | .02 |
| Experiment 2a | $766.41(105.80)$ | $786.73(127.22)$ | n.s. |
| Experiment 2b | $713.66(126.76)$ | $686.14(111.44)$ | .01 |

Complementary analysis We calculated the mean effect size differences between Compatible and Incompatible situations by subtracting the mean response times related to the Incompatible situation from the ones related to the Compatible situation for each subject and for each experiment. This allowed us to produce a value for each subject (i.e. positive if he was faster to respond in Compatible situations and negative if he was slower) and the size of this difference (see Figure 3). Regrouping those values for each experimental condition, we conducted a oneway between subjects ANOVA to compare the effect of

Experiments 1, 2 a and 2 b on mean effect size differences. There was a significant effect of the Experimental Conditions on subject Response times for the three conditions, $F(2,57)=7.42, p<.002, \eta_{\mathrm{p}}^{2}=.21$.

Post hoc comparisons using the Tukey HSD test indicated that the mean effect size for the Experiment 1 condition ( $M$ $=36.93, S D=64.62$ ) was significantly different than the one of the Experiment 2 b condition $(M=-27.51, S D=$ 46.54). The test indicated too that the mean effect size of the Experiment 2 a condition ( $M=20.32, S D=52.07$ ) was significantly different than the one of the Experiment 2b condition. However, the mean effect size of the Experiment 1 condition did not significantly differ from the one of the Experiment 2a.

Finally, considering that in the three experiments, our samples were quite inhomogeneous with respect to age, we performed the same analysis while excluding subjects older than 30 years. The results stayed unchanged.


Figure 3: Mean effect size differences (in milliseconds) as a function of the experimental conditions. Errors bars depict standard deviations.

## General Discussion

In Experiment 2b, subjects manipulated a real mug with constraining gloves before performing the categorization task. The only way to proceed was to grasp the mug by its tank and not by its handle which is in accordance with the purpose of the shape applied to the gloves. Results showed that a facilitation effect emerged from the incompatible situation regarding the classic stimulus-response compatibility paradigms. Thus, the training seems to have strengthened a potential reach to grasp gesture oriented to the mug tank and facilitated subsequent categorizations in situations wherein the responses and the mug tank were on the same side.

Taken together, the results suggest that facilitation effects observed in stimulus-response compatibility paradigms could well arise from specific motor activations instead of abstract coding like Philipps and Ward (2002) suggested it. Indeed, if the results observed in Experiment 1 were due to a representational abstract coding, no differences would be found applying a hand constraint during foot-press responses. Yet, in experiment 2 a , the effect was still in the
right direction but this time, it was not significant suggesting an interference in the subject disposition to respond. Furthermore, a real manipulation of a mug with a specific constraint inverted the location of the effect in Experiment 2b. This implies that the manipulation seems to have led to a sensorimotor integration changing the location of the potentiating part of the object. This result is also particularly interesting because it seems that no rationales can be found in the attentional shift hypothesis. Indeed, in all of our experiments, the stimulus stayed unchanged and hence, the perceptual asymmetry cannot be taken as the origin of these results.

Nevertheless, an alternative explanation could be as well proposed. Indeed, due to the manipulation phase, participants may have learnt to pay attention to the tank of the mug which could conduct to an attentional shift in the direction of this one. For further investigations, it would be interesting to expound on this by proposing a new experiment in which participants would realize the training but then, take off the gloves for the SRC experiment. This design should allow knowing if the results of the Experiment 2 a and 2 b are due to the wearing of the gloves or by sensorimotor integrations.

Regarding this possibility and considering that only a short training impacted the SRC effect directions. It is nevertheless possible that such attention-related effects be rooted in a sensorimotor process. For instance, a possible explanation could emerge if we replace our results within the framework of the premotor theory of attention (see Craighero, Fadiga, Rizzolatti \& Umiltà, 1999). According to this theory, orienting of attention implies an activation of basic circuits associated with the action goal. Therefore, the results of the Experiments 2 a and 2 b could arise from an attentional effect determined by the motor preparation induced by the training phase. This interpretation is in phase with both the specific motor activation account and the general ecological approach to perception. In this context, attentional shifts could be consecutive to premotor activations and be constitutive parts of the action-perception coupling.

In conclusion, the present study represents a further argument in favor of specific motor activations during perception of graspable objects. Nevertheless, it carries also some questions about its results and further works will be necessary to investigate such SRC modulation effects. More broadly, it underscores some imprecisions about the original propositions made by Gibson (1979). For instance, how a specific affordance can be perceived instead of another and how much the physical disposition to act in a given time impacts the subject's tendency to perceive them.

## Acknowledgments

Authors want to thank Jean-François Loisel for his help to design the visual stimuli used in this study.

## References

Anderson, S. J., Yamagishi, N., \& Karavia, V. (2002). Attentional processes link perception and action. Proceedings. Biological Sciences / The Royal Society, 269(1497), 1225-32.
Buccino, G., Sato, M., Cattaneo, L., Rodà, F., \& Riggio, L. (2009). Broken affordances, broken objects: A TMS study. Neuropsychologia, 47(14), 3074-3078.
Cho, D. T., \& Proctor, R. W. (2011). Correspondence effects for objects with opposing left and right protrusions. Journal of Experimental Psychology. Human Perception and Performance, 37(3), 737-49.
Craighero, L., Fadiga, L., Rizzolatti, G., \&Umiltá, C. (1999). Action for perception: A motor-visual attentional effect.Journal of Experimental Psychology: Human Perception \& Performance, 25, 1673-1692.
Fischer, M. H., \& Dahl, C. D. (2007). The time course of visuo-motor affordances. Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, 176(3), 519-24.
Gibson, J.J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.
Goslin, J., Dixon, T., Fischer, M. H., Cangelosi, A., \& Ellis, R. (2012). Electrophysiological examination of embodiment in vision and action. Psychological Science, 23(2), 152-7.
Michaels, C. (1988). SR compatibility between response position and destination of apparent motion: evidence of the detection of affordances. Journal of Experimental Psychology: Human Perception and Performance, 14(2), 231-40
Palamar, T. (2014). Mastering Autodesk Maya 2015 : Autodesk Official Press. Indianapolis, Indiana: John Wiley \& Sons, Inc.
Phillips, J. C., \& Ward, R. (2002). S-R correspondence effects of irrelevant visual affordance: Time course and specificity of response activation. Visual Cognition, 9(45), 540-558.

Stoffregen, T. A. (2003). Affordances as properties of the animal - environment system. Ecological Psychology.
Tucker, M., \& Ellis, R. (1998). On the relations between seen objects and components of potential actions. Journal of Experimental Psychology. Human Perception and Performance, 24(3), 830-46.
Tucker, M., \& Ellis, R. (2001). The potentiation of grasp types during visual object categorization. Visual Cognition, 8(6), 769-800.
Warren, W. H. (1984). Perceiving affordances: visual guidance of stair climbing. Journal of Experimental Psychology. Human Perception and Performance, 10(5), 683-703.

# People toss coins with more vigor when the stakes are higher 

Gregory L. Dam (gdam@iue.edu)<br>Department of Psychology, Indiana University East<br>2325 Chester Blvd., Richmond, Indiana, 47374, USA.


#### Abstract

We trust that the uncertainty regarding the outcome of a coin toss makes it a fair procedure for making a decision. Small differences in the force used to toss a coin should not affect this uncertainty. However, the voluntary movement involved in tossing a coin is subject to motivational influences arising from the anticipation of the value of the outcome of the toss. Presented here are measurements of hand velocities during coin tossing when the outcomes entail monetary gains and losses. Finger position measurements show that hand velocities are proportional to the amount of money at stake. Coin toss movements are faster and larger for higher stakes than for smaller monetary stakes.


Keywords: motor control, decision-making, affect, behavioral economics

## Introduction

Animals move faster to acquire larger rewards than to acquire smaller rewards (Kawagoe, Takikawa and Hikosaka, 1998; Choi, Pavan and Shadmehr, 2014). Delays in acquiring a reward decrease the probability of being successful and, since future rewards are temporally discounted, they may also decrease the subjective value of the reward. It has been demonstrated that the vigor of a movement, which is overtly expressed as the reaction time and velocity of the movement, are influenced by motivation effects arising from the cost or value of the outcome (Turner and Desmurget, 2010). For example, people show faster eye saccades when they are in more rewarding environments (Haith, Reppert \& Shadmehr, 2012) and monkeys make faster arm movement to higher reward targets (Opris, Lebedev \& Nelson, 2011). It is hypothesized that the purpose of the larger vigor observed in movements for greater rewards is to increase the probability of success and decrease the time to acquire the reward (Choi, Pavan and Shadmehr, 2014; Turner and Desmurget, 2010; Guitart-Masip, Duzel, Dolan \& Dayan, 2014). An open question is whether this vigor effect would be observed even when it is independent of the movement outcome. Such as in situations where changes in movement vigor do not influence the probability or timing of reward acquisition. For example, when tossing a coin for a monetary wager.

Coin tosses are used in sporting events such as cricket and American football to determine which team goes first. They are even used to settle the results of tied mayoral races in accordance with the law of many US states. When a game of chance is used to make a decision, the assumption is that there is sufficient uncertainty about the outcome to make the procedure fair. Even though coin tosses are the textbook example of uncertainty, coin tosses are entirely deterministic
physical processes. The coin trajectory can be predicted from initial conditions with Newton's laws of motion and Euler's equation for rigid body dynamics. However, small differences in the initial velocity and spin of the coin result in different outcomes. Mahadevan and Yong (2011) performed a phase space analysis of the probability distributions of coin outcomes as function of initial spin and vertical speed. The analysis reveals a high sensitivity of the coin's outcome to its initial spin and vertical velocity. We can assume that most coin flippers have no knowledge of how initial toss conditions map onto outcomes. Stewart (2014) suggested that perhaps the precision of human hand control is not accurate enough to reliably affect the outcome given the thin alternating regions in the phase space between heads and tails. Perhaps the smallest possible motor error in voluntary movement to toss a coin is spread across two or more outcome distributions in the phase space. Considering this possibility, one might conclude that the coin toss procedure itself is deterministic, but the initial conditions are random. Therefore, coin tossing is a special case where the movement vigor is independent of the outcome.

The outcomes of bodily movements produce the substantive consequences of behavior. It is therefore not surprising that organisms have adapted to perform movements precisely and efficiently. The escape vectors of cockroaches (Domenici, Blagburn \& Bacon, 2008), the foraging paths of bees (Reynolds et al., 2007) and the reaching arm movements of humans (Flash \& Hogan, 1985) all demonstrate optimal or near optimal movement performance. Experimental measures of human movement performance are predicted well by mathematically optimal models of movement behavior (Körding and Wolpert, 2006; Todorov and Jordan, 2002; Dam and Körding, 2009).

The success of optimal models for understanding movement stands in contrast to the descriptive models used to understand human judgment and decision-making. People demonstrate a variety of persistent and systematic biases in many domains of decision-making. A large body of research has shown that people are particularly prone to error during economic decisions. For example, in many situations people are loss averse, where they are about twice as unhappy with a monetary loss than they are happy with an equal magnitude monetary gain (Kahneman and Tversky, 1979). This leads to errors in decision making, such as a greater willingness to take a risk when potential losses are looming than when there is an equal potential gain to be had. Deviations from normative models have been traditionally attributed to distortions of judgments of value and probability made by the decision maker. Interestingly, the irrational distortions
observed in economic decision tasks are only partially present in mathematically equivalent motor tasks and demonstrate qualitatively different judgements of probability (Wu, Delgado and Maloney, 2009). This suggests that decision makers use information about the outcome value and probability differently when making economic decisions than when making motor decisions.

Tossing a coin for a monetary stake combines a voluntary movement with a cognitive evaluation of the outcome. For a person that has learned the motor skill of coin tossing, the movement is simple and involves little effort or deliberate planning. The hypothesis here is that the vigor of coin tosses will be affected by both the amount and valence (loss or gain) of the outcome. Specifically, the prediction is that the velocity and size of the coin tossing movements would be larger when the monetary stakes where higher. Additionally, considering loss aversion, movement velocities should be roughly twice a high when tossing a coin for a potential loss than for a potential gain of equal value.


Figure 1: An illustration of the coin toss movement made by participants during the experiment. Tosses were made with the coin placed on the back of the hand in a palm down position. This procedure for coin tossing increased the accuracy of measurements of hand velocities by allowing the tracking of individual finger tips.

## Method

## Participants

The experimental protocol was approved in accordance with Indiana University's policy statement on the use of human participants. Informed consent was obtained from fifty righthanded participants ( 21 male, 29 female). Participants were compensated a $\$ 5$ stipend prior to beginning the experiment. In addition to this stipend, participants were compensated according to the outcomes of the coins tosses as described below.

## Design

Each participant tossed a standard US quarter dollar coin with their dominant hand. The hand movements of the toss were
performed above an instrument designed to measured finger and hand position. There were six conditions resulting from a 2 (valence of outcome: loss or gain) X 3 (amount: 10ф or $25 \phi$ or $\$ 1$ ) factorial within-subjects design. The dependent variable was the vertical velocity of the right hand during the coin tossing movement.

## Materials

A Leap Motion Controller was used to measure hand movements during the experiment. The controller uses three infrared LED emitters and two cameras to track finger and hand position, directions of movement and velocities. The device's resolution is below 0.78 mm of movement (Oliveira \& Andrade, 2015). The cameras capture more than 290 frames a second from which position, velocity and direction of movement are computed to provide a sampling rate of 145 Hz .

## Procedure

The experiment was designed to measure whether the stakes of a coin toss would influence how the coin is tossed. At the beginning of each trial participants placed a quarter on the back of their dominant hand near the fingernails while standing. Participants were instructed to position their dominant arm with their elbow bent at 90 degrees and with their hand extended in the pronated position (palm down). The device was attached to a tripod that was adjustable vertically in height and placed 5 cm below the hand so that the tossing movement would take place within its effective workspace. This was done to assure accurate measures of hand position.

Figure 1 illustrates how the coin tosses were performed. Participants tossed the coin by accelerating it vertically into the air and then allowing the coin to bounce on the floor. The result was read from the coin as it lay on the floor. Each participant performed 30 coin tosses. Altogether 1378 coin tosses were measured from 1500 trials. The missing trials are due to participants performing the movement outside the device's effective workspace. Although some participants had a higher tendency to perform the toss outside of the devices' effective range, the missing trials are randomly distributed across conditions with $23,24,31,16,24$ and 34 trials missing data from conditions $+10 \phi,-10 \phi,+25 \phi,-25 \phi$, $+\$ 1$, and -\$1 respectively

During each trial the outcome of a single coin flip determined a monetary gain or loss to the participant. Half of the trials were gain conditions, where the participant stood to make money contingent on the outcome of the coin toss. During the other 15 trials, participants faced a potential monetary loss. The outcomes of all coin tosses were added to determine the total stipend for participation with a mean stipend of $\$ 5.67$ ranging from $\$ 5$ to $\$ 9.20$. The presentation of trial condition was randomized and fully balanced with


Figure 2: Hand velocities during three coin tosses. The two graph panels labeled 'typical toss' were chosen to represent the most common coin toss velocity profiles. The third panel labeled 'atypical toss' includes a pre-toss preparatory movement. Such pre-movement artifacts were excluded from the statistical analysis by using the maximum velocity as a measure of coin toss velocity.
each condition appearing exactly 5 times during the experiment.
At the beginning of each trial, the participant was informed of the trial condition by text that appeared on a 21 ' computer screen at a distance of 32 cm from their face. For example, the text: "If you lose this coin flip, you will lose $\$ 1$ " would appear during the $-\$ 1$ trial conditions. Additionally, the text was read aloud by the experimenter prior to commencing the coin toss movement. Before each toss movement, the participant called the toss, choosing either 'heads' or 'tails' by pressing the corresponding key ' H ' or ' T ' on a computer keyboard. If the coin landed heads up and the participant called heads, then the toss was regarded as a win. If the coin result did not match the call, the coin toss was regarded a loss. During gain conditions $(+10 \phi,+25 \phi$ or $+\$ 1)$, if the coin toss was won, the monetary amount at stake was added to the total stipend. If the outcome didn't match the call, the stipend remained unchanged. During loss conditions ( $-10 \phi,-25 \phi$ and $-\$ 1$ ), participants stood to lose the monetary amount if the toss was lost, or leave the stipend unchanged in the case of a match in outcome and call.

The current total monetary stipend amount was displayed on the screen and updated after each coin toss. If the total stipend at the completion of 30 coin tosses was higher than $\$ 5$, then participants were paid the difference in cash before completing the experiment. On the other hand, if participants finished with a total amount less than $\$ 5$, they were allowed to keep the $\$ 5$ show-up stipend. This was done to ensure that participants were not penalized for the outcomes of their tosses. On average, participants received $\$ 5.42$ ( $\mathrm{SD}=1.51$ ).

Economic decision making is often studied by measuring preferences between two or more lottery choices (Kahneman and Tversky, 1979). A lottery is a probability of obtaining an outcome that has an explicit value and can be expressed with the notation: [probability(outcome); value(outcome)]. For example, consider which lottery you would prefer: (a) five dollars for sure, or (b) a $50 \%$ chance of winning ten dollars and nothing otherwise. This choice between lottery can be
expressed as a choice between $(1, \$ 5)$ or $(0.5, \$ 10 ; 0.5,0)$, and most people prefer the sure outcome although according to rational choice theory we should be indifferent to the choice. Physically flipping a coin to decide an outcome is analogous to having selected a single lottery choice. If we assume that the procedure used for the coin toss is fair, then the probability will approach 0.5 as the number of tosses increases. For example, if a person cares about the outcome of a coin toss during the beginning of a sporting event, the coin toss lottery can be expressed as ( 0.5 , my team gets the ball; 0.5 , the other team gets the ball). How people value different prospects can be estimated by their preferences in lottery selections. The assumption in the current experiment is that the utility assigned to the potential outcome of the coin toss is reflected in the manner in which the movement is made.

## Data collection and measures

Position and velocity measurements were collected for 10 seconds after the participant called the toss and indicated that they were going to make the movement. The average duration of the coin tossing movement for all participants across all conditions was 157.37 milliseconds ( $\mathrm{SD}=65.94$ ), providing an average of 22.81 position measurements per movement. Since the instrument's measurements of fingertip positions are more accurate than for palm position, participants performed tosses with their palm down which allowed for the independent tracking of an average of 3.87 fingertips ( $\mathrm{SD}=$ 1.23 ) on each toss. The vertical velocity of the hand was calculated as the mean vertical velocity of all fingertip positions captured by the cameras during the toss.

## Results

Figure 3 shows mean hand velocities as a function of time with $95 \%$ confidence intervals. The mean hand velocity profiles are similar in shape to typical toss movements. However, hand movements in preparation for the toss are


Figure 3: Mean hand velocities with $95 \%$ CIs across participants as a function of time. The six conditions are collapsed by amount into three groups: $\pm 10 \phi, \pm 25 \phi$ and $\pm \$ 1$.
included in the analysis. In order to quantify the velocity of individual coin tosses, the maximum vertical velocity of the hand was computed for each toss. Analyses using maximum acceleration and integrated velocity across the entire movement produced similar results to those described below.

Overall, the tossing movement is highly conserved across participants as is evident in the aligned and averaged coin tosses displayed in Figure 3. However, there is systematic variance in toss velocities evident in the tendency by some participants to consistently toss the coin with more or less vigor than the average participant.


Figure 4: Mean maximum hand velocities with $1 \pm$ SEM bars for all six toss conditions.

Maximum hand velocities were analyzed using linear mixed-effects regression (LMER) with the lme4 package in R (Bates, Maechler, Bolker \& Walker, 2013). LMER allows for the controlling of random factors, such as the systematic
variability in toss force between participants. The monetary amounts at stake for each toss were modeled as fixed effects. A main positive effect of monetary value on hand velocities was observed, $F(1,1297)=3.82, \mathrm{p}=0.022$. Figure 4 shows the mean maximum velocity $\pm 1$ SEM for all six conditions ordered so that an effect of loss aversion is visible. However, the effect of valence (gain or loss) on hand velocity was not significant, $F(1,1297)=2.15, \mathrm{p}=0.14$. The interaction effect was non-significant, $F(1,1297)=0.35, p>.55$.
The mean maximum hand velocities for the $\pm 10 \phi, \pm 25 \phi$ and $\pm \$ 1$ conditions were $1463.17,1482.23$ and $1603.01 \mathrm{~mm} / \mathrm{sec}$ with $95 \%$ CIs [1335.44, 1590.71], [1090.95, 1873.56] and [906.69, 2299.48] respectively. The Cohen's $d$ effect sizes for monetary amount on maximum movement velocity were $d=$ 0.022 , for the difference between $\pm 10 \phi$ and $\pm 25 \phi$ conditions, $d=0.13$, for the difference between $25 \phi$ and $\pm \$ 1$ conditions, and $d=0.15$ for the difference between the $\pm 10 \phi$ and $\pm \$ 1$ conditions. Overall, the money at stake explains less than $8 \%$ of the variance in the maximum velocity of hand movements.

## Discussion

This study was designed to investigate the effects of the anticipation of monetary outcomes on the movements for coin tossing a coin. As hypothesized, larger monetary stakes resulted in higher velocities of the hand during coin flipping. Prospect theory predicts that people are often twice as sensitive to monetary losses than to equal magnitude gains. In the context of coin tossing for a potential loss, prospect theory predicts twice as large an effect of outcome valence than outcome magnitude on movement velocity. This prediction was not observed. One possibility is that the current experiment may lack sufficient statistical power to demonstrate a loss aversion effect on coin tossing movements. However, the results suggest that the hypothesized effect is much smaller than predicted by the results from traditional behavioral economics experiments (Kahneman, Knetsch and Thaler, 1990). Previous studies
have demonstrated that voluntary movements are not subject to distortions of value to the same degree as financial decisions as predicted by prospect theory, but are instead more subject to distortions of judgments of probability ( Wu , Delgado and Maloney, 2009). Evidence from the current experiment provides further evidence that simple movements are not influenced by the same mechanisms that are involved in making explicit financial decisions.

An impressive analysis of over 2.5 million golf putts of professional golfers at PGA golfing championships demonstrated that golfers putt more accurately when they are behind than when they are leading (Pope and Schweitzer, 2011). The analysis of laser measured golf putts revealed that golfers are more likely to make a shot when they are in a loss frame (i.e. a par putt) than when they are in a gain frame (i.e. a birdie putt). Unlike coin flipping, golf putting is a complex skill that is made possible by a variety of cognitive functions such as attention to task, planning, and explicit decisions about the movement. Pope and Schweitzer (2011) suggested that golfers deliberately consider that hitting the ball too softly may decrease the probability of a success, but nearly guarantee good placement for the subsequent putt, a desired position if one is ahead putting for birdie, but not desired if one is behind putting for bogey. These explicit forecasts of potential outcomes are subject to influences from the current and predicted affective states of the golfer (Kermer, DriverLinn, Wilson \& Gilbert, 2006). However, it is reasonable to assume that no such deliberate movement planning occurs during coin tossing.

Russell (1980) proposed a two-dimensional circumplex model of emotion where emotions are classified according to arousal and valence. Applying this model to the current experiment, the monetary value of the outcome corresponds to the arousal associated with the outcome. Valence corresponds to whether the outcome is a loss or a gain. A large body of research on approach-avoidance motor behavior has shown an effect of emotional valence on the automatic activation of movements (Markman \& Brendl, 2005; Lavender and Hommel, 2007 Maxwell \& Davidson, 2007). Research has shown that reaction times are quickest for approach movements towards positive stimuli (i.e. high valence) and for avoidance movements away from negative stimuli (i.e. low valence). In these studies, images are most often used to elicit emotions rather than money.

Experimentally manipulating both valence and arousal with image stimuli is difficult, because images that are rated as low valence are typically, if not exclusively, also rated as high arousal. For example, there are no images or sounds rated as low valence and low arousal in the over 1200 stimuli available in International Affective Systems (IAPS and IADS). It remains a possibility that arousal ratings of emotional stimuli may have a larger effect on approachavoidance movement response times and force than do valence ratings.

The experiment presented here was designed to independently measure the effects of monetary amount and valence on the velocity of coin tossing movements. The
results show that the magnitude of the outcome has a small effect on movement vigor. However contrary to prospect theory's prediction, the valence of the outcome has little or no additional effect on the coin tossing movement. This provides further evidence that outcome value is assessed differently during movement planning than during financial decision-making.

## Notes

The movement data, MATLAB analysis code and experimental protocol code are available at:
https://www.researchgate.net/project/People-toss-coins-with-more-vigor-when-the-stakes-are-higher

## Acknowledgments

I would like to thank Michael Frisby for his generous help with the statistical modeling of results.

## References

Bates, B., Maechler, M., Bolker, B. M. \& Walker, S. C. (2013) Fitting Linear Mixed-Effects Models using lme4, ArXiv e-prints, eprint arXiv: 1406.5823
Choi, J. E. S., Pavan, A., V., \& Shadmehr, R. (2014). Vigor of movements and the cost of time in decision making. Journal of Neuroscience, 34(4), 1212-1223.
Cronin, B. (2017, January 19). Super Bowl coin toss betting. Retrieved from https://www.pinnacle.com/en/betting-articles/football/super-bowl-coin-toss-betting
Dam, G \& Körding, K. (2009). Exploration and exploitation during sequential search. Cognitive Science, 33(3), 530541.

Diaconis, P., Holmes, S., \& Montgomery, R. (2007). Dynamical bias in the coin toss. Society for Industrial and Applied (SIAM) Review, 49(2), 211-235.
Domenici, P. B., Blagburn, J.M. \& Bacon, J.P. (2008). Cockroaches keep predators guessing by using preferred escape trajectories. Current Biology 18: 792-796.
Flash, T., \& Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. Journal of Neuroscience, 5 (7), 1688-1703.
Guitart-Masip, M., Duzel, E., Dolan, R., \& Dayan, P. (2004). Action versus valence in decision making. Trends in Cognitive Sciences, 18(4), 198-202.
Haith, A.M., Reppert, T.R. \& Shadmehr, R. (2012) Evidence for hyperbolic temporal discounting of reward in control of movements. Journal of Neuroscience 32: 11727-11736.
Jaynes, E. T. (1996), Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, UK.
Kahneman, D., \& Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.

Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263292.

Kahneman, D., Knetsch, J. L., \& Thaler, R. (1990) Experimental tests of the endowment effect and the coase theorem. The Journal of Political Economy, 98(6), 13251348.

Kawagoe, R., Takikawa, Y., \& Hikosaka, O. (1998). Expectation of reward modulates cognitive signals in the basal ganglia. Nature Neuroscience, 1(5), 411-416.
Körding, K. \& Wolpert, D. M. (2004). The loss function of sensimotor learning. Proceedings of the National Academy of Sciences, 9839-9842.
Körding, K. \& Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. Trends in Cognitive Sciences, 10, 319-326.
Kermer, D. A., Driver-Linn, E., Wilson, T. D. \& Gilbert, D. T. (2006). Loss aversion is an affective forecasting error. Psychological Science. 17 (8): 649-653.
Lavender, T., \& Hommel, B. (2007). Affect and action: Towards an event-coding account. Cognition \& Emotion, 21, 1270-1296.
Mahadevan, L. \& Yong, E. Y. (2011). Probability, physics and the coin toss. Physics Today, 66-67.
Markman, A. B., \& Brendl, C. M. (2005). Constraining theories of embodied cognition. Psychological Science. 16, 6-10.
Maxwell, J. S. \& Davidson, R. J. (2007). Emotion as motion: Asymmetries in approach and avoidant actions. Psychological Science, 18(12), 1113-1119.
Niv, Y., Daw, N. D., Joel, D. \& Dayan, P. (2006) Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl.) 191, 507-52.
Oliveira, F. H. M. \& Andrade, A. O. (2015) Preliminary evaluation of Leap Motion Controller as a human tremor record device. VIII Simpósio em Engenharia Biomédica (pp. 241-245). Uberlândia, Brazil.
Opris, I., Lebedev, M. \& Nelson, R. J. (2011) Motor planning under unpredictable reward: modulations of movement vigor and primate striatum activity. Frontiers in Neuroscience, 5(61), 1-12.
Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, Lau, H., Dolan, R. J. \& Frith, C. D. (2007). How the brain translates money into force: A neuroimaging study of subliminal motivation. Science, 316, 904-906.
Pope, D. G. \& Schweitzer, M. E. (2011). Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes. American Economic Review, 101:129-157.
Reynolds, A.M., Smith, A. D., Menzel, R., Greggers, U., Reynolds, D. R. \& Riley, J. R. (2007) Displaced honey bees perform optimal scale-free search flights. Ecology 88:1955-1961.
Russell, James (1980). A circumplex model of affect. Journal of Personality and Social Psychology. 39: 1161-1178.

Stewarts, I. (2011). Sources of uncertainty in deterministic dynamics: an informal overview. Philosophical Transactions of the Royal Society A, 369, 4705-4729.
Todorov, E. (2004). Optimality principles in sensorimotor control. Nature Neuroscience, 7, 907-915.
Todorov, E., Jordan, M. I. (2002). Optimal feedback control as a theory of motor control. Nature Neuroscience, 5, 12261235.

Turner, R. S., and Desmurget, M. (2010). Basal ganglia contributions to motor control: a vigorous tutor. Current Opinion Neurobiology, 20, 704-716.
Wu, S. W., Delgado, M., \& Maloney, L. T. (2009) Economic decision-making compared to an equivalent motor task. Proceedings of the National Academy of Sciences, 106, 6088-6093.

# Perspective-Taking in Referential Communication: Does Stimulated Attention to Addressee's Perspective Influence Speakers' Reference Production? 

Debby Damen (d.j.damen@uvt.nl), Per van der Wijst (per.vanderwijst @uvt.nl), Marije van Amelsvoort (m.a.a.vanamelsvoort@uvt.nl), Emiel Krahmer (e.j.krahmer@uvt.nl)

Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE, The Netherlands


#### Abstract

We investigated whether speakers' referential communication benefits from an explicit focus on addressees' perspective. Dyads took part in a referential communication game and were allocated to one of three experimental settings. Each of these settings elicited a different perspective mindset (none, selffocus, other-focus). In the two perspective settings, speakers were explicitly instructed to regard their addressee's (otherfocus) or their own (self-focus) perspective before construing their referential message. Results indicated that eliciting speakers' self- versus other-focus did not influence their reference production. We did find that speakers with an elicited egocentric perspective reported a higher perspective-taking tendency than speakers in the other two settings. This tendency correlated with actual referring behavior during the game, indicating that speakers who reported a high perspective-taking tendency were less likely to make egocentric errors such as leaking information privileged to speakers themselves. These findings are explained using the objective self-awareness theory.


Keywords: perspective-taking; referential communication; egocentricity bias; privileged information.

## Introduction

Engaging in successful referential communication implies that addressees are able to select the intended referent on the basis of speakers' descriptions. For this, speakers are expected to design their message optimally (i.e., audience design in Clark \& Murphy, 1982), adhering to addressees’ informational need (Clark, 1992). Speakers are supposed to exchange just the right amount of information, neither too little nor too much (Grice, 1975), and base their contributions on the knowledge, beliefs and assumptions that are shared or salient between themselves and their addressee (i.e., common-ground information). This is necessary, because addressees will rely on this shared, salient knowledge when interpreting the referential message (Arnold, Kaiser, Kahn, \& Kim, 2013). Referential communication thus relies a great deal on interlocutors' ability to accurately engage in the process of perspective-taking; the ability to take into account the knowledge and attentional state of their interaction partner at each step in the conversation. The questions that arise here are whether interlocutors are inclined to regard the other's perspective accurately during interaction, and if this is not the case, whether a stimulated attention to another's perspective would be beneficial for the referential communication process.

The literature shows a puzzling picture with regard to speakers' ability and propensity to accurately regard addressee's perspective and, thus, to engage in an accurate
audience design. On the one hand, studies evidenced that speakers succeed at assessing and adapting their communication to their addressees' knowledge (needs) (Heller, Gorman, \& Tanenhaus, 2012; Nadig \& Sedivy, 2002), whereas others have indicated that these adjustments are not always accurate (Horton \& Keysar, 1996; Keysar, Barr, Balin, \& Brauner, 2000). According to these latter studies, language production is not necessarily anchored to addressees' needs, but more to speakers' own knowledge and attentional state, resulting in utterances that are based on information immediately accessible to speakers themselves. Following this approach, addressee's knowledge is only considered in a later, optional stage in which speakers can consciously choose whether to adjust their language production to the common ground status (Horton \& Keysar, 1996). Scholars defending the latter view argue for speakers’ egocentricity bias (Keysar, Barr, \& Horton, 1998), entailing that speakers use their own mental state as a representational default to infer the one of their addressee (Epley, Keysar, Van Boven, \& Gilovich, 2004). Engaging in perspective-taking is then considered to be a cognitive effortful process that can result in egocentric anchor mistakes when speakers do not correct their automatic response. Research indicated that these errors are likely to occur in social interactions, as speakers sometimes refer to information not known to their addressee (Horton \& Keysar, 1996), or even leak privileged information that should have stayed confidential (Kaland, Krahmer, \& Swerts, 2014; Wardlow Lane, Groisman, \& Ferreira, 2006).

In a referential communication task, Wardlow Lane et al. (2006) evidenced speakers' informational leakage even when it bore negative consequences. During the task, speakers described geometrical objects to their addressees, with the goal of earning both of them points if the addressee correctly identified the referent. Before every description, speakers hid one object from their addressee's view. This object always differed in size from the target object speakers had to describe. Addressees could earn additional points by correctly guessing the identity of the hidden object. Although speakers were instructed not to let their addressee gain additional points, results showed that speakers were likely to cue the identity of their privileged object by referring to the size contrast they themselves were seeing. This was especially the case when the target object and speakers' privileged object were similarly rather than differently shaped, as the size contrast presented to speakers was then most relevant (i.e., salient) for speakers to discern.

Subsequent studies replicated findings of Wardlow Lane et al. (2006) by showing that speakers also leak information
non-verbally (Kaland et al., 2014), and especially when they do not have enough cognitive resources left to correct perspective mistakes (Wardlow Lane \& Ferreira, 2008). Intriguingly, speakers are even more likely to refer to privileged information when they are motivated to keep it confidential. The motivation to keep private information privileged further enhances its salience which, as a consequence, can result in it being revealed (Wardlow Lane \& Liersch, 2012).

It seems that despite their efforts speakers are not always able to monitor for perspective mistakes or to adjust their egocentric errors to addressees’ informational need. The question raised here is whether speakers' audience design would benefit from a constant reminder of interlocutor's informational need (i.e., perspective). Research has suggested that audience design is more likely to occur when speakers are made aware that such design is needed (Horton \& Gerrig, 2002). We therefore suggest that guiding speakers through a perspective-taking process might inhibit egocentric anchoring, and might boost their monitoring for perspective mistakes. This might incite speakers to correct for egocentric errors such as the leakage of privileged information (Horton \& Keysar, 1996), resulting in a references that are more accurately based on addressee's perspective, and less on speakers' own knowledge and attentional state.

## Current Study

This study examines whether speakers' elicited attention to addressee's perspective influences their reference production. Following the assumptions of the egocentricity hypothesis (Keysar et al., 1998), we expect speakers in a baseline setting (i.e., in which perspectives are not induced) to automatically anchor their referential expressions to their private knowledge, increasing the likelihood they will refer to this information, compared to other-focused speakers whose attention is focused on their interlocutor's perspective. We further hypothesize that self-focused speakers who are made explicitly aware of their own perspective will be more likely to leak privileged information than speakers referring in the baseline setting.

Based on the findings of Wardlow Lane et al. (2006), we additionally expect that speakers will be more likely to leak privileged information when this information is salient versus non-salient to them. That is, if speakers refer to a commonground figure (e.g., a circle) that has a size-contrasting match (e.g., a bigger circle) in their privileged ground, the size difference is relevant and, thus, salient to speakers themselves to discern. This in contrast to situations in which speakers are presented with a size-contrasting mismatch (e.g., a bigger triangle) in their privileged ground. The salience of the size contrast presented by matching rather than mismatching figures makes speakers more likely to add contrasting adjectives in their description of the target figure (e.g., "the small circle"), by which they leak privileged information. Finally, we expect that the salience of privileged information will interact with the induced perspective. Self-focused speakers are expected to be more likely to leak information
when it is salient versus non-salient, compared to the baseline setting. Since other-focused speakers explicitly focus on addressee's perspective, we expect these speakers to be less influenced by the salience of their private information, compared to the baseline setting.

## Method

## Participants

In total, 93 student-dyads $(N=186)$ participated in this study. The data of three dyads were excluded from analyses, due to an error in the experimental procedure $(N=2)$, or due to a low proficiency in the language of the experiment (Dutch) ( $N$ $=4)$. The analyses were based on 90 dyads in which the participants were randomly assigned either the role of the speaker ( 55 women, 35 men, $M_{\text {age }}=22.0$ years; age range 1834 years) or the role of the addressee ( 59 women, 31 men, $M_{\text {age }}=21.3$ years; age range 17-27). All participants were fluent in Dutch, did not experience problems at discerning the colors used in the study, and received a small remuneration for their participation.

## Design

The experimental design and procedure were replicated from Kaland et al. (2014), which in turn were inspired by Wardlow Lane et al. (2006). The experiment consisted of a referential communication task in which speakers were asked to describe mutually visible geometrical figures in such a way that the addressee could indicate the intended one out of a set of four. These four figures were physically presented on the table in between both interlocutors, and depicted on speakers' private computer screen. From their private computer screen, speakers were instructed to block one figure and, subsequently, to identify another figure on the table in front of them (figure 1). The occluded figure differed either in size or color from the three mutually visible figures. In our experiment, we replicated Kaland et al. (2014) privileged situation and added a perspective-taking manipulation. In this privileged setting, one object was always blocked from addressee's view and thus belonged to speaker's privileged ground.


Figure 1: The experimental setting in which the speaker (on the bottom) identified figures to the addressee (on the top).

## Materials

Eliciting Self- Versus Other-Focus Speakers' self- versus other-focus was manipulated by asking them explicitly to either regard their own (self-focus) or their addressee's (other-focus) perspective before they identified the target object. Participants were randomly assigned to one of the three communication settings (self-focus, other-focus, baseline), resulting in 30 speakers per setting. The selfversus other-focus was operationalized by asking speakers to answer a perspective question portrayed on the computer screen next to them. In the self-focus setting, speakers answered the question reinforcing their egocentric perspective: "Which four figures are visible to you?". This in contrast to the speakers in the other-focus setting who were asked to regard the perspective of their addressee: "Which three figures are visible to your addressee?". Speakers answered the question by selecting the figures on their private computer screen. To eliminate the possibility that the selffocused speakers would simply select all figures as a response to the question, a fifth figure was added to the figures presented on the computer screen.

To investigate the influence of our perspective manipulation, we allocated one third of the speakers to a baseline setting. In this setting, we did not reinforce speakers' self- versus other-focus. In this way, we were able to examine how speakers' reference production in the self- versus otherfocused settings would diverge from a baseline situation.

Salience of Privileged Information The salience of speakers' privileged knowledge was manipulated within communicative settings. Participants were confronted with 40 experimental trials, consisting of 20 salient and 20 nonsalient trials. In the salient trials, speakers' privileged figure was identically shaped to the target figure (e.g., both were circles) whereas in the non-salient trials both figures were differently shaped (e.g., a circle and a triangle). The salient trials were designed to elicit utterances that contrasted the target figure with the privileged one, whereas the non-salient trials assessed how often speakers included adjectives irrespective of the contrast presented. Figures in successive trials were never identically shaped, and half of the figures contrasted in size (big, small) and the other half in color (red, blue, green, black, grey, yellow) (Kaland et al., 2014). The figures' shape, color, and position were balanced across all trials. This resulted in $3 \times 2 \times 2$ design, with communication setting (self-focus, other-focus, baseline) as a between subjects' factor, and trial type (salient, non-salient), and contrast type (color, size) as within subject factors.

## Procedure

A throw of a dice decided which participant took the role of the speaker. Participants were told that, when the addressee was able to correctly identify the target figure, both the speaker and the addressee would obtain one point. Participants were told that failing to identify the target figure would result in zero points obtained, and the goal of the game was to obtain the maximum number of points.

Speakers and addressees sat down on opposite sides of a table. Speakers were seated next to a computer screen on which the experimental trails were presented using E-Prime version 2. At the beginning of each trial, addressees closed their eyes while the experimenter placed four cards on the table. When the four cards were put in place, speakers (a) hid one figure from their addressee's view by placing an occluder between the figure and their addressee. Subsequently in the other- and self-focused setting, speakers (b) answered a perspective question by selected either the three figures visible to their addressee (other-focus) or the four figures visible to them (self-focus). Hereafter, speakers (c) described the target object with just enough information so that their addressee was able to identify the intended figure. Speakers were instructed to look at the four cards on the table when referring to the target object. While hearing speakers refer to a figure, addressees opened their eyes and pointed at the intended figure on the table in front of them. Speakers subsequently (d) informed their addressee whether their selection was correct. Since speakers in the baseline setting were not confronted with a perspective-taking manipulation, these speakers only performed actions (a), (c), and (d). To ensure all steps of the procedure were executed correctly, the experimenter was present during the entire game.

The experimental game ended after 40 rounds. After the final round, speakers indicated on a ten-point scale to what extent they took into account their addressee's perspective during the game $(1=$ not at all, $10=$ very much $)$. Since audio recordings were made of all sessions, participants' consent to making these recordings and using them for scientific purposes were collected. Afterwards, all participants were debriefed.

## Coding

To measure speakers' reference to privileged information (RPI), we counted the adjectives that matched the contrast between the target and privileged figure. Adjectives that did not contrast the target figure to the privileged one were not taken into account. Speakers' RPI was calculated as a proportion ( $1=$ contrasting adjective uttered; $0=$ no contrasting adjective uttered).

## Results

All dyads obtained the maximum of 40 points, indicating that they were able to correctly identify all targets. In figure 2, the mean proportions of speakers' informational leakage (RPI) as a function of the perspective manipulation (baseline, otherfocus, self-focus), whether the target and speakers' privileged figure were similarly (salient trials) or differently (nonsalient trials) shaped, and whether these contrasts were presented in either color or size are shown. Overall, speakers in the baseline setting referred to privileged information in half of the produced references ( $50 \%$ ), followed by the otherfocused ( $45 \%$ ), and self-focused speakers ( $29 \%$ ). Across the three communicative settings, speakers seem to have referred to privileged information to the same degree for salient (43\%) and non-salient (40\%) trials.


Figure 2: Mean proportions of speakers' RPI. Error bars represent $95 \%$ confidence intervals.

The influence of the perspective manipulation and the interplay with the salience of speakers' privileged information on the probability of privileged information to be mentioned was analyzed using a generalized linear mixed model analysis with a binomial distribution. For this we used the GLMER function from the lme4 package (Bates, Maechler, \& Bolker, 2011) in R (version 3.3.0; www.rproject.org). We constructed a maximal model that included a full random effect structure (Barr, Levy, Scheepers, \& Tily, 2013). This maximal model included the perspective manipulation (self-focus, other-focus, baseline), the salience of the trials (salient, non-salient), and the contrast (color, size) presented in the trials as fixed factors. We included random intercepts and slopes for both speakers and experimental trials. The probability distribution was set on binomial with a logit link function and we used parametric bootstrapping over 100 iterations to estimate the confidence intervals and $p$-values. When the maximal model did not converge, we excluded random slopes with the lowest variance until convergence was reached. We report the results of the models that were the first to converge (Barr et al., 2013). An alpha level of .05 was used for all statistical tests. The models treated the baseline setting as the reference category, to which speakers' RPI in the other- and selffocused settings were contrasted.

Influence of Perspective on Speakers' RPI Speakers' RPI in the self- and other-focused setting did not significantly differ from speakers' RPI in the baseline setting. For nonsalient size trials, speakers in the other-focused ( $M=.33, S D$ $=.45, b=0.80, S E=2.07$, CI: [-3.02, 5.11]), and self-focused setting ( $M=.24, S D=.41, b=1.28, S E=1.66, \mathrm{CI}:[-1.98$, 4.52]), were just as likely as the baseline speakers ( $M=.44$, $S D=.50$ ) to refer to privileged information. The same held for non-salient color trials: other-focused ( $M=.55, S D=.47$, $b=1.24, S E=1.90, \mathrm{CI}:[-1.67,5.76])$, and self-focused speakers' $\operatorname{RPI}(M=.31, S D=.43, b=-3.31, S E=4.41$, CI: [$12.06,5.21]$ ) did not significantly differ from the baseline ( $M$
$=.54, S D=.50$ ). This pattern also held for salient size trials: speakers' RPI in the other- $(M=.34, S D=.44, b=-0.07, S E$ $=1.61$, CI: $[-2.92,3.38]$ ), and self-focused setting ( $M=.24$, $S D=.41, b=0.98, S E=1.56, \mathrm{CI}:[-1.86,4.25])$ did not significantly differ from the baseline ( $M=.46, S D=.50$ ). Finally, speakers' RPI on salient color trials in the other- ( $M$ $=.58, S D=.46, b=0.57, S E=2.36, \mathrm{CI}:[-3.38,5.88]$ ), and self-focused setting ( $M=.35, S D=.42, b=-3.26, S E=4.41$, CI: [-12.64, 4.67]) did also not significantly differ from the baseline ( $M=.55, S D=.49$ ).

Influence of Salience on Speakers' RPI In the baseline setting, the salience of privileged information did not influence speakers' RPI. Baseline speakers were just as likely to refer to privileged information on non-salient ( $M=.44, S D$ $=.50)$ and salient $(M=.46, S D=.50)$ size trials $(b=1.53, S E$ $=0.81$, CI: $[-0.33,2.86])$, and on non-salient $(M=.54, S D=$ $.50)$ and salient $(M=.55, S D=.49)$ color trails $(b=0.46, S E$ $=2.04$, CI: [-3.05, 4.93]).

Baseline speakers' RPI was also not influenced by the contrast presented in the trials. Speakers were just as likely to refer to privileged information on non-salient size ( $M=.44$, $S D=.50$ ) and non-salient color ( $M=.54, S D=.50$ ) trials ( $b$ $=-1.91, S E=1.93$, CI: $[-6.59,0.97])$, as on salient size $(M=$ $.46, S D=.50)$ and salient color $(M=.55, S D=.49)$ trials $(b$ $=-1.91, S E=2.23$, CI: $[-6.81,1.95])$.

When the two perspective settings were contrasted to the baseline setting, no significant differences were found. Like the speakers in the baseline setting, other-focused speakers' RPI did not differ between salient ( $M=.34, S D=.44$ ) and non-salient $(M=.33, S D=.45)$ size trials $(b=-0.88, S E=$ 0.90 , CI: [-2.61, 0.93]), nor between salient ( $M=.58, S D=$ .46) and non-salient ( $M=.55, S D=.47$ ) color trials ( $b=-$ $0.79, S E=1.13, \mathrm{CI}:[-3.10,1.32])$. The same held for the selffocused speakers. Their RPI did not differ significantly between salient ( $M=.24, S D=.41$ ) and non-salient $(M=.24$, $S D=.41$ ) size trials $(b=-0.30, S E=0.90, \mathrm{CI}:[-2.03,1.51])$, nor between salient ( $M=.35, S D=.42$ ) and non-salient ( $M=$ $.31, S D=.43$ ) color trials $(b=-0.24, S E=1.16, \mathrm{CI}:[-2.58$, 1.97]).

Like the baseline speakers, other- and self-focused speakers' RPI did not depend on the contrast presented in the trials. Other-focused speakers' RPI did not significantly differ between salient size ( $M=.34, S D=.44$ ) and salient color $(M=.58, S D=.46)$ trials $(b=0.64, S E=2.71, \mathrm{CI}$ : $[-$ 4.33, 6.27]), nor between non-salient size ( $M=.33, S D=.45$ ) and non-salient color $(M=.55, S D=.46)$ trials $(b=0.64, S E$ $=3.45$, CI: $[-5.32,8.20])$. Further, self-focused speakers' RPI did not significantly differ between salient size ( $M=.24, S D$ $=.41)$ and salient color $(M=.35, S D=.42)$ trials $(b=-4.24$, $S E=4.43$, CI: [-13.71, 3.64]), nor between non-salient size ( $M=.24, S D=.41$ ) and non-salient color trials $(M=.31, S D$ $=.43)(b=-4.24, S E=5.11, \mathrm{CI}:[-14.67,5.37])$.

Speakers' Self-Reported Perspective-Taking A one-way between-subjects ANOVA revealed that speakers' selfreported perspective-taking tendency significantly differed
between settings, Welch's $F(2,57)=4.43, p<.05$. Tukey HSD post-hoc comparisons revealed that self-focused speakers ( $M=7.73, S D=2.94$ ) reported a significant higher perspective-taking tendency than both other-focused ( $M=$ 5.62, $S D=3.63$ ) and baseline speakers $(M=5.60, S D=3.51)$ (both $p<.05$ ). Perspective-taking tendencies did not significantly differ between the other-focused and the baseline setting ( $p>.05$ ). To investigate whether speakers' self-reported perspective-taking tendency corresponded with their actual behavior during the game, a follow-up logit mixed model analysis was conducted. This model included speakers' SELF-REPORT as fixed effect, a random intercept for subjects, and a by-subject random slope for the effect of SELFREPORT. $P$-values were obtained using the Likelihood Ratio Test (LRT). The LRT revealed that speakers' SELF-REPORT was a significant predictor of their actual RPI, $\chi^{2}(2)=9.90$, $p<.001$. As speakers' perspective-taking tendency increased, they were less likely to have leaked privileged information during the game, $b=-2.75, S E=0.45, p<.001$.

## Discussion

In this paper we studied whether eliciting speakers' selfversus other-focus would influence their subsequent reference production. We found that speakers in the otherand self-focused settings were just as likely to refer to privileged information as the speakers whose perspectivetaking was not manipulated (i.e., in the baseline setting). Further, we did not replicate the results of (Kaland et al., 2014; Wardlow Lane \& Ferreira, 2008; Wardlow Lane et al., 2006; Wardlow Lane \& Liersch, 2012) who found that the salience of privileged information can boost the probability of it being leaked. In our study, speakers were just as likely to refer to private information, regardless of its salience. Perhaps speakers' tendency to retain a certain reference strategy throughout the game could have interfered with their audience design (Horton \& Gerrig, 2002), and the extent to which they were influenced by the elicited perspective and the salience of their privileged knowledge.

In our study, $66 \%$ of speakers $(N=59)$ either referred to color and size contrasts on all trials, or they refrained from including any adjectives throughout the game. Speakers' consistent referring behavior has been supported by previous research (Brennan \& Clark, 1996), and is strengthened by addressees' ability to identify the referent on the basis of speakers' descriptions (Clark \& Krych, 2004). Addressees partaking in our study were always able to correctly identify the intended object, regardless of the presence or absence of speakers' informational leakage. Each time addressees correctly identified the target object, they signaled to speakers that the reference had been successful. This could have inspired speakers to keep hold of their referential tactic and to base their references on previous formulated descriptions. This tendency to be consistent could have interfered with our perspective-taking manipulation and the extent to which speakers were influenced by (the salience of) their privileged knowledge. Furthermore, in our study speakers' leakage did not bear negative consequences. As a
result, egocentric errors were not detected and speakers were not encouraged to adjust their reference production. This implies that increasing speakers' awareness of the negative consequences associated with their leakage could reduce the extent to which they would leak such information. However, as previous research has shown (e.g., Kaland et al., 2014; Wardlow Lane et al., 2006) incentives to keep privileged information confidential might increase speakers' attention to this information, thereby ironically boosting the likeliness of it being mentioned. Enhancing speakers' awareness of the negative consequences of their leakage thus might not be the right solution. There are, however, other factors that should be considered with regard to addressing speakers' consistency in reference production.
One of these factors is the self-paced method by which speakers were confronted to the instructions and perspective manipulation. The self-paced method could have induced the routineness by which speakers performed the instructions and completed the trials. Moreover, the fact that the perspectivetaking manipulation was posed on speakers' private computer screen in which perspectives were not visibly marked could have reduced the intrusiveness of the elicited mindsets. Although speakers were explicitly trained to return their attention from their private screen to the physical context shared between them and their addressee before they identified the target figure, the possibility exists that speakers were still regarding their private screen (in which perspectives were not marked) while formulating their reference. These issues could be addressed in a future study by allowing the experimental leader to pace the experiment and to expose speakers in the shared physical context to the perspective-taking manipulation. For example, speakers could be explicitly asked to indicate which figures are visible to their addressee (i.e., eliciting an other-focus) or visible to themselves (i.e., eliciting a self-focus) by using the figures lying between them and their interlocutor.

Moreover, following the design of Kaland et al. (2014), speakers were confronted with six color manipulations compared to the two size manipulations. The obtrusive use of color could have induced speakers to refer to color contrasts on all of the experimental trials (Koolen, Goudbeek, \& Krahmer, 2013), irrespective of the elicited perspective or the salience of privileged information. A future study could explore this possibility by equalizing the number of colors used to the number of size contrasts employed in the game.

An interesting finding of this study that merits further attention is the result of speakers' self-reported perspectivetaking tendency and its relation to their reference production. Ironically, speakers with an elicited self-focus reported to have regarded their addressee's perspective more than the speakers in the other two settings. This self-reported tendency correlated negatively with speakers' previous leakage behavior, indicating that speakers with a high perspectivetaking tendency were less likely to have leaked private information during the game. It thus seems that not an elicited other- but instead a self-focus activated speakers' awareness
of their interlocutor's informational need, reducing the likelihood of egocentric perspective errors to occur.

Differences in speakers' self-report and leakage between the self-focused and baseline setting can be explained by the presence or absence of the perspective manipulation. In the self-focused setting, speakers answered a question that enhanced their own mental representation of the scene, whereas in the baseline setting, perspective enhancements were absent. A more intriguing finding, however, is the occurrence of a stronger perspective-taking tendency by the self-focused speakers than by the other-focused speakers. This tendency can be explained using the objective selfawareness theory (Wicklund, 1975). According to this theory, self-aware persons reflect on themselves as if they are an object under scrutinization. Under this scrutinization, the difference between their actual and required behavior, derived from the standards that apply to the interaction, becomes salient. Our self-focused speakers could have found themselves in such a reflective state, especially since a cue of their addressee's different perspective was present (Gendolla \& Wicklund, 2009). Speakers were able to see which figures were available for addressee's selection process (and which one was not). As a consequence, self-focused speakers could have been more aware of addressees' informational need than other-focused speakers, reducing the extent to which they were influenced by privileged information. This possible explanation needs further examination by exploring how much the self- versus other-focus perspective questions used in this study elicited speakers' self-awareness. For this, the validated Situational Self-Awareness Scale can be employed (Govern \& Marsch, 2001).

## References

Arnold, J. E., Kaiser, E., Kahn, J. M., \& Kim, L. K. (2013). Information structure: linguistic, cognitive, and processing approaches. Wiley Interdisciplinary Reviews: Cognitive Science, 4(4), 403-413.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bates, D., Maechler, M., \& Bolker, B. (2011). Ime4: Linear mixed-effects models using S4 classes. R package version 0999375-38. Retrieved from htttp.CRAN.Rproject.org/packages $=$ Ime4.
Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology-Learning Memory and Cognition, 22(6), 1482-1493.
Clark, H. H. (1992). Arenas of Language Use. Chicago: University of Chicago Press.
Clark, H. H., \& Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language, 50(1), 62-81.
Clark, H. H., \& Murphy, G. L. (1982). Audience design in meaning and reference. Advances in Psychology, 9, 287299.

Epley, N., Keysar, B., Van Boven, L., \& Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. Journal of Personality and Social Psychology, 87(3), 327-339.
Gendolla, G. H., \& Wicklund, R. A. (2009). Self-focused attention, perspective-taking, and false consensus. Social Psychology, 40(2), 66-72.
Govern, J. M., \& Marsch, L. A. (2001). Development and validation of the situational self- awareness scale. Consciousness and Cognition, 10(3), 366-378.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. L. Morgan (Eds.), Syntax and semantics. Volume 3: Speech acts. New York: Seminar Press.
Heller, D., Gorman, K. S., \& Tanenhaus, M. K. (2012). To name or to describe: shared knowledge affects referential form. Topics in Cognitive Science, 4(2), 290-305.
Horton, W. S., \& Gerrig, R. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. Journal of Memory and Language, 47, 589-606.
Horton, W. S., \& Keysar, B. (1996). When do speakers take into account common ground? Cognition, 59(1), 91-117.
Kaland, C., Krahmer, E., \& Swerts, M. (2014). White bear effects in language production: Evidence from the prosodic realization of adjectives. Language and Speech, 57(4), 470-486.
Keysar, B., Barr, D. J., Balin, J. A., \& Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. Psychological Science, 11(1), 32-38.
Keysar, B., Barr, D. J., \& Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. Current Directions in Psychological Science, 7(2), 46-50.
Koolen, R., Goudbeek, M., \& Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. Cognitive Science, 37(2), 395-411.
Nadig, A. S., \& Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. Psychological Science, 13(4), 329336.

Wardlow Lane, L. W., \& Ferreira, V. S. (2008). Speakerexternal versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6), 1466-1481.
Wardlow Lane, L. W., Groisman, M., \& Ferreira, V. S. (2006). Don't talk about pink elephants! Speakers' control over leaking private information during language production. Psychological Science, 17(4), 273-277.
Wardlow Lane, L. W., \& Liersch, M. J. (2012). Can you keep a secret? Increasing speakers' motivation to keep information confidential yields poorer outcomes. Language and Cognitive Processes, 27(3), 462-473.
Wicklund, R. A. (1975). Objective self-awareness. Advances in Experimental Social Psychology, 8, 233-275.

# Sex-Dependent Effects of Emotional Subliminal Visual Stimuli on a Decision-Making Task 

Christopher L. Dancy (christopher.dancy@bucknell.edu)<br>Department of Computer Science, Bucknell University, Lewisburg, PA 17837

Frank E. Ritter (frank.ritter@psu.edu)
College of IST, Penn State, University Park, PA 16802
Frank G. Hillary (fhillary@psu.edu)
Department of Psychology, Penn State, University Park, PA 16802
Kevin Voller (krvoller@gmail.com)
Senior UX Designer, Dick's Sporting Goods eCommerce, Pittsburgh, PA


#### Abstract

How do covert emotional stimuli affect decisionmaking? We investigated this question by exposing participants to subliminal visual stimuli during a computerized version of the Iowa Gambling Task (IGT) to assess whether different categories of images (negative, neutral, or positive emotional evaluations) would influence deci-sion-making behavior. Results did show sex-group interactions for IGT scores. In decision learning model simulations, it was found that different models were more appropriate to explain the task performance for different sex-group pairs. Overall, women showed more of an ability to integrate the additive negative signals from the stimuli to make more advantageous decisions than the men; consequently, this made the men more resilient to the negative effects of the positive stimuli on taskperformance. When taken with existing research, the results indicate that subliminal emotional stimuli can have subtle, potentially sex-dependent, effects on behavior during the decision-making process.


## Keywords: Decision-Making, IGT, Emotion, Simulation

## Introduction

How do covert emotional stimuli affect decisionmaking and choice behavior? There have been several studies that have explored the processes involved in, and the outcomes of, decision-making behavior (e.g., see Lerner et al., 2015; Weber \& Johnson, 2008), but relatively few studies that explore decision-making have also explicitly introduced emotional stimuli (Phelps \& Sokol-Hessner, 2012) and even fewer have sought to understand the interaction between unconsciously presented emotional stimuli and decisionmaking. One decision-making study by Winkielman et al. (2005) found that subliminally presented images (emotional faces) influenced judgment and choice during a series of decisions directly following the masked image exposure (with happy faces increasing the
amount of a beverage poured and consumed, and the purchase price a participant would be willing to pay). These images affected the decision despite not being consciously perceived or semantically related to the series of decisions made after the exposure. Subliminal image presentation can also cause changes in peripheral physiology that may not be perceived, particularly measures related to activation of the LC-Noradrenergic system and amygdala, for example heart-rate and eyeblink response (e.g., Ruiz-Padial et al., 2011). These effects on peripheral physiology are important, as the areas of the brain that are shown to respond to these subliminal stimuli are likely causing these cascades of changes (Öhman \& Mineka, 2001; Tamietto \& de Gelder, 2010).

The Iowa Gambling task (IGT) has been used to better explain links between changes in peripheral physiology and choice behavior, as well as to better understand some of the brain areas key to decision-making and related physiological behavior during the decisionmaking process (Bechara et al., 1999).

IGT subjects repeatedly chooses cards from 4 decks of cards. The payoff per card varies, and the subject is asked to maximize their payoff. The decks differ in the payoff they give for each card; some decks give better average payoffs than others, although all have variability. The task is used to study how subjects learn to use the payoffs in their decision-making. For some cognitive deficits the choices are not learned very quickly.

An important finding from Bechara et al. (1999) is that normal participants exhibited different skin conductance response (SCR) behavior than clinical patients with amygdala lesions and those with lesions in the ventromedial prefrontal cortex (VMPFC). These distinct clinical groups exhibited different SCRs both prior to selecting a card from a deck and in response to receiving a net gain or loss after selecting a card; this difference is especially apparent in the disadvantageous
decks (those decks that had a negative average payoff). The group with amygdala lesions exhibited both a reduced SCR prior to selecting a card from a deck and a reduced SCR after receiving a reward or loss, while the VMPFC group showed a more attenuated SCR prior to the selection of a card from a deck, indicating that amygdala nuclei may play an important role in giving affective value to representations in decision-making. SCR response patterns by those with amygdala lesions indicated a difficulty with coupling an affective value with the different decks and the cards from those decks.

We sought to better understand behavioral effects of this unconscious emotion perception and decisionmaking interaction by exposing study participants to subliminal emotional stimuli while they completed the IGT. Behavioral responses to the affective value of objects are mediated by cognitive processes that are modulated by neural processing in the amygdala (Moscarello \& LeDoux, 2013; Panksepp et al., 2011; Phelps, 2006). Given that the amygdala is also important for the processing of unconsciously presented emotional stimuli (Tamietto \& de Gelder, 2010), the unconscious perception of emotional stimuli may have behavioral effects on decision-making even if the stimuli are not integral to the decisions being made (e.g., Winkielman et al., 2005).

We expected that decision-making would differ depending on whether the subliminal image presented had negative, neutral, or positive evaluations. We present a study to test this hypothesis. Normally, in nonpathological populations, IGT performance is largely dependent upon learning deck contingencies over-time. This can be represented somewhat as a reinforcement learning process (e.g., with the expectance-valence model or the prospective-valence model; Ahn et al., 2008). To further explore the potential differences between groups (and, later, participant sex), we developed decision learning models (e.g., Ahn et al., 2008) that were run in simulations ${ }^{1}$; this gave us the opportunity to understand potential computational processes affected by the treatment.

## Method

97 undergraduate students were recruited as participants for this study ( 52 males; 45 females). The average ages of males and females were similar at 20.7 and 19.8 (respectively). Electrodermal Activity (EDA) data were collected for the final 66 ( 37 males and 29 females) participants (data not reported here). All participants were given college course extra credit.

A filter process that removed participants who completed less than $20 \%$ of their trials due to time re-

[^311]strictions (max 3.5 s per trial) resulted in the removal of 4 participants' data from further analysis; data from 93 total participants were analyzed. The negative, neutral, and positive (image) groups each had 31 participants. We ceased participant enrollment in the study after we crossed a 31 per-group threshold for task-related behavioral analysis and all volunteers had the opportunity to participate.

Participants used a version of the IGT that included a fixed reward and punishment schedule for each deck that was the same as the schedule used for the original IGT by Bechara et al. (2000). A modified computerized version of the IGT was used that runs in Matlab and uses the Psychtoolbox Matlab extensions (Brainard, 1997), which were used due to their high timing accuracy, community support, and cross-platform availability. The specific software used has had IGT-specific timing tests done to confirm timing accuracy (Dancy \& Ritter, 2017).

The visual stimuli presented during the IGT were obtained from the International Affective Picture System (IAPS; Lang et al., 1997). Table 1 lists the images used in the image sets used by the groups. Male and female pictures were matched so that, for each group, they had similar valence/arousal/dominance ratings and had a similar content subject; for example, some snake pictures had different ratings between sexes within the IAPS manual, so those images with lower valence/higher arousal ratings among the same category were chosen. Given that picture ratings in all categories differed between sexes, this method allowed more consistency in mean measured quantitative ratings among participant sexes.

Table 1. The IAPS images used in the experiment.

| Picture-Set | Picture Numbers |
| :---: | :---: |
| Negative $_{\text {Male }}$ | $1050,1202,1220,1304,1525$ |
| Negative $_{\text {Female }}$ | $1050,1120,1201,1202,1525$ |
| Neutral $_{\text {Male }}$ | $1670,7006,7010,7080,7175$ |
| Neutral $_{\text {Female }}$ | $1670,7004,7010,7012,7175$ |
| Positive $_{\text {Male }}$ | $4180,4210,4232,4664,8501$ |
| Positive $_{\text {Female }}$ | $4505,4525,4660,8001,8501$ |

## Procedure

Before participating in the study, all participants read and signed a consent form approved by the Office of Research Protections (ORP) at Penn State. After consenting to the form, all participants filled out a Positive and Negative Affect Schedule (PANAS) questionnaire. All participants who had their EDA recorded were then fitted with a Q sensor EDA device (Ming-Zher et al., 2010).

Each participant was assigned to one of three groups that determined which set of images they were shown: (a) a negative image group that consisted of images
with a low rated valence, and a high arousal; (b) a neutral image group that consisted of images with a medium rated valence and a low arousal (c) a positive image group that consisted of images with a high rated valence and a high arousal.

In this version of the IGT, participants had a maximum of 3.5 seconds to select a card from one of the four decks and if they failed to make a selection in the allotted time on any trial, a random deck was selected for them. After a card was selected from a deck, a red or black card was shown for 4 seconds. A (groupdependent) image was shown in place of the background image of the box where the reward and losses were shown for 17 ms when the participant selected from deck A or B. If the participant made a selection from deck C or D , a plain gray background image was shown for 17 ms . Directly after this 17 ms , the reward and loss that the participant received in response to their deck selection was presented in the same box for 3.5 seconds. All images used from the image set, as well as the background image used throughout the task, were converted to gray scale. Each intertrial break lasted 3.5 seconds, except every $20^{\text {th }}$ trial, after each of which the participant was asked two questions to assess their awareness of the task itself.

After the IGT was completed, participants filled out a second PANAS questionnaire and then (as needed) had the EDA device removed. Participants were then asked "Did you discover anything new by the end of the game?" and were partially debriefed on the task itself. Participants then completed the Affective Neuroscience Personality Scales (ANPS) questionnaire and were fully debriefed before ending the study session.

## Results

To understand the results, we used both traditional statistical techniques (e.g., one-way and repeated measure ANOVAs), as well as results from decision learning models. While the ANOVAs were useful for understanding general differences between participant groups, the decision-learning models allowed us to explore potential differences in the computational processes that may govern group differences.

## Analysis of Initial Hypotheses

We hypothesized that the score would be different between the negative, neutral, and positive image groups, and that performance would be highest in the negative image group and lowest in the positive image group. A one-way ANOVA for cumulative score (total score at the end of the IGT) did not show a statistically significant difference between groups $(F(2,90)=$ $0.81, \eta^{2}=.02$ ). We also hypothesized that score would improve over the course of the task (indicating a learning of the advantageous decks) and that this im-
provement would also differ between groups. A 3 (image group) x 5 (block) mixed-model ANOVA for score revealed a statistically significant block factor, showing that learning occurred $(F(4,360)=13.22, p<$ $.0001, \eta_{p}^{2}=.13$ ), but did not show a statistically significant block:group interaction $\left(F(8,360)=0.40, \eta_{p}^{2}=\right.$ .01).

## Post-hoc Analysis

Additional analysis of the data indicated that participant sex was a behavioral factor. Figure 1 shows mean cumulative scores by group for both males and females. The distribution of mean scores on the task among groups is mirrored between males and females; the mean cumulative score for the positive group among females ( -2.4 ) is closer to the negative group mean cumulative score in male participants ( -3.4 ) than the positive group mean score in male participants (11.6).


Figure 1 Cumulative score for male (left) and female (right) participants at the end of the task. (The error bars represent the standard error means)

A 2 (sex) x 3 (group) ANOVA for cumulative score showed a marginally statistically significant sex:group interaction $\left(F(2,87)=2.77, p=.07, \eta_{p}^{2}=.06\right)$. A $2 \times 3 \times 5$ mixed-model ANOVA for score showed a statistically significant block factor $(F(4,348)=$ $\left.13.55, \eta_{p}^{2}=.13\right)$ and a statistically significant block:sex:group interaction $(F(8,348)=2.12, p=$ $.034, \eta_{p}^{2}=.05$ ) indicating a difference in trends between sex:group pairs.

Thus, we see that the effect of stimuli valence had an effect on the cumulative score on this task, but that the positive and negative valence images appear to have different effects on men and women.

## Using Decision-Learning Models

Though using methods such as those used above are useful for finding differences between groups, simulations of decision-learning models can also be useful as they allow one to explore theoretical aspects of the computation leading to learning and decision-making performance. We ran simulations of decision-learning models to explore how different groups may have evaluated positive and negative payoffs (utility), how they learned these utilities after experiencing them (learning rule), and how these learned expectations may have influenced participants' actual choice (choice probability rule). This resulted in simulation parameter sweeps on 8 total models ( 2 per category); each model was run 100 times during the parameter sweep using the MindModeling@Home cognitive research system.

Functions Used to Construct the Models The two utility functions used were the expectancy utility function (hereinafter referred to as EU ) and the prospect utility function (hereinafter referred to as PU). EU contains a parameter meant to specify a model's attention to loss ( $w$ in eq. 1). Instead of assuming a subjective utility that is linearly proportional to the payoff amount, PU uses a non-linear function (e.g., Tversky \& Kahneman, 1992) to better account for the gain-loss frequency effect (whereby multiple small losses have a larger effect on choice behavior than a single loss equal to the sum of the smaller losses. In (eq. 2) net ( $t$ ) represents the net amount gained (or lost) on trial $t, w$ is a loss-aversion parameter, and $\gamma$ governs the shape of the equation.

$$
\begin{align*}
& u(t)=(1-w) * \operatorname{rew}(t)-w * \operatorname{loss}(t)  \tag{1}\\
& u(t)= \begin{cases}\operatorname{net}(t)^{\gamma} & \forall \operatorname{net}(t) \geq 0 \\
-w *|\operatorname{net}(t)|^{\gamma} & \forall \operatorname{net}(t) \geq 0\end{cases} \tag{2}
\end{align*}
$$

For learning, the Rescorla-Wagner, or delta, rule (Rescorla \& Wagner, 1972) and the decay reinforcement rule (Erev \& Roth, 1998) were used in separate decision models. In the Rescorla-Wagner rule (eq. 3) $\alpha$ represents a learning rate that determines the effect of the the prediction error (the utility minus the current expectancy). The same parameter provides a slightly different representation for the decay reinforcement rule (shown in e. 4). Here, the rule represents a recency parameter, which determines discount of past expectancy when updating with the new utility. Both the delta and decay rules are represented in Table 2 as Del and Dec, respectively.

$$
\begin{align*}
& E_{i}(t)=E_{i}(t-1)+\alpha *\left(u(t)-E_{i}(t-1)\right)  \tag{3}\\
& E_{i}(t)=\alpha * E_{i}(t-1)+u(t) \tag{4}
\end{align*}
$$

Finally, every model had one of two choice rules: tri-al-dependent and trial-independent. These rules define a parameter that affects the probability of selecting a card from each deck $\theta$ in equation 5 . In this case, $\theta$ affects the propensity to explore or exploit the problem space. When the parameter is low, the model is more likely to explore and select a random deck, and when it is higher it will exploit its knowledge and be more likely to select the decks that have a higher utility. The trial-dependent rule (eq. 6) is affected by the number of trials which the model has completed and the consistency parameter c, while the trial independent rule (eq. 7) is only affected by the parameter c (and thus static during the task).

$$
\begin{align*}
& P\left(D_{i}(t+1)\right)=e^{\theta(t) * E_{i}(t)} / \sum_{j=1}^{4} e^{\theta(t) * E_{j}(t)}  \tag{5}\\
& \theta(t)=(t / 10)^{c}  \tag{6}\\
& \theta(t)=3^{c}-1 \tag{7}
\end{align*}
$$

Model Results As one may predict from the human results reported above, the models that best matched human behavior differed between sex-group pairs. To find the best matching models we calculated the $R^{2}$ for each model-parameter-set combination using the proportions of cards selected from each deck during that particular block (i.e., four proportions adding to 1.0 in each of the five blocks). This measure was chosen because it allowed us to further specify how different processes (i.e., models) may explain not only the overall performance (i.e., score), but the proportions of cards selected from decks in each block that define the overall performance. Table 2 lists the top model (and related parameters) for each sex-group combination.

Table 2. Models and corresponding parameters that best matched each sex:group pair. Dec = Decay; Del = Delta; TI = Trial Independent; TD = Trial Dependent. All $\mathrm{R}^{2}(19) \mathrm{p}<.01$

| Sex:Group | Model | $\mathbf{c}$ | $\boldsymbol{w}$ | $\boldsymbol{\gamma}$ | $\boldsymbol{\alpha}$ | $\mathbf{R}^{\mathbf{2}}$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Male:Neg | PU-Dec-TI | -9.25 | 5.2 | .35 | .15 | .87 |
| Male:Neu | PU-Dec-TD | -8.25 | 4.1 | .20 | .93 | .89 |
| Male:Pos | PU-Del-TI | -1.50 | .13 | .00 | .75 | .88 |
| Female:Neg | PU-Del-TI | -5.50 | 2.5 | .65 | .43 | .87 |
| Female:Neu | PU-De1-TI | -6.75 | 6.8 | .15 | .30 | .92 |
| Female:Pos | PU-Dec-TI | -2.50 | 0.5 | .25 | .25 | .89 |

While there were varying parameters for all models a variant of the prospect utility (PU) model showed the greatest match to all of the sex:group pairs. The same model had the highest $\mathrm{R}^{2}$ for the negative scoring per-
formance group for each sex (females in the positive group and males in the negative group). The male:neutral group was the lone group pair to show the highest R2 with a trial dependent model.

## Discussion and Conclusion

These data indicate that the subliminal emotional stimuli had an effect on decision-making. There appears to be an important interaction between sex and emotional decision-making. Even though the stimuli were presented subliminally and were non-integral to the choices made, participants exposed to affectively charged stimuli responded differently to the outcomes of deck selections and performed better or worse on the task, depending on sex and the valence of the stimuli.

We did not find statistically significant evidence for a between group (negative, neutral, or positive) difference in IGT scores. However, we did find that there were significant differences between groups for IGT scores when factoring in participants' sex. What's more, mean scores among males showed a trend opposite of females across groups. These results seem to indicate that the stimuli had opposite effects on males and females.

This may be due to these stimuli affecting portions of the affect-memory coupling portion of the decisionmaking process that can go unnoticed without conscious reflection by the decision-maker. This seems likely given the mirrored distributions, but with similar performance between men and women in the neutral group. Indeed, the simulation model results showed that males in the negative group and females in the positive group were similar to the same class of model.

Similar to the results from a previous study by Aïte et al. (2013), the image-deck congruency also affected the participant's decision-making behavior, albeit differently in men and women. Though females exhibited a pattern similar to Aïte et al. (2013) with the cumulative score for the negative image group being the highest and the cumulative score for the positive group being the lowest, males exhibited the opposite behavior and the image effect was intensified. Indeed, a more recent review points to a difference between men and women in decision-making behavior during the IGT (van den Bos et al., 2013). In the study presented here, women perhaps showed more of an ability to integrate the additive negative signals from the stimuli to make more advantageous decisions than the men; this explanation, would also apply to men, making them more resilient to the negative effects of the positive stimuli on taskperformance. The difference in this affective signal integration may be partially due to the differences in amygdala activity found in men and women (e.g., Cahill, 2006; Hamann et al., 2004). These differences may have also led to a difference in memory processes
predominantly used to make decisions, as the differences in models (particularly learning processes) may suggest. A decay-based learning rule would better lend itself to a more hippocampal/declarative memory, timedependent (e.g., Anderson et al., 1999) emphasized decision-making process.

While this study yields interesting and worthwhile results, there were limitations in the study that restricted the scope of analysis and discussion. Our study is somewhat limited in that we were unable to compliment the results with neuroimaging data (e.g., fMRI). Neuroimaging data could allow more comment on the neural process mediated reasons that we found a difference in decision-making performance between groups that was dependent on participant sex.

Furthermore, the model analysis could be expanded in the future. Indeed, it may also be interesting to integrate an affective component into the simulations to more directly account for the stimuli. This would allow a finer analysis of the computational processes at work, albeit with a more complex model.

The aim of this study was to better understand how non-integral, subliminal stimuli may affect decisionmaking behavior and physiological responses during decision-making. Though we found some expected image-deck congruency effects, these were not as prevalent as originally hypothesized and participant sex also played a role in how decision behavior was unconsciously moderated by the stimuli. More study is necessary to better understand how these unconsciously perceived stimuli are affecting the process of decisionmaking.

Nonetheless, this work provides evidence that nonintegral subliminal stimuli may affect decision-making behavior at several points in the process depending on stimuli characteristics relative to the decision-maker, and reward and punishment contingencies present in the series of decisions. The work also provides evidence that methods of affective intervention during decisionmaking (e.g., presentation of an emotionally charged image to an individual as a part of a decision to purchase an item) should take into consideration the potential effects of the stimulus on males and females. The stimulus will likely have dissimilar effects and may have completely contrasting effects on individual choices based upon the sex of the decision-maker; this could lead to unintended behavioral consequences.

## Acknowledgements

This work was partially funded through support from the Bunton-Waller Fellowship (CLD), the Alfred P. Sloan Foundation Graduate Scholarship (CLD), and the Pennsylvania Space Grant Consortium Graduate Research Fellowship (CLD). The authors thank Jack

Harris at AFRL (711 HPW) for his advice with regards to using the MindModeling@Home platform.

## References

Ahn, W., Busemeyer, J. R., Wagenmakers, E., \& Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. Cognitive Science, 32(8), 1376-1402.
Aïte, A., Borst, G., Moutier, S., Varescon, I., Brown, I., Houdé, O., \& Cassotti, M. (2013). Impact of emotional context congruency on decision making under ambiguity. Emotion, 13(2), 177-182.
Anderson, J. R., Fincham, J. M., \& Douglass, S. (1999). Practice and retention: A unifying analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25(5), 1120-1136.
Bechara, A., Damasio, H., \& Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. Cerebral cortex, 10(3), 295.
Bechara, A., Damasio, H., Damasio, A. R., \& Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. Journal of Neuroscience, 19(13), 5473.

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10(4), 433-436.
Cahill, L. (2006). Why sex matters for neuroscience. Nature Reviews Neuroscience, 7(6), 477-484.
Dancy, C. L., \& Ritter, F. E. (2017). IGT-Open: An open-source, computerized version of the Iowa Gambling Task. Behavior Research Methods, 49(3), 972-978.
Erev, I., \& Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. The American Economic Review, 88(4), 848-881.
Hamann, S., Herman, R. A., Nolan, C. L., \& Wallen, K. (2004). Men and women differ in amygdala response to visual sexual stimuli. Nat Neurosci, 7(4), 411-416.
Lang, P. J., Bradley, M. M., \& Cuthbert, B. N. (1997).
International Affective Picture System (IAPS): Technical manual and affective ratings. The Center for Research in Psychophysiology, University of Florida. Gainesville, FL.
Lerner, J. S., Li, Y., Valdesolo, P., \& Kassam, K. S. (2015). Emotion and decision making. Annual Review of Psychology, 66(1), 799-823.
Ming-Zher, P., Swenson, N. C., \& Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Transactions on Biomedical Engineering, 57(5), 1243-1252.
Moscarello, J. M., \& LeDoux, J. E. (2013). The contribution of the amygdala to aversive and
appetitive pavlovian processes. Emotion Review, 5(3), 248-253.
Öhman, A., \& Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. Psychological Review, 108(3), 483522.

Panksepp, J., Fuchs, T., \& Iacobucci, P. (2011). The basic neuroscience of emotional experiences in mammals: The case of subcortical FEAR circuitry and implications for clinical anxiety. Applied Animal Behaviour Science, 129(1), 1-17.
Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. Annual Review of Psychology, 57, 27-53.
Phelps, E. A., \& Sokol-Hessner, P. (2012). Social and emotional factors in decision-making: Appraisal and value. In R. J. Dolan \& T. Sharot (Eds.), Neuroscience of preference and choice: Cognitive and neural mechanisms. Waltham, MA: Academic Press.
Rescorla, R. A., \& Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black \& W. F. Prokasy (Eds.), Classical Conditioning II: Current Research and Theory. New York, NY: Appleton-Century-Crofts.
Ruiz-Padial, E., Vila, J., \& Thayer, J. F. (2011). The effect of conscious and non-conscious presentation of biologically relevant emotion pictures on emotion modulated startle and phasic heart rate. International Journal of Psychophysiology, 79(3), 341-346.
Tamietto, M., \& de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. Nature Reviews Neuroscience, 11(10), 697-709.
Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297-323.
van den Bos, R., Homberg, J., \& De Visser, L. (2013). A critical review of sex differences in decisionmaking tasks: Focus on the Iowa Gambling Task. Behavioural Brain Research, 238(0), 95-108.
Weber, E. U., \& Johnson, E. J. (2008). Mindful judgment and decision making. Annual Review of Psychology, 60(1), 53-85.
Winkielman, P., Berridge, K. C., \& Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. Personality and Social Psychology Bulletin, 31(1), 121-135.

# Novice to Expert continuum may affect System Response Time 

Arindam Das ${ }^{1}$, Olivia Das ${ }^{2}$<br>York University (Alumni) ${ }^{1}$, Toronto, Canada; Ryerson University ${ }^{2}$, Toronto, Canada.


#### Abstract

To familiarize herself with a user-interface of a software system, a user needs practice. With practice, a user's think time gradually decreases-the novice to expert transition. We propose a queueing model that accounts for this transition in analyzing the performance of a distributed software system. We solve the model using deterministic simulation. Our model captures system performance in terms of system response time. We use the model to demonstrate how userswho are at various experience levels in the novice to expert continuum-may affect the system response time.


Keywords: User Think Time; System Response Time; Queueing Network, Novice to Expert Transition.

## Introduction

Nowadays, distributed systems are getting deployed in cloud since cloud computing allows for dynamic scaling of computational resources as required on a pay-per-use basis. This relieves the providers of cloud system services from buying and maintaining data centers thereby reducing the operational cost.

The performance of a cloud system can get affected due to novice to expert transition as the end users of the system learn to use the system through its user interface. When a user learns, gradual decrease in her think time occurs. Once a software system has responded to a request that was submitted by an end user (through a user interface), the user think time (UTT) refers to the number of seconds the user takes to plan before submitting the next request to the system. The request submission is usually accomplished by clicking an icon on a computer screen. This think time is negatively correlated to the user's expertise level-lower expertise level leads to larger think time and higher expertise level leads to smaller think time. The system workload is affected as a user continues to learn using a system. A novice (one having lower expertise level) requires larger UTT and therefore submits less number of requests per unit time (to the system) compared to an expert (one having higher expertise level). Consequently, as UTT of a user gradually decreases with practice (i.e. repeatedly using the system), the number of requests submitted to the system gradually increases, thereby impacting the system response time (SRT).

Performance evaluation of distributed systems has always considered the UTT as a random variable with a constant mean that does not change with practice (Trivedi, 2005). It never considered novice to expert transition during evaluation.

In this work, we focus on cloud-based systems where the number of computer terminals is fixed at any given point of time. Examples abound-ATMs for a banking system, check-in terminals for an airline at an airport, navigation
training simulators for aircraft pilots-that are deployed in cloud. For our modeling purpose, we assume a hypothetical scenario of a tutorial that is run from inside a classroom. The tutorial consists of learning the fixed locations of buttons on a graphical user interface (GUI) of a cloud system. The classroom has a fixed number of computer terminals, one per student, for the tutorial. No sooner a student finishes the requirements of the tutorial, she is replaced by a new one who is assumed to be at the lowest expertise level.

To realize the above tutorial scenario, we choose a closed queuing network as our system performance model. Choice of a closed queuing network helps us in two ways-one, it allows us to conform to the number of terminals being constant at any given point of time during the tutorial. Secondly, it enables us to account for the decreasing UTT of a user that could occur with repeated use of the interface.

The key contribution of our work is as follows: We demonstrate the effect of novice to expert transition on SRT of distributed systems. We accomplish this through a queueing model. The model accounts for novice to expert transition. Through our model, we show how users-who are at various experience levels in using a system-may affect the system response time.

## Novice to Expert Continuum: In the Context of Human Computer Interaction

We briefly explain what a learning curve is in the traditional context of human computer interaction (HCI).

Any new skill takes time to learn. End users take a while to ramp up on a new user interface; software designers take a while to ramp up on a new project. Learning refers to the acquisition of skill in performing a task through repeatedly executing the same task over time. People get faster and make fewer errors with practice-i.e. with repeated task execution.

In the domain of user interface, the core focus is always a human-centered approach to design-be it the design of a smartphone interface or the interface of a desktop screen. By doing so, we concede that the target of our user interfaces is a population of users with a wide span of skills. We must be aware that these skills change over time as a result of learning. If there are multiple users, they may operate at different expertise levels at the same time due to differences in their experience (Ritter, Baxter, Kim, and Srinivasmurthy, 2013). Such variation in user expertise often calls for a planning of computational resources that would ensure usability satisfaction with respect to SRT for users across all expertise levels.

To take learning into account, what we need is a graph that plots UTTs at different practice sessions for a given
task. A graph like this is popularly referred to as the learning curve-the novice to expert continuum. Figure 1 elucidates the three level hypothesis of learning postulated by Fitts (1964), Anderson (1982), and Kim and Ritter (2015). The hypothesis posits that a learning curve is roughly divided into three levels of user expertise. The first level is where a user is a novice trying to acquire the knowledge to execute a trial. The UTT is usually high at this level. The next level is the intermediate level. At this level, the user tries to consolidate the knowledge acquired in the novice stage. The final level is the expert level. At this third level, the user fine tunes the existing knowledge-users still get faster at the trial, although the improvements get diminishingly smaller (Ritter, Baxter, Kim, and Srinivasmurthy, 2013).


Figure 1. Three Level Hypothesis-Change in user think time across three levels of learning. The thick continuous line indicates continuous practice. (Figure adapted from Kim and Ritter (2015)).

In HCI, a learning curve for a user interface task is often obtained through empirical studies. Here, an interface under study is evaluated in a standalone mode of the client device such that the client software does not have to depend on anything other than the device it is hosted on. As a result, the delay between the submission of a user request (in form of a finger-press or mouse-click on the interface) and the corresponding response is assumed 0 . The interface is evaluated through an interactive task. Multiple human subjects are sampled from a population of novice users of the task. The task involves completing a set of trials-let a trial be the submission of a user request to the software system (in the context of this paper). Each of the users is given equal number of practice sessions to perform the task. The time to complete every trial of the task is measured at each practice session. This measurement is taken for every subject over all the practice sessions. The mean time to complete a trial at a given practice session-mean trial completion time-is then obtained by averaging over the trial completion times measured across all the subjects at that session. Since the delay between every user request and
its response is assumed 0 , the mean trial completion time at a given practice session (which normally would have been the sum of mean user think time and mean system response time at the session) reduces to the mean user think time at that session.

## Closed Queueing Network

A queueing network can be thought of as a network of servers with a queue of jobs (e.g. requests submitted due to mouse-click actions) at each server (Trivedi, 2001). We think of a server as being a world-wide-web server (web server), an application server (app server) or a database server (DB server).

In this work, we exploit a type of queueing network known as closed queueing network. An interactive terminal driven system is often modeled as a closed queueing network with a fixed set of terminals assumed to be part of the network. Each terminal models a delay center in the network. A terminal submits a request and waits for the response. It cannot submit another request until the response of the previous request returns. At any given point of time, the number of terminals in the network remains fixed.

A notion of think time exists in a closed queueing network. It is the time at each terminal between receiving the response of a request and sending out the next request. The idea of think time in a closed queueing network makes it a natural candidate for modeling an interactive system where a user ponders between the completion of a service request and the initiation of a new service request-the time to ponder thus being the UTT.

## A hypothetical Scenario to predict the Learning Effect on System Performance

Our hypothetical scenario consists of the followings: a tutorial that enables a student to learn the fixed locations of buttons on a GUI; a location learning task that is repeatedly executed by a student in the tutorial; and a software system that is used to conduct the tutorial. The tutorial involves the learning of button locations on the GUI only. It does not involve learning of any other type of content.

Note that the task of learning stable layouts of GUI items is not uncommon in real life. While such tasks are routinely executed on banking ATMs and interactive kiosks on a daily basis, they are also prevalent in aviation training, command and control scenarios, and process control plants (Waldron et al., 2005).

## Tutorial scenario

The focus of this paper is to demonstrate the effect of a learning curve on the performance of a 3-tier cloud system. To do so we imagine a tutorial scenario. We assume that a student attending the tutorial interacts with the system through a web-based user interface at a dedicated computer terminal.

The aim of the tutorial is to learn the location of buttons on a stable GUI of the system. The tutorial involves a given
number of practice sessions. The practice sessions are assumed to be separated from one another by a constant period of inactivity. Each student is required to complete all the practice sessions.

The tutorial is conducted inside a classroom having a fixed number of computer terminals. We assume that one student uses one terminal only for her practice sessions. The tutorial begins with one student at every terminal.

Different students may complete their practice sessions at different points of time. It is assumed that when a student completes all her practice sessions, she is replaced by a new student joining the tutorial at the lowest expertise level-the first practice session.

## Location learning task

We assume a simple location learning task that a student will repeatedly perform across multiple practice sessions in the aforementioned tutorial scenario. A location learning task is one where a student learns the locations of graphical items present on a user interface, given that the position (i.e. location) of the items on the interface do not change. We adopt such a task from Ehret's empirical study (Ehret, 2002). The interface on which the task is performed is a graphical layout that consists of 12 unlabelled square buttons arranged along the periphery of a circle. The locations of these square buttons are to be learned through practice. We refer to this interface as Unlabelled Interface. The twelve square buttons are mapped to twelve distinct colors. The colors are not visible; they stay hidden. The circle of square buttons surrounds a centrally located rectangular button. While every square button in the periphery acts as a potential target, the rectangular button at the centre of the circle acts as a cue color.

We refer to the task performed in learning the Unlabelled Interface as "Ehret's task". One practice session of Ehret's task consists of twelve trials. Each trial involves first locating and then clicking a square button that corresponds to a color displayed on the rectangular cue button. In a given practice session, the cue color is different for each of the twelve trials-every trial in a practice thus involves finding a target that is different from the rest eleven targets. In a trial, if the cue color is the color that is associated with the clicked square button, the user has found the target-the trial is therefore considered complete; otherwise the trial is to be repeated. For example, in a trial when the color in the rectangular cue button is green, the trial would be deemed complete only if the square button mapped to green is clicked by the user, not otherwise.

When Ehret conducted this task, he considered only the completed trials. We do the same while considering human learning in our model-we assume that every trial ends up finding the desired target. This helps us keep our model simple.

In Ehret's study, several human subjects had performed multiple practice sessions of Ehret's task on a standalone desktop computer with no internet connection. As a result the delay between the submission of a user request (in form
of a mouse-click) and the corresponding response was assumed 0 . We therefore consider a trial completion time in Ehret's study to be essentially a user think time (UTT) for the modeling purposes in this paper.

In our model, we utilize the mean trial completion time corresponding to each practice session of Ehret's task as the mean user think time for that session. We incorporate a 3tier, cloud-based distributed system that is responsible for processing a submitted user request (mouse-click on a square button). We assume that this system processes the mouse-click and returns the response-the cue color for the next trial-after a non-zero delay (the delay being the system response time, SRT).

## Software System

Figure 2 shows the software architecture of our hypothetical cloud-based distributed system that is used for conducting the aforementioned tutorial. We analyze this system in this work.


Figure 2. Our hypothetical 3-tier, cloud-based distributed system that is used to accomplish our tutorial scenario.

The system consists of 3 tiers. One or more Web servers run in the first tier (tier-1), one or more application servers (App servers) run in the second tier (tier-2) and one or more database servers (DB servers) run in the third tier (tier-3). The servers run of virtual machines (VMs). VM is a term used in cloud computing. It refers to a virtual (i.e. emulated) host. Similar to a physical host (computer), a VM can run an operating system as well as other processes. A VM can be acquired, networked, or released on demand through software based mechanisms. We assume that at any given tier, one or more VMs can be provisioned, each running a single instance of a server relevant to that tier.

We assume that the workload is equally distributed among the servers at any given tier. We indicate this in Figure 2 using the phrase "balanced load".

The student users access the application at the web servers through their web-based user interfaces.

## Control flow of a trial

A trial refers to first locating and then clicking a target square button on Ehret's Unlabelled Interface. In a trial, first a user spends time in reasoning and planning where the target square button would be. Then she submits a request to the system by clicking the potential target. Here, a request refers to a job that is generated due to a button-click and that is to be subsequently processed by the software system (Figure 2) starting from tier-1 and until tier-3. We assume that a request will be processed exactly once (in a server) at each tier. After completion of processing at the third tier, a response corresponding to the processed request is returned to the user. We assume that a request incurs a waiting time in the server's queue before being processed, if the server is busy. The request then incurs a service time for getting processed in the server.
The request is first sent to a Web server in tier-1 for processing. If the Web server is busy then the request needs to wait in the server's queue before getting processed.
Once the processing of the request at tier- 1 is finished, the request is redirected to an App server in tier-2. If the App server is busy then the request needs to wait in the server's queue before getting processed.
Once the processing of the request at tier-2 is finished, the request is redirected to a DB server in tier-3. If the DB server is busy then the request needs to wait in the server's queue before getting processed.

Once the processing of the request is finished at the DB server, the response corresponding to the request is sent back to the user. At this point, the trial is complete. We assume that the response returned to the user contains the information about a new cue color whose associated button is to be located (on the interface) in the next trial.

A trial thus incurs two delays. One is the time spent by a user in reasoning and planning where the target square button is located, given a cue color. This delay period is the user think time (UTT). The other is the system delay due to waiting times and service times incurred by the request between the click of a target button on the user interface and the return of the response. This second delay is the system response time (SRT).

Once all the trials of a practice session are complete, there is period of inactivity before the first trial of the next practice session begins-this period of inactivity is the inter practice time.

## A Queueing Model Considering the Effect of Novice to Expert Continuum

A user with lower expertise level requires larger UTT and therefore submits less number of requests per unit time to the system. In contrast, a user with higher expertise level
requires smaller UTT and therefore submits more number of requests to the system. Thus, as UTT of a user gradually decreases with practice, the number of requests submitted to the system gradually increases. The decreasing UTT thus influences the system workload which in turn affects the waiting times of the requests and consequently, the SRT. Keeping this in mind, we model our hypothetical distributed system as a closed queuing network.

The queuing model parameters, their meaning and their values are summarized in Table 1. The parameter values are specified inside bold parenthesis in the right column of the table. Each parameter is explained in due context as our work unfolds. Since the queueing network is a closed one, the total number of terminals (concurrent users) $\boldsymbol{N}$ in the system is constant at any point of time. An individual terminal user initiates a practice session $p$ by first thinking for a certain amount of time with mean $u_{p}$ (mean user think time per trial of practice session $p$ ) and then submitting the first request of that session to the system. After the completion of the request, the user thinks again for a time with mean $u_{p}$ and then submits the subsequent request of the practice session $p$.

Once a user finishes $\boldsymbol{T}$ number of trials needed to complete a practice session, she takes a break for some time with mean $\alpha$ (mean inter practice time). The user then proceeds with the next practice session. The user completes $\boldsymbol{P}$ practice sessions in total before leaving the system. A departing user is replaced by a new novice user who begins her practice at practice session 1 .
We assume that $\mu_{1}$ is the mean service time of each Web server replica at tier-1, $\mu_{2}$ is the the mean service time of each App server replica at tier-2, and $\mu_{3}$ is the mean service time of each DB Server replica at tier-3.

Table 1. Model Parameters

| Parameter | Meaning |
| :---: | :---: |
| P | Total number of practice sessions assumed to be completed by a user before leaving the system $(1 \leq p \leq P)$. ( $P=15$ sessions) |
| $N$ | Number of computer terminals (concurrent users). Once a user completes $\boldsymbol{P}$ practice sessions, she leaves the system and a new novice user occupies the terminal. The new user begins her practice at practice session $1 . N$ thus stays fixed during a simulation run, thereby abiding by the "constant number of customers" requirement for a closed queueing network. $(1 \leq i \leq N)$. (Simulation data collected at $N=120$ simulated concurrent users) |
| $N_{p}$ | Number of concurrent users at practice session $p$ at the start of a simulation run ( $1 \leq p \leq P$ ) where $N_{1}+N_{2}+\ldots+N_{P}=N$. This is applicable when human learning is considered. |
| $T$ | Number of trials to be completed in a practice session. This is assumed to be equal for all practice sessions. ( $\boldsymbol{T}=\mathbf{1 2}$ trials per session) |
| $r_{i, p}$ | Actual number of completed trials in practice session $p$ at terminal $i$ during the simulation. |


| $\alpha$ | Mean inter practice time (1 sec) |
| :--- | :--- |
| $\mu_{1}$ | Mean service time at tier-1 (0.5 sec) |
| $\mu_{2}$ | Mean service time at tier-2 $(\mathbf{0 . 5} \mathbf{~ s e c})$ |
| $\mu_{3}$ | Mean service time at tier-3 (0.5 sec) |
|  | Mean User Think Time per trial of <br> practice session $p \quad\left(u_{1}=\mathbf{1 2 . 5}, u_{2}=\mathbf{1 0 . 6}, u_{3}=\right.$ <br> $u_{p}$ |
| $\mathbf{8 . 9}, u_{4}=\mathbf{6 . 8}, u_{5}=\mathbf{6 . 5}, u_{6}=\mathbf{6 . 1}, u_{7}=\mathbf{5 . 1}, \quad u_{8}$ <br> $=\mathbf{4 . 2}, u_{9}=\mathbf{4 . 3}, u_{10}=\mathbf{4 . 3}, u_{11}=\mathbf{3 . 1}, u_{12}=\mathbf{2 . 7}$, <br> $u_{13}=\mathbf{2 . 9}, u_{14}=\mathbf{2 . 5}, u_{15}=\mathbf{2 . 2}$. The values are in <br> $\mathbf{s e c o n d s . ~ T h e y ~ a r e ~ o b t a i n e d ~ f r o m ~ F i g u r e ~ 3 ) ~}$ |  |

The system response time of a request is the time between the arrival of the request at a tier-1 server to the completion of the request at a tier-3 server. This time includes the waiting times at the queues of the relevant servers at different tiers and the service times of those servers. This implies that the SRT of a request is affected by the rate of request submissions (i.e. the number of requests submitted to the system per unit time) in addition to the service times of the servers.

User Think Time when human learning is not considered: When human learning is not considered, the think times of a user across all practice sessions will be identically distributed random variables with same mean $u_{1}$ $=u_{2} \ldots=u_{P}$.

User Think Time when human learning is considered: When human learning is taken into account, the think times of a user across all practice sessions will be identically distributed random variables with unequal means $u_{1} \neq u_{2} \ldots$ $\neq u_{P}$. Here, we take unequal means instead of purely decreasing means because of the following reason: Although a learning curve obtained through empirical studies show an overall decreasing trend in user think time with practice, sometimes it may exhibit exceptions in form of increased user think times at some practice sessions possibly owing to user fatigue.

We accomplish the analysis of our queuing model through deterministic discrete-event simulation. For simplicity, we assume that the user think times, the service times of the servers, and the inter practice time are deterministically distributed.

Let $s_{i, j, p}$ denote the system response time for a trial $j$ of practice session $p$ by a user at terminal $i$. During every simulation run, we record the response times $s_{i, j, p}$. Let $r_{i, p}$ denote the number of completed trials of practice session $p$ at terminal $i$.

The Mean System Response Time (Mean SRT) per trial $\overline{S_{p}}$ of practice session $p$, where $p=1,2, \ldots, P$ can be estimated as:

$$
\overline{s_{p}}=\frac{\sum_{i=1}^{N} \sum_{j=1}^{r_{i, p}} s_{i, j, p}}{\sum_{i=1}^{N} r_{i, p}}
$$

The numerator of the above equation denotes the total system response time of all the trials of practice session $p$
completed from all the terminals. The denominator represents the number of those trials.

## Model Results

We use the human learning curve observed by Ehret (2002) as an input to our model. This empirical curve of human learning was measured when human subjects executed Ehret's task-the task to learn the locations of square buttons on an Unlabelled Interface explained earlier. Figure 3 shows the learning curve. The curve is in terms of the mean user think times across the first 15 practice sessions completed by the sixteen human subjects. Subjects' point of regard was measured as they performed the task. The eye data was collected via an ASL 5000 eye-tracker.

Our simulated users are assumed to execute the aforementioned Ehret's task. The simulation emulates the hypothetical tutorial scenario explained earlier.

The simulated tutorial is assumed to start with a fixed number of computer terminals-one student using one terminal only. Once a user completes all the 15 practice sessions she leaves the system. A departed user is then replaced by a new novice user who begins her practice at practice session 1.


Figure 3. Human learning curve for Ehret's task (Ehret, 2002).

We show how users-who are at various experience levels in using a system-may affect the system response time. To do so, we run our simulation model with an initial proportion of users who are at various expertise levels.

We refer to the learning curve of Ehret's task (Figure 3). Let $\left[N_{1} / N_{6} / N_{11}\right]$ denote the initial proportion of users where $N_{1}$ users begin their practice at session $1, N_{6}$ users begin their practice at session 6 , and $N_{11}$ users begin their practice at session 11 at the start of a simulation run. We assume that there are no users at other expertise levels at the start of the simulation run, i.e. $N_{1}+N_{6}+N_{11}=N$. Here, we choose the practice sessions 1,6 and 11 assuming that novice-level experience starts at session 1, intermediatelevel experience starts at session 6 and expert-level experience starts at session 11 . Our choice is dictated by the
three level hypothesis of learning (discussed earlier; see Figure 1) applied on the learning curve of Figure 3.

We consider 120 concurrent users in the system, i.e. $N_{1}+$ $N_{6}+N_{11}$ is 120 . We perform one logical-hour analysis for a VM configuration where every tier of our three-tiered system has 6 VM replicas.

Table 2 shows mean SRTs $\overline{s_{\text {novice }}}, \overline{S_{\text {intermedıate }}}$ and $\overline{S_{\text {expert }}}$ for two initial proportions of users [120/0/0] and [40/40/40]. Here $\overline{s_{\text {novice }}}$ refers to $\overline{s_{1}}$, the Mean SRT at practice session $1 ; \overline{s_{\text {intermediate }}}$ refers to $\overline{s_{7}}$, the Mean SRT at practice session 7 ; and $\overline{S_{\text {expert }}}$ refers to $\overline{s_{15}}$, the Mean SRT at practice session 15 .

With respect to analyzing the initial user proportion [120/0/0] for one logical-hour, the reason for low mean SRT per novice trial $(2.25 \mathrm{sec})$ but high mean SRT per expert trial ( 7.22 sec ) is as follows: In this case, the system is transiting from all-novices to all-experts. The user think times (UTTs) at the novice level are substantially higher than those at the expert level. Therefore when all the users are at novice level, the rate of request submissions (to the system) is lower compared to all-experts. This leads to less waiting times during novice request executions and higher waiting times for expert request executions.

Table 2. Mean SRTs $\overline{\boldsymbol{s}_{\text {novıce }}}$ (i.e. $\overline{\boldsymbol{s}_{1}}$ ), $\overline{\boldsymbol{s}_{\text {intermedıate }}}$ (i.e. $\overline{\boldsymbol{s}_{7}}$ ) and $\overline{\boldsymbol{s}_{\text {expert }}}\left(\right.$ i.e. $\overline{\boldsymbol{s}_{15}}$ ) for different initial proportions of users. One logical-hour analysis. VM configuration consists of 6 VM replicas per tier. $N=120$ users.

| Initial User <br> Proportion <br> $\left[\mathbf{N}_{\mathbf{1}} / \mathbf{N}_{\mathbf{6}} / \mathbf{N}_{\mathbf{1 1}}\right]$ | $\overline{\boldsymbol{S}_{\text {novice }}}$ <br> (sec) | $\overline{S_{\text {Intermedıate }}}$ <br> (sec) | $\overline{\boldsymbol{s}_{\text {expert }}}$ <br> $(\mathrm{sec})$ |
| :---: | :---: | :---: | :---: |
| $[120 / 0 / 0]$ | 2.25 | 5.12 | 7.22 |
| $[40 / 40 / 40]$ | 4.36 | 4.5 | 5.03 |

On the contrary, the mean SRT of 5.03 sec per expert trial is less in case of [40/40/40] in comparison to 7.22 sec for the proportion [120/0/0]. This is found from a one logicalhour analysis. The reason is as follows: In case of [120/0/0], all the users start at the novice level. They then transition to the intermediate level almost at the similar time. Finally, all the users transition from the intermediate to the expert level-again, almost at the similar time. Once at the expert level, the rate of request submissions by a user is higher in comparison to her rate of submissions either at the intermediate or at the novice level. On top of that, since almost all the users have transitioned to the expert level, an expert trial has no choice but to compete for resources against majority of the trials which are also occurring at the expert level. In contrast, for the case [40/40/40], an expert trial competes for resources against a mixture of novice, intermediate and expert trials. Consequently, the mean SRT for an expert trial is higher in case of [120/0/0] than the proportion [40/40/40].

Let an example SRT requirement be as follows: "The mean SRT should be less than or equal to 5.5 sec ". Table 2 suggests that the VM configuration ( 6 VM replicas per tier) will satisfy the threshold of 5.5 sec for only the proportion
[40/40/40] in one logical-hour analysis since all of $\overline{s_{\text {novice }}}$, $\overline{S_{\text {Intermedıate }}}$ and $\overline{S_{\text {expert }}}$ are below the threshold for that proportion. The other proportion [120/0/0] will not satisfy the threshold since $\overline{S_{\text {expert }}}=7.22 \mathrm{sec}$ being more than 5.5 sec , the proportion will not be able to meet the SRT threshold for the expert trials.

This analysis of the effect of user proportion on SRT can be used by the system analysts when the workload trend for the system under analysis is known. An example workload trend could be a plot of initial user proportions of 120 concurrent users of the system against different times of the day obtained from the historical data of the system's usage. Suppose the plot shows an initial user proportion [120/0/0] at 8 am and $[40 / 40 / 40]$ at 3 pm . Our aforementioned one logical-hour analysis predicts that from 8am to 9am, the configuration of 6 VM replicas per tier would not satisfy the threshold SRT of 5.5 sec . But, the same configuration would ensure usability satisfaction (with respect to SRT) across all expertise levels from 3 pm to 4 pm .

## Conclusions

We propose a queueing model that accounts for novice to expert transition in analyzing the performance of a distributed software system. Our model captures system performance in terms of system response time. We use the model to demonstrate how users-who are at various experience levels in the novice to expert continuum-may affect the system response time.

## Acknowledgment

We would like to thank NSERC Canada for their financial support through the Discovery Grant of Olivia Das.

## References

Anderson J. R. (1982). "Acquisition of cognitive skill," Psychological review, vol. 89, no. 4, pp. 369-406.
Ehret, B. D. (2002). "Learning where to look: Location learning in graphical user interfaces," CHI, pp. 211-218.
Fitts, P. M. (1964). "Perceptual-motor skill learning," in Categories of human learning, New York, NY: Academic Press, pp. 243-285.
Kim, J. W., \& Ritter, F. E. (2015). "Learning, Forgetting, and Relearning for Keystroke-and Mouse-Driven Tasks: Relearning is Important," Human-Computer Interaction, vol. 30, pp. 1-33.
Ritter, F. E., Baxter G., Kim, J. W., \& Srinivasmurthy, S. (2013). "Learning and retention," in John D. Lee and Alex Kirlik (eds.). The Oxford Handbook of Cognitive Engineering, New York, NY: Oxford, pp. 125-142.
Trivedi, K. S. (2001). Probability and Statistics with Reliability, Queuing, and Computer Science Applications, 2nd ed. John Wiley and Sons.
Waldron, S.M., Duggan, G.B., Patrick, J., Banbury, S., \& Howes, A. (2005). "Adaptive information fusion for situation awareness in the cockpit," in Proc. 49th Annual Meeting of the Human Factors and Ergonomics Society, pp. 49-53.

# Modeling Comprehension Processes via Automated Analyses of Dialogism 

Mihai Dascalu (mihai.dascalu@cs.pub.ro)<br>Department of Computer Science, University Politehnica of Bucharest<br>Bucharest, 060042 Romania<br>Laura K. Allen (laurakallen@asu.edu)<br>Danielle S. McNamara (dsmcnama@asu.edu)<br>Department of Psychology, Arizona State University<br>Tempe, AZ, 85287-2111 USA<br>Stefan Trausan-Matu (stefan.trausan@cs.pub.ro)<br>Department of Computer Science, University Politehnica of Bucharest Bucharest, 060042 Romania

Scott A. Crossley (scrossley@gsu.edu)<br>Department of Applied Linguistics/ESL, Georgia State University<br>Atlanta, GA, 30303 USA


#### Abstract

Dialogism provides the grounds for building a comprehensive model of discourse and it is focused on the multiplicity of perspectives (i.e., voices). Dialogism can be present in any type of text, while voices become themes or recurrent topics emerging from the discourse. In this study, we examine the extent that differences between self-explanations and thinkalouds can be detected using computational textual indices derived from dialogism. Students $(n=68)$ read a text about natural selection and were instructed to generate selfexplanations or think-alouds. The linguistic features of these text responses were analyzed using ReaderBench, an automated text analysis tool. A discriminant function analysis using these features correctly classified $80.9 \%$ of the students' assigned experimental conditions (self-explanation vs. think aloud). Our results indicate that self-explanation promotes text processing that focuses on connected ideas, rather than separate voices or points of view covering multiple topics.


Keywords: comprehension; discourse analysis; dialogism; polyphonic model; self-explanation; think-aloud

## Introduction

Research on text comprehension suggests that skilled and less skilled readers differ in the frequency and type of strategies they employ while processing texts (Millis, Magliano, \& Todaro, 2006; Oakhill \& Yuill, 1996). Skilled readers, for instance, generate more inferences while reading, which allows them to establish connections between information in the text and their prior knowledge (Kintsch, 1998). Although not all readers naturally make these connections while reading, students can be prompted to generate inferences through instructions to self-explain (McNamara, 2004). Self-explanation is a response to text or discourse that is directed toward oneself, with an explicit purpose to construct meaning from the text. Explanations are statements generated aloud, through text, or silently to oneself, that go beyond the information provided explicitly
in the text to explain the ideas, their relations, and their underlying meaning.

In the context of text comprehension, self-explanation can improve readers' understanding of complex topics (McNamara, 2004, in press). From the point of view of theories within the field of text and discourse comprehension, the benefits of self-explanation have been attributed to increased bridging and elaborative inferences (i.e., making connections to prior ideas in the text or to prior knowledge) and to increased causal inferences (e.g., making connections to causal events; (Allen, McNamara, \& McCrudden, 2015)

Instructions to self-explain can be contrasted with those to think-aloud, which ask readers to report whatever thoughts are available to them while reading a text (e.g., readers report these thoughts after reading each sentence). Asking a reader to think-aloud reveals their use of comprehension strategies, including any inferences they generate while reading, but does not alter a reader's natural comprehension processes (McNamara \& Magliano, 2009).

In this study, we compare the comprehension processes associated with self-explanation and think-aloud from the lens of dialogism. Dialogism refers to a wider perspective of dialogue which is assumed to be present within any verbal or non-verbal language activity. Dialogism originates from the Russian philosopher and philologist Mikhail Bakhtin ( 1981,1984 ) who proposed that there is an implicit and multi-voiced dialogue underlying sense-making, communication, actions, and interactions (Linell, 2009). Accordingly, voices represent distinct positions, points of view, or ideas, that impact the nature and outcome of a discourse. Multi-voicedness, may drive to polyphony, which is a central concept within dialogism, and a focus of the polyphonic model, which is essential to this study (TrausanMatu \& Rebedea, 2009).

Dialogue traditionally refers to communication between two or more individuals. Indeed, within the context of

Computer Supported Collaborative Learning, dialogism has been considered better suited as a theoretical framework for multi-party conversations than classic Natural Language Processing theories that focus on phone-like interactions between two interlocutors (Trausan-Matu \& Rebedea, 2009).

Bakhtin, however, stressed the point that all text is multivocal, wherein our speech (all of our utterances) is filled with others' words (Bakhtin, 1986, p. 89). In this view, the concept of dialogue can be extended to include a wider variety of language activities. For instance, it may also refer to an internal dialogue within oneself or a dialogue amongst inner voices contrasting and debating ideas (Marková, Linell, Grossen, \& Salazar Orvig, 2007, ch. 6).

The polyphonic model is a generalization of Bakhtin's ideas in the sense that voices may not be only associated to an individual person. Voices may also be themes, or recurrent topics emerging from the discourse. They enter in inter-animation patterns which generate a polyphonic weaving characterized by a multitude of voices, each with its individuality, but which give birth to a coherent whole (Trausan-Matu \& Rebedea, 2009).

In this study, we apply the polyphonic model to analyze the presence of voices and their interactions within students' think-alouds and self-explanations. Think-aloud, by definition, is the externalization of the inner voice of the student, including voices that correspond to ideas, justifications, and assumed positions. Expressions of the text comprehension process by a student, in the form of think aloud or self-explanation, can be expected to include positions, reasons, ideas, all which may be viewed as voices.

We have operationalized the polyphonic model within the ReaderBench framework, which provides automated text and conversational analysis (Dascalu, 2014; Dascalu, Dessus, Bianco, Trausan-Matu, \& Nardy, 2014; Dascalu, Trausan-Matu, McNamara, \& Dessus, 2015). In ReaderBench, voices are identified using Natural Language Processing (NLP). Lexical chains (Galley \& McKeown, 2003), or sequences of repeated or related words, are merged into semantic chains by using relatedness calculated using Latent Semantic Analysis (Landauer \& Dumais, 1997) and Latent Dirichlet Allocation (Blei, Ng, \& Jordan, 2003). The previous semantic models play an important role in our polyphonic model as they are used to identify voices through semantic relatedness, thus highlighting cohesive contexts.

In addition, polyphonic inter-animation considers relations between voices, or points of view, along two dimensions: longitudinal along time, and transversal across time, using voices' co-occurrences within and across text segments (Trausan-Matu, Stahl, \& Sarmiento, 2007). The longitudinal dimension follows the continuation of ideas throughout the discourse, similar to a voice's individual melodic line in music. Simultaneously, voices co-occur in a vertical manner and, as in polyphonic music, this generates specific discourse contexts consisting of potential dissonances that need to be solved toward consonances.

This transversal effect, or voice overlap, supports the integration process that can create both unity across various themes, as well as differences or variations in points of view.

Specifically, we examine the extent to which differences between participants' expressions of self-explanation and think-aloud can be detected using the computational text analyses provided by ReaderBench, and in turn, how this theoretical perspective informs our understanding of text and discourse comprehension.

## Discourse Analysis within the Polyphonic Model

The polyphonic model can be used to analyze discourse in both conversations and plain texts (Trausan-Matu \& Rebedea, 2009). Bakhtin (1984) stated that polyphony occurs in any text, similarly to polyphonic music, composed under counterpoint rules. That means that there is a multitude of voices, each with its own individuality, whose sum comprises a coherent whole: "the voices of others become woven into what we say, write, and think" (Koschmann, 1999, p. 308). Meanwhile, the polyphonic discourse should also bring novelty, voices should interanimate in order to foster creativity. Following this perspective, the polyphonic approach to discourse analysis identifies voices in text and then investigates how voices are woven and how they inter-animate (Trausan-Matu, Stahl, et al., 2007).

Our automated process of voice identification starts by building lexical chains that are merged into semantic chains through semantic relatedness (Dascalu et al., 2015). Lexical chains can be identified using a disambiguation graph in which nodes are word instances having assigned their most probable sense, while weighted edges are semantic distances from WordNet (Galley \& McKeown, 2003). However, this approach is inherently limited because it only includes words from the same part of speech. Thus, we have used an iterative agglomerative hierarchical clustering algorithm that begins with the identified lexical chains as groups of clustered words and uses the semantic similarity between lexical chains as a distance function (Dascalu et al., 2015). If the semantic relatedness value is greater than an imposed threshold or if identical lemmas are identified in two word clusters, the latter are automatically merged.

Voices emerge as central topics of each text and rely on the occurrences of the underlying cohesive and semantically related words. The longitudinal dimension of voices becomes the context in which the voices span throughout the entire discourse. In contrast, the transversal dimension highlights different co-occurrence and inter-animation patterns of voices present within the same textual element, i.e., sentence or paragraph.

After voices are identified, a cohesion (or utterance) graph is constructed from the links between utterances (Dascalu, 2014; Dascalu, Dessus, Trausan-Matu, Bianco, \& Nardy, 2013; Trausan-Matu, Dascalu, \& Dessus, 2012; Trausan-Matu, Rebedea, Dragan, \& Alexandru, 2007) Within the cohesion graph, utterances are the nodes and
links consist of adjacency pairs, repetitions, or lexical and semantic chains, which are detected using NLP.

As such, voices can be identified as threads in the graphs (Trausan-Matu, Dascalu, \& Rebedea, 2014). Each utterance has an inner voice that inter-twines with other voices from the same thread or from different ones, but with less strength. Any new utterance in a dialogue is expressed as a voice, including its degree of interconnection with other utterances, relevance within the discourse, and potential impact within the overall discussion. Examining different semantic chains within the same textual fragment (sentences or paragraphs) reveals the transversal dimension of voice inter-animation.

## Current Study

This study comprises an analysis of a corpus of selfexplanations and think-alouds previously described in Allen et al. (2015). University students $(n=68)$ read a text about natural selection and were randomly assigned to one of two conditions related to their reading instructions: selfexplanation $(n=33)$ and think-aloud ( $n=35$ ). Students in the self-explanation and think-aloud conditions were prompted to generate typed responses on 16 occasions (i.e., on 16 of the 41 sentences). The self-explanation instructions asked students to explain the information they had just read to themselves, whereas the think-aloud instructions asked students to state whatever they were thinking. We aggregated students' 16 text responses (their self-
explanations or think-alouds) following the procedure described in Allen et al. (2015).

The individual files were then analyzed using ReaderBench. We calculated 29 voice indices related to:
a) span (distances between word occurrences within the same semantic chain),
b) recurrence (average and standard deviation in terms of distance between two consecutive words pertaining to the same voice, measured in number of in-between words from the initial text),
c) coverage of these semantic chains (average number of contained concepts per sentence or paragraph), and
d) information theory entropy (Shannon, 1948) based on the probability that a voice appears in a given text segment.

The previous indices relate to the longitudinal dimension of our analysis, while voice inter-animation relates to the transversal effect, which is computed in terms of cooccurrence patterns. As operationalization of the transversal dimension, we rely on: a) counting the number of concepts pertaining to different voices, but present in the same text segment, and b) pointwise mutual information (PMI) that measures the degree of association between voice distributions (Dascalu et al., 2015). These dialogic indices provide insights in terms of a text's overall cohesion, as voices help build a higher cohesion through lexical and semantic relatedness.

| Self-explanation (SE) | Think-aloud (TA) |
| :---: | :---: |
| In our lives, there are so many types of people around us to our lives colorful. also, in our daily lives, we meet different people who have different story to tell. some of them are happy, wealthy. Some of them have to worry about how to survive in this society. they are components to make our lives fascinates. <br> Life around us is fascinating because of the force of nature. those creatures around us are differently designed. Some of them are capable of seeing stuffs because they are given an unique thing-eye. that's one of the things to make them special, to make their lives fancy. <br> Life fascinates us because we have eyes. And eyes have precise arrangement so that eyes make us see things. This is also true for other organs, they are complexly design to make our body function. <br> Organs are not designed in advance for a specific purpose. organs are formed by the activities people do in their everyday-lives. organs are formed for the destination to make people survive, to make people's body function well. the two animals with cloudy lenses must give their next generation cloudy lens. and the generation of cloudy lens animal and clear lens animal will be hard to tell. because the offspring are given clear lens due to those who gave birth to it. Because a replicator is something that can make a copy of itself, with most of its traits duplicated in the copy, including the ability to replicate. The offspring's parents survived and pass this replicator to it, so that the offspring's eyes are the same as its parent's. | The surroundings we cannot to change, but we can our heart to adopt. <br> In my mind, the human also as one of the animals in the world, we have only different from the other animals because we have a thought. <br> The eyes is difference with the other organs. The animals eyes may be less important than other organs. <br> The author cannot to believe that the organs must have been designed in advance for a specific purpose is right. Used an example to support his ideal. According to the example, I feel that the offspring has clear lenses and can see well which is incorrect. in some way, their eyes has different with their parents' eyes. <br> That's mean the their another eyes is usedness. the better vision can help these animals to reproduce and get better generation. It's to point out the ideal which is the author want to explain. <br> The living surrounding makes natural selection in order to get better next generation. that's mean we can change or direct the selection to product. replicators try to use-up material to find their the great copies and energy to power replication. the most of the copying is worse that causing less efficient just the less of the copying is better and useful, back to the viewpoint. the apparent well engineered body is result by the replicator to make the natural selection. <br> Organisms is the standard by the natural selection. |

Figure 1. Sample inter-animation of voices within a self-explanation and a think-aloud protocol.

Table 1: MANOVA results.

| Index | Self-explanations <br> $\mathrm{M}(\mathrm{SD})$ | Think-alouds <br> $\mathrm{M}(\mathrm{SD})$ | $F$ | $p$Partial Eta <br> Squared |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Average span of lexical chains <br> Average paragraph voice co- <br> occurrence <br> Average sentence voice co-occurrence | $\mathbf{3 . 0 8}(0.81)$ | $1.84(0.94)$ | 33.212 | $<.001$ | .407 |
| Standard deviation of paragraph voice | $\mathbf{2 . 9 3}(1.08)$ | $1.93(0.97)$ | 15.984 | $<.001$ | .335 |
| co-occurrences | $\mathbf{6 . 0 0}(2.22)$ | $2.99(1.51)$ | 43.296 | $<.001$ | .195 |
| Average sentence entropy of voices <br> Average span of voices | $\mathbf{1 . 3 2}(0.23)$ | $1.14(0.25)$ | 10.202 | $<.01$ | .134 |
| Standard deviation of paragraph voice <br> mutual information (PMI) | $\mathbf{6 . 3 7}(1.64)$ | $5.25(1.41)$ | 9.184 | $<.01$ | .122 |
| Percentage of words that are included in <br> lexical chains | $\mathbf{0 . 5 8}(0.07)$ | $\mathbf{0 . 6 5}(0.13)$ | 5.755 | $<.05$ | .080 |
| Standard deviation of distributions per <br> paragraph | $\mathbf{0 . 7 5}(0.04)$ | $0.09(0.03)$ | 5.291 | $<.05$ | .074 |
| Standard deviation of sentence voice co- <br> occurrences | $\mathbf{1 . 5 1}(0.35)$ | $0.72(0.06)$ | 4.994 | $<.05$ | .070 |

Figure 1 presents an example of inter-animation of voices within a self-explanation and a think-aloud verbal protocol. Several threads that can be considered as voices, ranging from simple word repetitions (i.e., "organs") to semantically related concepts (i.e., "eye - lens", "generation - parent offspring", "copy - replication - duplicate - replication"), co-appear and inter-animate. Additional voices can be identified in both texts, but it is important to observe differences in terms of distributions: denser, more cluttered and more elaborated voices are present in self-explanations versus more varied and more spread-out voices in thinkalouds.

In addition, the texts exhibit different discourse structures: longer, more elaborated and more cohesive paragraphs in self-explanations versus shorter, more condensed phrases introducing multiple ideas in think-alouds. These latter discourse specific traits also directly influence the distribution of voices within the underlying analysis element (paragraph or sentence) as the chance of voice cooccurrence inherently increases in longer texts (e.g., selfexplanations).

## Results

Statistical analyses were conducted to assess the extent to which the dialogic indices related to voices and in turn, accurately classified students based on their experimental condition. Because ReaderBench reports raw voice counts, indices were also checked for multicollinearity with text length. Any index that was highly collinear ( $r>.90$ ) with text length was removed. We then conducted a MANOVA to identify which indices exhibited significant differences between the self-explanation and think-aloud conditions. Indices that were highly collinear ( $r>.90$ ) were flagged, and the index with the strongest effect size in the MANOVA was retained while the other indices were removed (see Table 1).

Longer spans of both lexical chains and voices, as well as higher paragraph and sentence voice co-occurrences, are indicative of longer, more elaborated texts (i.e. selfexplanations). Self-explanations also have higher standard deviations of co-occurrence patterns at both sentence and paragraph levels which reflect a greater variety of voices, as well as a more diverse and unequal overlap of voices. Higher voice entropy at sentence level and higher standard deviation of voice distributions at paragraph level also support the latter finding. Moreover, self-explanations have a slightly higher coverage of words that are integrated in longer semantic chains, thus denoting a more connected discourse. In contrast, think-alouds exhibited a higher standard deviation of paragraph voice pointwise mutual information. This was specific to the generation of new ideas which may or may not be intertwined with other voices. This result is indicative of a wider spread of synergic effects - either new, more isolated voices, or ones that interanimate more with the previous voices.

Based on the MANOVA, we selected the four indices with the strongest effect sizes to enter into a stepwise discriminant function analysis (DFA): Average span of lexical chains $(S E>T A)$, Average paragraph voice cooccurrence $(S E>T A)$, Average sentence voice cooccurrence ( $S E>T A$ ), and Standard deviation of paragraph voice co-occurrences ( $S E>T A$ ).

Table 2: DFA results.

|  | Type | Predicted Group <br> Membership |  | Total |
| :--- | :--- | :---: | :---: | :---: |
|  |  | SE | TA |  |
| Original | SE | 25 | 8 | 33 |
|  | TA | 5 | 30 | 35 |
| Cross- | SE | 25 | 8 | 33 |
| validated | TA | 6 | 29 | 35 |

Note: $\mathrm{SE}=$ self-explanation; $\mathrm{TA}=$ think-aloud

The DFA yielded a significant model, $\chi 2(d f=1$, $n=68)=34.243, p<.001$, correctly allocating $55(25+30)$ of the 68 students (accuracy $=80.9 \%$, see Table 2). To test the stability of our model, we conducted a leave-one-out cross-validation analysis, which also yielded an accuracy of $79.4 \%$. The measure of agreement between the actual instructional group and that assigned by our model produced a weighted Cohen's Kappa of .616, demonstrating substantial agreement.

## Discussion and Conclusions

In the current study, we examined the differences from a dialogism perspective between self-explanations and thinkalouds generated in response to a text. In previous research on this dataset (Allen et al, 2015), we examined the causal and referential cohesion differences between selfexplanation and think-aloud. The results of the latter study indicated that causal cohesion, but not referential cohesion differentiated students who were in the self-explanation and think-aloud conditions. In the current study, we build on this prior research by examining how textual indices related to dialogism relate to students' processing of text based on their text reading instructions.

Our results indicate that students who self-explained the text generated longer voices (lexical or semantic chains with a higher span) that inter-animate more (higher voice cooccurrences at sentence and paragraph levels). This suggests that students who were prompted to self-explain responded to the text by maintaining semantic connections of the concepts within the text.

We interpret these results to indicate that self-explanation promotes specific comprehension processes that are fundamentally different from responses generated during think-aloud protocols, evidenced by students' generation of more conceptually related and cohesive text responses, rather than multiple, separate "voices" or points of view covering multiple topics. Previous research indicates that self-explanation can enhance students' understanding of complex concepts; however, it is less clear how these instructional differences manifest in the linguistic properties of students' text responses.

By adopting a Natural Language Processing approach, this study examines on-line comprehension processes at a more fine-grained level and also contributes to a better understanding of how these processes may be automatically detected via computational linguistic analyses. The polyphonic model, built on dialogism and integrating advanced NLP techniques, represented a viable alternative to analyze students' discourse and differentiate among instructional settings.

As a limitation of our approach, voices need to account for more than word clustering based on distance, lexical and semantic overlaps which are currently used to operationalize our polyphonic model. In addition, many voice indices are multicollinear with text length. We need to develop methods to normalize raw voice score to help control for text length constraints. Voices represent a generalization of emerging
topics and should consider the corresponding sentiment valences in order to create a clearer perspective whether convergence or dissonances are encountered in the discourse. In order to address this issue, opinion mining techniques will be integrated in a follow-up iteration of our implemented model.

The dialogical framework offers new perspectives in the context of this study because both self-explanation and think-aloud be perceived as different kinds of dialogues. Self-explanations on the one hand include positions, reasons, ideas, all of which may be viewed as voices while the overall discourse can be regarded as an 'internal dialogue within the self' (Linell, 2009, ch. 6). On the other hand, think-alouds are more condensed, centered on generating new ideas and can also be perceived a 'dialogue between ideas' (Marková et al., 2007, ch. 6), a dialogue in which the debating voices are ideas. However, in both cases reflexive and cognitive processes are needed in order for students to express themselves.

As Linell (2009) stated, dialogues occur 'in and through words.' Certainly, there is more to dialogue and communication - for example, gestures, facial expressions, emotions, movement, all play crucial roles in dialogue; but here, because it is printed text, we can only examine words. Nonetheless, dialogism is tightly connected to the notion of sense-making as meaning is constructed by interacting with others and with the world, as well as with oneself via internal dialogue. As such, dialogism has a strong connection to cognition, which is sometimes ignored. Figure 2 represents this viewpoint. Cognition reflects prior and intrapersonal (individual) information and knowledge about the world, while meaning is constructed through social interactions and language within the dialogical context. Communication, both explicit interpersonal dialogue and implicit to oneself (i.e., internal dialogue), becomes a facilitator in terms of interaction, therein generating meaning in discourse.


Figure 2. Dialogical framing and interdependencies with core concepts.

## Acknowledgments

The work presented in this paper was partially funded by the University Politehnica of Bucharest through the "Excellence Research Grants" Program, UPB - GEX Contract number 12/26.09.2016, by the EC H2020 project RAGE (Realising
and Applied Gaming Eco-System) http://www.rageproject .eu/ Grant agreement No 644187, as well as funding to Arizona State University (IES 305A130124, IES R305A120707, NSF 1417997, NSF 1418378, ONR 12249156, ONR N00014140343). Any opinions or conclusions expressed are those of the authors and do not represent views of the IES, NSF, or ONR.

## References

Allen, L. K., McNamara, D. S., \& McCrudden, M. T. (2015). Change your mind: Investigating the effects of self-explanation in the resolution of misconceptions. Paper presented at the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015), Pasadena, CA.
Bakhtin, M. M. (1981). The dialogic imagination: Four essays (C. Emerson \& M. Holquist, Trans.). Austin and London: The University of Texas Press.
Bakhtin, M. M. (1984). Problems of Dostoevsky's poetics (C. Emerson, Trans. C. Emerson Ed.). Minneapolis: University of Minnesota Press.
Bakhtin, M. M. (1986). Speech genres and other late essays (V. W. McGee, Trans.). Austin: University of Texas.

Blei, D. M., Ng, A. Y., \& Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4-5), 993-1022.
Dascalu, M. (2014). Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence (Vol. 534). Cham, Switzerland: Springer.
Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., \& Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), Educational Data Mining: Applications and Trends (pp. 335-377). Cham, Switzerland: Springer.
Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., \& Nardy, A. (2013). ReaderBench, an environment for analyzing text complexity and reading strategies. Paper presented at the 16th Int. Conf. on Artificial Intelligence in Education (AIED 2013), Memphis, USA.
Dascalu, M., Trausan-Matu, S., McNamara, D. S., \& Dessus, P. (2015). ReaderBench - Automated Evaluation of Collaboration based on Cohesion and Dialogism. International Journal of Computer-Supported Collaborative Learning, 10(4), 395-423. doi:10.1007/s11412-015-9226-y
Galley, M., \& McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. Paper presented at the 18th International Joint Conference on Artificial Intelligence (IJCAI'03), Acapulco, Mexico.
Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge, UK: Cambridge University Press.
Koschmann, T. (1999). Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration. Paper presented at the Int. Conf. on Computer Support for Collaborative Learning (CSCL'99), Palo Alto.

Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104(2), 211-240.
Linell, P. (2009). Rethinking language, mind, and world dialogically: Interactional and contextual theories of human sense-making. Information Age Publishing: Charlotte, NC.
Marková, I., Linell, P., Grossen, M., \& Salazar Orvig, A. (2007). Dialogue in focus groups: Exploring socially shared knowledge. London, UK: Equinox.
McNamara, D. S. (2004). SERT: Self-Explanation Reading Training. Discourse processes, 38, 1-30.
McNamara, D. S. (in press). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. Discourse Processes.
McNamara, D. S., \& Magliano, J. P. (2009). Selfexplanation and metacognition: The dynamics of reading. In J. D. Hacker, J. Dunlosky, \& A. C. Graesser (Eds.), Handbook of Metacognition in Education (pp. 60-81). Mahwah, NJ: Erlbaum.
Millis, K., Magliano, J. P., \& Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and Latent Semantic Analysis. Scientific Studies of Reading, 10(3), 225-240.
Oakhill, J., \& Yuill, N. (1996). Higher order factors in comprehension disability: Processes and remediation. In C. Cornaldi \& J. Oakhill (Eds.), (pp. 69-72). Mahwah, NJ: Erlbaum.
Shannon, C. E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal, 27, 379-423 \& 623-656.
Trausan-Matu, S., Dascalu, M., \& Dessus, P. (2012). Textual complexity and discourse structure in ComputerSupported Collaborative Learning. Paper presented at the 11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012), Chania, Grece.

Trausan-Matu, S., Dascalu, M., \& Rebedea, T. (2014). PolyCAFe-automatic support for the polyphonic analysis of CSCL chats. International Journal of ComputerSupported Collaborative Learning, 9(2), 127-156. doi:10.1007/s11412-014-9190-y
Trausan-Matu, S., \& Rebedea, T. (2009). Polyphonic interanimation of voices in VMT. In G. Stahl (Ed.), Studying Virtual Math Teams (pp. 451-473). Boston, MA: Springer.
Trausan-Matu, S., Rebedea, T., Dragan, A., \& Alexandru, C. (2007). Visualisation of learners' contributions in chat conversations. In J. Fong \& F. L. Wang (Eds.), Blended learning (pp. 217-226). Singapour: Pearson/Prentice Hall.
Trausan-Matu, S., Stahl, G., \& Sarmiento, J. (2007). Supporting polyphonic collaborative learning. Indiana University Press, E-service Journal, 6(1), 58-74.

# Amortized Hypothesis Generation 

Ishita Dasgupta ${ }^{1}$, Eric Schulz ${ }^{2}$, Noah D. Goodman ${ }^{3}$ \& Samuel J. Gershman ${ }^{4}$<br>${ }^{1}$ Department of Physics and Center for Brain Science, Harvard University<br>${ }^{2}$ Department of Experimental Psychology, University College London<br>${ }^{3}$ Department of Psychology, Stanford University<br>${ }^{4}$ Department of Psychology and Center for Brain Science, Harvard University


#### Abstract

Bayesian models of cognition posit that people compute probability distributions over hypotheses, possibly by constructing a sample-based approximation. Since people encounter many closely related distributions, a computationally efficient strategy is to selectively reuse computations - either the samples themselves or some summary statistic. We refer to these reuse strategies as amortized inference. In two experiments, we present evidence consistent with amortization. When sequentially answering two related queries about natural scenes, we show that answers to the second query vary systematically depending on the structure of the first query. Using a cognitive load manipulation, we find evidence that people cache summary statistics rather than raw sample sets. These results enrich our notions of how the brain approximates probabilistic inference.


Keywords: Amortization; hypothesis generation; Bayesian inference; Monte Carlo methods

## Introduction

Many theories of probabilistic reasoning assume that people are equipped with a general-purpose inference engine that can be used to answer arbitrary queries for a wide variety of probabilistic models (Griffiths et al., 2012). The flexibility and power of a general-purpose inference engine trades off against its computational efficiency: by avoiding any assumptions about the query distribution, the inference engine relinquishes the opportunity to reuse computations across queries. Conversely, an inference engine may gain efficiency by incurring some amount of bias due to reuse-a strategy we refer to as amortized inference (Stuhlmüller et al., 2013; Gershman \& Goodman, 2014). We propose that people employ some form of this strategy, flexibly reusing past inferences in order to efficiently answer new but related queries.

The experiments reported in this paper explore amortization in sets of related queries that involve probabilistic inference over a very large space of possibilities. These possibilities are not all explicitly provided and have to be generated by the participant in order to carry out the inference. We frame amortization as the reuse of hypotheses that have already been generated in response to previous queries. We model the process of hypothesis generation with a stochastic sampling mechanism (Lieder et al., 2012; Dasgupta et al., 2016). One way to implement amortization in this framework is to directly reuse samples across different queries. Alternatively, amortization could be implemented by reusing some summary statistic compiled from previous samples. One goal of our experiments is to tease apart these different mechanistic assumptions. The basic logic of our experiments is to hold
one query fixed while manipulating an earlier query, allowing us to interrogate reuse of computations across queries.

## Stochastic hypothesis generation

For complex hypothesis spaces, exact probabilistic inference is typically intractable. Several lines of evidence point to the idea that humans approximate inference by constructing a Monte Carlo approximation of the posterior distribution (see Griffiths et al., 2012; Sanborn \& Chater, 2016, for a review). This "sampling hypothesis" can be realized algorithmically in a number of ways. Recent studies have shown how a number of apparent probabilistic fallacies can be understood as a consequence of resource-bounded sampling using a Markov chain Monte Carlo (MCMC) algorithm (Lieder et al., 2012; Dasgupta et al., 2016). Because we build on those studies in this paper, we briefly describe the theoretical framework.

A Monte Carlo approximation uses a set of $N$ samples $\left\{h_{1}, \ldots, h_{N}\right\}$, drawn from a hypothesis space $\mathcal{H}$, to approximate a target distribution. In our case, the target is the posterior distribution over hypotheses given data $d, P(h \mid d) \propto$ $P(d \mid h) P(h)$. The Monte Carlo approximation is defined by:

$$
\begin{equation*}
P(h \mid d) \approx \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left[h_{n}=h\right], \tag{1}
\end{equation*}
$$

where $\mathbb{I}[\cdot]=1$ if its argument is true ( 0 otherwise). MCMC algorithms generate samples by simulating a Markov chain whose stationary distribution is the posterior (MacKay, 2003). This approach is asymptotically exact (the approximation converges to the posterior as the number of samples approaches infinity) but under time or resource constraints only a small number of samples may be generated. Although this gives rise to systematic deviations from exact inference, it may in fact be the computationally rational sampling policy (Lieder et al., 2012; Vul et al., 2014; Gershman et al., 2015).

In our prior research (Dasgupta et al., 2016), we applied this model to a scene inference domain, using a database of natural object co-occurrence statistics compiled by Greene (2013). The task facing subjects in our experiments was to judge the probability of a particular set of latent objects in a scene conditional on observing another object (the cue). By manipulating the framing of the query, we showed that subjects gave different answers to formally equivalent queries. Specifically, by partially unpacking the queried object set (where fully unpacking an object set means to present it explicitly as a union of each of its member objects) into a small set of exemplars and a 'catch-all' hypothesis (e.g., "what is the probability that there is a book, a box, or any other object
beginning with B ?"), we found that subjects judged the probability to be higher when the unpacked exemplars were typical (a "subadditivity" effect; cf. Tversky \& Koehler, 1994) and lower when the unpacked exemplars were atypical (a "superadditivity" effect; cf. Sloman et al., 2004) compared to when the query is presented without any unpacking. To give a concrete example, in the presence of say a 'table', the typically unpacked query "what is the probability that there is also a chair, a curtain, or any other object beginning with C?" generates higher probability estimates, and the atypically unpacked query "what is the probability that there is also a cow, a canoe, or any other object beginning with C?" generates lower probability estimates, when compared to the packed query "what is the probability that there is also another object beginning with C ?".

These effects could be accounted for by the MCMC model under the assumption that the unpacked exemplar(s) initialize the Markov chain(s) that form the sample set. Because the initialization of the Markov chain transiently determines its future trajectory, initializing with typical examples causes the chain to tarry in the high probability region of the queried object set, thereby increasing its judged probability (subadditivity). However, initializing with atypical examples causes the chain to get easily derailed into regions outside the queried object set and thus generate more hypotheses that are not in the queried object set. This decreases the judged probability of the queried object set (superadditivity). The strength of these effects diminishes with the number of samples, as the chain approaches its stationary distribution (which is the same for all conditions). Accordingly, response time and cognitive resource availability modulate the effect size (Dasgupta et al., 2016).

## Amortized inference in hypothesis generation

We will consider two simple amortization schemes within the MCMC framework described above. ${ }^{1}$ In sample reuse, people may simply add samples generated in response to one query ( $Q 1$ ) to the sample set for another query ( $Q 2$ ). So if $N_{1}$ samples were generated in response to $Q 1$, and $N_{2}$ new samples are generated in response to $Q 2$, the response to $Q 2$ is given by a calculation carried out over all $N_{1}+N_{2}$ equally weighted samples. Under this scheme, all the computations carried out for $Q 1$ are available for flexible reuse in the computation for $Q 2$. In summary reuse, people may reuse a summary statistic computed from $Q 1$. This strategy is only applicable to problems where the answer to $Q 2$ can be expressed as the composition of the answer to $Q 1$, and an additional simpler computation. To make this viable in our experiments, $Q 2$ queries a hypothesis space that is the union between the hypothesis space queried in $Q 1$ and a disjoint hypothesis space. For example if $Q 1$ is "What is the probability that there is an object starting with a C in the scene?",

[^312]Q2 could be "What is the probability that there is an object starting with a C or an R in the scene?". In this case, the $N_{1}$ samples generated in response to $Q 1$ are summarized into one probability ("the probability of an object starting with C"), $N_{2}$ new samples are generated in response to a simpler query ("the probability of an object starting with R"), and these two numbers are then composed (in this case simply added) to give the final estimate for $Q 2$ ("the probability of an object starting with $C$ or R"). Under this scheme, only the final product of the computation carried out for $Q 1$ is reused in the calculations for $Q 2$. These two models are the two extremes between very flexible and very rigid reuse; intermediates are of course possible.


Figure 1: Simulation of subadditivity and superadditivity effects under sample-based (top) and summary-based (bottom) amortization schemes. In all panels, the y-axis represents the effect size for $Q 2$. Left panels show the effects of changing the sample size for $Q 1$; right panels show the effects of changing the sample size for $Q 2$. When, sample size for $Q 2$ is changed, sample size for $Q 1$ is held fixed at 230 , and vice versa.

Sample-based and summary-based amortization schemes make different predictions about how subadditivity and superadditivity change as a function of the sample size (Figure 1). Increasing the sample size for $Q 1$ amplifies the effects for $Q 2$ under a sample-based scheme, because this leads to more $Q 1$ samples being reused for $Q 2$. This effect can be counteracted by increasing the sample size for $Q 2$, which pushes the effects down (the effects go to 0 as the sample size for Q2 tends to infinity, since the Markov chain will converge to the same posterior for all conditions). Under a summarybased scheme, increasing the sample size for $Q 1$ will actually diminish the effects for $Q 2$, because the bias from $Q 1$ is strongest when the chain is close to its starting point. In other words, the subadditivity and superadditivity effects for Q2 derive from the same effects in $Q 1$, which themselves are primarily driven by the initialization (see Dasgupta et al., 2016). In Experiment 2, we test these different predictions by placing people under cognitive load during either $Q 1$ or

Table 1: Experimental stimuli and queries.

| Cue | $Q 1$ | Q1-Typical | Q1-Atypical | $Q 2$ |
| :--- | :---: | :--- | :--- | :---: |
| Table | C | chair, <br> computer <br> curtain | cannon, <br> cow, <br> canoe | C or R |
| Tele- <br> phone | D | display case, <br> dresser, <br> desk | drinking fountain, <br> dryer, <br> dome | D or L |
| Rug | B | book, <br> bouquet, <br> bed | bird, <br> buffalo, <br> bicycle | B or S |
| Chair | P | painting, <br> plant, <br> printer | porch, <br> pie, <br> platform | P or A |
| Sink | T | table, <br> towel, <br> toilet | trumpet, <br> toll gate, <br> trunk | T or E |
| Lamp | W | window, <br> wardrobe, <br> wine rack | wheelbarrow, <br> water fountain, <br> windmill | W or F |

Q2, a manipulation that is thought to reduce the number of samples (Dasgupta et al., 2016; Thaker et al., 2017).

## Experiment 1

Our first experiment pursued a basic carryover effect from one query $(Q 1)$ to the next ( $Q 2$ ). We assessed two putative signatures of sampling-subadditivity and superadditivityfor a fixed $Q 2$ while changing the structure of $Q 1$. Specifically, we compared three conditions that differed only in how $Q 1$ was framed: packed, unpacked-typical, and unpackedatypical. In order to encourage amortization, we specified $Q 2$ as a union of the hypothesis space queried by $Q 1$ and another hypothesis space-i.e., a disjunctive query. The design is summarized in Table 1.

## Participants

84 participants ( 53 males, mean age $=32.61, \mathrm{SD}=8.79$ ) were recruited via Amazon's Mechanical Turk and received \$0.5 for their participation plus an additional bonus of $\$ 0.1$ for every on-time response.

## Procedure

Participants were asked to imagine playing a game in which their friend sees a photo and then mentions one particular object present in the photo (the cued object). The participant is then queried about the probability that another class of objects (e.g., "objects beginning with the letter B") is also present in the photo.

Each participant completed 6 trials, where the stimuli on every trial corresponded to the rows in Table 1. On each trial, participants first answered $Q 1$ given the cued object, using a slider bar to report the conditional probability (Figure 2). The Q1 framing (typical-unpacked, atypical-unpacked or packed) was chosen randomly. Participants then completed the same


Figure 2: Experimental setup. Participants were asked to estimate the conditional probability using a slider bar within a 20 -second time limit.
procedure for $Q 2$, conditional on the same cued object. The framing for $Q 2$ was always packed.

## Results

Consistent with our prior studies (Dasgupta et al., 2016), we find both subadditivity and superadditivity effects for $Q 1$, depending on the unpacking (Figure 3): probability judgments were higher for unpacked-typical queries than for packed queries (a subadditivity effect; $t(77)=4.029, p<0.001$ ) and lower for unpacked-atypical than for packed queries (a superadditivity effect; $t(77)=-6.4419, p<0.001$ )


Figure 3: Experiment 1 results. Mean probability estimates by condition. Error bars represent the standard error of the mean.

Crucially, we also saw effects of $Q 1$ unpacking on response to $Q 2$, even though these queries were all presented as packed hypotheses. In particular, estimates for $Q 2$ were lower when $Q 1$ was unpacked to atypical exemplars $(t(77)=$ $-5.0789, p<0.001)$, demonstrating a superadditivity effect
that carried over from one query to another. We did not find an analogous carry-over effect for subadditivity $(t(77)=$ $0.72, p>0.4$ ), possibly due to the subadditivity effect "washing out" more quickly (i.e. with fewer samples) than superadditivity, at least within this domain (see Dasgupta et al., 2016).

Next, we explored whether responses to $Q 1$ predicted trial-by-trial variation in responses to Q2. As shown in Figure 4, we found significant positive correlations between the two queries in all conditions when aggregating across participants ( $p<0.01$ ). The same conclusion can be drawn from analyzing correlations within participants and then testing the average correlation against 0 (average correlation: $r=0.55$, $p<0.01$ ). Moreover, the within-participant effect size (the response difference between the unpacked conditions and the packed conditions) for $Q 1$ was correlated with responses to $Q 2$ for both atypical ( $r=0.35, p<0.01$ ) and typical ( $r=0.21, p<0.05$ ) unpacking conditions. This means that even though the subadditive condition did not significantly differ from the unpacked condition for $Q 2$ overall, participants who showed greater subadditivity or superadditivity for Q1 also showed correspondingly greater effects for $Q 2$.


Figure 4: Experiment 1 trial-by-trial analyses: Relationship between aggregated $Q 1$ and $Q 2$ responses. Lines show the leastsquares fit with standard error bands.

## Experiment 2

Experiment 1 showed strong evidence for reuse of inferential computations across queries when the evidence is fixed. Two questions naturally arise from this finding. First, how adaptive is amortization? Are samples reused promiscuously across queries (potentially leading to rampant memory-based biases), or is reuse sensitive to conditions where it is likely to be accurate? This is a delicate point, since it is impossible to know with certainty whether amortization is useful without knowing some properties of the problem (e.g., decomposability of the conditional distribution). Nonetheless, humans
may be able to utilize heuristics for constructing amortization strategies whose errors can be corrected by additional experience or computation (Stuhlmüller et al., 2013). We address this question by manipulating the "amortizability" of $Q 1$, in order to test the hypothesis that carry-over effects across queries will only occur under high amortizability conditions. We operationalize amortizability in terms of whether or not the hypothesis space queried by $Q 1$ is a subset of the hypothesis space queried by $Q 2$.

The second question concerns resource allocation. Theories of computational rationality argue that computations are selected to balance accuracy and cost (Lieder et al., 2012; Gershman et al., 2015). In the context of sampling, this means that fewer samples will be generated when cognitive resources are scarce. This hypothesis is consistent with the observation that subadditivity (Sprenger et al., 2011) and order effects (Thaker et al., 2017) are amplified under cognitive load. We pursue this question by manipulating cognitive load at both $Q 1$ and $Q_{2}$. As discussed in the Introduction, the different amortization schemes make different predictions for these manipulations (see Figure 1).

## Participants

80 participants ( 53 males, mean age $=32.96, \mathrm{SD}=11.56$ ) were recruited from Amazon Mechanical Turk and received \$0.5 as a basic participation fee and an additional bonus of $\$ 0.1$ for every on time response as well as $\$ 0.1$ for every correctly classified digit during cognitive load trials.

## Procedure

The procedure in Experiment 2 was largely the same as in Experiment 1, with the main difference being that participants had to remember a sequence of digits. On half of the trials the cognitive load manipulation occurred at $Q 1$ and on half at $Q 2$. The digit sequence was presented prior to the query, and then following their response to the query they were asked to make a same/different judgment about a probe sequence. Half of the probes were old and half were new.

To probe adaptive amortization, we added several $Q 2$ queries to the list shown in Table 1. These queries were deemed less amortizable because they lack any of the letters queried in Q1 (for example, ' T or R ' instead of ' C or R ' in the trial shown in the first row in Table1). In other words, these queries could not be decomposed and hence could not benefit from reuse. Half of the Q2 trials were randomly chosen to provide hypotheses with low amortizability.

## Results

As shown in Figure 5, we replicated and extended the results from Experiment 1, showing both subadditivity and superadditivity effects for $Q 1$ that carried over to $Q 2$. Analyzing only amortizable queries (averaging across load conditions), we found that probability judgments for Q1 were higher for unpacked-typical compared to packed (subadditivity; $t(79)=4.38, p<0.001)$ and lower for unpackedatypical compared to packed (superadditivity $t(79)=-4.94$,
$p<0.001$ ). These same effects occurred for $Q 2$ (unpackedtypical vs. packed: $t(79)=2.44, p<0.01$; unpackedatypical vs. packed: $t(79)=-1.93, p<0.05)$.


Figure 5: Experiment 2 results. Mean probability estimates for Q2 by condition. Error bars represent the standard error of the mean.

As in Experiment 1, there was strong correlation between responses to $Q 1$ and $Q 2$ overall conditions ( $r=0.44, p<$ 0.001 ), for the packed ( $r=0.44, p<0.001$ ), the typically unpacked ( $r=0.45, p<0.001$ ), as well as the atypically unpacked condition ( $r=0.35, p<0.01$ ); see Figure 6. Moreover, $Q 1$ and $Q 2$ were also highly correlated within participants (mean $r=0.31, p<0.01$ ) and participants who showed higher subadditivity or superadditivty effects for $Q 1$ also showed higher effects for $Q 2$ overall ( $r=0.31, p<0.001$ ), for the superadditive ( $r=0.3, p<0.001$ ), and for the subadditive condition ( $r=0.29, p<0.001$ ). This replicates the effects of amortization found in Experiment 1.


Figure 6: Experiment 2 trial-by-trial analyses: Relationship between aggregated $Q 1$ and $Q 2$ responses. Lines show the leastsquares fit with standard error bands.

Finally, we assessed how the carryover effects were mod-


Figure 7: Experiment 2: differences between responses for each condition and an average packed baseline. Bars represent a mean within-participant absolute effect. Error bars represent the standard error of the mean.
ulated by cognitive load and amortizability. To highlight the effects more clearly, we calculated each participant's mean response to all packed hypotheses for $Q 2$ over all trials as a baseline measure. We then calculated the difference between each condition's mean response and this mean packed response. This provides us with a measure of an average effect size within $Q 2$-responses (how much each participant underor overestimates different hypotheses as compared to an average packed hypothesis). Results are shown in Figure 7.

If cognitive load occurred during $Q 2$ and amortizability was low, none of the conditions produced an effect significantly different from 0 (all $p>0.5$ ). If cognitive load occurred during $Q 2$ and amortizability was high, only typically unpacked hypotheses produced an effect significantly higher than $0(t(38)=2.14, p<0.05)$. If cognitive load occurred during $Q 1$ and amortizability was low, again none of the conditions significantly differed from 0 (all $p>0.05$ ). Crucially, if cognitive load occurred during $Q 1$ and amortizability was high, both conditions showed the expected subadditive $(t(38)=4.18, p<0.05)$ and superadditive $(t(46)=$ $-1.89, p<0.05)$ effects. Moreover, calculating the average effect size for the different quadrants of Figure 7, the high amortizability-cognitive load at $Q 1$-condition produced the highest overall effect ( $d=0.8$ ), followed by the high amortizability-cognitive load at $Q 2$-condition $(d=0.56)$ and the low amortizability-cognitive load at $Q 1$-condition ( $d=0.42$ ). The low amortizability-cognitive load at $Q 2$ condition did not produce an effect higher than 0 . Moreover, highly amortizable trials were more strongly correlated with responses during $Q 1$ than trials with low amortizability ( 0.15 vs $0.41, t(157)=-2.28, p<0.05)$.

Intriguingly, on trials with cognitive load at $Q 2$, participants were on average more likely to answer the probe correctly for high amortizability trials compared to low amortizability trials $(t(36)=3.16, p<0.05)$. This is another signature of amortization: participants are expected to have more resources to spare for the memory task at $Q 2$ if the computations they did for $Q 1$ are re-usable in answering $Q 2$. This also indicates that these results cannot be explained by simply initializing the chain for $Q 2$ where the chain for $Q 1$ ended, which would not have affected computation time. Our results suggest that the transfer actually makes the second computation easier by re-using previous computations.

In summary, Experiment 2 replicates the effects found in Experiment 1 and the increased effect for the high amortizability condition provides further evidence that this effect is actually driven by amortization. These experiments also give us some insight into how amortization is implemented. Based on our simulations (Figure 1), we argue that the effect of cognitive load at $Q 1$ on $Q 2$ responses is more consistent with summary amortization than with sample amortization. These results suggest an active process of $Q 2$ being expressed in terms of the results to $Q 1$, when possible. This approach is more structured and less flexible than sample amortization but trades in this inference limitation for an increase in memory-efficiency and is thus consistent with beliefs about cost-efficient resource-rational inference strategies in humans.

## Discussion

In two experiments, we provided empirical support for amortized hypothesis generation. Participants not only exhibited subadditive and superadditive probability judgments in the same paradigm (replicating Dasgupta et al., 2016), but also carried over these effects to subsequent queries. Importantly, Experiment 2 demonstrated that such carry-over effects only occur when amortization can exploit shared structure across queries. Experiment 2 also demonstrated that cognitive load exerts its strongest effect when applied to the first query, suggesting (based on our simulations) that the carry-over effects are driven by some kind of summary-based amortization, whereby a summary statistic is computed from the samples and then reused to answer subsequent queries that can be expressed in terms of already completed calculations. This implies a structured amortization strategy, over one that reuses all old samples, and thus gives up some flexibility for memory-efficiency. Building on earlier results (Gershman \& Goodman, 2014), our results support the existence of a sophisticated inference engine that adaptively exploits past computations. While reuse can introduce error, this error may be a natural consequence of a resource-bounded system that optimally balances accuracy and efficiency (Lieder et al., 2012; Vul et al., 2014; Griffiths et al., 2015; Gershman et al., 2015). The incorporation of reuse into a Monte Carlo sampling framework allows the inference engine to preserve asymptotic exactness while improving efficiency in the finite-
sample regime
Future studies could use similar methods to study amortization in other domains, such as in concept learning (Goodman et al., 2008) or reinforcement learning tasks (Daw et al., 2011). There is also a much larger space of more sophisticated amortization schemes (e.g., Stuhlmüller et al., 2013; Rezende et al., 2014; Paige \& Wood, 2016) that we have not yet tried to test. Pinning down the computational details of amortization will be an important task for future work.

## References

Dasgupta, I., Schulz, E., \& Gershman, S. J. (2016). Where do hypotheses come from? CBMM Memo 56 .
Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., \& Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron, 69(6), 1204-1215.
Gershman, S. J., \& Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. In Proceedings of the 36th Annual Conference of the Cognitive Science Society, (pp. 517-522).
Gershman, S. J., Horvitz, E. J., \& Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349(6245), 273-278.
Goodman, N. D., Tenenbaum, J. B., Feldman, J., \& Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. Cognitive Science, 32(1), 108-154.
Greene, M. R. (2013). Statistics of high-level scene context. Frontiers in Psychology, 4(1), 777.
Griffiths, T. L., Lieder, F., \& Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Topics in Cognitive Science, 7, 217-229.
Griffiths, T. L., Vul, E., \& Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. Current Directions in Psychological Science, 21(4), 263-268.
Lieder, F., Griffiths, T. L., \& Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. In Advances in Neural Information Processing Systems, (pp. 2690-2798).
MacKay, D. J. (2003). Information theory, inference and learning algorithms.
Paige, B., \& Wood, F. (2016). Inference networks for sequential monte carlo in graphical models. In Proceedings of the 33rd International Conference on Machine Learning, vol. 48.
Rezende, D. J., Mohamed, S., \& Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of The 31st International Conference on Machine Learning, (pp. 1278-1286).
Sanborn, A. N., \& Chater, N. (2016). Bayesian brains without probabilities. Trends in Cognitive Sciences, 20(12), 883-893.
Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., \& Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(3), 573-582.
Sprenger, A. M., Dougherty, M., Atkins, S. M., Franco-Watkins, A. M., Thomas, R., Lange, N., \& Abbs, B. (2011). Implications of cognitive load for hypothesis generation and probability judgment. Frontiers in Psychology, 2, 129.
Stuhlmüller, A., Taylor, J., \& Goodman, N. D. (2013). Learning stochastic inverses. In Advances in Neural Information Processing Systems, (pp. 3048-3056).
Thaker, P., Tenenbaum, J. B., \& Gershman, S. J. (2017). Online learning of symbolic concepts. Journal of Mathematical Psychology.
Tversky, A., \& Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. Psychological Review, 101, 547-567.
Vul, E., Goodman, N., Griffiths, T. L., \& Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. Cognitive Science, 38(4), 599-637.

# The Causal Sampler: A Sampling Approach to Causal Representation, Reasoning and Learning 

Zachary J. Davis (zach.davis@nyu.edu)<br>Bob Rehder (bob.rehder@nyu.edu)<br>Department of Psychology, New York University<br>6 Washington Place, New York, NY, USA 10003 USA


#### Abstract

Although the causal graphical model framework has achieved success accounting for numerous causal-based judgments, a key property of these models, the Markov condition, is consistently violated (Rehder, 2014; Rehder \& Davis, 2016). A new process model-the causal sampler-accounts for these effects in a psychologically plausible manner by assuming that people construct their causal representations using the Metropolis-Hastings sampling algorithm constrained to only a small number of samples (e.g., $<20$ ). Because it assumes that Markov violations are built into people's causal representations, the causal sampler accounts for the fact that those violations manifest themselves in multiple tasks (both causal reasoning and learning). This prediction was corroborated by a new experiment that directly measured people's causal representations.


Keywords: causal learning, causal reasoning, sampling

## Introduction

The representation and use of causal knowledge is a central object of investigation in the cognitive sciences. Causal models have been found to affect cognition in a wide variety of inference problems, from reasoning and learning to decisionmaking and categorization (for a summary, see Rottman \& Hastie, 2014; Waldmann \& Hagmayer, 2013). One formal model of the representation of causal information - causal graphical models - has achieved success in modeling behavior across these tasks.

A foundational feature of causal graphical models is the Markov condition, which stipulates that the value of a node is independent of its non-descendants, conditional on its parents. This principle is crucial for statistical inference from causal graphical models (Pearl, 1988; Koller \& Friedman, 2009), and has been argued to be necessary for a rigorous account of interventions (Hausman \& Woodward, 1999).

Given the success of the causal graphical model formalism, one might expect to find the Markov condition satisfied in human behavior. In contrast, the causal inferences that people draw consistently violate the independence relationships implied by the Markov condition (Rehder, 2014; Rehder \& Burnett, 2005; Rehder \& Waldmann, 2016).

One explanation for Markov violations is that they represent a flaw in people's causal reasoning process. On this account, Markov violations would not necessarily manifest themselves on other causal-based tasks (e.g., causal learning). Rehder and Davis (2016) investigated this possibility by testing whether people honor the Markov condition during a causal hypothesis testing task. In fact, the hypothesis they favored reflected the same independence violations that characterize causal reasoning (details below).

Together, these findings pose a problem for current theories of causal cognition. We propose that the generality of these errors suggests that a reorientation is needed in our understanding of how people represent causal relationships. To this end, we propose a process model that conceives of causal cognition as based on simulation, rather than analytic calculation. The model outperforms traditional Bayes nets across tasks, and we test its predictions in a novel task.

## Process Model

Building on recent work in cognitive science that investigates the role of sampling methods in accounting for judgments in a variety of domains (Hertwig \& Pleskac, 2010; Lieder, Griffiths, \& Goodman, 2012; Vul et al., 2014), we propose a model for resource-constrained inference using causal models. In particular, we propose that, when reasoning about causal systems, people attend to concrete cases and shift attention between those cases systematically. This process yields a joint distribution as a representation of the causal system, which can be used for inference in any task that can be modeled with causal graphical models.

## Formalization

The proposed model is a variant of Metropolis-Hastings (MH) Markov Chain Monte Carlo, a computationally efficient rejection sampling method (Hastings, 1970). MH is defined by two components: a proposal distribution $\mathbb{Q}\left(q^{\prime} \mid q\right)$ and a transition probability $a\left(q^{\prime} \mid q\right)$, where $q$ is the current state and $q^{\prime}$ is the proposal state in the random walk. Whereas MH models often deal with a continuous state space, the proposed model samples over the discrete states of a causal model. Figure 1B presents the eight states for the three variable graph shown in Figure 1A.

The sampling process uses the standard MH transition probability:

$$
a\left(q^{\prime} \mid q\right)=\min \left(1, \frac{\pi\left(q^{\prime}\right)}{\pi(q)}\right)
$$

where $\pi(q)$ is the joint probability of the graph being in state $q$ given the graph's parameters (see the Appendix for an example of how $\pi(q)$ is calculated). The parameters reflect the particular beliefs of the participant (e.g. the causal strength between cause and effect).

We assume a proposal distribution $\mathbb{Q}\left(q^{\prime} \mid q\right)$ that restricts reachable states $q^{\prime}$ to those that differ from the current state $q$ by one binary variable. Each reachable state has an equal
probability of being selected. Edges in Figure 1B denote reachable states for a node.

Note that this proposal distribution confers additional efficiency benefits. Because only one variable is changed, the ratio $\frac{\pi\left(q^{\prime}\right)}{\pi(q)}$ simplifies to

$$
\frac{\pi\left(v_{i}^{\prime}, v_{-i}\right)}{\pi\left(v_{i}, v_{-i}\right)}=\frac{\pi\left(v_{i}^{\prime} \mid v_{-i}\right) \pi\left(v_{-i}\right)}{\pi\left(v_{i} \mid v_{-i}\right) \pi\left(v_{-i}\right)}=\frac{\pi\left(v_{i}^{\prime} \mid v_{-i}\right)}{\pi\left(v_{i} \mid v_{-i}\right)}
$$

where $v_{i}$ is the value of node $i$ in $q$, and $v_{i}^{\prime}$ is the value in $q^{\prime}$. This reduces the problem to calculating the relative conditional probabilities of two states, rather than representing the entire joint distribution. That calculating conditional probabilities only requires consideration of the node's Markov blanket further aids efficiency (Koller \& Friedman, 2009).

The model thus far is simply an efficient MH model for estimating a causal graph's joint distribution. Importantly, however, we introduce a bias in the starting point for sampling: It always starts sampling from 'prototype' states, those in which nodes are either all 0 or all 1 (bottom left and top right corners of Figure 1B). This assumption is inspired by JohnsonLaird's influential Mental Models theory, in which the most easily represented state is the one where antecedent and consequent are both true (Johnson-Laird \& Byrne, 2002). We propose that prototype states are the most easily represented states of a causal graph.

Regardless of our proposal distribution and biased initial samples, with many samples (e.g., $10^{6}$ ), the causal sampler will converge to the normative distribution. However, we assume that people are resource-constrained and thus can only take a few samples (on the order of less than twenty). In this range, an MH model will overestimate the probability of states near the starting point (as it did not have time to fully explore the state space) and underestimate the remaining states. This effect is shown in Figure 2.


Figure 1: (A) Common effect network. (B) Possible concrete states of a common effect network. Filled in circles indicate a value of 1 , empty circles indicate a value of 0 .


Figure 2: Joint distributions with causal strength $=.5$, causal prevalence $=.5$, strength of background causes $=.33$. The blue line (solid points) represents the joint distribution entailed by the normative model. Red lines (open points) represent the joint distributions simulated by the causal sampler, with thicker lines meaning fewer samples (thick $=4$ samples, medium $=8$, thin $=32$ ).

An important prediction of the causal sampler model is that Markov violations are not resultant from a particular reasoning or learning process. Instead, these violations are built into the representations of causal graphs and so will propagate to any task that used the representation. To test this prediction, we compared our model to standard Bayes nets on existing data sets in causal learning and reasoning, as well as a new task introduced at the end of this paper.

## Task 1: Causal Reasoning

The causal sampler model accounts for the independence violations found in human causal reasoning. For example, Rehder and Waldmann (2016) assessed the inferences people draw with the simple common effect graph in Figure 1A. Subjects were first instructed on two causal relationships that formed a common effect graph in the domains of economics, sociology, or meteorology (see the new experiment below for examples of these materials). The causal relationships were described as generative (a cause makes the effect more likely) and independent (each cause can bring about the effect on its own). Subjects were then asked to draw a number of causal inferences. For example, they were asked to estimate both $p\left(Y_{A}=1 \mid Y_{B}=1\right)$ and $p\left(Y_{A}=1 \mid Y_{B}=0\right)$ and also the same questions with the role of $Y_{A}$ and $Y_{B}$ reversed; thus, these inferences will be referred to as $p\left(Y_{i}=1 \mid Y_{j}=1\right)$ and $p\left(Y_{i}=1 \mid Y_{j}=0\right)$. The Markov condition stipulates that the two $Y \mathrm{~s}$ should be conditionally independent, that is, that $p\left(Y_{i}=1 \mid Y_{j}=1\right)$ should equal $p\left(Y_{i}=1 \mid Y_{j}=0\right)$. The empirical results shown in the left hand side of Figure 3 (gray bars)
reveal that subjects judged that $p\left(Y_{i}=1 \mid Y_{j}=1\right)>p\left(Y_{i}=\right.$ $\left.1 \mid Y_{j}=0\right)$ instead. This violation of independence is also illustrated by the normative fit of the common effect graphical model in Figure 1A (blue solid line) to the ratings of Rehder and Waldmann's subjects (which included conditional probability queries other than those shown in Figure 3) ${ }^{1}$. As expected, the normative model is constrained to predict that $p\left(Y_{i}=1 \mid Y_{j}=1\right)=p\left(Y_{i}=1 \mid Y_{j}=0\right)$. This apparent expectation that the causes of a common effect graph are positively correlated has been observed in other studies (e.g. Rehder \& Burnett, 2005; Rehder, 2014; Rottman \& Hastie, 2016) and violations of the Markov condition have been observed with other graph topologies (see Hagmayer, 2016, and Rottman \& Hastie, 2014, for reviews).


Figure 3: Data from Rehder \& Waldmann (2014), Experiment 1. Sampler (red lines) and normative (blue lines, solid points) fits to conditional probability judgments. Error bars denote $95 \%$ confidence intervals.

Figure 3 also presents the best fit of the causal sampler to these data (red solid line) and shows that an average of 17.9 samples in fact reproduces subjects' belief that $p\left(Y_{i}=1 \mid Y_{j}=\right.$ 1) $>p\left(Y_{i}=1 \mid Y_{j}=0\right)^{2}$. It does so because the two prototype

[^313]states are such that the causes are congruent $\left(Y_{i}=1 \& Y_{j}=1\right)$ or $\left(Y_{i}=0 \& Y_{j}=0\right)$. As was shown in Figure 2, the causal sampler overestimates these states, resulting in an inflated probability for states where the causes are congruent (e.g. $p\left(Y_{i}=1 \mid Y_{j}=1\right)$ ), and underestimates states where the causes are incongruent (e.g. $p\left(Y_{i}=1 \mid Y_{j}=0\right)$ ).

Note that the sampler also accounts for another reasoning error that subjects commit with the graph in Figure 1A. Explaining away is a signature property of common effect graphs. If $X$ is observed to occur then the probability that $Y_{A}$ is present of course increases. But if it is then further observed that the second cause $Y_{B}$ is present then the probability that $Y_{A}$ is present should decrease. (Conversely, if $Y_{B}$ is observed to be absent then the probability of $Y_{A}$ should increase.) In fact however, research finds that subjects often explain away too little or not at all (Morris \& Larrick, 1995; Rehder, 2014; see Rottman \& Hastie, 2014, for a review). The right three bars in Figure 3 illustrate the three conditional probability judgments relevant to explaining away: $p\left(Y_{i}=1 \mid X=1, Y_{j}=0\right)$, $p\left(Y_{i}=1 \mid X=1\right)$, and $p\left(Y_{i}=1 \mid X=1, Y_{j}=1\right)$. The fits of the normative model to these data points reveal that explaining away with Rehder and Waldmanns subjects was indeed too weak (the slope of the blue line is steeper than the empirical ratings). In contrast, the fit of the sampler predicts this too-weak explaining away (the slope of the red line is shallower). Because it predicts both independence violations and weak explaining away, the sampler achieves a better fit to these data according to a measure (AIC) that corrects for its extra parameter ( 30.3 vs. 33.6 ).

## Task 2: Causal Learning

The causal sampler also outperforms the normative model in a causal learning experiment. Rehder and Davis (2016) tested for the presence of independence violations in a hypothesis testing task by presenting subjects with a candidate theory that took the form of the graph in Figure 1A (again, in either the domain of economics, meteorology, or sociology). Subjects were then presented with hypothetical data and asked to rate the likelihood of observing the data if the theory was true. The correlation between $Y_{A}$ and $Y_{B}$ that obtained in the data was manipulated to be either negative, zero, or positive (all other aspects of the data, e.g., causal strengths, were held constant). The empirical results shown in Figure 4 (gray bars) revealed that subjects' ratings were largest when the inter-Y correlation was positive and smallest when it was negative.

The normative model's predictions for this task were derived by, for each of the three data sets, identifying the maximum likelihood parameters for the graph in Figure 1A to that data set. Using simple linear regression, the three maximum log-likelihoods were then scaled and translated onto the subjects' $0-100$ ratings. The fitted predictions averaged over subjects (blue line in Figure 4) show the expected result that the data set with a zero inter-Y correlation is more likely than

[^314]those with non-zero correlations, reflecting the independence between the causes stipulated by the normative model. ${ }^{3}$

The same process was followed for the causal sampler with the elaboration that we performed a grid search on the number of samples from 1 to 32. The fitted predictions (red line in Figure 4) reveal that the model, like the subjects, judged that the data set with the positive $Y_{A}-Y_{B}$ correlation is most likely to be generated by the candidate theory (chain length averaged over subjects was 2.3). As in conditional probability judgments, it makes this prediction because biased sampling (starting at the prototypes) combined with a limited number of samples naturally generates the expectation that $Y_{A}$ and $Y_{B}$ will be positively correlated.


Figure 4: Rehder \& Davis (2016). Sampler (red lines) and normative (blue lines, solid points) fits to data likelihood judgments. Error bars denote $95 \%$ confidence intervals.

## Task 3: Expected Distributions

Recall that when the causal sampler's number of samples is limited, it warps a causal graph's joint distribution, overestimating prototype states and underestimating others (Figure 2). The following experiment tests this account using a novel methodology, one that directly asks participants to generate a distribution for a causal graph.

## Method

Materials. Participants were presented with causal hypotheses in one of three domains: meteorology, sociology, or economics. Each domain had three variables (in economics: interest rates, trade deficits, and retirement savings; in meteorology: ozone levels, air pressure, and humidity; in sociology:

[^315]urbanization, interest in religion, and socioeconomic mobility). Each variable could take on two possible values. One of these values was described as "Normal" and the other was either "High" or "Low". The values of the variables were mixed to prevent domain-specific beliefs from affecting the results (alternate values were either all "High", all "Low", or a mixture of "High" and "Low"). All hypotheses were of the form shown in Figure 1A, with two causes of one effect.
Procedure. Participants first studied screens of information that defined the variables, provided a mechanism describing how each cause could independently generate the effect, and a diagram of the causal relationships. They were then required to pass a multiple-choice test of this knowledge.

Next, participants were asked to generate a data set that they would expect to result from the causal graph. The causal relationship between smoking and lung cancer was used as an example. Subjects were shown the four cells formed by crossing smoker/non-smoker with lung cancer/no-lung cancer and how (in terms of how hypothetical people were allocated to the four cells) a greater proportion of smokers had lung cancer as compared to non-smokers. Subjects were asked to generate an analogous distribution in their assigned domain (economics, etc.). Specifically, they were given 50 pennies and asked to distribute them among the cells formed by crossing the three binary variables. They did so by placing the coins on a large sheet that contained the eight possible states (the position of the states on the sheet was randomized).
Design and Participants. The experiment consisted of a 3 (domain) by 4 (variable senses, e.g., all "High") betweensubjects design. 60 New York University undergraduates received course credit for participation.

## Results

Figure 5 presents how subjects allocated the 50 pennies to the eight states of the graph in Figure 1A (gray bars). Because these raw data are difficult to interpret, we computed measures that reflect the statistical relationships among the three binary variables implied by the pennies. In particular, we first normalized a subject's distribution and then computed the phi coefficients between a $Y$ and an $X\left(\phi\left(Y_{i}, X\right)\right.$; the pennies were aggregated so that the two $Y$ s are interchangeable), between the $Y$ s themselves $\left(\phi\left(Y_{A}, Y_{B}\right)\right)$, and between the $Y \mathrm{~s}$ conditioned on the presence of $X\left(\phi\left(Y_{A}, Y_{B} \mid X=1\right)\right)$. These measures averaged over subjects are presented in Figure 6. First, the fact that $\phi\left(Y_{i}, X\right) \gg 0$ indicates that subjects understood that the $Y$ s were generative causes of $X$. Of greater theoretical importance is the fact that $\phi\left(Y_{A}, Y_{B}\right)$ was also significantly greater than $0, t(59)=3.62, p<.001$. This corroborates our claim that the violations of independence that obtain during causal reasoning (Figure 3) and hypothesis testing (Figure 4) are also manifested in peoples' causal representations (Figure 5).

The best fit of the normative model is shown superimposed


Figure 5: Causal sampler (red line) and normative (blue line, solid points) fits to participant-generated expected distribution judgments. Error bars denote 95\% confidence intervals.
on the empirical data in Figure 5 (blue line) ${ }^{4}$. The figure indicates that the normative model tends to underpredict subjects' judgments for the two prototype states (111 and 000) and overpredict the remaining states. Phi coefficients computed for these fits (blue line in Figure 6) show the expected result that the normative model requires that $\phi\left(Y_{A}, Y_{B}\right)=0$, at odds with subjects' distributions. Moreover, it sharply underpredicts $\phi\left(Y_{A}, Y_{B} \mid X=1\right)$. Because of the explaining away phenomenon described above, the normative model requires that $\phi\left(Y_{A}, Y_{B} \mid X=1\right)$ is negative (one cause is less likely when the other is present). Figure 6 shows that subjects' distributions implied a value of $\phi\left(Y_{A}, Y_{B} \mid X=1\right)$ that is less negative (i.e., explaining away was again too weak).

The best fit of the causal sampler (red lines in Figs. 5 and 6) shows that it accounts for the fact that, relative to the normative model, the number of pennies is generally too large for the prototype states and too small for other states ${ }^{5}$. Of course, this pattern was expected on the basis of the theoretical predictions in Figure 2. Like the subjects, the causal sampler predicts that $\phi\left(Y_{A}, Y_{B}\right)>0$ and that explaining away (as represented by $\phi\left(Y_{A}, Y_{B} \mid X=1\right)$ is too weak relative to the normative model. As a result, it achieved a better fit to these data than the normative model (AIC of 3.2 vs. 10.8).

## Discussion

Although causal graphical models have enjoyed success in explaining causal cognition, people consistently violate key predictions of these models. That independence violations manifest themselves in multiple tasks suggests that they arise from the causal representations that people construct. This

[^316]

Figure 6: Causal sampler (red line) and normative (blue line, solid points) fits to participant-generated expected distribution judgments. Error bars denote $95 \%$ confidence intervals.
conjecture was confirmed in an experiment using a new methodology that assessed, in a relatively direct way, people's causal representations. This result suggests that the fault lies not in how we reason or learn but how we represent.

This paper has proposed a process model that naturally constructs faulty causal representations. Importantly, it does so in a manner that is computationally efficient and psychologically plausible. The Metropolis-Hasting rule combined with the proposal distribution we advocate implies that at any one time reasoners only need to consider the relative likelihood of two graph states that differ by one variable, a computation that can be carried out very efficiently (because it involves only those nodes in the variable's Markov blanket; Koller \& Friendman, 2009). Yet further efficiencies can be achieved for conditional probability queries (because sampling can be limited to those graph states that instantiate a query's antecedent). Note that this view suggests that humans could construct veridical causal representations-if only they had the cognitive resources to do so. The fault thus lies not in our causal representations per se but rather in the fact that causal judgments must be computed in finite time and with limited resources. Independence violations are thus an unavoidable consequence of the tradeoff between accuracy, speed, and effort.

The causal sampler perhaps gains some credence given the property it shares with the well-known Mental Model theory, namely, that reasoning is based on concrete states of the world (Goldvarg \& Johnson-Laird, 2001; JohnsonLaird \& Byrne, 2002). There are, however, some differences. Whereas the model theory never represents cause-present/effect-absent situations, the causal sampler, as a probabilistic model, merely asserts that such situations are un-
likely (depending on the causal graph's parameters) and thus rarely sampled (cf. Khemlani, Barbey, \& Johnson-Laird, 2014). There are also differences regarding which states reasoners initially consider (initial mental models are similar but not identical to the causal sampler's starting samples).

The causal sampler accounts for independence violations with other graph topologies. For example, suppose the direction of causality in Figure 1A is reversed, yielding a common cause graph. Independence is then captured by the screening off principle whereby the effects ( $Y_{A}$ and $Y_{B}$ ) are independent conditioned on the cause $X$. In fact, people judge that $p\left(Y_{i}=1 \mid X=1, Y_{j}=1\right)>p\left(Y_{i}=1 \mid X=1, Y_{j}=0\right)$ instead (Rehder, 2014; Rehder \& Waldmann, 2016; Rehder \& Burnett, 2005). The causal sampler predicts this result as well (because biased sampling induces a positive correlation between the $Y$ s conditioned on $X$ ).

There are many possible directions for future research. For one, current models do not attempt to model the substantial variability in peoples causal inferences (Rehder, 2014; Rottman \& Hastie, 2016). The stochastic nature of sampling may shed light on this important aspect of behavior. The causal sampler also makes predictions about reaction times. For example, it would predict that longer reaction times implies a less warped joint distribution (because more samples were taken).

Research in the causal graphical model tradition has rarely considered the cognitive processes involved in causal-based judgments. A limited sampling approach to building causal representations (a) is psychologically plausible, (b) accounts for the key discrepancy between graphical models and human judgments (Markov violations), and (c) explains why those discrepancies manifiest themselves in multiple causalbased tasks. Yet, it doesn't deny that people are sophisticated causal reasoners-they are, however, limited ones. As a process model, the causal sampler allows the causal graphical model framework to be extended to new phenomena, such as within- and between-subject variability and response times.

## References

Goldvarg, E., \& Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. Cognitive Science, 25(4), 565-610.
Hausman, D. M., \& Woodward, J. (1999). Independence, invariance and the causal Markov condition. The British journal for the philosophy of science, 50(4), 521-583.
Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 97-109.
Hertwig, R., \& Pleskac, T. J. (2010). Decisions from experience: Why small samples? Cognition, 115
Johnson-Laird, P. N., \& Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. Psychological Review, 109(4), 646.
Khemlani, S. S., Barbey, A. K., \& Johnson-Laird, P. N. (2014). Causal reasoning with mental models. Frontiers
in Human Neuroscience, 8, 849.
Koller, D., \& Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.
Lieder, F., Griffiths, T. L., \& Goodman, N. D. (2012). Burnin, bias, and the rationality of anchoring. In P. PBartlett, et al. (Eds.), Advances in neural information processing systems (Vol. 25, pp. 2699-2707). Cambridge, MA: MIT Press.
Morris, M. W., \& Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. Psychological Review, 102, 331-355.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.
Rehder, B. (2014). Independence and dependence in human causal reasoning. Cognitive Psychology, 72, 54-107.
Rehder, B., \& Burnett, R. C. (2005). Feature inference and the causal structure of object categories. Cognitive Psychology, 50, 264-314.
Rehder, B., \& Davis, Z. (2016). Evaluating causal hypotheses: The curious case of correlated cues. In Papafragou, A., et al. (Eds.) (2016). Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Rehder, B., \& Waldmann, M. R. (2016). Failures of explaining away and screening off in described versus experienced causal learning scenarios. Memory and Cognition, 1-16.
Rottman, B., \& Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. Psychological Bulletin.
Rottman, B. M., \& Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. Cognitive Psychology, 87, 88-134.
Vul, E., Goodman, N. D., Griffiths, T. L., \& Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. Cognitive Science, 38, 599-637.
Waldmann, M. R., \& Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), Oxford Handbook of Cognitive Psychology (pp. 733-752). New York: Oxford University Press.

## Appendix

To calculate $\pi(q)$ (the probability of being in some state $q$ ), we simply use the normative calculation for each potential state. For example, when causal relations are generative, operate independently, and combine according to a noisy-or integration rule, $\pi(q)$ is defined as:

$$
1-\left(1-b_{j}\right) \prod_{q_{i} \in P a_{k}\left(q_{j}\right)}\left(1-m_{i j}\right)^{i n d\left(q_{i}\right)}
$$

where $b_{j}$ is the strength of causes exogenous to the model on the node, $\operatorname{Pa} k\left(q_{j}\right)$ denotes the parents of $q_{j}$ in the causal model, $m_{i j}$ denotes the causal strength between node $j$ and parent $i$, and $\operatorname{ind}\left(q_{i}\right)$ is an indicator function that yields 1 if feature $q_{i}$ is present, 0 otherwise.

# Understanding the Role of Perception in the Evolution of Human Language 

Isaac Davis (isaacd@andrew.cmu.edu)<br>Department of Philosophy, Carnegie Mellon University


#### Abstract

In this paper, we propose a flexible modeling framework for studying the role of perception in language learning and language evolution. This is achieved by augmenting some novel and some existing evolutionary signaling game models with existing techniques in machine learning and cognitive science. The result is a "grounded" signaling game in which agents must extract relevant information from their environment via a cognitive processing mechanism, then learn to communicate that information with each other. The choice of cognitive processing mechanism is left as a free parameter, allowing the model to be tailored to a wide variety of problems and tasks. We present results from simulations using both a Bayesian perception model and a neural network based perception model, which demonstrate how perception can "preprocess" environmental data in a way that is well suited for communication. Lastly, we discuss how the model can be extended to study other roles that perception may play in language learning. Keywords: Evolutionary Signaling Games, Perception, Language Evolution, Reinforcement Learning


## Introduction

In this paper, we are interested in studying three broadly defined types of interaction between perception and language learning. The first, and perhaps most obvious, is how our perceptions of the world constrain and affect our ability to learn language. The second type of interaction is the reverse of the first- how does learning a language shape our perceptions of the world? The third interaction, with a long history of philosophical inquiry, is how differing cognitive representations between agents affects their ability to communicate with each other. We propose a flexible modeling framework that can be used to represent and study all three types of interactions between perception and language learning.

This framework has two core components. The first are evolutionary signaling games which have been used extensively to explain a wide variety of communication phenomena in animals, from mating calls and warning calls in mammals and birds, to pheromone signaling in insect communities. More recent work has applied these models to the evolution of human language and human linguistic phenomena, such as compositionality (Franke 2015) and convex perceptual categories (Jager 2007, O'Connor 2014). These models typically represent a situation in which two agents must coordinate their behavior in such a way as to achieve some common goal, but without any pre-defined common language with which to communicate.

We shall see, however, that the standard signaling game model is not well suited for studying interactions between cognition and language learning. In order to represent perception and language in the same model, we derive a "grounded" signaling game, in which agents respond to raw sensory inputs, rather than cleanly-defined information states. We
achieve this using unsupervised learning techniques from Machine Learning and Cognitive Science.

In the following section, we present the relevant background in signaling games and reinforcement learning, and briefly discuss why the standard model is not yet equipped for studying interactions between perception and language learning. We then present our grounded signaling game model, and draw on literature in Deep Reinforcement Learning to derive an effective learning rule. In order to avoid introducing too much complexity at once, we first present the model with a "trivial" perception mechanism, consisting of the identity map on sensory inputs. After deriving the learning rule for this model, we then outline how to incorporate a non-trivial perception mechanism, and briefly discuss the two perception mechanisms we will test. We then present results from a battery of experiments in which the agents must learn to communicate structured visual information from a synthetic image environment. These results provide insights into the role of perception as a "pre-processing" step for communication, and reveal some interesting interactions between the type of environment and perception mechanism. Lastly, we discuss how the current model can be extended to study all three types of interaction addressed in the introduction.

## Background and Related Material

## Signaling Games

In a two-player signaling game, the sender observes an information state $d \in \mathcal{D}$, drawn from some probability distribution $P(\mathcal{D})$. The sender then chooses a signal $x \in X$ according to some decision rule, and transmits the signal to the receiver. The receiver then observes the signal, and chooses some act $a \in \mathcal{A}$ according to the receiver's decision rule. Both players then receive a payout $R(d, a)$, which is a function of the state, and act, but not the signal. The game is fully cooperative, in that both players receive the same payout every round.

The sender's decision rule can be represented by a function assigning each state $d$ to a probability distribution $P_{s}(x \mid d)$ over signals, and likewise for the receiver's rule $P_{r}(a \mid x)$. With this notation, the expected reward for both players is given by

$$
\begin{equation*}
E\left[R \mid P_{s}, P_{r}\right]=\sum_{d \in \mathcal{D}, x \in X, a \in \mathcal{A}} P(d) P_{s}(x \mid d) P_{r}(a \mid x) R(d, a) \tag{1}
\end{equation*}
$$

An evolutionary signaling game is played for many rounds, and after each round, players update their strategies according to some learning rule. The most well-studied and often used form of reinforcement learning in signaling games is Roth-Erev reinforcement learning. A Roth-Erev learning agent (the sender, for this example) is represented by a set of decision parameters $W^{s}$, one parameter for each state-signal
pair. Each weight $w_{i, j}^{s}$ represents the sender's unnormalized probability of choosing signal $x_{i}$ given state $d_{j}$, and the RothErev update rule is very simple. If in one round of play the sender observes state $d_{j}$, chooses signal $x_{i}$, and receives reward $R$, then the sender updates weight $w_{i, j}^{s}$ by some fixed proportion of the reward, $\Delta w_{i, j}^{s}=\alpha R$, and leaves all other weights unchanged.

In a signaling game with two equiprobable states, RothErev reinforcement learning will always converge to a signaling system equilibrium, in which the correct act is always chosen. In games with more than two states, or with nonuniformly distributed states, Roth-Erev learning will sometimes converge to a partial pooling equilibrium, in which some states are communicated accurately, and others are "pooled" into a single signal. Such outcomes are Nash Equilibria, meaning that neither player can improve the outcome by unilaterally changing their own strategy. The probability of converging to such partial pooling equilibria increases with the number of states (Huttegger, Skyrms, Smead, \& Zollman 2010).

While our general framework is compatible with any signaling game model, the experiments for this paper will be based on the Sim-Max game (Jager 2007), a variation of the standard signaling game. In a Sim-Max game, the receiver's goal is to guess which state the sender observed, and the reward function is a distance metric representing "similarity" between states. Such models are used when the state space is much richer than the signal space, and one-to-one state-signal mappings are no longer possible. An example of this is coloreven though colors vary continuously, we employ a relatively small number of words for describing them.

## Limitations of the Standard Model

The Roth-Erev reinforcement learning model is favored in signaling games for its simplicity and minimality of cognitive assumptions. However, there are two factors that prevent us from using the standard model for our purposes. The first factor is entirely practical. In particular, Roth-Erev reinforcement learning, while simple and well studied, does not scale well to larger problems, as it requires a single decision parameter for each state-signal pair. This makes Roth-Erev learning prohibitively slow to converge on large problems, and restricts us to relatively small simulations.

The second problem with the standard model is more conceptual. In the standard treatment of signaling games, states are represented as uniform, discrete "labels," or in the usual Sim-Max game as a uniformly distributed compact subset of Euclidean space. The key feature in these models is that there is no internal structure to the states themselves. That is, the players cannot discriminate between two states except through the reward function- two different states are either identical or not, but any further delineation between states can only be inferred from their effect on the payout. This, however, is often regarded as an advantage of the standard representation, rather than a limitation. The agents need not be
endowed with an inner mental language (Skyrms 2010), they need not know what the game is "about," or even that they are playing a game at all. It is certainly important, if we wish to study interactions between perception and language-learning, that we not make any restrictive cognitive assumptions which would "screen off" the cognitive details of interest. However, outside of very simple or tightly controlled experimental settings, the standard representation actually imposes some very strict assumptions about the agents' perceptions, albeit implicitly.

To illustrate this, consider the following two very similar, but not identical signaling games: in Game 1 , the sender observes one of ten cards $C_{1}, \ldots, C_{10}$, each of which depicts a digit $0-9$, and sends one of ten signals. The receiver must pull on one of ten levers, each of which bears a digit $0-9$. If the card and chosen lever show the same number, both players receive a reward of 1 , otherwise they receive no payout. This game fits exactly into the signaling game framework described above.

Now consider Game 2: in Game 2, the signals and the receiver's actions are the same as Game 1. However, instead of observing one of ten cards, the sender now observes a card depicting a handwritten digit $0-9$. How would we represent the state space for Game 2? A first guess might be to represent it as having the same state space as Game 1 , since each observation depicts one of 10 digits. But unlike Game 1, we cannot guarantee uniformity across separate instances of the same digit. Not all instances of 0 will look the same, and so they are distinct information states. To assume that they are not distinct information states is to assume that the sender perceives them as the same, or recognizes that they are of the same category. But these are acts of cognition. Seeing a digit handwritten on a card is not the same an recognizing which digit it depicts, or even recognizing that it is a digit at all. In this sense, the standard approach of representing states and acts as finite sets of distinct labels implicitly imposes presupposes a fixed way in which agents perceive, recognize, and process their environments.

This brings us to the two underlying principles of our framework. First, in order to represent all of the information available to agents, without explicitly assuming how they perceive or represent that information, we must represent states as raw sensory inputs, rather than discrete information states. Second, because sensory inputs tend to be high-dimensional, it is no longer the case, as with the standard signaling game, that states cannot be distinguished except through the reward function. In particular, Unsupervised Learning algorithms can infer informational structure from high-dimensional sensory inputs without any feedback or supervision. Thus, by integrating an unsupervised perception model into a "grounded" signaling game, we can represent the evolution of both the external signaling language and the agents' internal representations in the same framework.

## The Model

In this section, we present the model used in the experiments for this paper. We first present the model with a "trivial" perception mechanism, which performs no significant cognitive processing. Once we derive the learning rule for this model, we then outline how a non-trivial perception mechanism can be incorporated. The framework is designed to place no restrictions on what perception model we use, and so we will test two mechanisms, representing two schools of thought on modeling cognitive processing.

## The Grounded Signaling Game

In the experiments we present here, the state space $\mathcal{D}$ will be a synthetic image environment, presented to the sender as a vector of raw pixel values, and the signal space $X=\{0,1\}^{k}$ will consist of binary vectors of length $k$. As with the SimMax game, the receiver's action will be produce a guess dout as to which state $d^{\text {in }}$, the sender initially observes, based on the sender's signal $x$. The reward function will be a distance metric representing similarity between images.

Recall that a Roth-Erev learning agent requires one parameter for each state-signal pair. In these experiments, however, we will use 36 -pixel binary images, and signals between 4 and 36 bits in length. Even though only a small number of possible images $d \in\{0,1\}^{36}$ will ever appear with non-zero probability, the players do not know ahead of time which images are present, or how many appear with positive probability. This is an important distinction, as the receiver must reconstruct the original image pixel-by-pixel, rather than simply guessing from a list of potential images. Therefore, defining a Roth-Erev reinforcement learning agent for this game would require up to $2^{72}$ separate decision parameters for each player. Clearly this is intractable even for small images, and we must look elsewhere to derive a tractable learning rule. To this end, we draw on the representational flexibility of Artificial Neural Networks (ANNs).


Figure 1: Information flow for one round of the signaling game

Figure 1 depicts a single round of the signaling game. Rather than defining the sender with one decision parameter per state-signal pair, we define a decision rule $P_{s}\left(x \mid d^{\text {in }}, W^{s}\right)$ with one parameter per state-signal coordinate pair. Given a state $d$, we define, for each signal-coordinate $x_{i}$ :

$$
\begin{equation*}
P_{s}\left(x_{i}=1 \mid d\right)=\sigma\left(\sum_{j=1}^{n} W_{i, j}^{s} d_{j}\right) \tag{2}
\end{equation*}
$$

where $\sigma$ is the standard sigmoid activation function $\sigma(x)=$ $1 /\left(1+e^{-x}\right)$. Those familiar with ANNs will note that this is the standard expression for the activation value of a unit with sigmoid activation function. In the stochastic network corresponding to this game, the activation value is treated as a probability, and the output of unit $x_{i}$ is sampled from the Bernoulli distribution $x_{i} \sim \operatorname{Bernoulli}\left(P_{s}\left(x_{i}=1 \mid d\right)\right.$ ). Each weight $W_{i j}^{s}$ determines, in some sense, the "importance" of coordinate $j$ in determining the value of signal component $x_{i}$. The sender then generates the signal $x$ by first computing the activation probability for each signal coordinate, then independently sampling each coordinate from the computed Bernoulli distribution. Thus, the sender's probability of sending signal $x$ given object $d$ factors as:

$$
\begin{equation*}
P_{s}(x \mid d)=\prod_{i=1}^{k}\left(P_{s}\left(x_{i}=1 \mid d\right)\right)^{x_{i}}\left(1-P_{s}\left(x_{i}=1 \mid d\right)\right)^{1-x_{i}} \tag{3}
\end{equation*}
$$

The receiver's distribution is defined similarly, with the roles of $d$ and $x$ being reversed. With $P_{s}$ and $P_{r}$ defined as above, the expected reward function in equation (1) can be interpreted as an objective function $J\left(W_{s}, W_{r}\right)=E\left[R \mid P_{s}, P_{r}\right]$ to a multi-agent optimization problem, where both players wish to maximize $J\left(W_{s}, W_{r}\right)$, but each player directly controls only one set of parameters. For this experiment in particular, the cooperative objective is to output an image that is most similar to the input image, which is the same objective used in training certain types of auto-encoders (a type of unsupervised ANN, trained to accurately reconstruct its own inputs). Thus, we can efficiently represent a single round of this signaling game as a single forward pass through the three-layer stochastic-sigmoid auto-encoder network shown in figure 1.

## The Learning Rule

An auto-encoder, like most feed-forward neural networks, is generally trained using some variation of gradient descent via back-propagation of errors ${ }^{1}$. In each step of the backpropagation algorithm, an input vector is passed through the network, generating a hidden representation (in this case signal) from which the latter half of the network attempts to reconstruct the original input. The error signal (difference between input and output) at each unit is "propagated" backwards, and each weight is adjusted according to its "effect" on the resulting error. Back-propagation algorithms have been extremely successful in training neural networks to perform highly complex tasks, so it is tempting to co-opt the backpropagation algorithm as a "learning rule" for our two agents. However, even though we can represent our signaling game with a three-layer feed-forward network, a key assumption of the model prevents us from using back-propagation directly. In particular, the back-propagation algorithm computes the update to layer $l$ as a function of the parameter and activation

[^317]values of layer $l+1$. In this scenario, however, each layer represents a separate human agent, who cannot share parameter information with each other, thus preventing the requisite gradient information from flowing across agents.

Because of this, we instead use a REINFORCE learning rule, first named in Williams (1992). Consider a single round of the signaling game in which sender observes state $d=\left(d_{1}, \ldots, d_{n}\right)$, sends signal $x=\left(x_{1}, \ldots, x_{k}\right)$, receiver guesses state $d^{\prime}=\left(d_{1}^{\prime}, \ldots, d_{n}^{\prime}\right)$, and both players receive reward $R\left(d, d^{\prime}\right)$. We define $\Delta W_{i, j}^{s}$ and $\Delta W_{i, j}^{r}$, the weight updates for sender and receiver, as

$$
\begin{align*}
\Delta W_{i, j}^{s} & =\varepsilon\left(R\left(d, d^{\prime}\right)-b_{i j}\right)\left(x_{i}-P_{s}\left(x_{i}=1 \mid d\right)\right) d_{j}  \tag{4}\\
\Delta W_{i, j}^{r} & =\varepsilon\left(R\left(d, d^{\prime}\right)-b_{i j}\right)\left(d_{i}^{\prime}-P_{r}\left(d_{i}^{\prime}=1 \mid x\right)\right) x_{j} \tag{5}
\end{align*}
$$

where $\varepsilon$ is a learning rate and $b_{i j}$ is a reinforcement baseline. The main property of REINFORCE rules, as shown in (Williams 1992), is that the weight updates shown in equations (4) and (5) are unbiased estimates of the true gradient of $J\left(W_{s}, W_{r}\right)$, the expected reward function. That is, $\left(R\left(d, d^{\prime}\right)-b_{i j}\right)\left(x_{i}-P_{s}\left(x_{i}=1 \mid d\right)\right) d_{j}$ is an unbiased estimate of $\partial J / \partial W_{i j}^{s}$, and $\left(R\left(d, d^{\prime}\right)-b_{i j}\right)\left(d_{i}^{\prime}-P_{r}\left(d_{i}^{\prime}=1 \mid x\right)\right) x_{j}$ is an unbiased estimate of $\partial J / \partial W_{i j}^{r}$. This allows the two players to cooperatively implement an approximate gradient descent algorithm, despite the fact that neither player is explicitly computing any gradients. This rule is both computationally inexpensive and avoids the information-sharing problem of backpropagation, so it is well suited to our task.

Even though equations (4) and (5) are unbiased estimates of the true gradient, they can be very high variance estimates, so outside of very simple tasks, the "pure vanilla" REINFORCE rule (i.e. $b_{i j}=0$ ) can be hopelessly slow to converge. We therefore use a minimum variance, unit-specific baseline derived in Bengio (2013), given by the expression.

$$
\begin{equation*}
b_{i j}=\frac{E\left[\left(h_{i}-\sigma\left(a_{i}\right)\right)^{2} R\right]}{E\left[\left(h_{i}-\sigma\left(a_{i}\right)\right)^{2}\right]} \tag{6}
\end{equation*}
$$

where $h_{i}$ is the output value and $\sigma\left(a_{i}\right)$ the activation value of unit $i$. This can be easily computed on the fly by maintaining moving averages of weight updates and rewards over time.

## Adding Perception to the Model

The signaling game model we just introduced is "grounded," in the sense that states are represented as sensory inputs, but we have yet to incorporate a perception mechanism. While perception is a broadly defined and widely studied subject, we will adopt a very general stance on what constitutes "perception." We will take perception to be any map $F: \mathcal{D} \rightarrow \mathcal{Z}$ from states (represented as sensory inputs) to lower-dimensional internal representations $Z$. These internal representations can be interpreted as the features, categories, concepts, patterns, rules, etc. from which our higher-level decisions are made. For these particular experiments, we shall use perception models that learn both a recognition map and a generative map. The recognition map $F: \mathcal{D} \rightarrow Z$ infers a latent
representation $z$ from an object $d$, while the generative map $F^{-1}: Z \rightarrow \mathcal{D}$ generates an object $d$ from an internal representation $z$ (this is a slight abuse of notation, as the generative map will not in general be the inverse of the recognition map).


Figure 2: Perception as a pre-processing step prior to signaling game

In the experiments we present here, perception will take the form of a pre-processing step (figure 2). Prior to playing the signaling game, each player will independently sample images from their environment, engaging in unsupervised learning, training their recognition maps $F_{s}, F_{r}$ as well as generative maps $F_{s}^{-1}, F_{r}^{-1}$. Once the perception mechanisms have been trained, the signaling game proceeds as usual, except that the sender now makes their signaling decision $P_{s}\left(x \mid z_{s}^{i n}, W^{s}\right)$ as a function of the sender's internal representation $F_{s}\left(d^{i n}\right)=z_{s}^{i n}$ of the state $d^{i n}$. The receiver then observes the signal, and first generates an internal representation $z_{r}^{\text {out }}$, which is then mapped to output image $d_{r}^{\text {out }}$. The reward value is then computed as usual, and we apply the same REINFORCE updates in (4)-(5) to the player's decision parameters.

## Representing Perception

There has been surge of interest in computational models of perception, from both Machine Learning and Cognitive Science. We will test two different models, representing two main approaches that have been taken in studying perception.

The first are Bayesian models, which represent perception as a rational inference problem. Given object $d$, we infer a latent representation $z$ by maximizing the posterior probability $P(z \mid d) \propto P(d \mid z) P(z)$ using Bayes' rule. This requires an object model $P(d \mid z)$, as well as a prior distribution $P(z)$ over all possible latent representations. We shall use an Infinite Latent Feature Model, which learns a binary feature representation of visual data, without having to pre-define a fixed number of latent features. This is achieved using an Indian Buffet Process (IBP) prior, which defines a probability distribution over binary vectors with an unbounded number of features (Griffiths \& Ghahramani 2005). While exact inference over this distribution is intractable, MCMC sampling methods can be used to perform tractable inference.

The second perception mechanism we shall test a Helmholtz Machine (Dayan et al 1995), representing a neurocomputing model of perception. A Helmholtz Machine is a type of variational auto-encoder, which learns a lowdimensional representation of sensory inputs by iteratively inferring latent representations from data, then reconstructing simulated data from internal representations. These two steps are iterated in an alternating "wake-sleep" cycle, with the objective of minimizing the Kullback-Liebler divergence between the true distribution and the generative distribution. The result is a low-dimensional binary representation of the data, encoded on the hidden units of the network. While the Helmholtz Machine in its original form has been rendered largely obsolete by more powerful methods, we use this architecture for its relative simplicity and pedagogical value in the context of our goals.

## Experiments and Results

In this section we describe the environments and experimental conditions we tested, present the results of these experiments, and discuss their implications.

## Experimental Conditions

We tested three different 36 -pixel synthetic image environments, each intended to represent a different kind of informational structure. For the first environment (PICTURE), we defined 8 specific 36-pixel images, each equally likely to appear, and assign 0 probability to all other images. These 8 images were chosen so as to avoid recurring components or features across images. This serves as a baseline evaluation for the perception mechanisms- we can think of the PICTURE environment as representing a "traditional" signaling game setup, in which the "true" state space consists of 8 discrete states that share no common internal structure. In the context of our project, "there are only 8 things here" represents prior knowledge or recognition, and so the players must learn that their environment contains only these 8 images directly through their sensory inputs.

The second environment (FEATURE) consists of compositionally distributed images. We define four $3 \times 3$ pixel patterns (features), and generate each $6 \times 6$ pixel image by randomly selecting any number of the four features and compos-
ing them into a single image. This construction is based on an experiment in Griffiths \& Ghahramani (2005). The third environment (HIERARCHY) is hierarchically distributed over two categories: there are 12 images with non-zero probability, depicting either horizontal or vertical bars. Images are divided into category $A$ (vertical) and category $B$ (horizontal). Within each category, each image is equally likely to appear, but images from category $A$ are twice as likely to appear as images from category $B$. The FEATURE and HIERARCHY environment test the agents' abilities to learn non-trivial informational structure from the environment.

For each environment, we test a noiseless version, in which images are presented to the sender with binary pixel values, as well as two levels of corrupting noise, in which each pixel is independently perturbed before being shown to the sender. For a reward function, we test three different distance metricsHamming $\left(L_{1}\right)$, Euclidean $\left(L_{2}\right)$, and a patch-specific function that depends only on certain regions of the image.

## Results

Figure 3 shows a summary of results across our experiments, using the Hamming metric reward function (we observed no significant differences across reward functions). Convergence rates indicate the number of iterations required to achieve a threshold $90 \%$ of optimal performance, averaged over 5 runs for each condition. The Bayesian and Helmholtz columns correspond to the two perception models, while the Identify column indicates the trivial perception mechanism that performs no cognitive processing.

| Conditions |  | Bayesian |  | Helmholtz |  | Identity |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Environment | Noise Level ${ }^{1}$ | Signal Size ${ }^{2}$ | Convergence | Signal Size ${ }^{2}$ | Convergence | Signal Size ${ }^{2}$ | Convergence |
| PICTURE | 0 | 8 | $1.1 \times 10^{5}$ | 8 | $1.4 \times 10^{5}$ | 8 | * $4.5 \times 10^{5}$ |
|  | 0.25 | 8 | $1.2 \times 10^{5}$ | 8 | $1.4 \times 10^{5}$ | 8 | *5.0x10 ${ }^{5}$ |
|  | 0.35 | 8 | $1.2 \times 10^{5}$ | 8 | $1.5 \times 10^{5}$ | 8 | * $4.8 \times 10^{5}$ |
| FEATURE | 0 | 4 | $0.3 \times 10^{5}$ | 6 | $0.8 \times 10^{5}$ | 6 | $0.8 \times 10^{5}$ |
|  | 0.25 | 4 | $0.4 \times 10^{5}$ | 6 | $0.9 \times 10^{5}$ | 6 | $0.8 \times 10^{5}$ |
|  | 0.35 | 4 | $0.4 \times 10^{5}$ | 6 | $0.9 \times 10^{5}$ | 6 | $1.1 \times 10^{5}$ |
| HIERARCHY | 0 | 4 | $1.0 \times 10^{5}$ | 4 | $0.6 \times 10^{5}$ | 4 | $1.0 \times 10^{5}$ |
|  | 0.25 | 4 | $1.2 \times 10^{5}$ | 4 | $0.8 \times 10^{5}$ | 4 | $1.5 \times 10^{5}$ |
|  | 0.35 | 4 | $1.3 \times 10^{5}$ | 4 | $0.8 \times 10^{5}$ | 4 | $1.3 \times 10^{5}$ |

${ }^{1}$ Value indicates SD of mean-zero Gaussian noise
${ }^{2}$ Smallest bit-size that converged to optimum (best pooling outcome if optimum not achieved) across all runs *Indicates partial pooling outcome

Figure 3: Results
In the PICTURE environment, both perception mechanisms drastically improved convergence speed, by effectively reducing the scale of the problem from $2^{36}$ to 8 states. This pre-processing also smoothed over irregularities that would occur under the trivial perception mechanism, where the receiver would fix the value of certain pixels across all images. Introducing the perception mechanism allows the receiver to reconstruct the image based on their own internal representations of the environment, rather than by individually choosing the value of each output pixel.

In the FEATURE environment, convergence to the optimum was fast and reliable through all levels of noise, even
with the trivial model. The fact that communication is easier to learn in the FEATURE environment than the PICTURE environment, even though the former contains twice as many states as the latter, shows that it is not just the number of distinct states that affects learning, but the content of the states themselves. However, the trivial model was only able to converge to the optimum using a 6-bit signal, which is not minimal. The Bayesian perception mechanism, however, allowed the agents to correctly identify 4 latent features in their environment, which enabled them to learn to communicate using a minimal 4-bit signal. The Helmholtz model did not lead to any reduction in signal size. This is because the Bayesian model learns the number of latent features from the data, while the Helmholtz Machine uses a fixed number of hidden units. Thus the Bayesian model was able to learn a more efficient 4-feature representation than the network-based model, enabling more efficient communication.

In the HIERARCHY environment, convergence was fast and reliable under all 3 models, using a minimal 4-bit signal. The Helmholtz model is able to learn the more efficient representation in this environment, using one hidden unit to code for the category, and 6 more for each image within a category. This reduced convergence time by up to half. The Bayesian model, however, learns a less efficient representation, identifying 1 binary feature for each of the 12 images in the environment, and does not significantly improve convergence speed.

## Discussion and Future Work

The results from the previous section demonstrate how a perception mechanism can be incorporated into a signaling game model, and shed some light on the first interaction we raised in the introduction. In particular, we saw that a perception pre-processing step can enable faster and more robust cooperative learning from high-dimensional sensory inputs. We also saw that certain perception models can learn more efficient representations in certain environments. In this section, we discuss how the existing model can be extended to address the other types of interactions we wish to study.

## Other Roles of Perception

In the experiments presented here, perception was used strictly as a pre-processing step, but in order to better understand how language-learning can affect perception, we must allow the perception mechanisms to be trained in parallel with the signaling game. While the REINFORCE rule we present here does not scale well to very deep models with multiple hidden layers, recent advances in Deep Reinforcement Learning and Deep Q-learning allow us to scale the basic architecture up to very large tasks. Additionally, we may be interested in fixing the behavior of one agent, so that the non-fixed player learns the language of the fixed player. This would allow us to observe any influence that the fixed player's language has on the non-fixed player's learned representations.

## Perceptual Similarity in Communication

The model we present here already has most of what we need to address the third type of interaction, relating to perceptual similarity across agents. That is, we already represent the evolution of both the external language and the internal representations, so all we need is a means of quantifying "perceptual similarity" across multiple agents. To this end, we can use cross-systems analysis techniques like Representational Similarity Analysis (Kriegeskorte, N., Mur, M., \& Bandettini, P. A. 2008), a method for quantifying "representational similarity" between two different representation systems, regardless of the underlying topologies of the systems themselves. This would allow us to study inter-agent learning performance as a function of the similarity between their internal representations, and perhaps identify a "communicability threshold" of perceptual similarity below which no communication is possible.

## References

Bengio, Y., Lonard, N., \& Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432.
Dayan, P., Hinton, G. E., Neal, R. M., \& Zemel, R. S. (1995). The helmholtz machine. Neural computation, 7(5), 889-904.
Franke, M. (2015). The evolution of compositionality and proto-syntax in signaling games. Journal of Logic, Language and Information, forthcoming.
Griffiths, T. L., \& Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. In NIPS (Vol. 18, pp. 475-482).
Gu, S., Levine, S., Sutskever, I., \& Mnih, A. (2015). MuProp: Unbiased backpropagation for stochastic neural networks. arXiv preprint arXiv:1511.05176.
Huttegger, S. M., Skyrms, B., Smead, R., \& Zollman, K. J. (2010). Evolutionary dynamics of Lewis signaling games: signaling systems vs. partial pooling. Synthese, 172(1), 177-191.
Jager, G. (2007). The evolution of convex categories. Linguistics and Philosophy, 30(5), 551-564.
Kriegeskorte, N., Mur, M., \& Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2, 4.
O’Connor, C. (2014). Evolving perceptual categories. Philosophy of Science, 81(5), 840-851.
Skyrms, B. (2010). Signals: Evolution, learning, and information. Oxford University Press.
Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4), 229-256.

# Leaping across the mental canyon: Analogical retrieval across disparate task domains 

Shir Dekel (sdek7598@uni.sydney.edu.au)<br>University of Sydney, School of Psychology, Brennan MacCallum Building (A18)<br>Camperdown, NSW 2006, Australia<br>Bruce D. Burns (bruce.burns@sydney.edu.au)<br>University of Sydney, School of Psychology, Brennan MacCallum Building (A18)<br>Camperdown, NSW 2006, Australia

Micah B. Goldwater (micah.goldwater@sydney.edu.au)
University of Sydney, School of Psychology, Brennan MacCallum Building (A18)
Camperdown, NSW 2006, Australia


#### Abstract

The present study provides evidence for far analogical retrieval, i.e., analogical retrieval across disparate task domains, as a result of analogical comparison. Participants read source stories, which were then retrieved after a filled delay through abstract letter-string cues that matched the relational form of key parts of stories. They then generated responses to an ambiguous letter-string analogy problem. Evidence was found for far analogical retrieval of higherorder relations because 1. comparison of letter-string analogies cued source stories specific to the relations showed in the letter-strings, and then 2. those same relations formed the basis for how subjects solved novel letter-string problems. The experiment offers support for the schema induction account of analogical retrieval, and suggests that people are more sensitive to relational structures than was previously thought.


Keywords: analogy; memory; reasoning; analogical retrieval; letter-string analogies

## Introduction

Analogical retrieval leads to many important insights in science and design. It appears that these insights often emerge as a result of analogical retrieval from vastly different domains to one's present situation. Despite this, studies of schema induction in analogical thinking have primarily focused on the relatively narrow domain of semantic differences between stimuli, and cross-domain effects have been rare. The schema induction account of analogical retrieval (Gentner et al., 2009) suggests that cross-domain analogical retrieval should be facilitated by a comparison of analogues, by promoting a structural alignment of common relations. A more general principle, or schema, is then assumed to be available as a memory probe for future mapping in analogous situations. Evidence for the schema induction account was established through a comparison of target analogues, i.e., late analogical abstraction. This effect has been demonstrated in both studies of cued-retrieval (Gentner et al., 2009) and problem solving (Kurtz \& Loewenstein, 2007).

The effect was demonstrated in the domain of negotiation with a controlled memory set (Experiment 4; Gentner et al.,
2009). Undergraduates read seven negotiation scenarios, with only one containing the target negotiation principle. After 30 minutes of a filled delay, half of the participants were given two example cases of a certain negotiation principle and were explicitly asked to compare them, noting the key parallels. The other half were given the two cases to read separately. They were subsequently asked to recall a source case that best matched the two target comparison cases. Participants that explicitly compared target cases were significantly more likely to retrieve the source cases than participants that read the target cases separately. It appeared the explicit comparison made the abstract schema directly available as a retrieval cue to the original story.

The main limitation to this literature is that most studies have only varied what Barnett and Ceci (2002) call the knowledge domain of the analogues, despite the possibility for retrieval across different task domains. Inherent to the schema induction account is the assumption that analogical comparison highlights relational structure regardless of surface features. That is, it allows a cross-domain mapping. However, as information is often relevant across different tasks, it is important to understand whether and how crosstask retrieval can occur. This is an important hole to fill in the literature.

In addition to investigating whether we can elicit remindings across task domains, we can further investigate whether the cross-task commonalities that can serve as the basis for analogical remindings are limited to specific levels of abstraction. That is, the present study investigated retrieval rooted in common surface features, first-order relations, and higher-order relations (respectively). Including these different levels of abstraction as controls for each other in the analyses ensured a more adequate test of analogical retrieval. That is, a retrieval based in surface features across task domains without relational controls is not very surprising, given people's sensitivity for retrieval of surface features (Gentner, Rattermann, \& Forbus, 1993). Further, retrieval of relational content across task domains is more valid when controlled for by the possibility of a surface feature retrieval. That is, it is more likely that a source story was retrieved because of a relational match to
the letter-string cue when a surface feature alternative was also possible.

The schema induction account assumes that analogical comparison abstracts one's stimulus representation, i.e., relational commonalities are highlighted, and mismatches of features are ignored. To further test this account it is important to examine the representation formed from the comparison independently from the test of retrieval. Gentner et al. (Experiment 1; 2009) used a post-retrieval transfer task to confirm participant representations, but scoring was based on how well participants' descriptions matched a target schema, not directly analyzing the schema used by the participants. Hofstadter's (1995) letter-string proportional analogies could be used as a clearer way of determining the type and level of representation a person currently has. For instance, if asked "Suppose the letter-string abc were changed to abd; how would you change the letter-string mrrjjj in 'the same way'?" (Hofstadter, p. 238), one answer could be mrrkkk, if succession relation is used because $k$ follows $j$, just as $d$ follows $c$. A higher-order response represents $a b c$ as 1-2-3 and $a b d$ as 1-2-4, as per their order in the alphabet. The quantity of different letters in the string $m r r i j j$ can also be represented numerically as 1-2-3. This higher-order relational mapping leads to the inference that the fourth term in the analogy should be a quantity successor of $m r r i j j$ that can be numerically represented as 1 -2-4, i.e., mrrijjj.

## Present study

The present study extends the late analogical abstraction effect (Gentner et al., 2009) to investigate far analogical retrieval, i.e., retrieval across disparate task domains. Letterstring analogies were used as cues to retrieve story narratives (see Figure 1). This will be referred to as far analogical retrieval, as analogues are retrieved across task domains, a significantly more disparate - and conceptually far - retrieval than in previous studies. Each comparison cue had one analogous initial source story that matches the underlying schema. A pilot study (Dekel, 2016) showed that these source stories could be retrieved by analogous stories, replicating the late analogical abstraction effect (Gentner et al.). In the present study, correct source story retrieval after the letter-string comparison provided evidence for far analogical retrieval. A subsequent transfer task with a novel letter-string analogy determined participant schema representation for each level of abstraction (surface features, first-order relations, and higher-order relations).

The main hypothesis was that participants that compare two target letter-string analogies that share a particular schema would retrieve the source story that emphasizes the same schema, significantly more than participants comparing target stories that do not share this schema. For the transfer task, it was hypothesized that participants will respond to the transfer task according to their schema condition (see Figure 4).


Figure 1: Comparison of simplified designs in Gentner et al. (2009) and the present study. While participants in Gentner et al. retrieved source stories from a comparison of story cues, participants in the present study retrieved source stories from a comparison of letter-string cues.

## Method

## Participants

One hundred and eighty-one first-year undergraduates from the University of Sydney subject pool were recruited online, and were given course credit for their participation. Participants were randomly allocated to one of three schema conditions: surface schema, first-order relational schema, higher-order relational schema. One participant did not complete the first filler task and another did not complete the analogy example page (both due to computer error), so their data was excluded from the analysis of retrieval rates.

## Materials

The experiment was completed online and all materials were webpages coded with HTML and JavaScript.

Source stories The three source stories, shown in Table 1 , were designed to differ in semantic content, but be equivalent in structure and length. Each story presented an initial conflict, and a subsequent resolution. Critically, the resolution of each story also provided the information that made up the target schema for that story, which would then either match or mismatch with the later letter-string analogies. The first story schema is simply changing an E to an $F$. It is considered a surface story because its similarity to the later cues is based on an identical change. The second story schema is succession (of Valerie by Sylvia), considered to be a first-order relational story because it is related to the later cues by virtue of one relation (succession) and no surface features. The third story schema is the correspondence of quantity to an order (number of staff to a day's order in the week), considered to be a higher-order story because it relies on a mapping of firstorder relations. That is, numerical representation connects the first-order relational structure of two types of succession: ordinal succession, as in the order of days in the week, and quantity succession, as in the number of staff allocated.

Table 1: Source Stories and Explicit Principles.

| Schema condition | Story text | Explanation text |
| :---: | :---: | :---: |
| Surface | John is an owner of a small-town computer company and wanted to advertise his company to the town. He printed out some flyers with large font size to put up. However, there was a typographical error in the flyers, with the title printing out as 'Elash Computers' instead of 'Flash Computers', which John knew would confuse potential customers if put up around town. As such, he had to rewrite the company name for the posters, changing the ' $E$ ' to an ' $F$ ', and printing them again. There was only a typo in the word 'Elash', so only the letter ' $E$ ' was changed from the letter ' $E$ ', to the letter ' $F$ ', correcting the word 'Elash' to 'Flash'. | Both pairs rely on the same rule: Change E to F. |
| First-order relation | Jerome is an advisor to the King of a large nation and wanted to confirm the successor to the throne. He thought of Valerie, who was the king's eldest daughter. However, the advisor found out that despite being the next in line to the throne, Valerie had run away to a mountain town because she did not want to take on the responsibilities associated with being a Queen. As such, he worked out that Sylvia should be the next in line to the throne as she is the second-oldest sibling. The order of succession in the kingdom is found by birth order, so if the first born child is not able to uphold the throne, then the second born is next in line. | Both pairs rely on the same rule: Succession. For example: <br> Triangle changes to square because of the number of sides, and G to H because of alphabetic order. |
| Higher-order relation | Julia is a manager at a local information centre and wanted to staff her centre efficiently. She usually has about three people working every day. However, the number of visitors to the centre increases consistently each day, with almost no visitors on Mondays and peak number of visitors on Sundays, so most days the centre is either overstaffed, or understaffed. As such, she decided that she will roster on an amount of staff that corresponds with the order of that day in the week. The centre will have one staff member on Monday, being the first day of the working week, two on Tuesday, and so on, with seven people working on Sundays. | Both pairs rely on the same rule: Order corresponds to quantity. For example: E (fifth in the alphabet) changing to F (sixth in the alphabet) = five symbols (letters or shapes) changing to six symbols. |

Letter-string analogue comparison Participants received one of three pairs of proportional letter-string analogies to compare, as shown in Figure 2. In the figure, all three pairs are presented together to facilitate comparison of the differences between each pair. The first pair was designed to induce the surface schema, the second the firstorder schema, and the third the higher-order schema. The same basic structure and symbols (letters and shapes) were used for all three of the comparisons. Below this comparison, participants read a short explanation of the target principle, shown in Table 1 and then completed a short test of the principle.

Procedure. The experiment was run as an online study through a series of webpages. Participants read three oneparagraph narratives and typed how each story was resolved. They then completed two minutes of an unusual uses task (Diamond, 2013) and a page designed to inform and train participants about the structure and function of proportional analogies.

Participants then completed the comparison task, as per their schema condition, and on the subsequent page were asked to retrieve the source story that matched the comparison they just did. They then completed two minutes of a new unusual uses task. Participants then responded to the letter-string proportional analogy Suppose that the letterstring $A B C$ was changed to $A B D$; how would you change the letter-string C S SNNN in the same way? Following this, participants rated some prototypical responses to the analogy, and then a subsequent follow-up page asking participants to indicate the extent to which they used any of the letter-strings or stories when generating the letter-string analogies. Figure 3 shows this procedure.


Figure 2: The three pairs of comparison cues. The surface comparison (a) makes it apparent that the critical change is that from $E$ to $F$, regardless of position in the string or presence of shapes. In the first-order relational schema comparison (b) the increase in the letters' alphabetic order and shape's number of sides, expresses the concept of succession. The higher-order relational schema comparison (c) connects the alphabetic succession of the initial strings to the ordinal succession of the latter shape or letter strings. The higher-order relationship numerical representation connects these two forms of succession.

## Results

## Far analogical retrieval

A chi-square test was conducted for the cross-tabulation of retrieval by schema condition (Table 2). The retrieval variable had four levels: null retrieval, surface story, firstorder story, and higher-order story. Schema condition had three levels: surface schema, first-order schema condition, and higher-order schema condition. The overall effect was significant, $\chi^{2}(6, N=179)=46.55, p<.001$, suggesting an association between people's schema condition and story retrieval rates. To investigate the specific effects, the conditions were collapsed into 2 x 2 tables for each predicted effect. For each effect, schema condition was recoded into a dichotomous variable of those in the target schema condition and those that are not. Retrieval condition
was recoded into a dichotomous variable of those that retrieved the target schema story and those that did not.


Figure 3: Experimental procedure.
Participants in the surface schema condition retrieved the surface schema story ( $86.4 \%$ ) significantly more than those not in the surface schema condition $(40 \%), \chi^{2}(1, N=179)=$ $34.51, p<.001$. Those in the first-order schema condition retrieved the first-order schema story ( $31.1 \%$ ) significantly more than those not in the first-order schema condition $(13.6 \%), \chi^{2}(1, N=179)=7.91, p<.001$. Those in the higher-order schema condition retrieved the higher-order schema story ( $37.3 \%$ ) significantly more than those not in the higher-order schema condition (10.8\%), $\chi^{2}(1, N=179)$ $=17.6, p<.001$. The main hypothesis was thus supported by these results as a comparison of letter-string analogies facilitated correct retrieval of source stories with the same underlying schema.

Table 2: Frequency of Story Retrievals by Schema Condition.

|  | Schema Condition |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| Story <br> retrieval | Surface | First- <br> order | Higher- <br> order | Total |
| Surface | 51 | 28 | 20 | 99 |
| First-order | 5 | 19 | 11 | 35 |
| Higher-order | 1 | 12 | 22 | 35 |
| Null | 2 | 2 | 6 | 10 |
| Total | 59 | 61 | 59 | 179 |

## Transfer response

Responses were coded through the schema they presumably expressed. As in Burns (1996), letter-string responses generated by two or fewer participants were collapsed into the category Other. Figure 4 shows the structural hierarchy of the three prototypical responses to the letter-string analogy task. Participant response of DSSNNN was considered a Surface response, since it only takes into account the $C$ changing into $D$. Participant responses of CSSOOO, CSSNNO, CSTNNO were collapsed into category First-order, since they all use the first-order principle of succession. Participant response of CSSNNNN was considered Higher-order, since it takes into the higherorder correspondence of numerical representation. Table 3 shows the frequencies of these responses for each schema condition.

A chi-square test was conducted of letter-string response by schema condition. Letter-string response had four levels: surface, first-order, higher-order, and other. Schema condition had three levels: surface schema, first-order schema condition, and higher-order schema condition. The overall effect was significant, $\chi^{2}(6, N=176.04)=176.04, p$ $<.001$, suggesting an association between people's schema condition and letter-string response rates. To probe the specific effects, the conditions were collapsed into $2 \times 2$ tables for each predicted effect. For each effect, schema condition was recoded into a dichotomous variable of those in the target schema condition and those that are not. Letterstring response condition was recoded into a dichotomous variable of those in that generated the target letter-string response and those that did not.

The surface schema response was generated significantly more by those in the surface schema condition (70\%) than those not in the surface schema condition $(0.03 \%), \chi^{2}(1, N$ $=181)=94.12, p<.001$. First-order schema responses were generated significantly more by those in the first-order schema condition ( $82 \%$ ) than those not in the first-order schema condition (26.7\%), $\chi^{2}(1, N=181)=49.91, p<$ .001. The higher-order schema solution was generated significantly more by those in the higher-order schema condition (60\%) than those not in the higher-order schema condition $(0.006 \%), \chi^{2}(1, N=181)=108.16, p<.001$. As per the initial hypothesis, the selective generation of letter-
string responses were congruent with one's schema condition.


Figure 4: A representation of three responses to the analogy $\mathrm{ABC}: A B D:: \mathrm{CSSNNN}:$ ?. The surface response (a) considers the change from $C$ to $D$, per se, so merely changes the $C$ in CSSNNN to a D. The first-order relational response (b)
considers the change from C to D as one of ordinal succession, as per their order in the alphabet. Since $C$ is the last term in the string, NNN, as the last string of CSSNNN, is also changed to its successor in the alphabet: OOO. The higher-order response (c), on the other hand, considers the entire string and each letter's position in the alphabet, representing ABC as $1-2-3$ and ABD as 1-2-4. The change is still an ordinal succession, as C and D are successors in the alphabet, but the letters have been represented numerically. The quantity of different letters in the string CSSNNN can also be represented numerically as 1-2-3. This numerical representation allows this first-order relation to map to the ordinal succession relation of ABC. This higherorder relational mapping leads to the inference that the fourth term in the analogy should be a quantity successor of CSSNNN that can be numerically represented as 1-2-4, i.e.,

CSSNNNN.

Table 3: Frequency of Letter-string Responses by Schema Condition.

|  | Schema Condition |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| Letter-string <br> response | Surface | First- <br> order | Higher- <br> order | Total |
| Surface | 42 | 4 | 0 | 48 |
| First-order | 11 | 50 | 21 | 77 |
| Higher-order | 0 | 1 | 36 | 35 |
| Other | 7 | 6 | 3 | 21 |
| Total | 60 | 61 | 60 | 181 |

## Discussion

Successful cross-domain analogical retrieval is rare. Despite this, the results of the present study provide evidence that schema induction can facilitate far analogical retrieval, i.e., analogical retrieval across disparate task domains. The effect was found for surface, first-order relational, and higher-order relational schemas. As well as providing support for a schema induction account of analogical retrieval, the results of the present study also address the three limitations in this literature were identified above. First, there is now evidence that analogue comparison can facilitate analogical retrieval across stimuli that share no surface features, except for the presence of alphabetic characters. Second, while prior research usually neglects to consider different levels of abstraction in analogical retrieval, the present study investigated retrieval of surface features, first-order relations, and higher-order relations. Third, the present study used a transfer task to probe the way participants were representing their schema.

The main limitation of the present study is that the apparent retrieval effects, might actually be mapping effects. The combination of a relatively small number of source stories and short delay might mean that participants were considering each source story as a potential match to their comparison cue and then actively deciding on the best perceived mapping. Future replications of the present study should therefore include a larger set of source stories and a longer delay between source story encoding and the retrieval phase. Further, it is not clear what exact role the explicit principle played in cuing the source stories. In general, it seems that explicit principles are not sufficient to induce a schema, but do seem to facilitate induction. Thus, future replications should systematically manipulate the explicit principle and its inclusion with comparison to determine its role as a retrieval cue for the far analogical retrieval.

Relational priming is sometimes used to explain analogical retrieval effects (Holyoak, 2012). It is unlikely to explain all of the present retrieval results because most relational priming effects are demonstrated using individual word pairs of highly familiar relations. There is little evidence to suggest that a higher-order relation can be primed in the same way, and the participants in our study had little to no previous experience with the specific
relations presented to them. Further, pilot data (Dekel, 2016) shows that changing the explicit principle in the higherorder condition to a more specific form does not significantly impact retrieval rates. This suggests a lesser role of explicit principle wording in any potential priming.

## References

Barnett, S. M., \& Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. Psychological bulletin, 128(4), 612-637.
Burns, B. D. (1996). Meta-analogical transfer: Transfer between episodes of analogical reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(4), 1032.
Dekel, S. (2016). Leaping across the mental canyon: Analogical retrieval across disparate task domains. The University of Sydney.
Diamond, A. (2013). Executive Functions. Annual Review of Psychology, 64, 135-168.
Holyoak, K. J. (2012). Analogy and Relational Reasoning. In K. J. Holyoak \& R. G. Morrison (Eds.), The Oxford Handbook of Thinking and Reasoning (pp. 234-259). New York: Oxford University Press.
Diamond, A. (2013). Executive Functions. Annual review of psychology, 64, 135-168.
Gentner, D., Loewenstein, J., Thompson, L., \& Forbus, K. D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. Cognitive Science, 33(8), 1343-1382.
Gentner, D., Rattermann, M. J., \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive psychology, 25(4), 524-575.
Hofstadter, D. R. (1995). Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought. New York: Basic books.
Kurtz, K. J., \& Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. Memory \& Cognition, 35(2), 334-341.

# Conversational topic connectedness predicted by Simplicity Theory 

Jean-Louis Dessalles (dessalles@telecom-paristech.fr)<br>LTCI, Telecom ParisTech, Université Paris-Saclay Paris, France


#### Abstract

People avoid changing subject abruptly during conversation. There are reasons to think that this constraint is more than a social convention and is deeply rooted in our cognition. We show here that the phenomenon of topic connectedness is an expected consequence of the maximization of unexpectedness and that it is predicted by Simplicity Theory.


Keywords: Conversation, topic change, simplicity, unexpectedness, interestingness.

## Introduction

A few decades ago, attempts were made to understand how and why conversational topics are almost systematically connected to each other, while abrupt topic shifts are avoided or even socially repressed. Jerry Hobbs (1990), for instance, describes several connection patterns that topics must respect to come next to each other in conversation and he wonders whether these mechanisms are due to cognitive constraints or are mere social conventions. The constraint of topic connectedness is so strong that abrupt topic change is classically considered characteristic of pathological conversation. The inability to respect topic connectedness during conversation has been for instance described in autism (Hale \& Tager-Flusberg, 2005; Volden et al., 1997) and in schizophrenia (Harrow et al., 1983). Even in these conditions, it is not clear whether patients merely ignore the constraint or still respect it covertly. The latter possibility would result from the patients' inability to realize that some elements needed for connecting topics lie only in their mind and have not been made public (Harrow et al., 1983).

Though authors had an intuition about what a conversational topic is and about how an utterance can be appropriate, the lack of proper definitions made the problem of topic connectedness difficult to address formally. When commenting on his maxim 'be relevant', Grice wonders

> "what different kinds and focuses of relevance there may be, how these shift in the course of a talk exchange, how to allow for the fact that subjects of conversation are legitimately changed, and so on. I find the treatment of such questions exceedingly difficult" (Grice, 1975).

Conversation has often been described either at the surface level, with notions like repairs and adjacency pairs, or at the sociological level, with notions like involvement, face preservation or gender talk. The present study adopts a rather different perspective, a cognitive one. The point is to show that the problem of topic connectedness can be derived as a natural consequence of conversational mechanisms operating at the cognitive level.

The remainder of the text starts by making a distinction between two fundamentally different conversational modes:
narratives and argumentative discussions. The way narratives connect to each other is particularly intriguing. We will observe patterns of topic connection using data taken from a corpus of spontaneous conversations. I will then introduce the Simplicity Theory (ST) framework and show how it can be used to explain the topic connectedness phenomenon.

## Conversational Topics

For long, it was hardly possible to study spontaneous conversations by making any hypotheses beyond what could be objectively observed. No assumptions were made about the participants' beliefs or desires. As a consequence, the existence of two sharply contrasted conversation modes, narratives and argumentative discussions, was considered unimportant. Though these two modes may be observed in pure form during conversations, they are sometimes intertwined, making the distinction less apparent at the surface level. If one takes a cognitive perspective, however, the distinction cannot be overlooked. The first explicit description of the narrative/argumentation dichotomy was apparently given by Jerome Bruner (1986), though many authors (e.g. Sacks, 1992; Tannen, 1984; Eggins \& Slade, 1997) implicitly distinguished between stories and other forms of verbal interaction.

## Conversational stories

People tell stories during conversations, i.e. they mention past events that are supposed to have occurred. Despite early studies (Sacks, 1992; Labov \& Fanshel, 1977:105; Polanyi, 1979; Tannen, 1984), the importance of the phenomenon has rarely been acknowledged until more recently (Norrick, 2000), as it only occurs among people who are already acquainted and is absent from most corpora recorded in the lab. Typical conversational narratives are easy to recognize. They are most often about past events (the past tense is used in English) for which the four w's (when, where, what, who) get instantiated. Consider the following conversation from my corpus. It involves two French women who had some trouble buying butter ('beurre' in French) during their stay in Spain [translated from French].

D:[...] she was with her cousin in Spain. And so... they wanted to buy butter. And then [laugh] her cousin said to her, she didn't speak one word of Spanish, but she said to her: "I can speak Italian; Italian and Spanish, that's the same", and then
O : Oh là là! Oh là là!
D: So she enters the store, and she says 'Burro'. And then [laugh] then everyone was staring at her, and so 'burro' means 'donkey'.
$\mathrm{O}:$ Oh! [laugh]. It means 'donkey'! She wanted to say 'Butter'! Burro. [laugh] It plays tricks, isn't it?

This story is definitely reporting a situated event: the 'when', 'where', 'what', 'who' are supposed to be constants and not variables. However, being situated is not sufficient for an event to be worth telling.

> "if you come home and report what the grass looked like along the way, that there were four noticeable shades of green some of which appeared just yesterday because of the rain, then there may well some tightening up on the part a your recipient. And if you were to do it routinely, then people might figure that there's something odd about you" (Sacks 1992:219)
> "We would intuitively reject such introductions as 'Let me tell you something ordinary that happened yesterday...' A narrative that is in fact judged to be ordinary may be rejected after it is told by expressions equivalent to 'So what!"" (Labov \& Fanshel 1977:105)

The missing ingredient that is required to turn an event into a story has been informally described by several authors in similar terms: narratable events should be 'problematic' (Ochs et al., 1992), 'different from ordinary experience' (Labov \& Fanshel 1977:105), 'unexpected, deviant, extraordinary, or unpredictable' (van Dijk, 1993), 'abnormal' (Schank, 1979), 'odd or unexpected',' rare', 'impossible or unheard of', be 'the violation of a norm' (Polanyi, 1979), 'depart from expectations', be a low probability event (Agar, 2005). The 'burro' story above definitely matches many of these criteria. We will subsume all these properties by saying that an event must be unexpected to be storyworthy. This notion will be refined below.

Languages offer means to emphasize unexpectedness, ranging from adjectives like odd, funny to specific markers like the wo particle in Cantonese (Luke 1990). Unexpectedness is the key element that controls storyworthiness. Emotional events are of course interesting, but happy or tragic situations do not arouse emotion unless they are unexpected (Saillenfest \& Dessalles, 2012). We will see that the unexpectedness requirement is also the very reason why narrative topics are connected.

## Argumentative discussion

A significant amount of conversational time is devoted to an activity that radically differs from story telling, namely argumentative discussion. The argumentative mode seems to be the prevalent one, at least in my corpus of family conversations. During an argumentative discussion, people deal with problems, i.e. incompatibilities between beliefs and/or desires ${ }^{1}$. The following conversation deals with TV and radio power consumption (translated from French).

P - When you put it into standby mode using the remote control with the small red dot on.
L-mmm
P- Does the TV remain switched on?
L-Yes.

[^318]P- So it is to be avoided,
L-No.
P- leaving it that way permanently?
L-No. People would say yes, but, it is quite irritating; you don't take advantage of having a remote control, and, uh, I mean, you will save six month or one year on the TV's life expectancy. Pff.
D-[to P] Not at all. And anyway, it sets..., it damages tubes a lot to set them on and off.
L-No but anyway, the tube is switched off when you put in standby mode.
D-I don't think so.
L-No, one should not compare consumer electronics and professional tubes.
D-Because, still, when you [really] switch it on, you can hear quite a discharge.
L-Yes, well, the tube warms up. When you put it..., No, no, the tube is switched off, but because it is consumer electronics, uh, otherwise you burn the tube, if it is consumer electronics.
O-A totally unrelated issue: when I put, I leave my radio plugged in, knowing that it is also a cassette recorder,
L-Yes.
O-I can hear something. [...] Should I switch... should I pull off the plug each time or it cannot damage the engine.

The problem here is the apparent incompatibility between the standby mode and the wish to keep the TV undamaged (the last utterances show the transition to the next topic). Discussions function as consistency maintaining devices: participants point to an incompatibility (like standby mode vs. no damage) or try to resolve a previously mentioned incompatibility.

## What counts as a topic?

Based on the argumentative/narrative dichotomy, the notion of topic can receive a proper definition.

- The topic of a narrative is the unexpected event it refers to.
- The topic of an argumentative discussion is the logical inconsistency that motivates it.

One could be tempted to consider that stories and argumentation are just two extremes in a continuum. An utterance like (talking about a toddler) "She is going down the stairs by herself!" might seem hard to classify either as narrative or as argumentative, as the event is both unexpected and potentially problematic (the child may fall down). And how would this injunction to a child: "Don't touch it!" or an exclamation like "Oh, that's wonderful!" be classified? There are reasons, though, to stick to the narrative $v s$. argumentation dichotomy. The most important one comes from the conviction that human conversational behaviour cannot be based on a wide gamut of unrelated cognitive devices. The narrative competence, as described by Simplicity Theory (see below), and the argumentative competence, as described in a minimal way in (Dessalles, 2016) can account for the relevance of most conversational utterances.

The crucial element that helps decide in which mode we are is negation. If the speaker has the negation of the state of affairs in mind (the child [should] not go down the stairs by herself; the child should not touch the object), then the
move is argumentative; if the event is regarded as unexpected, it is a narrative move ${ }^{2}$.

## Topic shift and topic drift

Respecting transitions between conversational topics seems almost as important as making appropriate moves.

> Not only are there socially sanctioned rules for appropriate topics of conversation, but also, in the course of a conversation, it is impolite to make an abrupt change of topic even to another socially sanctioned topic. To make a change of topic one must usually create some link to the previous topic, or one must drift to another topic in a stepwise fashion. (Shiller 1995:184)

Hobbs (1990) identified several patterns through which successive topics connect to each other. One of them is 'semantic parallelism'. Two topics may share a common predicate $p$ applied to different (but similar) arguments $a_{1}$ and $a_{2}: p\left(a_{1}\right)$ and $p\left(a_{2}\right)$. For instance, two stories about an accident share this common feature, though the different roles (driver, victim...) would differ. Conversely, two stories may be connected by an argument instead of by the predicate: $p_{1}(a)$ and $p_{2}(a)$, e.g. if successive stories involve a same protagonist.

Hobbs then considers connections that apply to argumentative discussions. It is often artificial to talk about topic change in argumentative discussion. Hobbs prefers to talk about topic drift. Since argumentative discussions go around problems (i.e. logical inconsistencies between beliefs and/or desires), a solution to a previous problematic issue may be regarded itself as problematic. This may lead to topic drift: People stack problematic topics on top on each other, and may or may not revert to a previous one. When there is a 'main issue', the topic can be clearly identified. For instance, in my main corpus of French conversation, one discussion about preparing a meal that would suit North American visitors consists of 255 utterances and lasts for twenty minutes. In many cases, however, discussions drift with no intent to reconsider the initial issue.

## Observing topic shifts

Stories tend to cling to each other, forming what Deborah Tannen (1984:100) calls story rounds. During a conversation among friends, she counted 48 narratives, 21 of which where told in five rounds: two stories about sex differences for language learning, five stories about adopted children, five about summer camps, five about strange accidents and four about child discovery of sexuality. The 'burro' story (see above) is part of a story round as well. This story round in detailed in Table 1. Transitions between topics (here, association or analogy) are shown.

As mentioned above, argumentative discussions tend to drift through logical connections. However, in some cases, discussions can be connected to each other in much the

[^319]same way as narratives. Table 2 shows an example of what we may call argumentation rounds. The above discussion about the TV in standby mode is included in this round. Note that a story is embedded in this sequence, as it is used as an argument (independently from its unexpected character that makes it a story in its own right).

Table 1: Example of story round


It is interesting to observe the different categories of topic transition, as they are described for instance in (Hobbs, 1990), at work in real conversations. The point is not to make precise quantitative assessments, as we expect significant variance depending on the kind of corpus we observe (number and age of participants, degree of acquaintance, situation and so on). Rather, we would like to get just some rough idea of the relative importance of the different forms of topic connections.
The corpus chosen here is a set of conversations recorded during family gatherings during three years. The total duration is 17 h 50 min . Participants are mostly the same
across recordings. In order to avoid selection bias, a sampling method has been used. Two-minute long slices were selected around 139 randomly defined time locations. 8 of these slices were ignored as no one was talking at the central time. 18 more were discarded as unintelligible. Reasons for unintelligibility are multiple and include noises, simultaneous loud conversations and child screams. The 113 remaining excerpts can be classified as shown in Table 3.

Table 2: Example of argumentation round


Several comments have to be made about these results. The main observation is that very few topics are introduced out of the blue in this kind of corpus. This may suggest that the number of abrupt topic change observed in other corpora (e.g. Nordenstam, 1992) might be overestimated, either because some crucial connecting piece of knowledge might be unknown to the external observers, or because conversations elicited in the lab might lack the spontaneous aspect of normal conversation.

Almost any connection seems possible between narratives (the term 'signal' refers to mentions of unexpected events that are 'here and now'3 ${ }^{3}$. The most represented topic connection in this corpus is analogy, which means that these narratives occurred in typical story rounds. The connection might be less tight, as when only one element is common with the parent topic.

[^320]Another observation is that the most basic pattern: problem-solution or solution-problem, is the majority argumentative connection but is not the only one. Problems may also refer to the current situation (e.g. I am missing a fork) or to an element from a preceding narrative. Surprisingly, a problem may also refer to another problem. This occurs in a problem-solution-problem-problem pattern, in which the last problem suggests that not adopting the solution is a problem as well (see (Dessalles, 2016) for a minimal model of relevant argument generation).

Table 3: Topic shift counts

|  | Topic type | Link with parent topic | \# / 113 |
| :---: | :---: | :---: | :---: |
|  | Problem | None | 1 |
|  | Problem | Situation | 12 |
|  | Problem | solution, refutation | 13 |
|  | Problem | Problem | 6 |
|  | Problem | Narrative | 3 |
|  | Solution | Problem | 33 |
|  | Signal | Situation | 14 |
|  | Narrative | None | 3 |
|  | Narrative | Situation | 2 |
|  | Narrative | Temporal | 1 |
|  | Narrative | Common item | 4 |
|  | Narrative | Close association | 3 |
|  | Narrative | Analogy | 13 |
|  | Narrative | Problem | 1 |
|  | Narrative | Explanation, refutation, solution | 4 |

## Explaining topic connectedness

In this section, we go beyond description and ask why conversation topics are so systematically connected. The suggestion will be that topic connectedness is the expected outcome of the quest for unexpectedness and that it is predicted by Simplicity Theory.

## Simplicity Theory

Simplicity Theory (ST) has been developed to explain event narratability. As discussed above, the core notion is unexpectedness: events must be unexpected to be tellable, and conversely unexpected events are systematically tellable. The notion of unexpectedness is not intrinsic to the event. It depends on the observer and on the current context. Previous attempts to define unexpectedness as 'low probability' failed to incorporate this necessary relation to the context.

ST is based on the notion of abnormal simplicity. Imagine that the numbers $1,2,3,4,5$ and 6 are drawn in the National Lottery. Tough the probability of this outcome is exactly the same as for any other draw, the news would be considerably more thrilling. Intuitively, the consecutive draw is interesting because it is abnormally simple. Simplicity is
obvious here as the sequence $1,2,3,4,5,6$ is compressible. The underlying theoretical notion is a cognitive (i.e. resource-bounded) version of Kolmogorov complexity ${ }^{4}$. Complexity here means 'minimal description length'. 1, 2, $3,4,5,6$ can be described using a much shorter code than a 'normal' draw like 3, 17, 27, 28, 33, 45. The consecutive draw can be described using the 'increment' operation, which is one of the simplest operations in the context of numbers, whereas the 'normal' draw cannot be 'compressed' down to a shorted description than itself.

Unexpectedness (or abnormal simplicity) $U$ results from a contrast between causal complexity $C_{w}$ (the circumstances or choice points that brought the event to happen) and description complexity C. Formally:

$$
\begin{equation*}
U=C_{w}-C . \tag{1}
\end{equation*}
$$

Though ST was initially developed to account for specific situations such as lottery draws or coincidences, we were surprised to find that the complexity drop between causality and description generalizes to all narratable situations (Dessalles, 2008). For instance, a fortuitous encounter is all the more narratable as the place of the encounter is complex (or hard to reach) and the encountered person is simple (a close acquaintance or a celebrity). The former parameter (location remoteness) controls the causal complexity, while the latter (minimal description of the person) controls the description complexity (see www.simplicitytheory.science for further examples). ST also accounts for problems underlying argumentative topics: the intensity of the problem corresponds to a high value of causal complexity $C_{w}$ (Dessalles, 2013).

Simplicity Theory accounts for many aspects of interestingness that are left otherwise unexplained. It explains why recent events are more interesting if they are closer in time and space, why 'round' anniversaries (after exactly 1,10 or 100 years) make past events worth talking about, why mishaps concerning celebrities might be as interesting for certain audiences as if they concerned own family, why people are fond of exceptions, norm-breaking behaviour and record-breaking performances, why collectors value items that are remarkable due to a simple feature (e.g. an inverted image on a stamp), and so on. All these predictions are derived from the equation $U=C_{w}-C$ (Dessalles, 2008; 2013; Saillenfest \& Dessalles, 2015; see www.simplicitytheory.science).

## Context and simplicity

Context plays a prominent role in the phenomenon of topic connectedness. ST not only defines this role, but also defines what the context is. In ST's framework, the context is defined as the set of properties that contribute to making the event unexpected. Formally, such properties can be written as:

$$
f\left(s, c_{1}, c_{2}, \ldots\right)
$$

[^321]where $s$ is the event, $f$ is a predicate and $c_{i}$ are constants. Note that a property may represent a conjunction of properties: $f\left(s, c_{1}, c_{2}, \ldots\right)=\wedge f_{i}\left(s, c_{i 1}, c_{i 2}, \ldots\right)$. For instance, a property of the event might be takes_place(s, Spain). The unexpectedness of $s$ is:
\[

$$
\begin{equation*}
U(s)=C_{w}\left(f\left(s, c_{1}, c_{2}, \ldots\right)\right)-C\left(f\left(s, c_{1}, c_{2}, \ldots\right)\right) \tag{2}
\end{equation*}
$$

\]

It is easy to define context based on (2):

$$
\begin{aligned}
& \qquad \text { Context }= \\
& \text { properties and constants involved in complexity drop. }
\end{aligned}
$$

Using complexity chain rule, we can write:

$$
\begin{equation*}
C\left(f\left(s, c_{1}, c_{2}, \ldots\right)\right) \leq C(f)+C\left(c_{1} \mid f\right)+\ldots+C\left(s \mid f, c_{1}, \ldots\right) \tag{3}
\end{equation*}
$$

Conditional complexity $C(x \mid y)$ means the minimal description length of $x$ when the description of $y$ is available. (3) generalizes easily to conjunctions $\wedge f_{i}$.

We see from (2) and (3) that when telling a story or pointing to a problematic situation, the mention of contextual elements such as $f$ or $c_{i}$ encroaches on unexpectedness, as it diminishes the gap between causal complexity and description complexity.

## Topic connectedness explained

Topic connectedness offers an opportunity for conversational narrators to save on the description side under the following hypothesis:

## Hypothesis: Elements of the previous context are available for free for further descriptions.

For instance, if $T_{-1}$ represents the preceding context and if $f$ and $c_{1}$ are part of it, then $C\left(f \mid T_{-1}\right)=0$ and $C\left(c_{1} \mid f, \mathrm{~T}_{-1}\right)=0$. They disappear from (3). Since $T_{-1}$ has no reason to have a causal effect on $s, C_{w}\left(s \mid T_{-1}\right)=C_{w}(s)$, and (2) becomes:

$$
\begin{equation*}
U\left(s \mid T_{-1}\right)=C_{w}(s)-C\left(s \mid f, c_{1}, T_{-1}\right) \tag{4}
\end{equation*}
$$

We can see that if $f$ and $c_{1}$ are part of the current context (i.e. contribute to the complexity drop in (1) and are thus relevant):

$$
\begin{equation*}
U\left(s \mid T_{-1}\right)>U(s) \tag{5}
\end{equation*}
$$

We can conclude that the situation is more unexpected after $T_{-1}$ than in the absence of any context, and that the presence of $f$ in $T_{-1}$ makes it (more) relevant. As a consequence, an event that would be impossible to introduce out of the blue may get enough unexpectedness to be worth telling.

For instance, in the 'burro' story, elements like 'trip', 'foreign country' or 'communication failure' are available for free from the previous story about attempts to ask for milk at breakfast in Italy (see Table 1). Moreover, 'butter' appears simple as 'milk' was previously mentioned, and 'Spain' would appear simpler than most other countries (seen from France) as 'Italy' was mentioned. The second story would have been less unexpectedness without these elements already available. Similarly, in the argumentation round of Table 2, the close analogy between the TV and the radio getting possibly damaged in standby mode makes the second discussion much easier to introduce. The analogy
spares the complexity of describing some elements of the second discussion (consumer electronics, standby mode, getting damaged), making the problem worth discussing about. In the excerpt, note that despite the close analogy, the speaker still feels the additional precaution 'A totally unrelated issue' (original: "une question tout à fait à côté") to be necessary when introducing her topic.

## Discussion

Hobbs (1990) wonders "to what extent topic drift [is] a matter of cognition and to what extent a matter of convention". The above development suggests that social conventions play hardly any role, beyond the mere control of the unexpectedness threshold (as a result, switching topic might be easier in relaxed or intimate situations).

Could it be that the reuse of contextual elements from one context to the next would just be due to some sort of cognitive laziness or inertia on the speaker's side? We can exclude this possibility: as we saw, topic connectedness is not only a fact, but also a requirement. A same topic that would be interesting during a story round or an argumentation round may appear inappropriate or even pathological when introduced out of the blue (Sacks, 1992). The present paper suggests that a cognitive determinism is involved.

When introducing a topic abruptly, bringing concepts and names into the new context adds to descriptive complexity and, as a result, diminishes unexpectedness, up to a point that the new topic is at risk of loosing all relevance. Already mentioned concepts or names are, by contrast, descriptively costless. The very existence of 'story rounds' or 'argumentation rounds' seems to be entirely due to the temptation of using elements of the previous context to enhance the unexpectedness of the next one. The demand for unexpectedness appears to be a universal property of human spontaneous conversation. Topic coherence through conversation seems to be a side-effect of this requirement.

## References

Agar, M. (2005). Telling it like you think it might be: Narrative, linguistic anthropology, and the complex organization. Emergence: Complexity \& Organization, 7 (3), 23-34.
Bruner, J. (1986). Actual minds, possible worlds. Cambridge, MA: Harvard University Press.
Dessalles, J.-L. (2008). Coincidences and the encounter problem. In B. C. Love, K. McRae \& V. M. Sloutsky (Eds.), 30th Annual Conf. of the Cognitive Science Society, 2134-2139. Austin, TX: Cognitive Science Soc..
Dessalles, J.-L. (2013). Algorithmic simplicity and relevance. In D. L. Dowe (Ed.), Algorithmic probability and friends - LNAI 7070, 119-130. Berlin, D: Springer.
Dessalles, J.-L. (2016). A Cognitive Approach to Relevant Argument Generation. In M. Baldoni, C. Baroglio, F. Bex, T. D. Bui et al. (Eds.), Principles and Practice of Multi-Agent Systems, LNAI 9935, 3-15. Springer.

Eggins, S. \& Slade, D. (1997). Analysing casual conversation. London: Equinox.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. L. Morgan (Eds.), Syntax and semantics, vol. III: Speech acts, 41-58. New York: Academic Press.
Hale, C. M. \& Tager-Flusberg, H. (2005). The relationship between discourse deficits and autism symptomatology. J. of Autism and Developmental Disorders, 35 (4), 519524.

Harrow, M., Lanin-Kettering, I. \& Prosen, M. (1983). Disordered thinking in schizophrenia: Intermingling and loss of set. Schizophrenia Bulletin, 9 (3).
Hobbs, J. R. (1990). Topic drift. In B. Dorval (Ed.), Conversational organization and its development, 3-22. Norwood, NJ: Ablex.
Labov, W. \& Fanshel, D. (1977). Therapeutic discourse. New York, NY: Academic Press.
Luke, K. K. (1990). Utterance particles in Cantonese conversation. John Benjamins Publishing.
Nordenstam, K. (1992). Male and female conversational style. Int. J. of the sociology of language, 94 (1), 75-98.
Norrick, N. R. (2000). Conversational narrative: storytelling in everyday talk. Amsterdam: John Benjamins Publishing Company.
Ochs, E., Taylor, C., Rudolph, D. \& Smith, R. (1992). Storytelling as a theory-building activity. Discourse processes, 15, 37-72.
Polanyi, L. (1979). So What's the point? Semiotica, 25 (3), 207-241.
Sacks, H. (1992). Lectures on conversation vol. 2. Oxford, UK: Blackwell.
Saillenfest, A. \& Dessalles, J.-L. (2012). Role of kolmogorov complexity on interest in moral dilemma stories. In N. Miyake, D. Peebles \& R. Cooper (Eds.), 34th Annual Conference of the Cognitive Science Society, 947-952. Austin, TX: Cognitive Science Society.
Saillenfest, A. \& Dessalles, J.-L. (2015). Some probability judgments may rely on complexity assessments. 37th Annual Conference of the Cognitive Science Society, 2069-2074. Austin, TX: Cognitive Science Society.
Schank, R. C. (1979). Interestingness: controlling inferences. Artificial Intelligence, 12 (3), 273-297.
Shiller, R. J. (1995). Conversation, information, and herd behavior. American Economic Review, 85 (2), 181-185.
Tannen, D. (1984). Conversational style - Analyzing talk among friends. Norwood: Ablex Publishing Corporation.
Volden, J., Mulcahy, R. \& Holdgrafer, G. (1997). Pragmatic language disorder and perspective taking in autistic speakers. Applied Psycholinguistics, 18 (2), 181198.
van Dijk, T. A. (1993). Stories and racism. In D. K. Mumby (Ed.), Narrative and social control: critical perspectives, 121-142. Newbury Park, CA: Sage Publications.

# The Cognitive Architecture of Recursion: Behavioral and fMRI Evidence from the Visual, Musical and Motor Domains 

Mauricio de Jesus Dias Martins (diasmarm@hu-berlin.de)<br>Berlin School of Mind and Brain, Humboldt Universität zu Berlin, Luisenstrasse 56<br>10117 Berlin, Germany<br>Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a 04103 Leipzig, Germany


#### Abstract

In this manuscript, we summarize the results of our research program aiming at describing the cognitive architecture underlying the representation of recursive hierarchical embedding. After conducting a series of behavioral and fMRI experiments in the visual, musical and motor domains, we found that, behaviorally, the acquisition of recursive rules seems supported by cognitive resources that are general across domains. However, when we test well-trained participants in the fMRI, their representation of recursion seems supported by activating schemas stored in (visual, musical and motor) domain-specific repositories. This suggests that the resources necessary to acquire recursive rules are different from those necessary to utilize these rules after extensive training.


Keywords: recursion; hierarchy; embedding; visual; motor; music

Recursion is a fascinating concept that has inspired researchers from many disciplines because of its associations with language, music and mathematics, which are uniquely available to the cognitive repertoire of humans (Hauser, Chomsky, \& Fitch, 2002).

The definition of recursion is much discussed, and the term is currently used with many different possible meanings (Fitch, 2010). In the original mathematical terminology, recursive functions are those that take their own output as input for the next iteration, such as the function generating the natural numbers:

$$
\begin{aligned}
& \mathrm{N}_{0}=1 \\
& \mathrm{~N}_{\mathrm{i}}=\mathrm{N}_{\mathrm{i}-1}+1, \text { for } \mathrm{i}>0
\end{aligned}
$$

One of the properties of these functions is the capacity to generate an infinite set of outputs.

Empirically, this property of infinity is impossible to verify (Lobina, 2011). However, there are other properties of recursion that make it interesting for empirical cognitive sciences. For instance, when we combine recursion with hierarchical embedding, we can generate complex hierarchies using simple rules (Martins, 2012). In fact, recursion has been discussed as a necessary condition to generate hierarchical structures of unbounded depth, and the most efficient procedure to generate multiple hierarchical levels (Berwick \& Chomsky, 2016). For instance, in language, even though the use of recursive computations cannot be directly verified, it is inferred from the ability to generate hierarchical structures, as it is thought to be the only plausible mechanism to generate sets of sets (Berwick \& Chomsky, 2016), without which there
can be no multiple levels of embedding. The same standard has been proposed for the visual-spatial domain (Martins, 2012).

Complex hierarchical structures occur in language, music and action planning (Fitch \& Martins, 2014). In these domains, it is difficult to establish the empirical boundaries of the generative capacity. This is especially true when external memory and recording devices are available, as for example, in written language or in large scale engineering projects, such as those involved in building a particle accelerator. Independently of how complex a base structure is, it is always possible to embed it within a higher-order hierarchy.

The investigation of these properties of human cognition pose several challenges, which we tried to address in a systematic 6-year long research program. The first challenge was the definition of a clear theoretical framework to make recursion empirically tractable and consistent with a number of different domains (Martins, 2012; Martins \& Fitch, 2014). Crucially, the availability of recursion must be tested experimentally and neither simply assumed nor deduced from pure analytical methods (Martins \& Fitch, 2015). The crucial behavioral signature of a computational capacity of recursion is the ability to generate muliple new hierarchical levels (Martins, 2014; Berwick \& Chomsky, 2016). This relation is independent of particular algorithmic and biological implementations (Berwick \& Chomsky, 2016).

Thus, the second challenge was the development of experimental techniques that could be used to test the ability to represent recursive hierarchical embedding in different domains. Once these two challenges were met, we could start answering two central questions: (1) in which domains of cognition recursion is available, and (2) if recursion were available in more than one domain, would it be instantiated by a domain-general capacity, or by multiple domain- specific abilities?

## Recursion in the visuo-spatial domain

In our previous work, we were able to establish that in addition to language (Roeper, 2011), the ability to represent recursive hierarchical embedding is available in the visual domain (Martins, Martins, \& Fitch, 2015) (Figure 1). We have shown that both human adults and children (Martins, Martins, \& Fitch, 2015; Martins, Laaha, Freiberger, Choi, \& Fitch, 2014) are able spontaneously acquire a hierarchical selfsimilarity rule of the kind $\mathrm{A} \rightarrow \mathrm{A}[\mathrm{A}]$ and to use it to make judgments about well-formed visual fractals and violations
(Figure 2A). Crucially, this ability differed cognitively from a control, iteration task (Martins, Martins, \& Fitch, 2015), in which participants made judgments about similar fractals which were not generated via a hierarchical embedding rule (Figure 2B). In particular, while visual iteration correlated strongly with visuo-spatial working memory and non-verbal intelligence, recursion correlated weakly with these measures, correlating instead with performance in the Tower of Hanoi, a recursive planning task (Martins, Martins, \& Fitch, 2015).


Figure 1: Examples of both 'recursive' and 'iterative' processes generating a visual fractal. While recursive hierarchical embedding steps generate new hierarchical levels, embedded iteration adds elements to fixed hierarchical levels, without generating new.

A. Visual Recursion Task

B. Visual Iteration Task

Figure 2: In both tasks participants are exposed to the first three steps generating a fractal (top row) and asked which image from the bottom row is the correct continuation. Despite using the same pairs of test images, the visual recursion and iteration tasks correlated with different abilities (Martins, Martins, \& Fitch, 2015).

As in language (Roeper, 2011), the development path towards the acquisition of visual recursion requires the induction of simpler iterative (conjunctive) representations first, before recursive embedding becomes available (Martins, Laaha, Freiberger, Choi, \& Fitch, 2014). However, we also found that the representation of visual recursion is independent of verbal resources (Martins, Mursic, Oh, \& Fitch, 2015) and that it is not instantiated by the classical brain networks supporting language (Martins, Fischmeister, et al., 2014). Using fMRI, we found that the capacity to represent recursion in the visual domain is supported by the visual ventral stream and by structures (e.g. Medial Temporal Lobe) associated with the episodic memory system (Figure 3). Interestingly, we found that in comparison with non-recursive iterative procedures, recursion hinges mostly on top-down, internal representations (Fischmeister, Martins, Beisteiner, \& Fitch, 2016), associated with the Default Mode Network and Semantic Memory. These findings provide some cues concerning the basic mechanisms underlying recursive hierarchical embedding and its usefulness for human cognition: by providing strong top-down priors, recursion can facilitate the processing of complex hierarchical structures.


Figure 3: Brain networks supporting the representations of Recursion $>$ Iteration (red) and Iteration>Recursion (blue) in the visiuo-spatial domain (thresholded at a voxel-wise $F D R$-adjusted $p<.05$ with a 10 -voxel extent threshold).

## Recursion in the music and motor domains

Recently, we have extended this research to the musical and motor domains, where the availability of recursion has been previously suggested (Corballis, 2014; Jackendoff \& Lerdahl, 2006). We found that both musicians and non-musicians can acquire rules governing recursive embedding in tonal hierarchies (Figure 4A) (Martins, Gingras, Puig-Waldmueller, \& Fitch, 2017), and that this capacity shares resources with visual recursion and with recursive action planning (Tower of Hanoi). This suggests some degree of domain-generality. However, when we measured the brain activity underlying this
musical capacity in well-trained participants (using fMRI), we found little overlap between the musical and visual domains.

Here, we exposed 15 non-musicians to the first three steps forming a tonal fractal, either using recursive or iterative rules (Figure 4B and 4C), then gave them 4 seconds to try to imagine how the correct continuation ( $4^{\text {th }}$ step) would sound like. This 4 -second period was the 'generation phase'. After the 4 seconds, they were exposed to a test stimulus ('test phase') and asked to judge whether this was a correct continuation or a foil. Overall, participants performed 4 sessions of 18 trials each (6 trials of Fractal, 6 of Iteration and 6 of Repetition)


Figure 4: (A) Example of music fractal with tonal relations between levels, which were identical across all levels. (B) Cross-level fractal (recursive) rule, (C) Withinlevel (iterative) rule.

By contrasting Recursion>Iteration in the "generation phase" (while controlling for the activations in steps I, II and III) we found that recursive hierarchical embedding in music is supported by the primary and secondary Auditory Cortices in the left hemisphere and by the right Superior Temporal Gyrus, an area known to encode complex tonal relations (Figure 5) (Martins, Fischmeister, et al., in prep.). Interestingly, despite differences in the specific pattern of activation, we found again evidence that the representation of recursive embedding is supported by top-down processing: These activations occur in anticipation to a certain tonal sequence before it is played ("generation phase"), but only when the sequence was generated recursively (vs. iteratively). There were no task differences in the "test phase", only a main effect of correctness (i.e. violation>well-formed structures).


Figure 5: Brain network supporting the representation of recursive hierarchical embedding in tonal sequences during the "generation phase" (FWE-adjusted $p_{\text {cluster }}<.05, p_{\text {voxel }}<$ .001). The Repetition task was a working memory task, in which participants simply had to buffer step 3 (a complete and well-formed fractal) and then determine whether step 4 was a repetition of step 3. The opposite contrasts (Iteration $>$ Recursion and Repetition>Recursion) yielded no activations.

Finally, we tested participants in the motor domain (Figure 6). Here we asked 20 (non-musician) participants to execute motor sequences on a 16-keys keyboard as depicted on a computer screen. Similar to the music domain, we exposed participants to the first two steps of a process generating motor fractals, then asked them to plan the next correct step (the "planning phase") for 6 seconds. After the planning phase they were asked to execute the correct sequence on the keyboard without visual assistance ('execution phase'). Iteration and Repetition baseline tasks were devised, similar to the music domain. Each participant performed 4 sessions of 20 trials each ( 8 Fractal, 8 Iteration and 4 Repetition trials).


Figure 6: A) Example of a recursive process generating sequences of (silent) finger movements. Red, Green and Blue denote key presses with the thumb, index and middle fingers, respectively. On each step N, each key press is substituted by a sequence of 3 key presses with less than one third of the duration $d_{n}$. B) Example of an Iterative process generating the same motor fractal.

We found that during the planning of hierarchical motor sequences using recursive rules, participants activated a network known to instantiate motor planning and imagining, comprising the Somato-Motor and Premotor cortices bilaterally, Cerebellum and Basal Ganglia (Figure 7) (Martins, Bianco, Sammler, \& Villringer, in prep.). Furthermore, we found that the underlying generating rule (Recursive>Iterative) changed how the execution of identical motor sequences were neurally represented: During the execution of a sequence formed using iterative rules, we found a strong activation in the primary motor hand area ( $x=-52$, $y$ $=-18, \mathrm{z}=50, \mathrm{Z}=6.06$, FWE $\mathrm{p}_{\text {cluster }}<.05, \mathrm{p}_{\text {voxel }}<.001$ ). In contrast, when the sequence was formed recursively, we did not find a direct activation in the motor cortex, but a modulation of this area from a fronto-striatal cluster (PPI: Recursion>Iteration: $x=-36, y=-24, z=48$, cluster extent $=$ 182 voxels, $Z=3.96, p_{\text {voxel }}=.016$ ).


Figure 6: Brain network supporting the representation of recursive hierarchical embedding in the motor domain (FWE-adjusted $p_{\text {cluster }}<.05, p_{\text {voxel }}<.001$ ).

## Discussion

Taken together, these results suggest several things: (1) The acquisition of recursive rules is probably supported by cognitive resources that are general across domains; (2) However, when we test participants that are well-trained and at ceiling performance in the fMRI, their representation of recursion is instantiated by domain-specific neural systems; (3) In contrast with other (iterative) rules applied to hierarchies, recursion seems to allow a controlled top-down processing, in both discrimination (visual and tonal) and production (motor) of well-formed hierarchical structures. This result is consistent across several domains.

The apparent contradiction between points (1) and (2) can be solved if we surmise that the resources necessary to acquire recursive rules are different from those necessary to utilize these rules after extensive training. We hypothesize that acquisition requires domain-general resources, which are perhaps slow and effortful, while expert use is instantiated by activating schemas stored in domain-specific repositories, which are formed after a process of automatization. The answer to this question requires novel research investigating the neural networks supporting the acquisition of recursive rules.

## Acknowledgments

I am grateful to my many collaborators, especially Arno Villringer, Roberta Bianco, Daniela Sammler, Florian Fischmeister, Roland Beisteiner, Bruno Gingras, Isabel Pavão Martins and W. Tecumseh Fitch.

## References

Berwick, Robert C. \& Noam Chomsky. (2016). Why Only Us: Language and Evolution. Cambridge, MA: MIT Press.
Corballis, Michael C. (2014). The recursive mind: The origins of human language, thought, and civilization: Princeton University Press.
Fischmeister, F.P., Martins, M.D., Beisteiner, R., \& Fitch, W.T. (2016). Self-similarity and recursion as default modes in human cognition. Cortex.
Fitch, W.T. (2010). Three meanings of "recursion": key distinctions for biolinguistics. In Richard K Larson, Viviane M Déprez, \& Hiroko Yamakido (Eds.), The evolution of human language (pp. 73-90). New York: Cambridge University Press.
Fitch, W.T., \& Martins, M.D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. Ann N Y Acad Sci, 1316, 87-104.
Hauser, M.D., Chomsky, N., \& Fitch, W.T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? Science, 298, 1569-1579.
Jackendoff, R., \& Lerdahl, F. (2006). The capacity for music: what is it, and what's special about it? Cognition, 100(1), 33-72.
Lobina, D.J. (2011). Recursion and the competence/performance distinction in AGL tasks. Language and Cognitive Processes, 26(10), 1563-1586.
Martins, M.D. (2012). Distinctive signatures of recursion. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1598), 2055-2064.
Martins, M.D., Bianco, R., Sammler, D., \& Villringer, A. (in prep.). Fractals in Action: the Brain on Motor Hierarchies.
Martins, M.D., Fischmeister, F.P., Gingras, B., PuigWaldmueller, E., Beisteiner, R., \& Fitch, W.T. (in prep.). Fractal music elucidates the neural representation of tonal hierarchies.
Martins, M.D., Fischmeister, F.P., Puig-Waldmuller, E., Oh, J., Geissler, A., Robinson, S., . . . Beisteiner, R. (2014). Fractal image perception provides novel insights into hierarchical cognition. Neuroimage, 96, 300-308.
Martins, M.D., \& Fitch, W.T. (2014). Investigating recursion within a domain-specific framework Language and Recursion (pp. 15-26): Springer.
Martins, M.D., \& Fitch, W.T. (2015). Do we represent intentional action as recursively embedded? The answer must be empirical. A comment on Vicari and Adenzato (2014). Consciousness and Cognition, 38, 16-21.

Martins, M.D., Gingras, B., Puig-Waldmueller, E., \& Fitch, W. T. (2017). Cognitive representation of "musical fractals": Processing hierarchy and recursion in the auditory domain. Cognition, 161, 31-45.
Martins, M.D., Laaha, S., Freiberger, E.M., Choi, S., \& Fitch, W.T. (2014). How children perceive fractals: hierarchical self-similarity and cognitive development. Cognition, 133(1), 10-24.

Martins, M.D., Martins, I.P., \& Fitch, W.T. (2015). A novel approach to investigate recursion and iteration in visual hierarchical processing. Behavioral Research Methods.
Martins, M.D., Mursic, Z., Oh, J., \& Fitch, W.T. (2015). Representing visual recursion does not require verbal or motor resources. Cognitive Psychology, 77, 20-41.
Roeper, T. (2011). The Acquisition of Recursion: How Formalism Articulates the Child's Path. Biolinguistics, 5(1-2), 57-86.

# The Temporal Cheerleader Effect: Attractiveness Judgments Depend on Surrounding Faces Through Time 

Alexia Toskos Dils (alexia.toskos@purchase.edu)<br>Scott Niedopytalski (scott.niedopytalski@purchase.edu) Jeffrey Arroyo (jeffrey.arroyo@purchase.edu) Stephen J. Flusberg (stephen.flusberg@purchase.edu)<br>SUNY Purchase College, Department of Psychology<br>735 Anderson Hill Road, Purchase, NY 10577, USA


#### Abstract

Previous research has found that people are seen as more attractive when they appear in a group rather than in isolation. The present study asks whether faces that surround us in time also affect how attractive we appear to be. Participants rated the attractiveness of famous female faces presented in a sequence of three and in isolation. We found that people do integrate information about attractiveness over time, but that temporal context has the opposite effect of static context. People perceived faces as less attractive in a series than in isolation. We also varied the attractiveness of surrounding faces in order to examine how the serial position of contextual information might figure into people's judgments. We found that faces presented earlier in the sequence figured more heavily into people's judgment than did faces presented later in the sequence. These findings highlight the role of temporal context in perceptions of attractiveness.


Keywords: face perception, attractiveness, serial position effects, ensemble coding, cheerleader effect

## Introduction

In a $5^{\text {th }}$ season episode of the American comedy television series The Office, the employees of Dunder Mifflin paper company spend an entire day debating whether actress Hilary Swank is "hot" or not. The office workers are torn on the issue, battle lines are drawn, and emotions get heated. Though in part a satirical referendum on the public's dark obsession with - and objectification of - celebrity, the plotline of this episode raises an important question: What factors influence perceptions of attractiveness?

The intrinsic features of individual faces certainly contribute to the perception of attractiveness, for both evolutionary and cultural reasons. Female faces, for example, are rated as more attractive the more sexually dimorphic and prototypically "female" they are (Valenzano, Mennucci, Tartarelli, \& Cellerino, 2006), and the more symmetrical they are (Perret et al., 1999). However, certain situational factors such as amount of exposure to a face (Rashidi, Pazhoohi \& Hosseinchari, 2012) as well as the perceived market value of the person making the judgment (Morgan \& Kisley, 2014) can also impact attractiveness judgments.

Researchers have also examined whether contextual factors can influence attractiveness ratings. Recent work, for example, has uncovered a so-called "cheerleader effect" in which people are rated as more attractive when they
appear in a group than when they appear in isolation (Walker \& Vul, 2013). Walker and Vul explain these findings as a sort of perceptual averaging phenomenon. The idea is that people spontaneously extract an ensemble code when viewing a group of faces, and because average faces are seen as highly attractive (Langlois, Musseman, \& Roggman, 1994), attractiveness ratings for the faces that contributed to the ensemble receive a boost. This suggests that perceptions of attractiveness are in part constructed online, in the moments we experience another person's face.

Often the faces we encounter in a crowded place are processed serially, rather than all at once. For example, faces come in and out of sight as we walk down the street, scan a room, or swipe through profile pictures on social media websites. Interestingly, evidence from studies of object perception suggests that the visual system is capable of constructing average representations over time in addition to space. Albrecht and Scholl (2010) found that people's estimates of the average diameter of a growing or shrinking disc depended on which part of the disc received the most screen time. Participants overestimated the average when frames on the larger end of the spectrum hung on the screen longer than did frames on the smaller end of the spectrum, and vice versa.

Is information from faces spontaneously integrated over time in a similar way? The present study asks whether the cheerleader effect extends to faces that appear near one another in time in addition to space. Importantly, the study was also designed to address whether serial position influences how information about attractiveness is integrated over time. Do all faces in a sequence figure equally into the ensemble code, or is the average representation that people extract weighted more heavily by faces appearing early or late in the series?

Though research suggests the visual system computes ensemble codes of information presented close in space and time, the data is often inconsistent with a simple averaging account. For example, studies of the perception of serially presented lines have reported a recency effect in people's judgments: Estimates of line length were biased toward lines that appeared toward the end of the sequence (Weiss \& Anderson, 1969). And while Walker and Vul (2013) concluded that an averaging effect best captured the data they observed in their work on the cheerleader effect, not all of their findings are consistent with this interpretation. For
example, if the cheerleader effect results from simple averaging, then the more faces that contribute to the ensemble code, the more attractive the resulting average should be (i.e., increasing the number of faces in a group photograph should strengthen the cheerleader effect). Walker and Vul (2013, Experiment 4) tested this hypothesis with set sizes ranging from two to 16 . Although they found a cheerleader effect within each set size, the magnitude of the effect did not increase with the number of faces in the set.

In light of these findings, the present study was designed to differentiate between different possible accounts of the temporal cheerleader effect we investigated. On the one hand, it is possible that faces appearing at the end of a sequence bias the perceived attractiveness of faces earlier in the sequence (i.e., a recency effect). This would be consistent with the work of Weiss and Anderson (1969) on line perception. However, it is also possible that the first face might influence the perceived attractiveness of faces appearing later in the sequence. There are at least two reasons we might expect such a primacy effect. First, there might be a contrast effect (Kenrick \& Gutierres, 1980; Pegors, Mattar, Bryan, \& Epstein, 2015), such that seeing a highly attractive face may make subsequent faces seem less attractive by contrast, and vice versa. Second, there might be an anchoring effect, such that subsequent faces appear to have a similar attractiveness level as a previously presented face (Pegors, Mattar, Bryan, \& Epstein, 2015; Taubert, Van Der Burg, Alais, 2016; Tversky \& Kahneman, 1974).

Both contrast and anchoring effects have been observed in recent studies examining the effects of the attractiveness of the previous face on online judgments of serially presented faces (Pegors et al. 2015; Taubert et al., 2016). Pegors and colleagues suggest that the contrast effect derives from perceptual components of the judgment, whereas the anchoring effect derives from a bias to respond in the same way as in the previous trial. The present study extends this work to offline judgments of facial attractiveness. Does the temporal context in which a face appears affect later memory of how attractive that face was? Offline judgments further allow us to examine any serial position effects more fully. Specifically, this design element allows us to test whether faces that follow us in time can retroactively meddle with how attractive we are judged to be.

In order to (a) determine whether or not an offline, temporal cheerleader effect exists, and (b) understand the mechanisms underlying such an effect, participants in the present study rated the attractiveness of a variety of famous female faces. On each trial, faces were presented either in isolation or in a series of three faces. Afterward, participants were cued by the name of each celebrity they saw and asked to rate how attractive the person looked in the photograph. Celebrity faces were used (as opposed to non-famous faces) to allow for offline attractiveness ratings to be collected after all faces disappeared from view. The celebrity's name uniquely picked out which face participants were being asked to rate.

The stimulus set included two versions of each celebrity: an "attractive" version where the celebrity was photographed favorably, and an "unattractive" version where the celebrity was photographed uncharitably (see Figure 1 for example stimuli). On trials where faces were presented in a series, the middle face was always an attractive face and was either preceded by or followed by an unattractive face. Therefore, in the unattractive first condition, participants first saw an unattractive face followed by two attractive faces. In the unattractive last condition, participants saw two attractive faces followed by an unattractive face.

To the extent that people do integrate information about attractiveness over time, offline judgments of a given face should depend on whether the face was presented in isolation or in the middle position of a series of three faces. If the mechanism by which information is integrated is a contrast effect, then viewing an unattractive face in the first position of the sequence should cause the middle face to seem more attractive than it does in isolation, and vice versa. Alternatively, if the mechanism is anchoring, then seeing an unattractive face in the first position of the sequence should make subsequent faces seem less attractive than they do in isolation, and vice versa. Finally, if the mechanism is simple averaging, then all faces should appear more attractive when they are presented in a sequence, since research suggests that averaged faces appear more attractive (Langlois, Musseman, \& Roggman, 1994). However, since there are only three faces being averaged together here and one was specifically chosen to appear highly unattractive, it could be the case that the average of this small set would actually appear less attractive. What's more, these effects may depend on where the unattractive face appears in the sequence. If the averaging effect is subject to primacy effects, then the first face in the sequence will figure more heavily into the average. However, if the averaging effect is subject to recency effects, then later faces in the sequence will figure more heavily into the average. This study will help to rule out some of these possibilities.

## Experiment

## Methods

Participants 50 "master-level" participants were requested from Amazon's Mechanical Turk worker pool for this study. 68 people actually followed the link to the experiment and began the study, but 17 of those people did not complete the task. The remaining 51 participants were included in the study. There were 27 males and 23 females in the sample, with a mean age of $41.20(S D=12.67)$. One participant declined to provide demographic information. Each participant received $\$ 1.50$ for their time spent on the study.

Stimuli and Procedure Participants rated the attractiveness of faces in three within-subjects conditions: isolation (a face shown by itself), unattractive first (an unattractive face followed by two attractive faces in sequence), unattractive
last (two attractive faces followed by an unattractive face in sequence).


Figure 1. Example stimuli used in this study. On the left is the "unattractive" photograph of Debra Messing that participants saw, and on the right is the "attractive" photograph of the same celebrity.

Photographs of 75 unique female celebrities made up the stimulus set for this study. Two photographs were collected for each celebrity: an "attractive" version in which the person was made up for an event, and an "unattractive" version in which the celebrity was captured poorly (i.e., photographs found in tabloids claiming to have "shocking" pictures of the person; see Figure 1 for example stimuli). Therefore, the final stimulus set consisted of 150 celebrity photographs. Each participant saw a subset of 75 faces sampled from this full stimulus set ( 50 attractive faces and 25 unattractive faces). Participants saw either the attractive version or the unattractive version of each unique celebrity, and never both. Whether a celebrity was presented as attractive or unattractive was counterbalanced across participants.

The study was conducted using Qualtrics survey building software. On each trial, participants saw female celebrity faces presented either in isolation or in a series of three faces. Each face was presented along with the celebrity's name and remained on the screen for one second. Afterward, participants were instructed to rate the attractiveness of each face they had just seen, cued only by the celebrity's name. Participants entered their responses using a continuous sliding scale without numbers on it. This was designed to prevent participants from tracking the specific ratings they assigned to faces throughout the study. The scale ranged from "very unattractive" to "very attractive". Participants were allowed to take as much time as they needed to enter their ratings. The order in which participants were prompted to rate each celebrity was randomized on each trial. See Figure 2 for a schematic representation of the trial structure in this task.

Each participant completed 125 trials. 75 trials consisted of faces presented in isolation, and 50 trials consisted of faces presented in a series. Of the 50 trials in which the faces were presented in a group, 25 were unattractive first trials, and 25 were unattractive last trials. The middle face in the sequence was always attractive. Which face appeared in which position in the sequence was counterbalanced
across participants. All trial types were interleaved, and trials were presented to participants in random order.

## A.


B.


Figure 2. Trial structure for the conditions in which faces were presented in a sequence (A) and in isolation (B). Part A depicts an example unattractive first trial.

Participants rated each attractive photograph at two critical points in the study: once in isolation and once in the middle position of a series of three faces. We wanted to know whether attractiveness ratings for each photograph depended on this contextual manipulation, so we limited our analyses to the 50 attractive faces each person saw, as well as to ratings of the face presented only in the middle position on group trials.

At the end of the study, participants supplied information about their gender, race/ethnicity, and age. We also included a manipulation check designed by Oppenheimer, Meyvis, and Davidenko (2009) to ensure that all participants in our sample read task instructions carefully.

## Results

All 51 participants in this study correctly answered the instructions manipulation check, so data from all participants were analyzed. One participant declined to provide demographic information and therefore was not included in analyses that examined effects of gender.

To confirm that our participants found the "attractive" faces to be more attractive than the "unattractive faces," we compared participants' raw attractiveness ratings between
these two stimulus types. Indeed, participants rated the attractive faces more favorably on a scale from 1-100 ( $M=$ 70.91; $S D=13.12$ ) than they did the unattractive faces, $M=$ 34.46; $S D=16.46 ; t(50)=15.83, p<.001$.

For each participant, responses on analyzed trials were converted to $z$-scores. For each attractive face that a given subject saw, we subtracted the standardized rating of that face presented in isolation from the standardized rating of that same image shown in a group. Therefore, positive difference scores indicate that faces were rated as more attractive when presented in a group compared to isolation, and negative difference scores mean faces were rated as less attractive when presented in a group compared to isolation. Overall, participants rated celebrity faces as significantly less attractive when they were presented in a series than when they were presented in isolation ( $M_{\text {group-isolation }}=-$ $\left.0.098, S D_{\text {group-isolation }}=0.16\right), t(50)=-4.24, p<.05^{1}$.

Figure 3: Shift in standardized attractiveness ratings for faces presented in a sequence from those same faces presented in isolation. Error bars represent 1 SEM.

The differences in z-scores were submitted to an ANOVA with one within-subjects factor (order of attractiveness: unattractive first vs. unattractive last) and one betweensubjects factor (gender of participant: male vs. female). The ANOVA revealed a significant main effect of order of attractiveness on the attractiveness deficit for faces presented in a group relative to isolation. Specifically, the effect was stronger when participants saw an unattractive face first in the sequence ( $M=-0.16, \mathrm{~S} D=0.23$ ) than it was when participants saw an unattractive face last in the sequence $(M=-0.03, S D=0.23), F(1,48)=9.78, p<.005$ (see Figure 3). These effects were similar in men and women: The attractiveness deficit did not depend on the

[^322]gender of the participant $(F(1,48)=0.04, p>.50)$, and there was also no interaction between gender and order of attractiveness, $F(1,48)=1.97, p>.05$.

Post-hoc analyses revealed that seeing an unattractive face at the beginning of the sequence caused the middle face to appear less attractive than it did in isolation ( $M_{\text {group-isolation }}$ $\left.=-0.16, S D_{\text {group-isolation }}=0.23\right), t(50)=-5.07, p<.001$. While the presence of an unattractive face at the end of a sequence produced a numerical reduction in attractiveness of the middle face relative to isolation, this shift did not reach statistical significance $\left(M_{\text {group-isolation }}=-0.03, S D_{\text {group-isolation }}=\right.$ $0.23), t(50)=-1.03, p>.05$.

## Discussion

The purpose of this study was to determine whether and how people integrate information from faces over time when judging the attractiveness of others. In our experiment, participants viewed a series of attractive and unattractive female celebrity faces in isolation and in a series of three images. They were then asked to rate the attractiveness of each face (cued only by their name) on a sliding scale from very unattractive to very attractive. The results indicate that people do integrate information over time when judging the attractiveness of faces. Contrary to the "cheerleader effect," which suggests that faces are perceived as more attractive when they are presented in a group, the results from this experiment suggest that faces are perceived as less attractive when presented in a series than when they are presented in isolation. The results further show that this effect is stronger when an unattractive face is presented first in the series compared to when it is presented last.

There are a number of reasons why attractiveness ratings might be different for faces presented in a series compared to isolation. For example, seeing faces in a sequence may lead to a contrast effect, where seeing an attractive face causes a subsequent face to appear less attractive (and vice versa). Or there could be an anchoring effect where people rate subsequent faces to be similarly attractive to the initial face. Finally, there could be an averaging effect where all faces in the sequence contribute to an ensemble code that figures into attractiveness judgments for each face in the sequence. The results from the present study help to rule out one of these three possibilities: the contrast effect. Had our manipulation produced a contrast effect, then seeing an unattractive face first would have caused the middle face in the series to seem more attractive. Instead, our results showed the opposite pattern: viewing an unattractive face made the subsequent face appear to be less attractive than when it was viewed in isolation.

The data are partially consistent with an anchoring effect. An attractive anchor face should cause subsequent faces to seem more attractive than usual, and an unattractive anchor should cause subsequent faces to seem less attractive than usual. Indeed, unattractive faces at the beginning of the sequence caused subsequent faces to appear less attractive than usual, but the opposite pattern did not obtain for
attractive faces presented early in the sequence. Therefore these data are only partly in line with the findings of Taubert et al. (2016).

Pegors and colleagues (2015) concluded from their study examining sequential attractiveness judgments that contrast effects result from stimulus (perceptual) bias, whereas anchoring effects result from response bias. The present findings are not consistent with this explanation as the study design itself rules out response bias as a possible contributor. On trials in which faces were presented in a group, participants saw all three faces before responding to any one of them. They then entered their responses in a different order than the one in which the faces were presented. We analyzed the data based on the stimulus order and not the response order, so the partial anchoring effect should not necessarily be interpreted as a result of response bias. Furthermore, the lack of a contrast effect is a departure from the findings of Pegors and colleagues, suggesting there may be a qualitative processing difference between online and offline judgments of serially presented faces that persists well beyond the presentation of the face.

The findings presented in this paper are perhaps most consistent with an averaging effect. By including an unattractive face in the sequence of three, we may have reduced the attractiveness of the ensemble code participants extracted from the series, which then biased subsequent ratings of individual faces in the sequence. Furthermore, it appears that the serial position of the unattractive face influenced how the average representation was constructed. Contrary to the findings of Weiss and Anderson (1969), our results revealed a primacy effect on perceptual averaging. Namely, unattractive faces presented at the beginning of the series figured more heavily into the average than did unattractive faces presented at the end of the series. However, because averaging faces together generally increases perceived attractiveness, this account depends on the possibility that averaging only a small number of faces together (three), one of which is especially unattractive, can sometimes reduce rating attractiveness (perhaps, for example, when the faces are of familiar people, as in the present study). Indeed, morphs containing familiar (celebrity) faces are rated as less attractive than are the component faces used to generate the morph (Halberstadt, Pecher, Zeelenberg, Ip Wai, \& Winkielman, 2013). Therefore, using celebrity face may itself have lowered the attractiveness of the average face people constructed over time in the present study. Future work is required to confirm this possibility, however.

In order to preserve the design element in this study that no participant saw both the attractive and unattractive versions of the same celebrity, we did not include trials in which all faces in a sequence were attractive, nor did we include trials in which all faces were unattractive. Future work will examine these trial types so as to build a more complete picture of the way in which attractiveness in faces is integrated over time.

One ongoing study attempts to further differentiate between an anchoring and averaging account of our findings by replicating the present study with unattractive faces in the target (middle) position on group trials. If our findings reflect a cheerleader effect, then the results from this new study should mirror the results in the present study. Specifically, the middle face in the sequence would be rated as less attractive than a face presented in isolation, but this effect would only appear in the unattractive first condition. However, if our current findings reflect an anchoring effect, we should predict that the unattractive last condition in this ongoing study would produce the strongest anchoring effect - the condition in which an attractive face appears in the first position of the sequence. After all, this is the condition where the difference in attractiveness is largest between the first and middle faces. Therefore, an anchoring effect in this new version of the task predicts that the middle face would be rated as more attractive in a series than it is in isolation.

Our judgments about the attractiveness of others factor into a variety of important decisions we make every day. For example, whether a candidate would make a good fit for a job (Dipboye, Fromkin, \& Wilback, 1975), or whether to ask a person out on a date can be influenced by how we perceive the attractiveness of that individual. This study highlights one way in which such important decisions might be influenced by the faces that happen to surround us in the moments leading up to them.

## Acknowledgments

This research was supported by undergraduate research support awards from the School of Natural and Social Sciences at SUNY Purchase to SN and JA.

## References

Albrecht, A. R., \& Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. Psychological Science, 21(4), 560-567.
Dipboye, R. L., Fromkin, H. L., \& Wilback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic in evaluation of job applicant resumes. Journal of Applied Psychology, 60(1), 39-43. doi:10.1037/h0076352
Halberstadt, J., Pecher, D., Zeelenberg, R., Ip Wai, L., \& Winkielman, P. (2013). Two faces of attractiveness: Making beauty in averageness appear and reverse. Psychological Science, 24(11), 2342-2346.
Kenrick, D. T., Gutierres, S. E. (1980). Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. Journal of Personality and Social Psychology, 38(1), 131-140.
Langlois, J. H., Musselman, L., Roggman, L. A. (1994). What is average and what is not average about attractive faces? Psychological Science, 5(4), 214-220.
Morgan, L. K., \& Kisley, M. A. (2014). The effects of facial attractiveness and perceiver's mate value on adaptive
allocation of central processing resources. Evolution and Human Behavior, 35(2), 96-102. doi:10.1016/j.evolhumbehav.2013.11.002
Oppenheimer, D. M., Meyvis, T., \& Davidenko, N. (2009). Instructional manipulation check: Detecting satisficing to increase statistical power. Journal of Experiment Social Psychology, 45, 867-872. doi:10.1016/j.jesp.2009.03.009
Pegors, T., Mattar, M., Bryan, P., \& Epstein, R. (2015). Simultaneous perceptual and response biases on sequential attractiveness judgments. Journal of Experimental Psychology: General, 114(3), 664-673. doi:doi:10.1037/xge0000069
Perrett, D. I., Burt, D. M., Penton-Voak, I., Lee, K. J., Rowland, D. A., \& Edwards, R. (1999). Symmetry and human facial attractiveness. Evolution and Human Behavior, 20, 295-307. doi:10.1016/S1090-5138(99)00014-8
Rashidi, M., Pazhoohi, F., \& Hosseinchari, M. (2012). Effect of facial stimuli exposure time on evaluation of facial attractiveness. Australian Journal of Psychology, 64(3), 164-168. doi:10.1111/j.1742-9536.2011.00050.x
Taubert, J., Van Der Burg, E., Alais, D. (2016). Love at second sight: Sequential dependence of facial attractiveness in an online dating paradigm. Scientific Reports, 6, 22740. doi:10.1038/srep22740
Tversky, A., Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.
Valenzano, D. R., Mennucci, A., Tartarelli, G., \& Cellerino, A. (2006). Shape analysis of female facial attractiveness. Vision Research, 46(8-9), 1282-1291. doi:10.1016/j.visres.2005.10.024
Walker, D., \& Vul, E. (2013). Hierarchical encoding makes individuals in a group seem more attractive. Psychological Science, 25(1), 230-235. doi:10.1177/0956797613497969
Weiss, D. J., \& Anderson, N. H. (1969). Subjective averaging of length with serial presentation. Journal of Experimental Psychology, 82(1), 52-63. doi:10.1037/h0028028

# Architectural process models of decision making: Towards a model database 

Cvetomir M. Dimov (cvetomir.dimov@unil.ch), Julian N. Marewski (julian.marewski@unil.ch)<br>Department of Organizational Behavior, Université de Lausanne,<br>Quartier UNIL-Dorigny, Bâtiment Internef, 1015 Lausanne, Switzerland

Lael J. Schooler (Ischoole@syr.edu)
509 Huntington Hall, Department of Psychology, Syracuse University, NY 13244, USA


#### Abstract

We present the project aimed at creating a database of detailed architectural process models of memory-based decision models. Those models are implemented in the cognitive architecture ACT-R. In creating this database, we have identified commonalities and differences of various decision models in the literature. The model database can provide insights into the interrelation among decision models and can be used in future research to address debates on inferences from memory, which are hard to resolve without specifying the processing steps at the level of precision that a cognitive architecture provides.


Keywords: inference from memory; process model; ACT-R; decision making; model database

## Introduction

How do we infer which of two cars will be more durable? Which company will be more successful in the coming year? To address such questions, in a typical two-alternative forced-choice task of inference from memory (Gigerenzer \& Goldstein, 1996), two objects (e.g., two companies) are presented on a computer screen. A subject has to infer which of the two objects scores higher on a criterion of interest (e.g., the company growth in the next year) by relying on knowledge stored in memory.

Models of inference describe how subjects make inferences by using attributes of objects (e.g., who is the company's CEO) as cues. Many inferential models have focused on describing not just what the outcome of the inference would be, but also which processing steps a decision maker would take to reach a decision. These models include, among others, the various fast-and-frugal heuristics from the adaptive toolbox of heuristics (Gigerenzer, Todd, \& the ABC Research Group, 1999), parallel constraint satisfaction (PCS; Glöckner \& Betsch, 2008) and sequential sampling models (e.g., Lee \& Cummins, 2004).

Such process models have increased substantially our understanding of how people make inferences (e.g., Bröder, 2012) and why the inferential process is successful (Gigerenzer \& Brighton, 2009), but perhaps more importantly they have raised other questions and fueled important debates: Do people rely on a repertoire of strategies or on a single strategy (e.g., Lee \& Cummins, 2004; Marewski, Schooler, \& Gigerenzer, 2010; Newell, 2005; Glöckner \& Betsch, 2008)? Which types of models
(e.g., heuristics vs. more complex models) describe better people's decision processes (e.g., Goldstein \& Gigerenzer, 2002; Newell \& Bröder, 2008) and under what circumstances? When do people rely on non-compensatory as opposed to compensatory strategies (Glöckner \& Bröder, 2011)?

One major barrier to addressing those and related questions is that many models are almost always underspecified compared to the data that they are tested against. Specifically, process models of decision making often remain silent about components of cognition that are the foundation of decision making, such as perception, motor action, or memory. We argue that specifying relevant cognitive-behavioral processes will help those models make more precise predictions about, for example, response time and other process data. The increased precision, in turn, will not only allow researchers to more easily tell potentially competing models apart, but also aid in addressing ongoing debates and open research questions.

In fact, a significant amount of research has already started to embed existing decision models into detailed cognitive theories (Dimov, Marewski, \& Schooler, 2013; Fechner, Pachur, Schooler, Mehlhorn, Battal, Volz, \& Borst, 2016; Marewski \& Mehlhorn, 2011; Marewski \& Schooler, 2011; Nellen, 2003; Thomas, Dougherty, Sprenger, \& Harbison, 2008; Schooler \& Hertwig, 2005). The aim of the current line of work is to expand upon these efforts by systematically implementing existing models of inference in the cognitive architecture ACT-R (Anderson, 2007).

In what follows, we will briefly introduce ACT-R and present a summary of the model database that we are in the process of constructing. We will then explain in detail what knowledge each of the decision strategies requires for its functioning. We will conclude by discussing the advantages and shortcomings of our models. Once finalized, we plan to make the database of architectural process models of decision making available to the public.

## ACT-R

ACT-R is arguably the most advanced integrated theory of cognition. It has been used to construct models of very diverse tasks and phenomena, which include, among others, associative recognition (Schneider \& Anderson, 2012), analogy making (Salvucci \& Anderson, 2001) and multitasking (Salvucci \& Taatgen, 2008).

Table 1: Outline of the database of architectural process models of decision making, together with summaries of hypothesized procedural and declarative, symbolic and subsymbolic knowledge.

| Model | Source | Declarative <br> knowledge | Procedural knowledge |  |
| :--- | :--- | :--- | :--- | :--- |
| Recognition <br> Heuristic |  <br> Gigerenzer <br> $(2002)$ | Alternatives | Try to retrieve chunks representing alternatives. <br> Select alternative corresponding to successfully <br> retrieved chunk. | Information at the subsymbolic <br> lever |
| Fluency Heuristic |  <br> Hertwig $(2005)$ | Alternatives | Retrieve chunks representing alternatives and time <br> retrieval using timing module <br> Select alternative with faster retrieval time. | environment) |

ACT-R describes cognition as a set of modules that communicate through a procedural module realized as a central production system. The production system consists of production rules (i.e., if-then rules) whose conditions (the "if"-parts) are matched against the modules. If a rule's conditions are met, then the rule can fire and the specified action can be carried out. Modules model different cognitive processes, such as vision (visual module), motor action (motor module), declarative memory (declarative module), short-term information storage (imaginal module) and time tracking (timing module; Taatgen, van Rijn, \& Anderson, 2007). Productions send commands to modules to perform an action or change their state, or to access content placed in modules' buffers. In fact, because productions can only access content placed in the buffers, these can be thought of as processing bottlenecks. For instance, a production rule cannot access all information stored in the declarative module, but only the information placed in its associated retrieval buffer.

Productions are the representation of choice for procedural knowledge, while declarative knowledge, such as factual and episodic knowledge, is represented as chunks. Perceptual and memory modules, respectively, perceive and retrieve information in the form of chunks. A chunk consists of a set of slots, where each slot is (a pointer to) another chunk. For example, a chunk containing information about a company's annual revenue will have a slot with the company's name and another slot with its revenue.

ACT-R distinguishes a symbolic and a subsymbolic system. Productions, modules and buffers constitute the symbolic system, whose dynamics are governed by a set of equations, describing ACT-R's subsymbolic system. At the subsymbolic level, chunks' activations determine, for example, retrieval time or recall probability; productions' utilities reflect which productions were more successful in the past and therefore more likely to fire; visual parameters determine the time needed to shift visual attention to an object in the visual field, while motor parameters determine the time to generate a motor response.

Each ACT-R model is essentially composed of specifications of how declarative and procedural knowledge interact, both at the symbolic and subsymbolic levels. We will now focus on describing the declarative and procedural knowledge used in defining the models in the database. We refer those interested in a detailed exposition of ACT-R to Anderson (2007).

## Model building blocks

The models of inference that we will consider are presented in Table 1. In implementing these models in ACT-R, we relied on the building blocks that this cognitive architecture provides.

## Perceptual and motor processes

All models have equivalent perceptual and motor processes, involving visual perception from a screen and manual action on a keyboard. The models first perceive
each of the alternatives presented on a computer screen and, after executing a sequence of cognitive steps, they make a response by pressing the appropriate key on a keyboard. The primary contribution to behavioral predictions of the perceptual and motor processes in our models is to add a realistic estimate of perceptual-motor latency.

## Declarative chunks

The factual knowledge (e.g., "Berlin is a capital") that a model relies upon to make a decision is stored in declarative memory. Ten types of chunks are needed to construct the models in the database. Table 2 provides a summary of those chunk types and examples in Lisp code for each. Note that the examples are given for the city-size task, in which cities act as alternatives and subjects need to infer which of two cities is larger.

The simplest chunk type contains just the name of the alternatives. For example, if the alternatives are cities, whose relative sizes need to be inferred, such a chunk contains the city name (e.g., "Berlin"). These chunks are all that is required for inferential models, which rely on accessibility information, such as the recognition and fluency heuristics.

The second chunk type contains an entire cue profile of an alternative (i.e., the set of cues associated with an alternative). Such chunks are used, among others, by exemplar and prototype models. Some exemplar models also require chunks with direct criterion knowledge in addition to the cue profile. Moreover, prototype models require not only cue profiles, but also a stored prototype of an object with a high criterion value.

Table 2: Declarative knowledge categories.

| Chunk type label | Chunk examples in Lisp code |
| :--- | :--- |
| Alternative | (berlin name Berlin) <br> (berlin name Berlin airport yes capital <br> yes ...) |
| Cue profile | (berlin name Berlin population <br> Cue profile with <br> direct criterion <br> knowledge |
| High criterion value <br> prototype | (big-city name prototype airport yes <br> capital yes ...) |
| Cue profile pair | (pair1 airportl yes airport2 no <br> capitall yes capital2 no ...) |
| Cue profile pair <br> prototype | (prototype-left airportl yes airport2 <br> no capitall yes capital2 no ...) |
| Cue | (cue1 type airport) |
| Cue value | (berlin-airport city Berlin cue airport <br> value yes) |
| Cue validity | (airport-validity cue airport validity <br> 90) <br> (cue-pair first airport second capital) |
| Cue validity pair |  |
| Note. In these examples, chunk names, used for convenience, are <br> presented in bold; slot names, indicating a specific attribute, are in <br> italics, while slot values, representing the attribute values, are in <br> normal font. |  |

Resembling exemplar and prototype models, instancebased learning theory and parallel constraint satisfaction consider cue configurations to make inferences. However, they differ from the former in that they require chunks, which contain pairs of cue profiles. For example, the model "Instance-based learning theory individual" retrieves the cue profiles of both alternatives and then retrieves a cue profile pair from a successful previous trial. It then makes an inference based on the decision outcome of the retrieved cue profile pair. Similarly, our implementation of the parallel constraint satisfaction model requires a prototype of a successful cue profile pair.

Unlike configural models, like exemplar models, cueabstraction models (Newell \& Bröder, 2008) operate on individual cues. Such models, like take-the-last, retrieve cues one by one. Take-the-last requires separate chunk types for a cue and for the values of the alternatives on that cue. In addition to these chunks, other models, like take-the-best, require information about cue validities (i.e., the probability of making a correct inference using only this cue if the cue discriminates; see, Gigerenzer, Hoffrage, \& Kleinbölting, 1991), which, if taught in the experiment (e.g., Bröder, \& Schiffer, 2003), are stored numerical values. Finally, in some experiments one is provided only with the validity hierarchy, which can be represented as validity pairs of subsequent cues.

## Procedural knowledge: The sequence of processing steps

The procedural knowledge of a model consists of a finegrained sequence of processing steps (i.e., productions) that the model relies upon to make a decision. In all models, the sequence of processing steps includes commands to the visual module to encode the information presented on the screen and to the motor module to press a key to respond in a computerized experiment. As for the rest, the exact sequence of processing steps follows the original model definitions.

For example, fast-and-frugal heuristics usually rely on separate cues, on which detailed search, stopping and decision rules operate. Those models often theorize about the order, in which cues are considered. This ordering can be modeled through productions. In addition, productions can also determine if the model weighs cues equally, as in tallying, or differently, as in the weighted additive model, and execute this process. If cues are weighted equally, productions are required to send a request to declarative memory to retrieve the cue values. Productions then increment by 1 the number, which tracks the count of cues with a positive cue value of the alternative of interest. Other models, such as exemplar models, rely on all available cue information stored in a single chunk to make a decision. In such models, procedural knowledge is more peripheral to the decision process and mostly focuses on retrieval attempts.

Productions not only initiate retrieval, but are also dependent on what is retrieved, because a key determinant of which productions can fire is the available declarative
knowledge. Specifically, at each point in time only those productions, whose condition match the buffer states, will be considered to fire. Ultimately, which chunks are retrieved from memory will determine what could be placed in the buffers and therefore which productions will match.

## Information at the subsymbolic level

At the subsymbolic level, there is continuously valued information, which is necessary for the execution of some inferential strategies. However, productions cannot directly read out subsymbolic values. Instead, the model needs to let subsymbolic values guide symbolic knowledge. Thus far, we have identified four ways in which subsymbolic values play a key role in the execution of strategies.

First, the activation of chunks representing alternatives contains information about the alternatives' occurrence frequency in the environment. Specifically, base-level activation is a function of prior history of a chunk, which partially depends on environmental occurrence frequency, which, in turn, is related to many criteria of interest (Hertwig, Herzog, Schooler, \& Reimer, 2008). Accessibility-based strategies, such as the fluency heuristic, track the retrieval speed of alternatives as determined by their activation and choose the alternative, which was retrieved noticeably faster.

Second, activation can order cues, because cues which have a higher occurrence in the environment likely will have a higher activation. Thus, these cues may be more likely to be considered first in lexicographic strategies, such as take-the-first-cue or a sequential sampling model.

A third way in which information at the subsymbolic level can be used is as an implicit cue weighting mechanism. This weighting can take place through spreading activation, which is determined by the degree of association between the chunks placed in buffers and the chunks in declarative memory. If the cue profile of one of the alternatives is currently placed in the imaginal buffer, then it will activate cue profiles in memory through spreading activation. Those cue profiles will then have precedence in retrieval. Exemplar models rely on this process to make an inference about the alternative's criterion value.

Finally, production utility contains information about prior success. Production utility determines which production is more likely to fire when two or more productions are competing. If such a competition takes place between productions, which select which cue will be considered next, the utility of these productions can act as a cue's importance (e.g., as its validity, see Gigerenzer, Hoffrage, \& Goldstein, 2008, for the hypothesis that such a reinforcement learning process can teach cue validities) in lexicographic cue-abstraction models. This is the mechanism used in the model "Take-the-best reinforcement", which encodes the selection of each cue with a separate production and then learns the success of those cues through trial and error.

## Discussion and conclusion

We aim to provide a database of ACT-R implementations of decision models used in the literature on inferences from memory. We have divided these models into their key components. The models can serve as a basis for model tests and further model developments. Specifically, this database can be used, first, in model comparison simulations on the outcome and process level, whereby one identifies regions in the parameter space where these models diverge. Second, this database can be used in future studies to identify decision processes using both behavioral and neural data. This is an important advantage of ACT-R, because any ACT-R model can generate fMRI predictions on top of behavioral process predictions, such as response time, because of the established module-to-brain mappings (for an introduction, see Borst \& Anderson, 2015).

In addition, we think that the systematic examination of the building blocks of existing decision models will help us gain insights into how the models are related to each other. For example, through these implementations, we see that the parallel constraint satisfaction model can be conceived as functionally similar to an instance-based learning model, which stores and retrieves prototypical cue profile pairs.

It is important to note that in creating our ACT-R models we were forced to work with the mechanisms that ACT-R provides. For example, the original parallel constraint satisfaction model is cast as a connectionist network, in which connection weights are iteratively updated after each decision. This leads to cues effectively changing their validities as trials progress. As currently conceived, our model does not reproduce this behavior. Nevertheless, the model "Instance-based learning theory average", which in our database is very similar, effectively provides such a mechanism and can be thought as functionally analogous to the original parallel constraint satisfaction.

Such redefinitions and novel distinctions introduced in our modeling endeavor were due to the partial overlap between the various decision models in the literature. Another such distinction that we decided to introduce was in the declarative representation, which cue-abstraction models, like take-the-best and the sequential sampling model introduced by Lee and Cummins (2004), rely on. Originally, both models were conceived as, first, considering a cue, and only then examining the values of that cue for both alternatives. We have kept this definition for take-the-best and other heuristics. However, we have decided to label those models, which retrieve cue values directly, in a manner purely determined by declarative principles, sequential sampling models. These models can, for example, consider the value of cue 2 for alternative A, followed by the value of cue 4 for alternative $B$, and so on.

Another remark concerns the high degree of detail, which ACT-R introduces when decision models are implemented in it. The fine-grained way in which ACT-R models are specified has forced us, in many cases, to make assumptions about processes, about which the original models remained silent. For example, we had to rely on assumptions about
how cues are ordered in take-the-best. We have considered two ways to order cues in this work. Our first implementation relies on declarative retrieval to order cues, while the second one relies on procedural knowledge and utility learning. These assumptions reflect, so we hope, realistic ways of learning. On the one hand, in many experiments on take-the-best, one is explicitly taught the cue hierarchy, which is then stored as declarative knowledge. On the other hand, in natural settings, ordering cues according to validity is likely to occur through reinforcement learning, whereby one has had significant experience with considering several cues in the same setting.

To conclude, we would like to stress that Table 1 does, naturally, not include all possible tweaks and modifications that one can introduce when constructing models in ACT-R. It will be left to input from the different researchers working on inference from memory to determine which of our current ideas will survive, and which ones will be replaced or extended by others.

## Acknowledgments

We thank the Swiss National Science Foundation for generously funding this research with Project 146702 "Architectural Process Models of Decision Making".

## References

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York: Oxford University Press.
Borst, J. P., \& Anderson, J. R. (2015). Using the ACT-R Cognitive Architecture in combination with fMRI data. In An introduction to model-based cognitive neuroscience (pp. 339-352). Springer New York.
Bröder, A., \& Schiffer, S. (2003). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. Journal of Experimental Psychology: General, 132, 277-293.
Bröder, A. (2012). The quest for take the best - Insights and outlooks from experimental research. In P. Todd, G. Gigerenzer, \& the ABC Research Group, Ecological rationality: Intelligence in the world (pp. 216-240), New York: Oxford University Press.
Dimov, C. M., Marewski, J. N., \& Schooler, L. J. (2013). Constraining ACT-R models of decision strategies: An experimental paradigm. In M. Knauff, M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 2201-2206). Austin, TX: Cognitive Science Society.
Fechner, H. B., Pachur, T., Schooler, L. J., Mehlhorn, K., Battal, C., Volz, K. G., \& Borst, J. P. (2016). Strategies for memory-based decision making: Modeling behavioral and neural signatures within a cognitive architecture. Cognition, 157, 77-99.

Gigerenzer, G., \& Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. Topics in Cognitive Science, 1, 107-143.
Gigerenzer, G., \& Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 104, 650-669.
Gigerenzer, G., \& Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigerenzer, P. M. Todd, \& the ABC Research Group, Simple heuristics that make us smart (pp. 75-95). New York: Oxford University Press.
Gigerenzer, G., Hoffrage, U., \& Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco-Watkins, \& Thomas (2008). Psychological Review, 115, 230-239.
Gigerenzer, G., Hoffrage, U., \& Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. Psychological Review, 98, 506-528.
Gigerenzer, G., Todd, P. M., \& the ABC Research Group. (1999). Simple heuristics that make us smart. New York: Oxford University Press.
Glöckner A., \& Betsch T. (2008a). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. Judgment and Decision Making, 3, 215-228.
Glöckner, A., \& Bröder, A. (2011). Processing of recognition information and additional cues: based analysis of choice, confidence, and response time. Judgment and Decision Making, 6, 23-42.
Goldstein, D. G., \& Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. Psychological Review, 109, 75-90.
Gonzalez, C., Lerch, J. F., \& Lebiere, C. (2003). Instancebased learning in dynamic decision making. Cognitive Science, 27, 591-635.
Hertwig, R., Herzog, S. M., Schooler, L. J., \& Reimer, T. (2008). Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 1191-1206.
Johansen, M. K., \& Kruschke, J. K. (2005). Category representation for classification and feature inference. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 1433-1458.
Juslin, P., \& Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. Cognitive Science, 26, 563-607.
Lee, M. D., \& Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the 'take the
best' and the 'rational' models. Psychonomic Bulletin \& Review, 11, 343-352.
Logan, G. D. (1988). Toward an instance theory of automatization. Psychological Review, 95, 492-527.
Marewski J. N. \& Mehlhorn K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. Judgment and Decision Making, 6, 439-519.
Marewski, J. N., \& Schooler, L. J. (2011). Cognitive Niches: An ecological model of strategy selection. Psychological Review, 118, 393-437.
Marewski, J. N., Schooler, L. J., \& Gigerenzer, G. (2010). Five principles for studying people's use of heuristics. Acta Psychologica Sinica, 42, 72-87.
Nellen, S. (2003). The use of the "take-the-best" heuristic under different conditions, modelled with ACT-R. In F. Detje, D. Dörner, \& H. Schaub (Eds.), Proceedings of the fifth international conference on cognitive modelling (pp. 171-176). Bamberg, Germany: Universitätsverlag Bamberg.
Newell, B. R. (2005). Re-visions of rationality? Trends in Cognitive Sciences, 9, 11-15.
Newell, B. R., \& Bröder, A. (2008). Cognitive processes, models and metaphors in decision research. Judgment and Decision Making, 3, 195-204.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 104114.

Salvucci, D. D., \& Anderson, J. R. (2001). Integrating analogical mapping and general problem solving: The path-mapping theory. Cognitive Science, 25, 67-110.
Salvucci, D. D., \& Taatgen, N. A. (2008). Threaded cognition: an integrated theory of concurrent multitasking. Psychological Review, 115, 101-130.
Schooler, L. J., \& Hertwig, R. (2005). How forgetting aids heuristic inference. Psychological Review, 112, 610628.

Schneider, D. W., \& Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. Cognitive Psychology, 64, 127-160.
Taatgen, N. A., van Rijn, H., \& Anderson, J. (2007). An integrated theory of prospective time interval estimation: the role of cognition, attention, and learning. Psychological Review, 114, 577-598.
Thomas, R. P., Dougherty, M. R., Sprenger, A. M., \& Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. Psychological Review, 115, 155185.

Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological Review, 79, 281-299.

# Talking to Ourselves to Engage Control? Testing Developmental Relations Between Self-directed Speech, Cognitive Control and Talkativeness 

Sabine Doebel (sabine.doebel@colorado.edu), Cleo Andersen-Green (cleo.andersengreen@colorado.edu)<br>and Yuko Munakata (munakata@colorado.edu)<br>Department of Psychology and Neuroscience, University of Colorado Boulder, Muenzinger Psychology 345 UCB<br>Boulder, CO 80309 USA


#### Abstract

Is self-directed speech critical to cognitive processes supporting complex, goal-directed behavior? If so, how? An influential developmental hypothesis is that children talk to themselves to support cognitive control processes, and that with age this speech becomes increasingly covert and strategic. However, while many studies suggest language supports cognitive control, evidence that self-directed speech gradually internalizes has been mixed. Moreover, extraneous factors that could co-vary with self-directed speech, age, and cognitive control, such as talkativeness, have not been systematically tested. In this cross-sectional study of 865 - to 7 -year-old children we measured overt, partially covert, inner, and strategic speech on four cognitive tasks, along with task performance and child talkativeness. We did not find consistent evidence that self-directed speech changes with age; however, we did find consistent associations between self-directed speech and talkativeness. Partially covert and strategic speech predicted performance on one task, and inner speech was implicated on another. Self-directed speech tended to correlate across tasks, and these correlations held controlling for talkativeness. Taken together, these findings suggest 5- to 7 -year-old children may use different forms of self-directed speech to support cognitive control, and that the form this speech takes depends in part on factors beyond age, such as the cognitive demands of a task and child characteristics like talkativeness.


Keywords: cognitive control; executive functions; selfdirected speech; language and thought

## Introduction

What role does language play in our ability to flexibly override impulses and achieve goals? An influential developmental hypothesis is that language is key to the emergence and exercise of cognitive processes supporting goal-directed behavior (Luria, 1961; Vygotsky 1934/2012; Winsler, Fernyhough, \& Montero, 2009). On this view, children's control processes are initially supported by the speech of others (e.g., parents and teachers), and later by children's own external speech, which is gradually internalized as inner speech (i.e., verbal thought) during childhood. Self-directed speech is thought to change qualitatively with internalization (e.g., becoming more condensed), allowing it to more effectively support cognitive control (Alderson-Day \& Fernyhough, 2015; Vygotsky 1934/2012).

This hypothesis fits with a large body of research indicating that language supports cognitive control across development. Children use their own speech to support many aspects of cognitive control, including planning (Al-

Namlah, Fernyhough, \& Meins, 2006; Fernyhough \& Fradley, 2005;), working memory (Al-Namlah et al., 2006; Flavell, Beach, \& Chinsky, 1966), and task switching (Karbach \& Kray, 2007). Moreover, linguistic interventions in which labels or other kinds of linguistic input are provided to children have been found to support cognitive control performance both in the moment (e.g., Kray, Eber, \& Karbach, 2008) and in the longer-term (Doebel, Dickerson, Hoover, \& Munakata, 2017; Doebel \& Zelazo, 2016). Experiments using articulatory suppression during cognitive tasks suggest older children and adults use inner speech when engaging cognitive control (e.g., Emerson \& Miyake, 2003; Kray, et al., 2008).

Key questions remain concerning the extent to which selfdirected speech changes with age, and the kinds of speech children use to support cognitive control. Evidence for the hypothesis that self-directed speech gradually internalizes has been mixed. For example, while some studies have found that overt, task-relevant speech decreases with age (Winsler \& Naglieri, 2003), others have not (Al-Namlah, et al., 2006; Flavell et al., 1966). And external forms of selfdirected speech do not always predict performance (e.g., Doebel, et al., 2017; Winsler \& Naglieri, 2003). Moreover, key variables that might account for the presence or absence of different forms of self-directed speech have not been systematically examined. For example, how talkative a child is may co-vary with age, performance, and self-directed speech, and thus could explain any relations found among these variables. Consistent with this idea, previous work has found correlations between self-directed speech and social speech/talkativeness (Al-Namlah et al., 2006; Fernyhough \& Fradley, 2005).

Gaining further insight into factors that predict selfdirected speech in childhood is critical both to understanding the role of language in cognitive control and how it can be improved in those who struggle with it. For example, if self-directed speech does generally internalize with age across a particular age window, this could suggest that training children to internalize their speech might help them better engage control.

The current study aimed to clarify relations between selfdirected speech, age, performance, and talkativeness in children 5 to 7 years of age, a developmental period posited to reflect key transitions in self-directed speech (Gathercole, 1998; Winsler \& Naglieri, 2003; Winsler et al., 2009). We assessed children's use of task-relevant overt, partially covert, inner, and strategic speech during four
cognitive tasks tapping control processes. The study evaluated two contrasting hypotheses. If self-directed speech becomes more internalized and strategic with age, then age-related changes in speech should be found across tasks, and self-directed speech should be associated with cognitive performance. However, if self-directed speech does not generally change across childhood and manifests differently depending on task demands and child characteristics, then inter-task correlations among selfdirected speech indices may be present, and possibly correlated with talkativeness, but relations with age should be less consistent.

## Method

## Participants

Eight-six 5- to 7-year-old children $\left(M_{\text {age }}=5.99\right.$ years $S D_{\text {age }}$ $=.61$, range $=5.0-7.1$; females $=47$ ) were recruited from a database of families who had previously indicated interest in participating in research. Four additional children were excluded from the study due to uncooperativeness $(n=3)$ and developmental delay $(n=1)$. Some children did not complete all tasks due to failure to demonstrate understanding during practice or uncooperativeness. In total, $84,83,72$, and 76 children completed the delayed serial recall task, selective attention task, Tower of London, and the immediate serial recall task, respectively. Most participants ( $>90 \%$ ) had at least one parent with a four-year college degree and were Caucasian and non-Hispanic.

## Measures

Children completed four cognitive measures across two test sessions. The tasks were completed in the following order: delayed serial recall, selective attention, Tower of London, and immediate serial recall. A fixed order was used to minimize variation between subjects in task performance due to differences in order (Friedman, et al., 2008). The first three measures were used to assess external self-directed speech in addition to cognitive performance. The last measure was used to index inner speech.

Delayed Serial Recall Task (adapted from Flavell, Beach, \& Chinsky, 1966; Fig. 1) Children were presented with pictures of objects serially on a computer screen, and, after an eight second delay, were asked to recall the order in which they were presented. At test, the three items were presented together in a new order, and children had to point to the pictures in the order that they saw them. Following three practice trials, children completed 10 test trials.

Selective Attention Task (adapted from Manfra \& Winsler, 2006; Fig. 2) Children were shown a page of three pictures that matched on one of three dimensions (shape, color, number), and were asked to search a box for a picture card that reflected the matching dimension. The box contained 18 picture cards depicting a single dimensional value (e.g., a purple splotch or a silhouette of a star).

Following three practice trials, children completed 12 test trials.


Figure 1: Delayed serial recall task.

Tower of London planning task (adapted from Fernyhough \& Fradley, 2005; Fig. 3) Children were presented with two apparatuses, each of which had three wooden pegs of different lengths and three colored wooden spheres on the pegs. The spheres were configured in a different arrangement on each apparatus, and children were instructed to make one apparatus look like the other in as few moves as possible. They were also instructed that they could only move one sphere at a time and had to keep all spheres on the pegs (i.e., not holding a sphere in their hand while making moves with another sphere). Children completed six trials in total, half of which could be completed in three moves, and the other half of which could be completed in four moves. Performance was indexed by the total number of moves children made in excess of the minimum number required, divided by the total number of trials completed. If children broke rules or asked to start over, the trial was restarted. Only the final attempt at a trial and trials that were successfully completed were included in our analyses.

Immediate Serial Recall (adapted from Al-Namlah et al., 2006; Tam, Jarrold, Baddeley, \& Sabatos-DeVito, 2010) This task was identical to the delayed serial recall task except that 1) there was no delay between the stimuli presentation and the test phase; 2) children were instructed not to label the pictures overtly while they were being initially presented (Al-Namlah et al., 2006); and 3) children completed two ten-trial blocks instead of one: a phonologically similar block and a phonologically dissimilar block. The phonologically dissimilar block involved the same items presented in the delayed serial recall task. The phonologically similar block involved items that had similar-sounding names (e.g., clock, clown, cat). Inner speech was indexed as the accuracy rate on the phonologically similar block subtracted from the accuracy rate on the phonologically dissimilar block, with the
expectation that children who used inner speech (i.e., verbal coding of the to-be-remembered objects) would perform worse on the phonologically similar block because verbal coding would make the items more confusable.


Figure 2: Selective attention task.


Figure 3: Tower of London planning task.

Talkativeness Parents were asked to rate their child's level of talkativeness with people he or she does not know on a 5point scale, with 1 indicating that the child is not at all talkative, and 5 indicating that the child is very talkative. This approach was adapted from prior work in which teachers were asked to rate children's general talkativeness (Fernyhough \& Fradley, 2005). We opted to ask more specifically about talkativeness with unfamiliar people in order to reduce the likelihood that parents' evaluations would reflect how talkative their child is at home, which we expected would result in a less sensitive measure.

## Self-directed Speech Coding

Our coding scheme was based on prior work in this area (Winsler \& Naglieri, 2003; Flavell et al., 1966). Speech during each trial of each cognitive task was coded from videos by a research assistant who was blind to the experimental hypotheses. Ten percent of the videos were coded by a second blinded research assistant and inter-rater agreement was high, $r \mathrm{~s}>.85$. Each task was coded for nonsocial overt speech, defined as normal volume speech not directed at another person that could support task performance (rather than meta-comments about the task or stimuli, or comments unrelated to the task) and partially covert speech, such as whispering, muttering, and lip movement.

In addition, task-specific speech strategies were coded. On the delayed serial recall task, labeling at the onset of a trial (when the stimuli were being presented) and rehearsal
(during the presentation and test interval) were coded. On the selective attention task, labeling the matching dimension (e.g., "They're all purple") at the onset of or during the test trial was coded. On the Tower of London task, labeling the sphere and the location the child was placing or planning to place it was coded (e.g., "This one goes here for now"). For analyses, the number of trials on which a child used each coded form of speech was scaled by the number of trials the child completed.

## Results

## Self-directed Speech Variability and Frequency

As expected, all cognitive tasks elicited some self-directed speech (Table 1), and there was variability across tasks in the kinds of speech children used. However, numerous children did not use overt or partially covert self-directed speech on the tasks: 20 of 84 on the delayed serial recall task; 21 of 84 on the selective attention task; and 29 of 77 on the Tower of London task. This is comparable to rates of self-directed speech found in prior work (Fernyhough \& Fradley, 2005; Flavell et al., 1966; Manfra \& Winsler, 2006; Winsler \& Naglieri, 2003). Children showed evidence of inner speech on the immediate recall task, performing significantly worse on the phonologically similar block ( $M_{\text {accuracy }}=65 \%, S D=22 \%$ ) than on the phonologically dissimilar block ( $M_{\text {accuracy }}=72 \%, \mathrm{SD}=24 \%$ ), $M_{\text {diff }}=.07$ $S D_{\text {diff }}=.17, t(76)=3.74, p<.001$, consistent with previous findings (Al-Namlah et al., 2006; Tam, et al., 2010).

Table 1: Prevalence of Different Forms of External Self-directed Speech Across Measures

| Self-directed Speech <br> Task and Index | Mean \% of <br> trials on which <br> speech used | $N$ <br> children <br> using <br> speech |
| :--- | :---: | :---: |
| Delayed serial recall | $.30(.42)$ | 35 |
| Overt speech | $.31(.35)$ | 52 |
| Partially covert speech | $.27(.35)$ | 42 |
| Rehearsal | $.44(.44)$ | 48 |
| Labeling | $.24(.35)$ | 38 |
| Selective attention | $.28(28)$ | 58 |
| Overt speech | $.28(.34)$ | 47 |
| Partially covert speech |  |  |
| Labeling | $.16(.27)$ | 28 |
| Tower of London | $.22(.26)$ | 44 |
| Overt speech | $.15(.24)$ | 28 |
| Partially covert speech |  |  |
| Labeling |  |  |

## Relations Between Self-directed Speech and Age

We found minimal support for the hypothesis that selfdirected speech changes with age (Table 2). Zero-order correlations indicated that only partially covert speech on the selective attention task was related to age, such that as children got older they used less partially covert speech

Table 2: Correlations Between Self-directed Speech Indices and Age, Talkativeness, and Task Performance

| Self-directed <br> Speech Measure <br> and Index | Age | Talkativeness | Task <br> Score |
| :--- | :--- | :--- | :--- |
| Delayed recall | -.17 | $.19^{\wedge}$ | .04 |
| Overt | .13 | .14 | $.33^{* *}$ |
| Partially covert | .03 | $-.32^{* *}$ | $-.28^{*}$ |
| No speech | .18 | .09 | $.31^{* *}$ |
| Rehearsal | -.13 | $.28^{*}$ | $.19^{\wedge}$ |
| Labeling |  |  |  |
| Selective attention | -.03 | $.32^{* *}$ | .01 |
| $\quad$ Overt | $-.28^{*}$ | .16 | -.06 |
| Partially covert | .18 | $-.32^{* *}$ | .06 |
| No speech | -.08 | $.31^{* *}$ | .02 |
| Labeling |  | .17 | .14 |
| Tower of London | .08 | .17 |  |
| $\quad$ Overt | . .03 | $.28^{*}$ | .05 |
| Partially covert | -.10 | $-.20^{\wedge}$ | -.10 |
| No speech | .05 | $.26^{*}$ | -.07 |
| Labeling | .00 | .02 | - |
| Immediate recall |  |  |  |

(Table 2). All other age/self-directed speech correlations were not significant.

These analyses were followed up with linear regressions predicting age from self-directed speech, with related speech indices simultaneously included in models to control for one another's effects, and the results were unchanged. On the selective attention task, partially covert speech remained a significant predictor, $B=-7.36, S E=2.82, t=$ $2.60, p=.01$, whereas overt speech was not, $t<1, p>.25$. On the delayed serial recall task and Tower of London, neither overt nor partially covert speech were associated with age, $t \mathrm{~s}<1.38, p \mathrm{~s}>.17$. Similarly, neither rehearsal nor labeling changed with age on the delayed serial recall task, $t \mathrm{~s}>1.5, p \mathrm{~s}>.13$.

## Relations Between Self-directed Speech and Talkativeness

Across tasks, children who tended to use external forms of self-directed speech also tended to be more talkative (Table 2). Talkativeness was correlated with overt speech on all tasks, partially covert speech on the Tower of London (but not the delayed serial recall or selective attention tasks), and labeling (but not rehearsal) on the delayed serial recall and selective attention tasks. Inner speech on the immediate recall task was not associated with talkativeness. Talkativeness was not associated with age, $r=-.06, p>.25$, nor was it associated with performance on any of the tasks, $r \mathrm{~s}<.16, p \mathrm{~s}>.14$. As such, it was not included as a control variable in any models involving these factors.

## Relations Between Self-directed Speech and Performance

Children performed well on the four tasks (Table 3) and examination of histograms did not reveal floor or ceiling effects. Self-directed speech predicted performance on the delayed serial recall task. Zero-order correlations indicate partially covert speech and rehearsal were associated with performance on the delayed serial recall task, and also indicated a marginally significant association between labeling and performance on that task (Table 2). These findings were confirmed with linear models. Partially covert speech was a significant predictor of performance on delayed serial recall, controlling for overt speech, $B=.22$, $S E=.07, t=3.2, p=.002$, whereas overt speech was not predictive when controlling for partially covert speech, $t<1$, $p>.25$. Similarly, rehearsal was a significant predictor of performance on the delayed serial recall task, controlling for labeling, $B=.18, S E=.07, t=2.62, \mathrm{p}=.02$, consistent with prior work (Flavell et al., 1966). There was also a nonsignificant trend such that labeling tended to predict performance on the delayed serial recall task, controlling for rehearsal, $B=.09, t=1.59, p=.11$. However, self-directed speech on the selective attention task and Tower of London was not predictive of performance on those tasks, $t \mathrm{~s}<1, p \mathrm{~s}$ $>.25$.

Table 3: Performance on Cognitive Measures

| Measure | $M$ | $S D$ | Range |
| :--- | :--- | :--- | :--- |
| Delayed recall | .75 | .22 | $0-1$ |
| Selective attention | .94 | .13 | $.33-1$ |
| Tower of London | .65 | 1.48 | $0-11$ |
| Immediate recall - <br> phonologically similar block <br> Immediate recall - <br> $\quad .65$ | .22 | $0-1$ |  |
| $\quad$ phonologically dissimilar block | .72 | .24 | $0-1$ |

## Correlations Between Self-directed Speech Indices Within and Across Tasks

Many self-directed speech indices were correlated across tasks (Table 3). For example, children who used partially covert speech on the delayed serial recall task also tended to use it on the selective attention and Tower of London tasks, and children who used rehearsal on the delayed serial recall task tended to label on the selective attention task. These findings generally held when controlling for talkativeness, with the exception that some of the correlations between partially covert speech on Tower of London and other indices (partially covert speech and labeling on delayed serial recall, and labeling on selective attention) were no longer statistically significant, $r \mathrm{~s}<.18, p \mathrm{~s}>.10$.

We also found many correlations between self-directed speech indices within tasks. Some correlations were very strong, suggesting that certain strategies tend to be expressed more or less covertly.

Table 3: Correlations Between Self-directed Speech Indices Within and Across Cognitive Tasks

| Task and Speech Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 1. DR overt |  |  |  |  |  |  |  |  |  |  |
| 2. DR partially covert | -.17 |  |  |  |  |  |  |  |  |  |
| 3. DR rehearsal | .06 | $.74^{* * *}$ |  |  |  |  |  |  |  |  |
| 4. DR labeling | $.82^{* * *}$ | $.26^{*}$ | $.27^{*}$ |  |  |  |  |  |  |  |
| 5. SA overt | .03 | $.25^{*}$ | $.21^{\wedge}$ | .13 |  |  |  |  |  |  |
| 6. SA partially covert | .03 | $.28^{*}$ | .09 | $.21^{\wedge}$ | .02 |  |  |  |  |  |
| 7. SA labeling | .00 | $.33^{* *}$ | $.24^{*}$ | $.19^{\wedge}$ | $.87^{* * *}$ | $.30^{* *}$ |  |  |  |  |
| 8. TOL overt | -.08 | .08 | .00 | -.08 | $.55^{* * *}$ | .08 | $.37^{* *}$ |  |  |  |
| 9. TOL partially covert | .10 | $.20^{\wedge}$ | .16 | $.22^{\wedge}$ | $.27^{*}$ | $.24^{*}$ | $.23^{*}$ | $.33^{* *}$ |  |  |
| 10. TOL labeling | -.09 | .13 | .09 | -.03 | $.33^{*}$ | $.22^{\wedge}$ | $.29^{*}$ | $.53^{* *}$ | $.66^{* *}$ |  |
| 11. IR difference score | .06 | .07 | .00 | .15 | .16 | .12 | .12 | .10 | .04 | -.09 |

Note. $\mathrm{DR}=$ Delayed recall task; $\mathrm{SA}=$ Selective attention task; TOL $=$ Tower of London; IR $=$ Immediate recall task
${ }^{*} p<.05 ; * * p<.01 ;{ }^{* * *} p<.001 ;{ }^{\wedge} p<.10$.

## Discussion

The current study provides several new findings related to the role of self-directed speech in developing cognitive control. We did not find evidence that self-directed speech undergoes a general internalization process with age. Instead, our findings suggest that the format of self-directed speech may depend in part on other factors like child talkativeness and the specific cognitive demands of a task. We found that 5- to 7-year-old children used inner, partially covert, and strategic speech while engaging cognitive control on different tasks, and that more overt forms of speech tended to be related to talkativeness. The finding that external forms of self-directed speech predicted performance on the delayed serial recall task but not the selective attention and Tower of London tasks suggests that children may have been supporting cognitive performance on the latter tasks with internalized speech. The delayed serial recall task likely had the highest working memory demands of all the tasks (given the need to maintain three items in mind in a particular order across time) and as such, external speech may have been necessary to support performance. Conversely, the working memory demands of the selective attention and Tower of London tasks may have been lower, and thus inner speech may have been sufficient to support performance on those tasks. For example, on the selective attention task children needed to identify a common dimension among three objects on a page and keep that dimension in mind to guide their searching, but they could always look back at the objects on the page to recall the dimension, and they only had to maintain one dimension in mind.

These findings are consistent with an alternative view of how linguistic input influences developing cognitive control, and highlight the possibility that inner speech may play a role in cognitive control from early in development. For example, teaching 5-year-old children labels that can be used to support cognitive control helps children later engage control; however, children's tendency to vocalize the labels
when engaging control does not predict performance (Doebel, et al., 2017), consistent with the possibility that children can rapidly internalize speech used to support control. Foundational cognitive control processes begin to develop very early in life (Munakata, 1998), and continue to develop rapidly in early childhood, between 3 and 5 years of age (Diamond, 2013). Internalized forms of self-directed speech could be critical to the emergence of these processes.

An alternative interpretation of our findings is that there are indeed robust age-related changes in self-directed speech, but that our age range and sample size were too constrained to detect them. For example, prior work has found age differences in the use of rehearsal to support serial recall when the age groups being compared were 5,7 and 10 years (Flavell et al., 1966), and that overt selfdirected speech decreases with age in a large sample aged 5 to 17 years (Winsler \& Naglieri, 2003). However, given the frequency and variability in speech use in the current study, and that our sample spanned an age range identified as a transition period in the use of self-directed speech (e.g., Winsler \& Naglieri, 2003; Winsler, et al., 2009), it was surprising that age was not a significant predictor of speech on most tasks. Another possibility is that age-related patterns only manifest when cognitive demands are high and inner speech is insufficient to maintain the goal representations guiding performance. Future experiments can test this by manipulating the maintenance demands in a task and testing associations between age and self-directed speech.

Our findings are correlational, leaving open alternative explanations for the relation between self-directed speech and cognitive control. For example, it is possible that developmental increases in cognitive control lead to changes in self-directed speech (and that self-directed speech is epiphenomenal). Or a third, unmeasured variable may explain the relation between self-directed speech and cognitive control. Experiments manipulating cognitive control and testing influences on self-directed speech, and vice versa, could address this causal issue.

Self-directed speech may be a good target for interventions to improve cognitive control. Given that cognitive control develops dramatically in early childhood and predicts success in life across a range of outcomes (e.g., academics, health, and wealth; Moffitt, et al., 2011), there has been great interest in developing effective interventions to improve it. However, results of interventions to date have been mixed (e.g., Melby-Lervag \& Hulme, 2013). One potential reason is that approaches to date have not optimally targeted the processes that support developing cognitive control. Training children to use different forms of self-directed speech to support cognitive performance, such as labeling, rehearsal, and partially covert speech, may be a fruitful approach to improving control in children. ${ }^{1}$

The current study advances knowledge on the role of selfdirected speech in cognitive control by suggesting that the kinds of speech children use to support cognitive performance in childhood may depend on a range of factors beyond age, such as child talkativeness and the cognitive demands of a task. Future work can further test how selfdirected speech relates to cognitive control and how it can be targeted in cognitive control interventions.

## Acknowledgments

This research was supported by the Eunice Kennedy Shriver National Institute of Child Health \& Human Development of the National Institutes of Health under Award F32HD079191.

## References

Alderson-Day, B., \& Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. Psychological Bulletin, 141, 931-965.
Al-Namlah, A. S., Fernyhough, C., \& Meins, E. (2006). Sociocultural influences on the development of verbal mediation: Private speech and phonological recoding in Saudi Arabian and British samples. Developmental Psychology, 42, 117-131.
Diamond, A. (2013). Executive functions. Annual Review of Psychology, 64, 135-168.
Doebel, S., Dickerson, J. P., Hoover, J. D., \& Munakata, Y. Using language to get ready: Labels help children engage proactive control. Manuscript submitted for publication.
Doebel, S., \& Zelazo, P. D. (2016). Seeing conflict and engaging control: Experience with contrastive language benefits executive function in preschoolers. Cognition, 157, 219-226.
Emerson, M. J., \& Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. Journal of Memory and Language, 48, 148168.

[^323]Fernyhough, C., \& Fradley, E. (2005). Private speech on an executive task: Relations with task difficulty and task performance. Cognitive Development, 20, 103-120.
Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., \& Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. Journal of Experimental Psychology: General, 137, 201-225.
Flavell, J. H., Beach, D. R., \& Chinsky, J. M. (1966). Spontaneous verbal rehearsal in a memory task as a function of age. Child Development, 283-299.
Gathercole, S. E. (1998). The development of memory. Journal of Child Psychology and Psychiatry, 39, 3-27.
Karbach, J., \& Kray, J. (2007). Developmental changes in switching between mental task sets: The influence of verbal labeling in childhood. Journal of Cognition and Development, 8, 205-236.
Keeney, T. J., Cannizzo, S. R., \& Flavell, J. H. (1967). Spontaneous and induced verbal rehearsal in a recall task. Child Development, 953-966.
Kray, J., Eber, J., \& Karbach, J. (2008). Verbal selfinstructions in task switching: a compensatory tool for action-control deficits in childhood and old age? Developmental Science, 11, 223-236.
Luria, A. R. (1961). The role of speech in the regulation of normal and abnormal behavior. London: Pergamon Press.
Manfra, L., \& Winsler, A. (2006). Preschool children's awareness of private speech. International Journal of Behavioral Development, 30, 537-549.
Melby-Lervåg, M., \& Hulme, C. (2013). Is working memory training effective? A meta-analytic review. Developmental Psychology, 49, 270-291.
Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., ... \& Sears, M. R. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences, 108, 2693-2698.
Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the AB task. Developmental Science, 1, 161-184.
Tam, H., Jarrold, C., Baddeley, A. D., \& Sabatos-DeVito, M. (2010). The development of memory maintenance: Children's use of phonological rehearsal and attentional refreshment in working memory tasks. Journal of Experimental Child Psychology, 107, 306-324.
Vygotsky, L. (1934/2012). Thought and language. MIT press.
Winsler, A., Fernyhough, C., \& Montero, I. (2009). Private speech, executive functioning, and the development of verbal self-regulation. Cambridge University Press.
Winsler, A., \& Naglieri, J. (2003). Overt and covert verbal problem-solving strategies: Developmental trends in use, awareness, and relations with task performance in children aged 5 to 17. Child Development, 74, 659-678.

# The Interplay Between Self-evaluation, Goal Orientation, and Self-efficacy on Performance and Learning 

Kit S. Double (kit.double@sydney.edu.au)<br>Department of Psychology, University of Sydney<br>Sydney, NSW, Australia<br>Damian P. Birney (damian.birney@sydney.edu.au)<br>Department of Psychology, University of Sydney<br>Sydney, NSW, Australia


#### Abstract

Objective Self-awareness Theory (Duval \& Wicklund, 1972) proposes that self-evaluation increases an individual's awareness of any discrepancy between their current performance and an internal goal. In the current study we prompted self-evaluation throughout an intelligence test (Analysis-Synthesis Test - AST) using confidence ratings (CR). AST performance, the extent to which participants incidentally learnt task-relevant rules (learning rules was unnecessary because they were provided), self-efficacy, and goals, were assessed. The results indicated an effect of performing CR on both performance and rule learning, but the effect depended on self-efficacy. Compared to matched controls ( $n=45$ ), participants who performed CR ( $n=41$ ) and had high self-efficacy performed better on the AST but learnt fewer rules. Performing $C R$ had no effect on participants low in self-efficacy. This suggests that selfevaluation interacts with self-efficacy to modify participants' goals, specifically CR appear to shift individuals high in self-efficacy from a mastery goal to a performance goal.


Keywords: reasoning, intelligence, reactivity, goal orientation, self-efficacy

## The Interplay Between Self-evaluation, Goal Orientation, and Self-efficacy on Performance and Learning

Accurate self-knowledge is a highly valued attribute and important to everyday functioning. Awareness of our own abilities and past performance facilitates realistic goal selection and allows us to better direct our future behavior. There is some evidence that self-evaluation occurs almost continuously when we perform a demanding cognitive activity, and this self-evaluation may occur both spontaneously and unconsciously. Self-evaluation in this context is often referred to as performance monitoring or alternatively error detection. Performance monitoring is vital to learning outcomes as it allows the learner to identify errors so that they can avoid repeating them in the future (Yeung \& Summerfield, 2012). Performance monitoring is also important for the effective allocation of cognitive
resources (Carter et al., 1998). Accurate performance monitoring is central to an individual's ability to regulate their own cognitive behavior (Nelson \& Dunlosky, 1991) and, in particular, effectively make decisions about study time (e.g. Metcalfe \& Finn, 2008; Son \& Metcalfe, 2000). Furthermore, learners need to continually evaluate the effectiveness of different study activities on their learning in order to select the best possible study behaviors (Flavell, 1979).

Given the importance of accurate self-knowledge, individuals such as students and employees are often encouraged to self-reflect and self-evaluate their performance so that they can better identify their strengths weaknesses and detect issues or errors (Carver \& Scheier, 2001). However, there is little direct evidence that selfevaluation leads to more accurate self-knowledge (Silvia \& Gendolla, 2001). Indeed, self-assessment is often systematically inaccurate and flawed, which can often lead to negative outcomes and ineffective decision making (Dunning, Heath, \& Suls, 2004). Objective self-awareness theory (OST; Duval \& Wicklund, 1972, 1973) contends that self-focused attention does not necessarily lead to accurate self-knowledge, instead it directs individuals' attention to discrepancies between their current performance or behavior and their internal standards or goals (Silvia \& Duval, 2001), referred to here as goal discrepancies. OST theory argues that when an individual becomes aware of a goal discrepancy, they either modify their behavior to bring it in line with their goal, modify their goal, or disengage from the activity, which reduces their awareness of the goal discrepancy (Silvia \& Duval, 2001). For example, upon reflecting on their studying, a student may determine that they have not been working hard enough to prepare for an upcoming exam, which may lead them to study more or set a more modest goal for the exam, alternatively they may try to distract themselves to avoid thinking about studying for the exam.

Which of these three strategies an individual adopts is largely determined by their self-efficacy, if an individual has high self-efficacy they will believe that they can improve their performance to match their goal and will act accordingly, whereas individuals low in self-efficacy may lower their goals or disengage from the task because they do not believe they can implement the necessary improvements
in their performance. We can conceptualize these expected changes in terms of the goals we expect participants to adopt. We hypothesize that self-evaluation will lead high self-efficacy participants to focus on improving their performance, whereas low self-efficacy participants will either disengage from the task or lower their goal, both of which are likely to impair their performance on the task. These hypotheses therefore relate to changes in a participants goal orientation.

Goal orientation is a well-studied concept that broadly concerns the distinction between mastery goals that concern development, improvement, and learning compared with performance goals which prioritize performing well and demonstrating ability (VandeWalle, 1997). Generally speaking mastery goals are considered advantageous for a range of outcomes including academic engagement (Ames \& Archer, 1988; Pintrich, 2000; Wolters, 2004), job performance (Janssen \& Van Yperen, 2004; VandeWalle, Brown, Cron, \& Slocum Jr, 1999), and cognitive ability (Eison, 1981). Although some have conceptualized goal orientation as a trait or at least a quasi-trait like concept (DeShon \& Gillespie, 2005), goal orientation is domain specific and can equally be considered as state-like (VandeWalle, Cron, \& Slocum Jr, 2001). The goal orientation that an individual selects for a particular situation is somewhat determined by an individual's selfefficacy (Diseth, 2011). Individuals with high self-efficacy tent to pursue their goals with more effort and endeavor to develop from the experience of goal pursuit (DeGeest \& Brown, 2011). Given this, a number of studies have expectedly shown a positive relationship between high selfefficacy and mastery goal orientation (Bell \& Kozlowski, 2002; Diseth, 2011).

H1: In the control group (without self-evaluation), high self-efficacy participants will display mastery goal orientated behavior, whereas low self-efficacy participants will show performance goal orientated behavior.

The rationale behind encouraging self-evaluation in schools and workplaces has often been that they will elicit greater effort and goal driven behavior. However, as previously mentioned, when an individual becomes aware of a goal discrepancy they can either change their behavior or alternatively they can modify their standard/goal. With regards to modifying a goal, this may occur quantitatively (e.g. changing from a goal of a 75 on an exam to a 65) or qualitatively changing the nature of the goal (e.g. changing from try to master the material in a class to focusing on passing the exam). Based on this OST framework we expect that self-evaluation will lead participants with high selfefficacy to shift to a performance orientation in order to improve their performance and reduce the goal discrepancy. As previously stated, participants with low self-efficacy are likely to naturally adopt a performance goal-orientation and therefore self-evaluation should have little effect on the goal-orientation they adopt.

There is some previous evidence that self-evaluative prompts lead to such changes in goal orientation. A recent study by Mitchum, Kelley, and Fox (2016) using a wordpair learning paradigm found that, if the list of word-pairs participants were learning contained both difficult and easy items, then performing judgments of learning (i.e. rating how likely it is that they will recall a word-pair on a later test) resulted in participants spending more of their study time on the easier items rather than the difficult items. This resulted in participants who performed judgments of learning recalling fewer word-pairs on a subsequent test. The authors suggested that in the presence of both easy and hard items, judgments of learning make participants aware of the fact that they will inevitably fail to remember all of the words on the list, so they adopt a performance orientation rather than a mastery orientation and over study the easier items.

H2: Self-evaluation will result in participants adopting a performance goal orientation, regardless of their level of self-efficacy

In the current study we induce self-focused attention by asking participants to self-evaluate their performance after each item on an intelligence test, by providing confidence ratings (CR). CR direct participants' attention to their current subjective belief in their performance by require participants to reflect on, evaluate, and quantify their performance. A previous study which examined the effects of eliciting confidence ratings from participants while they completed an intelligence test found that participants who provided CR, performed better than participants who performed the task without providing ratings (Double \& Birney, 2017). Crucially, a subsequent experiment showed that this effect depended on the confidence/self-efficacy of participants, with CR facilitating the performance of high self-confidence participants but hindering the performance of low self-confidence participants.

H3: Performing CR will facilitate performance in high-self-efficacy participants and impair performance in low self-efficacy participants

In the current experiment participants completed a deliberately difficult task so that they would be likely to experience a performance discrepancy when they selfevaluated their performance. Of primary interest is whether this self-evaluation causes participants to improve their performance and/or change their goal orientation.

## Method

## Participants

A community sample of 85 participants ( $80 \%$ female) was recruited using an advertisement placed in a newsletter of the Australian Broadcasting Corporation as part of a research partnership with the University of Sydney $\left(M_{\text {age }}=\right.$
63.75, $\mathrm{SD}=9.83$ ). Participants received no remuneration for participating in the study. Participants were randomly allocated to the confidence ratings group (CR group; $n=42$ ) or a control group that did not provide confidence ratings (No-CR group; $n=43$ ).

## Materials and procedure

Participants completed the following measures online from their own personal computers using Qualtrics (Qualtrics, 2015) and Inquisit (Inquisit, 2016). All materials were programmed to display in a standardized fashion.

Analysis Synthesis Task (AST; Woodcock, McGrew, Mather, \& Schrank, 2001): A modified computerized version of the AST tasks was performed by participants. The AST task requires participants to solve problems by combining a series of tiles using a set of simple rules (e.g. a red triangle and purple square make blue circle). The rules are displayed continuously in the form of a key at the top of the screen. Figure 1 presents a typical question. One tile is blank and participants must decide which tile correctly fills the blank. Participants could combine any two tiles that are next to each other either horizontally or vertically.

As the task was expected to be difficult for a community sample, participants were given a series of practice items and were allowed to continue to practice until they felt they were ready to progress to the test phase (minimum of 12 practice items, maximum 36). The test block consisted of 20 items that were approximately ordered according to their difficulty. There were 5 rules in each of the practice and test blocks (different rules/stimuli were used in the practice and test phases).

Participants in the CR group were asked to rate their confidence that they answered the previous item correctly using a scale from $0-100 \%$. In order to reduce any response time effect on performance, participants in the control group were shown a blank screen for 2000 ms after each trial.


Figure 1: A typical question drawn from the AST task. that were used in the test block.

Self-efficacy: As self-efficacy is domain specific we used a particularly proximal measure of self-efficacy by having participants predict their score on the test block as a percentage after completing the practice block.

Rule Recall Test: After finishing the test block participants unexpectedly performed a recall test of the rules Participants had been given no prior warning that they would need to recall the rules on the later task and it was not necessary that participants memorize them as they had been displayed on the screen continuously during the AST task. The recall test asked participants questions such as "what color is the combination of the red triangle and blue circle?".

## Results

All data analysis was performed using $R$ version 3.2.3 ( R Core Team, 2015). Table 1 presents summary statistics for key study variables. The number of practice trials performed did not differ significantly between the CR group and the No-CR group, $F(1,83)=.17, p=.683$. Participants' initial predictions of their performance did not differ significantly between the CR group and the No-CR group, $F(1,83)=$ $.002, p=.967$.

Table 1: Descriptive statistics for key study variables.

| Variable | N | $M$ | SD |
| :--- | :---: | :---: | :---: |
| Predicted performance | 85 | 31.5 | 21.7 |
| Number of practice trials | 85 | 23.1 | 9.7 |
| AST practice score | 85 | 10.6 | 4.9 |
| AST test score | 85 | 10.3 | 2.7 |
| Rules recalled | 74 | 2.2 | 1.7 |

## AST Performance

Performance was analyzed using a linear regression model with number of correct items as the criterion variable. Experimental group and self-efficacy were entered as predictors along with the relevant interaction. In addition, as we were primarily interested in the moderating effect of self-efficacy, over and above ability, we included participants' practice scores and number of practice trails as covariates. Overall there was no main effect of experimental group, $\beta=.08, t(79)=.89, p=.378$. Self-efficacy was a significant negative predictor of performance, $\beta=-.27, t(79)$ $=2.02, p=.047$. Practice score and practice trial count were both positive predictors of test performance; $\beta=.68, t(79)=$ $7.19, p<.001$ and $\beta=.35, t(79)=4.03, p<.001$ respectively. Crucially, the group X predicted performance interaction was a significant predictor of test performance, $\beta$ $=.27, t(79)=2.07, p=.042$, see Figure 2 .
A simple slopes analysis indicated that self-efficacy was a significant negative predictor for the No-CR group, $\beta=-.18$, $t(79)=2.02, p=.047$, but not a significant predictor of performance for the CR group, $\beta=.06, t(79)=.74, p=.464$.

As shown in Figure 2, participants with high self-efficacy performed better in the CR group than the NO-CR group, whereas there was no group difference for participants with low self-efficacy.


Figure 2: Average number of correct items on the analysis synthesis task as a function of experimental group and predicted performance. The values used for high and low predicted performance are 1 standard deviation above and below the mean respectively. Error bars represent $\pm 1$ standard error of the mean.

## Rule Recall

11 participants did not complete the rule recall test and were therefore excluded from the analysis. Rule recall was analyzed using a second linear regression with number of correctly recalled rules as the criterion variable.

Experimental group, self-efficacy, and the interaction between the two were entered as predictor variables. In addition, to control for performance on the analysis synthesis task, AST test performance was entered as a covariate. Again there was no significant main effect of experimental group, $\beta=-.16, t(69)=1.54, p=.128$. Selfefficacy and AST performance were both significant positive predictors of rule recall performance, $\beta=.47, t(69)$ $=2.94, p=.004$ and $\beta=.33, t(69)=3.03, p=.003$ respectively. Again the hypothesized group X self-efficacy interaction was significant, $\beta=-.36, t(69)=2.25, p=.028$, see Figure 3. Simple slopes analysis indicated that selfefficacy was a significant positive predictor of recall performance for the No-CR group, $\beta=.31, t(69)=2.94, p=$ .004 , but not for the CR group, $\beta=-.01, t(69)=.11, p=$ .911. As shown in Figure 3, participants with high selfefficacy in the No-CR group outperformed all other groups in terms of rule recall.


Figure 3: Average number of rules recalled as a function of experimental group and predicted performance. The values used for high and low predicted performance are 1 standard deviation above and below the mean respectively. Error bars represent $\pm 1$ standard error of the mean.

## Discussion

Self-evaluation is often assumed to be an effective method to obtain accurate self-knowledge about one's abilities and performance. Organizations and educators often pursue formal and informal opportunities for feedback and evaluation, such as performance reviews, testing etc. These procedures may have many benefits such as improving communication, identifying ongoing issues, and providing feedback. However, in terms of the effect of self-evaluation on performance the current results reveal two important caveats in determining whether there is a benefit to performance/learning outcomes as a result of selfevaluation. The first is the importance of self-efficacy, our results show that self-evaluation improved the performance of participants with high self-efficacy but had no effect on participants with low self-efficacy. The second finding of note is that the effect of self-evaluation may depend on the nature of the outcome you are assessing. For participants high in self-efficacy, self-evaluation improved performance but impaired incidental learning, whereas there was no effect on either outcome for low self-efficacy participants.

According to OST, self-efficacy moderates the way in which we behave when confronted with goal discrepancies. The theory suggests that when high self-efficacy participants are confronted with goal discrepancies they work to improve their performance, whereas low selfefficacy participants may disengage from the task (thereby reducing awareness of the goal discrepancy). Our results conform to this general pattern with high self-efficacy participants improving their performance in response to CR
and low self-efficacy participants obtaining no benefit. This is in keeping with the previous finding by Double and Birney (2017), who demonstrated that performing CR during Raven's Progressive Matrices improved the task performance of high-confidence participants and impaired the performance of low-confidence participants. Our results have similarly found an asymmetry in the effect of CR on performance as a function of self-efficacy. Although we found no evidence of impaired performance in low selfefficacy participants, our results demonstrated that CR were beneficial only to participants high in self-efficacy. The difference in findings between this and Double and Birney (2017) in terms of low self-efficacy participants needs further exploration but may be a result of differences in how self-efficacy/self-confidence was assessed or the nature /difficulty of the task.

The current results suggest that high self-efficacy participants ordinarily adopt a mastery goal orientation, but shifted to a performance orientation when asked to perform CR. The finding that high self-efficacy participants tend to adopt a mastery goal is in keeping with the established relationship between goals and self-efficacy (Bell \& Kozlowski, 2002; DeGeest \& Brown, 2011; Diseth, 2011). Importantly, to our knowledge the results of the current study are the first to show that self-evaluation prompts high self-efficacy participants to adopt a performance orientation, which benefits performance but hinders incidental learning. When individuals high in self-efficacy are made aware of goal discrepancies they are likely to be motivated to reduce that discrepancy by attempting to improve their performance (Silvia \& Duval, 2001). The current results indicate that this focus on performance may come at the cost of a shift away from learning in line with the classic goal orientations paradigm (VandeWalle, 1997). It may be that, as a result of the evaluative nature of CR , they prompt high self-efficacy participants to direct attention to task relevant information and ignore task-irrelevant (rule) information. This suggests that appropriateness of using self-evaluation may depend on the outcomes of interest. In classrooms and workplaces self-evaluation has often become commonplace, but this may be problematic given the current body of evidence suggesting that a mastery orientation has many relative advantages in such settings (Janssen \& Van Yperen, 2004; Pintrich, 2000; Wolters, 2004). Although selfevaluation may be beneficial, the current results indicate that both self-efficacy and whether the valued outcomes are learning or performance based need to be considered before advocating self-evaluation.

Metacognitive interventions are often encouraged in the education literature and have obtained some positive results (Berardi-Coletta, Buyer, Dominowski, \& Rellinger, 1995; Desoete \& Roeyers, 2006). The current results, however, raise the possibility that such interventions are selectively benefiting students with high self-efficacy. Metacognitive prompts encourage individuals to monitor and evaluate their performance. Although the metacognitive literature has argued that such behaviors are important for error
monitoring, strategy selection and allocating cognitive resources, such behaviors also induce self-focused attention and may have an interactive effect with self-efficacy. Metacognitive prompts encourage individuals to monitor and evaluate their performance but do not necessarily provide a framework for doing so accurately. Although individuals are generally able to monitor their performance effectively, there are significant individual differences in the accuracy of such monitoring. It may be that in the current study, participants' evaluations of their own performance were shaped by their self-efficacy (i.e. high self-efficacy participants were likely to evaluate their performance positively and vice versa) and as such self-evaluation may reinforce existing beliefs in ability and thereby benefit only those who have a positive view of their own ability.

The current study has provided evidence that selfevaluative practice interact with self-efficacy to affect performance and incidental learning in an intelligence test. Given that self-evaluation is widely encouraged in both schools and workplaces, these results provide much needed research into the factors that affect the benefits of selfevaluation. The current results suggest that self-evaluation is beneficial to the performance of high self-efficacy individuals, but impairs incidental learning, most likely as a result of encouraging such individuals to adopt a performance goal orientation. For individuals low in selfefficacy, however, self-evaluation appears to have no effect on either their performance or learning outcomes.

## Acknowledgments

This study was funded by the Australian Research Council (grant number DP140101147).

## References

Ames, C., \& Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. Journal of Educational Psychology, 80(3), 260.
Bell, B. S., \& Kozlowski, W. (2002). Goal orientation and ability: Interactive effects on self-efficacy, performance, and knowledge. Journal of Applied Psychology, 87(3), 497.
Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., \& Rellinger, E. R. (1995). Metacognition and problem solving: A process-oriented approach. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(1), 205.
Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., \& Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. Science, 280(5364), 747-749.
Carver, C. S., \& Scheier, M. F. (2001). On the selfregulation of behavior. Cambridge, UK: Cambridge University Press.

DeGeest, D., \& Brown, K. G. (2011). The role of goal orientation in leadership development. Human Resource Development Quarterly, 22(2), 157-175.
DeShon, R. P., \& Gillespie, J. Z. (2005). A motivated action theory account of goal orientation. Journal of Applied Psychology, 90(6), 1096.
Desoete, A., \& Roeyers, H. (2006). Metacognitive macroevaluations in mathematical problem solving. Learning and Instruction, 16(1), 12-25.
Diseth, $\AA$. (2011). Self-efficacy, goal orientations and learning strategies as mediators between preceding and subsequent academic achievement. Learning and Individual Differences, 21(2), 191-195.
Double, K. S., \& Birney, D. P. (2017). Are you sure about that? Eliciting confidence ratings may influence performance on Raven's Progressive Matrices. Thinking and Reasoning, in press.
Dunning, D., Heath, C., \& Suls, J. M. (2004). Flawed selfassessment implications for health, education, and the workplace. Psychological Science in the Public Interest, 5(3), 69-106.
Duval, T. S., \& Wicklund, R. A. (1972). A theory of objective self-awareness. New York, NY: Academic.
Duval, T. S., \& Wicklund, R. A. (1973). Effects of objective self-awareness on attribution of causality. Journal of Experimental Social Psychology, 9(1), 17-31.
Eison, J. A. (1981). A new instrument for assessing students' orientations towards grades and learning. Psychological Reports, 48(3), 919-924.
Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitivedevelopmental inquiry. American Psychologist, 34(10), 906.
Inquisit. (2016). Inquisit 5 (Version 5.04). Retrieved from http://www.millisecond.com
Janssen, O., \& Van Yperen, N. W. (2004). Employees' goal orientations, the quality of leader-member exchange, and the outcomes of job performance and job satisfaction. Academy of Management Journal, 47(3), 368-384.
Metcalfe, J., \& Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. Psychonomic Bulletin \& Review, 15(1), 174-179.
Mitchum, A. L., Kelley, C. M., \& Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. Journal of Experimental Psychology: General, 145(2), 200.
Nelson, T. O., \& Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". Psychological Science, 2(4), 267-270.
Pintrich, P. R. (2000). The role of goal orientation in selfregulated learning. In M. Zeidner, P. R. Pintrich, \& M. Boekaerts (Eds.), Handbook of self-regulation. San Diego, CA: Academic Press.

Qualtrics. (2015). Qualtrics (Version July, 2015). Provo, Utah, USA Retrieved from http://www.qualtrics.com
R Core Team. (2015). R: A Language and Environment for Statistical Computing (Version 3.2.1) [Computer software]. Vienna, Austria R Foundation for Statistical Computing. Retrieved from ttp://http://www.R-project.org/
Silvia, P. J., \& Duval, T. S. (2001). Objective selfawareness theory: Recent progress and enduring problems. Personality and Social Psychology Review, 5(3), 230-241.
Silvia, P. J., \& Gendolla, G. H. (2001). On introspection and self-perception: Does self-focused attention enable accurate self-knowledge? Review of General Psychology, 5(3), 241.
Son, L. K., \& Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(1), 204.
VandeWalle, D. (1997). Development and validation of a work domain goal orientation instrument. Educational and Psychological Measurement, 57(6), 995-1015.
VandeWalle, D., Brown, S. P., Cron, W. L., \& Slocum Jr, J. W. (1999). The influence of goal orientation and self-regulation tactics on sales performance: A longitudinal field test. Journal of Applied Psychology, 84(2), 249.
VandeWalle, D., Cron, W. L., \& Slocum Jr, J. W. (2001). The role of goal orientation following performance feedback. Journal of Applied Psychology, 86(4), 629.

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. Journal of Educational Psychology, 96(2), 236.
Woodcock, R. W., McGrew, K. S., Mather, N., \& Schrank, F. (2001). Woodcock-Johnson III NU tests of achievement. Rolling Meadows, IL: Riverside Publishing.
Yeung, N., \& Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. Phil. Trans. R. Soc. B, 367(1594), 1310-1321.

# It's not just what we say, it's how we move: An examination of postural activity during a disclosure event 

Hannah M. Douglas (douglahh@mail.uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards<br>Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA<br>Stacie Furst-Holloway (furstse@ucmail.uc.edu)<br>Human Resource Effectiveness Lab, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA<br>Michael J. Richardson (richamo@ucmail.uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA<br>Rachel W. Kallen (rachel.kallen@uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA


#### Abstract

The current study incorporates concepts from dynamical systems theory (DST) and embodied cognition to propose a novel method of answering traditional questions in social psychology. Namely, we were interested in understanding postural sway complexity during the important interpersonal task of disclosing a hidden stigmatized identity (e.g., mental health disorder, history of sexual abuse). Using detrended fluctuation analysis and multifractal detrended fluctuation analysis, we captured postural activity while people shared their personal secrets to an imagined other. Results suggest that disclosure context, defined by both disclosure confidant and antecedent goals, is indeed embodied in our complex postural activity.


Keywords: Postural Sway; Concealable Stigmatized Identities; Detrended Fluctuation Analysis; Multifractal Detrended Fluctuation Analysis

## Introduction

The current project applied concepts from dynamical systems theory (DST) to common social-psychological phenomenon through the analysis of complex postural activity. Postural sway refers to subtle, unintentional movements that all people exhibit even when standing still. These nearly imperceptible fluctuations have demonstrated a functional role in maintaining balance and even efficiently exploring the environment (i.e., detecting depth) (Era \& Heikkinen, 1985). Generally, healthy adults tend to sway approximately 1 cm in the anterior-posterior (AP) direction and .5 cm in the mediolateral (ML) direction during quiet stance leading to great variability in postural activity within individuals (Baldan et. al., 2014). Research has found that there is meaningful structure to this movement variability in both
the AP and ML planes that exhibits fractal scaling, or self similarity across different timescales (Delignières, Torre, \& Bernard, 2011). This complex structure of postural sway allows us to adapt to different types of constraints-either personal, task relevant, or environmental-that exist across different time scales. The current project utilized two nonlinear data analytic techniques well suited to postural sway time series including detrended fluctuation analysis (DFA), and multifractal detrended fluctuation analysis (MFDFA) to characterize the spatio-temporal structure of postural activity during a social psychological event.

The complex (i.e., fractal) structure of postural variability can be influenced by a number of factors including schizophrenia (Kent et al., 2002), age, and movement disorders such as Parkinson's disease and Huntington's disease (lipsitz, 2004). A change in complexity is characterized by shifts from persistent pink noise to either anti-persistent white noise, or deterministic Brownian motion. As such, fractal, or pink noise, in postural sway has been consistently found in healthy adult populations, and a decline in complexity towards either white or Brownian noise is associated with a decline in health (Lipsitz, 2004).

While a change in the complex structure of postural variability is typically associated with poor health, recent research has found that cognitive activity can also impact postural behavior. For example, Riley, Baker, and Schmit (2003) found that postural sway standard deviation was reduced when participants were asked to complete a difficult digit rehearsal task. This change in postural sway as a function of a cognitive tasks, paired with the fractal nature of sway suggests a functional link between the brain and the body whereby the dynamics of human perceptual, motor, and cognitive processes are interaction-dominant (Riley, Shockley, van Orden,
2012). Interaction dominant-dynamics further suggest that each component system are coupled and therefore are reciprocally linked. This means that the behavior of each component, in this case the brain, the motor system, and the environment, depends on the activity of the other components (van Orden, Hollis, \& Wallot, 2012). To examine this phenomenon in a social psychological context, we will determine how postural activity changes while people disclose a concealable stigmatized identity (CSI) to an imagined other.

## Concealable Stigmatized Identity Disclosure

A CSI is any identity that is not immediately available to others, but could be socially devaluing if revealed, for example a mental health disorder, LGBT status, or a history of sexual abuse. While avoiding discrimination through concealment seems like an ideal solution, the extant literature has noted the numerous positive outcomes to disclosing (e.g., building trust, greater quality of life, etc.) as well as the negative impact of concealing (e.g., social isolation, anxiety, etc.) (Chaudoir \& Quinn, 2010).

Disclosure of a CSI, or the interpersonal process of sharing personal information, is a complicated process. The discloser must first decide how and when they want to share their identity with someone. Further, the discloser should be constantly evaluating their confidant's reaction to determine if they can expect a positive reaction with the desired social support, or a negative reaction and little or no support. Research suggests that, when disclosing a CSI to a confidant people will have specific goals for disclosing such as to build intimacy in a relationship, or to explain certain behaviors. Research suggests that the numerous goals for disclosure are either approach oriented-focused on achieving positive outcomes-or avoidance orientedfocused on avoiding negative outcomes (Chaudoir \& Fisher, 2010).

Approach and Avoidance Goal Motivation Research on goal motivation suggests that approach and avoidance systems result in differential exploration of the environment such that those who possess approach goals are interested in "reducing the discrepancy between themselves and their goal" (e.g., closing the gap between the discloser and the confidant; Chaudoir \& Fisher, 2010). Further, individuals who utilize approach goals in their disclosure may attend to positive stimuli in the environment. Conversely, when utilizing avoidance goals, individuals are interested increasing the distance between themselves and potential negative outcomes (e.g., increasing distance between the discloser and the confidant; Carver \& White, 1994). As research from embodied cognition suggests, changes in emotional or motivational systems would be reflected in behavioral
outcomes. Therefore, postural sway behavior provides a unique look into the embodiment of goal during the disclosure of a CSI. Further, as research has found a loss of complexity in postural sway as a function of increased cognitive load, it is likely the case that avoidance motivation, which is associated with attuning to negative environmental cues and less relaxed behaviors, would also lead to a loss of complexity. Further, there are many people in our lives with whom we can disclose such as with our friends and family (close others) and with our coworkers and bosses (professional others).

Disclosure Confidant Disclosure of a CSI can occur across all life domains and within different types of relationships. Our relationships with others can vary greatly as a function of domain context (e.g., workplace, family life, and social setting, etc.). Often, our relationships with family members will be different from our relationships with a boss or a coworker due to social norms associated with these contexts. Therefore, the level of detailed disclosure of a CSI is likely less for those we have a professional relationships with compared to our close friends or family members. In fact, many people may feel motivated to keep a CSI hidden completely from their coworkers as revealing such information could have a detrimental impact on their career path and job outcomes (Jones \& King, 2014).

Despite the potential for negative outcomes due to CSI disclosure, disclosure in the workplace should not be discounted. Research suggests potential negative workplace consequences of concealing including less job satisfaction and attention (Day \& Scheonrade, 1997). With a large portion of the workforce continuously making decisions about the information they reveal and conceal in a workplace setting, it is becoming increasingly apparent that a better understanding of workplace disclosure is necessary. However, the increased tension and threat involved with disclosing across different life domains might also impact the behavioral expression via postural sway behavior.

Postural Activity While the current literature has noted the importance of positive interpersonal disclosure outcomes across multiple life domains (i.e., home life, work life) utilizing different goals, little is known about how these different contexts impact the embedded nature of our cognitive and behavioral systems within the world. The present study is the first of its kind to examine the disclosure experience through the lens of embodied cognition in order to understand how the disclosure context is differentially manifested in measurable behavioral outcomes (i.e., postural activity).

Despite attempts to understand the impact of nonverbal behaviors on personal self-disclosure (see Derlega \& Berg, 2013), the existent literature has
focused on general self-disclosure, not disclosure of a CSI specifically. Further, nonverbal behaviors have typically been characterized by discrete, observable behaviors (e.g., facial expression, nods, and openness). The current project examines time dependent postural sway by utilizing dynamic data analytic tools that can capture the disclosure process as it occurs. By examining postural sway behavior during the disclosure of a CSI we can gain a better understanding of how our mental processes are manifested in our bodies relationship with the environment. Further, support for this claim would suggest that shifting motivation systems might lead to more positive behaviors, both verbal and nonverbal, during a disclosure event, and therefore more positive disclosure outcomes.

## Current Project

This project hopes to be the first to bridge the gap between the three discussed areas of research: disclosure context (i.e., close other and professional other disclosure), antecedent goals for disclosure events, and embodied cognition during the disclosure of a CSI. With this project, we hope to integrate theory from stigma, embodied cognition, and interaction-dominant dynamics to capture a holistic understanding of the cognitive and movement processes at play during CSI disclosure. As such, this project will utilize theory unique to postural sway literature, and measurement and data analytic techniques novel to disclosure. Finally, our results and discussion will be presented in such a way that both social and ecological psychologists might be able to utilize theory and methods from each other in future research endeavors. Based on previous postural sway research, we expect disclosures utilizing approach goals to close others would exhibit pink noise compared to avoidance disclosures to professional targets.

## Method

## Design

This study employed a 2 (goal motivation: approach/avoidance) $\times 2$ (target: close other/professional other) mixed design with goal motivation is the between subjects variable and disclosure target the within subjects variable. The primary dependent variables are postural sway dynamics measured at the head and waist (via mono-fractal and multi-fractal scaling) and responses on the Behavioral Approach System/Behavioral Avoidance System (BIS/BAS) scale.

## Participants

43 undergraduates were recruited from a large Midwestern University to participate in this study. Prior
to recruitment, participants were prescreened to determine their eligibility. In order to participate in this study, participants had to self-identify as living with a CSI. One participant was excluded from data analysis due to technical errors resulting in a sample of 42 participants. The majority of participants were female (36) and identified as white (35). The mean age was $20.21(S D=3.09)$. See table 1 for a breakdown of each CSI represented in this study.

Table 1. Table 1 shows the number of participants with each CSI type

| CSI Type | $N$ |
| :--- | :---: |
| Mental Health Disorder | 16 |
| Sexual Assault | 7 |
| Gender/Sexual Minority | 10 |
| Eating Disorder | 4 |
| Multiple CSI's | 2 |
| Other | 3 |

## Procedure

In the first portion of the study participants were seated at a computer equipped with Media Lab software (Empirisoft, 2014) where they completed the majority of the experiment. They were first asked to think about and describe a secret that they often keep hidden. Each participant was then instructed to write two disclosure letters sharing this secret to a close friend/family member and the other to someone with whom they have a professional relationship. Specifically, they were asked to think about a person in their life that they have not told this secret, but would like to. Prior to writing each letter, participants were told to write $3-5$ goals they have for their disclosure. To manipulate approach and avoidance goals, participants were simply told to either "think about achieving positive outcomes with their letter" or "think about avoiding negative outcomes with their letter" respectively.
After writing both disclosure letters, participants acted out their disclosure as if the person they wrote the letter to was standing in the room. During the disclosure event, two Polhemus sensors (one attached to a headband on the back of the head, the other attached to a belt just bellow the belly button) recorded postural activity at 60 Hz (FASTRAK, Polhemus, VT, USA). The experimenter explained that they should act as though they were talking to the person that they chose, using their letter as a guide. After completing the disclosure for both written letters, participants completed a number of self-report measures including the BIS/BAS scale (Carver \& White, 1994).

## Data Analysis

To capture the time dependent structure of postural variability as a function of both goal priming and disclosure confidant during the disclosure of a CSI, both DFA and MFDFA were used. Because postural data exhibits non-stationary, time-dependent variation, these data are characterized by fractional Brownian motion (fBm) making it particularly well suited to DFA (et al., 2000; Delignières, Torre, \& Bernard, 2011).

DFA provides the scaling exponent, $\alpha$, which describes the fractal scaling of a time series whereby: $\alpha$ $\approx .5$ indicates random, white noise scaling; $\alpha \approx 1$ suggests persistent pink noise scaling; and $\alpha \approx 1.5$ indicates Brownian motion.

MFDFA is an extension of the DFA and examines differences in the scaling exponents between small and large fluctuations. The relevant outcome parameter of interest in MFDFA is a characterization of the width of the multifractal spectrum $\mathrm{h}_{\text {MAX-MIN. }}$. Because MFDFA tells us whether there are different scaling exponents that exist at fast and slow fluctuations, if $\mathrm{h}_{\text {MAX-MIN }}$ is greater than 0 , we can assume the time series exhibits multifractality. See Ihlen (2012) for a detailed description of both DFA and MFDFA procedures.

Finally, Prior to analyzing each postural sway time series in the AP plane at the head ( $\mathrm{AP}_{\mathrm{HEAD}}$ ) and waist ( $\mathrm{AP}_{\text {WAIST }}$ ) and the ML plane at the head ( $\mathrm{ML}_{\text {HEAD }}$ ) and waist ( $\mathrm{ML}_{\text {WAIST }}$ ), the data were downsampled from 60 Hz to 30 Hz , linearly detrended, and, low-pass filtered at 20 Hz using a $2^{\text {nd }}$ order Butterworth filter. A surrogate analysis (detailed below) was also performed for DFA and MDFA for validation purposes.

## Results

A series of separate mixed method ANOVA's were performed on all relevant outcome parameters for DFA, and MFDFA to test our hypotheses that approach and avoidance goal motivation and target confidant would impact the dynamical structure of postural activity during a disclosure event. Four separate 2 (goal: approach/avoidance) $\times 2$ (target: close other/professional other) ANOVA's were performed on all outcome parameters, one each for $\mathrm{AP}_{\text {HEAD }}, \mathrm{ML}_{\text {HEAD }}, \mathrm{AP}_{\text {WAIST }}$, and $\mathrm{ML}_{\text {WAIST }}$ sway. Prior to statistical analysis, outliers 3 SD above and below the mean were identified and replaced with the mean value.

## Detrended Fluctuation Analysis

A series of ANOVAs were performed to capture differences in sway as a function of $\alpha$. To verify that there was a difference between the original time series and the randomly reshuffled, surrogate time series, a third 2 level term in the ANOVA (data: original/randomly reshuffled) was included making the analysis a $2 \times 2 \times 2$ design. There was a main effect of
data type for all $\mathrm{AP}_{\text {HEAD }}, \mathrm{AP}_{\text {WAIST }}, \mathrm{ML}_{\text {HEAD }}$, and $\mathrm{ML}_{\text {WAIST }}$, (for all $F(1,40)>2097, p<.0001$ ) such that the original data results were significantly larger from the randomly reshuffled, surrogate time series. That is, the original time series produced an average $\alpha$ around 1.3 for all directions of sway and the reshuffled time series produced $\alpha$ of .5 for all directions of sway. There were 2-way interactions of goal and data type for $\mathrm{ML}_{\text {HEAD }}$, and $\mathrm{ML}_{\text {WAIST }}(F(1,40)=5.52, p=.024$, and $F(1,40)=4.74, p=.035$, respectively), however, these results simply reflect a main effect of goal priming for the original data; no differences emerged in the reshuffled time series as a function of goal priming. As such, below is the planned $2 \times 2$ ANOVAs on the analysis of real (non-shuffled) data.
The ANOVA comparing the $\alpha$ exponent for $\mathrm{AP}_{\text {HEAD }}$ revealed a significant target by group interaction, $F(1,40)=4.32, p=.04, \eta_{\mathrm{p}}^{2}=.098$. Bonferroni post hoc comparisons were used to examine differences between $\alpha$ for close other and professional other disclosures for each approach and avoidance primed disclosures separately. Results indicate a marginally significant difference in the $\alpha$ exponent between close other and professional other target disclosures in the avoidance primed condition, whereby close other disclosures exhibited less persistent fractal scaling in their postural sway $(M=1.36, S D=.11)$ compared to professional other disclosures; which were more persistent and closer to pink noise $(M=1.3, S D=.1)$. There was no difference between close other and professional other disclosures during approach primed disclosures $(T(21)=$ $.49, p>.05)$. There were no other main effects for $\mathrm{AP}_{\text {HEAD }}$ sway (all $\left.F(1,40)<1.8, p>.05\right)$.

Next, an ANOVA comparing the alpha exponent from $\mathrm{ML}_{\text {HEAD }}$ revealed a significant main effect of goal priming $\left(F(1,40)=5.81, p=.02, \eta_{\mathrm{p}}^{2}=.13\right)$ such that approach primed disclosures exhibited more persistent fractal scaling in their postural sway $(M=1.28, S D=$ .13) compared to avoidance primed disclosures; which were less persistent and closer to Brown noise ( $M=1.35$, $S D=.13$ ). There were no other main or interaction effects for $\mathrm{ML}_{\text {HEAD }}$ (all $F(1,40)<2.1, p>.05$ ). Similar to the results found in $M L_{\text {HEAD }}$ sway, a significant main effect of goal priming emerged in $\mathrm{ML}_{\text {WAIST }}$ sway $\left(F(1,40)=4.56, p=.04, \eta_{\mathrm{p}}^{2}=.1\right)$, whereby those in the avoidance condition exhibited a loss of complexity compared to those in the approach primed condition ( $M$ $=1.34, S D=.14$ and $M=1.28, S D=.14$ respectively). No other main effects or interactions were significant for $\mathrm{ML}_{\text {WAIST }}$ (all $F(1,40)<1.9, p>.05$ ). Finally, there were no significant effects of $\alpha$ on $\mathrm{AP}_{\text {WAIST }}$ sway (all $F(1,40)$ $<2.47, p>.05$ ) (see Figure 1).


Figure 1: This figure represents mean $\alpha$.

$$
* p<.05, * * p<.01
$$

## Multifractal Detrended Fluctuation Analysis

The final series of ANOVA's compared the $h_{\text {MAX-MIN }}$ value for each independent variable. To check that there is a difference between the original time series and the surrogate time series, a third 2 level term (data: original/surrogate) was included in the initial analysis. The surrogate time series was developed by shuffling the time series using an inverse amplitude-adjusted Fourier transform to maintain the same scaling relation $\alpha$ (see Ihlen \& Vereijken, 2010 for detailed description) There was a significant main effect of data type for all $\mathrm{ML}_{\text {HEAD }}$, $\mathrm{ML}_{\text {WAIST, }}, \mathrm{AP}_{\text {HEAD }}$, and $\mathrm{AP}_{\text {WAIST }}$ (all $F(1,40)>54.7, p<$ $.0001)$ whereby the $\mathrm{h}_{\text {MAX-MIN }}$ was greater in the original data compared to the phase reshuffled time series. There were no 2-way interactions including data type suggesting there was no impact of goal priming or target confidant on results of the surrogate analysis (all $F(1,40)$ $<3.9, p>.05$ ). Therefore, results of the planned 2-way ANOVA examining goal priming and target confidant are reported below.

The analysis of $\mathrm{h}_{\text {MAX-MIN }}$ for $\mathrm{AP}_{\text {HEAD }}$ revealed a main effect of goal priming $\left(F(1,40)=4.95, p=.03, \eta_{\mathrm{p}}{ }^{2}=.11\right)$ such that the width was larger for approach primed disclosures ( $M=.97, S D=.04$ ) compared to avoidance primed disclosures ( $M=.85, S D=.04$ ). There were no other significant results for $\mathrm{AP}_{\text {HEAD }}$ (all $F(1,40)<3.01, p$ $>.05$ ). The same pattern of significant results emerged for $\mathrm{ML}_{\text {HEAD }}$ and $\mathrm{ML}_{\text {WAIST }}$, such that a main effect of goal motivation was revealed for both $(F(1,40)=8.57, p=$ $.006, \eta_{\mathrm{p}}^{2}=.18$ and $F(1,40)=7.62, p=.009, \eta_{\mathrm{p}}^{2}=.16$ respectively). The width for $\mathrm{ML}_{\text {HEAD }}$ was larger for
approach-primed disclosures $(M=1.04, S D=.22)$ than avoidance primed disclosures $(M=.89, S D=.22)$. Similarly, the ML ${ }_{\text {WAIST }}$ width was larger for approachprimed disclosures $(M=1.02, S D=.18)$ than avoidance primed disclosures $(M=.9, S D=.18)$ (Figure 2).


Figure 2: This figure demonstrates mean $\mathrm{h}_{\text {MAX-MIN }}$.

$$
p<.05, * * p<.01
$$

## Discussion and Conclusion

Taken together, these results support our hypotheses that both goal motivation and disclosure confidant would impact unintentional postural activity. We sought to examine the disclosure event on a very small scale (i.e., postural behavior) in order to understand how context shapes the way people communicate through behavior. These results broadly support the idea that our cognition and emotional content are manifested and embodied in measureable behavioral outcomes (Marsh, Ambady, \& Kleck, 2005). Most notable in these results is the influence of antecedent goal priming on the structure of postural variability. By utilizing nonlinear data analytic techniques novel to disclosure research, we have provided support that our cognitive and motor systems are functionally linked as a complex dynamical system.

Specifically, the significant interaction of the scaling exponent $\alpha$, which revealed that close other disclosures exhibited more deterministic behavior than professional other disclosures at $\mathrm{AP}_{\text {HEAD }}$, is contrary to our hypothesis that professional other disclosures would be more deterministic. However, since this effect was only found in the avoidance condition, which is associated with negative outcomes, these results may indicate that participants expected greater threat to their intimate relationships during close other disclosures when they
were utilizing avoidance goals. Because this effect was only found in the AP direction, it is important that future work seek to replicate these results.

Finally, results of the MFDFA support the mounting evidence that postural sway behavior exhibits multiple scaling exponents, as well as our hypothesis that disclosure context would functionally impact movement variability. Importantly, a significant difference in the $\mathrm{h}_{\text {MAX-MIN }}$ parameter suggests that approach primed disclosures exhibit a wider range of scaling exponents than avoidance primed disclosures. This supports the theory that approach systems are associated with attuning to more positive stimuli in the environment. As theory of postural sway variability suggests, our postural system aids in efficiently exploring the environment. Because we see differences in the smallest and largest scaling exponents, this suggests participants are able to explore different stable states in the approach condition compared to the avoidance condition. This could make approach primed disclosures more adept at adjusting behaviors with new information.

The results of this project provide evidence that both disclosure confidant and antecedent goals can affect the disclosure event itself. Further, this research suggests that postural sway behavior is an emergent property of a complex system and serves a functional role in both attaining environmental information and embodying ones cognitive and emotional processes. This has implications for developing tools for people who want to disclose a CSI. For example, by simply shifting internal motivation from avoidant to approach a reciprocal distribution across behaviors at different time scales could cascade, from very fast processes including postural sway, to slower timescale behaviors such as gross body movement, language, and confidant reactions. Future research should examine this relationship as well as how number of disclosures or fear of disclosure impacts these effects.

## References

Baldan, A. M. S., Alouche, S. R., Araujo, I. M. G., \& Freitas, S. M. S. F. (2014). Effect of light touch on postural sway in individuals with balance problems: a systematic review. Gait \& posture, 40(1), 1-10.
Carver, C. S., \& White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. Journal of personality and social psychology, 67(2), 319.
Chaudoir, S. R., \& Fisher, J. D. (2010). The disclosure processes model: understanding disclosure decision making and postdisclosure outcomes among people living with a concealable stigmatized identity. Psychological bulletin, 136(2), 236.
Chaudoir, S. R., \& Quinn, D. M. (2010). Revealing concealable stigmatized identities: The impact of
disclosure motivations and positive first-disclosure experiences on fear of disclosure and well-being. Journal of Social Issues, 66(3), 570-584.
Delignières, D., Torre, K., \& Bernard, P. L. (2011). Transition from persistent to anti-persistent correlations in postural sway indicates velocity-based control. PLoS Comput Biol, 7(2), e1001089.
Derlega, V. J., \& Berg, J. H. (Eds.). (2013). Selfdisclosure: Theory, research, and therapy. Springer Science \& Business Media.
Eke, A., Herman, P., Bassingthwaighte, J., Raymond, G., Percival, D., Cannon, M., ... \& Ikrényi, C. (2000). Physiological time series: distinguishing fractal noises from motions. Pflügers Archiv, 439(4), 403-415.
Era, P., \& Heikkinen, E. (1985). Postural sway during standing and unexpected disturbance of balance in random samples of men of different ages. Journal of Gerontology, 40(3), 287-295.
Ihlen, E. A. (2012). Introduction to multifractal detrended fluctuation analysis in Matlab. Fractal Analyses: Statistical And Methodological Innovations And Best Practices, 97.
Jones, K. P., \& King, E. B. (2014). Managing Concealable Stigmas at Work A Review and Multilevel Model. Journal of Management, 0149206313515518.

Kent, J. S., Hong, S. L., Bolbecker, A. R., Klaunig, M. J., Forsyth, J. K., O'donnell, B. F., \& Hetrick, W. P. (2012). Motor deficits in schizophrenia quantified by nonlinear analysis of postural sway. PLoS one, 7(8), e41808.
Lipsitz, L. A. (2004). Physiological complexity, aging, and the path to frailty. Science's SAGE KE, 2004(16), pe16.
Marsh, A. A., Ambady, N., \& Kleck, R. E. (2005). The effects of fear and anger facial expressions on approach-and avoidance-related behaviors. Emotion, 5(1), 119.
Riccio, G. E., \& Stoffregen, T. A. (1991). An ecological theory of motion sickness and postural instability. Ecological psychology, 3(3), 195-240.
Riley, M. A., Baker, A. A., \& Schmit, J. M. (2003). Inverse relation between postural variability and difficulty of a concurrent short-term memory task. Brain Research Bulletin, 62(3), 191-195.
Riley, M. A., Shockley, K., \& Van Orden, G. (2012). Learning from the body about the mind. Topics in Cognitive Science, 4(1), 21-34.
Smart, L., \& Wegner, D. M. (2000). The hidden costs of hidden stigma. The social psychology of stigma, 220242.

Van Orden, G., Hollis, G., \& Wallot, S. (2012). The blue-collar brain. Scale-free Dynamics and Critical Phenomena in Cortical Activity, 822, 114.

# A theory of the detection and learning of structured representations of similarity and relative magnitude 

Leonidas A.A. Doumas ${ }^{1}$ (alex.doumas@ed.ac.uk); Aaron Hamer ${ }^{1}$ (aaron.hamer@ed.ac.uk); Guillermo Puebla ${ }^{1}$ (pueblaramirezg@gmail.com); Andrea E. Martin ${ }^{1,2}$ (andrea.martin@mpi.nl)<br>${ }^{1}$ Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, United Kingdom<br>${ }^{2}$ Max Planck Institute for Psycholinguistics; Nijmegen, The Netherlands


#### Abstract

Responding to similarity, difference, and relative magnitude is ubiquitous in the animal kingdom. However, humans seem unique in the ability to represent relative magnitude and similarity as abstract relations that take arguments (e.g., greater-than $(x, y)$ ). While many models use structured relational representations of magnitude and similarity, little progress has been made on how these representations arise. Models that use these representations assume access to computations of similarity and magnitude a priori. We detail a mechanism for producing invariant responses to "same", "different", "more", and "less" which can be exploited to compute similarity and magnitude as an evaluation operator Using DORA (Doumas, Hummel, \& Sandhofer, 2008), these invariant responses can serve to learn structured relational representations of relative magnitude and similarity from pixel images of simple shapes.


## Introduction

Reacting to similarity, and magnitude ("same"/ "different", "more"/"less"; SDML) are hallmarks of complex organisms. For example, gerbils use the retinal size of a stimulus to estimate its distance (Goodale, Ellard, \& Booth, 1990), rats choose the larger of two food rewards (Kim et al., 2015), and pigeons learn to group pictures of 16 identical items in one set, and pictures of 16 different items in a different set (Young, Wasserman, \& Garner, 1997).
Humans, however, go beyond simple detection of relative magnitude and similarity. We make an analogies between a nucleus and the sun because they are both larger than their orbiting bodies (electrons and planets). We infer this relationship because we represent relative magnitude and similarity as abstract relations that take arguments (i.e., as predicates; see Holyoak, 2012).
Our ability to reason about abstract SDML manifests in a variety of domains such as analogy (e.g., Holyoak \& Thagard, 1995), categorisation (e.g., Medin, Goldstone, \& Gentner, 1993), and concept learning (e.g., Doumas \& Hummel, 2013). While models that use structured representations have had success in accounting for how humans use abstract SDML, these models say little about where the representations they use come from in the first place. For example, SME (Falkenhainer, Forbus, \& Gentner, 1989), STAR (Halford et al., 1998), and LISA (Hummel \& Holyoak, 1997, 2003) account for many phenomena from the analogy literature, but require the relations they use to make these analogies be hand-coded by the modeler. Similarly, Bayesian models of concept development and learning (e.g., Kemp, 2012; Kemp \& Tenenbaum, 2007, 2009; Lake et al., 2016) assume relational structures a priori, starting with a vocabulary
of objects and relations and learning new concepts by building new combinations of these innate elements.
Some models attempt to account for the origins of abstract concepts without assuming innate representations of relational concepts. For example, BART (Lu, Chen, \& Holyoak, 2012) uses feature lists generated by human subjects or corpora analysis to find properties associated with items in the world which instantiate particular relations. BART has difficulty with some edge cases of relational cognition (e.g., reasoning about something like an atom being bigger than something else when it has not experienced instances where an atom was bigger than anything), but the model makes a serious effort to account for development of analogy-making with minimal assumptions about the starting representations of the learning system.
In a similar vein, DORA (Doumas, Hummel, \& Sandhofer, 2008) explains how structured representations (i.e., predicates) can be acquired from unstructured representations (i.e., feature vectors). While DORA learns relational representations that can take any arguments (including edge cases and completely novel arguments; Doumas et al., 2008), DORA assumes a system to detect the invariant features that underlie the abstract concepts that it learns.
A complete account of how people acquire structured representations of abstract SDML relations must solve three problems. First, there must be some invariant features which remain constant across instances of the relation which the perceptual/cognitive system can learn to detect. Second, the system must isolate these invariants from other properties of the objects engaged in the relation to be learned. Third, the system must learn a predicate representation of the relational properties (i.e., an explicit entity that can be bound to arbitrary, novel arguments).
We solve the first problem with an extension to DORA which produces invariant responses to similarity and relative magnitude. We have previously shown how DORA can solve the second and third of these problems (Doumas et al., 2008). We begin with a brief overview of DORA, describe the process which produces invariant features for SDML, and provide simulations demonstrating how DORA solves all three problems to learn structured relational representations of SDML.

## Model

## DORA

DORA (Doumas, et al., 2008) is a symbolicconnectionist model, based on the LISA (Hummel \&

Holyoak, 1997, 2003) model of analogy. DORA learns structured relational representations from unstructured representations of objects (e.g. feature vectors).

LISAese Representations We begin by describing the end state of DORA's representations (i.e., its representations after it has gone through learning). Relational propositions are represented by a hierarchy of distributed and localist codes (see Figure 1). At the bottom, semantic units code the features of objects and roles in a distributed fashion. In the next layer, localist predicate-object (PO) units representing individual predicates (or roles) and objects, are connected to these distributed semantic representations. In the next layer, localist role-binding (RB) units link predicates and objects into specific role-filler pairs. At the top of the hierarchy, localist proposition ( P ) units link RB units into complete relational propositions. Importantly, while we use different names for the units in different layers, and different shapes to distinguish these units in diagrams, we do so only for the purposes of expositional brevity. These are just nodes in different layers of a network. RB units are just like PO units, except for the fact that they are in a different layer, and, therefore, take input from and pass input to different layers of units.


Figure 1. Complete relational proposition in DORA. Units in different layers are coded using different shapes for the purposes of exposition.

Propositions in DORA are divided into four mutuallyexclusive sets of layered networks: a driver, one or more recipients, long-term memory (LTM), and the emerging recipient (EM). Each set consists of a layered network of $\mathrm{PO}, \mathrm{RBs}$, and P units (i.e., there are specific layers coding for PO, RB, and $P$ units in the driver, and another set of layers coding for $\mathrm{PO}, \mathrm{RB}$, and P units in the recipient). Semantic units are shared across all networks (i.e., driver and recipient units are connected to the same pool of semantic units). The driver corresponds to the current focus of attention and controls the flow of activation. Units in the driver pass activation to the semantic units. Because the semantic units are shared by all sets, activation flows from the driver to the other three sets. DORA operations (e.g., mapping and relation learning, detailed below) proceed as a product of units in the driver activating their semantic units, which in turn activates units in the various other sets.
When a relational representation enters the driver the binding of roles to their fillers must be represented
dynamically without violating their independence (i.e., it is not sufficient to represent bindings using only conjunctive units; see, e.g., Doumas \& Hummel, 2005; von der Malsburg, 1999). DORA uses systematic asynchrony of firing to dynamically bind roles to their fillers (see Doumas et al., 2008). As a relational representation in the driver becomes active, bound objects and roles fire in direct sequence. Information about role-filler bindings is carried by proximity of firing (e.g., with roles firing directly before their fillers). This sequence-based binding keeps roles and their fillers distinct and thus independent. Using the example in Figure 1, in order to bind bigger to block and smaller to ball (and so represent larger (block, ball)), the units corresponding to bigger fire directly followed by the units corresponding to block, followed by the units for coding smaller followed by the units for ball.

Mapping DORA uses LISA's mapping algorithm (see Hummel \& Holyoak, 1997; Doumas et al., 2008). DORA learns mapping connections between units of the same type in the driver and recipient (e.g., between PO units in the driver and PO units in the recipient). These connections grow whenever corresponding units in the driver and recipient are active simultaneously. The connections act as mappings between corresponding structures in separate analogs. They also permit correspondences learned in mapping to influence correspondences learned later.

Relation Learning DORA uses comparison to isolate shared properties of objects and to represent them as explicit structures. DORA begins with simple featurevector representations of objects (i.e., a node connected to a set of semantic features describing that object). When DORA compares two objects, the two representations are activated simultaneously. For instance, if DORA compares a block that is larger than some object to a plate that is larger than some other object (e.g., when the block is larger than a ball and the plate is larger than a fork), then the nodes representing the block and plate fire together (Figure 2a). Semantic features shared by the compared objects (i.e., features common to the block and the plate) receive twice as much input and thus become roughly twice as active as features connected to one but not the other (Figure 2b). DORA then learns connections between a newly recruited PO unit and active semantic units via Hebbian learning (Figure 2c). In Hebbian learning the strength of a learned connection is a function of unit activation (i.e., stronger connections are learned to more active units). Consequently, the new PO unit becomes most strongly connected to the highly active semantic units. The new PO becomes an explicit representation of the feature overlap between the block and plate. In this example, DORA forms an explicit representation of the semantics of bigger things (i.e., the features common to both the block and plate). The new PO functions as a predicate representation of bigger because it can be dynamically bound to fillers via an RB unit (Figure 2d).


Figure 2. Comparison-based predication in DORA. DORA learns a representation of bigger by comparing a block that is bigger than some object to a plate that is bigger than some other object. (a) DORA compares a block and a plate. Units representing both become active. (b) Feature units shared by the block and the plate become more active than unshared features (darker grey). (c) A new PO unit learns connections to features in proportion to their activation (solid lines indicate stronger connection weights). The new unit codes the featural overlap of the block and plate (i.e., the role "bigger"). (d) This new PO unit functions as a predicate when dynamically bound to fillers.

DORA learns representations of multi-place relations by linking sets of co-occurring role-filler pairs into hierarchical relational structures. Continuing the example, when DORA compares a plate that is larger than a fork to a block that is larger than a ball, it will map larger (plate) to larger (block) and smaller (fork) to smaller (ball) (Figure 3a). When constituent sets of rolefiller pairs are mapped, a distinct pattern of firing emerges-namely, mapped RB units fire together and out of synchrony with any other RB units; Figure 3b-d). This pattern is a reliable signal that DORA exploits to combine sets of role-filler pairs into multi-place relations. In response to the pattern, DORA recruits a P unit that learns connections to any active RB units in the recipient (Figure $3 \mathrm{e}-\mathrm{g}$ ) via Hebbian learning. The result is a P unit linking the RB units in the recipient into a complete relational structure (larger (block, ball); Figure 3i).

## Producing invariant responses for basic SDML

A comparison-based solution to the problem of learning an invariant feature coding for "more", "less", and "same" requires the assumption that initially available magnitude information is coded by a direct neural proxy: All else being equal, higher magnitude items are coded (at least early in processing) by more neurons than comparatively lower magnitude items. For example, a larger item will be coded by more neurons
than a smaller item. There is a preponderance of evidence for this assumption. In visual processing, larger items take up more space on the retina (e.g., Wandell, 1995) and are coded by larger swaths of the visual cortex (e.g., Engel et al., 1994).


Figure 3. DORA learns a representation of the whole relation larger (block, ball) by mapping bigger(plate) to bigger(block) and smaller(fork) to smaller(ball). (a) The units coding bigger fire; (b) the units for plate and block fire; (c) the units for smaller fire; (d) the units for
fork and ball fire. (e) DORA recruits a P unit in the recipient. (f-g) DORA learns a connection between the new $P$ unit and the active RB unit (the unit coding for $\operatorname{bigger}(\mathrm{block})$ ). (h-i) The P unit learns connections to the active RB unit (coding for smaller(ball)). The result is a structure coding for larger(block, ball).

Basic magnitude calculation is accomplished by comparison. When the model attends to two representations with specific magnitude values (e.g., two POs attached to absolute size are present in the driver together; Figure 4a), the representations of the absolute magnitude semantics are co-activated and the PO units attached to these semantic units compete via lateral inhibition (Figure 4b). The POs will eventually settle, with either one PO becoming more active and inhibiting the other to inactivity, or, when both POs code for the same absolute magnitude, with both POs in a steady state of co-activation. More semantic units can then respond to the particular pattern of firing in the driver POs. Some units are excited by two active POs in the driver, others
are excited by a single highly active PO early in firing, or by a single highly active PO late in firing (these regions of excitement are easily learnable via simple neural threshold tuning). The active POs learn connections to the active semantic unit by Hebbian learning. If a single PO is active, that unit will learn connections to the semantics that are activated by a single highly active driver PO early in firing (which becomes the invariant signal for "more"; Figure 4c). When the active PO becomes inhibited (because of asynchronous binding), the second PO (the one inhibited by the winning PO) will become active (Figure 4d). That unit learns connections to the semantics that are activated by a single highly active driver PO late in firing (which becomes the invariant signal for "less"; Figure 4d). Otherwise, if two POs are co-active (i.e., they code the same magnitude), then they will learn connections to the semantics which are activated by two active driver POs (which becomes the invariant signal of "sameness".
(a)

(c)

(b)

(d)


Figure 4. The SDML detector working on POs coding different values on a dimension. For the purposes of clarity, only the predicate POs and their semantics are depicted in this figure. (a) Two POs coding for different heights are in the driver. (b) The semantics coding for absolute dimensional information become active and the two POs compete to become active. (c) The unit coding for the greater value on the dimension (here height-6) becomes active first, thus marking it as "more". The PO learns a connection to the semantic that responds to winning the SDML competition (i.e., the invariant of "more"). (d) The unit coding for the lesser value on the dimension (here height-3) will become active last, thus marking it as "less". The predicate is connected to the semantic unit coding for losing the SDML competition, or the invariant of "less".

In short, comparing different magnitudes in a network in which magnitude information is coded by an absolute
proxy (as in the human neural system) produces one of three patterns. (1) Both units settle into a state of similar co-activation-which occurs when two representations of the same magnitude are compared. (2) One unit becomes more active and forces the second unit to inactivity-which occurs when a unit codes for a greater magnitude. (3) One unit becomes active after it has been inhibited by a winning unit-which occurs when a unit codes for a lesser magnitude. Whatever units respond to these patterns naturally or through tuning become implicit invariant codes for the presence of "sameness", "moreness", and "lessness", respectively. Vitally, the same patterns will emerge and the same codes will become active when specific relative magnitudes are present even cross dimensionally. That is, the same patterns emerge and units become active during an instance of different absolute height, or width, or colour. What is left for the system is to learn explicit representations of these invariant semantics that are not tied to any specific magnitudes (e.g, a PO connected to semantics encoding 'more' \& 'height', without strong connections to any specific height) and can take other POs as arguments. In other words, exactly the learning that DORA does.

## Simulations

## Simulation 1

We tested whether DORA could learn structured representations of relative SDML relations starting with information about sets of shapes with features representing absolute values on dimensions. This simulation mirrored what happens during development when a child learns from experience without a teacher or guide.

The model began with pixel images of basic shapes (differing in shape, colour, size, width, and height). These images were pre-processed with a feedforward neural network that learned via back-propagation to deliver absolute shape, colour, size, width, and height information (akin to that information delivered by early visual processing). Each processed image was represented by a PO attached to the delivered features. In addition, each shape was also attached to a set of 10 extraneous features selected randomly from a set of 100 features, included as noise (as objects in the world contain several features extraneous to any particular learning goal). Each shape was then randomly paired with another to create pairs of shapes over which relations were learned. We created 100 pairs of objects in this manner and placed them in DORA's LTM.

We then allowed DORA to attempt to learn from these basic representations. On each learning trial, DORA selected one pair of objects from LTM at random and ran (or attempted to run) retrieval, mapping, SDML comparison, predication, and multi-place relation learning, and stored any representations that it learned in LTM. In short, we are testing whether unguided learning from simple shape objects is sufficient for DORA to
learn structured representations of relative SDML relations.

We defined a relational quality metric as the mean of connection weights to relevant features (i.e., those defining a relative magnitude on some specific dimension (e.g., 'more'+'height', or 'less'+'width')) divided by the mean of all other connection weights +1 (1 was added to the mean of all other connection weights to normalize the quality measure to between 0 and 1 ). A higher quality denoted stronger connections to the semantics defining a specific SDML relation relative to all other connections. We measured the relational quality of the last 100 items DORA had learned after each 100 learning trials for 1000 total learning trials. Importantly, we tested all representations that the model learned (not just those that instantiated the relevant relations) and included these in the relational selectivity calculation.
Figure 5 shows the quality of the representations that DORA learned. DORA learned representations of whole relational structures encoding relative magnitudes and similarity on all the encoded dimensions. DORA learned representations of bigger (one predicate PO connected most strongly to the semantics 'more' \& 'size', the other connected to 'less' \& 'size'), wider (predicate POs connected to 'more' \& 'width, and 'less' \& 'width'), taller (predicate POs connected to 'more' \& 'height, and 'less' \& 'height), same-size (predicate POs both connected most strongly to 'same' \& 'size;), same-width (predicate POs both connected most strongly to 'same' \& 'width'), same-height (predicate POs both connected most strongly to 'same' \& 'height'), same-colour (predicate POs both connected most strongly to 'same' \& 'colour'), and same-shape (predicate POs both connected most strongly to 'same' \& 'shape'). The results indicate that DORA can learn structured representations of relative SDML relations from objects that include only absolute values on dimensions even with the addition of extraneous noise.


Figure 5. Results of DORA's learning.

## Simulation 2

A crucial question remains: do the representations DORA learns meet the requirements of relational representations? Some hallmark of relational representations (see Holyoak, 2012) are that they, (i) form the basis of solving cross mappings; (ii) support
mapping similar, but non-identical predicates; and (iii) form the basis of overcoming the n -ary restriction.

During cross-mapping, an object (object1) is mapped to a featurally less similar object rather than a featurally more similar object because it (object1) plays the same role as the less similar object. Cross-mappings serve as a stringent test of the structure sensitivity of a representation as they require violating featural or statistical similarity.
We tested the relations that DORA had learned in the previous simulations for their ability to support finding cross-mappings. We selected two of the refined relations that DORA had learned during the previous simulation at random. We bound the relations to new objects, creating two new propositions, P1 and P2 such that the agent of P1 was semantically identical to the patient of P2 and patient of P1 was semantically identical to the agent of P2, and allowed DORA to attempt to map P1 and P 2 . We repeated this procedure 10 times, each time with a different randomly-chosen pair of relations. All 10 times DORA successfully mapped the agents and patients of P1 and P2. The relations DORA learned in the first simulation satisfy the requirement of crossmapping.

We also tested whether the relations that DORA has learned would support mapping to similar but nonidentical relations (such as mapping higher to greaterthan). Humans successfully map such relations (e.g., Bassok, Wu, \& Olseth, 1995; Gick \& Holyoak, 1983), an ability that Hummel and Holyoak $(1997,2003)$ have argued depends on the semantic-richness of relational representations. We selected one of the refined relations that DORA had learned during the previous simulation, R1, and constructed a new relation, R2, that shared $50 \%$ of its semantics (in each role) with the selected relation. So that mappings could not be based on object similarity, none of the objects that served as arguments of the relations had any semantic overlap. We repeated this process 10 times. Each time, DORA mapped the agent role of R1 to the agent role of R2 and the patient role of R1 to the patient role of R2, and, despite their lack of semantic overlap, corresponding objects always mapped to one another (because of their bindings to mapped roles).
Finally, we tested the model's ability to find mappings that violate the $n$-ary restriction: the restriction that an $n$ place predicate may not map to an $m$-place predicate when $n \neq m$. Almost all models of structured cognition follow the $n$-ary restriction (namely, those that represent propositions using traditional propositional notation and its isomorphs; see Doumas \& Hummel, 2005). However, the restriction does not appear to apply to human reasoning, as evidenced by our ability to easily find correspondences between bigger (Sam, Larry) on one hand, and small (Joyce), big (Susan), on the other (Hummel \& Holyoak, 1997).

To test DORA's ability to violate the n -ary restriction, we randomly selected a refined relation (R1) that DORA had learned in the previous simulation. We then created a single place predicate (r2) that shared $50 \%$ of its
semantics with the agent role of R1 and none of its semantics with the patient role. The objects bound to the agent and patient role of R1 each shared $50 \%$ of their semantics with the object bound to r2. DORA attempted to map R1 to r2. We repeated this process 10 times, and each time DORA successfully mapped the agent role of R1 to r2, along with their arguments. We repeated the simulation such that r 2 shared half its semantic content with the patient (rather than agent) role of R1. In 10 additional simulations, DORA successfully mapped the patient role of R1 to r2 (along with their arguments). In short, in all our simulations DORA overcame the $n$-ary restriction, mapping the single-place predicate r 2 onto the most similar relational role of R1.

## Conclusion

We have shown how structured relational representations of magnitude and similarity can be learned from objects with only absolute magnitude values. Our model exploits regularities that emerge in a connectionist network when distributed representations are compared or co-activated. These regularities serve as invariant signals that the model can learn to exploit to bootstrap the detection of relative magnitude differences and similarities. When linked with the DORA predicate learning algorithm, the system learns structured predicate representations of these relative magnitudes and similarities, and then can exploit the resulting representations to solve problems.
Our account provides a trajectory for similarity cognition that maps to cognitive complexity across species and maturational trajectories in humans. This trajectory reveals three distinct levels of abstraction in SDML computation; (i) implicit detection of SDML (responding based on the regular firing that occurs when absolute magnitudes are compared), (ii) implicit generalization of SDML (or learning based on the presence or absence of a particular feature; e.g., learning to respond based on the presence or absence of the 'more' feature), and (iii) predicate representations of SDML (or full-fledged relational representations that support complex cognitive capacities like analogy and reasoning).
This distinction may explain why humans solve some tasks involving similarity judgments without the extensive training that other animals require (e.g., Young, Wasserman, \& Garner, 1997). Humans may solve the task relationally rather than relying on generalized implicit similarity judgments.
Many cognitive architectures and task models rely on stimulus recognition. This theory explains how stimulus recognition might be computed. We believe that providing a computational account for a function existing models depend on represents a significant architectural contribution.

## References

Bassok, M., \& Olseth, K. L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. Journal of Experimental

Psychology: Learning, Memory, and Cognition, 21(6), 1522.

Doumas, L. A., \& Hummel, J. E. (2005). Approaches to modeling human mental representations: What works, what doesn't and why. The Cambridge handbook of thinking and reasoning, ed. KJ Holyoak \& $R G$ Morrison, 73-94.
Doumas, L. A., \& Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. PloS one, 8(6), e63889.
Doumas, L. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological review, 115(1), 1-43
Engel, S.A., Rumelhart, D.E., Wandell, B.A., Lee, A.T., Glover, G.H., Chichilnisky, E.J., \& Shadlen, M.N. (1994). fMRI of human visual cortex, Nature, 369, 525.
Falkenhainer, B., Forbus, K. D., \& Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial intelligence, 41(1), 1-63.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive psychology, 15(1), 1-38.
Goodale, M. A., Ellard, C. G., \& Booth, L. (1990). The role of image size and retinal motion in the computation of absolute distance by the Mongolian gerbil (Meriones unguiculatus). Vision Research, 30(3), 399-413.
Halford, G. S., Wilson, W. H., \& Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences, 21(06), 803831.

Holyoak, K. J. (2012). Analogy and relational reasoning. The Oxford handbook of thinking and reasoning, 234-259.
Holyoak, K. J., \& Thagard, P. (1995). Mental leaps.
Hummel, J. E., \& Holyoak, K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological review, 110(2), 220.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological review, 104(3), 427.
Kemp, C. (2012). Exploring the conceptual universe. Psychological review, $119(4), 685$.
Kemp, C., \& Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. Psychological review, 116(1), 20.
Kim, K. U., Huh, N., Jang, Y., Lee, D., \& Jung, M. W. (2015). Effects of fictive reward on rat's choice behavior. Scientific reports, 5, 8040.
Lake, B. M., Salakhutdinov, R., \& Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science, 350(6266), 1332-1338.
Lu, H., Chen, D., \& Holyoak, K. J. (2012). Bayesian analogy with relational transformations. Psychological review, 119(3), 617.
Medin, D. L., Goldstone, R. L., \& Gentner, D. (1993). Respects for similarity. Psychological review, 100(2), 254.
Von der Malsburg, C. (1999). The what and why of binding: the modeler's perspective. Neuron, 24(1), 95-104.
Wandell, B. (1995). Foundations of Vision. Sinaur Associates Inc.: Sunderland, MA.
Young, M. E., Wasserman, E. A., \& Garner, K. L. (1997). Effects of number of items on the pigeon's discrimination of same from different visual displays. Journal of Experimental Psychology: Animal Behavior Processes, 23(4), 491.

# Co-ordinating Non-mutual Realities: The Asymmetric Impact of Delay on Video-Mediated Music Lessons 

Sam Duffy (s.duffy@qmul.ac.uk)<br>Patrick G. T. Healey (p.healey@qmul.ac.uk)<br>Cognitive Science Group<br>School of Electronic Engineering and Computer Science<br>Queen Mary University of London<br>London, E1 4NS, United Kingdom


#### Abstract

During a music lesson, participants need to co-ordinate both their turns at talk and their turns at playing. Verbal and musical contributions are shaped by their organisation within the turntaking system. When lessons are conducted remotely by video conference, these mechanisms are disrupted by the asymmetric effects of delay on the interaction; in effect a "non-mutual reality" comprised of two different conversations at each end of the link. Here we compare detailed case studies of a copresent and a remote music lesson, in order to show how this effect arises, and how it impacts conduct during the lesson.


Keywords: video mediated communication; conversation analysis; music education; distance learning

## Introduction

When a student and tutor come together for the purpose of an instrumental music lesson, intuition suggests that the principal activity would be playing. However conversation is important, not just as a way to analyse musical contributions, but to organise them within the lesson flow. Participants may respond to talk with performance and vice versa, or even spend periods of time exchanging purely musical contributions (Duffy \& Healey, 2014). For example the tutor could give a verbal instruction that the student should action through performance, or the student could ask a question that the tutor answers through demonstration with their instrument. Activities are managed conversationally; discussion interleaved with performance, demonstration and musical experimentation, resulting in a rich multi-modal social interaction. The musical contributions include unscripted exchanges of short musical fragments intertwined with lesson dialogue. Analysis of their shape and timing shows that they are managed in ways analogous to conversational turn-taking. For example, a tutor's musical contribution can be used to initiate student self-repair in their performance (Duffy \& Healey, 2013). Non-verbal communication such as gaze, or maintaining spatial configurations with respect to each other and the music stand, are also an important part of student-tutor interaction (Duffy \& Healey, 2012).

The transition between speakers is an essential part of the organisation of turn-taking in conversation (Sacks, Schegloff, \& Jefferson, 1974). The preference for just one person to talk at a time requires participants to work together to minimise gaps and overlaps. Anticipating the possible end of a speaker's turn allows a listener to prepare to take the floor when an opportunity presents itself. Interactive turn-taking
phenomena such as backchannels, or making a bid for the floor for a turn at talk, require very precise timing. The timing of the transition between speakers is sometimes referred to as turn offset. It is usually reported as positive if there is a pause between speakers, and negative if there is an overlap (Stivers et al., 2009). Longer pauses and overlaps do occur, but the average turn offset in natural speech tends towards a short pause. A positive turn offset in the range of $0-200 \mathrm{~ms}$ is most likely to be perceived as a smooth turn transition (Stivers et al., 2009; Heldner \& Edlund, 2010).

Remote music tuition using video conferencing is a popular way to support music education in geographically remote areas but has also become an important part of urban mainstream conservatoires, for example to manage temporary separation when students or tutors have to travel to perform, or to manage international auditions. However the medium of communication is known to change aspects of conversational turn-taking, and this has important implications for videomediated remote music tuition (Duffy et al., 2012). Even minor disruptions to the transmission characteristics of the medium of communication, such as the latency and delay associated with video mediated communication, can seriously affect turn-taking (Whittaker, 2003).

Qualitative video analysis and conversation analysis (CA) have been used to examine video-mediated workplace communication (Heath \& Luff, 1991) and how participants complete collaborative tasks in video-mediated environments (O’Conaill, Whittaker, \& Wilbur, 1993; Heath, Luff, \& Sellen, 1997; Ruhleder \& Jordan, 2001). However there have been relatively few studies of the detailed effects of videomediated communication (VMC) on the timing of conversational turn-taking, and the results are inconsistent, driven by subtle differences in experimental set up. For example, some studies include signal delay (Ruhleder \& Jordan, 2001) whilst others exclude it (Sellen, 1992); in some cases specifically to isolate other interactional factors. Some studies compare video-mediated interaction (with or without delay) to same-room conditions, whilst others compare it to other lower quality remote communication systems or audio only scenarios (O'Conaill et al., 1993; O'Malley, Langton, Anderson, Doherty-Sneddon, \& Bruce, 1996). Studies which have included delay as part of their experimental set-up (O'Conaill et al., 1993; Ruhleder \& Jordan, 2001) suggest a further subtle effect; changes in the time of arrival of utterances with
respect to 'local' sound. Ruhleder and Jordan (2001) suggest that two people having two fundamentally different conversations with each other raises serious questions about what it means to 'share' a conversation in distributed settings. This leads to some interesting questions in terms of remote music lessons. How might the medium change the turn transitions observed in co-present lessons when they are mediated by video conference? How might the transition between 'speakers' be affected by the inclusion of musical contributions?
In order to investigate these questions, a detailed study was made of student-tutor interaction during a co-present and a remote music lesson, using CA and qualitative video analysis. CA has previously been used to examine aspects of instrumental music tuition (Ivaldi, 2014; Nishizaka, 2006; Szczepek Reed, Reed, \& Haddon, 2013), as well as the effect of medium on conversational turn-taking. This fine grained analysis of a same-room and a separated lesson allows us to examine both the turn-taking characteristics unique to a music lesson, and how these are affected by the medium of video. This work is part of a larger study of a number of co-present and remote music lessons (Duffy, 2015).

## Methodology

Two one-to-one lessons featuring woodwind instruments were observed, filmed and analysed in detail; a co-present ('same room') lesson and a video-mediated remote lesson. The co-present lesson featured a male student studying ABRSM grade 8 clarinet performance and was filmed during one of his regular weekly lessons at the junior school of a London Conservatoire. The female tutor had taught the student for many years. The remote lesson featured a female oboe student taking part in an ensemble residency with Aldeburgh Young Musicians, filmed during a remote music tuition study at Aldeburgh Music in Suffolk (Duffy et al., 2012). The student had been working with the tutor during the residency, but had not previously taken regular lessons with her. Both students had advanced to a similar level of proficiency; they were largely comfortable with the technical challenges of their instrument and capable of exploring musicality and expression. Both tutors were experienced in one-to-one tuition, but not video mediated tuition. The scope of this study was to examine student-tutor interaction, and did not consider teaching effectiveness between conditions.

Conversational turns are defined as the period during which a participant holds the floor, until there is a change in speaker (Sacks et al., 1974, pp.702-703). Turns in the footage from each lesson were coded using ELAN (Brugman, 2004). A separate tier was created for analysis of each of the following types of contribution: student talk, tutor talk, student play and tutor play. This data was exported as a transcript with timecode information so that calculations could be made such as turn frequency, mean turn duration and turn onset in relation to the preceding turn. Pauses between turns were coded as a positive offset, and overlap as a negative offset, similar to the approach used by Stivers et al. (2009). This allowed calcula-
tion of a net offset for a period of time or subset of turn types. Backchannels were excluded from the distribution, similar to the approach used by Sellen (1992), since they are not a bid for the floor or intended to initiate a change in speaker. As discussed, a difference between this analysis and existing literature is that we consider the transitions between musical, as well as verbal contributions. As a result, the following categories of turn transition were identified:

1. Talk following talk.
2. Talk following play.
3. Play following talk.
4. Play following play.

Established notation for conversation analysis, as described in the appendix of Sacks et al. (1974), was adapted to analyse musical contributions to lesson dialogue (Table 1).

| $(0.2 \mathrm{~s})$ | Elapsed time (seconds) used to denote pauses or silence <br> $-(1.4 \mathrm{~s})$ |
| :--- | :--- |
| Long single note and duration |  |
| $\uparrow--(2.3 \mathrm{~s})$ | Individual notes in a musical phrase and phrase duration <br> Rising passage of notes |
| ,,$-(1.2 \mathrm{~s})$ | Falling passage of notes <br> in-breath in preparation to play, and duration <br> onset of 'talk over play' overlap |
| $\overline{\{\text { first octave }\}}$ | Additional information for music notation <br> duration of period of overlap |
| $=0.6 \mathrm{~s}]$ | Latching (no interval between two pieces of talk) |

Table 1: Transcription notation.
The two rooms used for the remote lesson were adjoining suites at the same organisation (see Duffy et al. (2012) for more details). A separate video camera was placed in each suite, in addition to the video conference equipment, in order to capture student and tutor position with respect to the screen and provide a separate audio recording for each location. There was a small delay in visual processing caused by additional software being tested during the lesson. A delay was added to the audio so that audio and visuals arrived synchronised in each location. Audio samples from each room were synchronised and analysed using clearly visible audio transients which did not overlap with local sounds. Whilst the rooms were geographically close, the delay was of the same order as the latency experienced in a typical transatlantic video call ( 0.9 s ). This delay was constant, but in reality the magnitude of the delay would vary somewhat over the duration of the call, depending on the signal journey through different servers and exchanges.

## Results

First we will look at some general effects of the medium on the lessons analysed. Whilst the co-present lesson was slightly longer than the consolidated sections of the videomediated class analysed, they both contained similar proportions of instances of turns at talk ( $73 \%$ and $71 \%$ table 2 ) and instances of musical contributions ( $27 \%$ and $29 \%$ table 2 ). However the turn structure within this was quite different. The co-present lesson contained 753 turns in total whilst the
video mediated lesson contained just 234, and the average length of both turns at talk and musical contributions were significantly longer for the video-mediated lesson. Net mean offset for the remote lesson was 337 ms , 143 ms longer than the co-present lesson offset of 194 ms (table 3). These results are consistent with findings that video-mediated conversations are characterised by fewer turns of greater length and reduced overlapping speech (Cohen, 1982; O'Conaill et al., 1993; Sellen, 1992).

Table 2: Turn structure of co-present vs. remote lesson.

|  | co-present | remote |
| :--- | :---: | :---: |
| instrument | clarinet | oboe |
| total duration (mins) | 36 | 27 |
| number of turns at talk | 550 | 165 |
| as a \% of total turns | $73 \%$ | $71 \%$ |
| number of musical contributions | 203 | 69 |
| as a \% of total turns | $27 \%$ | $29 \%$ |
| total lesson contributions | $\mathbf{7 5 3}$ | $\mathbf{2 3 4}$ |
| average length of turns at talk (s) | 2.0 | 4.4 |
| average length of musical contributions (s) | 4.7 | 6.7 |
| total lesson average contribution length (s) | $\mathbf{2 . 8}$ | $\mathbf{5 . 1}$ |

Table 3: Net offset duration (ms) by transition type.

|  | talk <br> following <br> talk | talk <br> following <br> play | play <br> following <br> talk | play <br> following <br> play | total <br> lesson |
| ---: | :---: | :---: | :---: | :---: | :---: |
| co-present |  |  |  |  |  |
| n | 345 | 129 | 134 | 50 | 658 |
| $\%$ | $52 \%$ | $20 \%$ | $20 \%$ | $8 \%$ |  |
| mean (ms) | 287 | -61 | 297 | -70 | 194 |
| remote |  |  |  |  |  |
| n | 64 | 55 | 62 | 4 | 185 |
| $\%$ | $35 \%$ | $30 \%$ | $34 \%$ | $2 \%$ |  |
| mean (ms) | 39 | -40 | 993 | 124 | 337 |

Next we examine the net offset by transition type. In the co-present lesson, the net offset for turn transition type talk following talk, representing periods of student-tutor discussion, was 287 ms (table 3). This was slightly outside the range of $0-200 \mathrm{~ms}$ from the literature, but still showed a preference for a pause of the same order of size. For transition type play following talk, for example a student performing in response to a verbal instruction from the tutor, the co-present net offset was again in line with the literature ( 297 ms , table 3 ). In the remote lesson, the net offset for transitions of talk following talk decreased to 39 ms in line with our expectations from the literature, but the net offset for play following talk lengthened considerably to 993 ms .
Looking specifically at overlap by participant, the student showed a preference to play over tutor talk (33 instances of student play over talk overlap compared to 2 tutor instances table 4). One explanation for this is that in co-present lessons tutors were found to make long instructional turns to initiate student play, comprised of several utterances separated by
pauses, interspersed with backchannels by the student. The backchannels were placed with precision to show attentiveness without making a bid for the floor or disrupting the tutor's turn. Non-verbal cues enabled the student to determine when these turns were complete and they should start to play (Duffy, 2015, pp. 140-148). As the next example shows, in the remote lesson the student found this more difficult.

Table 4: Overlap duration (ms) by activity by participant

|  |  |  | talk <br> over <br> talk | talk <br> over <br> play | play <br> over <br> talk | play <br> over <br> play | $\begin{aligned} & \text { total } \\ & \text { lesson } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| co-present | student | n | 47 | 1 | 33 | 15 | 96 |
|  |  | mean (ms) | 330 | 65 | 637 | 524 | 463 |
|  | tutor | n | 42 | 46 | 2 | 9 | 99 |
|  |  | mean (ms) | 317 | 489 | 391 | 436 | 409 |
|  | total | n | 89 | 47 | 35 | 24 | 195 |
|  |  | mean (ms) | 324 | 480 | 623 | 491 | 436 |
| remote | student | n | 12 | 1 | 5 | 2 | 20 |
|  |  | mean (ms) | 780 | 1,326 | 615 | 466 | 735 |
|  | tutor | n | 10 | 16 | 1 | - | 27 |
|  |  | mean (ms) | $1,113$ | 478 | 1,234 | - | 741 |
|  | total | n | 22 | 17 | 7 | 2 | 48 |
|  |  | mean (ms) | 877 | 528 | 647 | 467 | 703 |

## Instructional turns

The tutor asked the student to play a scale. When the tutor paused after her first utterance, the student made physical preparations to play such as stepping back from the screen and raising her hands to the instrument body (transcript 1: line 1 and transcript 2 : line 2). However the tutor retained the turn, choosing to demonstrate by playing the scale herself (transcript 1 and 2: line 3). Towards the end of the scale the student raised her clarinet to her mouth again, this time placing the reed in her mouth (transcript 1: line 4 and transcript 2: line 5). However the tutor started a new utterance "so I mean you go C sharp to C sharp" and the student lowered her oboe again. This was the second abandoned attempt to play. At the end of this utterance the student nodded and raised her oboe for a third time. She placed the reed in her mouth and took an in-breath whilst the tutor talked (transcript 1: line 6). The tutor made no further utterances and finally the student moved into playing the scale. From the tutor room footage it was not clear why the student made two preparations to play which could not be followed through. From the student room footage it was clear that the student was placing these actions in the pauses that were interpreted as the end of the tutor's instructional turn. This happened several times, and towards the end of the lesson the student exclaimed "sorry sorry it's hard to know when to play". This example is analysed in more detail in Duffy (2015, pp. 350-359).

## Bidding for the floor to provide feedback

Transitions involving turns following play did not follow the literature. Both talk following play and play following play

```
1. T: can you just play me a scale [starting on top A?]
                                    [((S steps back from screen))]
    [(1.0s)]
    [((S lifts second hand onto instrument body))]
2. T: [u::m in fact]
    [((S steps back from screen))]
    [((T raises oboe))]
    (0.4s)
    ((sucks reed loudly twice))
    (0.9s)
    [((S raises oboe))]
```



```
    [((*))] [ ((**))]
```



```
5. T: [so I mean you go C#] to [C# [it's still A major]]
    [((** ))] [ (( S nods )) ]
6. T: [it's just A major] [let's just have a listen]
    [ (( *** )) ] [((S takes an in-breath))]
```

Transcript 1: The tutor initiates a scale - student room audio.

```
1. T: can you just play me a scale starting on top A?
2. T: [(1.0s)]
    [((S steps back from screen))]
    [uuum] in fact
    [((S lifts second hand onto instrument body))]
    (0.4s)
    ((sucks reed loudly twice))
    (0.9s)
```




```
5. T: [so I mean you go C#] [to C#] [its still A major]
    [((***))] [((**))] [((S nods))]
6. T: [[it's just A major]] [let's just have a listen]
    [ (( S nods )) ]
    [ ((*)) ] [ ((*** )) ]
        *S lifts oboe to playing position
        **S lowers oboe, keeping both hands on keys
        ***S places reed in mouth
```

Transcript 2: The tutor initiates a scale - tutor room audio
tended towards overlap in the co-present lesson, rather than a short pause ( -61 ms and -70 ms table 3). Talk following play tended towards overlap in the remote condition ( -40 ms table 3). Play following play tended towards a pause in the remote condition (net offset 124 ms ) but the proportion of this type of turn was significantly reduced to just $2 \%$, or 4 turns. All but one incidence of talk over play overlap was made by the tutor, in both the co-present and remote lesson (table 4), evidencing the tutor's preference to talk over student play when a problem was been diagnosed in order to provide feedback. This did not appear to be as disrupted by the medium as the previous example, perhaps because the length of the note during which the tutor bid for the floor was often of the same order as the duration of the delay (Duffy, 2015, pp. 245-263), so the tutor's interruption still arrived before the student could start the next musical phrase. What is beginning to emerge is asymmetry in the preferences for taking a turn to talk or play between the participants, some of which are disrupted more by the medium than others.

## Local differences in turn placement

Next we will look at an example which demonstrates the effect of the delay on the placement of a single turn. Audio waveforms from each room illustrate the effect in addition to the transcripts (figure 1). Coloured blocks have been annotated using Logic Pro 9 to highlight the different position of parts of the dialogue shown in transcript 3. The tutor waveform is narrower because the camera in the tutor room was further away from where the tutor was positioned, as a result the waveform has smaller amplitude (vertical height representing volume). This does not affect our analysis. The two audio samples were synchronised using the visual transients in the tutor's utterance "ba ba ba ba ba ba" (line 3 of transcript 3). Transcript 3 shows the difference in turn transition and sequence between the two rooms.

```
Audio from the camera in the student's room
1. T: 'cause the A sharp is always there from the crotchet rest ((*))
2. S: yeah fine
    (0.2s)
3. T: You've just got it so actually you think ba ba ba ba ba ba
Audio from the camera in the tutor's room
1. T: 'cause the A sharp is always there from the crotchet rest ((*))
        (0.8s)
2. S: [yeah fine]
3. T: [You've just got it] so actually you think ba ba ba ba ba ba
* a door slams shut as an observer leaves the room
```

Transcript 3: Turn sequence discrepancy between rooms
There are two main differences between the audio samples. The first relates to the student's utterance "Yeah fine" in line 2 (circled section of the waveform in figure 1). This utterance was made with respect to the tutor's turn in line 1 . In the student room audio the response followed straight on, after the noise of a door slamming at the end of the tutor's turn. In the


Figure 1: Audio discrepancy between rooms.
tutor room audio there was a 0.8 second pause after the tutor said "Cause the A sharp is always there from the crotchet rest". When the student utterance "Yeah fine" arrived, the tutor had already started talking again, so it overlapped with the start of the tutor's next comment "You've just got it". From the student's perspective, she had replied as soon as she heard the tutor's comment. However her reply was delayed in its return to the tutor by 0.9 s . When the student's response arrived, the tutor had already started to talk again having only heard silence, and so she talked over the student's response. As a result, the student's utterance "Yeah fine" was placed within a pause in the tutor's speech, but transformed into an overlap with tutor speech when received in the tutor's room. Several examples of similarly misapplied feedback are reported in Ruhleder and Jordan (2001).

The next example shows how turn sequence can be changed. Examining the student audio first, the musical phrase in line 1 of transcript 4 includes a pause notated in the score before a phrase is repeated. The student makes this pause 0.4 seconds in duration and starts the repeated phrase in line 2 . However the tutor appears to talk over this second phrase with "May-maybe a" (line 3). This is unusual, the tutor usually waits until the end of a musical phrase to start talking, the only overlap being with the final note (Duffy \& Healey, 2013); here the tutor starts talking mid-phrase. It is also unusual that the student does not stop playing, instead the tutor stops talking and the student continues. The tutor interjects again with "yeah" but the student still continues. The tutor then talks again straight after the last note of the phrase. Now the student stops playing, immediately looking up from the music and at the screen.
Looking at the tutor room audio, shown in the second half of transcript 4 , we see that the tutor started the utterance "May-maybe a" during the notated pause in the student's performance (transcript 4: line 2a). However the delay in trans-


Transcript 4: Relative position of tutor interruption.
mission of this utterance to the student room, meant that it arrived after the student had started to play the next phrase in line 2. In the tutor's room "May-maybe a" was interrupted by the student starting her second phrase in line 3 a and the tutor stopped talking. Her next utterance "yeah" in line 4a started during a long note played by the student, which could be interpreted as a bid for the floor. However when this utterance arrived in the student room, the long note was already complete and the student had moved on to the next phrase. From the tutor's perspective she had tried unsuccessfully to take the floor at the end of the first phrase.

## Discussion

The short fragments of music which occur during an instrumental lesson have been shown previously to be managed conversationally, and share some characteristics with turns at talk. Here we see that participants exhibit different preferences for how they manage transitions between verbal and musical contributions. The tutor more often leads lesson
flow, placing more of the responsibility for turn placement onto the student in their responses. The tutor is also more likely to bid for the floor during student play, whereas the student rarely interrupts the tutor in talk or play. Differences in preferences have also been reported in turn-taking associated with the roles of the teacher and students in a classroom (McHoul, 1978). The signal delay associated with VMC disrupts these preferences, exhibiting a greater effect on the student. Ruhleder and Jordan (2001) suggest that the mechanisms which are most affected by signal delay are conversational turn-taking, sequence organisation and repair; affecting trust and confidence between the participants. The phenomenon analysed here may explain student frustrations previously reported during remote music lessons (Duffy \& Healey, 2012). This study highlights a number of opportunities for further work. For example, it is not known if participants could acclimatise to aspects of the disruption to lesson interaction over time. A longitudinal study is recommended which follows student-tutor pairs taking both co-present and remote lessons. In this way, any effect caused by change in participants across conditions will also be controlled. There may also be different, more effective, ways to represent the naturalistic teaching interaction remotely through alternative technologies (Duffy \& Healey, 2017).

## Acknowledgments

This work is supported by the Media and Arts Technology programme, EPSRC Doctoral Training Centre EP/G03723X/1.

## References

Brugman, H. (2004). Annotating multimedia/multi-modal resources with ELAN. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004). Portugal.
Cohen, K. (1982). Speaker interaction: video teleconferences versus face-to-face meetings. In Proceedings of teleconferencing and electronic communications (pp. 189-199).
Duffy, S. (2015). Shaping Musical Performance Through Conversation. Doctoral thesis, Queen Mary University of London.
Duffy, S., \& Healey, P. (2012). Spatial Co-ordination in Music Tuition. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (pp. 1512-1517). Sapporo: Cognitive Science Society.
Duffy, S., \& Healey, P. (2013). Using Music as a Turn in Conversation in a Lesson. In Proceedings of the 35th annual conference of the cognitive science society (pp. 22312236). Berlin: Cognitive Science Society.

Duffy, S., \& Healey, P. (2014). The Conversational Organisation of Musical Contributions. Psychology of Music, 42(6), 888-893.
Duffy, S., \& Healey, P. G. (2017). A New Medium for Remote Music Tuition. Journal of Music, Technology and Education, 10(1), (in press).

Duffy, S., Williams, D., Stevens, T., Kegel, I., Jansen, J., Cesar, P., \& Healey, P. (2012). Remote Music Tuition. In Proceedings of the 9th sound and music computing conference (pp. 333-338). Copenhagen: smcnetwork.org.
Heath, C., \& Luff, P. (1991). Disembodied conduct: communication through video in a multi-media office environment. In J. S. Robertson, Scott P., Olson, Gary M. and Olson (Ed.), Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology (pp. 99-103). New Orleans: ACM.
Heath, C., Luff, P., \& Sellen, A. J. (1997). Reconfiguring media space: Supporting collaborative work. Video-mediated communication, 323-347.
Heldner, M., \& Edlund, J. (2010). Pauses, gaps and overlaps in conversations. Journal of Phonetics, 38(4), 555-568.
Ivaldi, A. (2014). Students' and teachers' orientation to learning and performing in music conservatoire lesson interactions. Psychology of Music, 44(2), 202-218.
McHoul, A. (1978, December). The Organization of Turns at Formal Talk in the Classroom. Language in Society, 7(2), 183-213.
Nishizaka, A. (2006). What to Learn: The Embodied Structure of the Environment. Research on Language \& Social Interaction, 39(2), 119-154.
O’Conaill, B., Whittaker, S., \& Wilbur, S. (1993). Conversations Over Video Conferences : An Evaluation of the Spoken Aspects of Video-Mediated Communication. Human Computer Interaction, 8, 389-428.
O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., \& Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. Interacting with Computers, 8(2), 177-192.
Ruhleder, K., \& Jordan, B. (2001). Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction. Computer Supported Cooperative Work (CSCW), 10(1), 113-138.
Sacks, H., Schegloff, E. A., \& Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50(4), 696-735.
Sellen, A. J. (1992). Speech patterns in video-mediated conversations. In CHI '92 (pp. 49-59). Monterey: ACM.
Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences of the United States of America, 106(26), 10587-92.
Szczepek Reed, B., Reed, D., \& Haddon, E. (2013). Now or Not Now: Coordinating Restarts in the Pursuit of Learnables in Vocal Master Classes. Research on Language \& Social Interaction, 46(1), 22-46.
Whittaker, S. (2003). Theories and Methods in Mediated Communication. In A. C. Graesser, M. A. Gernsbacher, \& S. R. Goldman (Eds.), The handbook of discourse processes (pp. 243-286). Lawrence Erlbaum Associates Inc.

# When Do Vehicles of Similes Become Figurative? Gaze Patterns Show that Similes and Metaphors are Initially Processed Differently 

Frank H. Durgin (fdurgin1@swarthmore.edu)<br>Swarthmore College, Department of Psychology, 500 College Avenue<br>Swarthmore, PA 19081 USA<br>Rebekah Gelpi (rgelpi1@swarthmore.edu)<br>Swarthmore College, Department of Psychology, 500 College Avenue<br>Swarthmore, PA 19081 USA


#### Abstract

Recent emphases on differences between metaphors and similes pose a quandary. The two forms clearly differ in strength, but often seem to require similar interpretations. In Experiment 1 we show that ratings of comprehensibility are highly correlated across simile and metaphor sentences differing only in the presence or absence of "like". In Experiment 2 we show that comprehensibility ratings for figurative forms predict both early (first pass) and late (second pass) fixation durations for metaphor vehicle, but only late fixation durations for vehicles in similes. Simile vehicles appear to initially be processed similarly to literal comparisons, with figurative interpretation occurring later. These observations are consistent with the different pragmatic strengths, and similar interpretations of the two forms.


Keywords: simile; metaphor; analogy; career of metaphor, implicature, eye-movements

## Introduction

Theories of figurative speech differ in emphasizing either (abstract) categorization (e.g., Glucksberg \& Keysar, 1990) or analogical comparison across domains (e.g., Tourangeau \& Sternberg, 1982). Bowdle and Gentner (2005) proposed that unfamiliar figurative usages tend to be preferred in simile form ("Moonlight is like bleach") whereas familiar ones are preferred in metaphor form ("Alcohol is a crutch"). They suggest that the surface form of a simile mirrors the cognitive processing (analogical reasoning) needed for an unfamiliar figurative meaning.

But it isn't the case that mere comparison captures analogical comparison. Aristotle is sometimes described dismissively as a comparison theorist (e.g., Tourangeau \& Sternberg, 1982), but Israel, Harding and Tobin (2004) rightly point out that Aristotle is a metaphor-first theorist. Aristotle has also been characterized as considering metaphors to be implicit analogies (Levin, 1982). Similes, on this view, might be figurative to the extent that they require cross-domain implicit analogical reasoning.

Glucksberg and Haught (2006) observed that adding an off-category adjective (e.g., "Moonlight is (*like) romantic bleach") disrupts the preference scheme identified by Bowdle and Gentner (2005). From this Glucksberg and Haught seek to argue that figurative categorization is the only plausible account of metaphor. They argue that their adjectival noun-phrases are dis-preferred in simile form
because similes refer to literal referents. But non-existing categories like "romantic bleach" block comparison generally, and do not differentiate literal from figurative comparison. Moreover, they seem to violate the need for distinct domains required for analogy to work.

In this paper we will adopt the strategy of contrasting online comprehension of similes both with metaphor and with literal comparisons (as recommended by Israel et al., 2004). By this means we will test whether similes are simply interpreted literally, as Glucksberg and Haught (2006) argued, or if they also differ from literal comparisons in ways that clarify their normal designation as figurative.

In Experiment 1 we will show how similar the comprehensibility ratings of associated metaphor and simile forms are. Despite previously recognized differences in strength (e.g., Glucksberg \& Keysar, 1990), aptness (e.g., Kennedy \& Chiappe, 1999), and preference (e.g., Bowdle \& Gentner, 2005), we find that comprehensibility judgments are relatively similar across both simile and metaphor forms for simple vehicles without adjectival modification. This suggests similar interpretations are reached for the figurative meaning of the vehicle in each figurative form.

In Experiment 2, we will use measures of gaze during reading to show that initial reading of similes more closely resembles that for literal comparisons than that for metaphors. Metaphors show reliable effects of figurative comprehensibility during initial first-pass reading; similes and literal comparisons do not show such comprehensibility effects early on. However, during second-pass reading, fixation durations on simile and metaphor vehicle both show strong correlations to ratings of figurative comprehensibility. This suggests that the figurative interpretation of a simile vehicle may simply be computed later than that of a metaphor.

## Experiment 1: Comprehensibility Ratings

## Methods

Materials (for Experiments 1 and 2) Seventy-five metaphors were gathered or updated from previous studies (Katz, Paivio, Marschark \& Clark, 1988; Bowdle \& Gentner, 2005; Thibodeau and Durgin, 2011). For gaze analysis (Experiment 2), each sentence was extended with a few words intended to be largely neutral with respect to
interpreting the sentence (e.g., "A smile is (like) a magnet for people."). A set of literal comparison statements was developed using the same vehicles (e.g., "A black hole is like a magnet in space."). To balance the number of sentences that did not include "like" and were not figurative, we also included 25 literal categorization statements filler sentences unrelated to the 75 experimental items. Three lists of 100 items were constructed in which one third of the 75 items appeared in each of the three forms (metaphor, simile, literal comparison) along with the 25 fillers.

Participants and Task A total of 91 adults were recruited through Amazon's Mechanical Turk to make ratings of comprehensibility of the critical word in each sentence. A third were assigned to each of the three lists of stimuli.

## Results and Discussion

As shown in Figure 1, comprehensibility ratings of metaphor and simile vehicles were highly correlated across items, $R_{73}=0.85, p<.001$. A common factor, produced by averaging the figurative rating sets by item accounted for $92 \%$ of the variance in the simile ratings and $93 \%$ of the variance in the metaphor ratings. There was, of course, no correlation between this common rating measure for figurative vehicles and the ratings given for the same words in literal comparisons with the same vehicles, $R_{73}=0.01$, ns.

It appears that ratings for both figurative sentence forms are typically based on similar figurative meaning, defined in relation to the topic.


Figure 1. Correlation between comprehensibility ratings for each figurative vehicle across figurative sentence forms.

## Experiment 2: Gaze Measures when Reading Comparisons, Similes and Metaphors

The rating data of Experiment 1 are consistent with the assumption of many scholars that there is good reason to suppose that similar figurative meanings are often achieved by sentences in simile and metaphor forms, even if they may sometimes diverge in understanding. However the rating data reflect post-interpretive evaluations and do not bear on the question of whether the initial cognitive encounter with
the vehicle (e.g., during reading of the metaphor) is substantially different in similes and metaphors. In order to better understand how comprehension may unfold differently for the two figurative forms, we next measured gaze patterns during reading of these same sentences.

## Methods

Participants Thirty-six Swarthmore College undergraduate students who were native English speakers participated in partial fulfillment of an introductory course requirement.

Design and Procedure The linguistic materials were identical to those used in Experiment 1, except that 6 practice items were developed to allow participants time to adapt to the task. One third of the 75 experimental items appeared in each of three forms (metaphor, simile, literal comparison) for each participant, according to one of three lists, and were randomly ordered and intermixed with 25 (filler) literal categorization sentences. Sentences were presented one at a time on a monitor in front of the participant after first establishing gaze on a fixation point just to the left of the presentation of the sentence. Participants were to read the sentence and respond by pressing a key when they had comprehended it. The sentence was then removed, and an easy multiple-choice comprehension test followed, asking which of four terms was most relevant to the meaning of the sentence they had just read (e.g., for "A smile is (like) a magnet for people", the correct answer was "attract"). Subjects made their choice using a game-pad with an appropriate spatial mapping to the choices on the screen. The entire experimental session took about 20 minutes.

Gaze Recording Gaze was tracked at 1000 Hz , using an Eyelink 1000 (SR Research). The experimental code was implemented in Experiment Builder (SR Research), and gaze parameters during reading were extracted using custom software in conjunction with the area-of-interest definition tools provided by Experiment Builder.

## Results and Discussion

Analysis Strategy Our principal interest was to compare the immediate processing of similes to the processing of metaphors, as measured by gaze parameters, and to secondarily use literal comparisons as an alternative comparison condition for similes. We used item-wise rating data from Experiment 1 as a predictor in LMER models. For the figurative items these ratings were averaged across figurative conditions. For gaze variables, we first used model comparisons to test whether the ratings had predictive value that differed between figurative sentence types. For example, if metaphor vehicles are treated as figurative words, while simile vehicles are treated as literal words, we might expect ratings of figurative interpretability from Experiment 1 to predict only the metaphor forms, and not the simile forms. To test for this we compared LMER models that included an interaction term (between ratings
and sentence type) with those that did not. Such model comparisons produce a Chi-square statistic. We also used LMER modeling to compare similes with literal comparisons. For overall comprehension time, we conducted a single overall analysis since comprehension

Gaze Behavior: Overview. The analysis of gaze behavior data is organized into five reading events relative to the target word (i.e., the vehicle) region: (1) Duration of initial fixation(s) on the critical word, (2) the subsequent frequency of regressions back out to the left (3) the duration of time


Figure 2. Mean comprehension times (mean response latency computed in log space by item) for metaphors (left), similes (middle), and literal comparisons (right) as a function of item comprehensibility (ratings). Best fit and SE shown.
time was expected to be correlated with comprehensibility ratings for all sentences.

Comprehension Time Participants' main task was simply to indicate comprehension of the sentence after reading it by pressing a key. The distribution of times was skewed, so centered $\log$ transformations of these times were used for statistical modeling. The main LMER model included sentence type (metaphor, simile and comparison) and comprehensibility ratings from Experiment 1 as predictors.

Error terms included subjects and item as well as the slopes for sentence form by subjects and by items, and the slopes of ratings by subjects. Model comparisons showed that including the interaction between sentence type and rating did not explain reliably more variance than a model without the interaction, $X^{2}(2)=0.27, p=.875$, indicating that the relation to ratings did not differ by sentence type. Rather for all three sentence types, ratings predicted comprehension time, $t(117.3)=3.54, p<.001$ (Satterthwaite approximations of $d f$ will be reported throughout; see Luke, 2006). However, compared to the similes, response times were reliably longer for the literal comparison sentences, $t(48.6)=3.94, p<.001$. Consistent with effects previously observed for apt or conventional figurative vehicles (e.g., Bowdle \& Gentner, 2005; Glucksberg \& Haught, 2006), comprehension time was marginally shorter for the metaphors than similes, $t(35.1)=$ 1.93, $p=.062$. The observed relationship of rated comprehensibility and comprehension time is shown in Figure 2 separately for each sentence type. These data reflect the expected relationships between comprehensibility ratings and comprehension time.
between first fixating the critical word and finally reading past the word, (4) the likelihood of refixating the word after passing it, and (5) if refixation occurred, the total time taken fixating the word after first passing it.

Our primary interest is in differences associated with the presence of "like" in advance of the figurative word, since the simile and metaphors forms used are otherwise identical. Comparisons between the literal comparisons and similes are also of interest, however, given that we are asking whether the figurative vehicle words in similes are initially processed literally.

Gaze Data Transformation and Truncation Duration data associated with gaze patterns were also log transformed to reduce skewing and were centered for analyses. Transformed durations that were more than 4 standard deviations above the transformed mean for that measure were truncated to 4 SDs. The proportion of data affected by this method was less than $1 \%$ each measure discussed. Ratings were centered (by subtracting off the mean) prior to analysis.

Gaze Behavior 1: First Fixation Duration. If participants initially seek to understand metaphor vehicles figuratively, but simile vehicles literally, the duration of their first fixation on the critical word might correlate with ratings of comprehensibility for figurative items only for metaphors.
Consistent with this hypothesis, comparison of LMER models of the figurative sentence data, with and without interactions terms, showed that the relationship of FFD to ratings differed for the two figurative forms, $X^{2}(1)=7.32$, p $=.007$. As expected, separate models of the two condition indicated that FFD for the figurative vehicle was related to
the rated comprehensibility in the metaphor form, $t(35.6)=$ $4.06, p<.001$, and not in the simile form, $t(68.0)=0.10$, ns. Models of FFD contrasting the two comparison sentence forms (literal comparisons and similes), found no differences in FFD between the two forms $\left(X^{2}(2)=2.74, p=\right.$
the various forms. Indeed, model comparisons indicated that the relation of GPT to rated comprehensibility differed reliably between simile and metaphor forms, $X^{2}(1)=9.8, p$ $=.002$. In these models, GPT was also reliably longer for metaphors than similes, $t(37.6)=2.84, p=.007$. Separate


Figure 3. Top row: First fixation duration (FFD). Bottom row: Go Past Time (GPT). Geometric means (by item and sentence type) for the figurative vehicle (or equivalent) are shown as a function of mean rated comprehensibility.
.254. The data for each form are plotted in the top panels of Figure 2. This pattern is consistent with the idea that the figurative vehicle in a simile is initially treated as a literal referent, as argued by Glucksberg and Haught (2006a). In contrast, initial encounters with metaphor vehicles produced effects consistent with an immediate search for a figurative interpretation.

Gaze behavior 2: Go past time. Go past time (GPT) is defined as the entire duration from first entering the critical word until finally passing to the right of the critical word. GPT for each sentence form is shown in bottom panels of Figure 3 as a function of comprehensibility rating. Because GPT includes FFD, FFD was included as a covariate in LMER analyses of GPT (the analyses come to the same conclusions without the covariate).

Again, we first sought to test whether there were reliable relationships between GPT and rated comprehensibility for figurative items, and, if so, whether these differed between

LMER models showed that, for metaphor sentences, GPT was reliably related to figurative comprehensibility ratings, $t(64.4)=2.60, p=.011$. This was not the case for similes, $t(74.1)=1.49, p=.141$ (where the trend was in the opposite direction, consistent with delays for highly conventional metaphors presented as similes).

Models comparing GPT for the similes and literal comparison sentences5 indicated that including the interaction of ratings and sentence type in the models made no reliable difference, $X^{2}(1)=2.75, p=.097$. A separate model of literal comparisons confirmed that here was also no reliable relationship between GPT and rated comprehensibility, $t(76.9)=0.70, p=.486$. Thus, prior to exiting the critical word to the right in the simile form, the figurative vehicle may still be treated primarily as a referent to the literal comparison category as the reader passes on the rest of the sentence.

Gaze Behavior 3: Returns to the Target Word From the Right Participants often returned to the critical word after they had already read beyond it. Indeed, this happened in roughly $59 \%$ of trials ( $60 \%$ of metaphor trials, $58 \%$ of literal comparison trials and $57 \%$ of simile trials). Is the likelihood of such Regressions In (RI: if the critical word was refixated on a given trial after exiting to the right) related to rated comprehensibility? LMER models of RI for the figurative sentences showed a reliable relationship existed between rated comprehensibility and the likelihood of RI, $t(59.9)=2.29, p=.026$, and that this effect did not reliably differ between metaphors and similes, $X^{2}(1)=$ 0.003 , $n s$. But LMER models of RI for comparison sentences also showed a reliable main effect of rated comprehensibility, $t(38.4)=2.27, p=.029$, and no evidence of an interaction, $X^{2}(1)=0.04, n s$. Thus, RI was more likely, for all sentence forms, as comprehensibility decreased.

Gaze Behavior 4. Second Pass Total Time (P2TT) Given that a reader had refixated the critical word after having read more of the sentence, was the total time spent refixating it before responding related to rated comprehensibility? Total comprehension time was included as a covariate because it was highly correlated with P2TT. Whereas FFD and GPT both distinguished similes from metaphors, understanding a simile typically requires reaching a similar understanding to the understanding required for a metaphor. When during reading comprehension might this happen?

LMER models of P2TT for the figurative sentences indicated that the relation between P2TT and comprehensibility ratings was highly reliable for these two forms, $t(63.0)=2.77, p=.007)$, but did not differ between the figurative sentence forms, $X^{2}(1)=0.37, p=.548$. Thus, P2TT appears to be similarly related to comprehensibility ratings for similes and metaphors, as shown in Figure 4. In contrast, LMER models of the comparison statements (i.e., similes and literal comparisons analyzed together) failed to show reliable relationship between ratings and P2TT, $t$ (39.9)
$=1.71, p=.095$, but also failed to detect reliably different effects of ratings for similes and literal comparisons, $X^{2}(1)=$ $0.36, \mathrm{p}=.548$.

To resolve the mixed evidence regarding similes in these two analyses, we modeled the effect of ratings on each sentence type separately, both with and without the total comprehension time as a covariate. For literal comparisons, there was no evidence that P2TT was related to comprehensibility ratings either with the covariate included, $t(64.3)=0.56, p=.580$, or without it, $t(59.9)=0.99, p=$ .327. Conversely, consistent with overall analyses of figurative sentences, in individual analyses of each of the figurative forms P2TT was similarly, but weakly related to comprehensibility when the covariate was included $($ metaphor: $t(49.3)=1.82, p=.075$; simile: $t(37.5)=1.94, p$ $=.060)$ and highly reliably related to comprehensibility when the covariate was not included in the models (metaphor: $t(34.4)=3.13, p=.004$; simile: $t(58.0)=2.92, p$ $=.005$ ). Recall that the overall relationship of comprehensibility and P2TT for figurative items in our combined analyses was reliable even with total response time included as a covariate (i.e., $\mathrm{p}=.007$, above). A similar analysis without the covariate also provided strong evidence of an overall relationship, $t(52.4)=2.99, p=.004$.

Does P2TT help to explain overall response time differences for figurative items? To test whether P2TT, itself, can account for longer overall overt comprehension responses, a new analysis of overall response time was conducted (for trials displaying RI) with and without P2TT as a covariate. Without P2TT included, there was strong evidence of a relationship between ratings of comprehensibility and comprehension time for these trials, $t(23.7)=3.08, p=.005$, but when P2TT was included as a covariate, no evidence of the relationship between response time and comprehensibility remained, $t(58.1)=1.50, p=$ .138. For figurative items, then, it appears that P2TT likely represents time used for computing a figurative interpretation of the sentence.


Figure 4. Geometric mean second pass total gaze duration (P2TT) by item as a function of rated comprehensibility for each sentence form. Only the 1431 trials where regression into the critical word occurred are included.

## General Discussion

In Experiment 1, we observed that ratings of comprehensibility for vehicles in simile or metaphor form were highly correlated. The average figurative ratings from Experiment 1 were used in Experiment 2 to try to predict gaze variables related to the figurative vehicle during reading. We reasoned that the similar comprehensibility ratings of similes and metaphors observed in Experiment 1 reflected similar ease or difficulty with deriving the appropriate figurative meaning of the vehicle; we sought evidence of when this might unfold during reading.

Gaze patterns for metaphor vehicles, but not simile vehicles, reflected rated figurative comprehensibility from the very first fixation. Metaphor vehicles that were judged less comprehensible were subjected to longer initial periods of analysis. In contrast, similarly-rated similes, did not show immediate effects. For similes, as for literal comparisons, initial measures of processing time for their vehicles in similes were unrelated to rated comprehensibility.

But simile processing resembled metaphor processing during second-pass reading of the sentence. Both simile and metaphor vehicles showed comprehension-related durations of fixation. These second-pass durations were related to the rated comprehensibility of the word in the sentence both for metaphors and for similes. This pattern was not found for literal comparisons.

The difference between similes and metaphors at first fixation might be regarded as reflecting the weaker pragmatic assertion involved in declaring that something is like something rather than that it is something (RubioFernández, Geurts \& Cummins, 2016). To say that something is like something else implies that it is also unlike it. In this sense similes are sensibly experienced as weaker than metaphors at first pass, even if the ultimate interpretation of what is being said about the topic ultimately requires accessing a figurative or abstract interpretation of the vehicle.

Overall, these data suggest that similes are initially treated similarly to literal comparisons, consistent with the arguments of Glucksberg and Haught (2006). However, the second pass data and the ratings data both suggest that sentence comprehension for simile forms still requires identifying a figurative interpretation. We think this supports Aristotle's assertion of the figurative nature of simile that is embedded within his longer discussion of metaphor. Aristotle ( $400 \mathrm{BC} / 1991$ ) wrote "A simile is also a metaphor, for there is little difference." This quote clearly implies that metaphor (literally a "carrying-over" of meaning) is the larger category.

Our study has used similes derived from metaphors. The data show that reading such similes differs substantially from reading their corresponding metaphors. However, the data also support the idea that the interpretive demands for the figurative vehicle may ultimately be similar in both forms. This distinguishes simile from literal comparison.

## Acknowledgments

This research was supported by a faculty research grant to FHD from Swarthmore College. RG was supported by a Frances Velay Fellowship.

## References

Aristotle (1991). On rhetoric: A theory of civic discourse (G. A. Kennedy, translator). New York: Oxford University Press. (Original work published $\sim 400 \mathrm{BC}$ ).
Bowdle, B. F., \& Gentner, D. (2005). The career of metaphor. Psychological Review, 112, 193-216.
Carston, R., \& Wearing, C. (2011). Metaphor, hyperbole and simile: A pragmatic approach. Language and Cognition, 3, 283-312.
Chiappe, D. L., \& Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. Psychonomic Bulletin \& Review, 6, 668-676.
Chiappe, D. L., \& Kennedy, J. M. (2000). Are metaphors elliptical similes? Journal of Psycholinguistic Research, 29, 371-398.
Glucksberg, S., \& Haught, C. (2006). On the relation between metaphor and simile: When comparison fails. Mind \& Language, 21, 360-378.
Glucksberg, S., \& Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. Psychological Review, 97, 3-18.
Israel, M., Harding, J. R., \& Tobin, V. (2004). On simile. In M. Achard and S. Kemmer (eds) Language, Culture, and Mind, (pp123-135). CSLI Publications.
Katz, A. N., Paivio, A., Marschark, M., \& Clark, J. M. (1988). Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. Metaphor and Symbol, 3, 191-214.
Levin, S. R. (1982). Aristotle's theory of metaphor. Philosophy \& Rhetoric, 15, 24-46.
Luke, S. G. (2016). Evaluating significance in linear mixedeffects models in R. Behavior Research Methods, 1-9. DOI:10.3758/s13428-016-0809-y
Rubio-Fernández, P., Geurts, B., \& Cummins, C. (2016). Is an apple like a fruit? A study on comparison and categorisation Statements. Review of Philosophy and Psychology, 1-24. DOI: 10.1007/s13164-016-0305-4
Thibodeau, P. H., \& Durgin, F. H. (2011). Metaphor aptness and conventionality: A processing fluency account. Metaphor and Symbol, 26, 206-226.
Tirrell, L. (1991). Reductive and nonreductive simile theories of metaphor. The Journal of Philosophy, 88, 337358.

Tourangeau, R., \& Sternberg, R. J. (1982). Understanding and appreciating metaphors. Cognition, 11, 203-244.

# Representing the Richness of Linguistic Structure in Models of Episodic Memory 

Melody Dye (meldye@indiana.edu) ${ }^{1}$, Michael Ramscar (michael.ramscar@uni-tuebingen.de) ${ }^{2}$, \& Michael N. Jones (jonesmn@indiana.edu) ${ }^{1}$<br>${ }^{1}$ Department of Psychological \& Brain Sciences, Indiana University, Bloomington<br>${ }^{2}$ Department of Linguistics, University of Tübingen


#### Abstract

The principal aim of a cognitive model is to infer the process by which the human mind acts on some select set of environmental inputs such that it produces the observed set of behavioral outputs. In this endeavor, one of the central requirements is that the input to the model be represented as faithfully and accurately as possible. However, this is often easier said than done. In the study of recognition memory, for instance, words are the environmental input of choice-yet because words vary on many different dimensions, and because the problem of quantifying this variation has long been out of reach, modelers have tended to rely on idealized, randomly generated representations of their experimental stimuli. In this paper, we introduce new resources from largescale text mining that may improve upon this practice, illustrating a simple method for deriving feature information directly from word pools and lists.


Keywords: recognition memory; word frequency; word length; feature frequency; orthographic similarity; semantic similarity; corpus analysis; vector space models

## Introduction

In cognitive modeling, there is a close interdependence between representation and process. A model consists in both a data structure (an abstract representation of environmental input), and an algorithm (the process that operates over the data to simulate behavior). The choice of structure for the underlying data is critical, as it can profoundly influence the choice of algorithm. Valid representational assumptions are of vital importance, in that they reduce the degrees of freedom available to the modeler, thereby constraining model selection.

Since the inception of memory research, psychologists have relied on verbal stimuli to study learning and forgetting (Ebbinghaus, 1885). In episodic and semantic memory, the majority of data has been-and is still-generated from experiments with word lists, and memory models are routinely assessed in terms of their ability to fit data on verbal remembering (Monsell, 1991). However, when it comes to words, the choice of data structure is complicated by the fact that words vary on a remarkable number of lexical and semantic dimensions (Baayen, Milin, \& Ramscar, 2016), which may or may not contribute to how they are learned and remembered. Historically, it has been impossible to reliably quantify all these points of variation. Memory modelers have thus tended to rely on randomly generated representations, which have been carefully selected to preserve the relevant properties of the data.

While this practice has been expedient, it is no longer strictly necessary. As large-scale corpora - and the technology to mine them-have become widely available (Gilquin \& Gries, 2009; Halevy, Norvig, \& Pereira, 2009; Recchia \& Jones, 2009), it has become not only possible, but relatively straightforward, to construct not merely plausible, but accurate representations of the stimuli used in a given experiment (Baayen, 2010). This is an important advance, as problems can arise when the selected
representation does not faithfully reflect the environment. For instance, global matching models of episodic memory have considerably more difficulty reproducing behavioral data when supplied with realistic semantic representations (Johns \& Jones, 2010). Improving the quality of our data structures could thus improve the quality of our process models.

Further, while the words selected for memory experiments are commonly assumed to vary randomly, in line with their selection procedure, this may not always be the case. For one, certain properties - such as semantic similarity - may be systematically skewed, and thus poorly represented by a normal distribution (Johns \& Jones, 2010). For another, there may be accidental variation between the word pools used by different research groups (van Heuven et al., 2014), which could produce conflicting results. Given renewed interest in replicability in the psychological and brain sciences (Open Science Collaboration, 2015), providing a more detailed account of the stimulus properties that produce a given effect should be a principal research aim (Ramscar, 2016).

The overarching goal of this paper is to enumerate a simple technique for investigating the lexical and semantic characteristics of a specific word pool, and to discuss how this can be fruitfully applied to the interpretation of empirical results in episodic memory.

## Word Frequency

Word frequency is a measure of a word's occurrence in the language, and a proxy for an individual subject's experience with that word. Frequency has long been a variable of central importance in cognitive models, as it is one of the strongest predictors of verbal processing and remembering (Baayen, Milin, \& Ramscar, 2016; Balota et al., 2007). In some models, frequency is treated as a causal variable-e.g., in a model of visual word recognition, frequency might function as an internal counter, in which each occurrence of an item increments its baseline activation upward (Coltheart et al., 2001). In others models, frequency is treated as an informative correlational variable, and items of a given frequency class are assigned specific feature values (Shiffrin \& Steyvers, 1997).

Setting the details aside, virtually all models incorporate frequency in one respect or another. Given the significance of frequency as an explanatory variable, its accuracy of measurement, relation to other lexical and semantic variables, and instantiation in cognitive models are all matters of some theoretical importance. Yet in spite of this, many researchers are still working with outdated measurements and methods, which are not being updated as the field advances. One particularly remarkable example of this is that the KučeraFrancis norms (1967), collected fifty years ago, are still widely used among psychologists to determine word frequency. This is the case even though they have been known for decades to be unreliable (particularly for lower frequency words), and are, on assessment, consistently the worst performing norms across an array of lexical processing tasks (Brysbaert \& New, 2009). Frequency values collected today are derived from corpora orders of magnitude larger.

Another source of concern is that word frequency itself is
routinely treated as a categorical variable, rather than a continuous one, even though dichotomizing a random variable can seriously jeopardize reliability (MacCullum et al., 2002; Hemmer \& Criss, 2013). Further contributing to this problem, the standard method for binning words into high and low frequency bands fails to take into account the skewed nature of the distribution. Indeed, in an analysis of several classic studies, high frequency items were found to have considerably larger standard deviations than their low frequency counterparts, and a sizeable percentage of 'low' frequency items were shown to fall at, or above, what should have been the border between the groups (van Heuven et al. 2014).

That influential psychometric tests have been predicated on such unreliable measures raises serious questions about their validity (Ramscar et al. 2014). Nevertheless, this binary division remains common in both experimental design and in modeling.

## Word Frequency Effects in Recognition

One domain in which it is still commonplace to bin experimental items into high (HF) and low frequency (LF) bands is recognition memory. Models of recognition offer an illustrative test case for why representational assumptions are important to cognitive modeling, and how they might be refined with simple data mining techniques. To clarify this example, we first briefly review recognition memory as an experimental paradigm and as a modeling domain.

In tests of single item recognition, subjects study a list of words, and then at test, are asked to discriminate words encountered at study (targets) from non-studied words (foils). The difficulty of the task lies in the fact that subjects must differentiate between words seen at study and words encountered in everyday life-i.e., they must distinguish between general familiarity with the test items and familiarity that is specific to the recognition task.

Global matching models have predominated as explanatory models of recognition performance (Hintzman, 1988; Murdock, 1982; Shiffrin \& Steyvers, 1997). These models are premised on the idea that item recognition depends not only on the characteristics of the item itself, but also on other items present concurrently in memory. When a specific item is presented at test, the available item and context cues form a joint probe of memory. This search process yields a match value between the test item and the contents of memory. If this value exceeds some threshold, the item is recognized as 'old'; if it fails to meet this criterion, the item is rejected as 'new'. A grounding assumption of global matching models is that studied items will have higher match values, on average, than unstudied lures. However, item recognition is rarely perfect, and much effort has been expended in identifying how interference can arise at retrieval. Noise sources are frequently categorized into two types: item noise (McClelland \& Chappell, 1996; Shiffrin \& Steyvers, 1997) and context noise (Dennis \& Humphreys, 2010). Item noise arises from spurious feature matches with other studied items; context noises arises from interference from extra-experimental contexts in which the tested item has occurred.

Among the findings that global matching models are designed to capture, one of the hallmarks is the mirror effect for word frequency: This is the finding that when HF and LF words are present in equal numbers at study, LF items are better recognized at test, garnering both more hits and fewer false alarms (Glanzer \& Adam, 1985). One way to capture this frequency effect is to assign different parameter values to HF and LF words, thereby generating different distributions of feature values, and hence, of featural similarity between items.

Such a representational choice reflects the fact that words are comprised of an array of surface and semantic properties that are known to vary with frequency, and to affect processing and remembering (Landauer \& Streeter, 1973; Schulman, 1967).

For example, in the Retrieving Effectively from Memory (REM) model, the parameter settings generate HF items with more common, overlapping features than LF items (Steyvers \& Shiffrin, 1997). Because these features are less diagnostic, the self-match between HF targets and their own memory traces is weaker than for LF targets; because they are more common, the likelihood of a chance feature match between HF targets and HF foils will be greater. This yields the canonical lower hit-rate and higher false-alarm rate for HF items.

## Representational Assumptions

Global matching models have shown considerable success in capturing the relevant empirical data, ranging from word frequency effects to differential forgetting (Clark \& Gronlund, 1996). Despite these undisputed successes, there are potential drawbacks in how they represent their list items. For one, these representations commonly lump together semantic, phonemic, and orthographic features into a single, indistinguishable feature set, making it impossible to tease apart how each dimension contributes to recognition performance. For another, representations are randomly generated, rather than empirically derived.

In the influential REM model, for example, a single parameter controls the mean and variability of the distribution that item features are sampled from (Steyvers \& Shiffrin, 1997). To capture qualitative differences in item similarity between word frequency bands, the parameter is adjusted separately for high and low frequency items. However, the specific parameter settings are unconstrained by the actual properties of the stimulus set. Instead, parameters are set either by convention or by best fit to the behavioral data.

Concerns have been raised with this type of practice. In particular, such flexibility leaves the resulting models open to the criticism that they could be made to fit a wide variety of results (Roberts \& Pashler, 2000). Conversely, they might require significant theoretical adjustments to account for the results when supplied with a realistic representation of the list items (see Johns \& Jones, 2010 for an illustration). Finally, if different experiments produce contradictory results, there is no straightforward way to trace back these differences to the characteristics of the lists.

The theoretical claims of this class of models could be strengthened by deriving the model parameters directly from the lexical and semantic characteristics of the experimental word pool, or test list. This could be accomplished in a number of ways. In the simplest case, the actual feature distribution of the stimuli could be used to determine the closest choice of parameter settings. Another option would be to generate the input representation directly from the stimuli, using either the real feature values, or adjusted feature values (which could be made more robust by incorporating noise, or various smoothing mechanisms; see e.g., Chen \& Goodman, 1999). Here, we detail a simple procedure for deriving feature information for lexical items as a function of their frequency class.

## Corpus Investigation

The following investigation was conducted 1) to illustrate how various lexical and semantic feature information can be derived directly from word pools and recognition lists, 2) to examine how these feature values can be expected to vary as a function of item frequency, and 3) to assess whether standard word pools mimic these differences (and each other).

## Verbal Properties and Frequency Class

In the study of semantic and episodic memory, different word pools make use of somewhat different sampling procedures and controls. Thus, our first goal was to establish a neutral, independent baseline, in which words were sampled without any special consideration other than frequency.

| Zipf value | fpmw | Examples |
| :---: | :---: | :---: |
| 1 | 0.01 | antifungal, bioengineering, farsighted, harelip, proofread |
| 2 | 0.1 | airstream, doorkeeper, neckwear, outsized, sunshade |
| 3 | 1 | beanstalk, cornerstone, dumpling, insatiable, perpetrator |
| 4 | 10 | dirt, fantasy, muffin, offensive, transition, widespread |
| 5 | 100 | basically, bedroom, drive, issues, period, spot, worse |
| 6 | 1000 | day, great, other, should, something, work, years |
| 7 | 10,000 | and, for, have, I, on, the, this, that, you |

Figure 1: The Zipf scale is a logarithmic scale that divides the frequency spectrum into seven discrete classes (van Heuven et al. 2014).

Word Frequency Words and their frequencies were extracted from the state-of-the-art 51 million word SUBTLEXus corpus (Brysbaert \& New, 2009). Frequency classes were assigned according to the Zipf scale, which is calculated for an individual item as $\log _{10}$ (frequency per billion words). The Zipf scale has a number of advantages over the typical binary division between HF and LF words, namely that it is a logarithmic scale reflecting the psychological interpretation of frequency, and its divisions are fine-grained, creating seven distinct classes rather than the traditional two (van Heuven et al., 2014). For purposes of comparison, a Zipf value of 3 or lower corresponds to LF words; 4 or higher to HF words (Figure 1).

Recognition Lists To create recognition lists, 10 items were selected at random (without replacement) from a given frequency bin. Half of these items were labeled targets, and the other half foils, replicating the standard list construction procedure. This sampling procedure was repeated until there were 1000 such lists for each frequency class.

The aim was to compare lists created in each band on four dimensions: word length, feature frequency, and orthographic and semantic similarity of targets to foils. These particular dimensions were chosen to be illustrative, and because they are known to be important contributing factors to item recognition. For word length and feature frequency, counts were computed for each item, and averaged over the entire list. For orthographic and semantic similarity, the similarity of each target to the distractors present at test was computed, and similarly averaged.

To preface, these analyses successfully replicate wellestablished findings on each of these dimensions, while providing a straightforward method for determining the actual empirical trends of a given frequency range, or item set.
Methodology Notably, the comparatively small number of types in the higher frequency ranges placed constraints on the construction of recognition lists (Figure 2). Specifically, list length was necessarily kept small, and while lists were created for Zipf values 1-6, 7 was excluded, as it comprised only 13 distinct word types, all of them function words.

This type distribution is a consequence of the universal scaling law for word frequencies, commonly known as Zipf's Law (1949). The idea is this: Say, an English text is selected, and each of the word types that occur in the text are arranged in order of their frequency, from most to least common, and assigned a numerical rank. Then, the full contents of the text that is, all of its word tokens - are thrown into a bag, shook,
and one word is selected at random. Zipf's Law states that the probability of drawing a given word is inversely proportional to that word's rank ordering. The law formalizes the notion that while a few words in a language are very common, the greater part are exceedingly rare.


Figure 2: The number of distinct word types in the SUBTLEXus corpus for each value of the Zipf scale.

Word Length Word length, whether computed in terms of letters or phonemes, has an inverse relationship with frequency, with word lengths tending to increase as frequency declines (Piantadosi et al., 2011; Sigurd, Eeg-Olofsson, \& Van Weijer, 2004; Wright, 1979; see Figure 3).


Figure 3: Average word length of list items increases as frequency declines.

Feature Frequency Feature frequencies represent the empirical n-gram frequencies of individual letters and letter combinations, and can be conceptualized as a measure of orthographic distinctiveness (Figure 4).

Feature frequency is known to vary with word frequency. On average, rarer words contain both more unusual letters, and more unusual combinations of letters (Malmberg et al. 2002; Zechmeister, 1969).

Orthographic similarity Orthographic similarity was computed as Levenshtein edit distance, a string metric that calculates the minimum number of edits (such as insertions, deletions, or substitutions) required to transform one word into the other (Figure 5).

Given that rare words are more orthographically distinctive (Landauer \& Streeter, 1973; Andrews, 1992), it stands to reason that in a recognition list context, they should be less orthographically similar to frequency-matched distractors than more common words (Hall, 1979).


Figure 4: The five panels depict the average feature frequencies of list items in SUBTLEXus as a function of their Zipf value. The overall trend indicates that higher frequency items are comprised of higher frequency features. Moreover, the larger the n-gram, the greater the separation between frequency classes. For unigrams, a more pronounced pattern of separation between Zipf bands is observable when minimum (rather than average) feature frequency is used.


Figure 5: Average orthographic similarity between targets and distractors declines as a function of frequency.

Semantic similarity Semantic similarity values were obtained from word2vec trained on the 300 billion word Google News corpus. word2vec is a two-layer neural network that produces word embeddings (Mikolov et al., 2013), and is considered state of the art in semantic space modeling (Baroni, Dinu, \& Kruszewski, 2014). word2vec was implemented with gensim, a Python framework for vector space modeling (Řehůřek \& Sojka, 2010), which adopts the continuous skip-gram architecture. The skip-gram model weights proximate context words more highly than distant ones, yielding better results for lower frequency words.

In a recognition task in which list items are randomly sampled from a given frequency band, the semantic similarity between targets and distractors should tend to decrease with frequency (Figure 6). This outcome is all but assured by the distributional properties of the lexicon: In the SUBTLEXus corpus, LF words comprise $80 \%$ of word tokens (van Heuven et al., 2014) and fully $94 \%$ of word types (Figure 2). The
semantic spread from which LF words are sampled will thus be far greater than that for HF items.


Figure 6: Average semantic similarity between targets and distractors declines across the HF range of the Zipf scale, implying that a set of randomly sampled words will be less semantically similar, on average, the lower their frequency class.


Figure 7: Average semantic similarity between targets and distractors across the LF range of the Zipf scale. While a slight ( $n s$ ) trend in the opposite direction is observable in the lower range of the scale, this is almost certainly a methodological artifact. If the missing data in Figure 8 is included as 0 -counts, the apparent trend reverses, and the pattern resembles that seen in Figure 6.

In making these calculations, there is an important methodological issue to consider-in particular, the problem, well-known to linguists, of data sparsity (Sinclair, 1997): While any given sample of language will provide ample evidence about its common words and phrases, it will provide little or none about its rarer, more informative elements (Church \& Gale, 1995). Not only will many perfectly legitimate words (and word co-occurrences) fail to occur in even very large swaths of text, but even most of those that do will occur only a few times, making their estimation unreliable. This is the basic problem of data sparsity and it is one that plagues semantic similarity analyses in the lower frequency ranges (Figures 7, 8).


Figure 8: Data loss for the semantic similarity analyses as a function of frequency class. Semantic similarity values were not available for all the words sampled, and the proportion of words with no data points grew as frequency decreased. For Zipf rank 1, fully $25 \%$ of data was lost.

Figures 6 and 7 show the similarity distributions for item pairs that were known to our word2vec model. However, given the significant data loss for LF items, looking solely at returned values constitutes selection bias, as it implies that unobserved pairs-for which the model cannot supply a score-likely have the same distributional properties as observed pairs. In fact, it is reasonable to assume that unobserved pairs are much less similar, on average. One way of addressing this issue is to assign item pairs with null values a similarity score of 0 . When these scores are included, the trend observable in the HF range (Figure 6) is also clearly observable in the LF range.

In the absence of knowledge, assigning 0 -counts is a useful heuristic. However, given that problems with data sparsity increase as frequency declines, this solution may disproportionately penalize the lowest frequency words. In future work, similarity-based smoothing techniques might be used to better estimate similarity values for unobserved pairs (c.f. Yarlett, 2007).

Interim Summary Our analyses of words in the SUBTLEXus corpus replicates and extends a number of well-known findings on the relationship between a word's frequency and its lexical and semantic features, including that:

1) word length increases as word frequency declines,
2) feature frequency increases with word frequency, with the rate of increase dependent on feature length,
3) orthographic similarity between targets and foils increases with word frequency,
4) semantic similarity between targets and foils increases with word frequency (though the calculation of similarity scores for LF item pairs requires careful consideration).

Available Analyses In the analyses reported here, pure lists were created for each frequency class, average feature information was extracted, and similarity measures were computed as a function of the mean similarity of a target to its foils. The purpose of this was largely illustrative; many variations on this procedure are possible, depending on the requirements of the model, or the empirical task.

One obvious choice point is the sampling method. For example, word selection could be constrained by specific lexical properties (e.g., limited to nouns, or words of length $n$ ), as is common practice in the design of word pools. Similarly, list composition could be varied by sampling specific proportions of words from different frequency bands.

Another matter of some importance concerns the choice of comparisons and statistical measures. Similarity can be computed relative to other targets, distractors, or both; it can also be calculated as an average, or in terms of "max" similarity (e.g., the top $10 \%$ of most confusable items). Likewise, when assessing the use of rare letters and rare letter combinations, it may be more useful to know the minimum feature frequency, or the median, rather than the mean.

Finally, while we chose to delimit our focus to just a few dimensions, there are many more lexical properties that systematically vary with frequency. For instance, rare words are more likely to be judged as abstract (Galbraith \& Underwood, 1973; Pavio, Yuille, \& Madigan, 1968), to be acquired later (Carroll \& White, 1973), and to be regular (Bybee \& Hopper, 2001).

## Word Pools

In the study of semantic and episodic memory, different word pools make use of somewhat different sampling procedures and controls. One concern is that different word
lists may vary in systematic ways from each other, producing variability in results; another is that they may have distinctly different properties from the language 'at large'. To check the validity of these worries, we compared the word pools of two representative cognitive memory labs, with an average h -index among the principle investigators of 20 , and published theoretical disagreements. These word pools were compared against a recognition word list devised by Dye, Jones, \& Shiffrin (2017) (Figures 9, 10).

The Dye et al. (2017) word list was deliberately constructed to increase the semantic and orthographic similarity of LF items, as reflected in Figures 9 and 10. In a recognition list experiment, this had the predicted effect of diminishing the standard mirror effect for word frequency, by bringing the false alarm rate for low and high frequency items into line.


Figure 9: A comparison of average semantic similarity of targets to foils across three word pools.

Notably, while the Dye et al. word list clearly differs from the two standard word pools, these word pools are not identical to each other either. In particular, though both pools are similarly distributed in terms of frequency and semantic similarity among items, in Word Pool 2, orthographic similarity among items is substantially increased compared to Word Pool 1 , and is matched across HF and LF items. This may produce differences in reported results, as orthographic similarity is known to modulate false alarm rates (Malmberg, Holden, \& Shiffrin, 2004).

Finally, it is worth noting that none of these 'controlled' word pools reflect the properties expected from random sampling, as illustrated in our exploration of the SUBTLEXus corpus. In particular, while the distribution of orthographic and semantic similarity values for LF and HF items are largely overlapping for the standard word pools (Figures 9, 10), a truly random selection of these items shows significant separation between frequency bands (Figures 5, 6).

These examples illustrate how the properties of word lists can be readily and fruitfully compared both to each other, and to larger corpora. In future work, we plan to expand this analysis to include more widely used word pools, such as the Toronto word pool (Friendly, Franklin, Hoffman, \& Rubin, 1982), a modified version of the Kucera \& Francis word pool (1967), and a categorized word pool (Murdock, 1976).


## References

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? Journal of Experimental Psychology: Learning, Memory, and Cognition, orthographic re
18(2), 234-254.
Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. The Mental Lexicon, 5(3), 436-461.
Baayen, R. H., Milin, P., \& Ramscar, M. (2016). Frequency in lexical processing. Aphasiology, 1-47.
Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. Behavior Research Methods, 39(3), 445-459.
Baroni, M., Dinu, G., \& Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 238-247.
Brysbaert, M., \& New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977-990.
Carroll, J. B., \& White, M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. Journal of Verbal Learning and Verbal Behavior, 12(5), 563-576.
Chen, S. F., \& Goodman, J. T. (1999). An Empirical Study of Smoothing Techniques for Language Modeling. Computer Speech and Language, 13, 359-394.
Church, K.W., \& Gale, W.A. (1995). Inverse Document Frequency (IDF): A Measure of Deviation from Poisson (pp. 121-130). In Proceedings of the Third Workshop on Very Large Corpora.
Clark, S. E., \& Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. Psychonomic Bulletin \& Review, 3(1), 37-60.
Coltheart, M., Rastle, K., Perry, C., Langdon, R., \& Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. Psychological Review, 108, 204-256.
Dennis, S., \& Humphreys, M. S. (2001). A context noise model of episodic word recognition. Psychological Review, 108(2), 452-478
Dye, M., Jones, M., \& Shiffrin, R. (2017). Vanishing the mirror effect: The influence of prior history \& list composition on recognition memory. Proceedings of the 39th Annual Conference of the Cognitive Science Society.
Ebbinghaus, H. (1885). Memory: A contribution to experimental psychology. New York: Dover
Friendly, M., Franklin, P.E., Hoffman, D., \& Rubin, D.C. (1982). The Toronto Word Pool: Norms for magery, concreteness, orthographic variables, and grammatical usage for 1,080 words. Behavior Research Methods and Instrumentation, 14(4), 375-399
Galbraith, R. C., \& Underwood, B. J. (1973). Perceived frequency of concrete and abstract words. Memory \& Cognition, 1(1), 56-60.
Gilquin, G. \& Gries, S.T. (2009). Corpora and experimental methods: A state-of-the-art review. Corpus Linguistics and Linguistics Theory, 5(1), 1-26.
Halevy, A., Norvig, P., \& Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2), 8-12.
Hall, J. F. (1979). Recognition as a function of word frequency. The American Journal of Psychology, 92(3), 497-505.
Hemmer, P., \& Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a oontinuous variable in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(6), 1947-1952.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. Psychological Review, 95, 528-551.
Johns, B. T., \& Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. Psychonomic Bulletin \& Review, 17(5), 662-672
Kucera, H., \& Francis, W. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.
Landauer, T. K., \& Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. Journal of Verbal Learning and Verb
Malmberg, K. J., \& Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(4), 616-630.
Malmberg, K. J., Steyvers, M., Stephens, J. D., \& Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. Memory \& Cognition, 30(4), 607-613
McClelland, J. L., \& Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory, Psychological Review, 105, 724-760.
Mikolov, T., Chen, K., Corrado, G. \& Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint: arXiv: 1301.3781
Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner \& G.W. Humphreys (Eds.), Basic processes in reading: Visual word recognition. Hillsdale, NJ: Erlbaum.
Murdock, B.B. (1976). Item and order information in short-term serial memory. Journal of Experimental Psychology, 105(2), 191-216.
Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. Psychological Review, 89, 609-626.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716-aac4716.
Paivio, A., Yuille, J. C., \& Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology, 76(1), 1-25.
Piantadosi, S. T., Tily, H., Gibson, E., \& Kay, P. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences of the United States of America, 108(9), 3526-3529
Ramscar, M. (2016). Learning and the replicability of priming effects. Current Opinion in Psychology, 12, 80-84
Recchia, G., \& Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. Behavior Research Methods, 41(3), 647-656.
Roberts, S., \& Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107(2), 358-367
Shiffrin, R. M., \& Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. Psychonomic Bulletin \& Review, 4, 145-166.
Schulman, A. I. (1967). Word length and rarity in recognition memory. Psychonomic Science, 9, 211-212.
Sigurd, B. Eeg-Olofsson, M. \& Van Weijer, J. (2004). Word length, sentence length and frequency Zipf revisited. Studia Linguistica, 58(1), 37-52.
Sinclair, J.M. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T McEnery, \& G. Knowles (Eds.), Teaching and Language Corpora. Longman.
van Heuven, W. J. B., Mandera, P., Keuleers, E., \& Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency dat
Psychology, 67(6), 1176-1190.
Wright, C. E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. Memory and Cognition, 7(6), 411-419.
word2vec: https://code.google.com/archive/p/word2vec/
Yarlett, D. (2007). Language learning through similarity-based generalization. Unpublished doctoral dissertation: Stanford University.
Zechmeister, Z.B. (1969). Orthographic distinctiveness. Journal of Verbal Learning and Verbal Behavior, 8(6), 754-761.
Zipf, G.K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley: Cambridge, MA.

# Due process in dual process: A model-recovery analysis of Smith et al. (2014) 

Charlotte E. R. Edmunds (ceredmunds@gmail.com)

School of Psychology, Plymouth University, Plymouth, PL4 8AA, UK

Fraser Milton (F.N.Milton@exeter.ac.uk)<br>School of Psychology, University of Exeter, Exeter, EX4 4QG, UK

Andy J. Wills (andy@willslab.co.uk)
School of Psychology, Plymouth University, Plymouth, PL4 8AA, UK


#### Abstract

Considerable behavioral evidence has been cited in support of the COVIS dual-system model of category learning (Ashby \& Valentin, 2016). The validity of the inferences drawn from these data critically depend on the accurate identification of participants' categorization strategies. In the COVIS literature, participants' strategies are identified using a model-based analysis inspired by General Recognition Theory (Maddox, 1999). Here, we examine the accuracy of this analysis in a modelrecovery simulation. We find that participants can appear to be using implicit, procedural strategies when their responses were actually generated by explicit rule-based strategies. The implications of this for the COVIS literature are discussed.


Keywords: categorization; COVIS; dual-systems accounts; model-recovery; GRT

## Introduction

Categorization studies rarely examine individual differences. Rather, researchers look at group performance to draw conclusions about the likely underlying mechanisms of category learning (Kurtz, 2015). For these inferences to be valid, the participants in each group must all learn in a qualitatively similar way (Maddox, 1999). Then, relatively little information is lost by averaging. However, severe interpretative difficulties can arise if participants learn in a variety of ways, as then the average will likely not represent the behaviour of any single person (Siegler, 1987).

This issue is more than hypothetical, as there is substantial evidence, and a degree of consensus, that different participants use qualitatively different strategies in categorization tasks (e.g., Nosofsky \& Zaki, 2002; Raijmakers, Dolan, \& Molenaar, 2001; Wills, Inkster, \& Milton, 2015). For example, some participants categorise stimuli on the basis of just one stimulus dimension (as in Figure 1B), or do so initially, even if optimum performance on the task requires using multiple stimulus dimensions (as in Figure 1B, where the participant's strategy is single-dimensional but the optimal classification strategy is diagonal).

COVIS (COmpetition between Verbal and Implicit Systems; Ashby \& Valentin, 2016) is one model that aims to predict when and why participants use different strategies.

COVIS assumes that categorization is mediated by two, parallel, competing systems: an Explicit System and a Procedural System. The Explicit System is assumed to implement rule-based strategies (such as in Figure 1B). Therefore, COVIS predicts this system will optimally learn category structures that are implementations of simple rules (such as in Figure 1A). If rule-based strategies result in poor accuracybecause the category structure is not rule-based and thus difficult to verbalise-COVIS predicts the Procedural System will gain control of responding. As the Procedural System is predicted to implement a variety of strategies (including the one demonstrated in Figure 1A), it is capable of implementing the optimum strategy for information-integration category structures (the structure, but not the strategy, shown in Figure 1B).

Typical COVIS-supporting experiments look for a differential effect of an experimental manipulation (e.g. feedback timing) on rule-based and information-integration category structure learning (for a review, see Ashby \& Valentin, 2016). The category structure manipulation is hoped to elicit a switch in the learning system controlling responding: because participants are learning a rule-based or informationintegration category structure they will use the appropriate strategy and so be using the Explicit or Procedural System, respectively. If the experimental manipulation affects one category structure condition more than the other, the experimenter infers that it affects the accuracy of one system more than the other, thereby providing evidence for a dual-system model of category learning.

For these experiments, the presence of subsets of qualitatively-different participants can be particularly problematic. Critically, the conclusion that the experiment supports a dual-system model depends on the assumption that the participants in each category structure condition used the most appropriate system to learn those structures. In other words, that participants used the Explicit System to learn the rule-based category structure, and the Procedural System to learn the information-integration category structure. If this is not the case, any differences in overall accuracy between category structure conditions might be due to varying rates of


Figure 1: Two example strategies implemented on two category structures: A) a unidimensional category structure with a diagonal (GLC) strategy applied, B) an information-integration category structure with a unidimensional rule applied. Category structure: dots represent Category A, and squares Category B. Responses: Filled symbols indicate a participant responded "A", open symbols indicate they responded " B ".
sub-optimum strategies between conditions, rather than to the existence of two category learning systems.

To illustrate this point, consider an experiment that examined the effect of deferring feedback and found it caused a reduction in performance by $10 \%$ for a particular category structure (such as in Smith et al., 2014). In the ideal case, all participants classifying a particular category structure would be using the same optimum strategy and all those in the relevant condition would be similarly affected by the manipulation; participants with deferred feedback would score $10 \%$ less than those with immediate feedback. Here, we could use standard group-accuracy analyses validly. However, if some participants were using other, sub-optimum strategies then drawing conclusions from the experiment is harder. One possibility is that the manipulation, within a given category structure, changes the relative proportions of different strategies used in each condition (feedback type). This would change average accuracy because, given a particular category structure, the highest accuracy for each strategy varies. A second possibility is the manipulation has a differential effect depending on the strategy type being used. For example, the manipulation could have had no effect on people using the optimum strategy, but could severely affect performance reliant on sub-optimum strategies (see Schnyer et al., 2009, for a similar argument).

To avoid the possibility that any dissociation in accuracy is due to the effects of sub-optimum strategies rather than
two competing systems, COVIS-supporting experiments use a strategy analysis informed by General Recognition Theory (GRT; Ashby \& Gott, 1988). This analysis is used as a manipulation check to determine which strategy each participant is using. This approach (hereafter, GRT analysis) assumes that strategies can be modeled by a (usually linear) decision bound that passes through stimulus space (such as those in Figure 1). For each participant, a variety of strategy models are fitted to their responses. The one that best represents that participant's pattern of responding is selected. Then, each participant's strategy is compared to the category structure they were assigned to learn. If enough participants are found to be using the optimum strategy for the category structure they were assigned, then the category structure manipulation is assumed to have elicited a corresponding shift in category learning system. Under this assumption, any dissociations in accuracy can be validly ascribed to the existence of two systems.

Using GRT analysis as a manipulation check is logically valid if and only if GRT analysis consistently and accurately identifies participants' strategies. In other words, GRT analysis must be able to correctly identify strategies under a variety of circumstances such as differing category structures, experimental manipulations and levels of noise. Unfortunately, recent evidence from our lab suggests that GRT analysis does not accurately recover the strategies participants use for information-integration category structures. For in-
stance, Edmunds, Milton, and Wills (2015) extended an experiment by Ashby, Maddox, and Bohil (2002) looking at feedback type by asking the participants to verbally describe their strategies. A substantial number of responders classified as using the Procedural System on the basis of GRT analysis nevertheless reported using an explicit rule-based strategy (which have been predictive of behaviour in other procedures; Wills, Milton, Longmore, Hester, \& Robinson, 2013).

One possible explanation for this contradiction is that participants did not accurately report their strategies. Two pieces of evidence speak against this interpretation. First, verbal reports successfully predict participants' performance in other tasks (Lagnado, Newell, Kahan, \& Shanks, 2006). Second, Carpenter, Wills, Benattayallah, and Milton (2016) found more frontal and medial temporal lobe activation for participants learning an information-integration category structure than for participants learning a rule-based structure. These brain regions are typically associated with explicit processing (Nomura et al., 2007), implying that classification of an information-integration category structure is at least as explicit in their study as classification of a rule-based structure.

A more interesting explanation for the disparity in strategies found by Edmunds et al. (2015) is that the GRT analysis is wrong. For example, because GRT analysis normally uses just the training stimuli rather than a broad range of transfer stimuli, perhaps it is biased towards the optimal strategy for each category structure? Work by Donkin et al. (2015) provides some support for this conjecture. Specifically, Donkin et al. found that including transfer stimuli from across the stimulus space reduced the proportion of participants classified as using the optimal (diagonal) strategy for an information-integration category structure.

The possibility that GRT analysis does not accurately recover the strategies participants use makes determining whether category learning is mediated by two learning systems more difficult. Consider an experiment that found that feedback delay harmed information-integration category learning but had no effect on unidimensional rule-based category learning. Furthermore, suppose that GRT analysis found that all the participants used the optimum strategy for the category structure they were presented with. If GRT analysis were accurate, we might conclude that the source of this interaction was the presence of two different systems. However, if GRT analysis was inaccurate this inference would not be the only one we could make. For example, if GRT analysis, in the information-integration conditions, falsely identified an explicit conjunction rule strategy as a diagonal (procedural) strategy, an alternative account might be that feedback delay impacts learning once participants are using sufficiently complex rules. This would be consistent with a single-system account and would potentially cast doubt on all of the COVISsupporting studies that used this method.

However, a limitation of all work to date is that one can never be sure whether GRT analysis contains significant flaws, because one does not know which strategy participants
were actually using. When employing data from real participants, all we have are multiple forms of assessment of their strategy (GRT analysis, verbal reports, brain activations etc.), all of which provide indirect and potentially flawed information. Using one measure to assess the quality of the others includes the circularity of assuming one of the measures is correct. In the current article, we use a model-recovery approach to break out of this loop.

Model recovery involves simulating hypothetical participants' responses according to the strategy models defined by the strategy analysis. By simulating responses we circumvent many of the problems with Donkin et al. (2015) and Edmunds et al. (2015), as now we know exactly which model each (simulated) participant is using. From these hypothetical, simulated participants we can then use GRT analysis to identify the strategies from the responses to see whether GRT analysis is capable of recovering the correct generating model. This model-recovery procedure is recommended as best practice for any cognitive modeling analyses (Heathcote, Brown, \& Wagenmakers, 2014) but has yet to be done for GRT analysis.

## Simulation of Smith et al. (2014)

Below, we use model-recovery techniques to demonstrate that current GRT analyses misidentify participants' strategies in the context of levels of performance accuracy reported in published work. Further, we demonstrate that it is possible for all participants to be using rule-based strategies but to still find a) an interaction between an experimental manipulation and category structure, and b) that the majority of participants are (incorrectly) identified by GRT analysis as using the optimum strategy for each category structure.

The experiment we chose for this demonstration is by Smith et al. (2014); a recent, representative example of empirical work within the COVIS framework (Ashby \& Valentin, 2016). This experiment investigated the effect of deferring feedback on category learning. Participants were randomly assigned to learn either a rule-based or informationintegration category structure (as in Figure 1) with one of two possible reinforcement schedules. In the immediate feedback condition, on each trial participants were shown a stimulus, then made their response and were immediately given corrective feedback for that trial. In the deferred feedback condition, the stimuli were shown in groups of six. The participants made responses for all six stimuli but only received corrective feedback at the end of the block. Smith et al. found that learning of the rule-based category structure was unaffected by this change in feedback timing, whereas learning of the information-integration category structure was "eliminated" (p. 454) with deferred feedback.

As well as being representative of the majority of COVIS experiments (Ashby \& Valentin, 2016), the work reported in Smith et al. (2014) is interesting to simulate as it is representative of the direction that the role of GRT analysis is beginning to take in newer COVIS experiments (see also, Smith
et al., 2015). In these newer studies, the authors move away from using the GRT analysis to ensure that participants were using the optimum strategy, and therefore category learning system, in each condition. Instead, they use the GRT analysis to determine the strategies that participants use in order to discern whether deferring feedback alters the strategies participants use in "a theoretically meaningful way" (p. 452). Smith et al. (2014) conclude that deferred feedback pushed participants in the information-integration condition away from classification via the Procedural system towards classification via the Explicit system. These conclusions would of course be substantially undermined if their GRT analysis failed to correctly identify the strategies participants used.

The possibility of a misidentification of participant strategies would also open the way for an alternative, singlesystem, account of their results. As previously discussed, verbal report data from Edmunds et al. (2015), and neuroscience evidence from Carpenter et al. (2016), indicate that participants sometimes learn information-integration category structures using complex, verbalisable rules-despite the GRT analysis pointing towards procedural (GLC) strategies in these cases. Perhaps this is also happening in Smith et al. (2014)? Specifically, we hypothesize that the majority of participants in the immediate information-integration category structure condition of Smith et al. are using a conjunction or another two-dimensional rule-based strategy, but this is mis-identified as an implicit (GLC) strategy by Smith et al.'s GRT analysis. The possibility of this kind of misidentification seems particularly acute in this study because those authors did not include a conjunction rule (or any other complex rule) in the set of models for their GRT analysis. Research by Donkin et al. (2015) suggests that failing to include complex rules in a GRT analysis increases the proportion of participants that are identified as procedural (GLC) responders.

## Method

To see whether it was possible that all the participants in Smith et al. (2014) were using rule-based strategies, we first generated a set of hypothetical participants. These participants' responses were generated from unidimensional and conjunction strategy GRT generating models that best fit either the unidimensional or information-integration category structures used by Smith et al. The unidimensional models where straight lines that passed perpendicularly through either the $x$-axis or the $y$-axis. Stimuli that lay on one side of the line were assigned "Category A" and those on the other "Category B." The conjunction models consisted of two lines perpendicular to each other that partitioned off a quarter of the space. The stimuli in that quarter were assigned "Category A" and those outside "Category B."

We then added various levels of noise to these hypothetical participants and calculated their accuracy. Twenty participants were generated for each level of noise, category structure and generating strategy. Then we performed the GRT analysis, which included three model types: unidi-


Figure 2: Simulation of Smith et al. (2014); bars are empirical data; plot points are the simulation. Smith did not report standard deviation.
mensional, diagonal (GLC) and random models (Maddox \& Ashby, 1993). Note that although some simulated participants' responses were generated by a conjunction strategy, this strategy type was not included in the GRT analysis. This was to keep the GRT analysis as similar as possible to the one conducted by Smith et al. (2014). We then selected 21 simulated participants (i.e. the same $N$ as Smith et al., 2014) for each condition such that, as far as was possible, they had a) the same average accuracy as that reported by Smith et al. (p. 451, their paper; Figure 2, current paper), b) the same number of "strong learners" (p. 541, their paper), and c) were identified by GRT analysis as using the same distribution of strategy types reported by Smith et al. (p. 452-453, their paper; Table 1, this paper).

## Results

In addition to the simulated participants having the same average accuracy (see Figure 2) and same distribution of GRT-recovered strategies (see Table 1) as the real participants in Smith et al. (2014), it was also possible to replicate Smith et al.'s statistical tests. For the simulated participants, the critical interaction between category structure and task was significant, $F(1,80)=10.64, p=.002$. Furthermore, as in Smith et al. (2014), performance in the two rulebased conditions were statistically indistinguishable, $t(40)=$ $0.44, p=.663$, as was the comparison between the unidimensional and information-integration immediate conditions, $t(40)=1.22, p=.228$. Whereas, the difference between the two information-integration category structure conditions did reach significance, $t(40)=4.98, p<.001$

Table 1 shows that it is possible to generate the statisti-
cal pattern and strategy model results reported in Smith et al. (2014), without resorting to a second Procedural System. Instead, all the so-called implicit responders found by Smith et al. (2014) could have been using rule-based strategies that were misidentified by the GRT analysis.

Table 1: GRT analysis of simulated participants for Smith et al. (2014). Counts in bold are from real participants, as reported by Smith et al. (2014), and are also the simulation results (the simulation exactly reproduces the observed distribution of recovered models). Remaining counts show how the two groups of generating models used in this simulation (UD and CJ) were recovered by the GRT analysis. So, for example, of the 18 UD generating models used in the UD-Imm condition, 13 were correctly recovered as UD.

|  | Recovered strategies |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | UDx | UDy | GLC | RND |
| UD-Imm. | $\mathbf{1 3}$ | $\mathbf{1}$ | $\mathbf{1}$ | $\mathbf{6}$ |
| Gen. model: UD | 13 | 1 | 1 | 3 |
| Gen. model: CJ | 0 | 0 | 0 | 3 |
| UD-Def. | $\mathbf{1 5}$ | $\mathbf{2}$ | $\mathbf{0}$ | $\mathbf{4}$ |
| Gen. model: UD | 15 | 2 | 0 | 2 |
| Gen. model: CJ | 0 | 0 | 0 | 2 |
| II-Imm. | $\mathbf{0}$ | $\mathbf{3}$ | $\mathbf{1 6}$ | $\mathbf{2}$ |
| Gen. model: UD | 0 | 3 | 0 | 1 |
| Gen. model: CJ | 0 | 0 | 16 | 1 |
| II-Def. | $\mathbf{2}$ | $\mathbf{1 3}$ | $\mathbf{3}$ | $\mathbf{3}$ |
| Gen. model: UD | 2 | 4 | 0 | 0 |
| Gen. model: CJ | 0 | 9 | 3 | 3 |

Strategies: UDx $=$ Unidimensional based on the x -dimension, $\mathrm{UDy}=$ Unidimensional based on the $y$-dimension, GLC $=$ General linear classifier, RND $=$ Random

## General Discussion

The influential COVIS model of category learning is supported by a great deal of behavioural data (Ashby \& Valentin, 2016). Predominantly, this evidence comes from a single experimental methodology which examines the effect of a factor on rule-based and information-integration category learning. COVIS predicts that its two systems can implement different strategy types, and so each will learn one of these category structures better than the other. Critically, the validity of the inferences from this paradigm hangs on correctly identifying the strategy each individual used to complete the learning task. This is because the experiments investigating COVIS cannot directly control which system participants use to respond. Instead, they manipulate the category structures and hope that this encourages participants to use the optimum system, and thus the correct strategy, for that category structure. Of course, participants may continue to use the sub-optimum system for a particular category structure. Thus, identifying the strategies participants use is crucial: if the participants are using the correct strategy for that category structure, then the experimenters assume that they must also be using the cor-
rect learning system for that structure. Then, any differential effects of a manipulation on each category structure can be attributed to the existence of two systems of category learning, not differing numbers of sub-optimal responders.

Despite its importance for the COVIS model, there is experimental (Edmunds et al., 2015) and modeling (Donkin et al., 2015) evidence to suggest that GRT analysis may be biased towards concluding that participants were using the optimum strategy for the category structure. To explore this possibility, we simulated an experiment by Smith et al. (2014) and showed that it was possible to reproduce their means, inferential statistics and strategy analysis using only participants who used rule-based strategies. Simulated participants classified the information-integration category structure using a conjunction rule, but were recovered by the strategy analysis as using a diagonal (GLC) strategy. This raises the possibility that participants in Smith et al. were, correspondingly, using rule-based strategies in classifying the informationintegration category structure. In other words, Smith et al. cannot be construed as clear evidence for dual-system accounts of category learning, as a single-system (rule-based) account also fits all the data (accuracy and GRT analysis) they presented.

## Implications for the COVIS model

The reported simulation demonstrates an inferential weakness in experiments argued to support COVIS: GRT analysis is not accurate enough to act as a manipulation check. It cannot determine whether manipulating the category structure successfully elicited a corresponding switch in the categorisation system underlying participants' responses. Consequently, it is difficult to judge whether a particular COVISsupporting dissociation is due to the existence of two distinct learning systems, or rather due to participants using different explicit strategies to learn each category structure. This increases uncertainty over conclusions of a swathe of COVIS-supporting studies that rely on comparing rule-based and information-integration category structures (see Ashby \& Valentin, 2016, for a partial list).

In relation to the experimental work by Edmunds et al. (2015), and Edmunds, Wills, and Milton (2016), this simulation also strengthens the evidence that participants can correctly report their categorisation strategies. In those experiments, participants learning information-integration category structures consistently reported using complex, rulebased strategies. In contrast, the GRT analysis identified these participants as using the correct (i.e. diagonal) strategy. In the above simulation, it was shown that participants using a conjunction rule were likely to be misidentified in GRT analysis as using a diagonal (GLC) strategy. Therefore, it seems plausible that all participants learn information-integration category structures explicitly, using rule-based approaches, but GRT analysis misidentifies some of these as using an implicit (GLC) strategy.

## Conclusions

The simulation reported above indicates that drawing conclusions from GRT analysis is risky. This has a knock on effect on the COVIS-supporting studies that rely on this analysis as a manipulation check. More investigations need to be done to understand which strategies participants use and how they are affected by the category structure being learned before we can be sure that experimental dissociations in this literature support a dual-system model of categorization. In other words, we advocate closer attention to due process in the evaluation of dual-system (and single-system) models.

## References

Ashby, F. G., \& Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(1), 33-53. doi: 10.1037/02787393.14.1.33

Ashby, F. G., Maddox, W. T., \& Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. Memory \& Cognition, 30(5), 666-677. doi: 10.3758/BF03196423
Ashby, F. G., \& Valentin, V. V. (2016). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen \& C. Lefebvre (Eds.), Handbook of categorization in cognitive science (2nd ed., pp. 547-572). New York, NY: Elsevier.
Carpenter, K. L., Wills, A. J., Benattayallah, A., \& Milton, F. (2016). A comparison of the neural correlates that underlie rule-based and information-integration category learning. Human Brain Mapping, 37, 3557-3574. doi: 10.1002/hbm. 23259

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., Nosofsky, R. M., Donkin, C., ... Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(4), 933-948. doi: 10.1037/xlm0000083
Edmunds, C. E. R., Milton, F., \& Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. The Quarterly Journal of Experimental Psychology, 68(2), 1203-1222. doi: $10.1080 / 17470218.2014 .978875$
Edmunds, C. E. R., Wills, A. J., \& Milton, F. N. (2016). Memory for exemplars in category learning. In A. Papfragou, D. Grodner, D. Mirman, \& J. C. Truesweell (Eds.), Proceedings of the 38th annual conference of the cognitive science society (pp. 2243-2248). Austin, TX: Cognitive Science Society.
Heathcote, A., Brown, S. D., \& Wagenmakers, E.-j. (2014). An introduction to good practices in cognitive $m$ odeling. New York, NY: Springer.
Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. In Psychology of learning
and motivation (Vol. 63, pp. 77-114). Academic Press. doi: 10.1016/bs.plm.2015.03.001
Lagnado, D. A., Newell, B. R., Kahan, S., \& Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. Journal of Experimental Psychology: General, 135(2), 162183. doi: 10.1037/0096-3445.135.2.162

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. Perception \& Psychophysics, 61, 354-74. doi: 10.3758/BF03206893
Maddox, W. T., \& Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. Perception \& Psychophysics, 53, 49-70. doi: 10.3758/BF03211715
Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, a. D., Gitelman, D. R., Parrish, T. B., ... Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. Cerebral Cortex, 17(1), 37-43. doi: 10.1093/cercor/bhj122
Nosofsky, R. M., \& Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(5), 924-940. doi: 10.1037/0278-7393.28.5.924
Raijmakers, M. E. J., Dolan, C. V., \& Molenaar, P. C. M. (2001). Finite mixture distribution models of simple discrimination learning. Memory \& Cognition, 29(5), 659677. doi: 10.3758/BF03200469

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J., \& Verfaellie, M. (2009). Prefrontal contributions to rule-based and information-integration category learning. Neuropsychologia, 47(13), 2995-3006. doi: 10.1016/j.neuropsychologia.2009.07.011

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. Journal of Experimental Psychology: General, 116(3), 250-264. doi: 10.1037/0096-3445.116.3.250

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. a., \& Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. Psychological Science, 25(2), 447-57. doi: 10.1177/0956797613509112

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., \& Church, B. A. (2015). The time course of explicit and implicit categorization. Attention, Perception, \& Psychophysics, 77(7), 2476-2490. doi: 10.3758/s13414-015-0933-2

Wills, A. J., Inkster, A. B., \& Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. Cognitive Psychology, 80, 1-33. doi: 10.1016/j.cogpsych.2015.04.002

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., \& Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? The Quarterly Journal of Experimental Psychology, 66(2), 299-318. doi: 10.1080/17470218.2012.708349

# Experimental Investigation into the Continuous Pattern of the Relationship Between Color Focality and Short-Term Memory Performance for Colors 

Siyuan Fang (siyuanfang@asagi.waseda.jp)<br>Graduate School of Human Sciences, Waseda University, 2-579-15 Mikajima<br>Tokorozawa City, Saitama Prefecture 359-1192, Japan<br>Tatsunori Matsui (matsui-t@waseda.jp)<br>Faculty of Human Sciences, Waseda University, 2-579-15 Mikajima<br>Tokorozawa City, Saitama Prefecture 359-1192, Japan


#### Abstract

Past studies reported that language-specific color focality has substantial influence on the short-term memory (STM) performance of colors of the speakers of the language, which we call the "focality effect." This study attempts to clarify the continuous pattern of this effect, that is, the manner in which correct recognition possibilities and misrecognition error distances of colors, which are two aspects of the STM performance for colors, change in a gradual fashion along the continuum of color focality. Our experiment, which tests the Japanese language, finds that a U-shaped relationship exists between the focality and the possibility of correct recognition, and that the misrecognition error distance increases as the focality decreases. We speculate that the subjects' frequent and conscious employment of the memorization strategy of coding colors using linguistic categories is one important cause of the detected effect patterns.


Keywords: color focality; short-term memory; continuous pattern; color discriminability; basic color terms

## Introduction

While color sensation changes gradually along the perceptual dimensions of hue, lightness, and chroma (Munsell, 1919), in languages, this is conceptualized into a series of categories. Every language contains a set of basic color terms in its lexicon, such as black, white, red, green, blue, yellow, brown, gray, orange, pink, and purple in the English language (Berlin \& Kay, 1991). The categories signified by the basic color terms (called "basic color categories" for short) are natural categories that have their inner structures formed around their prototypes. This means that within a basic color category, the member colors differ in their focality, namely their "closeness" to the prototype, or in other words, their goodness as a typical example of the category (Rosch, 1973).

In this case study, which tests the Japanese language, we aim to experimentally evaluate the universality of the phenomenon that language-specific color focality influences short-term memory (STM) for colors of the speakers of the language, which we call the "focality effect." More important, we attempt to clarify the continuous pattern of this effect if its existence turns out to be supported in our experiment.

The focality effect first appears in the English language in Heider (1972). In her experiment, Heider used a simplified version of the color array that Berlin and Kay (1991) used. The array was composed of 160 Munsell color chips, 24 of which were selected as test chips. Eight of these chips were focal colors, that is, the colors of the highest focality for
each of the eight chromatic basic color categories that were shared by numerous languages but generally corresponded to the English categories Red, Green, Yellow, Blue, Brown, Purple, Pink, and Orange (Roberson, Davies, \& Davidoff, 2000; Roberson, J. Davidoff, \& Shapiro, 2005). The other 16 chips were of lower focality for these categories, and thus were classified as nonfocal colors. The selection and categorization of the test chips were based on the color-naming data gathered by Berlin and Kay (1991). In each trial in Heider's experiment, a subject was required to watch a color chip for 5 s and then search for it in the color array after a 30-s interval, where the chip was hidden from the subject. For either stimulus type, two indexes of STM performance were measured. The first index was the "memory accuracy score (MAS)," which was defined as the mean number of correct recognitions for this stimulus type across the subjects. The second index was the "error distance score (EDS)," which measured the mean error distance across the incorrect trials of this stimulus type. The English-speaking subjects showed superior performance for both measures of the focal colors relative to the nonfocal ones. Roberson et al. (2000) employed the same experimental paradigm and stimuli. Regarding the English-speaking subjects, a focality effect similar to that reported by Heider was detected in terms of MAS. However, no focality effect was found in terms of EDS. Roberson et al. (2005), which also used this experimental paradigm and stimuli, found that the mean $d^{\prime}$ score (a modified version of MAS) of the test chips that were focal only in Himba (a language mainly spoken in Southern Africa) were significantly higher than that of the test chips that were focal only in English. This effect was also detected in the language of Berinmo, which is mainly spoken in Papua New Guinea. The index of EDS was not used in this study.

Overall, these studies have provided some evidence for the universal existence of the focality effect across languages in terms of MAS. On the other hand, no robust focality effect has been observed in terms of EDS. More empirical evidence is necessary to test whether the focality effect exists for these two STM performance measures. Because no Asian language has been studied in this field, we regard the Japanese language to be suitable as our target language. Furthermore, in these studies, color focality was treated as a categorical variable with only two values: "focal" and "nonfocal." This precluded
any elaborate descriptions of the focality effect. Therefore, in this study, we quantified the concept of color focality in a continuous fashion and delved into the continuous pattern of the focality effect, that is, how STM performance for colors changes gradually along the continuum of color focality.

## Experimental Settings

## Participants, Materials and Environment

Twenty-two subjects ( 11 males and 11 females of ages $M=$ 31.45 and $S D=14.34$, all native Japanese speakers with no color-related art experience), who are either undergraduate or graduate students at Waseda University, took part in the experiment. They all passed the Ishihara Color Vision Test (38 plates, International Edition), and no one reported having color vision deficiencies. Hence, these subjects were considered to have normal color vision.

A color array of Heider (1972)'s design was used. Its layout is shown in Figure 1. This array was made of cardboard $(58.5 \mathrm{~cm} * 28.5 \mathrm{~cm})$, and had color chips embedded in its white surface. Thirty colors (called "test colors" for short) were tested in the formal trials of Session 1. These colors were mounted on the white surface of a $5.0 \mathrm{~cm} * 5.0 \mathrm{~cm}$ piece of cardboard when being presented to the subjects. Chips in the Munsell Book of Color (Glossy Edition) were used.


Figure 1: Layout of color array.
The experiment was performed indoors with fluorescent lighting (type: National FHF 32EX-N-H, daylight color, color temperature: 5000 K ). The experimenter and subject being tested sat opposite each other at a table. The distance between the stimuli and the subject's eyes was controlled at 50 cm . A cardboard separating the two persons was erected along the middle of the table, making the subject unable to see the experimenter's face when observing the stimuli, waiting during the 30 -s interval, and filling out the answer sheets.

## Procedure

The entire experiment, which was carried out in Japanese, consisted of two sessions.

Session 1, which used a procedure similar to that used by Heider (1972), aimed to measure the subjects' STM performance for the test colors. This consisted of 33 trials. In each trial, a test color was presented to the subject for 5 s and then
retrieved by the experimenter. After a 30 -s interval, the color array was presented to the subject, and the subject was asked to report which color in the array he/she thought was the previously presented one by writing the coordinates of the color on an answer sheet. There was no conversation between the experimenter and the subject. Each test color was tested at least once with each subject, and for each subject, the order of color testing was randomly determined. Thus, for each subject, there were three repeated trials, which were intended to prevent the subject from using a strategy of excluding the already tested colors. Before the formal experiment began, a two-trial training session using a different set of test colors was conducted. For each subject, the colors tested during the training were randomly selected.

After all 33 formal trials were completed, a questionnaire was given to the subject. This questionnaire asked the subject to report freely on the strategies that he/she adopted to memorize the test colors during this session.

Session 2 was targeted to elicit the coverage of six basic color categories corresponding to the six Japanese basic color terms akairo (red), pinkuiro (pink), kiiro (yellow), orenjiiro (orange), chairo (brown), and murasakiiro (purple) (Uchikawa \& Boynton, 1987). Then, the focality of each test color was quantified using a modified version of Berlin and Kay (1991)'s method. First, the subject was required to write on six answer sheets (one for each basic color term) all colors that he/she thought could be named by the term. The answer sheets were provided to each subject in random order. Next, the subject was asked to report the colors they thought were the best examples of each of the six basic color terms. This was accomplished by writing the coordinates of the colors on an answer sheet. Multiple answers were allowed for each term, but the subject was instructed to narrow his/her selections as much as possible.

## Statistical Analysis and Results

## Variable Definitions

Focality Score We used the data obtained from Session 2 to specify the coverages of the six basic color categories over the array, and quantified the focality of the test colors. We first computed the six attributes for each test color: Red Index, Pink Index, Yellow Index, Orange Index, Brown Index, and Purple Index. These attributes measured the intersubject naming consistency of the color in terms of each basic color term. The Red Index was defined as the percentage of subjects who named the color as red, and the other five indexes were similarly defined. Then, we designated the Overall Index (OI) of a color as the largest of the six single-term-based indexes of the color. We classified a color into the color category Red if its OI was its Red Index, the color category Pink if its OI was its Pink Index, and so forth. Figure 2 shows the distribution of the nonzero OIs and the partition of the six basic color categories.

For each basic color category, most of the colors having the largest OIs were also frequently selected as the best examples
during the second part of Session 2. Thus, it is reasonable to deem the OI of a color as reflecting the appropriateness of the color as a typical example of the category to which the color belongs. In this manner, we defined the focality score (FS) of a color as its OI value.


Figure 2: Distribution of Overall Indexes of test colors (colors within area covered by thin diagonal stripes) and other relevant colors, and partition of the six basic color categories. Color depth represents Overall Index magnitude. [*: Orange Index = Brown Index]

Discriminability Score We defined the discriminability score (DS) of a test color on the color array as the average of the color differences between the test color and its eight adjacent colors. In this study, a color difference was defined as a Euclidean distance in the CIE L*a*b* color space. Thus, before calculating color differences, we transformed the Munsell coordinates of all relevant colors into CIE xyY coordinates using the O.S.A.-developed conversion tables (Newhall, Nickerson, \& Judd, 1943), then XYZ coordinates, and finally the $L^{*} a * b *$ coordinates. In the final transformation step, the parameter values of the CIE D50 standard illuminant, which resembled the light source used in this experiment, were used.
STM Performance Index 1: Memory Accuracy Score We adopted MAS as one index of STM performance. It measures the probability for which a color can be accurately recognized. Since FS is continuous in our study, it is necessary to take MAS also as continuous. We defined the MAS of a test color as the percentage of the trials where the subjects correctly recognized the color.
STM Performance Index 2: Error Distance Score The EDS is adopted as another index of STM performance. EDS measures the expected error extent in the case of misrecognition. As in Heider (1972)'s and Roberson et al. (2000)'s studies, the EDS for a test color is defined as the mean of the color differences between the test color and the colors mistaken for the test color in the incorrect recognition trials.

## The Relationship Between FS and MAS

In order to determine the continuous pattern of the relationship between FS and MAS, we first conducted regression analyses on the FS data and the original MAS data of the
test colors to obtain a general impression of the relationship pattern. No statistically significant linear relationship could be detected ( $R^{2}=0.066, P=0.171\left[B_{F S}=0.152, P=0.171\right]$ ), but a significant quadratic relationship was found $\left(R^{2}=0.237\right.$, $P=0.026\left[B_{F S}=-1.064, P=0.045 ; B_{F S * F S}=1.073, P=\right.$ $0.021]$ ).

Brown and Lenneberg (1954) found that color discriminability could facilitate STM performance for colors, which we call the "discriminability effect." Later, Heider (1972) and Lucy and Shweder (1979) pointed out that because the colors in the color array were unequal in discriminability, it was possible that it was color discriminability, not color focality, that caused the detected variance in STM performance for colors. This possible source of distortion was checked by using the following procedure. First, we looked into the relationship between DS and MAS. A significant positive linear regression model could be established between these two variables ( $R^{2}=0.353, P=0.001\left[B_{D S}=0.024, P=0.001\right]$ ). Then, regression analyses investigating the relationship between FS and DS were conducted. These analyses produced neither a significant linear model $\left(R^{2}=0.014, P=0.534\left[B_{F S}\right.\right.$ $=1.710, P=0.534])$ nor a significant quadratic one $\left(R^{2}=\right.$ $0.029, P=0.675\left[B_{F S}=-7.032, P=0.618 ; B_{F S * F S}=7.712, P\right.$ $=0.527])$. Nevertheless, we noticed that a slight U-shaped relationship could be recognized when we scrutinized the scatter plot. This means that the possibility that DS mediated the FS-to-MAS relationship could not be ruled out. Hence, we conducted partial linear and quadratic regressions on FS and MAS while excluding the influence of DS on MAS. No significant linear relationship was found $\left(R^{2}=0.054, P=0.217\right.$ [ $\left.B_{F S}=0.111, P=0.217\right]$ ), but a significant quadratic one was detected $\left(R^{2}=0.233, P=0.028\left[B_{F S}=-0.893, P=0.037\right.\right.$; $\left.B_{F S * F S}=0.885, P=0.018\right]$. This is plotted in Figure 3A). This is similar to the results of the initial regressions, and indicates that a significant quadratic relationship exits between FS and MAS even if DS has been treated as a control variable.


Figure 3: (A)Relationship between Focality Score and Memory Accuracy Score (corrected);(B)Relationship between Focality Score and Error Distance Score (corrected) in the general case, with a data point unexplained by the regression model depicted by a magenta-colored rhombus.

## The Relationship Between FS and EDS

The continuous pattern of the relationship between FS and EDS was investigated in the following manner. First, to obtain a preliminary impression of what the relationship pattern looks like, we carried out linear and quadratic regression analyses on the FS data and the original EDS data. Only the quadratic model reached statistical significance (quadratic model: $R^{2}=0.250, P=0.027\left[B_{F S}=-30.113, P=0.019\right.$; $\left.B_{F S * F S}=22.376, P=0.041\right]$; linear model: $R^{2}=0.110, P$ $=0.084\left[B_{F S}=-4.694, P=0.084\right]$ ). Then, a positive linear relationship was found between DS and EDS through a regression analysis ( $R^{2}=0.299, P=0.003\left[B_{D S}=0.692, P=\right.$ $0.003])$. In order to remove the distorting influence of the discriminability effect (as in the case of MAS), the regressions on the FS and EDS data were repeated but with the impact of DS on EDS partialed out. A significant quadratic model appeared $\left(R^{2}=0.221, P=0.044\left[B_{F S}=-22.166, P=0.041\right.\right.$; $\left.\left.B_{F S * F S}=15.895, P=0.085\right]\right)$, but not a linear one ( $R^{2}=0.121$, $P=0.070\left[B_{F S}=-4.110, P=0.070\right]$ ). This resembles the results of the initial regressions.

Nevertheless, there exists a test color that appears to be isolated from the cluster of other high-FS test colors at the EDS coordinates. Owing to the employment of the least squares method, this data point could have exerted a disproportionately strong influence on the relationship pattern. In order to determine what pattern the relationship actually takes in the general case, we reran the regressions on the corrected dataset but did not include this data point. This time we obtained a linear relationship ( $R^{2}=0.236, P=0.010\left[B_{F S}=-5.669, P=\right.$ 0.010], which is plotted in Figure 3B), instead of a quadratic one. (When adding FS*FS to the regression as a predictive variable, neither $B_{F S}$ nor $B_{F S * F S}$ achieved significance, although the model remained significant). The removed color is $5 \mathrm{YR} 4 / 8$. It was mistaken as the color one-unit above it ( 5 YR 5/12) in all its misrecognition cases. This misrecognition pattern is difficult to explain by the strategy of linguistic categorical color coding which will be discussed later. We thus conjecture that it might result from other memorization strategies, which needs further exploration.

## Memorization Strategies

In the questionnaire conducted at the end of Session 1, the subjects reported a total of six memorization strategies. For each strategy, Table 1 offers a brief description and shows how many subjects reported it.

## General Discussion

## The Continuous Patterns of the Focality Effect

Our experimental results demonstrated that in the Japanese language, color focality can affect STM performance for colors in a statistically significant way in terms of both correct recognition possibility and misrecognition error distance. With regard to the continuous patterns of the focality effect, a significant U -shaped quadratic regression function can be established between FS and MAS, which implies that STM per-

Table 1: Memorization Strategies Reported by Subjects in Questionnaire.
\(\left.\left.$$
\begin{array}{cl}\hline \text { Number of Reports } & \text { Brief Description } \\
\hline 16 & \begin{array}{l}\text { Use basic color concepts as refer- } \\
\text { ence points, and then fine-tune along } \\
\text { the dimensions of lightness and/or } \\
\text { saturation }\end{array} \\
\text { Associate the test color with the } \\
\text { color of a familiar object, e.g., the } \\
\text { banner of Waseda University, a Bor- } \\
\text { deaux wine, or lipstick, and then } \\
\text { fine-tune along the dimensions of } \\
\text { lightness and/or saturation }\end{array}
$$\right\} \begin{array}{l}Directly memorize the visual image <br>

of the test color\end{array}\right\}\)| Use the degree of preference for the |
| :--- |
| test color as a cue |

formance is best for colors at the two terminals of the focality continuum, and begins to decrease as the focality moves toward the intermediate level. In addition, a significant negative linear regression function can be established between FS and EDS under general circumstances. This suggests that the average error distance in the case of misrecognitions for a color decreases as its focality increases.

## One Cause of the Focality Effect Patterns

To determine what caused the continuous patterns, we examined the memorization strategies reported by the subjects (Table 1). We noticed that the strategy of encoding colors using linguistic color categories, which has the highest number of reports, might have played an important role in the formation of the detected focality effect patterns.
A detailed description of the procedure of this strategy in a single trial is as follows: The subjects consciously encoded the test color using the basic color terms while observing the test color. The basic color terms were used as reference points, which means that the subject anchored the test color to the central points of the basic color categories, namely, the most typical colors of these categories. The subject then retained this linguistic code in his/her STM during the waiting
period. Finally, during the phase of color searching, the subject decoded the code to recover the test color.

For convenience of discussion, color focality is generally divided into the levels of "high," "medium," and "low," and their respective ways of being coded are described as follows: A high-focality color can be encoded using only one basic color term since it is, or is substantially close to, the central point of the basic color category. With regard to a mediumfocality color, its coding needs some modifiers in hue and (or) lightness besides a basic color term, for example, bright orange, dark brown, and purplish pink. For a low-focality color, because it is situated at the border region between two basic color categories, the two basic color terms corresponding to the two categories are used to constitute the code for this color.

With the employment of this strategy, it is obvious that the correct recognition possibility for a color is mainly determined by 1) how easily the code for the color can be retained in STM during the waiting period, and 2) the semantic ambiguity of the code for the color, or in other words, how accurately the encoded color can be recovered from the code. Since codes for colors of all three types can be formed by just a few words, they will not cause a memory burden. This implies that the rate of successful retaining should be high for each color type. On the other hand, the variable of semantic ambiguity bears a much larger intertype variance, which indicates its chief role in mediating the impact of color focality on STM performance for colors. For a high-focality color, its code generally consists of a sole basic color term, which possesses a fairly plain meaning since any Japanese speaker is able to understand the definition of a basic color term. Thus, during the searching phase, the signifier of the code can be pinpointed in high precision. By contrast, the modifiers in the code for a medium-focality color have much vaguer meanings. Even if the subjects have carried the code into the decoding phase without mistakes, they will find themselves lost in numerous candidates, all of which more or less match the description. This will surely lower their chances of finding the one that they have actually coded. Following this logic, the code for a low-focality color, which involves basic color terms but no modifiers, should also be regarded as unequivocal in meaning. The central points of the basic color terms, as in the case of a high-focality color, can serve as reliable reference points for the localization of the encoded color. In brief, the semantic ambiguity of color codes, which negatively influences the likelihood of correct recognition for colors, is low for high- and low-focality colors and high for medium-focality colors. Thus, high- and low-focality colors tend to have higher rates of correct recognition than those of medium-focality ones. This is exactly what our experimental results have shown. In addition, owing to the fact that for any color the semantic ambiguity of its code is a languageinherent and thus subject-independent attribute, this continuous pattern can be expected to have a high degree of intralanguage consistency, or in other words, a high likelihood to be
replicated if the experiment is repeated using the same language.

With regard to the misrecognition error distance of a color, within the framework of linguistic categorical coding, this mainly depends on which parts of the code the subjects have forgotten, and how many times each of these parts have been forgotten. For a high-focality color, once a subject has forgotten the sole basic color term during the waiting period, in the searching phase he/she is unable to tell the basic color category to which the test color belongs. His/her selection will thus be random, although other memory clues, such as the visual image of the test color, can be of help. It is easy to imagine that under this circumstance, a large error will occur. The loss of the basic color term for a medium-focality color or both of the basic color terms for a low-focality color will lead to similar consequences. For a medium-focality color, when only the modifiers have been forgotten, given that the basic color term has become the only guide, the central point of this basic color category may pull the subjects' selections toward it. In this case, a misrecognition is expected to occur, but within a moderate error range that is approximately half the "category radius." Following the same logic, with regard to the code for a low-focality color, when one of its two basic color terms has been forgotten, the remaining one will tend to drag the subjects' selections toward the central point of the category it represents. On this occasion, because the test color is situated at the border region of the category, a selection with a error distance of approximately one categoryradius long might take place. Note that owing to the small total number of memory losses suggested by the small memory burden imposed by the color codes, it is possible that some of these "forgetting types" did not occur in our experiment. Thus, one explanation for the focality effect pattern that we detected is that our subjects have never forgotten the basic color terms in the codes for the high- and medium-focality colors. In addition, the small sample size of memory losses means that the distribution of occurrence frequency across the forgetting types can hardly be consistent across experiments even when using the same language. In other words, if the experiment is repeated, a substantially different frequency distribution across the forgetting types will occur, which will lead to a very different focality effect pattern.

## The Universality of the Focality Effect Patterns

Several past studies on STM performance for colors, which used English-speaking subjects, also recorded their subjects' memorization methods. Lucy and Shweder (1988) recorded the subjects' incidental remarks on memorization strategies during the course of their experiments, and they carried out a questionnaire on memorization strategies when the experiments were finished. They provided a quantitative report which showed that the strategy of linguistic categorical coding was the most frequently adopted, followed by the strategies of direct retention of visual image, present object association, and absent object association. This coincides well with the results of our questionnaire. Brown and Lenneberg
(1954), Lucy and Shweder (1979), and Garro (1986) also reported the use of linguistic categorical coding by their subjects, although they did not provide detailed statistics. The fact that linguistic categorical coding is employed as a chief memorization strategy by both Japanese speakers and English speakers suggests that its applicability is possibly universal across languages. Moreover, considering the hypothesized close ties of this strategy to the formation of the continuous patterns of the focality effect, this further implies that all languages may share a common language-based mechanism for focality-effect generation.

In terms of the focality effect pattern for the possibility of correct recognition, considering its presumed intralanguage consistency, it can be expected to be observed in other languages. This conjecture is supported by the agreement between the FS-to-MAS relationship detected in our experiment and the superiority of focal colors to nonfocal colors in correct recognition possibilities reported by Heider (1972), Roberson et al. (2000) and Roberson et al. (2005). On the other hand, even if the use of the strategy of linguistic categorical coding is universal across languages, because of the lack of intralanguage consistency, it is difficult to find a consistent focality effect pattern across languages in terms of the misrecognition error distance. A comparison of the results of Heider (1972)'s and Roberson et al. (2000)'s studies and our study demonstrated such interlanguage inconsistency.

A trivial case as it might be, reports show that there exit languages that possibly lack basic color terms, e.g., Piraha and Warlpiri (Everett, 2005; Wierzbicka, 2008; but see Regier, Kay, \& Khetarpal, 2009). This means that the linguistic definition of color focality and the memorization strategy of linguistic categorical coding possibly cannot be applied to such languages. Thus, the discussions in this section may be unsuitable for these languages.

## Conclusion and Implications for Future Work

This study is the first to probe into the continuous patterns of the focality effect. Our experiment confirmed the existence of the focality effect in the Japanese language, and clarified its continuous patterns in terms of correct recognition possibility and misrecognition error distance, which were two aspects of short-term memory performance. Correct recognition possibility is highest at the ends of the continuum of color focality, and decreases as color focality moves toward the medium region from either end. In addition, misrecognition error distance for colors, in the general case, decreases as color focality increases. We speculate that the subjects' frequent and conscious use of memorization strategies, especially the strategy of encoding colors using linguistic color categories, played an important role in the formation of focality effect patterns. The interstudy agreement on the recordings of memorization strategies suggests that the employment of linguistic categorical coding is possibly universal across languages. For this reason, and also owing to its likely high intralanguage consistency, we expect that the focality effect pattern
for correct recognition possibility that we detected can also be found in other languages. Empirical evidence for more languages is needed to evaluate this hypothesis. In addition, it is also interesting to see whether this focality effect pattern can also be found in the categories of other domains. These domains can be simple perceptual categories such as shapes and phonemes, complicated multimodal concepts such as animals and tools, or even emotionally or socially meaningful signals such as human facial expressions.

## References

Berlin, B., \& Kay, P. (1991). Basic color terms: their universality and evolution. Oakland, California: University of California Press.
Brown, R. W., \& Lenneberg, E. H. (1954). A study in language and cognition. Journal of Abnormal and Social Psychology, 49(3), 454-462.
Everett, D. (2005). Cultural constraints on grammar and cognition in Piraha: another look at the design features of human language. Current Anthropology, 46(4), 621-646.
Garro, L. C. (1986). Language, memory, and focality: a reexamination. American Anthropologist, 88(1), 128-136.
Heider, E. R. (1972). Universals in color naming and memory. Journal of Experimental Psychology, 93(1), 10-20.
Lucy, J. A., \& Shweder, R. A. (1979). Whorf and his critics: linguistic and nonlinguistic influences on color memory. American Anthropologist, 81(3), 581-615.
Lucy, J. A., \& Shweder, R. A. (1988). The effect of incidental conversation on memory for focal colors. American Anthropologist, 90(4), 923-931.
Munsell, A. H. (1919). A color notation. New York: Munsell Color Company.
Newhall, S. M., Nickerson, D., \& Judd, D. B. (1943). Final report of the OSA subcommittee on the spacing of the Munsell colors*. Journal of the Optical Society of America, 33(7), 385-418.
Regier, T., Kay, P., \& Khetarpal, N. (2009). Color naming and the shape of color space. Language, 85(4), 884-892.
Roberson, D., Davies, I. R. L., \& Davidoff, J. (2000). Color categories are not universal: replications and new evidence from a stone-age culture. Journal of Experimental Psychology: General, 129(3), 369-398.
Roberson, D., J. Davidoff, I. R. L. D., \& Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. Cognitive Psychology, 50(4), 378-411.
Rosch, E. H. (1973). Natural categories. Cognitive Psychology, 4(3), 328-350.
Uchikawa, K., \& Boynton, R. M. (1987). Categorical color perception of Japanese observers: comparison with that of Americans. Vision Research, 27(10), 1825-1833.
Wierzbicka, A. (2008). Why there are no 'colour universals' in language and thought. Journal of the Royal Anthropological Institute, 14(2), 407-425.

# Deconstructing Social Interaction: The Complimentary Roles of Behaviour Alignment and Partner Feedback to the Creation of Shared Symbols 

Nicolas Fay (nicolas.fay@gmail.com)<br>School of Psychological Science, University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia<br>Bradley Walker (bradley.walker@uwa.edu.au)<br>School of Psychological Science, University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia<br>Nik Swoboda (nswoboda@fi.upm.es)<br>Department of Artificial Intelligence, Universidad Politecnica de Madrid, Calle Ramiro de Maeztu, 7, 28040 Madrid, Spain


#### Abstract

This paper experimentally tests the contribution of two distinct aspects of social interaction to the creation of shared symbols: behaviour alignment and concurrent partner feedback. Pairs of participants ( $\mathrm{N}=120$, or 60 pairs) completed an experimental-semiotic game, similar to Pictionary, in which they tried to communicate a range of recurring meanings to a partner by drawing on a shared whiteboard (without speaking or using numbers of letters in their drawings). The opportunity for sign alignment and/or concurrent partner feedback was manipulated in a full factorial design. Each process made a distinct contribution to the evolution of shared symbols: sign alignment directly influenced communication success, and concurrent partner feedback drove sign simplification and symbolization. These complimentary processes led to the interactive evolution of effective and efficient human communication systems.


Keywords: Human Communication, Interaction, Icon, Symbol, Cultural Evolution, Language Evolution

## Introduction

Human cognition and behaviour is dominated by symbol use, evident from our everyday use of numeric and linguistic systems. But where do these symbols come from? This question is presented by Harnad (1990) as the symbol grounding problem: how shared meanings can arise from arbitrary symbols in the absence of a pre-established shared symbol system. A solution to the symbol grounding problem was offered by Peirce (1931), who suggested that symbols evolved from iconic signs that share a non-arbitrary correspondence between the sign and its meaning.

This icon-to-symbol transition has been convincingly demonstrated in experimental-semiotic communication games. These experiments examine the creation of novel human communication systems under controlled laboratory conditions (for reviews see Fay, Ellison, \& Garrod, 2014; Galantucci, 2017; Tamariz, 2017). They do this by using a paradigm in which human participants communicate without using their existing shared language. Typically, participants communicate in a novel modality, for example, through drawing (Galantucci, 2005; Garrod, Fay, Lee, Oberlander, \& MacLeod, 2007; Healy, Swoboda, Umata, \& King, 2007; Roberts, Lewandowski, \& Galantucci, 2015) or
by gesture (Christensen, Fusaroli, \& Tylén, 2016; Fay, Arbib, \& Garrod, 2013; Schouwstra \& de Swart, 2014; Stolk, Verhagen, \& Toni, 2016) and the experimenters examine how the communication systems arise and evolve over repeated interactions.

A key finding is the importance of iconic signs and social interaction to the creation of shared symbols (Galantucci, 2005; Garrod et al., 2007). In Garrod et al. (2007), pairs of participants tried to communicate a set of recurring meanings to their partner by drawing on a shared whiteboard. Like the game Pictionary, participants were not allowed to speak or use letters or numbers in their drawings. This procedure forced participants to create a novel communication system from scratch. Over repeated interactions three things happened: communication success improved, the signs used were transformed from complex iconic signs to simpler, more symbolic signs, and participants increasingly used the same signs to communicate the same meanings (i.e., their signs aligned; see Figure 1). This pattern, the creation of an effective inventory of shared symbols, has been widely replicated (Caldwell \& Smith, 2012; Fay, Garrod, Roberts, \& Swoboda, 2010; Garrod, Fay, Rogers, Walker, \& Swoboda, 2010; Theisen, Oberlander, \& Kirby, 2010).

These studies indicate that social interaction is crucial to the creation of effective and efficient human communication systems, but they are not clear on the precise mechanisms driving these outcomes. To better understand this, the present experiment isolates two important aspects of social interaction - behaviour alignment and concurrent partner feedback - and investigates the contribution of each to the evolution of shared symbols.

Pickering and Garrod (2004) argue that linguistic alignment drives successful communication. While there is a correlation between referential alignment and communication success (Fay, Lister, Ellison, \& Goldin-Meadow, 2014; Fusaroli et al., 2012; Reitter \& Moore, 2014), the causal role of referential alignment on communication success is unclear. If referential alignment directly influences communication success, then prohibiting interacting participants from aligning their signs will lower communication success.

Concurrent partner feedback can take a variety of forms. During conversation, listeners are co-narrators who provide verbal feedback (e.g., saying "mhm" while listening to a speaker) and visual feedback (e.g.,
frowning or nodding), that improves the flow of conversation (Bavelas, Coates, \& Johnson, 2000; Clark \& Krych, 2004; Mein, Fay, \& Page, 2016). Like listeners in a conversation, participants engaged in an experimental-semiotic game can signal their attention and understanding by annotating their partner's sign, e.g., by adding a tick mark (see Figure 1). During conversation listeners can indicate a communication breakdown and initiate a repair (e.g., by asking the speaker for clarification; Dingemanse et al., 2015; Schegloff, 2000). In addition to these information expansion requests, listeners can drive information contraction by indicating their understanding (e.g., by saying "yeah, yeah"). So, concurrent partner feedback during an experimental-semiotic game may drive communication success and sign simplification/symbolization.

The present experiment examined the influence of sign alignment and concurrent partner feedback on communication success and sign symbolization. It also tested if each process operates independently or if they interact.


Figure 1. Sign simplification and alignment for the meaning 'Museum' across 6-games between a pair of participants in the present experiment. Participants alternated drawing and identifying roles from game to game. At Game 1 Museum was communicated using a complex iconic sign that included a dinosaur, an exhibit space and two viewers. By Game 6 the sign has lost much of its initial iconicity, evolving into a simpler, more symbolic representation, communicated by the dinosaur's spine. In addition, partners' signs became increasingly similar, or aligned, across games.

## Method

The experiment received approval from the University of Western Australia Ethics Committee. All participants viewed an information sheet before giving written consent to take part in the study. The information sheet and consent form were both approved by the aforementioned Ethics Committee.

## Participants

One-hundred and twenty undergraduate students ( 84 females) participated in exchange for course credit or payment. Participants were tested in unacquainted pairs in testing sessions lasting 1 hour. All participants were free of any uncorrected visual impairment.

## Task and Procedure

Participant tried to graphically communicate a series of confusable meanings to their partner. Like the game Pictionary, participants were prohibited from speaking or using letters or numbers in their drawings. The Director would draw each meaning from their ordered list ( 16 targets plus 4 distractors; see Table 1 for a complete listing) and their partner, the Matcher, would
try to identify each meaning from their randomly ordered list of the same meanings.
The task was administered using a virtual whiteboard tool (Healy, Swoboda, \& King, 2002), which recorded all drawing activity. Each participant sat at a computer terminal where drawing input and meaning selection was made via a standard mouse. For the Director, each to-be-communicated meaning was highlighted in white text on a dark background at the top of the interface. Holding down the left mouse button initiated drawing. Director drawing was restricted to black ink and Matcher drawing was restricted to green ink (to distinguish between participants). By clicking an erase button on the interface participants were able to erase parts of the drawing. All drawing and erasing activity was displayed simultaneously on the Director and Matcher's shared virtual whiteboards. When the matcher believed they had identified the director's intended meaning they clicked the relevant button at the top of their interface, where there was a list of buttons corresponding to the competing meanings. Meaning selection brought the current trial to an end and initiated the next trial. No time limit was imposed, and participants were given no explicit feedback with regard to their communication success. Participants communicated remotely across networked computers and were unaware of their partner's identity.

Table 1. The set of meanings that Directors communicated to Matchers (distractor meanings given in italic). Target and distractor meanings were fixed across conditions and throughout the experiment.

| Places | People | Entertain- <br> ment | Objects | Abstract |
| :--- | :--- | :--- | :--- | :--- |
| Art Gallery | Arnold <br> Schwarzenegger | Cartoon | Computer <br> Monitor | Homesick |
| Parliament | Brad Pitt | Drama | Microwave | Loud |
| Museum | Hugh Grant | Sci-Fi | Refrigerator | Poverty |
| Theatre | Russell Crowe | Soap Opera | Television | Sadness |

The experiment examined the contribution of behaviour alignment and concurrent partner feedback to communication success and sign symbolization. Participants were randomly assigned to one of four conditions that represented a combination of the factors of interest: +Alignment + Feedback ( $\mathrm{N}=30$, or 15 dyads), +Alignment -Feedback ( $\mathrm{N}=30$, or 15 interacting), -Alignment +Feedback ( $\mathrm{N}=30$, or 15 dyads) and -Alignment -Feedback ( $\mathrm{N}=30$, or 15 dyads). In the -Alignment conditions participants were instructed not to copy their partner's drawings. They were told they would have to use a different sign to that used by their partner to communicate each meaning. In the -Feedback conditions Matchers were unable to provide within-trial feedback. Specifically, they were unable to draw while acting as the Matcher (this functionality was removed from the virtual whiteboard tool). In this condition the Director clicked a send button when they had finished their drawing. Once done the list of competing meanings became available for selection by the Matcher. Thus, Matchers were unable to interrupt the Director's communication and bring the trial to an end.

## Results

Participants followed the instructions not to align their signs（manipulation check）．Not being able to align their signs reduced communication success．By contrast，eliminating the opportunity for concurrent partner feedback did not directly affect communication success．Concurrent partner feedback affected sign simplification；when feedback was eliminated the signs produced were more complex．Sign alignment also affected sign simplification，but the effect was much weaker compared to the effect of concurrent partner feedback．See Figure 2 for examples of sign alignment and simplification in the different conditions．

| ＋Alignment <br> ＋Feedback |  | $T \\| 11 F$ | P |  | $ص$ | $5$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ＋Alignment <br> －Feedback |  |  |  | $\begin{aligned} & 80 \\ & 8_{8}^{\circ}= \end{aligned}$ |  | $\begin{aligned} & \text { Aógo } \\ & \uparrow \rightarrow \gamma^{\prime} \end{aligned}$ |
| －Alignment <br> ＋Feedback |  |  |  | ${ }^{8}$ |  | Tn |
| －Alignment <br> －Feedback | $\%$ |  | 回田䭪 | $10$ |  | (我 |
|  | Game 1 | Game 2 | Game 3 | Game 4 | Game 5 | Game 6 |

Figure 2．Sign alignment，simplification and symbolization for the meaning＇Parliament＇across 6－ games between participants in the different experimental conditions．Participants instructed not to copy their partner＇s sign for each meaning did so：one participant drew a building with a flag to communicate ＇Parliament＇and their partner drew a speaker at a podium（－Alignment＋Feedback condition）；another drew a parliamentary speaker with a hammer，and their partner drew a series of buildings（－Alignment－ Feedback condition）．When permitted to copy their partner＇s signs，sign alignment was observed：onto a flag（＋Alignment＋Feedback condition），or people seated around a table（＋Alignment－Feedback condition）．These examples highlight the diversity of signs used to communicate the same meaning in the present study．Concurrent partner feedback had a strong effect on sign simplification and symbolization： with feedback the signs were dramatically simplified across games（＋Feedback conditions），and without feedback they retained considerable sign complexity （－Feedback conditions）．

## Manipulation Check：Sign Alignment

Participants in the－Alignment conditions were instructed not to copy the drawings produced by their partner．Sign alignment was quantified by rating the similarity of pairs of drawings of the same meaning from each pair（at Game 1－2，2－3，3－4，4－5，5－6）on a Likert scale from 0－9，where $0=$ very dissimilar and $9=$ very similar（BW）． 4800 pairs of drawings were rated for similarity（16 meanings X 5 pairs of adjacent games X 15 pairs X 4 conditions）．Sign alignment scores for the drawings produced in the different conditions are shown in Figure 3．The results indicate that participants followed the－Alignment instructions： those permitted to copy their partner＇s signs showed
increasing sign alignment across games，whereas those not permitted to copy their partner＇s signs returned lower overall sign alignment scores that did not change across games．

Drawing alignment scores were entered into a mixed－design ANOVA that treated Alignment （＋Alignment，－Alignment）and Feedback（＋Feedback， －Feedback）as between－participant factors and Game （1－2，2－3，3－4，4－5，5－6）as a within－participant factor． This returned a statistically significant Alignment by Game interaction $\left[F_{\text {Linear }}(1,56)=50.849, p<0.001, \eta_{p}^{2}=\right.$ $0.564]$ ．The interaction effect is explained by the increase in sign alignment scores across games in the ＋Alignment conditions $\left[F_{\text {Linear }}(1,29)=131.622, \quad p<\right.$ $\left.0.001, \eta_{p}^{2}=0.819\right]$ and a null effect of Game in the －Alignment conditions $\left[F_{\text {Linear }}(1,29)=0.851, p=0.364\right]$ ． The Alignment manipulation worked．


Figure 3．Change in sign alignment scores（plotted for each pair）for the different conditions across games 1－6．The horizontal dashed red line indicates neutral sign alignment．The dark blue straight line is the linear model fit and the light grey shaded area is the $95 \%$ confidence interval．

No outliers were identified using the Interquartile Range rule（Moore，McCabe，\＆Craig，1993）．Drawing alignment scores were entered into a mixed－design ANOVA that treated Alignment（＋Alignment， Alignment）and Feedback（＋Feedback， －Feedback）as between－participant factors and Game （1－2，2－3，3－4，4－5，5－6）as a within－participant factor． This returned a statistically significant Alignment by Game interaction［ $F_{\text {Linear }}(1,56)=50.849, p<0.001, \eta_{p}^{2}=$ $0.564]$ ．The interaction effect is explained by the strong increase in sign alignment scores across games in the＋Alignment conditions［ $F_{\text {Linear }}(1,29)=$ 131．622，$p<0.001, \eta_{p}^{2}=0.819$ ］and a null effect of Game in the－Alignment conditions［ $F_{\text {Linear }}(1,29)=$ 0．851，$p=0.364]$ ．

The Alignment manipulation worked：participants who were allowed to copy their partner＇s drawings did so，and increasingly did so across games，whereas those who were prohibited from doing so did not copy their partner＇s drawings．

## Communication Success

Communication success was operationalized as the percentage of meanings accurately identified by the Matcher. Figure 4 shows the change in communication success (\%) across games 1-6 in the different conditions. The results show an increase in communication success across games in all conditions, but the increase is stronger in the +Alignment conditions compared to the -Alignment conditions.

One outlier ( $0.28 \%$ of data) was identified using the Interquartile Range rule (see Moore et al., 1993). This value was replaced by next lowest value. The communication success scores were then entered into the same mixed-design ANOVA used previously. This returned a statistically significant Alignment by Game interaction $\left[F_{\text {Linear }}(1,56)=135.151, \quad p<0.001, \quad \eta_{p}^{2}=\right.$ $0.707]$. In all conditions communication success improved across games: +Alignment conditions $\left[F_{\text {Linear }}(1,29)=117.268, p<0.001, \eta_{p}^{2}=0.802\right]$ and -Alignment conditions $\left[F_{\text {Linear }}(1,29)=38.435, p<0.001\right.$, $\left.\eta_{p}^{2}=0.570\right]$. However, the improvement in communication success (differences score: game 6 game 1) was stronger in the +Alignment conditions ( $M=24.17, S D=12.031$ ) compared to the -Alignment conditions $(M=13.96, S D=13.993), \quad t(58)=3.030$, $p=0.004, d=0.782$. The same pattern of results was returned when the communication success data was analyzed using logistic mixed effects modeling.

Sign alignment improved communication success, establishing a causal link between behaviour alignment and communication success. By contrast, concurrent partner feedback did not directly influence communication success [ $p=0.871$ ].


Figure 4. Change in communication success (plotted for each pair) for the different conditions across games $1-6$. The dark blue straight line is the linear model fit and the light grey shaded area is the $95 \%$ confidence interval.

## Sign Simplification and Symbolization

Following Garrod et al., (2007) simpler signs were considered to be more symbolic. Sign complexity was measured using Pelli et al.'s (2006) information
theoretic measure of perimetric complexity [Perimetric complexity $=$ (inside + outside perimeter) ${ }^{2} /$ ink area]. Previous work indicates this to be an effective scalefree measure of drawing complexity (Fay et al., 2010; Garrod et al., 2007; Tamariz \& Kirby, 2014). Sign complexity scores for the drawings produced in the different conditions are shown in Figure 5. Sign complexity tended to decrease across games in all conditions, but sign complexity was lower in the +Feedback conditions compared to the -Feedback conditions.


Figure 5. Change in sign complexity (plotted for each pair) for the different conditions across games 1-6. The dark blue straight line is the linear model fit and the light grey shaded area is the $95 \%$ confidence interval.

Ten outliers ( $2.78 \%$ of data) were identified using the Interquartile Range rule. These values were replaced by the next highest value. The sign complexity scores were then entered into the same mixed-design ANOVA used previously. This returned a statistically significant three-way Alignment by Feedback by Game interaction $\left[F_{\text {Linear }}(1,56)=4.140, p=\right.$ 0.047, $\eta_{p}^{2}=0.069$ ]. To understand the three-way interaction separate Alignment by Game ANOVAs were carried out for each level of Feedback. For the + Feedback conditions this returned a main effect of Game [ $\left.F_{\text {Linear }}(1,28)=73.809, p<0.001, \eta_{p}^{2}=0.725\right]$ with no other effects reaching statistical significance (ps> 0.304). So, both +Feedback conditions showed a similarly strong decrease in sign complexity scores across games, and there was no statistical evidence that sign alignment affected sign symbolization. A different pattern was returned by the -Feedback conditions. ANOVA returned a statistically significant Alignment by Game interaction $\left[F_{\text {Linear }}(1,28)=6.608\right.$, $\left.p<0.016, \eta^{2}=0.191\right]$. This interaction effect is explained by the statistically significant decrease in sign complexity scores across games in the +Alignment -Feedback condition $\left[F_{\text {Linear }}(1,14)=34.912\right.$, $\left.p<0.016, \eta_{p}^{2}=0.714\right]$ and the null effect of Game in the -Alignment -Feedback condition [ $F_{\text {Linear }}(1,14)=$ 2.825, $p=0.115]$. So, in the absence of concurrent partner feedback, sign alignment reduced sign
complexity. Without either interactive process there was no statistical evidence of a reduction in sign complexity across games.

Receiving concurrent partner feedback was important to sign simplification and symbolization. In the absence of concurrent partner feedback sign alignment reduced sign complexity, but not to the extent of concurrent partner feedback.

## Discussion

The present study investigated the precise role played by two distinct aspects of social interaction to the evolution of effective and efficient human communication systems: behaviour alignment and concurrent partner feedback. By experimentally manipulating the opportunity for behaviour alignment and concurrent partner feedback in a full factorial design, the experiment demonstrated that each process made a distinct contribution to the evolution of shared symbols: sign alignment directly influenced communication success and concurrent partner feedback drove sign simplification and symbolization. See Lister and Fay (in press) for a theoretical model of this process. Together, these complimentary processes explained the interactive evolution of effective and efficient human communication systems.

Our findings provide a solution the symbol grounding problem (Harnad, 1990). Complex iconic signs ground shared meanings. Once grounded, social interaction drives sign simplification and alignment, the mechanisms through which effective and efficient shared symbols arise. This explanation offers a convincing candidate process through which iconic signs evolve into symbols, as originally proposed by Charles Sanders Peirce over 100 years ago.

Other-initiated repairs are a frequent feature of conversation, and similar repair mechanisms are seen across a range of different languages (Dingemanse et al., 2015). Repairs - from a generic 'huh', to specific information requests - signal trouble and correct breakdowns in communication (Schegloff, 2000; Schegloff, Jefferson, \& Sacks, 1977). Other-initiated repairs were a frequent feature of communication in the +Feedback conditions, especially in the early games of the task ( $25.83 \%, 17.92 \%, 13.75 \%, 11.25 \%$, $7.91 \%, 2.91 \%$ of trials at Game 1-6). Yet, there was no evidence that this feedback directly affected communication success (19.16\%-point improvement in communication success from game 1 to 6 with partner feedback, and a 19.95\%-point improvement in communication success from game 1 to 6 without partner feedback (collapsed across the alignment conditions). By contrast, concurrent partner feedback was crucial to sign simplification and symbolization.

Why might other-initiated repairs not directly affect communication success? A simple answer is that people may not be sensitive to problems in communication in the first place. This was examined in a study in which conversation partners, who communicated via text-chat, were swapped with participants engaged in a separate and unrelated conversation (Galantucci \& Roberts, 2014). Participants failed to notice their conversation partner had changed (beyond chance level), despite the incoherent change in topic. This finding suggests that
communication is noisy and error-prone, and that people tend to be insensitive to communication problems. Perhaps our task is too simple to be able to detect the positive influence of other-initiated repairs on communication success. Against this, our experimental paradigm was sensitive to the positive influence of behaviour alignment on communication success.

Our experimental findings demonstrate that behaviour alignment directly influenced communication success. By contrast, there was no statistical evidence that other-initiated repairs directly affected communication success. This pattern of results supports models of dialogue that downplay the role of high-level cognitive processes, and stress the importance of behaviour alignment, via low-level processes such as priming, to successful communication (Garrod \& Pickering, 2004; Pickering \& Garrod, 2004). In the present study, although partner feedback did not directly affect communication success, it proved crucial to sign simplification and symbolization, which improved the smooth and efficient flow of communication (see also Bavelas et al., 2000; Mein et al., 2016).

## Acknowledgments

N.F. acknowledges support by an ARC Discovery grant (no. DP120104237).

## References

Bavelas, J. B., Coates, L., \& Johnson, T. (2000). Listeners as co-narrators. Journal of Personality and Social Psychology, 79(6), 941-952.

Caldwell, C. A., \& Smith, K. (2012). Cultural Evolution and Perpetuation of Arbitrary Communicative Conventions in Experimental Microsocieties. PLoS ONE, 7(8), e43807. https://doi.org/10.1371/journal.pone. 0043807

Christensen, P., Fusaroli, R., \& Tylén, K. (2016). Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization. Cognition, 146, 67-80. https://doi.org/10.1016/j.cognition.2015.09.004

Clark, H. H., \& Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language, 50(1), 62.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., ... others. (2015). Universal principles in the repair of communication problems. PloS One, 10(9), e0136100.

Fay, N., Arbib, M., \& Garrod, S. (2013). How to Bootstrap a Human Communication System. Cognitive Science, 37(7), 1356-1367. https://doi.org/10.1111/cogs. 12048

Fay, N., Ellison, M., \& Garrod, S. (2014). Iconicity: From sign to system in human communication and language. Pragmatics \& Cognition, 22(2), 244-263.

Fay, N., Garrod, S., Roberts, L., \& Swoboda, N. (2010). The interactive evolution of human communication systems. Cognitive Science, 34(3), 351-386.

Fay, N., Lister, C. J., Ellison, T. M., \& GoldinMeadow, S. (2014). Creating a Communication System from Scratch: Gesture Beats Vocalization Hands Down. Frontiers in Psychology, 5:354.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., \& Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. Psychological Science, 23(8), 931-939.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. Cognitive Science, 29(5), 737-767.

Galantucci, B. (2017). Experimental Semiotics. In Oxford Research Encyclopedia of Linguistics.

Galantucci, B., \& Roberts, G. (2014). Do We Notice when Communication Goes Awry? An Investigation of People's Sensitivity to Coherence in Spontaneous Conversation. PLOS ONE, 9(7), e103182. https://doi.org/10.1371/journal.pone. 0103182

Garrod, S., Fay, N., Lee, J., Oberlander, J., \& MacLeod, T. (2007). Foundations of Representation: Where Might Graphical Symbol Systems Come From? Cognitive Science, 31(6), 961-987.

Garrod, S., Fay, N., Rogers, S., Walker, B., \& Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols? Interaction Studies, 11(1), 33-50.

Garrod, S., \& Pickering, M. J. (2004). Why is conversation so easy? Trends in Cognitive Sciences, 8(1), 8-11.

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346.

Healy, P. G. T., Swoboda, N., \& King, J. (2002). A tool for performing and analysing experiments on graphical communication. In X. Faulkner, J. Finlay, \& F. Detienne (Eds.), People and computers XVI: Proceedings of HCl2002: The 16th British HCI Group Annual Conference (pp. 55-68). London: SpringerVerlag.

Healy, P. G. T., Swoboda, N., Umata, I., \& King, J. (2007). Graphical language games: Interactional constraints on representational form. Cognitive Science, 31, 285-309.

Lister, C. J., \& Fay, N. (in press). How to Create a Human Communication System: A Theoretical Model. Interaction Studies.

Mein, C., Fay, N., \& Page, A. C. (2016). Deficits in joint action explain why socially anxious individuals are less well liked. Journal of Behavior Therapy and Experimental Psychiatry, 50, 147-151. https://doi.org/10.1016/j.jbtep.2015.07.001

Moore, D. S., McCabe, G. P., \& Craig, B. A. (1993). Introduction to the Practice of Statistics. New York: W. H. Freeman.

Peirce, C. S. (1931). Collected Papers of Charles Sanders Peirce (Vol. 1-8). Cambridge, MA: Harvard University Press.

Pelli, D. G., Burns, C. W., Farell, B., \& Moore-Page, D. C. (2006). Feature detection and letter identification. Vision Research, 46(28), 4646-4674.

Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-226.

Reitter, D., \& Moore, J. D. (2014). Alignment and task success in spoken dialogue. Journal of Memory and Language, 76, 29-46. https://doi.org/10.1016/j.jml.2014.05.008

Roberts, G., Lewandowski, J., \& Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. Cognition, 141, 52-66.

Schegloff, E. A. (2000). When'others' initiate repair. Applied Linguistics, 21(2), 205-243.

Schegloff, E. A., Jefferson, G., \& Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. Language, 53(2), 361-382. https://doi.org/10.2307/413107

Schouwstra, M., \& de Swart, H. (2014). The semantic origins of word order. Cognition, 131(3), 431-436.

Stolk, A., Verhagen, L., \& Toni, I. (2016). Conceptual Alignment: How Brains Achieve Mutual Understanding. Trends in Cognitive Sciences. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364 661315002867

Tamariz, M. (2017). Experimental Studies on the Cultural Evolution of Language. Annual Review of Linguistics, 3(1), null. https://doi.org/10.1146/annurev-linguistics-011516-033807

Tamariz, M., \& Kirby, S. (2014). Culture: copying, compression, and conventionality. Cognitive Science, 39(1), 171-183.

Theisen, C. A., Oberlander, J., \& Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. Interaction Studies, 11(1), 14-32.

# A Neural Network Model for Taxonomic Responding with Realistic Visual Inputs 

Giorgia Fenoglio (giorgia.fenoglio@edu.unito.it)<br>Dipartimento di Informatica, Corso Svizzera 185, 10149, Torino, Italy<br>Roberto Esposito (roberto.esposito@unito.it)<br>Dipartimento di Informatica, Corso Svizzera 185, 10149, Torino, Italy

Valentina Gliozzi (valentina.gliozzi@unito.it)<br>Center for Logic, Language, Cognition<br>Dipartimento di Informatica, Corso Svizzera 185, 10149, Torino, Italy


#### Abstract

We propose a neural network model that accounts for the emergence of the taxonomic constraint in early word learning. Our proposal is based on Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint and overcomes one of its limitations, namely the fact that it considers artificially built, simplified stimuli. In fact, while in the original model the visual stimuli are random, sparse dot patterns, in our proposed solution they are photographic images from the ImageNet database. In our model the represented objects in the image can be of different size, color, location in the picture, point of view, etc.. We show that, notwithstanding the augmented complexity in the input, the proposed model compares favorably with respect to Mayor and Plunkett (2010)'s model.


## Introduction

A central issue in the current understanding of early lexical acquisition concerns how infants learn the reference of words. Quine (1960) famously raised the point that for every word heard in a given circumstance, there are several possible references: in order to infer the appropriate one, infants have to rule out several possible alternatives. An influential solution to the issue has been proposed by Markman (1989), suggesting that infants rule out inappropriate references by means of three constraints. By the whole object constraint children assume that novel words refer to objects as a whole, rather than to their parts, substance, color, or other properties. By the mutual exclusivity constraint children assume that two labels usually do not refer to the same object. Last, but central to this paper, by the taxonomic constraint children extend words to taxonomically-related objects (at the level of basic categories): when a child hears the word "dog" pronounced by a caregiver while pointing at a specific dog, she generalizes the reference of "dog" to all dogs, not just to the one in front of her.

Here we propose a neural network model that accounts for the emergence of the taxonomic constraint in early word learning, and can process realistic visual stimuli ${ }^{1}$. This is the first step towards the development of a model able to cope with visual and auditory stimuli that are both realistic.

[^324]Our starting point is Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint. The model consists of two self-organizing maps (a visual and an acoustic map) connected with Hebbian connections. The model successfully explains how it is possible to generalize a single word-object association to a whole class of objects. Essentially, this is the result of Hebbian learning creating wordobject associations over a previous conceptual organization of the visual and acoustic space.

Here we want to go beyond one limitation of Mayor and Plunkett (2010)'s model, namely the fact that it considers artificially built, simplified stimuli: in their model the visual stimuli are random, sparse dot patterns, in the style of (Posner, Goldsmith, \& Welton Jr, 1967), whereas the acoustic stimuli are manipulations of acoustic signatures extracted from sounds produced by a speaker, leading to a simplified acoustic input stimulus.

Would the model still work if we considered realistic visual inputs, instead? In order to address this question, we have expanded the original model's visual component making it able to process realistic visual stimuli, that in our case are images taken from the ImageNet dataset. More precisely, we have added to the visual component of Mayor and Plunkett (2010) an InceptionV3 deep network (Szegedy et al., 2015) which is at the state of the art in the image classification task. The deep network processes the visual scene in the image, builds a representation for it, and feeds the representation to the visual self-organizing map.

In order for the whole model to work, these representations need to contain a description of the main object of the visual scene, independent from the context. Understanding the nature of the image representations built at the various levels of the network is indeed one of the main points of debate in deep neural networks (Zeiler \& Fergus, 2014; Zhou, Khosla, Lapedriza, Oliva, \& Torralba, 2014; Agrawal, Girshick, \& Malik, 2014). In order to assess whether the InceptionV3 deep network feeds into the visual self-organizing map meaningful object representations, we performed several clustering experiments. These experiments investigated whether representations deriving from images of objects can be clustered
together by off-the-shelf clustering algorithms. They showed that the representations provided by InceptionV3 are reasonably well organized. A further test investigated whether the visual self-organizing map could organize the representations received from the deep neural networks in a topologically satisfactory manner.

We then tested the whole model to see if it still exhibits a taxonomic responding, generalizing learned word-object associations to the whole category. Results show that our model, despite starting from more realistic visual stimuli, does replicate Mayor and Plunkett (2010)'s success on taxonomic responding when few joint word-object associations are considered.

## Mayor and Plunkett (2010)'s model

Mayor and Plunkett (2010) neurocomputational model of taxonomic constraint (Figure 1) is based on two Self-Organizing Maps (SOMs): a visual map and an acoustic map, representing the primary visual cortex and the primary auditory cortex respectively.

The stimuli presented to the two maps are artificially built: the visual stimuli are random dot patterns, whereas the auditory stimuli are extracted from the acoustic signatures of uttered words; the acoustic signatures are manipulated in order to create simpler inputs.

Learning is a two-phase process. First, the two maps are independently trained to learn to categorize the visual and the acoustic stimuli. This first learning phase is preliminary to word learning, and unsupervised. The two maps are trained using the standard learning algorithm for self-organizing maps. In short, a stimulus $x$ is presented to each neuron of the map, and the Best-Matching Unit (BMU) is selected: this is the unit $i$ whose weight vector $w_{i}$ is closest to the stimulus $x$ (i.e. $i=\arg \min _{j}\left\|x-w_{j}\right\|$ ).

The weights of the best matching unit and of its surrounding units are updated in order to maximize the chances that in the future the same unit (or the surrounding units) will be selected as the best matching unit for the same stimulus or for similar stimuli. At iteration $n+1$, the weights for neuron $j$ are updated as follows:

$$
\begin{equation*}
w_{j}(n+1)=w_{j}(n)+\eta(n) h_{i, j}(n)\left(x-w_{j}(n)\right) \tag{1}
\end{equation*}
$$

where $\eta$ is the learning rate, and $h_{i, j}$ is the neighborhood function between $i$ and $j h_{i, j}(n)$ is defined as $h_{i, j}(n)=$ $\exp \left(-d_{i, j}^{2} / 2 \sigma(n)^{2}\right)$, where $d_{i, j}$ is the distance between $i$ and $j$ on the map's grid, and $\sigma(n)$ is the width of the gaussian.

After a while, the two maps learn to adequately represent the stimuli of their training set in a topologically significant way: close units respond similarly to similar stimuli. The neural activation $a_{j}$ of a neuron $j$ in response to a stimulus $x$ is defined as: $a_{j}=e^{-\frac{q_{j}}{\tau}}$, where $q_{j}$ is the quantization error
(i.e., the distance between the input vector $x$ and $j^{\prime}$ s weight vector: $\left.q_{j}=\left\|x-w_{j}(n)\right\|\right)$ ), and $\tau$ is a normalization constant.

Once the visual and acoustic maps have stabilized into a topological organization, proper word learning can start. This is the Hebbian Learning phase, in which the two kinds of stimuli are simultaneously presented to the model. For each joint presentation of a visual and acoustic stimulus, the synapses between the two maps are strengthened. In particular, for each neuron $v$ on the visual map and neuron $p$ on the acoustic map, the Hebbian connection $u_{v, p}$ is strengthened proportionally to the resulting neural activations $a_{v}$ and $a_{p}$, as follows:

$$
\begin{equation*}
u_{v, p}(n+1)=u_{v, p}(n)+1-e^{-\lambda a_{v} a_{p}} \tag{2}
\end{equation*}
$$

where $\lambda$ is the Hebbian training learning rate.
A single Hebbian learning event, combined with the previously acquired categorization capabilities of the visual and acoustic SOMs, allows the model to generalize the association to other stimuli belonging to the same category.

Once training is complete, the model is tested for its ability of comprehension and production. Comprehension is assessed by considering what visual category is retrieved when a word is presented to the auditory map and activation is propagated via Hebbian connections. Production is assessed by considering what word is produced by the auditory map when a visual stimulus is presented to the visual map, and activation is propagated through Hebbian connections.

The ability of the model to extend the learned word-object associations to other words and objects belonging to the same category is measured by the Taxonomic Factor which is the percentage of correct word-object associations generated by the model. Results show that when the SOMs are adequately trained the Taxonomic Factor reaches $80 \%$ after a single Hebbian learning trial.

One of the limitations of Mayor and Plunkett (2010)'s original model is that it uses artificially built input stimuli that are much simpler than what would derive from realistic contexts. Here we address this limitation, for what concerns the visual module, by introducing deep convolutional neural networks as shown in the next sections.

## Deep Convolutional Neural Networks

In the last few years research on deep networks contributed to reach human (sometimes super-human) performances on several difficult tasks (Hinton et al., 2012; Li \& Wu, 2015; Socher, Bauer, Manning, \& Ng, 2013; Yue-Hei Ng et al., 2015). In particular, in 2011 a deep convolutional model achieved for the first time super-human performances in a visual pattern recognition task and, in the following year, the AlexNet Convolutional Neural Network (CNN) model won the ImageNet competition by a significant margin (Krizhevsky, Sutskever, \& Hinton, 2012) over traditional competitors. These successes contributed to a growing interest in deep networks and today deep-network-based models are at the forefront of research in many different areas and are
setting performance records in tasks of interest for the cognitive sciences community such as image (e.g., (Russakovsky et al., 2015)) and speech recognition (e.g., (Xiong et al., 2016)).

Despite unheard performances achieved in many different tasks, deep models present important shortcomings that are far from being completely addressed. The most important problem from the point of view of the forthcoming discussion is the difficulty they present for what concerns the understanding of their internal working. Consequently, recent research investigated ways to make sense of the contents of the network providing interesting insights. For instance, Zeiler and Fergus (2014) use "deconvolution networks" to visualize the patterns that causes the activation of nodes in each layer; in Zhou et al. (2014) scenes are iteratively simplified or occluded to investigate which image patches and which objects contribute to the activation of nodes in a given layer; in (Agrawal et al., 2014) the authors investigate the presence of grand-mother-cells and of distributed representations in deep networks.

While the understanding of the representations built by these networks is still scattered and incomplete, some of the insights seem to be well supported. An important one concerns the hierarchical organization of the features: low-level (coarser) features are nearest to the network input, while higher-level (more abstract) features are nearest to the output (for an idea of the kind of features extracted at the different levels see for instance (Zeiler \& Fergus, 2014)). Interestingly this organization mirrors a well known characteristic of the representations in the primate inferior temporal (IT) cortex, and hence it hints at a possible cognitive justification of this computational model. To this regard it is interesting to mention that recent research investigated the connection between the representations built by several computational models and the representations in the IT cortex and found that deep neural networks are among the best models (Serre, 2016; Kriegeskorte, 2015). For instance, in (Khaligh-Razavi \& Kriegeskorte, 2014), the authors investigate a wide range of computational models and suggest that deep CNNs are, not only the best performing in term of accuracy, but also the best at explaining the IT representation (albeit still in an incomplete way).

Given the great accuracy they achieve and the possible cognitive plausibility of the CNNs, we have chosen to use these particular models as the visual component of the word-object association model we propose in the next section.

## Proposed model

In order to solve the problem of the lack of realism in the visual stimuli in the Mayor and Plunkett (2010) model, we propose to replace the input of the visual SOM with a representation built by a CNN as shown in Figure 2.

The long term objective of this research is to find a cognitively plausible model able to reproduce the word-object association abilities observed in infants using realistic image and audio stimuli. In this paper we keep a simplified auditory input and focus instead on providing a visual module capable


Figure 2: The proposed model: the visual component contains a deep convolutional network (InceptionV3) in order to process realistic images. The representation built by the deep neural network is then fed into the visual SOM. The acoustic component, on the contrary, only contains an acoustic SOM, as in Mayor \& Plunkett (2010)'s model, and can only process simplified acoustic stimuli.
of handling realistic images. In fact, in this proposal the auditory input is a mere placeholder that does not provide any real processing ability.

In practice, we shall assume that an oracle provides the auditory SOM with a perfect representation of the auditory stimulus, or label, in the form of a binary vector. The vector contains a 1 in position $i$ if the utterance provided to the auditory module corresponds to the $i$-th label; it contains a 0 otherwise.

The visual module shall, on the contrary, be able to cope with realistic images and, while we still assume that each image contains a main object corresponding to the concept to be learned, we pose no additional constraints. Images for a given concept can, for instance, be of different size, color, location in the picture, point of view, etc.. For instance, the "dogs" concept may be represented by images of dogs of different size, color, breed and be portrayed in different contexts, under different illuminations and poses.

The visual module is the concatenation of the InceptionV3 network and the SOM network we already introduced. InceptionV3 is a stack of Inception Modules, which parallelize and combine several convolution and pooling operations providing a richer output while still maintaining a small number of parameters. At the end of the stack of inception modules the
model contains a pooling layer of length 2048 which is fully connected with a shallow feed-forward neural network which is then used to classify the input image.

In the Inception architecture a representation of the input is propagated through the layers up to the top of the network where it is used to train the classifier. A question worth investigation is if and where a good representation of the concept in the input image is created by the deep network. In this paper we work under the assumption that such representation exists and argue that it has to be found in the last pooling layer (just before the fully connected neural classifier). Based on this assumption, we propose to use the vector containing the value of the 2048 neurons in that layer as the representation of the stimuli for the visual SOM. To verify the validity and the consequences of this assumption, we performed two sets of experiments: in the "Representation Quality" sub-section of the Experiments section we investigate the nature of the proposed representation, while in the subsequent section (Word Learning) we investigate the quality of the complete model. In our simulations we used the pre-trained Inception network provided by the TensorFlow library ${ }^{2}$.

The complete system is trained as outlined in the "Mayor and Plunkett (2010)'s model" section. In summary, given the representation from the CNN and the simplified auditory input, the two SOMs (composed by $20 \times 30$ neurons each) are trained to cluster together similar representations using the standard SOM training algorithm. In our tests, the two SOMs attain their best topological organization of the objects in the training set after 60 epochs (the learning rate is set to 0.3 and decreases linearly at each epoch). Afterwards, the association between the visual and the auditory input is created using Hebbian connections between the two maps: two stimuli belonging to the same category are presented together to the model, the visual stimulus is processed to extract its representation and presented along with the auditory stimulus to the corresponding SOMs. Finally the SOMs activations are used to update the Hebbian connections using the update rule in Formula 2.

To better cope with the variability in the input representation, we introduce two variations to the Hebbian training (with respect to the procedure outlined in (Mayor \& Plunkett, 2010)): $i$ ) we allow the network to learn from an increasing number of stimuli pairs (in the original paper a single pair of stimuli is presented to the network), this allows us to study how performances increase as the number of presentations grows; $i i$ ) we suppress the activation of a neuron in a SOM if its activation value is below 0.6.

## Experiments

In the following two sub-sections we investigate two important facets of the proposed model. In particular, in the "Representation Quality" Section we show that the representation found in the last pooling layer of the InceptionV3 network al-

[^325]lows one to cluster the input images into groups that correlate well with the classes assigned with the images themselves. This is arguably an evidence that such a representation can be usefully exploited as the input of the SOMs. In the "Word Learning" Section we focus on the complete model, replicate part of the experiments in (Mayor \& Plunkett, 2010), and compare our results with those reported in that paper.

All the experiments have been performed on two datasets. A first dataset is composed by 10 classes associated with 100 stimuli each, for an overall 1.000 stimuli. A second dataset contains 100 classes associated with 100 stimuli each, for an overall of 10.000 stimuli. Since the results for the two datasets are very similar, for the sake of readability we focus on the smaller dataset and refer to (Fenoglio, 2016) for the details of the experiments on the larger dataset. The code for the complete model along with the datasets used can be found at https://github.com/ ml-unito/nNsTaxonomicResponding.

## Representation Quality

In order to assess the quality of the representation found in the last pooling layer of the InceptionV3 model, we investigate how well these representations can be clustered together. For each image we extract the representation found in the last pooling layer of the deep network, we then cluster the resulting representations using a K-means and an agglomerative algorithm. For both algorithms the number of clusters is set to 10 . The clustering experiments have been conducted using the scikit-learn python library ${ }^{3}$.

Figures 3 and 4 report results for K-means clustering. Analogous results hold for agglomerative clustering. In particular, Figure 3 reports, for each class, a bar showing how the class objects are partitioned among clusters; Figure 4 reports, for each cluster, a bar showing the distribution of the classes within it. We then investigate the topological organization provided by the visual SOM out of the representations created by the deep model. We report in Figure 5 a representation of the topology found by the visual SOM after 60 learning epochs.

Discussion The experiments show that the two clustering algorithms are able to find good, albeit not perfect, partitions for the representations. In particular, Figure 3 shows that the objects in 7 out of 10 classes are mostly assigned homogeneously to a single cluster: in two of the remaining cases the objects are almost all distributed among two classes, while in a single case (and only for the k-means clustering algorithm) the objects are distributed on three clusters. Figure 4 shows a similar picture, but from the point of view of the clusters: in almost all cases ( 8 out of 10) we have clusters which are almost pure. The remaining two clusters conglomerate objects from different classes acting almost as folders where all uncertain objects are put.

Overall, it seems that the clustering algorithms do find a way to partition the representations of the objects into co-

[^326]

Figure 3: Per class distribution of objects into clusters (Kmeans clustering). Colors represents different clusters. Blue is used to represent the cluster containing the majority of the objects of a given class; orange, yellow, and green are used to represent the second, third, and fourth most represented clusters.


Figure 4: Distribution of classes among clusters (K-means clustering).
herent clusters. This is consistent with what happens for the topological organization that the SOMs create for the representations provided by the deep networks. Figure 5 shows that, with the exception of few cases, the visual SOM is able to group all related objects into nearby spaces.

## Word Learning

In order to evaluate the performance of our model in the task of word learning, we calculate the Taxonomic Factor of the model as defined in (Mayor \& Plunkett, 2010). We do so by testing the model for its production skills: for each class, 100 images are presented to the visual module, the activation is propagated through the deep neural network, then fed into the visual SOM. The activation of the visual SOM's best matching unit is propagated through the Hebbian connections up to the acoustic SOM. At the end of the process the resulting most active unit on the acoustic map is identified. It will be considered correct if it belongs to the area of the acoustic map


Figure 5: SOM clustering of the visual stimuli representations
associated to that word ${ }^{4}$. The percentage of correct words produced by the model when tested through all the classes is the Taxonomic Factor.

We have performed a number of experiments where we varied the number of presentations per class used to update the Hebbian connections. Specifically we let the number of presentations vary from 1 to 15 . For each experiment we repeated the test over 1.000 different training sets (we kept fixed the SOM and let vary the images presented to the Hebbian learning module) and report the average taxonomic factor over an independent test set composed by additional 1000 images (100 images per class). Results are shown in Figure 6.


Figure 6: Taxonomic factor of the model, using an increasing number of pairs of stimuli per class during the training of the Hebbian Connection (on the x -axis).

Discussion The experiments show that the Taxonomic Factor steadily grows as more word-object associations are presented and reaches an accuracy above $80 \%$ (which is comparable with results in Mayor and Plunkett (2010)) at the fourth joint presentation.

[^327]
## Conclusions

In this paper we have proposed an extension of the Mayor and Plunkett (2010) model for taxonomic responding. We have addressed the issue of adding realism to the visual stimuli. As a difference with respect to the original model in which these inputs were random dot patterns, the model can now deal with realistic images as those in the ImageNet dataset. This is possible thanks to the insertion of a deep convolutional neural network in the visual component of the model. Notwithstanding the higher complexity of the stimuli considered, our model exhibits taxonomic responding with performances comparable to the original one.

In our future work we will address the issue of making the acoustic module work with realistic stimuli. It can be interesting to explore whether a deep neural network for acoustic processing, as for instance the one proposed in (Xiong et al., 2016), could be nested into the acoustic part of the model in a way similar to what we already did for the visual component.

We will also explore whether the model proposed here can be used to provide a mechanistic account of the whole object constraint proposed by Markman (1989) by which a word is associated to the whole object instead of anyone of its properties. We conjecture that a model as the one proposed, with the deep component that extracts a representation of an object out of a more complex visual scene, can be adequate to the purpose: the whole object constraint may naturally emerge from the association of the word to the object's representation formed by the deep network. Important to this regard is the current discussion about the nature of the object representation built by deep networks (Ullman, Assif, Fetaya, \& Harari, 2016; Tang \& Kreiman, 2017).

## References

Agrawal, P., Girshick, R., \& Malik, J. (2014). Analyzing the performance of multilayer neural networks for object recognition. In European conference on computer vision (pp. 329-344).
Fenoglio, G. (2016). A computational model for word learning on real world data through Deep Neural Networks. Unpublished master's thesis, Turin University. (https:// github.com/gfbfenoglio/NNsTaxonomicResponding/ blob/master/Documents/MasterThesis.pdf, chapters $6,8)$
Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., .. Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. Signal Processing Magazine.
Khaligh-Razavi, S.-M., \& Kriegeskorte, N. (2014, November 6). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLOS Computational Biology, 10(11).
Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual Review of Vision Science, 1, 417-446.

Krizhevsky, A., Sutskever, I., \& Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
Li, X., \& Wu, X. (2015). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In Acoustics, speech and signal processing (icassp), 2015 ieee international conference on (pp. 4520-4524).
Markman, E. M. (1989). Categorization and naming in children: Problems of induction. Mit Press.
Mayor, J., \& Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. Psychological review, 117 1, 1-31.
Posner, M., Goldsmith, R., \& Welton Jr, K. (1967). Perceived distance and the classification of distorted patterns. Journal of Experimental Psychology, 73(1), 28-38.
Quine, W. V. O. (1960). Word and object. Cambridge,Mass.: MIT Press.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.
Serre, T. (2016). Models of visual categorization. Wiley Interdisciplinary Reviews: Cognitive Science, 7(3), 197213.

Socher, R., Bauer, J., Manning, C. D., \& Ng, A. Y. (2013). Parsing with compositional vector grammars. In In proceedings of the ACL conference.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 1-9).
Tang, H., \& Kreiman, G. (2017). Recognition of occluded objects. In Computational and Cognitive Neuroscience of Vision. (ed Zhao, Q). Singapore: Springer-Verlag, 14, 5777.

Ullman, S., Assif, L., Fetaya, E., \& Harari, D. (2016). Atoms of recognition in human and computer vision. Proceedings of the National Academy of Sciences, 113(10), 2744-2749.
Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2016). The microsoft 2016 conversational speech recognition system. CoRR, abs/1609.03528.
Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., \& Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 4694-4702).
Zeiler, M. D., \& Fergus, R. (2014). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833).
Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., \& Torralba, A. (2014). Object detectors emerge in deep scene cnns. CoRR, abs/1412.6856.

# How Many People Know? Representing the Distribution of Knowledge 

## Stanka A. Fitneva (Fitneva@queensu.ca) <br> partment of Psychology, Queen's University

Kingston, ON K7L 3N6, Canada

Abstract
The representation of the distribution of knowledge guides research aimed to explore adults' and 4-year-olds representation of the distribution of common (conventional and procedural) knowledge and expert knowledge associated
with five occupations in their community. In addition, we examined estimates of occupation-related everyday (nonexpert) knowledge. Both groups estimated that common
knowledge is more widely held than expert knowledge wit knowledge is more widely held than expert knowledge, with
everyday knowledge in between. For adults, but not children, everyday knowledge in between. For adults, but not children,
the distribution of expert knowledge was correlated with estimates of the proportion of people in each occupation.
Keywords:
development developmen

## Introduction

People act competently and adaptively in their physical and cial surroundings. Yet, their understanding of the physica his paradox likely rests in the social behavior: we rely on each other to fill the holes in our understanding (Keil, 2003). Understanding who knows what and how knowledge is structured is thus a critical aspect of human social cognition. Equally important, though less wel in the population. Distinguishing widely distributed from arrowly distributed knowledge could affect both the selectivity of our interpersonal interactions and the structure of our social networks.
Knowledge is clearly unevenly distributed. Some nowledge, such as conventional knowledge of common object labels and functions, is shared by all members of a nowledge, e.g, that grass is green, can be expected to be known by all in virtue of shared bodily experience. In contrast to this "common knowledge," other knowledge is privy to only groups of people within the community. The ivision of labor that characterizes most moder communities leads to conce different people
Surprisingly, the question of the relative spread of
different kinds of knowledge has received little attention Most of the literature on expertise has focused on explaining he attainment of expertise in various domains of activity (Feltovich, Ford, \& Hoffman, 1997) and on understanding of the clustering of knowledge in different kinds of experts Rozenblit, 2008). For instance, preschoolers appear to recognize that expertise is topic-specific and that being an
expert in one domain does not entail being an expert in other domains (Koenig \& Jaswal, 2011). Furthermore, by age 5, children appear to be able to link their skeletal understanding of knowledge domains (psychology, physics,
biology) with how knowledge is clustered in individual minds. That is, they can make an inference from what a person knows to what other things the person is likely to know (Keil et al., 2008).
With respect to the spread of knowledge, previous evidence exclusively for the existence of understanding of non-overlaps in knowledge. For instance, Lutz and Keil (2002) presented 3- to 5 -year-old children with a list of items representing the expertise of doctors and car
mechanics. Children were asked questions like "Who would mechanics. Children were asked questions like "Who would
know more about how to fix a broken arm?" While children identified above chance the relevant expert, this suggests sensitivity to non-overlaps in knowledge and that not everyone knows a given item. The data do not speak to the question of whether children recognize that expert knowledge is relatively narrowly distributed in the population
children' understanding of differences in havewledge For instance, a number of studies address children's understanding of the difference between child and adult knowledge (Aguiar, Stoess, \& Taylor, 2012; Fitneva, 2010; Fitneva, Ho, \& Hatayama, 2016; Taylor, Cartwright, \&
Bowden, 1991; VanderBorght \& Jaswal, 2009). Bowden, 1991; VanderBorght \& Jaswal, 2009).
Nevertheless, these studies only reveal that children identify Nevertheless, these studies only reveal that children identify
non-overlaps of knowledge between social groups. They don't address children's representation of the size of social groups and therefore fail to capture children's representation of the spread of different kinds of knowledge. Perhaps the only study that allows such an inference, Burton and mitchente knowledge to the self and denied its possession by a range of adults and children.
The question of children's representation of the distribution of knowledge has been also examined for conventional knowledge such as word meaning (Graham, Stock, \& Henderson, 2006; Henderson, Weighall, Brown, \& Gaskell, 2013). The dominant method here involves
examining whether children extend novel conventional knowledge (e.g., of an object label or a game rule) to a new person. Although the young children in these studies appear to appropriately extend conventional knowledge to others but restrict idiosyncratic knowledge such as desires to individuals, it is not clear that they see conventional knowledge as widely distributed. The reason is that in these
studies, new individuals are generally not introduced as randomly sampled from the population.
Recent evidence does suggest, however, that children are examples, Cimpian and Scott (2012) tested the belge. For to 7 year olds on how many people would know generic and non-generic facts. Children associated generic facts with "many" and non-generic facts with "few" people. Presumably this performance implies that children assume that generic information concerns how the world works and see adults as expe
2012). The
he question of how knowledge is distributed in the population highlights the social embeddedness of expertise.
Experts function within a community. The division of labor - in many cases the motivator for the development of expertise - would not be feasible and workable if it were not services, and ideas. Communicating and interacting with various experts both rests upon and develops relevan knowledge. This knowledge is neither conventional nor shared in the same way perception of color is. In contrast to conventional and common experiential knowledge, it can show considerable individual differences based on experience with the problem domains and or access
To sum up the goal of this study was to examine yo children's understanding of the distribution of expert, everyday, and conventional knowledge. In particular, we had three questions: 1) Do children associate exper knowledge with a smaller proportion of the population than common knowledge? 2) Do they associate it with a smaller Is the perceived distribution of expert knowledge related to the perception of number of experts in the community?
Even though past research has documented that by age four children understand that the knowledge of adult experts is not co-extensive (Lutz \& Keil, 2002), they may nevertheless associate expert knowledge with large portions more narrowly distributed than common and everyday knowledge. This would be consistent with children seeing adults as omniscient (Piaget, 1959), or at least people capable of exceptional performance in more than one domain. Alternatively, we expected that even 4 -year-olds might associate expert knowledge with a smaller portion of is because even very young Canadian children have first hand contact with experts (e.g., doctors) and observe the exchange of goods, services, and ideas ensuing from the division of labor.
There are a number of ways in which people can develop nderstanding of the spread of expertise. One of them is occupations. If this is the case, participants' estimates of the proportions of experts among adults would correlate with their estimates of the proportion of people with expert
knowledge. Thus, the study included questions asking participants to estimate the prevalence of differe ccupations in their community
mong the youngest to demonstrear-old children who ar overlaps in the knowledge of adults (Lutz \& Keil, 2002) At this age, children are also sensitive to relative magnitude.

## Method

## Participants

Thirty-six 4 -year-old children and 18 adults participated in the study. The children lived in the mid-size urban
community of Kingston Ontario and the adults were community of Kingston, Ontario and the aduits were
students at Queen's University in the same city. Six children were excluded due to not completing the study (2), self professed silly attitude (1) and clear pattern in responding (i.e., going up / down the scale, 3). Thus the final sample included 30 children (average age 54 months, range 48-60; 19 girls, 11 boys).

## Materials

We asked participants to indicate their perception of the distribution of expert knowledge, everyday knowledge in he same domains, and common knowledge. The expert Iledge pertained to five occupations: farmers, builder pilots, car mechanics, and doctors. These occupation vary in frequency in the community (more builders, car mechanics and doctors relative to farmers and pilots). An example of an expert knowledge question is "How many corresponding everyday knowledge question was "How many grown-ups know how to take their temperature?" The five common knowledge items referred to conventional knowledge, e.g., How many grown-ups know many grown-ups know how to lock their front door?" To examine whether the reported spread of expert knowledge corresponded to participants' perceived number of experts in the community, we also asked an occupation-focused question in each domain, e.g., "How many grown-ups in Kingston are doctors?"' Participants were also presented individual characteristics and behaviors, e.g., "How many grown ups in Kingston go to work / have pets?" These property questions aimed to further prompt thinking about wider and narrower sets of the population
Children answered the questions on a 5 -point scale with a slider. The five points depicted with pie charts $0 \%, 25 \%$,
$50 \%, 75 \%$, and $100 \%$ of the population. Note that this scale $50 \%, 75 \%$, and $100 \%$ of the population. Note that this scale
was not designed for recording realistic estimates of prevalence. This would have required a logarithmic scale and we were not aware of evidence of successful use of such a scale with children.
For the adults, the questions were presented on a piece of paper. The instructions and a figure representing a $0-100$
scale with the pie charts used with the slider appeared on op of the page.

## Procedure

Adults Adult participants answered the questions by writing down their answers next to each question. They were free to 100 range. Adults received course credit for participating.

Children The experiment began by explaining to children hat they will be asked questions about grown-ups. Afte that, the experimenter informed them that they were to
answer the questions by moving the slider to the pie chart hat showed the relevant proportion of grown-ups who know. As a warm-up, children were asked to position the slider in the all, none, and half positions. Children were also asked "How many grown ups in Kingston are shorter/talle han you?"' to provide practice answering with the slider
Children were asked the experimental questions in the
same random order, with the property questions interspersed among them. Although children were free to position the slider anywhere on the $0-100$ scale, they used the five points, consistent with the directions they received.
Subsequently, children were asked whether or not thei arents knew the items, e.g., Do mom and dad know how o fix a broken arm?" These questions aimed to provide an information in their homes. The study included several additional questions the data from which have not yet bee nalyzed. These questions were presented later and do not affect the current results. Parents also answered question regarding their child's familiarity with a large set occupations.

## Results

Figures 1 and 2 show respectively adults' and children's builder, car mechanic, doctor, farmer, and pilot, and relate expert and everyday knowledge. In addition, they show the groups' estimates of the prevalence of common knowledge.
The research questions identify two key comparisons in the data: expert - common knowledge and expert everyday knowledge. In addition, we examined the proportion of people in the five occupations and the proportion of people in the five occupations and the
distribution of expert knowledge related to these occupations in the population. Thus, the analytical approach ncluded a combination of targeted t -tests and analyses of ariance and correlation

## Adults

As Figure 1 suggests, on average, adults associated expert nowledge with a significantly smaller proportion of the than common knowledge ( $M_{\text {eppert }}=9.93 ; M_{\text {comm }}$ $=92.16 ; t(17)=36.36, p<.001)$.


Figure 1. Adults' estimates of the proportion of the population in five occupations and of the distribution of related expert and everyday knowledge. Adults' appears on the right. Error bars represent $\pm 1 \mathrm{SE}$


Figure 2. Children's estimates of the proportion of the population in five occupations and of the distribution of related expert and everyday knowledge. Children's estimate of the distribution of common knowledge appears on the right. Error bars represent $\pm 1$ SE.

We conducted an area (5) x question (occupationfocused, expert knowledge, everyday knowledge) repeated measures ANOVA to assess the differences in adults assessments of the prevalence of the five occupations, expert knowledge, and everyday knowledge. Both main of area $(F(4,68)=42.7, p<.01)$ reflected that some occupations and related knowledge were perceived as more common in the community than others. Of key interest, was the effect of question, $F(2,34)=138.2, p<.01$. As Figure 1 suggests, there was a significant difference in the estimates of the distribution of expert and everyday
knowledge $(M)=9.93 ; M$ $<.001)$. Furthermore, there was no difference in participants' responses to the occupation and expert
knowledge questions, $t(17)=1.08, p=0.295$. Given the significant interaction effect between area and question $(F(8,136)=47.7, p<.01)$, we conducted two follow-up
analyses. First, we examined the difference between expert and everyday knowledge items in each area. Although always in the expected direction, this difference was significant in three of the five areas (the exception being farmer and pilot).

Second, as we were interested in the relationship between participants' perceptions of the proportion of we calculated the correlation between these variables (rathe than their difference). The correlations ranged from .45 to .89 ( $p$ 's $\geq .06$ ) suggesting an overall significant relation between these variables.
In sum, adults recognized that expert knowledge is less prevalent than common knowledge. Furthermore, their linked to their beliefs about the proportion of people in each occupation.

## Children

As Figure 2 suggests, 4 -year-olds associated the common knowledge items on average with $79 \%$ of the population, which was significantly larger than their estimate of the prevalence of expert knowledge $53 \%, t(29)=5.1, p<0.01$. (professional, expert knowledge, everyday knowledge) repeated measures ANOVA on children's responses to the questions about the distribution of expert and everyday knowledge and occupations. The analysis only showed a significant effect of question type, $F(2,58)=6.2, p<.01$. Children associated expert knowledge with a significantly smaller proportion of the population than everyday
knowledge $\left(M_{\text {evpert }}=53 ; M_{\text {evervay }}=67 ; t(29)=3.61, p<\right.$ .01 ). The difference in children's estimates of the number of professionals in the population and the distribution of expert
knowledge was not significant $t(29)=135, p=0.19$. The knowledge was not significant, $t(29)=1.35, p=0.19$. The correlations between children's answers to these two questions in the individual areas ranged between .3 and .45 to reach significance. Interestingly, chil
people in the five occup was higher than their estimate of the proportion of people with related expert knowledge ( $53 \%$ ) but lower than thei estimate of the distribution of everyday knowledge ( $67 \%$ ). The next analysis examined 4 -yea--old's responses to the answers were averaged across area. Common knowledge was attributed to parents on average $90 \%$, significantly more often than either expert or everyday knowledge (both $t \mathrm{~s}<.01)$. Everyday knowledge was more likely to be attributed to parents than expert knowledge ( $M_{\text {expert }}=35$; $M_{\text {everpdy }}=64,(29)=4.86, p<.0$ ). . except car mechanic.

## Discussion

The present findings enrich our understanding of children's and adults' representation of the spread of different kinds of knowledge. Adults showed clear differentiation between expert knowledge and related everyday knowledge as well Furthermore, their estimates of the distribution of expert knowledge closely corresponded to their estimates of the frequency of different occupations. Children also indicated that expert knowledge is less widely distributed than everyday knowledge and common knowledge. Past research has revealed that children recognize that different adult The current study extends these findings to demonstrate the both children and adults see differences in the spread of different kinds of knowledge.
It is important to note that children's responses in the present study are unlikely to be affected by generic language that distinguishes knowledge that most people have from idiosyncratic knowledge of individuals. A widely and narrowly known novel facts (Cimpian \& Scott, 2012). Generic language did not distinguish the stimuli in the different conditions in present study. Thus, 4 -year-olds not only judge widely and narrowly held knowledge based on linguistic cues but have built representations of how knowledge in their environment varies in its spread.
How do people arrive at these representations, especially
with regards to expert knowledge? One possible route to representing the spread of expert knowledge is through considering the frequency of different occupations. Indeed, there was a clear relationship between adults' estimates of the proportion of people winh he target occupations and proportion of people with occupation-related knowledge. However, no such relationship was evident in the of different occupations were in-between the spread of expert and everyday knowledge. This finding suggests that children may not arrive at a representation of the distribution of knowledge considering the frequency of experts. It is possible that children's estimates of the distribution of expert knowledge and people in relate occupation questions focused on social actors while the expert knowledge questions focused on mental states associated with activities. For young children, tracking activities may be easier (given that their estimates were lower and thus more realistic) than the social agents associated with those activities.
Another route children can take to developing nnderstanding of whether something is widely or narrowny
known is through observations of adults in the family. Indeed, the analyses revealed that children crisply differentiated expert, everyday, and common knowledge when asked whether their parents have that knowledge. Children appear to believe that the number of people in
$50 \%$. This conflicts with a number of assumptions adult make about expertise, e.g., that a person does not have a pportunity to develop professional expertise in many areas extend expertise to other domains (Taylor et al., 1991). It may not be warranted, however, to make much of the absolute value of these numbers. First, children may have very well been providing relative answers (e.g., "many" vs. "few") and, second, the scale was not conducive to capturing realistic estimates of the small proportion of explore in future research whether children can work with ogarithmic scale.
On the flip side, even though high, children's estimate of he number of people in different occupations suggests tha hey realize that not everyone has the same profession (only half do!!). In other words, although they may consider adul to be well rounded, they do not consider them omniscient.

Clearly, the present conclusions are limited by th cupations that were represented and the associated item chosen for the study. As the analyses suggested, the magnitude of the difference between expert and everyda nowledge varied substantially across areas. The domain ere intentionally chosen to vary in the representation of he different occupations in the community. This could have farmers could have led to floor effects in the answers to both knowledge questions. It is also possible that the variability by area is due to the particular items chosen for the study Nevertheless, even though the effect sizes varied across occupation for both age groups, their direction was | consistent. |
| :--- |
| In conc |

conclusion, the present study is one of the first to provide clear evidence pertaining to adults' and children' eliefs about the relative spread of common, expert, and恠 been what everyone in their community is likely and expert knowledge, i.e., knowledge obtaine complement our current understanding of people' epresentation of the clustering of knowledge (Danovitch \& Keil, 2004; Keil et al., 2008) and help build comprehensive picture of the social landscape in people's minds which supports adaptive behavior in the face of ncomplete knowledge

## Acknowledgments

Thanks to Megan Loma who assisted with designing the and and stimuli and to Megan Loma, Katherine Lillico, Thanks also to the families who supported the research.

## References

guiar, N. R., Stoess, C. J., \& Taylor, M. (2012). The development of children's ability to fill the gaps in the 3(4), 1368-1381.

Burton, S., \& Mitchell, P. (2003). Judging who knows best about yourself: Developmental change in citing the self , -426-443.
Cimpian, A., \& Scott, R. M. (2012). Children expect generic
knowledge to be widely shared. Cognition, 123(3), 419433.

Danovitch, J., \& Keil, F. C. (2004). Should you ask a fisherman or a biologist?: Developmental shifts in ways of clustering knowledge. Child Development, 75(3), Feltovich $P$
Feltovich, P. J., Ford, K. M., \& Hoffman, R. R. (Eds.).
(1997). Expertise in context. Boston: MIT Press Retrieved from https://mitpress.mit.edu/books/expertisecontext
tneva, S. A. (2010). Children's representation of child and adult knowledge. Journal of Cognition and Development, II(3), 458-484.
(2016). Japanese knowledge: A case for developmental equifinality? PLOS ONE, Il(9), e0163018.
Graham, S. A., Stock, H., \& Henderson, A. M. E. (2006). Nineteen-Month-Olds' Understanding of the Infancy, 9(3), 341-350. Henderson, L Weighall
Henderson, L., Weighall, A., Brown, H., \& Gaskell, G.
(2013). Online lexical competition recognition and word learning in children and adults. Child Development, 84(5), 1668-1685. Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. Trends in Cognitive Sciences, $7(8), 368-375$.
Keil, F. C., Stein, C., Webb, L. Billings, V D.
Keil, F. C., Stein, C., Webb, L., Billings, V. D., \&
Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. Cognitive Science, 32(2), 259-300.
Koenig, M. A., \& Jaswal, V. K. (2011). Characterizing children's expectations about expertise and
incompetence: Halo or pitchfork effects? Child Development, 82, 1634-1647.
Lutz, D. J., \& Keil, F. C. (2002). Early understanding of the division of cognitive labor. Child Development, 73(4), 1073-1084
Piaget, J. (1959). The language and thought of the child New York: Harcourt, Brace \& Co, Inc.
Taylor, M., Cartwright, B. S., \& Bowden, T. (1991).
Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? Child Development, 62(6), 1334-1351.
VanderBorght, M., \& Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. Infant and Child Development, 18(61-71).

# The Effects of Shared Storybook Reading on Word Learning: A Meta-Analysis 

Zoe Flack<br>University of Sussex

Andy Field<br>University of Sussex

Jessica Horst<br>University of Sussex


#### Abstract

Although a rich literature documents pre-literate children's word learning success from shared storybook reading, a full synthesis of the factors which moderate these word learning effects has been largely neglected. This meta-analysis included 37 studies with 2,256 children, reflecting 104 effect sizes, investigating how the number of target words, tokens, story repetitions, reading styles and related factors moderate children's word comprehension. Dialogic reading styles, tokens, the number of words tested and story repetition all moderated word learning effects. Children's age, who read, number of target words and time between story and test were not moderators. These results provide information to guide researchers and educators alike to the factors with the greatest impact on improving word learning from shared storybook reading.


# Children Learn Words Better From One Storybook Page at a Time 

Zoe M. Flack (z.flack@sussex.ac.uk)<br>School of Psychology, University of Sussex, Pevensey 1 Building, Brighton, BN1 9QH UK<br>Jessica S. Horst (jessica@sussex.ac.uk)<br>School of Psychology, University of Sussex, Pevensey 1 Building, Brighton, BN1 9QH UK


#### Abstract

Two experiments tested how the number of illustrations in storybooks influences 3.5-year-old children's word learning from shared reading. In Experiment 1, children encountered stories with either two illustrations, one illustration or one large illustration (in the control group) per spread. Children learned significantly fewer words when they had to find the referent within two illustrations presented at the same time. In Experiment 2 a gesture was added to guide children's attention to the correct page in the two illustrations condition. Children who saw two illustrations with a guiding gesture learned words as well as children who had seen only one illustration per spread. Results are discussed in terms of the cognitive load of word learning from storybooks.


Keywords: word learning; cognitive load; extraneous information; storybooks; illustrations.

Sharing illustrated storybooks is a common activity for parents and young children (e.g., Rideout, Vanderwater, \& Wartella, 2003) and provides a richer source of vocabulary than everyday conversation (Montag, Jones, \& Smith, 2015). Several studies demonstrate that the styles of illustrations influence how well children learn from books (e.g., Tare, Chion, Ganea, \& DeLoache, 2010). However, little is known about how the number of illustrations influences learning. The current experiments investigate how well children learn new words from storybooks when they view one or two illustrated scenes at a time.

Pre-literate children rely on illustrations to help them make sense of the story content (for a review see, Wagner, 2013). Specifically, in an eye-tracking study, Justice, Skibbe, Canning, and Lankford (2005) found 4 -year-old kindergarten children looked longer at the illustrations than the print that accompanied texts, indicating that even with some emerging print awareness, children look primarily at illustrations. In another eye-tracking study, Evans and SaintAubin (2005) found that even with a range of illustration styles, preschool children spent the majority of their time looking at illustrations and only $6 \%$ of their time looking at the printed text (see e.g., Roy-Charland, Perron, Boulard, Chamberland, \& Hoffman, 2015; Roy-Charland, SaintAubin, \& Evans, 2007).

Pre-literate children have a growing awareness of reading conventions, such as, print conveys meaning and is read from left-to-right and top-to-bottom (for a review see International Reading Association \& The National

Association for the Education of Young Children, 1998; Snow, Burns, \& Griffin, 1998). However, because they cannot yet read, young children are unlikely to know when the reader has moved from the left-hand page to the righthand page. That is, children may be unable to determine which illustrated scene represents which part of the story. Thus, multiple illustrated scenes displayed simultaneously may make it more challenging to associate new words with their illustrated representations.

Evidence suggests that word learning is even more challenging for children when increasing amounts of perceptual information are presented simultaneously. For example, children struggle to learn object names when target objects are presented in less predictable locations (Benitez \& Smith, 2012), with many extraneous objects (Horst, Scott, \& Pollard, 2010) and with multiple combinations of extraneous objects, rather than the same combinations repeatedly (Axelsson \& Horst, 2014). Such findings are consistent with cognitive load theory (Sweller, 1998, 1989 or see, Paas, Renkl, \& Sweller, 2003 for a review), which explains how working memory capacity is inherently limited and is especially problematic in situations with extraneous information. Thus, reducing extraneous perceptual information helps children focus on the target information, which then improves learning. For example, Son, Smith, and Goldstone (2008), reduced cognitive load by providing simplified depictions of novel objects and found that this promoted better generalization of novel objects than more complex examples. Whether decreasing the number of illustrated scenes presented simultaneously in a storybook also decreases the cognitive load of word learning from shared storybook reading remains unknown.

In the current experiments we investigate whether decreasing the number of storybook illustrations presented simultaneously increases preschool children's ability to learn words incidentally from shared storybook reading. All children were presented with three storybooks that included illustrated scenes of a family's activities. The same two novel objects were included across the scenes and were named on the pages on which they were depicted (four pages for each object). Critically, all children heard the same three stories and saw the same 10 illustrations per story, however, the number of illustrations presented simultaneously and guidance varied across conditions. In Experiment 1, children saw either two illustrations (one
scene on each page of the open book) or one illustration (only on the right-hand page with the other side blank). Children in a control condition saw a large storybook (cf. Big Book Reading, Tse \& Nicholson, 2014) with one illustrated scene the same size as the two illustrations combined because we were concerned that another difference between the other conditions was the length of time needed to visually scan the two illustrations rather than one illustration. If decreasing the number of illustrations also decreases the cognitive load of word learning from storybooks then children should learn more words when they see only one illustration at a time. In contrast, if the number of illustrations does not affect cognitive load, then children should learn words equally from one- or twoillustration books. In Experiment 2, we investigate whether guiding children's attention to the correct page with a simple gesture helps children focus on the correct page and improves word learning-even with two illustrations.

## Experiment 1

## Method

Participants Thirty-six 3.5-year-old children ( $M=41.99$ months, $S D=1.76$ months, range $=38.87-45.14$ months) participated. Children were monolingual, British-English speakers from predominantly middle-class families. All children were typically developing with no reported speech or language difficulties. Twelve children each were randomly assigned to one of the three conditions: one illustration ( $M=41.87, S D=0.65,6$ girls), two illustrations ( $\mathrm{M}=42.85, S D=0.43,6$ girls), or control condition (one large illustration, $M=41.92, S D=0.45,6$ girls). There was no difference in maternal education levels between conditions, Fisher's Exact Test $=3.71, p=.98$. Two mothers each in the one and two illustrations conditions and three mothers in the control condition had completed high school (GCSEs and/or A-levels) and/or completed a vocational diploma or access course. Eight mothers each in the one and two illustrations conditions, and six in the control condition had an undergraduate degree and/or an undergraduate degree with a postgraduate certificate (e.g., Postgraduate Certificate in Education (PGCE), an additional teaching qualification). One mother each in the one illustration and control conditions had a Master's degree and one mother in each condition had a doctoral degree. One mother in the two illustrations condition and one mother in the control condition declined to answer this question. Parents were reimbursed for travel costs and children chose a small gift as a thank you for participating (e.g., a colouring book).

Storybooks. Stimuli included three 10-page storybooks slightly modified from Horst, Parsons, and Bryan (2011) The Very Naughty Puppy, Nosy Rosie at the Restaurant, and Rosie's Bad Baking Day. Each storybook depicted and named the same two novel objects four times. Each object had a function: the orange inverted slingshot functioned like
a hand mixer (tannin) and the metal kinetic wheel was used like a rolling pin (sprock). Throughout each story, objects were named incidentally and were not the focus of the story. The objects appeared twice on their own pages and twice together. We used real photographs edited with the poster edges feature in Photoshop to make them look like drawings typical of a commercially available children's book. Across storybooks there was no difference in the number of words per page, $M=45, S D=9.34, F(2,24)=0.98, p=.39$.

All children heard the same stories and saw all the illustrations for each story. The only difference between conditions was the way storybooks were printed (see Figure 1): children either heard stories with two A4 illustrated scenes per open spread, one A4 illustrated scene per spread (i.e., the left-hand page was always blank), or one A3 illustrated scene per spread. In the ISO A-series paper system (i.e., European standard), A3 pages ( 29.70 x 42.00 cm ) are twice the size of A4 pages ( $21.00 \times 29.70 \mathrm{~cm}$ ), thus the A3 condition served as a control condition where the storybooks included only one illustration per spread (as in the one A4 illustration condition) but included the same overall illustrated area as the two A4 illustrations condition). Because the one illustrations condition differed from the two illustrations condition in both surface area and amount of items/details, we wanted to include a control condition to disentangle which of these was driving any potential effects. Equating the number of items/details would have precluded presenting all children with the same illustrations; therefore, we tested surface area as the control condition. Data from all three conditions were collected at the same time.


Figure 1. Page 5 in Rosie's Bad Baking Day as seen by children in the 2 illustrations, 1 illustration and 1 large illustration conditions, respectively. Note, in the 2 illustrations condition page 4 is viewed at the same time as page 5.

Test stimuli. An A4 test booklet with images of four novel objects per right-hand page was used on the test trials (the left-hand pages were blank). On each page, four objects were presented on a plain white background without any other contextual information. Across test trials the targets (tannin and sprock) were presented with four additional novel objects that the children had not previously seen, so that each trial would present children with a different combination and it would not appear that a question was being repeated. Finally, a practice trial page included images of four known objects: a dog, a plane, a duck and a chair.

Procedure. Each child was tested individually in a children's lab at the university. During the reading phase, the experimenter sat opposite the child and held the storybook upright, to her side, with the pages facing the child, like a preschool teacher would when reading to a group of children during "circle time." The parent sat on a seat in a different corner of the room. All children were read each of the three stories. For each child all three stories were presented in the same format (e.g., two illustrations per spread). No dialogic techniques, such as giving definitions for novel words or pointing, were used during the readings. Story-order was counterbalanced across children.

After reading the final story, the experimenter proceeded to the test phase, which began with four warm-up trials to get the child used to pointing to pictures in the test booklet and to ensure the child understood the task. The experimenter opened the test booklet to one of the warm-up trial pages and asked the child to point to one of the familiar objects (e.g., "can you point to the plane?"). Across the four counterbalanced warm-up trials, children were asked to point to an object in each quadrant of the page. Next, the experimenter tested word learning. On each trial the experimenter turned to a different test page and asked the child to point to one of the novel objects. In total children were asked to point to each target novel object twice (see also Werchan \& Gómez, 2014). On half of the trials one target was present (e.g., sprock with three other novel objects) and on half of the trials both targets were present (e.g., sprock and tannin with two other novel objects). These latter trials ensured that children were not simply choosing the correct object because it had been the only one they had previously seen (for a review of this issue see (Axelsson \& Horst, 2013). Trial order, page and quadrant were counterbalanced across participants.

To confirm that children do not dislike books with only one illustration per page, we also asked children to rate their enjoyment of the individual stories and found no differences in enjoyment across stories or conditions, therefore for brevity these data are not included here but are available from the authors upon request.

## Results

Children in the one illustrations condition $(M=0.75, S D=$ $0.34, t(11)=5.14, p<.001, d=1.48$ ) and in the control (one large) condition ( $M=0.75, S D=0.30, t(11)=5.75, p<$ $.001, d=1.66$ ), chose the target object more than expected by chance (.25) see Figure 2, Left Panel. However, with Bonferroni's correction ( $p=.017$ ), children in the two illustrations condition did not chose the target object more than expected by chance $(M=0.44, S D=0.28, t(11)=2.28$, $p=.04, d=.66$ ). To test for differences between illustration formats, children's proportions of correct choices were entered into an ANOVA with illustration format (two, one, one large) as between-subjects factor. The ANOVA yielded a main effect of illustration format, $F(2,33)=4.10, p=.03$, partial $\eta^{2}=0.20$. Planned contrasts showed that children who saw two illustrations learned fewer words than children
who saw one illustration per spread, $t(33)=2.87, p=.007$, partial $\eta^{2}=0.20$. There was no difference in word learning between one illustration in A4 or one illustration in A3 $t(33)$ $=0.00$, ns. Thus, illustration size did not affect word learning, but the number of illustrations did.


Figure 2. Proportion of correct word learning trials for Experiment 1 (left) and Experiment 2 (right). Error bars represent +1 SEM.

## Discussion

Many illustrated storybooks are printed with two illustrations per spread (e.g., In the Night Kitchen by Maurice Sendak or Dinosaur Roar! By Paul and Henrietta Stickland)—if not more (e.g., The Incredible Book Eating Boy by Oliver Jeffers contains 6 illustrations on pages 7-8). Further, some books include a combination of one or more illustrations per spread (e.g., The Smartest Giant in Town by Julia Donaldson). Our goal is not to suggest that all of these books be reprinted. Therefore, we conducted a supplementary experiment to provide children with additional support so that they can learn from storybooks with multiple illustrations.

Because young children do not necessarily know when the text is referring to the left- or right-hand page, they may benefit from a non-verbal gesture to look to the correct page. Specifically, a non-verbal signal may help children to focus on the correct illustration at the correct time, thus improving their chances of learning new words from the storybook (cf. Booth, McGregor, \& Rohlfing, 2008).

Thus, in Experiment 2 we again read children storybooks with two illustrations per spread, but included a quick sweeping hand gesture to indicate which page we were reading from. We chose a sweeping gesture over other possible techniques to keep the manipulation visual without additional auditory information. We did not use a pointing gesture because we wanted to perform the same gesture on every page and some pages did not include a novel object, while others included both novel objects. Thus, this
sweeping gesture allowed us to maintain an incidental word learning task, rather than providing ostensive reference. If storybooks with one illustrated scene per spread are more helpful than storybooks with two illustrated scenes because children do not know which page to look at, then guiding them towards the correct page should improve word learning to similar levels as those from single illustration displays.

## Experiment 2

## Method

Participants An additional twelve 3.5 -year-old children ( $M$ $=40.45$ months, $S D=1.30$ months, range $=38.45$ to 45.03 months, 6 girls) participated. Children were monolingual, British-English speakers with no reported speech or language difficulties. Two mothers had completed high school (GCSE's and/or A-levels), seven had an undergraduate degree or an undergraduate degree with a postgraduate certificate. One mother had completed a Master's degree, one a doctoral degree and one declined to provide this information. Parents were reimbursed for travel costs and children chose a small gift as a thank you for participating (e.g., a colouring book).

Stimuli. The same stimuli were used as in the two illustrations condition in Experiment 1.

## Procedure

All children were read the three storybooks with two illustrations per spread. The procedure was the same as in Experiment 1 except that before reading each page, the experimenter smoothly swept her open hand from the top of the page to the bottom, thereby drawing children's attention to the correct page.

## Results

Children learned the words from the story (see Figure 2, Right Panel). Specifically, children chose the target object more than expected by chance $(M=0.88, S D=0.17, t(11)=$ 12.84, $p<.001, d=3.71$ ). Our goal was to determine whether adding a simple gesture would be sufficient to improve children's word learning from storybooks with two illustrations per spread. Thus, we compared the word learning performance of children in the current study to children in the two illustrations condition of Experiment 1. Children who had the additional support to guide their attention to the correct page learned words significantly better than children who did not have that support, $t(22)=$ $4.58, p<.001, d=8.78$.

## Discussion

In Experiment 2 we investigated whether orienting children's attention to the correct storybook page with a simple gesture while reading could diminish the effects of cognitive load from multiple illustrated scenes found in Experiment 1. Adding the gesture did not significantly
increase the amount of time needed to read the story, but did significantly improve children's word learning compared to reading without a guiding gesture.

The rates of word learning observed in Experiment 2 are similar to other studies using dialogic reading techniques, such as pointing or asking questions (Walsh \& Blewitt, 2006). For example, Ard and Beverly (2004) read storybooks to 3 - and 4 -year-old children either verbatim or with one of three dialogic techniques; added questions, added comments, or both questions and comments. Children learned approximately $75 \%$ of the new vocabulary with the dialogic reading techniques included but only $53 \%$ with verbatim readings. Although the efficacy of the use of dialogic techniques to improve children's word learning from storybooks is not in doubt, multiple dialogic techniques are often employed in combination, making it harder to compare effects across the literature for individual techniques (see Wasik, Hindman, \& Snell, 2016 for a recent review). It is therefore particularly exciting to see that such a simple gesture could have such powerful effects on children's learning.

## General Discussion

Across two experiments we investigated whether decreasing the number of storybook illustrations presented simultaneously increases preschool children's ability to learn words from shared storybook reading. In Experiment 1 we read children 10-page stories with either one, two, or one large illustration per spread. Children learned the new words better when presented with only one illustration per spread, regardless of the image size, even though all children saw the same number of illustrations overall. In Experiment 2 we read children the same stories with two illustrations per spread, but added a small sweeping gesture to indicate which page we were reading. Although children in this condition were presented with multiple illustrations at once, they were able to learn more words than expected by chance and more words than children presented with the same number of illustrations but no guidance on which page to attend to. Taken together these findings suggest that children's word learning is improved by helping children focus on the relevant information by either reducing the number of illustrations presented (Experiment 1) or directing their attention to the correct illustration (Experiment 2).

These findings are consistent with cognitive load theory (Paas et al., 2003; Sweller, 1988, 1989), which suggests that extraneous information can prevent optimal learning. The more information children need to think about, the more challenging the task. Consequently, removing extraneous perceptual information may improve learning (see, e.g., Son et al., 2008). For example, kindergarten children are better able to learn information from science lessons when the extraneous information of a highly-decorated classroom is removed (Fisher, Godwin, \& Seltman, 2014). Similarly, reducing the amount of extraneous information in graphs improves children's mathematics skills (Kaminski \&

Sloutsky, 2013) and removing extraneous information in ABC books improves alphabet learning (Chiong \& DeLoache, 2012). In the current study, in the two illustrations format, children are faced with processing additional materials-which in some cases may even provide conflicting information-slowing down the process of word learning. Children do not know when the story moves from one illustration to the other. In contrast, in the one illustration format, the child is provided with only the relevant scene, which corresponds with the text they are currently hearing, thereby reducing the cognitive load associated with understanding the story and the new words. Similarly in Experiment 2, children are directed towards the relevant scene, thereby reducing cognitive load.

Although children in the current studies learned target words better when presented with single illustrations, there may be benefits for other types of learning from multiple illustrations. For example, story comprehension may be better supported by having more to look at, particularly as visual attention to illustrations during storybook reading predicts story comprehension (Kaefer, Pinkham, \& Neuman, 2016).

The current findings may also be informative for research comparing e-books (i.e., storybooks presented on screens) with traditional two-illustration paper storybooks. Some studies report a deficit in learning from e-books (e.g., Segers, Takke, \& Verhoeven, 2004) while others do not (e.g., Korat \& Shamir, 2007). One explanation for this discrepancy is that e-books often contain added manipulative features, which may hinder learning. For example, e-books often contain additional games (e.g., de Jong \& Bus, 2002) or interactive dictionary features (e.g., Korat, 2009). Previous research indicates that added manipulative features such as pull-tabs hinder learning from paper books (Tare et al., 2010), however some features of ebooks may be helpful in the same way as dialogic techniques by highlighting key information at the right time. Another explanation is that e-books are often presented only one illustration at a time (e.g., Verhallen \& Bus, 2011), which could be an additional confounding factor when comparing between storybook media types. The current findings suggest that such single illustrations help children focus their attention on relevant information and may aid learning especially when children are exploring books without an adult.

The current experiments demonstrate that reducing the number of simultaneous illustrations to just one at a time improves children's word learning from shared storybook reading. This has important implications for educational research and suggests that even seemingly minor differences in illustration format can result in significant differences in how well children learn. These findings should help shape future storybook research design, and provide useful practical solutions, which could be used by teachers and parents alike and may inform our understanding of how to create eBooks and other media that children may encounter without an adult. Furthermore, in an age of seemingly
endless possibilities, they provide a stark reminder that less is sometimes more.

## References

Ard, L. M., \& Beverly, B. L. (2004). Preschool word learning during joint book reading: Effect of adult questions and comments. Communication Disorders Quarterly, 26(1), 17-28. doi: 10.1177/15257401040260010101

Axelsson, E. L., \& Horst, J. S. (2013). Testing a word is not a test of word learning. Acta Psychologica, 144(2), 264-268. doi: 10.1016/j.actpsy.2013.07.002
Axelsson, E. L., \& Horst, J. S. (2014). Contextual repetition facilitates word learning via fast mapping. Acta Psychologica, 152, 95-99. doi: 10.1016/j.actpsy.2014.08.002

Benitez, V. L., \& Smith, L. B. (2012). Predictable locations aid early object name learning. Cognition, 125, 339-352. doi: 10.1016/j.cognition.2012.08.006
Booth, A. E., McGregor, K. K., \& Rohlfing, K. J. (2008). Socio-pragmatics amd attention@ Contributions to gesturally guided word learning in toddlers. . Language Learning and Development, 4(3), 179202. doi: 10.1080/15475440802143091

Chiong, C., \& DeLoache, J. (2012). Learning the ABCs: What kinds of picture books facilitate young chidlren's learning? Journal of Early Childhood Literacy, 13(2), 225-241. doi: 10.1177/1468798411430091
de Jong, M. T., \& Bus, A. G. (2002). Quality of bookreading matters for emergent readers: An experiment with the same book in a regular or electronic format. Journal of Educational Psychology, 94, 145-155. doi: 10.1037/00220663.94.1.145

Donaldson, J. (2002). The smartest giant in town. UK: Macmillan Children's Books.
Evans, M. A., \& Saint-Aubin, J. (2005). What children are looking at during shared storybook reading: Evidence from eye movement monitoring. Psychological Science, 16(11), 913-920. doi: 10.1111/j.1467-9280.2005.01636.x

Fisher, A. V., Godwin, K. E., \& Seltman, H. (2014). Visual environment, attention allocation, and learning in young children. When too much of a good thing may be bad. Psychological Science, 25(7), 13621370. doi: 10.1177/0956797614533801

Horst, J. S., Parsons, K. L., \& Bryan, N. M. (2011). Get the story straight: contextual repetition promotes word learning from storybooks. Frontiers in Psychology, 2. doi: 10.3389/Fpsyg. 2011.00017

Horst, J. S., Scott, E. J., \& Pollard, J. A. (2010). The role of competition in word learning via referent selection. Developmental Science, 13(5), 706-713. doi: 10.1111/j.1467-7687.2009.00926.x

International Reading Association, \& The National Association for the Education of Young Children.
(1998). Learning to read and write: Developmentally appropriate practices for young children. The Reading Teacher, 52(2), 193-216.
Jeffers, O. (2007). The incredible book eating boy. NY, USA: Penguin Group.
Justice, L. M., Skibbe, L. E., Canning, A., \& Lankford, C. (2005). Pre-schoolers, print and storybooks: An observational study using eye movement analysis. Journal of Research in Reading, 28(3), 229-243. doi: 10.1111/j.1467-9817.2005.00267.x
Kaefer, T., Pinkham, A. M., \& Neuman, S. B. (2016). Seeing and knowing: Attention to illustrations during storybook reading and narrative comprehension in 2-year-olds. Infant and Child Development, e2018. doi: 10.1002/icd. 2018
Kaminski, J., A, \& Sloutsky, V. M. (2013). Extraneous perceptual information interferes with children's acquisition of mathematical knowledge. Journal of Educational Psychology, 105(2), 351-363. doi: 10.1037/a0031040

Korat, O. (2009). The effects of CD-ROM stroybook reading on Israeli children's early literacy as a function of age group and repeated reading. Education and Information Technologies, 14(1), 39-53. doi: 10.1007/s10639-008-9063-y
Korat, O., \& Shamir, A. (2007). Electronic books versus adult readers: effects on children's emergent literacy as a function of social class. Journal of Computer Assisted Learning, 23(3), 248-259. doi: 10.1111/j.1365-2729.2006.00213.x

Montag, J., L., Jones, M. N., \& Smith, L., B. (2015). The words children hear: Picture books and the statistics for language learning. Psychological Science, 26(9), 1489-1496. doi: 10.1177/0956797615594361

Paas, F., Renkl, A., \& Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. Educational Psychologist, 38(1), 14. doi: 10.1207/s15326985ep3801_1

Rideout, V. J., Vanderwater, E. A., \& Wartella, E. A. (2003). Zero to six. Electronic media in the lives of infants, toddlers and preschoolers. Menlo Park, CA: The Kaiser Family Foundation.
Roy-Charland, A., Perron, M., Boulard, J., Chamberland, J., \& Hoffman, N. (2015). "If I point, do they look?": The impact of attention-orientation strategies on text exploration during shared book reading. Reading and Writing, 28(9), 1285-1305. doi: 10.1007/s11145-015-9571-2

Roy-Charland, A., Saint-Aubin, J., \& Evans, M. A. (2007). Eye movements in shared book reading with children from kindergarten to Grade 4. Reading and Writing, 20(9), 909-931. doi: 10.1007/s11145-007-9059-9
Segers, E., Takke, L., \& Verhoeven, L. (2004). Teachermediated versus computer-mediated storybook reading to children in native and multicultural
kindergarten classrooms. School Effectiveness and School Improvement, 15(2), 215-226. doi: 10.1076/sesi.15.2.215.30430

Sendak, M. (1971). In the night kitchen (Red Fox 2001 ed.). UK: Random House Group Ltd.
Snow, C. E., Burns, M. S., \& Griffin, P. (Eds.). (1998). Preventing reading difficulties in young children. Washington, DC: National Academy Press.
Son, J. Y., Smith, L. B., \& Goldstone, R. L. (2008). Simplicity and generalization: Short-cutting abstraction in children's object categorizations. Cognition, 108(3), 626-638. doi: 10.1016/j.cognition.2008.05.002

Stickland, P., \& Stickland, H. (1994). Dinosaur Roar! UK: Random House Childrens Publishers.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 11, 65-99.
Sweller, J. (1989). Cognitive technology: Some procedures for facilitating learning and roblem solving in mathematics and science. Journal of Educational Psychology, 81, 457-466.
Tare, M., Chion, C., Ganea, P., \& DeLoache, J. (2010). Less is more: How manipulative features affect children's learning from picture books. Journal of Applied Developmental Psychology, 31(5), 395400. doi: doi:10.1016/j.appdev.2010.06.005

Tse, L., \& Nicholson, T. (2014). The effect of phonicsenhanced Big Book reading on the language and literacy skills of 6 -year-old pupils of differen reading ability attending lower SES schools. Frontiers in Psychology, 5, 1222. doi: 10.3389/fpsyg. 2014.01222

Verhallen, M. J. A. J., \& Bus, A. G. (2011). Young second language learners' visual attention to illustrations in storybooks. Journal of Early Childhood Literacy, 11(4), 480-500. doi: 10.1177/1468798411416785
Wagner, L. (2013). By the numbers: a quantitative content analysis of children's picturebooks. Frontiers in Psychology, 4(850). doi: 10.3389/fpsyg.2013.00850

Walsh, B. A., \& Blewitt, P. (2006). The Effect of Questioning Style During Storybook Reading on Novel Vocabulary Acquisition of Preschoolers. Early Childhood Education Journal, 33(4), 273278. doi: 10.1007/s10643-005-0052-0

Wasik, B. A., Hindman, A. H., \& Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. Early Childhood Research Quarterly, 37, 39-57. doi: 10.1016/j.ecresq.2016.04.003
Werchan, D. M., \& Gómez, R. L. (2014). Wakefulness (not sleep) promotes generalization of word learning in 2.5-year-old children. Child Development, 85(2), 429-436. doi: 10.1111/cdev. 12149

# A Unified Model of Speech and Tool Use Early Development 

Sébastien Forestier ${ }^{1,2}$ \& Pierre-Yves Oudeyer ${ }^{2,3}$ (sebastien.forestier@inria.fr) Université de Bordeaux ${ }^{1}$, Inria Bordeaux Sud-Ouest ${ }^{2}$, Ensta-ParisTech ${ }^{3}$, France


#### Abstract

Some studies hypothesize a strong interdependence between speech and tool use development in the first two years of life. To help understand the underlying mechanisms, we present the first robotic model learning both speech and tool use from scratch. It focuses on the role of one important form of body babbling where exploration is directed towards self-generated goals in free play, combined with imitation learning of a contingent caregiver. This model does not assume capabilities for complex sequencing and combinatorial planning which are often considered necessary for tool use. Yet, we show that the mechanisms in this model allow a learner to progressively discover how to grab objects with the hand, how to use objects as tools to reach further objects, how to produce vocal sounds, and how to leverage these vocal sounds to use a caregiver as a social tool to retrieve objects. Also, the discovery that certain sounds can be used as a social tool further guides vocal learning. This model predicts that the grounded exploration of objects in a social interaction scenario should accelerate infant vocal learning of accurate sounds for these objects' names.


Keywords: tool use; speech development; free play; exploration; imitation learning; social tool use; goal babbling

## Introduction

Some studies hypothesize that there might be a strong interdependence between speech and tool use development in the first two years of life (Gibson, Gibson, \& Ingold, 1994). Tool use and language seems to require similar information processing capabilities allowing the production and perception of sequential combinations of increasing complexity, from reaching to spoon self-feeding and from words to stories. In addition to showing similar compositional properties, speech and tool use might share some neural correlates involving Broca's area (Higuchi, Chaminade, Imamizu, \& Kawato, 2009). Those common neural correlates could have an evolutionary origin in the hominid lineage, where a selection pressure for complex tool use, language and social behaviors might have together driven the increase in brain planning capabilities (Morgan et al., 2015). In particular, the development of tool use precursors follows several overlapping phases of behaviors: 1) body babbling, where babies learn to control their body parts, 2) interacting with a single object, and 3) exploring object-object interactions (Guerin, Kruger, \& Kraft, 2013). From pointing movements to the control of a rake, new representations and physical understanding are developed to allow the planning of tool use actions composed of combinations of more simple actions, e.g. grasping the rake. During the same period, infants progressively learn how to efficiently use their vocal tract, comprising many complex actuators from the larynx to the lips. At birth, they produce immature protophones like squeals, growls or quasi-vowels, and by the end of their first year they are able to produce the speech-like syllables of their native language (Oller, 2014). Those syllables then form words which become the basis
of syntactic combinations essential to language expressiveness. Infants do not only explore tool use and vocalizations by themselves, driven by intrinsic motivations (Moulin-Frier, Nguyen, \& Oudeyer, 2013), but also spend a great part of their time interacting with their parents and other social peers where imitation is thought to be one of the important developmental pathways (Meltzoff, 1999). For instance, infants imitate the vowels produced by an adult speaker by 6 month of age (Kuhl, 2004), and 1.5-year-olds imitate demonstrations of a rake-like tool function to retrieve an out-of-reach toy (Chen \& Siegler, 2000).

In order to investigate hypotheses about the joint development of speech and tool use, we seek to build an embodied model of tool use and speech learning. Existing robotic models of tool use showed first insights into how relations between tools and other objects could be learned from grounded experimentation. In (Stoytchev, 2005), a robotic arm focused on learning rake-like tool affordances from the exploration of already implemented stereotyped arm behaviors. In (Tikhanoff, Pattacini, Natale, \& Metta, 2013), the iCub robot was given its arm's forward model and inverse optimization methods which led to stereotyped grasping. A recent series of robotic models considered the learning of tool use from scratch, without any kind of pre-programmed reaching skills (Forestier \& Oudeyer, 2016a, 2016b, 2016c). Those models studied the developmental progression of robotic agents between phases of behaviors with objects, and the evolution of their strategies to reach goals. They have shown interesting similarities with infant development in terms of developmental trajectories and strategy choice dynamics.

Recent computational models of vocal development make use of a simulated vocal synthesizer that the learning agent must control in order to produce vocalizations, with human sounds as targets to be imitated (Warlaumont, Westermann, Buder, \& Oller, 2013; Philippsen, Reinhart, \& Wrede, 2014). In (Moulin-Frier et al., 2013), the agent chooses the strategy that shows the best competence progress: either autonomously training to reach phonetic goals, or trying to imitate human sounds. They show that the intrinsic motivation for learning progress self-organizes coherent infant-like developmental sequences. Those models of language acquisition study several developmental pathways to the learning of forward and inverse models of a simulated vocal tract, from autonomous exploration to human sounds imitation. However, agents are not situated into a physical environment where vocalizations have a meaning related to objects.

Several works study joint action and language learning (Cangelosi et al., 2010), but give an advanced knowledge of the linguistic interaction protocol to the learning agent who has to associate predefined actions or objects to predefined
labels and learn the semantic compositionality. Also, agents learn actions without a nested tool use property.

In this paper we describe the first model that jointly considers the early development of both tool use and speech. Such a model could allow the investigation of hypotheses about the mechanisms underlying the observed links between tool use and speech development. In a previous work, we showed that the Model Babbling learning architecture (Forestier \& Oudeyer, 2016b) allows the development of tool use in a robotic setup, through several fundamental ideas. First, goal babbling is a powerful form of exploration to produce a diversity of effects by self-generating goals in a task space (Baranes \& Oudeyer, 2013). Second, the possible movements of each object define a task space in which to choose goals, and the different task spaces form an object-based representation that facilitates prediction and generalization, as explained by (Chang, Ullman, Torralba, \& Tenenbaum, 2016). Also, cross-learning between tasks updates all skills while exploring one in particular. A novel insight was that early development of tool use could happen without a combinatorial action planning mechanism: modular goal babbling in itself allowed the emergence of nested tool use behaviors.

Here we extend this architecture so that the agent can imitate caregiver's sounds in addition to autonomously exploring. We hypothesize that these same algorithmic ingredients allow a joint unified development of speech and tool use. Our learning agent is situated in a simulated environment where a vocal tract and a robotic arm are to be explored with the help of a caregiver. The environment is composed of three toys, one stick that can be used as a tool to move toys, and a caregiver moving around. The caregiver helps in two ways. If the agent touches a toy, the caregiver produces this toy's name, but otherwise produces a distractor word as if it was talking to another adult. If the agent produces a sound close to a toy's name, the caregiver moves this toy within agent reach.

We show that our learning architecture based on Model Babbling allows agents to learn how to 1) use the robotic arm to grab a toy or a stick, 2) use the stick as a tool to get a toy, 3) learn to produce toy names with the vocal tract, 4) use these vocal skills to get the caregiver to bring a specific toy within reach, and 5) choose the most relevant of those strategies to retrieve a toy that can be out-of-reach. Also, the grounded exploration of toys accelerates the learning of the production of accurate sounds for toy names once the caregiver is able to recognize them and react by bringing them within reach, with respect to distractor sounds without any meaning in the environment. Our model is the first to allow the study of the early development of tool use and speech in a unified framework.

## Methods

## Learning Environment

The learning environment ${ }^{1}$ is composed of a simulated 2D robotic arm and a simulated vocal tract that the agent controls

[^328]to interact with a caregiver and toys. In each trial, the agent observes the current environmental state and then executes a motor trajectory, either corresponding to moving the motors of the arm or of the vocal tract, and gets the associated sensory feedback composed of the trajectory of each object and the sound produced by the agent or the caregiver (see Fig.1).

Simulated Robotic Arm The simulated 2D robotic arm has 3 joints, with its base fixed at position $[0,0]$. Each joint rotates from $-\pi \mathrm{rad}$ to $\pi \mathrm{rad}$ and the 3 segments of the arm have length $0.25,0.15$ and 0.1 , so the arm has length 0.5 . The framework of Dynamical Movement Primitives (Ijspeert, Nakanishi, Hoffmann, Pastor, \& Schaal, 2013) is used to generate smooth joint trajectories from motor parameters. Each of the 3 joints is controlled by a DMP starting at the rest position of the joint (position 0 ) and parameterized by 7 weights: one weight on each of 6 basis functions and one weight representing the end position of the joint trajectory. To sum up, the agent provides a set of 21 trajectory parameters which are translated through DMPs to a set of smooth 50-steps trajectories for the arm's joints which gives a smooth 2D trajectory to the robotic hand.

Tool and Toys In the environment of the robotic arm, 3 toys can be grasped with the hand or with the help of a stick. The stick has length 0.25 and is considered grasped as soon as the hand reaches the handle side (orange) within a distance of 0.1 . At the end of the movement the stick is dropped and stays at its current position while the arm is reset to its rest position for the next iteration. The toys are reset to a random location every 20 iterations, at a distance between 0 and 1 from the center so possibly at an unreachable position.

Simulated Vocal tract A vocal tract is simulated through the DIVA model (Guenther, 2006) and allows the production of different sounds that we can classify into vowels. In the DIVA model, a set of parameters defines a vocal tract contour where each represents one component of a Principal Component Analysis of midsagittal MRI vocal tract profiles (see Fig.1b), from the jaw and tongue to the lips position. Here we use only the first 7 articulatory parameters, controlling most of the vocal tract shape's variability. From a vocal tract contour defined by a set of parameters, the DIVA software computes the corresponding sound and outputs its first 2 formants, which are often considered to give enough information to distinguish common English vowels. The DMP framework generates smooth trajectories of vocal parameters, as described above for arm parameters, to allow the simulated vocal tract to produce simple words composed of several vowels. Each of the 7 articulators is controlled by a DMP parameterized by 4 weights: the starting and end position of the parameter trajectory, and weights on 2 basis functions. Given a set of 28 trajectory parameters provided by a learning agent, the DMPs output a set of smooth 50 -steps trajectories for the 7 articulators that we use in the DIVA model, which through the DIVA software generates a smooth trajectory of the first two formants (called $F 1$ and $F 2$ ).


Figure 1: Agent's robotic and vocal environment. (a) Agent's 3 DOF arm, controlled with 21 parameters, grabs toys with its hand, or uses the stick to reach toys. Caregiver brings a toy within reach if the agent says its name. (b) Agent's vocal environment representing sounds as trajectories in the two first formants space. Agent's simulated vocal tract produces sounds given 28 parameters. When agent touches a toy, caregiver says toy's name. Some sounds corresponding to random parameters are plotted in red, and some sounds produced when imitating caregiver's /uye/ word in blue (best imitation in bold, error 0.3).

Sounds: from Vowels to Words The simulated vocal tract controlled through DMPs has the potential to produce words composed of a sequence of 3 vowels in the set $\{/ \mathrm{o} /, / \mathrm{u} /, / \mathrm{i} /$, $/ \mathrm{e} /, / \mathrm{y} /\}$. See Fig. 1 (b), "Motor babbling" condition, for an example of 200 trajectories corresponding to random sets of 28 parameters. We define a set of 6 words that the caregiver produces perfectly: $\{/ \mathrm{yeo} /$, /euy/, /iuo/, /uye/, /eou/, /oey/\}. A sound trajectory produced by the vocal tract is recognized if its distance to the perfect word is lower than 0.4.

Caregiver's guidance A simulated caregiver is given two roles to help the learning agent. First, at the beginning of the experiment, the caregiver chooses randomly a label for each toy from the set of 6 words. When the agent touches a particular toy with its hand, the caregiver then produces the sound trajectory corresponding to the label of this toy. If the agent does not touch any toy with the arm, the caregiver produces one of the distractor sounds, as if she was talking to another adult. Second, if the agent produces a sound trajectory recognized by the caregiver as the label of a toy, the caregiver moves the corresponding toy in between herself and the agent so that it becomes reachable by the agent with the hand. The caregiver is reset to a random position at each iteration.

Sensory Feedback Before choosing a motor command, the agent receives the state of the environment (or context) as the 2 D position of the caregiver, the stick and the 3 toys (so 10D). At the end of the movement, the agent receives a sensory feedback $s$ in the sensory space $S$ (60D), from the different objects in the environment. First, the trajectory of the hand is represented as its $x$ and $y$ positions at 5 time points: steps 1 , $13,25,38,50$ of the 50 -steps trajectory ( $S_{\text {Hand }}, 10 \mathrm{D}$ ). Similarly, the trajectories of the stick and the 3 toys during the movement are represented in 10 dimensional sensory spaces
( $S_{S t i c k}, S_{T o y_{1}}, S_{T o y_{2}}, S_{T o y_{3}}, 10 \mathrm{D}$ each). Sound, either produced by the agent or by the caregiver, is represented by the position of the first two formants at 5 time points ( $S_{\text {Sound }}, 10 \mathrm{D}$ ).

## Unified Modular Learning Architecture

The goal of a learning agent is to use its robotic arm and vocal tract to discover a diversity of sensory effects, and collect data to learn repertoires of skills in the form of inverse models allowing to reproduce these effects. Consequently, the agent is not given a priori a single target task to be solved, but a modular object-based representation of task spaces. The agent learns a set of sensorimotor models mapping a motor space to one particular sensory space (see Fig. 2). For instance, model 1 learns to move the hand from arm parameters, model 2 learns to move the stick, model 3,4 , and 5 learn to move one of the toys, and model 6 how to produce sounds with the arm, which will be possible by touching one of the toys with the hand so that the caregiver produces the corresponding label. Controlling vocal tract, model 7, 8 and 9 learn to move one of the toys by involving caregiver's help, and model 10 learns to produce diverse sounds autonomously.

Exploration through Model Babbling In order to actively explore and learn the 10 sensorimotor models from experimentation with the environment, learning agents use the Model Babbling architecture developed in (Forestier \& Oudeyer, 2016b) that we extend to handle the 2 motor spaces: the robotic arm and the vocal tract. First, the agent performs some random exploration of motor spaces, 500 with the robotic arm and 500 with the vocal tract, to get an initial sampling of those spaces. Then, at each iteration, the learning agent first chooses to train one of the 10 models, chosen randomly (e.g. from the robotic arm to the hand sensory space). A particular goal is then randomly chosen in


Figure 2: Learning Architecture. Agent controls 2 motor spaces and receives sensory feedback about 6 objects. Each arrow represents one of the 10 sensorimotor models learned.
the sensory space corresponding to the chosen model (e.g. a particular 2D trajectory of the hand). The agent then uses the corresponding inverse model to infer a motor command in the corresponding motor space (e.g. arm parameters) to reach the goal. Exploration happens in goal choice and in the new motor parameters that inverse models infer with generalization mechanisms and adding exploration noise.

Imitation of Sounds When the agent is choosing to train to produce sounds with its vocal tract (model 10), instead of always choosing random goals, it does this for half of the iterations (chosen randomly), and the other iterations are focused on trying to imitate the caregiver, by randomly choosing one of the sounds previously produced by the caregiver as a goal.

Forward and Inverse Models Each sensorimotor model provides a forward model and an inverse model, with the same implementation as in (Forestier \& Oudeyer, 2016b). The forward model predicts which sensory trajectory would be observed given the current context and a motor command to execute. The inverse model infers a motor command that could reach a desired goal given the current context. When a motor command $m$ is executed (either 21 parameters for the robotic arm or 28 for the vocal tract) in a context $c$ and a sensory feedback $s$ is received in $S$, all the sensorimotor models that share the same motor space are updated. New information comes as a tuple ( $m, c_{i}, s_{i}$ ) with $s_{i}$ being a subset of $s$ variables corresponding to the respective sensory space, and $c_{i}$ being the subset of $c$ relevant for this sensorimotor model. The relevant context for models 1 and 10 is empty, which means that hand trajectories and vocal sounds produced by the agent do not depend on the current position of other objects. The context for model 2 is the position of the stick, and for models 3,4 , and 5 , the position of the stick and of the corresponding toy. For model 6 , all toys are relevant, and for models 7, 8 and 9, the caregiver and the toy is useful. Given a database of ( $m, c_{i}, s_{i}$ ) experiments, an inverse model infers a probable motor command $m$ to reach a goal $s_{g}$ in a context $c_{i}$ by looking for the nearest neighbor $s_{N N}$ in $S_{i}$ of $s_{g}$ and retrieving the associated motor parameters $m_{N N}$ that were used to reach $s_{N N}$. It then outputs $m_{N N}$ plus Gaussian noise ( $\sigma=0.05$ ) to explore new parameters.

## Results

We ran 500 independent trials of 80000 iterations (or robot experiment) each. We measured how agents learned to move objects by giving them new goals in new contexts, and we analyzed the accuracy of the learned vocalizations.

## Competence to Reach Toys

After 80000 iterations of training, we measured the performance of each agent to retrieve a toy depending on its current position with its preferred method: with the hand, with the stick used as a tool or involving caregiver's help. Fig. 3 shows the mean competence of all agents to retrieve toys depending on the current position of the toys. The competence error to retrieve a toy is measured by the distance between a goal trajectory given to the agent, where the toy is moved towards the center, and the actual trajectory that the agent succeeds to give to the toy. The agent chooses the strategy expected by its inverse models to best reach the goal trajectory for the toy given the current context (position of the stick, toys and caregiver) and its past experience of 80000 iterations.

In most toy locations, the normalized competence of learning agents is significantly better ( $46 \%$ on average) than the normalized competence of a random agent ( $0 \%$ ). Our learning architecture thus allows to successfully reach new goals in multiple sensory spaces with multiple available strategies. Local variations reflects differences in strategy preferences and performances. For instance, where the hand cannot reach for the toy anymore, the agent still thinks this is a good strategy as it worked in a similar context (before the limit), but the hand strategy leads there to a bad performance. More training would refine the inverse models and the choice of strategy.


Figure 3: Competence after 80000 iterations. $0 \%$ means that competence to retrieve a toy there is as bad as with random agents, $100 \%$ says that agents perfectly retrieve a toy there.

## Three Strategies to Reach Toys

Fig. 4 shows the preference for the hand, tool and vocal strategies to retrieve a toy depending on the distance of the toy. In the center region, where agents can retrieve toys with all three strategies, agents choose most often the hand strategy (around $65 \%$ of the trials) and less the other two (around $15 \%$ to $20 \%$ each). In the second region, unreachable with
the hand, this strategy is still used around $50 \%$ of the trials, and the two other between $20 \%$ and $30 \%$. In the last region where the only useful strategy is to say the name of the toy so that the caregiver brings it closer, the vocal strategy is used more often: at distance 1 from center, it is used in $49 \%$ of trials, hand strategy in $38 \%$, and tool strategy in $13 \%$.


Figure 4: Strategy preferences depending on the distance of the toy. The two vertical bars shows the hand and stick limits.

## Vocal Learning with Caregiver's Feedback

The agents learn to produce vocalizations both with goal babbling and imitation of the caregivers' sounds. For each agent, three of caregiver's sounds (randomly selected) are toy names and the three others are distractors: sounds that have no special meaning for the agent. We measure errors to reproduce caregiver's sounds as the distance between the sound trajectory produced by the caregiver and the best imitation of the agent. We group the results into two categories: errors of sounds that serve as toy names and as distractors. From the 500 runs we could retrieve error data for 1482 toy names and 1482 distractors. Fig. 5 shows the distribution of errors after 80000 iterations. First, $88 \%$ of sounds have an error lower than 0.4 , and thus are successful imitations. Second, the median error for toy names is 0.23 and for distractors is 0.30 . Imitations of toy names are more accurate than of distractors: a Mann-Whitney U test gives $p<10^{-72}$. Errors distribution above 0.4 is similar for the two categories, but few toy name sounds have an error just below 0.4 compared to distractors: their distribution is shifted towards smaller errors.

## Discussion

This unified robotic model allows to study the interaction between the early development of tool use and speech. Results show that agents learn to 1) use the robotic arm to grab a toy or a stick, 2) use the stick as a tool to get a toy, 3) learn to produce toy names with the vocal tract, 4) use these vocal skills to get the caregiver to bring a specific toy within reach, and 5) choose the most relevant of those strategies to retrieve a toy, for instance preferring to use caregiver's help when the toy is out-of-reach. Interestingly, learning the production of accurate sounds for toy names was faster than for distractor sounds because inverse models often select the use of vocal-


Figure 5: Distribution of accuracy of imitations of caregivers' sounds after 80000 iterations. Below 0.4 vocal error, sounds are recognized as imitations by the caregiver. Imitations of toy names are more accurate than imitations of distractors.
izations to retrieve toys through the caregiver. Grounding vocal interaction between agent and caregiver in a play scenario thus accelerated the learning of toys' names production.

The proposed unified Model Babbling architecture does not integrate sequencing and combinatorial planning mechanisms and agents were not given initial teleological understanding of speech or tool use. However, with goal babbling and an object-based representation of task spaces, our architecture still allowed agents to learn behaviors showing a nested tool use structure, e.g. reusing movements of the stick to move a toy, or sound trajectories produced with the vocal tract so that the caregiver brings a toy. This suggests that observing infants using tools or asking for help with toys should not necessarily be interpreted as a correlate of capabilities for combinatorial sequencing and planning of actions.

It should be noted that for the agents in our model, involving the caregiver to move toys through vocalizations is a strategy with no special status with respect to the other strategies. This social interaction emerges from the same drive to refine sensorimotor models as in the learning of hand or stick movements. The production of sounds that can be understood by the caregiver as toy names to make it react and help can thus be interpreted as an emergent form of social tool use.

Those results offer a new prediction: exploration and play with objects in a grounded interaction scenario with a caregiver accelerates infant vocal learning of accurate sounds for the names associated to these objects. This hypothesis is consistent with experimental data from infant development research. First, (Clerkin, Hart, Rehg, Yu, \& Smith, 2017) shows that the objects that are frequent in the visual field of $81 / 2$ to $101 / 2$ mouth-old infants are also the objects for which infants acquire the name early. They explain that the particular distribution of object frequency in visual field can help linking the heard label to the good object in a scenario where the caregiver says the name of an object. However, this data is also consistent with our hypothesis: the most frequent objects in the visual field are the ones that the infant will most often choose goals for, and will engage caregiver's help by trying to
vocalize those toys' names. Infants could thus receive more vocal feedback for those words and learn to produce them earlier. This view also fits with recent data about the bodyobject interaction measure. In (Thill \& Twomey, 2016), the authors use a measure of the extent to which adults could easily interact with a named item and show that it predicts better the age of acquisition of the name of an item than its concreteness or imageability. In other words, the easier the interaction with an object is, the sooner its name will be acquired. Furthermore, caregiver's nonvocal feedback can also help vocal learning. Indeed, (Goldstein, King, \& West, 2003) provides evidence that a nonvocal feedback mechanism such as reacting to infant's vocalizations by smiling, or touching the infant can shape vocal babbling in real time. In our experiment, the caregiver reacts to a toy's name by giving the toy to the agent, which guides vocal learning. Such a mechanism could also be an important pathways to infant vocal development.

Our unified robotic model of speech and tool use gives a basis for future research in modeling interactions between their early development. From this study, we derived experimental predictions that could drive new experiments with infants and allow us to test and refine the model.

## References

Baranes, A., \& Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. Robotics and Autonomous Systems, 61(1).
Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., . . others (2010). Integration of action and language knowledge: A roadmap for developmental robotics. IEEE Transactions on Autonomous Mental Development, 2(3), 167-195.
Chang, M. B., Ullman, T., Torralba, A., \& Tenenbaum, J. B. (2016). A compositional object-based approach to learning physical dynamics. arXiv preprint arXiv:1612.00341.
Chen, Z., \& Siegler, R. (2000). Across the great divide: Bridging the gap between understanding of toddlers and older childrens thinking. Monographs of the Society for Research in Child Development, 65, 1, 108.
Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., \& Smith, L. B. (2017). Real-world visual statistics and infants' firstlearned object names. Phil. Trans. R. Soc. B, 372(1711).
Forestier, S., \& Oudeyer, P.-Y. (2016a). Curiosity-driven development of tool use precursors: a computational model. In 38th annual conference of the cognitive science society (cogsci 2016) (pp. 1859-1864).
Forestier, S., \& Oudeyer, P.-Y. (2016b). Modular active curiosity-driven discovery of tool use. In 2016 ieee/rsj international conference on intelligent robots and systems (iros) (pp. 3965-3972).
Forestier, S., \& Oudeyer, P. Y. (2016c). Overlapping waves in tool use development: A curiosity-driven computational model. In 2016 joint ieee international conference on development and learning and epigenetic robotics (icdlepirob) (pp. 238-245).

Gibson, K. R., Gibson, K. R., \& Ingold, T. (1994). Tools, language and cognition in human evolution. Cambridge University Press.
Goldstein, M. H., King, A. P., \& West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. Proceedings of the National Academy of Sciences, 100(13), 8030-8035.
Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. Journal of communication disorders, 39(5), 350-365.
Guerin, F., Kruger, N., \& Kraft, D. (2013). A survey of the ontogeny of tool use: from sensorimotor experience to planning. IEEE Transactions on Autonomous Mental Development, 5(1), 18-45.
Higuchi, S., Chaminade, T., Imamizu, H., \& Kawato, M. (2009). Shared neural correlates for language and tool use in broca's area. Neuroreport, 20(15), 1376-1381.
Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., \& Schaal, S. (2013). Dynamical movement primitives: learning attractor models for motor behaviors. Neural computation, 25(2), 328-373.
Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. Nature reviews neuroscience, 5(11).
Meltzoff, A. (1999). Born to learn: What infants learn from watching us. In Fox, N. \& Warhol, JG (Eds.), The Role of Early Experience in Infant Development, Skillman. NJ: Pediatric Institute Publications.
Morgan, T., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S., Lewis, H., ... others (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. Nature communications, 6.
Moulin-Frier, C., Nguyen, S. M., \& Oudeyer, P.-Y. (2013). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. Frontiers in psychology, 4.
Oller, D. K. (2014). The emergence of the speech capacity. Psychology Press.
Philippsen, A. K., Reinhart, R. F., \& Wrede, B. (2014). Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In 2014 joint ieee international conferences on development and learning and epigenetic robotics (icdl-epirob).
Stoytchev, A. (2005). Behavior-grounded representation of tool affordances. In Proceedings of the 2005 ieee international conference on robotics and automation. icra 2005.
Thill, S., \& Twomey, K. E. (2016). What's on the inside counts: A grounded account of concept acquisition and development. Frontiers in psychology, 7.
Tikhanoff, V., Pattacini, U., Natale, L., \& Metta, G. (2013). Exploring affordances and tool use on the icub. In 2013 13th ieee-ras international conference on humanoid robots (humanoids) (pp. 130-137).
Warlaumont, A. S., Westermann, G., Buder, E. H., \& Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. Neural Networks, 38, 64-75.

# What counts as math? <br> Relating conceptions of math with anxiety about math 

Ruthe Foushee (foushee@berkeley.edu)<br>Rachel Jansen (racheljansen@berkeley.edu)<br>Mahesh Srinivasan (srinivasan@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94704


#### Abstract

What do people think of when they think of "math?" We propose that individuals may have very different working definitions of the category of math, and that those with broader math conceptions may have less math anxiety. In Study 1, we introduce a method for indexing the "breadth" of individuals' math conceptions, and show that there is an inverse relation between conception breadth and math anxiety. These results suggest that math anxiety is related both to how expansive individuals perceive math to be, and how skillful they feel at the activities they think it could involve. Study 2 attempts an intervention on students' conceptions of math with a sample of middle school students. We find the same inverse relationship in students between math conception breadth and math anxiety as found in adults. We discuss ongoing work that further explores qualitative variation in math conceptions, and the lessons this may hold for intervening on math anxiety.


Keywords: math anxiety, conceptual structure, intervention

## Introduction

Recent U.S. initiatives in early science, technology, engineering, and math (STEM) highlight the growing importance of STEM education (e.g., White House Press Briefing, 2016), as well as the need for professionals in those fields to better represent the population. However, multiple barriers to an educated and diverse STEM workforce remain. One such barrier is psychological: an estimated $25-50 \%$ of U.S. college students are math anxious (Jones, 2001; Yeager, 2012), with women disproportionately affected (Hembree, 1990). Math anxiety refers to the tension or fear associated with the prospect of doing math (Ashcraft, 2002). In addition to being associated with lower math performance, math anxiety causes math-anxious individuals to generally avoid math. Given the national goal of broadening STEM participation, math avoidance might be the most devastating byproduct of anxiety about math, as it implies that math-anxious individuals will choose to end their formal math training as soon as possible.

Here, we are interested in how individuals' ideas of what math is-i.e., their math conceptions-might be a factor in their math anxiety and avoidance. "Math" can be used to refer to a wide range of activities, involving diverse skill sets and forms of reasoning. Individuals may differ in how they implicitly define the category of math, however, and properties of those definitions may be linked to their math anxiety.

Of particular interest in the present studies is what we will call the "breadth" of an individual's math conception. Guided by the idea that category structures can differentially license inferences (e.g., Ross \& Murphy, 1999), our studies test the
hypothesis that having a working math conception that is narrow (i.e., limited to a few branches of the math taxonomy, like arithmetic operations and numeric notation) might facilitate generalization of negative associations across the category. If this makes individuals confident about disliking math, rather than disliking only arithmetic or algebra, it could make them wary of future topics labeled as "math" that might have otherwise been appealing. In contrast, anxiety about the math category, and any new topics that are labeled as "math," might be harder to maintain if it encompasses many diverse subtopics and skills, ranging from the concrete (e.g., algebraic notation) to the abstract (thinking about infinity). In other words, insofar as math anxiety consists of anxiety generalized across the category of things construed as math, having a "broad" math conception may serve as a protective factor against the propagation of math anxiety.

As a first test of these ideas, we explore whether adults and children have different conceptions of what counts as math, and whether individuals with broader math conceptions may be less susceptible to math anxiety, such that math conception breadth and math anxiety will be inversely related.

## Origins of Math Anxiety

The origins of math anxiety are unclear. While research on math anxiety is motivated in large part by its impact on math performance, there is evidence suggesting the reverse direction of causation, as well. Much of this evidence comes from longitudinal studies where performance in an earlier year is more strongly correlated with math anxiety in a later year than earlier anxiety is with later performance (see Carey, Hill, Devine, \& Szücs, 2016, for a review). The relation between math anxiety and performance might be most accurately described as reciprocal, with early math difficulty leading to math anxiety, and math anxiety in turn leading to low performance, via avoidant behavior and increased constraints on processing (Carey et al., 2016).

In thinking about the relation between math conception and anxiety, we have thus far focused on a particular direction of causality, namely that narrow math conceptions might be a risk factor for developing math anxiety. But one could imagine a reciprocal relationship here, too. A child could acquire a math conception that is narrow, maybe via their early schooling, and find that they dislike or struggle with the contents of the category of math, leading them to become math anxious. Their math anxiety could in turn lead them to avoid engaging with new aspects of math that they might otherwise like or ex-
cel at, leading them to maintain both their narrow conception of math and their math anxiety.

In light of recent findings that math anxiety can be transmitted between generations, it is just as important to alleviate math anxiety in adults (i.e., so that they don't transmit it to children) as it is to intervene directly in children. Prior work has found that teachers' math anxiety may "spread" to their students (Beilock, Gunderson, Ramirez, \& Levine, 2010). This is especially problematic because aspiring teachers with math anxiety tend to gravitate toward teaching earlier grades, where they will be able to engage less with math (see Hadley \& Dorward, 2011), but where they will also be interacting with students in the early school years, when children are most impressionable. Parents' math anxiety can also affect their children. In one study, children of math anxious parents learned less during the school year than did children of non-math-anxious parents-but only if these parents frequently gave their children homework help (Maloney, Ramirez, Gunderson, Levine, \& Beilock, 2015). Given this evidence for the intergenerational transmission of math anxiety, our studies focus both on adults and children.

## Relating Math Conceptions and Math Anxiety

In principle, individuals could have 'math conceptions' that range from narrow (Math is the symbolic operations one learns in school) to broad (Math relies on logic, spatial reasoning, and pattern recognition).

Here, we develop a new measure to characterize the breadth of math conceptions. This measure presents participants with a diverse list of activities or topics, ranging from "sewing" to "playing soccer" or "physics." Participants are asked to indicate whether each item "could involve math," and, in some cases, to explain why. The idea is that when asked to answer whether a given activity "could involve math," individuals will be encouraged to come up with some rationale for how it could or could not involve math, and that their flexibility in categorizing activities as "math" will depend on the breadth of their (implicit) definition of the category. The point here is not that individuals typically construe an activity like "playing soccer" as involving math. Instead, our interest is in whether individuals vary in how flexible they are in categorizing activities that are not conventionally thought of as "math" as involving math. We present a diversity of activities to math, that can be related to math via diverse aspects of mathematical reasoning or subtopics, thus revealing the capacity and/or bounds of an individual's math conception. If an individual's conception of math is itself broad and diverse, we expect that it will be able to support explanations for the math-involvement of a wide range of activities. We thus operationalize breadth of math conception in the following studies as the number of activities that individuals say "could involve math." In Study 1, we also ask participants to rate their own skill at these same activities.

We hypothesize that broader math conceptions will relate to lower math anxiety in that they will afford individuals with more opportunities to recognize their own math engagement
or expertise, and dilute the negative impact of components of math that individuals have negative associations with. Related to this, we expect self-assessed skill with activities classified as involving "math" to mediate the proposed relation between conception breadth and math anxiety. Study 1 examines the relation between math conception and math anxiety in adults, taking subjective skill into account. Study 2 tests for the same relation in middle-school children, within the context of a intervention study that tests the effect of broadening math conceptions.

## Study 1: Adult Math Conception \& Anxiety

Study 1 examined the relation between math conception and math anxiety in adults via an online survey composed of seven counterbalanced blocks probing participants' math attitudes and associations.

## Stimuli \& Methods

Participants A total of 62 U.S. adults were recruited via Amazon's Mechanical Turk ( 31 female, 19-74 years, $M=$ $33.24, S D=10.25$ ). Participants were compensated for their participation, and the study took approximately 15 minutes to complete.
Math Conception In one block, participants saw a randomized list of topics and activities (e.g., "architecture," "cooking," "exercising"). Participants were asked to indicate whether. . e each activity or topic listed involves math or does not involve math. They responded by dragging each item into one of three boxes, labeled "Math," "Not Math," and "Not Sure." The more items categorized as involving math, the broader we considered their math conception to be (see above). We included the item "Math" as a control.
Activity Skill In another block, participants saw the same items in a new randomized order, and rated their skill at each item (How good would you say you are at each of these things?). They responded on a five-point Likert scale from 'Not at all good' to 'Very good.' We included a control item (For this question, respond 'Good'), as well as an opt-out scale option ('NA') for participants who had no experience with the item.
Math Anxiety We assessed participants' math anxiety using the single item math anxiety scale (SIMA; Núñez-Peña, Guilera, \& Suárez-Pellicioni, 2013). This measure asks simply, On a scale from 1 to 10, how math anxious are you? The SIMA has been validated on a large sample of U.S. college students. It shows the expected negative correlation with math achievement measures, high test-retest reliability, and is consistent with lengthier, established measures of math anxiety, like the Shortened Math Anxiety Rating Scale (sMARS; Alexander \& Martray, 1989, $r=.77$ ).
Other Measures We collected several other measures of participants' attitudes toward and history with math. One block assessed participants' "math mindset:" an analogy
to intelligence mindsets made specific to math (Yeager \& Dweck, 2012). Five items probed participants' beliefs about the fixedness of math ability (e.g., Math is a gift: you either have it or you don't.), which they responded to using a fivepoint Likert scale of agreement. Two blocks consisted of a single, open-ended question, one asking participants for an informal definition of math (Please describe what you think math is in the space below), and one eliciting their personal math history (Please write a brief summary of your experience with math from childhood until now). In the final block, we collected demographic information, including the number of semesters of college they had completed, and a list of all math classes they had taken.

## Results \& Discussion

Qualtitative Variation in Math Conceptions There was substantial variation in the activities that participants categorized as involving math (Figure 1). All participants appropriately responded that "Math" involved math, which we took as confirmation of their attention to the task. Items obviously involving math were categorized as such by the vast majority of participants (e.g., finance), while those representing related disciplines (e.g., biology), daily activities (e.g., cooking), and abstract, creative and language-related tasks (e.g., composing music, reading) received the fewest math-categorizations. In a separate study, we elicited explanations for participants' categorizations of a similar list of items. In that study, both adults and children frequently used contrast categories (e.g., "No, that's music!"), often from the humanities, to explain why items could not involve math. This type of explanation implies that participants perceived the categories of music, art, and even science as exclusive with math. Such a picture of what math is (and isn't) is consistent with the idea of a narrow math conception, and echoes what mathematician Paul Lockhart famously lamented as the sorry byproduct of American math education:

The first thing to understand is that mathematics is an art. The difference between math and the other arts, such as music and painting, is that our culture does not recognize it as such. [...] Nevertheless, the fact is that there is nothing as dreamy and poetic, nothing as radical, subversive, and psychedelic as mathematics. It is every bit as mind-blowing as cosmology or physics (mathematicians conceived of black holes long before astronomers actually found any), and allows for more freedom of expression than poetry, art, or music (which depend heavily on properties of the physical universe). Mathematics is the purest of arts as well as the most misunderstood. (Lockhart, 2009).

Math Conception \& Anxiety To answer whether breadth of math conception and math anxiety are related, we conducted a linear regression on individuals' math anxiety and the number of items they categorized as math, controlling for the number of semesters of college they had completed.

Table 1: Descriptive statistics for four blocks in Study 1. 'Items Categorized as Math' is out of a total of 32, and was analyzed as a proxy for the breadth of participants' math conceptions. 'Math Anxiety' is on a 10-point self-report scale. 'Self-Assessed Skill' represents the mean skill rating on a 5point Likert scale, across all items for all participants. 'Math Mindset' is coded to be on a 5-point scale indexing how fixed individuals believe math ability to be, with larger values indicating more fixed mindsets.

| Variable | $M$ | $S D$ |
| ---: | ---: | ---: |
| Items Categorized as Math | 13.10 | 5.35 |
| Math Anxiety | 4.44 | 3.04 |
| Self-Assessed Skill | 3.28 | 0.44 |
| Math Mindset | 2.13 | 0.99 |

In accordance with our predictions, math anxiety was negatively related to the number of items participants categorized as math, even controlling for education $(F(1,61)=6.44$ $p<.05$ with an $R^{2}$ of .082 ; see Figure 2). This supports the idea that individuals with broader math conceptions are less likely to experience math anxiety, and that this relation may not be attributable to exposure to topics in math alone.

To address whether the relation between math conception and anxiety is due in part to individuals' perception of their own skill at things they think might involve "math," we analyzed self-assessed skill and anxiety. For each individual, we took the mean skill of the items they had categorized as involving math and those they had categorized as not involving math. We dropped items for which participants reported having had no experience. A linear regression on self-reported skill and math anxiety revealed a significant negative correlation between math anxiety and mean self-assessed skill for items the individual was able to relate to math ( $\beta=-1.98$, $S E=0.60, t=-3.29, p<.01$ ), but no correlation between math anxiety and self-assessed skill for items judged to not involve math ( $\beta=0.11, S E=0.69, t=0.154, p=.88$ ). This asymmetry is important because it suggests that it is not just individuals who are less confident overall who suffer from math anxiety-if this were the case, we would have expected to find that lower skill related to higher anxiety for both items judged to involve math and items judged to not involve math.

In Study 1, both the number of items construed as involving math and participants' perceived skill at those items were related to math anxiety. As discussed above, one of the most dangerous features of math anxiety is its tendency to make individuals avoid math and thus fail to take advantage of opportunities to discover new aspects of mathematics they might excel at or appreciate. The fact that mean self-assessed skill at activities categorized as involving math was negatively related to math anxiety lends support to the idea that broad conceptions may be a protective factor in math anxiety, attenuating the impact of negative associations that individuals might have with activities they think could involve 'math.' Having a


Figure 1: Number of participants in Study 1 who labeled each activity as involving math.


Figure 2: Plot of linear regression line showing relationship between breadth of conception and math anxiety in Study 1, controlling for education ( $\alpha=6.93, \beta=-0.18, p<.05$ ).
broad math conception does not mean that an individual has to feel confident and have positive associations with all activities that they think involves math, but it could mean that negative associations with specific topics (like geometry or algebraic notation) will have less of an impact on their associations with the category as a whole.

## Study 2: Middle School Intervention

We were interested in whether students would exhibit the same qualitative variation in math conceptions and link between breadth and anxiety that we had seen with adults in Study 2. Additionally, as a first pass at investigating the causal relation between math conception and anxiety, and potential educational implications, we designed a brief intervention intended to broaden students' math conceptions.

## Stimuli \& Methods

Study 2 consisted of an interactive origami activity followed by four measures administered to participants in two
between-subjects conditions, BASELINE and Broad. Only participants in the BROAD condition received an explanation for the ways in which the activity had involved math before completing the other assessments.

Participants A total of 80 6th, 7th, and 8th grade students at a school in Gujarat, India participated (33 6th-graders, 7 girls; 21 7th-graders, 9 girls; 26 8th-graders, 9 girls). All 6thgrade participants were excluded for sharing answers ( $n=$ 33), leaving 477 th- 8 th grade students in our sample. Participants were tested in groups of $10-15$ assigned to the BASELINE or BROAD conditions in a classroom at their school.

Origami Activity Students sat in a circle on the floor around two experimenters who guided them through folding an origami crane. A third experimenter circulated to answer any questions, and students could also refer to printed, diagrammatic instructions distributed before the activity. All experimenters avoided using explicit math language during the folding instruction (e.g., reference to "angles," "half," "diagonal"), opting instead for generically narrated demonstration (e.g., "fold the paper like this"). Each student folded a paper crane, which they got to take home.

Construal Following the origami activity, students in both conditions answered whether the activity they just did could involve math (Yes/No/Not Sure), and to explain why. In addition, they rated how enjoyable and difficult they had found the activity, on a five-point Likert scale (from 'Not at all-' to 'Extremely-').

Intervention In the Broad condition-but not in the BASELINE condition-an experimenter then gave a brief explanation of how the origami activity involved and related to math (e.g., ... you have to think about spatial relations, and things like measurements of the different sides and angles. When designing new pieces of origami, you have to think creatively and flexibly, and use what you already know to come to new conclusions, like you have to do in math).

Avoidance The next measure participants completed was intended to indirectly access their math avoidance. The survey consisted of 6 items, each asking about a different school subject (e.g., How excited are you to learn a new topic in [math/Hindi] class?). Participants responded on a 5-point scale (from 'Not at all excited' to 'Extremely excited'). ${ }^{1}$

Math Anxiety We administered a child math anxiety questionnaire adapted from Ramirez, Gunderson, Levine, and Beilock (2013) by Barner et al. (2016), for use in India. The questionnaire consisted of 16 questions regarding students' experiences with math, which students responded to using a 5-point face scale (from 'Not nervous at all' to 'Very, very nervous'). The experimenter explained the scale and completed three warm-up questions with the students beforehand

[^329]

Figure 3: Number of participants in Study 2 who answered "yes" when asked whether each item could involve math.

Table 2: Means and standard deviations for each condition.

|  | Concept |  | Anxiety |  |
| ---: | ---: | ---: | ---: | ---: |
| Condition | $M$ | $S D$ | $M$ | $S D$ |
| BASELINE | 20.96 | 7.24 | 1.78 | 0.47 |
| BROAD | 24.61 | 5.96 | 1.60 | 0.36 |

to ensure understanding of the measure.
Math Conception The math conception measure was a variant of the one used in Study 1. We included 40 age- and place-appropriate items, and adjusted the wording used in the prompt from Study 1. Here, participants answered Could this activity involve math? (Yes/No/Not Sure), which we anticipated would encourage flexible thinking about the items and about math.

## Results \& Discussion

Qualitative Variation in Math Conceptions Participants indicated that an average of 22.74 out of 40 items could involve math. As in Study 1, there was considerable variation across items in the proportion of participants who judged them as involving math (Figure 3).

Math Conception \& Anxiety Participants received an average math anxiety score of 1.69 (out of 5 ). We were interested again in whether math anxiety scores were related to conception breadth, which we examined in our total sample, collapsing across condition. In middle-schoolers, as with adults, math anxiety was negatively related to the number of activities students categorized as math, $(F(1,44)=4.15$, $p<.05$ with an adjusted $R^{2}$ of 0.07 ; see Figure 4).

Conception Intervention We next analyzed math conception and math anxiety for our two conditions separately. If such a brief intervention were successful, we should expect conception scores to be higher in the Broad condition, and anxiety scores to be lower. While conception and anxiety


Figure 4: Linear regression showing relationship between breadth of conception and math anxiety in Study $2(\alpha=2.10$, $\beta=-0.02, p<.05)$.


Figure 5: (a) Boxplot of anxiety scores by condition. (b) Boxplot of conception score by condition.
might be slightly different in the anticipated direction between the two conditions (Table 2), the differences between group means is not significant (as determined by one-way ANOVAs for math conception: $F(1,45)=3.54, p=.066$, and anxiety: $F(1,45)=2.31, p=.14)$. The trend for math conceptions in particular is promising (see Figure 5): in the BaSELINE condition, there was more spread in the magnitude of participants' conception scores, while those in the Broad condition had generally 'broader' conceptions. Thus, it may be that with a different or merely more sustained intervention, students' math conceptions could be broadened.

Influence of Construal Out of the 47 participants analyzed, 36 said that the origami activity could involve math. Participants on average enjoyed the activity $(M=4.43, S D=$ 0.62 ) and did not find it difficult ( $M=2.28, S D=0.71$ ). This raises the possibility that we may not have found a robust intervention effect because our elicitation of construals of the origami activity as math itself served as an intervention on breadth of conception. In particular, given that all participants-including those in the BASELINE condition-
were asked to consider whether an enjoyable and easy activity could involve math before completing any of the surveys, they may have been primed to think more broadly and favorably of math.

## Discussion \& Future Directions

The above studies offer preliminary evidence for the intuition that individuals may have substantially different ideas of what constitutes math. Here, we have introduced the idea of math conceptions to describe these qualitatively different definitions of the category of math, and focused especially on their "breadth" to explain why certain types of math conceptions might make math anxiety more or less likely. Strikingly, the measure we introduced as a proxy for the breadth of individuals' math conceptions showed the hypothesized inverse link to math anxiety, in both adults and children, though it should be noted that the samples for Studies 1-2 differed in more than age. We see the remarkable dissimilarity of the two populations and contexts as adding strength to our results.

While this link between our measure of math conception breadth and math anxiety is promising, we imagine there is a great deal of additional variation among math conceptions that could be captured in future studies. Eliciting and analyzing participants' explanations for their categorization decisions may be one especially fruitful way to access other qualitative dimensions of math conceptions, alongside canonical methods to access category structure, like primed similarity judgments.

Without robust evidence for the efficacy of our intervention (Study 2), we cannot speak to the potential directionality of the math conception-anxiety relationship. Our ongoing work is exploring this question through an interactive intervention on adults' math conceptions, as well as an adaptation of the math conception measure for use with young children prior to being formally educated in math. Exploring math conceptions in young children, as well as directly assessing math skill in future studies with adults, will also address the heretofore unconsidered possibility that a third variable (like actual proficiency in math) is responsible for both responses on our current conception measure and levels of math anxiety. The ultimate goal of these lines of research is to understand and describe the character of individuals' implicit math categories, and leverage this knowledge to inform interventions aimed at reducing math anxiety in adults and children.

## Acknowledgments

We would like to thank the Amrit Vidlaya school in Gujarat, India for their generosity and enthusiasm about our project and Catherine Berner for assistance with data collection.

## References

Alexander, L., \& Martray, C. (1989). The development of an abbreviated version of the Mathematics Anxiety Rating Scale. Measurement and Evaluation in Counseling and Development, 22, 143-150.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. Current Directions in Psychological Science, 11(5), 181-185.
Barner, D., Alvarez, G., Sullivan, J., Brooks, N., Srinivasan, M., \& Frank, M. C. (2016). Learning mathematics in a visuospatial format: A randomized, controlled trial of mental abacus instruction. Child Development, 1-13.
Beilock, S. L., Gunderson, E. A., Ramirez, G., \& Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. Proceedings of the National Academy of Sciences of the United States of America, 107(5), 18601863.

Carey, E., Hill, F., Devine, A., \& Szücs, D. (2016). The chicken or the egg? The direction of the relationship between mathematics anxiety and mathematics performance. Frontiers in Psychology, 6, 1-6.
Hadley, K. M., \& Dorward, J. (2011). The relationship among elementary teachers' mathematics anxiety, mathematics instructional practices, and student mathematics achievement. Journal of Curriculum and Instruction, 5(2), 27-44.
Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. Journal for Research in Mathematics Education, 21, 33-46.
Jones, W. G. (2001). Applying psychology to the teaching of basic math: A case study. Inquiry, 6(2), 60-65.
Lockhart, P. (2009). A mathematician's lament: How school cheats us out of our most fascinating and imaginative art form. Bellevue Literary Press.
Maloney, E. A., Ramirez, G., Gunderson, E. A., Levine, S. C., \& Beilock, S. L. (2015). Intergenerational effects of parents' math anxiety on children's math achievement and anxiety. Psychological Science, 26(9), 1480-1488.
Núñez-Peña, M. I., Guilera, G., \& Suárez-Pellicioni, M. (2013). The single-item math anxiety scale (SIMA): An alternative way of measuring mathematical anxiety. Journal of Psychoeducational Assessment, 20(10), 1-12.
Ramirez, G., Gunderson, E. A., Levine, S., \& Beilock, S. (2013). Math anxiety, working memory, and math achievement in early elementary school. Journal of Cognition and Development, 14, 187-202.
Ross, B. H., \& Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. Cognitive Psychology, 38, 495-553.
The White House. (2016). STEM for all. Retrieved from https://obamawhitehouse.archives.gov/blog/2016/02/11/stemall.
Yeager, D. S. (2012, April). Productive persistence: A practical theory of community college student success. paper presented at the annual meeting of the American Educational Research Association. Vancouver, Canada.
Yeager, D. S., \& Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. Educational Psychologist, 47, 302-314.

# After braking comes hasting: reversed effects of indirect associations in $2^{\text {nd }}$ and $4^{\text {th }}$ graders 

Nicole Franke (nicole.franke@fernuni-hagen.de)<br>General Psychology (Judgement, Decision Making), Fernuniversität Hagen, Germany<br>General and Biological Psychology, University of Wuppertal, Germany<br>André Roelke (roelke@uni-wuppertal.de)<br>General and Biological Psychology, University of Wuppertal, Germany<br>Ralph R. Radach (radach@uni-wuppertal.de)<br>General and Biological Psychology, University of Wuppertal, Germany<br>Markus J. Hofmann (mhofmann@uni-wuppertal.de)<br>General and Biological Psychology, University of Wuppertal, Germany


#### Abstract

The Associative Read-Out Model (AROM) suggests that associations between words can be defined by the log likelihood that they occur together more often in sentences than predicted by their single-word frequency. Moreover, semantic relations can be defined by associative spreading across many common associates. Here, we addressed developmental effects of associative and semantic priming. Thus, we manipulated sentence-co-occurrencebased direct (syntagmatic) and common (paradigmatic) associations between prime and target words in $2^{\text {nd }}$ and $4^{\text {th }}$ graders. Syntagmatic associations decreased response times and error rates in both, $2^{\text {nd }}$ and $4^{\text {th }}$ graders. Paradigmatic associations increased errors rates in 2nd graders, whereas they decreased errors rates in 4 th graders. These results suggest that $2^{\text {nd }}$ graders profit from syntagmatic, i.e. contiguity-based associations, while a benefit from paradigmatic-semantic relationship probably develops from generalizing across many of these simple associations.


Keywords: Interactive Activation Model, Associative Read-Out Model, Semantic Priming, Computational Models, SyntagmaticParadigmatic shift

## Introduction

Starting with pioneer work of Meyer and Schvaneveldt (1971) a wide range of studies revealed that a target word (e.g. "chair") is processed faster and more accurate when a semantically associated prime word (e.g. "table") was presented before (e.g. Bentin et al., 1985; Neely, 1976). Interestingly, studies differentiating between various types of associations (e.g. Becker, 1980; Lucas, 2000; McNamara, 2005) and/or individual differences like age (e.g. review Chapman et al., 1994, McCauley et al., 1976) revealed inconsistent results with regard to the size and direction of the semantic priming effect. From a developmental perspective, the presence (or absence) of the semantic priming effect may be an indicator of the development and organization of semantic knowledge (e.g. Lucas, 2000; McCauley et al., 1976; Meyer \& Schvaneveldt, 1971). On the one hand, recent research revealed greater semantic priming effects (i.e. greater difference between primed and
non-primed condition) for younger children and elder people (see review Chapman et al., 1994). On the other hand, various studies, investigating processes of different types of relations in semantic priming tasks, revealed that younger children show priming effects if words are directly associated only, and not if they exclusively provide a category relation (e.g. McCauley et al., 1976). So far, empirical evidence points towards a greater facilitation by functional/associative relations in comparison to pure semantic relations in children, but also an increasing sensitivity for thematic and taxonomic relationships over age (Arias-Trejo \& Plunkett, 2013). In line with this, younger children tend to freely associate words that have a syntagmatic relation from mere common occurrence in sentences relation (e.g. "good" and "boy") rather than a paradigmatic relation due to the same form class (e.g. "good" and "bad") in comparison to older children and adults (e.g. White, 1985; Woodrow \& Lowell, 1916). This effect is also known as the syntagmatic-paradigmatic shift that occurs in an age range between 5 and 9 years (e.g. Brown \& Berko, 1960; Entwisle, 1966, Nelson, 1977).

To our knowledge, however, the relative reliance on syntagmatic and/or paradigmatic information has not yet been addressed during visual word recognition. As German children start reading at the age of 6 , we hypothesized that word-decoding abilities sufficient for syntagmatic effects should be apparent around the age of 7 years, i.e. in the $2^{\text {nd }}$ grade, while a stronger reliance on paradigmatic information should be observable around the age of 9 , i.e. in the $4^{\text {th }}$ grade. So far, most semantic priming studies used free association performance of adults to predict semantic priming (e.g., Lucas, 2000), though already Jung (1905) stated that free associations are diagnostic for interindividual differences. Therefore, it is questionable whether its usage as an independent variable for the study of children is appropriate (see Hofmann \& Jacobs , 2014). To derive a semantic long-term memory structure from
experience with a sample of text (Hofmann et al., 2011), a recent interactive activation model (IAM; McClelland \& Rumelhart, 1981) relies on co-occurrence statistics. The Associative Read-Out Model (AROM, Hofmann et al., 2011) is the first IAM with an implemented semantic layer. It defines two words as associated, if they co-occur more often together in sentences than predicted by their single occurrence frequency (Dunning, 1993; Quasthoff et al., 2006). Thus, it reflects a symbolic Hebbian learning approach (Hebb, 1949) by suggesting higher association strengths for words that are occurring more often together than predicted by the frequency-driven orthographic activation.

The AROM already successfully predicted behavioral and electrophysiological data for tasks, in which memory is explicitly required. Hofmann et al. (2011) showed that the correct identification of studied words as well as the false recognition of non-studied words is significantly higher for words with many associations in a recognition memory task. This result has recently been extended by Stuellein et al. (2016) in an EEG study by showing significant response time, P200 and N400 effects for words with many associations. It was an open question, however, to what extend these results were induced by pure direct associations and/or indirect associations like semantic feature overlap (Stuellein et al., 2016). As the AROM defines words as associated by the frequency of their common occurrence, it is in line with localist theories proposing direct associative links between symbolic representations to capture the meaning of a word (e.g., Anderson, 1983; Collins and Loftus, 1975). Whereas distributed models define the meaning of a word by a distribution across subsymbolic 'hidden' units (e.g. McClelland and Rogers, 2003), this assumption is in line with other co-occurrence based models, defining the meaning of a word by latent factors determining with which words they co-occur (e.g. Landauer and Dumais, 1997). In the tradition of distributed models, one can assume that words that often occur together in similar sentence contexts might share similar semantic features. In line with the idea, that the meaning of a word is determined by its surrounding context (Firth, 1957; Harris, 1951), common associates of two words can possibly be considered as common features (Hofmann and Jacobs, 2014). As a consequence, a more complex AROM, that would be able to simulate the dynamic co-activation of such semantic features, was discussed to be a plausible option to accommodate both perspectives (Stuellein et al., 2016).

In a recent study, Roelke et al. (2016, subm.) also tested the AROM in an implicit memory task. During primed lexical decision, a full factorial manipulation of direct association (strong/no) and the number of common associates (many/no) of prime and target revealed strong effects in adult participants. Prime and target words with direct and many common associates facilitated visual word recognition. In contrast, we also have preliminary evidence
of inhibitory priming effects at a very long SOA (Schmidt, 2015). These results are in line with recent studies, showing only facilitating effects for pure associative relations and inhibitory or facilitatory effects for semantic relations that are dependent on the time that the prime is processed (e.g. see Plaut \& Booth, 2000).

## The present study

To investigate whether the AROM can be used to address the syntagmatic-paradigmatic shift by relying on direct (associative, syntagmatic) and indirect (semantic, paradigmatic) relations, we tested $2^{\text {nd }}$ and $4^{\text {min }}$ grade students from two German elementary schools, using primed lexical decision. In line with recent results, we expected smaller semantic priming effects for younger children for semantic (indirect) relations. Furthermore, we expected greater semantic priming effects for associative (direct) relations in comparison to semantic (indirect) relations for all children, because these depend on an abstraction of experience-based knowledge.

## Methods

## Subjects

For all participating students, parents signed written consent in advance.

Second grade. Behavioral data were collected for $952^{\text {nd }}$ grade students of two primary schools in Solingen, Germany. Two students did not complete the experiment due to the task difficulty. Another eleven participants had to be excluded because of reading/writing disorders and two participants because of lacking German skills. The mean age of the remaining 75 students (female $=45$ ) was 7.46 years ( $S D=.502$ ). According to their parents, 62 ( $82.7 \%$ ) participants had learned German as their first language. Three children (4\%) learned Turkish as their first language, followed by Italian ( $\mathrm{N}=2,2.7 \%$ ) and Russian ( $\mathrm{N}=2,2.7 \%$ ). The remaining students came from a variety of linguistic background. For one student, data for the native language was missing.

Fourth Grade. Behavioral data were collected for $864^{\text {th }}$ grade students of two primary schools in Solingen, Germany. Ten students had to be excluded because of reading/writing disorders. The mean age of the remaining 76 students (female=52) was 9.54 years $(S D=.738)$. According to their parents, 62 participants ( $81.6 \%$ ) had learned German as their first language. Five children ( $6.6 \%$ ) learned Turkish as their first language, followed by Italian $(\mathrm{N}=4,5.3 \%)$ and Polish ( $\mathrm{N}=2,2.6 \%$ ). The remaining students came from a variety of linguistic background.

## Materials

Corpora. The word stimuli were taken from the word corpus "childLex", which is based on approximately 5000 German books for children between 6 and 12 years (Schroeder et al., 2015; status: September 2014). The books vary in length and content with about 5000 to 15,000 words per book. We used words that were among the list for 6 to 10 year old children. As the childLex corpus is not openly available for analyses, co-occurrence statistics were taken from the German corpus of the "Wortschatz" project (status: December 2006; Quasthoff et al., 2006). This corpus is largely composed of online newspaper (1992-2006). Based on 800 million tokens and 43 million sentences, two words were considered to be directly associated when they co-occurred more often together in sentences than predicted by their single occurrence frequency (Dunning, 1993). Indirect associations were defined as the number of common direct associates.

Stimuli. The stimulus set consisted of 160 primes and 160 targets. The 160 primes and 80 targets were German nouns. The remaining 80 targets consisted of 40 pronounceable pseudowords and 40 random letter strings. Pseudowords were created by changing one to three consonants of real nouns. 80 targets were German nouns that were split into four word conditions in a $2 \times 2$ design with the factors direct association (high vs. low) and indirect relation (high vs. low). Prime and target were considered to provide a low direct association, when they were not associated at all (association strength $=0$ ) and as high directly associated, when they were beyond a $2,5 \%$-quantile criterion (association strength > 3) of all possible stimuli ( $\mathrm{N}=6,975$; cf. Hofmann et al., 2011, for a formal definition of association strength). They were considered to provide a low indirect relation, when they had less than 65 common associates (below a $2,5 \%$-quantile criterion of all possible stimuli) and were considered to provide a high indirect association, when they had more than 300 common associates (beyond a $2,5 \%$-quantile criterion of all possible stimuli; cf. Bordag, 2007, for counts of common associates). From the childLex corpus, the word features frequency, word length and Orthographic Levenstein Distance (Yarkoni et al., 2008) were counterbalanced between the four word conditions for prime and target words to rule out confounding effects (condition differences $\mathrm{p}>.05$ ). Raw Lemma-Frequency was $\log 10$ transformed and words below and beyond a $2,5 \%$-quantile frequency criterion of all possible stimuli were excluded. Word and nonword length was limited from 3 to 6 letters. Before counterbalancing to rule out confounding variables, a manual examination of the stimulus set excluded inappropriate words for children (e.g. those with sexual content), prime and target pairs with the same first letter and compounds (e.g. "snowball").

## Procedure

Cover story. The instruction was embedded in a cover story, adapted from a children's lexical decision task by Richter et al. (2013). Children were asked to help an extraterrestrial named Reli, who came to earth to learn the language of the earthlings, to distinguish between real words and nonwords (Target). To further explain the appearance of the prime words, students were told that another extraterrestrial named Gudra also wanted to learn the language of the earthlings (Prime). Students were told that other children were helping Gudra, so that they had to read her word but that they did not have to react.

Semantic Priming Task. The semantic priming task was performed by groups of eight to ten students at the same time in a quiet room, separated from the rest of the class. Before the experiment started, the time course of the experiment was written on the blackboard and the task was explained to the children in front of the class. Each student worked on his/her own on a separate laptop. Students were asked to put on headphones and to leave on the headphones during the whole task. Before the task started, a detailed instruction was presented once more in a videoclip with the extraterrestrial Reli.

First a fixation cross was presented for 1000 ms on the screen. Then a prime word was presented in grey letters for 600 ms . The students were asked to read the prime but not to press a button. After the prime word, a blank screen appeared for 200 ms , after which the target word was presented in black letters. Students were asked to press a green button with their right forefinger on the keyboard ("K"), if the presented stimulus was a real word and to press a red button with their left forefinger on the keyboard ("D") if the presented stimulus was a nonword. The target word stayed on the screen until the student pressed one of the two buttons. Following another blank screen for 500 ms , the word "Bereit?" ('ready?') was presented in red letters on the screen and students were asked to press a yellow button ("space") with one of their thumbs on the keyboard, if they wanted to go to the next trial (s. Figure 1).
To get used to the task, five exercise trials were presented at the beginning. For the exercise trials, feedback was provided whether the response was correct or not. For the main task, no feedback was provided. During the main task, two breaks were included, each after 56-57 trials. The students decided on their own by pressing the yellow button when to continue with the main task.


Figure 1: Time course of the experiment
At the beginning of the main task and after every break two "icebreaker trials" were included, that were excluded from data analyses. For every participant, the order of the presented prime-target pairs was randomized. Students were asked to react as fast and as accurate as possible.

Data Analysis. Results were analyzed using general linear mixed-effect models with the fixed effects grade ( $2^{\text {nd }}$ vs. $4^{\text {th }}$ class), direct association (low vs. high) and indirect association (low vs. high), their interaction terms and the random intercepts subject and item. The dependent variables were accuracy and response times. For accuracy analysis we used binary logistic regression and for response time analysis we used linear model. Because the degrees of freedom are not exactly known in LMM analyses, we chose 2 standard errors as significance criterion (i.e. $\mathrm{t}>=2$; cf . Baayen et al., 2008, footnote 1; Masson \& Kliegl, 2013). Incorrect responses and those plus/minus a 3 standard deviation criterion from average for each subject and condition were excluded from response time analyses. We only report main effects and interactions between the experimental factors that are significant. When models revealed significant interactions between at least two of the experimental factors, post-hoc t-tests were conducted.

## Results

## Accuracy

Grade and direct association and the significant interactions grade*direct association and grade*indirect association led to significant contributions to the model (all $t^{\star} \mathrm{s}>=2$ ). The positive effects of grade ( $\beta=1.273, t=7.08, S E=.18$ ) and direct association ( $\beta=1.191, t=3.12, S E=.382$ ) indicate that direct associations increased accuracy, and that $4^{\text {th }}$ grade students made fewer mistakes than $2^{\text {nd }}$ grade students (s. also Figure 2). The analysis also revealed a significant interaction between direct associations and grade ( $\beta=-$ $0.360, t=-2.95, S E=.122$ ). Moreover, we obtained an interaction of indirect association and grade ( $\beta=-0.395$, $t=-3.33 S E=.119$ ).
Post-hoc t-tests revealed that for $2^{\text {nd }}$ graders words that were
high directly associated ( $M=3.50, S D=1.90$ ) led to fewer errors ( $t=-9.89, p=.000$ ) than words that were low directly associated ( $M=5.71, S D=2.60$ ). For $4^{\text {th }}$ graders high directly associated words ( $M=1.74, S D=1.24$ ) led also to significantly fewer errors ( $t=-12.32, p=.000$ ) than words that were low directly associated ( $M=3.95, S D=2.08$ ).
Furthermore, for $2^{\text {nd }}$ graders words with many common associates ( $M=4.84, S D=2.29$ ) led to significantly more errors ( $t=2.481, p=.015$ ) than words with few common associates ( $M=4.37, S D=2.16$ ). Whereas for $4^{\text {th }}$ graders words with many common associates ( $M=2.68, S D=1.82$ ) led to significantly fewer errors ( $t=-2.116, p=.038$ ) than words with few common associates ( $M=3.01, S D=1.48$ ).

## Response time

Grade and direct association led to a significant contribution to the model (all $t$ 's>=2). The positive effects of grade $(\beta=1.103, t=8.99, S E=.12)$ and direct association ( $\beta=0.19, t=3.06, S E=.08$ ) indicate that $4^{\text {th }}$ grade students responded faster than $2^{\text {nd }}$ grade students and in general students responded faster for words that were directly associated (s. Figure 3).


Figure 2: Mean accuracy (error rates) in $2^{n d}$ and $4^{\text {th }}$ grade students
Note: Error bars are standard errors.


Figure 3: Mean response times in $2^{n d}$ and $4^{\text {mit }}$ grade students Note: Error bars are standard errors.

## Discussion

To test whether the AROM can account for a developmental shift from associative-syntagmatic (direct) to semanticparadigmatic (indirect) relations during visual word recognition, we analyzed the performance of $752^{\text {nd }}$ grade and $764^{\text {n }}$ grade children of two German elementary schools in a semantic priming task. Direct (syntagmatic) associations decreased errors in $2^{\text {nd }}$ graders as well as in $4^{\text {th }}$ grade students. The analysis of indirect (paradigmatic) relations revealed a significant interaction of grade and paradigmatic associations: while paradigmatic associations led to inhibitory effects in $2^{\text {nd }}$ graders (more errors), they led to facilitating effects in $4^{\text {th }}$ graders (fewer errors). Thus, children may develop the ability to generalize across common associations between the second and the fourth grade.

Our results fit into a reading development model, in which category knowledge is gradually abstracted and develops from functional, event-based knowledge. Response differences may result from the addition from new structures within an associative network, instead of a complete reorganization (e.g. McCauley et al., 1976). Consistent with this, Nelson (1977) assumed that children first represent semantic knowledge as spatial or temporal scripts (e.g. "eating lunch") and gradually abstract and define categories from this script-based knowledge. Our results also show that the Associative Read-Out Model (Hofmann et al., 2011) is sufficient to define both, associative-syntagmatic and semantic-paradigmatic perspectives by co-occurrence statistics, and thus provides a computational window into developmental effects of visual word recognition. Future more explicit simulations with an

AROM thus may capture individual differences such as age by differential associative excitation and inhibition scaling parameters within the semantic representation layer.

The syntagmatic-paradigmatic shift in children is well known (e.g. Entwisle, 1966). De Saussure's (1959) coined the term "syntagmatic" as an associative relation between words that typically co-occur in a linear combination (cf. Hofmann \& Jacobs, 2014). He further proposes a second type of relation, i.e. that words are associated when "they have something in common" (1959, p. 123). In computational linguistics, the number of common associates is used to define paradigmatic relations: "For example, the semantic similarity of the words red and blue can be derived from the fact that they both frequently co-occur with words like color, flower, dress, car, dark, bright, beautiful, and so forth" (Rapp, 2002, p. 1). We think that simple withinsentence co-occurrence provides an intelligible, transparent and performance-independent explanation of differential effects during reading development.

We are aware of the fact, that the priming effects might also be driven by factors like positional-syntactic information (e.g. Hofmann, Biemann, \& Remus, 2017). Thus, future studies may also investigate the influence of syntactic information by using not only simple nouns from the word corpora, but also words from other syntactic classes or prime-target pairs spanning differential word classes (e.g., verbs, adjectives etc.). Further studies may also investigate whether computational models that reduce the amount of latent semantic dimensions can provide generalization capabilities that may account for more variance than the simple amount of common associates (e.g. Landauer \& Dumais, 1997).

Acknowledgments We like to thank all children and schools for their contribution to this study. Moreover, we thank Chris Biemann for giving advice concerning the word corpus and our student Azize Elia Walldorf for all her help during data acquisition. This work was partially supported by a grant by the Deutsche Forschungsgemeinschaft to MJH (HO5139/2-1).

## References

Arias-Trejo, N., \& Plunkett, K. (2013). What's in a link: associative and taxonomic priming effects in the infant lexicon. Cognition, 128(2), 214-217.
Anderson, J.R. (1983). A spreading activation theory of memory. Journal of Verbal Learning and Verbal Behavior, 22, 261-295.
Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics. Cambridge: Cambridge University Press.
Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. Memory \& Cognition, 8(6), 493512.

Bentin, S., McCarthy, G. \& Wood, C.C. (1985). Event-related potentials, lexical decision and semantic priming. Electroencephalography and Clinical Neurophysiology, 60, 343-355.
Bordag, S. (2007). Elements of Knowledge-free and Unsupervised Lexical Acquisition. Unpublished dissertation, Department of mathematics and informatics, University of Leipzig, Germany.
Brown, R. W. \& Berko, J. (1960). Word association and the acquisition of grammar. Child Development, 31, 1-14.
Chapman, L. J., Chapman, J. E, Curran, T. E., \& Miller, M. B. (1994). Do
children and the elderly show heightened semantic priming? How to answer the question. Developmental Review, 14, 159-185.
Collins, A.M., Loftus, E.F. (1975). A spreading-activation theory of semantic processing. Psychological Review, 82, 407-428.
Dunning, T.(1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19, 61-74.
Entwisle, D. R. (1966). The word association of young children. Baltimore: Johns Hopkins University Press.
Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955, in Studies in linguistic analysis. Oxford, England: Blackwell Publishers, pp. 1-32.
Harris, Z. S. (1951). Methods in Structural Linguistics. University of Chicago Press, Chicago, http://archive.org/details/structurallingui00harr
Hofmann, M. J., Kuchinke, L., Biemann, C., Tamm, S., Jacobs, A. M. (2011). Remembering words in context as predicted by an associative read-out model. Frontiers in Psychology, 2, 252,1-11.
Hofmann, M.J., Jacobs,A.M., (2014). Interactive activation and competition models and semantic context: from behavioral to brain data. Neuroscience \& Biobehavioral Reviews, 46,85-104.
Hofmann, M.J., Biemann, C., \& Remus, S. (2017). Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, EEGs and eye movements. In B. Sharp, F. Sedes, \& W. Lubaszewski (Eds.), Cognitive Approach to Natural Language Processing (pp.197216). Amsterdam, London: ISTE Press - Elsevier.

Hebb, D., (1949). The Organization of Behavior. Wiley: New York.
Jung, C.G. (1905). In: Ueber das Verhalten der Reaktionszeit beim Assoziationsexperiment. Ambrosius Barth: Leipzig.
Lucas, M. (2000). Semantic priming without association: A meta-analytic review. Psychonomic Bulletin \& Review, 7(4), 618-630.
Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104, 211-240.
Masson, M. E., \& Kliegl, R. (2013). Modulation of additive and interactive effects in lexical decision by trial history. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(3), 898-914.
McClelland, J.L., Rogers, T.T. (2003). The parallel distributed processing approach to semantic cognition. Nature, 4, 310-322.
McClelland, J. L., Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological Review, 88, 375-407.
McCauley, C., Weil, C. M., \& Sperber, R. D. (1976). The development of memory structure as reflected by semantic-primary effects. Journal of Experimental Child Psychology, 22, 511-518.
McNamara, T. (2005). Semantic Priming - perspectives from memory and word recognition. New York and Hove: Essays in Cognitive Psychology.
Meyer, D.E., \& Schvaneveldt, R.W. (1971). Facilitation in recognizing
pairs of words: evidence of a dependence between retrieval operations. Journal of Experimental Psychology. 90(2), 227-234.
Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. Memory \& Cognition, 4(5), 648-54.
Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: a review of research and theory. Psychological Bulletin, 84, 93-116.
Plaut, D. C., \& Booth, J. R. (2000). Individual and Developmental Differences in Semantic Priming: Empirical and Computational Support for a Single-Mechanism Account of Lexical Processing. Psychological Review, 107(4), 786-823.
Quasthoff, U., Richter, M. Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. Proceeding of the fifth Int. Conf. Language Resour. Eval., LREC, Genova, Italy (pp. 1799-1802).
Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. In: Association for Computational Linguistics, Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, pp. 1-7
Richter, T., Isberner, M., Naumann, J., \& Neeb, Y. (2013). Lexical Quality and Reading Comprehension in Primary School Children. Scientific Studies of Reading, 17(6), 415-434.
Roelke, A., Franke, N., Radach, R., Jacobs, A. \& Hofmann, M. (2016). Semantic higher order but not direct associations prime ventral visual stream activation. Manuscript submitted for publication.
de Saussure, F. (1959). Course in General Linguistics. Philosophical Library, New York http://books.google.de/books?id=FSpZAAAAMAAJ
Schmidt, U. (2015). Direkte und indirekte Assoziationen beim semantischen Priming. Unpublished bachelor thesis, Department of General Psychology, University of Wuppertal, Germany.
Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., \& Kliegl, R. (2015). childLex: Eine lexikalische Datenbank zur Schriftsprache für Kinder im Deutschen. Psychologische Rundschau, 66, 155-165.
Stuellein, N., Radach, R., Jacobs, A. \& Hofmann, M. (2016). No one way ticket from orthography to semantics in recognition memory: P200 and N400 effects of associations. Brain Research, 1639, 88-98.
Yarkoni, T., Balota, D., \& Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. Psychonomic Bulletin \& Review, 15 (5), 971-979.
White, H. (1985). The Syntagmatic-Paradigmatic Shift and Long-Term Memory Activation. The Journal of Genetic Psychology, 146, 561-562.
Woodrow, H. \& Lowell, F. (1916). Children's association frequency tables. Psychological Monographs, 22, No. 97.

# TRACX2: a RAAM-like autoencoder modeling graded chunking in infant visual-sequence learning <br> Robert M. French ${ }^{1}$ \& Denis Mareschal ${ }^{2}$ <br> ${ }^{1}$ LEAD-CNRS UMR 5022, UBFC, Dijon, FR \{robert.french@u-bourgogne.fr\} ${ }^{2}$ CBCD Birkbeck University of London, UK \{d.mareschal@bbk.ac.uk\} 


#### Abstract

Even newborn infants are able to extract structure from a stream of sensory inputs and yet, how this is achieved remains largely a mystery. We present a connectionist autoencoder model, TRACX2, that learns to extract sequence structure by gradually constructing chunks, storing these chunks in a distributed manner across its synaptic weights, and recognizing these chunks when they re-occur in the input stream. Chunks are graded rather than all-or-none in nature. As chunks are learned their component parts become more and more tightly bound together. TRACX2 successfully models the data from four experiments from the infant visual statistical-learning literature, including tasks involving lowsalience embedded chunk items, part-sequences, and illusory items. The model also captures performance differences across ages through the tuning of a single learning rate parameter. These results suggest that infant statistical learning is underpinned by the same domain general learning mechanism that operates in auditory statistical learning and, potentially, in adult artificial grammar learning. ${ }^{1}$


## Introduction

We live in a world in which events evolve over time. Consequently, our senses are bombarded with information that varies sequentially over time. One of the greatest challenges for cognition is to find structure within this stream of experiences. Even newborn infants are able to do this (Teinonen, et al. 2009; Bulf, Johnson \& Valenza, 2011), and yet, how this is achieved remains largely a mystery.
Two possibilities have been suggested (see Theissen, et al., 2013 for a detailed discussion). The first, characterised as statistical learning, involves using frequency and transition probabilities to construct an internal representation of the regularity boundaries among elements encountered. The second possibility suggests that elements that co-occur are recalled and simply grouped together - or chunked - into single units. Over time, these chunks can themselves be grouped into super-chunks or super-units. According to this view behaviour is determined by the recognition of these chunks stored in memory and associated with particular responses. What distinguishes these accounts

[^330]is that the former argues that it is the probabilistic structure of the input sequence that is represented and stored, whereas the later argues that specific cooccurring elements are stored, rather than the overarching statistical structure. Ample evidence in support of both of these views has been reported.
We will argue that this is a false dichotomy: both transitional probability learning (statistical learning) and chunking co-exist in one system that smoothly transitions between these apparent modes of behaviour. The appearance of two modes of learning is an illusion because only a single mechanism underlies sequential learning; namely, Hebbian-style learning in a partially recurrent distributed neural network. Such a system encodes exemplars (typical of chunking mechanisms) while drawing on co-occurrence statistics (typical of statistical learning models). An important corollary of this approach is that chunks are graded in nature rather than all-or-none. Moreover, interference effects between chunks will follow a similarity gradient typical of other distributed neural network memory systems.
Chunks are most frequently thought of as all-ornothing items. Who thinks of "cups" and "boards" when they see the word "cupboard"? Or "foot" and "ball" when they encounter the word "football"? Indeed, chunks like these have essentially the same status as "primitive" words like "boat" or "tree", which are not made of component sub-words. But new chunks do not suddenly appear ex nihilo in language. Rather, they are generally formed gradually, their component words becoming more and more bound together with time and usage. For example, when we encounter the words "smartphone", "carwash", or "petshop", we still clearly hear the component words. We hear them less in words like "sunburn" and "heartbeat". We hear them hardly at all in "automobile." How long did it take for people to stop hearing "auto" and "mobile" when they heard or read the word "automobile"? Like "automobile", it is likely that in a few years the current generation will no longer hear "smart" and "phone" when they hear the word "smartphone". This simple observation involving the graded nature of chunks is at the heart of the chunking mechanism in TRACX2.
These ideas were implicit in our initial presentation of the TRACX model (French et al., 2011). In TRACX we showed that a connectionist autoencoder, augmented with conditional recurrence, could extract chunks from a stream of sequentially presented symbols. TRACX
had two banks of input units, which it learned to autoencode onto two banks of identical output units. Sequential information was encoded by presenting successive elements of the sequence, first on the right input bank, then on the left input bank on the next time step. Thus, the sequence of inputs was presented in a successive series of right-to-left inputs, with learning occurring at each time step. However, if the output autoencoding error was below some pre-set threshold value (indicating successful recognition of the current pair of input elements), then, on the next time step, instead of the input to the right input bank being transferred to the left input bank, the hidden unit representation was put into the left input bank. The next item in the sequence was, as always, put into the right input bank. Weights were updated and the input sequence would then proceed as before. The result of this was that TRACX learned to form chunks of elements that it recognised as co-occurring (see French et al., 2011 for full details). TRACX successfully captured a broad range of data from the adult and infant auditory statistical learning literature and outperformed existing models of both chunking, notably, PARSER (Perruchet \& Vinter, 1998) and statistical learning (SRNs, Cleeremans \& McClelland, 1991).
TRACX2 (French \& Cottrell, 2014), which we use in this paper to segment and chunk sequential visual items, is an updated version of TRACX. TRACX2 removes the use of an all-or-nothing error threshold that determines whether or not the items on input are to be chunked. This effectively removes a conditional jump (i.e. an if-then-else) statement from the model, jump statements of this kind not being natural to neural network computation. In TRACX2, the contribution of the hidden-unit activation vector to the left bank of input units is graded and depends on the level of learning already achieved. TRACX (French et al., 2011) and TRACX2 (French \& Cottrell, 2014) were used to successfully model the segmentation of syllable (i.e., auditory) streams. In the present article, we use TRACX2 to model four experiments from the infant visual statistical learning literature. Visual statistical
learning paradigms involve showing infants sequences of looming colored shapes with varying degrees of statistical regularity embedded in the sequences. It was first developed as a visual analogue of the auditory statistical learning experiments (Kirkham, Slemner \& Johnson, 2002) and has yet to be captured by any modeling paradigm.

## The TRACX2 Architecture

TRACX2 was initially introduced by French and Cottrell (2014). The key to understanding TRACX2 is to understand the flow of information within the network. Over successive time steps, the sequence of information is presented item-by-item into the righthand bank (RHS) of input units. The left-hand bank (LHS) of input units is filled with a blend of the righthand input and the hidden unit activations at the previous time step, as shown in the following equation:

$$
\operatorname{LHS}_{\mathrm{t}+1}=\left(1-\tanh \left(\alpha \Delta_{\mathrm{t}}\right)\right) * \text { Hiddens }_{\mathrm{t}}+\left(\tanh \left(\alpha \Delta_{\mathrm{t}}\right)\right) * \mathrm{RHS}_{\mathrm{t}}
$$

where $\Delta_{t}$ is the absolute value of the maximum error across all output nodes at time $t, \mathrm{LHS}_{\mathrm{t}}$ is the activation across the left-hand bank of input nodes, Hiddens ${ }_{\mathrm{t}}$ are the hidden-unit activations at time $t, \mathrm{RHS}_{\mathrm{t}}$ is the activation across the right-hand bank of input nodes, and $\alpha$ is the sigmoid-"steepness" parameter, always set to 1 in the simulations presented here. If at time $t, \Delta_{t}$ is small, this means that the network has learned that the items on input are frequently together (otherwise $\Delta_{\mathrm{t}}$ could not be small). The contribution to the left-hand bank of input units at time $t+1$ of the hidden-unit activations, which constitute the network's internal representation of the two items on input at time $t$, is, therefore, relatively large and the contribution from the right-hand inputs will be relatively small. Conversely, if $\Delta_{t}$ is large, meaning that the items on input have not been seen together often, the hidden-layer's contribution at time $t+l$ to the left-hand input bank will be relatively small and that from the right-hand inputs will be relatively large. At each time step, the weights are updated to minimise output error (Fig. 1).
In layman's terms, this means that as you experience


Figure 1. Architecture and information flow in TRACX2. In all simulations reported in this paper, $\alpha=1$. When $\Delta$ is large (items not recognized as having been seen together before on input), almost all contribution to LHS comes from RHS. When $\Delta$ is small (items recognized as having been seen together before on input), almost all contribution to LHS comes from the Hidden layer (Hid).
items (visual, auditory, tactile) together over and over again, these items become bound to each other more and more strongly into a chunk until we no longer perceive its component parts.

## Modeling infant statistical learning

In this section we report on a total of four different simulations using TRACX2 of infant visual statistical learning behaviour. We used $\eta$ (the learning rate) as a proxy for development, with $\eta$ set to 0.0005 for newborns, 0.0015 for 2 -month-olds, 0.0025 for 5-month-olds, and 0.005 for 8 -month-olds. This is a typical parameter used to model age related differences in early learning (e.g., Thomas \& Johnson, 2006). There was a bias node on the input and hidden layers and momentum was always set to 0 . The key developmental hypothesis here is that, with increasing age, infants are progressively better at taking up information from an identical environment. This is consistent with the well-established finding that the average rate of habituation increases with increasing age during infancy (e.g., Bornstein et al., 1988; Colombo \& Mitchell, 2009; Westermann \& Mareschal, 2013). Finally, as has been used repeatedly elsewhere, we take network output error as a proxy for looking time in the infant (Mareschal \& French 2000; Mareschal, French, Quinn, 2000; Mareschal, Quinn, \& French, 2002; Mareschal \& Johnson, 2002; French, Mareschal, Mermillod \& Quinn, 2004; Westermann \& Mareschal, 2013). The idea here is that the amount of output error correlates with the number of cycles required to reduce the initial error, which corresponds to the amount of time or attention that the model will direct to a particular stimulus.
We begin by modeling the seminal Kirkham et al. (2002) visual statistical learning experiment demonstrating that age-related effects in the efficacy of learning can be accounted for by a simple and plausible parameter manipulation in TRACX2. We then show that TRACX2 can capture statistical learning in newborns, as well as their dependency on the complexity of the information stream (Bulf et al., 2011).

Finally, we show that, like 8 -month-olds (Slone \& Johnson, 2015), TRACX2 forms illusory conjunctions, normally taken as evidence of a statistical learning mechanism, but also shows decreased salience of embedded chunk items, normally taken as evidence of chunking. It, therefore, reconciles two apparently paradoxical behaviours within a single common mechanism.

## Visual statistical learning

Kirkham et al (2002) developed a visual analogue of the auditory statistical learning tasks initially developed by Saffran et al. (1996) and Aslin et al. (1998). Instead of listening to unbroken streams of sounds, infants were shown continuous streams of looming colorful shapes in which successive visual elements within a "visual word" were deterministic, but transitions between words were probabilistic (see Fig. 2, leftmost panel). Infants at three different ages were first familiarized to this stream of shapes, then presented with either a stream made up of the same shapes but with random transitions between all elements, or a stream made up of the identical visual words as during habituation. Kirkham et al. found that infants from 2 months of age subsequently looked longer at the random sequence than the structured sequence (even though elements are identical between streams) suggesting that the infants had learned the statistical structure of the training sequence.
We modelled this experiment by training the model with a sequence of inputs containing the identical probability structure to that used to train infants. The training sequence was identical in length to that used by Kirkham. The transitional probability within a visual word was $\mathrm{p}=1.0$, and between visual words $\mathrm{p}=.33$. Shapes were coded using localist, bipolar (i.e., $-1,1$ ) orthogonal encodings in order minimize effects due to input similarity. The RHS and LHS input vectors were comprised of 12 units.
Network performance was evaluated by averaging output error over all three of the possible two image "visual words" in the sequence. This was then


Figure 2. (leftmost panel) Illustration of visual sequences used to test infants (after Addyman \& Mareschal, 2013). (middle and rightmost panels) Left-hand panel: Infant performance reported in Kirkham et al. (2002) and, right-hand panel: TRACX2 performance with the familiar structured and novel non-structured sequences. (Error
is the maximum error of the network over all output units; SEM error bars.)
compared to the average output error for a set of three randomly selected two-image "visual non-words" that were neither words nor part-words, and, consequently, occurred nowhere in the training sequence. This is analogous to the word/non-word testing procedure used in auditory statistical learning studies (e.g., Saffran et al., 1996), and completely equivalent to testing the networks with a structured sequence (from which they would have extracted visual words) and a fully random sequence (in which no previous words or part-words exist). The model, like infants of all ages, looked longer at the randomised sequence than the structured sequence (Fig. 2, rightmost panel).

## Visual statistical learning in newborns

Bulf, Johnson, and Villenza (2011) asked whether the sequence-learning abilities demonstrated by Kirkham et al (2002) were present from birth. They tested newborns (within 1 week of birth) on black and white sequences of streaming shapes. In their "High Demand Condition", the sequence had the same statistical structure as in Kirkham et al. That is, the sequences were made up of 3 visual words, each made up of two shapes with a constant transition probability of 1.0 defining the word, and transitional probabilities of . 33 between words. They also introduced a "Low Demand Condition" in which the sequences were made up of only two words ( each consisting of two shapes with internal transition probabilities of 1.0) leading to transition probabilities at word boundaries of 0.5 (instead of the .33 previously used). The reasoning here was that newborns had more limited information processing abilities and may therefore struggle with a more complex sequence, already proving to be a challenge for 2 month olds.


Figure 3. Newborn performance as reported in Bulf \& Johnson (2011) in left panel and TRACX2 performance in right panel for familiar structured and novel nonstructured sequence.
Again, we modelled this study using TRACX2, in the same way as above, but by (1) reducing the learning rate to 0.0005 , and (2) creating both high-demand and low-demand sequences. In the low-demand condition
(LDC), there were two pairs of images, each made up of two different images (i.e., a total of 4 separate images). In the high-demand condition (HDC) there were three pairs of images, each made up of two different images (i.e., a total of 6 separate images). In the simulation for both the high-demand and lowdemand conditions, TRACX2 saw sequences of 120 words. Statistics were averaged over 30 runs of the program, with each run consisting of 10 simulated subjects. Figure 3 shows both the infant data and the model results. As with the infants, TRACX2 did not discriminate between the structured training sequence and the random sequence in the high demand condition (with the lower learning rate) but did discriminate between the two sequences in the low demand condition.

## Learning embedded and illusory items.

One consequence of chunking is that elements within a chunk become less salient as the chunks are increasingly consolidated. In contrast, statistical learning mechanisms predict that learners should form illusory items from elements that accidentally appear together on occasion. Slone \& Johnson (2015) explored whether infants' learning mechanisms would lead to the reduced salience of embedded items or to the emergence of illusory chunks, as a means of testing whether chunking or statistical learning underpins infant learning. To do this, they presented 8 -monthsolds with sequences structured as depicted in Figure 4a. Infants in the "Embedded Pair Experiment" did not differentiate embedded pairs from part-pairs that crossed word boundaries, but both were differentiated from the word pairs. Infants in the "Illusory Item Experiment" did not differentiate the illusory triplets from the part triplets, but both were differentiated from the actual triplets. This is perplexing because one result suggests that infants utilize chunking, whereas the other results suggests that they engage in statistical learning. TRACX2 captures both of these results well, with the caveat that the model is designed to produce the smallest error on the best learned patterns. (Figs. 4b, 4c). If we consider output error to be a measure of attention (the higher the error, the more attention the infant pays to that item), then we can say that TRACX2 is designed to orient to novel test patterns most (i.e., shows a novelty preference). In short, when modeling a novelty preference, the greater TRACX2's Error on output, the longer the looking time for infants.
Familiarity preferences are, in some sense, the inverse of novelty preferences. This means that the smaller the error for an item, the more attention the infant pays to that item. Thus, to model familiarity preferences we subtract the error on output from the maximum possible
error and call this "Inverse Error" (Fig. 4c). So, when modeling a familiarity preference, the greater TRACX's Inverse Error, the longer the infants' looking time.
Such shifts in orienting behaviour are common in infant visual orienting, and have been related to the complexity of the stimuli and the depth of processing (Roder, Bushnell, \& Sassville, 2000; Hunter \& Ames, 1988; see Sirois \& Mareschal, 2004, for a process account of the familiarity-to-novelty shift in a neural network model of habituation). Thus, TRACX2 captures both the reduced salience of embedded chunk items and the appearance of illusory conjunctions within a single mechanism, thereby reconciling apparently paradoxical infant behaviours.

## Discussion

TRACX2 (French \& Cottrell, 2014) is an updated version the TRACX architecture (French et al. 2011). As in the original architecture, TRACX2 is a memorybased chunk-extraction architecture. Because it is implemented as a recurrent connectionist autoencoder in the RAAM family of architectures (Pollack, 1989), it is also naturally sensitive to distributions statistics in its environment. In TRACX2, we replace the arbitrary all-or-none chunk-learning decision mechanism with a smooth blending parameter. TRACX2 learns chunks in a graded fashion as a function of its familiarity with the material presented. An implication of this is that chunks are no longer to be thought of as "all-or-none" entities. Rather, there is a continuum of chunks whose elements are bound together more or less strongly.
TRACX2 was used to model a representative range of infant visual statistical learning phenomena. No

previous models of these behaviours exist. As with the auditory learning behaviours, TRACX2 captures infants' apparent use of forward and backward transitional probabilities, the diminishing sensitivity to embedded items in the sequence, and the emergence of illusory words. However, it is important to understand that TRACX2 is not simply internalising the overall statistical structure of the sequence, but encoding, remembering and recognizing previously seen chunks of information. This is a fundamentally different account of infant behaviours than has previously been proposed (Krogh, Vlach \& Johnson, 2013).
TRACX2 can use frequency of occurrence or transitional probabilities equally well and fluidly to learn a task (as is the case with 8-month-olds; Marcovitch \& Lewkowicz, 2009). This would suggest that categorizing learning either as statistical or memory-based is a false dichotomy. Both can happen in a single system, with different behaviours seeming to appear depending on the constraints of the task, the level of learning and the level of prior experience. Moreover, the idea that infant looking time is determined by the recognition of regularly re-occurring items (chunks or individual items) is consistent with the recent evidence suggesting that local redundancy in the sequences is the prime predictor of looking away in infant visual statistical learning experiments (Addyman \& Mareschal, 2013).

TRACX2 also suggests that there are no specialised mechanisms in the brain dedicated to sequence learning. Instead, sequences emerge from the application of fairly ubiquitous associative mechanisms, coupled with graded top-down re-entrant processing.


(b)

(c)


Test type

Figure 4. (a) Familiarisation and testing items for embedded pairs (left panel) and illusory items (right panel) (after Slone \& Johnson, submitted). (b) Infant data (left-hand side of figure, familiarity preference, Experiment 1) and TRACX2 performance (right-hand side, SEM error bars). (c) Infant data (left-hand side of figure, novelty preference, Experiment 2) and TRACX2 performance (right-hand panel, SEM error bars). (Figure (a) permission pending).

Although there may be differences in the speed and richness of encoding across modalities, there is nothing intrinsically different in the way TRACX2 processes visual or auditory information. This suggests than any modality-specific empirical differences observed can be attributed to encoding differences rather than core sequence-processing differences.
In conclusion, we believe that chunking cannot be viewed as an all-or-nothing phenomenon. Chunks are learned and over the course of being learned their component parts become more and more tightly bound together. This is a fundamental principle of TRACX2. The results of the present paper suggest that infant statistical learning is underpinned by the same domain general learning mechanism that operates in auditory statistical learning and, potentially, also in adult artificial grammar learning. TRACX2, therefore, offers a parsimonious account of how infants find structure in time.

## Acknowledgments

This work was funded by a grant from the Agence Nationale de la Recherche Scientifique (ANR), ANR-14-CE28-0017, to the first author and an Economic and Social Research Council Grant RES-360-25-056 to the second author. DM is further funded by a Royal Society Wolfson Research Merit Award.

## References

Addyman, C. \& Mareschal, D. (2013) Local redundancy governs infants' spontaneous orienting to visual-temporal sequences . Child Development, 84, 1137-1144.
Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8 -month-old infants. Psychological Science, 9, 321-324.
Bornstein, M. H., Pecheux, M.G, Lecuyer, R. (1988) Visual habituation in human infants: development and rearing circumstances. Psychological Research, 50, 130-133.
Bulf, H., Johnson, S. P., \& Valenza, E. (2011) Visual statistical learning in the newborn infant. Cognition, 121, 127-132.
Cleeremans, A. and McClelland, J. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. Machine Learning, 7, 161-193.
Colombo, J. \& Mitchell, D. W. (2009) Infant visual habituation. Neurobiology of Learning \& Memory, 92, 225-234.
French, R. M., Addyman, C. \& Mareschal, D. (2011) TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. Psychological Review, 118, 614-636.
French, R. M. and Cottrell, G. (2014). TRACX 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction. In P. Bello, M. Guarini, M. McShane and B. Scassellati (Eds.), Proc of the 36th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society, 2016-2221.
French, R. M., Mareschal, D., Mermillod, M. \& Quinn, P. (2004) The role of bottom-up processing in perceptual
categorization by 3- to 4-month old infants: Simulations and data. JEP:G, 133,_382-397.
Hunter, M. A. \& Ames, E. W. (1988) A multifactor model of infant preferences for novel and familiar stimuli. Advances in Infancy Research,5, 69-95.
Kirkham, N., Slemmer, J.A., \& Johnson, S. P. (2002). Visual statistical learning in infancy: evidence of a domain general learning mechanism. Cognition, 83, B35-B42.
Krogh, L., Vlach, H. A \& Johnson, S. P (2013) Statistical learning across development: flexible yet constrained. Frontiers in Psychology, 3, art. 598.
Mareschal, D., Quinn, P. C. \& French, R. M. (2002) Asymmetric interference in 3- to 4-month-olds' sequential category learning. Cognitive Science, 26, 377-389.
Mareschal, D. \& Johnson, S. P. (2002) Learning to perceive object unity: A connectionist account. Developmental Science, 5, 151-172.
Mareschal, D., French, R. M. \& Quinn, P. (2000) A connectionist account of asymmetric category learning in infancy. Developmental Psychology, 36, 635-645.
Mareschal, D. \& French, R. M. (2000) Mechanisms of categorisation in infancy. Infancy, 1, 59-76.
Marcovitch, S. \& Lewkowicz (2009) Sequence learning in infancy: the independent contributions of conditional probability and pair frequency information. Developmental Science, 12, 1020-1025.
Perruchet, P., \& Vinter, A. (1998). PARSER: A model for word segmentation. J. of Mem. and Lang., 39, 246-263.
Pollack, J. (1989) Implications of Recursive Distributed Representations. In David S. Touretzky (ed.) Advances in Neural Information Processing Systems I (pp. 527-536). Morgan Kaufmann, Los Gatos, CA.
Roder, B.J., Bushnell, E.W., \& Sassville, A.M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing Infancy, 1, 491-507.
Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996) Statistical learning by 8 -month-old infants. Science, 274, 1926-1928.
Sirois, S. \& Mareschal, D. (2004). An interacting systems model of infant habituation. J. Cog. Neuro, 16, 1352-62.
Slone, L. K. \& Johnson, S. P. (2015) Statistical and chunking processes in infants' and adults' visual statistical learning. Poster presented and the Biannual Conf. of the SRCD, April 2015, Philadelphia, USA.
Teinonen, T., Fellman, V., Näätänen, R., Alku, P., and Huotilainen, M. (2009). Statistical language learning in neonates revealed by eventrelated brain potentials. BMC Neurosci. 10:21. doi:10.1186/1471-2202-10-21
Thiessen, E. D., Kronstein, A. T., and Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. Psychological Bulletin, 139, 792. DOI: 10.1037/a0030801

Thomas, M. S. C. \& Johnson, M. H. (2006). The computational modelling of sensitive periods. Developmental Psychobiology, 48, 337-344.
Westermann, G. \& Mareschal, D. (2013) From perceptual to language-mediated categorization. Philosophical Transactions of the Royal Society B, 369: 201220391

# Risk, Cognitive Control, and Adolescence: Challenging the Dual Systems Model 

Joanna Fryt (jfryt@up.krakow.pl) Tomasz Smolen (tsmolen@up.krakow.pl) Karolina Czernecka (czernecka.karolina@gmail.com)<br>Department of Psychology, Pedagogical University of Krakow ul. Podchorazych 2, Krakow, Poland

Amelia La Torre (amelia.latorre@gmail.com)<br>Monika Szczygiel (szczygiel89@gmail.com)<br>Institute of Psychology, Jagiellonian University, Krakow<br>ul. Ingardena 6, Krakow, Poland


#### Abstract

According to the dual systems model, adolescence is a period of imbalance between cognitive and motivational systems that results in increased tendency towards risk. In the study, we investigated the effects of rewards on risk-taking and cognitive control in 90 adolescents (13-16) and 96 adults (1835 ). Our results challenge the assumptions of the model as we observed that rewards lead adolescents to more conservative decisions in one of the risk tasks used in the study. We also observed that in cognitive control tasks, rewards influenced reaction latencies, but not the efficiency of control processes.


Keywords: risk taking strategy, cognitive control, sensitivity to rewards, dual systems model

## Introduction

As a developmental period, adolescence is commonly characterized by risk-taking, sensation seeking, impulsivity and the importance of peers. Such characteristics clearly serve an adaptive function during the transition to adulthood, fostering tendencies towards independence, novel experiences and social networks (Spear, 2000). At the same time, they expose adolescents to the negative consequences of their actions, with typical examples being reckless driving, experimenting with psychoactive substances or unprotected sex. In our study, we investigated adolescent sensitivity to rewards and its consequences for risk-taking and cognitive control.

## The Dual Systems Model

The dual systems model by Steinberg (2008) is one of the most influential propositions attempting to explain adolescent behavior that is well established in cognitive and neurodevelopmental research (Defoe, Dubas, Figner, \& van Aken, 2015; Geier, 2013). According to the model, adolescence can be characterized by a functional imbalance between the hyperactive motivational system, responsible for increased sensitivity to rewards, and the still maturing cognitive control system, responsible for reaction inhibition and effective management of information. In both human and animal adolescents, greater sensitivity towards pleasure, positive feedback and rewarding effects of social interactions are observed (Somerville \& Casey, 2010; Spear,
2011). In humans, it has been established that early (11-13 years old) and middle adolescence (14-16 years old) are the periods of highest sensitivity to rewards. The presence of salient incentives coupled with the immaturity of control processes is believed to result in increased tendency towards risk.

Risk-taking, defined as a propensity towards actions "with the highest outcome variability" (Defoe et. al., 2015), is the most studied consequence of adolescent sensitivity to rewards. More precisely, risk-taking is a preference for actions leading to a big gain of low probability over actions leading to a small gain of high probability. According to the dual systems model, adolescents take more risks in the presence of salient incentives and when they are emotionally aroused. Studies focusing on age differences in risk-taking show that adolescents do manifest stronger tendency towards risk than adults, but only under specific task demands or in specific social contexts. A meta-analysis by Defoe et al. (2015) revealed that in studies using probabilistic gambling tasks (e.g. Iowa Gambling Task, Columbia Card Task, Balloon Analogue Risk Task), these specific task demands include primarily immediate outcome feedback, i.e. participants are informed of their gains and losses immediately after each decision. In studies using fastpaced driving tasks (e.g. Stoplight Task, driving simulators), it is usually the presence of a peer observer that encourages adolescents to take risks (e.g. Chein, Albert, O'Brien, Uckert, \& Steinberg, 2011; Cascio et al., 2015). It seems unclear whether these two types of risk task measure one or more types of risky behavior. The use of probabilistic gambling tasks allow a better understanding of economic risk preference. In fast-paced driving tasks, the risk is more impulsive and more similar to everyday situations.

Beyond the risk context, the assumptions of the dual systems model are tested in a rewarded vs. neutral antisaccade paradigm. Interestingly, some results show that cognitive control is enhanced in adolescents, but not in adults, where it is financially rewarded (Geier, Terwilliger, Teslovich, Velanova, \& Luna, 2010; Padmanabhan, Geier, Ordaz, Teslovich, \& Luna, 2011). Such an effect does not correspond to the implication of the model that adolescent risk-taking stems from weaknesses of control processes. Rather, it indicates that adolescent sensitivity to rewards can
in fact be adaptive and promote cognitive efficiency. Unfortunately, the effect seems to be difficult to replicate. Subsequent studies show individual differences such as increase, decrease or no change in cognitive control in adolescent response to rewards (Geier \& Luna, 2012; Paulsen, Hallquist, Geier, \& Luna, 2015). One study using the Continuous Performance Test showed a similar increase of performance in children, adolescent and adults in the rewarded condition (Strang \& Pollack, 2014).

## The Challenges for the Dual Systems Model

Despite the fact that the dual systems model does not specify whether tendency towards risk is adaptive or maladaptive (Strang, Chein, \& Steinberg, 2013; Shulman et al., 2016), it is criticized mainly for its generality and the fact that it adopts a deficit perspective on adolescence (Pfeifer \& Allen, 2012; Telzer, 2016). Actually, results from many studies contribute to the image of adolescence as the period of greatest lability, vulnerability to social evaluation, and decision making which may be suboptimal or even lifethreatening. However, a high propensity towards risk may not be the domain of all but the most susceptible adolescents (Bjork \& Pardini, 2015). An interesting new development for the model might be offered by research demonstrating that adolescent sensitivity to rewards can lead not only to risk-taking, but can also be channeled towards safe (Cascio et al, 2015; Telzer, Ichien, \& Qu, 2015) or prosocial behavior (Telzer, Fuligni, Lieberman, \& Galvan, 2013; 2014).

The conceptualization of adolescent risk-taking within the model and beyond it remains, however, the most intriguing issue. Are adolescents impulsive risk-takers, who, due to the immaturities of their control processes, cannot override risky tendencies in the presence of salient incentives (Willoughby, Good, Adachi, Hamza, \& Tavernier, 2013)? Or is risk-taking rather a decision strategy adopted whenever it seems profitable? When we view risk-taking as a strategy, we can also see adolescents as having more control over their behavior than is assumed in the model. Decision strategies can vary depending on the task and the type of risky behavior (e.g. economic risk, driving risk). Adolescent sensitivity to rewards (e.g. financial rewards, immediate outcome feedback) can be similar or different in various risk tasks. Nevertheless, it seems to be associated with emotional arousal. Finally, an issue worth examining is whether cognitive control in adolescents is indeed weaker than in adults and more sensitive to rewards.

## Hypotheses

To sum up, we expected that adolescents would be more sensitive to rewards than adults and that the difference in sensitivity would manifest in more efficient cognitive control and a higher tendency towards risk when performance is rewarded. Also, as the dual systems model does not provide a direct link between the presence of reward and its possible effect on risk-taking and cognitive
control, we hypothesized that the effect may be mediated by the most obvious variable: emotional arousal.

Taking into consideration all of the above, we can formulate the following predictions.
Risk (1) When rewarded according to their performance, people will manifest more risk-taking compared to a noreward condition.
Cognitive control (2) When rewarded, people will exhibit more efficient cognitive control.
Developmental changes (3) The simple effects expected in hypotheses (1) and (2) will be larger for adolescents than for adults.
Arousal (4) People will report higher arousal when rewarded according to their performance, compared to a noreward condition. (5) The arousal level will be a substantial mediator between the type of condition, risk level and cognitive control efficiency.

## Procedure

The one hundred and eighty six subjects ( 81 men) were recruited either via parent-teacher conferences in local schools (adolescents) or online advertisements (adults) from two groups: adolescents $(N=90$, mean age $=13.82, S D=$ 0.89 , range $=[13,16])$, and adults $(N=96$, mean age $=$ 25.04, $S D=4.03$, range $=[18,35])$. Parental consent was obtained for all under age participants. The study was conducted in schools (adolescents) and the university psychological laboratory (adults). Participants were informed that the anonymized data would be used only for the scientific purposes of the study and that they could ask questions, withdraw their participation at any moment, and receive performance feedback after the study was completed.

The session lasted for about 90 minutes and consisted of two conditions, with a fifteen-minute break in between: (a) a set of tasks with rewards depending on the performance and (b) a set of tasks without any rewards. In each condition, participants performed four tasks, each preceded by a training session. Two tasks were cognitive control tasks (Stroop task, Antisaccade task), while the other two measured the tendency to take risks (Spaceride task and Stock Market task). The order of conditions and of the tasks within sets were randomized. In the middle of each condition (after performing two computer tasks), participants were asked to complete the SUPIN arousal scale. Therefore, each person performed four computer tasks and the SUPIN scale twice, once in a rewarded and once in an unrewarded condition.

Participants were paid for their attendance with vouchers (to a clothing store, a bookstore, or a movie theatre). The value of the vouchers depended on performance in each task in the rewarded condition and varied from $\$ 5$ to $\$ 15$ (mean \$10, equivalents in PLN).

## The Tasks

The tasks were selected so as to measure different aspects of risk-taking and cognitive control. In contrast to the Stock

Market task, which investigates the tendency to make risky decisions based mostly on deliberative thinking, the Spaceride task was designed to detect the tendency to take risk in emotionally stimulating conditions. As two aspects of cognitive control, interference inhibition and response inhibition were measured separately by the Stroop task and the Antisaccade task. The SUPIN Scale was introduced to control the level of positive and negative affect during each research condition, as a possible moderator of task results.
Stock Market The task resembles a financial game in which participants use virtual currency to buy shares in two fictitious companies. In each turn of the game it was possible to buy a number of shares of one or two companies or no shares at all. The only restriction was the amount of money the participant had at a given moment, which was shown on the right side of the screen. The participant had 60 seconds to take a single decision, and 20 decisions to make during the game, which was also displayed on the screen. During the game the participant could see the history of changes of the prices of each companies' shares displayed on a chart. After each decision they also saw a table showing the current values of stocks and how much money they had earned or lost so far. The price changes were probabilistic (independent and normally distributed). The expected gain (mean price change) from investing in any of the companies was the same; however, the variance of the price changes was small for one company (safe) and large for the other (risky). The difference between the companies was revealed to participants at the beginning of the task.
Spaceride The task fulfilled a function similar to the "Stoplight" task (Chein et al., 2011), in which participants in a car-driving context had to quickly decide whether or not to take a risk to reach their destination as quickly as possible. The Spaceride task has the form of a game in which the participant controls a spaceship seen from above. The task was to fly as quickly as possible to the end of the cosmic route. There were a number of danger zones where there was a risk of collision with asteroid. Those zones were marked by a sound signal, a light on a radar, and the appearance of distant asteroids in the background. A cloud of fog also sometimes appeared and covered the spaceship and its surroundings, making it impossible to see asteroids approaching. In each danger zone, the participant had to decide whether to slow down and avoid a collision or speed ahead, risking a collision with an asteroid. A collision would immobilize the spaceship for longer than it would take to fly through a danger zone.
Stroop task The task (Stroop, 1935) was used to evaluate participants' ability to inhibit interference. In each trial of the task, one of four words ("red", "brown", "blue" or "green") appeared on the screen displayed in one of the four colors (also red, brown, blue, or green). In congruent trials ( $50 \%$ of trials), the color was the same as the meaning of the word (e.g. the word "brown" written in brown), while in incongruent trails the word was written in one of the other three colors (e.g. the word "red" written in blue, meaning interference was present). The participants had to press one
of four keys corresponding to the displayed color of the word as quickly as possible and ignore the meaning of the word. To motivate the participants for a better response, a status bar visible on the top of the screen was additionally introduced. After every response the bar changed color to green when the response was correct, or to red when it was wrong. The faster the response, the shorter the bar, so the participant could see the accuracy and speed of every reaction during the game.
Antisaccade task The task (Unsworth, Schrock, \& Engle, 2004) served as a measure of response inhibition. The participant had to inhibit the tendency to look at a sudden presentation of a peripheral lure stimulus and instead look at its mirror location in order to perceive the target stimulus (arrow) and correctly react to it (press one of three keys depending on the direction of the arrow). Feedback was additionally introduced in the present task to inform participants of their accuracy. The feedback took the form of a screen-wide rectangle displayed in green (in the case of a correct response) or red (when the reaction was wrong). Modified SUPIN Scale The scale (Brzozowski, 2010) was derived from Watson \& Clark's Positive and Negative Affect Schedule. Both positive and negative affect were measured, forming two subscales of the questionnaire. The scale consisted of 20 adjectives describing various emotions. Participants indicated on a five-point Likert scale (from 1 - "very slightly or not at all" to 5 - "extremely") how well each adjective described their current state. On the basis of the results of our preliminary study, we altered seven items of the scale to achieve better psychometric characteristics. The modified version of the scale was used in the present study.

## Results

Statistics and data analysis One person did not finish the whole set of tasks, while 60 had their results removed for one task due to low accuracy (in the Antisaccade task) or outlying value; however, their remaining results were still used in the analysis.

A generalized linear mixed model using binomial distribution was fit to the Antisaccade task. The "Mediation" package in R was used for mediation analysis. Multi-factor analysis of variance with repeated measures was applied in all other analyses.

Condition (rewarded or not rewarded), group (adolescents or adults), and interaction between condition and age group were independent variables. The condition factor was applied within subjects while the age group was applied between subjects. We also controlled for position in series (first or second) and performance in analysis concerning risk tasks. Dependent variables (DV) were: number of correct responses in the Antisaccade task; Stroop effect in the Stroop task; proportion of high risk stocks in all stocks purchased (risk measure), and number of stocks purchased (alternative DV) in the Stock Market task; duration of pressing the break button (risk measure) and duration of pressing the break or accelerate button (alternative DV) in
the Spaceride task. Alternative DV in the Spaceride task was logarithmized due to its skewness ( $\gamma_{1}=2.17$ before and 0.29 after transformation). The performance measure in the Stock Market task was the amount of "money" in the last trial, and in the Spaceride task it was the negative time of the journey. DV in the Antisaccade task was accuracy, and in the Stroop task it was Stroop effect. We also examined reaction latencies in Antisaccade and Stroop tasks.
Cognitive control There was neither an effect of condition nor an interaction between condition and age group in the Antisaccade task $(\beta=0.027, p=.47$; and $\beta=-0.03, p=.56$ respectively, deviance $=1695.8$ ) and the Stroop task $(F[1,136]=0.2, \mathrm{p}=.66$; and $F[1,136]=1.75, p=.19$, respectively, $\eta^{2}=.037$ ). However, there was a significant difference between adolescents and adults in reaction latencies in the Antisaccade task ( 734 ms for adolescents and 695 ms for adults, $F[1,164]=10,84, p=.001)$ and a nearly significant difference between conditions ( 707 ms for not rewarded and 719 ms for rewarded, $F[1,152]=3.58, p=$ $.06, \eta^{2}=.19$ ). There also was a significant difference between the rewarded and unrewarded condition in reaction latencies in the Stroop task ( 936 ms for unrewarded and 906 ms for rewarded, $F[1,136]=101.97, p<.001, \eta^{2}=.12$ ).
Risk The performance in the Stock Market task did not depend on condition $(F[1,180]=0.1, p=.75)$, age group $(F[1,182]=0.59, p=.44)$, nor interaction between these two factors $\left(F[1,180]=1.65, p=.2, \eta^{2}=.032\right)$. Neither did the performance in the Spaceride task depend on any of these predictors $(F[1,182]=0.13, p=.72 ; F[1,183]=2.52$, $p=.11 ; F[1,182]=1.24, p=.27$, respectively, $\left.\eta^{2}=.051\right)$.

There also was neither effect of condition $(F[1,182]=$ $0.81, p=.37)$, age group $(F[1,183]=0.87, p=.35)$, nor interaction between condition and age group on risk $\left(F[1,182]=0.008, p=.93, \eta^{2}=.0078\right)$ in the Stock market task or $(F[1,182]=0.028, p=.88 ; \mathrm{F}[1,183]=1.57, p=.21$; and $F[1,182]=0.26, p=.61$, respectively, $\eta^{2}=.01$ ) in the Spaceride task.

However, there was a significant effect of condition (242 for the unrewarded condition and 223 for the rewarded condition, $F[1,181]=8.96, p=.0031$ ) and age group (200 for adolescents and 262 for adults, $F[1,182]=18.65, p<$ .001 ) and a nearly significant effect of interaction between condition and age group $\left(F[1,181]=3.43, p=.065, \eta^{2}=\right.$ .35) when alternative DV was used in the Stock Market task (see Figure 1), as well as an effect of interaction between condition and age group in the Spaceride task $(F[1,182]=$ $4.75, p=.031)$. The effect of condition or age group in the latter task was not significant $(F[1,182]=1.08, p=.3$; $F[1,183]=0.44, p=.14$, respectively, $\eta^{2}=.019$, see Figure 2).

Arousal The arousal differed significantly depending on condition ( 2.34 for not rewarded condition and 2.5 for rewarded condition, $F[1,180]=43, p<.001, \eta^{2}=.034$ ), but it was not a mediator between condition and alternative DV in the Stock Market task (proportion mediated $=-.06,95 \%$ $\mathrm{CI}=[-.44, .25], p=.63)$.


Figure 1: Quantity of stock purchased in the Stock Market task (alternative DV) in unrewarded and rewarded condition for adolescents and adults. Error bars indicate $95 \%$ confidence intervals.


Figure 2: Logarithm of total time for which the accelerate or break buttons were pressed in the Spaceride task (alternative DV) in unrewarded and rewarded conditions for adolescents and adults. Error bars indicate 95\% confidence intervals.

## Discussion

The first important observation made in the present study is that participants were sensitive to rewards in risk tasks, but this sensitivity leads adolescents and young adults to different decision strategies, depending on the context of the task. Adolescent decisions, however, cannot be interpreted as an increase in tendency towards risk, which challenges the assumption of the dual systems model (Geier, 2013). In the Stock Market task we observed that adolescents generally purchased less stocks than adults and the number of purchased stocks decreased even more in the rewarded condition (Fig. 1). In the Spaceride task there were no differences between adolescents and adults in time taken to press the break or accelerate button in danger zones in the no-reward condition. However, in the rewarded condition adolescents pressed the brake and accelerator buttons more than adults, making the difference between the groups significant (Fig. 2). It is interesting why the presence of rewards led adolescents to purchase less stocks in the market and steer the spaceship more boldly through danger zones. Possibly, when they had the opportunity to earn real money, participants chose a strategy that leads, as they believe, to better performance in the task. If purchasing stocks in the market is generally perceived as leading to both big gains and big losses-the option with "the highest outcome variability" as Defoe et al. (2015) define riskthen purchasing less stocks when real money is earned can be interpreted as a strategy that protects participants from loss. Otherwise, flying more boldly through danger zones cannot be seen as a strategy preventing collisions. It should be noted that while in the Stoplight task (Chein et al., 2011) participants decide whether to stop at a yellow light or drive through the crossroads, in the Spaceride task it is possible to brake and accelerate through the entire length of danger zones. Flying more boldly (such as the "speed-brake-speed" strategy) in dangerous areas is related to maintaining high speed and attempting to slow down just before asteroids. Less bold flying is slower, but makes attempts to avoid collision more effective. Summing up these results, it appears that adolescents made more conservative decisions than adults in one of the tasks and more risky decisions in the other. The context of tasks is therefore a variable that determines whether adolescents manifest risk-taking or riskaversion. We can speculate that more conservative decisions could be caused by a lack of familiarity with the contexts in which risk can occur (e.g. economic risk).

According to our hypothesis, participants reported higher emotional arousal in the rewarded condition. Such results suggest that the presence of a salient incentive leads to a greater motivational effort that manifests itself in higher reported arousal. We failed, however, to show that arousal mediates the relation between the presence of reward and risk-taking (or other decision strategy). As adolescents are viewed as impulsive risk-takers (Willoughby et al., 2013), the dual systems model predicts that high arousal in the rewarded condition enhances risk-taking because highly aroused adolescents cannot override risky tendencies. In our
study, however, participants seemed to be able to make decisions irrespective of their arousal and did not allow it to negatively influence their performance. It might be the case that arousal leads to impulsive decision-making only in specific circumstances. For example, high arousal may trigger risk-taking only in individuals in a negative emotional state (such as anxiety) or under high cognitive load (see, e.g., Zangeneh, Blaszczynski \& Turner, 2008). If the participants were in optimal emotional and cognitive state, reward-related arousal alone might not have been sufficient to cause a break-down in control processes and an increase in risk-taking. It is also possible that rewarding participants resulted in a higher but still optimal level of arousal, increasing not risk-taking, but effort. These explanations remain speculative and need further studies, but it seems that the dual systems model may oversimplify the proposed link between arousal and risk-taking.

Interesting results that challenge the dual systems model assumptions were also observed for the cognitive control measures. Firstly, adolescents were less accurate and slower in the Antisaccade task, while no differences between adolescents and adults were observed in the Stroop task. Thus, the antisaccade task seems to be more difficult for adolescents, a result which is consistent with previous studies (Geier \& Luna, 2012; Paulsen et al., 2015) showing that performance in the Antisaccade task improves with age. Secondly, we found that reward had no effect on both the accuracy in the Antisaccade task and the Stroop effect. However, in the rewarded condition participants exhibited longer latencies in the Antisaccade task and shorter latencies in the Stroop task. These results are not surprising given the fact that participants were informed that they were being rewarded for accuracy in the first task and for response speed in the second. The intriguing issue here is why the presence of rewards influenced not the measures of cognitive control efficiency (reaction inhibition and interference control), but reaction latencies in the tasks. It is possible that rewards enhance not a measured skill (i.e. control processes) that might be difficult to improve, but the motivational effort to do well in the task. Such an interpretation seems to be consistent with the effects of reward observed in the risk tasks, where again not the performance (e.g. money earned in the Stock Market, driving time in the Spaceride), but the decision strategies (e.g. purchasing more or less stocks, driving more or less dynamically) were enhanced. Additionally, we failed to observe interaction between age, condition and cognitive control efficiency, which is contrary to the dual systems model and consistent with the behavioral results of Paulsen et al. (2015). The effects of reward on reaction latencies in both tasks were similar in adolescents and adults.

To conclude, the results obtained in the study challenge the assumptions of the dual systems model about the universality of adolescent risk-taking. Risk-taking as a consequence of the weakness of control processes and sensitivity to incentives possibly manifests itself in certain circumstances. In our study, adolescents made decisions
which cannot be considered unequivocally risky or impulsive, despite the rewards. Further studies should help determine more precisely what set of circumstances triggers different behavioral responses in the presence of incentives and thus contribute to the development of the model.

## Acknowledgments

This work was sponsored by The National Science Centre (NCN) of Poland (grant 2015/18/E/HS6/00152).

## References

Bjork, J. M., Pardini, D. A. (2015). Who are those "risktaking adolescents"? Individual differences in neurodevelopmental research. Developmental Cognitive Neuroscience, 11, 56-64.
Brzozowski, P. (2010). Skala Uczuć Pozytywnych i Negatywnych (SUPIN). Polska adaptacja skali PANAS Dawida Watsona i Lee Anny Clark. Podręcznik. Warszawa: Pracownia Testów Psychologicznych.
Cascio, C. N., Carp, J., O’Donnel, M. B., Tinney, F. J., Bingham, C. R., Shope, J. T., Ouimet, M. C., Pradhan, A. K., Simons-Morton, B. G., Falk, E. B. (2015). Buffering social influence: neural correlates of response inhibition predict driving safety in the presence of a peer. Journal of Cognitive Neuroscience, 27(1), 83-95.
Chein, J., Albert, D., O’Brien, L., Uckert, K., Steinberg, L. (2011). Peers increase adolescent risk-taking by enhancing activity in brain's reward circuitry. Developmental Science, 14(2), F1-F10.
Defoe, I. N., Dubas, J. S., Figner, B., van Aken, M. A. (2015). A meta-analysis on age-differences in risky decision-making: Adolescents versus children and adults. Psychological Bulletin, 141(1), 48-84.
Geier, C. F. (2013). Adolescent cognitive control and reward processing: implications for risk taking and substance use. Hormones and Behavior, 46, 333-342.
Geier, C. F., Luna, B. (2012). Developmental effects of incentives on cognitive control. Child Development, 83(4), 1262-1274.
Geier, C.. F., Terwilliger, R., Teslovich, T., Velanova, K., Luna B. (2010). Immaturities in reward processing and its influence on inhibitory control in adolescence. Cerebral Cortex, 20, 1613-1629.
Padmanabhan, A., Geier, C. F., Ordaz, S. J., Teslovich, T., Luna, B. (2011). Developmental changes in brain function underlying the influence of reward processing on inhibitory control. Developmental Cognitive Neuroscience, 1(4), 517-529.
Paulsen, D. J., Hallquist, M. N., Geier, C. F., Luna, B. (2015). Effects of incentives, age, and behavior on brain activation during inhibitory control: A longitudinal fMRI study. Developmental Cognitive Neuroscience, 11, 105115.

Pfeifer, J. H., Allen, N. B. (2012). Arrested development? Reconsidering dual-systems models of brain function in adolescence and disorders. Trends in Cognitive Science, 16, 322-329.

Shulman, E. P., Smith, A. R., Silva, K., Icenogle, G., Duell, N., Chein, J., Steinberg, L. (2016). The dual systems model: Review, reappraisal and reaffirmation. Developmental Cognitive Neuroscience, 17, 103-117.
Somerville, L. H., Casey, B. J. (2010). Developmental neurobiology of cognitive control and motivational systems. Current Opinion in Neurobiology, 20, 2, 236241.
Spear, L. P. (2000). The adolescent brain and age-related behavioural manifestations. Neuroscience and Biobehavioral Reviews, 24(4), 417-463.
Spear, L. P. (2011). Rewards, aversions and affect in adolescence: emerging convergences across laboratory animal and human data. Developmental Cognitive Neuroscience, 1(4), 390-403.
Steinberg, L. (2008). A social neuroscience perspective on adolescent risk taking. Developmental Review, 28, 78106.
Strang, N. M., Chein, J. M., Steinberg, L. (2013). The value of the dual systems model on adolescent risk-taking. Frontiers in Human Neuroscience, 7:223, 1-4.
Strang, N. M., Pollak, S. D. (2014). Developmental continuity in reward-related enhancement of cognitive control. Developmental Cognitive Neuroscience, 10, 3443.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643662.

Telzer, E. H. (2016). Dopaminergic reward sensitivity can promote adolescent health: A new perspective on the mechanism of ventral striatum activation. Developmental Cognitive Neuroscience, 17, 57-67.
Telzer, E. H., Fuligni, A. J., Lieberman, M. D., Galvan, A. (2013). Ventral striatum activation to prosocial rewards predicts longitudinal declines in adolescent risk taking. Developmental Cognitive Neuroscience, 3, 45-52.
Telzer, E. H., Fuligni, A. J., Lieberman, M. D., Galvan, A. (2014). Neural sensitivity to eudaimonic and hedonic rewards differentially predict adolescent depressive symptoms over time. Proceedings of the National Academy of Sciences of the United States of America, 111, 6600-6605.
Telzer, E. H., Ichien, N. T., Qu, Y. (2015). Mothers know best: redirecting adolescent reward sensitivity toward safe behavior during risk taking. Social Cognitive and Affective Neuroscience, 10(10), 1383-1391.
Unsworth, N., Schrock, J. C., \& Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 1302-1321.
Willoughby, T., Good, M., Adachi, P. J. C., Hamza, C., Tavernier, R. (2013). Examining the link between adolescent brain development and risk-taking from a social-developmental perspective. Brain and Cognition, 83, 315-324.
Zangeneh, M., Blaszczynski, A., Turner, N. (2008) (Eds). In the pursuit of winning: problem gambling theory, research and treatment. New York: Springer.

# On an effective and efficient method for exploiting "wisdom of crowds in one mind" 

Itsuki Fujisaki (bpmx3ngj@gmail.com)<br>Graduate School of Arts and Sciences, The University of Tokyo<br>3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902

Hidehito Honda (hitohonda.02@gmail.com)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902
Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902


#### Abstract

Previous studies have shown that one can exploit "wisdom of crowds" by oneself. This is achieved by aggregating multiple "quasi-independent" estimates from the same person. However, previous methods were not necessarily easy to utilize. Therefore, we propose an efficient method based on perspective-taking. The procedure is as follows: First, one makes her/his own estimation. Second, one estimates again based on a different perspective ("general public"). Then these two estimations were averaged. Two experiments revealed that our method effectively induced the wisdom of crowds by oneself. More importantly, participants in our method made estimations more quickly than those in a previously proposed method, suggesting that our method required a relatively diminished cognitive load for participants. Further investigation suggested that our method was immune to adverse effects of confidence. Therefore, the present findings show that our method could be effective and efficient method for inducing the wisdom of crowds in one mind.


Keywords: Estimation; Judgments under uncertainty; Perspective taking; Judgment aggregation; Wisdom of crowds; Wisdom of crowds in one mind

## Introduction

"Wisdom of crowds" (Surowiecki, 2004) is the well known phenomena, such that the aggregation of multiple estimates made by large number of people is more accurate than the estimate of a single individual. Recently, an intriguing concept, termed "wisdom of crowds in one mind" has been discussed in the research field of judgment and decision making (e.g., Rauhut \& Lorenz, 2011). In examining this issue, researchers have discussed how a single person can exploit "wisdom of crowds" in her/his mind. Herzog and Hertwig (2014a) argued that this is achieved by averaging "quasi-independent" multiple estimates from the same person. A person's estimate is not always constant and has some variance, even for the same problem. Using such inconstancy and variance, s/he can exploit the "wisdom of crowds in one mind". For example (see Figure 1), there is a question with correct answer $50 \%$. A person's first estimate was $30 \%$. Imagine that $\mathrm{s} /$ he was asked to make the second estimate, then her/his second estimate was $80 \%$. The average of two estimates would be $55 \%$; a result more
accurate than the first estimates. Previous studies have proposed some methods about how to exploit the wisdom of crowds in one mind. Vul \& Pashler (2008) proposed a method in which individuals make estimations for the same problem twice, with a time lag (2-weeks) between estimations. Herzog \& Hertwig (2009) proposed the method called dialectical bootstrapping. In this method, people are asked to make an estimation twice. In the second estimation, they are provided an instruction (shown in Table 1, see "dialectical" condition). This instruction asks individuals to provide a different estimation from the first one, by considering new knowledge that was once overlooked, or searching out incorrect assumptions or considerations present in the first estimate. These studies generally showed that the average of the first and second estimates were more accurate than the first estimate, and that the benefit of averaging was larger than in the control condition (i.e., just making estimations twice without any time lag or instructions). Effectiveness of these methods has been


Figure 1. An example of wisdom of crowds in one mind. The red numbers indicate the distance between first or averaged estimate and correct answer.
confirmed repeatedly in subsequent studies (see Herzog \& Hertwig, 2014b for review).

However, these methods are not necessarily easy to utilize. In the method proposed by Vul and Pashler (2008),
the time lag suggested (i.e., 2 weeks) is necessary to exploit the wisdom of crowds within the mind of a single individual. In the dialectical bootstrapping proposed by Herzog and Hertwig (2009), a rather complicated instruction (Table 1) is necessary. Furthermore, the first estimate has to be presented to the person, which may not be necessarily efficient as a method for inducing the second estimate. In the present study, we propose a new method for exploiting the wisdom of crowds in one mind, based on previous findings on perspective-taking.

## Perspective-taking of the "general public" for exploiting the wisdom of crowds in one mind

## Perspective-taking

Perspective-taking is the cognitive action to take another point of view. This topic has been examined mainly in the research field of social psychology. Previous studies have revealed that perspective-taking changes peoples' subjective judgment or preference, such as stereotypes about a person (Galinsky \& Moskowitz, 2000). Furthermore, recent studies have indicated that perspective-taking prompts people to change estimates about objective values (e.g., the population of the city or the date of historical event). Yaniv \& Choshen-Hillel (2012) implemented a perspective-taking procedure in the advice-taking paradigm (for more information on the advice-taking paradigm, see Bonaccio \& Dalal, 2006 for review). In this experiment, participants were asked to generate an estimate about questions by taking into account estimated values by others (these values were given as advice). At this point, participants who were asked to generate estimation at the point of another person's perspective tended to accept advice more than those who were asked to generate their own estimation.

Based on these findings, it is predicted that estimations vary depending on the perspective an individual takes. Accordingly, we propose a method for utilizing wisdom of crowds in one mind by taking others' perspective. In our method, we implemented a perspective-taking procedure in the multiple estimates for the same question, to exploit the wisdom of crowds in one mind. Specifically, we propose the following procedure:
(1): A person makes her/his own estimation.
(2): S/he makes a second estimation based on a different perspective. Specifically, s/he takes perspective of the "general public" (Table 1, see "self-others" condition).
(3): Averaging the first and second estimations.

## Merits of perspective-taking of the "general public"

We believe that taking the perspective of the "general public" has some merits. First, different estimation may be easily induced. Although estimations by the same person tend to be analogous (e.g., control condition in Herzog \& Hertwig, 2009), people tend to believe that they differ from the general public in some ways. For example, in comparing driving ability, people tend to think that their ability is better than the general public (e.g., Svenson, 1981), suggesting that they believe "I am different from the general public!" Thus, a different estimation may be induced by taking the perspective of "general" individuals. Second, a different estimate may be induced irrespective of confidence about the first estimate. In making estimations, if people are confident about their estimates, they may not change their estimate when they are asked to make the second estimate. Previous studies on overconfidence (e.g., Koriat, Lichtenstein, \& Fischoff, 1980) reveal that people tend to be overconfident about the accuracy of their estimation. Thus, in inducing a different estimate from the same person, confidence (especially, as individuals tend to be overconfident) may adversely affect results, as s/he may be reluctant to change her/his first estimate when $s / h e$ is confident about the first estimate. For example, in the dialectical bootstrapping method (e.g., Herzog \& Hertwig, 2009), people are asked to make the "self" second estimate, even though they are provided an instruction to make a "different" estimate. Therefore, this method may be affected by the confidence in the first estimate, and an appropriately different estimate may not be induced in the second estimate. However, our method may correct for this. The current procedure asks individuals to make the second estimate from "other" people's perspective. Hence, it is expected that our method is relatively immune to the adverse effect of confidence, and the nature of changing the estimate, irrespective of confidence, may result in the effective utilization of the wisdom of crowds in a single mind. Third, anyone can imagine the general public. Leboeuf, Shafir, \& Bayuk (2010) showed that when participants were asked to

Table 1. Full text in instruction about three conditions.
Condition Instruction in the second estimate

Self-others How do you think people in general estimate about the following question? Make a second estimate after considering fully how people in general estimate about this.

Dialectical First, assume that your first estimate is off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Was the first estimate rather too high or too low? Fourth, based on this new perspective, make a second, alternative estimate.

Self-twice No instruction
take the perspective of a particular group ("family" in Study 2), the participants who actually identified with this group ("have a family" in Study 2) preferred the choice corresponding to the perceived group perspective (e.g., a family vacation). However, participants who were not associated with this type of group (e.g., single) were not influenced by this perspective-taking, suggesting that a perspective-taking procedure cannot work effectively for individuals that are not actually associated with the group. Given that people can compare their driving ability with general public (Svenson, 1981), people may be able to imagine the "general" public.

In the following sections, we shall report two behavioral experiments and discuss the effectiveness of our method.

## Experiment 1

We conducted a web-based behavioral study in order to examine whether our method effectively induced wisdom of crowds in one mind.

## Method

Participants A total of 452 participants were recruited for this experiment through a Japanese internet research company. Participants were randomly assigned into one of three experimental conditions (self-others, $n=150$; dialectical, $n=151$; self twice, $n=151$ ).
Tasks and materials Participants were asked to answer eight general knowledge questions, such as "What percent of the world's airports are in the United States?" (Vul \& Pashler, 2008). Participants answered questions twice following instructions (specific content will be reported the below).
Procedure In all conditions, participants first provided their own estimates about the questions. After answering all the eight questions, participants provided second estimates following the instruction for each condition (see Table 1). In the self-others condition, we instructed participants to take a "general public" perspective. In the dialectical condition, we gave the instruction of the dialectical bootstrapping based on Herzog \& Hertwig (2009). In the self-twice condition, no instruction was provided and participants just made second estimate again.

The order of the questions for the first estimate was randomized across participants and that for the second estimate was the same as in the first estimate.

## Analysis

In the following analyses, we calculated "\% MAD (= mean absolute distance) reduction averaging" (Herzog \& Hertwig, 2014a) as an index for the gained accuracy of averaging. First, absolute distance between an estimate and the correct answer was calculated per question for each participant. The mean values of these were computed as MAD. MAD indicated MAD of the first estimates, and $M A D_{\text {avg }}$ represents the MAD of the averaged estimates. Then, "\% MAD reduction averaging" was calculated for each
participant level as follows: $\left(\mathrm{MAD}_{1}-\mathrm{MAD}_{\text {avg }}\right) / \mathrm{MAD}_{1} .{ }^{1}$ See Figure 2 for examples.


$$
\% \text { MAD reduction averaging }=\frac{\text { Mean } A D_{1}-\text { Mean } A D_{\text {ave }}}{\text { Mean } A D_{1}}
$$

Figure 2. Examples of AD (absolute distance) and formula of $\%$ MAD reduction averaging. (a) is an example when averaged estimate is more accurate than first estimate and (b) is an example when averaged estimate is less accurate than first estimate.

## Results and discussion

Figure 3 shows \% MAD reduction averaging for each condition. In the self-others and dialectical conditions, \% MAD reduction averagings were significantly higher than zero (self-others: $\mathrm{M}=2.52, \mathrm{CI}=[0.08,4.75]$; dialectical: M $=2.20, \mathrm{CI}=[0.50,4.03])$. Therefore, averaged estimates reduced error compared to first estimates in these conditions. In contrast, in the self-twice conditions, \% MAD reduction averaging was not significantly higher than zero ( $\mathrm{M}=$ $1.34 \%, 95 \% \mathrm{CI}=[-3.43,0.67])$. Thus, this method did not


Figure 3. \% MAD reduction averaging.

[^331]significantly reduce error compared to the first estimates. A pairwise Wilcoxon rank sum test using a Bonferroni correction revealed that in the self-others condition and dialectical condition, \% MAD reduction averagings were significantly higher than that in self-twice condition ( $p$ $<.01 ; p<.05$, respectively). These results indicate that our method can exploit accuracy of averaging more effectively than the method without any instruction, as with the method proposed in Herzog \& Hertwig (2009).

Although no significant differences between self-others and dialectical condition was found ( $p>.1$ ), our method can exploit accuracy of averaging at least as effectively as the method proposed by Herzog \& Hertwig (2009) given that the mean value of $\%$ MAD reduction averaging in the selfothers condition was higher than that in the dialectical condition.

## Experiment 2

In Experiment 2, we used the same procedure with the following two exceptions. First, we measured the response time for the second estimate. Second, participants rated confidence about their first estimates, and we analyzed the relationship between the confidence and difference in the two estimates (i.e., first and second estimates).

## Method

Participants Japanese graduates and undergraduates from the University of Tokyo ( $N=77 ; 56$ men and 21 women; age $\mathrm{M}=20.90$, sd $=2.52$ ) participated in this experiment. They were randomly assigned into one of three experimental conditions (self-others, $n=25$; dialectical, $n=$ 24; self-twice, $n=28$ ).
Tasks and materials Participants were asked to answer twenty questions about general knowledge based on Herzog \& Hertwig (2014a). Questions were answered twice, with instructions, as in Experiment 1. In addition, in making first estimations, participants were also asked to rate their confidence for each estimation.
Procedure The experiment was individually conducted using a computer. In all conditions, participants first answered their own estimates about the questions, and rated confidence about their estimates on a 100-point scale. After answering all 20 questions, second estimates were made, following instructions, as in Experiment 1. In the self-others condition, we instructed participants to take the "general public" perspective. In the dialectical condition, we gave the instruction based on Herzog \& Hertwig (2009). In the selftwice condition, no instruction was given, and participants simply provided a second estimation.

The order of the questions for the first estimate was randomized across participants and that for the second question was the same as in the first estimate.

## Results and discussion

## Accuracy of averaging

In the following analysis, as in Experiment 1, we calculated "\% MAD reduction averaging" as an index for the gained accuracy of averaging. Figure 4 shows \% MAD reduction of the averaging for each condition. In the self-others condition, \% MAD reduction averaging was significantly higher than zero ( $\mathrm{M}=5.51 \%, 95 \% \mathrm{CI}=[1.26,9.56]$ ). In contrast, in the dialectical and self-twice conditions, \% MAD reduction averagings were not significantly higher than zero (dialectical: $\mathrm{M}=2.14, \mathrm{CI}=[-1.18,5.42]$; selftwice: $\mathrm{M}=1.60, \mathrm{CI}=[-0.24,3.54]$ ). Thus, these methods did not significantly reduce error compared to the first estimates.

Although a pairwise Wilcoxon rank sum test using a Bonferroni correction revealed that in the self-others condition, \% MAD reduction averaging was not significantly higher than that in the dialectical and the selftwice conditions ( $p s>.1$ ), the $\%$ MAD was higher in the self-others condition compared to the dialectical and the self-twice conditions. Given that the \% MAD reduction of


Figure 4. \% MAD reduction averaging.


Figure 5. Total response time in the second estimate.
the averaging was significantly higher than zero only in the self-others condition, our method could exploit accuracy of averaging effectively, compared to other methods.

## Response time in the second estimates

We analyzed total response time for the second estimate. Figure 5 shows total response time for each condition (selfothers: $\mathrm{M}=262.27, \mathrm{CI}=[222.80,303.69]$; dialectical: $\mathrm{M}=$ 412.19, CI = [327.23, 499.90]; self-twice: $\mathrm{M}=193.30, \mathrm{CI}=$ [164.67, 234.09].

Total response times were log-transformed and a pairwise t-test using a Bonferroni correction was conducted. It was found that in the self-others condition, participants finished second estimates more quickly than the dialectical condition ( $p<.05$ ). Total response time in self-others condition was longer than that in self-twice condition ( $p<.001$ ). These results suggest that participants in the self-others condition could exploit wisdom of crowds in one mind with diminishing more cognitive load than those in the dialectical condition.

## Further examination about the three methods

The difference in first and second estimates To confirm that our method exploited the difference of estimates, in the following analysis, we calculated median absolute distance (AD) to examine the distance between first and second estimates (Herzog \& Hertwig, 2014a). Median AD refers to the median absolute distance between the first and second estimates across 20 questions.

Figure 6 shows Median AD for each condition. In the self-others and dialectical conditions, median ADs were both larger than that in the self-twice condition (self-others: $\mathrm{M}=8.18, \mathrm{CI}=[7.02,9.28]$; dialectical: $\mathrm{M}=10.02, \mathrm{CI}=$ [8.48, 11.79], self-twice; $\mathrm{M}=2.45, \mathrm{CI}=[1.46,3.46]$; pairwise Wilcoxon rank sum test using a Bonferroni correction: $p s<.001$ ). There was no difference between self-others and dialectical condition ( $p>.1$ ). Therefore, these results showed that our method could induce the difference of estimates as in the dialectical bootstrapping.

Relationship between the difference in first and second estimates and confidence We examined the relationship between differences in the first and second estimates and confidence. Generally, if a person is confident about the first estimate, s/he may not change the second estimate. We predicted that since the participants in the dialectical and self-twice were asked to make "self" estimations, a negative correlation might be observed between the difference in the first and second estimates and confidence (i.e., s/he might not change the second estimate when $\mathrm{s} / \mathrm{he}$ was confident about the first estimate). However, this might not be true for the self-other condition because a person was asked to make estimate from other people's perspective in her/his second estimate.

We analyzed the relationship between the difference in the first and second estimates and the confidence about first
estimate. Absolute distance between the two estimates was calculated for each question within participants, and a correlation coefficient between the absolute difference and confidence in the first estimate was calculated for each participant. ${ }^{2}$

Figure 7 shows distributions of correlation coefficients for each condition. In the dialectical and self-twice conditions, $95 \%$ confidence intervals about correlation coefficients were less than zero (dialectical: $\mathrm{M}=-0.19, \mathrm{CI}=$ [-0.29, -0.086]; self-twice: $\mathrm{M}=-0.13, \mathrm{CI}=[-0.21,-0.060]$ ). In contrast, for the self-others condition, $95 \%$ confidence intervals about correlation coefficients included zero (selfother: $\mathrm{M}=0.03, \mathrm{CI}=[-0.040,0.10])$. These results indicate that participants in the self-other condition tended to make different estimations between the first and second estimations, irrespective of their confidence. However,


Figure 6. Median absolute distance.


Figure 7. Correlation between confidence and distance.

[^332]participants in the dialectical condition made analogous estimation in the second estimate as in that in the first estimate when they were confident in the first estimate. Given that wisdom of crowds in one mind tends to work when a person makes different estimations in the two estimations (e.g., Herzog \& Hertwig, 2009), our method is relatively immune to the adverse effect of confidence compared to the dialectical bootstrapping.

Furthermore, we analyzed the relationship between the accuracy and the confidence in the first estimate. The absolute distance between the correct answer and the first estimate were calculated for each question within participants, as an index for accuracy, and then we calculated a correlation coefficient between these two values for each participant. Figure 8 shows distributions of correlation coefficients. $95 \%$ confidence intervals included zero ( $\mathrm{M}=-0.0032$, $\mathrm{CI}=[-0.013,0.0080]$ ). This result indicates that confidence was not related with the actual accuracy.


Figure 8. Distributions of correlation coefficient between confidence and distance.

## General discussion

In the present study, we proposed a new method for utilizing the wisdom of crowds in one mind, and we examined whether our method was effective when compared to another method proposed in previous studies. Our findings were summarized as follows: First, we found that our method effectively induced the wisdom of crowds in one mind. Second, it was found that participants in our method made estimations more quickly compared to those in the previous method, suggesting that our method diminished cognitive load for participants. Third, we found that our method was relatively immune to adverse effects (e.g., confidence), given that the previous methods require a time lag or presentation of the first estimate (Vul \& Pashler, 2008; Herzog \& Hertwig, 2009).

Taken together, we believe that our method can be a more effective and efficient method for inducing wisdom of crowds in one mind.

## Acknowledgments

This research was supported by Grants-in-Aid for Scientific Research (A, 16H01725) and Young Scientists (B, 16K16070).

## References

Bonaccio, S., \& Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational Behavior and Human Decision Processes, 101, 127-151.
Galinsky, A. D., \& Moskowitz, G. B. (2000). Perspectivetaking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. Journal of Personality and Social Psychology, 78, 708-724.
Herzog, S. M., \& Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. Psychological Science, 20, 231237.

Herzog, S. M., \& Hertwig, R. (2014a). Think twice and then: Combining or choosing in dialectical bootstrapping? Journal of Experimental Psychology Learning Memory and Cognition, 40, 218-232.
Herzog, S. M., \& Hertwig, R. (2014b). Harnessing the wisdom of the inner crowd. Trends in Cognitive Sciences, 18, 504-506.
Koriat, A., Lichtenstein, S., \& Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6(2), 107-118.
LeBoeuf R. A., Shafir E., Bayuk J.B. (2010) The conflicting choices of alternating selves. Organizational Behavior and Human Decision Processes, 111, 48-61.
Surowiecki, J. (2004). The wisdom of crowds. New York, NY: Doubleday.
Svenson, O. (1981). "Are we all less risky and more skillful than our fellow drivers?". Acta Psychologica. 47, 143148.

Rauhut, H., \& Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. Journal of Mathematical Psychology, 55, 191-197.
Vul, E., \& Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. Psychological Science, 19, 645-647.
Yaniv, I., \& Choshen-Hillel, S. (2012). When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advicegiving. Journal of Experimental Social Psychology, 48, 1022-1028.

# Experimental Investigation on Top-down and Bottom-up Processing in Graph Comprehension and Decision 

Misa Fukuoka (fukuoka@cog.human.nagoya-u.ac.jp)<br>Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)<br>Akihiro Maehigashi (mhigashi@cog.human.nagoya-u.ac.jp)<br>Graduate School of Informatics, Nagoya University, Japan


#### Abstract

Graph comprehension requires both bottom-up processing from the graph representation and top-down processing guided by knowledge and attitude. In the current study, we investigated which of the bottom-up process phases: extraction, interpretation, and decision: were affected by the top-down processing derived by the impressions and social attitudes. The experimental results showed that the top-down processing driven by impressions temporarily formed in specific contexts affected both the extraction of information and the following decision phase whereas top-down processing driven by attitude formed over a long time based on social norms affected only the decision phase. In the latter case, a decision was made without any need for bottom-up processing.


Keywords: graph representation, decision, bottom-up processing, top-down processing

## Introduction

People make decisions by perceiving and understanding external resources. Visual representations such as graphs are effective for such processes. Visual representations are known for making the understanding of information easier, as graphs and diagrammatic presentations including scatter plots reduce cognitive errors (Ancker, Senathirajah, Kukafka, \& Starren, 2006; Lipkus \& Hollands, 1999). On the other hand, graphical representations can also bias interpretation: e.g., Woller-Carter et al. (2012) confirmed that an intentionally biased graph produces information reading errors.

Previous studies have confirmed that understanding graphs is best achieved from both top-down and bottom-up processing. Freedman and Shah (2002) proposed a CI model for graph comprehension based on the text understanding CI model proposed by Kintsch (1988). The model assumed that there was an interaction between the two information processing stages; bottom-up from the visually represented information encoded in the external resources and top-down from knowledge stored in long-term memory.

Many studies have shown that the bottom-up processing of graphs depends on the graphical representations. Shah and Carpenter (1995) found that the contents of reading information changed depending on which of the independent variables were assigned to the x -axis or the graph legend. Sanchez and Wiley (2006) found that fascinating and attractive visual information distracted participant focus from the crucial information related to the primary task, which implied that such information should be carefully restrained to ensure participant focus on the target information.

In top-down processing, many experimental findings have been reported that have found that a participant's knowledge and attitudes toward the topics significantly influence an understanding of the information in the graphs. Freedman and Smith (1996) found in their experiments with scatter plot graphs that the participants read the information not only using the bottom-up processing that arises from perceiving plot patterns in the graph, but also using top-down processing which is based on pre-formed knowledge. Kanzaki and Miwa (2012) found that in experiments with line graphs, information reading was performed with both bottom-up and top-down processing, but the top-down processing did not violate the reading of the bottom-up processing.

CaMeRa (Tabachneck-Schijf, Leonard, \& Simon, 1997) demonstrated two stages of bottom-up processing. In the preceding extraction stage, primary symbolic information was drawn from the visual representation of the graph: e. g., a $y$-axis value in the experimental condition in which an experimental manipulation was made was greater than the value in the control condition, and there was a substantial difference between the two conditions. Such difference perceptions at this stage were formed based mainly on the perceptual information processing in the working memory without accessing domain knowledge in declarative memory. In the following interpretation stage, the experimental results represented on the graph were interpreted with an integration of the symbolic information drawn from the graphical representation and knowledge stored in declarative memory: e.g., a medical material x is effective for improving activity y .

In actual situations, the final decision stage, may follow from an understanding of the information drawn from the graph: e. g., sales promotion for the medical material x was decided on.

Figure 1 shows the bottom-up processing series: extraction, interpretation, and decision.

The first aim of the current study was to confirm the bottom-up processing series shown above. In particular, we examined whether internally extracted information, such as mentally represented difference between experimental conditions, that is constructed from information represented on the graph as an external resource, drove the following bottom-up information processing. We should note that how information is extracted from external resources is generally different for each person who reads the graph, implying that there is a possibility that different internal information could be extracted from identical external resources. Therefore, in


Figure 1: Top-down/ bottom-up processing model of graph comprehension and decision.
this study, we set up difference perception stage in which internal information is constructed from "extracted" information on the graph.

In Experiment 1, we constructed an experimental setting in which we had participants represent perceived difference independently from the external resource. Through the experiment, we explicitly confirmed that based on the internally represented information, the following interpretation and decision stages were performed.

Next, we considered impressions and attitudes to be the factors that drove the top-down processing. Impressions are usually temporarily constructed with insufficient information based on stimuli presented in a situation or a context (Wang \& Nelson, 2014). On the other hand, attitudes seem to be more continuously formed based on social and moral norms, and are directed toward people, places, and social policies (Greenwald \& Banaji, 1995).

The second aim of this study was to understand which of the initial or final bottom-up processing stages was affected by the top-down processing derived from impressions and attitudes. The research question was whether top-down processing affected only the final decision stage, or also affected the initial extraction. Many studies on human perception have indicated that top-down processing derived from beliefs and knowledge bias human perception; however, previous studies on graph comprehension have investigated top-down processing using the dependent variables related to the latter bottom-up processing phases.

In Experiment 2, we performed an experiment to examine the top-down processing derived from impressions; and in Experiment 3, we further investigated the top-down processing based on attitudes.

## Experiment 1

Experiment 1 consisted of Experiments 1 a and 1 b .

## Participants

56 undergraduates ( 29 males and 27 females: $M_{\text {age }}=18.64$, $S D_{\text {age }}=0.82$ ) from Nagoya University participated in Experiment 1a. The experiment was performed in small groups of at most eight members. 54 undergraduates from Nagoya University ( 30 males and 24 females: $M_{\text {age }}=18.46$, $S D_{\text {age }}=0.97$ ) participated in Experiment 1b. The experiment was performed as class practice in a cognitive science class.

## Materials

In Experiment 1a, an experimental context was introduced in which an ingredient $A$ that was expected to improve biological vitality was assumed, and the effect of ingredient A was examined using a laboratory rat experiment with a hamster wheel. Participants were presented with a graph that indicated the number of cases out of 20,000 in which the rats continued to perform the hamster wheel task for more than three minutes. Figure 2 shows the presented graphs in which there is a substantial difference in the number of cases for the experimental rat group in which ingredient A was given to the rats and the control group in which it was not.

Figure 3 shows the graphs used in Experiment 1b. The graphs had perceptual features identical to those in Experiment 1a but the information content was different. The perceptual features of the graphs implied that there was a difference in the number of cases between the two conditions; however, the reality was the same. The scalability of the $y$ axis of the graphs was manipulated; as a result, across all graphs, the numbers of cases were only 1190 (5.95\%) for the experimental group and $1110(5.55 \%)$ for the control group.

## Procedure

Based on the presented graphs, participants were instructed to indicate their opinion as an expert advisor about a nursing tonic that an assumed company was developing to improve biological vitality.

The experimental procedure consisted of the following four stages.
Graphical presentation Participants were assigned to one of three groups. Each group consisted of 18, 19, and 19 participants respectively in Experiment 1a and 19, 18, and 17



Figure 2: Graphs used in Experiment 1a.


Figure 3: Graphs used in Experiment 1b.
participants in Experiment 1b. The participants in each group were presented with one of the three graphs shown in Figures 2 or 3 . A score of 1 to 3 was assigned to each of the three types of graphs for the following regression analysis.
Difference perception After the graphical presentation, participants estimated the degree of difference in the number of cases between the two conditions and were asked if there were any differences in the two conditions on a five-point scale from 1 (strongly disagree) to 5 (strongly agree).
Interpretation Then, the participants were required to estimate to what degree ingredient A was effective in improving biological vitality on a five-point scale.
Decision Finally, the participants were asked whether ingredient A should be included in the nursing tonic the company was developing on a five-point scale.

The experiment took about 30 minutes.

## Result

Single regression analyses were performed with the score recorded in one of the four experimental stages as the independent variable and the score in the following stage as the dependent variable. Figure 4 shows the results of Experiment 1a. There were significant relationships found between the presented graphs and the difference perceptions ( $\beta=.51, t(54)=4.36, p<.001, R^{2}=.26$ ), the difference perceptions and the interpretations ( $\beta=.42, t(54)=3.35, p$ $<.01, R^{2}=.17$ ), and the interpretations and the decisions ( $\beta$


Figure 4: Results of the single regression analyses for Experiment 1a.


Figure 5: Results of the single regression analyses for Experiment 1b.
$\left.=.48, t(54)=4.01, p<.001, R^{2}=.23\right)$. The same analyses were performed for Experiment 1 b and Figure 5 shows the results, in which there were significant relationships found between the difference perceptions and the interpretations ( $\beta$ $\left.=.42, t(52)=3.30, p<.01, R^{2}=.17\right)$, and the interpretations and the decisions ( $\beta=.58, t(52)=5.11, p<.001, R^{2}=.33$ ); however, no significant relationships were detected between the presented graphs and the difference perceptions $(\beta=.14$, $\left.t(52)=1.01, p=.32, R^{2}=.02\right)$.

## Experiment 2

Experiment 2 was performed to investigate how top-down processing driven by impressions affected each of the bottom-up processing phases confirmed in Experiment 1. Experiment 2 consisted of the former and the latter sessions. For each participant, the former session was followed by the latter session; and between the two sessions, a 5 minute break was inserted.

## Participants

60 undergraduates ( 30 males and 30 females: $M_{\text {age }}=18.87$, $S D_{\text {age }}=0.65$ ) from Nagoya University participated in Experiment 2. The experiment was performed in small groups of at most eight members.

## Materials

The former session in Experiment 2 was performed to replicate Experiment 1b. In the latter session, the graphs were the same as those in Experiment 1b, but a further experimental setting was introduced to manipulate the participants' impressions toward the medical material in which a fictional ingredient, a "proten" or a "rubison," rather than ingredient $A$ was assumed. For the manipulation, reading material that described one of the two pharmaceutical companies was used to give the participants positive impressions about the one company developing the "proten" and negative impressions of the other company developing the "rubison."

In Experiment 2, two experimental contexts were introduced, with the assignment of each of the two contexts to the former or the latter session being counterbalanced. One context was the same as that in Experiment 1: i.e., the development of a medical ingredient to improve biological vitality was introduced. The other context was the
development of an ingredient to recover physical strength after fatigue. In the latter case, a rat experiment was assumed with a 100 meter running test.

## Procedure

Based on the graphs that indicated the experimental results, participants were instructed to indicate their opinion as an expert advisor about a nursing tonic to improve biological vitality, or as an expert advisor about an energy drink to recover physical strength after fatigue.

The experimental procedure consisted of the following five phases.
Impression manipulation Before the graphical presentation stage, in the former session, participants answered a questionnaire about their impressions of an ingredient A that had been developed by an company X they belonged to. In this stage, no information about the ingredient was provided. They estimated, on a five-point scale, their impressions about ten items, such as "ingredient A is reliable," with a higher estimation score meaning more positive impressions toward ingredient A .

The following four experimental stages, graphical presentation, difference perception, interpretation, and decision, were the same as in Experiment 1. Participants were assigned to one of three groups. Each group consisted of 20 participants respectively. The participants in each group were presented with one of the three graphs. In the final decision stage, participants were asked to decide whether ingredient A should be included in the nursing tonic or the energy drink, on a five-point scale.

In the latter session, the impressions of the ingredients "proten" or "rubison" were manipulated using reading materials in which information about company Y which is developing the ingredient was given. First, the participants were presented with a text describing the information about the company Y. One text included characteristics of an excellent company (e.g., There is an excellent welfare program in company Y.), to persuade participants to have a positive impression of the ingredient "proten." The other text had characteristics of an evil company (e.g., There is no welfare program in company Y.), and persuaded participants to have negative impressions toward the ingredient "rubison." One of the two texts was presented to each participant. Then, the participants answered the same questionnaire as used in the former session, in which they were asked to give their impressions of the ingredient "proten" or "rubison."

From the graphical presentation through to the decision stages, the same procedure was utilized as in the former session.

The total time for Experiments 2 was about an hour.

## Result

Multiple regression analyses were performed with two independent variables; i.e., a score recorded in one of the four experimental stages and an impression score; and a dependent variable; i.e., a score from the following phase. Figure 6 shows the results from the former session. The
results replicated Experiment 1b. There were significant relationships found between the difference perceptions and interpretations ( $\beta=.59, t(57)=5.50, p<.001, R^{2}=.38$ ), and the interpretations and the decisions $(\beta=.66, t(57)=6.30, p$ $<.001, R^{2}=.42$ ); however, no significant relationship was detected between the presented graphs and the difference perceptions ( $\beta=.23, t(57)=1.82, p=.07, R^{2}=.11$ ). There were no relationship found between the impressions and any of the three bottom-up processing phases; difference perception, interpretation, or decision $(\beta=.24, t(57)=1.95$, $p=.06, R^{2}=.11 ; \beta=.09, t(57)=.84, p=.40, R^{2}=.38 ; \beta=.05$, $\left.t(57)=.49, p=.63, R^{2}=.42\right)$.

The same analysis was performed for the latter session. Figure 7 shows the results, from which it can be seen that the difference perceptions were affected by the impressions ( $\beta$ $\left.=.28, t(57)=2.25, p<.05, R^{2}=.11\right)$ but not by the graphical presentations $\left(\beta=.18, t(57)=1.42, p=.16, R^{2}=.11\right)$, the interpretations were not affected by the impressions ( $\beta=.21$, $\left.t(57)=1.99, p=.05, R^{2}=.44\right)$ but by the difference perceptions ( $\beta=.58, t(57)=5.61, p<.001, R^{2}=.44$ ), and the decisions were affected by both the impressions ( $\beta=.31$, $\left.t(57)=3.27, p<.01, R^{2}=.57\right)$ and the interpretations $(\beta=.58$, $\left.t(57)=6.20, p<.001, R^{2}=.57\right)$.


Figure 6: Results of the multiple regression analyses for the former session in Experiment 2.


Figure 7: Results of the multiple regression analyses for the latter session in Experiment 2.

## Experiment 3

Experiment 3 was performed to investigate how top-down processing driven by attitudes affected each of the bottom-up processing phases. In Experiment 3, we investigated how the participants' social attitudes toward smoking affected their decisions about a smoking cessation policy.

## Participants

55 undergraduates ( 33 males and 22 females: $M_{\text {age }}=18.51$, $S D_{\text {age }}=0.86$ ) from Nagoya University participated in Experiment 3. The experiment was performed as class practice in a cognitive science class.

## Materials

An experimental context was introduced in which a health survey was conducted in an assumed city X. Participants were presented with graphs that indicated the survey results. Specifically, the graph showed how many of the 20,000 respondents suffered from pulmonary problems. The perceptual features of the graphs were the same as those in the preceding experiments. It was assumed that one group had a family with a smoker and the other group did not.

## Procedure

One week before the experiment, the participants answered a questionnaire to measure their social attitudes toward smoking. They estimated, on a five-point scale, their attitudes toward smoking behavior for 10 items, such as "smoking is only malevolent for society," with a higher estimation score indicating greater negative attitudes toward smoking.

Participants were assigned to one of three groups. Each group consisted of 18,17 , and 19 participants respectively. The participants in each group were presented with one of the three graphs. In the final decision stage, participants were asked to indicate their opinion, as a health consultant, about whether or not employees in an assumed company should be prohibited from smoking both inside and outside the company.

## Result

The same multiple regression analyses as those in Experiment 2 were performed. Figure 8 shows the result. The difference perceptions were not affected by the graphical presentations or by the attitudes $(\beta=.03, t(52)=.23, p=.82$, $\left.R^{2}=.002 ; \beta=.03, t(52)=.20, p=.84, R^{2}=.002\right)$, the interpretations were affected only by the difference perceptions ( $\beta=.66, t(52)=6.61, p<.001, R^{2}=.48$ ) but not by the attitudes ( $\left.\beta=.16, t(52)=1.63, p=.11, R^{2}=.48\right)$, and the decisions were affected only by the attitudes ( $\beta=.50$, $t(52)=4.18, p<.001, R^{2}=.28$ ) but not by the interpretations ( $\beta=.09, t(52)=.71, p=.48, R^{2}=.28$ ).


Figure 8: Results of the multiple regression analyses for Experiment 3.

## Discussion

The first aim of this study was to confirm a series of the bottom-up process phases, information extraction from graphs as an external resource, interpretation, and decision.

In Experiment 1a, bar graphs with assumed experimental results in two conditions were used. In Experiment 1a, we manipulated the differences in the values of the independent variable in two experimental conditions, and confirmed that bottom-up processing was driven by internally represented differences extracted from the graph as the external resource. In Experiment 1b, we used another set of graphs in which the values for the independent variable were equivalent, and obtained the same results as in Experiment 1a, indicating that bottom-up processes also arise based on the participants' internally represented differences. It is important that bottomup processing is performed from internally represented information that is extracted from an external resource, and independent of the actual information represented in the external resource.

In Experiment 1b, we used graphs in which the visually represented differences were equivalent to the actual differences in the graphs used in Experiment 1a, even thought there was no actual difference in the two conditions. As a result, no correlation was found between the internal differences represented by the participants and the pseudo differences in the graphs. This indicated that the participants in this study were not affected by the visual biases included in the external resource when extracting the information in the graphs. Previous studies have found that the reading quality of graphical information depends on critical thinking capabilities (e.g., Woller-Carter et al., 2012). The critical thinking abilities of the participants in our experiments appeared to be relatively high, which needs to be considered when interpreting the experimental results in this current study.

In Experiment 2, we manipulated the participants' impressions of the target topic. First, we found the same bottom-up process that had been confirmed in Experiment 1b in which top-down processing was not assumed.

The first finding was that the final decision was made based on both top-down processing derived from impressions and bottom-up processing from the extraction of information from the external resource. Previous studies have consistently confirmed that impressions significantly affect decision making. Kostopoulou et al. (2017) experimentally found that in medical diagnoses by home doctors, the first impression of the patients affected diagnostic planning decisions. Jaros et al. (2000) reported that a perceptual impression of foods at a glance affected food selection.

The latter session in Experiment 2 confirmed that topdown processing using impressions affected the extraction phase in the initial bottom-up processing phase. The halo effect is when the impressions about one specific characteristic affect the estimation of other characteristics that may not even be related to the initial characteristic (e.g., Murphy et al., 1993). In the latter session, we manipulated the impression of the office environments of an assumed
company that had developed a medical ingredient. The participants' estimation about the ingredient depended on the manipulated impressions, even though the efficacy of the ingredient had no explicit relationship with the office environments.

In Experiment 3, we manipulated another factor that drives top-down processing; that participant attitudes are formed over a long period based on moral and social norms. The experimental results showed that different from Experiment 2, the final decision was made only based on the participants' attitudes toward the target topic, rather than depending on the interpretation drawn from the bottom-up processing.

This finding about the relationship between attitude and decision was consistent with findings in previous studies. It has been found that attitudes are crucial in predicting behavior (Conner \& Armitage, 1998), and that there is a strong relationship between attitudes and behavior (Fazio et al, 1982).

An important point from Experiment 3 is that when making the final decision, top-down processing was not concerned with the bottom-up processing output. The topic dealt with in the current study was smoking; therefore, as this was a familiar social topic for everyone, this may have driven the strong top-down processing.

The results in Experiment 3 showed that the top-down processing did not affect the bottom-up processing initial extraction phase, implying that such initial bottom-up processing may be isolated by top-down processing.

In the current study, we examined two factors; impressions and social attitudes; that drive top-down processing. In summary, when one factor that is formed temporarily is followed by a specific context, impressions take a central role in the comprehension and decision making about graphical information, indicating that bottom-up processing functions are compatible with top-down processing. On the other hand, when another top-down factor is socially formed over a long period, such as social attitudes, bottom-up processing tends to be separated from top-down processing, with top-down processing predominating.

## References

Ancker, J. S., Senathirajah, Y., Kukafka, R., \& Starren, J. B. (2006). Design features of graphs for communicating health risks: A systematic review. Journal of the American Medical Informatics Association, 13(6), 608-618.
Conner, M., \& Armitage, C. J. (1998). Extending the theory of planned behavior: A review and avenues for further research. Journal of Applied Social Psychology, 28(15), 1429-1464.
Fazio, R. H., Chen, J., McDonel, E. C., \& Sherman, S. J. (1982). Attitude accessibility, attitude-behavior consistency, and the strength of the object evaluation association. Journal of Experimental Social Psychology, 18(4), 339-357.
Freedman, E. G., \& Smith, L. D. (1996). The role of data and theory in covariation assessment: Implications for the
theory-ladenness of observation. Journal of Mind and Behavior, 17(4), 321-343.
Freedman E. G., \& Shah P. (2002). Toward a model of knowledge-based graph comprehension. Lecture Notes in Computer Science, 18-30.
Garcia-Retamero R., \& Cokely, E T. (2013). Communicating health risks with visual aids. Current Directions in Psychological Science, 22(5), 392-399.
Greenwald, A. G., \& Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4-27.
Jaros, D., Rohm, H., \& Strobl, M. (2000). Appearance properties - A significant contribution to sensory food quality? -. LWT - Food Sci. Technol., 33, 320-326.
Kanzaki, N., \& Miwa, K. (2012). Experimental investigation of the effects of representations and perspectives for the comprehension of line graphs. The Japanese Journal of Psychology, 83(3), 163-173.
Kintsch, W. (1988). The role of knowledge in discourse comprehension. A construction-integration model. Psychological Review, 95(2), 163-182.
Kostopoulou, O., Sirota, M., Round, T., Samaranayaka, S., \& Delaney, B. C. (2017). The role of physicians' first impressions in the diagnosis of possible cancers without alarm symptoms. Medical Decision Making, 37(1), 9-16.
Lipkus, I. M., \& Hollands, J. G. (1999). The visual communication of risk. Journal of National Cancer Institute Monographs, 25, 149-163.
Murphy, K. R., Jako, R. A., \& Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. Journal of Applied Psychology. 78(2), 218-225.
Sanchez, C. A., \& Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. Memory and Cognition, 34(2), 344-355.
Shah, P., \& Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. Journal of Experimental Psychology: General, 124(1), 43-61.
Tabachneck-Schijf, H. J. M., Leonardo, A. M., \& Simon, H. A. (1997). CaMeRa: A computational model of multiple representations. Cognitive Science, 21(3), 305-350.
Wang, Z., \& Nelson, M. R. (2014). Tablet as human: How intensity and stability of the user-tablet relationship influences users' impression formation of tablet computers. Computers in Human Behavior, 37, 81-93.
Woller-Carter, M. M., Okan, Y., Cokely, E. T., \& GarciaRetamero, R. (2012). Communicating and distorting risks with graphs: An eye - tracking study. Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting, 1723-1727.

# Measures and mechanisms of common ground: backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions 

Riccardo Fusaroli (fusaroli@cc.au.dk)<br>Kristian Tylén (kristian@dac.au.dk)<br>School of Communication and Culture \& The Interacting Minds Center<br>Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark<br>Katrine Garly (katrinegarly@hotmail.com)<br>Jakob Steensig (linjs@cc.au.dk)<br>School of Communication and Culture, Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus C, Denmark<br>Morten H. Christiansen (morten@cas.au.dk)<br>Department of Psychology, Cornell University, Ithaca, NY 14853<br>School of Communication and Culture \& The Interacting Minds Center Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

## Mark Dingemanse (mark.dingemanse@mpi.nl)

Max Planck Institute for Psycholinguistics
PO Box 3106500 AH Nijmegen The Netherlands


#### Abstract

A crucial aspect of everyday conversational interactions is our ability to establish and maintain common ground. Understanding the relevant mechanisms involved in such social coordination remains an important challenge for cognitive science. While common ground is often discussed in very general terms, different contexts of interaction are likely to afford different coordination mechanisms. In this paper, we investigate the presence and relation of three mechanisms of social coordination - backchannels, interactive alignment and conversational repair - across free and task-oriented conversations. We find significant differences: task-oriented conversations involve higher presence of repair - restricted offers in particular - and backchannel, as well as a reduced level of lexical and syntactic alignment. We find that restricted repair is associated with lexical alignment and open repair with backchannels. Our findings highlight the need to explicitly assess several mechanisms at once and to investigate diverse social activities to understand their role and relations.


Keywords: Social coordination, common ground; conversational repair; interactive alignment; backchannel.

## Introduction

A key question in cognitive science is how people coordinate knowledge and behavior in social interaction, a process sometimes referred to as grounding (Clark \& Brennan, 1991; Dale, Fusaroli, Duran, \& Richardson, 2013). Research over the past decades has highlighted processes like backchannels, conversational repair, and interactive alignment, but progress has been hampered by two challenges. First, these processes are rarely considered together, limiting our view of possible interrelations. Second, the study of such processes has been spread across disciplines and data types, limiting the possibilities for
prediction and generalization (de Ruiter \& Albert, 2017). Here we report on a principled comparison of backchannels, repair, and interactive alignment in two quite different types of contexts: free (FC) and task-oriented conversations (TOC). Engaging in conversation is a collaborative effort involving timely coordination at many levels. In their seminal 1991 article, Clark and Brennan (1991) suggest such coordination to be contingent on common ground, comprising mutual knowledge, beliefs and assumptions. The main mechanism for establishment and maintenance of common ground explored in their work is backchannels (Yngve, 1970). Backchannels are phatic signals such as head nods, eye blinks and vocal expressions of the type $u h$ huh, yeah, and okay (Bangerter \& Clark, 2003; Schegloff, 1982). In this study, we are concerned with vocal backchannels only. Even if such signals are often quite subtle, research suggests that speakers are very sensitive to these kinds of cues as ways of providing and monitoring positive evidence of mutual understanding, and their interruption can have detrimental effects on communication (Clark \& Krych, 2004).

A related phenomenon is conversational repair. While backchannels are mostly concerned with positive evidence of understanding, conversational repair refers to the interactional practices by which people signal and solve trouble in conversation (Schegloff, Jefferson, \& Sacks, 1977). Here we focus on the most interactive form of repair: other-initiated repair, a highly frequent conversational sequence where one participant initiates the repair procedure by means of a request for clarification like huh? or who?, and the other completes it. Formats for other-initiated repair frequently show lexical and syntactic repetition (Jefferson, 1972; Sacks, 1992), with a recent cross-linguistic study of informal conversation finding that $48 \%$ of all repair
initiating turns repeated part or whole of the prior turn (Dingemanse et al., 2015). Repair initiations can be ordered along a cline from weak (providing little indication of what or where the problem is, such as huh?) to strong (highlighting a specific element of a prior turn for clarification or confirmation, as in who?). While weak repair initiation is always possible, and so might be expected to be a default option, it has been suggested that the selection of repair formats follows a 'strongest initiator rule' (Clark \& Schaefer, 1989), according to which people select the most specific repair format possible, given constraints like noise and joint attention.

While backchannels and conversational repair are often thought to be of a more explicit, inferential character, the theory of interactive alignment in conversation suggests common ground to be established through low-level automatic priming processes (Pickering \& Garrod, 2004). It has been observed across many contexts and studies that interlocutors engaged in conversation often tend to adapt to each other's linguistic behaviors on many levels from prosody to syntax (Fusaroli \& Tylén, 2016). If an interlocutor, for instance, uses the phrase "I'm sure" to express confidence, there is an enhanced probability that the conversational partner will use similar wording later in the conversation even if other expressions would work just as well in that context (Fusaroli et al., 2012). Alignment is thought to percolate between levels of linguistic representation. Lexical alignment can for instance facilitate the alignment of other levels (e.g. syntactic choices), eventually leading to alignment of situation models of the ongoing activity, that is, common ground (Pickering \& Garrod, 2004). According to Pickering and Garrod, more explicit negotiation of common ground such as repair and backchannels are only recruited in cases of communication problems or misunderstandings. It should be noticed that a few alternative perspectives have been suggested: Brennan and Clark (1996) associate alignment with more explicit negotiations of shared conceptual representations, while others suggest a context-sensitive mechanism, which strategically selects for alignment or divergence according to the functional needs of the ongoing activity (Fusaroli, Raczaszek-Leonardi, \& Tylén, 2014; Healey, Purver, \& Howes, 2014). Common ground is often discussed as a unified concept foundational to conversation in general. However, different contexts of conversation are likely to afford different degrees of explication as well as different processes and mechanism for the establishment of common ground. Conversations among pilots and airport control towers thus require high levels of referential precision (Prinzo \& Britton, 1993), while the average dinner conversation may be more concerned with maintenance of social relations (Dunbar, Marriott, \& Duncan, 1997), to the point of ignoring referential misalignments (Galantucci \& Roberts, 2014).

To establish a more refined framework for the investigation of common ground, we propose an integrative approach comparing backchannel, alignment and repair
across diverse social activities. In particular, we focus on free spontaneously occurring interactions - traditionally favored by conversation analytic approaches - and a welldefined spatial navigation task to be jointly solved through conversation - traditionally favored by more cognitive and quantitative approaches. The investigation aims to determine how common ground is negotiated and maintained, and whether these processes are modulated by the social context. Moreover, we will investigate how the suggested mechanisms of common ground relate to each other: e.g. if repair and alignment are associated (Dingemanse et al., 2015), then measures of alignment will be tapping into both mechanisms.

We predict that (i) the different social contexts involve distinctive patterns in the dynamics and mechanisms of common ground, such that they allow an accurate classification of conversations as free or task-oriented. More specifically, (ii) baseline frequency of repair, interactive alignment, and backchannels may be higher in task-oriented interactions due to requirements for referential precision. (iii) The quality of the dynamics at work is also predicted to be different. Particularly, task-oriented conversations, aimed at coordinative precision and more tightly constrained by the lab context, will feature more restricted forms of repair. By contrast, free interactions, which often happen in more noisy environments and incorporate a wider range of activities, will involve more open forms of repair. (iv) The different indices of common ground are not independent from each other. For instance, we expect repair and alignment to be related because the repetition of linguistic forms across speakers is a key formal measure of both. The latter point underlines the importance of considering possible relations between repair and alignment when discussing measures and mechanisms of social coordination.

## Methods

## Free Social Conversations

We used 18 conversations from the DanTIN corpus (Steensig et al., 2013), 10 minutes per conversation, for a total of 4954 speech turns. Data collection was limited to spontaneous, naturally occurring conversations between families and friends. Participants often engaged in additional activities during these conversations (e.g., eating, or playing games). The average conversation involved 275 speech turns ( $\mathrm{SD}=55$ turns), with an average of 58 turns per interlocutor ( $\mathrm{SD}=52$ ). All conversations were in Danish. The corpus reflects the diversity of free social interactions, with seven conversations involving 2 participants, seven with 3, two with 4 , two with 5 participants, and one with 7 participants. We are currently extending this corpus to be better able to model the effects of number of interlocutors and participation framework.

## Task-oriented Conversations

We used the 44 task-oriented conversations that make up the DanPass corpus (Grønnum, 2009), totaling 9448 speech
turns. The conversations were aimed at solving the Map Task (Anderson et al., 1991). The average length of each conversation was 7.4 minutes ( $\mathrm{SD}=3$ ), with an average 214 turns per conversation ( $\mathrm{SD}=85$ ), and 107 per interlocutor ( $\mathrm{SD}=42$ ). All conversations were in Danish and involved only 2 participants in separate booths. We are currently extending this corpus to include co-present interlocutors.

## Backchannels

Backchannels were manually coded in $10 \%$ of the transcripts (2 free social and 5 task-oriented conversations). Based on this, an automatic procedure for coding backchannels was developed, based on turn length ( $<4$ words) and presence of the words "ja", "nej", "okay", "nå", "jo", "mmhm", "jamen", "mmm", and "åh". The system achieved substantial intercoder agreement with the manual coding (Kappa=0.62). As we are currently extending and validating the coding scheme, results based on the current version of this measure should be treated with caution.

## Conversational Repair

A trained analyst identified sequences of other-initiated repair and classified them according to three crosslinguistically attested format types: open request, which
 ra. restricted request, which restricts the problem space by requesting clarification of a specific element of the problematic turn; and restricted offer, which offers a candidate perception or understanding for confirmation (Dingemanse \& Enfield, 2015). In addition, 10\% of the transcripts ( 3 free social and 5 task-oriented conversations) were analyzed by a second coder. We obtained substantial intercoder reliability, corresponding to a Kappa of 0.67 for repairs in general, and 0.79 respectively for open and restricted repairs (the latter breaking down to 0.4 for restricted requests and 0.38 for restricted offers).

## Interactive Alignment

We calculated lexical and syntactic interactive alignment on a turn-by-turn basis. Each turn was lemmatized using the CST lemmatizer for Danish (Jongejan \& Haltrup, 2005) and parts of speech tagged using DKIE (Derczynski, Field, \& Bøgh, 2014). Lexical alignment was calculated as the cosine similarity between lemmatized words in adjacent speech turns uttered by different interlocutors. Syntactic alignment was calculated as the cosine similarity between 2 -grams

parts-of-speech in adjacent speech turns uttered by different interlocutors. To avoid possible lexical alignment confounds we regressed it out of syntactic alignment as in (Hopkins, Yuill, \& Keller, 2016).

## Data Analysis

To assess whether the free social vs. task-oriented nature of the conversations affected the development and maintenance of common ground, we employed mixed effects regression models to predict the presence of conversational repair (binomial variable), interactive alignment (continuous variable) and backchannels (binomial variable) on a turn-by-turn basis. We employed Task (binomial variable, FC vs. TOC), and Time within conversation (count variable, quantified as number of turns from the start) as fixed effects and conversation as random effects, including a random slope for Time. When the model did not converge, we removed first the random slope, then the fixed effect of Time. To determine whether the different social coordinative mechanisms are related to each other, we employed two mixed effects regressions. The first predicted the overall amount of repair initiations in a conversation (count variable) from the amount of backchannels and alignment, controlling for the offset of overall amount of conversational turns. The second predicted the presence of repair at a turn level (binomial variable) from the presence of backchannel and the level of alignment of that same turn. Finally, to establish how distinctive these mechanisms are, we produced a 5 -fold cross-validated predictive regression assessing whether one could use the presence and amount of conversational repair, interactive alignment and backchannels to identify the nature of the conversation. All analyses were run using R 3.3.2, RStudio 1.0.136, lme4 1.112, irr 0.84 and tidyverse 1.1.0.

## Results

## Backchannels

Backchannels were highly frequent in the corpora ( $54 \%$ of the speech turns), and more so in TOC ( $58 \%$ of speech turns), than in FC ( $48 \%$ of speech turns): $\beta=0.6, S E=0.04$, $p<.001$ (see Figure 1). Backchannels also increased over time ( $\beta=0.07, S E=0.03, p=.03$ ), but not differently in the two corpora ( $\beta=-0.02, S E=0.04, p=.57$ ).


## Conversational Repair

Conversational repair was highly frequent across both corpora, in line with previous findings in 12 other languages (Dingemanse et al., 2015). Repair initiations made up 3\% of speech turns, with an average 45.59 seconds ( $S D=54.8$ ) and 34.03 speech turns ( $S D=41.83$ ) between successive repairs. Task Oriented Conversations showed a higher frequency of repair ( $\beta=0.4, S E=0.18, p=.0274$ ) than Free Conversations, with 31.13 turns ( 51.13 seconds) between repair initiations in the former and 39.65 turns (61.3 seconds) between repair initiations in the latter; see Figure 1. Time was not a significant main effect $(\beta=-0.0007, S E=$ $0.0006, p=.3$ ), nor did it interact with Task ( $\beta=0.0007, S E$ $=0.001, p=.6$ ).

Open repair was much more frequent ( $38.5 \%$ of repair) in free social interactions, than in task oriented interactions (4\% of repair): $\beta=-2.82, S E=0.39, p<.001$. Open repair tended to decrease in frequency as conversations proceed ( $\beta$ $=-0.005, S E=0.003, p=.0607$ ), with no interaction with interaction type $(\beta=0.007, S E=0.004, p=.1275)$. Restricted request repair was not significantly different between corpora (FC: $17 \%$, TOC: $22 \%$ ): $\beta=0.4, S E=0.34$, $p=.23$ ). There was a marginal tendency for restricted request repair to increase over time ( $\beta=0.003, S E=0.0016$, $p=.0804$ ), with no interaction with interaction type ( $\beta=$ $0.002, S E=0.004, p=.5$ ), but the model did not converge with these factors. Restricted offer repairs were more frequent in TOC ( $74 \%$ of repair) than in FC ( $45 \%$ of repair): $\beta=1.97, S E=0.48, p<.001$. There was no significant main effect of time $(\beta=0.003, S E=0.0024, p=.185)$, but a significant interaction with interaction type ( $\beta=-0.006, S E$ $=0.003, p=.0415$ ) indicating a decrease in the TOC over time, but not in FC (see Figure 2).


Figure 2: Distribution of repair types in the two corpora.

## Interactive alignment

As illustrated by Figure 1, syntactic and lexical alignment is significantly lower in TOC than FC (lexical: $\beta=-0.14, S E=$ $0.15, p<.001$; syntactic: $\beta=-0.04, S E=0.009, p<.001$ ). Alignment significantly decreases over time for lexical ( $\beta=$ $-0.0002, S E=0.00001, p<.001)$, but not syntactic $(\beta=-$ $0.00005, S E=0.00004, p=.8$ ), and the decrease significantly interacts with Task, being smaller in TOC (lexical: $\beta=0.0001, S E=0.00007, p=.008$; syntactic: $\beta=$ $0.007, S E=0.00005, p<.001)$. These patterns hold when varying distance between speech turns (alignment over longer stretches of conversation, up to 5 turns of distance); increasing the unit of analysis (up to 4-grams of lexical or
syntactic units); or controlling for increased alignment in repair turns.

## Relations between repair, alignment and backchannels

The general level of conversational repair in a conversation was positively associated with the level of backchannel ( $\beta=$ $0.01, S E=0.003, p<.001$ ) and syntactic alignment $(\beta=$ $0.44, S E=0.18, p=.014$ ) and negatively associated with lexical alignment ( $\beta=-0.46, S E=0.19, p=.014$ ). At a turn level, conversational repair was associated with increased lexical alignment ( $\beta=0.98, S E=0.19, p<.001$ ), a result driven by the two restricted repair formats. A follow-up explorative analysis indicates that in TOC, alignment is indeed much higher for turns containing repair initiations than for other turns (Lexical: 0.168 vs. 0.102; Syntactic 0.091 vs. 0.048 ), but not so in FC (Lexical: 0.113 vs. 0.112 ; Syntactic: 0.074 vs. 0.087 ). Interactions between the different indexes did not significantly improve the likelihood of the model.

## Social coordinative mechanisms as discriminative patterns

Employing a combination of repair, interactive alignment and backchannels information, we were able to classify the transcripts according to their interaction type with an accuracy of $83.82 \%$ ( $95 \%$ CIs: $77.46 \%-88.97 \%$ ), a sensitivity of $84.71 \%$ and a specificity of $82.95 \%$, over chance accuracy of $51.46 \%$. General levels of repair in a conversation alone gave an accuracy of $61.4 \%$ ( $95 \%$ CIs: $53.7 \%-68.74 \%$ ). Interactive alignment gave an accuracy of 80.7\% (95\% CIs: 73.98\%-86.33\%). Backchannel gave an accuracy of $63.16 \%$ ( $95 \%$ CIs: $55.46 \%-70.39 \%$ ).

## Discussion

In this study, we compared different ways in which common ground may be established and maintained across both taskoriented and free conversations. We predicted that (i) we would find distinctive patterns of repair, backchannels and alignment. In particular, (ii) task-oriented conversation should show higher rates of repair, backchannels and alignment; (iii) task-oriented conversation should show lower rates of open requests for clarification; (iv) repair and alignment should be correlated because of the high frequency of repetition in restricted repair formats.

We found full support for (i): knowing the amount of repair, backchannels and alignment present in a conversation enables accurate ( $>80 \%$ ) discrimination between task-oriented and free conversations. We also found partial support for (ii): higher rates of backchannels and repair in TOC but not alignment; full support for (iii): lower rates of open requests, making restricted offers the most frequent in task-oriented interaction; partial support for (iv): lexical, but not syntactic alignment, is correlated with repair, an effect driven particularly by restricted repair formats. As such, our preliminary findings shed new light
on the relations between backchannels, conversational repair and interactive alignment as measures of social coordination. If further confirmed with a more controlled dataset, they might also help clarify the relations between informal and task-oriented interactions.

Four findings stand out. First, FC and TOC present clear differences in the mechanisms employed to negotiate and maintain common ground. As solving the MapTask requires the construction of a shared representation of the space to navigate and its landmarks, we observe more explicit negotiation (repair) and confirmation of common ground. In line with our previous work, we also find that alignment seems less crucial in TOC. This could be a consequence of a division of labor leading to complementary rather than repeated lexical and syntactic structure among individuals solving a task (Fusaroli, et al., 2012; Fusaroli \& Tylén, 2016). The higher alignment in FC might also indicate the prevalence of less explicit mechanisms to negotiate common ground, less likely to lead to face-loss (Bjørndahl, Fusaroli, Østergaard, \& Tylén, 2015; Brown \& Levinson, 1978). However, as a previous study reports an opposite result with TOC showing higher alignment than FC, ongoing work is implementing more conservative and comparable techniques, such as the use of surrogate pair (composed of interlocutors from different conversations) baselines (Healey, Purver, \& Howes, 2010; Hopkins, et al., 2016). Analogously, further investigation of the temporal decrease of alignment is warranted.

Second, repair in task-oriented interaction is strongly skewed towards restricted formats and particularly the restricted offer format. This provides novel support for the 'strongest initiator rule’ (Clark \& Schaefer 1989), according to which participants initiate repair using the strongest repair initiator possible given the circumstances. Prior work based on informal interaction found that noise and parallel involvements increased the likelihood of open repair (Dingemanse \& Enfield, 2015; Dingemanse et al., 2015), essentially by making it comparatively harder to initiate repair using restricted formats, which require having heard and understood as least part of the problematic source turn. Here we replicate this finding in an informal corpus of Danish interaction, and add a direct comparison with taskoriented interaction. The task-oriented condition takes away some common causes of perceptual and attentional difficulties, which should push people towards using more specific repair formats. This prediction is indeed met: in task-oriented interaction, the most specific ('strongest') repair initiation format is also the most commonly used.

Third, repair and alignment are intertwined. For instance, consider the following example, in which the restricted request consists in the repetition of all the words of the previous sentence, albeit in a slightly different order:

A: Vi var i Ikea [We were at Ikea]
B: Var I i Ikea, dig og ? [You were at Ikea, you and ?]
A: $m m h$

While alignment has often been cast as an implicit, automated background process, and repair as its explicit, and much rarer, "friend in need," our parallel investigation of repair and alignment reveals that widely used formal measures of alignment also pick up many restricted repair sequences. This is no surprise -after all, the crucial role of repetition in repair sequences has long been known- but it does point to the need for a reappraisal of the relationship between repair and alignment. Our findings suggest that the evidential base for a large part of the alignment literature may include many explicit repair sequences, belying the common assumption that alignment is an automated, lowlevel process.

Fourth, the combination of backchannels, repair and alignment allows us to classify interaction type with a high degree of accuracy. While these results still need to be generalized to a wider range of social activities and contexts, they open up new avenues for the possibility of classifying discourse data and contributing to the growing field of computer-assisted studies of dialog structure.
Although very encouraging, our findings should be viewed as preliminary. The two corpora differ in several aspects: the presence of a task: interlocutors' physical copresence (in FC but not TOC), number of interlocutors involved in the conversations ( 2 in TOC, 2 or more in FC), and familiarity (possibly higher in FC ). All these aspects are likely to affect conversational dynamics. We are currently extending the corpora to include full variation along these dimensions and account for them in the statistical analyses.

## Conclusions

A comparative assessment of three mechanisms for the negotiation of common ground - backchannels, conversational repair and interactive alignment - highlights important differences in free and task-oriented conversations, plausibly related to situational features and task demands. Our results point to interactions between these mechanisms, e.g. with restricted repair feeding lexical alignment, which suggests future research should further disentangle their reciprocal role. As a theory-driven quantitative comparative study of conversations, our approach shows how insights from conversation analysis, cognitive science and natural language processing can be combined to contribute to a cumulative science of human interaction.

## Acknowledgments

We thank Amalie Staggemeier for her help with the coding of the materials, Nina Grønnum for sharing the DanPass corpus, and Andreas Roepstorff for insightful discussions. The project was funded through a seed grant from the Interacting Minds Center, Aarhus University.

## References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., . . . Miller, J. (1991). The

HCRC map task corpus. Language and speech, 34(4), 351-366.
Bangerter, A., \& Clark, H. H. (2003). Navigating joint projects with dialogue. Cognitive Science, 27(2), 195-225.
Bjørndahl, J., Fusaroli, R., Østergaard, S., \& Tylén, K. (2015). Agreeing is not enough: The constructive role of miscommunication. Interaction Studies, 16(3), 495-525.
Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory and Cognition, 22(6), 1482-1493.
Brown, P., \& Levinson, S. C. (1978). Universals in language usage: Politeness phenomena Questions and politeness: Strategies in social interaction: Cambridge University Press.
Clark, H. H., \& Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine \& S. D. Teasley (Eds.), Perspectives on Socially Shared Cognition. Washington, DC: American Psychological Association.
Clark, H. H., \& Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of memory and language, 50(1), 62-81.
Clark, H. H., \& Schaefer, E. F. (1989). Contributing to discourse. Cognitive Science, 13, 259-294.
Dale, R., Fusaroli, R., Duran, N., \& Richardson, D. C. (2013). The self-organization of human interaction. Psychology of Learning and Motivation, 59, 43-95.
de Ruiter, J. P., \& Albert, S. (2017). An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology. Research on Language and Social Interaction, 50(1), 90-107.
Derczynski, L., Field, C. V., \& Bøgh, K. S. (2014). DKIE: Open Source Information Extraction for Danish. Paper presented at the EACL.
Dingemanse, M., \& Enfield, N. J. (2015). Other-initiated repair across languages: towards a typology of conversational structures. Open Linguistics, 1(1), 96-118.
Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., . . . Manrique, E. (2015). Universal principles in the repair of communication problems. PLOS ONE, 10(9), e0136100.
Dunbar, R. I. M., Marriott, A., \& Duncan, N. D. C. (1997). Human conversational behavior. Human Nature, 8(3), 231-246.
Fusaroli, R., Bahrami, B., Olsen, K., Rees, G., Frith, C. D., Roepstorff, A., \& Tylén, K. (2012). Coming to terms: an experimental quantification of the coordinative benefits of linguistic interaction. Psychological Science, 23, 931-939.
Fusaroli, R., Raczaszek-Leonardi, J., \& Tylén, K. (2014). Dialog as interpersonal synergy. New Ideas in Psychology, 32, 147-157.

Fusaroli, R., \& Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. Cognitive Science, 40(1), 145-171.
Galantucci, B., \& Roberts, G. (2014). Do we notice when communication goes awry? an investigation of people's sensitivity to coherence in spontaneous conversation. PLOS ONE, 9(7), e103182.
Grønnum, N. (2009). A Danish phonetically annotated spontaneous speech corpus (DanPASS). Speech Communication, 51(7), 594-603.
Healey, P., Purver, M., \& Howes, C. (2010). Structural divergence in dialogue. Paper presented at the In Proceedings of the Conference on Architectures and Mechanisms for Language Processing.
Healey, P., Purver, M., \& Howes, C. (2014). Divergence in dialogue. Plos One, 9(6).
Hopkins, Z., Yuill, N., \& Keller, B. (2016). Children with autism align syntax in natural conversation. Applied Psycholinguistics, 37(2), 347-370.
Jefferson, G. (1972). Side sequences. In D. N. Sudnow (Ed.), Studies in social interaction. New York, NY: Free Press.
Jongejan, B., \& Haltrup, D. (2005). the CST Lemmatiser. Center for Sprogteknologi, University of Copenhagen version, 2.
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27, 169-190.
Prinzo, O. V., \& Britton, T. W. (1993). ATC/pilot voice communications-a survey of the literature: DTIC Document.
Sacks, H. (1992). Lectures on Conversation, 2 Vols, ed. G. Jefferson, with introductions by EA Schegloff: Oxford: Blackwell.
Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. Analyzing discourse: Text and talk, 71, 93.
Schegloff, E. A., Jefferson, G., \& Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. Language, 361-382.
Steensig, J., Brøcker, K. K., Grønkjær, C., Hamann, M. G. T., Hansen, R. P., Jørgensen, M., . . . Pedersen, H. F. (2013). The DanTIN project-creating a platform for describing the grammar of Danish talk-ininteraction. Paper presented at the New Perspectives on Speech in Action: Proceedings of the 2nd SJUSK Conference on Contemporary Speech Habits, Samfundslitteratur, Frederiksberg.
Yngve, V. H. (1970). On getting a word in edgewise. Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, 567-577.

# How do Speakers Coordinate Planning and Articulation? Evidence from GazeSpeech Lags 

Chiara Gambi (cgambi@exseed.ed.ac.uk)<br>Department of Psychology, 7 George Square<br>Edinburgh, EH8 9JZ UK<br>Matthew Crocker (crocker@coli.uni-sb.de)<br>Department of Computational Linguistics and Phonetics, C7.1 University Campus<br>Saarbrücken, 66123 Germany


#### Abstract

How do speakers coordinate planning and articulation of more than one word at the same time? Here, we test whether they dynamically estimate how long it takes to (i) plan and (ii) articulate the words they intend to produce as a means of achieving such coordination. German speakers named two pictures without pausing, while their eye-movements were recorded. In line with previous reports, after their gaze left the first picture, speakers took longer to start speaking (i.e., the gaze-speech lag was longer) when the name of the first picture was shorter. But while gaze-speech lags were also longer when the second picture was harder to name, the two effects did not interact. We argue that speakers' flexible planning abilities might be accounted for by reactive, rather than proactive planning mechanisms.


Keywords: planning; estimation; duration; coordination; gaze-speech lag.

## Introduction

Speakers plan ahead of articulation (Griffin \& Bock, 2000; Levelt, Roelofs, \& Meyer, 1999) and usually complete lexico-semantic planning for at least a whole phrase (Martin, Crowther, Knight, Tamborello, \& Yang, 2010; Smith \& Wheeldon, 1999), phonological planning for a whole word (Meyer, 1996; Smith \& Wheeldon, 2004), and articulatory planning for a whole syllable (Meyer, Roelofs, \& Levelt, 2003) before beginning to speak. While this allows for rapid and fluent speech production, the incremental nature of planning also raises the question of how speakers manage the timely coordination of planning and articulatory processes.

## Mechanisms for Flexible Advance Planning

Several studies have uncovered regularities in speakers' amount of advance planning; for example, suggesting that speech onsets are comparable across short and long words because speakers usually complete articulatory processing for only the first syllable of a word prior to speech onset (Damian, Bowers, Stadthagen-Gonzalez, \& Spalek, 2010; see also Meyer, Roelofs, \& Levelt, 2003). Importantly, in recent years it has also become clear that the amount of advance planning speakers perform is not fixed, but rather varies with properties of both the utterance and the speaker's recent experience (Konopka, 2012; Van de Velde,

Meyer, \& Konopka, 2014), or task context (Meyer et al., 2003).

We thus know a great deal about factors that can influence the coordination of planning and articulatory processes. On the contrary, we know very little about the mechanisms that underlie the timely coordination of these processes. For example, we know that planning style is influenced both by the accessibility of linguistic units, and by the ease-of-apprehension of the referent (Konopka \& Meyer, 2014). This suggests that the mechanism involved must be sensitive to the difficulty of the planning process at different stages, but there are at least two ways in which such a mechanism could operate.

One possibility is that a flexible planning system operates reactively. Once speakers encounter some difficulty, a compensatory mechanism is triggered. For instance, if accessing a particular word is difficult (i.e., takes time), attention might be (temporarily) shifted to another process (e.g., retrieving a different word). But in addition, the planning system might, at least in part, allocate resources to different processes in a proactive manner. Such a proactive planning mechanism could be learnt from previous experience with producing language (in general, or within a particular task), and indeed there is evidence that planning style can be primed by previous experience with the same sentence structure (Van de Velde et al., 2014). A proactive planning mechanism would of course be beneficial in maximizing fluency, as it may allow speakers to avoid future difficulties, by anticipating their likely occurrence and taking appropriate steps before they even arise. This idea is reminiscent of some models of motor control (e.g., Wolpert and Flanagan, 2011).

## Proactive Planning: Candidate Evidence?

To our knowledge, no language production study has investigated this issue directly. However, in one seminal study, Griffin (2003) suggested that speakers might estimate how long both planning and articulating a word will take, and then combine such estimates to determine how to time one process with respect to the other in order to minimize future disfluencies (i.e., to plan proactively).

To illustrate, imagine a speaker of German preparing to produce Abschlussballkleider (dresses for the high-school prom). Let us assume, for the purpose of illustration, that
the speaker retrieves Abschlussball (prom) and Kleider (dresses) separately (Sandra, 1990) ${ }^{1}$. If so, the speaker will need to get the first syllable of Kleider ready to be articulated by the time articulation of Abschlussball is ending. To do this, the speaker could estimate both how long it will take her to get Kleider ready (i.e., retrieval difficulty) and how long it will take her to say Abschlussball (i.e., articulation duration). She could then determine that she will probably have enough time to prepare Kleider while saying Abschlussball (a long word), so she can start speaking right away. But if the first word is short, such as Sport in Sporttitelseite (sport title page), she might instead have to delay speech onset in order to prepare more of the second word before starting to speak. Similarly, she may need to delay speech onset if the second word is particularly difficult to retrieve.

However, the evidence in support of Griffin's proposal is currently somewhat mixed. Griffin (2003) asked speakers to name two pictures one after the other, without pausing, while their eye-movements were recorded. Critically, the name of the picture that was mentioned first (word ${ }_{l}$ ) could be either short (monosyllabic) or long (plurisyllabic). In this task, speakers usually shift their gaze from the first to the second picture as soon as they have retrieved the phonological representation for word $_{1}$ (Griffin, 2001; Meyer \& Van der Meulen, 2000). The gaze shift generally occurs before overt articulation of word $_{l}$. Importantly, the interval between this gaze shift and speech onset (i.e., the gazespeech lag) is longer when word ${ }_{l}$ is shorter (a reversed word-length effect). According to Griffin (2003), this shows that speakers estimate word $_{1}$ duration: they begin speaking earlier (with respect to the gaze shift, thus leading to a longer gaze-speech lag), when word $_{l}$ is shorter in order to have more time to retrieve the second picture's name ( word $_{2}$ ) before, rather than during, articulation of word ${ }_{l}$.

However, while Meyer, Belke, Häcker, and Mortensen (2007) replicated Griffin's finding ${ }^{2}$, they also provided a different explanation. We know that speakers may begin retrieving the articulatory code of the first syllable of a word as soon as they complete phonological processing for this syllable (i.e., without waiting for phonological processing of the whole word to be completed); in turn, as soon as they have retrieved such code, they can begin speaking. But if word $_{1}$ is monosyllabic, the moment of the gaze shift (which coincides with completion of phonological processing for the whole word; see above) also happens to coincide with the start of articulatory retrieval. As a result, the gazespeech lag will last at least the time it takes to perform

[^333]articulatory retrieval for one syllable. By contrast, for a polysyllabic word $_{l}$ the gaze shift occurs only later (once articulatory retrieval of the first syllable is already underway), thus leading to a shorter lag.

## This study

If Meyer et al.'s (2007) explanation is correct, then the reversed-length effect on gaze-to-speech lags is not evidence that speakers estimate duration, contrary to Griffin's (2003) suggestion. Moreover, neither study demonstrates that speakers can combine estimates of retrieval difficulty with estimates of duration, because they did not manipulate the difficulty of retrieving word $_{2}{ }^{3}$. Here, we provide a test of this hypothesis: If speakers take into account not only word ${ }_{1}$ length (monosyllabic vs. polysyllabic words), but also word $2_{2}$ retrieval difficulty, gaze-speech lags should be affected by both variables. Moreover, the effects of the two variables should interact, reflecting the workings of a proactive planning mechanism underlying the tight coordination of articulation (of word ${ }_{l}$ ) and planning (of word ${ }_{2}$ ).
For word $_{2}$, we chose a manipulation that is both known to reliably affect the earliest stages of picture naming, and very easy to identify for participants: Pictures were either visually intact or degraded (see Figure 1). We reasoned this would provide the most favorable test of Griffin's proposal, as speakers were placed in ideal conditions for estimating the difficulty of retrieving word $_{2}$; although degradation does not affect the difficulty of retrieving word ${ }_{2}$ directly, it makes accessing the corresponding concept more difficult, which then has a knock-on effect on the time it takes to fully prepare word $_{2}$. To give speakers ample opportunity to adjust to the relevant level of difficulty, and to avoid carryover effects, degradation varied between participants.
As in previous studies, the gaze-speech lag should be longer when word ${ }_{1}$ is shorter. In addition, if speakers can estimate retrieval difficulty, it should also be longer when word $_{2}$ takes longer to retrieve. Crucially, there should be a significant interaction, with word $_{2}$ difficulty having a larger effect when word ${ }_{l}$ is shorter. As there is less scope for completing word ${ }_{2}$ retrieval during the articulation of word $_{1}$ when word $_{1}$ is short, speakers should aim to complete most of word $_{2}$ retrieval before speech onset; instead, when word ${ }_{1}$ is long, the speaker can benefit from extra time after the onset of articulation, and increases in word ${ }_{2}$ retrieval difficulty may not affect the gaze-to-speech lag as strongly.
If Meyer et al.'s proposal is correct, however, the gazespeech lag should only depend on word ${ }_{l}$ length, and the reversed-length effect on gaze-speech lags would not be evidence for a proactive planning mechanism. Given the potential theoretical relevance of Griffin's (2003) original interpretation of her findings, testing her claim in full, as we do in this study, would advance our understanding of the

[^334]mechanisms underlying flexible planning in language production.

## Method

## Participants

Thirty-two native speakers of German ( 24 female, $\mathrm{M}_{\text {age }}=$ 23.8 yrs, $\mathrm{SD}=2.6$ ), with self-reported normal vision and no language impairments, were paid 8 euros/hour to participate in this and another eye-tracking experiment (not reported here). One participant was replaced because of excessive head movements. Sample size was determined on the basis of previous research (Griffin, 2003; Meyer et al., 2007)

## Materials

We selected 128 black and white line drawings from the picture naming norms of Bates, et al. (2003). Of these, 64 pictures with high name agreement were used as left pictures. Left pictures were named first, so we refer to the left picture names as word ${ }_{l}$. For half the items (Long), word $_{1}$ ranged from 2 to 4 syllables ( 152 -syllable words, 11 3 -syllable words, and 64 -syllable words) ${ }^{4}$, with a mean length of 2.31 syllables ( $\mathrm{SD}=0.64$ ). The other 32 pictures had monosyllabic names (Short). Short and long names were yoked in pairs matched for name agreement (Short: $.93(.10)$, Long: $.91(.10) ; \mathrm{t}(31)=1.17, \mathrm{p}>.2$ ), log-frequency (Short: 2.62(.46), Long: 2.54(.45); $t(31)=1.51, p>.1)$ in SUBTLEX-DE (Brysbaert et al., 2011), and initial phoneme.

Sixty-four additional pictures were used as right pictures, and were always named second in the task (word $)^{2}$ ). Two right pictures were associated with each pair of left pictures. Right pictures had high name agreement ( $\mathrm{M}=.94$, $\mathrm{SD}=$ .11); word ${ }_{2}$ had a mean length of 2.14 syllables ( $\mathrm{SD}=.71$ ), a mean frequency of $2.53(\mathrm{SD}=.64)$, and was semantically and phonologically unrelated to each word ${ }_{1}$ it was paired with. We created degraded versions of all right pictures by superimposing a mask of ten parallel white lines (about 35pt apart, and about 15pt-thickness; see Figure 1); on average the mask deleted $35 \%$ of all non-white pixels ( $\mathrm{SD}=2.3 \%$, $\min =30 \%, \max =41 \%$ ).

## Design and Procedure

Length varied within participants and items, whereas Degradation varied within items but between participants. To control for differences due to uninteresting properties of the right pictures, we constructed two lists of items, reversing pairings of left and right pictures (e.g., if in list 1 Bank was paired with Hund, and Brücke with Krone, list 2 featured Bank - Krone and Brücke - Hund); 8 random orders were generated for each list.

[^335]Participants were first familiarized with picture names. After identifying their dominant eye, they were seated about 60 cm from a 24 -inch LCD monitor. A head-mounted Eyelink 2000 recorded data from the dominant eye (pupilonly, sampling frequency: 250 Hz ). Participants were asked to avoid head movements and blinking, and named the pictures in left to right order. It was stressed they should avoid pausing between the two words. A high-quality microphone (Philips SBC ME 570) recorded participants' productions for the entire duration of the trial ( 5.5 seconds); speech onset latencies, and the duration of the pause between names (if present) were then measured offline (in Praat; Boersma \& Weenink, 2010).


Figure 1: A sample trial illustrating manipulations of word $_{1}$ length (short vs. long) and word $_{2}$ retrieval difficulty (intact vs. degraded pictures).

Presentation was controlled using Experiment Builder (Version 1.10.165). Before each trial, a fixation dot was presented where the left picture would subsequently appear. As soon as the participant fixated it, the experimenter triggered presentation of the stimuli (this was also used for drift correction). The left and right pictures were then displayed simultaneously on opposite sides of the screen, 324 pixels (or about $9^{\circ}$ of visual angle) apart. All pictures were scaled to a dimension of $290 \times 290$ pixels, with surrounding interest areas measuring $405 \times 307$ pixels (i.e., $11^{\circ}$ of visual angle horizontally, $9^{\circ}$ vertically).

The eye-tracker was calibrated twice using a ninepoint calibration grid, first after two practice trials, and then halfway through the session. The first trial after the practice session was a warmup trial, and was not analyzed. A session lasted 15-20 minutes.

## Results

Only trials in which both pictures were named fluently (i.e., with no repetitions or filled pauses, and with a silent pause no longer than 200 ms between the words) and using the expected names were analyzed (intact group: 87.99\%; degraded: $83.01 \%$ ). Following Meyer et al. (2007), we also discarded trials on which the pictures were not fixated in the order of mention (only one trial, degraded group), and trials on which the right picture was not fixated before speech onset (intact: 148 trials, or $16.43 \%$; degraded: 34 trials, or
$4.00 \%)^{5}$, as on such trials the gaze-to-speech lag would have been negative.

For the remaining trials we analyzed speech onset latencies, first-pass gaze to the left picture (the sum of all fixations to the left picture before the shift of gaze to the right picture), and the gaze-speech lag (time between the end of the first-pass gaze to the left picture and speech onset). In all analyses, we fit linear mixed-effects models using the lme4 package (D. Bates, Maechler, \& Dai, 2014) in R (R, Version 3.1.3). Fixed effects were contrast coded and centered. Random effects structure was maximal (Barr, Levy, Scheepers, \& Tily, 2013). All $p$ values are from loglikelihood ratio tests; $95 \%$ confidence intervals for model estimates are from the confint function (method="Wald"). We report the critical speech-gaze lag analyses first.

## Gaze-Speech Lag

As expected, the gaze-speech lag was both shorter when word $_{l}$ was long than when it was short $(\mathrm{B}=65 \mathrm{~ms}, \mathrm{SE}=12, \mathrm{t}=$ 5.56, $\chi 2(1)=21.46, p<.001, C I=[42,88])$ and longer for participants naming degraded than intact right pictures $(\mathrm{B}=-$ $140 \mathrm{~ms}, \mathrm{SE}=62, \mathrm{t}=-2.24, \chi 2(1)=4.63, p=.031, C I=[-262,-17]$; see Table 1, top). Crucially, however, there was no interaction between Length and Degradation ( $B=-14 \mathrm{~ms}$, $\mathrm{SE}=21, \mathrm{t}=-.69, \chi 2(1)=0.47, p=.491, C I=[-55,26])$.

## Speech Onset Latencies

After removing a further 7 ( $0.40 \%$ ) outliers (longer than 2500 ms ), we found speech onset latencies were not affected by Length, whether alone $(\mathrm{B}=-14 \mathrm{~ms}, \mathrm{SE}=27, \mathrm{t}=-.50$; $\chi 2(1)=0.21, p=.645, C I=[-67,40])$, or in interaction with Degradation ( $\mathrm{B}=5 \mathrm{~ms}, \mathrm{SE}=25, \mathrm{t}=.19 ; \chi 2(1)=0.04, p=.846$, $C I=[-44,54])$. However, speech onset latencies were longer for participants in the degraded than in the intact group ( $\mathrm{B}=$ $125 \mathrm{~ms}, \mathrm{SE}=60, \mathrm{t}=-2.09 ; \chi 2(1)=4.80, p=.028, C I=[-243,-8]$; see Table 1, middle).

Table 1: Mean gaze-speech lag, speech onset latency, and first-pass gaze to the left picture, in milliseconds (standard deviation of participants' means in brackets), as a function of word $_{l}$ Length and Degradation.

| Gaze-speech lag |  |  |
| :---: | :---: | :---: |
|  | Degraded | Intact |
| Long | 437(201) | 311(150) |
| Short | 504(210) | 362(158) |
| Speech onset latency |  |  |
|  | Degraded | Intact |
| Long | 1108(169) | 985(166) |
| Short | 1102(211) | 979(201) |
| First-pass gaze to the left picture |  |  |
|  | Degraded | Intact |
| Long | 678(77) | 699(100) |
| Short | 619(71) | 635(100) |

[^336]
## First-Pass Gaze to the Left Picture

The time spent looking at the left picture before gaze was shifted to the right was not affected by Degradation, whether alone $(\mathrm{B}=17 \mathrm{~ms}, \mathrm{SE}=27, \mathrm{t}=.63 ; \chi 2(1)=0.35, p=.555$, $C I=[-36,71]$; see Table 1, bottom) or in interaction with Length ( $\mathrm{B}=-10 \mathrm{~ms}, \mathrm{SE}=24, \mathrm{t}=-.43 ; \chi 2(1)=0.18, p=.668$, $C I=[-58,37])$. However, left pictures were fixated for longer if they had long than short names $(B=-66 \mathrm{~ms}, \mathrm{SE}=25, \mathrm{t}=-$ 2.70; $\chi 2(1)=7.21, p=.007, C I=[-115,-18])$, confirming that speakers shifted their gaze as soon as they completed phonological retrieval for word ${ }_{l}$.

## Discussion

We asked speakers to produce fluent two-word utterances and showed that the way they coordinate planning of the second word and articulation of the first word depends on both the length of the first word and the difficulty associated with retrieving the second word. The gaze-speech lag was shorter when participants were preparing to produce a long wordl and longer when word 2 was harder to retrieve.

However, we found no evidence for an interaction between word ${ }_{1}$ length and word ${ }_{2}$ retrieval difficulty. As expected, speakers in both groups took longer to articulate word $_{l}$ when it was polysyllabic ( 554 ms for the intact group, 539 ms for the degraded group) than when it was monosyllabic ( 401 ms for the intact group, 393 ms for the degraded group). This difference (about 150 ms ) is actually larger than the difference in speech onset times between the two groups of speakers (about 125 ms ). So, speakers in the degraded group could have had sufficient extra time during the production of a long word ${ }_{1}$ to compensate for the additional difficulty associated with retrieving the name of a degraded picture. In other words, if these speakers had planned proactively, they could have started speech earlier (with respect to the gaze shift) when word $_{l}$ was long than when it was short, as only in the latter case delaying speech onset would have benefitted fluency. Had they done so, gaze-speech lags would have been longer for participants in the degraded group than participants in the intact group (as we observed) but more so when participants were preparing to produce a short word $_{l}$, than when they were preparing a long word ${ }_{l}$.

This is not what we observed. Instead, participants in the degraded group appear to have used a different strategy, delaying speech onset regardless of word $_{l}$ length. Therefore, a strong version of Griffin's (2003) proposal is ruled out by our findings, as our speakers did not appear to be able to combine estimates of articulation duration with estimates of retrieval difficulty in order to precisely time articulation (of word $_{1}$ ) with respect to planning (of word $_{2}$ ).

Meyer and colleagues (2007)'s proposal, instead, is compatible with our results. First, it provides an alternative explanation of the reversed word-length effect on the gazespeech lag, which does not require a proactive planning mechanism. In addition, it may also explain the later speech onsets for speakers in the degraded group, as Meyer et al. (2007) recognized that speakers may not always start


Figure 2: First-pass gaze to the left picture, gaze-speech lag and speech onset latency (means and 95\% bootstrap CI), as a function of word $_{1}$ Length and Degradation.
articulation as soon as the articulatory code of the first syllable of a word has been retrieved.

We suggest that speakers in the degraded group buffered the first syllable of word $_{1}$ when word $_{2}$ representations failed to reach some activation threshold sufficiently early, or levels of competition within the production system (see Nozari, Dell, \& Schwartz, 2011) remained too high. ${ }^{6}$ Importantly, this type of planning mechanism can be considered reactive rather than proactive: It deals with difficulties (with word $_{2}$ retrieval) as they arise. It need not involve a mechanism that dynamically anticipates the likelihood of future difficulties, deploying different planning strategies depending on this likelihood being higher (i.e., when word ${ }_{l}$ is short) or lower.

Interestingly, based on our findings, it appears that speech is not planned proactively at the level of whole words. This appears to contrast with what we know about planning at the level of single sounds or syllables (e.g., Hickok, 2012), where there is evidence that speakers build forward models of upcoming speech movements that allow them to anticipate (and quickly correct, if necessary) what they are going to sound like (e.g., Niziolek, Nagarajan, \& Houde, 2013).

What might account for such discrepancy? We see at least two possibilities. First, research into forward models for speech has largely focused on speakers' ability to correct a

[^337]distortion in the spectral properties of the sounds they generate. We are not aware of any studies that investigated whether speakers anticipate and correct for the duration of a sound in a similar way as they do for spectral properties (e.g., pitch).

Second, in order to show the expected behavior under a proactive planning account, our speakers would have had to anticipate not just duration, but also retrieval difficulty. The latter is, unlike duration or pitch, a property of the process of planning itself, rather than an externally perceivable outcome of the planning process. As such, anticipating retrieval difficulty might involve a kind of "second-order" forward model. Speakers might be able to learn such forward models, but perhaps only with extensive training.

In conclusion, the reversed word-length effect cannot be interpreted as evidence that the flexibility of speakers' planning reflects the workings of a proactive mechanism. However, speakers are able to reactively compensate for retrieval difficulty, delaying speech onset when the need arises.

## Acknowledgments

We thank Raphael Morschett for assistance with preparation of the word lists and data collection. This research was supported by research funds assigned to the Chair of Psycholinguistics by the University of Saarland. Chiara Gambi is supported by a Leverhulme Trust Research Project Grant to Hugh Rabagliati and Martin Pickering (RPG-2014253).

## References

Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. (2013). Random effects structure for confirmatory hypothesis
testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bates, D., Maechler, M., \& Dai, B. (2014). Lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7). Retrieved from http://lme4.r-forge.r-project.org/
Bates, E., D’Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., . . . Pléh, C. (2003). Timed picture naming in seven languages. Psychonomic Bulletin \& Review, 10(2), 344-380.
Boersma, P., \& Weenink, D. (2010). Praat: Doing phonetics by computer (Version 4.6.22) [Computer Software]. Retrieved from http://www.praat.org/
Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., \& Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. Experimental Psychology, 58(5), 412-424. doi:10.1027/16183169/a000123
Damian, M. F., Bowers, J. S., Stadthagen-Gonzalez, H., \& Spalek, K. (2010). Does word length affect speech onset latencies when producing single words?. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(4), 892.
Experiment-Builder. (Version 1.10.165) [Computer Software]. Ottawa, Ontario, Canada: SR Research. Retrieved from www.pstnet.com
Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. Cognition, 82(1), B1-B14.
Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. Psychonomic Bulletin \& Review, 10(3), 603609.

Griffin, Z. M., \& Bock, K. (2000). What the eyes say about speaking. Psychological Science, 11(4), 274-279.
Hickok, G. (2012). Computational neuroanatomy of speech production. Nature Reviews Neuroscience, 13(2), 135145.

Jacobs, C. L., \& Dell, G. S. (2014). 'hotdog', not 'hot' 'dog': the phonological planning of compound words. Language, Cognition and Neuroscience, 29(4), 512-523.
Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. Journal of Memory and Language, 66(1), 143-162.
Konopka, A. E., \& Meyer, A. S. (2014). Priming sentence planning. Cognitive Psychology, 73, 1-40.
Levelt, W. J. M. (1989). Speaking: From intention to articulation. Cambridge, MA: MIT Press.
Levelt, W. J. M., Roelofs, A., \& Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22(1), 1-75.
Martin, R. C., Crowther, J. E., Knight, M., Tamborello, F. P., \& Yang, C.-L. (2010). Planning in sentence production: Evidence for the phrase as a default planning scope. Cognition, 116(2), 177-192.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture-word interference experiments. Journal of Memory and Language, 35(4), 477-496.
Meyer, A. S., Belke, E., Häcker, C., \& Mortensen, L. (2007). Use of word length information in utterance planning. Journal of Memory and Language, 57(2), 210231.

Meyer, A. S., Roelofs, A., \& Levelt, W. J. M. (2003). Word length effects in object naming: The role of a response criterion. Journal of Memory and Language, 48(1), 131147.

Meyer, A. S., \& Van der Meulen, F. F. (2000). Phonological priming effects on speech onset latencies and viewing times in object naming. Psychonomic Bulletin \& Review, 7(2), 314-319.
Niziolek, C. A., Nagarajan, S. S., \& Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. Journal of Neuroscience, 33(41), 16110-16116.
Nozari, N., Dell, G. S., \& Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflictbased account of monitoring in speech production. Cognitive Psychology, 63(1), 1-33.
R. (Version 3.1.3) [Computer Software]. Vienna, Austria: R Development Core Team. Retrieved from http://www.Rproject.org
Roelofs, A., \& Piai, V. (2011). Attention demands of spoken word planning: A review. Frontiers in Psychology, 2, 10.3389/fpsyg.2011.00307.
Sandra, D. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. Quarterly Journal of Experimental Psychology, 42(3), 529-567.
Smith, M., \& Wheeldon, L. (1999). High level processing scope in spoken sentence production. Cognition, 73(3), 205-246.
Smith, M., \& Wheeldon, L. (2004). Horizontal information flow in spoken language production. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(3), 675-686.
Van de Velde, M., Meyer, A. S., \& Konopka, A. E. (2014). Message formulation and structural assembly: Describing "easy" and "hard" events with preferred and dispreferred syntactic structures. Journal of Memory and Language, 71(1), 124-144.
Wolpert, D. M., \& Flanagan, J. R. (2001). Motor prediction. Current biology, 11(18), R729-R732.

# Automated Generation of Cognitive Ontology via Web Text-Mining 

Richard Gao (rigao@ucsd.edu) Thomas Donoghue (tdonoghue@ucsd.edu) Bradley Voytek (bvoytek@ucsd.edu)<br>Department of Cognitive Science, UC San Diego<br>9500 Gilman Drive, La Jolla, CA 92093-0515, USA


#### Abstract

A key goal of cognitive science is to understand and map the relationship between cognitive processes. Previous works have manually curated cognitive terms and relations, effectively creating an ontology, but do they reflect how cognitive scientists study cognition in practice? In addition, cognitive science should provide theories that inform experimentalists in neuroscience studying implementations of cognition in the brain. But do neuroscientists and cognitive scientists study the same things? We set out to answer these questions in a data-driven way by text-mining and automated clustering to build a cognitive ontology from existing literature. We find automatically generated relationships to be missing in existing ontologies, and that cognitive science does not always inform neuroscience. Thus, our work serves as an efficient hypothesis-generating mechanism, inferring relationships between cognitive processes that can be manually refined by experts. Furthermore, our results highlight the gap between theories of cognition and the study of their implementation.


Keywords: ontology; cognitive neuroscience; text-mining neuroinformatics, meta-analysis

## Introduction

## Ontology: Key Challenge in Cognitive Science

One of the fundamental goals of cognitive science is to study the set of processes that combine to give rise to "cognition". These processes can be thought of as abstractions to common, overlapping sets of behaviors. Constrained by methodological behaviorism, we can only observe behavior and label underlying cognitive processes after the fact. As such, they do not have direct grounding in the physical world, and thus need to be defined by the relational structure that link each other - an ontology. For example, attention and working memory are processes with different labels but are nonetheless woven together through behavior: one cannot allocate "working memory" without "paying attention". Thus, as we collect more observations to fill up the space of cognitive processes, we must be attentive in organizing what we know. This is the problem of mapping the ontological structure of cognitive processes, and has received extensive consideration previously (see, e.g., Poldrack \& Yarkoni, 2016).

## Neuroscience: Studying the Substrate of Cognition?

If cognitive processes are viewed as algorithms performing a set of computations, there must then exist a
physical substrate that is performing the computations. In the case of computer algorithms, the substrate consists of transistor elements. The brain, on the other hand, is a large part of the computational substrate of human cognition (along with our body and environment). Cognitive neuroscience, with the aid of neuroimaging, has revealed much about our cognitive processes, such as timing between consecutive steps in a cascade of processes. However, neuroimaging studies are almost always conducted in the laboratory, with specific physical and task constraints. Hence, one cannot be certain that cognitive neuroscience actually measures, or even attempts to measure, the full array of cognitive processes at play. For example, the consolidation of long-term memory is quite difficult to measure within the span of a single experiment, while visual perception can be easily studied. Conversely, observations in neuroscience may provide constraints for cognitive theories, but only if there is an overlap of interest in the same processes. Thus, we should understand the degree to which we are over- and under sampling cognitive processes while measuring the brain. This is the problem of adequate sampling of cognitive processes in cognitive neuroscience.

## Frameworks for Ontology-Mapping

The problem of ontology has been addressed previously. Notably, Poldrack and colleagues (2011) started a monumental effort in charting the ontological space of cognitive processes, as well as their related experimental tasks and disease correlates, aptly named the Cognitive Atlas. These authors hand-crafted hundreds of terms and their relations with each other, and invited researchers to contribute to documenting new relations - like a Wikipedia for cognitive science. While quality-controlled, curating these processes by hand relies on massive participation of the community, and must match the speed at which new evidence linking old processes is published. A complementary approach to human-generated relations is to let experts judge the validity of machine-generated relations, which can cover much more ground very quickly, saving human time and resources.

## Automated Generation of Cognitive Ontology

Here, we present an automated text-mining algorithm that scans through relevant literature databases and builds an
ontology through co-occurrences of cognitive terms mined from the Cognitive Atlas. In particular, we apply the mining algorithm to PubMed, as well as the proceedings to the Annual Meeting of the Cognitive Science Society, in an attempt to automatically generate an ontological structure supplementing the Cognitive Atlas. Furthermore, we search PubMed for cognitive terms in conjunction with neuroimaging terms to establish the cognitive ontology viewed through neuroscience. We note here that previous neuroinformatic works have tackled related challenges. In particular, Yarkoni et al. (2011) created Neurosynth as a meta-analysis of fMRI studies. Its strength lies in providing voxel-level identification of the neural support of cognition, though it necessarily ignores the massive body of electrophysiological research (EEG, MEG, etc.) in favor of certainty in spatial location. In addition, Voytek \& Voytek (2012) built BrainSCANR, an automated PubMed textmining application for similar purposes. However, that work focused primarily on aspects of neuroscience, with inclusion of brain regions and neurochemicals as keywords, while having a limited set of cognitive terms.

In the following sections, we describe the text-mining procedure, as well as an analysis of the word-relations constructed from the automatically generated databases. We present similarities of term-frequency in 4 databases: CogSci (CS), PubMed Cognitive (PMCog), and PubMed Neuro (PMNeu \& PMNeuMeth). We further explore latent structures within each database via hierarchical clustering to automatically generate an ontology of cognitive processes.

## Methods

All code available online at:
https://github.com/voytekresearch/IdentityCrisis

## Data Collection

Term Collection 803 cognitive terms were scraped from the "Concepts" page from the Cognitive Atlas. These were used as the main search terms below, and will thus be referred to as "cognitive terms."

CogSci Abstracts This database is constructed from the title and abstracts of the Presentations, Tutorials, Symposia, and Papers of the Annual Meeting of the Cognitive Science Society from 2010 to 2016. We look for the cognitive terms in each document, constructing a term-document matrix. We then built a co-occurrence matrix by noting all pair-wise co-occurrences of cognitive terms in each document. Data from all 7 years are aggregated. Terms with 50 or more occurrences are included in the clustering analysis (86).

PubMed Cognitive This database is constructed by searching in PubMed for pairs of cognitive terms in quotations, such as "attention"AND"working memory", plus a base phrase: ('AND("cognitive"OR"cognition")'), to ensure searches are constrained to hits relevant to cognition. Counts are recorded as the number of articles that include the search terms in the title or abstracts. Prior to pairs
search, we built a term-frequency vector measuring the occurrence of all 803 cognitive terms. Only individual terms with 500 or more hits ( 217 terms) were included in the pairs search to decrease search time. The number of hits for each pairs of terms ( $\mathrm{i} \& \mathrm{j}$ ) are recorded in the co-occurrence matrix as element $\mathrm{a}_{\mathrm{ij}}$. Search code was built upon the PubMed EUtils Tool API.

PubMed Neuro Method $\mathcal{\&}$ Neuro These databases are created as the one above, but in conjunction with a base phrase reflecting neuroimaging techniques,
('AND('+ '("fmri"OR"neuroimaging")OR'+
'("pet"OR"positron emission tomography")OR'+
'("eeg"OR"electroencephalography")OR'+
'("meg"OR"magnetoencephalography")OR'+
'("ecog"OR"electrocorticography")OR'+
'("lfp"OR"local field potential")OR'+
'("erp"OR"event related potential")OR'+
'("single unit"OR"single-unit"OR"single neuron")OR'+
'("calcium imaging")")').
138 terms remained after thresholding at 500 hits.
As suggested by reviewers, we further included a "general neuroscience" database that was not exclusively techniques, using ("neural"OR"neuroscience") as base phrase.

## Data Analysis

Term-Frequency Term-frequency for each cognitive term were calculated as a fraction by dividing the number of hits a term generated by the total results returned for the base phrase alone (for PubMed databases) or the total number of abstracts (for CogSci database). To visualize differences in term usage, we take the term-frequency difference between pairs of databases and find the terms with the highest absolute difference.

Hierarchical Clustering We use the SciPy hierarchical clustering module (scipy.cluster.hierarchy) to cluster terms based on their normalized co-occurrence matrix, where each row is divided by the diagonal of that row (co-occurrence with self). Dendrograms are generated and leaves are cut (colored) to generate $\sim N / 4$ clusters, where $N$ is the total number of terms in tree.

## Results

In summary, we find that:

1) there are discrepancies between prevalent terms discovered in the CogSci database and the PubMed Neuro database, with the former leaning towards more theoretical constructs, and the latter, experimentally tangible;
2) hierarchical clustering reveals reasonable yet novel groupings of cognitive terms that are undocumented in the Cognitive Atlas.

## Term-Frequency Across Databases

First, we address the question: do cognitive scientists and neuroscientists study the same underlying processes? Table

1 presents the top 20 most frequent cognitive terms in each database.

Table 1: Proportion of term occurrence for the top 20 terms in each database. Green boxes denote terms unique to that database, while red boxes denote terms unique to Neuro.

| CogSci |  | PM Cog |  | PM Neuro |  | PM Neuro Meth |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| learning | 0.257 | memory | 0.201 | activation | 0.114 | activation | 0.11 |
| search | 0.228 | risk | 0.131 | learning | 0.071 | detection | 0.068 |
| action | 0.218 | attention | 0.108 | memory | 0.065 | memory | 0.06 |
| language | 0.185 | learning | 0.104 | loss | 0.054 | risk | 0.059 |
| logic | 0.151 | association | 0.081 | action | 0.053 | attention | 0.052 |
| knowledge | 0.128 | anxiety | 0.066 | inhibition | 0.052 | monitoring | 0.048 |
| concept | 0.126 | stress | 0.059 | attention | 0.044 | sleep | 0.043 |
| memory | 0.122 | loss | 0.054 | risk | 0.041 | association | 0.041 |
| context | 0.116 | activation | 0.053 | perception | 0.038 | movement | 0.036 |
| decision | 0.098 | knowledge | 0.052 | detection | 0.037 | loss | 0.034 |
| attention | 0.091 | language | 0.051 | knowledge | 0.036 | action | 0.031 |
| reasoning | 0.081 | context | 0.046 | association | 0.034 | learning | 0.03 |
| judgment | 0.081 | working | 0.044 | context | 0.034 | perception | 0.029 |
| focus | 0.079 | focus | 0.04 | stress | 0.033 | inhibition | 0.029 |
| lying <br> inference | 0.074 | perception | 0.04 | movement | 0.032 | knowledge | 0.028 |
| perception | 0.072 | recognition | 0.039 | mood | 0.038 | recognition | 0.032 | focus $\quad 0.027$.

First, we note the general trend that CogSci proceedings are much more likely to contain one of several popular terms, with 3 terms appearing in more than $20 \%$ of the abstracts and 9 terms in more than $10 \%$. In contrast, only one word in the PMNeuro database is contained in more than $10 \%$ of the abstracts ("activation"), which may be artificially inflated due to usages of the word in contexts not describing cognitive activation (e.g., fMRI activation). This suggests that the terms we deem to describe "cognitive processes" do indeed see more usage in the cognitive science community.

On an individual term level, several striking patterns prevail. First, "learning" appears in about $25 \%$ of CogSci abstracts, but only $10 \%$ in PMCog, $7 \%$ in PMNeu, and $3 \%$ in PMNeuMeth. This reveals that the concept of "learning" is a rather popular theoretical construct, while being harder to study empirically via neuroimaging. Additionally, "search", "language", and "logic" all appear in more than $15 \%$ of CogSci abstracts, but do not crack the $5 \%$ mark in PMNeuro, further suggesting the difficulty or reluctance in studying these theoretically important but empirically illdefined concepts in a neuroscientific context.

On the other hand, "attention", "perception", and "movement" occur in all 4 databases with relatively low but similar proportions. This is unsurprising, as physical processes are much more easily studied in neuroscientific experiments.


Figure 1: Term frequency results for each database. Note that $y$-axis is in log scale.

We saw from Table 1 that term usage distribution for the most frequent terms are not the same across the 4 databases. Figure 1 plots these distributions for the top 250 terms used. We see that CogSci proceedings are not very diverse in terms of their topics of investigation, as the more common terms are much more represented in the abstracts. This may be due to the small number of CogSci abstracts available, compared to around 100 times more results returned from PubMed searches. However, PMCogs is less drastic but follows a similar trend, suggesting that cognitive science as a whole refers to these cognitive terms much more frequently than neuroscience.


Figure 2. Terms used most differently between Cognitive Science (CogSci, left, PMCogs, right) and Neuroscience.

Finally, we find terms with the biggest usage proportion difference between cognitive science and neuroscience. These results recapitulate the top terms we see in Table 1, where an overwhelming proportion of high-level, conceptual terms are overrepresented in the 2 cognitive datasets. Overall, these analyses demonstrates that, while cognitive terms are adopted more frequently in CogSci
abstracts than the general body of literature in PubMed, many of the processes that are focused on in the CogSci community has not seen as much empirical investigation in the neuroscience community.

## Hierarchical Clustering

Having shown a difference in the frequency of cognitive term usage between cognitive science and neuroscience, we turn to the term co-occurrence data to generate ontologies. Here, we can address the question of, in addition to being used with differing frequencies, whether the terms are used in different ways in relation to each other, which suggests a difference in term "meaning". Figure 3 shows dendrograms generated from the CogSci database and PMNeuMeth database. The length of the colored lines (starting from the right) when they merge reflect the similiarity of the the merged terms: the shorter the lines, the more similar they are. As such, pairs of terms like "acuity" and "visual acuity", or "memory" and "working memory" are merged very early on due to the overlap in words, which is a limitation of the text-mining method employed.

Barring these overlapping terms, very reasonable clusters emerge at the mid-level. For example, at the lower end of the CogSci tree exists a language group (red \& green) and a learning group (teal). Interestingly, "learning" and "generalization" are very closely tied. Moving up a few clusters, a reasoning cluster emerges (black), including "reasoning", "inference", "induction", and "rule". Similar clusters existing in the PMNeu tree, where the top clusters reflect all forms of perception, then attention, transitioning to speech processing, and finally to language understanding. "Theory of mind" is grouped with "empathy" and "social cognition", while "discrimination" is grouped with "categorization" and "judgement".

Due to the difference in term prevalence between these two databases, some clusters exclusive to one or the other appear. "Logic", "analogy", and "schema" exist as one cluster in the CogSci database, while "anxiety", "fear" and "extinction" emerge as a cluster in the PMNeu database. These clusters clearly reflect the theoretical vs. experimental nature of works published in these two fields. Furthermore, "learning" in CogSci, as mentioned above, talks about a high-level, mental process (tied to "category learning"), while it is linked to "skill", "navigation", and "expertise" in neuroscience. Overall, these examples qualitatively demonstrate that an automated mining and clustering process can tease out: 1) similarity of cognitive terms by grouping them within clusters, and 2) contextualized meaning of terms by grouping them into different clusters specific to cognitive science or neuroscience.
Finally, in keeping with our original goal, we examine whether clusters discovered with our automated process can be used to supplement information in the Cognitive Atlas. Figure 4 demonstrates one example concept: "learning". We observe that the only populated relationship is "are kinds of", in which more specific types of learning are listed. However, the ontological mapping does not capture
categorically similar terms described above, such as "generalization" or "categorization". Other examples of missed relationships are more nuanced. For example, under "addiction", the Cognitive Atlas currently includes "reward processing" as a part of addiction (also discovered in our clustering). However, it does not mention "anticipation" and "impulsivity", both of which are key factors in the continuation of addictive behavior. Hence, we conclude that automated clustering of related concepts can greatly aid in the curation of an extensive cognitive ontology.


Figure 4: "learning" and "addiction" as curated by Cognitive Atlas, supplemented by clusters generated automatically (from Fig. 3).

## Discussion

## Summary

In this study, we created a text-mining and clustering pipeline that aims to automate the process of aggregating information from existing literature to create an ontological structure for cognitive processes. We searched for cognitive keywords curated by the Cognitive Atlas, and analyzed databases created by scraping the proceedings to the Annual Cognitive Science meeting, as well as PubMed articles, containing these keywords. We find a prevalent usage of these terms in all the databases, particularly so in the CogSci abstracts. The frequency of term usages differ between CogSci abstracts and PubMed neuroscience articles, likely reflecting the methodological preferences in each field. Hierarchical clustering on pairwise term co-occurrence data group terms relating to each other, demonstrating practicality in serving as a hypothesis-generating procedure to further populate manually-maintained ontologies, such as the Cognitive Atlas.


Figure 3: hierarchical clustering results for CogSci and PubMed Neuro Method database.

## Implications for Cognitive Science

The current work presents two main contributions. First, the tool itself is completely open-sourced and depends on publicly available databases. Domain-specific researchers can utilize this tool to find common associations to their process of choice, such as addiction. This will be especially useful for beginner researchers, like undergraduate and early graduate students, to quickly situate their topic in the broader context. Furthermore, on a larger scale, this tool can serve as a complementary approach to hand-curated ontologies, saving experts time from manually filling in blanks. One point worth noting is that our work does not attempt to build the ontological structure of cognitive processes as it exists in our minds, similar to ideas suggested by Newell's universal theory of cognition (UTC). Rather, it is a meta-analysis of how cognitive scientists decide to investigate the latent structure of our cognitive processes through their work, with no claims on whether or how this ontological structure actually exists.

Second, the theoretical contribution of this work is that it points to the discrepancy between how cognitive science and neuroscience study cognition. One simple explanation is that neuroscience only partially overlaps with cognitive science, as genetic, molecular, and cellular investigations often do not relate to cognitive phenomena. This is clearly true, however, given that the PubMed Neuro Method database is built specifically with keywords relating to animal neuroimaging techniques, this is unlikely to be the explanation here. Furthermore, the gap similarly exists between PubMed Cognitive and PubMed Neuro databases, so it is not simply a difference in the source of data. Thus, this gap raises the alarming possibility, as one reviewer pointed out, that theories in cognitive science are not testable in the realm of neuroscience, and/or that neuroscience is simply not interested in or ready for the grand theories of cognition.

## Limitations \& Future Work

While the algorithm returns reasonable and novel results, a few methodological and data-collection limitations must be raised. First, in building the databases, CogSci abstract were collected only up to the annual meeting in 2010, as further archives were unavailable. In contrast, PubMed searches return all hits dating back 30 or more years. As such, it is possible that trends observed in the term-frequency analysis may be due to a temporary peak in interest in certain areas of research, such as "learning". This can be easily ameliorated, however, by rebuilding the PubMed databases while constraining the included search years. In fact, we can analyze different decades (or other periods of time) to see how ontological structure develops over time.

Second, due to the scraping method applied, terms with overlapping words, such as "memory" and "procedural memory" will co-occur with much higher frequency, possibly leading to inflated inferred relationships. Since
terms with overlapping words are very likely to have a superset-subset relationship, the over-interpretation of relationship is unlikely to create false positives. However, the artificial increase in co-occurrence may lead clustering to exclude related but now suppressed terms, leading to false negatives. This may be circumvented by making queries for specific terms, i.e., accessing specific rows in the co-occurrence matrix, and ranking related words in their rate of co-occurrence. Hierarchical clustering is simply one method to visualize the co-occurrence data, and many others may be applied on the same dataset to further tease out latent structures, such as Multi-Dimensional Scaling.

Lastly, the co-occurrence matrix is built on the assumption of a bag-of-words model, i.e., word-order and semantic relations don't matter, simply their shared presence in a document. This may lead to spurious linkages, if a document contained a phrase like "attention is not a type of memory." This is likely to be rare, and ultimately, still useful knowledge, as it implies that at some point these terms were wrongfully linked. This last point, however, raises a larger, philosophical question: can automated text mining of existing literature get at the ontology of cognitive science, and if so, is that the same ontology that exists in our minds? We may never know the answer to the latter, but the former is certainly an issue worth investigating. Regardless of whether or not the structure can be recovered from the model presented here, the knowledge structure clearly exists within the minds of practicing cognitive scientists. As such, we may leverage other sources of information, such as citation links, to trace out the ontology, which ultimately just represents a consolidation of knowledge across the broad, interdisciplinary study of cognition.

## Acknowledgments

We would like to thank Arturs Semenuks and Torben Noto for creative inspiration, reviewers for their helpful comments and suggestions, as well as the Alfred Sloan Foundation and the Natural Sciences and Engineering Research Council of Canada for funding.

## References

Poldrack, R. A. \& Yarkoni, T. (2016) From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. Annul Rev Psychol 67, 587-612 (2016).
Poldrack, R. A. et al. (2011) The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. Front. Neuroinform.
Voytek, J. B. \& Voytek, B. (2012) Automated cognome construction and semi-automated hypothesis generation. Journal of Neuroscience Methods 208, 92-100.
Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. \& Wager, T. D. (2011) Large-scale automated synthesis of human functional neuroimaging data. Nature Methods 8, 665-670.

# "If It Matters, I Can Explain It": Social Desirability of Knowledge Increases the Illusion of Explanatory Depth 

Christian Gaviria (cdgaviriam@unal.edu.co)<br>Javier Corredor (jacorredora@unal.edu.co)<br>Zadkiel Zuluaga-Rendón (zzuluagar@unal.edu.co)<br>Department of Psychology, Universidad Nacional de Colombia<br>Cr. 30 \# 45-03, Ed. 202 Of. 219, Bogotá, 111321. Colombia


#### Abstract

This paper explores whether social desirability affects the illusion of explanatory depth (IEOD) by comparing the magnitude of this illusion in topics with different levels of social desirability within several domains. This question was chosen because prior literature shows that social expectations about how much a person should know about a certain topic affect the magnitude of the IOED. Previous research shows also that social desirability has an effect on a similar illusion related to argumentation, and that the IOED is affected by the way a person thinks knowledge is distributed in his or her social group. In order to do so, 184 participants were assigned randomly to three knowledge domains (history, economics, and devices) and in each domain they rated their understanding of a high-desirability and a low-desirability topic following a standard IOED procedure. Results show that social desirability has an effect on the IOED magnitude and that overestimation of understanding varies among domains. Particularly, participants tend to overestimate their understanding of high desirability topics only. This effect was stronger in the historical domain.


Keywords: Illusion of explanatory depth; social desirability of knowledge; feeling of knowing; metacognition; motivated cognition.

## Introduction

There is extensive evidence that people are often overconfident regarding the quality and accuracy of their knowledge (Moore \& Healy, 2008; Zell \& Krizan, 2014). This metacognitive bias has been consistently found in the context of tasks as diverse as recalling memorized information (e.g., Koriat, Lichtenstein, \& Fischhoff, 1980), solving general-knowledge questions (e.g., Atir, Rosenzweig \& Dunning, 2015), evaluating text comprehension (e.g., Jaeger \& Wiley, 2015), and making consumer decisions (e.g., Alba \& Hutchinson, 2000).

The discrepancy between what people think they know and what they really know seems to be more conspicuous in certain kinds of knowledge. Thus, Rozenblit and Keil (2002) found that college students are prone to overestimate their ability to explain the mechanisms of devices or natural processes, but not their understanding of facts, narratives or procedures. This "Illusion of Explanatory Depth" (IOED), have been robustly documented in recent years, both in experts and non-experts (Lawson, 2006; Fisher \& Keil, 2015) as in young children (Mills \& Keil, 2004). The IOED
has also been found in both physical (Lawson, 2006; Fernbach, Sloman, St. Louis \& Shube, 2013; Fisher \& Keil, 2015) and social mechanisms (Alter, Oppenheimer \& Zemla, 2010; Fernbach, Rogers, Fox \& Sloman, 2013). These previous works have focused on demonstrating how pervasive is the IOED in different domains of knowledge and on searching for conceptual properties of objects/topics associated to different degrees of this phenomenon. However, when people evaluate how much they know about a particular topic, it is still possible they do not just keep in mind what they do know, but also the representation of what they should know. This representation can be inferred explicitly or implicitly from contextual and motivational cues, such as the perceived social desirability of knowledge. In the following sections, we review some empirical results from research both on IOED and metamemory that are consistent with this interpretation.

## IOED and Social Desirability of Knowledge

Fisher and Keil (2015) investigate whether expertise in a given domain of knowledge is associated to a more accurate self-evaluation of the understanding of objects and processes. In order to do so, they distinguish between passive and formal expertise; passive expertise refers to knowledge coming from the "exposure through life experience and the position one occupies in a society or culture" (p.1251; e.g., specific knowledge or skills culturally associated to gender or age), whereas formal expertise is the final outcome of systematic, continued and deliberate training in a specific domain with definite milestones (e.g., academic degrees). Fisher and Keil found that participants with less formal expertise (e.g., with no college major) overestimated more their understanding of topics related to their passive expertise than they did regarding their understanding of other topics. This difference was not replicated in the group with more formal expertise. In this group, participants overestimated their understanding of topics related to their area of formal expertise more than of other topics. These results suggest that people tend to overestimate their ability to explain topics related to their area of expertise. In the case of people with formal education, it happens with topics related to their formal expertise; in the case of people with no formal training, it happens with topics related to their passive one. In both cases, a critical factor affecting the IOED magnitude
seems to be the participants' beliefs about how much they should know about certain topics because of his or her type of expertise, regardless of how that knowledge was acquired.

In the same vein, Fisher and Keil (2014) asked their participants to write their arguments supporting their position about controversial topics. Before and after this task, they were asked to evaluate how well they could support their own positions. Additionally, participants judged how important the topics were for them. Using an experimental paradigm quite similar to that of the IOED research, they found evidence of an "Illusion of Argumentative Justification" (IAJ): participants' ratings of their ability to support their point of view decreased after writing their arguments. Importantly, caring for each topic was positively associated with both previous and posterior evaluations of the ability to rationally justify their own position, and this pattern was not replicated when arguments were rated by a different group of participants. In short, IAJ seems to be stronger in topics that matter to participants.

In apparent contradiction with these more recent findings, Rozenblit and Keil (2002, study 11) reported that perceived social desirability of explanations was not associated with the magnitude of overconfidence in any domain of knowledge (facts, procedures, narratives, or explanations). They even claim that "if anything, high desirability may cause people to more carefully assess their self-knowledge in a domain and, therefore, be more accurate." (p. 547). However, at least one important difference between Rozenblit and Keil (2002) and Fisher and Keil (2014) studies can account for this discrepancy: whereas Rozenblit and Keil compare differences in overconfidence between kinds of knowledge (e.g., facts, procedures, narratives, and explanations), Fisher and Keil contrast the IAJ magnitude between topics with different degrees of personal significance, within a same kind of knowledge (e.g., arguments). From a methodological point of view, comparing between kinds of knowledge could be not the optimal strategy to establish whether social desirability and IOED are related, as far as the latter is a phenomenon essentially linked to explanations. In this context, comparing the IOED magnitude between more or less socially desirable topics or explanatory domains might be more informative than contrasting the effect of social desirability between explanations and other kinds of knowledge (e.g., arguments). Exploring this alternative is the main purpose of this study.

The influence of social cues in the process of knowledge self-assessment is not an exclusive finding of the IOED paradigm. In the next section, we review some evidence from metamemory research suggesting the inferential nature of such process, and identifying a number of contextual factors affecting perceptions about how likely some specific content is to be recalled from memory.

## Social Desirability and the Feeling of Knowing

The research on metacognitive judgments in memory tasks has inquired about the sources of information people use to infer whether a particular content can be learned or recalled. Specifically, the "feeling of knowing" (FOK) has been extensively investigated. In general, this feeling is experienced by an individual when he or she thinks to have certain items stored in memory and the ability to recall or recognize them in the future, even when they cannot do it at the present (Hart, 1965).
In order to elucidate the metacognitive mechanism underlying to FOK, researchers have explored factors associated to the accuracy of these judgments. Consistently, it has been found that FOK is not the output of a unique mechanism. Instead, diverse factors can affect the metacognitive processes driving to it, depending on both recovery timing and task restrictions (for review, see Thomas, Lee \& Hughes, 2016). For example, whereas perceived familiarity with items can increase the FOK before the recall phase, related information accessibility has a major role when recalling is not successful (Koriat \& Levy-Sadot, 2001). These findings support the hypothesis that the FOK mechanism is not an encapsulated directaccess module (Hart, 1965), but the result from multiple inferential processes, working with information derived from cues previously or simultaneously generated along with the recall process (Thomas et al., 2016).
Supporting this hypothesis, Costermans, Lories and Ansay, (1992, exp. 2) explore several cues related to the magnitude of FOK judgments. In particular, they find that confidence is a better predictor of answers accuracy than FOK. Interestingly, both question familiarity and the estimated amount of people knowing the correct answer were positively associated with the FOK magnitude. In the same vein, De Carvalho and Yuzawa (2001) report that the FOK magnitude in college students with low levels of metacognitive ability increases when they are provided with information about fictitious students having high performances in a similar task and, correspondingly, diminishes it when these fictitious performances were presented as low. These results are compatible with the Self Consistency Model of Subjective Confidence (SCM; Koriat, 2012), which postulates that correlation between confidence and accuracy in FOK judgments is positive when people agree on the correct answer. In contrast, the confidenceaccuracy correlation is negative when there is a similar level of consensus about an answer that is ultimately wrong. Once again, these results confirm that FOK is not directly computed, but inferred from internal cues such as familiarity, processing fluency, and the perceived distribution of knowledge in the population.
Although the effect of social desirability of knowing the correct answer has not been directly tested on the FOK paradigm, there are reasons to think it might well be an informative cue about how likely a content in memory is to be recalled (Gruneberg, Monks \& Sykes, 1977). In a related area of research, Soderstrom and McCabe (2011) found that
college students judge that they will be more able to learn items whose successful recall is better rewarded in the experimental setting. In the same vein, the predicted grades of college students in a course exam are biased by their desired level of performance (Serra \& Demarree, 2016).

On the other hand, though experimental paradigms of FOK differ from IOED in that they had used pieces of nonexplanatory knowledge (e.g., historical events, dates, names, places, etc.), this fact does not rule out the possibility that both IOED and FOK engage analogous or common metacognitive mechanisms. If that is the case, factors related to the FOK magnitude might be responsible of differences in the IOED magnitude between topics and domains. Examining the influence of inferential cues such as social desirability on the IOED will allow us to identify the conditions that lead to the overestimation of the explanatory knowledge about a certain topic. In turn, this information would be useful in creating cognitive strategies to help people to re-calibrate their understanding and monitoring their own learning processes of specific contents in more accurate and effective ways (Dunlosky \& Thiede, 2013). In this context, the aim of this study is to determine whether social desirability of knowledge is used by participants as an informative cue when they are assessing their understanding of mechanisms in different domains of explanatory knowledge.

## Method

The experiment has two goals: First, we intend to establish whether social desirability of knowledge about a specific topic predicts the IOED magnitude. Second, we want to know whether the relationship between IOED magnitude and social desirability differs among explanatory domains (e.g., historical, economic, and devices).

## Participants

In this study participated one hundred and eighty-four students from a large research university (88 women) attending different undergraduate programs, with ages ranging from 18 to 42 years $(M=20.7, S D=2.04)$. Most of them received academic extra-credit for their participation in this study.

## Design

A mixed experimental design, $3 \times 2 \times 2$, was used, with explanatory domain (historical, economic and devices) as the between-subjects factor, and social desirability of topics (high and low), and pre-post measures as within-subjects factors. The dependent variable was the IOED magnitude, measured as the rating of understanding of each topic.

## Materials and Procedure

The same procedure used by Rozenblit and Keil (2002, study 11) was used to select high and low desirability topics for each domain. In a preliminary study, one hundred and ninety-four participants (117 women) evaluated the
perceived social desirability of knowledge about 21 topics (seven in each domain). Specifically, they reported how embarrassed they would feel if they did not have a good understanding on each topic in a 7-point scale, ranging from 1 ("If someone asked me to explain this topic and I had a poor understanding of that item, I would not feel embarrassed at all") to 7 ("If someone asked me to explain this topic and I had a poor understanding of that item, I would feel very embarrassed"). Six items -the two topics showing greater difference in the desirability scale within each domain- were selected for the main study (see Table $1)$.

Table 1: Means of social desirability of each topic in the preliminary study

| Domain | Topic | M | SE | 95\% CI |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | Lower bound | Upper bound |
| Historical | Long duration of the Colombian armed conflict | $5.02^{\text {a }}$ | . 135 | 4.76 | 5.29 |
|  | Creation of the FARC-EP guerilla | $4.58{ }^{\text {b }}$ | . 132 | 4.32 | 4.84 |
| Economic | Why inflation rises in Colombia | $4.37{ }^{\text {b }}$ | . 133 | 4.11 | 4.63 |
|  | How the stock market works | $3.65{ }^{\text {c }}$ | . 128 | 3.39 | 3.90 |
| Devices | How a fishing rod works | $3.40^{\text {c }}$ | . 151 | 3.10 | 3.70 |
|  | How a jet engine works | $2.43{ }^{\text {d }}$ | . 130 | 2.18 | 2.69 |

Note: $M=$ mean; $S E=$ standard error; CI = Confidence interval; Means marked with different letters differed significantly from each other $(p<.01)$

In an isolated and noise-free room, participants did the experimental task at their own pace in individual cubicles. For task presentation and response recording, the Qualtrics web-based survey software was used. During the session, it was verified that participants did not check other websites. After registering their demographic information, participants completed an instructional manipulation check (Oppenheimer, Meyvis \& Davidenko, 2009) to ensure the careful reading of the instructions. The following phases were aligned with the IOED experimental paradigm: initially, participants evaluated their knowledge about twelve topics (six of which were not part of the design), using a 7 -point scale, with 1 meaning "vague or poor understanding" and 7 "detailed and fine-grained
understanding". The instructions to use this rating scale were adapted from Rozenblit and Keil (2002) and Fernbach, Rogers, et al. (2013). Next, participants were randomly assigned to one of the three domains (historical, economic, and devices), and they were asked to explain in a step-bystep way the causal mechanism of one of the object/topic in this domain (explanation phase). In particular, they were provided with the following instruction:
"We want to know your explanation of some topics. The aim of this explanation is to show clearly how each step causes the next one, placing them in a sequence from the emergence of the causes until the moment when the phenomenon occurs. In other words, try to tell a story as
complete as you can (with no plot holes) that might be understood by anyone."
When the explanation was completed, participants evaluated again their understanding of the object/topic they had previously explained (post-evaluation phase). The sequence explanation-post evaluation was then repeated for the second object/topic. The presentation order of high and low desirability topics within each domain was randomly assigned.

## Results

A mixed ANOVA was conducted as the main analysis, with judgment timing (pre and post explanation) and perceived social desirability of knowledge on the topic (high and low) as within-subject factors, and both domain of knowledge and presentation order as the between-subjects factors. The dependent variable was the rating in the 7-point understanding scale.

Replicating the IOED phenomenon, a main effect of evaluation time was found. Ratings of understanding before the elaboration of explanations $(M=3.42, S E=.105)$ were higher than those produced after explanations ( $M=2.81$, $S E=.10), F(1,178)=52.43, p<.001, \eta_{\mathrm{p}}^{2}=.23$. Additionally, there was a significant interaction between judgment timing and domain of knowledge, $F(2,178)=$ 3.33, $p<.05, \eta_{p}^{2}=.03$. Post hoc analysis (Tukey's HSD) revealed that the decrease of understanding ratings was higher for the historical domain, $p<.01$ (see Table 2).

Additionally, social desirability of knowledge interacted with judgment timing, $F(2,178)=56.27, p<.001, \eta_{\mathrm{p}}^{2}=.24$. In particular, it was found a decrease on understanding ratings between judgments before and after the elaboration of explanations, only for high desirability topics (see Figure 1).

It was also found a marginally significant three-way interaction between judgment timing, social desirability and domain of knowledge, $F(2,178)=3.06, p=.049, \eta_{\mathrm{p}}^{2}=.03$. Specifically, the reduction of understanding after generating explanations in low social desirability topics is slightly greater in the historical domain (why FARC-EP guerrilla was created), than both in devices (how a jet engine works) and economic topics (how stock markets work; see Figures 2 and 3).

Table 2: Means of understanding in each domain of knowledge by judgment timing.

| Domain | Time | $M(S E)$ | $95 \%$ CI |  |
| :---: | :---: | :---: | :---: | :---: |
|  |  |  | Lower <br> bound | Upper <br> bound |
| Historical | Pre | $4.20(.182)^{* *}$ | 3.84 | 4.56 |
|  | Post | $3.35(.171)^{* *}$ | 3.01 | 3.68 |
| Economic | Pre | $2.90(.182)$ | 2.54 | 3.26 |
|  | Post | $2.26(.171)$ | 1.92 | 2.59 |
| Devices | Pre | $3.16(.185)$ | 2.79 | 3.52 |
|  | Post | $2.83(.174)$ | 2.48 | 3.17 |
| $* * p<.01$ |  |  |  |  |



Figure 1. Means of understanding in high and low social desirability topics by judgment timing. Bars represent 95\% confidence intervals.


Figure 2. Means of understanding in each domain on high social desirability topics) by judgment timing. Bars represent $95 \%$ confidence intervals.


Figure 3. Means of understanding in each domain on low social desirability topics by judgment timing. Bars represent 95\% confidence intervals.

Finally, there was a robust and unexpected three-way interaction between judgment timing, social desirability, and topic order, $F(2,178)=31.03, p<.001, \eta_{\mathrm{p}}^{2}=.14$. Particularly, when low social desirability topics were evaluated first, the understanding ratings for the second (high desirability) topic showed a lower IOED ( $M_{\mathrm{pre}}=3.85$, $S E=.186 ; M_{\text {post }}=3.05, S E=.171$ ) than when high desirability topics were evaluated first ( $M_{\mathrm{pre}}=4.35, S E=.187 ; M_{\text {post }}=2.68$, $S E=.172$ )

## Discussion

In the present study, we examined the relationship between IOED magnitude and social desirability of knowledge about specific topics in three different domains. Our results show that people overestimate their knowledge about causal mechanisms related to physical devices, as well as to economic and historical phenomena. Furthermore, the IOED seems to be stronger in the historical domain and that difference might be related to the higher social desirability of this domain.

Within each domain, the IOED was exhibited for the highly desirable but not for the less desirable topics. This finding confirms that perceived social desirability of knowledge is a relevant cue in the processes of knowledge self-assessment, as it is suggested by previous research on metamemory judgments. It is possible however that other factors like familiarity, accessibility or perceived distribution of information about topics play a role moderating the IOED effect. This could explain, for instance, why the initial understanding of unfamiliar and non-accessible topics (like the low desirability topic in the domain of devices) could be underrated rather than overrated. In this vein, future studies should separate the effect of social desirability from that of potential confounds as far as possible. Even if the influence of other informative cues is demonstrated, it would support the idea that the IOED is not only a consequence of the coarseness of intuitive theories (Rozenblit \& Keil, 2002), but also a by-
product of the inferential nature of metacognitive mechanisms. In other words, people overestimate their ability to explain objects and phenomena because they use multiple cues to assess how well they know them (including social desirability of that knowledge), and not only because they confuse their skeletal understanding with full-detailed representations of mechanistic knowledge.

Fernbach, Rogers, et al. (2013) found that the IOED magnitude correlated positively with the moderation of extreme political attitudes on controversial issues. Accordingly, if the social desirability of knowledge about a political issue enhances the related IOED, it is possible that extreme attitudes about more desirable topics to be also more likely to be moderated after trying to explain them. However, if an individual holds an extreme position about a socially relevant topic (e.g., abortion, gay marriage, gun control, etc.) and this position is relevant to his or her identity, previous evidence suggests that he or she will engage in a form of ideologically motivated cognition, making the related attitude more resistant to change (Kahan 2013). Eventually, this motivational bias could affect the metacognitive processes involved on the IOED. Thus, in some cases, social desirability of knowledge and motivated cognition could influence the IOED magnitude in opposed directions, depending on the personal relevance of topics related to extreme political attitudes. Testing empirically this potential interaction would shed light on the motivational mechanisms involved in the self-assessment of explanatory knowledge. This is important not only in theoretical terms, but also in applied situations like the decision making on complex policies in core political moments (e.g., Brexit referendum or Colombia's peace plebiscite).

Finding that highly desirable knowledge about relevant topics is more likely to be overestimated is not encouraging for deliberative democracies. However, our results suggest that asking participants to explain less desirable topics first can make them less willing to re-calibrate their initial ratings of knowledge about highly desirable topics. Further studies manipulating social desirability of topics betweenrather than within-subjects- would be useful to determine whether previous exposure to low IOED magnitudes can improve the accuracy of understanding estimation about socially desirable topics.

Our purpose in this paper is to bring together the FOK and IOED literatures in order to identify social desirability as an inferential cue in the process of understanding selfassessment. Exploring other interactions between cognitive, motivational and pragmatic factors in metacognitive processes can provide us with a more comprehensive picture of how we know that we know. Although social desirability cannot be randomly assigned, this study shows how it relates to the IEOD in natural settings. Separating social desirability from other factors might be impossible in natural settings and non-ecological in experimental ones. To give an extreme example, separating social desirability from social relevance be done if the former depends intrinsically
of the latter. So we consider that the manipulation here exposed is enough for to establish the relationship between both variables. Further experimental research is required to check if the relationship stands in experimental environments.

## Acknowledgments

Christian Gaviria was funded by the National Doctoral Programs Scholarship 647 from the Colombian Administrative Department of Science, Technology, and Innovation (COLCIENCIAS). We thank Daniela Bermúdez and Camila Garzón for their help on the data collection process.

## References

Alba, J. W., \& Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. Journal of Consumer Research, 27, 123-156.
Alter, A. L., Oppenheimer, D. M., \& Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. Journal of Personality and Social Psychology, 99(3), 436-451.
Atir, S., Rosenzweig, E., \& Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. Psychological Science, 26(8), 1295-1303.
Costermans, J., Lories, G. \& Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(1), 142-150.
De Carvalho Filho, M. K., \& Yuzawa, M. (2001). The effect of social influences and general metacognitive knowledge on metamemory judgments. Contemporary Educational Psychology, 26, 571-587.
Dunlosky, J., \& Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding, students' judgments can improve their achievement. Learning and Instruction, 24, 58-61.
Fernbach, P. M., Rogers, T., Fox, C. R., \& Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. Psychological Science, 24(6), 939-946.
Fernbach, P. M., Sloman, S. A., St. Louis, R., \& Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. Journal of Consumer Research, 39(5), 1115-1131.
Fisher, M., \& Keil, F. C. (2014). The illusion of argument justification. Journal of Experimental Psychology: General, 143(1), 425-433.
Fisher, M., \& Keil, F. C. (2015). The curse of expertise: When more knowledge leads to miscalibrated explanatory insight. Cognitive Science, 40(5), 1251-1269.
Gruneberg, M. M., Monks, J., \& Sykes, R. N. (1977). Some methodological problems with feeling of knowing studies. Acta Psychologica, 41, 365-371.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 56, 208216.

Jaeger, A. J., \& Wiley, J. (2015). Reading an analogy can cause the illusion of comprehension. Discourse Processes: A Multidisciplinary Journal, 52(5), 376-405.
Kahan, D. M. (2013). Ideology, motivation and cognitive reflection: An experimental study. Judgement and Decision Making, 8, 407-424
Koriat, A. (2012). The self-consistency model of subjective confidence. Psychological Review, 119(1), 80-113.
Koriat, A., \& Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and the accessibility heuristics to feelings of knowing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 3453.

Koriat, A., Lichtenstein, S., \& Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118.
Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. Memory \& Cognition, 34(8), 1667-1675.
Mills, C., \& Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. Journal of Experimental Child Psychology, 87, 1-32.
Moore, D. A., \& Healy, P. J. (2008). The trouble with overconfidence. Psychological Review, 115(2), 502-517.
Oppenheimer, D. M., Meyvis T. \& Davidenko N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45(4), 867-872.
Rozenblit, L, \& Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 92, 1-42.
Serra, M. J., \& DeMarree, K. G. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. Memory \& Cognition, 44(7), 1127-1137.
Soderstrom, N. C., \& McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda basedmonitoring. Journal of Experimental Psychology: Learning, Memory and Cognition, 37, 1236-1242.
Thomas, A. K., Lee, M., \& Hughes, G. (2016). Introspecting the elusive: The uncanny state of the feeling of knowing. In J. Dunlosky \& S. K. Tauber (Eds.). The Oxford Handbook of Metamemory. New York, NY: Oxford University Press.
Zell, E., \& Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. Perspectives on Psychological Science, 9(2), 111-125.
Zeveney, A. S., \& Marsh, J. K. (2016). The illusion of explanatory depth in a misunderstood field: The IOED in mental disorders. Proceedings of the 38th Annual Meeting Cognitive Science Society (pp. 1020-1025). Austin, TX: Cognitive Science Society.

# Reasoning with Fundamental Rights 

Lupita Estefania Gazzo Castañeda (Estefania.Gazzo@psychol.uni-giessen.de) Antje Stemmler (Antje.Stemmler@psychol.uni-giessen.de) Markus Knauff (Markus.Knauff@psychol.uni-giessen.de)<br>Justus-Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany


#### Abstract

People often withdraw previously drawn conclusions in light of new information. This defeasible reasoning is also important for law, where judges often have to change their verdicts in light of new evidence. Here we investigate defeasibility in the context of conflicting fundamental rights. When, for instance, law to property conflicts with law to information, can one of these rights be "defeated" by the other? We embedded conflicting fundamental rights in inference tasks (Experiment 1) and in elaborated vignettes (Experiment 2). Results show that people decide between two conflicting fundamental rights in a rational way. Case by case, participants protected that fundamental right whose violation evoked the highest moral outrage (Experiment 1) or whose violation was considered to be more serious (Experiment 2). We discuss the implications of our findings for law theory and psychology.


Keywords: defeasibility, legal reasoning, conditionals

## Introduction

Are humans rational? This question has concerned psychologists and philosophers for a long time. Philosophers have developed norms for rational thinking and psychologists have tested them empirically. In many of these experiments, classical logic was used as a norm for rationality. Participants were confronted with inference tasks, consisting of a conditional and a fact, and asked to indicate what follows necessarily. One example is Modus Ponens (MP):

```
If Ann is hungry (p) then she gets something to eat (q).
Ann is hungry (p).
Ann gets something to eat (q).
```

MP is a valid inference because in classical logic the antecedent $(p)$ is sufficient (but not necessary) for the consequent (q) (e.g., Thompson, 1994; 1995). Yet, many participants made logical errors in such inference tasks, rejecting otherwise valid conclusions. Nowadays, however, it is known that these "errors" are only a consequence of the complexity of human everyday reasoning. In everyday situations, many factors that are irrelevant for classical logic have to be considered and weighted in order to arrive at a reasonable conclusion. For instance, if Ann is on a strict diet, it may be rational to conclude that she will not get something to eat even if she is hungry. Contrary to classical logic, where no additional information can make a conclusion false, everyday reasoning is non-monotonic and defeasible (e.g. Oaksford \& Chater, 1995; 2013; Stenning \& van Lambalgen, 2005).

This phenomenon called defeasibility is also very important in legal reasoning (e.g., Bäcker, 2010; Prakken \&

Sartor, 2004). Judges are often confronted with complex cases, in which they have to arrive at rational verdicts. At first sight, we might thus think that classical deduction is an appropriate norm for legal reasoning. For instance, considering that the penal code says that "If a person kills another human, then the person has to be punished for manslaughter" we can conclude from the fact that a person killed another human that the person has to be punished for manslaughter. However, in many cases there are exculpatory circumstances that make it rational to reject this conclusion. There are different exculpatory circumstances defined in penal code, such as self-defense, necessity or psychological disorders. In light of those circumstances, judges know that the otherwise valid conclusion of punishment has to be defeated in favor of acquittal. But what happens if there are no clear rules on how to reason? Although this might sound counterintuitive for legal contexts, this nonetheless happens in federal constitutional courts, when two fundamental rights are in conflict. Imagine, for instance, that you live in a foreign country and the only way to hear news from your hometown is to mount a parabolic antenna on the facade of your rented flat. The landlady nonetheless prohibits it to you. Your right to receive information thus conflicts with the property law of the landlady. In general terms, all fundamental rights are equally important and have to be granted. So, which fundamental right has to be preferred over the other? Can one fundamental right "defeat" the other?

The aim of this paper is to investigate how people reason with conflicting fundamental rights. For this, we will first discuss the psychological literature on defeasible reasoning and then the law theoretic framework on fundamental rights.

## Withdrawing from valid conclusions

Many factors influence defeasible reasoning. One important factor is background knowledge. Just as shown in the introduction of this paper, when people know circumstances that prevent the consequent to occur although the antecedent is true, they reject otherwise valid conclusions. These circumstances are often called disablers or defeaters. But not only the availability of defeaters is important, also is their amount (e.g., Cummins, 1995; De Neys, Schaeken, \& d'Ydewalle, 2003a), their relative strength (De Neys, Schaeken, \& d'Ydewalle, 2003b), and their frequency of occurrence (Geiger \& Oberauer, 2007). The more defeaters there are, the more associated or salient these defeaters are, and the more often they occur, the more readily a conclusion is withdrawn. However, another important factor are utilities. Bonnefon (2009; Bonnefon \& Hilton, 2004) showed
that when people make inferences about actions, they consider the costs and benefits of this action given the consequences. For instance, when presented with the conditional "If Mary's TV is broken, she will have it fixed" and the fact that "Mary's TV is broken" participants refuse to conclude that Mary will have their TV fixed when presented with the additional information that "If Mary has her TV fixed, she will not be able to pay the electricity bill". The consideration of utilities during reasoning illustrates the closed connections between reasoning and decision making. When people reason in their daily lives, it is not just for the sake of reasoning per se, but to reach a goal - may this goal be something simple like getting a TV fixed or something relevant for society like reaching a legal verdict. Because goals are also relevant in law we expect utilities to play a similarly important role in legal reasoning.

## Legal reasoning

Law is defeasible in several ways (cf. Prakken \& Sartor, 2004): during police investigations when new evidence "defeats" previous insights, during trials when attorneys and prosecutors defeat each other's arguments, and in the application of legal rules in light of exculpatory evidence. The role of utilities in the application of legal rules has been tested by Gazzo Castañeda and Knauff (2016). In several experiments, laypeople and lawyers were confronted with legal conditionals embedded in MP inferences, which were presented together with exculpatory evidence (e.g., "If a person kills another human, then the person should be punished for manslaughter; Bob killed another person; Bob is schizophrenic and had a delusion of an attack against him; Should Bob be punished for manslaughter?"). As expected, lawyers considered exculpatory circumstances as prescribed by the penal code and irrespective of how morally outraging the offence was, deciding not to punish in light of exculpatory circumstances. Laypeople, instead, had difficulties in accepting exculpatory circumstances when the offence was highly morally outraging (e.g., maltreatment of wards), but not if the moral outrage was only low (e.g., illegal gambling). We argued that utilities might be responsible for this moral outrage effect. People can only feel secure in a society where they can be sure that the important rules are respected and offenders punished. The benefit of saving one's own feeling of security is thus weighted more than the costs of punishing somebody erroneously. This overweighing of one's own feeling of security is known from the belief in a just world literature (see Lerner, 1970), where people even tend to blame the victims of offences only to preserve their belief that people get what they deserve and that bad things only happen to bad people. From a utilitarian point of view, the punishment of offenders is thus of high utility - and the higher the moral outrage, the higher this utility is. Is it therefore possible that moral outrage also affects the weighing of fundamental rights?

Fundamental rights are generally coded in the constitution. The most known examples are right to dignity, liberty, freedom of thought and of expression, or right of property.

All of these have to be respected and protected. However, there are instances when two or more fundamental rights are in conflict, such as in the introductory example when the right to information conflicts with right to property. Judges in the federal court are thus faced with the problem that they have to decide which one deserves more importance, although both are theoretically equally important. This weighting of fundamental rights is called balancing and is an important case by case decision with no clear rules on how to decide. That is, cases with the same conflicting fundamental rights can (and should) end up with different verdicts due to case-specific details. Because of these missing rules, some law theorists argue that balancing cannot be rational (e.g., Habermas, 1992). Alexy (2003), however, argues that balancing can be rational by comparing for every single case the detriment of one fundamental right $i$ with the importance of satisfying the other fundamental right $j$. This is done by the so-called weight formula, which computes the ratio between the case-specific weights $\mathrm{I}_{\mathrm{i}}$ and $\mathrm{I}_{\mathrm{j}}\left(\mathrm{G}_{\mathrm{ij}}=\mathrm{I}_{\mathrm{i}} /\right.$ $\mathrm{I}_{\mathrm{j}}$ ). The first weight $\mathrm{I}_{\mathrm{i}}$ stands for the violation intensity of fundamental right $i$ by protecting fundamental right $j$, and the second weight $\mathrm{I}_{\mathrm{j}}$ stands for the importance of protecting fundamental right $j$ by violating fundamental right $i$. Applied to our concrete example, this would result in the following two questions: How serious is the invasion of the right to information by prohibiting the installation of the parabolic antenna? How important is it to protect the right to property by prohibiting the installation of the parabolic antenna? Already Darley showed in several experiments that the perceived seriousness or severity of offences is highly correlated by moral outrage (e.g., Alter, Kernochan, \& Darley, 2007; Carlsmith, Darley, \& Robinson, 2002; Darley, Carlsmith, \& Robinson, 2000). Therefore, we also expect that the case-specific weights of fundamental rights will depend on moral outrage: If the invasion of fundamental right A is considered more morally outraging than the invasion of fundamental right B , then fundamental right A should be protected over B.

In this paper, we combine the domains of defeasible reasoning from psychology with the concept of balancing from legal theory. In Experiment 1, we embedded two fundamental rights into conditional reasoning tasks and asked participants what should follow. In Experiment 2, we embedded conflicting fundamental rights into longer vignettes and asked participants for the case-specific weights $I_{i}$ and $I_{j}$.

## Experiment 1

## Methods

Participants We tested 40 people ( 21 male) without legal expertise. They were on average 26.62 years old ( $S D=6.93$ ). Material We took 16 real conflicts of fundamental rights from the German constitutional court and embedded them in defeasible inference tasks. Each problem started with a conditional containing one fundamental right A. Next, we presented a concrete situation as second premise in which the fundamental right A was involved, followed by a third
premise in which the fundamental right A is applied to this concrete situation (MP). Then, the second fundamental right B was presented as a defeater that is in conflict with the previous information. Finally, the conclusion was presented as a question asking either for the application of fundamental right A (Example 1) or fundamental right B (Example 2):
If a person's personal security is endangered, then its protection has to be warranted.
Person A's house has to be searched and seized because A is suspected to have death threatened person B.
To protect B this search and seizure can be authorized.
The suspect A has nonetheless right to privacy.
Should the house of suspect A be searched and seized?
If the personality rights are in danger, then their protection has to be warranted.
A celebrity is photographed without permission.
Due to the personality rights, all people's privacy has to be protected.
The press has nonetheless right to freedom of the press.
Should the celebrity be photographed by the press without permission?
Participants gave ratings from 1 (not at all) to 7 (definitely). Therefore, the higher a rating was, the more the participants preferred one fundamental right over the other. We refer to this as "preference rating".

We created two versions of the experiment by changing the order of the fundamental rights A and B (version 1 and 2). If in one version one fundamental right was presented as the conditional and the other as the defeater, then in the other version it was the other way around. The conclusion, however, always asked for the same fundamental right. This allowed us to control for order effects.

To measure moral outrage we conducted a norming study in which participants $(N=34)$ rated on a seven point Likertscale how morally outraged they would feel if the fundamental rights from the inference tasks were violated (e.g., "How outraging do you find it when a celebrity is photographed without permission?"). Because in each conflict situation there were two fundamental rights involved, this resulted in 32 violation ratings, ranging from 2.65 to $6.09 .{ }^{1}$

Procedure and Design The experiment was programmed on Superlab 4.5. Participants were tested individually on a desktop computer and were instructed that no right or wrong answers exist. The instructions included one practice problem. All 16 problems were presented randomly and separated by fixation crosses. The single premises were presented sequentially on separate screens. Participants could switch to the next premise by pressing the space bar. The last premise was always the question about the conclusion. It was

[^338]written in red font and was presented together with the 7 -point-Likert scale. The experiment was thus one factorial with "version" as a between subject variable.

## Results

Comparisons between the two versions of the experiment revealed no differences in preference ratings, $t(38)=1.36$, $p=.181$. That means that regardless of whether a fundamental right was presented as the conditional or as the defeater, this did not affect its evaluation in the conclusion. This allowed us to test the effect of moral outrage on preference ratings. For this, we first compared the two fundamental rights in each problem on the basis of the moral outrage ratings they got in the norming study. We looked at which fundamental right violation got higher moral outrage ratings and should thus be preferred. These predictions were then compared with the actual preference ratings participants gave in the experiment. Mean preference ratings over 4 (i.e., the scale midpoint) were classified as in favor, and ratings below 4 against the fundamental right presented in the conclusion (no mean preference rating $=4$ ). Descriptively, the moral outrage ratings allowed us to correctly predict 11 out of the 16 conflict situations. To corroborate this statistically, we tested the preference ratings of each inference task against 4 with one sample $t$-tests and a Bonferroni adjusted alpha of 0.0031 . Results are in Table 1. Of the 16 comparisons, 6 were not significantly different from 4 , meaning that participants were neither in favor nor against the fundamental right presented in the conclusion. From the remaining 10 problems, however, we were able to predict statistically 8 conflict situations.

Table 1: Predicted and actually preferred fundamental rights. Predictions were based on the moral outrage (MO) ratings from the norming study. Preference ratings of the actually preferred rights were tested against the scale midpoint 4 (Sign., Bonferroni adjusted alpha: 0.0031).

| Item | MO of A | MO of B | Predicted | Preferred | Sign. |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 3.15 | 5.44 | B | A | $<.001$ |
| 2 | 4.88 | 5.59 | B | A | $<.001$ |
| 3 | 3.88 | 3.32 | A | A | .001 |
| 4 | 3.85 | 4.21 | B | B | .001 |
| 5 | 3.71 | 4.59 | B | A | .147 |
| 6 | 3.29 | 4.71 | B | B | $<.001$ |
| 7 | 5.79 | 4.38 | A | A | $<.001$ |
| 8 | 3.76 | 5.26 | B | A | .008 |
| 9 | 4.41 | 5.06 | B | B | .386 |
| 10 | 3.24 | 5.53 | B | B | $<.001$ |
| 11 | 3.82 | 3.74 | A | A | .103 |
| 12 | 5.76 | 4.06 | A | A | $<.001$ |
| 13 | 5.18 | 5.44 | B | B | .309 |
| 14 | 2.65 | 6.09 | B | B | $<.001$ |
| 15 | 3.39 | 3.26 | A | B | .305 |
| 16 | 5.32 | 5.29 | A | A | $<.001$ |

As an additional measure of the relevance of moral outrage for balancing, we took the moral outrage ratings from the norming study and used these ratings to classify the fundamental rights in each problem as either low (ratings from 2.65 to 3.76 ), medium (from 3.82 to 4.88 ), or high (from 5.06 to 6.09 ) morally laden (the cut offs resulted from the division of our fundamental rights into these three groups).

An analysis of the respective preference ratings showed that when high and medium morally laden fundamental rights were in conflict, participants gave higher preference ratings for conclusions asking for the highly morally laden right ( $M=5.39 ; S D=1.20$ ) than for conclusions asking for the medium morally laden right ( $M=4.46 ; S D=1.36$ ), $t(39)=3.06$, $p=.004$. The same was the case for problems where medium and low morally laden fundamental rights were in conflict: conclusions asking for the medium morally laden right got higher preference ratings ( $M=4.5$; $S D=1.89$ ) than their counterpart $(M=3.25 ; S D=1.01), t(39)=3.72, p=.001$. We also found the same pattern for problems with high and low morally laden fundamental rights, but the difference in preference ratings ( $M=4.15$; $S D=1.09$ vs. $M=3.73$; $S D=1.36$ ) did not reach significance, $t(39)=1.38, p=.176$ (Bonferroni adjusted alphas: 0.0167).

## Discussion

Moral outrage was an important predictor when deciding between two fundamental rights in a defeasible reasoning paradigm. Participants protected more often that fundamental right whose violation provokes the highest moral outrage. A fundamental right was therefore only considered as a defeater if its violation was morally outraging enough.

Our results demonstrate the defeasibility of human reasoning. Even in contexts where we expect people to reason deductively - such as in law - reasoning is often defeasible. In fact, it is difficult to describe balancing through deduction. Deduction would imply that it should be (in principle) possible to enumerate all defeaters beforehand as part of the antecedent (e.g., If right to property is in danger and it does not conflict with right to information, then we have to protect it). Yet, this is not possible for balancing because cases with the same conflicting fundamental rights can end up with different verdicts due to case-specific details. An important task for cognitive psychologists therefore is to understand the cognitive processes behind balancing. Our research is a first step in this direction.

An open question is, however, whether participants are indeed capable to consider these case-specific details during balancing. Is balancing just a theoretical concept from legal theory? Or do people in fact balance and defeat fundamental rights differently depending on case-specific details? We tested this in Experiment 2.

## Experiment 2

In Experiment 2 we used a new experimental paradigm. First, we embedded the conflicting fundamental rights in elaborated vignettes. The vignettes described many casespecific details and were thus more realistic than the inference tasks from Experiment 1. Second, half of our participants were people with legal expertise. This allowed us to investigate balancing in a more realistic "court like" setting.

## Methods

Participants We tested 40 laypeople ( 17 male) and 40 lawyers (already graduated ones and advanced law students; 18
male). On average, laypeople were 24.3 years old ( $S D=4.2$; one missing value) and lawyers 24.8 years old ( $S D=3.1$ ).
Material We constructed our material by summarizing 8 real cases from the German federal constitutional court (i.e., BVerfGE), and embedding these into vignettes. Each vignette contained case-specific details such as the matter of facts, the case history (e.g., accusations, levels of jurisdiction involved), and the parties' arguments in favor or against the different fundamental rights. We selected our cases in such a way that two of them always contained the same conflicting fundamental rights, but received different verdicts from the constitutional court. These final verdicts were, however, not included in the vignettes. We had thus four pairs of cases: two cases of right to information vs. right to property, two cases of personality rights vs. right to freedom of press, two cases of personality rights vs. right to freedom of speech, and two cases of right to bodily integrity vs. the public interest to legal action. The vignettes were 324 to 506 words long and were developed by an advanced law student with the supervision of an experienced legal researcher.

The participants' tasks were (1) to come to a final verdict and (2) to determine the specific weights for $I_{i}$ and $I_{j}$. The question about the final verdict was formulated according to the legal theoretic tradition: "How would you decide? Which interest should resign: [Right A] or [Right B]?" Participants could select between " $[$ Right A] should resign", " $[$ Right B] should resign", and "Both interest deserve equal protection (standoff)". The question about the specific weights was split into two parts: First, participants had to judge the intensity of violation of right $A$ (e.g., "How intense do you think is the violation of T's right to information by prohibiting him to install a parabolic antenna?"). Second, participants had to judge the importance of protecting right $B$ (e.g., "How important is it to protect the right to property of the landlady by prohibiting the installation of the parabolic antenna?"). Participants had to select between "little", "medium", "very".
Procedure and Design The experiment was conducted via paper and pencil. Each vignette was presented on a small booklet containing on the first page the vignette and on the second page the questions about (1) the verdict and (2) the specific weights (in this order). Participants were instructed to imagine they were judges in the constitutional court. Each participant received 4 of the 8 vignettes, one of each pair. The order of the vignettes was randomized. The experiment thus followed a 4(type of conflict) x 2(pair) design, with "pair" as a between subjects variable and the type of conflict as a within subjects variable.

## Results

We first analyzed in how many cases the specific weights predicted the overall verdicts. As a correct prediction we counted (1) cases in which the fundamental right protected in the final verdict was also the one with the highest specific weight, and (2) cases in which participants weighted both rights equally in the questions about the specific weights and selected "standoff" as the verdict. This analysis showed
that correct predictions were significantly above chance: we could predict $71.9 \%$ of the laypeople's, $t(39)=5.31, p<.001$, and $81.9 \%$ of the lawyers' verdicts, $t(39)=8.64, p<.001$.

In a second step, we analyzed whether participants considered the case-specific details. Therefore, we looked at the four pairs of conflicting fundamental rights and compared within each pair how often Right A, Right B, or standoff were selected. We compared the frequency distributions of the three kinds of verdicts with Freeman-Halton tests. Indeed, results showed that in light of different case-specific details, participants gave different verdicts for the same conflicting fundamental rights. Laypeople did so for 2 of the 4 pairs of cases (personality rights vs. freedom of speech: $p=.002$; bodily integrity vs. public interest: $p=.001$ ), and lawyers did so for 3 of the 4 pairs of cases (personality right vs. freedom of speech: $p=.009$; personality rights vs. freedom of press: $p=.002$; bodily integrity vs. public interest: $p<.001$ ). However, only in $38 \%$ (laypeople) and in $55 \%$ (lawyers) of the cases the participants' final verdicts was the same as the actual verdicts from the constitutional court.

## Discussion

Experiment 2 shows that participants often decide between conflicting fundamental rights by considering case-specific details. This supports our main assumption that balancing is defeasible. In our study, participants did not apply some general rule (e.g., right to information deserves more importance than right to property), but decided on a case by case manner whether a specific fundamental right counts as a defeater or not. This defeasibility seems to be wellcaptured by Alexy's (2003) weight formula. An interesting question now is whether the basic idea of this formula is also helpful for understanding defeasibility outside the legal context. Take our initial example of Ann being hungry. Maybe people decide to defeat the conclusion that people eat when they are hungry by comparing the weights of "if hungry then eating" and "if hungry then not eating". Interestingly, this comparison of weights is similar to the concept of conditional probabilities (e.g., Evans \& Over, 2004). Many theories on conditional reasoning argue that defeasibility results from the fact that conditionals are understood as the conditional probability $\mathrm{P}(q \mid p)$, which is computed by dividing $\mathrm{P}(p \& q)$ with $\mathrm{P}(p \& q)+\mathrm{P}(p \& n o t-q)$. That is, similar to the weight formula, the weight (here the probability) of $p$ and $q$ is compared with the one of $p$ and not- $q$. This similarity, we think, deserves more attention from psychology and also from law theory.

Another point that also deserves attention is the mismatch between the final verdicts of our participants (laypeople and lawyers) and the actual verdicts of the constitutional court. On the one hand, participants followed the weight formula. Thus, they weighted the single fundamental rights in a rational way. On the other hand, our results indicate that they used specific weights that differed from those used by the constitutional court. This might be a result of our specific task setting, relatively low test power, the small number of vignettes we used, and the limited ecological validity of our
study. However, another interpretation is that the ethical values and moral principles that drive people's decisions differ from that of our legal system. We think that this is an important research topic at the intersection of cognitive science, social psychology, legal theory, and moral philosophy.

## General Discussion

We used methods from cognitive psychology to investigate the concept of balancing from legal theory. Our results show that people are willing to defeat single fundamental rights if they are in conflict with other fundamental rights. This defeasibility happens in a case-specific manner and not only when conflicting fundamental rights were presented in inference tasks, but also when they were embedded in ecologically more valid vignettes.

Our findings are important for several reasons. First, they show the importance of defeasible reasoning in many areas of real life. Defeasibility is important if we judge how severe violations of fundamental rights are and when we weight the importance of conflicting fundamental rights. Interestingly, however, some law theorists do not consider balancing as defeasible. According to Bäcker (2010), defeasibility describes the capacity to accommodate legally relevant exceptions. Therefore, only "normal" legal rules would be defeasible (e.g., those from penal code), but not fundamental rights. Fundamental rights are legal principles that have to be optimally achieved taking into account all possible circumstances, including other conflicting rights. Therefore, one fundamental right cannot be an exception to another fundamental right (Bäcker, 2010). From a psychological perspective two reasons speak against this view. The first is an empirical: the important aspect of psychological defeasibility is that people change their conclusions in light of new evidence and this certainly happens when one decides against one right in light of another right. The second reason is a theoretical: that one fundamental right cannot be an exception to another one does not speak against the defeasibility of balancing. In fact, the case by case weighting of fundamental rights is precisely what makes balancing defeasible and non-monotonic. Would one fundamental right be considered an "exception" of another, then we could theoretically enumerate it as part of some rule and reason deductively.

Second, our findings also help to understand the psychological variables behind the weight formula. According to Alexy (2003), balancing depends on specific weights, which reflect how serious it is not to protect one right or the other. It is, however, not clear how exactly judges determine this "seriousness". We operationalized this seriousness through moral outrage, which resulted to be a good predictor for the final verdicts. Is it thus possible that judges' balancing of fundamental rights is influenced by the level of moral outrage? The fact that both, moral outrage (Experiment 1) and the specific weights (Experiment 2), were good predictors for the final verdicts suggests this. Certainly, most judges and lawyers will not accept this view and it is indeed too early to come to this conclusion. However, we think it is
worthwhile to further study the relation between balancing and moral outrage. In these studies, the role of associative strength should also be considered. The concept of associative strength was introduced by Quinn and Markovits (1998) and applied to defeaters by De Neys et al. (2003b). According to De Neys and colleagues, a defeater has a highly associative strength if it is represented in memory as a good reason to prevent $q$ although $p$ is true. This could be also applied to balancing. A participant will probably only defeat a fundamental right $A$ by another fundamental right $B$, if $B$ is highly associated in one's memory as a reason to prevent A. For instance, one would defeat right to privacy by right to personal security, if personal security is represented in one's memory as more important than the right to privacy. The only problem with this approach is, however, that it is not clear whether associative strength captures all the casespecific details necessary for balancing. As already described, according to law theory, cases with the same conflicting fundamental rights do not have to end up with the same verdicts. Whether these case-specific circumstances that are decisive for balancing - are represented in our memory to influence their associative strength requires further investigation. Maybe the associative strength provides some general, case independent, overall weight to balancing, whereas moral outrage is responsible to tune the specific weights in accordance to the case-specific details.

Finally, our studies also show that paradigms from cognitive psychology are useful to investigate questions from other fields. Conditional inference tasks were originally introduced to test people's capacity to reason according to classical logic. In our study, however, we showed that inference tasks are also useful to test accounts from legal theory. We think that this is true for many other areas as well. For instance, inference tasks can also be helpful to study moral reasoning, where - similar to balancing - people also have often to decide between two conflicting principles (e.g., telling the truth or lying to not hurt someone's feelings).

## Acknowledgments

This work was supported by DFG grants KN 465/10-1 and KN 465/10-2 to Markus Knauff within the DFG Priority Program "New Frameworks of Rationality".

## References

Alexy, R. (2003). Die Gewichtsformel. In v. J. Jickeli, P. Kreutz, \& D. Reuter (Eds.), Gedächtnisschrift für Jürgen Sonnenschein (pp.69-78). Berlin: De Gruyter Recht.
Alter, A. L., Kernochan, J., \& Darley, J. M. (2007). Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. Law and Human Behavior, 31, 319-335.
Bäcker, C. (2010). Rules, principles, and defeasibility. In M. Borowski (Ed.), On the nature of legal principles, ARSPBeiheft 119 (pp. 79-91). Stuttgart: Franz Steiner Verlag.
Bonnefon, J.-F. (2009). A theory of utility conditionals: Paralogical reasoning from decision-theoretic leakage. Psychological Review, 116, 888-907.

Bonnefon, J.-F., Hilton, D. J. (2004). Consequential conditionals: Invited and suppressed inferences from valued outcomes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 28-37.
Carlsmith, K. M., Darley, J. M., \& Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology, 83, 284-299.
Cummins, D. D. (1995). Naive theories and causal deduction. Memory \& Cognition, 23, 646-658.
Darley J. M., Carlsmith, K. M., \& Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. Law and Human Behavior, 24, 659-683.
De Neys, W., Schaeken, W., \& d'Ydewalle, G. (2003a). Inference suppression and semantic memory retrieval: Every counterexample counts. Memory \& Cognition, 31, 581-595.
De Neys, W., Schaeken, W., \& d’Ydewalle, G. (2003b). Causal conditional reasoning and strength of association: The disabling condition case. European Journal of Cognitive Psychology, 15, 161-176.
Evans, J. St. B. T, \& Over, D. E. (2004). If. Oxford, England: Oxford University Press.
Gazzo Castañeda, L. E., \& Knauff, M. (2016). Defeasible reasoning with legal conditionals. Memory \& Cognition, 44, 499-517.
Geiger, S. M., \& Oberauer, K. (2007). Reasoning with conditionals: Does every counterexample count? It's frequency that counts. Memory \& Cognition, 35, 2060-2074.
Habermas, J. (1992). Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaat. Frankfurt a. M: Suhrkamp.
Lerner, M. J. (1970). The desire for justice and reactions to victims. In J. Macaulay \& L. Berkowitz (Eds.), Altruism and helping behavior. New York: Academic Press.
Oaksford, M., \& Chater, N. (1995). Theories of reasoning and the computational explanation of everyday inference. Thinking \& Reasoning, 1, 121-152.
Oaksford, M., \& Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. Thinking \& Reasoning, 19, 346-379.
Prakken, H., \& Sartor, G. (2004). The three faces of defeasibility in law. Ratio Juris, 17, 118-39.
Quinn, S., \& Markovits, H. (1998). Conditional reasoning, causality and the structure of semantic memory: Strength of association as a predictive factor for content effects. Cognition, 68, B93-B101.
Stenning, K., \& van Lambalgen, M. (2005). Semantic interpretation as reasoning in nonmonotonic logic: The real meaning of the suppression task. Cognitive Science, 29, 919-960.
Thompson, V. A. (1994). Interpretational factors in conditional reasoning. Memory \& Cognition, 22, 742-758.
Thompson, V. A. (1995). Conditional reasoning: The necessary and sufficient conditions. Canadian Journal of Experimental Psychology, 49, 1-58.

# The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents 

Felix Gervits (felix.gervits@tufts.edu)<br>Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02155<br>Gordon Briggs (gordon.briggs.ctr@nrl.navy.mil)<br>NRC Postdoctoral Fellow, Naval Research Laboratory, Washington, DC 20375<br>Matthias Scheutz (matthias.scheutz@tufts.edu)<br>Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02155


#### Abstract

One of the hallmarks of human natural language (NL) interaction is the ability for people to balance a variety of social and communicative goals when choosing how to realize their speech actions. These goals can include pragmatic criteria such as correctness, informativeness, and brevity (i.e., Gricean conversational maxims) or social factors such as politeness. However, there currently does not exist a general algorithmic method to explicitly modulate language generated by artificial agents based on an arbitrary number of pragmatic and social criteria. We propose a novel method to accomplish this task, in which rankings of candidate utterances by different pragmatic or social criteria are fused by use of a voting algorithm. We then give a proof-of-concept demonstration of the application of this method in the context of directive generation for human-robot interaction.


Keywords: Human-Robot Interaction; Pragmatics; Natural Language Generation; Politeness

## Introduction

One of the key strengths of humans as social agents is the ability to adapt our language to the communicative norms and needs of the present situation. When giving directives and making requests, we know when it is appropriate to be terse and direct (e.g., "Move out, double-time!"), and when it is appropriate to be polite and circumspect (e.g., "Would you mind passing the salt, please?"). In all our natural language (NL) interactions, we are faced not only with the complex problem of what to say, but also how to say it. Much of this complexity originates from the fact that the intended meaning of utterances in different situational contexts often differs with the literal meaning. For example, asking a waiter, "Can I have a steak?" is not a literal query as to one's physical ability to possess a particular menu item, but rather a means to convey an order.

Dialogue interaction for artificial agents is often viewed from a plan-oriented standpoint, in which the key planoperators are speech actions used to achieve some high-level set of task goals. The precise way in which these speech actions are realized (in so far as it does not affect the efficacy of the speech act) is often of secondary concern. As NLenabled agents become more prevalent in society, and as their manufacturers increasingly market these devices as "social" agents ${ }^{1}$, the disparity between the state-of-the-art in compu-

[^339]tational NL systems and the richness of human-generated language will become increasingly apparent. As such, the ability for an NL-enabled agent to consider and modulate their generated language in human-like ways will become correspondingly more relevant and important.

There is a sizable literature that draws inspiration from pragmatics and socio-linguistics in order to address specific subproblems in natural language generation (NLG) at the subsentential, sentential, and discourse levels. For example, there has been extensive work in operationalizing Gricean pragmatic criteria (Grice, 1975) at the subsentential level, specifically in the area of referring expression (RE) generation (Dale \& Reiter, 1995; Krahmer \& Van Deemter, 2012), in which considerations of correctness, informativeness, and brevity are addressed. There also exists a small body of work that seeks to modulate NLG at the sentential level (Briggs \& Scheutz, 2013; Gupta, Walker, \& Romano, 2007; Miller, Wu, \& Funk, 2008). These approaches seek to operationalize the notion of face-threat from politeness theory, and adjust the behavior of an agent accordingly.

Much of the previous work at the intersection of pragmatics, socio-linguistics, and NLG focuses on tackling specific subproblems in NLG or on modulating language based on a small set of criteria, such as politeness, e.g., Gupta et al. (2007). Yet, in order to generate more human-like language, a much more general framework is necessary. Below we propose some features that such a framework should possess:

1. The method of NLG modulation should be able to explicitly consider an extensible number of pragmatic and sociolinguistic criteria.
2. The method of NLG modulation should be adaptable such that the current situational context may affect the relative importance of communicative criteria.
3. The method of weighing communicative criteria should be agnostic to the choice of the underlying semantic representations used by the system.

At present, there exists no framework that meets all of these criteria. Much of the work in RE generation implicitly considers pragmatic criteria in the design of its algorithms (i.e., RE generation algorithms often search in order of shortest to
longest solution and terminate when a sufficiently informative solution is found (Bohnet \& Dale, 2005)), but does not provide an extensible framework for pragmatic and sociolinguistic modulation. Work such as Briggs and Scheutz (2013) is extensible, but it sorts potential utterances according to a fixed preference ordering of communicative goals, and its adaptability is limited. The work in Bayesian cognitive models of pragmatics (Goodman \& Stuhlmüller, 2013) can be extended to account for social communicative criteria, but it is tightly coupled to semantic representations and small domains amenable to Bayesian computational algorithms. Finally, there are promising approaches which meet some of the requirements, but they are limited to specific domains such as tutoring (e.g., Moore, Porayska-Pomsta, Varges, and Zinn (2004); Nye, Graesser, and Hu (2014)), and do not offer general solutions outside of that context.

In the following section, we present an approach that possesses all of the above desired features. We focus, in this paper, on the problem of modulating generated language at the sentential level, though we hope to apply similar techniques to NLG problems at subsentential and discourse levels. We first begin by examining various communicative goals that NL-enabled agents may need to consider. Next, we present a novel method of balancing these communicative criteria based on techniques from social choice theory (specifically, voting algorithms). Finally, we demonstrate our approach in the context of a human-robot interaction (HRI) scenario, and discuss directions for future work.

## Utterance Selection

In this section we describe an utterance selection algorithm designed to achieve the sort of linguistic modulation we have proposed. In Figure 1 we outline the key components to this approach, which bridges, within the context of an NLG architecture, the output of a dialogue planning component (responsible for selecting an appropriate sequence of speech actions to achieve some agent goal) and the input of an NLG surface realizer component, which is responsible for translating some symbolic linguistic representation into text to be displayed or to be output via text-to-speech. In many architectures, this connection is direct. However, as we have previously addressed, there are multiple ways of realizing speech actions. To effectively consider them, we need the following components:

- A component that factors situational context to produce multiple potential candidate utterance realizations for a given speech action. Examples of NLG pipelines that include such a component are Briggs and Scheutz (2013) and Gupta et al. (2007).
- A set of pragmatic or social criteria $\mathbf{P}$, each with a corresponding utility function $U_{\rho}(\rho \in \mathbf{P})$, that generates a weak preference order over candidate utterances ( $\Upsilon$ ). These criteria include correctness (Maxim of Quality), informativeness (Maxim of Quantity), directness and brevity (Maxim of Manner), and politeness.


Figure 1: Diagram outlining an architecture for flexible NLG that is modulated by an extensible number of pragmatic criteria. The dotted line represents the architectural components we focus in detail on in this paper.

- A component that factors in the agent's beliefs about the current situational context, current goals, and potentially any "personality" model given to the agent in order to produce a set of weights for each pragmatic criterion: $\mathbf{W}=$ $\left\{W_{1}, \ldots, W_{|\mathbf{P}|}\right\}$, where $W_{\rho} \in \mathbb{N}$ denotes the current strength of criteria $\rho$.
- A component that merges the rankings of candidate utterances $\Upsilon$ produced by the pragmatic criteria evaluations $\left(U_{1}, \ldots, U_{|\mathbf{P}|}\right)$ in accordance with the weights generated by the communicative norm reasoner.

In order to merge the rankings of candidate utterances, we used the Schulze voting method (Schulze, 2011), where each ordering produced by $U_{\rho}$ was counted $W_{\rho}$ times. This voting method is a ranked single-winner election system from social choice theory, which is used by many organizations to select a candidate that maintains voters' individual preferences. While this approach has not been previously applied to the domain of computational pragmatics, we find that it offers a robust, computationally-tractable solution to the problem of balancing communicative goals in natural language generation. In the following sections, we present a proof of concept demonstration of our framework, and show how it can be used to generate socially-appropriate directives in the context of human-robot interaction.

Table 1: Utterance selections for various communicative criteria priority orderings

| Relative Criteria Weightings | Utterance Selected | Utterance Output |
| :--- | :--- | :--- |
| Directness $>$ Brevity $>$ Politeness | Instruct $(R, \beta, \operatorname{do}(\beta$, plug_in $(R)),\{ \})$ | "Plug me in" |
| Directness $>$ Politeness = Brevity | Instruct $(R, \beta, \operatorname{do}(\beta$, plug_in $(R)),\{$ please $\})$ | "Plug me in"//"Plug me in, please" |
| Brevity $>$ Politeness $>$ Directness | AskYN $(R, \beta$, capableOf $(\beta$, plug_in $(R)),\{ \})$ | "Could you plug me in?" |
| Politeness $>$ Brevity $>$ Directness | AskYN $(R, \beta$, capableOf $(\beta$, plug_in $(R)),\{$ please $\})$ | "Could you plug me in, please?" |
| Directness = Politeness = Brevity | Instruct $(R, \beta$, do $(\beta$, plug_in $(R)),\{ \})$ | "Plug me in"/"Plug me in, please" |
| Politeness $>$ Directness = Brevity | AskYN $(R, \beta$, capableOf $(\beta$, plug_in $(R)),\{$ please $\})$ | "Could you plug me in, please?" |

## Demonstration: Directive Generation

In order to demonstrate the generality of this framework, we describe how our proposed framework has been integrated with the NL pipeline in a cognitive, robotic architecture, DIARC (Schermerhorn, Kramer, Middendorff, \& Scheutz, 2006). There has been growing interest in the field of HRI in the ways in which robots could phrase requests for assistance from human interaction partners with respect to politeness and other social norms (Gupta et al., 2007; Srinivasan \& Takayama, 2016; Strait, Canning, \& Scheutz, 2014; Torrey, Fussell, \& Kiesler, 2013). Below we present how our framework can be used to address this challenge.

## Framework Configuration

In DIARC, utterances are represented in the following form:

$$
v=\text { UtteranceType }(\alpha, \beta, X, M)
$$

where UtteranceType denotes the speech act classification, $\alpha$ denotes the speaker, $\beta$ denotes the addressee, $X$ denotes an initial semantic analysis, while $M$ denotes a set of sentential modifiers (e.g., "please"). The pragmatic reasoning component in the architecture associates an utterance $v$ in context $C$ with a set of implications:

$$
v_{C}:=\left\langle\mathbf{B}_{\text {lit }}, \mathbf{B}_{\text {int }}, \theta\right\rangle
$$

Each rule associates a particular utterance form $v$ in context $C$ with a tuple containing the set of beliefs $\mathbf{B}_{\text {int }}$ to be inferred based on the intended meaning of the utterance, the set of beliefs to be inferred based on the literal meaning of the utterance $\mathbf{B}_{\text {lit }}$, as well as the degree $\theta$ to which the utterance can be considered a face-threatening act (i.e., a threat to a person's self-image or autonomy) in context $C$ (Brown \& Levinson, 1987).

We define the criterion of correctness as:

$$
U_{\text {correct }}\left(v_{C}, \beta\right)=-\left|\left\{x: x \in \mathbf{B}_{\text {int }}\left(v_{C}\right) \wedge \beta \nvdash x\right\}\right|
$$

where $\beta$ consists of the agent's current set of beliefs. Therefore, utterances that imply more facts unsupported by the agent's beliefs are considered less correct than those that imply fewer unsupported facts. We define the criterion of informativeness as:

$$
U_{\text {inform }}\left(v_{C}\right)=\left|\mathbf{B}_{\text {int }}\left(v_{C}\right)\right|
$$

such that utterances that imply more facts are considered more informative than those that imply fewer facts. We define the criterion of directness as:

$$
U_{\text {direct }}\left(v_{C}\right)= \begin{cases}1 & \mathbf{B}_{\mathbf{l i t}}=\mathbf{B}_{\mathbf{i n t}} \\ 0 & \mathbf{B}_{\mathbf{l i t}} \neq \mathbf{B}_{\mathbf{i n t}}\end{cases}
$$

such that utterances in which the literal and intended meanings are the same are considered more direct than those in which they differ. We define the criterion of politeness as:

$$
U_{\text {polite }}\left(v_{C}\right)=-\theta\left(v_{C}\right)
$$

such that utterances in which the associated face-threat value $(\theta)$ are lower are considered more polite than those in which in it is higher. Finally, we define the criterion of modifierbrevity such that:

$$
U_{m-b r e v i t y}\left(v_{C}\right)=-|M|
$$

utterances with fewer sentential modifiers are considered briefer than those with more sentential modifiers ${ }^{2}$.

## Example Scenario

In this section, we present a proof-of-concept demonstration of the pragmatic modulation framework as applied to a directive formulation problem. Consider a scenario in which an NL-enabled robot is low on charge and needs a human to plug it in (want (bob, plug_in(self))). This will require a directive to be formulated and communicated to the human in order to accomplish the end goal of being plugged in. We consider four main directive formulation strategies in this scenario, realized in the following pragmatic rules in the architecture's dialogue component ${ }^{3}$ :

$$
\begin{align*}
& \operatorname{Instruct}(\alpha, \beta, X, M):= \\
& \quad\langle\{\operatorname{want}(\alpha, \operatorname{bel}(\beta, \operatorname{want}(\alpha, X)))\}, \\
& \left.\quad\{\operatorname{want}(\alpha, \operatorname{bel}(\beta, \operatorname{want}(\alpha, X)))\}, \theta_{\text {instruct }}\right\rangle \tag{1}
\end{align*}
$$

represents a literal directive from $\alpha$ to $\beta$. In the case of no politeness softeners, $M=\emptyset$, where in the case of softeners,

[^340]

Figure 2: Ratings of social context dimensions from behavioral data. Error bars represent SEM.
$M=\{$ please $\}$. In contrast, an indirect request can be represented by:

$$
\begin{align*}
& \text { AskYN }(\alpha, \beta, \operatorname{capableO} f(\beta, X), M):= \\
& \quad\langle\{\text { want }(\alpha, \text { informif }(\beta, \alpha, \operatorname{capableOf}(\beta, X)))\}, \\
& \left.\quad\{\operatorname{want}(\alpha, \operatorname{bel}(\beta, \text { want }(\alpha, X)))\}, \theta_{\text {AskYN }}\right\rangle \tag{2}
\end{align*}
$$

which represents the query "Can you $X$ ?" It is literally a query about one's capability, but can be interpreted as an indirect request. In the case of no politeness softeners, $M=\emptyset$, where in the case of softeners, $M=\{$ please $\}$. The relative face-threat values for each strategy are: $\theta_{\text {AskYN-p }}<\theta_{\text {AskYN }}<$ $\theta_{\text {instruct }-p}<\theta_{\text {instruct }}$, where " p " indicates the presence of politeness softeners.

Table 1 contains the utterance forms selected by the voting algorithm given the relative weights of the communicative goals of directness, politeness, and brevity. Correctness and informativeness were weighted above these criteria, but for the purposes of this scenario were irrelevant (as all candidate utterances were equally correct and informative). Our framework allows for socially-appropriate directive generation, as the various directive strategies, including: Direct - "Plug me in", Direct with softener - "Plug me in, please", Indirect "Could you plug me in?", and Indirect with softener - "Could you plug me in please?" were generated in different potential contexts. For example, if directness is the top priority (e.g., in a task-oriented environment) then a direct utterance will be selected. However, if politeness is required (e.g., in casual conversation or a service-robot scenario) then a more indirect utterance will be selected. The results of the demonstration show how our framework can be integrated in a dialogue system in order to produce robust socially-sensitive natural language utterances in a variety of contexts.

## Setting the Pragmatic Criteria Weightings

Next, we conducted an empirical investigation to establish an initial set of weights for the model (see 'Pragmatic Criteria Weightings' component in Figure 1) that is consistent


Figure 3: Ratings of pragmatic criteria from behavioral data. Error bars represent SEM.
with human judgments. To this end. we conducted a crowd sourcing study on Amazon Mechanical Turk in which people were shown hypothetical human-robot interactions and asked to rate various features of the interactions. A total of 42 people participated in the study - 23 of the participants were male, 17 were female, and 2 did not specify a gender. The average age was 35.9. All participants had US zip codes and received $\$ 1$ for their participation. The study was approved by the Tufts Institutional Review Board and all participants gave informed consent. In the study, participants were shown a text description of four scenarios ${ }^{4}$ and were asked to rate various social context dimensions (potential for harm, time pressure, interlocuter authority, and formality) and pragmatic criteria (directness, politeness, brevity) associated with the robot's speech in each scenario on a sliding scale from 0 (Strongly Disagree) to 100 (Strongly Agree).

Analyses of the data were carried out in order to establish a mapping between the pragmatic criteria, weightings, and utterance selection. First, the results for social context dimensions (see Figure 2) showed that each scenario had a distinct feature profile. Consequently, people expected the robot to adopt a different set of pragmatic criteria in each scenario (see Figure 3). The link between these contextual dimensions and the corresponding pragmatic criteria is important for determining the model weights in new contexts, but this will require future investigations that address the problem directly (see Discussion section). For the present work, we focus on using people's ratings for the pragmatic criteria to set the initial weights of our model. In order to rank these weights, we conducted a repeated measures ANOVA (with Bonferroni

[^341]Table 2: Candidate utterance types with corresponding directives from Scenario \#2

| Utterance Type | Robot Directive |
| :--- | :--- |
| $\left(u_{1}\right)$ Direct | "Plug me in." |
| $\left(u_{2}\right)$ Direct with softener | "Plug me in, please" |
| $\left(u_{3}\right)$ Indirect statement | "I would like you to plug me in. |
| $\left(u_{4}\right)$ Indirect statement with softener | "I would like you to plug me in, please." |
| $\left(u_{5}\right)$ Indirect question | "Could you plug me in?" |
| $\left(u_{6}\right)$ Indirect question with softener | "Could you plug me in, please?"" |

correction) to tease out the ordering of the pragmatic criteria for each scenario. In scenario $1(F(2,82)=18.237, p<$ .001 ), post-hoc tests revealed that people expected the robot to be more direct $(89 \%)$ vs polite $(71 \%, p<.005)$ and brief ( $62 \%, p<.005$ ). There was no significant difference between politeness and brevity in this scenario ( $p=.309$ ). This corresponds to criteria weightings of Direct $>$ Polite $=$ Brief, which would result in a tie in the selected utterance: "Hand me the red pills"/"Hand me the red pills, please" (see Table $1)$. In scenario $2(F(2,82)=4.470, p<.05)$, post-hoc tests revealed that people expected the robot to be slightly more direct ( $87 \%$ ) vs polite ( $74 \%, p<.05$ ). However, there was no significant difference between directness and brevity in this scenario ( $p=.092$ ) or between politeness and brevity ( $p=.673$ ). This corresponds to criteria weightings of Direct = Polite $=$ Brief, and a tie in the selected utterance: "Plug me in"/"Plug me in, please". In scenario $3(F(2,82)=44.334$, $p<.001$ ), post-hoc tests revealed that people expected the robot to be more polite ( $92 \%$ ) vs direct ( $58 \%, p<.001$ ) and brief $(56 \%, p<.001)$. There was no significant difference between directness and brevity in this scenario ( $p=1.00$ ). This corresponds to criteria weightings of Polite $>$ Direct $=$ Brief, and a selected utterance of "Could you hand me your coat, please". Finally, in scenario $4(F(2,82)=32.004, p<$ $.001)$, post-hoc tests revealed that people expected the robot to be more direct ( $85 \%$ ) vs polite $(42 \%, p<.001)$ and brief ( $77 \%, p<.005$ ). People also expected the robot to be more brief vs polite ( $p<.001$ ). This corresponds to criteria weightings of Direct $>$ Brief $>$ Polite, and a selected utterance of "Move out of the way". The utterance output corresponding to each of these criteria weightings is listed in Table 1, and was selected from a list of 6 possible utterance types (see Table 2). Overall, these empirical results serve as a starting point by which to set the weights of our model for sociallyappropriate utterance selection. Extensions of this approach as well as suggestions for future work are discussed in the Discussion below.

## Discussion

In the previous section, we demonstrated how the application of our novel, pragmatically-sensitive framework can result in richer, more human-like modulation of NL. The method of explicit operationalization of pragmatic and socio-linguistic criteria into functions that can produce preference orderings over candidate NLG representations holds advantages over many of the pre-existing approaches. For example, the merg-
ing of preference orders produced by utility functions rather than the direct merging of utility values avoids tricky questions about the direct quantitative comparisons of different pragmatic and socio-linguistic criteria ${ }^{5}$. Additionally, the explicit operationalization of criteria allows for more extensibility and flexibility compared to algorithms in which communicative criteria are factored in implicitly. Nonetheless, this extensibility and flexibility leads to a variety of challenges for future work.

## Computing and Learning Criteria Weights

While we used an empirical approach to initially set the weights for utterance selection, there still exists the normative challenge of determining what the most appropriate orderings/weightings of pragmatic and social goals are in any given communicative context. We allude to possible sources of information that could be used to compute these weights in Figure 1. These include the current beliefs of the agent about the situational context, the agent's goals (task-goals and social-goals), and potentially even models of personality (Mairesse \& Walker, 2011) or culture (Endrass \& André, 2014) that a designer may wish to imbue in the agent (e.g., a social robot configured to be impolite for entertainment purposes). The dynamics of how weights change within a single interaction and context are also a matter for future investigation. For example, a robot could become more polite if it detects that its interlocutor is distressed. The appropriate solution for this component would be entirely dependent on the particular interaction purpose, context, and desired effect. We view the present work as the first necessary step to opening up this rich area of future research.

We envision the process of computing criteria weights as a two-step process. First, various observable or inferable social context factors are evaluated in the given interaction scenario. These contextual features may include factors such as those in Figure 2. These in turn govern the weights that modulate utterance selection. The mapping between social context features and communicative criteria weights could potentially be learned in the following ways. Explicit feedback: the human interactant could provide explicit negative or positive feedback about the agent's recently-produced utterance with respect to a particular communicative criterion (e.g., "That was rude!" would indicate that weights for politeness should be increased in the present context). More subtle cues from facial expression, body language, or affect could also be used to modulate politeness, as in Moore et al. (2004). Passive observation: in a given interaction context, the agent could observe the utterances generated by other agents. An assumption of appropriateness could be made, in which case hypotheses for the possible criteria weights that the agent utilized in the present scenario could be abduced. These hypotheses can be used by the agent itself as constraints that in turn govern its own utterance selection in similar social contexts.

[^342]
## Improved Operationalization of Criteria

Because our proposed framework relies on explicit operationalization of communicative criteria in order to rank candidate utterances, adapting and refining these operationalizations to new criteria, semantic representations, and NLG architectures will be an ongoing task. Adaptation will likely be fairly straightforward for criteria such as correctness, but other pragmatic and socio-linguistic criteria are more complex and leave room for future work. In particular, within DIARC the operationalizations of politeness and brevity can be improved and expanded. As alluded to earlier, brevity will require architectural integration with the lower-level NLG components such as the surface realizer and text-to-speech in order to calculate metrics for lexical and auditory brevity. This will be especially important when the spoken tempo of utterances can be manipulated (one can imagine a speed vs. intelligibility trade-off). Politeness is another criterion ripe for refinement. For example, though we modeled a scenario in which positive face (agent standing) was potentially threatened, a general framework to detect and evaluate threats to positive face is still needed (Briggs \& Scheutz, 2014).

## Conclusion

It is important that socially-embedded artificial agents generate speech in human-like ways in order for interaction with such agents to be truly natural. To this end, we have introduced and demonstrated a general method for modulating utterance selection based on an arbitrary number of social and pragmatic criteria. Our approach possesses an important set of novel features, including extensibility to additional sociolinguistic criteria, adaptability to changing situational context, and agnosticism with respect to underlying semantic representations. In a proof of concept demonstration, we showed how our approach can be integrated with a cognitive robotic architecture in order to generate flexible, socially-appropriate directives in a variety of contexts. Future work will be needed to extend the operationalization of the communicative criteria and devise mechanisms to learn the weights of the model through natural interaction. Overall, the present work moves us a step closer towards the goal of artificial agents that can communicate in the kinds of robust and socially-sensitive ways found in human language.

## References

Bohnet, B., \& Dale, R. (2005). Viewing referring expression generation as search. In Proceedings of the 19th International Joint Conference on AI (pp. 1004-1009).
Briggs, G., \& Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In Proceedings of the 27th AAAI Conference on Artificial Intelligence (pp. 1213-1219).
Briggs, G., \& Scheutz, M. (2014). Modeling blame to avoid positive face threats in natural language generation. In Proceedings of the 8th International Conference on Natural Language Generation (pp. 1-5). Philadelphia, PA.

Brown, P., \& Levinson, S. C. (1987). Politeness: Some universals in language usage (Vol. 4). Cambridge University Press.
Dale, R., \& Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science, 19, 233-263.
Endrass, B., \& André, E. (2014). Integration of cultural factors into the behavioural models of virtual characters. Natural Language Generation in Interactive Systems, 227-251.
Goodman, N. D., \& Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. Topics in Cognitive Science, 5, 173-184.
Grice, P. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and Semantics, 3: Speech Acts.
Gupta, S., Walker, M. A., \& Romano, D. M. (2007). How rude are you?: Evaluating politeness and affect in interaction. In Affective Computing and Intelligent Interaction (pp. 203-217). Springer.
Krahmer, E., \& Van Deemter, K. (2012). Computational generation of referring expressions: A survey. Computational Linguistics, 38, 173-218.
Mairesse, F., \& Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. Computational Linguistics, 37, 455-488.
Miller, C., Wu, P., \& Funk, H. (2008). A computational approach to etiquette: Operationalizing brown and levinson's politeness model. Intelligent Systems, 23, 28-35.
Moore, J. D., Porayska-Pomsta, K., Varges, S., \& Zinn, C. (2004). Generating tutorial feedback with affect. In Proceedings of the FLAIRS Conference (pp. 923-928).
Nye, B. D., Graesser, A. C., \& Hu, X. (2014). Autotutor and family: A review of 17 years of natural language tutoring. International Journal of Artificial Intelligence in Education, 24, 427-469.
Schermerhorn, P. W., Kramer, J. F., Middendorff, C., \& Scheutz, M. (2006). DIARC: A testbed for natural humanrobot interaction. In Proceedings of the AAAI Mobile Robot Workshop (pp. 1972-1973).
Schulze, M. (2011). A new monotonic, cloneindependent, reversal symmetric, and condorcet-consistent single-winner election method. Social Choice and Welfare, 36, 267-303.
Srinivasan, V., \& Takayama, L. (2016). Help me please: Robot politeness strategies for soliciting help from humans. In Proceedings of the 2016 Conference on Human Factors in Computing Systems (pp. 4945-4955).
Strait, M., Canning, C., \& Scheutz, M. (2014). Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In Proceedings of the 9th International Conference on Human-Robot Interaction (pp. 479-486).
Torrey, C., Fussell, S. R., \& Kiesler, S. (2013). How a robot should give advice. In Proceedings of the 8th International Conference on Human-Robot Interaction (pp. 275-282).

# A Two-Stage Model of Solving Arithmetic Problems 

Vladimir V. Glebkin (gleb1514@gmail.com)<br>Russian Presidential Academy of National Economy and Public Administration, Prospect Vernadskogo, 82, Moscow, Russian Federation 119571;<br>Gymnasium 1514,<br>Krupskoi Street, 12, Moscow, Russian Federation 119311

Alexander E. Kovtunenko (sanyka390@yandex.ru),<br>MGIMO University,<br>76, Prospect Vernadskogo, Moscow, Russia, 119454

Ekaterina A. Krysova (katerina_a_a@mail.ru), UWC Red Cross Nordic,<br>Flekke 6968 Norway


#### Abstract

This paper examines a process of solving different types of counterfactual arithmetic problems (problems contradicted a visual experience, an experience of temperature, encyclopedic knowledge, etc.) in comparison with their 'real' counterparts by different types of subjects (e.g., educated in math and educated in humanities). As a result, a two-stage model of solving arithmetic problems is outlined in the paper.


Keywords: situated cognition; counterfactual reasoning; arithmetical problem; two-stage model; four-level-cognitive-development theory.

## Introduction

This research has been carried out at the junction of two sets of problems. The first one is concerned with counterfactual reasoning ${ }^{1}$. This issue has been discussed intensively in recent decades (e.g., Pearl 2000; Fauconnier \& Turner 2002, p. 17-59; Hiddleston 2005; de Vega et al. 2007; de Vega 2008; Ferguson \& Sanford 2008; de Vega \& Uritta 2011; Rips \& Edwards 2013), but some aspects thereof have not been touched so far. 'Situated cognition' is a blanket term for the second set (e.g., Clancey 1997; Kirshner \& Whitson 1997; Watson \& Winbourne 2007; Robbins \& Aydede 2009). This paper examines the role of situated cognition in counterfactual reasoning. A distinguished work of A. Luria (1976) was a starting point for that. When investigating cognitive skills of dekchans of Central Asia in $1930^{\text {th }}$ he encountered a curious phenomenon. His subjects were not able to solve counterfactual problems, whereas they solved quite easily similar problems consistent with their everyday life. Importantly, trying to solve counterfactual problems subjects addressed their day-to-day experience. Some striking examples thereof are given in Luria's monograph (1976, p. 131):

[^343]
## A 'conditional' problem that conflicts with actual experience is given:

[Exp.] Suppose it were to take six hours to get from here to Fergana on foot and a bicycle was twice as slow?
[Sub.] Then a bicycle would get there in three hours!
Solution on a level corresponding to practical reality.
[Exp.] No, a teacher gave this problem as an exercise - suppose that the bicycle were twice as slow.
[Sub.] If the cyclist makes good time, he will get to Fergana in two and a half or three hours. According to your problem, though, if the bicycle brakes down on the way, he'll arrive later, of course. If there's a breakdown, he'll be two or three hours late.
From the perspective of cultural psychology (Vygotsky, Luria, Cole, Tulviste, etc.), this and similar facts are usually interpreted as a difference between 'situational' thinking and 'abstract' thinking. The subjects of Luria's investigation were people of so-called 'sympractical' culture, in which 'situational' thinking based on day-today experience is supposed to be the only way of reasoning (Cole \& Scribner 1974, p. 160-168, 178-179; Luria 1976; Tulviste 1991). In that, Luria's dekchans endeavored to reinterpret abstract problems as stories from their everyday life.

Indeed, people of modern industrial cultures can solve counterfactual problems; meanwhile, situated cognition is an important part of their everyday life, and, therefore, it may influence their way and speed of solving counterfactual problems. The overall objective of our research was to test some obstacles they may encounter in this process.

It is worth noting that there are two basic models of counterfactual reasoning in contemporary cognitive science (Pearl 2000; Hiddlesston 2005; Rips \& Edwards 2013). The first model is called 'pruning theory'. From this perspective, when modeling counterfactual situation, subject changes the only element (in particular, in famed If Clinton were the Titanic, the iceberg would sink

Clinton replaces iceberg), all others being the same. The second model is named 'minimal network theory'. In its scope, a change of one element entails a number of changes in elements close to this one. Importantly, both models look formal and do not distinguish between situated cognition and abstract knowledge.

A preliminary hypothesis of this research was as follows. Given some discrepancies between situated cognition and a counterfactual situation in arithmetic problems, people of modern industrial cultures face a number of difficulties when solving counterfactuals. These difficulties would engender an extended period of time needed for solving counterfactual problem in comparison with 'real' one and also more errors in that. Perhaps, the most intriguing issue in this scope is a particular way of how situated cognition is involved in the process of solving. There are two basic options. From the first perspective, situated cognition is actual for the whole period of solving; from the second perspective, it is at work only in the first stage, in which an abstract model of the task is built, whereas in the second stage only formal operations are processed. If the first option is true, difficulties caused by a counterfactual situation can be drawn forth at any moment of reasoning; if the second option works, they are present early in stage, and then there is no difference, other things being equal, in solving counterfactual and 'real' problems.

One of the ways to test these options is to compare mean $\Delta \mathbf{t}_{\text {cr }}$ (the difference between the time needed to solve a counterfactual problem and the time needed to solve its 'real' equivalent) for people who solve problems faster (aka 'experts') and slower (aka 'amateurs'). 'Experts' superiority over 'amateurs' may be a result of more developed computational skills (factor $a$ ), of higher speed of transformation of a task into a system of abstract symbols (factor $b$ ), and of a combination of these factors. If mean $\Delta \mathbf{t}_{\mathbf{c r}}$ for 'experts' is less than for 'amateurs', the factor $b$ is important; if the difference between mean $\Delta \mathbf{t}_{\mathrm{cr}}$ is not significant, the factor $a$ dominates.

The lack of difference between mean $\Delta \mathbf{t}_{\mathrm{cr}}$ can be also an argument for the two-stage model with the following interpretation: in the first stage there is no difference between 'experts' and 'amateurs', the process in this stage is determined by situated cognition; in the second stage situated cognition is out of work, subjects only perform computational operations with formal objects. Importantly, it does not exactly mean that perception is also out of work in the second stage; it may also mean that in this stage perception works in quite a specific way, distant from day-to-day experience.

Another intriguing issue is a correlation between mean $\Delta \mathbf{t}_{\mathrm{cr}}$ and a type of problem. Counterfactual situation can be concerned with various perceptive channels (vision, hearing, taste, etc.) as well as with some theoretical knowledge. To find out which channel causes maximum $\Delta \mathbf{t}_{\mathrm{cr}}$ is useful to clarify the structure of basic elements of situated cognition. The priority of perceptual channels over encyclopedic knowledge was our working hypothesis in this case (e.g., Zacks 2015, p. 95-107).

The framework of the experiments to carry out was determined by a problem field represented above.

## Experiment 1

## Method

Subjects. A total of 25 students of Moscow high schools, $15-16$-year-old were the subjects of Experiment 1. We argued for this choice by fewer discrepancies in solving arithmetical problems for students than it would be for adult participants.
Material. The subjects were suggested to solve 16 simple arithmetical problems divided into 8 groups: a) problems which contradict a social experience, and their 'real' counterparts (e.g., A 24-page notebook costs 20 roubles more than a 96-page notebook. What is the price of the 24-page notebook if the price of 96-page notebook is 10 roubles?, and A 96-page notebook costs 20 roubles more than a 24-page notebook. What is the price of 96-page notebook if the price of 24-page notebook is 10 roubles?); b) problems which contradict a visual experience, and their 'real' counterparts (e.g., A cyclist moves 10 times faster than a car driver. Please, work out the speed of the car if the speed of the cyclist is $80 \mathrm{~km} / \mathrm{h}$; and A cyclist moves 10 times slower than a car driver. Please, work out the speed of the car if the speed of the cyclist is $8 \mathrm{~km} / \mathrm{h}$ ); c) problems which contradict hearing, and their 'real' counterparts (e.g., In normal conditions, a shout covers a distance of 10 metres; this is 40 metres less than the distance covered by a whisper. Please work out the distance that a whisper covers, and In normal conditions, a shout covers a distance of 50 metres; this is 40 metres more than the distance covered by a whisper Please work out the distance that a whisper covers); d) problems which contradict an experience of temperature, and their 'real' counterparts; e) problems which contradict an experience of taste, and their 'real' counterparts; f) problems which contradict laws of biology, and their 'real' counterparts (e.g., A father is 20 years younger than his son. How old is the son if the father is now 17 years old?; and A mother is 20 years older than her daughter. How old is the daughter if the mother is now 37 years old?); g) problems which contradict encyclopedic knowledge, and their 'real' counterparts; h) problems which contradict an experience of weight, and their 'real' counterparts. Each problem was printed on a special card.
Procedure. Each subject worked out the problems individually. Each subject solved firstly 16 counterfactual problems, randomly given to him/her (2 from each group), and then, a month after, 16 their 'real' counterparts. We endeavored, as seen, to minimize any text difference in each pair of problems in order to avoid 'noise interference'. Before the main procedure the subjects solved two control problems to make sure that they had no difficulties in that. Time from receiving a card to reporting the answer was measured for each problem.

The data was processed as follows. Firstly, $\Delta \mathbf{t}_{\mathrm{cr}}$ was counted for each pair of problems. Then, a total of positive and negative $\Delta \mathbf{t}_{\mathbf{c r}}$ was calculated, and the significance of the difference between positive and negative $\Delta \mathbf{t}_{\mathrm{cr}}$ alongside with the significance of the difference between a total of 'counterfactual' errors and that of 'real' errors was estimated with Pearson's chisquared test. After that, for more detailed analysis
comparative data for all pairs of groups of problems were processed with one-way ANOVA.

## Results \& Discussion

The results of Experiment 1 provided strong evidence for supporting the hypothesis that subjects will encounter more obstacles when solving counterfactual problems than real ones: only $5 \Delta \mathbf{t}_{\mathrm{cr}}$ from 400 had a negative value ( $\chi^{2}(1)=380.25 ; p<0.0001$ ).

Some complementary evidence for supporting general hypothesis was also provided by the analysis of the errors: a total of 34 for counterfactual problems; a total of 10 for 'real' ones $\left(\chi^{2}(1)=380.25 ; p<0.01\right)$.

Mean $\Delta \mathbf{t}_{\mathbf{c r}}$ and standard deviation for each group of problems are presented in Table 1.

Table 1: Mean $\Delta \mathbf{t t}_{\mathbf{c r}}$ for the groups of problems (sec.)

| a) | b) | c) | d) |
| :---: | :---: | :---: | :---: |
| $8.12 \pm 5.89$ | $11.95 \pm 15.66$ | $10.62 \pm 6.04$ | $8.32 \pm 10.93$ |
| e) | f) | g) | h) |
| $10.30 \pm 12.62$ | $6.30 \pm 2.95$ | $6.15 \pm 5.25$ | $8.39 \pm 12.70$ |

The first remarkable result in this table is a gross standard deviation engendered by a big dispersion of results for different participants. The minimal standard deviation holds for the groups f), g), a), and c). This may be connected with a computational complexity of a particular problem: in particular, other things being equal, multiplication leads to higher dispersion than addition; multiplication by eight - to higher dispersion than multiplication by two, etc.

Because of high dispersion, there is no significant difference in $\Delta \mathbf{t}_{\mathrm{cr}}$ for almost all pairs. The only exception is pairs (c, f) and (c, g) (p<0.01). However, this information is also useful as an argument for the hypothesis of the perceptual channels priority over encyclopedic knowledge: problems contradictory with hearing need more time to comprehend the task than problems contradictory with biological laws and encyclopedic knowledge.

By and large, the results of Experiment 1 gave clear evidence to support the basic hypothesis of more difficulties in solving counterfactuals than in solving their 'real' counterparts. Meanwhile, they raised a number of significant questions for further research. Firstly, Experiment 1 gave no evidence pro or contra the two-stage-model. The way of how a counterfactual situation matters the process of reasoning needed a fine-grained analysis. Secondly, the hypothesis of the perceptual channels priority over encyclopedic knowledge required a more detailed investigation.

The framework of Experiment 2 was determined by these issues. To examine the two-stage-model by engaging two groups of subjects with different skills of solving arithmetical problems was its objective. In that, students specialized in mathematics (SM) and their peers specialized in humanities (SH) were chosen to participate Experiment 2. As mentioned above, SM were expected to solve both 'real' and counterfactual problems faster than SH. Given that, the comparison of mean $\Delta \mathbf{t}_{\mathrm{cr}}$ for SM and

SH was under discussion. If significant difference between them could be interpreted in different ways and demanded further investigations to draw a more minute description, then the lack of such difference would testify for the two-stage model.

Let us take a more detailed look at this. The influence of a counterfactual situation is obviously concerned with situated cognition. The same $\Delta \mathrm{t}_{\mathrm{cr}}$ show that such influence does not depend on computational skills being an invariant, at least, for a particular age. Then, if situated cognition is actual for the whole process of working out a problem, its equal influence on SM and SH will be represented in equal $\boldsymbol{\varepsilon}$ (the ratio $\Delta \mathbf{t}_{\mathrm{cr}}: \mathbf{t}_{\text {real }}$ ), but not $\Delta \mathbf{t}_{\mathrm{cr}}$. Equal $\Delta \mathbf{t}_{\mathrm{cr}}$ is an argument for the two-stage model.

## Experiment 2

## Method

Subjects. A total of 40 students of Moscow high schools, 15-16-year-old - 20 SM students and 20 SH students were the subjects of Experiment 2. None of them participated Experiment 1.
Material. The subjects were suggested to solve ten arithmetical problems: five problems contradicting a visual experience (type b) and five problems contradicting encyclopedic knowledge (type g). The problems suggested were the same as the problems of this type in Experiment 1.
Procedure. It was the same as that of Experiment 1.
The data was processed in a following way. Firstly, similar to Experiment 1, Pearson's chi-squared test was applied to estimate the significance of the difference between positive and negative $\Delta \mathbf{t}_{\text {cr }}$ alongside with the significance of the difference between a total of 'counterfactual' errors and that of 'real' errors. After that, a correlation between a group of subjects (SM and SH ) and $\mathbf{t}_{\text {real }}, \mathbf{t}_{\mathbf{c}}, \Delta \mathbf{t}_{\mathrm{cr}}$ was checked with one-way ANOVA Finally, one-way ANOVA was used to check a correlation between $\Delta \mathrm{t}_{\mathrm{cr}}$ and the type of problem (type b vs. type g).

## Results

In accordance with Experiment 1, both SM and SH needed more time to solve counterfactual problems in comparison with their 'real' counterparts ( $\mathrm{p}<0.0001$ )

As predicted, SM solved both 'real' and counterfactual problems faster than SH ( $\mathrm{p}<0.001$ ).

The difference between $\Delta \mathbf{t}_{\text {crSm }}$ and $\Delta \mathbf{t}_{\text {crSH }}$ was not significant. In Table 2 mean $\Delta \mathbf{t}_{\text {crSm }}, \Delta \mathbf{t}_{\text {crSH }}$ and standard deviation are presented for each problem (problems 1-5 are concerned with visual experience, problems 6-10 with encyclopedic knowledge).

Table 2: Mean $\Delta \mathbf{t}_{\mathbf{c r S M}}, \Delta \mathbf{t}_{\mathbf{c r S H}}, \mathbf{P}$ for each problem

| № | $\Delta \mathbf{t}_{\text {crsm }}$, sec. | $\Delta \mathbf{t}_{\text {crSH, }}$, sec. | $\mathbf{P}$ |
| :---: | :---: | :---: | :---: |
| 1 | $5.9 \pm 5.9$ | $11.2 \pm 10.0$ | $\mathbf{0 , 0 3 7}$ |
| 2 | $4.5 \pm 4.0$ | $4.2 \pm 9.6$ | 0,883 |
| 3 | $3.8 \pm 4.0$ | $5.8 \pm 7.5$ | 0,521 |
| 4 | $7.8 \pm 6.2$ | $3.5 \pm 3.2$ | $\mathbf{0 , 0 1 0}$ |
| 5 | $5.2 \pm 5.1$ | $8.1 \pm 7.2$ | 0,123 |


| 6 | $4.6 \pm 3.3$ | $7.0 \pm 7.3$ | 0,197 |
| :---: | :---: | :---: | :---: |
| 7 | $4.7 \pm 3.6$ | $4.6 \pm 7.5$ | 0,970 |
| 8 | $4.6 \pm 2.5$ | $5.4 \pm 2.6$ | 0,319 |
| 9 | $4.8 \pm 4.8$ | $4.5 \pm 7.3$ | 0,782 |
| 10 | $4.5 \pm 3.7$ | $4.6 \pm 4.1$ | 0,951 |

As seen, only the results for problem 1 and problem 4 are more or less significant; at that, for problem 1 $\Delta \mathbf{t}_{\text {crSM }}<\Delta \mathrm{t}_{\text {crSH }}$ and for problem $4 \Delta \mathrm{t}_{\text {crSM }}>\Delta \mathrm{t}_{\text {crSH }}$.

Mean $\Delta \mathbf{t}_{\mathbf{c r}}$ for problems contradicting a visual experience (№ 1-5) is more than that for problems contradicting encyclopedic knowledge (№ 6-10)( $\Delta \mathbf{t}_{\text {cr1- }}$ ${ }_{5}=6.24 \pm 7.07 ; \Delta \mathrm{t}_{\text {cr6-10 }}=4.91 \pm 5.42 ; \mathrm{p}=0.046$ ).

## Discussion

The results of Experiment 2 provide some evidence to support the two-stage-model. A higher level of computational skills entailing a higher speed to work out a problem does not lead to less $\Delta \mathbf{t}_{\mathrm{cr}}$. As noticed above, constant $\Delta \mathbf{t}_{\text {cr }}$ is evidence of the same - at least, for the groups of subjects involved in the experiment - stage of the process. This stage is likely to be connected with the constructing a formal model of the problem, put another way, with the transforming a particular situation into the system of abstract symbols. Situated cognition dominates in this stage, whereas the next stage is concerned with computational operations with such system.

The results of Experiment 2 also support the hypothesis of the priority of perceptual channels over encyclopedic knowledge. The subjects face more difficulties in the situation contradicting their visual experience than in the situation contradicting their encyclopedic knowledge. These data are consistent with some observations in different fields, e.g., with the decisive role of perception in categorization (this idea is represented by the concept of basic level category; see, e.g., Rosch 1978; Lakoff 1987).

As mentioned, subjects of Experiment 1 and Experiment 2 were 15-16-year-old students. Such a choice was determined by a higher level of homogeneity in computational skills for that group in comparison with adults. Nevertheless, in order to verify the results on another age group, Experiment 3 was carried out.

## Experiment 3

## Method

Subjects. A total of 20 high-educated adults (age 35-60; mean age -48 ), half with education in math and physics (EM), and half with education in humanities (EH) were subjects of Experiment 3.
Material and procedure coincided with that of Experiment 2.

## Results and discussion

As it was hypothesized, dispersion for adults was much more significant than for students, because of notable
difference in practice. Nevertheless, the main results of Experiment 1 and Experiment 2 were confirmed.

Subjects solved counterfactual problems longer than 'real' ones $\left(\chi^{2}(1)=26.27 ; p<0.0001\right.$ for $E M ; \chi^{2}(1)=7.19$; $\mathrm{p}<0.01$ for EH ).

Although EM subjects solved the problems of both types faster than EH ones ( $\mathrm{p}<0.005$ ) the difference between $\Delta \mathbf{t}_{\mathbf{S M}}$ and $\Delta \mathbf{t}_{\mathbf{S H}}$ was not significant ( $\mathrm{p}=0.33$ ).

## General discussion

Returning to the issues represented in the introduction it is worth stressing again that the obstacles which Luria's dekchans faced when working out counterfactual problems also characterize people of modern industrial societies in similar situation. These obstacles are not as crucial, however, they lead to longer time needed to solve counterfactual problem in comparison with their 'real' counterparts as well as to more solving errors. These results are consistent with some data from other research fields. Thus, works by Frumkina and colleagues (see Frumkina \& Mikheev 1996 as a summary) gave clear evidence that 'complex thinking', which is, according to Vygotsky, a feature of preschool-age children and people of hunter-gatherer cultures, can also characterize people of modern industrial culture in some situations (e.g., in classification tasks). The only difference is that modern people can change their mind and shift from complex thinking to more abstract cognitive models after some clarifications of an experimenter.

In order to generalize these and similar observations, it is worth addressing the four-level-cognitivedevelopment theory (Glebkin 2015, Glebkin 2015a). This theory singles out four basic cognitive levels, which hold also a framework for cultural-historical typology: Level A characterizes great apes; Level B - prehistoric culture and hunter-gatherer culture; Level C - early theoretical cultures; Level D - Modernity in Europe and modern industrial cultures. Importantly, these levels build on each other, but do not interchange with each other; modern people, guided by circumstances, can operate on all levels. In particular, a majority of everyday skills (swimming, navigation in space, etc.) demand Level A and Level B; Level C is actual for, e.g., working out problems of school geometry; Level D - for abstract algebraic operations. By and large, conceptual systems on Level C operate with objects of natural/social world and their direct representations (historical events, social and political actions, natural objects, etc.). Unlike that, systems on level D operate with abstract objects which have no direct connections with natural/social world (e.g., non-Euclidean geometry, quantum field theory, etc.).

From this perspective the two-stage-model of working out arithmetic problems, developed in this paper, might be interpreted as a shift from Level C, basic in the first stage, to Level D dominating in the second stage. It means that humans can change cognitive levels not only by changing problems but also when solving the same problem.

Finally, it is worth noting that a two-stage model was also suggested by Maruyama et al. (2012) to account for a process of performing nested calculations (e.g. $8+(5-(3$
$+1)$ ) ). Both an analysis of eye-movements of subjects and magneto-encephalography data give some evidence for that. This means that such a model may work not only for arithmetical problems but also for other types of problems in mathematics.

## References

Clancey, W. (1997). Situated cognition: on human knowledge and computer representations. Cambridge; N.Y.: Cambridge University Press.

Cole, M. \& Scribner, S. (1974). Culture and thought. New York: Wiley.
de Vega, M. (2008). Levels of embodied meaning: From pointing to counterfactuals. In M. de Vega, A. Glenberg \& A. Graesser (eds.), Symbols and embodiment: debates on meaning and cognition (pp. 285-308). Oxford; New York: Oxford University Press.
de Vega, M. \& Uritta, M. (2011). Counterfactual sentences activate embodied meaning: An action-sentence compatibility effect study. Journal of Cognitive Psychology, 23 (8), 962-973.
de Vega, M., Uritta, M. \& Riffio, B. (2007). Canceling updating in the comprehension of counterfactuals embedded in narratives. Memory and Cognition, 35, 1410-1421.
Fauconnier, G. \& Turner M. (2002). The Way We Think. Conceptual Blending and the Mind's Hidden Complexities. N. Y.: Basic Books.
Ferguson, H. \& Sanford, A. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. Journal of Memory and Language, 58, 609-626.
Frumkina, R. \& Mikheev, A. (1996). Meaning and categorization. New York: Nova Science.
Gibbs, R. W. Jr. (2006). Embodiment and cognitive science. Cambridge; New York: Cambridge University Press.
Glebkin, V. (2015). The Problem of Cultural-Historical Typology From the Four-Level-Cognitive-Development Theory Perspective. Journal of Cross-Cultural Psychology, 46 (8), 1010-1022.
Glebkin, V. (2015a). A cognitive view on culturalhistorical typology. In G. Airenti, B. G. Bara, G. Sandini (Eds.) Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science. Torino, Italy, September 25-27, 738-743.
Hiddleston, E. (2005). A causal theory of counterfactuals. Nous, 39, 632-657.
Kirshner D. \& Whitson J. A. (eds.) Situated cognition: social, semiotic, and psychological perspectives. Mahwah: L. Erlbaum, 1997.

Lakoff, G. (1987). Women, Fire and Dangerous Things. Chicago and London: The University of Chicago Press.
Landy, D. \& Goldstone, R. L. (2009). How much of symbolic manipulation is just symbol pushing? In N.A. Taatgen \& H. van Rijn (eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society (pp. 1318-1323). Austin, TX: Cognitive Science Society. Luria, A. (1976). Cognitive Development: Its Cultural and Social Foundations. Cambridge, Mass.: Harvard University Press.

Maruyama, M., Pallier, C., Jobert, A., Sigman, M. \& Dehaene, S. (2012).The cortical representation of simple mathematical expressions. Neuroimage, 61, 1444-1460.
Pearl, J. (2000). Causality. Cambridge, UK: Cambridge University Press.
Rips, L. \& Edwards, B. (2013). Inference and Explanation in Counterfactual Reasoning. Cognitive Science, 37, 1107-1135.
Robbins, Ph. \& Aydede, M. (eds.) The Cambridge handbook of situated cognition. Cambridge; N.Y.: Cambridge University Press, 2009.
Roese, N. (1997). Counterfactual thinking. Psychological Bulletin,121, 133-148.
Roese, N. \& Olson, J. (1995). Counterfactual thinking: A critical overview. In What might have been: The social psychology of counterfactual thinking (pp. 1-55). Mahwah: Erlbaum.
Rosch, E. (1978). Principles of Categorization. In Cognition and categorization (pp. 27-48). Hillsdale, N.J.: L. Erlbaum Associates; N.Y.: Halsted Press.

Shapiro, L. (2008). Symbolism, embodied cognition, and the broader debate. Symbols and embodiment: debates on meaning and cognition. In M. de Vega, A. Glenberg \& A. Graesser (eds.), Symbols and embodiment: debates on meaning and cognition (pp. 55-74). Oxford; New York: Oxford University Press.
Tulviste, P. (1991). The cultural-historical development of verbal thinking. Commack, N.Y.: Nova Science Publishers.
Vankov, I. \& Kokinov, B. (2011). Embodied Comparison of Functional Relations. In B. Kokinov, A. KarmiloffSmith \& N. J. Nersessian (eds.), European Perspectives on Cognitive Science. Sofia: New Bulgarian University Press.
Watson, A. \& Winbourne, P. 2007. (eds.). New directions for situated cognition in mathematics education. N.Y.: Springer.
Zacks, J. (2015). Flicker. Your Brain on Movies. Oxford, New York: Oxford University Press.

# Discourse Acquisition in 'Pear Stories' of Preschool-aged Children 

Vladimir V. Glebkin (gleb1514@gmail.com)<br>Russian Presidential Academy of National Economy and Public Administration, Prospect Vernadskogo, 82, Moscow, Russian Federation 119571; Gymnasium 1514, Krupskoi Street, 12, Moscow, Russian Federation 119311

Nikita A. Safronov (nikita1997.08.21@mail.ru),<br>Moscow Region State University, Radio Street, 10, Moscow, Russian Federation 105005

Varvara A. Sonina (varvarasonina@mail.ru), Gymnasium 1514, Krupskoi Street, 12, Moscow, Russian Federation 119311


#### Abstract

This work focuses on an issue situated at the intersection of two domains: the oral mode of communication vs. the written mode of communication, and language acquisition. The backbone of this research is a conjecture that, for some age groups (babies, toddlers and preschool-aged children), to explore the acquisition of discourse as a whole (including gestures, facial expressions, prosody, pauses and discursive markers, etc.) is more appropriate than explore the acquisition of language exclusively. "The Pear Film" experimental line underpins the method of this research. The database comprises 74 'pear stories' of Moscow preschool-aged children and high school students. Three parameters of the discourse are of interest for the authors: a logical structure and a coherence of the narrative; gestures and spontaneous movements lost any communicative meaning; discourse words and pauses.


Keywords: multimodal communication, discourse, language acquisition, narrative, pear stories.

## Introduction

The study of spoken language in contrast with its written form has been one of the most intriguing issues in cognitive science over recent decades. At the beginning of the paper it is worthwhile to outline the key points which underpin this avenue of research (Tannen 1982; Chafe 1985; Chafe \& Tannen 1987; Miller \& Weinert 1998; Holie \& Adger 1998; Li \& Hombert 2002; Linell 2005; Tomasello 2008; Fais et al. 2012).

Since its emergence as an independent branch of science, linguistics by default has been based on the structure of written language as a paradigm for language in general. This has deep roots in the backstory of linguistics. Although for the Ancient Greeks and the Ancient Romans speech and argument were important elements of both politics and everyday life, theoretical approaches to language developed by Plato and, especially, by Aristotle bore on written language. This held true, without any alternatives, in the philosophy of the Middle Ages. Linguistic views of patristics and, then, scholasticism were almost exclusively concerned with literacy. Afterwards, the ancient and medieval tradition strongly influenced early modern scholars in their view of
language (Pascal, Descartes, Leibniz, Wilkins, etc.). As a result, some models of language as a formal system of symbols were developed which, in turn, inspired modern linguists to develop formal theories of language (Chomsky's works include abundant quotations from Descartes (Chomsky 1966, 2006), NSM theory by Wierzbicka and Goddard is, more or less, a replica of Leibniz's language of thought (Wierzbicka 1972, 1980, 1996), etc.). All these theories are based, as a matter of fact, on the framework of written language, even if their authors claim the opposite.

Although the view of language represented in the previous paragraph is likely to be a common place for linguists, this is anything but the truth. Importantly, unlike written language, spoken language is an element of multimodal communication, and it is absolutely senseless to explore spoken language beyond its links with other elements (gestures, facial expressions, prosody, pauses and discursive markers, etc.). Therefore, the only way to account for the framework and functions of spoken language in different communicative situations is to tackle the structure of discourse as a whole.

Such a change of perspective helps, in particular, to shed new light on the question of the origin of language. Thus, Tomasello (e.g., 2008) and colleagues take the social nature of humans as a basis for the research framework. For them, the demands of social nature cause a growth of both intensity and complexity of social communication which, in turn, leads to emergence of more complicated tools to perform that. These tools include gestures, prosody and, then, meaningful combinations of sounds as a part of multimodal communication.

This approach also changes standard views of the problem of language acquisition. Babies, toddlers and preschool-aged children acquire not exclusively language, but rather different types of discourse in which language is an extremely important but not the only element. Language acquisition goes for them in line with gesture acquisition, prosody acquisition, mimicry acquisition, etc.

This provides a theoretical framework of our research, whereas its experimental framework is based on "The Pear Film" research line. It is reasonable to outline
the milestones of "The Pear Film" story before proceeding to the structure and the results of the experiment.
"The Pear Film" is a six-minute movie made by Wallace Chafe and his colleagues in 1975. The film includes actions, pictures and sounds, but no words. In that, it deploys the same chain of events for all viewers. Therefore, a comparison of "the Pear Film" retellings, i.e. 'pear stories', by people of different cultures and languages can provide the researcher with important data of how language and culture influence a way people conceptualize a stream of events. Since 1975 a lot of investigations have been carried out to compare retelling strategies for people of different cultures (Du Bois 1980; Tannen 1980; Orero 2008; Matzur \& Mickievicz 2012; Blackwell 2015), for people with intellectual disability (Cummings 2015, 59-63); some investigations have explored peculiarities of referential choice in retellings (Downing 1980; Clancy 1980), work of consciousness in narration (Chafe 1980; Bernardo 1980), a structure of multimodal discourse (Fon et al. 2011; Kibrik et al. 2015).

As mentioned, a structure of multimodal discourse is also the main object of interest for the authors of this paper. In general, we follow the model developed by Kibrik and colleagues (2015), but we are interested in a process dynamics rather than a static picture.

Let us proceed directly to the research presented in this paper. The 'pear stories' of Moscow preschool-aged children in comparison with similar stories of high school students were in the focus of our interest in this research. In other words, we addressed a particular type of multimodal discourse in order to explore the process of discourse acquisition by focusing on a logical structure and a coherence of the narrative, gestures and spontaneous movements lost any communicative meaning, discourse words and pauses. Before dealing this issue at hand, some clarifications are needed.

Firstly, the discourse of the pear film retellings is not an informal situation for subjects, especially, for kids. Indeed, in natural contexts kids communicate and, in particular, retell stories in different way. At the same time, this situation is not completely unnatural for them. This is a type of a public talk they encounter in kindergartens, at schools and some other public places. They acquire this as they acquire many other types of multimodal communication. So, despite obvious restrictions, this kind of discourse can provide important data on the way of acquiring particular elements of multimodal communication by preschool and early school-age children.

In order to monitor quality of acquiring particular discourse skills, a sample for comparison is needed. Adults are expected to be such sample; however, the work with 14-16 year old students shows that they are as skillful in "The Pear Film" retellings as adults are. In that, because of some practical reasons a group of such students was taken as a control group for our research.

In our analysis we focused on three discourse elements which need more precise description.
A. A logical structure and a coherence of the narrative. There are a lot of works investigating the narrative development in preschool age children which use a wide range of criteria to check this process (Peterson
\& McCabe 1983; Stein \& Albro 1997; Sedov 2004; Nicoladis et al. 2009; Laurent et al. 2015; Levy \& McNeill 2015). A commonly used parameters to check narrative development are as follows: the length of story in words; the number of different words used to tell the story; a total of scenes in retellings; presence of basic semantic components of the story (beginning, setting and ending) (Nicoladis et al. 2009; Laurent et al. 2015). However, these characteristics seem to provide only a coarse-grained picture of the process failing to verify how subjects represent a logical structure of the narrative, i.e. causal links connecting the events within it. A more precise model to evaluate exactly this factor was elaborated in Sedov 2004. In order to check to what extent subjects represent a logical structure of stories they retell, the author examines such variables as the frequency of deictic words in the retellings, the frequency of anaphoric repetitions, the frequency of introductory model words, referential models the subjects apply, the appearance of retrospective and perspective views in the retellings, etc.

Taking into account these and similar works, in our research we focused on the following characteristics: the total number of words exploited in retelling, discounting selfrepetitions and false starts (TW); a total of scenes presented in retelling (TS); a total of errors in action description standardized on 100 words (FA) (e.g., 'guys picked up pears' instead of 'the boy hands pears to one of the guys'); a total of errors in object description standardized on 100 words (FO) (e.g., 'apples' instead of 'pears'); a total of incorrect description of causal chain of events and sub-events standardized on 100 words (FC) (e.g., ambiguous reference, missing connections within an event and between events); a total of interpretations (TI) (e.g., 'stole a basket of pears' instead of 'picked up a basket of pears'); a total of dependent words standardized on 100 words (TD) (such as 'who', 'which', 'because', etc.).
B. Gestures and spontaneous movements lost any communicative meaning. As mentioned, a number of works has been published over recent years to explore various aspects of correlation between gesture development and spoken language development in narratives of preschool age children (Nicoladis et al. 2009; Laurent et al. 2015; Levy \& McNeill 2015). Meanwhile, in our knowledge there are no works focusing on spontaneous movements lost any communicative meaning. Of great importance is the fact that kids, when retelling the story, perform a lot of unconscious movements which do not address their interlocutors. These movements are not gestures in the strict sense; the only function of such movements is to help kids in their reasoning and speaking. This is no metaphor to say that kids not only think with their brains and speak with their tongues, but they also think and speak with their bodies. Some evidence of this can be also found in students' and adults' retellings, but in this case such movements are presented in a restricted mode.

Again, in our knowledge this is a novel research domain which demands, first of all, a correct typology of spontaneous movements. A version of this is suggested below. Another important task is to measure the difference
in spontaneous movements of preschool age kids and that of high school students. This procedure is also presented below in the description of experimental method.
C. Discourse words and pauses. An important aspect of language acquisition is also, so to speak, smoothness of speech. Adults avoid pauses in communication and use different strategies to fill them (stretching out first and last sounds of the word, use of discourse words and single sounds, etc.). At the same time, kids are not embarrassed by gaps in communication. Their speech, at least, in this particular type of discourse, is, as it were, ragged. In the experiment we measured this difference. Also we compared a number of and a mean length of EDU (elementary discourse units) for the two groups of participants.

## Experiment

## Method

Subjects. 50 5-7 year old children ( $22 \mathrm{~m}, 28 \mathrm{f}$ ) attending Moscow kindergartens $(\mathrm{CH})$, and 24 ( $10 \mathrm{~m}, 14 \mathrm{f}$ ) 14-16 year old Moscow high school students (S). All subjects were monolingual.
Material. "The Pear Film" by Wallace Chafe ( 6 min 32 sec ).
Procedure. Each subject was processed individually. At the beginning the subjects were asked for watching the film closely in order to retell it as precisely as they can. Then they watched the film and after a minute retold it to some people who have not seen this film before. For kindergarten kids it was their kindergarten teacher, for students - their peers. At the same time, in order to check how the choice of addressee may influence the results, five students were asked to retell the story to the school principal in her office. When retelling, all subjects were sitting on high chairs for recording not only hand movements but also leg movements.

The retellings were filmed by a hidden camera for CH subjects and overtly for S subjects. Because of some technical problems only 37 from 50 CH video-retellings fitted for further examination (at the same time, all CH tape recordings were made properly).

Then the data were processed with ELAN to examine gestures and spontaneous movements. Also the retellings were recorded with unite-based discourse transcription system.

In order to work out TS we asked 10 independent participants to divide the film into episodes, and premised on their choice singled out eight basic scenes which formed the narrative framework.

## Results

A. The results for narrative skills are presented in Table 1 and Table 2.

Table 1. Mean TW, TS, TI, EC for CH and S

|  | TW | TS | TI | TD |
| :---: | :---: | :---: | :---: | :---: |
| $\mathbf{C H}$ | $88.2 \pm 45.9$ | $5.3 \pm 1.6$ | $1.5 \pm 1.0$ | $1.0 \pm 1.2$ |
| $\mathbf{S}$ | $298.6 \pm 138.2$ | $7.8 \pm 0.7$ | $4.7 \pm 1.9$ | $2.4 \pm 1.1$ |

Table 2. Mean FA, FO, FC for CH and S

|  | FA | FO | FC |
| :---: | :---: | :---: | :---: |
| $\mathbf{C H}$ | $1.9 \pm 2.1$ | $1.5 \pm 2.2$ | $2.0 \pm 2.2$ |
| $\mathbf{S}$ | $0.1 \pm 0.4$ | $0.1 \pm 0.3$ | $0 \pm 0.2$ |

All these results are statistically significant ( $\mathrm{p}<0.001$ ).

Concerning five students retelling the pear story to the school principal, there were no significant difference between them and other students in the scope of this experiment. Retellings in the principal office were likely to be less detailed and more formal, but in comparison with CH subjects these discrepancies were not important.

The quantitative data can be complemented by some qualitative analysis. Let us begin with TW. As this can be seen from Table 1, the mean length of the narrative is over three times more for the students than for the preschoolaged children. It means that the CH subjects lose a lot of content when describing any episode. Importantly, they only point at actions not focusing on appearance, clothes, scenery, etc. In turn, S subjects provided more or less detailed description of the picker, and also briefly characterized appearance of other characters.

FO values are consistent with this observation. In contrast with S subjects, CH subjects confused not only pears and apples, but also, and much more often, they confused age ('man' instead of 'boy') and gender ('boy' and even 'man' instead of 'girl') of the characters.

TS, FA and FC values points to notable difficulties in representing causal chain of the narrative. TS values show that two, in average, basic scenes get lost in 'pear stories' of preschool-aged children. This means that the story is often broken up into independent fragments which are linked with the conjunction 'then'. To be more precise, a majority of CH subjects missed the episode with a goat, which does not 'work' later on in the film (only 8 subjects from 50 remembered this), many of them missed also the girl on a bike approaching the boy, and sometimes the final scene was missed as well. Also, almost in all stories of CH subjects there were some ambiguities in reference because their use of pronouns sometimes did not allow determine a subject correctly.

Data for EC are consistent with that point. S subjects often used dependent words (mainly, 'who', 'which', but also 'when', 'where', 'what', 'why', etc.) in order to clarify reference. If there are some objects with the same nomination (e.g., 'boy'), exploiting of such constructions is useful tool for reference clarification. Meanwhile, CH participants can hardly use this tool.

Let us proceed to exploring some correlations within CH group. First of all, there is a positive correlation between the age of participants (AP) and TW, TI, TD and a negative correlation between the age of participants and FA, FO, FC. However, only the correlation between AP and TW; AP and FO are significant $(\mathrm{r}(\mathrm{AP}, \mathrm{TW})=0.35$, $\mathrm{r}(\mathrm{AP}, \mathrm{FO})=-0.27, \mathrm{p}<0.05 ; \mathrm{r}(\mathrm{AP}, \mathrm{TI})=0.14, \mathrm{r}(\mathrm{AP}, \mathrm{TD})=$ $0.23, \mathrm{r}(\mathrm{AP}, \mathrm{FA})=-0.25, \mathrm{r}(\mathrm{AP}, \mathrm{FC})=-0.13)$.

In the meantime, there is a significant positive correlation between TW and TI, TW and TD, and also a significant negative correlation between TW and FA, TW and FC, TW and FO (r(TW, FA) $=-0.44, r(T W, F C)=-$ $0.46, \mathrm{p}<0.01 ; \mathrm{r}(\mathrm{TW}, \mathrm{TI})=0,34, \mathrm{r}(\mathrm{TW}, \mathrm{TD})=0.33, \mathrm{r}(\mathrm{TW}$, $\mathrm{FO})=-0,29, \mathrm{p}<0.05)$. This gives some evidence for the point that all narrative skills represented by the variables of the subset A develop consistently.

In order to complete the spectrum of problems CH subjects encountered it is worth noting that ten of them $(20 \%)$ were not able to retell the pear film on their own, and the experimenters were made to help them with leading questions.
B. As mentioned, spontaneous movements were in focus of our interests. We singled out three general types of body movements according to body parts which provide them: hand movements (HM), leg movements (LM), torso movements (TM). Also for hands and legs we distinguished implicit movements (e.g., slight finger movements) and explicit movements (e.g., open movements of the whole hand). So, we used five combinations: IHM, EHM, ILM, ELM, TM. We measured the total time of each kind of movement in ratio to the time of the whole story presented in percent. The results are expressed in Table 3.

Table 3. Mean IHM, EHM, ILM, ELM, TM for CH and S

|  | IHM | EHM | ILM |
| :---: | :---: | :---: | :---: |
| $\mathbf{C H}$ | $31.9 \pm 22.2$ | $5.5 \pm 10.1$ | $7.0 \pm 11.6$ |
| $\mathbf{S}$ | $20.0 \pm 21.4$ | $0.2 \pm 0.8$ | $4.6 \pm 9.6$ |
|  | $\mathbf{E L M}$ | $\mathbf{T M}$ | $\mathbf{\Sigma}$ |
| $\mathbf{C H}$ | $6.9 \pm 16.9$ | $15.7 \pm 17.3$ | $65.8 \pm 18.2$ |
| $\mathbf{S}$ | 0 | $0.6 \pm 1.1$ | $24.5 \pm 21.1$ |

As can be seen from Table 3, there is a significant difference in performing spontaneous movements during the retelling between CH and S groups (the use of oneway ANOVA to compare $\boldsymbol{\Sigma}$ for these groups gives $\mathrm{p}<0.001$ ). CH subjects perform such movements during over a half-time period of the retelling, and over a quartertime period of the retelling the movements are explicit. The spectrum of their movements is really wide: kids put their hands under the legs, fidget in their seats, lift their legs up to their mouths, etc. All this almost totally disappear in retellings of $S$ subjects. The only spontaneous movements they perform are implicit hand movements
such as to finger over and, to some extent, implicit leg movements. Other types of movements are extremely rare.

Importantly, there is no significant correlation between $\Sigma$ (the sum of IHM, EHM, ILM, ELM, TM) and TW for CH subjects ( $\mathrm{r}=-0.14$ ). This can be interpreted as some evidence against the conjecture that narrative skills and body experience in discourse develop coherently.

The picture of gestures for CH and S groups is strictly opposite. Only 8 from $50(16 \%) \mathrm{CH}$ subjects exploited gestures as a more or less important tool in communication. At the same time, almost all $S$ subjects resorted to the permanent use of gestures during their retellings. The total time of gesture performance in ratio to the time of the whole story presented in percent is $2.4 \pm 5.2$ for $C H$ subjects and $68.5 \pm 23.7$ for $S$ subjects ( $p<0.001$ ).

Addressing again $S$ subjects who retold the pear story to the school principal, it is worthwhile to note that they performed less gestures and more spontaneous movements than their peers but this difference cannot change the picture drawn above.
C. The last group of parameters we worked out is concerned with discourse words, pauses and an EDU length. We measured the total number of discourse words in ratio to TW (DW, \%), the total length of pauses in ratio to the time of the whole story (LP, \%) ${ }^{1}$, and a mean EDU length (EDUL, words). The data are presented in Table 4.

Table 4. Mean DW, LP, EDUL for CH and S

|  | DW | LP | EDUL |
| :---: | :---: | :---: | :---: |
| $\mathbf{C H}$ | $3.0 \pm 2.0$ | $38.2 \pm 11.2$ | $3.2 \pm 0.5$ |
| $\mathbf{S}$ | $4.1 \pm 3.1$ | $25.3 \pm 10.4$ | $4.2 \pm 0.6$ |

These data give clear evidence that CH subjects in comparison with $S$ ones are less skillful in filling pauses in communication ( $\mathrm{p}<0.001$ ). Also it is worthwhile to point at more extensive EDU for S subjects ( $\mathrm{p}<0.001$ ). The difference between DW for CH subjects and S subjects is not significant.

Importantly, there is significant positive correlation between TW and DW for CH subjects ( r (TW, DW) = $0,32, \mathrm{p}<0.05$ ). Meanwhile, the correlation between TW and LP and between TW and EDUL is not significant ( r $($ TW, LP $)=-0,23 ; r($ TW, EDUL $)=-0,02)$.

## Discussion

The results of the experiment support the basic hypothesis of significant obstacles which preschool-aged children encounter when acquiring the discourse of retelling story in formal situation. In the experiment three basic components of this discourse were examined: a logical structure and a coherence of the narrative; gestures and spontaneous movements lost any communicative meaning; and EDU, discourse words and pauses. In all these components CH subjects experienced more or less serious difficulties in comparison with a control group presented by high school students. Furthermore, the

[^344]experiment provided some evidence that these components are acquired coherently, that is, the progress in one component correlates with positive shifts in others.

It is worth classifying the problems CH subjects face and cognitive skills behind these problems. Two basic domains can be picked out in "The Pear Film" retelling discourse. The first one is concerned with the 'content' of the discourse (to make sense of the story and to present this correctly in speech), the second one - with 'right' mode of communicative behavior. Cognitive skills behind the first domain can be also divided into two parts. The first part is based on situated cognition. A lack of practical knowledge entails increasing FO value which characterizes flaws in recognizing particular objects (as mentioned, kids confused pears with apples, goat with caw, girl with boy, boy with man, etc.) The second set of problems points at an inability to figure out a causal chain of events and to represent this chain in the retelling. In consistency with Sedov's (2004) results our data show that a majority of CH subjects in their retellings get plunged into the stream of events, and they are unable to change the perspective and to look at the story from the bird's eye view. This also determines their view of an addressee. They usually take for granted that an addressee is also familiar with all details of the story and he can easily reconstruct those following restrained comments of a storyteller. TS, TI, FA, and FC values characterize this issue.

CH subjects have also obvious problems with the use of language. "The Pear Film" retelling as a kind of discourse is close to writing, and writing skills are widely used by S subjects in their retellings. More or less consciously, they rest on texts studied at school as a paradigm for 'pear stories' they make up. CH subjects have no such experience. An influence of literacy on their retellings is trifling, if it is at all. As a result, their language is extremely poor, with minimum of extended and subordinate constructions. TC value is responsible for this set of factors.

The 'communicative' domain of the discourse is characterized by values of variables which constitute subsets B and C. In particular, LHM, EHM, LLM, ELM, and TM values point at a flaw in discourse competence connected with some lack of body control in communication.

Although three sets of factors presented above address different cognitive domains and function more or less independently, there is some coherence in their development (e.g., significant positive correlation between TW and DW give some evidence for such coherence).

Also LHM, EHM, LLM, ELM, TM data need more detailed commentary. This sounds nowadays as a common point that body movements in discourse have nothing but communicative function. Nevertheless, this is not so for kids. Again, the behavior of CH subjects gives robust evidence that they not only think with their brains and speak with their tongues, but they also think and speak with their bodies. The spontaneous movements which they perform intensively during the retelling are not directed to their interlocutors, but these movements do rather produce, as it were, 'nutrient medium' for the process of speaking. Their speech is a vivid and striking illustration
of theoretical postulates of embodiment theory (Barsalou 1999; Krois 2007; Barsalou 2010).

Finally, we venture to make a conjecture in this scope. The comparison of communicative models of CH and S subjects indicates a substantial shift from spontaneous movements to gestures. This shift is likely to be consistent with Vygotsky's (1986) theory of inner speech. According to Vygotsky, cognitive development in ontogeny is concerned with the transition from egocentric speech to inner speech which cognitive function is to mediate between speech and thought. From this perspective, inner speech is interpreted as interiorization of egocentric speech. Similarly, implicit spontaneous movements can be treated as interiorization of explicit spontaneous movements. Such interiorization comes to the end in thought which includes words and movements in a 'converted' mode similarly to how synthesis includes thesis and antithesis in Hegel's philosophy. On the opposite stage, thought deploys into a communicative utterance directed to an addressee and formed by words, gestures, prosody and other elements of multimodal communication. In this scope, gestures not just accompany words in communication, but rather they are equally meaningful element of communicative behavior.

Another possible domain to apply the results of our research is the theory of origin of language developed by Tomasello and colleagues. As mentioned, they point at gestures as at an important predecessor of vocal communication, but they do not take into account spontaneous movements. A precise analysis of spontaneous movements of children and great apes might shed new light on this issue.

## References

Barsalou, L. (1999). Perceptual symbol systems. Behavioral and brain sciences, 22, 577-660.
Barsalou, L. (2010). Grounded cognition: Past, Present and Future. Topics in Cognitive Science, 2 (4), 716-724.
Bernardo, R. (1980). Subjecthood and Consciousness. In In Chafe, W. (ed.). The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production (pp. 275300). Norwood, New Jersey: Ablex.

Blackwell, S. (2015). Porque in Spanish Oral Narratives: Semantic Porque, (Meta)Pragmatic Porque or Both? In A. Capone, \& J. Mey (eds.), Interdisciplinary Studies in Pragmatics, Culture and Society (pp. 615-632). New York: Springer Berlin Heidelberg.
Chafe, W. (1980). The Development of Consciousness in the Production of a Narrative. In W. Chafe (ed.). The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production (pp. 9-50). Norwood, New Jersey: Ablex.
Chafe, W. (1985). Linguistic Differences Produced by Differences between Speaking and Writing. In D. Olson, A. Hildyard, \& N. Torrance (eds.), Literacy, Language, and Learning (pp. 105-123). Cambridge: Cambridge University Press.
Chafe, W. \& Tannen, D. (1987). The Relation between Written and Spoken Language. Annual Review of Anthropology, 16, 383-407

Chelliah, S. (2013). Predicting reference form: A Pear Story Study of information status, thematic role and animacy in Meithei (Manipuri, Meiteiron). In T. Thornes, Andvik E., G. Hyslop, \& J. Jansen (eds.), Functionalhistorical approaches to exploration (pp. 223-236). Amsterdam; Philadelphia: John Benjamins Publishing Company.
Chomsky, N. (1966). Cartesian linguistics. New York, Harper \& Row.
Chomsky, N. (2006). Language and mind. Cambridge etc.: Cambridge University press.
Clancy, P. (1980). Referential Choice in English and Japanese Narrative Discourse. In W. Chafe (ed.), The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production (pp. 127-202). Norwood, New Jersey: Ablex.
Cummings, L. (2015). Pragmatic and Discourse Disorders. Cambridge: Cambridge University Press.
Downing, P. (1980). Factors Influencing Lexical Choice in Narrative In W. Chafe (ed.), The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production (pp. 89-126). Norwood, New Jersey: Ablex.
Du Bois, J. (1980). The Search of a Cultural Niche: Showing the Pear Film in a Mayan Community. In W. Chafe (ed.), The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production (pp. 1-8). Norwood, New Jersey: Ablex.
Fais, L., Leibowich, J., Hamadani, L. \& Ohira, L. (2012). Infant movement as a window into language processing. In J.-M. Colletta \& M. Guidetti (eds.), Gesture and Multimodal Development (pp. 99-127). Amsterdam; Philadelphia: John Benjamins Pub. Co.
Fon, J., Johnson, K. \& Chen, S. (2011). Durational Patterning at Syntactic and Discourse Boundaries in Mandarin Spontaneous Speech. Language and Speech, 54, 5-32.
Holie, S. \& Adger, C. (eds.). (1998). Kids talk: strategic language use in later childhood. New York: Oxford University Press.
Kibrik, A., Fedorova, O., \& Nikolaeva, Ju. (2015). Multimodal Discourse: In Search of Units, in G. Airenti, B. Bara \& G. Sandini (eds.), Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science, 4th European Conference on Cognitive Science, 11th International Conference on Cognitive Science, Torino, Italy, September 25-27, 2015 (pp. 662-667). Torino: University of Torino.
Krois, J. (ed.) Embodiment in cognition and culture. (2007). Amsterdam; Philadelphia: John Benjamins Pub. Co.
Laurent, A, Nicoladis, E. \& Marenette, P. (2015). The development of storytelling in two languages with words and gestures. The International Journal of Bilingualism, 19 (1), 56-74.
Levy, E \& McNeill, D. (2015). Narrative development in young children: gesture, imagery, and cohesion. Cambridge, UK: Cambridge University Press.
Li, Ch. \& Hombert, J.-M. (2002). On the evolutionary origin of language. In M. Stamenov \& V. Gallese (eds.), Mirror neurons and the evolution of brain and language (pp. 175-206). Amsterdam; Philadelphia: John Benjamins Pub.

Linell, P. (2005). The written language bias in linguistics: its nature, origins, and transformations. London; New York: Routledge.
Matzur, I. \& Mickievicz, A. (2012). Pear Stories and Audio Description: Language, Perception and Cognition across Cultures. Perspectives, 20 (1), 55-65.
McNeill, D. (2005). Gesture and Thought. Chicago: The University of Chicago Press.
Miller, J. \& Weinert, R. (1998). Spontaneous spoken language: syntax and discourse. Oxford: Clarendon Press; New York: Oxford University Press.
Nicoladis, E., Pika, S. \& Marentette, P. (2009). Do French-English bilingual children gesture more than monolingual children? Journal of Psycholinguistic Research, 38, 573-585.
Orero, P. (2008). Three different receptions of the same film: 'The Pear Stories Project' applied to audio description. European Journal of English Studies, 12 (2), 179-193.
Peterson, C. \& McCabe, A. (1983). Developmental psycholinguistics: three ways of looking at a child's narrative. New York: Plenum Press.
Sedov, K. (2004). Diskurs i lichnost': evolyutsiya kommunikativnoi kompetentsii [Discourse and personality: the evolution of communicative competence]. Moscow: Labirint.
Stein, N. \& Albro, E. (1997). Building complexity and coherence: Children's use of goal-structured knowledge in telling stories. In M. Bamberg (ed.), Narrative development: Six approaches (pp. 5-44). Mahwah, N.J.: L.Erlbaum Associates.

Tannen, D. (1980). A Comparative Analysis of Oral Narrative Strategies: Athenian Greek and American English. In W. Chafe (ed.), The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production (pp. 1-8). Norwood, New Jersey: Ablex.
Tannen, D. (ed.). (1982). Spoken and written language: exploring orality and literacy. Norwood: ABLEX Pub. Corp.
Tomasello, M. (2008). Origins of human communication. Cambridge, Mass.: MIT Press.
Vilaró, A., Duchowski, A, Pilar, O., Grindinger, T., Tetreault, S. \& di Giovanni, E. (2012). How sound is the Pear Tree Story? Testing the effect of varying audio stimuli on visual attention distribution. Perspectives, 20 (1), 55-65.

Vygotsky, L. (1986). Thought and language. Cambridge, Mass.: MIT Press.
Wierzbicka, A. (1972). Semantic Primitives. Linguistische Forschungen. №22. Frankfurt/M: Athenäum.
Wierzbicka, A. (1980). Lingua mentalis: The Semantics of natural Language. Sydney etc.: Academic Press.
Wierzbicka, A. (1996). Semantics: primes and universals. Oxford, N. Y.: Oxford University Press.

# Inductive reasoning influences perception of interspecies disease transmission risk 

Micah B. Goldwater (micah.goldwater@sydney.edu.au)<br>School of Psychology<br>University of Sydney

Molly E. Ireland (molly.ireland@ttu.edu)
Department of Psychological Sciences
Texas Tech University

Nick Gaylord (nlgaylord@ gmail.com)<br>Independent researcher<br>Oakland, CA USA

Jason Van Allen (jason.vanallen@ ttu.edu)<br>Department of Psychological Sciences<br>Texas Tech University

Tyler Davis (tyler.h.davis@ttu.edu)<br>Department of Psychological Sciences Texas Tech University


#### Abstract

Zoonoses (diseases that enter the human population via animal contact) are a major global health concern. Because of how zoonoses emerge, understanding human reasoning about the risk factors associated with animal contact is central to combating their spread. However, little is known about the factors that influence perception of these risks. We present an inductive account of zoonosis risk perception, suggesting that it is influenced by beliefs about the range of animals that are able to transmit diseases to each other. In Study 1, we find that participants who endorse higher likelihoods of cross-species disease transmission have stronger intention to report animal bites. In Study 2, adapting real world descriptions of Ebola virus from the WHO and CDC, we find that communications conveying a broader range of animals as susceptible to a disease increase intentions to report animal bites and decrease perceived safety of wild game meat. These findings suggest that cognitive factors may be harnessed to modulate zoonosis risk perception and combat emerging infectious diseases.


Keywords: Induction; categorization; risk perception; public health; premise number; premise diversity

## Introduction

Emerging infectious diseases are a major economic and public health concern. A majority of such diseases are of zoonotic origin (i.e. come from animals, Jones et al., 2008), with drivers including animal bites, consumption of wild game meat, and contact with livestock (Daszak, Cunningham, \& Hyatt, 2000). Human-animal interaction is central to all these drivers, but little is known about how people reason about potential risks in such scenarios (Janes, Corbett, Jones, \& Trostle, 2012). Similarly, research on cognitive factors is largely absent from public health initiatives targeting zoonoses, including interdisciplinary approaches such as One Health (Heymann \& Dar, 2014). The present work aims to bridge this gap by examining cognitive principles that influence zoonosis risk perception and how they can be harnessed to shape communications regarding disease transmission risk.

The literature on zoonoses lacks extensive research on the role of human reasoning, though several recent studies have examined factors that determine whether people will eat wild game meat (Kamins et al., 2015) and report adverse animal contact, such as bites, to a health professional. One particularly suggestive study (Bingham, Budke, \& Slater, 2010) found that survey respondents were more likely to report dog bites if they knew that bats could transmit rabies to humans. At first glance, this seems surprising - people's inferences about the risk associated with one species appear to be influenced by their knowledge of a completely different species.

The finding that knowlege about one animal can affect beliefs about other animals may be partly accounted for by two principles from the literature on inductive reasoning, namely premise number and premise diversity (Hayes, Heit, \& Swendsen, 2010; Osherson, Smith, Wilkie, Lopez, \& Shafir, 1990). According to the premise number principle, people are more confident in inferences that apply to a large number of category members (Li, Cao, Li, Li, \& Deak, 2009; McDonald, Samuels, \& Rispoli, 1996), where for example a property known to hold for both lions and giraffes will be more likely to hold for rabbits as well. According to the premise diversity principle, people find inferences sound to the extent that they hold for a wider range of category members (Heit \& Feeney, 2005; Lopez, 1995), where for example a property known to hold for lions and giraffes is more likely to generalize to rabbits, compared to one that holds for lions and tigers. In terms of zoonosis risk perception, knowing that both dogs and bats can transmit rabies may increase perceptions of human risk because they are often viewed as very different members of the mammal category.

Although premise number and diversity are plausibly related to the previous observations surrounding bite reporting
intentions, research on inductive reasoning has not been extended to many concrete domains such as work on risk perception in health or real-world decision making about health behaviors. Similarly, while people's judgments regarding contagion have been studied in social and health psychology research (Nemeroff, 1995), these studies have focused on affective and cultural factors as opposed to underlying cognitive processes such as inductive reasoning.

In the present work, we test two specific hypotheses from our theory that inductive reasoning principles influence zoonosis risk perception. First, consistent with the aforementioned rabies study, individual differences in perceived risk from animal contact should be associated with individual differences in beliefs about interspecies disease transmission. Second, perceptions of risk to humans should increase as a result of being presented with communcations depicting transmissibility amongst a wider range of species.

## Study 1

The goal of Study 1 was to examine whether perceived disease risk (measured by intentions to report animal bites) is associated with beliefs about interspecies disease transmission likelihood. Based on the premise number principle, we hypothesized that individuals who endorse stronger likelihoods of disease transmission between a number of different animal species would be more likely to perceive human risks from animal bites. To test this hypothesis, we conducted a survey measuring intentions to report bites from common mammals and birds along with judgments of interspecies disease transmission likelihood for a ficticious novel disease.

Method. Participants were 289 adults ( $55 \%$ men; mean age $=33.6, S D=10.2$ ) who completed an online survey and were recruited through the Mechanical Turk crowdsourcing platform. The survey was available to Mechanical Turk workers in the following countries where English is the primary language: USA, Australia, Canada, Great Britain, Ireland, New Zealand, and the Bahamas. The majority of participants had undergraduate (48.8\%) or advanced degrees (8.7\%). The sample was predominantly White (80.6\%), with $5.9 \%$ Asian, $3.8 \%$ Black, $6.9 \%$ Hispanic, $1 \%$ Native American or Alaska Native, and $1.7 \%$ other ethnicities. A majority of the sample ( $73.4 \%$ ) reported currently owning a pet. Participants were compensated $\$ 2$ for participation in the survey. Informed consent was obtained from all individual participants in the study, no participants were excluded from the survey results, and all protocols were approved by the Texas Tech University IRB.

Design. The study materials consisted of an electronic survey containing sections on demographics, bite reporting intentions, and species-to-species disease transmission beliefs. Demographics questions included sex, sexual orientation, ethnicity, education level, parents' education level, language(s) spoken, and pet ownership.

In the bite reporting section, participants were asked to judge their likelihod of reporting bites from various target animals to a health professional (of any type). Participants judged likelihood of reporting for each animal using a slider that could be adjusted in units of 1 from $0-100$ and also contained descriptive labels ranging from "Very Unlikely" to "Very Likely". Mammal and bird reporting were presented in a random order on separate screens. Mammals included dogs, skunks, monkeys, bats, and squirrels. Birds included grackles, swans, robins, blue jays, and peacocks.

The species-to-species disease transmission beliefs section employed the same sliding scales as the bite reporting section, but participants were asked to rate the likelihood of betweenanimal disease transmission for a hypothetical new disease. Each question took the following form:

Scientists discover that a new disease can infect the liver tissue of [premise animal]. How likely is it that this disease can infect the following animals?

The conclusion animals were listed on separate lines, each with their own response slider. Premise animals included bats, dogs, skunks, monkeys, grackles, blue jays, swans, and peacocks. Conclusion animals included bats, dogs, skunks, monkeys, squirrels, grackles, robins, blue jays, swans, and peacocks. Fewer premise animals were used so that less time would be required to complete the survey and to reduce participant attrition. Animals only appeared as conclusion categories when they were not the premise. Premises were presented in a random order on separate screens.

Results. Intentions to report bites were highly reliable within person (mammals: Cronbach's $\alpha=0.86$; birds $\alpha=$ 0.95 ), as were judgments of interspecies disease transmission likelihood (mammal-to-mammal: $\alpha=0.96$, bird-to-bird: $\alpha$ $=0.97$; between birds and mammals: $\alpha=0.99$ ). Nonetheless, linear mixed effects models revealed that intentions to report bites varied considerably between different species [Mammals: $F(4,1152)=111.1, p<.001, \eta_{p}{ }^{2}=0.28$; Birds: $\left.F(4,1152)=35.23, p<.001, \eta_{p}^{2}=.11\right]$ and ratings of interspecies disease transmission likelihood varied between the different premise types [mammal-to-mammal, bird-tobird, between birds and mammals; $F(2,576)=356.3, p<$ $\left..001, \eta_{p}{ }^{2}=0.55\right]$. Intentions to report bites were stronger for mammals than for birds $[t(288)=27.06, p<.001, d=$ 1.59; Figure 1A], and diseases were rated as more likely to be transmissible within mammals or birds than between them [mammal-to-mammal vs. between birds and mammals, $t(288)=18.77, p<.001, d=1.10$; bird-to-bird vs. between birds and mammals, $t(288)=23.42, p<.001, d=1.38]$. Consistent with previous work suggesting bats are viewed as similar to both mammals and birds (Davis et al., 2013), bats were rated as more likely than other mammalsto share diseases with birds $[t(288)=7.03, p<.001, d=0.41]$.


Figure 1: (A) Intentions to report animal bites. (B) Association between intentions to report mammal bites and mammal-to-mammal disease transmission ratings. (C) Association between intentions to report bird bites and between bird and mammal disease transmission ratings. Error bars reflect 95\% within-subject confidence intervals.

In support of our primary hypotheses, we found that individual differences in endorsement of bird-to-bird and mammal-to-mammal disease transmission were both positively associated with individual differences in intentions to report mammal bites [Mammal-to-mammal: Kendall's $\tau=$ $.147, p<.001$ (Figure 1B); Bird-to-bird, $\tau=.140, p<.001$; Between birds and mammals: $\tau=.009$ (Pearson's r of .21 , .21 and .009 respectively)].

Consistent with the premise number principle, endorsing greater odds of interspecies disease transmission was associated with stronger intentions to report mammal bites. For bird bites, only ratings of disease transmission between birds and mammals were associated with reporting intentions [Mammal-to-mammal $\tau=.043$, Bird-to-bird $\tau=.077$, Between birds and mammals $\tau=.219, p<.001$ (Pearson's $r$ of
$.04, .05$, and .26 respectively) (Figure 1C)]. Coupled with weaker intentions to report bird bites overall, these results suggest that people may only judge birds as risky to the extent that they believe birds and mammals can share diseases.

Discussion. Study 1's results suggest that inductive reasoning principles may underlie people's perceptions of zoonosis risk. Although the correlations are between a small and medium correlation given Cohen's (1992) criteria, they are within those expected between general health attitudes and behaviors (Azien \& Timko, 1986; Glasman \& Albarracin, 2006). However, because the results are correlational, it is difficult to infer the causal direction between the beliefs about interspecies disease transmission risk and bite reporting. It is possible that both are influenced by a common underlying factor, such as beliefs about contagion (Haidt, McCaluey, \& Rozin, 1994) or risk attitudes (Dohmen et al., 2011). Moreover, because the results examine individual differences, it is not clear from Study 1 whether such inductive reasoning principles could be harnessed to influence people's beliefs about the risks associated with animal contact.

## Study 2

The goal of Study 2 was to test whether it is possible to influence people's perceptions of zoonosis risk through framing communications to portray a greater number of animals as susceptible to a disease. As a case study, real-world communications about Ebola virus vary in terms of how they describe the range of animals susceptible to the disease. The Centers for Disease Control's factsheet (CDC, 2016) lists contact with fruit bats and nonhuman primates (apes and monkeys) as sources of human Ebola infection. Contrastingly, the World Health Organization's factsheet (WHO, 2016) lists a much wider range of animals: chimpanzees, gorillas, fruit bats, monkeys, forest antelope, and porcupines.

According to the premise diversity principle, the WHO's factsheet should lead to stronger perceptions of Ebola risk from animal conact because it lists a broader range of animals as sources of human Ebola infection. To test this hypothesis, in Study 2 we gave particpants two different communications about Ebola derived from the CDC and WHO factsheets (tailored to control all other differences in wording).

Method. Participants were 152 adults recruited from Mechanical Turk in the same manner as for Study 1. Sample demographics were comparable to those in Study 1; additionally $94.7 \%$ of the sample in Study 2 reported eating meat. No participants were excluded from the results, and all protocols were approved by the Texas Tech University IRB.

Design. The study materials consisted of an electronic survey containing a demographics section, an experimentally manipulated reading prompt about Ebola (derived from CDC and WHO factsheets), an Ebola susceptibility section, a bite reporting intentions section, and a meat safety section.

For the reading prompt, participants were given the following description about Ebola and asked to fill in a blank box by detailing the animals listed in the description:

> The Ebola virus causes an acute, serious illness which is often fatal if untreated. Ebola virus disease (EVD) first appeared in 1976 in 2 simultaneous outbreaks, one in what is now Nzara, South Sudan, and the other in Yambuku, Democratic Republic of Congo. The latter occurred in a village near the Ebola River, from which the disease takes its name. Ebola is introduced into the human population through close contact with the blood, secretions, organs, or other bodily fluids of infected animals such as [animal 1], [animal 2], [animal 3], and [animal 4].

The animals listed in the description were experimentally manipulated between participants. Participants were randomly assigned to read either a CDC-inspired set of animals with lower premise diversity (fruit bats, gorillas, monkeys, and chimpanzees; $\mathrm{n}=81$ ) or a WHO-inspired set with higher premise diversity (fruit bats, monkeys, forest antelope, and porcupines; $\mathrm{n}=70$ ). To verify that these prompts did indeed differ in premise diversity, we had a separate group of participants ( $\mathrm{N}=53$ ) provide pairwise similarity judgments between each of the premise animals. Consistent with our expectations, participants judged the CDC prompt animals to be significantly more similar (i.e. less diverse, $\mathrm{t}(52)=14.56, \mathrm{p}$; .001).

Next participants completed the Ebola susceptibility questionnaire. For each question, participants were asked "How likely is it that [animal] can get Ebola?" ( $1=$ Very Unlikely, 7 = Very Likely). Animals included both mammals and birds: bats, monkeys, zebras, meerkats, anteaters, giraffes, gazelles, storks, flamingos, cranes, vultures, and parrots.

Next participants completed the bite reporting questionnaire. Participants were told to "imagine that you are on safari and get bitten by an animal, but the bite just barely breaks the skin" when considering whether they would report a bite to a health professional. Each question asked them to rate ( $1=$ Very Unlikely, 7 = Very Likely), "how likely would you be to report being bitten by a [animal]?"

Last, participants completed the meat safety questionnaire. Participants were asked to rate ( $1=$ Very Unsafe, $7=$ Very Safe), "how safe you think it is for people in general to eat meat from each animal" and to "consider only immediate health risks from disease transmission."

Results. The results were consistent with predictions based on the premise diversity principle. Participants in the WHO (diverse) wording condition rated individual mammals as more susceptible to Ebola $[t(150)=3.70, p<.001, d=$ 0.6 ; Figure 2A], were more likely to report mammal bites $[t(150)=2.85, p=.005, d=.46$; Figure 2B], and perceived mammal meat as less safe $[t(150)=2.66, p=.009, d=.434$; Fiture 2C].

The WHO (diverse) wording condition also increased perception of birds' susceptibility to Ebola $[t(150)=2.06, p=$ $.040, d=0.33$ ] but did not significantly increase intentions to report bird bites $[t(150)=1.10, d=0.18]$ or lower perceptions of meat safety $[t(150)=1.28, d=0.21]$.

We additionally used linear regression to test whether the effect of wording condition on bite reporting and perceptions of meat safety was mediated by its effect on Ebola susceptibility ratings. First, we found that Ebola susceptibility was significantly associated with bite reporting and meat safety perceptions for both mammals and birds, even after taking into account the effect of wording condition [Mammal bites: standardized $b=0.47 ; t(149)=6.40 ; p<.001$; Mammal meat: standardized $b=-0.43 ; t(149)=5.67, p<.001$; Bird bites: standardized $b=0.51 ; t(149)=7.32, p<.001$; Bird meat: standardized $b=-0.45 ; t(149)=6.01, p<.001]$. Next, we found that including Ebola susceptibility in the regression model with the effect of wording condition made the effect of condition non-significant for all models [Mammal bites: $b=0.18, t(149)=1.20$; Mammal meat: $b=$ $-.18, t(149)=-1.16$; Bird bites: $b=.009 ; t(149)=0.063$; Bird meat: $b=-0.06 ; t(149)=0.43]$, suggesting that the effects of condition on meat safety and bite reporting were fully mediated by the effect of the different wordings on participants' perceptions of Ebola susceptibility. Finally, using a bootstrapping procedure (Preacher \& Hayes, 2008), we found that the indirect pathways between wording condition and the bite reporting and meat safety ratings were significant for both birds and mammals.

## General Discussion

Results from both studies indicate an important role that cognitive research can play in combating emerging zoonoses. Although rarely studied in the public health literature, humans' inferences about risk are central to their interactions with potential disease vectors. We found that cognitive principles related to premise number and diversity impact perceptions of zoonotic disease transmission risk and associated health behaviors. To the extent that people believe it is possible for many diverse species to transmit diseases to one another, they become more wary of their own risk of infection.

An experiment based on CDC and WHO Ebola factsheets further revealed that individuals' inductive reasoning strategies can be harnessed to make communications about disease risk more effective. Through the use of cognitive framing strategies, it may be possible to reduce adverse contact with animals and increase rapid reporting of potential disease exposure. Such approaches may be particularly effective for rural communities that are difficult to reach with other interventions. These results have the potential to contribute goals of identifying low-cost strategies for reducing emerging disease risk before outbreaks occur (Heymann \& Dar, 2014).

To our knowledge, the present results are the first to suggest that inductive reasoning processes studied in cognitive psychology also influence health behaviors. With such connections established, future studies on disease transmission risk perception would benefit from even stronger connections with cognitive research. One question is how people judge risks from different species. Here we focused on person-level characteristics that relate to perceived risk of animal contact


Figure 2: Effect of communication wording on (A) Perceived susceptibility of animals to Ebola, (B) Intentions to report animal bites, and (C) Perceived meat safety. Error bars reflect $95 \%$ within-subject confidence intervals.
(bites and game meat), averaging over differences between species. However, not all animals are associated with the same zoonosis risk, and it will be important to understand how to tailor communications to impact species selectively. For example, bats have a very strong association to emerging zoonosis (Calisher, Childs, Field, Holmes, \& Schountz, 2006), and it may be useful to tailor messages to focus on bats specifically. Although bats were associated with high levels of intended bite reporting and were perceived as being unsafe to eat, participants also may have underestimated the risks bats pose to other animals - indeed, participants rated disease transmission risk between bats and other mammals as lower than for more typical mammals. Because wildlifelivestock interactions are a major driver of emerging zoonosis (Jones et al., 2008), this finding suggests that people may underestimate the risk of keeping livestock near bat habitats.

One limiation of our second study is that much of the sample is not at high risk for Ebola virus. However, because zoonoses are common within the countries surveyed and can be transmitted via many different interactions with animals all of our participants were at some risk of zoonosis exposure. Still, future research should examine whether risk level or other variables may moderate the effect of inductive reasoning principles on risk perception. We anticipate that people's personal experience with zoonosis, as opposed to pure risk level per se, may strengthen the relationship between beliefs and health intentions. Indeed, in the broader attitudes and public health literatures, many associations between attitudes and behaviors are rather weak in the general population, but are much stronger in groups with direct experience (Fazio \& Zanna, 1978; Glasman \& Albarracin, 2006). Thus while
many people in these studies do not have direct experience with Ebola virus, we would expect attitudes and health intentions to be even stronger in those who do.

The present research is primarily aimed at building interdisciplinary connections between public health research (particularly inerdisciplinary efforts such as One Health) and cognitive psychology. Still, the current results may have implications for basic psychological research on contagion and induction as well. The law of contagion is a prominent social psychology construct that describes people's tendencies to believe that negative (and positive) properties, including diseases and social ills, can be transmitted to objects or people through mere contact (e.g. Rozin and Royzman (2001)). Current theories of sympathetic magical thinking often make distinctions between the law of contagion and the law of similarity, a separate construct that describes the belief that objects that share surface features also share deeper common essences (e.g., leading to disgust with fudge shaped like dog feces, and beliefs that voodoo dolls can affect the person they resemble; Rozin, Markwith, and Ross (1990)).

The present results suggest that the laws of contagion and similarity may not be fully separate, and similarity-based effects may influence perceptions of contagion. Indeed, theories suggest that inductive reasoning principles like premise number and diversity can increase generalization of properties (such as disease susceptibility) via similarity relationships between known and novel/unknown examples. For example, the diverse prompts in our second experiment may have increased perceptions of Ebola susceptibility by increasing the likelihood that the unknown examples would match the known examples in some respect. A major question in
cognitive psychology is how different respects in which items can be similar (Medin, Goldstone, \& Gentner, 1993) impact generalization of novel/unknown properties. Although our data does not distinguish between different candidate theories for similarity-based transfer of contagion, the results are suggestive that beliefs about contagion can be transferred via such similarity relationships.

In conclusion, emerging diseases from animals pose a substantial public health concern, yet little is known about how people judge risks associated with different drivers of zoonosis. The present studies illustrate that basic cognitive principles related to inductive reasoning not only impact individuals' perceptions of disease risk and associated health behaviors, but also can be harnessed for tailoring messages to properly convey risks associated with emerging diseases.

## References

Azien, I., \& Timko, C. (1986). Correspondence between health attitudes and behavior. Basic and Applied Social Psychology, 7, 259-276.
Bingham, G. M., Budke, C. M., \& Slater, M. R. (2010). Knowledge and perceptions of dog-associated zoonoses: Brazos County, Texas, USA. Preventative Veterinary Medicine, 93, 211-221.
Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. V., \& Schountz, T. (2006). Bats: Important reservoir hosts of emerging viruses. Clinical Microbiology Reviews, 19, 531-545.
Cohen, J. (1992). A power primer. Psychological Bulletin, 1, 155-159.
Daszak, P., Cunningham, A. A., \& Hyatt, A. D. (2000). Emerging infectious diseases of wildlife - threats to biodiversity and human health. Science, 287, 443-449.
Davis, T., Goldwater, M. B., Gaylord, N., Worthy, D., Otto, A. R., \& Glass, B. D. (2013). The cognitive psychology of human-bat interactions: Implications for ecological policy and zoonotic disease transmission. In Bats: Phylogeny and evolutionary insights, conservation strategies and role in disease transmission (pp. 1-17). Hauppage: Nova.
Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., \& Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. Journal of the European Economic Association, 9, 522-550.
Fazio, R. H., \& Zanna, M. P. (1978). On the predictive validity of attitudes: The roles of direct experience and confidence. Journal of Personality, 46, 228-243.
for Disease Control, C. C., \& Prevention). (n.d.). About ebola virus disease. (Retrieved from http://www.cdc.gov/vhf/ebola/about.html)
Glasman, L. R., \& Albarracin, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. Psychological Bullletin, 132, 778-822.
Haidt, J., McCaluey, C., \& Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven
domains of disgust elicitors. Personality and Individual Differences, 16, 701-713.
Hayes, B. K., Heit, E., \& Swendsen, H. (2010). Inductive reasoning. Wiley interdisciplinary reviews: Cognitive Science, 1, 278-292.
Heit, E., \& Feeney, A. (2005). Relations between premise similarity and inductive strength. Psychonomic Bulletin and Review, 12, 340-344.
Heymann, D. L., \& Dar, O. A. (2014). Prevention is better than cure for emerging infectious diseases. The BMJ, 348, g1499.
Janes, C. R., Corbett, K. K., Jones, J. H., \& Trostle, J. (2012). Emerging infectious diseases: the role of social sciences. The Lancet, 380, 1884-1886.
Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., \& Daszak, P. (2008). Global trends in emerging infectious diseases. Nature, 451, 900-993.
Kamins, A. O., Rowcliffe, J. M., Ntiamoa-Baidu, Y., Cunningham, A. A., Wood, J. L., \& Restif, O. (2015). Characteristics and risk perceptions of Ghanaians potentially exposed to bat-borne zoonoses through bushmeat. EcoHealth, 12, 104-120.
Li, F., Cao, B., Li, Y., Li, H., \& Deak, G. (2009). The law of large numbers in children's diversity-based reasoning. Thinking and Reasoning, 15, 388-404.
Lopez, A. (1995). The diversity principle in the testing of arguments. Memory \& Cognition, 23, 374-382.
McDonald, J., Samuels, M., \& Rispoli, J. (1996). A hypothesis-assessment model of categorical argument strength. Cognition, 59, 199-217.
Medin, D., Goldstone, R. L., \& Gentner, D. (1993). Respects for similarity. Psychological Review, 100, 254-278.
Nemeroff, C. J. (1995). Magical thinking about illness virulence: conceptions of germs from "safe" versus "dangerous" others. Health Psychology, 14, 147-151.
Organization), W. W. H. (n.d.). Ebola virus disease (fact sheet no. 103). (Retrieved from http://www.who.int/mediacentre/factsheets/fs103/en/)
Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.
Preacher, K. J., \& Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40, 879-891.
Rozin, P., Markwith, M., \& Ross, B. (1990). The sympathetic magical law of similarity, nominal realism and neglect of negatives in response to negative labels. Psychological Science, 1, 383-384.
Rozin, P., \& Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review, 5, 296-320.

# High-Performing Readers Underestimate Their Text Comprehension: Artifact or Psychological Reality? 

Stefanie Golke (stefanie.golke@ezw.uni-freiburg.de)<br>Department of Educational Science, University of Freiburg<br>Rempartstr. 11, 79098 Freiburg, Germany

Jörg Wittwer (joerg.wittwer@ezw.uni-freiburg.de)<br>Department of Educational Science, University of Freiburg<br>Rempartstr. 11, 79098 Freiburg, Germany


#### Abstract

We focused on the controversy whether high-performing readers consistently underestimate their comprehension or are prone to detrimental overestimations as much as less skilled readers are. Therefore, we conducted an experiment ( $N=105$ university students) to investigate judgment bias as a function of reading skill and text difficulty in terms of text cohesion. Results showed that the easy text produced underestimation of comprehension, whereas the hard text led to overestimation. Furthermore, readers with higher reading skills were less prone to overestimate their comprehension of a hard text than less skilled readers. However, we also found that more skilled readers showed lower sensitivity in discriminating between correct and incorrect answers than less skilled readers. Overall, our results do not support the idea that high-performing readers consistently underestimate their text comprehension. Findings are discussed with respect to readers' awareness of different text-based judgment cues and their (beliefs about their) reading skill.


Keywords: judgment bias; metacognitive sensitivity; text difficulty; reading skill; high-performing readers

## Introduction

Successful learning from text requires readers to accurately judge their text comprehension because false judgments (e.g., overestimation) can hamper the learning process. It is well acknowledged that readers in general are prone to overestimations, whereas particularly high-performing readers (i.e., readers who achieve high scores on a text comprehension test) might more likely underestimate their comprehension (de Bruin et al., 2016; Dunlosky \& Rawson, 2012). However, the methodology used to unveil underestimation by high-performing readers is not fully undisputed. Therefore, it remains unclear whether highperforming readers' underestimation is psychological reality or rather an artifact. We present an experiment that was conducted to advance our understanding about how highperforming readers judge their text comprehension.

## Comprehension Judgments and the Learning Process

Learning from text involves constructing a mental representation of the information provided in a text and retrieving the learned information at a later time. The learning process heavily depends on a reader's metacognitive ability
to monitor comprehension (i.e., metacomprehension), which is mirrored in the correspondence between a reader's comprehension judgment and actual performance on a comprehension test (Wiley, Griffin, \& Thiede, 2005).

Comprehension judgments can occur at different times in the learning process. Accordingly, research uses different types of comprehension judgments (Griffin, Jee, \& Wiley, 2009). The first type is the prospective judgment of comprehension that readers make after reading a text to predict how well they will perform on yet unknown test questions about the text. Furthermore, when readers complete test questions, they can use information about their (perceived) performance in answering the test questions to evaluate their comprehension. Thus, the second type of comprehension judgments is readers' confidence in their retrieved answers on single test questions (i.e., response confidence). The third type is the retrospective judgment of comprehension that refers to a whole set of test questions (i.e., how many of the test questions were answered correctly). The three types of judgment are assumed to reflect (slightly) different aspects of metacomprehension but complement each other (Schraw et al., 2014).

When readers make comprehension judgments, they normally use available cues (Koriat, 1997). These cues can arise from the learning material (e.g., text difficulty), a reader's (self-perceived) skills and resources (e.g., prior knowledge, reading ability) and a reader's experiences when reading the text or answering test questions. All types of cues can be useful for precise judgments when they are valid indicators of the required level of comprehension.

To support learning, judgments of comprehension need to be precise because they influence readers' subsequent learning activities. Imprecise judgments, especially overestimations, have a detrimental effect on learning (Dunlosky \& Rawson, 2012). For example, overestimation means that readers do not realize that their comprehension of text is worse than they think. Therefore, they might abstain from engaging in remedial activities. In contrast, underestimation might be less problematic for learning but it can hamper learners in allocating their learning time appropriately.

## Controversy Over High-Performing Readers' Underestimation

Numerous studies have shown that readers typically provide imprecise judgments. Most readers overestimate their comprehension of text and are overconfident in the correctness of their retrieved information when answering test questions about the text (Dunlosky \& Rawson, 2012; Maki et al., 2005).

However, concerning high-performing readers, it is sometimes reported that they tend to underestimate their comprehension (de Bruin et al., 2016; Zabrucky, 2010). This underestimation is often interpreted as a result of specific metacognitive or cognitive processes. For example, high performers are assumed to not give very high judgments of their comprehension to avoid being perceived as arrogant or to have negatively skewed misperceptions of their abilities, both resulting in underestimation (Zabrucky, 2010).

A completely different explanation of this phenomenon refers to a statistical bias of the measure used to unveil overestimation and underestimation (i.e., judgment bias) that becomes relevant when readers' level of performance is determined by their performance in the experimental comprehension test. More specifically, judgment bias uses the signed difference between a reader's prospective or retrospective judgment of comprehension and his/her actual performance on a comprehension test. Therefore, the reader's judgment bias is constrained by his/her performance (Griffin et al., 2009; see also Kruger \& Dunning, 1999). That is, readers who achieve the maximum or a very high performance score on a comprehension test (i.e., highperforming readers) are much more likely to show underestimation than readers with lower performance scores. Conversely, readers who have a very low performance score are much more likely to overestimate their comprehension. Furthermore, if the performance-level of readers is determined by their performance on the comprehension test - that is also part of the measure of judgment bias - both measures are statistically dependent on each other and normally show high negative correlations (i.e., higher performance on the comprehension test is associated with lower/more negative scores of judgment bias). Thus, the finding that high-performing readers underestimate their comprehension could also be a statistical artifact and, hence, might not reflect their actual ability to judge comprehension. To disentangle the effect of the level of comprehension on judgment bias, it seems useful to investigate judgment bias as a function of both readers' general reading skill and test/text difficulty.

## The Effect of Text Difficulty and Reading Skill

Maki et al. (2005) investigated judgment bias (i.e., overestimation or underestimation) as a function of text difficulty - determined by the readability of the texts - and students' general reading skill. Their findings did not support the view that high performers generally underestimate their
comprehension. Instead, for difficult texts (i.e., lower readability), it was found that high-ability readers were precise when making prospective judgments. Only when making postdictions, they underestimated their comprehension but so did medium-ability readers as well. Conversely, for easier texts (i.e., higher readability but still in the range between difficult and standard texts), all readers provided overoptimistic predictions of comprehension but precise postdictions.

This latter finding on easier texts is intriguing with regard to Schraw and Roedel's (1994) study that determined difficulty by the mean item difficulty of the test questions. They found that readers were overconfident on their answers in response to difficult and moderately difficult items but precise on items with low difficulty. Because high-ability readers in Maki and colleagues' study (2005) solved about $70 \%$ of the test items on the easier text, these test items were of low difficulty for them. Hence, their postdiction judgments were precise. But why did the (high-ability) readers overestimate their comprehension when making prospective judgments on the easier text? It appears as if the higher readability of the text might have induced readers - at any level of reading ability - to be overoptimistic. This interpretation is supported by findings from Weaver and Bryant (1995) who revealed that predictions of comprehension are not highly correlated to actual performance for texts with high or low readability.

Thus, previous studies showed that high-performing readers do not consistently underestimate their comprehension. Therefore, these studies provide useful hints about the controversy on high-performing readers. However, at the same time, the studies only provide information about the effects of item difficulty (Schraw \& Roedel, 1994) or text difficulty in terms of readability (Maki et al., 2005; Weaver \& Bryant, 1995). Readability that depends on, for example, word length, number of words per sentence, or passive/active structure is a salient text-based cue and a more distal indicator of the difficulty of the text content than, for example, text cohesion. With regard to theories on text comprehension (see e.g., Wiley et al., 2005), varying text difficulty in terms of readability might not discriminate well enough between readers with different levels of reading proficiency. Hence, it would be interesting to focus on cohesion as a different indicator of text difficulty and investigate judgment bias as a function of this text feature and reading skill.

## The Present Study

We examined the precision of comprehension judgments as a function of text difficulty and reading skill. In contrast to previous studies, we determined text difficulty in terms of text cohesion. To assess judgment bias, we used the signed difference between a reader's prospective or retrospective judgment of comprehension and his/her actual performance on a comprehension test. Moreover, we assessed readers'
metacognitive sensitivity and response bias as additional indicators of metacomprehension.

As main effects of text difficulty on performance, we expected that the easy text resulted in higher performance on the test questions than the hard text. Regarding the effect of text difficulty on judgment bias, we based our hypotheses on findings about item difficulty instead of texts' readability. Therefore, given the statistical dependence of performance level and judgment bias, we hypothesized that the easy text would lead to significant underestimation whereas the hard text should result in significant overestimation. Hence, using a within contrast, the easy text should result in a lower bias score of prospective and retrospective judgments than the hard text. Furthermore, we investigated in an exploratory way how reading skill was linked to readers' judgment bias for the easy text compared with the hard text. To do so, we computed multiple linear regressions that included reading skill and prior knowledge as relevant predictors of prospective judgment bias for the easy text and the hard text. In case of the retrospective judgment bias, we also used readers' metacognitive sensitivity and response bias that are based on readers' response confidence for the test questions as additional predictors. With regard to the relationship of reading skill with metacognitive sensitivity and response bias, we inspected their correlations with each other.

## Method

## Design

The experiment followed a two-factorial design with reading skill as a metric between-subjects factor and text difficulty as the within-subjects factor with two levels: one text with lower text difficulty (easy text) and one text with higher text difficulty (hard text) in terms of cohesive relations within the text (see also Materials). The order of the texts was counterbalanced across all participants.

As dependent variables, we assessed: 1) text comprehension (i.e., number of correctly answered questions about the text), 2) the bias of prospective and retrospective judgments, 3) metacognitive sensitivity, and 4) response bias. Furthermore, we assessed participants' prior knowledge about the topics of the text materials.

## Participants

Participants were 105 university students from educational science. They had a mean age of $22.78(S D=4.95)$ years and $82 \%$ of them were female.

## Materials

Table 1 displays the main characteristics of both texts. Given the scope of this study, we selected texts that represented different levels of text difficulty in terms of cohesion. Cohesion refers to the extent to which relations between ideas in a text are made explicit by using, for example, textual features such as causal, temporal, or additive connectives. We
determined cohesion by the proportion of sentences that contained a cohesive device on how the sentence is connected to previous ones. As displayed in Table 1, the cohesion score for the hard text was considerably lower than the score for the easy text. Thus, the hard text required readers to engage more deeply in comprehending the text compared with the easy text. Apart from cohesion, the texts were equivalent with respect to other characteristics including surface cues, such as readability or text length, as well as the domain of the texts (i.e., biology, see Table 1).

We used six open-ended comprehension questions for each text. The questions tapped information explicitly stated in the text.

Table 1: Characteristics of the texts.

| Characteristic | Easy text | Hard text |
| :--- | :---: | :---: |
| Topic | Reproduction | Immunology |
| No. of words | 380 | 397 |
| No. of sentences | 25 | 30 |
| Flesch-Index | 46 | 41 |
| Cohesion | 0.67 | 0.38 |

Note. ${ }^{\text {a }}$ Texts with a Flesch-Index (i.e., flesch reading ease score) between 30 and 50 reflect difficult texts in terms of readability that are typically used in higher education.

## Instruments and Measures

Prospective and Retrospective Judgments Participants indicated how many of the six text comprehension questions they think they would answer correctly (= prospective judgment) or had answered correctly (= retrospective judgment; value between 0 and 6).
Judgment Bias We used the signed difference between a reader's prospective or retrospective judgment of comprehension and the actual performance on the text comprehension test. Hence, the bias score could range between -6 (i.e., maximum underestimation) and +6 (i.e., maximum overestimation).
Response Confidence For each question, participants indicated how confident they were that their answer was correct (Likert scale from $1=$ very uncertain to 7 = very certain).
Metacognitive Sensitivity ( $\boldsymbol{d}^{\prime}$ ) Sensitivity reflects the ability of readers to distinguish between correct and incorrect responses on test questions. It uses readers' performance on single test questions and their response confidence on these test questions. We determined metacognitive sensitivity via $d^{\prime}$ that is based on signal detection theory (see Fleming \& Lau, 2014; Schraw et al., 2014) using the hit rate (i.e., number of questions that a reader answered correctly and rated as correct, divided by the total number of correctly answered questions) and the false alarm rate (i.e., number of questions that a reader did not answer correctly but rated as correct, divided by the total number of incorrect answers). The measure of $d^{\prime}$ is the difference between the standardized hit
rate and the standardized false alarm rate. A value of zero means that the reader could not discriminate between correct and incorrect responses, a positive value (i.e., higher hit rate than false alarm rate) reflects good sensitivity, and a negative value (i.e., higher false alarm rate than hit rate) suggests that the reader considered rather a false answer as correct than a correct answer.
Response Bias (c) The response bias $c$ is based on the sensitivity measure $d^{\prime}[c=-0.5$ * (standardized hit rate + standardized false alarm rate)]. The response bias represents the tendency of a reader to accept false alarms $(c<0)$ or to be cautious when giving confidence judgments on single test questions in order to avoid false alarms ( $c>0$ ).
Reading Skill We used a subtest of a computer-based German reading comprehension test for adults (ELVES; Richter \& van Holt, 2005). The subtest assessed higher-order processes of text comprehension.
Prior Knowledge There was a total of 12 open-ended questions that assessed readers' prior knowledge on immunology and reproduction. These questions were not identical to the text comprehension questions.

## Procedure

At the beginning, participants answered the prior knowledge test and proceeded with the reading comprehension test ELVES. After that, participants read the first experimental text and then judged their comprehension by predicting how many of the six text comprehension questions they think they would answer correctly. After the judgment, they answered the comprehension questions and rated their response confidence for each question. After answering all comprehension questions, participants made a retrospective comprehension judgment by indicating how many of the six questions they thought they had answered correctly. Subsequently, participants proceeded with the second experimental text in the same manner as they did for the first one.

## Results

To test the hypotheses regarding the main effect of text difficulty on performance and judgment bias, we performed (paired) $t$-tests (for descriptive statistics, see Table 2). In line with our hypotheses, we found that the easy text resulted in higher performance on the text comprehension questions, $t(104)=13.73, p<.001$, Cohens $d=1.49$ (large effect), than the hard text. Moreover, the mean scores of prospective and retrospective judgment bias for both texts (see Table 2) were significantly different from zero (i.e., the value of perfect judgment), all $p$ 's $<.004$. Thus, the easy text resulted in significant underestimation for both prospective and retrospective judgments. In contrast, the hard text resulted in significant overestimation for both types of judgment. A paired $t$-test confirmed that the easy text resulted in lower bias scores of prospective judgments, $t(104)=-12.96, p<.001$, Cohens $d=-1.42$ (large effect), and lower bias scores of
retrospective judgments, $t(104)=-6.13, p<.001$, Cohens $d=$ -0.68 (medium effect), than the hard text.

Furthermore, we performed multiple linear regressions to examine our research question regarding the relationship of reading skill with judgment bias for the easy and the hard text. For each type of judgment bias (i.e., prospective vs. retrospective bias), we computed separate multiple regressions for the easy and the hard text. Predictors were entered in one step.

Table 2: Means (and standard deviations) for dependent variables as a function of text difficulty.

| Dependent variable | Easy text | Hard text |
| :--- | :---: | :---: |
| Text comprehension | $4.90(1.31)$ | $2.99(1.25)$ |
| Prospective judgment <br> bias | $-0.79(1.42)$ | $1.21(1.40)$ |
| Retrospective judgment <br> bias | $-0.35(1.18)$ | $0.49(1.29)$ |

Regarding prospective judgment bias, we included prior knowledge on the topic of the text and reading skill as predictors. The results (see Table 3) showed that neither prior knowledge nor reading skill were statistically relevant predictors of prospective judgment bias for the easy text. However, for the hard text, reading skill was a statistically significant negative predictor of prospective judgment bias. That is, participants with higher reading skills were less likely to overestimate their comprehension of the hard text. However, as descriptive statistics revealed (see Table 2), we cannot conclude that these participants generally showed underestimation because only $12 \%$ of the total sample underestimated their comprehension of the hard text when making prospective judgments.

Table 3: Predictors of prospective judgment bias for easy and hard text.

| Predictor | $b$ | $S E b$ | $t(101)$ | $p$ |
| :--- | ---: | :---: | :---: | :---: |
| Easy text |  |  |  |  |
| Constant | -0.39 | 0.53 | -0.74 | .462 |
| Reading skill | -0.02 | 0.03 | -0.65 | .516 |
| Prior <br> knowledge | 0.00 | 0.01 | -0.32 | .749 |
| Hard text |  |  |  |  |
| Constant | 2.43 | 0.49 | 4.96 | $<.001$ |
| Reading skill <br> Prior <br> knowledge <br> -0.08 | 0.03 | -2.83 | .006 |  |

Note. For easy text: $R^{2}=.01, F(2,102)=0.32, p=.730$. For hard text: $R^{2}=.08, F(2,102)=4.16, p=.018$.

Moreover, regarding retrospective judgment bias, we included prior knowledge, reading skill as well as
metacognitive sensitivity and response bias as predictors. The multiple regression analyses revealed (see Table 4) that metacognitive sensitivity and response bias significantly predicted the retrospective judgment bias for the easy text. That is, the better a reader discriminated between correct and incorrect responses and the more the readers avoided false alarms in the confidence rating, the less likely this reader was to overestimate comprehension when making retrospective judgments on questions about an easy text. This result was also found for the hard text. Additionally, reading skill also predicted retrospective judgment bias for the hard text.

Table 4: Predictors of retrospective judgment bias for easy and hard text.

| Predictor | $b$ | $S E b$ | $t(99)$ | $p$ |
| :--- | ---: | ---: | ---: | :---: |
| Easy text |  |  |  |  |
| Constant | -0.02 | 0.41 | -0.05 | .958 |
| Reading skill | -0.03 | 0.02 | -1.30 | .222 |
| Prior knowledge | 0.00 | 0.01 | 0.31 | .761 |
| Sensitivity | -0.24 | 0.08 | -2.96 | .004 |
| Response bias | -0.59 | 0.14 | -4.13 | $<.001$ |
| Hard text |  |  |  |  |
| Constant | 1.51 | 0.46 | 3.30 | .001 |
| Reading skill | -0.06 | 0.02 | -2.34 | .022 |
| Prior knowledge | 0.00 | 0.01 | -0.48 | .632 |
| Sensitivity | -0.16 | 0.09 | -1.78 | .078 |
| Response bias | -0.78 | 0.16 | -5.04 | $<.001$ |

Note. For easy text: $R^{2}=.22, F(4,104)=6.83, p<.001$. For hard text: $R^{2}=.23, F(4,99)=7.30, p<.001$.

Furthermore, we explored the relationship of reading skill with metacognitive sensitivity and response bias, respectively. As displayed in Table 5, we found that participants with higher reading skills were less cautious (measure of response bias, $c$ ) when giving confidence ratings on the comprehension questions about the easy text. In addition, they were less able to discriminate between correct and incorrect answers (measure of metacognitive sensitivity, $d^{\prime}$ ) in response to questions about the hard text. Given the magnitude of the correlation coefficients, these relations are small effects. However, it appears that more-skilled readers were metacognitively less aware and, therefore, more overconfident when answering the test questions.

Table 5: Pearson's $r$ correlations between reading skill, sensitivity ( $d^{\prime}$ ), and response bias (c) for easy and hard text.

|  | Easy text |  | Hard text |  |
| :---: | :---: | :---: | :---: | :---: |
| Measure | $d^{\prime}$ | $c$ | $d^{\prime}$ | $c$ |
| Reading skill | .10 | $-.22^{*}$ | $-.27^{* *}$ | -.16 |

Note. ${ }^{*} p<.05 .{ }^{* *} p<.01$.

To sum up, we found that reading skill was a relevant predictor of prospective and retrospective judgment bias in case of the hard text, but not in the case of the easy text. Hence, participants with higher reading skills were less likely to overestimate their comprehension of the hard text. Moreover, we found that response bias and sensitivity influenced retrospective judgment bias for the easy and the hard text. Thus, the better participants discriminated between correct and incorrect answers or the more cautious they were when rating their answers as correct, the less likely they made overoptimistic retrospective judgments. In addition, we found that sensitivity and response bias were more negative for readers with higher reading skills, although these effects were rather small.

## Discussion

This study aimed to shed further light on the question whether high-performing readers adhere to judgment processes that lead them to consistently underestimate their comprehension across materials with different levels of difficulty. The results of our study do not support this assumption. Instead, our results suggest that readers with higher reading skills are better calibrated because they are less prone to overestimate their comprehension of a hard text compared with readers with lower reading skills. Kwon and Linderholm (2014) also found this relationship for texts with standard readability.

The finding that participants with higher reading skills were better calibrated supports the notion that higher reading skills include better monitoring during reading. Readers who actively monitor their text comprehension obtain a more comprehensive mental model of the text and are therefore more precise at judging their comprehension (Wiley et al., 2005). Furthermore, although a relationship between reading skill and judgment bias is evident, the magnitude of the relationship we found in our study is rather small. This indicates that other characteristics of the reader are also or even more relevant for judgment bias, for example, the selfperceived reading skill (Kwon \& Linderholm, 2014).

In contrast to the hard text, there was no relationship between reading skill and judgment bias on the easy text. This finding can be explained by the low difficulty of the test questions. Therefore, general reading skill was not predictive of test performance on the easy text and, thus, reading skill was not related to judgment bias on the easy text.

Another important finding in our study were the negative relations of reading skill with metacognitive sensitivity and response bias. This finding suggests that readers with higher reading skills may be metacognitively unaware when responding to the type of test questions we used in the present study. Therefore, despite their good calibration with respect to the hard text, participants with higher reading skills showed a flawed discrimination performance. To explain this lower discrimination, it can be speculated that their beliefs about their reading skill tempted high-ability readers to
proceed less mindfully with the test questions and, thus, to be overconfident on their answers.

This interpretation does not necessarily contradict the findings on the positive influence of reading skill and the negative impact of sensitivity and response bias on retrospective judgment bias because the strength of these relations was rather small. Moreover, it can be assumed that other factors influence judgment bias as well. Therefore, the seemingly contradicting relations between reading skill, discrimination performance, and retrospective judgment bias might simply indicate complex interactions between readers' characteristics and judgment processes that still need to be further uncovered (Schraw et al., 2014).

The findings of this study also contribute to the understanding of the effects of text-based cues on judgment bias. In our study, the easy text (i.e., higher cohesion) resulted in underestimation. Given that performance on test questions about the easy text was rather high, this underestimation was very likely to occur due to probabilistic assumptions (Schraw \& Roedel, 1994). Likewise, the observed overestimation on the test questions about the hard texts was also expected. In contrast, the easy text (i.e., higher readability) in Maki and colleagues' (2005) study resulted in overestimation of prospective judgments for all readers. Only when readability of texts was low, readers, except for weak readers, adjusted their comprehension judgments. Thus, we can conclude that texts that are easy to read - and, therefore, often preferred in instructional contexts because they increase performance are more likely to seduce readers to be overoptimistic. Conversely, high text cohesion does not seem to have such an effect on metacognitive judgment. Therefore, readers, including high-ability readers, are apparently unaware of the low validity of good text readability as a cue to judge their comprehension. With respect to readers' sensitivity for text cohesion, we aim to analyze our data in more depth addressing possible anchor effects based on the withinsubjects design and also examine the role of reading skill in this regard.

## Acknowledgments

This study is based on data collected by Nadine Bianchi, Nina Krämer, and Andreas Leiprecht as part of their unpublished Bachelor theses.

## References

de Bruin, A. B. H., Kok, E. M., Lobbestael, J., \& Grip, A. de. (2016). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. Metacognition and Learning. Advance online publication doi:10.1007/s11409-016-9159-5
Dunlosky, J., \& Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. Learning
and Instruction, 22, 271-280.
doi:10.1016/j.learninstruc.2011.08.003
Fleming, S. M., \& Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, 19. doi:10.3389/fnhum. 2014.00443

Griffin, T. D., Jee, B. D., \& Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. Memory \& Cognition, 37, 1001-1013.
doi:10.3758/MC.37.7.1001
Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126, 349-370. doi:10.1037/0096
Kruger, J., \& Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77, 1121-1134.
Kwon, H., \& Linderholm, T. (2014). Effects of selfperception of reading skill on absolute accuracy of metacomprehension judgements. Current Psychology, 33, 73-88. doi:10.1007/s12144-013-9198-x

Maki, R. H., Shields, M., Wheeler, A. E., \& Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. Journal of Educational Psychology, 97, 723-731. doi:10.1037/00220663.97.4.723

Richter, T., \& van Holt, N. (2005). ELVES: Ein computergestütztes Diagnostikum zur Erfassung der Effizienz von Teilprozessen des Leseverstehens [A computer-based instrument to assess efficiency of reading processes]. Diagnostica, 51, 169-182.
Schraw, G., Kuch, F., Gutierrez, A. P., \& Richmond, A. S. (2014). Exploring a three-level model of calibration accuracy. Journal of Educational Psychology, 106, 1192-1202. doi:10.1037/a0036653
Schraw, G., \& Roedel, T. D. (1994). Test difficulty and judgment bias. Memory \& Cognition, 22, 63-69. doi:10.3758/BF03202762
Weaver, C. A., \& Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. Memory \& Cognition, 23, 12-22.
Wiley, J., Griffin, T. D., \& Thiede, K. W. (2005). Putting the comprehension in metacomprehension. The Journal of General Psychology, 132, 408-428. doi:10.3200/GENP.132.4.408-428
Zabrucky, K. M. (2010). Knowing what we know and do not know: Educational and real world implications. Procedia - Social and Behavioral Sciences, 2, 12661269. doi:10.1016/j.sbspro.2010.03.185

# Computational modeling of auditory spatial attention 

Edward J. Golob (Edward.Golob@utsa.edu)<br>Department of Psychology, University of Texas, San Antonio<br>San Antonio, TX USA

K. Brent Venable (kvenabl@tulane.edu)<br>Department of Computer Science, Tulane University<br>New Orleans, LA USA

Jaelle Scheuerman (jscheuer@tulane.edu)
Department of Computer Science, Tulane University
New Orleans, LA USA

Maxwell T. Anderson (manders7@tulane.edu)<br>Department of Psychology, Tulane University<br>New Orleans, LA USA


#### Abstract

Attention plays a fundamental role in higher-level cognition. In this paper we develop a computational model for how auditory spatial attention is distributed in space. Our model builds on the assumption that attentional bias has bottom-up and top-down components. We represent each component and their synthesis as a map, associating a level of attentional bias to locations in space. The maps and their interaction are modeled using an artificial intelligence approach based on constraints. We describe the behavioral task we have designed to measure the attentional bias and discuss the results. We then test different hypotheses on the shape and interaction modalities of the maps in terms of how well they fit our behavioral data. The findings showed that combining top-down and bottom-up spatial attention gradients that differ in their spatial properties produced the best fit to behavioral data, and suggested several novel mechanisms for future testing.


Keywords: auditory attention; computational modeling; saliency map; constraints.

## Cognitive engineering problems and attention

Humans evolved in a dynamic environment of shifting opportunities and threats. Consequently, we are wellequipped to organize and frequently change goals and priorities to effectively deal with events in the natural and social worlds. High-level cognitive attributes, such as intelligence, creativity, and imagination presumably evolved to capitalize on these dynamics to promote survival (Flinn, Geary, \& Ward, 2005). A key aspect of higher-level cognition is attention. An important role for attention-like selection in information processing may not be limited to human cognition. For example, Helgason and colleagues proposed that attention is an essential element for systems to exhibit generalized intelligence, regardless of whether it is a biological or artificial intelligence system (Helgason, Thorisson, Garrett, \& Nivel, 2014).

In this article we broadly consider attention as a flexible means of enhancing specific aspects of information processing, as determined by factors such as the current goal (top-down) or stimulus characteristics important to the organism (such as unexpected loud sounds)(Chun, Golomb, \& Turk-Browne, 2011). This flexibility is assumed to be implemented by specific cognitive processing routines that were selected during the course of human evolution (Cosmides \& Tooby, 2013). Differences between sensory modalities in terms of how the adequate stimulus and receptor transduction relate to the kinds of information that can be detected in the environment is one factor relevant to the design of attentional processes. Consequently, in at least some respects attentional processing may sharply differ between sensory modalities.

We focus on the auditory system, and consider implications of the idea that the auditory system has a comparative advantage over other modalities in the ability to panoramically monitor the environment. Hearing provides an early warning system (Scharf, 1998) that allows organisms to prepare for, or evade, threats and to capitalize on prey or mating opportunities. This "3-D sphere" of spatial sensitivity for hearing is unique among sensory modalities because it can detect environmental events that are at a distance from the body (cf. somatosensation, gustation, to some degree olfaction) and out of sight (vision).

## Stability-flexibility dilemma and attentional systems

Most attention models consider attention that is directed by a conscious choice ("top-down" or "voluntary") to differ in important ways from attention that is involuntarily "captured" by an event in the world that has a salient property (Petersen \& Posner, 2012). Salience can be due to physical properties, such as a loud sound, or by having personal meaning such as one's name (Moray, 1959) or taboo words (Arnell, Killman, \& Fijavz, 2007),
and other aspects that may depend on the situation (Gygi \& Shafiro, 2011). The distinction between top-down and bottom-up attentional functions is both useful and
meaningful, even though top-down and bottom-up attentional processes are highly interactive (Folk, Remington, \& Johnston, 1992).

One of the defining features of the top-down aspect of attention is that it is limited. Either by design, such as matching the limitations in the number of actions that can be done at one time, or by overload from having too much information to be processed at one time, or both, voluntary attention is limited (Allport, 1989; Posner, 1978). Spatial attention has been intensively studied, in part, because it vividly illustrates limitations in attentional capacity. The limited capacity of spatial attention can be expressed as a spatial gradient relative to an attended location (reviewed in (Cave, 2013). The classic way to consider this gradient is that it reflects decreased investment of attentional resources with greater distance from an attentional focus, and the extent of the gradient can be adjusted based on the current task (Eriksen \& St. James, 1986).

The fundamental problem with including top-down and bottom-up attention in one general attention system that distributes attentional resources across space is that attention cannot be simultaneously both focused and diffuse. This kind of trade-off has been termed the "stability-flexibility dilemma (Liljenström, 2003), the "shielding-shifting dilemma" (Thomas Goschke \& Bolte, 2014), and a trade-off between organization and flexibility (Baars, 1997). The problem is compounded by not knowing when something will happen outside of the attentional focus that is critical for survival, thus preventing an anticipatory shift by top-down attention. Both top-down and bottom-up attention have clear survival value, but limited attention capacity implies tradeoffs between resources devoted to top-down vs. bottom-up attention functions. Similar issues concerning cognitive trade-offs have been explored in the context of cognitive control and task switching (Goschke, 2000), automatic vs. controlled processing (Schneider \& Chein, 2003), various dual process models of cognition (Evans, 2008), long-term knowledge (Caramazza \& Shelton, 1998), and memory systems (Sherry \& Schacter, 1987).

## Methods and computational modeling

The present study addresses the stability-flexibility dilemma posed by needing attention to be both focused on a task while also monitoring the environment for potential threats or opportunities by modeling auditory spatial attention bias as the net result of two attention modules and their output (Figure 1). Our aim is to develop a rigorous quantitative theory of auditory spatial attention. One module, called the "goal map" is devoted to top-down attention necessary to perform the current task. The other module, termed the "saliency map", is specialized to monitor, in parallel, the environment and, when needed, engage bottom-up orienting that overrides current attentional focus based on top-down processes. We combine novel parametric behavioral measures to map-out auditory attention over space with a computational model to explain how specific top-down and
bottom-up mechanisms jointly determine the shape of auditory spatial attention gradients.

Top-down goals


Figure 1. Proposed attentional model architecture.
Relative to existing models of auditory attention, the current model is designed to help understand somewhat higher levels of cognitive processing. Others have modeled perceptual features and how they are combined to generate a saliency map. There is overlap with our model at the level of saliency map. Prior work computes perceptual features such as stimulus location, frequency, intensity, and saliency as an output that is computed from a raw sensory input (Coensel \& Botteldooren, 2010; Kayser, Petkov, Lippert, \& Logothetis, 2005). Instead, we start with perceptual features as a given input and model how top-down and bottom-up modules interact in the context of working memory. Note also that the choice of modeling auditory spatial attention in the frontal plane has the benefit of needing to explain attentional bias in only one-dimension (the azimuth plane at a constant distance from center of head), which simplifies modeling. In contrast, visual studies of saliency maps use two-dimensional models (Kalinli \& Narayanan, 2007).

Model design The model is designed using constraints, a very general and powerful artificial intelligence framework for problem modeling and solving. (Rossi, Van Beek, \& Walsh, 2006). Constraints lie at the core of many successful applications in several domains such as scheduling, planning, vehicle routing, configuration, networks, and bioinformatics. The basic idea in constraint-based modeling is that the user states the constraints and a general-purpose constraint solver is used to solve them. Constraint solvers take a real-world problem, represented in terms of decision variables and constraints, and find, if it exists, an assignment to all the variables that satisfies all the constraints. A constraint concerns a subset of variables and defines which simultaneous assignments to those variables are allowed. Solutions are found by searching the solution space either systematically, as with backtracking or branch and bound algorithms, or use forms of local search which may be incomplete, that is, there is no guarantee they will return a solution. Systematic methods often interleave search and inference, where inference consists of propagating the information contained in one constraint to other constraints via shared variables. Constraints have been used before in the context of human cognition for example to model skilled behavior (Howes, Vera, Lewis, \& McCurdy, 2004). Recently an implementation of the cognitive architecture ACT-R
based on constraint handling rules, which are a closely related to constraints, has been proposed in (Gall \& Frühwirth, 2014). To the best of our knowledge, this is the first time constraints are employed at this level of cognitive modeling and in the context of attention.

The model makes several assumptions regarding proactive and reactive control. According to Braver's dual mechanisms framework (Braver, 2012), proactive control generates a sustained attentional bias in accordance with task goals, such as focusing on a pianist about to begin their recital. Reactive control, as the name suggests, is attentional orienting in response to a stimulus, such as if the pianist plays their first chord and everybody realizes that the piano is out of tune. In our model the goal map is the mechanism for proactive control. The spatial focus of the goal map can also be redirected in response to stimuli, and so could have a role in reactive control too. In contrast, the saliency map codes for reactive control. The relation between the saliency map and proactive control is only indirect. The focus of the saliency map is designed to be away from the goal map focus, thus any proactive shifts in the goal map focus will consequently lead to a similar shift in the saliency map focus.

Task and data to be modeled Young adult subjects ( $\mathrm{n}=42$ ) listened to 25 and 75 Hz amplitude modulated white noise, and responded with left/right hand (counterbalanced across subjects). Virtual stimuli were delivered via headphones to one of 5 locations in the frontal horizontal plane ( $\mathrm{L} \rightarrow \mathrm{R}$ locations: $-90^{\circ},-45^{\circ}, 0^{\circ},+45^{\circ},+90^{\circ} ; 2.4 \mathrm{sec}$ SOA). In each 6 min block subjects attended to a standard location (either $90^{\circ}, 0^{\circ}$, or $+90^{\circ}$ ). Most stimuli were given at the standard location ( $\mathrm{p}=.84$ ), with occasional shifts to the other 4 locations ( $\mathrm{p}=.04 / \mathrm{location}$ ). Analysis of variance (ANOVA) was used to examine reaction time as a function of standard condition (3) and stimulus location (5). Data were collapsed across AM rates (ns). We note that the following model is designed from general principles based on the attention and working memory literature, but the actual modeling here is very specific to our task. This is common in other areas such models of canonical visual search tasks. Future work will expand this model to include other tasks and situations.

The model Behavioral results were modeled using a constraint-based representation made up of three components: goal map, saliency map, and priority map. The maps represent the attentional bias across the horizontal frontal plane $\left(-90^{\circ}\right.$ to $+90^{\circ}$ ) (see heat-map in Figure 2, topleft). The priority map is the weighted sum of the goal and saliency maps and represents the total attentional bias at each degree location. Operationally, attention bias in the priority map relates to reaction time by equation 1 :

Eq. 1 Attentional bias $=(2,000$-reaction time $) /(2,000)$
The " 2,000 " value was chosen as an upper limit on reaction times to be analyzed (both in ms), and included nearly every correct trial in every subject. The units of attention bias are
arbitrary, but index reaction time with a range of between approximately 0.90 , which corresponds to a an extremely fast reaction time of 200 ms , to 0.0 , which indicates a $2,000 \mathrm{~ms}$ reaction time. Thus, larger attention bias values in the priority map reflect short reaction times and efficient processing, and longer reaction times have smaller values.


Figure 2: Variables and constraints that represent the interactions between the three maps in the model.

Each map is represented by a collection of variables, one for each 2-degree (the minimum distinguishable by a human ear) location in the range $\left\{-90^{\circ}, \ldots,+90^{\circ}\right\}$, and a set of constraints over the variables. Figure 2 (left) shows a portion of the constraint graph of the model where nodes correspond to variables and edges to constraints. These constraints limit the simultaneous assignments of the constrained variables as indicated in the equations below, where $V_{G}{ }^{i}, V_{S}{ }^{i}$, and $V_{P}{ }^{i}$ represent the $i$-th variables of the goal, saliency and priority maps. The constraints defining each map involve the variable corresponding to the attended location $(A)$ and the variables corresponding to a location. The variables associated with the goal map (blue nodes in Figure 2) are constrained to represent a standard Gaussian distribution with its peak at level $G_{G}$ and located at the attended location A. An example of one such distribution is shown in Figure 2 right above the set nodes representing the goal map variables. Each node represents the 2 degree portion of the x -axis right above it and the associated attentional bias value (y-axis) is the value assigned to the variable corresponding to the node. Similarly, for the saliency (where the distribution is shown below the nodes) and the priority map. Note that parameter $d_{G}$ is the standard deviation of the Gaussian and is used to model a symmetrical decrease in top-down attentional resources away from the goal location. Likewise, the variables corresponding to the saliency map have values compatible with an inverted Gaussian distribution with peak level $-G_{S}$ at attended location A and standard deviation $d_{S}$ representing a symmetrical increase in bottom-up attention away from the attended location. Finally, each priority map's variable takes as value the weighted sum of the values of the corresponding goal map and saliency map variables. The
graph at the bottom-right of Figure 2, shows an example of a goal map bias (green), saliency map bias (blue) and the associated priority map (purple). Weights $\alpha$ and $\beta$, are used to model the magnitude of contribution of, respectively, the goal and saliency maps to the priority map.

$$
\begin{gathered}
\text { Goal Map: }\left(A=a, V_{G}^{i}=G_{G} e^{\frac{-|a-i|^{2}}{2 d_{G}^{2}}}\right) \\
\text { Saliency Map: }\left(A=a, V_{S}^{i}=G_{S}-G_{S} e^{\frac{-|a-i|^{2}}{2 d_{G}^{2}}}\right) \\
\text { Priority Map: }\left(V_{G}^{i}=u, V_{S}^{i}=v, V_{P}^{i}=\alpha u+\beta v\right)
\end{gathered}
$$

We note that the constraint based model allows an easy extension to a 2D or 3D bias distribution. This can be achieved in two ways: either by increasing the number of variables (e.g. for the planar case a set for each concentric hemisphere), or by increasing the complexity of the domain values, e.g. 2D bias distributions over 2 degree sectors.

## Results

We first describe and discuss the results of the behavioral experiment in the previous section, and then the results of applying our computational model to the behavioral data. The goal map and saliency map each had two parameters for fitting: attention bias and standard deviation.

## Behavioral results

Reaction time curves for angular shifts had an inverted-u shape at all 3 standard locations ( $p$ 's $<.01$ )(Figure 3). Attending to the right ( $+90^{\circ}$ standard) had an attenuated inverted-u curve vs. $-90^{\circ}$ and $0^{\circ}$ standards ( $\mathrm{p}<.01$ ). Results show comparable reaction time increases to the nearest shift location for the $0^{\circ}$ and $-90^{\circ}$ standards, and then decreases in reaction time at the most distant locations. For the $+90^{\circ}$ standard there was a more gradual increase and decrease in reaction time across shift locations. Accuracy was high for all stimulus locations and conditions (> 94\%) and will not be analyzed here.


Figure 3. Behavioral results showing mean reaction times for standard locations at the far left $\left(-90^{\circ}\right)$, midline $\left(0^{\circ}\right)$, and far right $\left(+90^{\circ}\right)$ locations. A) Reaction time as a function of location for the three standard locations. B) Normalized reaction times plotted relative to the standard location, here denoted by " 0 ".

## Computational modeling results

Stochastic local search was used to find parameter values for $d_{S}, d_{G}, G_{S}$ and $G_{G}$, that minimize the root-mean-square (rms) error between the priority map and behavioral data. Bootstrapping methods were used to compare model fit as the parameters for the goal and saliency maps varied. There were 100 runs for each standard location to assess the consistency of results. On each run half of the subjects ( $\mathrm{n}=21$ ) were randomly selected to train the model. The model was then tested for fit using root-mean square error on the grand average of the remaining subjects ( $n=21$ ).

Comparison of two vs. three-component models Having attention bias centered on the standard location and decreasing with distance was modeled with only the goal map having input to the priority map. This two-component model had a poor fit to the reaction time data, with rms values nearly 100x worse than models with both goal and saliency map inputs to the priority map (Figure 4). Models with both topdown (goal map) and bottom-up (saliency map) spatial attention bias fit the data well, with rms values ranging from 0.0040 to 0.0035 for left or right standard locations $\left( \pm 90^{\circ}\right)$ and 0.0011 and 0.0012 for the $0^{\circ}$ standard. The fits at each standard location were all significantly different from each other ( p 's $<.001$ ). By contrast, rms values with only the goal map in the model were 0.3137 ( (-90 ${ }^{\circ}$ standard), $0.3060\left(+90^{\circ}\right.$ standard), and 0.1191 ( $0^{\circ}$ standard). We note that a model based only on the saliency map was not tested as it would have not been able to model the increased bias at the attended location. The results clearly show that a simple attention gradient that decreases with distance from the attended standard location (goal map only) is unable to account for the behavioral data. Models with both goal and saliency maps provided a good fit to the behavioral results. It is unclear why


Figure 4. Model fit with only the goal map (GM) representing top-down attention bias vs. the addition of a bottom-up component (saliency map, SM). Models examined whether goal and saliency maps had either equal influence on the priority map ("GM \& SM) or their levels were included as a parameter in the model ("free levels"). Model fit was measured using root-mean-square error (rms).
the fit for the $0^{\circ}$ standard is even better than the $\pm 90^{\circ}$ standards, but this may relate to the truncated range of locations on either side $\left( \pm 90^{\circ}\right)$.

Standard deviation parameters The range of spatial attention bias for the goal and saliency maps was quantified with separate standard deviation parameters (Figure 5). When only the goal map was included in the model the best fits had standard deviations of $\sim 100^{\circ}$, which produced a gradual decrease of attentional bias from the standard location. As shown above, only including the goal map produced a poor fit to the behavioral data. In all models with goal and saliency maps the standard deviations had, large, progressive reductions from standard locations on the left, to midline, and to the right ( $\mathrm{p}<.001$ ). This pattern was evident for both the goal and saliency map SD parameters. Analysis of both fixed and free bias models showed main effects of map type, with significantly larger SD values in the saliency map (p's $<.001$ ). There were interactions between standard location and map type, indicating that the difference between the SD of goal and saliency maps varied among standard locations ( p 's $<.001$ ).


Figure 5. Standard deviation (SD) parameters in the models with goal map (GM) and saliency map (SM) components. Standard deviation units are in degrees.

Attention bias level Lastly, we tested a model where the attention bias levels from the goal and saliency maps to the priority map were free to vary. The findings from when bias parameters were added to the model are shown in Figure 6 for each standard location. For the $\pm 90^{\circ}$ standards the goal map had a significantly greater bias than the saliency map, indicating a greater influence over the priority map outcomes. This was most evident for the $-90^{\circ}$ standard, which had little variability among modeling runs ( $\mathrm{p}<.001$ ). In contrast, for the $0^{\circ}$ standard there was substantial variability over modeling runs, and there was no significant difference between goal and saliency map bias.

Note that the range of attentional bias levels in the goal and saliency maps is much larger than the priority map (data not shown). This is the result of the model solutions having SD


Figure 6. Attention bias level results in the free level models. Bias indicates the overall level of inputs from the goal and saliency maps to the priority map. Greater bias indicates more influence over the priority map values.
parameter values that were both narrow enough to individually have bias levels near asymptote within the degree range tested. The model sums the contributions of goal and saliency maps to generate the priority map, which in turn is proportional to reaction time. The goal and saliency curves over space overlapped such that when one map had low bias the other had a large amount of bias. This additive approach in combination with moderate SD ranges forces many locations to have large differences between goal and saliency map values while retaining a much smaller range of priority map values. For perspective, the range of biases of between . 76-. 70 corresponds to reaction times between 480600 ms .

## Discussion and conclusions

In this paper we have studied spatial attention of the auditory system from a behavioral and computational modeling point of view. The main findings were that a traditional top-down attention gradient could not account for the behavioral data, but a model with two gradients corresponding to top-down and bottom-up bias worked well. The model is based on structuring the overall allocation of attentional bias as the sum of bottom-up and a top-down components. We have presented behavioral results aimed at describing the effect of the overall attentional bias and we have provided an experimental evaluation of different model hypothesis in terms of how well they fit the data. There was a pronounced left-right asymmetry in the reaction time profiles as a function of location that was accounted for by progressive reductions in the SD parameters of goal and saliency maps. The results support our approach which constitutes, to the best our knowledge, the first computational model that integrates top-down and bottom-up auditory spatial attention processes.

## Acknowledgments

This study was supported by NIH grant DC014736.

## References

Allport, A. (1989). Visual attention. In M. I. Posner (Ed.), Foundations of Cognitive Science (pp. 631-682). Cambridge, MA: MIT Press.
Arnell, K. M., Killman, K. V, \& Fijavz, D. (2007). Blinded by emotion: target misses follow attention capture by arousing distractors in RSVP. Emotion (Washington, D.C.), 7(3), 465-477.

Baars, B. J. (1997). The Global Workspace Theory of Consciousness. Journal of Consciousness Studies Studies, 4(4), 292-309.
Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. Trends in Cognitive Sciences, 16(2), 106-13.
Caramazza, A., \& Shelton, J. R. (1998). Domain-specific knowledge systems in the brain the animate-inanimate distinction. Journal of Cognitive Neuroscience, 10(1), 1-34.
Cave, K. R. (2013). Spatial attention. In The Oxford Handbook of Cognitive Psychology (pp. 117-130). New York, NY: Oxford University Press.
Chun, M. M., Golomb, J. D., \& Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. Аппи Rev Psychol, 62, 73-101.
Coensel, B. De, \& Botteldooren, D. (2010). A model of saliency-based auditory attention to environmental sound. Proc 20th Intl Cong Acoustics, 20(August), 18.

Cosmides, L., \& Tooby, J. (2013). Evolutionary Psychology: New Perspectives on Cognition and Motivation. Annu. Rev. Psychol, 64, 201-29.
Eriksen, C. W., \& St. James, J. D. (1986). Visual attention within and around the field of focal attention: a zoom lens model. Percept Psychophys, 40, 225-240.
Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology, 59, 255-278.
Flinn, M. V., Geary, D. C., \& Ward, C. V. (2005). Ecological dominance, social competition, and coalitionary arms races: Why humans evolved extraordinary intelligence. Evolution and Human Behavior, 26(1), 10-46.
Folk, C. L., Remington, R. W., \& Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. J Exp Psychol Hum Percept Perform, 18(4), 1030-1044.
Gall, D., \& Frühwirth, T. (2014). A formal semantics for the cognitive architecture ACT-R. In International Symposium on Logic-Based Program Synthesis and Transformation (pp. 1-18).
Goschke, T. (2000). Intentional Reconfiguration and Involuntary Persistence in Task Set Switching. Control of cognitive processes:, 18, 331. In S. Monsell \& J. Driver (Eds.), Attention and performance XVIII (pp. 331-355). Cambridge, MA: MIT Press.
Goschke, T., \& Bolte, A. (2014). Emotional modulation of control dilemmas: the role of positive affect, reward, and dopamine in cognitive stability and flexibility.

Neuropsychologia, 62, 403-423.
Gygi, B., \& Shafiro, V. (2011). The incongruency advantage for environmental sounds presented in natural auditory scenes. Journal of Experimental Psychology. Human Perception and Performance, 37(2), 551-65.
Helgason, H. P., Thorisson, K. R., Garrett, D., \& Nivel, E. (2014). Towards a General Attention Mechanism for Embedded Intelligent Systems. International Journal of Computer Science and Artificial Intelligence, 4, 17.

Howes, A. H., Vera, A., Lewis, R. L., \& McCurdy, M. (2004). Cognitive Constraint Modeling: A Formal Approach to Supporting Reasoning About Behavior Cognitive Constraint Modeling: A Formal Approach to Supporting Reasoning About Behavior. In Proceedings of the Annual Meeting of the Cognitive Science Society (pp. 595-600).
Kayser, C., Petkov, C. I., Lippert, M., \& Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology: CB, 15(21), 1943-1947.
Liljenström, H. (2003). Neural stability and flexibility: a computational approach. Neuropsychopharmacology 28 Suppl 1, S64-73.
Moray, N. (1959). Attention in dichotic listening: affective cues and the influence of instructions. Quarterly Journal of Experimental Psychology, 11(1), 56-60.
Petersen, S. E., \& Posner, M. I. (2012). The attention system of the human brain: 20 years after. Annual Review of Neuroscience, 35, 73-89.
Posner, M. I. (1978). Chronometric explorations of mind. New York: Halsted Press.
Rossi, F., Van Beek, P., \& Walsh, T. (2006). Handbook of Constraint Programming (Foundations of Artificial Intelligence). Amsterdam: Elsevier.
Scharf, B. (1998). Auditory attention: The psychoacoustical approach. In H. Pashler (Ed.), Attention (pp. 75-117). East Sussex, UK: Psychology Press.
Schneider, W., \& Chein, J. M. (2003). Controlled \& automatic processing: Behavior, theory, and biological mechanisms. Cognitive Science, 27(3), 525-559.
Sherry, D. F., \& Schacter, D. L. (1987). The evolution of multiple memory systems. Psychological Review, 94(4), 439-454.

# Recruitment of the motor system in the perception of handwritten and typed characters 

Chelsea L. Gordon (cgordon7@ucmerced.edu)<br>University of California at Merced, Department of Cognitive Science, 5200 North Lake Road, Merced, CA, 95343<br>Ramesh Balasubramaniam (ramesh@ucmerced.edu)<br>University of California at Merced, Department of Cognitive Science, 5200 North Lake Road, Merced, CA, 95343


#### Abstract

Many different functional roles have been ascribed to the motor system due to its prevalent recruitment in perceptual and cognitive tasks other than motor production. We discuss findings that suggest the motor system might take on multiple roles that vary with context and the brain networks involved. Using single-pulse TMS, we measured the corticospinal excitability of the FDI muscle in primary motor cortex as participants viewed words that were either typed or handwritten. We observed consistent facilitation of corticospinal excitability during reading of handwritten text. Although we observed facilitation in corticospinal excitability during the presentation of typed text, this effect decreased with repetitive presentations of stimuli. We suggest that the facilitation during presentation of typed words is a case of action simulation, and that the diminishing facilitation in the case of typed stimuli is representative of sensory prediction by the motor system. These findings suggest that we should consider multiple roles for motor recruitment during the observation of visual stimuli, taking context into consideration.


Keywords: Action observation, motor involvement in reading, sensorimotor prediction.

## Introduction

The motor system is involved in a large number of cognitive and perceptual domains, including action observation, perception of object affordances, speech perception, language and metaphor, and social cognition. There are many theories aimed at explaining this widespread use of the motor system. We will introduce some of these theories here and work supporting each of them. Then we provide an alternative hypothesis: that there is no one particular role for the motor system in perception and cognition, but that it plays many roles decided, in part, by situational context.

Outside of its role in moving the body, the most common and widely posited role of the motor system is in observation-execution mapping. A network which includes several motor regions of the brain is responsible for mapping observed actions onto one's own motor system, which is said to underlie action understanding. The neurological underpinnings of this system rely on particular neurons in motor cortex, called mirror neurons, that fire during both observation and performance of a motor act in macaque monkeys and in humans. (Rizzolatti et al. 1988, di Pellegrino
et al., 1992; Gallese et al., 1996; Mukamel et al., 2010). Umiltà et al. (2001) found that some subset of mirror neurons fire during the final part of an observed action, even if that final part of the action is occluded from view, suggesting that mirror neurons code the goal-related execution of an action. This also suggests that mirror neurons respond to action-related situations where determining the actor and situation involves more inference, suggesting a role in deeper understanding of action. Kohler and colleagues (2002), recording from single neurons in monkey premotor cortex, found that some of the same neurons that fire during observed action will also fire when monkeys are only hearing the auditory information from the action (i.e., the cracking of a peanut). Again, this involves inference to the presence of the actor without visual recognition.

Motor activation during action observation is also called motor resonance (Iacoboni, 1999), due to its timedependent and effector-specific nature. It is said that the motor system of the observer "resonates" with that of the actor, allowing the observer to use their own body to understand the action being performed. Gangitano, Mottaghy, and Pascual-Leone (2001) applied transcranial magnetic stimulation (TMS) to record motor-evoked potentials (MEPs) from the first dorsal interosseus (FDI) muscle on the right hand during the observation of a cyclic hand movement. They found that at the time when the FDI muscle of the observed hand was most contracted, MEPs in the observer were highest, and when the muscle was least contracted, MEPs were lowest. Thus, the motor resonance occurring in the observer is timelocked with specific muscle activity in the observed agent.
A related theory of motor system involvement is that of overt action simulation (Barsalou, 2009; Gallese and Lakoff, 2005). This is related to the above mentioned position and not mutually exclusive, as observationexecution matching could involve low-level simulation of an actor. Simulation theories, however, posit explicit ongoing simulation underlying perceptual and cognitive processes as a sort of online enactment, particularly for understanding semantics of action language. In other words, linguistic phrases such as "the boy caught the ball" are understood by low level simulation of the action in the
sentence. Numerous studies have shown that action language activates the motor system.
In an fMRI experiment, Hauk, Johnsrude, and Pulvermuller (2004) found that when participants read action words related to the arm, leg, or face, the corresponding regions of the motor somatotopy were active. Oliveri et al. (2004), using single pulse TMS, found that motor cortex activation increased for both action nouns and action verbs when compared to activation during non-action words. Candidi and colleagues (2010) found that verbs conjugated in the future tense induce higher corticospinal excitability than verbs conjugated in the past. Finally, Yang and Shu (2016) performed a meta-analysis on a large number of fMRI experiments where subjects were listening to literal action sentences, fictive motion sentences, metaphorical action sentences, and idioms, and found increased activation in motor regions during metaphorical action sentences. This activation is thought to contribute to understanding and mapping metaphors onto their concrete reference. Simulation theories are often associated with embodied cognition, proposing that we use our brains and bodies to ground conceptual and abstract content in sensorimotor systems.

The third theory we discuss is sensory prediction. In this case what is being coded for in the motor system is sensorimotor prediction, or continuous online prediction of the very next state of the stimulus. A predictive role for the motor system is suggested in Clark's (2015) theory of "embodied prediction", in which motor activation during action observation would entail prediction of the upcoming sensory signal based on the current sensory information. In this case, motor cortex would be active during the observation of a grasping action, because the observer's brain would be actively predicting the very next movement via motor regions. Thus, a predictive role can account for the findings from the action-observation network literature. Wilson and Knoblich (2005) outline a version of the perceptual prediction role of motor areas that uses what they call perceptual emulators. An emulator is a mental simulation that models the external world. The emulator continues updating the model online, and the output from this emulator can be compared to the actual state of the external world to verify that expectations are being met. Emulators running in the motor system would internalize a model of the biomechanics of the human body, allowing observers to model the movements of an observed agent as they unfold in time. Importantly, these emulators are running one step ahead of sensory input, predicting the upcoming external state before it happens and then comparing real to modeled state afterward.

If the motor system uses these emulators, it should also be able to model predictable sensory information that isn't human-made, such as rhythmic waves or the bouncing of a ball, by internalizing a model for the observed system. Supporting research comes from Schubotz (2007), whose work suggests that even observation of movements coming
from non-animate entities recruits the motor system. In a number of fMRI experiments, they find that particular types of perceptual prediction tasks involving such things as pitch identification, spatial or object-related identification tasks activates premotor areas in a somatotopic way, similar to effector-related observation/execution tasks. For instance, object-related tasks recruited regions of premotor cortex that share activation in hand-related execution and observation tasks. As there isn't a common repertoire to humans and rolling waves, these findings could not be explained under the motor resonance account.

We propose that the role of the motor system varies depending on context. For instance, during the perception of action language, the motor system might serve to provide the motor component of covert simulation that occurs in the embodied processing of language. During the observation of very well practiced movements by an outside actor, the motor system may serve the purpose of driving motor resonance in the observer to quickly map the actions to the observer's body and understand the action. Finally, during perceptual processing of non-human movements, the motor system might serve to assist in perceptual processing by way of predictive emulator models.
One potential way of differentiating between prediction and simulation is by observing how the modulation of the motor system changes over repetitive viewing of stimuli. If the observer is simulating the observed action, then we should see a steady facilitation of MEPs across repetitions of a stimulus, indicating simulation at each occurrence. If motor system facilitation is due to predictive processes, however, might expect a different pattern of modulation. Because less error correction takes place as a stimulus becomes more predictable, we can predict that the neural populations underlying the predictive processes will be less active for more predictable sensory stimuli. Thus, we should see a decrease in corticospinal excitability over multiple repetitions of a stimulus, as it becomes more predictable and leads to lower error correction..
In this experiment we look at corticospinal excitability using single-pulse transcranial magnetic stimulation (TMS) during the perception of written language to examine the extent of motor involvement in a few variations of the stimuli. Subjects viewed videos of words being written out with a stylus and of words being typed letter by letter. Previous analyses in our lab have shown that observation of handwriting leads to motor simulation, while observation of typed words does not. We proposed that this is because while it is apparent that the handwritten text are human created, this is less apparent for text created on a keyboard. We repeat all stimuli four times over the course of the experiment. We predicted that MEPs in the handwritten stimuli trials would show an even facilitation across all presentations of the stimuli, because simulation should be consistent no matter how predictable it is. We hypothesized that the MEPs in the typed stimuli trials would show initial facilitation, which would lessen as the stimuli are repeated
and there is less prediction error. This would be expected because the first presentations of the stimuli, appearing letter by letter, should be difficult to predict, resulting in a large error in perceptual prediction. As stimuli are presented more often, perceptual prediction should become easier and less effort required on the part of error correction.

## Methods

## Participants:

Twenty-four right-handed normal participants (8 males, 16 females, mean age $\sim 19.5$ ) were recruited in this study through UC Merced's SONA research system. All participants passed a safety screen and gave written, informed consent. The experimental procedure was approved by the UC Merced Institutional Review Board. Participants received 2 research credits that can be used for credit in some undergraduate courses.

## TMS and EMG recording:

Corticospinal excitability was measured by the amplitude of motor evoked potentials (MEPs) recorded using electromyography (EMG) on the first dorsal interosseus (FDI) muscle of the right hand. Two small adhesive electrodes $\left(1 \mathrm{~cm}^{\wedge} 2\right)$ were placed over the belly of the recorded muscle and a ground electrode was placed over a bone on the participant's elbow. A bandpass filter (50 $\mathrm{Hz} 1,000 \mathrm{~Hz}$ ) was applied to the EMG signal, which was digitized at $1,000 \mathrm{~Hz}$ for offline analysis. MEPs were elicited by applying single-pulse TMS to the FDI region of the left motor cortex. Pulses were delivered using a Magstim Rapid ${ }^{2}$ with an attached $70-\mathrm{mm}$ figure-of-eight coil positioned over the optimal scalp location with the handle pointing backward at 45 degrees from the midline. The procedure was as follows. Subjects were fitted with a swim cap that was covered by a grid of dots placed $1 \mathrm{~cm}^{2}$ apart. Optimal scalp position was determined by moving the coil by one centimeter intervals until the location eliciting the best MEPs was identified. This location was marked on the swim cap worn by the participant. Resting motor threshold was determined as the percent of machine output that produced 5 out of 10 MEPs of at least $50 \mu \mathrm{~V}$ peak-topeak amplitude. The stimulation intensity during the experiment was set to $120 \%$ of a participant's resting motor threshold. The coil was held steady at the optimal position throughout the experiment. Subjects were instructed to keep their head still and remain relaxed for the duration of the experiment.

## Experimental paradigm:

The visual stimuli consisted of videos of either handwritten or typed words or nonwords appearing letter by letter at a variable typing speed averaging 3-4 letters per second. Words were chosen that did not relate to any actions or manipulable objects, to ensure that our measurement would not be influenced by the effects of semantic processing of
action. We also included 10 baseline trials, which consisted of a single black box for the same duration as the stimuli. We chose to randomize the baseline trials in with the rest of the trials so that the baseline measure would not be biased by a lack of attention that can occur when baseline measures are all recorded pre-experiment. Stimuli included ten linguistic stimuli, which appeared four times in each of the conditions. This resulted in 80 stimuli trials and 10 baseline trials, or a total of 90 trials. Eight seconds passed in between individual trials, and the total experiment length was approximately 12 minutes. We chose to apply stimulation two seconds into the ongoing video, so that as the stimuli were repeated, they were more highly predictable (by the presence of the first few letters) by the time stimulation occurred. Because TMS stimulation would occur two seconds into the video, we ensured that the typed stimuli would display one of the following letters at that time [ $\mathrm{N}, \mathrm{H}$, $\mathrm{U}, \mathrm{M}, \mathrm{J}, \mathrm{I}]$, so that if subjects were simulating the typing in proper typing position, FDI would be the simulated muscle.
The stimuli appeared on a computer screen in front of the participants. Participants were instructed to attend to the stimuli on the screen and were given notice when the experiment was one-third and two-thirds of the way finished to prevent loss of attention.
TMS pulses were delivered 2 seconds after video onset. The interval between trials was 8 seconds, to avoid any cumulative effects of single-pulse TMS. After the experiment, subjects were asked whether they were able to stay attentive during the length of the experiment. Participants who said they were not were excluded from analyses (5 subjects).

## volume volume mjknmw mjknmw

Figure 1: Examples of stimuli used in the experiment. Handwritten stimuli are on the left and typed stimuli on the right. In the experiment, participants saw the stimuli appear as a video as they were written or typed.

## Results

The average raw MEP amplitude for handwritten stimuli was 1.126 mV , with a standard deviation of 1,303 . The average for typed words was 1098 mV , with a standard deviation of 1.295 . Because of the large variations between participants, raw MEP amplitude values were $z$-scored to allow inter-individual comparisons. The resulting z-scores indicate the distance (in standard deviations) that a particular MEP score is from the mean. Figure 2 shows the average $z$ score in each condition. The average $z$-score for handwritten stimuli was .1, while that for typed stimuli was -.06.
A two-way repeated-measures analysis of variance (ANOVA) was computed on the standardized MEPs to test
for significant effects. The considered factors were condition (handwritten or typed) by order (nth time that a stimulus appeared). We observed a significant main effect for condition, with handwritten stimuli producing greater facilitation of MEP amplitude with respect to typed stimuli, $\mathrm{F}(1,23)=7.62, \mathrm{p}<.01$. We also observed a significant interaction effect of condition by order of presentation, $F(3,184)=3.77, p=.05$. In particular, there was a consistent facilitation in MEPs in the handwritten stimuli regardless of how many times the stimulus has been presented. In the typed stimulus condition, however, there was an initial facilitation in the MEP amplitude that decreased with each repetition of the stimuli. This pattern of results confirms our hypothesis of typed stimuli showing an initial facilitation of corticospinal excitability, followed by a decrease in that facilitation. This also confirms our hypothesis that the handwritten words would induce consistent facilitation of corticospinal excitability.

A linear regression of presentation number on baseline zscore was performed in order to evaluate whether the baseline MEPs changed with multiple presentations of the stimuli. The regression came out non-significant $(\mathrm{t}=-1.1$, p $>.3$ ). This indicates that overall MEP amplitudes are not varying as a function of time or number of repetitions to stimuli.

MEPs of Condition by order of presentation


Figure 2: Standardized (Z-scored) MEP amplitudes for each condition. X-axis shows presentation number (nth time a stimulus was presented). Motor evoked potentials in the handwritten condition show consistent facilitation, while those in the typed condition show initial facilitation that decreases with presentation number.

## Discussion

In this experiment we observed a differential pattern of motor facilitation dependent on word reading condition. In particular, the observation of actively handwritten words produced a persistent facilitation in MEP amplitudes. This is consistent with the action observation research, where
subjects view actions produced by others over multiple trials and produce consistent MEP facilitation. When subjects are exposed to actively typed words, however, the pattern of MEP facilitation changes, with repetitive exposure to the stimulus resulting in a decrease in observed corticospinal excitability. In previous work, we hypothesized that typed stimuli might not show simulation because of two reasons. Either the act of typing has weak or no sensorimotor association, or the discrete nature of typed words does not invoke simulation the same as the continuous strokes of handwriting.

Evidently the motor system is doing something different from motor simulation during the observation of words that are actively typed. One potential hypothesis is that corticospinal excitability in the typed condition is influenced by attentiveness. As subjects are repeatedly exposed to words, they might lose interest and thus exhibit lower attention. We included in the experiment a baseline measure appearing randomly throughout, consisting of a solid black box that appears instead of the language stimuli. There were 10 baseline trials used. If corticospinal excitability was picking up on a measure of attention, we should see a predictable decreasing trend in MEP amplitudes across repetitions of the baseline trials as well. No such decreasing trend was observed over the repeated baseline trials. Though we cannot rule out the possibility entirely, this does suggest that there is something happening for the typed stimuli other than decreased attentiveness.

We suggest that the decrease in excitability across repetitions of stimuli is due to sensory prediction by the motor system. When the stimuli are less predictable (i.e., the first presentations), the sensory prediction error is large, resulting in higher motor activation. As the stimuli are repeated and become more predictable, the sensory prediction error becomes lower and we observe less corticospinal excitability in the motor system. This account is consistent with Schubotz's (2007) findings of motor activation during serial prediction tasks and Wilson and Knoblich's (2005) emulator account.
If our theoretical formulation is correct, this implies that the study of motor involvement in perception and cognition should take into account that the motor system is playing multiple processing roles that are network and contextdependent. The action observation based recruitment of the motor system is well established. Strong evidence suggests that this is due to motor resonance that is both effectorspecific and time-dependent. We contend that the role of motor cortex in action-observation is for low-level activation of one's own motor repertoire. Under our account, motor activation during perceptual processing of non-human-created stimuli, reported by Schubotz and colleagues, is not at odds with the resonance account of action observation. The particular information processing role of motor regions does not need to be identical across contexts. The functional network underlying action observation includes bilateral mid-temporal gyrus (MTG) and left
inferior parietal lobule as well as left premotor cortex. (Gazzola, Aziz-Zadeh, \& Keysers, 2006). Other brain regions active during figurative language include the left and right inferior frontal gyrus (IFG), bilateral medial frontal gyri (medFG), left temporal lobe, and amygdala. (Bohrn, Altmann, \& Jacobs, 2012). The function of motor activation in each of these different networks can be defined by its connections and interactions, allowing a motor predictive system or motor simulation system when appropriate.

How would this region have multiple functional roles? Evidence from single-unit recording of neurons in premotor areas suggests that there is a wide variety of neurons that respond to different contexts. For example, during the discovery of mirror neurons, many types of such neurons were identified (Di Pellegrino et al., 1992). Some of these are called "strictly congruent" mirror neurons, which respond to action observation and action execution only to the same exact movement. More common were "broadly congruent" mirror neurons, which respond to action observation and action execution during similar types of movements, encompassing a broader response range. We postulate that the first type is responsible for driving motor resonance-related activation, while the latter type could potentially underlie the sort of sensory prediction we discuss. Finally, a third type of neuron they observed was called a "canonical neuron", which respond to the observation of manipulable objects. Perhaps these neurons could play a role in mental simulation, or affordance processing. These examples are all speculative and not grounded by any evidence in the present work, but they aim to push intuitions toward a fresh perspective. Future work using single-neuron recording would be needed to directly test such hypotheses. At a brain region level, however, we can learn more by observing how activation in local regions changes with repetition of sensory stimuli or changes in stimuli.

Future research that we are currently engaged aims to explore how sensorimotor contingencies are learned by training participants on novel sensory to motor mappings. We will then use these controlled artificial mappings to explore sequential prediction and/or simulation using the motor system.

## References

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364(1521), 1281-1289.
Bohrn, I. C., Altmann, U., \& Jacobs, A. M. (2012). Looking at the brains behind figurative language-A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. Neuropsychologia, 50(11), 2669-2683.
Candidi, M., Leone-Fernandez, B., Barber, H. A., Carreiras, M., \& Aglioti, S. M. (2010). Hands on the future:
facilitation of cortico-spinal hand-representation when reading the future tense of hand-related action verbs. European Journal of Neuroscience, 32(4), 677-683.
Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., \& Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Experimental Brain Research, 91(1), 176-180.
Gallese, V., Fadiga, L., Fogassi, L., \& Rizzolatti, G. (1996). Action recognition in the premotor cortex. Brain, 119(2), 593-609.
Gallese, V., \& Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology, 22(3-4), 455479.

Gangitano, M., Mottaghy, F. M., \& Pascual-Leone, A. (2001). Phase-specific modulation of cortical motor output during movement observation. Neuroreport, 12(7), 1489-1492.
Gazzola, V., Aziz-Zadeh, L., \& Keysers, C. (2006).
Empathy and the somatotopic auditory mirror system in humans. Current Biology, 16(18), 1824-1829.
Hauk, O., Johnsrude, I., \& Pulvermüller, F. (2004).
Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41(2), 301-307. Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., \& Rizzolatti, G. (1999). Cortical mechanisms of human imitation. Science, 286(5449), 2526-2528.
Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., \& Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. Science, 297(5582), 846-848.
Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., \& Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. Current Biology, 20(8), 750-756.
Oliveri, M., Finocchiaro, C., Shapiro, K., Gangitano, M., Caramazza, A., \& Pascual-Leone, A. (2004). All talk and no action: a transcranial magnetic stimulation study of motor cortex activation during action word production. Journal of Cognitive Neuroscience, 16(3), 374-381.
Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., \& Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey. Experimental Brain Research, 71(3), 491-507.
Umilta, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., \& Rizzolatti, G. (2001). I know what you are doing: A neurophysiological study. Neuron, 31(1), 155-165.

# A Spiking Independent Accumulator Model for Winner-Take-All Computation 

Jan Gosmann (jgosmann@uwaterloo.ca)<br>Aaron R. Voelker (arvoelke@uwaterloo.ca)<br>Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Centre for Theoretical Neuroscience, University of Waterloo<br>Waterloo, ON, Canada N2L 3G1


#### Abstract

Winner-take-all (WTA) mechanisms are an important component of many cognitive models. For example, they are often used to decide between multiple choices or to selectively direct attention. Here we compare two biologically plausible, spiking neural WTA mechanisms. We first provide a novel spiking implementation of the well-known leaky, competing accumulator (LCA) model, by mapping the dynamics onto a population-level representation. We then propose a two-layer spiking independent accumulator (IA) model, and compare its performance against the LCA network on a variety of WTA benchmarks. Our findings suggest that while the LCA network can rapidly adapt to new winners, the IA network is better suited for stable decision making in the presence of noise.


Keywords: Neural Engineering Framework; Nengo; winner-take-all; decision making; mutual inhibition; neural competition; dynamical systems

## Introduction

Winner-take-all (WTA) networks are mechanisms that select the largest value among a number of inputs. More precisely, given a $D$-dimensional vector corresponding to the non-negative utility of $D$ different choices, the desired output is positive for the dimension with highest utility (i.e., the "winner") and zero for all others. This mechanism is regularly employed as a component of cognitive models involving selective attention (e.g., Itti, Koch, \& Niebur, 1998; Standage, Trappenberg, \& Klein, 2005) and decision making, where the action with the highest utility is selected to drive behaviour (e.g., O’Reilly, 1998).

A large body of literature examines the optimality of WTA mechanisms and their consistency with neurobiological and psychological data (e.g., Bogacz, Brown, Moehlis, Holmes, \& Cohen, 2006; Gold \& Shadlen, 2007; Smith \& Ratcliff, 2004). Here, we investigate the suitability of two different WTA mechanisms in the context of neurally plausible cognitive modelling. In particular, we map each mechanism onto a network of spiking neurons, and then compare them using a set of functional benchmarks that are normative in nature. The first mechanism we consider is an implementation of the leaky, competing accumulator (LCA) model from Usher and McClelland (2001). The LCA model and variants have been widely used, for example in versions of the Temporal Context Model (Sederberg, Howard, \& Kahana, 2008), and in work on the Remote Associates Test models (e.g., Kajic, Gosmann, Stewart, Wennekers, \& Eliasmith, 2017). The second mechanism we consider is the independent accumulator (IA) model that we propose here, which involves a secondary thresholding layer that is recurrently connected to a primary integrating layer.

To implement each model, we use the Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003) to map the model's dynamics onto populations of spiking neurons. In the remainder of the paper, we provide a short introduction to the NEF, describe our implementation of the two WTA mechanisms, present our benchmarks, and finally discuss some resulting implications for cognitive modelling.

## Methods

## The Neural Engineering Framework

The Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003) is a method for mapping a cognitive model, described using mathematical equations, onto a spiking neural network. We now describe the aspects of this framework that are relevant to this work, by summarizing its three principles: representation, transformation, and dynamics.

Principle 1: Representation We define the representation of a scalar value $x(t)$ by an encoding and decoding with respect to some population of neurons. The encoding of $x(t)$ into a spike train $a_{i}(t)$ for neuron $i$ is given by:

$$
\begin{equation*}
a_{i}(t)=G_{i}\left[\alpha_{i} e_{i} x(t)+J_{i}^{\mathrm{bias}}\right] \tag{1}
\end{equation*}
$$

where $\alpha_{i}$ is a gain factor, $e_{i}$ is an encoder that determines a neuron's tuning curve, $J_{i}^{\text {bias }}$ a bias current, and $G_{i}[\cdot]$ is the neural nonlinearity. Here, we use spiking, leaky integrate-and-fire (LIF) neurons for $G_{i}[\cdot]$, and set the encoders to one. Decoding weights $d_{i}$ are then used to approximate the represented value $\hat{x}(t)$ from the activity of the population of neurons by:

$$
\begin{equation*}
\hat{x}(t)=\sum_{i} d_{i}\left[\left(a_{i} * h\right)(t)\right] \tag{2}
\end{equation*}
$$

where $h(t)=\tau_{\mathrm{s}}^{-1} \exp \left(-t / \tau_{\mathrm{s}}\right)$ is an exponentially decaying synaptic filter with time-constant $\tau_{\mathrm{s}}$, and $*$ is the convolution operator. The decoding weights are obtained by least-squares optimization of the error $E_{x}=|x-\hat{x}|$. For the transmission of a value from one population to another, the connection weights are given by:

$$
\begin{equation*}
W_{i j}=\alpha_{j} e_{j} d_{i} \tag{3}
\end{equation*}
$$

Principle 2: Transformation By finding alternate decoding weights $d_{i}^{f}$ with the error given by $E_{f(x)}=|f(x)-\hat{x}|$, arbitrary linear and nonlinear functions $f(x)$ can be approximated in the connections between neural populations.

Principle 3: Dynamics Given some desired nonlinear dynamics for the state variable $x(t)$ :

$$
\begin{equation*}
\frac{\partial x}{\partial t}=g(x) \tag{4}
\end{equation*}
$$

we can map these dynamics onto a recurrent transformation, by harnessing the synaptic filter mentioned in Principle 1. In particular, for the exponentially decaying $h(t)$ we may apply Principle 2 to the recurrent transformation $f(x)=\tau_{s} g(x)+x$ to ensure that $x(t)$ obeys Equation 4.

## Leaky, competing accumulator model

Using the principles of the NEF, we have implemented the widely-used leaky, competing accumulator (LCA) model proposed by Usher and McClelland (2001). The dynamics (see Fig. 1a) for each state variable $x_{i}(t), 1 \leq i \leq D$, where $D$ is the number of choices, are given by:

$$
\begin{equation*}
\frac{\partial x_{i}}{\partial t}=\left(\rho_{i}-k x_{i}-\beta \sum_{j \neq i} x_{j}\right) \frac{1}{\tau}, \quad x_{i} \geq 0 \tag{5}
\end{equation*}
$$

where $\rho_{i}$ is each external input, $k$ is the leak rate, $\beta$ the lateral inhibition, and $\tau$ the integration time-constant. This model essentially integrates each input $\rho_{i}$ with a leak term $\left(-k x_{i}\right)$, minus competition from every other variable $\left(\beta \sum_{j \neq i} x_{j}\right)$. Supposing $\rho_{i}>\rho_{j}$ for all $j \neq i$, a WTA mechanism should indicate that $i$ is the winning choice. Setting $k=\beta=1$ will guarantee that the winning state $x_{i}$ converges to the value of the largest input $\rho_{i}$, while each losing state $x_{j}(j \neq i)$ converges to zero. Other choices of $k$ merely alter the effective $\tau$ and the effective gain on the input, while other choices of $\beta$ will produce unwanted behaviours (see supplementary analysis).

We implement Equation 5 with the NEF by using one population of neurons for each $x_{i}$, and applying Principle 3 to each population. By appropriately selecting the gain and bias parameters from Principle 1, we ensure that each state variable is rectified $\left(x_{i} \geq 0\right)$ as required. We believe this implementation is novel as it does not interpret each $x_{i}$ as a distinct neural firing rate, but rather as a population-level representation distributed across any number of spiking neurons. In effect, heterogeneous spike trains are weighted by optimal decoding weights to precisely implement the stated dynamics. This allows us to attain greater biological realism without altering the dynamics prescribed by Equation 5.

## Independent accumulator model

The other WTA mechanism that we investigate is our proposed independent accumulator (IA) model (see Fig. 2). We use the term 'independent' to refer to the fact that there are no direct interactions between each accumulator, unlike in the LCA model which has direct competition between states. To enable a form of competition, we add a second thresholding layer that projects back to self-excite and mutually-inhibit the first layer. We now provide the details of each layer - again implemented using the principles of the NEF.


Figure 1: Example time course of the state variables in the LCA (top) and IA (bottom) networks with three choices $(D=3)$. The vector of inputs is $[0.8,0.7,0.6]$.

The first layer consists of a separate integrator population (i.e., accumulator) for each state variable $x_{i}(t), 1 \leq i \leq D$. The second layer consists of independent, non-recurrent populations that receive input from the first layer in a one-toone fashion. From the second layer, we decode the function $\bar{x}_{i}:=\Theta\left(x_{i}-\vartheta\right)$ where $\Theta$ is the Heaviside function, and $\vartheta=0.8$ is a fixed threshold value that determines how much evidence needs to be accumulated to produce an output. The Heaviside decoded output of layer 2 projects back to layer 1 to add $\bar{x}_{i}-\bar{\beta} \sum_{j \neq i} \bar{x}_{j}$ to the input of $x_{i}$. Since intuitively the largest input will accumulate the fastest, once this reaches the threshold $\vartheta$ it will self-excite and inhibit all other state variables. Fixing $\bar{\beta}=2$ ensures that the losing state variables will


Figure 2: Independent accumulator (IA) network. Neural populations are denoted by circles labelled with their represented state variable. Arrows denote excitatory connections, while lines ending in circles denote inhibitory connections. The second layer computes $\bar{x}_{i}:=\Theta\left(x_{i}-\vartheta\right)$.
go to zero (see supplementary analysis). This is summarized more precisely by the following dynamics (see Fig. 1b):

$$
\begin{equation*}
\frac{\partial x_{i}}{\partial t}=\rho_{i} \frac{1}{\tau_{1}}+\left(\bar{x}_{i}-\bar{\beta} \sum_{j \neq i} \bar{x}_{j}\right) \frac{1}{\tau_{2}}, \quad x_{i} \geq 0 . \tag{6}
\end{equation*}
$$

Notably, this takes the form of Equation 5 after substituting $\tau=\tau_{1}, k=-\tau_{1} / \tau_{2}$, and $\beta=\bar{\beta} \tau_{1} / \tau_{2}$, with the only remaining difference being the Heaviside nonlinearity applied to the state feedback. Thus, in contrast to the continual competition occurring in the LCA model, the threshold $\vartheta$ is a free parameter that controls how much evidence needs to be integrated before enabling any competition between states. Instead of directly manipulating $\vartheta$, we opt to change $\tau_{1}$, which has a comparable effect due to the leak-free integration.

Note that the decoded output from layer 2 of the IA network has higher variance than the LCA network (Fig. 1), but the separation of the output for different choices is more relevant than the variance in interpreting the output.

## Benchmarks

To test and compare the two WTA mechanisms we provide an input of $\rho_{i}=u-s\left(1-\delta_{1 i}\right)+\eta_{i}$, where $u$ is the magnitude of the largest input, $s>0$ is the target separation relative to all other inputs, $\delta$ is the Kronecker delta, and $\eta_{i}$ is Gaussian (white) noise with standard deviation $\sigma$. Thus, without loss of generality, the first state variable receives the largest input $u$ plus noise, and all other state variables receive a noisy input that is smaller by $s$. It is important to note that $u$ not only determines the size of the largest input, but also the general baseline of inputs. Since all of the runner-ups have equal magnitude, this represents the most difficult scenario for the networks, where all potential choices must be considered. As $s \rightarrow 0$ the problem also becomes more difficult because the utilities of the choices are closer together. We use $u=1$ unless indicated otherwise, and set the number of choices to $D=10$. Furthermore, we increment the noise variance to highlight successes and failures as the task becomes increasingly difficult with more noise. This allows us to determine which functions are performed robustly by each network. In both WTA models we use 200 neurons per choice. In the IA network this is split into 150 neurons for each layer 1 population and 50 neurons for each layer 2 population. All remaining network parameters are summarized in Table 1.

To evaluate the two mechanisms on the previously defined input, we use a number of separate metrics. First, we determine whether the model is able to form a clear decision within one second. To be counted as 'clear', at least one output must remain above 0.15 across the time interval $(1 \mathrm{~s}, 2 \mathrm{~s}$ ] while all other outputs remain below this threshold. This lower bound of 0.15 was chosen to ensure that noise on a zero output is not mistaken for a non-zero output. Note that this metric requires that the decision does not change throughout the time interval. This does not take into account whether the winning output actually corresponds to the largest input. However, for

Table 1: Summary of parameter values.

| LCA time-constant | $\tau=0.1 \mathrm{~s}$ |
| :--- | :--- |
| LCA recurrency parameters | $k=\beta=1$ |
| IA accumulation time-constant | $\tau_{1}=0.1 \mathrm{~s}, 0.5 \mathrm{~s}$ |
| IA feedback time-constant | $\tau_{2}=0.1 \mathrm{~s}$ |
| IA threshold | $\vartheta=0.8$ |
| IA recurrency parameters | $\bar{\beta}=2$ |
| Recurrent synaptic time-constant | $\tau_{\mathrm{s}}=0.1 \mathrm{~s}$ |
| Feed-forward synaptic time-constant | $\tau_{\mathrm{s}}=0.005 \mathrm{~s}$ |
| Output decoding synaptic time-constant | $\tau_{\mathrm{s}}=0.01 \mathrm{~s}$ |

some models it is more desirable to produce a clear incorrect decision than an unstable incorrect decision. In other situations, though, the correctness of the decision may be of higher importance. Thus, we consider a trial 'correct' if the model forms a clear decision, and the largest output corresponds to the true largest input. Measurement of correct trials forms our second benchmark.

We use two additional benchmarks on the set of all trials with a clear decision. First, it is important to consider the speed at which the network can make decisions. We therefore define the 'decision time' as the length of time it takes to fulfil the conditions of a clear decision. Second, the correctness metric only considers the final averaged output during a time interval. It is possible for a network to produce transient outputs before the final decision is reached. In the context of a larger model, this can become a problem because the transient output might be prematurely interpreted as a decision. Thus, we define it as the 'highest output of a losing choice' during the whole simulation.

## Results

We find that the ability to form a clear decision of the LCA network decreases with more noise and less target separation (see Fig. 3). Also, the magnitude of the inputs has an important influence. For a small inputs with $u=0.2$ the winner will mostly not exceed the 0.15 threshold with noise present. The best performance is achieved with medium inputs with $u=0.6$. Higher inputs make it more likely that runner-ups will exceed the 0.15 threshold, especially with a small target separation. In contrast to the LCA network, the IA network manages to form a clear decision in every trial (not explicitly shown in Figure 3, but all data points fall on the grey line).

Interestingly, for all clear decisions the correct winner was determined by the LCA network. Thus, a plot of correct trials looks identical to Figure 3, with slightly different error bars. While always reaching a clear decision, the decisions of the IA network are not always correct. Overall, the IA performance tends to be worse than the LCA performance for a high input magnitude, but better for smaller inputs (Fig. 4a, b). We can greatly improve the IA performance by increasing $\tau_{1}$ to 0.5 s which slows down the integration of evidence (Fig. 4c). This improves the IA performance to be above the LCA performance for high baselines, but it will also increase the de-


Figure 3: Fraction of trials with a clear decision for the LCA network (which also exactly matches the fraction of correct trials). Each plot shows data for a different input magnitude $u$. Error bars denote bootstrapped $95 \%$ confidence intervals. The grey horizontal line indicates the optimum, which coincides with the performance of the IA network.
cisions times of the IA network. For a low baseline with $u=0.2$, it makes the IA network unable to decide within the allocated time frame, but given more time it would still reach a decision.

As shown in Fig. 5a, the time required to reach a decision in the LCA network depends mostly on the size of inputs and target separation. We averaged over the different noise conditions because the noise influence on the timing was minor. In the IA network the largest input is the most important factor (dashed vs. solid lines in Fig. 5b). Depending on this magnitude, the network can either be faster or slower than the LCA network, but it will need more time given a value of $\tau_{1}$ that achieves the same fraction of correct responses as the LCA network. There is also a slight influence of target separation and input noise, with an interaction of these two parameters (solid lines in Fig. 5).

Finally, looking at the transient responses indicates that both models might produce outputs of losing choices. For the LCA network the magnitude of the transient response mainly increases with the amount of noise (Fig. 6a). For the IA network, transient outputs are smaller in noisy conditions, but can be higher than for the LCA network in less noisy conditions with small target separations. The magnitude of such transient responses is reduced by adjusting $\tau_{1}$ to 0.5 , at the cost of a slower decision.

## Discussion

We have shown that neither network performs better on all benchmarks, but rather each is better suited for different purposes. For instance, the LCA network can determine the correct winner more quickly, and, given that a conclusive decision was made, it always selects the correct winner. However, under noisy conditions it may fail to produce a clear output at all, and thus not make a decision. This can be problematic for cognitive models that must form a clear decision (even when it may be incorrect). The IA network might not be able to identify the correct winner as quickly or as reliably (depending on the choice of $\tau_{1}$ ), but given enough time it will eventu-
ally arrive at a decision and produce a clear output. This is a direct consequence of the thresholding on the state feedback, which prevents competition from occurring until a sufficient amount of evidence has accumulated. This also enables the IA network to react to small inputs. In addition, the IA network is easily extendible to allow dynamic control of the decision speed, by supplying an external bias to layer 2 to adjust the $\vartheta$ threshold.

The LCA network is especially well-suited for situations where a decision needs to be continuously adjusted. The dynamics constantly push the winning state variable to the magnitude of the largest input, while adapting to input changes along a time-scale of $\tau$. This makes it quick to respond to changes in the input for smaller $\tau$, but leads to randomly switching outputs due to noise.

In contrast, the IA network is better suited for situations where a discrete sequence of decisions is required. After selecting a winner, the model's decision will persist due to selfexcitation, even in the absence of input. Thus, after making a decision, it is necessary to reset the model by inhibiting the winning accumulator. This limits how quickly successive decisions can be made and reduces the ability to react to changing inputs. However, once a decision is made, the network provides a stable output. For example, we intend to use the IA network in a model of free recall to output a sequence of WTA responses. In this case, the model requires stable recall in the presence of noise, even if each response may not be the "true" winner.

Both models might produce a transient response that may be interpreted as a decision, which can make the detection of decisions problematic. This is of special relevance when incorporating the networks into larger cognitive models. For the LCA network, the transient behaviour is inherent to its design; there will always be an initial rise of all state variables (that receive input) before the mutual inhibition grows strong enough to push them back to zero. In the IA network, however, such transient responses may be avoided by choosing appropriate $\tau_{1}$ and $\vartheta$. It should be noted that other recur-


Figure 4: Fraction of correct trials for the IA network. Each plot shows data for a different combination of input magnitude $u$ and integration time-constant $\tau_{1}$. Error bars denote bootstrapped $95 \%$ confidence intervals. The grey horizontal line indicates the optimum.


Figure 5: (Left) Mean decision times for the LCA network. Shown data is averaged across all noise levels, since noise had minimal effect on decision times. (Right) Mean decision times for the IA network with input magnitude $u=0.2$ (solid) and $u=1$ (dashed). Error bars denote bootstrapped $95 \%$ confidence intervals.
rently connected readout mechanisms have been proposed for the two choice case that produce bursting outputs somewhat comparable to the IA network (Lo \& Wang, 2006). However, the evidence integration in the Lo and Wang (2006) model uses competition via pooled inhibition, making it more similar to the LCA then the IA network.

We have not yet investigated the agreement of these mechanisms with neurobiological and behavioural data, although this has been done before for other WTA networks (Gold \& Shadlen, 2007; Smith \& Ratcliff, 2004). The implementation in spiking neurons, however, provides some basic biological plausibility and more readily permits comparisons with neural data. In particular, the IA network predicts that the firing rates for neurons in the first layer will rise up to a threshold, and that neurons in the second layer will not become active or inhibit the first layer until reaching this threshold. Neurons in the macaque lateral intraparietal area exhibit a similar step response (Latimer, Yates, Meister, Huk, \& Pillow, 2015). In contrast, the LCA network predicts that the firing of any neurons will proportionately inhibit all other neurons that do not represent the same state. With regard to behavioural data, different effects for decision times are predicted as the number
of choices increases: for the LCA network decisions become slower due to the mutual interaction, but for the IA network decisions will take less time because only a single accumulator needs to exceed the threshold. Nevertheless, evidence for one network does not exclude the possibility that the other network is employed for different tasks or by different brain areas. Relatedly, it might be more plausible to distribute the representation of all state variables over a single population of neurons. This is directly supported by the NEF, but we chose to leave this to future work to keep the current analysis free from potential interactions of the state variables introduced by such a distributed representation.

We also did not look at at the influence of the number of dimensions $D$ in detail. For higher $D$, we see reduced accuracy overall since each additional choice has a baseline chance to win due to noise. Nevertheless, the results that we discuss here are qualitatively similar.

One critique of non-leaky accumulator models is that their ability to discriminate the largest input increases indefinitely with time (Usher \& McClelland, 2001) and that there is no sensible stopping criterion. However, this assumes that the time to reach a decision has no cost. If time-to-decision has a


Figure 6: Transient response (highest output of losing choices) for the LCA and IA network. Error bars denote bootstrapped 95\% confidence intervals. The grey horizontal lines show the optimum.
cost, then it will at some point exceed the gain achieved from making a correct decision. Furthermore, this argument assumes integration with perfect accuracy. But, with networks built using the NEF, the representation of each state variable has limited precision, and so an ideal trade-off must be found.

To conclude, we investigated two spiking neural networks computing a winner-take-all function based on the leaky, competing accumulator model and a novel two-layer independent accumulator model. While both perform the same basic tasks, they fail in different ways as each task scales in difficulty via increased noise or less separation between choices. From a modelling perspective, this makes each network more useful for different situations. The LCA model is better for continuous updating of decisions, whereas the IA network is better suited for more discrete decisions in the presence of noise. It is left to future work to investigate whether these two distinct mechanisms can be identified from either behavioural or neurophysiological data.

## Notes

Source code and supplementary analysis are available at https://github.com/ctn-waterloo/cogsci17-decide.

## Acknowledgments

This work was supported by the Canada Research Chairs program, the NSERC Discovery grant 261453, Air Force Office of Scientific Research grant FA8655-13-1-3084, CFI, OIT, and NSERC CGS-D.

## References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., \& Cohen, J. D. (2006, October). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review, 113(4), 700-765. doi: 10.1037/0033-295X.113.4.700
Eliasmith, C., \& Anderson, C. H. (2003). Neural engineering: Computation, representation, and dynamics in neurobiological systems. Cambridge, MA: MIT Press.

Gold, J. I., \& Shadlen, M. N. (2007). The neural basis of decision making. Annual Review of Neuroscience, 30(1), 535-574. doi: 10.1146/annurev.neuro.29.051605.113038
Itti, L., Koch, C., \& Niebur, E. (1998, November). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254-1259. doi: 10.1109/34.730558
Kajic, I., Gosmann, J., Stewart, T. C., Wennekers, T., \& Eliasmith, C. (2017). A spiking neuron model of word associations for the Remote Associates Test. Frontiers in Psychology, 8(99). doi: 10.3389/fpsyg.2017.00099
Latimer, K. W., Yates, J. L., Meister, M. L. R., Huk, A. C., \& Pillow, J. W. (2015, July). Single-trial spike trains in parietal cortex reveal discrete steps during decisionmaking. Science, 349(6244), 184-187. doi: 10.1126/science.aaa4056
Lo, C.-C., \& Wang, X.-J. (2006, July). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nature Neuroscience, 9(7), 956-963. doi: 10.1038/nn1722

O'Reilly, R. C. (1998, November). Six principles for biologically based computational models of cortical cognition. Trends in Cognitive Sciences, 2(11), 455-462. doi: 10.1016/S1364-6613(98)01241-8

Sederberg, P. B., Howard, M. W., \& Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. Psychological Review, 115(4), 893-912. doi: 10.1037/a0013396

Smith, P. L., \& Ratcliff, R. (2004, March). Psychology and neurobiology of simple decisions. Trends in Neurosciences, 27(3), 161-168. doi: 10.1016/j.tins.2004.01.006
Standage, D. I., Trappenberg, T. P., \& Klein, R. M. (2005, July). Modelling divided visual attention with a winner-take-all network. Neural Networks, 18(5-6), 620-627. doi: 10.1016/j.neunet.2005.06.015

Usher, M., \& McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review, 108(3), 550-592. doi: 10.1037/0033-295X.108.3.550

# Folk intuitions about consciousness 

Katharina Graben<br>Justus Liebig University, Giessen, Hessia, Germany<br>Kai Hamburger<br>Justus Liebig University, Giessen, Hessia, Germany

Markus Knauff
Justus Liebig University, Giessen, Hessia, Germany


#### Abstract

In science and philosophy, there is still no general agreement on what 'consciousness' is. But how do normal people (with no education in psychology or philosophy) use the term in their everyday life? What is the folk understanding of the word "conscious"? We conducted an online study on how the general public uses the word "consciousness" in their daily life. Participants ( $\mathrm{n}=445$ ) answered the question "What is consciousness?" in four different formats: they (1) generated free definitions in their own words, (2) generated many synonyms, (3) generated one synonym, or (4) selected one alternative description in a multiple choice task. The most frequent words were: alertness, clarity, I-sensation, knowledge, perception, reflecting, and thinking. The word perception was provided most often across all formats. There was also a high correlation between all response formats. We discuss these findings and their implications for the scientific study of consciousness.


# Pragmatic aspects of spatial language acquisition and use across languages 

Myrto Grigoroglou (Mgrigor@Udel.Edu)<br>Department of Linguistics and Cognitive Science, University of Delaware<br>125 East Main Street, Newark, DE 19716 USA

Megan Johanson (Megan.johanson@Gmail.Com)<br>Department of Psychological and Brain Sciences, University of Delaware<br>105 The Green, Newark, DE 19716

Anna Papafragou (APapafragou@Psych.Udel.Edu)<br>Department of Psychological and Brain Sciences, University of Delaware 105 The Green, Newark, DE 19716


#### Abstract

Across languages, back is produced earlier and more frequently than front. This asymmetry has been attributed either to a conceptual/semantic asymmetry in the early meanings of these locatives (with back being more basic than front; conceptual immaturity account) or to the fact that Back configurations are inherently more 'noteworthy' than Front configurations (pragmatic account). Here, we tested the two accounts. In Study 1, children and adult speakers of English and Greek described Front/Back motion events. In Study 2, adult speakers of 10 additional languages described the same events. Despite cross-linguistic differences, speakers of all age and language groups typically used more Back than Front adpositions; furthermore, they often encoded Back information in occlusion verbs (e.g. hide) but no such verbs were available for Front. Thus, the front/back asymmetry is not due to children's conceptual immaturity but should be linked to pragmatic factors that also shape adult spatial language production cross-linguistically.


Keywords: front; back; motion events; spatial cognition; language production; pragmatics; theories of acquisition

## Introduction

It is widely recognized that children acquire spatial locatives in a consistent order cross-linguistically (e.g., E. Clark, 1980; E. Clark, 1977; Johnston \& Slobin, 1979; Parisi \& Antinucci, 1970). In many cases, patterns of language use in children, especially when these emerge cross-linguistically, have been argued to point to shared (possibly universal) conceptual asymmetries in underlying representations (Bowerman, 1996). For instance, the early emergence of prepositions such as in and on has been considered to reflect the early development of the notions of containment and support (Johnston \& Slobin, 1979; Piaget \& Inhelder, 1967).

However, it remains unclear whether patterns of spatial language use in children can be attributed merely to conceptual factors. For instance, unlike their positive counterparts in and on, 'negative' containment and support prepositions such as out and off are used extremely infrequently by children to mark locations, although, in principle, both 'positive' and 'negative' prepositions should present the same level of conceptual difficulty for the learner (compare The egg is in the cage vs. The egg is out of the cage; Papafragou, Viau, \& Landau, 2013). By contrast, a pragmatic explanation seems more adequate to account for these facts: 'negative' prepositions are used less frequently because their
informational contribution is low (they do not specify where something is) unless they can be interpreted as indicating a change of location (The bird is out of the cage is more felicitous than The egg is out of the cage; ibid.). Even though it is often acknowledged that both conceptual and pragmatic factors shape the way spatial language is used and acquired (e.g., E. Clark, 1973; Levinson \& Wilkins, 2006), the exact contribution of each factor remains open. The objective of the current study is to contribute to this debate.

## Case study: the acquisition of front and back

The acquisition of the locatives front and back presents a particularly good domain to explore the division of labor between pragmatic and conceptual explanations of spatial language acquisition and use. A number of studies have shown that, across different languages, the locative back is produced earlier and appears more frequently in children's speech than the locative front (Johnston, 1984; Johnston \& Slobin, 1979; Kubena, 1968). Some researchers have suggested that this asymmetry should be attributed to the conceptual immaturity underlying children's early representations of the relations front and back. According to this view, although the adult meanings of front/back are geometric and semantically symmetrical ("first/second in-line-of-sight"), children's early meanings are immature, function-based and asymmetrical: front means "visible" and back means "occluded". This conceptual/semantic asymmetry in the early meanings of Front and Back results in the asymmetric acquisition of the locatives (Johnston, 1984; Johnston \& Slobin, 1979).

Other researchers have argued that the asymmetry in the acquisition of front and back should be attributed to the pragmatics of these spatial expressions (Hill \& Vandeloise, 1991; Tanz, 1980). On this view, occlusion and visibility characterize typical Back and Front configurations, with occlusion (the functional corollary of Back) being, typically, more informative than visibility (the functional corollary of Front). Thus, children use back more frequently than front because the communicative need to mark that an object is occluded is (in most cases) more pressing than the need to mark that an object is visible.

The two explanations converge on the view that children should encode Back more frequently than Front but, crucially, they differ on whether adults should also exhibit a
similar asymmetry. If the asymmetry is due to early, immature meanings for the prepositions front and back (i.e., visibility and occlusion), as posited by the conceptual immaturity account, adults, having mature spatial semantics, should use the two terms equally frequently. By contrast, if the front/back asymmetry is driven by the inherent 'noteworthiness' of occlusion, as proposed by the pragmatic account, adults-just like children-might also exhibit the asymmetry.

Furthermore, the two accounts share the assumption that the meanings of the prepositions front and back involve the notions of visibility and occlusion, but they disagree on whether adults should entertain such meanings. On the conceptual immaturity account, adults (unlike children) should not have a bias to mark occlusion (as opposed to visibility) with expressions beyond front and back (e.g., with verbs such as hide, etc.) because such function-based representations should characterize only the early immature child semantics. By contrast, on the pragmatic account, adults-just like children-might also show a bias to mark occlusion.

## Current study

In the current study we test the predictions of the conceptual immaturity and pragmatic explanations. In Study 1, we elicit descriptions of Front/Back motion events from children and adult speakers of two typologically distinct languages (English and Greek). In Study 2, adult speakers of 10 additional languages describe the same events.

To evaluate the predictions of the conceptual immaturity and pragmatic explanations we compare Front/Back motion descriptions in both children and adults. Although previous research, surprisingly, did not include adults, adult data, especially if they represent a wide cross-linguistic sample, are crucial for validating theories of spatial language acquisition and use.

Furthermore, we look at elicited descriptions of motion paths. Unlike prior work on locative front/back, the choice of motion paths allows us to compare not only the use of 'front' and 'back' adpositions (prepositions and postpositions) across the languages in our sample but also the use of functional information (occlusion/visibility) encoded in verbs (see Landau \& Jackendoff, 1993; Miller \& JohnsonLaird, 1976; Talmy, 1983, on the role of spatial verbs). To ensure cross-linguistic validity, we include both satelliteframed languages that tend to encode motion paths in particles/non-verb elements, and verb-framed languages that tend to encode motion path information in the verb (Talmy, 1985). In study 1, we compare English (a satellite-framed language) to Greek (a verb-framed language). In study 2, our language sample contains an equal number of verb-framed and satellite-framed languages. For completeness in path descriptions, each Front or Back path has a goal variant (e.g., figure moving in front of/ behind the reference object) and a source variant (e.g., figure moving from front of/ from behind the reference object). In the case of Back paths, figures undergo a dynamic change of state from visibility to
occlusion (for goals) or vice versa (for sources). In the case of Front paths, figures undergo a change of location without a change of visibility (e.g., when figure X moves to/from front of $\mathrm{Y}, \mathrm{X}$ moves along a trajectory while being continuously visible).

## Study 1

## Methods

Participants Participants were 40 native English speakers and 40 native Greek speakers. They fell into two age groups (Children, Adults) with 20 participants in each age group for each language. The English-speaking children were recruited at daycares in Newark, DE and ranged between the ages of $3 ; 8$ and $5 ; 5(\mathrm{M}=4 ; 6)$. The English-speaking adults were undergraduate students at the University of Delaware and received course credit for their participation. The Greekspeaking children were recruited at daycares in Evia, Greece and ranged between the ages of $3 ; 9$ and $5 ; 3(M=4 ; 6)$. The Greek-speaking adults were University students or young professionals and were recruited in Evia and Athens, Greece.

Materials The stimuli consisted of a total of 48 dynamic motion events presented in Microsoft PowerPoint. Each event consisted of a Figure, which was always the same soccer ball, and a Reference object, which was selected from a set of simple, abstract 3D objects. We chose to use very simple schematic stimuli to elicit only or mainly path information (even from speakers of a language such as English which regularly encodes manner of motion).

The motion events depicted a total of eight different spatial relations, each with a source and a goal version. This battery was part of a larger project investigating cross-linguistic descriptions of motion paths. In the present study, we focus on Front (IN FRONT OF/FROM IN FRONT OF) and Back (BACK/FROM BACK). Stimuli depicting the Back relation always involved occlusion but stimuli depicting the Front relation never did (see Fig.1).


Figure 1: Schematic examples of test events. In FRONT/FROM FRONT scenes, the ball is always visible but in BACK/FROM BACK scenes it is occluded at the endpoint (goal scenes) or the beginning (source scenes) of the event.

The remaining relations were Containment (IN/OUT OF), Cover (UNDER/FROM UNDER), Support (ONTO/OFF OF), Contact (TO/FROM), Vertical Proximity (TOWARDS THE SIDE OF/AWAY FROM THE SIDE OF), and Horizontal Proximity (TOWARDS THE TOP OF/AWAY FROM THE TOP OF).

Six events were shown for each relation ( 3 exemplars, each with a goal and a source version). The source and goal versions of the same exemplar were identical except for the color of the Reference object and the direction of the motion path. The motion events lasted for three seconds and then the end state of the event remained on the screen until a key was pressed.

Procedure The adult participants were told that they would see a series of motion events involving a ball and another "toy." After viewing each event, the participants had to describe, in their native language, what the ball did in each event. Events remained on the screen until a key was pressed. The adult participants performed one practice trial.
Children were told that they were going to play a game where animals play with balls and "toys." They were then shown a screen with all Reference objects, and were told to call them all "toys." In order to help children maintain attention, a slide with a small cartoon animal in one of the bottom corners was presented before each motion event. The children's attention was drawn to the animal by the experimenter saying "Look at the (animal)! Let's see what the (animal)'s ball will do!". The motion clip was then played and remained on the screen; then the experimenter asked the child to describe what the animal's ball did. The children completed at least three practice trials before beginning the experiment.

Coding Each linguistic description for the Front and Back relations was first coded for the presence of a target preposition that had to correspond to the type of scene (goal/source). For Front (goal scenes), the target prepositions were in front of/to front of in English and brosta apo/brosta sto 'in front of' in Greek. For Front (source scenes), target prepositions included from (in) front (of) in English and (apo) brosta (apo) '(from) front (of)' in Greek. For Back (goal scenes), the target prepositions included behind, to back/in back in English and piso (apo)/apo piso 'behind' in Greek. For Back (source scenes), the target prepositions included from behind in English and apo piso (apo) 'from behind (of)' in Greek.

The linguistic descriptions were also coded for the presence of spatial expressions of visibility or occlusion. For the Front relation, there were no expressions encoding visibility. This fact is highly significant, and we return to it in the Results section. For the Back relation, we coded predominantly appearance/disappearance verbs that encoded occlusion (or, more accurately, a change of state from or to occlusion): disappear and hide (goal scenes), appear, emerge (source scenes) in English, and hanome 'disappear', krivome 'hide' (goal scenes) and emfanizome 'appear', apokaliptome 'reveal oneself' (source scenes) in Greek.

Finally, all linguistic descriptions of Front and Back relations were coded in terms of the total target spatial information they contained (i.e., target preposition or occlusion expression). This was done because there was often overlap in the use of target prepositions and other expressions to describe an event (e.g., in Greek I bala krivete piso apo to pehnidi 'the ball is hiding behind the toy'), so analyzing each separately might not accurately represent the way Front and Back relations are linguistically represented.

## Results and discussion

In three separate analyses, we test the competing predictions of the conceptual immaturity and the pragmatic account in terms of (a) the use of front/back prepositions, (b) the use of a broader set of visibility/occlusion expressions, (c) the use of the devices in (a) and (b) combined.

## Use of front vs. back prepositions

Beginning with prepositions, we performed a mixed $2 \times 2 \times 2$ ANOVA with Relation (Front, Back) as a within subjects factor, Age (children, adults) and Language (English, Greek) as between subjects factors, and the proportion of target prepositions as the dependent variable. The analysis yielded a significant main effect of Relation $(F(1,76)=11.29, p=$ $.001, \eta^{2}=.13$ ): participants, overall, mentioned Back prepositions more frequently than Front prepositions ( $\mathrm{M}_{\mathrm{F}}=$ $\left..35, \mathrm{M}_{\mathrm{B}}=.45\right)$. The analysis also yielded a main effect of Age $\left(F(1,76)=70.65, p<.001, \eta^{2}=.48\right)$ : unsurprisingly, adults used more prepositions than children ( $\mathrm{M}_{\mathrm{CH}}=.17, \mathrm{M}_{\mathrm{AD}}=.65$ ). Finally, the analysis returned a marginally significant effect of Language $\left(F(2,76)=3.55, p=.063, \eta^{2}=.05\right)$ in the expected direction: English speakers exhibited a small tendency to use more target prepositions than Greek speakers $\left(\mathrm{M}_{\mathrm{ENG}}=.50, \mathrm{M}_{\mathrm{GR}}=.29\right)$. The ANOVA did not show any other effects or interactions.

## Use of visibility vs. occlusion terms

We moved beyond front/back prepositions in the two languages under study to consider a broader set of expressions encoding visibility for Front and occlusion for Back relations. As already mentioned, across ages and languages, there was a great variety of expressions encoding occlusion in Back scenes in the present dataset, but no expressions encoding visibility in Front scenes. We, thus, analyzed only expressions marking occlusion (since visibility was not encoded). A two-way factorial ANOVA, with the proportion of occlusion expressions as the dependent variable and Age and Language as factors, returned a main effect of Language $\left(F(1,76)=10.51, p=.002, \eta^{2}=.12\right)$ : because the occlusion expressions were mainly verbs, Greek speakers used occlusion expressions more frequently than English speakers $\left(\mathrm{M}_{\mathrm{ENG}}=.38, \mathrm{M}_{\mathrm{GR}}=.60\right)$. Crucially, the ANOVA did not yield an effect of Age $(F(2,76)=0.87, p=.769$, n.s. $)$ : adults used occlusion expressions as frequently as children.

## Use of total Front vs. Back information

Finally, to explore the predictions of the conceptual immaturity and pragmatic accounts at a more comprehensive level of spatial encoding, we analyzed the proportion use of total spatial information (target prepositions and occlusion expressions) to mark Front and Back relations. We conducted a mixed $2 \times 2 \times 2$ ANOVA with Relation (Front, Back) as a within subjects factor, Language (English, Greek) and Age (children, adults) as between subjects factors, and the proportion of total spatial information as the dependent variable (see Fig.2). Results yielded a significant main effect of Relation $\left(F(1,76)=137.42, p<.001, \eta^{2}=.64\right)$ and a main effect of Age $\left(F(1,76)=52.19, p<.001, \eta^{2}=.41\right)$, qualified by an Age by Relation interaction $(F(1,76)=11.53, p=.001$, $\eta^{2}=.13$ ). T-tests within each relation revealed that the interaction was due to the fact that although adults used a higher amount of spatial information for both the Front and Back relations than children ( $p \mathrm{~s}<.001$ ), this difference was smaller for the Back relation $\left(\mathrm{M}_{\text {FRONT_DIFF }}=.47 \mathrm{vs}\right.$. $\mathrm{M}_{\text {BACK_DIFF }}=.22$ ). Importantly for present purposes, spatial information was used more frequently to encode Back compared to Front relations by both age groups (children: $t(39)=-10.33, p<.001$; adults: $t(39)=-6.22, p<.001)$.

Total spatial information


Figure 5: Proportion of total spatial information given by English and Greek speakers for the Front and Back relations. Error bars represent standard error.

Overall, these results show that both children and adults encode Back information more frequently than Front information at the three levels of spatial encoding. The analysis of Front and Back prepositions showed an asymmetry in the use of 'front'- and 'back'-denoting prepositions in both age and language groups. Furthermore, the asymmetry generalized to a broader range of occlusion/visibility expressions: Back scenes often elicited expressions encoding the change to or from occlusion (e.g., verbs denoting appearance/disappearance), and these expressions were used equally frequently by adults and children. By contrast, Front scenes did not elicit any expressions encoding visibility in any age or language group. Finally, an examination of the total spatial information offered in Front and Back paths confirmed the conclusion that
asymmetries in encoding the two types of path are largely informational: both children and adult speakers of English and Greek marked Back paths more frequently than Front paths. These findings are in accordance with the pragmatic hypothesis, which allows for the asymmetry to be present in various age groups and at any level of spatial encoding.

## Study 2

In Study 2, we tested 14 native speakers of 10 languages (Cantonese, Dhivehi, German, Javanese, Korean, Pashto, Malay, Spanish, Swahili, Turkish) on a paradigm almost identical to that in Study 1. We included both satellite-framed languages that, like English, tend to encode motion paths in particles/non-verb elements, and verb-framed languages that, like Greek, tend to encode motion path information in the verb (Talmy, 1985). Our new sample was split almost evenly between these two language types (Spanish, Turkish, Korean, and Swahili are verb-framed, German, Cantonese, and Javanese are satellite-framed, and Dhivehi, Malay, and Pashto are of unknown type).

## Methods

Participants Native speakers of 9 languages (Cantonese, Dhivehi, German, Korean, Malay, Pashto, Spanish, Swahili, Turkish) were recruited from the graduate student population of the University of Delaware. All students were proficient in English as well as their native language and had spent on average 5 years in the US. Data from one additional language (Javanese) were collected at a site abroad (Jakarta, Indonesia); see Table 1 for all 10 languages and language families. One to two informants from each language were tested. The average age of the informants was 26 years. Participants received a $\$ 10$ gift certificate as compensation for their participation.

Table 1: Languages sampled in the cross-linguistic survey (with number of participating speakers), language families, countries of origin and typological classification in the motion domain

| Language | Language Family | Country | Motion Typology |
| :---: | :---: | :---: | :---: |
| Cantonese $(\mathrm{n}=1)$ | Sino-Tibetan | China | S-Framed |
| Dhivehi ( $\mathrm{n}=1$ ) | Indo-Aryan/IndoEuropean | Maldives | Unclassified |
| German ( $\mathrm{n}=2$ ) | Indo-European | Germany | S-Framed |
| Indonesian/ <br> Malay ( $\mathrm{n}=2$ ) | Austronesian | Malaysia | Unclassified |
| Javanese $(\mathrm{n}=1)$ | Austronesian | Central Java | S-Framed |
| Korean $(\mathrm{n}=1)$ | Altaic | Korea | V-Framed |
| Pashto $(\mathrm{n}=1)$ | Indo-Iranian/IndoEuropean | Pakistan | Unclassified |
| Spanish ( $\mathrm{n}=2$ ) | Indo-European | Mexico, Columbia | V-Framed |
| Swahili $(\mathrm{n}=1)$ | Niger-Congo | Tanzania | V-Framed |
| Turkish $(\mathrm{n}=2)$ | Altaic | Turkey | V-Framed |

Materials The same motion events as in Study 1 were used but with one additional event for each relation (shown in both a source and goal version) for a total of 64 stimuli.

Procedure The procedure was the same as in Study 1 except that participants entered the descriptions of the events in a spreadsheet using their native language. These descriptions were glossed at a later stage by the participants and coded by the experimenters. Further interviews with participants were held to resolve any coding questions.

## Results

To compare the two contrasting predictions of the conceptual immaturity and pragmatic accounts, we recorded how frequently 'front' and 'back' adpositions (prepositions and postpositions) and visibility and occlusion expressions were used in Front and Back scenes cross-linguistically, averaging across informants of the same language (see Table 2). Beginning with adpositions, numerical data showed that in 6 out of 10 languages (German, Javanese, Korean, Malay, Spanish, and Turkish) 'back'-denoting adpositions were mentioned more frequently than 'front'-denoting adpositions, while in 3 languages (Dhivehi, Pashto, and Swahili) both types of adpositions were mentioned equally frequently. The opposite pattern was exhibited in the remaining language (Cantonese).

We also inspected the proportion of expressions indicating visibility (in the context of Front scenes) and occlusion (in the context of Back scenes). This inspection revealed that there were no visibility expressions for Front in any of the languages surveyed but occlusion expressions for Back occurred in 8 of the 10 languages in the sample (e.g., verbs with meanings such as 'hide', 'appear'/‘disappear' etc.).

Table 2: Percentage of adpositions and expressions of visibility and occlusion used for the FRONT and BACK relations across languages

| Language | FRONT <br> Adpositions | BACK <br> Adpositions | Visibility <br> Expressions | Occlusion <br> Expressions |
| :--- | :---: | :---: | :---: | :---: |
| Cantonese <br> $(\mathrm{n}=1)$ | 62.5 | 50 | 0 | 0 |
| Dhivehi <br> $(\mathrm{n}=1)$ | 100 | 100 | 0 | 50 |
| German <br> (n=2) | 75 | 100 | 0 | 0 |
| Javanese <br> $(\mathrm{n}=1)$ | 0 | 50 | 0 | 37.5 |
| Korean <br> (n=1) | 12.5 | 100 | 0 | 100 |
| Malay <br> $(\mathrm{n}=2)$ | 25 | 100 | 0 | 12.5 |
| Pashto <br> (n=1) | 100 | 100 | 0 | 100 |
| Spanish <br> $(\mathrm{n}=2)$ | 87.5 | 100 | 0 | 100 |
| Swahili <br> $(\mathrm{n}=1)$ | 100 | 100 | 0 | 12.5 |
| Turkish <br> $(\mathrm{n}=2)$ | 87.5 | 100 | 0 | 62.5 |

Overall, the present cross-linguistic data largely replicated the key findings from Study 1. Adult speakers of 9 different languages (with the exception of the Cantonese speaker) used 'back'-denoting adpositions and/or occlusion expressions more frequently than 'front'-denoting adpositions. Similarly to the English and Greek data, there were no expressions denoting visibility for Front scenes in any of the languages in this wider cross-linguistic dataset. Despite the limitations of working with such low numbers of informants, this set of data presents suggestive evidence that our developmental conclusions from Study 1 generalize across languages.

## General discussion

Previous research shows that the acquisition of spatial language follows a stable, potentially universal, crosslinguistic timetable. However, the precise factors involved are not always clear. The acquisition of the locatives front and back is a case in point. Across languages, the locative back is produced earlier and is more frequent than the locative front. This asymmetry has been attributed either to a conceptual/semantic asymmetry in the early meanings of these locatives (with back being more basic than front; conceptual immaturity account) or to the fact that Back configurations are inherently more 'noteworthy' than Front configurations (pragmatic account). The present study put these two accounts to test.

Results showed that, in Study 1, both children and adult speakers of English and Greek typically used more Back than Front prepositions. Furthermore, speakers of all age and language groups often encoded Back information in occlusion verbs (e.g. hide) but no such verbs were available for Front. Study 2 provides suggestive evidence that the English and Greek developmental findings extend to a wider cross-linguistic sample of adult speakers of 10 additional languages. Taken together, these data support the predictions of the pragmatic hypothesis over those of the conceptual immaturity hypothesis.

Why do speakers prefer to encode Back over Front? On a pragmatic account that treats spatial language production as a form of communication governed by broadly Gricean (1975) or post-Gricean (Herskovits, 1985; Levinson, 2000; Sperber \& Wilson, 1985/1995) pragmatics, speakers need to mark occlusion (or the change to/from occlusion in our dynamic stimuli) so that the location (or path) of the Figure can be identified correctly by a hearer, even if the hearer has no visual access to the scene. By contrast, visibility for Front (or no change of visibility in our stimuli) is a default situation that speakers are less likely to mark. In our stimuli, speakers used many other alternatives instead of 'front' (e.g., 'beside', 'near', 'to/from the middle of'). It is possible that the bias favoring Back might be supported by non-linguistic factors relating to how occlusion is represented (e.g., see Hespos, Gredebäck, von Hofsten, \& Spelke, 2009; Spelke \& von Hofsten, 2001).

The present results have intriguing implications about the nature of spatial language acquisition and use. According to traditional theories of linguistic and cognitive development,
the order in which children acquire spatial locatives (and other non-spatial vocabulary) is considered as an index of conceptual growth (e.g., see E. Clark, 1973; Bowerman, 1996; Huttenlocher, Smiley, \& Charney, 1983; Johnston \& Slobin, 1979). Our results raise the possibility that pragmatic pressures, which are active in adult communicators as well, can also shape the way spatial language is acquired and used. Furthermore, our findings suggest that pragmatic factors may also yield cross-linguistically stable, and potentially universal, patterns of spatial language use. Interestingly, the pragmatic preference to encode Back over Front may also affect the shape of cross-linguistic spatial semantic systems. In an extensive cross-linguistic report, Levinson and Wilkins (2006) state that, if a language has a 'front'-denoting locative it will necessarily have a 'back'-denoting locative but the reverse pattern does not occur. The way pragmatic considerations interact with conceptual and other factors to shape spatial language acquisition and use crosslinguistically is a rich avenue for future research.

## Acknowledgement

We wish to thank Gaby Hermon and Tom Conners for help with the Javanese data. This material is based upon work supported by the National Science Foundation under grant no. 1632849. M.J. is now with Community Research Partners in Columbus, Ohio.

## References

Bowerman, M. (1996). Learning how to structure space for language: A crosslinguistic perspective. In P. Bloom, M. A. Peterson, L. Nadel, \& M. F. Garrett (Eds.), Language and space (pp. 385-436). Cambridge, MA: MIT press.
Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), Cognitive development and the acquisition of language (pp. 65-110). New York: Academic Press.
Clark, E. V. (1980) Here's the top: Nonlinguistic strategies in the acquisition of orientational terms. Child Development, 51, 329-338. doi:10.2307/1129265
Grice, H., P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), Syntax and semantics: Speech acts, (Vol. 3, pp. 41-58). New York: Academic Press.
Herskovits, A. (1985). Semantics and pragmatics of locative expressions. Cognitive Science, 9(3), 341-378.
Hespos, S., Gredebäck, G., von Hofsten, C., \& Spelke, E. S. (2009). Occlusion is Hard: Comparing predictive reaching for visible and hidden objects in infants and adults. Cognitive Science, 33(8), 1483-1502. doi: 10.1111/j.1551-6709.2009.01051.x

Hill, C., \& Vandeloise C. (1991). Recherches interlinguistiques en orientation spatiale [Crosslinguistic studies on spatial orientation]. Communications, 53, 171-207. doi: 10.3406/comm.1991.1806

Huttenlocher, J., Smiley, P., \& Charney, R. (1983). Emergence of action categories in the child: evidence from verb meanings. Psychological Review, 90, 72-93. doi: 10.1037/0033-295X.90.1.72
Johnston, J. R. (1984). Acquisition of locative meanings: Behind and in front of. Journal of Child Language, 11, 407-422. doi:10.1017/S0305000900005845
Johnston, J. R., \& Slobin, D. I. (1979). The development of locative expressions in English, Italian, Serbo-Croation and Turkish. Child Language, 6(3), 529-545. doi: 10.1017/S030500090000252X

Kubena, M. D. (1968). An experimental study of the comprehension and expression of prepositions of location and direction of movement in the speech of children (Unpublished master's thesis). University of Texas at Austin, Austin, TX.
Landau, B., \& Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences, 16, 217-265. doi:10.1017/S0140525X00029733
Levinson, S. C. (2000). H. P. Grice on location on Rossel Island. Proceedings of the 25th Annual Meeting of the Berkeley Linguistic Society (pp. 210-224). Berkeley, CA: Berkeley Linguistics Society.
Levinson, S. C., \& Wilkins, D. P. (Eds.). (2006). Grammars of Space: Explorations in Cognitive Diversity. Cambridge: Cambridge University Press.
Miller, G. A., \& Johnson-Laird P. N. (1976). Language and perception. Cambridge, MA: Harvard University Press.
Papafragou, A., Viau, J., \& Landau, B. (2013, November). The ins and outs of spatial language: Paths, places, and negative spatial prepositions. Paper presented at the $38^{\text {th }}$ Annual Meeting of the Boston University Conference on Language Development, Boston, MA.
Parisi, D., \& Antinucci, F. (1970). Lexical competence. In G. B. Flores d'Arcais \& W. J. M. Levelt (Eds.), Advances in psycholinguistics. Amsterdam: North-Holland.
Piaget, J., \& Inhelder, B. (1967). The child's conception of space. New York: Norton.
Spelke, E. S., \& von Hofsten, C. (2001). Predictive Reaching for Occluded Objects by 6-Month-Old Infants. Journal of Cognition and Development, 2(3), 261-281. doi:10.1207/S15327647JCD0203_2
Sperber, D., \& Wilson, D., (1986/1995). Relevance: Communication and Cognition (2nd ed. 1995). Cambridge, MA: Harvard University Press.
Talmy, L. (1983). How language structures space. In H.L. Pick \& L.P. Acredolo (Eds.), Spatial orientation: Theory, research and application (pp. 225-282). New York: Plenum Press.
Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), Language typology and syntactic description (pp. 57-149). New York: Cambridge University Press.
Tanz, C. (1980). Studies in the acquisition of deictic terms. Cambridge: Cambridge University Press.

# Cognitive Style Predicts Magical Beliefs 

Giorgio Gronchi (giorgio.gronchi@unifi.it)<br>Department of Neuroscience, Psychology, Drug Research and Child's Health - Sect. Psychology<br>via di San Salvi 12, pad 26-50100 Firenze, Italy

Jeffrey C. Zemla (zemla@wisc.edu)<br>Department of Psychology<br>University of Wisconsin-Madison, Madison, WI 53706 USA

## Marco Brondi (marcobrondi@gmail.com)

La Corte dei Miracoli - House of Magic
via G. Piombanti 9d-57122 Livorno, Italy


#### Abstract

Magicians often rely on misdirection to fool their audience. A common way to achieve this is for the magician to provide a plausible and intuitive (but false) account of how an effect is performed in order to prevent spectators from uncovering the truth. We hypothesized that analytical thinkers would be more likely than intuitive thinkers to seek alternative explanations when observing a mental magic effect because generating a coherent explanation requires analytical thought. We found that while intuitive thinkers often espoused explanations for a magic trick similar to one provided by the magician, analytical thinkers tended to generate new explanations that echoed rational principles and relied on physical mechanisms (rather than mental capabilities). This difference was not predicted by differences in numeracy skills or need for cognition.


Keywords: cognitive style, misdirection, CRT, dual process theory of reasoning.

## Introduction

Renewing a research program as old as scientific psychology itself (Binet, 1894; Triplett, 1900), in the last decade a new research program has emerged that uses illusionism and magical effects to investigate how the mind works (Kuhn, Amlani, \& Rensink, 2008; Kuhn \& Land, 2006; Kuhn, Olson \& Raz, 2016; Macknik, MartinezConde, \& Blakeslee, 2010; Rensink \& Kuhn, 2015). Such efforts have mainly focused on the roles of perception, attention, and visual cognition, though a limited number of studies have examined the relationship between higher-level cognition and illusionism (Danek et al., 2014; Olson et al., 2015; Subbotsky, 2010). Nonetheless, there may be a deep connection between cognitive styles of thinking and the way in which people explain a magical effect.

Ekroll and Wagemans (2016) wrote "[illusionists'] ultimate aim is to design miracles, not mere illusions. That is, the magician's first question is how they can create the illusion of impossibility. Relatedly, the magician's second question is how they can make sure that nobody is able to figure out how it was done. That is, they are essentially aiming to construct a problem that is as difficult to solve as possible, given the fundamental principles of human
problem solving" (p. 486). These goals are mainly addressed by using misdirection: the act of manipulating the spectator's attention away from the actual cause of a magic effect (Kuhn et al., 2014). From a psychological point of view, misdirection can be achieved through the manipulation of variables related to at least three different processes: perception, memory and reasoning.

Kuhn et al. (2014) describe the role of perception and memory in detail, but the authors pointed out that misdirection of reasoning and beliefs is loosely defined and harder to describe. However, virtually all areas of illusionism employ techniques based on the dual process theory of thought (Evans \& Over, 1996; Sloman, 1996; Stanovich \& West, 2000) distinction between analytical (i.e., a slow and effortful form of deliberative thought) versus intuitive (i.e., a fast and effortless form of associative thought) thinking. When a magician interacts directly with a spectator, their actions, dialogue, and other aspects of the performance are aimed at prompting fast, effortless, associative, and nearly-automatic responses. For example, in the classic force, a spectator is asked to pick a card. The choice appears to be at the discretion of the spectator; however, the magician has actually chosen a predetermined card that he "forces" the spectator to pick. The force is obtained by placing the chosen card directly in the hand of the spectator in a way that seems as if it were a random choice. The timing and the naturalness of the magician's movements are crucial factors in getting the participant to "choose" the predetermined card. From a psychological point of view, the success of the force depends on triggering an intuitive-based response-that the participant actually has a choice-and crucially, avoiding an analytical response that could lead the participant to choose a different card (such as one located at the extremities of the fan of cards). The renowned card magic conjuror Roberto Giobbi suggests a series of verbal and non-verbal techniques that stimulate a quick and automatic response (Giobbi, 1995). For example, make a person feel comfortable and then abruptly ask for a card. When this happens in front of an audience, the person may feel pressured to comply quickly or risk embarrassment in front of the public for not having completed such a simple assignment.

Mental magic attempts to create the illusion of impossibility by simulating supernatural mental abilities (e.g., telepathy, clairvoyance, psychokinesis, mediumship, and so on) and, as opposed to more traditional magic areas (such as card magic), sleight-of-hand or object manipulation skills cannot be taken into account as a possible explanation. More recently, given the cultural and educational changes in Western society, supernatural-based explanations have become unrealistic for a general audience-although intuitive thinkers are more likely to hold supernatural (Bouvet \& Bonnefon, 2015) and paranormal (Pennycook et al., 2012) beliefs. Instead, the mental magic effects are explained in terms of natural skills, such as the ability to reliably read non-verbal signals, body language and subliminal manipulation of others behaviors by means of psychological suggestion. Even if such abilities are not $100 \%$ reliable and in many cases are not sufficient for explaining the observed effect, people generally accept that mental magic is due to highly trained psychological skills. Mentalists will adopt many subtle techniques in order to promote the default, most intuitive or automatic explanation and to avoid promoting alternative explanations. Analytical thinking could help the observer to contemplate alternative hypotheses (such as the use of physical devices or the presence of an accomplice) as well as the weaknesses of the assumed explanation based on the highly developed psychological ability of the mentalist (such as the unreliability of method).

Because magicians rely on intuitive explanations to sell the illusion, individuals prone to analytical thinking may be less susceptible to these tricks. Can individual differences in cognitive style predict the explanations given to a mental magic performance? Adopting a common methodology adopted from the dual process theory literature (Gronchi et al. 2016; Zemla, Steiner, \& Sloman, 2016), it is possible to investigate the relation between cognitive style (predisposition to adopt analytical vs intuitive thinking) with the explanation given to a mental magic effect. We seek to establish whether analytical thinking affects the explanations produced by spectators of a magic trick. We predict that observers adopting an analytical cognitive style are more able to inhibit the default, mental power-based explanation suggested by the mentalist.

## Experiment

We investigated whether an individual's cognitive style affects judgments about a mental magic effect. Participants watched a video where an expert mentalist performed a prediction effect (predicts a purported "free" choice made by the spectator), and were then prompted to explain the effect they just witnessed. In addition, participants made several judgements, such as whether it was easy to generate an explanation, whether they were surprised by the outcome of the effect, and whether they enjoyed the trick. On a separate day, participants completed an extended version of the cognitive reflection test (CRT; Frederick, 2005) used to measure the cognitive style and other related measures.

## Method

Participants 335 freshmen college students (71 male, 29 unknown) enrolled in the Psychology major of the University of Florence were recruited for course credit. The sample mean age (in years) was $19.5(\mathrm{sd}=2.3)$, range $18-46$ (29 of unknown age).

Materials and procedure Participants completed a Mental Magic Task and four other questionnaires: an extended version of the CRT questionnaire, the Need for Cognition (NFC) questionnaire, an abbreviated Numeracy Scale, and three questions about Science Interest. All materials were presented to participants in Italian and are translated here for the reader.

On day 1 of the experiment, participants completed an extended version of the CRT questionnaire (Toplak, West \& Stanovich, 2014; Zemla, Steiner, \& Sloman, 2016). The CRT scale is comprised of questions that have a wrong but intuitive answer in addition to a correct answer that requires analytical thinking. For example, one question states: In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? The default, intuitive response is 24 days (if 48 days is the time necessary to cover the entire lake, half of 48 should be intuitively the time necessary for covering half the surface). However, if the responder inhibits such response, it is relatively easy to see that every day the patch doubles in size, so on the $47^{\text {th }}$ day the lake was half covered and on the $48^{\text {th }}$ the lily pads will cover the entire surface. So, the inhibition of the most obvious response allows to adopt a deliberation-based form of thinking. The number of correct answers on the CRT measures the degree to which participants engage in analytical thinking and inhibit intuitive responses. Participants completed the 7 -item CRT scale in an open-ended format (except for the last item which was a multiple choice question).

On the same day, participants also completed three multiple choice questions about Science Interest inspired by the Science Curiosity Scale (Landrum et al., 2016): 1) Which of the following do you most like to read? Possible answers: fiction, science, sports, politics, history, other. 2) Let's suppose that it's necessary to take a mandatory class (which will not influence your grade; only attendance is necessary). Which of the following topic would you like to take? Possible answers: Contemporary history, Creative writing, Physics and Astronomy, Cinema and Arts. 3) Let's suppose that you are travelling for business in an abroad city that you have already visited and known. You have a free afternoon and you have time to do only one of the following activities (free and equally near to your hotel): Possible answers: Visiting the science museum, visiting the contemporary art museum, watching a show in a square, relaxing in a park, stay in the hotel to rest. Thus, for each question, there was an answer indicating interest toward science. The scale is intended to avoid socially desirable
responses that may be elicited from other science interest scales.

On day 4, participants completed a six-item numeracy scale (Weller et al., 2013) that evaluates one's competence in solving numerical problems (i.e., basic numerical operations, percentages). Given that CRT also measures numerical abilities as well as of inhibition of the intuitive response, we included the Numeracy Scale in order to exclude the possibility that an effect is due to numeracy and not cognitive style. On the same day, participants also completed the Need for Cognition scale (NFC; Cacioppo \& Petty, 1982). The NFC measures the tendency to enjoy and engage in challenging cognitive activities. Indeed, some people have little motivation and tend to avoid effortful cognitive activities whereas other individuals consistently seek opportunities to engage such kind of tasks.

On day 8, participants watched a video about a mentalist's performance. The mentalist is talking with a spectator about free will and the possibility of predicting others' behavior. The mentalist then shows a chessboard with the pieces in the starting position and asks the spectator to choose one of the pieces. The spectator, following the mentalist's directions, announces the chosen piece (a white bishop) and places it in the middle of the chessboard. The mentalist takes the chosen piece in his hands and begins to speculate about the factors that could have influenced the participant's choice: Did the white bishop have a particular significance? Was that particular bishop closer to the spectator compared to other bishops? Finally, the mentalist declares that he knew in advance that the spectator would take that particular piece and then invites the spectator to take another piece. The video ends with the revelation that every other piece besides the white bishop is stuck to the chessboard, so it would appear impossible that the participant could have chosen any other piece. Participants observed the video in a group, but they were also able to re-watch it using a smartphone or tablet. After watching the video, participants were individually asked to provide an explanation for the effect they just saw in an open-ended format. In addition, they answered each of the following questions on a 5-point Likert scale (from "not at all" to "extremely" for questions 1,2 and 4 and from "strongly disagree" to "strongly agree" for questions $3,5,6,7): 1$ ) How much did you enjoy the effect you just saw? 2) How much did it surprise you? 3) I tried to predict what the magician would do before he did it 4) How confident are you in your explanation? 5) It is easy to think of many alternative explanations for this effect 6) I would like to know how the effect actually works 7) I would like to see other magic effects.

## Data Analysis

For each participant, four scores were computed: a CRT score (the number of correct analytical responses out of 7), an NFC score, and a science interest score (giving 1 point for each answer related to science), and a numeracy score. The open format question ("How do you explain the effect you just saw?"), was coded using two different criteria. The
first criterion was based on physical explanation vs mental explanation. Physical explanations often contained reference to a physical device, such as glue or a magnet, but also include trivial solutions such as collusion between the mentalist and the spectator. Mental explanations were based on the possibility of behavior conditioning, the power of gestures, and the possibility of genuinely predicting in advance the spectator's choice (including supernatural powers). A third "other" category included no response, incomplete responses, and explanations not suitable to be categorized in previous terms (such as "I don't know"/"no idea").

A second criterion coded explanations as rational or irrational. Rational explanations included all the physical explanations that are actually feasible in practice, but also included statistical-based reasoning such as the possibility of predicting behavior with high probability on the basis of modal choices (obtained from experience). Irrational explanations were the same as the mental explanations excluding interpretations based on the statistical properties of people's choices (by means of previous empirical observations). Since all reported physical explanations were feasible ways of obtaining the effect, no physical explanations were coded as irrational. Again, an "other" category was included for coding no response, incomplete responses or other explanations not suitably categorized in previous terms. Two independent judges coded each explanation according to the two criteria. A third, independent judge broke any ties ( 5 out of 335 for the first criterion and 22 out of 335 for the second criterion).

## Results

Explanations of the magic effect With regard to the physical/mental dichotomy, the most common explanation (about 70\%) was a mental explanation based on the belief that the magician could systematically influence other people choices, such as "the mentalist implicitly and secretly conditioned the choice of the spectator." In several cases, participants added details such as "with gestures", "with his voice", "with his gaze", "with his mind", "with his movements," and so on. Only $13 \%$ of participants explained the effect in terms of a physical device (e.g., a special, delayed-effect glue; a magnet-based mechanism) or with a trivial solution (the spectator is an accomplice). The remaining $17 \%$ of responses were classified as "other".

With regard to the rational/irrational dichotomy, all of the physical explanations were included in the rational category and the majority of the mental explanations were categorized as "irrational". However, some of the mental explanations were included in the "rational" category. These were explanations based on the belief of influencing other's choice by subtle psychological techniques together with some rational considerations. For example: "The mentalist did a lot of trials before with other people to look for the most chosen piece. When facing the spectator, the mentalist someway suggested with his gaze to take the most likely piece. I think the mentalist has been very lucky." or "The
mentalist knows which piece is generally chosen most often. He exploited such knowledge together with his ability to influence spectator choice. Maybe he was not certain of the final result, but it was the most likely result." According to the rational/irrational dichotomy, the rational explanations were about $20 \%$ and the irrational explanations were about $63 \%$. The remaining $17 \%$ were classified as "other".

Explanations types and analytical style We evaluated whether CRT scores were correlated with the type of explanation produced (Figures 1 and 2). As predicted, those who generated a physical explanation (compared to a mental explanation) tended to have higher CRT scores, $t(70.1)=$ $12.89, p<.001$. Likewise, participants who wrote a rational explanation had higher CRT scores compared to those who wrote an irrational explanation $t(105.7)=13.61, p<.001$.


Figure 1: Participants who generated physical explanations typically had higher CRT scores (more analytical).


Figure 2: Participants who generated rational explanations typically had higher CRT scores (more analytical).

Explanations types and Numeracy Scale Unlike CRT, numeracy scores did not predict the explanation type that
participants generated. Participants who generated physical compared to mental explanations did not differ in numeracy, $t(34.2)=.33, p=.74$, nor did participants who generated rational compared to irrational explanations, $t(73.9)=.27, p$ $=.79$. This rules out the possibility that the observed CRTrelated differences can be due to disparity in numerical ability.

Explanations types and Need for Cognition There were no differences in the NFC score between those who generated physical and mental explanations, $t(33.2)=.11, p$ $=.91$, or between those who generated rational/irrational explanations, $t(80.8)=1.26, p=.21$. This rules out the possibility that the observed CRT-related differences can be due to different inclination towards effortful cognitive activities.

Explanations types, science interest and analytical style Participants that wrote a physical explanation had a greater score in Science Interest compared to those that wrote a mental explanation, $t(55.3)=2.65, p=.011$. Similarly, participants who wrote a rational explanation had a greater interest in science, $t(94.97)=2.36, p=.021$. Science Interest was also significantly correlated with the CRT, $\mathrm{r}=$ $.35, p<.001$.

Explanation types and correlations among the questions of the Mental Magic task We expected that the way in which participants explained the effect would influence their response and enjoyment of the effect. However, we found that the type of explanation produced did not predict enjoyment of the effect, the surprise experienced, the attempt to predict the effect, the confidence in the explanation given, the ease of thinking of different explanations, the desire to know how it works, or the desire to see a new effect (all $p>.1$ ). Similar results were obtained for the rational/irrational dichotomy. However the questions about the Mental Magic effect were correlated among themselves (Figure 3). Having enjoyed the effect correlated with all other variables in the mental magic task except the confidence of the given explanation. The degree of surprise was highly correlated with enjoyment of the trick ( $\mathrm{r}=.45$ ), the desire to know how the trick was performed $(\mathrm{r}=.33)$ and desire to see new magic effects ( $\mathrm{r}=.34$ ). Attempts to predict the mentalist's actions during the effect was weakly correlated with the ease of generating an alternative explanation after the effect $(r=.07)$, the desire to know how the trick was performed ( $\mathrm{r}=.14$ ), and desire to see new magic ( $\mathrm{r}=.12$ ). The highest correlation was between the desire to know how the trick was performed and the desire to see new magic $(r=.65)$. All correlations reported above were significant with $p<.05$, uncorrected for multiple comparisons.


Figure 3: Correlations among the enjoyment of the effect (enjoyment), the surprise experienced (surprise), the attempt to predict the effect (prediction), the confidence in the explanation given (confidence), the easiness of thinking of alternative explanations (ease of explanation), the desire to know how it was performed (know the trick) and the desire to see a new effect (new). Correlations with CRT, NFC,
Numeracy skills and Science Interest are also reported.
Analytical style and Mental Magic task questions The CRT was marginally correlated with the interest in seeing new magic ( $\mathrm{r}=.10, p=.08$ ), significantly correlated with attempts to predict how the trick was done $(\mathrm{r}=.11, p=$ .004 ), and significantly correlated with enjoyment of the mental effect ( $\mathrm{r}=.14, p=.01$ ).

NFC and Mental Magic task questions Similarly to CRT, Need for Cognition was correlated with the interest in seeing new magic ( $\mathrm{r}=.21, p=.01$ ), attempts to predict the next move of the mentalist ( $\mathrm{r}=.21, p=.01$ ), and with the enjoyment of the mental effect ( $\mathrm{r}=.18, p=.04$ ). The NFC was also marginally correlated with the interest in how the trick was performed ( $\mathrm{r}=.15, p=.07$ ).

Science Interest and Mental Magic task questions Science Interest was significantly correlated with attempts to predict the actions of the mentalist $(\mathrm{r}=.12, p=.029)$.

## Discussion

We found that analytical thinking predicts the way people explain a mental magic effect: intuitive thinkers were more inclined to explain a mental magic effect in the same terms suggested by the mentalist (e.g., conditioning the spectator's choice or advanced psychological ability), whereas analytical thinkers were more likely to explain the observed
effect by referring to a physical device or trivial tricks (such as collusion, i.e., a previous agreement between the mentalist and the spectator). The same pattern also held when "rational" considerations were taken into account in the categorization of the explanations: analytical thinkers were more inclined to seek an alternative explanation and reject an irrational explanation offered by the mentalist. We also found that these same differences in explanation type were predicted by interest in science, perhaps due to an aversion towards non-scientific (irrational) explanations. Moreover, we observed that such differences between analytical and intuitive thinkers were not due to related constructs such as numeracy skills or need for cognition

Although the explanation that a participant offered did not affect the way they perceived the trick (i.e., it was not correlated with any of the mental magic task questions), analytical style did predict differences. Analytical thinkers were more likely to predict the next step of the performance, suggesting that they were trying to go deeper into the effect not only at the end of the performance but also during the trick itself. This deeper level of engagement may be the reason why analytical thinkers also enjoyed the magic effect more, and expressed more interest in seeing new magic. We also found that being an analytical thinker was associated with an interest in science.

A limit of the present study is its reliance on correlational data. We are planning to further investigate this topic by manipulating the cognitive style to verify whether inducing analytical thinking will prompt participants to generate alternative rational explanations. Moreover, it is important to note that the sample was composed of Italian freshmen psychology students in their first days of college: commonly, those students have high expectations about the capabilities of psychology, including the skills that mentalist's performance suggests. This may explain the bias to provide mental explanations in our sample, although it does not explain the main effect of analytical thinking on explanation type. Another critical aspect is that we employed a single performance that was either interpretable in terms of advanced psychological skills and in physicaldevice terms in a relatively easy way. In other kinds of performances, it could be more difficult to think of rational/physical device-based explanations.

This work represents a first-step in the psychological investigation of magical effects in cognitive terms going beyond the more common perception and attention-based perspectives. The already existent but still limited literature about high-level cognition and magic may greatly benefit from our understanding of reasoning-based misdirection in terms of dual process theory of thought. In particular, such benefits can go in two opposite directions: psychological studies on cognitive styles (and dual process theory) may employ mental magic effects and magician's misdirection techniques to create unique and innovative experimental settings and, at the same time, conjurors may rely on the dual process theory of thought to improve their performance.

## Acknowledgments

Many thanks to Stefania Righi and Fiorenza Giganti for judging the open format questions. We are also in debt with Rachele Massini (who was also the third judge), Giacomo Fini, Daniele Lari, and Eleonora Falciani for collecting data. This work has been motivated also by fruitful discussions during La Corte dei Miracoli - House of Magic meetings.

## References

Binet, A. (1894). Psychology of Prestidigitation. Annual Report of the Board of Regents of the Smithsonian Institution, Washington, DC: U.S. Government Printing Office.
Bouvet, R., \& Bonnefon, J-F. (2015). Non-reflective thinkers are predisposed to attribute supernatural causation to uncanny Eexperiences. Personality and Social Psychology Bulletin, 41(7), 955-961.
Cacioppo, J. T., \& Petty, R. E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42(1), 116.
Danek, A. H., Fraps, T., Von Mueller, A., Grothe, B., \& Öllinger, M. (2014). Working wonders? Investigating insight with magic tricks. Cognition, 130(2), 174-185.
Ekroll, V., \& Wagemans, J. (2016). Conjuring deceptions: Fooling the eye or fooling the mind? Trends in Cognitive Sciences, 20(7), 486-489.
Evans, J. St. B. T., \& Over, D. E. (1996). Rationally and Reasoning. Hove, UK: Psychology Press.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19, 25-42.
Giobbi, R. (1995). Card College, Vol. 1: A Complete Course in Sleight-of-hand Card Magic. Seattle: Hermetic Press.
Gronchi, G., Righi, S., Parrini, G., Pierguidi, L., \& Viggiano, M. P. (2016). Dual process theory of reasoning and recognition memory errors: Individual differences in a memory prose task. In Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 331-335). Austin, TX: Cognitive Science Society.
Kuhn, G., Amlani, A. A., \& Rensink, R. A. (2008). Towards a science of magic. Trends in Cognitive Sciences, 12(9), 349-354.
Kuhn, G., Caffaratti, H. A., Teszka, R., \& Rensink, R. A. (2014). A psychologically-based taxonomy of misdirection. Frontiers in Psychology, 5.
Kuhn, G., \& Land, M. F. (2006). There's more to magic than meets the eye. Current Biology, 16(22), R950.
Kuhn, G., Olson, J. A., \& Raz, A. (2016). Editorial: The psychology of magic and the magic of psychology. Frontiers in Psychology, 7:1358.
Landrum, A. R., Hilgard, J., Akin, H., Li, N., \& Kahan, D. M. (2016). Measuring interest in science: The Science Curiosity Scale. In Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 16191624). Austin, TX: Cognitive Science Society.

Macknik, S. L., Martinez-Conde, S., \& Blakeslee, S. (2010). Sleights of Mind: What the Neuroscience of Magic Reveals about our Everyday Deceptions. London: Macmillan.
Olson, J. A., Amlani, A. A., Raz, A., \& Rensink, R. A. (2015). Influencing choice without awareness. Consciousness and Cognition, 37, 225-236.
Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J. \& Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. Cognition, 123(3), 335346.

Rensink, R. A., \& Kuhn, G. (2015). A framework for using magic to study the mind. Frontiers in Psychology, 5.
Sloman, S. A. (1996). The empirical case for two system of reasoning. Psychological Bullettin, 119(1), 3-22.
Stanovich, K. E., \& West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate. Behavioural and Brain Sciences, 23, 645-726.
Subbotsky, E. (2010). Magic and the Mind: Mechanisms, Functions, and Development of Magical Thinking and Behavior. Oxford: Oxford University Press.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking \& Reasoning, 20(2), 147-168.
Triplett, N. (1900). The psychology of conjuring deceptions. American Journal of Psychology, 11, 439-510.
Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., \& Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. Journal of Behavioral Decision Making, 26(2), 198-212.
Zemla, J. C., Steiner, S. M., \& Sloman, S. A. (2016). Analytical thinking predicts less teleological reasoning and religious belief. In Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 12171222). Austin, TX: Cognitive Science Society.

# Stopping Rules in Information Acquisition at Varying Probabilities and Consequences: an EEG Study 

Roberto Guedes de Nonohay (roberto.nonohay@ufrgs.br)<br>Cognitive Biosignals Lab - Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos, 2600<br>Porto Alegre, 90035003, Brazil<br>Gustavo Gauer (gustavo.gauer@ufrgs.br)<br>Cognitive Biosignals Lab - Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos, 2600<br>Porto Alegre, 90035003, Brazil<br>Richard Gonzalez (gonzo@umich.edu)<br>Biosocial Methods Collaborative, - University of Michigan, 426 Thompson Street<br>Ann Arbor, MI 48106 USA<br>Guilherme Lanning (guilherme.lannig@gmail.com)<br>Cognitive Biosignals Lab - Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos, 2600<br>Porto Alegre, 90035003, Brazil


#### Abstract

An experiment aiming to assess the use of stopping rules in information acquisition was performed. Participants were requested to make a decision in 24 financial scenarios with the possibility of buying information pieces. Behavioral and EEG data were recorded for analysis. Results showed that participants followed Bayesian calculations in order to determine a stop on information acquisition and decide. Moreover, the information acquisition strategies were consistent with prospect theory, in which participants will weigh information pieces differently and seek more or less information given different manipulations in scenario probability and consequences. EEG data suggest Slow Cortical Potentials at fronto-central electrodes.


Keywords: decision making; information acquisition; EEG; slow cortical potential.

## Introduction

As Taghavifard, Damghani and Moghaddam (2009) discuss, it is only possible to know the risks inherent in a decision if the individual has a relatively small degree of uncertainty. One way to diminish levels of uncertainty is by reducing residual uncertainty (Courtney, Kirkland \& Viguerie, 1997) through information acquisition. To acquire information is to search both internally and externally for elements that can affect the decision process. In their daily lives individuals receive a considerable amount of information through various modalities. Auditory, visual, tactile, emotional stimuli can be sources of new information. Each piece of information has some importance toward deciding either by improving the quality and quantity of information or by impairing an individual's ability to decide given that the amount of information is so great that the performance will be deteriorated (Di Caprio, Santos-Arteaga, \& Tavana, 2014). When information reveals itself and is processed by the decision maker, we find a transition from a situation of uncertainty to a situation of risk. In other words, the
decision maker now knows enough information about the problem so that he is able subjectively infer a probability for each outcome (Di Caprio et al., 2014).

Pretz, Naples, and Sternberg (2003) discuss the role of experts and the fact that too much information can actually impair the decision process. They propose that when an expert (a person that possesses a great deal of knowledge, acquired by experience and information gathering) in chess plays with slightly different rules, his performance might actually be worse than that of a player that is new to chess and plays the same modified game as the expert. This suggests that when an otherwise static environment becomes dynamic, a difficulty in deciding might appear. Too much information may be suboptimal for a decision maker (Di Caprio et al., 2014), whereas not enough information will prevent him from calculating risks properly and brings the decision process to one of most uncertainty (Taghavifard, Damghani \& Moghaddam, 2009). On the other hand, Frey, Hertwig and Rieskamp (2014) propose that there is no way to determine when the right amount of information is reached and no further acquisition needs to be done, at least in decisions from experience, although they also say that there may be benefits in small samples and frugal search. The question that remains is: how does a decision maker knows that he/she acquired enough information to go through with the process?

Many researchers are investigating the subject of information acquisition and how and when individuals stop searching for new information and proceed to decide. Gigerenzer (2000) proposes a fast and frugal way to decide in environments where both time and knowledge are restricted. By searching past information and knowledge in order to recognize elements regarding the decision and cues about those elements, the Take the Best (TTB) heuristic searches for the best cue in order to make a choice. In the experiments by Gigerenzer (2000), when people where
asked which of two German cities was the most populated, individuals would most likely use TTB in order to decide. Even so, the individual might seek other cues about each city from memory (i.e., perhaps if he saw the city on the news). According to the subjective validity of the cues, the one with the highest ranking is considered the best and thus appropriate for a decision. Little information search and acquisition are performed. Stern, Gonzalez, Welsh, and Taylor (2010) conducted and experiment in which individuals were presented with two decks with varying proportions of red and blue cards. Four draws of cards were made and at each draw the individual would have to state from which deck the card had been drawn from. Each draw represented acquiring a new piece information about the decision. After all four draws the individual would have to make a final decision between the decks or they could decline to choose. It is clear that each new information presented changed or reaffirmed the decision made by the individual. When conflicting information was presented (two draws were red cards and two were blue) individuals mostly declined to choose, inferring a $50 \%$ chance to each deck. When all draws were the same color, by the third draw individuals were already $100 \%$ confident from which deck the draws were made. This experiment poses that information acquisition can update individual beliefs about the outcome and that searching for information might improve the decision making process by incrementing it with a better view about the problem at hand.

Fifić and Buckmann (2013) probed the use of stopping rules by individuals. Stopping rules might determine the moment where the decision maker stops, or should stop, searching for information and actually decide. The authors reviewed some options of stopping rules that might require higher or lower cognitive demands. The first one is the socalled optimal stopping rule for evidence accumulation. It is based on Bayesian inference and implies that there should be an optimal number of pieces of information that need to be acquired. In their example the optimal stopping rule is 3 . This number represents that the individual will search for positive $(+1)$ and negative (-1) pieces of information and will only stop searching when the sum of the search reaches either +3 or -3 , in which case the individual will choose the option represented by the positive or negative sum, in their example to proceed or not with a risky cancer treatment. There is criticism regarding this rule, in order to calculate the optimal number there is a need to have a perfect knowledge of the situation and enough calculating skills to solve it through Bayesian probability (Fifić \& Buckmann, 2013). This option requires great amounts of time, knowledge and cognitive effort. In most cases in the real world there are limited amounts of each available to the decision maker. They then propose a stopping rule selection theory based on bounded rationality.

Two rules are suggested that do not depend on high amounts of knowledge about the environment and the situation. The first one is the fixed sample size. This rule entails that the decision maker will determine a sample size
before the beginning of the information search process, for example five. The individual will then search for information and will make the choice based on the valence that appears the most (positive or negative). The other rule is called runs stopping rule. In this case the decision maker will begin the search for information without determining a fixed sample. She will stop searching when a streak of either positive or negative pieces of information is found, three consecutive positive opinions for example.
The stopping rule selection theory proposes that each individual might use different stopping rules given time and cognitive efforts available (Fifić \& Buckmann, 2013). That is because there is no evidence that one single stopping rule can account for all responses from individuals. According to Fifić and Buckmann (2013) each individual will search a decision operative space in which the rules and values are stored. Given a decision situation the individual will then retrieve a stopping rule - a process that the authors call castnet retrieval. Much like fishing, each individual will select a space and a net size to cast and retrieve a stopping rule that will be applied. What is considered in order to cast a net in the decision operative space is the level of uncertainty with the environment, time frame, cognitive demand, and accuracy expectancy (Fifić \& Buckmann, 2013). After the stopping rule is selected, the individual will then proceed to collect information and finally decide.
Cognitive demand and the search for a stopping rule might reflect high levels of task engagement. That is, the individual is fully focused on solving the problem and anticipates the outcomes of the decision given each new information. This situation represents higher use of brain resources, especially in frontal areas. Few studies focus their analysis on pre-stimulus ERPs, especially when decision making is concerned. Böckner, Bass, Kenemans and Verbaten (2001) studied one form of Slow Cortical Potential (SCP). They found a Stimulus-Preceding Negativity at fronto-central electrodes in fear-induced trials. Oswald and Sailer (2013) found fronto-central SCPs before and after response in a temporal discounting task.

Other elements also influence the information acquisition process. Frey, Hertwig and Rieskamp (2014) found that both a facial expression of fear or the subjective feeling of fear may cause an individual to search more information. Söllner, Bröder, Glöckner and Betsch (2014) discovered that when intruding incompatible information appears, individuals trained in the TTB heuristic would not stop searching for information when they were supposed to if following TTB. Individuals rather adapted their information search, choice and confidence judgment processes to the content of such intruding information. It is widely recognized that the amount of information available and acquired by each individual will augment complexity levels in the decision situation, much like what happened with the intruding information.

## Methods

The objective of this study was to probe, based on the models of Fifić and Buckmann (2013), Stern et al. (2010) and Söllner et al. (2014), the use of stopping rules in the information acquisition and evidence accumulation processes and its electrophysiological correlates. A financial decision task was devised so that the use of stopping rules could be measured by the amount of information acquired by the individuals in each of the scenarios. As with real world decisions, scenarios were presented with varying levels of risk, uncertainty and consequences. During the task, EEG was continuously recorded to investigate correlates of information acquisition and decision behavior processes. A total of 47 (mean age: 18.89, SD: 1.68, 33 females) undergraduates from the University of Michigan Pysch Pool participated. Data was collected from 50 participants, however 3 were discarded because of poor electrode readings interfering with the EEG data. This study was approved by the University of Michigan's Institutional Review Board.

## Financial decision task

Each participant was presented with all 24 financial decision scenarios. The scenarios were presented written in a single paragraph. In all scenarios participants would have to choose whether to accept or reject the proposed situation, but they could also choose not to decide at all (a procrastination behavior). For every scenario there were 20 information pieces (or advices) that a participant may or may not buy in order to help them decide. Participants were instructed to press the "I" key on the keyboard whenever they wanted to buy information in a scenario. All information was presented in a crescent and pseudorandom order. The order of information appearance was made to resemble the stopping rules tested by Fifić and Buckmann (2013). Each new information was presented using simply the words "positive" or "negative", thus diminishing the probability of bias. The words mean a positive or negative opinion about accepting or rejecting the proposition in the scenario. Each information had a price ( $\$ 1$ for the first 10 pieces and $\$ 2$ for the other 10 ). There was a fixed fictional amount of $\$ 480$ available to any participant to complete the experiment - this amount was created specifically to refrain participants from always buying all 20 pieces of information. They were instructed not to use all the money available.

Each scenario showed a situation involving aspects of financial decisions such as investments, purchases, asset management, losses, etc. After reading the description of the situation, participants could obtain (buy) information regarding that scenario. Even if not buying any information, participants would be required to make a decision for each scenario. They could decide to buy/invest/pay (Positive), not to buy/invest/pay (Negative) or to not decide at the moment (Procrastination). After a decision, there was no feedback on the success of it, and the next scenario was presented. Participants did not receive any instructions regarding a
maximum period of time to decide at each scenario. They were free to use as much time as they wanted to read the scenario description, seek information and make a decision. The 24 scenarios were divided as such: 1) 12 scenarios with stated probabilities (risk scenario) in the description, composed of 3 scenarios with low negative consequences, 3 with high negative consequences, 3 with low positive consequences and 3 with high positive consequences; 2) 12 scenarios with unstated probabilities (uncertainty scenario) in the description, composed of 3 scenarios with low negative consequences, 3 with high negative consequences, 3 with low positive consequences and 3 with high positive consequences.
One example of a stated probability, low positive consequence scenario is: "You are thinking about buying a bicycle. There is a model that is $35 \%$ better than the alternative. You don't know what the average maintenance costs might be. You must decide if you: buy the bicycle, don't buy the bicycle or rather not decide now.", as shown in Figure 1. The stated probability is the $35 \%$ chance depicted, low consequence is due to the amount ( $35 \%$ is considered a low chance), positive consequence is the referred chance of being better than the alternative. Scenarios differ in the presence or not of the stated probability, consequences and valences of consequences. That means that the example above might be presented in another form, representing an unstated probability, high negative consequence scenario, like: "... There is a model that is much worse than the alternative ". Phrasings of probabilities and consequences were randomized. That means that the object of the scenario would be the same (bicycle, student loan, car fixing, etc.), but the probabilities (stated or not), and consequences (high or low and positive or negative) were randomized across participants for any given object.
EEG data was recorded through Acknowledge 4.4 software using an ABM B-Alert X10, with a 9 channel setup (Fz, F3, F4, Cz, C3, C4, Pz, P3 and P4) using linked mastoid as reference. Data was collected at a sampling rate of 256 Hz . Electrode scalp impedances were kept below 5 $\mathrm{k} \Omega$. Behavioral data and stimulus presentation was made via PST E-Prime Professional 2.0. The data was analyzed using ERPLab (Lopez-Calderon \& Luck, 2014). Data went through moving window artifact detection and filtered for both low and high pass $(0.1 \mathrm{~Hz}$ and 30 Hz , respectively). ERP epoch was from -2000 ms before the decision was made and 200 ms after the decision was made, giving the possibility of observing variations that occurred in a window of time before the actual decision. The use of this epoch is justified given that the information acquisition process is over before the decision is actually made, so in order to analyze event related potentials of stopping rules it is necessary to observe what happens before the decision. Target electrodes were located at fronto-central sites in order to search for SCPs (Oswald \& Sailer, 2013). Mean voltage over a specific time epoch was used to analyze the data.


Figure 1: Example of a scenario.

## Results

In order to determine the use of stopping rules and strategies for information acquisition we focus our analyses on two measures: information quantity (QTY) and balance (BAL). Information quantity is the mean amount of information pieces that each individual bought during each scenario. The balance is, just as Fifić and Buckmann (2013) proposed, one of the stopping rules, the Bayesian calculation of the valences for each information bought. That is, if an information is positive, then the value considered is +1 , if an information is negative, then the value considered is -1 . At the end of a given scenario, for example, if the pieces of information acquired were 3 positives and 2 negatives (independent of order of appearance), the balance will be +1 . The conditions compared to the two measures were: decision (positive, negative and procrastination), probability (risk and uncertainty), and the combination of consequences (high or low) and valence of consequences (positive or negative) in risky and uncertainty.

## Information acquisition

Of the total of possible scenarios, $40.63 \%$ were decided without any kind of information acquisition, thus without the use of stopping rules. This behavior might emerge given the objects of the scenarios at hand. In order to better control the conditions, the objects of decision (car, bicycle, motorcycle purchase, student financial aid, home and car repair, investments) were less complicated. That might have made the decisions easier based on each individual set of preferences. However, there is no data to back this hypothesis. Next, there were $44.88 \%$ of the scenarios that were decided using 1 through 5 information pieces. The $14.50 \%$ of cases left used 6 through 20 information pieces.

## Decision

Regarding the decisions available for the participants, the mean information quantity gathered when a decision was positive is 4.11 , when a decision was negative also 4.11 and when participants decided to procrastinate the mean quantity was 5.04. That shows that, despite the fact that participants had up to 20 information pieces available they sought only a small amount. Also it shows that the procrastination behavior was observed with more acquisition of information. On the other hand, when the balance is considered, a positive decision was made with a mean balance of +1.13 , negative decisions -0.73 and
procrastination decisions -0.05 . That means that the information acquisition stopping point behavior is more influenced by the so called balance of the valences, regardless of the quantity of information acquired. A oneway ANOVA was conducted to test for differences between each decision. The test revealed that there is a difference between the decisions both for QTY and BAL, $\mathrm{F}(2,1146)=189.9, \mathrm{p}<0.001$ and $\mathrm{F}(2,1149)=6.35, \mathrm{p}<0.01$, respectively. Post-hoc analysis using Tukey HSD test revealed significant differences between all interactions: positive-negative $\quad(\mathrm{p}=0.001)$, negative-procrastination ( $\mathrm{p}<0.001$ ) and positive-procrastination ( $\mathrm{p}<0.001$ ) for the BAL measure and only negative-procrastination ( $\mathrm{p}<0.05$ ) and positive-negative ( $\mathrm{p}<0.05$ ) for the QTY measure.

## Probability

Analyzing only if the scenario presented risk or uncertainty, the only significant difference was observed in the BAL measure, $\mathrm{F}(1,1150)=4.75, \mathrm{p}<0.05$. The mean BAL for risk scenario was 0.031 . For uncertainty scenario the mean BAL was 0.262 . As for the QTY measure the mean value for the risk scenario was 4.263 and 4.100 for the uncertainty scenario.

## Combining the conditions

The conditions were not presented isolated to the participants. Combining the conditions yielded 8 possible scenarios, as it was previously explained, that were randomly presented three times each for the participants. If all conditions are analyzed there is a significant difference for the BAL measure $(\mathrm{F}(7,1128)=8.090, \mathrm{p}<0.001)$. A post hoc Tukey HSD test revealed significant differences between several of the possible combinations. However, two differences between conditions are of particular interest. The first one is between scenarios with uncertainty, low positive consequence $(\mathrm{M}=0.118)$ and scenarios with uncertainty, low negative consequence ( $\mathrm{M}=0.326$ ), with $\mathrm{p}<0.001$. The second one is between scenarios with risk, high positive consequences $(\mathrm{M}=0.007)$ and scenarios with risk, high negative consequences ( $\mathrm{M}=0.181$ ) with $\mathrm{p}<0.001$.

## EEG

EEG analysis focused on risky and uncertain scenarios and both of the combined conditions highlighted previously. As was discussed earlier SCP might emerge in a situation where there might be prolonged use of cognitive control and resources in fronto-medial electrodes (Oswald \& Sailer, 2013). As it was seen, BAL has significant differences in risky and uncertain scenarios and also in scenarios with different valences and consequences. That might point to the fact that prior to a decision individuals may exert more thought and allocate more cognitive resources to decide given the conditions presented.
The comparison between risky and uncertain conditions showed SCP negativity for the uncertain condition and a positivity for the risky condition in F4 between -950 ms and
-500 ms , with statistically significant difference $(\mathrm{F}(1,98)=5.847, \mathrm{p}<0.05)$ as shown in figure 2 .


Figure 2: SCPs in risk x uncertain condition in F4. Black line represents uncertain condition, red line risk condition. The ellipsis shows the point of the significant difference. Y axis represents micro voltages, X axis represents the epoch in milliseconds.

As for the comparison between risky and uncertain scenarios in a low consequence condition, we found a SCP negativity for the uncertain condition and a positivity for the risky condition in Fz and F4 between -1290 ms and -490 ms , with statistically significant differences for both electrodes $(F(1,98)=3.631, \quad \mathrm{p}=0.05$ and $\mathrm{F}(1,98)=4.720$, $\mathrm{p}<0.05$, respectively) as shown in figure 3 .


Figure 3: SCPs in risk $x$ uncertain, low consequence condition. Top part represents Fz electrode. The bottom panel depicts voltages in the F4 electrode. Black line is uncertain condition, red line is risk condition. The ellipsis shows the point of the significant difference. $Y$ axis represents micro voltages, X axis represents the epoch in milliseconds.

When high consequences are observed, there is a marginal statistical significance between a SCP positivity in risky conditions and a negativity in uncertain conditions in F4 between -920 ms and $-500 \mathrm{~ms}(\mathrm{~F}(1,98)=3.517, \mathrm{p}=0.06)$ as shown in figure 4.


Figure 4: SCPs in risk x uncertain, high consequence condition in F4. Black line represents uncertain condition, red line risk condition. The ellipsis shows the point of the significant difference. Y axis represents micro voltages, X axis represents the epoch in milliseconds.

## Discussion

Behavioral data suggests that the balance of acquired information (BAL), according to Bayesian calculations (Fifić \& Buckmann, 2013), is a preferred stopping rule. EEG data supports this conclusion given the fact that where BAL represented significant differences, there was the emergence of SCPs. According to Oswald and Sailer (2013), the SCPs are task-related and the negativity might mean conflict processing and the usage of cognitive resources to resolve such conflicts. Even though there was also a significant difference for the quantity of information bought and the decisions, consciously or not participants behave according to Bayesian calculation in order to determine the end of the information acquisition process.
This holds up even if the conditions are considered (combined or isolated). This means that the participants will take into account the valences of the information pieces acquired and when they reach a particular threshold (depending on the scenario characteristics), the decision is made. That becomes clearer when the threshold is approximately +1 for a positive decision, approximately -1 for a negative decision and approximately zero for a procrastination decision. The procrastination decisions show that even though there are more pieces of information acquired, participants often would feel more uncertain and would rather skip the decision. This means that that particular scenario and the set of information acquired would not diminish the residual uncertainty acknowledged by the participant, thus making it harder to assess which decision is better given the probabilities and consequences.

Uncertain scenarios needed less QTY and a higher BAL in order to reach a decision than risk scenarios. The appearance of the SCP negativity for uncertain scenarios can reflect a higher conflict in this condition given that, even though participants seek less information, they need higher valences to resolve the conflict. This conflict may arise due to the difficulty to assign a value to the unstated probability described in the scenarios. As in Stern et al. (2010) each new information can change the subjective probability that the participant assigns to the outcome. These changes can require more BAL and result in more use of cognitive resources in order to decide.

When conditions were combined, especially the two highlighted previously, the same effect is also present. In uncertain low negative consequence conditions there is the need for more BAL and there is also a SCP negativity although with higher amplitude than the one described on the last paragraph. This, according to Oswald and Sailer (2013), mean that there is an expanded cognitive effort in resolving the conflict that the valences and the condition might imply.

Lastly, in high negative consequence conditions, risky scenarios need more BAL, however, in high consequence conditions the SCP negativity is seen for uncertain scenarios. We hypothesize that the lack of stated probability in a high consequence scenario might mean that the information has a higher weight for the participants and therefore there is no need to allocate as much cognitive effort as with risky conditions. In this case, a stated probability might introduce some level of ambiguity given that the risk is apparent and the consequences can be large.

## Conclusion

We developed an experiment aiming to observe different strategies, or stopping rules, that individuals might use in order to cease information acquisition and make a decision in a given scenario. Departing from the stopping rules proposed by Fifić and Buckmann (2013), we manipulated scenarios in order to show or not show probabilities, high or low consequences and positive or negative consequences. The data suggests that individuals do not actually follow a particular stopping rule, rather they tend to use, consciously or not, Bayesian calculations in order to consider all the information that was bought in a scenario, when considering the decisions participants made. Moreover we found SCP waves for different conditions in the experiment. That can mean that for those conditions there was an expanded allocation of cognitive resources in order to solve conflicts that emerged from the information acquisition and the scenario description. Those manipulations showed that the information acquisition behavior resembled prospect theory (Tversky \& Kahneman, 1992) in that different levels of risk or uncertainty combined with high/low and positive/negative consequences will directly affect the quantity of information bought and the weight that the information will have in order for a participant to feel satisfied and proceed to a decision.

This was an exploratory experiment in order to study the moments leading to a decision in an information acquisition task. Further studies should focus on confirming the behavior and electrophysiological correlates of each condition separately. Also, there is an opportunity for the use of integrated psychophysiological measures in order to confirm task engagement and cognitive effort in those conditions (ECG and eyetracking, for example)

## Acknowledgments

This work was partially supported by CAPES/Brazil. We would like to thank Alicia Carmichael, Syed Tahsin, David

Zhang, Lisa Jin, Donna Walter, Karen Nielsen, Jeannette Jackson and Debra Bourque for help with this study.

## References

Böcker, K.B.E, Baas, J.M.P, Kenemans, J.L., \& Verbaten, M.N. (2001). Stimulus-preceding negativity induced be fear: a manifestation of affective anticipation. International Journal of Psychophysiology, 43, 77-90.
Courtney, H., Kirkland, J., \& Viguerie, P. (1997). Strategy under uncertainty. Harvard Business Review, (December), 81-90.
Di Caprio, D., Santos-Arteaga, F. J., \& Tavana, M. (2014). Information acquisition processes and their continuity: Transforming uncertainty into risk. Information Sciences, 274, 108-124. doi:10.1016/j.ins.2014.02.144
Fifić, M., \& Buckmann, M. (2013). Stopping Rule Selection (SRS) Theory Applied to Deferred Decision Making. In Proceedings of the 35th Annual Meeting of the Cognitive Science Society (pp. 2273-2278). Retrieved from http://mindmodeling.org/cogsci2013/papers/0415/paper04 15.pdf

Frey, R., Hertwig, R., \& Rieskamp, J. (2014). Fear shapes information acquisition in decisions from experience. Cognition, 132(1), 90-99. doi:10.1016/j.cognition.2014.03.009
Gigerenzer, G. (2000). Adaptive thinking: Rationality in the real world. New York: Oxford University Press.
Lopez-Calderon, J., \& Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience, 8:213. doi: 10.3389/fnhum. 2014.00213

Oswald, F., \& Sailer, U. (2013). Slow cortical potentials capture decisions processes during temporal discounting. European Journal of Neuroscience, 37, 1159-1169. doi:10.1111/ejn. 12108
Pretz, J. E., Naples, A. J., \& Sternberg, R. J. (2003). Recognizing, defining and representing problems. In The Psychology of Problem Solving. Cambridge University Press.
Söllner, A., Bröder, A., Glöckner, A., \& Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. Acta Psychologica, 146, 84-96. doi:10.1016/j.actpsy.2013.12.007
Stern, E. R., Gonzalez, R., Welsh, R. C., \& Taylor, S. F. (2010). Updating beliefs for a decision: neural correlates of uncertainty and underconfidence. The Journal of Neuroscience, 30(23), 8032-41. doi:10.1523/JNEUROSCI.4729-09.20104
Taghavifard, M. T., Damghani, K. K., \& Moghaddam, R. T. (2009). Decision Making Under Uncertain and Risky Situations. In Enterprise Risk Management Symposium Monograph Society of Actuaries. Chicago, IL.
Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297-323.

# Solving additive word problems: Intuitive strategies make the difference 

Katarina Gvozdic (katarina.gvozdic@gmail.com)<br>Paragraphe Lab, EA 349, University Paris 8, Department of Psychology, 2 Rue de la Liberté, 93526 Saint-Denis Cedex 02, France

Emmanuel Sander (emmanuel.sander@univ-paris8.fr)<br>Paragraphe Lab, EA 349, University Paris 8, Department of Psychology, 2 Rue de la Liberté, 93526 Saint-Denis Cedex 02, France


#### Abstract

Young children use informal strategies to solve arithmetic word problems. The Situation Strategy First (SSF) framework claims that these strategies prevail even after instruction. The present study was conducted with second grade students in order to investigate the persistence of intuitive, situationbased strategies, on word problems that do not involve dynamic temporal changes. This is challenging for the SSF framework, since the lack of this dimension might bypass intuitive strategies. The results revealed that intuitive strategies persist, are valid for these types of problems, and impact the problems' difficulty. Indeed problems that require the application of arithmetic principles remain hard, even though they have been practiced at school. These findings provide complementary evidence to how mental calculation strategies articulate with arithmetic word problem solving and call for the extension of the $S S F$ framework.


Keywords:arithmetic word problems; problem solving; informal strategies; solution strategies; education.

## Introduction

Even before instruction young children can solve arithmetic word problems by usinginformal strategies (Verschaffel\& De Corte, 1997). These informal strategies reflect the situation described in the problem and preclude the flexible application of mathematical principles like commutativity, inversion or distributivity (Verschaffel \& De Corte, 1997).During the early years of elementary school, children improve their numerical competencies and acquire certain mathematical principles, which could lead us to expect that newly acquired arithmetic competencies would take place over the informal strategies.

Indeed, numerous mental calculation strategies that schooled children develop to solve problems when presented in their arithmetic expression (e.g.'8-5=') have been documented (e.g. Carpenter, Ansell, Franke, Fennema, \& Weisbeck, 1993; Torbeyns, De Smedt, Ghesquière, \& Verschaffel, 2009). They are mainly determined by the arithmetic operation that provides the solution. For subtraction problems, the principal distinction bears between direct subtraction strategies in which the subtrahend is straightforwardly taken away from the minuend (e.g. in which '42-39 =' is solved by '42-39'), and indirect addition strategies in which the calculation consists in finding how much needs to be added to the minuend to reach the subtrahend (e.g. in which ' $42-39=$ ' is solved by '39 + . = 42'). In both of these strategies, the arithmetic
operation that is used is subtraction, it is just the arithmetic format that is different (Campbell, 2008). In order to describe how students use the two strategies, Peters, DeSmedt, Torbeyns, Ghesquière, and Verschaffel (2013) provided empirical support for their Switch model. According to this model, students solve two-digit subtraction problems by switching between direct subtraction and indirect addition depending on the combination of the magnitude of the subtrahend and the numerical distance between the subtrahend and the minuend.

Brissiaud and Sander (2010) investigated how these mental calculation strategies articulate with the informal strategies students use on arithmetic word problem solving. They proposed a Situation Strategy First (SSF) framework which posits that the initial representation of a problem activates asituation-based strategy, both before and after instruction. Only when this strategy is not efficientthe representation of the problem may be modified and a set of arithmetic principles may be applied in order to provide an adequate solution in a more efficient way. In their experiments, each problem was presented to second and third grade students in two versions. The first version could be efficiently solved by mentally simulating the actions described in the problem - situation strategy problems ( $\mathbf{S i}$ problems). For example:
I. Luc is playing with his 42 marbles at recess. During the recess, he loses 3 marbles. How many marbles does Luc have now? [42-3= .]
Problem I is an Si-problem because simulating the action of losing 3 marbles through mentally counting down from 42 is easy to perform( 41 (1), 40 (2), 39 (3)). Thus, a situation-based solving strategy, modeling the described situation - the Si-strategy - is efficient.

For each Si-problem, a Mental Arithmetic counterpart was introduced (MA-problem). MA-problems are problems for which mental simulation is too costly to attain the result - thus for which the Si-strategy is not efficient. On the contrary, the use and application of arithmetic knowledge is efficient and makes the problem easy (MA-strategy). For example:
II. Luc is playing with his 42 marbles at recess. During the recess, he loses 39 marbles. How many marbles does
Luc have now? [42-39 = .]
The solution to problem II cannot be efficiently obtained by using the same procedure as for the first one; mentally simulating the action by counting down 39 marbles would be too costly. However the mental subtraction $42-39$ is
easy when the complement principle is mastered and leads to counting up from 39 to 42 .

The findings revealed that even after instruction, the Si problems remained systematically and significantly easier than the corresponding MA-problems. Furthermore, a higher use of informal strategies was observed on Siproblems, while arithmetic principles were almost exclusively used on MA-problems. For instance, when students succeeded to solve an Si-problem such as Problem I, they exclusively used a direct subtraction strategy, which is the Si-strategy in this case. However, when they succeeded to solve an MA-problem such as Problem II, even though Si-strategies were (scarcely)observed, they were solved to a greater extent by MA-strategies, such as looking for a missing addend in the previous example (' $39+$. $=$ 42'). This was never observed for Si-problems: no child tried to solve a problem such as Problem I by the missing addend' $3+$. $=42$ '.

Indeed, the arithmetic computations of both Si - and MAstrategies on subtraction problems are executed by the aforementioned mental computation strategies. The Switch model could accurately account for how the various arithmetic characteristics of the problems tested so far by the $S S F$ framework yield a clear computational advantage for one strategy over another. However, the Switch model does not provide an explanation for why students fail to apply arithmetic principles, such as it is observed through the significantly lower success rates on MA-problems. Indeed, even though the Switch model accurately describes the numerical conditions that require a switch between direct subtraction and indirect addition, it does not account for the mental re-representation needed in order to make this switch when a presented strategy cannot be easily performed in the same format as the one it is presented in.

We propose that the attainment of a mental rerepresentation would reflect an underlying conceptual metaphor that guides the interpretation and application of arithmetic principles. Conceptual metaphors are based on everyday human experience. The underlying mathematical ideas are constructed through cognitive mechanisms called fictive motion, which refer to the conception of static entities in dynamic terms (Lakoff \& Núñez, 2001). One of the main representations of arithmetic is object collection (Lakoff \& Núñez, 2001). The most widespread conceptual metaphor of subtraction that can be drawn from itis "taking away" (Fischbein, 1989; Lakoff and Núñez, 2001). Alternatively, arithmetic can be considered as motion along a path (Lakoff \& Núñez, 2001). The conceptual metaphor of subtraction that can be drawn from this conception is subtraction as a measuring stick (Lakoff \& Núñez, 2001), or as "determining the difference" (Selter, Prediger, Nührenbörger \& Hußmann, 2012). As Selter and collaborators (2012) pointed out, the "taking away" model might be more widespread, however seeing subtraction
solely as "taking away" is too one-sided, and both models are required in order to be flexible in mental arithmetic.

We consider that the failure to apply arithmetic principles on MA-problems is due to a restrictive representation of arithmetic, an intuitive representation(such as the "taking away" model), which entails a limited interpretation of the arithmetic situation embedded in the problem statement. Such an extension of the SSF framework would also challenge the most commonly used classification of arithmetic word problems introduced by Riley, Greeno and Heller (1983). Their classification determines the difficulty of a problem based on the semantic category it belongs to, while the $S S F$ framework puts emphasis on situation-based strategies and proposes thatthe efficiency of such strategies would be also a determining factor of difficulty.

Yet, all the subtraction problems that were tested by Brissiaud and Sander (2010) belonged to one same category of subtraction problems from Riley, Greeno and Heller's (1983) classification - change problems. These problems are dynamic in nature and describe an action with a temporal dimension,soliciting a mental simulation. However, the other problem categories do not involve this temporal dimension. They have been identified as more difficult than change problems, especially "compare" problems, in which a comparison between two quantities is involved and the question bears on the difference or on one of the compared quantities. It therefore remains an open issue if the mental simulation advocated by the $S S F$ framework is still relevant for problems that do not unfold along a temporal dimension.

Indeed, if the mental simulation of the problem was not solicited, then we could expect that the distinction between Si- and MA-problems among these categories would lose its relevance. It therefore remains an open issue if the mental simulation advocated by the SSF framework is still relevant for problems that do not unfold along a temporal dimension. If it would be demonstrated that the efficiency of the mental simulation influences a problem's difficulty even when it does not develop along a temporal timeline, it would warrant a broader view of Si -strategies and provide a new criterion for the assessment of problem difficulty, not based only on the semantic category, but also on the efficiency of the Si- strategy.

## Aim of the study

The purpose of the study was to demonstrate that the mental simulation of the arithmetic relations is not a mere consequence of a dynamic semantics of the problem, but an intrinsic property of arithmetic problem solving. Firstly, we conducted a longitudinal study in order to test the distinction between Si - and MA-problems in contexts less favorable for a mental simulation. Secondly, we conducted individual verbal reports in order to gather confirmatory evidence of situation-based strategies for Si-problems and a switch to non-situation-based strategies for MA-problems.

Table 1: Example of the problems for the number set (31, 27, 4) presented with different contexts

| Problem categories |  | Si problems | MA problems |
| :---: | :---: | :---: | :---: |
| Comparison problems | $\mathrm{D}[\mathrm{b}+.=\mathrm{a}]$ | There are 27 roses and 31 daisies in the bouquet. How many daisies are there more than roses in the bouquet? | There are 4 roses and 31 daisies in the bouquet. How many daisies are there more than roses in the bouquet? |
|  | $\mathrm{D}[\mathrm{a}-.=\mathrm{b}]$ | There are 31 oranges and 27 pears in the basket. How many pears are there less than oranges in the basket? | There are 31 oranges and 4 pears in the basket. How many pears are there less than oranges in the basket? |
|  | $C\left[b+b^{\prime}=.\right]$ $C[a-b=$. | James has 27 marbles. Steve has 4 marbles more than James. How many marbles does Steve have? <br> Anna has 31 euros. Susan has 4 euros less than Anna. How many euros does Susan have? | James has 4 marbles. Steve has 27 marbles more than James. How many marbles does Steve have? <br> Anna has 31 euros. Susan has 27 euros less than Anna. How many euros does Susan have? |
| Equalizingproblem S | $\mathrm{E}[\mathrm{b}+.=\mathrm{a}]$ | There are 27 oranges and 31 pears in the basket. How many oranges should we add to have as many oranges as we do pears? | There are 4 oranges and 31 pears in the basket. How many oranges should we add to have as many oranges as we do pears? |
|  | $\mathrm{E}[\mathrm{a}-.=\mathrm{b}]$ | There are 31 roses and 27 daisies in the bouquet. How many roses should we take away in order to have as many roses as we do daisies? | There are 31 roses and 4 daisies in the bouquet. How many roses should we take away in order to have as many roses as we do daisies? |
| Combine problems | $\mathrm{S}[\mathrm{b}+.=\mathrm{a}]$ | Mary has 27 euros in her piggybank and she has euros in her pocket. In total, Mary has 31 euros. How many euros does Mary have in her pocket? | Mary has 4 euros in her piggybank and she has euros in her pocket. In total, Mary has 31 euros. How many euros does Mary have in her pocket? |
|  | $S\left[b+b^{\prime}=.\right]$ | There are 27 blue marbles and 4 red marbles in Marc's bag. How many marbles are there in Marc's bag? | There are 4 blue marbles and 27 red marbles in Marc's bag. How many marbles are there in Marc's bag? |

## Experiment 1

## Method

## Participants

269 second grade students from 13 classes in 7 schools from working-class neighborhoods participated in the study. The average age of the children in January, when the first test was passed, was 7.62 years ( $s d=0.32,138$ girls).

## Material

There were 8 addition and subtraction problem types belonging to 3 major categories:

- compare problems: difference $\operatorname{set}(\mathrm{D}[\mathrm{b}+\ldots=\mathrm{a}$, $\mathrm{D}[\mathrm{a}-\mathrm{F}=\mathrm{b}])$ and compared set $\left(\mathrm{C}\left[\mathrm{b}+\mathrm{b}^{\prime}=.\right], \mathrm{C}[\mathrm{a}-\right.$ $\mathrm{b}=$.$] ),$
- equalizing problems $(\mathrm{E}[\mathrm{b}+.=\mathrm{a}], \mathrm{E}[\mathrm{a}-.=\mathrm{b}])$,
- combine problems $\left(\mathrm{S}[\mathrm{b}+.=\mathrm{a}], \mathrm{S}\left[\mathrm{b}+\mathrm{b}^{\prime}=.\right]\right)$.

The subtraction problems involved two numbers, $a$ and $b$ $(a>b)$. The numerical values for $a$ were either 42, 41, 33 or 31 , while in order to differentiate between Si - and MAproblems the values for $b$ were either kept small ( 3 or 4) or were close to $a(39,38,29$ or 27). To create Si-problems the small value of $b$ was used for the $\mathrm{C}[\mathrm{a}-\mathrm{b}=$.$] , while the b$
value close to $a$ was used for $\mathrm{D}[\mathrm{b}+\ldots=\mathrm{a}], \mathrm{D}[\mathrm{a}-.=\mathrm{b}]$,
 $\mathrm{S}\left[\mathrm{b}-\mathrm{b}^{\prime}=\right.$. ]. To create MA-problems the opposite $b$ value was respectively used for each problem, since it would make the Si-strategy costly.

Addition problems $\mathrm{S}\left[\mathrm{b}+\mathrm{b}^{\prime}=\right.$. ] and $\mathrm{C}\left[\mathrm{b}+\mathrm{b}^{\prime}=\right.$. $]$ involved two numbers, $b$ and $b^{\prime}$. Both numbers had the same characteristics as $b$ for subtraction problems, while the unknown value was equivalent to $a$. To create Si-problems the $b$ value close to $a$ was presented first, while the small $b$ value ( $b^{\prime}$ ) was presented second. To create MA-problems they were presented in the opposite order.

Thus the numbers involved in the data and the solution are $(31,27,4),(33,29,4),(41,38,3)$, and $(42,39,3)$.

Note that the number size was not the determining factor in the Si vs. MA-problem distinction. In the Si - versions one problem had the $b$ value close to $a$, while others had small $b$ values ${ }^{1}$. Also the second experiment was conducted to support this, by directly investigating students' strategy use.

[^345]Four different contexts were used for the wording of the problems: marbles, euros, flowers and fruits.

Table 1 provides examples of each problem category.

## Design

Children solved a total of 8 problems created by combining the 8 problems categories in either their Si - or MA- version. Each student therefore solved 4 Si-problems and 4-MA-problems. To control for the impact of position, numerical sets and context, 8 different problem sets were created. Another 8 problem sets were 'mirror' sets in which the Si-version of one problem would be presented in its MA counterpart, while the MA-problem would be presented in its Si-counterpart. Thus, 16 groups of problem sets were created altogether and counterbalanced across classrooms.

## Procedure

The experiment was composed of two sessions. The first conducted in January and the second one, strictly identical to the first one, 6 months later, in June. It was administered in the students' classrooms. Each child received an 8 pages booklet. There was a square in the middle of each page in which they wrote their answer. Each problem was read aloud twice to the whole classroom and children had one minute to write down the number that was the solution.

## Scoring

The solutions provided by the children were scored with 1 point when the numerical answer was exact, or within the range of plus or minus one of the exact value, in order to take into account mistakes in counting procedures. Any other answer received 0 points. The average of the sum of the scores on Si- and MA-problems was used as the dependent variable and analyses on these scores were carried out.

## Results

A first analysis was conducted in order to compare children's average success rates on Si - and MA-problems at the beginning of the year, followed by a second set of analyses in order to compare the success rates on the problems at the end of the school year. A third analysis bore on the progression over the year.

Repeated measure ANOVAs, with the 'Si- versus MAproblems' variable (further referred to as Problem type)as within-participant independent variables, were conducted for each session. The analyses of the scores obtained in January showed a highly significant main effect of Problem type on performance $\quad(F(1, \quad 268)=98.39, \quad p<$ $0.001, \eta^{2}=0.11$ ).Table2displays the average success rates. Indeed the Si-problems had a $19.57 \%$ higher success rate than MA-problems.

In June, there was a significant difference in performance between the two times of testing on Si-problems $(F(1,268)=$ 86.39, $\quad p<.001, \quad \eta^{2}=0.06$ ) and on MA-problems $\left(F(1,268)=36.58, p<.001, \eta^{2}=0.05\right)$. Yet, in accordance with our hypotheses, the results still revealed a highly significant main effect of Problem type on performance in June $\left(F(1,268)=119.57, p<.001, \eta^{2}=0.13\right)$.As displayed in Table2, the Si-problems had a $24.38 \%$ higher success rate than MAproblems in June (experiment 2).

Table 2: Average success rates

| Averagesuccess rate | January | June |
| :---: | :---: | :---: |
| Si-problems | $47.86 \%$ | $64.53 \%$ |
| MA-problems | $28.25 \%$ | $40.15 \%$ |

Indeed, after performing a repeated measure ANOVA with the Problem type and the times of testing as withinparticipant independent variables, the results confirmed that there was a significant main effect of Problem type $\left(F(1,268)=171.64, p<.001, \eta^{2}=0.12\right)$ and a main effect of the Time of testing $\left(F(1,268)=106.19, p<.001, \eta^{2}=0.05\right)$, but most importantly there was no interaction between the two variables $\left(F(1,268)=3.51, \quad p>.1, \quad \eta^{2}=0.001\right)$. Thus, as hypothesized, despite the progress made on each problem type throughout the year, the gap in performance persisted between Si - and MA-problems.

In order to test the hypotheses problem per problem, univariate ANOVAs, with the Problem type variable, were conducted for each of the eight problem categories and showed that almost all of the Si-problems were significantly easier than the corresponding MA-problems both in January and in June: $\mathrm{D}[\mathrm{a}-.=\mathrm{b}], \mathrm{C}\left[\mathrm{b}+\mathrm{b}^{\prime}=.\right], \mathrm{C}[\mathrm{a}-\mathrm{b}=],. \mathrm{E}[\mathrm{b}+$. $=\mathrm{a}], \mathrm{E}[\mathrm{a}-.=\mathrm{b}]$ and $\mathrm{S}[\mathrm{b}+.=\mathrm{a}](3.843<F(1,267)<72.501$, $p<.01,0.01<\eta^{2}<0.20$. The $\mathrm{D}[\mathrm{b}+.=\mathrm{a}]$ seemed to be particularly hard in January when no difference was observed $\left(F(1,267)=0.23, p>.1, \eta^{2}=0.001\right)(27.6 \%$ success rate on Si - and $25 \%$ on MA-problems), but the Si- versus MA-distinction was valid at the end of the year $\left(F(1,267)=15.63, p<.001, \eta^{2}=0.06\right)(47 \%$ success rate on $\mathrm{Si}-$ and $24 \%$ on MA-problems). The single exception for which no difference was observed on either time of testing was the combine superset problem $\mathrm{S}\left[\mathrm{b}+\mathrm{b}^{\prime}=\right.$. ] (January $\mathrm{F}(1,267)=2.69, p>.1, \eta^{2}=0.01$, June $F(1,267)=1.223, p>.1$, $\eta^{2}=0.005$ ), for which a difference was observed in the expected direction but not confirmed by the test ( $75 \%$ and $83 \%$ success rate on Si-problems and $66 \%$ and $77 \%$ on MAproblems, in January and June respectively).

## Discussion

The results revealed that the distinction between Si - and MA-problems remain relevant for subtraction and addition word problems that do not evolve along a temporal timeline. Our study shows that indeed, problems efficiently solved by direct modeling strategies remain easier for students even after they acquired more advanced skills in mathematics at the end of the year. The progression between the two sessions did not obliterate the distinction between Si -and MA-problems. The similar progression on Si - and MA- problems might be explained by the advances children made in computational execution of the calculations, or regarding their general comprehension skills.

A second experiment was conducted in order to provide confirmatory evidence that the difference in difficulty between Si - and MA-problems actually results from the preferential use of Si-strategies when they are efficient and for the lack in the application of arithmetic principles when this strategy is inefficient.

Table 3 : solving strategies for each problem category (with an example of the number set (31, 27, 4))

| Problem category | Si-problems |  |  |  | MA-problems |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | \% correct responses with described strategy | Si-strategy | Non-SI-strategies |  | \% correct responses with described strategy | Si-strategy <br> Direct modelling strategy | Non-SI-strategies |  |
|  |  | Direct modelling strategy | MA- <br> strategy | Other |  |  | MA- <br> strategy | Other |
| $\mathrm{D}[\mathrm{b}+.=\mathrm{a}]$ | 66.67\% | $27+$. $=31$ | 31-27= | 31-. $=27$ |  | $4+$. $=31$ | 31-4= | 31-. $=4$ |
|  |  | 92.86\% | 0\% | 7.14\% | 19.05\% | 50\% | 50\% | 0\% |
|  |  | 31-. $=27$ | 31-27= | $27+$ = 31 |  | 31-. $=4$ | 31-4= . | $4+$. $=31$ |
| $\mathrm{D}[\mathrm{a}-.=\mathrm{b}]$ | 55.56\% | 50\% | 0\% | 50\% | 33.33\% | 16.67\% | 83.33\% | 0\% |
|  |  | 31-4= | $4+.=31$ | 31-. $=4$ |  | $31-27=$. | $27+$. $=31$ | 31-. $=27$ |
| $\mathrm{C}[\mathrm{a}-\mathrm{b}=$. | 47.62\% | 100\% | 0\% | 0\% | 19.05\% | 50\% | 50\% | 0\% |
|  |  | $27+$. $=31$ | 31-27= . | 31-. $=27$ |  | $4+.=31$ | 31-4= . | $+4=31$ |
| $\mathrm{E}[\mathrm{b}+.=\mathrm{a}]$ | 47.62\% | 100\% | 0\% | 0\% | 38.10\% | 50.00\% | 25.00\% | 25.00\% |
|  |  | 31-. $=27$ | 31-27= . | 27+. $=31$ |  | 31-. $=4$ | 31-4= . | $4+$. $=31$ |
| $\mathrm{E}[\mathrm{a}-.=\mathrm{b}]$ | 66.67\% | $71.43 \%$ | 14.29\% | 14.29\% | 38.10\% | 12.50\% | 75\% | 12.50\% |
|  |  | $27+$. $=31$ | $31-27=$. | 31-. $=27$ |  | $4+.=31$ | 31-4= | 31-. $=4$ |
| $\mathrm{S}[\mathrm{b}+.=\mathrm{a}]$ | 47.62\% | 90\% | 0\% | 10\% | 33.33\% | 28.57\% | 71.43\% | 0\% |
|  |  | $27+4=$ | $4+27=$. |  |  | $4+27=$. | $27+4=$. |  |
| $S\left[b+b^{\prime}=.\right]$ | 85.71\% | 100\% | 0\% |  | 90.48\% | 5.26\% | 94.74\% |  |
|  |  | $27+4=$. | $4+27=$. |  | 23.81\% | $4+27=$. | $27+4=$. |  |
| $\mathrm{C}\left[\mathrm{b}+\mathrm{b}^{\prime}=.\right]$ | 61.90\% | 100\% | 0\% |  |  | 0\% | 100\% |  |

## Experiment 2

We collected additional information concerning the strategies children actually use when solving Si- and MAproblems. We asked them to solve problems and then to describe their solving strategy aloud. We predicted that the solution strategies which directly model the problem would be predominant for Si-problems but that alternative strategies would emerge for MA-problems.

## Method

## Participants

42 Grade 2 students from 4 classes in2 different schools from working-class neighborhoods participated in the study. The test occurred in June and the average age of the children on the test was 7.93 years ( $\mathrm{sd}=0.26,23$ girls). None of the participants participated in the previous experiment.

## Material \& Design

The same material and design was used as in the first experiment.

Concerning the evaluated strategies, if we take the $\mathrm{D}[\mathrm{a}-$. $=b]$ problem as an example, the Si- strategy used to solve it is to start from the largest presented quantity (31) and to double-count downwards until the second quantity is reached: in the Si-problem this would not be costly. The students would describe their solving process as starting
from 31 and counting down30(1), 29(2), 28(3), 27(4), bearing the answer 4 , and noted by the experimenter as 31-. $=27$.Yet using the same Si-strategy of double-counting downward in the MA-problem to get from 31 to 4 is a costly procedure. When students would use this strategy they would describe the same solving process: starting at 31 and counting down30 (1), 29(2), 28(3), ... 5(26), 4(27), bearing the answer 27 and noted by the experimenter as $31-.=4$.

Nevertheless, when applying arithmetic knowledge we can easily know that taking away 4 from 31 provides the correct numerical answer to this MA-problem. One of the possible descriptions of the students' solving process would be to start from 31 and take away 4 , with the result no longer being the number of times they counted down, but the number they reached. This Non-Si-,mental arithmetic strategy (MA-strategy) would be noted as '31-4 = .'.

## Procedure

The procedure was identical to the first experiment, except that the test was conducted individually in the school library and that after writing down the numerical answer, the student was asked to explain aloud how he or she found the solution. The possible strategies were established beforehand and there was no ambiguity in their coding. The strategies that the students reported were classified according to table 3 into Si-strategies when the strategy directly modeled the wording of the problem, or into Non-

Si-strategies when the strategy that the student described did not directly model the problem.

## Scoring

For both Si- and MA-problems, we computed a score of Si-strategies (Si-score) and Non-Si-strategies (Non-Siscore). If a pupil provided a correct answer and explained a strategy, the nature of this strategy was assessed and contributed 1 point to either the Si -strategy score or the Non-Si-strategy score of the problem type. No points were attributed if a student did not provide a correct response and/or did not describe any strategy after providing the correct answer (only $7.5 \%$ of the correct responses were not accompanied by a strategy description). Given that children solved 4 problems of each type, the scores ranged from 0 to 4.

## Results

The experiment replicated the previous findings, confirming that Si-problems were easier for children than MA-problems. The success rates were $67.2 \%$ and $41.5 \%$ respectively, and the variance analysis revealed that this difference was significant $\left(F(1,41)=17.86, p<.001, \eta^{2}=0.13\right)$.

Table 3 shows strategy use for each problem category and the disparities between the two kinds of strategies, among students that provided the right numerical solution and described a strategy.

We further performed two variance analyses using the Si strategy score and the Non-Si-strategy score as the dependent variables, and Problem type as the within factor variable. The average scores are presented in table 4.As expected, both differences were significant. Si-strategies were used significantly more on Si-problems $(F(1,38)=79.1$, $\left.p<.001, \eta^{2}=.5\right)$, as well as Non-Si-strategies on MAproblems $\left(F(1,38)=20.06, p<.001, \eta^{2}=.2\right)$.

Table 4 : Si- and Non-Si-score for solving strategies

| Problem type | Si-score | Non-Si-score |
| :---: | :---: | :---: |
| Si-problem | 2.05 | 0.31 |
| MA-problem | 0.26 | 1.13 |

## Discussion

The variance analyses confirmed that solution strategies which directly model the situation were predominant for Si problems and that solution strategies which required arithmetic knowledge were predominant for MA-problems. These findings suggest that the selected strategy drives the difference in performance on top of the problem category or factors such as mental calculation competences.

## General Discussion and Conclusion

The experiments conducted in the present study account for the spontaneous and intuitive modeling of the situation described in arithmetic word problems, which leads to a primary use of situation strategies and the application of arithmetic principles only when the first one is too costly. The significant difference that was observed between Siand MA-problems fits with the previous findings on change problems (Brissiaud \& Sander, 2010), and confirms that this
problem distinction is not specific to problems that evolve along a temporal timeline. These findings complement the traditional classification of arithmetic word problems according to which problem difficulty depends mostly on the problem category. They also provide evidence that situation strategies are not only tied to the semantic wording of a problem, but could be a fundamental property solicited by arithmetic problems.

## Acknowledgments

This study was supported by the "Fonds d'expérimentation pour la jeunesse" through the funding grant ACE_HAP_10for the project "ACE-ArithmEcole".

We thank the teachers and classes who participated in the the study, Adeline Lucchesi for her help with data collection and Sébastien Puma for his helpful comments.

## References

Brissiaud, R., \& Sander, E. (2010). Arithmetic word problem solving: a Situation Strategy First framework. Developmental Science, 13(1), 92-107.
Campbell, J. I. (2008). Subtraction by addition. Memory \& Cognition, 36(6), 1094-1102.
Carpenter, T.P., Ansell, E., Franke, M.L., Fennema, E., \&Weisbeck, L. (1993). Models of problem solving: a study of kindergarten children's problem-solving processes. Journal for Research in Mathematics Education, 24, 428-441.
Fischbein, E. (1989). Tacit Models and Mathematical Reasoning. For the Learning of Mathematics, 9(June), 914.

Lakoff, G., \& Núñez, R. (2001). Where mathematics comes from: How the embodied mind brings mathematics into being. New York,NY: Basic Books
Peters, G., Smedt, B., Torbeyns, J., Ghesquière, P., \& Verschaffel, L. (2013). Children's use of addition to solve two-digit subtraction problems. British Journal of Psychology, 104(4), 495-511.
Riley, M.S., Greeno, J.G., \& Heller, J.I. (1983). Developmentof children's problem solving ability in arithmetic.In H.P.Ginsburg (Ed.), The development of mathematical thinking. New York: Academic Press.
Selter, C., Prediger, S., Nührenbörger, M., \& Hußmann, S. (2012).Taking away and determining the difference - A longitudinal perspective on two models of subtraction and the inverse relation to addition. Educational Studies in Mathematics, 79(3), 389-408.
Torbeyns, J., De Smedt, B., Ghesquière, P., \&Verschaffel, L. (2009). Acquisition and use of shortcut strategies by traditionally schooled children. Educational Studies in Mathematics, 71,1-17.
Verschaffel, L., \& De Corte, E. (1997). Word problems: avehicle for promoting authentic mathematical understanding and problem solving in the primary school? In T. Nunes \& P. Bryant (Eds.), Learning and teaching mathematics: An international perspective(pp. 69-97). Hove, UK: Psychology Press.

# Age differences in language comprehension during driving: Recovery from prediction errors is more effortful for older adults 

Katja Häuser (khaeuser@coli.uni-saarland.de)<br>Saarland University<br>Campus C7.4<br>66123 Saarbrücken<br>Germany<br>Vera Demberg (vera@coli.uni-saarland.de)<br>Saarland University<br>Campus C7.4<br>66123 Saarbrücken<br>Germany<br>Jutta Kray (j.kray@mx.uni-saarland.de)<br>Saarland University<br>Campus A1.3, 66123 Saarbrücken<br>Germany


#### Abstract

Prior research yielded conflicting findings regarding whether older adults show a greater processing cost than younger adults when encountering unpredicted semantic material during language processing. Here, we investigated whether age-related differences in recovery from prediction error are influenced by increased demands on working memory. We used a dual task design: a primary sentence comprehension task in which semantic predictions were fulfilled or violated, and a concurrent driving task, thought to limit working memory resources in resolving prediction errors. In the dual task, older participants showed an increase in comprehension accuracy for sentences with semantic violations, while demonstrating a decrease in driving accuracy. Thus, when working memory resources were limited, older adults focused exclusively on the language task and neglected the driving task. This could be related to an age-related increase in generating semantic predictions, or to a general inability among older adults to divide attention between two cognitively demanding tasks.


Keywords: aging, semantic expectancy, dual tasking, attention allocation

## Introduction

Prediction of upcoming linguistic material is pervasive during language comprehension. Recent theories hold that expectations at higher levels of processing (e.g., syntactic, contextual) generate hypotheses and facilitate low-level processing, for example in word recognition (Kuperberg \& Jaeger, 2016).

Frequently, however, people encounter unpredicted linguistic content and must recover from unexpected events that violate their expectations. Indeed, research has shown that this recovery phase often involves a processing cost. For example, Federmeier and Kutas (1999) analyzed the N400 EEG component (for review, see Kutas \& Federmeier, 2000) to index comprehension difficulties when participants were reading unexpected sentence continuations. Sentences contained either an expected word, an unexpected word from the same semantic category, or an unexpected word from a different semantic category (e.g., They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of palms (expected) / pines (unexpected same category) / tulips (unexpected different category)). According to the results, the N400 was reduced for expected and semantically related words, indicating that processing of predicted (palms) and semantically related words (pines) was facilitated. In contrast, for semantically unexpected words (tulips) the N400 amplitude was high, suggesting comprehension difficulties among participants when predictions based on context were violated.

However, an unanswered question is whether older adults (65 years or older) use context to anticipate upcoming content during language processing in a similar fashion as younger adults do. Some studies have shown that older adults are more disturbed by unpredicted semantic material than younger adults, which suggests that older adults may rely more heavily on prediction making during language processing (DeLong, Groppe, Urbach, \& Kutas, 2012; DeDe, 2014; Rayner, 2006; Borges \& Coco, 2015). For example, Borges and Coco (2015) investigated age differences in visual object detection by using a priming paradigm in which prime and search scene were either
congruent (e.g., kitchen-kitchen) or not (e.g., bathroomkitchen). In addition, visual target objects (e.g., bread basket) were presented in a semantically consistent condition (e.g., on a restaurant table) or in an inconsistent condition (e.g., on a pool table). According to the results, older adults were less successful at detecting target objects when prime and search scene were semantically congruent, but the target was inconsistent with the search scene. The authors concluded that older adults rely more heavily on contextual expectations than younger adults by generating very specific predictions based on consistent information. Consequently, they showed a greater processing cost when expectations based on context and new information are inconsistent.

Other studies, in contrast, have shown that older adults generally appear less likely than younger adults to use context and engage in pre-activating information during sentence processing (e.g., Federmeier \& Kutas, 2005; Federmeier, Kutas, \& Schul, 2010). For example, Federmeier and Kutas (2005) compared younger and older adults' N400 amplitude for sentence-final words in highly and weakly constraining contexts (e.g., highly constraining: No one at the reunion recognized Dan because he had grown a beard; weakly constraining: At the children's park next to the beach she saw a man with a beard). Even though both age groups showed a similar N 400 for weakly constraining sentences, the older adults' brain response for strongly constraining sentences was delayed and diminished in shape. The authors concluded that older adults were unable to make use of the richer information available from strongly constraining contexts to guide semantic processing; possibly because age-related declines in working memory prevented older adults from quickly constructing and updating an ongoing message-level representation while at the same time processing new input (see Huettig \& Janse, 2016, for a similar account).

Given Federmeier and Kutas' (2005) implication of working-memory capacity, the goal of the present study was to examine whether age-related changes in predictive processing are influenced by increasing demands on working memory. We used a dual-task paradigm with a primary language comprehension task and secondary driving task, thought to limit cognitive resources that participants can expend to resolving semantic prediction errors. To our knowledge, only one previous study has investigated how aging affects dual-task performance during language processing and driving, and that particular study found an age effect that was limited in scope.

Becic et al. (2010) investigated story-retelling ability in younger and older adults while participants were engaged in a secondary driving task. According to the results, younger adults achieved high accuracy in both story retelling and driving, suggesting high capacity in this participant group to divide attention between the language and the driving task. For older adults, no reliable effects emerged in the primary analysis. However, there was a trend in the data (revealed by post-hoc tests), which suggested that better driving (less
variability in velocity and lane keeping) was associated with worse retelling. In other words, older adults who drove better also performed more poorly in the language task. Since the group of older adults showed worse story-retelling ability overall, it seemed that older adults primarily focused on getting the driving task right, while neglecting the language task. The authors suggested that, due to agerelated declines in working-memory capacity, older adults may be more likely to protect their driving by giving up on the story retelling task. However, the Becic et al. (2010) study remains somewhat mute with respect to age differences in predictive processing (the primary focus of the present study), since this question was not specifically addressed by that paper.

In the present study, we sought to adress age-related differences in recovery from prediction error more directly, by presenting stimuli in a low- vs. high-surprisal condition (e.g., Since Petra didn't have anything to wear for the barbeque, she bought a dress (low surprisal) in a nearby shop; vs. Since Petra didn't have anything to drink for the barbeque, she bought a dress (high surprisal) in a nearby shop). High-surprisal sentences were thought to induce a strong cognitive conflict since the second clause violated semantic predictions based on contextual information provided by the first clause (i.e., drink-dress).

Based on prior research on semantic surprisal in younger adults (DeLong, Troyer, \& Kutas, 2014; Kutas \& Federmeier, 2000), we predicted that younger adults should be sensitive to violations of semantic expectancies (probably indexed by lower accuracy for high-surprisal sentences in the language comprehension task). In addition, we expected stable performance in this participant group regardless of whether sentences were processed in the single or dual task, indicating high capacity in younger adults to divide attention even under conditions of high linguistic load (cf. Becic et al., 2010).

In contrast, for older adults our predictions were less clear based on previous research. If, on the one hand, older adults generate more specific predictions during language processing, we expected to find large processing costs in response to high-surprisal sentences, in particular under dual-task conditions, when less cognitive resources are available to resolve the semantically unexpected event. If, on the other hand, older adults are less efficient at generating predictions, we expected only minimal processing costs for high-surprisal sentences, with only small differences between single and dual-task condition.

## Method

## Participants

Thirty-six older adults (mean age $=72$ yrs; 18 female) from the Saarbrücken community participated for compensation. The control group consisted of 34 younger adults (mean age $=23 ; 20$ female), mostly students at UdS. All participants were native speakers of German, reported no hearing problems and had normal or corrected-to-normal vision.

## Procedure

We investigated age-related differences in recovery from prediction errors while participants were engaged in a single and dual task. The dual task consisted of a language comprehension and continuous driving task. The single task consisted of the driving or language task only. Overall, the experiment consisted of six major blocks - two dual-task blocks for simultaneous language comprehension and driving, two single-task blocks for single driving, and two single task blocks for language comprehension.

Language Task. The language comprehension task consisted of a sentence verification task for 192 spoken sentences, half of them presented in a low-surprisal condition (low processing effort) and high-surprisal condition (high processing effort). Each sentence was constructed of two clauses, with the verb of the first clause providing a semantic context and the noun of the second clause either matching (low-surprisal condition) or violating this semantic context (high-surprisal condition), for example, Since Petra didn't have anything to wear (low surprisal) / drink (high surprisal) for the barbecue, she bought a dress in a nearby shop. Participants were instructed to carefully listen to the sentences, which were presented to them over speakers, and asked to judge whether each sentence was meaningful and correct by verbally answering "Yes" or "No", while the researcher recorded their responses. In order to minimize prosodic differences among items, all sentences were synthesized prior to the experiment using MARY TTS (Schröder, Charfuelan, Pammi, \& Turk, 2008) and pauses manipulated so that the duration of the disambiguating word (dress) was always identical. To avoid stereotyped responses, we also presented 72 filler items in a low- and high-surprisal condition, involving syntactic violations. All items were randomized using a Latin Square randomization, with surprisal (highlow) as blocking factor, to ensure that each participant encountered each experimental item in only one of its experimental conditions.

Driving Task. We used the Continuous Tracking and Reaction Task (ConTRe Task; Mahr, Feld, Moniri \& Math, 2012), a highly controlled driving task which measures rapid changes in steering deviation from a target. As such, the ConTRe task allows for continuous and very finegrained measurement of online changes in task performance over time (e.g., Becic, Dell, Bock, et al., 2010; Demberg, Sayeed, Mahr, \& Müller, 2013). Participants were seated in front of a steering wheel and saw a 3D road on a screen, with two vertical color bars moving laterally across the screen at a continuous speed. Participants were instructed that they could control one of the bars (the blue one) by turning the steering wheel whereas the other bar (the yellow one) was controlled by the computer. Their task was to continuously track the yellow bar so as to keep the distance between the two bars minimal at all times. Participants' driving performance was assessed by measuring their
steering deviation (indicated in meters) when processing low- and high-surprisal sentences.

## Results

We constructed separate linear mixed effects models for response accuracy and steering deviation, as implemented in the lme4 library (Bates \& Sarkar, 2007) in R (R Development Core Team, 2013). Fixed effects for response accuracy were sentence type (low surprisal vs. high surprisal), task condition (single vs. dual), and age group (younger vs. older). Fixed effects for steering deviation were sentence type and age group. Since raw steering deviation was coded in positive and negative values, indicating left- and right-sided deviations, we squared its values to obtain a final measure. For the LMER model for steering deviation, $p$-values were approximated from the model coefficients using the normal distribution (see Barr, Levy, Scheepers, \& Tily 2013). Categorical predictors were sum coded. All models contained participants and items as crossed random effects, and random slope adjustments for subjects and items. In the event that a model failed to converge, we simplified the random slope structure progressively until convergence was achieved (for guidelines, see Barr et al., 2013). Higher-order interactions involving the factor age group were followed up with planned model splits between younger and older adults.

## Response accuracy

The model for response accuracy showed a significant interaction between sentence type and age group, as well as a significant interaction between sentence type and task condition (see Table 2). To locate the source of these interactions, we computed two follow-up models in which items were split by age group. Thus, we computed one model for younger adults, and another model for older adults.

The model for the younger adults showed nothing but a significant main effect of sentence type ( $b=-0.62, S E=$ $\left.0.14, t=-4.29, p<.001^{* * *}\right)$, indicating that, regardless of task condition, younger adults responded less accurately to high-surprisal than low-surprisal sentences (see Figure 1, left panel). In contrast, the model for the older adults showed a significant interaction between sentence type and task condition ( $b=-0.86, S E=025, t=-3.4, p<.001 * * *$ ). An inspection of the plot for this interaction (see Figure 1, right panel) suggested that older adults responded equally accurately to high- than low-surprisal sentences in the single task condition, but showed a selective increase in response accuracy for high-surprisal sentences in the dual-task condition. These observations were confirmed by additional follow-up models, in which we split items by task condition: As predicted, only the model for the dual-task condition showed a significant main effect of sentence type ( $b=0.77$, $S E=0.18, t=4.28, p<.001 * * *)$, indicating an increase in response accuracy for high-surprisal sentences.

Thus, the data for response accuracy showed two main things of interest: First, younger adults responded less
accurately to high-surprisal sentences, regardless of task condition, indicating stable performance in this participant group even when working-memory load was high (i.e. the dual-task condition). Older adults, in contrast, responded more accurately to high-surprisal sentences in the dual-task condition, indicating that they selectively focused on resolving the semantic conflict in these items (cf. Becic et al., 2010), presumably by giving up driving. To support this view, we now turn to the driving performance in both age groups.


Figure 1: Response accuracy ( $\pm$ SEM) in younger and older adults, depending on task condition and sentence type.

Table 1: Effect sizes (b), standard errors (SE), t-values, and p-values for the logistic LMER model for response accuracy. Significance codes: $\left.\left.{ }^{* * *} .001\right|^{* *} .01\right|^{*} .05$

|  | $b$ | $S E$ | $t$ | $p$ |
| :--- | :---: | :---: | :---: | :---: |
| Sentence Type | -0.12 | 0.10 | -1.26 | $n s$ |
| Task Condition | -0.03 | 0.10 | -0.35 | $n s$ |
| Age Group | 0.34 | 0.15 | 2.24 | $*$ |
| SentType:Task | -0.58 | 0.19 | -3.03 | $* *$ |
| SentType:Group | -0.98 | 0.19 | -5.11 | $* * *$ |
| Task:Group | -0.24 | 0.19 | -1.27 | $n s$ |
| SentType:Task:Group | 0.59 | 0.38 | 1.53 | $n s$ |
| Random Effects | Variance |  |  |  |
| Subject | 0.23 |  |  |  |

## Steering deviation

The model for squared steering deviations showed a significant interaction between sentence type and age group (see Table 2). The plot of this interaction (see Figure 2) suggested that younger adults showed stable driving
performance regardless of whether sentences were highly surprising or not, whereas older adults demonstrated higher steering deviations when high-surprisal sentences were presented. To confirm these observations, we again computed follow-up models in which we split items by age group. As expected, only the model for older adults showed a main effect of sentence type $(b=0.24, S E=0.06, t=4.05$, $p<.001^{* * *}$; younger adults: $b=-0.02, S E=0.03, t=-0.70$, $p>.05)$.

Thus, the analysis of the driving data supported our hypothesis based on the response data. First, younger adults showed constant steering deviations regardless of semantic violations, suggesting that even under conditions of high linguistic load, they maintained high driving acuity. Older adults, in contrast, demonstrated greater steering deviations in response to high-surprisal sentences, suggesting increased effort to recover from semantic violations.

In sum, whereas younger adults maintained a stable pattern of performance even under conditions of high linguistic load, older adults devoted all attentional resources to resolving semantic violations, while neglecting the driving task.


Figure 2: Steering deviations in the dual task ( $\pm$ SEM) for younger and older adults, depending on sentence type.

Table 2: Effect sizes (b), standard errors (SE), t-values, and bootstrapped p-values for the logistic LMER model for steering deviation.

|  | $b$ | $S E$ | $t$ | $p$ |
| :--- | :---: | :---: | :---: | :---: |
| Sentence Type | 0.11 | 0.03 | 3.53 | $* * *$ |
| Age Group | 0.31 | 0.17 | 1.80 | $n s$ |


| SentType:Group | 0.26 | 0.06 | 4.18 |
| :--- | :---: | :---: | :---: |
| *** |  |  |  |
| Random Effects |  | Variance |  |
| Subject |  | 0.50 |  |
| SentType $\mid$ Subject | 0.06 |  |  |
| Item | 0.01 |  |  |
| Age Group \| Item |  | 0.05 |  |

## Discussion

Prior research has yielded conflicting findings with respect to predictive processing in aging. Some studies have shown that older adults are impaired at using context to generate predictions about upcoming content during language comprehension. Other studies have indicated that older adults form strong and semantically specific predictions during language processing, resulting in effortful recovery when such predictions are violated.

In this study, we investigated age-related differences in recovery from prediction error under conditions of increased working-memory load. By using a secondary driving task, we limited working-memory resources participants could devote to resolving prediction errors. To manipulate prediction error, we presented sentences in a high- and lowsurprisal condition. In high-surprisal sentences, participants were expected to experience integration difficulties when encountering unpredicted semantic content. Low-surprisal sentences, in contrast, were thought to induce only minimal processing effort.

Two key findings emerged. First, even though younger adults were sensitive to violations of semantic predictions overall (indicated by lower response accuracy for highsurprisal sentences), they maintained a stable behavioral pattern in both response accuracy and driving performance. Thus, younger adults were able to resolve the semantic violation in high-surprisal sentences without experiencing trade-off effects between primary and secondary task. This suggests high working-memory capacity in this participant group to split attention even under conditions of maximal linguistic load. Second, we found that older adults allocated all processing resources towards resolving the unexpected sentence continuation in high-surprisal sentences. This increased their response accuracy in the sentence verification task, but it came at the expense of driving accuracy: When high-surprisal sentences were presented, older adults demonstrated a strong increase in steering deviation.

Thus, our results are more in line with studies suggesting that older adults form strong predictions during language processing, and that violations of these predictions induce maximal processing effort to resolve the prediction error. Unlike younger adults, however, older adults may not have sufficient working-memory capacity to integrate semantically unexpected material into an unfolding sentence context and to additionally perform a secondary task without a substantial drop in task performance.

A second interpretation of our results is that older adults are inable to successfully divide attention between two cognitively demanding tasks. Thus, they might globally
shift attentional resources towards one cognitive goal when multiple tasks have to be performed at the same time. This interpretation is in line with prior research suggesting that older adults can relevel their task priorities in a case-by-case manner that follows principles of selective optimization, by taking into account the subjective difficulty of each task and choosing the one which is most likely to garner success (see Li, Baltes, Staudinger, \& Lindenberger, 1999; Miles \& Stine-Morrow, 2004; Stine-Morrow, Miller \& Hertzog, 2006). Here, older adults might have adopted a strategy of selective performance optimization, by neglecting the high demands in the bar-tracking task and focusing exclusively on the sentence verification task. Overall, the language task may have seemed more likely to yield success, given older adults' increased verbal knowledge and linguistic capacity (Glisky, 2007).

Finally, a somewhat open question is to what extent our data have real-life implications on older adults' car driving security. On the one hand, our results are supported by prior studies using simulated but also naturalistic driving scenarios, suggesting that driving ability suffers most under conditions of high working-memory load (Cantin, Lavallière, Simoneau, et al., 2009; Strayer, Cooper, Turrill, et al., 2013), and that older adults are more likely to adapt to such situations by selectively focusing their attention on one task and disregarding the other (Becic et al., 2010). In addition, there is evidence from research on car driving safety (Strayer et al., 2013) indicating that behavioral results obtained in simulated driving environments are largely identical to real-life driving.

On the other hand, car driving involves a range of cognitive-behavioral demands the bar tracking task in the present study was lacking, for example reactions to road signs or traffic lights, overtaking maneuvers and lane changes, or braking for other cars. In fact, the contrast between the present results and those obtained by an earlier study using a similar dual-task paradigm (Becic et al., 2010), shows that differences in design might impact the results to a large extent. According to the results of Becic et al. (2010), older adults showed a reversed pattern of task prioritizing than observed here, by focusing exclusively on the driving task and neglecting the language task. However, the experimental set-up in Becic et al. (2010) was more naturalistic than the one in the present study, since that study used an actual car dummy and a wrap-around projection screen that displayed realistic images of road situations and naturalistic driving scenarios. This set-up might have induced a more realistic feeling of car driving, where accidents can actually be fatal. The older adults in that study might have employed a task solving strategy that followed the rule of safety-first, by focusing their attention on the task which seemed most dangerous to them. In contrast, poor performance in the bar tracking task in the present study had no such real-life implications, possibly rendering this task somewhat negligible to older adults.

In sum, our data support studies arguing that recovery from prediction errors is more effortful for older adults, and
that older adults allocate attentional resources differently from younger adults when task demands are high, by prioritizing one cognitive goal over others. We discussed two possible causes for these age-related differences, i.e. older adults' increased rate of forming semantic predictions based on context, and/or impaired working-memory resources normally associated with aging. Future work in our lab will further investigate these possibilities by also exploring the pupillary response as a measure of cognitive load, and by taking into account individual differences in executive functions.

## References

Bates, D.M. \& Sarkar, D. (2007). Ime4: Linear mixedeffects models using S 4 classes, R package version 0.99875-6.

Becic, E., Dell, G. S., Bock, K., Garnsey, S. M., Kubose, T., \& Kramer, A. F. (2010). Driving impairs talking. Psychonomic Bulletin \& Review, 17, 15-21.
Borges, M., \& Coco, M. I. (2015). Access and Use of Contextual Expectations in Visual Search during Aging. In G. Airenti, B. Bara, \& G. Sandini (Eds.). Proceedings of the EuroAsianPacic Joint Conference in Cognitive Science.
Cantin, V., Lavallière, M., Simoneau, M., \& Teasdale, N. (2009). Mental workload when driving in a simulator: Effects of age and driving complexity. Accident Analysis and Prevention, 41, 763-771.
DeDe, G. (2014). Sentence comprehension in older adults: evidence for risky processing strategies. Experimental Aging Research, 40, 436-454.
DeLong, K. A., Groppe, D. M., Urbach, T. P., \& Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. Brain and Language, 121, 226-239.
DeLong, K. A., Troyer, M., \& Kutas, M. (2014). Pre- processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. Language and Linguistics Compass, 8, 631-645.
Demberg, V., Sayeed, A., Mahr, A., \& Müller, C. (2013). Measuring linguistically-induced cognitive load during driving using the ConTRe task. Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 176-183). ACM.
Federmeier, K. D., \& Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. Journal of Memory and Language, 41, 469495.

Federmeier, K. D., \& Kutas, M. (2005). Aging in context: Age-related changes in context use during language comprehension. Psychophysiology, 42, 133-141.
Federmeier, K. D., Kutas, M., \& Schul, R. (2010). Agerelated and individual differences in the use of prediction during language comprehension. Brain and Language, 115, 149-161.

Glisky, E. L. (2007). Changes in cognitive function in human aging. In D. R. Riddle (Ed.), Brain Aging: Models, Methods, and Mechanisms. New York: CRC Press.
Huettig, F., \& Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. Language, Cognition \& Neuroscience, 31, 80-93.
Kuperberg, G. R., \& Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? Language, Cognition and Neuroscience, 31, 32-59.
Kutas, M., DeLong, K. A., \& Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), Predictions in the brain: Using our past to generate a future. New York: Oxford University Press.
Li, K. Z., Lindenberger, U., Freund, A. M., \& Baltes, P. B. (2001). Walking while memorizing: Age-related differences in compensatory behavior. Psychological Science, 12, 230-237.
Mahr, A., Feld, M., Moniri, M. M., \& Math, R. (2012). The contre (continuous tracking and reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 88-91). ACM.
Miles, J. R., \& Stine-Morrow, E. A. (2004). Adult age differences in self-regulated learning from reading sentences. Psychology and Aging, 19, 626.
R Development Core Team. (2013). R: A language and environment for statistical computing. [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.
Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., \& Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. Psychology and Aging, 21, 448-465.
Schröder, M., Charfuelan, M., Pammi, S., \& Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In Proceedings of the Blizzard Challenge

Stine-Morrow, E. A., Miller, L. M. S., \& Hertzog, C. (2006). Aging and self-regulated language processing. Psychological Bulletin, 132, 582.
Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J., Medeiros-Ward, N., \& Biondi, F. (2013). Measuring cognitive distraction in the automobile.

# Asymmetric detection of changes in volatility: Implications for risk perception 

Nigel Harvey (n.harvey@ucl.ac.uk), Matt Twyman (m.twyman@ucl.ac.uk) and Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk)

Experimental Psychology, University College London, Gower Street, London, WC1E 6BT, UK


#### Abstract

Variance of the outcomes associated with an option often provides a measure of the riskiness of that option. Hence, it is important for organisms are able to detect any sudden changes in outcome variance. In Experiment 1, we presented people with graphs of share price time series or water level time series. In half the graphs, variance (financial or flooding risk) changed at some point. People were better at detecting increases than decreases in risk - maybe because it is more important to detect increases in danger than decreases in it. However, in Experiment 2, people were still better at detecting increases than decreases in variance even when those changes did not reflect altered levels of risk. Our findings may reflect the fact that the actual change in variance exceeds the change needed to identify a regime change in variance by a larger amount for upward than for downward changes.


Keywords: volatility; variance; risk; change detection; judgment

## Introduction

In many domains, variance of outcomes associated with an option is taken as a measure of level of risk of that option. For example, in modern finance theory, level of risk associated with an asset is defined as the standard deviation of the returns on that asset (Jorion, 2006). Similarly, as variability in water levels increases, so does the risk of flooding or drought (Crowell, Coulton, Johnson, Westcott, Bellomo, Edelman, and Hirsh, 2010). Finally, in foraging theory, the risk associated with different food sources is defined in terms of the variance of the energy gains that an animal can derive from those sources (Kacelnik and Bateson, 1996). In all these cases, higher variance in the data is treated as a signal that risk levels are higher.
Most work in these and other domains has been based on the assumption that the riskiness of different options remains constant over time. For example, Diacon and Haseldine (2007), Duxbury and Summers (2004, 2017), Sobolev and Harvey (2016), and Weber, Siebenmorgen and Weber (2005) have used various methods to examine the relation between volatility of financial indicators (e.g., returns) and financial risk perception. However, level of risk can change: variance of outcomes may increase or decrease, often quite suddenly. As far as we are aware, there have been no studies of people's ability to perceive a change in volatility and, hence, to detect onset of a new level of risk.
Here we ask how easily people are able to detect such a change when they are given a graphical record of the
outcomes that have occurred. More specifically, we examine how well people are able to detect a structural break in the variance of a time series and study whether the level of their ability is influenced by whether that variance is framed as representing level of risk.

We varied task frame. In Experiment 1, any structural break in the series signified an increase or decrease in the level of risk over time. Changes in financial trading risk and water flooding risk were of this type. In Experiment 2, any structural break in the series did not represent any change or difference in risk level. Instead, participants needed to detect it because it represented an opportunity rather than a risk. These experiments were used to address two questions.

First, is there any asymmetry in ability to detect increases and decreases in volatility? Second, is any such asymmetry limited to tasks in which changes in volatility should be interpreted as temporal changes in level of risk? It can be argued that it is more important to detect an increase in risk so that protective measures can be adopted. Removing those protective measures when there is a decrease in risk is likely to be less critical.

## Experiment 1

In this first experiment, participants performed the task within a temporal risk frame. They were presented with one of two scenarios: a finance scenario and a flooding scenario.

## Method

Participants One hundred and sixty-five students acted as participants: 59 were assigned to the financial risk scenario and 106 were assigned to the flooding risk scenario.

Stimulus materials Each graphically presented series comprised 50 data points generated uniquely for each participant. They were drawn from a Gaussian distribution with a mean of 500 and a standard deviation of either 5.00 (low volatility) or 15.0 (high volatility). Of the 60 graphs seen by each participant, 15 were of low volatility throughout, 15 were of high volatility throughout, 15 contained a change from low volatility to high volatility, and 15 contained a change from high volatility to low volatility. The 60 graphs were presented in random order. When there was a change in volatility, it occurred between points 11 and 40 inclusive and with equal likelihood. One third of the graphs of each of the four types contained no trend, one third contained a shallow upward trend, and one
third contained a shallow downward one. When there was a trend, the series still started at 500 but was then incremented or decremented by 0.1 on each successive point. Labelling of graphs depended on the task frame.

Procedure In the financial risk scenario, the vertical axis was labelled as 'price' and the horizontal axis as 'hours' (Figure 1). Participants were told that the series represented a record of recent stock prices and told that increased volatility represented increased trading risk. They needed to detect whether a change in risk had occurred because their trading strategy would need to change if it had done.

Figure 1: Example graph from the finance scenario in Experiment 1 showing prices that change every hour for a period of 50 hours and volatility shifting from high to low.


In the flooding risk scenario, the vertical axis was labelled as water depth and the horizontal axis as 'hours'. Participants were told that each graph represented a record of water levels in various locations and that increased volatility represented increased risk of flooding. They needed to detect whether a change in flood risk had occurred in order to implement flood control measures if it had increased or to stand them down if it had decreased.
For each graph, participants first gave a yes/no response to signal whether they had detected a change in the volatility in it. They then estimated the likelihood that their response was correct on a $50-100 \%$ scale.

## Results

Here we report analyses of participants' detection responses using signal detection theory (Macmillan and Creelman, 1991). We extracted measures of sensitivity ( $\mathrm{d}^{\prime}$ ) and response criterion ( $\beta$ ) for a) trials starting with low volatility on the left of the graph that either stayed low or that changed to high volatility and b) trials starting with high volatility on the left that either stayed high or that changed to low volatility. Data were analysed in this way so that we could use the signal detection measures to compare detection of change when the series started with low volatility to that when it started with high volatility. To obtain $\mathrm{d}^{\prime}$ and $\beta$, the z -transformations of the hit rate $(\mathrm{z}(\mathrm{H})$ ) and false alarm rate $(\mathrm{z}(\mathrm{F}))$ were first obtained. Then

$$
\begin{aligned}
& \mathrm{d}^{\prime}=\mathrm{z}(\mathrm{H})-\mathrm{z}(\mathrm{~F}) \\
& \beta=\exp \left(\left(\mathrm{z}(\mathrm{~F})^{2}-\mathrm{z}(\mathrm{H})^{2}\right) / 2\right)
\end{aligned}
$$

The sensitivity measure d' reflects how discriminable signal (change) trials are from noise (no change) trials, with higher values indicating better detection performance. The response criterion measure $\beta$ reflects the relative strength the evidence has to reach in order for the organism to respond that the trial was a change trial, with a value of 1 indicating no response bias, while values below 1 indicating a bias towards responding 'change' (i.e., the evidence for 'no-change' has to be stronger than the evidence for 'change').

As we are interested only in the effect of increasing as compared to decreasing volatility, we collapse the data over the presence and types of trend. Also, note that the signal detection measures are based on both signal (change) and noise (no change) trials, and hence we cannot compare sensitivity and response bias between change and no-change trials.

Table 1: Mean values of sensitivity ( $\mathrm{d}^{\prime}$ ) and response criterion ( $\beta$ ) in the two types of scenario for detection of changes in volatility in graphs that started with low volatility and in those that started with high volatility.

|  | Sensitivity (d') |  | Response criterion $(\beta)$ |  |
| :--- | :--- | :--- | :--- | :--- |
|  | Low | High | Low | High |
|  | Starting <br> Volatility | Starting <br> Volatility | Starting <br> Volatility | Starting |
| Volatility |  |  |  |  |
| Financial <br> risk <br> scenario <br> $(\mathrm{n}=59)$ | .95 | .26 | .22 | .19 |
| Flooding <br> risk | .79 |  |  |  |
| scenario <br> $(\mathrm{n}=106)$ |  |  |  |  |

Mean values of $\mathrm{d}^{\prime}$ and $\beta$ are shown in Table 1. A two-way analysis of variance on $\mathrm{d}^{\prime}$ using starting volatility as a within-participant variable and temporal frame as a between-participant variable revealed a strong main effect of starting volatility $\left(\mathrm{F}(1,163)=43.82 ; \mathrm{p}<.001 ; \eta^{2}=.21\right)$ and some evidence of an interaction between this variable and frame type $\left(\mathrm{F}(1,163)=4.57 ; \mathrm{p}=.034 ; \eta^{2}=.03\right.$ ).

An ANOVA using the same variables on $\beta$ failed to reveal any significant effects.

## Discussion

The experiment showed that people find it easier to detect increases in volatility than decreases in volatility. Given that increases in volatility in the task scenarios corresponded to increases in risk, this result can be interpreted as showing that people are better at detecting increases than decreases in risk. This corresponds to what would be expected from a functional perspective: it is more important to be sensitive to increases in risk (so that protective measures can be
implemented) than to decreases in risk (as removal of protective measures is less urgent). Differences in the size of the effect in the two scenarios may be related to beliefs about the nature of the risks and the ease of managing them in the two cases.
Before committing to this risk-based interpretation of the effects, it is important to ascertain whether they appear when the same graphs are presented within a scenario that does not involve risk.

## Experiment 2

In this experiment, participants were presented with a version of the task in which risk assessment was not involved. Results were then compared to those obtained in the previous experiment.

## Method

Participants A total of 80 new participants drawn from the same pool as before performed a risk-free version of the task.

Procedure Participants were told that the data points represented the contours of a mountain range. The vertical axis represented height in meters and the horizontal one degrees of visual angle. Mountains could be formed of soft rock that had eroded (low variance) or harder rock that had not (high variance). They were told that they needed to detect differences in the contours of the mountains because mineral deposits tended to occur at the interface of hard and soft rocks. Identifying such interfaces would trigger groundbased surveys to confirm the presence of mining opportunities. Thus, a left/right difference in variance was associated with identification of an opportunity rather than a risk.

In all other respects, the experiment was the same as Experiment 1.

## Results

In the same way as before, the $\mathrm{d}^{\prime}$ and $\beta$ values were extracted from the data (Table 2). Then an ANOVA was used to compare the values obtained from the temporal risk scenarios of Experiment 1 with those obtained from the risk-free scenario in the current experiment. Starting volatility (low versus high volatility on the left side of the graph) was a within-participants variable and task frame (risk-free versus temporal risk scenarios) was a betweenparticipants variable.
Again, there was a strong main effect of starting volatility $\left(\mathrm{F}(1,243)=30.00 ; \mathrm{p}<.001 ; \eta^{2}=.11\right)$. However, in this case, though there was an effect of frame type $(F(1,243)=$ $10.34 ; p=.001 ; \eta^{2}=.04$ ), there was no interaction between frame type and starting volatility. Thus, while people were better at detecting differences in volatility in the risk-free scenario, they were better in both types of scenario at detecting changes in volatility from low to high (assuming left-to-right scanning in the risk-free scenario) than at detecting volatility changes from high to low.

As before, an ANOVA using the same variables on $\beta$ failed to reveal any significant effects.

Table 2: Mean values of sensitivity ( $\mathrm{d}^{\prime}$ ) and response criterion ( $\beta$ ) in the two types of scenario for detection of changes in volatility in graphs that started with low volatility and in those that started with high volatility.

|  | ${ }^{2}$ |  | Sensitivity $\left(\mathrm{d}^{\prime}\right)$ |  |
| :--- | :--- | :--- | :--- | :--- |
| Response criterion $(\beta)$ |  |  |  |  |
|  | Low | High | Low | High |
|  | Starting | Starting | Starting | Starting |
|  | Volatility | Volatility | Volatility | Volatility |
| Temporal | .85 | .37 | .15 | .21 |
| risk |  |  |  |  |
| scenario <br> $(\mathrm{n}=165)$ |  |  |  |  |
| Risk-free <br> scenario <br> $(\mathrm{n}=80)$ | 1.01 | .75 | .17 | .24 |

## Discussion

We obtained the same effect reported in Experiment 1 when participants performed the task within a risk-free scenario. Assuming left-to-right attentional scanning of the graphs (Bergen and Lau, 2012; Eviater, 1995; Maas and Russo, 2003), we can say that they were more sensitive to an increase in volatility than to a decrease in volatility. Furthermore, this was true whether or not greater volatility represented greater risk. The asymmetry uncovered in Experiment 1 is of a more general nature than we originally assumed. However, its implications for detection of changes in levels of risk remain.

There was also a main effect of scenario type on $d^{\prime}$ : sensitivity was higher in the risk-free scenario. Focusing on opportunities rather than risks appears to have made the task simpler for participants.

## General discussion

The experiments show that people find it easier to detect an increase than a decrease in the variance of a graphically presented time series. Though changes in risk are realized as changes in variance in many domains, Experiment 2 indicated that increases in variance are easier to detect than decreases in variance even when changes in variance do not correspond to changes in risk level. Here we will outline two possible explanations for our findings: an explanation in terms of the processes needed to detect upward and downward changes in variance and a functional explanation based on the relative importance of upward and downward changes in variance.

## A process-based account

It is possible that our findings arose because increases in variance are statistically easier to detect than decreases in
variance. For example, we could ask whether it is statistically easier to detect the presence of a data point outside a given distribution (an outlier) than to detect the absence of a data point expected within that distribution. Conceivably, more data might be needed to perform the latter detection reliably.

In fact, to detect an increase in variance, it is not sufficient to detect a single anomaly: in normal distributions, we expect one in 22 data points to be more than two standard deviations away from the mean. To detect a change in variance, the presence of unexpected data points outside a reference distribution or the absence of expected data points within that reference distribution must be persistent. In other words, there must be evidence of a regime shift in the variance of the distribution.
There are many different approaches to detecting regime shifts in the mean of time series but relatively few have been developed for detecting shifts in the variance of series. Downton and Katz (1993) developed a non-parametric bootstrap technique to compute confidence intervals for discontinuities in variance. However, their approach requires the series containing the putative regime shift in variance to be compared to a separate reference series known to be characterized by homogeneous variance. We presented our participants with series in which variance did not change but we did not inform them of this constancy for particular series. Thus they had no series that they could treat as a reference series in the manner that Downton and Katz (1993) require.

Rodionov (2004) developed a sequential algorithm for early detection of regime shifts in the mean of series. The advantage of his approach is that it does not require large amounts of data to be accumulated and can automatically detect regime shifts in real time. Later, Rodionov (2005) extended his approach so that it could be used to detect regime changes in variance in short series in real time. These features of his approach render it a suitable one for modeling detection of variance change in our experiments.

The first step is to identify the regime length ( $l$ ). In our task, this value would initially be set to 10 because participants knew there was no shift in the first 10 data points. The next step is to use an F-test to determine the critical variance ratio ( $\mathrm{F}_{\text {crit }}$ ) of two successive regimes that would be statistically significant. For an $l$ value of 10 and a p -value of 0.05 (one-tailed), this ratio is 4 . The variance of the initial $l$ values of the series is then used to estimate the variance of the current regime ( $\mathrm{V}_{\text {cur }}$ ). For the new regime to be statistically different from the current regime, its variance ( $\mathrm{V}_{\text {new }}$ ) should be equal to or greater than the critical variance ( $\mathrm{V}_{\text {crit }}$ ) if the variance is increasing or equal to or less than the critical variance $\left(\mathrm{V}_{\text {crit }}\right)$ if the variance is decreasing, where

$$
\begin{aligned}
& \mathrm{V}_{\text {crit }}=\mathrm{V}_{\text {cur }} \cdot \mathrm{F}_{\text {crit }} \\
& \mathrm{V}_{\text {crit } \downarrow}=\mathrm{V}_{\text {cur }} / \mathrm{F}_{\text {crit } \downarrow}
\end{aligned}
$$

The variance, $\mathrm{V}_{\text {cur }}$, is the sum of squares of $\mathrm{z}_{\mathrm{i}}$, where i spans from the first point of the current regime to $i=t_{\text {cur }}-1$. If, at time $\mathrm{t}_{\text {cur }}$, the current value $\mathrm{z}_{\text {cur }}$ satisfies either $\mathrm{z}^{2}$ cur $>$ $\mathrm{V}_{\text {crit } \uparrow}$ or $\mathrm{z}_{\text {cur }}^{2}<\mathrm{V}_{\text {crit } \downarrow}$, this time is marked as a potential point
where a regime shift in the variance has occurred. Subsequent values ( $\mathrm{z}_{\mathrm{cur}+1}, \mathrm{Z}_{\mathrm{cur}+2} \ldots$ ) are used to verify this hypothesis by using a Residual Sum of Squares Index (RSSI).

$$
R S S I=1 / l \sum_{i=t_{\text {cur }}}^{m}\left(z_{i}^{2}-V_{\text {crit }}\right),
$$

where $\mathrm{m}=\mathrm{t}_{\mathrm{cur}}, \mathrm{t}_{\mathrm{cur}}+1, \ldots, \mathrm{t}_{\mathrm{cur}}+l-1$.
If, at any time during the testing period from $\mathrm{t}_{\text {cur }}$ to $\mathrm{t}_{\mathrm{cur}}+l$ -1 , the index turns negative for the case where $\mathrm{V}_{\text {crit }}=\mathrm{V}_{\text {crit }}$ or positive for the case where $\mathrm{V}_{\text {crit }}=\mathrm{V}_{\text {crit }}$, the hypothesis of a regime shift in variance at time $t_{c u r}$ is rejected and $z_{\text {cur }}$ is included in the current regime. Otherwise, time $t_{\text {cur }}$ is taken as a break point at which a regime shift in variance occurred.

In essence, Rodionov's (2005) approach first detects an anomaly and then goes on to determine whether that anomaly persists over time. A regime shift in variance is identified only when it does. Because his approach is simple and requires little accumulated data, it is appropriate for the statistical detection of regime changes in variance in the type of task that our participants completed.

In our task, the value of the lower variance was 25 and, hence, $\mathrm{V}_{\text {crit }}=25 \times 4=100$. The value of the higher variance (225) exceeded this critical value by a large amount (125). The value of the higher variance was 225 and, hence, $\mathrm{V}_{\text {crit } \downarrow}=225 / 4=56.25$. The value of the lower variance (25) was less than this critical value by only a small amount (31.25). However, the relative difficulty of two comparative judgments does not depend on the size of the absolute difference between the stimuli.

According to Weber's Law, "The stimulus increase which is correctly discriminated in any specified proportion of attempts (except 0 and 100 per cent) is a constant fraction of the stimulus magnitude" (Thurstone, 1959, p. 61). In the case of upward changes in variance, the change in variance that participants had to detect (125) as a proportion of the critical variance (100) was 1.25 . In the case of downward changes in variance, the change in variance that participants had to detect (31.25) as a proportion of the critical variance (56.25) was 0.56 . Hence the task of deciding whether there was evidence of a new variance regime would have been more difficult when the variance decreased from the high to the low value than when it increased from the low to the high value.

In terms of Rodionov's (2005) approach, for each current value, $\mathrm{z}_{\mathrm{cur}}$, it would have been harder to determine whether $\mathrm{z}^{2}$ cur was less than $\mathrm{V}_{\text {crit }}$ than to determine whether it was greater than $\mathrm{V}_{\text {crit }}$. As a result, the initial assessment of whether a potential anomaly had occurred at $t_{\text {cur }}$ would have been harder for a downward than for an upward anomaly. Furthermore, using the RSSI to verify whether the potential anomaly should be confirmed would have been less effective for a downward than for an upward anomaly.

We have outlined this process-based account using the parameters of our experimental task but it could be applied to any task in which comparative judgments of variance are made. Of course, other process-based accounts are possible:
the strategy outlined by Rodionov (2005) is not the only statistical approach to detecting regime change in variance. Indeed, it is possible that no unitary process-based explanation would be appropriate to account for the asymmetry in our data. We may have evolved so that the characteristics of the processes that detect upward and downward changes in variance are different. It is to this possibility that we turn next.

## A functional explanation

A sudden increase in volatility can be regarded as a signal onset and a sudden decrease in volatility as a signal offset. Work in psychophysics indicates that people are better at detecting the onset of a signal than the offset of one (e.g., Ahumuda, Marken, and Sandusky, 1975). This phenomenon can be given a functional interpretation, albeit a more general one than that we proposed when discussing the results of Experiment 1. The onset of a signal is likely to be of greater importance to an organism than the offset of one. Signal onsets (e.g. the appearance of a predator) are more likely to require urgent and rapid action than signal offsets (e.g., the disappearance of a predator).

One objection to this account is that differences in signal importance should be expected to affect response bias ( $\beta$ ) rather than sensitivity ( $\mathrm{d}^{\prime}$ ). If a signal is more important, the response criterion should be shifted to the left to increase the proportion of hits. In other words, there should be no difference in $\mathrm{d}^{\prime}$ values for detecting signal onsets and offsets. Instead, responses should be more biased in favour of saying there is a change when signals start low but may change to high (potential signal onset) than when they start high but may change to low (potential signal offset).
The problem with this approach is that shifting the response criterion to the left will also serve to increase the proportion of false alarms. Responding to these false alarms is likely to be costly. For example, animals reacting to a non-existent predator may lose foraging time and flee into a more dangerous environment. These high costs would tend to force the response criterion rightwards and so counteract the benefit-driven increase in hit rate arising from moving it leftwards. According to this functional account, evolution resolved this dilemma over time by increasing sensitivity to signal onsets. Such a strategy would avoid the increased costs arising from the additional false alarms associated with a laxer response criterion while still assuring the benefits of a high hit rate.

## Implications

Although the phenomenon that we have identified is not specific to identification of changes in risk, it still has implications for risk perception. In finance, sudden changes in series variance occur (Hammoudeh and Li, 2008; Todea and Petrescu, 2012). Although attempts to predict these changes have been made using autoregressive conditional heteroskedasticity ( ARCH ) and generalized autoregressive conditional heteroskedasticity (GARCH) models
(Bollerslev, 1986; Engle, 1982), severe problems in forecasting them remain.

For Mandelbrot (1997), this was not surprising. He argued that bursts of high volatility are inherently unpredictable and emerge naturally as a consequence of the nonlinear processes responsible for generation of financial series. He claimed that these series do not meet the assumptions of modern financial theory (e.g., Markowitz, 1959; Sharpe, 1964; Black and Scholes, 1973) but are, instead, fractal. If he is correct, technical analysts and traders cannot possibly predict sudden volatility changes in financial series. Instead, all they can do is to be alert to the possibility that such changes will occur and then react to them appropriately as soon as possible.

Assuming that sudden volatility changes in financial series are not predictable, how would the asymmetry that we have identified here affect trading behavior? Increases in risk may lead investors to sell winning shares to lock in their profits but to keep losing ones in the hope that high volatility will provide an opportunity of selling them later at a higher price. Decreases in risk should lead to investors keeping their winning shares because nothing untoward will happen but to sell their losing shares because there is no chance of their bringing in a higher price later if they are retained. Easier detection of an increase than a decrease in volatility will lead responses to increases in risk to dominate responses to decreases in risk. In other words, the tendency to sell winning shares but to retain losing ones will dominate. This is the disposition effect (Shefrin and Statman, 1985). While we would not wish to claim that easier detection of increases than decreases in risk is the only driver of the effect, it may be contributory.

In our experiments, we presented time series graphically. We could explain our results by assuming a) that graphs were scanned left to right so that earlier data points were encountered before later ones, and b) that signal onsets are easier to detect than signal offsets. Both these assumptions are supported by existing evidence in the literature.

Consider now the case where the data points are encountered sequentially in real time. We would no longer need to make the first assumption: the earlier points would be encountered before later ones anyway. Hence, given that the second assumption holds, we would expect the asymmetry to be maintained. In other words, our findings could be expected to generalize to situations in which people experience data points successively over a period in real time.

For example, situations in which operators of some system receive readings in this way but assess volatility judgmentally rather than formally may produce a greater tendency to implement measures to protect against increased risk than to remove those measures once the period of increased risk has passed. Such situations could include those associated with natural hazards, such as evacuation decisions in the case of potential volcanic eruptions or hurricanes.

We would not wish to claim that asymmetric tendencies to respond to increases and decreases in risk in such cases
should be characterized as cognitive biases. In line with the functional approach discussed above, they may represent sensible ways of responding to changes in risk levels.

## Acknowledgements

This work was supported by Economic and Social Research Council (ESRC) Grant RES-062-23-2735.

## References

Ahumada, A. Jr, Marken, R. \& Sandusky, A. (1975). Time and frequency analyses of auditory signal detection. Journal of the Acoustical Society of America, 57, 385390.

Bergen, B. K. \& Lau, T. T. C. (2012). Writing direction affects how people map space on to time. Frontiers in Psychology, 3, Article 109
Black, F. \& Scholes, M. (1973). The pricing of options and corporate liabilities. Journal of Political Economy, 81, 637-654.
Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics. 31, 307-327.
Crowell, M., Coulton, K., Johnson, C., Westcott, J. Bellomo, D., Edelman, S. and Hirsh, E. (2010). An estimate of the US population living in 100-year coastal flood hazard areas. Journal of Coastal Research, 26, 201211.

Diacon, S. \& Hasseldine, J. (2007). Framing effects and risk perception: The effect of prior performance presentation format on investment fund choice. Journal of Economic Psychology, 28, 31-52.
Downton, M.W. \& Katz, R.W. (1993). A test for inhomogeneous variance in time-averaged temperature data. Journal of Climate, 6, 2448-2464.
Duxbury, D. \& Summers, B. (2004). Financial risk perception: Are individuals variance averse or loss averse? Economics Letters, 84, 21-28.
Duxbury, D. \& Summers, B. (2017). On perceptions of financial volatility in price sequences. The European Journal of Finance, Published online 6 March 2017.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation. Econometrica. 50, 987-1008.
Eviater, Z. (1995). Reading direction and attention: Effects of lateralized ignoring. Brain and Cognition, 29, 137-150.
Hammoudeh, S. \& Li, H. (2008). Sudden changes in volatility in emerging markets: the case of Gulf Arab stock markets. International Review of Financial Analysis, 17, 47-63.
Jorion, P. (2006). Value at risk: The new benchmark for measuring financial risk (3rd Ed.) New York: McGrawHill
Kacelnik, A. \& Bateson, M. (1996). Risky theories: The effect of variance on foraging decisions. American Zoologist, 36, 402-434.

Maas, A. \& Russo, A. (2003). Directional bias in the mental representation of spatial events: nature or culture? Psychological Science, 14, 296-301.
Macmillan, N.A. \& Creelman, C. D. (1991). Detection theory: A user's guide. Cambridge: Cambridge University Press.
Mandelbrot, B.B. (1997). Fractals and scaling in finance: Discontinuity, concentration, risk. New York: SpringerVerlag.
Markowitz, H.M. (1959). Portfolio selection: Efficient diversification of investments. New Haven, CT: Yale University Press.
Rodionov, S. E. (2004). A sequential algorithm for detecting climate change regime shifts. Geophysical Research Letters, 31, L09204.
Rodionov, S.N., (2005). A sequential method for detecting regime shifts in the mean and variance. In: Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability, V. Velikova and N. Chipev (Eds.), UNESCO-ROSTE/BAS Workshop on Regime Shifts, 1416 June 2005, Varna, Bulgaria, 68-72.
Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, 19, 425-442.
Shefrin, H. \& Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. The Journal of Finance, 40, 777-790.
Sobolev, D. \& Harvey, N. (2016). Assessing risk in graphically presented financial series. Risk Analysis, 36, 2216-2232.
Thurstone, L.L. (1959). The Measurement of Values. Chicago: The University of Chicago Press.
Todea, A \& Petrescu, D. (2012). Sudden changes in volatility - the case of five financial investment companies in Romania. Procedia Economics and Finance, 3, 40-48.
Weber, E.U., Siebenmorgen, N, Weber, M. (2005). Communicating asset risk: How name recognition and the format of historic volatility information affect risk perception and investment decisions. Risk Analysis, 25, 597-609.

# Compound effects of expectations and actual behaviors in human-agent interaction: Experimental investigation using the Ultimatum Game 

Yugo Hayashi (y-hayashi@acm.org)<br>College of Comprehensive Psychology, Ritsumeikan University<br>2-150 Iwakura-cho, Ibaraki, Osaka, 567-8570, Japan<br>Ryo Okada (lt0601hf@ed.ritsumei.ac.jp)<br>College of Letters, Ritsumeikan University<br>56-1 Kitamachi, Toji-in, Kita-ku, Kyoto, 603-8577, Japan


#### Abstract

This study investigated how the expectations of others (i.e., top-down processes) and actual perceived behavior (i.e., bottom-up processes) influence negotiations during humanagent interactions. Participants took part in several sessions of the ultimatum game; we investigated the bargaining strategies directed toward the computer agent. To investigate the influence of top-down and bottom-up processes on performance, we designed an experiment wherein (1) participants expected their partners were humans or agents, and (2) agents used different types of algorithmic behavior. Results revealed that irrational decisions, which are characteristic of human-human interactions, emerged when participants believed their opponents were human and when opponent behaviors were ambiguous. Further, we found participants adopted different bargaining strategies according to their expectations and the agent's specific algorithmic behavior. We discuss interplay of the two types of cognitive processing in human-agent interaction.


Keywords: human-agent interaction; top-down/bottom-up processes; social interaction; ultimatum game

## Introduction

Studies in human-computer interaction have revealed that how people engage with systems depends on how the agents are perceived (Nass, Moon, Fogg, Reeves, \& Dryer, 1995). The human user responds to social cues (Johnson, Veltri, \& Hornik, 2008) and to the apparent level of agency of the system (Blascovich et al., 2002). Studies have focused on how users adaptively interact based on their developing representation of the agent, which can be driven by the use of prior knowledge, such as using heuristics (top-down processing), and which can be modified based on the agent's actual behavior (bottom-up processing) (Hayashi \& Miwa, 2008). However, it is still unclear how the interdependence of these cognitive processes emerges, and it is not fully understood in which situations such interdependence occurs. To investigate these issues, we conducted a human-agent experimental study that involved negotiation in an ultimatum game.

## Two types of cognitive processing in human-agent interaction

Under what circumstances human-like traits such as agency are assigned to computers has been investigated in the fields of human computer interaction and interfaces (Kiesler, Waters, \& Sproull, 1996; McEneaney, 2013; Nass et al., 1995; Johnson et al., 2008; Blascovich et al., 2002). Theoretical studies of human computer interaction (e.g., Nass et al.
(1995)) have noted that people unintentionally respond to technology that exhibits social traits as if it were human, as a way to conserve cognitive resources and maximize response efficiency. HCI studies also suggest that how people perceive computers depends on the social cues that are designed into the system. For example, human facial features (Gong, 2008), embodied gestures (Buisine \& Martin, 2007), and language use (McLaren, DeLeeuw, \& Mayer, 2011) provide for a human-like agent that evokes social responses. However, there is controversy associated with this theory: such automatic responses have been suggested to be aberrant behaviors that result from situational inattention or inappropriate overgeneralization (McEneaney, 2013).
Recent studies in human-agent interaction (HAI) have pointed out the importance of top-down and bottom-up cognitive processing (Miwa \& Terai, 2006). Top-down processing is based on the socialized knowledge of others, i.e., interpersonal schemas or stereotypes (Fisk \& Taylor, 1991). Such processing is essential for developing representations of others in the initial stage of interaction, and can be used as supplemental information when representations are difficult to develop based on other's behaviors. However, the representation of others may change over time due to their ongoing behavior and the context in which the interaction occurs (Hayashi \& Miwa, 2008). Such behavior-based processes are examples of bottom-up processing.

It is important to note that in interpersonal communication between humans, people flexibly use both types of cognitive processing to economically process information when developing representations of others and deciding upon a response. However, few studies have investigated the relationship between the two types of processing in HAI, and it is unclear how such processing plays a role in interactions. Accordingly, in this study we used the Ultimatum Game (UG), a bargaining game that is commonly used in behavioral economics (Guth \& Tietz, 1990), to investigate how the combination of expectations and actual behavior influences cognitive processing during decision making.

## Influence of top-down and bottom-up processing in an ultimatum game

The ultimatum game is often used to investigate behaviors that are not self-regarding, such as choice inequity and reciprocity (Yamagishi et al., 2009). This game is played by two
players a proposer and a responder. Typically, one individual actively participates at any given time (i.e., it is a turntaking game).

First, the proposer receives a sum of money from the experimenter and then makes a proposal concerning how to share the money with the responder. The responder is given two alternatives, namely to either reject or accept the proposal. If the proposal is accepted by the responder, both players receive money according to the proposal, but if the responder rejects the proposal neither receives any money. As such, the self-regarding profit-motivated behavior is to accept any proposal.

Interestingly, respondents tend to reject proposals that are not distributed fairly, even when doing so results in a loss of profit for both players (Guth \& Tietz, 1990). In the current study, it is assumed that if the respondent (participant) perceives the proposer (agent) as human, the former may react accordingly, such as by rejecting proposals and abandoning profit as in human-human studies. We controlled the expectations (i.e., top-down processing) of the participants and determined whether expectations of their partner, such as believing the partner is human or non-human, would produce irrational behavior.
H1: When given an unfair proposal, the rejection rate by the respondent will increase when he/she thinks the partner is human compared to a computer agent.

However, as mentioned previously, actual behavior during interactions is used to update the representation of others (i.e., bottom-up processing). To investigate this issue, we used a multi-period version of the ultimatum game (mUG) (Guth, 1995). Studies have revealed that over repeated trials, players learn to expect that the proposer will suggest a fair deal in some future trial; as such, proposal rejections tend to decrease. That is, the number of rejections decreases due to understanding the strategy of the opponent (Slembeck, 1999). Therefore, we hypothesized that if the agents (proposers) showed concessional bargaining behaviors, and participants could perceive such behavior, respondents would perform more rationally by reducing the frequency of rejections.

H 2 : The rejection rate will decrease when participants understand that the proposer will provide concessional proposals.

Assuming that top-down and bottom-up processing are interdependent, it can be further assumed that the effect of expectations will emerge only when others' behaviors can be explicitly interpreted. To investigate this issue, we produced agents with different algorithmic behaviors, which will be described in more detail in the following section.

## Method

## Participants and procedure

Seventy-six (male: 30, female: 46, Mage: 21.38 , SD: 1.03 ) Japanese university students majoring in psychology voluntarily participated in the task; 3 were subsequently excluded
from data analysis because they discovered that their partner was not human.
Participants collected in small groups in a computer room and were instructed how to play the mUG game. They were told that they would play the role of either the proposer or responder; however, all were actually assigned the role of responder and the computer agent played the role of the proposer.


Figure 1: Example screenshot the task.
After the brief introduction to the task, participants were told to start the program, which appeared to connect to a randomly chosen peer in the computer room. They were told that 1,000 Japanese yen (approximately 12 dollars) was provided to the proposer. On the left hand side of the screen, the participant was required to input his or her IP address, which was nominally for connection to the opponent. Below were simple instructions including what he or she would/would not receive based on his or her decision. On the right hand side, the current proposal was shown. Below were decision buttons and a send button to transmit the result to the proposer.

First, a screen appeared that prompted the participant to wait until the proposer finished entering the amount of the proposal. After a short delay, the screen changed to that shown in Figure 1. Then, the participant chose to either accept or reject the proposal.

A proposal and subsequent decision constituted one trial and a total of 15 trials were conducted in one set of this task; two sets of this task were conducted in total. After completing the task, the participant wrote down a description of how he or she felt about his or her partner.

## Experimental conditions

This study examined mUG performance changes due perception of the partner as human or non-human and the partner's actual behavior. We used a 2 (perceived partner: human vs agent) X 4 (actual behavior: random vs adaptive [simple, egocentric, exocentric]) experimental design. The perception of the partner was controlled by telling the participant that the partner was either human or a computer agent. The former was called the human condition and the latter the agent condition.

In each set of the task, the announcement that the partner was human or computer was announced and the order of such announcements was counterbalanced between the small groups. There were no differences in rejection rates according to the order.

To investigate the effect of agent behavior, we implemented agents that utilized (1) algorithmic behavior or (2) no such algorithmic behavior (random condition). To determine how participants change their interactive strategies based on perceived behavior and ongoing interactions, we implemented three different types of behavior for (1). We examined different algorithms, including those that were likely to be perceived as offering generous or fair proposals. If bottom-up processing predominated in this task, the participant would likely adopt the rational strategy of accepting all proposals from these types of agent.

## Behavior of agent

In this section we describe the parameters that defined the agent behaviors. Table 1 shows all possible responses that could be generated by the agent for each trial. In the first trial, the agent always selected response type 4 in all conditions. Then, in the next and subsequent trials, the probability of generating each different response type differed according to the condition.

Table 1: Types of response(proposals) by the agent

| response type | amount of proposal (yen) |  |
| :---: | :---: | :---: |
|  | proposer(agent) | responder(participant) |
| $r 1$ | 100 | 900 |
| $r 2$ | 200 | 800 |
| $r 3$ | 300 | 700 |
| $r 4$ | 500 | 500 |
| $r 5$ | 700 | 300 |
| $r 6$ | 800 | 200 |
| $r 7$ | 900 | 100 |

In the random condition, the agent selected fair/unfair proposals (response type 1-7) randomly, with equal probability. This allowed for the investigation of ambiguous behaviors (i.e., restricting bottom-up processing). In egocentric, exocentric, and adaptive conditions, the agent proposed concessional and generous responses based on the participant's decisions.

In the simple adaptive condition(hereinafter referred to as adaptive condition) the agent repeated the proposal if it was accepted, and otherwise proposed the completely opposite monetary strategy (i.e., fair versus unfair). This was based on the Pavlof strategy in social games, wherein the basic rules are "win-stay" and "lose-shift" (Nowak \& Sigmund, 1992). In Figure 2, SAME denotes repeating the same proposal as in the prior trial.

The egocentric and the exocentric conditions were based on the adaptive condition. In the egocentric condition, the agent responded such that the proposal was clearly biased toward the computer agent(see Figure 3). More specifically, the

```
[r1-r3]
"accept" - > %SAME%
"reject" - > r5-r7:33.33%
[r4]
"accept" OR "reject" - >r1-r3,r4-r7: 16.66%
[r5-r7]
"accept" - > %SAME%
"reject" - >r1-r3:33.33%
```

Figure 2: Algorithm schematics of the adaptive condition.
agent reacted economically, such as proposing r3 if the participant kept accepting this proposal. The agent behavior in the egocentric condition is shown below. The agent decided on the next proposal depending on whether the participant accepted or rejected the previous proposal. For example, in the first trial the agent always proposed r 4 (see Table 1). On trial 2, if the participant selected accept, then the agent generated the next proposal based on the following probabilities: r1 (10 \%), r2 (20 \%), r3 (70 \%).

```
[r1]
"accept" -> r1:10%,r2:20%,r3:70%
"re ject" - > r5: 10%,r6: 20%,r7:70%
[r2]
"accept" - >r2:30%,r3:70%
"reject" -> r5:10%,r6:20%,r7:70%
[r3]
"accept" - > r3: 100%
"reject" - > r5: 10%,r6: 20%,r7:70%
[r4]
"accept" - > r1:10%,r2:20%,r3:70%
"reject" - > r5:10%,r6:20%,r7:70%
[r5]
"accept" - > r5:10%,r6:20%,r7:70%
"reject" - >r1: 10%,r2:20%,r3:70%
[r6]
"accept" - > r6:30%,r7:70%
"reject" -> r1: 10%,r2:20%,r3:70%
[r7]
"accept" - > r7 : 100%
"reject" - >r1:10%,r2:20%,r3:70%
```

Figure 3: Algorithm schematics of the egocentric condition.

In the exocentric condition the agent responded such that it sought less profit than in the egocentric condition(see Figure 4). If the participant kept accepting the proposals, the agent gradually proposed r1 more frequently, and even unfair, agent-biased proposals were most often r5 (i.e., relatively modestly favoring the agent).

```
[r1]
"accept" - > r1 : 100%
"reject" - >r5:70%,r6:20%,r7: 10%
[r2]
"accept" - > r1:70%,r2:30%
"reject" - > r5:70%,r6:20%,r7: 10%
[r3]
"accept" - > r1:70%,r2:20%,3:10%
"re ject" -> r5:70%,r6:20%,r7: 10%
[r4]
"accept" - >r1:70%,r2:20%,r3:10%
"reject" - >r r: 70%,r6:20%,r7: 10%
[r5]
"accept" - > r5:100%
"reject"->r1:70%,r2:20%,r3:10%
[r6]
"accept" - > r5:70%,r6:30%
"reject" - >r1:70%,r2:20%,r3:10%
[r7]
"accept" - > r5:70%,r6:20%,r7:10%
"reject" - >r1:70%,r2:20%,r3: 10%
```

Figure 4: Algorithm schematics of the exocentic condition.

## Results

## Performance of participant: Rejection rate

The participants' percentage rejections are shown in Figure 5. The vertical axis represents the average percentage of proposals rejected during the 15 trials, the horizontal axis shows each behavioral condition, and the different bar shading denotes the different instructions.

A 2 instructions (human or agent) x 4 agent behaviors (random, adaptive, egocentric, or exocentric) mixed factorial ANOVA revealed a significant interaction between the two factors $(F(3,72)=4.535, p=.0057)$. Analysis of simple main effects indicated that in the random condition, proposals by an apparently human opponent were rejected more often than those of a computer opponent $(F(1,72)=18.144, p=$ .0001 ), whereas there were no differences for the adaptive, egocentric, and exocentric conditions $(F(1,72)=0.504, p=$ $.4800 ; F(1,72)=2.016, p=.1600 ; F(1,72)=0.165, p=$ .6862, respectively).
The simple main effect of instruction (human or agent) was also significant for each behavior condition $(F(3,72)$ $=9.543, p=.0001 ; F(3,72)=3.388, p=.0198)$. Multiple comparisons using Ryan's method for the human instruction and showed that rejections were higher for the random condition than the adaptive, egocentric, and exocentric conditions ( $p=.0001 ; p=.0001 ; p=.0076$, respectively). For the agent instruction, the random condition only differed from the egocentric condition ( $p=.0052$ ). Also, when they were instructed that their partners were agents, the egocentric con-
dition was associated with less rejections than the exocentric condition ( $p=.0092$ ).

To summarize, the effect of instruction was significant when the behavior of the agent did not have any intention (i.e., the agent engaged in non-adaptive behavior). This indicates that H1 is supported only when others' behaviors cannot be used to understand their strategy (i.e., bottom-up processing is not possible). In contrast, the effect of the behavior markedly influenced the participants' performance; therefore, H2 is supported. However, participants' performance changed contingent on how they perceived their partner. That is, instruction and behavior interacted.


Figure 5: Ratio of rejections.

## Behavior of agent: ratio of proposal types

To further understand how the agents adaptively changed their behavior due to the participants' decisions, we examined the actual proposals made by the agents. Figure 6 shows the distribution of proposals for each condition. We then conducted an ANOVA that included the three behavioral conditions that adaptively changed their behavior based on the participants' decisions.

For the human condition, we conducted a $7 \times 3$ mixed factorial ANOVA with the seven selected responses (r1, r2, r3, r4, r5, r6, or r7) and adaptive conditions (adaptive, egocentric, or exocentric) as independent factors. There was significant interaction between the two factors $(F(12,324)=22.147$, $p=.0001$ ). Since we wanted to investigate which response appeared most frequently within each condition we only conducted simple main effects analysis for each level of condition. Significant main effects were present for all conditions (adaptive: $F(6,324)=5.211, p=.0001$; egocentric: $F(6$, $324)=45.798, p=.0001$; exocentric: $F(6,324)=18.403, p$ $=.0001$ ).

Next, multiple comparisons using Ryan's method were conducted for the adaptive condition. Response types r1, r2, and r 3 were used more frequently than r 3 , r 4 , r 5 , and r 6 ( $p$ $=.0001$, for each comparison). For the egocentric condition, response r 3 was used more often than all other responses ( r 1 , $\mathrm{r} 2, \mathrm{r} 4, \mathrm{r} 5, \mathrm{r} 6$, and $\mathrm{r} 7 ; p=.0001$, for each comparison). For the exocentric condition, response r 1 was chosen more frequently
than $\mathrm{r} 2, \mathrm{r} 3, \mathrm{r} 4, \mathrm{r} 6$, and $\mathrm{r} 7(p=.0001$, for each comparison) and response r 5 was used more frequently than $\mathrm{r} 2, \mathrm{r} 3, \mathrm{r} 4, \mathrm{r} 6$, and r7 ( $p=.0001$, for each comparison).

For the agent condition, we conducted the same analysis and found a significant interaction between the two factors $(F(12,324)=27.581, p=.0001)$. Focusing on the same simple main effects, responses differed according to condition $(F(6,324)=4.541, p=.0001 ; F(6,324)=52.996, p=$ $.0001 ; F(6,324)=22.469, p=.0001)$. Multiple comparisons revealed exactly the same pairwise differences were significant as in the human condition ( $p=.0001$, in each case).

To summarize: (1) in the adaptive condition, r1, r 2 , and r3 were used most frequently; (2) in the egocentric condition, r3 was most commonly used; and (3) in the exocentric condition, r1 and r5 were the most frequent proposals. This shows that agents responded differently to the participants' decisions and that the agent frequently generated proposals that did not favor itself in the exocentric condition.


Figure 6: Ratio of generated proposal(top: human condition, bottom:agent condition).

## Discussion

## Influence of the expectation of the other

The rejection rate data revealed that when the opponent had no strategy (i.e., random condition) the effect of expectations played an important role (i.e., human condition vs. agent condition). This shows the influence of top-down and bottom-up processing and their interdependence, whereby participants used initial expectations to generate a representation of their
opponent when the opponent's behavior was not clearly interpretable.

However, why did participants reject the proposer's offer most frequently when the proposer was believed to be human? Past research on economic behaviors using the UG has provided various explanations as to why participants reject proposals, even when doing so is not rational (Guth \& Tietz, 1990). Fehr and Schmidt (1999) proposed "inequity aversion theory," which posited that people are sensitive to unfair proposals, regardless of who profits most. People aim to balance inequities by rejecting unequal proposals. Furthermore, Falk, Fehr, and Fischbacher (2003) suggested that following unfair proposals, rejections will rise due to the interpretation of how the proposal was decided upon. I.e., there is an attribution of intentionality or animosity by others. As such, participants may have attributed the same types of intentions to their opponent in this study. However, when they believed their opponent was non-human, such human-specific effects did not occur and rejections decreased.

## Influence of the types of adaptive behaviors

The agent's behavior strongly affected rejection rates, whereby participants tended to reject proposals less frequently when the opponent adopted consistent and adaptive strategies, compared to the inconsistent random condition. This tendency was most pronounced when the partner was believed to be human. This indicates that participants decided upon a strategy based on their understanding of the adaptive behavior (i.e., using bottom-up processing), but relied on initial expectations (i.e., using top-down processing) when the opponent's behavior was unpredictable.

Interestingly, participants tended to behave more rationally (i.e., accepting the proposals) when they expected to interact with an agent only when the agent used an egocentric strategy. This indicates that expectations of such egocentric agents may have suggested that the system was non-negotiable toward fairer proposals, and thus the best strategy was to accept their proposals.

Surprisingly, compared to the egocentric condition, participants behaved more irrationally in the exocentric condition by rejecting proposals that were beneficial to them, such as r1. Figure 6 shows that participants oscillated between r1 and r5 as a consequence of their pattern of rejection and acceptance of proposals. However, why did they reject proposal r1? This can be interpreted as rejection to reduce the dissonance (Festinger, 1957) associated with an unfair proposal, regardless of who profits. Further, this could be a result of adopting social norms, such as inequity aversion (Fehr \& Schmidt, 1999). Such a socially interactive approach may be the result of perceiving the agent as a social actor (Nass et al., 1995).

These findings cast new light on how decisions in humanagent interaction change based on the compound effects of who an actor believes his or her opponent is, and the actual behavioral strategy observed.

## Conclusions

This study investigated the influence of top-down (i.e., expectations of others) and bottom-up processing (i.e., the observation of human-like strategic behavior) on human-agent interaction. This aim was to determine the interdependence of such processing, and to investigate how these processes influence rational decision making in a mUG.

Based on evidence that people reject unfair proposals in human-human interactions, we hypothesized that believing one's partner is human will influence the rejection of other's proposals, if the other's intentions are difficult to interpret (i.e., bottom-up processing cannot be used). By conducting a virtual human-agent experiment, we controlled participants' expectations via agent behavior that followed simple algorithms. The results supported our hypothesis and show that people rely on expectations of the opponent's behavior when the latter's actual behavior is ambiguous. This highlights the interdependent relationship of top-down and bottom-up processing in human-agent interaction.

In addition to the effects of the two types of processing in the mUG, results suggest that people try to avoid inequity; that is, to reject unfair proposals even if they are profitable for themselves. Such a tendency was observed here, even when the participant believed their opponent was a computer agent. This indicates that people treat their counterparts as social actors, even when the goal of the interaction is selfregarding.
In summary, this study supports the interdependent influence of two types of cognitive process, and captures the emergence of irrational decision making in human-agent interaction.

## Acknowledgments

This work was supported by the Grant-in-Aid for Scientific Research (KAKENHI), No. 16KT0157.

## References

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., \& Bailenson, J. (2002). Immersive virtual environment technology as a methodological tool for social psychology. Psychological Inquiry, 13(2), 103-124.
Buisine, S., \& Martin, C. J. (2007). The effects of speech gesture cooperation in animated agents' behavior in multimedia presentations. Interacting with Computers, 19(4), 484-493.
Falk, A., Fehr, E., \& Fischbacher, U. (2003). On the nature of fair behavior. Economic Inquiry, 41(3), 20-26.
Fehr, E., \& Schmidt, K., M. (1999). A theory of fairness, competition, and cooperation. Oxford JournalsSocial Sciences Quarterly Journal of Economics, 114(3), 817-868.
Festinger, L. (1957). A theory of cognitive dissonance. Stanford University Press.
Fisk, T. S., \& Taylor, E. S. (1991). Social cognition. McGraw-Hill Education.

Gong, L. (2008). How social is social responses to computers? the function of the degree of anthropomorphism in computer representations. Computers in Human Behavior, 24,(4), 1494-1509.
Guth, W. (1995). On ultimatum bargaining experiments? a personal review. Journal of Economic Behavior and Organization, 27(3), 329-344.
Guth, W., \& Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. Journal of Economic Psychology, 11(3), 417-449.
Hayashi, Y., \& Miwa, K. (2008). Schema-based and evidence-based communication in human-human and human-agent interaction. In Proceedings of 6th international conference of cognitive science (p. 285-288).
Johnson, R. D., Veltri, N. F., \& Hornik, S. (2008). Attributions of responsibility toward computing technology: The role of interface social cues and user gender. International Journal of Human Computer Interaction, 24(6), 595-612.
Kiesler, S., Waters, K., \& Sproull, L. (1996). A prisonor's dilemma experiment on cooperation with people and human-like computers. Journal of Personality and Social Psychology, 70(1), 47-67.
McEneaney, E. J. (2013). Agency effects in human-computer interaction. International Journal of Human-Computer Interaction, 12(29), 798-813.
McLaren, M. B., DeLeeuw, E. K., \& Mayer, E. R. (2011). Polite web-based intelligent tutors: Can they improve learning in classrooms? Computers and Education, 56(3), 574-584.
Miwa, K., \& Terai, H. (2006). Analysis of human-human and human-computer agent interactions from the viewpoint of design of and attribution to a partner. In Proceedings of the 28th annual conference of the cognitive science society (p. 597-602).

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., \& Dryer, D. C. (1995). Can computer personalities be human personalities? International Journal of Human Computer Studies, 43(2), 223-239.
Nowak, A. M., \& Sigmund, K. (1992). Tit for tat in heterogenous populations. Nature, 355(6357), 250-253.
Slembeck, T. (1999). Reputations and fairness in bargaining: experimentalevidence from a repeated ultimatum game with fixed opponents. In University of st.gallen, technical report (p. 1-21).
Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., \& Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. In Proceedings of the national academy of sciences of the united states of america pnas (p. 11520-11523).

# The role of presentation order and orientation on information search and evaluations: An eye-tracking study 

Claire L. Heard (claire.heard@kcl.ac.uk)<br>Department of Psychology, King's College London, Guy's Campus, London, SE1 1UL<br>Tim Rakow (tim.rakow@kcl.ac.uk)<br>Department of Psychology, King's College London, Guy's Campus, London, SE1 1UL<br>Tom Foulsham (foulsham@essex.ac.uk)<br>Department of Psychology, University of Essex, Wivenhoe Park, Colchester, CO4 2SQ


#### Abstract

Previous research conducted by Bergus et al. (2002) identified that treatment evaluations are more negative when risks are presented last. Extending discussion of this order effect, the current studies investigate this effect in tabular style displays, manipulating both order and orientation; and using eye-tracking methodology, explores the effect of these variables on the information search process. Analysis from eye-tracking data revealed a tendency to read information sets sequentially (i.e. read all risk information before transitions to the other set), which is stronger for the vertical orientation where switching between information sets is less common. Further, while balanced search was observed when benefits presented first, when presented with the risks first, search becomes more riskheavy. Eye-tracking measures did not strongly predict treatment evaluations, although, when holding other variables constant, time proportion spent on benefits positively predicted treatment evaluations.


Key words: Eye-Tracking; Information Search; Order Effects; Information Design

## Introduction

Previous research conducted by Bergus et al. (2002) investigated the role of order on judgements and decisions about a treatment. In their study, they investigated the effect that presenting either the risks last (benefits then risks presentation) or benefits last (risks then benefits presentation) had on favourability ratings of a treatment and choice about whether they would consent to a treatment. For their low risk Aspirin scenario, they found that when the last set of information read was about the risks (i.e. those presented with benefits then risks), ratings of favourability decreased (from pre-task favourability) compared to those who learned about the benefits last (i.e. presented with risks then benefits). Those presented with the risks last were also less likely to consent to a treatment.

Further evidence of this type of reliance or influence of the last piece of information processed (potentially because of its prevalence in one's memory) can be seen in a similar study by Ubel et al. (2010; who presented breast cancer patients with information about the benefits and risks of tamoxifen) and in other decision making tasks involving sequential information processing where individuals play an active role in searching out the information for themselves (Rakow, Denes \& Newell, 2008; Ashby \& Rakow, 2014).

When presented with this type of information in a medical setting or when people search for health information online, this information can often be presented using tabular-style displays (where risk and benefit information is separated into clear columns or rows either using a lined table or bullet point display). One such example, where this type of tabular style information format has been used is on the UK National Health Services (NHS) Choices website, where a fact sheet for prostate cancer screening (PSA) testing, which uses such separated, and bullet pointed display of the risks and benefits, is used (NHS Choices, 2016).

With such tabular displays (or any display allowing people to see the information simultaneously about the risks and benefits), the order in which people read the information is to some extent open to the individual. Research on information and picture search however reveals that people tend to examine information in an order consistent with their reading system, with those in western cultures starting in the top left and showing a bias to the left side of space and horizontal saccades made more frequently than vertical saccades (Foulsham et al., 2013; Foulsham, Kingstone \& Underwood, 2008).

These differences in search lead to questions of how these differences may generalize to more specific differences in perceptions of health information. For example, how such a bias to reading the left side (or top) first may indicate that information read in this position is likely to be read first and thus make order important. Or, how, with horizontal saccades being more common than vertical saccades, a choice of orientation (whether blocks are presented side-byside (i.e. horizontal arrangement) or above-and-below (i.e. vertical arrangement) may impact on search differences across information blocks.

## Current Study

Focusing on these types of tabular displays where information is presented simultaneously, the current research investigates whether such effects replicate across four low-risk (non-invasive) medicinal scenarios and whether orientation interacts with order.

Further, to understand the effect that such design choices of order and orientation have on pre-decisional information search strategies when presented with this information, eye-
tracking methodology was used to investigate information acquisition (search) processes. In particular, measures of looking order, proportion of time spent on the benefits number of transitions between reading the benefits and risks were calculated from the eye-fixation data.

Such measures allow a range of questions to be answered about people's search:

1) Is search consistent with the manipulated order or do people switch between searching between information sets frequently when given the choice?
2) How is search behavior affected by changes to the order and orientation?
3) Does this search behavior map onto subsequent treatment evaluations?

Predictions: Three main predictions were made, based on the literature presented above:

- Based on Bergus et al (2002), a recency effect is predicted. Thus, people should be more influenced by the information presented last.
- Despite simultaneous presentation, people will read information according to the manipulated order (i.e. in the order consistent with reading patterns).
- From findings that horizontal saccades occur more frequently than vertical saccades (Foulsham et al., 2008), reduced switching in the vertical orientation is predicted, which may lead to a stronger order manipulation in this orientation.


## Methods

## Participants

One hundred and fifty two students (108 in Study 1; 44 in study 2) from the University of Essex participated for course credit or payment.

## Materials and Procedure

## Evaluation Task

In both experiments, participants were presented with four hypothetical situations:

- Aspirin therapy treatment for mild carotid stenosis
- Statins for high cholesterol
- ACE (angiotensin converting enzyme) inhibitors for high blood pressure
- Anticoagulant medicine for deep vein thrombosis

Each scenario began with an introductory page, explaining the situation which led to the hypothetical medical diagnosis, the medical diagnosis is (and means) and what one of the recommended treatments is.

Next, participants were presented with three risk and three benefit statements for that treatment scenario (which were closely matched for characters/word length). These were presented in one of four orientation $X$ order presentations, such that either the risk information or benefit information
is presented first (either on the left or top) and with either information presented in a vertical (figure 1a: up/down) or horizontal (figure 1b: left/right) orientation.


Figure 1c: Example Scenario: Statins Horizontal: Benefits First

Irrespective of presentation order (orientation) after each risk (benefit) presentation, participants were presented with six treatment evaluation questions ( 3 positive \& 3 negative):

- P1: How favourable would you rate the treatment?
- P2: Would you choose the treatment?
- P3: Would you recommend this treatment?
- N1: How concerned would you be about the side effects?
- N2: Would you avoid this treatment?
- N3: If you were to choose this treatment, how likely do you think it is that you would experience one of its side effects?

For each of these questions, rating responses were made on a 7-point scale. For the analysis, because of a high Cronbach's alphas ( $\alpha$-range . $87-.92$ in study 1 ; $\alpha$-range . 85 .92 in study 2), responses from across these six questions (after reverse coding the scores for the three negative questions) were combined for each participant to form an overall treatment evaluation rating.

## Eye tracking

In Study 2, eye-tracking was conducted during the risk and benefit presentation phase of the study using the EyeLink 1000. The study was conducted within the associated Experiment Builder software application. Interest areas were defined around the six statements ( 3 risks, 3 benefits) and the two titles (Benefits, Risks).

From the eye-tracking data that was recorded, three main measures were calculated: SMRD order scores, time proportion (on Benefits) and number of transitions, and are explain below:

SMRD Order: This was chosen as a way of measuring whether actual looking order is consistent with the "manipulated order" (i.e. that when the risks are presented first according to our design, people look at the risks first). To create our measure of looking order, the formula used by Johnson, Häubl and Keinan (2007) was adapted ${ }^{1}$. While they used it to examine the order of thoughts (which they had people write down), we adapted it to examining looking order by replacing thoughts with fixations in the formula.

To calculate this, the formula below (Figure 2) was used (where MR represent median rank):

$$
\frac{2\left(M R_{\text {Eenefit }}-M R_{\text {Risk }}\right)}{\text { Number of Fixations }}
$$

Figure 2: SMRD Order Formula
To allow this calculation of ranks to be conducted, fixations were coded for order. For example, the first fixation coded as " 1 ", the second as " 2 " and so forth until all fixations included. Taking an example, looking mainly at the risk first lead to a positive SMRD score, and mainly benefits first a negative SMRD score.

Time Proportion: Another potentially relevant variable in determining people's subsequent choices is the proportion of time spent looking at the different types of information (i.e. risk and benefits), which helps to represent a measure of attention (i.e. amount of attention paid to each type of information).
Using the interest areas that were preset into the eye-link analysis software, this measure calculated the time spent looking in each interest area. This was then transformed into a proportion of time by dividing the time spent in the interest area by total time spent. From this, the proportion of time spent on the benefits was calculated by summing the proportion of time spent on the four relevant interest areas (i.e. the three benefits and benefit title for time spent on the benefits). Thus, at the end, a percentage score out of 100 was calculated and represented the balance of time spent on the benefits (versus the risks).

No of Transitions: This also represented a measure of attention, but this measured how people switched their attention between information sets. This was chosen to investigate the findings from the search literature of a tendency to make horizontal rather than vertical saccades (see introduction and prediction 3 for details).

For our purposes, this measured the number of times people switched between reading the risk information block to reading the benefit information block. A transition was

[^346]coded every time two adjacent fixations in the time-ordered fixation sequence were from different information blocks (i.e. one was from the risk block, while the other was from the benefits block).

## Results

## Eye-tracking Information Search Analysis

Each analysis was conducted collapsing the four scenarios ${ }^{2}$ and using a generalized estimating equations (GEE) model ${ }^{3}$ (with exchangeable correlation matrix, robust standard errors and Gaussian identity matrix).

## Order SMRD Score

As can be seen from the graph in figure 3 (note: scores of 1 denote reading all risks before all benefits, while -1 denotes reading all benefits before all risks), our manipulation of order was successful ( $B=0.80, \mathrm{Z}=24.44$, $\mathrm{p}<.001$ ), with those presented with the risks first (i.e. on top or on the left hand side) reading the risks first (and therefore having a positive SMRD scores) and those shown the benefits first showing a negative SMRD order score.


Although no significant main effect of orientation was found, a significant interaction between order and orientation was identified ( $\mathrm{B}=0.08, \mathrm{Z}=2.59, \mathrm{p}=.009$ ). As this shows, what is happening is that SMRD scores are closer to 0 in the horizontal orientation. As such, supporting our third prediction that the effect of order would be stronger in the vertical orientation (i.e. SMRD scores closer to extremes of +1 and -1 ).

## Time Proportion Spent on Benefits

As can be seen from the graph in figure 4, only small differences are seen between the different orientations. One

[^347]would predict that people would look evenly at the information and, as figure 2 shows, this is the case when the benefits are presented first (Horizontal: $t(40)=1.09, p=.283$. Vertical: $\mathrm{t}(46)=-0.79, \mathrm{p}=.43)^{4}$. However when risks are presented first, the time spent on the benefits drops and search becomes risk heavy (Horizontal: $\mathfrak{t}(32)=-5.17 \mathrm{p}<.001$; Vertical: $\mathrm{t}(54)=5.65, \mathrm{p}=.001)$.


Figure 4: The effect of order and orientation on time proportion spent searching the benefits information

## No of Transitions

As Figure 5 below demonstrates, while order (whether risks or benefits first) played little role in how many times participants transitioned between risk and benefit information, orientation made a big different to how many time people switched between reading the different information sets ( $\mathrm{B}=-0.85, \mathrm{Z}=-4.09, \mathrm{p}<.001$ ). When participants were presented with the information side by side (the horizontal orientation), participants were more likely to switch between reading information about the risks and reading information about the benefits. For the vertical orientation (where information was presented above and below), switching occurred less commonly.


Figure 5: The effect of order and orientation on number of transitions between information sets

## Effect on Overall Treatment Evaluations

"Manipulated Order" Analysis: As figure 3 reveals, people look in the manipulated order (i.e. generally looked

[^348]at the benefits before the risks in the benefits then risks conditions), as such it is appropriate to investigate the effect of these different order X orientation conditions on treatment evaluations. Across both experiment 1 and 2, only in 1 of the 8 scenarios did an effect of order reach significance (ACE Study 1, $F(1,102)=7.90, p=.006$, $\eta^{2}=.072$ ) all others $\mathrm{F}<1.30$, $\mathrm{p}>.275$ ). No main effects of orientation (all $\mathrm{F}<2.27, \mathrm{p}>.135$ ) or interaction between order and orientation (all F<1.96, p>.170) were found. ${ }^{5}$ Thus, in most cases, the treatment evaluation ratings across these four conditions were similar, often sitting close to the middle of the scale.

Eye-tracking Analysis: Holding all other eye-tracking variables constant (i.e. SMRD order and orientation), time proportion on benefits positively predicted treatment evaluations ( $\mathrm{B}=0.02 \mathrm{Z}=2.28, \mathrm{p}=.023$ ).

## Discussion

Considering the original study by Bergus et al. (1992), they found a recency order effect with the most recent information having the biggest effect on subsequent perceptions of a treatment. Unlike those researchers, our results did not support such a recency order effect, finding instead no consistent pattern of recency or primacy. The search data discussed later however does hint at a primacy advantage for negative information as a more likely possibility. Three suggestions are made to explain why such a disparity in results may have been found.

First, returning to Bergus et al.'s (2002) aspirin scenario, it is not clear that the list lengths (i.e. lists of risks and benefits) were matched for either word or character length. In particular, the risk list length appeared longer ${ }^{6}$. Such differences may have enhanced any order effects.

Second, such difficulties in finding consistency in order effects has been discussed by Hogarth and Einhorn (1992). One factor that they highlight of particular relevance for comparing this study to the previous study is the role of evaluation task (or response mode) differences, particularly in short information scenarios. These researchers have argued that differences in the response mode can change the way people evaluate information and what information is used as an anchor. Considering this anchor in particular, while Bergus et al (1992) had an initial evaluation question which provides people with an initial anchor, our study use only an end of sequence response, with no initial anchor specified. With no anchor provided, it is the first piece of information which provides the anchor value. Figure 6a provides an illustration of how these differences in anchor and processing strategy may predict the different pattern of results found in these studies. In particular, as Figure 6

[^349]shows, while a clear recency order effect is seen for Bergus et al. (2002), for our study, by using the first piece of information as an anchor, very little difference between final evaluation scores is predicted.


Third, in particular, it is worth noting that in keeping our scenarios simple and thereby short, this may have attenuated the size of any potential overall order effects on treatment evaluations. This may have occurred since with such short scenarios, it is not unreasonable to think that people could keep all 6 pieces of information (i.e. 3 risks and 3 benefits) in their memory at once, at least once the items' "gist" meaning had been processed (Miller, 1955; Reyna, 2008). Further, such differences in response mode appear to become less important as the scenario length increases and primacy becomes the predicted order effect irrespective of response mode.
Despite these factors, which may affect people's processing of the information in order to make a decision, the search data responses should be relatively unaffected by the strategy choice variable (which should affect starting position and information integration rather than search). Further, at longer scenario lengths, differences would be predicted to be more pronounced with the effect of the response mode diminished.
Looking at this search data reveals that changes in order and orientation do appear to change how people search the information presented. First, providing support that our external manipulation of order was successful, actual search order (as measured by the order SMRD) mapped onto the manipulated order that the scenario was assigned to. When participants were presented with the benefits "first" (at top or left side), they generally read the benefit information first, therefore having a negative SMRD value. When risks were shown "first", risk information was read first and participants had a positive SMRD value.

Thus, even with simultaneous presentation of information (and therefore no external constraints on order of search), people are still affected by a decision aid designer choice of where to place the information in a table. Rather than switching between reading the risks and benefits, participants generally chose to read each information set sequentially. Such effects suggest that typical reading
patterns (i.e. the tendency to start reading at the top left) in some way constrains how people will read information even when presented simultaneously (Foulsham et al., 2013). Thus, suggesting that use of simultaneous presentation format do not automatically remove presentation order effects from consideration.

Of note within this effect of order, the addition of orientation as a variable in our study highlights a further dimension to consider with this effect. In particular, orientation appears to affect the strength of the search order effect, with a stronger order effect seen in the vertical orientation than the horizontal orientation. Thus, a clearer sequential processing strategy - following the manipulated order (thereby reading all risks then all benefits, or vice versa) - is seen in this vertical orientation. When risks and benefits are presented side-by-side however in the horizontal orientation, this is weakened and switching between information sets (i.e. between risk and benefit information) becomes more common. Such results support our third prediction of a stronger order effect in the vertical orientation, based on previous search evidence by Foulsham et al. (2008) which found that in picture search horizontal movements are more common than vertical movements. Such an effect is further supported in the analysis of transitions between information sets where, switching between sets is significantly higher in the horizontal orientation than the vertical orientation.

Considering our final attention-based search measure taken, time proportion spent reading the benefits, this revealed that while an equal proportion of time is spent (approximately $50 \%$ ) on the risks and benefits when benefits are presented first, when risks are presented first, the time spent on the benefits drops and search becomes risk heavy (closer to a $60 / 40$ split). This asymmetric difference may suggest that risk information is particularly attention "grabbing" and difficult to engage from, thereby sustaining attention for longer and reducing the time left to spend reading the benefits. Support for such a finding can be seen in the negativity bias literature, where a propensity to attend to, learn from and use negative information more than positive information has been found, Vaish, Grossmann, \& Woodward, 2008).

Further, such a finding would predict a primacy advantage for the risk information rather than a recency effect, since the extra attention placed on this information as the negativity literature suggests, should lead people to "learn" and "use" this negative information to a greater extent. Thus, leading to more negative treatment evaluations. Such an effect is supported by analysis of the effect of the eyetracking variables, where only time-proportion was a significant predictor of treatment evaluations when holding the other variables constant.

## Future Directions

In future research looking at more complex scenarios, we predict that the search order and orientation differences
would become more pronounced as the amount of information presented is increased and aggregating information in a sequential fashion becomes essential.

Given more complex risk and benefit presentation scenarios, we predict that with such scenarios should lead to a primacy rather than a recency advantage. Such is predicted from our finding of a risky heavy search when risks are presented first (a primacy advantage), evidence that longer scenarios force toward a primacy advantage (Hogarth and Einhorn, 1992) and a reduced role of response mode in these longer scenarios.

For a second type of more complex scenario, multi-attribute (multi-option) choice, the role of orientation on search may be of particular importance. We predict that such orientation changes may change whether a more withinoption search or between option-search processing strategy is taken. Such differences in search tend to lead to the adoption of different choice strategies, which may ultimately affect which option is preferred (Hills \& Hertwig, 2010).

## Conclusion

These results highlight the role that seemingly arbitrary choices about the design of a decision aid, informational leaflet or website, such as order or orientation of the information can affect how information is searched. In certain situations, these search differences may subsequently affect judgements and choices made using such information as a basis for knowledge about a choice scenario.

## Acknowledgments

Financial support for this study was provided entirely by a collaborative PhD studentship received by the first author from the University of Cambridge (Winton Fund) and Economic Social Research Council (ESRC). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

## References

Ashby N. J. S., \& Rakow T. Forgetting the past: individual differences in recency in subjective valuations from experience. Journal of Experimental Psychology: Learning, Memory, \& Cognition. 2014;40:1153-1162.

Bergus G. R., Levin I. P., \& Elstein AS. Presenting Risks and Benefits to Patients. Journal of General Internal Medicine. 2002; 17(8): 612-517.

Foulsham T., Gray A., Nasiopoulos E., \& Kingstone A. Leftward biases in picture scanning and line bisection: a gaze-contingent window study. Vision Research. 2013; 78: 14-25

Foulsham T., Kingstone A., \& Underwood G. Turning the world around: patterns in saccade direction vary with
picture orientation. Vision Research. 2008; 48: 17771790.

Hardisty, D.J., Johnson, E.J., \& Weber, E.U. A dirty word or a dirty world? Attribute framing, political affiliation and query theory. Psychological Science, 2010; 21(1): 8692.

Hills T.T., \& Hertwig R. Information search in decisions from experience: do our patterns of sampling foreshadow our decisions?. Psychological Science. 2010; 21(12):1787-1792.

Hogarth R. M., \& Einhorn H. J. Order effects in belief updating: the belief-adjustment model. Cognitive Psychology. 1992; 24(1): 1-55.

Honish, GG., Edwards, E.P., Eiden, RD., \& Leonard, KE. Analysing family data: A GEE approach for substance use researchers, Addictive Behaviours, 2010:; 35(6): 558-563.

Johnson, Haübl \& Keinan (2007). Aspects of endowment: a query theory of value construction. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(3), 461-474.

Miller G. A. The magic number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review. 1955; 101(2); 343-352.

NHS Choices. Should I have a PSA test?. 2015, March 25 [Online Article] [Accessed 2016 August 27]. http://www.nhs.uk/Livewell/Prostatehealth/Pages/psatest.aspx

Rakow T., Demes K. A., \& Newell B. R. Biased samples not mode of presentation: re-examining the apparent underweighting of rare events in experience-based choice. Organizational Behavior and Human Decision Processes. 2008; 106(2):168-179

Reyna V. F. A theory of medical decision making and health: fuzzy trace theory. Medical Decision Making. 2008;28(6): 850-865.

Ubel P. A., Smith D. M., Zikmund-Fisher B. J., Derry H. A., McClure J., Stark A., Wiese C., Greene S., \& Jankovic A. Testing whether decision aids introduce cognitive biases: results of a randomised trial, Patient Education and Counseling. 2010; 80:158-163.

Vaish, A., Grossmann, T., \& Woodward, A. Not all emotions are created equal: The negativity bias in socialemotional development, Psychological Bulletin. 2008; 134(3): 383-403.

# Risk and Rationality in Decisions to Commit Crime 

Rebecca Helm

Cornell University

Valerie Reyna<br>Cornell University

Renee Williamson<br>Cornell University

Allison Franz<br>Cornell University


#### Abstract

Criminal behavior and related disorders have been associated with abnormal neural activity when experiencing or anticipating risks and rewards, as well as when exercising inhibition. However, behavioral and neural substrates of risk preferences and criminality have received scant attention when unconfounded with experience, anticipation, and inhibition. We test predictions of fuzzy-trace theory (FTT) in two experiments using a risky-choice framing task. Behavioral results show that individuals with a greater history of criminal behavior are less likely to engage in simple meaning-based processing and are less confident when doing so. These findings are supported by fMRI results showing a greater history of criminal behavior is associated with increased activation in regions associated with cognitive control when engaging in simple meaning-based processing. These results provide insight into the cognitive processes and brain mechanisms that are associated with criminal behavior.


# Planning in Action: Interactivity Improves Planning Performance 

Emma Henderson, Gaëlle Vallée-Tourangeau<br>Department of Management, Kingston University, Kingston-upon-Thames UNITED KINGDOM KT2 7LB<br>e.henderson/g.vallee-tourangeau@kingston.ac.uk

Frédéric Vallée-Tourangeau<br>Department of Psychology, Kingston University, Kingston-upon-Thames UNITED KINGDOM KT1 2EE f.vallee-tourangeau@kingston.ac.uk


#### Abstract

Planning is an everyday activity that is extended in time and space, yet is frequently studied in the absence of interactivity. Successful planning relies on an array of executive functions including self-control. We investigated the effects of interactivity and self-control on planning using a sequential-task paradigm. Half of the participants first completed a video-viewing task requiring self-control of visual attention, whereas the other half completed the same task without the selfcontrol constraint. Next, and within each of these groups, half of the participants manipulated cards to complete their plan (high-interactivity condition); for the other half, plans were made with their hands down (lowinteractivity condition). Planning performance was significantly better in the high- than in the lowinteractivity conditions; however the self-control manipulation had no impact on planning performance. An exploration of individual differences revealed that long-term planning ability and non-planning impulsiveness moderated the impact of interactivity on planning. These findings suggest that interactivity augments working memory resources and planning performance, underscoring the importance of an interactive perspective on planning research.


Keywords: personal planning, time management, distributed cognition, self-control, ego depletion

## Introduction

Planning is an essential cognitive process that is key to achieving productive time management. Successful planning depends on the ability to anticipate a sequence of operations intended to achieve one or more goals, requiring the capacity to effectively delay and resume the pursuit of goals, according to current resources and constraints (Hayes-Roth \& Hayes-Roth, 1979; Patalano \& Seifert, 1997). The planning process may involve initial planning; a systematic, rational approach where solutions are formulated ahead of plan execution (Morris \& Ward, 2004). As such they are subject to constraints on processing, including working memory resources. An alternative to this top-down model is opportunistic planning where the plan develops in situ driven by incoming information, rather than being entirely goaldirected in advance of any moves (Davies, 2005). Selection of initial or opportunistic planning is based on the problem complexity and environment, and individual differences (Davies, 2005).
Efficient planning depends on the coordination of a
variety of executive functions, ranging from formulating a sequence of sub-goals that together embody a plan, storing and updating the plan, consciously monitoring, up to controlling and coordinating the plan to effect the desired outcome (Morris \& Ward, 2004). This view suggests that successful planning depends on two key cognitive concepts: working memory and self-control. Self-control is defined as "the exertion of control over the self by the self" (Muraven \& Baumeister, 2000, p. 247), occurring when a person attempts to override or inhibit the way they would otherwise think or behave. Working memory is a memory system for temporal information, and is a key theoretical concept for understanding how a limited amount of information is kept temporarily highly available, integrating external and previously-stored information in order to facilitate cognition and complex behaviour (Logie \& Cowan, 2015). Working memory capacity has been linked with the ability to control attention (Engle \& Kane, 2004) and avoid impulsive interferences. These are important requisites of selfcontrol (Broadway, Redick, \& Engle, 2010). With multiple perspectives and potential actions held in working memory during planning, self-control seems key to planning appropriate actions and suppress inappropriate ones. This conception of planning, however, assumes that people solely rely on their mental resources when they engage in planning tasks. Yet reallife planning admittedly involves more than just mental processing: people who plan do so by making notes, writing and rearranging "to-dos" in lists, emails or index cards. In other words, they not only access but also interact with external information by manipulating to-dos while they plan.

## Distributed Cognition and Interactivity

Traditionally, thinking is considered to occur in the head, sandwiched between perceptual inputs and behavioural outputs (to adapt Hurley, 2001). More recently an epistemological shift to a distributed cognition perspective proposes a dynamic cognitive system whose structure is distributed across the internal resources of the individual (such as acquired knowledge) and the resources external to the individual (such as material representations and tools; e.g., Kirsh, 2010). Studies of planning are conventionally conducted in the absence of a distributed cognitive system and with a focus on the
"temporal ordering of action" (Kirsh, 1995, p. 31). Interactivity, by contrast, configures a dynamic agentenvironment system scaffolded from resources internal and external to the agent. In all likelihood, interacting with the physical feature of a problem results in a simpler problem configuration that engages perceptual and pattern matching processes. In addition, by changing the spatial rearrangement, non-strategic manipulations may serendipitously determine what to do next. In mental arithmetic, for example, moving number tokens when performing long sums enhances accuracy and improves efficiency (Vallée-Tourangeau, 2013); congenial, easy-toremember interim totals can be identified and physically segregated, action affordances shift as the problem configuration is transformed, the allocation of attentional resources is governed by dynamic changes in the problem. Similarly, in a Bayesian reasoning task, manipulating cards representing elements of a statistical sample led to a sharp increase in performance (ValléeTourangeau, Abadie, \& Vallée-Tourangeau 2015). One possible explanation for the positive impact of interactivity on performance is that increasing interactivity reduces the processing burden on an agent's working memory. If this is the case, we should expect that higher levels of interactivity and resulting opportunities to manipulate and rearrange information in a planning task should promote better planning performance.

## Self-control

Self-control is required to maintain goal intentions and plans over time, and resist the conflicts of immediate impulses such as attending to tempting stimuli. (Hofmann, Friese, \& Strack, 2009). Research into failures of self-control (acting on impulses) proposes that it is a finite mental resource that limits self-regulatory capability. Motivated by this approach Baumeister and colleagues (e.g., Baumeister, Bratslavsky, Muraven, \& Tice, 1998) developed the strength model of self-control with the central tenet that willpower is a limited resource akin to energy, which becomes fatigued or depleted with use, temporarily reducing the capacity for subsequent self-control. Baumeister et al. (1998) termed this state of reduced self-control ego depletion. Support for the model comes from research using a sequential-task paradigm; participants are required to engage in an initial task of self-control, and decrements in their performance are then measured on a second, unrelated task of self-control. A meta-analysis of 83 sequential-task studies reported a medium effect size (Cohen's $d=.62$ ) of ego-depletion (Hagger, Wood, Stiff, \& Chatzisarantis, 2010).
More recently, the ego-depletion effect has been called into question with studies failing to detect the phenomenon (e.g., Lurquin et al., 2016). A reanalysis of the Hagger et al. (2010) data and another meta-analysis identified small-study bias and an inflated effect size (Carter \& McCullough, 2014). This was followed by a
multi-lab Registered Replication Report (RRR) involving 23 labs worldwide which found an overall ego-depletion effect of close to zero (Hagger et al., 2016). This RRR used just one combination of tasks, and the present study responds to the recent calls for further replications using different combinations of tasks, increased sample sizes, and to investigate potential moderating variables (e.g., Lurquin et al., 2016).
Since planning involves self-control activities such as monitoring and coordinating actions, we should expect ego depletion to impair planning performance. Yet, as interactivity may offer a platform for offloading some of the cognitive processing required to monitor and coordinate action, we anticipate that the impact of ego depletion on planning performance will be tempered when cognitive agents are free to physically interact with their plan.

## The Present Experiment

Despite its ubiquity and importance in everyday life, planning research to date has tended to focus on planning dysfunctions and the order of actions, ignoring both the environment in which planning takes place and the cognitive state with which the participant comes to the task. This research typically uses one of two general types of tasks: puzzle-based tasks, which involve simple, mechanistic, easily-controlled procedures (e.g., the Tower of London), or real-world planning tasks, which invoke familiar procedures and contexts of a complexity analogous to everyday activities (Morris \& Ward, 2004; e.g., the Virtual Planning Task, Miotto \& Morris, 1998). The present experiment adopted a distributed-cognition perspective to determine how the manipulation of task interactivity and ego-depletion would affect performance on a real-world planning problem. We designed a lowinteractivity condition using a static paper presentation during which participants had to keep their hands down, while in the high-interactivity conditions cards corresponding with to-be-executed tasks could be manipulated and re-arranged as participants saw fit. Both conditions required working memory and self-control to switch and control attention between remembering plans and actioning them. We hypothesised that the inflexible, unmodifiable environment offered in the low-interactivity condition would lead to poorer performance relative to the high-interactivity condition. Since planning is reflected in physical changes in the environment, we hypothesised that the high-interactivity condition would lead to improved performance relative to the lowinteractivity condition.
A secondary aim of this experiment was to test whether the offloading of cognitive processing afforded by highly interactive environments could act as a buffer for the negative impact of depleted self-control resources on planning performance. We selected a widely-used egodepletion task (e.g., Schmeichel, 2007), along with a
planning task that is demanding of executive control in order to maximise the chances of demonstrating the egodepletion effect. The experimental set-up replicated as closely as possible that of the original author (Brendan Schmeichel), including a verbatim script of the original task that was obtained with the help of the authors of the recent replication study (Lurquin et al., 2016). We nearly doubled the sample size used in typical ego-depletion studies. Participants were allocated to either a control or an experimental ego-depletion group, the latter requiring self-control in order to direct attention towards an interviewee and away from distracting words presented on the screen. In order to understand what participants in both conditions were really doing during the videoviewing task, and to ensure the two conditions were indeed distinct, supplementary measures such as a word memory task to test adherence to instructions and additional questions regarding the ego depletion task were included. Based on the resource model of self-control participants in the ego-depletion condition, who performed the initial act of self-control, should perform worse in the subsequent planning task than participants in the control condition. Though, given the conflicting findings in the extant literature, and our new combination of tasks, we set out with an exploratory perspective on the effect of this frequently-used video-viewing task combined with our planning task.
In addition, we included self-report measures of flow, planning, and impulsivity to explore whether they would moderate the impact of interactivity on planning performance.

## Method

## Participants

One hundred participants ( 73 women, $27 \mathrm{men}, M_{\text {age }}=$ 31.90 years, $S D=11.77$ ) were recruited; some received course credits. All participants were naive to the purpose of the research.

## Procedure

The experiment employed a 2 (interactivity: high or low) x 2 (ego depletion: depletion or control) between-groups design. Participants were randomly allocated to one of the four experimental conditions ( $n=25$ per group). All tasks were completed in a single testing session, lasting approximately 45 min , which was divided into three phases. Participants first watched the ego-depletion video. This was followed immediately by the planning task. Finally, participants completed a series of self-report measures.
Ego-depletion task. Participants watched a 6 min , silent video featuring a woman being interviewed by an offscreen interviewer, as the initial task in a sequential-task paradigm (Schmeichel et al., 2003). During the video 36
common, one-syllable words (e.g., "play") appeared at the bottom of screen for 10 s each. Words appeared in black font on a white background and took up approximately one quarter of the screen.
In the ego-depletion condition participants were instructed to focus attention on the woman's face and not to look at the words that appeared on the screen. The control condition was identical except, crucially, no instructions were given regarding the words that appeared on the screen and participants were asked to watch the video as if they were "sitting at home watching TV". While participants viewed the video, the experimenter moved outside the room. Two modifications were made to the original task. First, distance between participant and screen was standardised at 40 cm . Second, to increase their saliency, the position of the words was changed from bottom right to bottom centre. The size, colour and font remained unchanged.
Planning task. Next, participants completed an adapted version of Miotto and Morris's (1998) planning task in either the low-interactivity list condition, or the highinteractivity board game condition. The object of the task was to plan and execute a sequence of specified activities on the four days of the week preceding a trip abroad. Twenty-eight activities were offered for completion, 16 of which were relevant to the trip. The remaining 12 activities were not relevant and were termed "distractors". To simulate the constraints of real world planning, once the participant had executed the activities for a given day they could not change their plan and had to move on to the next day. Participants were seated at a desk and instructed on the main features and rules of the task. They were advised that they could carry out four tasks per day; two in the morning and two in the afternoon and that not all tasks could be completed. There was no time limit but participants were instructed to complete the task as quickly and as accurately as possible. Two measures were used to calculate planning performance: (1) accuracy - a choice of task was considered correct if it was one of the 16 relevant activities, and where applicable, completed on the specified day and time; (2) latency per correct task calculated as overall latency divided by accuracy.


Figure 1: The experimental setting, high-interactivity condition (left), and low-interactivity condition (right).

High-interactivity condition. The 28 activities were printed on individual action cards ( 55 x 88 mm ). These action cards could be selected by moving the card into a
central day frame split into sections representing the morning and afternoon. As in the low-interactivity condition, a summary card which specified which activities needed to be done during the week was always available. Finally, an execution board was used to place cards that had been selected and representing tasks that have been completed (see Fig. 1, left panel). Participants were handed the 28 activity cards in a randomly ordered pack and were free to move the cards as desired in the working area. Participants could monitor their goal progress at any stage by checking the execution board.
Low-interactivity condition. The low-interactivity condition used a list of 28 activities to be performed (see Fig. 1, right panel). In this condition participants were instructed not to touch any task materials and to keep their hands on the desk for the duration of the task. To choose a task for completion participants verbally instructed the experimenter of their selection.
Additional Measures. Upon completion of the planning task, participants answered a flow questionnaire developed to gauge participant's enjoyment and engagement during a task (Vallée-Tourangeau et al., 2015). Next, all participants were given a surprise memory test for the words presented during the video in the first part of the experiment. The test compromised 36 words: 18 that appeared during the initial video-viewing task, and 18 that did not (the same test designed by J. H. Lurquin, personal communication, 2 March 2016). Participants then judged whether they had seen the words previously by circling either yes or no. The memory test was followed by a series of manipulation checks for the ego depletion task. Participants were asked to rate the difficulty of complying with the video task instructions they were given prior to watching the video $(1=$ not at all difficult to $10=$ very difficult). Following Lurquin et al. (2016) participants also rated how much effort they had put in to the task, and how hard they had tried to ignore or remember the words $(1=$ none to $10=a(\operatorname{lot})$.
Finally, participants completed individual differences measures of planning and impulsivity as an independent measure of planning ability. Participants completed Simons and Galotti's (1992) planning survey, a 31-item scale measuring everyday planning style, and Lynch, Netemeyer, Spiller and Zammit's (2009) propensity to plan for time short run and long run 6 -item scales. This is a $30-$ item scale, scored using three sub-scales: attentional, motor and non-planning.

## Results

Depletion participants rated the video task as more effortful ( $M=7.38, S D=1.99$ ) than did the control participants ( $M=6.74, S D=2.32$ ), however the difference was not significant, $t(98)=-1.48, p=.142$. Participants in the depletion condition remembered significantly fewer words ( $M=4.66, S D=4.31$ ) than control participants ( $M=$ $12.78=S D=3.72), t(98)=10.09, p<.001$, which
suggests that they complied, in part, with the task instructions. Although, if the depletion participants had fully complied with the instructions then they would not have remembered any words.
The main dependent measure in the planning task was accuracy, the maximum possible score being 16; the data are reported in Figure 2. Participants in the highinteractivity condition were more accurate than those in the low-interactivity condition, but the ego-depletion paradigm appeared to have no effect on planning performance. A 2 x 2 between subjects analysis of variance (ANOVA) revealed a significant main effect of interactivity, $F(1,96)$ $=78.03, p<.001$, but neither the main effect of ego depletion nor the interaction effect were significant ( $F \mathrm{~s}<$ $1)$.


Figure 2: Mean accuracy (left panel) and mean latency per correct task (in seconds, right panel) as a function of level interactivity and the experience of the ego depletion task (error bars are standard error of the means).

Latency per correct task was calculated as total latency divided by accuracy; lower scores reflect better performance. Figure 2 shows that generally participants in the high-interactivity condition were faster than those in the low interactivity condition. A $2 \times 2$ independent ANOVA revealed a significant main effect for interactivity $F(1,96)=25.72, p<.001$. Again, neither the main effect of ego depletion nor the interaction effect were significant ( $F \mathrm{~s}<1$ ).
We explored whether individual differences in planning, impulsiveness, and experience of flow moderated the impact of interactivity on planning performance. Planning performance was positively associated with flow, $\beta=.22$, $t(96)=2.45, p=.016$, although flow was not a moderator of the interactivity effect on planning performance; $\beta=$ $.03, t(96)=0.36, p=.72$. More interestingly, the impact of interactivity on planning performance was moderated by individuals' propensity to plan in the long-run, $\beta=-.15$, $t(96)=-2.08, p=.04$. Specifically, higher propensity to plan in the long-run was associated with higher performance under low interactivity but it did not predict performance under high interactivity (see Figure 3, left panel). The impact of interactivity on performance was also moderated by non-planning impulsiveness, $\beta=.16$, $t(96)=2.22, p=.03$. In this case, higher scores of nonplanning impulsiveness were associated with higher
planning performance under high levels of interactivity and lower planning performance under low levels of interactivity (see Figure 3, right panel).


Figure 3: Relationship between planning accuracy (Performance) and propensity to plan in the long run (left panel) and non-planning impulsiveness (right panel) as a function of interactivity levels (High vs. Low). Note. $\mathrm{CF}=$ Centred Form or mean deviation form.

## Discussion

In this experiment participants completed a planning task in two different interactivity conditions, one which permitted spatial rearrangement of the task, and one which did not. Generally, participants were more accurate and achieved faster latency per correct answer in the highinteractivity condition. These results can be explained in terms of the affordances provided by the different task environments. In the low-interactivity condition participants were forced to manipulate information mentally, relying on executive function-directed initial planning. As such, performance was limited by the participants' working memory capacity.
In contrast, the dynamic interface of the highinteractivity condition facilitated new affordances for task completion and paved the way for opportunistic planning, where task selection was guided, in part, by the physical changes in the configuration of the problem. Planning was interspersed with physical execution, alleviating the load on the participants' working memory compared to the complex initial planning in the head required by the lowinteractivity constraints. The significant increase in task performance in this condition could not be attributed to individual differences since there were no condition differences on measures of planning. Instead, the highinteractivity environment allowed participants to restructure and simplify the problem presentation in a way that was conducive to solving the task. For example, reorganising the activity cards made it possible to collate related activities, discard distractor cards and constantly track the state of the task. Furthermore, when participants planned all time-specific tasks first, they then perceived the gaps left in the plan, thus placements for single item tasks were physically discovered (Kirsh, 2010).
We sought to explore cognitive individual differences that would predict performance in both conditions of the planning task. We hypothesised that the skills implicated when planning in the head would be different from the
skills employed when planning in a distributed environment. Long Run Planning ability only mattered for those participants in the low interactivity condition; it had no effect on performance in the high interactivity condition. Conversely, high levels of non-planning impulsivity put people at an advantage under high interactivity, and at a clear disadvantage in the low interactivity condition. The fact that planning performance was superior in the high-interactivity, in the absence of a difference in self-reports of planning ability, suggests that the manipulation of cards augmented planning abilities (via working memory) above those measured with planning scales. This finding supports previous work (e.g., Vallée-Tourangeau, Sirota, \& Vallée-Tourangeau, 2016) and suggests that interactivity may functionally augment cognitive resources.

## Self-Control and the Elusive Ego Depletion Effect

The present experiment also examined the effect of ego depletion on planning. As advocated by Lurquin et al. (2016) we used a larger than average sample size ( $N=$ 100) and explored a new combination of tasks. Despite making these and other modifications to this highlyreplicated depletion task, the main effect of ego depletion was not significant. This result is inconsistent with the strength model (Baumeister et al., 1998), and many previous studies, including those from the laboratory where the term ego depletion was first coined, and where the video-viewing task was developed. However, it is consistent with more recent research that has failed to detect the ego-depletion effect, and most notably Lurquin et al. (2016), which used the same initial video-viewing task.
A critical pre-requisite of the sequential-task paradigm is that both tasks require the use of self-control. The present planning task has not been used as a second task in the sequential task paradigm: one possible explanation for the absence of ego depletion is that the outcome planning task did not require self-control. Yet, there is little doubt that planning requires the deliberate control of actions across time, and Baumeister and Vohs (2016) argue that planning draws on the same limited resource as self-control. Additionally, in the present study, task constraints such as adhering to activities stipulated on the summary card required that impulses to follow habitual holiday-planning responses be overridden using self-control. Further, it could be argued that the low-interactivity instructions demand self-control by requiring participants to keep their hands down on the table. Despite apparently meeting the conditions necessary to induce the ego-depletion, we found no evidence of an effect. Since the video-viewing task lacks an objective measure of task performance (Lurquin \& Miyake, 2017), we included the word memory task and assume that performance here points to adherence to the video task instructions. Participants in the ego-depletion condition remembered some words, indicating that they
looked at them and so did not fully adhere to the task instructions. If participants in this condition did not inhibit their natural impulse to respond to the attention-capturing words, they were not using self-control in the first task and thus would not be, and indeed were not, depleted in the second task. Without substantial modification to the task procedure and an objective measure of performance, our findings indicate that the video-viewing task does not operationalise ego-depletion as intended.

## References

Baumeister, R., \& Vohs, K. (2016). Strength model of selfregulation as limited resource: Assessment, controversies, update. Advances in Experimental Social Psychology, 54, 67-127.
Baumeister, R. F., Bratslavsky, E., Muraven, M., \& Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? Journal of Personality and Social Psychology, 74, 1252-1265.
Broadway, J. M., Redick, T. S., \& Engle, R. W. (2010). Working Memory Capacity: Self-control is (in) the goal. Self control in society, mind, and brain, 1, 163174.

Carter, E. C., \& McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? Frontiers in Psychology, 5, 823.
Davies, S. P. (2005). Planning and problem solving in well-defined domains. In R. Morris \& G. Ward (Eds.), The cognitive psychology of planning (pp. 35-52). New York: Psychology Press.
Engle, R. W., \& Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. Psychology of learning and motivation, 44, 145-199.
Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., \& Zwienenberg, M. (2016). A multi-lab pre-registered replication of the ego-depletion effect. Perspectives on Psychological Science, 11, 546-573.
Hagger, M. S., Wood, C., Stiff, C., \& Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta- analysis. Psychological Bulletin, 136, 495-525.
Hayes-Roth, B., \& Hayes-Roth, F. (1979). A cognitive model of planning. Cognitive Science, 3, 275-310.
Hofmann, W., Friese, M., \& Strack, F. (2009). Impulse and self-control from a dual-systems perspective. Perspectives on Psychological Science, 4, 162-176.
Hurley, S. (2001). Perception and action: Alternative views. Synthese, 129, 3-40.
Kirsh, D. (1995). The intelligent use of space. Artificial Intelligence, 73, 31-68.
Kirsh, D. (2010). Thinking with external representations. AI \& Society, 25, 441-454.

Logie, R. H., \& Cowan, N. (2015). Perspectives on working memory: Introduction to the special issue. Memory \& Cognition, 43, 315-324.
Lurquin JH, Michaelson LE, Barker JE, Gustavson DE, von Bastian CC, et al. (2016) No Evidence of the EgoDepletion Effect across Task Characteristics and Individual Differences: A Pre-Registered Study. PLoS ONE, 11, e0147770.
Lurquin, J. H., \& Miyake, A. (2017). Challenges to EgoDepletion Research Go Beyond the Replication Crisis: A Need for Tackling the Conceptual Crisis. Frontiers in Psychology, 8, 568.
LynchJr., J., Netemeyer, R., Spiller, S., \& Zammit, A. (2009). A generalizable scale of propensity to plan: The long and the short of planning for time and for money. Journal of Consumer Research, 37, 108-128.
Miotto, E. C., \& Morris, R. G. (1998). Virtual planning in patients with frontal lobe lesions. Cortex, 34, 639-657.
Morris, R., \& Ward, G. (2004). The cognitive psychology of planning New York: Psychology Press.
Muraven, M., \& Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? Psychological Bulletin, 126, 247259.

Patalano, A. L., \& Seifert, C. M. (1997). Opportunistic planning: Being reminded of pending goals. Cognitive Psychology, 34, 1-36.
Patton, J. H., \& Stanford, M. S. (1995). Factor structure of the barratt impulsiveness scale. Journal of Clinical Psychology, 51, 768-774.
Schmeichel, B. J. (2007). Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. Journal of Experimental Psychology: General, 136, 241-255.
Schmeichel, B. J., Vohs, K. D., \& Baumeister, R. F. (2003). Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing. Journal of Personality and Social Psychology, 85, 33-46.
Simons, D., \& Galotti, K. (1992). Everyday planning: An analysis of daily time management. Bulletin of the Psychonomic Society, 30, 61-64.
Vallée-Tourangeau, F. (2013). Interactivity, efficiency, and individual differences in mental arithmetic. Experimental Psychology, 60, 302-311.
Vallée-Tourangeau, G., Abadie, M., \& Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. Journal of Experimental Psychology: General, 144, 581-603.
Vallée-Tourangeau, F., Sirota, M., \& Vallée-Tourangeau, G. (2016). Interactivity mitigates the impact of working memory depletion on mental arithmetic performance. Cognitive Research: Principles and Implications, 1:26.

# Head and Heart Metaphors for Moral Decision Making: Conceptual or Communicative? 

Rose K. Hendricks (rhendricks@ucsd.edu)<br>Department of Cognitive Science, 9500 Gilman Dr. La Jolla, CA 92093 USA<br>Paul H. Thibodeau (paul.thibodeau@oberlin.edu)<br>Department of Psychology, 120 W. Lorain St. Oberlin, OH 44074 USA


#### Abstract

When faced with a moral dilemma, following your head versus your heart can result in very different decisions. Earlier work has argued that people who "self-locate" in the head tend to make more rational and less emotional decisions to moral dilemmas than those who "self-locate" in the heart. We replicate this finding, suggest an alternative interpretation of the result, and then extend it with a novel experiment. In a metaphor framing task, we manipulated the salience of the head/heart metaphors-by using them (a) in a single sentence, (b) a more elaborate paragraph, or (c) by emphasizing one in contrast to the other. We found that people who received the head metaphor made more rational decisions than those who received the heart metaphor, but only in the high salience condition that contrasted the two metaphors. This finding illustrates the communicative value of metaphor, which can be enhanced through comparison.


Keywords: metaphor; decision making; rationality; emotion

## Introduction

In the novel and movie Sophie's Choice, a Polish woman, Sophie Zawistowska, is arrested by the Nazis and sent to the Auschwitz death camp. On arrival, she is "honored" for not being a Jew by being allowed a choice: One of her children will be spared the gas chamber if she chooses which one should be killed. If she does not choose, both of them will be killed.

Many moral dilemmas, like Sophie's, can be construed as a contrast between two extremes, involving a rational, utilitarian option (choose one child to die so that only one life is lost) and an emotional option (forgo choosing; both children die, but you did not play a direct role in either death). Why do some people decide to use their head to make the rational choice, while others follow their heart in choosing the emotional option?

One possibility for why some people make more rational decisions than others appeals to a role for conceptual metaphor (Fetterman \& Robinson, 2013; Lakoff \& Johnson, 1980). In English, the "head" is associated with cold, rational decision making. We use instructions like "use your head" to encourage emotional detachment in favor of carefully deliberated judgment. The "heart," on the other hand, is associated with hotter, more emotional thinking. Telling someone to "follow their heart" often implies that
they should ignore a cost-benefit calculus in favor of a more impulsive decision.

Recent work has argued that these "head" and "heart" metaphors (or metonymies) do more than describe different modes of thinking. They may also represent different ways of thinking about the self: some people "self-locate" in their head; others "self-locate" in their heart (Fetterman \& Robinson, 2013). On this view, people who conceptualize thinking as a process that happens in their head will tend to make more rational decisions, while people who conceptualize thinking as something that happens in their heart will tend to make more emotional decisions.

Evidence for this theory comes from a series of studies in which people were asked: "Irrespective of what you know about biology, which body part [the head or heart] do you more closely associate with your self?" Then participants completed personality measures, general knowledge questions, or they answered a series of moral dilemmas. Fetterman and Robinson (2013) found roughly a $50-50$ split in how people identified with the head versus the heart, which, in turn predicted responses to the other measures: head-locators characterized themselves as more rational and interpersonally cold on the personality measures, answered more of the general knowledge questions correctly, and suggested more utilitarian responses to the moral dilemmas, compared to the heart-locators.

Given the study design, however, it is difficult to know whether people really self-locate in the head or heart, and whether individual differences in self-location tendencies predict behavior. That is, an alternative interpretation of the finding is that people have some sense of their typical cognitive style-whether they tend to base their decisions on more rational or emotional motivations-which is what people report for the self-location question. On this view, one might expect the same results if participants had been asked if they consider themselves to be more rational or emotional decision makers (as opposed to a question about self-location). In addition, how people respond to the selflocation question may influence their performance on subsequent measures. People who say that they self-locate in their "head" may be inclined to demonstrate their headiness by adopting a more rational strategy to the moral dilemmas, for example.

These concerns relate to long-standing questions about what can be inferred about mental representation from patterns of language use (Keysar \& Bly, 1995; McGlone, 2011; Murphy, 1996). For example, when someone says, "I followed my heart" are they really imagining that their decision was made in their heart? Or is this phrase merely a conventional expression that has come to mean something like "I made the emotional choice"?

In the current paper, we explore these concerns, and address novel theoretical questions about metaphor framing, by manipulating the salience of instructions to "use one's head" or "follow one's heart" in moral decision making. There were three conditions in the experiment. In the lowsalience condition, the phrase "use your head" or "follow your heart" was embedded in the instructions of the taskwhich involved responding to the five moral dilemmas that were used by Fetterman and Robinson (2013). In the medium salience condition, participants were presented with a discussion about Plato's theory of the self, which was said to emphasize the head or the heart; the given metaphor was repeated in different ways throughout a paragraph that preceded the moral dilemmas. In the high salience condition, participants received the same information as those in the medium salience condition, with an additional explicit contrast: they were told either that the "head and not the heart" or the "heart and not the head" is where the self is located.

We expected that the high-salience condition would elicit the strongest effect: with the emphasis on head-location, in explicit contrast to heart-location, leading to more rational responding (and vice versa). An explicit comparison between the two metaphors should highlight the underlying difference between a rational and emotional approach to the moral dilemmas (Edwards, Williams, Gentner, \& Lombrozo, 2014; Markman \& Gentner, 1996).

This result would support an alternative interpretation of Fetterman and Robinson (2013)'s work. First, it would illustrate that, at least in some circumstances, more salient metaphors are more influential. In the original study, the metaphors were highly salient, since they were explicitly contrasted with one another in a forced choice task. Second, it would suggest that these particular metaphors are informative because of their conventional, idiomatic meaning, rather than their role in the mental representation of self-location (Keysar, Shen, Glucksberg, \& Horton, 2000; Thibodeau \& Durgin, 2008).

This result would also represent a novel contribution to the metaphor framing literature, which has found that linguistic metaphorical frames can shape how people think about issues like immigration (Landau, Sullivan, \& Greenberg, 2009; Jia \& Smith, 2013), cancer (Hauser \& Schwarz, 2013; Hendricks \& Boroditsky, 2015), and crime (Thibodeau \& Boroditsky, 2011, 2013). For instance, metaphorically framing crime as a "virus" has been found to increase support for societal reform as a means of crimereduction, whereas a "beast" frame leads people to support more enforcement-oriented approaches to crime-reduction
(Thibodeau \& Boroditsky, 2011, 2013). The current work extends these findings by investigating whether an explicit contrast, designed to make the underlying entailments of the metaphor more salient, leads to stronger metaphor framing effects.

Before conducting the experiment, we first replicated the original study (Fetterman \& Robinson, 2013). We present the results of the replication, which confirm the original findings, and then discuss the results of our follow-up experiment.

## Methods

## Participants

500 and 1,000 people were recruited from Amazon's Mechanical Turk to participate in Studies 1 and 2, respectively. Data was excluded from participants who did not submit a correct completion code and from participants who answered more than 3 (of 5) attention check questions incorrectly, leaving data from 484 and 945 participants for analysis in Study 1 and Study 2, respectively. Participants who completed Study 1 were not eligible to participate in Study 2.

## Procedure

Study 1 Study 1 was a replication of Fetterman \& Robinson (2013, Study 5). At the beginning of the study, participants were asked: "Irrespective of what you know about biology, which body part do you more closely associate with your self?" They were required to choose either the head or the heart. Next participants considered five moral dilemmas, similar to and including the Sophie's Choice example from the introduction. Each dilemma had one rational response and one emotional response (see Appendix). After the five dilemmas they answered an attention check question about each dilemma, and finally completed the Big Five Inventory (BFI; John \& Srivastava, 1999). The BFI measures individuals' extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. In the original studies, Fetterman and Robinson (2013) included the measure of conscientiousness as a predictor of how people responded to the moral dilemmas. For consistency, we also include conscientiousness in the analyses below, although the results do not differ if this measure is excluded.

Study 2 The procedure for Study 2 was identical to Study 1, except that instead of choosing the body part that they most associate with the self, participants randomly received one of the two metaphorical frames at one of three salience levels.

In the low salience condition, the metaphor was instantiated only in the instructions for responding to the moral dilemmas:

Next you will read short scenarios and should report what you would do if you were in them. There are no right or wrong answers to the
questions. Just [follow your heart/use your head] to make the judgment that you think is right. Please read each carefully because you will be asked to answer other questions about them later.

The medium and high salience conditions included passages that instantiated the metaphors more explicitly than the low salience condition. The beginning of the medium and high salience passages were identical:

Plato said that there are three parts of the soul. The first is our appetites or desires; the second is hotblooded emotion; and the final is rational, conscious awareness. But these three parts of ourselves do not play equal roles in making us who we are. The [head/heart] is the most crucial for defining who we are. The [head/heart] is where we find our true self.

The medium salience passage continued:
If we are to live a long and prosperous life, we must always listen to our [head/heart]. George Washington, Abe Lincoln, and Michelle Obama are just a few of the incredibly successful people who have followed Plato's advice in never losing sight of the fact that their [head/heart] holds the key to who they truly are.

In the high salience condition, the emphasized metaphor was explicitly contrasted with the alternative. Thus, the high salience passage instead said:

If we are to live a long and prosperous life, we must always listen to our [head/heart, even if it conflicts with our heart/head]. George Washington, Abe Lincoln, and Michelle Obama are just a few of the incredibly successful people who have followed Plato's advice in never losing sight of the fact that their [head/heart] holds the key to who they truly are, even if it means disregarding what their [heart/head] tells them.

After the passages, all participants responded to the same five dilemmas used in Study 1 and answered the same attention check questions. They were then asked whether they remembered encountering the phrase "follow your heart" or "use your head" earlier in the experiment. This recognition memory question was included as a test of the salience manipulation. As expected, participants were more likely to remember the metaphor in the more salient conditions, $B=.75, S E=.10, p<.001$.

Analysis Mixed effect logistic regression models were used to analyze the data from both studies (Jaeger, 2008). Metaphor (head versus heart) was treated as a betweensubjects fixed effect in both studies; salience (low, medium, high) was treated as a between-subjects fixed effect for the analysis of Study 2; participants and moral dilemmas were treated as random effects in both studies. We compare
nested models and present standardized regression coefficients to conduct hypothesis tests (Menard, 2002).

## Results

## Study 1: Replication

In Study 1, more people identified with the head metaphor ( $63 \%$ ) than the heart metaphor ( $37 \%$ ), $\chi^{2}(1)=32.80, \mathrm{p}<$ .001. We tested whether participants' choice of metaphor predicted how they solved the moral dilemmas by comparing two nested models. In the first, conscientiousness was included as a predictor of participants' judgments; in the second model, participants' chosen metaphor was added, which significantly improved fit, $\chi^{2}(1)=19.84, p<.001$. People who identified with the heart metaphor solved the moral dilemmas more emotionally ( $M=.52, S D=.25$ ) than people who identified with the head metaphor $(M=.41, S D=.29), B=.68, S E=$ $.15, p<.001$, as did more conscientious participants, $B=$ $.28, S E=.11, p=.012$. These findings replicate the basic patterns reported by Fetterman \& Robinson (2013).

We also conducted analyses by item to test whether particular dilemmas were driving the effect. We found an effect of metaphor for dilemmas that elicited more ambivalent responses overall. That is, there was a stronger consensus among participants on how to respond to dilemmas 1 (rationally) and 3 (emotionally); there was no effect of participants' choice of metaphor on these dilemmas, $p \mathrm{~s}>.3$. There was less consensus among participants on how to respond to dilemmas 2, 4, and 5; these dilemmas showed differences as a function of which metaphor people chose, $p \mathrm{~s}<.001$ (see Table 1).

Table 1. Proportion of emotional responses overall, for head-locators, and for heart-locators by dilemma.

| Dilemma | Overall | Heart | Head |
| :---: | :---: | :---: | :---: |
| 1 | .70 | .73 | .69 |
| 2 | .36 | .49 | .29 |
| 3 | .14 | .13 | .14 |
| 4 | .50 | .61 | .44 |
| 5 | .57 | .66 | .51 |

In other words, head-locators did not simply choose the rational response to each dilemma (and vice versa for heartlocators). They were also sensitive to the content of the dilemmas. For this reason, we focus on responses to dilemmas 2,4 , and 5 in the experiment.

## Study 2: Metaphor Framing

We tested whether the metaphor used to describe the task and the salience of the metaphor affected participants' moral judgments. We focus on data from dilemmas 2, 4, and 5, since these dilemmas elicited more ambivalent responses overall, and were influenced by participants' choice of metaphor in Study 1.

The analysis revealed no main effect of metaphor, $\chi^{2}(1)=$ 1.73, $p=.189$, or salience condition, $\chi^{2}(1)=0.53, p=.467$. But it did reveal an interaction between metaphor and salience condition, $\chi^{2}(1)=4.34, p=.037$, as well as an effect of conscientiousness, $\chi^{2}(1)=9.52, p=.002$.

As shown in Figure 1, there was no effect of the metaphor frame in low, $B=.01, S E=.19, p=.964$, or medium, $B=$ $.08, S E=.21, p=.681$, salience conditions. There was an effect of the metaphor in the high-salience condition, $B=$ $.70, S E=.27, p=.011$. When the instructions emphasized the "heart" in explicit contrast to the "head," people responded to the dilemmas more emotionally (and vice versa).


Figure 1. Proportion of dilemmas solved emotionally in Study 1 (choice) and in Study 2 by metaphor and salience conditions. Error bars denote standard errors of the means.

The effect of the salience manipulation appeared to be fairly linear for the "heart" condition-with people responding more emotionally as the salience of the "heart" metaphor increased, $B=.21, S E=.10, p=.030$. The effect of the salience manipulation seems to have been more abrupt in the "head" condition. There was no difference in how participants responded to the low- and medium-salient versions of the instructions that emphasized the "head" metaphor, $p=.311$; participants responded marginally more rationally to the high-salient version, compared to mediumsalient version, of the instructions that emphasized the "head" metaphor, $p=.057$.

Figure 2 illustrates the effect of metaphor preference (Study 1) or metaphor frame (from the high-salience condition of Study 2) on each of the moral dilemmas. It shows that the metaphor people identified with in Study 1 had the biggest effect on judgments of the 2nd, 4th, and 5th moral dilemmas. These were the same dilemmas that were most influenced by the salient metaphor frames in Study 2.


Figure 2. Effect size by item (moral dilemmas 1-5) for the metaphor preference task in Study 1 and bin the highsalience condition of Study 2. The further the bar extends to the right (from 0 ), the more congruent the responses (i.e. heart and emotional responding; head and rational responding). Bars extending to the left (of 0 ) indicate a pattern of incongruent responding (i.e. heart and rational; head and emotional).

## Discussion

In this work, we first replicated prior work by Fetterman \& Robinson (2013) showing that people who identified with a heart metaphor for the self responded more emotionally to moral dilemmas, while people who identified with a head metaphor for the self responded more rationally to moral dilemmas. The original finding was interpreted as evidence for an individual difference in self-location grounded in conceptual metaphor. However, we have argued that there are alternative interpretations of the finding. Most notably, the heart and head metaphors are conventional expressions that correspond to emotional and rational modes of thinking, respectively. People have some self-awareness about how they make decisions-more rationally or more emotionally. When asked to choose between identifying with the heart or head, emotional decision makers choose the heart, while rational decision makers choose the head.

In a follow-up experiment, we examined whether metaphorically framing the locus of a person's decisions as either in the head or in the heart would lead them to make more rational decisions (in the case of the head) or emotional decisions (in the case of the heart). We also explored the role of salience in this process: using the metaphors in a single phrase (low salience), a more elaborate paragraph (medium salience), or by emphasizing one in direct contrast with the other (high salience).

We found an effect of the metaphor framing manipulation in the high-salience condition but not the low or medium salience conditions, suggesting that an explicit contrast between the metaphors was important for influencing behavior on the decision making task. In the high salience condition, since people were exposed to both metaphors, they had the opportunity to compare the two metaphors. In fact, in order to truly comprehend the passage, they needed to compare their passage's dominant metaphor to the alternative. In Fetterman \& Robinson's (2013) work and in our Study 1, choosing the locus of the self also encourages, and perhaps even requires, participants to explicitly compare the two metaphors' entailments in order to choose the one they believe describes them most accurately. Comparison has been found to be particularly effective in communicating the intended meaning of analogies (Edwards, Williams, Gentner, \& Lombrozo, 2014; Markman \& Gentner, 1996). This work suggests that explicitly comparing metaphor frames to each other may similarly highlight their differences and amplify their effects on cognition. In other words, we found that the head and heart metaphors used in this work were both conceptual and communicative.

All participants in these experiments were in the United States, so the implications about heart and head metaphors for decision making may not generalize to members of other cultures. It may be productive for future research to investigate interactions between cultural background and metaphor frames for decisions.

In addition, this work may have implications for the development of Deliberate Metaphor Theory (DMT; Steen, 2008), which argues that metaphors are most influential when they are used deliberately. That is, DMT emphasizes the social and pragmatic context in which figurative language is used, although the details of the theory (e.g., what constitutes a deliberate metaphor?) have yet to be ironed out, and empirical tests of the predictions made by the theory have received limited support (see, e.g., Gibbs, 2015a, 2015b; Thibodeau, In press). Thus, the current work may give researchers a novel case for thinking about one pragmatic signal-explicitly negating one metaphor in favor of another-that a metaphor has been used deliberately. Explicitly contrasting metaphors clearly signals deliberate use.

To advance Deliberate Metaphor Theory, it would be worthwhile to try and provide a more mechanistic account of the effect we have demonstrated. For instance, one might argue that the metaphors were used "deliberately" in all three salience conditions of the experiment. But we only found an effect when the two metaphors were contrasted with one another. An open question, therefore, is whether the contrast served as more of a pragmatic cue for participants to use the emphasized metaphor, or whether the contrast served to bring out the meaning of the head and heart metaphors more clearly-by highlighting an alignable difference between and the underlying meaning of the phrases (Gentner \& Markman, 1994).

Future work should also explore the role of comparison in metaphor processing more generally. Experiments that examine metaphor framing-for persuasion, instruction, and explanation-typically present participants with only one frame (e.g., Jia \& Smith, 2013; Landau, Sullivan \& Greenberg, 2009; Thibodeau \& Boroditsky, 2011, 2013). The current work suggests that explicitly contrasting metaphors may facilitate metaphorical reasoning.

## Acknowledgments

We thank Adam Fetterman for sharing the stimuli from the work we replicated (Fetterman \& Robinson, 2013). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650112.

## References

Edwards, B. J., Williams, J. J., Gentner, D., \& Lombrozo, T. (2014). Effects of comparison and explanation on analogical transfer. In P. Bello, M. Guarini, M. McShane, \& B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 445450). Austin, TX: Cognitive Science Society.

Fetterman, A. K., \& Robinson, M. D. (2013). Do you use your head or follow your heart? Self-location predicts personality, emotion, decision making, and performance. Journal of Personality and Social Psychology, 105(2), 316-334.
Gentner, D., \& Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. Psychological science, 5(3), 152-158.
Gibbs, R. W. (2015a). Does deliberate metaphor theory have a future? Journal of Pragmatics, 90, 73-76.
Gibbs, R. W. (2015b). Do pragmatic signals affect conventional metaphor understanding? A failed test of Deliberate Metaphor Theory. Journal of Pragmatics, 90, 77-87.
Hauser, D. J., \& Schwarz, N. (2014). The War on Prevention: Bellicose Cancer Metaphors Hurt (Some) Prevention Intentions. Personality and Social Psychology Bulletin.
Hendricks, R.K. \& Boroditsky, L. (2016). Metaphor \& emotion: Metaphorical frames for coping with hardship. In Papafragou, A., Grodner, D., Mirman, D., \& Trueswell, J.C. (Eds.) (2016). Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Jaeger, T.F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. Journal of Memory and Language, 59(4), 434-446.
Jia, L., \& Smith, E. R. (2013). Distance makes the metaphor grow stronger: A psychological distance model of metaphor use. Journal of Experimental Social Psychology, 49(3), 492-497.
John, O. P., \& Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and
theoretical perspectives. In L. A. Pervin \& O. P. John (Eds.), Handbook of personality: Theory
and research (Vol. 2, pp. 102-138). New York: Guilford Press.
Keysar, B. \& Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans?
Journal of Memory and Language, 34, 89-109.
Keysar, B., Shen, Y., Glucksberg, S., \& Horton, W. S. (2000). Conventional language: How metaphorical is it?. Journal of Memory and Language, 43(4), 576-593.
Lakoff, G. \& Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.
Landau, M. J., Sullivan, D., \& Greenberg, J. (2009). Evidence That Self-Relevant Motives and Metaphoric Framing Interact to Influence Political and Social Attitudes. Psychological Science, 20(11), 1421-1427.
Markman, A. B., \& Gentner, D. (1996). Commonalities and differences in similarity comparisons. Memory \& Cognition, 24(2), 235-249.
McGlone, M.S. (2011). Hyperbole, homunculi, and the hindsight bias: An alternative evaluation of conceptual metaphor theory. Discourse Processes, 48(8), 563-574.
Menard, S. (2002). Applied Logistic Regression Analysis Second Edition. Thousand Oaks, CA: Sage Publications, Inc. Murphy, G.L. (1996). On metaphoric representation. Cognition, 60, 173-204.
Steen, G. (2008). The paradox of metaphor: Why we need a three-dimensional model of metaphor. Metaphor and Symbol, 23(4), 213-241.
Thibodeau, P. H. (In press). The Function of Metaphor Framing, Deliberate or Otherwise, in a Social World. Metaphor and the Social World.
Thibodeau, P. H., \& Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. PLoS One, 6(2), e 16782.
Thibodeau, P. H., \& Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. PloS One, 8(1), e52961.
Thibodeau, P., \& Durgin, F. H. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. Journal of Memory and Language, 58(2), 521540.

## Appendix: Moral Dilemmas

1.You are an inmate in a concentration camp. A sadistic guard is about to hang your son who tried to escape and wants you to pull the chair from underneath him.

He says that if you don't he will not only kill your son but some other innocent inmate as well. You don't have any doubt that he means what he says. What would you do?

Rational = I would pull the chair
Emotional = I would NOT pull the chair
2.A pregnant woman leading a group of people out of a cave on a coast is stuck in the mouth of that cave. In a short time
high tide will be upon them, and unless she is unstuck, they will all be drowned except the woman, whose head is out of the cave. Fortunately, (or unfortunately,) someone has with him a stick of dynamite.

There seems no way to get the pregnant woman loose without using the dynamite which will inevitably kill her; but if they do not use it everyone will drown. What would you do if you were in this situation?

Rational = I would let them light the stick of dynamite
Emotional = I would NOT let them light the stick of dynamite
3.A trolley is running out of control down a track. In its path are five people who have been tied to the track by a mad philosopher. Fortunately, you could flip a switch, which will lead the trolley down a different track to safety. Unfortunately, there is a single person tied to that track. Would you flip the switch or do nothing?

Rational $=$ I would flip the switch
Emotional = I would do nothing
4. In the novel and movie Sophie's Choice, a Polish woman, Sophie Zawistowska, is arrested by the Nazis and sent to the Auschwitz death camp. On arrival, she is "honored" for not being a Jew by being allowed a choice: One of her children will be spared the gas chamber if she chooses which one should be killed. If she does not choose, both of them will be killed. Would you choose one of your children to be killed in the same situation?

Rational = I would choose a child to be killed
Emotional = I would NOT choose a child to be killed
5. In 1842 , a 23 ship struck an iceberg and more than 30 survivors were crowded into a lifeboat intended to hold 7 . As a storm threatened, it became obvious that the lifeboat would have to be lightened if anyone were to survive. The captain reasoned that the right thing to do in this situation was to force some individuals to go over the side and drown or everyone would drown. Would you support pushing some people off the boat so at least some people could survive?

Rational = I would support pushing people off the boat
Emotional $=$ I would NOT support pushing people of the boat

# When metaphors in the mind become metaphors in the mouth: Documenting the emergence of a new system of linguistic metaphors for time 

Rose K. Hendricks (rhendricks@ucsd.edu)<br>Benjamin K. Bergen (bkbergen@ucsd.edu)<br>Cognitive Science, UC San Diego, La Jolla, CA 92093 USA<br>Tyler Marghetis (tmarghet@indiana.com)<br>Psychological and Brain Sciences, Indiana University, Bloomington


#### Abstract

Languages exhibit striking semantic diversity, but different languages often share core metaphors. Conceptual Metaphor Theory (Lakoff \& Johnson, 1980) claims that universal human experiences give rise to conceptual representations that are then expressed in language. But languages change slowly, making it difficult to observe implicit conceptualization affecting linguistic convention in real time. Here, we describe a shared conceptualization previously absent from speech that has now become conventionalized in linguistic metaphors. In two studies, we document how members of the US military talk about time using conventionalized lateral metaphors (e.g., 'push the meeting right' to mean 'move the meeting later'). We show that military members, unlike civilians, consider such sentences to be acceptable-sometimes even more acceptable than more standard phrases. Moreover, military personnel seem unaware that these lateral metaphors are specific to their linguistic sub-community. Our findings suggest that implicit mental representations can become conventionalized metaphors in language.


Keywords: time; metaphor; linguistic convention; semantic change

## Introduction

Every language uses metaphor (Kövecses, 2005), and many of these metaphors appear universal or nearly universal. Core metaphors like AFFECTION is WARMTH occur in many languages, while the reverse, AFFECTION is COLD, does not (Kövecses, 2010). Why? The leading explanation is that these cross-linguistic regularities reflect universal, or nearly universal, human experiences (e.g., Clark, 1973; Lakoff \& Johnson, 1980). For instance, since being held by a caregiver is likely a nexus of warmth and affection, human infants learn to associate these experiences. On this account, recurring correlations in experience lead to mental representations that relate the two domains. And these representations, in turn, spill out into language, so that the
same experiences come to be described similarly across the world's languages.
However, this is not the only way in which mental representations and linguistic metaphors could have come into alignment. For instance, the causal direction may have been reversed: widespread, recurring patterns of thought may have arisen from exposure to linguistic metaphors, rather than the other way around (see Gibbs, 2011 for review). Indeed, there's substantial evidence that, at least in the short term, exposure to linguistic metaphors primes conceptualization (e.g., Thibodeau \& Boroditsky, 2011).
In fact, there is a surprising dearth of evidence that shared mental representations can give rise to novel linguistic conventions. While the historical record is replete with changes in linguistic semantics that appear, in retrospect, to be driven by conceptualization (Sweetser, 1991), we cannot assess the conceptual representations of historical people in a lab. If implicit patterns of thought give rise to linguistic conventions, we should be able to observe the emergence of new linguistic metaphors in real time. But the pervasiveness of core, cross-linguistic metaphors means that it's rare to find people who lack either the linguistic or conceptual manifestation of an otherwise universal metaphor.
To determine whether metaphor spreads from thought to language, we need to observe a linguistic community beginning to use metaphorical language that aligns with their prior conceptual representations. This would shed light on why so many languages share linguistic and conceptual metaphors. We believe we have found one such case.

## The coupling of language and thought about time

Of all the concepts that we understand metaphorically, the best studied is time. Within a given culture or community, individuals think and talk about time in ways that are both stable and shared. This often involves using space to structure their speech and understanding of time (for reviews, see Boroditsky, 2011; Núñez \& Cooperrider, 2013).

Specifically, English speakers think and speak about time as though it were represented along the sagittal (front-back)
axis. The future is 'ahead' and the past is 'behind,' and these manners of speaking align with physical behavior. When English speakers make decisions about events, they are faster to respond to future events by moving forward, and faster to respond to past events by moving backward (Sell \& Kaschak, 2011; Ulrich et al., 2012; Rinaldi et al., 2016). They are faster to make time judgments when future-related words are shown in front of an image of a person and past words behind (Torralbo et al., 2006). When imagining the future, people lean forward, and when thinking about the past, they lean back (Miles et al., 2010). And they gesture forward when talking about the future, but backwards when talking about the past (Casasanto \& Jasmin, 2012).

This use of the sagittal axis has been documented in many languages around the world (Núñez \& Cooperrider, 2013). Most follow the same pattern as English, but not all: In Aymara, past events are in front and future events behind, in both language and gesture (Núñez \& Sweetser, 2006).

Similarly, Mandarin Chinese uses vertical (up/down) terms systematically and productively to talk about time. Earlier events are $u p$ and later events are down. Native Mandarin speakers also think about time vertically, with earlier events above later ones (Boroditsky et al., 2010; Fuhrman et al., 2011; Miles et al., 2011; Yang \& Sun, 2015).

Thus, there is often a tight coupling of the spatial language used to talk about time and spatial thought used to conceptualize time. There is empirical evidence that this alignment can originate in linguistic metaphors (Hendricks \& Boroditsky, 2015). After English-speaking participants learned metaphors that placed earlier events either above or below later ones (i.e., breakfast is above dinner or breakfast is below dinner; Hendricks \& Boroditsky, 2015), they then exhibited metaphor-consistent responses on an implicit measure of their mental space-time associations. Language can be a formative force for mental representations.
Despite overlaps between temporal language and temporal thought, there is not a one-to-one correspondence between the two. In addition to their sagittal (front-back) mental time-line, English speakers also map time to the lateral (left-right) axis. For instance, when asked to arrange physical depictions of sequences of events, they arrange them from left to right (Tversky, Kugelmass \& Winter, 1991; Boroditsky \& Gaby, 2010). During natural speech, they gesture to the left for earlier events and to the right for later ones (Cooperrider \& Núñez, 2009; Casasanto \& Jasmin, 2012). And English speakers are faster to indicate that one event occurred earlier than another by responding on their left side, and faster for later events when responding on their right (Fuhrman \& Boroditsky, 2010; Miles et al., 2011; Weger \& Pratt, 2008; Walker et al., 2014).

However, these left-right mental timelines are absent from language. English speakers can look 'back' on the past, but never 'to the left.' There is a linguistic gap, therefore, where a widespread conceptual metaphor has not surfaced as a linguistic metaphor. If a cognitive representation were to leap to a linguistic representation, it would most likely first take hold in a sub-community of English speakers who use left-right linguistic metaphors to talk about time.

## Case Study: The US Military

Anecdotal evidence from members of the authors' own families suggests that one community of English speakers has started to use left-right metaphors when talking about time: members of the United States military. According to anecdote, one might propose to move a meeting to a later time by asking to 'move the meeting to the right.' If these reports were true, then this group would provide an opportunity to study a system of conceptual metaphors as it takes linguistic hold in a language community. Here, in two empirical studies, we sought to document this apparent linguistic innovation. In particular, we aimed to understand whether left-right linguistic metaphors are more acceptable to members of the military than to civilians, as well as the linguistic conventions association with these metaphors.
In Study 1, military and civilian participants rated the acceptability of sentences about time. The sentences were presented in four main conditions: Standard (The meeting was moved two days earlier, from Friday to Wednesday), Dynamic-Lateral (The meeting was moved two days to the left from Friday to Wednesday), Static-Lateral (The meeting on Wednesday is two days to the left of the meeting on Friday), and Ungrammatical (From the meeting was two earlier days, Friday to Wednesday pushed). We included lateral metaphors in two conditions (Dynamic and Static) because anecdotal evidence suggested that Dynamic uses may be more acceptable than Static.
Study 2 replicated the findings from Study 1 in a new sample of military and civilian participants. In addition, we investigated whether military personnel are aware that these metaphors are specific to their linguistic subcommunity. Together, these studies create a snapshot of the early stages of a shift from an exclusively cognitive representation to a novel linguistic convention.

## Study 1

In Study 1, participants rated the acceptability of sentences. Features of the sentences allowed us to measure whether, when, and to whom lateral (left-right) linguistic metaphors are acceptable.

## Methods

Participants: Active Duty members of the US military ( $n=$ 23) participated for $\$ 10$, and civilian undergraduates at UC San Diego ( $n=31$ ) participated for course credit. The military participants included 4 Army, 1 Navy, and 18 Air Force. They included 8 Officers and 15 Enlisted members. Materials: There were four types of sentences: a) Standard: using earlier or later to reschedule; b) Dynamic-Lateral: using left and right to reschedule; c) Static-Lateral: using left and right, but no rescheduling; d) Ungrammatical.
An equal number of sentences referred to events that were a) earlier and later; b) on the timescale of hours, days, and months; and c) required crossing a temporal boundary (events take place in different days, weeks, or years; i.e., Friday to Monday) and were within a temporal boundary (i.e., Tuesday to Friday). These additional features allowed us to examine whether conventionality differed in these various contexts.
Procedure: The study was completed on a computer. Participants were instructed to imagine a new colleague whose native language was not English and rate the acceptability of sentences $(n=48)$ uttered by this colleague, based on how participants would normally talk at work. Acceptability ratings used a 7 point Likert scale ( $1=$ totally unacceptable, $7=$ totally acceptable). Each sentence was presented on its own page. Participants then supplied standard demographic information (education, age), and military participants reported their service branch (Army, Navy, etc.), rank, and years of service. No other measures were collected.
Exclusions and Analyses: We subtracted each person's mean Ungrammatical rating from their mean Standard rating. Three participants ( 1 military, 2 civilians) did not rate Standard sentences at least one point higher than Ungrammatical ones, and were eliminated. Ratings were standardized by participant (i.e., z-scored), and then analyzed in a linear mixed-effects model. All hierarchical models used the maximal converging random effects structure justified by the experimental design (Barr et al, 2013), with random intercepts and slopes for both participants and items.

## Results

Figure 1 shows military personnel and civilians' acceptability ratings for the four sentence types. We first verified that participants from both populations rated the Standard phrases as most acceptable and the Ungrammatical phrases the least acceptable, with the Lateral phrases in between. Ratings were analyzed with a mixed-effects model with a fixed effect of Sentence Type (Standard > Dynamic-Lateral $>$ Static-Lateral $>$ Ungrammatical); we
used forward difference coding to test for pairwise differences between consecutive levels. Ratings did not differ by timescale (i.e., hours, days, and months; $p=.80$ ), so we collapsed timescales for all subsequent analyses. Standard items were rated as more acceptable than Dynamic-Lateral phrases ( $b=0.75 \pm 0.14$ SEM, $t=5.5, p<$ .001), which were more acceptable than Static-Lateral phrases ( $b=0.44 \pm 0.10 \mathrm{SEM}, t=4.6, p<.001$ ), which in turn were more acceptable than Ungrammatical phrases ( $b=$ $0.71 \pm 0.13 \mathrm{SEM}, t=5.3, p<.001)$.
We next tested our critical prediction: That these patterns of acceptability would differ systematically by population. We thus added a fixed effect of population (civilian $=1$, enlisted $=2$, officer $=3$; centered so the mean was 0 ). Standard phrases were used as a baseline. Once again, every type of sentence (Dynamic-Lateral, Static, Ungrammatical) was less acceptable than Standard (all $b \mathrm{~s}<-0.75, p<.001$ ). As predicted, there was no evidence that military and civilian participants differed in their ratings of Standard phrases ( $b=-0.07 \pm 0.08 \mathrm{SEM}, t=-0.8, p=.4$ ), nor was there evidence that the relative unacceptability of Ungrammatical phrases differed by population ( $b=-0.14 \pm$ 0.12 SEM, $t=-1.3, p=.2$ ). Similarly, military and civilian participants did not differ in the acceptability of Static-Lateral phrases, compared to Standard phrases ( $b=$ $-0.14 \pm 0.12$ SEM, $t=-1.3, p=.2$ ).

Acceptance of Dynamic-Lateral phrases, however, differed across populations ( $b=0.32 \pm 0.14 \mathrm{SEM}, t=2.3, p$ $=.027$ ). To make sense of this effect, we zoomed in on Dynamic-Lateral items and used forward-difference coding to compare the three populations. Civilians gave the lowest rating to the Dynamic-Lateral items $(M=0.0)$, while enlisted military participants rated these items as more acceptable $(M=0.17, b=-0.22, p=.07)$, and military officers rated these items as even more acceptable than enlisted military personnel ( $M=0.51, b=-0.43, p<.001$ ).

## Discussion

Study 1 confirmed that a specific subset of lateral metaphors has become conventionalized among a subculture of American English speakers: members of the US Military. If this reflected a more general conceptual difference-perhaps a willingness among military personnel to think about time along a left-to-right timeline-then this should have been reflected in increased acceptability for all lateral expressions. Instead, military personnel were especially accepting of lateral metaphors that used the dynamic language of movement, suggesting that this is a genuine linguistic convention, subject to the quirks and idiosyncrasies of cultural norms.


Figure 1. In Study 1, compared to civilians, military personnel were more accepting of Dynamic-Lateral phrases (e.g., 'meeting was moved to the left'). There was no difference in civilians' and military members' acceptance of Static-Lateral phrases. Error bars $=$ SEM.

## Study 2

Study 2 was designed to replicate and further explore this linguistic conventionalization. In addition, we sought to determine whether military personnel are aware that lateral metaphors are specific to the dialect of English spoken in the military, and not shared with the larger civilian population. Anecdotal evidence suggested that military personnel might be unaware that this system of linguistic metaphors is not shared widely among English speakers.

## Methods

Participants: Members of the US military $(\mathrm{n}=14)$ participated for $\$ 10$. Civilian undergraduates ( $\mathrm{n}=27$ ) participated for course credit. Military participants included 3 Army, 1 Navy, and 10 Air Force; 5 were Officers, 5 were Enlisted, and 4 did not identify their rank.
Materials: Materials were identical to Study 1 with two differences. First, based on anecdotal evidence that lateral metaphors were commonly used with the verb push, we included items with push (e.g., pushed two months). Second, to reduce the total number of items, we did not vary the timescale, since it had no effect on acceptability in Study 1.
Procedure: Participants completed two randomly-ordered tasks: the Acceptability Rating task from Study 1, and a forced choice Sentence Completion task.

The Acceptability Rating task was based on Study 1, with one critical difference: participants completed two randomly-ordered blocks of acceptability ratings, one in which they were asked to imagine all their colleagues were
in the military, and another in which they imagined all of their colleagues were civilians. Manipulating the utterances’ context in this way allowed us to test whether military participants were sensitive to the community-specificity of the lateral-dynamic metaphors.

In the Sentence Completion task, participants read the same sentences as in the Acceptability Rating task, but with a blank in place of month (e.g., The meeting was moved two months to the right, from November to $\qquad$ .) Choices included all odd-numbered months (January, March, etc.) and I don't know. Because of space limitations, we must analyze these data elsewhere.

To refresh participants between these tasks, they completed a brief "spot the differences" game, in which they had 45 seconds to count as many small differences as possible between two nearly identical images.

Exclusions and Analyses were unchanged from Study 1. No participants were excluded.

## Results

As in Study 1, we first verified that participants from both populations rated the Standard phrases as most acceptable and the Ungrammatical phrases the least acceptable, with the Lateral phrases in between. Once again, Standard items were rated as more acceptable than Dynamic-Lateral phrases ( $M=0.91$ vs. $M=0.21 ; b=0.73 \pm 0.14 \mathrm{SEM}, t=$ 5.4, $p<.001$ ), which were more acceptable than Static-Lateral phrases $(M=-0.10 ; b=0.31 \pm 0.07 \mathrm{SEM}, t=$ 4.3, $p<.001$ ), which in turn were more acceptable than Ungrammatical phrases ( $M=-1.03 ; b=0.83 \pm 0.16$ SEM, $t$ $=5.3, p<.001$ ).

We next attempted to replicate our main finding from Study 1: That military participants had a selective preference for lateral metaphors, compared to civilians. First, we replicated the finding that, overall, every type of sentence (Dynamic-Lateral, Static-Lateral, Ungrammatical) was rated as less acceptable than the Standard phrases (all $b \mathrm{~s}<-0.69, p \mathrm{~s}<.001$ ). Next, we replicated our critical finding that the populations differed in their preference for Dynamic-Lateral phrases $(b=0.57 \pm 0.05$ SEM, $t=10.6, p$ $<.001$ ). While civilians thought the Dynamic-Lateral phrases were far worse than the standard ones $(b=-0.96$ $\pm 0.05 \mathrm{SEM}, t=-20.2, p<.001$ ), for enlisted military personnel the difference was more than half of what it was for civilians $(b=-0.47 \pm 0.08 \mathrm{SEM}, t=-5.9, p<.001)$, while for military officers the Dynamic-Lateral phrases were actually considered to be significantly better than the Standard ones $(b=0.18 \pm 0.06 \mathrm{SEM}, t=-3.0, p<.01)$.

There was no evidence that the populations differed in their judgements of Ungrammatical phrases ( $p>.7$ ). However, unlike in Study 1, the populations did differ in
their acceptance of Static-Lateral phrases ( $b=-0.43 \pm 0.05$ SEM, $t=-8.1, p<.001$ ). As with the Dynamic-Lateral phrases, civilians were the most dismissive ( $M=-0.28$ ), followed by enlisted personnel $(M=0.15)$, and finally officers ( $M=0.43$ ) - although note that, here, officers did not prefer Static-Lateral phrases to the Standard phrases.

Finally, we investigated whether an utterance's context whether it was uttered in a military or civilian context - had a significant effect on its acceptability (Figure 2). This full model replicated the findings reported above. There were only two other significant effects. The first was an interaction between context and population ( $b=0.18$ $\pm 0.06 \mathrm{SEM}, t=2.9, p<.001$ ), driven by an 'opposite context' effect: military participants thought phrases were generally less acceptable in a civilian context, and vice versa. This may be due to participant uncertainty about norms in unfamiliar environments. The second was a three-way interaction between population, context, and Static-Lateral (vs. Standard) phrases ( $b=-0.20 \pm 0.09$ SEM, $t=2.3, p=.02$ ). This was driven by the fact that officers did not exhibit the 'opposite context' effect for Static-Lateral phrases - they liked them equally in both contexts. Thus, even in a hypothetical civilian context, military participants thought the lateral phrases were significantly more acceptable than actual civilians did.


Figure 2. In Study 2, military personnel were again more accepting of lateral phrases, this time for both static and dynamic versions. This pattern was repeated both when the phrases were uttered in a civilian context (left panel) or a military context (right panel). In other words, military participants thought that rescheduling a meeting to the right' is acceptable among civilians. Error bars $=$ SEM.

## Discussion

Conventionalized linguistic structures and cultural patterns of thought are often consistent (e.g., Boroditsky et al., 2010; Fuhrman et al., 2011; Winter, Marghetis, and Matlock, 2015), and prior work has provided evidence that learning new metaphors for talking about time can create new mental representations (Hendricks \& Boroditsky, 2015). However, there is little direct evidence of the purportedly more pervasive, reverse relationship: cultural
patterns of thought creating new conventionalized linguistic structures. In two studies, we provided evidence that one subset of American English speakers-members of the US military-have adopted a conventionalized system of metaphors for expressing mental representations that are pervasive in the minds of English speakers in general, but otherwise absent from language. Specifically, we found that members of the US military, especially Officers, consider sentences containing left-right metaphors (and only these types of sentences) as more acceptable than civilians do.
We also explored the nuances of the left-right metaphors based on military members' acceptability ratings for sentences with different features. In general, military personnel judged Dynamic-Lateral sentences - which express an event moving from one time to another - to be more acceptable than similar Static-Lateral sentences that express the same relationship without movement. This nuance in military members' acceptance of left-right metaphors is evidence that these new left-right linguistic conventions do not merely reflect a broad association of time with the lateral axis, but instead reflect specific linguistic conventions.
By showing that at least one subculture of American English speakers has conventionalized left-right linguistic metaphors for time, we demonstrate that the relationship between conventionalized structures in language and patterns in thought is bidirectional: not only can language shape mental representations, but our mental representations can also make their way into conventionalized language.
Why have military members adopted lateral metaphors that are absent from civilians' language? One hint may lie in the artifacts they use. Duty Rosters are documents that keep track of the work assignments for each each member of a military unit. Duty Rosters are standardized and governed by instruction manuals (Army Regulation [AR] 220-45). Each row represents an individual. Each column represents a successive date, ordered from left to right. Each cell thus indicates the task that was assigned to that individual (row) on that day (column). Unlike a standard American calendar, in which each row only has 7 days across, Duty Rosters arrange days in a continuous line extending rightward, potentially endlessly. One duty roster, for instance, had hundreds of columns, each representing a successive day.
The current work cannot distinguish between two explanations for how this linguistic innovation has spread. On the one hand, new linguistic conventions could have emerged spontaneously through the interaction with frequently consulted artifacts like Duty Rosters in a relatively linguistically encapsulated community. On a complementary account, these conventions may be the product of top-down, institutional decrees, where linguistic decision-makers within a community-affected no doubt by
the same convergence of conceptualization and material artifacts-themselves influenced linguistic habits through prescriptions for ways to talk about time. Future work will continue to document the history and use of these left-right linguistic conventions to distinguish between these two accounts. Similarly important to explore are the cognitive consequences of adopting lateral metaphors for time. Can adopting lateral linguistic metaphors facilitate reasoning about temporal change? Does it reduce miscommunication (e.g., allowing speakers to avoid ambiguous descriptions like Wednesday's meeting was moved forward two days)? Or might it increase other kinds of miscommunication, for example when English speakers communicate with Hebrew and Arabic speakers, whose mental timelines run right to left, counter to English speakers' (Tversky et al., 1991)? In sum, this work reveals a potentially fertile new way to study the give and take between individual conceptual metaphors and community-wide, conventionalized linguistic structure.

## Acknowledgments

We thank Steven Billington for recruitment help. RH was supported in part by NSF GRFP Grant No. DGE-1650112.

## References

Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing. J. of Memory and Language, 68, 255-278.
Boroditsky, L. (2011). How Languages Construct Time. In Space, Time and Number in the Brain. Elsevier.
Boroditsky, L., Fuhrman, O., \& McCormick, K. (2010). Do English and Mandarin speakers think about time differently? Cognition, 118(1), 123-129.
Boroditsky, L., \& Gaby, A. (2010). Remembrances of Times East. Psychological Science, 21(11), 1635-1639.
Brysbaert, M. \& New, B. Moving. (2009). Moving beyond Kucera and Francis, Behav. Res. Meth, 41, 977-990.
Casasanto, D., \& Jasmin, K. (2012). The hands of time. Cognitive Linguistics, 23, 643-674.
Clark, H. H. (1973). Space, time, semantics and the child. In T. E. Moore (Ed.), Cognitive development and the acquisition of language. New York: Academic Press.
Cooperrider, K., \& Núñez, R. (2009). Across time, across the body. Gesture, 9, 181-206.
Fuhrman, O., McCormick, K., Chen, E., Jiang, H., Shu, D., Mao, S., \& Boroditsky, L. (2011). How Linguistic and Cultural Forces Shape Conceptions of Time. Cognitive Science, 35, 1305-1328.
Gibbs, R. W. (2011). Evaluating Conceptual Metaphor Theory. Discourse Processes, 48(8), 529-562.

Hendricks, R. K., \& Boroditsky, L. (2015). New space-time metaphors foster new mental representations of time. Proceedings of the 37th Annual Meeting of the Cognitive Science Society.
Kövecses, Z. (2005). Metaphor in culture: Universality and variation. Cambridge University Press.
Kövecses, Z. (2010). Metaphor, language, and culture. DELTA: Documentação de Estudos Em Lingüística Teórica E Aplicada, 26(SPE), 739-757.
Kucera, H. \& Francis, W.N. (1967). Computational analysis of present-day American English. Brown U. Press.
Lakoff, G. \& Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.
Miles, L. K., Nind, L. K., \& Macrae, C. N. (2010). Moving Through Time. Psychological Science, 21(2), 222-223.
Miles, L. K., Tan, L., Noble, G. D., Lumsden, J., \& Macrae, C. N. (2011). Can a mind have two time lines? Psychonomic Bulletin \& Review, 18, 598-604.
Núñez, R. E., \& Sweetser, E. (2006). With the future behind them. Cognitive Science, 30, 401-450.
Núñez, R., \& Cooperrider, K. (2013). The tangle of space and time in human cognition. TiCS, 17, 220-229.
Rinaldi, L., Locati, F., Parolin, L., Bernardi, N. F., \& Girelli, L. (2016). Walking on a mental time line: Temporal processing affects step movements along the sagittal space. Cortex, 78, 170-173.
Schutze, C. T. (1996). The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology. University of Chicago Press.
Sell, A. J., \& Kaschak, M. P. (2011). Processing time shifts affects the execution of motor responses. Brain and Language, 117, 39-44.
Sweetser, E. (1991). From Etymology to Pragmatics. Cambridge: Cambridge University Press.
Thibodeau, P. H., \& Boroditsky, L. (2011). Metaphors we think with. PLoS One, 6(2), e16782.
Torralbo, A., Santiago, J., \& Lupiáñez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. Cognitive Science, 30(4), 745-757.
Tversky, B., Kugelmass, S. and Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. Cognitive Psychology, 23, 515-557.
Ulrich, R., Eikmeier, V., de la Vega, I., Ruiz Fernández, S., Alex-Ruf, S., \& Maienborn, C. (2012). With the past behind and the future ahead: Back-to-front representation of past and future sentences. Memory \& Cognition, 40(3), 483-495.
Winter, B., Marghetis, T, Matlock, T. (2015). Of metaphors and magnitudes. Cortex, 64, 209-224.
Yang, W., \& Sun, Y. (2016). A monolingual mind can have two time lines. PBR 23, 857-864.

# Boosting Knowledge-Building with Cognitive Dialog Games 

# Jonathan S. Herberg (jonathan.herberg@gmail.com) Ilker Yengin (yengini@ihpc.a-star.edu.sg) Praveena Satkunarajah (praveenas@ihpc.a-star.edu.sg) Margaret Tan (mtan@ihpc.a-star.edu.sg) 

Institute of High Performance Computing, A*STAR
1 Fusionopolis Way, \#16-16 Connexis, Singapore 138632


#### Abstract

Dialog game tools are text chat applications which aim to structure and promote students' collaborative learning by having them select a label and sentence-opener for each message they type to their learning partner. In this experiment, we compared students' learning from discussions via a dialog game tool to their learning via a standard freechat application. Students discussed topic questions with a learning partner. They then individually completed a multiple choice test, for assessing knowledge-gain, and a short-answer test, to assess readiness for knowledge-building. Results suggest that dialog games applications lead to increased readiness for knowledge-building, in the form of integrating distinct pieces of learned knowledge, than freechat applications. Follow-up analyses suggest that the degree of concept overlap between students' dialog messages and topic keywords, as measured by a "semantic fingerprint" system, is a potentially useful metric for predicting students' knowledge-building. Implications and potential applications of our findings are discussed.


Keywords: collaborative learning; generative learning; knowledge-building; metacognition; dialog games

## Introduction

One technique that aims to enhance collaborative learning activities among students, and to promote their communicative interaction skills, is to employ the dialog games approach. Dialog-game applications are computerized education-tools that structure students' interactive text chats by having them select the function of each dialog act they make. For each dialog act they also choose a sentence opener "scaffold" from a set of options available for the dialog act type. Such applications have been demonstrated to facilitate construction of structured communication behavioral patterns such as helping, information-seeking, probing, and instructing, between online learners (e.g., Ravenscroft, Wegerif, \& Hartley, 2007; Wells, 2014).

Analyses of learners' dialog patterns in their use of dialog games applications suggest several avenues by which they potentially may lead to more effective collaborative learning. In particular, the structure of communication promoted by dialog games implementations may improve common understanding of the knowledge perspective of one's dialog partners, more effective and coherent
argumentation, and more critical thinking (e.g., Carlson, 2012; Weigand, 2016).

Along these lines, one possibility is that dialog games applications may encourage more metacognition. Metacognition in this context refers to thinking about knowledge states, including insufficient knowledge, whether one's own or one's learning partner. It is a core factor for self-regulated learning patterns, which involve targeting one's misconceptions and effectively integrating newly learned information with prior knowledge (Azevedo et al., 2009). In a collaborative learning context, in addition to metacognition encouraging self-correction of one's misconceptions, it may elicit explanation and re-representation of one's knowledge to one's learning partners that in turn may support the construction of more robust knowledge-representations.

In other words, several patterns of behavior encouraged by dialog games applications may align with those that promote generative learning. Generative learning is learning which goes beyond mere memorization, involving deeper cognitive processing, manipulation, and restructuring of information (e.g., Fiorella \& Mayer, 2015). The outcome is new knowledge that can be applied in novel situations. Selfexplaining and re-representing information in order to teach others are examples of learning strategies which can lead to generative learning. Experimental evidence supports the notion that self-explaining can increase one's integration of learned knowledge and inferring of new knowledge (e.g., Ainsworth \& Burcham, 2007).

Tied to the notion of generative learning are the levels of learning in Bloom's taxonomy that go beyond remembering and understanding learning-domain information (Bloom, 1956). In particular, the "apply" and "analyze" levels involve transferring learned knowledge in order to solve problems and infer new knowledge. Related to this notion, in tutor-learner dialogs, tutor behaviors that encourage knowledge-building, or inference of new knowledge from existing knowledge, rather than shallow knowledge-telling behaviors (e.g., when the tutor immediately jumps to correct a learner's misconception, rather than eliciting the learner to figure out his or her own misconception) entail more generative learning (Roscoe \& Chi, 2007). The analysis of tutor-learner dialogs by Chi et al. (2001) indicates that certain dialog patterns, namely those which are interactive in nature, (which means that they contain joint-actions), encourage more generative learning, whereas dialogs that
are dominated by the tutor lead to more shallow learning. Whereas self-explaining is a constructive learning activity, i.e. one that encourages knowledge inference, it is not an interactive constructive activity. According to Chi et al., (2001) behaviors that are at-once both interactive and constructive are the core drivers of effective tutor-learner interactions. That is, the most effective tutor-learner dialogs are ones in which new knowledge is jointly constructed for the learner. In particular, their extensive analysis of tutorlearner dialogs suggests distinct interactive patterns that define effective knowledge-construction. An example of such a pattern is a tutor providing scaffold prompts (e.g., hints and highlights of relevant information) for the learner to figure out the solution to a learning-domain question or problem. Chi et al. (2001) further crafted categories of questions intended to assess whether a learner has acquired information laid out in a learning-text (text-explicit questions), has effectively integrated information from different places in the learning text (text-implicit questions), or has successfully constructed and applied a mental model for the learning domain, not explicitly described in the learning text (model-implicit questions). Ainsworth et al. (2007) in their self-explanation learning studies have adopted some of these questions, referring to the latter two categories as "implicit" and "knowledge-inference" questions. However, we would argue that the integration of disparate pieces of domain knowledge toward figuring out the answer to a question, as opposed to arriving at the answer by mere recall, is itself also a form of knowledge-inference, even if it does not involve an implicit mental model. Thus, we regard successful answering of both text-implicit and model-implicit questions as entailing some form of knowledge-building.

Our hypothesis for the study was that the patterns of communicative interaction promoted by a dialog-games application would elicit more generative learning among peer-learners than a free chat application. We developed a dialog-games application and a control free chat application and designed an experiment to evaluate students' collaborative learning outcomes. This included evaluating students' basic knowledge-gain through their performance on multiple-choice items assessing (text-explicit) recall and understanding of the learning material. Critically, to test our generative learning hypothesis we assessed participants' readiness for knowledge-building, through their performance on short-answer items that required them to either integrate pieces of existing knowledge (recalled from the learning-material) or to infer the answer by applying an accurate implicit mental model based on recalled learning-material information. We utilized three text-implicit questions for the first category of knowledgebuilding questions, and three model-implicit questions for the second kind. Our study thus extends prior research by investigating whether specifically the scaffold functions of dialog game applications enhance collaborative learning and increase the potential for knowledge-building. Additionally, we also explored the possibility of applying a natural-
language processing system to obtain dialog metrics that effectively predict better knowledge-building from students, in order to investigate the feasibility of integrating such features into dialog-games tools.

## Method

## Participants

Participants included in the analyses were $569^{\text {th }}$ grade students across three secondary schools in Singapore. Signed parental consent was obtained for these students to participate at a pre-scheduled school-day time in school classrooms or computer labs that had been made available for the experiment, with laptops set up at desks in the rooms. ${ }^{1}$

## Materials

The learning domain was the human circulatory system that was adapted from the Chi et al. (2001) peer-tutor dialog study. Each of the 13 subsections was designed on the computer screen to describe each topic (e.g., "The Blood Flow in the Heart"). Diagrams were added to facilitate comprehension. Bullet points under the diagrams described the main concepts.

The topic questions that students discussed via text chat are shown in Table 1. Topic 1 corresponds to questions in the Chi et al. taxonomy which require integration of distinct pieces of explicitly learned knowledge. That is, the two subquestions for Topic 1 are text-implicit knowledge-building questions. Topic 2 question, in addition, requires the correct mental model of circulation (as a double-loop) to answer the question effectively. Thus, it is a model-implicit knowledgebuilding question. Topic 3 provided students a general discussion about the learning-domain concepts.

There were pretest and posttest multiple choice questions to gauge students' prior knowledge, and their recall and understanding of the learning material. Also for the posttest, students received six knowledge-building questions. They are shown in Table 2. The first 3 are text-implicit questions, and the latter 3 are model-implicit questions, as developed and utilized by Chi et al. (2001) and Ainsworth (2007).

There were two conditions of the dialog games text chat tool employed for this experiment. For the Scaffold condition, the application included dialog act labels for

[^350]students to select, and corresponding sentence openers. These message types were based on speech-act theory and were adopted from those used in other dialog game implementations (Weigand, 2016). Students were also provided with a sheet that defined the different dialog acts to guide them (see Table 3). Table 3 also shows examples of sentence-openers that students could choose for each dialog act. Figure 1 illustrates the design of the dialog game window, with labels numbered indicating the steps for entering and sending a dialog message, as follows: (1) The topic question that defines the parameters of a given dialog is at the top of the screen. (2) Users may click on a bubble next to a dialog message to make a reply to the specific message (which can also be used to reply to earlier messages in the chat history). Reply messages are indented relative to the original message. If no reply bubble is clicked, the entered text message will appear below all the text messages in the chat window, with no indentation. (3) Users select one of the six communicative act labels, and then select a linked sentence opener from a dropdown menu. The selected sentence opener appears at (4). (5) Users type in the rest of their message into the text box. Note that only one user may type into his or her text entry box message at a time. If it is the other user's turn, the shadow text in the box says "Please wait your turn." If it is the given user's turn, it says "Enter your text." In addition, if a user has failed to first select a dialog act label and sentence opener, on clicking the text entry box a reminder message will appear, and the user is unable to type into the box until making these selections. (6) When a user has completed a message, he or she clicks the "Send" button.

The Freechat application was of similar design and appearance as the Scaffold application, and included the turn-by-turn use features, but did not feature the dialog act label buttons and sentence-opener display. Thus, users took turns simply entering messages, without the scaffold steps.


Figure 1: Dialog Game Screen.

Table 1: Topic discussion questions

| Topic No. | Discussion Question(s) |
| :---: | :--- |
| 1 | a) Why do we have valves in veins, but not in arteries <br> and capillaries? <br> b) Why don't we have valves in pulmonary veins? |
| 2 | Why do we sometimes refer to the heart as a "double <br> pump"? |
| 3 | What do you think are the most interesting aspects of <br> the structure and function of the human circulatory <br> system? Please discuss. |

Table 2: Posttest questions to assess Knowledge-Building

| Item <br> No. | Short-Answer Question |
| :---: | :--- |
| 1 | Why is there an artery that carries deoxygenated blood? |
| 2 | Why do vessels get increasingly smaller as they get close to <br> the body cells, and increasingly larger as they get nearer to the <br> heart? |
| 3 | In which kind of blood vessels (arteries, veins, or capillaries) <br> is the blood pressure the lowest? Why? |
| 4 | Why is your right ventricle less muscular than your left <br> ventricle? |
| 5 | The artery that carries blood from the right side of the heart to <br> the lungs (the pulmonary artery) carries about the same <br> amount of blood as the artery that carries blood from the left <br> side of the heart to the rest of the body (aorta). Why do they <br> carry the same amount of blood? |
| 6 | What would happen if the valves between the atria and the <br> ventricles got stuck open and wouldn't close? |

Table 3: Descriptions of communicative act labels, with example sentence-opener choices (Scaffold condition)

| Dialog Act | Description | Example Sentence Opener <br> Choices |
| :---: | :---: | :---: |
| Information | To provide or <br> describe relevant <br> facts or knowledge. | Let me explain... <br> Some facts are... <br> My understanding is that... |
| Propose | To bring up a new <br> idea to consider. | I suggest that... <br> Let us focus on... <br> I think it makes sense to... |
| Challenge | To argue against, or <br> provide evidence <br> against a dialog <br> statement. | I disagree because... <br> A counter-argument is... <br> An alternative view is... |
| Question | To ask your dialog <br> partner about <br> something you don't <br> know. | Why is it... <br> Can you explain... <br> What do you think about... |
| Agreement | To agree with a <br> statement made by <br> your dialog partner. | I agree, $\ldots$ <br> Good point, $\ldots$ |
| Support | To argue for, or <br> provide evidence for <br> a dialog statement. | I think this view is supported <br> by, $\ldots$ <br> To give an example, .... |

## Procedure

Students were randomly assigned to the Scaffold condition, involving the text chat application that required them to select dialog act labels and sentence openers, or to the

Freechat condition. There were 28 students for each condition. Within each condition, the students were again randomly assigned to dialog-discussion pairs. Students in each condition were taken to separate rooms for the study (Scaffold or Freechat). To minimize verbal and indirect interaction, no student sat next to any other student. Each of two experimenters was also randomly assigned to conduct the session for each condition.

In each study session room, pre-arranged laptops were placed on the desks. The experimenter overviewed the session, which consisted of the following tasks:

1) Students were given up to 7 minutes to individually complete the multiple-choice pretest (could click "submit" if they finished early). (The timer for all tasks was viewable at the top of the application window).
2) Following the pretest, students were taken to the learning material screen where they had 15 minutes to read and study the learning material.
3) Then the experimenter went over how to use the system. For the Scaffold condition, the experimenter went over the different communicative act labels, and the steps for entering in a message including a sentence opener. Students also received a dialog act description sheet (Table 3).
4) The students were given a five-minute demo dialog session to help them get accustomed to the application.
5) Next, the students (with their randomly assigned learning partner) discussed the dialog questions for the 3 topics. For both conditions, students took turns entering in a dialog message. They could also open a pop-up window that contained the learning-material, which they could refer to for the discussions. For each topic, students had a 10 minute dialog discussion.
6) Following the end of their dialogs discussion, the students completed the post-test individually. These consisted of the same multiple-choice questions as in the pre-test ( 6 minutes). In addition, they had to answer the short-answer questions (as in Table 2) to assess knowledge-building, for which they were given 25 minutes. For each portion of the posttest, students could click a "submit" button if they finished early.

## Measures

Knowledge-gain. To assess students' knowledge gain from reading the learning material and engaging in the dialog discussions, their scores on the posttest multiple choice (out of 10 points) were compared to their pretest scores.

Knowledge-building. A scoring guide was developed that allowed for 2 points maximum on each of the three text-implicit questions, and 3 points maximum on each of the three model-implicit questions. Basically, a point was awarded for each piece of information relevant for inferring the answer to the question, and for each correct inference.

For example, for Question 4, one point would be awarded for an accurate description of the function of the left ventricle, one for the right ventricle, and one point for the inference that the right ventricle doesn't need to pump blood with as great force as the left, as the blood travels less distance. Two raters, familiar with the scoring guide and the learning material and related concepts, scored participants' answers to these questions. They were kept naïve to the experimental condition for all the short-answer logs. The scores were averaged across the two raters. The intraclass correlation for absolute agreement on the items was computed as $\operatorname{ICC}(1,128)=0.87$ for the text-implicit items and $\operatorname{ICC}(1,128)=0.96$ for the model-implicit items.

Topic adherence. We conducted exploratory follow up analyses that utilized the "semantic fingerprint" system developed by the Cortical.io Company (with the API available on their website). The goal was to assess the feasibility of utilizing natural-language processing methods to predict students' capacity for knowledge-building (short-answer performance) from their dialog messages. Such functions, if predictive, could be useful to incorporate into dialog game applications, for teachers and students to track (in an automated fashion) learning outcomes implicitly from dialogs. The Cortical.io system represents the meaning of words in terms of their distributional overlap in a large linguistic corpus (i.e., Wikipedia). Its theoretical basis is the notion of distributional semantics, or "word spaces" (e.g., Sahlgren, 2006). The more frequently that words co-occur in near proximity in the corpus, the higher is their computed "semantic fingerprint overlap." The metric can also be extended by the system to compute the degree of semantic fingerprint overlap among text segments and documents, rather than of single words. For implementation details, refer to De Sousa Webber (2015).

Dialog file inputs were first corrected for spelling errors and abbreviations. What we refer to as "topic adherence" is, for each topic dialog and participant, the semantic fingerprint overlap between the participant's dialog messages (entered into the system as a single "document") and pre-selected keywords intended to represent important concepts related to the topic question. Refer to Table 1 for the Topic questions. For Topic 1, the keywords were: "valves," "veins," "arteries," "capillaries," "pulmonary," and "pressure." For Topic 2, they were: "heart," "lungs," "oxygen," "blood," and "pump." For Topic 3, they were "valves," "veins," "arteries," "heart," "lungs," "oxygen," "blood," and "circulatory." The additional dialog metrics of mean number of words-per-turn, and total number of turns, were used.

## Findings and Discussion

## Knowledge-gain scores

Figure 2, on the two pairs of bars on the left, shows the mean scores across the Scaffold and Freechat conditions on the pre-test and post-test multiple choice for assessing
students' level of recall and understanding of the domain material. It also displays the proportion-scores, so that tests with different scales can be displayed on the same chart. Participants did not differ significantly on their pretest scores, $t$ (53) < 1. Across both conditions, participants showed improvement on their post-test multiple-choice scores relative to their pre-test scores, $t(55)=5.22, p<$ .001 , with an effect size of $d=0.76$. The knowledge-gain (post- minus pre- test score difference) in the Scaffold condition ( $\mathrm{M}=1.07$ ) did not significantly differ from the Freechat knowledge-gain $(\mathrm{M}=1.14), t(53)<1$. The two conditions also did not differ significantly on the mean posttest multiple-choice scores, $t(53)<1$.

## Knowledge-building scores

To assess our hypothesis of increased knowledge-building for the Scaffold condition, we first conducted a MANOVA on the text-implicit and model-implicit scores. There was an overall effect of condition, $F(2,53)=3.19, p<.05$. Figure 2 , on the two pairs of bars on the right, shows the mean proportion-scores across the two sets of knowledge-building questions (text-implicit and model-implicit). The follow-up tests indicated no effect of condition on the model-implicit questions, $t$ (54) < 1 . However, for the text-implicit questions, the mean score was higher in the Scaffold than the Freechat condition, $t(54)=2.39, p=.02$, with an effect size of $d=0.64$.


Figure 2: Mean scores ( $+/-\mathrm{SE}$ ) in the Scaffold and Freechat for the multiple-choice pretest and posttest, and the knowledge-building tests (text-implicit and model-implicit).

## Dialog metrics for knowledge-building

We conducted follow-up multiple-regression analyses to explore whether the topic adherence scores obtained by the semantic fingerprint system, along with the metrics of words-per-turn and number of turns, could be of use for predicting students' readiness for knowledge-building (i.e., their short-answer scores). Scores were averaged for each participant across the three dialog topics. Tables 4 and 5
show the regressions separately on the Scaffold and the Freechat cases, respectively. For the Scaffold condition, the overall regression trends toward statistical significance, and the coefficient for topic-adherence reaches statistical significance. Total-turns trends in the direction of predicting increased knowledge-building scores. For the Freechat condition, the overall regression also trends toward statistical significance, but with a non-significant coefficient for topic adherence, and with the total-turns coefficient trending in the direction of predicting reduced knowledgebuilding. Overall, across both regressions words-per-turn appears to be a relatively weak predictor.

Table 4: Multiple regression for predicting knowledgebuilding (Scaffold condition)

| Predictor | B | SE B | $\boldsymbol{\beta}$ | $\boldsymbol{t}$ | $\boldsymbol{p}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Topic adherence | 19.53 | 7.64 | 0.55 | 2.56 | $0.02^{*}$ |
| Words-per-turn | 0.06 | 0.07 | 0.24 | 0.97 | 0.34 |
| Total turns | 0.42 | 0.21 | 0.52 | 1.96 | 0.06 |
| $\mathrm{R}^{2}=0.27, F(3,22)=2.72, p=0.07$ |  |  |  |  |  |

Table 5: Multiple regression for predicting knowledgebuilding (Freechat condition)

| Predictor | B | SE B | $\boldsymbol{\beta}$ | $\boldsymbol{t}$ | $\boldsymbol{p}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Topic adherence | 4.62 | 6.09 | 0.16 | 0.76 | 0.46 |
| Words-per-turn | -0.01 | 0.04 | -0.08 | -0.34 | 0.74 |
| Total turns | -0.13 | 0.08 | -0.41 | -1.62 | 0.12 |
| $\mathrm{R}^{2}=0.23, F(3,24)=2.32, p=10$ |  |  |  |  |  |

## Discussion

Our hypothesis was partially supported. Namely, students in dialog-games interactions to discuss topic questions in the learning domain exhibited a higher readiness for knowledge-building, in the form of making text-implicit inferences, than students in the freechat discussions. There was no significant improvement on model-implicit questions. The increased knowledge-building readiness was also over and above any knowledge-gain, which did not significantly differ between conditions.

In addition, the multiple-regression results suggest that natural-language processing methods may hold some promise in producing dialog metrics with predictive utility for knowledge-building. The predictive value (in terms of the standardized Beta coefficient) of our topic-adherence metric was particularly more prominent for the Scaffold condition than for Freechat condition. Also of interest, though more caution is warranted for interpretation of nonstatistically significant trends, is that the total-number of dialog turns went in the direction of predicting more knowledge-building for Scaffold condition, and less knowledge-building for Freechat condition. These trends
may be indicative of qualitative differences in the nature of dialogs with versus those without scaffolds, with a tendency for scaffolds to raise the potential learning value of each dialog turn, and to increase the potential knowledgebuilding when dialog partners jointly discuss core concepts in the learning domain. One possibility is that the dialog game scaffold functions in effect promote more selfexplanation in the process of developing explanations and arguments to one's dialog partner. An extensive study of collaborative learning dialogs by Asterhan \& Schwarz (2009), on the other hand, suggests that the process of argumentation itself may be essential for driving conceptual change from the joint construction of explanations. In the current context, if the dialog game scaffold functions encouraged more structured argumentation, this would open the door for dialogs that are more focused on the main topic concepts to generate improved conceptual understanding of the learning domain.

The conceptual foundation for applying the framework of dialog-games to learning is grounded in the notion of learning as a dialectical, social, and interactive process (cf. Mercer \& Littleton, 2007). Structuring a learning-discussion as dialog-game is therefore seen as a means to encourage effective argumentation and critical thinking (e.g., McAlister, Ravenscroft, \& Scanlon, 2004). In terms of Bloom's taxonomy, the potential, more immediate benefits of dialog-games can be viewed as focused on the application and analysis levels of learning. However, effective learning at these levels requires first a solid groundwork of basic understanding of concepts in a learning domain, and in turn takes time. Reaching even higher levels of learning, and unlocking creativity, is an ever increasing long-term process (cf. Bloom, 1956). Thus, dialog games may be beneficial for developing students' creativity, but this would need to be evaluated by an extended use of such applications for learning, e.g. over weeks or months.

Along these lines, one limitation of the current study is that it was a "single-shot" learning and evaluation session. For generative learning more time for absorbing, processing, and transforming information may be an essential element (Fiorella \& Mayer, 2015). Thus, even on the text-implicit questions, for which there was a medium-sized effect for the difference between conditions, the mean proportion of total points obtained was for both conditions only about half of the total possible. In addition to being constrained by time for the current study, another note is that dialog-games are often applied for conversations among small-groups (Ravenscroft, 2007). It is possible that learning-dialogs for groups of 3 or 4 may allow for more argumentation and perspective-taking opportunities than two-way dialogs. Future research directions are indicated for "scaling" up dialog-games applications for knowledge-building, both in terms of time (over a long-term learning period) and in terms of group-size (e.g., from learning-pairs to learning-groups). Such extensions may lead to larger-scale knowledge-building effects, and increase
the predictive value of dialog metrics for knowledge-building.

## References

Ainsworth, S., \& Burcham, S. (2007). The impact of text coherence on learning by self-explanation. Learning and instruction, 17(3), 286-303.
Ainsworth, S., \& Loizou, A.T. (2003). The effects of selfexplaining when learning with text or diagrams. Cognitive Science, 27(4), 669-681.
Asterhan, C. S., \& Schwarz, B. B. (2009). Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. Cognitive Science, 33(3), 374-400.
Azevedo, R., Moos, D. C., Johnson, A. M., \& Chauncey, A. D. (2009). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. Educational Psychologist, 45(4), 210223.

Bloom, B. S., Engelhart,, M. D. Furst, E. J., Hill, W. H.; Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York, NY: David McKay Company.
Carlson, L. (2012). Dialogue games: An approach to discourse analysis (Vol. 17). Springer Science \& Business Media.
Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., \& Hausmann, R. G. (2001). Learning from human tutoring. Cognitive Science, 25(4), 471-533.
De Sousa Webber, F. (2015). Semantic folding theory and its application in semantic fingerprinting. White paper retrieved from http://www.cortical.io.
Fiorella, L. \& Mayer, R. E. (2016). Eight ways to promote generative learning. Educational Psychology Review, 28(4), 717-741.
Mercer, N., \& Littleton, K. (2007). Dialogue and the development of children's thinking: A sociocultural approach. Routledge.
Ravenscroft, A. (2007). Promoting thinking and conceptual change with digital dialogue games. Journal of Computer Assisted Learning 23(6), 453-465.
Ravenscroft, A., Wegerif, R. \& Hartley, R. (2007). Reclaiming thinking: dialectic, dialogic and learning in the digital age. BJEP Monograph Series II, Number 5Learning through Digital Technologies, 1(1), 39-57.
Roscoe, R. D. \& Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. Review of Educational Research, 77(4), 534-574.
Weigand, E. (2016). The dialogic principle revisited: Speech acts and mental states (pp. 209-232). Springer International Publishing.
Wells, S. (2014). Supporting argumentation schemes in argumentative dialogue games. Studies in Logic, Grammar and Rhetoric, 36(1), 171-191.

# Leveraging mutual exclusivity for faster cross-situational word learning: A theoretical analysis 

Shohei Hidaka (shhidaka@jaist.ac.jp) and Takuma Torii (tak.torii@jaist.ac.jp)<br>Japan Advanced Institute of Science and Technology<br>1-1 Asahidai, Nomi, Ishikawa, Japan

George Kachergis (george.kachergis@gmail.com)<br>Dept. of Artificial Intelligence / Donders Institute, Radboud University<br>Nijmegen, the Netherlands


#### Abstract

Past mechanistic accounts of children's word learning claim that a simple type of cross-situational learning is powerful enough to match observed rates of learning, even in quite ambiguous situations. However, a limitation in some of these analyses is their reliance on an unrealistic assumption that the learner only hears a word in situations containing the intended referent. This study analyzed a more general type of cross-situational learning based on the relative frequency of word-object pairs, and found it to be slower than the simple mechanism analyzed in prior work. We then analytically explored whether relative-frequency learning can be improved by incorporating the mutual exclusivity (ME) principlean assumption that words map to objects 1 -to-1. Our analyses show that with a certain type of correlation in word-to-word relationship, ME makes relative frequency learning as efficient as fast-mapping, which can learn a word in one exposure.


Keywords: Word learning; Cross-situational learning models; Mutual exclusivity; Language acquisition

## Introduction

To a new learner of a language with a completely unknown word-referent mapping system, determining which words refer to which referents in any given scene may seem impossible on the face of it, since a word could refer not only to an object (e.g., 'apple'), but to a class of objects (e.g., 'fruit'), a property ('red'), or any one of endless possible combinations or configurations of features in the scene-an unconstrained problem of logical induction (Quine, 1960). In contrast to this theoretical observation about referential uncertainty, children are thought of as efficient learners, and in fact most human children do learn to understand and use an impressive number of words within the first years of life, achieving a vocabulary of roughly 60,000 by 18 years of age (Bloom, 2000). Developmental researchers have theorized that children use a variety of lexical constraints to limit the number of possible mappings they consider, and a number of empirical studies support these claims (Clark, 1987; Markman, 1990, 1992; Golinkoff, HirshPasek, Bailey, \& Wegner, 1992). One lexical constraint, used here as in past theoretical accounts-and supported by empirical developmental data, is that learners are biased to map words to entire objects, rather than to a feature of an object, or a group/configuration of objects (Markman, 1990).

Beyond lexical constraints that reduce the number of hypothesis meanings considered for a given word in a given situation, another possible remedy for the contradiction between the difficulty of the unconstrained inductive account of word learning and the ease of the observed process is that learners also reduce uncertainty in the word-object map by statistical inference over time, based on observing word-object pairs across multiple situations. Cross-situational learning (Pinker, 1984; Akhtar \& Montague, 1999; Siskind, 1996) is a type of learning based on this idea, which has been analyzed both empirically and theoretically over decades (Yu, 2008; Blythe, Smith, \& Smith, 2010). Blythe et al. (2010) formally quantified the effect of a type of crosssituational learning in terms of the rate of vocabulary growth. More recent studies (Blythe, Smith, \& Smith, 2016; Vogt, 2012) further showed that this type of crosssituational learning can be considerably slowed down for certain types of word co-occurrence distributions, including power-law distributions in which most words are seen relatively rarely, which describe word frequency distribution in natural languages (Zipf, 1949).

These theoretical analyses are still quite limited in their generality. The class of cross-situational learning analyzed in these past studies is called eliminative learning. In this scheme, when a learner is exposed to a set of referents, a correct word is spoken-never is a word spoken when its intended referent is not present. In this case, the learner can safely "eliminate" the possibility of word A being associated to the object B, if he or she experiences one episode that the word A is spoken without the object $B$. As this special assumption does not generally hold in real-world learning, the estimates on the speed of cross-situational learning in past studies give only an optimistic upper bound for its learning efficiency.

In this study, we consider a more general type of crosssituational learning, called relative frequency learning, of which eliminative learning is a special case. In the relative frequency learning scheme, it is assumed that a language system encodes the word-object pair with frequency higher than the other candidate pairs as the correct one, and the learner infers such relatively more
frequent word-object pairs from the sample. Under this assumption, the eliminative learning scheme is identified with the special case of seeing the correct word-object pair with probability 1 . In general, however, the eliminative learning rule cannot apply (or will mislead the learner if it is forced to apply) in word learning of a relative-frequency language system.

Therefore, relative frequency learning is generally slower than the eliminative learning. Thus, the main problem considered in this study is what plausible factor might make this type of learning more efficient - and can it be made efficient enough to be a realistic account for children's word learning? Specifically, we analyze the beneficial effect of learners applying a general principle of mutual exclusivity (ME), an assumption of a wordobject regularity requiring that no two objects are associated to one word. Application of a ME principle has long been theorized to be a constraint that can speed children's word learning (Markman \& Wachtel, 1988), and has found empirical support in both children (Halberda, 2003) and adults (Yurovsky \& Yu, 2008; Kachergis, Yu, \& Shiffrin, 2012). We then consider a word-word statistical relationship in which a group of distractor objects tend to co-occur with a word and thus slow learning.

In the following, we first outline the theoretical framework in which we provide a series of analyses of relative frequency learning. Second, we evaluate the basic learning efficiency in this scheme. Then we extend this evaluation of learning efficiency to multiple scenarios with different word-to-word statistical relationships.

## Relative-frequency learning

## Basic framework

In this study, we consider the following word learning scenario. The learner is exposed to multiple words and objects in each situation. In each situation, the learner does not know which word refers to which object, and the correct word-object mapping can only be inferred by integrating evidence across observations of multiple situations. Let $W=\{1, \ldots, n\}$ be a set of words and $O=\{1, \ldots, m\}$ a set of objects (or referents) which appear in these situations. In this study, we consider a language structure with one-to-one word-object mapping, in which $n=m$ and the word $i$ refers to object $i$. This is a quite strong assumption, which may not be considered entirely realistic as it is. It offers, however, a first approximation upon which we can base the analysis and later extend it.

Here, we consider a particular word learning scheme, called relative frequency learning, in which each object's to-be-associated (i.e., 'correct') word is spoken in its presence with greater frequency than any other word. This is a code in the sense of information theory - the signal, the correct word-object mapping, is encoded in the statistical regularity in observation across situations
(channel), and the learner decodes (infers) the correct word-object map using the underlying regularity: the correct word-object pair is the most frequent among the others.

There are theoretical analyses of a special case of this relative frequency learning, in which the correct word is spoken only in the presence of the corresponding object $($ i.e., $p($ object $\mid$ word $)=1$ ). In this special case, the learner can use not only the knowledge that the correct pair is more frequent, but also the quite strong rule that any object which does not appear with a spoken word cannot be the intended referent of that word. Thus, this learning scheme, which eliminates any word-object pair with probability less than 1 is called eliminative learning (Blythe et al., 2010). In this study, beyond this special case, we analyze a more general case of language and learning coded on the basis of relative frequency.

## Formulation

Denote the frequency of object $o$ given word $w$ by $f(o \mid$ $w)$. Then, suppose the learner (decoder) declares that the object $o \in O$ is the referent of the word $w \in W$ with probability

$$
P(o \mid w)=\frac{e^{f(o \mid w)}}{\sum_{o \in\{O\}} e^{f(o \mid w)}}
$$

In this scheme, the error, wrong declaration of the correct object, for word $w$ with the number of observed situations $n$ is proportional to $\epsilon(n, w)$ := $\sum_{o \neq w} e^{f(o \mid w)-f(w \mid w)}$. The sum of the errors for all words $\epsilon(n, w):=\sum_{w \in W} \epsilon(n, w)$ is an exponential function of the number of situations. Let us denote the rate of the exponential function as $R$, and thus $\epsilon(n)=e^{-R n}$. For a code with rate $R$ encoding less than $e^{R n}$ signals, $\lim _{n} \sum_{w \in W} P(o \mid w)=1$, and thus it is said to be learnable (reachable in information-theoretic terminology). If the rate satisfies $\epsilon(n)=e^{-R n}<e^{-C n}$ for any code, the constant $C$ is said to be the capacity of this channel in information-theoretic terms (Shannon, 1948). The rate, or the exponent coefficient of the error function, is a fundamental characteristic of the language-learning system when viewed as a signal transmitting process.

## Efficiency

In the relative frequency learning scheme, the object $o$ with the second largest probability given the word $w$, $p_{w \mid w}>p_{o \mid w}>p_{o^{\prime} \mid w}$ for $o^{\prime} \neq o, w$, is a key parameter giving the asymptotic time to learn the word $w$. With objects with the largest and second largest probability, the sample frequency can be written as follows. Let $\bar{p}=$ $1-p$. Specifically, consider that the sample frequency $f_{\text {now }}=f_{n}(o \mid w)$ follows the binomial distribution

$$
P\left(f_{\text {now }} \mid n, p_{\text {ow }}\right)=\binom{n}{f_{\text {now }}} p_{o w}^{f_{\text {now }}} \cdot \bar{p}_{\text {ow }}^{n-f_{\text {now }}}
$$

with probability $p_{o w}$.
Given this, the error probability in learning is characterized as follows. The probability for the word $w$ to be associated with the object $o$ is proportional to $e^{f_{\text {now }}}$. For a sufficiently large $n$, the difference between the two random variables asymptotically approaches

$$
\lim _{n \rightarrow \infty} \frac{e^{f_{n o w}-f_{n o^{\prime} w}}}{e^{n \Delta_{o, o^{\prime} \mid w}}}=C,
$$

where $\Delta_{o, o^{\prime} \mid w}:=\frac{p_{o w}-p_{o^{\prime} w}}{p_{o w} \bar{p}_{o w}+p_{o^{\prime} w} \bar{w}_{o^{\prime} w}}$. If there are $m$ objects with the second largest probability $p_{o w}>q>$ $\max _{o^{\prime} \neq w}, p_{o^{\prime} w}$ for the word $w$, the error probability is $1-P(w \mid w) \rightarrow C m e^{-n \frac{p_{o w-p_{o^{\prime}} w}}{p_{o w} \overline{\bar{o}_{o} w+p_{o^{\prime}} w^{\overline{p_{o}}{ }^{\prime} w}}} \text {. Thus, the rate of }}$ the relative-frequency code is $R=\min _{w} \Delta_{w \mid w}$ where

$$
\Delta_{o \mid w}:=\frac{p_{o w}-\max _{o^{\prime} \neq o} p_{o^{\prime} w}}{p_{o w} \overline{p_{o w}}+\max _{o^{\prime} \neq o} p_{o^{\prime} w} \overline{\max _{o^{\prime} \neq o} p_{o^{\prime} w}}}
$$

This analysis implies that the word-object pair with the smallest margin to second largest probability decides the learning rate in the relative frequency code.

## Incorporating mutual exclusivity (ME)

In the above analysis of the relative frequency code, the lexical constraint of one-to-one word-object mapping is not taken into consideration in the learning process. However, if the learner exploits the fact that no two objects are associated with the same word, namely correct word-object pairs are mutually exclusive, the learning is expected to be more efficient than the alternative without the knowledge. Let us call this ME learning. The difference in the rate of learning assuming ME and general relative frequency would be the effect of introducing a ME constraint in cross-situational learning.

With ME, the learner can exclude object $o$ when learning word $w$, if the object $o$ is likely to be associated with some other word $w^{\prime} \neq w$. Thus, the learning order of the words has considerable impact in learning under ME. As the previous analysis shows that the second most probable objects for word $w$ is the key factor giving the learning rate, let us call them distractors against the word $w$, and denote the set of distractors by $D(w):=\left\{w^{\prime} \mid \max _{o \neq w} f_{o \mid w}=f_{w^{\prime} \mid w}\right\}$.

## Best- and worst-case scenarios

Here let us analyze ME learning under a simplification that the learning time for the words with no distractor is $T_{0}$ and that for the words with one more distractors is $T_{1}$. The former case with no distractor is said to be fast mapping, in which a particular word-object pair is presented alone in a situation, and the learner learns the pair in a single shot (Carey \& Bartlett, 1978). The latter case is analyzed in the previous section in case of the relative frequency learning. In this case, if all the distractors has been eliminated, by the effect of ME, the
corresponding object can be uniquely identified, which is effectively the same as fast mapping. Thus, the worstcase learning time approaches that of relative frequency learning, and the best-case learning time approaches that of fast mapping, as the number of words is sufficiently large.

## Randomly distributed distractors

Random learning order Consider the case that each word is learned in a serial order and each has $k$ distractors. Furthermore suppose that the learning order is a random permutation, namely any order is uniformly sampled. Figure 1 shows a schematic co-occurrence matrix of the five such word-object pairs (filled markers) with $k=2$ randomly distributed distractors (open markers) for each pair. In this case, one expects that one word is likely to be learned after the $k$ distractors with probability $1 /(k+1)$. This is exactly true, if the number of words $n$ approaches to infinitely large. Therefore, the sum of expected learning time for all the words is

$$
\begin{equation*}
T=n\left(\frac{k}{k+1} T_{1}+\frac{1}{k+1} T_{0}\right) . \tag{1}
\end{equation*}
$$

Thus, when the learning order is a random permutation, the expected learning time is only the factor of $\frac{1}{k+1}$ shorter than the original time $n T_{1}$ at shortest.

| Word | Objects |  |  |  | $\# D$ |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| "Circle" | $\bullet$ | $\triangle$ |  | $\star$ |  | 2 |
| "Triangle" |  | $\Delta$ | $\square$ |  | $\diamond$ | 2 |
| "Square" | $\bigcirc$ | $\triangle$ | $\square$ |  |  | 2 |
| "Star" | $\bigcirc$ |  |  | $\star$ | $\diamond$ | 2 |
| "Diamond" |  | $\triangle$ | $\square$ |  | $\diamond$ | 2 |

Figure 1: A schematic word-object co-occurrence matrix in the case with random learning order and randomly distributed distractors.

## Shared distractors

Best and worst learning order Let us consider the best and worst case by manipulating which words the $k$ distractors are associated. In one of the best cases, every word shares the same set $D$ of $k$ distractors. Figure 2 shows a schematic co-occurrence matrix of the five such word-object pairs (filled markers), and each pair has $k=2$ distractors (open markers) and most of words share the same two distractors. In this case, the shortest learning time is obtained by a sequence of learned words in which the $k$ words with the $k$ distractors as their correct objects first (required about $T_{1}$ time each) and the others later (required $T_{0}$ time each). In the example (Figure 2), one of the best order is to learn the word "Circle" and "Triangle" at the first two rows in the
matrix, and then learn the other words. In this case, the total learning time is

$$
T=k T_{1}+(n-k) T_{0}
$$

As the number of words $n$ gets larger with a constant $k$, the learning time approaches to that of the fast mapping ( $T_{0}$ per word), which is the lower bound of learning time.

In one of the worst cases, on the other hand, the longest learning time is obtained by the reversed sequence, in which the words with the $k$ distractors as their correct objects are learned last. In total, the longest learning time is

$$
T=n T_{1} .
$$

As the number of words $n$ gets larger with a constant $k$, learning time approaches that of relative frequency learning, which is the upper bound of learning time.

Random learning order Thus, this analysis with the best and worst case scenario suggests that the learning order of words has a large impact on learning time. However, the expected learning time with the shared distractors is, again, exactly $1 /(k+1)$, which is no better than the learning time of the case with $k$ random distractors (Equation (1)):

$$
T=n\left(\frac{k}{k+1} T_{1}+\frac{1}{k+1} T_{0}\right)
$$

This analysis suggests that even systematically shared distractors cannot improve the learning time on average, if the learning order is uniformly at random.

| Word | Objects |  |  | $\# D$ |
| :--- | :--- | :--- | :--- | ---: |
| "Circle" | $\bigcirc$ | $\triangle$ | $\square$ |  |
| "Triangle" | $\bigcirc$ | $\Delta$ | $\square$ |  |
| "Square" | $\bigcirc$ | $\triangle$ | $\square$ |  |
| "Star" | $\bigcirc$ | $\triangle$ |  | $\star$ |
|  | 2 |  |  |  |
| "Diamond" | $\bigcirc$ | $\triangle$ |  |  |

Figure 2: A schematic word-object co-occurrence matrix in the case with random learning order and $k=2$ distractors shared by all the words systematically.

## Correlation in word-to-word relationship Mixture of two groups of words

As the previous analysis suggests that the relative frequency learning of a one-to-one word-object map in the cross-situational setting is as slow as independent learning even by incorporating ME. This result is largely due to the statistical structure of the word-word relationship - in the previous analysis, each word has $k$ other random words as distractors. In this section, we consider a
specific class of statistical regularity in the word-word relationship. Specifically, suppose there are two groups of words: in the one group of words, each word has no distractor, and in the other group of words, each word has $k$ distractors, whose referring words have no distractor (Figure 3). Thus, the learner is exposed to a mixture of two groups of words with and without distractors. Figure 3 shows a schematic co-occurrence matrix of such five word-object pairs, in which each of the first group of words ("Circle" and "Star") has no distractors, and each of the other group of words has two distractors whose referring words are the members of the first group.
Although this statistical regularity in word-to-word relationships looks similar overall to the previous case (compare Figure 2 and 3 ), this new case is substantially different from the previous cases. The key observation here is that no distractor words have any distractors against themselves. Thus, the first group of words (potential distractors to the other group of words) would be learned via fast mapping, and the other group would be learned also via fast mapping after their distractors are learned before their learning. The learning timing of these two groups are probabilistic, but the first group of words are expected to be learned earlier on average than the other group.

| Word | Objects |  |  | \# D |
| :---: | :---: | :---: | :---: | :---: |
| "Circle" | $\bigcirc$ |  |  | 0 |
| "Triangle" | $\bigcirc$ | A | is | 2 |
| "Square" | $\bigcirc$ |  | is | 2 |
| "Star" |  |  | $\star$ | 0 |
| "Diamond" | $\bigcirc$ |  | is | 2 |

Figure 3: A schematic word-object co-occurrence matrix in the case with the two groups of words. Each of the first group of words ("Circle" and "Star") has no distractors, and each of the second group of words ("Triangle", "Square" and "Diamond") has $k=2$ distractors, whose referring words ("Circle" and "Star") has no distractors.

## Efficiency analysis

Specifically, suppose that each word in the group with distractors is learned at the time step $t$ by the probability

$$
p_{t}=\left(q_{t}+\overline{q_{t}} p\right) \overline{p_{t-1}}
$$

where $p$ is the probability to learn this word with distractors at each step, and $q_{t}$ is the probability to learn it without distractor at step $t$, or is said the probability for the learning at step $t$ to be fast mapping. By setting $\sum_{t=1}^{\infty}(1-p) p^{t-1} t=T_{1}$ and $q_{t}=0$ for any $t$, this learning time with $k>0$ distractors is identified with the previous analysis.

Suppose that there are $n_{0}$ words without distractors, and $t_{0}<t$ samples out of the all $t-1$ samples are drawn from this group of words with equal probability. Then, according to Hidaka (2014), as $n_{0} \rightarrow \infty$, the probability to learn the $m$ words of this group with the $t_{0}$ samples asymptotically approaches to the binomial distribution

$$
\sum_{m=0}^{n_{0}}\binom{n_{0}}{m} r_{t}^{m}{\overline{r_{t}}}^{n_{0}-m}
$$

where $r_{t}:=1-\left(1-1 / n_{0}\right)^{t_{0}}$. If each word in the group with distractors is associated to $k$ distractive words uniformly at random, the fast-mapping probability is

$$
q_{t}=\sum_{m=0}^{n_{0}}\binom{n_{0}}{m} r_{t}^{m}{\overline{r_{t}}}^{n_{0}-m}\binom{m}{k} /\binom{n_{0}}{k} .
$$

As the hypergeometric distribution ${ }^{1}$ approaches the binomial distribution as $n_{0} \rightarrow \infty$, we obtain

$$
\left\|\binom{m}{k} /\binom{n_{0}}{k}-\binom{m}{k}\left(\frac{k}{n_{0}}\right)^{k}\left(1-\frac{k}{n_{0}}\right)^{m-k}\right\| \rightarrow 0
$$

Using these asymptotic distributions for $n_{0} \rightarrow \infty$, we obtain the binomial distribution

$$
q_{t} \rightarrow \frac{n_{0}!}{k!\left(n_{0}-k\right)!}\left(r_{t} \frac{k}{n_{0}}\right)^{k}\left(1-r_{t} \frac{k}{n_{0}}\right)^{n_{0}-k}
$$

With further transform for a sufficiently large $n_{0}$, we obtain the fast-mapping probability to be

$$
q_{t} \approx\left(\frac{t_{0}}{n_{0}}\right)^{k}
$$

This expression thus implies that the probability $q_{t}$ of learning via fast mapping with $k$ distractors approaches 1 , if the sample of the words without distractors $t_{0}$ is comparable to the number of such words $n_{0}$.

## Implications

Suppose the number of words without distractors is $n_{0}=$ $\gamma n$ with a certain constant $0<\gamma<1$, and the number of samples $t_{0}=\gamma t$. In this case, as $t_{0} / n_{0}=t / n$, after the point when the number of samples is comparable with the number of words, this learning is sufficiently treated as the fast mapping. Thus, the learning time of a word with $k$ distractors asymptotically approaches the speed of fast-mapping after some constant number of samples for each word. In other words, in the long run, any words would be considered learned in the fast-mapping manner, if any distractor word has no distractors against itself.

[^351]This analytic implication is striking in that crosssituational learning on the basis of relative frequency, which itself is as slow as independent learning with a random word-word relationship, can become as efficient as fast-mapping, up to a constant time per word. At the very least, this analysis implies that the nature of the word-to-word relationships is a critical factor in determining the efficiency of relative-frequency based crosssituational learning.

## Discussion

In this paper, we studied cross-situational word learning from a theoretical perspective as the formation of a one-to-one word-object map. Our formulation of crosssituational learning is defined as learning on the basis of the relative frequency of objects for each word, which is a more realistic alternative model than eliminative learning, a model analyzed in past studies (Blythe et al., 2010, 2016) that is anyhow a special case of relative frequency learning. Thus, our analysis of relative frequency learning is both more general and more realistic than previously-proposed frameworks. Our analysis shows that its total learning time depends on the minimal difference between the most frequent and the second-most frequent objects among all the words, and that it is quite slow.

Given that relative frequency learning alone is inefficient, we next analyzed the case when the learner applies the lexical constraint that no two referents are associated to a single word. This principle of mutual exclusivity (ME) has been hypothesized to be an important means of reducing ambiguity for children learning language (Markman \& Wachtel, 1988; Markman, 1990, 1992), and empirical work has found that both children (Golinkoff et al., 1992; Halberda, 2003; Markman, Wasow, \& Hansen, 2003) and adults in cross-situational word learning experiments (Yurovsky \& Yu, 2008; Kachergis et al., 2012) show a preference for learning mappings consistent with ME. Using ME, a word can be learned via fast mapping (learned on its first sample), if all the distracting words appearing with it are already learned. However, the effect of ME on the average learning time is quite limited - the same (up to a constant multiplier) as that of independent relative frequency learning, if the distractors for each word are distributed uniformly. In summary, this analysis suggests that the order in which words are learned is related to the statistical nature of the word-to-word relationship-i.e., the structure among the cooccurring distractors.

Therefore, we finally analyzed the case in which a set of words is composed of two word groups: in one group, each word has no distractors, and in the other group each word has $k$ distractors, which are the words without any distractors. Here, it is not just a mixture of two types of words, but the distracting words have no
distractors to themselves, and thus they are likely to be learned earlier than the other group. Thus, in this schematic word structure, the expected learning order is correlated to the number of distractors for the group of words. We hypothesize that, with this statistical regularity, relative frequency learning can be as efficient as learning via fast-mapping, which has been observed in young children (Mervis \& Bertrand, 1994; GershkoffStowe \& Hahn, 2007). Our analysis suggests that this hypothesis is supported: the learning time is comparable with that of fast mapping learning up to a constant number of samples per word, when a certain ratio of words has no distractors. We expect that this analytic result can be extended to a more general case, such that there are multiple groups with different numbers of distractors up to $k$ and a group of words with $k$ distractors that has no distractors which have $k$ or more distractors against themselves.

In summary, we have analyzed a more general and more realistic class of word learning models, relative frequency learning. Although we showed that learning in this more general framework can be quite slow, we then examined learning under assumptions of mutual exclusivity and word-to-word correlations that might more closely approximate learning situations in the natural language environment. By modifying situations to include realistic variants of these two factors, we showed that learning a full-sized vocabulary could be accomplished on a realistic timescale. Although this work is preliminary, the analytical techniques employed here can be applied to other, yet more realistic cross-situational learning schemes, incorporating better approximations of the language environment, of the problem faced by the learner, and of the biases employed by the learner.

## Acknowledgments

This study is supported by the JSPS KAKENHI Grant-in-Aid for Young Scientists JP 16H05860.

## References

Akhtar, N., \& Montague, L. (1999). Early lexical acquisition: the role of cross-situational learning. First Language, 19, 34-358.
Bloom, P. (2000). How children learn the meaning of words. Cambridge, MA: MIT Press.
Blythe, R. A., Smith, A. D. M., \& Smith, K. (2016). Word learning under infinite uncertainty. Cognition, 151, 18-27.
Blythe, R. A., Smith, K., \& Smith, A. D. M. (2010). Learning times for large lexicons through crosssituational learning. Cognitive Science, 34(4), 620642.

Carey, S., \& Bartlett, E. (1978). Acquiring a single new word. Papers and Report on Child Language Development, 15, 17-29.
Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney
(Ed.), Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum.
Gershkoff-Stowe, L., \& Hahn, E. R. (2007). Fast mapping skills in the developing lexicon. Journal of Speech, Language, and Hearing Research, 50, 682-697.
Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., \& Wegner, N. R. (1992). Young children and adults use lexical principles to learn new nouns. Developmental Psychology, 28(1), 99-108.
Halberda, J. (2003). The development of a word-learning strategy. Cognition, 87(1), B23-B34.
Hidaka, S. (2014). General type token distribution. Biometrika, 101(4), 999-1002.
Kachergis, G., Yu, C., \& Shiffrin, R. M. (2012). An associative model of adaptive inference for learning wordreferent mappings. Psychonomic Bulletin and Review, 19(2), 317-324.
Markman, E. M. (1990). Constraints children place on word meanings. Cognitive Science, 14, 57-77.
Markman, E. M. (1992). Constraints on word learning: Speculations about their nature, origins and domain specificity. In M. R. Gunnar \& M. P. Maratsos (Eds.), Modularity and constraints in language and cognition: The minnesota symposium on child psychology (pp. 59-101). Hillsdale, NJ: Erlbaum.
Markman, E. M., \& Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20, 121-157.
Markman, E. M., Wasow, J. L., \& Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. Cognitive Psychology, 47(3), 241-275.
Mervis, C. B., \& Bertrand, J. (1994). Acquisition of the novel name - nameless category (n3c) principle. Child Development, 65, 1646-1662.
Pinker, S. (1984). Learnability and cognition. Cambridge, MA: MIT Press.
Quine, W. V. O. (1960). Word and object. Cambridge, MA: MIT Press.
Shannon, C. E. (1948). A mathematical theory of communication. Bell System Tech. Journal, 27, 379-423.
Siskind, J. M. (1996). A computational study of crosssituational techniques for learning word-to-meaning mappings. Cognition, 61, 39-91.
Vogt, P. P. (2012). Exploring the robustness of crosssituational learning under Zipfian distributions. Cognitive Science, 36(4), 726-739.
$\mathrm{Yu}, \mathrm{C} .(2008)$. A statistical associative account of vocabulary growth in early word learning. Language Learning and Development, 4(1), 32-62.
Yurovsky, D., \& Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In Proc. of cogsci 30. Austin, TX: Cognitive Science Society.

Zipf, G. (1949). Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley.

# Perceptions of Psychological Momentum in Basketball 

Thomas Hilbig ${ }^{1}$ (thomas.hilbig@student.adelaide.edu.au)<br>Matthew B. Welsh ${ }^{1,2}$ (matthew.welsh@adelaide.edu.au)<br>Paul H. Delfabbro ${ }^{1}$ (paul.delfabbro@adelaide.edu.au)<br>1. School of Psychology, 2. Australian School of Petroleum, University of Adelaide, Australia


#### Abstract

Psychological momentum (PM) and the hot hand are related concepts describing people's beliefs regarding streaks of superior performance. This study examined the susceptibility of perceptions of PM to changes in the streakiness of otherwise equivalent series. Fifty-five male participants (31 basketballers and 24 control) completed a 'hot-cognition' experiment where they rated individual and team momentum and assessed the likelihood of a future shot's success after watching sequences of basketball shots. The experimental manipulation of the order of shots strongly affected participants' ratings of momentum and, less strongly, the probability they assigned to the future shot (i.e. the hot hand effect). Basketballers showed stronger reactions to manipulations of order than the controls, which could be attributed to greater investment in the task. The results demonstrate the importance of distinguishing between PM and the hot hand and also provide a valuable extension of prior work showing such effects into more realistic scenarios.


Keywords: hot hand; psychological momentum; basketball.

## Introduction

The 'hot hand' is regarded as a crucial determinant of success by coaches (Raab, Gula, \& Gigerenzer, 2012), fans (Markman \& Guenther, 2007) and players (Gilovich et al., 1985) - with players altering the frequency and difficulty of their shot attempts after making a series of shots in a row.
Early research, however, mostly suggested that the hot hand in basketball was a 'fallacy', finding field and 'freethrow' shooting streaks did not significantly deviate from what was expected by chance (Gilovich, Vallone, \& Tversky, 1985). Conversely, some studies support the hot hand in intercollegiate (Mace, Lalli, Shea, \& Nevin, 1992) and professional basketballers, but some had issues of limited sample size and questionable method of analysis.

Failures to detect a hot hand, however, have also been questioned on several fronts. For example, the complexity of the basketball environment, wherein the 'hot' player may start to take lower probability shots due to their increased confidence or the opposing team may pay additional attention to a 'hot' player thereby disguising any effect.
Attempts to counter such objections include analysis of free-throws (e.g., Gilovich et al., 1985) but others (Koehler \& Conley, 2003) have argued that free throws are not conducive for a hot hand due to their relatively high probability of success ( $\sim 75 \%$ for professionals) and the time lag between free throw attempts for the same individual. In fact, given the hot hand is considered a temporary phenomenon (Hamberger \& Iso-Ahola, 2004), which breaks
disrupt (Mace et al., 1992), the conditions of ordinary NBA games - with time outs, substitutions and a single player rarely making 15 shots in a game - may not be conducive to its occurrence. This suggests that, if the hot hand exists, its existence is overgeneralized - that is, occurs less often than it is perceived to have. Indeed, Koehler and Conley's (2003) analysis of the National Basketball Association (NBA) Long Distance Shootout Contest - in which a shooter is unguarded but the available time and number of shots is constrained - failed to detect non-random shooting patterns, despite commentator's accounts to the contrary.

However, people have demonstrated an ability to discriminate between streaky and steady shooters in basketball shot sequences where statistical tests could not (Hammack, Cooper, Flach \& Houpt, 2017). While observers have the tendency to be sensitive to runs, this does not necessarily indicate cognitive error, but perhaps rational mechanisms for processing complex information. That is, observers will act as if the hot hand exists and they are capable of accurately perceiving and harnessing its effects.

Iso-Ahola and Mobily (1980) proposed psychological momentum (PM) as a construct to account for these perceptions and subsequent behaviors: "an added or gained psychological power that changes a person's view of him/herself or of others, or others' views of him/her and themselves" (p. 392). In competitive scenarios, PM is a zero-sum game: obtained at the expense of a competitor.

Importantly, PM does not reflect superior performance i.e. a hot hand, as suggested by Avugos and Bar-Eli (2015), but rather a psychological phenomenon (Iso-Ahola \& Dotson, 2015). The key distinction stems from an individual experiencing improved neurophysiological performance, as opposed to just changes in psychological components (e.g. confidence, internal attributions, perceived superiority over opponents). For example, an athlete may experience improved belief in their ability due to previous success, but not have this result in meaningful changes in skill execution.

Furthermore, it cannot be assumed that improvements to confidence necessarily result in a greater probability of subsequent success. As noted above, initial success could lead to 'over-confidence', causing athletes to make riskier decisions than normal (Jones \& Harwood, 2008).

With this in mind it is important to distinguish between psychological momentum, the perception of hot hand effects, and actual hot hand effects in experimental tasks.

Examining the thoughts and attitudes of athletes during live play is, of course, unfeasible. Accordingly, this study focusses on the perception of PM by spectators and the
implications of this for their predictions/behaviors - that is, their expectations regarding the effects of such psychological momentum (e.g., the Hot Hand). Previous studies examining people's perceptions about sequences have presented hypothetical scenarios (e.g. Ayton \& Fischer, 2004), but context is important to making inferences about sequences (Matthews, 2013) and actual sporting experiences are thought to be more conducive to perceptions of PM (Jones \& Harwood, 2008). Therefore, the intent was to maximize participant engagement, without the difficulties arising from assessing participants during live play. A 'hot cognition' experiment was, therefore, devised wherein both basketballers and non-basketballers watched actual footage of basketball games with varied presentation of sequences of successful and unsuccessful shots.

## Aims and Hypotheses

1. Reordering series of basketball plays with the same number of successes and failures will alter the psychological momentum assigned by observers to a featured team/player.
2. This will alter the probability assigned to a future outcome following the observed sequence (Hot Hand).
3. Basketballers, with greater investment in the game, will react more strongly to manipulations of momentum.

## Method

## Participants

Participants were 55 male, English speakers with at least a basic understanding of basketball rules and terminology, aged 18-31 ( $M=21.4, S D=3.2$ ) and recruited from three sources: local basketball clubs ( $\mathrm{n}=22$ ), $1^{\text {st }}$ year Psychology students ( $n=7$ ) and the general public ( $n=26$ ). Participants were grouped as basketballers ( $N=31$ ) or control $(N=24)$ by their self-reported frequency of involvement in basketball. The basketballers were somewhat younger ( $M=$ 20.6, $S D=2.6$ ) than the control group ( $M=22.4, S D=3.7$ ).

The psychology students participated for course credit. Additional participants, recruited via emails to basketball clubs, flyers posted on the Adelaide University campus and Facebook advertising, received a $\$ 10$ gift card for their participation or chance to win a $\$ 50$ gift voucher.

## Materials

## Online Survey

Prepared in SurveyMonkey, the survey asked for demographic details and required participants to indicate how often (daily, several days a week, weekly, fortnightly, monthly and rarely/never) they engaged with various aspects of basketball: playing, watching or taking an interest in (e.g., reading about). The survey also included measures of: representation bias, numerical reasoning, perception of sequences, susceptibility to outcome bias, risk-attitudes, impulsivity and hot hand beliefs as described below. These were included as potential covariates/confounds that might differ between the groups and thus need to be controlled for:
Representativeness Bias. Four items were used to assess
respondent's beliefs about sequences in random processes: Lambos, Delfabbro and Puglies' (2007) coin toss scenario, where participants judge which of three series of outcomes (e.g. HTHTTHTHTHTH) is most likely; and three items adapted from Ayton and Fischer (2004) asking whether sequences of 16 -digit long binary outcomes with equal hits and misses but different alternation rates $(0.81,0.31$ and $0.19)$ were generated by random or human processes.

Cognitive Reflection. Frederick's (2005) 3-item CRT was used to measure participants' tendency to override predictable, but incorrect intuitive responses. Lower CRT scores indicate greater susceptibility to decision-making biases (Toplak, West \& Stanovich, 2011) and lower numeracy (e.g., Welsh, Burns \& Delfabbro, 2013).

Outcome Bias. Two scenarios described a physician's decision to conduct surgery on a suffering patient (based on Baron \& Hershey, 1988). These were near identical but the first described an $8 \%$ chance of death but a good outcome (successful operation) while the second gave a $2 \%$ chance of death but bad outcome (patient death). Rating the $1^{\text {st }}$ decision as better therefore displays outcome bias.

Risk Attitude. As belief in a hot hand is greater in those who regularly gamble and demonstrate a willingness to take greater risks in these scenarios (Wilke, Scheibehenne, Gaissmaier, McCanney \& Barrett, 2014), the 12-item gambling Domain-Specific Risk-Attitude Scale (Weber, Blais, and Betz, 2002) was used to assess risk attitudes towards: likelihood of gambling; perception of gambling risk; and expected benefits of gambling.

Impulsivity. The BIS-15 (Spinella, 2007) was used to assess impulsivity.

Hot Hand Beliefs. A 2-item, self-report measure developed by Gilovich, Vallone and Tversky (1985), which assesses a respondent's endorsement of sequential dependence among shots in basketball.

## Hot Cognition Task

The experimental task was composed of four sequences of basketball edited from footage of American college basketball games. The broadcast scoreboard was blurred out in the video footage to control for outcome bias, and the audio was removed to prevent crowd and broadcast commentator reactions influencing participant's responses. All the plays within a sequence featured the same player.

Each sequence condition included three made shots ( $H$ ) and three missed shots ( $M$ ) but a different order of shot outcomes (see Table 1). All were followed by the identified player being fouled in the process of making another, successful shot: resulting in a free throw (AND1 outcome), but the outcome of this was not shown. Given the absence of audio, the researcher indicated when the identified player was fouled - signaling the end of the video.

Looking at Table 1, one sees that the conditions convey varying senses of psychological momentum (PM). The Positive Recency (PR) and Negative Recency (NR) conditions have low alternation rates of hits versus missed shots and, therefore, longer outcome runs. Prior to the

AND1 outcome, the PR sequence contains a streak of three hits, while the NR sequence contains a streak of three misses to convey lower momentum (although weakened by the need to have the same successful, fouled basket at the end of the sequence). The two other conditions were intermediate between these - with greater alternation.

Table 1: Shot sequence outcomes for experimental conditions.

|  | Condition | Order of Shot sequence |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 1 | , | 3 | 4 | 5 | 6 | 7 |
|  | Negative Recency | H | H | H | M | M | M | H |
|  | Alternation | H | M | H | M | H | M | H |
|  | Weak Positive |  | H | M | H | M | H | H |
|  | Positive Recency | M | M | M | H | H | H |  |

Note: $\mathrm{H}=$ hit, $\mathrm{M}=$ miss.

## Procedure

Participants completed the survey detailed above prior to taking part in the experiment - either online or in person. The experiment was conducted individually for each participant to avoid confounds arising in groups (e.g. verbal commentary influencing responses). Participants were provided information regarding the nature of the experiment and screened for (basic) understanding of basketball rules and terminology used for various self-report measures.
The experiment was conducted within-subjects, with participants shown the four, Hot Cognition Task sequences (in a randomized order). Following each, participants were asked 4 questions assessing their beliefs around:

1) a player's likelihood of making the free throw resulting from the last play (Free Throw);
2) the player being 'on a roll' (Individual Momentum);
3) how difficult his shots were (Difficulty); and
4) the team having momentum (Team Momentum).

Responses were scored on a 5-point Likert scale: 1 (not at all) to 5 (extremely). These questions - based on a pilot study and previous qualitative research e.g. Koehler \& Conley (2003) - measure the participant's perception of individual and team momentum, and perceived difficulty due to theorized mediation effects.

## Results

The first two Hypotheses were that perceptions of psychological momentum would vary with the patterns of hits and misses in the four different conditions. To examine this, the mean ratings given by participants to each of the four dependent measures under each of the four conditions are shown in Figure 1. Looking at this figure, a clear distinction can be seen between the pattern of results for the measures of psychological momentum (Individual and Team Momentum) and the remaining measures - Free Throw likelihood and Shot Difficulty. Starting with the last it seems that, as would be hoped, participants' perceptions of shot difficulty did not vary across conditions in any obvious manner. A One-Way RM ANOVA, however,
indicated that the differences across conditions were significant, $\mathrm{F}(3,162), p<.001$, indicating that Shot Difficulty needed to be included as a covariate in the analyses described below. Analysis of the Free Throw ratings look similar, $\mathrm{F}(3,162), p<.001$, but here Bonferroni post hoc tests confirmed that two positive conditions (Weak Positive and Positive Recency) produced significantly higher ratings than the conditions with more recent negative outcomes.


Figure 1. Mean ratings of dependent measures by condition. Note: NR = negative recency; Alt = alternating; WP = weak positive recency; and $\mathrm{PR}=$ positive recency.

By contrast, the measures of individual and team momentum both show clear, linear trends with participants giving higher ratings in those conditions with more, recent positive outcomes. One-way Repeated Measures ANOVAs confirmed these differences as significant $\mathrm{F}(3,162)=63.4$ \& 26.7, respectively, $\mathrm{p}<.001$ in both cases and Bonferroni post hoc tests indicated that all conditions differed significantly from all others.

## Covariates

Correlational and principal component analyses (excluded for reasons of space) were used to determine which covariates should be accounted for in comparisons between basketballers and controls. This indicated only five variables/factors related significantly to the dependent measures: 1) Representation bias (coin toss); 2) Representation bias (high alternation rate); 3) CRT; 4) Outcome bias; and 5) Outcome perceptions (a factor composed of beliefs about hot hand and momentum).

## Basketballers vs Non-Basketballers

Our third hypothesis was that basketballers, due to their relative investment in the sport, thus responding to the experimental manipulations more strongly. That is, that their ratings would tend to be more extreme than nonbasketballers - lower in the conditions with more negative
outcomes and/or higher in conditions with more positive outcomes prior to the final observation.

To examine potential differences, the ratings provided by the two groups for the dependent measures are shown in Figure 2. Looking at the three subplots of Figure 2, one sees two distinct patterns. The first is in the Free Throw data (subplot a), where, in every condition, the basketballers rate the likelihood of the free throw being successful as higher than the non-basketballers - reflecting perhaps a better understanding of the difference in accuracy between field shooting and free throw shooting.

In subplots b and c, by comparison, we see the pattern predicted by Hypothesis 3 - with Basketballers' responses being more extreme than control subjects - i.e., lower when there has been a run of missed shots (NR condition) and higher following a series of successful shots (PR condition).


Figure 2. Comparisons between Basketballer and Control subject ratings of dependent measures by condition. Note:
$\mathrm{NR}=$ negative recency; Alt $=$ alternating; $\mathrm{WP}=$ weak positive; and $\mathrm{PR}=$ positive recency.

Group by Condition Repeated Measures ANCOVAs were run for each dependent measure, incorporating the covariates noted above. The results of these are shown in Table 2. Looking first at the data for Free Throw probability, one sees that, despite the pattern in Figure 2, the main effect of Group in the ANCOVA just fails to reach significance ( $p=.07,2$-tailed). The effect of condition was clearly non-significant ( $p=.84,2$-tailed) and there was no interaction between the two factors.
For Individual Momentum perceptions, by contrast, a significant main effect was found for condition ( $p<.001,2$ tailed) but not between groups ( $p=.67,2$-tailed). However, in line with our hypothesis, there was a significant Group $\times$ Condition interaction ( $p=.03,2$-tailed). Bonferroni post hoc tests indicated that Basketballers perceived more individual momentum than the Controls in the $\mathrm{PR}(d=0.16)$ and WP $(d=0.30)$ conditions, and perceived less in the NR
( $d=0.14$ ) and Alt ( $d=0.20$ ) conditions. There was also a significant covariate interaction: Condition $\times$ Representation bias (high alternation), $\mathrm{F}(3,46)=2.60, \mathrm{p}<.05$. This suggests that Individual Momentum is predicted by susceptibility to attribute random outcomes to human action.
For Team Momentum, a significant difference was found between groups ( $p<.05,2$-tailed), but the main effect of condition and the Group $\times$ Condition interaction just failed to reach significance ( $p=.07$ and .13 , respectively, 2 tailed). Given the directionality of our hypotheses, these near significant results were examined with post hoc Bonferroni tests, which indicated that Basketballers perceived less momentum than the control group in the NR ( $d=0.25$ ) and Alternation conditions $(d=0.35)$, but no more in the PR $(d=0.11)$ and WP $(d=0.04)$ conditions, which partially support the hypothesis.

Table 2: Summary of ANCOVAs for dependent measures

|  | Group |  |  |  | Condition |  |  |  | Interaction |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $F$ | $p$ | $\eta^{2}$ | $F$ | $p$ | $\eta^{2}$ | $F$ | $p$ | $\eta^{2}$ |  |  |
| FT | 3.48 | .07 | .07 | 0.29 | .84 | .01 | 0.07 | .98 | .00 |  |  |
| IM | 0.19 | .67 | .00 | 6.89 | $<.001$ | .13 | 3.02 | .03 | .06 |  |  |
| TM | 4.12 | .05 | .08 | 2.60 | .07 | .05 | 2.01 | .13 | .04 |  |  |

Note: FT = Free throw, $\mathrm{IM}=$ Individual momentum, $\mathrm{TM}=$ Team momentum. Greenhouse-Geisser corrections applied. Degrees of Freedom. Two-tailed $p$ values in all cases.

## Discussion

The above results provide support for all three Hypotheses. There is strong evidence that reordering the same number of successful and failed basketball plays to produce streaks of hits and misses affected the participants' perceptions of psychological momentum - for both the individual player and their team (H1). The evidence that this perception of momentum translates into a belief in a hot hand in the statistical sense (H2) - that is, altering the probability of a future shot is, however, weaker. Finally, there are significant differences between the responses of basketballers and non-basketballers and clear interaction effects between group membership and the strength of our psychological momentum manipulation (H3). These results are discussed, individually and in greater detail, below.

## Perception of Psychological Momentum

As noted above, Hypothesis 1, that perceptions of psychological momentum (PM) could be influenced by simple reordering of sets of basketball plays, was supported by the results. Specifically, participants rated the momentum of both the individual player and their team as significantly higher when the plays were ordered so as to have longer strings of hits at the end.

Parker, Paul \& Reinholtz (2016) similarly found that changes in perceived momentum of a contrived guessing game were greater as outcomes alternated. While perhaps not surprising, building upon the findings from hypothetical
manipulations is a valuable extension of such work demonstrating that the effect holds in a task more closely approximating real world situations but which removed a number of cues for momentum that would exist in realworld situations. For example, the broadcast scoreboard in the game footage shown to participants was blurred to control for potential outcome bias and prevent the score being used as a reference point by participants regarding a team's actual momentum. The footage was also played without sound so as remove the crowd reaction which might provide another cue to a team's momentum.
While these were confounds for the present study, their omission is also expected to have dampened the extent to which participants identified patterns of team momentum. That is, in equivalent, real world situations their effects seen here might well be stronger.

## Expectations Regarding the 'Hot Hand'

Given the clear distinctions drawn by participants between the momentum of the individual players and their teams across the different order conditions, the weaker effect of the experimental manipulation on their predictions of future success requires some explanation. While there was some evidence that people who had seen longer sequences of successful shots tended to rate the probability of the following free throw succeeding more highly than those who had seen more failures at the end of the task, this relationship was weak and non-significant when controlling for several covariates and did not follow the clearly linear pattern seen in the perceptions of momentum.
Had no relationship been seen, that could have supported the notion that PM is just a performance label used to evaluate whether past performances were successful or not (Cornelius et al., 1997) with no relevance to the future. The partial relationship, however, requires more explanation.

A possibility is that the experimental task, which had participants watch a series of seven field shots but then asked them to rate the likelihood of a following free throw being successful acted to limit the perceived transferability of momentum. That is, not only does the foul and subsequent free throw provoke a break in play (thereby potentially ending a hot hand effect, as described in Hamberger \& Iso-Ahola, 2004) but also introduces a change in the type of task. Participants may have recognized that a free-throw is a markedly different shot than any field goal attempt and thus, regarded the player's shooting form as less relevant, reducing the strength of any effect.

## Group Differences

As predicted in our third hypothesis, basketballers' responses differed significantly from those of nonbasketballers. The first observation, while not hypothesized, is that basketballers rated the chance of the free throw being successful higher in every case than the non-basketballer, reflecting their superior understanding of the actual success rates for elite level athletes. Other than this, though, their pattern of free throw predictions across the four conditions
is near-identical to that of the non-basketballers.
Of course, the fact that results supporting Hypothesis 3 are seen for perceptions of individual and team momentum - with basketballers being more strongly affected - but this fails to be converted into greater predicted likelihood of free throw success could fit with explanation given above regarding the overall weakness of these results. That is, if basketballers have a stronger belief in the separation between field shooting and free throw shooting performance, that would tend to flatten out their estimates of free throw likelihood more than is seen in the nonbasketballers - thereby counteracting their stronger perceptions of momentum.

As to why basketballers showed these stronger effects: perhaps simple interest in the game increases investment and thus cues greater attention to the scenario and patterns within it; or seeing such patterns calls to mind prior experiences of momentum and, within the experimental context, basketballers have more than non-basketballers.

## Covariates

As a brief note: as ANCOVAs were used to eliminate the possibility that results might be driven by differences between the groups. In these analyses, one covariate (a measure of representativeness bias) was highlighted as predicting individual momentum ratings: i.e., participants who attributed random sequences to human agents were more likely to rate the player as being 'on a roll'. However, none of the four measures of representativeness bias differed significantly between the groups.

## Caveats and Future Directions

While providing interesting results and at least some support for all of our hypotheses, there are a number of limitations of the study, which could be addressed in future work.

The first is the limited sample size. This resulted from difficulties in recruiting sufficient basketballers and was exacerbated by the decision to limit recruitment to males so as to eliminate the potential for gender moderation effects of PM (Iso-Ahola \& Dotson, 2014). Extending the study to include women (while taking into account how perceptions of may PM differ between men and women) and widening the recruitment (via online participation, for instance) could address this and assist in determining whether the less convincing results herein result from insufficient power.

A secondary concern lies in how the experimental measures were scored - on a 5-point scale from 'not at all' to 'extremely'. The question arising here is whether 'not at all' was regarded as a neutral (e.g., not on a roll) or negative (e.g., on a losing streak) response by participants. This should be clarified in future work.

Participants' responses to the dependent variables may have been influenced by undertaking several cognitive bias measures. To avoid this potential confound the hot cognition experiment could be conducted prior to completing relevant individual differences measures.

Finally, as noted above, while the use of a free throw as the shot to be predicted was done purposely - in order to
minimize other contextual factors for the predicted shot (e.g. in any differences in distance, angle, and opposition actions) - this may have undermined the transference of perceived momentum into future outcomes. This could be avoided in a number of ways - none, however, simple. For example, an exhaustive pre-test assessment of plays could have experts rate their equivalency prior to constructing the scenarios.
Alternatively, it might be possible to stage specific plays - either using real players or within a basketball game, for example. All of these, however, would require a significant amount of pilot work prior to any experimentation.

## Conclusions

While most research into momentum and the hot hand has been concerned with directly substantiating or refuting their existence, the present study aimed, instead, to explore participant's beliefs and perceptions regarding these - within the context of basketball shot sequences. Consistent with much of literature regarding PM, the ordering of sequential outcomes affected participant perceptions: positive recency sequences increased the likelihood that the focal team and player had momentum; while negative recency sequences were considered by participants as evidence of the player and team not being on a roll. These effects were more strongly reported by basketballers across conditions in the present study, compared to the control group, suggesting that domain-specific experience influences the perception of these patterns. The results also illustrate that further research is warranted to clarify why differences exist between participant's perceptions of momentum and their predictions of future success (i.e., hot hand beliefs).

## Acknowledgements

MBW thanks Santos and Woodside for their support of the CIBP at the Australian School of Petroleum.

## References

Avugos, S., \& Bar-Eli, M. (2015). A second thought on the success-breeds success model: Comment on Iso-Ahola and Dotson (2014). Review of General Psychology, 19, 106-111.
Ayton, P., \& Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: two faces of subjective randomness? Memory and Cognition, 32, 1369-1378.
Baron, J., \& Hershey, J. (1988). Outcome bias in decision evaluation. Journal of Personality and Social Psychology, 54(4): 569-579.
Cornelius, A., Silva, J. A., Conroy, D. E., \& Petersen, G. (1997). The projected performance model: Relating cognitive and performance antecedents of psychological momentum. Perceptual and Motor Skills, 84, 475-485.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 25-42.
Hamberger, M., \& Iso-Ahola, S. (2004). Psychological momentum and athletic performance: A critical review of research. Journal of Contemporary Athletics, 1, 207-226.

Hammack, T., Cooper, J., Flach, J. M., \& Houpt, J. (2017). Toward an Ecological Theory of Rationality: Debunking the Hot Hand "Illusion". Psychology, 29(1), 35-53.
Iso-Ahola, S., \& Dotson, C. (2014). Psychological momentum: Why success breeds success. Review of General Psychology, 18, 19-33.
Iso-Ahola, S. E., and Dotson, C. O. (2015). Psychological momentum-Not a statistical but psychological phenomenon. Rev. Gen. Psychol. 19, 112-116.
Iso-Ahola, S., \& Mobily, K. (1980). Psychological momentum: A phenomenon and empirical (unobtrusive) validation of its influence in sport competition. Psychological Reports, 46, 391-401.
Jones, M. I., \& Harwood, C. G. (2008). Psychological momentum within competitive soccer: Players' perspectives. Journal of Applied Sport Psychology, 20, 57-72.
Koehler, J. J., \& Conley, C. A. (2003). The "hot hand" myth in professional basketball. Journal of Sport \& Exercise Physiology, 25, 253-259.
Lambos, C., Delfabbro, P., \& Puglies, S. (2007). Adolescent Gambling in South Australia. Adelaide, South Australia: Department for Education and Children's Services for the Independent Gambling Authority of South Australia.
Mace, F., Lalli, J., Shea, M., \& Nevin, J. (1992). Behavioral momentum in college basketball. Journal of Applied Behavior Analysis, 25, 657-663.
Markman, K., \& Guenther, C. (2007). Psychological momentum: Intuitive physics and naïve beliefs. Personality and Social Psychology Bulletin, 33, 800-812.
Matthews, W. J. (2013). Relatively random: Context effects on perceived randomness and predicted outcomes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 1642-1648.
Parker, J. R., Paul, I., \& Reinholtz, N. (2016). Perceived
Momentum Influences Responsibility Judgments. Proceedings of the 38th annual conference cognitive science society.
Raab, M., Gula, B., \& Gigerenzer, G. (2012). The hot hand exists in volleyball and is used for allocation decisions. Journal of Experimental Psychology: Applied, 18, 81-94.
Spinella, M. (2007). Normative data and a short form of the Barratt Impulsiveness Scale. International Journal of Neuroscience, 117, 359-368.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. Memory \& Cognition, 39, 1275-1289.
Welsh, M., Burns, N., \& Delfabbro, P. (2013). The Cognitive Reflection Test: how much more than Numerical Ability? Cognitive Science Society (pp. 15871592). Austin, TX: Cognitive Science Society.

Wilke, A., Scheibehenne, B., Gaissmaier, W., McCanney P., \& Barrett, H.C. (2014). Illusionary pattern detection in habitual gamblers. Evolution and Human Behavior, Volume 35, Issue 4, 291-297.

# Translating a Reinforcement Learning Task into a Computational Psychiatry Assay: Challenges and Strategies 

Peter Hitchcock (pfh26@drexel.edu)<br>Clinical Psychology, Department of Psychology, Drexel University<br>Angela Radulescu (angelar@princeton.edu)<br>Department of Psychology, Princeton University<br>Yael Niv (yael@princeton.edu)<br>Department of Psychology, Princeton University<br>Chris R. Sims (chris.sims@drexel.edu)<br>Applied Cognitive \& Brain Sciences, Department of Psychology, Drexel University


#### Abstract

Computational psychiatry applies advances from computational neuroscience to psychiatric disorders. A core aim is to develop tasks and modeling approaches that can advance clinical science. Special interest has centered on reinforcement learning (RL) tasks and models. However, laboratory tasks in general often have psychometric weaknesses and RL tasks pose special challenges. These challenges must be addressed if computational psychiatry is to capitalize on its promise of developing sensitive, replicable assays of cognitive function. Few resources identify these challenges and discuss strategies to mitigate them. Here, we first overview general psychometric challenges associated with laboratory tasks, as these may be unfamiliar to cognitive scientists. Next, we illustrate how these challenges interact with issues specific to RL tasks, in the context of presenting a case example of preparing an RL task for computational psychiatry. Throughout, we highlight how considering measurement issues prior to a clinical science study can inform study design.


Keywords: computational modeling; reinforcement learning; measurement; psychometrics; computational psychiatry

A core aim of the emerging field of computational psychiatry is to translate tasks and modeling approaches from computational neuroscience into sensitive assays that can advance clinical treatment, diagnosis, practice, and theory (Hitchcock, 2017; Redish \& Gordon, 2016). New assays may advance clinical science by facilitating early illness detection, predicting illness progression, separating patients into subgroups, predicting type and extent of treatment indicated, and allowing measurement of the effects of emotion regulation strategies (Huys, Maia, \& Frank, 2016).
The effort to develop laboratory tasks into assays has been ongoing for years, but the use of computational cognitive models that describe the trial-by-trial behavior of subjects (Daw, 2011) is newer to clinical science. In theory, parameters derived from these models should compactly describe individual or group differences by revealing aspects of cognitive processing that are obscured in behavioral measures (Huys et al., 2016). An especially promising domain in this regard is reinforcement learning (RL). RL refers to a broad class of trial-and-error learning tasks wherein learning is driven mainly by a scalar reinforcement
signal (Sutton \& Barto, 1998). Over the past twenty years, computational models of RL have grown in sophistication and maturity (O’Doherty, Cockburn, \& Pauli, 2017). In addition, there has been a string of successful applications of RL modeling to clinical problems. These early successes may portend widespread use of RL assays in clinical science (Maia \& Frank, 2011).
Yet the history of converting laboratory tasks to clinical assays suggests caution is warranted. Laboratory tasks tend to have substantial (and often underappreciated) psychometric weaknesses (Lilienfeld, 2014). Consider the example of the dot probe task, an attention paradigm introduced over 30 years ago (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, \& Van Ijzendoorn, 2007). By 2007, 35 clinical studies using the task had been conducted. A metaanalysis that year concluded the task reliably detects attention differences between anxious and non-anxious groups (BarHaim et al., 2007). Dozens of studies subsequently tested "modification" variants of the task (which aim to retrain attention) (Hallion \& Ruscio, 2011). Yet recent metaanalyses suggest modification training produces very small effects and that extant modification studies evince publication bias (e.g., Heeren, Mogoașe, Phillippbot, \& McNally, 2015). These disappointing results prompted reexamination of the evidence for reliable, stable group differences per the original dot probe. Recent critiques, which have referenced a slew of null findings since 2007, concluded that the evidence for such differences is weak (Rodebaugh et al., 2016; Van Bockstaele, Verschuere, Tibboel, De Houwer, Crombez, \& Koster, 2014).
What went wrong? It is noteworthy that, although researchers have been employing the original dot probe since the 1980s, the first examination of its test-retest reliability was not published until 2005 (Schmukle, 2005). That study and others (e.g., Price et al., 2015) found the dot probe exhibits close to 0 test-retest reliability when analyzed using standard methods. These results suggest it is not possible to extract stable measures of differences in attention using the standard versions/analyses of the task (Rodebaugh et al., 2016; Van Bockstaele et al., 2014).


Figure 1. Pipeline for a computational psychiatry assay proposed by Paulus et al. (2016).

## Developing Computational Psychiatry Tasks with Strong Psychometric Properties

The dot probe paradigm provides a cautionary tale about pushing too quickly from a lab paradigm to applied research. The computational psychiatry community can learn from this example. Fortunately, the community appears aware of the challenges posed by laboratory tasks. For instance, Paulus, Huys, and Maia (2016) proposed a pipeline (Figure 1) for turning a task into an assay that can ultimately be used for assessment or as a treatment target in randomized control trials (RCTs). The authors emphasize establishing psychometric properties early in the pipeline-before relying on the task as a primary measure in RCTs.

Yet researchers entering computational psychiatry from the cognitive sciences may be unfamiliar with how the psychometric challenges of laboratory tasks interact with clinical design issues. Thus, this paper offers an overview of the relevant issues. Specifically, the rest of paper is part theoretical overview and part annotated case example of preparing a specific RL task for use in clinical science. We begin with a theoretical issue that may be unfamiliar to many in cognitive science.
General Psychometric Challenges Associated with Laboratory Tasks. A general-and formidable-challenge for extrapolating from laboratory task behavior is that subjects naturally vary in the state that they are in (e.g., tired, distraught, cognitively taxed) when they arrive at the laboratory. A classic solution to this random state variation problem-a problem that confounded social and personality psychologists for decades (Kenrick \& Funder, 1988)—is to assess the same subject at many time points and average over measurements. This approach can dramatically increase the convergent validity of lab tasks with self-report measures, presumably because an average over many time points yields a more stable, trait-like measure than one-time measurement (as the latter is often biased by state variation; Epstein, 1979).

However, assessing a single subject at many time points can be infeasible. First, much time is often needed to complete lab tasks, and thus repeating assessments on many occasions can substantially raise subject burden. Second, in some cases it is unrealistic to ask subjects to complete the task more than once. For example, a researcher may wish to examine how


Figure 2. The Dimensions Task (Niv et al., 2015), designed to investigate the role of attention in reinforcement learning.
depressive rumination-repetitive, negative, self-referential thinking-alters cognitive processing. This could be done by asking depressed subjects to complete a laboratory task while under the effects of a rumination induction. Many past studies have experimentally induced rumination in this way but, as far as we are aware, none has asked subjects to ruminate on more than one occasion. Indeed, it seems unreasonable and unrealistic to ask depressed subjects to undergo more than once a manipulation that-by design-provokes distress.
The effects of random state variation can be mitigated through study design. For example, a researcher investigating the effects of rumination on some task might ask subjects to perform the task once before and once while under the effects of rumination. Such a design should increase the ratio of systematic variability (variability due to induced rumination) to unsystematic variability (variability due to subjects being in different states when they enter the lab) because it delivers pre- as well as post-induction measures for each subject. A subject is unlikely to dramatically change the state she is in from pre- to post-induction (an unusually tired subject at baseline will likely remain so while under the effect of rumination induction). Thus the within-subject design controls for some of the unsystematic variability due to state variation. However, note that individual differences in susceptibility to the experimental perturbation (e.g., propensity to ruminate upon receiving the induction) will be affected by subjects' states. Thus this approach is helpful in minimizing noise but does not solve the random state variation problem.
The random state variation problem entails that a subject's parameter estimates in a laboratory measure will be corrupted by noise with respect to the subject's "true" parameter value, when the true parameter value is conceived of as a psychological variable akin to a trait. This noise will limit the predictive power of measures. Thus, when random state variation is expected (e.g., when a design only permits administering the task once or a few times), it is critical that the psychometric properties of the task are strong so that other sources of noise are minimized. For a more general discussion of how computational modeling may help remedy the random state variation problem, see Hitchcock (2017).
In the rest of this paper, we give a case example of preliminary efforts to establish the psychometric properties of a multidimensional RL task known as the Dimensions Task (Niv et al., 2015; Leong, Radulescu, Daniel, DeWonskin, \& Niv, 2017; Radulescu et al., 2016). The task itself is not the paper's focus, but we briefly describe it, our approach to modeling it, and its promise for clinical science in the next
section. The description will make subsequent sections, on the task's measurement properties and their relation to modeling issues, easier to follow.
The Dimensions Task. Trial-and-error learning in the real world often requires learning about a small set of stimulus features embedded in a milieu of irrelevant stimuli. Imagine telling (what you hope is) an amusing story to a friend and attempting to learn about the effects of specific actionsdramatic pauses, rhetorical flourishes, funny faces, etc. Learning about the effect of these actions requires attending to just a few fleeting features on the face of and in the body language of your friend while ignoring many irrelevant features-pimples on your friend's forehead, your computer screen flickering behind you, your internal dialogue about what to say next, etc. (Niv et al., 2015).
The Dimensions Task was designed to study such a scenario where only some aspects of the task are relevant and most can be ignored, as is so often required in the real world. Briefly (see Niv et al., 2015 for details), on each trial of the task subjects must select one of three possible stimuli. Each stimulus is composed of 3 features defined on 3 stimulus dimensions (for example, color, shape, and pattern) (Figure 2). Subjects play a set of games that can vary in length from 15-30 trials. Within a game, features of only one dimension (e.g., color) determine the probability of reward. Within this relevant dimension, one target feature (e.g., red) leads to reward with $75 \%$ chance whereas the other 2 features in the dimension (e.g., yellow, green) lead to reward with $25 \%$ chance. The target feature and relevant dimension change every game. The start of a new game is signaled to subjects.
Computational Model. Previous work (e.g., Niv et al. 2015; Radulescu et al., 2016) tested various computational models designed to reproduce subjects' trial-by-trial behavior in the task and found that human behavior is well described by a feature-level $R L(f R L)+$ decay model. The $f R L+$ decay model maintains weights reflecting the values of each of the 9 features. It linearly sums these weights to calculate the estimated value of each (3-feature) stimulus

$$
\begin{equation*}
V(S)=\sum W(f) \quad \forall f \in S \tag{1}
\end{equation*}
$$

For example, the model's estimate of the value of yellow-waves-triangle in the above trial is equal to the sum of the weights of yellow, waves, and triangle. Once a reward is received ( 0 or 1 points), the weights of the 3 features of the selected stimulus are updated based on the discrepancy between the obtained reward, $R_{t}$, and the model's estimate of the chosen stimulus's value, $V\left(S_{\text {Chosen }}\right)$, with update rate controlled by a learning rate free parameter, $\eta$

$$
\begin{equation*}
W^{\text {new }}(f)=W^{\text {old }}(f)+\eta\left[R_{t}-V\left(S_{\text {chosen }}\right)\right] \quad \forall f \in S_{\text {chosen }} \tag{2}
\end{equation*}
$$

For the other 6 features on a trial-those comprising the 2 stimuli not selected-the model decays the associated weights with a second free parameter, $d$

$$
\begin{equation*}
W^{\text {new }}(f)=(1-\mathrm{d}) W^{\text {old }}(f) \quad \forall f \notin S_{\text {chosen }} \tag{3}
\end{equation*}
$$

The decay parameter reflects the fact that subjects are selectively attending to (and learning about) few dimensions (Leong, Radulescu et al., 2017). The "forgetting" of the weights of unchosen features allows the model to "undo" learning about features not chosen on a trial.
Finally, the model assumes that the subject's probability of choosing each stimulus is proportional to the estimate of the value of the stimulus, as defined by a softmax equation with a third free parameter, $\beta$

$$
\begin{equation*}
p\left(\text { choose } S_{i}\right) \propto e^{\beta V\left(\mathrm{~S}_{i}\right)} \tag{4}
\end{equation*}
$$

The model thus has three free parameters: softmax action selection noise $\beta$, learning rate $\eta$, and decay parameter $d$. See Niv et al. (2015) for more details.
Stage in the Assay Development Pipeline. With respect to Paulus et al.'s (2016) pipeline (Figure 1), most prior studies using the Dimensions Task and $f R L+$ decay model fall into the Preclinical and Phase1a phases.

Notably, Radulescu et al. (2016, study 2) also provided a test of the task's promise for measuring group differences. Radulescu and colleagues found older adults were less accurate ( $p=.001, g=.94$ ) than younger adults. These behavioral results appeared to derive in part from differences in the decay parameter (median $=.52 \mathrm{v} .42$ for older vs. younger adults, respectively), implying that differences in this parameter may reflect meaningful differences in selective attention. These results suggest the task has promise as a sensitive measure of neuropsychological and clinical differences. Per Paulus et al.'s (2016) pipeline, this study marks the entrance into Phase 1 b : examining clinical validity (see Radulescu et al., 2016 for discussion).

Although the task has promise as a computational psychiatry assay, a number of modeling and psychometric obstacles must first be overcome. In the following sections, we report on efforts to explore the properties of the Dimensions Task and fRL+decay model using two previously collected datasets. The results have implications for the use of the Dimensions Task in computational psychiatry and thus are of specific interest to researchers interested in the construct of attention learning in computational psychiatry. But the more general interest aim of the following sections is to use this case study to illustrate some of the issues that arise in translating RL tasks to computational psychiatry.

## Methods

Datasets are from Niv et al. (2015; hereafter D1) and Radulescu et al. (2016, study 2; hereafter D2).
Specifications. In D1 $(N=22)$, subjects played 500 trials (number of trials per game was drawn from a Uniform $(15,25)$ distribution, for a total of $M=22.27, S D=1.45$ games per subject). In $\mathrm{D} 2(N=54)$, subjects played $\sim 1400$ trials ( $M=46.43, S D=5.41$ games; subjects stopped playing after exactly 40 min .; all games 30 trials).

## Results

Parameter Identifiability. A challenge in fitting RL parameters to individual subject behavior is that parameters
can be coupled and thus not fully identifiable. In the $f R L+$ decay model, equations $1-4$ show that the role of each parameter depends on the settings of the other parameters. Specifically, the values of the stimuli in equation 4 (in which choice is governed by $\beta$ ) depend-via equation 1 -on the weights of the chosen and non-chosen stimuli. Those weights are in turn respectively governed by the learning rate ( $\eta$, equation 2 ) and decay rate ( $d$, equation 3 ).

Coupling of the parameters modulating value estimation and choice is characteristic of many RL algorithms (Daw, 2011; Gershman, 2016). Coupling comes in two flavors: severe and moderate (Daw, 2011). Under severe coupling, parameters can trade off; for example, increases in one parameter can be perfectly compensated by decreases in another. As a result, parameter values may not-even in principle-be uniquely identifiable. Severe coupling can be tested for by repeatedly run an off-the-shelf optimizer from different initial parameter settings and checking whether optimization converges on the same estimates every time. If parameters are structurally coupled (i.e., there is no unique set of estimates), the optimizer will find different estimates on different runs, provided initializations allow the optimizer to cover sufficient territory in likelihood space. In D1 and D2, an optimizer repeatedly converged on the same parameter estimates, suggesting identifiability issues are not too severe to prevent finding a unique optimum.

However, there may still be more moderate identifiability issues. Intuitively, this is because maximum likelihood/maximum a posteriori (ML/MAP) estimates are tantamount to finding the highest point on the "hill" that defines the parameter surface in likelihood/posterior probability space. Yet they do not reveal the shape of the hill below: specifically, the shape of equal-likelihood ridges in the 3D likelihood space. If these ridges are diagonally shaped, they indicate covariance between the parameters. Intuitively, if changing a parameter in one direction (e.g., $\eta$ from 0.08 to 0.1 ) can be compensated for by changing another (e.g., $\beta$ from 6.2 to 5.1 ), with only miniscule changes in the likelihood, then one cannot safely draw conclusions from the point estimate of either parameter.
Identifiability and Computational Psychiatry. Identifiability poses a special challenge in the computational psychiatry domain, wherein the aim is often to derive parameters that can be used as predictors or outcome measures (Huys et al., 2016). Derived parameters whose point estimates have much uncertainty about them due to identifiability issues are unlikely to be useful for precision applications, such as prediction or diagnostic subtyping.
Probing Identifiability. A first helpful step for probing identifiability is to examine and visualize the Pearson correlations between pairs of estimates. Figure 3 plots point estimates for pairs of parameters in D1 and D2, with regression lines drawn to aid visualization.

Sets of parameters can fall along an elliptical contour in the likelihood space if there are identifiability issues, in which case the parameters will correlate. Thus, if parameter pairs closely correlate for most subjects in a dataset, this may


Figure 3. Parameter estimates in D1 (blue) and D2 (green).
indicate identifiability issues. However, correlations should only be a first step in checking for identifiability issues, for a couple reasons. First, to the extent that the parameters reflect meaningful psychological differences between individuals, we should expect they will correlate to some degree, because psychological variables often correlate within-subject (Lykken, 1968). Thus it can be difficult to determine whether correlations reflect modeling noise or true correlations between parameters. Second, correlations will not detect nonlinear relationships between parameters or other subtle identifiability issues (Gershman, 2016). Still, correlations are easily interpretable and a good place to start.
Figure 3 shows that, in both D1 and $\mathrm{D} 2,\{d$ and $\beta\}$ and $\{d$ and $\eta\}$ modestly correlate whereas $\{\eta$ and $\beta\}$ strongly correlate. In particular, in D2, $\{\eta-\beta\}$ estimates are nearly perfectly collinear for many subjects. Note also that, for all parameter pairs, the correlations are higher in D2, where there were more data, than in D1. In the test-retest reliability section below, we will present evidence suggesting that the parameter estimates may be more reliable in D2 than D1. Yet the higher correlations may also suggest more identifiability problems in D2. In fact, both possibilities-better parameter estimates and more identifiability problems - may be true. As noted, the equations in which the parameters are embedded dictate dependencies-and hence identifiability issuesbetween the parameters. If the true parameters are correlated, then when the model does a better job of recovering their values from noisy behavioral data, the observed data will also correlate more strongly. Thus, the increased correlations may actually be good news from a parameter recovery perspective. However, the $\{\eta-\beta\}$ collinearity does mean we should not treat these variables as independent.
Diagnosing Issues with Model Fit. Plots also allow visualization of outlying values, which may reflect model fit issues for specific subjects. For example, the arrows in the second row, second column plot in Figure 3 point to subjects
with outlying $\eta$ values. Outlying values might indicate $f R L+$ decay does not well describe specific subjects' choices in some or all of the task. However, the points could also reflect important individual differences, so additional checks are necessary to make a differential diagnosis.

We do not delve into model fit issues for individual subjects (as such specifics would not generalize beyond the samples/data in D1 and D2), but offer some general guidelines for probing these issues. First, another useful diagnostic is to plot likelihoods for each potentially problematic subject. For example, Daw (2011) provides an example of a 2 D heat map of likelihood values. A more quantitative assessment is the variance-covariance structure of parameter estimates; these structures can be examined by taking the inverse of the Hessian from optimization. On- and off- diagonal elements of $H^{-1}$ respectively give the variance and covariance of parameters. Large values indicate poor parameter estimates (Daw, 2011). Finally, problematic subjects' behavioral (and physiological, if available) data can be checked to see if these data are informative about the source of outlying parameter values (e.g., if reaction times were recorded, it can be useful to check if a subject responded atypically quickly or slowly during a subset or all of the task).
Ultimately, if outlying parameter values for a subject do not appear to be due to individual differences, but rather to issues with model-fit, the researcher may wish to treat these parameters as missing: Parameter estimates derived from a model that poorly describes a subject's behavior are meaningless. However, such decisions should be made-then adhered to-before inferential statistics, to avoid the "garden of forking paths" (Gelman \& Loken, 2013).
Subject-Specific Model Fit Issues and Computational Psychiatry. Our identification of apparent model fit issues among subjects illustrates the value of collecting data under different tasks specifications prior to attempting to develop a computational psychiatry assay. For instance, in the Dimensions Task, the presence of multiple individuals with apparently poor model fits suggests that some subjects in future clinical science designs will likely have missing data for model parameters (because, as noted, values from a model that poorly describes participant behavior should not be used). This is important information in the design phase of a clinical science study, as it may influence factors such as recruitment target, or collection of other data to aid estimation of anticipated missing values.
Test-retest reliability. As the cautionary tale of the dot probe task suggests, it is critical to establish the test-retest reliability of potential outcome measures. High test-retest reliability scores increase confidence that the measure is tapping a stable psychological construct (Hitchcock, Radulescu, Niv, \& Sims, 2017). Establishing stability of a measurement is a prerequisite for computational psychiatry designs that seek to use the measure to assess the effects of some experimental perturbation or group or individual differences. Nevertheless, the basic requirement of establishing test-retest reliability goes unmet with striking frequency in laboratory tasks (Lilienfeld, 2014).

| Dataset | $d$ | $\eta$ | $\beta$ |
| :--- | ---: | ---: | :---: |
| D1 | .17 | .36 | .71 |
| D2 | .68 | .79 | .69 |

Table 1. Intraclass correlation coefficients of parameters.
Table 1 presents test-retest reliability data for D1 and D2. These estimates were derived from splitting the data into approximately equal halves (specifically at the first game change after half of trials elapsed) and fitting the model to each (approximate) half. The test-retest reliabilities for $\{d$ and $\eta\}$ in D1 were quite low. This is likely because subjects only played 500 trials, and $\sim 250$-the approximate number of trials per half - may be too few trials to reliably estimate the parameters. In contrast, the D2 data suggest that $\sim 700$ trials allows for better parameter estimation, as reflected in the fact that test-retest scores for $\{d$ and $\eta\}$ are much higher.

Universal norms for intra-class correlation coefficients (ICCs) are arguably not justifiable (Weir, 2005) and at present there are no ICC benchmarks for RL tasks. But, in all domains, uncertainties around parameters increase as ICCs decrease (Weir, 2005). Thus, the above data are relevant to clinical science designs because they show how ICCs can increase with more data (see also Hitchcock et al., 2017). Gathering this information before designing a computational psychiatry assay is useful because computational psychiatry designs must often balance competing goals. On one hand, parameter estimates tend to improve with more trials. On the other, it may be infeasible to have subjects complete too long a task. For instance, individuals with certain disorders may fatigue easily. Experimental manipulations (e.g., rumination inductions) may also quickly dissipate. Test-retest reliability data can help negotiate the tradeoff between optimizing parameter estimates and keeping time on task feasible.

## Conclusions

Computational psychiatry promises to improve measurement and refine theory in clinical science (Hitchcock, 2017). Ultimately it may advance understanding of psychiatric disorders (Redish \& Gordon, 2016). Yet there are significant barriers to developing computational psychiatry assays. These barriers are diverse; hence this paper was part theoretical overview and part case study. The overview part of the paper first built motivation by discussing the dot probe paradigm, a case in which failure to attend to measurement issues in a laboratory task had disastrous results. Dozens of studies were conducted and vast resources were expended, over decades, before the poor properties of the task measures were realized. Next, we reviewed why laboratory tasks are so vulnerable to measurement issues: Task performance is often skewed by random state variation. That is, behavior collected only once or a few times from a single subject is often corrupted by situational factors. These review parts of the paper highlighted that minimizing noise in laboratory task measures is imperative. In the case study part of the paper, we overviewed modeling issues in RL tasks that can add
noise to parameter estimates, using two datasets for illustration. We concluded by presenting test-retest reliability data from the Dimensions Task, using this example to illustrate how time-on-task can improve reliability.

We should note that we have presented only some of the steps that should be taken when applying an RL task in clinical science. Other options include applying empirical priors (Gershman, 2016), using physiological data to aid parameter estimation (e.g., Leong, Radulescu, et al, 2017), and employing hierarchical modeling to weight parameter estimates by group statistics (Gelman \& Hill, 2006), which can reduce the variance of parameter estimates (Daw, 2011). As computational psychiatry develops, we predict that psychometric, study design, and parameter estimation issues will come increasingly to the fore.

## Acknowledgments

This work was supported by NSF research grant DRL1560829 (CRS) and ARO grant W911NF-14-1-0101 (YN).

## References

Bar-Haim, Y., Lamy, D., Pergamin, L., BakermansKranenburg, M. J., \& Van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. Psych. Bull., 133(1), 124.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models (pp. 3-38). Decision making, affect, and learning: Attention and performance XXIII, 23.
Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. Journal of Personality and Social psychology, 37(7), 1097-1126.
Gelman, A., \& Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press.
Gelman, A., \& Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.
Technical report, Department of Statistics, Columbia University, New York, NY.
Gershman, S. J. (2016). Empirical priors for reinforcement learning models. J. Mathematical Psychology, 71, 1-6.
Hallion, L. S., \& Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. Psychological Bulletin, 137(6), 940-958.
Heeren, A., Mogoașe, C., Philippot, P., \& McNally, R. J. (2015). Attention bias modification for social anxiety: a systematic review and meta-analysis. Clin. Psych. Rev., 40, 76-90.
Hitchcock, P.F. (2017). Computational Modeling and Reform in Clinical Science. [preprint; osf.io/mvxkf] OSF.
Hitchcock, P.F., Radulescu, A., Niv, Y., Sims, C.R. (2017). Assessing the Potential of Computational Modeling in Clinical Science. In The $3^{\text {rd }}$ Multidisciplinary Conference on Reinforcement Learning and Decision-Making.

Huys, Q. J., Maia, T. V., \& Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neuro., 19(3), 404-413.
Kenrick, D. T., \& Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. American Psychologist, 43(1), 23-34.
Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., \& Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron, 93(2), 451-463.
Lilienfeld, S. O. (2014). The Research Domain Criteria ( RDoC ): an analysis of methodological and conceptual challenges (pp. 13-14). Beh. Res. and Ther., 62, 129-139.
Lykken, D. T. (1968). Statistical significance in psychological research. Psych. Bull., 70(3), 151-159.
Maia, T. V., \& Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. Nature Neuroscience, 14(2), 154-162.
Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., \& Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. J. of Neurosci., 35(21), 8145-8157.
O'Doherty, J. P., Cockburn, J., \& Pauli, W. M. (2017). Learning, Reward, and Decision Making. Annual Review of Psychology, 68, 73-100.
Paulus, M. P., Huys, Q. J., \& Maia, T. V. (2016). A roadmap for the development of applied computational psychiatry. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1(5), 386-392.
Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., ... \& Amir, N. (2015). Empirical recommendations for improving the stability of the dotprobe task in clinical research. Psychological Assessment, 27(2), 365-376.
Radulescu, A., Daniel, R., \& Niv, Y. (2016). The effects of aging on the interaction between reinforcement learning and attention. Psychology and Aging, 31(7), 747-757.
Redish, A. D., \& Gordon, J. A. (2016). Computational Psychiatry: New Perspectives on Mental Illness. Cambridge, MA: MIT Press.
Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., \& Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. Journal of Abnormal Psychology, 125(6), 840-851.
Schmukle, S. C. (2005). Unreliability of the dot probe task. European Journal of Personality, 19(7), 595-605.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge: MIT press.
Van Bockstaele, B., Verschuere, B., Tibboel, H., De Houwer, J., Crombez, G., \& Koster, E. H. (2014). A review of current evidence for the causal impact of attentional bias on fear and anxiety. Psychological Bulletin, 140(3), 682-721.
Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. The J. of Str. \& Cond. Res., 19(1), 231-240.

# The Meanings of Morality: Investigating the psychometric properties of distributed representations of latent moral concepts 

Joe Hoover<br>University of Southern California, Los Angeles, California, United States

Katie Horton
University of Southern California

## Morteza Dehghani

University of Southern California


#### Abstract

People's beliefs about what is morally right and wrong vary widely between individuals, contexts, and cultures; however it is thought that they are governed by core latent constructs. While there is evidence that these constructs are reflected in natural language, this requires further testing. We demonstrate that the structure of moral values in natural discourse can be modeled by applying factor analyses to distributed representations of morally relevant terms learned by a neural network. We first demonstrate that robust latent constructs can be estimated from the covariance of distributed representations of construct exemplars. We then test whether the factor structure proposed by Moral Foundations Theory (MFT) is reflected in natural language. Finally, we conduct a bottom-up investigation of the structure of moral values in natural language using freeresponses reported by participants. Ultimately, we find evidence that the representation of moral values in natural language partially corresponds to MFT.


# Probability judgement from samples: accurate estimates and the conjunction fallacy 

Rita Howe (Rita.Howe@ucdconnect.ie)<br>School of Computer Science, Belfield, Dublin 4

Fintan J. Costello (Fintan.Costello@ucd.ie)
School of Computer Science, Belfield, Dublin 4


#### Abstract

This paper investigates a fundamental conflict in the literature on people's probability estimation. Research on 'perception' of probability shows that people are accurate in their estimates of probability of various simple events from samples. Equally, however, a large body of research shows that people's probability estimates are fundamentally biased, and subject to reliable and striking fallacies in reasoning. We investigate this conflict in an experiment that examines the occurrence of the conjunction fallacy in a probability perception task where people are asked to estimate the probability of simple and conjunctive events in a presented set of items. We find that people's probability estimates are accurate, especially for simple events, just as seen in previous studies. People's estimates also show high rates of occurrence of the conjunction fallacy. We show how this apparently contradictory result is consistent with a recent model of probability estimation, the probability theory plus noise' model.


Keywords: Conjunction fallacy; Disjunction fallacy; Perceptual probabilities; Probability estimation

## Introduction

The ability to reason under uncertainty (to estimate probabilities) is fundamental to human cognition. Humans exist in a world of stochastic processes, both stationary and nonstationary. They are regularly required to produce estimates for discrete events often with their own hidden parameters. It shouldn't be surprising then that humans are often very accurate in the probability judgements that they provide. This paper investigates a fundamental contradiction in the literature on people's probability estimation. Research on people's perception of probability shows that people are quite accurate when required to give estimates of the probability of simple events. Equally, however, a large body of research shows that people's probability estimates are fundamentally biased, and subject to reliable and striking fallacies in reasoning (such as the conjunction fallacy). To investigate this contradiction, we present an experiment that examines conjunction and disjunction fallacy rates and accuracy of probability estimates simultaneously. We find that while probability estimates are accurate, they are biased in specific ways. We also find that high conjunction and disjunction fallacy rates can co-exist with accurate probability estimation.

## Probability perception

Early research on probabilistic reasoning involved presenting participants with sequences or sets of simple events that varied on one particular dimension (sets of different shapes,
for example), and asked participants to estimate the probability of one particular event or outcome in that set (the probability of seeing a triangle in that set, for example). Results from these studies of 'probability perception' showed that the relation between subjects' mean estimates of probabilities and the sample proportions are described well by the identity function: people's probability estimates agreed well with the true objective probabilities (Peterson \& Beach, 1967). Later work on perceptual probabilities has suggested that humans have computational mechanisms that provide them with reasonably accurate judgements of simple probabilities (Balci, Freestone \& Gallistel, 2009). Participants are both accurate in their probability judgements and quick to detect large step changes in probabilities when required to give repeated estimates for non-stationary Bernoulli processes in real time (Gallistel, Krishan, Liu \& Miller, 2014). Similarly, Zhao, Shah, and Osherson (2009) used this 'probability perception' paradigm to examine people's judgements of conditional probability. Their participants observed shapes of different colours on screen for 4 seconds. These were static but appeared at new coordinates after a second had elapsed. Relatively small discrepancies between objective probabilities and conditional probability estimates were observed in this task.

## Fallacious reasoning

By contrast, research on errors in probabilistic reasoning (mainly in the 'heuristics and biases'framework) has uncovered many reliable and systematic errors or biases in people's judgement of probability. Over 50 such biases have been recognised, including the conjunction fallacy and disjunction fallacy (Baron, 2008). The conjunction fallacy, which arises when subjects judge some conjunction of events $A \wedge B$ to be more likely (more probable) than one of the constituent events of that conjunction, A, has gained the most attention since its discovery. Probability theory, which requires that $P(A \wedge B) \leq P(A)$ and $P(A \wedge B) \leq P(B)$ must always hold (simply because $A \wedge B$ cannot occur with A or B themselves occurring). The conjunction, $A \wedge B$, under the probabilistic laws, cannot be more likely than the single constituent A , thus when a participant chooses the conjunction $A \wedge B$ as more probable, they are committing a fundamental violation of rational probabilistic reasoning. The 'Linda problem'of Tversky and Kahneman (1983) is probably the best known example of this fallacy. The Linda problem is as follows:

Linda is 31 years old, single, outspoken, and very bright.

She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

## A. Linda is a bank teller <br> $A \wedge B$. Linda is a bank teller and is active in the feminist movement

Tversky and Kahneman found that when presented with simple conjunction problems over $80 \%$ of their participants made an erroneous judgement, judging $A \wedge B$ as more likely than $A$. A similarly reliable disjunction fallacy occurs when participants judge the constituents $\mathrm{A}, \mathrm{B}$ as more likely than the disjunction, $A \vee B$ (Bar-Hillel \& Neter, 1993). These widely replicated fallacy results were taken as an indication that humans do not reason in a normative fashion that is, they dont apply probabilistic rules to real-life contexts. Instead, it was suggested that people employ heuristics mental short cuts - to solve these problems. The conjunction fallacy, for instance, was suggested to occur because people employed a "representativeness heuristic" when reasoning about conjunctive problems (Tversky et al., 1983). Under this theory, the fallacy occurs as the person described in the conjunction is more representative of the information presented in the character sketch. However, research has called the validity of the representativeness account into question (Bonini, Tentori \& Osherson, 2004; Sides et al, 2002). Experiments that manipulated class inclusion, for instance, demonstrated that the fallacy occurs regardless of whether the conjunction is representative or not (Gavanski \& Roskos-Ewoldsen, 1991). While fallacy rates are generally quite high, a frequent observation among this research is that a small number of participants do not seem overly susceptible to the fallacy. In addition, over a number of conjunction problems, participants rarely have $100 \%$ error rates. Stanovich and West (1998) recognized that individuals can differ greatly on their performances on cognitive bias eliciting tasks. They found that subjects with higher cognitive ability were disproportionally likely to avoid committing a number of cognitive biases including the conjunction fallacy.

Previously, weighted models based on component probabilities such as the Signed Summation and Low-component models were popular as a means to explain the range of results that were consistently observed in fallacy research (Thüring \& Jungermann, 1990; Yates \& Carlson, 1986). However, these were limited in the scope of results that they could predict. A more successful iteration of these weighted models is the Configural Weighted Average (CWA) model (Nilsson, Winman, Juslin \& Hansson, 2009). This sophisticated weighted model includes a "noise component" that randomly disturbs probability judgements. More recently formal probabilistic models have sought to show that a range of biases can be explained as a function of quasi-rational
probabilistic reasoning instead of a heuristic process. Hilbert (2012) proposed a theoretical framework based on noisy information processing. Under this framework, memory based processes convert observations stored in memory into decisions. By assuming that these processes are subject to noisy deviations and that the noisy deviations are a generative mechanism for fallacious decision-making it provides an explanation for a number of cognitive biases . Costello and Watts (2016) proposed the Probability Theory plus Noise (PTN) model which can account for this variability in occurrence of the conjunction fallacy. In this model, people do reason in a normative fashion according to probability theory but are subject to random error in the reasoning process. The reasoner's decision-making processes, which are memory based, reliably apply the conjunction rule during the probability estimation process, but random noise causes fluctuations in judgement that sometimes lead to the subjective probability of a conjunction exceeding the subjective probability of the constituent. Costello and Watts showed that a simulation implementing this model produced a wide range of fallacy rates (from less than $10 \%$ to close to $70 \%$ ) and produced conjunction fallacy rates for a given set of materials that closely matched those seen in experimental studies for the same materials.

## Aims of this paper

These two lines of research use different paradigms (direct perception of probability in controlled sets of events versus questions about the probability of events given descriptions) and lead to two contradictory conclusions (people's probability estimates are fundamentally accurate versus people's estimates are fundamentally flawed). In this paper, we describe an experiment that aims to reconcile these two strands of research by using using a perceptual probability task to examine conjunctive and disjunctive probability judgements and the occurrence of the conjunction and disjunction fallacy. We ask whether these fallacies will occur reliably in direct probability perception even though people's estimates in probability perception tasks are typically accurate. We also investigate the predictions of the (PTN) model that attempts to simultaneously explain both relatively accurate estimation and reliable fallacy occurrence (Costello et al, 2016).

This model assumes that people estimate probabilities using a fundamentally rational process which is, however, subject to the systematic biasing effects of random noise in the reasoning process. Importantly, this model proposes that the rate of random noise is greater for more complex conjunctive and disjunctive events than for simple events (as a consequence of simple propagation of error: because conjunctions and disjunctions are more complex, they have more points of 'failure' at which random noise can have an effect). This model thus predicts relatively accurate probability estimates, especially for simple events (as seen in the 'probability perception' literature), but stronger systematic bias due to noise for conjunctive and disjunctive events (producing the conjunction and disjunction fallacy).

## Experiment

This experiment involves repeatedly presenting participants with images where each image contains a relatively large number of shapes differing in colour (red, white or green) and configuration (solid or hollow). For each image participants are asked to estimate the probability of some event (e.g., a randomly selected shape being red, for example). The true probability of events in these images were held constant across multiple presentations (but with the images themselves varying as to the position of the different shapes on the screen each time), as described below. Each participant saw multiple presentations of the same probability question (multiple questions for which the objectively correct probability was the same), allowing us to estimate the degree of random variation in participants estimates. Some questions asked about simple events (a shape being red, being hollow, etc.) while other questions asked about conjunctive and disjunctive events (a shape being red and solid, a shape being white or hollow, etc.) Two distinct sets of images were used, with objective probabilities of single and conjunctive probabilities held constant in each set (see below). The images from these two sets were interspersed with each other. Participants answered questions about 460 images in total. Images were only on screen for a short time ( 2 seconds), and so participants did not have time to count the occurrence of shapes of different types. Images were presented in randomised order.

## Materials

The images consisted of shapes of three colours - colours A, $B$, and $C$ respectively - and 2 shape configurations - $X$ and Y - with fixed probabilities. The actual colour varied from image to image, so sometimes colour A was white, sometimes colour A was red and sometimes colour A was green. The colours varied in the same way for colour B and colour C. The actual configuration of the shapes also varied from image to image so sometimes configuration X was solid shapes and sometimes configuration X was hollow shapes. Conjunction and disjunctions were created for a number of combinations of colour and configuration such as $\mathrm{P}(\mathrm{A} \wedge \mathrm{X}), \mathrm{P}(\mathrm{A} \wedge \mathrm{Y})$ and $\mathrm{P}(\mathrm{B} \vee \mathrm{X})$.

For each type $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{X}, \mathrm{Y}, \mathrm{A} \wedge \mathrm{X}, \mathrm{A} \wedge \mathrm{Y}$ etc, there were 20 images asking people to estimate the probability of that type. In practise, this meant that the participants saw 20 im ages asking them to estimate the probability of colour A, 20 images asking them to estimate the probability of colour B , 20 images asking them to estimate the probability of configuration X and so on.

## Set 1

In set 1 , colour A had a fixed probability of 0.7 , colour B had a fixed probability of 0.2 and colour C had a fixed probability of 0.1 . Configuration X had a fixed probability of 0.9 and configuration $Y$ had a fixed probability of 0.1 . The conjunctions for set 1 were created using the following colour and configuration combinations: $\mathrm{P}(\mathrm{A} \wedge \mathrm{X}), \mathrm{P}(\mathrm{B} \wedge \mathrm{X})$, and $\mathrm{P}(\mathrm{B} \wedge \mathrm{Y})$. These corresponded to the objective probability values of $0.63,0.18$
and 0.02 . The disjunctions for set 1 were created using the following colour and configuration combinations: $\mathrm{P}(\mathrm{A} \vee \mathrm{X})$, $\mathrm{P}(\mathrm{B} \vee \mathrm{X}), \mathrm{P}(\mathrm{A} \vee \mathrm{Y})$, and $\mathrm{P}(\mathrm{B} \vee \mathrm{Y})$. These disjunction combinations corresponded to the the objective probability values of $0.97,0.92,0.73$, and 0.12 . Participants viewed 220 images of 20 geometric shapes on a computer screen.

## Set 2

For set 2, colour A had the fixed probability of 0.333 , colour $B$ had the fixed probability of 0.333 and colour $C$ had the fixed probability of 0.333 . Configuration X had the fixed probability value of 0.5 and configuration Y had the fixed probability value of 0.5 . The conjunction for set 2 had the value of 0.17 . Any combination of $A, B, C$ and $X, Y$ would give this value. The disjunction had the objective probability value of 0.67 . Again, any combination of $A, B, C$ or $X, Y$ would give this value.

Participants viewed 240 images of geometric shapes in a computer screen. Each image consisted of 12,24 , or 36 shapes. Each objective probability values of $0.333,0.5,0.17$, and 0.67 were presented 20 times for each of the 12,24 and 36 shape images.

Each image presentation included a question to elicit a probability judgement. For the colour probability questions, participants were presented with questions in the form "What is the probability of picking a shape that is [colour A]?" or "What is the probability of picking a shape that is [colour B]?" Colour C was excluded from the probability questions. For the configuration questions, participants were presented with questions in the form: "What is the probability of picking a shape that is [configuration X]? or "What is the probability of picking a shape that is [configuration Y]?" The conjunction and disjunction questions took the same form. For instance, the question to elicit a probability judgement for the objective probability of 0.63 would be: "What is the probability of picking a shape that is [colour A AND configuration X]?".

## Procedure

Participants were seated at a screen. Each participant began with a training trial of sample stimuli to familiarize themselves with the task (see figure 1). Once the participants were comfortable with the task, they moved onto the experimental trials. The static image and the probability question appeared on screen simultaneously. The image was replaced with a blank screen once 2 seconds had elapsed to prevent the participants from counting the shapes. The associated question remained on-screen until the participants had made their guess. The participants indicated their estimate by moving dial on a slider using their mouse or arrow keys. This slider had a minimum value of 0 and a maximum value of 1 . A box in the corner indicated the exact value of the participants' estimate and dynamically updated as they moved the slider. When the participant was satisfied with their answer, they submitted it by clicking on a Next button. This also triggered the succeeding image and probability question.


Figure 1: The figure above displays example stimuli image from set 1 in grey scale. While the shape types changed between sets, the underlying proportions remained constant. The image above has a shape configuration of 0.9 for solid shapes and 0.1 for hollow shapes. The colours have fixed probabilities of 0.7, 0.2 and 0.1.

## Results

A total of 9 participants made 460 probability judgements each. Their responses and response time was recorded for each judgement. Two of the participants were excluded from the final analysis for failing to answer $20 \%$ of the questions. The number of participants is in line with other studies of probability perception (e.g. Gallistel et al, 2014).

Error and variance The initial analysis determined whether an estimate was an actual estimate or whether an error had occurred in the response (such as the participant mistakenly submitting an estimate). To do this, the standard deviation for each participants' estimates were calculated. Any estimate that fell $\pm 3$ standard deviations from the mean estimate of a probability was excluded. In total, this comprised of $1 \%$ of responses.

## Estimated probability vs true probability

For each of the 11 probability values in set 1 , each participant gave 20 estimates for its value. In set 2 , the 4 probability estimates were questioned at 3 different levels; 20 estimates were given for each probability value at each level. The relationship between mean probability estimates and objective probability are displayed in figure 2. For each probability value, the participants' average estimate and standard deviation was calculated. The average estimate and standard deviation were also calculated for the sample. The average deviation from the true probability was calculated in terms of percentage points. Some noticeable trends were observed, participants tended to overestimate the low probabilities and underestimate the higher probabilities. The degree of overestimation for the low constituents was much less than for the low complex statements. For instance, the constituent with a true probability of 0.1 have an average estimate of 0.13 , while


Figure 2: The above graph displays the average probability estimate vs the objective probability value by type. Any value falling above the line represents an overestimation of the probability value, while the values falling below the line represent underestimations of the true value. Largely, conjunctions were overestimated and disjunctions were underestimated. Constituents tended to have accurate estimates.
the conjunction with a probability of 0.02 had an average estimate of 0.14 and the disjunction of 0.12 had an average estimate of 0.26 . Overall, conjunctions were overestimated and disjunctions were underestimated. The conjunctions and disjunctions averaged 10 percentage points away from their true probability while the constituent average was 4 percentage points. Figure 2 shows the average estimate for each type.

For set 2, the conjunctions were overestimated on all occasions, with the average estimate increasing as the stimulus set became more complex. The disjunctions were consistently underestimated. Participants were more accurate in their estimates for the constituents. The 12 -shape combinations had the lowest average estimates, the 24 -shape estimates were higher than the 12 -shape and lower than the 36 -shape estimates. The 36 -shape images had the highest mean estimates.

## Fallacy rates

Each conjunction and constituent was presented 20 times to each participant. To evaluate the rate at which the participant had committed the conjunction fallacy, each conjunction judgment $1 \ldots 20$ was matched in order with its corresponding constituent judgments ( $1 \ldots 20$ ), so the first conjunction judgement was matched with the first constituent judgements, and so on. If a particular conjunction judgement exceeded the estimate of either of the corresponding constituent values, an instance of the conjunction fallacy was recorded. For each participant, there were six conjunction questions where the fallacy could be committed, three from set 1 and three from set 2 . The average conjunction fallacy rate was $30 \%$. Fallacy rates ranged from $13 \%$ to $69 \%$ : a range that is in-line with those seen in description based studies (e.g. StolarzFantino, Fantino, Zizzo, \& Wen, 2003). The set-up of this experiment allows us to categorise conjunctions based on their

Table 1: Fallacy rates occurrences by objective probability for the conjunction and disjunction problems.

| Set 1 |  |  |  |  | Set 2 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Conjunction fallacy |  |  |  |  |  |  |  |
| Conjunction probability | 0.02 | 0.18 | 0.63 | - | $\begin{gathered} 0.17 \\ \text { (12-shapes) } \end{gathered}$ | $\begin{gathered} 0.17 \\ \text { (24-shapes) } \end{gathered}$ | $\begin{gathered} 0.17 \\ \text { (36-shapes) } \end{gathered}$ |
| Fallacy rate | 43\% | 21\% | 68\% | - | 19\% | 13\% | 14\% |
| Disjunction fallacy |  |  |  |  |  |  |  |
| Disjunction probability | 0.12 | 0.73 | 0.92 | 0.97 | $\begin{gathered} 0.67 \\ \text { (12-shapes) } \end{gathered}$ | $\begin{gathered} 0.67 \\ \text { (24-shapes) } \end{gathered}$ | $\begin{gathered} 0.67 \\ \text { (36-shapes) } \end{gathered}$ |
| Fallacy rate | 34\% | 28\% | 49\% | 71\% | 40\% | 25\% | 24\% |

Fallacy rates varied quite significantly for the conjunctive and disjunctive judgements. The highest fallacy rates were observed for the highest objective probabilities. However, the objective probability value was necessarily an accurate predictor of fallacy occurrence as high fallacy rate were also observed for low objective probabilities.
actual probabilities and their underlying constituent probabilities. The participants showed marked differences in performances for each of the six conjunctions they were presented with. Table 1 displays the fallacy rate breakdown by conjunction type. The highest fallacy rates are seen for the conjunction with the highest probability value. High fallacy rates were also observed for the conjunction with the lowest probability value. The other conjunctions had low fallacy rates.

As for the conjunction fallacy, each disjunction judgement was matched with the constituent judgements in sequence, so the first disjunction judgement was matched with the first instances of the relevant constituent judgements. If a disjunctive estimate was less than either of its constituent estimates then it was counted as a disjunction fallacy. The average disjunction fallacy rate was $39 \%$. The fallacy rate ranged from $25 \%$ to $71 \%$, which is consistent with the results from description based research. The average fallacy rate for the each of the 7 possible disjunctions is displayed table 1 . As for the conjunctions, the objective probability value of the disjunction was not an indicator of fallacy rate occurrence.

Conjunction and disjunction fallacy occurrence varied over the course of presentation, however, there was no obvious trend of improvement or deterioration in the participants ability to avoid committing the fallacies (fallacy rates did not decline with familiarity).

## Variance

Since each conjunction, disjunction and constituent was presented 20 times to each participant, we can estimate the degree of variance (standard deviation) in estimates for type. Recall that the PTN model predicts greater variance would exist for the complex combinations than the constituents. Results showed that the complex combinations were more variable than their constituent counterparts for $68 \%$ of the comparisons. The conjunctions were noisier than their con-

Table 2: Average standard deviation for constituents, conjunctions and disjunctions

| A | B | $\mathrm{A}(\mathrm{SD})$ | $\mathrm{B}(\mathrm{SD})$ | $\mathrm{A} \wedge \mathrm{B}(\mathrm{SD})$ | $\mathrm{A} \vee \mathrm{B}(\mathrm{SD})$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |
| 0.1 | 0.2 | 0.050 | 0.072 | 0.070 | 0.102 |
| 0.1 | 0.7 | 0.050 | 0.083 | - | 0.091 |
| 0.9 | 0.2 | 0.073 | 0.072 | 0.071 | 0.120 |
| 0.9 | 0.7 | 0.073 | 0.083 | 0.085 | 0.095 |
| 0.5 | $0.33(12)$ | 0.088 | 0.090 | 0.086 | 0.145 |
| 0.5 | $0.33(24)$ | 0.101 | 0.080 | 0.075 | 0.126 |
| 0.5 | $0.33(36)$ | 0.060 | 0.098 | 0.096 | 0.111 |

The table above displays the average variability scores for the single constituents, conjunctions and disjunctions. In nearly all the cases, the complex judgement (conjunction or disjunction) was more variable than one if not both of its constituents.
stituents counterparts for $50 \%$ of the comparisons. Disjunctions were more variable than their constituent counterparts for $83 \%$ of the comparisons. The average variance for constituents was 0.77 , conjunctions had an average variance of 0.81 and disjunctions had an average variance of 0.11 .

Variance and fallacy rate Participants variability was positively correlated with their fallacy rates, small positive correlations were observed for the conjunction fallacy rates, $r=$ 0.25 and the disjunction fallacy rates, $r=0.36$ for set 1 . For set 2, a strong positive correlation was observed for the conjunction fallacy rate and variability, $r=0.89$ and a mild positive correlation for the disjunctions, $r=0.32$. This supports the PTN model assumption that conjunction and disjunction fallacies arise due to variability in conjunction and disjunction estimates. Table 2 displays the average standard deviation values for each constituent, conjunction and disjunction. Overall, the complex combinations had higher average standard deviations than the constituents.

## Timings

Participants had slower response times for their initial estimates but these decreased and plateaued rapidly. To investigate whether there was a difference in response times for type - constituent, conjunction, and disjunction - the average time for each type was calculated. Then a repeated measures ANOVA was performed to examine whether a difference existed for the average reaction times for judgement type. The ANOVA found a significant difference in reaction speed for judgement type, $\mathrm{F}(2,278)=8.478, p<0.05$. Pairwise comparisons showed that constituents judgements were significantly faster than conjunctions judgements. Constituent judgements were also significantly faster than disjunctions judgements. No significant different was observed between the judgements speeds for the conjunctions and disjunctions.

## Discussion

This paper investigated an apparent conflict in the literature on probability estimation, which shows accurate estimates when people estimate probabilities from directly presented samples, but systematic occurrence of the conjunction and disjunction fallacy when people estimate the probability of described events. Our experimental results show accurate estimates, and frequent fallacies in judgement for conjunctions and disjunctions, occurring simultaneously when probabilities are estimated from samples. This pattern of results is predicted by the PTN model (Costello \& Watts, 2016), in which people estimate probabilities by following standard frequentist probability theory, but with random noise in judgement. Fallacies in judgement such as the conjunction and disjunction fallacy are caused by increased rates of random error for conjunctions and disjunctions, while accurate estimates for constituents are produced because people follow standard probability theory in making estimates, and because constituent estimates are subject to lower rates of random error.

The results here demonstrate that producing accurate probability estimates (especially for constituents) and producing conjunction and disjunction fallacy responses are not mutually exclusive states: both patterns of results can occur simultaneously, and are naturally explained in a model where people reason according to standard probability theory but are subject to random noise in the reasoning process

While the wide range of fallacy results observed in the literature (approximately $10 \%$ to $80 \%$ depending on task type) is challenging, this experiment demonstrates that a range of fallacy rates from low to high can occur for the same task and are a consequence of the objective values of the constituents, the conjunctions or disjunctions and the rate of variability in estimates. We observed complex combination estimates that were consistently more variable than the constituent estimates. While both the CWA and PTN models predict that fallacy rates will be effected by noise, they make divergent predictions. The CWA model predicts a negative correlation with variability and fallacy rates, that is, the rate of fallacy errors should increase as the rate of noise decreases. Here, both conjunction and disjunction fallacy rates correlated positively with variability. These results are consistent with the PTN model which predicts that we should observe a positive relationship between fallacy rates and variability.

The results of this experiment seem to resolve the conflict between studies of probability perception and the studies of fallacies in reasoning. Participants were accurate probabilistic reasoners - their judgements of probabilities were good especially for the constituents. However, these judgements are still systematically biased and depending on the problem type, noisy. People's estimates are more variable for conjunctions and disjunctions than the constituents and this variance causes the occurrences of the conjunction and disjunction fallacy. This account of accurate estimates with consistent fallacy occurrences is consistent with the PTN model.

## References

Balci, F., Freestone, D., \& Gallistel, C. R. (2009). Risk assessment in man and mouse. Proceedings of the
National Academy of Science, USA, 106, 2459-2463.
Bar-Hillel, M., \& Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. Journal of Personality and Social Psychology, 65(6), 1119-1131.
Baron, J. (2007). Thinking and deciding (4th ed.). New York: Cambridge University Press.
Bonini, N., Tentori, K., \& Osherson, D. (2004). A different conjunction fallacy. Mind and Language, 19(2), 199-210.
Costello, F., \& Watts, P. (2016). Explaining High Conjunction Fallacy Rates: The Probability Theory Plus Noise Account. Journal of Behavioral Decision Making.
Gallistel, R., Krishan, M., Liu, Y., Miller, R., \& Latham, P. E. (2014). The perception of probability. Psychological Review, 121(1), 96-123.
Gavanski, I., \& Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. Journal of Personality and Social Psychology, 61(2), 181.
Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. Psychological Bulletin, 138(2).
Nilsson, Winman, A., Juslin, P., \& Hansson, G. (2009). Linda Is Not a Bearded Lady: Configural Weighting and Adding as the Cause of Extension Errors. Journal of Experimental Psychology: General, 138(4), 517-534.
Peterson, R., \& Beach, R. (1967). Man as an intuitive statistician. Psychological Bulletin, 68(1), 29-46.
Sides, A., Osherson, D., Bonini, N., \& Viale, R. (2002). On the reality of the conjunction fallacy. Memory \& Cognition, 30(2), 191-198.
Stanovich, K. E., \& West, R. F. (1998). Individual Differences in Framing and Conjunction Effects. Thinking \& Reasoning, 4(4), 289-317.
Stolarz-Fantino, S., Fantino, E., Zizzo, D. J., \& Wen, J. (2003). The conjunction effect: New evidence for robustness. American Journal of Psychology 116(1)
Thüring, M., \& Jungermann, H. (1990). The conjunction fallacy: Causality vs. event probability. Journal of Behavioral Decision Making, 3(1), 61-74.
Tversky, A., \& Kahneman, D. (1983). Extensional Versus Intuitive Reasoning - the Conjunction Fallacy in Probability Judgment. Psychological Review, 90(4), 293-315.
Yates, J. F., \& Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including ‘signed summation'. Organizational Behavior and Human Decision Processes, 37(2), 230-253.
Zhao, J., Shah, A., \& Osherson, D. (2009). On the provenance of judgments of conditional probability. Cognition, 113(1), 26-36.

# Disfluencies in dialogues with patients with schizophrenia 

Christine Howes (christine.howes@gu.se)<br>Philosophy, Linguistics and Theory of Science University of Gothenburg, Sweden

Mary Lavelle<br>Florence Nightingale School of Nursing and Midwifery King's College London, UK

Patrick G.T. Healey<br>Queen Mary University of London

Julian Hough<br>Bielefeld University

Rose McCabe<br>Exeter University


#### Abstract

Disfluencies such as self-repairs, filled pauses such as 'um' and silent pauses are pervasive in dialogue, but there is no consensus in the literature as to whether they reflect internal production pressures, or interactive issues - or how their effects are manifest in dialogue. It is well-known that patients with schizophrenia have problems with language and social cognitive skills, yet little research has investigated how these impact interaction. We report a study on the disfluency behaviours of patients with schizophrenia and their interlocutors who were unaware of the patient's diagnosis, compared to healthy control groups. Results show that patients use fewer self-repairs than either their partners or controls and fewer filled pauses ('er', 'um') than controls. Furthermore, the presence of the patient also affects patients' partners, who use fewer filled pauses than controls and more unfilled pauses than both patients and controls. This suggests that smooth coordination of turns is problematic in patients' dialogues.


Keywords: Disfluency; Dialogue; Schizophrenia

## Introduction

Disfluencies, such as self-repair, pauses and filler non-words such as er and um (filled pauses) are pervasive in dialogue (Schegloff, Jefferson, \& Sacks, 1977). Such disfluencies are conventionally regarded as symptomatic of problems with communication, caused by self-monitoring or production issues (Levelt, 1983). However, disfluencies also highlight the interactive nature of dialogue - many disfluencies occur as we tailor our talk for specific addressees, or as a direct result of feedback from our interlocutors (Goodwin, 1979).

Furthermore, different types of disfluencies have been hypothesised to contribute differently to the individual and shared actions that must be coordinated in successful dialogue. For example, in route following experiments, different distributions of filled pauses and self-repairs suggest that self-repairs occur because of production difficulties, but filled pauses fulfil an interpersonal function (Bortfeld, Leon, Bloom, Schober, \& Brennan, 2001; Nicholson, Eberhard, \& Scheutz, 2010; Brennan \& Schober, 2001).

Other research suggests that disfluencies should be categorised according to whether they make changes to the meaning of an utterance or not, e.g. reformulations and false starts are backwards-looking disfluencies, whilst word repetitions and filled pauses are forwards-looking (Ginzburg, Fernández, \& Schlangen, 2014; Allwood, Nivre, \& Ahlsén, 1990).

In addition to signalling difficulties that the speaker may be experiencing, different types of disfluency have been shown to have conventionalised meanings with respect to turn-taking. Filled pauses may indicate a break in the information (for example while the speaker searches for a word
or phrase) but an intention to retain the floor, whilst unfilled pauses may signal that the speaker does not intend to continue speaking (Clark \& Fox Tree, 2002; Allwood et al., 1990).

Recent work suggests that disfluencies have measurable effects on the dialogue - even if this is not necessarily the intention of the speaker (Finlayson \& Corley, 2012; Ginzburg et al., 2014). For example, in contexts where informational exchange is key, such as the Map Task (Anderson et al., 1991), more self-repair may be indexing one's own production difficulties, and how hard people are working to be understood by (and for) their interlocutors (Colman \& Healey, 2011).

Further evidence that disfluencies do not just function as markers of miscommunication but contribute to improving the effectiveness of interaction comes from psycholinguistic studies. For example, referential success and ambiguity resolution are aided by the presence of disfluencies (Brennan \& Schober, 2001; Bailey \& Ferreira, 2007). Communications training interventions also indicate that talk between psychiatrists and patients with schizophrenia is improved when the psychiatrist uses more self-repair (McCabe et al., 2016).

As can be seen from the above discussion, there is no general consensus in the literature about either which disfluencies should be focused on, how different types of disfluency should be categorised, or the effects they have in dialogue.

It is well documented that people with a diagnosis of schizophrenia have problems with language and social cognitive skills, including with self-monitoring (Johns et al., 2001) and turn-taking (using role-play; Mueser, Bellack, Douglas, and Morrison, 1991), yet little research has investigated how these impact interaction. The few studies that do find that less patient self-repair is associated with verbal hallucinations (Leudar, Thomas, \& Johnston, 1992), more patient otherinitiated repair (clarification of the doctor's talk) is associated with better adherence to treatment (McCabe et al., 2013), and clinicians' use of self-repair has positive clinical consequences (McCabe et al., 2016). Research into disfluencies therefore has the potential to be used in diagnostic tools, and feed into training for psychiatrists to detect when a patient is in difficulty or shape their own talk more effectively.

However, there is conflicting evidence regarding disfluencies in patients with schizophrenia and whether this differs from non-clinical populations. In consultations, patients use more self-repair than psychiatrists (McCabe et al., 2013) and this is higher than found in general dialogue in the demographic portion of the British National Corpus, or Map Task dialogues (Howes, Purver, McCabe, Healey, \& Lavelle, 2012;

Colman \& Healey, 2011). This might be expected in the clinical domain where patients are explaining and providing updates on their health and treatment. In contrast, in a controlled study where subjects described the experimenters actions, frequency of repair did not differ between patients and matched controls (Leudar et al., 1992).

These differences may reflect interactional factors, such as domain, or role, and not differences between schizophrenia patients and non-clinical populations per se. Additionally, in these studies, patients' interlocutors were aware of their diagnosis, which could have influenced the way they interacted with the patient (Doyen, Klein, Pichon, \& Cleeremans, 2012).

We report an analysis of a unique corpus of 40 triadic dialogues (Lavelle, Healey, \& McCabe, 2013) that avoids these potential confounds (see Method, below). Given that the literature suggests patients produce more self-repairs than their psychiatrists in dialogues during clinical consultation, and more so than people in natural corpora from the general population, and with the assumption that turn-taking cues (including offering the floor to another participant, or retaining it for oneself) partially consist of filled and unfilled pauses (Maclay \& Osgood, 1959), we expect the following:

## Hypotheses

1. Compared to healthy control conversational groups and their healthy conversational partners, patients will produce more disfluent talk, with more self-repairs.
2. Compared to controls and their conversational partners, patients will produce fewer turn-taking cues (filled and unfilled pauses).
3. Compared to patients and controls, patients' partners will produce more turn-taking cues.

## Method

## Participants

The data consist of transcripts of twenty patient interactions, involving one patient conversing with two healthy controls who were unaware of the patient's diagnosis, and twenty control interactions (3 healthy participants). Twenty patients with a diagnosis of schizophrenia ( 6 male, 14 female) and one hundred non-psychiatric healthy participants, forty in the patient condition ( $21 \mathrm{~m}, 19 \mathrm{f}$ ) and sixty in the control condition ( 34 $\mathrm{m}, 26 \mathrm{f}$ ), participated. Participants within each triad were unfamiliar to each other. Due to technical issues one patient interaction and one control conversation could not be transcribed and are excluded from the analysis, resulting in data from 57 individuals in control groups ( 19 triads), and 19 patients and 38 healthy controls in patient interactions.

Non-psychiatric healthy participants were recruited through advertising on local community websites. Of those who responded to the advertisement, $40 \%$ participated. Participants with a diagnosis of psychosis or affective disorders in themselves, or any first-degree relatives, and those who were not fluent English speakers were excluded.

Patients were recruited at routine psychiatric outpatient clinics under supervision of their psychiatrist. $25 \%$ of all patients approached agreed to participate. Patients were taking anti-psychotic medication which fell within the low dose range (Chlorpromazine equivalents $50-200 \mathrm{mg} / \mathrm{day}$ ). Nonnative English speakers and patients presenting with motor side effects from antipsychotic medication were excluded based on a clinician's assessment.

The distribution of gender did not significantly differ between patient and control conditions ( $\mathrm{P}: n=60$ : female $=53.33 \%, \mathrm{C}: n=60:$ female $=43.33 \% ; \chi^{2}=1.20, p=$ 0.27 ). Patients were significantly older than controls ( P : $M=41 s . d .=8.6, \mathrm{C}: M=31$ s.d.$\left.=9.6 ; t_{119}=4.51, p<0.01\right)$

Symptoms were assessed using the Positive And Negative Symptom Scale for Schizophrenia (Kay, Flszbein, \& Opfer, 1987). Patients displayed relatively low PANSS scores for positive symptoms - additional features that occur with the disorder such as hallucinations or delusional beliefs ( $M=$ $15.8 ;$ s.d. $=6.76$ ), and negative symptoms, which represent a reduction in usual function such as social withdrawal, diminished affect, apathy and anhedonia $(M=9.95 ;$ s.d. $=3.36)$.

## Ethics

All procedures were approved by a UK NHS Research Ethics Committee (07/H0711/90). All participants gave written informed consent and were free to withdraw at any time.

## Procedure

Participants were brought into the laboratory in threes and seated in a triangular formation so that they all had good visual access to each other (see Figure 1). The researcher read aloud a fictional moral dilemma, the 'balloon task' (see Task section, below), which has been used for studying dialogue, and is known to stimulate discussion (Howes, Purver, Healey, Mills, \& Gregoromichelaki, 2011). The group was provided with an opportunity to ask questions before the researcher left the interaction space and the task began. Interactions ended when participants reached a joint decision. Groups that failed to reach agreement had their interaction terminated at approximately 450 seconds ( 7 minutes 30 seconds).


Figure 1: Participants engaged in triadic interaction

## Task

The balloon task is an ethical dilemma requiring agreement on which of four passengers should be thrown out of a hot air balloon, which is losing height and about to crash into some mountains killing all on board unless one of them jumps to their certain death in order to save the other three. The four passengers are described to the participants as follows:

Dr. Robert Lewis - a cancer research scientist, who believes he is on the brink of discovering a cure for most common types of cancer.

Mrs. Susanne Harris - who is not only widely tipped as the first female MP for her area, but is also over the moon because she is 7 months pregnant with her second child

Mr. William Harris - husband of Susanne, who he loves very much, is the pilot of the balloon, and the only one on board with balloon flying experience.

Miss Heather Sloan - a 9 year-old music prodigy, considered by many to be a "twenty first century Mozart".

Participants were instructed to debate the reasons for and against each person being saved, and reach mutual agreement about who should jump.

## Analysis

Participants' speech was transcribed in ELAN (Brugman \& Russel, 2004), allowing us to map the transcriptions to the video and precisely time pauses.

Self-repair Self-repairs were annotated automatically using STIR (STrongly Incremental Repair detection; Hough and Purver, 2014); ${ }^{1}$ which detects speech repairs on transcripts word-by-word incrementally. It uses a pipeline of classifiers to tag each word of the transcript as either fluent, or in an element of the three-part repair structure below, according to the manual by (Meteer, Taylor, MacIntyre, \& Iyer, 1995):

$$
\begin{equation*}
\text { John } \underbrace{[\text { likes }}_{\text {reparandum }}+\underbrace{\{\mathrm{er}\}}_{\text {interregnum }} \underbrace{\text { loves }]}_{\text {repair }} \text { Mary } \tag{1}
\end{equation*}
$$

STIR uses features from n-gram language models in a pipeline of classifiers which classify whether the current word constitutes a boundary of each part of the repair structure. STIR is trained on the Switchboard corpus (Godfrey, Holliman, \& McDaniel, 1992) and achieves an F-score accuracy for self-repair detection of 0.81 on conversational data (Howes, Hough, Purver, \& McCabe, 2014). It has previously been applied to therapeutic dialogue, with high rates of correlation to human coders in terms of self-repair rate (Howes et al., 2014), so is adequate for our annotation purposes.

Filled pauses In this data, filled pauses were found to be inconsistently spelt (aammm, er, eerrrrmm, uhmmm etc). A find-and-replace operation was applied to the corpus prior to analysis to give these a standardised spelling, i.e. 'er’ (Howes et al., 2014). For the analyses we used a count of the number of filled pauses used by each participant.

Unfilled pauses Following e.g. Zellner (1994), we defined unfilled pauses as speech-free spaces between segments of speech by the same speaker of greater than 200 milliseconds. Pause segments were automatically extracted from the ELAN transcripts. For the analyses we used a count of the number of unfilled pauses used by each participant.

As patients produce fewer turns than their interlocutors, per-turn rates for these measures were calculated for each individual participant as the total number of self-repairs, filled or unfilled pauses produced divided by the total number of turns. Patients' turns are also typically shorter - as longer turns are expected to have more repair (Bortfeld et al., 2001), we also calculated measures per 100 words (see Table 1).

Statistics Analyses were run in SPSS using Generalised Linear Mixed Models (GLMMs) to control for both fixed and random effects. In all reported models, condition was a fixed effect and participant ID was a random effect with individuals clustered by their conversation group. For each model we used random intercepts and the maximal random effect structure justified by the sample (Barr, Levy, Scheepers, \& Tily, 2013), using a gamma distribution with a log link function. We report exact p -values throughout, but take $p<0.05$ to be the criterion of significance.

## Results

Table 1 shows statistically significant main effects of condition. Pairwise comparisons are reported in the text below.

Number of turns Patients ( P ) produced significantly fewer turns than their partners (PP) ( $t_{1,102}=-3.247, p=0.001$ ). There was no significant difference between patients and controls (C) or patients' partners and controls ( $\mathrm{P} / \mathrm{C} t_{1,102}=$ $-1.574, p=0.118 ; \mathrm{PP} / \mathrm{C} t_{1,102}=-0.213, p=0.832$ ).
Number of words Patients produced fewer words in total and per turn than either their partners or controls (Total: $\mathrm{P} / \mathrm{PP} t_{1,102}=-3.914, p<0.001 ; \mathrm{P} / \mathrm{C} t_{1,102}=-2.481, p=$ $0.015 ; \mathrm{PP} / \mathrm{C} t_{1,102}=0.172, p=0.864 ;$ Per turn: P/PP $t_{1,102}=$ $-3.823, p<0.001 ; ~ \mathrm{P} / \mathrm{C} t_{1,102}=-2.183, p=0.031$; PP/C $\left.t_{1,102}=0.979, p=0.330\right)$.

Table 1: Overview.

| ¢ |  | Patien | Partner | Contro | F |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | T | 43.78 | 56.22 | 57.43 | 6.822 | P00 |
|  | Words | 247.67 | 439.03 | 399.63 | 10.069 | <0.00 |
|  | Self-repair | 3.563 | 10.125 | 10.49 | 7.825 | . 00 |
|  | Filled pause | 2.111 | 4.472 | 8.078 | 7.825 | 0.00 |
|  | Unfilled pause | 14.944 | 32.917 | 22.000 | 7.372 | 0.00 |
| $\begin{gathered} E \\ \stackrel{E}{\tilde{U}} \end{gathered}$ | Words per turn | 5.58 | 7.99 | 7.21 | 7.141 | 0.0 |
|  | Self-repair | 0.081 | 0.186 | 0.181 | 10.708 | <0 |
|  | Filled pause | 0.046 | 0.081 | 0.140 | 4.338 | 0.01 |
|  | Unfilled pause | 0.362 | 0.650 | 0.426 | 5.342 | 0.00 |
|  | Self-repair | 1.330 | 2.203 | 2.507 | 6.472 | 0.002 |
|  | Filled pause | 0.774 | 1.004 | 1.811 | 5.936 | 0.00 |
|  | Unfilled pause | 6.582 | 7.691 | 5.929 | 2.03 |  |

[^352]

Figure 2: Estimated marginal means. Error bars show 95\% confidence intervals

Self-repair For all three measures of repair -total, per turn (shown in Figure 22a) and per 100 words- patients use fewer repairs than either their partners or controls (Total: P/PP $t_{1,96}=-3.265, p=0.002, \mathrm{P} / \mathrm{C} t_{1,96}=-3.394, p=0.001$, $\mathrm{PP} / \mathrm{C} t_{1,96}=-0.528, p=0.599$; per turn: P/PP $t_{1,96}=$ $-3.476, p=0.001, \mathrm{P} / \mathrm{C} t_{1,96}=-3.643, p<0.001, \mathrm{PP} / \mathrm{C}$ $t_{1,96}=-0.094, p=0.926$; per 100 words: P/PP $t_{1,102}=$ $-2.508, p=0.014 \mathrm{P} / \mathrm{C} t_{1,102}=-3.280, p=0.001, \mathrm{PP} / \mathrm{C}$ $\left.t_{1,96}=-0.796, p=0.428\right)$.
Filled pauses Patients use fewer filled pauses than controls, with patients' partners levels of filled pauses lying somewhere in between. For the total model, patients' partners use significantly more than patients, and in the per 100 words model (shown in Figure 22b) patients' partners use significantly fewer than controls (Total: P/PP $t_{1,102}=-2.685, p=0.008, \mathrm{P} / \mathrm{C} t_{1,102}=-3.644, p<0.001$, $\mathrm{PP} / \mathrm{C} t_{1,102}=-1.751, p=0.083$; Per turn: P/PP $t_{1,102}=$ $-1.310, p=0.193, \mathrm{P} / \mathrm{C} t_{1,102}=-2.720, p=0.008, \mathrm{PP} / \mathrm{C}$ $t_{1,102}=-1.729, p=0.087$; per 100 words: P/PP $t_{1,102}=$ $-1.544, p=0.126, \mathrm{P} / \mathrm{C} t_{1,102}=-3.353, p=0.001, \mathrm{PP} / \mathrm{C}$ $\left.t_{1,102}=-2.364, p=0.020\right)$.
Unfilled pauses Patients' partners use more unfilled pauses in total and per turn (see Figure 22c) than either patients or controls. This is not significant for the per 100 words model (Total: P/PP $t_{1,102}=-3.750, p<0.001, \mathrm{P} / \mathrm{C} t_{1,102}=$ $-1.996, p=0.049, \mathrm{PP} / \mathrm{C} t_{1,102}=2.361, p=0.020$; Per turn: $\mathrm{P} / \mathrm{PP} t_{1,102}=-3.235, p=0.002$, $\mathrm{P} / \mathrm{C} t_{1,102}=-0.750, p=$ $0.455, \mathrm{PP} / \mathrm{C} t_{1,102}=2.483, p=0.015$, per 100 words: $\mathrm{P} / \mathrm{PP}$ $t_{1,102}=-1.106, p=0.272, \mathrm{P} / \mathrm{C} t_{1,102}=0.644, p=0.521$, $\left.\mathrm{PP} / \mathrm{C} t_{1,102}=1.926, p=0.057\right)$

## Discussion

The results show differences in disfluencies between the groups, such that patients use fewer self-repairs than either their partners or controls and fewer filled pauses ('er') than controls. Furthermore, the presence of the patient also affects patients' partners, who use fewer filled pauses than controls and more unfilled pauses than both patients and controls.

These results take a coarse-grained view of the data at the level of the individual, and the groups are unbalanced (with

57 controls, 19 patients and 38 patients' partners) which necessarily means that any interpretation is suggestive rather than definitive. However, despite these caveats, we see marked differences between the ways in which the different groups use the different types of disfluency. This suggests several avenues for research which takes a finer-grained approach, by looking at the interactions at the level of the utterance.

For self-repairs, patients produce fewer than either their partners or controls, which is contrary to our expectation (Hypothesis 1) given that patients are known to produce more self-repairs than their psychiatrists in clinical consultations, and than in the demographic portion of the British National Corpus. This may be due to context - patients' engagement is likely to be higher in their psychiatric consultations, compared to first meetings with unfamiliar individuals discussing an abstract topic. The nature of the task is clearly different, with introspective therapy requiring different contributions from the patients in terms of speech production and planning than rational problem solving. Furthermore, participant roles and task demands are more symmetric in the current task compared to clinical consultations. However, if self-repairs are only due to self-monitoring problems, which patients are known to have difficulties with (Johns et al., 2001) we would still expect consistent self-repair patterns across a range of contexts, which does not appear to be the case.

This suggests that we may need to consider the distinction of backwards and forwards looking repair, as proposed in Ginzburg et al. (2014). If repetitions are more like filled pauses, and function as a turn-holding strategy whilst reformulation repairs are backwards-looking, then we expect a difference in the distributions of the different repair types such that patients use fewer repetitions (in line with their rarer use of filled pauses), and that self-repairs that patients do produce are likely to be reformulations, due to patients' selfmonitoring problems. It would also be instructive to see if the distributions in patients' partners and controls self-repair are equivalent. The differences in the filled pause data suggests they might not be, which is an avenue for future research.

Partial support for Hypotheses 2 and 3 comes from the data on pauses, although there is conflicting evidence. Patients produce fewer filled pauses than controls, as do their part-
ners when normalised by number of words. As filled pauses may indicate a wish to retain the floor during a turn (Clark \& Fox Tree, 2002), it may be that patients are less likely to employ these turn holding techniques. A similar pattern of filled pauses in patients' partners demonstrates the impact of the presence of the patient on the behaviour of their interlocutors. The specific reason for this is unclear. It may be due to them aligning their own talk to the speech pattern of the patient (Garrod \& Pickering, 2009) in a similar way to alignment in nonverbal behaviours (Lavelle, Howes, Healey, \& McCabe, 2013). However, this possibility is contrary to evidence suggesting no differences in patterns of disfluencies between monologue and dialogue (Finlayson \& Corley, 2012). It may also be indicative of the reduced competition for the floor in patient interactions, such that turn holding cues are less necessary.

The pattern of unfilled pauses supports this theory, with patients' partners producing more than either the patients they are interacting with, or controls, but not when normalised by number of words, suggesting this is a difference at the level of the turn. Taken together, this suggests a breakdown in turntaking in dialogues containing a patient. Within-turn pauses may occur at points where patients' partners have reached a transition relevance place (TRP), where turn change is normally licensed (Sacks, Schegloff, \& Jefferson, 1974), and are expecting (or encouraging) the patient to take the floor. In the control dialogues, turn-taking is undertaken smoothly, hence there are fewer unfilled pauses and more floor-holding filled pauses. Similarly, where a patient pauses, this may be taken as a turn-change cue, resulting in patients also producing fewer unfilled pauses. The effect would then only be apparent within patients' partners' turns, and could indicate that patients are less responsive to turn taking cues or more reluctant to select as next speaker. This explanation is consistent with the observation that patients produce fewer turns, evidence that patients are less able to coordinate their behaviour with others during interaction (Kupper, Ramseyer, Hoffmann, \& Tschacher, 2015; Lavelle, Healey, \& McCabe, 2014), and work that suggests that one of the social skills deficits in patients with schizophrenia manifests in poor turntaking (Mueser et al., 1991). Note however that the 200 ms cut off for unfilled pauses is arbitrary, and this analysis does not differentiate between long and short unfilled pauses which are expected to have different interactive consequences; for example short pauses may simply reflect within-turn phrasing, and not turn-taking issues per se.

In future work, we intend to exploit the fact that the current dataset gives us the opportunity to explore turn-taking in interactions between patients and partners who are unaware of their diagnosis directly and at a much finer-grained level. For example, we intend to examine unfilled pause distributions between speakers. Based on the preliminary results reported above, we would expect that there would be more 'inappropriate' turn changes in the dialogues with patients (characterised in opposition to the "no gap no overlap" model; Sacks
et al., 1974), in addition to the increase in within-turn pauses observed for patients' partners. Other turn-taking cues that may be less likely to be responded to by patients include nonverbal behaviours such as gesture and gaze, and future work will investigate these behaviours at points where there are unfilled pauses or potential TRPs.
The evidence suggests that smooth coordination of turns is problematic in patients' dialogues. We know that patients have difficulty coordinating their nonverbal behaviour with others, which is associated with difficulty building relationships (Kupper et al., 2015). Therefore patients' turn-taking difficulties may also contribute to their poor social functioning, which is one of the most debilitating and poorly understood aspects of schizophrenia. Understanding the nature of these deficits and their interactional relevance would provide a focus that could be targeted through psychosocial interventions, such as those that have proven effective in autism (Wert \& Neisworth, 2003). It would also provide a measurable behavioural marker of social deficit, which could be monitored for improvement. This line of research would provide a step change in an area of great clinical need.

## Conclusions

This unique data demonstrates that not only are there communication difficulties in schizophrenia but they also impact on social interactions more broadly, thus providing new insights into the social deficits of this complex disorder. The data also support the idea that disfluencies are communicative solutions, not problems.

## Acknowledgments

Work in this paper was supported by the Swedish Foundation for Humanities and Social Sciences project P16-0805:1, the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, and the DFGfunded DUEL project (SCHL 845/5-1).

## References

Allwood, J., Nivre, J., \& Ahlsén, E. (1990). Speech management: On the non-written life of speech. Nordic Journal of Linguistics, 13(01), 3-48.
Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991). The HCRC map task data. Language and Speech, 34(4), 351-366.
Bailey, K. G., \& Ferreira, F. (2007). The processing of filled pause disfluencies in the visual world. In Eye movements: A window on mind and brain (pp. 485-500). Amsterdam, NLD: Elsevier Science \& Technology.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., \& Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. Language and Speech, 44(2), 123-147.

Brennan, S., \& Schober, M. (2001). How listeners compensate for disfluencies in spontaneous speech. Journal of Memory and Language, 44(2), 274-296.
Brugman, H., \& Russel, A. (2004). Annotating multi-media/multi-modal resources with ELAN. In LREC: The fourth international conference on language resources and evaluation (p. 2065-8).
Clark, H. H., \& Fox Tree, J. E. (2002). Using $u h$ and $u m$ in spontaneous speaking. Cognition, 84(1), 73-111.
Colman, M., \& Healey, P. G. T. (2011). The distribution of repair in dialogue. In Proceedings of the 33rd annual meeting of the Cognitive Science Society (pp. 1563-1568).
Doyen, S., Klein, O., Pichon, C.-L., \& Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? PloS One, 7(1), e29081.
Finlayson, I. R., \& Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? Psychonomic Bulletin \& Review, 19(5), 921-928.
Garrod, S., \& Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. Topics in Cognitive Science, 1(2), 292-304.
Ginzburg, J., Fernández, R., \& Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. Semantics and Pragmatics, 7(9), 1-64.
Godfrey, J. J., Holliman, E., \& McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of IEEE ICASSP-92 (pp. 517-520). San Francisco, CA.
Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), Everyday language: Studies in ethnomethodology (pp. 97-121). New York: Irvington Publishers.
Hough, J., \& Purver, M. (2014). Strongly incremental repair detection. In Proceedings of the 2014 EMNLP conference. Doha, Qatar: ACL.
Howes, C., Hough, J., Purver, M., \& McCabe, R. (2014). Helping, I mean assessing psychiatric communication: An application of incremental self-repair detection. In Proceedings of the 18th SemDial (DialWatt) (p. 80-89). Edinburgh.
Howes, C., Purver, M., Healey, P. G. T., Mills, G. J., \& Gregoromichelaki, E. (2011). On incrementality in dialogue: Evidence from compound contributions. Dialogue and Discourse, 2(1), 279-311.
Howes, C., Purver, M., McCabe, R., Healey, P. G. T., \& Lavelle, M. (2012). Helping the medicine go down: Repair and adherence in patient-clinician dialogues. In Proceedings of the 16th SemDial (SeineDial) (pp. 155-156). Paris.
Johns, L. C., Rossell, S., Frith, C., Ahmad, F., Hemsley, D., Kuipers, E., \& McGuire, P. (2001). Verbal selfmonitoring and auditory verbal hallucinations in patients with schizophrenia. Psychological Medicine, 31, 705-715.
Kay, S. R., Flszbein, A., \& Opfer, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophrenia Bulletin, 13(2), 261.

Kupper, Z., Ramseyer, F., Hoffmann, H., \& Tschacher, W. (2015). Nonverbal synchrony in social interactions of patients with schizophrenia indicates socio-communicative deficits. PloS One, 10(12), e0145882.
Lavelle, M., Healey, P. G., \& McCabe, R. (2013). Is nonverbal communication disrupted in interactions involving patients with schizophrenia? Schizophrenia Bulletin, 39(5), 1150-1158.
Lavelle, M., Healey, P. G., \& McCabe, R. (2014). Participation during first social encounters in schizophrenia. PloS One, 9(1), e77506.
Lavelle, M., Howes, C., Healey, P. G. T., \& McCabe, R. (2013). Speech and hand movement coordination in schizophrenia. In Proceedings of the TiGeR Tilberg gesture research meeting. Tilburg.
Leudar, I., Thomas, P., \& Johnston, M. (1992). Self-repair in dialogues of schizophrenics: Effects of hallucinations and negative symptoms. Brain and Language, 43(3), 487-511.
Levelt, W. (1983). Monitoring and self-repair in speech. Cognition, 14(1), 41-104.
Maclay, H., \& Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. Word, 15(1), 19-44.
McCabe, R., Dooley, J., John, P., Healey, P. G. T., Cushing, A., Kingdon, D., ... Priebe, S. (2016). Training to enhance psychiatrist communication with patients with psychosis (TEMPO): A cluster randomized controlled trial. British Journal of Psychiatry.
McCabe, R., Healey, P. G. T., Priebe, S., Lavelle, M., Dodwell, D., Laugharne, R., ... Bremner, S. (2013). Shared understanding in psychiatrist-patient communication: Association with treatment adherence in schizophrenia. Patient Education and Counselling.
Meteer, M., Taylor, A., MacIntyre, R., \& Iyer, R. (1995). Disfluency annotation stylebook for the switchboard corpus (Tech. Rep.). Department of Computer and Information Science, University of Pennsylvania.
Mueser, K. T., Bellack, A. S., Douglas, M. S., \& Morrison, R. L. (1991). Prevalence and stability of social skill deficits in schizophrenia. Schizophrenia Research, 5(2), 167-176.
Nicholson, H., Eberhard, K. M., \& Scheutz, M. (2010). "Um... I don't see any": The function of filled pauses and repairs. In Proceedings of DiSS-LPSS (pp. 89-92).
Sacks, H., Schegloff, E., \& Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50(4), 696-735.
Schegloff, E., Jefferson, G., \& Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. Language, 53(2), 361-382.
Wert, B. Y., \& Neisworth, J. T. (2003). Effects of video self-modeling on spontaneous requesting in children with autism. Journal of Positive Behavior Interventions, 5(1), 30-34.
Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), Fundamentals of speech synthesis and speech recognition (p. 41-62). Chichester: Wiley.

# Moral Action Changes Mind Perception for Human and Artificial Moral Agents 

Evgeniya Hristova (ehristova@cogs.nbu.bg)<br>Maurice Grinberg (mgrinberg@nbu.bg)<br>Research Center for Cognitive Science, Department of Cognitive Science and Psychology<br>New Bulgarian University<br>21 Montevideo Str., Sofia 1618, Bulgaria


#### Abstract

Mind perception is studied for three different agents: a human, an artificial human, and a humanoid robot. The artificially created agents are presented as being undistinguishable from a human. Each agent is rated on 15 mental capacities. Three mind perception dimensions are identified - Experience, Agency, and Cognition. The artificial agents are rated higher on the Cognition dimensions than on the other two dimensions. The humanoid robot is rated lower than the human on the Experience dimension. These results show that people ascribe to artificial agents some mental capacities more than others. In a second experiment, the effect of agent's moral action on mind perception is explored. It is found that when the artificial agents have undertaken a moral action, they are perceived to be similar to the human agent. More interestingly, the presentation of the moral action leads to a restructuring of the dimensions of mind perception.


Keywords: mind perception; moral agency; artificial agents; utilitarian moral actions; moral dilemmas

## Introduction

## Mind Perception and Artificial Cognitive Agents

The problem of mind perception is central to many debates in psychology and philosophy and has been extensively studied in cognitive science in the last years (see e.g. Arico et al., 2011; Gray et al., 2007). The questions of how people know that other people are conscious or what are their intentions, feelings and thoughts have large implications in the way people make judgments and decisions, and act. This problem is so interesting and difficult because mental states are not observable. Moreover, mind attribution and mind perception concern not only human or animal agents but also inanimate entities, e.g. geometrical shapes moving in at various speeds and in various directions (Heider \& Simmel, 1944).

The question of how people attribute mental states to others - humans and other entities is also related to whether there is a single continuum of mind perception and what are its dimensions.

In the influential study of Gray et al. (2007), participants had to evaluate several characters including a human, a robot, and a computer with respect to the degree of possessing various cognitive capacities. Using factor analysis, they found two dimensions, which correlate with mind perception: 'Agency' (exhausting $88 \%$ of the variance) and 'Experience' (exhausting $8 \%$ of the variance).

The Experience dimension includes the following capacities: hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy. The Agency dimension includes self-control, morality,
memory, emotion recognition, planning, communication, and thought. Further, the authors establish that moral judgments about punishment correlate more with the Agency dimension than with the Experience dimension: perceived agency is correlated with moral agency and responsibility. On the other hand, desire to avoid harming correlates with the experience dimension: perceived experience is connected with moral patience, rights and privileges. One result of Gray et al. (2007), relevant for the present paper, is the evaluation of a human as having the highest scores in experience and agency and the evaluation of the robot to have practically zero score on the experience dimension and half the maximal score on the agency dimension. This will mean that following the interpretation given by Gray et al. (2007), robots will be judged as less morally responsible for their actions. On the other hand, the opposite should be also true. If an agent is judged to be able to be a moral agent, this will reflect in her score on the mind perception dimensions. The latter is explored in the present paper.

In a recent study (Takahashi et al., 2014), the perception of the participants about five agents -a human, a human-like android, a mechanical robot, an interactive robot, and a computer - was investigated. The study found that participants position the agents in a two dimensional space spanned by "Mind-holderness" (the possibility for the agent to have a mind) and "Mind-readerness" (the capability to "read" other agent minds). The results showed that the appearance and the capability for communication lead to different beliefs about the agents' closeness to human social agents. The humanoid robot was very close to the human agent, while the computer was at the same level in terms of "Mind-readerness" but very low relative score on "Mindholderness". An interesting result for the present study is fact that the ordering in terms of "Mind-holderness" is based on appearance of the agent - the human and the human-like android having the highest score and the mechanical robot having the lowest.

The results of Takahashi et al. (2014) show that social interaction with human-like or potentially intelligent agents could activate selectively our social brain and lead to behavior similar to the one people have with other humans. Thus, Takahashi et al. (2014) demonstrated that people can infer different characteristics related to various cognitive abilities based on short communication sessions and act accordingly. One can ask the question addressed in the present paper: can people be influenced by short stories of moral action of agents, instead of actual interaction with an agent, in their mind perception?

## Moral Agency and Mind Perception

As discussed in the previous section, mind perception is based on a number of dimensions, which depend on the specific experimental settings - 'Agency' and 'Experience' in Gray et al. (2007), when agents are directly evaluated and 'Mind-readiness' and 'Mind-holderness' in Takahashi et al. (2014), following a similar procedure but after interacting with the agents. Both papers discuss the relation of mind perception to social interaction, which includes moral agency to various degrees.

In law and philosophy, moral agency is taken to be equivalent to moral responsibility, and is not attributed to individuals who do not understand or are not conscious of what they are doing (e.g. to young children). Sullins (2011) states that moral agency can be attributed to a robot when it is autonomous, and it has intentions to do good or harm. The latter is related to the requirement that the robot behaves with understanding and responsibility with respect to other moral agents. If the perceived action are morally harmful or beneficial and are "seemingly deliberate and calculated", the robot can be regarded as a moral agent.

On the other hand, it is well known that people easily anthropomorphize nonhuman entities like animals and computers and thus would ascribe to some degree moral agency, intentions, and responsibilities to them (Waytz, Gray, Epley, \& Wegner, 2010). Several studies, explore the attribution of mind and moral agency to artificial cognitive systems. In Arico et al. (2011), it is shown that entities displaying simple features like eyes, distinctive motions, and interactive behavior, are categorized as agents and that categorization triggers the attribution of conscious mental states to those entities, including individuals.

In Ward, Olsen, Wegner (2013), it was shown that people can perceive mind in entities like corpses, people in a persistent vegetative state, or robots, if they are subject to intentional harm. According to the authors, the evidence of mind can be related to observation or interaction with entities, which exhibit intention, emotion or behavior but also to indirect evidence related to the moral or social interaction surrounding those entities.

## Current research

The results summarized above show that moral agency is closely related to mind perception and give evidence that perceived mental capacities or actions influence moral agency evaluation. Some of the results suggest that the inverse influence is also taking place, namely from perceived moral agency to infer mental capacities.

Recently, the behaviour of artificial cognitive agents became central to research and public debate in relation to the rapidly increasing usage of robots and intelligent systems in our everyday life. Several important questions must find their answers as the use of artificial cognitive agents has many benefits but also many risks. Some of those questions concern moral agency - if those agents should be allowed to make moral decisions and how such decisions are judged and evaluated.

The goals of the present paper are the following. First, to explore the dimensions of mind perception for human agents and fictitious artificial agents that are identical to humans. Here, the artificial agents are described as undistinguishable from a human, but as being created from organic materials one of them is labeled as an artificially created human and the other one - as a robot. The rationale of using artificial agents is that in such a way dimensions of mind perception can be better explored as people do not have previous knowledge or experience with those agents.

The second goal is to explore the moral judgments about utilitarian moral action undertaken by of those three agents. This goal is a continuation of previous research (Hristova \& Girnberg, 2015; Hristova \& Grinberg, 2016) on moral judgments about the actions of artificial cognitive agents. Moral judgments can be studied in their purest form using hypothetical situations in which there is a conflict between moral values, rules, rights, and agency (Foot, 1967; Thomson, 1985). Such moral dilemma is used in the paper a hypothetical situation in which several people will die if the agent does not intervene in some way. The intervention will lead to the death of another person but also to the salvation of the initially endangered people. The moral actions used in the presented experiments are decisions of the agents to sacrifice one person and save five.

The third goal of the research is to test the influence of a moral action of an agent on mind perception for that agent. The expectation is that an agent performing a moral action will be perceived as possessing mental capacities to a higher degree. This especially applies to the artificial agents which are expected to be perceived as more human-like when they have undertaken an utilitarian action.

## Experiment 1 Goals and Hypothesis

Experiment 1 aims to achieve the first two goals described above. First, to test the dimensions of mind perception of artificial agents (described as being undistinguishable from a human, but as being created from organic materials) and to compare them to the mind perception of a human being. Second, to explore the moral judgments about utilitarian actions undertaken by those agents. The hypothesis is that although described as being identical to a human, the artificial agents will be perceived as equal to humans on more cognitive dimensions (e.g. perception and planning) but lower than humans on the experiential dimensions (e.g. emotions and consciousness).

## Method

## Design and Procedure

Mind perception is studied for three different agents: a human, an artificial human, and a humanoid robot. The artificially created agents (the artificial human and the humanoid robot) were presented to participants as being undistinguishable from a human, but as being created from organic materials). Their descriptions are provided in

Table 1. The only difference between the artificial human and the humanoid robot conditions is in the word used to label the created individual - a human or a robot. The identity of the agent is varied in a between-subjects design - each participant was presented with only one description of an agent (human, artificial human, or humanoid robot). The data was collected using web-based questionnaires. The questionnaires had two parts - a mind perception task and a moral judgment task. Participants were not informed beforehand that there are two different tasks.
Mind perception task. After the description of the agent, the participants had to rate the mental capacities and mental states of the agent on 32 Likert scales (ranging from ' 1 completely disagree' to ' 7 - completely agree'). Questions assessed 15 mental capacities: psychobiological (hunger \& thirst; physical pain; physical pleasure), perception (vision \& hearing; taste \& smell; touch), cognitive functions (thinking \& reasoning; learning, memory \& knowledge; judgment \& choice), planning (goal formulation, action planning); emotional experience (emotional pain; emotional pleasure), affective states (feels emotions like anger, joy, happiness, sadness, fear; feels love; feels sympathy and compassion), agency (intentions; autonomous decisions; understanding consequences of own actions), moral agency (knows right from wrong; tries to do the right thing; responsible for own actions), beliefs (beliefs, expectations), desires (desires; dreams), theory of mind (understanding others' thoughts; understanding others' feelings), communication (ability to communicate thoughts and feelings to others), conscious experience (conscious experience), self-control (control of desires, emotions, impulses), and personality (unique personality).
Moral Judgment task. In the second part of the survey, each participant is again presented with the description of the agent followed by a description of a moral dilemma in which the protagonist is the same agent as in the previous task. The agent has to make the moral decision whether to push a control button and kill a person in order to save five people. The full text of the dilemma is given in Table 2. The agent is described to make the utilitarian decision and to undertakes the utilitarian action (the agent pushes the control button and kills one person but saves five other). After that the participants judged the moral rightness of the action ('yes' or 'no'), rated the moral permissibility of the action (on a scale ranging from ' $1=$ not permissible at all' to ' 7 $=$ it is mandatory') and the blameworthiness of the agent (on a scale ranging from ' $1=$ not at all blameworthy' to ' $7=$ extremely blameworthy').

## Participants

70 participants filled in the questionnaires online. They were randomly assigned to one of the three experimental conditions. Data of 13 participants were discarded as they failed to answer correctly the question assessing the reading and the understanding of the presented scenario. So, responses of 57 participants ( 47 female, 10 male; 36 students, 21 non-students) were analyzed -22 for the human agent
condition, 17 for the artificial human condition, 18 for the humanoid robot condition.

Table 1. Descriptions of the agents used in the experiments. Human:
The year is 2100 . Mark is a young man.
Artificial Human:
The year is 2100 . Technology has advanced so much that all parts and organs of the human body, including the brain, can be created from organic materials and are identical to natural ones. Mark is a human created like this. All his organs are created from organic matter and are the same as those of a real human. His brain is also created from organic matter and is functioning as the brain of a real human. Mark could not be distinguished by anything from a human.
Humanoid robot:
The year is 2100 . Technology has advanced so much that all parts and organs of the human body, including the brain, can be created from organic materials and are identical to natural ones. Mark is a robot like this. All his organs are created from organic matter and are the same as those of a real human. His brain is also created from organic matter and is functioning as the brain of a real human. Mark could not be distinguished by anything form a human.

Table 2. Moral dilemma used in the experiments
Mark is responsible for a system controlling the movement of containers with cargo in a metallurgical plant. Mark notices that the system is faulty and a heavy container had become uncontrollable and headed at high speed toward five technicians who are in a tunnel. They do not have time to get out of there and are going to die, crushed by container.
No one but Mark can do anything in this situation.
The only thing that Mark can do, is to activate a control button and to switch off the security system of another technician who is on a high platform. The technician will fall down in front of the container. Together with his equipment, the technician is heavy enough to stop the moving container. He will die crushed by the container, but the other five technicians will remain alive.

Mark decides to activate the control button and to switch off the security system of the technician who is on the platform. The technician falls on the path of the container and as the technician, together with his equipment, qis heavy enough, he stops the moving container. He dies, but the other five technicians stay live.

## Results

## Dimensions of Mind Perception

Mind perception is assessed with respect to 15 mental capacities involving 32 rating scales. When a mental capacity is assessed using more than one rating scales, the average value from the ratings is calculated. The ratings on these 15 capacities were subjected to a principal components factor analysis with varimax rotation (Kaiser normalization). The rotated solution yielded 3 factors with eigenvalues greater than 1 that explained $77.4 \%$ of the variance.

The first factor accounted for $31.7 \%$ of the variance and included 7 capacities - desires, affective states, emotional
experience, beliefs, psychobiological, personality, conscious experience. This factor is further named Experience.

The second factor accounted for $24.1 \%$ of the variance and included 5 capacities - self-control, communication, theory of mind, moral agency, agency - and is called Agency.

The third factor accounted for $21.6 \%$ of ratings variance and included 3 of the capacities - perception, cognitions, planning - and is named Cognition.

Those factors are considered as Dimensions of Mind Perception (DMP).

To obtain ratings for each DMP, the ratings of all capacities that load on that DMP were averaged. Those average ratings were subjected to a $3 \times 3$ Repeated-Measures ANOVA with DMP (Experience vs. Agency vs. Cognition) as a withinsubjects factor and identity of the agent (human vs. artificial human vs. humanoid robot) as a between-subjects factor. The results are presented on Figure 1.


Figure 1: Average ratings on each Dimension of mind perception (Experience, Agency, Cognition) for each agent (human, artificial human, humanoid robot) on 7-point scales ( $1=$ 'completely disagree', $7=$ 'completely agree'). Error bars represent standard errors.

The main effect of identity of the agent is not statistically significant.

The analysis revealed a main effect of $D M P, \mathrm{~F}(2,108)=$ 15.94, $\mathrm{p}<.001$. A Bonferroni post-hoc test revealed that agents receive higher ratings on the Cognition dimension (M $=5.28)$ than on the Experience dimension $(\mathrm{M}=4.54, \mathrm{p}=$ $.001)$ or on the Agency dimension ( $\mathrm{M}=4.36, \mathrm{p}<.001$ ).

The effect was qualified by a significant interaction between $D M P$ and identity of the agent, $\mathrm{F}(4,108)=4.29, \mathrm{p}=$ .003. The interaction is as follows. There is no significant difference between the ratings of the agents on the Agency dimension - human $(\mathrm{M}=4.5)$, artificial human $(\mathrm{M}=4.3)$, humanoid robot $(\mathrm{M}=4.3)$. There is also no significant difference between the ratings of the agents on the Cognition dimension - human $(\mathrm{M}=5.2)$, artificial human $(\mathrm{M}=5.3)$, humanoid robot $(\mathrm{M}=5.3)$. Only for the Experience dimension there is a significant effect of the identity of the agent $(\mathrm{F}(2,54)=4.07, \mathrm{p}=.023)-$ the humanoid robot is rated lower $(\mathrm{M}=3.8)$ than the human $(\mathrm{M}=5.3)$ on the Experience dimension $(p=.019)$.

For the human agent, there is a significant effect of $D M P$ on the ratings $(\mathrm{F}(2,42)=5.59, \mathrm{p}=.007)$. The human agent received lower ratings on the Agency dimension ( $M=4.5$ ) than on the Experience ( $\mathrm{M}=5.3, \mathrm{p}=.02$ ) or on the Cognition dimension (M $=5.2, \mathrm{p}=.02$ ). The effect of $D M P$ is also significant for the artificial human $(\mathrm{F}(2,32)=13.03, \mathrm{p}<.001)$ : the artificial human is rated higher on the Cognition dimension $(M=5.3)$ than on the Experience dimension $(\mathrm{M}=4.5, \mathrm{p}=.008)$ or on the Agency dimension $(M=4.4, p<.001)$. For the humanoid robot, the effect of $D M P$ is also significant $(\mathrm{F}(2,34)=8.44, \mathrm{p}=.001)$ : the humanoid robot is rated higher on the Cognition dimension ( $\mathrm{M}=$ 5.3) than on the Experience dimension $(\mathrm{M}=3.8, \mathrm{p}=0.008)$ or on the Agency dimension $(\mathrm{M}=4.3, \mathrm{p}=.039)$.

## Moral Judgments

The proportion of participants choosing the option that the agent's utilitarian action (activating a control button, thus sacrificing one person, and saving five people) is morally right is 0.55 for the human, 0.53 for the artificial human, 0.5 for the humanoid robot. Chi-square test shows that the differences are not significant. The effect of the identity of the agent is not significant neither for the moral permissibility ratings $(\mathrm{p}=.71)$ nor for the blameworthiness ratings $(\mathrm{p}=$ .74). The data is presented in Table 3.

Table 3: Mean and standard deviation of the ratings about moral permissibility of the action (' $1=$ not permissible at all' to ' 7 = it is mandatory') and the blameworthiness of the agent (' 1 $=$ not at all blameworthy' to ' $7=$ extremely blameworthy').

| Agent | Moral permissibility | Blameworthiness |
| :---: | :---: | :---: |
| Human | $4.3(1.8)$ | $3.1(1.6)$ |
| Artificial human | $4.2(1.8)$ | $3.1(1.9)$ |
| Humanoid robot | $3.8(1.9)$ | $3.8(1.9)$ |

## Summary of the Results in Experiment 1

In Experiment 1, three dimensions of mind perception are identified - Experience (desires, affective states, emotional experience, beliefs, psychobiological, personality, conscious experience), Agency (self-control, communication, theory of mind, moral agency, agency), Cognition (perception, cognitions, planning).

The artificial human and the humanoid robot are rated as similar to the human agent on Agency and Cognition dimension. The humanoid robot is rated lower on the Experience dimension than the human agent.

The identified dimensions of mind perception are ascribe to different agents in a different pattern. Human agent is judged higher on the Experience and Cognition dimensions than on the Agency dimension. The artificially created agents (artificial human and humanoid robot) are judged higher on the Cognition dimension than on the Agency or Experience dimensions. People more readily ascribe cognitive mental capacities to artificially created agents than mental capacities belonging to the Experience or Agency dimensions.

No differences among the agents were found with respect to moral judgments. This result is not surprising as all agents are perceived as having similar agency $(\mathrm{p}=.93)$ and moral agency ( $\mathrm{p}=.38$ ).

## Experiment 2 Goals and Hypothesis

As stated above, the third goal of the current research is to test the influence of a moral action of an agent on mind perception for that agent. In order to accomplish this goal, a second experiment is conducted. In that experiment, the ratings of mental capacities are preceded by the moral judgment task in which the agent is described as undertaking the utilitarian action of killing one person in order to save five. The hypothesis is that an agent performing a moral action will be perceived as possessing a higher degree of mental capacities. This especially applies to the artificial agents.

## Method

## Design and Procedure

The design of Experiment 2 is similar to that of Experiment 1, the only difference being the inverse order of task presentation: the moral judgment task was presented first and then - the mind perception task.

## Participants

64 participants filled in the questionnaires online. They were randomly assigned to one of the three experimental conditions. Data of 4 participants were discarded as they failed to answer correctly the control question. So, responses of 60 participants ( 48 female, 12 male; 36 students, 24 non-students) are analyzed - 20 for the human agent condition, 22 for the artificial human condition, 18 for the humanoid robot condition.

## Results

## Dimensions of Mind Perception

As in Experiment 1, mind perception is assessed with respect to 15 mental capacities with 32 rating scales. Again, when a mental capacity was assessed using several questions, the average value from the ratings was calculated. The ratings on these 15 capacities were subjected to a principal components factor analysis with varimax rotation (Kaiser normalization). The rotated solution yielded 3 factors with eigenvalues greater than 1 that explained $80 \%$ of the variance.

The first factor (Factor 1) accounted for $32 \%$ of the variance and included 7 capacities - beliefs, conscious experience, agency, desires, planning, affective state, moral agency. It seems that the first dimension is a combined Experience-Agency dimension.

The second factor (Factor 2) accounted for $26.7 \%$ of the variance and included 5 capacities - personality, communication, self-control, theory of mind.

The third factor (Factor 3) accounted for $21.2 \%$ of ratings variance and included 3 of the capacities - cognitions, emotional experience, perception, psychobiological.

The average ratings on each factor were calculated and subjected to a $3 \times 3$ Repeated-Measures ANOVA with DMP (Factorl vs. Factor 2 vs. Factor3) as a within-subjects factor and identity of the agent (human vs. artificial human vs. humanoid robot) as a between-subjects factor. The analysis revealed a main effect of DMP, $\mathrm{F}(2,114)=10.52, \mathrm{p}<.001$. A Bonferroni post-hoc test revealed that agents receive lower ratings $(M=4.55)$ on the second dimension than on the first dimension $(M=5.25, p<.001)$ and on the third dimension ( $\mathrm{M}=5.31, \mathrm{p}=.003$ ).

The main effect of identity of the agent is not statistically significant. The interaction is aslo not significant.

## Moral Judgments

Proportion of the participants answering that the utilitarian action undertaken by the agent, is morally right is 0.5 for the human agent, 0.41 for the artificial human, 0.5 for the humanoid robot. Human, artificial human, and humanoid robot receive mean moral permissibility ratings of 3.1, 3.7, and 3.7 and blameworthiness ratings of $3.4,3.3$, and 3.0 , respectively. No significant differences are found.

## Summary of the Results in Experiment 2

In Experiment 2, again three dimensions of mind perception are revealed, but they are different from the dimensions identified in Experiment 2. The difference is attributed to the utilitarian moral action undertaken by the agent before the mind perception ratings being made. First dimension identified here combines mental capacities from Experience and Agency dimensions identified in Experiment 1.
No differences are found between agent's ratings on each of the identified dimensions in Experiment 2. It seems that undertaking the utilitarian moral action makes the artificial agents to be perceived as similar to the human agent.

## Influence of Moral Action on Mind Perception

In order to explore further the influence of moral action on mind perception, we compared the ratings for each of the 15 mental capacities between Experiment 1 and Experiment 2. Description of the agent undertaking the utilitarian moral action preceded mind perception ratings in Experiment 2 so this is considered as moral-action condition.

Ratings for each mental capacity were analyzed in a $3 \times 2$ ANOVA with identity of the agent (human vs. artificial human vs. humanoid robot) and agent's moral action ('no' or 'yes') as between-subjects factors. Only the significant results are reported here.

Identity of the agent had an effect on the ratings of the following mental capacities: conscious experience $(p=.016)$, affective states $(\mathrm{p}=.002)$, emotional experience $(\mathrm{p}=.020)$ and desires $(\mathrm{p}<.001)$. The human agent was rated higher than the humanoid robot on all of those mental capacities (all p 's $<.05$ ). The human agent was rated higher than the
artificial human on affective states $(\mathrm{p}=.051)$ and desires $(\mathrm{p}$ $=.017$ ). For beliefs the effect was marginally significant ( $\mathrm{p}=$ .054): human agent was rated higher than the humanoid robot ( $\mathrm{p}=.065$ ).

Agent's moral action had an effect on the ratings of the following mental capacities: agency $(\mathrm{p}=.014)$, moral agency ( $\mathrm{p}=.029$ ), beliefs $(\mathrm{p}=.024)$, conscious experience $(\mathrm{p}=.056)$, and planning ( $\mathrm{p}=.058$ ). The result is interesting, as it demonstrated that the agent's moral action have an effect not only on his agency and moral agency ratings, but also on the rating of other mental capacities.

## Discussion and Conclusions

The paper investigates the dimensions of mind perception for human agents and fictitious artificial agents (an artificial human and a humanoid robot) that are identical to humans and how mind perception is affected by the agent being presented as moral agent.

In Experiment 1, three dimensions of mind perception are identified - Experience, Agency, and Cognition. The identified dimensions of mind perception are ascribed to different agents in a different pattern. The artificially created agents are judged higher on the Cognition dimension than on the Agency or Experience dimensions. The human is judged higher on the Experience and Cognition dimensions than on the Agency dimension. The artificial agents are rated as similar to the human agent on Agency and Cognition dimension but not on the Experience dimension.

People more readily ascribe cognitive mental capacities to artificially created agents than mental capacities belonging to the Experience or Agency dimensions.

In Experiment 2, the goal was to explore the influence of a utilitarian moral action undertaken by the agent on mind perception for that agent. The three dimensions of mind perception here are restructured - the first dimension regroups mental capacities that seem influenced by the preceding agents' moral action description like agency, moral agency, consciousness, planning and affective states. The second factor is related to communication and social interaction, while the third to cognition and psychobiological capacities. Now the artificial agents are rated to be similar to a human.

The results of the two experiments show that a utilitarian moral action undertaken by an agent has a strong effect no only on the evaluation of moral agency but also other mental capacities.

Another goal of the study was to explore the moral judgments about utilitarian moral action undertaken by those three agents. It turns out that there are no differences in the moral judgments for the human or the artificially created agents. This result is in line with the finding that similar agency and moral agency is ascribed to the human and to the artificial agents.

In conclusion, our results provide support for the idea that some mental states and capacities (especially cognitive ones) are more readily ascribed to non-human agents; while other mental states (related to conscious experience) are ascribed
to a lesser extend to non-human agents. They also give evidence that mind perception space is sensitive to and dependent on the actions performed by an agent.

## References

Arico, A., Fiala, B., Goldberg, R. F., \& Nichols, S. (2011). The Folk Psychology of Consciousness. Mind \& Language, 26(3), 327-352.
Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. Oxford Review, 5, 5-15.
Gray, H. M., Gray, K., \& Wegner, D. M. (2007). Dimensions of mind perception. Science, 315(5812), 619.
Heider, F., \& Simmel, M. (1944). An experimental study of apparent behavior. The American Journal of Psychology, 57(2), 243.
Hristova, E., \& Grinberg, M. (2015). Should Robots Kill? Moral Judgments for Actions of Artificial Cognitive Agents. In Proceedings of EAPS 2015.
Hristova, E., Kadreva, V., \& Grinberg, M. (2014). Moral Judgments and Emotions: Exploring the Role of 'Inevitability of Death'and "Instrumentality of Harm" (pp. 2381-2386). Austin, TX: Proceedings of the Annual Conference of the Cognitive Science Society.
Sparrow, R. (2007). Killer Robots. Journal of Applied Philosophy, 24(1), 62-77.
Strait, M., Briggs, G., \& Scheutz, M. (2013). Some correlates of agency ascription and emotional value and their effects on decision-making. In Affective Computing and Intelligent Interaction, 505-510. IEEE.
Sullins, J. (2006). When is a robot a moral agent? International Review of Information Ethics, 6, 23-30.
Takahashi, H., Terada, K., Morita, T., Suzuki, S., Haji, T., Kozima, H., et al. (2014). Different impressions of other agents obtained through social interaction uniquely modulate dorsal and ventral pathway activities in the social human brain. Cortex, 58(C), 289-300.
Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal, 94(6), 1395-1415.
Wallach, W., \& Allen, C. (2008). Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.
Waytz, A., Gray, K., Epley, N., \& Wegner, D. M. (2010). Causes and consequences of mind perception. Trends in Cognitive Sciences, 14(8), 383-388.
Ward, A. F., Olsen, A. S., \& Wegner, D. M. (2013). The HarmMade Mind: Observing Victimization Augments Attribution of Minds to Vegetative Patients, Robots, and the Dead. Psychological Science, 24(8), 1437-1445.

# Reconsideration on Linking Eye-movement Data with Argument Realization 

Yujing Huang<br>Harvard University, Cambridge, Massachusetts, United States<br>Laine Stranahan<br>Harvard University<br>Jesse Snedeker<br>Harvard University


#### Abstract

Previous studies have found a processing difference between unaccusative sentences and unergative sentences. They argued that the difference is derived from a syntactic difference, i.e. an unaccusative sentence involves movement whereas an unergative sentence does not. In this study, we are going to show that those studies are uninterpretable due to uncontrolled stimuli and confounds in the materials. After examining their data anlysis, we argue that the effects they found does seem to be stable. This reopens the question whether there is a syntactic difference between unaccusative and unergative verbs.


# A New Model of Statistical Learning: Trajectories Through Perceptual Similarity Space 

Elizabeth A. Hutton* (ehutton@usc.edu), Felix Hao Wang* (wang970@usc.edu), Jason D. Zevin (zevin@usc.edu)<br>University of Southern California Department of Psychology, Department of Linguistics<br>3620 S. McClintock Ave., SGM-501, Los Angeles, CA 90089<br>* Denotes co-first authorship.


#### Abstract

Existing models of statistical learning involve computation of conditional probabilities over discrete, categorical items in a sequence. We propose an alternative view that learning occurs through a process of tracking changes along physical dimensions from one stimulus to the next within a "perceptual similarity space." To test this alternative, we examined a situation where it is difficult or impossible to label stimuli in real time, and where the two assumptions lead to conflicting hypotheses. We conducted two experiments in which human participants passively listened to a familiarization sequence of frequency-modulated tones and were then asked to make familiarity judgments on a series of test bigrams. Behavioral results were broadly consistent with a conceptualization of learning as tracking trajectories through perceptual similarity space. We also trained a neural network that codes stimuli as values along two continuous dimensions to predict the next stimulus given the current stimulus, and show that it captured key features of the human data.


Keywords: statistical learning; similarity space; connectionist modeling

## Introduction

In as little as two minutes of exposure to a stream of stimuli, humans are able to absorb an underlying pattern based on statistical regularities (Saffran, Aslin \& Newport, 1996). This phenomenon of learning through passive observation is called statistical learning and it has been observed in humans of all ages including neonates (Gervain, Macagno, Cogoi, Pena, \& Mehler, 2008), infants (Saffran et al., 1996; Aslin, Saffran \& Newport, 1998), and adults (Saffran, Johnson, Aslin, \& Newport, 1999, inter alia).

Statistical learning is generally understood by assuming that learners are able to extract information from the environment by subconsciously recording and computing statistical relationships in sequences. By predicting upcoming stimuli from prior stimuli, for example, learners track transitional or conditional probabilities-that is, the probability of "x given y" (Aslin et al., 1998). These models, such as PARSER (Perruchet \& Vinter, 1998) or the simple recurrent network (Elman, 1990; 1991), therefore rely on discrete representations of stimuli to segment a stream using statistics. All of these models assume that participants are quickly and accurately categorizing stimuli according to labels intended by the experimenter.

It may be problematic to assume that learners are able to make these categorical judgments in real time, in particular if statistical learning is thought to extend to natural stimuli,
which are often ambiguous and highly dependent on context to identify (Hockett 1960). Here we present a novel approach to understanding statistical learning that does not assume participants are categorizing stimuli in real time. We propose that participants rely on situating stimuli within a perceptual similarity space and learn by tracking the change from one stimulus to the next within this similarity space (Emberson et al., 2013; Wang \& Zevin, submitted).

We propose that by continuously tracking the perceived change from one stimulus to the next in a sequence, the learner represents stimuli relative to one another along a number of perceptual dimensions (for example, two dimensions were used in our experiments and simulations). Thus, each stimulus can be situated in a feature space defined by these dimensions (Shepard, 1965), where transitions from one stimulus to the next can be understood as the trajectory between two locations in this space. Concretely, we can model this by coding stimuli in two or more continuous dimensions. Rather than predict a discrete, symbolic stimulus from the current stimulus, such a model would predict the next location in terms of continuous values on its dimensions. A simple connectionist model provides a logical approach to simulating the phenomenon.


Figure 1: The angles of all possible trajectories from each point in the acoustic space following the grammar for Experiment 1 ( ABCD ).

For our experiments, we adapted stimuli from Holt and Lotto's studies of auditory categorization (2006). Stimuli were frequency-modulated tones uniformly distributed over a two-dimensional acoustic space that can be visualized as a grid with carrier frequency on the $y$-axis and modulation
frequency on the x -axis. Each stimulus was assigned a category based on its location in this space by dividing the grid into four quadrants, labeled A, B, C, and D (see Figure $1)$. During the familiarization phase of the experiments, stimulus tones were presented as a stream of bigrams organized by the four experimenter-defined categories (e.g. a tone from quadrant A was always followed by a tone from B). In this way, the sequences can be described alternatively as a sequence of category labels, or as a sequence of trajectories through similarity space, leading to different predictions about how participants should process the test stimuli. For example, some test stimuli violate predictions based on a sequence of category labels, but are broadly consistent with the direction of change in similarity space.

A neural network simulation provided a qualitative fit to the results from two experiments with different sequences. In fact, the model fit a difference between the two experiments that we did not predict when designing the stimuli.

## Experiment 1: ABCD

Experiment 1 was motivated by a desire to replicate, with more power, an earlier study on the same topic (Wang \& Zevin, submitted). Twice as many subjects were recruited and an extraneous test condition was excluded for the new version of the experiment. The experiment was designed to test the different predictions made by the two accounts discussed in the introduction: the categorization-based approach and the similarity space approach. Specifically, two different types of non-words were created: one for which the items violate the grammar but whose transition trajectory was similar to other transition trajectories in the training (Correct Trajectory Non-Word), and one for which the items never occurred in the training and the transitional trajectory was very dissimilar to other transitional trajectories in the training (Incorrect Trajectory Non-Word). If participants relied on identifying the incoming units as categories $\mathrm{A}, \mathrm{B}, \mathrm{C}$ or D , they would treat words better than non-words and treat both types of non-words as equally unfamiliar. If participants made use of the transition trajectories, Correct Trajectory Non-Words should not be as good as Words but Incorrect Trajectory Non-Words should be much worse than both Words and Correct Trajectory Non-Words.

## Methods

Participants: 78 undergraduate students from The University of Southern California were recruited from the Psychology Department subject pool. They received either course credit or a payment of $\$ 5$ for their participation. Due to technical errors, data was only collected for 72 of the 78 who participated.

Stimuli: The stimuli were frequency-modulated tones adapted from the studies of Holt and Lotto (2006). 64 tones were uniformly distributed over a two-dimensional acoustic space in perceptually equivalent steps $(30 \mathrm{~Hz}$ in carrier
frequency, 18 Hz in modulation frequency). The stimuli were divided into four even quadrants each containing 16 tones and labeled A, B, C, and D. Each stimulus comprised 300 ms of sound and 300 ms of silence.

Familiarization Phase: The entire experiment was controlled using Paradigm (Perception Research Systems, 2007) on a Windows desktop computer. Participants were allowed to read the material of their choice while passively listening through headphones to 10.5 minutes of a sound stream. The sound stream consisted of a total of 512 AB words and 512 CD words, such that all possible A-B transitions and C-D transitions were presented twice. Only half of all possible part-word transitions (from B to A or C and from D to A or C) occurred. The stimuli were chosen using a recursive algorithm to ensure even sampling from the distribution. Consequently, the transitional probability of a tone from B following one from A is 1 , while the probability of a tone from A following one from $B$ is 0.5 .

Testing Phase: Immediately following the training phase, participants were instructed to make a series of familiarity judgments on 36 pairs of tones. During each trial, participants clicked anywhere on the screen to begin and a consecutive sequence of two tones was played. Following presentation of the sequence, participants were asked to indicate their familiarity with the pair of tones. A text prompt was displayed ("Do you think that you heard this sequence in the previous section?") and participants responded by clicking on one of five ratings ("Definitely", "Maybe", "Not Sure", "Maybe Not", "Definitely Not"), ending the trial. There were a total of 36 trials, 12 of each from 3 test conditions: Word, Correct Trajectory Non-Word, and Incorrect Trajectory Non-Word. Each test category had 4 unique test items that were repeated 3 times each, for a total of 12 trials per condition. To maintain consistency across conditions, all test items were novel (i.e. none of the bigrams were present in the familiarization sequence) and followed trajectories with a length of 3 arbitrary units from the first to second tone in the bigram. The Word condition contained two AB and two CD pairs, where bigrams that started in quadrant A followed the median angle for 3 units from the starting stimulus, terminating in quadrant $B$. In the Correct-Trajectory Non-Word condition, each pair of stimuli began and ended in the same quadrant (e.g. AA or BB) but followed a trajectory along the median angle established during the training phase (in general, towards the center of the acoustic space). The Incorrect Trajectory Non-Word condition contained the same pairs of sounds from the Correct Trajectory condition, but reversed the order in which they were played such that they followed the opposite, more unfamiliar trajectory (i.e. outwards from the center of the acoustic space). To reiterate, although the distance in feature space between each tone of a bigram remained at a constant 3 units, only items in the Word condition crossed a quadrant boundary.


Figure 2: Visualization of the 4 test items from each of the 3 test conditions for Experiment 1 (ABCD).

## Results and Discussion

Inferential tests for both experiments are based on linear mixed effects models created in Stata (StataCorp, 2013). Words were rated as significantly more familiar than items in both of the non-word conditions: Correct Trajectory ( $\beta=$ $0.27, z=5.01, p<0.05)$ and Incorrect Trajectory $(\beta=0.37$, $\mathrm{z}=6.91, \mathrm{p}<0.05$ ). This result demonstrates that learning has occurred, as participants treated the grammatical bigrams as different and more familiar than the other sequences. The difference between ratings for Correct Trajectory Non-Words and Incorrect Trajectory Non-Words was marginally significant ( $\beta=0.10, z=1.90, p=0.057$ ). Although the increase from Correct to Incorrect Trajectory Non-Words was only marginally significant, it is important to note the overall trend of increasing unfamiliarity across the 3 conditions (see Figure 3) is consistent with data from Wang \& Zevin (submitted).


Figure 3: Ratings by test category for Experiment 1. Each dot in the scatter represents a subject's mean rating on a scale from 1 to 5 (where 1 is most familiar and 5 is most unfamiliar) for that category. The line and shadow indicate the mean rating and $95 \%$ confidence interval for all subjects in that category.

## Experiment 2: ABDC

In Experiment 1 ( ABCD ), words were defined as transitions from a tone in quadrant A to one in B or from a tone in quadrant C to one in D , such that words could always be recognized as going down in carrier frequency. In other words, participants could have used a single dimension to learn the regularities in Experiment 1. However, we wanted
to examine how participants would learn when the grammar was more complicated. So, in Experiment 2 (ABDC), words were defined as transitions from A to B or D to C , making it necessary to use both carrier frequency and modulation frequency to identify grammatical bigrams. This more complicated grammar should be harder for subjects to learn because it requires tracking two dimensions rather than one.

## Methods

Participants: 84 undergraduate students from The University of Southern California were recruited from the Psychology Department subject pool. They received either course credit or a payment of $\$ 5$ for their participation. Due to technical errors, data was only collected for 72 of the 84 who participated.

Stimuli: Stimuli were taken from the same acoustic space as Experiment 1. Each stimulus comprised 300 ms of sound and 300 ms of silence.

Familiarization Phase: In Experiment 2, words were defined as transitions A-B and D-C (rather than C-D as in Experiment 1). The sound stream contained a total of 512 AB words and 512 DC words, such that all possible A-B transitions and D-C transitions were presented twice. The procedure used was identical to Experiment 1, where participants listened to the familiarization stream passively.


Figure 4: The angles of all possible trajectories from each point in the acoustic space following the grammar for Experiment 2 (ABDC).

Testing Phase: The testing procedure for Experiment 2 was consistent with Experiment 1, but used a different set of test items. There were a total of 36 trials, 12 of each from the same 3 test conditions: Word, Correct Trajectory NonWord, and Incorrect Trajectory Non-Word. Each test category had 4 unique test items that were repeated 3 times each, for a total of 12 trials per condition. All test items were novel and followed trajectories with a length of 3 arbitrary units from the first to second tone in the bigram. As before, the Correct-Trajectory Non-Word pairs of stimuli began and ended in the same quadrant (e.g. AA or DD) but followed a trajectory along the median angle established
during the training phase (in general, towards the center of the acoustic space). The Incorrect Trajectory Non-Word condition contained the same pairs of sounds from the Correct Trajectory condition, but reversed the order in which they were played such that they follow the opposite, more unfamiliar trajectory (i.e. outwards from the center of the acoustic space).


Figure 5: Visualization of the 4 test items from each of the 3 test conditions for Experiment 2 (ABDC).

## Results and Discussion



Figure 6: Ratings by test category for Experiment 2. Each dot in the scatter represents a subject's mean rating on a scale from 1 to 5 (where 1 is most familiar and 5 is most unfamiliar) for that category. The line and shadow indicate the mean rating and $95 \%$ confidence interval for all subjects in that category.

As in the previous experiment, there is robust evidence of statistical learning in Experiment 2. Unlike in Experiment 1, however, there was only a marginally significant difference between average ratings for Words and Correct Trajectory Non-Words ( $\beta=0.11, \mathrm{z}=1.96, \mathrm{p}=0.05$ ). As before, Words were rated as significantly more familiar than Incorrect Trajectory Non-Words ( $\beta=0.30, \mathrm{z}=5.62, \mathrm{p}<0.05$ ). Further, Correct Trajectory Non-Words were rated as significantly more familiar than Incorrect Trajectory NonWords $(\beta=0.20, \mathrm{z}=3.66, \mathrm{p}<0.05)$, which indicates sensitivity to the direction of change.

Thus, results from both Experiment 1 (ABCD) and Experiment 2 (ABDC) follow the same general trend: words were rated as most familiar, followed by Correct Trajectory Non-Words, with Incorrect Trajectory Non-Words rated as most unfamiliar, although particular pairwise contrasts
differ in significance across experiments. Curiously, and contrary to our initial predictions, the difference between ratings for Words and Correct Trajectory Non-Words is smaller in Experiment 1 than in Experiment 2, $(\beta=-0.17$, z $=-2.20, \mathrm{p}<0.05$ ).

## Computational Modeling

## Design and Procedure

In order to simulate learning in our experiments, we developed a simple feed-forward back-propagation, neural network using PDPTool (McClelland 1986; 2015). The neural network used a logistic activation function and had two input units, two output units, a two-unit hidden layer and a bias. Two versions of the model were trained ten times each: ABCD and ABDC , which were identically constructed but received different inputs corresponding to the 1024 stimulus sequences from Experiment 1 and 2, respectively. Each stimulus was coded as a pair of coordinates representing its location in the acoustic space. Inputs and outputs were scaled to fit within the valid range of input $[-1,1]$ and output $[0,1]$ values. The model was trained to predict the next stimulus from the current stimulus as bigram pairs, including both Words and PartWords, so for example the sequence ABCD would be presented to the model in three discrete trails: $\mathrm{AB}, \mathrm{BC}$, and CD. As an initial measure of learning, we trained multiple runs for 100 epochs, collecting pattern sum of squares (pss) on the training items after each of the first ten epochs, and every fifth epoch thereafter. We observed that the model reached asymptote by this measure after ten epochs, to an error of 0.12 for $A B C D$ and 0.14 for $A B D C$.

Error scores for the test items were generated by presenting the first stimulus in each test pair to the model and calculating the summed squared error (pss) for the model's output relative to the second item in the test bigram. This measure was taken for all 12 test items every 5 epochs from the 10th to the 50th, and the mean of these observations taken for each run. Means of all ten runs and standard deviations across runs are reported in Tables 1 and 2.

## Results and Discussion



Figure 7: Average error by test category over 10 runs of the model for ABCD (right) and ABDC (left).

The computational models for ABCD and ABDC qualitatively replicated the human data. Figure 7 displays each model's error by test condition, measured as the squared distance between the model's prediction and the target (i.e. the second point in the test item). Thus, the higher the error for a test item, the further away in feature space the second tone in the pair was from the model's prediction. As such, the model's error on each test item parallels the measures of familiarity collected from the human data. As in the human data, both ABCD and ABDC display the overall increasing trend across the three test conditions, with the Incorrect Trajectory Non-Words treated as significantly different than the Words. Furthermore, there is a difference in the Correct Trajectory Non-Word condition between ABCD and ABDC . The ratio of the errors between the categories Word and Correct Trajectory Non-Word is larger in ABCD (0.0955) than in ABDC (0.0387), and it is clear from Figure 7 that this increase is larger for ABCD than ABDC . This difference qualitatively mimics the observed discrepancy between ratings for Correct Trajectory Non-Words in Experiments 1 and 2.

Table 1: Average error and standard deviation by test category over 10 runs of the simulation for ABCD .

| Condition | Avg. Error | Std. Deviation |
| :---: | :---: | :---: |
| Word | 0.0229 | 0.0094 |
| Correct Tr. NW | 0.2399 | 0.0165 |
| Incorrect Tr. NW | 0.5484 | 0.0333 |

Table 2: Average error and standard deviation by test category over 10 runs of the simulation for ABDC .

| Condition | Avg. Error | Std. Deviation |
| :---: | :---: | :---: |
| Word | 0.0052 | 0.0007 |
| Correct Tr. NW | 0.1342 | 0.0051 |
| Incorrect Tr. NW | 0.5060 | 0.0039 |

## General Discussion

The behavioral results from the two experiments presented here are broadly consistent with our conceptualization of statistical learning as occurring by situating stimuli in a perceptual similarity space. Further, the computational model we designed according to this conceptualization fits the data quite well. The auditory stimuli were specifically designed to be difficult to categorize, yet participants were able to distinguish between words and non-words after brief, passive familiarization with a sequence of grammatical bigrams. Although results for the Correct Trajectory Non-Word condition differed between Experiments 1 and 2, the overall trend of increasing unfamiliarity across conditions indicates that learners are sensitive to the trajectory from one stimulus to the next in feature space.

Using the same stimuli - indeed, the same ABCD familiarization sequence used in the current Experiment 1 -

Wang and Zevin (submitted) observed a small difference between Words and Correct Trajectory Non-Words, and a much larger difference between Correct and Incorrect Trajectory Non-Words. Across a number of experiments we are not reporting here due to space limitations, the general pattern of decreasing familiarity from Words to Correct to Incorrect Trajectory Non-Words is always present, although different contrasts are significant by inferential tests under different conditions. We therefore suggest that this overall pattern is the most critical feature of the data to simulate.

Interestingly, there are more subtle differences between Experiments 1 and 2 that are also captured by the simulation. Both the model and the human participants treated Correct Trajectory Non-Words as more similar to Incorrect Trajectory Non-Words in Experiment 1, but more similar to Words in Experiment 2. Until examining the simulation results, we failed to consider an idiosyncrasy with how the test items were chosen between experiments. The test items for ABCD and ABDC differed slightly in how they were sampled from throughout the feature space. As shown in Figures 2 and 5 above, the four non-word pairs for ABCD were taken from each of the four quadrants while in ABDC the four non-word items were drawn from only two quadrants (two from A and two from D). Therefore, half of the Correct Trajectory Non-Words in ABCD followed the correct trajectories for words and the other half for part-words while in ABDC they all followed trajectories for words. This could explain why the Correct Trajectory items were rated as more unfamiliar in ABCD than in ABDC for both the human experiments and the computational models.

Interestingly, the simulation's overall error, especially for Words, is lower in ABDC than ABCD . One possible explanation is that having two meaningful dimensions to define words provides the model with more information over which it can track probabilities, increasing its ability to learn the grammar. In contrast, the extra dimension introduces additional complexity that makes the sequences more difficult for humans to learn. This gets at one of the problems with the model: it is almost too good at learning the pattern. While humans must approximate each stimulus's location in similarity space, the model receives exact coordinates so naturally the model will produce more accurate and precise predictions. A further problem with the simulation lies in the fact that connectionist models like the SRN (Elman, 1990) and the one presented here all learn with supervision. While the model receives feedback on its predictions for every stimulus, human learners are thought to be dependent on unsupervised mechanisms under similar conditions (McClelland, 2006).

Furthermore, because the model was designed for a very specific experimental setting, it has limited applications. We have proposed elsewhere (Wang \& Zevin, submitted) that the trajectory-tracking approach may provide an explanation for statistical learning phenomena hitherto unaccounted for by existing models. For example, word segmentation during initial language acquisition is a real-life situation in which category labels are not readily available and the sequence
signal may be ambiguous due to natural variation in human speech (Shannon, 1948; Hockett, 1960).

However, there is no reason to believe that the trajectorytracking model tells the whole story. It is more likely that learners utilize different mechanisms, either simultaneously or individually, depending on the situation and the information that is readily available in the stimuli sequence. Relying on perceptual similarities is useful when stimuli are defined on the same dimensions and low-level physical features are readily extracted. When it is easy to abstract and divide stimuli into categories, however, there may be situations in which stimuli are readily recognizable, and it is simpler (i.e. involves lower computational load) to compute transitional probabilities over labels.

In conclusion, the results of this series of experiments and their remarkably close fit to the simulations provide overwhelming support for our theory that learning occurs by tracking changes in perceptual features from one stimulus to the next in a sequence. Although we observed a difference in one of the test conditions between the two experiments, the simulations reproduced the phenomenon, leading us to believe that it was a result of an idiosyncrasy in our test stimuli. Results from both experiments were otherwise consistent with our assertion that participants are situating stimuli within a perceptual similarity space and learn the pattern by tracking their trajectories through this space.

## Acknowledgments

We would like to thank Guo Dong for help with the recursive algorithm for generating balanced stimulus sequences, and Jay McClelland for suggestions on coding the two-dimensional space for the simulations.

## References

Aslin, R. N., Saffran, J. R., \& Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. Psychological science, 9(4), 321-324.
Elman, J. L. (1990). Finding structure in time. Cognitive science, 14(2), 179-211.
Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. Machine learning, 7(2-3), 195-225.
Emberson, L. L., Liu, R., \& Zevin, J. D. (2013). Is statistical learning constrained by lower level perceptual organization?. Cognition, 128(1), 82-102.
Fiser, J., \& Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(3), 458.
Gervain J, Macagno F, Cogoi S, Pena M, Mehler J. The neonate brain detects speech structure. Proc Natl Acad Sci U S A 2008, 105:14222-14227.
Gómez, R. L. (2002). Variability and detection of invariant structure. Psychological Science, 13(5), 431-436.
Hockett, C. D. (1960). The origin of speech. Scientific American, 203, 88-111.

Holt, L. L. \& Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. Journal of the Acoustical Society of America, 119, 3059-3071.
Kleinschmidt, D. F., \& Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. Psychological Review, 122(2), 148.
McClelland, J. L., Rumelhart, D. E., and the PDP Research Group (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models. MIT Press, Cambridge, MA.
McClelland, J. L. (2006). How far can you go with Hebbian learning, and when does it lead you astray. Processes of Change in Brain and Cognitive Development: Attention and Performance XXI, 21, 33-69.
McClelland, J.L. (2015). PDPTool [Computer Software]. Retrieved from http://web.stanford.edu/group/pdplab/pdphandbook
Perception Research Systems. 2007. Paradigm Stimulus Presentation. Retrieved from http://www.paradigmexperiments.com
Perruchet P, Vinter A. (1998). PARSER: a model for word segmentation. J Mem Lang, 39:246-263.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926-1928.
Saffran, J. R., Johnson, E. K., Aslin, R. N., \& Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70(1), 27-52.
Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379-423.
Shepard, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. Stimulus Generalization, 94-110.
StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP. Retrieved from http://www.stata.com/
Wang, F. H., \& Zevin, J.D. (submitted) Statistical learning of unfamiliar sounds as trajectories through a perceptual similarity space.

# The role of prior knowledge and expertise on choice of referring expression 

Alyssa Ibarra (aibarra@ur.rochester.edu)<br>Departments of Brain \& Cognitive Sciences and Linguistics, 363 Meliora Hall, Box 270268<br>Rochester, NY 14627<br>Jeffrey Runner (jeffrey.runner@rochester.edu)<br>Departments of Linguistics and Brain \& Cognitive Sciences, Box 270096<br>Rochester, NY 14627<br>Michael K. Tanenhaus (mtanenha@ur.rochester.edu)<br>Departments of Brain \& Cognitive Sciences and Linguistics, 363 Meliora Hall, Box 270268<br>Rochester, NY 14627


#### Abstract

Referential success depends on choice of referring expression. The choice of referring expression will depend on contextual factors as well as factors related to speaker and addressee knowledge. A shared-learning paradigm was used in which partners learned names of objects together and separately before a referential task. Items differed on commonality, with some independently rated as more common and some as more rare. Speakers were less likely to use names versus other forms when items were rare than common ( $\mathrm{p}<0.001$ ) and less likely to use names when items were new than learned together ( $\mathrm{p}<0.001$ ). Asymmetry effects showed that speakers were more likely to use a name when the addressee was deemed more knowledgeable in post-test ratings ( $\mathrm{p}<0.01$ ). Together, we take this to show speakers choose to use a name versus a description based on the likelihood that their interlocutor will know the name. Factors affecting the likelihood include prior knowledge of what a typical addressee will know and shared experience, which includes inferring an interlocutor's expertise, as dynamically updated during a dialog.


Keywords: interactive conversation; referring expressions; common ground; expertise, belief updating

## Introduction

In interactive conversation, the likelihood of a speaker's referential success depends on choice of referring expression. A speaker's choice to refer to a picture as, say, a "dog" versus a "Bernese Mountain dog" or "the large black dog with a white chest and tan marking", depends on factors relevant to the context, the speaker, and the addressee. If, for example, the referential domain includes multiple dogs, a more specific label will be needed to pick out a unique object; whereas a domain with a single dog and several cats is likely to elicit the basic label "dog". A speaker's knowledge, or lack of knowledge, of dog breeds will restrict the name alternatives the speaker will have readily available, as will the speaker's assessment of the addressee's knowledge-e.g., if the speaker is aware that the addressee does not know a breed of dog, the speaker may choose to describe the dog rather than use the name of the breed.

Speakers readily distinguish between differences in knowledge when they learn novel names for novel objects with a partner (Wu \& Keysar, 2007; Heller et al. 2012;

Gorman et al., 2013). In this shared-learning paradigm, participants learn some novel names together (shared names) and then one participant, who is subsequently the director in a referential task, learns additional names alone (privileged names). This creates a situation of knowledge asymmetry that the speaker may use to inform choice of referring form in the referential task. Speakers indeed track shared experience when the objects are novel, i.e., when they have not seen the objects prior to the experiment. They use more names than descriptions for those objects that have been learned together. (Wu \& Keysar, 2007); and rarely use the name-alone form for privileged names (Heller et al. 2012; Gorman et al., 2013).

In related work, Gegg-Harrison (2016, also see GeggHarrison \& Tanenhaus, 2016) embedded name learning in the context of a toy world. In a role-playing game, certain levels were always encountered before others. The participant's choices made regions of the world and the information contained there inaccessible. Therefore, a participant who displayed knowledge of a particular name would implicate that she would know some names but not others. The participant then interacted with a game expert in several tasks that involved characters from the toy world. Interactions with the expert showed that the participant modified her name use and assessment of what the speaker knew based on the expert's use of names.
In most conversations, especially with a relatively unfamiliar addressee, a speaker will not have direct, shared experience. If we assume that speakers choose to use a name because it is the least resource demanding, shortest, and richest referring expression, then a rational speaker would take into account the likelihood that the addressee would know the name. That likelihood would be based on both the likelihood that any addressee would know that name (baseline likelihood) and evidence specific to that addressee, much of which is gleaned from the ongoing conversation.

To lay the groundwork for explicitly evaluating the likelihood hypothesis, we modify the shared-learning paradigm by using pictures of real entities that vary in baseline likelihood and by having the learning be interactive, which allows the director to have assessed the expertise of the matcher. We hypothesize that name use will be affected
by both baseline likelihood and shared experience, including inferred level of expertise, with larger effects of shared learning for less commonly known (rare) names and lower name usage for matchers who are judged to have lower expertise, especially for rare names, even when they have been learned together. If we can establish that the paradigm is sensitive to expertise, then this allows for more targeted questions about the factors, including prosody and choice of lexical expressions, that interlocutors use to signal and infer which names are likely to be know to each other.

Using novel names and novel objects is well-suited for asking basic questions about whether speakers can form item-specific memories of shared experience that can be used as a basis for common ground. However, limitation of using novel objects and novel names is that it abstracts away from two important characteristics that influence choice of referential forms. The first is the speakers' prior beliefs about how likely any interlocutor is to know a name. For example, any speaker of English can assume that her interlocutor will know the general category of dogs and its base-level name. The second is that it doesn't capture the dynamic aspect of interactive conversation. A speaker is unlikely to know the full extent of her addressee's knowledge about a topic prior to an interaction. Rather she may draw inferences based on what her interlocutor reveals during the interaction. For example, if a speaker learns that his interlocutor is a gourmet cook, he can assume that she will know the names for even relatively rare kitchen utensils. This assumption is possible even without direct evidence of knowledge of particular names by attributing to that interlocutor knowledge that is likely known by most gourmet cooks.

The present study extends the shared-learning paradigm in two important ways. First, we use real objects drawn from categories, in particular dog breeds and kitchen utensils, in which there are commonly known names (e.g., "tongs") and less common (rare) names (e.g., "mandoline"). We normed items as common or rare to the average person but chose categories that might differ in expertise given broad designations of communities (i.e., cooks and dog lovers). Second we modify the shared-learning paradigm to create conditions where participants have the opportunity to assess each other's expertise in a domain. This will allow one to tease apart whether assessment of addressee knowledge is acquired throughout the interaction by cues separately from the shared experience of learning names together.

Specifically, we ask whether the evidence that shared learning of novel objects informs referential choice can be interpreted as a part of a larger likelihood computation that also incorporates common knowledge, shared experience and inferred expertise. It could be the case that shared experience effects seen with novel objects can be attributed to the triggering of episodic traces, such that when a speaker chooses to refer to an object, the memory of having learned it with a specific addressee is activated, which in turn informs referential choice (Horton \& Gerrig, 2005). The low-level trace is enough to explain the shared learning effects. A speaker, however, may come into the experiment with a prior
belief about a partner's likely knowledge that gets updated as more evidence is presented. Thus, when the partners learn novel names together, the speaker updates the belief that the addressee knows the name, having learned it together. In the paradigm with novel objects, this belief is likely binary: they either learn it together in the experimental context or not. The present study, however, in its incorporation of real-world objects, creates a situation in which interlocutors not only may have different prior beliefs about the addressee's knowledge at the beginning of the experiment but may also dynamically update those beliefs throughout the interaction. The assessment of a partner's knowledge via shared experience as well as its updating, of course, also involves memory processes. However, it is not clear how memory traces with a specific partner could account for how beliefs of the overall commonality of an object interact with that shared experience. Furthermore, as the interaction unfolds, it is not clear whether further updating of this belief occurs given evidence from the interaction. An episodic trace would not predict its effect on referential choice for new items, for example. If episodic traces are driving the effects of shared learning, then speakers will be equally sensitive to the shared experience of learning names in the context of the experiment. However, if commonality and inferred expertise combine with shared experience to determine the overall likelihood of an addressee knowing a name, then one cannot only appeal to episodic-based explanations.

The present study concerns itself with the following questions: 1) Does the commonality of an object impact the effect of shared learning with a partner? 2) Do interlocutors dynamically update their beliefs about their partner's expertise in a subject? If so, does this expertise interact with commonality and shared learning to inform the speaker's choice of referring form?

We predict that if items are more likely to be known prior to the experiment, shared experience in the experiment is less likely to affect choice of referring expression. Conversely, if rare items are similar to novel items in past studies (i.e., associated with low or no prior knowledge), then the shared experience effects are likely to be strongest in this group. Furthermore, if expertise is inferred throughout the interaction, then the effects of commonality and/or shared experience, if any, should differ according to whether the addressee is deemed to be knowledgeable in the domain or not. Finally, although the effects of shared experience will be smaller for common names, we should still see some effects; this would suggest that shared experience enters into likelihood calculations, even for commonly known names.

## Methods

The experiment consisted of three parts: a learning phase, in which the participants learned the name of items in two categories--dog breeds \& kitchen utensils; a test phase, in which one participant directs the other to pick out a target item in a referential task; and a post-test, in which participants rate their partner's, and their own, knowledge of
the items as well as identify the context of learning. See Figure 1 for an illustration of these three parts.


Figure 1: Experimental procedure consisting of the Learning phase (i.e., Training), the Test phase, and Post-test phase.

## Norming

The items used in this experiment were normed and rated as either common or rare from two categories on which people often differ in expertise (dog breeds and kitchen utensils). See Figure 2 for sample stimuli of kitchen utensils. The norming procedure consisted of a presentation of 8 images ( 4 dog breeds, 4 kitchen items, with 2 common and 2 rare of each category). Each image was presented one at a time and participants were tasked to 1) label the image, 2) provide a confidence rating for that label, and 3) indicate how likely it is that the average person would know the label.

There were 80 unique images normed, separated into 10 lists of 8 items. 600 total participants were tested (roughly 60 people per list) with an average of 58.8 data points per item, due to some blank responses for single items.

Of the 80 unique items normed, we chose 36 experimental items: 12 for shared learning; 12 for director-alone learning; and 12 new items to be tested but not trained on. Half of all items $(\mathrm{n}=18)$ were dogs and the other half $(\mathrm{n}=18)$ were kitchen items. Within each of these categories, half ( $\mathrm{n}=9$ ) were rare, and half ( $n=9$ ) were common.

We chose the items by first sorting by largest sample. For common items, we then took the items with the highest average rating (i.e., the knowledge rating for the "average person") with the highest accuracy of labeling (above 70\% accuracy). For rare items, we took the items with the lowest average rating (i.e., the knowledge rating for the "average person") with the highest accuracy (above $10 \%$ accuracy). A minimum of $10 \%$ accuracy was implemented in order to remove items that had a name that was incorrect but confidently rated (e.g., "Greyhound" was a highly repeated label for an Azawakh, resulting in 0\% accuracy).


Figure 2: Sample stimuli of rare and common kitchen utensils.

## The Learning Phase

The learning phase allowed participants to observe each other's expertise. We manipulated whether names were shared or privileged by dividing the phase into shared learning among the two partners and director-alone learning. This allowed us to observe how a speaker combines shared experience with general likelihood of knowing a name and modulates choice of referring expression by perceived differences in expertise.

At the beginning of the learning phase, participants sit together at a table with the experimenter standing in front of them. Based on seating arrangement, each participant is assigned as either the Director or the Matcher.

For the shared-learning portion of the learning phase, the experimenter explains to the participants that they will be learning names of items together and that the images will be of different dogs and different kitchen items which they may or may not be familiar with. The experimenter then presents a flashcard with an individual image one at a time. After presenting the image, participants discuss with one another whether they know the name of the image. If they know it, they say the name aloud. After given some time to guess, the experimenter states the correct name of the item. This procedure is done for 12 items, presented in three blocks of four items. After each block, the items are repeated once more before moving to the next block. After all three blocks, the participants go through the whole stack.

For the second half of the learning phase (Director-alone learning), the Matcher sits at a computer in the same room and is instructed to wear headphones playing instrumental music while engaging in a game of Solitaire. During this time, the experimenter presents the Director with 12 additional images. The procedure is as above, with three blocks of four items. Thus, together with the shared learning portion, a total of 24 items were presented in the learning phase, holding out the last 12 items for the test phase.

Both portions of the learning phase (shared and Directoralone) are recorded and transcribed.

## The Test Phase

Following the learning phase, the participants began the test phase, which was a referential task in which the Director verbally leads the Matcher to pick out a target item from an array. The test phase created a situation for the Director to refer to the items of varying commonality and ground status that were introduced in the learning phase in a controlled referential task.

Both Matcher and Director sat at their own computer facing one another. The Director was shown one image on the screen, which could be an image that was learned together with the Matcher, learned alone, or never presented during the learning phase. The Matcher, on the other hand, was presented with three images. All three images were of the same category (i.e, all dogs or all kitchen items), and they are of same commonality (i.e., all rare or all common). The difference among the images is in their ground status (i.e., one is shared, one is learned by Director alone, and one is
new to both participants). The Director was instructed to identify the image on her screen with explicit instruction that there is no restriction on language (i.e., she can name or describe as needed), and the Matcher was tasked to click on the target image. The Matcher was also allowed to ask questions and interact freely with the Director as needed.

This procedure was followed for 24 test items: 8 Shared, 8 Alone, and 8 New targets. Half the targets in each ground status ( $n=4$ ) are dog breeds; half ( $n=4$ ) are kitchen items. Furthermore, half the dog items $(\mathrm{n}=2)$ are rare; the other half ( $n=2$ ) are common; half the kitchen items ( $n=2$ ) are rare; the other half ( $\mathrm{n}=2$ ) are common.

Each trial was recorded and the form of referring expression was coded as either Name Alone, Description, Name + Description, or Description + Name.

The data below are from 24 pairs of participants.

## The Post-test

After the test phase, both participants individually took a post-test. Participants were asked to rate their partner's knowledge of each domain (dog breeds/kitchen utensils), as well as rate their own knowledge in the domain. Next, participants were shown a single item and were asked to type in the label for the item, if known, and rate their confidence in their label. They were also asked to rate their confidence that their partner would know the label for that object. This procedure was repeated for all test items as well as new items. These measures of confidence were implemented to measure general expertise as well as presumed relative expertise between the two participants (i.e., whether the Director was more knowledgeable than the Matcher, or vice versa) in order to assess knowledge asymmetry.

Lastly, they were asked to identify the context of learning the item: learned with partner, alone (applicable for director only), learned prior to the experiment, or never learned. The items tested varied in context of learning; they could be an item from shared learning, Director-alone learning, or New items. This measure was used to observe whether participants were tracking context of learning by specific item.

## Results

We focus on speakers' Name-Alone use. We fit a GLM model predicting Name use against a single category combining other forms. Fixed effects were Commonality (Rare/Common), Ground Status (Privileged/Shared/New), and Knowledge Asymmetry (More/Less knowledgeable Matcher, as determined by post-test ratings) with Pair as a random effect (Table 1). Below we address our particular predictions.

| Name $\sim$ Asymmetry + Ground <br> * Commonality + (1\|Pair) | $\beta$ | Std <br> Error | P-value |
| :--- | :--- | :--- | :--- |
| AsymmetryM+ | 0.73 | 0.29 | $<0.01$ |
| AsymmetrySAME | -0.22 | 0.33 | 0.49 |
| CommonalityRare | -3.36 | 0.43 | $<0.001$ |
| GroundShared | 1.13 | 0.69 | 0.10 |
| GroundNew | -1.53 | 0.42 | $<0.001$ |
| GroundPriv:CommonalityRare | -0.47 | 0.76 | 0.53 |
| GroundNew:CommonalityRare | 0.67 | 0.56 | 0.23 |

Table 1: GLM model ouput

## Does the commonality of an object impact the effect of shared learning with a partner?

A main effect of Commonality demonstrates less name use for Rare than Common names ( $p<0.001$ ). In regard to Ground status, speakers are less likely to use a Name when the item is New than when Shared ( $\mathrm{p}<0.001$ ), but for learned names, the main effect of Ground was not significant.

Figure 3 shows Name use by the speaker across Commonality and Ground. The main effect of commonality can be seen by the larger proportion of name use in the right three columns of the graph (Fig 3). Although there was no effect of Ground Status across learned items, when looking at Rare items in comparison to Common items, one can see that differences in name use across Ground status conditions are more apparent for Rare than Common objects, similar to results in past work with novel objects. Indeed, in separate models, effects of Ground were significant for Rare items, such that names are used less for privileged items than for shared items. Common objects show the same pattern but to a much lesser extent. Thus, the speaker is less likely to use names when referring to rare objects than common ones and this preference is further reduced if the rare objects are privileged or new.


Figure 3: Overall use of names by the Director across Ground Status (Shared, Privileged, New) and Commonality (Common, Rare) conditions.

Does inferred expertise inform the speaker's choice of referring form?

A main effect of Asymmetry showed that speakers were more likely to use a Name when the Director presumes the Matcher to be more knowledgeable than herself ( $\mathrm{p}<0.01$ )

Figure 4 shows Expertise effects across Ground Status and Commonality conditions. The main effect is shown by an overall larger proportion of name use when the Matcher is deemed more knowledgeable than the speaker $(\mathrm{M}+$ ) than when the Matcher is deemed less knowledgeable (M-). These effects are carried most strongly in two conditions: for Rare objects that are learned together (Shared), and for Common objects that are New (not learned at all in experiment).


Figure 4: Proportion of Names when Matcher deemed more $(\mathrm{M}+$ ) or less (M-) knowledgeable than the Director.

Data gathered from the post-test, in which participants had to identify the context in which a particular item was learned (e.g., alone, shared, or neither) revealed high accuracy for tracking ground status (Table 2). This measure was recorded to assess memory of ground status across the course of the experiment. Table 2 shows highest accuracy of context identification for Rare Shared items and lowest accuracy for Common Shared, but this is not significantly different across the categories.

| Context Identification Accuracy (Post-Test) |  |  |
| :---: | :---: | ---: |
| Common | Alone | $86.1 \%$ |
|  | Shared | $85.4 \%$ |
| Rare | Alone | $86.1 \%$ |
|  | Shared | $91.7 \%$ |

Table 2: Post-test accuracy

## Discussion

Speakers' choice of names is strongly affected by the prior likelihood that an interlocutor will know a name. When items are more common, shared learning has weaker effects on
name use than when an object is rare. As in previous studies with novel objects, common ground effects are more apparent for less commonly known items. This is taken to show that a speaker's reliance on shared learning as a means to assess partner knowledge is reduced when the objects are likely to be commonly known. This provides evidence that the basis of common ground in an interaction relies on assessments of prior knowledge as well as shared experience. The strongest version of the memory-based account would expect ground effects regardless of commonality, as long as partners have shared experience. However, our post-test data show that speakers are highly accurate in identifying the context of learning, suggesting that they are not less sure about whether a name was shared but they are using that information in combination of other information: i.e, prior knowledge.

Furthermore, we see general expertise effects, such that name use is increased when the addressee is deemed to be more knowledgeable than the speaker. This provides evidence of dynamic updating of knowledge assessments. Future directions would include a more controlled way of assessing when exactly expertise judgments as this might help in teasing apart whether this is confined to the learning phase, or whether participants indeed continue to update beliefs throughout the test phase. However, even with assessments taken post test, we see expertise effects driven from interaction.

Having established expertise effects using this paradigm, we are currently carrying out follow-up work that uses the paradigm to explore further questions. For instance in an ongoing experiment we ask whether expertise assessments derived from the interaction are more strongly weighted than expertise assessments derived from top-down knowledge (e.g., telling the speaker his partner is an expert). Our current study asks this very question by having a director complete the collaborative task in the test phase without any prior interaction with her partner and given only top-down information about the partner's status as an expert in the domain.

Another avenue of research study explores the signals in the interaction that contribute to the assessments of expertise. Specifically, we are testing normed markers of uncertainty modeled on the types of utterances we observed during the learning phase. A confederate matcher will, in the learning phase, reveal her expertise through use of these uncertainty cues. If naïve directors attribute expertise to their partner as a function of these cues, it will be an important step in further understanding the particular components of an interaction that inform beliefs of partner knowledge.

Overall, the contribution of the current study is in embedding the previous evidence on shared experience into a larger computation of the likelihood of addressee knowledge. The effect of commonality on choice of referring expression may not be surprising on its own but together with evidence of inferences of expertise throughout the interaction, one can get a better understanding of how speakers may be combining these different sources of information to compute
this likelihood. Future studies will examine the factors that shift around the likelihood rather than to only appeal to lowlevel processes (i.e., memory traces) that contribute to the individual factors. We then can evaluate quantitative models to compare likelihood models that combine multiple cues with other classes of models.

Lastly, this study explores expertise in domains given interactions about particular items. We argue that the generalization of presumed expertise to a larger domain given updated beliefs about particular objects is tied to beliefs about groups of people and the presumed knowledge of particular communities. For example, dog kennel owners are likely to be presumed to know a lot of about dog breeds and not so much about cat breeds. However, if the relevant community were veterinarians, a member of that community might be presumed to know about both dog breeds and cat breeds. We argue that assessments about expertise are necessarily tied to the community applied, and this constraint is utilized by speakers to narrow in on the dimensions that are relevant for both generalization to new interlocutors and generalization to new items in a given domain.

## Acknowledgments

We are grateful to the members of the KurTan lab meeting and the Experimental Semantics and Pragmatics (ESP) reading group who provided helpful input towards the design and implementation of this study. We especially want to acknowledge Whitney Gegg-Harrison. Her thesis work first explored many of the questions we are pursuing in this line of research. Whitney also provided valuable advice on the current project. This work was supported by NIH grant HD 27206 to MKT.

## References

Clark, H. H., and Marshall, C. (1981). "Definite reference and mutual knowledge," in Elements of Discourse Understanding, eds A. K. Joshi, B. L. Webber, and I. A. Sag (New York, NY: Cambridge University Press), 1063.

Gegg-Harrison, W. (2014). Knowledge and naming in interactive conversation (Doctoral dissertation, University of Rochester). Retrieved from http://hdl.handle.net/1802/28853.
Gegg-Harrison, W.M. \& Tanenhaus, M.K. (2016). What's in a name? Interlocutors dynamically update expectations about shared names. Frontiers in Psychology.7, Paper 212.
Gorman, K. S., Gegg-Harrison, W. M., Marshall, C. R., \& Tanenhaus, M. K. (2013). What's learned together stays together: speakers' choice of referring expression reflects shared experience. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(3), 843-854.
Heller, D., Gorman, K. S., \& Tanenhaus, M. K. (2012). To name or to describe: shared knowledge affects referential form. Topics in Cognitive Science, 4, 290-305..
Horton, W., \& Gerrig, R. (2005). Conversational common ground and memory processes in language production.

Discourse Processes, 20(1), 1-35.
Wu, S., \& Keysar, B. (2007). The effect of information overlap on communication effectiveness. Cognitive Science, 31, 1-13.

# Beyond Almost-Sure Termination 

Thomas F. Icard (icard@stanford.edu)<br>Department of Philosophy, Stanford University


#### Abstract

The aim of this paper is to argue that models in cognitive science based on probabilistic computation should not be restricted to those procedures that almost surely (with probability 1 ) terminate. There are several reasons to consider nonterminating procedures as candidate components of cognitive models. One theoretical reason is that there is a perfect correspondence between the enumerable semi-measures and all probabilistic programs, as we demonstrate here (generalizing a better-known fact about computable measures and almostsurely halting programs). One practical reason is that the line between almost sure termination and non-termination is elusive, as well as arbitrary. We argue that this matters not only for theorists, but also potentially for a learner faced with the task of inducing programs from experience.


## Introduction

The metaphor of cognition as computation provides a fruitful and flexible foundation for cognitive science. While computation can be understood broadly to encompass many different paradigms and formats (Rescorla, 2015), it is generally presumed that an upper bound on what can be computed by the human mind is that which can be computed by a Turing machine, or a program in any other universal model of computation, such as lambda calculus, recurrent neural networks with rational weights, combinators, Java programs, and so on.

Some early proponents of the computational theory of mind (e.g., Putnam 1967) focused attention on probabilistic computation, allowing randomization in state transitions; and random mechanisms have been central in psychological models going back at least to stimulus-response theory, which had formal connections to probabilistic automata (Suppes, 1969). In recent work, computation with random elements has taken on new significance, where mental representations themselves are characterized in terms of probabilistic procedures or programs (Goodman et al., 2014), and noise is seen not just as a nuisance, but as deeply tied to an agent's capacity for prediction and induction. Although probabilistic machines cannot compute any more functions than deterministic machines, this shift in emphasis raises new and distinct questions. For instance, how expressive is a given class of probabilistic machines for representing useful distributions?

Much of the recent theory of probabilistic computationparticularly that motivated by application to cognition-has focused on computable probability distributions, specifically restricting to procedures that terminate almost surely (a.s.), that is, with probability 1 . This has given rise to a rich body of work. For instance, it can be shown that the computable distributions correspond to the a.s.-terminating probabilistic Turing machines (see, e.g., Freer et al. 2012), as well as to the a.s.-terminating stochastic lambda terms (Dal Lago and Zorzi, 2012). The limits of computability in the context of
conditioning continuous distributions have also been thoroughly investigated (Ackerman et al., 2011).

These important advances notwithstanding, the aim of the present paper is to argue that in cognitive science the focus on a.s.-termination is overly restrictive. As a foundation for theorizing about cognitive processes we should consider the class of all probabilistic computations, not just those that a.s. halt. To use terminology introduced more formally below, cognition should be modeled on the more general class of enumerable semi-measures, rather than the smaller class of computable probability measures. We offer two arguments for this claim, one practical and one theoretical.

The practical argument is that the line between a.s. termination and non-termination is elusive and arbitrary. This point is illustrated with a simple example, where the boundary can be studied concretely. The theoretical argument is that the correspondence between the semi-measures definable by probabilistic machines and enumerable semi-measures is more basic and canonical than that between measures definable by a.s.-terminating machines and computable measures. We give a simple, self-contained proof of this first correspondence (Theorem 1), which subsumes the second as a special case (Corollary 1). This proof is elementary, and is arguably simpler than direct proofs of the corollary. We also discuss some connections to program induction, and possible repercussions for probabilistic inference.

## Background on Probabilistic Computation

Consider any universal language for describing computations. Allowing programs in one of these languages access to an unlimited source of iid samples from a $\operatorname{Bernoulli}(0.5)$ distribution brings us to the setting of universal probabilistic languages. For instance, a Turing machine might have an additional read tape with an infinite sequence of random bits, while a lambda term might make use of a choice operator $\oplus$, where $M \oplus N$ reduces to $M$ or $N$, each with probability 0.5 . Just as the Church-Turing Thesis states that any two reasonable deterministic models of computation will be equivalent, one might hypothesize that any reasonable way of adding fair coin flips to these models will give rise to an equivalent model of probabilistic computation. For the rest of this paper we will remain neutral about which of these versions we adopt.

Non-termination, even for very simple, e.g., monitoring processes, is of course a desirable feature of many mechanisms involved in control, where inputs are continually processed (Botvinick and Cohen, 2014). However, our interest here is non-termination even for stand-alone programs without input, so we restrict attention to this setting.

A probabilistic program $\pi$, in any machine language, generates an output $w$-let us suppose outputs are (or at least
encode) binary sequences, so that $w \in\{0,1\}^{*}$-with some probability, which we will write $\mu_{\pi}(w)$. That is, $\mu_{\pi}(w)$ is the sum of the probabilities of all the execution sequences that halt with output $w$. The program $\pi$ implicitly represents a distribution on binary strings; however, this distribution may not be a proper probability distribution on $\{0,1\}^{*}$, as it may be that $\sum_{w} \mu_{\pi}(w)<1$. This will happen if the program fails to halt with some probability $1-\sum_{w} \mu_{\pi}(w)>0$. A function $\mu$ for which $\sum_{w} \mu(w) \leq 1$ is called a (discrete) semi-measure, and it is called a probability measure if this holds with equality.

A semi-measure $\mu$ is (computably) enumerable if it is approximable from below; that is, if for each $w \in\{0,1\}^{*}$ there is a computably enumerable weakly increasing sequence $q_{0}, q_{1}, q_{2}, \ldots$ of rational numbers, such that $\lim _{i \rightarrow \infty} q_{i}=\mu(w)$. Most semi-measures are not enumerable, but for any probabilistic program $\pi$, the semi-measure $\mu_{\pi}$ will be enumerable. To approximate $\mu_{\pi}(w)$ from below, consider the set $W_{i}$ of strings $v$ with length $l(v) \leq i$, such that $\pi$ accesses (exactly) the bits of $v$ before terminating with output $w$. Letting $q_{i} \stackrel{\Delta}{=} \sum_{v \in W_{i}} 2^{-l(v)}$, it is then evident that $\lim _{i \rightarrow \infty} q_{i}=\mu(w)$. Theorem 1 below states the converse of this observation, that in fact every enumerable semi-measure $\mu$ is $\mu_{\pi}$ for some $\pi$.

A semi-measure $\mu$ is called computable if it is enumerable and for each $w$ there is also a computably enumerable weakly decreasing approximating sequence $q_{0}, q_{1}, q_{2}, \cdots \rightarrow \mu(w)$. There are computable semi-measures that are not probability measures, but every enumerable probability measure is computable: we can enumerate the sum $\sum_{w^{\prime} \neq w} \mu\left(w^{\prime}\right)$ by dovetailing to obtain $q_{0}^{*}, q_{1}^{*}, q_{2}^{*}, \ldots$, and then $1-q_{0}^{*}, 1-q_{1}^{*}, 1-q_{2}^{*}, \ldots$ converges from above to $1-\sum_{w^{\prime} \neq w} \mu\left(w^{\prime}\right)=\mu(w)$. Corollary 1 below states that the computable probabilities measures are exactly those of the form $\mu_{\pi}$ for some a.s.-terminating program $\pi$ (see, e.g., Freer et al. 2012; Dal Lago and Zorzi 2012).

In addition to encompassing the space of randomized algorithms, probabilistic programs are of special interest in cognitive science because of their ability to provide compact representations of quite complex distributions, e.g., over combinatorially rich spaces (Goodman et al., 2014; Piantadosi et al., 2016). By encoding these distributions only implicitly through the program's objective probability of returning different outputs, they make an attractive candidate for plausible representations of subjective probability (see Icard 2016 for discussion). Moreover, it is often possible to define programs for automatically representing conditional distributions, and thus to apply and adapt the tools of Bayesian statistics to this setting (Tenenbaum et al., 2011; Freer et al., 2012).

## Why Non-Terminating Programs?

The enumerable semi-measures form a larger, and arguably more natural, class than the computable measures, but what is the reason to include them in our study of cognitive agents?

The claim of this section is that there is a tension between allowing rich, interesting programs and ensuring those programs a.s. terminate. It is well known that testing whether a deterministic program halts is an undecidable $\left(\Sigma_{1}^{0}\right)$ problem,
and verifying a.s. termination of a probabilistic program is of even higher complexity ( $\Pi_{2}^{0}$, see Kaminski and Katoen 2015). It follows that the only way to ensure a.s. termination is to restrict to smaller, controlled fragments. This is confining both for the cognitive scientist proposing psychological models, and for the learning agent who may need to construct and induce programs on the fly.

## Between Termination and Non-Termination

The boundary between a.s.-terminating programs and nonterminating programs often looks quite arbitrary. To illustrate, we use a very simple example close to recent work on intuitive physics (e.g., Sanborn et al. 2013; Battaglia et al. 2013). This work models people's ability to understand and predict physical events using probabilistic programs for constructing internal "simulations" that operate in rough accordance with physical laws. The example here is far less sophisticated, merely concerning speed along a single spatial dimension. Consider a proverbial tortoise-and-hare scenario, where the tortoise is moving ahead at a constant rate following a slight head start, and the erratic rabbit is nonchalantly racing to catch up. We might suppose that the hare leaps forward some random distance about every fourth step that the tortoise takes. The question is when, if ever, the hare will catch up. Imagine this prediction arising from mental simulations of something like the following program $\pi \downarrow$ :

```
t = 5; h = 0
while (h < t):
    t = t + 1
    if (flip(0.25)): h = h + Uniform(1,7)
return t
```

For instance, a person observing a rabbit chasing a tortoise might extract a program like $\pi^{\downarrow}$ in order to make predictions about what will happen some number of steps later.

Where $H_{k}$ is the distance traveled by the hare at stage $k$, consider the random variable $X_{k}=(5+k)-H_{k}$, measuring the extent of the tortoise's current lead. It is easy to show that the sequence $\left\{X_{k}\right\}$ forms a random walk martingale, and specifically that $\mathbb{E}\left[X_{k+1} \mid X_{1}, \ldots, X_{k}\right]=X_{k}$, and so $\mathbb{E}\left[X_{k}\right]=5$ for all $k$. By the recurrence property for symmetric random walks, we reach $X_{k} \leq 0$ at some stage $k$ almost surely. Thus this program $\pi^{\downarrow}$ halts with probability 1. (Cf. Chakarov and Sankaranarayanan 2013 for powerful a.s.-termination proof techniques that cover examples like this.)

While $\pi^{\downarrow}$ as written terminates, small changes in the parameters of the program lead to positive probability of nontermination. For instance, if $t$ is instead incremented by $1+\varepsilon$ at each step, or if the increase in h is drawn uniformly from an interval $(1,7-\varepsilon)$, for $\varepsilon>0$, then the resulting program $\pi^{\uparrow}$ may not halt because the expected distance between the tortoise and hare constantly increases. In particular, there will be a constant $C>0$, such that for any fixed $x_{k}$, we have $\mathbb{E}\left(X_{k+1} \mid x_{k}\right)-x_{k}=C$. Hence $\mathbb{E}\left(X_{k+1}\right)=\mathbb{E}\left(\mathbb{E}\left(X_{k+1} \mid X_{k}\right)\right)=$ $\mathbb{E}\left(X_{k}\right)+C$, from which it follows $\mathbb{E}\left(X_{k}\right)=5+k C$ for each $k$.

This means that the long run expected value of $X_{k}$ is infinite, and the program fails to halt with some positive probability.

One might suspect that this theoretical distinction could have practical repercussions. Would we not want some guarantee that our program would eventually halt? The problem with this line of thought is that, from a practical perspective, non-termination is not any worse than eventual termination but only after an inordinate amount of time. Simulating the program $\pi^{\downarrow}$ above-and terminating computation whenever the number of steps reaches an upper bound of, say, $10^{7}$ —we see that the program reaches this upper bound about $.01 \%$ of the time. Though a large majority ( $\sim 75 \%$ ) of runs end within 100 steps, the empirical average runtime is in the tens of thousands. ${ }^{1}$ Thus, in some small number of cases we would presumably have to terminate computation anyway. From this simulation perspective, the behavior of $\pi^{\uparrow}$, taking $\varepsilon>0$ to be very small, is empirically nearly indistinguishable. The fact that some of these runs might never terminate is immaterial, practically speaking.

This argument is about possible non-termination, and it does not distinguish between computable and merely computably enumerable distributions. If we increment $t$ by a computable real number $1+\varepsilon$, then, though $\pi^{\uparrow}$ might never halt, $\mu_{\pi^{\uparrow}}$ is actually a computable semi-measure, with a computable probability of not halting. However, this situation again may be practically no different from a situation in which $1+\varepsilon$ can only be approximated from below. This paper is a plea mainly for non-terminating programs, and one could in principle accept non-termination but still insist on computability. There may be contexts where insistence on computability may be appropriate (see the section below on conditioning); the claim of this paper, however, is that we ought not make this restriction in general.

## A Remark on Levels of Analysis

The argument that $\pi^{\downarrow}$ and $\pi^{\uparrow}$ are practically indistinguishable assumes that we may have to terminate computation beyond a certain point no matter which one we run, and that the resulting behavior will look nearly indistinguishable. A possible objection at this stage is that by enforcing an upper bound on computation time, we are effectively only considering programs that a.s. (in fact, always) halt anyway. That is, the larger program encompassing both the simulation model itself and whatever controls the simulations always terminates after a bounded amount of time.

This objection is fine as far as it goes, but it undercuts the motivation for considering rich, e.g., recursive, probabilistic programs to begin with. When we write the program $\pi \downarrow$ above in Java, for example, we understand it as encoding an abstract procedure that could in principle run for any amount of time, even though we know no concrete implementation of $\pi \downarrow$ has this property. Indeed, $\pi^{\downarrow}$ abstracts away from many de-

[^353]tails about how the program might be implemented. The idea that we can construe some psychological models in a similar manner is very familiar in cognitive science (Marr and Poggio, 1976). Characterizations of mental phenomena using grammars, recursive constructions, and other devices that license unbounded computations are legitimized by potential gain in conceptual clarity and modularity. We understand while-loops, models of Newtonian mechanics, and so on, in a very general way: we have a good sense of what they can do, what problems they can be used to solve, and how they can be combined with other tools to form even more powerful devices. From this perspective it is unsurprising that such devices would make their way into our cognitive models.

Such issues about levels of analysis are beyond the scope of this paper. The present suggestion is simply that the best arguments favoring liberal use of a.s.-terminating probabilistic machines as cognitive models should extend to the class of all probabilistic machines. Just as there may be practical reasons to avoid computable, but algorithmically intractable, procedures in practice, so it may make sense in many cases to avoid use of procedures that might not terminate. That does not delegitimize their use in cognitive modeling.

## Program Induction

The argument up to this point has been largely negative, that there is no reason to exclude effective semi-measures as possible components of a cognitive model. But there also may be good positive reasons to include them when we consider the learning problem of inducing programs from observations (see, e.g., Lake et al. 2017 for application of this idea to human cognition). Given the high complexity of verifying a.s.termination, the learner seems to be faced with a dilemma: either restrict search to a small fragment of possible programs or risk hypothesizing programs that may not halt.

For example, it is difficult to imagine a sufficiently flexible class of programs-say, a class built out of a few primitive constructions such as while-loops and simple arithmetical operations like addition-from which one could easily obtain the program $\pi^{\downarrow}$ above, but not one of the variants $\pi^{\uparrow}$ that might have some probability of not halting. It is not that one would prefer to construct $\pi^{\uparrow}$ over $\pi^{\downarrow}$, but that they are equally preferable and that separating them in a principled way might be difficult, no matter what method is used to perform the induction. That is, even if the goal is to construct an a.s.-halting program, flexibility in program construction might require the possible construction of non-halting programs.

In light of this possibility, a natural suggestion is to consider enriching program induction frameworks with more expansive classes of programs. Consider Bayesian approaches to program induction. Where $\mathcal{C}$ is some class of (semi-) measures on a space $X$, e.g., on $\{0,1\}^{*}$-so that we can ask about $P(X)$ for any $P \in \mathcal{C}$ and $X \in X$-we could have a prior measure $v$ over $\mathcal{C}$ that induces a mixture distribution $P_{V}$ on $X$ :

$$
P_{\mathrm{v}}(X) \stackrel{\Delta}{=} \sum_{P \in \mathcal{C}} v(P) P(X)
$$

Thanks to Theorem 1, we can always think of each element of $\mathcal{C}$ as a semi-measure $\mu_{\pi}$ defined by a probabilistic program $\pi$ from some class $\Pi$, so $v$ defines a prior on programs in $\Pi$.

Hierarchical Bayesian models fit this description, where $\mathcal{C}$ is typically a parametrized family of distributions and $v$ is a hyperprior over those parameters (though hierarchical models may include more levels), and they are often explicitly encoded as probabilistic programs. Provided one can define appropriate likelihood functions $v(Y \mid P)$ and $P(Y \mid X)$, it makes sense to condition such a mixture distribution on data $Y$ using Bayesian inference:

$$
\begin{equation*}
P_{\mathrm{V}}(X \mid Y)=\sum_{P \in \mathcal{C}} v(P \mid Y) P(X \mid Y) \tag{1}
\end{equation*}
$$

By updating $v$ alongside candidate ground-level distributions $P$, such methods capture effects of learning at multiple levels of abstraction, such as the ability to transfer general principles inferred in one domain to novel but related domains.

Probabilistic programs generalize hierarchical Bayesian models to allow wider classes $C$ of measures. For instance, work by Piantadosi et al. (2016) considers learning in a context where $\mathcal{C}$ is defined by logical expressions of the sort typically used in natural language semantics. Evidently, there is no reason we could not consider classes that include enumerable semi-measures as well. An alluring possibility is to take $\mathcal{C}$ to be the class of all enumerable semi-measures-i.e., all programs-with $v$ assigning a weight to each. Because $\mathcal{C}$ is then computably enumerable, there are many effective semimeasures $v$ assigning positive weight to all probabilistic programs, and even here $P_{V}$ is guaranteed to be an enumerable semi-measure, and thus definable by a program. Learning in this setting is somewhat fraught (see below), but at least such a semi-measure can be represented. By contrast, when $\mathcal{C}$ is the class of computable measures there is no computable $v$ with support exactly $\mathcal{C}$, since that set is undecidable.

## Simplicity Bias

In this broader setting of program induction, as in hierarchical models, it is presumed that a good prior on $\mathcal{C}$ is one that favors simpler hypotheses. This might be achieved, for instance, by defining $v$ with a probabilistic grammar so that shorter programs are automatically given higher probability. A very general proposal for biasing simpler functions, known as Solomonoff induction, is based on ideas from Kolmogorov complexity. In brief, the proposal is to assign probability to a string $w$ in proportion to the shortest (deterministic) program that, when run on a universal Turing machine $U$, produces $w$ as output. The intuition is, data that could be produced by simpler mechanisms should be a priori more likely.

As an aside, there are other applications of simplicitybased constructions inspired by Kolmogorov complexity that make use of enumerable semi-measures. As an example, in their generalization of Shepard's Universal Law of Generalization, Chater and Vitányi (2003) assume enumerable "confusability" probabilities, $P\left(R_{a} \mid S_{b}\right)$-specifying how likely it is that a subject will give a response appropriate to $a$ when
presented with $b$-to develop a notion of similarity between arbitrary representations. The basic idea is that similarity is roughly proportional to the length of the shortest (deterministic) program that would be required to transform one representation into the other. Enumerability is exactly what is needed to derive (a generalized version of) the Universal Law.

But what about simplicity-based Solomonoff induction? There are several issues with Solomonoff induction (including the variant here for semi-measures, due to Zvonkin and Levin 1970). One well known problem is that the resulting prior is very sensitive to the choice of universal Turing machine $U$. In fact, it has been shown that there is a perfect correspondence between weightings $v$ on the class of enumerable semi-measures and choices of universal Turing machines $U$ for the Solomonoff prior (Wood et al., 2011). In other words, the class of Solomonoff priors just is the class of mixture semi-measures $P_{V}$ in which $v$ assigns positive weight to all the enumerable semi-measures. It is therefore questionable whether this framework really does provide a foundation for understanding simplicity, since it is not clear what an "unbiased" choice of $U$ or $v$ would be (see Sterkenburg 2016).

A larger problem with Solomonoff induction, however, concerns the complexity of conditional inference. Whereas each Solomonoff prior is itself computably enumerable, conditioning on data leads to a function that is not even enumerable (specifically, we go from $\Sigma_{1}^{0}$ to $\Delta_{2}^{0}$, see Leike and Hutter 2015 , cf. also the next section). Since the whole point of Solomonoff induction is to learn, this is a disappointing result. In particular, it means that no probabilistic program can represent a conditioned Solomonoff prior. Given the centrality of induction, we would like to understand better what we can do with conditioning, and whether Bayesian program induction is even possible when some of the candidate programs have positive probability of not halting. This leads us to the next section.

## Conditioning Enumerable Semi-Measures

The fact previously mentioned-that effective semi-measures are not closed under conditioning-appears problematic, at least for Bayesian applications of probabilistic programs. It is especially noteworthy given that the computable measures are closed under conditioning in the discrete setting. While a full discussion of conditionalization is beyond the scope of this paper, it is worth briefly clarifying the issue. (Of course, for non-Bayesian approaches to learning programs, e.g., Neelakantan et al. 2016, this may not even be problematic.)

Conditioning an effective semi-measure may produce a function that is only "limit computable" (Leike and Hutter, 2015), meaning the conditional probability of each string can be approximated, but the sequence of rationals need not approach its limit (even weakly) monotonically. The intuition behind this is clear. To determine $\mu(X \mid Y)$ we must compute $\mu(X, Y) / \mu(Y)$. If all we can do is approximate each of $\mu(X, Y)$ and $\mu(Y)$ from below, and we know nothing about how fast we are converging to the correct values, then we know absolutely
nothing about the ratio $\mu(X, Y) / \mu(Y)$ at any finite stage.
In the computable setting, as Freer et al. (2012) explain, it is straightforward to define a single Turing machine QUERY that takes a program $\pi$ and a Boolean condition $\kappa$ (also represented as a probabilistic program), and produces a representation of the posterior distribution $\operatorname{QUERY}(\pi, \kappa)$. The idea is to divide the infinite random bit stream into infinitely many random bit streams, and find the first one that satisfies $\kappa$. Then run $\pi$ using this bit stream to generate an output $w$. As long as $\kappa$ a.s. terminates and returns 'true' with positive probability, $\mu_{\mathrm{QUERY}(\pi, \kappa)}$ correctly defines the posterior distribution.

By the aforementioned result, we know there can be no machine that conditions an arbitrary semi-measure $\mu_{\pi}$ on an arbitrary condition $\kappa$. Nonetheless, provided $\kappa$ stipulates a computable condition, even when $\mu_{\pi}$ is merely computably enumerable, $\operatorname{QUERY}(\pi, \kappa)$ will correctly represent the conditioned semi-measure, which shows that the conditional distribution is also enumerable. Thus, enumerable semi-measures are closed under conditioning with computable queries.

In this sense the situation for enumerable semi-measures is no worse than that for computable measures: both can be conditioned with computable observations in a uniform way. For many settings this does not seem at all limitative. As a simple example, we could imagine conditioning the program $\pi^{\uparrow}$ on the statement that the tortoise reached at least 15 steps. This is an easily, indeed finitely, verifiable proposition.

The main limitative result is rather the one we already mentioned: though we can represent complex semi-measures such as Solomonoff priors, the probabilities of even basic observations like "the first object is a 0 " are not computable. Nonetheless, we can hope for something in this direction. Sufficient conditions for a conditional mixture semi-measure $P_{\mathrm{v}}$ to be semi-computable are not terribly stringent. First, the prior $v$ over semi-measures $\mu \in \mathcal{C}$ should be computable (e.g., this holds if the semi-measures/programs are generated by a probabilistic grammar). Second, as above, the specific data $Y$ must be computably verifiable for each $\mu \in \mathcal{C}$. If both of these are satisfied, then the adapted version of (1)

$$
\begin{aligned}
P_{v}(X \mid Y) & =\sum_{\mu}\left(\frac{v(\mu) \mu(Y)}{\sum_{\mu^{\prime}} v\left(\mu^{\prime}\right) \mu^{\prime}(Y)} \frac{\mu(X, Y)}{\mu(Y)}\right) \\
& =\frac{\sum_{\mu} v(\mu) \mu(X, Y)}{\sum_{\mu} v(\mu) \mu(Y)}
\end{aligned}
$$

is enumerable, and thus representable by a single program, e.g., using an operation like QUERY. Though this falls short of full Solomonoff-style induction, it does generalize what is usually done with Bayesian program induction. It also reveals a distinctive positive reason to entertain specific effective semi-measures as candidate cognitive models. Granted our previous suggestion that it might be beneficial for learning to consider wide classes of programs, putting a computable prior on such a class will result in a enumerable mixture semi-measure $P_{V}$ that can be conditioned.

Questions about conditioning in this more general setting clearly merit further attention, especially in relation to more
realistic inference methods such as MCMC (see, e.g., Goodman et al. 2008). Algorithmic tractability is an obvious worry, but this is already a worry when everything is computable, and it is not obvious that including effective semi-measures exacerbates the problem. Moreover, even in the computable case, for the continuous setting conditionalization is not in general a computable operation (Ackerman et al., 2011). Consequently, as computational-level models of learning and inference, it appears that the effective semi-measures fare no worse than the computable measures.

## Universality

In this final section we offer a proof of the correspondence between probabilistic machines and enumerable semimeasures. Specifically, we show that every enumerable semimeasure can be represented by a probabilistic machine. The proof is similar to proofs for the computable case (e.g., Freer et al. 2012), but we can only use enumerability and must allow for probability of non-halting executions. The exposition is intended to be accessible and intuitive.

Let $\mu$ be an enumerable semi-measure on $\{0,1\}^{*}$. That is, for each word $w \in\{0,1\}^{*}$, there is a computably enumerable weakly increasing sequence $q_{0}, q_{1}, q_{2}, \ldots$ of rational numbers such that $\lim _{i \rightarrow \infty} q_{i}=\mu(w)$. Assume without loss that $q_{0}=0$. Note then that $\mu(w)=\sum_{i=0}^{\infty}\left(q_{i+1}-q_{i}\right)$. Our aim is to show that $\mu=\mu_{\pi}$ for some machine $\pi$.

Let $\left\langle \_, \quad\right\rangle: \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ be a fixed (computable) bijective pairing function with first projection $\rho_{1}(n)=k$ when $n=$ $\langle k, i\rangle$. Let $w_{0}, w_{1}, w_{2}, \ldots$ be a fixed enumeration of $\{0,1\}^{*}$, each with a fixed enumeration of approximating rationals $q_{0}^{k}, q_{1}^{k}, \ldots$ converging from below to $\mu\left(w_{k}\right)$. We define a sequence of rational (thus computable) numbers as follows:

$$
\begin{array}{rll}
r_{0} & \stackrel{\Delta}{=} & q_{0}^{0}=0 \\
r_{n+1} & \stackrel{\Delta}{=} & r_{n}+\left(q_{i+1}^{k}-q_{i}^{k}\right)
\end{array}
$$

where we assume $0=\langle 0,0\rangle$ and $n+1=\langle k, i\rangle$.
Our machine $\pi$ works in stages, observing a random sequence of bits $a_{0}, \ldots, a_{j-1}$ while producing an enumeration $r_{0}, \ldots, r_{j-1}$. At each stage $j$, we observe a bit $a_{j}$ and add a rational $r_{j}$, then check whether, for any $n$ with $0 \leq n<j$, the following condition (2) is satisfied:

$$
\begin{equation*}
r_{n}<\sum_{i=0}^{j} a_{i} 2^{-i}-2^{-j} \text { and } r_{n+1}>\sum_{i=0}^{j} a_{i} 2^{-i}+2^{-j} \tag{2}
\end{equation*}
$$

That is, where $\tilde{p}=\sum_{i=0}^{j} a_{i} 2^{-i}$ is the rational generated so far, we know our randomly generated real number will lie somewhere in the interval $(\tilde{p}-\varepsilon, \tilde{p}+\varepsilon)$, and (2) tells us that this interval sits inside the interval $\left(r_{n}, r_{n+1}\right)$. If this holds, output $w_{\rho_{1}(n+1)}$. Otherwise, move on to stage $j+1$.

Each word $w$ has its probability mass $\mu(w)$ distributed
across different intervals in $[0,1]$. Specifically:

$$
\begin{aligned}
\mu\left(w_{k}\right) & =\sum_{n: \rho_{1}(n+1)=k} r_{n+1}-r_{n} \\
& =\sum_{i=0}^{\infty}\left(q_{i+1}^{k}-q_{i}^{k}\right) .
\end{aligned}
$$

The procedure generates approximations $\tilde{p}=\sum_{i=0}^{j} a_{i} 2^{-i}$ to a random real number, and as soon as we are guaranteed that this random number is in one of our intervals between $r_{n}$ and $r_{n+1}=r_{n}+\left(q_{i+1}^{k}-q_{i}^{k}\right)$, i.e., that no further bits will take us out of that interval (condition (2) above), we halt and output the string $w_{k}$ corresponding to the interval, with $k=\rho_{1}(n+1)$. Clearly, the probability of outputting $w$ is exactly $\mu(w)$, and the probability of not halting at all is $1-\sum_{w} \mu(w)$.
Theorem 1. Probabilistic machines correspond exactly with the enumerable semi-measures.

As every enumerable probability measure is computable, we have the following well-known corollary.
Corollary 1. A.s.-terminating probabilistic machines correspond exactly with the computable measures.

## Conclusion

Defining distributions by means of programs in a universal probabilistic language yields exactly the computably enumerable semi-measures. Our claim has been that this wider class, going beyond a.s.-terminating programs, provides a sensible foundation for theorizing about representation, inference, and learning in cognitive science. Assuming we want to make a clean separation between computational and algorithmic levels of analysis-which is evidently necessary to justify use of anything beyond (probabilistic) finite-state automata in the first place-we see no reason to restrict attention to programs that a.s. terminate, neither for the theorist nor for the learner.

## References

Ackerman, N. L., Freer, C. E., and Roy, D. M. (2011). Noncomputable conditional distributions. In Proceedings of Logic in Computer Science (LICS).
Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45):18327-18332.
Botvinick, M. M. and Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. Cognitive Science, 38:1249-1285.
Chakarov, A. and Sankaranarayanan, S. (2013). Probabilistic program analysis with martingales. In International Conference on Computer Aided Verification (CAV).
Chater, N. and Vitányi, P. (2003). The generalized universal law of generalization. Journal of Mathematical Psychology, 47:346-369.
Dal Lago, H. and Zorzi, M. (2012). Probabilistic operational semantics for the lambda calculus. RAIRO - Theoretical Informatics and Applications, 46(3):413-450.

Freer, C., Roy, D., and Tenenbaum, J. B. (2012). Towards common-sense reasoning via conditional simulation: Legacies of Turing in artificial intelligence. In Downey, R., editor, Turing's Legacy. ASL Lecture Notes.
Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., and Tenenbaum, J. B. (2008). Church: A language for generative models. In Uncertainty in Artificial Intelligence.
Goodman, N. D., Tenenbaum, J. B., and Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. In Margolis, E. and Laurence, S., editors, The Conceptual Mind: New Directions in the Study of Concepts. MIT Press.
Icard, T. F. (2016). Subjective probability as sampling propensity. Review of Philosophy and Psychology, 7(4):863-903.
Kaminski, B. L. and Katoen, J.-P. (2015). On the hardness of almost-sure termination. In Mathematical Foundations of Computer Science, pages 307-318.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences. forthcoming.
Leike, J. and Hutter, M. (2015). On the computability of Solomonoff induction and knowledge-seeking. In 26th International Conference on Algorithmic Learning Theory.
Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry. MIT Memo 357.
Neelakantan, A., Le, Q. V., and Sutskever, I. (2016). Neural programmer: Inducing latent programs with gradient descent. In International Conference on Learning Representations (ICLR).
Piantadosi, S. T., Tenenbaum, J. B., and Goodman, N. D. (2016). The logical primitives of thought. Psychological Review, 123(4):392-424.
Putnam, H. (1967). Psychophysical predicates. In Capitan, W. H. and Merrill, D. D., editors, Art, Mind, and Religion. Pittsburgh University Press.
Rescorla, M. (2015). The computational theory of mind. In Zalta, E. N., editor, Stanford Encyclopedia of Philosophy.
Sanborn, A. N., Mansinghka, V. K., and Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. Psych. Rev., 120(2):411-437.
Sterkenburg, T. F. (2016). Solomonoff prediction and Occam's razor. Philosophy of Science, 83:459-479.
Suppes, P. (1969). Stimulus-response theory of finite automata. Journal of Math. Psych., 6(3):327-355.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331:1279-1285.
Wood, I., Sunehag, P., and Hutter, M. (2011). (Non-)Equivalence of Universal Priors. In Dowe, D., editor, Solomonoff 85th Memorial Conference, pages 417-425. LNCS.
Zvonkin, A. K. and Levin, L. A. (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. Uspekhi Matematicheskikh Nauk, 25(6):85-127.

# The Influence of Speaker's Gaze on Sentence Comprehension: An ERP Investigation 

Torsten Kai Jachmann (jachmann@coli.uni-saarland.de)<br>Heiner Drenhaus (drenhaus@coli.uni-saarland.de)<br>Maria Staudte (masta@coli.uni-saarland.de)<br>Matthew W. Crocker (crocker@coli.uni-saarland.de)<br>Department of Language Science and Technology, Campus Saarbrücken, Germany<br>Cluster of Excellence MMCI, Saarland University, Germany


#### Abstract

Behavioral studies demonstrate the influence of speaker gaze in visually-situated spoken language comprehension. We present an ERP experiment examining the influence of speaker's gaze congruency on listeners' comprehension of referential expressions related to a shared visual scene. We demonstrate that listeners exploit speakers' gaze toward objects in order to form sentence continuation expectations: Compared to a congruent gaze condition, we observe an increased N400 when (a) the lack of gaze (neutral) does not allow for upcoming noun prediction, and (b) when the noun violates gaze-driven expectations (incongruent). The later also results in a late (sustained) positivity, reflecting the need to update the assumed situation model. We take the combination of the N400 and late positivity as evidence that speaker gaze influences both lexical retrieval and integration processes, respectively (Brouwer et al., in press). Moreover, speaker gaze is interpreted as reflecting referential intentions (Staudte \& Crocker, 2011).


Keywords: ERP; N2; N400; late sustained positivity; gaze; prediction; referential expressions

## Introduction

The gaze of a speaker toward objects present in a shared scene in a face-to-face interaction provides a visual cue that expresses the speaker's focus of visual attention and may draw the listener's attention as well (Emery, 2000; Flom, Lee, \& Muir, 2007). This visual cue can be used by the listener to ground and disambiguate referring expressions, infer the speaker's intentions and goals, and thus facilitate comprehension (Hanna \& Brennan, 2007). As most research conducted on the influence of speakers' gaze on listeners' language comprehension focused on behavioral data (e.g.: reaction times, eye movements), little is known about the precise time course for the integration of visual and linguistic information, or which underlying mechanisms are involved.

We therefore conducted an ERP-study to investigate how listeners integrate cues provided by the speaker's gaze when it is time-aligned to an utterance containing statements about the visual context. We monitored listeners' event-related potentials (ERPs) as they observed a stylized face performing gaze actions toward simple objects preceding their mentioning in a simultaneously presented utterance that compared objects in the scene with one-another. The gaze cue corresponding to the second object in the sentence was either Congruent (toward the named object), Incongruent (toward the object
that remains unnamed in the sentence) or Neutral (toward an empty position at the bottom of the screen). This manipulation was intended to shed light on how listeners use speakers' gaze to anticipate and integrate subsequently mentioned referents.

Previous eye-tracking studies have shown that, when a visual context is present, speakers orient their gaze toward an object about $800-1000 \mathrm{~ms}$ before mentioning it (Griffin \& Bock, 2000; Kreysa, 2009). However, it is less clear to what extent such speaker gaze affects listeners' sentence comprehension.

Staudte and Crocker (2011) showed in an eye-tracking study that participants used gaze cues to disambiguate a sentence with multiple same-type referents as soon as it was provided, expressed by a higher inspection rate to the gazed-at object compared to the competitor. Furthermore, a misleading gaze cue lead to a longer reaction time while judging whether the sentence was true or false given the visual scene.

For our ERP study, we hypothesized that listeners integrate gaze cues in a situation model to anticipate which objects are likely to be subsequently mentioned, influencing the retrieval and integration of the noun. Specifically, we expect an N400 modulation will occur as a function of predictability, such that neutral gaze, and even more so incongruent gaze, will increase the amplitude of the N400 compared to a congruent gaze condition. Additionally, we hypothesized that the naming of objects that were previously eliminated based on (incongruent) speaker gaze should lead to a higher cost of integration, possibly reflected by a increased late positivity (Brouwer, Crocker, Venhuizen, \& Hoeks, in press).

## Experiment

In our experiment, German native speakers judged the truthfulness of an auditorily presented sentence given a visual context while their EEG was recorded. Each trial contained a stylized face that performed gaze actions timed to the sentence that was to be evaluated. Every gaze action was performed 800 ms prior to the naming of the corresponding noun. The first gaze was always Congruent toward the object that was named first in the sentence. Gaze to the second object was manipulated such that is was either toward the second named object (Congruent), toward a distractor object that was
never named during the course of the trial (Incongruent) or toward the bottom of the screen where no object was situated (Neutral).

We hypothesized that (1) Congruent gaze toward the upcoming object leads to facilitated retrieval of the corresponding noun, and a reduced N 400 , as it is highly predictable given the visual scene. (2) Incongruent gaze on the other hand is hypothesized to evoke an increased N400 modulation, as the visual information favors predictions of the unnamed object and thereby hinders word retrieval. Additionally, the elimination of the named object based on the visual information demands an update of the situation model and thereby might increases integration costs reflected by a late positivity. (3) As Neutral gaze does not highlight one object more than the other, both remaining objects are equally predictable in the sentence. This might lead to an intermediate retrieval cost of the noun.

## Participants

Forty-five right-handed native speakers of German (Mean age: 24; Age range: [18, 32]; SD: 3.39; Male: 8; Female: 37) took part in the ERP experiment. 15 participants were removed from the analysis due to their behavioral data (3) and too high numbers of eye artifacts (12). ${ }^{1}$ Participants gave informed consent. All participants had normal or corrected-tonormal vision and had no hearing problems. All participants were compensated with $€ 15$ for their participation.

## Stimulus Materials and Procedure

We created 24 pictures of objects of masculine, feminine and neuter gender ( 8 per gender). The pictures were pretested to ensure that they (a) were recognized as the intended objects and (b) were equally complex in their appearance.

Participants were presented with a picture containing three objects of the same gender that varied either in size or brightness arranged in positions above, left and right of the center of the screen. Each screen contained a large, medium, and small object (or bright, medium, and dark object respectively). After 3000 ms , a stylized face appeared in the middle of the screen with a straight gaze toward the participant. The face then performed gaze actions timed to an auditory presented sentence of the form "Verglichen mit dem Auto, ist das Haus verhältnismäßig klein, denke ich" (Compared to the car, the house is relatively small, I think). The utterance was a synthesized German sentence using the CereVoice TTS systems Alex voice (Version 3.2.0). We created different versions of example utterances that varied in intonation contour and turn internal pause length. A Google Form was used to collect responses of seven participants, who listen to those examples with the task to rate their naturalness and order them from most natural to least natural. We selected the version with the most natural rating for the experiment.

In order to keep the influence of the first noun on the second gaze cue as well as on the second noun the same across

[^354]all items, a pause of variable length was introduced after the first noun, so that the distance of the onset of the first noun to the onset of the second half of the sentence always was about 1000 ms . At sentence onset, the face retained its straight gaze but opened the mouth to evoke the impression of the face being the speaker of the sentence. The first gaze cue appeared approximately 800 ms before the first noun was mentioned. This gaze cue was always Congruent toward the named object for all experimental trials. Also, in order to ensure the participants' attention throughout the entire sentence, the first named object in the experimental items was always the medium sized object (or object of medium brightness when brightness was manipulated). If the first mentioned object were the smallest/brightest or biggest/darkest object in the scene, it would not matter which of the other objects were named second, as for both the same comparative adjective would render the sentence true or false.

An example of the visual scene provided in Figure 1 displays the time line of an example trial, with a small house, medium car and a large $t$-shirt. If the $t$-shirt was mentioned first in this context, both of the remaining objects would be smaller. The second, manipulated gaze cue then appeared again 800 ms prior to the onset of the second noun. The gaze was redirected toward the participant 400 ms before the end of the sentence, and the mouth closed on the offset of the sentence.

Each item appeared in three conditions (Congruent / Incongruent / Neutral). In the Congruent condition, the gaze preceding the second noun was directed toward the subsequently named object. In the Incongruent condition, the gaze cue went toward the object that remained unnamed in the sentence. In the Neutral condition, gaze was directed toward the bottom of the screen where no object was present, in order to still present a gaze cue induced by the eye-movement of the face. Additionally, we created versions of those manipulations that were counterbalanced for naturalness. Naturalness was defined as the truth value of the utterance in reality. For example, the in-reality invalid utterance "compared to the car, the house is relatively small, I think" was counterbalanced with the utterance "compared to the house, the car is relatively small, I think". This counterbalancing also led to a swap of the size of the named objects in the visual scene. Using a Latin-square design, this led to a total of six lists.

Each list contained 72 experimental items ( 24 per condition) and 72 fillers with mentioning of an object other then the medium object as the first noun and gaze patterns different from the gaze patterns in the experimental items. As in the experimental items only the second gaze cue was manipulated, $25 \%$ of the fillers (18) contained a manipulation of the first gaze cue instead of the second gaze cue. This version of the fillers still started with a mentioning of the medium object as the first noun in the sentence. The first gaze cue was always neutral and never incongruent in order to enforce the validity of the gaze cues. The remaining fillers were of the same form as the experimental items with the difference


Figure 1: Timeline of an item in the Congruent condition.
that the first mentioned object was either the small or large (bright/dark) object. The second named object in these fillers was the medium object half of the time and the remaining size/brightness the other half. The gaze patterns performed on these fillers always started with a congruent gaze, as in the experimental items, followed by another congruent gaze towards the second named object half of the times (36) and a quarter of the times each by an incongruent or neutral gaze cue (18). This distribution of gaze patterns throughout the experiment led to an overall ratio of congruent gaze actions ${ }^{2}$ of $70 \%$ (204). Another $18 \%$ (51) of the gaze actions were Neutral and only about $12 \%$ (33) of the gaze actions were Incongruent. This way, the validity of the gaze cue was strongly enforced in order to avoid that participants would start to ignore the gaze cues altogether throughout the course of the experiment.

The stimuli were presented using the E-prime software (Version 2.0.10. Psychology Software Tools, Inc.). Each participant was seated in a sound-proof, electro-magnetically shielded chamber in front of a 24 " Dell U2410 LCD monitor (resolution of $1280 \times 1024$ with a refresh rate of 75 Hz ). The distance between the participant and the screen was always 100 cm in order to keep all objects in a $5^{\circ}$ visual angle from the center of the screen. This was done to minimize eye movements throughout the experiment. While the participants were prepared for the recording, they were presented with all objects that occurred throughout the experiment and their naming. The Alex voice of the CereVoice TTS was also used for the naming of the objects. After this, participants were presented with written instructions and completed six practice trials. The items were pseudo randomized for each list and presented in 7 blocks with breaks after each block. After each item, the participants were asked to indicate whether the sentence was true given the visual context they were presented with by pressing one of two buttons. Answers were recorded using a Response Pad RB-834 (Cedrus

[^355]Corporation). The experiment lasted approximately 45 min .

## Data Analysis

The EEG was recorded by $24 \mathrm{Ag} / \mathrm{AgCl}$ scalp electrodes (actiCAP, BrainProducts) and amplified with a BrainAmp (BrainVision) amplifier. Electrodes were placed according to the 10-20 system (Sharbrough et al., 1995). Impedances were kept below $5 \mathrm{k} \Omega$. The ground electrode was placed at AFz . The signal was referenced online to the reference electrode FCz and digitized at a sampling rate of 500 Hz . The EEG files were re-referenced offline to the average of the mastoid electrodes. The horizontal electrooculogram (EOG) was monitored with two electrodes placed at the right and left outer canthi of each eye and the vertical EOG with two electrodes below both eyes paired with Fp 1 and Fp 2 . During recording an anti-aliasing low-pass filter of 250 Hz was used. The EEG data was band pass filtered offline at $0.01-40 \mathrm{~Hz}$ in order to attenuate skin potentials and other low voltage changes as well as line noise and EMG noise (Luck, 2014). Single-participant averages were computed for a 1100 ms window per condition relative to the acoustical onset of the noun following the manipulated gaze cue and the manipulated gaze cue itself. All segments were aligned to a 100 ms pre-stimulus baseline. We semi-automatically screened offline for artifacts.

Due to the nature of the task and the experimental setup containing various eye movements performed by the displayed face, the number of eye artefacts was relatively high. Therefore, we set a threshold of $30 \%$ rejection rate per condition for participant exclusion (i.e.: participants' data with more than 7 rejected trials out of 24 in one or more conditions were removed. On average 5.3 trials per participant and condition ( $22 \%$ ) were rejected due to eye movements). This led to the removal of 12 participants from the analysis. Additionally, the data of 3 participants was removed due to their behavioral data. Participants' data was removed if they gave wrong answers to more than $10 \%$ of the questions. Overall, the two criteria led to the removal of the data of 15 participants. The averaged data of the remaining 30 participants (Mean age: 23.7; Age range: [18, 32]; SD: 3.49; Male: 4) was
exported using BrainVision Analyzer (Version 2.1) BESA export function.

We analyzed the ERP data time-locked to the onset of the second noun following the manipulated gaze cue. We used R ( R Core Team, 2015) to perform repeated measures analysis of variance (ANOVA) using Greenhouse-Geisser correction. We report F values, Greenhouse-Geisser corrected p values and $\eta^{2}$ (partial eta-squared) values as a measure of effect size. All ANOVAs were computed on the F3, Fz, F4, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1 and O2 electrodes including ROIs for frontal (F3, Fz, F4, FC5, FC1, FC2, FC6), central (C3, Cz, C4, CP5, CP1, $\mathrm{CP} 2, \mathrm{CP} 6$ ) and posterior (P7, P3, Pz, P4, P8, O1, O2) distributions.

## Critical Region (Second Noun)

We analyzed the influence of experimental condition (Congruent, Incongruent, Neutral gaze) time locked to the onset of the second noun, including electrode site (frontal/central/parietal) as withinsubject factors. An ANOVA of time window between 150 and 400 ms showed a main effect of condition $\left(F(2,58)=3.99, p<0.05, \eta^{2}=0.12\right)$. There is a globally distributed, significantly larger negativity for both the Incongruent and Neutral condition compared to the Congruent condition $\left(\left(F(1,29)=5.66, p<0.05, \eta^{2}=0.16\right)\right.$ and $\left(F(1,29)=4.86, p<.05, \eta^{2}=0.14\right)$ respectively). However, further Visual inspection revealed that the Neutral condition contains two distinct, frontally distributed peaks within this time window (see Figure 2 for comparison). This is coherent with previous findings by, e.g., Hagoort and Brown (2000). This led to the analysis using a moving time window in this epoch in order to determine whether those peaks are indeed distinct. We split the data of the previous time-window in four overlapping sub-time-windows of 100 ms length with a distance of 50 ms each and introduced those time-windows as a factor in an ANOVA, where the interaction of time-window, longitude and condition showed a significant effect $\left(F(12,348)=1.95, p<0.05, \eta^{2}=0.06\right)$. We find a main effect of condition only in the time windows between $150-300 \mathrm{~ms}\left(F(2,58)=3.93, p<0.05, \eta^{2}=0.12\right)$ and $300-400 \mathrm{~ms}\left(F(2,58)=3.38, p<0.05, \eta^{2}=0.1\right)$. This indicated that the two peaks are indeed distinct. A pairwise comparison of the conditions in the time window of the earlier peak ( $150-300 \mathrm{~ms}$ ) showed that both the Incongruent and Neutral condition retain their significantly larger negativity $\left(\left(\left(F(1,29)=5.82, p<0.05, \eta^{2}=0.17\right)\right.\right.$ and $\left(F(1,29)=4.93, p<.05, \eta^{2}=0.15\right)$ respectively $)$. A pairwise comparison in the time window of the second peak also shows that both the Incongruent and Neutral condition contain a significantly larger negativity $\left(\left(F(1,29)=4.22, p<0.05, \eta^{2}=0.13\right)\right.$ and $\left(F(1,29)=5.99, p<.05, \eta^{2}=0.17\right)$ respectively $)$. Additional to the findings in the early time-window, we analyzed the time-window between $600-1000 \mathrm{~ms}$. The analysis revealed a main effect of experimental condition


Figure 2: ERP time-locked to the Second Noun Onset separated by the Experimental Conditions (Congruent (black), Incongruent (red) and Neutral (blue)). Reported regions are highlighted by boxes. The data presented shows the electrode subset Fz and Pz filtered at 20 Hz (low-pass) for presentation purposes only.
$\left(F(2,58)=3.38, p<0.05, \eta^{2}=0.16\right)$. A pairwise analysis of the conditions showed that the late long-lasting positivity is only present in the Incongruent condition $\left(F(1,29)=7.24, p<0.05, \eta^{2}=0.2\right)$.

Table 1: Summary of the pair-wise computed differences between the (C)ongruent, (I)ncongruent and (N)eutral conditions split by the analyzed epochs. Significance is indicated by *

| Time Window | C - I | C - N | N - I |
| :---: | :---: | :---: | :---: |
| $150-400 \mathrm{~ms}$ | $*$ | $*$ | - |
| $150-300 \mathrm{~ms}$ | $*$ | $*$ | - |
| $250-350 \mathrm{~ms}$ |  | - |  |
| $300-400 \mathrm{~ms}$ | $*$ | $*$ | - |
| $600-1000 \mathrm{~ms}$ | $*$ | - | $*$ |

## Discussion

Research from Van Berkum, Koornneef, Otten, and Nieuwland (2007) suggests that comprehenders predict the upcoming course of a sentence based on the previously gathered information that they integrated in a situation model (Zwaan \& Radvansky, 1998). Various studies have further shown that not only linguistic information is used to form predictions
about upcoming sentence content but also visual information (Staudte \& Crocker, 2011; Ferreira, Foucart, \& Engelhardt, 2013). It therefore seems reasonable to suggest that such visual cues contribute to the construction of the situation model. We interpret the two main components (N400, late positivity) found in terms of the retrieval-integration approach (Brouwer et al., in press): The N400 is modulated by the retrieval difficulty of the upcoming noun, influenced by its predictability given the visual context. Further, the positivity is influenced by the integration difficulties founded in the need to update the situation model.

## N400

Parviz, Johnson, Johnson, and Brock (2011) showed that the N 400 m , representing the N 400 as measured by magnetoencephalography (MEG), is modulated by the information content that is conveyed by a word in a given context. A similar interpretation can be attributed to findings from Willems, Frank, Nijhof, Hagoort, and Van den Bosch (2015). In their study participants listened to spoken narratives. Their results show that words with a higher surprisal let to an increased activation in the left temporal lobe, which has been identified as source of the N400 effect (e.g., Van Petten and Luka (2006)). Additionally, an ERP study from Frank, Otten, Galli, and Vigliocco (2015) revealed a correlation between the amplitude of the N400 and word surprisal. In the current study, the gaze cue preceding the second noun leads to predictions for the upcoming noun. In the Congruent condition, those predictions are fulfilled, which leads to an effortless retrieval of the noun and thus leads to a reduced N 400 amplitude compared to both Neutral and Incongruent conditions. In the Neutral condition, participants have two possible upcoming nouns active. The information conveyed by the noun therefore is higher than in the Congruent condition, as the set of candidates is reduced to the actual target. In the Incongruent condition, the information conveyed by the noun contradicts the prediction made using the visual information. This results in an increase of the retrieval cost of the noun and thus to an increased N400 effect compared to the congruent condition.

The early onset of the negativity ( 150 ms after noun onset) suggests we may in fact be observing modulation of the N 2 component, as well as N400. This is supported by the analysis using a moving time window, which revealed two distinct peaks between $150-300 \mathrm{~ms}$ and $300-400 \mathrm{~ms}$ respectively, especially prominent in the Neutral condition (see Figure 2 for comparison), and is discussed below.

## N2

The globally distributed, early negative component between 150 and 300 ms can be interpreted as a reminiscent of the Phonological Matching Negativity (PMN) as described by Connolly and Phillips (1994). Similar results have been found by Hagoort and Brown (2000). They explain this early effect with a peak around 250 ms as a mismatch between the expected word form given a context and the actual activated word candidates given the speech signal listeners perceive. In
this study, the context was built up by the gaze towards an object present in the visual scene. The following word now either confirms the expectation (Congruent), which in turn leads to no PMN modulation, or disconfirms them (Incongruent), which evokes a large PMN modulation.

Following the account of Hagoort and Brown (2000), which states that 'the N250 effect might reflect the lexical selection process that occurs at the interface of lexical form and contextual meaning', the effect in our Neutral condition could also be explained as such an selection process. Given our visual scene, at the second noun, two of the three objects are still valid targets. The Neutral gaze cue, directed downwards, does not provide any further information about the upcoming word. As both remaining objects are equally plausible, a decision for either one has to be made using the first phoneme of the uttered word, which leads to the discard of one of the two predictions. This selection process elicits the negativity in the N2 region found in the Neutral condition. It is important to highlight that all of the previously named studies establish the predictive context using language. Our study differs in that predictive context is determined solely on the basis of visual information: the linguistic context does not contribute a preference for either of the valid nouns.

## Positivity (600-1000ms)

The relation between updating of a situation model and the occurrence of a late positivity has been demonstrated in various studies (Burkhardt, 2007; Donchin, 1981). Following those accounts, we can interpret our findings in the later time window starting at 600 ms as similarly reflecting the cost of updating the situation model, and integration more generally (Brouwer et al., in press). In both the Congruent gaze and Incongruent gaze condition participants can exploit the gaze cue towards an object in their situation model to make predictions about the upcoming noun. In the Congruent condition, this leads to no violation of those predictions and thereby doesn't require an update of the situation model.

In the Incongruent condition however, the violation of the predictions leads to the need to update the situation model: the gaze cue toward an object leads to the listener's interpretation of the gazed-at object to be the upcoming noun. This in turn leads to the elimination of the remaining object as relevant to the situation model. As the upcoming noun however shows that the previously discarded object is in fact relevant, the situation model has to be updated, which leads to higher integration costs expressed by the late positivity.

The Neutral condition does not draw the focus to one single object but leaves two objects (the so far unnamed objects) as equally possible targets. The prediction of this set of objects is not violated and therefore does not require an update of the situation model.

## Conclusion

We suggest that the N 400 and late positivity are most naturally interpreted in terms of the retrieval-integration approach (Brouwer et al., in press): The N400 findings suggest that
gaze is used to anticipate the upcoming noun, resulting in increased retrieval cost when gaze is absent or incongruent. Interestingly, the late positivity for incongruent gaze, suggests that gaze is interpreted as conveying referential intentions, resulting in integration difficulty only when gaze is misleading. This is consistent with eye-tracking data from Staudte and Crocker (2011). Additionally, we found an N2 modulation preceding the N400, suggesting gaze leads listeners to anticipate specific word forms. More specifically, we argue that the gaze cue preceding the second noun is used in combination with the unfolding situation model to make predictions about the continuation of the sentence. Those predictions are then matched with the auditory input. If the initial phoneme of the input is in line with the prediction, this phoneme provides little new information and therefore facilitates word retrieval. If however the phoneme provides more information, either by helping to reduce the set of predictions to a single target (Neutral condition) or through violation of the predictions (Incongruent condition), an N2 modulation is elicited. In both cases, a subsequent N400 modulation is evoked. If the predictions are completely violated (Incongruent condition), the situation model needs to be updated, which increases the integration cost of the corresponding noun, expressed by a late positivity. Given the findings in the N4 time-window and the late positivity, a classical semantic integration (N4) and reanalysis (P6) account seems unlikely. The integration of the word in the Neutral condition should not lead to a strong N4 modulation as both possible words fit the context without a semantic violation. This predicted modulation for only the Incongruent condition however can be found in the later time-window reflected in the late positivity. Additionally, the late positivity should not be evoked, according to the classic account, as no syntactic reanalysis is needed in any condition. In sum, our findings demonstrate a robust influence of non-verbal gaze cues on several underlying processes, including auditory processing, lexical retrieval, and integration with sentence meaning.

## Acknowledgments

This work was funded by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative.

## References

Brouwer, H., Crocker, M. W., Venhuizen, N., \& Hoeks, J. C. J. (in press). A Neurocomputational Model of the N400 and the P600 in Language Processing. Cognitive Science.
Burkhardt, P. (2007). The p600 reflects cost of new information in discourse memory. Neuroreport, 18(17), 18511854.

Connolly, J. F., \& Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. Journal of Cognitive Neuroscience, 6(3), 256-266.

Donchin, E. (1981). Surprise!... surprise? Psychophysiology, 18(5), 493-513.
Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience \& Biobehavioral Reviews, 24(6), 581-604.
Ferreira, F., Foucart, A., \& Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. Journal of Memory and Language, 69(3), 165-182.
Flom, R. E., Lee, K. E., \& Muir, D. E. (2007). Gazefollowing: Its development and significance. Lawrence Erlbaum Associates Publishers.
Frank, S. L., Otten, L. J., Galli, G., \& Vigliocco, G. (2015). Brain \& Language The ERP response to the amount of information conveyed by words in sentences. Brain and Language, 140, 1-11.
Griffin, Z. M., \& Bock, K. (2000). What the eyes say about speaking. Psychological science, 4(11), 274-279.
Hagoort, P., \& Brown, C. M. (2000). Erp effects of listening to speech: Semantic erp effects. Neuropsychologia, 38(11), 1518-1530.
Hanna, J. E., \& Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. Journal of Memory and Language, 57(4), 596-615.
Kreysa, H. (2009). Coordinating speech-related eye movements between comprehension and production. The University of Edinburgh.
Luck, S. J. (2014). An introduction to the event-related potential technique. MIT Press.
Parviz, M., Johnson, M., Johnson, B., \& Brock, J. (2011). Using language models and latent semantic analysis to characterise the n 400 m neural response. In Proceedings of the australasian language technology association workshop 2011 (pp. 38-46).
R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Staudte, M., \& Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. Cognition, 120(2), 268-291.
Van Berkum, J. J., Koornneef, A. W., Otten, M., \& Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. Brain Research, 1146, 158-171.
Van Petten, C., \& Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. Brain and language, 97(3), 279-293.
Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., \& Van den Bosch, A. (2015). Prediction during natural language comprehension. Cerebral Cortex, bhv075. doi: 10.1093/cercor/bhv075

Zwaan, R. A., \& Radvansky, G. A. (1998). Situation models in language comprehension and memory. Psychological bulletin, 123(2), 162.

# In search for the relevant space of implicit memory deficit in dyslexia 

Sagi Jaffe-Dax (jaffedax@princeton.edu)<br>Department of Psychology<br>Princeton University<br>Princeton, NJ 08540 USA

Merav Ahissar (msmerava@gmail.com)<br>Department of Psychology and Edmond and Lily Safra Center for Brain Sciences<br>The Hebrew University of Jerusalem<br>Jerusalem 91905 ISRAEL


#### Abstract

Studies of dyslexics, whose implicit memory is impaired, suggest that their implicit inference of sound statistics and its integration into perception is inefficient. Specifically, dyslexics' implicit memory decays faster and consequently only accumulates information over shorter temporal windows. We now ask whether this abnormal dynamic is domain general by measuring its cortical distribution. We measure the dynamics of behavioral context effects and the concurrent neural adaptation during fast acquisition fMRI. We find a similar pattern of fast decay of adaptation across a broad range of cortical areas, though most significant effects are found in auditory cortex. This broad neural distribution suggests that the relevant aspect of implicit statistical inferences is not the nature of stimuli, but their temporal distribution.


Keywords: implicit memory; adaptation; fMRI; dyslexia, anchoring hypothesis of dyslexia; Bayesian inference; statistical learning.

## Background

Implicit statistical learning has a profound measurable effect on performance in perceptual discrimination tasks. Seemingly simple and unbiased, serial discrimination places an asymmetric memory load on sequentially presented stimuli. The stimulus presented first has the highest memory load, since it needs be retained until the subsequent stimulus is presented. Hence, relying on priors to compensate for noisy representations is crucial for accurately inferring the perceived stimulus. Priors can be utilized to modify noisy representations into a maximum likelihood estimate of stimuli, those based on previous exposures that led to the formation of priors. The integration of priors has been modeled computationally in the context of auditory perception for signal detection tasks (Treisman \& Williams, 1984) and recently for 2-tone frequency discrimination (Raviv et al., 2012, 2014; reviewed in Bausenhart, Bratzke, \& Ulrich, 2016). Raviv et al. (2012) model asserts that participants do not compare the representations of the first and second tones, as requested to do and as they introspectively do. Rather, they compare the representation of the second tone to the integrated representation of the first tone with the estimated mean (prior) of previous stimuli. Thus, the representation of the first tone is contracted towards the mean of previous trials, and when this contraction is in
the direction of the correct response, the success rate will increase.

Accordingly, trials in this serial discrimination task can be divided into those in which contraction increases the perceived difference between the two stimuli, hence increasing success rate (Bias + ), and trials in which contraction decreases the perceived difference, and decreases the success rate (Bias-). The magnitude of the contraction can be quantified as the difference in performance on these two types of trials.

Contraction bias in serial comparison tasks has been observed in the visual (Ashourian \& Loewenstein, 2011; Fischer \& Whitney, 2014; Lages \& Treisman, 1998; Liberman, Fischer, \& Whitney, 2014), auditory (Lu, Williamson, \& Kaufman, 1992; Raviv et al., 2012; Treisman \& Williams, 1984) and tactile (Hairston \& Nagarajan, 2007) modalities, and was even observed in tactile velocity discrimination tasks in rats (Fassihi, Akrami, Esmaeili, \& Diamond, 2014) and vibro-tactile discriminations in monkeys (Romo, Hernández, Zainos, Lemus, \& Brody, 2002).

Dyslexia is a wide spread learning disability which poses an obstacle in acquiring academic education. It is probably the most prevalent learning disability. Defined as a "specific and significant impairment in the development of reading skills that is not accounted for by mental age, visual acuity problems, or inadequate schooling" (WHO, 2010), it affects $5 \%$ of the world's population (Lindgren, De Renzi, \& Richman, 1985).

Dyslexics are diagnosed by their reading difficulty but are also consistently found to have difficulties in non-linguistic perceptual tasks (McAnally \& Stein, 1996; Sperling, Lu, Manis, \& Seidenberg, 2005). Specifically, dyslexics exhibit poorer utilization of implicit memory to compensate for noisy observation (Ahissar, Lubin, Putter-Katz, \& Banai, 2006; Jaffe-Dax, Raviv, Jacoby, Loewenstein, \& Ahissar, 2015; Oganian \& Ahissar, 2012). We recently accounted for this deficient memory usage by faster recovery from neural adaptation, as measure in Event Related Potential (ERP; Jaffe-Dax, Frenkel, \& Ahissar, 2017). Importantly, dyslexics also exhibit poorer utilization of implicit memory in visual discrimination tasks (Jaffe-Dax, Lieder, Biron, \& Ahissar, 2016). Namely, their implicit memory impairment is not modality-specific, but rather general multi-modal.

Importantly, even in linguistic tasks, compared to good readers, dyslexics perceptual performance is adequate when the stimuli are equally unfamiliar to both groups, and only differ from peers in familiar contexts (Perrachione, Del Tufo, \& Gabrieli, 2011). This accumulative body of research stands in contrast to both the traditional phonological deficit theory of dyslexia (Snowling, 2000) and the claim for dyslexics' overall noisy neural representation (Hancock, Pugh, \& Hoeft, 2017).

Implicit memory is an inaccessible cognitive module which has precise resolution, large capacity and long term retention (Schacter, 1987). Contemporary studies define implicit memory as a memory which does not depend on Medio-Temporal lobe activity, or inaccessible to subject's awareness. This negative definition has been recently challenged by a general theory of wide-spread cortical plasticity in response to perceived events (Reber, 2013). Implicit memory utilization has been recently quantified using a heuristic approximation of Bayesian inference in simple perceptual task (Raviv et al., 2012). This well-defined grounded model of high level cognitive aptitude paved a way for a search for well-defined neural mechanism which would give rise to implicit memory.

While neural adaptation has a long history of research and has been studied intensively in human and animals (Khouri \& Nelken, 2015), its cognitive role and behavioral implication are yet poorly understood. The time scale of recovery from neural adaptation has been linked with the time scale of implicit memory trace ( Lu et al., 1992). Our preliminary findings of fast decay suggest that this rate of decay of adaptation is a crucial machinery for statistical learning. Though traditionally adaptation was considered as reflecting fatigue, we gradually understand that it is used for statistical learning and updating of categories (Kleinschmidt \& Jaeger, 2016).

In the current study, we used fMRI to localize dyslexics' cortical site of fast decay of adaptation. We expected that, if their deficit was specific to speech sounds, the difference in dynamics would affect only specific areas of the temporal lobe, and perhaps also some parietal areas. However, if their deficit was domain general, then all areas activated by these stimuli would show a similar pattern compared with controls.

Our results show, that all active regions exhibit faster recovery from adaptation among dyslexics. We also found additional support to the relation between implicit memory and the time-scale of recovery from adaptation by replicating the parallel finding of shorter adaptation and shorter implicit memory in a special population - dyslexics. Our findings suggest that neural adaptation in cortical areas that correspond to the specific perceptual event carries implicit memory.

This current study shows that dyslexics' implicit memory impairment is dictated by the dynamic decay of their representation throughout the cortex and not by a modalityspecific difficulty.

## Results

We administered two-tone frequency discrimination to 20 dyslexics and 19 good readers in four conditions of trial intervals: 3, 6, 9 and 15 seconds (trial onset to trial onset; TOA). Implicitly, the representation of the first tone is degraded and contracted towards the mean of previously experienced tones (contraction bias). In Bias + trials this bias extended the perceived difference between the current pair and improve their detectability; while in Bias- trials this bias decreased the perceived difference between the current two tones and hence hampered performance (Raviv et al., 2012). We previously found that this implicit memory impact decayed with the extension of inter-trial intervals - it was weaker in conditions of longer inter-trials intervals than in conditions with short inter-trial intervals (Jaffe-Dax et al., 2017). We measured implicit memory as the difference in sensitivity ( $d^{\prime}$ ') between Bias + and Bias- trials and modelled its decay using an exponential decay: $\Delta d^{\prime}(T O A)=a+$ $b \exp (-T O A / \tau)$, where $a$ denotes the estimated $\Delta d^{\prime}$ at $t \rightarrow$ $\infty$ (asymptotic level); $\beta$ denotes the difference between the $\Delta d^{\prime}$ at $t=0$ and at $t \rightarrow \infty$ (decay magnitude); $\tau$ denotes the time it takes for $\Delta d^{\prime}$ (at $t=0$ ) to decay to $1 / e(\sim 37 \%)$ of its initial value (temporal slope parameter). A small $\tau$ indicates fast decay.

In this current study, we replicate faster implicit memory decay among dyslexics outside of the magnet (in a training session), but surprisingly not in-scan. In the training session dyslexics' implicit memory decay was faster than controls' (group $\tau \pm$ IQR: controls: $2.3 \pm 3.4$; dyslexics': $0.4 \pm 0.4 ; z=$ $2, p<0.05$, Mann-Whitney U-test). Figure 1 shows the difference in $d^{\prime}$ between the trials that benefit from prior integration and trials that loose from it as a function of intertrial interval. For both groups, we did not observe a significant decay of implicit memory in the scanner, Perhaps the scanner noise forced the subjects to compensate for their noisy perception in all TOA conditions.


Figure 1. Implicit memory decay is faster among dyslexics in the training session, but not in-scan. Controls in blue; dyslexics in red.

We expected an attenuation of neural activity in the conditions with short inter-trial interval relative to the conditions with long inter-trial intervals, as neural adaptation decays. Specifically, we fitted their BOLD activity in the four ITI conditions to an exponential decay model $(\beta(T O A)=$ $a+b \exp (-T O A / \tau))$ and searched for regions in which the


Figure 2. Estimated groups' model parameters for exponential BOLD decay across the cortex. A. Controls' asymptote level (a). B. Dyslexics' asymptote level (a). C. Controls' magnitude of decay (b). D. Dyslexics' magnitude of decay (b). E. Controls' decay time ( $\tau$ ). F. Dyslexics' decay time ( $\tau$ ).
model parameters ( $a, b$ or $\tau$ ) differed significantly between the groups. Figure 2 shows the average groups' parameters of the exponential decay model. Throughout the cortex, dyslexics' estimated decay rate was faster than controls' (i.e. shorter $\tau$ ). This confirms our prior hypothesis regarding their faster recovery from adaptation. Unexpectedly, the asymptote level (a) of the model also differed between the groups. Specifically, the overall BOLD level of controls was higher than dyslexics'. This could be attributed to overall lower signal-to-noise ratio in dyslexics' BOLD signal, but surely worth further investigation and verification.
We compared the individually estimated parameters of BOLD decay between the groups and found that the decay time differed significantly in the left primary auditory cortex (group $\tau \pm$ IQR: controls: $25.4 \pm 990.9$; dyslexics': $9.1 \pm$ 113.2; $z=2.2, p<0.05$, Mann-Whitney U-test). Unexpectedly, the asymptotic level differed significantly in this region (group $a \pm \mathrm{IQR}$ : controls: $10.4 \pm 298$; dyslexics': $9.1 \pm 113.2 ; z=2.5, p<0.05$ ), in the left central gyrus (controls: $7.7 \pm 230.9$; dyslexics': $3.9 \pm 12.5 ; z=2, p<0.05$ ) and in the right ventral frontal cortex (controls: $6.6 \pm 180.1$; dyslexics': $2.2 \pm 55 ; z=2.1, p<0.05$, Mann-Whitney Utests). Figure 3 shows the clusters which revealed significant difference in estimated parameters between the groups.

These significant group differences were mostly apparent in the cortical regions that are known to be most active during this task (Daikhin \& Ahissar, 2015). Recent study with multiple types of stimuli reported reduced stimulus specific adaptation in dyslexics across all related cortical regions (Perrachione et al., 2016). In line with these findings, we
argue that shorter time scale of recovery from adaptation is a general property of dyslexics' cortex, which accounts for their shorter retention of implicit memory.

## Discussion

Adaptation is a simple, well defined candidate for the neural basis of implicit memory. Assuming a labelled line of input units, adaptation from previous stimuli should attenuate responses to upcoming stimulus as a function of its proximity to the previous stimuli. This proximity has both temporal and parametric dimensions, such that similar stimuli should yield a more attenuated response than dissimilar stimuli, and rapid repeating stimuli should yield a more attenuated response than slow repetitions.

In this work, we have presented evidence that neural adaptation in modality specific cortical region decayed faster in a special population of subjects, who also exhibited faster decay implicit memory to the same modality. Taken together, these parallel findings support the purposed model for implicit memory.

A recent imaging study compared adaptation to similar stimuli within a category among dyslexics and good readers. Our suggestion is in line with the researchers' findings of deficient adaptation in dyslexics in every category-specific cortical region that was measured (Perrachione et al., 2016).

Our findings suggest that the property of adaptation which is impaired in dyslexics is the temporal dimension rather than the similarity between stimuli. Namely, their adaptation decay as a function of time and not as a function of representational similarity (which was recently found intact


Figure 3. Group effect in estimated decay model parameters. A. Controls' estimated decay time ( $\tau$ ) was longer than dyslexics' in superior temporal gyrus. B. Controls' estimated asymptote level (a) was higher than dyslexics' in central sulcus and in ventral frontal cortex.
in dyslexics ;Boets et al., 2013). Thus, it is this dynamic feature that governs implicit memory skills and deficits.

## Methods

The demographic, cognitive and reading assessments of this cohort is described in Jaffe-Dax et al., 2017.

In the two-tone frequency discrimination task, subjects were asked to indicate which of the two tones had a higher pitch. The tones were drawn from a uniform distribution between $800-1250 \mathrm{~Hz}$ and the frequency difference within each pair was randomly drawn between 1-20\%. Each subject performed 12 blocks of 16 trials. Each block had a constant Trial Onset Asynchrony (TOA) of 3, 6, 9 or 15 seconds (in random order). Subjects performed the task outside of the scanner (training) and in-scan.

Stimuli were digitally constructed using Matlab 2015b (The Mathworks Inc., Natwick, MA) and administered
through inserted sound attenuating MR compatible S14 earphones (Sensimetrics Corporation, Malden, MA).

Prior to the functional scan, a high-resolution $(1 \times 1 \times 1$ mm resolution) T 1 -weighted magnetization-prepared rapid acquisition gradient-echo (MPRAGE) images were acquired using a 3-T Magnetom Skyra Siemens scanner and a 32channel head coil, at the ELSC Neuroimaging Unit (ENU). The functional MRI protocols were based on a multislice gradient echo-planar imaging and obtained under the following parameters: $\mathrm{TR}=1 \mathrm{~s}, \mathrm{TE}=30 \mathrm{~ms}$, flip angle $=90^{\circ}$, imaging matrix $=64 \times 64$, field-of-view $=192 \mathrm{~mm} ; 42$ slices with 3 mm slice thickness and no gap, were oriented in ACPC plane, covering the whole brain, with functional voxels of $3 \times 3 \times 3 \mathrm{~mm}$ and multiband parallel imaging with acceleration factor $=3$ (Moeller et al., 2010).

Each condition was modelled separately to account for its contribution to the measured BOLD signal in each voxel, i.e., its $\beta$ values for each TOA condition. A spherical searchlight with 4 voxels radius was performed centered at each voxel to
average $\beta$ values for each condition. An exponential decay model (see Results) was fitted to the smoothed $\beta$ values and its parameters were estimated for each center voxel.

## Acknowledgments

This study was supported by the Israel Science Foundation (ISF grant no. 616/11 and Canada-Israel grant no. 2425/15), the Gatsby Charitable Foundation, The German-Israeli Foundation for Scientific Research and Development (grant no. I-1303-105.4/2015), Canadian Institutes of Health Research (CIHR), The International Development Research Center (IDRC) and the Azrieli Foundation.

## References

Ahissar, M., Lubin, Y., Putter-Katz, H., \& Banai, K. (2006). Dyslexia and the failure to form a perceptual anchor. Nature Neuroscience, 9(12), 1558-64. https://doi.org/10.1038/nn1800
Ashourian, P., \& Loewenstein, Y. (2011). Bayesian inference underlies the contraction bias in delayed comparison tasks. PloS One, 6(5), e19551. https://doi.org/10.1371/journal.pone. 0019551
Bausenhart, K. M., Bratzke, D., \& Ulrich, R. (2016). Formation and representation of temporal reference information. Current Opinion in Behavioral Sciences, 1-7. https://doi.org/10.1016/j.cobeha.2016.01.007
Boets, B., Op de Beeck, H. P., Vandermosten, M., Scott, S. K., Gillebert, C. R., Mantini, D., ... Ghesquière, P. (2013). Intact but less accessible phonetic representations in adults with dyslexia. Science (New York, N.Y.), 342(6163), 1251-4. https://doi.org/10.1126/science. 1244333
Daikhin, L., \& Ahissar, M. (2015). Fast learning of Simple Perceptual Discrimination Reduces Brain Activation in Working Memory and in High-level Auditory Regions. Journal of Cognitive Neuroscience, 27(7), 1308-1321.
Fassihi, A., Akrami, A., Esmaeili, V., \& Diamond, M. E. (2014). Tactile perception and working memory in rats and humans. Proceedings of the National Academy of Sciences of the United States of America, 111(6), 23316. https://doi.org/10.1073/pnas. 1315171111

Fischer, J., \& Whitney, D. (2014). Serial dependence in visual perception. Nature Neuroscience, 17(5), 738747. https://doi.org/10.1016/j.cub.2014.09.025

Hairston, I. S., \& Nagarajan, S. S. (2007). Neural mechanisms of the time-order error: an MEG study. Journal of Cognitive Neuroscience, 19(7), 1163-74. https://doi.org/10.1162/jocn.2007.19.7.1163
Hancock, R., Pugh, K. R., \& Hoeft, F. (2017). Neural Noise Hypothesis of Developmental Dyslexia. Trends in Cognitive Sciences, $x x$, 1-15. https://doi.org/10.1016/j.tics.2017.03.008
Jaffe-Dax, S., Frenkel, O., \& Ahissar, M. (2017). Dyslexics' faster decay of implicit memory for sounds and words is manifested in their shorter neural adaptation. eLife, 6, 1-19. https://doi.org/10.7554/eLife. 20557
Jaffe-Dax, S., Lieder, I., Biron, T., \& Ahissar, M. (2016).

Dyslexics' usage of visual prior is impaired. Journal of Vision, 16, 1-9. https://doi.org/10.1167/16.9.10.doi
Jaffe-Dax, S., Raviv, O., Jacoby, N., Loewenstein, Y., \& Ahissar, M. (2015). A Computational Model of Implicit Memory Captures Dyslexics' Perceptual Deficits. Journal of Neuroscience, 35(35), 1211612126. https://doi.org/10.1523/JNEUROSCI.130215.2015

Khouri, L., \& Nelken, I. (2015). Detecting the unexpected. Current Opinion in Neurobiology, 35, 142-147. https://doi.org/10.1016/j.conb.2015.08.003
Kleinschmidt, D. F., \& Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? Psychonomic Bulletin \& Review, 23, 678-691. https://doi.org/10.3758/s13423-015-0943-z
Lages, M., \& Treisman, M. (1998). Spatial frequency discrimination: Visual long-term memory or criterion setting? Vision Research, 38(4), 557-572. https://doi.org/10.1016/S0042-6989(97)88333-2
Liberman, A., Fischer, J., \& Whitney, D. (2014). Serial dependence in the perception of faces. Current Biology, 24(21), 2569-2574. https://doi.org/10.1016/j.cub.2014.09.025
Lindgren, S. D., De Renzi, E., \& Richman, L. C. (1985). Cross-National Comparisons of Developmental Dyslexia in Italy and the United States. Child Development, 56(6), 1404-1417.
Lu, Z.-L., Williamson, J., \& Kaufman, L. (1992). Behavioral Lifetime of Human Auditory Sensory Memory Predicted by Physiological Measures. Science, 258(5088), 1668-1670.
McAnally, K. I., \& Stein, J. F. (1996). Auditory Temporal Coding in Dyslexia. Proceedings of the Royal Society B: Biological Sciences, 263(1373), 961-5. https://doi.org/10.1098/rspb.1996.0142
Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., \& Uğurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16 -fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain FMRI. Magnetic Resonance in Medicine, 63(5), 1144-1153. https://doi.org/10.1002/mrm. 22361
Oganian, Y., \& Ahissar, M. (2012). Poor anchoring limits dyslexics' perceptual, memory, and reading skills. Neuropsychologia, $\quad 50(8)$, 1895-905. https://doi.org/10.1016/j.neuropsychologia.2012.04.01 4
Perrachione, T. K., Del Tufo, S. N., \& Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. Science (New York, N.Y.), 333(6042), 595. https://doi.org/10.1126/science. 1207327
Perrachione, T. K., Del Tufo, S. N., Winter, R., Murtagh, J., Cyr, A., Chang, P., ... Gabrieli, J. D. E. (2016). Dysfunction of Rapid Neural Adaptation in Dyslexia. Neuron, 92(6), 1383-1397. https://doi.org/10.1016/j.neuron.2016.11.020

Raviv, O., Ahissar, M., \& Loewenstein, Y. (2012). How recent history affects perception: the normative approach and its heuristic approximation. PLoS Computational Biology, 8(10), e1002731. https://doi.org/10.1371/journal.pcbi. 1002731
Raviv, O., Lieder, I., Loewenstein, Y., \& Ahissar, M. (2014). Contradictory Behavioral Biases Result from the Influence of Past Stimuli on Perception. PLoS Computational Biology, 10(12), e1003948. https://doi.org/10.1371/journal.pcbi. 1003948
Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. Neuropsychologia, 51(10), 2026-2042.
https://doi.org/10.1016/j.neuropsychologia.2013.06.01 9
Romo, R., Hernández, A., Zainos, A., Lemus, L., \& Brody, C. D. (2002). Neuronal correlates of decision-making in secondary somatosensory cortex. Nature Neuroscience, 5(11), 1217-1225. https://doi.org/10.1038/nn950
Schacter, D. L. (1987). Implicit Memory: History and Current Status. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(3), 501-518.
Snowling, M. J. (2000). Dyslexia. Wiley-Blackwell.
Sperling, A. J., Lu, Z.-L., Manis, F. R., \& Seidenberg, M. S. (2005). Deficits in perceptual noise exclusion in developmental dyslexia. Nature Neuroscience, 8(7), 862-3. https://doi.org/10.1038/nn1474
Treisman, M., \& Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. Psychological Review, 91(1), 68-111. https://doi.org/10.1037/0033-295X.91.1.68
WHO. (2010). International Classification of Deseases (ICD10). Retrieved from http://apps.who.int/classifications/icd10/browse/2016/ en\#/F81

# Cognition Influencing Auditory Perception in SLD Children: Revisiting the Models of Auditory Processing 

Saransh Jain (saranshavi@gmail.com)<br>PhD Research Scholar, JSS Institute of Speech and Hearing, M.G.Road, Mysuru-570004<br>N.P. Nataraja (npnataraja@gmail.com)<br>Director and Professor, JSS Institute of Speech and Hearing, M.G.Road, Mysuru-570004


#### Abstract

The study assessed the auditory processing abilities and the cognitive skills in children with specific learning disability. It investigates the top-down or bottom-up influence on auditory processing. Using a test battery approach, the association between cognitive skills (verbal working memory and attention) and auditory processing abilities (auditory closure, binaural integration and temporal processing skills) has been measured. The results revealed that cognitive processes significantly affect the bottom-up auditory perception. The effect of cognition was more evident in speech processing than non-speech signal processing. These findings may be useful in designing appropriate therapeutic protocol for children with specific learning disability.


Keywords: dyslexia; learning disability; psychoacoustics; speech perception.

## Introduction

Auditory processing involves the ability of the auditory system to localize and lateralize sounds, discriminate and recognize auditory patterns, temporal aspects of signal, and understanding the auditory information in degraded listening environments (ASHA, 1996; Bellis, 2003; Chermak \& Musiek, 1997), efficiently and effectively. Any disturbance in perceptual processing of the auditory information is referred to as auditory processing disorders (ASHA, 1996).

Auditory processing is affected in individuals with peripheral hearing loss (Neijenhuis, Tschur, \& Snik, 2004), elderly population (Atcherson, Nagaraj, Kennett, \& Levisee, 2015), with certain neurological disorders (Klein et al., 1995), psychological disorders (Iliadou et al., 2013), developmental disabilities like attention deficit hyperactive disorders (Chermak, Somers, \& Seikel, 1998), dyslexia (Hugdahl et al., 1998), learning disability (Kraus et al., 1996), specific language impairment (Cohen, Campbell, \& Yaghmai, 1989), and others. Studies have indicated that children with learning disability show inability in processing complex auditory information (Merzenich et al., 1996). This processing problems have been attributed to the neurophysiological encoding of the speech stimuli and higher level processing deficits (Studdert-Kennedy \& Mody, 1995). Lui et al. (2009) have suggested top-down processing deficit of semantic tasks in auditory modality in children with
reading disability. Verbal working memory, which is the ability to store acoustic information for a short period and plays important role in speech perception (Ingvalson, Dhar, Wong, \& Liu, 2015), is affected in LD children (Alloway \& Alloway, 2010; Wiguna, Wr, Kaligis, \& Belfer, 2012).

Attention deficits have also been found to be prominent in children with learning disability (Finneran, Francis, \& Leonard, 2009). Pinheiro et al. (2010) have reported that LD children have displayed poor divided attention abilities in dichotic listening tasks. In auditory stroop task, selective attention, i.e., the ability to focus on relevant auditory information while ignoring the irrelevant information, has been found to be affected in these children (Faccioli, Peru, Rubini, \& Tassinari, 2008).

Researchers have reported that deficits in the cognitive abilities in the form of verbal working memory and auditory attention have been found in case of LD children. Therefore, it was hypothesized that the cognitive abilities may be associated with auditory processing disorders in LD children. Hence, in the present study, some auditory processing abilities and cognitive skills were assessed in children with specific learning disability.

## Methodology

## Participants

A standard group comparison research design was adapted. 31 children ( 17 males and 14 females) diagnoses as specific learning disability (SLD) by qualified speech language pathologist as per DSM-5 criterion (American Psychiatric Association, 2013) were included. Equal number of typically developing children (TD) were also selected ( $\mathrm{n}=31$ ). The children in both SLD and TD group were native Kannada speakers and belonged to similar socio-economic and cultural background. All the children were within the age range of 810 years and were having normal hearing sensitivity (PTA $\leq 15 \mathrm{dBHL} ; \mathrm{SRT} \pm 10 \mathrm{~dB}$ of PTA; SIS $>90 \%$ ). All children had average or above average intelligence (I.Q. $\geq 90$ ) as assessed by the school psychologist. None of the child had any associated speech, language, otological, psychological and/or neurological problems. The study adhered to the rules of the institutional ethical board and approved to test human
subjects. An informed written consents were obtained from either parents or teachers of all the participant before commencing the tests.

## Assessment of Auditory Processing Abilities

Tests to assess auditory closure, binaural integration, temporal resolution, temporal pattern recognition and temporal masking were selected. Auditory closure is the ability to fill-in the missing auditory information when the external redundancy in the acoustic signal is reduced. Time compressed speech test (TCST) and word recognition in noise test (WRS) were used to assess auditory closure abilities. TCST comprised of 40 standardized Kannada sentences (a Dravidian language) with 3-4 words. These sentences were divided into 2 sets by randomly assigning the sentences into sets, i.e. 20 sentences per set. The sentences were processed to have $50 \%, 60 \%, 70 \%$ and $80 \%$ of temporal compression. The participants were expected to repeat the complete sentence as the compressed sentences were presented.

WRS consisted of five lists with 30 standard Kannada words per list. Each list was processed with steady-state noise to obtain a SNR of $-9,-6,-3,0$ and +3 dB . The participants were expected to repeat the words as they heard. A detailed description of the stimulus parameter is available elsewhere (Jain, Vasudevamurthy, \& Raghavendra, 2015).

Auditory fusion test was used to measure binaural integration skills. The test comprised of standardized Kannada bisyllabic words, where first syllable of the word was presented in to one ear and the corresponding second syllable was presented in other ear, simultaneously. The participants were expected to say the whole word. Two lists of 30 words each, were presented randomly. The lists were constructed in such a way that a syllable which was in the initial position in one word would also occur in the final position of any other word. This reduced the syllabic position effect.

Temporal resolution abilities were measured using temporal modulation transfer function (TMTF) at 8,60 and 200 Hz modulation frequencies. The stimulus was a 500 ms Gaussian noise that was modulated at specific frequency. Using a two alternative force choice method, the participants were asked to identify the interval containing a modulated noise. 90 sound sequences were presented by adapting the maximum likelihood procedure which was implemented using Matlab (Grassi \& Soranzo, 2009).

Temporal pattern recognition was measured using duration pattern test. The test stimuli, as suggested by Musiek (1994), consisted of a 1000 Hz pure tone generated using audacity software (ver. 1.3.14 beta). Two tones, one with 500 ms and another with 250 ms , were used. The tones were patterned in six different combinations such that one tone was presented once while other tone was presented twice, with an inter-tone interval of 300 ms . The participants were asked to repeat the
sequence in which tones were presented. Each tone sequence was presented at least five times.

Temporal masking skills were measured using backward masking test by following the maximum likelihood procedure implemented using Matlab (Grassi \& Soranzo, 2009). The test stimulus was a 1000 Hz tone of 20 ms duration which was presented immediately before a 300 ms band pass noise $(400-1600 \mathrm{~Hz})$. The participants were asked, using a two alternative force choice method, to identify the noise interval which had a tone. 90 pair of sounds were presented.

## Assessment of Cognitive Skills

The cognitive skills were assessed in terms of verbal working memory, divided attention and selective attention. Auditory digit span test (Blackburn, 2011) and operation span test (Kane et al., 2004) were used to assess verbal working memory. In digit span test, sequence of digits were presented binaurally and participants were asked to repeat the sequence in either the same order (forward digit span) or in the reverse order (backward digit span) of presentation. The stimuli were presented in the increasing order of the number of digits. The testing started from two digits level and moved up to ten digit level. Three trials at each level were given and when the participant responded $2 / 3$ trials correctly, the next level of test was administered. The maximum number of digits repeated correctly were considered as the thresholds.

The operation span test based on the study of Kane et al. (2004), has been standardized in Kannada by Jain and Kumar (2016). In this test, the target stimuli (phrases varying from two to five bi-syllabic words) were presented along with a secondary task (a mathematical operation). Participant's task was to solve the mathematical problem and label it as correct and incorrect and subsequently say the word in the order of presentation. For two word sentences, each correct word repeated was given a score of 0.5 ; for three word sentence, each correct word repeated was given a score of 0.33 ; and so on, till five word sentence where each correctly repeated word was given a score of 0.2 . In total, 12 such sentences were presented (three in each phrase length) and the scores were given out of 12 (each sentence carried a total score of one, when all the words were correctly repeated).

The attention skills were measured for divided attention and selective attention task using dichotic digit test (Musiek, 1983). The test comprised of pair of digits presented binaurally. In the divided attention task (free recall), the participants were expected to repeat all the digits presented to them. In selective attention tasks (force recall), the participants focused their attention to one ear only and repeated the digits presented to that ear while ignoring the digits being presented to other ear. Total 30 pair of digits were presented randomly for each task.

## Procedure

The testing was carried out in a sound treated room. The stimuli were presented binaurally at the participant's comfortable loudness level which varied from 65 dB to 80 dB HL, using TDH-39 headphones connected to computer based audiometer (Interacoustics AD-629). The testing took at least 2-2.5 hours for each participant and was conducted in two sitting for the consecutive days. Using this procedure, participants of both the groups were tested and data was collected.

## Data Analysis

Logistic regression with linear or non-linear interpolation was used to measure SNR-50 (SNR level at which participants' responded correctly, at least for $50 \%$ stimuli) for word recognition scores in noise and compression-50 (compression level at which $50 \%$ correct identification of sentences was obtained) for time compressed speech test. The data was normally distributed across groups as per ShapiroWilk test for normalcy ( $\mathrm{p}>0.05$ ) and hence parametric statistics was used. One way analysis of variance was used estimated the significance of differences of scores for auditory processing and cognitive tests between SLD and TD children. The test scores were dependent variables whereas group distribution was independent variable. The partial least square regression-structured equation modeling was used to note the relationship between the test carried out for cognitive skills and auditory processes. Further, it was also used to find out the correlation between cognitive abilities and auditory processes in SLD children.

## Results

The data obtained from descriptive statistical analysis for auditory processing tests and cognitive tests are presented in Figure 1 and 2, respectively as box plots. The mean scores for all the tests (except DPT) were better for TD children in comparison to SLD children. The results of one way ANOVA are shown in Table 1. Statistically significant differences, between TD and SLD groups were found in all the tests of auditory processing and cognitive skills, except for DPT. It was also specific learning disability accounted for more than $50 \%$ variance in the test scores (partial eta square was greater than 0.5 ). An exception to this was word recognition in noise scores and forward digit recall scores, where the effect size was greater than 0.3 only.

The correlation between cognitive skills and auditory processing abilities were measured using partial least square regression. A formative model was created where working memory and attention were considered as latent variables and the measures to assess working memory (digit and operation span) and attention skills (dichotic digit test scores) as observed variables. The effect was seen on three measures of auditory processing i.e., auditory closure, binaural integration and temporal processing. The model had good fit with
standardized root mean residual of 0.036 (Hu \& Bentler, 1998).


Figure 1: Box plots are representing the scores obtained for tests to assess auditory processing abilities.


Figure 2: Box plots are representing the scores obtained for tests to assess cognitive skills.

Table 1: The F-values and significance of difference (pvalues) for tests to assess auditory processing abilities and cognitive skills, between SLD and TD children.

| Test Procedures | df | F-value | p-value |
| :--- | :--- | :--- | :--- |
| Tests to assess auditory processing abilities |  |  |  |
| Time Compressed Speech Test | 1 | 126.34 | 0.000 |
| Word recognition Scores (in noise) | 1 | 51.13 | 0.000 |
| Auditory Fusion Test | 1 | 160.75 | 0.000 |
| TMTF (8 Hz) | 1 | 579.00 | 0.000 |
| TMTF (60 Hz) | 1 | 753.11 | 0.000 |
| TMTF (200 Hz) | 1 | 446.50 | 0.000 |
| Duration Pattern Test | 1 | 1.87 | 0.176 |
| Backward Masking Test | 1 | 579.55 | 0.000 |
| Tests to assess cognitive skills |  |  |  |
| Digit Span Test (Forward) | 1 | 27.79 | 0.000 |
| Digit Span Test (Backward) | 1 | 166.70 | 0.000 |
| Operation Span Test | 1 | 57.64 | 0.000 |
| Dichotic Digit Test (Free Recall) | 1 | 256.47 | 0.000 |
| Dichotic Digit Test (Force Right) | 1 | 149.95 | 0.000 |
| Dichotic Digit Test (Force Left) | 1 | 94.89 | 0.000 |

The regression model is presented in Figure 3. It was noted from the figure that the adjusted $\mathrm{R}^{2}$ indicated $71.6 \%$ of variance in the auditory closure abilities were associated with attention skills and working memory. Similarly, $67.5 \%$ variance in binaural integration abilities, and $91 \%$ variance in temporal processing abilities were attributed to cognitive skills. It was also noted that attention was highly correlated to auditory processing than with working memory. Among the working memory tests, backward digit recall represented the working memory skills maximally. Attention skills were better represented in terms of divided attention. Similarly, TCST was found to be a more reliable measure to assess auditory closure. The temporal processing abilities were better represented by backward masking test.

## Discussion

The present study measured the association between cognitive skills and auditory processing in SLD children. The results of ANOVA revealed significant differences on all the measures of auditory processing and cognition between SLD and TD children (except DPT). Many researchers have reported disorders of auditory processing in LD children (Cohen et al., 1989; Dawes \& Bishop, 2010; Kraus et al., 1996). Therefore the assessment of SLD, using series of tests of auditory processes, like in the present study may provide better information about SLD. Further, cognitive abilities have also being examined previously in such children (Alloway \& Alloway, 2010; Faccioli et al., 2008; Finneran et al., 2009; Pinheiro et al., 2010; Wiguna et al., 2012), but the relationship between cognition and auditory processing has not been investigated intensively. Such investigations, like in the present study, would lead to better understanding of the
relative contribution of top-down or bottom-up processes involved in auditory perception.

At times, it seems that the structured equation modeling used in the present study is under powered, as the sample size is small. However, the power analysis run with effect size of 0.5 and the power coefficient of 0.95 , for five predictor variables (the variables assessed the cognitive skills) indicated that total sample size should be 42 . In the present study, although the sample size for SLD children is 31 only, is it is not much lesser than the suggested sample size. Thus, it was considered that SEM should be a reasonable tool to assess the association between auditory processing abilities and cognitive skills.

The association between cognitive skills and auditory processing are highly significant. Both attention and working memory seems to be influencing auditory processing, and the contribution of attention seems to be more than working memory, especially for temporal processing. Therefore, the findings of the present study may be considered as suggesting the influence of cognitive skills on auditory processing. Similar findings have been reported by other researchers (Moossavi et al., 2014; Murphy et al., 2013) in normal children. Based on the present results, it may be extended to SLD children also.

In the present study, most of the tests used speech stimuli, and the results suggest that the cognition was influencing the auditory processing of speech than auditory processing of tonal/noise perception. ....speech test and cognition (correlation) Similar findings have been investigated by several other investigator (Fedorenko, 2014; Hällgren, Larsby, Lyxell, \& Arlinger, 2001; Larsby, Hällgren, Lyxell, \& Arlinger, 2005). Word recognition in noise required the processing of both speech and non-speech stimuli, showed the path coefficient of WRS (in noise) was 0.379 . This also suggest the contribution of cognition in auditory processing of speech more than for non-speech stimuli. This further strengthen the conclusion that the cognition has greater influence on auditory processing of speech.

## Conclusion

The present study examined the association between cognitive abilities and auditory processing, and highlights the cognitive influence on auditory processing. The findings of the study indicate that the cognitive abilities are associated with auditory processing in SLD children also like in normal. It has also been found that the cognition is associated with auditory processing of speech more than non-speech signal. These findings may be useful in understanding speech perception in SLD children and may be used in designing appropriate speech and language intervention techniques.


Figure 3: The structure equation model showing the association of cognitive skills (in terms of working memory and attention skills) with auditory processing (in terms of auditory closure, binaural integration and temporal processing). The eclipse and rectangle are used to represent latent and observed variables, respectively. In the eclipse, adjusted R2 values are given, and those between the arrow bars are path coefficients of the model.

## References

Alloway, T. P., \& Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. Journal of Experimental Child Psychology, 106(1), 20-29.
American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition). American Psychiatric Association. Retrieved from http://psychiatryonline.org/doi/book/10.1176/appi.books. 9 780890425596
ASHA. (1996). Central Auditory Processing: Current Status of Research and Implications for Clinical Practice. American Journal of Audiology, 5(2), 41.
Atcherson, S., Nagaraj, N., Kennett, S., \& Levisee, M. (2015). Overview of Central Auditory Processing Deficits in Older Adults. Seminars in Hearing, 36(03), 150-161.
Bellis, T. J. (2003). Assessment and management of central auditory processing disorders in the educational setting: from science to practice (2nd ed). Clifton Park, NY: Delmar Learning.
Blackburn, J. (2011). Digit Span Tester (Version 2.1.3) [Windows]. Retrieved from https://sourceforge.net/projects/digitspantester

Chermak, G. D., \& Musiek, F. E. (1997). Central auditory processing disorders: new perspectives. San Diego: Singular Pub. Group.
Chermak, G. D., Somers, E. K., \& Seikel, J. A. (1998). Behavioral signs of central auditory processing disorder and attention deficit hyperactivity disorder. Journal of the American Academy of Audiology, 9(1), 78-84; quiz 85.
Cohen, M., Campbell, R., \& Yaghmai, F. (1989). Neuropathological abnormalities in developmental dysphasia. Annals of Neurology, 25(6), 567-570.
Dawes, P., \& Bishop, D. V. M. (2010). Psychometric profile of children with auditory processing disorder and children with dyslexia. Archives of Disease in Childhood, 95(6), 432-436.
Faccioli, C., Peru, A., Rubini, E., \& Tassinari, G. (2008). Poor readers but compelled to read: Stroop effects in developmental dyslexia. Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence, 14(3), 277-283.
Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. Frontiers in Psychology, 5.
Finneran, D. A., Francis, A. L., \& Leonard, L. B. (2009). Sustained attention in children with specific language
impairment (SLI). Journal of Speech Language and Hearing Research, 52(4), 915.
Grassi, M., \& Soranzo, A. (2009). MLP: a MATLAB toolbox for rapid and reliable auditory threshold estimation. Behavior Research Methods, 41(1), 20-28.
Hällgren, M., Larsby, B., Lyxell, B., \& Arlinger, S. (2001). Cognitive effects in dichotic speech testing in elderly persons. Ear and Hearing, 22(2), 120-129.
Hu, L., \& Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 3(4), 424-453.
Hugdahl, K., Heiervang, E., Nordby, H., Smievoll, A. I., Steinmetz, H., Stevenson, J., \& Lund, A. (1998). Central auditory processing, MRI morphometry and brain laterality: applications to dyslexia. Scandinavian Audiology. Supplementum, 49, 26-34.
Iliadou, V. V., Apalla, K., Kaprinis, S., Nimatoudis, I., Kaprinis, G., \& Iacovides, A. (2013). Is central auditory processing disorder present in psychosis? American Journal of Audiology, 22(2), 201-208.
Ingvalson, E. M., Dhar, S., Wong, P. C. M., \& Liu, H. (2015). Working memory training to improve speech perception in noise across languages. The Journal of the Acoustical Society of America, 137(6), 3477-3486.
Jain, C., \& Kumar, A. U. (2016). Relationship among Psychophysical abilities, Speech Perception in Noise and Working memory in Individuals with Normal hearing sensitivity across different age groups ( PhD in Audiology). University of Mysore, Mysore, India.
Jain, S., Vasudevamurthy, \& Raghavendra, A. P. (2015). Maturation of temporal processing in children: measurements using speech and non-speech stimuli. Journal of Hearing Science, (2), 23-35.
Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., \& Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. Journal of Experimental Psychology. General, 133(2), 189-217.
Klein, S. K., Kurtzberg, D., Brattson, A., Kreuzer, J. A., Stapells, D. R., Dunn, M. A., ... Vaughan, H. G. (1995). Electrophysiologic manifestations of impaired temporal lobe auditory processing in verbal auditory agnosia. Brain and Language, 51(3), 383-405.
Kraus, N., McGee, T. J., Carrell, T. D., Zecker, S. G., Nicol, T. G., \& Koch, D. B. (1996). Auditory neurophysiologic responses and discrimination deficits in children with learning problems. Science (New York, N.Y.), 273(5277), 971-973.
Larsby, B., Hällgren, M., Lyxell, B., \& Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. International Journal of Audiology, 44(3), 131-143.

Liu, L., Friedman, E., Bolger, D., Bitan, T., \& Booth, J. (2009). Children with reading disability show deficits in top-down and bottom-up processing during semantic tasks in both visual and auditory modalities. NeuroImage, 47, S165.
Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., \& Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. Science (New York, N.Y.), 271(5245), 77-81.
Moossavi, A., Mehrkian, S., Lotfi, Y., Faghihzadeh, S., \& Sajedi, H. (2014). The relation between working memory capacity and auditory lateralization in children with auditory processing disorders. International Journal of Pediatric Otorhinolaryngology, 78(11), 1981-1986.
Murphy, C. F. B., La Torre, R., \& Schochat, E. (2013). Association between top-down skills and auditory processing tests. Brazilian Journal of Otorhinolaryngology, 79(6), 753-759.
Musiek, F. E. (1994). Frequency (pitch) and duration pattern tests. Journal of the American Academy of Audiology, 5(4), 265-268.
Neijenhuis, K., Tschur, H., \& Snik, A. (2004). The effect of mild hearing impairment on auditory processing tests. Journal of the American Academy of Audiology, 15(1), 616.

Pinheiro, F. H., Oliveira, A. M. de, Cardoso, A. C. V., \& Capellini, S. A. (2010). Dichotic listening tests in students with learning disabilities. Brazilian Journal of Otorhinolaryngology, 76(2), 257-262.
Smith, S. L., \& Pichora-Fuller, M. K. (2015). Associations between speech understanding and auditory and visual tests of verbal working memory: effects of linguistic complexity, task, age, and hearing loss. Frontiers in Psychology, 6.
Studdert-Kennedy, M., \& Mody, M. (1995). Auditory temporal perception deficits in the reading-impaired: A critical review of the evidence. Psychonomic Bulletin \& Review, 2(4), 508-514.
Wiguna, T., Wr, N. S., Kaligis, F., \& Belfer, M. L. (2012). Learning Difficulties and Working Memory Deficits among Primary School Students in Jakarta, Indonesia. Clinical Psychopharmacology and Neuroscience, 10(2), 105-109.

# Failure to use probability of success in deciding whether to pursue one goal or two. 

Warren R. G. James (r01wj16@abdn.ac.uk)<br>School of Psychology, The Univerity of Aberdeen

Alasdair D. F. Clarke (a.clarke@essex.ac.uk)<br>Department of Psychology, The University of Essex<br>Amelia R. Hunt (a.hunt@abdn.ac.uk)<br>School of Pscyhology, The University of Aberdeen


#### Abstract

Difficult tasks should be attempted one at a time, while easy tasks can be undertaken in parallel. Reinforcing our previous conclusion that people are surprisingly poor at applying this logic, we find people fail to select standing positions that maximize their probability of success in throwing a beanbag into one of two possible hoops. We asked participants to explicitly report their odds of successfully throwing a beanbag into each hoop from the location they had chosen to stand, and estimates were highly accurate. Nonetheless, participants failed to use estimates of success appropriately to maximize success, suggesting a failure of insight, rather than limited or inaccurate information, can account for suboptimal decisions about standing position.


Keywords: Bounded Rationality; Optimal Behaviour; Awareness; Decision Making.

## Introduction

Human skill is limited, and effective decisions must take these limitations into account. In Chess and Go, for example, it is impossible to select the optimal move by mentally simulating every possibility. An effective strategy must take into account the constraints of ones own memory capacity. Simon (1990) used the term bounded rationality to describe decisions that are rational given known constraints.

We recently reported a surprising failure to make effective decisions about whether to pursue one goal or two (Clarke and Hunt, 2016). In one experiment from that study, participants had to throw a beanbag into one of two hoops. The distance between the hoops varied and participants were told which one of the two hoops had been randomly selected to be the target only after they chose a place to stand. The optimal strategy when the hoops are relatively close together is to choose a standing position equidistant from both hoops, making an accurate throw possible irrespective of which hoop is the target. However, as the hoops move further apart, the
probability of a successful throw from the center position drops below $50 \%$. Now the best strategy is to stand close to one of the two hoops and hope it is the target on that trial. Despite the availability of this simple strategy, the distance between the hoops had no systematic effect on where people stood, demonstrating a profound failure to optimize throwing accuracy. The same failure to adjust strategy in response to difficulty was observed in deciding where to fixate to detect one of two targets (see also Morvan and Maloney, 2012) and in allocating attention when trying to memorize digit strings. These experiments took into account each individuals performance limitations by measuring throwing ability, visual acuity, and memory capacity in a separate session involving only single targets with no decisions. This allows for an individualized estimate of when a given participant should switch from attempting both goals to prioritising one. Nonetheless, bounded rationality could explain this failure as participants could be unaware of, or incorrect about, their own abilities (Schraw and Dennison, 1994).

In the current experiment, we repeated the beanbagthrowing task but explicitly asked participants to report on their expected accuracy before each throw. This allows direct comparison of estimated and actual throwing performance. Three outcomes are possible, all of which are informative about the cause of suboptimal decisions in this task. First, consistent with bounded rationality, participants may be unaware of their throwing ability and therefore fail to account for it when selecting standing positions. Second, participants may not adequately attend to their own abilities. Self-talk can improve throwing (Chang et al., 2014), suggesting explicitly reporting about ones own ability can improve task performance. Third, participants may have an accurate representation of their own skill, and drawing their attention to this information may not influence their standing position decisions. In this case, the results would suggest failure to maximize success in this paradigm is an example of a more fundamental limitation in decision competence.

## Methods

## Participants

Twenty-four participants ( 17 female) were recruited via the SONA systems at the University of Aberdeen and took part in return for course credit. The sample size of 24 (12 in each group) was based on our previous experiment (Clarke and Hunt, 2016). Average age was 19 years ( $\mathrm{SD}=1.5$ ).

## Procedure

We used a similar protocol to that of Experiment 2 in Clarke and Hunt (2016). The experiment was carried out over two sessions, with participants carrying out Session 2 a week after Session 1. Both sessions took place in the same sheltered paved area. The paving slabs were used as a convenient unit for measuring distances as they were approximately the same size as the hoops (the slabs measured $0.46 \times 0.61 \mathrm{~m}$ and the hoops had a diameter of 0.4 m ).

The aim of Session 1 was to measure how well each participant could throw a bean bag into hoops placed at seven evenly-spaced distances from three (1.38m) to fifteen ( 6.90 m ) slabs away, in two different directions (direction was counter balanced). A total of 84 throwing trials were completed in this session ( 12 beanbags for each of the 7 target distances). The data gathered from the first session were used to model how participants accuracy decreases as the distance to the target hoop increases (see Figure 1).

In Session 2, participants were again throwing beanbags into hoops, but there were now two hoops, and either one could be the target. Participants were asked to choose a place to stand before throwing, and were not told which hoop would be their target until after they had made this decision. To avoid having to re-position hoops from trial to trial, the area contained six hoops of three different colours, with blue hoops at the furthest distance, yellow at an intermediate distance, and red at the closest distance. The actual distance of each hoop colour depended on performance in the first session, with each colour corresponding to an estimate of the participants throwing accuracy from the centre: Blue hoops were placed at the slab distance where accuracy was expected to be closest to $10 \%$, yellow hoops at $50 \%$, and red hoops at $90 \%$ (see Figure 2). The beanbag on each trial was randomly drawn from a bag, which initially contained nine beanbags, three of each colour. The colour of the beanbag drawn from the bag determined which hoop pair was the target on that trial. Once all nine had been thrown they were replaced. This process was repeated 5 times for a total of 45 trials.

On each trial in Session 2, participants retrieved a beanbag and then chose somewhere to stand (the participants were told that they could stand anywhere on the paved area). The experimenter then informed them about which hoop (north or south) was their throwing target on that trial. The direction was randomised, with each direction equally likely on every trial. It was made clear to the participants that the direction


Figure 1: This shows the set up for session 1 from the participant's point of view.
of each throw was random and had been predetermined, with each direction being equally likely. They then performed the throw and their standing position and throwing accuracy were recorded.

All participants followed the protocol described above, and half of the participants were also instructed to give an estimate (in percentages) of their expected throwing accuracy for both hoops from the location they had chosen to stand. They were prompted to provide this estimate after they chose a place to stand but before they actually threw the beanbag. This group will be referred to as the Online Estimation Group. This condition was included to test whether drawing the participants attention to this component of the problem they were faced with would help them to perform better.

Upon completion of the 45 trials in Session 2, all participants then performed a task similar to Session 1, but instead of throwing, they were required to give an estimate of their accuracy, in percentages, for each distance that had been tested in the first session. Participants stood in one spot, and one hoop was moved to different distances from them in either direction. The distances were split into two sets: 3, 7, 11, 15 and $5,9,13$. One set was presented to the participants first, in ascending order (i.e. getting further from the participant); then the other set was presented to them in a descending order (i.e. getting closer). The order of sets was counterbalanced across participants so each set was presented first equally often. Results were analysed in R ( R Core Team,


Figure 2: The setup for session 2
2016) and modelled using the lmer function from the lme4 package (Bates et al., 2014).

## Results

## Actual accuracy vs. Estimated accuracy

Participants throwing performance in Session 1 is shown in Figure 3. The relationship between accuracy and distance for each participant was modelled using logistic regression. Participants estimate of their own throwing ability is superimposed in blue. The majority of our participants were accurate in their ability to estimate their own throwing ability. This can be summarised by looking at the correlation between actual and estimated accuracy for each individual. This gives a median Pearsons correlation coefficient $r$ of 0.89 ( $\min =0.72$, $\max =0.96$ ).


Figure 3: These graphs illustrate the accuracy (proportion correct) over the various distances for each participant (actual, in red), and their estimates of their own accuracy over the same distances (in blue).

## Standing position

The optimal strategy for the closest hoop distance in Session 2 would be to stand in the middle, as the expected accuracy is $90 \%$ regardless of which hoop was selected as the target. For the farthest hoop distance, the optimal strategy would be to stand next to one of the two blue hoops, as this means that they would be approximately $100 \%$ accurate for that hoop and $0 \%$ accurate for the other hoop. Considering that each hoop was equally likely to be selected, this means that standing next to one blue hoop gives the participant a $50 \%$ chance of success, which is much greater than the $10 \%$ accuracy they would achieve by simply standing in the middle. The point where participants should switch between a centre and a side strategy is marked by a blue line in Figure 3 (the $50 \%$ point in their reported estimate of accuracy is shown by the red line). For the majority of the participants, it makes no difference whether we use their actual accuracy or their estimated accuracy to determine the ideal switch point, given the resolution of our experiment. The black dots illustrate the chosen standing positions on each trial: it is clear from these results that participants do not switch their strategy at either point; in fact, generally speaking, participants do not alter their standing position systematically with the distance of the hoops. These data are similar to those from the throwing task reported in Clarke and Hunt (2016). Interestingly, we can see that one participant (participant B11 in Figure 4) approaches the optimal strategy, particularly with respect to their estimated accuracy. Taken in aggregate, however, participants did not tend to stand closer to the hoop when they were further apart (a paired samples t-test comparing standing position for farthest to the closest hoop distance was non-significant; $t(23)=-.49)$.

## Standing positions across groups

To explore whether being asked to estimate the probability of successfully completing both possible throws had an effect on participants decisions, we compared the standing positions for the control group (Group A) to the online estimate group (Group B). For the closest hoops, participants in Group A stood on average 0.13 (standard deviation $=0.21$ ) of the way from the central point to one of the two hoops, and participants in Group B stood at 0.14 (standard deviation $=0.27$ ). Similarly, there was little difference between the two groups for the blue (far) hoops: participants chose to stand slightly further away from the central point (Group A: $\mathrm{M}=0.128, \mathrm{SD}=$ 0.17 , Group $\mathrm{B}: \mathrm{M}=0.2, \mathrm{SD}=0.22$ ). We conclude that being prompted to verbally report estimated accuracy for each trial had no effect on the strategies participants used to complete the task (close; $\mathrm{t}(22)=.09$, far; $\mathrm{t}(22)=.895)$.

Analysis was also carried out to examine whether the accuracy of a participant's estimate of their own ability was correlated with how closely they followed the optimal strategy. To do this, the r value for each participant (representing the accuracy of their estimate) was correlated with a normalised value for the average distance of the standing position from


Figure 4: Black dots show the standing positions of each participant on each trial in Session 2 as a function of distance of the hoops from the centre (x-axis). Standing position was normalised so that 0 represents having stood at the centre and 1 being one of the side hoops. Red lines represent the distance at which participant should have switched between standing at 0 and standing at 1 based on their actual accuracy, and the blue lines is the switch point based on their own estimates of their accuracy.
the optimal position (in the $10 \%$ accuracy and $90 \%$ accuracy trials). The Pearson's correlation coefficients were very weak for both the $10 \%$ accuracy $(r=-0.12)$ and the $90 \%$ accuracy conditions ( $\mathrm{r}=-0.12$ ). To clarify, had there been an effect of accuracy of estimates, there should have been a negative correlation for both accuracy conditions. The lack of a relationship suggests that participants with a greater awareness of their ability were not better at making using this information.

## Discussion

Participants do not select standing positions that maximize their throwing performance, reinforcing the conclusion that participants fail to solve this task, as previously found by Clarke and Hunt (2016). This experiment was designed to test whether limitations in self-awareness of throwing ability could account for participants poor choices in standing positions. The results suggest participants were highly accurate in reporting on the expected outcome of their decision, but failed to make use of this information when deciding where to stand.

The second possibility we considered in the introduction is that participants have an accurate representation of their own skill, but fail to use this information in making their decision. If this is the case, asking participants to explicitly judge their expected accuracy before each throw should prompt them to make better decisions about where to stand. However, compared to the control group, the standing positions selected by participants who were explicitly asked about their accuracy before each throw were not more optimal. We asked partici-
pants only to state their expected accuracy for each hoop from the position they had selected; it may be that this was too indirect to cause participants to actually use this information in deciding where to stand. In future studies, it may be of interest to ask more probing questions. For example, asking them to explain why they choose to stand may lead to better decisions.

We conclude that participants fail to adopt an optimal strategy despite having highly accurate information about the expected outcome of their decisions. It is possible that the participants were not aware that this information was relevant to making a decision and therefore did not pay it due attention (Gegenfurtner et al., 2011) or were unable to use the information in an effective manner (Hardman and Cowan, 2016). It is also possible that participants were distracted by other, less relevant information. Gaissmaier and Schooler (2008) suggest that some people may engage in searching for a pattern that they may be able to exploit in order to form their decisions even when there is no pattern to the task. With this in mind, it may be of interest to investigate what information people deem to be relevant to a task they are performing. Finally, our sample did not widely vary in throwing ability and in self-awareness. A sample of participants with a wider range of throwing ability, and particularly including highly skilled throwers, may provide further insight into whether confidence in the relevant information can elicit optimal decisions for maximising accuracy.

We encounter situations with multiple goals and targets frequently in daily life, from deciding which locations to monitor while driving, to deciding how to invest time and resources in various projects. It is therefore surprising that our participants are so poor at making these decisions. In daily life these situations tend to be far more complex than the situation we have constructed here, and the expected outcomes of possible decisions would be similarly complex to calculate. Heuristics are simple rules of thumb to cope with decision making in complex environments. They can often lead to near-optimal behavior using less computation and information. For example, an effective rule for intercepting a high ball (e.g. in baseball) is to keep the ball at a fixed gaze position as you run towards it (Gigerenzer and Brighton, 2009). This simple heuristic allows a fielder to behave as if they had solved the differential equations that govern the balls movement. However, in our task the optimal solution can be fully described by a simple heuristic (i.e., always stand in the center when the hoops are close, and switch to one hoop when they are far apart). Our results suggest that participants are nonetheless unsystematic in their decisions.

## Acknowledgements

This research was supported by a James S. McDonnell scholar award (A.R.H.).

## References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Chang, Y.-K., Ho, L.-A., Lu, F. J.-H., Ou, C.-C., Song, T.-F., and Gill, D. L. (2014). Self-talk and softball performance: The role of self-talk nature, motor task characteristics, and self-efficacy in novice softball players. Psychology of Sport and Exercise, 15(1):139-145.

Clarke, A. D. F. and Hunt, A. R. (2016). Failure of intuition when choosing whether to invest in a single goal or split resources between two goals. Psychological Science, 27(1):64-74.

Gaissmaier, W. and Schooler, L. J. (2008). The smart potential behind probability matching. Cognition, 109(3):416422.

Gegenfurtner, A., Lehtinen, E., and Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. Educational Psychology Review, 23(4):523-552.

Gigerenzer, G. and Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. Topics in Cognitive Science, 1(1):107-143.

Hardman, K. O. and Cowan, N. (2016). Reasoning and memory: People make varied use of the information available in working memory. Journal of experimental psychology. Learning, memory, and cognition, 42(5):700-722.

Morvan, C. and Maloney, L. T. (2012). Human visual search does not maximize the post-saccadic probability of identifying targets. PLoS Comput Biol, 8(2):e1002342.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schraw, G. and Dennison, R. S. (1994). Assessing metacognitive awareness. Contemporary educational psychology, 19(4):460-475.

Simon, H. A. (1990). Invariants of human behavior. Annual review of psychology, 41(1):1-20.

# Algebra is not like trivia: Evaluating self-assessment in an online math tutor 

Rachel A. Jansen (racheljansen @berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94720 USA<br>Anna N. Rafferty (arafferty@carleton.edu)<br>Department of Computer Science, Carleton College<br>Northfield, MN 55057 USA<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94720 USA


#### Abstract

Appraising one's own performance after a task, known as selfassessment, has been studied from a cognitive science perspective in domains such as humor, trivia, and logic. Previous studies have found that participants are systematically poor at judging their own performance, though sometimes self-assessment varies based on actual performance. We explored calibration of self-assessment on algebra problems, a domain where people have typically received explicit instruction. In this domain, we found that people do not behave as they do in other domains previously studied: they are generally well-calibrated in judging their algebra performance. This suggests that in the course of learning to solve algebra problems, people have also learned to accurately judge their performance, both absolutely and relative to others.


Keywords: self-assessment; algebra; intelligent tutor; calibration

## Introduction

Providing personalized and well-designed educational tools for online learners is a necessity. One important feature of online learning is that it is self-directed. Learners they guide their own way through the plethora of available materials (e.g., Song \& Hill, 2007). How can we design effective tools to help these kinds of learners be even better at gaining new knowledge through this medium? In order to be selfdirected, learners need to know what information they are lacking. Thus, it is important to find out what learners actually know and use this data to motivate them in an online setting. We can learn more about what learners know simply by asking them to evaluate their performance on a task after completing it, known as self-assessment. Self-assessment has been studied in both cognitive science (e.g., Dunning \& Kruger, 1999; Krueger \& Mueller, 2002) and educational domains (e.g., Bol \& Hacker, 2001). Here we seek to use the methods from cognitive science on an education-related task, focusing on online learners.

In psychological studies of self-assessment, it has been observed that people systematically misjudge how they perform relative to others. These studies have used tasks not formally taught and subsequently tested in a school setting, such as humor and logical reasoning (Dunning \& Kruger, 1999) or
trivia (Burson, Larrick, \& Klayman, 2006). Miscalibration has been observed across all of these domains.

The dependence on tasks such as trivia knowledge raises the question of whether similar patterns of perceived ability exist for domains in which people have already received a good deal of instruction. Given the increasing opportunities for people to engage in self-directed study online, we are interested in self-assessment in an instructed online setting. Based on previous research, we might expect that people will be poorly calibrated to their own performance.

We investigate this question through two online experiments. Experiment 1 replicates previous findings with an online population, specifically by using methods from the second study of Burson et al. (2006). As in the original paper, we find that people are poorly calibrated when self-assessing their performance on trivia problems. In Experiment 2, we turn to an instructed domain to see what links might exist between actual performance, perceived performance (both absolute and relative), and perceived difficulty. Specifically, we study algebraic equation solving, an area where we would expect our participants to have had much practice and instruction. In contrast to the trivia domains, we find that people are relatively accurate in their self-assessment about their algebraic equation solving abilities.

## Background

In the cognitive science literature, there is a general finding that people are miscalibrated in their performance judgments (e.g., Krueger \& Mueller, 2002). Yet there have been different interpretations of who is driving the trend of poor calibration and why. Dunning and Kruger (1999) originally explored this quandary, finding that those in the lowest quartile of performance appeared to judge themselves as performing much better than they truly did and that those in the highest quartile were more accurate in their judgments. They interpreted the poor perceived performance on the part of the lowest-scoring individuals as a metacognitive deficit - the worst performers lacked both the skills needed to correctly do the task and also to judge their performance on the task. Yet in later studies, participants have been observed to systematically misinter-
pret their performance regardless of actual score on a task (e.g., Krueger \& Mueller, 2002). While Dunning and Kruger found that calibration improved with actual performance, this is not the case in other studies such as in Burson et al. (2006).

In Burson et al. (2006), twelve trivia-like domains with varying levels of difficulty were studied. The authors found that regardless of the task and difficulty, participants at all levels of performance were equally inaccurate in judging their ability relative to others. However, they did find that for easier tasks, participants judged themselves as performing better than on more difficult tasks. This makes it appear that those with higher actual scores were more accurate in their judgments on easier tasks and that those with lower scores were more accurate in their judgments on difficult tasks.

Knowledge of one's own performance has been explored in educational contexts. In a study with graduate students in education, Bol and Hacker (2001) found that low-achieving students were less able to accurately calibrate ratings of their own performance on their final exam than high-achieving students. This is consistent with results from Dunning and Kruger (1999). However, they did not ask students to evaluate their performance relative to others. In another study, Bol, Hacker, O'Shea, and Allen (2005) observed that overt practice with self-assessment does not help increase accuracy. Yet they did see that high-achieving students are more accurate than low-achieving students in their performance predictions. They also found that higher achieving students are underconfident in their predictions while lower achieving students are overconfident.

Metacognitive skills have been found to be helpful for allowing students to improve their own learning processes (e.g. White \& Frederiksen, 2005). White and Frederiksen (2005) argue that working towards metacognitive understanding of one's own learning process motivates them to learn. This is important for online learners as well. As they are selfdirected, they need motivation to feel capable of learning on their own. In one study, White and Frederiksen (1998) found that including metacognitive training in a curriculum significantly increased low-achieving students' performance.

## Experiment 1: Trivia

First, we sought to replicate previous findings from Burson et al. (2006) that showed people were poorly calibrated in a trivia task. We aimed to confirm that the same results held in an online population. Our experiment replicates Study 2 from Burson et al. (2006). Plots (a) and (c) of Figure 1 show recreated versions of their original findings. In this study, all participants were poor at estimating their performance, regardless of true performance on a task. Burson et al. (2006) also found that difficulty had an effect on self-assessment accuracy, where estimated performance was on average lower for the more difficult domains than for the easier domains.

## Methods

Participants. A total of 40 participants (19 female, mean age $=30.9)$ in the USA were recruited from Amazon's Me-
chanical Turk and compensated $\$ 1.50$.
Materials. Materials from two of the five domains in the original study were included. Two domains were excluded based on Burson et al.'s (2006) findings that they were too difficult or too easy, resulting in floor or ceiling effects, and a final domain about the length of time pop songs remained on the charts was excluded due to inconsistent data from Billboard.com. We were thus left with two domains: college acceptance rates and dates of Nobel prizes in literature. For each domain there were two subsets of 10 questions each, one easy and one difficult. The more difficult version required participants' estimates to fall within a narrower range to be considered correct (e.g., within 5 years of the correct date for the harder version vs. within 30 years for the easier version).
Procedure. Participants responded to all four sets of questions, and the order of difficulty was counterbalanced across participants. For each subset, participants answered 10 questions about one domain with instructions stating they would get credit for an answer if it was within a certain range of the correct answer. Then, they were asked to rate their percentile performance, or how well they believed they performed relative to others on that set (out of 100), as well as how difficult it was for themselves and for other participants in the study (out of 10). Following the four sets of questions, they completed a survey about their demographics. The entire study took participants an average of 12.8 minutes.

## Results and Discussion

We performed similar analyses to Burson et al.'s to confirm that our findings were consistent (see Figure 1 (a) and (c)). For both tasks (estimates of years a Nobel prize in literature was received and of college acceptance rates), scores were much lower on the difficult versions than for the easy versions (Nobel: $M_{\text {hard }}=1.60$ vs. $M_{\text {easy }}=6.93$; College: $M_{\text {hard }}=1.63 \mathrm{vs} . M_{\text {easy }}=6.53$, all out of 10 ). A two-way analysis of variance (ANOVA) on true score with domain and difficulty as within-participant variables shows a main effect of difficulty $(F(1,156)=294.54, p<.001)$. Consistent with the difference in scores, harder trivia sets were rated as more difficult for participants than easy trivia (Nobel: $M_{\text {hard }}=9.08$ vs. $M_{\text {easy }}=8.15$; College: $M_{\text {hard }}=8.13$ vs. $M_{\text {easy }}=6.90$, all out of 10). An ANOVA on perceived difficulty for oneself shows a main effect of true difficulty $(F(1,157)=8.06$, $p<.05)$ and of domain $(F(1,157)=8.44, p<.05)$. Additionally, the average Nobel prize estimates were perceived as more difficult for the self than the college acceptance rate estimates $\left(M_{\text {nobel }}=8.6125 \mathrm{vs} . M_{\text {college }}=7.5\right)$.
Percentile estimates. Users were asked to rate their percentile estimate after completing a task, or how well they think they did relative to others on a scale of 0 to 100 . Overall, a participant's true score was weakly correlated with their percentile estimate (Pearson's $r=.17, p<.05$ ). The mean percentile estimate across all four tasks was 34.04 , which is consistent with the average found in the original study of


Figure 1: Perceived percentiles broken out by domain ( $a$ and $b$ ) and difficulty ( $c$ and d). (a) Participants' estimates of percentile by quartile of actual performance for questions about Nobel prize winners and college acceptance rates in Burson et al. (2006) and (b) in Experiment 1. 'Nobel' refers to Nobel prizes in Literature, and 'College' refers to college acceptance rates. (c) Participants' estimates of percentile by quartile of true performance for easier and more difficult tasks in Burson et al. (2006) and (d) in Experiment 1. Note that the difficult tasks in the original study included a third domain, number of weeks pop songs were on the charts, which is not included in our study. Vertical bars represent one standard error. This information was unavailable for the original study by Burson et al.
37.04. An ANOVA on percentile estimate showed a main effect of difficulty $(F(1,157)=11.95, p<.05)$ and of domain $(F(1,157)=6.74, p<.05)$. For the Nobel tasks, percentile ratings on the difficult version were lower than for the easy version $\left(M_{\text {hard }}=22.28\right.$ vs. $\left.M_{\text {easy }}=34.98\right)$. The college acceptance rate tasks showed the same pattern ( $M_{\text {hard }}=31.88$ vs. $\left.M_{\text {easy }}=44.05\right)$. On average, the percentile ratings for the Nobel tasks were lower than for the college tasks ( $M_{\text {nobel }}=$ 28.63 vs. $M_{\text {college }}=37.96$ ). As found in the original study, perceived performance was lower for more difficult tasks.

Quartiles. As in Burson et al. (2006), we divided all participants into four quartiles based on performance. As shown in Figure 1, we see very similar results to the original study - estimates of percentile performance on the test sets about dates Nobel prizes were won tended to be lower than estimates of percentile performance on the test sets about college acceptance rates. Additionally, the easier test sets were given higher percentile estimates than the more difficult ones.

Just as Burson et al. (2006) replicated Krueger and

Mueller's (2002) result that participants of all skill levels miscalibrate their performance relative to others, we observe a similar characteristic pattern in online users. On the easier tasks, participants at all skill levels are equally inaccurate in their estimates and on the difficult tasks, the highest performers do even worse than the lowest performers on judging their relative performance.

## Experiment 2: Algebra

In our next study, we aimed to compare results from previously researched trivia-based domains to a school-taught domain: algebraic equation solving. What is interesting about this domain, as opposed to others previously used in experimental psychology studies of self-assessment, is that participants have all received feedback about their performance in the past. We could thus imagine that participants might have more awareness of how well they have historically done compared to their peers and calibrate their estimates based on how much time has passed since they last solved algebraic equations.


Figure 2: Interface of Emmy's Workshop.

We made use of Emmy's Workshop (Rafferty \& Griffiths, 2015), an adaptive algebra tutor designed to glean more information about users than just number of problems solved correctly (see Figure 2). Participants enter in their work step by step when solving each problem. The goal is to determine where in problem-solving users are faltering and then to offer them personalized feedback on a skill they are struggling with (Rafferty, Jansen, \& Griffiths, 2016).

## Methods

Participants. A total of 41 participants in the USA were recruited from Amazon's Mechanical Turk and compensated $\$ 6$. They had not completed postsecondary mathematics courses beyond algebra. Two were excluded who accidentally exited the study and had to start again from the beginning. We thus had 39 participants ( 17 female, mean age $=$ 33.2 years).

Procedure. Participants first completed a survey where they rated their knowledge of algebraic equation solving and how important it is for them to know a great deal about this domain. Then they completed 24 problems in Emmy's Workshop. They received no feedback about their problem solving. Next, they estimated their performance in both absolute terms ("How many of the 24 algebraic equations you just completed do you think you answered correctly?") and in relative terms ("Think about the 24 equations you solved. Compared to other participants in this study, how good are you at solving algebraic equations? Marking $90 \%$ means you will do better than $90 \%$ of participants, marking $10 \%$ means you will do better than only $10 \%$, and marking $50 \%$ means that you will perform better than half of the participants."). They also rated how difficult the task was for them and how difficult they thought it was for others. Finally, they completed the same demographics survey as in Experiment 1, but with additional questions about their mathematics education background.

## Results and Discussion

On average, participants solved 9.28 problems correctly (out of 24). The average perceived score was 10.38 , and the average percentile estimate was 39.38 . Overall, participants accurately estimated both number correct and percentile rankings
(a)

(b)


Figure 3: Perceived scores and percentiles in Experiment 2. (a) Participants' estimates of score (out of 24) by quartile of actual performance. (b) Participants' estimates of percentile by quartile of actual performance. Vertical bars represent one standard error.
(see Figure 3). While both total score and percentile estimates are examining distinct measures of performance, we see a similar pattern. The correlations between true score and estimated score, and between true score and percentile estimates were both high, unlike in the previous study (Pearson's $r=0.66$ for both comparisons, $p<.001$ ). Algebra is a domain where people have received feedback in the past, which has trained them to know how they compare to their peers. In contrast, people have not generally practiced and received feedback about their trivia performance to the same degree. In a school-taught domain where a learner might have a better sense of how they have done in the past, they are better able to estimate their performance, unlike in the domains tested in previous cognitive studies of self-assessment.

Difficulty. On average, participants perceived the task as being easier for others than for themselves: average perceived difficulty was 8.18 for the self and 7.36 for others (out of 10). As shown in Figure 4 (a), the more someone finds the task to be difficult for themselves relative to others, the more they
underestimate their performance $(F(1,37)=8.1, p<.05$, $R^{2}=0.18$ ).

Experiment 1 included easy and difficult sets of questions in each domain. Mirroring this design would be difficult for algebraic equation solving because skills are likely to vary widely across participants. Instead, we divided the participants into two groups based on a median split of their perceived difficulty. Perceived difficulty was measured by taking perceived difficulty for the self minus perceived difficulty for other participants. The easy group perceived the task as easier for them than for others $\left(N=18, M_{\text {difficulty }}=-1.28\right)$ and the hard group perceived the task as harder for them than for others $\left(N=21, M_{\text {difficulty }}=2.52\right)$. In the easy group, scores were 10.6 on average, score estimates were 15 on average, and the mean percentile estimate was 54.83 . In the difficult group, scores were 8.14 on average, score estimates were 6.43 on average, and the mean percentile estimate was 26.14. Those with a positive perceived difficulty (who believed the task was more difficult for themselves than for others) tended to underestimate their performance, while those with a negative perceived difficulty (who believed the task was easier for themselves than for others) tended to overestimate their performance (see Figure 4 (b)). Though these results suggested that users are accurate at estimating their performance, we see that this is actually not the case - self-assessment is adjusted either positively or negatively based on perceived difficulty of the task.

There are qualitative characteristics of these data which are consistent with the findings of the first experiment - percentile estimates are lower for tasks perceived as more difficult. However, people are much better calibrated in this domain than in the trivia domains. It is not that users systematically have metacognitive deficits, but that if they perceive a difference between their own ability and that of others, then they demonstrate systematic miscalibration, either positively or negatively.

## General Discussion

In these two studies, we aimed to explore how online participants perceive their performance in an algebra setting, assuming we would discover poor calibration in participants' estimates. Interestingly, we see that people are well-calibrated in judging their algebra performance, both absolute and relative to others. Crucially, we do not see overestimation by the worst performers as observed by Dunning and Kruger (1999) and in other studies: people seem in particular to know when they are performing poorly.

## Possible Explanations

One explanation is that people have been well-trained to selfassess in school-taught domains such as math, both in terms of raw scores and occasionally with respect to others (e.g. via standardized tests and classes that are curved). Better accuracy in self-assessment tasks through training has been noted in work on superforecasters (e.g., Mellers et al., 2015). In
(a)

(b)


Figure 4: Interaction of perceived difficulty relative to others and amount of over or underestimation in Experiment 2.
(a) Plot of linear regression equation predicting amount of overestimation (measured by taking estimated number correct minus actual number correct) from perceived difficulty (measured by taking perceived difficulty for oneself minus perceived difficulty for others). (b) Participants' estimates of percentile by quartile of true performance grouped by perceived difficulty.
this body of work, a small subpopulation has demonstrated high predictive ability about international events. Members of this group exhibit a variety of good habits and have largely been able to train to be well-calibrated in their judgments. If people can be trained to make accurate judgments about the world, they can also conceivably be trained to make accurate judgments about themselves. In the domain of algebra, we seem to have trained, through feedback on performance, to link feelings after a task to true performance. This enables us to calibrate more accurately. Perhaps if we had similar kind of experience in doing trivia quizzes, then we would be better calibrated in that domain too. We can ask questions similar to those posed by Mellers et al. (2015), such as whether it is possible to transform students into top-performing algebra problem-solvers via labeling them as "high potential latebloomers" meaning capable of gaining expertise later in life.

This mindset-related intervention (e.g., Yeager \& Dweck, 2012) or other interventions may be effective at impacting a learner's self-assessment and thus metacognitive skill. This does, however, come in conflict with the results of Bol et al. (2005) who saw that practicing self-assessment did not help increase accuracy.

Effective self-directed learners are aware of what they need to learn. Training learners to accurately evaluate their ability has the potential to help them seek out necessary materials. Knowing that through training learners have the ability to properly self-diagnose in a domain means they have the opportunity to select what is necessary for them to learn. With education being increasingly made available online, self-motivated learners need to be well-calibrated to their knowledge of domains in general.

## Future Directions

We would like to further investigate what types of people are miscalibrated in their performance judgments in an online algebra setting. In light of the results presented in this paper, we will run another study with an increased sample size, primarily to see if there are gender differences in self-assessment. At present, there is a trend of high-performing women underestimating their performance in comparison to high-performing men, but an increased sample size will be necessary to judge the validity of this conclusion.

To probe further into students' perceptions of their ability, we will run a similar study asking how well students believe they perform on individual skills relevant to algebraic equation solving. Emmy's Workshop contains an inverse planning algorithm that assesses ability on six different skills such as arithmetic and distribution (Rafferty et al., 2016), so we will be able to compare actual ratings on these skills by said algorithm to a user's perceived ability on each individual skill.

Additionally, develop models of self-assessment, in a similar vein to Labutov and Studer (2016). As self-assessment involves making an inference about one's own ability based on one's performance, we can think about using Item Response Theory (IRT), a family of models commonly used by education researchers, to estimate the ability of students, both overall and on individual skills. This will help inform how perceived performance on each problem individually will predict actual performance on subsequent problems.

## Conclusion

Self-assessment has been studied in both cognitive science and educational contexts. Our experiment connects methods from the self-assessment literature to applications in education, specifically aimed at studying the self-evaluations made by online learners of varying ability and backgrounds. We find that, on average, participants solving algebraic equations are well-calibrated in their estimates of their own performance, both absolute and relative. This stands in contrast to previous work in both cognitive science and education where miscalibrations have been observed by participants of all ability levels. However, participants who perceive the task as
excessively difficult tend to underestimate their performance, marking them as a possible group to develop intervention for improving their self-assessment skills.

## Acknowledgments

This work was supported by NSF grant DRL-1420732 to Thomas Griffiths and Anna Rafferty. Thanks to Priyanka Bhoj for assisting with analyses for Experiment 1 and to Katherine Burson for sharing her study materials.

## References

Bol, L., \& Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. The Journal of Experimental Education, 69, 133-151.
Bol, L., Hacker, D. J., O'Shea, P., \& Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. The Journal of Experimental Education, 73, 269-290.
Burson, K. A., Larrick, R. P., \& Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. Journal of Personality and Social Psychology, 90, 60-77.
Dunning, D., \& Kruger, J. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77, 1121-1134.
Krueger, J., \& Mueller, R. A. (2002). Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. Journal of Personality and Social Psychology, 82, 180-188.
Labutov, I., \& Studer, C. (2016). Calibrated self-assessment. In Proceedings of the 9th international conference on educational data mining. (pp. 119-126).
Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., ... Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science, 10, 267-281.
Rafferty, A. N., \& Griffiths, T. L. (2015). Interpreting freeform equation solving. In Proceedings of the 17th international conference on artificial intelligence in education (pp. 387-397). Springer International Publishing.
Rafferty, A. N., Jansen, R. A., \& Griffiths, T. L. (2016). Using inverse planning for personalized feedback. In Proceedings of the 9th international conference on educational data mining. ( pp . 472-477).
Song, L., \& Hill, J. R. (2007). A conceptual model for understanding self-directed learning in online environments. Journal of Interactive Online Learning, 6, 27-42.
White, B., \& Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. Cognition and Instruction, 16, 3-118.
White, B., \& Frederiksen, J. (2005). A theoretical framework and approach for fostering metacognitive development. Educational Psychologist, 40, 211-223.
Yeager, D. S., \& Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. Educational Psychologist, 47, 302-314.

# Reasoning ability predicts irrational worldview but not conspiracy belief 

Jan Jastrzębski (jastrz.jan@gmail.com)<br>Institute of Psychology, Jagiellonian University<br>Grodzka 52, 31-044 Krakow, Poland

Adam Chuderski (adam.chuderski@uj.edu.pl)<br>Institute of Philosophy, Jagiellonian University<br>Grodzka 52, 31-044 Krakow, Poland


#### Abstract

Previous research showed that individual tendency to believe in conspiracy theories is related to numerous social, personality, and cognitive variables. Moreover, such a tendency may reflect a broader trait for epistemic irrationality, which drives other pseudoscientific and paranormal beliefs. However, the relationship between conspiracy belief and reasoning ability (fluid intelligence; Gf) was not sufficiently studied to date, even though Gf level strongly influence the way in which individuals think and reason. Using confirmatory factor analysis, we found the robust link between conspiracy belief and other irrational beliefs. All those irrational beliefs were also substantially related to the closeminded cognitive style. However, even though Gf significantly predicted other irrational beliefs, it explained less than $2 \%$ of variance in conspiracy belief. This result suggests that effective reasoning cannot prevent even highly intelligent people from endorsing conspiracy theories.


Keywords: rationality; intelligence; conspiracy theory; paranormal beliefs; pseudoscience;

## Introduction

Conspiracy theory is an explanation of a significant event, like the sudden death of famous person, terrorists attack, or catastrophe, as resulting from some secret plot made by a powerful organization or a group of powerful individuals. Although, in principle, such theories may be true (e.g., the Watergate scandal), usually they are insufficiently supported by facts, disregarded by experts, and based on pseudoscientific assumptions.

Importantly, the belief in conspiracy theories (henceforth, conspiracy belief) pertains not only to advocates of extreme ideologies or to paranoid and delusional individuals, but is prevalent in diverse cultures and societies (Raab, Ortlieb, Auer, Guthmann, \& Carbon, 2013). Although to date most of research on the topic was conducted in Western countries, some studies showed that conspiracy belief is a widespread phenomenon among people all over the world (e.g. Raab et.al., 2013; Goertzel, 1994; Swami et al., 2011; Bruder, Haffke, Neave, Nouripanah, \& Imhoff, 2013).

Certain people are more likely to hold conspiracy belief than others. What is important, this tendency may be a part of more general mindset, worldview or mentality (Goertzel, 1994; Dagnall, Drinkwater, Parker, Denovan, \& Parton, 2015; Imhoff \& Bruder, 2014). People who believe in one conspiracy theory are also more likely to believe in another one, even if the theories are unrelated (Swami, Chamorro-

Premuzic, \& Furnham, 2010), contradictory (Wood, Douglas, \& Sutton, 2012), or the second one is fictional and encountered for the first time (Swami et al., 2011). Such a kind of conspiratorial mentality was associated with numerous socio-psychological variables, including anomie, powerlessness, feeling of meaninglessness, distrust, authoritarianism, political cynicism, low self-esteem, and schizotypy (Goertzel, 1994; Abalakina-Paap, Stephan, Craig, \& Gregory, 1999; Swami et al., 2010; Swami et al., 2011). Moreover, conspiracy belief is closely related to other epistemically dubious beliefs, like paranormal and pseudoscientific claims and theories (Lobato, Mendoza, Sims, \& Chin, 2014). Altogether, paranormal, pseudoscientific, and conspiracy beliefs may result from one and the same irrational worldview/mindset.

The problem that we investigated was how are various instances of irrational thought, and particularly conspiracy belief, related to reasoning ability (fluid intelligence; Gf), which is defined as the ability to solve novel problems by means of abstract reasoning. As Gf strongly predicts many socio-psychological variables (see Deary, 2012), can Gf also predict individual tendency for irrational beliefs? At least intuitively, it seemed reasonable to expect that more intelligent people, because of their more powerful reasoning, would be more sceptical toward dubious, unsupported beliefs, including conspiracy theories. Besides intuition, numerous premises can be found in existing literature.

First, positive correlations between intuitive thinking style and conspiracy belief were reported. Also, experimentally induced willingness to engage in analytic reasoning reduced this belief (Swami, Voracek, Stieger, Tran, \& Furnham, 2014). Thus, a disposition for reflective thinking may help to embrace more sceptical stance towards irrational claims and theories, and may affect one's worldview even counter to cultural factors (Pennycook, Fugelsang, \& Koehler, 2015). Moreover, tendency for analytic thinking may prevent people from relying on intuitions and "gut feelings" that often lead to cognitive biases and heuristics, which may, at least to some extent, drive conspiracy belief. For example, Clarke (2002) argued that such a belief may stem from attribution bias, which consists of overestimating the influence of personal factors, and ascribing responsibility primarily to agents, instead of explaining events in terms of situational factors and coincidence. Likewise, it was argued that this belief may be related to representativeness heuristic, which leads people to seek explanations that
possess salience proportional to the very significance of events (Leman \& Cinnirella, 2007). Since major events need major explanations, people may see ordinary causes of great-impact events as unsatisfying and thus unlikely, and thus may embrace conspiracy theories instead. Finally, Brotherton and French (2014) showed the people displaying conspiratorial mentality to be more prone to conjunction fallacy, which is a reasoning error consisting of assessing the probability of two co-occurring events as being more likely than the joint probability that these events will occur alone. As the tendency for analytical thinking is at least moderately related to intelligence (Pennycook et al., 2015), a negative relationship between intelligence and conspiracy belief may also exist.

On the other hand, the relationship between intelligence and biased/irrational thinking is not straightforward. Although some biases might be attenuated by higher reasoning ability, some may not be related to intelligence at all (Stanovich \& West, 2008). High reasoning ability may not be enough to prevent people from embracing dubious theories. A research program aimed at understanding the relationship between intelligence and rationality, started by Stanovich, put emphasis on the need to distinguish between these two mental qualities. Although there can be a positive relationship between the two, what is essential for rationality may be such thinking dispositions as the willingness to think reflectively and open-mindedly. Consistently, the relationship between intelligence and irrational beliefs might be at least partially mediated by cognitive style. Although such a possible mediation so far has never been studied in the context of conspiracy theories, some supporting evidence comes from studies on paranormal and religious beliefs (Pennycook, 2014).

Finally, studies showed a moderate negative link between paranormal/pseudoscientific beliefs and intelligence (e.g., Rindermann, Falkenhayn, \& Baumeister, 2014), but the relationship between intelligence and conspiracy belief in conspiracies has not been studied sufficiently enough.

Only one study to date examined this relationship (Swami et al., 2011). First, it showed a negative, though weak, correlation between conspiracy belief and self-assessed intelligence. However, this result does not seem reliable, as this measure of intelligence had low validity. Second, the study reported weak negative correlation of conspiracy belief and crystallized intelligence (Gc) - the ability to use acquired experience and knowledge. However, Gf may be even more important for the rejection of conspiracy theories than Gc, because higher Gf levels allow more effective processing of relations among objects, events, and facts (Chuderski, 2014). Such relations can be used to create counterexamples in a reasoning process (Johnson-Laird, 2006).

## Study

The main goal of the study was to fill in the gap in existing data on the relationship between Gf and conspiracy belief. We expected reasoning ability to at least weakly predict
belief in conspiracy theories. Furthermore, we intended to replicate the results that show moderate positive correlations between different kinds of dubious beliefs: conspiracy, paranormal, as well as pseudoscientific ones (Labato et al., 2014; Brotherton \& French, 2014; Swami et al., 2011).

In order to test the strength of the conspiracy-reasoning link as well as to examine the strength of relationship between conspiracy and paranormal/pseudoscientific belief we used multiple measures of each belief, as well as we applied latent variable modelling by means of confirmatory factor analysis (CFA). Applying more than one measure of each construct. and calculating latent variables enables a more valid and reliable measurement of the constructs in question as well as the relationships between them, as compared to using single measures (see Kline, 1998).

We applied two measures of conspiracy belief to a large sample of Polish adults. Because many conspiracy theories are strongly culture-specific, one scale was created to measure belief in a particular conspiracy theory pertaining to political situation in Poland: the theory about catastrophe in Smolensk. The Smolensk conspiracy is probably the most distinctive case of conspiracy theory in the Polish society, and it is similar to conspiratorial themes that are vivid in other societies (e.g., the death of President Kennedy and Princess Diana). However, it is possible that some specific factors may play a crucial role for the Smolensk conspiracy (e.g. most of its advocates are right-wing/conservative), which might not drive other conspiracy theories, and which thus may bias the relationship between reasoning ability and conspiracy belief. To avoid such a bias, we also applied a measure of general conspiratorial beliefs and attitudes - The Generic Conspiracist Beliefs Scale (GCB) (Brotherton et al., 2014). Importantly, GCB does not concern any particular conspiracy theory, but deals with common conspiratorial themes (e.g. governments totally controling the information flow), that enables broader generalization of the results. Also, we used one questionnaire to measure paranormal beliefs, and another for pseudoscientific beliefs.

In addition, we measured open-minded cognitive style, understood as mental flexibility and openness toward the alternative views, perspectives, and counter-evidence. To do so we applied two questionnaires: NEO-openness subscale and open-minded thinking scale. Open-minded thinking was previously shown to be negatively (though rather weakly) related to conspiracy belief (Swami et al., 2014). Finally, Gf was measured with two visuospatial tests and one numerical test that involved abstract reasoning.

## Method

## Participants

A total of 318 voluntary participants ( 218 women, 100 men) were recruited via ads in publicly accessible websites. The participants were paid an equivalent of 20 euros in Polish currency. The mean age was 24.4 years ( $S D=6.02$, range 18-45). Four participants did not complete all the applied questionnaires and were excluded from the analysis.

## Materials

## Smolensk Conspiracy Scale

Conspiracy theories are cultural phenomena. Studying particular conspiracy theories requires that the participants are familiar with them and their cultural context. Our choice of a theme for well-known conspiracy theory was the 2010 catastrophe in Smolensk (the Russian Federation), in which the Polish President's plane crashed, and all of the 96 crew and passengers, including President Lech Kaczynski, died. The Smolensk catastrophe was judged by official aviation experts (PCINAA, 2011) to result from the pilot's error as well as the improper organization of the flight. The crash had a specific political context: The death of President Kaczynski, who travelled across Russian territory, in order to commemorate Polish officers killed by Soviets during WWII, despite his tense political relations with the Russian government of President Vladimir Putin. Furthermore, President Kaczynski and his conservative camp strongly opposed the Polish government and its supporting liberal party, while the presidential campaign in Poland was about to start. All of this made an excellent context for various accusations and plot hypotheses, even though the explanation of the catastrophe is straightforward. Consequently, five years after the catastrophe, a public opinion survey (CBOS, 2015) showed that about $30 \%$ members of the Polish society considered the hypothesis of assassination of Lech Kaczynski plausible (among them 8\% were convinced it was true). Thus, the Smolensk catastrophe made a crucial and interesting case of conspiracy belief (henceforth we call it the Smolensk conspiracy). So, a twelve-item questionnaire was developed, with seven items measuring belief in the Smolensk conspiracy, and five reverse-scored items probing belief in the official explanation of the catastrophe.

## The Generic Conspiracist Beliefs scale

We used the validated 15 -item scale of Brotherton et al. (2014) to measure the general tendency for conspiracy belief. The scale covers general conspiratorial assumptions such as beliefs in prevalent government misconduct, secret groups exerting the control over global events, dangers to personal health and liberties (e.g., the mind control experiments), extraterrestrial cover-up, and the full censorship over information. The sample item was "The governments are involved in the murder of innocent citizens and/or well-known public figures, and keep this a secret".

## Pseudoscientific Belief Scale

We created an 18 -item questionnaire to measure pseudoscience belief and disapproval of scientific knowledge. The items covered range of topics (medicine/health, natural science, evolution, psychology, sexuality), and were mixed with 9 filler items dealing with general scientific knowledge. Sample test items were "Mercury in vaccines may increase probability of acquiring autism among small children" and "Crystals possess qualities which protect against negative influence of electromagnetic radiation".

## Paranormal Belief Scale

Our measure of paranormal belief was based on Revised Paranormal Belief Scale (Tobacyk, 2004). We removed four items concerning religious belief, as we applied a separate religious beliefs questionnaire in the session (not analysed in the present study). We removed another three items concerning extra-ordinary life forms (e.g. Loch Ness monster), as being outdated and possibly unfamiliar to our participants. The final version contained 20 items such as "In some cases it is possible to communicate with the dead".

## Fluid intelligence tests

We applied three Gf tests. The classic Gf test - Raven's Advanced Matrices (Raven, Court, \& Raven, 1983), as well as Figural Analogies (Chuderski \& Necka, 2012), were administered in shortened versions (18 items each). Each of the two tests was composed of odd numbered items from respective standard 36 -item versions. Their administration time was half of the standard one ( 20 and 15 minutes respectively). The third test was Number series, in which the task was to find the rule according to which the number sequence or the array is constructed, and to complete the sequence/array with the missing number. Participants were given 18 minutes to solve the 18 number series problems.

## Open-mindedness cognitive style questionnaires

The first questionnaire measuring open-minded thinking included 14 items from Actively Open-minded Thinking scale (Stanovich \& West, 2007), selected on the basis of our previous data. All 14 items were scored in such a way that higher scores represented a larger tendency toward rigid, dogmatic, categorical thinking, as well as the trend for sticking to one's beliefs even in the face of counterevidence (e.g., "Changing your mind is a sign of weakness"). The total score on the scale was reversed, so that higher total scores indicated more open-minded, flexible thinking. The second questionnaire was an 12-item openness to experience subscale of the Polish adaptation of the NEO-Five Factor Inventory (Costa \& McCrae, 1992).

In all of the questionnaires except the cognitive style measures, participants judged whether the given statements are true or false using a seven-point scale $(0=$ false for sure, $3=$ uncertain, $6=$ true for sure $)$. Five-point $(1=$ definitely disagree, $5=$ definitely agree) and a four-point scale were used in the Open-minded Thinking and the NEO-openness scales, respectively.

## Procedure

Participants were tested in a psychological laboratory, in groups of six participants on average. The Gf tests were applied in the fixed order (Raven APM, Figural Analogy Test, Number series test). All the questionnaires were completed via computers at the end of the study session. At the course of the session participants completed other tasks (working memory and cognitive control tests, religiosity questionnaires, etc.) unrelated to the topic of this study.

## Results

All the measures applied had at least satisfactory internal consistency (Cronbach's alphas > .71), including Smolensk conspiracy and GCB scales (Cronbach's alpha $=.88$, and .94 , respectively), and all of them fitted well the normal distribution (max. skew $=-0.28$, max. kurtosis $=-0.82$ ).

Firstly, the endorsement of the most extreme form of conspiracy theory, the assassination theory, was examined. Answers on the respective item ("The cause of the catastrophe was an assassination.") of the Smolensk Conspiracy scale showed that about $10 \%(N=32)$ of the participants considered it a possible option (answered "it is probably true"), $5 \%(N=16)$ answered "it is true", and $4 \%$ ( $N=12$ ) answered "it is true for sure". Thus, the support for Smolensk conspiracy in our sample was rather low.

The CFA model (Figure 1) correlated four latent variables: Conspiracy Belief (loading Smolensk Conspiracy scale and GCB), Irrational Belief (Paranormal and Pseudoscience), Gf (the three reasoning tests), and Openmindedness (NEO-openness and Open-minded Thinking).

The model fit was assessed with three indices (see Kline, 1998): $\chi^{2}$ statistic (its value divided by the number of degrees of freedom should not exceed $\chi^{2} / d f=2.0$ ), Bentler's comparative fit index (CFI should exceed .92), and the root mean square error of approximation (RMSEA should be less than .08 ). The fit of the model was good: $\chi^{2}(21)=36.01$, $\mathrm{CFI}=.981, \mathrm{RMSEA}=.047(90 \% \mathrm{CI}=[.017, .074])$.

All factor loadings (see Table 1) showed satisfactory validity of the applied measures, except for NEO-openness. Importantly, belief in Smolensk conspiracy was substantially related to GCB. Thus, Smolensk Conspiracy scale seems to be a valid measure of conspiracy belief.
In line with our expectations, the correlation between the Conspiracy Belief and Irrational Belief factors was strong, $r=.72, p<.001$. However, the negative link between the Conspiracy Belief and Gf was very weak ( $r=-.13$ ), and despite our large sample it was not statistically significant ( $p=.08$ ). Thus, Gf predicted only a negligible amount of variance ( $2 \%$ ) in conspiracy belief. However, as expected, there was a negative correlation between Gf and Irrational Belief, $r=-.31, p<.001$. In addition, the open-minded

Table 1. Factor loadings from the CFA model (all $p \mathrm{~s}<.001$ )

| Latent variable | Measure | Factor <br> loading |
| :---: | :--- | :---: |
| Gf | Raven Matrices | 0.84 |
|  | Analogies | 0.78 |
|  | Numbers | 0.66 |
| Irrational | Pseudoscience | 0.84 |
| Belief | Paranormal | 0.74 |
| Conspiracy | Generic conpiracist beliefs | 0.87 |
| Belief | Smolensk conspiracy | 0.50 |
|  | NEO-openness | 0.36 |
| Open-mindedness | Open-minded thinking | 0.96 |

cognitive style showed the substantial negative correlation with Conspiracy Belief and Irrational Beliefs. Thus cognitive style was a much stronger predictor of conspiracy and irrational beliefs than Gf.

## Discussion

We aimed to test whether conspiracy belief weakens with an increased reasoning ability (Gf). Contrary to our expectations, results showed that it virtually did not; Gf explained less than $2 \%$ of variance in conspiracy belief, and despite our large sample the link was not significant. On the other hand, Gf predicted about $9 \%$ of variance in paranormal and pseudoscience belief. Although the relationship was weaker than in previous studies, it is in line with these studies (Rindermann, Falkenhayn, \& Baumeister, 2014). Moreover, irrational beliefs shared half of variance with conspiracy belief, also replicating similar findings (Lobato et al., 2014; Brotherton \& French, 2014; Swami et al., 2011).
The robust relationship between conspiratorial, paranormal, and pseudoscientific beliefs suggests that they rely on a common underlying mindset/worldview, which reflects the tendency to believe in irrational, epistemically dubious claims and theories. There probably are specific social and cognitive factors (e.g. anomie, political cynism, distrust, radicalism) that seem to induce the conspiratorial mindset, and, to a lesser extent, the other kinds of dubious beliefs. However, the general tendency to believe in the irrational most likely stems primarily from interrelated personality traits and thinking dispositions, such as intuitive thinking and close-mindedness (the latter shown by the present data).
If so, why is conspiracy belief unrelated to Gf, as compared to paranormal and pseudoscientific beliefs? First, although more intelligent people more frequently hold to proper scientific explanations of facts (what makes them less likely to believe in pseudoscience), most of conspiracies (also the Smolensk conspiracy) needn't be inconsistent with the body of scientific knowledge (though are unsupported by facts). Plots, evil politicians, and secret organisations undoubtedly exist, but usually they are not the reasonable


Figure 1. Correlations between four latent variables in the CFA model. All correlation are significant at $p \mathrm{~s}<.001$, except for the one presented in the dashed line.
explanations of complex phenomena. Thus, even though intelligent people may easily suppress their intuitions favouring paranormal/unscientific phenomena as being unlikely and contradicting the general scientific knowledge, they may let their intuitions about political/social issues develop more freely.

Second, endorsing conspiracy theories may be seen as a process of motivated reasoning (Kunda et al., 1990; Saunders, State, \& Farhart, 2016), which is a kind of biased reasoning directed by motivation to arrive at the desired conclusions. In context of conspiracy theories, this process may satisfy the ideological and psychological needs such as the loyalty toward ideological groups (Saunders, State, \& Farhart, 2016). Conspiracy belief seems to be strongly motivated personally as well as engaging, as it touches the basic political and social opinions and values. Thus, people may have stronger motivation to rationalize their conspiracy beliefs, comparing to paranormal or pseudoscientific beliefs. Importantly, myside/confirmation bias - the tendency to evaluate and provide arguments in a manner biased towards our own views - is basically unrelated to intelligence (Stanovich, West, \& Toplak, 2013). Also, the more subjectively important the issue, the more strongly motivated the reasoning process becomes. Thus, intelligent and curious individuals may perform better at finding quality arguments supporting their worldview and prior beliefs, but they are not more inclined to objectively consider counterevidence and alternative perspectives, especially in cases of highly engaging issues. Moreover, more politically knowledgeable individuals may be even more likely to embrace conspiracy theories than the less knowledgeable ones (Saunders et al., 2016). Similarly, although providing relevant scientific information may change people's opinion on global warming (Ranney \& Clark, 2016), general scientific literacy/numeracy is unrelated to differences in opinion on global warming risk, but is related to a greater opinion polarization on the issue (Kahan et al., 2012). The opinion is instead well predicted by different values sets/worldviews. Also, at least in some cases, a high level of reasoning ability/reflective thinking may actually lead to more motivated reasoning (Kahan, 2013), and thus leading to greater polarization of prior beliefs, rather than alleviating their influence. To sum up, whether an individual embraces a conspiracy theory may be primarily dependent on his/her prior worldview and mindset, which directs the reasoning process to conclusions consistent with this worldview, and high intelligence may rather serve this process instead of hindering it. Consequently, sheer high intelligence may not be enough to prevent people from endorsing dubious conspiracy theories. Even some highly intelligent individuals may believe in conspiracies, as did some of our participants who scored really high on intelligence tests, but regardless of their high ability believed in assassination as the major cause of the Smolensk catastrophe.
More generally, our results serve as another example that intelligence and rationality should be treated as dissociable constructs (see Stanovich et al., 2013). Although some
irrationality indices (e.g., paranormal and pseudoscience) may be moderately related to intelligence, other may be weakly related, as probably is in the case of at least some conspiracy beliefs.
On the other hand, we should notice that conspiracy theories are not homogenous phenomena, and are also not irrational by definition. In some cases lack of healthy skepticism toward official information from seemingly reliable sources may be as harmful as unreflective belief in dubious conspiracy theories, and thus we do not think the less ones score on a conspiracy questionnaire the better. However, confidence in questionable conspiracy beliefs may be interpreted as irrational, as is confidence in dubious paranormal or pseudoscientific beliefs. Secondly, although we think that use of two measures of conspiracy belief dealing with different conspiratorial attitudes and beliefs provides a good measure of general conspiratorial mindset, we cannot exclude that the relation between belief in conspiracies and cognitive dispositions might be different in cases of some particular conspiracy theories.
In conclusion, our results make an important contribution to the conspiracy and rationality research, by showing conspiracy belief to be virtually unrelated to reasoning ability. Given the present data as well as numerous other cognitive, social, and personality variables that play role in prevalence of such complex socio-cultural phenomena as conspiracy theories, it seems that intelligence alone cannot prevent people from believing in conspiracy theories. Additionally, they provide more evidence for the strong conspiracy-irrationality relationship, supporting the view that the individual tendency to think in an irrational/ conspiratorial way may constitute a stable and important personality trait/cognitive style.

## Acknowledgments

This work was sponsored by the Foundation for Polish Science (the MISTRZ project sponsored to A. Chuderski).

## References

Abalakina-Paap, M., Stephan, W. G., Craig, T., \& Gregory, W. L. (1999). Beliefs in conspiracies. Political Psychology, 20(03), 637-647.
Brotherton, R., \& French, C. C. (2014). Belief in conspiracy theories and susceptibility to the conjunction fallacy. Applied Cognitive Psychology, 28, 238-248.
Bruder, M., Haffke, P., Neave, N., Nouripanah, N., \& Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy Mentality Questionnaire. Frontiers in Psychology, 4, 4: 225.
Cacioppo, J. T., Petty, R. E., Feinstein, J., \& Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological Bulletin, 119, 197-253.

CBOS (2015) Survey report. Before the fifth anniversary of the Smolensk catastrophe. Available at: http://cbos.pl/SPISKOM.POL/2015/K_049_15.PDF
Chuderski, A. (2014). The relational integration task explains fluid reasoning above and beyond other working memory tasks. Memory \& Cognition, 42, 448463.

Clarke, S. (2002). Conspiracy theories and conspiracy theorizing. Philosophy of the Social Sciences, 32, 131150.

Dagnall, N., Drinkwater, K., Parker, A., Denovan, A., \& Parton, M. (2015). Conspiracy theory and cognitive style: a worldview. Frontiers in Psychology, 6(February), 1-9.
Goertzel, T. (1994). Belief in conspiracy theories. Political Psychology, 15, 731-742.
Imhoff, R., \& Bruder, M. (2014). Speaking (Un-)truth to power: Conspiracy mentality as a generalised political attitude. European Journal of Personality, 28, 25-43.
Johnson-Laird, P. N. (2006). How we reason? Oxford: Oxford University Press.
Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. Judgment and Decision Making, 8(4), 407-424.
Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., \& Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. Nature Climate Change, 2(10), 732-735.
Kunda, Z., Dunning, D., Jones, E., Jussim, L., Miller, D., Nisbett, R., \& Petty, R. (1990). The case for motivated reasoning, Psychological Bulletin, 108(3), 480-498.
Leman, P. J., \& Cinnirella, M. (2007). A major event has a major cause: Evidence for the role of heuristics in reasoning about conspiracy theories. Social Psychological Review, 9, 18-28.
Lobato, E., Mendoza, J., Sims, V., \& Chin, M. (2014). Examining the relationship between conspiracy theories, paranormal beliefs, and pseudoscience acceptance among a university population. Applied Cognitive Psychology, 625, 617-625.
Newheiser, A. K., Farias, M., \& Tausch, N. (2011). The functional nature of conspiracy beliefs: Examining the underpinnings of belief in the Da Vinci Code conspiracy. Personality and Individual Differences, 51, 1007-1011.
Oberauer, K., Süß, H.-M., Wilhelm, O., \& Wittman, W. W. (2008). Which working memory functions predict intelligence? Intelligence, 36, 641-652.
Pennycook, G. (2014). Evidence that analytic cognitive style influences religious belief: Comment on Razmyar and Reeve (2013). Intelligence, 43, 21-26.
Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., \& Fugelsang, J. a. (2012). Analytic cognitive style predicts religious and paranormal belief. Cognition, 123(3), 335346.

Pennycook, G., Fugelsang, J. A., \& Koehler, D. J. (2015). Everyday consequences of analytic thinking. Current Directions in Psychological Science, 24, 425-432.
Polish Committee for Investigation of National Aviation Accidents (2011). Final report. Available at: http://komisja.smolensk.gov.pl/kbw/komunikaty/8877,Fi nal-report-of-the-Committee-for-Investigation-of-National-Aviation-Accidents-i.html
Raab, M. H., Ortlieb, S. A., Auer, N., Guthmann, K., \& Carbon, C. C. (2013). Thirty shades of truth: Conspiracy theories as stories of individuation, not of pathological delusion. Frontiers in Psychology, 4(JUL), 1-10.
Ranney, M. A., \& Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. Topics in Cognitive Science, 8, 49-75.
Rindermann, H., Falkenhayn, L., \& Baumeister, A. E. E. (2014). Cognitive ability and epistemic rationality: A study in Nigeria and Germany. Intelligence, 47, 23-33.
Saunders, K. L., State, C., \& Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust, American Journal of Political Science, 60(4), 824-844.
Stanovich, K. E., \& West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. Journal of Personality and Social Psychology, 94(4), 672-695.
Stanovich, K. E., West, R. F., \& Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. Current Directions in Psychological Science, 22, 259-264.
Swami, V., Chamorro-Premuzic, T., \& Furnham, A. (2010). Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11conspiracist beliefs. Applied Cognitive Psychology, 24, 749-761.
Swami, V., Coles, R., Stieger, S., Pietschnig, J., Furnham, A., Rehim, S., \& Voracek, M. (2011). Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. British Journal of Psychology, 102, 443-463.
Swami, V., Voracek, M., Stieger, S., Tran, U. S., \& Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. Cognition, 133(3), 572-585.
Tobacyk, J. J. (2004). A Revised Paranormal Belief Scale. International Journal of Transpersonal Studies, 23, 94 98.

Wood, M. J., Douglas, K.M., \& Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. Social Psychological and Personality Science, 3, 767773.

# The Motor System Does Not Use a Curvilinear Impetus Belief: Folk Physics and Embodied Cognition 

Jay Jennings (jayjennings@cmail.carleton.ca) Jim Davies (jim@jimdavies.org)

Institute of Cognitive Science, 1125 Colonel By Drive<br>Ottawa, ON K1S5B6 Canada


#### Abstract

Previous work shows that people often believe, contrary to actual physics, that objects travelling in a curved path through a tube will continue to travel in a curved path after exiting the tube. In the present study, previous work was replicated, but accuracy increased in a new condition in which people were asked to catch an actual ball emerging from a tube. That is, in this case there is a discrepancy between how we believe the world works, and how our motor system responds to events in the world. This finding supports the theory that the perception and action systems of the brain use different methods to predict how things move in the world, and that the abstract reasoning systems used to explain how the world works are often in conflict with the action systems.


## Introduction

People are not only brains, they are bodies too, and these bodies experience the world and the laws that govern its physics every day. Both children and adults make similar mistakes when verbally describing how they would crawl on their hands and knees - even after they have just physically crawled (Piaget, 1976). With something so fundamental to our everyday experience, we should easily be able to describe its physics and procedures. The systematic discrepancies observed between the ability to do a physical task and the inability to accurately describe or depict it has been studied under the umbrella term of folk or "naïve physics."

Since Piaget (1976) detailed several of these phenomena, one of the most notable instances of this effect is the curvilinear impetus belief: the incorrect assumption that an object travelling in a circular motion will continue this curved path upon exiting a spiral. In previous studies the average percentage of curved lines predicted had been around $36 \%$, and $49 \%$ for participants that had no formal physics education (McCloskey, Caramazza, \& Green, 1980; McCloskey \& Kohl, 1983; Cook \& Breedin, 1994; Freyd \& Jones, 1994; Kaiser, Jonides, \& Alexander, 1986).

Some have suggested that the abstractness of the task affected accuracy (Freyd \& Jones, 1994; McCloskey \& Kohl, 1983). For instance, when people are presented with the spiral problem on paper, but are told it is a garden hose and water as opposed to a ball and tube, they accurately predict the straight trajectory of the water, but continue to make incorrect predictions about the ball (Catrambone, Jones, Jonides \& Seifert, 1995). Freyd and Jones (1994) theorized that in order to understand the abstract diagrams
presented on paper, the participants may be generating abstract theories that are separate from their experiences in the real world, or on patterns of motion observed in living objects as opposed to inanimate objects. Thus, even though the problem is essentially the same, a physical example that has likely been experienced by the participants before (the garden hose), generates more correct answers. Therefore, previous experience from seeing an object in motion may be recruited to solve abstract problems. If this is true, then it would indicate that people can make good predictions about situations they have seen, but have trouble transferring that skill to new domains.

To test whether observing motion could influence accuracy, McCloskey and Kohl (1983) presented participants with 3 conditions designed to alter the degree of motion in a curvilinear task. Participants viewed three training conditions: no motion (a paper and pencil diagram), dynamic rotation (where the ball on-screen simply orbited around the circumference), and dynamic trajectory (where the ball left the orbit). The trajectory condition produced both correct and curvilinear trajectories. Surprisingly, there was no significant difference between these groups. The perceptual experiences failed to facilitate accurate trajectories. The authors argued that visual information was not sufficiently embodied to produce accurate responses.

This led to postulating whether physical touch improved the accuracy of participants' responses. McCloskey and Kohl (1983) asked participants to physically push a puck through a slightly curved ' $C$ ' shaped path. The task could be accomplished if a straight-line method was used. The task was designed to add a motor component to test whether participants used curvilinear strategies to try to complete the exercise, which would result in a failure to complete the task. Still, 25 percent of the participants demonstrated some curvilinear impetus belief, and participants who failed to pass the puck through the curved area with the correct straight-line method demonstrated on a post-test questionnaire that an alternative method might have been curvilinear. Previous research, such as work on the Ebbinghaus Illusion, showed that the physical grasping of two identically sized objects was completed accurately by the motor system, while the perceptual system simultaneously perceived one as substantially bigger. This would suggest that the action system of the brain would not be fooled by perceptual illusions (Haffenden, Schiff \& Goodale, 2001). However, curvilinear impetus types of
problems are not visual illusions, but internal misconceptions involving object motion, which seem to form the basis of our understanding of object physics.

Reliance on curvilinear impetus concepts today may seem counterintuitive, but in the past they were part of a dominant theory of physics in Medieval Europe. It was believed that moving objects were imbued with their own force, or impetus, that compelled them to behave in a certain way (Kozhevnikov \& Hegarty, 2001). But the fact that most present-day people are ignorant of this theory suggests that these impetus beliefs are systematic and consistent between human beings. Impetus theories may provide a way of conceptualizing the world when no analogy or previous experience is accessible - a default physics in want of physical experience (Kozhevnikov \& Hegarty, 2001). If this is true, then these errors should persist even when people have abstract, rule-based physics knowledge. Kozhevnikov and Hegarty (2001) tested this by giving physicists and non-physicists a spiral diagram and asking them to draw the trajectory. In one condition, participants were given as much time as they needed to answer, and in the other condition they had to answer as quickly as possible. Interestingly, both groups fell back on curvilinear impetus beliefs in the time-pressured condition. Thus, even people who possess expert physics knowledge still fall back on impetus theory when under time constraints.

There is now some evidence that impetus theories, specifically curvilinear impetus theories, are part of a default physics. However, many experiments tested impetus beliefs using abstract stimuli, and were testing the production or prediction of trajectories rather than how people responded to them. For instance, people do not regularly predict where water will exit from a coiled hose; however, they do respond to it physically and accurately without effort.

Many of the experiments on curvilinear impetus belief overlook the way our motor system, which tends to respond to the environment accurately, predicts the motion of objects in the world (Oberle, McBeath, Madigan \& Sugar, 2005; Zago \& Lacquaniti, 2005). The ability of the motor system to respond to moving objects may be an entirely embodied phenomena, one that has been explained with complex abstract terms, but is easily accomplished with simple, embodied perception-action rules (Wilson \& Golonka, 2013).

For instance, complex motion problems like the outfielder problem can be easily explained with perception-action rules, rather than with abstract, representational explanations. The outfielder problem asks the question, 'how does an outfielder know where to be to catch a fly ball?' The traditional cognitive approach would suggest reasoning based on a model or rules-for example, that the outfielder calculates initial speed and angle, and uses laws of projectile motion to predict where the ball will land and moves there in a straight line. The embodied approach, on the other hand, would state that as the outfielder moves, they utilize a combination of their motion through space as
well as the ball's motion through space. The embodied approach claims that the outfielder simply follows two elementary, perceptually based rules; first, move in a curved path that mirrors the path of the ball, so that it appears that it is moving in a straight line, and second, match speed so that it appears that the ball is moving at a constant velocity (Wilson \& Golonka, 2013). Furthermore, most outfielders run in curved lines. There have been many elaborations on these simple embodied rules to catch a fly ball that rely entirely on reactionary rules (Chapman, 1968; McLeod, Reed, Dienes, 2006; Tresilian, 1995). To further add to the evidence that the motor system responds in real time to the environment, rather than following a previously simulated prediction, outfielders can catch fly balls in virtual reality whose paths actually defy physics, where it is impossible to predict trajectories (Fink, Foo \& Warren, 2009).

The apparent discrepancy between the motor system responding to a trajectory and the abstract prediction or production of one could explain why multiple studies report surprisingly high prediction errors in schematized spiral problems. This effect may be due to there being separate pathways for perception and action of visual stimuli (Goodale \& Milner, 1992; Haffenden, Schiff \& Goodale, 2001). For example, when a ball is dropped from a certain height, regardless of the weight, participants will react accordingly and catch the ball, and their implicit motor knowledge will even account for the mass by demonstrating stronger muscle activity in anticipation of a heavier object (Oberle et al, 2005; Zago \& Lacquaniti, 2005). Interestingly, the same participants will make incorrect Aristotelian assumptions about which ball will hit the ground first when posed the question abstractly, but will demonstrate implicit motor knowledge of Newton's Laws of Dynamics when responding physically. These results demonstrate that the motor system is responsible for accurately responding and accomplishing a task that the same participants are unable to accurately describe conceptually.

This discrepancy between being able to do something and being able to accurately describe it has produced some interesting theories to account for it. Rather than the embodied/abstract distinction, Tresilian (1995) proposes a dual system of object motion that is treated separately in the mind, using different mechanisms to process information: the cognitive-perceptual and the action-oriented. The cognitive-perceptual pathway would deal with more abstract information, based on prediction or using rule-based algorithms, whereas the action-oriented pathway mirrors what embodiment researchers have called the perceptionaction loop. This action-oriented pathway would consist of simple, automatic, and reactionary rules that utilize relational information between the body and the environment - information that sophisticated robots use to locomote or otherwise interact with their surroundings, or how an outfielder catches a fly ball (Raibert et al., 2008, Tresilian, 1995; Wilson \& Golonka, 2013). Tresilian proposed that we naturally conceptualize the world via the cognitive-perceptual pathway, and physically respond to the
world via the action-oriented pathway. The errors observed in folk physics research may be due to the cognitiveperceptual pathway processing information that is more naturally suited to processing by the action-oriented pathway. In Tresilian's model, the errors that are observed in curvilinear impetus belief problems are the result of the cognitive-perceptual pathway being forced to perform an inherently artificial task that it is not biologically suited to, whereas a real spiral and ball would have the action-oriented pathway perform as intended - quickly, accurately, and situated in the real world. Thus, a cognitive-perceptual pathway set opposite to an action-oriented pathway would account for all types of object motion perception, as well as the errors observed in the literature.

The goal the present study was to test the distinction between the cognitive-perceptual pathway and the actionoriented pathway by systematically increasing the degree of embodiment on variations of the spiral tasks used in McCloskey et al. (1980). The curvilinear impetus belief was tested in the abstract sense on paper, as well as in an embodied sense where participants were instructed to reach for a ball as it rolled out of a physical spiral. An intermediate condition was presented so that the spiral device was present, but the participant chose from multiplechoice correct and incorrect trajectories drawn on paper in front of the physical spiral. We hypothesize that, in the abstract condition, the brain will use the cognitiveperceptual pathway, which will result in many errors. In the prediction condition, both the cognitive-perceptual pathway and the action-oriented pathway will be engaged, which will result in fewer errors. In the action condition, only the action-oriented pathway will be engaged, resulting in the fewest errors of the three conditions.

## Method

The first group of participants were given a diagram from McCloskey et al. (1980) to control and replicate the findings, and to provide an abstract condition. They were asked to draw the trajectory of an imagined ball exiting a spiral on the diagram.

The second group was presented with a spiral device designed to carry a small metal ball, and the participants were asked to select a labeled trajectory similar to McCloskey and Kohl (1983), thus providing data for a prediction condition between abstract and action conditions, as the spiral no longer had to be imagined but was physically instantiated in front of them.

Participants in the third group, the action condition, were presented with the device and were asked to catch the ball as it exited the spiral. We recorded the distance from their hands to the correct trajectory of the ball.

All responses were recorded categorically as correct or incorrect.

## Participants:

A total of 72 adults, all undergraduate students were recruited to participate from February 28, 2015, to

November 27, 2016 through the online recruitment site at Carleton University in exchange for extra credit in a class. Two participants were excluded from the dataset due to a misunderstanding of the instructions. From the remaining participants, there were 40 females and 30 males. The average number of physics courses taken was 1.6 , and the average year of university of the participants was 2.2 . Participants were split into three groups: an abstract condition, a prediction condition, and an action condition. There were 24 participants for the abstract condition, 23 for the prediction condition, and 22 for the action condition, with a mean age of 20.6. All participants had normal or corrected to normal vision.

## Materials:

Participants in the abstract condition were provided with a pen, and a spiral diagram that was used in McCloskey et al. (1980) on $8.5 "$ x $11 "$ white standard weight paper.

The prediction and action conditions were presented with the physical device sitting on a table, positioned approximately 45 cm to the left of the participant. The device was 90 cm tall, with a diameter of 44 cm (See Figure 3). For the prediction condition, the device was set alongside pre-drawn predictions of correct and incorrect trajectories on a large sheet of $61 \times 90.2 \mathrm{~cm}$ graph paper, placed so that they appeared to emerge from the exiting end of the tube. The spiral was made of one inch clear tubing, 3-inch PVC pipe, and wood fittings. The metal bearing used in Condition 3 weighed 28.3 grams.

To determine where the participants were reaching in Condition 3, a Nikon P7700 Powershot video camera was placed above the area on a tripod so as to include the participant's hand, as well as 1 -inch measurement marks to measure correct or incorrect responses.

## Stimuli:



Figure 1: Spiral figure for the Abstract condition
The stimuli for the abstract condition was taken from McCloskey et al. (1980) in order to verify that the groups of students used in our experiment produced similar errors to the students in the previously mentioned paper, as well as provide data for an abstract condition (Figure 1).

In the prediction condition, participants were seated in the same orientation to the apparatus as the action condition to select from set trajectories. The participants were presented with the spiral apparatus (rather than solely a picture) and
were asked to verbally make a selection from the set of predefined trajectories. The trajectories were taken from the most likely errors in McCloskey et al. (1980) that were


Figure 2: Spiral figure used in conjunction with spiral device for the Prediction condition
subsequently used in the multiple-choice condition in McCloskey and Kohl (1983) (Figure 2).

In the action condition, participants were presented with the spiral apparatus and were asked to catch the ball as it exited the tube (Figure 3). A camera was employed to measure the reactions and catalogue the accuracy of the motor response.


Figure 3: Spiral device for the prediction and action conditions. Side view (left) and top view (right)

## Procedure:

In the abstract condition, the participants were presented with a paper and pencil test with the diagram from McCloskey et al. (1980), hidden beneath the short demographic questionnaire. They were then given the following instructions:

[^356]In the prediction condition, after completing the questionnaire, the participants were shown the physical apparatus and were asked to choose from 6 set trajectories from McCloskey and Kohl (1983). The participants in this condition were instructed as follows:
"This device will allow a small metal ball to travel down the tube. The ball will then exit the tube and travel on its trajectory. Your task is to pick which labeled trajectory the ball will most likely follow without using your hands. Please verbally indicate your answer."

In the action condition, after completing the questionnaire, the participants were shown the physical spiral device and given the following instructions:

> "This device will allow a small metal ball to travel down the tube. Once the ball is inserted into the tube, your task is to reach for the ball only when it passes the marked area in red so as to catch it in the palm of your hand. Please move your hand in a straight line away from you."

Participants were asked to wait until the ball had passed the marked area so that their hand would have to move late and fast, to prevent them from moving their hand in response to the perceived trajectory-we wanted to measure where their motor system thought the ball would go, not how it might change trajectory on the fly.

## Measures:

We counted any curved lines drawn in the abstract condition as erroneous and took them to indicate a curvilinear impetus belief. Similarly, we categorized as incorrect any curved responses selected in the prediction condition.

In the action condition, we coded any deviations away from the correct trajectory as erroneous. This was accomplished by measuring, on the video, how far the participant's knuckle on the index finger was from the correct trajectory. When the participants reached for the ball after it had passed the marker, any deviations from the correct trajectory exceeding 1 inch away from the zero line were recorded as errors. If while attempting to catch the ball, the knuckle of the hand was within one inch on either side of the zero line, the response was categorically correct.

## Design:

Participants were assigned to the three conditions in the order of abstract, prediction and action, in a between subjects design. A Chi Square Test for Independence was conducted for all three conditions, and between paper and prediction, paper and action, and prediction and action as post hoc analyses. The number of physics courses was split into three arbitrary groups of None (0), Some (1-3), and Many ( 3 or more), and were analyzed with the Chi Square test for independence as well. All post-hoc tests were corrected using the Bonferroni correction.

## Results

Participants in the action condition, where people actually reached for the ball rolling out of a curved tube, were much more accurate than participants who merely predicted the path of the ball on paper.

Over all three conditions, the accuracy of responses differed significantly between the conditions, though not in the predicted increasing fashion. A Chi Square test for Independence between the three groups produced $\chi^{2}=(2, N$ $=68)=26.31, p<.001$, Cramer's $V=.43$, which is statistically significant and indicates a large effect size.


Post hoc analysis illustrates that the main effect, between all three conditions, was primarily driven by the difference between the prediction condition and the action condition, as can be observed in Figure 4. Post hoc testing using the Bonferroni Correction for pairwise comparisons with the Chi Square test determined that there was a significant difference between the abstract and prediction comparison, $\chi^{2}=(1, N=47)=10.55, p<.017$, the abstract and action comparisons, $\chi^{2}=(1, N=47)=16.42, p<.017$, as well as the prediction and action comparisons, $\chi^{2}=(1, N=46)=$ 25.66, $p<.017$.

There was no relationship between the number of physics courses taken, $\chi^{2}=(2, N=70)=1.59, p=.45$. Post Hoc tests with the Bonferroni Correction also revealed no significant differences between groups, $p>.017$.

## Discussion

These results have provided some support for the original hypothesis that the degree of embodiment significantly influences the accuracy of responses. However, we did not observe the stepped increase from the abstract condition to the action condition. There was no effect of the number of physics courses on the number of correct responses, which supports Kozhevnikov and Hegarty's (2001) finding that physicists and non-physicists alike seem to rely on an incorrect default curvilinear physics for abstract problems. In the abstract condition, $62.5 \%$ of the respondents produced incorrect trajectories compared to the $36 \%$ and 49\% reported in McCloskey et al. (1980).

The original hypothesis that the difference between the abstract condition and the action condition would produce significantly different responses was supported. Surprisingly, however, the prediction condition had the fewest correct responses, with most participants choosing between two of the 6 options (E and D in Figure 2), whereas it was hypothesized that the prediction condition would have more correct responses than the abstract condition. Lastly, the action condition had the most correct responses, which supported our hypothesis that there is a dissociation between our action-oriented pathway and our cognitiveperceptual one.

With this paradigm in mind, the finding that many of the prediction condition responses were incorrect might be explained by the fact that both the cognitive-perceptual and the action-oriented pathways were in conflict, causing more errors. Alternatively, there was no participant motion involved, and this lack of a motor aspect in the prediction condition might have hindered its validity as a medium embodied condition. Also, the relatively high number of correct responses in the action condition suggests that physical responses are more biologically adapted for accuracy. Taking these findings into account, the act of drawing in the abstract condition may have allowed the action-oriented pathway to provide more correct paths of the imaginary ball, while the stationary aspect of the prediction condition may have only allowed the cognitive-perceptual pathway to be used. Thus, the abstract condition may have been more embodied than the prediction condition, which if true, would demonstrate the stepped increase hypothesized.

The significant differences observed in this experiment have illustrated the degree to which our minds are divided depending on the task and level of physical action. However, future research could expand the prediction condition of this experiment by investigating how participant motion, for example pointing to the correct trajectory or producing it alongside the spiral might increase the number of correct responses by activating the actionoriented pathway. Other areas of research may include asking participants to imagine the trajectory of the ball before choosing or drawing their prediction, to more accurately determine the capacities of the cognitiveperceptual pathway. Furthermore, the method for coding accuracy was done categorically and was measured differently between conditions, posing some reliability concerns. Because of these limitations, future projects should investigate the production of the trajectories, as in the abstract condition, in a setup identical to the prediction and action conditions, thereby increasing the generalizability of the results.

The implications of this research are broad. The pedagogical implications for teaching elementary physics could be substantial. Participants were largely incorrect when presented with abstract, as well as non-dynamic examples of this physics problem (i.e. prediction condition), so it might be beneficial for educational authorities to encourage more embodied examples of physics problems to
facilitate student understanding. The results partially support Tresilian's (1995) model that the cognitive-perceptual pathway demonstrates poorer performance on abstract physical tasks, whereas the action-oriented pathway does better on responsive and physical tasks.

These results have provided evidence for one of the cornerstones of embodied cognition: that cognition evolved for action (Wilson, 2002). This is not usually the case in many research papers, as much embodied research that is publically received involves associations between mental representations and bodily postures, types of movement, and even the types of clothes people wear (Adam \& Galinsky, 2012; Markman \& Brendl, 2005; Reutner, Hansen, \& Greifeneder, 2015). Although the results from these associational studies are interesting, they support a weaker version of embodied cognition; one that implies that cognition is merely influenced by the brain, the body and the environment. This study suggests a distributed view of the mind, one that engages the action-oriented pathway when solving embodied motion problems, yet falls back on the cognitive-perceptual pathway when motion is absent.

At present there is evidence for a dual-pathway system: one utilized when the mind is processing abstract conceptualizations, and one that utilizes the motor system when the mind is processing motion problems, using perception-action loops in a reactionary fashion. The classical view of the mind within cognitive science as a symbol manipulator falls apart when real world motion is involved, and embodied cognition and Tresilian's dual pathway system for object motion may provide a new way of conceptualizing how minds react and think about object motion.

## References

Adam, H., \& Galinsky, A. D. (2012). Enclothed cognition. Journal of Experimental Social Psychology, 48(4), 918925.

Catrambone, R., Jones, C. M., Jonides, J., \& Seifert, C. (1995). Reasoning about curvilinear motion: Using principles or analogy. Memory \& Cognition, 23(3), 368373.

Chapman, S. (1968). Catching a baseball. American Journal of Physics, 36(10), 868-870.
Cook, N. J., \& Breedin, S. D. (1994). Constructing naive theories of motion on the fly. Memory \& Cognition, 22(4), 474-493.
Fink, P. W., Foo, P. S., \& Warren, W. H. (2009). Catching fly balls in virtual reality: A critical test of the outfielder problem. Journal of Vision, 9(13), 14.
Freyd, J. J., \& Jones, K. T. (1994). Representational momentum for a spiral path. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(4), 968-976.
Goodale, M. A., \& Milner, A. D. (1992). Separate visual pathways for perception and action. Trends in neurosciences, 15(1), 20-25.

Haffenden, A. M., Schiff, K. C., \& Goodale, M. A. (2001). The dissociation between perception and action in the Ebbinghaus illusion: Nonillusory effects of pictorial cues on grasp. Current Biology, 11(3), 177-181.
Kaiser, M. K., Jonides, J., \& Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. Memory \& Cognition, 14(4), 308-312.
Kozhevnikov, M., \& Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. Memory \& Cognition, 29(5), 745-756.
Markman, A. B., \& Brendl, C. M. (2005). Constraining theories of embodied cognition. Psychological Science, 16(1), 6-10.
McCloskey, M., \& Kohl, D. (1983). Naive physics: the curvilinear impetus principle and its role in interactions with moving objects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(1), 146-156.
McCloskey, M., Caramazza, A., \& Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. Science, 210(4474), 1139-1141.
McLeod, P. Reed, N. Dienes, Z. (2006). The generalized optic acceleration cancellation theory of catching. Journal of Experimental Psychology: Human Perception and Performance, 32, 139-148.
Oberle, C. D., McBeath, M. K., Madigan, S. C., \& Sugar, T. G. (2005). The Galileo bias: A naive conceptual belief that influences people's perceptions and performance in a ball-dropping task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(4), 643-653.
Piaget, J. (1976). The grasp of consciousness: Action and concept in the young child. Cambridge: Harvard University Press.
Raibert, M., Blankespoor, K., Nelson, G., Playter, R., \& Team, T. B. (2008, July). Bigdog, the rough-terrain quadruped robot. In Proceedings of the 17th World Congress (Vol. 17, No. 1, pp. 10822-10825). Proceedings Seoul, Korea.
Reutner, L., Hansen, J., \& Greifeneder, R. (2015). The Cold Heart Reminders of Money Cause Feelings of Physical Coldness. Social Psychological and Personality Science, 1948550615574005, 1-6.
Tresilian, J. R. (1995). Perceptual and cognitive processes in time-to-contact estimation: analysis of prediction-motion and relative judgment tasks. Perception \& Psychophysics, 57(2), 231-245.
Wilson, A. D., \& Golonka, S. (2013). Embodied cognition is not what you think it is. Frontiers in Psychology, 4, 58.
Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin \& Review, 9(4), 625-636.
Zago, M., \& Lacquaniti, F. (2005). Cognitive, perceptual and action-oriented representations of falling objects. Neuropsychologia, 43(2), 178-188

# Network Analysis of a Large Sample of Typical and Late Talkers 

Eva Jimenez (E.Jimenez-Mesa@warwick.ac.uk)<br>Department of Psychology, University of Warwick, University Road, Coventry, CV4 7AL, United Kingdom<br>Thomas Hills (T.T.Hills@warwick.ac.uk)<br>Department of Psychology, University of Warwick, University Road, Coventry, CV4 7AL, United Kingdom


#### Abstract

The focus of this paper is to examine differences in semantic network structure of late talkers and typical talkers to elucidate potential learning strategies used by late talking children. To address this question, we conducted network analysis on the vocabularies of 2,912 children, with 566 of those being late talkers. Contrary to previously reported findings, the results show that late talkers have well-connected vocabularies as measured by median degree, clustering coefficient, and mean distance, with more well-connected networks in some cases than their typical talking peers. Further analysis of word order suggests that late talkers may be selecting based on frequency and connectivity of the words in the learning environment, more so than typical talkers. The language processing difficulties in late talkers appear not to be associated with their semantic network properties. In sum, late talkers may initially benefit from using word frequency and word connectivity strategies to build well-connected vocabularies.


Keywords: semantic networks; network analysis; corpus analysis; language acquisition; late talkers; word frequency

## Introduction

Children start learning words within the first and second year of life. Some children are slower than others, and some of the slowest children go on to show lifelong learning difficulties. These children, called late talkers, have been the subject of extensive research, both to understand how they learn but also to understand how their learning might be better facilitated to prevent lifelong problems. One of the outstanding questions in late talker research is to what extent late talkers are simply 'slower' versions of typical talkers? The alternative is that they show different learning strategies and therefore not only learn words more slowly but learn different words. In the present work we address this question using network analysis on the largest currently available sample of children's vocabularies. The aim is that by identifying similarities and differences in vocabulary acquisition, we can better identify the strategies that late talkers might be using, if indeed they are using strategies different from typical talkers. Before we explain our methods, we first briefly review the literature on network analysis and late talkers.

## Network Analysis in Language Acquisition

Semantic network analysis has allowed researchers to explore language processing in adults (Wachs-Lopes \& Rodrigues, 2016) and language acquisition in children (Hills,

Maouene, Riordan \& Smith, 2010; Hills, Maouene, Maouene, Sheya \& Smith, 2009). In network analysis, also known as graph theory, words are modelled as vertices and relationships between words are modelled as edges. Semantic relatedness amongst words is the focus of the present study, however, other relationships have been used in the past, including features, phonology, and free associations (e.g., Li , Farkas \& MacWhinney, 2004, Hills, Maouene, Maouene, Sheya \& Smith, 2009).

According to Watts \& Strogatz (1988) small world networks are "highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs". Smallworld properties have been reported not only in adult vocabulary but also in toddlers as young as 15 months (Beckage, Smith, \& Hills, 2011). Local structure, where words are connected in clusters, may represent semantic categories (Hills, Maouene, Maouene, Sheya \& Smith, 2009b). Words between clusters may facilitate transitions between clusters, and therefore are believed to be critical in language processing (Cancho \& Solé, 2001; Banavar, Maritan \& Rinaldo, 1999).

Networks statistics can be computed to evaluate the connectivity of a lexical network. Three of them are considered in this work. The degree of a word is the number of ties that word has with other words. Calculating the mean or median of this measure provides the overall level of cohesion in a network. The clustering coefficient explores the degree of clustering of nodes within a network. Finally, the mean distance of a network shows the average of the shortest paths between all pairs of words which gives an idea of its global access. Many studies have included these statistics in their analysis to examine research questions related to language acquisition, e.g. studies on lexical growth (Hills, Maouene, Riordan \& Smith, 2010), categorical memberships in the lexicon (Hills, Maouene, Maouene, Sheya \& Smith, 2009b), and the influence of bilingual first-language learning on early English acquisition (Bilson, Yoshida, Tran, Woods \& Hills, 2015).

## Language Acquisition in Typical and Late Talkers

Extensive research has already taken place around those children with small vocabularies compared to their normed peers (for a review, see Desmarais, Sylvestre, Meyer, Bairati \& Rouleau, 2008). Whereas many late talkers catch up with their peers in word production (so called 'late bloomers'; Thal, Tobias \& Morrison,1991), some others will have language difficulties that drag on, to be later diagnosed with Specific Language Impairment (SLI) (Leonard, 2000). For
the latter group, language problems will continue with comprehension, production and/or pragmatics (Leonard, 2000). Despite late bloomers' vocabulary improvement, they may still be more likely to experience difficulties in language-related tasks, such as reading (Rescorla, 2009). Defining and understanding the characteristics and strategies of late talker's vocabulary could help to develop effective early interventions.

The starting point for our research is Beckage, Smith and Hills (2011). In their study, network analysis was used to characterize the vocabularies of 66 typical and late talking children. Semantic relatedness of words, computed from word co-occurrence derived from the CHILDES database (MacWhinney, 2000), was used to connect the words in the child's vocabulary. Results showed that both typical and late talkers exhibit small-world structure, although late talkers present this to a lesser degree. The study suggests the existence of a relationship between the child's rate of lexical development and the connectivity of her individual network. This finding led the authors to hypothesize the possibility of an 'oddball' strategy used by late talkers: a preference to learn words that have lower semantic relatedness with words they already know. Thus, late talkers may use different learning strategies or have differences in their ability to discriminate word referents.

A later study by Nematzadeh, Fazly \& Stevenson (2014) challenged these results. By means of computational modelling, the authors simulated typical and late talking word learners to explore differences in their semantic networks. Surprisingly, neither type of simulated word learner showed a small-world structure. Referring to Beckage, Smith and Hills' work (2011), the authors questioned the use of the same edges to link words of the networks of both typical and late talkers as it assumes that both groups learn the same knowledge about words. Moreover, the authors called into question the 'oddball strategy' alleging that late talking children do not possess enough information about the words to discern similarities and dissimilarities between words. However, their methodological differences may also explain the differences in the results. Whereas in Beckage, Smith and Hills' study (2011) the co-occurrence of words in childdirected speech generated semantic relatedness between words, Nematzadeh, Fazly \& Stevenson (2014) used associative semantic information provided by a custom lexicon to link words with similar meanings.

Word frequency has also been investigated as one of the main influences in word learning. In Stokes (2010), two-year old typical talking children tended to acquire more high frequency words than late talking children, who in turn learned more words with a higher phonological neighborhood density. However, the question remains whether there is a difference between the two groups with respect to preferences for acquiring some words earlier than others, which may have given a different perspective of the learning strategies used by each group.

The present work further investigates the difference between early and late talkers using vocabularies taken from
the open repository website Wordbank (Frank, Braginsky, Yurovsky \& Marchman, in press), providing a sample of 2,912 children, of which 566 are considered late talkers. The methodology followed and the words selected are identical to those in Beckage, Smith and Hills (2011), the only differences being the number of children and the diversity of their backgrounds as they come from nine different American studies. Our principle question is, how does the semantic vocabulary structure differ between typical and late talkers? We ask two additional questions: (1) what is the relationship between word frequency and order of word learning for the two groups, and (2) what is the relationship between word frequency in the language learning environment and the connectivity between the words. These allow us to investigate additional pathways for language learning in late talkers.

## Methods

## Vocabulary

Publicly available vocabulary data for 5,450 children aged 16 to 30 months was downloaded on October 2016 from Wordbank (Frank et al., in press). Data is contributed by various researchers using the MacArthur-Bates Communicative Development Inventory (MCDI) (Fenson, Dale \& Reznick, 1993). The data set used in this work was downloaded by selecting 'Words \& Sentences' under forms and 'English' under language. To facilitate comparisons between late and typical talkers, we limited vocabulary sizes to between 20 and 220 words, a range where typical and late talkers overlapped that also allowed for meaningful network statistics. Few late talkers had a productive vocabulary size greater than 220 words. After limiting the vocabulary size, the final total number of children remaining was 2,912 . Of the 2,912 children (aged 15 to 30 months), 566 have a vocabulary size atypical for their age and 2,346 presented a normal vocabulary size for their age. Late talkers were at or under the $20^{\text {th }}$ percentile of their same-age peers. To calculate this, each child was assigned to a decile grouping according to their age and vocabulary size reported in the MCDI instrument. The decile grouping was looked up from a table of estimated percentiles on the Wordbank website created using a quartile regression with monotonic polynomial spline as the base function. Although the total number of words that can be recorded in the MCDI questionnaire is 680, in Beckage, Smith and Hills' study (2011) only 291 words were used which appeared on both the toddler and infant forms, allowing comparison across ages. The same words were selected in this study. The 291 words consist of 207 nouns, 50 verbs, 14 adjectives, 12 pronouns, 6 adverbs, 1 quantifier and 1 demonstrative. All categories in the MCDI were included except for 'Sound Effects and Animal Sounds', 'Helping Verbs', and 'Connecting Words'.

## Semantic Relationships Between the Words

To link the words in each child's productive vocabulary, semantic relatedness between words was computed using co-
occurrence statistics derived from an American English corpus of child-directed speech, CHILDES (MacWhinney, 2000). A surface proximity approach (see Evert, 2008) was used to determine the frequency in which each distinct word (node) in the corpus co-occurred with other words (collocates). An empty co-occurrences matrix was created and then populated by moving a window of span size 5 words forward through the corpus. As co-occurrences were encountered the count for that pair was incremented. A subset of this large matrix was created where the rows and columns intersected with the 291 words selected from the MCDI forms. Finally, the count values were converted to a simple binary representation.

## Random Acquisition Networks

In order to compare each individual child's network with similar size networks, 300 random acquisition networks were generated for each child. These networks have the same number of words $n$ as the child's network, but the words are selected randomly from the set of 291 words. Then, each random network was linked using the values from the CHILDES matrix explained above. The same statistical properties computed for each child's vocabulary network were also computed on the 300 random acquisition networks and then averaged. These random network statistics provide the structure inherent in the language context without including the particular word learning pattern of the child, thus providing a point of comparison for each child's network, allowing us to compare children with different size vocabularies against a 'random' learner.

## Word Frequency and Connectivity

Preferences for learning certain words over others was assessed with respect to vocabulary size. First, the sample was divided into two groups: late talkers and typical talkers. Then each group was ordered by increasing vocabulary size, creating subgroups. Within each subgroup and for each word a count was made of children that produced the word. Words were ranked based on their count. Differences in ranking between late and typical talkers was calculated by subtracting the respective ranking value for each word, allowing us to identify differences between the groups in their preference for learning certain words.

Only words with a minimum ranking difference of 20 or more are presented, but the results are not sensitive to this number. The frequency of each word was taken from CHILDES and compared between the two groups (see MacWhinney, 2000; Li \& Shirai, 2000). Within the $291 \times 291$ matrix of co-occurrences, some words are better connected (have higher degree) than others. This was calculated directly from the matrix by counting the total number of occurrences of a word with other words.

Table 1: Distribution of the children's vocabulary and mean age (in parenthesis)

| Vocabulary size |  |  |  |  |  |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $20-40$ | $40-60$ | $60-80$ | $80-$ | $100-$ | $120-$ | $140-$ | $160-$ | $180-$ | $200-$ |
|  |  |  |  | 100 | 120 | 140 | 160 | 180 | 200 | 220 |
| LT | 133 | 70 | 74 | 60 | 55 | 46 | 36 | 46 | 25 | 21 |
|  | $(23)$ | $(24.5)$ | $(25)$ | $(26.1)$ | $(27.2)$ | $(27.7)$ | $(28.3)$ | $(28.5)$ | $(29.2)$ | $(29.9)$ |
| TT | 427 | 304 | 213 | 197 | 191 | 159 | 187 | 190 | 205 | 262 |
|  | $(17.1)$ | $(17.9)$ | $(18.5)$ | $(19.6)$ | $(20.4)$ | $(20.8)$ | $(21.4)$ | $(22.7)$ | $(23.4)$ | $(24.5)$ |
| Total | 560 | 374 | 287 | 257 | 246 | 205 | 223 | 236 | 230 | 283 |

## Results

## Network Analysis

Analysis was carried out using R and the igraph package, version 1.0.1 (Csárdi \& Nepusz, 2006). Connectivity was assessed by computing three statistics of each directed network: median in-degree, clustering coefficient, and mean distance. Late talkers are unequally distributed across the sample: they have higher representation at lower vocabulary sizes and low or no representation at the higher vocabulary sizes. Therefore, the sample was divided into bins of children with similar vocabulary sizes in ranges of 20 words. The size of the bin does not influence our results, but does facilitate their visual presentation. Table 1 shows the number of late talkers (LT) and typical talkers (TT) in each bin, as well as the mean age in each group.

Stepwise linear mixed effects analysis ( $\operatorname{lm} 4$ package, version 1.1-12; Bates, Mächler, Bolker \& Walker, 2015) was performed to explore the relationship between each of the three network statistics and type of talker. The final three models includes vocabulary size and type of talker as fixed effects to predict each network statistic. No collinearity of these predictors was found (VIFs < 1.02). Intercepts were allowed to vary across vocabulary sizes as the values of each network statistic varies across the size contexts. The inclusion of vocabulary size as a fixed effect allowed us to control for network size and led to a better statistical fit. No significant difference in the model resulted when allowing slopes to vary by either vocabulary size or type of talker. The significant main effect for type of talker reveals a positive relationship between LT and median indegree (Estimate $=.061, S E=.025$, $p=.014$ ). A marginal main effect for type of talker indicates that being a late talker leads to higher clustering coefficient (Estimate=.003, $S E=.002, p=.063$ ). A negative main effect was observed for late talkers on mean distance (Estimate $=$ $.014, S E=.004, p=.0008$ ). These results indicate that LT's vocabularies are better connected and have better global access than TT's vocabularies when considering all the vocabulary sizes together.

We compared the observed networks statistics with those from the vocabulary size-matched random acquisition networks by calculating ratios (Figures 1, 2 and 3). Linear mixed effect analysis was also carried out to examine the relationship between these ratios and type of talker. The model had the same structure as used in the previous analysis.

Results indicate no main effect for the type of talker on indegree ratio (Estimate $=.024, S E=.014, p=.078$ ). Significant main effects were found on clustering coefficient (Estimate $=.009, S E=.005, p=.044$ ) and mean distance ratios (Estimate $=-0.008, S E=0.002, p=.0007$ ), indicating that late talking children tend to have higher clustering coefficient and lower mean distance in their networks.


Figure 1. Median degree ratio of the observed data to the RAN. Note: * p < 0.05


Figure 2. Clustering coefficient ratio of the observed data to the RAN. Note: * p < 0.05


Figure 3. Mean distance ratio of the observed data to the RAN. Note: * p < 0.05, ** p < 0.01

Table 2: Difference between typical and late talkers compared to random acquisition networks

|  |  | Networks M(SD) | $\begin{gathered} \text { RAN } \\ \text { M(SD) } \end{gathered}$ | t (df) | d |
| :---: | :---: | :---: | :---: | :---: | :---: |
| LT | In-degree | $\begin{gathered} 23.89 \\ (12.56) \end{gathered}$ | $\begin{gathered} 20.81 \\ (12.61) \end{gathered}$ | $\begin{gathered} 12.6 \\ (565) \end{gathered}$ | . 530 |
|  | Clustering coefficient | $\begin{aligned} & .58 \\ & (.07) \end{aligned}$ | $\begin{gathered} .54 \\ (.01) \end{gathered}$ | $\begin{gathered} 13.9 \\ (565) \end{gathered}$ | . 584 |
|  | Mean distance | $\begin{aligned} & 1.68 \\ & (.1) \end{aligned}$ | $\begin{aligned} & 1.73 \\ & (.02) \end{aligned}$ | $\begin{aligned} & 10.33 \\ & (565) \end{aligned}$ | . 434 |
| TT | In-degree | $\begin{gathered} 26.7 \\ (13.44) \end{gathered}$ | $\begin{gathered} 25.01 \\ (14.56) \end{gathered}$ | $\begin{aligned} & 13.81( \\ & 2345) \end{aligned}$ | . 285 |
|  | Clustering coefficient | $\begin{aligned} & 0.57 \\ & (.06) \end{aligned}$ | $\begin{aligned} & 0.55 \\ & (.01) \end{aligned}$ | $\begin{gathered} 18.81 \\ (2345) \end{gathered}$ | . 388 |
|  | Mean distance | $\begin{gathered} 1.7 \\ (.09) \end{gathered}$ | $\begin{aligned} & 1.72 \\ & (.02) \end{aligned}$ | $\begin{gathered} 10.23 \\ (2345) \end{gathered}$ | . 211 |

Standardized residual plots were visually inspected for the linear mixed models to check for homoscedasticity and normality, and a violation of these requirements was noted. To confirm these results, we carried out more analysis. Ttests were conducted to detect differences between LT and TT regarding the ratio of the observed statistics to the statistics of the size-matched RAN. Late talker's vocabularies have higher values of median in-degree ( $M=1.28, S D=.39$ ) and clustering coefficient ( $M=1.08, S D=.140$ ) than typical talking children (in-degree: $M=1.20, S D=.37$ ), $t(817)=$ 4.54, $p<.001, d=-.22$; clustering coefficient: $M=1.05, S D=$ $0.13), t(820)=-4.72, p=p<.001, d=-.23$ ). Late talking children also had lower values of mean distance ( $M=.97$, $S D=.06$ ) than typical talking children ( $M=.99, S D=.063$ ), $t(815)=5.18, p<.001, d=.25$.

Further analysis using the ratio data was conducted to check whether the same differences between LT and TT are also observed within each bin of vocabulary size. Significant results are signaled with an asterisk in Figures 1, 2 and 3. Late talkers obtained higher in-degree and higher clustering coefficient than TT in the vocabulary range 40 to 60 words. Significant differences in mean distance were found in three groups of vocabulary size, all of them present lower values for LT: 40 to 60 words, 60 to 80 words, and 140 to 160 words. The same t-test analysis in each bin was performed using only the observed data. Apart from mean indegree, in which any difference between the type of talkers were found, similar results were obtained in clustering coefficient and mean distance in the same vocabulary sizes. These results show that LT's vocabularies are better connected than TT's
vocabularies in certain vocabulary ranges, and also that TT resembles their RAN more than LT do in these ranges.

Paired t-tests were conducted to compare the properties of LT and TT networks with their corresponding size-matched RAN. Results can be seen in Table 2. Both LT and TT showed significantly higher values of median in-degree and clustering coefficients and significantly lower values for mean distance than their size-matched RAN. Thus, LT and TT seem to present vocabularies which are well connected and have good global access.

## Word Frequency and Connectivity

To investigate differences in the order in which LT and TT learn words, we examined the correlation between word order and word frequency in CHILDES. The order in which all 291 words are learned across vocabulary size is not correlated to word frequency for either type of talker (LT: $r_{s}=-.10, p=.082$, TT: $r_{s}=-.063, p=.28$ ). When the same analysis is performed on each vocabulary size bin, word order in LT was significantly related to word frequency in vocabulary sizes of 20 to 40 words ( $r_{s}=-.13, p=.023$ ), and 40 to 60 words ( $r_{s}=-$ $.16, p=.006$ ). Thus, the order in which LT learn words seems to be related to their frequency during the first stages of vocabulary development. Further correlation analysis revealed a relationship between connectivity and word order ( $r_{s}=-.13, p=.026$ ) in LT but not in TT ( $r_{s}=-.11, p=.071$ ). Analysis on each bin shows that, only for LT, this relationship is present in vocabulary sizes of 20 to 40 words ( $r_{s}=-.16, p=.008$ ), 40 to 60 words ( $r_{s}=-.18, p=.002$ ), 100 to 120 words ( $r_{s}=-.13, p=.032$ ), 120 to 140 words ( $r_{s}=-.13, p=$ .024 ), 180 to 200 words ( $r_{s}=-.12, p=.043$ ), and 200 to 220 words ( $r_{s}=-.12, p=.041$ ). In view of these results, it seems that connectivity and frequency of the 291 words are somehow related to word order in LT but not in TT. In addition, word connectivity seems to be more strongly related to word order in LT than word frequency.

Further analysis considered only those words that differed considerably in word order between LT and TT. That is, there were some words that were learned earlier by one of the type of talker groups compared to the other group, we refer to them here as 'preferred words'. Figure 4 shows significant differences between the two types of talkers: LT learned more words that are highly frequent in the language environment $(M=3588, S D=897)$ than TT $(M=1060, S D=687), t(16.9)=$ $7.08, p<.001, d=3.16$. The connectivity within the matrix of co-occurrences of preferred words was compared between LT and TT. Children with language delay learn words that are well connected in the matrix ( $M=101, S D=11.4$ ), more so than TT $(M=75.8, S D=12.01), t(17.9)=4.79, p<.001, d$ $=2.14$. These results are consistent with this studies analysis of the network statistics, with LT learning more wellconnected words than TT. Results from a logistic regression using the preferred words indicated that frequency is not a good predictor of the type of talker ( $B=.007, S E=.008, p=.38$, Homes-Lemeshow $R_{2}=.85$ ), contrary to connectivity, which was a significant predictor $(B=.14, S E=.055, p=.009$, Homes-Lemeshow $R_{2}=.50, \quad 95 \%$ CI [1.06, 1.33]).

Additionally, we wanted to see whether word frequency in CHILDES and the connectivity of the 291 words within our matrix of co-occurrences are correlated. Results showed that these two measures are strongly related ( $r_{s}=.91, p<.001$ ). These results suggest that LT children may be more susceptible to these two word properties, however only the connectivity of the preferred words can predict the type of talker.


Figure 4. Word frequency by vocabulary size. Numbers at each point reflect number of preferred words used for analysis in each group.

## Discussion

The present study uses network analysis on a large sample of children's vocabulary to explore the idea that late talkers may have a different word learning strategy than typical talkers. Multiple results from different analysis fail to agree with the findings in Beckage, Smith and Hills (2011). The authors reported that children with language delay have networks with less clustering coefficient and less mean distance than their vocabulary size-matched RAN. Furthermore, the authors hypothesized that this may be due to LT using an 'oddball strategy' to learn words, i.e. late talkers may be attracted to those words that are not well connected in the learning environment, as opposed to the idea of 'preferential acquisition', in which the tendency is to learn earlier on the most contextually diverse words in the learning environment (Hills, Maouene, Riordan \& Smith, 2010). However, the use of a larger sample in this study suggested that LTs do not use an 'oddball strategy', rather they seem to learn words that are well connected in the environment.

Analysis of the frequency of these preferred words in the learning environment showed that LT learn a good proportion of highly frequent words earlier than TT. However, the relationship between word frequency and language delay is still unclear as it seems to not be a good predictor for type of talker. Nevertheless, these results seem to contradict the findings by Stokes (2010), who found that two-year old typical talkers learn more high frequency words than LT do. However, the inclusion of function words may partly explain
the difference between the present study's results and Stokes' results as their frequency in the learning environment is higher than open class words. The connectivity of the preferred words within the 291 words used in the present study indicate that LT also happened to learn well connected words earlier.

One of the reasons behind the findings may be that late talking child are more passively influenced by word frequency, consequently, learning more highly frequent words that happen to be well connected in the learning environment. As knowing these high frequency words can deliver a degree of communication success, the requirement to acquire advanced strategies may be further delayed.

In sum, the evidence reported in this study suggests that children with language delay have well connected vocabularies and good global access, in many cases better than the typically developing children. Late talkers may be more influenced by word frequency or connectivity, perhaps using a strategy to learn words that are contextually diverse in the learning environment as noted in previous work (Hills et al., 2009a). In order to elucidate whether these two types of talkers are using different word learning strategies, future research will need to examine the longitudinal development of vocabularies in LT and TT and a different approach to assign a more individualistic semantic relatedness between the words.

## References

Banavar, J. R., Maritan, A., \& Rinaldo, A. (1999). Size and form in efficient transportation networks. Nature, 399(6732), 130-132.
Bates, D., Mächler, M., Bolker, B. M., \& Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4, Journal of Statistical Software, 67(1). http://doi.org/10.18637/jss.v067.i01
Beckage, N., Smith, L., \& Hills, T. (2011). Small Worlds and Semantic Network Growth in Typical and Late Talkers. PLoS ONE, 6(5). http://doi.org/10.1371/
Bilson, S., Yoshida, H., Tran, C. D., Woods, E. A., \& Hills, T. T. (2015). Semantic facilitation in bilingual first language acquisition. Cognition, 140, 122-134. http://doi.org/10.1016/j.cognition.2015.03.013
Cancho, R. F. i, \& Solé, R. V. (2001). The Small World of Human Language. Proceedings: Biological Sciences VO 268, (1482), 2261
Csárdi, G., \& Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Systems, 1695, 1695.
Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., \& Rouleau, N. (2008). Systematic Review of the Literature on Characteristics of Late-Talking Toddlers. International Journal of Language \& Communication Disorders, 43(4), 361-389. Evert, S. (2008). Corpora and collocations. Corpus Linguistics. An International Handbook, 1-53. http://doi.org/10.1515/9783110213881.2.1212

Fenson, L., Dale, P. S., \& Reznick, J. S. (1993). Macarthur communicative development inventories: User's guide and technical manual. San Diego, CA: Singular Publishing Group.
Frank, M. C., Braginsky, M., Yurovsky, D., \& Marchman, V. A. (in press). Wordbank: An open repository for developmental vocabulary data. Journal of Child Language.
Hills, T. T., Maouene, J., Riordan, B., \& Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. Journal of Memory and Language, 63(3), 259-273. http://doi.org/10.1016/j.jml.2010.06.002
Hills, T. T., Maouene, M., Maouene, J., Sheya, A., \& Smith, L. (2009a). Longitudinal Analysis of Early Semantic Networks Preferential Attachment or Preferential Acquisition? Psychological Science, 20, 729-739.
Hills, T. T., Maouene, M., Maouene, J., Sheya, A., \& Smith, L. (2009b). Categorical structure among shared features in networks of early-learned nouns. Cognition, 112(3), 381396. http://doi.org/10.1016/j.cognition.2009.06.002

Leonard, LB. (2000). Children with Specific Language Impairment. Cambridge, MA: MIT Press
Li, P., \& Shirai, Y. (2000). The acquisition of lexical and grammatical aspect. Berlin and New York: Mouton de Gruyter.
Li, P., Farkas, I., \& MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. Neural Networks, 17(8-9), 1345-1362. http://doi.org/10.1016/j.neunet.2004.07.004
MacWhinney, B. J. (2000). The CHILDES project: Tools for analyzing talk (3rd ed.). Mahwah, NJ: Erlbaum.
Nematzadeh, A., Fazly, A., \& Stevenson, S. (2014). Structural Differences in the Semantic Networks of Simulated Word Learners. Proceedings of the 36th Annual Conference of the Cognitive Science Society, 1072-1077.
Rescorla, L. (2009). Age 17 language and reading outcomes in Late-Talking Toddlers: Support for a dimensional perspective on language delay. Journal of Speech, Language and Hearing Research, 52,16-30.
Stokes, S. F. (2010). Neighbourhood density and word frequency predict vocabulary size in toddlers, Journal of speech, language and hearing research, 53, 670-683.
Thal, D., Tobias, S., \& Morrison, D. (1991). Language and gesture in late talkers: A 1-year follow-up. Journal of Speech and Hearing Research, 34, 604-612.
Wachs-Lopes, G. A., \& Rodrigues, P. S. (2016). Analyzing natural human language from the point of view of dynamic of a complex network. Expert Systems With Applications, 45, 8-22. Retrieved from http://10.0.3.248/j.eswa.2015.09.020
Watts, D. J. \& Strogatz, S. H. (1998). Collective dynamics of small-world networks. Nature, 393, 440-442.

# Spatial language: Meaning, use, and lexical choice 

Kristen Johannes (kjohann@WestEd.org)<br>Science Technology Engineering and Mathematics Program, WestEd<br>300 Lakeside Dr., Oakland, CA 94612 USA

Barbara Landau (landau@cogsci.jhu.edu)<br>Department of Cognitive Science, Johns Hopkins University, 3400 N Charles St., Baltimore, MD 21218 USA


#### Abstract

Accounts of spatial language aim to address both the meaning of a spatial term and its usage patterns across diverse cases, but do not always clearly distinguish these from one another. Focusing on the case of English prepositions in and on, we set out to disentangle spatial language meaning from spatial language use by comparing judgments on a series of linguistic tasks designed to tap each aspect of spatial language. We demonstrate that judgments of truth-conditional meaning and patterns of naturalistic use show different distributional signatures, with judgments of meaning giving rise to a more uniform distribution than use patterns. We explore a third aspect of spatial language: lexical choice, and propose that choice is a key factor in shaping the distribution of spatial expression use. Our analyses reveal that the distribution of lexical choice judgments is highly correlated with the distribution of expression use in spatial descriptions for the same spatial scenes, supporting a model of spatial language that differs from traditional accounts of meaning and categorization.


Keywords: Spatial cognition; spatial language; semantics; language use

## Introduction

Spatial terms in languages of the world tend to constitute a small closed class set (Landau \& Jackendoff, 1993; Talmy, 1985). In English, for example, this set is typically limited to the spatial prepositions, including in, on, over, above, etc. To linguistically encode spatial relations with this limited inventory, a speaker must systematically abstract over fine-grained properties of objects and configurations and attend to coarse-grained spatial and/or mechanical properties of their relations. Modeling the nature of this abstraction remains a long-standing problem in the cognitive sciences. The systematic ways in which speakers encode relations (i.e., generate descriptions) is often confounded with the ways in which they decode spatial descriptions (i.e., understand the meaning of descriptions).

This problem has been exacerbated by a lack of separation between definitional questions about the meaning of a spatial term like in or on and categorization questions about the use of a term by a population of speakers questions that may ultimately have different answers. Meaning and use represent distinct and separable aspects of many semantic domains (Cruse, 2011). In keeping with this observation, we suggest that the task of formally defining
spatial terms such as in and on is separate from, albeit related to, the task of specifying the conditions under which speakers will use a spatial term to describe a location or configuration. For example, formal accounts of spatial meaning come under fire when proposed meanings cannot accommodate peripheral uses (see e.g., Bennett, 1975 for examples and e.g., Feist, 2000, and Herskovits, 1986 for commentary), while accounts of spatial categorization based on language usage patterns often propose all-or-none category boundaries that mimic binary truth conditional judgments (Regier, Khetarpah1, \& Majid, 2013). In this paper, we aim to disentangle spatial language meaning from spatial language use by comparing judgments on a series of linguistic tasks across the same sets of spatial stimuli, including a task designed to directly assess speakers' lexical choices, which we propose are key in accounting for spatial term use but are not necessarily active in spatial term meaning.

Below, we review a selection of research on spatial language categorization, focusing mainly on the prepositions in and on. We organize the review into work that explicates the formal meaning of spatial terms and work that targets speakers' use of spatial terms for categorization. We then introduce recent work that suggests that speakers' choice of spatial term from among candidates is a critical variable in reconciling categories of spatial term meaning with patterns of speakers' spatial term use. The current study addresses these relationships - between meaning, choice, and use - directly for the English prepositions in and on, evaluating two complementary hypotheses, outlined below.

## Defining spatial terms

Past and present, accounts of spatial meanings have also had to shoulder the burden of accounting for detailed patterns of spatial expression use (and, in some cases abstract uses of spatial expressions, see e.g., Jamrozik \& Gentner, 2015). Traditional simplified accounts of prepositional meaning such as Bennett (1975) ${ }^{1}$ attempt to define spatial prepositions as a function of geometric

[^357]properties of configurations as a means of abstracting away from specific objects. These definitional theories have been consistently criticized for being, on the one hand, too vague and allowing unlikely cases into the definition (e.g., an apple under an upside-down bowl fits Bennett's denotation of "in the bowl") and for failing, on the other hand, to predict the range of peripheral cases for which in and on can apply (e.g., an apple on top of other fruit contained by a bowl, cf. Feist, 2000). These accounts have been replaced by proposals that incorporate large sets of features in order to narrow and specify the meaning of in and on based on usage patterns (see e.g., Feist, 2000; Vandeloise, 2010; Xu \& Kemp, 2012), and by proposals that prioritize world knowledge and pragmatic inference (Herskovits, 1986), so as to preserve narrow denotations for in and on while accounting for peripheral cases that depend on additional processes such as chaining (Malt, Sloman, Gennari, Shi, \& Wang, 1999). The current study examines whether accounting for frequent or infrequent uses of a spatial term is a necessary goal for accounts of spatial meaning.

## Spatial categories inferred from language use

Studies of spatial language categorization typically measure speakers' usage of spatial terms for different spatial scenes and, based on these data, one can infer possible category boundaries for single terms and/or semantic structure across multiple terms. Many of these accounts do not start from any initial hypotheses about the semantic content or meaning of particular spatial terms, and instead use spatial descriptions to infer systematic groupings of scenes under spatial terms. One prevailing assumption, however, is that a given spatial scene will fall "all-or-none" into only one spatial term category (e.g., the same scene cannot be categorized as both in and on).

For example, Levinson and colleagues (Levinson et al, 2003) and Regier and colleagues (Regier et al., 2013) examined spatial descriptions for a diverse set of spatial scenes from the Topological Relations Picture Series (Bowerman and Pederson, 1993). Across a large sample of languages, both groups analyzed the spatial term(s) used by the majority of speakers in a language group to encode a given scene - a point we will return to shortly.

Levinson et al. used multidimensional scaling on these data and proposed underlying spatial categories that are shaped by a handful of "attractors" - salient spatial scenes that are encoded in similar ways across languages. Similarly, Regier et al. employed an inferential (semantic map) analysis to come to a similar solution. Both studies are agnostic to the lexical content of particular spatial terms, but the researchers' analytical choices reflect a critical assumption about how spatial language use relates to underlying spatial categories. Specifically, researchers in both studies identified the modal term used by the majority speakers of each language for each scene, treating languageinternal variation as noise. The result of this modal assumption is binary, all-or-none categorization of a scene by spatial terms in a language, partitioning spatial scenes
into language-based equivalence classes, reminiscent of binary truth-conditional meaning.

This all-or-none semantic category structure limits the inferences that can be made about the relationship between the meaning of a spatial term and its use in encoding different spatial scenes. In particular, it ignores the possibility that spatial terms might overlap in the spatial scenes they apply to, leading to probabilistic use of multiple spatial terms, and, in a similar vein, precludes the idea that spatial terms can compete with one another to encode the same spatial scene.

Recent work from Johannes and colleagues (Johannes, Wilson \& Landau, 2016; Johannes 2015; Landau, Johannes, Skordos, \& Papafragou, 2016) demonstrates that multiple spatial terms are used by English speakers to encode the same spatial scenes. Moreover, they find that tracking the fine-grained use of a single spatial term across a diverse set of spatial relation scenes reveals a graded, non-uniform distribution of expression use across scenes, suggesting that some terms are a "better fit" to a spatial scene than others. In this paper, we extend the observations of Johannes and colleagues, proposing that speakers' choice of spatial term, among many candidates, to describe a configuration is a critical variable in accounting for the non-uniform distribution found in spatial expression usage patterns.

## The Current Study

In the current study, we pursue two related hypotheses aimed at exploring how speakers evaluate the meanings of spatial terms and how this process differs from their decisions to use specific terms in spatial descriptions. We propose implicit lexical competition - speakers' choice of a particular spatial term among viable candidates - as a way of accounting for differences in speakers' judgments of spatial expression meaning and patterns of spatial expression use.

We test these hypotheses using the spatial terms in and on as a case study and compare data from three different linguistic tasks, outlined in Table 1, conducted using the same diverse sets of containment and support scenes (originally from Johannes, 2015, and Johannes, Wilson, Landau, 2016; see Figures 1 and 2). A truth-value judgment task is used to assess speakers' binary truth conditions for different expressions by simply asking whether a given expression applies to a given spatial scene. A spatial description task is used to observe speakers' self-generated spatial descriptions for each spatial scene. Finally, a forcedchoice judgment task is used to measure speakers' judgments about which of two (true) spatial expressions is a better fit to a given spatial scene.

Hypothesis 1: Speakers' judgments of the truth-conditional meaning of spatial expressions are subject to different criteria than their decisions to use these expressions in spatial descriptions. We predict that tasks that separate these two types of judgments (see Table 1) will show different distributional signatures across the same set of diverse
scenes, with usage patterns yielding a more articulated, nonuniform distribution. Moreover, speaker's truth value judgments for a given spatial expression are not necessarily predicted to correlate with their use of the same spatial expression for the same spatial scenes, as tasks are hypothesized to engage different linguistic processes.

Hypothesis 2: Speakers' truth-value judgments and patterns of spatial descriptions differ due to implicit competition among felicitous candidates, which at play in spatial description tasks but not in truth-value judgment tasks. We predict that the distribution of judgments from a lexical choice task (Table 1), wherein speakers must choose between two felicitous spatial terms for diverse spatial scenes, will align with speakers' usage pattern in a spatial description task for those same scenes, but are not predicted to correlate with truth-value judgments.

## Experiment

## Methods

Design. The experiment was structured as a betweensubjects design with five separate groups of adult participants. Each group completed a different pairing of a linguistic task with a spatial stimulus set (see Table 1).

Linguistic tasks. Table 1 provides an example of each of the linguistic tasks, along with the range of possible responses.

Table 1. Linguistic tasks used with each stimulus set, including example prompts and possible responses.

| Task | Example Prompt | Responses |
| :---: | :---: | :---: |
| Truth value judgment (Stimulus sets $1 \& 2$ ) | Is the following sentence true of the scene? <br> "The sandwich is on the plate." | Binary judgment: \{Yes or No\} |
| Spatial description (Stimulus sets $1 \& 2$ ) | Where is object A in relation to object B in the scene? <br> [A: strawberries; B: bag] | Natural language description: "The strawberries are in the bag." |
| Forcedchoice judgment (Stimulus set 1 only) | Which of these two sentences is a better description of the scene? <br> A: "The tape is on the box" <br> B: "The tape is stuck to the box" | Binary judgment: \{A or B \} |

Participants. A total of 175 adults (mean age $=19.6$ years) participated in the experiment through a series of self-paced online interfaces in return for course credit. Table 2 shows the number of participants that provided data for each linguistic task.

Table 2. Participant breakdown across tasks and stimuli sets.

| Stimulus set 1 Tasks | $\mathbf{N}$ |
| :---: | :---: |
| Truth value judgment | 50 |
| Spatial Description | 50 |
| Stimulus set 2 Tasks | $\mathbf{N}$ |
| Truth value judgment | 25 |
| Spatial Description | 25 |
| Forced-choice judgment | 25 |

Materials. We used two sets of stimuli to elicit linguistic judgments and descriptions. Stimulus set 1 was developed by Johannes (2015) and consisted of 64 containment scenes and 64 support scenes, for a total of 128 items (Figure 1). Stimulus set 2 came from Johannes, Wilson, and Landau (2016; adapted from Landau et al., 2016) and consisted of 18 containment scenes and 15 support scenes, for a total of 33 items (Figure 2).


Figure 1. Example containment (left) and support (right) scenes from Stimulus set 1.


Figure 2. Example containment (left) and support (right) scenes from Stimulus set 2.

Procedure. Participants completed each task using a selfpaced online interface. Tasks included critical trials, which probed linguistic judgments for the prepositions in and on for containment and support scenes (see Figures 1 and 2), respectively, as well as filler trials, which elicited judgments for other prepositions or descriptions for scenes depicting other types of spatial relationships (e.g., proximity). The number of critical trials in each task varied depending on the stimulus set: tasks employing Stimulus set 1 had 128 critical trials and 62 filler trials, while tasks that employed Stimulus set 2 had 33 critical trials and 11 filler trials.

## Results

We first examined the relationship between truthconditional meaning and expression use by comparing the patterns of participants' truth-value judgments to their expression usage patterns from the spatial description task, both carried out using Stimulus set 1 . We then explored the relationship between participants' truth-value judgments, spatial descriptions and lexical choice patterns for tasks carried out with Stimulus set 2 . We compared patterns of spatial expression use, from the spatial description task, across spatial scenes to patterns of truth-value judgments, from the truth value judgment task, and patterns of lexical choice, from the forced-choice task. Although participants judged, or described, both containment scenes and support scenes, we present, analyze, and discuss these spatial categories separately.

Comparing distributions of truth value judgments to spatial descriptions: Stimulus set 1. Figure 3 presents a subset of containment and support items side by side and respectively shows participants' average rates of use of in or on (i.e., proportion of descriptions using in or on) in the spatial description task (top black bars) and average truthvalue acceptance rates (i.e., proportion of "True" judgments) on the truth-value judgment task (bottom white bars). Truth-value acceptance rates were greater than or equal to rates of expression use for all but 5 containment items and all but 2 support items ${ }^{2}$.

We tested whether measures of spatial expression meaning and spatial expression use show similar distributional signatures across the same spatial scenes. Our reasoning was as follows: if participants are using the same knowledge in similar ways to make judgments about spatial expression meaning and decisions about expression use, then the resulting pattern of truth-value judgments for in and on should systematically relate to the pattern of in and on use in descriptions of the same scenes. That is, scenes that are frequently described with in or on should also show higher rates of acceptance on the truth value judgment task, and scenes for which in and on are used infrequently should show low rates of truth value acceptance. We tested this

[^358]prediction using Pearson correlations, computed separately for containment and support items, between rates of in and on use in the spatial description task and rates of in and on acceptance in the truth value judgment task. The pattern of spatial descriptions and the pattern of truth-value judgments for containment items showed a weak, negative, but reliable correlation ( $r=-.383, n=64, p<.01$ ), while support items showed no reliable correlations between usage and acceptance judgment patterns ( $r=.059, n=64, n s$ ). The weak relationship between participants' truth-value judgments and spatial descriptions aligns with the picture in Figure 3, wherein truth-value judgments show a uniform distribution across scenes, while spatial expression use in descriptions shows a more articulated usage profile.

Our analysis supports a disconnect between participants' judgments about the meaning and felicity of in and on, on the one hand, and their decision to use the expressions to describe containment and support scenes, on the other. While it is clear that meaning and use must be linked in some way (that is, speakers must have implicit knowledge of the meaning of a spatial expression in order to successfully use it to communicate), we suggest that this link is not direct and explore lexical choice - a speaker's decision about which of multiple expressions apply to a given situation - as an intervening process between meaning and use.


Figure 3. Patterns of expression use and truth-value judgments (bottom) of in (left panel) and on (right panel) across a subset of containment and support items.

Exploring lexical choice as an intervening variable between truth value judgments and spatial descriptions:
Stimulus set 2. We collected descriptions and judgments for items in Stimulus set 2 (33 items total). Participants' responses are displayed separately for containment (Figure
4) and support (Figure 5) items, which present patterns of responses from the spatial description task, forced-choice judgment task, and truth-value judgment task. As before, participants produced descriptions with in and on nonuniformly across containment and support items (top panels of Figures 4 and 5) and showed near-uniform truth value judgments across the same items (bottom panels of Figures 4 and 5). Participants' patterns of responses on the forcedchoice judgment task, like their patterns of spatial expression use, showed a non-uniform distribution across items (middle panels of Figures 4 and 5).


Figure 4. Patterns of use (top), forced choice judgments (middle), and truth value judgments (bottom) for in across containment items. Items on the $y$-axis are presented in the same order in all three plots.

Following our previous analysis, we first measured the relationship between participants' average rates of in and on use in their spatial descriptions and acceptance rates in their truth-value judgments. Pearson correlations between language use and truth value judgments were nonsignificant for both containment items ( $r=.306, n=18, n s$ ) and support items ( $r=.159, n=15, n s$ ).

Next, we explored the hypothesis that lexical choice, operationalized here as forced-choice judgments, serves as an implicit process in the generation of spatial descriptions
but not truth-value judgments. Forced-choice judgments were not reliably correlated with truth value judgments for either containment or support. However, forced-choice judgments were strongly related to patterns of in and on use for both containment ( $r=.538, n=18, p<.01$ ) and support items ( $r=.736, n=15, p<.01$ ), suggesting similar variation in speakers' constrained (forced-choice) decisions about which of two expressions best applies to a spatial scene and their unconstrained decisions about how to describe the same spatial scene.


Figure 5. Patterns of use (top), forced choice judgments (middle), and truth value judgments (bottom) for on across support items. Items on the $y$-axis are presented in the same order in all three plots.

## Discussion

In this paper, we compared behavior across three commonly used linguistic tasks in order to examine and elucidate the relationship between judgments of meaning, lexical choice and language use as they apply to spatial terms like in and on. We found that truth-value judgments of the meaning of in and on are nearly uniform across diverse containment and support scenes, demonstrating that these terms are true of the scenes. Speakers' use of these terms, however, is not uniform: some scenes are described more frequently by in and on than others.

Consistent with this distributional difference, we found no strong reliable statistical relationship between judgments of meaning vs. use for the same spatial scenes. However, when we measured judgments of lexical choice (between in and on and other truth-conditionally feasible alternatives), we discovered a non-uniform distribution of choices, similar to the distribution evidenced for spatial expression use. Our analyses confirmed a strong statistical relationship between participants' responses on these tasks across the same set of containment and support scenes. These results support a view of spatial expression meaning as partially distinct from spatial expression use.

## Consequences for the possible meanings of spatial

 expressions. Early accounts of the meaning of terms like in and on (e.g., Bennett, 1975) came under fire (and were subsequently replaced) owing to the underspecified nature of their proposed denotation. The reasoning behind the critical reception of these theories was that a useful definition of a term like in should apply to exactly those cases that we most often use the term for and should rule out cases for which the term is rarely used. However, including a layer of lexical choice in the spatial encoding system, as we suggest here, allows for underspecified meanings that may over-extend to cases where the term is rarely used precisely because other better-suited terms are used in its place. For example, Johannes $(2014,2015,2016)$ suggests an underspecified account of meaning for spatial terms like in and on, whereby speakers' use of these preposition is blocked by the presence of more informative lexical verbs (e.g., hang, attach).
## Consequences for the study of spatial categorization

 through language use. The majority of studies on spatial categorization start by identifying a single form class - for example, prepositions in English - that serves as the primary vehicle for spatial meaning. In contrast to this, the results of the current study suggest that fine-grained spatial categorization is a function of speakers' choices between multiple felicitous expressions and not only dependent on the truth-conditional meaning of a single expression. Thus, future work on spatial categorization should expand the spatial language inventory (beyond e.g., prepositions, see Johannes, Wilson, \& Landau, 2016) and focus on how categories carved out by individual spatial terms may overlap to give rise to a complex graded semantic space for this domain.
## Conclusions

We have demonstrated that, for English, speakers' judgments of the truth-conditional meanings of a spatial term are not necessarily aligned with their use of that term to describe the same spatial scene. We propose that the process of choosing a spatial term among a set of felicitous competitors gives rise to speakers' non-uniform distribution of spatial expression use.

## References

Cruse, A., (2011). Meaning in Language: An Introduction to Semantics and Pragmatics. Oxford University Press.
Feist, M.I. (2000). On in and on: An investigation into the linguistic encoding of spatial scenes. Doctoral dissertation, Northwestern University.
Herskovits, A. (1986). Language and spatial cognition: an interdisciplinary study of the prepositions in English. Cambridge, England: Cambridge University Press.
Jamrozik, A., \& Gentner, D. (2015). Well-hidden regularities: Abstract Uses of in and on retain an aspect of their spatial meaning. Cognitive Science, 39, 1881-1911.
Johannes, K. (2015). Geometric and functional knowledge in the acquisition of spatial language. Doctoral dissertation, Department of Cognitive Science, Johns Hopkins University.
Johannes, K., Wang, J., Papafragou, A., \& Landau, B. (2015). Systematicity and variation in the distribution of spatial expressions in three distinct languages. Proceedings of the $37^{\text {th }}$ Annual Meeting of the Cognitive Science Society, 997-1002.
Johannes, K., Wilson, C., \& Landau, B. (2016). Systematic feature variation underlies adults and children's use of in and on. Proceedings of the $38^{\text {th }}$ Annual Meeting of the Cognitive Science Society, 2429-2434.
Johannes, K., Wilson, C., \& Landau, B. (2016). The importance of lexical verbs in spatial language acquisition: The case of in and on. Cognition, 157, 174189.

Landau, B. \& Jackendoff, R. (1993). 'What' and 'where' in spatial language and spatial cognition. Behavioural and Brain Sciences, 16(2), 217-265.
Landau, B., Johannes, K., Skordos, D., \& Papafragou, A. (2016). Containment and Support: Core and complexity in spatial language learning. Cognitive Science, 40, 1-32.
Levinson, S.C., Meira, S, et al. (2003). 'Natural concepts' in the spatial topological domain - adpositional meanings in cross-linguistic perspective: An exercise in semantic typolgogy. Language, 79, 485-516.
Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., \& Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. Journal of Memory and Language, 40, 230-262.
Regier, T., Khetarpahl, N., \& Majid, A. (2013). Inferring semantic maps. Linguistic Typology, 17, 89-105.
Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In Timothy Shopen (ed.), Language typology and syntactic description (pp. 57-149).
Cambridge: Cambridge University Press.
Vandeloise, C. (2010). Genesis of spatial terms. In V. Evans \& P. Chilton (Eds.), Language, cognition, and space: The state of the art and new directions (pp. 171-192). London: Equinox.
Xu, Y. \& Kemp, C. (2010). Constructing spatial concepts from universal primitives. Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

# Belief Digitization in Economic Prediction 

Samuel G. B. Johnson ${ }^{1}$ \& Faith Hill ${ }^{2}$<br>(sgbjohnson@gmail.com, fahill@vassar.edu)<br>${ }^{1}$ Department of Psychology, Yale University, New Haven, CT 06520 USA<br>${ }^{2}$ Department of Psychological Science, Vassar College, Poughkeepsie, NY 12604 USA


#### Abstract

Economic choices depend on our predictions of the future. Yet, at times predictions are not based on all relevant information, but instead on the single most likely possibility, which is treated as though certainly the casethat is, digitally. Two sets of studies test whether this digitization bias would occur in higher-stakes economic contexts. When making predictions about the future asset prices, participants ignored conditional probability information given relatively unlikely events and relied entirely on conditional probabilities given the more likely events. This effect was found for both financial aggregates and individual stocks, for binary predictions about the direction and continuous predictions about expected values, and even when the "unlikely" event explicitly had a probability as high as $30 \%$; further, it was not moderated by investing experience. Implications for behavioral finance are discussed.


Keywords: Judgment \& decision-making; probabilistic reasoning; explanatory reasoning; behavioral economics.

## Introduction

Investors aim to buy low and sell high. Alas, this adage requires investors to predict the future-a feat known to be difficult for mortals (and even for economists).

People are famously biased in making predictions (Kahneman \& Tversky, 1973), relying on a variety of useful but fallible heuristics. In economic contexts, a particularly worrisome bias would be belief digitization, as found in some other contexts (Johnson, Merchant, \& Keil, 2015; Murphy \& Ross, 1994). That is, when a reasoner is presented with data more consistent with one hypothesis than another, the reasoner acts as though the higher-probability hypothesis is certainly true when making predictions following from the hypothesis.

For example, in one study (Johnson et al., 2015), participants read about a pond that had ecological problems explainable either by an infestation of one type of snail (a simple explanation), or by an infestation of two types simultaneously (a complex explanation). The simple explanation was, reasonably, seen as more likely (about a $66 \%$ chance) than the complex explanation (about 34\%). Yet, when using those explanations to make further predictions (e.g., about bacteria proliferation), people ignored this uncertainty. Manipulating the probability of bacteria proliferation given the simple explanation had a large effect on predictions about proliferation, but manipulating the probability given the complex explanation had no effect at all. People digitized the simple explanation, tacitly assigning $100 \%$ of their probabilistic weight to that possibility. Even though
people often explicitly quantify uncertainty, this uncertainty does not propagate to subsequent computations but is instead rounded, in effect, to 0 or 1.

Such findings pose a challenge to probabilistic theories of cognition that treat humans as Bayesian thinkers who integrate across possibilities rationally (e.g., Anderson, 1991). Nonetheless, in many contexts, this strategy may a reasonably adaptive way to solve an otherwise intractable problem. The inference in this case (from ecological problems to snail infestation to the probability of bacteria proliferation) involves a fairly short chain of reasoning, yet people treated the first step in the inference as certain when making the second step. But we often rely on lengthy chains of reasoning, and propagating uncertainty through the entire chain may well be beyond our cognitive limits. If we must limit the complexity of these computations by prohibiting the consideration of multiple possibilities at each stage (e.g., thinking only about the consequences of the one-snail explanation or the twosnail explanation, but not integrating across both), then it is best to focus on the single most likely possibility. A person could do worse than this kind of belief digitization, even as it leads us astray relative to the optimal answer.

The current studies test whether such a digitization bias would influence judgments in economic contexts. In particular, digitization could affect predictions about future asset prices. Consider the impact of some piece of news, such as information about the government budget. Such information often has uncertain implications for future valuations, so rational investors would assign distributions over these possible futures and value assets according to their expected value. If voters elect a conservative populist (to take an example that is, of course, entirely hypothetical), this introduces uncertainty about the probability of fiscal stimulus. Perhaps there is a $70 \%$ probability of stimulus (with one set of implications for future valuations) and a $30 \%$ probability of fiscal austerity (with a different set of implications). Investors should rationally incorporate both possibilities into their valuations of the market, with a $70 \%$ weight on one possibility and a $30 \%$ weight on the other. Yet, if investors digitize, they would treat the likely event as certain when predicting the future value of the market. Rather than considering both possible futures, they would value assets assuming only the single most likely future.

Although previous studies using non-financial stimuli are consistent with this possibility, it is not clear that digitization effects would generalize to these contexts. First, people are likelier to rely on multiple categories in category-based prediction tasks when the categories are
dangerous or threatening rather than emotionally neutral (Zhu \& Murphy, 2013). If people adopt a more reflective, normative strategy under higher-stakes situations, perhaps they also do so when making economically relevant predictions. Second, and related, people are sometimes more rational when making decisions rather than logically equivalent inferences (Johnson, Zhang, \& Keil, 2016). These two factors could lead people to integrate probabilities across potential futures.

Two sets of studies test whether people nonetheless make digitized predictions in economic contexts. Experiment 1 provides an initial test, asking participants to make probabilistic predictions about the direction of asset prices, given uncertain information. Experiment 2A tests whether digitization effects would occur only for binary predictions (i.e., will a price go up or not?) or would instead extend to predictions of expected value on a continuous scale. Finally, Experiment 2 B tests a possible boundary condition by giving participants explicit posterior probabilities for the market's future direction. After examining these studies individually, we pool the data to examine whether expertise can combat digitization biases. In the General Discussion, we assess the implications of these findings for behavioral finance.

## Experiment 1

Participants in Experiment 1 made predictions about the future prices of financial assets in light of information with uncertain implications. Experiment 1A looked at predictions about market aggregates (e.g., the S\&P 500) and Experiment 1B looked at predictions about individual stocks (e.g., GE). Given that individual stocks seem to be priced more efficiently than the market as a whole (see Shiller, 2005), perhaps digitization mechanisms do not apply as robustly to predictions about individual stocks.

Participants predicted the probability of an increase in an asset price, denoted as $\mathrm{P}(Z)$, based on information about two mutually exclusive possibilities, $A$ and $B$. For instance, $A$ might represent a stimulatory fiscal policy and $B$ an austere fiscal policy. Participants were given information implying that $\mathrm{P}(A)>\mathrm{P}(B)>0$, so that both $A$ and $B$ are possible even as $A$ is likelier-the government may not have made a decision on its fiscal policy, but a stimulus is probable. In addition, participants were given information about the probability of $Z$ conditional on $A$ and $B-\mathrm{P}(Z \mid A)$ and $\mathrm{P}(Z \mid B)$. If people take both more and less likely possibilities into account, then they should rely on both $\mathrm{P}(Z \mid A)$ and $\mathrm{P}(Z \mid B)$ when predicting $\mathrm{P}(Z)$. In contrast, if people digitize, relying only on the single most likely possibility, then only manipulations of $\mathrm{P}(Z \mid A)$ should propagate to predictions of $\mathrm{P}(Z)$.

## Methods

We recruited 200 participants from Mechanical Turk, divided between Experiments 1A and 1B.
Participants each completed three items. For each item, participants read about an uncertain event, where one
possibility ( $A$ ) seemed more likely than the other ( $B$ ), given the available information. These likely and unlikely possibilities differed in their implications for future prices of financial assets. In the high/low condition, the more likely event $A$ would have a high chance of leading to an increase in asset values (i.e., $\mathrm{P}(Z \mid A)$ is high), whereas the less likely event $B$ would have a low chance of leading to an increase (i.e., $\mathrm{P}(Z \mid B)$ is low). One item in the high/low condition of Experiment 1A read:

Imagine that a foreign government is deciding what level of spending to adopt in the next fiscal year.
If they increase public spending, the value of the US stock market is likely to go up.
If they decrease public spending, the value of the US stock market is unlikely to go up.
Suppose that the leader of this government is concerned about the distribution of wealth in the country and is considering increasing public spending.
Participants reading this information should conclude that possibility $A$ (public spending increase) was likelier than possibility $B$ (public spending decrease). For instance, an investor might assign an $80 \%$ probability to possibility $A$ and a $20 \%$ probability to possibility $B$.

Whereas $\mathrm{P}(Z \mid A)$ was high and $\mathrm{P}(Z \mid B)$ was low in the high/low condition, both $\mathrm{P}(Z \mid A)$ and $\mathrm{P}(Z \mid B)$ were low in low/low condition:

If they increase public spending, the value of the US stock market is unlikely to go up.
If they decrease public spending, the value of the US stock market is unlikely to go up.
Rationally, the probability of a price increase is lower in the low/low than the high/low condition, since possibility $A$ has positive (indeed, high) probability of being correct. Thus, both rational and digitizing investors would distinguish between the low/low and high/low conditions.

A third condition, however, generates different predictions for these two groups of investors. In this low/high condition, $\mathrm{P}(Z \mid A)$ is low and $\mathrm{P}(Z \mid B)$ is high:

If they increase public spending, the value of the US stock market is unlikely to go up.
If they decrease public spending, the value of the US stock market is likely to go up.
In this low/high condition, a rational investor would judge the probability of a price increase likelier than in the low/low condition, since possibility $B$ has positive probability (albeit lower than $A$ ). In contrast, if people digitize, tacitly assigning $0 \%$ weight to $B$, then the low/high and low/low conditions would not differ.

After reading each item, participants rated $\mathrm{P}(A)$ and $\mathrm{P}(B)$ (e.g., "Government intends to increase public spending" and "Government intends to decrease public spending") on a 0 to 100 scale. This measure was taken to ensure that people did not explicitly place a $0 \%$ weight on $B$, in which case rational prediction and digitization do
not diverge in their predictions. Further, explicitly quantifying uncertainty in the task produces a task demand to incorporate this uncertainty into predictions, working against our hypothesis.

Finally, on the same page, participants predicted $\mathrm{P}(Z)$ ("What do you think is the probability that the US stock market will go up?") on the same scale used above.

Experiments 1A and 1B differed only in the asset being judged. In Experiment 1A, the asset was the overall value of the US stock market and in Experiment 1B, the asset was the share price for stock in specific corporations.

The three probability conditions were counterbalanced with three different vignettes (one on fiscal policy, one on monetary policy, and one on regulatory policy) using a Latin square. The items were presented in a random order.

After the main task, participants completed 10 check questions and were excluded from analysis if they answered more than one-third incorrectly $(N=19)$. Another 14 participants were excluded because their total probability ratings for at least one item were not between $80 \%$ and $120 \%$. However, including these two types of participants does not alter the pattern of results. Finally, 49 participants were excluded because they did not rate the $A$ more likely than $B$ for at least one of the items, since our predictions are predicated on participants' belief that $\mathrm{P}(A)>\mathrm{P}(B)$. (See Experiment 2 B for a version that did not require the latter two categories of exclusions.)

Table 1: Results of Experiment 1

| Condition |  | Predicted P(Z) |  |
| :---: | :---: | :---: | :---: |
| $\mathbf{P}(\boldsymbol{Z} \mid \boldsymbol{A})$ | $\mathbf{P}(\boldsymbol{Z} \mid \boldsymbol{B})$ | Exp. 1A | Exp. 1B |
| Low | Low | $28.8(28.7)$ | $30.1(27.4)$ |
| High | Low | $73.0(17.7)$ | $75.6(12.8)$ |
| Low | High | $30.3(26.5)$ | $32.3(26.0)$ |

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

## Results and Discussion

As shown in Table 1, participants digitized in both Experiments 1A and 1B.

In both experiments, participants relied on $\mathrm{P}(Z \mid A)$ for in their predictions of future asset prices. The high/low and low/low conditions differed only in $\mathrm{P}(Z \mid A)$, and these conditions differed sharply in predictions $[t(62)=10.38, p$ $<.001, d=1.85$ and $t(54)=10.98, p<.001, d=2.13$ for Experiments 1A and 1B]. Thus, people take account of high-probability possibilities when making predictionsconsistent with both rational and digitizing strategies.

These two strategies differ, however, in their predictions about the low/high condition. This condition differs from the low/low condition only in $\mathrm{P}(Z \mid B)$. Thus, if people take account of less likely possibilities, they should differentiate between these two conditions, but if they digitize, these conditions should be rated similarly.

Supporting the latter possibility, there was no difference between these conditions for either experiment $[t(62)=$ $0.50, p=.62, d=0.06$ and $t(54)=0.47, p=.64, d=0.08$, respectively]. Since we predicted null effects for these comparisons, we also computed Bayes Factors (Rouder et al., 2009; scale factor 1), which strongly favored the null hypothesis $\left[B F_{01}=8.9\right.$ and 8.5 , respectively]. Further, based on participants' other judgments, we can calculate the normative mean difference between the low/high and low/low conditions (which would produce $M \mathrm{~s}=37.6$ and 40.3 for low/high, respectively). In both cases, the actual differences were less than these benchmarks $[t(62)=2.41$, $p=.019, d=0.30$ and $t(57)=1.73, p=.089, d=0.23]$.

Together, these results show that people fail to account for low-probability possibilities when making economic predictions. This was true both when predicting the overall level of financial aggregates as well as the value of stock shares in individual companies.

That said, one may raise some concerns about these results. Perhaps of most concern, the information given in the problem could plausibly have implied near-certainty in its predictions (e.g., "the leader of this government is concerned about the distribution of wealth in the country and is considering increasing public spending"). To assess this possibility, we looked at participants' explicit judgments about $\mathrm{P}(A)$ and $\mathrm{P}(B)$. Unlike their implicit judgments, which assigned essentially a $100 \%$ probability to $A$, participants assigned more reasonable probabilities when asked explicitly ( $83 \%$ and $82 \%$ in Experiments 1A and 1 B , respectively). Nonetheless, we address this concern head-on in Experiment 2B.

## Experiment 2

Experiment 2 examines two possible boundary conditions on belief digitization in economic contexts.

First, Experiment 1 asked for predictions about the probability of binary events (increases or decreases in value). The direction of future gains or losses is likely to be the dominant factor in real investing decisions, but the extent of these predicted gains or losses is also important. In some cases, people are better at reasoning about continuous rather than binary events (e.g., in covariationbased causal reasoning; Alloy \& Tabachnik, 1984). Experiment 2 therefore tests whether digitization effects extend to continuous judgments of expected value.

Second, participants in Experiment 1 arrived at estimates of $\mathrm{P}(A)$ and $\mathrm{P}(B)$ on the basis of other, ambiguous information, as has been the case in most prior work finding digitization effects (Johnson et al., 2015; Murphy \& Ross, 1994). Would such effects occur even when the problem explicitly quantifies the uncertainty? Experiment 2B addresses this question by assigning a $30 \%$ probability to the less likely event. This further rules out the concern that participants may have rationally ignored a low probability. This also addresses the concern that participants in Experiment 1 may have actually assigned extremely low explicit probabilities to the
unlikely events and reported biased explicit judgments due to task demands. In that case, it would not be their implicit judgments that are biased (for interesting reasons) but their explicit judgments (for deflationary reasons).

## Methods

We recruited 200 participants from Mechanical Turk, divided between Experiments 2 A and 2B.

The procedure of Experiment 2 A was identical to Experiment 1A, except that the dependent measure was a continuous price, on either the NASDAQ, DJIA, or S\&P 500 , instead of the probability of a directional change. Participants were given approximately the current value of one of these indices (e.g., "Suppose the current value of the United States stock market, as indexed by the S\&P 500 , is $\$ 2,000$ ") and then asked to predict the future value of that index ("Please estimate what you think the value of the S\&P 500 will be 3 months from today") on a scale ranging from $10 \%$ lower than its current value (e.g., $\$ 1800$ ) to $10 \%$ higher than its current value (e.g., \$2200).

Experiment 2B was identical, except explicit probabilities were given for $A$ and $B$ ("Analysts say there is a $70 \%$ chance that this foreign government will increase public spending, and a $30 \%$ chance that it will decrease public spending") and thus participants were not asked to rate the probabilities of these events.

After the main task, participants completed 10 check questions and were excluded from analysis if they answered more than one-third incorrectly $(N=14)$. Another 7 participants from Experiment 2A were excluded because their total probability ratings for at least one item were not between $80 \%$ and $120 \%$. Finally, 31 participants from Experiment 2A were excluded because they did not rate $\mathrm{P}(A)$ higher than $\mathrm{P}(B)$ for at least one of the items. Since Experiment 2B explicitly provided these probabilities, participants were not excluded for this reason. Analyses including all participants found similar results for both experiments.

Table 2: Results of Experiment 2

| Condition |  | Predicted Change |  |
| :---: | :---: | :---: | :---: |
| $\mathbf{P}(\boldsymbol{Z} \mid \boldsymbol{A})$ | $\mathbf{P}(\boldsymbol{Z} \mid \boldsymbol{B})$ | Exp. 2A | Exp. 2B |
| Low | Low | $-0.21 \%$ | $0.33 \%$ |
|  |  | $(2.96 \%)$ | $(3.44 \%)$ |
| High | Low | $2.86 \%$ | $3.57 \%$ |
|  |  | $(3.50 \%)$ | $(2.89 \%)$ |
| Low | High | $-0.32 \%$ | $0.37 \%$ |
|  |  | $(3.18 \%)$ | $(3.54 \%)$ |

Note. Entries are predicted changes in stock market value. SDs in parentheses.

## Results and Discussion

As shown in Table 2, participants once again digitized.
In Experiment 2A, participants predicted a significantly higher change in asset price in the high/low condition than
in the low/low condition $[t(56)=6.59, p<.001, d=0.95]$. Thus, participants did consider the likely event when making predictions. However, participants again ignored the less-likely event $B$, since they did not use $\mathrm{P}(Z \mid B)$. Predicted changes did not differ across the low/high and low/low conditions $[t(56)=-0.23, p=.82, d=-0.04$, $\left.B F_{01}=9.4\right]$. Further, as in Experiment 1, the difference in predicted changes between the low/high and low/low conditions was marginally lower than it normatively should have been (for a low/high mean of $0.59 \%$ ), based on the other judgments $[t(56)=1.93, p=.059, d=0.26]$. Thus, digitization occurs even for predictions made on a continuous scale rather than probabilities of binary events.
Experiment 2B provided explicit probabilities of $\mathrm{P}(A)$ and $\mathrm{P}(B)$, ensuring that the "unlikely" event $B$ had a rather serious chance of occurring ( $30 \%$ ). Nonetheless, the results are similar to Experiment 2A. While participants again differentiated between the high/low and low/low conditions $[t(90)=7.06, p<.001, d=1.02]$, they did not differentiate between the low/high and low/low conditions $\left[t(90)=0.08, p=.93, d=0.01, B F_{01}=12.0\right]$. Further, the difference between conditions was dramatically lower than it normatively should have been (for a low/high mean of $1.75 \%)$ [ $t(90)=3.41, p<.001, d=0.36]$. Thus, people are willing to ignore even a $30 \%$ probability of an event's occurrence when predicting assets' future value.

One possible objection is that participants may have been giving an appropriate answer, depending on their interpretation of the question. That is, whereas participants' judgments of probabilities in Experiment 1 normatively should accommodate the possibility of lowerprobability events (as is provable from the laws of probability), predictions of future value may be reports of the most likely single value, rather than the expected value. In fact, the single most likely value of the market does depend greatly on $\mathrm{P}(Z \mid A)$, given that $A$ is the single most likely event, but to a much lesser degree on $\mathrm{P}(Z \mid B)$.

However, there are two reasons to doubt this interpretation. First, although the maximum-probability and expected value interpretations of the question are both reasonable, participants would have to uniformly adopt the maximum-probability interpretation to produce our results. That is, if half of participants took the maximumprobability interpretation and therefore did not use $\mathrm{P}(Z \mid B)$ in their predictions, the other half of participants were still making a mistake in failing to use $\mathrm{P}(Z \mid B)$.

Second, even though ignoring $\mathrm{P}(Z \mid B)$ is appropriate in estimating the maximum-probability value of the price, people tend to probability-match rather than maximize in tasks of this sort. For example, suppose there is one button that has a $70 \%$ chance of giving a positive payoff and another button that has a $30 \%$ chance of giving the payoff. If you are supposed to predict which button will produce the payoff on a given trial, the rational thing to do would be to choose the $70 \%$ button every time. In fact, people will predict the $30 \%$ button a significant fraction (roughly $30 \%$ ) of the time. The only way to reconcile this
result with the current task is to assume that participants have tacitly assumed that the $30 \%$ probability event has a $0 \%$ chance of occurrence and can thus be safely ignored.

Overall, Experiment 2 helps to address alternative interpretations of Experiment 1, and shows that people do not need to arrive at estimates of event probabilities themselves in order to digitize them. Together, these two experiments demonstrate that digitization effects may be a pervasive force in investors' judgments of future value.

## Expertise Effects

Amateur investors are often referred to as "noise traders" in financial models (Shleifer \& Summers, 1990) and the behavior of these models depends greatly on these traders' beliefs and choices (Shleifer, 2000). Although professional investors may use different strategies from amateurs (but see Tuckett, 2011), the behavior of amateurs contributes to market dynamics and is therefore important to characterize. Given that our participants are laypeople, but some have investing experience whereas others do not (about half of Mechanical Turk participants own financial assets and about half have taken at least one finance course; Johnson \& Tuckett, 2017), would we see expertise effects within this sample?

Participants in both studies were asked to rate their investing experience and knowledge. If people who have more domain expertise are likelier to consider lowprobability events in making predictions, then the effect of $\mathrm{P}(Z \mid B)$-converted to a $z$-score to aggregate data across studies-ought to be larger for individuals with more experience and knowledge. This was not the case, either for self-reported experience $[r(264)=.02, p=.72]$ or for knowledge $[r(264)=-.02, p=.70]$.

This result, although preliminary, suggests that domain expertise may not be sufficient to overcome digitization effects even in a context like financial prediction that has obvious real-world implications. This does not necessarily undermine the argument often advanced by economists that highly incentivized individuals can avoid such biases, nor the possibility that in market contexts corrective forces can emerge if a subset of investors exploit the suboptimal behavior of others. Nonetheless, this result does suggest that quite extensive expertise-outside the range of experience of our sample-is necessary for such mechanisms to apply. Digitization appears to be a robust cognitive bias at the individual level, and is therefore likely to cause suboptimal performance from investors at a variety of skill levels unless explicitly checked.

## General Discussion

Economic choices, such as investment allocations, depend on our predictions about the future. Rational predictions require us to integrate over multiple uncertain possibilities; failing to do so leads to overconfident predictions that are too near to $0 \%$ or $100 \%$. Yet, participants in our studies consistently failed to account for lower-probability possibilities in making predictions.

Digitization is broadly consistent with conviction narrative theory (e.g., Tuckett, 2011), the idea that decisions in highly uncertain environments are made by constructing a narrative to explain the past, projecting this narrative into the future, and using emotional reactions to the projected narratives to guide choices. For example, amateur investors use company performance news to guide predictions and choices even once the market has had time to "price in" that information, particularly if the news concerns the future rather than the past (Johnson \& Tuckett, 2017). This follows from narrative thinking, since narratives are emotionally valenced and temporally oriented. Another important feature of narrative thinking is that it is linear-it concerns a single sequence of events rather than a web of possibilities. The current work shows that people indeed focus on a single narrative to explain the past and project the future, rather than integrating across multiple possible narratives.

In addition to this theoretical contribution, these results have two kinds of practical implications. First, these biases may persist at the market level, leading to mispricing. A previous study examined explanatory biases in the context of Wall Street Journal headlines (Johnson, 2016). For instance, one headline read "ECB Move Crushes Hopeful Markets." There had recently been a downturn in European markets because the European Central Bank (ECB) had chosen to follow a less inflationary monetary policy than markets had expected. Had investors been "counting on" monetary expansion, tacitly assigning it a $100 \%$ probability? Or had the market priced in this uncertainty already (as mainstream financial theory suggests; e.g., Malkiel \& Fama, 1970)?

Investors made an uncertain diagnosis (the meaning of the ECB chair's statements) and a prediction based on that diagnosis (the implications for monetary policy). Normatively, uncertainty about the interpretation of ECB statements should propagate to any predictions based on such inferences. If the market digitizes at an aggregate level, however, this could have led the market to react strongly to disconfirmed expectations: If the expectations are formed based on uncertain information treated as certain, the market would be overconfident. This could lead prices to be either too high or too low-and indeed to oscillate between those extremes. New information may cause an investor to rationally move from predicting, say, a $70 \%$ probability to a $30 \%$ probability of some event. If these probabilities are treated as $100 \%$ and $0 \%$, respectively, this will lead to a much larger shift in asset valuation than is justified by fundamentals.

That said, such an interpretation of these experimental results is controversial, as are many efforts in behavioral finance to generalize from individual behavior to marketlevel behavior (Shleifer, 2000). A common rejoinder from a neoclassical approach is that behavioral biases can often be neutralized in market contexts. Markets create incentives for accuracy, which are often lacking in behavioral experiments. Markets allow for specialization
so that investors can learn over time to correct their biases (though our expertise analysis suggests that such learning is non-trivial). And perhaps most importantly, selfcorrecting market-level phenomena may emerge. If some, potentially small, subset of investors comes to understand the biases of other investors, they can trade against that bias and capitalize on others' irrationality. Because of these mechanisms, market prices may be less likely to be seriously afflicted by digitization biases than are individual investors' decisions. However, given that stock markets appear to be more volatile than is justified by fundamentals (de Bondt \& Thaler, 1985; Shiller, 1981), digitization of hypotheses could be a partial explanation of this excess volatility. Nonetheless, this issue will not be adjudicated by lab experiments alone.

Second, however, these biases are troubling not only because of potential market inefficiencies they might cause. Even if financial markets do have self-correcting forces that lead experienced investors to profit from the errors of novice investors, the losses of these novices are still cause for concern. Digitized predictions of asset prices can lead to several errors in the investing strategies of amateur "noise traders." First, if one has a high valuation of an asset relative to the market, one may overpay for that asset. For instance, if one is purchasing a house and has an unreasonably high valuation of that house, the buyer may not adequately negotiate the price. Second, extreme asset valuations may potentially lead to suboptimal patterns of diversification. A very bullish assessment of the tech industry accompanied by a very bearish assessment of the financial sector may lead one to prioritize the former over the latter, when a diversified investor would spread her exposure over all sectors. Third, if one's valuations are oscillating faster than the market's valuations, this may lead investors to overtrade, which leads to portfolio value loss due to transaction costs. Finally, overconfident predictions about asset prices may lead investors to inadequately hedge: If the cost of insurance is high relative to the perception of the risk being insured, there is less incentive to insure. This may lead some investors to be overexposed to unexpected downturns in the market-why hedge against something that is deemed, at some level, to be impossible?

Nassim Taleb (2010) warns of "black swans""unknown unknowns" of high impact that we discount on the basis of their low probability. Our participants exemplified this problem, and indeed took it one step further: An event with a $30 \%$ chance is not exactly on the tail of a distribution. Investors would do well to consider all the swans-both black and white.

## References

Alloy, L.B., \& Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. Psychological Review, 91, 112-149.
Anderson, J.R. (1991). The adaptive nature of human
categorization. Psychological Review, 98, 409-429.
De Bondt, W. F. M., \& Thaler, R. (1985). Does the stock market overreact? The Journal of Finance, 40, 793805.

Johnson, S.G.B. (2016). Explaining December 4, 2015: Cognitive science ripped from the headlines. In Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Johnson, S.G.B., Merchant, T., \& Keil, F.C. (2015). Predictions from uncertain beliefs. In Proceedings of the 37th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Johnson, S.G.B., \& Tuckett, D. (2017). Narrative decision-making in investment choices: How investors use news about company performance. In preparation.
Johnson, S.G.B., Zhang, M., \& Keil, F.C. (2016). Decision-making and biases in causal-explanatory reasoning. In Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Kahneman, D. (2011). Thinking, fast and slow. New York, NY: Farrar, Straus, and Giroux.
Kahneman, D., \& Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.
Malkiel, B.G., \& Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25, 383-417.
Murphy, G.L., \& Ross, B.H. (1994). Predictions from uncertain categorizations. Cognitive Psychology, 27, 148-193.
Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., \& Iverson, G. (2009). Bayesian $t$-tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin \& Review, 16, 225-237.
Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? The American Economic Review, 71, 421-436.
Shiller, R. J. (2005). Irrational exuberance (2nd Ed.). Princeton, NJ: Princeton University Press.
Shleifer, A. (2000). Inefficient markets: An introduction to behavioral finance. Oxford, UK: Oxford University Press.
Shleifer, A., \& Summers, L. H. (1990). The noise trader approach to finance. Journal of Economic Perspectives, 4, 19-33.
Taleb, N. N. (2010). The black swan: The impact of the highly improbable (2nd Ed.). New York, NY: Random House.
Tuckett, D. (2011). Minding the markets: An emotional finance view of financial instability. London, UK: Palgrave Macmillan.
Zhu, J., \& Murphy, G. L. (2013). Influence of emotionally charged information on category-based induction. PLoS ONE, 8, e54286.

# Measuring Abstract Mindsets through Syntax: Improvements in Automating the Linguistic Category Model 

Kate M. Johnson<br>University of Southern California

Reihane Boghrati
University of Southern California
Cheryl Wakslak
University of Southern California
Morteza Dehghani
University of Southern California


#### Abstract

The Linguistic Category Model (LCM) was developed as a manual coding scheme for quantifying abstract mindsets in human language. Previous attempts to computationally automate the LCM have relied primarily on pre-coded semantic features, which fail to incorporate important contextual information integral to the LCM coding scheme. In this paper, we introduce Syntax-LCM, a novel method for automating LCM coding using syntax and dependency tree features as predictors of construal level. We compare the accuracy of Syntax-LCM to that of two previously used automated methods: LIWC LCM and Brysbaert concreteness ratings. We find support that the Syntax-LCM approximates the hand-coded LCM with higher accuracy compared to both the Brysbaert and the LIWC LCM. We also provide evidence that the syntactic features accounted for by Syntax-LCM mirror the inclusion criteria in the original coding manual and support theoretical relationships between distance and abstract thinking.


# Iconicity in Word Learning: What Can We Learn from Cross-Situational Learning Experiments? 

Matthew Jones (m.jones.12@ucl.ac.uk)<br>Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)<br>Department of Experimental Psychology, University College London<br>26 Bedford Way, LONDON, WC1H 0AP UK


#### Abstract

Iconicity, i.e. resemblance between form and meaning, is a widespread feature of natural language vocabulary (Perniss, Thompson, \& Vigliocco, 2010), and has been shown to facilitate vocabulary acquisition (Imai, Kita, Nagumo, \& Okada). But what kind of advantage does iconicity actually give? Here we use cross-situational learning ( Yu \& Smith, 2007), to address the question for sound-shape iconicity (the so-called kiki-bouba effect, Ramachandran \& Hubbard, 2001). In contrast to Monaghan, Mattock, and Walker (2012), Experiment 1 suggests that the iconicity advantage comes from referential disambiguation rather than more efficient memory encoding. Experiments 2 and 3 replicate this result, and moreover show that the kiki-bouba effect is roughly equally strong for sharp and rounded shapes, a property that classic experiments were unable to confirm, and which has implication for the effect's mechanism


Keywords: iconicity; cross-situational learning; kiki-bouba, vocabulary acquisition; artificial language learning; soundsymbolism

## Introduction

## Iconicity as Widespread in Natural Language

The meaning of a word does not determine its form, but wordforms are often motivated by iconic relationships with meaning. In English, iconicity can be found in onomatopoeia, (e.g. bang, miaow). Outside the IndoEuropean family, iconicity is more pervasive. Large iconic lexica are reported for many unrelated languages, signed and spoken (see Perniss et al., 2010).

Such iconicity is not limited to sounds. In Japanese reduplication of syllables indicates repetition of an event, and voicing of an initial consonant indicates object size (e.g. gorogoro - heavy object rolling repeatedly; korokoro - light object rolling repeatedly; Perniss et al., 2010).

## Iconicity and Word Learning

Experimental work shows that Japanese sound-symbolic words are easier for 3 -year-olds to learn than non-iconic words, whether the children are Japanese or English speakers (Imai et al., 2008; Kantartzis, Imai, \& Kita, 2011; Yoshida, 2012).

Observational research suggests a role for iconicity in vocabulary acquisition outside the lab. Japanese children acquire iconic words early (Maeda and Maeda, 1983), and in keeping with this Saji and Imai (2013) find that Japanese caregivers use more sound-symbolic and onomatopoeic words speaking to their toddlers than to adults.

Perry, Perlman, and Lupyan (2015), analysing English and Spanish, found a negative correlation between iconicity and age of acquisition: even in Indo-European languages, it may play a role in acquisition.
However, substantial questions remain about what advantage iconicity confers on word learning. Does iconicity kick in after the problem of identifying a word's meaning has already been solved, with iconic words being encoded in memory more quickly or efficiently? Or does iconicity help by facilitating referential disambiguation? Experiment 1 will begin to address this question.

## Sound-Shape Iconicity

A near-universal form of iconicity is the association between certain sounds (e.g. back vowels and high sonority consonants) with heavy, slow, rounded objects; and others (e.g. front vowels and low sonority consonants) with small, quick, jagged objects (Ramachandran \& Hubbard, 2001). In standard demonstrations, participants are given images of two 2 -dimensional shapes, one round, the other spiky. The majority pairs 'kiki' with a spiky shape, and 'bouba' with a rounded shape. (Dingemanse \& Lockwood, 2015).
The mechanism of sound-shape iconicity is, however, uncertain. The effect could arise from correlated input from different sensory modalities. Alternatively, Ramachandran and Hubbard (2001) suggest it is a reflection of cross-modal analogy between the articulatory gestures required to produce the labels and the visual properties of the shapes ( p . 19). They also suggest that 'cross-wiring' (p. 21) of auditory and visual brain maps may create an unmediated link.
Another possible explanation is more literal correspondences between speech sounds and lip shape. 'Bouba' involves literal rounding of the lips - visual or motoric representations of lip rounding could mediate between 'round' sounds and objects. This account predicts an asymmetry: round sound-shape associations should be stronger than spiky ones, because round sounds involve literal rounding of an articulator, whereas spiky sounds do not involve any comparable spikiness. Some prior ERP evidence suggests that the round association may be stronger than the spiky one in processing (Kovic, Plunkett, \& Westermann, 2010 - though their paradigm could not separate the associations behaviourally). This dissociation is not something that the classic kiki-bouba experiment is able to test: with two words and two shapes, one (hypothetically stronger) sound-shape pairing would automatically determine the other (weaker or absent) pairing. However our

Experiments 2 and 3 will represent some of the first work ever to address this question.

## Cross-Situational Learning

Monaghan, Mattock, and Walker (2012) established that the classic kiki-bouba effect is found using the cross-situational learning (CSL) paradigm. CSL takes the form of a series of trials where a word is appears along with a number of possible referents (Yu \& Smith, 2007). Any single trial is ambiguous, and initially participants must guess, but information can be integrated across trials to solve this triallevel referential ambiguity.

Monaghan et al.'s referents were round and spiky shapes, and their names were iconically round or iconically spiky nonwords. Half of shapes received iconically congruent names (e.g. rounded shape-round name), and the other half iconically incongruent (e.g. rounded shape-spiky name). In each trial the participant saw two shapes, heard one name, and indicated which shape the name belonged to. Would accuracy in choosing the correct referent would be higher for congruently named items?

Monaghan et al. found that congruence was no advantage in the first block, but became advantageous in later blocks. Moreover the advantage was only present in trials where the unnamed shape (the foil) was from the opposite category to the target. From the first result they concluded that iconicity indeed supports word learning (perhaps e.g. in the sense of facilitating more rapid or robust memory encoding of iconic names); from the second they that the advantage pertains to category level information, and not to information distinguishing individual words within categories.

These results are somewhat surprising. The classic kikibouba experiment involves guessing names. If iconicity is expressed there, then why wouldn't it be expressed in the first block, when participants are forced to guess namereferent pairings? If that bias were expressed from the start, then iconicity might support referential disambiguation. Experiment 1 takes up this question. Experiments 2 and 3 attempt to tease apart effects of round vs. spiky iconicity.

## Experiment 1

## Methods

Participants 24 adult native English monolinguals (13 women, $M=29.7 \pm 10.0$ ).

Visual Stimuli (Shapes) Sixteen shapes were created using the GNU Image Manipulation Program. Eight 'spiky' shapes were created using randomised parameters. Eight 'rounded' shapes were created by taking each spiky shape and using its corners as fixed points for Bezier curves, then scaled by eye to match for perceived size (see Figure 1). Stimuli were 600*600 pixel images comprising the shape in black on a white background.

Auditory Stimuli (Names) Names were constructed on the basis of LetterScore, a text-based index of sound-shape iconicity: All consonant-vowel pairings in English
orthography that feature consonants with only one canonical pronunciation ( $N=85$; c, g, q, and x were excluded) were rated by monolingual Anglophones who did not participate in other studies ( $N=28,12$ women, $28.5 \pm 12.0$ years old) on a ten-point scale anchored by a circle (1) and a star (10).

Eight of the names were constructed using syllables that received the spikiest ratings (example: tikiza), eight using the syllables that received the roundest ratings (example: mujo). For each category of name, two were one syllable long, four were two, and two were three. Recordings were made by a female native speaker of North American English, pronouncing the words as she considered natural.

Subsequently, word recordings were normed as part of a wider norming study. 101 native English speakers ( $M=32.4$ $\pm 9.7,41$ women) were each given 118 speech tokens to rate (largely from another study), meaning that each speech token was rated about ten times. The study was performed using Qualtrics (2015). In each trial, the participant saw a seven-point ratings scale. ' 1 ' represented the roundest rating, and ' 7 ' the spikiest (counterbalanced for half of participants). The mean of each token's ratings was then taken. This was its WordScore. Names for Experiments 2 and 3 were also rated for WordScore (see below). T-tests confirm that spiky names $(M=4.71 \pm 0.53)$ were rated as significantly spikier by WordScore than round names ( $M=$ $2.90 \pm 0.43)(p<.001, t(13.4)=7.54$, difference $=1.81$, $95 \% \mathrm{CI}$ [1.29,2.33]; Cohen's $d=3.77$ ).

Apparatus and Procedure The study was run using Matlab 7.4.0 on an IBM compatible PC equipped with a 15 " monitor (resolution: $1024 \times 768$ ). For each participant, half the shapes in each category received congruent names (e.g. round names for round shapes). The other half received incongruent names (e.g. spiky names for round shapes). Assignment of names to shapes was counterbalanced between participants.

The experiment took the form of a series of 256 trials, each featuring two shapes on screen (one to the left and one to the right - see Figure 1) and one name (played through headphones). The name belonged to one of the two shapes (this shape was the target, the unnamed shape being the foil). The participants stated which shape the name belonged to (by pressing the left or right arrow). Participants received no feedback and had to guess at first, but in time could infer which name belonged to which shape by noting that each name only consistently appears with one shape.


Figure 1: A cross-situational learning trial (note that names were presented aurally, not in text)

Trials were grouped into four blocks of 64 trials, as in Monaghan et al.. Within each block each name appeared four times, and each shape appeared four times as a target and four times as a foil. The number of times each shape appeared on each side of the screen in each role was counterbalanced, as was the number of appearances by each shape as a foil for a target from its own category vs. the opposite category. The same name was not permitted two trials in a row. Otherwise trials were randomised.

## Results

Trials with reaction times of less than 0.5 seconds or more than 25 seconds were removed.

Statistical Methods Data was analysed using the LMEM package lme4, version 1.1-12 (Bates, Maechler, Bolker, \& Walker, 2015) running in R version 3.2.1 ( R Core Team, 2015). In addition to random intercepts for names and participants, we also included random slopes. We aimed for a design-driven maximal random effects structure (see Barr, Levy, Scheepers, \& Tily, 2013), but were limited in the number of random effects we could fit. For participants we included random slopes for linear block, congruence, category of foil (coded as same or different to category of target), and the congruence-category of foil interaction. For names, we were limited to random effects slopes for congruence, category of foil, and their interaction. Block was coded linearly $(1=-1.5,2=-0.5,3=0.5,4=1.5)$, and both other variables were contrast coded (incongruent $=$ 0.5 , congruent $=0.5$; same category foil $=-0.5$, different category foil $=0.5$ ). Our predictor was accuracy: i.e. whether participants answered correctly on given trials.

Overall Analysis The omnibus model showed reliable effects only of linear block $(\beta=0.84,95 \%$ CI [0.658, 1.022], $z=9.066$ ): participants learned; and congruence ( $\beta=$ $0.417,95 \%$ CI [0.128, 0.706], $z=2.826$ ): participants performed better with congruent names (see Figure 2). The congruence-category of foil interaction was also significant ( $\beta=0.702,95 \% \mathrm{CI}[0.203,1.201], z=2.759$ ): congruence represents more of an advantage when the foil shape is from the opposite category to the target.


Figure 2: Graph of the predictions by block and congruence of the final omnibus model for Experiment 1. Error bars represent 95\% CIs.

See https://github.com/JMJofficial/Jones_Vigliocco_2017 more graphs, and for graphs of Experiments $2 \& 3$.

Block 1 Monaghan et al. found that congruence interacted with block. Crucially, there was no congruence advantage in the first block, implying that the benefit of congruence was to memory encoding rather than kiki-bouba style response bias. By contrast, we found no interaction between congruence and block ( $z<0.7$ ), suggesting an advantage from the first block. To test this, we fitted a model for the first block only. There were reliable effects of congruence ( $\beta=0.328,95 \% \mathrm{CI}[0.047,0.609], z=2.288$ ): performance was better in congruent trials; and of the interaction between congruence and category of foil ( $\beta=0.842,95 \% \mathrm{CI}[0.368$, 1.316], $z=3.484$ ): the benefit of congruence was stronger in different-category-foil trials. Note that this cannot be attributed to differences in design, as the number and structure of our trials was identical.
To exclude the possibility that this is the result of learning within the first block, we took the 187 trials with a different category foil where a participant encountered a name for the first time, and fitted a LMEM featuring only a fixed intercept, and random intercepts by participant. On $56.1 \%$ of trials participants chose the iconically congruent referent for the name. The model's intercept was not reliably different from zero under two-tailed interpretation $(\beta=0.247,95 \%$ CI [-0.048, 0.549], $z=1.678$ ), but under a one-tailed interpretation, the intercept was significantly different from zero at $p=.047$. Thus though this analysis has low power, it suggests a sizable bias towards iconic matches before learning has taken place, which can only be explained by the bias/referential disambiguation account.

In conclusion, we largely replicated Monaghan, Mattock, and Walker's (2012) findings, but found an advantage of iconicity from the first block. This difference with Monaghan et al. - and the fact that iconic congruence is only an advantage when the foil is from the opposite category - is consistent with the possibility that iconicity biased participants towards the right answer in trials where they were forced to guess, effectively assisting with referential disambiguation. The discrepancy with Monaghan et al. may be due to name stimuli: while we tailored ours to maximize iconicity, they created theirs on the basis of phonetic features, which do not correlate perfectly with iconicity (e.g. Monaghan et al. used plosives as spiky sounds, but [b] - a plosive - is widely deemed to sound round, cf. bouba). Next we move on to two further experiments aimed at testing the relative contribution of roundness and spikiness to sound-shape iconicity.

## Experiments 2 and 3

Experiment 2 and 3 aim to clarify the mechanism of soundshape iconicity by modifying Experiment 1 in order to test
the effect of round-to-round and spiky-to-spiky iconicity separately, something previous experiments have been unable to do. This is achieved by using iconically neutral names as well as round and spiky names. Experiment 2 yielded marginally significant results, so Experiment 3 was a replication to attempt to verify whether the effect was real. Both were then submitted to omnibus Bayesian statistics.

We opted for a two-condition design. Each condition is of the same format as Experiment 1, and each features both round and spiky shapes, but one condition features round and neutral names only, the other features spiky and neutral names only, thus avoiding problems related to tasks that involve discrimination between a round and spiky alternative. If one class of name is less iconic then we would expect minimal benefit of one class of shape being paired with that class of name vs. a neutral name.

## Experiment 2: Methods

Participants were 32 adult native English monolinguals (17 women, $M=23.3 \pm 4.4$ ).

Visual Stimuli (Shapes) The eight round and eight spiky shapes used in Experiment 1 were combined with an additional eight of each, created in the same manner.

Auditory Stimuli (Names) 32 names were generated using previously normed syllables (see Experiment 1) - eight from round syllables, eight from spiky, and 16 from neutral; and recorded as in Experiment 1. T-tests confirm that the spiky names $(M=4.71 \pm 0.53)$ were rated as spikier than the round names $(M=2.90 \pm 0.43)(p<.001, t(13.4)=7.54$, difference $=1.81,95 \%$ CI [1.29, 2.33]; Cohen's $d=3.77$ ). Moreover, neutral names $(M=3.77 \pm 0.80)$ were rated as less spiky than spiky names $(p=.002, t(20.0)=3.46$, difference $=0.94,95 \%$ CI $[0.37,1.51]$; Cohen's $d=1.31$ ), and less round than round names $(p=.002, t(21.8)=3.46$, difference $=1.04,95 \%$ CI [0.35, 1.39]; Cohen's $d=1.24$ ).

Apparatus and Procedure Every participant took part in a round and a spiky condition. Each condition was of identical form to Experiment 1 . One of the two conditions was the 'round' condition. In this condition half of the shapes were round and half spiky (eight of each), and, crucially, half of the names were round and half neutral. The other condition was the 'spiky' condition - which again had eight round and eight spiky shapes, but by contrast had eight neutral names and eight spiky names (fresh shapes and neutral names were used in the second condition). Shapes, neutral names, and condition order were counterbalanced across participants.

Here congruence is defined within whichever half of the putative round-spiky spectrum of sounds the condition in question covers. E.g. in the round condition, round nameround shape pairings were considered congruent and round name-spiky shape pairings were considered incongruent. However, neutral name-spiky shape pairings were considered congruent for the purposes of the following analysis. The reverse was done for the opposite condition.

This format was so that we could apply the same kinds of analysis as for Experiment 1 to keep results comparable.

## Experiment 2: Results

Data were analysed as in Experiment 1. The additional variable of condition was coded Round $=-0.5$, Spiky $=+0.5$.

In the omnibus model, both linear ( $\beta=0.722,95 \%$ CI [0.619, 0.824], $z=13.835$ ) and quadratic block $(\beta=-0.11$, $95 \%$ CI $[-0.179,-0.041], z=-3.116$ ) were reliable predictors: performance improved over the blocks, with improvement being faster between early than late blocks. Category of foil was also a reliable predictor ( $\beta=0.162$, 95\% CI [0.015, 0.309], $z=2.163$ ): performance was better on trials with foils from the opposite category to the target. Finally, the interaction between congruence and category of foil was reliable ( $\beta=0.286,95 \%$ CI $[0.013,0.56], z=2.05$ ): performance was better on congruent trials as long as the target and foil were from different categories (the main effect of congruence was not reliable: $\beta=0.027,95 \%$ CI [$0.14,0.194], z=0.317$; perhaps because iconic contrast was less pronounced than in Experiment 1).

The crucial interaction between condition and congruence was marginally reliable in the expected direction, implying that congruence was more of an advantage in the round condition ( $\beta=-0.208,95 \%$ CI $[-0.482,0.066], z=-1.486$ ): an inconclusive result (though the same was not true for the three way interaction adding category of foil: $\beta=-0.047$, $95 \%$ CI [-0.377, 0.283], $z=-0.277$ ).

As with Experiment 1, we analysed Block 1 in isolation. There was a reliable effect of condition $(\beta=0.177,95 \%$ CI [0.001, 0.354], $z=1.967$ ): performance was better in the spiky condition. Though there was no overall effect of congruence ( $z<1.3$ ), there was a reliable effect of category of foil ( $\beta=0.173,95 \%$ CI [0.022, 0.325], $z=2.246$ ): performance was better when the foil and target were from different categories, and a reliable interaction between congruence and category of foil ( $\beta=0.39,95 \%$ CI [ 0.105 , $0.675], z=2.679$ ): that congruence was advantageous when foil and target were from different categories. Given that previous results suggest that the congruence advantage is in different-category-foil trials, this means that as in Experiment 1, and in contrast to the results of Monaghan et al., iconic congruence was an advantage from the outset. Finally, there was a reliable interaction between congruence and condition: for Block 1 (as was marginally the case for the omnibus model), the effect of congruence was stronger in the round condition ( $\beta=-0.315,95 \%$ CI $[-0.628,-0.002]$, $z=-1.975$ ).

To summarise, Experiment 2 largely replicated the results of Experiment 1. Additionally, we did not find an unambiguously reliable difference between round and spiky conditions in terms of iconicity advantage. However, the marginally reliable interaction is possible evidence for round iconicity being stronger. Experiment 3 is a nearreplication aimed at clarifying this.

## Experiment 3: Methods

Participants were 32 adult native English monolinguals (21 women, $M=21.8 \pm 3.2$ ).

Visual Stimuli (Shapes) Were as Experiment 2.
Auditory Stimuli (Names) A fresh set of 32 names (eight round, eight spiky, and 16 neutral) were generated as in Experiment 2. An additional factor was controlled: number and distribution of phonemes in each category of name. Names were recorded as in Experiments 1 and 2.

T-tests confirm that the spiky names $(M=4.80 \pm 0.39)$ were rated for WordScore as spikier than the round names $(M=2.90 \pm 0.63)(\mathrm{p}<.001, t(11.7)=7.19$, difference $=$ 1.90, $95 \%$ CI [1.32, 2.47]; Cohen's $d=3.60$ ). Moreover, neutral names $(M=3.80 \pm 0.86)$ were rated as less spiky than spiky names $(p<.001, t(22.0)=3.91$, difference $=$ $1.00,95 \% \mathrm{CI}[0.47,1.53]$; Cohen's $d=1.35$ ), and less round than round names $(\mathrm{p}=.01, t(18.5)=2.89$, difference $=0.90$, $95 \%$ CI [0.25, 1.55]; Cohen's $d=1.24$ ).

Apparatus and Procedure Were as in Experiment 2.

## Experiment 3: Results

Data were analysed as in Experiment 2. The omnibus model featured reliable effect of linear block ( $\beta=0.739$, $95 \%$ CI [0.589, 0.889], $z=9.641$ ): participants learned. However, it featured no other significant predictors $(|z|<2.0)$. In this respect, it was different from Experiments 1 and 2, both of which showed some advantage of congruence. However, the coefficients for both congruence ( $\beta=0.177,95 \%$ CI [$0.003,0.357], z=1.929$ ) and the congruence-category of foil interaction ( $\beta=0.178$, $95 \%$ CI $[-0.081,0.437], z=$ 1.346) were in the expected direction, with congruence qualifying as marginally reliable. Crucially, the congruencecondition interaction did not approach reliability ( $\beta=$ $0.047,95 \%$ CI $[-0.326,0.233], z=-0.327)$, and neither did the interaction between congruence, category of foil, and condition ( $\beta=-0.147,95 \%$ CI $[-0.464,0.17], z=-0.908$ ) suggesting that if there was an effect of congruence, it was no stronger in the round condition.

Again we analysed Block 1 in isolation. This time there were no reliable predictors ( $z<1.3$ in each case). However note that this parallels the omnibus model, which featured no reliable predictors except block. Thus these results are silent on the question of the nature of the iconic advantage as the advantage failed to show up overall (probably due to a smaller iconic differences between words than in Experiment 1 leading to a weaker effect and Type II error).

Thus Experiments 2 and 3 gave somewhat contradictory results, with Experiment 2 showing a marginally reliable interaction between condition and congruence in the expected direction, and Experiment 3 showing no such thing. To attempt to resolve this, we submitted both sets of results to Bayesian statistics, which have the capability to confirm the null, and make it unproblematic to add more data to an analysis as one goes along (Kruschke, 2011).

Models Not having a clear prior for the alternative hypothesis, we opted for Bayesian parameter analysis (Kruschke, 2011). We use the R package rstanarm (Gabry \& Goodrich, 2016). We examined 95\% Highest Density Intervals for parameter estimates (HDIs): the highest average density continuous interval containing $95 \%$ of posterior probability distribution. If this region excludes zero we can treat a predictor as reliable.

We based our priors on Gelman, Jakulin, Pittau, and Su's (2008) recommendations. All variables were centred at zero and scaled so as to have a standard deviation of 0.5 . Priors (which were defined for the $\log$ odds ratios used as the models' parameters rather than for raw probabilities) took the form of Cauchy distributions.

Models were similar to the models used for Experiments 2 and 3, but a predictor and a by-subjects random slope were added for condition order. All two- and three-way interactions were included.

Results There were credible effects of linear $(\beta=1.638$, $95 \%$ HDI $[1.45,1.835]$ ) and quadratic ( $\beta=-0.155,95 \%$ HDI $[-0.255,-0.06]$ ) block, condition order $(\beta=0.472,95 \%$ HDI $[0.322,0.620])$, and category of foil ( $\beta=0.091,95 \%$ HDI [0.019, 0.164]): participants performed better on trials where the target and foil shapes came from different categories. The HDIs for the main effect of congruence encompass zero ( $\beta=0.087,95 \%$ HDI [-0.022, 0.198]), albeit narrowly. However, there is a credible interaction between congruence and category of foil ( $\beta=0.242,95 \%$ HDI [0.114, 0.369]), indicating that an advantage for congruence is present when the target and foil are from different categories. There were interactions between condition order and both linear ( $\beta=0.329,95 \%$ HDI [0.196, $0.462]$ ) and quadratic ( $\beta=-0.213,95 \%$ HDI [-0.333, $0.094]$ ) block, indicating that initial learning was faster in the second condition. There was also a difficult-to-interpret interaction between quadratic block, condition order, and congruence ( $\beta=-0.253,95 \%$ HDI $[-0.498,-0.009]$ ). However, overall, the Bayesian analyses confirm the earlier inferential statistics.

Turning to the crucial interaction of congruence with condition type: the posterior distribution for the interaction between congruence and condition type is narrow compared to the prior, and centred close to zero ( $\beta=-0.022,95 \%$ HDI $[-0.165,0.125])$. If we assume the largest absolute value in the HDI, and take the intercept as our baseline, the difference between the levels of the interaction is $84.2 \%$ versus $85.3 \%$. This is the same as the difference when the main effect of congruence is examined in the same way, assuming the mean of the posterior.

Even if we assume that the extreme values of the HDI are correct, the effect of the congruence-condition interaction is no bigger than that of congruence tout court (which applies to both conditions). Thus there is clearly less support for the interaction between congruence and condition type than for the effect of congruence across conditions, and given the

## Experiments 2 and 3: Bayesian Analysis

HDIs encompass zero, our results are consistent with their being no congruence-condition interaction.

## Discussion

We presented evidence that iconicity enhances performance in a statistical learning paradigm. Experiment 1 was a replication of Monaghan et al. (2012), and thus in a sense not novel, but (as small but theoretically interesting differences in our respective results underscore) the value of replication is increasingly recognised in cognitive science. Close analysis of the beginning of Experiment 1 (supported by the results of Experiment 2), and the consistent tendency for iconicity to be a greater advantage when the foil presented during the trial does not also match the name, suggest that the benefit of iconicity in these experiments is to do with picking out the right referent during a particular trial rather than in to do with learning in some other sense (contra Monaghan et al.). Thus one role of iconicity in vocabulary learning may be in referential disambiguation, in line with evidence that people guess iconic word meanings in unfamiliar languages above chance (Imai et al., 2008).

The second set of findings relate to the relative importance of rounded-rounded and spiky-spiky mappings in sound-shape iconicity. One hypothesis is that rounded sounds are associated with rounded lip shape, and that the iconicity arises from sound-shape correspondences during speech production and comprehension (Ramachandran and Hubbard, 2001). If this is indeed the mechanism for soundshape iconicity, we would expect rounded associations to be primary, and spiky associations to arise later through something like a principle of contrast. If this were the case then rounded associations should be stronger than spiky associations (as suggested in Kovic. et al., 2010).

Experiment 2 and 3 tested this possibility by separating round and spiky iconicity into two separate conditions, and seeing whether iconic congruence exerted a stronger effect in one or the other. Experiment 2 appeared to suggest (with marginal reliability) that iconicity improved performance in the round condition, but not the spiky. However, this asymmetry failed to replicate in Experiment 3. We therefore submitted data from both experiments to a Bayesian analysis. Though the results were somewhat inconclusive, they suggest that any asymmetry between the conditions is a smaller effect than the overall influence of iconic congruence, and indeed they are consistent with there being no asymmetry at all. However, this may be different in the case of production, which would force motoric and perceptual engagement with lip shape (Jones et al., in prep.).

Our results advance our understanding of iconicity's role, suggesting it supports referential disambiguation.

## References

Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Gabry, J. and Goodrich, B. (2016). rstanarm: Bayesian Applied Regression Modeling via Stan. R package version 2.9.0-3
Gelman, A., Jakulin, A., Pittau, M. G., \& Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2(4), 1360-83.
Imai, M. Kita, S., Nagumo, M., and Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109, 54-65.
Jones, M., Vinson, D., Clostre, N., Lau Zhu, A., Santiago, J., and Vigliocco, G. (in prep.). Iconicity emerges in a model of language change.
Kantartzis, Imai, \& Kita (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. Cognitive Science, 35, 575-586.
Kovic, V., Plunkett K., and Westermann, G. (2010). The shape of words in the brain. Cognition, 114, 19-28.
Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. Perspectives on Psychological Science 6(3), 299-312.
Lockwood, G., \& Dingemanse, M. (2015). Iconicity in the lab: a review of behavioural, developmental, and neuroimaging research into sound-symbolism. Frontiers in Psychology, 6(1246).
Maeda, T., and Maeda, K. (1983). Yoji no goihattatsu no kenkyu [Investigation of a Child's Lexical Development]. Tokyo: Musashino Shoin.
Monaghan, P., Mattock, K., \& Walker, P. (2012). The role of sound symbolism in word learning. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 38(5), 1152-1164.
Perniss, P., Thompson, R. L., \& G. Vigliocco (2010). Iconicity as a general property of language: evidence from spoken and signed languages. Frontiers in Language Science, l(227).
Perry, L. K., Perlman, M., Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. PLoS 1, 10(9): e0137147. DOI:10.1371/journal.pone. 0137147
Ramachandran, V., \& Hubbard, E. (2001). Synaesthesia: A window into perception, thought, and language. Journal of Consciousness Studies, 8(1), 3-34.
Saji, N. \& Imai, M. (2013). Onomatope kenkyu no shatei chikadzuku oto to imi [Sound Symbolism and Mimetics]. In K. Shinohara \& R. Uno (Eds.), Goishutoku ni okeru ruizousei no kouka no kentou (pp. 151-166). Tokyo: Hituji Syobo.
Yu, C., \& Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(15), 414-420.

# Mindfulness and Fear Conditioning 

Fergal W. Jones (Fergal.Jones@canterbury.ac.uk)<br>Salomons Centre for Applied Psychology, Canterbury Christ Church University, Runcie Court, Broomhill Road, Tunbridge Well, Kent TN3 0TF, UK

Rossy P. McLaren (R.P.McLaren@exeter.ac.uk)<br>Department of Psychology, College for Life and Environmental Sciences, University of Exeter, Washington Singer Building, Perry Road, Exeter EX4 4QG, UK

Ian P. L. McLaren (I.P.L.McLaren@exeter.ac.uk)
Department of Psychology, College for Life and Environmental Sciences, University of Exeter, Washington Singer Building, Perry Road, Exeter EX4 4QG, UK


#### Abstract

During mindfulness-based interventions participants can be invited to bring aversive stimuli to mind while practicing mindfulness. This is thought to help the stimuli become less aversive. However, the mechanisms underlying this process are not fully understood. In this study we explored these by examining the effects of mindfulness practice and stimulus visualization on stimuli associated with electric shocks. Participants were trained on a discrimination between two visual stimuli using a standard electrodermal conditioning procedure, in which one stimulus (CS+) was paired with shock and the other (CS-) was not. They then visualized either the CS+ or CS-, while practicing mindfulness or performing a control activity. Following a number of extinction trials, the impact of these manipulations was assessed during a reacquisition test-phase. Both mindfulness and visualization of the CS+ led to slower reacquisition of the CS+/shock association, when measured physiologically, and their effects were additive. Moreover, these effects dissociated from participants' expectancy of shock. If confirmed in future work, these findings may have implications for the treatment of stimulus-specific anxiety.


Keywords: mindfulness, associative learning, extinction, reacquisition

## Introduction

In recent years there has been a rapid growth of interest in mindfulness (e.g. Mindfulness All-Party Parliamentary Group, 2015). This has been driven in part by the growing evidence for the efficacy of mindfulness-based therapeutic interventions, such as mindfulness-based cognitive therapy (Segal, Williams \& Teasdale, 2013), which has been shown to reduce the risk of relapse of depression relative to treatment as usual and more active controls (Kuyken et al., 2015). However, much remains to be understood about the nature and mechanisms of action of mindfulness (Tang, Holzel \& Posner, 2015; van der Velden et al., 2015). In this paper, we attempt to further this understanding in one particular area, namely how mindfulness interacts with basic human learning processes (cf. Treanor, 2011). To set the stage for this, it is helpful to first consider mindfulness in more detail.

## Mindfulness

Kabat-Zinn's (1994) frequently cited definition of mindfulness describes it as 'paying attention in a particular way: on purpose, in the present moment, and nonjudgmentally'. Mindfulness meditation practice is seen as a means of cultivating this way of attending. In a typical practice, 'mindfulness of the breath', participants are invited to pay attention to and be curious about their moment-bymoment experience of breathing, and to be gentle with themselves should their attention wander away from this (Kabat-Zinn, 1990).

In mindfulness-based interventions, such as mindfulnessbased cognitive therapy, after participants have developed some experience at practicing mindfulness, they are invited to deliberately bring attention to a difficult experience during mindfulness practice (Segal et al., 2013). Frequently this can be a memory or image associated with feelings of anxiety and/or low mood. This is thought to help participants learn to not engage in unhelpful rumination and worry when faced with a difficulty (cf. Segal et al., 2013), and to help them to build their ability to tolerate distress (cf. Lotan, Tany \& Bernstein, 2013). In addition, theories of associative learning would suggest that basic associative learning processes should be in play (cf. Treanor, 2011). However, the latter aspect has yet to be adequately investigated empirically. A laboratory model that can be used to examine such processes further is the electrodermal fear conditioning paradigm.

## Electrodermal Fear Conditioning

In this paradigm, a neutral stimulus, referred to as the conditioned stimulus (CS), becomes capable of eliciting fear through its repeated pairing with an aversive unconditioned stimulus (US), such as electric shock (see McAndrew et al., 2012 for details of the procedure used here). Participants' learning of this CS-US association is typically measured in two ways. The first is through their 'conditioned response' (CR), which in this case usually includes increased arousal following the CS, due to increased anxiety at the prospect of being shocked. This can be detected by measuring the

Table 1: The study's design

| Group | Training | Manipulation | Extinction | Reacquisition |
| :---: | :---: | :---: | :---: | :---: |
| MV+ | A+ B- | Mindfulness \& A | A- B- | A+ B- |
| MV- | A+ B- | Mindfulness \& B | A- B- | A+ B- |
| CV+ | A+ B- | Control \& A | A- B- | A+ B- |
| CV- | A+ B- | Control \& B | A- B- | A+ B- |

conductance of electricity between two electrodes on the skin, and is typically referred to as the skin conductance response (SCR). Secondly, when presented with the CS, participants can be asked to rate how much they expect a shock.
Under certain conditions, dissociations can be obtained between SCRs and expectancy ratings (e.g. Knight, Nguyen, \& Bandettini, 2003; Tabbert, Stark, Kirsch, \& Vaitl, 2006; McAndrew et al., 2012). These can be explained by dual process models of human learning, such as that proposed by McLaren, Green and Mackintosh (1994), and further developed in McLaren, Forrest, McLaren, Jones, Aitken and Mackintosh (2014). They argue that people can learn using both associative processes, which are similar to those found in other animals, and through rule-based processes capable of symbolic manipulation (though Lovibond and Shanks, 2002 take an alternative view). Thus, as well as having the potential to provide a useful laboratory model to investigate mindfulness further, the fear conditioning paradigm enables an examination of the degree to which mindfulness interacts differently with different learning processes.

## The current study

Therefore, in the current study we aimed to embed mindfulness practice in a human electrodermal, fear conditioning procedure. More specifically, we planned to train people to learn a 'CS'-> shock association (along with an appropriate comparison) before inviting them to practice mindfulness while visualizing the CS , in a similar way to how distressing events can be brought into attention during mindfulness practice. We then planned to examine what effects, if any, this mindfulness visualization had on the learning of the CS -> shock relationship, relative to various comparison groups. Furthermore, we sought to examine whether such mindfulness-based visualization had differential effects on the different learning processes tapped by SCR and expectancy. We hoped that this investigation would provide us with a better understanding of the learning processes in play during mindfulness practice, which in turn could help contribute to improving the efficacy of mindfulness-based interventions for anxiety.

## Method

## Participants

Ninety-six University of Exeter students participated in this experiment. There were 72 women and 24 men and their
ages ranged from 18 to 30 years, with a mean of 20.4 years ( $\mathrm{SD}=2.75$ ). All were paid $£ 6$. Participants were randomly allocated to the groups described below, constrained to ensure equal group sizes ( $\mathrm{N}=24$ in each group). The study received ethical approval from the University of Exeter, Psychology Ethics Committee.

## Design

The study began with a training phase during which all the participants learnt an $\mathrm{A}+\mathrm{B}$ - discrimination; that is, they learnt that one conditioned stimulus (the CS+ ), which we will refer to as A, was always followed by an electric shock, while a second, the CS- (B), never was.

Following this, each of the four groups received a different manipulation, as illustrated in Table 1. In the mindfulness visualization plus (MV+) condition, participants were invited to visualize the CS (i.e. A) that had been previously paired with shock while they practiced mindfulness. In the mindfulness visualization minus (MV-) condition, they visualized the unpaired CS (i.e. B) while practicing mindfulness. The two control conditions were identical to the mindfulness ones, with the exception that, instead of practicing mindfulness, participants were asked to listen to an excerpt from an audio book. Thus, in the control visualization plus ( $\mathrm{CV}+$ ) condition, participants listened to the audiobook while visualizing stimulus A , and in the control visualization minus (CV-) condition, participants listened to the audiobook while visualizing stimulus B.

Following this manipulation, all the groups received an extinction test phase, during which A and B were presented but neither were paired with shock. This was followed by a reacquisition test phase, during which all participants were again trained on the original $\mathrm{A}+\mathrm{B}$ - discrimination; that is, A was once again followed by a shock, whereas $B$ was not. This second test phase was included, as pilot work suggested it was more sensitive to the effects of the manipulation. This may be because extinction happens relatively quickly in this paradigm, resulting in very few trials that provide useful data from the extinction test phase.

Therefore, there were three independent variables; namely practice type (mindfulness vs. control), stimulus visualized (A vs. B), and stimulus tested (A vs. B), with the former two being between-subject factors and the later being a withinsubject factor.

## Stimuli

The two CSs were a brown cylinder $4.5 \times 6 \mathrm{~cm}$ onscreen, and a pink square $5.5 \times 5.5 \mathrm{~cm}$ onscreen. Each CS presentation lasted for 5 seconds. The use of these stimuli as CS A and CS B was counterbalanced across participants.
The unconditioned stimulus (US) was a 500 ms shock administered with a PowerLab 26T generator using stainless steel electrodes attached to the left proximal and medial phalanges of the index finger. At the beginning of the experiment, participants set their own shock level between 5 and 20 mA , to a level that was "definitely uncomfortable but not painful".

## Visualization Guidance

Mindful Visualization This period of practice began with 7.5 minutes of 'mindfulness of the breath' guided by audio CD. This practice was of the sort used in mindfulness-based cognitive therapy (Segal et al., 2013), and the CD had been recorded by the first author (who has a postgraduate qualification in teaching mindfulness). The intention of this initial period of practice was to help establish a more mindful state of mind before the participant began stimulus visualization. After 7.5 minutes, the CD guidance asked the participant to open their eyes, if they were closed, and look at the name (either 'pink square' or 'brown cylinder', as appropriate to the condition) that the experimenter had placed in front of them. The audio CD then invited them to close their eyes and continued as follows:
'... as best you can, remembering what this shape looked like in the first part of the study, and seeing if it is possible to hold an image of this shape in your mind. [pause] Don't worry if you find it hard to picture this shape, as this is more difficult for some of us than others, [pause] just doing your best to bring to mind whatever memories and images you have of this shape from the first part of the study. [pause] And if it seems to help, you might want to sometimes say the name of the shape to yourself. [pause] And if there are any feelings or bodily sensations that accompany the memory or image of the shape, just acknowledging those and allowing them to be present as you hold this shape in mind. [pause] And if at any point you forget which one of the shapes you are being invited to hold in mind, opening your eyes again briefly and re-reading the piece of paper.'

The CD guidance subsequently invited participants to expand their attention so that they both held in mind an 'image or memory' of the shape and attended to 'the experience of breathing'. This was followed by periods of silence, interspersed with guidance to the same effect. In total, the audio CD lasted 13 minutes 20 seconds.

Control Visualization In the control visualization conditions, participants were asked to listen to an excerpt from an audiobook by Bill Bryson. This material was chosen as the calm tone of delivery was similar to that for the mindfulness visualization guidance and the content was likely to be experienced as engaging but uncontroversial. As
with the mindfulness visualization conditions, after 7.5 minutes participants were asked to look at the name of the shape that the experimenter had placed in front of them, and then hold an image of this shape in mind. The wording and timing of the stimulus visualization instructions were identical to the mindfulness visualization conditions, with the exception that guidance pertaining to attending to the breath was omitted. In between visualization guidance, the audiobook continued to play. Each control visualization condition lasted for the same amount of time as each mindfulness visualization condition.

## Measures

Skin conductance Skin conductance response (SCR) was measured using LabChart software via MLT116F GSR electrodes attached to the medial phalanges on the left third and fourth fingers.

Expectancy Expectancy ratings for the US were recorded using a Contour Shuttle Xpress device. Participants were required to make an expectancy rating about the extent they thought the shock would happen during presentation of the CS. The device had five buttons and fitted nicely into one hand such that one button corresponded to one finger. The different expectancy values were: 1 "There will definitely not be a shock", 2 "There might not be a shock", 3 "Not sure either way", 4 "There may be a shock", and 5 "There will definitely be a shock". A continuously available legend explained which buttons represented which ratings.

State mindfulness As a manipulation check, the State Mindfulness Scale (Tanay \& Bernstein, 2013) was administered to all participants immediately after the visualization stage. This 21 -item, self-report measure asks participants to rate how well each item (e.g. 'I felt closely connected to the present moment') describes their experience over the past 15 minutes. It has satisfactory psychometric properties, for example Cronbach's $\alpha=.95$ (Tanay \& Bernstein, 2013). Higher scores indicate higher levels of state mindfulness. Therefore, if the mindfulness visualizations were successful at inducing a more mindful state of mind than the control visualizations, participants in the former conditions should score significantly more highly on this measure than those in the latter.

## Procedure

With the exception of the visualization stage (see below), the participants were told they would receive shocks to some of the visual stimuli throughout the experiment. They were asked to rate their expectancy that the shock would occur during each stimulus presentation, using the Shuttle Xpress device. Otherwise they were asked to remain still to avoid motion artefacts in the SCR. On shock (A+) trials, a 500 ms US was administered after 4500 ms of CS A being on screen, whereas on no shock (B-) trials no US occurred.

SCR recordings were taken on every trial, during the five seconds prior to CS onset (Pre-CS), five seconds while the

CS was on screen and five seconds after the CS (Post-CS). The inter-trial interval (ITI) was randomly varied between 30 and 40 seconds in order to stop participants timing the onset of the CS. Long ITIs were required to allow the SCR recording to reach baseline after the previous US.

This experiment had an initial training phase of 12 trials: six each of $\mathrm{A}+$ and B -, in a random order. This was followed by either the mindfulness visualization guidance or control visualization guidance, as appropriate for the condition. Participants were advised that there would be no stimuli on the screen and no shocks, during this stage of the experiment. Participants were then given a copy of the state mindfulness scale, with its name removed. They were asked to tell the experimenter their rating of each statement, rather than write it, so that they could keep their hands still and remain wired to the electrodes.

After this, participants were advised that the stimuli would be appearing on the screen again, and that some of these would be accompanied by shock. They were asked to rate their expectancy of shock in the same manner as previously. There were four extinction trials, in the order A, $\mathrm{B}, \mathrm{B}, \mathrm{A}$, (counterbalanced) during which no shocks were administered. These were followed immediately by eight trials of re-acquisition, comprising four A+ trials and four $B$ - trials, in a random order. During this reacquisition phase, CS A was accompanied by shock and CS B was not, in exactly the same manner as during the initial training. Participants were then asked whether they had previous experience of mindfulness practice. The word mindfulness had not been used up to this point, in case it influenced participants' responding, given the frequency with which this topic is currently covered in the UK mainstream media. Finally, participants were debriefed, thanked for their time and paid.

## Data Preparation

The SCR data were recorded in micro-Siemens in LabChart and exported to Excel. For each trial, a mean SCR was calculated for both the 'pre-CS' and 'CS' periods. These data were then transformed using a log transformation to reduce the variability between participants. In order to measure the change in SCR associated with the occurrence of the CS, for each trial a 'CS-SCR minus pre-CS-SCR' difference score was then calculated. This score was taken to be a measure of the conditioned response to the CS, and henceforth is simply referred to as the $\triangle$ SCR (change in skin conductance response). For the expectancy data, the rating in the CS period was used as the participant's expectancy of the US on that trial. Subsequent data analysis was conducted using SPSS version 22.

## Results

## Training Data

The $\triangle \mathrm{SCR}$ data from training were analyzed using a stimulus (A vs. B) by group (MV+, MV-, CV+, CV-) ANOVA. The main effect of stimulus was significant
( $\mathrm{F}(1,92)=12.46, \quad \mathrm{p}<0.005$ ), with participants exhibiting higher SCR scores after $\mathrm{A}+$ (mean $=0.069, \mathrm{SE}=0.007$ ) than B- (mean $=0.048, \mathrm{SE}=0.007$ ). This difference in SCR scores did not significantly differ between the groups $(\mathrm{F}(3,92)=1.14$, n.s.), as their training regime was the same.

This pattern was also observed in a stimulus by group analysis of the expectancy ratings from training. Specifically, participants had a significantly $(\mathrm{F}(1,92)=396.56, \mathrm{p}<0.001)$ higher expectation of shock on $\mathrm{A}+$ training trials (mean $=4.1, \mathrm{SE}=0.06$ ) than on B - training trials (mean $=1.9, \mathrm{SE}=0.07$ ). Furthermore, this difference did not significantly differ across the groups $(\mathrm{F}(3,92)=0.37$, n.s.). Thus, as expected, participants showed learning of the $\mathrm{A}+\mathrm{B}$ - discrimination in both the $\triangle \mathrm{SCR}$ and expectancy data, and this did not differ between groups.

## Manipulation Check

The state mindfulness scale data were analyzed using a practice type (mindfulness vs. control) by stimulus visualized (A vs. B) ANOVA. Participants in the mindfulness conditions had significantly higher state mindfulness scores than those in the control conditions ( $\mathrm{F}(1,92)=9.37, \mathrm{p}<0.01$; respective means: 76.2 ( $\mathrm{SE}=1.72$ ) and 68.7 ( $\mathrm{SE}=1.72$ )). This difference did not significantly differ across stimulus visualized $(\mathrm{F}(1,92)=0.01$, n.s.). Thus, as intended, the mindfulness practice appeared to have increased state mindfulness levels relative to control, and regardless of whether stimulus A or B was visualized.

## Test Data

The focus here is on the $\Delta \mathrm{SCR}$ and expectancy data from the reacquisition test-phase, as pilot work suggested this would provide a more sensitive test of any effects than the data from the extinction phase. It also included twice as many trials as the extinction phase, and so should produce less noisy data.


Figure 1: Mean change in skin conductance by Group and by CS tested, from the reacquisition test-phase. Error bars represent the standard error.

Considering the $\triangle \mathrm{SCR}$ data first, these were analyzed using a stimulus tested (A vs. B) by practice type (mindfulness vs. control) by stimulus visualized (A vs. B)

ANOVA. The means can be seen in Figure 1; note that in this figure the four combinations of practice type and stimulus visualized are represented by the four groups on the x-axis. As would be expected given the contingencies, there was a significant main effect of stimulus tested ( $\mathrm{F}(1,92)=18.32, \mathrm{p}<0.001$ ), with higher $\triangle \mathrm{SCRs}$ to stimulus A than stimulus B. In addition, there were significant two-way interactions between stimulus tested and practice type $(\mathrm{F}(1,92)=6.71, \mathrm{p}<0.05)$, and between stimulus tested and stimulus visualized $(\mathrm{F}(1,92)=6.10, \mathrm{p}<0.05)$. None of the other main effects, nor the three-way interaction, were significant (all $\mathrm{p}>0.4$ ). Thus the findings suggest that practice type and stimulus visualized had additive effects on the difference in $\triangle \mathrm{SCR}$ between A and B . More specifically, practicing mindfulness appeared to decrease the difference in $\triangle \mathrm{SCR}$ between the two, as did visualizing stimulus A . Hence the condition (MV+) containing both mindfulness and visualization of stimulus A had the smallest difference between A and B ; the condition containing neither (CV-) had the biggest difference between the two; and the other two conditions were somewhere in between.


Figure 2: Mean expectancy ratings by Group and by CS tested, from the reacquisition test-phase. Error bars represent the standard error.

Turning to the reacquisition expectancy data (Figure 2), these were analyzed using the same three-way ANOVA as above. As with the $\triangle$ SCR data, there was a main effect of stimulus tested $(\mathrm{F}(1,92)=677.26, \mathrm{p}<0.001)$, with a correct, higher expectation that a shock would follow stimulus A than B. However, none of the other main effects or interactions were significant (all $\mathrm{p}>0.1$ ). Thus, in contrast to the $\triangle \mathrm{SCR}$ data, none of the manipulations had a measurable effect on the participants' expectation of shock.

## Discussion

There are a number of results that emerge from this study. Perhaps the first point to make is that training was very effective, and produced good conditioning both in terms of conscious cognitive expectancy and in terms of change in skin conductance. There were no confounding differences during training across groups, and the manipulation check at
the end of the interposed activity indicated that mindfulness practice was also successful. We can be fairly confident, then, that the study we set out to conduct has actually taken place. We can now ask what the effects of mindfulness practice and visualization are on differential fear conditioning.

Starting with expectancy ratings for shock to the CS + and CS- during re-acquisition, the answer is equally straightforward. Our manipulations had no differential effect. All groups showed the same (highly significant) level of differential conditioning on this measure. Any account that would claim that conscious expectancy is what drives changes in skin conductance would thus have to postulate a similar pattern of results in the $\triangle \mathrm{SCR}$ measure, but this was not what we observed. Instead, $\Delta \mathrm{SCR}$ varied across groups, and in particular the extent to which differential conditioning was re-acquired differed across groups. Group MV+, which received mindfulness practise and visualised the CS+ at the same time, showed no differential conditioning. Group CV-, which listened to the audio book and visualised the CS- at the same time, showed strong differential conditioning, actually stronger than that during initial training. The difference, then, between these groups was considerable, and our analysis produced results that suggested that both mindfulness practise and visualisation of the CS+ had some protective effect against re-acquisition of differential fear conditioning.

This contrast between effects on expectancy and on $\triangle$ SCR suggests that after mindfulness practise and visualisation of the CS+, even though people knew that a shock was likely to occur to A and not to B , this had no effect on their physiological response to those stimuli. In some sense, then, their autonomic response has become decoupled from their conscious cognitive appraisal of the situation. We can argue that this has happened to a lesser extent for Groups MV- and $\mathrm{CV}+$, and not at all for CV-. This is not the only possible interpretation of these results, however, and we need to consider others that might generate the same pattern on our two measures.

One such possibility is that rather than visualisation of the CS+ having a protective effect, it was in fact visualisation of the CS- that simply extinguished any fear generalising to that stimulus and so led to stronger differential conditioning. We cannot rule this possibility out, but would expect visualising the CS + to have had an even stronger effect than visualising CS-. This is because the CS+ would have had stronger associations to shock after training, and the extinction would be expected to have been proportional to the strength of the association. This mechanism could, then, explain why visualising the CS+ seemed to impair differential conditioning, and visualising the CS- seemed to (relatively speaking) help it. On this account, the visualisation effect was just one of imagined extinction feeding through into re-acquisition. But understanding the effect of mindfulness practise in this way is probably not helpful. The effect was additive with visualisation, which indicates a different source for it, and claiming that listening
to Bill Bryson potentiated fear conditioning does seem a little unlikely (as well as unkind).

We can ask, however, why the effect we observed was on differential conditioning, and why the $\Delta \mathrm{SCR}$ for B was so high in the MV+ group. This was undoubtedly a contributory factor to our result, though the stronger effect across groups may have been the effect of our manipulations on A . In fact, it is quite striking that the average $\triangle \mathrm{SCR}$ to A and B was roughly the same in each group. Thus, we can argue that the overall physiological reactivity of each group was approximately constant, it was just how that was distributed over A and B that varied. If it was the case that the training data were mostly based on conscious expectancy of shock (and that would not be surprising in such a simple preparation), then the implication is that this factor was no longer effective in Group MV+ during reacquisition, since if participants' expectancies had driven their physiological reactivity in that case, we would have seen a difference in $\triangle \mathrm{SCR}$ to A and B equivalent to that in the other groups, but we did not. It is possible that instead we observed the effects of the underlying associative learning as a result of training, extinction and re-acquisition and that this had become decoupled from control by conscious cognitive expectancy. It is also possible that one belief about the contingencies had been replaced by another, though our expectancy rating data argues against this. Further research will be needed to establish exactly what the effective contribution of the mindfulness practise is here.

## Conclusion

We ran this experiment to try to understand what effect mindfulness might have on the processes underlying conditioned fear (and hence anxiety). We can be confident that it does have an effect, and that this effect appears to be separate from that of visualisation. Whatever the exact mechanism involved, mindfulness practise appears to protect against re-acquisition of conditioned fear after extinction, a result that must be worth pursuing and that may have implications for the treatment of stimulus-specific anxiety.

## Acknowledgments

This research was funded by British Academy / Leverhulme Small Research Grant SG150007.

## References

Bryson, Bill. (2003). A short history of nearly everything. USA: Broadway Books.
Kabat-Zinn (1990). Full catastrophe living: using the wisdom of your body and mind to face stress, pain and illness. New York: Bantam Dell.
Kabat-Zinn, J. (1994). Wherever you go, there you are: mindfulness meditation in everyday life. New York: Hyperion.

Knight, D.C., Nguyen, H.T., \& Bandettini, P.A. (2003). Expression of conditional fear with and without awareness. PNAS, 100(25), 15280-15283.
Kuyken, W., Warren, F.C., Taylor, R.S., Whalley, B., Crane, C., et al. (2016). Efficacy of mindfulness-based cognitive therapy in prevention of depressive relapse: an individual patient data meta-analysis from randomized trials. JAMA Psychiatry, 73, 565-574.
Lotan, G., Tanay, G., \& Bernstein, A. (2013). Mindfulness and distress tolerance: relations in a mindfulness preventive intervention. International Journal of Cognitive Therapy, 6, 371-85.
Lovibond P.F., \& Shanks, D.R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. Journal of Experimental Psychology: Animal Behavior Processes, 28(1), 3-26.
McAndrew, A., Jones, F. W., McLaren, R., \& McLaren, I. P. L. (2012). Dissociating expectancy of shock and changes in skin conductance: an investigation of the Perruchet effect using an electrodermal paradigm. Journal of Experimental Psychology: Animal Behavior Processes, 38, 203-208.
McLaren, I. P. L., Forrest, C.L., McLaren, R.P., Jones, F.W., Aitken, M.R.F. and Mackintosh, N.J. (2014). Associations and Propositions: The case for a dualprocess account of learning in humans. Neurobiology of Learning and Memory, 108, 185-95.
McLaren, I.P.L., Green, R.E.A., \& Mackintosh, N.J. (1994). Animal learning and the implicit/explicit distinction. In N.C. Ellis (Ed.), Implicit and explicit learning of languages. Cambridge, MA: Academic Press.
Mindfulness All-Party Parliamentary Group (2015). Mindful nation UK. London: The Mindfulness Initiative.
Segal, Z. V., Williams, J.M.G., \& Teasdale, J.D. (2013). Mindfulness-based cognitive therapy for depression: a new approach to preventing relapse ( $2^{\text {nd }}$ ed.). New York: Guilford.
Tabbert, K., Stark, R., Kirsch, P., \& Vaitl, D. (2006). Dissociation of neural responses and skin conductance reactions during fear conditioning with and without awareness of stimulus contingencies. Neuroimage, 32(2), 761.

Tanay. G., \& Bernstein, A. (2013). State mindfulness scale (SMS): development and initial validation. Psychological Assessment, 25, 1286-99.
Tang, Y. Y., Holzel, B. K., \& Posner, M. I. (2015). The neuroscience of mindfulness meditation. Nature Reviews Neuroscience, 6, 213-225.
Treanor, M. (2011). The potential impact of mindfulness on exposure and extinction learning in anxiety disorders. Clinical Psychology Review, 31, 617-25.
van der Velden, A. M., Kuyken, W., Wattar, U., Crane, C., Pallesen, K. J., Dahlgaard, J., Fjorback, L. O., \& Piet, J. (2015). Attentional orienting and executive control are affected by different types of meditation practice. Consciousness and Cognition, 46, 110-126.

# A Study on the Impact of Chess Training on Creativity of Indian School Children 

Ebenezer Joseph (emmanuelchess@gmail.com)<br>University of Madras,<br>Department of Psychology, WCC, College Road, Chennai, 600006, India.

S. Sundar Manoharan (ssm@karunya.edu)

Karunya University
Coimbatore, 641114, India.

Veena Easvaradoss (veenadoss@gmail.com)<br>Department of Psychology, Women's Christian College<br>College Road, Chennai, 600006, India.

David Chandran (davidbeulah@gmail.com)
Emmanuel Chess Centre
Kasthuri Ranga Road Chennai, 600018, India.


#### Abstract

Creativity is the ability to produce work that is both novel and appropriate. The study, funded by Indian government, analyzed the effect of one-year chess training on the creativity of children. A pretest and posttest with control group design was used, with 31 children in experimental and 32 in control group. The experimental group underwent weekly chess training. Wallach-Kogan Creativity Test (Indian Adaptation) was used. Analysis revealed that only the experimental group had statistically significant gains in total creativity and two nonverbal subtests. The authors conclude that systematic chess training inculcates in the child the ability to think divergently and creatively.


Keywords: Abstract Thinking; Chess Training; Creativity; Innovation; Divergent Thinking

## Introduction

Creativity is defined as the tendency to generate or recognize ideas, alternatives, or possibilities that may be useful in solving problems, communicating with others, and entertaining ourselves and others (Franken, 1982). Typically creativity is defined as "the ability to produce work that is both novel (i.e. original, unexpected) and appropriate (i.e. adaptive concerning task constraints)" (Sternberg, 1999). These definitions emphasize both the concept of fluency and novelty in the responses that have been generated.

Most theorists agree that the creative process involves a number of components, most commonly:

1. Imagination
2. Originality (the ability to come up with new and original ideas and products)
3. Productivity (the ability to generate a variety of ideas through divergent thinking)
4. Problem solving (application of knowledge and imagination to a given situation)
5. The ability to produce an outcome of value and worth

Creativity is commonly utilized divergent thinking. A creative or divergent thinker is described as the person who
pushes the boundaries of ability and knowledge and is able to reconsider the problem to find different perspectives and solutions and ignore distractions that can negatively affect his or her productivity (Saccardi, 2014). Creativity among children emerges gradually between grades one to three (Torrance, 1964). In general, the broad and complex multidimensional concepts of creativity can be measured by the Torrance Tests of Creative Thinking (TTCT: Torrance, 1964, 1990a, 1990b) and the Wallach-Kogan Creativity Tests (WKCT: Wallach \& Kogan, 1965).

There is a fairly common belief that creativity can be developed through training. Various recent studies that have assessed the effects of programs for stimulating creativity confirm this belief (Antonietti, 2000; Fleith, Renzulli, \& Westberg, 2002; Komarik \& Brutenicova, 2003; Saxon, Treffinger, Young, \& Wittig, 2003). Consequently, many countries are increasingly placing a high priority on stimulating creative thinking at the school level.

Since chess helps in developing strategic thinking and problem-solving skills of children, it may also be effective in improving their cognitive skills (Sigirtmac, 2016). Chess builds problem-solving abilities, enhance strategic thinking skills, and even improves self-esteem as well as higherorder thinking skills, which are known as meta-cognitive skills. In countries, where chess is intensely played by students, practicing students become among the top students in mathematics and science and they are able to recognize complicated patterns (Milat, 1997).

While a number of other models of creativity have brought out the steps involved in the creative process, Avni (1998) posited a four-step model specific to chess playing. According to him, an intelligent process in playing chess consists of four different steps: synthesis (opinion forming and plan shaping), gathering (collecting the raw materials during position evaluation), enlightenment (a sudden observation of an idea), and realization (translating the idea into practical lines of play). Thus, these four steps can be
used for a creative process that could also work in some other areas (Bushinsky, 2009).

India has a long history of chess playing but there are only a few studies on chess as a strategy to increase cognitive abilities. Further, there are no studies assessing the impact of chess intervention on the creativity of children. If research can establish that chess training can facilitate creativity, it can significantly impact educational programs to increase creative thinking.

The objective of the study was, therefore, to analyze the effect of one-year chess training program on the creativity of school-going children of both genders and to assess its effect on the verbal and nonverbal components of creativity. It was hypothesized that chess training would significantly increase creativity in children.

## Methodology

The research design used for the study was pretest and posttest with control group design. The independent variable was the Chess training program, and the dependent variable was Creativity of children.

The sample consisted of 63 children, 31 in the experimental group and 32 in the control group. The children in the experimental group were selected purposively and comprised children who volunteered for the chess program. The children in the control group were randomly selected using random numbers table generated online. The children in the control group were selected on the basis of no chess knowledge and were not given chess training. During the time of chess intervention for the experimental group, the control group children were engaged in other activities such as music, arts and in outdoor sports such as cricket, football, basketball, etc. The mean age for experimental group was 11.86 years ( $\mathrm{SD}=$ 1.44 ) and for control group was 12.03 years ( $\mathrm{SD}=1.14$ ). The experimental group consisted of 9 girls and 22 boys, and the control group consisted of 7 girls and 25 boys.

## Tools

Creativity was assessed by Indian adaption of WallachKogan Creativity Test. The WKCT (Wallach \& Kogan, 1965) is similar to the TTCT in that it focuses on divergent thinking and assesses both visual and verbal content. It includes three verbal subtests-Instances (e.g., name all the round things you can think of), Alternative Uses (e.g., for a newspaper), and Similarities (e.g., How are a cat and mouse similar?)—and two figural subtests—Pattern Meanings and Line Meanings (interpreting abstract patterns and lines). It is scored for fluency (number of ideas) and uniqueness (ideas not offered by others in the group being tested). Wallach and Kogan's (1965) major contribution was their belief that standardized test procedures were not conductive to creative performance and their insistence on a more relaxed and game-like atmosphere. The test is given individually, and no time limits are imposed. However, in the present administration, a time limit of three minutes was given for each subtest. The number of valid responses for each subtest
was summed to obtain the subtest totals. The total creativity scores comprised the sum of the subtest scores.

## Chess Training Methodology

The children were grouped into small clusters based on the chess ability and learning capacity and were trained for an hour starting from the basics. The training methodology comprised Winning Moves Chess Learning Program (Joseph, 2008) Episodes 1-22, lectures with the demonstration board, on-the-board playing and training, chess exercise through workbooks (Chess School 1A, Chess School 2, and tactics), and working with chess softwares. Further students' games were mapped and analyzed using score sheets and Chess software. The children were taught the ideas behind chess openings, and exposure to classical games was also given. The children participated in mock as well as regular tournaments. On an average, the children underwent one hour per week chess intervention for about 25-30 sessions. One coach was assigned for 8 students.

## Procedure

Baseline creativity assessment was done after obtaining informed consent, from the parents and the school authorities. The research was carried out on the approval of government of India, department of science and technology, Task force and the doctoral committee. Reassessment was carried out after an average duration of one year. The assessment environment was quiet without any disturbance and kept standardized. Psychologists were trained to administer the test in a uniform standardized method to minimize the testing error.

Clustering technique was used to form the training groups of six to eight children. The chess training consisted of once-a-week chess classes conducted for one hour during the end of school hours for a year (about 30 hours of chess training). The children were given a standardized Winning Moves Chess Learning Program (Joseph, 2008), and they played at tournaments also.

## Results

The analysis was carried out using SPSS. Paired $t$-test was carried out to analyze differences within groups, and independent $t$-test was used to assess differences between groups in the mean total creativity scores and mean subtest scores. Pre-intervention equivalence of groups on creativity was established for total creativity scores and the subtest scores.

Table 1: The Significance of the Difference between the Means of the Experimental and Control Groups on the Creativity Test using the Independent $t$-Test.

| Scores | Assess ment | Mean and Standard Deviation |  | $t$ |
| :---: | :---: | :---: | :---: | :---: |
|  |  | Experi mental | Control |  |
| Total creativity | Pre | $\begin{aligned} & 54.19 \\ & 16.98 \\ & \hline \end{aligned}$ | $\begin{aligned} & 53.93 \\ & 12.38 \\ & \hline \end{aligned}$ | 0.06 |
|  | Post | $\begin{aligned} & 16.90 \\ & 18.77 \end{aligned}$ | $\begin{aligned} & 52.40 \\ & 17.12 \end{aligned}$ | 2.09* |
| Instances | Pre | $\begin{aligned} & 13.74 \\ & 6.34 \end{aligned}$ | $\begin{gathered} 15.81 \\ 5.26 \end{gathered}$ | -1.41 |
|  | Post | $\begin{aligned} & 17.41 \\ & 6.79 \end{aligned}$ | $\begin{gathered} 16.78 \\ 5.92 \end{gathered}$ | 0.39 |
| Alternate Uses | Pre | $\begin{aligned} & 9.09 \\ & 3.66 \end{aligned}$ | $\begin{gathered} 10.09 \\ 3.03 \end{gathered}$ | 1.17 |
|  | Post | $\begin{aligned} & 10.87 \\ & 3.66 \end{aligned}$ | $\begin{aligned} & 9.46 \\ & 4.22 \end{aligned}$ | 1.40 |
| Similariti es | Pre | $\begin{aligned} & \hline 7.74 \\ & 3.51 \end{aligned}$ | $\begin{aligned} & 7.96 \\ & 3.52 \end{aligned}$ | 0.25 |
|  | Post | $\begin{aligned} & 9 \\ & 4.47 \end{aligned}$ | $\begin{aligned} & 7.93 \\ & 3.74 \end{aligned}$ | 1.01 |
| Line <br> Drawing | Pre | $\begin{aligned} & 11.77 \\ & 4.98 \end{aligned}$ | $\begin{aligned} & \hline 9.96 \\ & 3.99 \end{aligned}$ | 1.58 |
|  | Post | $\begin{aligned} & 12.12 \\ & 4.22 \end{aligned}$ | $\begin{gathered} 8.65 \\ 4 \end{gathered}$ | 3.34** |
| Pattern <br> Meaning | Pre | $\begin{aligned} & 11.83 \\ & 4.68 \end{aligned}$ | $\begin{gathered} 10.09 \\ 3.74 \end{gathered}$ | 1.63 |
|  | Post | $\begin{aligned} & 12.80 \\ & 4.81 \end{aligned}$ | $\begin{aligned} & \hline 9.87 \\ & 4.11 \\ & \hline \end{aligned}$ | 2.59** |

*p<.05; **p<.01.
Table 1 indicates that there was a significant difference between the means of the post-intervention total creativity scores ( $p<.05$ ) Cohen's d indicated an effect size of ( 0.52 ), indicating that chess training had significantly increased creativity. Significant differences between the postintervention means were observed on the Line Drawing subtest ( $p<.01$ ) Cohen's d effect size ( 0.84 ) and the Pattern Meaning subtest ( $p<.01$ ) Cohen's d effect size (0.68) , indicating that chess training had significantly increased the scores on these two subtests. No significant differences were observed on any other subtest.

## Discussion

It can be inferred from Table 1 that systematic chess intervention increases creativity in children. As research has clearly established, chess is a game that stimulates cognitive processes and strengthens intellectual abilities and cognitive skills (Aciego, García, \& Betancort, 2012; Bilalic, McLeod, \& Gobet, 2007; De Bruin, Kok, Leppink, \& Camp, 2014). Moreover, it has shown that the intellectual gains have translated into increases in both IQ and academic scores (Aydın, 2015; Barrett \& Fish, 2011; Joseph, Easvaradoss, \& Solomon, 2016; Romano, 2011). Large Effect Sizes for Total Creativity (0.52), Line Drawing (0.84) and Pattern Meaning (0.68) where seen, indicating that chess had a significant impact on Total Creativity, Line and Pattern
subsets of the experimental group. This finding was in line with Sigirtmic (2016), findings who found a statistically significant difference between elaboration, resistance to premature closure and total creativity score of children in favour of those who received chess training.

In the present study, the children were taught chess systematically. They did not merely play chess but were strongly encouraged to challenge their own standards and also to play competitively. They analyzed their own games, identified their strengths, and understood their mistakes. They were also given opportunities to pit their skills against others as they played in tournaments. It is clear that the outcome of this rigorous, yet enjoyable, training methodology was the enhanced cognitive abilities that were reflected in increased creativity scores.

The intellectual strategies underlying chess playing have been spelt out by Avni (1998). According to him, chess playing involves an intelligent process that consists of four different steps: synthesis (opinion forming and plan shaping), gathering (collecting the raw materials during position evaluation), enlightenment (a sudden observation of an idea), and realization (translating the idea into practical lines of play). The child thinks beyond the usual solutions using divergent thinking, thinking abstractly, weighing options, evaluating outcomes, and making decisions. Insightful thinking also appears to play a role.

The Wallach-Kogan Test, which was used in the present study, requires the child to think divergently, quickly, and fluently, generating as many responses as possible on the different tasks. It is evident that similar abilities are utilized in playing chess where innovativeness and accuracy and both broad-based and precise thinking are required. The experimental group, which had undergone one-year training, in chess appears to have acquired these skills as indicated by a significant increase in overall creativity compared to the control group. Earlier studies have pointed to the positive impact that chess has had on academic scores, especially language and reasoning (Joseph et al., 2016). The components of creativity studied on the test are the ability to name objects that have common properties involving abstraction ability (Instances), to identify multiple uses for common objects involving divergent thinking (Alternate Uses), to perceive similarities between two different objects utilizing generalizing and abstracting ability (Similarities), to perceive meaning in meaningless stimuli involving innovativeness (Line Drawing), and to perceive meaning in structures stimuli involving the ability to form association (Pattern Drawing). The children in the experimental group have shown increases in all the post-intervention scores, though not all increments have reached significance.

Significant increases have been observed on the Line Drawing subtest ( $p<.01$ ) and the Pattern Meaning subtest ( $p<.01$ ) as seen in Table 1. On the Line Drawing subtest, the child is shown a line drawing for 30 seconds and is asked to generate as many responses as possible about what the drawing means to him or her. On the Pattern Meaning subtest, the child is shown a design (which is more
structured) and is asked to generate as many responses as possible about what the design means to him or her. This test measures fluency and the ability to uncritically generate ideas and possibilities, both commonplace and unique. The game of chess uses primarily visuospatial strategies. Systematic chess training inculcates in the child the ability to think divergently, visualizing the pros and cons of the various chess moves.

Garaigordobil (2006) studied the impact of a play program on the verbal and graphic-figural creativity. Results showed a positive effect of the intervention, as the experimental participants significantly increased their verbal creativity and graphic-figural creativity. This research primarily focused on structured cooperative play. The chess intervention in the present study also has structural characteristics that corroborate the finding of other studies that had indicated positive effects of play on the development of creativity. This structured quality helps the child to systematically visualize all the possible options and outcomes available to him or her. This ability, which has been acquired through chess training, has led to the increased total creativity scores and the increases on the visuospatial subtests.

## Implications

It is evident that systematic chess intervention increases creativity in children. The child thinks beyond the usual solutions-using divergent thinking, thinking abstractly, weighing options, evaluating outcomes, and making decisions. Significant improvement in the Line Drawing and Pattern Meaning subtest substantiates the fact that the game of chess primarily uses visuo-spatial strategies. Systematic chess training inculcates in the child the ability to think divergently, visualizing the pros and cons of various chess moves. It allows the child to conceptualize all the possible options and outcomes available to him or her. Increasing the creativity of children has possible far-reaching benefits for academic performance and generally for life skills. Systematically learning chess as part of school activities appears to have a broad spectrum of positive outcomes. The child who develops the ability to think in creative ways in playing chess is likely to transfer this learning to dealing with life challenges creatively.

## Acknowledgement

This research was supported by (1) the funding provided by the Department of Science and Technology, Cognitive Science Research Initiative, Government of India; (2) travel grant received from Ministry of Sports and Youth Affairs, HRDS, Government of India. Authors acknowledge the financial assistance.

## References

Aciego R., García, L., \& Betancort, M. (2012). The benefits of chess for the intellectual and social-emotional enrichment in schoolchildren. The Spanish Journal of Psychology, 15(2), 551-559.
Antonietti, A. (2000). Enhancing creative analogies in primary school children. North American Journal of Psychology, 2, 75-84.
Avni, A. (1998). Creative chess. London: Everyman Publishers.
Aydin, M. (2015). Examining the impact of chess instruction for the visual impairment on mathematics. Educational Research and Reviews, 10(7), 907-911.
Barrett, D. C., \& Fish, W. W. (2011). Our move: Using chess to improve math achievement for students who receive special education services. International Journal of Special Education, 26(3), 181-193.
Bilalic, M., McLeod, P., \& Gobet, F. (2007). Does chess need intelligence?-A study with young chess players. Intelligence, 35, 457-470.
Bushinsky, S. (2009). Deus Ex Machina-A higher creative species in the game of chess. AI Magazine, 30(3), 63-70. Retrieved from http://dx.doi.org/10.1609/aimag.v30i3.2255
De Bruin, A. B. H., Kok, E. M., Leppink, J., \& Camp, G. (2014). Practice, intelligence, and enjoyment in novice chess players: A prospective study at the earliest stage of a chess career. Intelligence, 45, 18-25.
Fleith, D. S., Renzulli, J. S., \& Westberg, K. L. (2002). Effects of a creativity training program on divergent thinking abilities and self-concept in monolingual and bilingual classrooms. Creativity Research Journal, 14, 373-386.
Franken, R. (1982). Human motivation. Monterey, CA: Brooks/Cole Publishing.
Garaigordobil, M. (2006). Intervention in creativity with children aged 10 and 11 years: Impact of a play program on verbal and graphic-figural creativity. Creativity Research Journal, 18(3), 329-345.
Joseph, E. (2008). Patent No. L-32958/2009, India.
Joseph, E., Easvaradoss, V., \& Solomon, N. J. (2016). Impact of chess training on academic performance of rural Indian school children. Open Journal of Social Sciences, 4, 20-24. Retrieved from http://dx.doi.org/10.4236/jss.2016.42004
Komarik, E., \& Brutenicova, E. (2003). Effect of creativity training on preschool children. Studia Psychologica, 45, 37-42.
Milat, M. (1997). The role of chess in modern education. Retrieved from http://southernchessclub.org/site/documents/TheRoleofCh essinMode rnEducation.pdf

Romano, B. (2011). Does playing chess improve math learning? Promising (and inexpensive) results from Italy. (Unpublished doctoral dissertation). University of Pennsylvania, Philadelphia.
Saccardi, M. (2014). Creativity and children's literature: New ways to encourage divergent thinking. Santa Barbara, CA: ABC-CLIO.
Saxon, J. A., Treffinger, D. J., Young, G. C., \& Wittig, C. V. (2003). Camp invention (R): A creative, inquiry-based summer enrichment program for elementary students. Journal of Creative Behavior, 37, 64-74.
Sigirtmac, A. D. (2016). An investigation on the effectiveness of chess training on creativity and theory of mind development at early childhood. Academic Journals, 11(11), 1056-1063.

Sternberg, R. J. (Ed.). (1999). Handbook of creativity. New York, NY: Cambridge University Press.
Torrance, E. P. (1964). The Minnesota studies of creative thinking: 1959-1962. In C. W. Taylor (Ed.), Widening horizons in creativity (pp. 125-144). New York: John Wiley \& Sons.
Torrance, E. P. (1990a). Torrance tests of creative thinking norms-Technical manual (figural). Bensenville, IL: Scholastic Testing Service.
Torrance, E. P. (1990b). Torrance tests of creative thinking norms-Technical manual (verbal). Bensenville, IL: Scholastic Testing Service.
Wallach, M. A., \& Kogan, N. (1965). Modes of thinking in young children: A study of the creativity intelligence distinction. New York: Holt, Rinehart \& Winston.

# Document Similarity Misjudgment by LSA: Misses vs. False Positives 

Kyung Hun Jung (kjung2@kennesaw.edu)<br>Department of Psychology, Kennesaw State University (Kennesaw Campus) 402 Bartow Ave Kennesaw (Rm 4030), GA 30144 USA

Eric Ruthruff (ruthruff@unm.edu)<br>Department of Psychology, Logan Hall MSC03-2020<br>1 University of New Mexico, Albuquerque, NM 87131-0001 USA

Timothy Goldsmith (gold@unm.edu)
Department of Psychology, Logan Hall MSC03-2020
1 University of New Mexico, Albuquerque, NM 87131-0001 USA


#### Abstract

Modeling text document similarity is an important yet challenging task. Even the most advanced computational linguistic models often misjudge document similarity relative to humans. Regarding the pattern of misjudgment between models and humans, Lee and colleagues (2005) suggested that the models' primary failure is occasional underestimation of strong similarity between documents. According to this suggestion, there should be more extreme misses (i.e., models failing to pick up on strong document similarity) than extreme false positives (i.e., models falsely detecting document similarity that does not exist). We tested this claim by comparing document similarity ratings generated by humans and latent semantic analysis (LSA). Notably, we implemented LSA with 441 unique parameter settings, determined optimal parameters that yielded high correlations with human ratings, and finally identified misses and false positives under the optimal parameter settings. The results showed that, as Lee et al. predicted, large errors were predominantly misses rather than false positives. Potential causes of the misses and false positives are discussed.


Keywords: text document relatedness; semantic similarity; latent semantic analysis (LSA)

## Introduction

Modeling how humans judge the semantic similarity of text documents is an interesting topic in cognitive science with numerous practical implications. In an effort to better model human document similarity judgments, Lee, Pincombe, and Welsh (2005) compared several models of document similarity, including latent semantic analysis (LSA). They found that LSA's cosine similarity scores yielded higher agreement ( $r=.60$ ) with aggregate human ratings than other models, such as Tversky's (1977) ratio model ( $r=.50$ ). Considering that the inter-rater correlation among human raters is also about .60 , LSA seems to judge about as well as a single human rater. However, the moderate correlation between humans and LSA also suggests that LSA does not fully capture human similarity judgments 1 .

[^359]To better understand the weaknesses of LSA, and thereby improve models of text document similarity, this study investigated the pattern of discrepancy between LSA and humans with respect to their document similarity ratings. Specifically, we examined the frequency and degree of underestimation (misses) and overestimation (false positives) made by LSA relative to humans under favorable parameter settings of LSA.

## Misses vs. False Positives

Regarding the nature of the misjudgment by models, Lee et al. (2005) suggested that extreme misses would be a stronger cause than extreme false positives. They made this suggestion based on an observation that the common features model (Lee \& Navarro, 2002) occasionally misses the high similarity between documents that is readily apparent to humans. In a scatterplot of the model ratings against human ratings for each document pair, the authors found a cluster of points (document pairs) with low model ratings but high human ratings. That is, the model missed some of the strong document similarities that humans detected.

Lee et al.'s (2005) analysis above was based on the common features model, but it might apply to LSA as well. The common features model judges document similarity primarily based on the proportion of common features (words) shared by two documents. Notably, LSA's underlying model, the vector space model, determines document similarity in a similar manner. Therefore, Lee et al.'s findings suggesting more extreme misses over extreme false positives may also apply to LSA. This interesting hypothesis, if validated, would provide a valuable clue to improving models of text document similarity. However, it has not yet been rigorously tested.

Testing the hypothesis seems straightforward at first glance: compare human and LSA's document similarity ratings for a set of documents pairs. Then, document pairs with especially low LSA ratings compared to human ratings should be considered as misses and the reverse as false
positives. However, LSA's document representation depends on the parameters used such as the quantity and quality of the background documents (Bullinaria \& Levy, 2006), the dimensionality (Dumais, 1991; Landauer \& Dumais, 1997), and the local-global weighting schemes (Lintean, Moldovan, Rus, \& Mcnamara, 2010; Nakov, Popova, \& Mateev, 2001). Therefore, LSA's misjudgments relative to human judgments could vary depending on the parameters.

In this study, we attempted to investigate the nature of misjudgment by LSA under its optimal parameter settings. Therefore, we first identified LSA's optimal parameter settings by employing as many as 441 unique parameter combinations. Then, under the selected optimal parameter settings, we identified misjudgments by LSA as misses or false positives. Finally, we measured the degree of misjudgment of the two types using normalized scores.

The remainder of this paper has the following structure: (1) introduction to LSA, (2) an experiment identifying optimal parameter settings, (3) identification of misjudgments as misses and false positives under the optimal parameter settings, and (4) discussion of the underlying causes of the misses and false positives.

## LSA

LSA is based on a vector space model in which documents are first transformed into a word-by-document matrix. Rows of the matrix correspond to the unique words across documents, whereas columns correspond to individual documents. Cell values are the frequencies of words within each document. The cell values can be weighted in two respects: to what degree a word is important in representing a document's topic (local weighting), and to what degree a word is important in distinguishing one document from another according to their topics (global weighting). Using the weighted cell values, each document can be represented as a vector in a multidimensional space, where the dimensions correspond the unique words. Finally, the sematic similarity between two document vectors is typically measured using the cosine similarity score.

The core process that distinguishes LSA from the vector space model is singular value decomposition (SVD) implemented on the word-by-document matrix. SVD is a matrix factorization method that decomposes an original matrix (A) into three sub-matrices, $\mathrm{USV}^{\mathrm{T}}$, where U is a unitary w * r matrix (word-by-dimension matrix), S is an r * $r$ diagonal matrix with non-negative real numbers on its diagonal (singular value matrix), and $\mathrm{V}^{\mathrm{T}}$ is a unitary $\mathrm{r} * \mathrm{~d}$ matrix (dimension-by-document matrix). By multiplying these three sub-matrices, the original matrix can be retrieved, and this type of SVD is called full SVD.

In a modified version of the full SVD, called reduced SVD, small singular values located in the lower right corner of S are intentionally discarded, while preserving the first k largest

2 Background documents are included in the original corpus subject to LSA, along with the target documents. They are employed only for constructing the multidimensional space in which the target
singular values. The corresponding columns and rows of U and $\mathrm{V}^{\mathrm{T}}$, respectively, are discarded, too. The original $\mathrm{USV}^{\mathrm{T}}$, after discarding some values, then can be denoted as $\mathrm{U}^{\prime} \mathrm{S}^{\prime}\left(\mathrm{V}^{\mathrm{T}}\right)^{\prime}$, where $\mathrm{U}^{\prime}$ is a $\mathrm{w}^{*} \mathrm{k}$ matrix whose columns are the first $k$ columns of $U, S^{\prime}$ is a $k^{*} k$ diagonal matrix whose diagonal elements are the $k$ largest singular values of $S$, and $\left(\mathrm{V}^{\mathrm{T}}\right)^{\prime}$ is a $\mathrm{k} * \mathrm{~d}$ matrix whose rows are the first k rows of $\mathrm{V}^{\mathrm{T}}$. By multiplying these three reduced sub-matrices, one can obtain the least squares approximation of the original matrix. Finally, documents can be represented as vectors on a $k$ dimensional singular-value-space, which has k orthogonal axes. These dimensions are constructed so that the first axis explains the largest amount of variance of A , and the second axis explains the second largest amount of variance of $A$, and so on.

Furnas et al. (1988) was the first to apply the reduced SVD to the vector space model. This method was later called latent semantic analysis by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), who also demonstrated that LSA retrieves information better than traditional word-matching methods. Deerwester et al. argued that SVD uncovers latent semantic relations across documents that are buried in the corpus by removing noise (small singular values) in the original word-by-document matrix.

## Identification of Misses and False Positives under LSA's Optimal Parameter Settings

## Stimuli and Procedure

Target Text Documents We used the 1,225 document pairs from Pincombe (2004), which Lee et al. (2005) also adopted. These document pairs were generated by pairing 50 target news articles selected from Australian Broadcasting Corporation's news mail service. Each news article had a single paragraph containing 51 to 126 words (average: 82 words). They covered a variety of topics, such as terrorism and hunger in Africa. For each of the 1,225 document pairs, Pincombe collected about 10 human ratings by asking 83 university students to each rate the relatedness of a subset of the document pairs. Participants used a five-point scale, with one indicating "highly unrelated" and five indicating "highly related".

Background Documents Lee et al. (2005) used 314 news articles from the same Australian news corpus as background documents2. In this study, to explore the optimal parameter settings of LSA, we employed 4,172 additional news articles from the same news corpus (total of 4,486). These new background documents contained a single paragraph (average: 152 words). They also covered a variety of topics as the 50 target news articles did. In addition to the background document size used in Lee et al. (314) and the maximum size available in this study $(4,486)$, we examined
documents are represented. It is generally regarded that LSA's performance improves as the number of background documents increases (Bullinaria \& Levy, 2006).
five intermediate background document sizes by randomly selecting the following numbers of documents from the new set of 4,172 articles: $314,750,1,000,2,000$, and 3,000 (see Table 1).

Table 1. Seven background document conditions.

| Size | Source |
| :---: | :---: |
| 314 | The same 314 news articles as in Lee et al. (2005) |
| 314 | Randomly selected from the 4,172 articles |
| 750 | Randomly selected from the 4,172 articles |
| 1,000 | Randomly selected from the 4,172 articles |
| 2,000 | Randomly selected from the 4,172 articles |
| 3,000 | Randomly selected from the 4,172 articles |
| 4,486 | Combination of the 314 news articles from Lee et <br> al. and the new 4,172 articles |

Dimensionality Regarding the dimensionality of the reduced SVD, the maximum possible dimension for a given background document size corresponds to the total number of documents subjected to SVD (50 + the number of background documents). For example, in the 314background document condition, the maximum dimension is $364(=50+314)$. In most of the background document conditions employed in this study, higher dimensions than 364 were possible. However, following some researchers' arguments for the importance of maintaining 300 dimensions (Landauer \& Dumais, 1997), we selected the following seven dimensions for the reduced SVD (i.e., LSA): 50, 100, 150, $200,250,300$, and 364.

Other LSA Parameters Stemming, normalization, and removal of stopwords and alphanumeric words are known to improve LSA's document representation (Pincombe, 2004; Stone, Dennis, \& Kwantes, 2011). Therefore, they were applied to all LSA runs. Three local weighting schemes (tf, $\log$, and alt-log) and three global weighting schemes (idf, entropy, and p-inverse) were selected based on their significant effects observed in a pilot study (not reported here).

LSA cosine scores were computed for every possible (441) combination of the above parameters: 7 background document sizes $* 7$ dimensions $* 3$ local weighting schemes * 3 global weighting schemes.

Identifying Misses and False Positives To classify LSA ratings as misses and false positives relative to human ratings, we first normalized the human ratings and LSA's cosine scores using z -score 3 . The degree of misjudgment was measured as the absolute difference between the two normalized scores for a given document pair. If a document pair's normalized cosine score was smaller than the

[^360]normalized human rating by at least 1.0 , then the LSA's cosine score was considered a miss. But if a document pair's normalized cosine score was greater than the normalized human rating by 1.0 , then the LSA cosine score was considered a false positive.

## Results and Discussion

Optimal Parameters of LSA To determine which parameter settings are optimal for LSA's document similarity representation, we examined the correlation between LSA cosine scores and human ratings. The correlation was affected more systematically and strongly by the interaction of background document size and dimensionality than the local-global weighting schemes. Therefore, for the sake of simplicity, we merged the correlations across the nine weighting schemes at a given background document size and dimensionality. As shown in Table 2, the correlation increased markedly as we added more background documents, consistent with previous research (Bullinaria \& Levy, 2006). But this effect was more prominent at relatively high dimensions than at low dimensions.

Table 2. Correlations between human ratings and LSA cosine scores as a factor of the background document size and dimensionality. Correlations were merged across the nine local-global weighting schemes at a given background document size and dimensionality. Relatively high correlations ( $r \geq .67$ ) are shaded.

|  | Dimension |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Background | 50 | 100 | 150 | 200 | 250 | 300 | 364 | Average |  |
| 314 | 0.53 | 0.55 | 0.55 | 0.56 | 0.56 | 0.57 | 0.59 | 0.56 |  |
| New 314 | 0.62 | 0.59 | 0.58 | 0.56 | 0.57 | 0.58 | 0.61 | 0.59 |  |
| 750 | 0.65 | 0.66 | 0.63 | 0.61 | 0.61 | 0.60 | 0.59 | 0.62 |  |
| 1000 | 0.62 | 0.66 | 0.66 | 0.64 | 0.62 | 0.61 | 0.60 | 0.63 |  |
| 2000 | 0.51 | 0.61 | 0.64 | 0.68 | 0.67 | 0.68 | 0.67 | 0.64 |  |
| 3000 | 0.53 | 0.61 | 0.64 | 0.64 | 0.66 | 0.67 | 0.69 | 0.63 |  |
| 4486 | 0.58 | 0.65 | 0.67 | 0.69 | 0.68 | 0.68 | 0.66 | 0.66 |  |
| Average | 0.58 | 0.62 | 0.62 | 0.63 | 0.62 | 0.63 | 0.63 | 0.62 |  |

To identify optimal parameter settings of LSA, we first selected the 10 combinations of background document size and dimension that yielded correlations of at least .67, averaged across all weighting schemes (see the shaded cells in Table 2). Then, for each of these 10 combinations, we chose the local-global weighting scheme that yielded the highest correlation with human ratings. Table 3 shows the specific parameter settings of these 10 selected combinations as optimal parameter settings. The table also shows the correlation, number of misses and false positives, and the average absolute z -score errors.
the z -score normalization yields more reliable results with respect to the frequency of misses and false positives.

Table 3. Ten optimal parameter settings of LSA selected for the identification of misses and false positives. The parameters, correlation with human ratings, number of misses and false positives, and the average of the absolute z -score errors are shown.

| Background <br> document <br> size | Dimension | Local <br> Weighting | Global <br> Weighting | Correlation | Number <br> of misses | Number <br> of false <br> positives | Average <br> misjudgment <br> (absolute z-score <br> error) for misses | Average <br> misjudgment <br> (absolute z-score <br> error) for false <br> positives |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2000 | 200 | tf | p-inverse | 0.70 | 112 | 86 | 1.57 | 1.43 |
| 2000 | 250 | tf | p-inverse | 0.68 | 121 | 67 | 1.60 | 1.56 |
| 2000 | 300 | tf | idf | 0.68 | 117 | 72 | 1.61 | 1.56 |
| 2000 | 364 | alt-log | p-inverse | 0.68 | 118 | 78 | 1.60 | 1.54 |
| 3000 | 300 | tf | p-inverse | 0.68 | 115 | 86 | 1.63 | 1.47 |
| 3000 | 364 | alt-log | entropy | 0.69 | 104 | 76 | 1.61 | 1.42 |
| 4486 | 150 | alt-log | p-inverse | 0.68 | 119 | 95 | 1.57 | 1.42 |
| 4486 | 200 | tf | p-inverse | 0.70 | 120 | 83 | 1.62 | 1.48 |
| 4486 | 250 | tf | p-inverse | 0.68 | 124 | 78 | 1.64 | 1.58 |
| 4486 | 300 | tf | idf | 0.68 | 119 | 69 | 1.61 | 1.49 |

Nature of Misjudgments by LSA To determine the nature of LSA's misjudgments under optimal parameters, we used the z-score errors obtained from the 10 parameter settings shown in Table 3. As shown at the bottom of the table, misses $\left(\mathrm{M}_{\text {Miss }}=117\right)$ were much more common than false positives $\left(\mathrm{M}_{\text {False Positive }}=79\right), \chi^{2}(1, N=1,959)=73.324, p<.001$, just as suggested by Lee et al. (2005). Also, as the error magnitude increases, the ratio of misses to false positives also increases, which is consistent across the 10 optimal parameter settings. Figure 1 shows the frequency of the two types of errors (misses vs. false positives) as a function of the absolute z-score error.


Figure 1. The frequency of the two types of misjudgment by LSA as a function of the absolute z-score error.

Effect of the Parameters on the Frequency of Misses and False Positives Although the distribution of the two types of errors by LSA at optimal parameters was the primary focus of this study, we also examined the ratio of misses to false positives across all the 441 parameter settings. The results showed that the ratios were systematically affected by the interaction between the background document size and dimensionality. That is, the ratio of misses to false positives increased as the dimensionality increased. However, the degree of increase is getting less prominent as the background document size increases. In other words, although there were more misses than false positives in general, the disproportion of misses over false positives is more prominent at high dimensions with small number of background documents.

Effect of the Number of Background Documents on Correlation between Humans and LSA One of the most striking findings above was the strong effect of background document size on LSA's document similarity representation. As shown in Table 2, employing more background documents (combined with an appropriate dimensionality) tends to significantly improve LSA's document similarity judgments. To illustrate the significant effect of background documents, we plotted the correlation between LSA and human ratings for three background document sizes $(0,314$, and 4,486 ) and the nine weighting schemes as a function of dimensionality (Figure 2). The graph illustrates (a) the strong effect of the number of background documents, (b) important effect of dimensionality when the background document size is small (i.e., the left side of the graph), and (c) the relative unimportance of weighting schemes.


Figure 2. The correlation of ratings between humans and LSA for three background document conditions ( 0,314 and $4,486)$ and nine weighting schemes as a function of dimensionality.

One may suspect that including even more background documents would further increase the correlation. However, to make a positive impact on LSA's performance, the background documents should not only be numerous but also relevant to the content of the target documents (Foltz, Britt, \& Perfetti, 1995). For example, Stone, Dennis, and Kwantes (2011) tested the effect of various kinds of background documents on LSA's document similarity judgments, using the same 50 target news articles examined in this study. They tested 55,021 Canada Toronto Star newspaper articles (miscellaneous gossip paragraphs, from the year 2005) and 10,000 articles selected by the researchers from online encyclopedia, Wikipedia (http://www.wikipedia.org/). However, the highest correlation between humans and LSA obtained was about .10 with the gossip news articles as background and . 40 with the Wikipedia background documents. These correlations were significantly lower than the highest correlation of .60 obtained in Lee et al. (2005) and .70 in the current study, despite utilizing only 314 and 4,486 background documents, respectively. Therefore, not only the size but also the relevance of background documents to the target documents seem to be critical for LSA's document similarity judgments.

If target documents came from a certain population (e.g., specific news corpus), we recommend using documents from the same population as background documents. In our case, employing background documents of 4,172 news articles that came from the same population as the target articles increased the correlation between humans and LSA from .60 to .70 .

## Conclusion

Lee et al. (2005) suggested that the primary weakness of computational models of document similarity is failing to pick up on some of the strong document similarities that humans easily detect. To test this hypothesis, we compared the document similarity ratings made by humans and LSA based on a range of parameter combinations. Then we
identified the frequency and degree of large misses and large false positives under optimal parameters of LSA. The results confirmed that LSA makes more misses than false positives, especially among the most severe errors.

The results also suggest that if one attempts to further improve models of text document similarity by reducing its errors relative to humans, the misses rather than the false positive would be the primary focus of the revision. More specifically, one should look for ways to help models pick up on some of the strong semantic similarities that they currently miss.

## Potential Causes of Misses and False Positives

An obvious follow-up question of this study is what causes LSA's greatest misses and false positives. Considering that LSA's basis, the vector space model, judges document similarity based on the overall word similarity between two documents, a potential cause of error is that LSA misses or falsely overestimates the semantic similarity of some word pairs from two documents. In fact, there are various cases where LSA cannot help but miss some of the word similarities, which in turn would cause one type of error, miss. For example, although "United States", "US", "U.S.", and "U.S.A" refer to the same country, they may not be recognized as the same entity in the word-by-document matrix for various reasons: because they are not a single word (United States), too short to be included (US or U.S. after the special character removal), or happen to match an excluded stop word (US and the pronoun us). However, humans would correctly recognize them and utilize these words for document similarity judgment.

Also, some words (especially proper nouns including human names) may occur in the target documents but not in the background documents, preventing LSA from utilizing those words in judging document similarity. However, those words could be critical for humans to judge the document similarity. Then, LSA may judge document pairs including those words to be less related than humans would do (i.e., leading to a miss).

The above-mentioned potential cause of misses (i.e., LSA misses document similarity because it misses word similarity in document pairs) could be further supported if LSA's document similarity scores do correspond to the overall word similarity between two documents. To confirm this, we calculated the correlation between the 1,225 document pairs' LSA cosine scores and the average LSA cosine scores of every possible word pair from each of the document pairs. We found a correlation of .73 from this analysis, indicating that LSA's document similarity is heavily relying on the overall word similarity in document pairs.

Similar to the potential cause of misses by LSA addressed in the above, a potential cause of false positives by LSA is that LSA mistakenly perceives semantic similarity between words that are in fact unrelated. Table 4 shows 10 word pairs that were judged to be highly related by humans and LSA, respectively in one of the document pairs used in this study. Although LSA does generally make reasonable judgments on
word relatedness, some word pairs judged to be highly related by LSA do not seem to have a meaningful relationship. For example, design and document were the most strongly related word pair to LSA, despite being seemingly unrelated. Thus, LSA will occasionally overestimate the relatedness of the document pairs that include this word pair.

Table 4. The 10 most related words pairs to humans and LSA from a pair of news articles.

| Highly related <br> word pairs by <br> human | Ratings <br> $(1-5$ scale) | Highly related <br> word pairs by <br> LSA | Ratings <br> (z-score) |
| :--- | :---: | :--- | :---: |
| dollar-money | 5.00 | design-document | 6.87 |
| job-money | 5.00 | increase-rise | 5.61 |
| angrily-attack | 4.90 | paid-worker | 3.97 |
| plan-target | 4.83 | effect-target | 3.38 |
| increase-profit | 4.80 | group-work | 3.37 |
| money-profit | 4.80 | effect-increase | 3.08 |
| cost-lawsuit | 4.78 | disclosure-profit | 2.89 |
| job-meet | 4.75 | disclosure-financial | 2.83 |
| agreement-plan | 4.73 | commonwealth-deal | 2.41 |
| job-paid | 4.73 | australia-target | 2.32 |

An alternative hypothesis regarding the misses and false positives of LSA is that, when judging document similarity, humans do not rely on the overall word similarity as much as LSA does. As Griffiths, Steyvers, and Tenenbaum (2007) suggested, humans may catch the gist of each document and compare the semantic representations of the gist rather than relying on the overall similarity of words in the documents. Then, two documents with a large overlap of words but with different topics would be regarded unrelated by humans although they could be highly related to LSA (resulting in false positives). To assess to what degree human document similarity judgments rely on the overall word similarity, one could examine the correlation between human document similarity ratings and the average of the human similarity ratings for all the possible word pairs in a given document pair. If humans do not rely on the overall word similarity as much as LSA does, then the correlation would not be as high as the corresponding correlation of LSA.

## References

Bullinaria, J. A., \& Levy, J. P. (2006). Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39, 510-526.
Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., \& Harshman, R. A. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391-407.
Dumais, S. T. (1991). Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, and Computers, 23, 229-236.

Foltz, P. W., Britt, M. A., \& Perfetti, C. A. (1995). Measuring text influence on learning history. Proceedings of the Fifth Annual Winter Text Conference, Jackson, WY.
Furnas, G. W., Deerwester, S. C., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., \& Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. Proceedings of the Eleventh Annual International ACM 81 SIGIR Conference on Research and Development in Information Retrieval. 465-480. Grenoble, France.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114, 211-244.
Jung, K. (2013). Mismatches between humans and latent semantic analysis in document similarity judgments (Doctoral dissertation).
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
Lee, M. D., \& Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. Psychonomic Bulletin \& Review, 9, 43-58.
Lee, M. D., Pincombe, B. M., \& Welsh, M. B. (2005). An empirical evaluation of models of text document similarity. Proceedings of Twenty Seventh Annual Conference of the Cognitive Science Society (pp. 1254-1259). Mahwah, NJ: Erlbaum.
Lintean, M., Moldovan, C., Rus, V., \& McNamara D. S. (2010). The role of local and global weighting in assessing the semantic similarity of texts using Latent Semantic Analysis. Proceedings of the Twenty Third International Florida Artificial Intelligence Research Society Conference. Daytona Beach, FL.
Nakov, P., Popova, A., \& Mateev, P. (2001). Weight functions impact on LSA performance. Proceedings of the Recent Advances in Natural Language Processing (pp. 187-193). Tzigov Chark, Bulgaria.
Pincombe, B. M. (2004). Comparison of human and LSA judgments of pairwise document similarities for a news corpus (Tech. Rep. DSTO-RR-0278). Adelaide, Australia: Australian Defense Science and Technology Organization, Intelligence, Surveillance and Reconnaissance Division.
Stone, B., Dennis, S., \& Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis, Topics in Cognitive Science, 3, 92-122.
Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.
Yeh, E., Ramage, D., Manning, C. D., Agirre, E., \& Soroa, A. (2009). WikiWalk: Random walks on Wikipedia for semantic relatedness. Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (pp. 41-49). Stroudsburg, PA, USA: Association for Computational Linguistics.

# "The Polar Express" is Bipolar: Critical Film Reviews Influence Uncanny Valley Phenomenon in Semi-Realistic Animation Films 

Jari Kätsyri (jari.katsyri@maastrichtuniversity.nl)<br>Department of Cognitive Neuroscience, Oxfordlaan 55, 6229 EV Maastricht, The Netherlands


#### Abstract

Previous research suggests that semi-realistic animation films such as The Polar Express are representative of the uncanny valley (UV) hypothesis, which predicts that highly humanlike artificial characters can appear eerie. In the present study, we investigated the extent to which critical film reviews can influence the perceived eeriness of such films. The reviews were adopted from authentic ones and expressed either negative or positive attitudes towards the animation techniques. Audiovisual speech asynchrony, which is known to induce eeriness, was included as an objective manipulation. Our results showed large review tone effects for both implicit and explicit eeriness evaluations. In contrast, speech asynchrony failed to elicit significant effects. These results demonstrate that critical film reviews representing opposite attitudinal poles can elicit consistent changes in the viewers' evaluations of semi-realistic animations. The present findings cannot, however, be taken as evidence against the UV hypothesis itself in computer-generated characters.


Keywords: Uncanny Valley hypothesis; anthropomorphism; social influence; animation films

## Introduction

Realistic computer-generated (CG) characters are commonly used in lieu of human actors when creating social signal stimuli for neurocognitive experiments. Although this practice has several advantages such as the ease of creating and manipulating stimuli, it is conceivable that subtle flaws in highly realistic CG characters could elicit unintended negative reactions in human participants. The Uncanny Valley (UV) hypothesis predicted such reactions already in the 1970s. The original hypothesis suggested that observing highly realistic robots or other mechanical devices can elicit negative feelings characterized by eeriness and lack of familiarity (Mori, 1970/2012).

## Empirical Evidence for Uncanny Valley

Although the UV hypothesis is compelling and it seems to have plentiful anecdotal evidence (see below), empirical evidence for its existence is still elusive (Kätsyri, Förger, Mäkäräinen, \& Takala, 2015; S. Wang, Lilienfeld, \& Rochat, 2015). A lack of commonly agreed-upon operationalization is a fundamental problem for testing the hypothesis. According to a conservative interpretation, any artificial-to-human stimulus continuum should elicit a nonlinear evaluation curve in which the most negative evaluations occur at levels preceding the full degree of human-likeness. The bulk of empirical studies have, however, shown that increasingly human-like stimuli simply tend to elicit more positive evaluations in a linear manner (Kätsyri et al., 2015).

Although the UV is apparently not triggered by all artificial-to-human continua, it could still occur under some specific conditions. Accumulating evidence has shown that negative evaluations can be elicited by a mismatch between artificial and realistic features (Seyama \& Nagayama, 2007; MacDorman \& Chattopadhyay, 2016) or by categorization difficulty (Burleigh \& Schoenherr, 2015; Yamada, Kawabe, \& Ihaya, 2013; however, see MacDorman \& Chattopadhyay, 2016). Almost all of this evidence comes from rigorous manipulations of CG and human faces. A few other studies using naturalistic - and plausibly more ecologically valid - stimuli have provided evidence for the UV in pictures of prosthetic hands (Poliakoff, Beach, Best, Howard, \& Gowen, 2013) and real-world robot faces (Mathur \& Reichling, 2016).

## Uncanny Valley and Animated Film Characters

The UV hypothesis was rediscovered at the beginning of the present millennium (e.g., MacDorman, 2005) and has since received increasing research interest. First fully computeranimated films with deliberately realistic characters, including Final Fantasy (Aida, Lee, Sakai, Sakaguchi, \& Sakakibara, 2001) and The Polar Express (Goetzman, Starkey, Teitler, \& Zemeckis, 2004), were released roughly around the same time. Perhaps not coincidentally, these and some other animation films have been adopted as anecdotal examples of the UV, with frequent citations in scientific research reports (e.g., Piwek, McKay, \& Pollick, 2014) and popular scientific magazines (e.g., Spinney, 2017).

It seems plausible that semi-realistic animated film characters could appear eerie in the sense of the UV. Early computer graphics methods in particular tended to suffer from shortcomings in modeling light reflections from the skin and eyes, for example (e.g., Wechsler, 2002). These and other subtle flaws on otherwise quite realistic characters could elicit sufficient featural mismatch to make them appear uncanny. In general, semi-realistic animated film characters stand out from the intentionally caricatured and exaggerated characters that are the norm in traditional animation (Kaba, 2012).

To our best knowledge, semi-realistic animated film characters, cartoonish animated film characters, and human actors have been explicitly compared with each other only recently in our own previous study (Kätsyri, Mäkäräinen, \& Takala, 2017). The results showed that semi-realistic characters are considered more eerie and selected more often as representative examples of the UV hypothesis than cartoonish characters or human actors. These results suggest that semi-realistic animated film characters capture some aspects of the UV hypothesis.

## Critical Film Reviews and Uncanny Valley

Semi-realistic animation films, which are a rare exception among the traditional computer-animated films, have attracted a mixed but predominantly negative critical reception. Some of the published reviews have explicitly characterized the realistic characters as "soulless" and "creepy" (e.g., Savlov, 2004). In the present study, we explore the extent to which this kind of critical reviews can influence the evaluation of animated characters in semirealistic films.

A research tradition in social psychology has shown that individuals' attitudes and behavior are influenced by those of others (Bohner \& Dickel, 2011). In one form of social influence, credible information from others is accepted as valid evidence about reality (Deutsch \& Gerard, 1955). Perceived expertise on the subject matter has long been considered as one source of credibility (Kelman, 1956). On the other hand, negative attitudes are known to exert greater influence than equally intense positive attitudes (Cacioppo, Gardner, \& Berntson, 1997). An excellent demonstration of this is that negative peer reviews exert greater influence than positive reviews on individuals' attitudes towards consumer products (Lee, Park, \& Han, 2008). Taken together, negative reviews from film experts should exert an influence. Consistently, negative film reviews are known to elicit decreased box profit especially during early presentation weeks (Basuroy, Chatterjee, \& Ravid, 2003).

## Present Study

The present study aims to compare the effects of critical film reviews and objective differences between films on the UV phenomenon in semi-realistic animation films. When comparing small sets of individual films (as in Kätsyri et al., 2017), confound effects cannot be fully excluded. Instead, in the present study, we decided to manipulate audiovisual speech asynchrony in the same set of films. Audiovisual speech asynchrony was selected as the objective manipulation since a previous study has shown that it elicits eeriness in virtual characters (Tinwell, Grimshaw, \& Nabi, 2015). Increased eeriness was observed at a $200-\mathrm{ms}$ asynchrony at the earliest, and it was more pronounced when the auditory stream preceded rather than followed the visual stream. These findings are consistent with psychophysics literature, in which simultaneity judgments for audiovisual speech occur roughly in the time frame of -130 ms (audio first) to +220 ms (vision first) (Conrey \& Pisoni, 2006).

Participants were asked both to rate the eeriness of films immediately after viewing them (implicit evaluations) and to score how representative each film was of UV after having received full debriefing (explicit evaluations). Our main hypotheses were:

H1: Majority negative reviews will elicit higher implicit and explicit eeriness than majority positive reviews.

H2: Asynchronous audiovisual speech will elicit higher implicit and explicit eeriness than synchronous audiovisual speech.

## Method

## Participants

Forty participants ( 20 women) with a mean age of 29.7 years ( $\mathrm{SD}=8.5$ ) took part in the study. All participants were native Finnish speakers with a good command of written English. Three participants who scored poorly in a postexperimental reading comprehension test were replaced with new participants. Participants reported having normal hearing, normal (or corrected) visual acuity, and no history of dyslexia. Participants were compensated with two movie tickets. Study protocol was approved by the Aalto University Research Ethics committee.

## Stimuli

Animation Film Scenes Stimuli were four film scenes from animation films Final Fantasy (Aida et al., 2001), The Polar Express (Goetzman et al., 2004), Beowulf (Rapke, Starkey, \& Zemeckis, 2007), and Tintin (Jackson, Kennedy, \& Spielberg, 2011). Films were selected on the basis of our previous evaluation (Kätsyri et al., 2017). All films were fully computer-animated and used motion-capture techniques for character animation (most character motions were captured from real actors). Film scenes were extracted from official DVD releases, depicted spoken dialogue, and did not contain violence or nudity. These copyrighted materials were presented under an external license.

Audiovisual Speech Synchrony Original audiovisual tracks for the predominantly spoken film scenes were used for the synchronous speech condition. For the asynchronous condition, auditory stream in each film scene was modified to precede the visual stream by 200 ms .

Film Reviews For each film, six fictional but plausible film reviews (three positive and three negative) were created. For authenticity, all reviews were displayed in English. Available reviews were first extracted from Metacritic database (http://www.metacritic.com). Brief statements expressing positive and negative attitudes towards animation techniques were then extracted and modified to produce an initial set of 12 reviews (half positive). These reviews were evaluated by 19 additional participants for expressed attitude ( -4 very negative to +4 very positive). Negative and positive reviews were clearly differentiated $(\mathrm{M}=-2.1$ and $+2.6, \mathrm{SD}=1.2$ and 0.8$), F(1,18)=114.95, p$ $<.001, \eta_{\mathrm{p}}^{2}=.87$.

Final selected stimuli were three positive and three negative reviews with roughly similar lengths ( $M=82$ and 79 words) and similar mean ratings across films. The reviews focused on either specific visual features (e.g., "[...] their eyes, supposedly the windows to the soul, are more often dead than alive") or the overall impression generated by highly realistic animations (e.g., "[...] the soft-edged, photorealistic style - suspended somewhere between liveaction and animation, fairy tale and reality - feels entirely appropriate in this context").

Table 1: Counterbalancing between films and conditions.

| Group | Negative/ <br> Asynch. | Positive/ <br> Asynch. | Negative/ <br> Synch. | Positive/ <br> Synch. |
| :--- | :--- | :--- | :--- | :--- |
| 1 | F1 | F2 | F3 | F4 |
| 2 | F2 | F3 | F4 | F1 |
| 3 | F3 | F4 | F1 | F2 |
| 4 | F4 | F1 | F2 | F3 |

## Procedure

Film preference evaluation was used as a distractor task to avoid making the true study objectives too obvious. Specifically, participants were explained that the study aimed to investigate the effects of critical film reviews on attitudes and consumer behavior. Full debriefing was given at the end of the study.

Participants were assigned randomly to the conditions of a 4 (group) $\times 2$ (majority review tone: negative, positive) $\times$ 2 (speech synchrony: asynchronous, synchronous) mixed model design. Majority review tone and speech synchrony were counterbalanced with films using a $4 \times 4$ Latin square as shown in Table 1. Participants were asked to read film reviews carefully and told that a memory task would follow the evaluation. Each trial proceeded as follows:

- Title and a brief description of the film were presented with a minimum reading time of 20 s .
- Four reviews were presented (min. 10 s per each). For the majority negative/positive condition, three reviews were negative/positive and one was of the opposite valence. Reviews were paired randomly with fictional reviewer names (half female) and well-known magazine titles.
- An overview of the film scene was presented (min. 20 s ).
- The movie scene was played back ( 164 s to 182 s ).
- The participant answered a series of self-report questions.

The experiment was carried out on a desktop computer running Psychtoolbox (Brainard, 1997) for Matlab. Film stimuli were displayed on a 24 " wide-screen display (Eizo ColorEdge CG241W) at horizontal resolution of 1024 pixels and vertical resolution depending on aspect ratio (424 to 548 pixels). Participants were seated 80 cm from the display. Auditory sound tracks were standardized at -5 dB and played on closed earphones at a loud but comfortable level. The whole experiment took approximately 60-90 mins.

## Dependent Variables

Film Preference For attitude towards the film (cf. Voorveld, 2011), participants rated whether they enjoyed each film, were content with it, and found it interesting. This scale had a good internal reliability (Cronbach's $\alpha=0.96$ ). For consumption intent (cf. A. Wang, 2006), participants rated whether they would like to see the full film (again if seen), recommend it to a friend, and pay for seeing it (good reliability: $\alpha=0.91$ ). These and all other ratings were given on a 7 -step Likert scale ranging from total disagreement (1) to total agreement (7).

Table 2: Means (and SEMs) by majority review tone.

| Variable | Negative | Positive | $F(1,36)$ | $p$ | $\eta_{\mathrm{p}}{ }^{2}$ |
| :--- | :---: | :--- | :---: | :---: | :---: |
| Film attitude | $4.6(0.2)$ | $5.0(0.1)$ | 6.94 | .027 | .13 |
| Consumption intent | $3.8(0.2)$ | $4.4(0.2)$ | 8.72 | .006 | .19 |
| Human-likeness | $4.7(0.2)$ | $5.0(0.1)$ | 8.08 | .007 | .18 |
| Eeriness | $3.2(0.2)$ | $2.8(0.2)$ | 7.46 | .010 | .17 |
| Representativeness | $10.5(0.7)$ | $8.2(0.8)$ | 6.52 | .015 | .15 |

Eeriness (Implicit) and Human-likeness For eeriness, participants rated whether the characters appeared eerie, creepy, and strange (good reliability: $\alpha=0.84$ ). Although not included in any hypothesis, human-likeness scale was also included because it is a focal dimension of the UV (Mori, 1970/2012). Participants were asked to rate whether the characters appeared realistic, cartoonish, and similar to real people (adequate reliability: $\alpha=0.71$ ).

Representativeness (Explicit Eeriness) After finishing the film evaluation task, participants received a full debriefing of the experiment and the UV hypothesis. Participants were then given four plastic cards depicting the four animation films and asked to place these cards on a cardboard depending on how representative the films were of the UV (left side: not at all, right side: perfectly). Responses were scored from 0 to 100 based on the cards' physical positions.

## Analysis

Data were subjected to a mixed-design GLM analysis in SPSS (version 24). Effect sizes were quantified using partial $\eta^{2}$ values and classified as large ( $\eta^{2} \geq .14$ ), medium ( $\eta^{2} \geq$ .06), or small ( $\eta^{2} \geq .01$ ) based on Cohen's (1992) guidelines.

## Results

## Manipulation Checks

After the experiment, participants were asked to read 12 review statements for each film, and to tell whether they had read these reviews during the experiment (yes/no). Statements included four correct and eight incorrect options and an equal number of positive and negative reviews. Recognition performance, as indexed by $d$ ' sensitivity score (Stevens \& Pashler, 2002), was well above chance level ( $M$ $=1.47, S D=0.59), t(39)=15.80, p<.001$, showing that participants had attended and comprehended the reviews.

## Negative Reviews

As can be seen in Table 2, majority review tone elicited significant effects for all dependent variables. Supporting H1, majority negative reviews elicited higher eeriness ratings and representativeness scores than majority positive reviews. In the distractor task, majority negative reviews elicited decreased film attitude and consumption intent ratings. Human-likeness ratings were decreased for majority negative review reviews. Review tone had a medium effect for film attitude and a large effect for all other variables.

Table 3: Means (and SEMs) by speech synchrony.

| Variable | Asynch. | Synch. | $F(1,36)$ | $p$ | $\eta_{\mathrm{p}}{ }^{2}$ |
| :--- | :--- | :---: | :---: | :---: | :---: |
| Film attitude | $4.8(0.2)$ | $4.8(0.2)$ | 0.02 | .884 | .00 |
| Consumption intent | $4.2(0.2)$ | $4.0(0.2)$ | 0.46 | .503 | .01 |
| Human-likeness | $4.8(0.2)$ | $4.9(0.2)$ | 0.21 | .650 | .01 |
| Eeriness | $3.0(0.2)$ | $3.0(0.2)$ | 0.07 | .790 | .00 |
| Representativeness | $9.9(0.8)$ | $8.7(0.9)$ | 1.26 | .269 | .03 |

## Speech Asynchrony

As shown in Table 3, speech synchrony failed to elicit significant main effects on eeriness, representativeness, or any other variable. Asynchrony elicited slightly higher representativeness scores than synchrony; however, this effect was small and clearly non-significant. Hence, the results did not support hypothesis H2.

## Interactions

Fig. 1 illustrates interaction effects between review tone and speech synchrony. Visual inspection suggests that review effects for some variables were qualified by an interaction with speech synchrony. Consistently, significant and large interaction effects were observed for film attitude, $F(1,36)$ $=7.05, p=.012, \eta_{\mathrm{p}}^{2}=.16$, consumption intent, $F(1,36)=$ $6.42, p=.016, \eta_{\mathrm{p}}^{2}=.15$, and human-likeness, $F(1,36)=$ $7.59, p=.009, \eta_{\mathrm{p}}^{2}=.17$. Simple effect tests revealed that for these variables, negative reviews elicited lower ratings than positive reviews only in the asynchronous condition ( $p \leq$ . $001, \eta_{\mathrm{p}}^{2} \geq .254$ ). Importantly, interaction effects were not significant for eeriness, $F(1,36)<1, p=.643, \eta_{p}^{2}=.01$, or representativeness, $F(1,36)=1.21, p=.278, \eta_{p}^{2}=.03$. Hence, H1 was not affected by this interaction.

## Confounds

The following confound effects were tested and excluded using mixed model analyses (Hoffman \& Rovine, 2007): previous familiarity (film seen or not seen), awareness of UV, awareness of review manipulation, and awareness of asynchrony manipulation. A mixed model equivalent to GLM was first specified, and confounds were then added and tested individually. The pattern of significant results was not changed by the inclusion of any tested confound into the model.

Awareness variables were derived from a semi-structured interview conducted the end of the experiment. Briefly, UV awareness meant that participants (18\%) had heard about the UV hypothesis, were able to explain it correctly, and associated it with this experiment. For review awareness, participants $(30 \%)$ noticed that films were preceded unequally by positive and negative reviews, were able to choose which films received majority positive/negative reviews, and were aware of the manipulation. For asynchrony awareness, participants (30\%) mentioned audiovisual asynchrony spontaneously and were aware of the manipulation. Notably, only $8 \%$ of participants were able to choose asynchronous films correctly, however.


Figure 1: Means (and SEMs) by majority review tone and speech synchrony. Note that representativeness scores are truncated to 0-20.

## Discussion

The aim of the present study was to provide an objective comparison of the effects of critical reviews and objective differences (audiovisual speech synchrony or asynchrony) on the UV phenomenon in semi-realistic animation films. We hypothesized that critical film reviews focusing on the appearance of realistic animation technologies would exert an influence on both implicitly and explicitly evaluated eeriness. Indeed, our results showed that review tone (majority positive or negative) elicited large effects both on eeriness ratings given immediately after film viewing (implicit evaluations) and representativeness scores given after full debriefing of the UV (explicit evaluations). Unexpectedly, audiovisual speech asynchrony failed to elicit significant effects.

The present results demonstrate that social influence (Bohner \& Dickel, 2011) originating from critical reviews can exert a large influence on the subjective eeriness of semi-realistic animated film characters. Results from explicit evaluations are particularly important because they demonstrate that critical reviews can affect individuals' attitudes specifically in the UV context. A closer inspection of the results suggests that participants did not consider the animation films as particularly eerie: $95 \%$ confidence interval for the overall eeriness ratings was [2.7, 3.3] on the 7 -step scale (i.e., below midpoint) and for the overall representativeness scores [8.1, 10.5] on the visual scale from 0 to 100 . In the absence of strong effects in the stimuli, participants may have relied on information from the film reviews instead.

Although not predicted beforehand, majority negative reviews also elicited lower human-likeness ratings than majority positive reviews - but only in film scenes with asynchronous audiovisual speech dialogue. This result differs clearly from eeriness evaluations, which were not sensitive to the interaction between review tone and synchrony. This suggests that review effects were more robust for evaluated eeriness than for evaluated humanlikeness.

The failure to find significant effects for the $200-\mathrm{ms}$ audiovisual speech asynchrony is surprising given that such delay should be noticeable on human faces (Conrey \& Pisoni, 2006) and elicit eeriness in virtual characters as well
(Tinwell et al., 2015). We note that several participants ( $30 \%$ ) commented spontaneously upon asynchrony in the films even though much fewer ( $8 \%$ ) were actually able to identify which films were truly asynchronous. A possible explanation for this discrepancy is that animated lip movements were sufficiently inaccurate to begin with in some of the films so that additional delay had little or no effect on the perceived asynchrony. It is also possible that asynchrony effects may have been concealed by the stronger review tone effects. In partial support, secondary evaluations - but not eeriness evaluations - were affected by the interaction between review tone and asynchrony.

A significant limitation of the present study is that, in the absence of statistically significant effect for asynchronous speech, the magnitude of critical review effect cannot be meaningfully interpreted. Hence, the present results can tell us that critical film reviews exert a large influence on the perceived eeriness of semi-realistic animated film characters, but it cannot tell us whether these effects are weaker or stronger than those elicited by genuine physical differences in the stimuli.

The present results are nevertheless relevant for the UV phenomenon because semi-realistic animated film characters have already been linked to this phenomenon previously (Kätsyri et al., 2017). Methodological limitations in this previous study warrant some caution in interpreting the findings; for example, eeriness ratings were close to minimum for all film characters. It is particularly clear that semi-realistic animation films do not elicit any such aversion that could be implicitly or explicitly likened to human corpses or zombies (cf. Mori, 1970/2012). Nevertheless, semi-realistic animated characters received slightly but statistically significantly higher eeriness ratings and were considered more often as being representative of the UV hypothesis than other types of films. Hence, these previous findings provide support for subtle UV effects in semi-realistic animated film characters. The present results add to this by demonstrating that such effects can also be elicited by critical film reviews. It should be emphasized that the present study did not aim to "find" the UV for animated characters, as this was already done previously.

Precisely for the same reason, the present results should be considered meaningful even though the review effect was not tested for cartoonish animated characters or real human actors. In fact, the present film reviews that focused explicitly on the disadvantages of realistic animation techniques could not have been paired with any other types of films, at least not without arousing suspicion in the study participants. Although fictitious negative reviews focusing on the disadvantages of traditional animation could possibly have been created, it is likely that such stimuli would have been considered implausible as this kind of reviews do not appear to exist in reality.

Taken together, the present findings demonstrate that critical film reviews representing opposite attitudinal poles negative or positive - towards realistic animation techniques can elicit consistent changes in individuals' evaluations.

In this sense, the UV phenomenon in semi-realistic animation films could be characterized as being bipolar.

The present findings should obviously be interpreted with caution when it comes to explaining the complexities of attitude formation in real life. Allowing some speculation, it is possible that critical film reviews might have contributed to the wide-spread adoption of specific animation films as anecdotal examples of the UV. The present results empathetically cannot be taken to mean that the UV would be just a media phenomenon, however. First, the present study does not allow comparing the effects of critical film reviews to genuine differences between stimuli. Second, empirical evidence for the UV phenomenon has already begun to accumulate, especially from studies with featurally mismatching (Seyama \& Nagayama, 2007; MacDorman \& Chattopadhyay, 2016) and naturalistic stimuli (Mathur \& Reichling, 2016; Poliakoff et al., 2013).

Overall, the present findings highlight the importance of social factors in evaluating contemporary technological artefacts, in particular those that involve human-like characteristics.

## Acknowledgments

This study was carried out at the Department of Computer Science, Aalto University, Finland, with an individual research grant from the Emil Aaltonen Foundation. The work was also supported by the Marie Skłodowska-Curie Individual Fellowship (H2020 programme) under grant agreement 703493. The author wishes to thank Prof. Tapio Takala, M.Sc.(Tech.) Meeri Mäkäräinen, and D.Tech. Klaus Förger for previous collaboration, which provided the foundation for the present study.

## References

Aida, J., Lee, C., Sakai, A. (Producer), Sakaguchi, H., \& Sakakibara, M. (Director). (2001). Final Fantasy: The Spirits Within [Motion Picture]. USA, Japan: Columbia Pictures.
Basuroy, S., Chatterjee, S., \& Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. Journal of Marketing, 67(4),

103-117. doi:10.1509/jmkg.67.4.103.18692
Bohner, G., \& Dickel, N. (2011). Attitudes and attitude change. Annual Review of Psychology, 62(1), 391417. doi:10.1146/annurev.psych.121208.131609

Brainard, D. H. (1997). The psychophysics toolbox. Spatial Vision, 10, 433-436.
Burleigh, T. J., \& Schoenherr, J. R. (2015). A reappraisal of the uncanny valley: categorical perception or frequency-based sensitization? Cognitive Science, 5, 1488. doi:10.3389/fpsyg.2014.01488
Cacioppo, J. T., Gardner, W. L., \& Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. Personality and Social Psychology Review, 1(1), 3-25.

Cohen, J. (1992). A power primer. Quantitative Methods in Psychology, 112(1), 155-159.
Conrey, B., \& Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. The Journal of the Acoustical Society of America, 119(6), 4065-4073.
Deutsch, M., \& Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. The Journal of Abnormal and Social Psychology, 51(3), 629.
Goetzman, S., Starkey, S., Teitler, W. (Producer), \& Zemeckis, R. (Director). (2004). The Polar Express [Motion Picture]. USA: Warner Brothers Pictures.
Hoffman, L., \& Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. Behavior Research Methods, 39(1), 101-117. doi:10.3758/BF03192848
Jackson, P., Kennedy, K. (Producer), \& Spielberg, S. (Director). (2011). The Adventures of Tintin: The Secret of the Unicorn [Motion Picture]. USA: Columbia Pictures.
Kaba, F. (2013). Hyper-realistic characters and the existence of the Uncanny Valley in animation films. International Review of Social Sciences and Humanities, 4(2), 188-195.
Kätsyri, J., Förger, K., Mäkäräinen, M., \& Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. Frontiers in Psychology, 6, 390. doi:10.3389/fpsyg. 2015.00390
Kätsyri, J., Mäkäräinen, M., \& Takala, T. (2017). Testing the "uncanny valley" hypothesis in computeranimated film characters: An empirical evaluation of natural film stimuli. International Journal of Human-Computer Studies, 97, 149-161.
Kelman, H. C. (1956). Processes of opinion change. Public Opinion Quarterly, 25, 57-78.
Lee, J., Park, D.-H., \& Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. Electronic Commerce Research and Applications, 7(3), 341-352. doi:10.1016/j.elerap.2007.05.004
MacDorman, K. F. (2005). Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it? (pp. 108-118). Presented at the CogSci-2005 Workshop: Toward Social Mechanisms of Android Science, Stresa, Italy.
MacDorman, K. F., \& Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition, 146, 190-205. doi:10.1016/j.cognition.2015.09.019
Mathur, M. B., \& Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. Cognition, 146, 22-32. doi:10.1016/j.cognition.2015.09.008

Mori, M. (1970/2012). The Uncanny Valley (K. F. MacDorman \& N. Kageki, Trans.) IEEE Robotics \& Automation Magazine, 19(2), 98-100. doi:10.1109/MRA.2012.2192811
Piwek, L., McKay, L. S., \& Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. Cognition, 130(3),

271-277. doi:10.1016/j.cognition.2013.11.001
Poliakoff, E., Beach, N., Best, R., Howard, T., \& Gowen, E. (2013). Can looking at a hand make your skin crawl? Peering into the uncanny valley for hands. Perception, 42(9), 998-1000. doi:10.1068/p7569
Rapke, J., Starkey, S. (Producer), \& Zemeckis, R. (Director). (2007). Beowulf [Motion Picture]. USA: Paramount Pictures.
Savlov, M. (2004, November 12). The Polar Express. The Austin Chronicle. Retrieved from http://www.austinchronicle.com/calendar/film/200 4-11-10/the-polar-express/
Seyama, J., \& Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. Presence: Teleoperators and Virtual Environments, 16(4), 337-351. doi:10.1162/pres.16.4.337
Spinney, L. (2017, October 29). Exploring the uncanny valley: Why almost-human is creepy. New Scientist.
Stevens, S. S., \& Pashler, H. E. (Eds.). (2002). Stevens' Handbook of Experimental Psychology (3rd ed). New York: John Wiley \& Sons.
Tinwell, A., Grimshaw, M., \& Nabi, D. A. (2015). The effect of onset asynchrony in audio visual speech and the uncanny valley in virtual characters. International Journal of the Digital Human, 2(2), 97-110.
Voorveld, H. A. M. (2011). Media multitasking and the effectiveness of combining online and radio advertising. Computers in Human Behavior, 27(6), 2200-2206. doi:10.1016/j.chb.2011.06.016
Wang, A. (2006). The effects of expert and consumer endorsements on audience response. Journal of Advertising Research, 45(04), 402. doi:10.1017/S0021849905050452
Wang, S., Lilienfeld, S. O., \& Rochat, P. (2015). The uncanny valley: Existence and explanations. Review of General Psychology, 19(4), 393-407. doi:10.1037/gpr0000056
Wechsler, L. (2002, January 6). Why is this man smiling? Wired, 10.06.
Yamada, Y., Kawabe, T., \& Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. Japanese Psychological Research, 55(1), 20-32. doi:10.1111/j.14685884.2012.00538.x

# Self-other distinction in the motor system during social interaction: A computational model based on predictive processing 

Sebastian Kahl (skahl@uni-bielefeld.de)<br>Stefan Kopp (skopp@uni-bielefeld.de)<br>Social Cognitive Systems Group, Faculty of Technology, Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany


#### Abstract

During interaction with others, we perceive and produce social actions in close temporal distance or even simultaneously. It has been argued that the motor system is involved in both processes, but how does it distinguish in this processing between self and other? In this paper, we present a model of self-other distinction within a hierarchical sensorimotor system that is based on principles of perception-action coupling and predictive processing. For this we draw on mechanisms assumed for the integration of cues to generate sense of agency, i.e., the sense that an action is self-generated. We report results from simulations of different social scenarios, showing that the model is able to solve the problem of the dual use of the sensorimotor system.


Keywords: perception-action coupling; social cognition; mirroring; dual-use; sense of agency; predictive processing

## Introduction

In everyday social interaction we constantly try to deduce and predict the underlying intentions behind others' social actions, like facial expressions, speech, gestures, or body posture. This is no easy problem and the underlying cognitive mechanisms and neural processes even have been dubbed the ,dark matter" of social neuroscience (Przyrembel, Smallwood, Pauen, \& Singer, 2012). Action recognition is commonly believed to rest upon principles of predictionbased processing (Clark, 2013), where predictions about expected sensory stimuli are continuously formed and evaluated against incoming sensory input to inform further processing. Such a predictive processing does not only inform our perception of actions of others, but also our action production in which we constantly predict the sensory consequences of our own actions and correct them in case of deviations.

Both of these processes draw on the human motor system constituting a perception-action coupling (Prinz, 1997). However, in dynamic social interaction, perception and production often need to be at work simultaneously and for both, actions of self and other. How does the sensorimotor system distinguish between self and other? And how does it interplay with higher-level cognitive processes like mentalizing to solve this social differentiation problem?

As of yet, it is not clear how exactly self-other distinction is implemented within the motor system, but there is evidence for a differentiated involvement supporting the motor system's key role in social cognition (Schütz-Bosbach, Mancini, Aglioti, \& Haggard, 2006). We aim to contribute a computational modeling perspective. In previous work we de-
vised a model of the interplay of mentalizing and predictionbased mirroring during social interaction. It demonstrated how mentalizing - even with minimal abilities to account for beliefs, desires and intentions - affords interactive grounding and makes communication more robust and efficient (Kahl \& Kopp, 2015). In that work two virtual agents interacted in a communication game, each of which equipped with models of a mirroring system and mentalizing system, respectively.

In this paper we present an extension of the predictionbased model of the sensorimotor system to enable it to differentiate actions of its own from the interaction partner's actions. We start with briefly introducing the hierarchical, prediction-based model of a sensorimotor system. Then we discuss how this model can be extended to deal with concurrent perception and production in social situations. This includes a basic ability to integrate predictive and postdictive cues to form a sense of agency (SoA) that helps to differentiate between self and other. Finally, we present and discuss results from simulation studies of different simple scenarios, which test the model's ability to infer SoA for its own actions.

## Computational model of a sensorimotor system

Like other attempts to model the motor system, we chose to make use of a hierarchical representation of increasing abstractions over motor commands (Wolpert, Doya, \& Kawato, 2003; Sadeghipour \& Kopp, 2010). In a three-level hierarchy (see Figure 1), we represent motor primitives on the lowest level (MPrim), followed by a motor sequence layer (MSeq), and motor schemas on the topmost level of abstraction (MSchema). Motor primitives represent single movement segments in space, motor sequences store lists of motor primitives, while motor schemas represent abstract clusters of motor sequences grouped by similarity. We assume that these representations are the basis for a prediction-based model of sensorimotor processing which underlies both action perception and production. To this end, we assume the representations to be multimodal, i.e., combining visual, motor and proprioceptive aspects of action, if available. Consequently, they are used as more or less high-level or visuomotor representations of action and their outcomes. During action perception, we further assume that the correspondence problem is solved in the sense that an observed action by another agent is mapped into one's own self-centered frame of reference. That is, we feed the perceived action trajectory directly and bottom-up into the sensorimotor system.


Figure 1: The predictive sensorimotor hierarchy, based on predictive processing and perception-action coupling. Predictions are sent top-down from state nodes $(S)$ and will be compared to sensory evidence in error nodes $(E)$ in order to drive updates within the hierarchy.

## Sequence matching

In the current model motor primitives are matched against sensory input, which is assumed to consist of a sequence of the last two perceived input coordinates. Motor sequences are matched against a temporary motor sequence concurrently collected from the motor primitive layer, yielding a best match and a prediction of the next motor primitive in the sequence. Motor schemas are likewise matched against the currently best matching motor sequence. In the case of the motor primitives, before the best match is searched, the input sequence is linearly interpolated to match the length of the motor primitives and it is scaled and translated to match the motor primitive's position and size in its coordinate system. Sub-sequence matching is solved by applying euclidean distance measures, which provides high accuracy in our domain size. The same matching algorithm is used for comparing motor sequences in the motor schema layer.

## Predictive sensorimotor hierarchy

The model realizes a predictive processing account resting upon assumptions of the predictive brain hypotheses (Clark, 2013). To that end, it stores representations in the form of discrete probability distributions that can be influenced both bottom-up, in the form of evidence for its last prediction from the next lower layer, and top-down in the form of a prediction by the next higher layer. Following the assumption that the main flow of information is top-down and that motor control is also just top-down sensory prediction, described as "actionoriented predictive processing", or "active inference" (Clark, 2013), all layers receive the next higher layer's prediction and evaluate it for their own bottom-up prediction in the next time step.

As shown in Figure 1, in any time step, the top layer is the first to update its discrete probability distribution in the state node ( $S$ ), given its prior distribution $\left(S_{t-1}\right)$ and the likelihood, calculated in the error node $(E)$ based on the evidence from the layer below. The updated state node $\left(S_{t}\right)$ will be used as a prediction for the current time step, influencing the layer below as a prior, and a copy will be stored in the error node for comparison in the next time step:

$$
S(\text { MSchema })_{t}=S(\text { MSchema })_{t-1} E(M S e q)_{t-1}
$$

Next, the state node at the layer of motor sequences will be updated given its prior distribution, the prediction from the motor schema layer and the likelihood, calculated from the evidence in the layer below:

$$
S(M S e q)_{t}=S(\text { MSchema })_{t} S(M S e q)_{t-1} E(\text { MPrim })_{t-1}
$$

The resulting posterior distribution will be sent as prediction to the layer below, and as evidence to the layer above. Finally, the state node at the motor primitive layer will receive an update given its prior from the last time step, the posterior from the motor sequence layer and the likelihood of the received sensory evidence ( $o$ ) given all motor primitives:

$$
S(\text { MPrim })_{t}=S(\text { MSeq })_{t} S(\text { MPrim })_{t-1} E(o)
$$

For a better understanding of the process of how the model matches the input to its hierarchical representation, see Figure 2 . We have recorded handwritten capital letters using a graphical tablet. All sequences of drawing the 26 characters of the alphabet are stored with a sampling rate of 25 frames per second. From this dataset (12 primitives, mapping onto 26 sequences, mapping onto 26 schemas) we can trigger the model to draw a character, and simulate the model perceiving somebody's drawing of a character in real-time (by simply feeding the trajectories into the system as input one coordinate after the other). Figure 2 depicts three steps in the prediction-based recognition process that leads to a high probability of perceiving the drawn character.

## Precision

The sensorimotor hierarchy learns motor sequences and motor schemas online, with each layer having to decide whether to add a new representation or not. One cognitively plausible


Figure 2: Prediction-based recognition process of the character $L$ being drawn: black dots indicate input coordinates, red dotted lines represent matched motor primitives.
way to determine the need to extend upon the available motor representations is to calculate the precision of each layer's prediction against the evidence in the next time step. Friston and Frith (2015) used precision (the inverse of the variance between prior and posterior probability distributions) as a sort of cortical gain control or neuro-modulation, as to control the influence of the feedback that their Songbird models received upon their song production. We use it as a measure of how well our model can predict its environment and update or extend it according to how similar the prior $\left(P=S_{t-1}\right)$ and posterior $\left(Q=S_{t}\right)$ are:

$$
\operatorname{Pr}(P, Q)=\frac{1}{\operatorname{Var}(Q-P)}
$$

This enables each layer to evaluate its predictive power and, by thresholding, to decide whether to extend its repertoire.

## Active inference

Following the assumption that overt action is basically actionoriented predictive processing (Friston, Daunizeau, Kilner, \& Kiebel, 2010), we allow high-precision prediction at the layer of motor primitives to be acted out. This leads to overt and automatic enactment of correctly predicted actions, similar to the automatic imitation seen in patients suffering from echopraxia (Ganos, Ogrzal, Schnitzler, \& Münchau, 2012). To control this motor execution, we introduced a signal into the top-layer of the hierarchy, which acts as a motor intention for a specific motor schema, including a strong boost of this motor schema's probability. This percolates down the hierarchy to boost associated representations and informs about the intention to act. Once it reaches the lowest layer of the hierarchy, and combined with the high precision threshold, the act to produce the motor representation will be allowed.

With the hierarchical model in place, we set out to find mechanisms to distinguish activations that stem from own action from those arising due to the interaction partner's action. One way is to make sure that the perceived action outcomes are correctly attributed. That is, we need to look at creating SoA, i.e., the sense that an action is self-generated.

## Self-other distinction and sense of agency

How does the human brain distinguish between information about ourselves and others? Or to be more specific, how can we distinguish ourselves from others so that we do not falsely attribute an action outcome to ourselves? These questions pertain to the more general mechanisms that give rise to a sense of "feeling of control", agency, and "self". Generally, a person's SoA is believed to be influenced through predictive and postdictive (inferential) processes, which when disturbed can lead to misattributions of actions as in disorders as for example in patients suffering from schizophrenia (van der Weiden, Prikken, \& van Haren, 2015). We aim to identify mechanisms in order to model these processes and their integration into a combined SoA.

The predictive process makes use of people's ability to anticipate the sensory consequences of their own actions. It allows to suppress, i.e., decrease the intensity of incoming signals which enables people to distinguish between self-caused actions and their outcomes and those actions and outcomes caused by others. One account to model these processes is based on inverse and forward models to account for disorders of awareness in the motor system and delusion of control (Frith, Blakemore, \& Wolpert, 2000). This view suggests that patients can no longer link their intentions to their actions, that is they are conscious of their intention, but not of the sensory consequences of the action. As research into schizophrenia has shown, reliable and early self-other integration and distinction is important not only for the correct attribution of SoA, but also in turn for the correct attribution of intentions and emotions in social interactions. This even suggests that the attenuation of perceived sensory outcomes correlates with activation in the mirror neuron system (van der Weiden et al., 2015). Weiss, Herwig, and Schütz-Bosbach (2011) showed that there is a social aspect to predictive processes that influence SoA by comparing perceived loudness of auditory action effects in an interactive action context. They found that attenuation occured also in the interactive context, comparable to the attentuation of self-generated sound in an individual context.

The postdictive process relies more on inferences drawn after the movement in order to infer whether our intentions are contingent and consistent with the observed events (Wegner \& Wheatley, 1999) and is also influenced by higher-level causal beliefs and thoughts. One important aspect of this inferential process relies on the temporal aspects of actionoutcome integration. It was shown that increasing action outcome delay decreases feeling of control (Sidarus, Chambon, \& Haggard, 2013). This is related to the "temporal binding window" (Colonius \& Diederich, 2004), in which the sensory signals related to the outcome of an action are integrated. The width of the window is dependend on the predictability of the action outcome. Since we have more experience in predicting our own body, the temporal binding window is more narrow for own action outcomes than for other people's actions. Being able to make such a distinction allows people to monitor,
infer and distinguish between causal relations for own and other's behavior. Another aspect informing the postdictive process relies on the valence attributed to an action outcome, where positive valence of an action outcome leads to stronger SoA (Yoshie \& Haggard, 2013).

In sum, there are two processes that can inform SoA and hence can help to distinguish actions of self and other in social interaction. A predictive process works on the content of the action, e.g., the motor command and utilizes forward models as a mechanism to predict the to-be-produced motor command. A postdictive process works with higher-level causal beliefs like the intention to act and temporal binding as mechanisms to infer the consistency of the action outcome.

But how do these two processes work together to inform SoA and what if their cues are unreliable? Cue integration was first studied by Moore, Wegner, and Haggard (2009) who found that when predictive cues are absent external cues become more influential. Nahab et al. (2010) found in an imaging study that there is a leading and a lagging network that both influence SoA prior to and after an action. The leading network would check whether a predicted action outcome would be perceived, while the lagging network would make use of these cues to further process SoA leading to its conscious experience. It seems that in order to generate SoA, both systems do not necessarily have to work perfectly together, as there is evidence for a weighted integration of cues for agency based on their reliability (Moore \& Fletcher, 2012). Also, if the reliability of the predictive process was reduced, the system put more weight on the postdictive inferential processes (Wolpe et al., 2014). Furthermore, the fluency of action-selection processes may also inform the selfother distinction, because the success of repeatedly predicting the next actions seem to accumulate over time to inform SoA (Chambon, Sidarus, \& Haggard, 2014).

## Modeling self-other distinction in social interaction

During online social interaction, the sensorimotor system potentially gets involved in simultaneous action perception and production processes. Our goal is to investigate how the prediction-based model can be extended to cope with the social differentiation problem during such dual-use situations. To this end, we integrate three cues from the predictive and postdictive processes into SoA for produced actions: In the predictive process, we have the match or mismatch of the predicted action-outcome. In the postdictive process, we have the intention to act and the delay in the action-outcome for temporal binding. However, it is not obvious how these cues are being integrated. As a first step, we test two simple ways to do so, namely, to connect them conjunctively or disjunctively. A conjunctive connection allows SoA to occur only if it is supported by both processes; in a disjunctive connection only one cue suffices to inform SoA, in a more flexible but potentially more error prone manner.

The predictive mechanism to differentiate self and other
works based on the content of the predictions that are being sent down the hierarchy. As described above, the predictability of actions by itself provides a predictive cue for a feeling of control, or SoA. Thus, the model needs to quantify the success of a prediction about the outcome of an acted-out motor representation. Since we already have a layer's precision as a measure of success of its predictions, we can directly utilize it as a cue to model SoA.

As the postdictive inferential mechanism we model the temporal binding window as a Gaussian with its mean $(\mu)$ at the predicted delay, which is informed by the perceived action duration during learning. The Gaussian's standard deviation $(\sigma)$ is scaled by the model's layer's predictive precision.


Figure 3: A model of self-other distinction based on the prediction of the consequence of an action and the postdictive integration of an intention to act with the perceived consequence of the action during a predicted temporal binding window, scaled by the predicting layer's precision.

In Figure 3 you can see a sketch of how the predictive and postdictive mechanisms work together to infer SoA for the produced action and its consequence that is perceived. The postdictive mechanism for temporal binding will be triggered by the sensorimotor hierarchy's intention to produce an action, in that it will receive a reference to the motor representation to anticipate. Information from this motor representation will then be used to model the temporal binding window. Now, when the anticipated motor representation is perceived the delay until this perception occurred $(x)$ will be used to calculate the likelihood in the temporal binding window's Gaussian,

$$
\operatorname{lh}(\sigma, \mu, x)=\frac{1}{l h_{\max }} \frac{1}{\sigma \sqrt{2 \pi}} e^{-\frac{x-\mu^{2}}{2 \sigma^{2}}}
$$

with $\sigma$ being scaled by precision (Pr), i.e., trust in the model's predictions. The resulting likelihood will be scaled to $l h_{\max }$, the Gaussian's peak. This postdictive cue has to be combined
with the predictive cue of the general precision of the predictions. We do this by assuming a winner takes all estimate for the predictive and postdictive cues, with a threshold at $50 \%$ probability. Postdictive and predictive cues will then be connected conjunctively or disjunctiveley to reach an estimate for a combined SoA. Since fluency is important for SoA (Chambon et al., 2014), we will integrate this estimate over time. The agency estimate will add to the overall SoA through a simplified Kalman filter,

$$
\begin{gathered}
\text { agency }_{t}\left(\text { agency }_{t-1}, \text { agency }_{\text {estimate }}\right)= \\
\text { agency }_{t-1}+\operatorname{Pr} *\left(\text { agency }_{\text {estimate }}-\text { agency }_{t-1}\right)
\end{gathered}
$$

By allowing the agency estimate (agency estimate ) to influence the overall SoA $\left(\right.$ agency $\left._{t}\right)$ only through this filter, strong fluctuations are dampened and with a gain controlled by precision ( $P r$ ) the influence of the estimate will strongly depend on the success of previous predictions. This means that an agency estimate will either have a strong influence if precision is high, or SoA can only accumulate slowly if precision is low.

This is our integrated model of predictive and postdictive mechanisms which will enable the hierarchical sensorimotor system to differentiate between actions initiated from self and others.

## Simulations and Results

To test the combined model's ability to solve the problem of the dual use of the sensorimotor system and differentiate between self and other we simulate a limited social situation. In this situation, the model will write a character while it either receives the correct action-outcome as feedback, or it receives delayed or different feedback than expected. This way we simulate the effect of concurrent perception of an interaction partner's action.

We test three scenarios for each combination of predictive and postdictive cues to form SoA. In the first scenario we trigger the intent to act out a motor schema and the model will perceive its own correct output as feedback. Here, the model will receive both cues correctly. In the second scenario we trigger the same intent to act, but now the model receives feedback with a delay, disrupting the postdictive cue. In the third scenario the model will again be triggered to act, while this time it receives unpredicted action-outcomes. Here, the model can receive correct cues only accidentally. The model will be triggered to produce and perceive the letter $L$ in scenario one and two. In scenario three, the model will perceive the letter $M$ being produced instead. We log the calculated SoA over time for each scenario.

The resulting plots in Figure 4 show SoA in the different scenarios. The upper row shows the resulting plots for the conjunctive and the lower row for the disjunctive cue integration. In scenario one, the integration of cues jumps strongly because predictive precision is high and small irregularities in timing have a strong effect. The conjunctive connection of cues does not allow for SoA in scenarios two and three, because both cues are not available simultaneously. The dis-


Figure 4: Resulting sense of agency over time for each scenario and each connection between predictive and postdictive cues.
junctive connection between cues fares better, since it allows for SoA even when only one of the cues is available.

In dynamic scenarios of concurrent perception and production, with feedback either from own or from other's actions, a more flexible distinction may be supported by results reported in the literature. A disjunctive integration is sound with regard to results where the reliability of the predictive process was reduced and the system put more weight on postdictive processes (Wolpe et al., 2014). Also, Moore and Fletcher (2012) found evidence for a weighted integration based on the cues' reliability. Another aspect which we found to also influence our results, was the fluent correct prediction of actions (Chambon et al., 2014). Especially in the conjunctive scenario one and the disjunctive scenario two, the accumulation of agency estimates over time was disrupted either through prediction-error or temporal binding errors.

Taken together, the results show that the model can correctly attribute its own action outcomes to itself, which contributes to distinguishing itself from an interaction partner. This differentiating role of the motor system and its strong involvement in social cognition was also proposed by (SchützBosbach et al., 2006). Our cognitive model and its simulation results support this suggestion.

## Conclusion

We have presented a predictive hierarchical model of the sensorimotor system, integrated with a model of self-other distinction that can solve the dual-use problem in dynamic social situations. Furthermore, we presented simulation results of different scenarios of simultaneous perception and production. We compared them to the literature on SoA and the influence of the motor system on social cognition. This comparison suggests that our model can correctly attribute SoA for its own actions, using a more flexible (disjunctive) integration of predictive and postdictive cues.

Taken together, our modeling approach supports the motor system's role in social cognition. Still, the literature on the social brain suggests that motor cognition, as well as the distinction of self and other are influenced by higher level
processes, causal beliefs, and by the mentalizing network.
We agree that the interplay between mentalizing and mirroring needs to be incorporated to meet the demands of truly social systems in interaction scenarios with multiple agents. In earlier work, we already combined our previous model of the sensorimotor system with a mentalizing model in a social scenario with two virtual agents (Kahl \& Kopp, 2015).

In future work, we want to improve our setup by making use of the differentiating information provided by the present model to inform higher-level cognition through an interplay with the mentalizing system. We conjecture this interplay can yield the distinction between one's own and an interaction partner's beliefs needed in social interaction, where informed reciprocity is the key to efficient and successful communication.

## Acknowledgments

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

## References

Chambon, V., Sidarus, N., \& Haggard, P. (2014, May). From action intentions to action effects: how does the sense of agency come about? Frontiers in Human Neuroscience, 8 , 439-9.
Clark, A. (2013, June). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181-204.
Colonius, H., \& Diederich, A. (2004, July). Multisensory Interaction in Saccadic Reaction Time: A Time-Window-of-Integration Model. Journal of Cognitive Neuroscience, 16(6), 1000-1009.
Friston, K., Daunizeau, J., Kilner, J., \& Kiebel, S. J. (2010, February). Action and behavior: a free-energy formulation. Biological Cybernetics, 102(3), 227-260.
Friston, K., \& Frith, C. (2015, November). A Duet for one. CONSCIOUSNESS AND COGNITION, 36, 390-405.
Frith, C. D., Blakemore, \& Wolpert, D. M. (2000, December). Abnormalities in the awareness and control of action. Philosophical Transactions of the Royal Society B: Biological Sciences, 355(1404), 1771-1788.
Ganos, C., Ogrzal, T., Schnitzler, A., \& Münchau, A. (2012, July). The pathophysiology of echopraxia/echolalia: Relevance to Gilles De La Tourette syndrome. Movement Disorders, 27(10), 1222-1229.
Kahl, S., \& Kopp, S. (2015). Towards a Model of the Interplay of Mentalizing and Mirroring in Embodied Communication. In G. Airenti, B. G. Bara, \& G. Sandini (Eds.), Euroasianpacific joint conference on cognitive science (pp. 300-305).
Moore, J. W., \& Fletcher, P. C. (2012, March). Sense of agency in health and disease: A review of cue integration approaches. CONSCIOUSNESS AND COGNITION, 21(1), 59-68.

Moore, J. W., Wegner, D. M., \& Haggard, P. (2009, December). Modulating the sense of agency with external cues. CONSCIOUSNESS AND COGNITION, 18(4), 10561064.

Nahab, F. B., Kundu, P., Gallea, C., Kakareka, J., Pursley, R., Pohida, T., ... Hallett, M. (2010, December). The Neural Processes Underlying Self-Agency. Cerebral Cortex, 21(1), 48-55.
Prinz, W. (1997). Perception and action planning. European journal of cognitive psychology.
Przyrembel, M., Smallwood, J., Pauen, M., \& Singer, T. (2012). Illuminating the dark matter of social neuroscience: Considering the problem of social interaction from philosophical, psychological, and neuroscientific perspectives. Frontiers in Human Neuroscience, 6(June), 190-15.
Sadeghipour, A., \& Kopp, S. (2010, November). Embodied Gesture Processing: Motor-Based Integration of Perception and Action in Social Artificial Agents. Cognitive Computation, 3(3), 419-435.
Schütz-Bosbach, S., Mancini, B., Aglioti, S. M., \& Haggard, P. (2006, September). Self and Other in the Human Motor System. Current Biology, 16(18), 1830-1834.
Sidarus, N., Chambon, V., \& Haggard, P. (2013, December). Priming of actions increases sense of control over unexpected outcomes. CONSCIOUSNESS AND COGNITION, 22(4), 1403-1411.
van der Weiden, A., Prikken, M., \& van Haren, N. E. M. (2015, October). Self-other integration and distinction in schizophrenia: A theoretical analysis and a review of the evidence. Neuroscience and Biobehavioral Reviews, 57, 220-237.
Wegner, D. M., \& Wheatley, T. (1999, July). Apparent mental causation. Sources of the experience of will. American Psychologist, 54(7), 480-492.
Weiss, C., Herwig, A., \& Schütz-Bosbach, S. (2011, July). The Self in Social Interactions: Sensory Attenuation of Auditory Action Effects Is Stronger in Interactions with Others. PLoS ONE, 6(7), e22723-3.
Wolpe, N., Moore, J. W., Rae, C. L., Rittman, T., Altena, E., Haggard, P., \& Rowe, J. B. (2014, January). The medial frontal-prefrontal network for altered awareness and control of action in corticobasal syndrome. Brain, 137(1), 208-220.
Wolpert, D. M., Doya, K., \& Kawato, M. (2003, March). A unifying computational framework for motor control and social interaction. Philosophical Transactions of the Royal Society B: Biological Sciences, 358(1431), 593-602.
Yoshie, M., \& Haggard, P. (2013, October). Negative Emotional Outcomes Attenuate Sense of Agency over Voluntary Actions. CURBIO, 23(20), 2028-2032.

# A transfer advantage of learning diagrammatic representations of mathematics 

Jennifer A. Kaminski (jennifer.kaminski@wright.edu)<br>Department of Mathematics and Statistics, Wright State University 3640 Colonel Glenn Hwy, Dayton, OH 45435 USA


#### Abstract

This study examined learning and transfer of a simple mathematical concept when learning a symbolic sentential format versus learning a diagrammatic format. Undergraduate college students learned an instantiation of a cyclic group and were then given a test of a novel isomorphic group of the same order followed by a test of a novel non-isomorphic group of a higher order. The results were that both the sentential and the diagrammatic formats led to successful learning and transfer to the novel isomorphic group. However, only learning from the diagrammatic representation produced successful transfer to the non-isomorphic group. These findings suggest that learning a diagrammatic representation of a mathematical concept can have transfer advantages over learning strictly sentential formats.


Keywords: Learning; Transfer; Mathematics; Diagrams.

## Introduction

Mathematical concepts are often difficult for students to acquire. Part of this difficulty may be related to the fact that mathematics is generally expressed with abstract symbols, such as variables. Mathematical symbols can be challenging for students to interpret and use, leading to misconceptions and obstacles to learning. For example, many algebra students believe that if $x$ is an integer, then $y$ is the next larger integer (Wagner, 1981, 1983). Another common misconception is that equivalent equations with different variables, such as $7 \times w+22=109$ and $7 \times n+22=109$, have different solutions (Wagner, 1981, 1983).

Other evidence for the difficulty of using symbols comes from comparing performance on purely symbolic tasks to analogous contextualized tasks and finding an advantage for reasoning and problem solving in the contextualized formats (e.g. Saxe, 1988; Koedinger \& Nathan, 2004; Koedinger, Alibali, \& Nathan, 2008). For example, students are frequently more successful solving simple algebra problems when presented as story problems than when presented as symbolic expressions (Koedinger \& Nathan, 2004; Koedinger, et al, 2008). The advantage of contextualized situations may be that when contexts are familiar to students, they can derivate mathematical structure from the context itself (Bassok, 1996, 2003). For instance, given a situation involving 12 tulips and 3 vases, students tend to divide 12 by 3 instead of performing another arithmetic operation because a group of flowers is typically divided between a number of vases. Familiar contextualization may also facilitate learning of new concepts (e.g. Kaminski, Sloutsky, \& Heckler, 2013).

Although contextualized representations of mathematics may sometimes facilitate reasoning, problem solving, and initial learning, such representations can hinder transfer of mathematical knowledge to novel situations (Kaminski, Sloutsky, \& Heckler, 2008, 2013). When college students learned an algebraic system through a familiar context that facilitated initial learning, they were unable to transfer knowledge to a novel analogous domain. However, students who learned the same concept through a generic symbolic format successfully transferred knowledge. Transfer failure may be due to the fact that contextualized real-world instantiations of mathematics communicate more nonessential information than simple symbolic instantiations (Kaminski et al, 2013). This nonessential information is often salient and may divert attention from the less salient mathematical structure, making it difficult to recognize the mathematical structure in novel, superficially dissimilar situations (Kaminski, et al, 2008, 2011, 2013).

However, an important question remains. Does all extraneous information hinder transfer? Perhaps some representations of mathematics have extraneous information that can facilitate transfer. One possible type of representation is a visual display that helps communicate the relevant global relational structure. Such displays are diagrams, and include graphs, matrices, tables, as well as some nonstandard representations. These visual displays instantiate a system with minimal extraneous information, but contain perceptual information that helps spatially organize the elements of the system according to the relevant relational structure. Previous research has demonstrated that diagrams may have advantages over sentential representations for reasoning and problem solving (e.g. Cheng, 2002; Lakin \& Simon, 1987), analogical transfer (e.g. Gick \& Holyoak, 1983; Pedone, Hummel, \& Holyoak, 2001), and non-isomorphic transfer (Novick \& Hmelo, 1994).

The advantage of an effective diagram over a sentential representation may be increased salience of the relations between elements and an ability to accommodate a different number of relevant elements. Consider the example of probability. A sentential format would list relevant probabilities of events $A, B, C$, as $P(A), P(B), P(A \mid C)$, etc. A tree diagram could visually highlight the relationships between events and could be modified to include additional events. By doing so, diagrams may help to communicate higher-order structure, which may allow the learner to transfer knowledge not only to isomorphic situations (i.e. structurally analogous situations) but also to non-isomorphic situations of the same structural class.

For pedagogical reasons, it is important to examine conditions that promote both isomorphic and nonisomorphic transfer because application of mathematical knowledge involves both isomorphic transfer (e.g. transfer of solution strategies to analogous story problems) as well as non-isomorphic transfer (e.g. solution techniques for systems of two variables applied to systems of more than two variables). From a theoretical perspective, it is important to understand how both types of transfer processes are related. Many theories of analogical transfer posit that successful transfer requires alignment of structure across a familiar domain and an isomorphic target domain; this alignment places analogous elements in a one-to-one correspondence across domains (e.g. Gentner, 1983, 1988; Gentner \& Holyoak, 1997; Holyoak \& Thagard, 1989, 1997). From multiple instances, learners may form abstract schematic representations that reflect commonalities of the instances (Gick \& Holyoak, 1983; Novick \& Holyoak, 1991; Reed, 1993). These theories can account for transfer of mathematical knowledge across isomorphs. However, it is unclear how they can account for transfer across nonisomorphic domains in which structure cannot be aligned across instances. If learning of diagrams allows transfer to non-isomorphic situations, do learners align structure when transferring to an isomorphic situation?

The goal of the present research was to examine learning and transfer of a novel mathematical concept from a strictly sentential symbolic representation versus a diagrammatic representation. This study examined both isomorphic transfer (transfer to another system with the same relevant structure and the same number of elements) and nonisomorphic transfer (transfer to another system with the same relevant structure but a different number of elements). When learning the strictly sentential representation, participants may acquire only knowledge of isolated relations between elements and may not gain insight into the higher-order mathematical structure. When learning the diagrammatic representation, participants may learn more than isolated relations between individual elements. They may acquire a structural representation of the concept that can be modified to include more elements than initially learned. As such, the diagram may help communicate higher-order structure, and learning this diagram may facilitate recognition of this structure in novel isomorphic domains as well as non-isomorphic domains of the same type of structure. Therefore, it is hypothesized that both the sentential and diagrammatic representations will result in successful learning and isomorphic transfer, but only the diagrammatic representation will result in successful nonisomorphic transfer.

The concept under consideration was that of a cyclic group (defined in the Method section). Participants learned an instantiation of a cyclic group of order 3 (i.e. three unique elements) with or without the inclusion of a diagram. Participants were then tested on a novel cyclic group of order 3 to examine isomorphic transfer. They were also asked to match analogous elements across domains to
investigate whether there are differences in structural alignment when learning a strictly sentential versus a diagrammatic representation. Afterward participants were tested on a novel cyclic group of order 4 (i.e. four unique elements) to examine non-isomorphic transfer.

## Experiment

## Method

Participants Fifty-eight undergraduate students from a large Midwestern university participated in the experiment and received partial credit for an introductory psychology course.

Materials and Design The experiment included three phases: (1) training and testing in a learning domain, (2) testing in an isomorphic transfer domain, and (3) testing in a transfer domain of the same structure as the learning domain but higher order. Participants were randomly assigned to one of three conditions (Diagram, No Diagram, or Baseline). Participants in the Diagram and No Diagram conditions learned different instantiations of a cyclic group of order 3 during the first phase of the experiment. Participants in the Baseline condition proceeded directly to phase 3, omitting phases 1 and 2. The purpose for the Baseline condition was to measure spontaneous performance in the non-isomorphic transfer domain, without prior instruction on the concept. The isomorphic transfer domain (phase 2) was used in several previous studies (Kaminski, et al, 2008, 2013); without first learning an isomorphic domain, participants were unable to score above chance on this transfer domain.

The learning domain and two transfer domains were artificially constructed instantiations of the concept of a cyclic group. The learning domains and the first transfer domain were of order 3 (i.e. had three unique elements), and the second transfer domain was of order 4 (i.e. had four unique elements). A Cyclic Group of Order $n$ is a set of $n$ elements, or equivalence classes, and an associated binary operation over which the following algebraic properties hold: associativity, commutativity, existence of identity, and existence of inverses. This means that if the operation is denoted by "+", then the following are true. The Associative Property states that for any elements, $x, y, z$, of the set, $(x+y)+z=x+(y+z)$. The Commutative Property states that for any elements $x, y$ of the set, $x+y=$ $y+x$. Also, there is an element, $I$, in the set called the Identity Element, such that for any element, $x, x+I=x$. Finally, for any element, $x$, there exists an Inverse Element, $y$, such that $x+y=I$. In addition, a cyclic group is a group that can be generated by a single element. This concept is equivalent to addition modulo $n$.

The concept of a cyclic group can be instantiated in an unlimited number of ways. The instantiations used for both the Diagram and No Diagram conditions involved three arbitrary symbols, $\boldsymbol{\square}$, and 1. Participants learned the principles of a cyclic group instantiated as associations
between the symbols. The difference between the conditions was the presence or absence of a diagram, the procedure for using the diagram, and the associated cover stories.

In the No Diagram condition, the instantiation was described to participants as rules of a symbolic language in which combinations of two or more symbols yield a predictable resulting symbol. Statements were expressed as symbol 1 symbol $2 \rightarrow$ resulting symbol. Table 1 shows the symbols, the specific rules, and examples. In the Diagram condition, the cyclic group was described to participants as rules for a code-breaking device that can be used to decode sequences of symbols. The decoding device appeared as a circle with three equally spaced positions marked. One symbol was placed at each position. Given a sequence of two or more symbols, the decoder could be used to determine a resulting symbol by starting at the first symbol and moving clockwise around the dial shown in Figure 1. Figure 1 also presents the procedure for using the decoder, the specific rules, and an example.

In both conditions, participants were taught the same associations between sequences of symbols and saw the same sentential statements, symbol 1 symbol $2 \rightarrow$ resulting symbol. The rules, examples, and test questions were identical in both conditions. Aside from different cover stories, the only difference between the conditions was the inclusion of the diagrammatic representation (i.e. the decoding device) and its associated procedure in the Diagram condition. At the end of phase 1, participants were tested with a 24 -question multiple-choice test.

The second phase of the experiment was testing of an isomorphic transfer domain. This transfer domain was identical for both conditions and was also a cyclic group of order 3 involving three images of perceptually rich objects. It was described as a children's game where children sequentially point to objects and "the winner" points to a final object (see Transfer Domain 1 in Table 1). Participants were told that the correct final object is specified by the rules of the game (which were the rules of a cyclic group). Furthermore, in both conditions they were told that the rules were like those of the system they just learned. No explicit training in the transfer domain was given; instead, participants were shown a series of examples from which the rules could be deduced (see operands and results for Transfer Domain 1 in Table 1). Participants were asked to figure out the rules of the game by using their knowledge of the learned system. Then they were tested with a 24 -question multiple-choice test, isomorphic to the test in the learning phase of the Diagram and No Diagram conditions, but using the elements of the transfer domain.

Following the test, participants were asked to match analogous elements across the learning and transfer domains. Correct matching of elements was taken as an indicator of correct structural alignment between the learning and transfer domains. For cyclic groups of order 3, there are two possible correct mappings between groups. The identity element is unique; therefore a correct mapping must align these two elements across domains. However,
the mapping between remaining two elements is not unique. Therefore, a response was considered correct if (a) the mapping was one-to-one and onto (i.e., each learning element corresponded to a single transfer element and each element of the transfer elements were used) and (b) the mapping preserved the identity element. In other words, if a participant used each of the group elements and mapped the identity element correctly, then the response was correct. Because the critical aspect was correctly choosing the identity element and most participants were expected to form mappings that were onto, $33 \%$ accuracy was used as a conservative measure of chance for a group of participants.

The third phase of the experiment was testing of a nonisomorphic domain of the same structure as the learning domain (i.e. a cyclic group of order 4). Participants were given a paper and pencil ten-question multiple-choice test (see Table 2) and told that the knowledge of the system they learned first can help them figure out the new system. They were also given five example statements (the operands and results shown in Table 2) from which the complete set of rules could be deduced.

Table 1: Stimuli for the learning and isomorphic transfer domains.



Figure 1: Diagram used in the Diagram Condition

Table 2: Stimuli for non-isomorphic transfer domain.

|  | Transfer Domain 2 <br> Cyclic group of order 4 |
| :---: | :---: |
| Elements | $\bigcirc$ ¢ $\square$ |
| Identity | $\bigcirc$ |
| Associations between elements | (Presented as examples) |
|  | Operands Result |
|  | - , $\star$ * |
|  | $\bigcirc$, $\odot \bigcirc$ |
|  | $\star$, $\star$ - |
|  | $\bigcirc, \bigcirc$ |
|  | $\star$, $\bigcirc \square \square$ |
| Example Test Question | Find the resulting symbol $\bullet, \star, \odot, \square \rightarrow \text { ? }$ <br> Answer: |

Procedure Participants were seen individually in a lab on campus. Phases 1 and 2 were presented on a computer. Participants proceeded at their own pace, with their responses recorded by the computer. The learning phase consisted of approximately 80 slides and required approximately 15 minutes to complete. The transfer phase (phase 2) consisted of 48 slides and took on average 10 minutes to complete. The second transfer phase (phase 3) took participants approximately 8 minutes to complete. All material in phase 3 was presented on paper.

## Results

Four participants (two Diagram and two No Diagram) were removed from the analysis because they failed to learn

Table 3. Mean accuracy (percent correct) on learning and isomorphic transfer. Note: Standard deviations are presented in parentheses. Chance performance is $37.5 \%$

|  | Learning | Transfer |
| :---: | :---: | :---: |
| No Diagram | $78.9(14.5)$ | $78.5(17.8)$ |
| Diagram | $85.2(8.95)$ | $74.5(22.1)$ |

the concept in phase 1 ; their learning scores were less than 11 and no different than the chance score of 9 (i.e. $37.5 \%$ ).

Note that for the following analyses, learning scores and isomorphic transfer scores were not normally distributed in the Diagram condition; the distributions were negatively skewed ( $S W s<.89, d f \mathrm{~s}=18, p \mathrm{~s}<.04$ ). In addition, nonisomorphic transfer test scores had bimodal distributions in both conditions. Therefore, non-parametric analyses were done to examine each of these scores.

Participants in both conditions successfully learned the concept (see Table 3). Learning scores were above chance in both conditions, Wilcoxon Signed Ranks Test, $Z \mathrm{~s}>3.72$, $p \mathrm{~s}<.001$. There were no significant differences in learning levels between conditions, Mann-Whitney $U$ test, $U=123.5$, $p=.22$.

In phase 2 (i.e. testing of an isomorphic transfer domain), participants also performed well in both conditions (see Table 3). Scores were above a chance score of $37.5 \%$, in both conditions, Wilcoxon Signed Ranks Test, $Z \mathrm{~s}>3.63$, $p \mathrm{~s}$ $<.001$. Note that previous research demonstrated that without initially learning an isomorphic domain, participants were unlikely to score above chance on this transfer task (Kaminski, et al, 2008, 2013). Therefore, it appears that participants in both conditions successfully transferred structural knowledge acquired in phase 1 to answer questions about the isomorphic domain in phase 2. No significant difference in transfer scores between the two conditions was found, Mann-Whitney $U$ test, $U=151.0, p=$ .74 .

In addition, most participants in both conditions accurately matched analogous elements across the learning and isomorphic transfer domains ( $83 \%$ in the Diagram condition and $78 \%$ in the No Diagram condition), suggesting that they successfully aligned analogous structure across the two domains. The percent of participants in both conditions was well above chance of $33 \%$ and not different between conditions, Fisher Exact test, $p \mathrm{~s}=1.00$.

While there were no differences in performance levels between conditions for phases 1 and 2, there were significant differences in performance in phase 3 (i.e. testing on a non-isomorphic domain of similar structure). Scores in the Diagram and No Diagram conditions had bimodal distributions; Table 4 presents the frequency of scores at different levels. Transfer scores were higher in the Diagram condition ( $M=67.8, S D=9.69$ ) than in the No Diagram condition ( $M=41.1, S D=7.79$ ), Mann-Whitney $U$ test, $U=$

Table 4. Percent of participants in each condition scoring at different levels on the non-isomorphic transfer task.

|  | Accuracy Level |  |  |
| :---: | :---: | :---: | :---: |
|  | $\underline{\text { Low }}$ | $\underline{\text { Middle }}$ | High <br> $(0-40 \%)$ |
| $(50-70 \%)$ | $(80-100 \%)$ |  |  |
| Baseline | 61 | 33 | 6 |
| No Diagram | 61 | 17 | 22 |
| Diagram | 39 | 0 | 61 |

$97.5, p<.04$. Furthermore, scores in the Diagram condition were higher than scores in the Baseline condition ( $M=$ $41.1 \%, S D=19.1 \%$ ), Mann-Whitney $U$ test, $U=107.0, p<$ .04 , one-tailed. However, scores in the No Diagram condition were not significantly different than those in the Baseline, $U$ $=144.5, p=.58$. This finding suggests that the majority of participants in the Diagram condition were able to transfer knowledge of the cyclic group order 3 to the non-isomorphic cyclic group of order 4, but the majority of participants in the No Diagram condition were not able to do so.

## Discussion

The goal of the present study was to investigate transfer of mathematical knowledge when learning a strictly sentential representation versus learning a diagrammatic representation. This study considered transfer to a novel isomorphic domain as well as transfer to a novel nonisomorphic domain of the same structural class. Both formats resulted in equally successful learning and isomorphic transfer. However, participants who learned the diagram were more successful at non-isomorphic transfer than those who learned only the sentential format. These findings suggest that although the diagram added nonessential information, this information did not hinder learning or isomorphic transfer. Moreover, the inclusion of this information resulted in a clear advantage for nonisomorphic transfer.
Previous research has demonstrated that learning instantiations that include extraneous information hinder transfer of mathematical knowledge to novel isomorphs because the extraneous information is generally salient and likely diverts attention from the relevant structure (Kaminski et al, 2008, 2011, 2013). Compared to strictly sentential representations, diagrams also communicate nonessential information to the learner. For example, in the present study, it is not necessary to include the diagram; the same rules and associations were learned equally well in the No Diagram condition. Clearly standard diagrams such as tree diagrams, matrices, and graphs also communicate nonessential information. However while the information added by a diagram is nonessential, it is not necessarily irrelevant. Effective diagrams use visual information to spatially organize elements of a system in a way that
highlights relations and relevant structure and does not divert attention from the structure. Such diagrams may help to communicate global structure of the system in manner that can be modified if necessary to incorporate a different number of elements.

In the current study, the diagram was circular and likely helped to communicate the cyclic nature of the group and the fact that any element can be obtained as a result of operations involving the other elements. It is more difficult to recognize the cyclic nature of the relationship between elements in the strictly sentential format. Even if learners had constructed a schematic representation of the concept from the sentential format without the diagram, the schema appears to reflect only local associations between three elements and not a more global structure that can be modified and applied to non-isomorphic situations.

With regard to structural alignment, participants in both conditions were equally accurate at matching analogous elements, possibly suggesting structural alignment. However it is not clear that this element-level matching is necessary when learning the diagram or whether global structure can be mapped from learned to target domains without one-to-one correspondence of elements.

Successful non-isomorphic transfer from learning the diagram suggests that participants have formed a more sophisticated internal representation of a structural class of mathematical entities, in this case cyclic groups of different orders. Recognizing that different mathematical entities can fall into the same structural categories is an important part of advancing mathematical knowledge. For example, algebra students should be able to modify techniques for solving systems of equations of two variables to solve systems of equations of more than two variables. Similarly, college students should recognize that slope of a line is an instance of derivative of a function. An effective diagram, if available, may help illuminate structure in a way that allows modification of the number of elements. Standard mathematical diagrams such as matrices, graphs, networks, and Venn Diagrams do precisely this.

At the same time, there may be limitations to the benefit of diagrams. The inclusion of diagrams may not always facilitate initial learning. Correct interpretation and use of diagrams requires additional learning beyond learning standard sentential representations. For some combinations of concepts, diagrams, and learners, such as those considered in this study, a diagram is easily learned. However, this is not always the case. For example, in middle school students, diagrams provided a benefit for solving algebraic word problems only for older students and high-achieving students, but not for younger students and lower-achieving students (Booth \& Koedinger, 2012). Some concepts may be simple enough to be learned without such additional representations. For more difficult concepts, some learners may be unable to fully learn the diagram and the relationship between the diagram and standard sentential formats such as equations.

It is also important to note that while a diagram is a visual representation of the elements and relations of a system, it is meaningless without knowledge of how to interpret it. The diagram used in the present study involved the visual representation in Figure 1 along with the procedure of how to use it. The same is true for common mathematical diagrams such as multiplication tables and Cartesian graphs. These well known diagrams easily communicate information to us only because we have been explicitly taught procedures for constructing and interpreting them.

Learning diagrams in addition to standard sentential mathematics may require additional effort. For some learners and some diagrams, this may be challenging. However, the benefit of well-designed diagrams is likely worth the effort. Once learned, diagrams likely can provide advantages for transfer to isomorphic situations and many non-isomorphic situations.

## Acknowledgments

This research was supported by a grant from the Institute of Education Sciences, U.S. Department of Education (\#R305A140214).

## References

Bassok, M. (1996). Using content to interpret structure: Effects on analogical transfer. Current Directions in Psychological Science, 5, 54-58.
Bassok, M. (2003). Analogical transfer in problem solving. In J. E. Davidson \& R. J. Sternberg (Eds.) The psychology of problem solving (pp. 343-369). New York: Cambridge University Press.
Booth, J. L, \& Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. British Journal of Educational Psychology, 82, 492-511
Cheng, P. C. H. (2002). Electrifying diagrams for learning: principles for complex representational systems, Cognitive Science, 26, 685-736.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170. doi:10.1207/s15516709cog0702_3
Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child Development, 59, 47-59. doi:10.2307/1130388
Gentner, D., \& Holyoak, K. J. (1997). Reasoning and learning by analogy. American Psychologist, 52, 32-34. doi:10.1037/0003-066X.52.1.32
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38. doi:10.1016/0010-0285(83)90002-6
Herscovics, N., \& Kieran, C. (1980). Constructing meaning for the concept of equation. The Mathematics Teacher, 73(8), 572-580.

Holyoak, K. J., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. American Psychologist, 52, 3544.

Holyoak, K. J., \& Thagard, P. (1997). The analogical mind. American Psychologist, 52, 35-44.
Kaminski, J. A., Sloutsky, V. M., \& Heckler, A. F. (2008). The advantage of abstract examples in learning math. Science, 320, 454-455.
Kaminski, J. A., \& Sloutsky, V. M. (2011). Representation and transfer of abstract mathematical concepts in adolescence and young adulthood. In V. Reyna (Ed.) The Adolescent Brain: Learning, Reasoning, and Decision Making (pp. 67-93). Washington, DC: APA.
Kaminski, J. A., Sloutsky, V. M., \& Heckler, A. F. (2013). The Cost of Concreteness: The Effect of Nonessential Information on Analogical Transfer. Journal of Experimental Psychology: Applied, 19, 14-29.
Koedinger, K. R., Alibali, M. W., \& Nathan, M. M. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. Cognitive Science, 32(2), 366-397. http://dx.doi.org/10.1080/03640210701863933.
Koedinger, K. R., \& Nathan, M. J. (2004). The real story behind story problems: effects of representations on quantitative reasoning. Journal of the Learning Sciences, 13, 129-164.
Larkin, J. H., \& Simon, H. A. (1987). Why a diagram is (sometimes) worth then thousand words. Cognitive Science, 11, 65-100.
Novick, L. R., \& Holyoak, K. J. (1991). Mathematical problem solving by analogy. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 398415.

Novick, L., \& Hmelo, C. E. (1994). Transferring symbolic representations across nonisomorphic problems. Journal of Experimental Psychology: Learning. Memory, and Cognition, 20, 1296-1321.
Pedone, R., Hummel, J. E., \& Holyoak, K. J. (2001). The use of diagrams in analogical problem solving. Memory and Cognition, 29, 214-221.
Reed, S. K. (1993). A schema-based theory of transfer. In D. K. Detterman \& R. J. Sternberg (Eds.) Transfer on Trial: Intelligence, Cognition, and Instruction. Norwood, NJ: Alex Publishing.
Saxe, G. B. (1988). The mathematics of child street vendors. Child Development, 59, 1415-1425.
Wagner, S. (1981). Conservation of equation and function under transformations of variable. Journal for Research in Mathematics Education, 12, 107-118.
Wagner, S. (1983). What Are These Things Called Variables? The Mathematics Teacher, 76(7), 474-479.

# Gricean epistemic reasoning in 4-year-olds 

Alyssa Kampa (akampa@udel.edu)<br>Department of Linguistics, 125 E. Main St, University of Delaware<br>Newark, DE 19716 USA

Anna Papafragou (annap@udel.edu)<br>Department of Psychological and Brain Sciences, 105 The Green, University of Delaware<br>Newark, DE 19716 USA


#### Abstract

Recent experimental evidence suggests that adults incorporate speaker knowledge into the derivation of pragmatic implicatures. Developmental studies report that 5 -year-old children also succeed in taking speaker knowledge into account in implicature computation, but 4 -year-olds fail. The present study investigated the pragmatic competence of 4-year-olds, specifically the ability to incorporate speaker knowledge into the derivation of ad hoc scalar implicatures. Using a simple paradigm inspired by referential communication, we found that 4 - year-olds are able to incorporate speaker knowledge into implicature derivation. These results have implications for our understanding of the linguistic, pragmatic, and epistemic abilities of young children.


Keywords: implicatures; pragmatics; speaker knowledge

## Background

As established by Grice (1989), communication involves partners in a conversation working towards the same cooperative goal (cf. also Sperber \& Wilson, 1986). To that end, speakers must be as informative as required by the purpose of the exchange. If a speaker is less than fully informative, as in (1), the listener will assume that - as far as the speaker knows - the stronger alternative in (2) is not true.
(1) Some chipmunks collect acorns.
(2) All chipmunks collect acorns.

The inference that the stronger statement in (2) does not hold is known as a scalar implicature and requires pragmatic reasoning. Scalar implicatures take their name from the fact that they rely on a comparison to a lexical item on an informativeness scale that the speaker could have used but did not (Grice, 1989). In the case of (1) and (2), the lexical alternatives involved form a scale ordered in terms of logical strength (Horn, 1998). Furthermore, this logically ordered scale is a feature of the language that needs to be accessed in order for the hearer to compute an implicature. Other types of scalar implicatures rely on ad hoc scales that depend upon contextual information. For instance, a speaker can utter (3) to communicate that the stronger statement in (4) does not hold:
(3) Chip and Dale collect acorns.
(4) Chip, Dale and Max collect acorns.

Past findings in the literature have indicated that children struggle with deriving scalar implicatures until late in development: unlike adults, they fail to reject weak (underinformative) statements when a stronger alternative is true (Chierchia et al, 2001; Noveck, 2001). Eye-tracking methods have also revealed weaknesses in early implicature computation (Huang \& Snedeker, 2009). However, 5-yearolds have increased success in computing scalar implicatures when task demands set up an expectation of a stronger utterance (Papafragou \& Musolino, 2003; Skordos \& Papafragou, 2016; Katsos \& Bishop 2011; see Papafragou \& Skordos, 2016 for a review), and even 3-yearolds succeed in deriving ad hoc scalar implicatures in a simple referent selection task (Stiller, Goodman, \& Frank, 2015; cf. also Barner, Brooks, \& Bale, 2011). At present, there is much discussion in the field about whether early failures with implicatures were due to children's increased pragmatic tolerance in judgment tasks (Katsos \& Bishop, 2011), lack of linguistic processing abilities (Chierchia et al, 2001; Reinhart, 2004), inability to access stronger lexical alternatives (Barner et al., 2011), failures in assessing which alternatives are conversationally relevant (Skordos \& Papafragou, 2016), or some combination of these factors.

The present study seeks to incorporate speaker knowledge into the task of implicature derivation (Sauerland, 2004; Fox, 2007; Chierchia, Fox, \& Spector, 2009). There is evidence that adults consult the speaker's knowledge state when computing implicatures (Bergen \& Grodner, 2012; Breheny, Ferguson \& Katsos, 2013). For instance, the hearer upon encountering a statement such as (1) or (3) is justified in concluding that the listener does not know whether the stronger alternative is true (or, in other cases, that the speaker knows that the stronger alternative is false). However, developmental studies have shown that young children have difficulties with such epistemic aspects of implicature computation. Hochstein and colleagues (2016) conducted a study with 4 - and 5 -year-olds investigating their ability to compute non-scalar "ignorance" implicatures which require the incorporation of the speaker's knowledge state into their derivation. They found that 5 -year-olds were able to succeed on this task but 4-year-olds failed.

In a study most closely related to the present experiment, Papafragou, Friedberg and Cohen (in press) found a similar pattern. In that study, 4- and 5-year-old children watched
short videos of two twins. In one video, an observer only saw part of one twin's action, and in the other video, the observer saw the whole action. Children themselves had access to the completed action that was the same in both videos (e.g., a girl colored a star). Children then heard a statement made by one of the observers about the action (e.g. 'The girl colored some/all of the star') and had to decide which observer said it. Five-year-olds were able to successfully incorporate speaker knowledge into their pragmatic reasoning, attributing weak statements to the partly informed observed and strong statements to the fully informed observer, but 4 -year-olds struggled. In later manipulations, when the observers' access to the actions was identical to the children's (and hence there was no need to reason about someone else's belief), 4-year-olds' performance improved.

The present paper revisits the issue of whether 4-year-olds can incorporate the speaker's knowledge state into the computation of implicatures, as 5 -year-olds and adults have been shown to do. The task designed for this experiment was created with the goal of keeping task demands as simple as possible. The design borrows from the referential communication paradigm (see, e.g., Nadig \& Sedivy, 2002). In this paradigm, speaker knowledge is established through the speaker's visual perspective without the need to set up an elaborate background scenario (cf. also Matthews et al., 2006). The task also has a clear goal (referent selection; see Stiller et al., 2014) and targets ad hoc scalar implicatures that rely on contextual knowledge set up within the experimental scene.

## Experiment

## Participants

Thirty-one 4-year-olds (mean age: 4;6, range: 4:0 to 4:11, 16 female) participated. Children were recruited from Newark (DE) preschools. A control group of 26 adult participants was also tested. Adult participants were recruited with a HIT posted on MTurk.

## Method

For the test phase, participants were shown pairs of pictures displayed side by side on a laptop screen (see Figure 1). Within a pair, each picture showed the same person sitting across a table behind a two-compartment box with identical objects (e.g. a spoon and a bowl), facing the camera. In one picture, the girl could see the contents of both compartments in her box (full access box), but in the other, she could only see the content of one compartment (e.g. the spoon) because the other compartment was blocked (limited access box). The participants could see the full contents of both boxes. Within a pair, the first (leftmost) picture was displayed for 2 seconds, followed by the appearance of the second picture that remained on the screen for 2 seconds. Then an audio recording of a sentence was heard (both pictures remained on the screen). The
sentences were either weak (e.g. "I see a spoon") or strong (e.g. "I see a spoon and a bowl").


Figure 1. An example stimulus.
Before the test phase, children were introduced to the limited access box and an explanation of how it worked. The explanation involved showing a picture of the girl with the limited access box in front of her and one object in the hidden compartment. The children were asked whether they could look through the open and closed compartments. For the closed compartment, the children were then asked why they could not look through it (answers typically mentioned that it was closed or blocked). They were asked if they thought the girl could look through the blocked compartment. The children were then asked whether the girl knew there was a hidden item in the blocked compartment. To remind children of the properties of the boxes, after the first 4 test trials, children were again asked whether they could look through both compartments of the limited access box and whether they thought the girl could look through both sides and knew there was an object there (for the importance of such reminders of the visual properties of the display, see Nadig \& Sedivy, 2002). Children who answered "yes" both times to the question of whether the girl knew there was an object in the hidden compartment were excluded because they did not understand the nature of the limited-access box $(\mathrm{N}=6)$.

The participants were given two pre-test trials. These trials involved a two-picture set-up, as in the test phase. Participants were told that they would see some pictures of the girl looking at a box that was open on both sides, and then looking at a box that was open on one side and closed on the other. The participants were told that the girl was going to talk about only one of the boxes. They were instructed that they would hear a sentence and would have to decide which box she was talking about. For the first pretest trial, the items in the two boxes were different across boxes rather than identical as in the test trials (a book and a cup in one box; an orange and a spoon in the other). The sentence unambiguously described the full access box ("I see a book and a cup"). For the second pre-test trial, the boxes had different objects again: the full access box had two objects but the limited access box only had an object in
the closed compartment, but no object in the open compartment. The sentence was: "I see nothing." Neither of the pre-tests involved perspective taking. One child failed both pre-tests and was excluded.
The test phase was identical to the pre-test trials, but with identical objects in the two boxes. After hearing either the weak or strong sentence, participants were asked, "Which box is she talking about?", and had to point to the correct picture. The pictures were counterbalanced in terms of whether the limited access box was on the left or the right within the pair. Participants were given 4 strong and 4 weak sentences in a mixed order. Two presentation lists were created; assignment of Type of Sentence (strong, weak) to pairs of pictures was counterbalanced across lists.

In order to succeed on the task, participants had to inhibit their own perspective, since both boxes had identical contents from the children's point of view. If participants successfully incorporated the perspective and knowledge of the speaker, they should say that the strong statement "I see a spoon and bowl" described the full access box, because the spoon was not visible to the speaker in the limited access box. For weak statements, such as "I see a spoon," participants should pick the limited access box because, although the full access box also had a spoon, it would be underinformative to only mention the spoon if the speaker could also see a bowl.

## Results

Results are presented in Table 1. In accordance with our predictions and adult judgements, correct answers were defined as choosing the full access box for the strong sentences and the limited access box for the weak sentences. For each of the participants, a mean score across the four trials was calculated for both the strong and weak conditions. Because most of the 4 -year-olds received a score of 0 or 1 in the critical weak condition ( 21 out of 26), participants were divided into passers (score $\geq .75$ ) and failers (score $\leq .50$ ).

Adults performed at ceiling in both the strong and weak conditions. For the 4 -year-olds, Fisher's exact test revealed a marginally significant difference in the number of passers and failers for the strong vs. weak condition ( $\mathrm{p}=.05$, 2tailed). Comparisons across age groups revealed a significant difference between adults and 4-year-olds in the weak condition ( $\mathrm{p}=.01$ ), but no significant difference in the strong condition $(p=1)$. Nevertheless, in the critical 4 -yearold weak condition the number of passers was significantly different from the expected ratio due to chance ( $\mathrm{p}=.029$ ).

Table 1: Task performance.

|  | Classification | Condition |  |
| :--- | :--- | :--- | :--- |
|  |  | Strong | Weak |
| Adults | Passers | 26 | 26 |
|  | Failers | 0 | 0 |
| Children | Passers | 25 | 19 |
|  | Failers | 1 | 7 |

## Discussion

This experiment investigated 4-year-olds' ability to incorporate speaker knowledge into the computation of ad hoc scalar implicatures. The results suggest that 4 -year-olds display the ability to incorporate speaker knowledge into implicature derivation. These findings lower prior age estimates of children's ability to take the epistemic step during implicature computation - but align with reports in the literature about the epistemic ability of very young children in non-linguistic tasks (Surian, Caldi \& Sperber 2007; Baillargeon, Scott \& He, 2010). Notice that epistemic stance per se was not as demanding (see strong sentences): when taking someone else's perspective was combined with computing an implicature that this person could have intended, given their knowledge state (weak sentences), performance dropped.

An interesting question is why 4 -year-olds were able to succeed at this task when they failed at prior studies targeting sensitivity to the speaker's epistemic stance in implicature-computation (Hochstein et al., 2016; Papafragou et al., in press). In the present experiment, 4-year-olds needed to compute implicatures, but they also needed to reason about what a person had access to, determine how that would affect the speaker's utterances, and inhibit their own perspective. Nevertheless, our paradigm was based on a simple, clear way of establishing that someone's knowledge differs from the child's own; furthermore, the present paradigm included a clear conversational goal (the identification of the box that the speaker is talking about). In both of these respects, the current study is simpler than past attempts to link the informativeness of a sentence to a speaker's mental state.

These findings and the paradigm used in this experiment provide fertile ground for a continued investigation into the pragmatic ability of young preschool children. It is possible that children younger than 4 could be found to demonstrate these abilities with an even simpler task. We are currently pursuing this possibility in ongoing work.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under grant no. 1632849.

## References

Baillargeon, R., Scott, R.M., \& He, Z. (2010). False-belief understanding in infants. Trends in Cognitive Sciences, 14, 110-118.

Barner, D., Brooks, N., \& Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. Cognition, 188, 87-96.
Bergen, L., \& Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 1450.
Breheny, R., Ferguson, H., \& Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. Cognition, 126, 423-440.
Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., \& Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In A. H.-J. Do et al. (eds.), BUCLD 25 Proceedings, 157168. Somerville, MA: Cascadilla Press.

Chierchia, G., Fox, D., \& Spector, B. (2009). Hurford's constraint and the theory of scalar implicatures. Presuppositions and Implicatures, 60, 47-62.
Fox, D. (2007). Free choice and the theory of scalar implicatures. In Uli Sauerland \& Penka Stateva (eds.), Presupposition and Implicature in Compositional Semantics. Palgrave Macmillan. Houndmills, Basingstoke.
Frank, M. C., \& Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. Cognitive Psychology, 75, 80-96.
Grice, P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.
Hochstein, L., Bale, A., Fox, D., \& Barner, D. (2016). Ignorance and Inference: Do Problems with Gricean Epistemic Reasoning Explain Children's Difficulty with Scalar Implicature? Journal of Semantics, 33, 107-135.
Horn, L. (1989). A natural history of negation. Chicago: University of Chicago Press.
Huang, Y. \& Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. Developmental Psychology, 45, 1723-1739.
Katsos, N., \& Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. Cognition, 120, 67-81.
Matthews, D., Lieven, E., Theakston, A., \& Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. Applied Psycholinguistics, 27, 403-422.
Nadig, A.S., \& Sedivy, J.C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. Psychological Science, 13, 329-336.
Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. Cognition, 78, 165-188.
Noveck, I.A. \& Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In N. Burton-Roberts (ed.) Advances in Pragmatics, 184212. Basingstoke: Palgrave.

Papafragou, A., Friedberg, C., \& Cohen, M. (in press). The role of speaker knowledge in children's pragmatic inferences. Child Development.
Papafragou, A. \& Musolino, J. (2003), Scalar implicatures: experiments at the semantics-pragmatics interface. Cognition, 86, 253-282
Papafragou, A., \& Skordos, D. (2016). Scalar implicature. In J. Lidz, W. Snyder \& J. Pater (eds.), Oxford Handbook of Developmental Linguistics. Oxford: Oxford University Press.
Papafragou, A., \& Tantalou, N. (2004). Children's computation of implicatures. Language Acquisition, 12, 71-82.
Reinhart, T. (2004). The processing cost of reference-set computation: Acquisition of stress shift and focus. Language Acquisition, 12(2), 109-155.
Skordos, D., and Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. Cognition, 153, 6-18.
Sperber, D., \& Wilson, D. (1986). Relevance: Communication and cognition. Cambridge, MA: Harvard University Press. $2^{\text {nd }}$ ed. 1995.
Stiller, A., Goodman, N. D., \& Frank, M. C. (2015). Ad-hoc implicature in preschool children. Language, Learning, and Development, 11, 176-190.
Sauerland, U. (2012). The computation of scalar implicatures: Pragmatic, lexical or grammatical? Language and Linguistics Compass, 6, 36-49.
Sauerland, U. (2004), Scalar implicatures in complex sentences. Linguistics and Philosophy, 27, 367-91.
Surian, L., Caldi, S. \& Sperber, D. (2007). Attribution of beliefs to 13-month-old infants. Psychological Science, 18, 580-586.

# A Hebbian account of entrenchment and (over)-extension in language learning 

Vsevolod Kapatsinski (vkapatsi@uoregon.edu) Zara Harmon (zforough@uoregon.edu)<br>Department of Linguistics, 1290 University of Oregon<br>Eugene, OR 97403 USA


#### Abstract

In production, frequently used words are preferentially extended to new, though related meanings. In comprehension, frequent exposure to a word instead makes the learner confident that all of the word's legitimate uses have been experienced, resulting in an entrenched form-meaning mapping between the word and its experienced meaning(s). This results in a perception-production dissociation, where the forms speakers are most likely to map onto a novel meaning are precisely the forms that they believe can never be used that way. At first glance, this result challenges the idea of bidirectional form-meaning mappings, assumed by all current approaches to linguistic theory. In this paper, we show that bidirectional form-meaning mappings are not in fact challenged by this production-perception dissociation. We show that the production-perception dissociation is expected even if learners of the lexicon acquire simple symmetrical form-meaning associations through simple Hebbian learning.


Keywords: Hebbian learning; word learning; mental lexicon

## Introduction

Extension of frequent forms to novel uses is one of the most common processes in language change, and results in the robust correlation between frequency and polysemy: upon examining a dictionary, it quickly becomes evident that it is the most frequent words that have the largest number of uses (compare get vs. obtain, Piantadosi et al., 2012; Zipf, 1949). Extension of familiar words and constructions to new uses is also one of the major mechanisms driving grammaticalization, a largely unidirectional process through which grammatical morphemes evolve out of lexical sources (Bybee, 2003, 2010). A well-studied example in English is the verb will, which was gradually extended from volitional lexical uses (e.g., I will it to happen) to grammatical future tense uses that no longer imply volition (e.g., I will get fired for suggesting this).

Extension can be observed not only in diachrony but also in online language use. In particular, novel extensions are frequently observed in children's use of both referential terms and verb-argument structure constructions. For example, a child may name a cow a kitty or extend the verb giggle to transitive use, as in don't giggle me (e.g. Naigles \& Gelman, 1995; Pinker 1989). The words that are so overextended tend to be the frequent ones, or else ones that are highly accessible in the moment because they have just been used (Gershkoff-Stowe et al., 2006; for adults, see also Ferreira \& Griffin, 2003; Burke et al., 2004). These patterns parallel the diachronic tendency for frequent words to acquire novel uses.

Crucially, a child can overextend a word in production without overextending it in comprehension. When presented with a word she over-extends in production and asked to pick out all the objects the word can refer to, the child often does not select the objects to which she over-extends the word in production as its possible referents (Naigles \& Gelman, 1995).

In fact, frequency appears to have opposite effects in comprehension and production. Whereas frequent words are extended to new uses in production, frequent words are likely to be restricted to the uses in which they have been experienced. For example, Xu \& Tenenbaum (2007) show that experiencing fep paired with a Dalmatian once leads children to think it plausible that fep refers to all dogs, but three fep-Dalmatian pairings are enough to restrict the set of referents to Dalmatians (see also Ambridge et al., 2008; Brooks \& Tomasello, 1999; Theakston, 2004; Wonnacott et al., 2008, for related results with syntactic constructions). Frequent exposure to a form-meaning pairing appears to convince learners that the form always co-occurs with this meaning.

## The Data

In recent work, we have confirmed the existence of this dissociation in adult learners of a miniature artificial language (Harmon \& Kapatsinski, submitted; Experiment 1). In our study, participants were exposed to a language with two plural suffixes ( $-d a n$ and $-s i l$ ) and two diminutive suffixes (-nem and -shoon). For each participant, one suffix was more frequent than others. Each participant was tested on both comprehension and production.

Participants experienced the language through passive exposure, with nouns bearing the suffixes (e.g. ostodan, zutishoon) presented auditorily, paired with pictures of their referents. Each trial began with a picture of the referent, followed 500 ms later by the spoken word. After the offset of the spoken word, the experiment advanced to the next trial, which began 400 ms later.

Nouns bearing plural suffixes were paired with pictures of multiple large creatures (with the kind of creature determined by the stem), whereas each noun bearing a diminutive suffix was paired with pictures of a single small creature.

For half of the participants $(\mathrm{n}=35)$, those in the Dan condition, the form -dan was more frequent than the others. For the other half, assigned to the Nem condition ( $\mathrm{n}=35$ ), the frequent form was -nem. The competing -sil and -shoon
forms were always equally frequent. The unsuffixed stems constituted the singular non-diminutive form of the noun.

After exposure, participants were tested on both production and comprehension. In the production test, were presented with meanings and asked to express them. Crucially, one of the meanings was a novel one, plural diminutive (multiple small creatures). Each trial began with the presentation of the picture of a novel singular object on the computer screen. The name of the novel object was presented auditorily over headphones as in the training stage. Once the sound finished playing, the picture was removed and replaced with a display of four pictures representing four different meanings: a single object of the same type, a miniature version of the same object, multiple objects of the same type, and multiple miniature objects of the same type. Three of these pictures disappeared, leaving the participants with the one target picture to name (i.e., meaning to express). Participants were asked to generate the form for the target meaning using the stem that was presented and say the form aloud. They had five seconds to do so.

Data were analyzed using logistic mixed-effects models with maximal random-effects structure using the lme4 package (Bates et al., 2015) in R (R Core Team, 2015). Significance was assessed by comparing models with and without a predictor using a log likelihood test. Participants were significantly more likely to use a given form if it was the frequent form during exposure (Figure $1 ; \chi^{2}(1)=21, p<$ .0001). This was not simply an effect of semantic feature frequency, i.e. DIM.PL in Dan, where PL was frequent, was not simply interpreted as PL: the synonym of a frequent form (-sil in Dan and -shoon in Nem) was no more likely to be used to express the novel meaning than the synonym of the infrequent form (-sil in Nem and -shoon in Dan; $p=.9$ ).


Figure 1: An illustration and results of the production test. A suffix is produced more often when it has been experienced more frequently (-dan in Dan and -nem in Nem), both to express the meaning with which it was experienced and to express the novel related meaning (DIMPL)

In the comprehension test, participants were presented with forms and asked to click on the corresponding meaning using the four-picture display briefly flashed in production. The meanings included the familiar meanings as well as the
novel meaning. In this task, participants were less likely to click on the novel meaning given a form that was frequent during exposure (Figure $2 ; \chi^{2}(1)=17, p=.000037$ ). As in production, these effects could not be accounted for by the relative frequencies of the meanings because the synonym of a frequent form was significantly more likely to be mapped onto the novel meaning than the synonym of the infrequent form ( $p<.001$ ). Thus, participants are not simply more likely to click on the more familiar meanings, rather they are more likely to click on familiar meanings in response to the forms that have been frequently paired with them in training. For forms that have been paired with the frequent meaning less frequently, the novel meaning is preferred, despite the fact that these forms are as frequent as synonyms of infrequent forms.


Figure 2: An illustration of the comprehension test and the corresponding results from Harmon \& Kapatsinski (submitted). Responding 'DIMPL' meant clicking on multiple small creatures; 'PL' meant clicking on multiple large creatures, and 'DIM' meant clicking on a single small creature. When a suffix occurred frequently in training (-dan in Dan Condition and -nem in Nem Condition), it became less likely to be mapped onto the novel meaning, DIMPL, and more likely to be mapped onto the meaning it was paired with during exposure.

The results therefore show a production-comprehension dissociation: the forms participants were most likely to use to refer to the novel meaning in production were the forms they were least likely to map onto the novel meaning in comprehension.

Thus, frequency of a form-meaning pairing appears to have opposite effects in production and comprehension. These results therefore appear, at first glance, to be problematic for simple Hebbian models of word learning (McMurray et al., 2012; Yu \& Smith, 2012) that learn symmetrical bidirectional form-meaning mappings based on form-meaning co-occurrence as well as for the notion, nearly universally accepted in linguistics (cf. Ramscar et al., 2010), that linguistic contructions, whether lexical or grammatical, are Saussurean signs - i.e., that there is a form representation that mediates the auditory-to-semantic mapping in comprehension and the semantic-to-articulatory mapping in proeuction. The aim of the present paper is to
show that, somewhat counterintuitively, the observed dissociation actually falls out of simple Hebbian learning of bidirectional form-meaning associations.

## The baseline model: Frequency counter

According to Hebb (1949), neurons that fire together wire together. We assume a distinction between cues and outcomes, where outcomes follow cues. On every trial, associations between the cues present on that trial and the following outcomes strengthen. How much they strengthen is determined by the salience of the cue, the salience of the outcome, and the learning rate. During the exposure trials in Harmon \& Kapatsinski (submitted), forms began 500 ms into the presentation of the referent. Therefore, we assume cues to be the semantic features of the referents (BIG, SMALL, MANY and ONE) plus a context cue, present on every trial (Pavlov, 1927; Rescorla \& Wagner, 1972). This order of presentation was chosen to reflect the temporal dynamics of real-life word learning (Pereira et al., 2014). Unlike error-driven models such as Rescorla \& Wagner (1972), we did not multiply the increment in association strength by prediction error. This is part of what makes the model Hebbian: it does not learn less on trials with unsurprising (or no-longer-surprising) outcomes, and would not exhibit cue competition effects such as blocking or overshadowing.

In essence, this base model is simply counting frequencies of form-meaning mappings. When it encounters a cue (meaning) followed by an outcome (form), it simply increases the weight of the link between them by a constant number, which we set to 1 in order to emphasize the model's nature as a simple frequency counter. The results do not change depending on what the number is.

## Linking hypotheses

In order to connect the model's knowledge to the experimental results, we need a set of linking hypotheses connecting the weights and activations of the model to the participants' responses in the experimental tasks. We assumed that production involves activating forms given the semantic features present on that test trial and the context cue. The activation of a form is simply the sum of connection strengths from the semantic and context cues present on the test trial to that form. The choice of the form is then determined stochastically (Luce, 1963): the form is chosen in proportion to its activation value relative to the sum of all forms' activation values given the cues present. Stochastic choice implements probability matching, an empirical universal in tasks that demand repeatedly choosing between the same alternatives (Azab et al., 2016).

The linking hypothesis for comprehension is more controversial. Note that the model, like the subjects, was trained only in the meaning $\rightarrow$ form direction. However, the comprehension task required participants to choose meanings given forms, reversing the cue $\rightarrow$ outcome mappings they were trained on. Participants were extremely accurate in the comprehension task, suggesting that they
were able to bring the knowledge they acquired to the task. The model must be able to do the same. We propose that the associations participants learn obey the Symmetry Principle: a cue $\rightarrow$ outcome association is as strong as the corresponding outcome $\rightarrow$ cue association (Asch \& Ebenholtz, 1962; Kahana, 2002). This is another way in which the proposed model differs from models that perform using prediction error, such as Rescorla \& Wagner (1972). This difference, however, is crucial for the model's ability to simulate the comprehension data.

We assume that a choice between two meanings depends on the difference in activations between the two meanings' contrasting features. For example, the probability of clicking on [small;plural] rather than [big;plural] when presented with -dan is proportional to the difference in association strengths between $-d a n \sim$ SMALL (=SMALL $\sim-d a n)$ and dan $\sim$ BIG (=BIG $\sim-d a n)$. The bigger this difference, the more likely participants are to click on the meaning that actually was paired with the form in training (Miller \& Matzel, 1988).

## Models between frequency and contingency

Besides the connections between the cues and outcomes present on a particular trial, there are three other sets of connections that could potentially be updated. Alternative theories of associative learning differ in their claims about whether these connections are indeed updated.

First, there are connections from the cues present on a trial to the outcomes absent from that trial. It is usually thought that these connections' weights are reduced, so that cues that are consistently paired with the absence of a certain outcome develop inhibitory connections to that outcome, with the subject learning the negative contingency present in the environment. Second, there are connections from the absent cues to the present outcomes. These connections are assumed not to be updated by Rescorla \& Wagner (1972). However, van Hamme \& Wasserman (1994) and Tassoni (1995) argued that - if participants know the set of cues that could occur on every trial - the absence of a cue could be salient. In other words, learners may notice the consistent absence of a cue on trials containing a certain outcome and develop a negative association between that cue and the outcome. Finally, one could argue that connections from absent cues to absent outcomes may also be updated, gaining strength: when a cue and an outcome are absent together, the learner is in a position to learn that absence of the cue predicts absence of the outcome (Tassoni, 1995). Thus, models of learning can be arranged from simplest (wiring together present cues and outcomes only) and least veridical - least able to faithfully reproduce environmental contingencies - to most complex and most veridical (updating all connections on every trial). In what follows, we examine the types of updating that are needed to capture the experimental results by independently varying whether each distinct set of connections undergoes updating. Table 1 summarizes the possible models from a simple frequency counter that updates only the connections
between present cues and present outcomes to a fully veridical contingency tracker that updates all four sets of connections (in the normative direction). We will refer to the models we evaluate with the abbreviations shown on the left sides of the table cells. For example, the RescorlaWagner model updates only the sets of connections in the top row and can therefore be abbreviated as $\left(p_{c}\right)$.

Table 1: The four distinct sets of cue-outcome connections on every trial and whether their weights should become more positive $(+)$ or more negative $(-)$ in a model that is able to capture environmental contingencies
veridically. The two subscripts c and o refer to cue and outcome respectively. Presence is denoted by p and absence by a.

|  |  | Outcome Present |  | Outcome Absent |  |
| :--- | :--- | ---: | :--- | ---: | :---: |
| Cue Present | $\left(\mathrm{p}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}\right)$ | + | $\left(\mathrm{p}_{\mathrm{c}} \mathrm{a}_{\mathrm{o}}\right)$ | - |  |
| Cue Absent | $\left(\mathrm{a}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}\right)$ | - | $\left(\mathrm{a}_{\mathrm{c}} \mathrm{a}_{\mathrm{o}}\right)$ | + |  |

## Extension of frequent forms in production

Table 2 shows predicted activations of the frequent suffix, its synonym, and the two other suffixes (which are always activated equally) by the semantic features of the novel meaning (MANY and SMALL) under all logically possible models of associative learning. The left column represents the simplest possible model, a frequency counter (Bybee, 2010). Columns 2-4 represent association sets that can be added to the frequency counter in order to make contingency learning more veridical, incorporating learning of connections involving absent cues and/or outcomes. Column 5 is the model that learns only from present cues (a Hebbian version of Rescorla \& Wagner, 1972). Column 6 is the full model that learns about all associations, including associations between absent cues and absent outcomes (a Hebbian version of Tassoni, 1995). Extension of frequent forms to novel meanings is predicted if the activation of the frequent form exceeds that of all other forms, including the frequent form's synonym. In other words, a preference to extend the frequent form to novel meanings is predicted whenever the largest number is in the top row.

As seen in Table 2, extension of the frequent form is predicted by increasing the weights of connections from present cues to present outcomes, as well as by decreasing the weights of connections from present cues to absent outcomes. Updating connections from absent cues (in the normative direction) acts against extension.

For the simulations reported in this table, it was assumed that an absence of a cue or outcome is noticed only half the time while its presence is always noticed. Associative learning in conditioning paradigms tends to be slower when reinforcement is signaled by the absence of a cue than when it is signaled by the presence of a cue (e.g. Wasserman et al., 1990). However, one might question whether absences are missed or ignored that often, and wonder whether noticing absences more would eliminate extension. It turns
out not to matter much: $a_{c}$ does not overpower $p_{c}$ even if absences are as salient as presences. All extant models of learning agree that absent stimuli are no more salient than stimuli that are actually presented and therefore all predict (over-)extension of frequent forms to related meanings in production.

Table 2: Activations of the frequent suffix, its synonym, and the other two suffixes given the novel diminutive plural meaning under alternative models.

| DIM.PL $\rightarrow$ | $\mathrm{p}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}$ | $\mathrm{p}_{\mathrm{c}} \mathrm{a}_{\mathrm{o}}$ | $\mathrm{a}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}$ | $\mathrm{a}_{\mathrm{c}} \mathrm{a}_{\mathrm{o}}$ | $\mathrm{p}_{\mathrm{c}}$ | all |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Frequent | 72 | -42 | -18 | 15 | 30 | 24 |
| Synonym | 24 | -66 | -6 | 21 | -42 | -12 |
| Other | 24 | -66 | -6 | 21 | -42 | -12 |

## Entrenchment in comprehension

Table 3 reports activation differences between features that distinguish the novel meaning from the familiar meaning paired with a form in training. Because of the Symmetry Principle, the activation differences correspond to meaning $\rightarrow$ form connection weights involving the semantic features in question. For example, the activation difference between the non-diminutive and diminutive plural for -dan is the weight of the connection between -dan and BIG minus the weight of the connection between -dan and SMALL. The activation difference between the singular and plural diminutive for -nem is the weight of the connection between -nem and ONE minus the weight of the connection between -nem and MANY (cf. Miller \& Matzel, 1985).

Entrenchment is observed if this difference is larger (more positive) for a frequent form compared to the 'other' forms, i.e. if the value in the top row in Table 3 is larger than the value in the bottom row.

Table 3: Comprehension effects. Each cell contains activation difference between the meaning paired with a form in training and the novel, diminutive plural, meaning.
Activations of shared features of the competing meanings cancel out. Therefore, for plural suffixes this is the difference in activations between BIG and SMALL, and for diminutive suffixes it is the difference between ONE vs.
MANY. Entrenchment is predicted if Frequent $>$ Other.

| Right-New | $\mathrm{p}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}$ | $\mathrm{p}_{\mathrm{c}} \mathrm{a}_{\mathrm{o}}$ | $\mathrm{a}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}$ | $\mathrm{a}_{\mathrm{c}} \mathrm{a}_{\mathrm{o}}$ | $\mathrm{p}_{\mathrm{c}}$ | all |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Frequent | 36 | 0 | 36 | 6 | 36 | 78 |
| Synonym | 12 | -12 | 12 | -6 | 0 | 6 |
| Other | 12 | 12 | 0 | 0 | 24 | 24 |

Table 3 shows that entrenchment is favored by strengthening $\mathrm{p}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}$ connections between present cues and present outcomes, weakening $\mathrm{a}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}$ connections between absent cues and present outcomes, and strengthening $a_{c} a_{0}$ connections between absent cues and outcomes. Because updating $\mathrm{p}_{\mathrm{c}} \mathrm{p}_{\mathrm{o}}$ and $\mathrm{p}_{\mathrm{c}} \mathrm{a}_{0}$ weights pull in different directions, entrenchment only occurs if absent outcomes are less salient
than present outcomes. In other words, weights of connections to absent outcomes must change less than the weights of connections to present outcomes. This appears to be a reasonable assumption (e.g. Tassoni, 1995), though not all extant models make it. For example, the Naïve Discriminative Learner (Baayen et al., 2011), which uses equilibrium equations for the Rescorla \& Wagner (1972) model from Danks (2003: 116), does not show entrenchment because the learning rates for present and absent outcomes in Danks' equations are equal, a simplifying assumption (Danks, 2003: 115-116).

## Conclusion

Studies of comprehension suggest that frequently encountering a form-meaning pairing convinces the learner than the form cannot be used in any other way (Braine \& Brooks, 1995; Brooks \& Tomasello, 1999; Regier \& Gahl, 2004; Stefanowitch, 2008; Xu \& Tenenbaum, 2007). Nonetheless, frequent forms are the ones most likely to be extended to new uses. Using a frequent form in a novel way seeds the process of language change because that novel use can then be picked up by others, spreading through the speech community. As the novel use diffuses through the community, it becomes conventional. Over historical time, extension of frequent forms results in the well-documented correlation between frequency of use and number of senses: in every language, it is the most frequent forms that are most polysemous (Piantadosi et al., 2012; Zipf, 1949).

Conventionalization of extensions is the primary mechanism behind the diachronic process of grammaticalization (Bybee, 2010; Heine, 2011). The importance of this diachronic process can hardly be overstated as it is the primary source of grammar: almost all grammatical morphemes, whether bound affixes or independent functors like prepositions, determiners or auxiliaries are former lexical words that have been gradually extended to new and new uses (Bybee, 2003; 2010; Christiansen \& Chater, 2016).

Despite the correlation between frequency and semantic extension, the causal mechanisms behind grammaticalization remain controversial. For example, Haspelmath (1999) has argued that increases in frequency seen in grammaticalization are caused by the extension of the grammaticalizing form to new uses, which are in turn caused by semantic broadening. Bybee (2003) agrees that semantic broadening causes extension but suggests that high frequency causes semantic broadening. Like Haspelmath (1999), Heine (2011) does not allocate frequency a causal role in the process but suggests that extensions result in broadening.

In our recent experimental work, we have documented that the same speaker may extend a frequent form to a new meaning in production despite being least likely to map it onto the new meaning in comprehension. This suggests that the speaker may extend a form to a new meaning, thereby seeding the process of language change, without necessarily considering the form to be the best way to express that new
meaning. Use in a new context can therefore be caused by high frequency and precede semantic broadening.

In the present paper, we have argued that this productioncomprehension dissociation falls out of simple, Hebbian associative learning models, which acquire symmetrical form-meaning associations based on cue-outcome cooccurrence (Hebb, 1949; Miller \& Matzel, 1985; see also McMurray et al., 2012; Yu \& Smith, 2012). While such dissociations have previously been used to support the idea that form $\rightarrow$ meaning associations are distinct from meaning $\rightarrow$ form associations (Kapatsinski, 2009), the present results indicate that a single set of bidirectional associations suffices.

Remarkably, all that is required to obtain the divergence between frequency effects in production and comprehension - entrenchment of the frequent in comprehension, and extension in production - is the assumption that cue and outcome absences to be less salient than present cues and outcomes, an uncontroversial assumption (Tassoni, 1995; Wasserman et al., 1990) that is also normatively justified: almost every stimulus is absent more often than it is present, hence the presence of a stimulus is typically more informative about the contingencies in the learner's environment than its absence (McKenzie \& Mikkelsen, 2007). Despite being surprising to human theorists, frequency-driven semantic extension is predicted by every associative learning theory.

## References

Ambridge, B., Pine, J. M., Rowland, C. F. \& Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgments of argument-structure overgeneralisation errors, Cognition, 106(1), 87-129.
Asch, S. E., \& Ebenholtz, S. M. (1962). The principle of associative symmetry. Proceedings of the American Philosophical Society, 106(2), 135-163.
Azab, H., Ruskin, D., \& Kidd, C. (2016). Adults’ guesses on probabilistic tasks reveal incremental representativeness biases. Proceedings of the Annual Conference of the Cognitive Science Society, 38, 2831-36.
Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., \& Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. Psychological Review, 118(3), 438-481.
Bates, D., Maechler, M., Bolker, B. M. \& Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Braine, M. D. S. \& Brooks, P. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In M. Tomasello \& W. Merriman (Eds.), Beyond names for things: Young children's acquisition of verbs (pp. 353-376). Hillsdale, NJ: Erlbaum.
Brooks, P., \& Tomasello, M. (1999). How children constrain their argument structure constructions. Language, 75(4), 720-738.

Burke, M. D., Kester Locantore, J., Austin, A. A., Chae, B. (2004). Cherry pit primes Brad Pitt: Homophone priming effects on young and older adults' production of proper names. Psychological Science, 15(3), 164-170.
Bybee, J. (2010). Language, usage and cognition. Cambridge, UK: Cambridge University Press.
Bybee, J. (2003). Cognitive processes in grammaticalization. In M. Tomasello (Ed.), The new psychology of language: Cognitive and functional approaches to language structure (Vol. 2, 145-167). Mahwah, NJ: Erlbaum.
Christiansen, M. H., \& Chater, N. (2016). Creating language: Integrating evolution, acquisition, and processing. Cambridge, MA: MIT Press.
Danks, D. (2003). Equilibria of the Rescorla-Wagner model. Journal of Mathematical Psychology, 47(2), 109121.

Ferreira, V. S., Griffin, Z. M. (2003). Phonological influences on lexical (mis)selection. Psychological Science, 14(1), 86-90.
Gershkoff-Stowe, L., Connell, B., \& Smith, L. (2006). Priming overgeneralizations in two-and four-year-old children. Journal of Child Language, 33(3), 461-486.
Harmon, Z., \& Kapatsinski, V. (In preparation). Putting old tools to novel uses: The role of form accessibility in semantic extension.
Haspelmath, M. (1999). Why is grammaticalization irreversible? Linguistics, 37(6), 1043-1068.
Heine, B. (2011). Grammaticalization in African languages. In H. Narrog \& B. Heine (Eds.), The Oxford handbook of grammaticalization (pp.696-707). New York: Oxford University Press.
Hebb, D. O. (1949). The organization of behavior: A neuropsychological approach. Malden, MA: Wiley.
Kahana, M. J. (2002). Associative symmetry and memory theory. Memory and Cognition, 30(6), 823-840.
Kapatsinski, V. (2009). Adversative conjunction choice in Russian (no, da, odnako): Semantic and syntactic influences on lexical selection. Language Variation and Change, 21(2), 157-173.
Luce, R. D. (1963). A threshold theory for simple detection experiments. Psychological Review, 70(1), 61-79.
McKenzie, C. R., \& Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. Cognitive Psychology, 54(1), 33-61.
McMurray, B., Horst, J. S., \& Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological Review, 119(4), 831-877.
Miller, R. R., \& Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. The Psychology of Learning and Motivation, 22, 51-92.
Naigles, L. G., \& Gelman, S. A. (1995). Overextensions in comprehension and production revisited: Preferentiallooking in a study of dog, cat, and cow. Journal of Child Language, 22(1), 19-46.

Pereira, A. F., Smith, L. B., \& Yu, C. (2014). A bottom-up view of toddler word learning. Psychonomic Bulletin \& Review, 21(1), 178-185.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122(3), 280-291.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.Rproject.org/
Ramscar, M., Yarlett, D., Dye, M., Denny, K., \& Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. Cognitive Science, 34(6), 909-957.
Regier, T., \& Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. Cognition, 93(2), 147-155.
Rescorla, R. A., \& Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black \& W. F. Prokasy (eds.), Classical conditioning II: Current research and theory, 64-99. New York: Appleton-Century-Crofts.
Stefanowitsch, A. (2008). Negative entrenchment: A usagebased approach to negative evidence. Cognitive Linguistics, 19(3), 513-531.
Tassoni, C. J. (1995). The least mean squares network with information coding: A model of cue learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(1), 193-204.
Theakston, A. L. (2004). The role of entrenchment in children's and adults' performance limitations on grammaticality judgment tasks, Cognitive Development, 19(1), 15-34.
van Hamme, L. J., \& Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. Learning and Motivation, 25(2), 127-151.
Wasserman, E. A., Dorner, W. W., \& Kao, S.-F. (1990). Contributions of specific cell information to judgments to interevent contingency. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 509521.

Wonnacott, E., Newport, E. L., Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. Cognitive Psychology, 56(3), 165-209.
Xu, F., \& Tenenbaum, J. B. (2007). Word learning as Bayesian inference. Psychological Review, 114(2), 245272.

Yu, C., \& Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. Psychological Review, 119(1), 21-39.
Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Oxford: Addison-Wesley.

# Effects of Delayed Language Exposure on Spatial Language Acquisition by Signing Children and Adults 

Dilay Z. Karadöller (dilay.karadoller@mpi.nl)<br>Center for Language Studies, Radboud University, Nijmegen, The Netherlands International Max Planck Research School for Psycholinguistics, Nijmegen, The Netherlands

Beyza Sümer (beyza.sumer@mpi.nl)
Center for Language Studies, Radboud University, Nijmegen, The Netherlands Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Aslı Özyürek (asli.ozyurek@mpi.nl)
Center for Language Studies \& Donders Center for Cognition, Radboud University, Nijmegen, The Netherlands Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands


#### Abstract

Deaf children born to hearing parents are exposed to language input quite late, which has long-lasting effects on language production. Previous studies with deaf individuals mostly focused on linguistic expressions of motion events, which have several event components. We do not know if similar effects emerge in simple events such as descriptions of spatial configurations of objects. Moreover, previous data mainly come from late adult signers. There is not much known about language development of late signing children soon after learning sign language. We compared simple event descriptions of late signers of Turkish Sign Language (adults, children) to age-matched native signers. Our results indicate that while late signers in both age groups are native-like in frequency of expressing a relational encoding, they lag behind native signers in using morphologically complex linguistic forms compared to other simple forms. Late signing children perform similar to adults and thus showed no development over time.


Keywords: sign language; late acquisition; spatial relations; left right

## Introduction

The most frequently preferred way of spatial encoding in sign languages requires the use of morphologically complex linguistic forms and use of signing space analogue to how the entities are located with respect to each other in the real space (e.g., Emmorey, 2002). In these forms, called classifier predicates (CLs), signers use their hands to represent the location and motion of the entities, as shown in Figure 1a below. Morphological complexity of these forms comes from the need for choosing the correct handshape for the entities (e.g., index finger for long and thin entities) and simultaneous coordination of both hands in the signing space to express their locations (e.g., Supalla, 1982). The relative spatial location of these forms in signing space represents spatial relations among entities in an analogue way - unlike the categorical forms (i.e., ad positions or spatial nouns) in spoken languages.
Earlier studies on spatial language acquisition of native signing deaf children, mostly focused on motion events, have claimed such morphological complexity to be a hindering factor compared to hearing children (e.g.

Engberg-Pedersen, 2003; Slobin, 2003). However, recent work has found that Turkish Sign Language (TID) acquiring deaf children learn to encode static location of the objects placed on a lateral axis (e.g., pen left to paper, apple right to box) earlier than their hearing peers (Sümer, 2015; Sümer, Perniss, Zwitserlood, Özyürek, 2014). Even though these children mostly used classifier predicates, they also used specific lexical signs (i.e., relational lexemes) for left and right, which are body-anchored in TID (see Figure 1b), as frequently as the native signing adults. Thus, classifier predicates and/or body anchored relational lexemes seem to facilitate the learning of static spatial relations for deaf children who have been exposed to sign language input since birth (native signers).


Figure 1: Descriptions from adult native signers of TiD for the spatial relation of the pen with respect to the paper using (a) classifier predicates and (b) body anchored relational lexeme for 'left' (Sümer et al., 2014).

It is not known however if similar patterns also emerge in language development of deaf children with delayed sign language exposure (late signers). A previous study with deaf children (ages 5-6 yrs.) who were never exposed to a sign language (i.e., home signers) found no evidence of relational encoding for spatial relation (Gentner, Özyürek, Gürcanli, Goldin-Meadow, 2013). Here, we investigate whether spatial encodings of a similar type studied by Sümer and her colleagues (2014) can also be learned within a short time by late signing children after brief exposure ( 2 years) to sign language (after the age of 6 ) or whether late exposure to sign language is a
drawback in mastering spatial language in general. Moreover, we investigate the effects of delayed language exposure on both late signing children as well as adults to see if any developmental pattern emerges in such delayed exposure to language and if yes how.

A body of evidence on the effects of late sign language acquisition by deaf individuals posits that adult late signers lag behind adult native signers in several domains such as general cognitive abilities (Bebko \& McKinnon, 1990; Mayberry \& Waters, 1991; Parasnis et al., 1996; Wilson, Bettger, Niculae, \& Klima, 1997) and sign language comprehension (Emmorey, 1993; Emmorey \& Corina, 1992; Mayberry \& Eichen, 1991). Nevertheless, most of the studies were restricted only to sign language comprehension, and there are only a few studies on sign language production which focus only on adult patterns. Question of how late signing children perform in production compared to late adult signers and also to their age-matched native signing peers could be informative in understanding the impact of delayed language exposure for adult and child late signers, and to what extent development plays a role in delayed language acquisition.

In series of studies, Newport (1988; 1990) investigated language production of adult late signers through descriptions of motion events. She found that they described motion events by using fewer classifier predicates (CLs) than native signers, but rather preferred simple forms such as lexical verbs. Given that the motion events consist of several components, such as Figure, Ground, Path, Manner (Talmy, 1985), it raises the question of whether similar patterns emerge for late signing children and adults in describing static events which has fewer components, such as only Figure and Ground. In addition, a previous study (Sümer, 2015) revealed that adult-like descriptions of static spatial relations are acquired earlier for static locations than for motion event descriptions in TİD by native signing deaf children. We do not know if these patterns also apply to late signing children and adults of TİD and if there are any developmental patterns in late signers.

## The Present Study

To answer above questions, this study investigates the effects of late sign language exposure on the ability to encode static spatial events for left-right spatial relations by late signers of TID and compares them to those of native TİD signers, which were already reported by Sümer et al. (2014) and Sümer (2015). We also compare the descriptions of late signing children and adults to those of native children and adults obtained earlier by Sümer (2015) to see if there is any developmental pattern in late signers.

## Participants

Seven adult late signers ( $30 ; 0-49 ; 0, M=39.14$ years), 7 child late signers $(7 ; 3-10 ; 9, \mathrm{M}=7 ; 8), 10$ native adult signers $(18 ; 5-45 ; 10, M=31 ; 4)$, and 10 native child
signers $(7 ; 2-9 ; 10, \mathrm{M}=8 ; 3)^{1}$ participated in the study. All late signers had learned TiD after age 6 when they started a school for the deaf. Late signers did not have any prior exposure to sign language because according to the Turkish Education System, age 6 is the earliest age for starting school in Turkey. Before starting school, these signers mainly stayed at home with their non-signing parents. As a result, late signing children participated in this study, had exposure of 2 years of sign language at the time of testing. Before starting to school, both children groups got two hours of speech therapy every week. At the end of the study, adult participants received a small monetary compensation; child participants received a gender neutral color pencil kit.

## Stimuli and Procedure

Stimuli included a set of 36 displays. In each display, there were 4 pictures ${ }^{2}$ consisted of two entities placed in various spatial configurations (Left, Right, In, On, Under, Front and Behind) to each other. Within each display (four-picture set), only one picture was considered to be the target, which was marked with a red outer frame. The experimental displays we focused on in this study consisted of 6 displays, in which the target picture depicted either Right or Left configurations (Figure 2).


Figure 2: Example of a display. The target picture (apple to the right of the box) to be described is framed in red.

Participants were seated across a confederate deaf addressee who was a native signer of TİD. Stimuli were presented through a $15^{\prime \prime}$ MacBook Pro computer. Computer screen was only visible to the participants. Participants were asked to describe the target pictures to the addressee and were not instructed to use any specific strategy. In order to create an interactive nature, addressee was given a booklet containing the same displays and was asked to point at the picture that the participant described. At the end of the experiment, participants filled out demographics and language background surveys.

[^361]
## Coding

Descriptions for the target items were coded for the presence of spatial encoding and choice of linguistic strategies used to describe spatial relationship between two entities by using ELAN, a free annotation tool (http://tla.mpi.nl/tools/tla-tools/elan/) for multimedia resources developed by the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands (Wittenburg et al., 2006). We categorized three main linguistic strategies used to encode, specifically, Figure object's relation to Ground object: Classifier predicates CLs (Figure 1a), Relational Lexemes like LEFT and RIGHT (Figure 1b), and other alternative linguistic forms. Other forms included showing the location of objects through pointing (59\% of the cases; Figure 3), placing the objects in the signing space through virtually drawing by hand (SASS), using a lexical verb to infer objects (e.g. using sign for "to sit" to represent location of a boy).

Annotation of data from the native signers were done by hearing research assistants and a native TID signer and coded by the second author of this study. Annotation of data from the late signers were done by hearing research assistant and coded by the first author of this study. Later, all annotation and coding was checked by the second author of this study. All annotators and coders had knowledge of TiD.

(a) RH: Point (cat)locR LH: CL (boat)locL

(b) RH: CL(house)locR LH: Point (horse)locL

Figure 3: Descriptions from late signers of TiD for two objects located on the lateral axis by using pointing by (a) an adult TİD signer (middle finger point) and (b) a child TID signer (index finger point).

## Results

Mean proportions of linguistic strategies were calculated for each participant. Arcsine transformation was applied to all data to ensure normality. The mean proportions and standard errors (SEs) in the table and the graphs are reported from the untransformed data.

First we investigated the frequency of encoding a spatial relation of entities by different groups in each language status. Results of the 2 (Age: Adults, Children) X 2 (Status: Native, Late) Between Subjects ANOVA on mean proportions of encoding a spatial relation revealed no main effect of age, $F(1,30)=2.918, p=.098, \eta^{2}=$ $.089, M S E=.381$, no main effect of status, $F(1,30)=$ $.373, p=.546, \eta^{2}=.012, M S E=.049$ and no interaction, $F$ $(1,30)=768, p=.388, \eta^{2}=.023, M S E=.100$. These results indicate that all groups of participants generated equal amount of relational encodings (Table 1).

Table 1: Mean Proportions and SEs of frequency of encoding a spatial relation as a function of Age and Status.

| Participants | Native Signers | Late Signers |
| :--- | :---: | :---: |
| Adults | $0.97(.02)$ | $0.98(.03)$ |
| Children | $0.92(.05)$ | $0.81(.11)$ |

As the next step, we investigated what types of linguistic strategies were preferred by each age group and status. The results of a 2 (Between Subject, Age: Adults, Children) X 2 (Between Subject, Status: Native, Late) X 3 (Within Subject, Linguistic Strategy: CLs, RLs, Other) Mixed ANOVA yielded a main effect of Linguistic Strategy, $F(1.32,12.42)=37.114, p<.001, \eta^{2}=.533$, $M S E=.314$, no main effect of age, $F(1,30)=.347, p=$ $.560, \eta^{2}=.011, M S E=.008$, no main effect of Status, $F$ $(1,30)=1.385, p=.249, \eta^{2}=.044, M S E=.008$. Due to a marginal interaction between Linguistic Strategy and Status, $F(1.32,12.42)=3.628, p=.053, \eta^{2}=.108, M S E=$ .314 , separate analyses were conducted for each Status.

## Linguistic Strategies used by Native Signers

The results of 2 (Between Subjects, Age: Adult, Child) X 3 (Within Subjects, Linguistic Strategy: CLs, RLs, Other) mixed ANOVA showed no main effect of age, $F(1,18)=$ 1.214, $p=.285, \eta^{2}=.063, M S E=.028$ but a main effect of Linguistic Strategy, $F(1.25,22.53)=60.186, p<.001$, $\eta^{2}=.770, M S E=.183$, without an interaction between them, $F(1.25,2.53)=.251, p=.674, \eta^{2}=.315, M S E=$ .183. Tests of within subject comparisons showed that classifier predicates were preferred more frequently than relational lexemes ( $p<.001$ ) and other linguistic forms ( $p$ $<.001$ ). The frequency of using relational lexemes and other forms found to be similar to each other ( $p>.05$ ). The lack of main effect for age indicated that deaf children used the linguistic forms in three different categories as frequently as deaf adults. See Figure 4 below.


Figure 4: Mean proportions and SEs of linguistic descriptions of native signers as a function of age.

## Linguistic Strategies used by Late Signers

The results of 2 (Between Subjects, Age: Adults, Children) X 3 (Within Subjects, Linguistic Strategy: CLs, RLs, Other) mixed ANOVA showed no main effect of age, $F(1,12)=.097, p=.761, \eta^{2}=.008, M S E=.015$ but a main effect of Linguistic Strategy, $F(1.18,14.10)=6.771$, $p=.017, \eta^{2}=.036, M S E=.588$, without an interaction between them, $F(1.18,14.10)=.450, p=.544, \eta^{2}=.036$, $M S E=.588$. Tests of within subject comparisons showed that relational lexemes were used less frequently than classifier predicates ( $p<.01$ ) and other forms ( $p<.05$ ). The frequency of using classifier predicates and other forms are found to be similar to each other $(p=573)$. The results indicate that late signers -unlike native signers- use other forms as frequently as classifier predicates. See Figure 5 below.

Results showed that native and late signers show different production patterns in their descriptions of the location of the objects when they are placed to left or right to each other. Namely, native signers show a significant preference over using the morphologically complex CLs in describing, however, late signers employ other simpler forms as frequently as CLs. Moreover, this tendency is significant for all age groups.


Figure 5: Mean proportions and SEs of linguistic descriptions of late signers as a function of age.

## Discussion and Conclusion

Our study has two key findings. First, late signing adults and children differ from native signing adults and children in their linguistic descriptions. Namely, late signers do not
show a preference for CLs. Rather they employ simpler other forms as frequently as CLs compared to native signers. These findings clearly indicate that late exposure to sign language by deaf individuals has long-term effects on their production patterns. Second, late signing children perform similar to late signing adults in their static spatial event descriptions and show no further developmental trajectory in their preferred linguistic devices.

Results of this study complement the previous literature - yet in another sign language - showing a tendency towards decreased preference for CLs and increased preference for other simpler forms not only for complex event descriptions, as found in ASL (Newport, 1988; 1990), but also for simple static events (Sümer et al., 2014; Sümer, 2015). In the case of native signers of TID, both adults and children prefer CLs more than RLs and other forms while the nature of this distribution is different for late signers of TİD, who use CLs and simpler other forms such as pointing in similar amounts while they prefer RLs for left and right less frequently. The body-anchored relational lexemes in TİD were very rarely used by late signing adults and not at all by the late signing children. Thus, the use of CLs and RLs did not seem to ease acquisition for these children as they did for native children. However, we should be cautious in generalizing the effects of these comparisons due to relatively few number of participants we could investigate especially from late signing group.

Considering the high proportion of relational encodings shown on Table 1, 2-year exposure to sign language after age 6 seems to be enough to initiate spatial language production in late signing children to become adult like. This is a rather striking finding since Gentner and her colleagues (2013) found that home signing children (deaf children never exposed to a sign language) in similar age range do not produce gestures in a way that would encode spatial relations, which indicates the necessity of sign language input for relational encoding to emerge at all. This pattern also shows that while relational encoding might emerge as the earliest feature of encoding a spatial event, and as less sensitive to delayed exposure, morphologically complex forms or body anchored relational lexemes such as LEFT and RIGHT might be more resistant to the timing of the input to develop.

The difference in the preference for linguistic forms to encode spatial relations between native signers and late signers provides evidence for the role of maturational constrains in language acquisition. Moreover, previous studies have been mostly on comprehension of sign language (e.g., Mayberry \& Eichen, 1991), our results indicate that these effects are not only restricted to comprehension, but also observed in production, as also shown by Newport (1988; 1990). Furthermore, it shows that these results on motion event expressions can be extended to those of static events.

Our study also uniquely displays these patterns for late signing children for the first time in the literature. Although, late signing adults of TİD have lengthier language experience compared to late signing children, who have only a 2 -year of sign language exposure, we still observed similar preferences in how frequently they prefer CLs and other forms. It seems that the hindering
effects of delayed language exposure persist in language production despite many years of language use. This finding is also in line with other studies that show the significance of age of acquisition, rather than the length of exposure, in both sign and spoken language development (Mayberry, 2010).

Finally, late signing children were not only exposed to sign language late but also the kind of language they were exposed to, that is the language used by their late signing peers in the primary school, was also non-native. Thus, future research should also investigate the effect of type of language input (i.e., non-native input) in addition to age of acquisition in late signing children and adults on late signers' language production.

## Acknowledgements

This work has been supported by European Research Council (ERC) Starting Grant and NWO VICI Grant awarded to the third author. We thank our deaf assistant Sevinç Yücealtay Akın and our hearing assistant Hükümran Sümer for their help with data collection and annotation. Moreover, we thank directors and teachers at ÇED and Rotary Tuzla Primary School for the Deaf in Istanbul, Turkey.

## References

Bebko, J. M., \& McKinnon, E. E. (1990). The language experience of deaf children: Its relation to spontaneous rehearsal in a memory task. Child Development, 61, 1744-1752.
Emmorey, K., \& Corina, D. (1992). Differential sensitivity to classifier morphology in ASL signers. Linguistic Society of America.
Emmorey, K. (2002). Language, cognition, and the brain: Insights from sign language research. Mahwah, NJ: Lawrence Erlbaum Associates.
Emmorey, K. (1993). Processing a dynamic visualSpatial language: Psycholinguistic Studies of American Sign Language. Journal of Psycholinguistic Research, 22, 153-187.
Engberg-Pedersen, E. (2003). How Composite is a Fall? Adult's and Children's Descriptions of Different Types of Falls in Danish Sign Language. In K. Emmorey (Ed.), Perspectives on Classifier Constructions in Sign Languages. Mahwah, NJ: Lawrence Erlbaum Associates.
Gentner, D., Özyürek, A., Gürcanli, Ö., \& GoldinMeadow, S. (2013). Spatial language facilitates spatial cognition: Evidence from children who lack language input. Cognition, 127, 318-330.
Mayberry, R. I. (2010). Early language acquisition and adult language ability: What sign language reveals about the critical. The Oxford handbook of deaf studies, language, and education, 2, 281.
Mayberry, R. I., \& Eichen, E. B. (1991). The long-lasting advantage of learning sign language in childhood: Another look at the critical period for language acquisition. Journal of Memory and Language, 30, 486512.

Mayberry, R. I., \& Waters, G. S. (1991). Children's memory for sign and fingerspelling in relation to
production rate and sign language input. Theoretical Issues in Sign Language Research, 2, 211-229.
Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. Language Sciences, 10, 147-172.
Newport, E. L. (1990). Maturational constraints on language learning. Cognitive Science, 14, 11-28
Parasnis, I., Samar, V. J., Bettger, J. G., \& Sathe, K. (1996). Does deafness lead to enhancement of visual spatial cognition in children? Negative evidence from deaf nonsigners. Journal of Deaf Studies and Deaf Education, 1, 145-152.
Slobin, D. I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. Language in mind: Advances in the Study of Language and Thought.
Sümer, B., Perniss, P.M., Zwitserlood, I.E.P. \& Özyürek, A. (2014). Learning to express "left-right" \& "frontbehind" in a sign versus spoken language. In P. Bello, M. Guarini, M. McShane \& B. Scassellati (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society. Austin,Tx: Cognitive Science Society.
Sümer (2015). Acquisition of Spatial Language by Signing and Speaking Children: A comparison of Turkish Sign Language (TiD) and Turkish. Unpublished doctoral dissertation, Radboud University Nijmegen, Nijmegen.
Supalla, T.R. (1982). Structure and acquisition of verbs of motion and location in American Sign Language. Unpublished doctoral dissertation, UCSD, The USA.
Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. Language Typology and Syntactic Description, 3, 57-149.
Wilson, M., Bettger, J. G., Niculae, I., \& Klima, E. S. (1997). Modality of language shapes working memory: Evidence from digit span and spatial span in ASL signers. Journal of Deaf Studies and Deaf Education, 150-160.
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

# A Meta-Analysis of the Joint Simon Effect 

April Karlinsky (april.karlinsky@alumni.ubc.ca)<br>School of Kinesiology, University of British Columbia<br>210 - 6081 University Boulevard, Vancouver, BC V6T 1Z1 Canada<br>Keith R. Lohse (rehabinformatics@gmail.com)<br>School of Kinesiology, Auburn University<br>301 Wire Road, Auburn, AL 36849 USA<br>Melanie Y. Lam (mlam@stfx.ca)<br>Department of Human Kinetics, St. Francis Xavier University<br>PO Box 5000 (Courier 1 West Street), Antigonish, NS B2G 2W5 Canada


#### Abstract

Since its design in 2003, the joint Simon task and corollary joint Simon effect (JSE) have been invaluable tools towards the study of joint action and the understanding of how individuals represent the action/task of a co-actor. The purpose of this meta-analysis was to systematically and quantitatively review the sizeable behavioural evidence for the JSE. Google Scholar was used to identify studies citing the first report of the joint Simon task (Sebanz, Knoblich, \& Prinz, 2003) up until June 23, 2015. After screening, thirtynine manuscripts were included in the meta-analysis, thirteen of which included individual go/no-go (IGNG) control data. Separate random-effects models were conducted for both the joint Simon and IGNG datasets, and meta-regression models were used to assess potential moderators that may impact the strength of the JSE. The results provide an important quantitative summary of the literature and serve as a foundation for future research surrounding the JSE.


Keywords: joint action; spatial compatibility; corepresentation

## Introduction

Throughout the day, people engage in a variety of social interactions that mold our behaviour, and even independent events can be shaped by those around us. In recent years, much research has been devoted to better understanding how individuals mentally represent the presence, tasks, and actions of others, and how such representations influence one's own behaviour, in contrast to matched behaviours performed alone. A valuable experimental paradigm towards this end has been a spatial compatibility task, more specifically the Simon task, which can be performed in an individual (e.g., Simon, 1969) or joint setting (e.g., Sebanz, Knoblich, \& Prinz, 2003).

## The (Joint) Simon Task

In a typical two-choice Simon task, stimuli are presented to the left or right of centre. A non-spatial stimulus feature (e.g., colour, shape, tone pitch) informs the participant whether to make either a left or right key press response. For example, a triangle requires a left key press response while a circle requires a right key press response. Even though the stimulus location (left, right) is irrelevant to the task, it nevertheless modulates responses, such that responses are faster and more accurate when the spatial
location of the stimulus and response are compatible (e.g., left-left) than when they are incompatible (e.g., left-right). This phenomenon, known as the spatial compatibility or Simon effect, has been shown to be robust, with this pattern of results replicated in many studies (for review, see Lu \& Proctor, 1995).

In a social variant of the Simon task, two people are each assigned a stimulus-response mapping, such that a go/no-go protocol is completed independent from, yet complementary to the other's task. The emergence of a spatial compatibility effect (henceforth referred to as a joint Simon effect, JSE) in the joint setting was taken as evidence that representations were formed for not only one's own part of the task but also their co-actor's (Sebanz et al., 2003), since the effect was noticeably absent when participants performed the same go/no-go protocol alone (individual go/no-go task, IGNG) (see Callan, Klisz, \& Parsons, 1974).

## Interpretations of the Joint Simon Effect

The JSE was originally interpreted in terms of the action corepresentation account (e.g., Sebanz et al., 2003; for elaboration and more detailed review of this and subsequent interpretations, see Dolk et al., 2014). According to this account, individuals represent a co-actor's task quasiautomatically; it is the representation of the alternative stimulus-response mapping that is thought to increase response conflict, eliciting the JSE. Other authors have posited the actor co-representation account, whereby response conflict emerges from the representation of the coactor, as opposed to the co-actor's specific task, such that conflict surrounds which agent should act when (Wenke et al., 2011). However, these accounts do not explain why the JSEs are induced in non-social contexts (e.g., Guagnano, Rusconi, \& Úmilta, 2010). In efforts to offer a more comprehensive explanation for the JSE, Dolk, Hommel, Prinz, and Liepelt (2013) formulated and tested the referential coding account. This account posits that greater similarity across action event representations can lead to a greater emphasis on their discriminating features (e.g., location). In a series of five experiments, they manipulated the social nature of the experimental setup in two ways: (1) absence of a biological co-actor, and (2) removing any event character (e.g., sound). They showed that the JSE could be
elicited by non-social action events (e.g., Japanese waving cat) but not if the "event-like character of the sounds and movements" are eliminated (Dolk, Hommel, et al., 2013, p. 1255). What makes the referential coding account particularly appealing is that it can explain not only the occurrence of the JSE in non-social contexts, but also the more pronounced JSEs observed when there is increased self-other integration (Colzato, de Bruijn, \& Hommel, 2012), as presumably within friendly partnerships (e.g., Hommel, Colzato, \& van den Wildenberg, 2009), in-group interactions (e.g., Iani, Anelli, Nicoletti, Arcuri, \& Rubichi, 2011; McClung, Jentzsch, \& Reicher, 2013; Müller, Kühn, et al., 2011), and cooperative contexts (e.g., Iani et al., 2011; Iani, Anelli, Nicoletti, \& Rubichi, 2014).

## Current Meta-Analytic Review

The current meta-analysis offers several novel contributions to the field of joint action. First, to our knowledge, it is the only application of quantitative methods to evaluate the substantial body of work that has emerged since the introduction of the joint Simon task (Sebanz et al., 2003). As such, it complements recent qualitative literature reviews (e.g., Dolk et al., 2014) while providing unique insights into the nature of co-representation, as indexed by the JSE. Second, we explored the size of: (1) the overall JSE, (2) the JSE when the original conditions were conceptually replicated (see Sebanz et al., 2003, Experiment 1), and (3) the JSE when an elimination or reversal of the effect was anticipated. The inclusion of these latter moderator analyses enhances our understanding of the JSE and its sensitivity to experimental manipulations. Third, we included an analysis of the IGNG task, which is considered an important control when investigating the JSE and enriches interpretations of joint effects (e.g., Sebanz et al., 2003).

## Methods

## Search Strategy and Study Selection

On June 23, 2015, two authors (AK and MYL) conducted an electronic search in Google Scholar for all citations of Sebanz et al. (2003). Following the addition of Sebanz et al. (2003) to the search results and removal of duplicates, 329 records were screened for eligibility. The following exclusion criteria were used to screen the articles: (a) manuscripts that were not published or translated into English; (b) manuscripts that did not include a joint Simon task; (c) studies that did not report response times (RT) and standard deviations (SD) or standard errors (SE); (d) studies examining children ( $<18$ years old). It should be noted that articles examining joint action in special populations (e.g., individuals with schizophrenia) were not excluded, but only the data for healthy controls were included in the analyses.

Two authors (AK and MYL) screened articles by title and abstract according to these criteria. These same authors then used the criteria to screen the remaining 61 articles by full text for inclusion. When there was disagreement, the authors discussed the articles in question until consensus was
reached. A total of 42 manuscripts remained eligible for inclusion in the quantitative analysis, but 3 of these manuscripts were subsequently excluded as they were doctoral dissertations whose eligible studies were also published (and included) as distinct manuscripts (Anelli, 2012; Müller, 2013; Sellaro, 2013). The 39 manuscripts remaining in the meta-analysis comprised 104 independent groups of participants (contributing 95 joint Simon datasets and 35 IGNG datasets), as some manuscripts contained multiple experiments and/or multiple groups of participants. ${ }^{1}$

## Data Extraction

Two authors (AK and MYL) independently extracted data from each manuscript relevant to sample size, experimental manipulation, and response time (means and SDs or SEs). ${ }^{2}$ When necessary, data were manually estimated from reported figures. These two authors discussed any discrepancies between their extractions until consensus was reached with respect to the data included in the analyses.

## Data Analysis

Cohen's $d$ was calculated directly from the extracted RT data and the pooled between-subject SD . In cases of repeated measures designs, data were averaged across conditions such that each independent group of participants contributed only one effect size to each analysis. The effect sizes and variances were entered into a random-effects meta-analysis using the 'metafor' package in R (R Core Team 2014; Viechtbauer, 2010) and the DerSimonian and Laird method of estimation (Borenstein, Hedges, Higgins, \& Rothstein, 2011). Effect size calculation was arranged such that effects favouring a JSE always had a positive value (i.e., incompatible mean RT - compatible mean RT). An effect size of zero indicated no difference between compatible and incompatible trials.

Custom scripts were written to test random-effects models for the overall effect of spatial compatibility within joint Simon and IGNG tasks (Cooper, Hedges, \& Valentine, 2009), and Egger's test of asymmetry was used to assess bias (Egger, Davey Smith, Schneider, \& Minder, 1997). Considering the wide range of experimental manipulations within the joint Simon task literature, we also conducted two moderator analyses using meta-analytic regression. First, conditions conducted as controls (control moderator) were compared to all other conditions, ${ }^{3}$ to provide an index of the JSE unmodulated by experimental variables. Second,

[^362]conditions hypothesized by the original authors to eliminate or reverse the JSE (wipeout moderator) were compared to all other effect sizes. ${ }^{4}$ Unlike the "overall" random-effects model of the JSE, in cases of repeated measures designs, we preferentially submitted a group's 'control' or 'wipeout' data (when available) towards the relevant meta-regression model (rather than averaging across within-group conditions). Details regarding the raw data, moderator coding, and analysis scripts are available online at https://github.com/keithlohse/social_simon_meta.

## Results

## No Spatial Compatibility Effect in IGNG Contexts

As expected, the IGNG studies $(n=35)$ yielded no evidence of a spatial compatibility effect (i.e., the RT difference between incompatible and compatible trials was not statistically different from zero), $d=0.07,95 \%$ confidence intervals (CI) $[-0.01,0.16]$. A statistical test of asymmetry revealed the distribution was not skewed, $t(33)=-0.76, p=$ .45.

## Evidence of Positivity Bias and Small Effect Sizes Across Joint Simon Studies

Prior to analysis, a funnel plot revealed an extremely positive and imprecise effect size (from Dolk et al., 2012, see data point in bottom right corner of Figure 1A) which was removed from all subsequent analyses.

Figure 1A shows the distribution of joint Simon task effect sizes as a function of the standard error in each study ( $n=94$ ). Even with the Dolk et al. (2012) data point removed, a statistical test of asymmetry confirmed the positive skew in these data, $t(92)=3.25, p=.002$, indicating significant bias, with more small, positive studies being published. The random-effects model summary effect size was $d=0.26,95 \% \mathrm{CI}[0.21,0.30]$.

Considering the significant positive skew across the dataset, we also ran a second random-effects model restricted to large samples in efforts to remove bias. ${ }^{5}$ Restricted to the largest studies $(n=20)$, the distribution was not skewed, $t(18)=0.96, p=.35$, and the summary effect size was reduced, $d=0.17,95 \%$ CI $[0.10,0.25]$.

## No Evidence Control Conditions Moderate the JSE

We used meta-regression to compare the effect sizes derived from control conditions $(n=23)$ to all other conditions ( $n=$ 71), to broadly assess any modulation of the effect by experimental manipulations. There was no significant difference between the effect sizes of control conditions, $d=$ $0.34,95 \% \mathrm{CI}[0.24,0.44]$, compared to non-control

[^363]conditions, $d=0.24,95 \%$ CI [0.19, 0.29], $p=.074$. As shown in Figure 1B, the distribution of the effect sizes remained significantly skewed, $t(92)=3.33, p=.001$.


Figure 1: The funnel plots for the JSE (incompatible (IC) mean RT - compatible (C) mean RT) showing effect sizes (d) as a function of precision (standard error) for the A) overall random-effects model; B) meta-regression of the control moderator (triangles $=$ controls; circles $=$ noncontrols); and C) meta-regression of the wipeout moderator (triangles = wipeouts; circles = non-wipeouts). Positive values show a difference in favour of a JSE (i.e., faster RTs on compatible trials)

## Wipeout Conditions Decrease the JSE

Considering that 'non-control' conditions encompass both those experimental designs hypothesized to augment and to diminish the JSE, we conducted an additional metaregression model to assess any moderating effects of conditions explicitly hypothesized by the original authors to eliminate or reverse the JSE $(n=16)$ compared to those that were not ( $n=68$ ). The summary effect size of wipeout conditions ( $d=0.12,95 \%$ CI $[0.01,0.22]$ ), was significantly smaller than that of non-wipeout conditions ( $d=0.33,95 \%$

CI [0.27, 0.38]), $p<.001 .{ }^{6}$ As shown in Figure 1C, the distribution remained skewed, $t(82)=2.77, p=.007$.

## Discussion

Since its design in 2003, many researchers have used the joint Simon task to explore the nature, extent, and boundaries of shared representations, as indexed by the JSE (for review, see Dolk et al., 2014). The present metaanalysis provides the first, much-needed quantitative summary of the literature, and serves both as a snapshot of the research to date, and a foundation on which to build future inquiries.

Across 39 manuscripts, our meta-analysis suggests the JSE is a reliable, albeit small, effect (summary $d=0.26$ ). However, this analysis also revealed significant asymmetry within the data, potentially indicative of publication bias. Specifically, the data are positively skewed (even after removing an outlier), such that more small "positive" studies are being published than those with "negative" results. When we limited our analysis to large samples, the distribution was no longer skewed, but it revealed that the "real" effect size is likely even smaller than it first appeared ( $d=0.17$ ). This has two principle implications: (1) researchers studying this effect need an adequate sample size to achieve statistical power, and (2) there is probably limited "practical" significance of this effect, although it is still useful as a behavioural assay to understand cognitive processes (when conducted with sufficient power).

The small JSE effect size also reinforces the importance of the IGNG random-effects model, where we confirmed that a compatibility effect did not arise under individual task conditions. It should be noted that of the 39 manuscripts eligible for the joint Simon analysis, only 13 included an IGNG condition. The failure to include such a condition is of potential concern as it has been shown that a small but significant spatial compatibility effect can be observed in a go/no-go task (see Callan et al., 1974). In the case that a significant effect is found in the IGNG condition, then this compromises interpretations of the JSE.

Given the sizeable body of research included in the random-effects model of the JSE, we sought to parcel out factors that could be moderating the size of the JSE. We began with an exploration of control versus non-control task conditions. The meta-regression analysis revealed no evidence that control conditions yielded JSEs that were reliably different to those of experimental conditions. A possible explanation for our finding is that task conditions have been manipulated to elicit a range of effects on the JSE (e.g., reverse, eliminate, decrease, increase), which could result in cancellation effects and account for the lack of statistical difference between the size of the JSE under control and non-control conditions. Another plausible

[^364]interpretation is that the JSE is sufficiently robust that there is some leeway in what one can do experimentally and still elicit the JSE. ${ }^{7}$

As a next step, we classified experimental conditions anticipated to eliminate or reverse the JSE as 'wipeout' conditions, and used a meta-regression model to assess their potential moderating effect on the size of the JSE. As anticipated, the summary effect size of the wipeout conditions was significantly smaller than the non-wipeout conditions. However, we wish to add a note of caution when interpreting this analysis. Our coding was based on the original authors' predictions, which we assume to represent a priori hypotheses, but it is possible some were made $a$ posteriori, reflecting post hoc justifications for the findings (Kerr, 1998; Lohse, Buchanan, \& Miller, 2016). An important message to convey to authors is to ensure they are transparent about whether their hypotheses are a priori or a posteriori. In the case that a hypothesis is generated based on theory or prior research, then they should be clear to outline why they believed a condition would eliminate or reverse the JSE. Alternatively, if after data collection potential explanations for what has been found are devised, then authors should be upfront about this. While a posteriori hypotheses tend to be looked at less favourably, they do offer a springboard to test other methods or experimental designs. Nevertheless, the current results confirm that the JSE is sensitive to manipulations 'designed' to diminish its presence.

As the first quantitative description of the joint Simon literature, a clear future direction would be to metaanalytically capture the studies not included here, for example with respect to special populations (e.g., de la Asuncion, Bervoets, Morrens, Sabbe, \& De Bruijn, 2015; Liepelt et al., 2012). Additionally, and particularly in light of the asymmetry present in the current data, subsequent researchers could attempt to solicit unpublished 'file drawer' data, which might help to counter the observed positivity bias, and provide a more accurate picture of the conducted research and estimate of the underlying effect. Also missing from the present analyses are studies not reporting enough data to calculate their associated effect sizes (e.g., no error bars on figures, or not specifying what measure the error bars represent). As such, we urge researchers and reviewers to be diligent towards the reporting of all results, to avoid such omissions in the future.

As a final note, we encourage researchers who are designing an experiment to investigate the JSE to perform (and report) an a priori power analysis (Cumming, 2012; Lohse et al., 2016). A shortcoming of some joint Simon studies is the inadequate sample size. Indeed, across all the

[^365]studies included in this meta-analysis, not one reported estimating sample size. The effect sizes presented in this paper could be used to conduct a power analysis, and this simple procedure will help ensure that the JSE that is (or is not) being detected is a real effect. Since the joint Simon task is commonly used to explore joint action and corepresentation, it is of great import to establish that the observed effect is appropriately powered if we are to infer its underlying mechanisms and influence on behaviour.

## References

## (* indicates manuscript included in the meta-analysis)

Anelli, F. (2012). Social cognition: New insights from affordance and Simon effects. Doctoral dissertation, Department of Philosophy, University of Bologna.
Borenstein, M., Hedges, L. V., Higgins, J. P. T., \& Rothstein, H. R. (2011). Introduction to Meta-Analysis. Hoboken, NJ: John Wiley \& Sons.
Callan, J., Klisz, D., \& Parsons, O. A. (1974). Strength of auditory stimulus-response compatibility as a function of task complexity. Journal of Experimental Psychology, 102(6), 1039-1045.
*Colzato, L. S., de Bruijn, E. R., \& Hommel, B. (2012). Up to "me" or up to "us"? The impact of self-construal priming on cognitive self-other integration. Frontiers in Psychology, 3, 1-4.
*Colzato, L. S., van den Wildenberg, W. P. M., \& Hommel, B. (2013). Increasing self-other integration through divergent thinking. Psychonomic Bulletin \& Review, 20(5), 1011-1016.
*Colzato, L. S., Zech, H., Hommel, B., Verdonschot, R., van den Wildenberg, W. P. M., \& Hsieh, S. (2012). Loving-kindness brings loving-kindness: The impact of Buddhism on cognitive self-other integration. Psychonomic Bulletin \& Review, 19(3), 541-545.
Cooper, H., Hedges, L. V., \& Valentine, J. C. (2009). The handbook of research synthesis and meta-analysis. New York, NY: Russell Sage Foundation.
*Costantini, M., Di Vacri, A., Chiarelli, A. M., Ferri, F., Romani, G. L., \& Merla, A. (2013). Studying social cognition using near-infrared spectroscopy: The case of social Simon effect. Journal of Biomedical Optics, 18(2), 1-6.
*Costantini, M., \& Ferri, F. (2013). Action corepresentation and social exclusion. Experimental Brain Research, 227(1), 85-92.
Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York, NY: Routledge.
*de la Asuncion, J., Bervoets, C., Morrens, M., Sabbe, B., \& De Bruijn, E. R. A. (2015). EEG correlates of impaired self-other integration during joint-task performance in schizophrenia. Social Cognitive and Affective Neuroscience, 10(10), 1365-1372.
*Dittrich, K., Dolk, T., Rothe-Wulf, A., Klauer, K. C., \& Prinz, W. (2013). Keys and seats: Spatial response coding
underlying the joint spatial compatibility effect. Attention, Perception, \& Psychophysics, 75(8), 1725-1736.
*Dittrich, K., Rothe, A., \& Klauer, K. C. (2012). Increased spatial salience in the social Simon task: A responsecoding account of spatial compatibility effects. Attention, Perception, \& Psychophysics, 74(5), 911-929.
*Dolk, T., Hommel, B., Colzato, L. S., Schütz-Bosbach, S., Prinz, W., \& Liepelt, R. (2011). How "social" is the social Simon effect? Frontiers in Psychology, 2, 1-9.
Dolk, T., Hommel, B., Colzato, L. S., Schütz-Bosbach, S., Prinz, W., \& Liepelt, R. (2014). The joint Simon effect: A review and theoretical integration. Frontiers in Psychology, 5, 1-10.
Dolk, T., Hommel, B., Prinz, W., \& Liepelt, R. (2013). The (not so) social Simon effect: A referential coding account. Journal of Experimental Psychology: Human Perception and Performance, 39(5), 1248-1260.
*Dolk, T., Liepelt, R., Prinz, W., \& Fiehler, K. (2013). Visual experience determines the use of external reference frames in joint action control. PLoS ONE, 8(3), 1-8.
*Dolk, T., Liepelt, R., Villringer, A., Prinz, W., \& Ragert, P. (2012). Morphometric gray matter differences of the medial frontal cortex influence the social Simon effect. Neuroimage, 61(4), 1249-1254.
Egger, M., Davey Smith, G., Schneider, M., \& Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. $B M J, 315(7109)$, 629-634.
*Ferraro, L., Iani, C., Mariani, M., Milanese, N., \& Rubichi, S. (2011). Facilitation and interference components in the joint Simon task. Experimental Brain Research, 211(3-4), 337-343.
*Guagnano, D., Rusconi, E., \& Umiltà, C. A. (2010). Sharing a task or sharing space? On the effect of the confederate in action coding in a detection task. Cognition, 114(3), 348-355.
*Hommel, B., Colzato, L. S., \& van den Wildenberg, W. P. (2009). How social are task representations? Psychological Science, 20(7), 794-798.
*Iani, C., Anelli, F., Nicoletti, R., Arcuri, L., \& Rubichi, S. (2011). The role of group membership on the modulation of joint action. Experimental Brain Research, 211(3-4), 439-445.
*Iani, C., Anelli, F., Nicoletti, R., \& Rubichi, S. (2014). The carry-over effect of competition in task-sharing: Evidence from the joint Simon task. PLoS ONE, 9(6), 1-8.
Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review, 2(3), 196-217.
*Kiernan, D., Ray, M., \& Welsh, T. N. (2012). Inverting the joint Simon effect by intention. Psychonomic Bulletin \& Review, 19(5), 914-920.
*Lam, M. Y. (2013). Modulation of joint action correspondence effects by task context: Examination of the contributions of social, spatial, and response discrimination factors. Doctoral dissertation, School of Kinesiology, University of British Columbia.
*Lam, M. Y., \& Chua, R. (2010). Influence of stimulusresponse assignment on the joint-action correspondence effect. Psychological Research, 74(5), 476-480.
*Liepelt, R. (2014). Interacting hands: The role of attention for the joint Simon effect. Frontiers in Psychology, 5, 113.
*Liepelt, R., Schneider, J. C., Aichert, D. S., Wöstmann, N., Dehning, S., Möller, H., ... Ettinger, U. (2012). Action blind: Disturbed self-other integration in schizophrenia. Neuropsychologia, 50(14), 3775-3780.
Lohse, K., Buchanan, T., \& Miller, M. (2016). Underpowered and overworked: Problems with data analysis in motor learning studies. Journal of Motor Learning and Development, 4(1), 37-58.
Lu, C.-H., \& Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. Psychonomic Bulletin \& Review, 2(2), 174-207.
*Malone, M., Castillo, R. D., Kloos, H., Holden, J. G., \& Richardson, M. J. (2014). Dynamic structure of jointaction stimulus-response activity. PLoS ONE, 9(2), 1-7.
*McClung, J. S., Jentzsch, I., \& Reicher, S. D. (2013). Group membership affects spontaneous mental representation: Failure to represent the out-group in a joint action task. PLoS ONE, 8(11), 1-9.
*Milanese, N., Iani, C., \& Rubichi, S. (2010). Shared learning shapes human performance: Transfer effects in task sharing. Cognition, 116(1), 15-22.
*Milanese, N., Iani, C., Sebanz, N., \& Rubichi, S. (2011). Contextual determinants of the social-transfer-of-learning effect. Experimental Brain Research, 211(3-4), 415-422.
Müller, B. C. N. (2013). Social moderators of action corepresentation. Doctoral dissertation, Behavioural Science Institute, Radboud University Nijmegen.
*Müller, B. C., Brass, M., Kühn, S., Tsai, C.-C., Nieuwboer, W., Dijksterhuis, A., \& van Baaren, R. B. (2011). When Pinocchio acts like a human, a wooden hand becomes embodied. Action co-representation for non-biological agents. Neuropsychologia, 49(5), 1373-1377.
*Müller, B. C., Kühn, S., van Baaren, R. B., Dotsch, R., Brass, M., \& Dijksterhuis, A. (2011). Perspective taking eliminates differences in co-representation of out-group members' actions. Experimental Brain Research, 211(34), 423-428.
*Ruys, K. I., \& Aarts, H. (2010). When competition merges people's behavior: Interdependency activates shared action representations. Journal of Experimental Social Psychology, 46(6), 1130-1133.
*Sebanz, N., Knoblich, G., \& Prinz, W. (2003). Representing others' actions: Just like one's own? Cognition, 88(3), B11-B21.
Sellaro, R. (2013). How does task sharing influence individual's performance? An investigation with interference paradigms. Doctoral dissertation, Doctoral School in Brain and Cognitive Sciences, University of Trento.
*Sellaro, R., Hommel, B., Paccani, C. R., \& Colzato, L. S. (2015). With peppermints you're not my prince: Aroma modulates self-other integration. Attention, Perception, \& Psychophysics, 77(8), 2817-2825.
*Sellaro, R., Treccani, B., Rubichi, S., \& Cubelli, R. (2013). When co-action eliminates the Simon effect: Disentangling the impact of co-actor's presence and task sharing on joint-task performance. Frontiers in Psychology, 4, 1-15.
Simon, J. R. (1969). Reactions towards the source of stimulation. Journal of Experimental Psychology, 81(1), 174-176.
*Stenzel, A., Chinellato, E., Bou, M. A. T., del Pobil, Á. P., Lappe, M., \& Liepelt, R. (2012). When humanoid robots become human-like interaction partners: Corepresentation of robotic actions. Journal of Experimental Psychology: Human Perception and Performance, 38(5), 1073-1077.
*Stenzel, A., Dolk, T., Colzato, L. S., Sellaro, R., Hommel, B., \& Liepelt, R. (2014). The joint Simon effect depends on perceived agency, but not intentionality, of the alternative action. Frontiers in Human Neuroscience, 8, 1-10.
*Tsai, C.-C., \& Brass, M. (2007). Does the human motor system simulate Pinocchio's actions? Coacting with a human hand versus a wooden hand in a dyadic interaction. Psychological Science, 18(12), 1058-1062.
*Tsai, C.-C., Kuo, W. J., Hung, D. L., \& Tzeng, O. J. (2008). Action co-representation is tuned to other humans. Journal of Cognitive Neuroscience, 20(11), 2015-2024.
*Tsai, C.-C., Kuo, W. J., Jing, J. T., Hung, D. L., \& Tzeng, O. J. (2006). A common coding framework in self-other interaction: Evidence from joint action task. Experimental Brain Research, 175(2), 353-362.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48.
*Welsh, T. N. (2009). When $1+1=1$ : The unification of independent actors revealed through joint Simon effects in crossed and uncrossed effector conditions. Human Movement Science, 28(6), 726-737.
*Welsh, T. N., Higgins, L., Ray, M., \& Weeks, D. J. (2007). Seeing vs. believing: Is believing sufficient to activate the processes of response co-representation? Human Movement Science, 26(6), 853-866.
*Welsh, T. N., Kiernan, D., Neyedli, H. F., Ray, M., Pratt, J., Potruff, A., \& Weeks, D. J. (2013). Joint Simon effects in extrapersonal space. Journal of Motor Behavior, 45(1), 1-5.
Wenke, D., Atmaca, S., Holländer, A., Liepelt, R., Baess, P., \& Prinz, W. (2011). What is shared in joint action? Issues of co-representation, response conflict, and agent identification. Review of Philosophy and Psychology, 2(2), 147-172.

# A Toolbox of Methods for Probabilistic Inference 

Charles Kemp (ckemp@cmu . edu)<br>Department of Psychology, Carnegie Mellon University<br>Pittsburgh, PA 15213 USA<br>Caleb Eddy (cweddy@andrew. cmu . edu)<br>Department of Psychology, Carnegie Mellon University<br>Pittsburgh, PA 15213 USA


#### Abstract

We propose that probabilistic inference is supported by a mental toolbox that includes sampling and symmetry-based reasoning in addition to several other methods. To flesh out this claim we consider a spatial reasoning task and describe a number of different methods for solving the task. Several recent process-level accounts of probabilistic inference have focused on sampling, but we present an experiment that suggests that sampling alone does not adequately capture people's inferences about our task.


Keywords: probability judgment; probability estimation; reasoning; sampling; symmetry

Certainty is often unattainable, and people must therefore maintain degrees of belief. A prominent tradition in cognitive science explores where these degrees of belief come from and how they are updated given evidence. One line of work focuses on normative accounts of reasoning under uncertainty, and many of these accounts rely on probability theory. A distinct but related line of work focuses on process-level accounts that attempt to characterize how probabilistic inference is implemented by the mind and brain.

Recent work on process-level accounts has emphasized the idea that the mind approximates probabilistic inference by sampling (Griffiths, Vul, \& Sanborn, 2012; Sanborn \& Chater, 2016; Bonawitz, Denison, Griffiths, \& Gopnik, 2014). We believe, however, that sampling is just one among many methods that people use for probabilistic inference. Other possible methods depend on symmetrybased reasoning (Strevens, 1998; Vasudevan, 2012), counting events (Johnson-Laird, Legrenzi, Girotto, Legrenzi, \& Caverni, 1999; Fox \& Levav, 2004), computing sums (Fischbein, 1975), products (Fischbein, 1975) and ratios (Zhao, Shah, \& Osherson, 2009), and ignoring irrelevant information (Grove \& Koller, 1991). This paper lays out an initial proposal about a mental toolbox of such methods. The long term challenge is to understand which methods belong in the toolbox, when they are applied, and how they are flexibly combined. Addressing this challenge is far from straightforward, but essential in order to understand probabilistic inference at the process level.

A longstanding debate in the reasoning literature pits model-based approaches against those that rely on mental proofs. Model-based inference relies on representations of concrete states of affairs, and mental proofs are constructed by applying abstract rules. We believe that both approaches have their merits, and that people draw on both in different
contexts. Our toolbox of methods therefore includes modelbased approaches such as sampling alongside alternatives that require the construction of mental proofs. A pluralist approach, of course, does not immediately resolve the issues at stake in the debate about models and proofs. Detailed work is needed to establish when people rely on model-based approaches and when they construct mental proofs.

## Spatial reasoning task

Because different tasks may elicit different reasoning methods, a comprehensive theory of probabilistic inference should be able to account for a wide range of tasks. As a starting point, we focus here on one simple task. Figure 1a shows a Tshaped rock in a square pond. Suppose that a blue beetle and a gold beetle are both located somewhere on the rock. If the blue beetle is north of the gold beetle, what is the probability that the blue beetle is also east of the gold beetle?

The inferences we consider can be formalized using the graphical model in Figure 1f. Variable $T$ specifies the topography of the pond, and $z_{b}=\left(x_{b}, y_{b}\right)$ and $z_{g}=\left(x_{g}, y_{g}\right)$ indicate the positions of the blue and gold beetles respectively. These positions depend on $T$ because both must fall on a rock rather than in the water. Variables $r_{x}$ and $r_{y}$ indicate relations between the beetles along the $x$ and $y$ axes respectively. For example, $r_{x}=1$ indicates that blue is east of gold, $r_{x}=-1$ indicates that blue is west of gold, and $r_{x}=0$ captures the rare case in which neither beetle is east of the other.

In this setting, a model is a pair $\left(z_{b}, z_{g}\right)$ that specifies the locations of both beetles. One example is shown in Figure 1 b . Our task is deliberately chosen so that it is impossible to enumerate all possible models. In contrast, some previous research on probabilistic inference focuses on problems for which the set of models is discrete and relatively small, which allows inference methods that depend on enumeration or counting (Fox \& Levav, 2004).

Our task has several other appealing properties. It is closely related to a family of tasks known as three-term series problems that have been prominent in the reasoning literature (Clark, 1969; Jahn, Knauff, \& Johnson-Laird, 2007). One such problem asks "if A is west of B and B is west of C , is A west of C ?" Compared to these problems, one advantage of our task is that it allows for a wide range of normative responses. For example, the normative responses to the questions in Figures 1a and 1e are 0.5 and 0.75 respectively, and by varying the shape of the rock it is possible to create a ques-
(a)

(b)

(c)

(d)

(e)

(f)


The blue beetle is north of the gold beetle. How likely is it that the blue beetle is east of the gold beetle?

Figure 1: (a) Our task requires participants to reason about the relative locations of two beetles in a pond. The rock surface is shown in grey, and both beetles are located somewhere on the rock. (b) A complete model that specifies the locations of both beetles. (c) A partial model that specifies the location of the gold beetle only. (d) The normative answer to the canonical question below the ponds can be computed by dividing the area marked with horizontal lines by the area marked with vertical lines. (e) A pond that produces a normative response of 0.75 . (f) Graphical model showing the relationships between variables described in the text.
tion with any desired probability as the normative response. A second advantage of our task is that it admits a range of variants that can potentially provide insight into probabilistic inference. One such variant is to ask the same question but to display the position of one beetle, as shown in Figure 1c. Finally, the next section illustrates that our task is useful for exploring probabilistic inference because it can be solved in principle by several methods.

## A toolbox of probability estimation methods

We now describe a toolbox that contains eight methods for estimating a conditional probability $P\left(r_{x} \mid r_{y}\right)$. This probability corresponds to the strength of an argument in which the premise is $r_{y}$ (e.g. "blue is north of gold") and the conclusion is $r_{x}$ (e.g. "blue is east of gold"). To simplify our notation we treat the topography $T$ as background knowledge and drop it from our equations.

Each method is intended to represent a family of related approaches rather than a single precisely-defined algorithm. After introducing each method, we describe one concrete instantiation of the method, but other instantiations of each method are possible.

1. Sample complete models. One way to estimate the probability $P\left(r_{x} \mid r_{y}\right)$ is to think of a number of models that make the premise true, and to consider how many of these models also make the conclusion true (Johnson-Laird et al., 1999). We refer to this approach as complete sampling, because each model considered provides a complete specification of the positions of the beetles. A normative version of complete sampling is:

$$
\begin{align*}
P\left(r_{x} \mid r_{y}\right) & =\int_{z_{b}, z_{g}} P\left(r_{x} \mid z_{b}, z_{g}\right) P\left(z_{b}, z_{g} \mid r_{y}\right) d z_{b} d z_{g} \\
& \approx \frac{1}{m} \sum_{i=1}^{m} P\left(r_{x} \mid z_{b}^{i}, z_{g}^{i}\right) \tag{1}
\end{align*}
$$

where each pair $\left(z_{b}^{i}, z_{g}^{i}\right)$ is a sample from $P\left(z_{b}, z_{g} \mid r_{y}\right)$. Equation 1 shows how sampling $m$ models is a way to approximate
an integral over all possible locations of the beetles, and the approach is normative in the sense that the approximation approaches the probability $P\left(r_{x} \mid r_{y}\right)$ as the number of samples becomes large.
2. Sample partial models. An alternative to complete sampling is to work with partial models such as the example in Figure 1c that specify the location of one beetle only. For example, a reasoner might imagine several possible locations of the gold beetle, and assess the probability of the conclusion in each case. A normative version of this method is:

$$
\begin{align*}
P\left(r_{x} \mid r_{y}\right) & =\int_{z_{g}} P\left(r_{x} \mid r_{y}, z_{g}\right) P\left(z_{g} \mid r_{y}\right) d z_{g} \\
& \approx \frac{1}{m} \sum_{i=1}^{m} P\left(r_{x} \mid r_{y}, z_{g}^{i}\right) \tag{2}
\end{align*}
$$

where each $z_{g}^{i}$ is a sample from $P\left(z_{g} \mid r_{y}\right)$. In Equation 1, the term $P\left(r_{x} \mid z_{b}^{i}, z_{g}^{i}\right)$ is either 1 or 0 , and can be computed by inspecting whether location $z_{b}^{i}$ lies to the east or the west of $z_{g}^{i}$. The analogous term in Equation 2 is $P\left(r_{x} \mid r_{y}, z_{g}^{i}\right)$, which can be computed using the ratio in Equation 3 below, or one of the other methods in the toolbox.

A premise such as "blue is north of gold" locates a figure object (blue beetle) with respect to a ground object (gold beetle), and Equation 2 could be used by reasoners who focus on the ground object. Another approach is to sample possible locations of the figure object. This approach can be formalized using a variant of Equation 2 in which $z_{g}$ is replaced by $z_{b}$.
3. Construct a model-based proof. If the process of sampling models (complete or partial) is accessible to awareness, then reflecting on this process may be enough to derive some conditional probabilities. Suppose for example that a reasoner samples the mental model shown in Figure 1ba model in which blue is north of gold (as required by the premise) and in which blue is east of gold (as stated by the conclusion). An alert reasoner may notice that this model can be paired with a twin that is identical except that the posi-
tions of blue and gold are reflected about the rock's axis of symmetry. In the twin model the premise still holds but the conclusion does not. Further reflection may establish the conviction that each model can be paired with a twin in this way, which means that every model that supports the conclusion is paired with a twin that rejects the conclusion. The predictions based on the models and their twins therefore "cancel out," revealing that $P\left(r_{x} \mid r_{y}\right)=0.5$. The overall chain of reasoning can be formalized as a mathematical proof that refers to models chosen "without loss of generality."
4. Exploit symmetry. The proof sketched in the previous section has two distinctive characteristics: it refers to specific models and it makes use of symmetry. Symmetry, however, can also be used to derive probabilities without needing to consider any specific models. In Figure 1a, a reflection in the T-shaped rock's axis of symmetry maps east onto west and vice versa, but leaves the shape of the rock unchanged. As a result, inverting east and west in any probability statement concerning the rock leaves the probability unchanged. For example, $P$ (blue east of gold|blue north of gold) must equal $P$ (blue west of gold|blue north of gold), and because these two probabilities sum to one both must equal 0.5 .

Symmetry can also be used to derive an unconditional probability such as $P($ blue east of gold $)=0.5$. It is vanishingly improbable that the beetles have identical x coordinates, which means that blue is either east or west of gold. Given that no available information distinguishes between these states, the principle of indifference (Strevens, 1998) implies that both must have a probability of 0.5 .
5. Ignore irrelevant information. A basic strategy for simplifying probabilistic inference is to ignore information that has no bearing on the conclusion. If the pond contains an upright square rock, for example, the $x$ and $y$ coordinates of a beetle are statistically independent-knowing one of these coordinates places no constraints on the other. It follows that $P\left(r_{x} \mid r_{y}\right)=P\left(r_{x}\right)=0.5$, where the final step follows from the principle of indifference as described in the previous section.
6. Apply the ratio rule. Suppose that $z_{g}$ (the position of the gold beetle) is known, as shown in Figure 1c. The conditional probability $P\left(r_{x} \mid r_{y}, z_{g}\right)$ can be computed using

$$
\begin{equation*}
P\left(r_{x} \mid r_{y}, z_{g}\right)=\frac{P\left(r_{x}, r_{y} \mid z_{g}\right)}{P\left(r_{y} \mid z_{g}\right)} \tag{3}
\end{equation*}
$$

Equation 3 is simple to compute by estimating the area of two regions in a diagram like Figure 1c. The denominator $P\left(r_{y} \mid z_{g}\right)$ is proportional to the area of rock that is north of $z_{g}$ (indicated with vertical lines in Figure 1d). The numerator $P\left(r_{x}, r_{y} \mid z_{g}\right)$ is proportional to the area that is north and east of $z_{g}$ (indicated with horizontal lines in Figure 1d).
7. Apply Bayes rule. Bayes rule can be applied as follows:

$$
\begin{equation*}
P\left(r_{x} \mid r_{y}\right)=\frac{P\left(r_{y} \mid r_{x}\right) P\left(r_{x}\right)}{P\left(r_{y}\right)}=P\left(r_{y} \mid r_{x}\right) \tag{4}
\end{equation*}
$$

where the final step follows from the observation above that $P\left(r_{y}\right)=P\left(r_{x}\right)=0.5$. In general $P\left(r_{y} \mid r_{x}\right)$ will be no easier to compute than $P\left(r_{x} \mid r_{y}\right)$, so applying Bayes rule may not be useful. There may be cases, however, in which one of these probabilities is easier to compute than the other.
8. Enumerate cases. One general strategy for solving a difficult problem is to break it down into a set of simpler subproblems. In Figure 1e, a reasoner may estimate $P\left(r_{x} \mid r_{y}\right)$ by considering 4 cases: either both beetles are on the bottom left rock, both are on the top right rock, blue is bottom left and gold is top right, or blue is top right and gold is bottom left. This strategy can be captured formally by introducing a variable $v$ that indicates which of the 4 cases obtains:

$$
\begin{equation*}
P\left(r_{x} \mid r_{y}\right)=\sum_{v} P\left(r_{x} \mid v, r_{y}\right) P\left(v \mid r_{y}\right) \tag{5}
\end{equation*}
$$

Each of the sub-problems is simpler than the original. For example, if both beetles are on the same rock, then $P\left(r_{x} \mid r_{y}\right)=$ 0.5 , as argued in our discussion of method 5 .

Using the toolbox. We suspect that all eight methods in the toolbox and possibly others are available to human reasoners. Given a problem, a reasoner must therefore decide which method or methods to try. Sometimes two or more methods will need to be combined: for example, methods 2 and 8 ("sample partial models" and "enumerate cases") both express the original probability as a function of several probabilities, which must be estimated in turn.

At present, a detailed mechanistic understanding of probability estimation seems remote. Establishing that people rely on one method for a given task is difficult, because numerous other methods must be ruled out. Establishing that people do not rely on a given method may be more tractable, because only one hypothesis must be ruled out. Given the recent emphasis on sampling as a mechanism for probabilistic inference, we designed a study to explore whether sampling is a plausible account of inference in our setting.

## Experiment

We suspect that people rely on sampling when other methods are unavailable, but are able to exploit symmetry when relevant. If so, then people's responses to symmetric ponds might be systematically different from their responses to other ponds. Our experiment was designed to test this possibility.
Participants. 36 participants were recruited using Amazon Mechanical Turk and paid for their participation.

Materials. Participants were asked to reason about the 26 ponds shown in Figure 2. The first 19 ponds are categorized as "double symmetry," "single symmetry" or "no symmetry" ponds depending on whether they have both vertical and horizontal symmetry, only one of these symmetries, or neither vertical nor horizontal symmetry. The next 5 ponds are "non50 " ponds, or ponds for which the normative response is other than 50.

## No symmetry

(a)
(b)
(c)
(d)
(e)
(f)
$(\mathrm{g}) \quad(\mathrm{h})$
$\begin{array}{llll}\text { (i) } & \text { (j) } & \text { (k) }\end{array}$
Catch
(y)
Non-50
(z)
(m)
(n)
(o)
(p)
(q)
(r)
(s)
(t)
(u)
(v)
(w)
(x)



Figure 2: Ponds used in the experiment. The normative response for all ponds in the first three categories (no symmetry, single symmetry, and double symmetry) is 50 . Ponds (l), (q), (s) and (x) all have rocks that enclose a body of water.

Procedure. Participants read an introduction that described an eccentric businessman who owned many square ponds. Each pond was said to contain a single gold beetle and a single blue beetle. Participants were told that the beetles could not swim, so each beetle was located somewhere on a rock. They then answered three simple questions that tested their comprehension of what they had just read. They remained on the introductory screen until they had answered all three questions correctly.

Each participant then saw the 24 ponds in Figures 2a-2x in a random order. For each pond, they read that "In this pond the blue beetle is $r 1$ of the gold beetle." They were then asked "How likely is it that the blue beetle is $r 2$ of the gold beetle?", and required to give their answer on a 0-100 scale with labels at 0 ("Not likely") and 100 ("Very likely"). For each pond and each participant, $(r 1, r 2)$ was a pair of perpendicular directions (e.g. (north, east), (north, west)) randomly drawn from the set of 8 such pairs.

After the 24 ponds participants responded to two catch trials that had unambiguous answers. One stated that "the blue beetle is east of the gold beetle" and asked participants to rate the likelihood that the blue beetle is west of the gold beetle. The second was similar but used the north-south instead of the east-west axis. The rocks used for these questions are shown in Figures 2y and 2z.
Results. We computed normative responses for each pond by assuming that the location of each beetle was generated from a uniform distribution over the rock surface. For single symmetry and double symmetry ponds, the normative response is always 50 . Normative responses for the non-50 and no symmetry ponds were computed by using complete sampling and drawing 100,000 samples. When creating the no symmetry ponds, the dimensions of the ponds (e.g. the relative lengths of the two T-segments in Figure 2a) were adjusted until complete sampling returned a normative result between 49.5 and 50.5.

Because the question associated with each pond was randomized, all responses were con-


Figure 3: (a) Mean human responses versus normative responses (b) Histogram of correlations achieved by individuals
verted to responses to the canonical question $P$ (blue north of gold|blue east of gold). Our conversion assumed that $P$ (blue north of gold|blue east of gold) $=$ $1-P$ (blue south of gold|blue east of gold), and similarly for other pairs of opposite directions. We also assumed that $P\left(r_{x} \mid r_{y}\right)=P\left(r_{y} \mid r_{x}\right)$, as discussed in Method 7 above. We expect that intuitive judgments do not always respect or even approximate this latter identity, but assuming that they do allows for a simple first look at our data.

16 participants failed to give ratings of 0 on both catch trials, and were dropped from all subsequent analyses. Figure 3a shows that mean responses among those who remained roughly tracked normative responses. Each point in the scatter plot corresponds to a pond. For example, the point at the top right of the plot corresponds to Figure 2t. The overall correlation between human and normative responses is 0.83 , and Figure 3b shows the correlations achieved by individual participants. Some participants had correlations near zero, but half had correlations exceeding 0.5. Overall, Figure 3 suggests that humans perform relatively well at the task.

The comparison of primary interest is between no sym-


Figure 4: Mean distance from 50 for no symmetry, single symmetry and double symmetry ponds. Results are shown for (a) experimental data (b) complete sampling model with $m=8$ (c) partial sampling model with $m=3$. These $m$ values were chosen to approximately match the variability in the human data.
metry ponds, single symmetry ponds and double symmetry ponds. The normative response for these ponds is always 50 , and we therefore analyzed the extent to which responses differed from 50. Figure 4a shows that responses for the double symmetry ponds tended to be closer to 50 than responses to the other two kinds of ponds. A Mann-Whitney test indicated that the distance from 50 was greater for no-symmetry ponds (median $=16, \mathrm{n}=120$ ) than for single-symmetry ponds (median $=13, \mathrm{n}=120$ ), $\mathrm{U}=5998, p=0.012$. A second test indicated that the difference between single-symmetry ponds and double-symmetry ponds (median $=0, \mathrm{n}=140$ ) was also statistically significant ( $\mathrm{U}=5260, p<0.001$ ). A natural interpretation of these results is that some participants relied on symmetry-based reasoning.

Complete and partial sampling can both be implemented in different ways, but the implementations suggested by Equations 1 and 2 are especially appealing. These implementations are relatively simple, and both approximate the normative response as the number of samples becomes large. Figures 4 b and Figures 4 c show results for these two implementations. In both cases, the number of samples is chosen so that the model matches the average distance from 50 in the human data. Although matched to humans in this respect, the two sampling models do not account for the special status of the double symmetry ponds in the human data. For example, the complete sampling model predicts no difference between the no symmetry and double symmetry ponds.

A second challenge for a sampling model is whether it can account for the human data given a psychologically plausible number of samples. For the sake of argument, assume that each of our participants is using complete sampling, and that each draws the same number of samples $m$ in Equation 1. Figure 5 shows how the predicted variability in the human data decreases as $m$ increases. If each participant drew one sample only, then some would give responses of 0 and others would give responses of 100 , and the average distance from 50 would be 50 for no symmetry, single symmetry and double symmetry ponds alike. If $m$ were very large, then each participant would give a response very close to 50 . Figure 5a shows that setting $m$ to 5 or 6 is enough to account for the variability in responses to the no symmetry and single symmetry ponds.


Figure 5: Distance from 50 predicted by (a) complete and (b) partial sampling as the number of samples increases. In (a) a single model curve is shown in black because model predictions are identical for no symmetry, single symmetry and double symmetry ponds. In (b) three model curves are shown because the model predictions for the three classes of ponds are close but not identical.

For double symmetry ponds, however, $m$ must be set higher than 20 in order to match the human data. A value this high does not seem psychologically plausible, and challenges the hypothesis that people rely on complete sampling when reasoning about double symmetry ponds.

Figure 5b shows the analogous plot for partial sampling. In this case, setting $m$ to 10 or so is enough to account for the variability in responses to the double symmetry ponds. This value seems high, but perhaps not high enough to definitively rule out partial sampling as a psychological account. The dif-
ference in the human data between double symmetry and both single and no symmetry ponds, however, remains a challenge for models that rely on partial sampling.

Discussion. Although our implementations of complete and partial sampling do not account well for our data, it is possible that other implementations of these methods will perform better. Our implementations assume that models are randomly sampled from the set of all models consistent with the premise of a given argument, but in reality people may sample some kinds of models more often than others. For example, perhaps people prefer to locate the beetle mentioned first towards the left of the pond (Jahn et al., 2007) or towards the top (Levelt \& Maasen, 1981). Previous accounts of spatial reasoning have documented effects like these (Jahn et al., 2007), and it seems likely that similar effects will emerge in our setting.

In addition to left-right and up-down preferences, people may also prefer to sample models in which the beetles are located along axes of symmetry. A preference of this kind could help to explain results that are also consistent with symmetrybased reasoning. In Figure 1a, for example, a partial sampling method that uses just one sample will generate the normative response of 50 provided that the single sample locates the gold beetle along the rock's axis of symmetry.

Throughout we have mostly considered inference methods that compute or approximate normative responses. Our data suggest that people's responses to our task are roughly consistent with normative inference, but in other settings people make inferences that are far from normative. For example, base-rate neglect may occur if people apply Bayes rule without including the prior (Kahneman \& Tversky, 1973). In other cases people may rely on sampling but sample from the "wrong" distribution-for example, some of our participants may have sampled from $P\left(z_{g}\right)$ rather than $P\left(z_{g} \mid r_{y}\right)$ in Equation 2. Each method in the toolbox can be applied in normative and non-normative ways, and detailed work is required to understand how a method is applied in any given setting.

## Conclusion

We suggested that people make use of a mental toolbox that includes several qualitatively different methods for probabilistic inference. Each of these methods has several variants, and some methods can be combined with each other. We therefore believe that people can draw on a set of inference methods that is relatively large, which makes understanding probabilistic inference at the process level very challenging indeed.

Like previous researchers we believe that behavioral experiments can provide some insight into the processes that support probabilistic inference. We described a spatial reasoning task that appears to be a natural candidate for inference by sampling, but our results suggest that any simple sampling method is unlikely to fully capture the way in which people approach the task. Ruling out one simple hypothesis about inference is one thing, but providing a comprehensive account
of probabilistic inference is another thing entirely. We confess to some scepticism about whether behavioral data alone are enough to reveal the mind's algorithms for probabilistic inference.

## Acknowledgments

Caleb Eddy was supported by a fellowship in computational neuroscience that was awarded by the Center for the Neural Basis of Cognition at CMU and the University of Pittsburgh and funded by NIH R90 DA023426.

## References

Bonawitz, E., Denison, S., Griffiths, T. L., \& Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. Trends in Cognitive Sciences, 18(10), 497-500.
Clark, H. (1969). Linguistic processes in deductive reasoning. Psychological Review, 76(4), 387.
Fischbein, H. (1975). The intuitive sources of probabilistic thinking in children. D. Reidel Publishing Company.
Fox, C. R., \& Levav, J. (2004). Partition-edit-count: Naive extensional reasoning in judgment of conditional probability. Journal of Experimental Psychology: General, 133(4), 626-642.
Griffiths, T. L., Vul, E., \& Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. Current Directions in Psychological Science, 21(4), 263268.

Grove, A. J., \& Koller, D. (1991). Probability estimation in face of irrelevant information. In Proceedings of the seventh conference on Uncertainty in Artificial Intelligence (pp. 127-134).
Jahn, G., Knauff, M., \& Johnson-Laird, P. (2007). Preferred mental models in reasoning about spatial relations. Memory \& Cognition, 35(8), 2075-2087.
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., \& Caverni, J. (1999). Naive probability: A mental model theory of extensional reasoning. Psychological Review, 106(1), 62-88.
Kahneman, D., \& Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), 237-251.
Levelt, W., \& Maasen, B. (1981). Lexical search and order of mention in sentence production. In W. Klein \& W. Levelt (Eds.), Crossing the boundaries in linguistics (pp. 221252). D. Reidel.

Sanborn, A. N., \& Chater, N. (2016). Bayesian brains without probabilities. Trends in Cognitive Sciences, 20(12), 883893.

Strevens, M. (1998). Inferring probabilities from symmetries. Noûs, 32(2), 231-246.
Vasudevan, A. (2012). Symmetry and probability. Unpublished doctoral dissertation, Columbia University.
Zhao, J., Shah, A., \& Osherson, D. (2009). On the provenance of judgments of conditional probability. Cognition, 113(1), 26-36.

# The Narrow Conception of Computational Psychology 


#### Abstract

One particularly successful approach to modeling within cognitive science is computational psychology. Computational psychology explores psychological processes by building and testing computational models with human data. In this paper, it is argued that a specific approach to understanding computation, what is called the 'narrow conception', has problematically limited the kinds of models, theories, and explanations that are offered within computational psychology. After raising two problems for the narrow conception, an alternative, 'wide approach' to computational psychology is proposed.


Keywords: narrow conception, individualism, computation, psychology, explanation

## Introduction

Cognitive science has gained a good deal of theoretical and methodological impetus from thinking about how psychological processes can be described, studied, and simulated using different types of models. One particularly successful approach to modeling is computational cognitive modeling or, more simply, computational psychology. Computational psychology explores psychological processes by building and testing computational models with human data (Sun, 2008).

In this paper, it is argued that a specific approach to understanding computation, what is dubbed the 'narrow conception', has problematically limited the kinds of models, theories, and explanations that are sometimes offered within computational psychology.

The impetus for the current study arises from a growing debate around the role, nature and status of computation within psychological investigations. Several authors have begun to re-examine what computationalism stands to offer the cognate disciplines (Piccinini, 2015; Milkowski, 2015). The current discussion stands to contribute to this growing trend by exploring and examining one important assumption that underwrites a notable swath of research within computational psychology. The goal is to show that computational psychology has overlooked an important constraining assumption.

## Computational Psychology

For many, computational theory provides a theoretically flexible and expressively powerful tool for exploring cognition (Anderson 1983; Pylyshyn, 1984; Newell, 1990; Anderson \& Lebiere, 1998, 2003). The computational approach allows researchers to construct detailed accounts of the mechanisms, structures, and processes that underwrite cognition. In testing and extending the theories of other domains, such as cognitive psychology and artificial intelligence, computational investigations offer a functionally viable yet mathematically rigorous way of exploring cognitive or psychological processes.

A good deal of the explanatory value of computational psychology lies not only in the ability to produce computer simulations, but also in using those simulations to make predictions about human data. By matching the 'fit' of human data with computer simulations, researchers establish systematic relationships between computational models and psychological processes, which can reveal the underlying structure and form of cognitive functionalities (Sun \& Ling, 1998).

Consider three illustrative examples of computational psychology in action. First, consider Shiffrin and Steyvers' (1997) REM model of episodic memory. Shiffrin and Steyvers' model is one instance of a class of abstract, computational models that attempt to explain recognition judgments. These models employ a 'global matching' procedure. The global matching procedure produces a familiarity signal that indicates whether or not an item has been previously presented to the model - a test cue, for example, that matches two features of one item will yield a higher familiarity judgment than a test cue that matches one feature of each of the two items.
Shiffrin and Steyvers' model puts a Bayesian twist on the global matching procedure. The REM model calculates the likelihood of whether a cue item matches or corresponds to particular stored memory traces by assigning values to each of the stored items. When the model is tested to see if it can identify whether cue items are new or old, the cues are compared with each trace item in memory such that the model calculates the likelihood of the retrieval cue and the trace item matching. Recognition judgment is explained in terms of a probabilistic familiarity process operating within memory.
According to Shiffrin and Steyvers, the REM model accounts for a number of distinct memory effects. One example is the word frequency mirror effect. The word frequency mirror effect says that subjects often make more false alarms on high-frequency lures (foils) versus lowfrequency lures and more correct "old" responses to lowfrequency targets versus high-frequency target when making recognition judgments (Glanzer et al., 1993). The REM is able to accommodate the word frequency effect in virtue of the fact that low-frequency words have more unusual features than high-frequency words (e.g., more syllables). The REM model is able to use a slightly lower value when generating low frequency lures items during matching, which results in these items having slightly higher feature values.

The relevance of the REM model is that in measuring the fit of the model to behavioral data and by adopting a Bayesian approach to the global matching procedure, the REM model focuses on both the essential interplay between modeling and experimental data and formalizing cognitive processes in a computationally rigorous way.

Next, consider Dienes’ (1992) connectionist model of implicit language learning. Dienes' model attempts to computationally instantiate how language users come to implicitly understand artificial grammars using artificial neural networks. In particular, Dienes' model uses a feedforward autoassociator network.

In a feed-forward autoassociator network - which is a version of the more standard multilayer network - activation passes through the connection weights of the network just once to produce the output activation. The feed-forward autoassociator contrasts with recurrent autoassociator networks, in which the output activation arrives back at each node and is passed through the weights again until a stable state is reached.
In terms of network training, Dienes' model was presented with the same learning material as human subjects, which included arbitrary strings of letters, such as MTTTTV or MTTVT. These features of strings were represented as activations to the network's input layer. Depending on whether the feature was present or absent, the unit coding the feature would have an activation of either 1 or 0 . Once the network learned the arbitrary training strings, similar to experimental tests, the model was then made to make grammaticality judgments on new strings of letters. The goal was to see if the network had learnt the underlying grammatical principles that implicitly structured the arbitrary strings being presented.

When Dienes' tested the model, it was found the network was able to distinguish grammatical versus ungrammatical strings. The network was able to reproduce the training strings by adding or subtracting strings from an exemplar case. The model predicted each feature of a string based on some set of the remaining features from an exemplar. When Dienes' compared network results to that of human subjects, it was found that the network could classify test strings as well as people could. The network tended to reproduce grammatical test strings more faithfully than nongrammatical test strings.
Dienes' connectionist model stands as a further interesting example of computational research, as it provides a computational account of implicit artificial grammar learning that measures the fit of the model with behavioral data. By investigating how artificial neural networks handle artificial grammar tasks, Dienes' attempts to undercover the computational processes and representations underlying implicit language learning.
A third example comes from Osherson et al.'s (1990) declarative model of inductive reasoning. Osherson et al.'s (1990) model attempts to investigate the computational underpinnings of 'inductive' reasoning - inductive reasoning is the process by which premises are thought to lend non-conclusive support to the truth of specific conclusions.
In Osherson et al.' model, inductive reasoning is explained in terms of the assessment of propositional statements according to the similarity between premise and conclusion categories For example, consider two inferential chains: (i)

Mice have property X/All mammals have property X and (ii) Horses have property X/All mammals have property X. The category of 'mammal' in the conclusion covers both mice and horses. For Osherson et al., understanding how humans are able to make inferences about mice and horses depends on understanding how structural relationships between different categories are established - for example, understanding that mice and horses are instances of the subordinate category mammal.
Two features allow Osherson et al.'s model to make sense of cases such as the above. The first is that the model assesses the similarity between premise categories and conclusion categories. The second is that the model measures how well the premise categories covers the superordinate category. Coverage between premise and conclusion categories is assessed in terms of the average similarity of the premise category to members of the superordinate category. For instance, to the extent that horses are more typical mammals than mice, and therefore more similar to other kinds of mammals, (ii) will have greater coverage than (i).
Osherson et al.'s model is interesting because it addresses a number of empirical phenomena. One example is similarity effects. Similarity effects occur when people make inferences based on the perceived similarity between items in different inferential chains. Osherson et al. (2008) found, for example, that when people were given a choice between two syllogistic arguments about $95 \%$ chose the argument that they perceived to contain the greater similarity between premise and conclusion categories, e.g., sparrows to robins and blue jays versus geese to robins and blue jays. Osherson et al.'s model was able to accommodate such cases by assessing the relationship between the subordinate and premise categories.

Similar to the previous models, Osherson et al.'s declarative model is an illustrative example of computational psychology, because it is not only informed by and tested against empirical data, but it also attempts to identify the computational procedures and properties underlying complex cognitive processes, such as inductive reasoning.

The point of the previous survey is that each of the three models provides a paradigmatic example of computational psychology. Each model attempts to undercover the computational underpinnings of various cognitive processes via the construction and testing of computer models with human data. These models help to tease out the underwriting assumptions within computational research.

## The Narrow Conception

With the domain of analysis laid out, the task now is to examine one approach to understanding computation that underlies a good deal of the research within computational psychology, what is labeled the 'narrow conception'.
In order to get a better handle on the narrow conception, consider what Segal (1991) says about computational cognitive systems, he writes: "It seems likely that whole
subjects (or whole brains) make up large, integrated, computational systems...the whole subject is the largest acceptable candidate for the supervenience base because it is the largest integrated system available" (p.492). For Segal, the individual or whole subject (which is plausibly identical to the whole brain) is the largest unit available for computational, psychological investigation. Newell et al. (1989) offer a similar view, writing: "Symbol systems are an interior milieu, protected from the external world, in which information processing in the service of the organism can proceed" (1989, p.107). Here, again, computational systems are limited to the boundary of the individual.

Consider Fodor (1983) next: "Mechanisms of transduction are thus contrasted with computational mechanisms: whereas the latter may perform quite complicated, inference-like transformations - the former are supposed - at least ideally - to preserve the information content of their input" (1983, p. 41). Fodor's contrast between sensory transducers and computational mechanisms is indicative of where he thinks computational systems are located. Computational systems are sandwiched between transducers and motor outputs. Finally, consider what Egan (2000) says on the matter: "A computational theory prescinds from the actual environment because it aims to provide an abstract, and hence completely general, description of a mechanism that affords a basis for predicting and explaining its behaviour" (p.191). Only by abstracting away from the embedding environment and focusing on the individual can one begin to provide successful computational analyses. Once again, the outer limit of formal analysis for computational systems is the individual.

Common to each of these views is the idea that the individual or some sub-module, conceived of in terms of the primary unit of action, constitutes the largest organizational system amendable to computational description (i.e. computational modeling). The individual marks the conceptual boundary for computational, psychological investigations. Here is one way the view might be formulated:

## The Narrow Conception: Computational cognitive

 systems are, and should be studied as if they were, located entirely within the individual or some submodule.Something in the spirit of this claim seems to have operated implicitly within a good swath of computational psychology. The narrow conception, if true, represents a principled claim about where and how computational cognitive systems should be studied. It constitutes a plausible and substantive proposal for computational psychology.
Consider the methodological implications of the narrow conception. If computational systems are wholly interior to the individual, then computational modeling should have as its target only those systems and processes that are
individual-centered. As Segal diagnosis the situation: "Whole subjects plus embedding environments do not make up integrated, computational systems" (1991, p.492). The embedding environment plus individual will always fail to be adequate for computational analysis. Only the individual or some sub-system will be sufficient for computational modeling.

One motivation for adopting the narrow conception is that it provides a powerful way of explaining the causal powers of cognition. If cognitive systems are computational systems and computational systems are located within the individual, then identifying the causal properties and powers of computational systems provides insight into causal power of cognitive processes and abilities. Memory effects, such as primacy and recency affects, for example, will be best explained by focusing on the computational search strategies used by individuals during various tasks (e.g., exhaustive versus terminal search) (Sternberg, 1969). Only by identifying the distinct functional and causal properties intrinsic to the individual are rigorous computational, psychological explanations provided.
What is interesting about the narrow conception, besides its relatively straightforward nature, is that it is plausibly supported by and conforms to a good deal of research within computational psychology. This is why authors such as Segal claim that it is "likely" that the whole subject is the largest unit of analysis. The narrow approach is an empirical wager on how computational cognitive systems are distributed in nature.
Return to the three previous models to see why. First, consider how Shiffrin and Stevyer describe their model: "This cued recall model is meant to illustrate one plausible way in which retrieval from episodic images and retrieval from lexical/semantic images could work hand in hand to allow recall to take place" (1997, p.160). The emphasis on retrieval and storage is indicative of the narrow conception: the computational processes under investigation are localized within the individual. It is only once items are learned and internalized that computational processes can operate over them. The Bayesian matching procedure applies to items stored internally within an individual's episodic memory.

Consider, next, how Dienes' conceives of his model, he writes: "[L]awful behaviour may be produced by a connectionist network in which rules or hypotheses are not explicitly represented" (1992, p.40). A little later he writes: "the subject of the models obeys the rules, but does not represent them symbolically"(1992, p.70). Again, the message is plain. The artificial neural network represents a cognitive system that employs internal representations and rules that solve artificial grammar tasks, and the human data helps to reveal these internal computational processes and structures. The connectionist network is meant to represent the internal computational system within a subject that is used to carry out the cognitive task.
Finally, consider Osherson et al.'s model. In studying inductive reasoning, Osherson et al. adopt the following
position: "The similarity-coverage model assumes that the existence of a pre-established hierarchy of categories that classify the instances figuring in an argument. The success of the model in predicting the qualitative phenomena...testifies to the approximate soundness of the model's assumptions" (1993, p.200). What emerges, again, is a particular interpretation of what has been revealed about the underlying computational system. Reasoning about inference chains is an internal computational process that requires the deployment of particular categorical hierarchies. The boundary of the cognitive system is once again fixed at the formal system detecting relationships between argument stimulus input and subordinate categories.

Each of the three examples conforms, in varying degrees, to the narrow conception. The individual or some subcomponent is the complete and natural unit of computational theorizing. The individual, in each case, is conceived of, and studied as if it were, the largest organized set of components capable of supporting computational investigation.

But notice that in addition to helping researchers to better understand models, the narrow conception also helps to structure the way in which researchers go about identifying and constructing investigations. The narrow conception also offers a means for thinking about where and what to look for when during investigation. It proposes methodological guidelines for studying computational cognitive systems.

Recall, for instance, that each of the three models addressed particular problems, proposed different solutions, and provided different explanations. Shiffrin and Stevyers' model, for instance, conceived of recognition as a problem of item matching. This meant that the computational processes involved searching through memory traces using a global matching procedure. Dienes' model, on the other hand, conceived of implicit learning as a form of pattern recognition. This led to looking for the internal exemplar representations and rules that allowed the network to identify and classify new letter strings. Finally, in Osherson et al.'s study, inference was taken to involve detecting structural category relations. This meant that it attempted to build a model around understanding how such categorical relationships could be structured.

One way to understand why each study offers the types of model it does and measures the fit of its model(s) against the types of experimental data that it does is as a result of the constraining influence of the narrow conception. In directing attention to the individual and its sub-components, the narrow conception sets up certain implicit conceptual boundaries. It limits which computational explanations are seen as viable, which properties and processes are taken to be necessary for investigation, and which solutions are considered plausible. The explanatory space of options surrounding computational theorizing is delimited. The narrow conception curbs the conceptual and methodological understanding of computation available for use within investigations.

## The Wide Conception

The discussion up until this point has been largely descriptive. The goal has been to articulate what the narrow conception amounts to and provide a sense of the way in which it imposes interpretative and methodological constraints on research. In this final section, the aim is to provide a critical analysis of the view. Two problems are raised.
The first problem follows on the heels of the constraining influence of the narrow conception. The issue is that if the narrow conception limits the theoretical and explanatory horizons of computational investigations, then it also limits the kinds of research that can conducted. This is an undesirable state of affairs insofar as a healthy domain of investigation should have the broadest range of alternatives available when conducting research. If researchers are limited in the potential avenues they might explore, then the range of theories, explanations, and models they end up offering may turn out to be impoverished. In an ideal world, there will be as few constraining or biasing assumptions as possible during investigation. Insofar as the narrow conception operates as a constraining assumption on computational psychology, it forms a barrier to conducting successful research.
The history of behaviorism offers an instructive example. In both its logical and philosophical forms, behaviorism eschewed recourse to 'mental' vocabulary. It held that only 'observable behaviour' was the proper subject of psychological investigation. One result of its constraining influence was North American psychology made little reference to mental structures and processes. It took almost 30 years to reclaim the conceptual territory lost to behaviorism (Gardner, 1985). The claim here is not quite so negative, but the moral is the same. The narrow conception has potentially closed off interesting avenues of computational research because of its constraining influence.
One might respond by arguing that the above concern is only a really problem if the narrow conception turns out to be false. But that given the wealth of empirical support the view enjoys, there is really no reason to think that the narrow conception is in fact not the right view to hold. The problem with this response is that gets the order of explanation backwards. It is not that the narrow conception is true because computational research conforms to its strictures. Rather, it is because the narrow conception imposes certain restrictions on research that computational investigations conform to its strictures. The narrow conception problematically limits the range of alternatives considered before, during and after investigation.
The second concern is that the narrow conception, on occasion, provides explanatory weaker accounts of psychological phenomena in virtue of its over emphasis on individual-bound systems. Because the narrow conception emphasizes the individual as the limit of computational explanations, investigations based on its strictures can fail to
identify the important computational role played by environmental elements.
Consider an example from the history of cognitive science. Problem solving was traditionally thought to involve a search through problem space (Newell, Shaw, \& Simon, 1960). One way this approach was computationally instantiated was to simulate agents searching mentally through a virtual problem space during various tasks (Newell \& Simon, 1976). One issue with these early approaches is that cognizers often interactively explore problems by physical manipulating external structures (Kirsh, 2009). These types of actions are more than just pragmatic, as they crucially help cognizers to simplify and transform complex problems. Computational models that focused narrowly on internal searches missed out on the simplifying computational role of epistemic actions (see Wilson, 2004; Clark, 2008).
Insofar as computational explanations fail to pay sufficient attention to elements of the environment that offload and distribute cognitive activities, they stand to provide weaker accounts of psychological phenomena. Computational explanations that are overly reliant on the narrow conception, such as in the above case, can supply explanatorily weaker accounts (Wilson, 2014). This is not to say that every computational explanation that subscribes to the narrow conception is explanatorily weaker. Rather, it is to point out that because there are blind spots imposed by the narrow conception, some computational explanations may, on occasion, be weaker than potential alternatives.

The previous two concerns should not be taken to undermine the narrow conception in its entirety. Instead, the concerns are better understood as forming a negative case against the sufficiency of the narrow conception as a global thesis. Given this, it will be worth exploring a possible alternative approach to understanding computation.

Wide computationalism is the idea that at least some of the elements of computational cognitive systems can reside outside the individual (Wilson, 1994, 1995; Hutchins, 1995; Kersten, 2016; Kersten \& Wilson, 2016). Wide computational systems are those systems that recruit computational units from the larger embedding environment. Similar ideas have also been offered about cognition under the label of 'situated, embedded and extended' cognition (see Wilson, 2004; Clark 2008).

The viability of wide computationalism follows from the location neutrality of computational individuation. Wilson, for example, writes: "There is nothing in the method of computational individuation itself... which implies that the class of physical features mapped by a realization function cannot include members that are part of the environment of the individual" (1994, p.355). Because formal systems are medium neutral, it is at least possible that some of the computational elements include parts outside the individual. Wide computationalism stands in contrast to the narrow conception insofar as it pushes computational analysis outside the individual. Wide computationalism also gains
support from a number of empirical studies in human and animal psychology (see Kersten, 2016).

Wide computationalism is a locational thesis about the realization or supervenience base of computational cognitive systems. It is a view about the scope of physical systems, processes, and components that are capable of supporting computational analysis. What this means is that although wide computationalism is compatible with either an individualist (Segal, 1991) or an externalist (Shagrir, 2001) interpretation, it is, strictly speaking, non-committal on issues of representational or semantic individuation.

For present purposes, the truth of wide computationalism is less important than the alternative it presents. This is because wide computationalism provides one potential alternative for understanding computation within computational psychology. In articulating a conception of computation that moves beyond the individual, wide computationalism stands to supply an importantly distinct approach to understanding computational investigations. By exploiting the location neutrality of computational individuation, wide computationalism re-conceptualizes the study of computational cognitive systems as at least partially requiring analysis of the embedding environment.
Investigations based on this wide approach stand to pay closer attention to the role of the environment, given their explicit focus on computational systems spreading out across the brain, body and world. Examples of the wide conception in action, for example, include agent-based models or swarm behaviour models (see Dawson, 2010). One way to view wide computationalism, then, is as an alternative conception of the underlying concept of computation that may be used within computational psychology.

Another way to make the point is to say that whereas the narrow conception might be construed as a restrictive monistic and a priori assumption about how cognitive states and processes are studied, wide computationalism provides an alternative pluralistic, empirical approach to investigation. Instead of viewing the narrow conception as exhausting the logical space of investigation, wide computationalism might be seen as a further, important additional explanatory strategy that can be used when thinking about computational investigations. Some phenomena may be more amendable to wide investigation, while others may conform more closely to the narrow conception. It may be that in some cases a narrow approach is preferable, while in others a wide approach is more suitable. In opening up the logical space, computational psychology is better positioned to precede both methodologically and theoretically.

This is only the briefest of sketches, but it should begin to provide a sense of how computational psychology may move beyond the narrow conception. However, the wide approach is not offered as a replacement to the narrow conception, but rather as a supplement. Wide computationalism is simply an extension of the logic inherent within computational psychology. The point is that
it can step in when computational investigations run up against the limits of the narrow conception. On the proposed view, research that conforms to the narrow conception, such as the three examples surveyed, still makes a valuable and important contribution to cognitive science and psychology.

The general point to note in concluding is that in demonstrating the commitment of three paradigmatic examples of computational research to the narrow conception and outlining two problems the view faces, the case for the existence and problematic influence of the view has been at least partially motivated. The narrow conception has, on occasion, problematically structured at least some of the thinking within computational psychology, and that in doing so it has laid down some of the conceptual track on which the computational research train has run. Given this, further examination of previously underexplored approaches, such as wide computationalism, may help enrich the range of theories, models, and explanations offered within computational psychology.

## References

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.
Anderson, J., \& Lebiere, C. (2003). The newell test for a theory of cognition. Behavioral and Brain Sciences, 26, 587-640
Anderson, J., \& Liebere, C. (1998). The atomic components of thought. Mahwah, NJ: Lawrence Erlbaum Associates.
Dawson, M.R.W., Dupuis, B., \& Wilson, M. (2010). From Bricks to Brains: The Embodied Cognitive Science of LEGO Robots. Edmonton: Athabasca University Press.
Boden, M. A. (1981). Minds and Mechanisms: Philosophical Psychology and Computational Models. Brighton, Sussex: The Harvester Press.
Clark, A. (2008). Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford: OUP.
Cleeremans, \& Dienes, Z. (2008). Computational Models of Implicit Learning. In R. Sun (ed.), The Cambridge Handbook of Computational Psychology (pp.396-421). Cambridge, MA: Cambridge University Press.
Dienes, Z. (1992). Connectionist and memory- array models of artificial grammar learning. Cognitive Science, 16, 41-79.
Egan, M. F. (2000). Computation and Content. The Philosophical Review, 104(2), 181-203.
Fodor, J. (1983). Modularity of Mind. Cambridge, MA: Bradford/MIT Press.
Gardner, H. (1985). The Mind's New Science. New York: Basic Books.
Glanzer, M., Adams, J. K., Iverson, G. J., \& Kim, K. (1993). The regularities of recognition memory. Psychological Review, 100, 546-567.
Hutchins, E. (1995). Cognition in the wild. Cambridge: MIT Press.
Kersten, L. (2016). A Mechanistic Account of Wide Computationalism. Review of Psychology and Philosophy, (online), 1-17. DOI: 10.1007/s13164-016-0322-3.

Kersten, L., \& Wilson, R.A (2016). The Sound of Music: Externalist Style. American Philosophical Quarterly, 53(2), 139-154.
Kirsh, D. (2009). Problem Solving and Situated Cognition. In P. Robbins and M. Aydede, The Cambridge Handbook of Situated Cognition (pp. 264-306). Cambridge, Mass.: Cambridge University Press.
Milkowski, M. (2015). Computational mechanism and models of cognition. Philosophia Scientiae, 18(3), 1-14.
Newell, A. (1990). Theories of Unified Cognition. Cambridge, MA: Harvard University Press.
Newell, A., \& Simon, H. (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.
Newell, A., \& Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. Communications of $A C M, 19,113-126$.
Osherson, D. N., Smith, E.E., Wilkie, O., Lopez, A., Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.
Piccinini, G. (2015). Physical Computation: A Mechanistic Account. Published to Oxford Scholarship Online. doi: 10.1093/acprof:oso/9780199658855.001.0001.

Pylyshyn, Z. (1984). Computation and Cognition: Toward a Foundation for Cognitive Science. Cambridge, MA: MIT Press.
Rowlands, M. (1999). The Body in Mind: Understanding Cognitive Processes. Cambridge: CUP.
Segal. G. (1991). Defence of a Reasonable Individualism. Mind, 10, 485-494.
Shagrir, O. (2001). Content, Computation, and Externalism. Mind, 110 (438), 369-400.
Shiffrin, R. M., \& Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. Psychonomic Bulletin and Review, 4, 145-166.
Simon, A. (1962). The Architecture of Complexity. Proceedings of the American Philosophical Society 106, 467-482.
Sternberg, S. (1969). Memory Scanning: Mental Processes Revealed by Reaction Time Experiments. American Scientist, 57, 421-457..
Sun, R. (2008). Theoretical status of computational cognitive modeling. Cognitive Systems Research, 1-17.
Sun, R., \& Ling, C. (1998). Computational cognitive modeling, the source of power and other related issues. AI Magazine, 19(2), 113-120.
Wilson, R. (1994). Wide computationalism. Mind 103(4): 351-372.
Wilson, R. (1995). Cartesian psychology and physical minds: Individualism and the sciences of the minds. Cambridge: Cambridge University Press.
Wilson, R. (2004). Boundaries of the mind: The individual in the fragile sciences. Cambridge: Cambridge University Press.
Wilson, R. (2014). Ten questions concerning extended cognition. Philosophical Psychology 27(1): 19-33.

# Resolving Two Tensions in 4E Cognition Using Wide Computationalism 

Luke Kersten (L.M.Kersten@sms.ed.ac.uk)<br>School of Philosophy, Psychology, and Language Sciences, University of Edinburgh<br>3 Charles Street, EH89AD, Edinburgh, UK<br>Joe Dewhurst (s0819473@sms.ed.ac.uk)<br>School of Philosophy, Psychology, and Language Sciences, University of Edinburgh<br>3 Charles Street, EH89AD, Edinburgh, UK<br>George Deane (s1670299@sms.ed.ac.uk)<br>School of Philosophy, Psychology, and Language Sciences, University of Edinburgh<br>3 Charles Street, EH89AD, Edinburgh, UK


#### Abstract

Recently, some authors have begun to raise questions about the potential unity of 4E (enactive, embedded, embodied, extended) cognition as a distinct research programme within cognitive science. Two tensions, in particular, have been raised: (i) that the body-centric claims embodied cognition militate against the distributed tendencies of extended cognition and (ii) that the body/environment distinction emphasized by enactivism stands in tension with the world-spanning claims of extended cognition. The goal of this paper is to resolve tensions (i) and (ii). The proposal is that a form of 'wide computationalism' can be used to reconcile the two tensions and, in so doing, articulate a common theoretical core for 4 E cognition.


Keywords: 4E cognition, wide computationalism, bodycentrism, extended functionalism, autopoietic theory

## Introduction

Enactive, embodied, embedded and extended cognition, or simply 4E cognition, has often been thought to form a collective challenge to traditional or classical cognitive science (Menary, 2010). Common to many of these views is the idea that cognitive processes are often integrated with and heavily dependent on bodily and environmental structures (Varela, Thompson, \& Rosch, 1991; Clark \& Chalmers, 1998; Haugeland, 1998; Hutto \& Myin, 2013).
More recently, some authors have begun to raise questions about the potential unity of 4 E cognition as a distinct research programme within cognitive science (Clark, 2008a; Clark \& Kiverstein, 2009; Menary, 2010). Two tensions, in particular, have been raised: (i) that the body-centric claims embodied cognition militate against the distributed tendencies of extended cognition and (ii) that the body/environment distinction emphasized by enactivism stands in tension with the world-spanning claims of extended cognition. These two tensions constitute a problem for cognitive science insofar as 4 E cognition is thought to form distinct field of study, and not merely a loose set of alphabetically related approaches (Ward \& Stapleton, 2012).
The goal of this paper is to resolve tensions (i) and (ii). The proposal is that a form of 'wide computationalism' can be used to reconcile the two tensions, and, in so doing,
articulate a common theoretical core for 4 E cognition. It is argued that wide computationalism satisfies the various demands of the embodied, enactive and extended theorists in virtue of placing a simultaneous emphasis on abstract analysis and functional mechanisms.

## Two Tensions in 4E Cognition

Following Clark and Kiverstein (2009), three 'strands' can be identified as generating the two tensions within 4E cognition, these include: body-centrism, extended functionalism, and autopoietic theory.

The first strand is body-centrism. This is the idea that the body has a non-trivial role in determining mental states and functioning, that the details of a creature's embodiment have a profound affect on the nature and functioning of the mind (Noë \& Reagan, 2001; Noë, 2004; Gallagher, 2005). Shapiro, for instance, writes: " $[\mathrm{P}]$ sychological processes are incomplete without the body's contributions. Vision for human beings is a process that includes features of the human body. This means that a description of various perceptual capacities cannot maintain body-neutrality" (2004, p.190). The body is depicted as 'intrinsically special.' Body-centrists hold that without discussion of the unique contribution of bodily structures and activities, cognitive explanations are crucially lacking. The view is also sometimes called the "constitutive-contribution claim" (Clark, 2008a).

Support for body-centrism comes from research highlighting the functional dependences of mental processes on bodily structures and activities (Clark, 2008a, b). Work on embodiment and conceptualization, for example, demonstrates that understanding abstract concepts, such as love, often depends on the metaphorical expansions of more familiar concepts, such as up and down or front and back (Lakoff \& Johnson, 1980).

The second strand is extended functionalism. For the extended functionalist, cognitive systems are functional wholes distributed across diverse sets of components and processes. Cognitive activities involve a complex balancing act between brain, body and world (Harman 1998, Clark \& Chalmers, 1998; Wilson 2004; Wheeler, 2010). The spirit of
extended functionalism is embodied in what Clark and Chalmer's (1998) call the 'parity principle', which says: "[i]f, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as a part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process" (p.29). The parity principle stresses the location neutrality of cognitive analysis. It highlights abstract, functional analysis in place of detailed physiological investigations. The view is also sometimes called the "distributor role" in discussions of embodied cognition (Wilson \& Foglia, 2016).

Support for extended functionalism comes from research focusing on the way in which cognizers often exploit, scaffold, and distribute cognitive activates across bodily and environment structures (Clark, 2005, 2008b). Work on problem solving, for example, shows that people often simplify and transform complex problems, such as the Tower of Hanoi, by manipulating physical environments (Kirsh \& Maglio, 1995; Kirsh, 2009).

The final strand to consider is 'autopoietic theory.' The central claim of autopoietic theory is that cognitive systems are created by the reciprocal interaction of internal and external components in the service of some larger function, such as homeostasis (Weber and Varela, 2002). Autopoietic theory connects to 4 E cognition via the notion of 'sense making'.

Autopoietic theory maintains that because living systems, such as cognition, are autonomous, self-regulating systems, and sense making is required for maintaining a system's boundary, autopoietic systems produce and maintain a physical boundary between the organism and its physical environment. Because sense making is a self-regulating act, organisms often bring forth meaning on the basis of their autonomy - autonomy in this context means actively sustaining identity under precarious circumstances. A system maintains its organization by regulating its interactions with the environment via sense making (Thompson, 2007; Wheeler, 2009).

Consider how each of the three strands fit within 4E cognition. First, enactivism and embodied cognition often endorse body-centrism in virtue of emphasizing what they take to be the unique contributions of bodily structures and activities, such as sensorimotor knowledge. Second, extended cognition often endorses extended functionalism by assigning a non-trivial role to environmental elements in sustaining cognitive activities. Third, enactivism is often framed in terms of autopoietic theory insofar as sense making is treated as a constitutive element of demarcating the organism/environment boundary.

Not every version of enactivism is committed to autopoietic theory, and not every version of embodied cognition is committed to body-centrism. There is, at least in principle, some compatibility between the various views. Nonetheless, because some versions of each view are, as a matter of fact, committed to the different strands, the two
tensions do represent a substantial challenge for 4E cognition.

Consider, then, how the three strands generate the two tensions. The first tension follows from the fact that if the body has a non-trivial role in determining mental states, then cognition cannot also be location neutral; the converse of which is that if cognition is location neutral, then the body cannot have a privileged status in cognition. If the body is simply an instrument through which larger functional complexes are realized, then bodily structures cannot form the exclusive realization base of cognitive activities. Extended functionalism precludes the constitutive contribution claims of body-centrism, while body-centrism precludes the possibility of cognitive systems extending beyond the boundary of the individual.

The second tension emerges from the idea that if extended functionalism is correct, and cognitive systems can stretch out into the world, then living systems cannot also be coextensive with cognitive systems, as per enactivism. Here is Clark and Kiverstein (2009) diagnosing the situation:

> If living systems and cognitive systems are identical, both systems must have boundaries that coincide. However, the boundaries of the living systems are the physical boundaries of the organism. If extended functionalism is correct, the boundaries of cognitive systems can criss-cross the physical boundaries of the organism. This is precisely what the enactivists cannot allow. (p.2).

Extended cognition requires that cognitive systems recruit resources outside the boundary of the individual. Enactivism, however, denies this possibility. It therefore undercuts the identification of cognitive systems with extended systems by maintaining a sharp distinction between the physical boundaries of the organism and the environment, assuming also that each view is taken to be a global thesis about cognition.

Tensions (i) and (ii) emerge as a function of the opposing elements within 4E cognition. Tension (i) emerges as a result of body-centrism's emphasis on the unique contribution of the bodily structures, while tension (ii) follows from the location-neutrality of extended functionalism. The tensions are important for at least two reasons. One is that they stand to undermine the collective thrust of 4 E cognition by showcasing fractures within the larger framework (Clark, 2008a). Another is that they reveal a lack of 'deep theoretical core' within 4E cognition. They expose a conceptual gap at the centre of an otherwise vibrant and animated collection of research (Clark \& Kiverstein, 2009).

One constructive proposal that has been offered to resolve the first of the two tensions is Clark (2008a). Clark's suggestion is that the body plays an enabling computational role within cognitive processes that selectively impacts both conscious and non-conscious computational strategies. Clark's view is that the first tension can be resolved by
viewing bodily structures in terms of enabling different kinds of information processing. The cognitive significance of the body resides in the functional role it occupies within 'intelligent' organization - this is what explains the intuition that the body makes a special contribution to cognition.

Clark's proposal, although not explicitly, also provides a solution to the 'deep theoretical core' problem. This is because it articulates, at least in principle, a common 'computational/functional core' for 4 E cognition. Embedded, embodied, and extended approaches are unified by a shared emphasis on distributed functional complexes supporting cognitive activities. What is important is that bodily or environmental structures are situated within a larger computational/functional framework during investigation. Some systems will be individual bound (as per enactive and embodied cognition), while others will spread out across brain, body and world (as per extended cognition).
One problem with Clark's response, despite its advantages, is that it fails to specify the relationship between physical mechanisms and computational systems finely enough. It fails to cash out what it is that allows the body to play its 'enabling role' in cognitive activities in the first place. If the body is merely one element within a larger brain-body-world complex, why should it have such a constraining and enabling role? The problem is not that Clark is wrong in proposing that the body has an enabling computational role, but that the suggestion alone does not suffice to specify what the role amounts to and why it should prove important.

## Two Tensions Resolved

In what follows, we argue that Clark's proposal can be supplemented and further developed by appealing to the notion of 'wide computationalism'. We begin by outlining and motivating wide computationalism and then turn to showing how the view addresses each of the two tensions.
Wide computationalism is the view that some of the units of computational cognitive systems reside outside the individual (Wilson, 1994, 1995, 2004; Hutchins, 1995; Clark \& Wilson, 2009; Kersten, 2016; Kersten \& Wilson, 2016). Wide computationalism stakes a claim on the scope of physical systems, processes, and components that are capable of supporting computational analysis. A wide computational perspective opens up the possibility of exploring computational units that include the brain and aspects of the beyond-the-head environment.
Wide computationalism gains a theoretical foothold via the location neutrality of computational individuation. Since formal systems are indifferent to physical medium and computation is a formal system, it is possible that at least some states and processes relevant to a computational system may reside outside the individual. Nothing in the method of computational individuation precludes the possibility of wide computational systems.
Traditionally, wide computationalism has been committed to what some call "causal mapping accounts" of
computation (Chalmers, 1994; Chrisley, 1995). Causal mapping accounts maintain that in order for a physical system to implement an actual computation there must be a mapping of computational states to physical states such that transitions between the physical states result in corresponding transitions between the computational states. Causal mapping accounts, whether wide or narrow, articulate the conditions for ascription of computational implementation in terms of isomorphic mappings between computational descriptions and physical descriptions via transitions between physical states.
More recently, some have argued that wide computationalism should adopt a 'mechanistic' approach to computation (Kersten, 2016). Wide mechanistic computation differs from causal mapping formulations in that it frames the conditions of concrete computation in terms of functional mechanisms (Milkowski, 2013, 2015; Piccinini, 2015; Dewhurst, 2016). Mechanistic accounts maintain that concrete computations occur wherever there is a physical system that has an organization of spatiotemporal components such that it computes an abstract function in virtue of manipulating medium-independent vehicles. The mechanistic approach emphasizes functionally integrated systems that compute at least one abstract function via vehicle manipulation.
The wide account of computation extends the mechanistic reasoning to brain-body-world systems. It maintains that whether or not functional mechanisms, ones that process medium-independent vehicles, are constituted by spatiotemporal components squarely localized within the individual or crisscrossing into the world is an a posteriori question. Since the mechanistic conditions on concrete computations are medium and location neutral, the question of wide computational systems is an open one - some physical computing cognitive systems may be ensconced within the body, while others may be spread out over brain, body and world.
In addition to its theoretical plausibility, wide computationalism also gains support from studies in animal and human psychology. There is a natural set of phenomena productively studied by wide computationalism.
Research in form perception, for example, shows that formal primitives in the environment are often relevant to computationally explaining the construction of complex, internal representations. Wilson $(1994,1995)$ takes such work to be indicative of a wide computational system, as it acknowledges the unique computational role of states beyond the individual within perceptual processing.
Research on the spatial navigation of bats has also been used to support wide computationalism. Kersten (2016), for instance, argues that bats' navigation system instantiates a wide computational system in virtue of employing a functional mechanism that spans the brain, body and world, and which processes medium-independent acoustic vehicles. Bat morphology, acoustic signals and neural processing conspire to support object detection along vertical planes using a wide computational system (MacIver, 2009).

Finally, Hutchins (1995) has argued for the presence of wide computation in the context of ship navigation. Hutchins' claim is that members of a navigation team carry out computational tasks that extend beyond the local actions of individual team members, which is indicative of a wide computational system.

The central message is that wide computational systems are not only theoretically plausible, but they are actually implemented in a number of cases. Research in human and animal psychology delivers several examples of concrete computational systems that extend beyond the boundary of the individual. In what follows, we adopt the wide mechanistic account of computation, though, for ease of exposition, we refer to it simply wide computationalism.

Consider how wide computationalism might address the first of the two tensions. Noë (2004), for instance, writes: "If perception is in part constituted by out and out possession and exercise of bodily skills...then it may also depend on our possession of the sorts of bodies that can encompass those skills, for only creatures with such a body could have those skills" (p.25). Noë and other body-centrists are at pains to highlight the role and contribution of bodily actions in cognitive processes. Such considerations motivate the claim that bodily-structures are constitutive of cognitive processes.

The wide mechanistic account can accommodate these types of considerations by focusing on functional mechanisms. A system is a functional mechanism when it consists of a set of spatiotemporal components that contribute to the system's overall function in virtue of the organization and interaction of its component parts (Piccinini, 2015, p.119). For the wide computationalist, the set of bodily and neural structures responsible for delivering visual perception are the functional mechanism that carry out the larger, computational task under investigation. This means that the claims of the body-centrist can be reframed in terms of 'wide' functional mechanisms; these are mechanisms whose component parts are spread out over internal and environmental elements (see Menary, 2007, ch.2). The constitutively embodied systems, such as in the case of vision, are the wide functional mechanisms localized to the body-brain complex. One way to understand bodycentric theorists, then, is as making fine-grained statements about wide functional mechanisms.

Consider the extended functionalist side of the equation. In applying the method of computational analysis to worldindividual spanning systems, wide computationalism maintains a commitment to the location neutrality of cognition. What matters for the wide computationalist is the functional capacity being investigated, not the physical medium through which it is realized. The implication is that extended functionalism's emphasis on medium independence and abstract analysis is preserved within wide computationalism. The view retains the abstract form of analysis crucial to the extended functionalist.

A resolution to the first tension is in sight. By reenvisaging body-centrism in terms of the implementation of
wide computational systems space is opened up for the tight, causal integration of bodily and neuronal processes in support of cognitive processes (i.e., wide functional mechanisms) and the locational neutrality of computational individuation (i.e., wide computational analysis). The special status of bodily structures turns out to be species of a more general class of wide mechanistic systems. The only difference is that whereas some wide functional mechanisms are instantiated within individuals (as highlighted by bodycentrism), others are instantiated by the brain, body and world (as highlighted by extended functionalism).
Consider the second tension. Enactivism, of the autopoietic variety, was unimpressed by the blurring of the organism/environment boundary within extended functionalism. The enactivist claimed that if cognitive systems were autopoietic systems then it followed that cognitive systems could not be extended, the underlying assumption being that autopoietic systems were organism bound.
One route to reconciling the two views is to show that autopoietic theory is compatible with computationalism more generally, assuming also that extended functionalism is congenial to wide computationalism more generally (see Wilson, 2004). Insofar as wide computationalism is a species of computationalism, the compatibility of computationalism with autopoietic systems theory suffices to show the compatibility of extended functionalism with autopoietic theory.
One reason to think that autopoietic theory is compatible with computationalism is a common emphasis on mechanistic explanation (Machamer, Darden, \& Craver, 2000; Bechtel, 2008). Consider, for instance, what Maturana and Varela (1980) write about the methodology of autopoietic theory:

An explanation [of autopoietic theory] is always a reformulation of a phenomenon showing how its components generate it through their interactions and relations...the elements used in the explanations are bodies and their properties...they are relations and their relations, independently of the nature of the bodies that satisfy them...This mode of thinking is not new, and is explicitly related to the very name of mechanisms. (pp. 75-76).

Autopoietic theory is, at root, a functional or mechanistic approach to explanation. What matters is that systems are explained in terms of the interaction of component parts with each other and the environment, regardless of whether the systems under investigation are biological or cognitive in character. There is a functional/mechanistic mode of explanation underlying autopoietic theory.

Compare this with computationalism. Computational analysis involves detailing how the arrangement and interaction of various components conspire to process information bearing vehicles. The mechanistic account requires showing how concrete, functional mechanisms compute medium-independent vehicles in virtue of
processing some portion of their physical structure. Computational analysis is also a species of mechanistic explanation (Milkowski, 2013).

There turns out, then, to be little incompatibility between computationalism and autopoietic theory. Both approaches employ something like a mechanistic explanatory strategy when investigating phenomena, although autopoietic theory is generally pitched at lower-level biological phenomena than computationalism. The common focus on mechanistic explanation ensures that there will be an overlapping set of phenomena productively studied by both approaches.

Why, then, the perceived tension? One reason is that many of the systems analyzed by autopoietic theory are located or ensconced within the organism. Most autopoietic systems are contained within the organism as a matter of empirical fact. However, this alone does not imply that all systems must be analyzed so as to localize within the individual; parts of the environment may still come to be included within the larger analysis. In principle, autopoietic theory, similar to computationalism, is location neutral.

Another reason for the perceived tension is that some within the enactivism literature assume that computationalism implies a commitment to representation and/or information processing theories (Di Paolo, 2009). These authors assume that because computationalism entails a commitment to representation and information processing theories, and autonomous, self-regulating systems stand in contrast to these views, enactivism must be opposed to computationalism.

However, as the previous discussion of mechanistic computation illustrates, there are a number of viable accounts of computation that are minimal in their commitment to representation or information processing theories (Stich, 1983; Egan, 1995; Piccinini, 2008). There is little reason to think that an opposition to representation and information processing commits enactivism to an opposition to computationalism more generally.

The point to note is that because wide computationalism places a greater emphasis on the way in in computational processes are grounded in particular physical mechanisms it creates a link between the body-centric claims of enactivism and embodied cognition, on the one hand, and the functional considerations of extended cognition, on the other. It is in virtue of analyzing mechanisms from several vantages some quite fine-grained, others quite coarse-grained - that wide computationalism is able to mediate the competing claims of the three strands. The view offers a philosopher's stone of sorts through which to translate the various claims of the enactivist, embodied and extended theorist.

To be a bit more specific, tension (i) is resolved by the fact that wide computationalism allows the claims of the body-centrist to be reframed in terms of wide functional mechanisms, while tension (ii) is resolved by the fact that wide computationalism, in virtue of being a species of computationalism more generally, is theoretically compatible with the basic methodology of autopoietic theory. The success of wide computationalism, therefore,
stems from the fact that it retains several of the central insights and elements that prove important to embodied, enactive and extended theorists.

One interesting implication of the preceding analysis is that it reveals what might be called the 'computational/mechanistic' core of 4E cognition. The discussion of wide computationalism goes some way to showing that something akin to 'computational/mechanistic' explanation may underwrite a fair amount of 4E cognition. Mechanistic explanations, which are explanations of systems in terms of the activities and organization of component parts, turn out to be important not only for cognitive science and psychology more generally, but for 4E cognition specifically (Craver, 2006). This point is only provisional, of course. But it does point to a promising future line of inquiry. A continued focus on the 'computational/mechanistic' underpinnings of 4E cognition may well serve to further clarify and unify the field as a whole.

## Conclusion

The preceding discussion is only the first step in a larger analysis. More still needs to be said. Nonetheless, the discussion is important because it offers one route to resolving the two tensions troubling 4E cognition. What's more, the discussion shows that a renewed focus on computationalism, particularly of the wide variety, may have key role to play in illuminating the conceptual foundations of 4 E cognition. This result is both interesting and novel, as discussions of 4 E cognition sometimes eschew mention of computation. By showing that a form of computationalism provides a theoretically flexible yet robust vehicle through which to understand and translate the various strands of 4 E cognition, the current discussion provides not only a partial vindication of wide comptuatiaonlism, but also helps to shed light on an important set of issues facing a growing research programme within cognitive science.

## References

Adams, F., \& Aizawa, K. (2008). The bounds of cognition. Malden: Blackwell.
Bechtel, W. (2008). Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience. London, Routledge.
Chalmers, D. (1994). On implementing a computation. Minds and Machines, 4(4), 391-402.
Clark, A. (2005). Intrinsic content, active memory, and the extended mind. Analysis, 65, 1-11.
Clark, A. (2008a). Pressing the Flesh: A Tension in the Study of the Embodied, Embedded Mind? Philosophy and Phenomenological Research, 76(1), 37-59.
Clark, A. (2008b). Supersizing the mind: Embodiment, action, and cognitive extension. New York: Oxford.
Clark, A., \& Chalmers, D. (1998). The Extended Mind. Analysis, 58, 7-19.

Clark, A., \& Kiverstein, J. (2009). Introduction: Mind Embodied, Embedded, Enacted: One Church or Many? Topoi, 28, 1-7.
Craver, C. (2006). When Mechanistic Models Explain. Synthese, 153(3), 355-76.
Chrisley, R.L. (1995). Why everything doesn't realize every computation. Minds and Machines 4: 403-430.
Dewhurst, J. (2016). Review of Physical Computation. Philosophical Psychology. DOI: 562 10.1080/09515089.2016.1150450.

Di Paolo, E. A. (2009). Extended Life. Topoi, 28, 9-21.
Egan, F. (1995). Computation and Content. Philosophical Review, 104, 181-203.
Gallagher, S. (2005). How the body shapes the mind. Oxford: OUP.
Haugeland, J. (1998). Mind Embodied and Embedded. In J. Haugeland (ed.), Having Thought: Essays in the Metaphysics of Mind (pp. 233- 267). Cambridge, MA: Harvard University Press.
Hutchins, E. (1995). Cognition in the wild. Cambridge: MIT Press.
Hutto, D. and Myin, E. (2013). Radicalizing Enactivism: Basic minds without content. Cambridge, MA: MIT Press.
Lakoff, G. (1980). The Metaphorical Structure of the Human Conceptual System. Cognitive Science, 4, 195208.

Machamer, P. K., Darden, L., \& Craver, C. (2000). Thinking about Mechanisms. Philosophy of Science, 67: 1-25.
MacIver, M.A. 2009. Neuroethology: From morphological computation to planning. In (eds.) P. Robbins and M. Aydede, The Cambridge Handbook of Situated Cognition (pp.480-504). New York, Cambridge University Press.
Maturana, H. R., \& Varela, F. J. (1980). Autopoiesis and cognition: the realization of the living. Dordrecht, Holland: Kluwer Academic Publishers.
Menary, R. (2007). Cognitive integration: Mind and cognition unbounded. Basingstoke: Palgrave Macmillan.
Menary, R. (2010). Introduction to the special issue on 4E cognition. Phenomenology and Cognitive Science, 9:459463.

Machamer, P. K., Darden, L., \& Craver, C. (2000). Thinking about Mechanisms. Philosophy of Science, 67, 1-25.
MacIver, M.A. (2009). Neuroethology: From morphological computation to planning. In (eds.) P. Robbins and M. Aydede, The Cambridge Handbook of Situated Cognition (pp.480-504). New York, Cambridge University Press.
Milkowski, M. (2013). Explaining the Computational Mind. Cambridge, MA.: MIT Press.
Milkowski, M. (2015). Computational mechanism and models of cognition. Philosophia Scientiae 18(3): 1-14.
Noë, A. (2004). Action in Perception. MIT Press.
Kersten, L., \& Wilson, R.A (2016). The Sound of Music: Externalist Style. American Philosophical Quarterly, 53(2), 139-154.
Kersten, L. (2016). A Mechanistic Account of Wide Computationalism. Review of Psychology and Philosophy, (online), 1-17.

Kirsch, D. (2009). Problem solving and situated cognition. In (eds.) P. Robbins and M. Aydede, The Cambridge Handbook of Situated Cognition, (pp. 264-306). New York, NY: Cambridge University Press.
Kirsh, D., \& Maglio, P. (1995). On distinguishing epistemic from pragmatic actions. Cognitive Science, 18, 513-549.
O'Regan, J. \& Noë, A. (2001). A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences, 25(4), 883-975.
Piccinini, G. (2007). Computing mechanisms. Philosophy of Science, 74(4): 501-526.
Piccinini, G. (2008). Computation without Representation. Philosophical Studies, 137(20), 205-241.
Piccinini, G. (2015). Physical Computation: A Mechanistic Account. Published to Oxford Scholarship Online. doi: 10.1093/acprof:oso/97801996588555.001.0001.

Stich, S. (1983). From folk psychology to cognitive science. Cambridge: MIT Press.
Thompson, E. (2007). Mind in Life: Biology, phenomenology, and the sciences of mind. Cambridge, MA: Harvard University Press.
Varela, F., E. Thompson, and E. Rosch. (1991). The embodied mind: Cognitive science and human experience. Cambridge: MIT Press.
Weber, A., \& Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundation of individuality. Phenomenology and the Cognitive Sciences, 1(2), 97-125.
Wheeler, M. (2009). Reconstructing the Cognitive World. MIT Press.
Wheeler, M. (2010). In defense of extended functionalism. In R. Menary (Ed.), The Extended Mind (pp. 245-270). MIT Press.
Wilson, R. (1994). Wide computationalism. Mind, 103(4), 351-372.
Wilson, R. (1995). Cartesian psychology and physical minds: Individualism and the sciences of the minds. Cambridge: Cambridge University Press.
Wilson, R. (2004). Boundaries of the mind: The individual in the fragile sciences. Cambridge: Cambridge University Press.
Wilson, R. (2010). Extended vision. In (eds.) N. Gangopadhyay, M. Madary, and F. Spicer, Perception, action and consciousness (pp.277-290). New York: Oxford University Press.
Wilson, R.A., and Clark, A. (2009). How to situate cognition: Letting nature take its course. In (eds.) M. Aydede and P. Robbins, The Cambridge handbook of situated cognition (pp.55-77). New York: Cambridge University Press.
Wilson, R. A., \& Foglia, L. (2016). Embodied cognition. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

# Processing Spatial Relations: A Meta-Analysis 

Ann-Kathrin Kießner (Ann-Kathrin.Kiessner@ venus.uni-freiburg.de)<br>Center for Cognitive Science, University of Freiburg, 79098 Freiburg, Germany<br>Marco Ragni (ragni@informatik.uni-freiburg.de)<br>Cognitive Computation Lab, Technical Faculty, University of Freiburg, 79110 Freiburg, Germany


#### Abstract

The ability to reason about relations is relevant for many spatial cognitive processes. This can involve: (i) to represent spatial information mentally, (ii) to manipulate the spatial representation, and (iii) to infer new spatial information. Several cognitive theories make assumptions and predictions about the underlying processes. A detailed and systematic overview and analysis of ireliable effects across studies is missing. This article presents a meta-analysis of 35 studies about spatial relational reasoning. Studies were classified according to different factors including the ambiguity of the spatial description, i.e., if it the description allows for more than one representation, the presentation of information, i.e., if the information has been presented auditorily or in a written form, and the task, i.e., if a conclusion or model of the premises needs to be generated or verified. Implications of the findings for the mental model theory and working memory are discussed.


Keywords: reasoning; spatial relations; meta-analysis.

## Introduction

Spatial cognition allows us to perform a variety of everyday actions, such as sharing spatial information, navigation, and even assembling diverse kinds of objects. A successful spatial interaction requires us to represent spatial relational information and to reason with and about this information. Human communication mainly uses qualitative descriptions to specify relationships between spatial objects ${ }^{1}$ instead of a numerical or quantitative data description that is used in robot navigation. Relations which are expressed linguistically by the comparative, such as 'greater than', have been extensively investigated in the past century by using behavioral experiments (e.g., Hunter, 1957; Störring, 1908). Within an experiment, problems are often reduced to their essential characteristics limiting irrelevant information. Consider the following example:

## Premise 1: The post office is left of the train station.

Premise 2: The train station is left of the main crossroad.
Conclusion: The post office is left of the main crossroad.
The premises contain spatial information ("left of") about the relationships among spatial objects (e.g.,"post office"). A deductive inference makes implicit given information, e.g., the relation between the post office and the main crossroad explicit. While this inference is easy, and most participants solve it correctly, such transitive inferences can be at the core of more difficult inference problems with more objects and more relations. In the following we will give a brief overview

[^366]about reported factors of reasoning difficulty in the literature and present briefly implications of two cognitive theories relevant to our analysis. Cognitive psychologists have disagreed about the exact character of underlying mental processes (Goodwin \& Johnson-Laird, 2005) and the main question is, what reliable findings need to be explained by cognitive theories?

## Factors of Reasoning Difficulty

In general reasoning difficulty can appear on all levels: in comprehending the presented information (the language level), generating a mental representation (the representational level including working memory) or reasoning about the representation (reasoning level). The literature reports several factors that can be related to these levels.

Behavioral findings support that the presentation format affects spatial relational reasoning (Van der Henst \& Schaeken, 2005): In comparison with simultaneous premise presentation, accuracy is significantly lower in sequential presentation of the premise information (Roberts \& Sykes, 2003; Schaeken \& Johnson-Laird, 2000; Van der Henst \& Schaeken, 2005). This difference presumably reflects that during simultaneous premise presentation reasoners have all information available until they respond in contrast to sequential premise presentation posing more demand on working memory (Ormrod, 1979; Schaeken \& Johnson-Laird, 2000). Models of working memory, e.g., Baddeley's Working Memory Model (WMM; Baddeley, 1986) support the assumption that human reasoning is restricted by the limited capacity of working memory (Klauer, 1997) and that there are different components with specific limitations and modalities. Based on Baddeley's WMM, factors such as presentation form, task type and number of terms may influence spatial relational reasoning. Similar to premise presentation, it can be assumed that the presentation form (written vs. spoken language) may affect the way a problem is processed (Ormrod, 1979).

If premises are presented auditorily the spatial information is presented sequentially and thus more load on working memory is placed (Ormrod, 1979). According to the WMM, the larger the number of terms within a problem, the higher is the amount of information that must be retained (Clevenger \& Hummel, 2014). Lastly, the task type may have an influence as well. In conclusion generation tasks reasoners have to generate a conclusion. Whereas during verification tasks, reasoners have to check if a putative conclusions follows. This type of task captures only the ability to recognize a solution, but not producing it and thus requires smaller amounts of work-
ing memory (Klauer, 1997; Kubinger \& Wolfsbauer, 2010).
Another factor is the so-called indeterminacy effect (Byrne \& Johnson-Laird, 1989). While Example 1 above allows only for one possible arrangement (that we call model) and is called a determinate problem, indeterminate problems are possible. Consider the following example:

Premise 1: The post office is left of the train station.
Premise 2: The train station is left of the police office.
Premise 3: The crossroad is right of the post office.
Conclusion: The police office is right of the crossroad.
An indeterminate problem is more difficult to solve than a determinate one (e.g., Boudreau \& Pigeau, 2001; JohnsonLaird \& Byrne, 1991). Determinate problems allow only one qualitative arrangement (in contrast to quantitative differences, e.g., metric distances), while indeterminate problems allow for multiple different arrangements - as explained by the (preferred) mental model theory (MMT: Byrne \& Johnson-Laird, 1989; Ragni \& Knauff, 2013):
(1) post office - crossroad - train station - police office
(2) post office - train station - crossroad - police office

Most people construct just one or two models at most and often neglect other models consistent with the premises. These models differ qualitatively, e.g., there is mentally a different arrangement of the train station and the crossroad possible from the indeterminate description.

So far no systematic review of the recent literature has been carried out and no uniform and unambiguous conclusions about differences in accuracy have been drawn. For this reason, a cross-study meta-analysis of behavioural data from spatial relational reasoning is conducted. The aim of this paper is to test whether the predictions of individual studies and the predictions of MMT and WMM hold generally and to give an detailed overview that can serve as a potential benchmark for theories about spatial reasoning. This analysis investigates differences in accuracy depending on indeterminacy, premise presentation, presentation form, and type of task as well as number of terms. Resulting from the theoretical background and empirical findings to spatial relational reasoning, predictions are:

1. In spatial relational reasoning, determinate problems are easier to solve than indeterminate problems (e.g., Boudreau \& Pigeau, 2001; Johnson-Laird \& Byrne, 1991).
2. Compared with problems presented in spoken language, problems in written language appear with higher accuracy (e.g., Ormrod, 1979; Van der Henst \& Schaeken, 2005).
3. Reasoners solve more problems correctly when the task is to verify instead of generating conclusions/models.
4. Problems consisting of three terms are less difficult to solve than four-term problems.
5. In case of simultaneous premise presentation, accuracy is higher than in case of sequentially premise presentation.

In the following, these predictions are analyzed with the aim of gaining differences in accuracy specific to the various types. The results of the analysis are evaluated and interpreted with respect to predictions of the mental model theory and implications from working memory limitations.

## The Meta-Analysis

Paper Acquisition In order to acquire sufficient and suitable data for the meta-analysis, we needed to find experiments in which the participants drew their own conclusions to all sorts of tasks in spatial relational reasoning. An initial set of eligible studies came from a meta-analysis database of coded studies about spatial relational reasoning from the Cognitive Computation Lab (University of Freiburg, Germany) the database incorporated a comprehensive search for studies reported until 2013. The database contained the literature using the online platforms PubMed and Google Scholar for entries by the following main query: '(relational) AND (reasoning) OR (reasoning) AND (about) AND (relations) OR (transitive reasoning)'. All the studies in that database were reviewed for eligibility and an independent search was conducted: Online literature searches were performed on the 29th of October, and 29th of November 2016 using PubMed and Google Scholar. For the first PubMed and Google Scholar search, the same term like in 2013. For the second PubMed search, conducted on the 29th of November, the query '(spatial) AND (reasoning) AND (relations) OR (spatial reasoning) AND (relations)' was used, since the initial query was too unspecific for this search engine.

Criteria for Inclusion of Studies Experiments were assessed and selected for this meta-analysis if they met not only the search terms but also the following criteria: Experiments containing spatial relations and in cases of visuo-spatial relations, experiments emphasizing on spatial representations were also considered. All the experiments had to involve healthy, adult participants and use a within-subjects design to keep the homogeneity among different conditions. Participants had to know beforehand that their task was to reach a conclusion and there had to be no secondary-tasks. These criteria were used to ensure that the reasoning process was actually taking place and to eliminate other cognitive processes as a biasing factor. Moreover, only peer-reviewed and published studies of both, behavioural and neurophysiological experiments conducted in any country were considered. Outcome results of accuracy must have been presented in a quantitative form that permitted computation or reasonable estimation of an effect size statistic representing the difference in accuracy. Finally, information on factors of interests had to be given in the study. The literature search identified 138 experiments of 84 articles reporting results from psychological studies. Of these, 32 experiments (23\%) (e.g., Knauff \& Johnson-Laird, 2002a) did not report behavioral data or did not present spatial realtions by means of language.

Twenty experiments ( $14 \%$ ) were rejected because they did not report or measured accuracy (e.g., Brüssow et al.,

Table 1: Means and standard deviations of the overall accuracy for the studies (in \%).

| Dataset with indeterminacy condition | determinate problems | indeterminate problems |
| :--- | :---: | :---: |
| simultaneous-verbal presentation | $70(7)$ | $52(8)$ |
| sequential-verbal presentation | $65(4)$ | $42(13)$ |
| sequential-auditory presentation | $63(10)$ | $34(19)$ |

Values are rounded to integers. Indeterminacy (simultaneous-verbal): Experiments with simultaneous premise presentation, generation task, verbal presentation form and 5-term problems. Indeterminacy (sequential-verbal): Experiments with sequential premise presentation, generation task, auditory presentation form and 5 -term problems. Indeterminacy (sequential-auditory): Experiments with sequential premise presentation, generation task, verbal presentation form and 5-term problems.
2013). Seventeen experiments (12\%) did not report information about either indeterminacy, presentation form, type of task or number of terms nor premise presentation (e.g., Fangmeier et al., 2006). For 13 experiments (9\%) the original study was not available (e.g., Hagert, 1984). Eight experiments ( $6 \%$ ) used secondary-task methods (e.g., Knauff et al., 2004), six experiments ( $4 \%$ ) used a between-subjects design (e.g., Boudreau \& Pigeau, 2001) and three experiments (2\%) included children or patients (e.g., Knauff \& May, 2006). In two experiments ( $1 \%$ ) a recognition task was performed (e.g., Mani \& Johnson-Laird, 1982) and two other experiments ( $1 \%$ ) used a visual presentation form (e.g., Knauff \& May, 2006). In total, 206 raw differences in means of accuracy between different types of spatial relational reasoning problems and other types of (relational) reasoning met the inclusion criteria for this meta-analysis. An asterisk precedes each of these reports in the reference list.

Paper Classification After paper selection, experiment characteristics were coded by the authors for the following characteristics: indeterminacy, premise presentation, presentation form, number of terms, task and sample size. The sample size for each experiment was defined as the number of participants at the time of the final measure of logical correct answer. The first division of data was made between determinate and indeterminate problems. Furthermore, the data was subdivided into groups of simultaneous and sequential premise presentation. All premises were either displayed at the same time and remained available (simultaneous), or were presented one at a time and disappeared with the onset of a new premise (sequential). Moreover, it was coded whether the participants had to listen to the premises in form of audio recording using spoken language (auditory form) or whether the premises were presented by means of literacy language on screen or on paper (verbal form). Likewise, the data was grouped into experiments with either three, four, or five terms. Finally, experiments were assigned to the group of verification task when the reasoners had to verify a outative conclusion or to select the correct model from a given set of models. If the participants had to generate a model (an arrangement of objects) or to draw a conclusion, tasks were characterized as generation tasks. For each factor, the data was divided into subgroups finding the combination of variants that had the
most values for comparison of two variants of a factor. For this purpose, cross-classifying factors were used to build a contingency table of the counts at each combination of factor levels resulting in eight combinations of data. In Table 2, the datasets for all types of factors and its characteristics are presented. The factor number of participants within an experiment was not included in matching since otherwise, the number of raw means within subgroups would have been too small for statistical analysis. For the results reported in this study, statistical analysis consisted of non-parametric tests using one-sided Wilcoxon Rank-Sum Tests and a significance level of $\alpha=.01$ was defined. All statistical analysis was performed using R (R version 3.2.2, 2015-08-14; R Core Team, 2015).

## Results and Discussion

The data was corrected for outliers by excluding percentages of accuracy with values greater than two standard deviations from the mean value. All the statistical analysis reported is based on the data corrected by outliers. We analysed the percentage of correctness for all eight combinations of data. ${ }^{2}$ Table 2 summarizes the descriptive statistics in all conditions.

The analysis shows that in the indeterminacy (simultaneous-verbal) sample, significantly more correct responses were given if the problem was determinate than indeterminate (Wilcoxon Rank-Sum Test $W=168$, $p<.0001$, one-sided). The same trend is also visible for the sequential-verbal and sequential-auditory indeterminacy cases (Table 1). One-sided Wilcoxon Rank-Sum Tests showed that both trends were statistically significant (indeterminacy (sequential-verbal): $W=30, p<.05$; indeterminacy (sequential-auditory): $W=30, p=.01$ ). These results supports several empirical findings (e.g., Byrne \& JohnsonLaird, 1989) and can be explained by MMT as well as Baddeley's WMM. For an indeterminate description, it is necessary to construct not one but several mental models in order to correctly represent its meaning. The larger the number of models that reasoners must consider, the higher is the load on working memory. The attempt to construct

[^367]Table 2: Descriptive statistics of the overall accuracy for the studies (in \%).

| Dataset | Factor | $n_{\text {rawmeans }}$ | $n_{\text {participants }}$ | Mean (SD) | Min | Q1 | Median (MAD) | Q3 | Max |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Indeterminacy*** | det | 12 | 2384 | $70(7)$ | 64 | 64 | $66(3)$ | 78 | 80 |
| (simultaneous-verbal) | indet | 14 | 2456 | $52(8)$ | 40 | 43 | $54(7)$ | 58 | 59 |
| Presentation form 1 (n.s.) | verb | 15 | 255 | $90(10)$ | 68 | 89 | $90(11)$ | 98 | 98 |
|  | audi | 18 | 264 | $87(7)$ | 77 | 81 | $88(9)$ | 94 | 96 |
| Presentation form 2** | verb | 37 | 702 | $72(15)$ | 40 | 61 | $77(17)$ | 84 | 92 |
|  | audi | 10 | 192 | $57(14)$ | 33 | 47 | $61(12)$ | 67 | 76 |
| Task** | veri | 15 | 255 | $90(10)$ | 68 | 89 | $90(11)$ | 98 | 98 |
|  | gen | 30 | 856 | $76(17)$ | 54 | 58 | $75(25)$ | 91 | 99 |
| Number of terms (n.s) | three | 30 | 856 | $76(17)$ | 54 | 58 | $75(25)$ | 91 | 99 |
|  | four | 37 | 702 | $72(15)$ | 40 | 61 | $77(17)$ | 84 | 92 |
| Premise presentation (n.s.) | sim | 24 | 3504 | $49(12)$ | 27 | 44 | $48(10)$ | 55 | 66 |
|  | seq | 15 | 255 | $90(10)$ | 68 | 89 | $90(11)$ | 98 | 98 |

All values were rounded to integers. Factor labels refer to auditory vs. verbal presentation form (audi/verb), determinate vs. indeterminate problems (det/indet), model generation vs. model verification task (gen/veri) and sequential vs. simultaneous premise presentation (seq/sim). $n_{\text {rawmeans }}$ : number of raw differences in means; $n_{\text {participants }}$ : number of participants. Indeterminacy (simultaneous-verbal): Experiments with simultaneous premise presentation, generation task, verbal presentation form and five-term problems. Presentation form 1: Experiments with sequential premise presentation, verification task, determinate and three-term problems. Presentation form 2: Experiments with sequential premise presentation, generation task, determinate and four-term problems. Task: Experiments with sequential premise presentation, verbal presentation form, determinate and three-term problems. Number of terms: Experiments with sequential premise presentation, verbal presentation form, verification task and determinate problems. Premise presentation: Experiments with verbal presentation form, verification task, determinate and three-term problems. ${ }^{* * *}$ significant $p<.001,{ }^{* *}$ significant $p<.01$
several models may overload working memory capacities so that no models would be constructed. Likewise, the accuracy was significantly higher if the problems were presented verbally than during auditory presentation (Wilcoxon Rank-Sum Tests, presentation form 2: $W=283, p<.01$, one-sided). A similar but not significant trend showed in the presentation form 1 condition (Wilcoxon Rank-Sum Tests, $W=176, p=.07$, one-sided). Spoken language can only be presented in a serial way. Thus, the spatial description has to be stored in the phonological loop using a language-based form. At the same time, mental models are built in the visuo-spatial sketchpad. When working memory load increases it becomes harder to keep track of all the premises and inferences separately. Additionally, written language already implies information about spatial relations and is, therefore, more similar to the information contained in the problem description than in case of auditory presentation. Furthermore, a one-sided Wilcoxon Rank-Sum Test supports the prediction that reasoning difficulty is lower in verification tasks ( $W=332, p<.01$ ). Thereby, reasoners have to build a model based on the premises presented and verify a proposed model or conclusion. In addition to model construction, reasoners have also to draw a conclusion to solve the generation task correctly. This requires a larger amount of working memory (Klauer, 1997). The descriptive results for the dataset number of terms are consistent with the predictions. Three-term problems were higher in accuracy than four-term problems (Table 2). However, the Wilcoxon Rank-Sum Test showed no significant difference ( $W=641$, $p>.1$, one-sided). One explanation could be the limited
capacity of the working memory that is roughly three to five objects or role bindings (Clevenger \& Hummel, 2014). Both three-term and four-term problems do not exceed the capacities of working memory. However, descriptive results assume a tendency for an increasing memory load in four-term problems. Contrary to the predictions, there was no significant difference between accuracy in case of simultaneous and sequential premise presentation ( $W=0$, $p>.1$, one-sided Wilcoxon Rank-Sum Test). Furthermore, descriptive results indicate that problems with sequential premise presentation are less difficult to solve than simultaneously presented problems (Table 2). This result was unexpected and contradicts previous findings (Roberts \& Sykes, 2003; Van der Henst \& Schaeken, 2005). With regard to the data used in this study, the following differences can be observed which may have influenced the result: In case of simultaneous presentation, the total number of participants tested is higher than in sequential condition ( $N=3504$ vs. $N=255$ ). The difference in sample size might have influenced the results. In addition, the factor premise order was not controlled in this study. The dataset of simultaneous premise presentation contained discontinuous problems. However, in the set of sequential premise presentation, the factor premise order was either unspecified or continuous. Studies have shown that accuracy is higher in case of continuous premise order (Knauff et al., 1998). Thus, an effect of premise order can not be excluded here. A further explanation may be that the amount of information must be processed is reduced as a result of sequential presentation. Hence, attention control may be facilitated and a model can
be constructed incrementally from the premises (principle of economicity, Manktelow \& Galbraith, 2012).

## General Discussion

Despite a century of research on spatial relational reasoning and more than a hundred articles there is still no systematic analysis of which factors do contribute to the human difficulty in reasoning about spatial relations. The quality of the reported experimental data differs, sometimes all relevant information is reported, sometimes the standard deviation is missing. Hence, only a limited number of experiments could be included in our review. We identified reliable differences in accuracy between determinate and indeterminate problems, auditory and verbal presentation form and also between tasks of model generation and verification. This meta-analysis confirmed previous empirical findings that are predicted by the mental model theory and influenced by the limited-capacity working memory as predicted by WMM (see Klauer, 1997): The effect of indeterminacy is the strongest effect (see Table 2) that is directly related with the number of models that need to be generated in the reasoning process. The same holds for the task where the verification is easier (the constructed models need to be compared with a given one), while in the conclusion generation process all models need to be checked. And, finally a generation task with four terms (Presentation form 2 in Table 2) can lead to more demands on both the construction and storing the model in memory.

This systematic review has, however, a few limitations related to the results and their interpretation and leads to some new questions. First, the number of studies considered is a limiting factor to the expressive power of the analysis results. Furthermore, most of the studies included in this metaanalysis did not report any information about the premise order (e.g., if the premise information is continuosly), so it was not possible to control for this factor. Likewise, with regard to the small number of raw means, it was not possible to factor in the sample size of each experiment. The next step of analysis in this study will focus on this particular point and investigates questions, such as the variability in effects across studies and how this variability can be explored in terms of moderator variables. Identification of the moderator variables that describe the study characteristics associated with larger and smaller effects is another kind of contribution meta-analysis can make to understanding difficulties in spatial relational reasoning. Of particular importance is the role such moderator analysis can play in ascertaining which variants of spatial descriptions are most effective for reasoning.

Taken together this study illustrates a use of meta-analysis for data interpretation beyond conventional statistical analysis. Some cross-experimental results can be formulated: First, determinate problems are easier to solve than indeterminate problems. Second, compared with auditory presentation, problems in form of written language are less difficult. Further, the accuracy was better for tasks that require the verification of conclusions or models than in tasks that require to
generate conclusions or models. This meta-analysis confirms some previous empirical findings, and supports predictions of the spatial mental model theory together with assumptions from a limited spatial working memory.

## Acknowledgments

This work has been supported by a DFG Heisenberg fellowship RA 1934/3-1 and projects RA 1934/2-1 and 4-1.

## References

Asterisks refer to studies included in the analysis.
Baddeley, A. (1986). Working memory. Clarendon Press, Oxford.
Boudreau, G., \& Pigeau, R. (2001). The mental representation and processes of spatial deductive reasoning with diagrams and sentences. International Journal of Psychology, 36(1), 42-52.
Brüssow, S., Ragni, M., Frorath, M., Konieczny, L., \& Fangmeier, T. (2013). Premise annotation in mental model construction: An ACT-R approach to processing indeterminacy in spatial relational reasoning. Cognitive Systems Research, 24, 52-61.
*Bucher, L., \& Nejasmic, J. (2012). Relocating multiple objects during spatial belief revision. In U. D. Stachniss C. Schill K. (Ed.), International conference on spatial cognition (Vol. 7463, pp. 478-491). Springer.
*Byrne, R. M., \& Johnson-Laird, P. (1989). Spatial reasoning. Journal of memory and language, 28(5), 564-575.
Clevenger, P. E., \& Hummel, J. E. (2014). Working memory for relations among objects. Attention, Perception, \& Psychophysics, 76(7), 1933-1953.
*De Soto, C. B., London, M., \& Handel, S. (1965). Social reasoning and spatial paralogic. Journal of Personality and Social Psychology, 2(4), 513-521.
*Ehrlich, K., \& Johnson-Laird, P. N. (1982). Spatial descriptions and referential continuity. Journal of verbal learning and verbal behavior, 21(3), 296-306.
*Fangmeier, T., \& Knauff, M. (2009). Neural correlates of acoustic reasoning. Brain research, 1249, 181-190.
Fangmeier, T., Knauff, M., Ruff, C. C., \& Sloutsky, V. (2006). fmri evidence for a three-stage model of deductive reasoning. Journal of Cognitive Neuroscience, 18(3), 320-334.
Goodwin, G. P., \& Johnson-Laird, P. (2005). Reasoning about relations. Psychological Review, 112(2), 468-493.
Hagert, G. (1984). Modeling mental models: Experiments in cognitive modeling of spatial reasoning. In T. O'Shea (Ed.), Proceedings of the 6th European Conference on Artificial Intelligence (pp. 179-188).
Hunter, I. M. (1957). The solving of three-term series problems. British Journal of Psychology, 48(4), 286-298.
Johnson-Laird, P. N., \& Byrne, R. M. (1991). Deduction. Lawrence Erlbaum Associates, Inc.
Klauer, K. C. (1997). Working memory involvement in propositional and spatial reasoning. Thinking \& Reasoning, 3(1), 9-47.
*Knauff, M., Bucher, L., Krumnack, A., \& Nejasmic, J. (2013). Spatial belief revision. Journal of Cognitive Psychology, 25(2), 147-156.
*Knauff, M., Fangmeier, T., Ruff, C. C., \& Johnson-Laird, P. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. Journal of Cognitive Neuroscience, 15(4), 559-573.
Knauff, M., \& Johnson-Laird, P. (2002a). Reasoning and the visual-impedance hypothesis. In C. Freksa, W. Brauer, C. Habel, \& K. F. Wender (Eds.), International Conference on Spatial Cognition (pp. 372-384). Springer.
*Knauff, M., \& Johnson-Laird, P. (2002b). Visual imagery can impede reasoning. Memory \& Cognition, 30(3), 363371.
*Knauff, M., \& May, E. (2006). Mental imagery, reasoning, and blindness. The Quarterly Journal of Experimental Psychology, 59(1), 161-177.
*Knauff, M., Mulack, T., Kassubek, J., Salih, H. R., \& Greenlee, M. W. (2002). Spatial imagery in deductive reasoning: a functional mri study. Cognitive Brain Research, 13(2), 203-212.
Knauff, M., Rauh, R., Schlieder, C., \& Strube, G. (1998). Continuity effect and figural bias in spatial relational inference. In M. A. Gernsbacher \& S. J. Derry (Eds.), Proceedings of the 20th Annual Conference of the Cognitive Science Society (pp. 573-578).
Knauff, M., Strube, G., Jola, C., Rauh, R., \& Schlieder, C. (2004). The psychological validity of qualitative spatial reasoning in one dimension. Spatial Cognition and Computation, 4(2), 167-188.
*Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., \& Knauff, M. (2011). A model for relational reasoning as verbal reasoning. Cognitive Systems Research, 12(3), 377392.

Kubinger, K. D., \& Wolfsbauer, C. (2010). On the risk of certain psychotechnological response options in multiplechoice tests. European Journal of Psychological Assessment, 26, 302-308.
Mani, K., \& Johnson-Laird, P. N. (1982). The mental representation of spatial descriptions. Memory \& Cognition, 10(2), 181-187.
Manktelow, K., \& Galbraith, N. (2012). Thinking and reasoning: An introduction to the psychology of reason, judgment and decision making. Psychology Press.
*Morra, S. (1989). Developmental differences in the use of verbatim versus spatial representations in the recall of spatial descriptions: A probabilistic model and an experimental analysis. Journal of Memory and Language, 28(1), 37-55.
*Morra, S. (2001). On the information-processing demands of spatial reasoning. Thinking \& reasoning, 7(4), 347-365.
*Nejasmic, J., Bucher, L., \& Knauff, M. (2015). The construction of spatial mental models-a new view on the continuity effect. The Quarterly Journal of Experimental Psychology, 68(9), 1794-1812.
*Nejasmic, J., Krumnack, A., Bucher, L., \& Knauff, M. (2011). Cognitive processes underlying the continuity effect in spatial reasoning. In L. Carlson, C. Hoelscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd annual conference of the cognitive science society (pp. 1127-1132).
Ormrod, J. E. (1979). Cognitive processes in the solution of three-term series problems. The American Journal of Psychology, 235-255.
*Prado, J., Van Der Henst, J.-B., \& Noveck, I. A. (2010). Recomposing a fragmented literature: How conditional and relational arguments engage different neural systems for deductive reasoning. Neuroimage, 51(3), 1213-1221.
R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
*Ragni, M., \& Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. Psychological Review, 120(3), 561-588.
*Roberts, M. J. (2000). Strategies in relational inference. Thinking \& Reasoning, 6(1), 1-26.
Roberts, M. J., \& Sykes, E. D. (2003). Belief bias and relational reasoning. The Quarterly Journal of Experimental Psychology: Section A, 56(1), 131-154.
*Ruff, C. C., Knauff, M., Fangmeier, T., \& Spreer, J. (2003). Reasoning and working memory: common and distinct neuronal processes. Neuropsychologia, 41(9), 1241-1253.
*Schaeken, W., Girotto, V., \& Johnson-Laird, P. N. (1998). The effect of an irrelevant premise on temporal and spatial reasoning. Kognitionswissenschaft, 7(1), 27-32.
Schaeken, W., \& Johnson-Laird, P. N. (2000). Strategies in temporal reasoning. Thinking \& Reasoning, 6(3), 193219.

Störring, G. (1908). Experimentelle Untersuchungen über einfache Schlussprozesse. W. Engelmann.
Van Der Henst, J.-B. (2002). Mental model theory versus the inference rule approach in relational reasoning. Thinking \& Reasoning, 8(3), 193-203.
Van der Henst, J.-B., \& Schaeken, W. (2005). The wording of conclusions in relational reasoning. Cognition, 97(1), 1-22.
*Vandierendonck, A. (1996). Evidence for mental-modelbased reasoning: A comparison of reasoning with time and space concepts. Thinking \& Reasoning, 2(4), 249-272.

# Decoding Virtual Agent's Emotion and Strategy from Brain Patterns 

Eunkyung Kim, Sarah I. Gimbel, Aleksandra Litvinova, Jonas T. Kaplan and Morteza Dehghani<br>\{eunkyung; sgimbel; alitvino; jtkaplan; mdehghan\}@usc.edu<br>University of Southern California, Los Angeles, CA 90089 USA


#### Abstract

Recent advances in technology have paved the way for humanagent interactions to become ubiquitous in our daily lives, and decades worth of research on virtual agents have enhanced these interactions. However, for the most part, the effect of different types of agents on the human brain is unknown, and the neuroscience of human-agent interactions is rarely studied. In this study, we examine the underlying neural systems involved in processing and responding to different types of negotiating agents. More specifically, we show that different brain patterns are observed for various types of virtual agents; consequently, we can decode the strategy and emotional display of the agent based on the counterpart's brain activity. Using fMRI data, we analyzed participants' brain activity during negotiations with agents who show three different emotional expressions and use two different types of negotiation strategies. We demonstrate that, using Multi-Voxel Pattern Analysis, we can reliably decode agents' emotional expressions based on the activity in the left dorsal anterior insula, and also agents' strategies based on the activity in the frontal pole.


Keywords: Human-Agent Interaction; Negotiation; Emotion; Decision-Making; fMRI

## Introduction

Virtual agents have become a part of our daily lives. From commercial websites that make use of chat agents for answering users' questions in personalized settings to educational and training software that incorporate virtual agents to provide better learning experiences, our numbers of interactions with virtual agents have dramatically increased in the past couple of years.

Parallel to this increase, various lines of research have studied the interactions between humans and virtual agents. For example, research has examined the contributing factors to user engagement (Bickmore, Schulman, \& Yin, 2010; Castellano, Pereira, Leite, Paiva, \& McOwan, 2009) and establishment of bonds with virtual agents (Cassell \& Thorisson, 1999; Wang \& Gratch, 2009). Bickmore et al. (2010) showed that people engage more with life-like virtual agents, such as a relational agent who remembers past history and relates to that history when communicating. Moreover, Wang and Gratch (2009) demonstrated that virtual agents who give nonverbal immediacy feedback, such as eye contact and gestures, are found to establish rapport with human partners.

Related to our research, the effects of emotion and strategies are among the most widely explored topics in humanagent interaction research (Maldonado et al., 2005; Kim, Dehghani, Kim, Carnevale, \& Gratch, 2014; Van Kleef, De Dreu, \& Manstead, 2004). These factors are particularly important because they are central in providing clues about the internal states and the intentions of the counterpart in any type of interactions (Jurafsky, Ranganath, \& McFarland, 2009; Rafaeli \& Sutton, 1987). Previous studies on
the effect of emotion include the work of Maldonado et al. (2005) where they demonstrated that people who interact with an emotional agent perform better on a test than those who interact with an emotionless agent in a web-based learning environment. Van Kleef et al. (2004) argue that automated agents who express emotions, such as anger or happiness, elicit different levels of concessions based on the type of expressed emotions. Also, several researchers have examined the effects of different types of negotiation strategies during human-agent interactions. For example, Das, Hanson, Kephart, and Tesauro (2001) demonstrated that the agreed trade prices from human-agent interactions are different for two types of agent strategies; one strategy is to maximize its expected surplus using trade history and the other strategy is to make small random adjustments to the trade price continuously. Similarly, Grosz, Kraus, Talman, Stossel, and Havlin (2004) demonstrated that there are particular agent strategies that elicit more concessions during negotiations.

These, along with numerous other studies, have helped the field establish sets of features that influence the quality of human-agent interaction, resulting in a more enhanced and realistic experience for the human user. However, with a few exceptions (e.g., Sanfey, Rilling, Aronson, Nystrom, and Cohen (2003)), the majority of these studies have for the most part treated the process and the mechanism through which these features affect the human counterpart as a black box; they demonstrate that a particular type of agent, with particular emotion and strategy, enhances a user's experience (e.g. performance on a test). However, the question of how these features affect the underlying neural mechanisms of the user that result in such enhancement still remains unanswered. For instance, even though it is established that interacting with an emotional agent results in better performance on a test than interacting with an unemotional agent (Maldonado et al., 2005; Karacora, Dehghani, Krämer-Mertens, \& Gratch, 2012), the neuroscience of these interactions are not fully understood.

In this paper, we investigate the underlying neural systems that are activated when participants interact with agents who show different emotional expressions and apply different negotiation strategies. As various emotions and strategies have been related to diverse reactions in previous studies, we assumed that we could find differences between the neural processes that are activated when these factors are manipulated. To examine these differences, we studied people's neural activation while they were interacting with a virtual agent. We hypothesize that distinct patterns of brain activity would be observed for each agent type.

To study brain activity during human-agent interaction,
we used functional magnetic resonance imaging (fMRI). We used a human-agent negotiation platform for the experiment in order to capture active interaction between a participant and a virtual agent while they are trying to reach an agreement. The virtual agent was designed to display three different emotional expressions and apply two fixed negotiation strategies, for six combinations of emotional expression and negotiation strategy. During the experiment, we had participants engage in several rounds of negotiations with the virtual agent inside the fMRI scanner. We then analyzed their brain patterns during the decision-making period using Multi-Voxel Pattern Analysis (Norman, Polyn, Detre, \& Haxby, 2006).

We focused our analyses on the anterior insula and the frontal pole. Anterior insula is a well-known emotion-related brain region that is consistently activated when processing basic emotions such as anger and sadness, as well as social emotions such as empathy and vicarious emotions (Kober et al., 2008; Lamm \& Singer, 2010). It has been repeatedly reported that both observing the emotional facial expression and feeling the emotion activate the anterior insula (Zaki, Davis, \& Ochsner, 2012). On the other hand, the frontal pole is one of well-known decision-making-related brain regions. It has been shown that frontal pole plays a significant role in thinking about the future (Okuda et al., 2003), and people with frontal pole impairment make disadvantageous decisions (Anderson, Bechara, Damasio, Tranel, \& Damasio, 1999).

We hypothesize that an agent's negotiation strategy can be predicted based on the activity in the frontal pole, and an agent's emotional expression can be predicted based on the activity in the anterior insula. To validate these hypotheses, we compared participants' brain activities in these regions during negotiations in terms of an agent's emotional expression (angry, neutral, and sad), and an agent's negotiation strategy (conceding and non-conceding).

Our research contributes to a fast growing field of humanagent interaction, and is one of the first lines of work that investigates the underlying neural systems involved in the process of human-agent negotiation. This paper is organized as follows. First, we introduce our negotiation platform, the Objects Negotiation Task. Next, we explain our experimental settings and the design of our virtual agent. Then, we describe the parameters used to record fMRI data and how this data were analyzed. Finally, we show our results and discuss our findings.

## Objects Negotiation Task

The Objects Negotiation Task is a multi-round human-agent negotiation platform (Dehghani, Carnevale, \& Gratch, 2014) where a human negotiator and a virtual agent can negotiate diverse items over multiple rounds. We used a modified version of this task tailored for use in the fMRI. Common fruits were used as negotiation items, and the payoff for each item for both players were explicitly specified on the screen. In order to make sure all participants had the same goal during
negotiations, they were asked to focus on maximizing their total payoffs. To ease the calculation of total payoffs, the system automatically calculated the player's total payoff as well as the agent's total payoff, and displayed whenever the items are redistributed.

When the negotiation starts, items are placed in the middle row indicating that they do not belong to anyone. After the participant distributes all the items, the 'Go!' button on the right bottom corner is enabled and the participant can propose his/her offer by clicking the button. The participant is then asked to wait until the virtual agent accepts or rejects the offer. If accepted, both parties get the proposed items and the negotiation ends. If rejected, the participant is asked to wait for the virtual agent to propose a counteroffer. Next, the virtual agent's offer is shown and the participant reviews the offer for five seconds. During this time, the participant simply observes the counteroffer and cannot relocate the items. For the fMRI version of the task, this review time was introduced to make sure that we can separate brain activity between the offer-making period and the non-offer-making periods. After the review, the participant decides whether to accept or reject the virtual agent's offer. If he/she accepts the virtual agent's offer, the negotiation ends. If rejected, the participant is again asked to wait for five seconds and then is redirected to the first step. Figure 1 shows an example of the timeline of the Objects Negotiation Task.

In our study, we used six types of agents characterized by the three types of emotions they expressed and the two types of offer sets representing their negotiation strategies. More details about these features are described in the following two sections.

## Agent's Emotional Expressions

The role of emotional displays in negotiation has been extensively documented (Lerner, Li, Valdesolo, \& Kassam, 2015). To find the neural mechanisms involved in processing different emotional displays in human-agent interactions, we used three types of facial expressions to express agents' emotions; angry, neutral (no emotion) and sad. Figure 2 shows agents' emotional expressions that were used in the experiment. In angry and sad conditions, the virtual agent's face starts as neutral and changes to the emotional expression for five seconds on the first, third, and fifth rounds of negotiation. In the neutral condition, the agent's face starts as neutral and does not change.

## Agent's Negotiation Strategies

We used two sets of pre-programmed agent offer strategies: non-conceder and conceder. In the non-conceder strategy, the agent starts with no concession and continues with gradually increased concession. In the conceder strategy, the agent starts with some concessions and keeps conceding further in the next rounds.

When the virtual agent decides whether to accept or reject a participant's offer, the agent calculates the summed payoffs and compares it with its next offer. If the summed pay-


Figure 1: Timeline of the Objects Negotiation Task used in the fMRI experiment. During a negotiation, a participant and a virtual agent take turns in making a proposal. If the proposed offer is accepted, the negotiation ends. If rejected, the player who rejected the offer makes a new proposal.


Figure 2: Agent's emotional expressions. Angry (left), neutral (middle), and sad (right).
offs are larger than the summed payoffs of the next offer, the agent accepts the offer. Otherwise, the agent rejects the offer and proposes a new offer. The same payoff matrix was used across all six negotiation tasks so that we could control for the potential effect of varying payoff values. However, we randomized the order of items shown on the screen in each task to give participants the impression that they were playing a new negotiation task every time.

## Experiment

## Participant

We recruited ten participants through an online bulletin board at the University of Southern California. Prior to the study, all participants completed a checklist to make sure they were eligible to take part in an MRI study. All procedures were approved by the USC Institutional Review Board and participants were provided with a written informed consent for the study. One participant's data was later excluded from the analyses because of technical problems with the obtained images.

## Procedure

Upon arrival, participants completed the informed consent and each participant was asked to read the following hypothetical scenario:

You are a restaurant owner in a small town. There has been a major fire in the market providing the necessary fruits for your restaurant and as a result only a limited number of fruits are available. Because of this you have to split the available fruits with another restaurant owner. You and the other owner value each fruit differently. In order to run your restaurant you need to get as many fruits as possible.

In the task that follows, you will negotiate about how to distribute the fruits between you and the other restaurant owner.

After reading the scenario, participants were first invited to play a trial negotiation so that they got used to the interface of the task. Then participants performed six negotiation tasks inside an fMRI scanner. The total scan time for each participant was approximately 45 minutes.

## Data Analysis

To study brain patterns during negotiations with various agents, we analyzed participants' brain activity using general linear models (GLM) analysis and multi-voxel pattern analysis (MVPA).

## General Linear Model Analysis

To extract brain activity of the offer-making period from the whole negotiation period, we ran a general linear model (GLM) analysis using tools from the FMRIB's Software Library (FSL) (Smith et al., 2004). First, we pre-processed
our fMRI data using FSL to reduce the noise of our dataset. Data pre-processing included the following mostly standard steps: (1) Motion-correction with MCFLIRT to fix head motion artifacts during scans, (2) Slice timing correction for interleaved acquisitions to compensate for timing difference between slices of functional images, (3) Non-brain structures removal with Brain Extraction Tool to remove non-brain regions, such as the scalp, (4) Spatial smoothing with a Gaussian kernel of full width at half maximum 5 mm to increase statistical power by improving the signal to noise ratio, (5) High-pass temporal filtering to let high frequencies, such as activities relevant to experimental conditions, pass and to remove low frequencies such as signal drifts.

For each negotiation for each participant, changes in the blood-oxygen-level dependent signal were modeled with regressors for the offer-making period. Then the regressors were convolved with a double-gamma hemodynamic response function. The non-offer-making periods were modeled as baseline.

## Multi-Voxel Pattern Analysis

When analyzing fMRI data, it is important to take into account the activity of the voxels, as well as the interactions between them because the activity in neighboring voxels is interdependent. However, given the univariate nature of GLM, the model fits to each voxel's time-course separately. To overcome this drawback, we used multi-voxel pattern analysis (MVPA) (Norman et al., 2006) which uses patternclassification techniques to extract the pattern of response across multiple voxels.

We preprocessed the GLM analysis results and used them as inputs for the MVPA. The preprocessing included linear de-trending which removes any bias resulting from scanner drift over the acquisition time, and z-scoring which normalizes the range of each voxel. We used leave-one-participantout cross-validation for MVPA, in which a classifier is trained on eight participants' data and then tested with the last participant's data.

Previous studies have shown that the anterior insula (AI) is activated when processing emotions (Kober et al., 2008; Lamm \& Singer, 2010), and the frontal pole (FP) is activated when making a decision that affects the future (Okuda et al., 2003). We therefore hypothesized that the agent's negotiation strategy can be predicted based on the participant's brain activity in the FP, and the agent's emotional expression can be predicted based on the participant's brain activity in the AI. To test our hypothesis, we chose the AI and the FP as our regions of interest (ROIs), and performed ROI MVPA. In the following section, we explain the ROI MVPA approach.
Region of Interest Multi-Voxel Pattern Analysis To find the relationship between an agent's expressed emotion and brain activity as well as an agent's negotiation strategy and brain activity, we performed region of interest (ROI) MVPA with both the AI and the FP. The AI on each side of the brain can be divided into two subregions with distinct patterns of
connectivity: dorsal anterior insula (dAI), connected to the dorsal anterior cingulate cortex; and the ventral anterior insula (vAI), connected to the pregenual anterior cingulate cortex (Deen, Pitskel, \& Pelphrey, 2011). We ran ROI analyses for all four AI regions. On the contrary, the FP does not have widely accepted subregions (Moayedi, Salomons, Dunlop, Downar, \& Davis, 2014). Thus, we ran ROI analysis for the whole FP labeled by the Harvard Center for Morphometric Analysis (Desikan et al., 2006).

To make sure brain activity in the AI or the FP is responsible either for agent's emotional expressions or negotiation strategies, we ran ROI analyses for both conditions, i.e., we calculated the prediction accuracy of agent's negotiation strategies using both the AI and FP as ROIs. We assumed that the prediction accuracy with their expected ROI would be significantly higher than the chance level, but the prediction accuracy with their unexpected ROI would be indistinguishable from chance.

We trained a linear Support Vector Machine (SVM) classifier using voxels from each of our ROIs separately using feature selection. Feature selection is a common approach used to reduce the number of features (voxels) by selecting only relevant features as input to a classifier. Classification performance improves with feature selection as it only picks features that vary significantly between categories (Guyon \& Elisseeff, 2003). In our analyses, we used the GLM analysis results to compute the $F$-score for each voxel, and then used an analysis of variance (ANOVA) measure to select the top $10 \%$ of features with the highest F-scores.

Each participant's brain was transformed into standard MNI space (Evans et al., 1993) to have a brain that is more representative of the population. After performing this process for all participants, individual-level analyses were combined for a group-level analysis.

## Results

As discussed previously, the AI is a brain region known to respond to emotional expressions, and it can be divided into two subregions with distinct patterns of connectivity. Therefore, we first ran ROI MVPA for all four (left/right $\times$ ventral/dorsal) AI regions separately. The prediction accuracies of emotional expression using our ROI MVPA with four AI regions indicate that ROI MVPA with the L-dAI has the best prediction accuracy ( $38.40 \%$ ), while the prediction accuracy of other AI regions are indistinguishable from the chance level of $33 \%$. A binomial test revealed that the prediction accuracy of the ROI MVPA with the L-dAI is significantly above chance ( $p=0.0566$ ). Therefore, we focus our analysis on the results of ROI MVPA with the L-dAI.

The prediction accuracy for decoding an agent's emotional expressions is higher for the L-dAI decoder ( $38.40 \%$, compared to $33.87 \%$ for the FP $(p=0.0960)$ or chance ( $p=$ 0.0566 ). The prediction accuracy using the FP ROI MVPA was not different from chance ( $p=0.4266$ ). This supports our hypothesis that an agent's emotional expression can be

Table 1: Prediction accuracy on an agent's emotional expression and an agent's strategy from ROI MVPA. One-tail binomial tests were performed for each condition compared to the chance level, and significant results $(p<0.05)$ were marked with (*).

| Region of <br> Interest (ROI) | Condition | Prediction <br> Accuracy | Chance <br> Level |
| :---: | :---: | :---: | :---: |
| Left Dorsal <br> Anterior Insula <br> (L-dAI) | All Emotions | $38.40 \%\left(^{*}\right)$ |  |
|  | Angry | $45.07 \%\left(^{*}\right)$ | $33.33 \%$ |
|  | Neutral | $29.60 \%$ |  |
| Frontal Pole <br> (FP) | All Strategies | $40.53 \%\left(^{*}\right)$ |  |
|  | Conceder | $47.96 \%\left(^{*}\right)$ | $50.00 \%$ |
|  | Non-conceder | $70.42 \%\left(^{*}\right)$ |  |

reliably predicted using brain activity in the AI which is an emotion-related brain region, but not with information in the FP.

Similarly, the accuracies for predicting an agent's negotiation strategy using the ROI MVPA with the FP was $58.96 \%$. A binomial test again confirmed that this performance is significantly higher than the chance level of $50 \%(p=0.0085)$. The prediction performance was almost at chance (52.71\%) for the L-dAI ( $p=0.2581$ ). This result validates our hypothesis that counterpart's negotiation strategy can be predicted based on brain activity in the FP, which is activated when people do active decision-making, but not with the insula, which is involved in emotion processing.

In order to examine these results in more detail, we broke down the predictions. Specifically, we analyzed the prediction accuracy for each agent's emotional expression and negotiation strategy from ROI MVPAs with anterior insular regions and frontal lobe (Table 1). Our results indicate that brain patterns in the L-dAI can predict angry and sad conditions but not neutral agent facial expressions. Also, brain patterns in the FP can predict the non-conceder negotiation strategy but not the conceder strategy.

Overall, our results confirm that negotiating with different types of agents results in activity in different brain regions, and these activity patterns can be used to further decode the specific type of interaction agent.

## Discussion

The neuroscience of human-agent interactions is a rarely studied topic and the majority of studies treat the processes by which various agent features affect human participants as a black box. To answer the question of how these features interact with underlying neural mechanisms, we investigated brain activity during human-agent interactions. More specifically, participants engaged with virtual agents who showed three different emotional expressions (angry, neutral and sad) and used two different types of negotiation strategies (conceding and non-conceding). Using a human-agent negotiation platform, participants interacted with virtual agents in an fMRI
scanner, and their brain activity during the interaction was recorded. We hypothesized that an agent's emotional expression could be predicted based on patterns in emotion-related brain regions, and an agent's negotiation strategy could be predicted based on patterns in decision-making-related brain regions. Therefore, we focused our analyses on the AI and the FP, as previous studies have shown that AI is activated when people engage in emotional tasks, and the FP is activated when people perform active decision-making tasks.

Our ROI MVPA results support our hypothesis; prediction accuracy of an agent's emotional expression based on brain patterns in the L-dAI, and that of agent's negotiation strategy based on brain patterns in the FP are well above the chance level. These results indicate that different features are likely processed in different brain regions. Finding which information is processed in certain brain regions would allow us to reliably decode the feature of the agent from users' brain activity. More detailed analyses revealed that brain patterns in the L-dAI could be used to predict angry ( $45.07 \%$ ) and sad ( $40.53 \%$ ) conditions, but not the neutral condition ( $29.60 \%$ ). This indicates that there are clear differences in brain patterns in the L-dAI between angry and sad conditions. We hypothesize that the reason why the patterns in this region failed to predict the neutral condition is that the neutral facial expression is the default expression throughout the experiment. The facial expression of the agent only changes when it morphs into sad or angry. We plan to tackle this problem by only showing the agent's face during the decision making phase.

With regard to agent's negotiation strategies, predictions using brain patterns from the FP showed significantly higher accuracy compared to the chance level ( $50 \%$ ) for the nonconceding condition ( $70.42 \%$ ), but not for the conceding condition $(47.50 \%)$. We assume that this is because participants expected to deal with a counterpart that acted like a conceding and fair agent, i.e. an agent who might start with a slightly unfair offer but over time it makes adjustments toward a fair offer. It is possible that the distinct patterns in the FP witnessed during negotiations with the non-conceding agent is because this agent acts in a very greedy and tough way that is not typical in social interactions. This could result in unique patterns of activity in the FP.

While our sample size could be considered small, we would like to note that sample size tends to be small in fMRI studies. Also, it is worth mentioning that the probability of finding the same effect as one found in the original experiment is not dependent on sample size, but dependent on $p$ value (Killeen, 2005). This is because large effect sizes produce significant results, even with small sample size.

In conclusion, our results indicate that there is a link between an agent's emotional expression and brain activity in the L-dAI, and also between an agent's negotiation strategy and brain activity in the FP. Even though the results are preliminary, our work sheds light on the links between certain brain regions and different agent features. In future studies, we plan to continue investigating these links with other
features such as voice tone and gestures, and hopefully over time construct a map of the brain regions activated by various agent features and compare these regions to human-human interactions.

## Acknowledgments

This research is supported by AFOSR FA9550-14-1-0364.

## References

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., \& Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. Nature neuroscience, 2(11).
Bickmore, T., Schulman, D., \& Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. Applied Artificial Intelligence, 24(6).
Cassell, J., \& Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. Applied Artificial Intelligence, 13.
Castellano, G., Pereira, A., Leite, I., Paiva, A., \& McOwan, P. W. (2009). Detecting user engagement with a robot companion using task and social interaction-based features. In 2009 international conference on multimodal interfaces.
Das, R., Hanson, J. E., Kephart, J. O., \& Tesauro, G. (2001). Agent-human interactions in the continuous double auction. In International joint conference on artificial intelligence (Vol. 17).
Deen, B., Pitskel, N. B., \& Pelphrey, K. A. (2011). Three systems of insular functional connectivity identified with cluster analysis. Cerebral Cortex, 21(7).
Dehghani, M., Carnevale, P. J., \& Gratch, J. (2014). Interpersonal effects of expressed anger and sorrow in morally charged negotiation. Judgment and Decision Making, 9(2).
Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... others (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage, 31(3).
Evans, A. C., Collins, D. L., Mills, S., Brown, E., Kelly, R., \& Peters, T. M. (1993). 3d statistical neuroanatomical models from 305 mri volumes. In Nuclear science symposium and medical imaging conference, 1993.
Grosz, B. J., Kraus, S., Talman, S., Stossel, B., \& Havlin, M. (2004). The influence of social dependencies on decisionmaking: Initial investigations with a new game. In Proceedings of the third international joint conference on autonomous agents and multiagent systems-volume 2.
Guyon, I., \& Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3.
Jurafsky, D., Ranganath, R., \& McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. In Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics.

Karacora, B., Dehghani, M., Krämer-Mertens, N., \& Gratch, J. (2012). The influence of virtual agents? gender and rapport on enhancing math performance. In Proceedings of the 34th annual meeting of the cognitive science society.
Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. Psychological science, 16(5).
Kim, E., Dehghani, M., Kim, Y. K., Carnevale, P. J., \& Gratch, J. (2014). Effects of moral concerns on negotiations. Proceedings of the 36th Annual Meeting of the Cognitive Science Society.
Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., \& Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: a metaanalysis of neuroimaging studies. Neuroimage, 42(2).
Lamm, C., \& Singer, T. (2010). The role of anterior insular cortex in social emotions. Brain Structure and Function, 214(5-6).
Lerner, J. S., Li, Y., Valdesolo, P., \& Kassam, K. S. (2015). Emotion and decision making. Psychology, 66.
Maldonado, H., Lee, J.-E. R., Brave, S., Nass, C., Nakajima, H., Yamada, R., .. Morishima, Y. (2005). We learn better together: enhancing elearning with emotional characters. In Proceedings of th 2005 conference on computer support for collaborative learning.
Moayedi, M., Salomons, T. V., Dunlop, K. A., Downar, J., \& Davis, K. D. (2014). Connectivity-based parcellation of the human frontal polar cortex. Brain Structure and Function.
Norman, K. A., Polyn, S. M., Detre, G. J., \& Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. Trends in cognitive sciences, $10(9)$.
Okuda, J., Fujii, T., Ohtake, H., Tsukiura, T., Tanji, K., Suzuki, K., ... Yamadori, A. (2003). Thinking of the future and past: The roles of the frontal pole and the medial temporal lobes. Neuroimage, 19(4).
Rafaeli, A., \& Sutton, R. I. (1987). Expression of emotion as part of the work role. Academy of management review, 12(1).
Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., \& Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. Science, 300(5626).
Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... others (2004). Advances in functional and structural mr image analysis and implementation as fsl. Neuroimage, 23.
Van Kleef, G. A., De Dreu, C. K., \& Manstead, A. S. (2004). The interpersonal effects of anger and happiness in negotiations. Journal of personality and social psychology, 86(1).
Wang, N., \& Gratch, J. (2009). Can virtual human build rapport and promote learning? In Artificial intelligence in education conference.
Zaki, J., Davis, J. I., \& Ochsner, K. N. (2012). Overlapping activity in anterior insula during interoception and emotional experience. Neuroimage, 62(1).

# Decoding Partner Type in Human-Agent Negotiation using functional MRI 

Eunkyung Kim (eunkyung@usc.edu) ${ }^{1}$, Jared Gilbert (jaredgil@usc.edu) ${ }^{1}$, Charlotte Horowitz (shchar@umich.edu) ${ }^{2}$, Jonathan Gratch (gratch@ict.usc.edu) ${ }^{3}$, Jonas T. Kaplan (jtkaplan@usc.edu) ${ }^{1}$ and Morteza Dehghani (mdehghan@usc.edu) ${ }^{1}$<br>${ }^{1}$ Brain and Creativity Institute, University of Southern California, Los Angeles, CA 90089<br>${ }^{2}$ Department of Psychology, University of Michigan, Ann Arbor, MI 48109<br>${ }^{3}$ Institute for Creative Technologies, University of Southern California, Playa Vista, CA 90094


#### Abstract

People interact differently with humans than they do with computers, but there is minimal research on what brings about these differences. Using agents labeled as either "another participant" or a "computer program", we investigated the differences in people's behavior and brain activity during the course of a negotiation paradigm. Our results indicate that people perceive human-labeled agents more human-like than computerlabeled agents, and the level of concession in the negotiations is dependent on agent type. We have also found that these differences can be captured in brain activation by showing that parts of the Theory of Mind neural correlates are activated in human-labeled agent conditions, but not in computer-labeled agent conditions. We further demonstrate that brain activity can predict whether the negotiation agent was introduced as a competing human player or a computer program. Overall, our study suggests that labeling an interaction partner as either another human or a computer program leads to significant impacts on one's decision making.


Keywords: Human-Agent Interaction; Negotiations; fMRI

## Introduction

Can computer agents act as substitutes for human beings during the course of an interaction? This has been a popular topic in sci-fi movies for decades. In fact, some computer agents that were thought to exist only in movies a few decades ago, are now widely used in daily life. Smartphones, for example, are commonly used to execute voice commands through programs like Siri, and movie theaters now have more ticket vending machines than guest personnel.

However, human-agent interactions are often quite different from human-human interactions (Gray, Gray, \& Wegner, 2007; Melo, Marsella, \& Gratch, 2016), and many factors contribute to these differences. Researchers have continuously tried to identify what these disparities are and why they occur, with the hopes to bridge the gap between humanhuman and human-robot/agent encounters.

A robot's appearance has been found to be paramount in the interaction style of the human subjects. For example, when people interface with robots that have mechanical, nonhuman like features, even when the robot performs human-like actions, they are often unable to overlook these traits (Hegel, Krach, Kircher, Wrede, \& Sagerer, 2008). Thus, robots designed to have eyes similar to humans (Banh, Rea, Young, \& Sharlin, 2015), or baby face-like heads (Powers \& Kiesler, 2006), were found to be more effective in evoking a more human-like interaction.

These differences have also been extensively studied using brain imaging techniques, especially regions associated
with Theory of Mind (ToM), due to their importance in social interactions. ToM refers to the ability of one person to reason about another person's mental states, including their intentions and beliefs (Premack \& Woodruff, 1978). Previous fMRI studies have demonstrated that cortical activity in the neural structures related to ToM tend to be more active when participants were told they were facing a human partner compared to a computer program (Kircher et al., 2009). Research also demonstrated that activity in these same regions scaled according to the human-likeness of their interaction partner when using computer-animated characters or nonhuman agents (Chaminade, Hodgins, \& Kawato, 2007; Krach et al., 2008).

Many previous behavioral and neuroimaging studies have used negotiation platforms to examine human-agent communication, because negotiations involve complex cognitive effort and established social interaction techniques. For example, studies show that when participants play the Ultimatum Game against computer partners, they are more likely to accept unfair offers compared to when they play against human partners, where they tend to be less willing to accept offers of unequal value (Sanfey, Rilling, Aronson, Nystrom, \& Cohen, 2003). Studies have also shown that people have stronger emotional reactions to unfair offers made by other humans (Vant Wout, Kahn, Sanfey, \& Aleman, 2006).

Using a multi-round, multi-object negotiation platform for our research, we explored whether a computer agent introduced as another human was perceived more anthropomorphically than one that was introduced as a computer program. We then investigated whether agent type produced behavioral differences, and whether one type of agent resulted in more concessions compared to the other. In a follow up experiment, we compared brain activity during interactions with humanlabeled and computer-labeled agents to determine whether these perceptual differences were also observable in brain patterns. Following collection of fMRI data, we investigated whether classifiers could be trained to determine whether the participant was playing against a human-labeled or computerlabeled agent.

We hypothesized that participant behavior and brain activity would be different during interactions with humanlabeled agents, compared to interactions with computerlabeled agents, even though both agents used exactly the same strategies and emotions. Our initial experiment consisted of an online negotiation task intended to explore perceptual and


Figure 1: Objects Negotiation Task Timeline
behavioral differences pertaining to anthropomorphic characteristics in human and computer-labeled agents. Next, we adapted our negotiation framework into an fMRI experiment, attempting to find neural differences for the two distinct partner conditions. In addition to these studies, we also ran a prediction algorithm and multi-voxel pattern analysis (Norman, Polyn, Detre, \& Haxby, 2006) based on the fMRI data.

This work is distinct from previous studies due to the use of an identical computer agent, regardless of what partner type was specified. The majority of previous studies employed computer-animated characters or robots that had differing levels of anthropomorphism. We demonstrate that even though the same computer agent is used, perceptual differences were captured in behavioral and brain data. To the best of our knowledge, this is one of the first lines of research that uses a multi-round negotiation platform to investigate perceptions of anthropomorphism. We believe that natural interactions take place over multiple rounds/sessions, and it is therefore important to investigate perception differences through multi-round negotiations.

This paper is structured as follows. First, we introduce and explain the Object Negotiation Task, the platform used for both experiments described. Next, we outline an exploratory, behavioral experiment performed to examine the differences between interactions with a human-labeled agent and a computer-labeled agent. After which we discuss our fMRI experiment, using the same paradigm as the first, trying to identify differences in brain activity. Lastly, we discuss the implications of our results and future work.

## Objects Negotiation Task

The Objects Negotiation Task is a web-based multi-round negotiation task where a participant and a computer agent can take turns distributing objects (Dehghani, Carnevale, \& Gratch, 2014). The original version was designed for behav-
ioral data collection only, so in this paper, a modified version of this task was used for collection of both behavioral and brain data. Figure 1 shows the timeline of the modified version of this task. Some of the modifications made included an emotion-reporting phase and offer-review phase, which were added to separate collection of brain data between differing phases. In addition, a partner introduction phase was added, allowing participants to receive a notification specifying whether their partner type was another participant or the computer program before the negotiation began.

The sequence of the modified Objects Negotiation Task is as follows. When the task begins, the negotiation partner type is displayed. Specifically, in the human-labeled agent condition, the message shown to the participant is 'In this task, you will be negotiating with the other participant.' In a computer-labeled agent condition, the same message was shown but 'the other participant' was changed to 'a computer program'. Next, a 'connection establishment' message for the human condition and a 'program setup' message for the computer condition appear on screen, to persuade participants of their partner setting. Throughout the negotiation, the partner type is constantly included on screen so the participant clearly recognizes his/her partner type. The partner is labeled as 'the other participant' or 'the computer program.'

In the first negotiation round, items are positioned in the middle row, indicating that those items belong to neither player. The participant is asked to propose an initial offer by moving items into his/her own set of boxes (bottom row) or their partner's set of boxes (top row). Once the initial offer is made, the partner (agent) chooses an emotion pertaining to the offer which is then displayed to the participant. Available emotions include: happy, content, neutral, angry and sad. The partner only shows the predefined emotion for each round. After the emotion is displayed, the partner decides whether to accept or reject the offer. This decision is based on a pre-
defined offer value; when the payoffs of the predefined offer are more than the participant's current offer, the partner rejects, when the payoff is less, the partner accepts.

If the participant's offer is accepted, the items are distributed as proposed and the participant is notified. If the participant's offer is rejected, the partner then proposes a counteroffer. When the counteroffer is received, the participant has 5 -seconds for review. During this time, the participant can only observe; no items can be transferred. The review time was specifically introduced for optimal brain activation, as we wanted to record an active decision-making process. For the same reason, our analysis was focused on data collected during this review phase. After the review phase, the participant reports his/her emotion about the proposed offer by choosing from the following descriptive options: happy, content, neutral, angry, or sad. The participant also decides whether to accept or reject the offer. If the participant rejects the offer, a new round begins, and all phases are repeated. The negotiation can last for a maximum of six rounds. If no agreement is made in six rounds, neither party receives anything.

## Study 1: Online Experiment

We designed an online experiment to determine whether people perceive human-labeled agents differently than computerlabeled agents during interactions in our negotiation game, as well as to find behavioral differences between agent type in concession-making.
Negotiation Partners Two sets of strategies and two types of emotions were used for the partner agents. Agent strategies included tough and soft. A tough strategy starts with a greedy offer, and a soft strategy starts with a relatively generous offer. Figure 2 shows payoffs for the agent and the participant when the agent uses tough or soft strategies.

Agent emotions included anger and neutral (no emotion). Anger was chosen because it was found to be the most effective emotion in yielding concessions during negotiation tasks (Van Kleef, De Dreu, \& Manstead, 2004). For the anger condition, the agent displayed an angry face in rounds 2,4 , and 6 , and a neutral face in rounds 1,3 , and 5 . For the neutral condition, the agent reported a neutral face in every round.
Procedure 420 subjects ( 237 male and 183 female; mean age $=33.5$ ) living in the United States were recruited via Amazon Mechanical Turk (MTurk). Each participant was asked to read a hypothetical scenario in which they acted as a restaurant owner, and negotiated for fruit with another restaurant representative due to a fruit shortage as a result of a recent fire in a local market. Each subject was then told to negotiate with either a computer program or another (hypothetical) MTurk player. Regardless of type label, the negotiation partner was always a pre-programmed computer agent. After completing all negotiations, subjects were asked to fill out an anthropomorphism questionnaire (Bartneck, Kulić, Croft, \& Zoghbi, 2009) about their partner, as well as a demographic


Figure 2: Payoffs for agents and participants across both agent strategies.


Figure 3: Anthropomorphism Scores for human-labeled agents and computer-labeled agents. Higher score means the agent is perceived as more human-like. The error bar shows standard errors.
questionnaire. In the anthropomorphism questionnaire, participants rated their impression of their partner using a scale from 1 to 7 , where 7 means human-like and 1 means machinelike. Subjects were also given a simple attention-check question, implemented to make sure the participants were paying attention; it merely asked what type of partner they were assigned during the task. Each participant was compensated $\$ 1$.
Data Analysis We excluded subjects who had participated in our previous negotiation studies or failed to give the correct answer to the attention-check question. After exclusion, we had data from 329 subjects. Scores from each condition were calculated for the anthropomorphism questionnaire to verify whether participants perceived human-likeness differently between agents. In addition, we calculated concessions across partner type in each condition to analyze behavioral differences. Concession was calculated by subtracting payoffs of agreed offers from payoffs of initial offers. A three-way between-subjects analysis of variance (ANOVA) was used to find the interaction between partner type, partner strategy, and partner emotion during concession.
Results The anthropomorphism scores of the humanlabeled agents and the computer-labeled agents are shown in Figure 3. 1-way ANOVA results show that people consis-


Figure 4: Concessions to human-labeled agents and computer-labeled agents. The error bars show standard errors.
tently thought their partner to be more human-like when told their partner was a human player, no matter what the negotiation strategy $(\mathrm{F}(1,327)=14.09, p<0.001)$.

Concessions to human-labeled agents and computerlabeled agents during negotiations are shown in figure 4. ANOVA results show that there is a $2 \times 2 \times 2$ interaction between agent type (human/computer) $\times$ agent emotion (angry/neutral) $\times$ agent strategy (tough/soft) for concession $(\mathrm{F}(1,321)=3.387, p=0.066)$. We also ran a $2 \times 2$ ANOVA after dropping each strategy. A two-way interaction was found for tough strategy $(\mathrm{F}(1,164)=4.699, p=0.031)$.
Discussion Our findings from anthropomorphism scores suggest that there are perception differences in the interactions between human-labeled agents and computer-labeled agents. Also, our ANOVA results suggest that there is an interaction between agent type $\times$ agent emotion $\times$ agent strategy for concession. This indicates that not only are people's perceptions of the two agents distinct, but their behaviors also vary depending on agent type. To study whether these behavioral differences have neural correlates, we designed the following fMRI experiment. Because the largest concession differences were found in the tough conditions, implying the tough strategy was best suited to observe those differences in behavior, we mainly employed tough agents in the following experiment.

## Study 2: fMRI Experiment

We hypothesized that perceptual and behavioral differences could be captured in brain activity, especially in ToM related brain regions, as they were found to be correlated with human-likeness of physically existing human-like robots (Chaminade et al., 2007; Krach et al., 2008). Each subject performed the negotiation task with both types of agent in order to compare brain activity from human-labeled vs. computer-labeled agent interaction. To make interactions with human-labeled agents more realistic, we introduced a confederate into the study, so participants believed they would be competing against another human player.

Participants 20 healthy American subjects (10 male and 10 female), recruited via the University of Southern California online bulletin board, took part in this study. Subjects were 21.4 years old on average ( $\mathrm{SD}=2.58$ ). All participants were right-handed and had no history of neurological or psychiatric disorders.

Negotiation Partners Although we were only interested in tough agent type, we used two types of soft agents on top of four types of tough agents (human/computer-labeled $\times$ angry/neutral agents). This is modification was implemented because subjects participated in a series of consecutive negotiations, unlike our online experiment where each subject only played in a single negotiation. Including soft agents ensured that participants did not play with the same agent over and over again. Each subject negotiated with six types of agents. While every subject negotiated with four types of tough agents, 10 subjects ( 5 male and 5 female) negotiated with two types of emotion-neutral soft agents, while the remaining 10 subjects negotiated with two types of emotionangry soft agents. Agent order was randomized.

Procedure Each participant was greeted by an experimenter and introduced to the confederate as the competing player. The participant and the confederate were guided to a preparation room where they filled out an informed consent form, incidental findings form, and safety screening form. After forms were completed, the confederate was guided to a separate MRI room for "setting up". The participant was given the instructions and rules regarding the negotiation task, and played a trial negotiation against a computer program before starting the experiment. During the trial, a trackball mouse similar to one used in the scanner environment was provided, so that the participant became familiarized with it's operation. The participant was then guided to the actual MRI room and was told that while in the scanner he/she would run through three negotiation tasks with the participant in the other MRI room, and three negotiation tasks with the computer program. The task was back projected on a screen, seen through a mirror attached to the head coil, and operated a trackball mouse to navigate negotiations. In each task, a different set of negotiation items were used, and payoffs for these items varied with position, in order to give an impression that each negotiation was unique. Participants answered a shortened version of the anthropomorphism questionnaire at the completion of each round. After a maximum of six negotiation rounds, participants filled out a handedness and demographic questionnaire. Before leaving, subjects were debriefed and compensated $\$ 30$.
fMRI Data Acquisition fMRI scans were performed at the USC Dana \& David Dornsife Cognitive Neuroscience Imaging center. Images were acquired using a 3-Tesla Siemens PRISMA MRI scanner with a 20-channel matrix head coil. Two sets of high-resolution anatomical images were acquired for registration purposes. Six sets of echo-planar images (EPI), one set for each negotiation, were acquired continu-


Figure 5: Frontal medial cortex from Harvard-Oxford atlas (left) and overlaid accuracy map for all participants (right). A part of frontal medial cortex was included in the accuracy map (white dotted box).
ously with the following parameters: $\mathrm{TR}=2,000 \mathrm{~ms}, \mathrm{TE}=$ 25 ms , flip angle $=90^{\circ}, 64 \times 64$ matrix, one shot per repetition, in-plane resolution $3 \times 3 \mathrm{~mm}^{2}$, 41 transverse slices, each 3 mm thick, covering the whole brain. Total scan time for each participant was approximately 50 minutes.
Data Analysis We conducted a general linear model (GLM) analysis and then used the results as input for multivoxel pattern analysis (MVPA). GLM analysis was performed using FMRIB's Software Library (FSL) to locate brain regions activated during proposal review phases. As mentioned earlier, this phase was specifically targeted due to the likelihood of collecting data pertaining to decision making for subsequent negotiation rounds. For GLM analysis, data preprocessing steps included motion correction, brain extraction, spatial smoothing, slice timing correction, and high-pass temporal filtering. After completing data pre-processing, we modeled brain activity during proposal review phases using a double gamma hemodynamic response function. Data collection from all other time points were used as baseline. We then performed MVPA to find brain regions that illustrated different patterns across agent type. In MVPA, neural representations were decoded by applying pattern-classification algorithms on fMRI data (Norman et al., 2006). We used detrended and $z$-scored GLM analysis results as inputs for MVPA, and trained a linear Support Vector Machine (SVM) classifier using feature selection, introduced to improve classification performance by picking the most relevant features as inputs for the classifier (Guyon \& Elisseeff, 2003). Searchlight analysis (Kriegeskorte, Goebel, \& Bandettini, 2006) was used as the feature selection method to analyze contents multivariately at each location in the brain. We implemented a leave-one-participant-out cross validation as balance for MVPA. More details on fMRI data analysis can be found in (Kim, Gimbel, Litvinova, Kaplan, \& Dehghani, 2016), as the same analysis methods were used.

Results The right side of Figure 5 shows an overlaid accuracy map for all participants. Accuracy maps were acquired from the top $5 \%$ of all t -maps, which were generated using t-tests versus chance. Interestingly, medial prefrontal cortex, one of the ToM brain regions, was included in the accuracy map, suggesting that people indeed perceived the
human-labeled agent to have more human-like qualities.
MVPA, with searchlight as a feature selection method, revealed that agent type (human/computer) can be predicted based on brain activity during proposal-review phases. Prediction accuracy for agent type was $58.41 \%$, with a standard error $0.01 \%$, where chance level is $50 \%$. The improvement was found to have statistical significance (Two-tailed t-test: $p$ $<0.001$ ).

Discussion The results of Study 2 demonstrates that differences in how we perceive the 'humanness' of an agent can be captured using fMRI. Specifically, our results show that parts of the ToM neural correlates are activated in humanlabeled agent conditions, but not in computer-labeled agent conditions. This finding is consistent with previous studies that reported increased brain activity in ToM brain regions corresponding to human-likeness of interacting partners (Chaminade et al., 2007; Krach et al., 2008). Our MVPA analysis further revealed that these differences are great enough that classifiers can be trained that can reliably distinguish brain activity between the two types of agents.

## Overall Discussion

Our goal was to investigate differences in behavior and brain activity during human-agent negotiations. Focusing on partner type, we hypothesized that both parameters would be distinct when comparing computer-labeled and human-labeled interactions.

Results from our online experiment indicate that people perceive human-labeled agents more human-like than computer-labeled ones, even though both used parallel strategies and emotions. This suggests that people's attitudes towards computer partners are distinguishable from those towards human partners. Furthermore, a 3-way interaction between agent type $\times$ agent emotion $\times$ agent strategy was found for concession.

Results from our fMRI experiment suggest that brain patterns observed during interactions with human-labeled agents are different from ones with computer-labeled agents. More specifically, the medial prefrontal cortex, part of the ToMrelated neural structures, was found to be included in accuracy maps, indicating neural activity during interactions with human-labeled agents are distinct from ones with computerlabeled agents. This is in line with a previous finding, where the medial prefrontal cortex was found to be activated while playing rock-paper-scissors with a human player, but not activated when playing the same game with a known, pre-defined computer algorithm (Gallagher, Jack, Roepstorff, \& Frith, 2002). Using a negotiation paradigm, a more complicated task than rock-paper-scissors due to the inclusion of multiple decision-making rounds, we found that the same effect exists with differently labeled agents. Negotiations require a more substantial cognitive effort than a game like rock-paperscissors; there are a larger amount of possible actions to consider, an increased opportunities for loss, and a greater obligation to compete, or cooperate and come to some sort of
agreement. The negotiation tasks used in the aforementioned experiments are able to advantageously measure interactions that require high levels of cognitive energy, and are therefore more useful when attempting to explore human-robot interactions.

Our findings are also consistent with previous studies that found ToM-related neural structures to be more activated when interacting with agents that had more human-like characteristics, regardless of whether those agents happened to be robots or computer-generated characters (Chaminade et al., 2007; Krach et al., 2008). These findings imply perceptual variance between interactions with human-like and nonhuman-like agents, and leads us to believe that human participants do engage in greater mentalising efforts when faced with human-labeled or human-like robots.

In conclusion, we examined the relationship between negotiation partner type and behavioral and neural measures regarding individual's perceptions of human-agent interactions. Our study suggests that when either by labeling agents as other humans or as computer programs significantly impacts one's perception of the situation; this is ultimately demonstrated through negotiation behavior and brain activity. The results give us further insight into the counter-play between emotional and cognitive processes, leading us to believe that our emotions may have greater impact on decision making than which we are consciously aware. Ultimately, these results inform us that there is a noticeable and consistent difference between the perceptual and emotional reactions that humans have towards other humans when compared to those same reactions with computer agents. Further research needs to be executed to more thoroughly understand why these differences in interaction occur, but this study has illustrated that computers and technology do indeed impact the way humans interact with the world, and each other. This is important to consider as computers continue to be increasingly implemented in everyday life and society.

## Acknowledgments

This research is supported by AFOSR FA9550-14-1-0364.

## References

Banh, A., Rea, D. J., Young, J. E., \& Sharlin, E. (2015). Inspector baxter: The social aspects of integrating a robot as a quality inspector in an assembly line. In Proceedings of the 3 rd international conference on human-agent interaction (pp. 19-26).
Bartneck, C., Kulić, D., Croft, E., \& Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics, 71-81.
Chaminade, T., Hodgins, J., \& Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. Social cognitive and affective neuroscience, 2(3), 206-216.

Dehghani, M., Carnevale, P. J., \& Gratch, J. (2014). Interpersonal effects of expressed anger and sorrow in morally charged negotiation. Judgment and Decision Making, 9(2).
Gallagher, H. L., Jack, A. I., Roepstorff, A., \& Frith, C. D. (2002). Imaging the intentional stance in a competitive game. Neuroimage, 16(3), 814-821.
Gray, H. M., Gray, K., \& Wegner, D. M. (2007). Dimensions of mind perception. Science, 315(5812), 619-619.
Guyon, I., \& Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3.
Hegel, F., Krach, S., Kircher, T., Wrede, B., \& Sagerer, G. (2008). Understanding social robots: A user study on anthropomorphism. In Robot and human interactive communication, 2008. ro-man 2008. the 17th ieee international symposium on (pp. 574-579).
Kim, E., Gimbel, S. I., Litvinova, A., Kaplan, J. T., \& Dehghani, M. (2016). Predicting decision in human-agent negotiation using functional mri. Proceedings of the 38th Annual Meeting of the Cognitive Science Society.
Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., ... Krach, S. (2009). Online mentalising investigated with functional mri. Neuroscience letters, 454(3), 176-181.
Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., \& Kircher, T. (2008). Can machines think? interaction and perspective taking with robots investigated via fmri. PloS one, 3(7), e2597.
Kriegeskorte, N., Goebel, R., \& Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America, 103(10).
Melo, C. D., Marsella, S., \& Gratch, J. (2016). People do not feel guilty about exploiting machines. ACM Transactions on Computer-Human Interaction (TOCHI), 23(2), 8.
Norman, K. A., Polyn, S. M., Detre, G. J., \& Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. Trends in cognitive sciences, 10(9).
Powers, A., \& Kiesler, S. (2006). The advisor robot: tracing people's mental model from a robot's physical attributes. In Proceedings of the 1st acm sigchi/sigart conference on human-robot interaction (pp. 218-225).
Premack, D., \& Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and brain sciences, 1(04), 515-526.
Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., \& Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. Science, 300(5626).
Van Kleef, G. A., De Dreu, C. K., \& Manstead, A. S. (2004). The interpersonal effects of anger and happiness in negotiations. Journal of personality and social psychology, 86(1).
Vant Wout, M., Kahn, R. S., Sanfey, A. G., \& Aleman, A. (2006). Affective state and decision-making in the ultimatum game. Experimental brain research, 169(4), 564-568.

# What Do We Learn from Dyslexia and Second Language Learners on the Difference Between Long-term Frequency and Short-term Sequence Repetition Effects? 

Eva Kimel<br>The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem<br>Itay Lieder<br>The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem<br>Luba Daikhin<br>Department of Psychology, The Hebrew University of Jerusalem<br>Hilla Jacoby<br>Department of Psychology, The Hebrew University of Jerusalem

Atalia Hai Weiss
Department of Psychology, The Hebrew University of Jerusalem
Merav Ahissar
The Edmond and Lily Safra Center for Brain Sciences and Department of Psychology, The Hebrew
University of Jerusalem


#### Abstract

Dyslexia is a common learning disability, but its core deficit is still under debate. The anchoring deficit hypothesis suggests that dyslexics' benefit from experimental stimuli statistics is impaired (e.g. Ahissar, 2007). In this study we asked whether dyslexia is also associated with reduced sensitivity to long-term statistics. Spans for lists of syllables were measured, and indeed, dyslexics benefited less than controls from syllabic frequency. However, dyslexics' benefit from sequence repetition was similar to controls'. In order to dissociate the impact of item familiarity from exposure unrelated factors, native English speakers performed the experiment. They were expected to benefit from repetition, but not from syllabic frequency (in Hebrew). Indeed, that was the case. These data suggest that benefits from long-term distributional statistics are impaired in dyslexia, whereas on-line benefits from sequence repetition are adequate. Moreover, our results suggest different underlying mechanisms for long-term distribution learning and short-term sequence learning.


# "Oops, I did it again." The impact of frequent behaviour on causal judgement 

Lara Kirfel (lara-christina.kirfel.15@ucl.ac.uk)<br>David Lagnado (d.lagnado@ucl.ac.uk)<br>Department of Experimental Psychology, UCL, 26 Bedford Way, London WC1E 6BT


#### Abstract

Current causal theories aim to incorporate the effect of statistical and prescriptive norms on causal judgements, stating that norm-violating actions are judged as more causal than norm-conforming ones. In this paper, we present two experiments that undermine this claim, showing that people attribute increased causality to agents who conform to the norm of frequent behaviour. Furthermore, we find that the time point at which a moral norm is introduced does not make a difference to causal attributions, but that the frequency of a norm violation further accentuates its causal rating. Because these findings present a challenge to current norm theories of causation, we argue for an extended counterfactual model of causal attribution.


Keywords: causal judgement; counterfactual reasoning, frequency; norms; moral judgement

## Introduction

Judgements about causal processes are a crucial skill to make sense of the word. While our ordinary concept of causation was long thought to be entirely objective in character, we know that factors that go beyond the causal facts of a situation, e.g. the moral status of an action, are hugely influential on how we think about the causality of an agent (Halpern \& Hitchcock, 2015; Hitchcock \& Knobe, 2009, Alicke, Rose \& Blome, 2011). For example, if two agents cause a negative outcome by jointly performing the same action, but only one of them is actually allowed to perform this action, people judge the agent who violated the rule as more of a cause (Knobe \& Fraser, 2008). A similar picture arises for the statistical properties of an action. Exceptional or atypical actions (Kahneman \& Miller, 1986), as well as actions with statistically unlikely outcomes (Kominsky et al., 2015) are judged more causal than typical actions or likely outcomes.

All these cases involve circumstances in which the factor that is viewed as a cause is 'abnormal': a person does something immoral, a person does something she usually does not do, or an action has a very unlikely outcome. It has therefore been suggested that our causal judgements are sensitive to 'normality', or 'norms', with the term norms encompassing a variety of norms like moral norms, social rules or statistical norms such frequent behaviour or the likelihood of an outcome. This view holds that normviolating actions, i.e. actions that deviate from norms or normal circumstances, are judged as more causal than normconforming ones (Hart \& Honoré, 1963; Hitchcock \& Knobe, 2009). The underlying assumption is that we assess a causal candidate in terms of its counterfactual relation for the
outcome (Gerstenberg \& Lagnado, 2010; Halpern \& Pearl, 2010) and that norms come into play concerning which counterfactuals we consider. Hence, when reasoning about the cause of an event, norms are thought to make normconforming counterfactuals more relevant (Hitchcock \& Knobe, 2009), more available (Kahneman \& Miller, 1986), more probable to be sampled (Icard \& Knobe, 2016) or to be ranked highest in the order of possible counterfactuals (Halpern \& Hitchcock, 2015).

Common to all these accounts is the idea that in a causal scenario with multiple causal candidates, we will single out the action or event that is norm-deviant, e.g. immoral or unlikely, because we have the tendency to counterfactually simulate an alternative scenario in which normality is in place and no norm is violated (Hitchcock \& Knobe, 2009; Kahneman \& Miller, 1986). In addition, it has been shown that the norm violation of one agent not only increases causal attribution, it can also reduce causal attribution to other, norm conforming agents, known as "causal superseding" (Kominsky et al., 2015). Although this account applies to normality and norms in a very broad sense, studies in causal attribution have mainly investigated the influence of violations of moral norms (Alicke, Rose \& Blome, 2011 Hitchcock \& Knobe, 2009; Knobe \& Fraser, 2008), statistical norms (Kahneman \& Miller, 1986; Kominsky et al., 2015) and norms of proper functioning (Hitchcock \& Knobe, 2009). We will subsume these theories under a general norm-framework of causal judgement.

In the causal cognition literature, frequent, repeated actions have often been considered as belonging to a special kind of statistical norms, so called 'agent-level' statistical norms or norms of 'typical behaviour'. According to the norm framework of causal judgement, actions that deviate from this norm, i.e. atypical actions that are rarely or almost never performed, are seen as more causal than actions that are performed frequently (Knobe \& Fraser, 2008). However, Sytsma and colleagues showed that these 'atypical', i.e. infrequent actions receive reduced instead of increased causal attribution (Roxborough \& Cumby 2014; Sytsma et al., 2012). The authors systematically varied the frequency and morality of actions in the 'pen case' scenario (Knobe \& Fraser, 2008). In this scenario, both a professor and an administrative assistant each take one of two available pens in the department office, with negative consequences. Sytsma et al. varied whether each agent frequently took pens from the office, and whether they were officially allowed to do so. While, as expected, they found that an agent was judged as more causal when violating the department policy, they also
found that an agent was attributed more causality when he frequently took pens rather than only once. This was even the case when there was no department policy prohibiting the use of pens (Sytsma et al., 2012).

By showing that an action that has been repeatedly performed previously receives higher rather than lower causal ratings, the authors provide evidence that agent-level frequent behaviour does not always follow the general schema attributed to the role of norms in causal reasoning. However, the influence of frequent behaviour on causal judgements in this particular case might not come as a surprise. The causal structure of the "pen case" scenario is sensitive towards the frequency of the actions - trivially, the more often a person takes pens, the more they contribute to the pens running out. This raises the question whether the frequency of a particular action, or non-action, also influences our judgement of its causal contribution in a causal structure where the frequency itself does not make a difference to the outcome. The causal situation in the pen case is cumulative: The more often the action is performed, the more likely the final outcome is to happen. However, let's imagine a case where at the end of each day the pen stock gets replenished to two pens. If only one of the agents repeatedly takes a pen on each day, then no problem occurs. But when both agents decide to take a pen on the same day, then there will be no pens left for emergency cases. This case represents a conjunctive causal structure - two actions are needed for the adverse outcome to occur. We were interested whether the frequency of an action still has an impact on our causal judgements in such a scenario:

1. Does the frequency of an action affect causal judgments in a conjunctive causal structure?

Norm-incorporating theories predict increased causal attribution to actions that violate moral norms, but make no prediction as to whether the frequency of the norm violation further accentuates its causal relevance. In terms of counterfactual dependence, if the joint actions of a frequent and a one-off moral norm violator lead to a negative outcome in a conjunctive causal structure, they are both equally causal - if either of them had not acted, the outcome would not have occurred (Halpern \& Pearl, 2010). Furthermore, they are both also 'equally abnormal', because both of them violated the norm in the situation of the final outcome. Current norm incorporating theories leave open whether we mentally undo only the final vs. all previous norm-violating actions in order to assess the causal relevance of a norm-violating agent. This motivated our second research question:
2. Is a frequent vs. one-off norm violation assigned a different amount of causality in a conjunctive causal structure?

Gershman et al. (2016) show that if a certain action leads to a negative outcome in a new context, e.g. the door at the new office breaks when you turn the doorknob clockwise, this action will be judged less blameworthy when it has been
a habitual action before, e.g. when the door knob at home runs clockwise. In fact, we can think of situations in which an action that has been frequently done before is suddenly not allowed anymore, for example smoking in bars after the UK smoking ban in 2007. This raises the question whether a norm-violating act that has been permissively performed before is judged less causal than a repeated norm violation:
3. How does a frequent norm violation compare to an action that has been done frequently, but violated a norm only once?

The two experiments in this paper aim to address these questions.

## Experiment 1

Experiment 1 examined the influence of the "frequency" and "morality" of an action on causal judgments in a conjunctive causal structure. Frequent and moral behavior were varied across the actions of one agent ('Agent 1') and held constant across actions of the other agent ('Agent 2'). In order to manipulate frequent behavior realistically, we used the time frame of a week in which information about the action of the agents is successively presented day by day. Additionally, we varied whether the official norm indicating which agent is allowed to perform the respective action is either introduced right at the beginning of the scenario, or just before the final outcome ("time point of moral norm"). By varying the time point at which the moral norm was introduced we could test actions that either frequently violate a norm, or violate the norm for the first time but have been frequently performed before.

## Methods

Participants 103 participants were recruited via Amazon Mechanical Turk ( $\$ 0.80$ for ca. 15 min ). Participants who answered more than four of the 32 manipulation check questions wrong were removed from the analysis, leaving 81 participants $($ male $=45$, female $=36$; mean age $=34.78$; $\mathrm{SD}=11.36$, age range $=19-74$ ).

Materials All conditions used variations of the same vignette. The scenario describes a week in a startup company, from Monday to Monday, with office days from Monday to Thursday. The company printer breaks as soon as it receives two printing orders at the same time [conjunctive causal structure with two actions], while individual printing does not make a causal difference. The participant acts as CEO with two co-workers, Agent 1 and Agent 2, whose names and gender are counterbalanced across scenarios.

[^368]Agent 1 either uses the printer frequently during the week ("frequent behaviour"), or never prints during the week ("infrequent behaviour"). Furthermore, there is a company policy that determines that Agent 1 and Agent 2 are either allowed to use the printer anytime they want, or only on selected days.

As the CEO of your startup, you officially rule that both Agent 1 and Agent 2 are allowed to use the printer whenever they want [Agent 1 is only allowed to print on Tuesdays and Thursdays, and Agent 2 only on Mondays and Wednesdays.]

By printing on e.g. a Monday, Agent 1 conforms to the company policy in case of the liberal printer policy ("moral norm conformation") or violates the company policy in case of the selective policy ("moral norm violation"). Combined with the frequency manipulation, Agent 1 therefore conforms to or violates the policy, and does so frequently or infrequently. In contrast, Agent 2 prints frequently during the week, and always only on the days on which she is allowed to print. Each day of the week is successively presented.

On Monday, $X$ uses the printer for printing the new flyer design.
On Tuesday, $X$ uses the printer for printing the program outline.
On Wednesday, $X$ uses the printer for the printing of the new logo.
On Thursday, $X$ uses the printer for printing out the schedule.
After the presentation of one week, i.e. the Monday after the start of the scenario, both agents simultaneously send printing orders to the printer and it crashes.

On the following Monday, you come into the office and the printer is broken. Both Agent 1 and Agent 2 have sent printing orders to the office printer today. This is bad because your start up currently does not have the budget to afford a new high quality colour printer.

The company policy is either introduced at the beginning of the scenario, as shown in the previous examples, or on Thursday, i.e. just before the final outcome ("time point of moral norm")

Design and Procedure The experiment was designed as 2 (Frequency) $\times 2$ (Moral Norm) x 2 (Time point of moral norm) within subject paradigm. Participants saw all eight variations of the scenario in a randomized order, and each office day was presented successively on a single slide. The participants then had to judge the causal contribution of both agents to the outcome on a 10-point causal rating scale ("To what extent did Agent X cause the outcome?": 0"- 'None at all'; "10"- 'Fully). Recent studies have highlighted that the term 'cause' in the test questions is ambiguous and can refer to both the causal mechanism and the agent's accountability (Samland et al., 2015). Therefore, we added a 7-point counterfactual relevance agreement scale ("If Agent X had not printed, the problem would not have occurred."; "1" 'Strongly disagree'; " 7 " - 'Strongly agree’) to directly test counterfactual reasoning. In addition, four manipulations check questions about the moral norm ("In the scenario you have just read, was Agent X allowed to print on Monday?") and the frequency of the agents' actions ("In the scenario you
have just read, did Agent X typically use the printer?') with the answer options ('yes', 'no') were given.


Figure 1. Experimental design with the norm introduced on beginning of the scenario (above) and on Thursday, i.e. at the end of the scenario (below).

## Results

Causal Rating | Agent 1 Causal ratings for Agent 1 were higher when they violated the moral norm ( $\mathrm{M}=6.35$; $\mathrm{SD}=2.43$ ) than when they conformed to it ( $\mathrm{M}=4.53$; $\mathrm{SD}=1.60), F(1,80)=67.71, p<.001, \eta_{p}^{2}=.46$. In contrast, when Agent 1 acted only once, they were judged as less causal ( $\mathrm{M}=4.80, \mathrm{SD}=2.46$ ) than when they performed a frequent behaviour $(\mathrm{M}=6.06, \mathrm{SD}=1.71), F(1,80)=25.08, p$ $<.001, \eta_{p}^{2}=.24$. The time point of the norm did not reveal any significant effects ( $\mathrm{p}=.83$ ).

Causal Rating | Agent 2 Agent 2 was seen as less causal when Agent 1 violated a moral norm ( $\mathrm{M}=3.61, \mathrm{SD}=2.33$ ) compared to when Agent 1 conformed to the norm ( $\mathrm{M}=5.40$, $\mathrm{SD}=1.78), F(1,80)=53.33, p<.001, \eta_{p}^{2}=.40$. Likewise, Agent 2 was seen as more causal when Agent 1 printed oneoff ( $\mathrm{M}=4.89, \mathrm{SD}=2.32$ ), compared to when Agent 1 printed frequently ( $\mathrm{M}=4.13, \mathrm{SD}=1.60$ ) , $F(1,80)=13.46, p<.001$, $\eta_{p}^{2}=.14$. The time point of the norm did not reveal any significant effects ( $\mathrm{p}=.16$ ).


Figure 2. Causal Rating. "M" indicates the moral norm and "F" the frequency of behaviour, with ' $\checkmark$ ' for 'conforming/frequent' and ' $\boldsymbol{X}$ ' for 'violating/one-off'. The error bars represent $\pm 1$ SE of the mean.

Counterfactual Relevance |Agent 1 The agreement ratings for the counterfactual relevance of Agent 1 were affected by both Agent 1's moral behaviour, $F(1,80)=21.60, p<.001$, $\boldsymbol{\eta}_{p}^{2}=.21$, and frequency of action, $F(1,80)=17.86, p<.001$, $\boldsymbol{\eta}_{\boldsymbol{p}}^{2}=.18$. The agreement for Agent 1 as relevant counterfactual was higher when they violated a moral norm ( $\mathrm{M}=5.20, \mathrm{SD}=1.47$ ) than when they did not $(\mathrm{M}=4.75$, $\mathrm{SD}=1.44$ ) and lower when they acted one-off ( $\mathrm{M}=4.69$, $\mathrm{SE}=1.63$ ) compared to when they acted frequently $(\mathrm{M}=5.26$, $\mathrm{SD}=1.32$ ).

Counterfactual Relevance | Agent 2 The violation of morality by Agent 1 decreased agreement with Agent 2 as counterfactually relevant (MD=-.45; SD=0.13), $F(1,80)=$ 12.67, $p<.001, \eta_{p}^{2}=.14$. Agent 2 was also seen less counterfactually relevant when Agent 1 acted frequently $(\mathrm{MD}=-.32 ; \mathrm{SD}=0.13), F(1,80)=12.67, p=.006, \boldsymbol{\eta}_{\boldsymbol{p}}^{2}=.14$

Moral Norm Violation A 2 (Frequency) $\times 2$ (Time point of moral norm) ANOVA for norm violating actions revealed a significant difference for the frequent norm violation ( $\mathrm{M}=7.09, \mathrm{SD}=1.24$ ), vs. infrequent norm violation ( $\mathrm{M}=5.62$, $\mathrm{SD}=1.36$ ) of Agent 1 on causal judgements, $F(1,80)=24.60$, $p<.001, \eta_{p}^{2}=.24$. A corresponding reverse rating was found for Agent 2 when Agent 1 violated the moral norm frequently $(\mathrm{M}=3.24, \mathrm{SD}=1.25)$ vs. once $(\mathrm{M}=3.99, \mathrm{SD}=1.33), F(1,80)=$ $8.43, p=.005, \eta_{p}^{2}=.09$. Counterfactual relevance ratings for Agent 1 were higher when Agent 1 frequently violated the norm ( $\mathrm{M}=5.49, \mathrm{SD}=.15$ ) vs. once once ( $\mathrm{M}=4.92, \mathrm{SD}=.20$ ), $F(1,80)=16.17, p<.001, \eta_{p}^{2}=.17$. Agent 2 was seen less counterfactually relevant when Agent 1 frequently violates the norm, compared to violating the norm once ( $\mathrm{MD}=-.47$, $\mathrm{SD}=.18), F(1,80)=7.23, p=.009, \eta_{p}^{2}=.08$.


Figure 3. Counterfactual relevance rating. " M " indicates the moral norm and " $F$ " the frequency of behaviour, with ' $\checkmark$ ' for 'conforming/frequent' and ' $X$ ' for 'violating/one-off'. The error bars represent $\pm 1 \mathrm{SE}$ of the mean.

## Discussion

Experiment 1 confirmed the effect of moral norms on causal attribution (Knobe \& Fraser, 2008), showing that an agent who violates a moral norm is judged as more causal and counterfactually relevant than a norm-conforming agent. We also found a reversed norm influence for agent-level statistical norms, with frequent behaviour being judged more instead of less causal than one-off actions. We found the same effects for counterfactual relevance ratings, showing that moral norm violations and frequent behaviour increase the extent to which an action is seen as counterfactually relevant to the outcome. An increase in causal attribution to Agent 1 due to norm violating and/or frequent behaviour was accompanied by a decrease in causal attribution to Agent 2, known as "causal superseding" (Kominsky et al., 2015). We did not find an interaction between the immoral norm behaviour and the time point at which the norm was introduced in the scenario. Hence, an action that has been violating the norm from the beginning is judged as causal as an action that has been frequently performed before, but only violated a norm after the new introduction of a moral norm.

Our finding that frequent behaviour increased causal attribution is consistent with Sytsma et al. (2012), but was surprising insofar as the causal structure of our scenario was designed to be causally insensitive to an individual action that occurs frequently, i.e. using the printer often. However, it might be argued that the influence of frequent vs. one-off actions in a conjunctive causal structure depends on the knowledge that both agents have about each other's behaviour. For example, when the agent who usually never uses the printer suddenly prints in a conjunctive causal structure, she might be seen as more causally responsible for the outcome if she is aware that the other agent frequently
prints. In addition, despite our attempt to implement a conjunctive causal structure, participants might have interpreted frequent behaviour in our scenario as gradually damaging the printer. In order to further investigate the role of knowledge, and accentuate the difference between a cumulative and conjunctive causal structure, we conducted a second experiment.

## Experiment 2

Experiment 2 examined the influence of frequent behavior in dependence of the underlying causal structure and state of knowledge of the agents.

## Methods

Participants 102 participants were recruited via Amazon Mechanical Turk ( $\$ 0.80$ for ca. 15 min ), and after removing participants who had more than four of the 32 check questions wrong, the data of 64 participants were analyzed ( male $=34$, female $=30$; mean age $=33.56 ; \mathrm{SD}=11.92$, age range $=21-74$ ).

Materials The same design as in Experiment 1 was used, except that the moral norm manipulation was removed by setting the moral status of all actions to "moral norm conforming" by a company policy that allowed the two agents to use the printer anytime. The frequency of the actions of Agent 1 was varied while holding the frequency of Agent 2 constant. Hence, either both of them frequently used the printer on different days during the week, or only Agent 2 used the printer. The underlying causal structure of the scenario was also varied by either containing a conjunctive causal structure (Experiment 1), or a cumulative causal structure, i.e. every action gradually increases the likelihood for the outcome to happen

The printer is quite worn out, and every printing order sent strains the printer system a bit more.

Finally, we varied whether the agents know about each other's behavior.

The printer management system does [does not] display the current user balance so that Agent 1 and Agent 2 know [do not know] who prints how much each day.

Design and Procedure The experiment was designed as a 2 (Frequency) $\times 2$ (Causal Structure) $\times 2$ (Knowledge) within Subject Design. The participants saw each of the eight scenarios in a randomized order, and answered two causal strength and counterfactual relevance questions, using the same scales as in Experiment 1. Following this, they were asked two manipulation check questions about the typical behavior ("In the scenario you have just read, did Agent X typically use the printer?") a knowledge manipulation question ("In the scenario you have just read, does the printer management system display the user balance?"), and a causal structure check question ("In the scenario you have just read, under which conditions does the printer crash?"), with the
options "Two or more printing orders at the same time" or "Overuse".

## Results

Causal Rating | Agent 1 Causal ratings for Agent 1 were generally higher when they acted frequently ( $\mathrm{M}=5.31$; $\mathrm{SD}=1.30$ than when they acted one-off ( $\mathrm{M}=2.90 ; \mathrm{SD}=1.89$ ), $F(1,63)=103.78, p<.001, \eta_{p}^{2}=.62$. However, the increase in causal contribution when Agent 1 used the printer frequently was greater in the cumulative causal structure than in the conjunctive causal structure, $F(1,63)=103.78, p=.001$, $\eta_{p}^{2}=.16$. No significant effect for knowledge was found ( $\mathrm{p}=.48$ ).

Causal Rating | Agent 2 Agent 2 was seen as less causal when Agent 1 acted frequently, ( $\mathrm{M}=5.32, \mathrm{SD}=1.30$ ), but more causal when Agent 1 performed a one-off action $(\mathrm{M}=6.93, \mathrm{SD}=1.90), F(1,63)=69.15, p<.001, \eta_{p}^{2}=.52$. The increase in causal attribution to Agent 2 when Agent 1 did not act frequently was greater in the cumulative than in the conjunctive causal structure, $F(1,63)=39.64, p<.001, \eta_{p}^{2}=$ .37. There was no significant effect for knowledge ( $p=.90$ ).


Figure 4. Causal Rating. " $F$ " indicates the frequency of behaviour, with ' 1 ' for 'frequent' and ' 0 ' for 'one-off'. The error bars represent $\pm 1 \mathrm{SE}$ of the mean.

Counterfactual Relevance Agent 1 was seen as more counterfactually relevant when they acted frequently ( $\mathrm{M}=4.86 ; \mathrm{SD}=.1 .33$ ) versus one-off $(\mathrm{M}=3.93 ; \mathrm{SE}=1.56), F(1$, 63) $=29.87, p<.001, \eta_{p}^{2}=.32$, and the increase in counterfactual relevance when acting frequently was greater in the cumulative structure, $\mathrm{F}(1,63)=6.90, p=.011, \eta_{p}^{2}=$ .10. Agent 2 is assigned less counterfactual relevance when Agent 1 acts frequently ( $\mathrm{MD}=-.56 ; \mathrm{SD}=.197, F(1,63)=$ $12.19, p=.02, \eta_{p}^{2}=.16$. There was no significant effect for the knowledge factor $(\mathrm{p}=.50)$.


Figure 5. Counterfactual Relevance Rating. "F" indicates the frequency of behaviour, with ' 1 ' for 'frequent' and ' 0 ' for 'one-off'. The error bars represent $\pm 1 \mathrm{SE}$ of the mean.

## Discussion

Experiment 2 confirmed the effect of frequent behavior on causal judgments. Frequent behavior was seen as more causal than one-off behavior, independent of the underlying causal structure. However, participants were aware that the change from frequent behaviour to a one-off action led to a greater reduction in causal contribution in the cumulative causal structure than in a conjunctive structure. The effect of frequent behaviour was independent of whether or not the agents knew about each other's behavior.

## General Discussion

Our two experiments found evidence for an increase in the causal attribution to an agent who acts in line with their frequent behaviour, even when the frequency of this behavior does not make a causal difference to the outcome. Our results present a challenge to the norm account of causal judgment, showing that a one-off action that deviates from frequent behavior is judged less instead of more causal for a negative outcome. Since the knowledge about the other agent's behavior did not make a difference to causal judgements, despite being crucial for estimating the outcome in a conjunctive causal structure, we rule out that people might have inferred bad intentions or foreseeability of the outcome from typical behavior (Lagnado \& Channon, 2008). Hence, we believe that the fact that people assign increased causality to a frequent, typical action goes against predictions of norm theories and shows the need for a new account of frequency of actions in causal attribution. Current counterfactual and structural equation models (Halpern \& Pearl, 2005; Chockler \& Halpern, 2014) fail to account for the asymmetry of causal attribution between a frequent vs. one-off actor in a conjunctive causal structure, given that previous behavior does not change the counterfactual dependency of the two actions in the actual situation.

We draw two conclusions from this. First, we argue for the need to include the previous history of actions of an agent
into counterfactual accounts of causal judgments. Second, we argue that an extension of current counterfactual theories is needed in order to capture the influence of frequency. One such extension could be probabilistic (Fenton-Glynn, 2016). While frequent behavior does not differ from one-off behavior in terms of single counterfactual dependencies, it does so probabilistically. If, in a counterfactual world, the other agent acts at a random time point in a conjunctive structure, the outcome is more likely to occur if I frequently perform the other action needed for the outcome. In contrast, it is less likely to occur if I act infrequently. The raised probability of the outcome due to my frequent behavior can even increase when we also vary whether the other agent acts frequently vs. one-off. As a result, we argue that in addition to counterfactually testing whether the undoing of an action makes a difference, we also need to test whether the variation of the frequency of the action would make difference to the likelihood of the outcome to occur.

Our results show that a frequent norm violation is judged more causal than a one-off moral norm violation. This is despite the fact that in the actual situation of the outcome, they are both equally counterfactually relevant, as well as equally. Hence, we think that current norm theories need to include the frequency of previous norm violations in order to fully capture the influence of moral norms on causal attributions. However, the time point at which the moral norm is introduced into the scenario does not change the way we assign causality to actions that violate this norm. To conclude, the influence of the frequency of an action proves to be a crucial factor in the attribution of causality, and calls for new theoretical frameworks of causation. We argue that this framework has to be probabilistic.

## References

Alicke, M. D., Rose, D., \& Bloom, D. (2011). Causation, Norm Violation, and Culpable Control. Journal of Philosophy, 108(12), 670-696.
Fenton-Glynn, L. (2016). A Proposed Probabilistic Extension of the Halpern and Pearl Definition of "Actual Cause." The British Journal for the Philosophy of Science, 78, 85-109.
Gershman, S. J., Gerstenberg, T., Baker, C. L., \& Cushman, F. A. (2016). Plans, Habits, and Theory of Mind. PloS One, 11(9), e0162246.
Gerstenberg, T., \& Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. Cognition, 115(1), 166-71.
Hart, H. L. A., \& Honoré, T. (1985), Causation in the Law, 2nd ed., Oxford: Clarendon.
Halpern, J. Y., \& Hitchcock, C. (2015). Graded Causation and Defaults. The British Journal for the Philosophy of Science, 66(2), 413-457.
Halpern, J. Y., \& Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. The British Journal for the Philosophy of Science, 56(4), 843-887.
Hitchcock, C., \& Knobe, J. (2009). Cause and Norm. Journal of Philosophy 106 (11): 587612.

Icard, T., Kominsky, J., and Knobe, J. (2016). Causality, Normality, and Sampling Propensity. In : Proceedings of the 38th Annual Meeting of the Cognitive Science Society.
Kahneman, D., \& Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. Psychological Review, 93(2), 136-153.
Knobe, J., \& Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. Moral Psychology. In: Walter Sinnott-Armstrong (ed.), Moral Psychology. MIT Press
Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., \& Knobe, J. (2015). Causal superseding. Cognition, 137, 196-209.
Lagnado, D. A., \& Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. Cognition, 108(3), 754-770.
Roxborough, C., \& Cumby, J. (2009). Folk psychological concepts: Causation1. Philosophical Psychology, 22(2), 205-213.
Samland, J., \& Waldmann, M. R. (2015). Highlighting the causal meaning of causal test questions in contexts of norm violations. In D. C. Noelle et al. (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Sytsma, J., Livengood, J., \& Rose, D. (2012). Two types of typicality: rethinking the role of statistical typicality in ordinary causal attributions. Studies in History and Philosophy of Biological and Biomedical Sciences, 43(4), 814-20.

# Variability in advice taking in decision making 

Miho Kitamura (kitamura.miho@aoni.waseda.jp)<br>Waseda University,<br>3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

Katsumi Watanabe (katz@waseda.jp)<br>Waseda University,<br>3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan


#### Abstract

We investigated how people would change and vary in accepting advice when the effectiveness of advice was unclear. In each trial, participants estimated a monthly rent of an apartment room based on the attribute list. Then, another estimate by a real-estate agent was given as advice. Participants made a final estimation, either by taking the advice fully, partially, or rejecting it totally. They repeated 48 estimations without feedback. The weight of advice index, representing how much each participant weighed a given advice, gradually decreased as the number of trials increased. Interestingly, the gradual reduction of acceptance was not observed in participants with high empathy and low depressive scores; they kept accepting advice even when the effectiveness of advice was unclear. These results suggest that the willingness of accepting and using advice depends on history of advice taking, the individual traits, and mood.


Keywords: decision making; advice taking; individual difference

## Introduction

When we cannot make a decision by ourselves, we frequently ask for advice. Assume you are about to purchase a house and have two candidate options. Both houses are similarly attractive but different in various aspects. You cannot decide which to buy. Here comes a friend and starts giving you a series of advice. Will you accept advice and, if so, how much will you use the advice for your final decision? Obviously, it depends. Generally, we do not know if advice is useful or not until the outcome of a decision comes out. However, under ambiguous situations with time constraints, we often need to decide whether we should accept advice. In the present study, we examined how people would change and vary in accepting advice when the effectiveness of advice is unclear. In particular, we were interested in whether advice taking would depend on history of experience of advice taking and individual traits and mood.

In general, people heed advice and adjust their estimate and/or judgment (Bonaccio \& Dalal, 2006; Yaniv, 2004b). However, this advice taking process is prone to various cognitive bias. For example, people tend to weight own judgment more even when they should take advice to improve their judgments (Yaniv, 2004a; Yaniv \& Kleinberger, 2000). One of the biasing factors is a
characteristic of advisor. For example, people follow advice more when their advisor is labeled as an expert by the experimenter (e.g, Meshi et al., 2001; Harvey \& Fischer, 1997), and when an advisor expresses great confidence in his/her advice (e.g., Sniezek \& van Swol, 2001).

Although the previous studies mainly focused on the characteristics of advisor, several studies showed that, the internal states of decision makers (e.g., emotion and confidence) also modulate advice taking behaviors (Cooper, 1991; Gino \& Schweitzer, 2008; Gino, Brooks, \& Schweitzer, 2012). For example, when anxiety is experimentally induced in participants, they show lower confidence and consequently a higher level of acceptance of advice (Gino et al., 2012). These previous findings suggest that there may be individual differences regarding advice taking. However, it is not clear how individual characteristics (i.e., emotion, personality, and interpersonal reactivity) would influence advice taking. Using three questionnaires of personality, emotion, and interpersonal reactivity, we examined individual variability of advice taking behaviors.

One of major motivation for acceptance of advice is to adjust and improve own judgments. Therefore, if the effectiveness of advice is unclear, the acceptance of advice will decrease over time. In the present study, we confirmed this tendency and then examined how the reduction of advice taking would vary across individuals.

## Method

## Participants

Fifty-four university students participated in the present study ( 18 females and 36 males). They gave written informed consent and were blind to the purpose of the study. Most of them lived in Tokyo and its vicinity and had some knowledge of rent in the Tokyo area.

## Price Estimation Task

Participants performed the price estimation task (Meshi et al., 2012), where they estimated rents of apartment rooms. Five attributes of each apartment room (room layout, area in square meters, distance form the nearest station, floor level, and building age) were presented on a computer monitor (Figure 1). Based on those set of information, participants estimated the monthly rent of each room (in Japanese yen).

After the initial estimation, another opinion, i.e., advice, on the rent of the room was given in one half of trials (advice condition). We instructed them, "Advice was derived from a real-estate agency who visited our laboratory to participate this experiment another day and judged the rental price for them on the same computer," but the price was an actual rent of the room around the Takadanobaba station, Sinjyuku, Tokyo. Then, participants were asked to decide whether they would consider changing their first estimation, if so how much. They reported the estimated price by adjusting the slide bar on the monitor with a computer mouse, and decided by clicking the mouse button. In the other half of trials, they did not receive any advice but still were asked whether or not they would like to change their opinions (no-advice condition). No feedback was given about their decision. Advice and no-advice conditions were randomly intermixed and presented 24 times, resulting in 48 trials for the whole session. The trials were divided into four sessions. The next session started after participants pressed the space bar; they were allowed to rest between trials.


Figure 1: Example of the trial flow of the price estimation task of the present study.

## Analysis

In the advice condition of the present task, we obtained two estimates for the rents: the fist estimate without advice and the second estimate after receiving advice. The difference between advice and the first estimate represents how much opinions are different between the participant and the advisor. The difference between the second rating and the first rating represents how much the participant changes his or her opinion. In order to examine how much each participant took into consideration or weighed the given advice, we calculated "Weight of Advice" index (WOA; Bonaccio \& Dalal, 2006).

$$
W O A=\frac{\text { Second estimate }- \text { First estimate }}{\text { Advice }- \text { First estimate }}
$$

For example, a participant's initial estimate was 1000 yen, and the advice was 2000 yen. Assume, after receiving the advice, the participant adjusts her estimate slightly higher, say, 1100 yen. Then, the weight of advice index would be calculated as 0.1 , indicating a small influence of advice. On the other hand, the participant may fully use the advice and adjust her estimate to 2000 yen. The calculated weight of advice index would then be 1 . It is theoretically possible that WOA is larger than 1 or takes a negative value. However, as such cases would represent behaviors different from normal advice taking or simple response errors, we excluded those data from the analyses (42 trials). Additionally, we also excluded the data from 89 trials whose first estimations were identical to the advice prices, or whose response time is 4 SD above the mean response time of all advice-taking trials. The resulted number of trials was 1165 .

## Questionnaires

In addition to WOA from each participant, to examine possible influences of participants' personality, mood, and interpersonal reactivity traits on the advice taking in the price estimation task, we used three questioners; Ten Item Personality Inventory (TIPI-J; Oshio, et al., 2012; Gosling, Rentfrow, \& Swann, 2003); Beck Depression Inventory (BDI; Beck et al., 1961), and Interpersonal reactivity index (IRI; Davis, 1980; 1983; 1996). These questionnaires are widely used to measure participants' personality, mood, and interpersonal reactivity traits in previous studies (e.g, Oshio et al., 2014). All participants were asked to answer those questionnaires.

## Results

In Figure 2, we plotted the difference between the first rating and advice, and the difference between the first and second ratings. The horizontal axis is the difference between the advice and the first estimate, representing how much opinions were different. The vertical axis is the difference between the second estimate and the first estimate, representing how much participants changed their estimates. Each dot represents a single trial. In this plot, two lines were apparent. One was a horizontal line at the 0 point of the vertical axis. These data points were from the trials where the participants did not change their estimates at all. Irrespective of the differences in opinions, they simply ignored the advice and did not change their estimates. The other was a diagonal line. The data points on the diagonal line were from the trials where the difference between the advice and the first estimate and the difference between the second estimate and the first rating estimate identical. That is, the participants fully accepted the advice. Then the data between these two lines represented the trials where the participants took the advice but adjusted their estimates partially.

To capture various advice taking behaviors, we calculated the Weight of Advice index (WOA). Figure 3 shows the frequency distribution of the WOA. The data at the WOA of zero correspond to the data on horizontal line in Figure 2, i.e., the participants did not take the advice at all. Out of the total trials of about 1165, 535 trials were this type. By contrast, the trials where the participants fully accepted the advice were at the WOA of 1 . These correspond to the diagonal line in Figure 2. 96 trials were this type. WOA values between 0 and 1 represent the trials where the participants partially accepted the advice. In total, advice was used $54 \%$ of trials. This acceptance ratio was similar to that observed in the previous study (Soll \& Larrick, 2009).


Figure 2: The difference between the second estimate and the first estimate as a function of the differences between advice and the first estimate.


Figure 3: Frequency distribution of the weight of advice (WOA) index.

## Variability of advice taking in time

Since the experiment contained 48 trials, we were able to examine how WOA would change as the experiment progressed. Figure 4 shows the averaged WOA scores from all the participants as a function of the trial number. We found that WOA, namely, the tendency to accept advice, gradually decreased over the course of the experiment $(r(48)$
$=-.524, p<.001)$. Then, we divided the trials into the first and second halves and found the significant difference of WOA scores between them $(t(1163)=2.70, p<.01)$, indicating that the participants followed the advice less in the second half than in the first half.
One may argue that the decrease in the acceptance of advice over time would be due to that the participants become better in the estimation task and could perform the task without advice in the second half. We did not provide feedback for their estimation. However, they repeatedly saw the rent estimation presented as advice, which were actual rents. In addition, the participants had some knowledge of rent in the Tokyo area. Thus, there was the possibility that participant's estimate would approach to the estimate presented as advice implicitly. To test such a possibility, we also calculated the difference between the advice and the first estimate as an index of the task performance (Figure 5). We did not find any improvement of the task performance as a function of trial number $(r(48)=.041, p=.786)$. Therefore, the decrease in WOA was not due to the better performance in the second half.


Figure 4: Averaged WOA scores as a function of the trial number. WOA decreased gradually as the number of trials increased. Error bars show the standard errors of the mean scores.


Figure 5: Averaged task performance scores (correct rent price, i.e., advice, minus the first estimate) as a function of the trial number. Error bars show the standard errors of the mean scores.

(mean WOA scores in Second half - mean WOA scores in First half)

Figure 6: Distribution of WOA changes of each participant. Positive values indicate that the participants accept the advice more in the latter half of the experiment (positive group), while negative values indicate that the participants accepted the advice less in the latter half of the experiment (non-positive group).

## Variability among individuals

The overall results suggested that, on average, the participants followed advice less in the second half than the first half of the task. However, this was not the case for all the participants; some participants did show larger WOA in the latter half. We calculated the change of WOA for each participant by subtracting WOA in the first half from WOA in the second half. Thus, a positive score would indicate that the participants tended to accept the advices in the second half. Figure 6 shows the distribution of WOA changes.

We divided the participants into the "positive-group" (WOA changes values were positive) and the "non-positive group" (WOA changes values were negative or equal to zero), and examined whether there would be any individual differences between the two groups of the participants. The questionnaires measured depressive state, big-five personality trait (Extraversion, Conscientiousness, Agreeableness, Openness to Experience, and Neuroticism), empathy for others (cognitive abilities on a Fantasy Scale (FS) and Perspective Taking (PT), and affective components through an Empathic Concern (EC) and Personal Distress (PD). We compared the scores of the questionnaires between the positive group and the non-positive group. Two participants were excluded from the analyses because significant parts of the questionnaires were incomplete. In total, the data from fifty-two participants were analyzed.

We found significant differences in depressive state $(t(50)$ $=2.30, p<.03$ ) and in perspective taking (PT) scores $(t(50)=-2.48, p<.02)$. The participants who showed the positive WOA changes had lower depression scores and higher perspective taking scores (Figure 7). These results suggest that individuals with less depression and higher perspective taking tend to keep taking the advice. There was


Figure 7: Averaged depression and perspective-taking scores for different WOA changes groups (non-positive and positive groups). Error bars show the standard errors of the mean scores.


Figure 8: The relationship between the depression and perspective-taking scores.
a significant negative correlation between the depression and the perspective taking scores $(r(52)=-.293, p<.05$; Figure 8). All the other scores did not show any significant difference.

## Discussion

In the present study, we examined how people would change and vary in accepting advice when the effectiveness of advice was not clear. We found that: (1) Participants took the advice less in the latter half of the experiment. (2) The decrease in WOA was not due to the change in task performance. (3) The decrease of WOA was pronounced in the participants with high depressive state and low perspective taking tendency of empathy. In other words, the participants with low depression and high empathy kept taking advice even when the effectiveness of advice was unclear. Our results suggest that the willingness of accepting and using advice changes with prior experience of advice taking and the individual traits and mood interact with the change in advice taking.

In about $54 \%$ of trials, the participants used advice either fully or partially, even though the usefulness of advice was not clear. This result suggests that it is difficult to ignore advice completely under uncertainty of usefulness of advice. Previous studies suggest that the major motivation of seeking advice is not only for accuracy in decision making but also for social reasons, for example sharing responsibility or to justify our decisions (Kennedy, Kleinmuntz, \& Peecher, 1997; Harvey \& Fischer, 1997). Our results might reflect this kind of social motivations in advice taking.

Since the participants were not sure about whether advice was useful or not in the present study, it was reasonable to assume that the acceptance of advice would reduce over time. However, this was not the case for all the participants; low depressive and high perspective taking participants kept taking advice. Also, there was the negative correlation between depressive scores and perspective taking scores. It has been reported that depression is associated with focused attention on the self rather than others (Ingram, 1990) and focusing increases the accruing periods and severity of depression (e.g., Just \& Alloy, 1997; Kuehner \& Weber, 1999; Nolen-Hoeksema, 2000; Nolen-Hoeksema, Parker, \& Larson, 1994). However, it still remains to be investigated how depressive mood and the ability of perspective taking influence and/or interact with advice taking behavior.

In the present study, we set the situation that advice was derived from a real-estate agency. There is the possibility that this information about advisor would help to increase the acceptance of advice even when the effectiveness of advice was not clear. A previous research showed that people valued expert advice than novice one when making decisions (e.g., Meshi et al., 2001). Therefore, the characteristics of an advisor have influences in advice taking, which include not only expertise but also how closely the advising person is related. In our daily-life, we often ask friends or family for advice, even though we can seek expert advice. Recent research suggests that advice would be accompanied by social and emotional support (i.e, regulating emotional distress), and decision makers prefer such emotional and social support when they make decisions (Horowitz et al., 2001; Dalal \& Bonaccio, 2010).

In the future research, it would also be interesting to examine the role of seeking or accepting advice derived from others that have close associations with them, such as friends or family.

In the present study, no feedback was provided, It is possible that providing feedback may affect the acceptance of advice over time. We found the decrease of WOA over time when the effectiveness of advice was uncertain. Providing feedback would make participants to explicitly adjust their estimates. In future research, it would be important to examine the effects of feedback on temporal variations for the acceptance of advice.

## Acknowledgments

This study was partly supported by Japan Science and Technology Agency and Genesis Research Institute Inc.

## References

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., \& Erbaugh, J. (1961). An inventory for measuring depression. Archives of general psychiatry, 4, 561-571..
Bonaccio, S., \& Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational Behavior and Human Decision Processes, 101, 127-151.
Cooper, R.S. (1991). Information processing in the judgeadvisor system of group decision-making. Unpublished master's thesis, University of Illinois, Urbana-Champaign.
Dalal, R. S., \& Bonaccio, S. (2010). What types of advice do decision-makers prefer?. Organizational Behavior and Human Decision Processes, 112, 11-23.
Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology, 10, 85.
Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. Journal of Personality and Social Psychology, 44, 113126.

Davis, M. H. (1996). Empathy: A social-psychological approach. Boulder, CO: Westview.
Gino, F., \& Moore, D. A. (2007). Effects of task difficulty on use of advice. Journal of Behavioral Decision Making, 20, 21-35.
Gino, F., \& Schweitzer, M. (2008). Blinded by anger or feeling the love: How emotions influence advice taking. Journal of Applied Psychology, 93, 1165-1173.
Gino, F., Brooks, A. W., \& Schweitzer, M. E. (2012). Anxiety, advice, and the ability to discern: feeling anxious motivates individuals to seek and use advice. Journal of personality and social psychology, 102, 497.
Gosling, S. D., Rentfrow, P. J., \& Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in personality, 37, 504-528.
Harvey, N., \& Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility.

Organizational, Behavior and Human Decision Processes, 70, 117-133.
Horowitz, L. M., Krasnoperova, E. N., Tatar, D. G., Hansen, M. B., Person, E. A., Galvin, K. L., \& Nelson, K. L. (2001). The way to console may depend on the goal: Experimental studies of social support. Journal of Experimental Social Psychology, 37, 49-61.
Ingram, R.E. (1990). Self-focused attention in clinical disorders: re- view and a conceptual model. Psychological Bulletin, 109, 156-176.
Just, N., \& Alloy, L. B. (1997). The response styles theory of depression: tests and an extension of the theory. Journal of abnormal psychology, 106, 221.
Kennedy, J., Kleinmuntz, D. N., \& Peecher, M. E. (1997). Determinants of the justifiability of performance in illstructured audit tasks. Journal of Accounting Research, 35, 105-123.
Kuehner, C., \& Weber, I. (1999). Responses to depression in unipolar depressed patients: An investigation of NolenHoeksema's response styles theory. Psychological medicine, 29, 1323-1333.
Meshi, D., Biele, G., Korn, C. W., \& Heekeren, H. R. (2012). How expert advice influences decision making. PLoS One, 7, e 49748.
Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. Journal of abnormal psychology, 109, 504.
Nolen-Hoeksema, S., Parker, L. E., \& Larson, J. (1994). Ruminative coping with depressed mood following loss. Journal of personality and social psychology, 67, 92.
Oshio, A., Shingo, A. B. E., \& Cutrone, P. (2012). Development, Reliability, and Validity of the Japanese Version of Ten Item Personality Inventory (TIPI-J). Japanese Journal of Personality, 21, 40-52.
Oshio, A., Abe, S., Cutrone, P., \& Gosling, S. D. (2014). Further validity of the Japanese version of the Ten Item Personality Inventory (TIPI-J). Journal of Individual Differences, 35, 236-244.
Sniezek, J. A., \& Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. Organizational behavior and human decision processes, 84, 288-307.
Soll, J. B., \& Larrick, R. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 780-805.
Yaniv, I. (2004a). The benefit of additional opinions. Current Directions of Psychological Science, 13, 75-77.
Yaniv, I. (2004b). Receiving other people's advice: Influence and benefit. Organizational Behavior and Human Decision Processes, 13, 1-93.
Yaniv, I., \& Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. Organizational Behavior and Human Decision Processes, 83, 260-281.

# Inner speech in post-stroke motor aphasia 

Vanja Kljajevic (vanja.kljajevic@gmail.com)<br>University of the Basque Country (UPV/EHU), Vitoria \&<br>IKERBASQUE, Basque Foundation for Science, Bilbao, Spain<br>Estibaliz Ugarte Gómez (estibaliz.ugarte@atecearaba.org)<br>Asociación de traumatismo cráneo-encefálico y daño cerebral adquirido de Álava (ATeCe), Vitoria, Spain

Cristina López (cristina.lopez@atecearaba.org)<br>ATeCe, Vitoria, Spain<br>Yolanda Balboa Bandeira (yolibalboa_b@yahoo.es)<br>ATeCe, Vitoria, Spain

Agustin Vicente (agustin.vicente@ehu.eus)
University of the Basque Country (UPV/EHU), Vitoria \&
IKERBASQUE, Basque Foundation for Science, Bilbao, Spain


#### Abstract

The goal of the present study was to determine whether chronic post-stroke patients with motor aphasia have impaired inner speech abilities and whether they use inner speech in everyday life. To answer these questions, we recruited eight chronic post-stroke aphasic patients and 13 cognitively healthy adults, who underwent testing on a range of evaluative tests and four experiments specifically designed for the purposes of this study. The experimental results suggest that post-stroke patients with motor aphasia have impaired inner speech. However, patients' subjective reports indicate that they use various types of inner speech, despite the observed deficit. Taken together, our data suggest that impairment of certain aspects of inner speech may still allow a degree of use of other aspects of inner speech, emphasizing a need to extend research on inner speech in aphasia to the variety of its forms.


Keywords: aphasia; inner speech; anomia; working memory.

## Introduction

Inner speech has been traditionally recognized as an important component of human mental life, and in particular its role in the relationship between language and thought has been debated (Kinsbourne, 2000). The interest in inner speech has been renewed recently, partly due to new perspectives on how language contributes to consciousness and whether conscious thought is possible without inner speech (de Guerrero, 2005, Martínez-Manrique \& Vicente, 2010), and partly due to recent developments in speech production theories (e.g., Indefrey \& Levelt, 2004) that propose that inner speech is a stage in speech production that precedes articulation (Levelt, 1995). Inner speech is often defined as silently talking to oneself or speech-foroneself (Vygotsky, 1986), "the little voice in the head" (Perrone-Bertolloti et al., 2014), an internalized verbal thought that can be consciously explored (Marverl \& Desmond, 2012), "the subjective experience of language in the absence of overt and audible articulation" (Alderson-

Day \& Fernyhough, 2015, p. 931) or more generally as a form of mental imagery (Oppenheim \& Dell, 2008).

The unique cognitive status of inner speech is reflected in its pervasive role in a variety of functions. It is involved not only in language, but also in working memory (e.g., in subvocal rehearsal) (Paulesu et al., 1993; Baddeley \& Loggie, 1999), complex reasoning (Baldo et al., 2015), selfregulation (Vygotsky, 1986), meta-cognition (Bermúdez, 2003), and self-awareness (Morin, 2009; Morin \& Michaud, 2007). For example, impairment of inner speech has been associated with the impairment of global self-awareness, self-related memories and emotional awareness, along with impaired "sense of individuality" and corporeal awareness (Morin, 2009). It also contributes to auditory verbal hallucinations in schizophrenia (Frith, 1992). In addition to establishing that inner speech is implicated in a variety of cognitive functions in healthy and impaired brains, research so far has discerned various forms of inner speech, such as condensed, dialogic, self-referent, involving others (Alderson-Day \& Fernyhough, 2015), with the evidence indicating that these different forms of inner speech may require support of different brain areas (Alderson-Day et al., 2015).

However, investigating directly and objectively the highly subjective and elusive psychological processes that support inner speech represents a methodological challenge (Vygotsky, 1986; De Bleser \& Marshall, 2005). The attempts to determine the elusive nature of inner speech fall roughly into two types: the accounts that try to view inner speech in its entirety and emphasize its subjective character, and accounts that focus on a specific aspect of inner speech that can be objectively tested. Thus, the hypotheses generated by the first type of accounts cannot be objectively tested, and the explanations based on the second type of accounts do not exhaust the concept of inner speech.

As an example, Vygotsky's (1986) concept of inner speech is characterized by a highly subjective meaning, predicative nature, and a reduced syntactic form. In
contrast, an often studied proxy of inner speech in contemporary research is phonological representation of a word, which is tested via a person's ability to silently judge whether two words rhyme, whether one word is longer than the other, or whether two words are homophones (Levine et al., 1982; Feinberg et al., 1986). Making silent judgments in such tasks requires the use of inner speech.

Another way of distinguishing between the two types of approaches to studying inner speech is in terms of abstract and concrete inner speech, where the former refers to inner speech as relating language and thought, and the latter considers inner speech a stage in speech production, i.e., phonological and/or phonetic level of inner speech (Levelt, 1995).

One interesting and still not well explored issue pertains to inner speech abilities in persons with impaired overt speech due to brain damage, such as post-stroke patients with motor aphasia.

The available evidence suggests a great variability in inner speech abilities in persons with post-stroke aphasia, with patterns of partial as well as complete deficit in inner speech (Levine et al., 1982; Feinberg et al., 1986; LanglandHassan et al., 2015). Furthermore, the degree of inner speech deficit in these patients typically coincides with a degree of overt speech impairment, although cases of aphasia with discrepant overt-covert speech abilities were also reported (Geva et al., 2011a).

In the present study, we investigated inner speech abilities in chronic post-stroke patients with motor aphasia. The goal was to determine whether these patients had impaired inner speech abilities and how they used inner speech in everyday life. To answer these questions, we tested chronic poststroke aphasic patients on a variety of evaluative tests and four experiments specifically designed to assess their inner speech abilities.

## Methods

## Participants

Eight post-stroke aphasic speakers (two males) were recruited at a local rehabilitation center (Asociación de traumatismo cráneo-encefálico y daño cerebral adquirido de Álava, henceforth ATeCe). All patients were at the chronic stage, with more than one year post-onset time. They all had suffered a single stroke in the left hemisphere, affecting the prefrontal or fronto-temporal areas, except one person, who had a lesion in the left basal ganglia (A07). Seven patients had motor aphasia, while one person's aphasia was characterized as mixed aphasia (A02). This person's aphasia was of a predominantly motor type, with a strong anomic component it was labeled as "mixed" aphasia by the speech pathologist (C.L.) on the basis of multiple extensive tests, subjective observations and neurological reports. Severity of patients' aphasia ranged from moderate to severe. They were all right-handed before the illness, except one patient
who was left-handed (A03). They had no other significant neurological or psychiatric conditions.

Thirteen cognitively healthy adults for the control group (four males) were recruited from the community. The participants in the control group had no history of neurological or psychiatric disease, drug abuse, and at the time of testing they were not using any medications that could affect cognition.

All participants had normal hearing and normal or correct-to-normal vision.

## Evaluative measures

To obtain more general information on patients' cognitive abilities relevant for the present study, we administered the following evaluative measures: Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005) to assess participants' general cognitive status, Boston Diagnostic Aphasia Examination (Goodglass \& Kaplan, 1983) - the Spanish version (García-Albea et al., 1996), Boston naming test (Kaplan et al., 1983), phonological discrimination test (Ardila et al., 1994), the Edinburgh handedness inventory (Oldfield, 1971), the Month ordering test (Almor et al., 2001) to test their verbal working memory, Raven's Progressive Color Matrices (Raven et al., 1990) to test their nonverbal intelligence, Beck depression inventory (Beck et al., 1961) to exclude severe depression, and the Varieties of Inner Speech Questionnaire (McCarthy-Jones \& Fernyhough, 2011) to obtain an insight into each patient's awareness of their use of different forms of inner speech in everyday life.

Unlike other types of inner speech questionnaires, the VISQ addresses important aspects of inner speech, such as condensation and dialogicality. The condensed form, which has by definition reduced syntax, and dialogic inner speech, which is exchange among different internalized perspectives, are of particular importance in studying inner speech of patients with motor speech disorders. Thus, the Questionnaire assesses four types of inner speech: dialogic, condensed, other people in inner speech, and evaluative/ motivational inner speech Examples of stimuli for each form are given in (1-4):
(1) My thinking in words is more like a dialog with myself rather than my own thoughts in a monologue. - Dialogic
(2) I think to myself in words using brief phrases and single words rather than full sentences. Condensed
(3) I hear other people's voice nagging me in my head. - Other people in inner speech
(4) I talk silently to myself telling myself to do things. - Evaluative/motivational inner speech (McCarthy-Jones \& Fernyhough, 2011).

Healthy controls (HCs) underwent MoCA and completed the Varieties of Inner Speech Questionnaire. All tests were administered in Spanish, which was the first language of all participants.

## Experimental measures

All participants were tested in four tasks, which require inner speech for correct completion: silent judgments of rhyming (experiment 1), syllable discrimination (experiment 2), identification of words in compounds (experiment 3) and identification of words in names for numbers (experiment 4).

Experiment 1: The rhyming task in our study required silent rhyming judgments of pairs of words associated with pairs of drawings. The silent rhyming judgments paradigm has been successfully used in previous studies on inner speech in other languages (Langland-Hassan et al., 2015). This paradigm is more appropriate for the Spanish language than the classical paradigm, in which rhyming of written words is judged, because judging whether Spanish written words rhyme can be done visually, on the basis of words' orthography and without evoking inner speech (e.g., avióncamión). There were 40 pairs of drawings in this experiment, with 20 pairs representing objects whose names rhyme in Spanish and 20 pairs representing objects whose names do not rhyme in Spanish. The drawings were selected from Snodgrass \& Vanderwart (1980), based on a standardization for Spanish (Sanfeliu \& Fernandez, 1996), considering name agreement, image agreement, familiarity, and visual complexity. The stimuli were created from the drawings with the highest ratings in Spanish-speakers.

Using the same pairs of words from this experiment, we tested participants' overt rhyming abilities at the end of experimental session. In this overt rhyming judgment task, the experimenter read aloud pairs of words and participants' task was to judge if the words within each pair rhymed.

Experiment 2: The syllable discrimination task required a silent discrimination of syllables in verbally presented words ( $\mathrm{n}=40$ ). The stimuli consisted of sets of randomized 2-syllable ( $\mathrm{n}=13$ ), 3-syllable ( $\mathrm{n}=15$ ), and 4 -syllable ( $\mathrm{n}=12$ ) highly frequent Spanish words, such as bueno. The experimenter read words and the participants were required to determine the number of syllables in each word.

Experiment 3: The following test required silently discerning words in compounds. The stimuli consisted of compounds ( $\mathrm{n}=20$ ) and simple words $(\mathrm{n}=20)$. Like in previous experiments, the stimuli included only highly frequent Spanish words, such as girasol and were presented in a randomized order. The experimenter read words and the participants were required to determine the number of words in each compound.

Experiment 4: The final test required silently discerning words in names for numbers. The stimuli ( $\mathrm{n}=20$ ) included trials that allowed 1, 2, 3, or 4 words to be discerned. For example, eight contains only one word, whereas in fifty-six two additional words can be discerned: fifty, six. The experimenter read number words and the participants were required to determine the number of words in each.

Except for the silent rhyming task, the stimuli in the experiments were presented verbally.

## Procedures

Before evaluation began, each participant signed informed consent. Prior to each experiment, participants completed two to four practice trials. They were instructed to avoid using strategies, moving the mouth or tongue, and to use inner speech in all tests. The experimenter would read the stimuli, except in Experiment 1 in which the stimuli were presented visually, and participants were required to respond as accurately as they could. There was no time limit for answers. Participants' responses for each experiment were recorded manually on a separate response form and later scored for accuracy. No feedback was provided during testing.

Testing was carried out individually with each participant in a quiet room, at ATeCe (patients) and at their homes (HCs). Each patient was assessed in 3 sessions. The first two sessions were devoted to evaluative measures (one session with the certified speech pathologist, C.L., one session with the neuropsychologist, E.U.G.) and the last session was devoted to the experiments. Each session lasted approximately 1 hour. HCs were tested in a single session.

The study was approved by the Basque Ethics Committee for Clinical Research as well as by the ethics committee of the University of the Basque Country. The study was conducted in accordance with the Helsinki Declaration guidelines on studies involving human subjects

## Statistical analyses

We performed Mann-Whitney test for group comparisons in experimental tasks and the modified -t-test (Crawford \& Howell, 1998) to compare performance in experimental tasks of the patient with mixed aphasia to the mean scores of the HCs group; the latter tests were one-tailed, as recommended (Crawford \& Howell, 1998), with $\alpha$ set at 05 . All statistical analyses were carried out in SPSS 22, except for the modified t-test which was performed using the SINGLIMS program (Crawford \& Howell, 1998).

## Results

The two groups did not differ considerably in age ( $\mathrm{t}(19)=-$ $1.433, \mathrm{p}=.168$ ) or in years of formal education ( $\mathrm{t}(19)=-$ $.560, \mathrm{p}=.582$ ). However, they differed in general cognitive status ( $\mathrm{t}(19)=-7.213, \mathrm{p}<.005$ ), with the aphasic group having achieved a mean score of $17.8( \pm 2.5)$ on MoCA and HCs having a mean of $25.1( \pm 2.1)$.

Each patient achieved $100 \%$ correct scores on the phoneme discrimination task. Their performance on Raven Progressive Color Matrices (RPCM) and verbal working memory (vWM) test varied, revealing different degrees of deficit in non-verbal intelligence and vWM capacity respectively across patients. The results of these tests are summarized in Figure 1.


Fig 1. Aphasic speakers' performance on RPCM and vWM.
Furthermore, patients' performance on the Varieties of Inner Speech Questionnaire revealed different degrees of use of the four types of inner speech (Fig. 2). The condensed inner speech featured prominently, while inner speech involving others was less present in their everyday spontaneous use of inner speech.


Fig 2. Aphasic speakers' use of four types of inner speech in everyday life.

Looking at the results from experimental measures, we found that, compared to HCs, the aphasic group performed considerably worse on both rhyming tests (silent and overt), and on the test of discrimination of syllables within words (Table 1).Their ability to discern numbers and words in compounds was comparable to that of HCs.

Table 1: Group differences across the experimental tasks.

|  | $U$ | $Z$ | $p$ |
| :--- | :--- | :--- | :--- |
| Silent ryyming | 4.500 | -3.475 | $.001^{*}$ |
| Overt rhyming | 1.500 | -3.701 | $.005^{*}$ |
| Syllables | 25.500 | -1.985 | $.047^{*}$ |
| Compounds | 32.000 | -1.513 | $.13 / .05^{*}$ |
| Numbers | 43.500 | -.631 | .53 |

However, since the modified $t$-test has shown that the person with mixed aphasia, A02, performed well in all tasks
relative to HCs (silent rhyming, $p=.16$; overt rhyming, $p$ $=.07$; syllables, $p=.33$; compounds, $p=.18$; numbers, $p$ $=.5$ ), we repeated analyses excluding this patient from the aphasic group.

The overall pattern of results remained unchanged, except that the $p$ value obtained in testing for differences in distinguishing words in compounds reached statistical significance ( $U=22.000, Z=-1.93, p=.05$ ). Thus, our experimental data indicate that the persons with motor aphasia had overall impaired inner speech, except when inner speech involved words for numbers.

## Discussion

The goal of the present study was to assess whether poststroke patients with motor aphasia have deficits in inner speech. To answer this question, we tested eight patients with chronic aphasia using four tests of inner speech, including the traditional tests such as silent rhyming judgments and syllable discrimination, together with the tests of word and number discrimination. The main finding of the present study is that post-stroke patients with motor aphasia have impaired inner speech. An additional, unexpected finding is their preserved ability to silently discern numbers in number words. The two findings are discussed in turn.

Since all the aphasic patients achieved $100 \%$ correct on the phoneme discrimination task (see Results), the deficit in silent rhyme judgments cannot be due to impaired ability for phonological discrimination. Instead of being solely due to inner speech impairment, the silent rhyme judgments impairment could be related to anomia, i.e. word finding difficulty which to some extent characterizes all types of aphasia (Benson \& Ardila, 1996). A deficit in retrieval of words referring to the presented images would preclude making judgments on whether a pair of words referring to a pair of drawings rhyme or not. Alternatively, a relatively high working memory load in the silent rhyming task could explain the deficit: the task requires interpreting pairs of drawings, keeping track of semantic information derived from the drawings, retrieving appropriate words, and maintaining the retrieved words in working memory, while judging whether the words rhyme.

Furthermore, the aphasic group had low scores in overt rhyme judgments, a task in which the experimenter read aloud pairs of words. Although this task imposes a smaller cognitive load compared to the silent rhyme judgment task, it still requires a certain amount of working memory capacity, a cognitive resource that appears to be deficient in the aphasic group (Fig. 1). Taken together, these results indicate that the aphasic group's deficit in judging rhymes both covertly and overtly may be due to the verbal working memory impairment. This explanation is compatible with the aphasic group's low scores on the subtests of MoCA related to language and/or memory for language, including repetition of sentences and verbal fluency. However, although limited vWM capacity may have affected silent
rhyming in the aphasic group, it does not explain their overall poor scores on inner speech tasks, involving also impaired discrimination of syllables in words and impaired discrimination of words in compounds. These scores indicate an impairment of inner speech, which could have been further exacerbated by an increased verbal WM load and anomia in the silent rhyming task.

The unexpected finding that the aphasic group was successful in discerning numbers in number words is not only interesting in itself, but it also has implications for the theories on number processing. It suggests that the aphasic speakers relied on a nonverbal, digital representation of numbers (e.g. "57") and not on spoken numerals representations ("fifty-seven"). This explanation is compatible with the proposals suggesting that number processing is mediated by modality-specific processes, e.g., verbal code vs. digits (Kadosh \& Walsh, 2009). While the neural substrates supporting the two types of processing differ, it remains unclear whether the differences pertain to these representations supporting inner speech as well.

Wernicke originally proposed that the left superior temporal gyrus (STG) supported auditory word-form recognition, monitoring of speech output generated by frontal regions as well as inner speech, while contemporary models suggest a functional distinction between the anterior STG, which supports auditory word-form recognition, and posterior STG, which regulates speech production, including inner speech (DeWitt \& Rauschecker, 2013). Neuroimaging studies have associated inner speech with a range of brain areas, including the left supramarginal gyrus, posterior STG, middle temporal gyrus and the inferior frontal gyrus (Paulesu et al., 1993; Indefrey \& Levelt, 2004; Geva et al., 2011b; Fama et al., 2017). A recent voxel-based lesion-symptom mapping study involving 40 lefthemisphere post-stroke patients indicates that inner speech, at the phonological access stage of speech production, is supported by the posterior STG and adjacent areas (Pillay et al., 2014), and that the left inferior frontal gyrus in fact supports working memory and control processes associated with inner speech. Our data appear to support this model, although we would interpret the working memory deficit in our aphasic group in terms of affected network connectivity (Kljajevic, 2014), because all the patients had deficient vWM , regardless of each case's specific site of lesion.

As pointed out in the Introduction, studying phonological word form as a proxy of inner speech provides only a part of the answer to the question of whether inner speech is impaired in motor aphasia. A quick look at the patients' subjective reports on their use of inner speech in everyday life (Fig. 2) suggests a rather complex picture. It suggests that, as in impaired overt speech - where some degree of communication may take place despite the deficit-, impairment of certain aspects of inner speech may still allow use of its other forms. It is not surprising that aphasic speakers in our study showed the most use of the condensed type of inner speech in everyday life. The condensed form is closest to the type of inner speech originally described by

Vygotsky (1986): compressed, and not necessarily resembling overt speech.

In conclusion, our data indicate a degree of inner speech impairment in post-stroke patients with motor aphasia, regardless of lesion distribution (frontal, fronto-temporal, deep grey matter structures). Like their overt speech, their inner speech appears to be affected by difficulties in word retrieval and reduced verbal WM capacity.

## Acknowledgments

This study was funded by a grant from IKERBASQUE, Basque Foundation for Science 111407EMDD (V.K.), IT769-13 (Basque Government) (A.V.) \& FFI2014-52196-P (Spanish MINECO) (A.V).

## References

Alderson-Day, B. \& Fernyhough, C. (2015). Inner speech: development, cognitive function, phenomenology and neurobiology. Psychological Bulletin 141, 931-965.
Alderson-Day, B., Weis, S., McCarthy-Jones, S., Moseley, P., Smailes, D. \& Charles Fernyhough, C. (2015). The brain's conversation with itself: neural substrates of dialogic inner speech. Social Cognitive and Affective Neuroscience, doi: 10.1093/scan/nsv094.
Almairac, F., Herbet, G., Moritz-Gasser, S., Menjot de Champfleur, N. \& Duffau, H. (2015). The left inferior fronto-occipital fasciculus subserves language semantics: a multilevel lesion study. Brain Struct Funct 220, 19831985.

Almor, A., MacDonald, M. C., Kempler, D., Andersen, E.S., \& Tyler, L. K. (2001). Comprehension of long distance number agreement in probable Alzheimer's disease. Language \& Cognitive Processes, 16(1), 35-63.
Ardila, A., Rosselli, M. \& Puente, A. E. (1993). Neuropsychological evaluation of the Spanish speaker. Plenum Press: New York.
Baddeley A. D., Logie R. H. (1999). Working memory: the multiple component model. In: Miyake A., Shah P. (Eds), Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. New York, NY: Cambridge University Press, pp. 28-61.
Baldo, J.V., Paulraj,S.R., Curran, B.C. \& Dronkers, N.F. (2015). Impaired reasoning and problem-solving in individuals with language impairment due to aphasia or language delay. Frontiers in Psychology 6: 1523.
Beck, A.T., Ward, C. H., Mendelson, M., Mock, J., \& Erbaugh, J. (1961). An inventory for measuring depression. Archives of General Psychiatry, 4, 561-571.
Benson, F.D. \& Ardila, A. (1996). Aphasia: a clinical perspective. Oxford: Oxford University Press.
De Bleser, R. \& Marshall, J.M. (2005). Egon Weigl and the concept of inner speech. Cortex, 41, 249-257.
De Guerrero, M.C.M. (2005). Inner Speech - L2: Thinking Words in a Second Language. New York: Springer.

DeWitt, I. \& Rauschecker, J.P. (2013). Wernicke's areas revisited: parallel streams and word processing. Brain \& Language 127, 181-191.
Fama, M.E., Hayward, W., Snider, S.F., Friedman, R.B., Turkeltaub, P.E. (2017). Subjective experience of inner speech in aphasia: preliminary behavioral relationships and neural correlates. Brain \& Language 164, 32-42.
Feinberg, T.E., Gonzalez-Rothi, L.J. \& Heilman, K.M. (1986). "Inner speech" in conduction aphasia. Arch Neurol 43, 591-593.
Frith C. (1992). The Cognitive Neuropsychology of Schizophrenia. Hover: Lawrence Erlbaum Associates.García-Albea, J.E., Sánchez Bernardos, M.L. \& Del Viso, S. (1996). Test de Boston para el diagnóstico de la afasia: adaptación española. Editorial Médica Panamericana, S.A., Madrid.
Geva, S., Bennett, S., Warburton, E.A. \& Patterson, K. (2011a). Discrepancy between inner and overt speech: Implications for post-stroke aphasia and normal language processing. Aphasiology 25, 323-343.
Geva, S., Jones, S., Crinion, J.T., Price, C.J., Baron, J.-C. \& Warburton, E.A. (2011b). The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. Brain, 134, 3071-3082.
Goodglass, H. \& Kaplan, E. (1983). Boston Diagnostic Aphasia Examination. Philadelphia: Lea \& Febiger.
Goral, M., Clark-Cotton, M., Spiro, A., Obler, L.K., Verkuilen, J. et al. (2011). The Contribution of Set Switching and Working Memory to Sentence Processing in Older Adults. Exp Aging Research, 375(5), 516-538.
Heilman, K.H. (2006). Aphasia and the diagram makers revisited: an update of information processing model. Journal of Clinical Neurology, 2, 149-162.
Indefrey, P \& Levelt, W.J. (2004). The spatial and temporal signatures of word production components. Cognition 92, 101-144.
Kaplan, E., Goodglass, H. \& Weintrab, S. (1983). The Boston naming test. Philadelphia: Lea \& Febiger.
Kadosh, C. \& Walsh, V. (2009). Numerical Representation in the Parietal Lobes: Abstract or not Abstract? Behav Brain Sci., 32, 313-28.
Kinsbourne, M. (2000). Inner speech and the inner life. Brain \& Language 71, 120-123.
Kljajevic, V. (2014). White matter architecture of the language network. Translational Neuroscience 5, 239-252
Langland-Hassan, P., Faries, F.R., Richardson, M.J. \& Dietz, A. (2015). Inner speech deficits in people with aphasia. Frontiers in Psychology 6:528, doi: 10.3389/fpsyg. 2015.00528

Levine, D.N., Calvanio, R., Popovics, A. (1982). Language in the absence of inner speech. Neuropsychologia 20, 391-409.
Levelt, W.J.M. (1995). Speaking. From intention to articulation. The MIT Press: Cambridge, MA.
Marvel, C.L. \& Desmond, J.E. (2012). From storage to manipulation: how the neural correlates of verbal working
memory reflect varying demands on inner speech. Brain \& Language 120, 42-51.
McCarthy-Jones, S. \& Fernyhough, C. (2011). The varieties of inner speech: links between quality of inner speech and psychopathological variables in a sample of young adults. Conscious Cogn. 20(4):1586-93.
Martínez-Manrique, F. \& Vicente, A. (2010) What the...!' The role of inner speech in conscious thought. Journal of Consciousness Studies, 17, 141-67.
Morin, A. (2009). Self-awareness deficits following loss of inner speech: Dr. Jill Bolte Taylor's case study. Consciousness and Cognition, 18, 524-529.
Morin, A. \& Michaud, J. (2007). Self-awareness and the left frontal gyrus: inner speech use during self-related processing. Brain Research Bulletin 74, 387-396.
Nasreddine, Z.S., Phillips N.A., Bédirian, V., Charbonneau, S., Whitehead, V. Collin, I. et al. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool for Mild Cognitive Impairment. Journal of the American Geriatrics Society, 53, 695-699.
Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropshycologia 9, 97-113.
Oppenheim, G.M. \& Dell, G.S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. Cognition 106, 528-537.
Paulesu, E., Frith, C.D. \& Frackowiak, R.S.J. (1993). The neural correlates of the verbal component of working memory. Nature 362, 342-345.
Perrone-Bertolloti, M., Rapin, L., Lachaux J.P., Baciu, M., Lœvenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. Behav Brain Res.261, 220-239.
Pillay, S.B., Stengal, B.C., Humphries, C., Book, D.S. \& Binder, J.R. (2014). Cerebral localization of impaired phonological retrieval during rhyme judgment. Ann Neurol, 76, 738-746.
Raven, J. C., Court, J. H., \& Raven, J. (1990). Manual for Raven's progressive matrices and vocabulary scalessection 2: Coloured progressive matrices. Oxford: Oxford Psychologists Press.
Sanfeliu, C.M. \& Fernandez, A. (1996). A set of 254 Snodgrass-Vanderwart pictures standardized for Spanish: Norms for name agreement, image agreement, familiarity, and visual complexity. Behavior Research Methods, Instruments, \& Computers, 28(4), 537-555.
Snodgrass, J.G. \& Vanderwart, M. (1980). A standardized set of 254 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6, 174-215.
Vygotsky, L. (1986). Thought and language. Cambridge, MA: The MIT Press.

# The Impact of Population Structure on Models of Language Change 

Jordan Kodner<br>University of Pennsylvania, Philadelphia, Pennsylvania, United States<br>Christopher Cerezo Falco<br>University of Pennsylvania, Philadelphia, Pennsylvania, United States


#### Abstract

The dynamics of language evolution and learning in individuals have been extensively studied. Our knowledge of the transmission process in particular has been advanced by the iterated learning model. Additionally, work has been done in the area of population structure and social networks. However, less has been described about the interaction between individuallevel transmission and network structures. We present a general framework for representing transmission and learning algorithms within social networks. We demonstrate that population structure interacts with the transmission process to influence the dynamics of change. Taking network effects into account, studies on language evolution will capture a fuller picture of the phenomenon.


# Eye movements during reference production: Testing the effects of perceptual grouping on referential overspecification 

Ruud Koolen (r.m.f.koolen@uvt.nl)<br>Tilburg center for Cognition and Communication (TiCC), Tilburg University PO Box 90153, 5000 LE, The Netherlands<br>Yorick Fliervoet (yorick@fliervoet.com)<br>Tilburg center for Cognition and Communication (TiCC), Tilburg University<br>PO Box 90153, 5000 LE, The Netherlands


#### Abstract

When referring to a target object in a visual scene, speakers are assumed to consider certain distractor objects that are visible to be more relevant than others. However, previous research that has tested this assumption has mainly applied offline measures of visual attention, such as the occurrence of overspecification in speakers' target descriptions. Therefore, in the current study, we take both online (eye-tracking) and offline (overspecification) measures of attention, to study how perceptual grouping affects scene perception, and reference production. We manipulated three grouping principles: region of space, type similarity, and color similarity. For all three factors, we found effects, either on eye movements (region of space), overspecification (color similarity), or both (type similarity). The results for type similarity provide direct evidence for the close link between scene perception and reference production.


Keywords: Reference production; Perceptual grouping; Eyemovements; Overspecification; Visual scene perception.

## Introduction

Suppose you want to point out the marked object in Fig. 1 to a listener. To complete this task, you should produce a referring expression such as "the small bowl" or "the small green bowl", to distinguish the target object from the other objects that are present in the visual scene (the distractors). Although both above example expressions allow the listener to identify the target, the second one is overspecified: it contains a color attribute that is unnecessary for unique identification.


Figure 1: An example visual scene (Koolen et al., 2014)
From prior research (e.g., Pechmann, 1989; Koolen, Goudbeek, \& Krahmer, 2013; Rubio-Fernàndez, 2016), it is known that speakers overspecify their referring expressions very frequently. Why do they do so? We argue that at least one of the
answers to this question is to be found in visual scene perception, and explore to what extent certain objects in a scene are more likely to be perceived than others. For example, in Fig. 1, the plate on the sideboard might be overlooked because it is placed on a different surface than the target (i.e., sideboard rather than table), or because it has a different type (i.e., plate rather than bowl). In these cases, the distractor set would be limited to the large bowl, making a minimal description such as "the small bowl" likely to be uttered. On the other hand, if the plate on the sideboard catches attention anyway, for example because it has a different color than the target object, the perceived color variation may cause speakers to overspecify with color (Koolen et al., 2013).

Although there is growing awareness that scene perception and language production are indeed closely linked, previous research in this direction has generally taken indirect, offline measures of visual attention. Therefore, in the current paper, we combine online (eye-tracking) and offline (occurrence of overspecification) measures to search for structural relations between scene perception and attribute selection for referring expressions.

## Theoretical background

The starting point of our research is the assumption that in a reference production task, speakers do not regard all objects in a visual scene to be relevant distractors, but rather rely on a subset of distractor objects. More specifically, speakers are expected to only consider the distractors that are in their focus of attention (Beun \& Cremers, 1998). One can think of various factors that determine whether an object is perceived or not, such as its physical distance to the target (i.e., proximity). Given that proximity predicts that only objects that are close to the target referent are in the speaker's focus of attention, it can influence the composition of the distractor set for a visual scene (Clarke, Elsner, \& Rohde, 2013a).

Proximity is one of the Gestalt laws of perceptual grouping that were originally introduced by Wertheimer (1923), next to similarity, closure, continuation, and pragnanz. These laws are principles of perceptual organization that serve as heuristics: mental shortcuts for how we perceive the visual environment (Wagemans et al., 2012) and create meaningful groups of objects that we see around us (Thórisson, 1996). On top of the classical laws of grouping, Palmer (1992) defined another principle, common region of space, which holds that objects
that fall within an enclosing contour, such as a table surface, are usually perceived as a group as well.
This study will apply a manipulation of common region of space, as well as two manipulations of similarity: color similarity and type similarity. Previous research that directly tests how these principles influence reference production is scarce. For color similarity, we know that speakers overspecify more often when they perceive color variation in a scene than when all objects are of the same color (Koolen et al., 2013; RubioFernàndez, 2016). This effect of color variation interacts with type similarity: the proportion of overspecification is highest when there is at least one distractor object that shares its type with the target, but not its color (Koolen, Krahmer, \& Swerts, 2016). Also common region of space has been found to affect referential overspecification, as revealed by Koolen, Houben, Huntjens, and Krahmer (2014). In their experiment, Koolen et al. used scenes such as the one depicted in Fig. 1, displayed in both 2D and 3D. The target was always on the table, and mainly for the 3D scenes - speakers overspecified more often when a differently colored distractor was also on the table (in the same group as the target) rather than on the sideboard (in a different group), although the physical distance between the objects was the same in both scenarios.

Crucially, the above papers, as well as many others studies on reference production (e.g., Clarke et al., 2013a), have used indirect measures of visual attention, such as the occurrence of overspecification. This is problematic in studying how the distractors in a visual scene shape attribute selection. For example, although the experiment by Koolen et al. (2014) suggests that region of space affects overspecification, there is no direct evidence that this result is due to the way in which speakers might ignore distractors that are not in the same region as the target referent. Therefore, in the current research, we collect eye movements as a direct, online measure of visual attention, and combine these data with a more traditional, offline analysis of referential overspecification.

While eye-tracking methodologies are very commonly used to investigate language comprehension (e.g., Tanenhaus, Spivey, Eberhard, \& Sedivy, 1995), they are still rare in language production research, initially because speech movements can disrupt eye movement data (Pechmann, 1989; Griffin \& Davison, 2011). After some early studies that explored the effect of object fixations on order of mention (e.g., Griffin \& Bock, 2000; Meyer, Sleiderink, \& Levelt, 1998), some researchers recently started to apply eye-tracking to test the effects of perceptual and conceptual scene properties on rather open-ended descriptions (Coco \& Keller, 2012; 2015) and object naming (Clarke, Coco \& Keller, 2013b). However, none of this work has tested systematically how perceptual grouping affects attribute selection for reference production.

## Current study

To study how different manipulations of perceptual grouping affect reference production, we conducted an experiment in which speakers described target objects in visual scenes. The stimuli were taken from Koolen et al. (2014), for the sake of comparability. We recorded both the participants' speech as
well as their eye movements during the reference production task. Speech data were annotated for the occurrence of overspecification; i.e., if descriptions contained a redundant color attribute. This variable served as a replication of Koolen et al. (2014). New in our study are the eye-tracking data. Here, we analyzed the number of fixations on the distractor we manipulated, and the total gaze duration for that object.

For region of space, we hypothesize that if a distractor is in the same region of space as the target, it is viewed more often and longer than if the region of space is different, and that this will eventually lead to more overspecification. The same goes for type similarity, with more views, longer viewing time and more overspecification for a distractor of the same rather than a different type than the target. Lastly, for color similarity, we expect to find that a distractor most likely attracts attention if it has a different color than the target, resulting in more views, longer viewing times, and again more overspecification than for a distractor that shares its color with the target.

## Method

## Participants

Thirty-one participants ( 26 female, mean age: 21.6) took part in the experiment. The participants were gathered randomly at the campus of Tilburg University, and received a piece of candy as a reward. All participants were native speakers of Dutch, the language of the experiment.

## Materials

The stimulus material consisted of near-photorealistic visual scenes like the example scenes presented in Fig. 2 on the next page. As noted above, the scenes were taken from the related previous study by Koolen et al. (2014). They depicted a living room containing a dinner table and a sideboard, and some objects such as chairs for a more realistic look. The scenes were modeled and rendered using Maxon's Cinema 4D.

The table and the sideboard formed the two surfaces (i.e., regions of space) that were important for our manipulations, since these were the spaces where the target and its two distractors were positioned. The target object always occurred on the table, in the middle of the scene, together with a distractor close next to it (either left or right). This first distractor object always had the same type and color as the target object, but a different size. This way, the distractor ensured that size was always needed for a distinguishing description, and that mentioning color thus resulted in an overspecified referring expression. The scenes also had a second distractor object, by means of which our three manipulations of perceptual grouping were realized.
Firstly, there was a manipulation of perceptual grouping in the law of common region of space. This manipulation was operationalized by positioning the second distractor either in the same region of space as the target (i.e., on the table, see the left scenes of Fig. 2), or in a different region (i.e., on the sideboard, see the right pictures of Fig. 2). It is important to note that the physical distance between the target object and the second distractor was the same in both scenarios.


Fig. 2: Examples of critical trials in our experiment. The distractor shares its region of space with the target (i.e., the table) in the left scenes, and is in a different region (i.e., the sideboard) in the right scenes. The distractor has the same type as the target in the upper four pictures, and a different type in the lower four pictures. The distractor has the same color as the target in the first, second, fifth and sixth picture, and a different color in the third, fourth, seventh, and eighth picture.

Secondly, we had two manipulations of perceptual grouping related to the law of similarity. The first one varied the type of the distractor: this type could be the same as the target's type, or different. Example scenes can again be found in

Fig. 2, where the second distractor object (the plate) has a different type than the target object (the bowl) in the lower four trials, while all relevant objects are of the same type in the upper four trials. Another manipulation of similarity was
employed by varying the color of the second distractor: this color could again be the same as or different than the color of the target object (see again Fig. 2 for example scenes).

While Fig. 2 depicts all visual scenes that were created for the bowl, the same was done for three other types of targets: a plate, a mug and a cutting board. The scenes for these four object types were manipulated in all conditions, resulting in eight trials for each object type. Participants were thus presented with thirty-two (four $x$ eight) critical trails. In all trials, the target object could only be distinguished by mentioning type and size; if participants included color, it made the description overspecified.

Two measures were taken to avoid participants from using the same strategy for all critical trials. Firstly, we had thirtytwo filler trials. Although the scenes for these fillers had the same basic set-up as the critical trails, with all kinds of objects placed on the table and the sideboard, there were more objects present, which could all be the target for that scene. Furthermore, since all objects in the filler trials were white, participants were discouraged to use color when referring to the target here.

Secondly, to prevent participants from developing a viewing strategy, we created two versions of the experiment. For both versions, half of the visual scenes for the critical trials were mirrored. In version 1, this was done for the scenes in which the second distractor was in the same region of space as the target object, while in version 2, all the scenes in the different region of space condition were mirrored. Thus, by taking this measure, all participants saw half of the critical trails mirrored.

## Procedure

The experiment took place in a soundproof booth, located in the SensoMotoric Instruments lab at Tilburg University. The eye-tracking measurements were made with a SMI RED250 device, operated by the IviewX and the ExperimentCenter software-packages. The eye-tracker had a sampling rate of 250 HZ . We used the microphone of a webcam to record the descriptions of the participants; the camera was taped off for privacy reasons. The stimulus materials were displayed on a 22 inch P2210 Dell monitor, with the resolution set to 1680x 1050 pixels, with 90.05 pixels per square inch.

After entering the laboratory, participants signed a consent form, and read a first basic instruction stating that they were going to act as the speaker in a language production experiment. Participants were then seated in the soundproof booth, in front of the eye tracker, and their eyes were calibrated using a 9-point validation method. When the calibration was completed successfully, participants were invited to read a second instruction, which was more detailed than the first one, and stated that participants were going to produce oral descriptions of target objects in visual scenes in such a way that these objects could be distinguished from the remaining objects in the scene. It was emphasized that using location information in the descriptions (e.g., "the bowl on the left") was not allowed. After this second instruction, participants completed two practice trials, and had the possibility to ask
questions. Once the procedure was clear, the experimenter left the booth, and the experiment started.
All participants were shown a total of 64 stimuli ( 32 critical trails and 32 fillers) in a random order. The visual scenes were depicted in the middle of the screen, filling $70 \%$ of the available space; the remaining $30 \%$ consisted of a grey border surrounding the scenes. Before every trial, a screen with an ' X ' appeared somewhere in the $30 \%$ contour area. When this X had been fixated for one second, the next visual scene appeared automatically. When fixating the X did not work, participants could make the next scene appear manually by pressing spacebar. The position of the X was different for all trials: they appeared in a random position in the grey border, again to make sure that participants did not develop a viewing strategy. There were 1.6 times more X triggers on the top and bottom row than on the left and right side, in proportion to the $1680 \times 1050$ screen resolution. Once all 64 trials had been completed, participants were instructed to leave the booth. It took around 30 minutes to complete the experiment.

## Research design

The experiment had a $2 \times 2 \times 2$ design with three within-participants factors: region of space (same, different), type (same, different), and color (same, different). Three dependent variables were measured: the occurrence of color in the target descriptions; the gaze duration upon the manipulated distractor in milliseconds per trial per participant; and the number of times that the manipulated distractor was fixated per trial per participant.

## Data coding and preparation for analysis

All recorded object descriptions were transcribed and coded for the presence of color ( 0 or 1 ). For the eye-tracking data, we first checked for ill measurements, and excluded the data recorded for one participant from further analysis. We then assigned all fixations to either one out of four areas of interest (AOIs) we defined. There was one AOI for the target, one for the sideboard, one for the central part of the table, and one remainder area. The AOIs for the sideboard and the central part of the table represented the areas where the manipulated distractor could be placed. The remainder area was used for fixations that were not on the target or distractor objects that were present in the scenes. The AOIs where the manipulated distractor could occur were central to our analyses.
The coding process resulted in a separate path file for every participant. These path files were converted into a single file, and loaded into SPSS for statistical analysis. Although there was supposed to be data for 960 target descriptions ( 30 speakers times 32 trials), the data for 24 trials could not be analyzed because either the description or the eye movements were not recorded correctly. The final analysis thus contained data for 936 trials.
While the data for all 936 trials was used to analyze the redundant use of color, we created subsets of the data to analyze gaze duration and the number of fixations. For both variables, we only analyzed the cases where speakers fixated - and thus saw - the manipulated distractor. This was the case in 680 out
of 936 cases. For gaze duration, we then calculated for every trial the total amount of time that the participant looked at the manipulated distractor object, and standardized this score by calculating the $z$-score per trial per speaker. Only the scores in the range of $-3 \leq z \leq 3$ were included in the analysis, which means that scores for 13 cases were filtered out.

For the number of fixations, we created a similar subset of the data, but this time we calculated the number of times that speakers looked at the manipulated distractor for every trial. Again, the $z$-score was calculated, which led to the exclusion of 12 trials that were not part of the final analysis for this variable.

## Results

To test for significance, we performed a series of univariate ANOVA tests. We only report on interactions when they are significant. Given that we used subsets of the data in our statistical analyses, performing repeated measures tests was not possible due to empty cells.

## Results for redundant color use

In general, our speakers included a redundant color attribute in $64 \%$ of the descriptions. The first ANOVA was performed to test if redundant color use was affected by our manipulations of perceptual grouping.

The first factor that we expected to affect the redundant use of color was region of space. However, we did not find a significant effect here ( $F_{(1,927)}=.11$, n.s.) : speakers redundantly used color equally often when the manipulated distractor was in the same $(\mathrm{M}=.64, \mathrm{SE}=.02)$ or a different $(\mathrm{M}=.64, \mathrm{SE}=$ .02) region of space as compared to the target.
For our two manipulations of similarity, we did find effects on the redundant use of color. In these cases, the main effects of type similarity $\left(F_{(1,927)}=9.94, p<.01, \mathfrak{y}_{\mathrm{p}}^{2}=.011\right)$ and color similarity $\left(F_{(1,927)}=5.44, p<.05, \mathfrak{y}_{\mathrm{p}}^{2}=.006\right)$ were due to an increase in redundant color use when the manipulated distractor had the same type as the target, and a different color ( $\mathrm{M}=$ $.77, \mathrm{SE}=.03$ ). The other three cells were practically indistinguishable (same type - same color: $\mathrm{M}=.61, \mathrm{SE}=.03$; different type - same color: $\mathrm{M}=.60, \mathrm{SE}=.03$; different type different color: $\mathrm{M}=.59, \mathrm{SE}=.03$ ). This pattern resulted in a significant interaction between type similarity and color similarity $\left(F_{(1,927)}=7.47, p<.01, \mathfrak{y}_{\mathrm{p}}^{2}=.008\right)$.

## Results for gaze duration

The second ANOVA was run to analyze if our manipulations of grouping on the total amount of time that speakers looked at the manipulated distractor.

Firstly, there was a main effect of region of space on gaze duration $\left(F_{(1,651)}=215.5, p<.001, \mathfrak{y}_{\mathrm{p}}{ }^{2}=.249\right)$, showing that the distractor object was looked at significantly longer when it occurred in the same $(M=1812.7, S D=60.87)$ rather than a different $(M=466.7, S E=68.6)$ region of space than the target. A similar effect was found for the manipulation of type similarity $\left(F_{(1,651)}=5.06, p<.05, \mathfrak{y}_{\mathrm{p}}^{2}=.008\right)$. For this factor, we found that distractors that shared their type with the target
$(\mathrm{M}=1242.9, \mathrm{SE}=66.2)$ were looked at longer than distractors for which this was not the case $(\mathrm{M}=1036.5, \mathrm{SE}=63.5)$. The third factor, color similarity, did not affect gaze duration: although the distractor was looked at slightly longer when it had the same $(\mathrm{M}=1176.6, \mathrm{SE}=61.6)$ rather than a different $(\mathrm{M}=1102.8, \mathrm{SE}=67.9)$ color than the target, this difference was not significant $\left(F_{(1,651)}=.65\right.$, n.s. $)$.

## Results for number of fixations

The third dependent variable in our experiment was the number of fixations on the manipulated distractor. Again, there were effects of region of space and type similarity, but not of color similarity.
Firstly, when the distractor was in the same region of space as the target object, participants looked at this object significantly more often $(\mathrm{M}=2.04, \mathrm{SD}=.06)$ than when it occurred in a different region of space $(\mathrm{M}=1.56, \mathrm{SD}=.06) ; F_{(1,652)}=$ $33.37, p<.001, \mathfrak{y}_{\mathrm{p}}^{2}=.049$. Similarly, when the distractor was of the same type as the target object, it was fixated more often ( $\mathrm{M}=1.93, \mathrm{SD}=.06$ ) than when it had a different type $(\mathrm{M}=$ $1.67, \mathrm{SD}=.06$ ). Again, we found no effect of color similarity: the distractor's color (same: $\mathrm{M}=1.85, \mathrm{SE}=.06$; different: M $=1.76, \mathrm{SE}=.06$ ) did not influence the number of fixations $\left(F_{(1,652)}=1.15\right.$, n.s. $)$.

## Discussion

The goal of this research was to test how perceptual grouping affects reference production. We combined both online (eyetracking) and offline (occurrence of referential overspecification) measures of visual attention to study the extent to which grouping causes speakers to ignore certain distractors that are present in a visual scene, aiming to connect the observed scan patterns referential overspecification. We had three manipulations of grouping (i.e., common region of space, color similarity, and type similarity), all realized by varying the location and characteristics of one specific distractor object in the visual scenes that were presented to the participants.

The first manipulation that was present in our stimuli made the manipulated distractor object appear either in the same or a different region of space as compared to the target referent. In Koolen et al. (2014), this manipulation led to a significant effect of grouping on overspecification, with more redundant color attributes in the 'same group' condition rather than the 'different group' condition. In the current study, we could not replicate this result: the proportions of overspecification that we found were the same in both conditions. However, we did find effects of region of space in the eye-tracking data: when the distractor was in the same region as the target referent, it was viewed longer and more often than when it was in a different region. This way, region of space (Palmer, 1992) influences the extent to which certain distractors are considered in a reference production task.

The question remains why the patterns for common region of space that we observed in the eye-tracking data were not reflected in effects on overspecification with color, such as found by Koolen et al. (2014). To explain this issue, we refer to some practical differences between the two studies. Firstly,

Koolen et al. (2014) displayed the stimuli on a big television screen, while the current experiment used only $70 \%$ of a computer screen. Perhaps more important was that Koolen et al. found a convincing effect of common region of space for 3D visual scenes, but that the effect was small for 2D scenes. In the current study, only 2D scenes were used, due to the eyetracking paradigm. Given that our 2D scenes led to clear effects of region of space in the eye-tracking data, it would be interesting to test how this grouping principle affects language on variables other than overspecification, such as fluency and speech onset time.

For type similarity, the effect of the manipulation in the reference production data resonates the pattern in the eye-tracking data. When the distractor had the same type as the target, it was viewed longer and more often than when the type was different, and the proportion of overspecified references was higher. These results show direct evidence for the close link between visual scene perception and language production, in line with the few previous studies in this direction (e.g., Coco \& Keller, 2012; 2015; Griffin \& Bock, 2000). For color similarity, we found a significant interaction with type similarity for the speech data, with an increase in overspecification with color when the distractor had the same type as the target, and a different color. This interaction is a replication of Koolen et al. 2016). For the eye-tracking data, there were no significant effects or interactions with color similarity involved, presumably since color differences "pop out" of the scene (Treisman \& Gelade, 1980). As such, there is no strict need for speakers to fixate distractors (repeatedly) in order to perceive their different color.

Finally, we would like to discuss our decision to use subsets of the data for the eye-tracking analyses. In these subsets, we only included data for the trials where the speaker fixated the manipulated distractor object (or at least the AOI where it was occurred) at least once. Thanks to this approach, we could be certain that speakers were most likely aware of the existence of this object, which makes the observed effects of perceptual grouping even more valid: it excludes, for example, measurement errors that occur when speakers change their position in front of the eye-tracker. However, one can also argue that our approach was too strict, because in order to form a description of a target object, it is not necessary to scan all objects in the scene. In future analyses, we aim to refine our paradigm, also by distinguishing various time windows for every trial to test both the structural and temporal relations between scene perception and reference production.

## Acknowledgments

We thank Jan Huntjens and Eugène Houben (www.eyetractive.com) for developing the stimulus materials, and Rein Cozijn for his assistance in programming the experiment and analyzing the data.

## References

Beun, R.J., \& Cremers, A. (1998). Object reference in a shared domain of conversations. Pragmatics \& Cognition, $6(1 / 2), 121-152$.

Clarke, A., Elsner, M., \& Rohde, H. (2013a). Where's Wally: the influence of visual salience on referring expression generation. Frontiers in Psychology, 4, article 329.
Clarke, A., Coco, M. \& Keller, F. (2013b). The impact of attentional, linguistic, and visual features during object naming. Frontiers in Psychology, 4: article 927.
Coco, M. \& Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. Cognitive Science, 36 (7), 1207-1223.
Coco, M. \& Keller, F. (2015). Integrating mechanisms of visual guidance in naturalistic language production. Cognitive Processing 16 (2), 131-150.
Griffin, Z. \& Bock, K. (2000). What the eyes say about speaking. Psychological Science, 11, 274- 279.
Griffin, Z. \& Davison, J. (2011). A technical introduction to using speakers' eye movements to study language. The Mental Lexicon, 6 (1), 53-82.
Koolen, R., Goudbeek, M., \& Krahmer, E. (2013). The effect of scene variation on the redundant use of color. Cognitive Science, 37 (2), 395-411.
Koolen, R., Houben, E., Huntjens, J., \& Krahmer, E. (2014). How perceived distractor distance influences reference production: Effects of perceptual grouping in 2D and 3D scenes. In Proceedings of the 36th annual meeting of the Cognitive Science Society (CogSci). Québec, Canada.
Koolen, R., Krahmer, E., \& Swerts, M. (2016). How distractor objects trigger referential overspecification: testing the effects of visual clutter and distance. Cognitive Science, 40 (7), 1607-1647.

Meyer, A., Sleiderink, A. \& Levelt, W. (1998). Viewing and naming objects: eye movements during noun phrase production. Cognition, 66, B26-B33.
Palmer, S. (1992). Common region: a new principle of perceptual grouping. Cognitive Psychology, 24 (3), 436-447.
Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics, 27, 89-110.
Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. Frontiers in Psychology, 7: 153.
Wagemans, J., Elder, J., Kubovy, M., Palmer, S., Peterson, M., Singh, M., \& Von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and Figure-ground organization. Psychological Bulletin, 138 (6), 1172.
Tanenhaus, M., Spivey, M., Eberhard, K. \& Sedivy (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.
Thórisson, K. (1994). Simulated perceptual grouping: an application to human-computer interaction. Proceedings of the $16^{\text {th }}$ annual conference of the Cognitive Science Society (CogSci), 876-881. Atlanta, Georgia, USA.
Treisman, A. \& Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12, 97-136.
Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. Psychologische Forschung, 4, 301-350.

# Modelling conceptual change as foraging for explanations on an epistemic landscape 

Ismo T Koponen (ismo.koponen@helsinki.fi)<br>Department of Physics, University of Helsinki, P.O. Box 64, FI-00014 Helsinki, Finland<br>Tommi Kokkonen (tommi.kokkonen@helsinki.fi)<br>Department of Physics, University of Helsinki, P.O. Box 64, FI-00014 Helsinki, Finland


#### Abstract

We discuss here conceptual change and the formation of robust learning outcomes from the viewpoint of complex dynamic systems, where students' conceptions are seen as context dependent and multifaceted structures which depend on the context of their application. According to this view the conceptual patterns (i.e. intuitive conceptions) may be robust in a certain situation but are not formed, at last not as robust ones, in another situation. The stability is then thought to arise dynamically in a variety of ways and not so much mirror rigid ontological categories or static intuitive conceptions. We use computational modelling in understanding the generic dynamic and emergent features of that phenomenon. The model shows how context dependence, described here through structure of epistemic landscape, leads to formation of context dependent robust states. The sharply defined nature of these states makes learning to appear as a progression of switches from state to another, given appearance of conceptual change as switch from one robust state to another.


Keywords: Conceptual change; concept learning; epistemic landscape; simulations

## Introduction

Learning scientific knowledge where learners initial, intuitive concepts gradually change towards more scientific ones is known as conceptual change. Conceptual change as an expression for such learning emphasizes the clear transition or even revolutionary-like transformation of learners knowledge during the learning process (Duit \& Treagust, 2003; Ozdemir \& Clark, 2007; Rusanen, 2014). The recently suggested complex dynamic systems view on conceptual change instead of such a picture views students' conceptions as multifaceted structures which depend on the context of their application. In the dynamic systems view the conceptual patterns (i.e. intuitive conceptions) may be robust in a certain situation but are not formed, at last not as robust ones, in another situation. The stability is then thought to arise dynamically in a variety of ways rather than mirroring rigid preconceptions or static intuitive conceptions (Brown, 2014; Gupta, Hammer, \& Redish, 2010; Koponen, 2013; Koponen \& Kokkonen, 2014). What we think as intuitive conceptions may be in fact so strongly dependent on context, instructional settings and individual learning history that such conceptions should be approached as emergent cognitive epiphenomena, which are situational and mirror partially the targeted scientific models forming the basis of the design of instructional settings. In what follows, we refer to such epiphenomenal conceptual structures simply as students explanatory schemes. In this study we discuss how the dynamic systemic view may change our ideas how conceptual change may accrue from emergent
robust learning outcomes. As a concrete example of learning we consider direct current (DC) electrical circuits and empirical results obtained in that context (Koponen \& Kokkonen, 2014; Kokkonen \& Mäntylä, 2017). In this case the target knowledge and learning situation can be modelled as learning a tiered structure of explanatory schemes, where students are expected to learn a simple set of concepts and relational schemes between the concepts. The model is highly simplified and idealized, but it shows how context dependence, described here through structure of epistemic landscape, leads to formation of context dependent robust learning outcomes. Due to sharply defined nature of these states, learning appears as a progression of switches from state to another, giving appearance of conceptual change as switch from one preexisting robust state to another, instead of gradual dynamic change.

## Empirical cases modelled

The research of learning DC-circuits has revealed that the students tend to use very similar types of explanatory schemes. Some researchers of conceptual change attribute these schemes to pre-existing ontological commitments, while some others see them more context dependent and possibly even artefacts of the empirical research setting (Brown, 2014; Gupta et al., 2010; Koponen, 2013; Koponen \& Kokkonen, 2014). Nevertheless, most empirical studies have revealed very similar collections of explanatory schemes although there are differences in details (see (Ozdemir \& Clark, 2007; Gupta et al., 2010; Koponen \& Kokkonen, 2014; Kokkonen \& Mäntylä, 2017) and references therein). The empirical data used here as starting point consists of three different contexts I-III (Koponen \& Kokkonen, 2014; Kokkonen \& Mäntylä, 2017):

- I: Light bulbs in series. Two variants (a single light bulb and two light bulbs) in terms of the brightness of the bulbs are compared. This comparison consists of events $e_{1}$ and $e_{2}$.
- II: Light bulbs in parallel. The first variant is again involves a single light bulb. The second variant involves two light bulbs in parallel. Comparisons yield events $e_{1}^{\prime}$ and $e_{2}^{\prime}$.
- III: Comparison of the brightness of light bulbs in series (I) and in parallel (II). In the first variant, participants compare the brightness of light bulbs in series, and parallel circuits
to the one-bulb case only. In the second variant, participants compare series and parallel cases to each other. This yields events $e_{1}^{\prime \prime}$ and $e_{2}^{\prime \prime}$.
All six different types of events are referred to as an event set $\varepsilon=e_{0}, e_{1}, e_{2}, e_{0}^{\prime}, e_{1}^{\prime}, e_{2}^{\prime}, e_{0}^{\prime \prime}, e_{1}^{\prime \prime}, e_{2}^{\prime \prime}$, with $e_{0}, e_{0}^{\prime}$ and $e_{0}^{\prime \prime}$ representing observations of the brightness of a single light bulb in each context (the brightest light bulb). This set thus describes (formally) the task and how it was sequenced and how students progressed from context to I to III. In what follows $\varepsilon$ is treated as exogenous variable describing the event set, scaled to range $\varepsilon \in[0,1]$ where 1 represent full set of events. Further details about the empirical setup, design and excerpts from the student interviews are reported elsewhere (Koponen \& Kokkonen, 2014; Kokkonen \& Mäntylä, 2017). When details are put aside, in all cases one finds similar types of explanatory schemes ,listed and characterised in Table I.

Table 1: The explanatory schemes $m_{k}$ inferred from the empirical studies (Koponen, 2013; Kokkonen \& Mäntylä, 2017).

| Model | Description |
| :--- | :--- |
| m 1 | The battery as a source of electricity <br> (current or voltage). |
| m 2 | m1+ components consume electricity <br> (current or voltage). |
| m 3 | $\mathrm{m} 2+$ voltage/current over components <br> creates/needs current/voltage. |
| m 4 | m3 refined as scheme based on Ohms law <br> + Kirchhoffs laws KI and KII. <br> m4+components consume electric <br> energy/power (Joule's law) |

Explanatory schemes $m_{1}$ and $m_{2}$ are well-known electric current-based intuitive explanatory schemes found in many empirical studies (Koponen \& Kokkonen, 2014; Kokkonen \& Mäntylä, 2017), while $m_{3}$ is partially correct explanation, which takes into account the role of components in determining the current. Finally, schemes $m_{4}$ and $m_{5}$ are complete and correct (scientific) schemes. The determination schemes D1 and D2 constraints (Kirchhoff's I and II laws, respectively) and D3 is relational scheme (Ohm's law) regulating the relationship between the pertinent concepts (voltage and current). A more detailed description of these cases and their representation are given elsewhere (Koponen \& Kokkonen, 2014; Kokkonen \& Mäntylä, 2017) and are not reproduced here. The structure of explanatory schemes can be schematically represented as in Fig. 1 as the generic tiered system, where more sophisticated explanatory schemes are at the highest tiers and the less sophisticated schemes at lower tiers can be seen as incomplete or partial versions of the higher tier explanatory schemes.

## The simulation model

The task we discuss here involves five explanatory schemes with ascending complexity and can thus be represented as


Figure 1: A Tiered system of five explanatory schemes. The different hierarchical levels consist of explanatory schemes $\mathrm{m} 1-\mathrm{m} 5$ of ascending level of complexity and expanding coverage of explanatory power. The symbols $C_{1}$ and $C_{2}$ are concepts (current and voltage) entering the models m1-m5.
a tiered structure shown in Fig. 1.The tiered system of explanatory schemes can be represented as an epistemic landscape, which is an abstract representation of the explanatory power of explanatory schemes. Such descriptions have been previously used in studies describing the cognitive and social effects of discovery and knowledge foraging (Weisberg \& Muldoon, 2009; McKenzie, Himmelreich, \& Thompson, 2015). Learning is then described as foraging for best expalining scheme in that landscape, based on utility guided probabilistic selection of the best explaining scheme.

## Epistemic landscape

A tiered system of explanatory schemes consists of schemes $m_{k}, k=1,2, \ldots, 5$, in which the hierarchical level $k$ is defined according to the complexity of the scheme. More complex schemes require greater proficiency from the user of the scheme, such as mathematical proficiency in deriving predictions from the scheme or making deductions based on it. The utility of a given scheme can be seen as a trade-off measure between the scheme's complexity and the amount of events which the learner needs to explain. The scheme $m_{1}$ is simple and, thus, its utility for a simple set of events is high, but decreases for many events to be explained. The scheme $m_{5}$ is the most complex one and requires great proficiency. Because it is complex to use, it has low utility in simple cases, but its utility increases with accumulation of events.

The system of explanatory schemes, as far as the explanatory power of schemes for given set of events is in focus, can be represented in idealized form of epistemic landscape. The epistemic landscape is a simplified description how increased information (in form of events) gives cues to select a given model, and on the other hand, it describes how much proficiency is required in using the model. There is at present
no detailed way to derive the epistemic landscape from the graph as described in Fig. 1 and the connection remains a qualitative one. With these restrictions, however, the epistemic landscape can be constructed by using utility functions $u_{k}(\varepsilon, \kappa)$, which describe the epistemic utility of schemes $m_{k}$. The detailed forms of the functions are, fortunately, not important here; it is enough that they can serve to describe the assumed generic features of the tiered system. Therefore, the mathematical description of the epistemic landscape adopted here is based on a set of suitably flexible functions. Convenient mathematical forms that are easy to use in simulations because the cumulative probability finction is invertible are provided by MinMax-distributions (Kumaraswamydistributions) (Jones, 2009) as given in Table 1. The epistemic landscape thus consists of five manifolds of which Fig. 2 show the schemes with the greatest utility in a given region.

Table 2: The utility functions $u_{k}(\varepsilon, \kappa)$ forming the epistemic landscape. The normalization factors $N_{1}-N_{5}$ are defined so that maximum value of each utility function is 1 . The functions $f_{n, m}(x)=x^{m}\left(1-x^{m-1}\right)^{n-m}$ are MinMax-distributions (Kuwaraswamy-distributions).

| State | Utility function |
| :--- | :--- |
| $m_{1}$ | $u_{1}(\varepsilon, \kappa)=N_{1} f_{n_{1}, m_{1}}(\varepsilon) f_{n_{1}^{\prime}, m_{1}^{\prime}}(\kappa)$ |
| $m_{2}$ | $N_{2}\left[a_{1} u_{1}(\varepsilon, \kappa)+a_{2} f_{n_{2}, m_{2}}(\varepsilon) f_{n_{2}^{\prime}, m_{2}^{\prime}}(\kappa)\right]$ |
| $m_{3}$ | $N_{3}\left[b_{1} u_{2}(\varepsilon, \kappa)+b_{2} f_{n_{3}, m_{3}}(\varepsilon) f_{n_{3}^{\prime}, m_{3}^{\prime}}(\kappa)\right]$ |
| $m_{4}$ | $N_{4}\left[c_{1} u_{3}(\varepsilon, \kappa)+c_{2} f_{n_{4}, m_{4}}(\varepsilon) f_{n_{4}^{\prime}, m_{4}^{\prime}}(\kappa)\right]$ |
| $m_{5}$ | $N_{5}\left[d_{1} u_{4}(\varepsilon, \kappa)+d_{2} f_{n_{5}, m_{5}}(\varepsilon) f_{n_{5}^{\prime}, m_{5}^{\prime}}(\kappa)\right]$ |

## Learning as foraging

The model of learning introduced here assumes that learning takes place as foraging for explanation schemes across the epistemic landscape. We assume that foraging is guided simply by the topography of the epistemic landscape, as a "hill climbing" (HC) in the direction of the steepest change of the gradient of the landscape (McKenzie et al., 2015; Weisberg \& Muldoon, 2009). When exogenous parameter $\varepsilon$ increases by $\delta \varepsilon$ (a new event or cue becomes available), the agent selects the most probable explanatory scheme from the neighborhood of its current state either by switching the state, "uphilling" by increasing the proficiency or, if more advantageous, "downhilling" by decreasing the proficiency. Proficiency is taken here as a skill-like property. A response to success and failure is modelled as logistic development (Steenbeck \& Van Geert, 2007; Van Geert, 2014) of learner's proficiency during the learning process in form

$$
\begin{equation*}
\kappa \leftarrow \kappa \pm \mu \kappa(1-\kappa) \tag{1}
\end{equation*}
$$

where $\mu$ is the effect of memory of success or failure. Here, success means that during foraging learner has upphilled, i.e. moved to direction of increased utility, failure, on the other hand, means that learner has downhilled, moved towards decreased utility.


Figure 2: The epistemic landscape corresponding to explanatory models from $m_{1}$ to $m_{5}$ as indicated in space spanned by events consisting of events $\varepsilon$ and learner's proficiency $\kappa$. The contours are shown for values $0.95,0.90,0.86,0.82,0.78$, $0.74,0.70,0.67,0.60,0.55,0.50$.

## Selection of explanatory scheme

The learners are assumed to select the best explanatory scheme $m_{k}$, one at a time, on basis how its utility compares to utilities of other schemes. The probability $P_{k}$ that scheme $m_{k}$ is selected is based on probabilistic decision theory (Laciana \& Oteiza-Aguirre, 2014; Yukalov \& Sornette, 2014) and is given as

$$
\begin{equation*}
P_{k}=\frac{u_{k} \exp \left[\beta u_{k}\right]}{\sum_{j \neq k} u_{j} \exp \left[\beta u_{j}\right]} \tag{2}
\end{equation*}
$$

where $\beta$ is parameter related to the confidence of choice, $\beta \ll$ 1 indicating low confidence (i.e. high noise or randomness) and $\beta \gg 1$ high confidence (i.e. low noise or randomness). In what follows, we use $\beta=5$ which represents high confidence.

## Implementation of simulations

The control (exogenous) variable is event $\varepsilon$. The output (endogenous) variables are the selected explanatory scheme $m_{k}$ and the learner's proficiency $\kappa$, which changes dynamically as a part of the learning process. The output variables depend on the parameters, which are the confidence $\beta$ and memory $\mu$.

The learning process as foraging across the epistemic landscape is simulated based on the probability of explanatory scheme selection $P_{k}$ in Eq. (2). At each instant when the value of $\varepsilon$ increases by $\delta \varepsilon$ (here $\delta \varepsilon=0.01$ ), it is decided whether: 1) the model switch happens, or 2) proficiency increases, decreases or remains unchanged. Both of these three
steps are characterised by a set of probabilities, and event selection is carried out by the roulette wheel -method (Lipowski \& Lipowska, 2012). In the roulette wheel -method a discrete set of $N$ possible events $k$ with probabilities $p_{k}$ are arranged with cumulative probability $\Phi_{k}=\sum_{i=1}^{k} p_{i} / \sum_{i=1}^{N} p_{i}$. The event $k$ is selected if random number $0<r<1$ falls in the slot $\Phi_{k-1}<r<\Phi_{k}$. In case 1) the probabilities $p_{k}$ are given by Eq. (5) and $p_{k}=P_{k}$ with $k=1,2,3$. In case 2 ) one has three probabilities $p_{1}=P_{k^{\prime}}(\varepsilon+\delta \varepsilon, \kappa)$, $p_{2}=P_{k^{\prime}}(\varepsilon+\delta \varepsilon, \kappa+\delta \kappa)$ and $p_{3}=P_{k^{\prime}}(\varepsilon+\delta \varepsilon, \kappa-\delta \kappa)$ for any given scheme $m_{k^{\prime}}$. All simulations are carried out for equally distributed set of all initial values of $\kappa$, for 100 steps with $\delta \varepsilon=0.01$ and $\delta \kappa=0.01$ in a mesh of $100 \times 100$ points and for 9000 repetitions.

## Results

The outcome of the simulations applied in case of learning the tiered theory structure is number density $n_{k}$ of choice of given scheme $m_{k}$ at given values of $\varepsilon$ and $\kappa$. The number density $n_{k}$ is related to likelihood that in an ensemble of students a given student holds the explanatory scheme $m_{k}$. In case a large set of students' explanatory schemes are collected in an empirical research the density $n_{k}$ would correspond the distribution of how different finding are classified in different categories, categories then roughly corresponding the peaks in the density distribution, while the slight differences in empirically found categories would corresponding the diffuse spread of seen in the density distribution. This association of empirical findings is not exact, of course, but provides a close enough interpretation of the density plots. Note that all density plots are shown as contour plots as in topographical maps.

The shift to hold or select more advanced schemes during the learning (or training sequence) when $\varepsilon$ increases from $\varepsilon=0$ (no events to be explained) to $\varepsilon=1$ (all events to be explained) is particularly clear when density $n_{k}$ of selected schemes in the $(\varepsilon, \kappa)$-space is examined. Such density distributions $n_{k}$ of preferred schemes are shown in Fig. 3 for strong ( $\mu=0.05$ ), intermediate ( $\mu=0.02$ ), and weak ( $\mu=0.01$ ) memory effects. Results are shown only for cases that initially have proficiencies $0.45<\kappa<0.55$ which represents a middle cohort of initial proficiencies, thus representing the assumed average student for which the learning task is designed. The results shown in Fig. 3 demonstrate how selection of given schemes $k$ accumulate to certain regions, different from but close to those regions where utilities (see Fig. 2) have peak values. These regions are shown as dark color in the figures, the darker the shade the higher the density. The dark regions where densities accumulate are the robust outcomes of learning. This behaviour is due to dynamic effects of foraging for best explanatory schemes in the epistemic landscape and how memory affects the development of proficiency.

The density distribution shows directly how likely a selection of given explanatory scheme is in comparison to other schemes. When the memory is weak $(\mu=0.01)$ the low-level schemes $m_{1}$ and $m_{2}$ are likely to be selected throughout the learning sequence. In addition, scheme $m_{3}$ is present through-
out the learning sequence because it is the most preferred initial scheme for mid-cohort learners. When memory increases from $\mu=0.01$ to 0.05 the dynamic evolution becomes more interesting. In the intermediate stage of learning (stage II) scheme $m_{4}$ begins to compete with $m_{3}$ and finally, in the end of the learning stage scheme $m_{5}$ is dominant. For the highest memory $\mu=0.05$ the development becomes very predictable. Schemes $m_{1}$ and $m_{2}$ are likely choices only at low proficiencies, and finally, in the end of the training sequence $\varepsilon_{i} 0.6$ the scheme $m_{5}$ is dominant. For high memory-effects and high confidence the robust learning outcomes are sharply defined, island- like and give expression of well-focused explanatory schemes with no overlap with other explanatory schemes. The overall picture is then that when new event becomes available, learner switches to better explaining schemes towards the end of the learning sequence. This is the successful learning path.

In high memory region, however, the polarization of learning outcomes happens; with increased preference of high level schemes $m_{5}$ also the preference for low level schemes $m_{1}$ and $m_{2}$ tend to increase. This is due to fact that success and failure affect in similar way and have equally strong memory-effect.; success feeds success but similarly also failure feeds failure. Of course, were the memory effect asymmetric, stronger memory effect for success than for failure, such polarization would disappear.

## Discussion and conclusions

In the complex systems view of conceptual change suggested here the formation of robust learning outcomes accrues from foraging on epistemic landscape, which represent the target knowledge as it is contained in the designed learning task. The interplay of learner's cognitive dynamics and the target knowledge as it appears in the design of the learning tasks leads to formation of stable and dense regions of preferred explanatory schemes in epistemic landscape. The origin of these robust states is on the learning dynamics and how it interacts with the context (structure of the learning task). In some cases, depending on the learner's proficiency and the development of the proficiency, learning outcomes may match the target knowledge, but in some other cases, may fall short of targeted outcomes. However, even those states, which do not match the targeted states, are robust, thus giving impression of pre-existing conceptual states of learner, as assumed in traditional conceptual change models. Accumulation of densities $n_{k}$ in certain regions are those areas, where empirical findings will be likely to associate the dynamically formed epiphenomenal robust state with a certain assumed misconception or pre-existing intuitive conception. If this interpretation is correct, the vision of conceptual change as switch between cognitively pre-existing static states to another needs to be revised and replaced by a more dynamic and fluid picture of dynamically formed robust but yet epiphenomenal states.

In the present study, the picture of conceptual change as


Figure 3: The effect of memory $\mu$ on explanatory scheme selection $n$ when events unfold (described as an increasing number of events $\varepsilon$ ). The cases with memory $\mu=0.01,0.02$ and 0.05 are shown. The range $\kappa \in[0.45,0.55]$ of initial proficiencies are considered (mid-cohort). The contours are shown for probabilities $P=0.80,0.70,0.50,0.25,0.15,0.10,0.05,0.02,0.01$, $0.0050,0.0025$. The number of repetitions for each of $100 \times 100$ data points is 9000 .
switch from intuitive conceptions to more scientific conceptions (or sometimes, to other intuitive conception) emerges as rapid but continuous dynamic development of one robust state to another state rather than as abrupt and discontinuous switch from one pre-existing static state to another. Moreover, such states are seen as epiphenomenal outcomes of interplay between learning dynamics and task design, rather than independent construct of mind, rooted in cognitively fundamental, e.g. substance-based ontological categories. The fact that for most of the training sequence there is little overlap between the different robust epiphenomenal states and periods of clearly continuous change are short, a picture of discontinuous switch from robust state to another is obvious. Superficially the course of events in the present model correspond the traditional view of conceptual change but the difference in interpretation of the underlying dynamics and nature of states in present view is fundamentally different from the traditional one; the present view strongly suggests that behind the observed behaviour is after all continuous learning dynamics and which, through designed epistemic landscape, is essentially context dependent.

In summary, the dynamic view provides fresh viewpoint on conceptual change and suggest new ways to conceptualise it. The results we have provided here are far from conclusive and are at best only suggestive, but we think that the view proposed here of learning outcomes as context dependent, dynamically robust but ultimately emergent epiphenomena deserves closer attention and prompts us to design very different empirical research settings. We expect that the main use of the abstract computational model as introduced here is on its potential uses in guiding attention in interdependencies of task structure and learning outcome, and in helping to focus on dynamic, time dependent features of conceptual change in empirical research settings.

## References

Brown, D. E. (2014). Students conceptions as dynamically emergent structures. Science \& Education, 23, 1463-1483. doi: 10.1007/s11191-013-9655-9
Duit, R., \& Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. International Journal of Science Education, 25, 671-688.
Gupta, A., Hammer, D., \& Redish, E. F. (2010). The case for dynamic models of learners' ontologies in physics. The Journal of the Learning Sciences, 19, 285-321. doi: 10.1080/105084062010491751

Jones, M. C. (2009). Kumaraswamys distribution: A betatype distribution with some tractability advantages. Statistical Methodology, 6, 70-81.
Kokkonen, T., \& Mäntylä, T. (2017). Changes in university students' explanation models of dc circuits. Research in Science Education, in print, 1-23. doi: 10.1007/s11165-016-9586-y

Koponen, I. T. (2013). Systemic view of learning scientific concepts: A description in terms of directed graph model. Complexity, 19, 27-37. doi: 10.1002/cplx. 21474
Koponen, I. T., \& Kokkonen, T. (2014). A systemic view of the learning and differentiation of scientific concepts: The case of electric current and voltage revisited. Frontline Learning Research, 4, 140-166. doi: 10.14786/flr.v2i2.120

Laciana, C. A., \& Oteiza-Aguirre, N. (2014). An agent based multi-optional model for the diffusion of innovations. Physica A, 394, 254-265. doi: 10.1016/j.physa.2013.09.046

Lipowski, A., \& Lipowska, D. (2012). Roulette-wheel selection via stochastic acceptance. Physica A, 391, 2193-2196. doi: 10.1016/j.physa.2011.12.004
McKenzie, A., Himmelreich, J., \& Thompson, C. (2015). Epistemic landscapes, optimal search and the division of cognitive labor. Philosophy of Science, 82, 424-453. doi: 10.1086/681766

Ozdemir, G., \& Clark, D. B. (2007). An overview of conceptual change theories. Eurasia Journal of Mathematics, Science \& Technology Education, 3, 351-361.
Rusanen, A.-M. (2014). Towards to an explanation for conceptual change: A mechanistic alternative. Science \& Education, 23, 1413-1425.
Steenbeck, H. W., \& Van Geert, P. L. (2007). A theory and dynamic model of dyadic interaction: Concerns, appraisals, and contagiousness in a developmental context. Developmental Review, 27, 1-40. doi: 10.1016/jdr.2006.06.002
Van Geert, P. (2014). Dynamic modelling for development and education. Mind, Brain, and Education, 8, 57-73. doi: 10.1111/mbe. 12046

Weisberg, M., \& Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. Philosophy of Science, 76, 225-252. doi: 10.1086/644786
Yukalov, V. I., \& Sornette, D. (2014). Selforganization in complex systems as decision making. Advances in Complex Systems, 17, 1450016. doi: 10.1142/S0219525914500167

# Action Understanding in High-Functioning Autism: The Faux Pas Task Revisited 

Joanna Korman (joanna.korman.ctr@nrl.navy.mil)<br>Navy Center for Applied Research in Artificial Intelligence<br>Naval Research Laboratory, 4555 Overlook Ave SW, Washington, DC 20375

Tiziana Zalla (tiziana.zalla@ens.fr)<br>Institut Jean Nicod-CNRS, Pavillon Jardin<br>Ecole Normale Supérieure, 29, rue d'Ulm, 75005, Paris, France

Bertram F. Malle (bertram_malle@brown.edu)
Department of Cognitive, Linguistic, \& Psychological Sciences
Brown University, 190 Thayer St. Providence, RI 02912


#### Abstract

Individuals with autism spectrum disorders (ASD) are said to have deficits in "theory of mind." The present paper explores two main accounts of the mechanisms underlying these deficits. On one account, high-functioning adults with ASD struggle to infer others' mental states. On another account, they lack an ability to integrate those mental states into a coherent understanding of action. We tested these two accounts by making several modifications to the Faux Pas task - a commonly used advanced theory of mind taskincluding the presentation of explicit mental state information. Surprisingly, in contrast to previous work, individuals on the autism spectrum exhibited both intact integration and intact inference.


Keywords: Theory of mind; intentional action; autism spectrum disorder; mental state inference

## Autism and Theory of Mind: Belief and Action Understanding

Autism spectrum disorder (ASD) is a developmental disorder characterized by deficits in reciprocal social interaction and communication (DSM-V, American Psychiatric Association, 2013). In contrast to typically developing children, autistic children are widely described as having deficits in "Theory of Mind" (ToM), or the ability to represent the mental states of other people. Characteristic of these deficits have been ASD children's early failures to pass false belief tasks at the normative age of four years (Baron-Cohen, 1985). But by the time they reach adulthood, many high-functioning adults on the autism spectrum succeed at traditional false belief tasks (Bowler, 1992). Because ASD individuals require higher verbal ability than typically developing adults to pass such tasks, researchers have proposed that ASD individuals do so by using deliberate, conscious calculation (Happé, 1995) and by relying on specific features of language, such as complement syntax (Lind \& Bowler, 2009).

However, high-functioning autistic adults who pass classic ToM tasks persist in their social deficits (e.g., Klin, Jones, Schultz, \& Volkmar, 2003). Therefore, such tasks cannot be capturing these social deficits' core features.

Hence, autism researchers have developed more "advanced" ToM tasks to highlight the persistence of broader theory of mind deficits in more naturalistic settings. Although researchers have succeeded at demonstrating that individuals on the autism spectrum struggle with these novel tasks (Baron-Cohen, 1999; Zalla, Sav, Stopin, Ahade \& Leboyer, 2009), the precise mechanisms underlying these struggles have gone largely unexplored. The present paper explores these mechanisms by examining a modified version of one such task.

## The Faux Pas Task: Revealing Deficits in Adults with ASD

The "faux pas" task (Baron-Cohen, O'Riordan, \& Jones, 1999; Zalla et al., 2009) presents a context in which one character (the speaker) makes a statement that is unintentionally offensive to the listener because the speaker has a false belief. For example, in one story, Jane moves into a new apartment and purchases new curtains for the windows. When her best friend Lisa comes over, she says to Jane, "Oh, I hope you're going to get new curtains! These ones are awful!" Typically developing individuals recognize that Lisa's comment is offensive to Jane, but when asked why Lisa said that, infer that Lisa did not know the curtains were picked out by Jane herself. In contrast, while individuals with ASD can detect that something was "wrong" or "awkward" with Lisa's comment, they struggle to detect that Lisa made the comment unintentionally - that she had a false belief and made the statement out of a positive or neutral desire (e.g., to be helpful with decorating). The outcome is not just an unfortunate side effect of an otherwise fulfilled intention; the complete falsity of the agent's belief actually precludes the fulfillment of the speaker's desire. Unlike controls, adults with ASD demonstrated mixed success in detecting the character's false belief and positive desire. They sometimes acknowledged that the speaker had a positive desire but often failed to correctly infer the speaker's belief state. And in a small number of cases ( $10 \%$ of all responses), they even incorrectly attributed a negative intention to the speaker
(e.g., Lisa wants to insult Jill's taste in décor) (Zalla et al., 2009).

## The Faux Pas Task: What Does it Measure?

What explains the struggles that high-functioning individuals with ASD encounter on these tasks? The faux pas task has previously been described as a more naturalistic and more robust way than 'false belief' tasks of measuring mental state understanding. But understanding the mental states behind the kinds of actions presented in the Faux Pas task can be broken down into several distinct subtasks, outlined below and in Figure 1.
Conceptual integration First, achieving a holistic understanding of the depicted complex behavior requires an understanding of the concept of intentional action. This concept requires the social perceiver to grasp more than just individual mental state concepts of belief and desire: rather, it requires understanding how these individual concepts are integrated to support the understanding of an action as intentional. That is, an action is performed intentionally only if the agent had a desire for the action's outcome and a belief that her action would lead to (serve as a means to achieving) that particular outcome (Malle \& Knobe, 1997). In the present example, for instance, understanding what action was intentional (making a certain remark) and what outcomes were unintentional (the remark offended Jane) requires understanding both that Lisa had a false belief (that the curtains were bought by the previous owner) and that she wanted to be critical only of the previous owner, and that she therefore could not have offended Jane intentionally.
Mental state inference In addition to requiring participants to have a concept of intentional action, the faux pas task requires another capacity: the ability to infer the specific contents of the character's mental states. For example, it is not explicitly stated in the story that Lisa has a false belief about who bought the curtains or that she wanted to be critical of the previous owner; the social perceiver must infer this belief and this desire from the story. Further, there are at least two different mechanisms by which individuals with ASD could be failing to correctly infer the characters' mental states (see Figure 1). On the consensus account, the faux pas task is simply an "advanced" way of revealing enduring deficits in theory of mind - the ability to generate the contents of specific mental states, most particularly, belief - that are already revealed in simpler false belief tasks earlier in life. We refer to this account as the enduring "theory of mind" deficit account.

On an alternate account, however, individuals may struggle on the faux pas task not because they have an enduring deficit in generating the contents of mental states, but because the faux pas task contains an increase in complexity over traditional false belief tasks, thereby disallowing the use of familiar (linguistic) compensation strategies. For example, individuals with ASD may struggle on the faux pas task because this task presents stories in which it is necessary to make a rich suite of background
assumptions about the agent's social roles and context (e.g., the idea that friends normally do not insult their friends' curtains). Such a task does not allow for the use of simple rules (such as "perception leads to knowledge") to generate mental state contents; it instead requires social perceivers to draw productively on their general knowledge to produce accurate inferences. Since individuals with ASD also have documented deficits in this type of knowledge-based inferential generativity (Loth, Gómez, \& Happé, 2008), it is possible that struggles on the faux pas task are due to a general knowledge-based inference deficit, and not any deficit in theory of mind per se.


Figure 1. Capacities required for action understanding in the Faux Pas task.

## The Present Study: Which Hypothesis Explains Low Performance?

To distinguish between the mental state inference hypothesis and the conceptual integration (or intentional action understanding) hypothesis, we developed 8 novel vignettes based on the faux pas task (Zalla et al., 2009; Baron-Cohen at al., 1999) and created four conditions of varying information availability. In the "No Information" condition, we presented participants with no explicit mental state information, as in the original Faux Pas task. In three additional conditions, we presented them with either the character's belief only (Belief condition), the character's desire only (Desire condition), or both the character's belief and desire (Full Information condition). A sample story in the Full Information formulation follows. In the "No Information" condition, the underlined text would be omitted:
Clara is very short and dresses plainly. One day she goes to pick up her son James from school early for a medical appointment. Clara enters the school and spots James's teacher, Mrs. Hayes. Mrs. Hayes thinks that Clara is a student lost in the hallway. [Belief] Mrs. Hayes wants to help [Desire]. Before Clara can ask after James, whereabouts, Mrs. Hayes looks at Clara and says, "Have you lost your class, honey?"

Mental State Inference vs. Conceptual Integration By comparing stories in the three explicit mental state information conditions with the No Information condition, we can broadly distinguish between the inference and
integration hypotheses. If adults with ASD are capable of conceptually integrating mental state information for action understanding but are not capable of inferring this information, they should show improved performance in the presence of explicit mental state information over the "No Information" case. In contrast, if they are not capable of integrating mental state information for intentional action understanding, they should struggle to accurately interpret the meaning of the agents' actions as depicted in the stories even when mental state information is explicitly presented to them.

Two Mechanisms of Mental State Inference In addition to distinguishing broadly between mental state inference and conceptual integration, we also sought to distinguish two possible mechanisms of mental state inference deficit: the commonly cited theory of mind deficit and a nonspecific general knowledge deficit. In previous versions of the Faux Pas task, pieces of general knowledge (such as the fact that, in the above sample story, Clara probably appears to be younger than she really is) were not included in the story and instead had to be inferred - in addition to mental state information. Low performance on the task may thus have been a result of failed general knowledge inferences, not a result of failed ToM inferences. In the present task, we sought to mitigate this ambiguity by explicitly providing such background information in every condition (e.g., "Clara is very short and dresses plainly") and requiring participants to infer only the missing mental state information (e.g., that, in light of Clara's appearance, Mrs. Hayes falsely believed Clara to be a student). If participants continue to struggle to produce belief inferences in the No Information condition in spite of these additions of inference-ready background information, then we can be confident that the present faux pas stories indeed measure only mental state inference (or "theory of mind") capabilities, and not any additional abilities.

Conceptual Integration Apart from inference, by explicitly presenting mental state information, we can test ASD individuals' abilities to integrate this information in the service of intentional action understanding. To demonstrate integration capacity, individuals on the autism spectrum must go beyond understanding the fact that Lisa had a false belief. They must also show understanding of how her false belief is relevant to the action and its outcome: that it was because she had a false belief that she made the remark, and that the remark resulted in offense because of that false belief (Lisa didn't realize that it would lead to a negative outcome when she said it) (Figure 2). Similarly, understanding that Lisa had a positive desire is not sufficient for a full conceptual understanding of the action. In the presence of a positive desire, individuals with ASD must recognize that the action (utterance) still may have caused a negative outcome, even though the desire motivating it was a positive one. In summary, to show
integration capacity, individuals with ASD must be able to see how a story character's mental states relate to her action and its outcomes.

Most centrally, in the presence of both a belief and a desire (Full Information condition), the ability to correctly understand that the action is intended to be a positive one, was caused by the story character's false belief, and has a negative outcome, demonstrates intact integration capacities. In addition, the inclusion of conditions in which either a belief or a desire alone was presented allowed us to test integration abilities under somewhat more difficult conditions. ${ }^{1}$


Figure 2. Conceptual integration requires the social perceiver to recognize the relevance of the story character's mental states to the action's meaning and outcome.

## Study

## Procedure and Measures

We presented control and ASD participants with eight faux pas stories each: two of each in each of the four mental state information conditions (No Information, Belief only, Desire only, Full Information). Participants also received six control stories. Participants read each story and then answered several forced-choice and open-ended questions about the story (detailed below), which served both as measures of inference and integration depending on information condition: in the No Information condition, measures of belief understanding, explanations for the action, and description of the action were a measure of inferential abilities, and in the explicit mental state information condition, these same measures demonstrated participants' abilities to integrate provided mental states into a coherent understanding of action.

Participants 20 participants with Autism Spectrum Disorder (ASD), as confirmed on the ADOS-2 (Lord, Rutter, DiLavore, Risi, Gotham, \& Bishop, 2012), were recruited in partnership with the Rhode Island Consortium for Autism Research and Treatment (RI-CART), $M_{\text {Age }}=31.90$ years; 5 Female; mean score on Ravens Progressive Matrices, 9-item short form intelligence test (Bilker at al., 2012), $M=$ 45.94/60, $S D=8.0 .20$ typically developing controls were

[^369]recruited to match the ASD group on age, gender, and intelligence $\left(M_{\text {Age }}=30.35\right.$; 6 Female; $M_{\text {Ravens }}=49.75, S D=$ 14.29).

Interpretation of the Utterance After reading each story, participants were first prompted to describe the main character's utterance. They were instructed to "check all that apply" among four options: "It was awkward," "It was nice," "It was mean," and "It was neutral."

Belief Question Participants answered whether the character who made the utterance possessed a false belief (e.g., "Did Mrs. Hayes believe that Clara was James's mother?)

Explanation Question Participants then answered in a text box the "explanation" question, which simply asked, "Look back at what [character] said. Why did [s/he] say that?" These open-ended responses were content coded for explanatory quality. Two coders classified each response into a single numbered category, 0-3. To receive a perfect score of 3 , the participant had to give an explanation that directly stated or otherwise implied that the speaker's action was caused by that character's false belief. Incorrect responses, such as those that cited a negative intention on the part of the speaker, received a score of " 0 ".

## Results and Discussion

Interpretation of the Utterance To achieve a data reduction of the 32 cells represented (4 descriptors for each of 4 information conditions, rated by two different participant groups), responses for each of the four variables were first aggregated across each cell of the 2 (autism vs. control) X 4 (mental state information) design. Values on each variable were aggregated across participant group and information condition, yielding one variable for each descriptor, and entered into a principal components analysis. Two orthogonal components were extracted. Component loadings for each of the four variables are provided in Table 1.

Table 1. Principal components analysis of four utterance interpretation options results in two components

| Descriptor | Unintentional | Motive |
| :--- | :---: | :---: |
| Awkward | 0.65 | 0.33 |
| Nice | -0.48 | 0.35 |
| Mean | -0.32 | -0.61 |
| Neutral | -0.08 | 0.45 |

The two components can be best interpreted as capturing (1) whether the utterance was classified as awkwardly unintentional (awkward) vs. intentional (nice or mean) and (2) whether the motive behind the action was classified as nice (or neutral) vs. mean. Component scores for each participant were computed from linear combinations of the
four constituent variables, and ranged in value from -2 to 2 . A score close to +2 on the Unintentional component indicated that "awkward" was checked for both stories in that condition (and that the slight was therefore unintentional), while a score close to -2 indicated that the utterance in that story had instead been classified as intentional. On the Motive component, scores closer to +2 indicated less endorsement of the "mean" descriptor and more endorsement of the "nice" or "neutral" descriptors.

For this and subsequent measures, we performed two main analyses. To test the inference hypothesis, we compared the two participant groups with $t$ tests in the No Information condition only. To test the integration hypothesis, we performed ANOVAs on the four conditions for both participant groups. Of primary interest to the integration hypothesis were interactions between participant group (Autism vs. Control) and information condition (primarily, the three explicit information conditions vs. the No Information condition).

On the Unintentional component (Figure 3), controls recognized the utterance as "awkward" with greater frequency in the No Information condition than did ASD participants, $t(34.72)=2.04, p<.05$. In addition, there was an interaction between Autism and Information condition, but it did not reach significance, $F(1,38)=3.14, p=.08$. On the Motive component (Figure 4), there were no group differences in the No Information condition, nor were there significant main effects or interactions of information condition with participant group.


Figure 3: Unintentional Component
Implications for Inference Hypothesis Compared to control participants, ASD participants were less capable of recognizing that the speaker's utterance could be described as awkward. However, the two groups did not differ in any of the explicit mental state information conditions with respect to this recognition, suggesting that the noinformation difference may be spurious or not due to a deficit on the part of ASD participants. When assessing the possible motives of the speaker, ASD participants exhibited no difficulties identifying the positive desire (lack of 'mean' intent) underlying the speaker's utterance.

Implications for Integration Hypothesis In spite of some differences with controls in the No Information condition, ASD individuals performed similarly to controls in the presence of explicit mental state information, identifying the utterance's awkwardness with similar frequency and the story character's positive desire with similar frequency.


Figure 4: Motive Component
Belief Question Correct responses to the belief question were aggregated across the two stories comprising each of the four information conditions, yielding a score of 0 to 2 per condition (see Figure 5).


Figure 5: Belief Question
To examine the inference hypothesis, we compared performance of ASD participants ( $M=1.65$ stories correct) and Control participants $(M=1.80)$ in the No Information condition. There was no difference in performance, Welch's $t(36.88)=1.05, p=0.30$.

To test the integration hypothesis, we examined whether adding mental state information affected the accuracy of correctly recognizing the belief, again performing a 4 (Information condition) X 2 (Autism vs. Control) mixed ANOVA. There was a main effect of information condition: Participants as a whole exhibited improved performance in the presence of explicit mental state information (in the three explicit mental state information conditions) as compared to the No Information condition, $F(1,38)=4.36$, $p=.04$. No interactions between Autism and Information conditions reached significance.

Implications for Inference Hypothesis In our No Information condition, all general knowledge was made explicit, and ASD participants performed as well as controls. By contrast, in comparable no-information conditions, previous Faux pas studies did not make such background knowledge available and documented performance decrements for ASD participants. This pattern of results supports the hypothesis that previously documented inferential deficits may not be due to deficits in "theory of mind" per se, but may instead have been caused by the inability to draw inferences from general knowledge.

Implications for Integration Hypothesis Neither participants with ASD nor control participants performed at ceiling in the No Information condition, leaving room for improvement (and a demonstration of integration capacities) in the presence of explicit mental state information. Although there was a main effect of information condition, there was no difference between ASD and control participants. Thus, ASD participants appear to be as capable as control participants at integrating explicitly presented belief and desire information.

Explanation Question Once more, there was no difference between ASD and control participants in the No Information condition, Welch's $t(36.52)=.40, p=.70$. We also found no main effects or interactions involving the comparison between ASD and control participants across all four conditions (all $p$ s > 0.36). There were, however, significant main effects of information condition, with all participants providing higher-quality explanations when receiving explicit information about both mental states than in the No Information condition, $F(38)=3.71, p=.06$, and higherquality explanations when receiving explicit information about both a belief and desire than when receiving information about either one of these mental states alone, $F(38)=4.44, \mathrm{p}<.05$.

Implications ASD and control participants provided equally accurate explanations in all information conditions, including the most challenging one (where no explicit belief or desire information was provided). Moreover, like control participants, ASD participants improved their explanation quality in response to explicit mental state information, suggesting in particular that integration capacities held by control participants are also held by those with ASD. Thus, we may conclude once more that previously documented deficits for ASD individuals - in both inference and integration - may have been caused by other aspects of the task, such as the requirement of general knowledge recruitment.

## General Discussion

We considered two main hypotheses that could explain deficits for individuals with ASD in demanding theory of mind tasks such as the Faux Pas task. One suggests that individuals with autism struggle to generate the contents of mental states (inference), while the other suggests that they
struggle to integrate mental state information to reach a full understanding of action (integration). Both hypotheses fall short of explaining our data.

## Mental State Inference Deficits?

Contrary to previous studies of the Faux Pas task (Zalla et al. 2009, Baron Cohen et al., 1999) as well as other similar advanced theory of mind tasks (e.g., Happé, 1994), ASD participants in our study performed comparably to control participants, even when receiving no explicit mental state information. This performance spanned a number of measures, including correct inferences of the story character's false belief and positive desire. It appears that, in the presence of enriched background information to afford inferences from general knowledge, ASD participants more capably inferred mental states than they did in previous studies in which stimuli lacked such enriched background information. Although the present study did not directly compare background-enriched stories with unenriched stories, this finding is suggestive: previously documented deficits on advanced theory of mind tasks may depend on a suite of inferential capacities, of which mental state inference, per se, or "theory of mind," is only one, and perhaps a less influential one.

## Integration Deficits?

In addition to demonstrating intact inferential abilities in the presence of enriched background information, ASD participants in our study also demonstrated intact abilities to integrate provided mental state information into a coherent understanding of intentional action. Even in response to a challenging, open-ended question about the character's utterance- "Why did he say that?" -participants with ASD accurately linked mental states with action as well as controls did.

With the addition of (1) enriched background information and (2) explicit mental state information, high-functioning adults with ASD exhibited a remarkable ability to comprehend the meaning behind a story character's complex intentional action. This finding is notable in light of previous work suggesting that individuals with ASD struggle to reach the requisite mental state inferences (Baron Cohen et al. 1985, Happé 1994), and to integrate a character's mental states with her action's outcome to reach a full comprehension of that action (Moran, Young, Saxe, Lee, O'Young, Mavros, \& Gabrieli, 2011).

## Acknowledgments

Thanks to Corey Cusimano, Baxter DiFabrizio, Stuti Thapa Magar, Lindsay Oberman, Stephen Sheinkopf, and all the staff of the Rhode Island Consortium for Autism Research and Treatment (RICART).

## References

American Psychiatric Association. (2013). Diagnostic and statistical manual (5th ed). Washington, D.C.

Baron-Cohen, S., Leslie, A. M., \& Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21(1),
Baron-Cohen, S., O’Riordan, M., \& Jones, R. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. Journal of Autism and Developmental Disorders, 29, 407-418.
Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., \& Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. Assessment, 19(3), 354-369.
Bowler, D. M. (1992). "Theory of Mind" in Asperger's Syndrome. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 33(5), 877-893.
Happé, F.G.E. (1994). An advanced test of theory of mind: Understanding story characters' thoughts and feelings by able autistic, mentally handicapped and normal children and adults. Journal of Autism and Developmental Disorders, 24, 1-24.
Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. Child Development, 66(3), 843-855.
Klin, A., Jones, W., Schultz, R., \& Volkmar, F. (2003). The enactive mind, or from actions to cognition: lessons from autism. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358(1430), 345-360.
Lind, S. E. \& Bowler, D. M. (2009). Language and theory of mind in autism spectrum disorder: The relationship between complement syntax and false belief task performance. Journal of Autism and Developmental Disorders, 39(6), pp. 929-937.
Lord, C., Rutter, M., DiLavore, P.C., Risi, S., Gotham, K., \& Bishop, S. (2012). Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1-4. Torrance, CA: Western Psychological Services.
Loth, E., Gómez, J. C., \& Happé, F. (2008). Event schemas in autism spectrum disorders: the role of theory of mind and weak central coherence. Journal of Autism and Developmental Disorders, 38(3), 449-463.
Malle, B. F., \& Knobe, J. (1997). The Folk Concept of Intentionality. Journal of Experimental Social Psychology, 33(2), 101-121.
Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O’Young, D., Mavros, P. L., \& Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. Proceedings of the National Academy of Sciences, 108(7), 2688-2692.
Zalla, T., Sav, A.-M., Stopin, A., Ahade, S., \& Leboyer, M. (2009). Faux pas detection and intentional action in Asperger Syndrome. A replication on a French sample. Journal of Autism and Developmental Disorders, 39(2), 373-382.

# Dependent Choices in Employee Selection: Modeling Choice Compensation and Consistency 

Antonia Krefeld-Schwalb (antonia.krefeld-schwalb@unige.ch)<br>Benjamin Scheibehenne (Benjamin.scheibehenne@unige.ch)<br>University of Geneva, Geneva School of Economics and Management, Uni Mail, Bd du Pont-d'Arve 40, CH-1112 Geneva 4, Switzerland

Jörg Rieskamp (joerg.rieskamp@unibas.ch) \& Nicolas Berkowitsch (nicolas.berkowitsch@unibas.ch)<br>University of Basel, Department of Psychology,<br>4055 Basel, Switzerland


#### Abstract

Past choices can influence subsequent choices in employee selection. Previous approaches rather described similar sequential effects with feedback learning or the misperception of randomness. However, in the selection of job candidates also the accumulation of the moral impact of previous choices might influence subsequent choices. We investigated that question by making two major contributions to the literature. First, we developed an experimental paradigm for measuring sequential choices in employee selection and second, we implemented a widely applicable computational model, the Dependent Sequential Sampling Model, for explaining sequential effects in choices. By using this methodological approach, we uncovered sequential effects in employee selection. Participants ( $\mathrm{N}=600$ ) were especially motivated to compensate for morally dubious choices, with some participants showing consistent choice behavior if their previous choices had been morally virtuous. These results support the assumption of asymmetric compensation of morally dubious choices, sometimes referred to as the moral cleansing hypothesis.


Keywords: sequential sampling model; preferential choice; sequential decision making; employee selection.

## Theoretical Background

Ethical and moral aspect play a major role, when managers select new employees. In order to enable a fair employee selection procedure, every candidate is supposed to be evaluated only on her skills and accomplishments relevant for the position in question. This does not only involve the prohibition of any kind of discrimination, but also demands the evaluation of every candidate, independent of other candidates and contextual factors. However, interviewers deviate from such a fair evaluation procedure. Discrimination in the application process for jobs, e.g. based on gender and skin color, is common (Gregory, 2003; Bertrand \& Mullainathan, 2004; Pager, Western, \& Bonikowski, 2009). And instead of evaluating candidates independently, previous candidates on the same day influence the evaluation of later candidates (Simonsohn \& Gino, 2013). The present article tackles this problem. We will present a theory driven experimental method and computational model for investigating sequential effects in the employee selection and other areas. Our approach aims at first identifying sequential effects and second quantifying the individual degree and direction of the effects.

## Sequential Effects in Employee Selection

Many sequential effects in decision making and choices are explained with the "gamblers fallacy", the misperception of randomness (Ayton \& Fischer, 2004; Clotfelter \& Cook, 1993). However, if moral aspects are relevant for the choice, other factors contribute to the sequential effects as well. It is often assumed that the moral credentials of previous choices are accumulated and influence subsequent choices (e.g. Monin \& Miller, 2001; Tetlock, Kristel, Elson, Green, \& Lerner, 2000). This can well be illustrated in the employee selection. Given the common problem of discrimination in the job market (Gregory, 2003; Bertrand \& Mullainathan, 2004; Pager et al., 2009), it is of higher moral value to choose a person belonging to a group discriminated against than a person belonging to a favored group. This can lead to various potential sequential choice effects. If individuals made a series of choices for the same group, e.g the group not discriminated against, they tend to compensate these choices in the following (Conway \& Peetz, 2012; Jordan, Mullen, \& Murnighan, 2011; Sachdeva, Iliev, \& Medin, 2009). While some individuals simply balance morally dubious and morally good choices (Dhar, Huber, \& Khan, 2007; Dhar \& Simonson, 1999; Huber, Goldsmith, \& Mogilner, 2008), less symmetric compensating choice behavior is also possible. For example, especially after morally dubious choices, people feel the urge to compensate for these choices with a morally virtuous choice subsequently, referred to as cleansing (Tetlock et al., 2000). And also the complementary effect has been observed, referred to as licensing (e.g. Monin \& Miller, 2001). Nonetheless, even contrarily to the idea of compensation, consistent choice behavior is also possible (Gneezy, Imas, Nelson, Brown, \& Norton, 2011; Zhang, Cornwell, \& Higgins, 2014).

## Modeling Sequential Choice Effects

In most domains of real life choices and decisions, sequential effects have either been explained with models involving reinforcement learning (e.g. Kruschke, 1992; Gluck et al., 1988; Rieskamp \& Otto, 2006; Simão \& Todd, 2002; Stewart, Brown, \& Chater, 2002; Stewart \& Brown, 2004; Todd, 2007) or the effects are explained with the "gamblers fallacy" (Thaler \& Johnson, 1990; Novemsky \& Dhar, 2005; Chen,

Moskowitz, \& Shue, 2016; Ayton \& Fischer, 2004; Clotfelter \& Cook, 1993). However, these models do not explain asymmetries in compensating previous choices, which are in the moral literature sometimes referred to as moral licensing or moral cleansing, that even occur without explicit feedback. We will introduce a computational model in the following that can account for these effects. The model involves one parameter that indicates the individual degree and direction of sequential effects, it quantifies the tendency for compensation or consistency with previous choices.

## Method

In order to investigate sequential effects in the employee selection, we developed and applied the experimental paradigm "'The Sequential Employee Selection Task". In previous experiments on sequential effects in the job application process and other domains, the investigated choice often followed a different task, e.g. a rating task. Contrarily, participants in our experiment were faced with repeated choice tasks of the same format and in the same context.

After a series of choices between two candidates, who clearly differed in the qualification for the job (dominated trials), subjects were to choose between two equally qualified candidates (ambiguous trials). In some of the conditions candidates from a group discriminated against (discriminated group) dominated the previous trials, and in other conditions candidates from the complementary favored group (non-discriminated group) dominated these trials. The following ambiguous trials were the same between the conditions. Thus, different choice probabilities in these ambiguous trials indicated the influence of the previous trials.

If the choice probabilities in the ambiguous trials systematically differ between the conditions, there exist sequential effects in this task $\left(\mathrm{H}_{1}\right)$. If the probability to choose a candidate from the same, previously dominating, group is decreased, compensating choices are observed $\left(\mathrm{H}_{2}\right)$. If the probability increases, participants make consistent choices $\left(\mathrm{H}_{3}\right)$. If the compensation differs with regard to whether the discriminated or the non-discriminated group dominated the first trials, the compensation is asymmetric and the moral impact of the choices is accumulated. This finding would indicate sequential effects referred to as moral licensing $\left(\mathrm{H}_{2 \mathrm{a}}\right)$ or moral cleansing $\left(\mathrm{H}_{2 \mathrm{~b}}\right)$. If previous choices are compensated symmetrically, choices are balanced $\left(\mathrm{H}_{2 \mathrm{c}}\right)$. In order to estimate initial choice biases and the weights of the candidates' attributes, the manipulated sequences can further be compared to control conditions, in which only the ambiguous trials, the same as in the experimental conditions, were tested. As an additional add-on to previous studies, the sequential effects were not only tested in one single trial, but four ambiguous trials, enabling estimation of the individual degree of compensation.

## Participants

We recruited participants living in the U.S. through amazon's Mechanical Turk (Amazon, 2013). In order to avoid inatten-


Figure 1: Screenshot of one of the ambiguous trials in a condition with skin color as the salient category
tive participants or computer programs filling out the questions, we included an additional test page at the beginning. Of the 635 recruited participants $600(95 \%)$ passed that screening ( $47 \% \mathrm{f}$, mean age $=36.06$ ). Those participants received USD 1.50 for their participation in the 10 minutes experiment.

## Design and Procedure

The Sequential Employee Selection Task was manipulated on 2 factors between the subjects, resuming in 6 between-subject conditions. The sequential effects were investigated in four of these conditions differing with regard to the salient category, skin color vs. gender, and with regard to whether the candidates from the discriminated (female or black candidates) or the non-discriminated group (male or white candidates) dominated the first eight trials. In the two control conditions the baseline choice probabilities for candidates belonging to the respective groups were tested.

Participants were presented with a hypothetical recruiting scenario and asked to repeatedly choose the most suitable candidate out of two job applicants, see Figure 1 for an example. All candidates were described by three attributes on scales between 0-100: "leadership skills", "social competence", and "typing speed". The information was presented in an information board and a profile picture above this board. On top of it, a fictitious company name served as a title of the trial. The first two of the attributes were described as important in the initial instructions, whereas the third was described as a less important attribute for the position. The candidates' individual levels on the attributes are indicated by the numbers in the respective cells of the information board. The profile pictures above the table indicated the gender or the skin color of the candidates, depending on the condition. Pictures from the Chicago Face Database (Ma, Correll, \& Wittenbrink, 2015), only smiling faces, were used for the present experiment. Participants received no instructions with regard to the relevance of these personal characteristics and other features for the position, beyond the instructions on the three

Table 1: Illustration of the series of trials in the gender conditions. In the skin color conditions female faces are replaced with black male faces.

| Dominating group | Dominated trials |  |  |  | Ambiguous trials |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1 | 2 | ... | 8 | 9 | 10 | 11 | 12 |
| Discr | $9>0^{7}$ |  |  |  | $9=0^{\prime}$ |  |  |  |
| Non-discr | $9<0^{7}$ |  |  |  | $9=0{ }^{\prime}$ |  |  |  |
| Control | - |  |  |  | $9=0^{\prime}$ |  |  |  |

Note. Discr/Non-discr $=$ female or black/male or white candidates dominate first 8 trials; Control $=$ only ambiguous trials; $[>,<,=]=$ better, worse, equally qualified.
attributes.

Each participant made twelve pairwise decisions in a row. The description of the task emphasized that the twelve candidate pairs applied for twelve different companies. Furthermore, showing fictitious company names at every decision stressed that each choice ought to be independent of the previous ones. The procedure of the experiment in the gender conditions is illustrated in Table 1. The choice pairs were constructed such that one of the two candidates was clearly superior on the two most important attributes (i.e. leadership skills and social competence) in the first eight pairs (= dominated trials), the two relevant attributes provided higher values for one candidate. Contrarily, the candidates in the last four pairs were equally well suited for the job (= ambiguous trials). In those trials, one of the relevant attributes had a higher value for one candidate, whereas the other attribute provided a higher value for the other candidate. The values of the third attribute were equal between the two. Thus, even when participants weighted all information for making a choice between the candidates, this weighted additive would lead to no preference for one candidate or the other. It was further controlled, that in two of these ambiguous trials the candidates from one group and in the remaining two trials the candidates from other group had a higher value on the first attribute. This procedure allows to control for the common use of a heuristic which places more weight on the first discriminating attribute (Take-the-best heuristic (TTB) Gigerenzer \& Goldstein, 1996). The order of the trials, within the dominated and the remaining ambiguous trials was randomized, and the position of the dominated candidates (left or right) was counterbalanced. However, the first of the ambiguous trials was fixed so that the the candidate from the discriminated group had a higher value on the first attribute in this trial. This was done in order to increase the contrast between the conditions. Subjects were randomly assigned to one of the conditions.

## Formalization of the Dependent Sequential Sampling Model

In order to explain sequential effects in choices it is assumed that some information of previous choices must be accumulated in memory and influence subsequent choices. As for modeling the evidence accumulation models within one trial (e.g. Busemeyer \& Townsend, 1993; Lee, 2004; Newell \& Lee, 2011), sequential sampling models might be used for modeling this accumulation process across trials as well.

In their simplest form, sequential sampling models assume that evidence in a given binary choice task is accumulated from a starting point across a fixed number of steps. After all fixed steps were made, this option with the highest accumulated evidence is chosen. For the purpose at hand, we adapted a similar sequential sampling model (Milosavljevic, Malmaud, \& Huth, 2010), for incorporating previous choices. The starting point of the evidence accumulation process usually represents only any previous bias for one of the options. In the present study, it was enlarged with incorporating previous choices as well. It was further considered 1) that the influence of previous choices decays exponentially with temporal distance and 2) that previous choices could potentially lead to compensation, confirmation or not influence subsequent choices. The latter, the direction and strength of the influence of previous choices, was governed by the compensationconsistency parameter $\kappa$.

Relying on a simple drift diffusion model (Milosavljevic et al., 2010) we assume that the evidence z in a binary decision task evolves according to the following equation across all time steps s:

$$
\begin{equation*}
z_{s}=z_{(s-1)}+\mu+\varepsilon_{s} \tag{1}
\end{equation*}
$$

The evidence which is added at every time step is defined by the speed of evidence accumulation $\mu$, representing the overall difference of the options' values in the trial $\mu=V_{A}-V_{B}$, and an error term $\varepsilon_{s} \sim N(0, .1)$. The values of the options V represent the sum of the option's attributes weighted with the attributes' weights w. If there is no previous bias nor any influence of previous choices, this evidence accumulation starts at 0 . For additionally indicating a systematic bias in a choice task $\beta$ is introduced. The parameter describes any initial response bias at the starting point $z_{0}$. In a similar vein $z_{0}$ can incorporate the influence of previous choices. We assume that the starting point in a given trial $\mathrm{t}, z_{0 t}$ is further influenced by $\rho=[-1,1]$ and a compensating-consistency parameter $\kappa$. The function $\rho$ incorporates the evidence for previous choices and the memory for it. It is determined as the inner product of a decay vector and the previous evidence. The decay vector determines that more recent trials have a greater impact. In line with previous research on memory, the memory further decreases as an exponential function. The impact of previous trials increases in those trials, in which the evidence was sparse.

$$
\begin{equation*}
\rho_{t}=\frac{1}{\left(V_{A t}-V_{B t}\right)} \cdot e^{- \text {decay }_{t}} \tag{2}
\end{equation*}
$$

As has been outlined in the introduction different hypotheses on the direction of the effect in sequential choices have been formulated. We incorporated these hypothesis in the compensation-consistency parameter $\kappa=[-1,1]^{1}$ which is multiplied with $\rho$ for determining the direction of the sequential effect. If $\kappa>0$, previous choices are compensated, if $\kappa<0$, individuals make consistent choices and if $\kappa=0$, previous choices do not influence the present choice. Following the starting point $z_{0}$ in trial t is defined as:

$$
\begin{equation*}
z_{0 t}=\beta-\kappa \times \rho_{t} \tag{3}
\end{equation*}
$$

Finally, a probit link is applied for predicting the choice probability in any trial $t$ based on the accumulated evidence from the starting point $z_{0 t}$ up to $z_{S t}$ in that trial.

$$
\begin{equation*}
P(A)_{t}=\phi\left(z_{S_{t}}, 0,1\right) \tag{4}
\end{equation*}
$$

Additionally we introduced a trembled hand error $\xi$ which indicates the probability to guess between the options.

$$
\begin{equation*}
P(A)_{t}=(1-\xi) \times \phi\left(z S_{t}, 0,1\right)+\xi \times .5 \tag{5}
\end{equation*}
$$

## Results



Figure 2: Probability to choose the minority candidate, as a function of the dominating group, and the position in the trial sequence (Trial Number). The choice probabilities of the dominated trials are averaged over the trials. The error bar represent $+/-1.96 \times S E$. The dashed line indicates guessing probability .5.

The dominating candidate was chosen on average in $94 \%$ of the dominated trials, indicating that the participants followed the instructions of the task very well. Figure 2 illustrate the choice probabilities of the candidates from the discriminated group. The data were collapsed over the categories skin color and gender, for reducing redundancy. As expected, we observed sequential effects in the task, because the probability to choose a female or black candidate in the ambiguous

[^370]trials, (Trial Number 9-10), differed between the conditions. The probability was decreased, compared to the other conditions, if the respective discriminated group had dominated the previous trials. The choice probabilities in the control condition serve as a reference for choice probabilities without sequential effects.

For explaining the observed sequential effects, the DSSM model was fitted to the data. The choice probabilities in the control conditions were used to inform the parameters of the DSSM model. The response bias parameter $\beta$ was fixed so that $\Phi(\beta, 0,1)$ corresponds to the averaged probability to choose a candidate from the discriminated group in the control conditions $(\mathrm{P}($ female $/$ black $)=.58)$. Likewise the decision weights were adapted for capturing the higher weight on the first attribute, by adapting the attributes weights w to the ratio of the choice probabilities, in accordance with the first attribute $\mathrm{P}($ female $/$ black $)=.78$, or not $\mathrm{P}($ female $/$ black $)=.38$, $\mathrm{w}=(2,1,0)$. The final model was fitted to the complete data set of the experimental conditions via grid search and minimizing logloss $\left(\log \operatorname{Loss}=-\frac{1}{n} \Sigma \log (\right.$ Likelihood $)$ ).

The best fitting parameters for the experimental conditions and the corresponding logLoss are illustrated in Table 2. The logLoss for a complete guessing model would be logLoss $=$ .69. In order to further validate the model we compared it via BIC to a complete Guessing and a Sequential Sampling Model SSM without dependencies between the trials. Across all participants, the DSSM provided a better fit than the guessing model and the SSM, $B I C_{D S S M}=1990, B I C_{S S M}=3423$, and $B I C_{\text {guess }}=6670$. The best fitting parameters indicate

Table 2: Best fitting parameters of $\kappa$ as a function of the conditions.

| Cat | Gender |  | Skin Color |  |
| :--- | :---: | :---: | :---: | :---: |
| Dom. <br> group | discr | non-discr | discr | non-discr |
| $\kappa$ | .08 | .48 | .12 | .56 |
| logLoss .25 | .31 | .29 | .29 |  |

Note $. \operatorname{logLoss}=$ mean negative $\log$ Likelihood
compensating choices in all conditions, however the compensation is higher in the conditions in which the candidates from the non-discriminating group dominated the first eight trials. In order to test the difference of $\kappa$ between the conditions, the best fitting parameters for $\kappa$ were additionally estimated on the individual level. This comparison revealed significant differences between the conditions in the gender condition $T(193.32)=-7.73, p<.001, \log (B F)^{2}>10$ and the skin color condition $T(164.09)=-8.99, p<.001, \log (B F)>$ 10. This indicates stronger compensation of morally dubious choices, than compensation of morally virtuous choices. There were large individual differences in the parameters, $\kappa$ $S D_{\text {discr }}=.69$ and $S D_{\text {nondiscr }}=.50$, and a considerable number of participants $\sim 30 \%$ applied extremely consistent choice

[^371]behavior in the condition in which candidates from the discriminated group dominated the first 8 trials.

To conclude, we observed sequential effects in the sequential employee selection task. In general, participants tended to compensate for previous choices, especially if these choices were morally dubious. If the previous choices were of high moral value, choosing the candidates from the group discriminated against, a large number of participants showed consistent choice behavior.

## Discussion

Present choices between job candidates are influenced by previous, unrelated, choices. Choosing a job candidate consistently from one group over the complementary group, defined by skin color or gender, decreased the probability to choose a candidate from the same group in subsequent trials. However, the sequential effects are not symmetric, because not all choices are equally compensated for. Thus, instead of balancing groups over a series of choices our data support the assumption of a stronger compensation for morally dubious choices, sometimes referred to as the moral cleansing hypothesis (Tetlock et al., 2000).

The mere existence of moral cleansing and moral licensing effects has been questioned recently via series of failed replications (Earp, Everett, Madva, \& Hamlin, 2014; Blanken, Van De Ven, Zeelenberg, \& Meijers, 2014). However, the experimental studies investigating these effects rarely observed multiple similar choices in a row. Those studies rather used different tasks, investigated behavioral intentions or used other experimental methods aiming at inducing a specific mindsets, in order to influence subsequent single choices. Furthermore, no computational model has been implemented for analyzing the data. We make two major contribution to this debate. First, we present an experimental paradigm for observing sequential effects in the employee selection and beyond. The task can easily be framed differently in order to test sequential effects in other contexts as well. Second, we formalized a computational model, the Dependent Sequential Sampling Model, for describing and explaining sequential choice effects and corresponding individual differences. Especially the individual estimates of $\kappa$ via the DSSM indicate large individual differences with regard to the compensation or confirmation of previous choices. While $\kappa$ on the group level indicated compensation also in the conditions which were dominated by candidates from the non-discriminating group, investigating individual estimates identified large individuals differences. Importantly, a meaningful number of participants showed actual consistent choice behavior (Zhang et al., 2014; Gneezy et al., 2011).

The DSSM relates to the well established application of accumulator models in choice tasks (Busemeyer \& Townsend, 1993; Lee, 2004; Newell \& Lee, 2011; Ratcliff, 1978). For the purpose at hand the model was used in a very simple version (Milosavljevic et al., 2010). A richer dataset, with within-subject manipulation of conditions and a larger num-
ber of sequential choices would further allow to increase the complexity of the model by estimating more parameters on the individual level, for example the individual initial decision bias. Nonetheless, the current simplicity is perfectly suited for the present research questions.

We provide strong evidence for compensation of morally choices in employee selection and the presented experimental and methodological approach further allows replicating our findings in other areas as well.

## References

Ayton, P., \& Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: two faces of subjective randomness? Memory $\{\&\}$ Cognition, 32(8), 1369-1378. Doi: 10.3758/BF03206327

Bertrand, M., \& Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American Economic Review, 94(4), 991-1013. Doi: 10.1257/0002828042002561

Blanken, I., Van De Ven, N., Zeelenberg, M., \& Meijers, M. H. C. (2014). Three attempts to replicate the moral licensing effect. Social Psychology, 45(3), 232-238. Doi: 10.1027/1864-9335/a000189

Busemeyer, J. R., \& Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review, 100(3), 432-459.
Chen, D., Moskowitz, T., \& Shue, K. (2016, feb). DecisionMaking under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires (Tech. Rep.). Cambridge, MA: National Bureau of Economic Research. Doi: 10.3386/w22026
Clotfelter, C. T., \& Cook, P. J. (1993). Notes: The "Gambler's Fallacy" in Lottery Play. Management Science, 39(12), 1521-1525. Doi: 10.1287/mnsc.39.12.1521
Conway, P., \& Peetz, J. (2012, jul). When Does Feeling Moral Actually Make You a Better Person? Conceptual Abstraction Moderates Whether Past Moral Deeds Motivate Consistency or Compensatory Behavior. Personality and Social Psychology Bulletin, 38(7), 907-919. Doi: 10.1177/0146167212442394

Dhar, R., Huber, J., \& Khan, U. (2007). The shopping momentum effect. Journal of Marketing Research, 44(203), $1-36$.
Dhar, R., \& Simonson, I. (1999, feb). Making Complementary Choices in Consumption Episodes: Highlighting versus Balancing. Journal of Marketing Research, 36(1), 29. Doi: 10.2307/3151913
Earp, B. D., Everett, J. A. C., Madva, E. N., \& Hamlin, J. K. (2014). Out, Damned Spot: Can the "Macbeth Effect" Be Replicated? Basic and Applied Social Psychology, 36(1), 91-98. Doi: 10.1080/01973533.2013.856792
Gigerenzer, G., \& Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded ratio-
nality. Psychological Review, 103(4), 650-69. Doi: 10.1037/0033-295X.103.4.650

Gluck, M. a., Bower, G. H., Donegan, N., Estes, W. K., Kahneman, D., Mcclelland, J., ... Schechter, N. (1988). From conditioning to category learning: an adaptive network model. Journal of experimental psychology. General, 117(3), 227-47. Doi: 10.1037/0096-3445.117.3.227
Gneezy, A., Imas, A., Nelson, L., Brown, A., \& Norton, M. (2011). Paying to Be Nice: Consistency and Costly Prosocial Behavior. Advances in Consumer Research, 39, 468470.

Gregory, R. F. (2003). Women and Workplace Discrimination: Overcoming Barriers to Gender Equality. Rutgers University Press.
Huber, J., Goldsmith, K., \& Mogilner, C. (2008, dec). Reinforcement versus balance response in sequential choice. Marketing Letters, 19(3-4), 229-239. Doi: 10.1007/s11002-008-9042-5

Jordan, J., Mullen, E., \& Murnighan, J. K. (2011). Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior. Personality and Social Psychology Bulletin, 37(5), 701-713. Doi: 10.1177/0146167211400208

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. (Vol. 99) (No. 1). Doi: 10.1037/0033-295X.99.1.22
Lee, K. (2004, dec). Age, neuropsychological, and social cognitive measures as predictors of individual differences in susceptibility to the misinformation effect. Applied Cognitive Psychology, 18(8), 997-1019. Doi: 10.1002/acp. 1075

Ma, D. S., Correll, J., \& Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. Behavior research methods, 47, 1122-1135. Doi: 10.3758/s13428-014-0532-5

Milosavljevic, M., Malmaud, J., \& Huth, A. (2010). The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure. Judgement and Decision Making, 5(6), 437-449. Doi: 10.2139/ssrn. 1901533
Monin, B., \& Miller, D. T. (2001). Moral credentials and the expression of prejudice. Journal of Personality and Social Psychology, 81(1), 33-43. Doi: 10.1037/0022-3514.81.1.33

Newell, B. R., \& Lee, M. D. (2011, dec). The right tool for the job? Comparing an evidence accumulation and a naive strategy selection model of decision making. Journal of Behavioral Decision Making, 24(5), 456-481. Doi: 10.1002/bdm. 703

Novemsky, N., \& Dhar, R. (2005, dec). Goal Fulfillment and Goal Targets in Sequential Choice. Journal of Consumer Research, 32(3), 396-404. Doi: 10.1086/497551
Pager, D., Western, B., \& Bonikowski, B. (2009). Discrimination in a Low-Wage Labor Market: A Field Experiment. American Sociological Review, 74(5), 777-799.

Doi: 10.1177/000312240907400505.Discrimination
Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59-108. Doi: 10.1037/0033-295X.85.2.59

Rieskamp, J., \& Otto, P. E. (2006). SSL: A Theory of How People Learn to Select Strategies. Heuristics: The Foundations of Adaptive Behavior, 135(2), 207-236. Doi: 10.1093/acprof:oso/9780199744282.003.0011

Sachdeva, S., Iliev, R., \& Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation: Research Article. Psychological Science, 20(4), 523-528. Doi: 10.1111/j.1467-9280.2009.02326.x
Simão, J., \& Todd, P. M. (2002, apr). Modeling Mate Choice in Monogamous Mating Systems with Courtship. Adaptive Behavior, 10(2), 113-136. Doi: 10.1177/1059712302010002003

Simonsohn, U., \& Gino, F. (2013). Daily Horizons Evidence of Narrow Bracketing in Judgment From 10 Years of M.B.A. Admissions Interviews. Psychological Science, 24(2), 219-224. Doi: 10.1177/0956797612459762
Stewart, N., \& Brown, G. D. A. (2004). Sequence Effects in the Categorization of Tones Varying in Frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(2), 416-430. Doi: 10.1037/0278-7393.30.2.416
Stewart, N., Brown, G. D. A., \& Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(1), 3-11. Doi: 10.1037/0278-7393.28.1.3
Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., \& Lerner, J. S. (2000). The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. Journal of Personality and Social Psychology, 78(5), 853-870. Doi: 10.1037/0022-3514.78.5.853
Thaler, R. H., \& Johnson, E. J. (1990). Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. Management Science, 36(6), 643-660. Doi: $10.1287 / \mathrm{mnsc} .36 .6 .643$
Todd, P. M. (2007). How much information do we need? European Journal of Operational Research, 177(3), 13171332. Doi: 10.1016/j.ejor.2005.04.005

Zhang, S., Cornwell, J. F. M., \& Higgins, E. T. (2014, jan). Repeating the Past: Prevention Focus Motivates Repetition, Even for Unethical Decisions. Psychological Science, 25(1), 179-187. Doi: 10.1177/0956797613502363

# The Influence of Prosody and Case Marking on Thematic Role Assignment in Ambiguous Action Scenes: Adults versus Children 

Julia Marina Kröger (jkroeger@cit-ec.uni-bielefeld.de)<br>Bielefeld University, Department of Linguistics, Germany<br>Katja Münster (katja.muenster@hu-berlin.de)<br>Pia Knoeferle (pia.knoeferle@hu-berlin.de)<br>Humboldt University Berlin, Department of German Language and Linguistics, Germany


#### Abstract

In two visual word eye tracking studies, we investigated the influence of prosody and case marking on children's and adults' thematic role assignment. We assigned an SVO/ OVS-biasing (vs. neutral) prosodic contour to unambiguously case marked German subject-verb-object (SVO) and object-verb-subject (OVS) sentences respectively. Scenes depicted ambiguous action events (e.g., donkey-paints->elephant-paints->cheetah) but case marking and prosody could, in principle, disambiguate. In adults, case marking but not prosody rapidly guided thematic role assignment. Children did not rely on case marking but exploited the biasing prosody to enhance their agent-first interpretation of the sentences. These results suggest that in scenes depicting fully ambiguous role relations, children's understanding of case marking at the age of five is not yet robust enough to enable thematic role assignment. Prosody did not overwrite the SVO preference, it rather enhanced it.


Keywords: Visual World Paradigm, eye movements; prosody; action scenes; age differences; language processing; thematic role assignment; case marking.

## Introduction

In rich contexts, listeners can rapidly exploit a range of different cues (e.g., prosody, non-linguistic information, case marking) during language comprehension. Such cues can, for instance, help them to efficiently identify the thematic roles of a sentence. Thematic roles distinguish the role an argument carries with regards to the predicate of the sentence (Carnie, 2002). Although children rapidly acquire language during the first years of their life, a full command of their native language takes, not surprisingly, time to develop. One aspect that is particularly challenging is the correct assignment of thematic roles. In German, for example, a transitive sentence typically includes a subject, a verb, and an object but speakers can arrange these constituents in more than one way, yielding, among others, subject-verb-object (SVO) or object-verbsubject (OVS) orders. To determine who is doing what to whom, comprehenders can rely on the case-marked determiners of the sentential noun phrases. Accusative case marks a sentence-initial noun phrase as the object (patient), and nominative case marks it as the subject (agent) of the sentence. But case marking can be
ambiguous in German (feminine and neuter nouns have the same form in the nominative / subject and the accusative / object case), resulting in temporary ambiguity as to who-does-what-to-whom. The resolution of this ambiguity can be difficult for children, not only because of their strong SVO word order bias. The following section describes a range of studies on how adults rapidly exploit visual, case, and prosodic cues for thematic role assignment while children sometimes struggle to exploit these cues.

## Visual Information

In rich contexts, non-linguistic information like the visual referential context, contrast between objects, depicted actions, or events can all rapidly influence the interpretation, syntactic structuring, and thematic role assignment of spoken utterances (e.g., Chambers, Magnuson, \& Tanenhaus, 2004; Knoeferle, Crocker, Sheepers \& Pickering, 2005; Sedivy et al., 1999; Tanenhaus et al., 1995). Tanenhaus and colleagues (1995) found that adults use referential context to disambiguate sentences (e.g., Put the apple on the towel in the box). The visual display contained either one or two referents (one referent: an apple on a towel, two referents: one apple on a towel and another apple on a napkin). Participants' gaze pattern suggested structural disambiguation in that they interpreted the towel as a destination (VP-analysis) for the apple in the one-referent context and as the location of the apple (NP modifier analysis) in the two-referent context (see Sedivy et al., 1999 for related effects of contrastive adjectives in establishing reference to objects).

Event relations depicted in the visual context can help listeners with anticipating thematic role relations. In Knoeferle et al. (2005), participants listened to locally ambiguous German SVO and OVS sentences (transl: 'the princess (agent/patient) washes/paints apparently the pirate (patient)/ the fencer (agent)'). The sentences did not provide information about who is the agent or the patient prior to disambiguation by case marking on the determiner of the second noun phrase. Scenes depicted the princess in both an agent and a patient role (the fencer was acting upon the princess while the princess was acting upon a pirate). During the verb ('washes / paints'), participants successfully anticipated the patient or the
agent role filler respectively in SVO and OVS sentences. Thus, the depicted event information resolved the ambiguity in the linguistic input.

Children, much like the adults, can rapidly exploit depicted actions to correctly identify thematic roles in German SVO and OVS sentences. Shortly after the verb had identified the action (and its associated role relations), the children anticipated the patient (vs. agent) in the scene for SVO sentences (Zhang \& Knoeferle, 2012) and the depicted agent (vs. patient) for OVS sentences (Münster, 2016; Zhang \& Knoeferle, 2012). However, children, unlike adults, did not exploit a referential context for disambiguating a VP-/NP-attachment ambiguity. When hearing Put the frog on the napkin... they interpreted the napkin as the frog's destination in both the one- and the two-referent context even though the latter biased towards a location interpretation (Trueswell et al., 1999). However, in the absence of real-time measures these results do not provide insight into children's online sentence processing.

## Case Marking

When it comes to linguistic information, case marking is believed to be a very strong cue for thematic role assignment with adult participants (e.g., Matzke, Mai, Nager, Rösseler \& Münte, 2002). However, studies in five-year old children report conflicting results. Dittmar et al., (2008), found that five-year-olds struggled to exploit case marking for thematic role assignment. The results of an act-out task revealed that children relied on (SVO) word order instead of case marking for interpreting ambiguous and unambiguous German SVO and OVS sentences. They interpreted the first noun phrase as the subject/agent even if it was case-marked as the object/patient of the sentence (agent-first/SVO bias). However, more recent evidence suggests that children at the age of four to five can rely on case marking for correct thematic role assignment in unambiguous German SVO and OVS sentences (e.g., Özge et al., 2015).

In Özge et al., (2015), information from the visual display likely supported the interpretation of the linguistic input. The scenes were created on the basis of world knowledge about who is the most likely agent and the most likely patient (typically the fox eats the hare and the hare eats the cabbage). The scenes thus showed animals (a hare, a fox, and a cabbage) between which stereotypical role relations exist, but they did not disambiguate who-does-what-to-whom (case marking did).

Even clearer effects of the visual context emerged in Zhang et al. (2012) and Münster (2016). Here, depicted actions (e.g., a bear painting a worm) disambiguated the role relations (only one animal performed the action mentioned in the linguistic input) but scenes did not depict stereotypical role relations; when the actions were absent, unambiguous case marking alone (i.e., in the absence of stereotypical role relations between bear and worm) was insufficient to disambiguate the role relations. Thus, the visual context seems to matter. By contrast, the effects of case marking alone on children's thematic role
assignment (i.e., when scenes do not support thematic role assignment through either stereotypical world knowledge associated with the characters or disambiguating action depictions) remain unclear.

## Prosody

Supra-sentential information can also be useful for establishing a link between the linguistic input and the visual world. Among others, prosody assigns focus to sentence constituents (e.g., via accentuation). Prosody can moreover rapidly disambiguate syntactic structure. In a visual world eye tracking study, participants rapidly exploited prosody to identify grammatical functions when the scene depicted role fillers such as a cat, a bird, and a dog, for which world knowledge implicated stereotypical thematic relations (e.g., cats chase birds and dogs chase cats, Weber et al, 2006). But the scenes did not otherwise disambiguate the upcoming thematic role relations. Feminine case marking (identical in nominative and accusative case) on the determiner of the first noun phrase created locally structurally ambiguous sentences: Die Katze (amb.) jagt womöglich den Vogel (acc/obj)/der Hund (nom/subj)- 'The cat (amb.) chases possibly the bird (obj/patient)/the dog (subj/agent)'. Biasing prosodic contours prompted listeners to make more anticipatory eye movements towards the agent (vs. patient) in the scene for ambiguous OVS sentences and towards the patient (vs. agent) in the scene for ambiguous SVO sentences. Prosody was the only information available for the correct anticipation of thematic roles prior to the disambiguating case marking on the second noun phrase; but at the same time, world knowledge associated with the scene may have provided a supportive background.

Children also exploited prosody for thematic role assignment (Grünloh et al., 2011). Two short videos showed thematic role relations in two orders, permitting a direct contrast of agent-patient and patient-agent events. However, prosodic effects (of an accentuated first noun phrase in OVS sentences) emerged only when case marking was also present. When case marking was absent (ambiguous OVS sentences) children relied on their (SVO) word order bias instead of prosody and thus interpreted the OVS sentences as agent-first sentences.

## The Present Research

This paper investigates the effects of prosody (biasing towards either the SVO or the OVS order as in Weber et al., 2006 vs. neutral) and case marking (SVO vs. OVS) in ambiguous action scenes. The actions were depicted but did not give away the specific role relations of the sentences (two characters performed identical actions and could both function as possible agents of the verb, Fig. 1, the elephant; the donkey, see Table 1 for sentences). Our scenes did not include stereotypical knowledge about who does what to whom. If a supportive context is necessary for children to exploit case marking, then we should see no effects of case marking on thematic role assignment in the present study (i.e., no anticipation of the agent / the
donkey in OVS but of the patient, the cheetah, in SVO and OVS sentences, Table 1 and Fig. 1, portraying an SVO bias). This prediction is based on the literature that revealed conflicting results regarding children's reliance on case marking (Dittmar et al., 2008; Özge et al., 2015). Alternatively, seeing event relations depicted could be helpful even if the characters perform identical actions (i.e., seeing a donkey as the patient in a painting event might help to integrate object case marking). If so, we should see effects of case marking.

If prosodic effects do not depend on a supportive scene context, then we should see effects of prosody on children's thematic role assignment and anticipation of role fillers in the scene. Alternatively, we may observe the effects of an SVO bias (more looks to the patient / the cheetah during the verb) independent of case marking and prosody in children.

Adults can rapidly use case marking for thematic role assignment (Matzke et al., 2002) and should thus anticipate the correct patient / agent depending on case marking. Prosody should also influence the timing and amount of anticipatory eye movements towards the target role filler during the time course of the sentence.

Participants further responded post-trial to questions about who-does-what-to-whom. Adults should answer these questions correctly in almost all cases since case marking was unambiguous (further effects of prosody may or may not emerge). For children, case-marking may affect the accuracy (reduced for OVS sentences). We also expected more correct responses for sentences assigned a biasing (vs. neutral) prosodic contour if children benefit from prosody for thematic role assignment.

## Experiments

Participants. 24 five-year old children (age range 4.5 to 5.10 years) and 24 young adult (mean age=25.5) monolingual (i.e., no acquisition of a second language before the age of 6) native speakers of German participated in this study. Children came from different kindergartens in the area of Bielefeld and the experiment was conducted at the kindergartens. Each child received a toy for participation. Young adults were students from Bielefeld University and were paid to participate. Participants had normal or corrected vision and hearing and all gave informed consent. The Bielefeld University ethics committee approved the experiments.
Materials. A linguistically trained female native speaker of German recorded 24 unambiguous transitive German subject-verb-object (SVO) and 24 unambiguous transitive German object-verb-subject (OVS) sentences. She was instructed to use the prosodic structure displayed in Table 1 for each sentence structure respectively. The sentences were all unambiguously case marked on the first noun phrase of the sentence. We emulated the prosodic contours reported by Weber et al. (2006) and these were either present or sentence intonation was even. In SVO sentences the main stress was on the verb ( $\mathrm{L}^{*}+\mathrm{H}$ accent on NP1, $\mathrm{H}^{*}$ accent on the verb) and in OVS sentences on the first noun phrase ( $\mathrm{L}+\mathrm{H}^{*}$ accent on NP1; Table 1).

Table 1. Overview of experimental conditions (ag=agent, pat=patient, subj=subject, obj=object).

| Sentence <br> Structure | Prosodic <br> Structure | Example <br> SVO <br> L*+H (NP1), <br> $\mathrm{H}^{*}$ (verb) |
| :--- | :--- | :--- |
| neutral | Der Elefant (subj/ag) <br> zeichnet sogleich den <br> Gepard (obj/pat). <br> The elephant (subj/ag) <br> draws immediately the <br> cheetah obj/pat). |  |
| OVS | Der Elefant (subj/ag) <br> zeichnet sogleich den <br> Gepard (obj/pat). <br> The elephant (subj/ag) <br> draws immediately the <br> cheetah obj/pat). |  |
| OVS | neutral | Den Elefanten (obj/pat) <br> zeichnet sogleich der <br> Esel (subj/ag). <br> The elephant (obj/pat) <br> draws immediately the <br> cheetah (subj/ag). |
| L+H*(NP1) | Den Elefanten (obj/pat) <br> zeichnet sogleich der <br> Esel (subj/ag). <br> The elephant (obj/pat) <br> draws immediately the <br> cheetah (subj/ag). |  |

For each of the 24 SVO and OVS sentences we created scenes depicting three clipart animal characters (Fig.1). The direction in which these characters were looking was the same for all three of them, either left or right. The middle character and one of the adjacent characters were depicted as performing the same action (e.g., for the verb 'draw', zeichnen, the two characters were depicted as holding a pencil on a canvas). The third character did not perform an action. Thus, the actions provided a context but did not permit comprehenders to unambiguously identify the correct thematic role relations upon hearing the verb.

The middle character was always role ambiguous because it could be the agent or the patient of the scene. We dub the other character depicted as performing an action the 'true agent' of the scene and the character not performing an action the 'true patient' of the scene. Each animal character filled both roles (that of a true agent and a true patient). The cheetah, for instance, is the true patient in the image shown in Figure 1. In a counterbalancing scene, the cheetah is the true agent. The elephant is depicted as drawing in the example scene (true agent); in a counterbalancing scene, the elephant is the true patient. Across all lists, each scene occurred once in each condition with all characters facing right and once
with all characters facing left. We added 8 fillers to the experiments. The number of filler was kept constant for the two experiments to maximize similarity in the materials across age groups. Post experiment questionnaires revealed no recognition of the experimental goal. After each trial participants were asked a comprehension question which was either presented in active or in passive voice (e.g., Who paints here? Or who is being painted here?). Prior to the actual experiment three practice trials familiarised the participants with the scenes and the task. The design of these items was identical to the experimental items. The sentences were structured either in SVO or OVS order and prosody was either biasing or neutral. All scenes were pre-tested with 20 five-year old children and the results confirmed that all depicted characters were correctly identified. Out of the 24 different depicted actions, only one was not correctly identified and was exchanged.

true agent ambiguous true patient
Figure 1. Example picture of a scene.
Procedure. Participants' eye movements were monitored with an Eyelink 1000 eye tracker with a sampling rate of 500 Hz Monocular, and an average accuracy of $0.5^{\circ}$ in the remote setup. Images were presented on a Dell laptop with a screen resolution of $1920 \times 1080$. Before starting the experiment, the experimenter manually calibrated the eye tracker using a five-dot calibration scheme. For each trial, the scene was presented for 2000 ms , followed by the auditory sentence. 1500 ms after sentence offset the question followed (Fig. 2). A drift correct point separated the trials to ensure calibration of the eye tracker and the same starting point for each trial. Participants first saw the practice trials. Next, the experimenter re-calibrated before starting the experiment. Each testing session lasted approximately 20 minutes.
Analysis. We defined two word regions for the analysis: verb and adverb (beginning of verb onset to adverb onset for the verb and beginning of adverb onset to NP2 onset for the adverb). These two regions were defined on the basis of the prosodic structure of the sentences. We focused on the verb region because this is where the prosodic structures can be distinguished. Whenever there is a main stress on the first noun phrase, the verb experiences a fall in stress. Otherwise the main stress is on the verb. We were further interested in the adverb region to examine post-verbal eye movements.

Visual input: 2000ms


Figure 2. Example for the time course of an experimental trial.

In the scenes, we defined two role fillers (the donkey, the true agent; the cheetah, the true patient) as areas of interest. The middle role filler was always mentioned at the beginning of each sentence and was thus not used for the analysis of anticipatory eye movements. We computed mean log-ratios of looks towards the agent and the patient of the scene (see Arai, van Gompel \& Sheepers, 2007; Carminati \& Knoeferle, 2013). Log-ratios are a relative measure that represents the looks towards one character over the other. On the basis of these mean log-ratios we conducted an Analysis of Variance (ANOVA) following a 2 (word order) x 2 (prosody) design by subject and by item for all word regions of the sentence (NP1, verb, adverb, NP2). All positive numbers in the log-ratios represent a preference of looking at the agent (vs. patient) in the scene and all negative numbers a preference towards the patient (vs. agent) in the scene. The postsentence questions show the number of correctly answered questions. We calculated percentages of correct answers of all possible answers and analysed the accuracy data using generalised linear mixed effects models (Bates, Mächler, \&Walker, 2015).
Eye movement results. The data for both age groups showed no significant effects of prosody in the verb and adverb regions (Figs 3 and 4). For the adults (only), the analyses revealed a main effect of word order for the verb and adverb (Fig. 3: adverb). The adults were more likely to inspect the patient (vs. agent) in SVO sentences (negative numbers) and the agent (vs. patient) in OVS sentences (positive numbers) during the adverb region (word order effect: $p<.001$ ).

The child data showed more looks towards the patient (vs. agent) in all four conditions (intercept $p<.05$ ). The preference to inspect the patient over the agent is only slightly higher in the biasing compared to the neutral prosody conditions (Fig. 4).
Accuracy Results. Adult's post-sentence answers revealed a high percentage of correct answers (99\%), with no difference between the conditions (Fig. 5). The child data revealed an overall accuracy of $71 \%$. The analyses revealed no clear difference between the two prosodic
conditions but a main effect of word order ( $p<.001$, Fig. $6)$.


Figure 3. Mean log-ratio of looks of agent over patient during the adverb region per condition in adults (Analysis by subjects).


Figure 4. Mean log-ratio of looks (agent over patient) during the adverb region per condition in children (Analysis by subjects).

## Adults: Correct Post-Sentence Answers

Figure 5. Accuracy results: percentage of correct postsentence answers per condition in adults.

## Children: Correct Post-Sentence Answers



Figure 6. Accuracy results: percentage of correct postsentence answers per condition in children.

## Discussion

We investigated children's and adults' thematic role assignment in unambiguously case marked German SVO and OVS sentences using scenes in which two role fillers performed identical actions. Thus, as participants heard the verb, it was unclear which of two events (one depicting the NP1 referent as the agent, the other depicting him as a patient) to rely on for anticipating upcoming role fillers. In brief, the scene did not disambiguate the thematic roles relations. But case marking and prosodic cues could, in principle, permit anticipatory thematic role assignment. We recorded participants' eye movements to the agent and patient in the scene while they inspected scenes and listened to related German SVO and OVS sentences (Table 1).
Previous research has reported effects of prosody on thematic role assignment in German children and adults (Grünloh et al., 2011; Weber at al., 2006). Unlike these previous findings, our results revealed no significant effect of prosody. Previous research further revealed that adults rapidly use case marking for thematic role assignment (Matzke et al., 2002). Results for children were, however, contradictory (Dittmar et al., 2008; Münster, 2016; Özge et al., 2016; Zhang \& Knoeferle, 2012). In line with previous findings, the adults in our study exploited case marking for real-time thematic role assignment. They directed more anticipatory looks towards the true patient (vs. agent) in SVO sentences and towards the true agent (vs. patient) in OVS sentences. Children, by contrast, did not exploit case marking for such visual anticipation. In both SVO and OVS sentences, they directed more anticipatory looks towards the true patient during the adverb. They thus seem to have interpreted OVS sentences as agent-first sentences, disregarding the unambiguous object case marking on the sentence-initial noun phrase.
In adults, the null effect of prosody might be explained by the fact that morpho-syntactic (case) information has stronger links to thematic role assignment than suprasentential information. Similar arguments have been made for prosodic marking and object color contrast (Sedivy et al., 1999). Sedivy et al. argued that color contrast effects enable a strongly contrastive interpretation already, eliminating further contrastive intonation effects. A similar argument might hold for case marking and prosody. Case marking in adults fully disambiguated thematic role assignment and prosody had no additional beneficial effects. Relatedly, adults in Grünloh et al. (2011, Exp. 1) failed to exploit prosody but used case marking for thematic role assignment.
Concerning the children, our prosody results differed from those by Grünloh and colleagues (2011) who reported that a contrastive OVS-biasing intonation in unambiguous sentences facilitated children's identification of patient-agent events. One reason for this might be that we used different scenes. In our study, participants saw one ambiguous scene that included both possible interpretations (e.g., true patient <-action ambiguous <-action true agent). In Grünloh and
colleagues (2011), by contrast, children selected one event picture from two adjacent ones (in which agent and patient roles appeared in reversed order, agent-patient vs. patient-agent). This direct contrast may have facilitated identifying the correct event by means of prosody (and case marking). In addition, in our scenes participants including children - inspected the NP1 referent when it was named. The verb then linked to a matching action of that referent and likely reaffirmed this was the correct agent. It is possible that the verb-action match and additional posture of the NP1 referent (facing the true patient), may have led children in particular to interpret an $\mathrm{L}+\mathrm{H}^{*}$ accent on NP1 as highlighting the NP1 as the agent rather than as the patient, even if case-marking indicated it was the object and patient.

In previous studies on children's use of case marking (Münster 2016; Özge et al., 2015; Zhang \& Knoeferle 2012), visual information likely provided a supportive background for exploiting case marking. Unlike prior research, our scenes did not constrain thematic role relations by means of stereotypical knowledge (Özge et al., 2016) or action depictions that - once the verb became available - permitted children to distinguish SVO and OVS sentences (Münster, 2016; Zhang \& Knoeferle, 2012). One of the reasons for why children in our study did not use case marking in real time might thus be the lack of supportive contextual information.

Children's comprehension mechanisms develop throughout the first years of their life. They learn from their immediate environment and from observing who interacts with whom. Perhaps they need an unambiguous visual background to exploit case for syntactic structuring. Without that, they fail to correctly interpret more demanding OVS sentences and fall back on default structures (e.g., SVO). Further research could examine when children start to abstract away from the visual display and begin to use case marking in an adult-like manner for syntactic structuring.

## Acknowledgements

This research was funded by the Cognitive Interaction Technology Excellence Center (277, DFG) and by the Project "FoTeRo" in the Focus center XPrag (DFG).

## References

Altmann, G. T. M., \& Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. Cognitive Science, 33(4), 583-609.
Arai, M., Van Gompel, R. P. G., \& Scheepers, C. (2007). Priming ditransitive structures in comprehension. Cognitive Psychology, 54, 218-250.
Bates, D., Mächler, M., Bolker, B. \& Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistics Software, 67 (1), 1-48.
Carminati, M. N., \& Knoeferle, P. (2013). Effects of speaker emotional facial expression and listener age on incremental sentence processing. PloS one, 8(9), e72559.

Carnie, A. (2002). Syntax: A Generative Introduction. Malden, MA: Blackwell Publishers.
Chambers, C. G., Tanenhaus, M. K. \& Magnuson, J. S. (2004). Actions and Affordances in Syntactic Ambiguity Resolution. JEP:LMC, 30(3), 687-696.
Dittmar, M., Abbot-Smith, K., Lieven, E., \& Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. Child Development, 79(4), 1152-1167.
Grünloh, T., Lieven, E., \& Tomasello, M. (2011). German children use prosody to identify participant roles in transitive sentences. Cognitive Linguistics, 22, 393-419.
Kamide, Y., Altmann, G. T. M., \& Haywood, S. L. (2003b). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. $J M L, 49(1), 133-156$.
Knoeferle, P., Crocker, M. W., Scheepers, C., \& Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment. Cognition, 95, 95-127.
Matzke, M., Mai, H., Nager, W., Rösseler, J., \& Münte, T. (2002). The costs of freedom: an ERP-study of noncanonical sentences. Clinical Neuropsychology, 113, 844-852.
Meroni, L. \& Crain, S. (2011). How Children Avoid Kindergarden-paths. In Edward Gibson and Neal Pearlmutter (Eds.), The processing and acquisition of reference. Cambridge, MA: MIT Press.
Münster, K. (2016). Effects of Emotional Facial Expressions and Depicted Actions on Situated Language Processing across the Lifespan. Doctoral Thesis. Bielefeld University.
Özge, D., Kornfilt J., Münster, K., Knoeferle, P., Küntay, A., \& Snedeker, J. (2016). Predictive Use of German Case Markers in German Children. Proceedings of the 40th Annual Boston University Conference on Language Development (pp. 291-303). Somerville, MA.
Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., \& Carlson, G.N. (1999). Achieving incremental semantic interpretation through contextual representation. Cognition, 71, 109-148.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632-1634.
Trueswell, J., Sekerina, I., Hill, N. \& Logrip, M. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. Cognition, 73, 89-134.
Weber, A., Grice, M., \& Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities. Cognition, 99, B63-B72.
Zhang, L., \& Knoeferle, P. (2012). Visual Context Effects on Thematic Role Assignment in Children versus Adults. In Proceedings of the 34th Annual Meeting of the Cognitive Science Society (pp. 2593-2598). Boston, USA

# Enhancing metacognitive reinforcement learning using reward structures and feedback 

Paul M. Krueger ${ }^{1}$ (pmk@berkeley.edu)<br>Falk Lieder ${ }^{1}$ (falk.lieder@berkeley.edu)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Department of Psychology, University of California Berkeley, Berkeley, CA 94720 USA<br>${ }^{1}$ These authors contributed equally.


#### Abstract

How do we learn to think better, and what can we do to promote such metacognitive learning? Here, we propose that cognitive growth proceeds through metacognitive reinforcement learning. We apply this theory to model how people learn how far to plan ahead and test its predictions about the speed of metacognitive learning in two experiments. In the first experiment, we find that our model can discern a reward structure that promotes metacognitive reinforcement learning from one that hinders it. In the second experiment, we show that our model can be used to design a feedback mechanism that enhances metacognitive reinforcement learning in an environment that hinders learning. Our results suggest that modeling metacognitive learning is a promising step towards promoting cognitive growth.


Keywords: Decision-Making; Planning; Metacognitive Reinforcement Learning; Cognitive Training

## Introduction

One of the most remarkable aspects of the human mind is its ability to improve itself based on experience. Such learning occurs in a range of domains, from simple stimulus-response mappings, motor skills, and perceptual abilities, to problem solving, cognitive control, and learning itself (C. S. Green \& Bavelier, 2008; Bavelier, Green, Pouget, \& Schrater, 2012). Demonstrations of cognitive and brain plasticity have inspired cognitive training programs. The success of cognitive training has been mixed and the underlying learning mechanisms are not well understood (Owen et al., 2010; Anguera et al., 2013; Morrison \& Chein, 2011). Feedback is an important component of many effective cognitive training programs, but it remains unclear what makes some feedback structures more effective than others, and there is no principled method for designing optimal feedback structures.

To address these problems, we model cognitive plasticity as metacognitive reinforcement learning. This perspective allows us to translate methods for accelerating reinforcement learning in robots (Ng, Harada, \& Russell, 1999) into feedback structures for cognitive training in humans.

Here, we evaluate this approach in the domain of planning. As a first step, we developed a metacognitive reinforcement learning model of how people learn how many steps to plan ahead in sequential decision problems, and we test its predictions empirically. The results of our first experiment suggest that our model can discern which reward structures are more conducive to metacognitive learning. In our second experiment, we find that feedback structures designed based on our model can accelerate learning to plan.

We start by introducing the theory of reinforcement learning that our approach is based upon. The following two sections apply this theory to model the problem of deciding how to decide and the process by which people learn to do so. We then use this theory to motivate a novel computational method for designing feedback structures that promote cognitive plasticity and experimentally test the predictions of our theory. We close with a discussion of the implications of our results for cognitive training.

## Planning and reinforcement learning

A sequential decision problem can be modeled as a Markov decision process (MDP)

$$
\begin{equation*}
M=\left(S, \mathcal{A}, T, \gamma, r, P_{0}\right), \tag{1}
\end{equation*}
$$

where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $T\left(s, a, s^{\prime}\right)$ is the probability that the agent will transition from state $s$ to state $s^{\prime}$ if it takes action $a, 0 \leq \gamma \leq 1$ is the discount factor on future rewards, $r\left(s, a, s^{\prime}\right)$ is the reward generated by this transition, and $P_{0}$ is the probability distribution of the initial state $S_{0}$ (Sutton \& Barto, 1998). A policy $\pi: \mathcal{S} \mapsto \mathcal{A}$ specifies which action to take in each of the states. The expected sum of discounted rewards that a policy $\pi$ will generate in the MDP $M$ starting from a state $s$ is known as its value function

$$
\begin{equation*}
V_{M}^{\pi}(s)=\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} \cdot r\left(S_{t}, \pi\left(S_{t}\right), S_{t+1}\right)\right] \tag{2}
\end{equation*}
$$

The optimal policy $\pi_{M}^{\star}$ maximizes the expected sum of discounted rewards, that is

$$
\begin{equation*}
\pi_{M}^{\star}=\arg \max _{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{\downarrow} \cdot r\left(S_{t}, \pi\left(S_{t}\right), S_{t+1}\right)\right] . \tag{3}
\end{equation*}
$$

Solving large planning problems is often intractable because the number of possible action sequences grows exponentially with the number of steps one plans ahead. When the state space $S$ is discrete and relatively small, dynamic programming can be used to find optimal plans in polynomial time (Littman, Dean, \& Kaelbling, 1995). But the highdimensional, continuous state spaces people have to plan with in real life are too large for these methods. Instead, people seem to rely on approximate planning strategies (Huys et al., 2015) and often decide primarily based on immediate
and proximal outcomes while neglecting the long-term consequences of their actions (Myerson \& Green, 1995). Despite its fallibility, looking only a few steps ahead can drastically simplify the planning problem, and this may often be a necessity for bounded agents with imperfect knowledge of the environment (Jiang, Kulesza, Singh, \& Lewis, 2015). Since cutting corners in the decision process is both necessary and problematic, good decision-making requires knowing when that is admissible and when it is not. Knowing how much to plan is therefore an important metacognitive skill to learn.

Previous work suggests that this metacognitive skill can be learned through trial and error (Lieder \& Griffiths, 2015). Learning through trial and error can be understood in terms of reinforcement learning (Sutton \& Barto, 1998). While certain reinforcement learning algorithms can, in principle, learn to solve arbitrarily complex problems, reinforcement learning can also be very slow-especially when rewards are sparse and the optimal policy is far from the learner's initial strategy. A common approach to remedy this problem is to give the algorithm pseudo-rewards for actions that do not achieve the goal but lead in the right direction ( Ng et al., 1999).While previous work has developed this idea to accelerate learning a direct mapping from states to actions, we will leverage it to accelerate learning to plan.

## Deciding how to decide

People can use many different decision strategies. This poses the problem of deciding how to decide (Boureau, SokolHessner, \& Daw, 2015). Previous research on meta-decisionmaking has focused on the arbitration between habits versus planning (Keramati, Dezfouli, \& Piray, 2011; Dolan \& Dayan, 2013). While this is an important meta-control problem, it is only one part of the puzzle because people are equipped with more than one goal-directed decisionmechanism. Hence, when the model-based system is in charge, it has to be determined how many steps it should plan ahead. Ideally, the chosen planning horizon should achieve the optimal tradeoff between expected decision quality versus decision time (Vul, Goodman, Griffiths, \& Tenenbaum, 2014) and mental effort (Shenhav et al., 2017).

Here, we make the simplifying assumption that people always choose the action that maximizes their sum of expected rewards over the next $h$ steps, for some value of $h$ that differs across decisions. A planning horizon of $h=1$ entails looking only at the immediate outcome of each action (myopic onestep planning) whereas a planning horizon larger than one entails solving a sequential decision problem to form a multistep plan. Under this assumption, the meta-decision problem is to select a planning horizon $h$ from a set $\mathcal{H}=\{1,2, \cdots$,$\} ,$ execute the plan, select a new planning horizon, and so on. More formally, this problem can be formalized as a metalevel MDP (Hay, Russell, Tolpin, \& Shimony, 2012). In our task, the meta-level MDP is

$$
\begin{equation*}
M_{\text {meta }}=\left(S_{\text {meta }}, \mathcal{H}, T_{\text {meta }}, r_{\text {meta }}\right) \tag{4}
\end{equation*}
$$

where the meta-level state $m \in \mathcal{S}_{\text {meta }}=\{0,1,2,3,4\}$ encodes
the number of remaining moves, and the meta-level action $h \in \mathcal{H}=\{1,2,3,4\}$ is the planning horizon used to make a decision. The meta-level reward function $r_{\text {meta }}$ integrates the cost of planning with the return of the resulting action:

$$
\begin{equation*}
r_{\text {meta }}\left(m_{k}, h_{k}\right)=-\operatorname{cost}\left(h_{k}\right)+\sum_{t=1}^{h} r\left(s_{t}, \operatorname{plan}_{t}^{\left(k, h_{k}\right)}\right) \tag{5}
\end{equation*}
$$

where $\operatorname{plan}_{t}^{(k, h)}$ is the $t^{\text {th }}$ action of the plan formed by looking $h$ steps ahead in the meta-level state $m_{k}$. The meta-decisionmaker receives this reward after the plan has been executed in its entirety. If the meta-decision-maker selects short planning horizons there can be multiple plan-act-reward-learn cycles within a single trial. The cost of planning $\operatorname{cost}\left(h_{k}\right)$ is determined by the branching factor $b$ of the decision tree according to

$$
\begin{equation*}
\operatorname{cost}\left(h_{k}\right)=\lambda \cdot b^{h_{k}} \cdot h_{k}, \tag{6}
\end{equation*}
$$

where $b^{h_{k}}$ is the number of plans, $h_{k}$ is the number of steps per plan, and $\lambda$ is the cost per planning step.*

## Metacognitive reinforcement learning

Solving the problem of deciding how to decide optimally is computationally intractable but the optimal solution can be approximated through learning (Russell \& Wefald, 1991). We propose that people use reinforcement learning (Sutton \& Barto, 1998) to approximate the optimal solution to the meta-decision problem formulated in Equation 4.

## Model

Our model of metacognitive reinforcement learning builds on the semi-gradient SARSA algorithm (Sutton \& Barto, 1998) that was develop to approximately solve MDPs with large or continuous state spaces. Specifically, we assume that people learn a linear approximation to the meta-level Q-function

$$
\begin{equation*}
Q_{\mathrm{meta}}\left(m_{k}, h_{k}\right) \approx \sum_{j=1}^{7} w_{j} \cdot f_{j}\left(m_{k}, h_{k}\right) \tag{7}
\end{equation*}
$$

whose features $\mathbf{f}$ comprise one indicator variable for each possible planning horizon $h\left(f_{1}=\mathbb{1}(h=1), \cdots, f_{4}=\mathbb{1}(h=4)\right)$, one indicator variable for whether or not the agent planned all $l$ steps until the end of the task $\left(f_{5}=\mathbb{1}(h=l)\right)$, the number of steps that were left unplanned $\left(f_{6}=\max \{0, l-h\}\right)$, and the number of steps the agent planned too far ( $f_{7}=$ $\max \{0, h-l\}$ ). The semi-gradient SARSA algorithm learns the weights of these features by gradient descent. To bring it closer to human performance, our model replaces its gradient descent updates by Bayesian learning. Concretely, the weights $\mathbf{w}$ are learned by Bayesian linear regression of the bootstrap estimate $\hat{Q}\left(m_{k}, h_{k}\right)$ of the meta-level value function onto the features $\mathbf{f}$. The bootstrap estimator

$$
\begin{equation*}
\hat{Q}\left(m_{k}, h_{k}\right)=r_{\text {meta }}\left(m_{k}, h_{k}\right)+\left\langle\mu_{t}, \mathbf{f}\left(m^{\prime}, h^{\prime}\right)\right\rangle \tag{8}
\end{equation*}
$$

[^372]is the sum of the immediate meta-level reward and the predicted value of the next meta-level state $m^{\prime}$. The predicted value of $m^{\prime}$ is the scalar product of the the posterior mean $\mu_{t}$ of the weights $\mathbf{w}$ given the observations from the first $t$ actions (where $t=\sum_{n=1}^{k} h_{n}$ ) and the features $\mathbf{f}\left(m^{\prime}, c^{\prime}\right)$ of $m^{\prime}$ and the planning horizon $h^{\prime}$ that will be selected in that state.

We assume that the prior on the feature weights reflects that it is beneficial to plan until the end $\left(P\left(f_{5}\right)=\mathcal{N}(\mu=1, \sigma=\right.$ $0.1)$ ), although planning is costly $\left(P\left(f_{1}\right)=P\left(f_{2}\right)=P\left(f_{3}\right)=\right.$ $P\left(f_{4}\right)=\mathcal{N}(\mu=-1, \sigma=0.1)$ ), and that planning too much is more costly than planning too little $\left(P\left(f_{7}\right)=\mathcal{N}(\mu=-1, \sigma=\right.$ $0.1)$ and $\left.P\left(f_{6}\right)=\mathcal{N}(\mu=0, \sigma=0.1)\right)$.

Given the posterior on the feature weights $\mathbf{w}$, the planning horizon $h$ is selected by Thompson sampling. Specifically, to make the $k^{\text {th }}$ meta-decision, a weight vector $\tilde{w}$ is sampled from the posterior distribution of the weights given the series of meta-level states, selected planning horizons, and resulting value estimates experienced so far. That is,

$$
\begin{equation*}
\tilde{w}_{k} \sim P\left(\mathbf{w} \mid \mathcal{E}_{k}\right) \tag{9}
\end{equation*}
$$

where the set $\mathscr{E}_{k}=\left\{e_{1}, \cdots, e_{k}\right\}$ contains the meta-decisionmaker's experience from the first $k$ meta-decisions; to be precise, each meta-level experience $e_{j} \in \mathcal{E}_{k}$ is a tuple $\left(m_{j}, h_{j}, \hat{Q}\left(m_{j}, c_{j} ; \mu_{j}\right)\right)$ containing a meta-level state, the computation selected in it, and the bootstrap estimates of its Qvalue. The sampled weight vector $\tilde{w}$ is then used to predict the Q-values of each possible planning horizon $h \in \mathcal{H}$ according to Equation 7. Finally, the planning horizon with the highest predicted Q-value is used for decision-making.

By proposing metacognitive reinforcement learning as a mechanism of cognitive plasticity, our model suggests that reward and feedback are critical for cognitive growth. Conceptualizing metacognitive reinforcement learning as a regression problem suggests that learning how to best think about a problem should require less practice the stronger the correlation between the features $\mathbf{f}(m, c)$ (i.e., the predictors) and the resulting reward net the cost of thinking (i.e., the criterion; Green, 1991). Here, we apply our model to predict how quickly people can learn that more planning leads to better results from the reward structure of the practice problems. According to the model, learning should be fastest when the reward increases deterministically with the planning horizon both within and across problems. By contrast, learning should be slower when this relationship is degraded by additional variability in the rewards that is unrelated to planning. The following experiments test this prediction and illustrate the model's utility for designing feedback structures that promote metacognitive learning.

## Experiment 1: Reward structures can help or hinder learning to plan

## Methods

We recruited 304 adult participants from Amazon Mechanical Turk. The task took about 25 minutes, and participants were paid $\$ 2.50$ plus a performance-dependent bonus of up to

## Round 1 of 13

Location: Smithsville Flight: 1 of 2 Earnings: \$0 Bonus: \$0


Figure 1: Screenshot of a problem from Experiment 1.
\$2.00. Participants played a series of flight planning games. The environment consisted of six different cities, each connected to two other cities (Figure 1). Participants began each trial at a given city, and were tasked with planning a specified number of flights. Each flight was associated with a known gain or loss of money, displayed onscreen. Thus, the participants' task was to plan a route that would maximize their earnings or minimize their losses, based on the number of planning steps required for that game.

The experiment comprised thirteen trials total: a sequence of three practice problems which required planning 2,3 , and 3 steps ahead, respectively, followed by ten 4 -step problems, with a break after trial eight. The order of the two 3-step problems was randomized, and the order of the ten 4-step problems was randomized across the last ten trials of the experiment. Participants were assigned randomly to one of two conditions: environments with reward structures designed to promote learning ("diagnostic rewards"), or environments with reward structures designed to hinder learning ("non-diagnostic rewards").

The problems of the diagnostic rewards condition were automatically generated to exhibit four characteristics:

1. For each $l$-step problem, planning $h<l$ steps ahead generates $l-h$ suboptimal moves. In other words, each myopic planner makes the maximum possible number of mistakes.
2. When the number of moves is $l$, then planning $l$ steps ahead yields a positive return, but planning $h<l$ steps ahead yields a negative return.
3. The return increases monotonically with the planning horizon from 1 to the total number of moves.
4. Each starting position occurs at least once.

The reward structures used for the non-diagnostic rewards condition were created by shifting the diagnostic reward structures so as to degrade the correlation between planning horizon and reward. Concretely, for half of the problems all rewards were shifted down such that no amount of planning could achieve a return better than $-\$ 10$. Since the original problems were such that the 1-step planner always performed worst, the shift was $\frac{-r_{1}+X}{l}$ where $r_{1}$ is the return of the 1 -step planner, $l$ is the number of steps in the planning problem, and $X$ is a random number between 10 and 20 that differed across problems $(X \sim \operatorname{Uniform}([10,20]))$. For the other half of the problems, all rewards were shifted up by $-\frac{r_{1}+X}{l}$ such that all planners achieve a return of at least $+\$ 10$. These reward structures make it extremely difficult for metacognitive reinforcement learning to discover that planning is valuable, because the random shifts greatly diminish the correlation between planning horizon and reward.

## Results

Both model simulations and human behavior demonstrated enhanced learning in environments with diagnostic rewards. Figure 2 shows the mean performance of the metacognitive reinforcement learning model, and the mean performance of human participants. Here, performance is measured as relative reward

$$
\begin{equation*}
R_{\text {rel }}=\left(R-R_{\min }\right) /\left(R_{\max }-R_{\min }\right), \tag{10}
\end{equation*}
$$

where $R$ is the total reward received during the trial, and $R_{\text {min }}$ and $R_{\max }$ are the highest and lowest possible total reward on that trial, respectively.

To measure the effects of condition and trial number on performance in human participants, we ran a repeatedmeasures ANOVA. This revealed a significant effect of both trial number $(F(9,2989)=3.44, p<0.001)$ and condition $(F(9,3029)=15.26, p<0.0001)$, such that participants improved over time, and participants with diagnostic feedback performed better than those without. To measure learning in each group, we ran a simple linear regression of the relative reward on the trial number. This revealed a significant regression equation for participants who received diagnostic rewards $(F(2,302)=11.28, p<0.01)$, with an $R^{2}$ of 0.59 , but not for participants who received non-diagnostic rewards $(F(2,302)=3.51, p>0.05)$, with an $R^{2}$ of 0.31 , suggesting that improvement in performance occurred with diagnostic rewards, but not without.

To analyze the frequency with which participants chose the optimal route, we performed a multinomial logistic regression of whether or not each participant chose the optimal route on trial number and group. This revealed significant effects of trial number $\left(p<10^{-6}\right)$ and group ( $p<0.0001$ ).

In addition, we found that participants interacting with a diagnostic reward structure learned to plan significantly further ahead than participants interacting with the non-diagnostic reward structure. When there were four steps left, the average planning horizon was 2.96 with diagnostic rewards compared to 2.65 with non-diagnostic rewards $(t(596)=2.94$,


Figure 2: Model predictions and human performance in Experiment 1. Error bars indicate the standard error of the mean. Model predictions were averaged over 500 simulations.
$p<0.01$ ). When the rewards were diagnostic of good planning, participants' choices in the first step of the 4 -step problems accorded $10.3 \%$ more frequently with 4 -step planning $(t(302)=3.57, p<0.001)$. For 3 remaining steps there was a significant increase in choices according with optimal 1-step $(p<0.01)$, 2-step ( $p<0.01$ ) and 4-step planning $(p<0.01)$. For 2 remaining steps, there was a significant increase in choices according with optimal 1-step planning ( $p<0.0001$ ) without a decrease in agreement with other planning horizons. Finally, on the last move participants' choices in the environment with diagnostic rewards corresponded $5.8 \%$ more frequently with optimal 1 -step planning $(t(302)=3.71, p<0.001)$, and significantly less frequently with 2-step and 3-step planning ( $p<0.01$ and $p<0.001$ ). In summary, diagnostic rewards led to better agreement between the planning horizon and the number of remaining steps.

## Experiment 2: Using feedback to promote learning to plan

When one has control over the reward structure of an environment, creating rewards tailored to faster learning may be feasible. However, often environmental rewards are fixed. In Experiment 2, we tested whether providing feedback may be an effective alternative approach to accelerating learning. When participants do not plan enough to find the optimal route, this could be because the time cost of planning an optimal route outweighs its benefits. To change that, we provided feedback in the form of timeout penalties for short-sighted decisions.

## Methods

We recruited 324 adult participants on Amazon Mechanical Turk. The task took about 30 minutes, and participants were paid $\$ 3.00$ plus a performance-dependent bonus of up to $\$ 2.00$. Participants played twenty trials of the flight planning game described above. These trials were divided into a training block and a testing block. The training block con-
sisted of six trials requiring 2-step planning, followed by ten trials requiring 3 -step planning. The testing block consisted of four additional 3-step trials. The order of the 2-step trials and the order of the 3 -step trials were randomized across subjects. Participants were randomly assigned to either the feedback condition or the control condition.

In the training block, participants in the feedback condition were told their apparent planning horizon at the end of every trial and penalized with a timeout that reflected the amount of planning they had eschewed. Concretely, we set the durations of the timeouts such that the cost of short-sighted decisions was proportional to the amount of necessary planning the participant had eschewed. Specifically, the forgone cost of planning was estimated by cost $=2^{l-\hat{h}}$, where $l$ is the number of moves for that trial, $\hat{h}$ is the participant's apparent planning horizon, and 2 is the branching factor since each step entailed a binary decision. The participant's planning horizon was estimated by the number of consecutive moves consistent with the optimal policy, beginning with the last move, followed by the second-to-last, etc. At the end of each trial of the first block, participants in the feedback group were penalized with a timeout delay for sub-optimal routes. The delay was calculated as $7 \cdot($ cost -1$)$ seconds. During this period, participants were unable to proceed to the next trial. If participants performed the optimal route, they were able to proceed immediately to the next trial.

The control group received no feedback and had to wait a fixed amount of time at the end of every trial in block 1 , regardless of their performance. This fixed period was set to 8 seconds, to match the mean timeout period for participants in the feedback group ( 7.9 seconds). Neither group received feedback or delays in the test block.

The planning problems presented in this experiment were created in two steps. In the first step, we created 2- and 3step problems with maximally diagnostic reward structures (according to the criteria used in Experiment 1) subject to the constraint that the first move with the highest immediate reward was optimal for exactly half of those problems. In the second step, we modified these problems so as to deteriorate the correlation between planning horizon and reward using the same method we employed to create the non-diagnostic reward structures used in Experiment 1.

## Model Predictions

We applied the metacognitive reinforcement learning model described above to the problem of learning how many steps one should plan ahead. We simulated a run of the experiment described above with 1000 participants in each condition. The simulations predicted a gradual increase in the relative return from the first 3-step problem to the last one (see Figure 3). With feedback, the relative return increased faster and reached a higher level than without feedback.

## Results

To quantify the effects of condition and trial number on performance (measured as relative reward), we ran a mixed-



Figure 3: Results of Experiment 2. The metacognitive RL model predicts that feedback accelerate learning to plan. Human behavior shows a similar pattern of results.
design repeated-measures ANOVA on participant performance during the 3 -step trials. This revealed a significant effect of feedback $(F(9,4521)=8.54, p<0.01)$ and trial number $(F(9,4521)=1.85, p<0.05)$ on relative reward. To measure learning in each group, we performed a simple linear regression of relative reward on trial number for the 3step trials in the training block (i.e., when participants in the feedback group received feedback). This revealed a significant regression equation for the feedback group $(F(2,322)=$ $5.28, p=0.05$ ), with an $R^{2}$ of 0.40 but not for the control group $(F(2,322)=1.57, p>0.05)$, with an $R^{2}$ of 0.16 . This suggests that participants who received feedback improved during the training block but the control group did not.

Feedback increased the model's average performance in both the training block and the transfer block. We next tested whether the enhanced learning of the feedback group during training resulted in better performance in the transfer block (trials 17-20) where they no longer received any feedback. A two-sample $t$-test revealed that the feedback group's advantage in the testing block was nearly significant $(t(1294)=1.53, p=0.063)$. Figure 3 compares our participants' performance to the model predictions.

As predicted by our model, a multinomial logistic regression of whether or not each participant chose the optimal route on trial number and feedback, revealed significant effects of trial number $(p<0.0001)$ and feedback ( $p<0.01$ ).

Feedback appeared to increase people's planning horizons: when there were two remaining moves, the choices of the feedback group accorded $4 \%$ less often with myopic choice $(t(1398)=-2.17, p<0.05), 7 \%$ more often with optimal 2step planning $(t(1398)=3.44, p<0.001)$, and $4 \%$ more often with optimal 3-step planning $(t(1398)=2.43, p<0.05)$.

## Discussion

In this article, we have introduced a computational model of how people learn to decide better. Its central idea is that
learning how to think can be understood as metacognitive reinforcement learning. Our model extends previous research on strategy selection learning (Lieder et al., 2014; Lieder \& Griffiths, 2015) by capturing that choosing cognitive operations is a sequential decision problem with potentially delayed rewards rather than a one-shot decision. The new model correctly predicted the effects of reward structure and feedback on learning to plan: Experiment 1 suggested that our model captures the effect of reward structures on the speed of metacognitive learning. We then applied our theory to design feedback for people's performance in environments whose reward structure is not diagnostic of good planning. Experiment 2 confirmed the model's prediction that this intervention would be effective.

Our results suggest two pragmatic approaches to promoting cognitive growth: first, designing reward structures that are diagnostic of the quality of reasoning, planning, and decision-making; second, providing feedback on the process by which a decision was made. In Experiment 2 we followed the latter approach by designing feedback based on the cost of planning; but other types of feedback may also be useful. If cognitive plasticity is based on model-free reinforcement learning as assumed by our theory, then its speed should critically depend on how well the feedback people receive upon performing cognitive operations reflects their value. Therefore, feedback structures that align immediate feedback with long-term value should be maximally effective at promoting cognitive plasticity and learning to make better decisions. Future experiments should test this hypothesis by designing feedback structures using the optimal gamification method introduced by Lieder and Griffiths (2016). Feedback designed using optimal gamification could be especially beneficial because the underlying method of reward shaping is designed to accelerate model-free reinforcement learning ( Ng et al., 1999). Critically, to promote learning how to decide, people should decide without any assistance and only receive feedback after their choice.

We hope that our theory of metacognitive reinforcement learning will be a step towards establishing a scientific foundation for designing feedback for cognitive training and other interventions for promoting cognitive growth. Future work will evaluate alternative forms of feedback, address the problem of transfer and retention, and design more effective training paradigms using tasks that are maximally diagnostic of how people think and decide.

Acknowledgments. This work was supported by grant number ONR MURI N00014-13-1-0341. The authors thank Thomas Hills, Peter Dayan, Rika Antonova, Silvia Bunge, Stuart Russell, Amitai Shenhav, and Sebastian Musslick for feedback and discussions.

## References

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., ... others (2013). Video game training enhances cognitive control in older adults. Nature, 501(7465), 97101.

Bavelier, D., Green, C. S., Pouget, A., \& Schrater, P. (2012). Brain plasticity through the life span: learning to learn and action video games. Annual review of neuroscience, 35, 391-416.
Boureau, Y.-L., Sokol-Hessner, P., \& Daw, N. D. (2015). Deciding how to decide: self-control and meta-decision making. Trends in cognitive sciences, 19(11), 700-710.
Dolan, R. J., \& Dayan, P. (2013). Goals and habits in the brain. Neuron, 80(2), 312-325.
Green, C. S., \& Bavelier, D. (2008). Exercising your brain: a review of human brain plasticity and training-induced learning. Psychology and aging, 23(4), 692.
Green, S. B. (1991). How many subjects does it take to do a regression analysis. Multivariate behavioral research, 26(3), 499-510.
Hay, N., Russell, S., Tolpin, D., \& Shimony, S. (2012). Selecting computations: Theory and applications. In N. de Freitas \& K. Murphy (Eds.), Uncertainty in artificial intelligence: Proceedings of the twenty-eighth conference. P.O. Box 866 Corvallis, Oregon 97339 USA: AUAI Press.
Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. Proceedings of the National Academy of Sciences, 112(10), 3098-3103.
Jiang, N., Kulesza, A., Singh, S., \& Lewis, R. (2015). The dependence of effective planning horizon on model accuracy. In Proceedings of the 2015 international conference on autonomous agents and multiagent systems (pp. 1181-1189).
Keramati, M., Dezfouli, A., \& Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol, 7(5), e1002055.
Lieder, F., \& Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In Proceedings of the 37th annual conference of the cognitive science society.
Lieder, F., \& Griffiths, T. L. (2016). Helping people make better decisions using optimal gamification. In Proc. 38th annu. conf. cogn. sci. soc., philadelphia (pp. 2075-80).
Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., \& Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, \& K. Weinberger (Eds.), Advances in neural information processing systems 27 (pp. 28702878). Curran Associates, Inc.

Littman, M. L., Dean, T. L., \& Kaelbling, L. P. (1995). On the complexity of solving markov decision problems. In Proceedings of the eleventh conference on uncertainty in artificial intelligence (pp. 394-402).
Morrison, A. B., \& Chein, J. M. (2011). Does working memory training work? the promise and challenges of enhancing cognition by training working memory. Psychonomic bulletin \& review, 18(1), 46-60.
Myerson, J., \& Green, L. (1995). Discounting of delayed rewards: Models of individual choice. Journal of the experimental analysis of behavior, 64(3), 263-276.
Ng, A. Y., Harada, D., \& Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In I. Bratko \& S. Dzeroski (Eds.), Proceedings of the 16th Annual International Conference on Machine Learning (pp. 278-287). San Francisco: Morgan Kaufmann.
Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., ... Ballard, C. G. (2010). Putting brain training to the test. Nature, 465(7299), 775-778.
Russell, S., \& Wefald, E. (1991). Principles of metareasoning. Artificial Intelligence, 49(1-3), 361-395.
Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J., \& Botvinick, M. (2017). Toward a rational and mechanistic account of mental effort. Annual Review of Neuroscience, 40.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA, USA: MIT press.
Vul, E., Goodman, N., Griffiths, T. L., \& Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. Cognitive science, 38(4), 599-637.

# Thinking and Guessing: Bayesian and Empirical Models of How Humans Search 

Marta Kryven<br>University of Waterloo, Waterloo, Ontario, Canada<br>Tomer Ullman<br>Massachusetts Institute of Technology<br>William Cowan<br>University of Waterloo, Waterloo, Ontario, Canada<br>Josh Tenenbaum<br>Massachusetts Institute of Technology


#### Abstract

Searching natural environments, as for example, when foraging or looking for a landmark, combines reasoning under uncertainty, planning and visual search. Existing paradigms for studying search in humans focus on step-by-step information sampling, without examining advance planning. We propose and evaluate a Bayesian model of how people search in a naturalistic maze-solving task. The model encodes environment exploration as a sequential process of acquiring information modelled by a Partially Observable Markov Decision Process (POMDP), which maximises the information gained. We show that the search policy averaged across participants is optimal. Individual solutions, however, are highly variable and can be explained by two heuristics: thinking and guessing. Self-report and inference, a Gaussian Mixture Model over inverse POMDP, consistently assign most subjects to one style or the other. By analysing individual participants' decision times we show that individuals solve partial POMDPs and plan their search a limited number of steps in advance.


# The Puzzle of Conditionals with True Clauses: Against the Gricean Account 

Karolina Krzyżanowska (k.krzyzanowska@lmu.de)<br>Munich Center for Mathematical Philosophy \& Center for Advanced Studies, LMU Munich Geschwister-Scholl-Platz 1, 80539 Munich, Germany

Peter J. Collins (pcolli09@mail.bbk.ac.uk) and Ulrike Hahn (u.hahn@bbk.ac.uk)<br>Department of Psychological Sciences, Birkbeck, University of London<br>Malet Street, London, WC1E 7HX U.K.


#### Abstract

Indicative conditionals, that is sentences of the form "If $p$, then $q$," belong to the most puzzling phenomena of language. On the majority of accounts of indicative conditionals, the truth of $p$ and $q$ suffices for "If $p$, then $q$ " to be true or highly acceptable. Yet, many conditionals with true clauses, even if there is a meaningful connection between them, sound odd. The most common reaction to this phenomenon is to attribute the oddity of conditionals with true clauses to natural language pragmatics. We present an experimental study investigating how the presence or absence of a connection between the clauses affects the assertability of conditionals and conjunction expressing generic and specific kind of content. The results refute the standard pragmatic explanation.


Keywords: indicative conditionals; conjunctions; relevance; specific content; generic content; assertability

## Introduction

Indicative conditionals ${ }^{1}$ are sentences that we use to express hypothetical thoughts. They are central to our reasoning, planning, and problem solving. We entertain them when making everyday decisions ("If I add too much chilly to the curry, John will complain"), discussing public policies ("If we lower taxes, we will not have sufficient resources to fund social security benefits"), or doing science ("if we do not curb carbon output, sea levels will rise dangerously").

Conditionals are usually defined as sentences of the form "If $p$, then $q$," such as:
(1) If Dora studied physics, then she knows how to solve differential equations.

Intuitively, the antecedent of a conditional, $p$ ("Dora studied physics"), expresses a condition under which $q$, the consequent ("Dora knows how to solve differential equations"), occurs or from which it can be inferred. Assuming that it is true that Dora studied physics and that she knows how to solve differential equations, and given that a degree in physics is a good reason to believe that a person can solve differential equations, (1) is rendered true on any account that allows conditionals to be true or false at all, and highly acceptable on those accounts that deny conditionals their truth aptness (that is, accounts that do not view conditionals as statements that are 'true' or 'false,' in the same way that questions or commands are not true or false). But what if both $p$ and $q$ are true, yet there is no connection between them, that is, one cannot

[^373]infer $q$ from $p$ nor $p$ makes it more likely that $q$ ? Let us suppose that Dora can solve differential equations, and that she also plays basketball. The fact that Dora is a basketball player does not allow us to predict anything about Dora's mathematical skills. The two facts do not seem to be connected at all, yet on many prominent accounts of conditionals, the sentence
(2) is rendered true or, at least, highly acceptable:
(2) If Dora plays basketball, then she knows how to solve differential equations.

This is due to the fact that the majority of the prominent theories of conditionals validate the Principle of Conjunctive Sufficiency, often simply referred to as the Principle of Centering, which allows us to infer "If $p$, then $q$ " from the conjunction of $p$ and $q$.

Centering has recently attracted attention in psychology of reasoning (Cruz et al. 2016), because this is an inference rule that distinguishes between some of the most popular philosophical and psychological accounts of conditionals, such as the Mental Models Theory (Johnson-Laird \& Byrne 2002) and the suppositional theory (Adams 1975; Edgington 1995; Evans \& Over 2004; Cruz et al. 2016), on the one hand, and the 'inferentialist' approach on which a connection between a conditional's antecedent and its consequent belongs to the literal, semantic meaning of a conditional, on the other hand. This connection may be defined in different ways, for instance, as a whole variety of inferential relations (Krzyżanowska et al. 2013, 2014) or in terms of probabilistic relevance (Skovgaard-Olsen et al. 2016b,a), but on no 'inferentialist' account of conditionals will a sentence like (2) be acceptable unless one can show that there is some kind of relationship between basketball and maths.

This is not to say that proponents of the 'centering' theories do not find sentences like (2) strange. They do, but assume that the oddity of missing-link conditionals can be explained in terms of pragmatics, that is, the aspect of language that allows speakers to infer the intended meaning of linguistic expressions even where it is different than what is literally said. For instance, "Some students passed the exam" pragmatically implicates that not all of them passed, although, from a purely logical point of view, it is consistent with "All students passed the exam" (Bott \& Noveck 2004). Along these lines, (Over et al. 2007, p. 92) make the following observation. Anyone who takes the natural-language conditional to
be probabilistic-its meaning exhausted by the Equation, i.e. $\operatorname{Pr}$ ("If $p, q$ ") $=\operatorname{Pr}(q \mid p)$-can argue that:
the use of a conditional pragmatically suggests, in certain ordinary contexts, that $p$ raises the probability of $q$ or that $p$ causes $q$.

A similar take on the connection between $p$ and $q$ can be found in the Mental Models literature. On this account, a language user interprets an assertion by constructing mental models (Johnson-Laird \& Byrne 2002, p. 653). On the most recent version of Mental Models Theory, the core meaning of a natural language conditional refers to a set of possibilities equivalent to the material interpretation of a conditional (p. 665). However, what kind of possibilities a language user envisages when interpreting a sentence is susceptible to the processes of semantic and pragmatic modulations. In particular:
modulation can establish an indefinite number of different temporal, spatial, and coreferential relations between the antecedent and consequent of a conditional. (Johnson-Laird \& Byrne 2002, p. 660)

But what kind of pragmatic mechanisms are responsible for the oddity of missing-link conditionals? The most famous pragmatic explanation of why some conditionals, and especially those that are true due to the truth of their consequents, appear odd was proposed by Grice (1989). He argued that:

To say that " $p \supset q$ " is to say something logically weaker than to deny that $p$ or to assert that $q$, and is thus less informative; to make a less informative rather than a more informative statement is to offend against the first maxim of Quantity, provided that the more informative statement, if made, would be of interest. There is a general presumption that in the case of " $p \supset q$," a more informative statement would be of interest (Grice 1989, p.61).

On this account, a sentence such as (2) is simply unassertable in a context in which its antecedent and consequent are known to be true, because a stronger statement, namely "Dora knows how to solve differential equations" is available and should have been asserted instead. Note, however, that even if an appeal to the maxim of Quantity explains why sentences like (2) are not felicitous things to say, it does not illuminate the fact that speakers seem to interpret conditionals as if their clauses were somehow connected. Moreover, it also does not allow us to distinguish between sentences such as (2) and those conditionals with true clauses that are, intuitively, perfectly fine, like (1).

Both (1) and (2) consist of a true antecedent and true consequent. The only difference between them is that there is an inferential connection between doing physics and possessing certain mathematical skills, whereas playing basketball, as far as we know, has no bearing on the latter at all. The truth of
$p$ and $q$ is clearly not enough for a conditional to be a reasonable thing to say or to accept, but neither does it suffice to render a conditional unassertable at all.

## Connecting antecedents and consequents

In the case of what we will label in the following as a TT conditional, that is, a conditional whose antecedent and consequent are (known to be) true, the connection between the clauses cannot be translated directly into 'possibilities' in the way envisaged by Mental Models Theory, or into the notion of the probabilistic relevance, understood in terms of the $\Delta p$ rule, if the conditional probability, $\operatorname{Pr}(q \mid p)$, is understood as the ratio of $\operatorname{Pr}(q \wedge p)$ to $\operatorname{Pr}(p)$ (Over et al. 2007; Oberauer et al. 2007; Skovgaard-Olsen et al. 2016b). On this widespread probabilistic approach, $p$ is said to be positively relevant for $q$ if $\Delta p>0$, where $\Delta p$ is defined as a difference between $\operatorname{Pr}(q \mid p)$ and $\operatorname{Pr}(q \mid \neg p)$. However, when both $p$ and $q$ are known to be true, $\operatorname{Pr}(p)=\operatorname{Pr}(q)=\operatorname{Pr}(q \mid p)=1$, whereas $\operatorname{Pr}(q \mid \neg p)$ is undetermined since $\operatorname{Pr}(\neg p)=0$, and hence $\Delta p$ cannot be calculated. Another, related, measure of probabilistic relevance is the difference between $\operatorname{Pr}(q \mid p)$ and $\operatorname{Pr}(q)$ (cf. Skovgaard-Olsen et al. 2016a). $\operatorname{Pr}(q \mid p)>\operatorname{Pr}(q)$ ensures that the antecedent has a probability-raising effect on the consequent. However, when both $p$ and $q$ are known to be true, $\operatorname{Pr}(q)=\operatorname{Pr}(q \mid p)$. In that case, $p$ is probabilistically irrelevant for $q$ just because $\operatorname{Pr}(q)$ cannot be raised any higher. Yet, as the example (1) illustrates, the clauses of a TT conditional may seem connected anyway, hence probabilistic relevance defined in this way is insufficient to capture the intuition behind that connection. By the same token, Johnson-Laird and Byrne's (2002) suggestion that the core meaning of the conditional is that the "antecedent describes a possibility, at least in part, and the consequent can occur in this possibility." ( p . 650), is of no help with TT conditionals, as true states of affairs are necessarily 'possible,' so that the notion of possibility is insufficient to distinguish between 2 and 1.

Some of these problems can be avoided when conditional probability, rather than unconditional probability, is treated as the primitive notion, and thus $\operatorname{Pr}(q \mid p)$ is not calculated from $\operatorname{Pr}(p \wedge q)$ and $\operatorname{Pr}(p)$ (e.g. Popper 1959; de Finetti 1970/1990). One may also consider a counterfactual notion of relevance, that is, e.g., $\Delta p$ calculated as if $p$ was not known to be true. ${ }^{2}$ In our experimental design, we assume an intuitive, pre-theoretic notion of the connection, which does not depend on any particular operationalisation of the notion of relevance.

## What antecedents and consequents are about

In order to understand the semantics and pragmatics of conditionals, one should arguably turn to the way conditionals are actually used in everyday language. Linguists have sought to provide extensive overviews of different types of conditionals

[^374]Jenny enters the room of her 14 year old son, Tom, to persuade him to clean up his room a bit. She even offers him her help. While sorting through the papers on Tom's desk, Jenny notices a math test with an " $F$ " on it, which, apparently, Tom has concealed from her. Tom notices what his mother found and decides to divert her attention to something else.

Tom: "Oh, look, my plant has dried up completely."
Jenny: "Did you water it at all?"
Tom: "Well, no..."
In this context, would it make sense for Jenny to say the following?


Figure 1: Example vignette used in the experiment.
(see, e.g, Declerck \& Reed 2001, for a comprehensive corpus based analysis). Such overviews reveal that conditionals can be characterised and classified in many different ways. Consideration of those differences suggests that the broad category of 'indicative conditional' (as in, 'in the indicative mood' as opposed to the 'subjunctive mood', e.g., I would, I could, etc.) which is the theoretical focus of the majority of psychological and philosophical work is far too broad. The way a conditional is interpreted and evaluated may be affected by the kind of content expressed by its clauses, even where the content is broadly 'indicative'. Among others, the content of a clause may be specific, that is, the clause can be about a specific object (a token) known to all participants of the conversation, e.g. "this book," or generic, concerning a type of an object, e.g. "a book." Consequently, we can distinguish between generic and specific conditionals:

## Generic (type):

(3) If a book is hardcover, it is expensive.

Specific (token):
(4) If this book is hardcover, it is expensive.

As noted by Declerck \& Reed (2001, p. 2), the unacceptability of conditionals with true antecedents:
is due to the fact that a speaker cannot process a fact as a supposition, except in 'inferential' conditionals, i.e., in conditionals expressing a conclusion $Q$ that is drawn from a premise $P$.
We investigate whether there is a difference in how people process generic and specific content, and consequently, if the two types of content may have an effect on people's assertability judgements. Additionally, we hypothesise that the presence or absence of an inferential connection may make the
conditional assertable, even if the antecedent is true and hence difficult to process as a supposition.

## The present experiment

We investigated whether people's assertability judgements depend on what a conditional is about. More specifically, we were interested in two factors that may be expected to affect people's evaluations of a conditional: the presence of an inferential connection between antecedent and consequent, and the kind of content the conditional expresses. We compared how people evaluate conditionals with how then evaluate conjunctions consisting of the same true clauses, such as:

Conditional: If you didn't water your plant, you failed your math test.
Conjunction: You didn't water your plant and you failed your math test.
Our test followed a $2 \times 2 \times 2$ factorial design. Sentence type (conditionals vs. conjunctions) was manipulated within subjects. Type of content (generic: type vs. specific: token) and the inferential connection (presence: C+ vs. absence: C-) were manipulated between subject. These are examples of conditionals belonging to each of the resulting four groups:

Token C+ If you didn't water your plant, it dried up.
Token C- If you didn't water your plant, you failed your math test.
Type C+ If you don't water a plant, it dries up
Type C- If you don't water a plant, you fail math tests.
Finally, we asked participants to evaluate the extent to which given sentences are reasonable things to utter in given contexts by means of two different questions:

Assertability: "In this context, to what extent would it be natural for X to assert the following sentences?"

Sensibleness: "In this context, would it make sense for X to say the following?"

## Methods

## Participants

244 individuals participated in the online survey on the MTurk platform (https://www.mturk.com/). We removed four participants who did not complete the survey and three participant whose reported first language was not English. Of the remaining 237 participants, 115 were female. The mean age of the participants was 33.97 (range 18-63). All participants received a small remuneration for their time and effort.

## Materials and procedure

Each participant has been randomly assigned to one of the four groups: Type C+, Type C-, Token C+, Token C-. Participants in each group were presented with 8 blocks, one at a time. The order of presentation was randomised. Each block contained a vignette consisting of a conversational context followed by two sentences: a conditional and a conjunction, presented in randomised order. There were four types of vignettes. Each participant saw each vignette twice: once followed by the question about the "assertability" of the two sentences, and once followed by the question about their "sensibleness." Figure 1 shows an example Token-C- item used in the experiment.

## Results

The data on 'sensibleness' mirrored exactly the pattern of responses to the 'assertability' question. Therefore, for brevity's sake, we report only the analysis of the latter question, the descriptive statistics for which are reported in Table 1 , below.

Table 1: Descriptive statistics for 'assertability.'

|  |  | Type |  | Token |  |
| :--- | ---: | ---: | :---: | :---: | :---: |
|  |  | C+ | C- | C + | C- |
| Conditionals | mean | 5.51 | 2.52 | 3.92 | 1.48 |
|  | sd | 0.94 | 1.08 | 1.35 | 0.74 |
| Conjunctions | mean | 3.79 | 4.59 | 5.98 | 6.04 |
|  | sd | 1.48 | 1.30 | 0.79 | 1.01 |

As figure 2 suggests, the absence of an inferential connection makes both type and token TT conditionals unassertable, although the presence of connection does not seem to be enough to make a token conditional assertable. By contrast, the presence or absence of a connection has little effect on the assertability of token conjunctions. When there is no connection at all, conjunctions are judged to be more assertable than conditionals. At the same time, while token $\mathrm{C}+$ conjunctions seem to be more assertable than token $\mathrm{C}+$ conditionals, type $\mathrm{C}+$ conjunctions are less assertable than type $\mathrm{C}+$ conditionals.

We performed analyses in R (R Core Team, 2016) using functions from Wilcox (2016) and the WRS2 package
(Mair, Schoenbrodt, \& Wilcox 2016). The robust analyses used $20 \%$ trimmed means. This difference between type and token is confirmed by significant 2-way interaction in a robust mixed ANOVA ('bwtrim' function) between the variables type/token and conditional/conjunction, both with an inferential connection $(\mathrm{C}+)(Q=111.74, p<.001)$ and without (C-) $(Q=68.11, p<.001)$. We thus conducted further statistical analyses separately for token and type materials.
Token A robust mixed ANOVA produced significant main effects of Sentence Type, $Q=485.34, p<.001$, and Connection, $Q=83.94, p<.001$, and a significant interaction, $Q=$ $77.98, p<.001$. Simple-effects analyses comprise robust t -tests. We compared the effect of sentence type separately for items with and without a connection using robust pairedsamples t-tests (the yuend function). When there was a connection, conjunctions were rated on average 2.11 higher than conditionals; this difference was significant, $t(26)=8.66$, $p<.001, r=.89$. When there was no connection, conjunctions were rated on average 4.94 higher than conditionals; this difference was significant, $t(35)=23.83, p<.001, r=.94$. We explored the effect of the connection for each type of sentence, using robust independent-samples t -tests (the 'yuen' function). For conditionals, ratings were on average 2.67 higher with a connection than without; this difference was significant, $t(43.78)=12.25, p<.001, r=.90$. For conjunctions, ratings were on average .16 higher without a connection than with; this difference was not significant, $t(57.79)=.80$, $p=.43, r=.11 .^{3}$

In sum, the token data showed that conjunctions were rated consistently higher than conditionals. Conditionals were rated higher with a connection; conjunctions nonsignificantly higher without one.

Type A robust mixed ANOVA produced a non-significant main effect of Sentence Type, $Q=1.79, p=.19$. The main effect of Connection was significant, $Q=40.21, p<.001$, as was the interaction, $Q=134.85, p<.001$. As above, simple effects of sentence type comprised robust paired-sample t -tests. With a connection, conditionals were rated on average 1.79 higher than conjunctions; this difference was significant, $t(35)=6.46, p<.001, r=.83$. Without a connection, conjunctions were rated on average 2.28 higher than conditionals; this difference was significant, $t(35)=9.20, p<.001$, $r=.87$. As above, simple effects of connection comprised robust independent-samples $t$-tests. For conditionals, ratings were on average 3.15 higher with a connection than without; this difference was significant, $t(65.41)=14.98, p<.001$, $r=.90$. For conjunctions, ratings were on average .92 higher without a connection than with one; this difference was significant, $t(68.26)=3.20, p=.002, r=.10$.

In sum, the type data showed that, when there was a connection, conditionals were rated higher than conjunctions. When there was no connection, conjunctions were rated

[^375]

Figure 2: Distribution of the responses to the 'assertability' question.
higher than conditionals. For conditionals, items with a connection were rated higher than items without; for conjunctions, the opposite was the case.

## Discussion

Our data clearly show that knowing that $p$ and $q$ are true is not sufficient for "If $p, q$ " to be assertable. The presence of an inferential connection between $p$ and $q$ is not sufficient either, yet it is necessary for a conditional to be assertable. However, the presence of a connection does not seem to affect the assertability of conjunctions in a similar way. In fact, conjunctions in which conjunct are inferentially connected tend to be rated lower than those without a connection. Our findings pose a problem for all theories of conditionals that treat the intuition that conditionals are about connections solely as a pragmatic aspect of their meaning. Our results undermine the standard pragmatic account of the oddity of missing link conditionals. On this account, asserting a TT conditional is a violation of the Maxim of Quantity, because when both $p$ and $q$ are known to be true, one is justified in asserting a stronger, more informative statement. That is, one should assert the conjunction of $p$ and $q$. As our data clearly show, when an inferential connection is present, generic conditionals are more
assertable than generic conjunctions which cannot be reconciled with the standard, Gricean account.

A Gricean explanation of these findings can follow two different paths: One can reject the most fundamental principles of Gricean pragmatics by denying that informativeness guides people's assertability judgements. Another, less costly, option is to rethink the semantics of conditionals and accept the possibility that the connection between antecedents and consequents is an important part of their meaning. On the latter approach, the connection should be taken as an additional piece of information conveyed by a conditional, but not by a conjunction. Our findings suggest that we need an account that renders conditionals more informative than conjunctions (on their standard, truth-functional interpretation). ${ }^{4}$ Neither Mental Models nor the suppositional account of conditional can follow the latter path, however. Johnson-Laird \& Byrne (2002, p. 651) deny that a relation between $p$ and $q$ is part of the core meaning of a conditional: the logic of a conditional on this interpretation is essentially the logic of

[^376]material implication. Although Johnson-Laird \& Byrne admit that content of the clauses of a conditional (semantic modulation) and contextual factors (pragmatic modulation) influence the interpretation of a conditional, the only thing that these mechanisms do is constrain the set of possibilities a speaker envisages when interpreting a sentence. On Mental Models theory, when $p$ and $q$ are known to be true, the interpretation of a conditional and the interpretation of a conjunction do not seem to be distinguishable at all. Apart from vague appeals at 'pragmatic modulations,' Mental Model theory has no way to explain our findings.

The data are no less problematic for the suppositional account. First of all, as observed earlier, none of the standard probabilistic measures of relevance is applicable when both clauses of a conditional are known to be true, if conditional probability is understood in accordance with Kolmogorov's axioms. Second, even if we grant the supporters of the probabilistic interpretation of a conditional that the relevance of $p$ for $q$, however it is formalised, is pragmatically implicated, this account does not have the means to explain the discrepancy between type and token TT conditionals.

Token conditionals are sentences such as "If you didn't water your plant, it dried up" or "If this book is hardcover, it is expensive." Although a hard cover $(p)$ is a good evidence that the book costs a lot $(q)$, i.e., one can infer $q$ from $p$ together with some general knowledge about the world, the conditional sounds strange to many respondents. In the same context, a generic conditional "If a book is hardcover, it is expensive" is evaluated as assertable. This discrepancy might be due to the fact that the specific antecedent, which refers to a particular book that is directly available to conversational partners, is considered a fact and hence, as suggested by Declerck \& Reed (2001, p. 2), is more difficult to process as a supposition from which one is to make an inference (cf. Elder \& Jaszczolt 2016 on the notion of remoteness). By contrast, the generic antecedent "a book is hardcover" requires an additional inferential step to be evaluated as true (if something is true about this book in front of us, it is also true about $a$ book), hence the evidence for the truth of the antecedent can be considered inferential and not direct.

Most fundamentally, however, our results suggest that an adequate account of the conditional and of reasoning with conditionals will have to engage more seriously with the circumstances in which conditionals can and cannot be used in everyday language.

## References

Adams, E. W. (1975). The logic of conditionals. Dordrecht: D. Reidel.

Bott, L., \& Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. Journal of Memory and Language, 51, 437-457.
Cruz, N., Over, D., Oaksford, M., \& Baratgin, J. (2016). Centering and the meaning of conditionals. In A. Papafragou, D. Grodner, D. Mirman, \& J. C. Trueswell (Eds.), Proceed-
ings of the 38th annual conference of the cognitive science society (p. 1104-1109). Austin, TX: Cognitive Science Society.
de Finetti, B. (1970/1990). Theory of probability: A critical introductory treatment (Vol. 1). Chichester: John Wiley \& Sons.
Declerck, R., \& Reed, S. (2001). Conditionals: A comprehensive empirical analysis. Berlin/New York: Mouton de Gruyter.
Edgington, D. (1995). On conditionals. Mind, 104(414), 235-329.
Elder, C.-H., \& Jaszczolt, K. M. (2016). Towards a pragmatic category of conditionals. Journal of Pragmatics(98), 3653.

Evans, J. S. B. T., \& Over, D. E. (2004). If. Oxford: Oxford University Press.
Grice, H. P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.
Johnson-Laird, P. N., \& Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. Psychological Review, 109(4), 646-678.
Krzyżanowska, K., Wenmackers, S., \& Douven, I. (2013). Inferential conditionals and evidentiality. Journal of Logic, Language and Information, 22(3), 315-334.
Krzyżanowska, K., Wenmackers, S., \& Douven, I. (2014). Rethinking Gibbard's riverboat argument. Studia Logica, 102(4), 771-792.
Mair, P., Schoenbrodt, F., \& Wilcox, R. (2016). WRS2: Wilcox robust estimation and testing.
Oberauer, K., Weidenfeld, A., \& Fischer, K. (2007). What makes us believe a conditional? the roles of covariation and causality. Thinking \& Reasoning, 13(4), 340-369.
Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., \& Sloman, S. A. (2007). The probability of causal conditionals. Cognitive Psychology, 54, 62-97.
Popper, K. (1959). The propensity interpretation of probability. British Journal of the Philosophy of Science, 10, 25-42.
R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.r-project.org/
Rott, H. (1986, 370). If s, Though, and Because. Erkenntnis, 25, 345.
Skovgaard-Olsen, N., Singmann, H., \& Klauer, K. C. (2016a). Relevance and reason relations. Cognitive Science. doi: 10.1111/cogs. 12462
Skovgaard-Olsen, N., Singmann, H., \& Klauer, K. C. (2016b). The relevance effect and conditionals. Cognition, 150, 26-36.
Wilcox, R. R. (2016). Understanding and applying basic statistical methods using R. Hoboken, NJ: John Wiley \& Sons.

# Cascading effect of context and competition on novel word learning 

Sarah C. Kucker (kuckers@uwosh.edu)<br>Aaron Bagley (baglea45@uwosh.edu)<br>The University of Wisconsin Oshkosh; Department of Psychology<br>800 Algoma Blvd; Oshkosh, WI 54901 USA


#### Abstract

Learning, especially in the case of language acquisition, is not an isolated process; there is ever-present competition between words and objects in the world. Such competition is known to play a critical role in learning. Namely, the amount and variability of competing items during word learning have been shown to change learning trajectories in young children learning new words. However, very little work has examined the interaction of competition amount, competition variability, and task demands in adults. The current study assesses adults' ability to map new word-referent pairs in varying amounts of competition and competitor variability. In addition, the effect of mapping context on retention was assessed. Results suggest that retention is weak in some cases and importantly, there are cascading effects of competitor variability in mapping on later retention of new words. Results are discussed in light of associative learning mechanisms and the implications of competition for learning.


Keywords: word learning; fast mapping; context; competition

## Introduction

Word learning always occurs in a context. New words are encountered in a conversation, new objects are alongside others on a shelf, and word-referent pairs are embedded in fleeting moments. Often, the referent for a word is not immediately clear. It is precisely this complexity and ambiguity that has puzzled researchers for decades - how do individuals, and young children in particular, correctly map a novel word to its referent amidst dozens of possible objects in the world around them? The last few decades of research have subsequently resulted in multiple pathways by which children overcome the problem of referential ambiguity to acquire new words and add them to their lexicon. We know that children appear to use their prior knowledge of objects (e.g. mutual exclusivity; Markman \& Wachtel, 1988), novelty (e.g. N3C; Mervis \& Bertrand, 1994), and social cues (Akhtar, Carpenter, \& Tomasello, 1996) to eliminate competitors and focus on the target. We know that the number of competitors present (Zosh, Brinster, \& Halberda, 2013) and the context in which objects are encountered (Horst, 2013) matter. And we know that as a child's vocabulary grows, so too does their performance in word learning tasks (Bion, Borovsky, \& Fernald, 2013). There has also been parallel research on how both adults (Alt \& Gutman, 2009; Greve, Cooper, \& Henson, 2014; Halberda, 2006; Warren \& Duff, 2014) and robots (Morse et al., 2015; Twomey, Morse, Cangelosi, \& Horst, 2016) acquire new words.

Most recently, there has been a further push to view word learning as a process that unfolds over multiple timescales instead of a one-shot learning episode (Kucker, McMurray, \& Samuelson, 2015). The initial mapping between a new word and its referent happens in real-time. This is an in-themoment process of comprehending "Where's the wif?" or "Get the dax." However, this initial association between the word and referent is often weak and transient. Over time and repetitions the word-referent association becomes honed and strengthened, partially by reinforcing the correct links but also by pruning away incorrect associations (McMurray, Horst, \& Samuelson, 2012). Regardless if it is children, adults, or robots, word learning requires real-time responses to interact with slower associative mechanisms. Successful long-term learning thus depends partially on the context in which the word-referent pair was initially encoded as well as the context in which it is stored and retrieved.

## Competition in Mapping

Across all work on word learning, there is one factor that is consistent - competition. Competition occurs between objects present in the world, between words in the lexicon, and between former knowledge and new. At times there exists a lot of competition, while at other times the competition is limited. Sometimes, the competing items that may be present are relevant and helpful (e.g. learning about forks while in the kitchen), and sometimes the competitors and context are not. In addition, the initial context in which a word is mapped to its referent can vary widely not just between people or words, but even across occurrences of a single word. That is, the word fork may be heard primarily in the kitchen, but occasionally could be used while outside. This competition and its variability have cascading effects on retrieval and retention of newly learned information.

One of the classic tests of word learning is Carey and Bartlett's (1978) study of "fast-mapping". Here, preschool children were presented with a novel word while in a classroom, surrounded by multiple people and objects and tested on their ability to accurately "Get the chromium tray, not the blue one". This referent selection task has been used for decades to test not just an individual's ability to map words and referents in ambiguous scenes, but also retain newly mapped word-referent pairs until a later time.

A critical part of the classic referent selection task is the presence of competition - in order to get the chromium tray, children had to decide against the blue tray and not pick up a cup or spoon. In many variations of the task, there is obviously competition between multiple items that may be present (e.g. tray vs. cup). Other versions of the task
highlight the competition between known items and novel (e.g. blue vs. chromium). And others try to eliminate direct competition by directly naming a single item. Regardless of the number or type of competitors, both children and adults are easily able to map a novel word to a novel referent in-the-moment (Horst, Scott, \& Pollard, 2009; Warren \& Duff, 2014; Greve, et al., 2014), becoming increasingly successful over development (Bion, Borovsky, \& Fernald, 2013). However, real-time eye-tracking behavior does suggest there are differences in mapping due to competition. For instance, when children and adults are presented with both a known and a novel object and hear a novel word, they reliably look to the competing known items before settling on the novel referent (Halberda, 2006). In addition, there is evidence that explicit selection of a competitor (in error) can still lead to later learning (Fitneva \& Christiansen, 2011). Thus, the simple presence of competitors can change even subtle selection behaviors and cognitive processing.

Furthermore, other work suggests that the details about how many competitors or the variability of the competitors also shift real-time processing, which has cascading effects on long-term learning. First, work with children suggests that the number of competitors present during mapping has a direct negative correlation on performance - fewer items seem to help mapping (Axelsson, Churchley, \& Horst, 2012), but more competitors benefit long-term learning (Zosh, et al., 2013). And, computational models of referent selection and retention propose that though more competition during mapping requires more cognitive processing, it also results in more opportunities for pruning away spurious word-referent links, thus allowing more honing in a single trial (McMurray, et al., 2012). This subsequently boosts retention (McMurray, Zhao, Kucker, \& Samuelson, 2013). In a recent study of adult word learning, encoding new words with competitors resulted in almost immediate lexical competition between the new word and known word. However, learning that occurred in an isolated, direct-naming context did not (Coutanche \& ThompsonSchill, 2014). Competitive initial learning also led to broader semantic networks days later. Taken together, the amount of competition influences real-time processing and has cascading effects on long-term retention.

In addition to the amount of competition present, the type of competition may also play a role. Specifically, the relevance and variability of competing known foil items interacts with new word-referent pairs. Recent work has demonstrated that young children often learn new words better when couched in a known context, such as when drawn from a category of objects they are familiar with (Borovsky, Ellis, Evans, \& Elman, 2016), or when an item is physically located in a related scene (Meints, Plunkett, Harris, \& Dimmock, 2004). It is also clear that both children and adults pay attention to competitors even if they are not the target. In some cases, children implicitly extract some minor representation or memory for competitors such that they are more likely to look at them later (Wojick, 2013). In other cases, adults recall semantic information about the
novel target, acquired while the target was presented in various scenarios (Alt \& Gutman, 2009). In addition, context diversity is more predictive of word knowledge in adults (and vocabulary acquisition in children) than simple frequency of occurrence (Adelman, Brown, \& Quesada, 2006; Hills, Maouene, Riordan, \& Smith, 2010). Put together, this suggest that the composition of, and specifically the relationship between, the competitors might matter not just for mapping, but also for retention.

## Competition in Retention

Generally, adults are considered to be good at learning new words, performing well above chance on laboratory tasks (Greve, et al., 2014). However, we also know that an individual's ability to retain newly learned words is reliant on a number of factors, such as the length of delay (Vlach \& Sandhofer, 2012; McGregor, 2014) and phonological neighbors (Tamminen \& Gaskell, 2008). Work with children has suggested that retention is also partially dependent on the number of competitors present during mapping (Zosh, et al., 2013), and the types of cues present during encoding (Capone \& McGregor, 2005). Furthermore, most work showing good retention with adults has thus far primarily used relatively easy task - pulling the target out of a three-item array. Thus, though adults are expected to be good at learning new words, it is likely that competition can still play a role. A recent study by Dautriche and Chemla (2014) found context effects in crosssituational learning where consistent repetition of competitors over the course of multiple blocks led to improvements in memory for both the broader context group and target. Learning here was tested through forcechoice selection of the target from four-item array. In another study, Warren and Duff (2014) tested retention after a zero competitor/ ostensive (direct) naming condition as well as a two-item referent selection condition. Critically, they looked at retention with both a three-alternative forcedchoice recognition test (3AFC) and free recall (recalling word-forms from test). Overall, typically developing adults selected the referent on the 3AFC test well above chance, near $65 \%$ correct, for both conditions. However, free recall varied between the mapping contexts - accuracy was higher for words learned directly without competitors than those with competition during mapping. That is, though context may influence initial encoding, the method for tapping retention is critical in assessing the robustness of the newly acquired word-referent pair.

Despite the lack of work on competition effects on the retention of new words, the evidence suggests that the associative learning processes of adults, like children, are highly variable and occur incrementally over multiple timescales (Kucker, et al., 2015). Such associative processing occurs over the lifespan and critically, some propose that the complex networks in which new words are learned are also linked with advances in other areas of cognition (Roembke, Wasserman, \& McMurray, 2016). Thus, the process of learning a single new word has
important implications for broader theories of associative learning and cognition.

## Current Study

Competition is widely accepted as highly relevant for word learning, but there is still much debate about how much competition is optimal and what effects initial competition has on long-term learning. Both the number and variability of competitors has been studied in word learning of young children (though with mixed results); much less work has been conducted in adults despite the fact that adults too, continue to learn new associations. In addition, the question of how adults learn simple associations (such as words) is at the crux of many computational models of learning and information processing (e.g. Regier, 2005; McMurray, et al., 2012). Thus, there is a gap in understanding both how an adult's real time processing changes across contexts and the impact the contexts have on their retention of new wordreferent associations. The current study fills this gap by combining word learning methods from work with children with the adult literature to test the impact of competition (amount and variability) on in-moment mapping, and the cascading effects mapping context has on retention. Furthermore, learning is assessed with two different methods to tap both weak associations that may be recognized in an array but not recalled as well as robust associations that can be easily recalled without cues.

## Methods

## Participants

A total of 149 adults participated. All individuals were monolingual English speakers and provided informed consent before participating. Individuals were recruited either through a current college course (receiving course credit), or through Amazon's MTurk (receiving \$1.75).

## Stimuli

Individuals saw a randomly selected subset of known and novel words and objects over the course of the experiment. Known items were drawn from pools of prototypical toys (book, ball, drum, block), kitchen utensils/tools (fork, spoon, bowl, cup), vehicles (car, boat, airplane, bike), clothing (shoe, hat, belt, purse), or furniture (chair, bed, lamp, clock). All images were previously normed by a separate set of adults to elicit their respective labels and be obvious members of their category. Novel items were drawn from a pool of unique items typically used in child word learning studies. These items were judged in prior studies to be highly unfamiliar and not easily named by most adults (Horst \& Hout, 2016). Novel words were legal words in English but had few or no known orthographic neighbors.

Table 1. Mapping conditions

| Condition | \# of <br> Comp <br> etitors | "Which is the cheem?" |
| :--- | :--- | :--- |
| OD none | 0 |  |

and all completed the entire study on-line in a single sitting. Conditions varied in the number of competitors present during initial mapping and the variability of competitors present (Table 1). The number of competitors varied from zero to four with the to-be-learned novel object present alongside no other items (ostensive definition trial; OD none), paired with two distinct known competitors ( 3 total items, alternative forced-choice trial; 3AFC), or paired with four distinct known competitors (5AFC) on each trial. Competitors also differed in their variability - competitor items were either variable, drawn from distinct known categories (vary), or clustered and drawn from the same category (same).

Each condition began with a mapping phase in which participants were initially exposed to novel word-referent pairs, followed by a retention phase in which they were tested on their memory for the word-referent pairs from mapping. During the mapping phase, participants were instructed to pay attention to a series of word-referent pairs presented on the screen. One each trial, a novel printed word was presented along with an image of a novel object. In the OD none ( 0 competitions) condition, no other items were present on the screen. In the 3 and 5 item conditions ( 3 AFC , 5AFC), 2 or 4 images of known items were respectively presented alongside the novel item. Items were presented equally spaced in a horizontal row with location randomized across trials. In each case, the participant had to click on the correct referent before proceeding to the next trial. Novel word-referent pairs were presented at least once (in the case of OD none) and no more than three times (for most AFC versions) over the course of all training trials (referred to as Novel Mapping trials). The 3AFC and 5AFC conditions also included filler/catch trials in which participants were asked to select one of the known items (referred to as Known Mapping trials). These were included both as catch trials and to draw the participants attention to the competitors in order to increase encoding of them. There were between 12 and 16 total mapping trials and each participant was trained on 4-5 novel words.

Immediately following the mapping phase, participants were tested for their comprehension and retention of the novel word-referent pairs. This was done via two methods free response recall and forces-choice (multiple choice) recognition. The free response retention trials asked participants to describe the referent for each of the novel words from mapping/training (e.g. "Describe the cheem"). Though prior work has asked participants to recall the wordform (thus testing phonological memory), here we ask for a description of the item to tap semantic-conceptual memory and thus allow a tighter connection to what is tested in the forced-choice trials. In addition, one current hypothesis is the semantic relationship (not names) between stimuli matters. In the forced-choice trials, participants were given one of the novel words from mapping and asked to select the correct referent from an array of all multiple novel objects from mapping. Words and objects for all phases were counter-balanced across conditions and participants.

## Results

Average percent correct for each participant on Known Mapping trials, Novel Mapping trials, Forced Choice (recognition) Retention, and Free Recall Retention were scored. Both Mapping trials and the Forced Choice Retention were calculated as percent of trials an individual correctly chose the target. Free Recall response was scored by 2-3 independent, blind coders who calculated a binomial score for each answer. If the written description was specific enough to uniquely identify the target object from the array of novels, it received a score of one. However, if the description was either too vague to refer to a single specific item or if it described a foil object, it received a score of zero. This coding scheme thus gave participants credit for a variety of responses (e.g. an overall description or a single unique feature) as it eliminated possible referents systematically based on the information given. For instance, if the participant responded "the blue flat thing", and there was only one novel item presented that was blue (and it was the correct answer), they would receive a one. If, however, a participant responded with "the round one", and multiple novel objects were spherical, then they would receive a score of zero. Only scores of one were counted as correct in calculating overall performance. All coding was conducted as a consensus of scores across two blind coders with additional 1-2 coders settling all discrepancies. Responses that required more than four coders to come to a consensus were thrown out.

First, performance on mapping across all conditions was analyzed. Each condition was compared against chance and each trial type was analyzed with a two-way ANOVA of performance on Number of Competitors (3AFC vs. 5AFC) by Competitor Type (vary vs. same). Unsurprisingly, adults are very good at the Known item filler trials on all conditions, selecting the known target item at nearly ceiling levels (see Figure 1). As suspected, there was no difference between conditions on these trials (Known Mapping; $\mathrm{F}(1,149)=.806, p=.493, \mathfrak{\eta}^{2}=.016$ ). Adults were also well


Figure 1. Performance on mapping and retention
above chance in all conditions at selecting the novel target (Novel Mapping). However, there was a main effect Type of Competitors, $\mathrm{F}(1,178)=9.027, p=.003, \eta^{2}=.048$, with adults performing significantly better when the competing items varied than when competitors were from the same category. There was no effect of competitor amount. This suggests that despite a very robust lexicon and a clear ability to find the referent when asked, adults real-time mapping of novel words may be influenced by subtle changes in the type of items present, regardless of the number of items.

Performance on retention was then analyzed. Most subjects were asked both a free recall retention question followed by a forced-choice recognition test. Due to a programming error, 29 subjects in the 3AFC same condition only answered the forced-choice test and 24 subjects in the 3 varies condition were asked the forced-choice first followed by free recall. ${ }^{1}$ In addition to testing performance against chance, a repeated measures ANOVA of Retention Test Type (forced-choice vs. free recall) on Number of Competitors (3AFC vs. 5AFC) by Competitor Type (vary vs. same) was run. First, adults were well above chance in all conditions on the forced-choice retention, though importantly, not at ceiling. Chance in both cases was set at $20 \%$ as presumably, adults were recalling one of the five just-learned items. On free recall, however, performance was much more mixed with performance in the vary conditions at or near chance, 3AFC: $\mathfrak{t}(45)=1.401, p=.168$; 5AFC: $\mathrm{t}(28)=2.042, p=.051$. In addition, there was a significant ANOVA of Retention for Type of Retention, Number of Competitors and Competitor Type, $\mathrm{F}(1,216)=5.865, \quad p=.016, \quad \eta^{2}=.026$, with a significant Competitor Type main effect, $\mathrm{F}(1,216)=13.286, p<.001$, $\mathrm{y}^{2}=.058$, and Test Type by Competitor Type interaction, $\mathrm{F}(1,199)=5.655, p=.018, \mathrm{n}^{2}=.028$. Thus, the variability of the context matters for mapping and has cascading effects on retention, especially when a more rigorous test of recall is used.

[^377]
## Discussion

Overall, the results show a unique pattern of responding for both in-the-moment mapping as well as retention. Though the type of competitors appears to matter overall for Novel Mapping, its effect on retention is confounded by the type of test, showing differential affects for forced-choice vs. free response retention. This could mean that the processes that support real time processing may not be the same mechanisms driving retention. Alternatively, it could suggest that associative learning is a complicated interaction of multiple factors and tasks and there are different ways of tapping knowledge to reveal those processes in different ways. That is, a free response retention test has no other items present. Thus, in order to respond, an individual has to have a robust memory to retrieve, but also may grasp onto any cues they can, such as vague memories from when they initially learned the word (Coutanche \& Thompson-Schill, 2014). Initial context that had more helpful cues (similar, same competitor condition) would give more relevant aids to retrieval, thus boosting performance.

Not only does the type of test matter, but subtle shifts in learning that are not always apparent with one test can be revealed with another. Specifically, the different retention tests give hints to the strength of the word-referent associations formed. Forced choice retention tests show only a slight effect of competitor variability, however, free recall for those exact same items is influenced in a different way by the variability of competitors from the previous mapping trials - variable competitors are equivalent to the Ostensive Naming, zero competitors condition on forcedchoice retention, (though lower than same competitor condition) but at chance when tested with free recall (at which point all other conditions are above chance).

These results also speak to the nature of the underlying association. During mapping, the initial association is weak, fragile, and likely has many spurious connections. More competitors, especially competitors that are vastly different and unique, take longer to process, and thus initial mapping is less successful. In some prior cases, these "harder" initial encoding scenarios have led to more robust learning because they require more processing (Vlach \& Sandhofer, 2014). If we only test learning through a forced choice recognition test, that is precisely what we see - above chance learning across the board. However, the free response retention reveals that the variable groups (the same groups who were relatively poorer on Novel Mapping) are at chance for recalling the word-referent link. This suggests that the wordreferent link that is being built during mapping has clearly been encoded and withstands weak tests of its viability, but has a long way to go before it is fully integrated into the lexicon.

At its core, word learning is a form of associative learning. As such, the results of the current study may suggest that domain-general associative mechanisms are influenced by the context and competition. Importantly, associations are constantly in flux and robust links are not guaranteed, even in supposedly "easy" cases of learning.

That means that the imperfect retention seen in the current study is not evidence for a lack of learning, but rather as an in-progress process which will continue to be strengthened over time. The type of competitive mechanism employed in these associative learning situations is not just beneficial for a single moment as demonstrated here, but likely also beneficial for the $2^{\text {nd }}, 3^{\text {rd }}, 4^{\text {th }}$ and all future encounters of the word (Benitez, Yurovsky, \& Smith, 2016; Dautriche \& Chemla, 2014; Yurovsky \& Yu, \& Smith, 2013). That is, learning is reliant not just on one-to-one links, but rather is couched within a larger network which has a direct influence on the development of a single association, which in turn may alter learning future associations.

As a whole, the current study examines two critical elements to associative learning -1 ) how does the amount and variability of context shape real-time processing, and 2) how does the amount and variability of the competitors shape the refinement of those associations over time. Results suggest that there is not a single pathway to mapping or retaining, but rather, like most cognitive processes, it is a complicated interaction.

## Acknowledgments

Funding to SCK by a Callier Postdoctoral Research Fellowship from the University of Texas at Dallas. Special thanks to the Oshkosh Development Lab for help coding.

## References

Adelman, J.S., Brown, G.D.A., \& Quesada, J.F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. Psychological Science, 17(9), 814-823.
Akhtar, N., Carpenter, M., \& Tomasello, M. (1996). The role of discourse novelty in early word learning. Child Development, 67, 635-645.
Alt, M. \& Gutmann, M.L. (2009). Fast mapping semantic features: Performance of adults with normal language, history of disorders of spoken and written language, and attention deficit hyperactivity disorder on a word learning task. Journal of Communication Disorders, 42(5), 347364.

Axelsson, E.L., Churchley, K., \& Horst, J.S. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. Frontiers in Psychology, 3(88).
Benitez, V.L., Yurovsky, D., \& Smith, L.B. (2016). Competition between multiple words for a referent in cross-situational word learning. Journal of Memory and Language, 90, 31-48.
Bion, R. A. H., Borovsky, A., \& Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel wordobject mappings in relation to vocabulary learning at 18, 24, and 30 months. Cognition, 126(1), 39-53.
Carey, S., \& Bartlett, E. (1978). Acquiring a single new word. Papers and Reports on Child Language Development, 15, 17-29.

Borovsky, A., Ellis, E. M., Evans, J. L., \& Elman, J. L. (2016). Lexical leverage: Category knowledge boosts real-time novel word recognition in 2-year-olds. Developmental Science, 19(6), 918-932.
Capone, N.C., \& McGregor, K.K. (2005). The effect of semantic representation on toddlers' word retrieval. Journal of Speech, Language, and Hearing Research, 48, 1468-1480.
Carey, S. \& Bartlett, E. (1978). Acquiring a single new word. Papers and Reports on Child Language Development, 15, 17-29.
Coutanche, M.N., \& Thompson-Schill, S.L. (2014). Fast mapping rapidly integrates information into existing memory networks. Journal of Experimental Psychology: General, 143(6), 2296-2303.
Dautriche, I., \& Chemla, E. (2014). Cross situational word learning in the right situations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(3), 892-903.
Fitneva, S.A., \& Christiansen, M,H. (2011) Looking in the wrong direction correlates with more accurate word learning. Cognitive Science, 35, 367-380.
Greve, A., Cooper, E., \& Henson, R.N. (2014). No evidence that 'fast-mapping' benefits novel learning in healthy older adults. Neuropsychologia, 60, 52-59.
Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. Cognitive Psychology, 53, 310-344.
Hills, T.T., Maouene, J., Riordan, B., \& Smith, L.B. (2014). The associative structure of language: Contextual diversity in early word learning. Journal of Memory and Language, 63, 259-273.
Horst, J.S. (2013). Context and repetition in word learning. Frontiers in Psychology, 4, 1-11.
Horst, J.S., \& Hout, M.C. (2016). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental research. Behavioral Research Methods, 48(4), 1293-1409.
Horst, J.S., Scott, E.J., \& Pollard, J.A. (2009). The role of competition in word learning via referent selection. Developmental Science, 13, 706-713.
Kucker, S. C., McMurray, B., \& Samuelson, L. K. (2015). Slowing Down Fast Mapping: Redefining the Dynamics of Word Learning. Child Development Perspectives, 9(2), 74-78
Markman, E. M., \& Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20(2), 121-157.
McGregor, K.K. (2014). What a difference a day makes: Change in memory for newly learned word forms over 24 hours. Journal of Speech, Language, and Hearing Research, 57(5), 1842-1850.
McMurray, B., Horst, J.S., \& Samuelson, L.K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological Review, 119(4), 831-877.

McMurray, B., Zhao, L., Kucker, S. C., \& Samuelson, L. K. (2013). Pushing the envelope of associative learning: Internal representations and dynamic competition transform association into development. In L. Gogate \& G. Hollich (Eds.) Theoretical and computational models of word learning: Trends in psychology and artificial intelligence. (pp.49-80). Hershey, PA: IGI Global.
Meints, K., Plunkett, K., Harris, P.L., \& Dimmock, D. (2004). The cow on the high street: Effects of background context on early naming. Cognitive Development, 19(3), 275-290.
Mervis, C. B., \& Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. Child Development, 65(6), 1646-1662.
Morse, A.F., Benitez, V.L., Belpaeme, T., Cangelosi, A., \& Smith, L.B. (2015). Posture affects how robots and infants map words to objects. PLoS One, 10(3), 1-17. https://doi.org/10.1371/journal.pone. 0116012
Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. Cognitive Science, 29, 819-865.
Roembke, T.C., Wasserman, E., \& McMurray, B. (2016). Learning in rich networks involves both positive and negative associations. Journal of Experimental Psychology General, 145(8), 1062-1074.
Tamminen, J. \& Gaskell, M.G. (2008). Newly learned spoken words show long-term lexical competition effects. Quarterly Journal of Experimental Psychology, 61(3), 361-371.
Twomey, K., Morse, A., Cangelosi, A., \& Horst, J. (2016). Children's referent selection and word learning: Insights from a developmental robotic system. Interaction Studies, 17(1), 101-127.
Vlach, H. A., \& Sandhofer, C. M. (2012). Fast mapping across time: Memory processes support children's retention of learned words. Frontiers in Psychology, 3. http://doi.org/10.3389/fpsyg.2012.00046
Vlach, H.A., \& Sandhofer, C.M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. Cognitive Science, 38, 757-774.
Warren, D.E., \& Duff, M.C. (2014). Not so fast: Hippocampal amnesia slows word learning despite successful fast mapping. Hippocampus, 24, 920-933.
Wojcik, E. H. (2013). Remembering new words: Integrating early memory development into word learning. Frontiers in Psychology,4. http://doi.org/10.3389/fpsyg.2013.00151
Yurovsky, D., Yu, C., \& Smith, L.B. (2013). Competitive processes in cross-situational word learning. Cognitive Science, 37(5), 891-921.
Zosh, J.M., Brinster, M., \& Halberda, J. (2013). Optimal contrast: Competition between two referents improves word learning. Applied Developmental Science, 17(1), 2028.

# Geometric Concept Acquisition in a Dueling Deep Q-Network 

Alex Kuefler (akuefler@stanford.edu)<br>Symbolic Systems Program

Mykel J. Kochenderfer (mykel@stanford.edu)<br>Department of Aeronautics and Astronautics

James L. McClelland (jlmcc@stanford.edu)<br>Department of Psychology<br>Stanford University, Stanford, CA 94305 USA


#### Abstract

Explaining how intelligent systems come to embody knowledge of deductive concepts through inductive learning is a fundamental challenge of both cognitive science and artificial intelligence. We address this challenge by exploring how a deep reinforcement learning agent, occupying a setting similar to those encountered by early-stage mathematical concept learners, comes to represent ideas such as rotation and translation. We first train a Dueling Deep Q-Network on a shape sorting task requiring implicit knowledge of geometric properties, then we query this network with classification and preference selection tasks. We demonstrate that scalar reinforcement provides sufficient signal to learn representations of shape categories. After training, the model shows a preference for more symmetric shapes, which it can sort more quickly than less symmetric shapes, supporting the view symmetry preferences may be acquired from goal-directed experience.


## Introduction

Mathematical concepts are formally definable and may be deduced from of axioms. In contrast, most human mental representations, such as visual categories and language concepts, resist precise definitions and only come to be known after considerable inductive experience. Similarly, deep neural networks have achieved human-level or state-of-the-art performance on tasks such as object recognition (He, Zhang, Ren, \& Sun, 2016), game playing (Mnih et al., 2015; Silver et al., 2016), and speech generation (van den Oord et al., 2016) by learning distributed (rather than symbolic) representations through inductive (rather than deductive) training. The empirical success of human learners and artificial neural networks contrasts sharply with the description of mathematical concepts as abstract, formal, and universal ideals.

A growing literature argues that the developmental details of embodied agents are not mere nuisance variables associated with the learning and deployment of mathematical concepts, but are necessary tools in facilitating cognition. For instance, some work suggests that the use of hand gestures aids learning by grounding the meaning of abstract principles such as continuity and magnitude in the willful motions of the body (Goldin-Meadow, Cook, \& Mitchell, 2009; Marghetis \& Núnez, 2013; Marghetis, Núñez, \& Bergen, 2014). The tuning of low-level responses of the visual system has also been associated with algebraic expertise. Marghetis, Landy, and Goldstone (2016) argue that a process of "regimented perception", implemented by object-based attention,


Figure 1: The top panels show example frames from the shape sorting environment. The blue cursor indicates grabbing, and the green cursor is not grabbing. The bottom panel shows examples of each shape in isolation: Hexagon (Hex.), Equilateral Triangle (E. Tri.), Trapezoid (Trap.), Square, and Right Triangle (R. Tri.).
lets viewers parse algebraic notation in such a way that makes salient its hierarchical structure, thus offloading much of the work of calculation from high-level cognition to perception. Taken together, this work suggests that for all its formality, mathematics is still an evolutionarily recent phenomenon, and in its comprehension, one marshals bodily and neural resources adapted for other purposes.

This account dovetails nicely with the Parallel Distributed Processing approach (Rogers \& McClelland, 2014), or more recently, advances in deep learning. Just as embodied cognition challenges the primacy of symbolic representations in mental processes, deep learning has been used to overcome a number of problems once thought solvable only through the manipulation of compositional tokens or formal logic. For example, the "Differentiable Neural Computer" of Graves et al. (2016) can answer queries involving textual and hierarchical reasoning, relying solely on a learned, neural memory device. Furthermore, these models learn "end-to-end", adjusting connection strengths between stimulus-facing neurons based on error signals propagated down from higher level de-
cisions (Mnih et al., 2015), much as Marghetis et al. (2016) argue that pursuing algebraic skills re-tunes the visual system. Despite these connections to the embodied approach in mathematical cognition, we are not aware of major work exploring how deep neural networks capture mathematical, and in particular geometric intuitions through a goal-driven, learning process of perception and actuation.

This paper provides a proof of concept and an exploratory analysis. We demonstrate that a domain-agnostic learning algorithm is able to represent geometric concepts that are only implicitly coded in its training task's structure. In particular, we develop a simulated shape sorting toy similar to those enjoyed by young children. Such a learning environment tasks an agent to carry out sequences of actions that require knowledge about some geometric properties (whether implicit or explicit) to maximize reward. We train an agent to expertise in this task, then evaluate its learned behavior and representation of shape categories.

Our primary contribution has been to show that reinforcement learning (RL) is sufficient to train a convolutional neural network agent to perform a shape sorting task expertly. This training leads to the development of shape-specific representations at the top convolutional layer, which feeds into network layers that compute value. We also found that our model exhibited a preference for objects with higher orders of symmetry, supporting the view that experience, rather than an innate symmetry bias, may be the basis for similar similarity preferences in both human infant and primate studies (Bornstein, Ferdinandsen, Gross, 1981; McMahon Olson, 2007). We also make available the source code for the pixels-to-actions shape-sorting task described in this paper.

## Approach

Our goal is to demonstrate that deep learning may provide a computational paradigm for building on psychological theory and generating new hypotheses about geometric concept acquisition. Blocks worlds have previously been used to study and model intuitive physics (Hamrick, Battaglia, \& Tenenbaum, 2011; Zhang, Wu, Zhang, Freeman, \& Tenenbaum, 2016). Although such environments feature a finite set of discrete entities adhering to rules of interaction, their broad properties and questions of investigation tend to differ. For instance, these environments simulate physical properties, like velocity. In contrast, we seek to understand abstract properties, such as shape and symmetry, and we thus introduce an environment with few physical constraints. As such, the experimental setup can be decomposed into an environment and learning agent, which we represent with a neural network.

## Environment

During the initial training stage of our neural network, the model interacts with a simulated shape sorting toy (see Figure 1), which may be interpreted as a finite horizon Markov Decision Process (MDP) with deterministic transitions and high-dimensional states (Kochenderfer \& Reynolds, 2015).

Environment interactions are divided into trials, which consist of at most 500 timesteps. At timestep $t$, the environment emits an image $s_{t} \in \mathbb{R}^{84 \times 84}$ depicting some combination of two types of objects: blocks and holes. Every object in $s_{t}$ is characterized by an orientation and a position vector, which remain fixed for holes, but are subject to change for blocks. Each object is also characterized by a convex, 2D shape drawn from the set $X$, which includes squares, trapezoids, equilateral triangles, right triangles, and hexagons. Once a shape is chosen for an object, it is held constant throughout the course of a trial. The image $s_{t}$ also includes a cursor used by an agent to manipulate the position and orientation of blocks.

The initial frame $s_{0}$ includes three blocks whose shape assignments $X_{b}$ are drawn uniformly with replacement from $X$ with randomized positions and orientations. Four holes with random orientations are also given shape assignments $X_{h}$ drawn without replacement from $X$. A constraint $X_{b} \subset X_{h}$ ensures that no block will be generated without a corresponding hole. The positions for holes are also randomized at the beginning of each trial, but only drawn from four possible, equidistant locations. This second constraint ensures that holes never overlap. In contrast, blocks may overlap (sometimes completely) but are manipulable, and can be disentangled by an agent.

Given $s_{t}$, an agent responds with action $a_{t}$ in $\mathcal{A}$, which includes: up, down, left, right movements, toggle grab, rotate clockwise or counterclockwise, or wait. Rotations are $30^{\circ}$ and a single cardinal movement covers $10 \%$ of the height or width of $s_{t}$. If the grab is active, blocks will "stick" to the cursor, changing position and orientation with the cursor. If the grab is inactive, movement and rotation actions do not influence the blocks.

The environment uses a reward function that assigns a small reward when the cursor grabs a block $r_{1}=0.001$, a small penalty $-r_{1}$ when the cursor contacts the border of the screen, a large reward $r_{2}=1.0$ when the cursor fits a block to a corresponding hole or completes a trial. A fit occurs when the cursor "releases" a block over a hole, and the block's vertex set is contained by the hole's vertex set. If a fit occurs, the block disappears and will not return for the rest of the trial.

## Agent

We desire a model of a learning agent that is (1) Deep, or capable of expressing multiple, hierarchical representations that could feasibly embody geometric invariants, given raw pixels, (2) Psychologically plausible, or sufficiently similar to animal decision making to suggest research directions for cognitive science, and (3) Powerful enough to solve the nontrivial MDP described in the previous section. The Deep QNetwork (DQN) (Mnih et al., 2015) is an attractive option that can accommodate these considerations.

DQN has attained state-of-the-art results on similar tasks that include discrete action spaces and high-dimensional state spaces. Sharing many architectural properties with convolutional neural networks, it learns a succession of hidden rep-


Figure 2: The Dueling Deep-Q Network architecture and the dimensionality of each layer. Boxed layers are involved in representation, output layers are involved in actuation. Dotted lines denote scalar-vector broadcasting that merge value and advantage streams. We use the same filter sizes and strides as Wang et al. (2016)
resentations that can be visualized and interpreted as an abstraction hierarchy (Zeiler \& Fergus, 2014). Furthermore, DQN features some desiderata as a model of animal decision making. A prioritized replay pool has been compared to hippocampal learning mechanisms (McClelland, McNaughton, \& O'Reilly, 1995), and the architecture is trained using temporal difference (TD) learning, which has been shown to underpin some forms of animal learning (Shah, 2012).

TD learning is here accomplished by minimizing the adaptive loss function

$$
\begin{equation*}
L_{i}\left(\theta_{i}\right)=\mathbb{E}_{\left(s, a, r, s^{\prime}\right) \sim D}\left[\left(y_{i}-Q\left(s, a ; \theta_{i}\right)\right)^{2}\right] \tag{1}
\end{equation*}
$$

with target value

$$
\begin{equation*}
y_{i}=r+\underset{a^{\prime}}{\gamma \max ^{\prime}} Q\left(s^{\prime}, a^{\prime} ; \theta_{i}^{-}\right) \tag{2}
\end{equation*}
$$

State-action-reward sequences ( $s, a, r, s^{\prime}$ ) observed during environment interactions are drawn from a replay buffer $D$ and used as training samples. $Q\left(s, a ; \theta_{i}\right)$ represents the sum of discounted future rewards if action $a$ is taken from state $s$ and is estimated at epoch $i$ by a DQN parameterized by $\theta_{i}$. Updating the model using estimates from a target network parameterized by $\theta_{i}{ }^{-}$has been shown to improve the stability of training (Mnih et al., 2015; Wang et al., 2016). A policy can then be induced from the DQN by selecting actions maximizing $Q\left(s, a ; \theta_{i}\right)$ with probability $\varepsilon$ and otherwise selecting exploratory, random actions.

In this work, $Q\left(s, a ; \theta_{i}\right)$ is represented by a Dueling Deep Q-Network (DDQN) which is subject to the same TD learning paradigm as DQN, but features a different architecture (Wang et al., 2016), shown in Figure 2. ${ }^{1}$ DDQN follows its convolutional layers with two disjoint, fully-connected streams that represent the scalar value of a state $V(s)$ and the advantage $A(s, a)=Q(s, a)-V(s)$ of a state-action pair

[^378]separately. These representations are merged with the broadcasting rule
\[

$$
\begin{array}{r}
Q(s, a ; \theta, \alpha, \beta)= \\
V(s ; \theta, \beta)+A(s, a ; \theta, \alpha)-\frac{1}{|\mathcal{A}|} \sum_{a^{\prime}} A\left(s, a^{\prime} ; \theta, \alpha\right) \tag{3}
\end{array}
$$
\]

where $\theta, \alpha$, and $\beta$ parameterize the convolutional, advantage, and value sub-networks respectively. DDQN has been shown to improve the state-of-the-art beyond the performance of DQN and has some favorable properties as a neurobiological model, as it extends to deep neural networks the advantage learning paradigm, which has been shown to correlate with striatal neural activity during instrumental learning tasks (O’Doherty et al., 2004).

## Learned Behavior

We trained our agent to complete the shape sorting task over the course of one week on a single GeForce GTX 980 graphics processing unit. It completed approximately 480,000 trials consisting of at most 500 timesteps each. After training, the agent completed two experimental tasks consisting of 50,000 trials. Although the agent trained using $\varepsilon$-greedy exploration with $\varepsilon=0.1$, testing tasks were performed using a pure greedy policy. On both experimental tasks, we found that the agent adopted the strategy of performing translation actions early in the trial, followed by rotation actions once the block was in place over the correct hole.

Single Block Performance. One block per trial was drawn from $X_{b}$ and initialized at a random position with four holes from which to choose. The agent's cursor was initialized at the center of the screen. Each trial ended when the agent fit the block to the appropriate hole, attempted to fit the block to an incorrect hole, or exceeded 500 timesteps. Incorrect fits and time outs together accounted for less than $4 \%$ of the total number of trials, with the vast majority of failed trials resulting from time outs.

Table 1 demonstrates the agent's efficiency at the task on the trials it successfully completed, in addition to the estimated value computed by the network upon first grabbing a block. Although the agent performed nearly optimally on all shapes, we found that the network assigned higher estimated value for shapes with greater symmetry (which also corresponds the minimum number of steps needed to fit the block to a hole). However, although the right triangle and trapezoid share the same symmetry order, the trapezoid is assigned slightly higher value.

Shape Preference. Observing that our model estimated higher value for some shapes over others, we tested to see whether the agent demonstrated preferences in a twoalternative forced choice. In this experiment, two blocks belonging to different shape categories were generated and placed equidistantly to corresponding holes. Trials terminated when the agent selected a "winner" by releasing a held

Table 1: Agent's performance on single-block trials, including value estimated by layer Val., and average number of steps needed to complete the trial against average number of steps actually taken by the agent, per each shape category.

| Shape | Value | Min. Steps | Act. Steps | Ratio |
| :--- | :--- | :--- | :--- | :--- |
| Hex. | 0.70 | $10.74 \pm(3.7)$ | $11.83 \pm(4.0)$ | 0.91 |
| Square | 0.67 | $11.37 \pm(3.8)$ | $12.57 \pm(4.1)$ | 0.91 |
| E. Tri. | 0.64 | $11.76 \pm(3.9)$ | $12.90 \pm(4.1)$ | 0.91 |
| Trap. | 0.57 | $13.73 \pm(4.2)$ | $15.06 \pm(4.7)$ | 0.91 |
| R. Tri. | 0.55 | $13.66 \pm(4.2)$ | $15.38 \pm(4.8)$ | 0.89 |



Figure 3: Each cell displays the fraction of choices between two shapes in which the shape on the $x$-axis was chosen. Marginal probabilities of choosing a given shape are shown above.
block over a corresponding hole. We observed a slight bias such that the policy selected the block appearing on the right hand side of the screen $58.65 \%$ of the time, but controlled for this effect by randomizing block position each trial.

The results shown in Figure 3 accord with the findings from the single block experiment. Blocks appear to be preferred on the basis of the number of steps needed to achieve a fit, which is in turn determined by their symmetry order. However, the trapezoid is again preferred to the right triangle, despite the fact that both blocks are equally symmetric. We explore a possible explanation for this result in the next section.

## Learned Representations

To gain insight on the agent's elicited behavior, we treat the network as a feature extractor and use classification techniques to explore how it internally represents shape categories in different layers. Within the context of computational neuroscience, linear classifiers have been used to decode information about categorical stimuli from neural responses (Naselaris, Kay, Nishimoto, \& Gallant, 2011). We adopt a
similar approach. Intuitively, because neural networks are universal function approximators (Hornik, Stinchcombe, \& White, 1989), the activation vectors of a well-tuned network corresponding to different categories should be discriminable up to a linear transformation. As such, we assess the classification accuracy of a Support Vector Machine (SVM) trained on encodings from different layers of the DDQN.

Our SVM implementation comes from the open source library, scikit-learn (Buitinck et al., 2013) and makes use of a linear kernel $K\left(x, x^{\prime}\right)=x^{\top} x^{\prime}$. Multiway classification is achieved using a "one-versus-all" scheme, such that for $n$ classes, $n$ separate binary classifiers are trained to discriminate its corresponding class from examples belonging to other categories. The final classification is made by the model that makes its predictions with the largest margin.

## Dataset

One might argue that shape representations in the network depend heavily on scene context. For example, when a scene contains multiple blocks, it may not be useful to encode any information about shape identity until the cursor has taken hold of a single block, as only then must it make a decision contingent on the identity of the shape. To test this hypothesis, we also repeat the discrimination experiment for conditions involving an absent cursor, and conditions in which the cursor is visibly grabbing the block.

We generated our training and validation image sets by enumerating all the possible positions and orientations for a single block, subject to the environment's translation and rotation step sizes, and excluding duplicate shape orientations resulting from symmetry. Each combination was used to produce a set of 100 unique examples by randomly permuting the background holes. The resulting data set consisted of 81,000 images, which we shuffled and partitioned into training and validation sets using a $25-75$ split. The data sets including a cursor were generated by the same process.

Each frame $s_{i}$ was replicated four times, producing an $84 \times 84 \times 4$ tensor, which was then encoded as an activation vector $z_{i}{ }^{j}$ at the $j$ th layer of the DDQN. Because the size of the layers differ greatly, we use principal component analysis (PCA) to enforce that all activation vectors have 300 dimensions. To establish a classification baseline, we repeat the same analysis on encodings produced by an untrained, randomly initialized neural network with the same architecture. We do not standardize encodings prior to classification or dimensionality reduction, as all input variables are ReLU-gated activations and are thus measured on the same scale.

## Results

Results shown in Figure 4 support the view that shape information is scene dependent, albeit slightly. At the level of Conv3 and beyond, classification accuracy was consistently greater by about $3 \%$ when the cursor was grabbing the block. Figure 5 visualizes the activation vectors in both conditions, and suggests that the categories become better separated during a grab. Interestingly, these results contradict the view


Figure 4: Average validation accuracy of the SVM. The top plot shows accuracy per encoding layer, with baseline accuracy on encodings produced by a DDQN with random weights. The bottom plot shows classification confusion on the "Grabbing" condition at Conv3 in the trained network.
that shape information plays a major role in the directions of greatest variance in higher layers. Whereas the classifiers achieve above $70 \%$ accuracy on frames encoded by Conv3, upstream encodings from $A d v$. and Val., were discriminable only about $45 \%$ of the time, dropping even below Convl and Conv2.
The most significant misclassifications are shown to be between the equilateral triangle and trapezoid, whereas the most easily discriminable shapes were the hexagon and square. These misclassifications may explain the preference for trapezoid blocks over right triangles, despite their similar orders of symmetry, as the network seems to "confuse" trapezoids with the easier triangle shape, whereas the right triangle is easy to classify as a difficult block.

## Conclusions

Learning mechanisms and computational principles underlying mathematical cognition are not well understood. However, deep neural networks provide opportunities for exploring this direction of inquiry. We hypothesized that reinforcement learning, which incorporates active probing of an environment, serves as a sufficient training signal for learning many geometric properties embodied implicitly in an interactive shape sorting task.


Figure 5: A sample of 2,500 encodings extracted from Conv3. Linear Discriminant Analysis was used to project the 300dimensional vectors onto a subspace in which the classes are well-separated. The separation is more clear when the cursor is grabbing the block.

On a representational level, we showed that shape identity can be recovered from the network's hidden layers using linear classifiers and that this information is more strongly encoded in later convolutional layers than in the final hidden layers needed to valuate states and possible actions. Recent work suggests an analogy between the hierarchical structure of convolutional neural networks and the hierarchical structure of the visual system (Yamins et al., 2014). If this analogy is to be taken seriously, we should predict that despite their simplicity, geometric forms may find representation in later visual areas when tied to one's goals, as when playing with a shape sorter or interpreting mathematical diagrams on an exam.

On a behavioral level, we found also that a preference for symmetric blocks emerged as a consequence of their ease of fitting. Past work indicates that a preference for symmetric shapes exists among both monkeys (McMahon \& Olson, 2007) and human beings, but emerges only later in development (Bornstein, Ferdinandsen, \& Gross, 1981). Bornstein et al. (1981) in particular suggest that this preference, which favors vertical over horizonal symmetry, arises not from the information redundancy present in such figures, but from their adaptive value. Symmetrical figures tend to be animate, and can thus act as adversaries or allies in an organism's pursuit of goals. In our domain as well, we found that a block's degree of symmetry influences a learning agent's discounted sum of expected rewards. This result further supports the adaptive view of symmetry preference over the redundancy view, and implies a number of testable predictions for future work.

Follow up studies should investigate whether children trained to play with shape sorters prefer different blocks on the basis of their symmetry properties, and if so, if this preference can be modulated by increasing the stakes of the task, whether by providing greater rewards or less time to react. Further simulation work should also explore the relationship
between symmetry and visual similarity. Despite both being completely asymmetric, we found that trapezoids were strongly preferred to right triangles, as they are visually closer to equilateral triangles. An experiment in which "adversarial" shapes attempt to look easier to fit than they really are may demonstrate how the constraints of perception (imposed by the visual similarity between blocks) and the constraints of actuation (imposed by the reward signal, or task) must be mutually satisfied in embodied, geometric concept acquisition.

The code associated with this paper can be found at https://github.com/akuefler/shape-sorting.

## Acknowledgements

We would like to thank Steven Hansen for useful discussions.

## References

Bornstein, M. H., Ferdinandsen, K., \& Gross, C. G. (1981). Perception of symmetry in infancy. Developmental Psychology, 17(1), 82.
Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning (pp. 108122).

Goldin-Meadow, S., Cook, S. W., \& Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. Psychological Science, 20(3), 267-272.
Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626), 471-476.
Hamrick, J., Battaglia, P., \& Tenenbaum, J. B. (2011). Internal physics models guide probabilistic judgments about object dynamics. In Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 1545-1550).
He, K., Zhang, X., Ren, S., \& Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).
Hornik, K., Stinchcombe, M., \& White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359-366.
Kochenderfer, M. J., \& Reynolds, H. J. D. (2015). Decision making under uncertainty: Theory and application. MIT press.
Marghetis, T., Landy, D., \& Goldstone, R. L. (2016). Mastering algebra retrains the visual system to perceive hierarchical structure in equations. Cognitive Research: Principles and Implications, $1(1), 25$.
Marghetis, T., \& Núnez, R. (2013). The motion behind the symbols: A vital role for dynamism in the conceptualization of limits and continuity in expert mathematics. Topics in Cognitive Science, 5(2), 299-316.
Marghetis, T., Núñez, R., \& Bergen, B. K. (2014). Doing arithmetic by hand: Hand movements during exact
arithmetic reveal systematic, dynamic spatial processing. The Quarterly Journal of Experimental Psychology, 67(8), 1579-1596.
McClelland, J. L., McNaughton, B. L., \& O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102(3), 419.
McMahon, D. B., \& Olson, C. R. (2007). Repetition suppression in monkey inferotemporal cortex: relation to behavioral priming. Journal of Neurophysiology, 97(5), 35323543.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Humanlevel control through deep reinforcement learning. Nature, 518(7540), 529-533.
Naselaris, T., Kay, K. N., Nishimoto, S., \& Gallant, J. L. (2011). Encoding and decoding in fMRI. Neuroimage, 56(2), 400-410.
O’Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., \& Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science, 304(5669), 452-454.
Rogers, T. T., \& McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. Cognitive Science, 38(6), 1024-1077.
Shah, A. (2012). Psychological and neuroscientific connections with reinforcement learning. In M. Wiering \& M. van Otterlo (Eds.), Reinforcement learning: State-of-the-art (pp. 507-537). Springer.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.
van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. CoRR abs/1609.03499.
Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lactot, M., \& de Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. In International Conference on Machine Learning (ICML) (p. 1995-2003).
Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., \& DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences, 111(23), 8619-8624.
Zeiler, M. D., \& Fergus, R. (2014). Visualizing and understanding convolutional networks. In European Conference on Computer Vision (pp. 818-833).
Zhang, R., Wu, J., Zhang, C., Freeman, W. T., \& Tenenbaum, J. B. (2016). A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. arXiv preprint arXiv:1605.01138.

# Perceived Difficulty of Moral Dilemmas Depends on Their Causal Structure: A Formal Model and Preliminary Results 

Barbara Kuhnert ${ }^{1}$ (kuhnertb@informatik.uni-freiburg.de), Felix Lindner ${ }^{2}$ (lindner@informatik.uni-freiburg.de) Martin Mose Bentzen ${ }^{3}$ (mmbe@dtu.dk), Marco Ragni ${ }^{1,2}$ (ragni@informatik.uni-freiburg.de)<br>${ }^{1}$ Cognitive Computation Lab, University of Freiburg, Freiburg im Breisgau, Germany<br>${ }^{2}$ Foundations of Artificial Intelligence, University of Freiburg, Freiburg im Breisgau, Germany<br>${ }^{3}$ Management Engineering, Danish Technical University, Lyngby, Denmark


#### Abstract

We propose causal agency models for representing and reasoning about ethical dilemmas. We find that ethical dilemmas, although they appear similar on the surface, differ in their formal structure. Based on their structural properties, as identified by the causal agency models, we cluster a set of dilemmas in Type 1 and Type 2 dilemmas. We observe that for Type 1 dilemmas but not for Type 2 dilemmas a utilitarian action does not dominate the possibility of refraining from action thereby constituting a conflict. Hence, we hypothesize, based on the model, that Type 1 dilemmas are perceived as more difficult than Type 2 dilemmas by human reasoners. A behavioral study where participants rated the difficulty of dilemmas supports the models' predictions.


Keywords: Moral Reasoning; Moral Complexity; Moral Dilemmas; Causal Agency Models; Ethical Principles

## Introduction

Currently, we experience a hot debate on moral reasoning and artificial intelligence (AI). In one respect, the discussion is about how to apply AI technology morally. In another respect, there is a requirement to enable AI technology itself to make moral decisions. Fields of application are self-driving cars (Bonnefon, Shariff, \& Rahwan, 2016), robots navigating in social environments (Lindner, 2015), and even robots that give moral advice (Lindner \& Bentzen, 2017). As a consequence, new research areas such as machine ethics (Allen, Wallach, \& Smit, 2006) and moral human-robot interaction (Malle, Scheutz, Arnold, Voiklis, \& Cusimano, 2015) arise.

To address the requirement for autonomous moral decision making, we recently introduced a software library for modeling hybrid ethical reasoning agents (short: HERA) ${ }^{1}$ (Lindner \& Bentzen, 2017). The goal of the HERA project is to provide theoretically well-founded and practically usable logicbased machine ethics tools for implementation in artificial moral agents such as (social) robots and software bots. To align human moral reasoning with moral reasoning by machines, our development of formal models and algorithms is informed by moral psychology and moral philosophy. We aim for the integration of various theories about human moral development, moral reasoning, and ethics.

There are several approaches to explain human moral reasoning. Kohlberg (1984), whose approach is based on Piaget's "genetic epistemology" claimed that individuals are passing through six invariant and universal stages in the development of moral reasoning. Reaching the next stage rep-

[^379]resents a qualitative advance in the ability to make consistent and differentiated judgments concerning moral norms and principles. Conversely, the theory of moral reasoning advocated by Mikhail (2007) assumes that there is moral grammar triggering certain moral judgments. He hypothesizes two rules for the grammar: the norm prohibiting intentional battery as a means, and the norm of double effect valuating battery as side effect. The research by Greene, Sommerville, Nystrom, Darley, and Cohen (2001); Haidt (2001) claims a prevalence of emotionally based moral intuition. Greene and Haidt (2002) are moving away from moral reasoning tending towards moral judgments caused by immediate affective intuitions and emotions. Greene et al. (2001) advanced the dual process model of moral judgment. They assume competitive moral subsystems in the brain resp. moral reasoning that is influenced by the mutual interaction and competition between two distinct psychological systems: (1) the emotional, intuitive, deontological judgment system and (2) the rational, calculated, utilitarian judgment system.

Throughout the literature, various hypothetical moral dilemmas are used to investigate questions concerning human morality, moral reasoning and moral judgments. We will make use of four dilemmas:

1. Runaway Trolley Dilemma A runaway trolley is about to run over and kill five people. If a bystander throws a switch then the trolley will turn onto a sidetrack, where it will kill only one person.
2. Pregnancy Dilemma A pregnant woman is about to give birth to her triplets. If the doctors treat the woman then her triplets will live, but she will die. Otherwise, the triplets will die, but the life of the pregnant women will be saved.
3. Boat Dilemma A boat is about to sink because of overweight. If the crew is told to throw the biggest person into the sea then the boat will not sink and the other three passengers will be saved (but the big person will die).
4. Hijacked Airplane Dilemma An airplane was hijacked by terrorists, and the terrorists threaten to crash the airplane against a populated area on the ground. If the military shoots the airplane the passengers will die but the airplane will crash in a deserted area thus not harming anyone else.

Several ways of classifying dilemmas and different moral reasoners have been proposed: Greene, Nystrom, Engell,

Darley, and Cohen (2004) differentiate between personal dilemmas and impersonal dilemmas. Among subjects, commonly more deontological judgments are produced with personal dilemmas, while impersonal dilemmas commonly produce consequentialist judgments (Moll \& de Oliveira-Souza, 2007). Crockett (2013) proposes a model-based system for consequentialist reasoning: The reasoner evaluates the best outcome of an action by starting from the current action and searching through a decision tree. In the model-free evaluation, which is associated with deontological reasoning, the forward searching is not activated. Shou and Song (2015) found that most of their subjects, irrespective of whether they chose a deontological decision or a consequentialist decision, evaluated consequences when information about outcome probabilities was provided. Wiegmann and Waldmann (2014) propose that the moral dilemmas' underlying causal structure supports moral intuition and thus is an important factor for the moral judgments humans make.

Thus, we observe that research has been focused on the effects of content and structure of moral dilemmas on human moral judgments. Our research focus is on moral complexity and adds evidence for how structural properties of moral dilemmas affect their perceived difficulty. The paper is structured as follows: First, we introduce causal agency models as a tool for representing moral dilemmas in terms of causes and utilities. Second, we define ethical principles within this framework. Third, the four aforementioned moral dilemmas are modeled using causal agency models. Based on structural commonalities and differences of these models, we distinguish two dilemma types, which we term Type 1 and Type 2 dilemmas. We hypothesize that Type 1 dilemmas are more difficult to solve for humans than Type 2 dilemmas. Fourth, we present an empirical study which shows that our model predicts human ratings about the perceived difficulty for the two types of moral dilemmas.

## Causal Agency Models

Ethical principles can be modeled as specifications of moral permissibility in causal agency models. Causal agency models are extensions of causal models that are used for counterfactual reasoning about causality, responsibility, blame, and related concepts (Halpern, 2016). In our HERA framework, an ethical principle is represented as a logical formula whose truth determines which actions are permissible according to the principle and which are not. Actions and their consequences are modeled as directed acyclic graphs showing causal influence. At the root of the graph will be actions and other independent variables influencing consequences further down the graph. Boolean structural equations capture all the information about the causal relationship between variables. For instance, to model that the trolley from the Runway Trolley Dilemma will turn onto a sidetrack when the bystander throws the switch, we may write the boolean structural equation turn $:=$ throw. The boolean variable turn will be true in the model whenever the boolean variable throw is true in the
model. The set of boolean structural equations in a model is called a causal mechanism. The truth assignment of the root node of the graph is called a world or an option. Formally, we define causal agency models as follows:

## Definition 1 (Causal Agency Models)

A (boolean) causal agency model, $M$, is a tuple $\langle U=A \cup$ $B, C, F, u, W\rangle$, where, $A=\left\{a_{1}, \ldots, a_{m}\right\}$ is a nonempty finite set of propositional variables called the actions. $B=\left\{b_{1}, \ldots, b_{k}\right\}$ is a (possibly empty) finite set of propositional variables called the background variables. Together the actions and background variables are the exogenous variables as defined above. $C=\left\{c_{1}, \ldots, c_{n}\right\}$ is a finite (possibly empty) set of propositional variables called the endogenous variables. $F$ is a causal mechanism explained above. $u:$ literals $\rightarrow \mathbb{Z}$ is a utility function assigning an integer value to each literal. W is a set of boolean interpretations of $(A \cup B)$.

We assume some familiarity on the part of the reader with classical propositional logic (and $(\wedge)$, or $(\vee)$, not $(\neg)$, and so on) and of truth functional semantics. A formula containing variables such as $\left(c_{1} \wedge a_{1}\right)$, is intended to mean that consequence $c_{1}$ and action $a_{1}$ both obtain. We write $M, w_{i}=$ $\left(c_{1} \wedge a_{1}\right)$ for $\left(c_{1} \wedge a_{1}\right)$ is true with option $w_{i}$ (or at world $\left.w_{i}\right)$ in the model $M$. Apart from propositional formulas we need simple arithmetic formulas expressing the utility of literals. We write $u\left(v_{i}\right)=z$, for an integer $z$, with the intended meaning that the utility of $v_{i}$ is $z$, similarly we write $u\left(v_{i}\right) \geq u\left(v_{j}\right)$ for the utility of $v_{i}$ is equal to or greater than the utility of $v_{j}$. We extend the utility function to conjunctions of literals by addition of the utilities of the conjuncts. The utility of other formulas (e.g., disjunctions) is undefined.

## Ethical Principles

Causal agency models play the role of representations of situations involving moral decisions. We now define ethical principles according to which moral permissibility of actions can be assessed based on the actions' consequences. For the following discussion, the principle of act-utilitarianism and the notion of Pareto dominance are of particular importance.

The utilitarian principle focuses on consequences of actions. It states that an agent ought to perform the action among the available alternatives with the overall maximal utility. We adopt an act-utilitarian interpretation which does not distinguish between doing and allowing, i.e., the causal structure of the situation is not taken into account. Thus the action which the agent ought to perform is the one which leads to the best possible situation, i.e., the highest utility, regardless of what the agent causes and intends.

## Definition 2 (Utilitarian Permissibility)

Let $w_{0}, \ldots, w_{n}$ be the available options, and $\operatorname{cons}_{w_{i}}=$ $\left\{c \mid M, w_{i} \models c\right\}$ be the set of consequences and their negations that obtain with these options. An option $w_{p}$ is permissible according to the utilitarian principle if and only if none of its alternatives yield more overall utility, i.e., $M, w_{i}=$ $u\left(\bigwedge \operatorname{cons}_{w_{p}}\right) \geq u\left(\bigwedge\right.$ cons $\left._{w_{i}}\right)$ holds for all $w_{w_{i}}$.

The utilitarian principle allows that an action brings about some bad consequences if it at the same time brings about more good consequences. For instance, it allows sacrificing some people if this sacrifice serves the good of many people. As an alternative to utilitarian permissibility we introduce the principle of Pareto permissibility. To this end, we first define the notion of Pareto dominance, which allows us to conclude that some action brings about a negative outcome in some respect, although it may be the optimal action from an utilitarian point of view. An option $w_{a}$ dominates another option $w_{b}$ if and only if $w_{a}$ is no worse in any aspect compared to $w_{b}$, and $w_{a}$ improves at least one aspect of $w_{b}$ either by making more good consequences obtain or less bad consequences obtain. Thus the agent does not change the world for the worse and will change it for the better by choosing the dominant action instead of the dominated one.

## Definition 3 (Pareto Dominance)

Let $w_{0}, w_{1}$ be two available options, let cons ${ }_{w_{i}}^{\text {good }}=$ $\left\{c \mid M, w_{i} \models c \wedge u(c)>0\right\}$ be the set of good consequences of option $w_{i}$, cons $\overline{w_{i}} \overline{\text { good }}=\left\{c \mid M, w_{i} \models \neg c \wedge u(c)>0\right\}$ the set of good consequences that does not obtain in option $w_{i}$, and cons $_{w_{i}}^{\text {bad }}=\left\{c \mid M, w_{i} \models c \wedge u(c) \leq 0\right\}$ the bad consequences of option $w_{i}$. Option $w_{0}$ dominant option $w_{1}$ if and only if the following conditions hold: 1) $w_{0}$ shares all the good consequences with $w_{1}\left(M, w_{0} \models \bigwedge\right.$ cons $\left._{w_{1}}^{\text {good }}\right)$, 2) $w_{0}$ either has at least one good consequence that does not hold in $w_{1}$, or $w_{1}$ has at least one bad consequence that does not hold in $w_{0}\left(M, w_{0} \models \bigvee \operatorname{cons}_{w_{1}}^{\overline{\text { good }}}\right.$ or $\left.M, w_{0} \models \neg \bigwedge \operatorname{cons}_{w_{1}}^{\text {bad }}\right)$, and 3) all the bad consequences of $w_{0}$ are also bad consequences of $w_{1}$ $\left(M, w_{1} \models \bigwedge\right.$ cons $\left._{w_{0}}^{b a d}\right)$.

Based on Pareto dominance, Pareto permissibility is defined. Pareto permissibility permits options not dominated by other options. Pareto permissibility can thus be understood as a principle of moral rationality: If there is an option that is better in all aspects compared to an alternative, then the only rational choice is to choose the better one. It would be irrational (and thus impermissible) to choose the worse alternative.

## Definition 4 (Pareto Permissibility)

Let $w_{1}, \ldots, w_{n}$ be the set of options available to an agent. Option $w_{i}$ is permissible according to the Pareto principle if and only if it is not dominated by some option $w_{j}$.

As will become apparent below, utilitarian permissibility and Pareto permissibility predict the same set of permissible actions for some dilemmas and different sets of permissible actions for other dilemmas. Generally, actions permissible from the utilitarian point of view are also permissible from the Pareto point of view. But the converse does not hold: For some dilemmas, the set of actions permitted by each principle differ. In those cases of disagreement the moral reasoner has to solve a conflict.

## Models of Moral Dilemmas

In this section, the four dilemmas presented in the introduction are modeled within the framework of causal agency mod-
els. Commonalities and differences are discussed both with respect to representation and ethical reasoning.

## Representations

Consider the Runaway Trolley dilemma (cf., p.1). We model this situation from the perspective of the bystander, who faces the decision to either throw the switch or to refrain from doing so. Let $a_{1}$ be the action variable representing the action of throwing the switch, and $a_{2}$ be the action variable representing refraining from throwing the switch. The consequence variable $c_{1}$ represents that the one person on the other track dies, and the consequence variable $c_{2}$ represents that the five persons on the current track die. The causal mechanism is expressed by structural equation in the following way: The structural equation $c_{1}:=a_{1}$ states that throwing the switch brings about the death of the one person on the other track, and the structural equation $c_{2}:=\neg a_{1}$ states that not throwing the switch will bring about the death of the other five persons. We assign utilities $u\left(c_{1}\right)=-1$ and $u\left(c_{2}\right)=-5$ to the consequences reflecting the number of deaths. For the lucky case that $c_{1}$ or $c_{2}$ do not obtain, we assume positive consequences, viz., $u\left(\neg c_{1}\right)=1$ and $u\left(\neg c_{2}\right)=5$. (One could argue that it is also appropriate to set $u\left(\neg c_{1}\right)=u\left(\neg c_{2}\right)=0$, because survival does not improve the persons' current state of being alive. On the other hand, to escape from danger intuitively bears positive utility. We consider this question as another empirical question that is out of the scope of this paper. For now it is important to note our findings do not depend on this choice.)

We consider now the Pregnancy dilemma and model the situation from the perspective of the doctor, wo faces the decision to either treat the woman or to refrain from doing so. Thus, we are assuming two actions $a_{1}$, treating the woman, and $a_{2}$, refraining from treating the woman. Moreover, we introduce consequence $c_{1}$ representing that the woman dies, and consequence $c_{2}$ representing that the triplets die. The structural equations are $c_{1}:=a_{1}$ and $c_{2}:=\neg a_{1}$. The utilities are set in accordance with the number of dying individuals: $u\left(c_{1}\right)=-1$ and $u\left(c_{2}\right)=-3$. As with the first dilemma, we assume that not dying yields positive utility, and hence we set $u\left(\neg c_{1}\right)=1$ and $u\left(\neg c_{2}\right)=3$.

Note that the Pregnancy dilemma is structurally isomorphic to the Runaway dilemma, i.e., the dilemmas can be mapped to each other. The only difference is the number of deaths in case of inaction (3 versus 5). Hence, we do not expect big differences regarding the complexity of reasoning about these dilemmas.

The Boat dilemma is modeled from the perspective of the crew, that has to decide whether to throw the biggest person into the sea. We assume two actions $a_{1}$, throwing the biggest person into the sea, and $a_{2}$, refraining from doing so. In contrast to the two previous dilemmas, it would be incorrect to model this dilemma as a choice between the one dying because of performing $a_{1}$ and the other three dying because of refraining from action. Instead, the model has to capture that the biggest person will die in both cases, viz., either because of being thrown into the sea or by drowning together with his
colleagues because of the sinking ship. To represent this situation appropriately, we assume three consequences: the ship sinks $\left(c_{1}\right)$, the biggest person dies $\left(c_{2}\right)$, and the three other passengers die ( $c_{3}$ ). The structural equations are $c_{1}:=\neg a_{1}$ (the ship will sink if the biggest person is not thrown into the see), $c_{2}:=a_{1} \vee c_{1}$ (the biggest person will die if she is thrown into the sea or if the ship sinks), and $c_{3}:=c_{1}$ (the three other passengers will die if the ship sinks). The utilities again reflect the number of deaths: $u\left(c_{2}\right)=-1$ and $u\left(c_{3}\right)=-3$, and as with the other two principles we assume that $u\left(\neg c_{2}\right)=1$ and $u\left(\neg c_{3}\right)=3$.

The Hijacked Airplane dilemma again is isomorphic to the Boat dilemma. It can thus be modeled accordingly: $a_{1}$ refers to the action of shooting the airplane, and $a_{2}$ to refraining from doing so. Consequence $c_{1}$ represents the airplane crashing, $c_{2}$ represents the death of the passengers, and $c_{3}$ corresponds to the death of people on the ground. The utilities can be set to any values such that $u\left(c_{2}\right)>u\left(c_{3}\right)$.

## Ethical Reasoning

The ethical principles "utilitarian permissibility" and "Pareto permissibility" defined above can now be applied to the outlined models of the four moral dilemmas. The first observation is that according to utilitarian permissibility taking action $\left(a_{1}\right)$ is permissible and refraining from action $\left(a_{2}\right)$ is impermissible in all four dilemmas, i.e., it is obligatory to throw the switch, to treat the woman, to throw the biggest crew member into the sea, and to shoot the hijacked airplane. This is rather easy to see by considering the sums of the utilities. E.g., throwing the switch in the Runaway Trolley dilemma yields utility $u\left(c_{1} \wedge \neg c_{2}\right)=-1+5=4$ whereas not throwing the switch yields $u\left(\neg c_{1} \wedge c_{2}\right)=1-5=-4$.

For the Runaway Trolley dilemma and the Pregnancy dilemma, performing action $a_{1}$ does not dominate refraining from action $\left(a_{2}\right)$ according to the definition of Pareto dominance. To see this, note that cons ${ }_{w_{a_{2}}}^{\text {good }}=\left\{\neg c_{1}\right\}$ (i.e., the good thing about not throwing the switch is that the one person will not die, and the good thing about not treating the woman is that the woman will not die) but $M, w_{a_{1}} \not \vDash \neg c_{1}$ (i.e., the one person will die in case of throwing the switch, and the woman will die in case of treatment). Conversely, using exactly the same argument refraining from action does not dominate acting. Thus, no matter how one decides someone will be harmed who will not be harmed under the alternative option. Because no action is dominated by the other, both the actions are permissible according to Pareto permissibility.

For the Boat dilemma and the Hijacked Airplane dilemma, performing action $a_{1}$ is the only Pareto permissible choice. The reason is that drowning the biggest person and shooting the airplane dominate the respective alternatives. Note that $w_{a_{1}}$ dominates $w_{a_{2}}$ according to the definition of Pareto dominance: First, observe that $\operatorname{cons}_{w_{a_{2}}}^{\text {good }}=\emptyset$ (i.e., refraining from action yields no positive consequences), $\operatorname{cons}_{w_{w_{2}}}^{\overline{g o o d}}=$ $\left\{\neg c_{2}, \neg c_{3}\right\}$ (i.e., when refraining from action none of the positive consequences hold), and $\operatorname{cons}_{w_{a_{1}}}^{b a d}=\left\{c_{2}\right\}$ (i.e., the nega-
tive consequence of $a_{1}$ is that the biggest person dies resp. the passenger die). Second, verify that indeed $M, w_{a_{1}} \models \top$ (satisfying condition 1 of the definition of Pareto dominance, all the good consequences of refraining are also good consequences of throwing, viz., there are none), $M, w_{a_{1}} \models \neg c_{2} \vee$ $\neg c_{3}$ (satisfying condition 2 of the definition of Pareto dominance, throwing (shooting) yields one of the good consequences not yielded by refraining, viz., $\neg c_{3}$ ), and $M, w_{a_{2}} \models c_{2}$ (satisfying condition 3 of the definition of Pareto dominance, the bad consequences of throwing (shooting) is also a bad consequence of refraining).

To sum up, for the isomorphic pair Runway Trolley dilemma and Pregnancy dilemma, both taking action and refraining are Pareto permissible but only the former is permitted by the utilitarian principle. Thus, the two principles are in conflict. For the isomorphic pair Boat dilemma and Hijacked Airplane dilemma, the two principles agree on only permitting taking action.

## Type 1 and Type 2 Dilemmas

Our formal investigations suggest that the moral dilemmas we are considering can be classified based on their formal properties. All the considered dilemmas are constituted by the choice between a big sacrifice as a consequence of inaction or a smaller sacrifice as a consequence of action. However, in case of the Runaway Trolley and the Pregnancy dilemma, the sets of negatively affected people are disjoint, whereas in case of the Boat dilemma and the Hijacked Airplane dilemma, the set of negatively affected people as a consequence of action is a subset of the set of negatively affected people as a consequence of inaction. This analysis yields that putting other people in danger by saving some raises moral conflicts, whereas saving a subset of people in danger does less so.

We take this difference to be a justification for subsuming dilemmas of the Runaway Trolley and Pregnancy dilemma type under Type 1 dilemmas, and dilemmas of the Boat and Hijacked Airplane type under Type 2 dilemmas. We conjecture that the utilitarian choice does Pareto dominate the alternative option in case of Type 2 dilemmas whereas it does not in Type 1 dilemmas. Thus, for Type 1 dilemmas, ethical principles predict different sets of permissible actions, and hence there is a conflict to resolve which is not present for Type 2 dilemmas. We therefore hypothesize that Type 2 dilemmas are easier to solve for humans, and we present a study which confirms our hypothesis.

## Hypotheses

The above theoretical analysis predicts that Type 2 dilemmas-due to the absence of a moral conflict-are easier to solve than Type 1 dilemmas. These considerations lead to two testable hypotheses:

- Hypothesis 1: Type 1 dilemmas such as the Pregnancy and the Runaway Trolley dilemma are rated as equally difficult.
- Hypothesis 2: Type 2 dilemmas such as the Boat dilemma
and Hijacked Airplane dilemma are rated as significantly easier to solve than Type 1 dilemmas.

Both hypotheses can be formally justified: The Type 1 dilemmas Pregnancy and Runaway Trolley are isomorphic, i.e., each one can be mapped to the other conserving the structure of the problem. Hypothesis 2 is justified for Type 2 dilemmas, as the utilitarian optimum dominates the possibility of refraining from action. This does not hold for Type 1 dilemmas. These hypotheses are investigated in the next section experimentally.

## Experiment

We report the second part of an experiment that focuses on rating the difficulty of moral dilemmas.

## Methods

Participants Participants were recruited on the online platform Amazon Mechanical Turk and received a monetary compensation for their participation. A total of 60 participants $(f=33)$ completed the study $\left(M_{\text {age }}=40.7, S D_{\text {age }}=\right.$ $8.86, \min _{\text {age }}=21, \max _{\text {age }}=70$ ). $33 \%$ of the participants reported to have finished high school or college, $12 \%$ stated to have an associate degree, $32 \%$ reported to have a bachelor degree while $23 \%$ stated to have a master or a higher academic degree.

Procedure, Design and Materials After the introduction to the setting participants received three problems. Each problem consisted of brief descriptions of two moral dilemas (c.f., Bucciarelli, Khemlani, \& Johnson-Laird, 2008), both presented at the same time on the left or the right part of the screen. Participant had to decide which of these two moral decision situations was more difficult to make, given that they should aim for saving lifes. More precisely, the participants had to decide between the Pregnancy and Runaway Trolley Dilemma, the Pregnancy and Boat Dilemma, and the Runaway Trolley and Boat Dilemma. Hence, participants were making a binary decision that was encoded in a dichotomous variable. After selecting the more difficult scenario the participants had to rate the perceived difficulty on a scale from 0 (hardly more difficult) to 100 (extremely more difficult) using a slider. This value was encoded in a second variable.

## Results

The frequencies of selections for the moral dilemma decision tasks can be found in Fig. 1. In the first problem the same number of participants rated either the Pregnancy Dilemma or the Trolley Problem to be the more difficult one. In the second problem 38 participants decided the Pregnancy Dilemma to be the more difficult decision scenario while 22 participants chose the Boat Dilemma. In the third problem 44 participants opted for the Trolley Dilemma and 16 for the Boat Dilemma. A two-tailed binomial test was used to compare the frequencies for the dichotomous variable.

As predicted, no reliable difference in the evaluation of the difficulty of the moral dilemmas Pregnancy and Runaway


Figure 1: Frequencies in the evaluation of the moral dilemma difficulty between two tasks $(* \leq .05, * * * \leq .001)$.

Table 1: Mean values for the participants rating of the difficulty to find a decision in the selected scenario.

|  | Mean $_{\text {difficulty }}$ |  |  |
| :--- | :--- | :--- | :--- |
| Decision Task | PW | RT | OB |
| PW-RT | $M=72.37$ | $M=60.23$ |  |
|  | $S D=28.37$ | $S D=32.40$ |  |
| PW-OB | $M=58.32$ |  | $M=51.05$ |
|  | $S D=32.48$ |  | $S D=37.22$ |
| RT-OB |  | $M=50.07$ | $M=43.94$ |
|  |  | $S D=32.99$ | $S D=32.91$ |

Note: PW: Pregnant Woman scenario; RT: Runaway Trolley scenario; OB: Overweight Boat scenario

Trolley can be found (exact binomial test, two-sided, n.s., $n=60$ ). There is a significant difference in the evaluation of the moral dilemmas Pregnancy and Overweight Boat (exact binomial test, two-sided, $p \leq .05, n=60$ ) and a significant difference in the evaluation of the moral dilemmas Runaway Trolley and Overweight Boat (exact binomial test, twosided, $p \leq .001, n=60$ ). Once more, Fig. 1 illustrates the differences of difficulty per decision task. The mean values of the participant's rating of their personal difficulty to find a decision in the previously selected scenario are shown in Table 1. Subsequent two-tailed t-tests showed no significant differences between the mean values $M_{P W}$ and $M_{R T}$ (decision task PW-RT), $M_{P W}$ and $M_{O B}$ (decision task PW-OB), and also not for $M_{R T}$ and $M_{O B}$ (decision task RT-OB) concerning their rating of the subjective difficulty.

## Discussion

As our theory predicted moral dilemmas can systematically differ in their perceived difficulty: When asking about the Pregnancy and Runaway Trolley dilemmas, as hypothesized, no significant difference in the relative difficulty rating could be identified. We explain this by the dilemmas' same complexity of the formal structure requiring a similar cognitive
effort. However, the questions concerning the decision difficulties between the ethical scenarios Pregnancy and Overweight Boat or the Runaway Trolley and Overweight Boat resulted in reliable differences in the evaluation of the difficulty of the moral decision situation. In both cases the Boat Dilemma was selected reliably less often. These results support our theory of a different formal structure implying a different cognitive effort and therefore a lower complexity of the Boat Dilemma.

Once the participants have selected the moral dilemma they perceived to be more difficult (the dichotomous decision), their subsequent rating of the difficulty in the interval from 0 to a 100 is statistically equal in comparison to the rating of the participants who chose the other dilemma confirming the result. Overall, there is a tendency towards a lower decision difficulty in the Boat Dilemma.

## General Discussion

The formally predicted distinction between Type 1 and Type 2 moral dilemmas have been empirically supported. Our results support the theoretical assumption that less the dilemma's content but the formal structure and the associated cognitive effort is a predicting factor affecting people's rating of a dilemmas' difficulty. We recall that a main difference between moral dilemmas of Type 1 and Type 2 are either based on action that the utilitarian choice does not or does Pareto dominate the alternative choices. This connects the presented formalism with ethical principles and a decision theoretic interpretation. For Type 1 dilemmas, ethical principles predict different sets of permissible actions, and hence there is a conflict to resolve which is not present for Type 2 dilemmas. The absence of such a conflict appear at least on the problems' surface to be easier to solve due to the lower cognitive effort they require. Further investigations ought to contain a replication of the results with balanced materials and higher sample sizes. In addition applying qualitative research such as interviews or thinking aloud techniques may give deeper insight in the complex human decision-making process particularly in morally difficult decision situations. This would offer additional insights about the motives, thoughts, and concepts people have when they have to solve tasks about moral principles and can provide the reasons for their decisions. By applying a qualitative content analysis of the different causal structure of dilemmas may improve the detection and categorization of the objective, systematic, and formal features of the dilemma's content. These categories in turn can be validated by an assignment of dilemmas as a possible task in a further experiment. Having a formal theory at hand allows to systematically analyze the implications of the objectives, concepts, and features relevant for moral decision making. Our formalism is able to distinguish between moral dilemmas and-at least for the reported cases-predict a perceived subjective difference between human raters.

## References

Allen, C., Wallach, W., \& Smit, I. (2006). Why machine
ethics? IEEE Intelligent Systems, 21(4), 12-17.
Bonnefon, J.-F., Shariff, A., \& Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573-1576.
Bucciarelli, M., Khemlani, S., \& Johnson-Laird, P. N. (2008). The psychology of moral reasoning. Judgment and Decision, 3(2), 121.
Crockett, M. J. (2013). Models of morality. Trends in cognitive sciences, 17(8), 363-366.
Greene, J. D., \& Haidt, J. (2002). How (and where) does moral judgment work? Trends in cognitive sciences, 6(12), 517-523.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., \& Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron, 44(2), 389-400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., \& Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108.
Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review, 108(4), 814-834.
Halpern, J. Y. (2016). Actual causality. Cambridge, MA: The MIT Press.
Kohlberg, L. (1984). Essays on moral development: Vol. 2. the psychology of moral development: Moral stages, their nature and validity. Harper \& Row.
Lindner, F. (2015). Soziale Roboter und soziale Räume: Eine Affordanz-basierte Konzeption zum Rücksichtsvollen Handeln. Doctoral dissertation, Department of Computer Science, University of Hamburg, Hamburg.
Lindner, F., \& Bentzen, M. M. (2017). The hybrid ethical reasoning agent IMMANUEL. In B. Mutlu, M. Tscheligi, A. Weiss, \& J. E. Young (Eds.), Proceedings of the 2017 conference on human-robot interaction (HRI2017). ACM/IEEE.
Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., \& Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In Proceedings of the tenth annual acm/ieee international conference on human-robot interaction (pp. 117-124).
Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. Trends in cognitive sciences, 11(4), 143-152.
Moll, J., \& de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. Trends in cognitive sciences, 11(8), 319-321.
Shou, Y., \& Song, F. (2015). Moral reasoning as probability reasoning. In D. C. Noelle \& P. P. Maglio (Eds.), Proceedings of the 37th annual meeting of the cognitive science society (pp. 2176-2181). Austin, TX.
Wiegmann, A., \& Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. Cognition, 131(1), 28-43.

# Communicative efficiency in language production and learning: Optional plural marking 

Chigusa Kurumada<br>ckuruma2@ur.rochester.edu<br>Department of Brain and Cognitive Sciences, Meliora Hall University of Rochester

Scott Grimm<br>scott.grimm@rochester.edu<br>Department of Linguistics, Lattimore Hall<br>University of Rochester


#### Abstract

Recent work suggests that language production exhibits a bias towards efficient information transmission. Speakers tend to provide more linguistic signal for meaning elements that are difficult to recover while reducing contextually inferrable (more frequent, probable, or expected) elements. This tradeoff has been hypothesized to shape grammatical systems over generations, contributing to cross-linguistic patterns. We put this idea to an empirical test using miniature artificial language learning over variable input. Two experiments were conducted to demonstrate that the inferrability of plurality information inversely predicts the likelihood of overt plural marking, as would be expected if learners prefer communicatively efficient systems. The results were obtained even with input frequency counts of the plural marker counteracting the bias, and thus provide strong support for a critical role of inferrability of meaning in language learning, production, as well as in typologically attested variations.


Keywords: language production, artificial language learning, optional morphology, plural marking, communicative efficiency

## Introduction

Speakers face a multitude of constraints when encoding their intended message as an actual utterance. On the one hand, speakers want to encode their meaning in a way that guarantees communicative success-it must be understood by the interlocutor. At the same time, speakers need to cope with difficulties associated with utterance planning and articulation. As researchers have shown, speakers regularly do this, e.g. by choosing shorter forms and/or elements that are readily retrieved and formulated (see, inter alia, Ferreira \& Dell, 2000; MacDonald, 2013).

A body of psycholinguistic work sees this negotiation between communicative success and effort minimization as a guiding principle of the computational system underlying language production and comprehension. Specifically, it is expected that there is an efficiency-based trade-off between the amount of information encoded and the amount of linguistic signal expended by the speaker. (e.g., Aylett \& Turk, 2004; Levy \& Jaeger, 2007; Buz, Tanenhaus, \& Jaeger, 2016). Communicative efficiency is predicted to be maximized when the speaker preferentially encodes components of meanings that are otherwise less likely to be inferred by the listener given prior expectations.

Against this backdrop, we consider the possible role of communicative efficiency in the organization of grammatical number marking. Grammatical number systems often have "markedness" contrasts between a default, uncoded value, and a value explicitly coded, e.g. through morphology. Typically, the singular value is uncoded while the plural value
is coded, as in $d o g$ vs. dogs. One question we will ask is if a plural value for a referent is likely to be inferred, will a speaker encode it? While it has been long observed that languages have preferences for what information is coded in default forms as opposed to explicitly coded (e.g., Greenberg, 1966), the causes underlying these preferences have remained obscure. The design of our study allows us to take a step towards distinguishing what sort of information forms the basis for these preferences. In particular, we investigate whether the active ingredient is the predictability of linguistic forms, i.e. frequency of occurrence of some element in produced language, or if predictability is related to the meaning.

## A case study: Optional plural marking

Unlike in English, grammatical encoding of plural meaning (e.g., dog vs. dogs) can be optional in some languages. Optional Plural Marking (OPM) is not uncommon crosslinguistically (e.g., Yucatec Maya (Butler, Bohnemeyer, \& Jaeger, 2017)) and has been investigated in linguistic work on grammatical systems (see Corbett (2000) and Haspelmath (2013) for general discussion). Yet, the mechanisms that predict when speakers would use (or would not use) the marker are not well understood.

A class of proposals, elaborated for number marking more generally, grounds the encoding of number values in conceptual properties related to entities (Prasada, Ferenz, \& Haskell, 2002; Wisniewski, Lamb, \& Middleton, 2003). This viewpoint suggests that singular (or plural) values might be more conceptually consonant for some entity types than for others. For instance, entities that are typically conceptualized as individuals (e.g., large animals) tend to be referenced in language as singular, rather than plural. For these entities, their occurrence in plural is limited, and therefore, plural coding is the unexpected or "marked" value. Conversely, for entities that are often conceptualized as collectives (e.g., small insects), plural coding is the expected or "unmarked" value. In sum, on this view formal (morphological) markedness corresponds to conceptual markedness.

When combined with a framework such as the communicative efficiency hypothesis, this "markedness" of plural meaning can predict biases seen in language production. Put simply, learning and production is guided by a consideration to communicate the plural meaning most efficiently. That is, learners should prefer systems in which markedness of plural meaning is inversely correlated with the production of plural marking. In relation to OPM, accounts based in communicative efficiency predict that when learners of an OPM language
refer to multiples of individualized items (e.g., large animals), they should be more likely to produce plural marking, compared to when referring to multiples of collective items (e.g., small insects).

Preliminary support for the conceptual markedness account comes from repeated observations across a number of studies on typologically-diverse languages which possess a singulative/collective morphology (e.g., Arensen (1998) on Murle, Mifsud (1996) on Maltese, Stolz (2001) on Welsh, see (Grimm, 2012) for discussion). In these languages, referents that are likely to be conceptualized and manipulated as collectives (e.g., fruits, grains, vegetables) or as a group/mass of individuals tend to be expressed with lexical items that have a plural meaning by default (e.g., psy "peas" in Welsh) and only through an additional singulative suffix can singletons be designated (e.g., psy-en "pea").

A difficulty arises, however, in determining "markedness" of plural meaning based on token counts of plural forms in a corpus. Haspelmath and Karjus (2017), for instance, collected token counts of singular vs. plural forms of a word (e.g., psy-en and psy) to argue that frequency asymmetries can predict the asymmetrical plural marking system such that the more frequent meaning (singular/plural) is often encoded in a simpler form. However, in this approach, one can only infer the frequency of meaning (e.g., How often does one talk about pea(s) as singular or plural?) based on the frequency of form (e.g., How often does one use a singular or plural form for pea(s)?). In other words, there is no simple way of dissociating predictions of the communicative efficiency account from an account based on form frequency: speakers may be simply reproducing the patterns heard in the input (e.g., They are more likely to hear psy than psy-en when they see peas and are faithfully representing the pattern in their own production).

To address this problem, we present two production experiments using an artificial language learning paradigm. Learners acquire 12 novel nouns and one novel verb to produce simple intransitive sentences with the Subject-Verb word order. As we describe below, the novel lexicon consists of two classes of referents: six Individuals and six Collectives that depict fictitious animals and insects, respectively. In the input, they were visually presented as either singletons or multiples at varying rates: Individuals are more likely to be singletons whereas Collectives are more likely to be multiples. Referents are optionally (stochastically) plural-marked and the probability of occurrence of the marker was constant across Individuals and Collectives.

This setup pinpoints an instance where frequency (inferrability) of meaning can be examined independent of frequency of forms. For instance, Individuals are less likely to appear as multiples compared to Collectives. This makes the plural meaning less inferrable for Individuals than for Collectives without the overt marking. Therefore, conceptual-markedness based accounts would predict that learners should be more likely to use the plural marker with

## a) singletons


b)multiples


Experiment 1


Experiment 2

Figure 1: Sample images of visual stimuli in Experiments 1 and 2.
the Individuals rather than Collectives. Critically, this bias is not predictable based purely on the frequency of forms. Notice that, given the fact that Collectives are more likely to appear as multiples, a larger proportion of token counts of Collectives appear with the plural marker than Individuals. If learners are simply reproducing the patterns observed in the input, they should produce the optional plural marker more with Collectives than with Individuals.

Results from this investigation may help to bridge the gap between the factors shaping sentence production in language processing and those that are shaping typological patterns. It has long been observed that the lexicon and grammar of languages across the world tend to exhibit many properties that would be expected if language was shaped by communicative pressures (e.g., Zipf (1949); Plotkin and Nowak (2000), also precisely those predicted by accounts of communicatively efficient language production Piantadosi, Tily, and Gibson (2011); Jaeger (2013)). Recent work on learning biases during (miniature artificial) language acquisition has also found similar biases to be active during artificial language learning (e.g., Culbertson, Smolensky, \& Legendre, 2012; Fedzechkina, Newport, \& Jaeger, 2016; Smith \& Wonnacott, 2010). Fedzechkina et al. (2012) found that native speakers of American English, when learning a miniature language with an optional case marking morphology, restructure the input and condition the uses of the marker on factors such as Animacy. This is in line with patterns observed in existing optional (or more categorical) case-marking languages, suggesting a tight link between observations in lab-based studies and typological pattern found in existing languages (e.g., Aissen, 2003; Kurumada \& Jaeger, 2015).


Figure 2: Schematic illustration of the flow of the experiment and proportions of singleton and multiple visual prompts.

## Experiment 1

We employ a miniature artificial language learning paradigm modifying Fedzechkina, Jaeger, and Newport (2012). Participants first learn 12 nouns and then learn to produce intransitive sentences in response to video clip prompts. We manipulated visual features of the referents (e.g., size, group size, movements) as well as the probability with which Individuals (animals) and Collectives (insects) appear as singletons and multiples, respectively. If optional number-marking is affected by a preference for communicative efficiency, speakers should be more likely to produce responses with a pluralmarker for Individual (animal) compared to Collective (insect) referents.

## Methods

Participants 48 native speakers of American English at University of Rochester participated in this study. They received $\$ 10$ for their participation.

## The language

Lexicon We constructed 12 nonce nouns. Six of them denote large animal characters and the other six denote small insect characters (e.g., Fig.1). To ensure that results did not include spurious phonological effects, we created two versions of character-noun combinations. All of the nouns were 1-2 syllables obeying English phonotactics (e.g, norg, velmick, zamper). When characters were presented as multiples, the noun was optionally suffixed with the plural-marker ( $-k a$ ) $2 / 3$ of the time.

We included only one verb - glim - meaning "moving up and down." In constructed sentences, the verb followed a noun, constituting a SV (intransitive) word order (e.g., Velmick-ka glim).

## Procedure

There were five phases in this experiment (Fig. 2). Participants went through phases (1) - (3) for six of the 12 noun types (three animals and three insects) and then repeated the same procedure to learn the other six words.
(1) Word exposure ( $\mathbf{1 2}$ characters $* 2=24$ trials total): During word exposure, participants were presented with pictures of each of the characters. Participants were instructed to repeat the names of the characters aloud. In this phase, all the characters were presented as singletons. An animal was depicted approximately three times as large as an insect.
(2) Word learning game ( 12 characters * $4=48$ trials total): The initial word presentation was followed by a word learning phase where participants were presented with four pictures (4 Alternative-Forced-Choice task) and asked to choose the correct match for the noun provided (48 trials total). Feedback was provided after each trial. In this phase, Individuals and Collectives were presented as singletons and multiples at different rates. Individuals occurred $75 \%$ of the time as a singleton (i.e., one animal, Fig. 1a), and $25 \%$ as multiples (Fig. 1b). Collectives had the inverse distribution (25\% singleton, $75 \%$ multiples). Both Individual (animal) nouns and Collective (insect) nouns were followed by the pluralmarker (ka) $2 / 3$ of the time when occurring as multiples.
(3) Word production ( 12 characters * $1=12$ trials total): Participants were shown 12 characters (singleton) one by one and asked to name each of them.
(4) Sentence comprehension ( $\mathbf{1 2}$ characters $* 4=48$ trials total): During the sentence comprehension phase, participants viewed short clips and heard their descriptions in the novel language. Participants were asked to repeat the sentences out loud. As in the word learning phase, Individuals and Collectives occurred as singletons $75 \%$ and $25 \%$ of the time, respectively, and they were followed by the pluralmarker (ka) $2 / 3$ of the time when occurring as multiples. Consequently, participants heard the animal and insect nouns with
ka 10 times and 30 times, respectively, by the end of this phase (Fig. 2). Importantly, this means that input frequency biases against the prediction of communicative-efficiency: the input in our experiment(s) provides more instances of training for plural-marked Collectives than Individuals.
(5) Sentence production ( 12 characters $* 2=24$ trials total): In the final test (sentence production) phase, participants saw silent videos of singletons and multiples and had to produce intransitive descriptions. In this phase, visual images for the multiples had three instances of the characters both for animals and insects. This was done to ensure that participants use -ka to signal plurality rather than the particular number of instances (two for animals and ten for insects) seen in the exposure input.

## Scoring

In the 4AFC comprehension test, participants' responses were scored as "correct" if they matched the intended referent. Following the standard used in similar studies (e.g., Fedzechkina et al. (2012)), we a priori decided to exclude participants who failed to achieve mean accuracy of $65 \%$ from all analyses.

We transcribed the production obtained in (5) and annotated if participants produced a given noun correctly and if a noun was produced with $k a$ or not. In the comprehension test, participants responses were scored as "correct" if it matched the provided input, while subtle phonological variations (e.g., velmick pronounced as belmick) were ignored.

## Results and Discussion

Comprehension Accuracy To ensure that participants had achieved a sufficient level of accuracy in identifying referents, we first measured their performance in the 4AFC word learning game. The average rate of correct response was $93.9 \%$ (animals, $93 \%$; insects, $94 \%$ ) and all the subject means were well above the pre-determined cut-off rate of $65 \%$. The mean accuracy of the word production phase (3) was above $85 \%$. This suggests that the task was feasible and the lexicon was acquired reasonably well before participants performed the production task.

Plural Marker Use in Production We excluded six (12.5\%) of the participants who failed to produce $50 \%$ of the sentences in the final sentence production phase. This was done to ensure that the data analyzed are produced by those who have mastered the language at a more or less sufficient level. All the results we report below remain unchanged, however, when we include all the participants. We then further removed $116(13 \%)$ sentences that included wrong nouns such as a different character's name or a noun that did not belong to the learned lexicon. The final dataset included 42 subjects and 773 sentences.

Proportions of participants' plural marker use in Experiment 1 are illustrated in Fig. 3. To analyzed the data, we used a mixed effect logit model in R, predicting the use of the optional plural marker. We included the noun classes (Individuals (animals) vs. Collectives (insects)) and visual prompts
(singleton vs. multiples) as fixed effects and participants and items as random effects. The model included the maximal random effects structure justified by the data based on model comparison (Jaeger, 2008). There was an expected significant main effect of visual prompts such that participants were more likely to produce the optional plural marker $k a$ for multiples ( $p<.001$ ). Critically, the interaction between the noun class and the visual prompts was also significant ( $p<.03$ ): learners (inversely) conditioned plural production on plural inferability. They did so despite the fact that they were exposed to three times as many instances of $-k a$ with the Collectives (insects) compared to the Individuals (animals).

## Experiment 2

What is driving the observed difference between Individuals and Collectives? Under our hypothesis, it is at least partially due to the expectation that animals are less likely to be represented with the plural meaning, and hence the meaning is less inferrable (and conversely for insects). In Experiment 1, however, it is not clear if the inferrability of the plural meaning (the conditional probability of multiples given the referent) is learned within the experiment or it is carried over from participants' prior semantic knowledge that insects are more likely to occur, and be referred to, as multiples.

To separate these two factors, in Experiment 2, we used the lexical items from Experiment 1 while associating them with novel geometrical shapes to minimize effects of prior semantic knowledge. If participants exhibit the same asymmetric use of the plural marker for Individuals and Collectives, that will yield support for the idea that the inferability is likely extrapolated in this experiment.

## Participants

52 native speakers of American English at University of Rochester participated in this study. They received $\$ 10$ for their participation.

## The language

The lexicon was identical to that used in Experiment 1. The only difference is that the visual images consisted of 12 geometrical shapes with no commonly known names. To equate the visual features of the referents (e.g., size, spacial distributions, complexity of visual scenes), we created two classes of referents (Fig. 1). Individuals consisted of six relatively large geometrical shapes spatially distributed in a manner similar to how the animals were presented in Experiment 1. On the other hand, Collectives consisted of six smaller shapes that replace the insects in Experiment 1.

## Procedure

The same as Experiment 1.

## Results and Discussion

Comprehension Accuracy The mean accuracy in the 4AFC task was $86 \%$ (Animals, 89\%; Insects, 83\%), suggesting that the word learning was slightly more difficult in Experiment


Figure 3: Proportions of plural marker use by conditions. Dots present by-participant averages (White $=$ singleton visual prompt; Black $=$ multiple visual prompt). Error-bars show 95\% Confidence Intervals. Dotted line indicates the input ratio of the $-k a$ marking for the multiples.

2 compared to Experiment 1, presumably due to the overall unfamiliarity with the geometrical shapes. One subject could not achieve the cut off rate of $65 \%$ and was removed from the analysis. The mean accuracy in the word production phase (3) was $80 \%$.

Plural Marker Use in Production We excluded ten (19.2\%) of the participants who failed to produce $50 \%$ of the sentences in the final sentence production phase. As in Experiment 1 , all the results we report below remain unchanged with the complete set of data. We then further removed 151 ( $15.5 \%$ ) sentences that included wrong nouns. The final dataset included 42 subjects and 823 sentences.

Proportions of participants' plural marker use in Experiment 2 are illustrated in Fig. 4. We constructed a combined model with the noun classes (Individuals vs. Collectives), visual prompts (singleton vs. multiples), and experiments as fixed effects and participants and items as random effects. As in Experiment 1, we found a significant main effect of visual prompts (= more $k a$ use for multiples) $(p<.001)$ and an interaction between the noun class and the visual prompts ( $p<.002$ ), indicating an inverse conditioning of $-k a$ production on plural inferrability. Importantly, there was no significant effect of the experiments. This suggests that the plural predictability is not necessarily tied to participants' prior knowledge of the semantic classes (animals vs. insects) and is learnable with respect to new classes of referents.

## General Discussion

Our results suggest that native speakers of American English prefer to produce an NP without overt marking of plurality when the meaning is more inferrable given the noun classes (e.g., animals vs. insects). The effect was present even with the nonce noun classes, when their within-experiment


Figure 4: Proportions of plural marker use by conditions. Dots present by-participant averages (White $=$ singleton visual prompt; Black $=$ multiple visual prompt). Error-bars show 95\% Confidence Intervals. Dotted line indicates the input ratio of the $-k a$ marking for the multiples.
statistics, as well as visual features of referents (size, spacial arrangements, and movement patterns), support differential plural predictability. We thus argue that learners have implicit knowledge of the relative inferrability of plural meaning (e.g., How often do you describe animals/insects as singletons vs. multiples?), and this knowledge supports the learning of morphological systems of a novel language. Critically, English does not have the optional plural marking (OPM) system. Still, when native speakers of English are exposed to an OPM language with no bias to mark plurality for low-inferrability items, they end up producing more plural marking for less inferrable items.

The current results constitute strong support for the view that language production is optimized to maximize the efficiency of information transmission (Levy \& Jaeger, 2007, Jaeger, 2013). The asymmetrical uses (and non-uses) of $k a$ cannot be accounted for in terms of availability of an upcoming linguistic element or other sources of speaker-internal production or planning difficulties (Ferreira \& Dell, 2000; MacDonald, 2013), since all the sentences were produced with the same verb and no participant failed to learn the verb.

It is an open question how learners compute the plural predictability. In the current experiment, we provided multiple cues to noun classes beyond the statistics of singleton vs. multiples. For instance, Individuals were always depicted larger in size than Collectives. In the sentence comprehension and production phases, each instance of Individuals moved independently while Collectives always showed a group motion. Future studies can manipulate these cues separately to delve into effects of spacio-temporal distributions of referents on conceptualization of noun classes and their plural inferrability.

Lastly, this study has broad implications for understand-
ing typologically attested morpho-syntactic variation. It has long been hypothesized that conceptual markedness plays a guiding role in grammaticalization of morpho-syntactic elements. The current experimental paradigm using an artificial language allows us to dissociate the effects of input in terms of the predictability of forms (e.g., How often do you hear a particular noun with $-k a$ ?) and the predictability/inferability of meaning (e.g., How likely is it that a given referent is described as a singleton vs. multiples?), making it possible to test a multitude of hypotheses put forward about effects of meaning-based predictability. For instance, it has been observed that functionally paired objects (e.g., glasses, chopsticks, a set of pillars) and body-parts (e.g., eyes, ears, hands) are often conceptualized as plural by default, and hence likely encoded without any additional plural marking morphology (Haspelmath \& Karjus, 2017). We can directly test this hypothesis in the current paradigm using objects that differ in their likelihood of appearing in pairs.

In summary, the inferrability of plurality information guides learners to restructure the input they receive, as would be expected if language users are biased towards communicatively efficient systems. Our results thus illuminate the critical role of distributional information of meanings on language learning, production, and typological variation across languages.

## References

Aissen, J. (2003). Differential object marking: Iconicity vs. economy. Natural Language and Linguistic Theory, 21, 435-483.
Arensen, J. E. (1998). Murle categorization. In G. J. Dimmendaal \& M. Last (Eds.), Surmic languages and cultures (pp. 281-318). Rüdiger Köppe.
Aylett, M. P., \& Turk, A. (2004). The smooth signal redundancy hypothesis. Language and Speech, 47(1), 31-56.
Butler, L. K., Bohnemeyer, J. B., \& Jaeger, T. F. (2017). Plural marking in Yucatec Maya at the syntax-processing interface. In A. Machicao y Priemer, A. Nolda, \& A. Sioupi (Eds.), Zwischen Kern und Peripherie (Studia Grammatica, volume 75). Berlin: Akademie-Verlag.
Buz, E., Tanenhaus, M. K., \& Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. Journal of Memory and Language, 89, 68-86.
Corbett, G. G. (2000). Number. Cambridge: CUP.
Culbertson, J., Smolensky, P., \& Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122(3), 306-329.
Fedzechkina, M., Jaeger, T. F., \& Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. PNAS, 44(109), 17897-17902.
Fedzechkina, M., Newport, E. L., \& Jaeger, T. F. (2016). Balancing effort and information transmission during language acquisition : Evidence from word order and case marking. Cognitive Science, 1-31.

Ferreira, V. S., \& Dell, G. S. (2000). The effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive Psychology, 40, 296-340.
Greenberg, J. H. (1966). Universals of language (2nd ed.). Cambridge: MIT Press.
Grimm, S. (2012). Number and Individuation. Unpublished doctoral dissertation, Stanford University.
Haspelmath, M. (2013). Occurrence of nominal plurality. In M. S. Dryer \& M. Haspelmath (Eds.), The world atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Haspelmath, M., \& Karjus, A. (2017). Explaining asymmetries in number marking: Singulatives, pluratives and usage frequency. Linguistics.
Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. JML, 59, 434-446.
Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. Frontiers in Psychology, 4, 1-4.
Kurumada, C., \& Jaeger, T. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. Journal of Memory and Language, 83.
Levy, R., \& Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, \& T. Hoffman (Eds.), Advances in Neural Information Processing Systems (NIPS) (Vol. 19, p. 849-856). Cambridge, MA: MIT Press.
MacDonald, M. (2013). How language production shapes language form and comprehension. Frontiers in Psychology, 4, 226.
Mifsud, M. (1996). The collective in Maltese. Rivista di Linguistica, 8(1), 29-51.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2011). The communicative function of ambiguity in language. Cognition, 122, 280-291.
Plotkin, J., \& Nowak, M. (2000). Language evolution and information theory. Journal of Theoretical Biology, 205(1), 147-159.
Prasada, S., Ferenz, K., \& Haskell, T. (2002). Conceiving of entities as objects and stuff. Cognition, 83, 141-165.
Smith, K., \& Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. Cognition, 116(3), 444-449.
Stolz, T. (2001). Singulative-collective: Natural morphology and stable cases in Welsh number inflexion of nouns. Sprachtypologie und Universalienforschung, 54(1), 5276.

Wisniewski, E. J., Lamb, C. A., \& Middleton, E. L. (2003). On the conceptual basis for the count and mass noun distinction. Language and Cognitive Processes, 18, 583-624.
Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. AddisonWesley.

# Behavioral Dynamics and Action Selection in a Joint Action Pick-and-Place Task 

Maurice Lamb ${ }^{1}$ (maurice.lamb@uc.edu), Tamara Lorenz ${ }^{1,2,3}$ (lorenztr@ucmail.uc.edu), Stephen Harrison ${ }^{4}$ (steven.harrison@uconn.edu), Rachel Kallen ${ }^{1}$ (rachel.kallen@uc.edu), Ali Minai ${ }^{2}$ (ali.minai@uc.edu), Michael Richardson ${ }^{1}$ (michael.richardson@uc.edu)

${ }^{1}$ Department of Psychology, 4150 Edwards 1<br>Cincinnati, OH 45221 USA<br>2Department of Electrical Engineering and Computer Science, 812 Rhodes Hall<br>Cincinnati, OH 45221 USA<br>${ }^{3}$ Department of Materials and Mechanical Engineering, 2600 Clifton Ave, Cincinnati, Ohio 45221, USA<br>${ }^{4}$ Department of Kinesiology, University of Connecticut, CT, USA


#### Abstract

Many common tasks require or are made more efficient by coordinating with others. In this paper we investigate the coordination dynamics of a joint action pick-and-place task in order to identify the behavioral dynamics that underlie the emergence of human coordination. More precisely, we introduce a task dynamics approach for modeling multi-agent interaction in a continuous pick-and-place task where two agents must decide to work together or alone to move an object from one location to another. Our aims in the current paper are to identify and model (1) the relevant affordance dynamics that underlie the selection of the different action modes required by the task and (2) the trajectory dynamics of each actor's hand movements when moving to grasp, relocate, or pass the object. We demonstrate that the emergence of successful coordination can be characterized in terms of behavioral dynamics models which may have applications for artificial agent design.


Keywords: Behavioral Dynamics, Affordances, Multi-agent Coordination, Dynamical Modeling, Joint action, Pick-andplace, Dynamical Systems Theory, Decisions

## Introduction

Often, everyday tasks can be accomplished more quickly and efficiently when individuals work together and coordinate their actions to accomplish task goals. However, increasing the number of individuals engaged in a task constructively increases the complexity of the task by expanding the degrees of freedom and interactions that define the task action space. Computational approaches to dealing with the increased complexity of joint action tasks largely focus on reducing complexity by identifying representational or neural structures that support successful joint action (Graf, SchützBosbach, \& Prinz, 2009; Rizzolatti \& Craighero, 2004; Sebanz \& Knoblich, 2009). Equally important, however, is understanding what aspects of successful multiagent coordination naturally emerge from the physical and informational dynamics of a given task context (Richardson \& Kallen, 2016; Richardson, Marsh, \& Schmidt, 2010; Richardson et al., 2015). The aim of the current project was to identify these task dynamics for a simple joint action pick-and-place task. Of particular interest, was the degree to which the complex patterns of interpersonal movement coordination and action (affordance) selection that emerge could be
captured by extending a low dimensional behavioral dynamics (Warren, 2006) model of individual environmental route navigation (Fajen \& Warren 2003) and pick-and-place behavior (Lamb et al., under review, Washburn, et al. 2015). Because the pick-and-place behaviors exhibited by the proposed low dimensional model emerge from the physical and informational dynamics of a given task context, the proposed model may be developed as a simple artificial agent system that can interact with human co-actors. An artificial agent system based on the model would be able to interact with human co-actors in the task without access to a coactor's cognitive states, i.e. without a theory of mind.

## Methods

## Participants

20 University of Cincinnati students (aged 18 to 28 years) were recruited to participate in the experiment. Participants received credit as a part of a class requirement for an undergraduate Psychology course. All participants provided written consent prior to completing the study, with the procedures and methodology employed reviewed and approved by the University of Cincinnati Institutional Review Board.

## Materials and Apparatus

An illustration of the experimental task setup is provided in Figure 1. Participants stood at a $1.65 \mathrm{~m} \times 0.89 \mathrm{~m} \times 0.995 \mathrm{~m}$ table in a laboratory room and completed a joint action pick-and-place task in a virtual environment. The virtual environment consisted of a room similar to the laboratory room and a table that was isomorphic in size and location to the table in the laboratory room. The virtual environment was displayed to each participant using Oculus Rift DK2 virtual reality headset (Oculus VR, Irvine, California). The physical table acted as a solid surface limiting the participants' movements within the virtual environment and creating a surface on which the participants could move a hand-held wireless Polhemus Latus motion-sensor (Polhemus Ltd, Vermont, USA) that tracked their right hand movements within the virtual environment at 96 Hz . The participants'
head movements were also tracked using Oculus Rift DK2 head tracking system.

The virtual environment, task objects, and task controllers were designed using the Unity 3D game engine (version 5.2.0; Unity Technologies, San Francisco, California) and Sketchup 2015 (Tremble Navigation Technologies, Sunnyvale, California). The maximum display latency between the participants' real-world movements and their movements in the virtual environment was 33 ms . The experimental task states were continuously recorded at 70 Hz .

As indicated in Figure 1, Participant locations were identified in terms of the "A" side or the "B" side of the table, where the body of the participant on the A side of the table (participant A ) is nearer to the center of the appearance range than the body of the participant on the B side (participant B). Participant A was positioned on the A side of the table, standing half way between the middle of the table and the pickup location. Participant B was positioned such that their right shoulder was directly across the table from the right shoulder of participant A (see Figure 1).

Within the virtual environment, the participants were represented as identical virtual avatars modeled after a crash


Figure 1: Illustration of experimental setup. At the beginning of each trial a virtual disc would appear on the left side of the table within the appearance range. Disc color indicated target location for that trial. The targets squares ( $\mathrm{M}=$ magenta, $\mathrm{Y}=$ yellow, $\mathrm{G}=$ green, $\mathrm{B}=\mathrm{Blue}, \mathrm{R}=$ red $)$ were always visible on the right side of the table. test dummy with a height of 1.8 m . Both the participants' right hands were represented by a semi-transparent blue sphere at the end of the dummy's right wrist in order to simplify interaction with the task environment. An inverse kinematics
controller (model and controller supplied by Root Motion, Tartu, Estonia) driven by the Polhemus motion sensor movements and the head movements of the participants controlled the right arm and body movements of the participants' virtual avatar, respectively. The resulting arm and body movements were not identical to the real world arm and body movements of the participants, but were close enough to render any differences between the real and virtual body postures of the participants unnoticeable or not functionally relevant.

## Experimental Task

The experimental task required participants to work together to move virtual disc objects (henceforth disc) that appeared on one end of a virtual tabletop to one of five evenly spaced target locations on the other end of the table (see Figure 1). The target location for a given trial was indicated by the color of the disc. A trial involved successfully moving a disc to the correct target location. Target colors and locations did not change during the task. However, discs appeared in random locations along the $y$ table axis within the middle third of the table (appearance range). Participants completed 2 blocks of 150 trials, 30 trials for each target color. Target colors were randomly presented.

The participants were instructed to pick up the disc when it appeared and attempt to move it to the target location. Either participant could pick up the disc, but they were instructed not talk or gesture to one another during the task. A pickup occurred when a participant's sphere came in contact with the disc. When picked up, the disc moved with the participant's sphere until it reached the target or the participant passed the disc. The participants were informed that if the target was either too far away or uncomfortable to reach, they could pass it to the other participant. A pass involved picking up the disc and then releasing it somewhere on the table by lifting their hand from the table. To complete a pass, the other participant would pick up the disc and move it to the target. A trial was completed when the disc reached the correct target.

## Procedure

Participants were told that the experiment was investigating the dynamics of joint action pick-and-place behavior and that they would be completing a simple pick-and-place task with one another. The participants were then embedded within the virtual environment using the HMD and viewing height and sensor calibration was performed. Task instructions were then provided to the participants and after participants indicated that they understood the task procedure and goal, they were given an opportunity to complete 2 practice blocks. The first practice block consisted of 12 trials where the disc always appeared in the center of the appearance region and indicated the middle (green) target. Each participant took 6 turns picking up the disc and either passed it or took it to the target. The second practice block involved 20 trials, 4 trials for each target location. In this practice block, the pickup location was randomly assigned within the appearance range on each trial.

As mentioned above, experimental trials were broken up into 2 blocks of 150 trials. After the first experimental block participants switched sides of the table, i.e. the participant on side A moved to side B and the participant on side B moved to side A. Participants switched HMDs and moved to the other side of the lab room table to effect this switch. Experimental blocks lasted between 10 and 15 minutes.

## Results

In order to model the emergence of successful joint action during the current pick-and-place task, the analysis was directed towards answering two general questions. First, what task variables determined the participants' decision to pickup and/or or pass. In other words, what were the affordance (action opportunity; Gibson, 1979) based action selection dynamics that characterized pickup and pass behavior? Second, what were the trajectory dynamics of the participant's hand movements when moving to grasp, relocate, or pass the disc within a two-dimensional task space. Below we consider each of these questions in turn.

## Decisions

For the pick-and-place task investigated here there were two affordance based action selection decisions that we examined. First, participants had to decide whether or not to pick up the disc when it appeared. In order to understand the basis for this decision we applied the C4.5 decision tree algorithm (Quinlan, 1993) using a 10 fold cross validation to participant pick decisions ( $n=2998$ ) in order to create a decision tree with a minimum node size of 50 instances. This analysis revealed that the only attributes used to build the tree were the current location of each actor's hand to the disc, with the resulting decision tree able to correctly predict $87 \%$ of the pick decisions. Attributes that were considered for each participant included: hand's current distance to the disc, hand's current distance to target, disc location, and target location. These attributes were not considered relevant to modeling the decision behavior if it was not included in the decision tree produced by the C 4.5 method or if its exclusion resulted in a change in predictive success of $<3 \%$.

The C4.5 decision tree algorithm was also applied using a 10 fold cross validation to a data set of 2998 passing decisions in order to create a decision tree with a minimum node size of 50 instances. When the only attribute used to build the tree was the distance of the resting location of one of actor's hand to the disc the resulting decision tree was able to correctly predict $79 \%$ of the pickup decisions. Resting hand location for each side was defined as a position 0.15 m from the edge of the table directly in front of the participant's right shoulder. The same set of attributes considered for the pickup decision were considered for the pass decision, with the addition of the previous pass decision. None of these other attributes

[^380]significantly increased pass prediction beyond that predicted by actor resting location alone.


Figure 2: Heat-maps of example participant (top) and model simulation (bottom) trajectories during the experimental task.

## Movements

An example of the complete set of participant pair trajectories are illustrated in figure 2 (top) as a heat map. This heat map plot was created by dividing the table into $125 \times 108$ grid and for each trial, the number of times a participant's location was recorded in a given grid cell was tallied to create a histogram of trajectory locations in table coordinates. A greyscale value was assigned to each cell from a scale of 64 shades. All participants exhibited a qualitatively similar sideways "spaghetti monster" heat-map, with concentrations of trajectories (brighter areas), corresponding to discs (far left side of heat-map plot), pass/rest locations (top and bottom left of center on the heat-map plot), and target locations (5 distinct points across the right of the heat map plot). Because of the number of subtask trajectories, trajectory heat maps provide a relatively straightforward tool for comparing qualitative similarities between both human participants and between human data and simulation data. ${ }^{1}$

Participant trajectories tended to curve slightly away from straight-line trajectories. This may be due to the fact that after completing a subtask goal participants employed a simple strategy of heading in the general direction of the next subtask goal instead of taking an initial heading defined by the straight-line angle from their current location to the goal location.

Participant subtask movements exhibited a bell shaped velocity profile with the peak velocity occurring around half
way through the trajectory. Across all subtask trajectories, the average peak velocity was $1.231 \mathrm{~m} / \mathrm{s}(\mathrm{Mdn}=1.252 \mathrm{~m} / \mathrm{s}, \mathrm{Q} 1=$ $0.924 \mathrm{~m} / \mathrm{s}, \mathrm{Q} 3=1.373 \mathrm{~m} / \mathrm{s}$ ) and the peak velocity occurred on average around $57 \%$ ( $\mathrm{SD}=15 \%$ ) of any given subtask trajectory. For the 14 subtask trajectories examined, average peak velocity for each subtask trajectory was significantly correlated, $r(14)=0.89, p<0.001$, with the average straightline distance of each subtask trajectory. Shorter trajectories had lower average peak velocities than longer trajectories.

In order to identify where participant's passed the disc on pass trials, cluster analysis was conducted, using the K-means cluster analysis algorithm, which finds cluster centers that minimize the sum of squared error (SSE) for a given number of clusters, k. We analyzed the release/pass locations to determine whether these locations typically clustered around 1,2 , or 3 cluster centroids. The optimal number of clusters was defined as the value of $k$ such that the difference of the SSE for a reference distribution, determined by Monte Carlo sampling of a reference distribution, was greatest compared to the other values of $k$. For each pair, separate evaluations were run for each side of the table. For side A, when a participant on side $A$ passed at least once during the experiment ( $n=8$ pairs), the optimal number of clusters was 1 for all passes on this side of the table. Likewise, when a participant on side $B$ passed at least once during the experiment ( $n=9$ pairs), the optimal number of clusters was 1 for most pairs $(n=7)$. When a participant on side A passed during the experiment $(n=8)$, the passes clustered around an average $(x, y)$ table location of $(0.24 \mathrm{~m}, 0.62 \mathrm{~m})$. When a participant on B side of the table passed ( $n=9$ ), the passes clustered around an average $(x, y)$ table location of $(0.33 \mathrm{~m}$, 0.18 m ).

## Model

The current study had two overall aims. The first aim was to identify the behavioral dynamics that underlie a continuous joint action pick-and-place task, in which two participants had to move objects from one tabletop location to another either alone or by passing the object to another co-actor. Our second aim was to develop a behavioral dynamics model that can characterize the joint action behaviors and choices (pickup or not; pass or not) of the participants engaged in during the joint action pick-and-place task. With respect to this aim we developed a model of both the participant's movement in the task space and their decisions to both pick up the object when it appears on a given trial and to pass the object.

## Movement Dynamics

In order to model the dynamics of each participant's, henceforth agent, hand movements throughout the task, a task specific parameterization of the Fajen and Warran model of human locomotory navigation was employed (Fajen \& Warren, 2003, 2004, 2007; Warren \& Fajen, 2008). This model has also been extended to model single actor pick and place behavior (Lamb et al., under review). In the current context, the model characterizes a heading direction or
angle, $\varphi_{i}$, of an agent's hand or end-effector (where each agent is indexed by $i$ ) during each task trial was defined by

$$
\begin{equation*}
\ddot{\varphi}_{i}=-b_{g_{i}} \dot{\varphi}_{i}-k_{g_{i}}\left(\varphi_{i}-\theta_{g_{i}}\right)\left(e^{-c_{1} d g_{i}}+c_{2}\right) \tag{1}
\end{equation*}
$$

where $\dot{\varphi}_{i}$, and $\ddot{\varphi}_{i}$, correspond to the velocity and acceleration of the agent's end-effector heading angle, respectively, and $b$ and $k$ are damping and spring/stiffness terms, such that $-b_{g_{i}} \dot{\varphi}_{A_{i}}$ acts as a friction force on turning rate, and the function $-k_{g_{i}}\left(\varphi_{i}-\theta_{g_{i}}\right)$ operates to minimize the difference between the current heading angle, $\varphi_{i}$, and the angle $\theta_{g_{i}}$, of the corresponding subtask goal/target location (i.e., the pickup location for pickup movements, the release/pass location for passing movements, and the target/drop-off location for target movements). The distance of the agent $i$ 's end effector to the current goal location is defined by $d_{g_{i}}$. The presence of the factor $\left(e^{-c_{1} d g_{i}}+c_{2}\right)$ in the second addend of the right-hand side introduces an exponentially decaying function characterized by a constant offset parameter $c_{2}$ and an exponential decay rate, which is a function of the constant parameter $c_{1}$ and the Euclidean distance, $d_{\mathrm{g}}$, between an agent's current hand location and the current goal location. The parameter $c_{2}$ ensures that the rate of change in heading direction never goes to zero (Fajen \& Warren, 2003). Note that the parameters $\theta_{g_{i}}$ and $d_{g_{i}}$ change continuously as the position of the agent's hand/end-effector moves through the task space.

Finally, in order for to capture the non-constant velocity profile observed in participants, $v_{i}$ is introduced to characterize the movement velocity of the agent's endeffector (hand). $v_{i}$ is defined by means of the additional $2^{\text {nd }}$ order differential equation

$$
\begin{equation*}
\ddot{v}_{i}=-b_{v_{i}} \dot{v}_{i}-k_{v_{i}}\left(v_{i}-C_{v_{i}}\left(1-e^{-d_{g_{i}}}\right)\right) \tag{2}
\end{equation*}
$$

where $b_{v_{i}}$ and $k_{v_{i}}$ act as damping and stiffness terms on the rate of change of $v_{i}$, which increases and decreases as a function of the target (goal) distance, $d_{g_{i}}$. When the agent's end-effector or hand is far away from the target location (1-$\left.e^{-d g_{i}}\right) \approx 1$ and $v_{A}$ increases. As the distance to the goal location decreases, however, $\left(1-e^{-d g_{i}}\right)$ approaches zero and $v_{i}$ decreases accordingly. $C_{v_{i}}$ is a constant parameter that specifies the maximum velocity in $\mathrm{m} / \mathrm{s}$, such that the same equation can be used for a wide range of different movement distances, with differential peak velocities resulting for shorter and longer distances (see Lamb et al., under review for more details on this velocity function).

## Action Selection Dynamics

In the experimental task there are two task defined choices. First, one of the agents must choose to pick up the task object while the other agent chooses to stay out of the way. Second, once an object is picked up, the agent with the object must decide to either take the object to the goal location or pass it to their co-actor. In both cases, the decision can be
characterized as a selection between action modes or affordances, i.e. pick up the object or wait and take the object to goal or pass. Moreover, previous research using a nonrandom pick-and-place task paradigm suggests that recent action modes may affect the current action mode selection (Lamb et al., Under Review). As a result, in the current context the action mode selection dynamics may be captured by

$$
\begin{equation*}
\dot{x}_{l}=-\alpha_{j_{i}}+x_{i}-x_{i}^{3} \tag{4}
\end{equation*}
$$

where $x_{i}$ represents the state variable for action section (i.e., affordance mode) of the previous action selection process and $\dot{x}_{l}$ is the action selection state variable for the current trial. $\alpha_{j_{i}}$ corresponds to the specific subtask action mode and agentnormalized E/A ratio where the decision to pick up an object can be defined for Agent 1 by

$$
\begin{equation*}
\alpha_{s_{1}}=\left(\sigma_{s_{1}}-\frac{d_{g_{s_{1}}}}{R_{1}}\right) \delta_{s_{1}}-\left(\sigma_{s_{2}}-\frac{d_{g_{s_{2}}}}{R_{2}}\right) \delta_{s_{2}} \tag{5}
\end{equation*}
$$

and for Agent 2 by

$$
\begin{equation*}
\alpha_{s_{2}}=\left(\sigma_{s_{2}}-\frac{d_{g_{s_{2}}}}{R_{2}}\right) \delta_{s_{2}}-\left(\sigma_{s_{1}}-\frac{d_{g_{s_{1}}}}{R_{1}}\right) \delta_{s_{1}} \tag{6}
\end{equation*}
$$

where $d_{g_{s_{i}}}$ is the distance from current location of the $i^{\text {th }}$ agent's end effector to the disc's location. Similarly, the decision to pass was defined by

$$
\begin{equation*}
\alpha_{p_{i}}=\left(\sigma_{p_{i}}-\frac{d_{g_{p_{i}}}}{R_{i}}\right) \delta_{p_{i}} \tag{7}
\end{equation*}
$$

$d_{g_{p_{i}}}$ is the distance of the agent's resting end-effector (hand) location to the target location, and $R_{i}$ is a measure of the agent's maximal preferred reach. In both equations equation, $\sigma_{j_{i}}$ and $\delta_{j_{i}}$ are constant scaling factors. In Eq. 5, 6, and 7, $d$ is a subtask action mode parameter that identifies the state of the subtask action relevant environmental property.

## Model Simulation

To determine whether systems defined by the movement trajectory dynamics (Eq. 1 and 2) and the action selection dynamics (Eq. 4, 5, 6, and 7) of the current model could complete the task independently complete the current pick-and-place task, a MATLAB (2016a) simulation was conducted. A flow diagram illustrating the structure of the simulation is provided in Figure 3. The simulated environment consisted of a $1.5 \mathrm{~m} \times 0.89 \mathrm{~m}$ rectangular space matching the experimental table's dimensions. The simulation target and disc locations were initialized in the same manner as in the experimental task. 10 different simulation sequences were conducted, with each simulation sequence consisting of 400 trials. Each simulation sequence
was initialized with the same pickup/target order used for a participant pair. The passing location centers corresponded to the observed passing location centers for each participant pair. Cluster centers corresponding to each participant pair were used to initialize the simulation sequence based on that pair's appearance/target order. Experimentally observed within pair variability in pass locations was likely due to the many complex interactions from which this passing behavior emerges (Holden, 2002; 2005; Stephen \& Mirman, 2010). However, in our simulations this variability is produced by a sequence of random values generated from a lognormal distribution that were added to the passing location centers in order to produce a pass location distribution that was similar to the experimentally observed data.

For each action selection the pickup and passing solutions to parameter equations, action selection dynamics, Eq. 4, were integrated for 1500 steps using the MATLAB ODE45 function with the end state of the integration used to drive the pickup and pass decisions (and return to rest position). The output state of the action selection equation was stored as an input for integration of the action selection equation in the next trial ( $x=0$ for the first trial in a sequence). Heading angles were initialized in the cardinal direction of the next subtask goal, e.g. pick up to target trajectories initialized with a heading angle of $0^{\circ}$ heading directly to the right side of the table. Random noise was added to the initial angle from a uniform distribution with $\mathrm{min} / \mathrm{max}$ values of $\pm 17^{\circ}$. The movement dynamics, Eq. 1 and 2, were integrated for each subtask movement using the Euler integration (. 01 time step), with integration terminated when the model location was within 4 cm of the target location. Random noise was added to the model heading direction, $\varphi_{i}$, at each time step of the integration using a uniform distribution with $\mathrm{min} / \mathrm{max}$ values of to $\pm 1.14^{\circ}$.


Figure 3: Structure of simulation. Eq. 1 and 2 are implemented in the upper loop and Eq. 4 through 7 are implemented in the lower goal selection loop.
An example simulation run is illustrated by the bottom heat-map in Figure 2. For all simulation runs, the simulation agents successfully completed the pick-and-place task within the task constraints. All simulation agent trajectories
remained within the task space and subtask trajectories were within the same regions as those produced by human participants. Simulation agent trajectories exhibited less variability as can be seen in figure 2, though the importance of this variability for a human co-actor engaged in the task is something that will require further research. To the extent that it is relevant, this variability may be replicated in an artificial agent implementation of the model by the addition of noise terms, through coupling to the human agent, and possibly by noise introduced by the agent's hardware instantiation (e.g. motor variability in a robotic system).

The simulation agents were also able to spontaneously select between picking up the object or not and between passing the object or completing the task alone in a manner similar to the real participants. For pickup trajectories it is notable that pickup decisions may be made and changed as co-actors move or do not move throughout the task space. If both simulation co-actors are roughly the same distance the pickup location at the beginning of a trial, noise and velocity profile variations still results in just one agent picking up the object. If a simulation agent picked up an object, that agent always passed for the farthest target and often did for the second farthest, with the decision to pass for this target fluctuating due to previous pass decisions and noise in the system, matching experimentally observed participant behaviors.

## Conclusions

The current model is useful for providing insight into how complex movement and decision dynamics might emerge from a system given relatively simple information structures. Notably, the model does not assume the need to understand or predict co-actor intentions or beliefs. This makes the model an ideal candidate for implementation in an artificial agent system that can interact in real-time with human coactors but does not have access to sophisticated sensory or computational systems for interpreting high level cognitive states. We are currently in the process of implementing a version of this model in virtual and robotic systems to test with human co-actors in order to validate the capabilities of behavioral dynamics models applied in this way.

## Acknowledgments

This research was funded by The National Science Foundation (NSF\#1513801) and National Institute of Health (R01GM105045-01).

## References

Fajen, B. R., \& Warren, W. H. (2003). Behavioral dynamics of steering, obstable avoidance, and route selection. Journal of Experimental Psychology: Human Perception and Performance, 29(2), 343.
Fajen, B. R., \& Warren, W. H. (2004). Visual guidance of intercepting a moving target on foot. Perception, 33, 689716.

Fajen, B. R., \& Warren, W. H. (2007). Behavioral dynamics of intercepting a moving target. Experimental Brain Research, 180(2), 303-319.
Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.
Graf, M., Schütz-Bosbach, S., \& Prinz, W. (2009). Motor Involvement in Action and Object Perception Similarity and Complementarity. In S. Semin \& G. Echterhov (Eds.), Grounding sociality: Neurons, minds, and culture. NY: Psychology Press.
Lamb, M., Kallen, R. W., Harrison, S. J., di Bernado, M., Minai, A. A., \& Richardson, M. J. (Under Review). Behavioral Dynamics of Object Movement and Passing During a Joint-Action Pick and Place Task.
Mark, L. S., Nemeth, K., Gardner, D., Dainoff, M. J., Paasche, J., Duffy, M., \& Grandt, K. (1997). Postural dynamics and the preferred critical boundary for visually guided reaching. Journal of Experimental Psychology: Human Perception and Performance, 23(5), 1365-1379. https://doi.org/10.1037/0096-1523.23.5.1365
Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
Richardson, M. J., \& Kallen, R. W. (2016). SymmetryBreaking and the Contextual Emergence of Human Multiagent Coordination and Social Activity. In Contextuality from Quantum Physics to Psychology (pp. 229-286).
Richardson, M. J., Marsh, K. L., \& Schmidt, R. C. (2010). Challenging the egocentric view of coordinated perceiving, acting, and knowing. The Mind in Context, 307-333.
Rizzolatti, G., \& Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience, 27, 169-192. https://doi.org/10.1146/annurev.neuro.27.070203.144230
Sebanz, N., \& Knoblich, G. (2009). Prediction in joint action: what, when, and where. Topics in Cognitive Science, 1(2), 353-367 https://doi.org/10.1111/j.1756-765.2009.01024.x
Warren, W. H., \& Fajen, B. R. (2008). Behavioral Dynamics of Visually Guided Locomotion. In A. Fuchs \& V. K. Jirsa (Eds.), Coordination: Neural, Behavioral and Social Dynamics (pp. 45-75). Springer Berlin Heidelberg.
Warren Jr., W. H., \& Whang, S. (1987). Visual guidance of walking through apertures: Body-scaled information for affordances. Journal of Experimental Psychology: Human Perception and Performance, 13(3), 371-383.
Washburn, A., Evans, J., Lamb, M., Kallen, R. W., Harrison, S. J., \& Richardson, M. J. (2015). Behavioral Dynamics of a Joint-Action Object Movement and Passing Task. Studies in Perception and Action XIII: Eighteenth International Conference on Perception and Action, 81.

# Analogies Emerge from Learning Dyamics in Neural Networks 

Andrew Lampinen (lampinen@stanford.edu)<br>Department of Psychology, Jordan Hall, 450 Serra Mall, Stanford CA 94305<br>Shaw Hsu (cshawhsu@stanford.edu)<br>Department of Biophysics, James H. Clark Center, 318 Campus Dr., Stanford CA 94305<br>James L. McClelland (mcclelland@stanford.edu)<br>Department of Psychology, Jordan Hall, 450 Serra Mall, Stanford CA 94305


#### Abstract

When a neural network is trained on multiple analogous tasks, previous research has shown that it will often generate representations that reflect the analogy. This may explain the value of multi-task training, and also may underlie the power of human analogical reasoning - awareness of analogies may emerge naturally from gradient-based learning in neural networks. We explore this issue by generalizing linear analysis techniques to explore two sets of analogous tasks, show that analogical structure is commonly extracted, and address some potential implications.


Keywords: neural networks; structure learning; representation; analogy; transfer;

## Introduction

Analogical transfer is often considered an essential component of "what makes us smart" (Gentner, 2003). However, there is a tension in the literature - Detterman (1993) has declared that "signifient transfer is probably rare and accounts for very little human behavior." Yet other authors have found that in some cases analogical transfer between superficially dissimilar systems can be so natural that it may not even require explicit awareness of the analogy (Day \& Goldstone, 2011). How can we reconcile these viewpoints?

One feature that often separates the researchers with these opposing viewpoints is the type of tasks and transfer they consider. When Detterman (1993) says that the manipulations necessary to show transfer have "the subtlety of [a] baseball bat", he cites work like that of Gick \& Holyoak (1980) which shows the difficulty of rapidly making an explicit mapping between two superficially disparate domains to explicitly solve a problem. By contrast, the Day \& Goldstone (2011) experiments show transfer when participants learn about a system by interacting with it over a longer period of time, and then transfer is measured implicitly on an analogous system. We believe that this distinction between fast-explicit analogical transfer and slower-potentially-implicit analogical transfer may explain much of the disagreement in the literature. (See also Bransford \& Schwartz (1999).)

Previous work has shown that neural networks can provide a good model for "slow" analogical transfer in domains as broad as artificial grammar learning (Dienes et al., 1999) and verbal analogies Kollias \& McClelland (2013). In particular, one line of work shows that neural networks are capable of extracting analogous structure from knowledge domains
that are completely non-overlapping in their inputs and outputs (Hinton, 1986; Rogers \& McClelland, 2008). In other words, if you train a neural network to solve two identical tasks, using separate sets of inputs and outputs but sharing the hidden units, in some cases it will generate representations that reflect the analogy (i.e. analogous items will generate more similar patterns of activity in the hidden units than non-analogous items) (Rogers \& McClelland, 2008). This can lead to the ability to correctly make analogical inferences about items not explicitly taught (Hinton, 1986). This extraction of shared structure sets neural networks apart from simple forms of statistical pattern recognition (Rogers \& McClelland, 2008) such as linear data analysis techniques like PCA.

Furthermore, recent work has shown that neural networks can show benefits of training on multiple tasks (Dong et al., 2015; Rusu et al., 2015, e.g.). Even a small amount of learning on distinct but related tasks has been shown to improve performance. For example, training a natural language translation system on image captioning and autoencoding improves translation performance (Luong et al., 2016). Learning on numerous language translation pairs can even give generalization without further training to unseen language pairs (Johnson et al., 2016). We suggest that these benefits may be due to neural networks ability to extract shared structure. Because human experience is filled with distinct tasks that share common elements (language, various perceptual modalities, etc.) understanding the way that structure is learned across tasks may be essential to understanding human intelligence and building better artificial intelligence systems.

However, we have little understanding of how, why, or when neural networks are able to extract structural analogies from their training data. Here, we describe a preliminary investigation into this question, and in the process describe a new approach to analyzing neural network representations that may yield more general insights. We begin with a very simple instantiation of a task with analogous structure.

## A Simple Task

In the original work of Hinton (1986), a neural network was taught to answer queries about the structure of two perfectly analogous family trees (one English and one Italian, see fig. 5), and was shown to generate representations that extract the analogy, in the sense that analogous people from differ-
ent families are represented similarly. Here, we pare this task down to its barest essentials: two perfectly analogous domains with separate inputs and ouputs. For our task, the inputs can be thought of as the set of letters $\{R, L, \rho, \lambda\}$, and the outputs as $\{P, D, S, \pi, \delta, \sigma\}$. The task can be seen as mapping an input letter onto the letters that it can follow (e.g. "R" can follow "D" as in "draw," but cannot follow "S"), where there is an analogy between the Latin and Greek letters. See below for the input-output (I/O) mapping:

|  | $P$ | $D$ | $S$ | $\pi$ | $\delta$ | $\sigma$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| $R$ | 1 | 1 | 0 | 0 | 0 | 0 |
| $L$ | 1 | 0 | 1 | 0 | 0 | 0 |
| $\rho$ | 0 | 0 | 0 | 1 | 1 | 0 |
| $\lambda$ | 0 | 0 | 0 | 1 | 0 | 1 |

When and how does a neural network extract the analogous structure across the domains in this simple task?

## Methods: Linear Networks?

There have been recent developments in the theory of linear neural networks which show that the process of learning is entirely driven by the Singular Value Decomposition (SVD) of the input-output correlation matrix (Saxe et al., 2013). The SVD can be seen as breaking the structure of the task into individual "modes" - linear structures in the dataset, somewhat like components in PCA. Specifically, a mode consists of an input pattern (which can be interpreted in this case as the input letters the mode responds to), a singular value (which roughly corresponds to the amount of variance explained by this mode), and an output mode (the output letters produced by the given pattern on the inputs). For example, see Fig. 1 for the SVD of the I/O mapping for the letter task above.

This decomposition tells us more about the task structure the network is using. There are three modes in the SVD. The first (left output mode/top input mode) represents the difference between the Latin and Greek letters, so it is positive for the Greek inputs and negative for the Latin outputs, and is positive for the Greek outputs and negative for the Latin outputs. The next two components represent the distinctions between the letters R and L , and the letters $\rho$ and $\lambda$, respectively. Saxe et al. (2013) showed these results have implications for the learning of non-linear networks as well, so linear neural networks can be a more tractable place to analyzee learning dynamics. In addition, using the I/O SVD allows the discovery of representational components which are distributed across units, so it is more general than simply examining what aspects of the task individual hidden units represent, or examining the weight matrices directly. Thus one might hope to answer our questions in a linear framework.

However, linear networks cannot represent analogous structure from non-overlapping inputs and outputs at convergence. With non-overlapping inputs and outputs, the I/O correlation matrix is block diagonal, and the SVD modes will thus occur within blocks (this is why in Fig. 1 the modes showing separation between the letters in each domain have
no input or output weights to the other domain). ${ }^{1}$ Thus, since the final representational components that a linear network learns are precisely the components of the SVD (Saxe et al., 2013), there will be no sharing of structure across domains.

Furthermore, the optimal rank $k$ approximation to a matrix is to take the top $k$ components from the SVD (Mirsky, 1960). If a linear network's hidden layers are restricted to rank lower than that of the I/O correlation matrix, detail within the domains will be lost. Thus a linear neural network cannot solve the task perfectly if any of its hidden layers has a number of units smaller than the rank of the I/O correlation matrix. By contrast, a non-linear network can exploit the analogy between the domains to find more parsimonious solutions. Is there a way to leverage linear insights in the non-linear case?

## Methods: A Linearized Approach

As we shall see, while a linear network cannot extract the analogous structure from the task, inserting a single nonlinearity after the output layer can allow it to do so. In the case that the non-linearity is a sigmoid, this essentially reduces the problem to logistic regression; here we will use rectified linear units in our analysis because their structure makes the output patterns more intuitively interpretable. Once this almost-linear network has solved the problem, consider its outputs immediately prior to the non-linearity. These are produced by the linear part of the network, and together with the non-linearity suffice to produce the desired outputs. We can use these to turn the problem into a linearly analyzable one - simply treat these pre-nonlinearity outputs as outputs of a linear network. Then the problem becomes susceptible to the types of linear analyses discussed above.

Thus we trained a neural network with a single hidden layer (4 units) and a single non-linearity (a rectifier at the output layer) to solve this task. See fig. 3 for a diagram of the network. No biases were used, weights were initialized uniformly between 0 and 0.1 , all training was done by Stochastic Gradient Descent (i.e. in each epoch the data are presented one at a time in a random order, and the weights are updated after each data point) with $\eta=0.01$ for 500 epochs.

## Results

The solution that the nonlinear network discovers the majority of the time (about $75 \%$ ) is to output the same pattern on both sets of output units, but offset the "incorrect" domain sufficiently negative so that it is hidden by the rectified, thus the task that the linear portion of the network is effectively performing at convergence is:

|  | $P$ | $D$ | $S$ | $\pi$ | $\delta$ | $\sigma$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $R$ | 1 | 1 | 0 | 0 | 0 | -1 |
| $L$ | 1 | 0 | 1 | 0 | -1 | 0 |
| $\rho$ | 0 | 0 | -1 | 1 | 1 | 0 |
| $\lambda$ | 0 | -1 | 0 | 1 | 0 | 1 |

[^381](Note that the network can actually map the first element of one domain onto either element of the other. We discuss the solution shown here one for clarity, the other just shuffles some rows and columns.)

The SVD of this linearized mapping shows a rank 2 solution (see fig. 2). The first component is similar to the first component of the regular SVD, in that it reflects the separation of the domains, but the second component collapses the other two components of the linear SVD. In other words, the analogy has been learned - the network is using the parallels between the two tasks to reach a more parsimonious solution. It is able to incorporate the analogy into its computations by allowing both the sets of outputs to vary, and simply suppressing the outputs from the "wrong" domain for its current task.

Because this solution is rank 2, a non-linear network with two hidden units should be able to solve the task, whereas a linear network will require three. We have verified these results empirically for this task. Thus the ability of a non-linear neural network to extract common structure from multiple tasks can allow it to find more parsimonious (i.e. lower-rank) solutions. We would like to highlight this point: the representation of the analogy in the SVD is not purely epiphenomenal - it makes a more parsimonious solution possible.

## Evolution of the I/O Mappings

When a non-linear network has only two hidden units, it must extract the analogy to be able to solve the task, but with more hidden units there are a variety of solutions that could potentially emerge (such as just learning the mapping of each input to its output pattern independently). However, our network extracted shared structure on about $75 \%$ of the runs we conducted (as measured by more than $20 \%$ score on the crossprojection metric described below). What drives this fairly consistent extraction of analogy? In this section we consider the evolution of the outputs over the course of learning.

The output structure of the network goes through a fairly consistent progression, which we will describe qualitatively at various key stages (the exact values depend on the initialization, so the matrices here are approximations to within about $\pm 0.1$ ). The outputs begin as small positive numbers, approximately 0 (because the weights are initialized uniformly between 0 and 0.1 ). Next, the network captures the base rate activations of each output unit, around epoch 75. (Note that this is already accounted for in the SVD, because the output variables are centered before computing the SVD).

$$
\text { base rates }=\left[\begin{array}{cccccc}
0.5 & 0.25 & 0.25 & 0.5 & 0.25 & 0.25 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0.5 & 0.25 & 0.25 & 0.5 & 0.25 & 0.25
\end{array}\right]
$$

Then the network captures the existence of the two domains but not the structure within them (around epoch 140). This corresponds to the first component of either SVD. Up to this point, a linear network follows a similar learning trajectory.

$$
\text { base rates by domain }=\left[\begin{array}{cccccc}
1 & 0.5 & 0.5 & 0 & 0 & 0 \\
1 & 0.5 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0.5 & 0.5 \\
0 & 0 & 0 & 1 & 0.5 & 0.5
\end{array}\right]
$$

Finally it learns the internal structure of the domains (they are not learned at exactly the same time, which is learned first depends on the initilization). Around epoch 400 it has solved the task completely, with some sort of offset structure in the non-linear case, or without in the linear case:

$$
\text { solution with offsets }=\left[\begin{array}{cccccc}
1 & 1 & 0 & 0 & 0 & -1 \\
1 & 0 & 1 & 0 & -1 & 0 \\
0 & 0 & -1 & 1 & 1 & 0 \\
0 & -1 & 0 & 1 & 0 & 1
\end{array}\right]
$$

For most of the learning process, the networks are extracting similar structure, so one might expect that even the linear network would show some representation of the analogy at intermediate stages of learning. Indeed, once the base rates by domain are learned, both the linear and non-linear networks begin to extract the analogy between the domains. See fig. 4 for a plot of how much each domain's input mode projects to the other domain's output mode, i.e. "cross-talk" between the domains. This is a simple measure of the extent to which the network is extracting shared structure. However, while both networks develop some representation of the analogy initially, this activity extinguishes rapidly in the linear network, while it persists in the non-linear network.

Why do both networks show some representation of the analogy initially? We will analyze this in the linear case. At the stage when the base rates by domain have been learned, adding a little bit of shared structure actually reduces meansquared error (MSE). If the network moves from the base rates by domain pattern to the pattern shown below, the small increase in MSE from the $\pm 0.1$ values is more than offset by the decrease from splitting the 0.5 values into 0.4 and 0.6.

$$
\left[\begin{array}{cccccc}
1 & 0.6 & 0.4 & 0 & 0.1 & -0.1 \\
1 & 0.4 & 0.6 & 0 & -0.1 & 0.1 \\
0 & 0.1 & -0.1 & 1 & 0.6 & 0.4 \\
0 & -0.1 & 0.1 & 1 & 0.4 & 0.6
\end{array}\right]
$$

Indeed, suppose there is a hidden unit which responds differentially within the domains (as they all will to some extent because of the random initialization). The updates of the output weights for this unit will point in the direction of analogy extraction once the base rates by domain have been learned. See below for the output error, hidden unit activity, and corresponding weight updates in the case that the hidden unit responds positively to the first element of each domain, and negatively to the other. ${ }^{2}$ (Note that the output weight updates for a hidden unit are proportional to the product of the output error and the hidden unit's activation.)

[^382]

Figure 1: SVD of I/O correlation matrix (colors are scaled to show qualitative features, red $=+$, white $=0$, blue $=-$ )


Figure 2: SVD of linearized I/O correlation matrix (colors are scaled to show qualitative features, red $=+$, white $=0$, blue $=-$ ). Note how Fig. 2a becomes Fig. 1a if the negative values are hidden by a nonlinearity.


Figure 3: Simple task network, showing a sample propagation of an input through the network with the single non-linearity at the output. (Circles represent inputs or fully connected units, squares represent non-linearities.)


Figure 4: I/O SVD component cross-projection (dot product between output mode of an SVD component and the response of the network to the other domain's input mode)


Figure 5: Family trees from Hinton (1986), (reproduced with permission).


Figure 6: Family tree task network (Circles represent inputs or fully connected units, squares represent non-linearities. Ellipses denote units omitted from the diagram - the hidden layer and all input and output groups had 12 units apiece.)

| output error |  |  |  |  | unit | unit output weight updates |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | + | - | 0 | 0 | 0 | + | 0 | + | - | 0 | 0 | 0 |
| 0 | - | + | 0 | 0 | 0 | - | 0 | + | - | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | + | - | + | 0 | 0 | 0 | 0 | + | - |
| 0 | 0 | 0 | 0 | - | + | - | 0 | 0 | 0 | 0 | + | - |
| net output weight update: |  |  |  |  | 0 | + | - | 0 | + | - |  |  |

Summing these updates captures the analogy between the domains. The network will exploit this analogy to reduce error, even if it must eventually discard it in the linear case.

## Reanalyzing Hinton's Family Tree Example

Next, we briefly turn our attention to the example of Hinton (1986). Hinton's task involves learning two isomorphic family trees, one English and one Italian (see fig. 5). This structure is taught implicitly by presenting a person (e.g. "Jennifer") and a relationship (e.g. "Father"), and training the network to produce the correct target person ("Andrew" in this case). There are 52 such relationships per family.

## Methods

Hinton used the same inputs for type of relationship for both families. To highlight the extraction of analogous structure we separated these into distinct input banks (these could be thought of as the English and Italian words for different relations, e.g. "uncle" vs. "zio" ). We also reduced his network down from 3 hidden layers to a single hidden layer with 12 units. Unlike the simple problem above, this problem is not linearly separable, so we included a non-linearity at the hidden layer as well as the output (see fig. 6). We trained this network by SGD with $\eta=0.005$ for 1000 epochs.

In a task which requires multiple non-linearities, we cannot perform as simple an analysis as in the earlier task. However, by definition each layer of the network has only a single nonlinearity, and so we can perform an analysis like the above on each layer. In this way we can understand something about the computations that layer is performing. However, the interpretation will not be as simple as above.

This difficulty is compounded by the complexity of the structure being learned in each family. In the simple problem above it was possible to "eyeball" the structure extraction, but here the structure is too rich. There are a variety of possible ways the families can be mapped onto one another (e.g. flipping the tree left to right and swapping all genders), and it's possible that the networks are extracting overlapping structure from several of these analogies. In this setting, how can we examine whether the network is learning the analogy?

As a first test of this, we looked for representation of the analogy in the input modes of the first layer SVD. To do this, we computed the dot product of each mode's weights for one family with that mode's weights for the other family, and then tested how significant this similarity was by comparing it to the null distribution generated nonparametrically by randomly permuting the columns of the input mode matrix 1000 times and computing the same dot product for each one.

We denoted a mode as showing significant extraction of the analogy if it showed a stronger similarity between the weights for the two families' inputs than $95 \%$ of its permutations did. We repeated this analysis across 100 network initilizations.

## Results

We found a great deal of analogous structure was extracted. The runs had a median of 4 modes showing significant analogous structure extraction, and all the runs had at least one significant mode (for comparison, if 5\% of the modes showed significant results by chance, we would still expect $54 \%$ of the runs to yield no significant results). To account for the symmetry of the tree under flipping, we repeated the same analysis after permuting the second family's input columns appropriately. Since the network has no way to distinguish the "regular" mapping from this "flipped" mapping during learning, we would expect to see a similar frequency of significant modes for each, and indeed the distributions are similar. Furthermore, the runs had a median of 6 modes showing significant extraction of either the regular or flipped mapping, and in all of the runs it had extracted 3 or more components that significantly represent one analogy or the other (if we assume 5\% false positives, we would expect results this extreme in only $0.01 \%$ and $3 \%$ of the runs, respectively). See fig. 7 . The frequency of analogy extraction suggests this may be a central feature of how neural networks solve tasks.

Although we have focused on broad analogies between the families here, we would like to note that analyzing the SVDs can give more detail. In some cases modes reflect an analogy only in the "person" inputs, or only in the "relationship" inputs. Within a family, analyzing the SVD modes can outline the structure the network is extracting, e.g. modes often appear which represent the gender of the target of a relationship like "mother". We have omitted these analyses due to length constraints.

## Disussion

We have outlined a new technique for analyzing neural network representations and their learning dynamics: analyzing the SVD of the "linearized" mapping at each layer (i.e. the mapping from the inputs to the pre-nonlinearity activity). This allows us to bring the power of linear analyses to bear on the rich phenomena that occur only in non-linear networks.

Using this technique, we have explored how a simple neural network can extract the analogy between simple tasks with non-overlapping inputs and outputs. We showed that, while a linear network cannot represent analogies, a single nonlinearity at the output layer can allow the network to represent the analogy, and that this structure emerges naturally (even in a linear network) from gradient descent once the base rates by domain have been learned. A linear network must discard this analogy to reach its optimal solution, but a non-linear network is able to retain it by simply offsetting the outputs to a sufficiently negative value, and does so the majority of the time in our results. Here we used rectifiers, but the same general solution is achievable with other nonlinearities.


Figure 7: How many of the input modes from the SVD showed significant projection onto the regular or flipped analogies (with null distribution for comparison)

We then broadened our approach to explore the family tree task originally proposed in Hinton (1986). Because this task is not linearly separable, we created a general network with two nonlinear layers, and applied our analysis to each layer. We found evidence of a great deal of extraction of two possible analogies between the families in the network (either the intended isomorphism between the family trees, or one in which one family tree was flipped left-to-right and genderreversed), and that networks seemed generally to be discovering elements of both analogies. Indeed, representation of the analogies seemed even more common than on the simpler task. On the simple task $25 \%$ of the networks showed no evidence of common structure extraction, but on the family tree task every network extracted at least three input modes that projected significantly onto one of the analogies.

These results suggest that sensitivity to analogy may be a natural feature of gradient based learning in nonlinear neural networks. This may underlie many of the "slow" analogical transfer effects we highlighted in the introduction. Furthermore, this may be a part of why learning multiple tasks facilitates more rapid learning and better performance in machine learning systems, and it may have important implications for cognition. The power and generality of human cognition may result from extracting common structure from the diverse but deeply related tasks we engage in throughout our lives.

## Future Directions

1. In our analysis we analyzed the input modes of the first layer and the output modes of the second layer. In the future it will be important to explore modes that map into and out of the hidden layer, and what they imply about the representations at the hidden layer. This would also allow us to apply this analysis to deep networks.
2. Learning representations that reflect analogies may provide amortized inference about potential analogical structure in the world. Can this support explicit analogical reasoning?

## Acknowledgments

This material is based upon work supported by the NSF GRF under Grant No. DGE-114747.

## References

Bransford, J. D., \& Schwartz, D. L. (1999). Rethinking Transfer : A Simple Proposal With Multiple Implications. Review of Research in Education, 24(1), 61-100.
Day, S. B., \& Goldstone, R. L. (2011). Analogical Transfer From a Simulated Physical System. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(3), 551-567. doi: 10.1037/a0022333
Detterman, D. K. (1993). The Case for the Prosecution: Transfer as an Epiphenomenon. In Transfer on trial: Intelligence, cognition, and instruction (pp. 1-24).
Dienes, Z., Altmann, G. T. M., \& Gao, S.-J. (1999). Mapping across Domains Without Feedback: A Neural Network Model of Transfer of Implicit Knowledge. Cognitive Science, 23(1), 53-82. doi: 10.1207/s15516709cog2301
Dong, D., Wu, H., He, W., Yu, D., \& Wang, H. (2015). MultiTask Learning for Multiple Language Translation. Acl, 1723-1732.
Gentner, D. (2003). Why We're So Smart. In Language in mind: Advances in the study of language and thought. (pp. 195-235).
Gick, M. L., \& Holyoak, K. J. (1980). Analogical Problem Solving. Cognitive P, 12, 306-355.
Hinton, G. (1986). Learning distributed representations of concepts. doi: 10.1109/69.917563
Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... Dean, J. (2016). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv, 1-16.
Kollias, P., \& McClelland, J. L. (2013). Context, cortex, and associations: A connectionist developmental approach to verbal analogies. Frontiers in Psychology, 4(NOV), 1-14.
Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., \& Kaiser, L. (2016). Multi-task Sequence to Sequence Learning. Iclr, 1-9.
Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. The Quarterly Journal of Mathematics, 11(1), 50-59. doi: 10.1093/qmath/11.1.50
Rogers, T. T., \& McClelland, J. L. (2008). A simple model from a powerful framework that spans levels of analysis. Behavioral and Brain Sciences, 31, 729-750.
Rusu, A. A., Gomez Colmenarejo, S., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., ... Hadsell, R. (2015). Policy Distillation. arXiv, 1-12. doi: 10.1038/nature 14236
Saxe, A. M., McClelland, J. L., \& Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Advances in Neural Information Processing Systems, 1-9.

# A Hierarchical Bayesian Model of Individual Differences in Memory for Emotional Expressions 

David Landy (dlandy@indiana.edu)<br>Indiana University, Bloomington<br>Department of Psychological and Brain Sciences, 1101 East 10th Street<br>Bloomington, IN 47405 USA

L. Elizabeth Crawford (Icrawfor@richmond.edu)<br>Jonathan Corbin (jcorbin@richmond.edu)<br>University of Richmond<br>Department of Psychology, 28 Westhampton Way<br>Richmond, VA, 23173

Keywords: category adjustment models; emotion memory; emotion perception; face perception; individual differences; Bayesian modeling


#### Abstract

When participants view and then reproduce simple objects that vary along a continuous dimension such as length or shade, or when they view images of faces that vary in emotional expression, their estimates tend to be biased toward the average value of the presented objects, a phenomenon that has been modeled as the result of a Bayesian combination of prior category knowledge with an imprecise memory trace (Corbin, Crawford \& Vavra, 2017; Huttenlocher, Hedges \& Vevea, 2000). Whereas previous work described a general cognitive strategy based on data aggregated across participants, here we examined individual differences in strategy. Thirty-six participants viewed and reproduced 496 morphed face stimuli that ranged from angry to happy. We found substantial variation in the bias patterns participants produced. Individuals' estimates were well fit by a model that posited attraction toward three categories, one at the happy end of the range, one at the angry end, and one that captured the entire range of presented stimuli, and by allowing the weight given to each category to vary by participant.


## Introduction

Memories are never pure. Memory of an object is determined not only by that individual object, but also by the set, or category, to which it belongs. Specifically, items tend to be remembered as being more like the typical (average) item in a set than they actually were. For example, Huttenlocher, Hedges and Vevea (2000) had participants view and immediately reproduce individual items that varied along a continuous dimension such as length, width, or shade. They manipulated the presented distribution of lengths, widths, and shades and found that estimates were biased toward the central value of the distribution shown. They proposed the bias is a byproduct of a Bayesian combination of a noisy, unbiased memory trace of the stimulus with a prior distribution that reflects the presented stimuli. Related Bayesian accounts have been developed to
account for bias in time perception (Jazayeri \& Shadlen, 2010), hue judgments (Olkkonen, McCarthy, \& Allred, 2014), and estimates of the sizes of familiar fruits and vegetables (Hemmer \& Steyvers, 2009). Here we extend this earlier work in two important ways. First, we apply this explanation to rich, socially relevant stimuli: faces that vary in emotional expression. Second, we model individual differences in how people rely on category knowledge when remembering facial expressions.

It is an open question whether memory for facial expressions can be characterized by the same principles that have been used to explain memory for length of a line. Facial expressions are socially meaningful and visually complex stimuli with which people have extensive prior experience, and unlike many other objects, faces are processed holistically (e.g., Maurer, Le Grand, \& Mondloch, 2002). Compared to simple geometric objects, it is more difficult to assess visual memory of real faces. One approach is to use morphing software to create gradations of faces that vary along a dimension of interest. By morphing pictures of the same actor making angry, neutral, and happy faces, we can create a continuum of emotional expression to be used in memory tasks like the immediate reproduction procedure described above. These morphed continua allow researchers to assess the degree to which a particular face is remembered as having an expression that is more or less happy or angry than it actually was.

Few studies have used face morphs to examine bias in memory for individual facial expressions (but see Haberman, Brady \& Alvarez, 2015; Haberman \& Whitney, 2009 for related work). In a study designed to examine the central tendency bias in face memory, Corbin, Crawford and Vavra (2017) ran several experiments in which participants viewed faces one at a time and, after each one, estimated its expression by adjusting a response morph. Estimates were consistently biased toward the central value of the stimulus distribution, whether it ranged from very sad to neutral, very happy to neutral, or moderately happy to moderately sad. Furthermore, the degree of this central tendency bias


Figure 1: Example stimuli. Shown are the original angry neutral, and happy faces used to generate the stimulus morphs as well as morphed images between angry and neutral and between neutral and happy.
increased with longer retention intervals between stimulus and response. Bayesian models predict such an effect because, as the trace memory distribution becomes noisier (i.e., more variable), the Bayesian combination of trace memory and category knowledge will give more weight to the category knowledge (see also Huttenlocher et al., 2000; Crawford, Huttenlocher \& Engebretson, 2000).

The Corbin et al. (2017) work was designed to allow for group-level conclusions and not for modeling of data from individual participants. This is typical of cognitive psychology, which usually characterizes the cognitive processing of a presumably generic, modal human mind without examining the variation between individuals. However, as we have noted elsewhere (Crawford, Landy \& Presson, 2014; Crawford, Landy \& Salthouse, 2016), that can lead to conclusions about aggregate tendencies that do not reflect the behavior or cognitive processing of any single individual. In fact, little is known about how people differ in their use of stimulus distributions to inform estimates of individuals. Building on the Corbin et al. findings, here we use Bayesian hierarchical modeling to examine both aggregate bias patterns and bias patterns at the level of individual participant. This approach allows us to estimate how each individual combines different category structures to arrive at estimates.

Emotional faces vary in physical dimensions such as mouth shape and brow orientation, as well as in affective significance, which can be processed automatically and unconsciously (e.g., Axelrod, Bar, \& Rees, 2015; Vuilleumier, 2005). A continuum of emotional expression is necessarily bound up with physical feature variations and we do not attempt to tease these apart. Instead, we capitalize on previous work (Corbin et al., 2017; Haberman et al, 2015, Haberman et al., 2019) showing that the continuum created by morphing emotional faces produces results that mirror those found in studies using simple dimensions such as size, color, or shade. This work suggests that, when shown a set of faces that vary on a morphed expression continuum, people are sensitive to the central tendency of the set along that dimension.

## Experiment

## Method

Participants Thirty-six (11 male) students from the undergraduate participant pool at the University of Richmond received course credit for participating.

Materials Images were from the NimStim face stimulus set ${ }^{1}$, a database of photographs of young adults depicting various emotional expressions. Sixteen models ( 8 male, 8 female) were chosen and the closed-mouth angry, neutral, and happy expressions of each were used to create the stimuli. Because in some cases, changes in hair position led to distracting artifacts in the morphed sets, we edited the initial images to maintain consistent hair placement. Using FantaMorph software (Abrosoft, 2002), each model's expressions were morphed from angriest to neutral to happiest, creating a set of 41 evenly distributed expressions that changed in $5 \%$ increments.

Procedure Each trial started with a crosshair at the center of the screen for 830 ms followed by a centrally presented single image frame taken from the morphed sets of faces and shown for 500 ms . The faces presented for study ranged from an angry expression (face \#5) to happy (\#35) and did not include the five most extreme images from either end of the continuum. After a blank screen ( 66 ms ), a response face of the same model was shown in the upper left hand corner of the screen. Participants were instructed to "use the right and left arrow keys to change the expression of the face to match the expression of the previous photograph." Pressing the right arrow key made the expression cycle through the entire morph (images $0-40$ ), cycling from happy to neutral to angry (or vice versa). Pressing the left arrow key cycled in the opposite direction. In a between subjects

[^383]manipulation, participants were randomly assigned so that the starting frame of the response morph was always the angriest face ( $\# 0$ ) or always the happiest face (\#40). Participants estimated each of the 31 facial expressions for each of the 16 models, for a total of 496 randomly ordered trials.

## Modeling

We modeled this data using a hierarchical Bayesian approach, simultaneously modeling individuals and group averages (see Figure 2). We assumed that each person was affected by a weighted combination of three potential biases: an overall inward bias toward the central category prototype ( $N$ ), and two attractive biases toward postulated extreme categories, representing the endpoints of happiness $(H)$ and anger ( $A$ ). We assumed equal variance for each category, and a logistic categorization boundary. Each category had a separate 'weight' $(W)$, which allowed the model to treat responses as the result of any number of categories from 0-3; best-fitting models uniformly predicted three categories (see Figure 5).

Explanations of bias are usually rooted in principles of Bayesian estimation: biasing responses toward a prior expectation reduces error (e.g., Feldman, Griffiths, \& Morgan, 2009; Huttenlocher et al., 2000). In this initial analysis, we simply assumed that each category attracted responses toward its center. This structure captures the relationships most often studied in category-based adjustment experiments, but abstracts away from the relationship between variability and category use-components of the model which have previously met with some predictive success (Crawford et al, 2016), but which were to the side of our primary concerns in this initial analysis

Model predictions were unbounded, but actual responses were bounded between an extreme happy face (valued as 1), on one end of the scale, and an extreme angry face (-1). To handle this, we assumed that when participants retrieved a face beyond the edges of the scale, they would select the most extreme face available.


Figure 2: Graphical model diagram of the Bayesian model. Rij is the response to stimulus j presented to subject i . The mean response is the sum of the stimulus value, sij, and three sources of bias, corresponding to the angry (A), neutral (N), and happy (H) prototypes. Each prototype has a weight (W) and a location (L). The category weights were potentially asymmetric, depending on the valence of the initial response slider (that is, whether the response face (f) was set to maximally angry ( -1 ) or maximally happy (1). The model only shows the first layer of fits: all top-level distributions were governed by population-level hyper parameters (see Table 1), which employed weak priors. In all cases, we assumed unbounded parameters to be normally distributed, and positive unbounded parameters to be gamma.


Figure 3: Aggregate and Individual Model Fits: (Left panel) Mean bias in response along with predictions averaged across participants. Errors reflect standard errors. (Right panel) Model fits for each individual participant. Use of all three categories is substantial, but starting side of the response strongly impacted the relative strength of these categories.


Figure 4: Individual model fits and data. Each dot is the bias in response to that stimulus, averaged across all times that participant viewed that expression. Each panel represents responses from one participant. Although different participants show quite different behaviors, the model treats each as a variation around a common theme of inward bias toward three weighted prototypes.

## Results

Aggregated and individual response patterns are plotted in Figures 3 and 4. As can be seen, there was a strong pattern, overall, of attraction toward the center of the distribution. However, this was tempered by strong outward trends among most individuals. These outward biases tended to be moderately strong, roughly comparable in size to the bias toward the center, and in some cases dominating it. Figure 3 shows the aggregated model fits across participants; Figure 4 the individual fits.

Parameters fit hierarchically are listed in Table 1 and include the weights attributed to each category and the locations of each category.

The magnitude of the individual differences in weights can be characterized by the posterior deviation parameters ( $\sigma$ ) governing weights. The $95 \%$ Highest Density Intervals for these excluded 0 (see Table 1), indicating that individuals differed in the weight given to these parameters (gamma shape parameters of roughly $<1$ correspond to high density around 0 ), and that these differences were not well explained by sampling noise.

Table 1: Priors and posteriors of population parameters. The $\mu$ values on the locations indicate mean locations of the categories, while the weight parameters have shape and rate values.

| Parameter | Population Prior | 95\% HDI |
| :---: | :---: | :---: |
| $W_{A}$ | $\Gamma$ (shape, rate) shape $\sim \Gamma(1,0.005)$ rate $\sim \Gamma(1,0.005)$ | $\begin{aligned} & \text { shape: }[11,168] \\ & \text { rate: }[6,92] \\ & \text { mean: }[1.6,2.4] \end{aligned}$ |
| $W_{N}$ | $\Gamma$ (shape, rate) <br> shape $\sim \Gamma(1,0.005)$ <br> rate $\sim \Gamma(1,0.005)$ | shape: $[39,196]$ rate: [42,175] mean: [.85,1.2] |
| $W_{H}$ | $\Gamma$ (shape, rate) shape $\sim \Gamma(1,0.005)$ rate $\sim \Gamma(1,0.005)$ | shape: 120,235$]$ rate: [57,124] mean: [1.6,2.6] |
| $L_{A}$ | $\begin{aligned} & N(\mu, \tau) \\ & \mu \sim N(-1,80) \\ & \tau \sim \Gamma(1,200) \end{aligned}$ | $\begin{aligned} & \mu:[-1.3,-1.15] \\ & \tau:[.2,2640] \end{aligned}$ |
| $L_{N}$ | $\begin{aligned} & N(\mu, \tau) \\ & \mu \sim N(0,80) \\ & \tau \sim \Gamma(1,200) \end{aligned}$ | $\begin{aligned} & \mu:[-0.06,-0.002] \\ & \tau:[68,560] \end{aligned}$ |
| $L_{H}$ | $\begin{aligned} & N(\mu, \tau) \\ & \mu \sim N(1,80) \\ & \tau \sim \Gamma(1,200) \end{aligned}$ | $\begin{aligned} & \mu:[1.225,1.325] \\ & \tau:[52,2040] \end{aligned}$ |
| tau | $\Gamma$ (shape, rate) <br> shape $\sim \Gamma(3,1)$ <br> rate $\sim \Gamma(3,1)$ | $\begin{aligned} & \text { shape: }[3600 \text {, } \\ & \text { 10000] } \\ & \text { rate: }[240,520] \end{aligned}$ |
| $\beta$ (side bias) | $\begin{aligned} & N(\mu, \tau) \\ & \mu \sim N(0,80) \\ & \tau \sim \Gamma(5,0.1) \end{aligned}$ | $\begin{aligned} & \mu:[0.01,0.02] \\ & \tau:[4,800,28800] \end{aligned}$ |



Figure 5: Simplex plot of the relative weights accorded to each category. A dot reflects a mean individual. Red indicates starting values on the happy side, blue on the angry side. Although in principle, the total weight could vary, in practice each individual showed a mean weight between 3.3 and 3.5 , making simplex plots a useful visualization of the three values.

One factor had a strong apparent impact on the weight given to the left and right categories: the starting location of the response. To quantify this effect, we modeled the left and right weights as symmetric, except for a mean shift determined by an individual splitting parameter. This splitting parameter was fit to individuals; the posterior fits are shown in Figure 5. The results suggest a moderate impact of start location on category weight, such that people more heavily weighted the category represented in the starting value.

## Discussion

Building on earlier work on inductive category effects on memory, we assume that estimates of an individual object combine an inexact memory trace of the object with knowledge of the set to which it belongs (e.g., Huttenlocher et al., 2000), producing estimates that are biased toward category prototypes. Such central tendency effects have been shown in studies using immediate reproduction tasks with simple stimuli that vary on one or two dimensions, such as size and shade (e.g., Crawford et al., 2001). Extending this work to more complex and socially relevant stimuli, Corbin et al. (2017) found that estimates of emotional expressions are also biased toward the center of the presented range of expressions, suggesting that participants used an inductively formed category to adjust estimate of faces.

Here we further examined the kinds of category structures involved in face memory and the degree to which individuals differed in their use of these structures.

As in previous work, estimates generally were biased toward the presented distribution's center (here a neutral expression). In addition, we found substantial variability between participants such that most participants were not well described by a model that treated estimates as resulting from adjustment toward a single, centrally located category. Good model fits at the participant level were achieved by positing that estimates could be adjusted toward two additional categories (centered on angry and happy values) and by allowing category weights to vary by participant. We note that this threecategory model reflects the structure that was used to generate the stimuli: pictures of faces that actors made when told to show happy, angry and neutral expressions.

Some of the difference in how participants weighted the different categories could be accounted for by the starting value of the response face, which was randomly assigned between subjects. On average, greater weight was given to the category that aligned with the starting position (either $100 \%$ happy or $100 \%$ angry). The effect of the starting value was not linear across the stimulus range, as would be expected by inadequate adjustment away from an anchor. Instead it appears that the starting value encouraged participants to rely more heavily on the closest emotion category. Although studies of inductive category learning typically focus on the distribution of test objects, this result suggests that response objects may also contribute to the category structure used during estimation.

It is common to analyze group-level data and describe the collective's average behavior, but this approach can miss meaningful variation in cognitive strategies used by individuals. Modeling responses at the individual level reveals similarities across participants as well as some systematic differences. From the current study, it is not known why people adopt the strategies that they do. The model's success in capturing the different data patterns produced by individuals makes it a valuable framework for future studies of how differences in cognitive, social, and affective processing may influence the reliance on categories when remembering emotional faces. The variation in bias that we observed suggests that models pitched at the level of group averages are likely to mislead us away from the best interpretations.

## References

Axelrod, V., Bar, M., \& Rees, G. (2015). Exploring the unconscious using faces. Trends in cognitive sciences, 19(1), 35-45. doi: 10.1016/j.tics.2014.11.003

Corbin, Crawford \& Vavra (2017). Misremembering emotion: Inductive category effects for complex emotional stimuli. Memory \& Cognition, pp. 1-8. doi: 10.3758/s13421-017-0690-7

Crawford, L. E., Huttenlocher, J., \& Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction? Psychological Science, 11, 280-284. doi:10.1111/1467-9280.00256
Crawford, L. E., Landy, D., \& Presson, A. N. (2014). Bias in spatial memory: Prototypes or relational categories? Proceedings of the 36th Annual Conference of the Cognitive Science Society. Quebec City, Quebec: Cognitive Science Society.
Crawford, L. E., Landy, D., \& Salthouse, T. A. (2016). Spatial working memory capacity predicts bias in estimates of location. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(9), 1434-1447. doi: 10.1037/xlm0000228
Feldman, N. H., Griffiths, T. L., \& Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. Psychological review, 116(4), 752-782. doi: 10.1037/a0017196

Haberman, J., Brady, T. F., \& Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. Journal of Experimental Psychology: General, 144(2), 432-446. doi: 10.1037/xge0000053
Haberman, J., \& Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. Journal of Experimental Psychology: Human Perception and Performance, 35, 718-734. doi: 10.1037/a0013899
Hemmer, P., \& Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. Psychonomic Bulletin \& Review, 16(1), 8087. doi: 10.3758/PBR.16.1.80

Huttenlocher, J., Hedges, L. V., \& Vevea, J. L. (2000). Why do categories affect stimulus judgment? Journal of Experimental Psychology: General, 129, 220-241. doi: 10.1037/0096-3445.129.2.220
Jazayeri, M., \& Shadlen, M. N. (2010). Temporal context calibrates interval timing. Nature Neuroscience, 13, 1020-1026. doi:10.1038/nn. 2590
Maurer D., Le Grand R., Mondloch C.J. (2002). The many faces of configural processing. Trends in Cognitive Sciences, 6, 255-260. doi: 10.1016/S1364-6613(02)01903-4
Olkkonen, M., McCarthy, P. F., \& Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. Journal of Vision, 14, 1-15. doi: 10.1167/14.11.5
Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. Trends in cognitive sciences, 9(12), 585-594. Doi: 10.1016/j.tics.2005.10.011

# Simulating performance in unconscious plagiarism 

Nicholas Lange (nick.lange@plymouth.ac.uk)<br>School of Psychology, Portland Square<br>Drake Circus, Plymouth University, PL4 8AA<br>United Kingdom

Timothy J. Hollins (thollins@plymouth.ac.uk)<br>School of Psychology, Portland Square<br>Drake Circus, Plymouth University, PL4 8AA<br>United Kingdom


#### Abstract

Studies of unconscious plagiarism have reported that people mistakenly include a partner's responses when trying to recall their own (recall-own task) and include own responses when trying to recall their partner's (recall-partner task). In a simulation, we tested if participants' memory performance at test, including source errors, can be explained by participants simply guessing items that come easily to mind. We show that guessing alone cannot account for the pattern of data participants show at test. Modifying the simulation by including memory for self-generated items allows us to replicate the pattern of responding in the recall-own but not the recall-partner task, even when we assume that participants in the recall-partner task strategically withhold more fluent items from report. This suggests that judgements of items' memory strength alone cannot explain performance in the unconscious plagiarism paradigm.


Keywords: source memory; free recall; unconscious plagiarism

## Background

In the standard unconscious plagiarism (or cryptomnesia) experiment (Brown \& Murphy, 1989), participants in groups take turns to generate solutions for a task. Following a delay participants are asked to complete a recall and/or a generatenew task. In the recall task, participants are asked to selectively recall the solutions they generated themselves, avoiding those generated by others in the group. In the generatenew task, participants are asked to generate novel solutions to the task, avoiding both previously self- and other-generated ones. Plagiarism errors (or source errors in the recall-own and generate-new task) are now solutions generated by other members of the group that participants falsely claim to have generated themselves, with plagiarism typically at abovechance rates for both the recall-own and generate-new task (Brown \& Murphy, 1989). More recently, Hollins, Lange, Berry, and Dennis (2016) showed that source errors in recall tasks are not limited to the recall-own task, but also occur during the recall of partner-generated items in the recall-partner task. Rather than participants being biased to simply claim ideas as their own, it appears that participants are simply confused about the source of the ideas they retrieve from memory (Hollins, Lange, Dennis, \& Longmore, 2015; Perfect, Field, \& Jones, 2009).

While source errors are typically treated as an instance of false memories, an alternative account is that they constitute accidental errors that occur by chance (Brown \& Murphy,

1989; Tenpenny, Keriazakos, Lew, \& Phelan, 1998). In the study phase, participants are asked to take turns generating responses to cues, such as category exemplars. Without further instruction to generate typical or atypical exemplars, it is likely that participants will first generate responses that are readily available to them, i.e. typical exemplars in the category. This would be in line with participants employing a fluency heuristic (Jacoby, Woloshyn, \& Kelley, 1989).

Brown and Murphy (1989) tested this non-memorial guessing account. They presented participants with the test phase of an unconscious plagiarism experiment without a preceding study phase. When treating this generation at test as recall from a study phase that participants did not participate in, participants still committed "source errors" to a high degree. This seems to suggest that reporting items based on fluency or the frequency or typicality of items could be responsible for source errors.

Critically, Brown and Murphy (1989) focused only on source errors, and only on source errors in the recall-own task. In the present paper, we adapted the unconscious plagiarism and anti-plagiarism paradigm as used in Hollins et al. (2015, 2016) for actions. We constructed guessing simulations that builds on Brown and Murphy but attempts to simulate performance across all task measures in both retrieval tasks. If source errors are in part the result of chance performance, the same would have to be true for the correct retrieval and the generation of novel items at test.

We constructed a base guessing model that samples, ignorant to study phase and task, from the possible items per category cue, with sampling weighted by the frequency or typicality of those items. In subsequent simulations, we modified this base guessing simulation by manipulating the number of items per cue available to participants at test, the memory for self-generated and partner-generated items, and the orientation towards self-generated or partner-generated items at test given the retrieval task.

## Experimental work

Unconscious plagiarism has been exclusively studied with verbalizable stimuli (for reviews see Perfect \& Stark, 2008; Gingerich \& Sullivan, 2013). We adapted the unconscious plagiarism paradigm with two retrieval tasks to motor actions to produce observed data. In this experiment, we asked par-
ticipants to take turns performing and observed actions with a partner in the study phase. In the test phase, participants then were asked to recall performed actions (equivalent to the Recall-own task for verbal material) or asked to recall observed actions (equivalent to the Recall-partner task).

## Method

Participants 40 members of the public participated for payment of $£ 12$. Three participants did not attend all sessions and their data were excluded from the analysis.

Procedure Participants were asked to attend two sessions in total, a day apart. For the first session, participants were paired and asked to take turns generating and acting out shapes with any part of their body or combinations of body parts. They were shown 15 shape cues $(=$, A, C, F, H, I, J, K, $\mathrm{L}, \mathrm{O}, \mathrm{P}, \mathrm{T}, \mathrm{V}, \mathrm{X}, \triangle)$ in total. Participants were cued with a printed label of each shape. Members of the pair took turns generating actions for each cue, interleaving performing and observing actions such that performing an action in response to a cue was followed by observing the other person perform an action in response to the same cue. Each participant generated a total of 4 actions per cue, resulting in 60 performed and 60 observed actions overall. Participants were told to observe their partners during partner-generation to avoid duplicating exemplars that had already been created for a cue. Participants observed their partner perform actions under a secondary task load for two-thirds of the shape cues. The assignment of shape cues to secondary task conditions was counterbalanced across participants. The focus of the present paper is on the control condition only, i.e. the one-third of actions participants performed and observed without secondary load, for the purposes of simulating guessing performance.

Participants returned the next day individually for a memory test. They were instructed to retrieve and re-perform either the actions they had generated themselves (Recall performed) or those they had observed their partner perform (Recall observed) the previous day. They were cued with the shape labels, and asked to re-perform as many actions as they could remember for each shape (free report). They were asked to avoid performing actions that did not comply with their retrieval task.

Preprocessing of observed data In the study phase, participants could commit two types of errors: self-plagiarism, that is repeating an item they had already generated for a particular cue, and other-plagiarism, repeating an item their partner had already generated for a cue. In the control condition, participants self-plagiarised on average 5.13\% ( $\mathrm{SD}=5.75 \%$ ) of items and plagiarised $9.73 \% ~(\mathrm{SD}=6.25 \%$ ) of partner-generated items. Items that both participants had generated at study for a cue were removed from analysis, since the source of the item, if retrieved, would be ambiguous.

## Results

Participants' mean performance in both retrieval tasks is shown by the bars in the figures in this paper. Correct source
retrieval was higher in the Recall performed than the Recall observed task, $t(34.80)=2.75, p=.009$. There was no evidence for source errors or intrusion errors being committed more frequently in one than the other retrieval task, $t(31.86)=1.29, p=.21$ and $t(35)=.04, p=.97$.

## Simulating frequency-based guessing

## Base guessing simulation

In the base frequency-based guessing simulation we extended the idea of a test phase without prior study phase tested by (Brown \& Murphy, 1989). We used a Monte Carlo procedure to simulate how many correct responses, source errors and intrusion errors participants would make if they were guessing and had just generated potential actions for each shape "on the fly" during the test phase, rather than genuinely retrieving them from what they had previously either seen or performed. We simulated the test phase of the experiment for each participant and each shape separately to take into account differences between individual participants, differing frequency profiles for the different shapes, and the typicality of individual items.

As a first step of the simulation process, we determined frequency norms for the different actions generated for each of the 15 shapes used in the experiment. We took into account all possible shapes all participants had generated across the experiment. Some actions were produced more frequently than others across participants, resulting in a frequency profile for each shape. Both self-plagiarised and partner-plagiarised items at study were included in the creation of these frequency profiles. For each shape, we converted these frequency profiles of the different actions into probability distributions, reflecting the relative probability that a particular action was produced for a given shape. For each shape, the probabilities across the shape cue summed to 1 to represent all possible actions for the shape.

We used these distributions as the basis for participants' guessing. Partner-plagiarised items were excluded from the distribution prior to guessing to match the analysis of the observed data. For each participant we sampled the number of items from each shape distribution that participants reported at test for that particular shape (excluding the partnerplagiarised items again). The sampling was weighted by the relative probability of the items, to implement that guessing was not random but biased by the frequency or typicality and therefore fluency of items. This sampling was done without replacement to match the experimental procedure of only retrieving an item once. Items were sampled sequentially, with the distribution re-normed after each draw. We repeated this for every participant, and ran the simulation over 500 iterations for stable estimates.

In addition to the items participants reported at test (correct responses, source errors, novel items), we now also have the same measures if participants were only guessing at test.

Figure 1 shows the results of the base guessing simulation (the full-length distribution indicated by the stars) relative to
the observed data. In both retrieval tasks, mere guessing may approximate the number of source errors but cannot account for the number of correct responses and intrusion errors. It is therefore unlikely that performance in this paradigm occurs simply because highly-frequent items are generated both at study and test, regardless of an influence of memory.

In the next section, we will modify this base guessing simulation by manipulating first the length of the frequency distribution, and then introducing memory and meta-cognition into the frequency-based guessing.

## Modifying the base guessing simulation

In the base guessing simulation we assume that participants' responses are based entirely on the overall frequency (or typicality) of items. This assumption leads to the following conditions for guessing: a) each participant has the entirety of each shape distribution available to them at retrieval, b) memory encoding in the study phase does not affect the frequency of items (i.e., there is no effect of memory) and c) participants' responding does not change with the instruction to retrieve items from one or the other source. In the next three steps, we therefore simulated the influences of the length of the distribution, memory effects and retrieval task orientation on participants' guessing. Strictly speaking, only the first modification still represents participants only guessing, i.e., responding without memory. The second modification introduces an effect of memory and the final modification an effect of metacognitive choices made at retrieval.

Length of the distributions In the base guessing distribution, the simulated participants sample their guesses from all possible ways a particular shape was produced in the experiment. This assumes that each participant has access to all possible ways a shape can be represented with the body that were produced throughout the experiment - this is a strong assumption that may inflate the number of novel items relative to items generated at study. It is more likely that each participant has only a subset of items for each shape cue available to draw on. In the first modification of the base guessing simulation, we therefore simulated the pattern of performance in the task if participants guess from frequency distributions for each shape that are a shorter, i.e., include fewer possible actions.

Table 1: Average number of items in the guessing distributions with shortened tails

|  | Number of items per cue |  |  |
| :--- | :--- | :--- | :--- |
| Length | Mean $(\mathrm{SD})$ | Max | Min |
| Full | $23.66(4.68)$ | 33 | 16 |
| 0.9 | $12.30(3.19)$ | 19 | 5 |
| 0.8 | $8.18(2.20)$ | 13 | 2 |
| 0.7 | $5.59(1.26)$ | 8 | 1 |
| 0.6 | $3.78(0.88)$ | 6 | 1 |

The probabilities associated with the items in the base


Figure 1: Observed data in bars (correct responses, source errors, intrusion errors) with $95 \%$ confidence intervals and data predicted by participants guessing with distributions of varying length relative to the full distribution (points)
guessing simulation sum to 1 , from most frequent items at higher probabilities to least frequent items with lower probabilities. We created shorter distributions by successively shortening the tail of each shape distribution, i.e., removing the least frequent items. This resulted in distributions representing the top $90 \%, 80 \%, 70 \%$ and $60 \%$ of items generated for each shape across participants in the study phase. The sampling procedure was otherwise identical to the one described above.

Table 1 shows the average number of items, as well as maximum and minimum number of items, that could be sampled across shapes for the different lengths of distributions. For the shorter distributions, in some cases the total number of guesses to be sampled was longer than the distribution to sample from. In those cases, the total possible number of items, i.e. all items in the shortened distribution, was sampled as a guess in lieu of the total number of responses participants in fact made in that case.

Figure 1 shows participants’ observed performance (bars) and simulated performance (points and lines) in both the Recall performed and Recall observed task. The stars indicate the sampling based on the full-length distributions for each shape, the remaining points the proportionally shorter distributions (relative to the full-length distribution). Comparing the simulated performance across the different lengths of the distribution shows that with shorter distributions, the number of novel items that are sampled during guessing decreases. There is only a minimal effect on the number of correct responses and source errors that are sampled.

While guessing even based on shorter distributions does not approximate performance in the Recall performed task, guessing based on drastically shortened distributions comes close to replicating the pattern of responding in the Recall observed task. Though note that the radically shortened dis-
tributions do not contain many items available for guessing. Naturally, these very short distributions not only contain only a minimum of novel items (hence the decrease in the sampled novel responses), they also do not contain many items participants generated in the study phase and hence do not drastically increase correct and source error responses.

Performance in the Recall-performed task (the original unconscious plagiarism paradigm) cannot be only the result of frequency-based guessing at test. In the Recall observed task, this type of guessing could potentially account for the pattern of responding. In the next step, we modified the distribution further by adding memory for items that were generated at study.
Memory after generation Pure guessing, here implemented by sampling based on the overall frequency or typicality of items, does not fully approximate performance in the memory test and therefore is not an explanation for unconscious plagiarism performance (when both correct responses and intrusions are considered alongside the number of source errors). In the next step, we tested if adding an effect of memory to the model by boosting the probability of items that were generated by participants could account for the pattern of data observed in the experiment.

We used the full-length distribution (rather than shortened distributions). We implemented memory for items by adding a second probability term to all items participants generated themselves (but not to items participants observed their partners perform - we added this modification in the final simulation). The additional probability terms for self-generated items were 0 (the base guessing simulation), $0.1,0.2,0.3,0.4$ and 0.5 . The final shape distributions were re-normed so all probabilities summed to 1 after this memory probability term was added to the prior probability of each item. Beyond the memory boost, the sampling procedure was identical to the one described in the previous simulations.

Figure 2 shows the observed data (bars) and the predicted responses based on the guessing simulation with memory boost. The memory boost results in good approximation of performance in the Recall performed task (even with the fulllength distribution used for guessing). This suggests that participants in the Recall performed task may simply successfully employ a fluency heuristic (Jacoby et al., 1989) by reporting items that are strongly represented at test by a combination of their base typicality and some memory.

In the Recall observed task, increasing the likelihood of generated (here: source error items) to be guessed leads to grave misfits of the pattern of data observed in the experiment. If participants were still simply reporting the most fluent exemplars at test, regardless of the task, the number of source errors (self-generated items with higher memory), should be higher than the number of correct responses (observed actions). This is clearly not what participants in the experiment are doing.

In the final modification, we therefore introduced a metacognitive modification to the simulation that has partici-


Figure 2: Observed data in bars (correct responses, source errors, intrusion errors) with $95 \%$ confidence interval and data predicted by participants guessing with self-generated items' probability to be sampled boosted by varying probabilities.
pants orient their report towards their retrieval task, i.e., deliberately withholding fluent items in the Recall observed task.

Orienting towards retrieval task Both the manipulation of the length of the distribution and boosting memory for self-generated items assumed that participants use a fluency heuristic in the test phase of an unconscious plagiarism task. Regardless of the the task instruction to retrieve self-generated or other-generated (here: observed) items, the fluency heuristic assumes that participants will base their responding entirely on what comes to mind at test. This means items associated with higher probabilities will be reported more readily, regardless of the retrieval task.

In a more nuanced approach, it is feasible that participants are able to regulate the memories they report (Marsh \& Bower, 1993; Hollins et al., 2016). In this case, participants in the Recall observed task could be able to withhold items that first come to mind from report if they assume that better memory/higher fluency would be indicative of self-generation and hence represents a source error. This type of source monitoring is based entirely on monitoring the memory strength of items, rather than any source features.

In the simulation, we implemented a task orientation by sampling double the total number of items a participant reported from each shape distribution. For the Recall performed task, we then used the top half of the sampled items as items guessed in the simulation. In the Recall observed task, we discarded the first few guesses (this is participants withholding items from report) and instead used the bottom half of guesses, the relatively less frequent items. The remainder of the simulation was identical to previous simulations, with sampling based on the full-length distribution.

Figure 3 shows the results for the frequency-based guessing if self-generated items are boosted in memory and participants in the Recall observed task withhold these items.


Figure 3: Observed data in bars (correct responses, source errors, intrusion errors) with $95 \%$ confidence intervals and data predicted by participants guessing with self-generated items' probability to be sampled boosted by varying probabilities.

The results for the Recall performed task are naturally equivalent to the simulation without retrieval task orientation, since in both cases the most frequent items are reported. For the Recall observed task, the simulated participants are now less likely to now report source errors, i.e. they successfully withhold those items. Rather than this boosting the correct retrieval (items they observed their partner perform), this modification only increases the number of novel items. Even with an orientation towards weaker items, a frequency-based guessing procedure with memory for self-generated items does not account for the observed pattern of data in the Recall observed task.

Remember, we implemented only increased memory for self-generated but not for observed items. It is possible that even small increases in memory for observed items could explain the correct responses in the Recall-observed task. In a final step modification, we therefore manipulated memory for observed actions. We sampled from distributions slightly limited in length ( 0.9 distribution from the length modification) and boosted memory for self-generated actions by 0.3 (the boost that most closely matches the pattern of responding in the Recall performed task). We boosted memory for observed actions by a probability term of $0,0.3 / 4,0.3 / 3,0.3 / 2$ and 0.3 , using the assumption that memory for self-generated items is not likely to be lower than memory for observed actions.

Figure 4 shows the results of the simulation and the observed data. Boosting memory for observed actions does not lead to a closer approximation of the data in the Recall performed task, in part because the effects of the shortened distribution and memory for observed actions both limit retrieval of novel items and increase retrieval of observed actions.

In the Recall observed task, with additional memory for observed actions, the number of correct responses observed in the experiment cannot be replicated. This may not be sur-


Figure 4: Observed data in bars (correct responses, source errors, intrusion errors) wit $95 \%$ confidence intervals and data predicted by participants guessing with self-generated items' probability to be sampled boosted by varying probabilities.
prising given the implementation of the task orientation in the simulation. We implemented participants' orientation towards observed actions in the Recall observed task as a withholding of the first few sampled items, in spirit of participants' performance being based on fluency and the interpretation of fluency alone. Increasing the probability of observed items now makes it more likely for those items to be sampled first, and therefore withheld. In other words, to replicate the pattern of observed data using a frequency-based sampling approach, memory for observed actions has to be lower than memory for self-generated actions, with the items participants observed needing to be of higher strength than novel items that were not generated. In the simulations, we did not achieve this balance. It is not clear if memory strength alone is sufficient to explain performance in the Recall observed task.

## Discussion

We adapted the extended unconscious plagiarism paradigm (Hollins et al., 2016, 2015) to motor actions. Participants took turns generating and observing actions in the study phase, and were asked to retrieve actions they performed themselves or actions they observed their partner perform in the test phase. We simulated performance in the task to test if guessing alone can account for the pattern of data we observed.

We simulated the experiment to test if frequency-based guessing can account for the observed results. This account is a variation of a fluency or memory strength account of unconscious plagiarism (Marsh \& Bower, 1993; Hoffman, 1997; Jacoby et al., 1989) and proposes that memory retrieval in the unconscious plagiarism paradigm is guided by the overall memory strength or availability of items at test. Items with higher memory strength are more likely to come to mind and hence be reported at test. We have shown that this approach
with modifications of length, memory and task orientation provides a reasonable description of the data in the Recall performed task. Notably, it does less well in accounting for the data in the Recall observed task.

Given the framework of this kind of memory strength account, the main difference between the retrieval tasks is that participants will do better to report highly-frequent items in one case (Recall performed task) and better to withhold them in another case (Recall observed task). In the Recall observed task, participants ideally report items of some memory strength. In our simulations we were not able to replicate that participants, in fact, are able to make correct responses in the Recall observed task that exceed source errors and intrusion errors.

While responding based on memory strength alone could explain performance in the Recall performed task, it is not sufficient to explain performance in the Recall observed task (for a similar conclusion using a signal detection approach, see Hollins et al., 2016).

There are two possibilities. Participants in the Recall observed task may simply be guessing. In particular if they can only generate very few items (or only very few items beyond self-generated items they remember and are potentially withholding from report), guessing without any memory boost may account for performance in the task. We showed that with very short distributions, performance in the task was approximated. Arguably, the shortest distributions that came closest in matching the pattern afford unrealistically few items to participants in the test phase.

Alternatively, performance in the unconscious plagiarism (and anti-plagiarism paradigm) may not be only based on the overall strength (fluency, familiarity or item memory) of items. In line with the source monitoring framework (Johnson, Hashtroudi, \& Lindsay, 1993) used to explain monitoring failures in other false memory paradigms, participants may judge source memory on a dimension separate to the overall memory strength. While memory strength alone may not allow participants to distinguish very typical items that were not generated from atypical items they observed, retrieving source features from the memory of the observed actions (visual, cognitive, affective, etc.) would allow them to report the observed action over the novel action when asked to do so by the task.

In conclusion, plagiarism errors are not simply the result of participants guessing and reporting typical exemplars at study and at test. While performance when asked to retrieve self-generated items may be explained by participants simply using overall memory strength to guide their responding, performance when asked to retrieve partner-generated items cannot. A source memory account that assumes that participants consider qualitative features of their memory alongside the overall memory strength would be more parsimonious in accounting for performance in both retrieval tasks.

## References

Brown, A. S., \& Murphy, D. R. (1989). Cryptomnesia: Delineating inadvertent plagiarism. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 432442. doi: 10.1037/0278-7393.15.3.432

Gingerich, A. C., \& Sullivan, M. C. (2013). Claiming hidden memories as one's own: A review of inadvertent plagiarism. Journal of Cognitive Psychology, 25, 903-916. (00003) doi: 10.1080/20445911.2013.841674

Hoffman, H. G. (1997). Role of memory strength in reality monitoring decisions: Evidence from source attribution biases. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 371-383. doi: 10.1037/02787393.23.2.371

Hollins, T. J., Lange, N., Berry, C. J., \& Dennis, I. (2016). Giving and stealing ideas in memory: Source errors in recall are influenced by both early-selection and latecorrection retrieval processes. Journal of Memory and Language, 88, 87-103. doi: 10.1016/j.jml.2016.01.004
Hollins, T. J., Lange, N., Dennis, I., \& Longmore, C. A. (2015). Social influences on unconscious plagiarism and anti-plagiarism. Memory, 1-19. doi: 10.1080/09658211.2015.1059857

Jacoby, L. L., Woloshyn, V., \& Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. Journal of experimental psychology: General, 118, 115-125. doi: 10.1037/0096-3445.118.2.115

Johnson, M. K., Hashtroudi, S., \& Lindsay, D. S. (1993). Source monitoring. Psychological Bulletin, 114, 3-28. doi: 10.1037/0033-2909.114.1.3

Marsh, R. L., \& Bower, G. H. (1993). Eliciting cryptomnesia: unconscious plagiarism in a puzzle task. Journal of Experimental Psychology. Learning, Memory, and Cognition, 19, 673-688. doi: 10.1037/0278-7393.19.3.673
Perfect, T. J., Field, I., \& Jones, R. (2009). Source credibility and idea improvement have independent effects on unconscious plagiarism errors in recall and generate-new tasks. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 267-274. doi: 10.1037/a0013936
Perfect, T. J., \& Stark, L. J. (2008). Tales from the Crypt... omnesia. In J. Dunlosky \& Bjork (Eds.), A handbook of metamemory and memory (pp. 285-314). New York: NY: LEA.
Tenpenny, P. L., Keriazakos, M. S., Lew, G. S., \& Phelan, T. P. (1998). In search of inadvertent plagiarism. The American Journal of Psychology, 111, 529-559. doi: 10.2307/1423550

# When do learned transformations influence similarity and categorization? 

Steven Langsford (steven.langsford@adelaide.edu.au)<br>School of Psychology, University of Adelaide<br>Andrew T. Hendrickson (d.hendrickson@tilburguniversity.edu)<br>Department of Communication and Information Systems, Tilburg University<br>Amy Perfors (amy.perfors@adelaide.edu.au)<br>School of Psychology, University of Adelaide<br>Daniel J. Navarro (d.navarro@unsw.edu.au)<br>School of Psychology, University of New South Wales


#### Abstract

The transformational theory of similarity suggests that when judging similarity, people are sensitive to the number of transformation operations needed to make two compared representations match. Although this theory has been influential, little is known about how transformations are learned and to what extent learned transformations affect similarity judgments. This paper presents two experiments addressing these questions, in which people learned categories defined by a transformation. In Experiment 1, when the transformations were directly visible, people had no trouble learning and applied their knowledge to similarity and categorization judgments involving previously unseen items. In Experiment 2, the task required transformations to be inferred rather than observed. People were still able to learn the categories, but in this more difficult case ratings were less strongly affected by training. Overall, this work suggests that newly learned transformations can impact similarity judgments but the salience of the transformation has a large impact on transfer.


Keywords: similarity; category learning; transformational similarity

## Introduction

Calculating similarities is a core process in cognition (Medin, Goldstone, \& Gentner, 1993) and plays a central role in categorization (Nosofsky, 1984). However, there is considerable debate about the fundamental building blocks for computing the similarity between objects that contain structured properties (Markman \& Gentner, 1993; Hahn, Chater, \& Richardson, 2003). One proposed basis for similarity is the transformational distance between items (Imai, 1977), which holds that the similarity between two objects is proportional to the number of steps required to transform one object into the other. Several papers outline the theoretical foundations of the approach (Chater \& Vitányi, 2003; Chater \& Hahn, 1997; Bennett, Gács, Li, Vitányi, \& Zurek, 1998), the empirical evidence for it (Hahn et al., 2003; Hodgetts, Hahn, \& Chater, 2009; Hahn, 2014), and the arguments against it (Larkey \& Markman, 2005; Müller, van Rooij, \& Wareham, 2009; Grimm, Rein, \& Markman, 2012).

Transformation distances are sensitive to the primitive transformations available, but it is unclear how people might determine the relevant set (Grimm et al., 2012). Some transformations may be innate, but Müller et al. (2009) argue that for computational tractability, transformations must be organized in relatively small domain-specific sets. This suggests
that where domain structure is learned, the relevant transformations for comparisons in that domain must also be learned.

We interpret the transformational approach as predicting a strong link between transformation learning and similarity judgments: learning a new transformation that directly connects two items should reduce the transformation distance between the items and thus increase the similarity between them. However, relatively little is known about how quickly transformations can be learned or how much new transformations impact similarity. The most relevant evidence comes from Hahn, Close, and Graf (2009), who found that people shown morphs from A to B rated similarity higher in the observed morph direction than the reverse direction. These results suggest that people are able to learn transformations over short timescales, and that there may be some impact on similarity. We extend this line of work using a transfer task, where test items are novel but instantiate the trained transformation. We manipulate whether transformations are directly observed or inferred, and separate measures of learning success from those of similarity judgment change.

## Experiment 1

Can people learn categories that are defined by a novel transformation, and do they apply this transformation to novel categorization and similarity judgments? Experiment 1 addresses these questions with a training task designed to maximize the salience of a transformation relationship linking objects that belong to the same category. This is accomplished by showing the transformation after each categorization judgment during training. After training, we compare category membership and similarity judgments for a common set of previously unseen test items, contrasting responses from participants who were trained on different transformations.

Our results suggest that people learned the transformations and that this learning influenced subsequent categorization and similarity judgments. Items related by the newly-learned transformation were rated as more similar and more likely to belong to the same category. Items related by a novel transformation sharing some higher-level properties with the trained one were also rated as more similar and more likely to belong to the same category, although to a lesser extent.


Figure 1: The two transformations used during the training phase of Experiment 1. In the movement training people learned a non-rigid clockwise rotation transformation (top row), whereas in the COLOR TRAINING condition they learned a color swapping rule (bottom row). For both, the image on the left shows how that transformation was defined, and the image on the right gives an example on a particular stimulus. In this figure we use textures to display the four possibilities for each cell. The actual stimuli were presented in color, with the four possible values being red, green, yellow and blue.

## Method

Participants Four hundred and forty-four participants were recruited via Amazon Mechanical Turk and paid US\$0.75. $62 \%$ were male, with ages ranging from 18 to 67 (mean: 33.3). Three hundred and eleven participants were from the USA, 120 were from India, and 13 were from other countries. Forty-seven were excluded from all analyses: 12 for self-reported color-blindness and 35 for failing to pass check questions during the test phase of the experiment.

The experiment used two different pre-defined exclusion criteria, one based on training phase responses and one based on test phase responses. For the training phase, if any participant took more than 40 trials to learn any category that participant's data would be excluded. No participants were excluded on this basis. For the test phase, we also excluded any participant who gave an average similarity/categorization rating of less than 6 (out of 7) to the test trials with identical stimuli: 35 people were removed on this basis. One hundred and eighty six people were assigned to an IDENTITY condition in which the transformation to be learned was the identity transformation (i.e., no change). These participants easily learned the categories but were at floor for all generalization questions. Their results are not analyzed further.

Procedure The experiment consisted of six training phases and a test phase. Within each training phase, participants were trained on a new category of objects until their accuracy reached criterion. In the test phase, participants were asked to make categorization or similarity judgments of novel stimuli. All stimuli in the experiment consisted of $3 \times 3$ grids of colored cells, where each cell was a single color: red, yellow, blue or green (see the right panel of Figure 1). The stimuli were approximately 200 pixels wide on each side.

In each training phase, participants were shown a 'base' stimulus and told that it belonged to a category (e.g., wugs).

Two items were displayed underneath with the question "Which of these is also a wug?" Participants were instructed to respond by clicking on the button located below their choice and were given feedback based on their choice. After an incorrect selection, the message "Sorry, try again" appeared and participants had to click the correct stimulus to proceed. After a correct selection, the message "Correct" appeared and an animation was presented morphing the base stimulus into the correct one. The next trial would then begin with the newly transformed item as the new target stimulus. For each category (e.g., wugs) this process continued until either the participant made four correct choices in a row or 40 trials had elapsed, at which point the experiment moved on to the next category (e.g., philbixes).

The set of stimuli in each category was determined by the base pattern and the transformation (shown in Figure 1). Each of the six training categories began with a unique 'base pattern' that was the same for all participants, and on each subsequent trial category members were generated by one application of the transformation. For participants in the COLOR TRAINING condition ( $\mathrm{n}=114$ ), the transformation from one item in the category to the next was a color-swapping rule in which cells that were colored red became green, green became red, blue became yellow, and yellow became blue. In the MOVEMENT TRAINING condition ( $\mathrm{n}=144$ ), the transformation that defined the set of items in the category was a non-rigid clockwise rotation of the cells in the grid. Applying this transformation caused the colors around the outside of the grid to shift one cell forward.

The test phase consisted of 20 test trials in which participants were asked to make judgments about pairs of novel stimuli that never appeared during training. The stimuli could be related to each other in one of six ways: identical $(\mathrm{n}=2)$, no simple relation ( $n=2$ ), related by the trained movement ( $n=4$ ) or trained color ( $n=4$ ) transformations, or related by the novel movement ( $n=4$ ) or novel color ( $n=4$ ) transformations. The


Figure 2: Effect of transformation training. The $y$-axis reflects the difference in responses given due to training condition, contrasting ratings given when test items do not match the training condition (NO MATCH) as compared to when they are related to the training, either as an exact MATCH or as a similar but novel NEAR MATCH. Thus, values above zero indicate effective training (in the case of MATCH) and generalization (in the case of NEAR MATCH). The left panel shows Experiment 1, which made the transformations explicit. In it, people learned and generalized the transformations for both categorization (light bars) and similarity (dark bars) questions, although the magnitude was smaller for similarity. The right panel shows Experiment 2, in which the transformations were less salient. In that case, learning and generalization were evident for categorization questions, but these were much larger than for similarity. Error bars express $95 \%$ credible intervals for a Bayesian t-test.
basis for these relations were not equally available to all participants: test items instantiating color transformations were unrelated for people given the movement training, and vice versa, manipulating the relation of the test items to the training while keeping the items themselves constant. The identical and no simple relation trials were of the same form as the test trials but only used for attention-check exclusions and not analyzed further. The novel movement transformation consisted of shifting all cells in the grid down by one row and moving the bottom row to the top. The novel color transformation swapped red with blue and green with yellow.

The order of test trials was randomized. Half the participants in each condition were asked to make CATEGORIZATION judgments by rating how likely it is that the two stimuli "have the same name" from "Not at all" to "Extremely" on a seven point scale. The other half were asked to rate the SIMILARITY of the two stimuli on a seven point scale.

In summary, there were two training conditions, each using a different transformation. There were four critical types of test item (excluding attention checks). The critical property of interest was the relationship between the test item and the training condition: did the test reflect the same or similar transformation as the training? The same items had different status for different participants depending on the training they saw: test items were considered to be MATCH trials when the two stimuli being compared were related by an application of the trained transformation, NEAR MATCH trials when
the test stimuli were related by a transformation similar but not identical to the trained one, and NO MATCH when the test stimuli were not related to the training. Thus, for a person who received COLOR TRAINING, a test item involving that same color transformation would be a MATCH, one involving the novel color transformation would be a NEAR MATCH, and the two movement-related test items would be NO MATCH. None of the test items were previously seen in training.

## Results

We first wish to establish whether the training phases were of comparable difficulty. We therefore looked at both how fast people reached the mastery criterion as well as the exclusion rates between conditions. People reached the criterion of four correct responses in a row in an average of 6.3 trials in the MOVEMENT TRAINING condition, and 5.8 in COLOR TRAINING, with $95 \%$ of all categories learned in eight trials or less. Participant inclusion rates were also comparable across conditions, at $86.0 \%$ and $87.5 \%$, as was average accuracy over all trials ( $85 \%$ and $88 \%$ in the MOVEMENT and COLOR training respectively). This suggests that people learned to effectively distinguish category members from foils and that both transformations were similarly difficult.

To test the impact of training on people's ratings, we examined the degree to which their responses were different on the MATCH and NEAR MATCH test items from the NO MATCH baseline. NO MATCH baseline ratings for test items related

Similarity Judgments

| Test item relationship | Test item average | NO MATCH item average | Difference | BF |
| :--- | ---: | ---: | ---: | ---: |
| MATCH items | $3.61(1.71)$ | $2.49(2.00)$ | 1.13 | $>1000$ |
| NEAR MATCH items | $3.14(1.71)$ | $2.58(1.91)$ | 0.569 | $>1000$ |

CATEGORIZATION Judgments

| Test item relationship | Test item average | NO MATCH item average | Difference | BF |
| :--- | ---: | ---: | ---: | ---: |
| MATCH items | $4.17(1.76)$ | $1.46(1.83)$ | 2.71 | $>1000$ |
| NEAR MATCH items | $3.15(2.11)$ | $1.7(1.922)$ | 1.45 | $>1000$ |

Table 1: Descriptive statistics and hypothesis tests for Experiment 1. For each of the MATCH and NEAR MATCH items (first column), we show the average responses for each (second column) compared to the NO MATCH baseline on the same items (third column). We performed a Bayesian $t$-test on the difference between these (fourth column) and found that in all cases there was a strong effect of training (fifth column).
by a COLOR transformation came from participants exposed to MOVEMENT TRAINING, baseline ratings for test items related by a MOVEMENT transformation came from participants exposed to COLOR TRAINING. In both cases the stimuli involved in the MATCH and contrasting NO MATCH groups were physically identical, likewise for NEAR MATCH items and their corresponding NO MATCH group. The left panel of Figure 2 illustrates these differences due to training experience. For instance, the MATCH bar reflects the difference between responses for the same item in the MATCH and NO MATCH conditions (thus, a value higher than zero indicates that the transformation training had an effect). Similarly, the NEAR MATCH bar reflects the difference between responses for the same item in the NEAR MATCH and NO MATCH conditions (thus, a value higher than zero indicates some generalization of training to a similar transformation).

Table 1 shows the absolute responses for the items of interest (i.e., the MATCH or NEAR MATCH items, in the second column) and the unrelated NO MATCH items in the third column. We used a Bayesian t-test (Morey, Rouder, \& Jamil, 2014; Rouder, Speckman, Sun, Morey, \& Iverson, 2009) to quantify the difference between them (fourth column), yielding a Bayes factor associated with the size of that different (fifth column). There was a strong ( $B F>1000: 1$ ) effect of training for both the categorization and similarity judgments. However, these two types of judgment were impacted to different extents. For instance, the overall difference in item ratings between training conditions was between 1.07 and 1.33 larger ( $95 \%$ credible interval) for categorization judgments than similarity judgments.

Similarly, both MATCH and NEAR MATCH test item ratings differed strongly due to training ( $B F>10^{3}: 1$ ), but to different extents. For instance, the difference due to training was between 0.67 and 0.93 rating points larger for MATCH as opposed to NEAR MATCH transformations. This suggests that people were less likely to generalize their responses as strongly to similar but not identical transformations.

## Conclusion

The results of Experiment 1 show that learning categories that are defined by a transformation can lead people to produce consistently different patterns of judgments for novel items.

Test items that were connected either by a learned transformation or a similar transformation were reliably rated higher. This increase in rating was found for similarity judgments as well as judgments about category membership.

This pattern of results is consistent with the predictions of the transformational account of perceptual similarity (Hahn et al., 2003). Furthermore, it suggests that by learning categories that are related by a transformation people can infer the transformation and apply it to novel items and categories.

That said, it is unclear to what extent the training in Experiment 1 is reflective of real-world transformation learning. In the experiment, objects were shown transforming into each other repeatedly; but in the real world, many transformations that define categories occur at a time scale that people cannot directly observe (e.g. seasons, aging, etc.). Experiment 2 aimed to test if the explicit presentation of the transformation was necessary to elicit quick learning and generalization of transformations.

## Experiment 2

Experiment 1 provides "in principle" evidence that people are capable of learning rich knowledge about classes of stimulus transformations and the categories to which they are applicable. However, the structure of our task made learning as easy as possible: during the training phase participants were explicitly shown the transformation at the end of every trial. When learning new categories in real life it is more typical for people to encounter a variety of exemplars. For example, when learning the transformations involved in the aging of human faces, people observe many faces at different ages, but do not directly observe the aging process. It is thus unclear how generalizable these results are.

This issue is particularly important for evaluating the transformational account of similarity. With a few notable exceptions, such as rotation, the majority of transformations plausibly involved in comparisons are unobservable. Experiment 2 addresses the question of how easily learnable transformations when they are more implicit. By increasing the difficulty of the task, this manipulation also allowed us to examine the extent to which variation in ease of transformation learning is reflected in similarity.

## Method

Participants Two hundred and fifty-two participants were recruited via Amazon Mechanical Turk and paid US\$1. $60 \%$ were male, with ages ranging from 19 to 67 (mean: 34.7). Two hundred and forty-seven participants were from the USA, with the remainder from India, South America, and the UK. Fifty-three were excluded from all analyses: 2 for self-reported color-blindness, 12 for not completing the experiment, and 39 for failing exclusion criteria.

The experiment used two different predefined exclusion criteria, one based on training phase responses and one based on test phase responses. For the training phase, if any participant took more than 30 trials to learn two of the last three categories that participant's data would be excluded (this number was arrived at based on pilot data). Twelve participants were excluded on this basis. For the test phase, any participant who gave an average similarity/categorization rating of less than 6 (out of 7) to the identical trials were excluded: Twenty-seven people were removed on this basis. Ninetyfive participants were in a COLOR TRAINING condition and 92 were in a movement training condition. Sixty-five participants were in an Identity condition and their results are not analyzed further.

## Procedure

As in Experiment 1, this experiment consisted of six training phases and a test phase. Within each training phase, a new category of objects was learned until a mastery criterion was reached. In the test phase, participants were asked to make categorization and similarity judgments of novel stimuli. However, a number of aspects of the experiment differed from Experiment 1.

Based on pilot testing of category learning, the stimuli were simplified by adding the constraint that each stimulus contained at least six cells that shared the same color. Furthermore, the COLOR TRAINING transformation was modified to increase the number of possible stimuli within the categories. Instead of changing all colors (red to green, green to red, yellow to blue, and blue to yellow) as a single transformation, this was broken into two transformations. A single transformation consisted of either swapping the colors of red and green, or swapping yellow and blue. This doubled the number of stimuli in each condition in the COLOR TRAINING condition to more closely match the number in the mOVEment training condition.

The structure of the training trials also differed from Experiment 1 . On each trial participants were shown two stimuli and asked if both items belonged in the category. There was always at least one category member displayed. In half the trials, the other stimulus was also in the category and related by one application of the transformation being trained. In the other half, the other stimulus was not in the category. After participants responded yes or no they were given feedback indicating if they were correct, but unlike in Experiment 1 they did not observe the actual transformation.

Participants proceeded to the next category when they were correct on 8 of 10 trials. Consecutive sets of six trials were constrained to contain three 'yes' and three 'no' trials (in shuffled order), meaning participants reaching criterion answered both 'yes' and 'no' correctly. The test phase was largely similar to Experiment 1 except that the UNRELATED trials were removed and trials were grouped into four blocks, with order of presentation randomized within each block to avoid runs of similar test items.

## Results

The results indicate that difficulty was higher than Experiment 1 , but comparable across conditions. People reached the accuracy criterion in an average of 14.12 trials in COLOR training and 13.8 trials in movement training. Inclusion rates were comparable between conditions at $74 \%$ and $80 \%$ respectively.

As in Experiment 1, we were interested in whether responses to test items were different based on whether the transformation involved was a MATCH, NEAR MATCH, or NO match to the trained transformation. The right panel of Figure 2 shows the differences in responses, analogous to the same analysis in Experiment 1, with the associated Bayesian t-test results shown in Table 2. In all cases we find strong evidence that participants' ratings for the same items were higher when they had a MATCH or NEAR MATCH relationship to training as opposed to NO match status ( $B F>49: 1$ at minimum). This suggests that training was effective and people were capable of learning the transformations even if they were not explicitly shown.

That said, the size of the difference depended on question type. Category learning showed a much larger effect: the difference was between 1.09 and 1.31 points larger ( $95 \%$ credible interval) for categorization questions than similarity ones. Unlike in Experiment 1, match and near match status were not strongly differentiated: the difference involving MATCH status items as opposed to NEAR MATCH status items plausibly included zero (with a $95 \%$ CI between -0.01 and $0.21)$.

## General Discussion

The results from Experiment 1 showed that people are capable of learning a novel transformation, recognizing that this transformation is relevant to determining category membership, and applying the learned transformation when assessing similarity between items belonging to novel categories. This finding is consistent with the learning effect seen in Hahn et al. (2009), but extends previous results in showing systematic generalization across related transformations.

Experiment 2 echoes these results and further finds that the effect is not limited to training in which people see objects transforming; seeing labeled category members can induce a change in judgments. However, there are two notable differences from Experiment 1. First, as the transformations become less prominent during training, they seem to have less impact on subsequent judgments, particularly for similarity.

|  | SimILARITY Judgments <br> Test item relationship |  |  |  | Test item average |
| :--- | ---: | ---: | ---: | ---: | ---: |
| NO MATCH item average | Difference | BF |  |  |  |
| MATCH items | $3.56(1.6)$ | $2.98(1.75)$ | 0.58 | $>1000$ |  |
| NEAR MATCH items | $2.66(1.53)$ | $2.31(1.62)$ | 0.35 | 49 |  |
|  |  |  |  |  |  |
|  |  | CATEGORIZATION Judgments |  |  |  |
| Test item relationship | Test item average | NO MATCH item average | Difference | BF |  |
| MATCH items | $3.72(1.85)$ | $2.14(2.15)$ | 1.57 | $>1000$ |  |
| NEAR MATCH items | $2.93(1.98)$ | $1.35(1.82)$ | 1.58 | $>1000$ |  |

Table 2: Descriptive statistics and hypothesis tests for Experiment 2. For each of the match and NEAR MATCH items (first column), we show the average responses for each (second column) compared to the NO мATCH baseline on the same items (third column). We performed a Bayesian $t$-test on the difference between these (fourth column) and found that in all cases there was a strong effect of training (fifth column).

Second, the novel and trained transformations were less well differentiated.

The attenuation of the training effects seems likely to be a result of task difficulty, with people less inclined to shift their judgments based on training that was less clear. However the lack of differentiation between trained and novel but similar transformations is harder to interpret. Possibly participants formed an incomplete representation of the transformation which was applicable to both near and exact matches to training, but the form of this representation is unclear. People's success in distinguishing targets from foils at training suggests they did not simply track which features remain invariant (e.g., noting in the MOVEMENT TRAINING condition that colors are preserved and in the COLOR TRAINING condition that configurations are preserved).

In terms of the predictions of transformational similarity, our results are somewhat mixed. It is clear that people learn transformations relevant to a new domain quickly, and that such transformations can be applied to categorization and similarity judgment. However, the pattern of generalization between exact matches and near-matches would seem to require some kind of graded availability of transformations based on family resemblances between them, complicating the computation of transformation distances.

Transformations as features are common in natural categories, for example growth and aging or characteristic movement. Despite this, their role in similarity judgments over structured representations remains unclear. Taking as a starting point predictions implied by tractability constraints on the transformational account of similarity, the two studies presented here examine the conditions under which transformation learning might influence similarity and categorization. Our results show that people can learn transformations quickly and use them in subsequent similarity and categorization judgments. However, productive use of the transformations depends to some extent on the ease with which the transformation was learned, and in both easy and difficult learning conditions involves generalization across related transformations.

## Acknowledgments

SL was supported by an Australian Government Research Training Program Scholarship. DN received salary support from ARC grant FT110100431 and AP from ARC grants DP110104949 and DP150103280.

## References

Bennett, C. H., Gács, P., Li, M., Vitányi, P. M., \& Zurek, W. H. (1998). Information distance. Information Theory, IEEE Transactions on, 44(4), 1407-1423.
Chater, N., \& Hahn, U. (1997). Representational distortion, similarity and the universal law of generalization. In Simcat97: Proceedings of the interdisciplinary workshop on similarity and categorization.
Chater, N., \& Vitányi, P. (2003). The generalized universal law of generalization. Journal of Mathematical Psychology, 47(3), 346369.

Grimm, L. R., Rein, J. R., \& Markman, A. B. (2012). Determining transformation distance in similarity: Considerations for assessing representational changes a priori. Thinking \& Reasoning, 18(1), 59-80.
Hahn, U. (2014). Similarity. Wiley Interdisciplinary Reviews: Cognitive Science, 5(3), 271-280.
Hahn, U., Chater, N., \& Richardson, L. B. (2003). Similarity as transformation. Cognition, 87(1), 1-32.
Hahn, U., Close, J., \& Graf, M. (2009). Transformation direction influences shape-similarity judgments. Psychological Science, 20(4), 447-454.
Hodgetts, C. J., Hahn, U., \& Chater, N. (2009). Transformation and alignment in similarity. Cognition, 113(1), 62-79.
Imai, S. (1977). Pattern similarity and cognitive transformations. Acta Psychologica, 41(6), 433-447.
Larkey, L. B., \& Markman, A. B. (2005). Processes of similarity judgment. Cognitive Science, 29(6), 1061-1076.
Markman, A. B., \& Gentner, D. (1993). Structural alignment during similarity comparisons. Cognitive psychology, 25(4), 431-467.
Medin, D. L., Goldstone, R. L., \& Gentner, D. (1993). Respects for similarity. Psychological review, $100(2), 254$.
Morey, R. D., Rouder, J. N., \& Jamil, T. (2014). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=BayesFactor (R package version 0.9.8)
Müller, M., van Rooij, I., \& Wareham, T. (2009). Similarity as tractable transformation. In Proceedings of the 31st annual conference of the cognitive science society (pp. 50-55).
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, memory, and cognition, 10(1), 104.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., \& Iverson, G. (2009). Bayesian $t$ tests for accepting and rejecting the null hypothesis. Psychonomic bulletin \& review, 16(2), 225-237.

# Preschoolers and Infants Calibrate Persistence from Adult Models 

Julia Leonard<br>Massachusetts Institute of Technology, Cambridge, MA, USA<br>Max Kleiman-Weiner<br>Massachusetts Institute of Technology, Cambridge, Massachusetts, USA<br>Yuna Lee<br>Massachusetts Institute of Technology, Cambridge, Massachusetts, USA<br>Josh Tenenbaum<br>Massachusetts Institute of Technology<br>Laura Schulz<br>Massachusetts Institute of Technology


#### Abstract

Perseverance, above and beyond IQ, predicts academic outcomes in school age children, however, little is known about what factors affect persistence in early childhood. Here, we propose a formal Bayesian model of how children might learn how to calibrate effort from observing adult models and then explore this idea behaviorally across two experiments in children and infants. Results from Experiment 1 show that preschoolers persist more after watching an adult persist, but only if the adult is successful at reaching their goal. Experiment 2 and a pre-registered replication extend these findings, showing that even infants use adult models to modulate their persistence, and can generalize this inference to novel situations. These results suggest that both preschoolers and infants are sensitive to adult persistence and use it to calibrate their own effort in far-reaching ways.


# You can take a noun out of syntax...: Syntactic similarity effects in lexical priming 

Nicholas A. Lester (nlester@umail.ucsb.edu)<br>Department of Linguistics, University of California, Santa Barbara, USA

Laurie B. Feldman (Ifeldman@albany.edu)<br>Haskins Laboratories, New Haven, CT, \& Department of Psychology, State University of New York, Albany, NY, USA

Fermín Moscoso del Prado Martín (fmoscoso@linguistics.ucsb.edu)<br>Department of Linguistics, University of California, Santa Barbara, CA, USA


#### Abstract

Usage-based theories of syntax predict that words and syntactic constructions are probabilistically interconnected. If this is true, then words that occur in similar distributions of syntactic constructions should prime each other. These effects should be fine-grained; even small differences between the syntactic distributions of pairs of words of the same grammatical category should cause variation in priming. Prior research from production suggests that this prediction should hold even in tasks without any syntactic requirement. In this study, we introduce a measure of the similarity between the syntactic contexts in which two nouns occur. We show that this similarity measure significantly predicts visual lexical decision priming magnitudes between pairs of nouns. This finding is consistent with the predictions of usage-based theories where fine-grained similarity of syntactic usages between prime-target pairs affects decision latencies, over and above any effects attributable to semantic similarity.


Keywords: syntax; priming; usage-based linguistics; visual lexical decision; information theory

## Background

Lexical priming experiments have a long history in psycholinguistic research. Though the bulk of this research has focused on semantic and orthographic effects, some studies have considered the role of syntax (henceforth grammatical priming). Early work looked at the effects of inflectional congruity across word classes. For example, in Serbian, inflected nouns are recognized faster when primed by case-appropriate adjectives (e.g., Gurjanov, Lukatela, Moskovljević, Savić, \& Turvey, 1985). More recent work has looked at contextualized reading effects. Nouns and verbs that are biased to occur in congruent syntactic constructions (e.g., direct-object vs. subordinate clause continuations; I need some coffeelto go to the market) facilitate processing of later content (Novick, Kim, \& Trueswell, 2003). Thus, accessing a noun primes expectations about its syntactic context. Congruity effects have been interpreted as evidence for robust, probabilistic syntactic specifications for lexical items.

The empirical evidence outlined so far is complemented by work in theoretical linguistics. Usage-based linguistic theories argue that all facets of grammar, including words
and syntactic structures, are potentially interconnected on the basis of one's experience with language (e.g., Diessel, 2015). Let us refer to this position as the probabilistic network hypothesis. Results such as those reported by Novick et al. (2003) are easily accounted for under this framework. To use the connectionist metaphor, connections between lexical and syntactic nodes are tuned as a function of their frequency of distinctive co-activation (e.g., Gries \& Stefanowitsch, 2004). Stronger connections are processed more efficiently. Further support for this hypothesis comes from work on word production: the probability distributions of words in particular syntactic structures influence picture naming latencies (Lester \& Moscoso del Prado Martín, 2016).

Direct, probabilistic relationships between words and syntactic structures are not universally accepted across linguistic models. Many models argue that syntax only enters the lexicon through general categorical specifications (i.e., most generative approaches to syntax). Accordingly, words may have a feature indicating the part-of-speech category to which they belong (noun, verb, adjective, and so on). More recent work in this vein has expanded the syntactic content of the lexicon to include more fine-grained syntactic categories. For example, in current mainstream generativist syntax (the Minimalist Program; Chomksy, 1995), words contain information about the syntactic frames with which they can combine as functional head (sometimes called subcategorization or c-selection; for a similar approach, see Bresnan, 2001). Crucially, these syntactic specifications represent categorical constraints on the possible distributions of words. We will call this the categorical constraint hypothesis. Under this account, probabilistic relationships are simply not available to the grammar. Any effects of probability are designated "extra-grammatical" (Stabler, 2013) and are instead usually attributed to relationships in other mental systems, such as the Conceptual-Intentional system.

This theoretical distinction leads to different predictions about the nature of grammatical priming. The probabilistic network hypothesis predicts that probabilistic information about the semantic and syntactic similarity of words should produce independent priming effects. The categorical constraint hypothesis predicts that probabilistic effects
should only arise for semantic similarity (as the syntactic system does not encode such relationships). We test this contrast using a simple lexical priming paradigm.

Research on grammatical priming has largely relied on syntactic or pseudo-syntactic contexts (e.g., using an adjective as a prime for a noun). However, the predictions of usage-based theory, along with recent evidence from production (e.g., Lester \& Moscoso del Prado Martín, 2016), suggest that syntactic information -all of it- should be automatically activated every time a word is accessed. This should be true even when the word is presented in isolation for purposes of the task, as in visual lexical decision (see also Durán and Pillon, 2011). We therefore use a simple overt lexical priming paradigm with visual lexical decision. We restrict our analysis to nouns to guard against intercategorical effects. We predict RTs based on the similarity of semantic and syntactic distributions across a range of words words. The probabilistic network hypothesis would be supported by evidence of priming for similar syntax and semantics, independently. The categorical constraint hypothesis would be supported by priming only in the domain of semantics.

## Methods

## Data

We used the response latencies contained in the Semantic Priming Project (SPP; Hutchison, et al., 2013). The SPP contains response times and accuracies, along with a host of norming data, that were collected using a visual lexical decision task with overt orthographic priming. On each trial, participants were shown a centered fixation cross for 500 ms , followed by a prime word (all caps) for 150 ms . The prime was followed by a blank screen lasting either 50 or 1050 ms . The target word was displayed (all lowercase) until a either decision was made or $3,000 \mathrm{~ms}$ elapsed, at which point the experiment would advance to the next trial.

We used only those trials containing primes and targets that also appear both in the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, \& Brysbaert, 2012) and the age of acquisition norming database of Kuperman, Stadthagen-Gonzalez, \& Brysbaert (2012). We limit the data in this way to take advantage of the additional lexical controls afforded by these databases. To ensure that all stimuli were understood primarily as nouns, we further limited the trials to include only those in which both prime and target received majority noun tags in the British National Corpus (BNC). In this way, we obtained a dataset consisting 1,305 unique primes and 821 unique targets (a total of 1,670 unique nouns).

## Syntactic space

To measure the relationship between the noun-pairs in the syntactic system, we first need to operationalize the syntactic system itself. Decades of research have failed to produce an exhaustive list of the syntactic constructions of English (let alone any other language), and we do not presume to offer such a list here. Instead, we rely on the set of low-level
relations as defined within Dependency Grammar formalisms (e.g., Mel'čuk, 1988; Nivre, 2005). Dependency Grammars model only relations (dependencies) between pairs of words. These relations are asymmetric: each extends from a head (the syntactic and conceptual core word) to a modifier (whose syntactic role is contingent on the head). Each dependency is labeled to reflect its syntactic function. For example, the and waffle in the noun phrase the waffle would be bound by the det relation, which attaches a determiner (the, the modifier) to a noun (waffle, the head). Other examples include the nsubj relation, which binds a noun (modifier) to a verb (head) as its subject, and the pobj relation, which binds a noun (modifier) to a preposition (head) as its object. A further detailed description and discussion of Dependency Grammar formalism is beyond the scope of this study. We adopt the dependency formalism implemented in the spaCy parser (Honnibal \& Johnson, 2015), one of the fastest and most accurate dependency parsers available.

We define the syntactic space for nouns as the set of dependencies for which at least one noun from our sample of SPP primes and targets has been attested either as head or as modifier. For each noun in our dataset, we extracted all sentences containing that noun from the BNC. We conditioned the search to include only sentences in which the word form was indeed tagged as a noun. Those sentences were parsed using spaCy. We then compute the frequency distribution of each noun across the dependencies for which it serves as head or modifier. To increase the reliability of our frequency estimates, we discard vectors for all nouns that occurred in fewer than 100 sentences in the BNC ( $\sim 1$ per million words). The total syntactic space is defined as a vector in which each column reflects one among the set of unique dependencies occurring across all nouns. Finally, we merge the individual frequency distribution of each noun into the total syntactic space, creating a matrix of $n$ rows by $m$ columns, where $n=$ the number of total unique dependency types (46) and $m=$ the number of unique SPP/BLP nouns $(1,241)$. The result is therefore a uniform syntactic space for all nouns, where individual nouns may or may not be attested in each possible dependency. In theoretical terms, we treat these vectors as reflecting the statistical connectivity between each noun and the syntactic structures in which it takes part, as is proposed in the usage-based literature. Psycholinguistic support for this treatment comes from an earlier study showing that these and similar dependency vectors affect processing latencies in noun production over and above the effects of other known factors (Lester \& Moscoso del Prado Martín, 2016).

## Measuring syntactic similarity

We are interested in the possibility that pre-activation of shared syntactic representations will affect the speed of word recognition. Therefore, we need some measure of the similarity between the syntactic distributions of primes and targets in our behavioral data. Note that similarity in syntactic space outlined above does not reduce solely to shared types of dependencies. For example, consider two
words, $w 1$ and $w 2$, that occupy the same set of 20 dependency types. Suppose that $w 1$ and $w 2$ have roughly equivalent overall frequencies and that those frequencies are distributed equally across the dependency types for both words. In this case, we would call them syntactically similar, and consider the number of overlapping types as an appropriate measure of the strength of their similarity. Now suppose that the two words have similar overall frequencies, but that these frequencies are distributed over complementary sets of the dependencies that they share, such that $w 1$ has a frequency of 1 wherever $w 2$ has a frequency >100 and vice versa. In this case, we would call them dissimilar. For this, we need to simultaneously account for shared types, as well their probability distributions. One measure well suited to this task is the Jensen-Shannon Divergence (JSD; Lin, 1991). JSD is a symmetric variant of the Kullback-Leibler Divergence (KLD). The KLD between two probability distributions $P$ and $Q$ is defined in Eq. 1.

$$
\begin{equation*}
K L D(P \| Q)=\sum_{\mathrm{i}} P(\mathrm{i}) \log \frac{P(\mathrm{i})}{Q(\mathrm{i})} \tag{1}
\end{equation*}
$$

This measure captures the average amount of additional information that one would need in order to recode an event from distribution $P$ as if it belonged to distribution $Q$. Importantly, $\operatorname{KLD}(P \| Q) \neq \operatorname{KLD}(Q \| P)$, meaning that one must decide a priori in which direction to take the distance. JSD provides a solution to the asymmetry problem by taking the midpoint between the two distributions, then taking the mean distance of the distributions to the midpoint. Formally, JSD is expressed as follows (Eqs. 2 and 3).

$$
\begin{equation*}
J S D(P \| Q)=\frac{1}{2} K L D(P \| M)+\frac{1}{2} K L D(Q \| M) \tag{2}
\end{equation*}
$$

where

$$
\begin{equation*}
M=\frac{1}{2}(P+Q) \tag{3}
\end{equation*}
$$

This measure has the advantage of being both symmetrical $[\mathrm{JSD}(P \| Q)=\mathrm{JSD}(Q \| P)]$ and bounded $(0 \leq \mathrm{JSD} \leq 1)$.

JSD measurements depend on estimates of the probability distributions of events within a distribution, rather than on their actual probability distributions. Maximum-likelihood estimates of information-theoretical measures are known to be biased. To guard against this bias we apply a bias-reducing frequency correction to our syntactic vectors, using the plug-in James-Stein shrinkage estimator (Hausser \& Strimmer, 2009).

The methods above provide an operationalization of syntactic similarity between primes and targets. For each prime-target pair in the sample, we compute the JSD between their syntactic vectors. A value of 0 indicates identity; a value of 1 indicates complete independence. According to usage-based theories, (at least the bulk of) syntactic structure is meaningful- that is, directly linked to semantic representations in the same way as words (e.g., Diessel, 2015). This means that any effect we uncover for our
measure may actually reflect semantic similarity, which is well known to affect priming magnitudes (e.g., Neely, 1991). Moreover, the contrast between the probabilistic network and categorical constraint hypotheses depends on a direct comparison of syntactic and semantic sources of similarity. Fortunately, the SPP contains annotation of the degree of semantic similarity between prime and target, indicated by cosine similarities in Latent Semantic Analysis space (LSA). LSA measures the extent to which words occur in similar texts, with higher cosine values indicating greater similarity (Landauer \& Dumais, 1997). We transformed the cosine similarities into distances (i.e., 1-cos).

Figure 1 plots the relationship between the syntactic distances (JSD) between pairs of words as a function of their semantic distances (LSA) values. As one would expect, there is a significant positive (linear) ${ }^{1}$ correlation between both measures, meaning that words that are similar in meaning tend to occur in similar syntactic contexts. However, an important feature of Figure 1 is the triangular shape of the variance: words that are very close in meaning vary only slightly in syntactic similarity, while words that are distant in meaning vary more widely. This relationship supports the account of Jackendoff (2013), who argues for the existence of syntactic generalizations (i.e., constructions) that allow structural inheritance among sets of semantically heterogeneous sub-constructions. In other words, nouns that are extremely similar in meaning (e.g., synonyms) will always appear in extremely similar syntactic contexts. However, there is large variability in the syntactic similarities of words with different meaning (or there is large variability in the semantic similarity between pairs of words that appear in very different syntactic contexts). This suggests that syntax and semantics are not as tightly coupled as some would argue (e.g., Goldberg, 1995), and their contributions can indeed be considered separately.


Figure 1: Relationship between syntactic and semantic distance measures

[^384]To disentangle the purely syntactic aspects of lexical similarity from what can be attributed to similarity in meaning, we residualized the semantic measure out of the syntactic measure. This was achieved by fitting a linear regression predicting the JSDs as a function of the LSA distances, and using the residuals of this regression as our measure of syntactic difference. This measure captures the information in JSD that is not attributable to semantics (cf., Hendrix, Bolger, \& Baayen, 2017; responding to the concerns expressed by Wurm \& Fisicaro, 2014).

## Further controls

A number of other factors are known to impact recognition latencies in the primed lexical decision paradigm. These fall into three categories: effects related to recognizing individual words, (other) effects based on the relationship between prime and target, and effects related to the nature of the task itself. From the first set, the most important predictor is the surface frequency of the target: i.e., more frequent words are recognized faster. We use the SUBTLEX-UK frequencies, which are based on movie subtitles and known to outperform estimates drawn from other corpora, including the BNC (van Heuven, Mandera, Keuleers, \& Brysbaert, 2014). We also include a measure of the density of the orthographic neighborhood of the target known as OLD20 (Yarkoni, Balota, \& Yap, 2008). The more similar the spelling of the word to its closest neighbors, the faster it is recognized. Another predictor that has been proposed is age of acquisition: the earlier a word is acquired in the lifespan, the faster it is recognized (e.g., Kuperman et al., 2012). Less important, but nevertheless known to exert an effect, is the orthographic length of the word: longer words take longer to recognize (New, Ferrand, Pallier, \& Brysbaert, 2006).

Besides our residualized syntactic measure, we included two additional predictors relating the prime and target: We included semantic distance (i.e., the LSA distances), as semantic similarity is known to facilitate access to targets (i.e., semantic priming). In addition, we considered the Levenshtein distance (LD; Levenshtein, 1966; van der Loo, 2014) between prime and target to account for possible effects of orthographic relatedness. We expect orthographically similar prime-target pairs to result in slower recognition latencies (cf., Adelman, et al., 2014). In addition to these main effects, we tested two-way interactions between the inter-stimulus interval (ISI) on the one hand, and LSA distance, LD, and residualized JSD on the other. This was done to account for the possibility that priming effects might change with the different offsets between prime and target.

Finally, we included the (log) sequential position of each trial in the overall experimental order of presentation. As participants move through the trials, we expected some degree of fatigue to set in (each participant performed over 800 trials).

## Results

We fitted a linear mixed-effect regression model predicting response latencies from the SPP primed lexical decision database as a function of the variables outlined above. In addition to the fixed effects, we included random effects for participants and prime-target pairs (i.e., random slopes). We discarded $6.7 \%$ of all trials as outliers (all latencies falling below 400 ms or 2 standard deviations above the mean). In addition, we corrected for a strong positive skew in the response times by taking the logarithm of RTS (as suggested by a Box-Cox power analysis; Box \& Cox, 1964). Visual inspection of the model residuals with and without the corrections confirmed the adequacy of these steps.

All main effects for the control predictors besides OLD20 surfaced as significant at the $\alpha=.05$ level, and in the expected direction. The model also revealed a significant ( $p<.001$ ) effect of the two-way interaction between LD and ISI: at 50 ms ISI, LD had a negative impact on response times $(-2.5 \mathrm{~ms}$ per unit increase in LD), with no effect at 1050 ms . Importantly, the model revealed a significant interaction ( $p<.01$ ) between ISI and LSA distance, consistent with what one would expect. Response latencies increased by about 5 ms per .1 increase in cosine distance at a short ISI. At a long ISI, this effect was reduced to $\sim 3 \mathrm{~ms}$ per .1 increase. As semantic distance between prime and target increased, so did target recognition latencies, with stronger effects at the shorter ISI.

Over and above the effects of the controls, and crucially over that of semantic similarity, the model revealed a statistically independent significant main effect ( $p<.001$ ) of the residualized syntactic distance. For every . 1 increase in residualized syntactic distance, response latencies were increased by $\sim 4 \pm \sim 3 \mathrm{~ms}$. As predicted by the probabilistic network hypothesis, the less related the prime and target in syntactic space, the longer it takes to recognize the target. There was also a marginal interaction of JSD with ISI ( $p=.07$ ). The trend resembled that observed for LSA: longer ISIs lead to an attenuated contribution of syntactic similarity. However, given the marginal status of the effect, we do not interpret it further.

## Discussion

The present study finds a relatively strong effect of syntactic similarity on lexical priming magnitudes. In fact, the effect was similar in strength -if anything stronger- to that of semantic similarity. To our knowledge, this study is the first to demonstrate that pre-activating a word's syntactic space affects access to that word in a prima facie non-syntactic task. This effect provides support for the probabilistic network hypothesis, which predicts that words and syntactic structures are interdependent, and that these connections are forged and tuned by experience. Crucially, these probabilistic relationships are at the core of the grammatical apparatus they are not simply attributable to the extra-grammatical conceptual system. If that were the case, we should have found no effect of syntactic similarity once semantics was accounted for.

The data we rely on here do not provide us with a non-primed baseline, meaning that we cannot distinguish a facilitation effect of syntactic similarity from an inhibitory effect of syntactic dissimilarity. We therefore leave this question for further research. However, the similarity in shape between the syntactic and semantic effects suggest that syntax -as is argued for semantics (e.g., Lam et al, 2015)constrains the set of lexical candidates prior to the lexicality judgment. Furthermore, it suggests that syntax, like semantics, is obligatorily accessed as soon as lexical forms become active. Crucially, the relationships between words and syntax become active even when (overt) syntactic structure is not built into the stimuli and not really necessary for performing the task. Recent psycholinguistic work on single-word production has echoed this point. For example, Lester and Moscoso del Prado Martín (2016) report chronometric findings suggestive of large-scale interactivation between syntax to lexicon in a bare-noun picture-naming task. Other studies have found that syntactic category information is likewise obligatorily activated in non-syntactic production tasks (e.g., Durán and Pillon, 2011). The present study extends these findings from production to comprehension, from spoken language to written language, and from a simple to a primed paradigm. Hence, the converging evidence suggests that obligatory syntactic access, along with bi-directional activation between syntax and lexicon, is a general, modality-independent property of language processing.

These data also speak to linguistic representation. Branigan and Pickering (in press) argue that, in order for priming to take place, some common connection must exist between the prime and target on the one hand, and the representations underlying the measurement of their similarity. This notion is applied to the relationship between words and conceptual content in the semantic priming literature (e.g., Lam, Dijkstra, \& Rueschemeyer, 2015). Likewise, our results can be interpreted as reflecting that each noun's representation is explicitly connected to the set of syntactic structures in which it participates and that these representations are shared across words. Moreover, the probabilistic nature of our measure suggests that connection weights -not just the set of shared syntactic types- are represented in the lexico-syntactic network, exactly as predicted by usage-based models of linguistic representation (Diessel, 2015) and as evidenced in sentence-reading paradigms (Novick et al., 2003). Importantly, these findings are not consistent with modular-syntactic models (e.g., Chomsky, 1995), which posit only categorical relationships between words and syntax. Adapting the old adage, "you can take the noun out of syntax, but you can't take the syntax out of the noun."

A possible limitation is that we used Latent Semantic Analysis as a proxy for semantic related when 'cleaning' our syntactic measure of its semantic component. It remains possible -albeit, in our opinion, unlikely- that part, or even all, of the effects of syntactic similarity could be accounted
for by a more fine-grained measure of semantic relatedness or similarity than that provided by LSA.

Another possible limitation concerns the morphological structure of the words in our study. While we only included monosyllabic and disyllabic nouns, some of the tokens contained derivational morphology (e.g., actor). Morphology is known to interact with priming from other domains (e.g., semantics; Feldman et al., 2015). Therefore, it remains unclear to what extent morphology was contributing to both the shapes of the distributions we computed from the corpus and/or aspects of the priming relationship. In future research, it will be necessary to account for possible derivational relationships between target and prime, and to explore how morphological structure impacts syntactic diversity.
The interaction between our measure and the temporal offset between the prime and the target was only marginally significant. The SPP contains only two such offsets: extremely fast and extremely slow. We suspect that a more robust interaction might arise if one considers offsets intermediate between these extremes. Furthermore, by incrementally increasing the offset between 50 and 1050 ms , we would allow considering the ISI as the numerical magnitude it is (cf., Feldman et al., 2015), rather than as a bi-valued factor.

In sum, our results suggest that, in line with the predictions of usage-based theories of grammar, the representation of words is inextricably tied to the grammatical contexts in which these words are encountered. The results indicate that even the extremely fine-grained differences in syntactic use that can be found between words of a single class (nouns) have detectable effects on their processing and representation. This is true even in tasks -such as visual lexical decision- that do not to involve any explicit involvement of the syntactic system. In other words, in comprehension, the activation of the syntactic properties of a word is automatic. The word comes with its whole syntactic baggage. Furthermore, this syntactic baggage goes well beyond mere grammatical category information, and includes a rich, fine-grained account of the syntactic contexts in which each particular noun is used.

## References

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., Perry, J. R., Lupker, S. J., Forster, K. I., Cortese, M. J., Scaltritti, M., Aschenbrenner, A., J., Coane, J. H., White, L., Yap, M. J., Davis, C., Kim, J., \& Davis, C. J. (2014). A behavioral database for masked form priming. Behavioral Research Methods, 46, 1052-1067.
Box, G. E. P., \& Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistics Society, Series B (Methodological), 26, 211-252.
Branigan, H. \& Pickering, M. (in press). An experimental approach to linguistic representation. Behavioral and Brain Sciences.
Bresnan, J. (2001). Lexical Functional Syntax. Oxford: Blackwell Publishers.

Chomsky, N. (1995). The minimalist program. Cambridge: MIT Press.
Diessel, H. (2015). Usage-based construction grammar. In E. Dabrowska and D. Divjak (Eds.), Handbook of Cognitive Linguistics (pp. 295-321). Boston: De Gruyter.
Duràn, C. P. \& Pillon, A. (2011). The role of grammatical category information in spoken word retrieval. Frontiers in Psychology, 2, 1-20.
Feldman, L. B., Milin, P., Cho, K. W., Moscoso del Prado Martín, F., \& O'Connor, P. (2015). Must analysis of meaning follow analysis of form? A time course analysis. Frontiers in Human Neuroscience, 11, 1-19.
Goldberg, A. E. (1995). Constructions: A Construction Grammar approach to argument structure. Oxford: Oxford University Press.
Gries, S. Th. \& Stefanowitsch, A. (2004). Extending Collostructional Analysis: A corpus-based examination of 'alternations.' International Journal of Corpus Linguistics, 9, 97-129.
GurJanov, M., Lukatela, G., Moskoljević, J., Savić, M. \& Turvey, M. T.. (1985). Grammatical priming of inflected nouns by inflected adjectives. Cognition, 19, 55-71.
Hausser, J. \& Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research, 10, 1469-1484.
Hendrix, P., Bolger, P. and Baayen, R. H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43, 128-149.
Honnibal, M. \& Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1373-1378). Lisbon, Association for Computational Linguistics.
Hutchison, K.A., Balota, D.A., Neely, J.H., Cortese, M.J., Cohen-Shikora, E. R., Tse, Chi-Shing, Yap, M. J., Bengson, J. J., Niemeyer, D., \& Buchanan, E. (2013). The Semantic Priming Project. Behavior Research Methods, 45, 1099-1114.
Jackendoff, R. (2013). Constructions in the parallel architecture. In T. Hoffmann \& G. Trousdale (Eds.), The Oxford Handbook of Construction Grammar (pp. 70-92), Oxford: Oxford University Press.
Keuleers, E., Lacey, P., Rastle, K., \& Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. Behavior Research Methods, 44, 287-304.
Kuperman, V., Stadthagen-Gonzalez, H., \& Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. Behavior Research Methods, 44, 978-990.
Lam, K. J. Y., Dijkstra, T., \& Rueschemeyer, S-A. (2015). Feature activation during word recognition: action, visual, and associative-semantic priming effects. Frontiers in Psychology, 6, 1-8.

Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
Lester, N. A. \& Moscoso del Prado Martín, F. (2016). Syntactic flexibility in the noun: Evidence from picture naming. In A. Papafragou, D. Grodner, D. Mirman, \& J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 2585-2590). Austin, TX: Cognitive Science Society.
Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Doklady Akademii Nauk SSSR, 163, 845-848.
Lin, J. (1991). Divergence measures based on the Shannon Entropy. IEEE Transactions on Information Theory, 37, 145-151.
Mel'čuk, I. (1988). Dependency syntax: Theory and practice. Albany: The SUNY Press.
Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theory. In D. Besner \& G. W. Humphreys (Eds.), Basic processes in reading: Visual word recognition (pp. 264-336). Hillsadale, NJ: Erlbaum.
New, B., Ferrand, L., Pallier, C., \& Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. Psychonomic Bulletin and Review, 13, 45-52.
Nivre, J. 2005. Dependency grammar and dependency parsing. Technical Report MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.
Novick, J. M., Kim, A., Trueswell, J. C. (2003).Studying the grammatical aspects of word recognition: Lexical priming, parsing, and syntactic-ambiguity resolution. Journal of Psycholinguistic Research, 32, 57-75.
Stabler, E. P. (2013). Two models of minimalist, incremental syntactic analysis. Topics in Cognitive Science, 5, 611-633.
Stefanowitsch, A. \& Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. International Journal of Corpus Linguistics, 8, 209-243.
van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. The R Journal, 6, 111-122.
Van Heuven, W.J.B., Mandera, P., Keuleers, E., \& Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. Quarterly Journal of Experimental Psychology, 67, 1176-1190.
Wurm, L. H., \& Fisicaro, S. A. (2014). What residualizing predictors in regression models does (and what it does not do). Journal of Memory and Language, 72, 37-48.
Yarkoni,T., Balota, D., \& Yap, M. (2008). Moving beyond Coltheart's $N$ : A new measure of orthographic similarity. Psychonomic Bulletin and Review, 15, 971-979.

# How the Mind Exploits Risk-Reward Structures in Decisions under Risk 

Christina Leuker (leuker@mpib-berlin-mpg.de)<br>Timothy J. Pleskac (pleskac@mpib-berlin.mpg.de)<br>Thorsten Pachur (pachur@mpib-berlin.mpg.de)<br>Ralph Hertwig (hertwig@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development<br>Lentzeallee 94, 14195, Berlin, Germany


#### Abstract

In many natural domains, risks and rewards are inversely related (Pleskac \& Hertwig, 2014). We sought to understand how people might use this relationship in choosing among risky gambles. To do so we, manipulated risk-reward structures of monetary gambles to be either negatively or positively correlated, or uncorrelated. After substantial exposure to these environments, participants completed a speeded choice task among non-dominated gambles. Eye-tracking data from this task suggests that participants often shifted their attention to mainly one attribute in the correlated conditions, in which the risk-reward relationship was present. This was an adaptive strategy that resulted in a similar proportion of expected-value maximizing choices, compared to a more compensatory processing strategy.


Keywords: risk-reward relationship; decisions under risk; attention; noncompensatory processing; adaptive cognition

## Introduction

How likely is it to win the jackpot in the state lottery? Although many people play this game for a small pay-to-play fee, most of them also know that they are unlikely to win it. In fact, the larger rewards that we desire are usually unlikely to occur. While such a negative relationship between risks and rewards or probabilities and payoffs exists across gambles in many monetary and nonmonetary domains in the environment, this relationship is hardly every present in empirical studies of risky choice (Pleskac \& Hertwig, 2014). In this study, we investigated how people's experience with different risk-reward relationships impact how they process explicitly stated payoffs and probabilities in decisions under risk. In particular, we studied how an environment in which risks and rewards are correlated would be conducive for the use of noncompensatory processing strategies, that ignore part of the attributes, in a situation where time was limited.

## Adaptive Decision Making

According to an adaptive view of cognition, people exploit statistical regularities in the environment (Simon, 1956). As Payne, Bettman and Johnson (1993) found, the extent to which people exploit structures in the environment can largely depend on "the structure of the available alternatives, and [...] the presence of time pressure" (p. 534). For instance, people can decide to rely on a subset of cues in the environment because cues are often interrelated (Brunswik, 1952). Despite using a reduced amount of information, this can lead to good choices (Gigerenzer, Todd, \& the ABC Research Group, 1999). Here, we propose that the risk-reward relationship is a key structure that people capitalize on to make fast, adaptive (or value-maximizing) decisions.

## Choice in Risk-Reward Environments

When should and do people rely on risk-reward structures to inform their decisions? One case is when information is missing, such as in decisions under uncertainty, where the probabilities of obtaining a reward are unknown. In this case, Pleskac and Hertwig (2014) showed that people use a riskreward heuristic, inferring the probability of a payoff from the magnitude of the payoff itself. In a new set of studies, we have also found that in using the risk-reward heuristic people appear to adapt to different risk-reward structures (Leuker, Pleskac, Pachur, \& Hertwig, in prep.). In particular, we exposed participants to different risk-reward environments by asking them to price gambles from different risk-reward environments. Then we asked participants to choose between an uncertain prospect (where the probabilities were unstated) and a certain payoff. Participants' preferences were again consistent with them using a risk-reward heuristic, inferring probabilities from payoff magnitudes. Moreover, their preferences depended on the environment they had been exposed to previously. For example, participants in the negative condition chose the lower payoff, uncertain options more often compared to the positive condition. Based on these results, we sought to examine if and how people adapt their decisionmaking processes to risk-reward structures in decisions under risk, when payoffs and probabilities of the option are known.

## The Current Study

Processing strategies. One way to distinguish between processing strategies is to consider the amount of attributes they rely on. Compensatory strategies process and trade off of all available and relevant information. Noncompensatory strategies "typically reduce processing demands by ignoring potentially relevant information" (Payne, Bettman, \& Johnson, 1988). Thus, one important reason to consider noncompensatory processing strategies (despite information being, in principle, available, as in risky choice) is when time or cognitive resources are limited.

Strategy-environment dependence. Early research on these two classes of strategies demonstrated that their success largely depends on the environment in which they are recruited. Specifically, in environments with nondominated options (e.g., gamble $A$ offers a higher payoff $x$, but gamble $B$ offers a higher probability $p: x_{A}>x_{B}$ and $p_{A}<p_{B}$ ), people should rely on compensatory strategies (see Table 2, Payne et al., 1988). A decision maker who processes the dimensions


Figure 1: Choice stimuli based on their relationship between probabilities and payoffs. Each point depicts one gamble from the choice phase. Across conditions, probabilities and payoffs were (A) negatively correlated, (B) positively correlated or (C) uncorrelated. Black circles are environment gambles (60 pairs). Triangles are common gambles interspersed in all three conditions (15 pairs). Dominated options not depicted (5 pairs).
in a noncompensatory fashion in these environments-for instance by relying on a simplifying heuristic that attends to outcomes only-will suffer "a substantial loss in accuracy" (Payne et al., 1993, p. 539). In contrast, such noncompensatory processing strategies have been shown to perform well when dominance is possible (that is, one gamble is better on all dimensions: if $x_{A}>x_{B}$ and $p_{A}>p_{B}$ ).

Local vs. global environment. By definition, nondominated options create an inverse risk-reward relationship in a given set of alternatives, because the gamble offering a higher payoff will always be associated with a lower probability relative to the other gamble $\left(x_{A}>x_{B}\right.$ and $\left.p_{A}<p_{B}\right)$. However, this "local" risk-reward relationship (within a pair of gambles) can differ from a "global" risk-reward structure (across a larger reference class of gambles). That is, nondominated alternatives can be drawn from globally structured or unstructured environments. We propose that both the use and performance of either type of strategy is also highly dependent on these global risk-reward structures. Global correlations between risks and rewards make one of the cues redundant (payoffs predict probabilities and vice versa). Therefore, we hypothesized that, when options are drawn from correlated riskreward environments, noncompensatory strategies can lead to accurate, expected-value maximizing choices even if neither option is dominated. For choices between nondominated options from globally uncorrelated environments, results may resemble those of Payne, Bettman and Johnson (1988).
To test these ideas, we employed a between-subjects design manipulating the global risk-reward relationship between the possible options participants experienced (Figure 1). In a first pricing phase, we showed participants individual gambles and asked them to state their willingness to sell each gamble. We used this phase to expose people to different risk-reward environments. Detailed data from this phase will be reported else-
where. Our focus in this paper is the second phase, where participants chose between pairs of risky options under moderate time pressure (Figure 2). The gambles in the choice phase were drawn from the same, condition-dependent risk-reward environments, and paired such that neither option was dominated. We tracked participants' eye movements to dissociate between processing strategies across the different risk-reward environments, as choice patterns alone may not be sufficient to do so. As an independent test of whether participants had picked up the different risk-reward relationships, we asked them to estimate probabilities from payoffs at the end of the experiment.

## Method

## Participants

Ninety-three ( 55 female) participants (mean age $=25.6 \mathrm{yrs}$, $S D=3.7 ; N=31$ per condition) from the participant pool at the Max Planck Institute for Human Development, Berlin, completed the experiment (duration $\sim 75 \mathrm{~min}$ ). All participants were paid a fixed rate of $€ 12$ plus a bonus based on their performance in a random subset of trials from the pricing phase and choice task (€3.53-11.67).

## Design

The experiment consisted of three phases. In the pricing phase, participants were presented with single gambles and asked to indicate their willingness to sell for each of them. Between subjects, we manipulated the types of gambles people were presented with such that payoffs and probabilities were positively or negatively correlated, or uncorrelated. In the subsequent choice task, these different risk-reward structures were maintained. People were asked to choose between gamble pairs within 3s. All gambles were in the gain domain (" $p_{1}$ chance of winning $x_{1}$, otherwise nothing"). We used an experimental currency, the $\mathrm{E} \$$ (conversion rate $2500 \mathrm{E} \$=€ 1$,
disclosed in the instructions). We collected eye-tracking data during the exposure phase and the choice task. As people are merely exposed to different risk-reward structures, participants picking up risk-reward structures despite not being told about the presence of any relationship in the data would constitute a form of unsupervised learning. Finally, in the third phase we asked participants to estimate the probabilities they thought were associated with various payoff levels. We did this to test whether participants had picked up the different risk-reward structures from the gambles they were exposed to throughout the study. Participants were not informed about the estimation task beforehand.

Gamble environments. The gambles from the pricing and choice phases were constructed such that across gambles, there was a negative, a positive, or no relationship between risks and rewards. For the negative condition, we drew random payoffs from a uniform distribution (range 1.01 $2500 E \$$ ). The probabilities for each payoff were inversely related to the payoff $x$ such that, $p=1-\frac{x}{2500}$. We added normally distributed noise to logit-transformed payoffs and probabilities. For the positive condition, we reversed the order of probabilities such that the highest probabilities were now associated with the highest rewards (and vice versa). For the uncorrelated condition, we re-linked payoffs and probabilities randomly.

Pricing task. The pricing task served to expose participants to different risk-reward environments. Briefly, participants were shown each of the 90 gambles from one of the environments and asked to state a price they would be willing to sell the gamble for. In addition to 90 condition-dependent gambles based on the aforementioned construction rule, participants were also asked to price 30 gambles that were common to each of the three conditions (triangles in Figure 1), yielding 120 gamble stimuli per condition. To motivate participants to report their true valuations of the gambles, we implemented a Becker-DeGroot-Marschak auction (Becker, Degroot, \& Marschak, 1964). In particular, ten gambles were selected at the end of the experiment and participants either played out the gamble or received their stated selling price.

Choice task. Gambles were created using the same construction rule as above. An initial set of 100 gambles yields 4950 possible gamble pairs. We randomly drew 60 nondominated gamble pairs per condition (see circles in Figure 1). By design of the study, expected value differences were largest in the uncorrelated and smallest in the positive condition (uncorrelated: $M d=173 E \$$, $.53-1374 \mathrm{E} \$$; negative condition: $M d=134 E \$$, $49-511 \mathrm{E} \$$; positive condition: $M d=23 E \$$, $43-146 \mathrm{E} \$$ ). In addition, we interspersed 15 gamble pairs that were common to each of the conditions in the second half of the choice task (triangles in Figure 1), and 5 choices with dominated options as catch trials, yielding 80 choices in total. Common gambles allowed us to examine condition-dependent processing differences on precisely the same stimuli. Participants were instructed to choose their


Figure 2: Typical choice trial. For trials that exceeded the time limit, we presented an additional screen informing participants that they had lost payoffs in that particular trial (not shown). Eye-tracking data was recorded throughout, analyses are based on the second screen.
preferred gamble within 3s (see Figure 2). Crucially, participants were informed in this task that the gambles were drawn from the same population of gambles they had experienced in the previous pricing task. Five randomly selected choices were played out at the end of the experiment.
Estimation task. We drew 20 payoffs (range 1.01 $2500 E \$$ ) and asked participants to estimate the probabilities that had been associated with these payoffs in the main experiment.

## Eye-tracking

During the pricing and choice tasks, we collected binocular eye position data with an EyeTribe tracker, sampled at 60 Hz . The experiment was implemented in PsychoPy 1.83.01 and the eye-tracking interface PyTribe (Dalmaijer, Mathôt, \& Van der Stigchel, 2013). Each participant's eye movements were calibrated using the Eyetribe UI with a 9-point grid before each task ( $<0.7$ ). Participants were seated approximately 60 cm from the screen using a chinrest affixated to the table, in a room with negligible ambient light. We preprocessed raw samples by parsing eye-tracking data into fixations and saccades using the saccades package in R (Saccades Version 0.1-1, 2015), based on a velocity-based algorithm (Engbert \& Kliegl, 2003). Eye-tracking analyses in this paper are based on fixation data.

## Analysis

The data were analyzed using Bayesian General Linear Models using Stan in R for regression analyses (RStanArm Version 2.9.0-4, 2016). We ran 3 chains ( 2500 samples each, burn-in of 500), and investigated (convergence of) our posteriors visually and with the Gelman-Rubin statistic (Gelman \& Rubin, 1992). We report the mean of the posterior distribution of the parameter of interest and two-sided $95 \%$ equal tail credible intervals (CI) around each value.

## Results

## Behavioral

We excluded one participant in the negative condition who chose the dominated option in 4 out of 5 catch trials.
Pricing task. For all participants, prices were strongly related to the expected values of the gambles (credible payoff $\times$ probability interaction, $b=.70, \mathrm{CI}=[.66, .74])$.

Choice task. Participants across all three conditions chose the expected-value-maximizing options above chance level ( $M=.71$, $\mathrm{CI}=[.56, .85]$ ). As expected, in the positive condition (in which the EV differences between the options were rather small) participants made fewer EV-maximizing choices $(M=.59, \mathrm{CI}=[.51, .67])$ than in the uncorrelated condition (in which the EV differences were larger; $M=$ $.70, b=.11, \mathrm{CI}=[.00, .22]$ ), and the negative condition ( $M=.74, b=.25, \mathrm{CI}=[.14, .37]$ ). Controlling for EV differences and individual variation, participants in the negative condition achieved a higher proportion of expectedvalue maximizing choices $(M=.70, \mathrm{CI}=[.52, .88])$ compared to the uncorrelated condition $(M=.31, b=-.39, \mathrm{CI}=$ $[-.63,-.16])$, and the positive condition ( $M=.39, b=-.32$, $\mathrm{CI}=[-.54,-.09])$. In both models, the highest accuracy was achieved in the negative condition. In the subset of gambles that were common across all conditions, there were no differences in accuracy between the conditions $(M=.53, \mathrm{CI}=$ [.35, .72]).

Response times were comparable across all conditions and gamble types. In addition, small proportions of timed-out trials (negative: .006, positive: .016, uncorrelated: .013) indicate that participants were well-adjusted to the speed instruction of $3 \mathrm{~s}(M d=1.63 \mathrm{~s}$ even suggest that people could have taken more time on many trials).

Estimation task. Participants' probability estimates reflected the risk-reward structure they had been exposed to previously. That is, participants in the negative condition provided lower probability estimates for gambles with higher payoffs ( $b=-.64, \mathrm{CI}=[-.68,-.60]$, \% per $100 \mathrm{E} \$$ ), and in the positive condition participants provided higher probability estimates for gambles with higher payoffs ( $b=.16$, $\mathrm{CI}=[.10, .14])$. In the uncorrelated condition, participants provided lower probability estimates for gambles with higher payoffs (weaker slope compared to the negative condition: $b=-.32, \mathrm{CI}=[-.37,-.26])$.

## Eye-tracking

We defined four areas-of-interest (AOIs), one for each payoff and probability. We visually inspected the quality of every participant's eye-tracking data by plotting their fixations over time. Seven participants whose fixations did not map onto the screen correctly were excluded, a possible result of the eyetracker being moved during the experiment. We excluded one further participant who was blind in one eye, leaving $N=84$ for the eye-tracking analyses ( 27,29 , and 28 in the negative,


Figure 3: Number of AOIs inspected per condition. Each dot represents one participant's average number of AOIs visited per trial. Dashed line $=$ mean number of AOIs visited across participants. Differences between conditions are driven by the composition of compensatory/noncompensatory strategies (percent compensatory in negative: $44 \%$; positive: $34 \%$, uncorrelated: 61\%).
positive, and uncorrelated conditions, respectively).
Number of AOIs viewed. To test whether the presence of a risk-reward relationship led to more noncompensatory processing, we averaged the number of AOIs each participant viewed (max. 4). Participants in the uncorrelated condition inspected the largest number of AOIs ( $M=3.46, \mathrm{CI}=$ [3.45, 3.48]). Participants in the positive condition inspected a credibly lower number of AOIs ( $M=3.20, b=-.20$, CI $=[-.23,-.18])$. The negative condition also inspected credibly fewer AOIs, but the difference was smaller ( $M=3.40$, $b=-.06, \mathrm{CI}=[-.08,-.03]$, model on the trial level). Using the average number of fixations as an alternative indicator resulted in the same pattern of results, and only a marginally higher count (uncorrelated condition $M=3.71$, negative condition $M=3.51$, positive condition $M=3.39$ ), likely because the time limit imposed in the experiment did not allow for many re-acquisitions (i.e., fixations back to a previously acquired AOI). Note that the mean number of fixations is rather low (i.e. <4). We ran the same AOI model using only common gamble data (see triangles in Figure 1). Again, the uncorrelated condition inspected most AOIs ( $M=3.21$, $\mathrm{CI}=[3.09,3.34]) . \quad$ This number was lower in the positive condition $(M=2.83, b=-.38, \mathrm{CI}=[-.57,-.20]$, difference credible), and in the negative condition ( $M=3.07$,
$b=-.14, \mathrm{CI}=[-.34, .04]$, difference however not credible). Because these gambles were identical across conditions, this suggests condition-dependent processing strategies are not merely a by-product of specific risk-reward environments that vary on crucial dimensions such as EV differences between gambles. Figure 3 suggests substantial individual differences among participants in the conditions (indeed, differences in numbers of AOIs inspected can be accounted for by including participant as a grouping factor). More importantly, however, visual inspection of the data suggests two subgroups that can roughly be split by the mean number of AOIs inspected across participants ( $M=3.11$, dashed line in Figure 3): Participants who tend to inspect all four AOIs ("compensatory") and participants who ignore some of the AOIs ("non-compensatory"). Thus, differences between conditions may be driven by the composition of compensatory/noncompensatory strategies (proportion compensatory in uncorrelated: .61, positive: .34, negative: .44). That is, participants in the uncorrelated condition were 2.85 times more likely to rely on a compensatory strategy than participants in the positive condition ( $b=1.05, \mathrm{CI}=[.07,2.04]$ ). The difference between the negative and positive risk-reward environments was not credible ( $b=.68, \mathrm{CI}=[-.29,1.73]$, OR $=1.97)$. A majority of participants in the correlated environments thus seemed to rely on a noncompensatory strategy (note that here such a strategy could also mean attending to three out of four AOIs per trial, see Figure 3).

Attention to attributes. Which attributes did participants attend to, especially when choosing to ignore some of the information? All participants fixated most on payoff information (. 57 of fixations, $\mathrm{CI}=[.53, .61]$ ). This proportion decreased for participants who inspected more AOIs ( $b=-.98$, $\mathrm{CI}=[-1.56,-.38]$; no credible effect of condition). At the extreme end, participants who, on average, inspected roughly two AOIs fixated on the payoff $80 \%$ of the time. An alternative viable noncompensatory strategy would have been to focus more on the information presented at the top or the bottom of the screen. We counterbalanced the location of attributes (between-participants). However, top/bottom fixation proportions were unrelated to the number of AOIs inspected ( $b=.03, \mathrm{CI}=[-.60, .66]$ ), suggesting that participants considered payoff information as more relevant when using noncompensatory strategies.

EV choices by strategy. Do compensatory or noncompensatory strategies differ in performance within the three environments? Figure 4 shows that users of a noncompensatory strategy (triangles) achieved similar levels of EV-maximizing choices compared to users of a compensatory strategy (circles), overall ( $M=-.10, \mathrm{CI}=[-.39, .18]$ ). Unexpectedly, this held irrespective of condition (no credible strategy $\times$ condition interaction). This result also held when controlling for differences between the condition in EV difference between the options. We expected that in the uncorrelated condition, decision performance would be compromised for users of a


Figure 4: Proportion of higher EV choices by a participant's average number of AOIs inspected per trial, condition, and decision strategy. Circles = noncompensatory strategy users, triangles $=$ compensatory strategy users. The gain in EVmaximizing choices associated with inspecting more AOIs is more pronounced for participants in the uncorrelated condition.
noncompensatory strategy-who should lack critical information to determine the EV maximizing option.

At the same time, across conditions, the proportion of EV-maximizing choices was higher for participants who inspected more AOIs (main effect of AOIs inspected on EV choice irrespective of processing strategy, $b=.044, \mathrm{CI}=$ [.001, .089]). Within each subgroup, the increase in EV choices with increasing numbers of AOIs inspected is more pronounced for participants in the uncorrelated condition (see black regression line, Figure 4). Yet, this interaction effect is not credible, potentially due to the small number of participants in each subgroup (AOI $\times$ condition interaction with the positive condition as a reference; compensatory: $b=.02$, CI $=[-.04, .09])$, noncompensatory: $b=.05, \mathrm{CI}=[-.01, .11])$. In general, one would indeed expect that the increase in the proportion of EV choices with higher number of AOIs inspected is more pronounced for the uncorrelated condition because this condition allows for least simplification.

## Discussion

Risk-reward relationships allow people to make fast, valuemaximizing decisions. A majority of people exposed to correlated risk-reward structures used noncompensatory processing strategies, likely as a result of time pressure. With fewer AOIs inspected, participants focused more on payoff
information. In turn, most people who experienced an uncorrelated risk-reward environment attempted to take into account all attributes, speaking in favor of a more compensatory processing strategy. This strategy use is adaptive given the affordances of the different environments. While correlated risk-reward environments made one of the attributes redundant, such a relationship did not exist for gamble problems in uncorrelated risk-reward environment. Condition-dependent processing differences (i.e., numbers of AOIs visited) persisted when restricting the analysis to a common set of gambles interspersed in each of the conditions. Surprisingly, these differences only had a minor impact on EV choice.
Earlier research suggested that noncompensatory strategies fare well when dominance is possible, but not when neither option is clearly dominated (Payne et al., 1988). We identify one qualification of this prediction, showing that noncompensatory processing strategies can also perform well for nondominated option pairs; namely when a risk-reward relationship is present in the global set that gamble pairs are drawn from. Researchers have studied the influence such contextual factors before. Birnbaum (1992) found that participants' certainty equivalents for gambles were larger when a set of certainty equivalents to choose from was positively skewed (vs. negatively skewed). In addition, the marginal distributions of payoffs, probabilities and delays can account for psycho-economic functions that are often described in the literature (Stewart, Chater, \& Brown, 2006). Here, we extend such considerations by manipulating the joint distribution of payoffs and probabilities.

Several limitations of the current study should be mentioned. First, it is currently unclear what underlies the strong individual differences in noncompensatory/compensatory strategy use in each condition. Potentially, some participants did not perceive a time limit of 3 s as pressing enough to opt for noncompensatory strategies, or turned to different simplification strategies. Overall, the dichotomous distinction between compensatory and noncompensatory processors may be too simplistic: For instance, some individuals attended to three attributes on average (i.e., more than one class of attributes such as two payoffs, one probability). Another possibility is that users of a noncompensatory strategy fixated on some but glanced the other attributes (covert attention), or changed strategies across trials. Lastly, more research is needed to study the process by which people learn about different risk-reward structures (Klayman, 1988).

## Conclusion

People's choices and processing strategies are impacted by the risk-reward structure in a given environment. Specifically, correlated risk-reward environments allow decision makers to use noncompensatory strategies when they need to reduce processing demands. This strategy use is adaptive, given that it does not need to compromise accuracy if it matches the environment. Many natural environments exhibit an inverse relationship between payoffs and probabilities that can thus be
exploited in a similar way, when time or cognitive resources are limited. These findings challenge theories of decision making under risk, that often treat payoffs and probabilities as independent attributes determining the value of an option. In comparison, an adaptive decision maker may often have good ecological reasons to process payoffs and probabilities dependently.

## Acknowledgments

This research was supported by the MaxNetAgeing Research School (CL). We would like to thank Chantal Wysocki and Lisa-Marie Jirak for assistance with data collection.

## References

Becker, G. M., Degroot, M. H., \& Marschak, J. (1964). Measuring utility by a single response sequential method. Behavioral Science, 9(3).
Birnbaum, M. H. (1992). Violations of Monotonicity and Contextual Effects in Choice-Based Certainty Equivalents. Psychological Science, 3(5), 310-314.
Brunswik, E. (1952). The conceptual framework of psychology. Journal of Consulting Psychology, 16(6), 475-475.
Dalmaijer, E. S., Mathôt, S., \& Van der Stigchel, S. (2013). PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. Behavior Research Methods, 1-16.
Engbert, R., \& Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. Vision Research(9), 10351045.

Gelman, A., \& Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, 7(4), 457-511.
Gigerenzer, G., Todd, P., \& the ABC Research Group. (1999). Simple heuristics that make us smart.
Klayman, J. (1988). On the how and why (not) of learning from outcomes. Advances in Psychology, 54, 115-162.
Leuker, C., Pleskac, T., Pachur, T., \& Hertwig, R. (in prep.). Risk and reward: Exploiting the environments risk-reward structures in decisions under uncertainty.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1988). Adaptive strategy selection in decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3), 534-552.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1993). The adaptive decision maker. Cambridge University Press.
Pleskac, T. J., \& Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. Journal of Experimental Psychology: General, 143(5), 2000-2019.
RStanArm version 2.9.0-4.
(2016).
cran.r-project.org/web/packages/rstanarm/.
Saccades version 0.1-1.
(2015).
cran.r-project.org/web/packages/saccades / .
Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological Review, 63(2), 129-138.
Stewart, N., Chater, N., \& Brown, G. D. A. (2006). Decision by sampling. Cognitive Psychology, 53(1), 1-26.

# Semantic vector evaluation and human performance on a new vocabulary MCQ test 

Joseph P. Levy (j.levy@roehampton.ac.uk)<br>Department of Psychology, University of Roehampton<br>London, UK

John A. Bullinaria (j.a.bullinaria@cs.bham.ac.uk)<br>School of Computer Science, University of Birmingham, Birmingham, UK

Samantha McCormick (Samantha.McCormick@roehampton.ac.uk)<br>Department of Psychology, University of Roehampton<br>London, UK


#### Abstract

Vectors derived from patterns of co-occurrence of words in large bodies of text have often been used as representations of some aspects of the meanings of different words. Generally, the distance between such vectors is used as a measure of the semantic similarity between the word meanings they represent. One important way of evaluating the performance of these vectors has been to use them to answer vocabulary multiple choice questions (MCQs) where the participant is asked to judge which of several choice words is closest in meaning to a stem word. The existing vocabulary MCQ tests used in this way have been very useful but there are some practical problems in their use as general evaluation measures. Here, we discuss why such tests remain useful evaluation measures, introduce a new vocabulary test, evaluate several current sets of semantic vectors using the new test and compare their performance to human data.


Keywords: Distributional semantics; vocabulary MCQ.

## Introduction

There are many potential applications for a method that can reliably form the basis for measuring the semantic distance between words or concepts. Many methods achieve this by placing each word/concept in a multidimensional space where the dimensions are defined by the way in which words co-occur in corpora of real language use (Schütze, 1993; Bullinaria \& Levy, 2007; Turney \& Pantel, 2010). The simplest such methods place a target word in a space defined by the count of how many times this word co-occurs with other words in the corpus with each of these context words defining a dimension. The resulting semantic vectors may also be smoothed by some kind of dimensionality reduction technique. Most current techniques retain only a small proportion of the number of initial dimensions (often 300) and refer to these dense vector sets as "word embeddings" (e.g., Mikolov et al., 2013). Two words are then judged to be semantically similar if the vector distance between them is small.

Various validation approaches have been used, but a particularly convenient way of evaluating such techniques is
to measure how well they perform in a vocabulary multiple choice question (MCQ) test where a participant is asked to choose which of a number of choice words is closest in meaning to a stem word (not to be confused with a morphological stem). Often this amounts to choosing a ("key") synonym, or the word that is closest to being a synonym, from the other choice words that act as distractors. For human participants, these tests are used to measure vocabulary knowledge. Such tests are ideal methods to use to evaluate co-occurrence techniques which construct semantic vectors for each word such that their distances are related to how substitutable the words are for each other. A linguistic intuition would be that if two words are substitutable for each other in everyday language then they are synonyms or at least very closely semantically related.

Landauer and Dumais (1997) used the performance of their LSA (Latent Semantic Analysis) method on the retired items of a once commercially available test of English vocabulary the TOEFL (Teaching of English as a Foreign Language). They reported a performance of $64.4 \%$, well above chance and equivalent to acceptable performance for entry to a US University. This MCQ test has been the most widely used vocabulary test to date for evaluating distributional semantic vectors ${ }^{1}$. Turney (2001) also describes a commercially available test, the ESL, and this has been used to evaluate such methods. Another candidate MCQ test would be the one used by Adelman et al. (2014) from Shipley (1940). This is a 40 item vocabulary MCQ using some now somewhat archaic US English usages. It has the advantage of being freely available in the appendix of a historic journal article. Such tests are valuable evaluation measures for semantic representations in that they are independently designed to measure the performance of human participants. However, they have a number of disadvantages if they are used as the only evaluation:

[^385]- They are at least relatively commercially sensitive as making a real-word test freely available would render it useless as a fair measure of human vocabulary knowledge. We have always found that researchers who report the use of such tests have been helpful and generous in making the test items available to other researchers but, inevitably, the commercial/practical sensitivity of the items is an obstacle for widespread open publication of them. This can prevent the clear reporting of slight changes made to the items due to low frequency word forms (or USA/UK spelling variants) not appearing, or not appearing frequently enough, in the corpora used, which can lead to papers reporting results of slightly different tests.
- They are relatively small as they are designed to be completed in a reasonably short period of time by the human testees.
- The individual questions may not be uniform in terms of difficulty or kind of semantic relationship being tested. A question may test knowledge of near-synonymy or one of a degree of some other kind of semantic similarity such as category membership.

In this paper, we describe a new 200 item MCQ vocabulary which we will make freely available. The test has been constructed using the lexicographical judgements implicit in the construction of the noun entries in WordNet (Fellbaum, 1998). Half of the stem words in the test are relatively high frequency (in the psycholinguistically relevant sense) and half are low frequency. Word frequency is a dominant lexical variable for human language processing and especially so in instruments, such as this one, that are designed to measure vocabulary knowledge. The 200 MCQ set is large enough to be performed comfortably by human participants and to be split into subsets for training and testing when using machine learning techniques.

As noted above, vocabulary MCQ tests have frequently been used as evaluation measures for distributional semantic vectors. However, some of the most recent methods for generating such semantic vectors (e.g., Mikolov et al., 2013; Pennington et al., 2014) have emphasised evaluation using sets of analogy problems. We would argue that both vocabulary and analogy tests are important in evaluating the semantic competence of distributional semantic vectors, as well as being useful in models of human semantic performance. Here, we therefore evaluate three promising recent semantic vector techniques using our new vocabulary test.

In constructing our new test, we use WordNet (Fellbaum, 1998), a freely available lexicographic database organised around lists of synonyms (synsets) for the different senses of each word in the database. This allows us to use the independent linguistic judgements of the WordNet team as a standard for competence in tests of synonymy judgements on the vocabulary MCQ items that we construct.

We consider that MCQ vocabulary tests are interesting psychological tasks in their own right. It is likely that word frequency measures will dominate any quantitative model of
relative question difficulty and that word familiarity (and proxies for this such as level of education or experience with English in the case of the data described here) is likely to dominate models of individual differences in performance on these tasks. If a participant has never or very rarely come across a stem or synonym then, apart from the possibility of sensitivity to form-meaning symbolism (e.g., Levy \& Thompson, 2008; Monaghan et al., 2014), they are unlikely to perform well on test items containing these words. However, there remains the strong possibility that semantic distances between stem and synonym, stem and distractors and synonym and distractors will affect question difficulty and the choice of distractor when an MCQ item is answered incorrectly. Semantic vector techniques are a good resource for calculating these distances. Thus, vocabulary MCQs are useful measures for evaluating the competence of semantic vector techniques, and semantic vectors are likely to be components of any complete model of human vocabulary MCQ performance.

In the rest of this paper, we outline how we constructed a new 200 item vocabulary MCQ test, show how well three recent methods for generating distributional semantics vectors perform on the test, and compare the performance of the semantic vectors with human performance on the same test items. We intend to make the MCQ items and human data freely available as a research resource.

## Construction of New Vocabulary Test

We constructed a set of 200 vocabulary MCQ items. This is larger than most of the evaluation benchmarks that have been suggested (allowing the set to be potentially split into independent training and testing components for reliable model selection purposes) but still a manageable number of items for individual human participants.

The words in the MCQs were chosen by using the entries for nouns in WordNet. All words considered for selection appeared in both the SUBTLEX-UK (Van Heuven et al., 2014) and WordNet databases, and were dominantly tagged in their noun form in both databases. SUBTLEX-UK is a database constructed from BBC TV subtitling records and so this ensured that the words chosen were in common usage in the UK.

We chose to generate stem-synonym pairs by using the synsets in WordNet because this gives us an independent benchmark for lexical semantic relations. By dividing the MCQ items into two subsets where one has relatively high stem frequencies and the other has relatively low ones, the vocabulary test controls one of the most important influences on human linguistic performance.

The potential candidate list of stem words was divided into lower-frequency (LF) and higher-frequency (HF) subsets using the "Zipf" scale (van Heuven et al., 2014), which is based on the $\log _{10}$ transform of word frequency. Those authors argue that this scale is a better way to control frequency in a psycholinguistically relevant way than frequency per million word (fpmw) counts. For example, using these counts to select stimuli results in an
underrepresentation of relatively low frequency words that are familiar to human participants.

These candidate lists were randomly sorted. Stems and synonyms (taken from the synsets associated with each stem noun) were selected from this list such that the final HF and LF subsets consisted of pairs that were matched for stem word length, synonym frequency and synonym length. Hyphenated stem words or synonyms were excluded from selection. Three distractor words were selected at random from the remaining nouns in WordNet and SUBTLEX-UK with a Zipf frequency greater than two. The mean distractor length and frequency (over the three distractors) was pairwise matched to the synonym. Mean stem, synonym and distractor characteristics are shown in Table 1.

Table 1: Mean Zipf Frequency (F) and word length (L) for Stem, Synonym (Syn) and Distractor (Distr) words

| MCQs | Stem | Stem | Syn | Syn | Distr | Distr |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | F | L | F | L | F | L |
| LF | 3.0 | 6.3 | 3.6 | 6.4 | 3.6 | 6.3 |
| HF | 4.8 | 6.4 | 3.7 | 6.5 | 3.7 | 6.6 |

## Human Performance on the Vocabulary Test

The vocabulary MCQ test was given to 85 monolingual English speaking undergraduate student participants and 77 non-monolingual students. Their performance is summarised in Table 2.

Table 2: Performance (\% correct) of monolingual and non-monolingual participants

| MCQs | mean | SD | range |
| :--- | :---: | :---: | :---: |
| Monolingual |  |  |  |
| All 200 | $79 \%$ | $10 \%$ | $56 \%-97 \%$ |
| 100 LF | $72 \%$ | $13 \%$ | $46 \%-96 \%$ |
| 100 HF | $86 \%$ | $8 \%$ | $63 \%-98 \%$ |
| Non-monolingual |  |  |  |
| All 200 | $71 \%$ | $10 \%$ | $49 \%-92 \%$ |
| 100 LF | $61 \%$ | $12 \%$ | $32 \%-88 \%$ |
| 100 HF | $81 \%$ | $10 \%$ | $52 \%-96 \%$ |

Mean performance is high but does not appear to be close to ceiling. The very best performance is close to maximal demonstrating that it is possible for humans to achieve a close to perfect score.

The mean scores for monolingual participants are higher than that for non-monolingual participants for the complete MCQ set and the two subsets.

As would be expected for human performance, performance for the high frequency stems exceeds that for the low frequency ones with the non-monolingual participants demonstrating a larger deficit for low frequency stems. Clearly, human performance has been affected by our manipulation of stem frequency whilst matching the
synonym and distractor frequencies between the low- and high-frequency subsets. The correlation between MCQ question item difficulty (as measured by overall percentage correct) and stem SUBTLEX-UK frequency is 0.41 , suggesting that factors other than word frequency may be affecting human vocabulary test performance.

Three participants achieved overall scores of 97\% - six errors from the 200 MCQ items. These few errors were sometimes made on the same question by more than one participant and were also sometimes also made by the semantic vector methods.

## Semantic Vectors for Evaluation

In addition to testing our new MCQ vocabulary test on human participants, we also used it to evaluate three available sets of semantic vectors, all derived from large text corpora but using contrasting methods to capture the patterns of word usage regularities. Our aim here is to illustrate how well a few recent methods that have been most successful on other semantic tasks are able to perform on this task, and not to make any claims about optimal methods or parameters.

We compared the levels of success on the new vocabulary test using three different kinds of semantic vectors that span the range of approaches available: vectors derived from the methods described by Bullinaria \& Levy (2012), the publicly available semantic vectors that were generated using one of the word2vec (Mikolov et al., 2013) methods (available at: https://code.google.com/archive/p/word2vec/) and the GloVe (Pennington, Socher \& Manning, 2014) vectors derived from the co-occurrence matrix from 6, 42 and 840 billion word corpora available at: http://nlp.stanford.edu/projects/glove/.

The Bullinaria \& Levy ( $\mathrm{B} \& \mathrm{~L}$ ) vectors are derived from the ukWaC (Baroni et al., 2009) (2 billion words) corpus by counting word co-occurrences in a context window of size one and using those counts to generate a starting matrix of positive pointwise mutual information (PPMI) values for about 50,000 target words and 50,000 context words. This choice of window size and co-occurrence statistic was previously shown to be optimal for a range of semantic tasks (Bullinaria \& Levy, 2007) and is now widely used. Singular Value Decomposition (SVD) dimensionality reduction is then used to generate orthogonal matrices $U$ and V and diagonal singular value matrix S such that $\mathrm{M}=$ $\mathrm{USV}^{\mathrm{T}}$, and dimensionality reduction is performed by taking the principal components of $\mathrm{US}^{\mathrm{P}}=\mathrm{MVS}^{\mathrm{P}-1}$ to produce semantic vectors with a dimensionality of 5,000 with the components weighted using a Caron (2001) P value of 0.25 . These parameter values were optimised so as to perform well on four different semantic evaluation measures including a version of the Landauer \& Dumais (1997) TOEFL MCQ vocabulary test, and achieved the current state-of-the art performance on the TOEFL MCQ ${ }^{1}$ test.

The word $2 \mathrm{vec}(\mathrm{W} 2 \mathrm{~V})$ vectors were generated using a 100 billion word sample of the Google News dataset. Word2vec uses supervised learning algorithms to train a simple but
very large neural network model to predict either which words will appear in a window around the current word (the context given the current word) or which word will appear given the current context words. There are a number of different methods and parameters that can be varied in what amounts to a family of techniques. We made use of the publicly available vectors which have 300 dimensions.

The GloVe (G6, G42, G840) vectors were extracted from the files linked to on the GloVe website. The G6 vectors were generated from a 6-billion-word corpus derived from Wikipedia. The 42B and 840B vectors were generated from 42 billion and 840 billion word corpora derived from Common Crawl archives (obtained by an automated process of systematically browsing the web). All the GloVe vectors used here have 300 dimensions. The GloVe method uses regression modelling to learn semantic vectors from the non-zero entries of a word co-occurrence matrix such that the dot product between the vectors for a pair of words equals the logarithm of their probability of co-occurrence. Pennington et al. (2013) show that their vectors perform well on the analogy problem set that was also used to evaluate the word2vec methods.

Levy, Goldberg and Dagan (2015) have argued that the three broad semantic vector techniques used here have similar levels of overall performance when appropriately tuned.

17 of the 1,000 words within the 200 MCQs did not appear in the word2vec vector sets due to differences in UK and USA English. For these words, we used the vectors derived from the USA spelling variants. 7 words did not appear in the GloVe vectors derived from their 6 billion word corpus. For these words, we substituted related words that did appear in the vector set. The other semantic vector sets contained all the 1,000 words used in this MCQ vocabulary set.

Clearly, there are a number of differences in the corpora and parameters used for the three methods and so this exercise cannot reliably compare the success of the different methods in general, but serves as a comparison of a number of different off-the-shelf semantic vector sets.

## Semantic Vector Performance on the Vocabulary Test

We compared the performance of the five different vector sets on all 200 items and on the LF and HF subsets. Mean performance is summarised in Table 3.

Table 3: Performance (\% correct) of the five different vector sets.

| MCQs | B\&L | W2V | G6 | G42 | G840 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| All 200 | 91.0 | 87.0 | 86.5 | 89.5 | 92.0 |
| 100 LF | 93.0 | 87.0 | 86.0 | 92.0 | 95.0 |
| 100 HF | 89.0 | 87.0 | 87.0 | 87.0 | 89.0 |

All three types of semantic vector perform well but not perfectly. None of them match the competence standard of
the WordNet-based MCQ test. The GloVe vectors trained on an 840 billion word corpus comes closest to matching the very best performance of human participants. However, all the vector sets exceed the mean performance of the human participants by large margins.

The LF and HF subsets are distinguished by the SUBTLEX-UK frequencies of their stem words. The frequencies (and word lengths) of synonyms and distractors were matched between subsets. Unsurprisingly, the human participants performed better on the HF subset than on the LF subset, presumably reflecting the association between word frequency and the familiarity that participants had with the stem words. However, in three of the five sets, the semantic vectors performed better on the LF subset than the HF one. Since even the smallest corpus used for generating the semantic vectors was 2 billion words in size, it is likely that all the words used in the vocabulary MCQ test were sampled a great many times and that this overcame any difference in the reliability of the semantic representations due to frequency differences. For the 2 billion word ukWaC corpus that we used (by far the smallest of the corpora used to train the methods compared here), Table 4 gives the frequency statistics for the vocabulary MCQ test synonyms. There is a very large variance in frequency values within each subset. The mean frequency for the LF subset is 3993 with the lowest stem frequency being sampled 98 times in the corpus. The correlation between the $\log _{10} \mathrm{ukWaC}$ corpus synonym frequencies and their SUBTLEX-UK Zipf frequencies is 0.93 .

Table 4: ukWaC frequencies for the stem words in the vocabulary MCQ test

| MCQs | mean | SD | range |
| :--- | :--- | :--- | :--- |
| All 200 | 119,809 | 198,796 | $98-1,057,891$ |
| 100 LF | 3993 | 5099 | $98-36,571$ |
| 100 HF | 235,625 | 228,725 | $12,291-1,057,891$ |

We suppose that any differences in performance for the semantic vectors on the LF and HF subsets is due to an inadvertent bias in the distribution of semantic distances between the MCQ words that is revealed when the statistical reliability related to word frequency differences is made irrelevant due to very high degrees of corpus sampling. It is likely that there is a higher degree of semantic ambiguity for the high frequency stems and this may have affected the MCQ results. We will explore these issues in further detail in modelling work in a future paper.

The corpora used to train the vector generation methods ranged from 2 to 840 billion words. Although performance of the different methods differed by only a few percent, it is notable that the $\mathrm{B} \& \mathrm{~L}$ vectors achieved one of the higher scores using a corpus of 2 billion words and that the GloVe vectors achieved higher scores as the corpus size used increased from 6 to 42 and then 840 billion words. The B\&L method was tuned for previous work on a different
vocabulary MCQ test whilst W2V and GloVe have been reported as having notable success on the rather different domain of analogy problems. It is likely that further parameter tuning would increase the scores of all three methods on this specific task if not in general for other tasks.

Landauer \& Dumais (1997) reported that their LSA semantic vectors performed at a level of $64.4 \%$ on a TOEFL vocabulary MCQ test. This matched the performance of a large sample of applicants to colleges in the USA from nonEnglish speaking countries who averaged a score of $64.5 \%$. This is close to the performance of our non-monolingual group on the 100 low frequency MCQs ( $61.4 \%$ ). These LSA vectors were trained on a much smaller corpus than the other semantic vectors describe here ( 6.4 million words) and so, arguably, are a better psychological model of attaining a degree of semantic competence from a realistic scale of linguistic input. However, they did not approach the high scores required to claim to be a model of idealised semantic competence.

As noted above, there were several MCQs where the same errors were made by some of the highest performing human participants and some of the vector methods. In some of these cases, it appears that the questions were made more difficult than expected by the random choice of distractor items leading to one of the distractors potentially being more closely semantically related to the stem than the intended synonym. An example is the intended stem, synonym, distractor1, distractor2, distractor3 MCQ: benefit, welfare, flask, advantage, lipstick. Here, two of the three highest performing participants and four of the five semantic vector methods made an error. Mean human accuracy was at below chance level. Because the vector methods have captured synonymy well, they show the potential for automatically measuring vocabulary MCQ difficulty in terms of semantic similarity over and above the effect of word frequency.

## Discussion

In cognitive science, we are often interested in building idealised or technologically useful models of intelligent behaviour as well as psychologically valid ones. Ideally, these are complementary aims. The development of methods to generate distributional semantic vectors over the past 20 years is an interesting example of the possible tensions between these two types of model. Landauer \& Dumais (1997) proposed LSA as a model of human semantic performance. LSA was partly validated by its success in matching human non-native performance in the TOEFL MCQ test. However, LSA was not capable of approaching perfect performance on the task. Current techniques have achieved very high levels of success on that task and similar ones such as the test proposed here. However, the amount of data used to train these models is very far in excess of the amount of text that a human could read in a lifetime. The use of distributional semantic vectors in the modelling of human performance (e.g., Pereira et al.,

2015; Mandera et al., 2017) and human brain activity (e.g., Mitchell et al., 2008; Bullinaria \& Levy, 2013) is becoming more widespread due to better and more easily available semantic vectors. However, it remains unclear how the balance between idealised competence and realistic human performance can be modelled by such techniques and which corpora and parameter settings should be used. Some of these issues can be addressed in the straightforward arena of vocabulary MCQ tests.
In this paper, we introduce a vocabulary test, based on WordNet synsets, that is both challenging enough for human participants not to be performed at ceiling and large and uniform enough to be useful as an evaluation measure for corpus-derived semantic vectors.
We argue that we are within reach of developing distributional semantic vectors that can demonstrate competence in the important, if narrow, domain of synonymy judgement. However, there is much to be done in using such representations as components of models that can successfully account for actual human performance on these same tasks.
Although the semantic vectors we tested were close to an idealised level of competence, they do not reflect the clear effect of synonym frequency in the human data. However, the ability of the semantic vectors to provide reliable measures of semantic similarity does show promise for modelling aspects of vocabulary MCQ question difficulty that are left after the influence of word frequency is accounted for.
A single set of semantic vectors cannot both account for idealised semantic competence as defined by a resource such as WordNet and provide a model of average imperfect human performance. For tasks such as vocabulary MCQ tests, it may be necessary to use semantic vectors as models of competence and account for varying performance using simple psychologically valid variables such as word frequency or familiarity. Certainly, current methods for the generation of semantic vectors only obtain very high performance scores after training on enormous corpora, orders of magnitude larger than any human would experience. This may make them poor or partial models of human semantic learning but useful technological tools and cognitive modelling components. It remains to be seen whether semantic vectors with somewhat lesser levels of competence, perhaps trained on much smaller corpora, are better tools for modelling ordinary levels of human performance.

Vocabulary MCQ tests have been useful measures of human word knowledge. In the past they have proved their worth as evaluation measures for semantic vector generation. They are psychological tasks in themselves and we have suggested here that semantic vector methods may allow us to model aspects of question difficulty that are related to relative semantic distances and this may also prove useful for the design of such instruments.
Vocabulary MCQ tests are an important component in the evaluation of representations of lexical semantics. We have
argued that it is important that such representations can account for idealised performance and so reach perfect performance in these tests. Current techniques have not yet reached this level of competence. It would also be highly desirable if these techniques contributed to our ability to model the imperfect performance of human participants on semantic tasks. We argue that vocabulary MCQ tests serve as useful psychological tasks to model. By making our new test freely available along with human data, we hope to stimulate further research.

## Acknowledgments

We acknowledge the funding given by our Universities and thank the student participants in our vocabulary MCQ experiment and our colleagues who facilitated this data collection. We also thank the authors of the $u \mathrm{kWaC}$ corpus, word2vec methodology and GloVe methodology for making their data and code available to us.

## Availability of the MCQ word list

We have made the new vocabulary MCQ test words, and various earlier test sets, available to be downloaded from: http://www.cs.bham.ac.uk/~jxb/corpus.html

## References

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., et al. (2014). A behavioral database for masked form priming. Behavior Research Methods, 46(4), 1052-1067.
Baroni, M., Bernardini, S., Ferraresi, A., \& Zanchetta, E. (2009). The waCky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3), 209-226. http://doi.org/10.1007/s10579-009-9081-4.
Bullinaria, J. A., \& Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39(3), 510-526.
Bullinaria, J. A., \& Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stoplists, stemming, and SVD. Behavior Research Methods, 44(3), 890-907.
Bullinaria, J. A., \& Levy, J. P. (2013). Limiting factors for mapping corpus-based semantic representations to brain activity. PLoS ONE, 8(3), e57191. http://doi.org/10.1371/journal.pone. 0057191
Caron, J. (2001). Experiments with LSA scoring: Optimal rank and basis. In: M. W. Berry (Ed.), Computational Information Retrieval, 157-169. Philadelphia, PA: SIAM.
Fellbaum, C. (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press..
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211.

Levy, O., Goldberg, Y., \& Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 3, 211-225.
Levy, J. P., \& Thompson, N. (2008). Using distributional methods to explore the systematicity between form and meaning in British Sign Language. From Associations to Rules - Connectionist Models of Behavior and Cognition Proceedings of the Tenth Neural Computation and Psychology Workshop, 100-111.
Mandera, P., Keuleers, E., \& Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. Journal of Memory and Language, 92, 57-78. http://doi.org/10.1016/j.jml.2016.04.001
Mikolov, T., Chen, K., Corrado, G., \& Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Mitchell, T. M., Shinkareva, S. V, Carlson, A., Chang, K.M., Malave, V. L., Mason, R. A., \& Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. Science, 320(5880), 1191-5. http://doi.org/10.1126/science. 1152876
Monaghan, P., Shillcock, R. C., Christiansen, M. H., \& Kirby, S. (2014). How arbitrary is language? Philosophical Transactions of the Royal Society B. 369: 20130299. http://dx.doi.org/10.1098/rstb.2013.0299

Pennington, J., Socher, R., \& Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.
Pereira, F., Gershman, S., Ritter, S., \& Botvinick, M. (2015). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. Cognitive Neuropsychology, 33, 175190. http://dx.doi.org/10.1080/02643294.2016.1176907

Shipley,W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. The Journal of Psychology, 9, 371-377.
Schütze, H. (1993). Word space. In: S. J. Hanson, J. D. Cowan \& C. L. Giles (Eds.) Advances in Neural Information Processing Systems 5, 895-902. San Mateo, CA: Morgan Kauffmann.
Turney, P.D. (2001). Mining the Web for synonyms: PMIIR versus LSA on TOEFL. Proceedings of the Twelfth European Conference on Machine Learning (ECML2001), Freiburg, Germany, pp. 491-502.

Turney, P. D., \& Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, 141-188. http://doi.org/10.1613/jair. 2934
Van Heuven, W. J., Mandera, P., Keuleers, E., \& Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. The Quarterly Journal of Experimental Psychology, 67(6), 1176-1190.

# "I'm Better than You at Labeling!": Preschoolers Use Past Reliability when Accepting Unexpected Labels 

Xiaoqian Li (xiaoqian_li@mymail.sutd.edu.sg) \& W. Quin Yow (quin@sutd.edu.sg)<br>Singapore University of Technology and Design, 8 Somapah Road, Singapore


#### Abstract

How do young children decide to trust testimony that contradicts their initial beliefs? The current study examined whether children rely on cues to informant credibility (i.e., history of accuracy) to determine if they would endorse an unexpected label from an informant. Three- and 4-year-olds $(N=60)$ saw a picture of a hybrid artifact that consisted of features of two typical familiar artifacts. Children made initial judgments about the name of the hybrid object and subsequently received a different name offered by an informant who had earlier either accurately or inaccurately named familiar objects. Children were more willing to revise their own judgment and accept the unexpected label if it was from a previously accurate informant than if it was from someone who had made obvious naming errors. This suggests that preschool-aged children selectively revise their own knowledge; they are more trusting toward sources proven accurate than inaccurate.


Keywords: selective trust; accuracy; reliability; unexpected testimony; preschoolers

## Introduction

Children not only learn from their own perceptions and observations of the world around them but also learn from information provided by others. Research on young children has underlined the importance of others' testimony in knowledge acquisition in the early stages of development (see Gelman, 2009; Harris, 2007; Mills, 2013, for reviews). In particular, extensive studies have demonstrated that children show selective trust in testimony depending on an informant's previously established credibility (e.g., Koenig, Clément, \& Harris, 2004; Sabbagh \& Baldwin, 2001; for a review, see Koenig \& Sabbagh, 2013).

Children show sensitivity to the prior accuracy of informants and make persistent use of such information to learn new words and new object functions (Birch, Vauthier, \& Bloom, 2008), learn novel rules (Rakoczy, Warneken, \& Tomasello, 2009), and solve problems (Palmquist \& Jaswal, 2015). This effect of accuracy has been established in studies that present children with two unfamiliar informants who consistently provide accurate or inaccurate information in the context of familiar objects and examine whether these children subsequently prefer the accurate over the inaccurate informant when the two informants offer conflicting novel information (e.g., two different names for the same novel object). By applying such a two-informant paradigm, research has shown that children aged 4 years, and 3 years under certain conditions, selectively endorse new information from the accurate informant over the inaccurate one (e.g., Koenig et al., 2004; Nguyen, Gordon, Chevalier, \& Girgis, 2016).

Although the two-informant paradigm has proven informative, children's selective trust toward one of two informants leaves open questions of how children evaluate and learn from a single informant-a situation in which children are typically involved in everyday interactions (e.g., Lane \& Harris, 2015). Furthermore, the mechanisms underlying children's assessment of the testimony from a single source versus two contrasting sources may be different. For example, 3- and 4-year-olds would endorse new information from a previously inaccurate informant as long as there was no other informant who proposed an alternative (Vanderbilt, Heyman, \& Liu, 2014). This suggests that in general children prefer to endorse testimony from an accurate versus an inaccurate informant, but yet they are willing to trust the testimony of a single inaccurate informant if that is the only testimony available.

The studies presented above are related to children learning something that they do not have any prior knowledge of. However, acquiring new knowledge is plausibly a simpler process than accepting something that conflicts with one's own knowledge or violates one's expectations (hereafter as unexpected testimony). It remains unknown to what extent children's acceptance of unexpected testimony would be influenced by the informant's past accuracy. Specifically, in the absence of another person's testimony, would children still trust a previously inaccurate (or accurate) informant's testimony when it is contrary to one that they have previously formed on their own? A related line of research has shown that 3- to 4 -year-olds are indeed credulous and tend to give up their own beliefs to accept other's statements that are counterintuitive or overtly misleading, particularly when the speaker's communicative intent is salient (e.g., Heyman, Sritanyaratana, \& Vanderbilt, 2013; Lane \& Harris, 2015). Research examining preschoolers’ advice-taking (Rakoczy, Enrling, Harris, \& Schultze, 2015) also demonstrated that 3to 6 -year-olds were more likely to adjust their own social judgments when receiving advice from an expert rather than an ignorant advisor, suggesting that children do keep track of informants' credibility and use such cues in adjusting their own inferences. However, no studies have examined whether an informant's history of accuracy would affect children's willingness to revise their own testimony in face of testimony different from their own beliefs, and in the context where there was no other informant offering alternative testimony.

The current study seeks to use the single-informant paradigm to explore the extent to which 3- and 4-year-olds use informants' past accuracy as a cue to evaluate testimony that contradicts their own beliefs about object labels. The
task was modeled after Vanderbilt et al. (2014) and Jaswal, Lima, and Small (2009). Children were presented with pictures of ambiguous hybrid artifacts each involving features of two typical and familiar objects (e.g., car-shoe), and were asked to provide a name for each hybrid artifact. We examined how children would react to a previously accurate or inaccurate informant who provided an unexpected label, which was not misinformation, but one that always contradicted the children's own beliefs. This would test whether children are willing to give up their own (prior) beliefs to accept the informant's testimony. Based on previous findings that preschoolers show robust sensitivity to informants' past accuracy, we predicted that children would be trusting toward the previously accurate informant and accept this informant's unexpected labels, and that they would be skeptical about the previously inaccurate informant and, therefore, rather not change their initial judgments about the object labels. Additionally, it has been found that children with larger vocabulary (Jaswal, 2007) and those with disadvantaged executive function (Jaswal et al., 2014) were more credulous toward others' testimony than those with smaller vocabulary and advanced executive function skills, respectively. Hence, children's vocabulary and executive function ability were also assessed to control for potential confounds in this study.

## Method

## Participants

A total of 603 - and 4-year-olds ( $M_{\text {age }}=49.64$ months, $S D=$ 6.69 , range $=35-61$ months; 26 girls) participated in this study. Half of them were randomly assigned to the accurate condition and the other half to the inaccurate condition. Children were recruited from private childcare centers in a middle-class neighborhood in Singapore. Only children whose parents had given their consent were included in the study. The majority of participants ( $96 \%$ ) were Asian and the rest were Eurasian. All children spoke English and the experiment was conducted in English. One additional child did not want to provide answers in the test phase and, therefore, was excluded from the final sample. Another eight children participated but were excluded due to experimenter error $(n=2)$ or failure to name familiar objects (see Design and Procedure, $n=6$ ).

## Materials

Video clips of a single informant naming photographs of objects were prepared and shown to children on a 13 -inch laptop computer. The informant was a college-age female actor and with neutral facial expressions. In each video clip, the informant was seated behind a table with a picture placed on it. Pictures of three familiar objects (i.e., apple, ball, and book) were used in the familiarization trials to establish the informant's accuracy. Eight typical exemplars of familiar categories were selected and paired to form four stimulus sets to be used in the test trials (i.e., key-spoon, car-shoe, toothbrush-pen, and hat-cup). For each stimulus
set, features from the pair of typical exemplars were integrated to form a "hybrid" exemplar. These hybrid exemplars were designed such that each hybrid object looked mostly like one of the two typical exemplars of that set (i.e., dominant exemplar) but also shared some features of the other exemplar (i.e., non-dominant exemplar) such as a spoon-like key (see Figure 1).

A total of 14 video clips was made, corresponding to three familiarization clips with the informant naming each of the familiar objects correctly (accurate condition), three familiarization clips with the informant naming each of the familiar objects incorrectly (inaccurate condition), four testimony clips featuring the informant naming the hybrid objects with the labels for the dominant exemplars, and four testimony clips featuring the informant naming the same hybrid objects with the labels for the non-dominant exemplars (see Figure 2).

## Design and Procedure

This study employed a between-participants design. The procedure for both the accurate and inaccurate conditions was the same, except that the history of the informant was established differently during the familiarization phase.

Children were tested individually in a quiet room at their childcare centers. Children were randomly assigned to one of the two conditions. All children were seated in front of a laptop computer and a female experimenter (a different


Figure 1. Hybrid objects used in the experiment.


Figure 2. Screenshots from familiarization video clips in the familiarization phase (top) and testimony video clips in the test phase (bottom).
person from the informant in the video clips) was seated beside the child. The whole procedure was videotaped and children's responses were coded from the videos. Each child received three familiarization trials (Familiarization Phase), followed by four test trials (Test Phase), and two explicit judgment trials. The experiment was designed and written using PsychoPy (Peirce, 2007).

Familiarization Phase In the familiarization phase, the experimenter introduced the children to a single informant, a still image of whom was shown on the screen. In the accurate condition, children were then presented with three familiarization video clips, each showing the informant accurately naming a familiar object (i.e., labeling an apple, a ball, and a book correctly). In the inaccurate condition, children were familiarized with the same informant providing inaccurate information (i.e., saying that the above three objects were a dog, a tree, and a chair, respectively). After each familiarization video clip, children saw a picture of the same familiar object presented in the previous video (without the informant) and were asked by the experimenter, "Can you tell me what this is called?" This question was to make sure that each child knew the correct labels of the familiar objects, implying that the child was able to tell whether the informant had made errors or not. Children did not receive any feedback about whether they or the informant was correct. The order of the presentation of the three familiar objects was the same for all participants.

Test Phase The experimenter proceeded to the test phase where children had to respond to a testimony that conflicted with their initial judgments (i.e., an unexpected but possible label for a hybrid object). Each test trial began with a picture of a typical object (e.g., a key) appeared on the computer screen and the experimenter asked the children for the name of the object. The purpose of the question was to ensure that the children knew the names of the typical objects. Six children were excluded because they failed to name one or more typical objects (two in accurate and four in inaccurate condition). The experimenter then showed a picture of the other typical object from the same stimulus set (e.g., a spoon) on the screen, and children were again asked what that object was called. Children were then presented with a picture of the hybrid object that included features of the two previously shown typical objects (e.g., a keyspoon), and were asked if they knew what that object was called (pre-testimony test). Children's answers to this pretestimony test were recorded as their initial judgments about the name of the hybrid object. The experimenter then told the children, "Okay. Now, let's hear what the girl will say about this." and played a testimony video clip, where the informant and the picture of the same hybrid object appeared in the clip, and the informant always provided a label that was different from what the children had indicated earlier. Subsequently, children were shown the picture of the hybrid object on the computer screen for a second time (without the informant) and were asked by the experimenter
what the hybrid object was called (post-testimony test). During the test phase, children received neutral feedback following a response, regardless of what their answers were (e.g., "Thank you!"). All children completed four test trials. The order of the two typical exemplars of each stimulus set was fixed, such that, for two sets of stimuli (key-spoon and hat-cup), the dominant exemplars were shown first, and for the other two sets (car-shoe and toothbrush-pen), the nondominant exemplars were shown first. The order of the four test sets was randomized for each participant.

Explicit Judgment Trials After children completed all four test trials, the experimenter asked two questions assessing children's evaluation about the informant. With the picture of the informant presented on the screen, children were asked, "Was this girl good or not good at telling the names of the pictures?" Finally, the experimenter showed a picture of a novel object and asked children whether they would seek the informant's help for the name of the object, "If you wanted to know what this new thing was called, would this girl be a good person to ask?" This question was included to replicate Vanderbilt et al.'s (2014) results on children's overwhelming judgment of the accurate/inaccurate speaker as being a good person to ask for the label of a novel object.

In addition to performing the above experiment investigating selective trust in unexpected testimony, all children completed the Peabody Picture Vocabulary Test (PPVT-4; Dunn \& Dunn, 2007), which measures receptive English vocabulary, and the Dimensional Change Card Sort task (DCCS-standard version; Zelazo, 2006), which assesses executive function. Both PPVT and DCCS tasks were administered and scored following standard procedures. For the PPVT, each child obtained a standardized score with a mean of 100 . For the DCCS, children were classified as passing or failing the task based on performance on the post-switch phase of the DCCS.

## Results

The number of times (out of 4 trials) when children adopted the different labels provided by the informant rather than persisting with their initial answers about the names of the hybrid objects on post-testimony test was calculated for each participant as the dependent variable. Thus, the scores indicate how children are willing to trust the informant and accept the informant's testimony that is contrary to their own. Preliminary analyses confirmed no effects or interactions involving children's age, gender, stimulus set, or test order; therefore, further analyses collapsed across these factors. We report $95 \%$ confidence intervals and effect sizes for our statistical tests. In the case of comparisons of group means these confidence intervals refer to the observed mean difference.

We first evaluated whether children in the two conditions had comparable language and executive function abilities. Vocabulary scores on PPVT-4 were not significantly different between the conditions (accurate: $M=95.07$; inaccurate: $M=96.17$ ), $t(58)=-0.35, p=.73$, Cohen's $d=-$
$0.092,95 \%$ CI of the difference $[-7.45,5.25]$. Regarding performance on the DCCS task, children in the two conditions were comparable in terms of the proportion of children who passed the task (accurate: $20 / 30$ passing, $67 \%$; inaccurate: $20 / 30,67 \%), \chi^{2}(1, N=60)=0.00, p=1.00$, Cramer's $V=.00$. Correlational analyses revealed that neither the PPVT scores nor the DCCS scores were related to children's likelihood of endorsing the informant's testimony across the two conditions, $r \mathrm{~s}<.042, p \mathrm{~s}>.74$.

In addition, all children, regardless of the experimental condition, correctly named the familiar objects and the typical exemplars during the familiarization and test phase respectively. Analyses of children's responses in the pretestimony test trials, where children were to name the hybrid objects for the first time, indicated that children provided labels matching the dominant exemplars on $92 \%$ of the trials and labels matching the non-dominant exemplars $8 \%$ of the trials. There were equivalent number of pre-testimony trials where children chose the non-dominant label in accurate and inaccurate conditions (7 vs. 13 trials, respectively). Nevertheless, the informant always provided the label that was different from the children's label for the hybrid object in the testimony video clips (thus unexpected).

Crucially, we were interested in whether 3- and 4-yearolds would respond less credulously to the unexpected testimony from an informant who demonstrated inaccuracy at naming familiar objects compared to an informant with a history of accuracy, that is, whether children would discard their own labels and accept the unexpected labels from the informant in the post-testimony test.

As seen in Figure 3A, children were more willing to revise their answers and accept the unexpected testimony in the accurate condition ( $M=3.60,95 \% \mathrm{CI}[3.15,4.06]$ ) than in the inaccurate condition $(M=1.33,95 \% \mathrm{CI}[0.88,1.79])$, $U=123.50, z=-5.12, p<.001, r=-.66$. Examining patterns of individual behavior revealed similar differences between accurate and inaccurate conditions. We calculated the number of children in each condition who endorsed the unexpected testimony on $0,1,2,3$, or 4 out of four trials. Chi-square tests showed that the distribution of children across various patterns of responses was different between the accurate and inaccurate conditions, $\chi^{2}(4, N=60)=$ 30.64, $p<.001$, Cramer's $V=.72$. There were more children who accepted the informant's testimony on 3 or 4 trials in the accurate condition $(n=28)$ than the inaccurate condition $(n=7)$. None of the significance levels were affected by removing data from test trials where the child chose the non-dominant label in the pre-testimony test before the informant providing different testimony.

Lastly, children's responses to the two explicit judgment questions were analyzed ${ }^{1}$. In line with the patterns of information endorsement, significantly more children in the accurate condition agreed that the informant was good at

[^386]naming the pictures compared with those in the inaccurate condition $(90.0 \%$ vs. $51.9 \%), \chi^{2}(1, N=57)=10.24, p=$ .003 , Cramer's $V=.42$ (see Figure 3B). However, there was no significant effect of condition on the question assessing children's willingness to seek help from the informant for the label of a novel object, $\chi^{2}(1, N=59)=2.59, p=.18$, Cramer's $V=.21$ (see Figure 3C). Children overwhelmingly judged that the informant, regardless of whether she showed a history of accuracy or inaccuracy, would be a good person to ask about the label of a novel object- $89.7 \%$ of children did so in the accurate condition, and it was the case for $73.3 \%$ of children in the inaccurate condition. These rates were in a similar range as those reported by Vanderbilt et al. (2014) where children were asked to judge whether an accurate or inaccurate source, either alone or paired with each other, would be a good person to ask for labels of novel objects ( $76 \%-90 \%$ ).

## Discussion

The current study examined the extent to which an informant's history of accuracy influenced children's endorsement of claims that conflicted with their independent beliefs about ambiguous hybrid artifacts. Using a single-informant paradigm, we found that 3- and 4-yearolds' trust toward unexpected testimony differed depending on the informant's past accuracy. Specifically, while children consistently revised their own initial judgments and endorsed unexpected testimony from an informant who appeared to be accurate and knowledgeable about common objects, they were less likely to do so in response to someone who made naming errors with these common objects. This study provided novel findings that children consider an informant's previous epistemic history when determining whether or not to revise their own prior beliefs in light of unexpected testimony.

The present results challenge the notion that in the absence of conflicting testimony from another informant, young children generally trust a single informant who has a record of inaccuracy. This notion of trust toward a single inaccurate informant may be true only if children themselves do not hold any conflicting information at all. Vanderbilt et al. (2014) found that 3- and 4-year-olds were willing to trust the testimony of an inaccurate informant when there was no other informant offering an alternative. In contrast, children in the current study were less willing to accept the testimony provided by a single inaccurate informant in the absence of competing testimony from another informant. An important difference between this study and Vanderbilt et al.'s (2014) work, however, is that children in this study held a different interpretation of an object than the informant, whereas children in the previous study did not. Thus, the present results suggest that children's own prior knowledge play an important role in selective trust, and that children would evaluate all available sources of information, including themselves, when determining whom to trust. In situations where children are mostly ignorant, such as labeling unfamiliar


Figure 3. Evaluation of the informant by condition: (A) number of trials (out of 4) children adopting unexpected testimony from the informant, (B) percentage of children agreeing the informant being good at naming pictures, and (C) percentage of children being willing to seek novel information from the informant. Error bars represent $95 \%$ confidence intervals.
objects, children are likely to perceive an informant who confidently provides testimony as more knowledgeable than them, even if the informant has made errors previously. Whereas in situations where children possess some prior knowledge, albeit loosely formed ones, such as labeling ambiguous hybrid objects in this study, children would evaluate the relative trustworthiness of the source of information against themselves.

How do children determine the trustworthiness of an informant in relation to themselves? Past accuracy is one important factor. In the current study, children were more willing to discard their own beliefs in favor of an adult informant's testimony when this adult had been proven accurate compared to an inaccurate informant. However, it remains an open question that whether the current results were due to a negative bias toward the inaccurate source of information, or due to both a preference for the accurate source and an avoidence of the inaccurate one. Future studies could further investigate this question by examining children's responses in a control condition where no history of accuracy/inaccuracy would be provided. Age is another factor. Children in the accurate condition might perceive the accurate adult as a more credible source of information about what artifacts are called than them. In fact, 3- and 4-year-olds were more willing to learn novel labels from an adult than from a child when both were equally reliable (Jaswal \& Neely, 2006). However, children appeared to weigh accuracy over age in selective trust; they were found to trust a previously accurate child more than a previously inaccurate adult when learning new words. This is also true in the inaccurate condition reported here, such that children's distrust toward inaccuracy was so robust that they assumed that an inaccurate adult was less reliable than them (a child). These results suggest that children consider multiple factors when evaluating the trustworthiness of another source of information compared with them, but weigh certain factors more than others (i.e., accuracy over age).

Even though the present results showed that children demonstrated a reduced tendency to accept the testimony against their own judgments in the inaccurate condition, it remains debatable whether it was because the children believed that the inaccurate informant was not trustworthy, or that they simply had alternative information available (i.e., their own), or both. Our results showed that children gave up their own beliefs and accepted the testimony from an accurate informant on an average of $90 \%$ of trials (nearceiling), yet they were still willing to give up their own beliefs and accept the inaccurate informant's testimony on an average of $33 \%$ of trials (a $0 \%$ would indicate absolute rejection). This implies that children may be more ready to accept an adult informant's testimony than to reject it, even when the adult informant had been inaccurate previously, and even when the testimony conflicted with their own, at least possibly until they are provided with stronger evidence of the negative credibility of the adult informant (Ronfard \& Lane, in press).

Children's degree of selective trust may thus be affected by various factors that reflect the extent of an informant's credibility. For instance, children were found to be more forgiving with errors in the episodic domain (e.g., locations of objects) than in the semantic domain (e.g., names of objects); they used semantic errors but not episodic errors when evaluating informants' trustworthiness in labeling objects (Palmquist \& Jaswal, 2015). It is unknown whether children would remain skeptical toward unexpected testimony from an inaccurate informant who made episodic errors. Furthermore, hybrid artifacts were used in this study and the unexpected labels provided by the informant were always possibly "correct" and not entirely wrong, as the labels did contain some features of the hybrid, although they were contrary to children's inferences. It is unknown whether children would still be willing to accept the accurate informant's unexpected labels if the labels were not possibly correct (e.g., calling a spoon-like key a cat). Further research is needed to investigate how children's
selective trust may change depending on the types of errors made by the informants and when the unexpected labels are not actually possible.

Last but not least, children's openness to alternative information may be dependent on the strength of their initial beliefs. Indeed, Chan and Tardif (2013) found that 6-yearolds were more accepting an alternative when they felt less certain about their own prior knowledge. In the current study, children who chose the non-dominant label in the pre-testimony trials might be less certain about their answer and more prone to revise. Therefore it was important to control for children's initial choice (there were only limited number of trials with the non-dominant label thus they were excluded).

To conclude, the current study showed that when confronted with different testimony from others, young preschoolers selectively revised their own inferences depending on the informant's past accuracy. Young children are savvy in that they can use such credibility cues to evaluate another individual who holds different opinions from them and decide whether to adjust their own beliefs or not accordingly. The ability to appropriately evaluate the reliability of various sources of information and update their own knowledge correspondingly is important, since reliable sources allow children to learn efficiently while unreliable sources increase the risk of being misinformed. Our findings suggest that this ability is emerging in 3- and 4 -year-old children.

## Acknowledgments

We are grateful to the children and parents who participated and to the teachers and staff of the Creative O Preschoolers' Bay, Red SchoolHouse, MacPherson Sheng Hong Childcare Centre, and Safari House Preschool. We thank Eilise Francesca for serving as the actor in our videos. This research was supported by Singapore University of Technology and Design President's Graduate Fellowship awarded to the first author.

## References

Birch, S. A. J., Vauthier, S. A., \& Bloom, P. (2008). Threeand four-year-olds spontaneously use others' past performance to guide their learning. Cognition, 107, 1018-1034.
Chan, C. C. Y., \& Tardif, R. (2013). Knowing better: The role of prior knowledge and culture in trust in testimony. Developmental Psychology, 49, 591-601.
Dunn, L. M., \& Dunn, D. M. (2007). Peabody Picture Vocabulary Test. Bloomington, MN: NCS Pearson Inc.
Gelman, S. A. (2009). Learning from others: Children's construction of concepts. Annual Review of Psychology, 60, 115-140.
Harris, P. L. (2007). Trust. Developmental Science, 10(1), 135-138.
Heyman, G. D., Sritanyaratana, L., \& Vanderbilt, K. E. (2013). Young children's trust in overtly misleading advice. Cognitive Science, 37, 646-667.

Jaswal, V. K. (2007). The effect of vocabulary size on toddlers' receptiveness to unexpected testimony about category membership. Infancy, 12(2), 169-187.
Jaswal, V. K., Lima, O. K., \& Small, J. E. (2009). Compliance, conversion, and category induction. Journal of Experimental Child Psychology, 102, 182-195.
Jaswal, V. K., \& Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. Psychological Science, 17(9), 757-758.
Jaswal, V. K., Pérez-Edgar, K., Kondrad, R. L., Palmquist, C. M., Cole, C. A., \& Cole, C. E. (2014). Can't stop believing: Inhibitory control and resistance to misleading testimony. Developmental Science, 17(6), 965-976.
Koenig, M. A., Clément, F., \& Harris, P. (2004). Trust in testimony: Children's use of true and false statements. Psychological Science, 15, 694-699.
Koenig, M. A., \& Sabbagh, M. A. (2013). Selective social learning: New perspectives on learning from others. Developmental Psychology, 49, 399-403.
Lane, J., \& Harris, P. (2015). The roles of intuition and informants' expertise in children's epistemic trust. Child Development, 86(3), 919-926.
Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. Developmental Psychology, 49, 404-418.
Nguyen, S. P., Gordon, C. L., Chevalier, T., \& Girgis, H. (2016). Trust and doubt: An examinication of children's decision to believe what they are told about food. Journal of Experimental Child Psychology, 144, 66-83.
Pamquist, C. M., \& Jaswal, V. K. (2015). Preschoolers' inferences about pointers and labelers: The modality matters. Cognitive Development, 35, 178-185.
Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. Journal of Neuroscience Methods, 162(1-2), 813.

Rakoczy, H., Enrling, C., Harris, P. L., \& Schultze, T. (2015). Young children heed advice selectively. Journal of Experimental Child Psychology, 138, 71-87.
Rakoczy, H., Warneken, F., \& Tomasello, M. (2009). Young children's selective learning of rule games from reliable and unreliable models. Cognitive Development, 24(1), 61-69.
Ronfard, S., \& Lane, J. (in press). Preschoolers continually adjust their epistemic trust based on an informant's ongoing accuracy. Child Development.
Sabbagh, M.A., \& Baldwin, D.A. (2001). Learning words from knowledgeable versus ignorant speakers: links between preschoolers' theory of mind and semantic development. Child Development, 72, 1054-1070.
Vanderbilt, K. E., Heyman, G. D., \& Liu, D. (2014). In the absence of conflicting testimony young children trust inaccurate informants. Developmental Science, 17(3), 443-451.
Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. Nature Protocols, 1, 297-301.

# How does Music Reading Expertise Modulate Visual Processing of English Words? An ERP study 

Tze Kwan, Li (saralis@hku.hk)<br>Hei Yan Veronica, Chan (veron919@connect.hku.hk)<br>Luhe Li (rhys@connect.hku.hk)<br>Janet H. Hsiao (jhsiao@,hku.hk)<br>Department of Psychology, University of Hong Kong<br>627 Jockey Club Tower, Pokfulam Road, Hong Kong SAR


#### Abstract

Music notation and English word reading have similar visual processing requirements. It remains unclear how the two skills influence each other. Here we investigated the modulation of music reading expertise on visual processing of English words through an ERP study. Participants matched English real, pseudo, and non-words preceded by musical segments or novel symbol strings in a sequential matching task. Musicians showed smaller N170 amplitude in response to English non-words preceded by musical segments than by novel symbol strings in the right hemisphere. This effect was not observed in real or pseudo-words, or in any of nonmusicians' responses. Similar to English non-words, musical segments do not have morphological rules or semantic information, giving rise to this modulation effect. This finding suggested a shared visual processing mechanism in the right hemisphere between music notation and English non-word reading, which may be related to serial symbol processing as suggested by previous studies.


Keywords: Music reading expertise; EEG; event-related potential (ERP); English word reading

## Introduction

Recent research has shown that different perceptual expertise domains can influence each other. For example, car perception was interfered by concurrent face perception in car experts (presumably also face experts) but not in car novices, suggesting shared neural processing mechanisms between car and face recognition expertise (Gauthier, Curran, Curby \& Collins, 2003). In an ERP study, Rossion, Kung, and Tarr (2004) showed that expertise with Greebles led to a decrease in N170 in response to faces with concurrent Greeble presentation, suggesting competition between expertise domains in early perceptual processing.

Similarly, music notation and English word reading expertise may influence each other due to their similarities in visual processing. For example, both music notation and English word reading involve decomposing visual input into components (i.e., letters or notes) for mapping to components in sounds (i.e., phonemes or pitches; Brown, Martinez \& Parsons, 2006; Hsiao \& Lam, 2013). The requirement of grapheme-phoneme conversion in English word reading has been suggested to lead to a strong left hemisphere (LH) lateralization. For example, a right visual field (RVF)/LH advantage has been found in word naming (e.g. Brysbaert \& d'Ydewalle, 1990). Consistent with these findings, fMRI studies have shown a region inside the left fusiform area
responding selectively to words (e.g. McCandliss, Cohen, \& Dehaene, 2003). ERP studies showed that English words elicited larger N170 amplitude in the LH than the RH in a repetition detection task (Maurer, Brandeis \& McCandliss, 2005). This LH lateralization may be attributed to the leftlateralized phonological processing (Rumsey et al., 1997).

Similarly, in music notation processing, Segalowitz, Bebout, and Lederman (1979) reported a RVF/LH advantage in chord playing, which may be related to the requirement of mapping individual notes to different pitches/fingerings. Indeed, music notation and English word reading are shown to have shared neural mechanisms in the LH. For example, musicians with brain lesions in the LH showed difficulties in both music and English word reading (Hébert \& Cuddy, 2006). Also, both English and music notations are read from left to right, and thus letters and music notes are recognized in the RVF more often than the left visual field (LVF) during reading, resulting in a similar RVF processing advantage due to perceptual learning (Wong \& Hsiao, 2012).

While the LH is shown to play an important role in English word and music notation reading, the RH is also involved, particularly in visual form processing of words and notes. For example, in a lexical decision priming task, English word processing in the LVF/RH was shown to benefit from orthographically similar primes, whereas that in the RVF/LH benefitted from phonologically similar primes. This result suggested that the RH and the LH had differential advantages in orthographic and phonological processing of English words (Lavidor \& Ellis, 2003). Consistent with this finding, English word processing in the RH has been reported to be more sensitive to variations in visual word forms. For example, the word length effect in English lexical decisions (i.e., faster and more accurate responses to shorter words) was only observed when words were presented in the LVF/RH but not the RVF/LH, suggesting that RH word processing involves more letter-by-letter recognition/serial processing than that in the LH (Lavidor \& Ellis, 2001). Similarly, in music note processing, a right lateralized or bilateral visual processing mechanism has been observed. For example, fMRI studies have shown that the right occipitotemporal region was associated with music sightreading (Schön, Anton, Roth \& Besson, 2002). Bilateral activations in the fusiform and inferior occipital gyri in musicians were also reported in a note selection task. (Proverbio, Manfredi, Zani \& Adorni, 2013). In a divided
visual field study, no lateralization effect was observed in sequential matching of notes and chords (Li \& Hsiao, 2015).

Although previous research has suggested similarities between English word and music notation reading processes, it remains unclear how they influence each other. We have previously found that, whereas non-musicians showed a typical RVF/LH advantage in naming English words, musicians showed an LVF/RH advantage and responded significantly faster than non-musicians when words were presented in either the LVF or the center position (Li \& Hsiao, 2015). This effect suggested a facilitation of RH English word processing due to music reading experiences. This phenomenon may be due to shared neural mechanisms between the two expertise domains in the LH that lead to resource competition, consequently making musicians rely more on RH processing for English word recognition. It may also be the similarities between music notation and English word reading processes in the RH accommodate each other, making the relevant processes more efficient and consequently facilitating RH English word processing.

While English word and music notation reading share similar visual processing requirements, they differ significantly in their involvement in lexical processing. More specifically, English words follow morphological and orthographic rules with clearly defined segment boundaries and lexical representations, whereas musical segments do not follow as strict sequencing rules as words and are not associated with specific semantic representations (Chan \& Hsiao, 2016). Since previous research has suggested that LH English word processing is more relevant to phonological processing of English words whereas RH English word processing is more sensitive to variations in visual word forms, the modulation of music reading experience on visual processing of English words is likely to be mainly due to a shared processing mechanism in the RH. In addition, this modulation may be stronger in English non-word processing than the processing of real or pseudo-words, since nonwords do not follow morphological rules or have meanings, similar to musical segments. To test these hypotheses, here we conduct an EEG study to examine how music reading expertise influences visual processing of English stimuli. A sequential matching task is used to focus on visual processing of English words. Following Rossion et al. (2004), here we examine how N170 responses to English words are influenced by the processing of music notes in musicians and non-musicians. We expect that musicians will have a stronger reduction in N170 response to English stimuli under the processing of music notes than non-musicians in the RH, particularly for English non-words.

## Methods

## Participants

Participants were 60 Cantonese (L1)-English (L2) bilinguals from Hong Kong, whose ages ranged from 18 to 29 ( $M=$ $21, S D=2.8$ ). They had similar language and college education backgrounds, with normal or corrected to normal vi-
sion. They were categorized as 30 musicians ( 14 males, 16 females) and 30 non-musicians ( 12 males, 18 females).

Musicians were well-trained pianists, who started music training at age 3-8 $(M=5.33, S D=1.47)$. All of them were either piano teachers, music major students, or frequent piano players. They had attained grade 8 or above in the graded piano examinations of the Associated Board of The Royal Schools of Music (ABRSM), with 8-25 year experience in piano playing ( $M=15.03, S D=3.89$ ) and regular music reading hours per week ( $M=7.16, S D=12.33$ ). Musicians outperformed non-musicians in musicality, as assessed by the Goldsmiths Musical Sophistication Index (Müllensiefen, Gingras, Musil, \& Stewart, 2014; $t(58)=9.97, p<.001)$. In contrast, non-musicians did not receive any music training.

Aside from their music background, musicians and nonmusicians were closely matched in handedness and language exposure. Most participants were right-handed, which was assessed using the Edinburgh Handedness Inventory (Oldfield, 1971; M: 54.33, $3^{\text {th }}$ right decile; NM: 64.33, $3^{\text {rd }}$ right decile, n.s.). All participants started learning English as a second language at age 3 , and have similar self-reported English reading hours (M: 27.48; NM: 18.77; n.s.). No participants had experience with the Tibetan language.

## Materials

Materials consisted of 3 types of English words (real, pseudo, and non-words with 4-6 letters) as target stimuli and two types of comparable pre-/post-stimulus masks: musical segments with 4 random notes without clefs ( $n=1323$ ) ranging from D4 to G5 and Tibetan letter strings with 4 random letters ( $n=1323$ ). Tibetan letter strings, a novel stimulus type that no participants had any experience with, were included as a control condition.

English real words ( $n=126$ ) were selected from the SUBTLEX-US corpus (Brysbaert, New \& Keuleers, 2012) and Wuggy (a word generator, Keuleers \& Brysbaert, 2010). To control information distribution within a word, the same number of high-frequency words and low-frequency words were selected within the informative beginning and informative end subsets in Bryden, Mondor, Loken, Ingleton \& Bergstrom (1990). Word frequency was closely matched between 'same' and 'different' trials in the matching task and between music and Tibetan conditions. For 'same' trials, two target stimuli were identical. For 'different' trials, half trials had shared beginnings (e.g. banker, banner), while the other half had shared ends (e.g. salary, notary).

English pseudo-words (i.e. non-existing words with legal letter strings at the word beginning and word end, $n=126$ ) were created by extracting and recombining word beginnings and ends from our English real word list. This is to control information distribution at the word beginnings and ends between real and pseudo-word stimuli. For 'same' trials, two target stimuli were identical. For 'different' trials, half trials had shared beginnings (e.g. banher, banord), while the other half had shared ends (e.g. saliew, supiew).

English non-words (i.e., illegal letter strings, $n=126$ ) were created by re-ordering the letters in the word begin-
nings and word ends from our English pseudo-word list such that the letter combinations do not follow morphological rules in English. This is to closely match the letters used in all conditions. For 'same' trials, two target stimuli were identical. For 'different' trials, half trials had shared beginnings (e.g. nbaerh, nbaodr), while the other half had shared ends (e.g. alsiwe, spuiwe). The non-words were checked against the morphologically ambiguous syllables in the ARC Nonword database (Rastle, Harrington, \& Coltheart, 2002) to ensure their suitability for our task.

## Design

To focus on visual processing of English words, a sequential matching task similar to Gauthier et al. (2003) was used. The design consisted of 2 within-subject variables: English word type (real/pseudo/non-words), stimulus mask (musical segments vs. Tibetan letter strings), and 1 between-subject variable: group (musicians vs. non-musicians). In the ERP data analysis, an additional variable hemisphere (LH vs. RH) was included. Participants completed the task with English real, pseudo, and non-word stimuli with either musical segment or Tibetan letter string masks (Fig. 1). For each mask type, 36 'same' and 36 'different' trials were included for each word type condition. Half of the stimulus pairs in 'same' and 'different' trials were different in the two mask conditions to avoid practice effects.

English words were displayed in Courier (a serif font with fixed width) to ensure constant center-to-center spacing between letters. Under the viewing distance 50 cm , each English word subtended a horizontal and vertical visual angle of $4.06^{\circ} \times 0.95^{\circ}$ (4 letters), $5^{\circ} \times 0.95^{\circ}$ ( 5 letters) and $6.35^{\circ} \times 0.95^{\circ}$ ( 6 letters). Musical segments with 4 random notes in crotches ( 1 beat) with the five-line staff subtended a horizontal and vertical visual angle of $6.90^{\circ} \times 1.62^{\circ}$. Tibetan letter strings with 4 random letters were presented in Himalaya font and subtended a horizontal and vertical visual angle of $6.90^{\circ} \times 1.62^{\circ}$. All stimuli were presented in black with a white background on a CRT monitor. Experiments were conducted using E-Prime v2.0 with 64-channel ANT EEG recording. A chinrest was used to reduce head movement. The block and trial orders were randomized.

## Procedure

Each trial started with a central fixation with a randomly determined presentation duration between $400-600 \mathrm{~ms}$. A pre-stimulus mask (a musical segment or a Tibetan letter string) was presented for 600 ms , followed by an 800 ms presentation of the first target stimulus (a real/pseudo/non word). Then, a post-stimulus mask (a musical segment or a Tibetan letter string) was presented for 600 ms , followed by an 800 ms presentation of the second target stimulus (a same or different real/pseudo/non-word; Fig. 1). All stimuli were presented at the center of the screen. Participants judged whether the two target stimuli were the same or not by pressing buttons with both hands. The trial did not proceed to the 800 ms 'blink' period until receiving partici-
pants' response. Accuracy (ACC) and response time (RT) were recorded by Eprime with EEG recording.

Prior to the English word sequential matching task, a demographic and music background questionnaire, the Goldsmiths Musical Sophistication Index (Müllensiefen et al., 2014) and Edinburgh Handedness Inventory (Oldfield, 1971) were conducted to assess participants' language, music background, and handedness.


Fig. 1. Procedure of the English word sequential matching

## Results

In the English word sequential matching task, no significant difference was observed between musicians and nonmusicians in ACC and RT of matching real (M: 97.27\%, 606.02 ms ; NM: $94.31 \%, 774.41 \mathrm{~ms}$ ), pseudo (M: 97.04\%, 619.90 ms ; NM: $93.29 \%, 698.58 \mathrm{~ms}$ ) and non-words (M: $95.88 \%$, 598.58 ms ; NM: $91.20 \%$, 727.38 ms ), suggesting that they had a similar performance level in the task.

The 64-channel EEG data were analyzed using EEGLAB (Delorme \& Makeig, 2004) and ERPLAB (LopezCalderon \& Luck, 2014) in MATLAB. Bin-based epochs were extracted from -100 ms to 600 ms of the stimulus onset and corrected from baseline deviations using a prestimulus window of 99 ms . The analyses of the N170 component were based on the electrode pairs with the largest N170 amplitude from the grand average data. Accordingly, electrodes PO7 (LH) and PO8 (RH) were selected for the analysis of N170 response to the pre-stimulus masks (musical segments vs. Tibetan letter strings), while electrodes P7 (LH) and P8 (RH) were selected for N170 responses to the first presentation of the English real, pseudo, and non-words preceded by musical segments or Tibetan letter strings, using repeated measures ANOVA. Note that we only analyzed the N170 responses to the first presentation of the English word stimuli since the EEG responses to the second stimulus may be contaminated by button responses..


Figure 2. Average N170 amplitude at PO7 and PO8 in response to musical segments and Tibetan letter strings (error bars $=+/-1 S E ; * * * p<.001, * p<.05)$.

In the ERP response to the pre-stimulus mask, a significant interaction between mask type (music vs. Tibetan) $\times$ group (musicians vs. non-musicians) was observed, $F(1,52)$ $=31.80, p<0.001$ : musicians had a larger N170 amplitude than non-musicians in response to musical segments, $t(52)=$ $-2.07, p=.044$ (Fig. 2), whereas no difference was observed between the two groups in response to Tibetan letter strings. When we split the data by group, musicians had a larger N170 amplitude in response to musical segments than to Tibetan letter strings, $F(1,27)=68.98, p<0.001$ (Fig. 2), whereas non-musicians did not show any significant differences in response to musical segments and Tibetan letter strings. These findings were consistent with the perceptual expertise literature showing that visual expertise increases the N170 amplitude in response to the stimuli in experts as an expertise marker (Rossion et al., 2004). No main effects or interactions with hemisphere were observed (Fig. 3).


Figure 3. N170 amplitude in response to musical segments and Tibetan letter strings between musicians and nonmusicians in $\mathrm{PO} 7(\mathrm{LH})$ and $\mathrm{PO} 8(\mathrm{RH})$.


[^387]Figure 4. N170 amplitude in response to English (a) real, (b) pseudo and (c) non-words preceded by musical segments and Tibetan letter strings between musicians and nonmusicians in P7 (LH) and P8 (RH) in sequential matching.

For N170 responses to English words (the first target stimulus), a significant four-way interaction, mask type (music vs. Tibetan) $x$ word type (real vs. pseudo vs. nonwords) x hemisphere (LH vs. RH) x group (musicians vs. non-musicians), was observed, $F(2,53)=3.32, p=.044$. To better understand this interaction, we examined the N170 amplitude in response to real, pseudo, and non-words separately (Fig. 4). A significant interaction among mask type, hemisphere, and group was found in English non-words, $F$ $(1,54)=6.27, p=.015$, but not in real or pseudo-words. This three-way interaction suggested that musicians and non-musicians had different N170 amplitudes in response to non-words preceded by musical segments and Tibetan letter strings in the LH and the RH. This effect was not found in real or pseudo-words.

When we examined the data of non-words in two participant groups separately, musicians showed a significant interaction between mask type (music vs. Tibetan) and hemisphere (LH vs. RH), $F(1,26)=10.60, p=.003$, whereas non-musicians did not. When we examined musicians' data in the two hemispheres separately, a significant main effect of mask type (music vs. Tibetan) was observed, $F(1,26)=$ 9.004, $p=.006$ : musicians had a smaller N170 amplitude in response to English non-words preceded by musical segments $(-2.17 \mu \mathrm{~V}, S D=3.88$, Fig. 5) than those preceded by Tibetan letter strings in the RH $(-4.11 \mu \mathrm{~V}, S D=2.11)$. This mask type effect was not observed in the LH. Note that this mask type effect was also not observed in either participants' N170 responses to real and pseudo-words, or nonmusicians' N170 responses to non-words. This phenomenon demonstrates a modulation of musicians' musical segment processing on English non-word processing in the RH.


Figure 5. Musicians had a greater reduction in N170 amplitude in response to non-words preceded by musical segments than that preceded by Tibetan letter strings in the RH. No reduction effect was observed in the LH or in nonmusicians. (error bars $=+/-1 S E ; * * p<.01$ ).

## Discussion

Here we examined how music reading expertise influences visual processing of English stimuli. Since music notation reading does not involve semantic processing as English word reading does, we hypothesized that the modulation of music reading experience on English word processing would be mainly in the RH, which is shown to be important for visual form processing of English words. In addition, the modulation would likely be stronger in English non-word processing than the processing of real or pseudo-words, since similar to musical segments, non-words do not follow morphological/orthographic rules. Consistent with our hy-
potheses, we showed that musicians had a reduced N170 amplitude in response to English non-words preceded by musical segments as compared with that preceded by novel symbol strings in the RH, whereas non-musicians showed no difference in N170 response to non-words preceded by either musical segments or Tibetan letter strings. In addition, this reduction in N170 in musicians was only observed in non-words, but not in real or pseudo-words. This result suggests a shared neural mechanism between English nonword and musical segment processing in the RH.

The RH N170 modulation effect of musical segments in musicians was only observed in English non-words but not in real or pseudo-words. This effect suggests that the interaction between visual English word and music notation processing depends on the similarities of the cognitive processes involved. More specifically, in contrast to English real and pseudo-words, non-words and musical segments do not follow any morphological or orthographic rules (Chan \& Hsiao, 2016). Given that they share similar global forms, containing components of similar sizes arranged horizontally, their recognition may both rely on component by component serial processing, giving rise to the modulation effect. Consistent with this speculation, a RH advantage in the perception of global forms has been consistently reported (Sergent, 1982). English word processing in the RH is also shown to be more sensitive to variations in visual word forms than the LH, such as words in case alternation (Lavidor \& Ellis, 2001). In particular, Lavidor and Ellis (2001) found that the word length effect in English lexical decisions (i.e., faster responses to shorter words) was observed only when words were presented in the LVF/RH but not in the RVF/LH. However, when words in MiXeD CaSe were used, encouraging letter-by-letter processing, the word length effect was observed in both visual fields. These results suggest a letter-by-letter, serial processing engaged in the RH word recognition, in contrast to a left-lateralized automated, whole-word lexical processing unaffected by word lengths (see also Lavidor, Ellis, \& Pansky, 2002). Similarly, patients with LH lesions retained letter-by-letter reading ability, suggesting that the nature of RH word processing involves letter-by-letter recognition (Cohen et al., 2004). Our results here were consistent with these findings, suggesting that RH English word processing was modulated by music notation reading experience due to their similarity in letter-by-letter or note-by-note visual processing. Consistent with our finding, in an fMRI study, Proverbio et al., (2013) reported that musicians recruited the right fusiform gyrus and the right inferior occipital gyrus in an orthographic letter recognition task, whereas non-musicians showed activations at the corresponding regions in the LH. This finding again suggests that music reading expertise modulates English word reading in the RH.

This RH modulation effect of music reading expertise was also consistent with our recent study showing that musicians named English words faster than non-musicians when words were presented in the LVF/RH (Li \& Hsiao, 2015). More specifically, this LVF/RH advantage in word naming
in musicians may be due to the facilitation of shared neural information processing mechanisms in the RH between music notation and English word reading, resulting in a transfer effect from music note to English word processing in the RH. Note that in the current study, the lack of the N170 modulation effect in real and pseudo-words does not necessarily mean that this modulation from music notation reading experience does not affect real word and pseudo-word processing. English word recognition involves the processing of visual word forms, phonology, and semantics. While the LH is shown to involve critically in lexical processing, the RH is reported to be important for the processing of visual word forms. Our current results suggest that the modulation of music experience is mainly in the RH. Since the processing of real and pseudo-words involves both visual word form and lexical/sublexical processing, these lexical effects may also influence N170 amplitudes measured in both hemispheres. Indeed, Ziegler et al. (1997) showed that real and pseudo-words elicited more negative early visual ERPs than non-words in bilateral posterior regions in a lexical decision task, with this difference appearing earlier in the LH than the RH. Thus, the RH N170 modulation effect of music reading expertise may have been contaminated by lexical/sublexical effects in real and pseu-do-word processing. It is also possible that the lack of the modulation effect in real and pseudo-word processing is because random musical segments were used. Future work will examine whether musical segments from real musical pieces (motifs) will have different modulation effects.

Note also that the current results do not rule out possible modulation effects of music reading experience on phonological processing of English words, since our task, sequential matching, involved mainly visual word processing. Previous studies have reported benefits of music training on the phonological processing of English words, as shown in phonological skill training (Degé \& Schwarzer, 2011). Thus, musicians' LVF/RH advantage in English word naming over non-musicians observed in our previous study (Li \& Hsiao, 2015) could also be related to modulation effects of music reading experience on English phonological processing in the LH. Future work will examine this possibility.

In short, here we show that music notation and English non-word processing share similar neural mechanisms in the RH, as demonstrated in the reduced N170 responses to English words under the processing of musical segments. This effect was not observed in real or pseudo-words. Similar to English non-words, musical segments do not follow orthographic rules. Their processing may rely on serial processing of horizontally arranged components of similar sizes, giving rise to the modulation effect. This effect demonstrates that the interaction between different perceptual expertise domains depends on the similarities of the cognitive processes involved. Future work may use Korean Hangul stimuli, in which letters are arranged into a square shape instead of horizontally, to examine whether the modulation effect of music reading expertise in the RH was restricted to words with a global form similar to music notations (i.e.,
components of a similar size arranged horizontally) or could be applied to words in alphabetic languages in general.

## Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (ECS scheme project \# HKU 758412H to J.H. Hsiao).

## References

Brown, S., Martinez, M. J., \& Parsons, L. M. (2006). Music and language side by side in the brain: a PET study of the generation of melodies and sentences. Eur. J. Neurosci., 23, 2791-2803.
Bryden, M. P., Mondor, T. A., Loken, M., Ingleton, M. A., \& Bergstrom, K. (1990). Locus of information in words and the right visual field effect. Brain Cognition, 14, 4458.

Brysbaert, M., \& D'Ydewalle, G. (1990). Tachistoscopic presentation of verbal stimuli for assessing cerebral dominance: Reliability data and some practical recommendations. Neuropsychologia, 28, 443-455.
Brysbaert, M., New, B., \& Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. Behav. Res. Methods, 44, 991-997.
Chan, A. B., \& Hsiao, J. H. (2016). Information distribution within musical segments. Music Perception, 34, 218-242.
Cohen, L., Henry, C., Dehaene, S., Martinaud, O., Lehéricy, S., Lemer, C., \& Ferrieux, S. (2004). The pathophysiology of letter-by-letter reading. Neuropsychologia, 42, 1768-1780.
Degé, F., \& Schwarzer, G. (2011). The effect of a music program on phonological awareness in preschoolers. Front. Psychol., 2, 7-13.
Delorme, A., \& Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Meth., 134, 9-21.
Gauthier, I., Curran, T., Curby, K. M., \& Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. Nat. Neurosci., 6, 428-432.
Hébert, S., \& Cuddy, L. L. (2006). Music-reading deficiencies and the brain. Adv. Cogn. Psychol., 2, 199-206.
Hsiao, J. H., \& Lam, S. M. (2013). The Modulation of Visual and Task Characteristics of a Writing System on Hemispheric Lateralization in Visual Word Recogni-tion-A Computational Exploration. Cognitive Sci., 37, 861-890.
Keuleers, E., \& Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. Behav. Res. Methods, 42, 627-633.
Lavidor, M., \& Ellis, A. W. (2001). Mixed-case effects in lateralized word recognition. Brain Cognition, 46, 192195.

Lavidor, M., \& Ellis, A. (2003). Orthographic and phonological priming in the two cerebral hemispheres. Laterality, 8, 201-223.
Lavidor, M., Ellis, A. W., \& Pansky, A. (2002). Case alter nation and length effects in lateralized word recognition:

Studies of English and Hebrew. Brain Cognition, 50, 257271.

LI T.K. \& Hsiao J.H. (2015). Music Reading Expertise Modulates Hemispheric Lateralization in English Word processing but not in Chinese Character Processing. CogSci 2015 proceedings, 1344-1349.
Lopez-Calderon, J., \& Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. Front. Hum. Neurosci., 8, 213.
Maurer, U., Brandeis, D., \& McCandliss, B. D. (2005). Fast, visual specialization for reading in English revealed by the topography of the N170 ERP response. Behav. Brain Funct., 1, 1.
McCandliss, B. D., Cohen, L., \& Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. Trends Cogn. Sci., 7, 293-299.
Müllensiefen, D., Gingras, B., Musil, J., \& Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. PloS one, 9, e89642.
Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia, 9, 97-113.
Proverbio, A. M., Manfredi, M., Zani, A., \& Adorni, R. (2013). Musical expertise affects neural bases of letter recognition. Neuropsychologia, 51, 538-549.
Rastle, K., Harrington, J., \& Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. Q. J. Exp. Psy-chol.- A, 55, 1339-1362.
Rossion, B., Kung, C. C., \& Tarr, M. J. (2004). Visual expertise with nonface objects leads to competition with the early perceptual processing of faces in the human occipitotemporal cortex. Proc. Natl. Acad. Sci. U.S.A., 101, 14521-14526.
Rumsey, J. M., Horwitz, B., Donohue, B. C., Nace, K., Maisog, J. M., \& Andreason, P. (1997). Phonological and orthographic components of word recognition. A PET-rCBF study. Brain, 120, 739-759.
Schön, D., Anton, J. L., Roth, M., \& Besson, M. (2002). An fMRI study of music sight-reading. Neuroreport, 13, 2285-2289.
Segalowitz, S. J., Bebout, L. J., \& Lederman, S. J. (1979). Lateralization for reading musical chords: Disentangling symbolic, analytic, and phonological aspects of reading. Brain Lang., 8, 315-323.
Sergent, J. (1982). The cerebral balance of power: Confrontation or cooperation? J. Exp. Psychol.: Human, 8, 253272.

Wong, Y. K., \& Hsiao, J. H. W. (2012). Reading direction is sufficient to account for the optimal viewing position in reading: The case of music reading. CogSci 2012 proceedings, 2540-2545.
Ziegler, J. C., Besson, M., Jacobs, A. M., Nazir, T. A., \& Carr, T. H. (1997). Word, pseudoword, and nonword processing: A multitask comparison using event-related brain potentials. J. Cognitive Neurosci., 9, 758-775.

# Metaphor Congruent Image Schemas Shape Evaluative Judgment: A Cross-Linguistic Study of Metaphors for Economic Change 

Patricia Lichtenstein (plichtenstein@ucmerced.edu)<br>Cognitive \& Information Sciences, University of California, Merced<br>5200 N. Lake Road; Merced, CA 95343 USA<br>Ekaterina Shutova (es407@cam.ac.uk)<br>Computer Laboratory, University of Cambridge<br>15 JJ Thomson Avenue; Cambridge CB3 0FD UK


#### Abstract

Metaphor pervades discussions of important socio-political topics. Recent research indicates that metaphorical language can influence how people reason about such topics, potentially affecting real-world decision-making. In this study, we report on research into the effects of metaphor on evaluative judgment, another aspect of decision-making that has been less well studied than reasoning. We use a cross-linguistic difference in the metaphors used by English and Spanish speakers to discuss economic change to investigate how metaphorical language affects evaluative judgment. We show that the image schematic information inherent in the semantics of the different metaphors performs a central role in shaping this process.


Keywords: metaphor; evaluative judgment; cross-linguistic variation; image schemas; socio-political discourse

## Introduction

I say, block those metaphors. America's economy isn't a stalled car, nor is it an invalid who will soon return to health if he gets a bit more rest. Our problems are longer-term than either metaphor implies. And bad metaphors make for bad policy. - Paul Krugman

We often use metaphorical language to express a stance toward an object or event. The metaphors invoked by Mr. Krugman operate by applying a basic image schema from the domain of spatial motion to the complex socio-political domain of the economy. The image schema is that of BLOCKAGE. Image schemas are dynamic representations of spatial relationships, force relationships, and motion in space (Langacker, 2001; Talmy, 1990). They develop through experience as encoded in our daily sensory-motor activities, and are thought to be the basis of the human conceptual system (Lakoff \& Johnson, 1999). The Blockage image schema entails an entity whose tendency towards motion is being impeded. It can be used in metaphorical expressions to signal evaluative judgment.

Evaluative judgment is the process of assessing a stimulus through the filter of an internal network of beliefs, values, and aesthetics. This process interacts with sensory-motor nervous system circuitry to influence how we interpret other people's behavior and whether we participate in select activities or events (Norman et al., 2011). It contributes to basic human impulses, such as avoidance and approach. It further serves to guide complex behavior. For example, evaluative judgment informs activities such as participating in a social movement,
contributing to a charity, and voting for a presidential candidate where the choice of candidate can be based more heavily on evaluative processes than on reasoning.

Evaluative processes pervade human language on many levels. People use linguistic information as an input into evaluative judgments at the social and interpersonal level: we assess our partners in communication positively or negatively based on accent, intonation, prosody, word choice, and grammatical complexity (Berger \& Calabrese, 1975; Fuertes et al., 2012). Evaluative information is, in turn, directly encoded in the lexical semantics of language. For instance, cross-linguistic research shows evaluative information to be the semantic component that factors most frequently and consistently into lexical expressions from a wide variety of language families (Triandis \& Osgood, 1958). The words in an utterance, then, potentially lead to entailments and evaluative inferences that are congruent with the stance of the speaker. Similarly, framing a topic to highlight and promote one's own evaluative judgment can influence how that topic is evaluated by others.

Framing can be achieved through metaphorical language, the semantics of which convey an evaluative component that can be as basic as avoidance versus approach: while the specter of urban violence must be strictly avoided, an epidemic of urban violence can be cautiously approached (Thibodeau \& Boroditsky, 2011). Metaphor theorists from ancient times to the present have considered evaluative judgment to be one of the primary functions of metaphorical language (Lakoff \& Johnson, 1980). This evaluative information is conveyed largely through the image schemas that are encoded in language and accessible to human cognition through our experiences of perceiving, interacting with, and emotionally responding to our environment (see for example Barsalou, 2010; Zwaan, 2008).

Interestingly, recent experimental research on metaphorical language provides evidence for the influence of image schematic information on evaluative judgment. In one study, participants were more likely to judge immigrants negatively after being primed to view their own country as a human body (e.g., a nation undergoing a growth spurt) rather than as an abstract entity (e.g., a nation undergoing a period of innovation), an effect that is congruent with a conventional metaphor that describes immigrants and immigration in terms of dis-
ease (Landau et al., 2009). Keefer et al., (2014) have characterized this effect as metaphoric fit. Metaphoric fit refers to maintaining consistent image schematic information across metaphors during discourse. In a test of the metaphoric fit hypothesis, they asked participants to read an article on depression and its symptoms, and then evaluate the effectiveness of a proposed antidepressant medication. Three versions of the article framed depression in terms of a spatial metaphor (DEpression is Down), or in terms of a visual metaphor (DEPRESSION IS DARK), or using non-metaphorical language. Participants in the spatial metaphor condition judged a drug called Liftix as more likely to be effective, whereas participants in the visual metaphor condition judged a drug called Illuminix more likely to be so. Recent results from another study (Thibodeau, 2016) support the metaphoric fit hypothesis, finding that people evaluated solutions to social problems more favorably when the solution and the problem were framed with congruent metaphoric schemas. For example, given a description of economic hardship in terms of a BALANCE schema, participants preferred a solution that involved "returning to equilibrium". When the description was given in terms of a Splitting schema, they preferred one that "narrowed the gap". Taken together, these findings present compelling evidence that the image schematic information specified by metaphorical language can impact how people evaluate situations. A question that naturally follows from such findings is whether cross-linguistic differences in metaphor usage produce similar effects. For example, if spatial metaphors for depression were predominant in one language, while visual metaphors for depression were in another, would speakers of the first language favor Liftix and speakers of the second Illuminix?

Although several studies have demonstrated that spatial metaphors for time vary across languages (e.g., Chan \& Bergen, 2005; Fuhrman \& Boroditsky, 2010; Nuñez \& Sweetser, 2006), and that these distinctions are reflected in differences in how speakers conceptualize and reason about time (e.g., Casasanto \& Boroditsky, 2008), less attention has been directed towards cross-linguistic contrasts in other conceptual domains. Nor has evaluative judgment, as opposed to reasoning, been the focus of previous cross-linguistic research on metaphor. Here, we discuss an investigation into cross-linguistic differences in metaphors for economic change and whether these differences are reflected in speakers' evaluative judgments about economic change. In this study, we focus on the domain of the economy for a few reasons. Behavior related to the economy and financial decisionmaking is largely motivated by evaluative judgment, and evaluative judgment is consequently a topic of great interest in behavioral economics. In addition, patterns of cross-linguistic differences in the metaphors used to discuss economic change and issues related to the economy have been identified in previous research (e.g., Boers \& Demecheleer, 1997). CharterisBlack \& Ennis (2001), for instance, compared metaphor use in news coverage of business and financial matters across

English and Spanish publications. They found many crosslinguistic similarities in metaphor use, including the tendency to characterize the state of the economy in terms of the physical or psychological health of an organism, to characterize change in the economy in terms of physical motion, and to characterize unusual economic events in terms of natural disasters. At a higher resolution of analysis, however, divergent patterns emerged. The salience of certain metaphors varied across the two languages, as did the types of subordinate concepts favored in metaphor use. Nautical metaphors, for example, were found to be common in English reports on the economy but rare in Spanish reports.

Recent developments in natural language processing approaches to metaphor research have allowed the analyses of much larger corpora to corroborate such hand-annotated corpus studies. Gutiérrez et al., (2017) present a method for detecting fine-grained, cross-linguistic textual differences through the automated analyses of large multilingual corpora. Employing data-driven techniques from natural language processing and machine learning, this method can identify crosslinguistic variation in the use of metaphor among multiple topics. Regarding the topic of the economy, the study found that English discussions of economic change most frequently employ metaphors based on locative motion verbs (e.g., "the economy was going backwards" to describe negative growth). In comparison, Spanish discussions typically rely on metaphors based on expanding/contracting motion verbs (e.g., "la economia se contrajo" to describe negative growth). The image schematic structure of locative motion verb metaphors as in the English example can be categorized as Source-Path-Goal with a focus on Path. Figure 1a illustrates how the English phrase instantiates this schema. The image schematic structure of volumetric motion verbs as in the Spanish example can be categorized as EQUILIBRIUM. Figure 1b illustrates how the Spanish phrase instantiates it. Notably, locative motion verbs feature movement out-of-place whereas expanding/contracting motion verbs feature movement in-place. The experiment in the present study was designed in accordance with these results. In the experi-


Figure 1: The Source-Path-Goal and Equilibrium image schemas prominent in English and Spanish metaphors for economic change.
ment, conducted across a sample of native English and native Spanish speakers, we tested whether people were more sensitive to information presented in a form congruent with the more salient metaphors for economic change in their language. For the experimental task, participants made an evaluative judgment ("Improved" or "Worsened") regarding economic change in a fictional country based on a representative graphic. Change was depicted in the graphic along the two dimensions characteristic of the Source-Path-Goal image schema (henceforth referred to as Motion) and the EQUILibrium image schema (henceforth referred to as Volume). That is, the figure in the graphic progressed from a starting point along a linear path to an end point at the same time that it expanded or contracted. If metaphor-congruent image schemas are active during the process of forming evaluative judgments, we expect English speakers' evaluations of the economy to align more closely with the direction of change in the linear dimension of the graph, and Spanish speakers' evaluations to align more closely with change in the volumetric dimension.

## Experiment

## Methods

Participants We recruited 60 participants from one English-speaking country (the US) and 60 participants from three Spanish-speaking countries (Chile, Mexico, and Spain) using the CrowdFlower crowdsourcing platform. ${ }^{1}$ All participants answered a demographic questionnaire, reporting gender, age, location, native language, level of education, color vision deficiencies, and use of touchscreen device during task. Results from nine US and three non-US participants were discarded for failure to meet the language requirement.

Materials \& Design Participants first read a brief description of the experimental task, which introduced them to a fictional country in which economists are devising a simple but effective graphic for representing change in the economy (see Figure 2 for the English version). Spanish speakers read a translation of the English text provided by a native speaker of Spanish who is also fluent in English. They then navigated to a new page to begin the task. Stimuli were presented in a 1200 -pixel by 700 -pixel frame. The center of the frame contained a sphere with a 64-pixel diameter. For each trial, participants clicked on a button to activate an animation of the sphere which involved (1) a positive displacement (in rightward pixels) of $10 \%$ or $20 \%$, or a negative displacement (in leftward pixels) of $10 \%$ or $20 \%$; and, (2) an expansion (in increased pixel diameter) of $10 \%$ or $20 \%$, or a contraction (in decreased pixel diameter) of $10 \%$ or $20 \%$. Participants saw each of the resulting conditions three times. The displacement and size conditions were drawn from a random permutation of conditions using a Fisher-Yates shuffle (Fisher \& Yates, 1963). Crucially, stimuli in half of the trials contained conflicting image schematic information with respect

[^388]
#### Abstract

Oxar is a country on the planet Xor. Oxarian economists are experimenting with new ways of modelling economic trends in their country. A group of economists are attempting to design a simple yet intuitive way of representing changes in the economy of the country. You are here to help them in their mission! During the experiment, you'll see a series of graphics representing change in Oxar's economy. Your task will be to decide, according to the dynamic graphic you have just seen, whether the economy has improved or worsened. Then, you will be asked to estimate how much the economy changed according to the graphic. Please try and give your judgments about the economy as accurately as possible.


Figure 2: Description of the experimental task in English.
to the displacement and volumetric metaphors for economic change (e.g., the sphere could both expand and move backwards).

Once the sphere had completed its movement across the screen, participants recorded a judgment of whether the economy improved or worsened by clicking on the appropriate radio button. They then indicated on a seven point scale how confident they were in their judgment and by how much they judged the economy to have changed.

## Results \& Discussion

Congruence \& Direction To analyze participant judgment of the direction of economic change, we first coded the responses as being Motion-congruent or Volumecongruent. For instance, if on a given trial the sphere was expanding while moving backwards and the response was "Economy Improved", the result was coded as Volumecongruent. If the response had been "Economy Worsened", the result would have been coded as Motion-congruent. Figure 3 presents the mean proportion of responses that fall into these categories for both languages, ${ }^{2}$ showing that Spanish speakers' judgments were more likely to align with the Volume dimension of the stimuli (i.e., the dimension encoded in the image schemas used most frequently in Spanish to discuss economic change metaphorically), and English speakers' judgments were more likely to align with the Motion dimension (i.e., the dimension encoded in image schemas used most frequently in English to discuss economic change metaphorically). To test whether people's judgments followed a language-specific, metaphor-congruent response more often than predicted by chance, we used a mixed effects logistic regression model to analyze the data using the lme4 package in the R statistical language. ${ }^{3}$ We compared two models: one that modeled the influence of language on

[^389]

Figure 3: Mean proportion of dimension-congruent responses by language for Both-congruent, Motion-congruent and Volume-congruent judgments.
judgment for the conflicting trials ( $n=2520$ ), and one that modeled the influence of language on judgment for the nonconflicting trials $(n=2520)$. We reasoned that, if the image schematic information in the experimental stimuli is influencing judgment in a manner that is both language-specific and metaphor-congruent, language would be a significant predictor of judgment in the conflicting trials but not in the nonconflicting trials.

Both models included random effects for participant and item to control for their associated intraclass correlation, and both used the Laplace approximation for parameter estimation. The results of the analyses show an effect of language in the conflicting trials $(\beta=3.96, p<0.001)$, but not in the nonconflicting trials ( $\beta=0.23, p=0.57$ ). To evaluate model fit, we randomized the subset of our data containing the results of the conflicting trials and split it into testing and training sets. We then modeled the data in the training set using the same mixed effects logistic regression equation from our original analysis, and used this model to predict participant judgment in the testing set. The prediction accuracy of our model was 0.84. The AUROC (area under ROC curve) for predicting judgment with the model was 0.93 (see Figure 4), indicative of the expected proportion of true positives ranked before a uniformly drawn random negative. The above analyses indicate that participants' judgments of whether the economy improved or worsened varied in a predictable manner based on their native language. The estimated odds that the judgment of a Spanish speaker aligned with changes in Volume image schematic information in the stimuli rather than MoTION image schematic information were 52.46 greater than the corresponding odds for an English speaker.

Congruence \& Magnitude We also examined whether image schematic information influenced participants' judgments of the magnitude of economic change. Here, we focused on the nonconflicting trials. Recall that for these trials when-


Figure 4: The ROC curve showing the ability of our model to correctly classify the judgments of participants.
ever the sphere was moving forward, it was expanding; and, whenever it was moving backwards, it was contracting. However, in half of the trials ( $n=1260$ ), the sphere was either expanding or contracting to a greater degree than it was moving forward or backwards, or vice versa. If metaphor-congruent image schemas are affecting speakers' judgments, then we would expect Spanish speakers to increase their estimates of the magnitude of change when expanding or contracting of the sphere is the dominant schema. Similarly, we would expect English speakers' judgments to be greater when forward or backward motion is the dominant schema.

On average, English and Spanish speakers estimated the amount of economic change to be greater when the sphere was expanding/contracting than when it was moving forward/backwards, as shown in Figure 5. To assess whether the differences across language and dominant schema were significant, we analyzed the data using cumulative link (probit) models via the ordinal package in R. ${ }^{4}$ We first modeled amount of change with additive predictors of language and dominant schema, with participant and item as random effects. To test for an interaction between these factors, we fit a second model with the interaction as a predictor and used the likelihood ratio test to compare the two models. Our analysis indicated no main effects for language or for the interaction between language and dominant schema. There were marginal effects for Volume as the dominant schema both when contracting $(\beta=0.3122, p=0.009)$ and when expanding ( $\beta=0.3150, p=0.009$ ). These results tell us that a $10 \%$ change in contraction of the sphere increases the probability of higher estimates of the amount of change by $31.22 \%$ for Spanish and English speakers. A $10 \%$ change in expansion of the sphere increases the probability of higher estimates of the amount of change by $31.50 \%$, again for both groups of speakers.

[^390]

Figure 5: Mean estimated amount of change per language and stimuli dimension.

Congruence \& Confidence With the last metric we collected from participants, that of confidence, we investigated whether one group of speakers found a particular set of stimuli (or the experimental task itself) to be more difficult or confusing than was the case for the other group of speakers. English speakers, for example, may have had more difficulty in assessing the stimuli with contradictory information since English employs Motion and Volume based metaphors in discussions of economic change.

We again analyzed the data using cumulative link (probit) models, here with additive predictors of language and stimuli class, with participant and item as random effects. To test for an interaction between the factors, we fit a second model with the interaction as a predictor and used the likelihood ratio test to compare the two models. We found no significant differences across languages and no significant interaction between language and stimulus condition. English speaking participants and Spanish speaking participants both showed a highly significant increase in confidence for judgments in three of the stimulus conditions: when the sphere was contracting by $20 \%$ and moving backward by $20 \%(\beta=0.4907, p<0.001)$, when the sphere was expanding by $20 \%$ and moving forward by $20 \% ~(\beta=0.6553, p<0.001)$, and when the sphere was expanding by $20 \%$ and moving forward by $10 \% ~(~ \beta=0.4815, p<0.001)$.

Both groups of participants, then, showed greater confidence in their judgments, with a $49.07 \%$ increased probability, given a stimulus congruent with a large decline in the economy across both image schemas. Given a stimulus suggesting a large improvement in the economy across image schemas, participant confidence increased with a probability of $65.53 \%$. Given a stimulus congruent with a large improvement in the economy per the Volume dimension of the stimuli and a moderate improvement per the MOTION dimension, confidence improved with an increased probability of $48.15 \%$. These results suggest that, for all participants,
confidence in judgment relied on magnitude. In the first two conditions where confidence increased, the sphere was increasing or decreasing to a maximum degree along both dimensions. Hence, it is unclear whether participants were more sensitive to language-specific, metaphor-congruent image schematic information. That participant confidence increased in the third condition noted above suggests that participants overall were more confident when the sphere was moving forward rather than backwards. Since the results were consistent across languages, we found no evidence that any particular combination of schemas were more difficult to parse for one, but not the other, group of speakers.

## General Discussion

Metaphors based on movement are regularly used in English and Spanish (amongst other languages) to discuss the economy and economic change. Past research suggests that in discussions of important socio-political topics, such as the economy, metaphorical language can influence people's reasoning. Here, we presented evidence that metaphorical language also affects people's evaluative judgment. Evaluative judgment, which differs from reasoning in that it is not logicbased, is a crucial component of decision-making. The results of our study thus serve to tease apart how these two mechanistically different processes influence decision-making, with our results indicating that metaphorical language affects the evaluative judgment component of decision-making through the image schematic information present in the semantics of the metaphor. These findings corroborate accounts of language and cognition that emphasize the role of language in the development of associative and representational routines. In line with such accounts, the current findings reflect Spanish speakers learning to associate economic change with shifts in Volume and habituating focus to volumetric image schematic information. English speakers, who learn to associate economic change with shifts in Motion, habituate focus to image schematic information involving displacement.

Our findings with respect to congruence and amount of change, and with respect to congruence and confidence, provide further nuance to this view. While judgments of economic improvement or decline aligned more consistently with language-specific, metaphor-congruent dimensions in the stimuli, judgments regarding the amount of change were more consistent with changes in Volume for both groups of speakers. A possible explanation for this effect is that visual experience of a change in amount is more strongly correlated with a change in density than with a change in location. Another possible explanation is that English speakers associate direction of change with motion through space but associate amount of change with manner of motion, and changes in the size of the sphere were more suggestive of manner of motion. As noted above, judgments of confidence were consistent with changes across both schemas for both groups of speakers, as long as the changes were consistent across schemas. This indicates that, while participant focus
may have been driven by metaphor-congruent schemas, they were still sensitive to both aspects of the stimuli. This latter finding supports the view that language probabilistically influences other cognitive functions as opposed to those that consider language to have a more discrete effect on other aspects of cognition.

Evaluative judgment is a core component of decisionmaking, especially in the socio-political domain, and it is highly reliant on the image schematic information shared across percepts and concepts. Such image schematic information is also key to the cognitive function of metaphor. The present work represents a preliminary step in linking evaluative judgment and its reliance on image schematic information to cross-linguistic variation in metaphor use. Future studies along these lines should shed further light on how the image schematic information intrinsic to metaphorical language shapes evaluative judgment in discussions of the economy, as well as discussions of other important socio-political domains.

## Acknowledgements

We thank the CrowdFlower workers who participated in our experiment and the CogSci reviewers for insightful feedback on our study. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship.

## References

Barsalou, L. (2010). Grounded cognition: Past, present, and future. Topics in Cognitive Science, 2, 716-724.
Berger, C., \& Calabrese, R. (1975). Some exploration in initial interactions and beyond: toward a developmental theory of interpersonal communication. Human Communication Research, 1, 99-112.
Boers, F., \& Demecheleer, M. (1997). A few metaphorical models in (western) economic discourse. In W. Liebert, G. Redeker, \& L. Waugh (Eds.), Discourse and perspective in cognitive linguistics. Amsterdam/Philadelphia: Benjamins.
Casasanto, D., \& Boroditsky, L. (2008). Time in the mind: Using space to think about time. Cognition, 106, 579-593.
Chan, T., \& B.Bergen. (2005). Writing direction influences spatial cognition. In Proceedings of the twenty-seventh annual conference of the cognitive science society (pp. 510516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Charteris-Black, J., \& Ennis, T. (2001). A comparative study of metaphor in Spanish and English financial reporting. English for Specific Purposes, 20, 249-266.
Fisher, R., \& Yates, F. (1963). Statistical tables for biological, agricultural, and medical research. Edinburgh: Oliver and Boyd.
Fuertes, J., Gottdiener, W., Martin, H., Gilbert, T., \& Giles, H. (2012). A meta-analysis of the effects of speakers accents on interpersonal evaluations. European Journal of Social Psychology, 42, 120-133.

Fuhrman, O., \& Boroditsky, L. (2010). Cross-cultural differences in mental representations of time: Evidence from an implicit non-linguistic task. Cognitive Science, 34, 401450.

Gutiérrez, E., Shutova, E., Lichtenstein, P., de Melo, G., \& Gilardi, L. (2016). Detecting cross-cultural differences using a multilingual topic model. Transactions of the Association for Computational Linguistics, 71-81.
Keefer, L. A., Landau, M. J., Sullivan, D., \& Rothschild, Z. K. (2014). Embodied metaphor and abstract problem solving: Testing a metaphoric fit hypothesis in the health domain. Journal of Experimental Social Psychology, 55, 12-20.
Lakoff, G., \& Johnson, M. (1980). Metaphors we live by. Chicago, IL: University of Chicago Press.
Lakoff, G., \& Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to western thought. Basic books.
Landau, M. J., Sullivan, D., \& Greenberg, J. (2009). Evidence that self-relevant motives and metaphoric framing interact to influence political and social attitudes. Psychological Science, 20(11), 1421-1427.
Langacker, R. (2001). Concept, image, and symbol. Berlin, Germany: De Gruyter Mouton.
Norman, G., Norris, C., Gollan, J., Ito, T., Hawkley, L., Larsen, J., ... Berntson, G. (2011). Current emotion research in psychophysiology: he neurobiology of evaluative bivalence. Emotion Review, 3, 349-359.
Nunez, R., \& Sweetser, E. (2006). With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. Cognitive Science, 30, 401-450.
Talmy, L. (2000). Toward a cognitive semantics. Boston, MA: MIT Press.
Thibodeau, P. H. (2016). Extended metaphors are the home runs of persuasion: Don't fumble the phrase. Metaphor and Symbol, 31(2), 53-72.
Thibodeau, P. H., \& Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. PloS one, 6(2), e16782.
Traugett, E. (1978). On the expression of spatiotemporal relations in language. In J. Greenberg (Ed.), Universals of human language. vol. 3: word structure. Stanford, CA: Stanford University Press.
Triandis, H., \& Osgood, C. (1958). A comparative factorial analysis of semantic structures in monolingual Greeks and American college students. Journal of Abnormal and Social Psychology, 57, 186-196.
Zwaan, R. (2008). Experiential traces and mental simulations in language comprehension. In M. de Vega, A. Glenberg, \& A. Graesser (Eds.), Symbols and embodiment: Debates on meaning and cognition. Oxford, UK: Oxford University Press.

# Keystroke Dynamics Predict Essay Quality 

# Aaron D. Likens (Aaron.Likens@asu.edu) <br> Laura K. Allen (LauraKAllen@asu.edu) Danielle S. McNamara (danielle.menamara@asu.edu) 

Arizona State University, PO BOX 872111<br>Tempe, AZ 85281 USA


#### Abstract

Language entails many nested time scales, ranging from the relatively slow scale of cultural evolution to the rapid scale of individual cognition. The nested, multiscale nature of language implies that even simple acts of text production, such as typing a sentence, entail complex interactions involving multiple concurrent processes. As such, text production may have much in common with other cognitive phenomena thought to emerge from multiplicative interactions across temporal scales, namely those that exhibit fractal properties. We investigated the relationship between fractal scaling and the quality of produced text. Participants ( $N=131$ ) wrote essays while their keystrokes were recorded. Fractal analyses were then performed on time series of interkeystroke intervals (IKIs). Results showed that fractal properties characterizing IKIs positively predicted expert ratings of essay quality, even after accounting for essay length. The results support our hypotheses concerning multiscale coordination and text production.


Keywords: text production; writing; keystroke; multifractal; essay quality

## Introduction

Recent theoretical and empirical work characterizes language as a complex, dynamic system that involves the coordination of multiple nested time scales (Dale, Kello, \& Schoenemann, 2016; Rączaszek-Leonardi, 2010; Rączaszek-Leonardi \& Kelso, 2008). Consider three time scales that have been highlighted extensively in the literature. Language can evolve on a relatively slow scale along with changes in cultures and significant historical events. On a faster scale, language can be altered throughout an individual's life, based on their experiences and knowledge. Lastly, language can change at more rapid scales in response to cognitive events that can span days, hours, or mere fractions of a second.

These time scales span several orders of magnitude and paint a complex picture of language. The picture is further complicated because each of the time scales implicates different systems (e.g., cultural, interpersonal, physiological) and suggests that language processes should be studied as complex, dynamical systems. The current work explores this idea in the context of text-based language production. We examine whether dynamic analyses of typing behaviors during essay writing provide empirical support for the notion of writing as a complex, dynamical system.

The nested, multiscale character of text production is apparent in the simple example of typing an essay. The
relatively fast time scale of word selection is nested within and constrained by the slower time scale of idea generation. Singular ideas are further nested within the subtopics and global topic of the essay that change at even slower rates. Beyond these examples, nesting can continue at both faster and slower time scales. Rapidly changing physiological processes influence and support the act of writing that would not be possible without a lifetime of learning or the evolution of a language within a culture. Thus, even the seemingly simple act of typing one sentence of an essay may entail complex interactions of any number of processes, each with its own characteristic rate of evolution. The implication is that language production involves the coordination of numerous systems over many different time scales. Our assumption is that the act of text production (i.e., typing an essay) provides a window into ongoing cognitive processes (Pinet, Ziegler, \& Alario, 2016). As such, we expect that keystroke dynamics will reveal the multiplynested character of text production.

## Multiscale Interactions in Human Behavior

A wide range of cognitive phenomena have been described as emerging from the interaction of multiply-nested time scales (Ihlen \& Vereijken, 2010). The principle evidence for that claim is the observation of fractal scaling. Fractal scaling typically refers to two qualities: long-range autocorrelation and scale dependence. Long-range autocorrelation implies that time series observations exhibit significant correlations over large timespans (Beran, 1994). That is, an observation made at one point in time is related to subsequent observations that extend into the future. Scale dependence suggests that measurements of time series (e.g., variance) depend on the temporal scale at which they are measured (Mandelbrot \& Van Ness, 1968).

There are numerous examples of behavioral time series known to exhibit fractal scaling: reaction times (Gilden, Thornton, \& Mallon, 1995; Van Orden et al., 2003), time estimation (Wagenmakers et al., 2004), eye movements (Stephen \& Anastas, 2011), hand movements (Anastas, Stephen, \& Dixon, 2011; Stephen, Arzamarksi, \& Michaels, 2010), arm movements (Chen, Ding, \& Kelso, 1997), postural corrections (Collins \& DeLuca, 1993), and various forms of tool-use (Likens, Fine, Amazeen, \& Amazeen, 2015; Nonaka \& Bril, 2014).

Much of the work on fractal scaling in cognition has emphasized interaction across scales as its primary theoretical contribution (Ihlen \& Vereijken, 2010; Kelty-

Stephen \& Wallot, in press). The basic idea is that the rich structure observed in behavioral time series is the product of many simultaneously occurring processes (e.g., physical, cognitive). Each process exists on its own time scale, with effects of slower time scales multiplicatively cascading to faster and faster time scales. As such, fractal scaling reflects on-the-fly cognitive organization during tasks (Van Orden et al., 2003; Wallot, Hollis, \& van Rooij, 2013). Further, variability in fractal properties reflects the flexibility and adaptability in typical cognitive tasks necessary for coordination across those levels (Anastas et al., 2011).

If fractal scaling reflects the flexibility and adaptability that stems from multiscale coordination, then reliable relationships should exist between fractal scaling and other meaningful aspects of behavior. The literature contains several such examples. Visual search is faster when eye movements exhibit fractal properties (Stephen \& Anastas, 2011). Fractal variability in hand movements predicts better perceptual estimates (Stephen et al., 2010). Moreover, fractal patterns distinguish between various forms of skilled and non-skilled behavior (e.g., Nonaka \& Bril, 2014). These examples are not exhaustive but hint at the large number of skillful behaviors that exhibit fractal characteristics.

The current work explores the idea that fractal scaling might also reflect the multiscale coordination involved in the skilled production of text. Across domains, the evidence implies being skilled means adapting to task demands, and fractal scaling characterizes flexibility (Gorman et al., 2010; Nonaka \& Bril, 2014). The observation of flexibility in skilled text production (Allen, Snow, \& McNamara, 2016) leads to the hypothesis that fractal scaling will reveal the flexibility required from the nested, multiscale act of composition. That is, we expect more skilled text production will be characterized by fractal variability.

We are not aware of any studies that have examined the time course of text production for evidence of fractal scaling; however, work related to reading and skilled typing provides some bases for exploration (e.g., Wallot \& Grabowski, 2013; Wallot et al., 2013; Wijnants et al., 2012). For example, Wijnants and colleagues (2012) showed that the presence of fractal scaling in word naming times distinguished dyslexic and non-dyslexic readers. They also found a positive relationship between fractal scaling and reading fluency. Another study involving skilled typing suggests that fractal properties may depend on task complexity/difficulty (Wallot \& Grabowski, 2013). The relevant finding in that study was that there was greater fractal variability over time when participants typed a set of directions than when they typed simple lyrics from memory or simply copied text.

## Current Study

This study investigates how fractal properties in keystroke logs are related to the quality of written text. Participants wrote timed, prompt-based argumentative essays while their keystrokes were recorded. Time series were constructed from the latencies between keystrokes and analyzed by
fractal analysis. Essays were scored by experts on holistic quality and analytical subscales. This study is exploratory and the first of its kind; nonetheless, our general expectation is that, like performance on other tasks, fractal properties will serve as reliable predictors of essay quality.

## Method

Participants Undergraduate students $(N=131$, Female $=$ 58 , mean age $=19.8$ years) were recruited from a large university in the United States. Students participated in the study in exchange for course credit.
Procedure Participants wrote a timed (25-minutes), promptbased, argumentative essay. Essay prompts were similar in structure to Scholastic Aptitude Test (SAT) prompts in that participants were asked to take either a supporting or contrary position on a given topic. Keystrokes and their respective time stamps were recorded while students composed their essays. Unsurprisingly, participants varied considerably in the number of keystrokes they produced ( $M$ $=3,385.40, S D=1,107.03$ ). To prevent bias, only the first 999 keystrokes were retained for further analysis, corresponding to lowest number of keystrokes in our sample. No other keystrokes (e.g., backspaces) were omitted. Keystroke timestamp series were then differenced to obtain time series of interkeystroke intervals (IKIs). Mouse movements were not recorded.

Text Analyses Pairs of raters evaluated the essays based on holistic quality and analytic subscales. Raters received extensive training before scoring and received compensation for their time. Holistic scores ranged from one (minimum) to six (maximum) and were based on a standardized rubric used in the assessment of SAT essays. Interrater reliability was good $(r=0.75)$. Raters were instructed to treat the distance between points (e.g., 1-2, 3-4, 4-5) as equal. The nine subscales, also based on a 6-point scales, were:

Introduction. $(M=3.97, S D=0.96)$ Demonstrates mastery in meeting the goals of an introduction (e.g., presenting a topic, providing a purpose, clearly stating a thesis, previewing arguments).

Body. ( $M=4.08, S D=0.90$ ) Demonstrates mastery in meeting the goals of body arguments (e.g., transition between arguments, using topic sentences, supporting arguments with evidence, and maintaining a flow throughout the arguments).

Conclusion. $(M=3.19, S D=1.32)$ Demonstrates mastery in meeting the goals of a conclusion (e.g., summarizing the essay, re-establishing the significance of discussion, capturing the reader's attention, and effectively closing the essay).

Organization. $(M=3.86, S D=0.98)$ Follows a logical structure, beginning with the introduction, through the arguments and evidence presented in the body arguments, and to the conclusion.

On-Topic/Global Cohesion. $(M=4.13, S D=0.85)$ Details presented throughout the essay support the thesis
and do not stray from the prompt and the main ideas and organizing principles presented in the introduction.

Grammar, Syntax, \& Mechanics. $(M=3.70, S D=0.79)$ Employs correct Standard American English, avoiding errors in grammar, syntax, and mechanics; the essay conveys strong control of the standard conventions of writing.

Voice. $(M=4.09, S D=0.76)$ The writer is expressive, engaging, and sincere, with a strong sense of audience.

Word Choice. $(M=4.07, S D=0.71)$ Word choice is precise and effective.

Sentence Structure. $(M=4.06, S D=0.75)$ Sentence patterns are varied effectively, enhancing the quality of the essay.
Fractal Analysis Fractal analysis comes in two forms, monofractal analysis and multifractal analysis, both of which were performed on the IKI time series. The goal of monofractal analysis is to understand how variability depends on scale (e.g., Eke, Herman, Kocsis, \& Kozak, 2002). In general, evaluating monofractality means estimation of scaling exponents from the relationship, $F^{2}(s)$ $\sim s^{H}$ where $H$ is the Hurst exponent, and $F^{2}(s)$ is a measure of fluctuation. Being a singular measure, the Hurst exponent provides a measure of typical scaling behavior in a time series. Moreover, $H$ ranges from zero to one and has useful interpretive ranges (Collins \& DeLuca, 1993; Gorman et al., 2010). When $H=0.5$, the time series exhibits random variation. When $\mathrm{H}>0.5$, the series contains long-range autocorrelation, and when $\mathrm{H}<0.5$, the series exhibits longrange anticorrelation such that small values generally follow large values and vice versa. Many series have been shown to require not one but a spectrum of exponents to characterize their variability (Ihlen \& Vereijken, 2010; Kantelhardt et al., 2002). Hence, the goal of multifractal analysis is to determine whether fractal scaling is fixed across time; that is, whether a time series exhibits multifractality (Kantelhardt et al., 2002).

We used Multifractal Detrended Fluctuation Analysis (MFDFA; Kantelhardt et al., 2002) to evaluate both monofractal and multifractal properties in IKIs. The outcome of MFDFA is the multifractal spectrum. MFDFA is the generalization of Detrended Fluctuation analysis (DFA) and has been used in diverse literature to characterize time-varying structure (Kantelhardt et al., 2002; Peng et al., 1994). The MFDFA procedure consists of five steps. The first step is to create the profile by integrating over a meancentered time series. In a second step, the time series of length, $N$, is divided into $N_{s}=\operatorname{int}(N / s)$ non-overlapping bins, such that each bin contains $s$ observations. To compensate for $N_{s}$ often being a non-integer multiple of $s$, the binning procedure is performed twice by starting from each end of the time series. The result partitions the time series into $2 N_{s}$ bins. In a third step, data in each bin is fit with a least squares regression line that is subtracted from the binned data to obtain local residuals. The bin-wise residuals are squared and averaged to obtain a measure of variance within
each segment, $v$. The fourth step averages over all the bins to obtain the $q$ th order fluctuation function as captured in

$$
\begin{equation*}
F_{q}(s)=\left\{\frac{1}{2 N_{s}} \sum_{v=1}^{2 N_{s}}\left[F^{2}(v, s)\right]^{q / 2}\right\}^{1 / q} \tag{1}
\end{equation*}
$$

where $F^{2}(v, s)$ is the variance calculated in Step 3 and $q$ takes on both positive and negative values. Steps 2 through 4 are repeated for several $s$, increasing $s$ by a power. The current work used a fractional power $(11 / 10)$ for varying $s$ which allowed for a larger range of scales over which scaling estimates were made. The maximum $s$ was $\leq N / 4$. Step 5 evaluates scaling behavior by performing a log-log regression of $F_{q}(s)$ on $s$ for each value of $q$. We used 101 values of $q$, ranging from -3 to 3 . When scaling properties are present, the result from Step 5 is a linear slope equal to the $q$-order Hurst exponent, $H(q)$. When $q=2$, the procedure is equivalent to standard DFA. $H(q)$ can then be used to estimate the width of the multifractal spectrum $d h(q)$. In contrast $H, d h(q)$ provides a measure of the variability in scaling over time.

## Results

Hierarchical multiple regression was used to explore the relations between the fractal properties in IKIs (i.e., $H$, $d h(q))$ and holistic essay scores $(M=3.85, S D=0.89)^{1}$. Table 1 presents the descriptive statistics for the predictor variables used in constructing regression models.

Table 1. Descriptive statistics

| Variable | $M$ | $S D$ |
| :--- | ---: | ---: |
| Number of Words (NW) | 412.67 | 162.22 |
| $d h(q)$ | 1.32 | 0.26 |
| $H$ | 0.51 | 0.06 |

In addition to fractal properties, we included the total number of words (NW) in each essay as a predictor in the regression model because of the known positive relationship between essay length and essay quality (e.g., McNamara, Crossley, \& Roscoe, 2013; McNamara, Crossley, Roscoe, Allen, \& Dai, 2015). Predictors were checked for multicollinearity and all variance inflation factors were less than $2\left(\mathrm{VIF}_{\mathrm{NW}}=1.09 ; \mathrm{VIF}_{H}=1.21 ; \mathrm{VIF}_{d h(q)}=1.26\right)$, indicating that multicollinearity was not a concern. Note that, NW, $H$, and $d h(q)$ were converted to $z$-scores to aid in interpretation. This was especially crucial in the case of $H$ as its theoretical domain is $(0,1)$. NW was entered in the first model step; $H$ and $d h(q)$ were both entered in the second model step. As expected, the initial model was significant, $\beta$ $=0.47, R^{2}=0.28, p<0.001$; however our interest was in characterizing whether fractal properties predicted essay quality over and above NW. The results showed that fractal properties improved model fit, $F(2,127)=6.68, p<0.01, R^{2}$ $=0.35$. As expected NW was a significant predictor such that a one standard deviation increase in essay length predicted a 0.54 increase in holistic score, $t(127)=8.09, p<$

[^391]0.001. After controlling for NW, the model also revealed that a one standard deviation increase in $H$ predicted a 0.21 increase in holistic score, $t(127)=2.92, p<0.01$. Furthermore, a one standard deviation increase in $d h(q)$ predicted a 0.29 increase in holistic score, $t(127)=3.19, p<$ 0.01 .

Following the analysis of holistic essay scores, nine additional sets of regression models were fit predicting each subscale from NW, $H$, and $d h(q)$. The modeling strategy for these additional models was the same as for overall essay quality. A summary of those models appears in Table 2. The table shows that fractal properties explain significant variance for seven out of nine subscales, with Conclusion and Organization being the exceptions. Of note is the fact that, for several outcomes, the fractal properties explain more than twice the variance explained by NW.

Table 2. Regression models for expert rated subscales Predictors

|  | Predictors |  |  |  |  |  | $F$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Model 1 |  | Model 2 |  |  |  |  |
| DV | NW | $R^{2}$ | NW | H | $d h(q)$ | $R^{2}$ |  |
| Intro | 0.30*** | 0.10 | 0.39*** | 0.27** | 0.29*** | 0.20 | 7.85*** |
| Body | 0.43*** | 0.22 | 0.38*** | 0.18* | 0.15 | 0.27 | 3.52* |
| Conc. | 0.63*** | 0.23 | 0.64*** | 0.24* | 0.13 | 0.26 | 2.46 |
| Org. | 0.39*** | 0.15 | 0.44*** | 0.16 | 0.16 | 0.19 | 2.45 |
| Coh. | 0.17* | 0.04 | 0.23** | 0.19* | 0.17* | 0.10 | 3.71* |
| Gram. | 0.15* | 0.04 | 0.24*** | 0.20** | 0.33*** | 0.18 | 11.30*** |
| Voice | 0.30*** | 0.15 | 0.34*** | 0.17* | 0.20** | 0.22 | 5.78** |
| Word | 0.22*** | 0.10 | 0.31*** | 0.22*** | 0.31*** | 0.27 | 15.06 *** |
| Sent. | 0.30*** | 0.16 | 0.37*** | 0.19** | 0.23*** | 0.25 | 7.79*** |

Note: ${ }^{* * *} \mathrm{p}<0.001,{ }^{* *} \mathrm{p}<0.01, * \mathrm{p}<0.05$. F test was based on 2 and 127 degrees of freedom.

## Discussion

In this study, we investigated how the multiscale characteristics of text production relate to essay quality. In general, we found that the Hurst exponent was a positive predictor of holistic essay quality and analytical scores. Similarly, we found that broader multifractal spectra predicted better quality essays, overall, and on several analytical subscales. The remainder of the discussion is structured as follows: First, we give an overview and basic interpretation of the scaling behavior observed for IKIs during essay production. Second, we speculate on how those interpretations inform patterns of prediction observed with respect to essay quality. Lastly, we offer ideas for potential applications and future research.

## Scaling Properties in IKIs

We found that IKIs in this study were characterized by global $H$ close to the value typical of random variation (i.e., $H=0.50$ ). The observed mean of $H$ was surprising given the prevalence with which $H$ s indicative of long-range correlation (i.e., $H>0.5$ ) have been observed in other tasks (Kello et al., 2010). The result was further surprising
because keystrokes observed during typing tasks have been previously characterized as being anti-correlated, where $H<$ 0.5 (Wallot \& Grabowski, 2013). One possible reason for the difference in results is those authors' use of power spectral density to estimate $H$ (labeled $\alpha$ in that study). Simulation work has shown power spectral density underestimates $H$ (Delignières et al., 2006). The reason for those differences may be the method used to estimate $H$.

A more likely and substantive reason relates to nature of the tasks used in each study. In Wallot and Grabowski (2013), participants performed one of three relatively simple tasks: type a nursery rhyme from memory; copy text from a page; and generate a novel set of directions from school to home. The latter condition was the closest to essay writing but still differs substantially in complexity and difficulty, factors known affect the scaling properties in basic motor control tasks and complex tasks like steering (e.g., Chen et al., 2001; Likens et al., 2015). Writing a timed essay is arguably more difficult and more complex than giving familiar directions. Perhaps, then, the Hurst exponents we observed in this study reflect those differences in task difficulty. The results concerning $d h(q)$ lends further, albeit tentative, support for that conclusion.

The trend across tasks observed in Wallot and Grabowski (2013) permits cautious speculation about the meaning of spectral widths within the current context. Note that a direct comparison between the widths we observed and those in Wallot and Grabowski is not possible because they used a wavelet form of multifractal analysis, and different methods are known to produce different widths (Ihlen \& Vereijken, 2013). In Wallot and Grabowski, the multifractal spectrum width increased as a function of task complexity, with the generative task producing the broadest spectrum. Similar results have also been reported in the motor control and social coordination literatures where an increase in task difficulty has been associated with widening $d h(q)$ (e.g., Davis, Brooks, \& Dixon, 2016; Romero, Coey, Beach, \& Richardson, 2013). A reasonable conclusion is that the relatively broad spectra we observed reflect the difficulty inherent in writing a timed essay.

Unlike $H$, the multifractal spectrum does not have the same useful interpretive indices concerning long range correlation and randomness. However, a few words are possible concerning why task complexity or task difficulty would affect the width of the multifractal spectrum. The multifractal spectrum provides a summary of scaling behaviors that evolve over time (Ihlen \& Vereijken, 2013; Kantelhardt et al., 2002). If the Hurst exponent were sufficient to describe the scaling behavior present in the IKI time series, then one would expect a narrow spectrum - a time-invariant monofractal process. Instead, we observed broad multifractal spectra that are more consistent with interpretation of a time-varying multifractal process. Timevarying scaling behavior is thought to reflect the ongoing dynamics in complex, dynamical systems that range from individual physiological processes to entire human teams (Likens et al., 2014). Time-varying scaling behavior in the

IKIs might reflect changes in cognitive state or changes in strategy that accompany the multiscale coordination involved in writing an essay (e.g., Stephen et al., 2009). That idea is elaborated in the following section in the context of essay quality.

## Scaling Properties as Predictors of Essay Quality

We have suggested that changes in scaling behavior may reflect changes in cognitive state or strategy. If so, then a broader multifractal spectrum could reflect flexibility in writing. Multifractal scaling is synonymous with flexibility and adaptability in other contexts (e.g., Collins \& De Luca, 1993), and flexibility in the use of cohesive devices (i.e., flexibility in writing) predicts higher quality essays (e.g., Allen et al., 2016; Snow et al., 2015). The implication is: if multifractal scaling reflects flexibility in writing, then wider multifractal spectra may also indicate higher quality essays. The current findings seem to support such reasoning. Results from regression analyses suggest fractal properties positively predict overall essay quality as well as quality on analytical subscales.

Another notable feature of the regression analyses was that $d h(q)$ did not predict the quality of either the Body or Conclusion. As a potential explanation of those results, we refer to our data preparation steps. The time series in our sample were truncated to accommodate participants with short essays. Given the average length of intact series was over three times the length of the truncated series we analyzed, there is a strong possibility the fractal analyses did not equally represent Body and Conclusion aspects of text. If true, then perhaps $d h(q)$ did not adequately capture variability with respect to Body and Conclusion sections. In addition, neither $H$ nor $d h(q)$ predicted the organization subscale. A similar interpretation could be made concerning the length of the time series analyzed with respect to the length of a typical essay in our sample.

## Applications and Future Directions

In this study, we have shown for the first time that fractal properties measured while writing an essay predict essay quality. Being the first of its kind, we have interpreted the results cautiously. However, the results are promising and suggest opportunities for future research and applications.

One promising area of research pertains to flexibility and adaptability in writing. As already discussed, multifractal scaling may suggest flexibility and adaptability in writing. If so, then it should be possible to link multifractal characteristics with other aspects of writing flexibility (Allen et al., 2016; Snow et al., 2015). In those studies, flexibility was characterized over several essays; however, if flexibility is important on the timescales of days and weeks, then flexibility should also be important within the context of a single essay. If so, then the fractal properties of keystrokes may also relate to the flexibility at those slower time scales.

Another related area of investigation involves the use of fractal properties in applied settings. The results of the
current study, if replicable, could inform applied educational settings such as those involving learning analytics and automated writing evaluation systems. The analyses we have presented here are algorithmically efficient enough to be implemented in real time. Real-time assessment of fractal properties is promising on several fronts. Real-time fractal properties could be monitored by instructors for early signs of writing difficulty and provide faster, targeted feedback. The same notion could apply within automated writing evaluation systems to augment automated feedback systems.

Lastly, the methods we have presented are not limited to the analysis of keystrokes. The use of physiological measurements and various movement registration devices is becoming more common in applied literature on intelligent tutoring systems (D’Mello, Picard, \& Graesser, 2007). Fractal analyses have proven beneficial in other settings involving physiological data, primarily because of relationship between fractality and flexibility (e.g., Chen et al., 2001; Ivanov et al., 2001). An open, empirical question is whether fractal analysis of physiological data may reveal flexibility in intentional forms of behavior. In conclusion, we have demonstrated that text production exhibits scaling properties like those observed in other cognitive phenomena. In doing so, we have also supported the idea that language is a complex, dynamical system involving coordination across many nested time scales. Going forward, our goal will be to further articulate time scales relevant to text production.

## Acknowledgments

This research is supported in part by the Institute of Education Sciences (R305A120707) and the Office of Naval Research (N00014140343 and N12249156). Any opinions or conclusions expressed are those of the authors and do not represent views of the IES or ONR.

## References

Allen, L., Snow, E., \& McNamara, D. (2016). The Narrative Waltz: The Role of Flexibility in Writing Proficiency. Journal of Educational Psychology, 108, 911-924.
Anastas, J., Stephen, D., \& Dixon, J. (2011). The scaling behavior of hand motions reveals self-organization during an executive function task. Physica A: Statistical Mechanics and its Applications, 390(9), 1539-1545.
Beran, J. (1994). Statistics for long-memory processes. Boca Raton, FL: Chapman and Hall/CRC Press.
Chen, Y., Ding, M., \& Kelso, J. (1997). Long memory processes (1/f $\alpha$ type) in human coordination. Physical Review Letters, 79(22), 4501-4504.
Chen, Y., Ding, M., \& Kelso, J. A. (2001). Origins of timing errors in human sensorimotor coordination. Journal of Motor Behavior, 33(1), 3-8.
Collins, J. J., \& De Luca, C. J. (1993). Open-loop and closed-loop control of posture: A random-walk analysis of center-of-pressure trajectories. Experimental Brain Research, 95, 308-318.

Dale, R., Kello, C., \& Schoenemann, P. (2016). Seekng synthesis: The integrative problem in understanding language and its evolution. Topics in Cognitive Science, 8, 371-381.
Davis, T. J., Brooks, T. R., \& Dixon, J. A. (2016). Multiscale interactions in interpersonal coordination. Journal of Sport and Health Science, 5(1), 25-34.
Delignières, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., \& Ninot, G. (2006). Fractal analyses for 'short' time series: a re-assessment of classical methods. Journal of Mathematical Psychology, 50(6), 525-544.
D'Mello, S., Picard, R., \& Graesser, A. (2007). Toward an affect-sensitive AutoTutor. IEEE Intelligent Systems, 22(4).
Eke, A., Herman, P., Kocsis, L., \& Kozak, L. (2002). Fractal characterization of complexity in temporal physiological signals. Physiological Measurement, 231), R1-R38.
Gilden, D. L., Thornton, T., \& Mallon, M. W. (1995). 1/f noise in human cognition. Science, 267, 1837-1839.
Gorman, J. C., Amazeen, P. G., \& Cooke, N. J. (2010). Team coordination dynamics. Nonlinear Dynamics, Psychology and Life Sciences, 14, 265-289.
Ihlen, E., \& Vereijken, B. (2010). Interaction-dominant dynamics in human cognition: Beyond $1 / f^{\alpha}$ fluctuation. Journal of Experimental Psychology: General, 139, 436463.

Ihlen, E., \& Vereijken, B. (2013). Multifractal formalisms of human behavior. Human Movement Science, 32, 633651.

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., \& Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. Physica A: Statistical Mechanics and its Applications, 316(1), 87-114.
Kelty-Stephen, D. G., \& Wallot, S. (in press). Multifractality versus (mono)fractality as evidence of nonlinear interactions across time scales: Disentangling the belief in nonlinearity from the diagnosis of nonlinearity in empirical data. Ecological Psychology.
Likens, A. D., Amazeen, P. G., Stevens, R., Galloway, T., \& Gorman, J. C. (2014). Neural signatures of team coordination are revealed by multifractal analysis. Social Neuroscience, 9(3), 219-234.
Likens, A., Fine, J., Amazeen, E., \& Amazeen, P. (2015). Experimental control of scaling behavior: what is not fractal? Experimental Brain Research, 233, 2813-2821.
Mandelbrot, B., \& Van Ness, J. (1968). Fractional Brownian motions, fractional noises and applications. SIAM review, 10(4), 422-437.
McNamara, D., Crossley, S., \& Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. Behavior Research Methods, 45, 499515.

McNamara, D., Crossley, S., Roscoe, R., Allen, L., \& Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing, 23, 35-59.

Nonaka, T., \& Bril, B. (2014). Fractal dynamics in dexterous tool use: The case of hammering behavior of bead craftsmen. Journal of Experimental Psychology: Human Perception and Performance, 40(1), 218.
Pinet, S., Ziegler, J., \& Alario, F. (2016). Typing is writing: Linguistic properties modulate typing execution. Psychonomic Bulletin \& review, 23, 18981906.

Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., \& Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. Physical Review E, 49(2), 1685-1689.
Rączaszek-Leonardi, J. (2010). Multiple time-scales of language dynamics: An example from psycholinguistics. Ecological Psychology, 22(4), 269-285.
Rączaszek-Leonardi, J., \& Kelso, J. (2008). Reconciling symbolic and dynamic aspects of language: Toward a dynamic psycholinguistics. New Ideas in Psychology, 26, 193-207.
Romero, V., Coey, C. A., Beach, A., \& Richardson, M. J. (2013). Effects of Target Size and Symmetry on the Structure of Variability in Precision Aiming. In $\operatorname{CogSci}$.
Snow, E. L., Allen, L. K., Jacovina, M. E., Crossley, S. A., Perret, C. A., \& McNamara, D. S. (2016). Keys to Detecting Writing Flexibility Over Time: Entropy and Natural Language Processing. Journal of Learning Analytics, 2(3), 40-54.
Stephen, D. G., \& Anastas, J. (2011). Fractal fluctuations in gaze speed visual search. Attention, Perception, \& Psychophysics, 73(3), 666-677.
Stephen, D. G., Arzamarski, R., \& Michaels, C. F. (2010). The role of fractality in perceptual learning: Exploration in dynamic touch. Journal of Experimental Psychology. Human perception and performance, 36(5), 1161.
Stephen, D. G., Broncoddo, R. A., Magnuson, J. S., \& Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. Memory \& Cognition, 37(8), 1132-1149.
Van Orden, G. C., Holden, J. G., \& Turvey, M. T. (2003). Self-organization of cognitive performance. Journal of Experimental Psychology: General, 132, 331.
Wagenmakers, E. J., Farrell, S., \& Ratcliff, R. (2004). Estimation and interpretation of $1 / \mathrm{f} \alpha$ noise in human cognition. Psychonomic Bulletin \& Review, 11, 579-615.
Wallot, S., \& Grabowski, J. (2013). Typewriting Dynamics: What Distinguishes Simple From Complex Writing Tasks?. Ecological Psychology, 25(3), 267-280.
Wallot, S., Hollis, G., \& van Rooij, M. (2013). Connected text reading and differences in text reading fluency in adult readers. PloS One, 8(8), e71914.
Wallot, S., \& Van Orden, G. (2011). Toward a lifespan metric of reading fluency. International Journal of Bifurcation and Chaos, 21(04), 1173-1192.
Wijnants, M. L., Hasselman, F., Cox, R. F. A., Bosman, A. M. T., Van Orden, G. (2012). An interaction-dominant perspective on reading fluency and dyslexia. Annals of Dyslexia, 62, 100-119.

# Visuomotor Adaptation and Sensory Recalibration in Reversed Hand Movement Task 

Jenny Lin ${ }^{1 *}$<br>jenny.h.lin98@gmail.com<br>Yixin Zhu ${ }^{1 \star} \quad$ James Kubricht ${ }^{2 \star} \quad$ Song-Chun Zhu ${ }^{1}$<br>yixin.zhu@ucla.edu kubricht@ucla.edu<br>${ }^{1}$ Department of Statistics ${ }^{2}$ Department of Psychology<br>* Equal Contributors University of California, Los Angeles


#### Abstract

Visuomotor adaptation plays an important role in motor planning and execution. However, it remains unclear how sensorimotor transformations are recalibrated when visual and proprioceptive feedback are decoupled. To address this question, the present study asked participants to reach toward targets in a virtual reality (VR) environment. They were given visual feedback of their arm movements in VR that was either consistent (normal motion) with the virtual world or reflected (reversed motion) with respect to the left-right and vertical axes. Participants completed two normal motion experimental sessions, with a reversed motion session in between. While reaction time in the reversed motion session was longer than in the normal motion session, participants showed the learning improvement by completing trials in the second normal motion session faster than in the first. The reduction in reaction time was found to correlate with greater use of linear reaching trajectory strategies (measured using dynamic time warping) in the reversed and second normal motion sessions. This result appears consistent with linear motor movement planning guided by increased attention to visual feedback. Such strategical bias persisted into the second normal motion session. Participants in the reversed session were grouped into two clusters depending on their preference for proximal/distal and awkward/smooth motor movements. We found that participants who preferred distal-smooth movements produced more linear trajectories than those who preferred proximal-awkward movements.


Keywords: Virtual reality; motor planning; scene representation; visual misalignment

## Introduction

## Virtual Reality

Virtual reality (VR) technology provides an analog experience in a three-dimensional environment similar to that of the real world. In the real world, certain environmental factors and physical constraints are fixed and cannot be modified. However, VR allows researchers to design controlled virtual environments with ease and precision. In addition, modern advancements in VR tracking allow for accurate measurements of human body movements. Thus, task success, motor error and correspondence with candidate trajectories can be accessed directly.

Although previous studies in VR have focused primarily on hardware problems in order to improve user experience (Shotton et al., 2013; Weichert, Bachmann, Rudak, \& Fisseler, 2013), simulation performance (e.g., Unreal Engine 4, Unity3d, and NVidia Flex), system integration (Lin et al., 2016; Shah, Dey, Lovett, \& Kapoor, 2017), and locomotion in immerse experience (Bruder \& Steinicke, 2014), recently efforts have been increasingly devoted to examining human perception and reasoning in virtual scenes (e.g., Azmandian, Hancock, Benko, Ofek, \& Wilson, 2016; Mehra et al., 2016;

Patney et al., 2016; Ye et al., 2017; Li, Liang, Quigley, Zhao, \& Yu, 2017).

## Motor Planning

The process of reaching toward an object in the environment involves minimizing the distance between the hand and target locations in the physical world (i.e., the hand and target states) over time. This is achieved by (1) planning a motor movement to achieve a desired task goal, (2) sending the associated motor command to the arm, and (3) comparing observed sensory feedback to predicted sensory feedback to infer the current hand state and form subsequent motor commands (i.e., sensorimotor transformation; BattagliaMayer et al., 2014; Wolpert, 1997). The present study examined how reaching movements change in response to misaligned sensory feedback in a VR environment. Specifically, how do reaching trajectories change as visual and proprioceptive feedback are decoupled?

When visual and proprioceptive feedback are inconsistent, new mappings between visual and proprioceptive inputs are reestimated (Cressman \& Henriques, 2009). Results from Cressman and Henriques's (2009) study suggest that in addition to sensorimotor recalibration, visuomotor adaptation involves partial proprioceptive recalibration: i.e., humans "realign proprioceptive estimates of hand position to match visual estimates." However, it has been demonstrated that visuomotor adaptation can occur in the absence of proprioceptive input, for example, in the case of deafferented individuals (Ingram et al., 2000; Miall \& Cole, 2007). It is therefore possible that proprioceptive recalibration does not underlie visuomotor adaptation and that the two processes are independent from one another. This hypothesis is consistent with empirical results showing that humans curtail the contribution of proprioceptive input in the case of misaligned visual feedback (Bernier, Burle, Vidal, Hasbroucq, \& Blouin, 2009; Wont \& Henriques, 2009) when performing motor movements. Thus, when visual and proprioceptive feedback are inconsistent, people could reduce the contribution of proprioceptive information to the motor planning process and form new visuomotor transformations to achieve extrinsic goals.

In this study, participants reached toward targets in a virtual environment where their hand movements were shown to be either consistent or reversed (in the vertical and leftright axes) with respect to virtual movement. If proprioceptive inputs are ignored (perhaps due to their unreliability in the reversed movement environment), participants should rely more heavily on visual inputs when planning and executing movements. Moreover, we expect participants will adapt to the reversed environment by constructing and implementing new visuomotor mappings. Although we pre-


Figure 1: Participants reached toward targets in a virtual environment where their hand movements were shown to be either (a) consistent, or (b) reversed (in the vertical and left-right axes) with respect to virtual movement. (Top) Real world actions. (Bottom) Virtual simulation.
dict proprioceptive inputs to be ignored or even suppressed, we expect proprioceptive feedback to be considered in cases where visually-guided movement is kinesthetically awkward. Using rich trajectory measurements from a VR system, we compared performance between participants who appeared to adopt different strategies guided either by visual or proprioceptive feedback. In summary, the purpose of the present study was to quantitatively compare reaching strategies in a novel VR task across normal- and reversed-motion environments and to determine whether changes in reaching strategies persist when visual and proprioceptive information are re-coupled.

## Experiments

In the present study, we examined whether humans can adapt to environments where visual estimates of objects' positions are inconsistent with proporioceptive input. Participants interacted with virtual targets using two motion controllers in a VR application, where the movement of the virtual controller either matched the motion of the physical controller or was flipped on certain axes (both vertical and left-right). Participants were instructed to touch a series of virtual targets with the virtual controllers and then return to a neutral pose in between targets. Response time and arm movement trajectories were recorded and analyzed.

## Participants and Apparatus

A total of 20 participants ( 10 female and 10 male) participated in the study. Participants were graduate students at the University of California, Los Angeles. The average age of participants was 22.8 years old with a standard deviation of 2.67. All participants had normal or corrected-to-normal vision. Of the 20 participants, 16 had never interacted with VR technology prior to participating in the experiment.

The VR system integrated Unreal Engine 4 with an HTC Vive headset and two motion controllers, one held in each hand. 3D meshes which matched the Vive motion controllers in size and shape were used to represent controller position in the virtual environment. To generate the visual display of reversed movement, the virtual controller in VR was moved in opposite directions (i.e., in both the vertical and left-right
axes) to the physical displacement of the controller moved by human participants in the real world (Fig. 1). Participants began the experiment by moving their hands into a neutral pose where the physical and virtual controllers were aligned to the same position. Movement along the depth axis (i.e., forward vs. backward) was not reversed.

The targets were cyan capsules of 20 cm height and diameter. We chose cyan as the color of the targets in order to ensure the targets would be visible against the background of the environment. The targets began glowing when touched by a controller, providing visual feedback to the subject indicating whether they had successfully touched the target. The color of the targets did not change between experimental sessions. To ensure that for any given target location each participant reached approximately the same distance, we required that the participants assume a neutral pose before the next target was spawned. We define the neutral pose as follows:

At the beginning of each testing block, participants were told to hold both controllers in front of them at waist level with their elbows held loosely at their sides. Participants were allowed to adjust their pose until they were comfortable, but were informed that they needed to be able to comfortably reach forward, up-down and side-to-side from this pose. Participants then started an experimental block by pressing the trigger button on the bottom of either motion controllers. A transparent rectangular prism was spawned such that its center was located at the midpoint of the two controllers. This rectangular prism defined each participant's neutral zone, and we considered the participant to be in a neutral pose when both controllers overlapped with the neutral zone for an uninterrupted 0.5 seconds. In order to provide feedback to the user about whether they were in a neutral pose, the neutral zone changed color to reflect how many controllers overlapped with it: black for zero controllers, grey for one, and green for two. The neutral zone only changed color when the participant needed to enter a neutral pose, and otherwise remained green.

Response time was defined as the duration between the initial spawning of the target to when it was deactivated. Trajectory was defined as the three-dimensional movement of the controllers over this time period. For a video demonstrating the experimental setup, please see https://vimeo.com/216580864.

## Procedure

The experiment was conducted in a quiet office, and all physical obstacles were removed from the testing area. Participants remained standing and stationary for the duration of the experiment. They received a warning signal if they moved near the boundaries of the virtual environment.
Practice Session First, participants familiarized themselves with the VR headset and motion controllers. Participants were given a demonstration of the neutral position and told to move both of their controllers to the indicated locations. After participants confirmed that they were capable of comfortably performing the required range of movements from their neutral pose, they were informed that both response time and movement trajectories would be recorded. Prior to the testing session, participants completed a practice session with five

(a)

Figure 2: Response time analysis for the normal- (N1 and $N 2$ ) and reversed-motion ( $R 1$ ) trial sessions. Red horizontal lines indicate median response times. The bottom and top edges of the blue boxes indicate the 25 th and 75 th percentiles, respectively. The whiskers extend to the most extreme data points that were not considered outliers, and red ' + ' symbols indicate outliers. (a) Session median times to reach targets: $1.29,5.71$ and 1.03 seconds. (b) Block median times to reach targets: $1.30,1.22,1.34,6.10,5.68,5.36,1.13,1.06$ and 0.94 seconds.
targets in the normal condition. This session served to familiarize participants with the experimental procedure and provide the experience of interacting with objects in the virtual environment.

Testing Session Participants completed nine blocks consisting of ten trials each. The first three blocks ( $N 1$ session) were completed with normal movements. The subsequent three blocks ( $R 1$ session) were completed with reversed movements. Participants were informed along which axes arm movement would be reversed (i.e., the left-right and updown directions). The last three blocks ( $N 2$ session) were completed with normal movements once again. Participants were given breaks between blocks to rest their arms. After indicating that they were ready to continue, participants proceeded to the subsequent block.

At the start of each block, the virtual meshes were aligned with the locations of the physical controllers. Each participant completed the same nine blocks. Target locations were evenly distributed throughout an $80 \times 20 \times 80 \mathrm{~cm}$ region located 35 cm in front of the neutral zone. The order of the target positions within each block was randomized between participants.

## Results

## Response Time Analysis

As expected, participants showed much longer response times (RT) in the reversed-motion condition than in the normalmotion condition. There was a four-fold increase between median RT for the $N 1$ relative to the $R 1$ session. Interestingly, upon returning to normal movement in the $N 2$ session, participants showed a $20.1 \%$ improvement in response time compared to the $N 1$ session $(t(600)=7.07$, $p<.001$; see Fig. 2a). Moreover, response times in the three blocks of the $N 2$ session displayed a decreas-


Figure 3: Trajectory analysis using DTW to quantify the discrepancy between human reaching behavior and a linear motion trajectory - the straight line between the hand's starting position and target location. Median distance scores for each session: $505.62,613.24$, and 483.55 cm . Block medians: 501.03, 525.70, 496.62, 671.75, 648.25, 582.34, 513.57, 490.28 , and 435.84 cm .
ing trend $(b=-0.0067[-0.0101,-0.0033])$, indicating a learning effect that was not present in the $N 1$ session ( $b=$ $-0.0009[-0.0053,0.0035])$; Fig. 2b).

## Trajectory Analysis

Next, trajectory analysis was performed to further quantify human performance relative to candidate trajectories. We define the baseline trajectory as the shortest linear path between the hand start position and the target location. All trajectories were interpolated to 500 3D points to account for variation in trajectory length. Dynamic Time Warping (DTW) was then utilized to determine the minimum distance mapping between the ideal and behavioral trajectories. DTW is a distance measure algorithm that has been used extensively in the speech recognition community (e.g. Berndt \& Clifford, 1994). By estimating a non-linear mapping between two time-dependent sequences, DTW provides a numerical representation of the similarity between any pair of spatiotemporal sequences. Other communities including robotics and biology have also adopted and modified this algorithm for various signal-comparison applications.

The DTW trajectory distance measure revealed closer correspondence to baseline trajectories in the $N 2$ session compared to the $N 1$ session (Fig. 3a), suggesting a learning effect through practice. There was also a clear decrease in DTW distance across the three blocks within the $N 2$ session that was not evident in the $N 1$ session (Fig. 3b), suggesting humans moved their arms more linearly (i.e., closer to the baseline linear trajectory) upon return to the normal motion environment. To rule out the possibility that the increasing linear movements in the $N 2$ session was due to familiarization with the VR system, we performed a linear regression on the median trajectory difference among participants (Fig. 4). Although there is no noticeable trend in the $N 1$ session, performance in the $N 2$ session shows a strong improvement that falls well outside the $95 \%$ confidence region for $N 1$. Moreover, the slope in $N 2$ was approximately equal to that in the $R 1$ session, although the regression coefficient in the $R 1$ ses-


Figure 4: Linear regression results using median DTW distance among 20 subjects across 90 trials divided into 3 sessions and 9 blocks. Red dashed lines represent 95\% confidence intervals for the regression coefficient estimates. Slopes in the three sessions are $0.216,-3.810$, and -4.436 .
sion is more uncertain: i.e., the confidence interval of the $R 1$ slope is greater than that of $N 2$. This suggests a large degree of within-group variability, which is further explored in the following sections.

After forming new visuomotor mappings in the the $R 1$ session, participants' movement trajectories became increasingly linear: i.e., closer to the baseline trajectory. If participants began relying on visual feedback when constructing and revising their motor plans (i.e., proprioceptive inputs were suppressed), we would expect them to execute linear movement paths. The increasingly linear motor movements over the course of the $R 1$ session are consistent with this prediction. Interestingly, reliance on visual inputs appeared to persist in the following $N 2$ session when proprioceptive and visual information were recoupled. We predict that with further exposure to the normal-motion environment, the linearity of participants' reaching patterns would return to the level measured in the $N 1$ session.

## Possible Planning Models in Reversed Motion Blocks

While the shortest linear path between two points is the most direct trajectory, it is not necessarily the most optimal reaching strategy: e.g., due to mechanical limb constraints. To examine this, we used DTW to compare against other candidate trajectories to assess their potential as possible movement strategies. One possible alternative strategy is to consider each axis independently in order to plan motor movements in the reversed motion condition. To examine this alternative strategy, human trajectories were compared to all six possible axis decompositions (Fig. 5a) using DTW. While some participants did demonstrate paths that were more similar to various axis decompositions, participants' trajectories were generally more similar to the shortest linear path (Fig. 3b), indicating that most participants were not considering each axis independently.

Another observation of participants' trajectories is that they were noisy, especially during the reversed motion session. Since participants were instructed to reach a set of given targets, their movements were goal-directed and partially guided. We compared participants' trajectories with predictions from a guided random walk model Pearson (1905).


Figure 5: (a) Six possible axis decompositions were generated by computing the shortest path along each axis. (b) Human trajectories were compared against all six axis decompositions using DTW, and the minimum value was reported. Session medians: $736.72,734.38$, and 661.90 cm . (c) 10 of the guided random walks generated between the given start and end point. (d) Human trajectories were compared against 100 guided random walks using DTW, and the most similar value was reported. Block medians: 315.03, 307.15, 300.95, $522.34,515.89,452.21,310.25,278.97$, and 247.46 cm .
Given a starting point, a set of 100 proposed moves were generated within a 5 cm radius. Next, the model computed the distance between each of the proposed movements and the end point. A movement was then chosen from two options: 1) the shortest distance with probability .2 , or 2 ) randomly chosen movement among the 100 (random) proposed movements with probability .8. Finally, after approximately a few hundred iterations, the guided random walk model converged and reached the end point, as shown in Fig. 5c. Measured by DTW, human movement trajectories were found to be more similar to the guided random walks not only during the reversed-motion session but also during both normal-motion sessions (Fig. 5d). The fit of the model predicted trajectories to human performance across all the three sessions suggests that participants' motor movements were goal-directed but executed with inherent motor noise.

## Movement Strategies in Reversed Motion Blocks

In the normal-motion sessions, participants consistently used both arms to perform the reaching task, while favoring the controller closest to the target. In the reversed-movement session, however, a variety of strategies emerged. Some participants predominantly used one hand regardless of the location of the target relative to their neutral zone. Others favored the hand that was furthest from the target. Thus, we further examined the distribution of participants' reaching strategies.

In certain experimental trials, touching the target with the nearest hand required the participant to reach across their body while looking in the opposite direction, due to the reversed axes. This pose is physically difficult to accomplish. In contrast, the participant could reach for the target with their opposite hand, resulting in a pose that was physically comfortable. However, this would require the participant to use the hand that was physically furthest away, which is highly nonintuitive (Fig. 6). The cost to execute a path is thus dependent on not just proximity but also kinesthetic ease of execution.

We examined the interplay between the two constraints (i.e., proximity and ease of motor execution) in planning motor movements. Criteria were defined as follows: a trajectory


Figure 6: Illustration of different movement strategies in the reversed-motion session. (a) Solid lines indicate the trajectories visualized in VR. Dashed lines indicate the corresponding real-world trajectories of participants' hands. Red trajectory indicates the path executed by the participant, and the blue trajectory indicates the shortest computed path from the opposite hand. In this case, the target is located to the left of the participant in the virtual environment. (b) The experimenter demonstrates the awkward pose with the shorter trajectory (top) and the equivalent comfortable pose with the longer trajectory (bottom).
is considered proximal if a participant uses the hand initially closest to the target, and considered distal if he uses the hand initially furthest from the target. The trajectory is considered awkward if it requires reaching across the body's center and smooth if it does not. These criteria result in four different trajectory categories: proximal-smooth, proximalawkward, distal-smooth, and distal-awkward (See Fig. 7). In the normal-motion sessions, participants strongly favored the proximal-smooth strategy, with the distal-awkward strategy occurring only in a few selected trials where the target was close to the mid-line. In the reversed motion session, participants demonstrated all three strategies except the distalawkward.

We performed k-mean clustering on participants' trajectories in the reversed-motion session and found that two stable clusters emerged. Cluster size was split evenly at ten participants each, indicating that half of the participants were more likely to use the proximal-awkward strategy and the other half


Figure 7: Four different trajectory categories. (a) Proximalsmooth.
(b) Proximal-awkward.
(c) Distal-smooth. Distal-awkward.


Figure 8: K-mean clustering ( $k=2$ ) results on reversed strategy. (Left) 10 participants favored distal-smooth reaching strategies, indicating that they were utilizing predictions about proprioceptive feedback and actively reasoning about whether the motions would lead to awkward movements, whereas (Right) the other 10 participants preferred proximalawkward reaching strategies, indicating that they primarily utilizing visual information.
were more likely to use the distal-smooth strategy. The former group favored visual proximity: i.e., they attempted to reach the target using the hand that was closest to the target. The latter group favored smooth motion: i.e., they used learned associations between proprioceptive feedback and visual movement to predict which hand choice would result in the least awkward pose. In this case, participants were required to imagine the potential trajectories and associated proprioceptive feedback to plan their movement. These findings suggest that humans adopt different strategies to cope with the novel task in the reversed motion session by focusing on either spatial proximity for efficiency or smooth motion to avoid impossible or awkward poses.

A linear regression analysis was performed on DTW measurements after separating participants into the two groups as shown in Fig. 9. It is clear that the pose-focused participants demonstrated greater improvement compared to proximityfocused participants, although this learning effect did not persist in the subsequent normal-motion session.

## Discussion

When planning motor movement according to misaligned visual feedback, proprioceptive feedback has been shown to be suppressed while attention to visual information is enhanced. We hypothesized that in the case of reversed virtual feedback, target-directed reaching movements would rely primarily on visual feedback and thus accord with candidate linear trajectories. This prediction is confirmed by participants in the reversed-motion session using only a single hand, which arguably arises due to the relative ease of forming new visuomotor mappings with a single arm compared to both arms simultaneously. We found that participants in the reversedmotion session ( $R 1$ ) exhibited a preference for linear trajectories, which agrees with increasing suppression of using proprioceptive information to guide motor movements. Interestingly, this increasing linear preference-and corresponding reliance on newly formed visuomotor mappings-persisted into the second normal motion session ( $N 2$ ) although it was not observed in the first normal session ( $N 1$ ). We predict that this bias toward linear movement strategies would diminish with


Figure 9: DTW distance to linear reaching trajectories for (a) reasoning-focused and (b) perception-focused participants. Those participants that utilized predictions about proprioceptive feedback to guide their reaching movements showed increasingly linear trajectories compared to those participants who primarily utilized visual information. Slopes in (a): -$0.11,-5.22,-4.58$. Slopes in (b): $-0.23,-2.10,-4.38$
further exposure to the normal-motion environment, as traditional sensorimotor mappings utilizing proprioceptive information are employed.

However, the main finding of the present study could have resulted from increased familiarity with the VR system and environment. Thus, a follow-up study to this experiment is to establish a second control condition where each of the three experimental sessions involve normal motion. If performance does not vary across the three normal sessions, the finding that reversed motion increases preference toward visuallyguided, linear motor movements would be strengthened. Additionally, movement in the virtual world was reversed on two axes (vertical and left-right) in the present study. Future work should examine how performance changes when a single axis-or different pairs of axes-are flipped. Moreover, would exposure to one reversed axis improve performance under a second (different) reversed axis?

Tactile signals are an important cue for planning and executing object interactions (Johansson \& Flanagan, 2009). One of the major disadvantages with current commercial VR products is that tactile feedback is missing in the virtual world. In the present study, we compensated for the lack of tactile feedback by using additional visual cues to indicate successful reach events; however this does not change the fact that a significant source of feedback is missing. For future studies it would be worth providing a haptic signal through the controller's actuators or using a tactile data glove to administrate more fine-grained feedback. We predict that implementing haptic feedback to the current experimental method would inhibit suppression of proprioceptive information and consequently interfere with the formation of new visuomotor mappings.

Future work should also examine sensorimotor recalibration in more complicated tasks than the present reaching movements: e.g., stacking blocks or completing towers of Hannoi problems. In these tasks, cognitive resources are devoted to planning a sequence of motor movements, which may yield strong interference to the visuomotor adaptation process and provide a unique window to study the interplay between motor planning and reasoning.
Acknowledgments The work reported herein was supported by DARPA XAI grant N66001-17-2-4029, DARPA SIMPLEX grant N66001-15-C-4035, ONR MURI grant

N00014-16-1-2007, NSF grant BCS-1353391, and a NSF Graduate Research Fellowship.

## References

Azmandian, M., Hancock, M., Benko, H., Ofek, E., \& Wilson, A. D. (2016). Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences. In Proceedings of the 2016 chi conference on human factors in computing systems (pp. 1968-1979).
Battaglia-Mayer, A., Buiatti, T., Caminiti, R., Ferraina, S., Lacquaniti, F., \& Shallice, T. (2014). Correction and suppresion of reaching movements in the cerebral cortex: Physiological and neuropsychological aspects. Neuroscience and Biobehavioral Reviews, 42, 232-251.
Berndt, D. J., \& Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In Kdd workshop (Vol. 10, pp. 359-370).
Bernier, P.-M., Burle, B., Vidal, F., Hasbroucq, T., \& Blouin, J. (2009). Direct evidence for cortical suppression of somatosensory afferents during visuomotor adaptation. Cerebral Cortex, 19(9), 2106-2113.
Bruder, G., \& Steinicke, F. (2014). Threefolded motion perception during immersive walkthroughs. In Proceedings of the 20th acm symposium on virtual reality software and technology (pp. 177185).

Cressman, E. K., \& Henriques, D. Y. P. (2009). Sensory recalibration of hand position following visuomotor adaptation. Journal of Neurophysiology, 102, 3505-3518.
Ingram, H. A., Van Donkelaar, P., Cole, J., Vercher, J. L., Gauthier, G. M., \& Miall, R. C. (2000). The role of proprioception and attention in a visuomotor adaptation task. Experimental Brain Research, 132(1), 114-126.
Johansson, R. S., \& Flanagan, J. R. (2009). Coding and use of tactile signals from the fingertips in object manipulation tasks. Nature Reviews Neuroscience, 10(5), 345-359.
Li, C., Liang, W., Quigley, C., Zhao, Y., \& Yu, L.-F. (2017). Earthquake safety training through virtual drills. IEEE Transactions on Visualization and Computer Graphics, 23(4), 1275-1284.
Lin, J., Guo, X., Shao, J., Jiang, C., Zhu, Y., \& Zhu, S.-C. (2016). A virtual reality platform for dynamic human-scene interaction. In Siggraph asia 2016 virtual reality meets physical reality: Modelling and simulating virtual humans and environments (p. 11).
Mehra, R., Hohnerlein, C., Perek, D., Gatti, E., DeSalvo, R., \& Keller, S. (2016). Hapticwave: directional surface vibrations using wave-field synthesis. In Acm siggraph 2016 emerging technologies (p. 11).
Miall, R. C., \& Cole, J. (2007). Evidence for stronger visuomotor than visuo-proprioceptive conflict during mirror drawing performed by a deafferented subject and control subjects. Experimental Brain Research, 176(3), 432-439.
Patney, A., Kim, J., Salvi, M., Kaplanyan, A., Wyman, C., Benty, N., ... Luebke, D. (2016). Perceptually-based foveated virtual reality. In Acm siggraph 2016 emerging technologies (p. 17).
Pearson, K. (1905). The problem of the random walk. Nature, 72(1865), 294.
Shah, S., Dey, D., Lovett, C., \& Kapoor, A. (2017). Aerial Informatics and Robotics platform (Tech. Rep. No. MSR-TR-2017-9). Microsoft Research.
Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., ... Moore, R. (2013). Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1), 116-124.
Weichert, F., Bachmann, D., Rudak, B., \& Fisseler, D. (2013). Analysis of the accuracy and robustness of the leap motion controller. Sensors, 13(5), 6380-6393.
Wolpert, D. M. (1997). Computational approaches to motor control. Trends in cognitive sciences, 1(6), 209-216.
Wont, T., \& Henriques, D. Y. P. (2009). Visuomotor adaptation does not recalibrate kinesthetic sense of felt hand path. Journal of Neurophysiology, 101(2), 614-623.
Ye, T., Qi, S., Kubricht, J., Zhu, Y., Lu, H., \& Zhu, S.-C. (2017). The martian: Examining human physical judgments across virtual gravity fields. IEEE Transactions on Visualization and Computer Graphics, 23(4), 1399-1408.

# The development of turn-taking: Pre-schoolers may predict what you will say, but they don't use those predictions to plan a reply. 

Laura Lindsay<br>University of Edinburgh, Edinburgh, United Kingdom<br>Chiara Gambi<br>University of Edinburgh<br>Hugh Rabagliati<br>University of Edinburgh, Edinburgh, UK


#### Abstract

Whereas adults exchange conversational turns very rapidly, children often respond after long gaps. However, it has been proposed that the infrastructure necessary to take turns develops in infancy. Why are children slow to respond to turns? Adults' turn-taking skills, it has been argued, rely on an ability to both predict when the current turn will end and prepare a response as soon as possible. In two experiments, we ask how these two abilities (prediction and early preparation) develop. Adults and 3-to-5-year-olds answered yes/no questions while playing an iPad-based maze game. Distributional analysis of answer latencies suggest that (i) neither children nor adults rely on fine-grained predictions of turn duration and (ii) both children and adults use predictions about turn content to prepare their answer early. In sum, by the age of three, children already have the cognitive architecture necessary to take turns successfully.,


# Mouse Tracking Shows Attraction to Alternative Targets While Grounding Spatial Relations 

Jonas Lins (jonas.lins@ini.rub.de) Gregor Schöner (gregor.schoener@ini.rub.de)

Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany


#### Abstract

Evidence that higher cognitive processes are coupled in a graded and time-continuous way to sensory-motor processes comes, in part, from mouse-tracking studies. In these, curved mouse trajectories toward one of two fixed response locations reveal the evolution of certainty about a cognitive task that participants solve. We present a paradigm in which selection of the response location is itself the cognitive task. From among items in a visual scene, participants select a target that is described by a spatial relation (e.g.,"the red to the left of the green"), where one target item (here, "red") matches the description better than alternative same-colored targets. In the mouse trajectories, we find clear evidence for attraction to the alternative targets, attraction to the reference item (here "green"), and an early biasing influence of the spatial term.


## Introduction

Over the last two decades, a major theme in cognitive science has been that cognitive processing is graded in nature, unfolds continuously in time, and is coupled to perceptual and motor processes (Schöner, Spencer, \& the DFT Research Group, 2015; Spivey, 2007). Movement preparation, for instance, was shown to occur in a graded and continuous form when the time interval between imperative stimulus and response was varied experimentally in the timed movement initiation paradigm (Ghez et al., 1997). The theoretical account (Erlhagen \& Schöner, 2002) linked this phenomenon to graded distributions of population activation in the motor and premotor cortex (Bastian, Schöner, \& Riehle, 2003; Cisek \& Kalaska, 2005). That higher-level cognitive processes, such as understanding spatial concepts, interact with perceptual and motor processes has been seen through interaction effects in reaction time tasks that probed potentially overlapping perceptual and motor representations (e.g., Richardson, Spivey, Barsalou, \& McRae, 2003). Mouse tracking and similar techniques have been a major tool to study and establish this link between cognitive and sensory-motor processes (for review see Freeman, Dale, \& Farmer, 2011; Song \& Nakayama, 2009).

In mouse tracking paradigms, participants are asked to solve a cognitive task that may engage high-level concepts. The response is usually ensured to begin early in relation to the decision process, analogously to the timed movement initiation paradigm. Typically, a computer mouse must be moved toward one of two (sometimes a few) response locations. What varies over time is the certainty over the response, which is reflected by the movement deviating somewhat into intermediate directions. A limitation of previous mouse tracking research with respect to the interaction of cognitive and sensorimotor processes is that the motor responses are usually fixed and assigned symbolically to the
solutions of the cognitive task (e.g., left button for "yes" and right button for "no"). When a small number of possible movement targets is known in advance, it is not the decision itself that specifies the metrics of the required movement.

We developed an experimental paradigm in which the spatial target of a mouse movement was directly specified by the cognitive task and unknown in advance. Participants read a target description invoking colors and spatial relations (e.g., "the yellow to the left of the green") and then saw a complex layout of colored objects (e.g., Fig. 1c). A selection decision had to be made among multiple same-colored target items ("yellow"), clicking the one best matching the spatial term. Although some aspects of the visual arrays are tightly coupled and impossible to vary independently, we were able to unravel their effects by counter-balancing those aspects that posed potential confounds, separately for each comparison.

We looked for three signatures of interaction among the task's cognitive and sensory-motor dimensions. First, the alternate targets (i.e., distractors) were predicted to metrically attract the trajectories, in analogy to the effect of alternate but incorrect choice alternatives in classic mouse tracking research. Second, while the reference item is never an alternate target, it likely engages attentional focus during processing (Franconeri, Scimeca, Roth, Helseth, \& Kahn, 2012), which we predicted to also cause attraction. Third, based on previous evidence (Tower-Richardi, Brunye, Gagnon, Mahoney, \& Taylor, 2012), we expected a bias into the direction described by the spatial term. In our recent neural process model of spatial language grounding (Richter, Lins, \& Schöner, 2017), discrete amodal representations of target, reference, and spatial term guide activation in continuous perceptual representations. In the model, target and reference must become active sequentially, because overlapping substrates are engaged to spatially index the corresponding visual items. We thus expected some temporal displacement in the biases toward these items. The spatial term, in contrast, impacts another substrate and is active early and in parallel to target and reference. We thus expected the spatial term effect to bias movement metrics globally and early on.

## Methods

Participants Twelve participants ( 5 female, 7 male, mean age 27.4 years $\pm 3.8$ s.d., one left-handed) were recruited by notices at the local department, receiving EUR 10 for participation. All were naïve to the experimental hypotheses, native German speakers, and had normal vision.

Material The experiment was run using Matlab and the Psychophysics Toolbox (Brainard, 1997) and presented on a 22 " LCD screen (Samsung, 226BW; visible image $475 \mathrm{~mm} \times$ 197 mm ) at a viewing distance of approximately 70 cm (subtending approx. $40.4^{\circ} \times 16.2^{\circ}$ of visual angle, v.a.). Trajectories were collected using a standard computer mouse (Logitech, M-UAE96, mean sampling rate 92.17 Hz ). Mouse speed was set such that movement on the tabletop translated to cursor movement over the same distance, to make motions more similar to natural arm movements and simplify cognitive transformation from hand coordinates to screen space.

Procedure A trial began with a black start marker in the bottom center of an otherwise gray screen. To proceed, the participant moved the mouse cursor (a white dot) onto the start marker. After resting there for 300 ms , a German spatial phrase appeared at a position somewhat random around the center of the stimulus region (up to $\pm 48 \mathrm{~mm} / 20 \mathrm{~mm}$ horiz./vert.), for instance, "Das Gelbe links vom Grünen." ("The yellow left of the green."), denoting a target item by its color and its position in relation to a reference item, which was also specified by color. Display duration ranged randomly from one to two seconds. The phrase then disappeared and a beep signaled the participant to start moving the cursor upwards within one second (if movement started too late, the trial was aborted and appended at the end). Movement onset was registered when a velocity of $20 \mathrm{~mm} / \mathrm{s}$ was exceeded. At that point, twelve colored items appeared above the start marker (e.g., Fig. 1c). Thus, movement was already in progress when the selection process started. One of the twelve items was the uniquely colored reference mentioned in the phrase, one was the target, and one was the main distracter, which had the same color as the target but provided a worse match for the spatial term, according to a spatial template described below. The participant had to select the item which in his or her opinion best matched the preceding phrase (participants could select any item). If time since movement onset exceeded two seconds, the trial was aborted and appended at the end. Participants were instructed that there were no incorrect responses, that items were not obstacles, and that response time was limited such that they had to respond promptly. After 13 practice trials, each participant completed 446 trials in random order (one completed eight more, to use the entire set of 5360 trials, described below).

Spatial phrases Spatial phrases were in German and of the form "Das art Grüne $_{\text {tgt }}$ rechts vom ${ }_{\text {spt }}$ Roten $_{\text {ref }}$ " ("The ${ }_{\text {art }}$ green $_{\text {tgt }}$ to the right of spt $^{\text {the }}$ red $_{\text {ref }}$ "), where art denotes the article, which was always "Das", tgt denotes a target component from the set \{Rote, Grüne, Blaue, Gelbe, Weiße, Schwarze\}, ref denotes a reference component from the set $\{$ Roten, Grünen, Blauen, Gelben, Weißen, Schwarzen\} ("the \{red, green, blue, yellow, white, black\} one"), and spt denotes a spatial term from the set \{links vom, rechts vom, über dem, unter dem \} (\{left of, right of, above, below \}). In all trials, the spatial phrase posed a valid
description of an item in the stimulus display.
Stimulus displays Visual items were irregular polygons with an outer diameter of 16.4 mm ( $1.34^{\circ}$ v.a.), each colored in one of the six colors that also occurred in the spatial phrases (red, green, blue, yellow, white, and black). Items were combined into stimulus displays as described in the following.

We generated multiple three-item configurations of a reference item, a target item, and a distracter, differing in how target and distracter were situated relative to the reference. Their positions were selected based on a spatial template, a fit function $f(\phi, r)$ with the reference at $[0,0]$ that indicates how well a given position, defined by angle $\phi$ and radius $r$, matches a spatial term. ${ }^{1}$ Fig. 1a (left panel) shows a plot over Cartesian coordinates for "right of". The shape of the spatial templates is inspired by behavioral data (e.g., Logan \& Sadler, 1996) which computational modeling work reproduced using similar functions (Lipinski, Schneegans, Sandamirskaya, Spencer, \& Schöner, 2012).

Targets were placed in a region where $f(\phi, r)>0.6$ and where the outer radices of reference and target were separated by at least $0.5 \mathrm{~mm}\left(0.04^{\circ}\right.$ v.a.; Fig. 1a, center panel). Within this region, targets were centered on the junctions of a square grid, resulting in sixteen evenly distributed target positions. For each of the 16 target positions a separate set of distracter positions was determined (out of which one distracter was used per trial, paired with the respective target). These were obtained with the same method as the targets, but the general region for distracters was constrained by $f(\phi, r)>0.4$ (green outline in Fig. 1a, center panel), and as an additional constraint distracters' fit had to be at least 0.03 lower than target fit (min. border-to-border distance again 0.5 mm ; see Fig. 1a, right panel). Hence, the shape of distracter regions differed between target positions so that distracter numbers differed as well, varying from 16 to 25 (mean 20.9) per target position. Colors for each three-item set were randomly picked, with target and distracter being colored alike.

A set of 335 different three-item configurations was obtained for each spatial term, differing between terms only in orientation. We thus arrived at 1340 configurations, each of which was presented at four different positions on the screen, such that the target item of each configuration appeared once in each of four different on-screen target locations (black X's in Fig. 1b). These were arranged in a square around the cen-

$$
\begin{aligned}
& { }^{1} \text { In polar coordinates the function is given by } \\
& f(\phi, r)=e^{\left[-\frac{\left(\phi-\phi_{0}\right)^{2}}{2 \sigma_{\phi}^{2}}\right]} \cdot e^{\left[-\frac{\left(r-r_{0}\right)^{2}}{2 \sigma_{r}^{2}}\right]} \cdot\left(1+e^{\beta\left(\left|\phi-\phi_{0}\right|-\phi_{\text {fex }}\right)}\right)^{-1}
\end{aligned}
$$

where $\phi$ denotes polar angle, $r$ is the radius, $\phi_{0}$ is the mean of a Gaussian function over angle, $\sigma_{\phi}$ is its standard deviation, $r_{0}$ and $\sigma_{r}$ are analogue parameters for a Gaussian over radius, $\beta$ is the steepness of a sigmoid function over angle, and $\phi_{\text {flex }}$ is the separation of its inflection point from the mean of the Gaussian over angle. We used $\sigma_{\phi}=1.05, r_{0}=0 \mathrm{~mm}, \sigma_{r}=47 \mathrm{~mm}, \beta=25$, and $\phi_{\text {flex }}=1.45$. Parameter $\phi_{0}$ depended on the spatial term, with "right of", "above", "left of", and "below" corresponding to $\phi_{0}=\left\{0, \frac{\pi}{2}, \pi, \frac{3}{2} \pi\right\} \mathrm{rad}$.


Fig. 1: (a) left panel: Spatial template for "right of". (a) center panel: General item regions defined by fit and distance constraints. White outline denotes outer radius of reference item. (a) right panel: One specific three-item configuration (target in red, distracter in green). The green dotted line here shows the region where distracters were placed for that specific target position. (b) Experimental screen area with regions and locations constraining stimulus placement; item arrangement for "left of". The red dot corresponds to the reference item, the closer green dot on the left corresponds to the target item, and the green dot below that to the (main) distracter. The green dot on the right is the opposite distracter. Gray dots are fillers. Black X's are potential target positions, the yellow diamond marks the center of mass of all items, the dotted square denotes the region generally eligible for item placement, and the dashed gray line illustrates the direct path to the target. The start marker is located at the bottom center (black dot). (c) The stimulus arrangement from (b) as viewed by the participants.
ter of the stimulus region, at a distance of $28.3 \mathrm{~mm}\left(2.32^{\circ}\right.$ v.a.) horizontally and vertically. Restricting target positions to a few fixed locations and having the stimulus array sample space around those locations alleviated the common problem of different movement metrics for different spatial locations.

Nine filler items were added to each trial, each colored randomly in one of the four remaining colors. Locations were restricted to a square region of 184 mm ( $15.1^{\circ}$ v.a.) side length, whose midpoint was 200.8 mm ( $16.6^{\circ}$ v.a.) straight above the start marker. The center of mass (CoM) across all 12 visual items had to be congruent with the center of that region ( $\pm 0.8 \mathrm{~mm}$ ) so that it was identical across conditions and positioned in the horizontal screen center, allowing to more easily partial out a putative bias to either one (a bias to the horizontal screen center was expected because the items appeared only as soon as upward movement was detected). Fillers retained a border-to-border distance of at least 0.5 mm . Locations were random otherwise. Finally, as an additional incentive to evaluate the spatial relation, in some trials ( $27 \%$ ) one filler was turned into an additional distracter by giving it the same color as target and (main) distracter. It had to be located on the side of the reference opposed to the spatial term, and separated from the reference along the term's axis (e.g., horizontal for "right of") by at least 28.3 mm ( $2.32^{\circ}$ v.a.).

The full stimulus set included 335 configurations $\times$ 4 spatial terms $\times 4$ target positions $=5360$ trials, which were randomly assigned to the twelve participants.

## Analysis

We analyzed only trials where participants selected the item best matching the spatial phrase according to the fit function (hereafter called target). We refer to the straight line from a trajectory's first point to the target item's center as direct path.

Sharply curved trajectories were discarded from analyses, in order to consider only trajectories exhibiting graded attrac-
tion while excluding re-decisions in mid-flight and mouseovershoots. Curvature was assessed by temporarily interpolating to a uniform segment length of 5 mm and then applying the osculating circle method (considering each vertex and its two neighbors). Trajectories exceeding a curvature of 0.1 were discarded. We chose this approach over other values such as area under curve (Hehman, Stolier, \& Freeman, 2015), as these are less informative in a setup with multiple potential effect sources on both sides of the direct path.

Trajectory preparation Trajectories were trimmed to start with movement onset and to end with the first data point after crossing the target border. They were then translated to start at $[0,0]$ and rotated around that point by the angle between the target's position vector and the $y$-axis. Positive $x$-values thus denote deviation from the direct path to the right, negative values indicate leftward deviation. Trajectories' spatial coordinates were linearly interpolated over 151 equidistant time steps to enable averaging (combining position data from identical values of elapsed proportion of total movement time).

Statistical analysis Mean trajectories were compared by testing for differences between x-coordinates at each of the 151 time steps using two-tailed paired-sample t-tests with $p<0.01$. Since data points in each mean trajectory are highly interdependent, and given the large number of tests, the informative value of each individual t -test is limited. To remedy this, we used the bootstrapping procedure introduced by Dale, Kehoe, and Spivey (2007), providing a criterion for how many t-tests in sequence must yield significance before a difference between trajectories can be considered overall significant. A separate criterion with $p<0.01$ was computed for each comparison based on 10,000 artificial experiments each.

Balancing to isolate main effects To obtain unbiased estimates of the individual effects of distracter, reference, and CoM position by comparing two conditions (e.g., all trials
where the distracter was left versus right of the direct path), the impact of the others (e.g., reference and CoM side) must be taken into account. For this, we distinguished trials into categories that indicated whether a potentially confounding item was on the same or opposite side of the direct path as the item of interest. There was a different set of four categories for each item of interest. For instance, balancing categories for the distracter effect were rs/cs, rs/co, ro/cs, and ro/co, " $r$ " denoting the reference, " $c$ " the CoM, and " $s$ " and " $o$ " denoting whether the respective item was on the same or opposite side of the direct path as the distracter. Corresponding categories for considering reference and CoM were named analogously ("d" denoting the distracter). When comparing two sets of trials for one effect, balancing out the other two then works by ensuring that each set is composed of an equal number of trials from each relevant balancing category. This requirement is not fully satisfied by our full set of trials and some comparisons. To allow judging how imbalances may have affected the respective data, Fig. 2 plots the distribution of trial numbers over the relevant balancing categories alongside each comparison. To further validate that observed effects were not due to imbalances, post hoc balancing was conducted: We did a second analysis for each comparison, identical to the one performed on the full trial set, but beforehand randomly discarded trials from over-represented categories such that a balanced distribution was obtained within each condition and participant. We report when this substantially changed effects. For the overall mean trajectory, categories were based on item sides relative to the direct path. Category labels used the same letters as above, in addition to "l" (left) and "r" (right; e.g., dr/rl/cl means that the distracter was on the right and the other items on the left side).

## Results

A total of 5245 trajectories was obtained ( 115 were lost due to technical problems) and participants selected the best-fitting item in 4710 (mean $89.82 \% \pm 3.3$ s.d.). Of these, 446 ( $9.47 \%$ ) exceeded curvature threshold, leaving 4264 (81.3\%) for analysis. Mean movement time was $1061 \mathrm{~ms}( \pm 116$ s.d.); noteworthy differences occurred only between upper and lower target positions ( $1140 \mathrm{~ms} \pm 119$ s.d. and 977 ms $\pm 114$ s.d.). Participants reported not to have noticed that target positions were limited to four locations.

Fig. 2 shows mean trajectories for all comparisons, along with trial distributions over balancing categories for each condition. The overall mean trajectory (Fig. 2a) slightly curved rightwards, likely reflecting kinematic bias. To provide an idea of this bias in relation to other effects, dotted gray lines in each panel of Fig. 2 show the mean over trials from the compared conditions. As expected based on the task instructions, a strong bias toward the CoM was evident (Fig. 2b).

We report statistical test results in this form: 46/8, 5-50\%, providing the number of successive significant time steps (46) along with the bootstrap criterion for overall significance (8), followed by the percentages of elapsed movement time at the
start and end of the sequence (5-50\%). Considering all trials, there was a significant bias away from the reference side in the first half of the movement (57/6, 1.3-38.4\%; Fig. 2c) and a significant bias in reference direction in the second half (52/6, 66.2-100\%). Assessing the effect of reference side separately for trials with horizontal-axis spatial terms ("left" and "right") and for vertical ones ("above" and "below") showed that the bias away from the reference was driven by the horizontal term trials (69/30, 1.3-46.4\%; Fig. 2d). Note that in these trials deviation away from the reference is congruent with spatial term direction. Correcting for the overrepresented distracter-opposite trials (cs/do, co/do) by post hoc balancing did not remove the effect (70/47, 1.3-47\%; Fig. 2f). The later bias toward the reference was driven by the vertical spatial term trials (94/31, 38.4-100\%; Fig. 2e). Post hoc balancing showed that it was not due to the overrepresented distracter-same trials and resulted in an earlier onset of the reference effect (102/31, 33.1-100\%; Fig. 2g).

As shown in Fig. 2h, there was a sustained, significant bias in distracter direction for the whole trial set (100/15, $33.4-100 \%$ ). Assessing the effect separately by spatial term axis showed that the effect's early component was driven exclusively by horizontal term trials (140/6, 8-100\%; vertical: 86/33, 43.7-100\%; Fig. 2i,j). These included a pronounced majority of reference-opposite trials (ro/cs, ro/co), suggesting that the distracter effect's early component may in fact be a bias in spatial term direction (i.e., away from the reference), as reported above. Post hoc balancing indeed reduced the distracter bias to the second half of the movement (81/8, $47-100 \%$; Fig. 2k). Post hoc balancing the vertical term trials left the effect largely unchanged (89/11, 41.7-100\%; Fig. 21).

Post hoc balanced vertical term trials (Fig. 2g) provide the most unbiased estimate of the reference effect. Comparing its onset in these trials to that of the distracter effect in the analogous comparison (Fig. 21) shows an earlier onset of the reference effect by $8.6 \%$ of movement time (equaling 91.1 ms , based on mean movement time in these trials).

## Discussion

We have described a mouse tracking paradigm in which unknown spatial targets were specified by the task through a relational description and demonstrated how influences from multiple effect sources in such a setup may be disentangled.

As predicted, distracters attracted mouse paths, similar to decision alternatives in classic mouse tracking studies (e.g., Dale et al., 2007). The predicted attraction toward reference items was observed as well, for the spatial terms "above" and "below". Moreover, as hypothesized, a bias in spatial term direction was present from early on for horizontal-axis spatial terms. We interpret this as a spatial term effect rather than repulsion from the reference based on the very early onset (note that the spatial term was not predictive of absolute target location in the paradigm) and in line with previous evidence (Tower-Richardi et al., 2012). Its apparent absence in vertical term trials is unsurprising, as it would act orthogonally to the


Fig. 2: Mean trajectories and comparisons between conditions. Trajectory plots show the distance from the direct path to the target $(x=0)$ against the proportion of elapsed movement time. Negative values indicate leftward deviation and vice versa. Conditions compared within a panel differ with respect to the position, relative to the direct path, of either the center of mass $(\mathrm{CoM})$, the reference item (Ref), or the distracter ( Dtr ). Blue and red circles in the top of each plot indicate the direction in which the item of interest was situated for the trajectory in the same color. Gray dotted lines represent the mean across all trajectories in the two compared conditions. Color-coded p-values and effect sizes (absolute Cohen's d) for each time step are shown on the left and right side of each plot. The black line on the left side of each plot indicates individual significant t-tests $(p<0.01)$ at the spanned time steps. Asterisks next to these lines indicate overall significance ( $p<0.01$ ) for a given sequence of individually significant time steps (based on bootstrap criterion; see Methods for details). A bar graph below each trajectory plot shows the distribution of trials over categories relevant for counter-balancing in the respective comparison; bars correspond to trajectory lines of the same hue. (a) Mean across all trials. (b) Means by CoM side, (c) by reference side, across all trials, (d) by reference side, only including trials with spatial terms "left" and "right", (e) by reference side, only including trials with spatial terms "above" and "below"; (f) and (g) show the same comparisons as (d) and (e), but with a post hoc balanced set of trials. (h) Means by distracter side, across all trials, (i) by distracter side, only including trials with spatial terms "left" and "right",(j) by distracter side, only including trials with spatial terms "above" and "below"; (k) and (l) show the same comparisons as (i) and (j), but with a post hoc balanced set of trials.
axis along which deviation was assessed. This may also explain why reference attraction is visible only in vertical term trials: If the spatial term effect impacts the entire length of trajectories as hypothesized, the two effects may cancel each other out in the late portion of horizontal term trials.

The attraction to the reference item confirms that it engages spatial attention during relational processing (e.g., Franconeri et al., 2012), even when it is unique in color. This may hint that spatial indexing (Logan \& Sadler, 1996) of its position is mandatory for grounding. It further suggests that computationally relevant non-targets can impact the motor level.

The observed distracter attraction is reminiscent of reaches to average locations under target uncertainty (e.g., Chapman et al., 2010). Aspects specific to relation grounding may as well play a role, though, for instance, through locations in a neural map being differentially activated by a relational template. A hint at this interpretation is the small extent of distracter attraction compared to a mean distracter distance to the direct path of 20.59 mm , calling into question mere averaging. The latter aspect, as well as the early spatial term effect, the mandatory reference selection, and the offset time courses of reference and distracter attraction, parallel our neural process model of grounding, in which item positions are sequentially stored in distinct neural substrates to apply a concurrently active, graded relational template (Richter et al., 2017).

There is ample room for new research in the direction suggested here. One step may be to clarify in how far distracter attraction is specific to relational processing. The temporal order of effects must be probed more formally. Finally, higher cognitive processes may further be unraveled through additional variations of spatial phrase structure or visual displays.

## Conclusion

As participants perceptually ground spatial phrases such as "the red to the left of the green", they attend to potential target objects (here, red ones) and typically select the one best matching the spatial relation. Mouse trajectories toward the ultimately selected target reveal transient biases in multiple directions. First, they show attraction to the alternative targets, consistent with previous evidence. Second, an attraction to the reference object ("green") begins somewhat earlier and may reflect allocation of spatial attention. Third, a bias in spatial term direction is present from early on. Overall, this study frames motor responses as direct reflections of the perceptual grounding of spatial phrases, bringing evidence for the coupling of cognitive to sensory-motor processes to a new level.

## References

Bastian, A., Schöner, G., \& Riehle, A. (2003). Preshaping and continuous evolution of motor cortical representations during movement preparation. Eur. J. Neurosci., 18, 2047-2058.
Brainard, D. H. (1997). The psychophysics toolbox. Spatial Vision, 10(4), 433-436.
Chapman, C. S., Gallivan, J. P., Wood, D. K., Milne, J. L., Culham, J. C., \& Goodale, M. A. (2010). Reaching
for the unknown: Multiple target encoding and realtime decision-making in a rapid reach task. Cognition, 116(2), 168-176.
Cisek, P., \& Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. Neuron, 3(45), 801-814.
Dale, R., Kehoe, C., \& Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. Mem. Cogn., 35(1), 15-28.
Erlhagen, W., \& Schöner, G. (2002). Dynamic field theory of movement preparation. Psychol. Rev., 109(3), 545572.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., \& Kahn, L. E. (2012, 02/2012). Flexible visual processing of spatial relationships. Cognition, 122, 210227.

Freeman, J. B., Dale, R., \& Farmer, T. A. (2011). Hand in motion reveals mind in motion. Front. Psychol., 2, 1-6.
Ghez, C., Favilla, M., Ghilardi, M. F., Gordon, J., Bermejo, R., \& Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. Exp. Brain Res., 115, 217-233.
Hehman, E., Stolier, R. M., \& Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. Group Process. Intergr. Relat., 18(3), 384-401.
Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., \& Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. J Exp Psychol Learn, 38(6), 1490-1511.
Logan, G. D., \& Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, \& L. Nadel (Eds.), Language and Space (pp. 493-529). Cambridge, USA: MIT Press.
Richardson, D. C., Spivey, M. J., Barsalou, L. W., \& McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. Cogn. Sci., 27(5), 767-780.
Richter, M., Lins, J., \& Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. Top Cogn Sci, 9(1), 35-47.
Schöner, G., Spencer, J. P., \& the DFT Research Group. (2015). Dynamic thinking: A primer on dynamic field theory. Oxford University Press.
Song, J., \& Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. Trends Cogn Sci., 13(8), 360-366.
Spivey, M. J. (2007). The continuity of mind. Oxford, UK: Oxford University Press.
Tower-Richardi, S., Brunye, T., Gagnon, S., Mahoney, C., \& Taylor, H. (2012). Abstract spatial concept priming dynamically influences real-world actions. Front. Psychol., 3, 361.

# Prediction and uncertainty in an artificial language 

Tal Linzen<br>LSCP \& IJN, CNRS, ENS, PSL Research<br>University and Johns Hopkins University<br>tal.linzen@gmail.com

Noam Siegelman<br>Department of Psychology<br>Hebrew University of Jerusalem<br>noam.siegelman@gmail.com

Louisa Bogaerts<br>Department of Psychology<br>Hebrew University of Jerusalem<br>bog.louisa@gmail.com


#### Abstract

Probabilistic prediction is a central process in language comprehension. Properties of probability distributions over predictions are often difficult to study in natural language. To obtain precise control over these distributions, we created artificial languages consisting of sequences of shapes. The languages were constructed to vary the uncertainty of the probability distribution over predictions as well as the probability of the predicted item. Participants were exposed to the languages in a self-paced presentation paradigm, which provides a measure of processing difficulty at each element of a sequence. There was a robust pattern of graded predictability: shapes were processed faster the more predictable they were, as in natural language. Processing times were also affected by the uncertainty (entropy) over predictions at the point at which those predictions were made; this effect was less consistent, however.


Keywords: Entropy, prediction, statistical learning, artificial language, psycholinguistics

## Introduction

Our environment is characterized by recurring temporal patterns; the sound of an ambulance siren, for example, tends to predict the appearance of an ambulance. Humans can quickly learn to exploit these contingencies between stimuli to anticipate future events and react to those events more effectively. The ability to track dependencies across the elements of a sequence is central to language processing: prediction of upcoming words is employed during language comprehension (DeLong, Urbach, \& Kutas, 2005) and may play a central role in acquisition (Gómez, 2002).

Prediction in natural language is rarely categorical: there is generally some uncertainty as to the upcoming word. Rather than predict a single word or avoid making predictions altogether, readers maintain a probability distribution over the upcoming words: words that are more likely to come up are activated to a greater extent (Smith \& Levy, 2013). Probability distributions over predictions are often difficult to study in natural language, due to the need to find sets of words that happen to have the desired probabilistic relations in a natural corpus. The present study builds on work that shows that the processing of temporal contingencies can be studied using artificial language learning experiments. These experiments typically consist of a familiarization phase, in which participants are exposed to the artificial language, and a test phase, in which they are requested to distinguish sequences that follow the patterns of the language from sequences that do not. We use this paradigm to study probabilistic prediction in sequence learning and processing.

Quantifying probabilistic prediction: A predictive dependency is made up of two parts: the point at which the
prediction is generated (the predictive item) and the point at which it is matched against the incoming input (the predicted item). We study both parts of the process. At the predictive item, multiple probabilistic predictions can typically be generated. Higher uncertainty over the correct prediction may lead to increased competition among those predictions and slower processing. We follow earlier work in quantifying uncertainty using the entropy of the distribution over possible predicted items (Linzen \& Jaeger, 2014; Hasson, 2017):

$$
\begin{equation*}
H=-\sum_{w \in W} P(w) \log _{2} P(w) \tag{1}
\end{equation*}
$$

where $W$ is the set of possible items and $P(w)$ is the probability of $w$ in the current context. At the point at which predictions are matched against the input, input items that were predicted with a higher probability may be processed more quickly. In natural language, processing difficulty at an item $w$ is proportional to its surprisal $\left(-\log _{2} P(w)\right)$ : more surprising words tend to be read more slowly (Smith \& Levy, 2013). A final expectation-based measure that has been argued to be a reliable predictor of reading times (RTs) in natural language is uncertainty reduction: words that reduce uncertainty about the sequence to a greater extent are predicted to be read more slowly (Hale, 2003; Frank, 2013).

The experiments: We report two experiments designed to examine these quantitative measures of probabilistic prediction in artificial languages. In what follows, we briefly discuss our general methodological strategy.

In many artificial language learning experiments, the familiarization phase consists of passive exposure; as such, the only behavioral measure collected in these studies is the proportion of correct grammaticality judgments given after the familiarization stage is over. Recently, a number of online paradigms have been proposed that track the learning process as it unfolds over the course of the familiarization phase (Siegelman, Bogaerts, Christiansen, \& Frost, 2017). Online paradigms also provide an index of processing time at each individual item of the sequence, making them particularly well-suited to studying the generation and validation of predictions. We adopt one of these paradigms, the selfpaced reading paradigm (Just, Carpenter, \& Woolley, 1982), adapted to artificial language learning by Karuza, Farmer, Fine, Smith, and Jaeger (2014). In this paradigm, the elements of each sequence are presented sequentially; the participant controls when the next sequence element is revealed.

Previous studies of prediction have focused on sequences with nonadjacent dependencies: the sequence is of the form
$A X B$, where $A$ predicts $B$ (Karuza et al., 2014; Misyak, Christiansen, \& Bruce Tomblin, 2010). Instead, we use dependencies of the form $A B$, without an intervening element; such dependencies are in general easier to learn (Newport \& Aslin, 2004). By increasing the likelihood that our participants will learn the language, we can ask more fine-grained questions than would be possible using nonadjacent dependencies.

Summary of goals: We address the following issues:

1. Does the probability of the second shape $B$ given the first shape $A$ affect the processing of $B$ ?
2. Are processing times at the point where the prediction can be made (shape $A$ ) affected by the uncertainty of the probability distribution over predictions?
3. Does the reduction of uncertainty about the sequence at shape $B$ entail greater processing difficulty?

## Experiment 1

## Stimuli

Following Karuza et al. (2014), we used sequences of letters from the Ge'ez script, which is used to write several Ethiopian and Eritrean languages. We refer to these letters as shapes since none of our participants were familiar with this script. As in Karuza et al. (2014), our sequences consisted of three shapes. As we have mentioned, we omitted the intermediate shape - the dependency was adjacent. To keep the structure of the stimuli similar to the stimuli used in the previous study and to avoid task effects related to the beginning of a new sequence, all of the sequences started with a fixed shape $r$ (distinct from the $A$ and $B$ shapes). This shape was the same in all trials for a given participant and was not analyzed.

The language used in this experiment is described in Table 1 . Each participant was exposed to two types of $A$ shapes. Low entropy $A$ shapes were followed by one of two $B$ shapes, with probability $1 / 4$ and $3 / 4$ respectively. High entropy $A$ shapes were followed by one of four $B$ shapes, each with probability $1 / 4$. There were two $A$ shapes of each type, for a total of four $A$ shapes. None of the $B$ shapes were repeated across $A$ shapes: there were 12 distinct $B$ shapes. We refer to the $B$ shapes with a probability of $1 / 4$ as high surprisal shapes, and to shapes with a probability of $3 / 4$ as low surprisal shapes.

To control for potential differences in the visual complexity of particular shapes, the shapes that served as $a_{1}, a_{2}$ and $b_{1}, \ldots, b_{6}$ were counterbalanced across participants.

## Participants

A total of 44 participants ( 24 women and 20 men; age range: 20-28, mean age: 23.4) from the Hebrew University of Jerusalem community completed the experiments.

## Procedure

The experiment consisted of three phases: familiarization, test and a post-test phase.

| Shape $R$ | Shape $A$ | Shape $B$ | TP | Surprisal |
| :--- | :---: | :---: | :---: | ---: |
| High entropy: $(H=2)$ |  |  |  |  |
| $r$ | $a_{1}$ | $b_{1}$ | $1 / 4$ | 2 |
| $r$ | $a_{1}$ | $b_{2}$ | $1 / 4$ | 2 |
| $r$ | $a_{1}$ | $b_{3}$ | $1 / 4$ | 2 |
| $r$ | $a_{1}$ | $b_{4}$ | $1 / 4$ | 2 |
| Low entropy: $(H=0.81)$ |  |  |  |  |
| $r$ | $a_{2}$ | $b_{5}$ | $1 / 4$ | 2 |
| $r$ | $a_{2}$ | $b_{6}$ | $3 / 4$ | 0.41 |

Table 1: Half of the language used in Experiment 1 (the other half is duplicated: a high entropy $a_{3}$ paired with high surprisal $b_{7}$ through $b_{10}$ and a low entropy $a_{4}$ paired with a high surprisal $b_{11}$ and a low surprisal $b_{12}$ ). TP indicates the transitional probability between the $A$ and the $B$ shape (e.g., $\left.P\left(b_{1} \mid a_{1}\right)=1 / 4\right) . H$ indicates the entropy of each distribution.

Familiarization phase: Each trial started with a sequence of dashes where the shapes would be; participants pressed the spacebar to reveal the next shapes one by one. When a shape was revealed, the previous shape was replaced by a dash again. Before this phase of the experiment began, participants were instructed to try to remember the sequences, since they would be tested on them later on.

There were 288 sequences in this phase. This contrasts with the familiarization phase in Karuza et al. (2014), which consisted of 432 sequences; we chose to have a shorter familiarization phase because prediction effects in Karuza et al. (2014) plateaued about half way through the experiment. We further simplified their design by eliminating the catch trials meant to ensure that participants were paying attention. These trials were not necessary because we analyzed data only from participants who successfully learned the language: our assumption was that participants who were not paying attention would fail to learn the language.

Test phase: This phase consisted of 24 trials, each of which elicited a judgment for one sequence. All three shapes of the sequence were presented at once (not in self-paced presentation). Half of the trials contained sequences that had been presented during familiarization; the other half contained the shapes from the familiarization phase arranged in unseen sequences. Participants were asked to press one button if the sequence appeared familiar given the sequences they had seen in the first phase, and another button if it did not.

Post-test phase: The test phase was followed by another self-paced presentation phase. This phase was somewhat shorter, consisting of 192 trials. Participants were again instructed to attempt to remember the shapes. The goal of this phase was to examine the behavior of participants who have already learned the language; for example, if predictability effects were found, are they restricted to the stages in which the participant has not yet mastered the language?


Figure 1: Reading times in Experiment 1 (above: $A$ shape; below: $B$ shape).

## Results

Accuracy: We briefly analyze the familiarity judgments from the test phase before moving on to the analysis of the processing time data from the familiarization phase, which is the focus of this study. On average, participants were more likely to judge a sequence as grammatical, leading to higher accuracy on grammatical than ungrammatical sequences ( $82 \%$ vs. $60 \%$ ). To test for differences across types of test sequences, we coded the test sequences based on the category of their $A$ and $B$ shapes (e.g., low entropy + high surprisal). Logistic mixed-effects models fitted separately to grammatical and ungrammatical sequences did not find significant differences across sequence types (grammatical: $\chi^{2}(2)=2.8, p=.24$; ungrammatical: $\left.\chi^{2}(2)=4.4, p=.11\right)$.

RT preprocessing and analysis: We refer to the sequential processing times measured by key press latencies as reading times (RTs) for consistency with the sentence processing literature. Following Karuza et al. (2014), we excluded shapes on which RTs were (1) longer than six seconds or (2) three standard deviations higher or lower than the participant's mean RT for shapes in the same position. This resulted in the exclusion of $1.1 \%$ and $2.5 \%$ of the shapes respectively.

We only analyzed RTs from participants who gave correct grammaticality judgments to at least 18 of the 24 sequences
(the lowest number for which $p<.05$ according to an exact binomial test). Of the 44 participants, 23 passed this threshold. Our statistical analysis largely followed Karuza et al. (2014). RTs were log-transformed and submitted to a linear mixed-effects regression with a random intercept for shape and a random intercept and slope for all fixed effects. Trial number and its interaction with the experimental factors were included in all models.

RT results: The time course of the results is plotted in Figure 1. Overall, RTs decreased markedly over the course of the familiarization phase, picked up in the beginning of the post-test phase, then decreased again.

The average difference in RTs between high and low entropy $A$ shapes in the familiarization phase was 119 ms ( 877 ms for high entropy and 758 ms for low entropy shapes). The linear mixed-effects model analysis indicated that this difference was statistically significant $\left(\chi^{2}(1)=4.2, p=.04\right)$. The effect of trial number was highly significant $\left(\chi^{2}(1)=31.5\right.$, $p<.001$ ). The interaction between trial number and entropy did not reach significance $\left(\chi^{2}(1)=.06, p=.81\right)$, suggesting that there was no clear evidence that the effect of entropy changed over the course of the experiment.

There were three types of $B$ shapes: high surprisal ones that followed a low entropy $A$ shape (e.g., $b_{5}$, see Table 1); high surprisal ones that followed a high entropy $A$ shape (e.g., $b_{1}$ ); and low surprisal ones that followed a low entropy $A$ shape (e.g., $b_{6}$ ). We first examined the effect of surprisal, collapsing across the two categories of high surprisal shapes. We found that high surprisal shapes were read more slowly than low surprisal shapes $\left(\chi^{2}(1)=17.8, p<.001\right)$; the average difference in RT was 200 ms ( 812 ms for high surprisal and 612 ms for low surprisal shapes). The effect of trial number was highly significant again $\left(\chi^{2}(1)=44.6, p<.001\right)$, and interacted with surprisal such that the effect of surprisal weakened over the course of the familiarization phase $\left(\chi^{2}(1)=9.9\right.$, $p=.002$ ).

Finally, we compared the two types of high-surprisal $B$ shapes, which were matched for surprisal but differed in the entropy of the $A$ shape that preceded them. The mean RTs were almost identical across these two types of shapes (812 ms after high entropy $A$ shapes and 813 ms after low entropy ones). This difference was not significant in the statistical analysis (main effect of entropy: $\chi^{2}(1)=1.2, p=.27$; interaction with trial number: $\left.\chi^{2}(1)=.9, p=.35\right)$.

## Discussion

In this experiment, participants were taught a language designed to assess the effect of measures of probabilistic prediction on sequence processing. Neither surprisal nor uncertainty reliably affected judgment accuracy in the test phase; they did, however, modulate processing times during the familiarization phase. First, predictability affected RTs in the expected way: high surprisal $B$ shapes were read more slowly than low surprisal ones. Second, uncertainty at the $A$ shape affected RTs in a way that is consistent with competition among
the predictions: higher entropy shapes were read more slowly than low entropy ones.

Finally, we did not find evidence for an effect of uncertainty reduction on the $B$ shape. To see why, note that the $B$ shapes are the last item in the sequence; as such, they reduce the uncertainty about the sequence to 0 . The amount by which uncertainty is reduced is therefore equal to the entropy of the distribution over predictions at the $A$ shape; yet there was no evidence for a difference in reading times between high surprisal $B$ shapes that followed a high entropy $A$ shape (and therefore reduced entropy by 2 bits) and high surprisal $B$ shapes that followed a low entropy $A$ shape (and reduced entropy by only 0.41 bits).

## Experiment 2

In Experiment 1, uncertainty was perfectly correlated with the number of possible predictions: high entropy $A$ shapes had four prediction options compared to two options in low entropy $A$ shapes. The goal of the current experiment is to examine whether we can find entropy effects when the number of options is kept constant. For a given number of options, entropy is highest when the distribution is uniform; we therefore compare a uniform distribution to a skewed one, that is, with one option that is more likely than the others.

## Participants

A total of 49 participants completed the experiment. Two participants were excluded for not completing the experiment and one for having prior exposure to Amharic, which uses the Ge'ez script; of the remaining participants, 35 were women and 11 men (age range: 19-31; mean age: 23.8).

## Materials

The language used in Experiment 2 is shown in Table 2. There were three types of $A$ shapes. Two of the $A$ shapes could be followed by three possible $B$ shapes (to avoid having to teaching participants a very low probability option, we used three options instead of four as in Experiment 1.) After $a_{1}$, the distribution of the $B$ shapes was uniform: each of the shapes had a probability of $1 / 3$. After $a_{2}$ the distribution was skewed: one of the shapes had a probability of $2 / 3$ and the other two $1 / 6$ each.

To control for the possibility that any difference between the two 3-option shapes could reflect skew rather than entropy as such, we additionally included a third type of $A$ shape that was followed by one of two $B$ shapes, each with probability $1 / 2$. As this distribution is uniform, we expect this shape to pattern with $a_{1}$ if the relevant factor is skew. Conversely, since its entropy is lower than either 3-option shapes, it should be processed faster than either of them if entropy is the relevant factor.

Due to the larger number of conditions and the need to provide sufficient exposure to lower probability $B$ shapes $(1 / 6$ compared to $1 / 4$ in Experiment 1), each type of $A$ shape was represented by a single shape only.

| Shape 1 | Shape 2 | Shape 3 | TP | Surprisal |
| :--- | :---: | :---: | :---: | :---: |
| Three options, uniform: $(H=1.58)$ |  |  |  |  |
| $r$ | $a_{1}$ | $b_{1}$ | $2 / 6$ | 1.58 |
| $r$ | $a_{1}$ | $b_{2}$ | $2 / 6$ | 1.58 |
| $r$ | $a_{1}$ | $b_{3}$ | $2 / 6$ | 1.58 |
| Skewed, three options: $(H=1.25)$ |  |  |  |  |
| $r$ | $a_{2}$ | $b_{4}$ | $4 / 6$ | 0.58 |
| $r$ | $a_{2}$ | $b_{5}$ | $1 / 6$ | 2.58 |
| $r$ | $a_{2}$ | $b_{6}$ | $1 / 6$ | 2.58 |
| Uniform, two options: $(H=1)$ |  |  |  |  |
| $r$ | $a_{3}$ | $b_{7}$ | $3 / 6$ |  |
| $r$ | $a_{3}$ | $b_{8}$ | $3 / 6$ | 1 |

Table 2: Language used in Experiment 2. $H$ indicates the entropy of each distribution.

## Procedure

The structure of the experiment was the same as in Experiment 1. The familiarization self-paced presentation phase consisted of 324 sequences. This phase was followed by 16 familiarity judgments, and an additional post-test self-paced presentation phase with 216 sequences.

## Results

Accuracy: Overall accuracy was higher than in Experiment 1, though the bias for marking sequences as grammatical remained: $93 \%$ of the grammatical sequences and of $77 \%$ of the ungrammatical sequences were identified correctly. We tested for an effect of the four types of sequences (see Table 2) on accuracy rates on grammatical sequences. A logistic mixed-effects model did not reveal an effect of sequence type $\left(\chi^{2}(3)=4.5, p=.21\right)$. Likewise, there was no effect of either $A$ or $B$ shape type on accuracy rates in ungrammatical sentences $\left(A: \chi^{2}(2)=2.65, p=.27 ; B: \chi^{2}(3)=3.2, p=.36\right)$.

RT preprocessing and analysis: As before, we restricted our analysis to participants who showed evidence of learning the language, defined as giving correct judgments more often than chance ( $p<.05$ according to the binomial test); this translates to performing at least 13 of the 16 trials correctly. Of the 46 participants, 33 passed this threshold. We excluded key presses with extreme RTs using the same criteria as before, resulting in the exclusion of $3.38 \%$ of the shapes. Analysis methods were in general identical to Experiment 1, with the exception that our mixed-effects models did not include a random intercept for shape in cases where there was only one shape in each condition (i.e., in the analysis of $A$ shapes).

RT results: The qualitative pattern of results was similar to Experiment 1: RTs globally decreased over the course of the familiarization phase, briefly increased in the post-test phase, then decreased again.


Figure 2: Condition means in the familiarization phase of Experiment 2 (above: $A$ shape; below: $B$ shape). Error bars represent $95 \%$ within-subject confidence intervals.

We first discuss the statistical analysis of familiarization phase RTs on $A$ shapes, starting with an analysis of entropy as a numerical predictor There was a main effect of entropy $\left(\chi^{2}(1)=5.1, p=.02\right)$, a main effect of trial num$\operatorname{ber}\left(\chi^{2}(1)=40.3, p<.001\right)$ and a nonsignificant interaction $\left(\chi^{2}(1)=3.3, p=.07\right)$. RTs in the individual conditions were longest on the uniform 3-option shape and shortest on the skewed 3-option shape; although the entropy of the 2-option shape was lowest of all three shapes, average reading times on this shape were somewhat higher than the skewed 3-option shape (see Figure 2). The difference in RTs between the two 3 -option shapes was significant $\left(\chi^{2}(1)=5.1, p=.02\right.$ ), but the interaction with trial number was not $\left(\chi^{2}(1)=2.3, p=.13\right)$. The difference between the two shapes with a uniform prediction distribution (3-option vs. 2-option) and the interaction between this difference and trial number did not reach significance (main effect: $\chi^{2}(1)=3.5, p=.06$; interaction: $\chi^{2}(1)=3.1, p=.08$ ), and neither did the difference between the skewed 3-option and uniform 2-option shapes (main effect: $\chi^{2}(1)=.5, p=.48$; interaction: $\chi^{2}(1)=.03, p=.87$ ).

We next discuss the $B$ shapes. Again, we first enter surprisal as a numerical predictor. The statistical analysis found a highly significant effect of this predictor $\left(\chi^{2}(1)=36.2\right.$, $p<.001$ ) and of trial number $\left(\chi^{2}(1)=75.3, p<.001\right)$, as well as an interaction between the two $\left(\chi^{2}(1)=20.3\right.$, $p<.001$ ). Inspection of the average RTs for each level of surprisal (see Figure 2) suggests that not all differences between
consecutive levels of surprisal are equally large; in fact, only the difference between the $p=2 / 6$ and $p=3 / 6$ shapes was statistically significant $\left(\chi^{2}(1)=21.3, p<.001\right)$.

## Discussion

RTs on the two 3-option $A$ shapes were consistent with the hypothesis that higher uncertainty leads to longer RTs. The difference was smaller than in Experiment 1 (around 60 ms ), though that is to be expected given the smaller difference in entropy between the two shapes in the current experiment. The same hypothesis, however, predicts that RTs on the 2option shape should be lower than either 3-option shape; there was no evidence for such an effect.

There was a strong effect of surprisal overall, but there was often no evidence for differences between consecutive levels of surprisal. The difference between the two $B$ shapes that followed the 3-option skewed $A$ shape was particularly large. Finally, since no two $B$ shapes were matched on predictability and at the same time differed in the entropy of the $A$ shape that predicted them, the design of Experiment 2 did not allow us to test for an effect of uncertainty reduction.

## General Discussion

Probabilistic prediction plays a central role in language processing: a predictive item sets up expectations for predicted items later in the sequence. We studied the reflexes of probabilistic prediction in two artificial languages, which allowed us to exert precise control over the distribution over predictions. We used self-paced presentation (Just et al., 1982; Karuza et al., 2014), which yields implicit measures of processing at every element of the sequence. Two experiments revealed graded predictability effects parallel to those found in natural language. They also suggested that higher uncertainty over predictions at the point where predictions are generated leads to longer processing times, although these effects were weaker. No clear support was found for an effect of uncertainty reduction, even when controlling for predictability.

To further investigate the results, we pooled the data from both experiments and plotted the mean RTs in the familiarization phase by numerical entropy and surprisal in Figure 3 (since Experiment 2 was slightly longer, we discarded the trials following the first 288 trials for the purpose of this analysis). The evidence for a linear effect across experiments of the numerical predictors appears stronger for surprisal than for entropy. In particular, there are no clear differences among low-entropy distributions (lower than 1.5), and the slight differences that do exist are in the opposite direction than predicted by a linear relationship between entropy and RTs. Statistical models including data from both experiments did not reveal overall entropy effects at the $A$ shape (entropy: $\chi^{2}(1)=0.4, p=.53$; trial number: $\chi^{2}(1)=63$, $p<.001$; interaction: $\chi^{2}(1)=2, p=.16$ ), but did reveal clear surprisal effects at the $B$ shape as well as an interaction with trial number (surprisal: $\chi^{2}(1)=25.5, p<.001$; trial number: $\chi^{2}(1)=104.1, p<.001$; interaction: $\chi^{2}(1)=21.7$, $p<.001$ ).


Figure 3: Comparison across the experiments: means of the first 288 trials of the familiarization phase (above: A shape; below: $B$ shape). Error bars represent within-subject confidence $95 \%$ confidence intervals based on two standard deviations from the mean.

While any conclusion from pooling together two experiments with a different design and a different set of subjects should be taken as tentative, the nonlinear relationship between entropy and processing times suggests that entropy may not be the best metric for difficulty in prediction generation; additional properties of the distribution over predictions, such as the number of options or the probability of the most likely option, may need to be taken into consideration.

Figure 1 suggests that RTs in Experiment 1 may have reached a plateau about 250 trials into the familiarization phase; differences among conditions appeared to grow increasingly small around this time (Karuza et al. (2014) report a similar pattern). RTs increased at the beginning of the posttest phase, and then plateaued again around 100 trials into the pre-test phase. We did not present an in-depth analysis of the post-test phase for reasons of space; however, the fact that the overall increase in RTs at the beginning of the post-test phase was accompanied by a re-emergence of predictability and entropy effects suggests that the convergence between the conditions at the end of the familiarization phase is due to a floor effect rather than due to participants abandoning predictive processes once the language has been learned.

We made relatively few modifications to the methodology developed by Karuza et al. (2014), with the goal of building on their established paradigm. This entailed in particular that
our sequences were made up of visual symbols rather than auditory or written words; none of the symbols had any semantic content. The encouraging results of the present study suggest that this method may be extended to richer artificial languages that are a closer approximation of natural languages.

Acknowledgments: We thank Lizz Karuza for sharing her materials. This research was supported by the European Research Council (grant ERC-2011-AdG 295810 BOOTPHON) and the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC).

## References

DeLong, K., Urbach, T., \& Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience, 8(8), 1117-1121.
Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. Topics in Cognitive Science, 5(3), 475-494.
Gómez, R. L. (2002). Variability and detection of invariant structure. Psychological Science, 13(5), 431-436.
Hale, J. (2003). The information conveyed by words in sentences. Journal of Psycholinguistic Research, 32(2), 101123.

Hasson, U. (2017). The neurobiology of uncertainty: implications for statistical learning. Philosophical Transactions of the Royal Society B, 372(1711), 20160048.
Just, M. A., Carpenter, P. A., \& Woolley, J. D. (1982). Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General, 111(2), 228238.

Karuza, E. A., Farmer, T. A., Fine, A. B., Smith, F. X., \& Jaeger, T. F. (2014). On-line measures of prediction in a self-paced statistical learning task. In Proceedings of the 36th Annual Meeting of the Cognitive Science Society (pp. 725-730).
Linzen, T., \& Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. In Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics (pp. 10-18).
Misyak, J. B., Christiansen, M. H., \& Bruce Tomblin, J. (2010). Sequential expectations: The role of predictionbased learning in language. Topics in Cognitive Science, 2(1), 138-153.
Newport, E. L., \& Aslin, R. N. (2004). Learning at a distance I: Statistical learning of non-adjacent dependencies. Cognitive Psychology, 48(2), 127-162.
Siegelman, N., Bogaerts, L., Christiansen, M. H., \& Frost, R. (2017). Towards a theory of individual differences in statistical learning. Philosphical Transactions of the Royal Society B, 372(1711), 20160059.
Smith, N. J., \& Levy, R. (2013). The effect of word predictability on reading time is logarithmic. Cognition, 128(3), 302-319.

# Explain, Explore, Exploit: Effects of Explanation on Information Search 

Emily G. Liquin (EmilyLiquin@Berkeley.Edu)<br>Tania Lombrozo (Lombrozo@Berkeley.Edu)<br>Department of Psychology, 3210 Tolman Hall<br>Berkeley, CA 94720 USA


#### Abstract

How does actively seeking explanations for one's observations affect information search over the course of learning? Generating explanations could plausibly lead learners to take advantage of the information they have already obtained, resulting in less exploration. Alternatively, explaining could lead learners to explore more, especially after encountering evidence that suggests their current beliefs are incorrect. In two experiments using a modified observe or bet task, we investigate these possibilities and find support for the latter: participants who are prompted to explain their observations in the course of learning tend to explore more, especially after encountering evidence that challenges a current belief.


Keywords: explanation; exploration; learning; decision making

In the decades leading up to his publication of On the Origin of Species, Charles Darwin recorded the titles of 687 books of English non-fiction that he read. According to analyses by Murdock, Allen, and DeDeo (2017), Darwin's reading fell into three epochs, each defined by a certain pattern of exploration, or broad search across new topic areas, and exploitation, or extended examination of texts within a similar topic area. Darwin's example raises questions about the relationship between explanation and information search. In searching for an explanation (in Darwin's case, a scientific explanation for the diversity of living things), do people pursue evidence broadly (i.e., exploring), or restrict their search to align with their current beliefs (i.e., exploiting)? Do these tendencies shift over time as new evidence is acquired? And if so, how?

Lombrozo and colleagues have proposed that when engaged in explanation, children and adults recruit explanatory considerations as evaluative constraints, rendering them more likely to generate and favor hypotheses that support "good" explanations - namely those that are simple, broad, and exhibit other explanatory virtues (Lombrozo, 2016; Williams \& Lombrozo, 2010, 2013). There is also evidence that the hypotheses one generates and considers influence information search (Bonawitz, van Schijndel, Friel, \& Schulz, 2012). Combining these proposals thus predicts that patterns of information search could be affected by engaging in the process of explanation.

To date, few studies have investigated the relationship between explanation and information search. In one study, Legare (2012) found that children's explanations for an unexpected piece of evidence predicted their subsequent exploratory behavior. In more recent work, Ruggeri, Lombrozo, and Xu (in prep) found that prompting children
to explain relationships in a target domain prepared them to ask more efficient questions on a subsequent 20 -questions task in that domain. Neither study, however, was designed to test the causal influence of generating explanations on decisions to explore in a dynamic learning task, nor were they designed to examine adults' exploratory behavior.

In two experiments, we investigate how explanation generation affects patterns of exploration by prompting adult learners to explain as they search for information over the course of a category learning task. To accomplish this, we draw on research from the reinforcement learning literature on the explore-exploit dilemma (Cohen, McClure, \& Yu, 2007; Kaelbling, Littman, \& Moore, 1996). As defined in this literature, exploration involves seeking new information, while exploitation involves seeking reward (by taking advantage of the information one has already acquired). For example, in the observe or bet task (Navarro, Newell, \& Schulze, 2016; Tversky \& Edwards, 1966), agents must choose between "observing" which of two bulbs lights up on a given trial (without receiving any reward) or "betting" which bulb they think will light up for the chance to earn a reward (without observing that trial's result). Each bulb lights up with some fixed probability that the learner must infer through a period of observation. In this task, observation is equated with exploration (i.e., information seeking with no potential for reward) and betting is equated with exploitation (i.e., reward seeking with no potential for information).

## The Present Research

In the present research, we propose a new method, the contextual observe or bet task (inspired by the contextual multi-armed bandit task; Langford \& Zhang, 2008). In this task, a set of "context variables" (i.e., features of the options that vary across trials) can be used to predict the option that will provide a reward on each trial. Successful performance depends on learning to identify and use these variables. This method integrates a more complex, real-world learning task into an active, dynamic learning environment. We can answer the question "when do explainers choose to explore?" by measuring when learners choose to observe rather than bet.

To develop a contextual observe or bet task well suited to exploring the effects of explanation on information search, we adapt the stimuli from Williams and Lombrozo (2010). In a set of three studies, Williams and Lombrozo presented learners with four exemplars from each of two novel categories. Category members could be classified by a


Figure 1: a. Typical trial (Expts. 1 and 2): Both robots can be classified by the $100 \%$ rule (foot shape) and the $75 \%$ rule (body/head shape); b. Non-obvious anomaly trial (Expts. 1 and 2): Both robots can be classified by the $100 \%$ rule but not the $75 \%$ rule; c. Obvious anomaly trial (Expt. 2): Both robots can be classified by the $100 \%$ rule, but only one robot can be classified by the $75 \%$ rule. Category labels were only displayed if a participant chose to observe and are included here for clarity.
salient rule that accounted for $75 \%$ of exemplars or a subtle rule that accounted for $100 \%$ of exemplars. Participants who were asked to explain the category membership of each exemplar were more likely to discover the $100 \%$ rule than participants who engaged in a control task.

For our contextual observe or bet task, we present learners with pairs of category exemplars over a number of trials. On each trial, learners can choose to "observe" the category labels of the exemplars or "bet" which exemplar they believe belongs to a given category. Learners are thus free to determine when to seek information (exploration/ observation) and when to seek reward (exploitation/betting) as they learn the features that predict category membership (context variables).

Prior work motivates two hypotheses regarding the effects of explanation generation on explore-exploit decision making. By Hypothesis 1, explaining could lead learners to greater exploitation. Previous research suggests that people use the first explanation they receive as a benchmark by which to judge subsequent explanations (Ihme \& Wittwer, 2015) and use their current explanation to decide between competing hypotheses for new data (Johnson \& Krems, 2001). Learners may thus prefer the first explanation they generate. This tendency towards accepting the first explanation in a series could lead people to switch to exploitation after arriving at an initial explanation, even if it is based on scant evidence. We suggest that learners may thus be more willing to quickly settle on a hypothesis that aligns with their initial beliefs based on the first pieces of information gathered, leading to increased exploitation.

By Hypothesis 2, explaining could lead learners to greater exploration. This hypothesis is consistent with one interpretation of the findings from Williams and Lombrozo (2010): when prompted to explain, participants continued to "search" the stimuli until they found a good explanation, rather than settling for the salient but imperfect $75 \%$ rule. Relatedly, Williams, Lombrozo, and Rehder (2013) found that explainers seemed to perseverate in looking for a perfect classification rule, even when none was available. If explainers explore until they find a good explanation, then evidence that a candidate explanation is inadequate could be a critical cue that leads explainers to engage in further exploration. Indeed, Williams and Lombrozo (2010) found that explaining anomalies (i.e., exceptions to the $75 \%$ rule) was particularly powerful in encouraging learners to reject
an imperfect rule and discover a better alternative (see also Williams, Walker, \& Lombrozo, 2012). However, this finding was not experimentally linked to an increase in exploration or information search, which makes it possible that anomalous evidence influenced discovery via other mechanisms. It is thus an open question whether explanation has a causal impact on exploration, and if so whether this impact is most pronounced when the evidence that is being explained contradicts one's current beliefs.

For our contextual observe-or-bet task, Hypothesis 1 thus predicts that relative to control participants, those who are prompted to explain will make more "bet" choices. In contrast, Hypothesis 2 predicts that relative to control participants, those prompted to explain will be more likely to observe, especially on trials following the observation of information that is anomalous with respect to initial beliefs, which we expect to correspond to the obvious rule that accounts for $75 \%$ of exemplars. In two experiments, we test these hypotheses.

## Experiment 1

## Method

Participants Participants for both experiments were recruited from Amazon Mechanical Turk and paid $\$ 0.85$ for participating in the 8.5 -minute study. Participation in both experiments was restricted to users in the United States with a $95 \%$ or higher approval rating based on at least 50 previous tasks. Participants in Experiment 1 were 302 adults ( 143 males and 159 females) ranging from 18 to 74 years of age $\left(M_{\text {age }}=34\right)$ and were randomly assigned to the explain condition $(N=151)$ or the control condition $(N=151)$. Ninety-four additional participants (44 in the explain condition and 50 in the control condition) were excluded for failing to pass two attention checks (see below).

Materials Thirty-two images of "alien robots" (see Figure 1) were designed based on the stimuli used by Williams and Lombrozo (2010). Robots varied along four dimensions: foot shape, body/head shape, left-half color, and right-half color. Twenty-two different foot shapes were used, each of which appeared on no more than two robots. All Glorps had feet that were pointy on the bottom surface, and all Drents had feet that were flat on the bottom surface. Overall, $75 \%$ of Glorps had round bodies/heads, $25 \%$ of Glorps had
square bodies/heads, $75 \%$ of Drents had square bodies/ heads, and $25 \%$ of Drents had round bodies/heads. The color dimensions were irrelevant to category membership.

Foot shape (pointy/flat) was a " $100 \%$ rule" that accounted for the category membership of all robots, and body/head shape (round/square) was a " $75 \%$ rule" that only accounted for the category membership of $75 \%$ of the robots.

Procedure Participants were introduced to Glorp robots and Drent robots. On each of 16 trials, participants were shown a Glorp-Drent pair. Robots were paired such that no color appeared more than once in a pair, and all atypical Glorps were paired with atypical Drents. The side on which Glorps and Drents appeared was counterbalanced across trials. Pairs were presented in a random order, aside from the first four trials. For these trials, typical exemplars were presented on trials one, two, and three ("typical trials"), and atypical exemplars were presented on trial four ("anomaly trial").

On each trial, participants were given the choice to "observe" - offering the opportunity to gain information but no reward - or "bet" - offering the opportunity to gain reward but no information. If a participant chose to observe, the participant was shown which robot from that pair was a Glorp and which was a Drent. Participants in the explain condition were asked to explain why the indicated robot was a Glorp robot, while participants in the control condition were asked to write down any thoughts they had about that trial. Participants were required to spend at least 10 seconds completing these tasks before advancing to the next trial. No points were awarded when a participant chose to observe.

If a participant chose to bet, the participant was asked to indicate which robot they thought was a Glorp. If their choice was correct, the participant would gain one point, and if their choice was incorrect, they would lose one point. However, no feedback was given on bet trials; participants were not shown their scores until the task was complete. ${ }^{1}$

Participants were instructed to attempt to maximize their score, but also to learn how to differentiate Glorps and Drents. All participants were explicitly told that they would be asked to report any patterns that could help differentiate Glorps and Drents at the end of the task. Participants were not incentivized on the basis of their score, a point to which we return in the General Discussion.

After the 16-trial observe or bet task, participants reported any patterns they had found that differentiated Glorps and Drents and indicated what percentage of robots they thought could be accurately characterized using that pattern. Participants could report up to eight patterns. Participants also completed an attention check in which they had to distinguish between a robot they had seen during the previous task and three robots that they had not seen before. All novel robots were obviously different in appearance from Glorps and Drents. A second attention check required

[^392]participants to read the instructions from the first attention check, which directed them to ignore the question that followed and instead type a specific word into the answer textbox.

## Results

Rule Discovery In the explain condition, $17 \%$ of participants reported the $100 \%$ rule after completing the observe or bet task, while only $6 \%$ of participants in the control condition reported this rule. A chi-square analysis revealed that this difference was significant, $\chi^{2}(1)=8.27, p$ $=.004$. While these discovery rates seem quite low, they are not inconsistent with previous research (Williams \& Lombrozo, 2010). Additionally, these results replicate Williams and Lombrozo's (2010) finding that generating explanations promotes the discovery of broad rules.


Figure 2: Experiment 1 choices by condition and trial. Vertical line indicates first anomaly trial. Error bars: 1 SE.

Explanation and Exploration We used a generalized linear mixed effects model to predict participants' observe/bet choices over time, with a random intercepts term to capture individual differences. The choice to observe rather than bet was significantly predicted by linear and quadratic effects of trial number, analysis of deviance: $\chi^{2}(1)=212.70, p<.001$ and $\chi^{2}(1)=78.04, p<.001$, respectively. Increasing trial number led to more betting at a decreasing rate over time. Condition was not a significant predictor of observe/bet choices (see Figure 2). This indicates no overall differences between the two conditions in explore/exploit decisions.

Next, we investigated the effect of explaining anomalies on subsequent exploration. For participants who observed on the first anomaly trial and thus received information contradicting the $75 \%$ rule, explain condition participants were more likely than control condition participants to continue to observe on the following trial, at a level approaching significance, $\chi^{2}(1)=3.49, p=.06$. There was no condition difference in observation on the trial following the first anomaly for participants who did not observe on the anomaly trial, and thus did not receive information contradicting the $75 \%$ rule, $\chi^{2}(1)=0.55, p=.46$. However, a logistic regression predicting observation following the anomaly trial by condition and observation on the anomaly
trial did not reveal a significant interaction ( $b=0.71$, $\mathrm{OR}=$ $2.03, z=1.24, p=0.21$ ), likely due to the small proportion of our sample that observed the anomaly ( $27 \%$ ).

To analyze whether the increased exploration following an observed anomaly led to increased discovery of the $100 \%$ rule, we performed a logistic regression predicting $100 \%$ rule discovery by condition and observation on the trial following the first anomaly. Both condition $(b=1.12, \mathrm{OR}=$ $3.05, z=2.73, p=.006)$ and post-anomaly observation $(b=$ $0.80, \mathrm{OR}=2.23, z=2.17, p=.03$ ) were significant predictors. Thus, while explanation had an effect on $100 \%$ rule discovery above and beyond the effects of postanomaly observation, this increase in exploration following the observation of an anomaly also predicted rule discovery.

## Discussion

In Experiment 1, we found that after observing information that contradicted the $75 \%$ rule, participants who were asked to explain tended to explore more often than control participants. This exploration increased the probability of discovering a broad rule that accounted for all observations. These findings support Hypothesis 2: explanation led learners towards greater exploration after receiving evidence that current beliefs were wrong or incomplete.
These results leave open two possibilities for how explanation interacts with anomalous information. We previously suggested that explaining anomalies encourages learners to reject their current (imperfect) hypothesis, prompting subsequent exploration in the service of finding a more satisfactory alternative. On this view, explanation affects the downstream processing that follows the detection of an anomaly. However, there is also evidence that generating explanations can help learners articulate and recognize their current beliefs, rendering a conflict between those beliefs and subsequent information more apparent (Chi, de Leeuw, Chiu, \& LaVancher, 1994). Extending these ideas, it could be that generating explanations on trials that preceded an observed anomaly helped learners recognize the anomalies as such, and that increased sensitivity to anomalies is what drove effects of explanation.

In Experiment 2, we evaluate these alternatives by introducing violations of the $75 \%$ rule that are detectable whether or not the participant chooses to observe on that trial ("obvious anomalies"). When an atypical exemplar from one category (e.g., a round Drent) is paired with a typical exemplar from the other category (e.g., a round Glorp), both robots have the same shape. Since each trial contains one Glorp and one Drent, the trial is a clear violation of the $75 \%$ rule. If explanation enhances anomaly detection by solidifying learners' initial beliefs, we would expect participants who are prompted to explain to observe at a higher rate on the first obvious anomaly trial relative to those who are not prompted to explain. On the other hand, if explaining an anomaly is instead what prompts learners to reject prior beliefs and seek out better alternatives, we would expect effects of explanation to emerge only after an anomaly has been observed, and to manifest as an increase
in observation on trials following obvious anomalies. In Experiment 2, we test these potential accounts.

We additionally varied the point at which the first obvious anomaly was introduced - on trial 4 versus trial 8 - to test whether the timing of anomalous information affects rule discovery and/or interacts with explanation. If the power of explaining anomalous information emerges from the conflict between the novel information and prior beliefs, then introducing anomalous evidence later in the task (after beliefs have been solidified) should lead to a larger effect of explanation on exploration. If, however, explanation biases learners towards their prior beliefs (Walker, Lombrozo, Williams, Rafferty, \& Gopnik, 2017; Williams \& Lombrozo, 2013), increasing the strength of learners' beliefs by increasing the amount of consistent evidence prior to introducing an anomaly could decrease the effect of anomalous evidence on subsequent exploration.

## Experiment 2

## Method

Participants Participants were 400 adults ( 192 males, 204 females, and 4 who did not specify gender) ranging from 18 to 73 years of age $\left(M_{\text {age }}=34\right)$ who were randomly assigned to the explain condition $(N=203)$ or the control condition ( $N=197$ ), as well as early anomaly timing ( $N=197$ ) or late anomaly timing ( $N=203$ ). One hundred fifty-three additional participants ( 73 in the explain condition and 80 in the control condition) were excluded for failing to pass two attention checks mirroring those used in Experiment 1.

Materials The 32 alien robot images used were identical to those used in Experiment 1.

Procedure The procedure was largely identical to Experiment 1. Three atypical Glorps were paired with atypical Drents ("non-obvious anomalies"). One atypical Glorp was paired with a typical Drent, and one atypical Drent was paired with a typical Glorp ("obvious anomalies").

For those assigned to early anomaly timing, the first obvious anomaly was presented on trial 4. For late anomaly timing, the first obvious anomaly was presented on trial 8. The second obvious anomaly was always on trial 15 . Nonobvious anomalies were randomly distributed throughout the remaining trials, excluding the first three trials.

## Results

Rule Discovery Within early anomaly timing participants, $33 \%$ of explain participants and $27 \%$ of control participants reported the $100 \%$ rule. Within late anomaly timing participants, $30 \%$ of explain participants and $14 \%$ of control participants reported the $100 \%$ rule. A logistic regression analysis revealed that participants in the explain condition were significantly more likely than participants in the control condition to discover the $100 \%$ rule ( $b=0.60$, $\mathrm{OR}=$ $1.81, z=2.55, p=.01$ ). Participants with late anomaly timing were somewhat less likely than participants with
early anomaly timing to discover the $100 \%$ rule ( $b=-0.43$, $\mathrm{OR}=0.65, z=-1.84, p=.07$ ).

Explanation and Exploration We used a generalized linear mixed effects model to predict participants' observe/bet choices over time, with a random intercepts term to capture individual differences. The choice to observe rather than bet was significantly predicted by condition (explain vs. control) and linear and quadratic effects of trial number, analysis of deviance: $\chi^{2}(1)=4.47, p=.03 ; \chi^{2}(1)=273.93, p$ $<.001$; and $\chi^{2}(1)=150.40, p<.001$, respectively. Anomaly timing (early vs. late) was not a significant predictor of observation. Explain participants were more likely to observe than control participants, and increasing trial number increased the likelihood of betting at a decreasing rate over time (see Figure 3).

Next, we analyzed exploration on the first obvious anomaly trial by performing a logistic regression with task (explain vs. control) and anomaly timing (early vs. late) as predictors. Participants with late anomaly timing were 56\% less likely to observe on the first anomaly trial relative to participants with early anomaly timing ( $b=-0.82$, $\mathrm{OR}=$ $0.44, z=-2.81, p=.005$ ), indicating that fewer participants observed the first obvious anomaly when it was presented later in the task. Condition was not a significant predictor of anomaly observation, nor was the interaction between condition and anomaly timing. These findings suggest that explanation did not exert effects on discovery by increasing the rate at which obvious anomalies were detected.

To analyze exploration following an anomalous observation, we performed a logistic regression predicting observation on the trial following the first obvious anomaly, with condition (explain vs. control) and anomaly timing (early vs. late) as predictors. This revealed a marginally significant interaction between task and anomaly condition ( $b=0.88, \mathrm{OR}=2.40, z=1.78, \mathrm{p}=.07$ ). For late anomaly timing, explain participants were more likely than control participants to observe on the trial following the first obvious anomaly, $\chi^{2}(1)=6.85, p=.009$. This difference was not significant for early anomaly timing, $\chi^{2}(1)=0.01, p$ $=.92$. These findings support the idea that explanation affects learning by increasing exploration in the face of anomalous evidence; they also challenge the idea that effects of explanation are restricted to anomaly detection. Explainers were no more likely to choose to observe an obvious anomaly, but were more likely (in the late anomaly condition) to follow up with additional observation.

To analyze whether this increased exploration following an observed anomaly led to increased discovery of the $100 \%$ rule, as well as whether condition had an effect on rule discovery above and beyond the effects of such exploration, we performed a logistic regression predicting $100 \%$ rule discovery by condition and observation on the trial following the obvious anomaly. Both condition ( $b=0.49$, $\mathrm{OR}=1.64, z=2.10, p=.04)$ and observation following the first obvious anomaly $(b=0.79$, $\mathrm{OR}=2.20, z=3.12, p=$ .002 ) were significant predictors of $100 \%$ rule discovery.

## Discussion

These results suggest that explanation generation leads to greater exploration after observing evidence that challenges a current hypothesis. This difference in exploratory behavior does not depend on simply noticing the presence of contradictory information, but instead depends specifically on one's attempts to explain this anomalous information.

We also found a suggestive difference between early and late anomaly timing. Further research is clearly warranted, but the effect of explaining an obvious anomaly may be more powerful as the degree of conflict between the anomaly and one's current beliefs is increased.


Figure 3: Experiment 2 choices by condition, anomaly timing, and trial. Vertical lines indicate obvious anomaly trials. Error bars: 1 SE.

## General Discussion

In two experiments, we investigated how explanation generation affects exploration over the course of a category learning task. Lombrozo and colleagues (Lombrozo, 2016; Williams \& Lombrozo, 2010, 2013) have proposed that generating explanations recruits a set of inductive constraints on hypothesis generation and selection, which can lead to the discovery of broad, simple, and generalizable rules and patterns. In the present research, we extend this account, suggesting that this learning outcome is partially dependent upon generating explanations for anomalous observations, which increases a learner's propensity to seek additional evidence.

Our results are consistent with Legare's (2012) finding that children's explanations for surprising events predicted their exploratory behavior. In the present research, however, we establish a causal link between explanation and exploration, and demonstrate that this link holds not only for children's causal learning (as proposed by Legare), but also for adults' category learning.

That said, many open questions remain. For example,
might explanation affect explore/exploit decisions by shifting participants' confidence on each trial (e.g., Auer, 2002)? Does explaining affect motivation, which could also be achieved by incentivizing reward? Equally important is identifying boundary conditions on our effects: are there situations in which explaining anomalies could lead learners to explain them away (Chinn \& Brewer, 1998), and thus engage in greater exploitation?

While some of the effects reported here failed to reach statistical significance, we did find similar results across two experiments. Unfortunately, the effect sizes are limited by the small proportion of participants who are able to discover the $100 \%$ rule. Future work might benefit from more sensitive paradigms. Additionally, the paradigm used here allowed participants to gain some information on each trial without engaging in exploration. Since exemplars from one category were always presented with exemplars from the other category, participants could identify the features that differed between the two categories without choosing to observe. Future work will limit potential learning to observation trials in order to better isolate the effects of explanation on information search.

In sum, our findings suggest that attempting to explain observations that are anomalous with respect to one's current beliefs encourages further exploration. This may be one mechanism by which generating explanations affects learning, and provides compelling evidence that top-down constraints on hypothesis generation and selection affect not only the conclusions that learners draw, but also the ways in which they seek information - whether that information comes from $19^{\text {th }}$ century texts or a robot classification task.

## Acknowledgments

This work was supported by a McDonnell Foundation award to TL. EL is also the recipient of a National Science Foundation Graduate Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the McDonnell Foundation or the National Science Foundation.

## References

Auer, P. (2002). Using confidence bounds for exploitationexploration trade-offs. Journal of Machine Learning Research, 3, 397-422.
Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., \& Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. Cognitive Psychology, 64(4), 215-234.
Chi, M. T., de Leeuw, N., Chiu, M., \& LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.
Chinn, C. A., \& Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. Journal of Research in Science Teaching, 35(6), 623-654.
Cohen, J. D., McClure, S. M., \& Yu, A. J. (2007). Should I
stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1481), 933-942.
Ihme, N., \& Wittwer, J. (2015). The role of consistency, order, and structure in evaluating and comprehending competing scientific explanations. Instructional Science, 43(4), 507-526.
Johnson, T. R., \& Krems, J. F. (2001). Use of current explanations in multicausal abductive reasoning. Cognitive Science, 25(6), 903-939.
Kaelbling, L. P., Littman, M. L., \& Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.
Langford, J., \& Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. Advances in neural information processing systems 20 (pp. 817-824).
Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesistesting behavior in young children. Child Development, 83(1), 173-185.
Lombrozo, T. (2016). Explanatory preferences shape learning and inference. Trends in Cognitive Sciences, 20(10), 748-759.
Murdock, J., Allen, C., \& DeDeo, S. (2017). Exploration and exploitation of Victorian science in Darwin's reading notebooks. Cognition, 159, 117-126.
Navarro, D. J., Newell, B. R., \& Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. Cognitive Psychology, 85, 43-77.
Ruggeri, A., Lombrozo, T., \& Xu, F. (in prep). Explaining prepares preschoolers to ask efficient questions.
Tversky, A., \& Edwards, W. (1966). Information versus reward in binary choices. Journal of Experimental Psychology, 71(5), 680-683.
Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., \& Gopnik, A. (2017). Explaining constrains causal learning in childhood. Child Development, 88(1), 229246.

Williams, J. J., \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. Cognitive Science, 34(5), 776806.

Williams, J. J., \& Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. Cognitive Psychology, 66(1), 55-84.
Williams, J. J., Lombrozo, T., \& Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. Journal of Experimental Psychology: General, 142(4), 1006-1014.
Williams, J. J., Walker, C. M., \& Lombrozo, T. (2012). Explaining increases belief revision in the face of (many) anomalies. Proceedings of the 34th annual conference of the cognitive science society (pp. 1149-1154). Austin, TX: Cognitive Science Society.

# Can Illness be Bright? Metaphor Comprehension Depends on Linguistic and Embodied Factors 

Pei Liu (p.liu1@lancaster.ac.uk)<br>Louise Connell (l.connell@lancaster.ac.uk)<br>Dermot Lynott (d.lynott@lancaster.ac.uk)<br>Department of Psychology, Fylde College,<br>Lancaster University, Bailrigg, Lancaster, LA1 4YF, UK


#### Abstract

Conceptual representations in language processing employ both linguistic distributional and embodied information. Here, we aim to demonstrate the roles of these two components in metaphor processing. The linguistic component is captured by linguistic distributional frequency (LDF), that is, how often the constituent words appear together in context. The embodied component, on the other hand, refers to how easy it is to generate an embodied simulation, operationalised by a previous norming study. In the current study, we looked at the interplay of these components in metaphor processing, and investigated their roles at different depths of processing in two experiments. Thus, we required participants to engage in shallow processing (Experiment 1: Sensibility Judgement), or deep processing (Experiment 2: Interpretation Generation). Results showed that the increase of both variables made it more likely to accept a metaphor. However, whereas ease of simulation (EoS) contributed to the speed of processing at both levels of depth, LDF only affected the speed in shallow processing. Specifically, LDF acted as a heuristic, both to speed up responses to accept metaphors as sensible when the frequency is high, and to flag up potentially unsuccessful processing when it is low. Overall, these results support views of language processing that emphasise the importance of both linguistic and embodied components according to task goals.


Keywords: metaphor processing; embodied cognition; linguistic distribution; simulation; depth of processing

## Traditional views of metaphor processing

In a metaphoric expression, a word or a phrase (the source) is applied to an object or an action (the target) to which it cannot be literally applied. In the expression "a bright student", a student is not an object to which the visual property of bright is usually applied. Nevertheless, we can comprehend it effortlessly meaning "clever or intelligent students". How is this comprehension achieved? Many factors have been implicated, namely the familiarity, conventionality and aptness of a metaphor. These factors can not only affect the speed of processing (Giora, 2007; Pierce \& Chiappe, 2008), but conventionality and aptness are also suggested to determine the mechanism of processing, whether by comparison or by categorisation (Bowdle \& Gentner, 2005; Jones \& Estes, 2006).

However, when it comes to understanding exactly how these three factors affect metaphor comprehension, they have problems with their theoretical specificity, and subsequently their operationalisation. Familiarity and conventionality are often used interchangeably, and they
face the same problem concerning their operational definitions. They are sometimes assumed to refer to how often people have encountered the metaphoric expression itself (e.g., how often is "bright" used to describe "students"? e.g., Cardillo, Watson, Schmidt, Kranjec, \& Chatterjee, 2012; Roncero \& de Almeida, 2014a) and sometimes to how accustomed people are to relating the expression to its metaphoric meaning (e.g., "bright" meaning intelligent and quick-witted: Campbell \& Raney, 2015; Mashal, Faust, Hendler, \& Jung-Beeman, 2009), but these are two very different and dissociable theoretical constructs. A particular linguistic expression might be encountered reasonably often but remain poorly understood (e.g., purple prose), or a metaphoric meaning might be encountered reasonably often via a different expression to the one supplied (e.g., "Solution can be bright").

Aptness also has received different definitions. It is sometimes assumed to reflect a very general, high-level quality or goodness of a metaphor and is often operationalised as such (Haught, 2014), whereas at other times represents a much more low-level specification of how well the metaphoric meaning (e.g., intelligent and quick-witted) fits or overlaps with the target (e.g., "student": Chiappe \& Kennedy, 1999). In addition, aptness appears to be theoretically confounded with familiarity and conventionality. Only apt metaphors are likely to become conventionalised or familiar, as a metaphor that does not work well is unlikely to become widely used by speakers of a language. Because familiarity and conventionality depend on usage patterns of metaphors across a language, and usage patterns depend to some extent on aptness, it means that there is a core dependency between the factors that is not trivial to disentangle. Indeed, ratings of aptness and familiarity are highly correlated ( $\mathrm{r}=.73-.82$ : Campbell \& Raney, 2015; Roncero \& de Almeida, 2014a), as are ratings of aptness and corpus frequency counts of the metaphoric expression ( $\mathrm{r}=.41-.57$ : Roncero \& de Almeida, 2014b; Thibodeau \& Durgin, 2011).

In summary, familiarity, conventionality and aptness have all been shown to affect metaphor processing. However, they have several theoretical and operational problems that mean they have limited utility in enhancing our understanding of what makes a metaphor easier to process. Rather than continuing to vary and refine how these factors are conceptualised, we propose a different approach to seek clearer predictors of metaphor processing that (a) are
theoretically and operationally distinct, and (b) are able to independently account for speed and accuracy performance in metaphor processing.

## Grounded views of language processing

Research in conceptual representation has tended to operate in parallel to that of traditional metaphor processing, and therefore takes quite a different perspective on how access to meaning takes place. Essentially, two components are employed in the mental representation of meaning when people process language (Barsalou, Santos, Simmons, \& Wilson, 2008; Connell \& Lynott, 2014). The first component relies on the statistical, distributional pattern of how words co-occur across contexts (Landauer \& Dumais, 1997). The second type of representation is the embodied (also known as the grounded, sensorimotor or situated) component, which relies on the process of simulation; that is the partial reactivation of past perceptual, motor, affective, introspective and other experiences (Barsalou, 1999).

Together, the linguistic and embodied components can explain language processing better than either alone (e.g., Andrews, Vigliocco, \& Vinson, 2009). In particular, research in the grounded linguistic-embodied approach has demonstrated that the linguistic distributional information provides a powerful tool for superficial language processing because activity of the linguistic component peaks earlier than that of the embodied simulation component (Louwerse \& Jeuniaux, 2008). People are more likely to rely on the embodied component when deeper processing is specifically cued in the task; but people will be reliant upon the linguistic component to generate a good-enough approximation (Ferreira, Bailey, \& Ferraro, 2002) when shallow processing can suffice.

In line with these arguments, Connell and Lynott (2013) proposed that information from the linguistic component could act as a cognitive triage mechanism during language processing by providing a guide to whether it is worth expending effort on costly embodied simulation. If the linguistic component indicates that future processing is likely to fail (e.g., the words rarely co-occur in the same context and so their combined meaning might not be simulated successfully), then the processing cold be abandoned before any more cognitive effort is expended by the embodied component. On the other hand, if the linguistic component indicates that future processing is likely to succeed (e.g., the words often co-occur in the same context and so their combined meaning can probably be simulated successfully), then it could either inform a response immediately (i.e., based on the linguistic shortcut alone) or allow the embodied component to continue developing a detailed simulation of meaning.

Connell and Lynott's proposal can be applied directly to the study of metaphor processing, because the interplay of the linguistic and embodied components, and the role of the linguistic shortcut as a cognitive triage mechanism, operate in theory across all types of language comprehension. In this study, we asked participants to process metaphors that
systematically varied on these two dimensions. The linguistic component is quantified by linguistic distributional frequency (LDF), that is how often constituent words ("bright" and "student") of a metaphor co-occur in a large corpus.

The embodied component, on the other hand, is operationalised by ease of simulation (EoS), a new normed metric that quantifies how easy it is to come to a mental representation of a metaphor (Liu, Connell, \& Lynott, 2016). This metric was extracted using a principle components analysis from the ratings on the sensibility (how much sense the sentence make if read or heard), usability (how easy it is to use it in conversation or writing), and imaginability (how easy it is to describe the concept) in the norming study. When combined into a single measure, these ratings offer a proxy for how easy it is to simulate a concept. That is, if people find it easy to make sense of and use the metaphor, as well as imagine the concept, they would find it easy to generate embodied simulations.

Although LDF and EoS may correlate to a certain degree, we expected them to play distinctive roles in metaphor processing after the common variance between them is removed. Both variables would independently affect the acceptance rate and speed of metaphor processing. Specifically, increase in both variables would make it more likely and faster for people to accept a metaphor, and meanwhile slower to reject the very metaphor. More crucially, participants performed one of two tasks: a sensibility judgement task (Experiment 1), which required relatively shallower processing because participants made only yes/no response; or an interpretation-generation task (Experiment 2), which was deeper because they specified the meaning verbally. We predicted that EoS would play a greater role in deep processing or when people accepted metaphors because it indicated successful simulation, while LDF should play a larger role in shallow processing especially when people rejected a metaphor as not being sensible.

## Experiment 1: Sensibility Judgement Task

## Method

Participants Twenty-eight participants took part (five male and 23 female), all of whom were students at Lancaster University and native speakers of English with mean age of 19.1 years ( $\mathrm{SD}=1.1$ ). The sample size was determined beforehand to achieve the same level of variability as the conceptual combination study in Connell and Lynott (2013).
Materials We used a total of 452 metaphoric sentences taken from Liu et al., (2016). All sentences took the form "Noun can be adjective" (e.g., Student can be bright), and were composed of 113 perceptual adjectives (e.g., bright: Lynott \& Connell, 2009), each paired with four nouns that were capable of eliciting metaphoric meanings that vary independently on the following two dimensions (see examples in Table 1):

Table 1: Sample metaphors, and their scores for ease of simulation and LDF.

| Metaphor | EoS | LDF |
| :--- | :---: | :---: |
| Illness can be bright. | -1.32 | 2.95 |
| Supply can be bright. | -1.02 | 3.72 |
| Solution can be bright. | 1.41 | 3.11 |
| Student can be bright. | 1.84 | 4.08 |

EoS for each sentence ranged from easy to difficult ( $M=$ $0.00, S D=1.00$ ), and was calculated in a novel norming study by Liu et al. (2016). The scores were a single principle component extracted from ratings of sensibility, imaginability and usability of the metaphors. LDF for each sentence ranged from low to high ( $M=2.95, S D=0.97$ ), and was calculated as the $\log$ of the summed bi- to fivegram frequencies of the sentence's noun and adjective in the Google Web1T Corpus (Brants \& Franz, 2006). To take the metaphor "Students can be bright" as an example, the LDF was the $\log$ of the sum of the frequencies of "student ... bright" and "bright ... student" with zero, one, two, and three intervening words.

The sentences were split into four lists, where each adjective appeared only once per list, and the distribution of easy/difficult to simulate and high/low distributional frequency was equal across lists (EoS: $F(3,440)=1.70, p$ $=.166$; LDF: $F(3,440)=0.43, p=.734)$. Each participant saw only one list.

Procedure Participants read one sentence in each trial and decided whether or not the sentence made sense. All trials had the same structure. Participants first saw a fixation cross for 1000 ms , followed by the noun for 500 ms , followed by the phrase "can be" for 500 ms , and then followed by the adjective. The adjective remained onscreen until participants made a response. Participants pressed either the comma key (",") if they judged that the sentence made sense; or the full stop key (".") if they judged that it did not make sense. The response could be made without a time limit; and participants were told explicitly that there were no right or wrong answers to the question. Both the response decisions ("yes" to accept the metaphor as sensible; or "no" to reject
the metaphor as nonsensical), and the response time in milliseconds (RT) from onset of the adjective were recorded as dependent variables.

Design and Analysis Response decisions were analysed in a mixed effects logistic regression (binomial distribution with logit link), with response as the dependent variable (coded as 1 for "yes", accepting the metaphor as sensible; and 0 for "no", rejecting the metaphor as nonsensical), participants and items as crossed random factors, and LDF, EoS, and their interaction as fixed factors. We only modelled random intercept because models with random slope failed to converge.

Response times (RTs) were analysed using mixed effects linear regressions, firstly in an omnibus analysis with crossed random factors of participants and items, and fixed factors of response, LDF, EoS, and their interactions. Then, we ran separate analyses on acceptance ("yes") and rejection ("no") responses because we expect the effects of the fixed factors to be opposite for "yes' and "no" responses. While the increase in LDF and EoS should make it faster to accept a metaphor, it should make it slower to reject a metaphor.

## Results and Discussion

All participants had mean response times within 3SD of the overall mean and so all were included in analysis. Two trials were removed because of motor error (RT < 200ms). Furthermore, individual trials with RT more than 3SD from each participant's mean per response decision were removed as outliers: $1.33 \%$ of "yes" responses and $2.20 \%$ of "no" responses.
Among 3105 valid trials, 1413 ( $45.51 \%$ ) were accepted as sensible ("yes" responses) and 1692 ( $54.59 \%$ ) were rejected as nonsensical ("no" responses). Logistic regression showed evidence for net suppression. This means that while metaphors with the "yes" response had a higher mean LDF than those with the "no" response (i.e. the higher LDF was, the more likely it should be to respond "yes"), the effect of LDF in the mixed effects logistic regression turned out negative (i.e., the higher LDF was, the less likely to respond "yes"). This suggested that LDF enhanced the effect of

Table 2: Results from the mixed effects linear regression of RT in Experiment 1.

| Variable | $\beta$ | $95 \%$ CI | df | $t$ | $p$ |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Intercept | 1129.66 | $[1011.04,1248.29]$ | 28.2 | 18.67 | $<.001$ |
| Response | 0.76 | $[-36.86,38.32]$ | 2157.5 | 0.04 | 0.969 |
| EoS | 67.06 | $[41.76,92.36]$ | 1257.7 | 5.20 | $<.001$ |
| LDF | 14.76 | $[-9.81,39.32]$ | 822.3 | 1.18 | 0.239 |
| Response * EoS | -181.55 | -38.55 | $[-74.91,-144.19]$ | 2277.1 | -9.52 |
| Response * LDF | $[1.18,47.61]$ | 2010.1 | -2.10 | 0.001 |  |
| EoS * LDF | $[-67.07,2.08]$ | 1192.1 | 2.06 | 0.046 |  |
| Response * EoS * LDF | -32.50 |  | 2318.1 | -1.84 | 0.066 |

Table 3: Results from mixed effects regressions of RT per response in Experiment 1.

| Response | Variable | $\beta$ | $95 \% \mathrm{CI}$ | df | $t$ | $p$ |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: |
| Accept | Intercept | 1172.17 | $[1051.79,1292.53]$ | 25.7 | 19.09 | $<.001$ |
| ("yes") | EoS | -134.88 | $[-161.98,-107.79]$ | 357.3 | -9.76 | $<.001$ |
|  | LDF | -27.52 | $[-53.10,-1.93]$ | 248.1 | -2.11 | .036 |
|  | EoS * LDF | -13.00 | $[-37.91,11.91]$ | 394.6 | -1.02 | .307 |
| Reject | Intercept | 1155.41 | $[1010.94,1299.88]$ | 26.7 | 15.68 | $<.001$ |
| ("no") | EoS | 72.28 | $[47.27,97.29]$ | 323.8 | 5.66 | $<.001$ |
|  | LDF | 17.73 | $[-6.40,41.83]$ | 189.1 | 1.44 | .152 |
|  | EoS * LDF | 23.33 | $[0.51,46.14]$ | 295.3 | 2.00 | .046 |

EoS by explaining the residuals of EoS rather than the variance of response decision. In order to establish the true relationships between response decision and our independent variables, we therefore removed the shared variance between LDF and EoS (currently correlating at $r$ $=.27$ ) by orthogonalising the variables using a principal components analysis with varimax rotation and Kaisar normalization on a covariance matrix (Glantz \& Slinker, 2000). What this did was to create two new orthogonal variables ( $r=0$ ), each correlating highly with one original variable $(r=.99)$. These two new variables were thus named orthogonal EoS and orthogonal LDF. We re-ran the logistic regression with them and obtained results as follows (further analyses all used orthogonal variables).

The logistic regression with orthogonal variables showed that both variables had a positive effect on response decision. As the EoS increased by one unit, the odds to accept a metaphor as sensible increased 3.42 times $(z(3101)=25.03, p<.001, \beta=1.23)$. As the LDF increased, the odds to accept a metaphor increased with a marginally significant effect (1.084 times, $z(3101)=1.88$, $p=.06, \beta=0.08$ ).

RT was also analysed using orthogonal variables. Table 2 shows full results of the omnibus analysis. Overall, EoS had a positive effect on RT ( $M=1139 \mathrm{~ms}, S D=587 \mathrm{~ms}$ ), and it critically interacted negatively with response decision, suggesting that the direction of the EoS effect differed by the response type, and was greater for "yes" than "no" responses. LDF had no overall main effect, but interacted with response decision to indicate that the direction of LDF differed for "yes" and "no" RTs.

Since we had separate hypotheses for "yes" and "no" RTs, we divided the dataset by response decision and analysed their RTs separately. Results are given in Table 3. For "yes" responses (i.e. metaphors that were accepted as sensible; RT: $M=1150 \mathrm{~ms}, S D=589 \mathrm{~ms}$ ), the easier a metaphor was to simulate, the less time people took to accept it as sensible. Also, the more often the words in a metaphor co-occurred in language, the faster people were to accept it as sensible.

For "no" responses ( RT : $\mathrm{M}=1114 \mathrm{~ms}, \mathrm{SD}=603 \mathrm{~ms}$ ),
the effects ran in the opposite direction (Table 3). As predicted, people were faster to reject metaphors that were normally regarded as difficult to simulate. Furthermore, it interacted with LDF positively, such that the effect of EoS was reduced when LDF was low (the $\beta$ for the interaction term was positive).

## Experiment 2: Interpretation Generation Task

## Method

In this study, we asked 40 participants (native speakers of English, 11 males, age: $M=19.65, S D=2.08$ ) to judge whether they could think of a meaning for the metaphoric sentences instead of judging their sensibility. The procedure was the same as Experiment 1, except that participants needed to provide their interpretation of the sentences after they responded "yes". To reduce the possibility of fatigue, each participant saw half (56-57) of the items of Experiment 1.

## Results and Discussion

Data cleaning was performed as in Experiment 1. Furthermore, we also identified accept ("yes") trials with invalid interpretations (e.g., blank, "I don't know"). Two participants were excluded from analysis for providing more than $50 \%$ invalid interpretations. Amongst the remaining participants, $2.33 \%$ of interpretations were identified as invalid. For individual trials, $2.10 \%$ of "yes" responses and $2.00 \%$ of "no" responses were identified as outliers.

Among 2103 valid trials, 1302 (61.91\%) were accepted as interpretable whereas 801 ( $38.09 \%$ ) were rejected as uninterpretable. The logistic regression showed that both EoS and LDF had positive effects on response decision. For every unit of increase in EoS, the odds of accepting a metaphor as interpretable increased 2.826 times ( $z(2099)$ $=17.49, p<.001, \beta=1.04$ ); and for every unit of increase in LDF, the odds of accepting a metaphor increased 1.286

Table 4: Results of the mixed effects linear regression of RT in Experiment 2.

| Variable | $\beta$ | $95 \% \mathrm{CI}$ | df | $t$ | $p$ |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Intercept | 2796.79 | $[2341.82,3251.76]$ | 43 | 12.05 | $<.001$ |
| Response | 58.38 | $[-143.64,260.40]$ | 1487.5 | 0.57 | .571 |
| EoS | 125.28 | $[-20.79,271.35]$ | 1466.9 | 1.68 | .093 |
| LDF | 44.47 | $[-97.02,185.95]$ | 1461.8 | 0.62 | .538 |
| Response * EoS | -589.63 | $[-774.55,-404.70]$ | 1464.3 | -6.25 | $<.001$ |
| Response * LDF | -13.52 | $[-193.60,166.55]$ | 1462.1 | -0.15 | .883 |
| EoS * LDF | -34.60 | 4.82 | $[-167.90,98.71]$ | 1463.8 | -0.51 |
| Response * EoS * LDF |  |  | 1463.3 | 0.06 | .611 |

Table 5: Results of mixed effects linear regression on RT per response in Experiment 2.

| Response | Variable | $\beta$ | $95 \% \mathrm{CI}$ | df | $t$ | $p$ |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: |
| Accept ("yes") | Intercept | 2961.19 | $[2507.85,3414.54]$ | 37.87 | 12.80 | $<.001$ |
|  | EoS | -538.01 | $[-665.18,-410.83]$ | 344.85 | -8.29 | $<.001$ |
|  | LDF | 17.48 | $[-103.06,138.02]$ | 275.02 | 0.28 | .776 |
|  | EoS * LDF | -34.12 | $[-149.71,81.46]$ | 319.23 | -0.58 | .563 |
| Reject ("no") | Intercept | 3245.15 | $[2507.68,3982.63]$ | 31.53 | 8.63 | $<.001$ |
|  | EoS | 213.38 | $[100.81,325.94]$ | 355.53 | 3.72 | $<.001$ |
|  | LDF | 44.11 | $[-62.70,150.92]$ | 321.04 | 0.81 | 0.419 |
|  | EoS * LDF | -56.58 | $[-158.14,44.98]$ | 413.74 | -1.09 | 0.276 |
|  |  |  |  |  |  |  |

times $(z(2099)=4.67, p<.001, \beta=0.25)$.
Linear regression of RT across both responses found no overall effects of either EoS or LDF (see Table 4). However, EoS interacted negatively with response decision, indicating the effect of EoS were in opposite directions for "yes" and "no" responses. Results separated by response decision are given in Table 5. As predicted, for "yes" responses (RT: M $=3083 \mathrm{~ms}, S D=2638 \mathrm{~ms}$ ), EoS had a negative effect, meaning that people were faster to accept a metaphor as interpretable when it was typically considered easy to simulate compared to difficult to simulate. LDF did not affect the speed of interpretation, nor was there an interaction. Also as predicted, for "no" responses (RT: $M=$ $2436 \mathrm{~ms}, S D=2105 \mathrm{~ms}$ ), people were faster to reject a metaphor as uninterpretable when it was normally considered difficult to simulate. LDF did not affect rejection speed, nor did it interact with EoS.

Since we had specific hypothesis with regards to the depth of processing, we examine such task differences further in cross-experiment analyses ( 0 coded for sensibility judgement task, and 1 for interpretation generation). As expected, the likelihood of accepting versus rejecting a metaphor varied by task: the odds to accept a metaphor increased 3.24 times in deep interpretation generation compared to shallow sensibility judgement; $(z(5200)=3.39$,
$p=.001, \beta=1.18$ ). As for response time, EoS interacted with task, showing that the effect is larger for deep interpretation generation than for shallow sensibility judgement, as predicted ("Yes": $t(2461.2)=-6.39, p<.001$, $\beta=-402.9$, "No" : $t(1936.6)=3.37, p=.001, \beta=145.00)$. The interaction between EoS and LDF also varied across tasks $(t(1933.3)=-2.10, p=.036, \beta=-81.67)$, which was larger for shallow than deep processing.

## General Discussion

Our goal in taking this grounded approach was to move the investigation of metaphor processing beyond the traditional factors of familiarity, conventionality, and aptness, which while having a long history of use - have been increasingly criticised for theoretical and operational problems that limit their utility in explaining what makes one metaphor easier to understand than another. Indeed, our study generated complex results that could not be accounted for by traditional theories with single factors.

The current study shows for the first time that both linguistic component (based on linguistic distributional frequency) and embodied component (based on ease of simulation norms) affect metaphor comprehension independently. Their roles are statistically distinct from each
other after we managed to remove their common variance with a principle components analysis. Whereas ease of simulation often had a large effect overall and was more prominent for the "yes" response because that was when simulation was eventually successful; linguistic distributional frequency represents a relatively coarsegrained, but nonetheless highly useful, approximation of whether a particular source and target have previously formed a metaphor. It informs people's responses, not only making acceptance more likely and faster when the words are likely to constitute a meaningful representation, but also flags up potentially unsuccessful simulation to be rejected right away without further processing when distributional frequency is low. Our findings are consistent with the grounded views which suggest that conceptual representation relies on both embodied simulation and linguistic distributional pattern ((Barsalou et al., 2008; Connell \& Lynott, 2014; Louwerse \& Jeuniaux, 2008).

However, against predictions, the main effect of linguistic distributional frequency did not differ between shallow and deep processing tasks according to the cross-experiment analysis (i.e., linguistic distributional frequency itself did not interact with task). This null effect could be because the tasks disincentivised using the linguistic shortcut by allowing people as much time as needed to make a response. That is, they had unlimited time resource to form a mental representation using the embodied component. In future research, we will impose time constraints on the task in order to further examine the utility of linguistic distributional information during metaphor processing, and provide an additional test of the linguistic shortcut hypothesis.

## References

Andrews, M., Vigliocco, G., \& Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. Psychological Review, 116(3), 463-498.
Barsalou, L. W. (1999). Perceptual symbol systems. The Behavioral and Brain Sciences, 22(4), 577-609-60.
Barsalou, L. W., Santos, A., Simmons, W. K., \& Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. M. Glenberg, \& A. C. Graesser (Eds.), Symbols and Embodiment: Debates on meaning and cognition. Oxford; New York: Oxford University Press.
Bowdle, B. F., \& Gentner, D. (2005). The career of metaphor. Psychological Review, 112(1), 193-216.
Brants, T., \& Franz, A. (2006). Web 1T 5-gram Version 1 LDC2006T13. DVD. Philadelphia: Linguistic Data Consortium.
Campbell, S. J., \& Raney, G. E. (2015). A 25-year replication of Katz et al.'s (1988) metaphor norms. Behavior Research Methods, 48(1), 330-340.
Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., \& Chatterjee, A. (2012). From novel to familiar: tuning the brain for metaphors. NeuroImage, 59(4), 3212-21.

Chiappe, D. L., \& Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. Psychonomic Bulletin \& Review, 6(4), 668-76.
Connell, L., \& Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. Topics in Cognitive Science, 6(3), 390-406.
Ferreira, F., Bailey, K., \& Ferraro, V. (2002). Good-enough representations in language comprehension. Current Directions in Psychological Science, 11(1), 11-15.
Giora, R. (2007). Is metaphor special? Brain and Language, 100(2), 111-4.
Glantz, S., \& Slinker, B. (2000). Primer of Applied Regression \& Analysis of Variance. McGraw-Hill Education.
Haught, C. (2014). Spain is Not Greece: How Metaphors are Understood. Journal of Psycholinguistic Research, 43(4), 351-356.
Jones, L., \& Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. Journal of Memory and Language, 55(1), 18-32.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
Liu, P. Q., Connell, L., \& Lynott, D. (2016). Ease-ofSimulation norms for 452 Adjective Metaphors. In Prep.
Louwerse, M., \& Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In A. Glenberg \& A. C. Graesser (Eds.), Embodiment and meaning: A debate. Oxford, England: Oxford University Press.
Lynott, D., \& Connell, L. (2009). Modality exclusivity norms for 423 object properties. Behavior Research Methods, 41, 558-564.
Mashal, N., Faust, M., Hendler, T., \& Jung-Beeman, M. (2009). An fMRI study of processing novel metaphoric sentences. Laterality, 14(1), 30-54.
Pierce, R. S., \& Chiappe, D. L. (2008). The Roles of Aptness, Conventionality, and Working Memory in the Production of Metaphors and Similes. Metaphor and Symbol, 24(1), 1-19.
Roncero, C., \& de Almeida, R. G. (2014a). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. Behavior Research Methods.
Roncero, C., \& de Almeida, R. G. (2014b). The importance of being apt: metaphor comprehension in Alzheimer's disease. Frontiers in Human Neuroscience, 8, 973.
Thibodeau, P. H., \& Durgin, F. H. (2011). Metaphor Aptness and Conventionality: A Processing Fluency Account. Metaphor and Symbol, 26(November), 206-226.

# Want to prime exercise? Calorie labels work better than activity ones! 

Dawn Liu (dliuxi@essex.ac.uk)<br>Marie Juanchich (m.juanchich@essex.ac.uk)<br>Department of Psychology, University of Essex<br>Wivenhoe Park, Colchester, CO4 3SQ, UK


#### Abstract

'Activity-equivalent' food labels are believed to encourage consumers to partake in exercise. This may occur by semantic priming, where featuring images of physical activity increases the mental accessibility of the concept of exercise, making it more 'fluent' and therefore more influential on people's behaviour. We tested how the format of labels (image vs. text) and representation of energy ('activity' vs. 'calorie) affected mental accessibility of exercise in a word-fragment completion task and participants' behavioural intentions for exercise $(N=142)$. Participants exposed to calorie labels produced more exercise-related words and viewed an imagined exercise scenario as shorter and more enjoyable. Images led to higher intentions to exercise than text when they described activities but they led to lower intentions to exercise than text when they described calories. Findings suggest that activity labels do not trigger more activity related thoughts, but could increase exercise intentions only if presented in pictorial format.


Keywords: priming; exercise; obesity; health; food labels; behavioural intentions

## Introduction

Obesity is a serious global issue, with nearly $40 \%$ of the world's population as overweight or obese in 2014 (World Health Organisation, 2016), a condition that is a major risk factor for noncommunicable diseases such as cardiovascular conditions and stroke (Grover et al., 2015). Physical inactivity has been identified as a main contributor to the obesity epidemic, due to the structure of work and transport become more sedentary in nature (WHO, 2004). People mostly agree explicitly that physical activity is good for health, but in the UK, about $40 \%$ of adults still do not achieve minimum recommended levels of physical activity in their daily life (Craig \& Mindell, 2012). While there may be various reasons for this discrepancy between attitudes and behaviour, closing the gap by increasing physical activity remains an important applied challenge.

The Royal Society for Public Health (RSPH, 2016) recently proposed to target obesity in the UK by the introduction of food labels that reflect the amount of physical activity required to burn off calories in the food. This label is intended to be easier to understand and to nudge people into exercising more. In this project, we assess which aspect of the new label would increase physical activities. We propose some hypotheses about the effect of using the concept of exercise instead of calories to represent food energy, and presenting an image instead of a text. We posit that mentioning exercise instead of calories could act as a semantic prime that increases the accessibility of
activity-related thoughts (Meyer \& Schvaneveldt, 1971) and therefore the intention to engage in physical activities. Compared to text, using an image may also influence the meta-cognitive processes involved in processing the label. Research suggests that the distinctive nature of pictures enhance their recollection over words (Curran \& Doyle, 2011), may generate more interactive information processing (Domke, Perlmutter, \& Spratt, 2002), and are superior to words in activating conceptual understanding (McBride \& Dosher, 2002). Images of a concept are thus more likely to increase its mental accessibility, priming it to be processed more quickly, or more 'fluently' when next encountered.
Concepts that have been activated are subsequently perceived more easily (i.e. more fluent) and are better liked (Winkielman, Schwarz, Reber, \& Fazendeiro, 2000). Factors that increase perceptual fluency, such as improved clarity of presentation, also increase liking of stimuli (Oppenheimer, 2008). Notably, exercise that was presented in a passage with easy-to-read font (high fluency) was estimated to take less effort to perform than when the font was illegible (low fluency) and this influenced people's willingness to engage in it (Song \& Schwarz, 2008).

Manipulating the accessibility of different options has also shown positive effects on healthy related behaviours. Environmental cues can increase the salience of healthy choices and subtly prime people to view these options more favourably (Marteau, 2011; Wilson, Buckley, Buckley, \& Bogomolova, 2016). For example, placing fruit instead of chocolate snacks at supermarket checkouts increased purchase of the healthier food options (Foster et al., 2014). Featuring healthier sandwiches in a more prominent and unhealthier ones in a less prominent spot on a menu made people more likely to order the healthy sandwiches (Wisdom, Downs, \& Loewenstein, 2010).

In sum, variables that increase perceptual fluency (such as familiarity and clarity) as well as variables that increase conceptual fluency (such as exposure to associatively related concepts) can influence the subjective ease with which a stimulus can be processed. We therefore hypothesised that image and activity labels would result in greater mental availability of exercise than text and calorie labels. Additionally, image and activity labels would result in exercise being rated as less effortful and more enjoyable, leading to greater willingness to engage in exercise behaviours than with text and calorie labels.

## Methods

## Participants

One hundred and forty-two English-speaking participants were recruited online through snowball sampling in the UK and Singapore, and sharing through social networking sites and online forums. Age ranged from 18 to 74 years ( $M=$ 40.30, $S D=15.47$ ). Ethnicity was $58 \%$ Asian and $23 \%$ percent white/Caucasian (19\% other races). Body Mass Index (BMI) estimates based on weight and height categories of participants ranged from 15.82 to 43.51 ( $M=$ $22.501, S D=4.66$ ).

## Materials ${ }^{1}$

Food labels We manipulated label format (image vs. text) and energy representation (activity vs. calorie) to create four labels as shown in Figure 1. The image labels were derived from existing activity-label depictions that have been proposed and tested in previous research (e.g. Swartz, Dowray, Braxton, Mihas, \& Viera, 2013; Van Kleef, Van Trijp, Paeps, \& Fernandez-Celemin, 2008). In order to keep image quality consistent, we standardised the walking image across both activity-image and calorie-image labels. Equivalent activity values were calculated from the chosen calorie value based on the mean weight of a UK individual (70kg).


Figure 1: Labels used as stimuli.
Word fragment completion task Participants read 17 word fragments, of which 12 could be completed with either exercise-related or neutral words. For example, the fragment 'S _ ORT' could be completed as SPORT (exercise-related) or $\overline{\text { SHORT }}$ (unrelated). This type of task has previously been used as a measure of mental availability (Tulving, Schacter, \& Stark, 1982). The fragments were chosen to keep the overall word frequency for potential completions as balanced as possible. Each word fragment was presented on a separate page. Participants were told that both speed and accuracy were important in the task.

Perception of exercise We measured participants' perception of exercise as an indicator of exercise favourability. Participants were asked to imagine they had agreed to go for a 5 km walk with their neighbour, and indicate how long (on a slider scale from 0 to 120 minutes)

[^393]and how enjoyable they thought it would be on a 5-point Likert scale (1: not at all, 5: extremely).

Exercise intentions In addition to direct questions about their intentions to exercise (defined as a sustained period of activity of at least 10 minutes) in the next week (likelihood and duration), participants were also given a role-play scenario where they could decide between a sedentary or active option to carry out the scenario task. For example, 'You leave your [2 ${ }^{\text {nd }}$ storey] flat and must go downstairs. You can either take the lift or the stairs. Both are equally accessible from your door. Which would you choose?' Participants' answers in this role-play scenario were scored on a 6-point scale between sedentary and active options (1: sedentary, 6 : active). This was averaged to create an overall score for active choices.

## Procedure

The experiment was delivered online via Qualtrics survey platform. Participants were randomly assigned to view one of the four labels and they answered filler questions about the labels to ensure that they processed it. Then the participants took the word fragment completion task presented as part of a separate study on language ability. After this, they completed the measures for perceptions of exercise, followed by the role-play and direct questions about their exercise intentions. Finally, they provided demographic information, which included scales assessing attitudes towards health (Steptoe, Pollard, \& Wardle, 1995) and exercise (Courneya, Conner, \& Rhodes, 2006).

## Results

The dependent variables in the experiment were analysed in a MANOVA with energy representation and label format as independent variables.

Mental availability of exercise As shown in Figure 2, more exercise-related word completions were observed for activity than calorie labels, $F(1,124)=3.94, p=.049, \eta_{\mathrm{p}}^{2}=$ .03. Label format had no significant effect on number of exercise-related words produced, $F(1,124)=.17, p=.682$, $\eta_{\mathrm{p}}^{2}<.01$. The interaction was not significant, $F(1,124)=$ $.12, p=.730, \eta_{\mathrm{p}}^{2}<.01$.


Figure 2: Number of exercise-related words produced by energy and format of label. Individual points reflect means and $95 \%$ confidence intervals.

Perception of exercise As shown in Figure 3, in the calorie condition, images led to higher walk estimates (suggesting greater perceived effort) than text, but the reverse was true in the activity condition, $F(1,124)=3.94, p=.049, \eta_{\mathrm{p}}^{2}=$ .03. Both main effects were not significant (energy, $F(1$, 124) $=.01, p .941, \eta_{\mathrm{p}}^{2}<.01$; format, $F(1,124)=.37, p=$ $\left..546, \eta_{\mathrm{p}}^{2}<.01\right)$. The MANOVA found no significant effect on enjoyability ratings, $F(3,124)=1.24, p=.300, \eta_{\mathrm{p}}^{2}=.03$.


Figure 3: Perception of exercise effort (in minutes to walk) by energy and format of label. Individual points reflect means and $95 \%$ confidence intervals.

Exercise intentions Participants' estimates of exercise time per day were multiplied by their judgements of exercise likelihood to form a combined measure for intention to exercise. We excluded 15 extreme scores with estimated intention to exercise of 10 hours or more (more than 1.5
standard deviations above the mean.) The resulting distribution is shown in Figure 4.
The analyses showed no significant effect of label conditions on scores for active choice, $F(3,124)=.37, p=$ $.776, \eta_{p}^{2}=.01$. However, we found that energy and format interacted to determine intention to exercise. In the activity condition, choices and intentions to exercise were greater when viewing images, but in the calorie condition, they were higher when viewing text, $F(1,124)=8.14, p=.005$, $\eta_{\mathrm{p}}^{2}=.06$. The individual main effects were not significant (energy, $F(1,124)=.29, p=.591, \eta_{\mathrm{p}}^{2}<.01$; format, $F(1$, 124) $\left.=.77, p=.382, \eta_{\mathrm{p}}^{2}<.01\right)$.


Figure 4: Exercise intentions (likelihood $x$ intended duration) by energy and format of label. Individual points reflect means and $95 \%$ confidence intervals.

We replicated the MANOVA with age, BMI, and attitudes towards health and exercise. The analysis showed that the interaction effect for label format and energy on exercise intentions remained significant, $F(1,117)=6.43, p=.013$, $\eta_{p}^{2}=.05$. The effect of energy on number of exerciserelated word completions and interaction effects for perception of exercise was no longer significant, $F(1,117)=$ $2.75, p=.100, \eta_{\mathrm{p}}^{2}=.03 ; F(1,117)=3.75, p=.055, \eta_{\mathrm{p}}^{2}=$ $.03)$. The effects of the covariates on the dependent variables were also non-significant (age, $F(5,113)=.66, p$ $=.652, \eta_{\mathrm{p}}^{2}=.03$; BMI, $F(5,113)=.28, p=.922, \eta_{\mathrm{p}}^{2}=.01$; health attitude, $F(5,113)=.06, p=.255, \eta_{\mathrm{p}}^{2}=.06$; exercise attitude, $\left.F(5,113)=.04, p=.422, \eta_{\mathrm{p}}^{2}=.04\right)$. We also analysed the effect of ethnicity. While Asian participants were less likely in general to pick active options in the scenario-based tasks, $F(3,112)=3.58, p=.016, \eta_{\mathrm{p}}^{2}=.09$, using Pillai's trace, we found no other main effects or interactions of ethnicity with label format, $F(10,218)=$ $1.16, p=.317, \eta_{\mathrm{p}}^{2}=.05$ or energy representation, $F(10,218)$ $=.97, p=.47, \eta_{\mathrm{p}}^{2}=.04$.

## Discussion

Two aspects of food labels were examined for their ability to increase intentions for exercise: representing food energy value in terms of activity time instead of calories, and using images instead of text. Building on empirical findings that greater perceptual and conceptual fluency from prior exposure increases liking, it was predicted that image and activity labels would prime greater mental availability of exercise and thus increase intentions to exercise. No significant main effects of label format were found, however calorie labels generated more exercise-related word completions. Exercise was perceived as more effortful for calorie labels in image format, but more effortful for activity labels in text format. For activity labels, intentions to exercise were greater for image than text formats, but the opposite was true for calorie labels.
It has been demonstrated that priming effects in word fragment completions occur when semantic information about the prime is retrieved (Smith, 1991), which we hypothesised to be facilitated by the perceptual and conceptual fluency of image and activity labels. However, we found instead a higher number of exercise-related words generated in calorie condition. It therefore seems propitious to consider what concepts are primed by the idea of calories. This term has been used as the current standard to represent energy values on packaged food for several decades (Wartella, Lichtenstein, \& Boon, 2010), and is also frequently associated with dieting and weight loss in popular media and health communications (e.g. Department of Health, 2015). While we assumed that people would have a more intuitive understanding of activity than calories, experience dealing with less intuitive but common conceptual representations may also affect semantic associations. For example, more people associate sugar amounts with grams instead of teaspoons, despite the latter being a more intuitive measurement (Vanderlee, White, Bordes, Hobin, \& Hammond, 2015). We cannot rule out that repeated exposure to calories in the context of exercise allowed participants to automatically generate the concept of exercise from viewing calorie labels.
The finding that format did not affect mental availability or intended behaviours directly, but had different effects depending on whether an activity or calorie label was presented, was unexpected. We speculate that activity and calorie labels have a processing advantage in image and text form respectively. Activity image labels are suggested to be visually and conceptually easier to understand (Campos, Doxey, \& Hammond, 2011), but the relative familiarity of calorie text labels (as opposed to calorie image ones) enhances its ease of processing (Zajonc, 1968). Since both calorie and activity labels showed an ability to prime exercise-related words, it is possible that the relative processing ease of activity image and calorie text labels further increased their fluency over activity text and calorie image labels and therefore the fluency of exercise associated with the label. This would subsequently drive more favourable emotions towards exercise (Song \& Schwarz,
2008) and explain the increased intentions to exercise in these conditions.

Practically, our results replicated previous survey findings of increased intentions to perform exercise after viewing activity labels (RSPH, 2016) for image labels but not for text labels, where calories outperformed activity. Overall, our data indicate that activity labels did not meeting the expectations of the RSPH, albeit with small effect sizes. Nonetheless, we question whether swapping existing calorie-based food labels for activity equivalent ones would be a wise investment. The use of images with the activity labels may generate a greater influence on people's association of energy with exercise over time, since pictures have been shown to improve memory performance over repeated trials more than words (Erdelyi \& Becker, 1974). However it is uncertain how long this effect would need to take hold in the population, or whether it would even surpass the current greater ability of calories than activity to generate exercise associations in text format.
However, before condemning activity labels, it would be prudent to compare activity and calorie labels to a control condition to determine if the mere presence of either label is indeed sufficient to prime exercise concepts and related behaviours. Previous work has indicated that the presence of either label on a menu can reduce energy ordered compared to a no-label condition (James, Adams-Huet, \& Shah, 2015), with activity labels being slightly (but not significantly) more effective than calorie labels. This also suggests that the effects of labels could extend beyond participants' intentions to influence their behaviour, which is not a foregone conclusion from our results. Indeed, a gap between intentions and behaviour often exists, especially for exercise-related behaviours (Sniehotta, Scholz, \& Schwarzer, 2005). The reporting of exercise intentions needs to be interpreted with caution, especially given the possibility that the social desirability of physical activity may prompt participants to overestimate their exercise intentions-although such an effect is likely to be consistent across experimental manipulations. Nevertheless, future research would do well to include measures that provide a more reliable indication of participants' actual exercise and food choice behaviours post exposure to different labels.
Further research should also look to replicate results with a variety of image samples for more robust consideration of the different types of activity images, as well as on a sample with more varied income and literacy levels. In our sample, nearly all respondents had completed tertiary education, which is often an indicator of higher literacy and income (McCoy, 2013). Groups with low literacy or income understand nutrition labels less well and are less familiar with calories, all of which could affect label fluency (for familiarity with calories, see Bleich \& Pollack, 2010; for income, see Rothman et al., 2006, also Viswanathan, Hastak \& Gau, 2009; for literacy, see Signal et al., 2008). Nevertheless, our study demonstrates the importance of psychological research in informing policy decisions and population-level health interventions.

## References

Bleich, S. N., \& Pollack, K. M. (2010). The publics' understanding of daily caloric recommendations and their perceptions of calorie posting in chain restaurants. BMC Public Health, 10, 121.
Campos, S., Doxey, J., \& Hammond, D. (2011). Nutrition labels on pre-packaged foods: A systematic review. Public Health Nutrition, 14, 1496-1506. doi: 10.1017/S1368980010003290

Courneya, K. S., Conner, M., \& Rhodes, R. E. (2006). Effects of different measurement scales on the variability and predictive validity of the "twocomponent" model of the theory of planned behavior in the exercise domain Psychology \& Health, 21, 557-570. doi: 10.1080/14768320500422857
Craig, R., \& Mindell, J. (2012). Health survey for England 2012: Health, social care and lifestyles. London, UK: Health and Social Care Information Centre.
Curran, T., \& Doyle, J. (2011). Picture superiority doubly dissociates the ERP correlates of recollection and familiarity. Journal of Cognitive Neuroscience, 23, 1247-1262. doi: 10.1162/jocn.2010.21464
Department of Health, U. K. (2015). How to diet. from http://www.nhs.uk/Livewell/loseweight/Pages/how-todiet.aspx
Domke, D., Perlmutter, D., \& Spratt, M. (2002). The primes of our times? An examination of the 'power' of visual images. Journalism, 3, 131-159.
Erdelyi, M. H., \& Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. Cognitive Psychology, 6, 159-171. doi: 10.1016/0010-0285(74)90008-5
Foster, G. D., Karpyn, A., Wojtanowski, A. C., Davis, E., Weiss, S., Brensinger, C., et al. (2014). Placement and promotion strategies to increase sales of healthier products in supermarkets in low-income, ethnically diverse neighborhoods: A randomized controlled trial. Am J Clin Nutr, 99, 1359-1368. doi: 10.3945/ajen.113.075572

Grover, S. A., Kauouache, M., Rempel, P., Joseph, L., Dawes, M., Lau, D. C. W., et al. (2015). Years of life lost and healthy life-years lost from diabetes and cardiovascular disease in overweight and obese people: a modelling study. The Lancet Diabetes \& Endocrinology, 3, 114-122.
James, A., Adams-Huet, B., \& Shah, M. (2015). Menu labels displaying the kilocalorie content or the exercise equivalent: Effects on energy ordered and consumed in young adults. American Journal of Health Promotion, 29, 294-302. doi: 10.4278/ajhp.130522-QUAN-267
Marteau, T. (2011). Judging nudging: Can nudging improve population health? British Medical Journal, 342. doi: 10.1136/bmj.d228

McBride, D. M., \& Dosher, A. (2002). A comparison of conscious and automatic memory processes for picture and word stimuli: a process dissociation analysis. Conscious \& Cognition, 11, 423-460.

McCoy, E. (2013). Lost for words: Poor literacy, the hidden issue in child poverty. A policy position paper. London, UK: National Literacy Trust.
Meyer, D. E., \& Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90, 227-234. doi: 10.1037/h0031564
Oppenheimer, D. M. (2008). The secret life of fluency. Trends in Cognitive Science, 12, 237-241. doi: 10.1016/j.tics.2008.02.014

Rothman, R. L., Housam, R., Weiss, H., Davis, D., Gregory, R., Gebretsadik, T., et al. (2006). Patient understanding of food labels: The role of literacy and numeracy. American Journal of Preventative Medicine, 31, 391398. doi: 10.1016/j.amepre.2006.07.025

Royal Society for Public Health, U. K. (2016). Introducing "activity equivalent" calorie labelling to tackle obesity.
Signal, L., Lanumata, T., Robinson, J. A., Tavila, A., Wilton, J., \& Ni Mhurchu, C. (2008). Perceptions of New Zealand nutrition labels by Māori, Pacific and low-income shoppers. Public Health Nutrition, 11. doi: 10.1017/S1368980007001395

Smith, M. C. (1991). On the recruitment of semantic information for word fragment completion: Evidence from bilingual priming. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 234-244. doi: 10.1037/0278-7393.17.2.234
Sniehotta, F. F., Scholz, U., \& Schwarzer, R. (2005). Bridging the intention-behaviour gap: Planning, selfefficacy, and action control in the adoption and maintenance of physical exercise. Psychology and Health, 20, 143-160. doi: 10.1080/08870440512331317670

Song, H., \& Schwarz, N. (2008). If it's hard to read, it's hard to do. Psychological Science, 19, 986-988.
Steptoe, A., Pollard, T. M., \& Wardle, J. (1995). Development of a measure of the motives underlying the selection of food: The food choice questionnaire. Appetite, 25, 267-284. doi: 10.1006/appe.1995.0061
Swartz, J. J., Dowray, S., Braxton, D., Mihas, P., \& Viera, A. J. (2013). Simplifying healthful choices: A qualitative study of a physical activity based nutrition label format. Nutrition Journal, 12, 72. doi: 10.1186/1475-2891-12-72

Tulving, E., Schacter, D. L., \& Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 336342.

Van Kleef, E., Van Trijp, H., Paeps, F., \& FernandezCelemin, L. (2008). Consumer preferences for front-ofpack calories labelling. Public Health Nutrition, 11, 203-213. doi: 10.1017/S1368980007000304
Vanderlee, L., White, C. M., Bordes, I., Hobin, E. P., \& Hammond, D. (2015). The efficacy of sugar labeling formats: Implications for labeling policy. Obesity, 23, 2406-2413. doi: 10.1002/oby. 21316

Viswanathan, M., Hastak, M., \& Gau, R. (2009). Understanding and facilitating the usage of nutritional labels by low-literate consumers. Journal of Public Policy \& Marketing, 28, 135-145. doi: 10.1509/jppm.28.2.135

Wartella, E. A., Lichtenstein, A. H., \& Boon, C. S. (2010). History of nutrition labelling. In E. A. Wartella, A. H. Lichtenstein \& C. S. Boon (Eds.), Front-of-package nutrition rating systems and symbols: Phase I report. Washington, D.C.: National Academies Press.
Wilson, A. L., Buckley, E., Buckley, J. D., \& Bogomolova, S. (2016). Nudging healthier food and beverage choices through salience and priming. Evidence from a systematic review. Food Quality \& Preference, 51, 4764. doi: 10.1016/j.foodqual.2016.02.009

Winkielman, P., Schwarz, N., Reber, R., \& Fazendeiro, T. (2000). Affective and cognitive consequences of visual fluency: When seeing is easy on the mind. In L. M. Scott \& R. Batra (Eds.), Persuasive imagery: A consumer response perspective (pp. 75-89). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Wisdom, J., Downs, J. S., \& Loewenstein, G. (2010). Promoting healthy choices: Information versus convenience. American Economic Journal: Applied Economics, 2, 164-178. doi: 10.1257/app.2.2.164
World Health Organisation, W. (2004). Global strategy on diet, physical activity and health.
World Health Organisation, W. (2016, June). Obesity and overweight: Fact sheet. Retrieved January 17, 2017, from
http://www.who.int/mediacentre/factsheets/fs311/en/
Zajonc, R. B. (1968). Attitudinal effects of mere exposure. Journal of Personality \& Social Psychology Monograph Supplement, 9, 1-27.

# Failure to replicate talker-specific syntactic adaptation 

Linda Liu ${ }^{1}$, Zachary Burchill ${ }^{1}$, Michael K. Tanenhaus ${ }^{1,2}$, T. Florian Jaeger ${ }^{1,2,3}$<br>\{lliu47, zburchil, mtanenha, fjaeger\} @ ur.rochester.edu<br>Department of Brain and Cognitive Sciences ${ }^{1}$, Department of Linguistics ${ }^{2}$, Department of Computer Science ${ }^{3}$,<br>University of Rochester<br>Rochester, NY 14627 USA


#### Abstract

Sentence understanding is affected by recent experience. An important open question is whether this reflects adaptation to the statistics of the input. Support for this hypothesis comes from the recent finding that listeners can simultaneously learn and maintain the syntactic statistics of multiple talkers (Kamide, 2012). We attempt-and fail-to replicate this finding. This calls into questions whether recency effects in sentence processing originate in the same adaptive mechanisms operating during speech perception (for which talker-specific adaptation is well-established).


Keywords: sentence processing; attachment ambiguity; priming; adaptation; talker specificity

## Introduction

Talkers differ in how they realize the same sound categories and words. This inter-talker variability is known to be one of the biggest challenges to speech perception, and has been investigated by a large body of research. This work has identified adaptation as a central mechanism by which listeners overcome inter-talker variability (e.g., Bradlow \& Bent, 2008; Kraljic \& Samuel, 2007). With exposure to a novel talker, listeners seem to be able to adapt their category boundaries to the statistics of the current input (Clayards, Tanenhaus, Aslin, \& Jacobs, 2008; Kleinschmidt \& Jaeger, 2015). This adaptation can persist over time (Eisner \& McQueen, 2006), and listeners can maintain separate adaptations for different talkers (Kraljic \& Samuel, 2007; Trude \& BrownSchmidt, 2012). Findings like these suggest that listeners continuously infer and store talker-specific information about the realization of phonological categories (for review, see Kleinschmidt \& Jaeger, 2015).

Here we focus on adaptation beyond speech perception, and the extent to which it exhibits properties similar to those demonstrated for speech perception. It has long been recognized that talkers also vary in, for example, their lexical and syntactic preferences, and that these differences can impede processing (for a recent review, see Fine, Jaeger, Farmer, \& Qian, 2013). For example, the efficiency of expectationbased processing (i.e., utilizing expectations based on previous input to facilitate integration of bottom-up input) depends on the degree to which comprehenders' expectations-based on previously experienced input-match the statistics of the present input. Thus, if talkers differ in their preferences, and thus the statistics of the input they provide, this should impede processing. On the flip side, expectations that match the statistics of the current input will facilitate processing.

This and related considerations have motivated work on adaptation beyond speech perception, including lexical (e.g., Creel, Aslin, \& Tanenhaus, 2008), prosodic (e.g., Kurumada, Brown, \& Tanenhaus, 2012), and-most relevant to the present purpose-syntactic adaptation (e.g., Fine et al., 2013). For syntactic processing, there is now some evidence that experience with a novel environment (e.g., an experiment) can lead to longer-lasting changes in processing, persisting for at least several days (Wells, Christiansen, Race, Acheson, \& MacDonald, 2009). Further, paralleling studies on adaptation in speech perception, a recent study found that listeners can adapt to multiple talkers simultaneously (Kamide, 2012). As one of the few studies to demonstrate talker-specificity in syntactic adaptation, Kamide's study constitutes an important contribution to our current understanding of whether syntactic adaptation is best understood in terms 'dumb' priming mechanisms (Tooley, 2009) or 'smarter' learning mechanisms (Fine et al., 2013). This result has received a fair amount of attention (with about 10 citations/year); however, as we will discuss later, there are some reasons to interpret the finding with caution.

Here we report efforts to replicate Kamide (2012). We begin by summarizing the original study. Then we introduce our own paradigm, which closely follows this study (we gratefully acknowledge Yuki Kamide's generous willingness to discuss details of her design; errors remain our own). We note here that we made a few changes to the procedure and stimuli of Kamide's paradigm, which we introduce and motivate as we go through our experiment.

## Overview of Kamide (2012)

Kamide (2012) employed a look-and-listen visual world eyetracking paradigm to investigate changes in listeners' syntactic expectations for specific talkers. Participants saw visual scenes like that depicted in Figure 1.

These scenes were paired with sentences containing a syntactic attachment ambiguity. For example, in the sentence "The master of the dog who will bury the treasure is quite old now.", the relative clause, "who will bury the treasure," can describe "the master" or "the dog". These two interpretations correspond to different syntactic parses. The former constitutes high attachment, the latter low attachment.

This attachment ambiguity is assumed to be resolved at "treasure" since MASTER is more likely to be the agent of the BURYING-TREASURE event. Kamide's analyses focuses on the temporarily ambiguous stretch between the introduction


Figure 1: Example experimental scene paired with example (high attachment) sentence, both from Kamide (2012). Underlining marks the syntactically ambiguous sentence region, during which saccades to the target (treasure) and competitor (bone) objects are taken to indicate high vs. low attachment, respectively. (Target and competitor objects are not labeled in actual stimuli shown to participants).
of the relative clause verb and the resolution of the ambiguity ("will bury"). During this stretch, eye-movements to the master or dog can be seen as reflecting expectations that the relative clause will attach high (or low). Specifically, listeners should show more anticipatory eye-movements to referents that are plausible objects for the expected agent. In Figure 1, if high (low) attachment is expected, listeners should show anticipatory eye-movements to the treasure (bone).

This property theoretically makes it possible to investigate changes in syntactic expectations based on recent exposure, including changes in expectations for specific talkers. Specifically, Kamide exposed participants to sentences from three different talkers. One talker always used high attachment sentences, one talker always used low attachment, and the final talker used high and low attachment equally often. The question then is whether participants' eye-movements during the temporarily ambiguous stretch begin to reflect talkerspecific syntactic preferences. If so, participants should be more likely to exhibit anticipatory saccades toward the target object for the two talkers that consistently attached either low or high (cued talkers), compared to the talker who used high and low attachment equally often (uncued talker). This advantage should emerge with increasing exposure throughout the experiment.

Kamide tested this prediction by analyzing the number of trials in which at least one saccade was launched to each object (target vs. competitor) for the cued vs. uncued talkers during each phase of the experiment (beginning [first two blocks] vs. end [last two blocks]). Supporting the hypothesis of talker-specific syntactic adaptation, Kamide identified a statistically significant three-way interaction between object $x$ cue $x$ phase: at the end of the experiment (compared to the start), participants made more saccades to the target for the
cued talkers, compared to the uncued talkers.
However, there are reasons to interpret this result with caution. First, there were relatively few trials with fixations to either the target ( $12.1 \%$, averaged across the beginning and end of the experiment) or the competitor (11.8\%). Second, the looks to the target and the competitor actually decreased over the course of the experiment (from $15.0 \%$ to $9.2 \%$ for the target and $15.6 \%$ to $8.0 \%$ for the competitor). Both of these effects result in a small signal to detect potential effects, which we try to address by using a task-based paradigm.

## Experiment

## Methods

Participants We recruited 24 participants at the University of Rochester (c.f. 48 participants in the original study). Although we had originally intended to run 48 participants, a power simulation (reported below) following 24 participants suggested we would be unlikely to detect an effect, even with the additional participants. All participants were native speakers of English with normal vision. The experiment took about 1 hour and participants were paid $\$ 10$.

Visual Stimuli We used the visual scenes (20 experimental and 40 filler) from Kamide (2012). Each experimental scene consisted of six images arranged against a colored background (e.g., Fig. 1). Two images corresponded to the target and competitor agents of the sentence, and four images corresponded to the potential objects for the sentence.

Sentences We used the original sentences from Kamide (2012), with minor vocabulary adjustments to accommodate American listeners (e.g., motorcycle instead of motorbike). Each scene was paired with four training sentences (A) and four test sentences (B), which involved novel objects referenced in the training sentences. Sentences 1 and 3 are sentences with high attachment interpretations (i.e., in 1 A , the master is the agent of the main verb bury). Sentences 2 and 4 are sentences with low attachment interpretations (i.e., in 2 A , the dog is the agent of the main verb bury).

1. The master of the dog who will (A) bury the treasure /(B) drink the brandy is quite old now.
2. The master of the dog who will (A) bury the bone / (B) drink the water is quite old now.
3. The dog of the pirate who will (A) bury the bone $/(B)$ drink the water is quite old now.
4. The dog of the pirate who will (A) bury the treasure /(B) drink the brandy is quite old now.

Additionally, we used the 40 original filler sentences which did not contain a syntactic ambiguity, and instead contained a relative clause containing an unambiguous single noun phrase antecedent (e.g. "The woman who will lift the pet carrier has never had a dog.").

Auditory Stimuli American listeners might have difficulty distinguishing between the varieties of UK English used in
the original study (conducted in Scotland). We thus recorded new materials from talkers selected to be clearly distinguishable by American listeners. We recorded one Britishaccented male (BM; average sentence duration $=3,668 \mathrm{~ms}$, $\mathrm{SD}=485 \mathrm{~ms}$ ), one British-accented female (BF; $3,944 \mathrm{~ms}$, SD $=430 \mathrm{~ms}$ ), and one Indian-accented Male (IM; $4,358 \mathrm{~ms}$, SD $=516 \mathrm{~ms}$ ). The BM and BF talkers served as the cued talkers, consistently producing either only high attachment sentences or only low attachment sentences. Between participants, we counterbalanced the assignment of high vs. low attachment to talkers, so that each talker served as the high attachment talker for half of participants. Following Kamide, one talker (Talker IM) always served as uncued talker, producing equal amounts of high and low attachment sentences.

All audio recordings were scaled to an average intensity of 70 dB . Following Kamide (2012), the shared portion of each low attachment sentence (up to the relative clause; i.e. "the master of the dog") was then spliced into the corresponding high attachment sentence to ensure that any effects found would not be due to early prosodic cues; i.e. both high and low attachment sentences shared the same low attachment prosody.

Presentation Lists Following Kamide (2012), we created four presentation lists, crossing whether Talker BM and Talker BF served as the high and low attachment talkers, and which items were used in the cued vs. uncued talkers. Each participant was tested on one of these lists.

Each list was divided into six blocks, each consisting of 30 trials (20 experimental, 10 filler). Each experimental scene was presented once per block, with a different sentence each time (i.e. in the first four blocks, sentences 1-4A were heard, and in the final two blocks, either sentences 1-2B or 3-4B were heard). During each block, the uncued talker (IM) produced both five low and five high attachment sentences, and the cued talkers (BM and BF) each produced either five low or five high attachment sentences, along with five filler sentences. Sentence order within the lists was pseudorandomized following the criteria described in Kamide (2012).

Procedure Our procedure closely followed that of Kamide (2012) with the exception that we used a task-based clicking paradigm (described below) as compared to a look-and-listen paradigm. This decision was made because the results of the original study suggest that there was very little signal to detect potential effects. First, in the original study, the percentage of trials in which a subject makes a saccade toward the target or the competitor in the critical window was small (fewer than $24 \%$ ). Second, participants show a (highly significant) decrease in saccades to both target and competitor over the course of the experiment: $15.0 \%$ to $9.2 \%$ for the target object and $15.6 \%$ to $8.0 \%$ for the competitor object.

In a comparison between a look-and-listen paradigm and an explicit task-based paradigm, Altmann and Kamide (1999) find that although results between the two paradigms were qualitatively similar, the task-based paradigm had less-
delayed anticipatory eye movements, and subjects were almost twice as likely to fixate relevant referents (see Salverda \& Tanenhaus, in press).

Expecting more looks to the target and competitor (as taskrelevant objects), we incorporate the task of clicking on the target into Kamide (2012). This carries the additional benefit of allowing us to test whether participants are paying attention during the experiment (see Results).

Participants were seated in front of a 19 inch computer screen. Eye movements were monitored using an Eyelink 1000 , sampling at 500 Hz from the left eye. Recalibration was performed at the beginning of each new block (every 30 trials), and drift correction was performed every five trials.

At the start of each trial, participants saw a task prompt against a white background on the screen, e.g. "Click on what will be buried." Crucially, the object that should be clicked on varied between high and low attachment sentences (i.e. treasure or bone). After the participant clicked to verify that they had read the prompt, a fixation cross appeared in the center of the screen. After the participant clicked on the fixation cross, the visual scene appeared on the screen. The sentence played over loudspeakers after the scene preview period of 1000 ms elapsed. The trial ended after participants clicked in the scene after the audio file ended.

## Results

We begin by assessing how participants engaged in the experimental task. First, we test whether participants performed the task as intended. Second, we assess whether the taskbased paradigm caused participants to have more saccades to the task-relevant referents during the critical time window, thereby increasing our power to detect an effect. Third, we test whether the number of trials containing saccades to target, compared to competitor, objects decreases over the course of the experiment (as reported in the original study).

All analyses presented below include the maximal random effects justified by the design (i.e. by-participant and by-item intercepts and slopes for cue, phase, and their interaction). We treat sentences sharing the same prefix up to the relative clause as one item (e.g., 1A-B and 2A-B form an item), such that cue, phase, and their interaction varied within items. We continue to use the term target object to refer to the object that is consistent with the intended attachment interpretation (e.g., the treasure in 1A), and the term competitor object to refer to the object consistent with the unintended attachment (e.g., the bone in 1A).

Task performance: Click responses We find that participants overwhelmingly ( $98.3 \%$ of all trials) clicked on the correct target object (e.g., treasure in Fig. 1). This suggests that participants understood and successfully performed the experimental task. One scene showed a high incorrect response rate, with over $40 \%$ of subjects ( 10 of 24 ) clicking on the wrong object. This was likely due to the use of target label (hat) that is also a valid label for the competitor (cap). Trials involving this scene, and all trials with incorrect clicks, are
excluded from subsequent analyses.
Task performance: Saccades to task-relevant referents Next, we analyze the number of trials containing looks to task-relevant objects (i.e., either of targets and competitors) and the changes in this number over the course of the experiment. We analyze the same time-window as in the original study: from the start of the onset of the verb of the relative clause to the onset of the object of the relative clause (e.g. "bury the" in Fig. 1). The mean duration for this timewindow is 371 ms (cued; $\mathrm{SD}=66 \mathrm{~ms}$ ) and 430 ms (uncued; $\mathrm{SD}=82 \mathrm{~ms}$ ). These durations are somewhat different from, but similar to, the corresponding time-windows in the original study (cued: 528 ms , uncued: 365 ms ).

Compared to Kamide, we find nearly twice as many trials containing saccades to the target $(24.1 \%$ here vs. $12.1 \%$ in the original) or competitor objects ( $22.6 \%$ here vs. $11.8 \%$ in the original), averaged over the first and last two blocks. Like in the original study, we find that the overall probability of saccades to task-relevant objects decreases over the course of the experiment. However, this decrease was less dramatic than in the original study. In the first two blocks, $46.8 \%$ of all trials contained saccades to task-relevant referents and only decreased to $44.6 \%$ of all trials in the last two blocks. In fact, the number of trials in which there were saccades to the target object actually increased slightly, going from $22.2 \%$ in the first two blocks to $25.9 \%$ in the last two. This constitutes an increase of 0.2 in log-odds for the target object (compared to a decrease of 0.56 in the original study) and a decrease of 0.28 in log-odds for the competitor object (compared to a decrease of 0.75 in the original).

In summary, the task-based paradigm increases saccades to the task-relevant referents, as intended. This means that our estimates of the relative proportions of eye-movements to target vs. competitor objects-the dependent variable for the main analyses presented below-are based on more data. This should result in more reliable statistical signal for the main analysis (but see our discussion for caveats).

Main analysis: Anticipatory saccades during original time window We first analyze saccades during the time window employed in the original study. Failing to find evidence for talker-specific learning, we then present additional post-hoc analyses. Specifically, we extend our analyses to other time windows, and to individual cued talkers. All our analyses analyze the proportion (of eye-tracking samples with) fixations to the target vs. competitor objects during the time window. This differs from Kamide's, who analyzed the number of trials containing at least one saccade to the target vs. competitor during the time window. The two measures yield the same conclusions for the present data (see below). We chose our analysis approach, because Kamide's measure resulted in a model that converged only under a drastically reduced maximal random effects structure (byparticipant random intercepts and slopes, and by-item random intercepts). We also note that our approach is known


Figure 2: Proportion of fixations over the course of a trial. Colors show the different referents in the scene. Panels show the cued and uncued talkers. Regions A, B, and C correspond to the time windows in which participants heard the equivalents to "The master of the dog", "who will", "bury the", respectively. Kamide (2012)'s analysis was restricted to region C , as was our original analysis. We also present post-hoc analyses of regions A and B+C. For this plot, we normalized times within each region (essentially aligning all trials) before averaging across trials and subject.
to be anti-conservative (because of auto-correlations between eye-tracking samples), but not known to be underpowered.

The full time course for fixations over trials is shown in Fig. 2. Following Kamide (2012), we compare the data collected during the first two blocks (beginning) to the data collected during the final two blocks (end) during the time window between the start of the verb of the relative clause to the noun (e.g. "bury the").

We perform mixed logistic regression predicting fixations to target versus competitor from cue (sum-coded: cued $=1$ vs. uncued $=-1$ ) and phase (sum-coded: end $=1$ vs. beginning $=-1$ ). If participants are able to learn the syntactic preferences of specific talkers, then we expect more looks to the target object during the end of the experiment compared to the beginning. We would only expect this for the cued talkers, in which the talkers produced high or low attachment sentences consistently, and not for the uncued talker, in which the talker produced a mixture of both high and low attachment sentences. The critical result to replicate is the interaction between cue and phase of the experiment.

None of the factors reached significance ( $p s>0.14$ ). The analysis is summarized in Table 1 (penultimate column).

Critically, we did not find a significant interaction between cue and phase ( $p=0.8$ ). Using the reduced model based on Kamide's measure, where the critical measure is a 3-way interaction between object $x$ cue $x$ phase, confirmed this pattern. Specifically, Kamide's measure returned a significant two-way interaction between object and cue in the unexpected direction ( $\hat{\beta}=-0.08, p<0.05$ ). The critical 3-way interaction trended in the same direction as in the original study, but did not reach significance ( $p=0.11$ ).

For the remaining analyses we only present analyses over proportions of fixations. Using Kamide's measure led to models with reduced random effect structures, but did not change the results.

Post-hoc analyses over additional time windows We entertain two possibilities for the observed null effect. First, it is possible that the introduction of a task-based paradigm affected the time course of eye-movements (see e.g., Altmann \& Kamide, 1999). Specifically, in our paradigm listeners saw the verb prior to hearing the sentence, which may have allowed them trigger anticipatory eye-movements even earlier, before the onset of the RC verb. We therefore conduct posthoc analyses over two additional time windows (see Figure 2):

1. Start of the sentence to start of the relative clause (SentenceStart to RCStart; e.g. "The dog of the pirate")
2. Start of the relative clause to start of the relative clause noun (RCStart to RCNounStart; e.g. "who will bury the")
3. Start of the relative clause verb to start of the relative clause noun (RCVerbStart to RCNounStart; e.g. "bury the")

A second possibility is that participants look towards the agent, rather than object, of the verb. We thus also analyze looks to the target versus competitor agents. For example, in the low attachment sentence ( 1 A ), the target agent is the dog, and the competitor agent is the master. We do so also for the original time window analyzed in the previous section.

This results in 6 analyses ( 3 time windows by 2 ways to operationalize target vs. competitor), of which one is the original analysis. The results from all analyses are presented in Table 1 . None of the critical effects reach significance in any of the analyses. However, for the agents, we identified marginal interactions of cue and phase in two of the time windows (Table 1); however, both were numerically in the opposite of the expected direction. Following Kamide, we also performed each of these analyses for trials containing saccades to target or competitor (with a reducted random effects structure). None of the additional analyses reached significance in the predicted direction (predicted direction: $p>0.7$ ), though the agents analysis for region C was significant in the unpredicted direction ( $p<0.05$ ).

We did not find a reliable effect of talker-specific syntactic adaptation between the cued and uncued talkers, even when considering post-hoc analyses over additional time-windows. All remaining analyses therefore use the same time windows as the original study.

So far, following Kamide (2012), all our analyses compared eye-movements for cued vs. uncued talkers. Such analyses group the two cued talkers together, ignoring potential differences in listeners' expectations for a male and a female (British English) talker. We conducted several followup analyses, comparing eye-movements in response to these two cued talkers. We found no evidence that participants were biased toward expecting more high (or low) attachment for either talker. Nor did we find any evidence that participants adapted more for one talker over the other.

## General Discussion

In contrast to the growing literature on talker-specific adaptation in speech perception, one of the few studies that has directly addressed talker-specific syntactic adaptation is (Kamide, 2012). Here we were unable to replicate the original study's findings. We are aware of one other ongoing effort within the same paradigm as the original study, albeit with different stimuli (ongoing work by Ryskin, Fine, \& BrownSchmidt). Ryskin and colleagues do not find evidence for talker-specific syntactic adaptation over the original time window, while finding weak evidence over a larger time window.

One possibility why the present experiment failed to find talker-specific syntactic adaptation is that it had less power than Kamide's original study: we used fewer participants (24 here vs. 48 in the original study). On the other hand, presumably because of the use of a task-based paradigm in the present experiment, the proportion of trials containing a look to either the target or competitor was almost twice as high in the present study than in the original study ( $42 \%$ in the present study vs. $<24 \%$ in the original study), which should increase the power. Power analyses suggested that power in our experiment was slightly lower than, but comparable to, Kamide's. For example, even when assuming an increase of 0.6 in log-odds in looks to the target in the cued vs. uncued condition from beginning to end of the experiment (constituting a 9 -fold increase in effect size), the present approach (24 participants, $40 \%$ looks to task-relevant objects) yields power of $69.4 \%$, compared to $82.5 \%$ for Kamide's study ( 48 participants, $20 \%$ looks to task-relevant objects).

Another possibility is that limited familiarity with the British- and Indian-accented English could have affected participants' ability to understand our talkers. If participants were unable to parse our talkers' sentences, this would make it impossible to learn talker-specific syntactic statistics. We note, however, that participants almost always clicked on the correct object ( $98.3 \%$ of all trials), indicating that they were able to successfully understand talkers. Additionally, the three talkers were easily distinguishable, which would be expected to support the learning of talker-specific statistics.

Additionally, it might be that our task encouraged participants to listen for the correct object without processing the syntactic structure. Participants were asked to click on the object of the verb; the verb was always provided at the very end of the sentence, and participants were not allowed to click un-

Table 1: Output of mixed effects logistic regression comparing fixations to target versus competitor objects and agents (positive coefficient estimates indicate more looks to the target object). Marginal effects in italics. No factors reached significance.

| Predictors | SentenceStart to RCStart |  |  |  | RCStart to RCNounStart |  |  |  | RCVerbStart to RCNounStart |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Object |  | Agents |  | Objects |  | Agents |  | Objects |  | Agents |  |
|  | $\hat{\beta}$ | $p$ | $\hat{\beta}$ | $p$ | $\hat{\beta}$ | $p$ | $\hat{\beta}$ | $p$ | $\hat{\beta}$ | $p$ | $\hat{\beta}$ | $p$ |
| Intercept | 0.49 | 0.11 | -0.07 | 0.37 | 0.09 | 0.57 | 0.17 | 0.06 | 0.23 | 0.30 | 0.27 | 0.07 |
| Cue (Cued=1 vs Uncued=-1) | 0.27 | 0.33 | -0.04 | 0.56 | -0.17 | 0.30 | -0.06 | 0.56 | -0.22 | 0.35 | -0.08 | 0.63 |
| Phase (End=1 vs Beginning=-1) | -0.12 | 0.71 | 0.00 | 0.97 | 0.20 | 0.25 | 0.14 | 0.15 | 0.42 | 0.14 | 0.14 | 0.30 |
| Cue:Phase | -0.27 | 0.45 | 0.04 | 0.57 | 0.01 | 0.93 | -0.16 | 0.10 | 0.07 | 0.80 | -0.25 | 0.08 |

til then. This meant that participants did not derive any time benefit from correctly tracking the talker's attachment preferences. However, we note that in the original experiment, as participants were asked only to look at the scene while listening to the sentence, there was also no explicit incentive to track the talker's attachment preferences.

An alternate possibility is that listeners do not exhibit talker-specific adaptation to attachment structure because they are not capable of it. Why might this be the case, particularly given the growing body of evidence for talker-specific phonetic adaptation? As talkers vary systematically in how they sound, both due to physiological factors (e.g. vocal tract size), and sociolinguistic factors (e.g. regional dialect), phonetic adaptation to talker-specific pronunciations can lead to a large benefit in later encounters with that same talker (Bradlow \& Bent, 2008).

By contrast, syntactic adaptation, particularly syntactic adaptation to high and low attachment structure, may not carry the same utility: subject relative clauses alone occur fewer on than $2 \%$ of all noun phrases (Roland, Dick, \& Elman, 2007). Non-local subject relative clauses with two animate (potential) heads-like those used in our experimentare even rarer. Thus, talker-specific syntactic adaptation to such a low-frequency structure may not lead to the same benefits of improved language understanding as phonetic adaptation to much more frequent phonemes. It may be beneficial to address questions of talker-specific adaptation by examining more frequent syntactic structures.

In sum, the current experiment failed to replicate the effect that listeners are able to track talker-specific syntactic preferences for attachment (Kamide, 2012). This raises questions about the extent to which syntactic adaptation may involve the same adaptive mechanisms used in phonetic adaptation.

## Acknowledgments

We thank Yuki Kamide for generously providing her visual stimuli and sentences, and for helpful discussion. We thank Anne Pier Salverda for technical and design advice, and members of HLP and KurTan lab at Rochester for feedback. This work was funded by NIH R01 HD075797 to TFJ. Views expressed here do not necessarily reflect those of the NIH.

## References

Altmann, G. T., \& Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3), 247-264.
Bradlow, A. R., \& Bent, T. (2008). Perceptual adaptation to non-native speech. Cognition, 106(2), 707-729.
Clayards, M., Tanenhaus, M. K., Aslin, R. N., \& Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3), 804-809.
Creel, S. C., Aslin, R. N., \& Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. Cognition, 106(2), 633-664.
Eisner, F., \& McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. J. Acoust. Soc. Am., 119(4).
Fine, A., Jaeger, T. F., Farmer, T., \& Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. PloS one, 8(10).
Kamide, Y. (2012). Learning individual talkers structural preferences. Cognition, 124(1), 66-71.
Kleinschmidt, D. F., \& Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. Psychol. Rev., 122(2), 148.
Kraljic, T., \& Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. J. Mem. Lang., 56(1), 1-15.
Kurumada, C., Brown, M., \& Tanenhaus, M. (2012). Prosody and pragmatic inference: It looks like speech adaptation. In Proc. 34th ann. conf. cognitive science society.
Roland, D., Dick, F., \& Elman, J. L. (2007). Frequency of basic english grammatical structures: A corpus analysis. JML, 57(3), 348-379.
Salverda, A. P., \& Tanenhaus, M. K. (in press). The visual world paradigm. In A. de Groot P. Hagoort (Ed.), Research methods in psychology. John Wiley \& Sons.
Tooley, K. (2009). Is syntactic priming in sentence comprehension really just implicit learning. In Proc. 22nd cuny.
Trude, A. M., \& Brown-Schmidt, S. (2012). Talkerspecific perceptual adaptation during online speech perception. Lang. Cognitive. Proc., 27(7-8), 979-1001.
Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., \& MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. Cognitive Psychol., 58(2), 250-271.

# Instrumental Representations of Sensorimotor Control: Representations at Intermediate Level 

Hsi-wen Daniel Liu (hwliu@pu.edu.tw)<br>Center for General Education, Providence University<br>Taichung, Taiwan


#### Abstract

In cognitive science, computation is largely accompanied with representational theory of mind. Yet, it remains unclear whether this companionship also appears in the realm of sensorimotor control. Grush's (2004) and Pezzulo's (2008, 2011) account of anticipatory representations provide a limited answer, as they are only suitable for forward models, but not the entire sensorimotor control. Rescorla's (2016) representational explanation for sensorimotor psychology addresses several intentional states considered in Bayesian inference and optimal modeling. However, the above accounts does not explain how motor commands are produced and chosen in the course of sensorimotor control for maintaining accuracy of goal-achievement. The present paper aims to explain it with a representational account by considering instrumental representations of sensorimotor control, which appear at the intermediate level and are exemplified by motor commands and costs. Such representations do not presume decouplability, as they need to run on-line in the maintenance of accuracy.


Keywords: Sensorimotor control; representation; optimal feedback control; Bayesian decision theory.

## Introduction

Within cognitive science there is a long-standing dispute between different paradigms concerning the role of representation in computation. Classical cognitive science understands cognition in terms of computation over mental representations, and considers the role of cognition to be deriving world-models that provide a database for thinking, planning and problem-solving. Decouplability between representations and their immediate environment is taken as intrinsic to representation. By contrast, a 'pragmatic turn' raises the 'action-oriented paradigm' that considers cognition to be providing skillful know-how in situated and embodied actions (Engel et. el., 2013). Clark (1997) raises the notion of action-oriented representations, which do not presume decouplability, in his 'minimal representationalism'. Action-oriented representations, yet, are mostly applied to reactive motor activities, regardless of various models of motor control.

The computational theory of mind is largely accompanied with a representational theory for explaining a variety of faculties, including perception, language, thinking, and problem-solving. However, the role of representation is highly debated for the faculty of motor action. Within models of motor control, the forward model is firmly associated with a role of representation, in the notions of emulating representation or anticipatory representations (Grush, 2004; Pezzulo, 2008, 2011). For those
representations, decouplability is claimed to be an intrinsic property. In addition, Bayesian models of motor control are seen as highly associated with a robust notion of mental representation (Rescorla, 2016). Decouplability is also seen as intrinsic to that notion of representation (Haselager et al., 2003). As to other models of motor control, however, the role of representation is unclear. The computational theory of sensorimotor control has been established (Wolpert and Ghahramani, 2000; Franklin and Wolpert, 2011; Orbán and Wolpert, 2011). It is suggested that motor control is conformed to the notion of pragmatic representation, as different from that of semantic representation (Jeannerod, 2006). Yet, questions remain as to what the pragmatic representation is and how it is related to computational models of motor control.

Given that computation of cognition is an explanatory account of cognition (Marr, 1982), an explanation can be found in the computational theory of sensorimotor control, as shown above. The sensorimotor control, as seen in the determination of an appropriate motor command, is considered to be fundamentally a decision process (Körding and Wolpert, 2006). The decision is to choose an appropriate motor command in order to achieve a given goal. The present paper contends that the explanation of the sensorimotor control is to explain why and how the decision could determine an appropriate motor command that turns out to achieve the goal. The explanandum is the way in which the goal is achieved by choosing an appropriate of motor command. The present paper characterizes this way of goal-achievement in terms of the end-means relation. The goal of a motor task is the end, and a chosen motor command is its means. Humans can rationally consider the appropriate means for an end; similarly, the sensorimotor system can choose an appropriate motor command for a given goal. This similarity, yet, is subject to two caveats. Firstly, while in economic decision-making or daily affairs in general, the rationality proceeds at the personal-level, the sensorimotor processes largely operate at subpersonal-level. Secondly, while characterization of human rationality need not be put in terms of probability, the actual performance of sensorimotor control is found to be very close to descriptions made with Bayesian decision theory (Körding and Wolpert, 2006), a theory with probabilistic measures.

Explanation of sensorimotor control can be pursued in anti-representationalist accounts, which hold certain explanatory perspectives (Turvey and Fonseca, 2009). Motor control, as contended in such accounts, is determined by interactions between the neural system, body, and the environment. Different from such accounts, the present
paper preserves an explanatory role for representation in terms of a novel notion of representation-instrumental representation-and regards the above interactions as complementary to computation and representation. Arrangement of that complementarity is for two reasons. One is the overall success of the computational approach to sensorimotor control (Franklin and Wolpert, 2011; Orbán and Wolpert, 2011; Todorov, 2004; Wolpert and Ghahramani, 2000). The other reason is that problems of sensorimotor control, to put it in Clark's (1997) terms, are not 'representational-hungry'. The neural system of sensorimotor control has tight interactions with body, and the environment. The notion of instrumental representation differs from the mainstream conceptions of representation in that it does not presume decouplability from the environment. Putting above two reasons together has an implication in the level of instrumental representations. Like that the present account of representation stands between the classic account of representation and those antirepresentationalist accounts, instrumental representations can be conceived of as standing at the intermediate level of the mind.

The present paper aims to raise a representational account of the sensorimotor control. This aim is to be achieved by explaining computation of sensorimotor control in terms of representation, on grounds that computation is a way to explain the mind, in the first place. Section two discusses computational explanations of sensorimotor control in terms of end-means relations. Section three specifies the notion of instrumental representations for explaining the sensorimotor control, on grounds of computational explanations of sensorimotor control.

## Computation in the Sensorimotor Control

The Bayesian decision theory in sensorimotor control holds a computational perspective with two components-estimation of environmental and bodily conditions, on the one hand, and decision made upon motor commands for the most desirable performance, on the other. To put it in an epistemological dichotomy, the former component is to measure environmental and bodily facts, while the latter one is to evaluate motor actions. To put it otherwise, the former is close to perception, while the latter to decision-making.

## Uncertainty

Measurement of states in the sensorimotor system is affected with various factors of uncertainty, and consequently it cannot be accurate like that we manage to measure the length of an object left on the table with a ruler. Sensory signals of the environment have inherent delays, which affect signals at all stages of sensorimotor system from the afferent (coming-in-from-the-outside) sensory information, to conduction along the neural fibers, together with the complexity of processing (face recognition being longer than motion perception) and 'slower' modality (vision being longer than proprioception). It can be said that we 'live in the past' by accessing the 'out-of-date
information' about the common world and our own bodies (Franklin and Wolpert, 2011, pp. 425-6). In addition, the nervous system is corrupted by noise, which also affects sensorimotor control at all stages, from the reception of sensory information, to planning, resulting in variability in movement endpoints. Noise, hence, contaminates our observation of the sensorimotor system internally and externally, by affecting estimation of body states and world conditions (ibid., p. 425). Noise in motor commands, in particular, increases propositionally to the size of their signal (the so-called 'signal-dependent noise'). Different motor commands would incur different degree of variability in the resulting endpoints. To put it more specifically, the motor performance is subject to speed-accuracy trade-off described by Fitt's law (Harris and Wolpert, 1998). Noise in the sensorimotor system corrupts not only estimation of internal and external states, on the one hand, but also performance of motor actions, on the other.

Apart from delay and noise, there are more factors of uncertainty residing in the sensorimotor system. Environmental conditions are constantly subject to change, for example, forces imposed upon the arm in the reaching movement (Shadmehr and Mussa-Ivaldi, 1994). It is also uncertain as to which actions or tasks would be beneficial in the real world outside the laboratory (Franklin and Wolpert, 2011). Furthermore, our motor system is non-stationary, for example, the length and weight of our limbs are changing when we grow up, and our muscles are getting stronger with larger forces (ibid.). Those uncertain factors would constantly contaminate information of the sensorimotor system, information which consequently needs to be optimized in order to achieve the given goal.

Given the aforementioned factors of uncertainty, the computation of sensorimotor system needs to optimize its information, in order to find the best resolution in view of goal-achievement. The sensorimotor system should maintain optimal estimation of world and body states, and should conduct optimal choice of motor commands. This need of optimality, in both state estimation and action choice, introduces a version of instrumental rationality that is immanent in the sensorimotor system at the subpersonallevel.

## Decision on Grounds of Utility

By contrast, another version of instrumental rationality turns up in the sensorimotor system's course of decision-making for optimal choice of a particular motor command, given a particular goal. Various particular goals can be given to a sensorimotor system, which are ends for the system to seek appropriate means for their fulfillment. Whenever a means is determined, a decision is made for attaining the end. The degree of fulfillment can be evaluated in positive terms, such as benefit, reward, utility or prospect; or alternatively in negative terms, such as loss or cost (Körding and Wolpert, 2006). The degree of fulfillment is a foundational notion, the present paper proposes, for explaining the sensorimotor control. In the realm of sensorimotor control, the utility of a
motor movement is the decision theory's way of evaluating the fulfillment of the goal. The fulfillment to a higher degree would receive the evaluation of a higher utility.

According to the decision theory, the expected utility of an action is defined quantitatively in terms of probability, as follows:

$$
E[\text { Utility }]=\sum_{\substack{\text { possible } \\ \text { outcomes }}} p(\text { outcome } \mid \text { action }) U(\text { outcome })
$$

where $p$ (outcome | action) is the probability of an outcome given an action, and $U$ (outcome) is the utility assigned to that outcome. An action is chosen for maximizing the expected value, which is put in terms of utility. A decision made in this way is defined to be a rational choice (Körding and Wolpert, 2006). This is a normative theory that defines the way in which people should behave; that way is put in terms of rationality. In other words, the rationality of action urges that people in their actions should pursue a higher degree of utility. When it is put in the context of sensorimotor control, a utility function evaluates how well a movement is performed. This way of evaluation quantifies, in terms of utility, the total desirability of a chosen movement. In addition, the decision theory is also considered to be a descriptive theory by assuming that people act rationally. The rationality assumed in this theory explains why people behave in the way they do. In fact, empirical findings indicate that the Bayesian decision theory shows successfully how people actually perform their sensorimotor control (ibid.).

The decision theory measures a motor movement, in terms of utility (measuring positively) or cost (measuring negatively), by evaluating its degree of fulfillment, that is, how well it achieves the goal. The end-means relation is assumed in the notion of fulfillment. The utility, in a descriptive term, defines how well a means attains the end.

The measurement of a cost depends on the immediate conditions of all relevant factors in the sensorimotor processing. Specifically, it depends on current states of the sensory system and properties of the to-be-chosen motor commands. The considered states include body states and the environmental conditions, for example, joint angles and velocities, and positions of relevant objects. Properties of motor commands can be measured with different emphases, for example, the jerk, torque change, energy, time, variance, of the to-be-chosen motor commands. The measurement, whatever the emphasis, takes the form of 'minimize X ' (Todorov, 2004). It is to minimize the size of relating factors, for example, the energy-to-be-consumed of the motor commands. In other words, the measurement of cost is sensitive to the size of the system's immediate response to its (bodily and environmental) conditions, which are embodied properties (i.e. jerk, torque change, energy, time, variance, etc.) of the sensorimotor system. To be noted, measuring the cost is not purely an internal matter, but is to be put in real situations, which consist of bodily and environmental conditions.

## The Basics of Motor Commands

Choice of motor commands, as above considered, is to be managed after a series of motor commands is organized. The computation of sensorimotor control is required to explain how to transform a higher-level goal into a series of motor commands, which are strictly constrained in the embodied sensorimotor system. This explanation consists of two parts: coordinate transformation and modular structure, both of which assume the end-means relationship. The former, coordinate transformation, is to convert sensory signals of the goal into motor commands. Sensory signals consist of visual information of the object in the goal-state together with the signals relating to the posture of bodily parts (hands, arms, shoulders, head, and eyes). Those signals need to be transformed into a set of motor commands that would bring about the goal-state when they are performed. This task is named to be sensorimotor transformations, which are accomplished with the mapping of a three-layered neural network: from the input layer of posture signals, to the intermediate layer that consists of population codes, to the output layer of the motor command that consists of the change in joint angles needed for the task (Pouget and Snyder, 2000). This is a way of reverse engineering, which is called the inverse model, as its direction of transformation is opposite to the forward model of motor control (Wolpert and Ghahramani, 2000). The sensorimotor transformations and their products can be regarded as means for the end of achieving the given goal.

As for the construction of movements, it remains controversial as to how much of movement might be controlled by modular processes (Zelik et. el., 2014). Insofar as modular organization is applicable, various complex motor movements are constructed through flexible combinations of a limited number of modules, in order to simplify computation by reducing degrees of freedom (Jing et. el., 2004). In other words, a complex motor command is organized with a combination of motor primitives. A motor command consists of a series of muscle activations for the needed changes of joint angles. With a study in the vertebrate spinal cord, it is shown that a complex motor command is produced by combining a few motor primitives, which are 'unit burst generators' organized in the spinal cord. Each burst generator is to control the activation of a small group of synergistic muscles, or motor synergies (Tresch et. el., 1999). Such a combination is a modular representation (Mussa-Ivaldi, 1999). A motor task is to produce a motor command with an appropriate modular structure. The motor commands produced in the above way are basic elements for the choice of cost function (Wolpert and Ghahramani, 2000). Given that a single motor command has the above modular representation, it would naturally be questioned as to how the series of motor commands leading to the achievement of a goal are organized.

The notion of modular structure appearing in sensorimotor control does not strictly follow Fodor's (1983) sense of modularity. Firstly, Fodor (1975) argues that
mental modules are combined with a language of thought (LOT, Mentalese). The modular structure of movement generation, however, does not seem to follow the structure of LOT, in the following two aspects. The weighted and graded combination of modules (Jing and Weiss, 2005) does not show the structure of LOT. Furthermore, generation of movements out of modular organization has practical limitations, as is found in generation of movements for diverse locomotor behaviors; sufficient flexibility needs a further basis of coordination (Zelik, 2014). There may be no clear distinction between planning and execution, because coordinated motor movements may emerge out of real-time optimal feedback control (Todorov and Jordan, 2002), as discussed below. Emergence of coordinated movements out of interaction with the environment makes generation of a movement deviant from the modular organization. To summarize, the modular organization is only loosely applicable in sensorimotor control.

## Coordination During Execution

The sensorimotor control on the basis of decision, cost and optimality, as aforementioned, can be managed, to a certain degree, in abstraction from the real situations. A version of optimality can be so pursued, by way of open-loop optimization, with detailed planning in advance of execution. The accuracy can be maintained to a certain degree, yet with serious limitations. The application of optimal principles seeks average optimality over previous performance. As it is detached from the real situations, the sensorimotor control is like playing "a prerecorded movement tap" and consequently the given goal is treated like a laboratory task. It would be unable to encounter the trial-to-trial variability in the real situations (Todorov, 2004, p. 2).

Such an abstract way of sensorimotor control is rather like the maintenance of thought, as it can run in abstraction from the real environment. It has the merit of a Popperian creature, that is, planning internally for a best solution of the considered problem before its execution. Decision can be made for a relatively optimal performance. Yet, an important way of sensorimotor control would be completely missing-coordination during execution. Understanding how this is done is a central problem in motor control for nearly 70 years (Todorov and Jordan, 2002).

The coordination during execution presumes optimal feedback control-the optimal control with on-line sensory feedbacks. It does not plan a desired trajectory before execution, but maintain the coordination on-line in response to all the task-specific contingencies in the real situations. Coordination in the sensory system is highly important as such a system is highly redundant, with a high number of ways over the combination of motor activations, and full of a variety of uncertain factors such as noise and delay, as aforementioned. The optimal feedback control produce "continuous trajectory of movement in response to contiguous stream of sensory input (Körding and Wolpert, 2006)". Costs are continuously generated with on-line control of sensory feedbacks. The evaluation of cost is put in terms of 'cost-
to-go'- the continuous and integral summation of costs (Todorov, 2004). The optimal feedback control responds to real situations of the body and the environment, and fully manifests the continuous way of motor decision in fastchanging conditions. This cannot be done in a detached model.

When the optimal feedback control operates as a way of coordination, motor synergies and the achievement of the given goal emerges. It only asserts what to achieve, without dealing with the how question in detailed. After the goal is given, the optimal feedback control can keep on seeking an appropriate resolution because of its coupling with the plant, in a way like the operation in the dynamic systems view. The stages of planning and execution are not separate (Todorov and Jordan, 2002).

The success of the coordination, in the optimal feedback control, relies on a normative property of the end-means relation, which is immanent in the sensorimotor system. That is, the sensorimotor system as a system with the endmeans relation would seek appropriate means for its given end. This property of the optimal feedback control can be seen as endowed in evolution. It is a process that makes possible the emergence of coordination in the sensorimotor system. After it encounters its contingencies, including the fast-changing environment (with body) with various uncertain factors, the sensorimotor system operates like the way present in the dynamic systems view. In that dynamic relations, motor synergies will eventually emerge.

## Instrumental Representations

Representations in the sensorimotor system are generally divided into three types: end, means, and cost. The goals in the sensorimotor system are regarded as ends, and motor commands are means for attaining those ends. An end represents a world state that is to be attained. A common element in the latter two types is the end-means relation: a means can attain its end, but therein lies a certain cost; in this sense, the means and the cost are called 'instrumental representations'. The means are first-order representations, which represent ways to attain their relating ends. Costs, by contrast, are second-order representations that represent prospects of the relating means in the processes of attaining their ends. Furthermore, those three types of representations hold different foundations of representation qua representation. An end represents a to-be-attained future state, in which the end refers to the to-be-attained future state on grounds of similarity. By contrast, the means represents the to-be-attained end in the way that the end is to be attained through the means. The cost, in addition, evaluates the prospect of a means in the way to attain its end. Finally, whether a means represents an end successfully, is measured in terms of accuracy (as opposed to truth), that is, accuracy with which the means attains the end. As accuracy can be measured with various degree, misrepresentation is subject to different degree.

## Representations as Stand-Ins

Representation, generally speaking, is something $R$ that stands in a system $S$ for something else $E$. That is, $R$ is a surrogate in the system $S$ for $E$. Representation sometimes serves as a Poppian creature: something that can run internally in a system before it is actually carried out. Representation in this sense simulates what will actually happen. It $(R)$ is a surrogate of $E$ 's actual performance. Based on Cummins (1996), a surrogate in this sense refers to its target with the informatiom of its content. A city map refers to the city streets according to structures of the map. A map user can simulate a feasible route in the map without actually walking in the street. Further, in order to account for the wits in the sensorimotor control, the notion of representation can be extended from the predicative relation to the end-means relations, insofar as they bring about a certain target (the end) with recourse to a certain content (the means). That is, the predicative representation $R b$ in a descriptive relation describes $E . R b$ refers to the target of $E$, and the content of $R b$ describes $E$. By contrast, the instrumental representation $R m$ in an instrumental (that is, end-means) relation brings about $E$, and the content of $R m$ guides the system $S$ to reach the goal-state $E$. Furthermore, a different surrogate $R i$ in the system $S$ would be likely to bring about a different state, as opposed to $E . R m$ and $R i$, hence, are alternative means generated in the system S , alternatives which can be compared for a higher degree of prospect $P$ to bring about the end-state $E$. The prospect $P$ is a second-order instrumental representation, as it evaluates the degree in which certain means would bring about the end. $R m$ and $R i$ are exemplified in the instrumental system by motor commands, and the prospect $P$ is exemplified by costs. Thus, instrumental representations are genuine representations because they stand in a system ( $S$ ) for something else $(E)$ that they bring about. The instrumental representations stand in the sensorimotor system, rather than merely serving as physical components of a mechanism, because they always have alternatives to be chosen in their way to bring about an end. The above profile will be discussed more specifically below.

The decouplability between representations and their environments is but a particular case for a system capable of generating alternatives. The system produces something else apart from a fixed representation. In a dark night, as I encounter a distal horse (a target) I may consider it to be a cow; in a darker night, I may even feel that it is like a unicorn. For achieving a given goal, sensorimotor representations produce various motor commands that are all likely to achieve the goal. The inverse model of a sensorimotor system transforms a single goal-state $G$ into various motor commands, which would all be likely, to a certain degree, to achieve the same goal G. The sensorimotor system always has alternatives to be chosen, in a decision for an optimal motor command, as indicated by the redundancy present in the musculoskeletal system. This is unlike a physical causal relation, such as knocking a group of billiard balls with a single ball, where a move in a
particular circumstance will determine a single result. Alternatives, as aforementioned, are made by the system $S$ in a non-physical connection, when the system $S$ encounters a fixed condition (e.g. encountering a horse, or given a particular goal-state). In the sensorimotor system, the inverse model generates various motor commands in a nonphysical connection, which is non-physical insofar as it is computational. The mechanism on the basis of Bayesian decision theory, in addition, makes a choice among redundant motor commands for optimality with a lowest cost. The choice from alternatives justifies that the sensorimotor control is not a 'merely physical' device.

Decoding of sensorimotor representations in the sensorimotor control is grounded on the use of those representations in the way to achieving the goal. Therein, the pragmatic dimension of sensorimotor control is considered in terms of end-means relations. The use consists of estimation over environmental conditions for applying them and choice between them, as manifest in the application of Bayesian decision theory in sensorimotor control (Körding and Wolpert, 2006). The generated motor commands in the inverse model, in addition, are made with alternatives, which are available for choice in their use dedicated to the achievement of the goal.

The motor commands are genuine representations because their way of bringing about goal-achievement is internally rich. Based on Cummins' (1996) notion of representation, the goal-state is the target while the content in use is the information employed, serving as guidance of the instrumental control, for achieving the goal. Specifically, the estimation of environmental conditions in the Bayesian model of sensorimotor control presumes the need of achieving the goal, and so is the model of optimality. Thus, the sensorimotor representations are internally rich, even compared to the classic representations, which are dubbed as representation-hungry. For example, the Bayesian inference, in order to estimate external conditions out of noise, needs to take account of priors, that is, previous experiential outcomes. This makes the Bayesian model of the sensorimotor control even no less emphasized on internal wits than the computation related to the classic theory of representation. In addition, the decision made in relation to cost, as discussed previously, making a choice from various alternative motor commands. Furthermore, the on-line measurement of cost adds more wits on the top of the computation with open-loop optimization. The instrumental representation based on alternatives of action command, together with their accompanying estimation and choice, provides computation with internal alternatives, evaluation and inference. Such a way of representation is internally rich.

Representations can serve as stand-ins of a system without being predicative. Representations are stand-ins for the existence of certain states, or for those states' activities. The former relates to representational production and the latter representational consumption. As considered above, the stand-ins can be instrumental, and consequently need
not be built with the function of imagination, on grounds of which counterfactual representations are possible. Imagination is surely a characteristic of human cognition. Cognition, however, can have other characteristics, for example, instrumental allocation, that is, arrangement of end-means relations. The end-means relation, for a given end, need not consist of a single string of causal chain, as it can produce alternative means for the same end. Those alternatives can be evaluated with different degree of prospects for attainment of the end. The choice from alternatives justifies the cognitive bearing of the instrumental representations.

Before concluding, it should be noted that instrumental representations can have a combinatorial structure only in a loose sense. Instrumental representations of sensorimotor control do not follow the LOT, basically because its modular organization is only loosely applicable, as discussed in a previous section. As a consequence, the combinatorial structure-with which mental representations can be generated recursively and systematically from primitive states-would not be generally salient in the realm of sensorimotor control. In particular, the costs of motor commands are continuously generated in on-line feedback control, as manifest in 'cost-to-go'-the continuous and integral summation of costs, as aforementioned. With this way of computing costs, the consequently chosen motor commands can only have modular structures (if there are) in a loose sense.

## Conclusions

Computation of sensorimotor control employs instrumental representations-representations with end-means relations-as exemplified by motor commands and costs. Motor commands represent ways to achieve the goal, and costs represent prospects of goal-achievement. They are intermediate-level representations, because the computation of motor commands does not rely on reactive machinery, and because they appear at the sub-personal level. Although they have modular structures, sensorimotor representations are initiated continuously and connected integrally. In order to maintain accuracy of goal-achievement, the sensorimotor system needs on-line incorporation of sensory feedbacks, and consequently sensorimotor representations cannot be detached from the body and the environment.

## References

Clark, A. (1997). Being There: Putting Brain, Body and World Together Again. Cambridge, MA: MIT Press.
Cummins, R. (1996). Representations, Targets, and Attitudes, Cambridge, MA: MIT Press.
Engel, A. K., Maye, A., Kurthen, M., and König, P. (2013). Where's the action? The pragmatic turn in cognitive science, Trends in Cognitive Sciences, v.17(5):202209.

Fodor, J.A. (1983), The Modularity of Mind,: MIT Press.
Fodor, J.A. (1975), The Language of Thought, Cambridge, Mass.: Harvard University Press.

Franklin, D. W. and Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control, Neuron, v. 72(3): 425-442.
Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. Behavioral and Brain Sciences 27: 377-442.
Harris, C. \& Wolpert, D. M. (1998). Signal-dependent noise determines motor planning, Nature, v.394: 780-784.
Haselager, P., de Groot, A, and van Rappard (2003). Representationalism vs. anti-representationalism: a debate for the sake of appearance, Philosophical Psychology, v. 16 (1): 5-23.
Jeannerod, M. (2006). Motor cognition, Oxford Univ. Press.
Jing, J. and Weiss, K. R., (2005). Generation of variants of a motor act in a modular and hierarchical motor network, Current Biology, v. 15, 1712-1721.
Jing, J., Cropper, E. C., Hurwitz, I., and Weiss, K. R. (2004). The construction of movement with behavior-specific and behavior-independent modules, The Journal of Neuroscience, 24(28): 6315-6325.
Körding and Wolpert (2006) Bayesian decision theory in sensorimotor control, Trends in Cognitive Sciences, v.10(7), pp. 319-326.

Marr, D., (1982). Vision, San Francisco: W.H. Freeman.
Mussa-Ivaldi, F. A. (1999). Modular features of motor control and learning. Current Opinion in Neurobiology, v.9, pp: 713-717.
Orbán, G. and Wolpert, D. M. (2011). Representations of uncertainty in sensorimotor control, Current Opinion in Neurobioloby, v.21, pp.629-635.
Pezzulo, G. (2008). Coordinating with the Future: the Anticipatory Nature of Representation. Minds and Machines, 18, 179-225.
Rescorla, M. (2016). Mind \& Language, v.3(1), pp. 3-36.
Rescorla, M. (2015). The Computational theory of mind, Stanford encyclopedia of philosophy.
Shadmehr, R., and Mussa-Ivaldi, F.A. (1994). Adaptive representation of dynamics during learning of a motor task. Journal of Neuroscience, 14, 3208-3224.
Todorov, E. (2004). Optimality principles in sensorimotor control, Nature Neuroscience, 7(9), pp. 907-915.
Todorov, E. and Jordan, M. (2002). Optimal feedback control as a theory of motor coordination, Nature Neuroscience, v. 5(11), pp. 1226-1235.
Tresch, M. C., Saltiel, P. \& Bizzi, E. (1999). The construction of movement by the spinal cord. Nat. Neurosci. 2, 162-167.
Turvey, M. T., and Fonseca, S. (2009). Nature of Motor Control: Perspectives and Issues, Progress in motor control, Springer, pp. 93-123.
Wolpert, D. M. and Ghahramani, Z. (2000). Computational principles of movement neuroscience, Nature neuroscience supplement, v.3, pp. 1212-1217.
Zelik, K. E., La Scaleia, V., Ivanenko, Y. P., and Lacquaniti, F. (2014). Can modular strategies simplify neural control of multidirectional human locomotion? Journal of Neurophysiology, 111: 1686-1702.

# Disentangling perceptual and linguistic factors in parsing 

David J. Lobina (davidjames.lobina@ub.edu)<br>Faculty of Philosophy/LOGOS, University of Barcelona, Barcelona 08001, Spain<br>\title{ Josep Demestre (josep.demestre@urv.cat) \& José E. García-Albea (jegarcia.albea@urv.cat) }<br>Department of Psychology, University Rovira i Virgili, Tarragona 43007, Spain


#### Abstract

We offer a re-evaluation of the tone-monitoring technique in the study of parsing. Experiment 1 shows that reaction times (RTs) to tones are affected by two factors: a) processing load, resulting in a tendency for RTs to decrease across a sentence, and b) a perceptual effect which adds to this tendency and moreover plays a role in neutralising differences between sentence types. Experiment 2 successfully discriminates these two factors by registering event-related brain potentials during a monitoring task, establishing that the amplitudes of the N1 and P3 components -the first associated with temporal uncertainty, the second with processing load- correlate with RTs. Experiment 3 then behaviourally segregates the two factors by placing the last tone at the end of sentences, activating a wrap-up operation and thereby both disrupting the decreasing tendency and highlighting structural factors.


Keywords: Tone monitoring; Processing load; ERPs; Position effect.

## Introduction

Monitoring tasks have long been employed in psycholinguistics, and the end-of-clause effect is possibly the better-known result in the study of parsing. According to Abrams and Bever (1969), the end of a clause exerts a particular cognitive load in the parser: reaction times (RTs) to tones are higher when placed at the end of the first clause of biclausal sentences than in between clauses or at the beginning of the second clause. Other structural effects have been reported in complex sentences with this technique, such as differences in processing load between subject and object relative clauses (Cohen \& Mehler, 1996).

In this paper, we track the processing load of parsing monoclausal sentences carefully and show that the issues are rather nuanced. In particular, we show that the RTs of a tonemonitoring task are affected by the additive effects of perceptual and psycholinguistic factors, and this has gone unnoticed until now. Interestingly, perceptual and psycholinguistic factors can be both successfully discriminated by registering event-related brain potentials (ERPs) during a monitoring task and behaviourally segregated by placing the last tone at the end of sentences, where a wrap-up operation presumably takes place. Perceptual factors are rather strong in monitoring tasks and they appear to have been operative in past studies too, which now need to be reconsidered.

## Experiment 1

Monoclausal, subject-verb-object Spanish sentences were constructed. Starting from a matrix proposition, two different types of sentences were created: type $A$ sentences exhibited a complex subject but a simple object, and the reverse was
the case for type $B$ sentences. By a complex subject or object we mean a noun phrase which is modified by another noun phrase, whilst a simple subject or object is composed of a determiner and a noun only. Three tone positions were determined to probe the processing load of the parser, both within a sentence and between sentence types. The materials of Experiment 1 are shown below, where the $\mid$ symbol identifies the boundaries under study:

Type A: El candidato $\mid$ del partido $\mid$ se preparó $\mid$ el próximo discurso.
'The party's candidate prepared his next speech'.
Type B: El candidato | ha preparado | un discurso | sobre la sanidad.
'The candidate has prepared a speech about the health service’.

We chose to focus on two central operations of parsing phrase completion and clausal integration- and hypothesised that a) the tone position and sentence type factors would each be significant, and $b$ ) there would be an interaction between them. This general hypothesis yields a number of more specific predictions. In the first position, the parser has processed the same material in type A and type B sentences, identifying the noun phrase the candidate as the subject of the sentence, following the canonical subject-verb-object(s) order in Spanish, and thereby predicting the appearance of the verb. Thus, the cognitive load should be equal and the RTs similar. In the second position, the verb prediction is borne out in type B sentences and the parser successfully completes the subject noun phrase, whereas in type A sentences the parser is processing a longer subject noun phrase and the verb prediction is still active. Moreover, in type B sentences the parser has integrated the verb and the subject noun phrase and now expects an object noun phrase, whilst in type A sentences the parser is yet to conduct any integration. In this case, the processing load should be greater in type A sentences and RTs higher to those of type B sentences, given the central role of the verb in a sentence. Finally, in the third position the parser has integrated subject and verb in type A sentences and now predicts an object noun phrase, whereas in type B sentences the parser has successfully integrated part of the object noun phrase. In this case too type A sentences should involve more processing load and therefore higher RTs at this position.

## Method

Participants. 80 psychology students participated for course credit. The mean age was 20 years, and participants had no known hearing impairments.
Materials. Two variants of monoclausal, active, declarative, subject-verb-object Spanish sentences were constructed from 60 matrix propositions. Type A sentences exhibited an [NP-[PP-NP]-[VP-NP]] pattern whereas type B sentences manifested a [NP-[VP-NP-[PP-NP]]] form -these are the structural conditions of the experiment. All sentences were unambiguous and composed of high- or very high-frequency words. Three tone positions per sentence were established, the three positional conditions of the experiment (1-2-3). Tones were placed on the vowel of the second syllable following the relevant boundary, had a frequency of 1000 Hz , a duration of 25 ms ., and a peak amplitude equal to that of the most intense sound of the materials ( 80 dBs ). Every sentence had one tone only.
Procedure. The design of the experiment was a 2 (sentence type factor) by 3 (tone position factor) within-subjects, within-items factorial, and therefore six lists of the task were created. The sentences were presented over the headphones binaurally and participants were instructed to hold a keypad with their dominant hand in order to press a button as soon as they heard the tone. They were told to be as quick as possible, but to avoid guessing. Once a sentence had finished, the next sentence would be presented upon pressing the space bar, giving subjects control over the rate of presentation.

## Results

Reaction times were collected and trimmed with the DMDX programme. A response that occurred before the tone or 3 seconds after the tone was not recorded at all, while responses deviating 2.0 SDs above or below the mean of each participant were eliminated (this affected $4.3 \%$ of the data). Table 1 collates the RTs per condition.

Table 1: Experiment 1. RTs per tone position per sentence.

|  | Tone Position |  |  |
| :---: | :---: | :---: | :---: |
| Sentence Type | 1 | 2 | 3 |
| A | 257.22 | 222.51 | 206.78 |
| B | 252.40 | 217.33 | 205.26 |

As shown in Table 1, RTs are greater in Position 1 and decrease thereon for each sentence type. A repeated-measures analysis of variance showed that the tone position factor was significant for both the subjects and items analyses $\left(F_{1}(2,158)=144, p<.001, n_{p}^{2}=0.647 ; F_{2}(2,118)=295\right.$, $p<.001, n_{p}^{2}=0.834$ ), while the sentence type factor was only significant for the subjects analysis $\left(F_{1}(1,79)=4.66\right.$, $p<.05, n_{p}^{2}=0.056 ; F_{2}(1,59)=2.48$, n.s.). There was no interaction between the two experimental factors (all $F s<1$ ).

Pair comparisons between the three positions of the tone position factor show that the differences in RTs among the three positions were significant (all $p s<.01$ ).

## Discussion

The results show a clear decreasing tendency in RTs, and whilst the two experimental factors were significant in the subjects analysis, this was not the case in the items analysis, where only the tone position factor was significant. Moreover, there was no interaction between the two factors. Thus, not all of our predictions were confirmed. The decreasing progression is rather robust, and the high significance of the (tone) position factor is further confirmation. This would be in line with the general expectation that processing load decreases as a sentence is presented, which follows from the incremental nature of parsing; in this case, the least linguistic material to process, the easier it will be to respond to the tone. However, the fact that there was no interaction between the two factors is surprising, as tone monitoring was expected to be sensitive to structural features. Each tone was placed in a different segment in each sentence type, and thus the parser, except for the first tone position, cannot be computing the same predictions at each position -i.e., the processing load cannot be the same. This ought to be especially significant when it comes to integrating nouns and verbs during the course of a sentence, but as the data show the earlier or later appearance of the verb and whether noun phrases were simple or complex do not appear to have had an effect. It is worth restating that tones were placed a syllable after the main boundary, but this very short sound (/de/ or /a/, in one case) would not be enough to predict the precise nature of the new phrase in the absence of further material (prepositional, verb, etc.), as various continuations are possible; this extra syllable would only indicate that the previous phrase had finished and needs to be completed, which is precisely what we were aiming to track in the experiment.

The decreasing tendency can be observed in Abrams and Bever (1969) too (and in other past studies). These authors established three positions in sentences such as since she was free that $\mid$ day $\mid$ her $\mid$ friends asked her to come (i.e., before the main clause break, in the clause break, and right after the clause break, all marked with |), and the RTs they obtained certainly exhibit a decrease: 243 ms ., 230 and 216. Relatedly, Cutler and Norris (1979) report that monitoring tasks in general exhibit a tendency of RTs to decrease across a sentence, and this needs to be taken into consideration. Crucially, the results reported in Abrams and Bever (1969) cannot be wholly explained in terms of processing load as they used biclausal, complex sentences and the course of incremental parsing in that study ought to have been different to what we obtained here.

Thus, we postulate that there is a perceptual factor at play in monitoring tasks; roughly stated, the later the tone appears, the more prepared the participants are to respond to it. If this is the case, there would be two types of uncertainties to track in monitoring tasks: one linguistic, stemming from incremen-
tality (viz., what linguistic material is it left to process?), and the other perceptual (viz., when will the tone appear?), which we shall call the position effect. As such, the results of our experiment -a decrease in RTs and no interaction between factors- would be the product of the additive effects of perceptual and psycholinguistic factors. If this conjecture is correct, then the greater RTs in the first tone position in Abrams and Bever (1969) may not have been due to an end-of-clause effect, but the result of the combination of perceptual and psycholinguistic factors. Indeed, given that past studies did not consider this perceptual factor and thus did not control for tone position, we are unsure as to whether the end-of-clause effect is well supported. That being so, the results reported in Cohen and Mehler (1996) are at first sight structural rather than perceptual, and as such tone monitoring must be sensitive to both factors (our own results yielded clear structural factors too). In order to delve deeper into this issue, we can combine tone monitoring with the recording of ERPs, which will allow us to track two different ERP components, one related to processing load (and linguistic uncertainty), the other to the position effect (and temporal uncertainty). If there is a correlation between these ERP waves and RTs, our analysis would be confirmed.

## Experiment 2

Only type A sentences from the previous experiment were employed, as there was no need to use both sentence types; the tone positions, however, remained the same. We concentrated on two ERP components, yielding two broad predictions. It was hypothesised that the N1 wave, a component associated with temporal uncertainty (Näätänen \& Picton, 1987), would correlate with the RTs, and thus its amplitude would be highest at the first tone position, the perceptual uncertainty of the participants being greatest at that point, and decrease thereon. This part of the experiment aimed to evaluate the significance of the position effect, and the N 1 is a pertinent component for such a task, given that it tracks perceptual processes rather than (higher) cognitive ones.

The second component is the P 3 , a component whose amplitude to a secondary task has been shown to be affected by the difficulty of the primary task in dual-task settings such as ours. Past results with dual-task experiments (e.g., Wickens, Kramer, Vanasse, \& Donchin, 1983) indicate that the P3 associated with a secondary task (in this case, reacting to the tone) will have a low amplitude if the primary task (here, parsing the sentence) is of considerable difficulty. In other words, there ought to be a correlation between the fluctuations in difficulty in a primary task and the amplitude of the P3 to a secondary task. In our experiment, as the primary task decreases in difficulty (as manifested by the linear decrease in RTs from the first to the third position), the amplitude of the P3 was predicted to increase from position 1 onwards.

Crucially for our purposes, the biphasic pattern we are hypothesising is well established in the dual tasks literature. Wickens et al. (1983) report an N1-P3 pattern when an au-
ditory probe is employed, and this is precisely what we are after: an N1 wave tracking perceptual processes and a P3 component tracking cognitive processes. In particular, we expect to obtain an N1 wave with a frontal distribution and a P3 with a more posterior-paretial distribution, thus singling out two independent components and confirming the processes that interest us. If these two waves turn out to be present in the data, and their amplitudes go in the direction we are postulating, we would have clear evidence for the two factors we have postulated. To our knowledge, moreover, this is the first time that the P3 is employed in a study of parsing as a metric of processing load, and we hope our results constitute evidence for its general usefulness. Naturally, these two hypotheses hold if and only if the pattern in RTs obtained in the previous experiment does not vary, and we hypothesised that this would be the case indeed.

## Method

Participants. 18 psychology students participated in the experiment. The mean age was 22 years, and subjects had no known hearing impairments.
Materials. The same as type A sentences from the previous experiment, but these now numbered 120 items.
Procedure. Participants were exposed to a total of 120 items, presented in three blocks. Apart from the electroencephalography (EEG) measures that were undertaken and the greater number of items, the task remained the same as in the previous experiment. The EEG was recorded continuously by $19 \mathrm{Ag} / \mathrm{AgCl}$ electrodes which were fixed on the scalp by means of an elastic cap (Electrocap International, USA) positioned in accordance with the 10-20 International system. ERPs were algebraically re-referenced to linked earlobes offline. Electrode impedances were kept below $5 k \Omega$. All EEG and EOG channels were amplified using a NuAmps Amplifier (Neuroscan Inc., USA) and recorded continuously with a bandpass from 0.01 to 30 Hz and digitised with a 2 ms . resolution. The EEG was refiltered off-line with a $25-\mathrm{Hz}$, low-pass, zero-phase shift digital filter. Automatic and manual rejections were carried out to exclude periods containing movement or technical artefacts (the automatic EOG rejection criterion was $\pm 50 \mu \mathrm{~V}$ ).

## Results

## Behavioural Data

The reaction times of the 18 participants were collected and trimmed with the DMDX programme. As before, responses deviating 2.0 SDs above or below the mean of each participant were eliminated, which in this case affected $3.6 \%$ of the data. The data are presented in Table 2.

As expected, the RTs manifest the exact same pattern as in Experiment 1: reaction times decrease from the first position onwards. A repeated-measures analysis of variance showed that the tone position factor was significant for both the subjects and items analyses $\left(F_{1}(2,34)=39, p<.001\right.$, $\left.n_{p}^{2}=0.698 ; F_{2}(2,238)=93, p<.001, n_{p}^{2}=0.441\right)$. Regarding pair comparisons between the different tone positions (1

Table 2: RTs per tone position.

| Tone Position |  |  |
| :---: | :---: | :---: |
| 1 | 2 | 3 |
| 325.05 | 266.53 | 247.60 |

vs. 2, etc.), the analyses showed that all comparisons were significant (all $p s<.01$ ).

## Electrophysiological data

The data were processed using BrainVision Analyzer 2 (Brain Products, Gilching, Germany). Average ERPs were calculated per condition and per participant from - 100 to 500 ms . relative to the onset of the tone, and before grand-averages were computed over all participants. A 100 ms . pre-tone period was used as the baseline. Only trials without muscle artefact or eye movement/blink activity were included in the averaging process. The analyses were based on 15 channels divided into five separate parasagittal columns along the anteroposterior axis of the head. The columnar approach to analysing the ERP data provides both an anterior-to-posterior as well as a left/right comparison of ERP effects. The electrodes in each of two pairs of lateral columns (Inner Column: F3/F4, C3/C4, P3/P4; Outer column: F7/F8, T3/T4, T5/T6) and on the Midline Column ( $\mathrm{Fz}, \mathrm{Cz}, \mathrm{Pz}$ ) were analysed with three separate ANOVAs. The analysis of the midline column included the position factor (position 1 vs. position 2 vs . position 3) and the location factor with three levels (Fz vs. Cz vs. Pz ). The analyses of the two pairs of lateral columns involved repeated measures ANOVAs with within-participants factors Position (position 1 vs. position 2 vs. position 3) and Location (anterior, central, and posterior). Omnibus ANOVAS were followed up with pairwise comparisons intended to discern whether there were differences among the three tone positions. All post-hoc analyses were Bonferroni corrected. Based on prior reports, two time windows were selected for analysis of the mean amplitudes of the components of interest: the N1 component was analysed from 120 ms . to 200 ms ., and the P300 component was evaluated from 230 ms . to 400 ms . In order to not clutter the presentation of our results, we only report the main effect of the Tone Position factor and the significant interaction effects.

Fig. 1 depicts brain potential variations in the three midline electrodes included in the analyses. As can be observed, the three tone positions exhibit a clear biphasic pattern, with a first modulation in the N1 time window in frontal and central electrodes, followed by a second modulation in the P300 time window in the central and posterior electrodes.

N1 epoch ( $\mathbf{1 2 0 - 2 0 0} \mathbf{~ m s}$ ). During the N1 epoch, there was a main effect of Position in the Midline, Inner, and Outer columns. Bonferroni corrected pairwise comparisons showed that all three positions differ from each other significantly in the three columns (all $p s<.05$ ), reflecting a more negative-

Figure 1: ERP waveforms for the three tone positions shown from a 100 ms . before tone presentation to a 500 ms . posttone presentation. The waveforms depict brain potential variations in the three midline electrodes included in the analyses. Negative voltage is plotted up.

going amplitude for position 1 relative to position 2 , and a more negative-going amplitude for position 2 relative to position 3. There was also a significant interaction between Position and Location in the Midline, Inner, and Outer columns (all $p s<.05$ ). In the Midline and Inner columns, posthoc comparisons revealed that whereas in frontal and central electrodes position 1 was more negative relative to position 2, and position 2 more negative relative to position 3 (all $p s<.05$ ), there were no differences in the posterior electrodes (all $p s>.20$ ). In the Outer column, post-hoc comparisons revealed that whereas in frontal electrodes position 1 was more negative than position 2, and position 2 more negative than position 3 (all $p s<.05$ ), there were no differences in central and posterior electrodes (all $p s>.52$ ).

P300 epoch ( $230-400 \mathrm{~ms}$ ). During the P300 epoch, there was a main effect of Position in the Midline, Inner, and Outer Columns. Bonferroni corrected pairwise comparisons in the three columns showed all three positions to differ from each other significantly (all $p s<.05$ ), reflecting a more positivegoing amplitude for position 3 relative to position 2 , and a more positive-going amplitude for position 2 relative to position 1. There was also a significant interaction between Position and Location in the Midline, Inner, and Outer columns (all $p s<.05$ ). In all three columns, post-hoc comparisons revealed that whereas in central and posterior electrodes position 3 was more positive relative to position 2, and position 2 more positive relative to position 1 (all $p s<.05$ ), there were no differences in the frontal electrode (all $p s>.11$ ).

## Discussion

As the behavioural data show, the prediction regarding the RTs pattern was confirmed; that is, RTs to the first tone are slowest, and then become faster thereon. This allows us to discuss the ERP data in the terms we had devised. The ERP data confirm the hypothesised topographical distributions and amplitudes for the N1 and P3 waves we expected. The N1 pattern indicates that participants are indeed uncertain as to when the tone is going to appear, and their uncertainty decreases as the sentence unfolds. We stated in the previous section that the linear decrease in RTs must be due to a combination of two factors and the N1 data confirm that there is indeed a purely perceptual factor at play, what we called earlier the position effect. Regarding the P3, its pattern can be explained in terms of processing load. As the amplitude of the P3 increases from position 1 onwards, and there is furthermore a negative correlation between RTs and the amplitude of the P3, this confirms that as the sentence is being processed the parser's unfulfilled predictions decrease, and thereby more resources can be allocated to monitoring the tone.

The biphasic pattern we have recorded confirms our analysis. First, the correlation between the amplitude of the N1 wave and tone position confirms that there is a strong perceptual factor and that it has an effect on performance. Second, the correlation between the amplitude of the P3 and tone position confirms two interrelated points: a) that tone monitoring is a dual task in which sentence processing is the primary task and tone monitoring the secondary; and, consequently, b) that the fluctuations in processing load are in part due to the decreasing uncertainty the parser experiences, thus dismissing alternative explanations in terms of response strategies, guessing the position of the tone, etc. All in all, we have succeeded in discriminating - that is, recording - the two factors we had posited. In the next experiment we shall show how they can in addition be behaviourally segregated.

## Experiment 3

In the previous experiments we did not examine the end-ofclause effect directly, as we used monoclausal instead of biclausal sentences and moreover none of the tones were placed
at the end of the sentences. In this experiment, we change the tone positions of type $B$ sentences from Experiment 1 to probe if by placing a tone at the end of a sentence the strong tendency for RTs to decrease is disrupted. The end of sentences is the locus of a wrap-up operation, which need not be the same as an end-of-clause effect; the wrap-up would involve operations that would not apply at the end of clauses (e.g., closing off all syntactic phrases, completing the sentence's semantic representation, etc.). We only used type B sentences because a) no across-sentence-type comparisons were relevant, and b) type B sentences exhibit a complex noun phrase in the object position, and this is a better configuration for our purposes.

Three tone positions are maintained, but their locations were changed: one at the beginning of the sentence and two within the verb's complex object, shown in the next section. It was hypothesised that the wrap-up effect would be indeed applicable at the end of a sentence and therefore that the pattern in RTs should be different from the pattern observed in the previous experiments. In particular, we expected a V-shape pattern in which RTs to the first position were highest, descending significantly for the second position, but then raising for the third and last position, the postulated locus of the wrap-up.

## Method

Participants. 37 psychology students participated in the experiment for course credit. The mean age was 22 years, and none of the subjects had any known hearing impairments.
Materials. Type B sentences from Experiment 1 were employed. The tone positions were modified to evaluate the wrap-up effect, as shown in the sentence below (where marks tone position). 60 fillers were also employed. In all other respects, the task did not change.
(1) El candi|dato ha preparado un di| scurso sobre la sani|dad.

Procedure. The same as in Experiment 1.

## Results

The reaction times of the 37 participants were collected and trimmed with the DMDX programme. Responses deviating 2.0 SDs from the mean of each participant were eliminated, affecting $3.8 \%$ of the data. Table 3 presents the final data.

Table 3: RTs per tone position.

| Tone Position |  |  |
| :---: | :---: | :---: |
| 1 | 2 | 3 |
| 414.16 | 351.88 | 365.45 |

In this experiment, RTs were greatest in the first position, and there was a slight increase from the second to the third position. A repeated-measures analysis of variance showed that
the tone position factor was significant for both the subjects and items analyses $\left(F_{1}(2,72)=98, p<.001 ; F_{2}(2,118)=\right.$ $110, p<.001)$. All post-hoc pairwise comparisons proved to be significant (all $p s<.001$ ).

## Discussion

As predicted, the wrap-up effect was detectable with the tonemonitoring task, thereby disrupting the linear decrease in RTs, as can be seen in Fig. 2.

Figure 2: RTs progression in Experiment 3


Indeed, even though RTs to the first position were greatest and there was a noticeable decrease from the first to the second position, the processing load associated with the wrapup effect resulted in an increase in RTs from the second to the third position, in clear contrast with what was obtained in the previous experiments, and resulting in the V-shape pattern observed in Fig. 2. This would seem to indicate that tone monitoring is not entirely hostage to perceptual factors such as the position effect; a design can be found so that structural properties are brought out more clearly, resulting in the clear segregation of the two factors that have animated the whole discussion. This is behavioural confirmation of what was observed on the ERP record, vindicating the usefulness of tone monitoring as a psycholinguistic technique.

Whether the wrap-up operation can be related to the end-of-clause effect apparently unearthed in previous studies is not so clear. In those studies, and as already stated, the end of a clause was in fact the end of a subordinate clause within complex, biclausal sentences, and that introduces a specific level of complication. Moreover, the end-of-clause position was also the first tone position in those studies, pointing to the probable impact of the position effect.

## Conclusion

We have here reported three main results with the tonemonitoring technique: a) a pronounced decrease in RTs for each sentence type (Experiments 1 and 2), which suggests that the parser's processing load decreases as the sentence is presented, thus releasing more cognitive resources to monitor the tone in so doing, in accordance with well-known prop-
erties of parsing (viz., incrementality); b) no interaction between the tone position and sentence type factors (Experiment 1), the potential result, in part, of what we have called here the position effect; and c) perceptual and psycholinguistic factors can be separately observed in an ERP recording (Experiment 2) and behaviourally segregated in a carefully designed experiment (Experiment 3).

The position effect, in particular, seems to have gone entirely unnoticed in all previous tone-monitoring studies. Abrams and Bever (1969) explained their data solely in terms of what they called the end-of-clause effect, but the two factors we have analysed here seem to be operative in their study too, and that muddies their data significantly. That is, even though these scholars placed a tone at the end of a clause, this tone position constituted the first of a decreasing tendency in a series of three tones, and thus the higher RTs to this (first) position may not have been the sole result of structural factors. There is, therefore, a very possible confusion and conflation between perceptual and psycholinguistic factors in their data, and this merits a closer look.

The two main factors we have identified here - the position effect and processing load- conspire to yield the RTs that can be obtained with the tone monitoring technique, and as a result future experiments employing this technique, we advise, will need to take this contingency into consideration. In our study we have shown that the two factors can be certainly separated, especially when one sets out to do so, but the combination of these factors may hide or obscure structural effects in tone monitoring tasks, requiring a more focused design if structural effects constitute the focus point.

## Acknowledgments

This research was funded by two AGAUR research grants (2011-BP-A-00127 and 2014-SGR-1444).

## References

Abrams, K., \& Bever, T. G. (1969). Syntactic structure modifies attention during speech perception and recognition. The Quarterly Journal of Experimental Psychology, 21(3), 280-290.
Cohen, L., \& Mehler, J. (1996). Click monitoring revisited: an on-line study of sentence comprehension. Memory and Cognition, 24(1), 94-102.
Cutler, A., \& Norris, D. (1979). Monitoring sentence comprehension. In W. E. Cooper \& E. C. T. Walker (Eds.), Sentence processing: psycholinguistic studies presented to Merrill Garrett (p. 113-34). Hillsdale, NJ: Lawrence Erlbaum.
Näätänen, R., \& Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology, 24(4), 375-425.
Wickens, C., Kramer, A., Vanasse, L., \& Donchin, E. (1983). Performance of concurrent tasks: a psychophysiological analysis of the reciprocity of informationprocessing resources. Science, 221(4615), 1080-2.

# Grasping Multisensory Integration: Proprioceptive Capture after Virtual Object Interactions 

Johannes Lohmann (johannes.lohmann@uni-tuebingen.de)<br>Department of Computer Science, Cognitive Modeling, Sand 14<br>Tübingen, 72076 Germany

Jakob Gütschow (jakob.guetschow@student.uni-tuebingen.de)<br>Department of Computer Science, Cognitive Modeling, Sand 14 Tübingen, 72076 Germany

Martin V. Butz (martin.butz@uni-tuebingen.de)<br>Department of Computer Science, Cognitive Modeling, Sand 14 Tübingen, 72076 Germany


#### Abstract

According to most recent theories of multisensory integration, weighting of different modalities depends on the reliability of the involved sensory estimates. Top-down modulations have been studied to a lesser degree. Furthermore, it is still debated whether working memory maintains multisensory information in a distributed modal fashion, or in terms of an integrated representation. To investigate whether multisensory integration is modulated by task relevance and to probe the nature of the working memory encodings, we combined an object interaction task with a size estimation task in an immersive virtual reality. During the object interaction, we induced multisensory conflict between seen and felt grip aperture. Both, visual and proprioceptive size estimation showed a clear modulation by the experimental manipulation. Thus, the results suggest that multisensory integration is not only driven by reliability, but is also biased by task demands. Furthermore, multisensory information seems to be represented by means of interactive modal representations.


Keywords: Multisensory Integration; Multisensory Conflict; Object Interaction; Virtual Reality

## Introduction

Adaptive interaction with the environment requires the combination of various sensory signals. According to theories of predictive coding, this integration is driven by a desire for consistency between internal models and the external world (Friston, 2010), as well as by a desire for consistency across different internal models (Butz, Kutter, \& Lorenz, 2014; Ehrenfeld, Herbort, \& Butz, 2013). Research on the mechanism of multisensory integration has shown that this consistency is achieved in terms of a maximum likelihood integration which combines different sensory signals based on their respective reliability estimates, resulting in a Bayesian estimate about the state of the external world (Ernst \& Banks, 2002; Ernst \& Bülthoff, 2004). It is still debated, however, whether this estimate is represented by means of an integrated representation (Cowan, 2001) or by means of separate, modality specific representations which are integrated on demand (Baddeley \& Hitch, 1974). Experimental results show strong interactions between modalities in the internal representation, for instance between visual and auditory working memory (Morey \& Cowan, 2005). Furthermore, unimodal retrieval from a multisensory representation is affected by pre-
vious modal encodings (Thelen, Talsma, \& Murray, 2015). Quak, London, and Talsma (2015) suggest that task requirements typically determine whether a unimodal or a complex, multisensory representation is formed.

Our aim in the present study was two-fold. First, we wanted to investigate whether multisensory integration is modulated by task relevance. Second, we wanted to probe the nature of the stored representations. To investigate these questions, we combined an object interaction task involving multisensory conflict with a size estimation task. We let participants perform a grasp-and-carry task in an immersive virtual reality, by tracking the hands of the participants. Conflict was introduced in terms of a visual offset, either expanding or shrinking the visual grip aperture, thereby dissociating vision and proprioception. Moreover, we augmented the object interaction with vibrotactile feedback, which signaled when the relevant object was grasped. After the object interaction, we let participants judge the size of the object they interacted with either visually or based on the grip aperture. If vision and proprioception are integrated, visual estimates should be biased in the same way as proprioceptive estimates. On the other hand, if there was no bias in visual estimates, this would imply an independent storage of modal information.

## Method

## Participants

Twenty students from the University of Tübingen participated in the study (seven males). Their age ranged from 18 to 34 years $(\mathrm{M}=22.1, \mathrm{SD}=3.9)$. All participants were righthanded and had normal or corrected-to-normal vision. Participants provided informed consent and received either course credit or a monetary compensation for their participation. Three participants could not complete the experiment due to problems with the motion capture system, only the data of the remaining 17 participants was considered in the data analysis.

## Apparatus

Participants were equipped with an Oculus Rift(C) DK2 stereoscopic head-mounted display (Oculus VR LLC, Menlo

Park, California). Motion capture was realized by the combination of a Synertial IGS-150 upper-body suit and an IGS Glove for the right hand (Synertial UK Ltd., South Brighton, United Kingdom). Rotational data from the suit's and glove's inertial measurement units was streamed to the computer controlling the experiment via a Wifi connection. The data was then used to animate a simplistic hand model in a virtual reality. Since the IGS system only provides rotation data, we used a Leap Motion(C) near-infrared sensor (Leap Motion Inc, San Francisco, California, SDK version 2.3.1) to initially scale the virtual hand model according to the size of the participants' hands. To allow participants to confirm their size estimates without manual interactions, participants were equipped with a headset. Speech recognition was implemented by means of the Microsoft Speech API 5.4. The whole experiment was implemented with the Unity ${ }^{\circledR}$ ) engine 5.0.1 using the C\# interface provided by the API. During the experiment, the scene was rendered in parallel on the Oculus Rift and a computer screen, such that the experimenter could observe and assist the participants.

To provide the participants with vibrotactile feedback during object interactions, we used two small, shaftless vibration motors attached to the tip of the thumb and the index finger of the participants. The diameter of the motors was 10 mm , the height was 3.4 mm . The motors were controlled via an Arduino Uno microcontroller (Arduino S.R.L., Scarmagno, Italy) running custom C software. The microcontroller was connected to the computer via a USB port which could be accessed by the Unity® program. If a collision between the virtual hand model and an object was registered in the VR, the respective motor was enabled with an initial current of 2.0 V . The deeper the hand moved into the object, the higher the applied current (up to 3.0 V ) and the according vibration. At a current of 3.0 V , the motors produced a vibration with 200 rotations per second, the resulting vibration amplitude was 0.75 g . The wiring diagram as well as additional information regarding the components are available online. ${ }^{1}$

## Virtual Reality Setup

The VR scenario put participants in a small clearing covered with a grasslike texture, surrounded by a ring of hills and various trees. A stylized container was placed in the center of the scene and served as target for the transportation task (see Fig. 1, left panel). The to-be-grasped and carried object was a cube rendered with a marble texture. The size of the cube varied from trial to trial but the cube always appeared at the same position in the scene. Textual information, like trial instructions and error feedback were presented on different text-fields aligned at eyeheight in the background of the scene.

Centered at the participants' hip $^{2}$, the task space covered

[^394]60 cm from left to right and 55 cm in depth. Corresponding to the data generated by the IGS suit an upper body rig was placed in the scene. It was positioned about 45 cm in front of the spawning position of the cube, slightly behind the the container. Hence, participants could reach both the container as well as the cube comfortably with their right arm. The rig itself was not rendered, only the right hand of the participants appeared in the scene visually.

The multisensory conflict between visual and proprioceptive grip aperture was realized in terms of a visual angular offset on the root joints of the thumb and index finger. They could be rotated either $10^{\circ}$ towards each other, or away from each other. To maintain the same aperture, this visual offset had to be compensated by an adjustment of the actual aperture in the opposite direction. To compensate for a visual offset shrinking the grip aperture, the grip aperture had to be wider, while a visual offset extending the grip aperture required a closer grip aperture. In one third of the trials, no manipulation was applied (the different offset conditions are shown in Fig. 1, right panel).

## Procedure

Participants received a verbal instruction at the beginning of the experiment regarding the use and function of the applied VR equipment. Then, they were equipped with the inertial motion capture system, consisting of the suit and the glove. If necessary, the finger sensors of the glove were fixated with rubber bands. After aligning the sensors and enabling the data streaming, the vibration motors were fastened underneath the thumb and index finger tip with rubber bands. Participants were then seated comfortably on an arm chair.

After this, participants were asked to hold their right hand over the Leap sensor to scale the virtual hand size according to their actual hand size. The control was then switched to the IGS system and participants put on the HMD to start the training phase. Participants could practice the grasping and carrying of the cube until they felt comfortable with the task. They had to complete at least 15 successful repetitions of the task before they were allowed to proceed. The grasp and carry task is described in detail in the next section.

After completing the training, the experimenter switched manually to the main experiment. The experiment consisted of eight blocks, each composed of 15 trials. The multisensory conflict between seen and felt grip aperture was introduced during the intertrial interval while the screen was blacked out. ${ }^{3}$ In each trial participants had to grasp a cube and put it into the target container. After the object interaction, the scene faded out and one of two possible reproduction scenes

[^395]

Figure 1: The left panel shows the VR scene and the initial position and fixation checks before the presentation of the target cube. Participants had to maintain a stable fixation on the fixation cross, the green spheres represent the starting position. The right panel shows the different offset conditions. Inward offsets are indicated by the light gray joints, dark gray joints indicate the outward offset condition.
appeared. This was independent of the success in the object interaction, the reproduction scene was also shown in case of error trials. In these scenes participants had to reproduce the size of the cube they interacted with either visually or by indicating the size in terms of a grip aperture. After each block, there was a break of at least ten seconds, after the fourth block, a longer break of at least two minutes was administered. Participants were allowed to put off the HMD during the breaks. After the experiment, participants were asked to complete a presence questionnaire (IPQ, Schubert, Friedmann, \& Regenbrecht, 2001). The whole procedure took 90 to 120 minutes, including the preparation and the practice trials.

Grasp and Transportation Task At the beginning of each trial, participants had to move their right hand into a designated starting position, consisting of red, transparent spheres indicating the required positions of the fingers and the palm. The spheres turned green when the respective joints were in position. Furthermore, participants had to maintain a stable looking direction on a fixation cross (see Fig. 1, left panel). When both requirements were met, the fixation cross as well as the visible markers of the initial position disappeared and the target cube appeared. Participants were instructed to grasp the cube with a pinch grasp and to move it into the target container. A successful pinch required the tips of the thumb and the index finger to be placed on opposite sites of the cube and to maintain a stable grip aperture. Participants received vibrotactile feedback whenever touching the cube. The feedback scaled with the depth of penetration, becoming more intense the deeper the fingers were moved into the cube. The task was successfully completed by placing or dropping the cube into
the container. Success was indicated by the cube bursting into an explosion of smaller green cubes. Interactions were canceled if the cube was penetrated overly strongly, dropped outside the container, moved outside the reachable space (e.g. by throwing it), or in case the interaction took more than 20 seconds. If one of the conditions was met, participants received error feedback and the trial progressed with the reproduction task.

After completing or failing the interaction, the markers for the initial position reappeared and participants had to move their hands back into the initial position. Then a visual mask was applied, accompanied by random vibrations on the finger tips. The visual and tactile masking commenced for one second. After the masking the scene faded to black and after one second, one of the two reproduction scenes appeared. The offset manipulation was removed during the blank interval.
Size Estimation In both versions of the size estimation task, participants had to reproduce the cube size. For the visual reproduction, the scene was similar to the one in which the interaction took place. However, the ground textures were replaced and different tree models were used to avoid possible comparisons between the cube size and external landmarks. A cube was placed at the center of the scene, at the same position where the cube during the interaction phase appeared. Above the cube, a slider was displayed, which allowed the participants to scale the cube by dragging the slider button with their fingertips. The slider spanned approximately 20 cm from left to right. The initial position of the slider button and thus the initial size of the visual reference cube was determined by the cube size during the interaction phase. For the smaller three sizes the slider started out at $10 \%$ and for
the two larger sizes it started out at $90 \%$ of the sliding range.
For the proprioceptive reproduction, all visuals were deactivated (including the hand model), only the horizon as well as small white sparks in the center of the scene remained active to remind the participants that the experiment was still running. Participants were instructed to indicate the size of the cube they interacted with by means of the grip aperture between thumb and index finger. To confirm their estimate, participants were requested to say the German word for "continue" or "done" ("weiter" or "fertig"). The voice control identified these commands and ended the trial, recording either the slider position - indicating the visual edge length of the cube - or the grip aperture as the size estimate.

## Factors

We varied three factors across trials. First, the edge length of the cube, which had to be interacted with and which size had to be estimated, was either $7 \mathrm{~cm}, 7.35 \mathrm{~cm}, 7.7 \mathrm{~cm}, 8.05 \mathrm{~cm}$, or 8.4 cm . Second, the visual grip aperture was either shrunk, or extended by $10^{\circ}$, or corresponded with the felt grip aperture. In the following, we will refer to visual offsets shrinking the aperture as inward offsets, conversely, we will refer to offsets extending the aperture as outward offsets. Third, we varied the reproduction modality, which could either be visual or proprioceptive. Hence, the experiment followed a $5 \times 3 \times$ 2 within-subject design. Each of the 30 conditions was repeated four times, resulting in 120 trials. The trial order was randomized.

## Dependent Measures

Besides the size estimates in the two different reproduction conditions, we obtained several time measures. Movement onset was determined as the time between the end of the fixation until leaving the starting position. Contact time refers to the time between movement onset and successful grasp. Interaction time refers to the time interval between the grasp and reaching the container.

## Results

Data was aggregated according to the $5 \times 3 \times 2$ withinsubject design. Seeing that the size estimation had to be performed after error trials as well, there are no missing data with respect to the size estimates. For the duration measures, only correct trials were considered. The overall error rate was high (nearly $30 \%$ ), due to the task complexity. In case of missing time data, the respective cell mean was interpolated within participants by the mean over all conditions with the same offset type. For all dependent measures, values differing more than two times of the standard deviation from the mean were excluded, which was the case for $2 \%$ of all data points. ${ }^{4}$

Size estimates, time measures, and error rates were analyzed with repeated measures ANOVAs using R ( R Core

[^396]Table 1: ANOVA table for the analysis of the size estimates. The assumption of sphericity was violated for the cube size factor and the interaction between offset and reproduction condition, the according p -values were subjected to a Greenhouse-Geisser adjustment.

| factor | df | F | p | $\eta_{p}^{2}$ |
| :--- | :---: | :---: | :---: | :---: |
| size | 4 | 34.84 | $<.001^{*}$ | .69 |
| offset | 2 | 17.55 | $<.001^{*}$ | .52 |
| repro. type | 1 | 0.48 | .50 | .03 |
| size $\times$ repro. type | 4 | 2.94 | $.027^{*}$ | .16 |
| offset $\times$ repro. type | 2 | 3.95 | $.045^{*}$ | .20 |
| size $\times$ offset | 8 | 1.03 | .42 | .06 |
| size $\times$ offset $\times$ repro. type | 8 | 2.35 | $.022^{*}$ | .13 |

Team, 2016) and the $e z$ package (Lawrence, 2015). All post-hoc t-tests were adjusted for multiple comparisons by the method proposed by Holm (Holm, 1979). Results from the presence questionnaire were compared with the reference data from the online database. ${ }^{5}$ There were no significant differences.

## Size Estimates

Data were analyzed with a 5 (cube size) $\times 3$ (offset) $\times 2$ (reproduction type) factors repeated measures ANOVA. Results are shown in Tab. 1. The analysis yielded significant main effects for cube size and offset. The main effect for cube size matches the actual cube size: larger cubes were estimated larger and smaller cubes were estimated smaller. To check if the estimates were veridical, we tested whether the estimated cube sizes differed from the actual cube sizes. None of the respective comparisons yielded significant results.
With respect to the main effect of offset, participants overestimated the cube size in case of inward offsets, compared to conditions with no offset $(t(16)=3.45, p=.007)$. For outward offsets participants underestimated the cube size, compared to conditions with no offset $(t(16)=2.98, p=.009)$. Finally participants provided larger estimates in case of inward, compared to outward offsets $(t(16)=5.23, p<.001)$.

Both, cube size and offset interacted with the reproduction condition. The interaction between cube size and reproduction type is due to a systematic overestimation of the larger cubes in case of the visual reproduction. In both cases, the estimates are significantly larger than the actual sizes of 8.05 $\mathrm{cm}(t(16)=4.26, p=.003)$, and $8.4 \mathrm{~cm}(t(16)=3.21, p=$ .022 ), respectively. ${ }^{6}$
The interaction between reproduction condition and offset was further analyzed with post-hoc t-tests. Estimates in case of outward offsets were significantly smaller than in case of

[^397]

Figure 2: Three-way interaction between reproduction condition, cube size and offset. Significant differences with $p<.05$ between estimates in case of inward and outward offsets are indicated by an asterisk. The respective $t$-tests were one-sided (inward $>$ outward) and were adjusted for multiple comparisons. The dashed line indicates the actual cube size.
inward offsets, both, for visual $(t(16)=-2.21, p=.021)$, as well as for proprioceptive $(t(16)=-5.48, p=.002)$ reproduction. However, the differences between the offset conditions were much more pronounced in case of proprioceptive reproduction, resulting in the observed two-way interaction.

This pattern of results was modified by a three-way interaction between cube size, offset and reproduction condition. Separate ANOVAs for the different cube sizes showed that the interaction between reproduction condition and offset was only present for cubes of intermediate ( 7.7 cm ) and large size $(8.05 \mathrm{~cm})$. For these two conditions, there were no significant differences between the offset conditions in case of visual reproduction. The differences for proprioceptive reproduction remained significant. The main effect of offset, however, remained significant for all of these separate analyses.

With respect to our hypotheses, the difference between inward and outward offsets is most relevant. To check whether inward offsets always result in larger estimates than outward offsets, we checked whether the respective difference is significant for the five different cube sizes, separately for the two reproduction conditions. In case of proprioceptive reproduction, the difference is significant for all cube sizes, except the smallest one of 7 cm . For visual reproduction the differences reached significance for all cube sizes, except the intermediate $(7.7 \mathrm{~cm})$ and large size $(8.05 \mathrm{~cm})$. The results are shown in Fig. 2.

## Time Measures

Data were analyzed with a 5 (cube size) $\times 3$ (offset) factors repeated measures ANOVA. No significant effects were found for the movement onset times. The analysis of object contact times yielded a significant main effect for off-
set $\left(F(2,32)=76.57, p<.001, \eta_{p}^{2}=.83\right)$. Slowest contact times were observed for outward offsets, while inward offsets yielded the fastest response times. All of the respective pairwise comparisons yielded significant results. The analysis of the interaction times yielded a significant main effect for offset as well $\left(F(2,32)=4.90, p<.014, \eta_{p}^{2}=.23\right)$. Participants were slower in transporting the cube in case of outward offsets. Post-hoc t-tests showed that the interaction times were significantly elevated in case of outward offsets, both compared to inward offsets $(t(16)=2.39, p=.042)$, as well as to trials without offset $(t(16)=2.42, p=.042)$.

## Error Rates

The analysis of the error rates yielded significant main effects for cube size $\left(F(4,64)=4.27, p=.004, \eta_{p}^{2}=.21\right)$ and offset $\left(F(2,32)=12.22, p<.001, \eta_{p}^{2}=.43\right)$. In general, participants made fewer errors during interactions with larger cubes. Furthermore, error rates were higher in case of inward offsets. Post-hoc t-tests showed that error rates increased for inward offsets, when compared to both outward offsets $(t(16)$ $=-3.67, p=.004)$, and no offsets $(t(16)=-4.56, p<.001)$.

## General Discussion

Previous studies on multisensory integration have shown a dominance of visual information in the perception of object size (e.g. Ernst \& Banks, 2002). To investigate whether task demands, which require to focus on another modality, can reduce this dominance, we let participants perform a grasp-and-carry task under multisensory conflict between vision and proprioception. In order to do so, we manipulated the mapping between seen and felt grip aperture. After the ob-
ject interaction we let participants estimate the size of the object they interacted with - either visually or by providing a proprioceptive estimate via grip aperture. Our results show a systematic bias in the size estimates due to the introduced offset between seen and felt grip aperture. A wider grip aperture resulted in object size overestimations, while a smaller aperture yielded underestimations. This was true for both, visual and proprioceptive size estimates. Hence, the adaptation of the size estimation followed the proprioceptive adaptation, which was necessary to compensate for the visual offset.

While the offset manipulation led to different actual grip apertures for cubes of the same size, the visual impression of both the cube size and the grasp of the virtual hand remained the same. Thus, if the size estimate was dominated by the visual impression, there should have been no effect of the offset condition in the visual reproduction trials. In contrast, our results show a clear influence of proprioceptive information on the size estimates in both modalities. However, this influence was much more pronounced in the case of the proprioceptive reproduction. Apparently, proprioceptive information dominated the resulting percept, even if proprioception was much noisier than vision, indicated by the comparatively large variance in the proprioceptive size estimates.
The combination of VR with motion capturing enabled us to dissociate vision and proprioception in an interactive setup. Compared to previous studies, which investigated the effects of mismatching sensory information regarding an object, the applied setup allows to manipulate the own body perception without affecting the visual impression of the external, virtual world. Some issues with respect to the experimental setup remain. The high error rates imply that even with the vibrotactile augmentation, the object interaction remained difficult for the participants. Especially in case of outward offsets, participants took quite long to grasp and carry the cube. The error rates were elevated for inward offsets, which were associated with the fastest grasping and interaction times, implying a speed accuracy trade-off. Furthermore, our setup did not comprise a control condition without grasping. Including trials which only require touching the object will clarify whether the mere presence of a graspable object yields a bias towards proprioceptive information, or if performing the actual interaction is necessary to induce the bias.
Despite these issues, the results allow us to draw the following two conclusions. First, visual and proprioceptive information regarding the object size seem to be stored separately, but are able to affect each other. If there was only a single percept reflecting the cube size across modalities, then the reproduced size should be independent of the reproduction modality. This is clearly not the case, given the huge difference in the variance of the visual and proprioceptive estimates and the stronger bias in proprioceptive compared to visual reproduction. This conclusion dovetails with results reported by (Ernst \& Banks, 2002), who showed that sensory data are stored separately, when they originate from different modalities. Second, the integration process that produces a
visual or a proprioceptive estimate is influenced by the type of reproduction. The considerable difference between the effect sizes implies a different weighting of the modality-specific encodings in the two reproduction conditions.

## References

Baddeley, A. D., \& Hitch, G. (1974). Working memory. Psychology of learning and motivation, 8, 47-89.
Butz, M. V., Kutter, E. F., \& Lorenz, C. (2014). Rubber hand illusion affects joint angle perception. PloS One, 9(3), e92854.
Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behavioral and brain sciences, 24(1), 87-114.
Ehrenfeld, S., Herbort, O., \& Butz, M. V. (2013). Modular neuron-based body estimation: maintaining consistency over different limbs, modalities, and frames of reference. Frontiers in Computational Neuroscience, 7(Article UNSP 148).

Ernst, M. O., \& Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429-433.
Ernst, M. O., \& Bülthoff, H. H. (2004). Merging the senses into a robust percept. Trends in Cognitive Sciences, 8(4), 162-169.
Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 65-70.
Lawrence, M. A. (2015). ez: Easy analysis and visualization of factorial experiments [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ez (R package version 4.3)
Morey, C. C., \& Cowan, N. (2005). When do visual and verbal memories conflict? the importance of working-memory load and retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(4), 703.
Quak, M., London, R. E., \& Talsma, D. (2015). A multisensory perspective of working memory. Frontiers in human neuroscience, 9 .
R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Schubert, T., Friedmann, F., \& Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. Presence, 10(3), 266-281.
Thelen, A., Talsma, D., \& Murray, M. M. (2015). Singletrial multisensory memories affect later auditory and visual object discrimination. Cognition, 138, 148-160.

# Goal-Directed Deployment of Attention in a Computational Model: A Study in Multiple-Object Tracking 

Andrew Lovett (andrew.lovett.ctr@nrl.navy.mil)<br>Will Bridewell (will.bridewell@nrl.navy.mil)<br>Paul Bello (paul.bello@nrl.navy.mil)<br>U.S. Naval Research Laboratory, 4555 Overlook Ave SW<br>Washington, DC 20375


#### Abstract

We present a computational model exploring goal-directed deployment of attention during object tracking. Once selected, objects are tracked in parallel, but serial attention can be directed to an object that is visually crowded and in danger of being lost. An attended object's future position can be extrapolated from its past motion trajectory, allowing the object to be tracked even when it is briefly occluded. Using the model, we demonstrate that the difficulty of tracking through occlusions increases with the number of objects because they compete for serial attention.


Keywords: attention; perception; cognitive model; multipleobject tracking; visual cognition

## Introduction

Our visual experience of the world is rich and exhilarating, full of a wide variety of objects that move either predictably or erratically. Making sense of what we see requires an ability to follow these objects over time, sometimes tracking two, three, or more at once. The criticality of this capability is reflected in the existence of low-level mechanisms in the vision system that can follow multiple objects in parallel, seemingly without explicit attention. However, when the paths of objects intersect or when one object occludes another, these mechanisms are insufficient, requiring that we attend to an object to disambiguate it from others.

What guides our attention toward any one particular object in the visual field? There is a considerable amount of literature that seeks to answer this question by appealing to visual salience (Borji \& Itti, 2013). That work emphasizes contrast effects in the early visual system that draw our eyes to regions of fast motion, bright lights, and flashes of color. The target phenomenon described by that literature is goalfree, visual perception-that is, where a person would look if told to freely examine an image. The corresponding results say little about top-down influences on visual attention, such as when a person adopts a goal to track an object over time.

Fortunately, there is a task that lets us study goal-directed, visual attention under exactly these conditions. Experiments on multiple-object tracking (MOT) have people distinguish one or more targets from a larger set of identical distractors as they move across an empty background (Pylyshyn \& Storm, 1988). In these studies, the only identifying
characteristic of each object is its motion history. As a result, the low-level mechanisms that maintain the identity of objects in parallel are taxed, which lets researchers study their limits and therefore determine the conditions that require visual attention.

We know, for instance, that when targets come close to other objects, they tend to draw attention (Iordanescu, Grabowekcy, \& Suzuki, 2009; Zelinsky \& Todor, 2010). This effect of crowding does not hold when distractors are clustered together, so the goal to track the targets must play a role in protecting them from possible confusion with the other objects. It also appears that the difficulty of MOT increases when multiple targets are crowded simultaneously (Srivastava \& Vul, 2016). These findings suggest that visual attention is deployed serially and can only disambiguate one threatened target at a time.

In this paper, we present a computational model that accounts for one of the roles that attention plays in object tracking. Our research builds on previous work by Bello, Bridewell, and Wasylyshyn (2016) that assumed attention is serially deployed to initially encode targets, after which a parallel process that does not require attention exclusively handles object tracking. In their model, the interaction between attention and the tracking goal was limited to keeping visual attention on the targets. The model's ability to track targets broke down when the targets' previous positions were insufficient for distinguishing them from other nearby objects (e.g., when the objects moved quickly or were close to each other). In the updated model presented here, the processes for serial shifts of attention are refined and contribute throughout the task.

Specifically, the updated model detects when objects flagged as targets are visually crowded and, in response, directs attention to them. Sustained attention on an object enables the construction of its motion trajectory, which can be used to predict its future position. This extra information gives the model the ability to follow a target through an occlusion event, where another object overlaps or covers the target. Because attention is deployed serially, only one target can be tracked in this way at a time. As a result, when two targets are crowded, they vie for attention and the one that is not selected remains in danger of being lost by the parallel, tracking process.


Figure 1. Flow of information between components, both bottom-up (black arrows) and top-down (gray arrows). Components in bold respond to the focus of attention. Components with dashed borders are task-specific.

We claim that tracking through occlusion is facilitated by goal-directed deployment of attention to the target involved. We support this claim by providing a computational model of visual attention described in the next section. Briefly, this model requires that crowded targets compete for attention and its associated computational benefits. To test the model, we apply it to stimuli drawn from the work by Luu and Howe (2015) showing that people are better at predicting target positions from past trajectories when there are fewer targets. We find that the model accounts for these results and is in accordance with a broader range of findings in the literature.

## Computational Model of Visual Attention

The computational model is implemented using ARCADIA (Bridewell \& Bello, 2016a), a cognitive system designed for exploring the role of attention. The system operates in cycles that correspond to 25 ms of activity in human perception. On each cycle components, which carry out all the computation in a model, place their results in a location called accessible content. ARCADIA uses an attentional strategy to select one of these results as a focus of attention, which directs processing in a subset of components. On the subsequent cycle, the components receive sense data (e.g., a video frame), accessible content, and the focus of attention as input and produce the next collection of accessible content as output.

Like other models built using ARCADIA, this model of visual attention consists of a set of components and an attentional strategy. Many of the components included in the current model were previously described by Bello, Bridewell, and Wasylyshyn (2016). Looking at Figure 1, these include image segmenter, object locator, object-file binder, and vstm (which implements visual short-term memory). In the rest of this section, we summarize these components, mention changes to object locator, detail the new components, and discuss the attentional strategy.

Beginning at the bottom of Figure 1, image segmenter polls a sensor that provides one frame of video input each


Figure 2: A: An image with eight objects. B: A priority map in which four of the objects are tracked.
cycle. The component then outputs a set of proto-objects, hypotheses for the locations of objects (Rensink, 2000), based on closed-contour regions in the frame. For protoobjects to lead to object representations, they must receive attention. To this end, a set of highlighters, described later in this section, proposes one or more candidate proto-objects for the focus of attention. If ARCADIA focuses on one of these candidates, then the object file binder constructs an object file, which is based on the ideas of Treisman and Gelade (1980) and is a representation that binds together any visual features found at the proto-object's location (e.g., color profile, size). If that object file receives attention, then the vstm component stores it in memory.

## Improvements to the Parallel Aspects of Tracking

If tracking objects always required attention, then it would take four ARCADIA cycles ( 100 ms ) to go from visual input to representing a single object in vstm. To update that object's location would take the same amount of time. Even accounting for the ability to pipeline parts of the process, two cycles ( 50 ms ) are required for shifting covert attention and representing an object. The timing needed to serially update the location of multiple objects, which is based on evidence from visual search in humans (Wolfe, 2003), is unrealistic. Therefore, the model needs a way to track objects in parallel.

To address this need, the model includes tracking functionality in object locator. In earlier ARCADIA models, this component kept location information up-todate by matching object files in vstm to the proto-objects nearest to each object's last known location. This approach was inspired by Pylyshyn's (1989) proposal that roughly four objects can be tracked in parallel using visual indices and by Dawson's (1991) work that identified a nearestneighbor constraint in apparent motion, which is likely related to tracking.

In this model, we refine object locator to provide an account of Dawson's constraint based on newer results in visual processing. This new implementation generates a two-dimensional priority map (Fecteau \& Munoz, 2006; Bisley \& Goldberg, 2010), with enhanced regions at each tracked object's last known location and suppressed regions around them. Evidence for this treatment of spatial regions


Figure 3. A: An occlusion event. B: Part of a priority map, updated with an attended object's predicted location, shown as the off-center darker red circle.
comes from early work on multifocal attention (Castiello \& Umiltà, 1992) and center-surround suppression (Tsotsos et al., 1995; Desimone \& Duncan, 1995). To track objects in parallel, each object file in vstm is matched to the protoobject that most overlaps its corresponding enhanced region on the priority map. Figure 2 B shows an example of such a map with enhanced red and yellow circles and suppressed outer green rings. Importantly, if two tracked objects are near each other, one object's suppressed region may overlap another object's enhanced region (see the two lower circles in Figure 2B), resulting in a smaller enhanced region and a greater chance of a tracked object being lost. ${ }^{1}$

## Goal-Directed Attention in Tracking

Adopting a goal to track specific moving objects, or targets, alters how attention is deployed. In particular, attention can be drawn to a target when there is a risk that the parallel, tracking mechanisms could fail for that object. For instance, as suggested by Figure 3A, when multiple objects overlap, they look like a single proto-object. After those objects move apart, it is unclear which one, if any, was previously a target. This problem arises because following a target through an occlusion event requires more information than only its previous location. On these occasions, the model uses an attended target's recent motion history to extrapolate its future position in order to track it through occlusions. We conjecture that serial attention is required for this process because it involves binding trajectory information and the corresponding extrapolated position to a particular object file.

Goal-directed deployment of attention is assisted by the highlighters mentioned earlier in this section. Recall that these components propose proto-objects as candidates for attention and therefore determine which objects will be

[^398]stored in vstm and tracked by object locator. There are three highlighters, one of which is task specific and the other two are generally important for tracking. First, color highlighter is used to identify targets and queries about objects in the multiple-object tracking videos, indicated by objects changing color in the videos.

The other two highlighters propose proto-objects corresponding to currently tracked objects. The crowding highlighter proposes each tracked object as a candidate for attention and includes as information the distance from each one to the nearest other proto-object. This value provides a measure of crowding and is based on the finding that tracked objects draw attention when they are visually crowded and in danger of being lost (Iordanescu, Grabowekcy, \& Suzuki, 2009; Zelinsky \& Todor, 2010).

The maintenance highlighter proposes maintaining attention on the object that was last in focus. If attention remains on the same object over a period of time, this component computes its motion trajectory from location changes over a window of two to three cycles. Additionally, maintenance highlighter detects occlusion events, where the focused object is partially or completely occluded by another object (e.g., Figure 3A). When the attended object is occluded, the component predicts the focused object's position based on its recorded trajectory. This information lets object locator update its priority map to enhance the object's predicted location (the off-center, red circle in Figure 3B), improving its ability to continue tracking the object after the occlusion event ends.

The final component, target object guesser, records the model's responses in the multiple-object tracking task. This component reports whether the model considers a probed item to be a target (tracked) or a distractor.

The model's attentional strategy is a priority list over the elements in accessible content. The highest priority is to focus on new object-files for storage in vstm. Below that, the strategy prefers proto-objects, which enables encoding them into object files. The preferences for proto-objects are ordered with color highlighter first, which ensures that targets are initially encoded and that probes are noticed when objects change color. The next highest priority is to maintain attention on a crowded target, one whose distance to the nearest other proto-object has fallen below a crowding distance threshold. The third highest priority is to attend to whichever target is the most crowded-the one with the lowest crowding distance. This ordering enables goaldirected deployment of attention to objects that are in danger of being lost, and it handles competition between simultaneously crowded objects. Once an attended object is endangered, attention will stay on it until the distance to nearby proto-objects exceeds the crowding threshold even if other targets are also in danger.

In summary, the model consists of eight components (Figure 1), four of which are new and one of which was substantially changed. Two components are task specific: target object guesser and color highlighter. The model includes three free parameters, two in object locator ${ }^{1}$ and


Figure 4. Simulated accuracy across conditions. Error bars are standard error. Speeds were calibrated separately for 2 and 4 targets to achieve about $75 \%$ accuracy.
the crowding distance threshold, which indicates when targets are too close to other proto-objects. In the next section, we report an experiment that supports the validity of this model in the context of multiple-object tracking.

## Evaluation

We evaluated the computational model by running it on MOT videos similar to those used in Luu and Howe's (2015) Experiment 1. In that work, participants tracked either two or four targets with either predictable or unpredictable motion trajectories. Luu and Howe's key finding was that people more accurately track objects with predictable trajectories than with unpredictable trajectories, but only in the two target condition. The model in this paper accounts for this effect, showing that goal-directed deployment of attention can be used to predict a target's location from its past trajectory. This ability enables tracking a single target through an occlusion, and when multiple targets are simultaneously crowded, they compete for attention. This competition for resources means that task difficulty increases with the number of targets.

## Experiment

In each trial of Luu and Howe's experiment, two or four out of eight total disks were highlighted in red to indicate that they were the targets. Afterwards, all disks turned black and the disks moved for 5.5 s while participants fixated on a center cross. During this time the disks could occlude (i.e., pass through) each other. At the end, two disks were highlighted in sequence, and participants indicated whether each one was a target. Each highlighted disk had a $50 \%$ chance of being a target, and participants needed to respond correctly on both for the trial to be coded as correct.

There were two movement conditions for the experiment. In the first condition, every disk moved predictably in straight lines and changed direction only after bouncing off the edge of the display. In the second condition, the disks moved similarly, but every $300-600 \mathrm{~ms}$, they would randomly change direction. This unpredictable movement was expected to reduce the reliability of any effort to compute and utilize motion trajectories.

At the beginning of the study, the motion speeds for each participant were calibrated to determine the speed where the participant achieved $75 \%$ accuracy. Calibration occurred separately for two and four targets and used only predictable motions. Afterwards, participants were tested over 120 randomly generated trials, 30 in each condition (number of targets $\times$ motion predictability), with conditions interleaved.

Luu and Howe reported data from 15 participants. Their results found, unsurprisingly, that tracking two targets was easier than tracking four, as indicated by a much higher speed when calibrating for two targets. Importantly, they observed a significant interaction between the number of targets and motion predictability. Pairwise comparisons indicated that predictable motions were easier than unpredictable motions for two targets but not for four targets. Luu and Howe proposed that object tracking is sensitive to motion trajectories for two targets, but less so for four targets, which is in line with findings by Fencsik, Klieger, and Horowitz (2007).

## Model

To evaluate the computational model using Luu and Howe's experiment, we randomly generated 120 trial videos each for 15 virtual "participants" (the model was the same in each case, so only the trial videos varied). The videos matched the description in the paper as closely as possible with five minor exceptions.
(1) There was no fixation cross, but center fixation was enforced in the model.
(2) It was impossible to match to the original study's display size $\left(15^{\circ} \times 15^{\circ}\right)$ because the model does not perceive the display from a quantifiable viewing distance. However, the study's proportion of disk size to display size was maintained.
(3) Videos were constrained to begin and end with all disks at least one radius apart (such constraints are common but were not mentioned in the paper).
(4) To save simulation time, disks were highlighted for a shorter duration.
(5) Disk colors differed from the original, which was incidental.
The model's crowding distance threshold was 1.6 diameters, meaning an attended target would need to be at least this distance away from all other disks before the model could swap attention to another target.

## Results

The calibration phase of the experiment differed slightly from Luu and Howe's approach. Because the model was held constant across virtual participants, we calibrated the speeds only once. We found that to ensure roughly equivalent accuracy close to $75 \%$, the speeds were eight pixels per cycle for two targets and four pixels per cycle for four targets.

Figure 4 displays the results for each condition. A twoway ANOVA with set size (two vs. four targets) and predictability was conducted. There was a significant main
effect of predictability, $\mathrm{F}(1,56)=14.3, \mathrm{p}<.001$, indicating that accuracy was higher with predictable trajectories. There was also a significant interaction between set size and predictability, $\mathrm{F}(1,56)=4.8, \mathrm{p}=.032$, indicating that the effect of predictability was greater for two targets. Unpaired comparisons confirmed that predictability had a significant effect on accuracy for two targets, $\mathrm{M}=79.8 \%$ (predictable) vs. $68.9 \%$ (unpredictable), $\mathrm{t}(28)=3.91, \mathrm{p}<.001$, but not for four targets, $M=75.3 \%$ (predictable) vs. $72.4 \%$ (unpredictable), $\mathrm{t}(28)=1.23, \mathrm{p}=.230$.

## Discussion

The model's results matched the human data, which suggests that the model accounts for two key findings. First, as evidenced in the speed discrepancies during the calibration phase, tracking two targets was easier for the model and for people than tracking four targets. Notably, increasing the number of targets increases both the number of possible occlusion events and the potential for simultaneous crowding. These effects are important for the model, which explains errors as resulting in part from failures to attend to targets during occlusion. As a result, slowing object movement reduces the number of occlusions and contributes to the ability to successfully track targets.

The second and more important finding is that the model more accurately tracked objects that moved predictably than those that moved unpredictably, but only for two targets. To understand this, we have to describe why the model could track some objects through occlusion events when the trajectories were unpredictable. Recall that objects changed direction only every $300-600 \mathrm{~ms}$, or $12-24$ cycles in ARCADIA, and that the model calculates motion trajectories over a 2 cycle window. As long as a target maintains course through the occlusion and the two cycles before it, tracking should work perfectly. In practice, this means that the unpredictable trajectories only disrupt a small proportion of occlusion events.

As an explanation, the model suggests that there are two sources of error: missed occlusion events and unpredictable trajectories for attended occlusions. With four targets there are more missed occlusion events due to simultaneous crowding than with two, so proportionally that has a larger effect on the error rate than the unpredictable trajectories. This difference explains why unpredictable trajectories are more harmful with two targets than with four, and the combination of this with the overall small proportion of occlusion events disrupted by unpredictable trajectories explains the lack of a significant effect with four targets.

## General Discussion

The model demonstrates the critical role of goal-directed visual attention in object tracking. Although attention is not always needed to update target locations, it provides key information to aid in tracking targets that are in danger of being lost due to visual crowding. In the reported model, attention provides a target's motion trajectory, which enables tracking through occlusions.

One explanation for how people track multiple objects is provided by the multifocal view of attention (Cavanagh \& Alvarez, 2005). Proponents of this view have argued for two theoretical limits on attention. First, attention may be a limited resource that must be distributed among targets (Holcombe \& Chen, 2012), which makes tracking more difficult when targets are crowded simultaneously and must compete for attention (Srivastava \& Vul, 2016). Second, attention may be subject to spatial interference between neighboring targets (Franconeri, Jonathan, \& Scimeca, 2010), which makes tracking more difficult when targets are nearer to each other (Shim, Alvarez, \& Jian, 2008; Holcombe, Chen, \& Howe, 2014).

The reported model offers a competing explanation that distinguishes between serial attention to a single object, which is used to bind features and compute motion trajectories; and parallel enhancement of multiple objectlocations, which is used to track objects. These separate mechanisms account for both apparent limits described above. First, tracking difficulty increases when targets are simultaneously crowded because they compete for the serial focus of attention. Second, difficulty increases when targets are near each other because the parallel tracking process uses center-surround suppression, with an enhanced region and a surrounding suppressed region at each target's location. When two targets are close such that one's suppressed region overlaps the other's enhanced region, the enhanced region shrinks and there is a greater chance of losing the target. Additionally, difficulty increases with the number of targets and with object speed (Alvarez \& Franconeri, 2007) because these manipulations increase the frequency of events where targets are simultaneously crowded or targets interfere with each other.

Although there are other computational models that have been applied to MOT, the reported model provides a novel explanation. Oksama and Hyönä (2008) relied solely on serial attention and Kazanovich and Borisyuk (2006) relied entirely on multifocal attention. Srivastava and Vul's (2016) Bayesian, multifocal model is similar to ours in that it distributes attention to visually crowded targets, which lets it predict greater tracking difficulty when targets are crowded simultaneously. However, their model makes no link between attentional distribution and computing motion trajectories. Additionally, the model cannot account for spatial interference between targets. Finally, their model is disconnected from video input, and it abstracts away the underlying correspondence problem.

In this paper, we demonstrated the role that goals may play in object tracking. In particular, the model's implicit goal to track targets enhances its ability by influencing where it attends. That is, selecting an object as a target recruits processes that monitor crowding and maintain focus when that target is endangered. Additionally, we note that the information made available by attending to an object is a form of indirect influence by the goal on visual processing (e.g., during the creation of the priority map). In the future we intend to explore other cases where the goal-directed
deployment of attention interacts with perception and eventually with motor control.

## Acknowledgments

This research was performed while the first author held an NRC Research Associateship award at the U.S. Naval Research Laboratory. The authors acknowledge support from the Office of Naval Research under grants N0001416WX01804 and N0001417WX00153. The views expressed in this paper are solely the authors' and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

## References

Alvarez, G. A., \& Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. Journal of Vision, 7(13):14.
Bello, P., Bridewell, W., \& Wasylyshyn, C. (2016). Attentive and pre-attentive processes in multiple object tracking: a computational investigation modeling object construction and tracking. In Proceedings of the 38th Annual Meeting of the Cognitive Science Society. Philadelphia, PA.
Bisley, J. W., \& Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. Annual Review of Neuroscience, 33, 1-21.
Borji, A., \& Itti, L. (2013). State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, 185-207.
Bouma, A. (1970). Interaction effects in parafoveal letter recognition. Nature, 226, 177-178.
Bridewell, W., \& Bello, P. (2016a). A theory of attention for cognitive systems. In Fourth Annual Conference on Advances in Cognitive Systems. Evanston, IL.
Castiello, U., \& Umiltà, C. (1992). Splitting focal attention. Journal of Experimental Psychology: Human, Perception, \& Performance, 18, 837-848.
Cavanagh, P., \& Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. Trends in Cognitive Sciences, 9, 349-354.
Dawson, M. (1991). The how and why of what went where in apparent motion: modeling solutions to the motion correspondence problem. Psychological Review, 98, 561603.

Desimone, R., \& Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18, 193-222.
Fecteau, J. H., \& Munoz, D. P. (2006). Salience, relevance, and firing: a priority map for target selection. Trends in Cognitive Sciences, 10, 382-390.
Fencsik, D., Klieger, S., \& Horowitz, T. (2007). The role of location and motion information in the tracking and recovery of moving objects. Attention, Perception, \& Psychophysics, 69, 567-577.
Franconeri, S. L., Jonathan, S. V, \& Scimeca, J. M. (2010). Tracking multiple objects is limited only by object
spacing, not by speed, time, or capacity. Psychological Science, 21, 920-925.
Holcombe, A. O., \& Chen, W.-Y. (2012). Exhausting attentional tracking resources with a single fast-moving object. Cognition, 123, 218-228.
Holcombe, A. O., Chen, W.-Y., \& Howe, P. D. L. (2014). Object tracking: absence of long-range spatial interference supports resource theories. Journal of Vision, 14(6), 1.
Iordanescu, L., Grabowecky, M., \& Suzuki, S. (2009). Demand-based dynamic distribution of attention and monitoring of velocities during multiple-object tracking. Journal of Vision, 9(4), 1.
Kazanovich, Y., \& Borisyuk, R. (2006). An oscillatory neural model of multiple object tracking. Neural Computation, 18, 1413-1440.
Kramer, A. F., \& Hahn, S. (1995). Splitting the beam: distribution of attention over noncontinguous regions of the visual field. Psychological Science, 6, 381-386.
Luu, T., \& Howe, P. D. L. (2015). Extrapolation occurs in multiple object tracking when eye movements are controlled. Attention, Perception, \& Psychophysics, 77, 1919-1929.
Oksama, L., \& Hyönä, J. (2008). Dynamic binding of identity and location information: a serial model of multiple identity tracking. Cognitive Psychology, 56, 237-283.
Pylyshyn, Z. (1989). The role of location indexes in spatial perception: a sketch of the FINST spatial-index model. Cognition, 32, 65-97.
Pylyshyn, Z. W., \& Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spatial Vision, 3, 179-197.
Rensink, R. A. (2000). The dynamic representation of scenes. Visual Cognition, 7, 17-42.
Shim, W. M., Alvarez, G. A., \& Jiang, Y. V. (2008). Spatial separation between targets constrains maintenance of attention on multiple objects. Psychological Bulletin Review, 15, 390-397.
Srivastava, N., \& Vul, E. (2016). Attention modulates spatial precision in multiple-object tracking. Topics in Cognitive Science, 8, 335-348.
Treisman, A. M., \& Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.
Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., \& Nuflo, F. (1995). Modeling visual attention via selective tuning. Artificial Intelligence, 78, 507-545.
Whitney, D., \& Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. Trends in Cognitive Sciences, 15, 160-168.
Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. Trends in Cognitive Sciences, 7, 70-76.
Zelinsky, G. J., \& Todor, A. (2010). The role of "rescue saccades" in tracking objects through occlusions. Journal of Vision, 10(14), 29.

# 'It's More Fun With My Phone': A Replication Study of Cell Phone Presence and Task Performance 

Ulrik Lyngs (ulrik.lyngs@cs.ox.ac.uk)<br>Department of Computer Science, University of Oxford<br>Oxford, OX1 3QD United Kingdom

Ryan McKay (ryan.mckay@rhul.ac.uk)<br>Department of Psychology, Royal Holloway, University of London<br>Egham, Surrey, TW20 0EX United Kingdom


#### Abstract

From distracted driving, to work focus on a computer, increasing amounts of research is investigating how digital technology influences users' attention. A couple of widely cited studies have found that the mere presence of cell phones interferes with social interactions and cognitive performance, even when not actively in use. These studies have important implications but they have not yet been replicated, and also suffer from methodological shortcomings and lack of established theoretical frameworks to explain the observed effects. We improved the methodology used in a previous study of phone presence and task performance (Thornton, Faires, Robbins, \& Rollins, 2014), while testing an 'opportunity cost' model of mental effort and attention (Kurzban, Duckworth, Kable, \& Myers, 2013). We were unable to replicate Thornton et al.'s finding that presence of cell phones reduces performance in a specific cognitive task (additive digit cancellation). Moreover, contrary to our expectations, we found that participants who used their phones more, and who were more attached to them, found the tasks more fun/exciting and effortless, if they completed them with their phone present.


Keywords: attention; distraction; cell phones; smartphones; effort; task performance

## Introduction

A growing amount of research in the cognitive science of attention studies how workers and students navigate a workspace with ample opportunities for distraction and/or multitasking (Cain \& Mitroff, 2011; Mark, Voida, \& Cardello, 2012; Ophir, Nass, \& Wagner, 2009; Pea et al., 2012; Ralph, Thomson, Cheyne, \& Smilek, 2014). Recently, this strand of research has moved on to how smartphones influence their users' attention. This is an important topic, since more than $72 \%$ of the US population own smartphones (Pew Research Center, 2016), and because it has very real consequences: the US Department of Transportation recently urged mobile companies to develop a simplified 'Driver Mode' for smartphones, due to an alarming rise in traffic accidents related to distracted driving (NTHSA, 2016).

A couple of widely cited studies have reported negative effects of the mere presence of cell phones on social interactions. Przybylski \& Weinstein (2012) varied whether or not a mobile phone was placed next to strangers engaged
in a conversation task and found that participants reported lower relationship quality and partner closeness when a cell phone was present. A follow-up observational study found a similar effect in a coffee-shop setting (Misra, Cheng, Genevie, \& Yuan, 2014). However, the results from these studies are open to a multitude of interpretations (e.g. various meanings of phone presence in a social context).

Our point of departure was a controlled study by Thornton et al. (2014) who in a non-social context investigated effects of cell phone presence on performance in simple cognitive tasks. They varied whether or not a cell phone was present on a participant's table while he/she completed a series of tasks (digit cancellation: searching for and crossing out target numbers among other numbers, or trail making: connecting consecutively numbered or lettered circles displayed in random order). They found that people performed worse in more challenging versions of these tasks (crossing out pairs of target numbers that add up to a specific number; connecting circles so that consecutive numbers and consecutive letters alternate, e.g. 1-A-2-B-3-C$\ldots$..), when a cell phone was present. The authors concluded that the mere presence of a cell phone, even when not in use, can be distracting and cause performance deficits on tasks that require full attention for optimal performance.

The experiments by Thornton et al. have potentially widereaching implications, from distracted driving to performance in schools and workplaces (Thornton et al., 2014). However, no replication studies have been conducted to establish the reliability of their findings. Moreover, their study had limitations: In their first experiment, they manipulated the presence of an experimenter's cell phone rather than the participant's own. In their second experiment, they varied the presence of participants' own phone but did not check whether their procedure for doing so made participants suspicious about the purpose of the experiment. They also did not test any theoretical frameworks that would explain their observed effects.

## The present research

We followed up on Thornton et al.'s study, addressing these limitations: We i) conducted a replication study using their
original stimuli for the digit cancellation task ${ }^{1}$ (responding also to general calls for replication studies, cf. Francis, 2012; Nosek, Spies, \& Motyl, 2012); ii) improved the original procedure to study effects of presence of participants' own smartphones while ruling out suspicion; and iii) tested a new theoretical framework for understanding the effects. In relation to the latter, we applied Kurzban et al.'s 'opportunity cost' model of attention and mental effort (Kurzban et al., 2013). According to this cognitive model, the human mind continuously computes the opportunity costs of our available tasks, i.e. the value of options that one is missing out on by persisting on the current task. The higher the perceived opportunity costs, the more the current task will feel mentally effortful and/or boring, with decreased quality of performance to follow. This model is well suited to predict effects of smartphones: Smartphones give immediate access to a virtual infinity of stimulating and relevant content, from global news to social gossip and video games. Insofar as they therefore afford opportunities for highly rewarding activities other than the task at hand, smartphone presence should increase the current task's opportunity costs. In turn, this might make one's current task feel more boring or effortful, and cause decreased quality of performance. Hence, our predictions were:

Prediction 1 (replication): Average scores in the additive digit cancellation task will be lower when a smartphone is present than when it is absent.

Prediction 2: The digit cancellation tasks will feel more effortful to complete when a smartphone is present than when it is absent.

## Methods

## Participants

53 participants ( 50 female) were recruited at the University of London, Royal Holloway ${ }^{2}$. Mean age was 18.8 years (SD $=1.4$, range 17-27).

## Materials

Digit Cancellation Task Participants completed two versions of a digit cancellation task, using Thornton et al.'s original stimuli. In both tasks, participants were given a piece of paper containing 20 rows of 50 -digit strings. In the 'simple' version, participants cross out every instance of the number specified at the beginning of each row (e.g. 3: $7301638 \ldots$ ). In the 'additive' version, participants cross out

[^399]every instance of two consecutive numbers that, when added, equal the digit specified at the beginning of each row (e.g. 5: 1237814...). In the 'simple' version participants cross out as many numbers as possible in 90 seconds; in the 'additive' version they cross out as many pairs of numbers as possible in 180 seconds.

Effort Measure The participants filled in a brief questionnaire about how effortful they thought each task was to do. Participants indicated a) how boring or exciting the task was ( $1=$ Very boring, $7=$ Very exciting), b) how effortless the task was (1 = Intensely effortful, $2=$ Completely effortless), c) how fun the task was ( $1=$ Not fun at all, 7 = Intensely fun), and d) how difficult the task was ( $1=$ Not difficult at all, $7=$ Intensely difficult). We constructed the questionnaire to probe the experiences mentioned by Kurzban et al. (2013) as dimensions of effort that correspond to perceived opportunity costs.

Individual Difference Questionnaires Following Thornton et al., participants completed a) the Attentional Behaviour Rating Scale (Ponsford \& Kinsella, 1991), a measure of general attentional difficulties, b) a Cell Phone Usage survey (Thornton et al., 2014), a measure of overall cell phone use, c) the Possession Attachment survey (Weller, Shackleford, Dieckmann, \& Slovic, 2013), a measure of how attached participants feel to their phone, and d) general demographics.

## Procedure



Figure 1: Experimental procedure
After signing a consent form, participants were asked to use their phone to photograph one of four objects placed on a desk. After the participant took the photo, an RA asked to
look at it and made a note of the object photographed and the photo's orientation. (What the participant photographed was irrelevant - the purpose of this initial task was to check if the participant had a smartphone, and to give the RA control of the phone's placement without making the purpose of the experiment obvious.) Next, participants were seated. In the phone-present condition, the RA placed the phone face-up near the edge of the table and said "I'll just leave this here, if that's okay". In the phone-absent condition, the RA placed a stack of post-it notes near the edge of the participant's table, and asked the participant to turn off their phone and put it away in their bag. Participants were then given one of the digit cancellation tasks to complete (order was counterbalanced). After completing the task, they filled in an effort measure. Then they completed the second digit cancellation task, and filled in another effort measure. Finally, the participants filled in an openended question about what they thought the purpose of the experiment was, followed by the questionnaires. The procedure is summarized in figure 1.

## Results

No participants reported suspicion that the purpose of the experiment was to study effects of phone presence.

## Prediction 1: Phone presence and cancellation score

In the simple digit cancellation task, there was no significant difference between cancellation scores in the phone-present $(\mathrm{Mdn}=65.0, \mathrm{IQR}=11)$ and phone-absent $(\mathrm{Mdn}=69.5$, $\mathrm{IQR}=14.2$ ) conditions, $\mathrm{W}=348.5, \mathrm{p}=0.30^{3}$. Similarly, in the additive cancellation task there was no significant difference between scores in the phone-present (Mdn = 20.0, $\mathrm{IQR}=4.5$ ) and phone-absent $(\mathrm{Mdn}=18.0, \mathrm{IQR}=6)$ conditions, $\mathrm{W}=259.5, \mathrm{p}=0.62$.

As will be discussed, scores in the additive cancellation task were highly left-skewed, with very few participants obtaining a score higher than 23 (see Fig. 2).


Figure 2: Distribution of scores in the additive cancellation task.

[^400]Table 1: Scores in the individual difference measures

|  | Phone present |  | Phone absent |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Mean | SD | Mean | SD |
| Attentional behaviour | 43.3 | 1.2 | 41.9 | 1.2 |
| Cell phone use | 57.0 | 2.3 | 58.3 | 1.6 |
| Possession attachment | 17.4 | 0.9 | 17.3 | 1.0 |

## Prediction 2: Phone presence and subjective effort

To test effects on subjective effort, we first did a principal component analysis of responses on the effort measure. Scores for the simple cancellation task clustered on a 'fun/excitement' and a 'difficult/effortful' factor, whereas scores for the additive cancellation task clustered on a single factor of 'effortlessness'. We computed a score for each participant on these three factors and used them as our measures of 'effort'.

There was no main effect of phone presence on how effortful participants found the tasks, neither in the simple cancellation task ('fun/excitement': phone-present, Mdn = 4.75, phone-absent: $\mathrm{Mdn}=4.50, \mathrm{~W}=285, \mathrm{p}=0.65$; 'difficult/effortful': phone-present: Mdn $=3.0$, phoneabsent: $\mathrm{Mdn}=3.5, \mathrm{~W}=358.5, \mathrm{p}=0.32$ ) nor the additive cancellation task ('effortlessness': phone-present: Mean = 3.81, $\mathrm{SE}=0.17$, phone-absent: $\mathrm{Mean}=3.48, \mathrm{SE}=0.21$, $t(43.58)=-1.25, \mathrm{p}=0.22) .{ }^{4}$

## Interactions: Effects of personality variables

To explore whether the personality variables interacted with effects of phone presence, we split participants into 'high' and 'low' scoring groups on the questionnaires (Attentional Behaviour, Cell Phone Usage, and Possession Attachment), separating the groups at the median. We conducted factorial ANOVAs for each effort dimension, using 'high'/'low' questionnaire category as predictors. In the simple cancellation task, there was a significant interaction between smartphone presence and Cell Phone Usage, $F(1,41)=$ $5.00, p=0.03$ : When a phone was present, participants high on Cell Phone Usage rated the task as more fun/exciting (Mean $=5.17, \mathrm{SD}=1.05$ ), than did those low on Cell Phone Usage (Mean $=4.47, \mathrm{SD}=0.72$ ), $p=0.039$. In other words, participants who generally use their phones more found the task less boring when they completed it with their phone next to them. See Figure 3.

[^401]

Figure 1: Interaction between phone use and phone presence on 'fun/exciting' ratings for the simple cancellation task. Error bars show standard errors.

Similarly, in the additive cancellation task, there was a significant interaction between smartphone presence and Possession Attachment, $F(1,41)=4.40, p=0.04$. When a phone was present, participants high on Possession Attachment found the task more effortless (Mean $=4.04$, $\mathrm{SD}=0.90$ ), than those less attached to their phones (Mean = $2.94, \mathrm{SD}=1.02$ ), $p=0.10$. In other words, participants more addicted to their phones felt that the task required less effort if they had their phone next to them. See Fig. 4.


Figure 2: Interaction between phone attachment and overall scores on 'effortlessness' in the additive cancellation task.

## Discussion

In terms of task performance, we, similarly to Thornton et al., did not observe any statistically significant effect of phone presence on performance in the simple digit
cancellation task. More importantly, however, we did not replicate Thornton et al.'s central finding that phone presence causes diminished performance on the additive version of the task. Two things should be noted: Even though our sample size was similar to the original study, it may have been too small to reliably detect this effect. We ran a post-hoc power analysis of Thornton et al. and found that their experiments $(\mathrm{n}=54$ and $\mathrm{n}=47)$ only had a power of .65 to detect an effect in a two-tailed $t$-test. Sample size should have been $n=66$ just to obtain a power of .8 . Note, however, that we did not even observe a trend towards replication - in fact, in our study, additive cancellation scores were marginally larger in the phone-present than the phone-absent condition. Moreover, recall that scores in the additive task were left-skewed with very few participants obtaining a score higher than 23 . When we went over Thornton et al.'s stimuli, we discovered that one row of numbers (row nine), located when most participants were running out of time, had no targets at all. Moreover, the two rows before this one contained only one target each, in contrast to the first six rows which contained from two to four targets each. This will have reduced variation in performance between participants in the higher end of the performance distribution. For example, if one participant just managed to cross out the single target in row eight before running out of time, whereas another were ahead and managed to search through also all of row nine, these two participants will still have been given the same score. Hence, the material design is likely to have reduced our power to detect an effect, because it will have masked some of the variation in performance between participants.

Another issue is that in our setup, each participant only completed two versions of the cancellation task, whereas in Thornton et al.'s original study each participant completed two versions of the cancellation task and two versions of the trail making test. Whereas Thornton et al. did not discuss this, it is possible that effects of phone presence on task performance in this particular lab scenario is contingent on some degree of mental fatigue or shift in motivation from performing more tasks (Baumeister, Bratslavsky, Muraven, \& Tice, 1998; Inzlicht, Schmeichel, \& Macrae, 2014). However, we did add an effort measure for participants to complete after each task, which reduced the difference in participant investment between the original study and ours.

In terms of subjective effort, there was also no main effect of smartphone presence. However, we observed an unpredicted effect in which participants using their phones more often, and participants more attached to their phones, found the tasks more fun/exciting and effortless, respectively, if they completed them with their phones next to them. We cannot draw any strong conclusions due to our limited sample size and the post-hoc nature of this analysis, but future studies should test if the relationship replicates.

The interaction ran in the opposite direction from what we had initially predicted from Kurzban et al. (2013)'s 'opportunity cost' framework. Nevertheless, we still expect
this framework to be useful to approach cognitive effects of information communication technologies. From 'fear of missing out' to popular anti-distraction apps like Freedom and SelfControl that people use to restrict the functionality of their own devices, many current phenomena suggest that opportunity cost models remain important to explore. However, if the interaction effect replicates, it might mean that factors like anxiety from separation from one's phone (Cheever, Rosen, Carrier, \& Chavez, 2014) or positive feelings from having more stimulation available (Gazzaley \& Rosen, 2016) provides better explanations than Kurzban et al. (2013)'s opportunity cost model.

Finally, Kurzban et al.'s paper offered a persuasive, but abstract model. The effort measure we developed here is the first attempt to operationalize their opportunity cost model for experimental studies. Despite the present paper's mixed findings, we encourage future studies to apply Kurzban et al.'s model to human-computer interaction research and to test the reliability and validity of our effort measure.

In sum, follow-up research should establish whether Thornton et al.'s finding of a detrimental effect of phone presence on performance in the additive cancellation task is valid, by using larger sample sizes and adjusting the experimental stimuli to better pick up variation between participants. Future studies should also test whether heavy phone users really do feel that tasks are less, rather than more, effortful to complete when they have their phones present. With smartphone use now ubiquitous, it should be a priority in cognitive science research on executive functioning to establish conclusive findings on how smartphones affect users' attention and performance, and why.

## Acknowledgements

This research would not have been possible without the hard work of our research assistants Alexandrina Vasilichi and Ken Petalcorin.

## References

Baumeister, R. F., Bratslavsky, E., Muraven, M., \& Tice, D. M. (1998). Ego depletion: is the active self a limited resource? Journal of Personality and Social Psychology, 74(5), 1252-1265.
Cain, M. S., \& Mitroff, S. R. (2011). Distractor filtering in media multitaskers. Perception, 40(10), 1183-1192.
Cheever, N. A., Rosen, L. D., Carrier, L. M., \& Chavez, A. (2014). Out of sight is not out of mind: The impact of restricting wireless mobile device use on anxiety levels among low, moderate and high users. Computers in Human Behavior, 37, 290-297.
Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. Psychonomic Bulletin \& Review, 19(6), 975-991.
Gazzaley, A., \& Rosen, L. D. (2016). The Distracted Mind: Ancient Brain in a High-Tech World. Cambridge,

MA: MIT Press.
Inzlicht, M., Schmeichel, B. J., \& Macrae, C. N. (2014). Why self-control seems (but may not be) limited. Trends in Cognitive Sciences, 18(3), 127-133.
Kurzban, R., Duckworth, A., Kable, J. W., \& Myers, J. (2013). An opportunity cost model of subjective effort and task performance. The Behavioral and Brain Sciences, 36(6), 661-79.
Mark, G. J., Voida, S., \& Cardello, A. (2012). "A Pace Not Dictated by Electrons": An Empirical Study of Work Without Email. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 555-564). New York: ACM.
Misra, S., Cheng, L., Genevie, J., \& Yuan, M. (2014). The iPhone Effect: The Quality of In-Person Social Interactions in the Presence of Mobile Devices. Environment and Behavior, 1-24.
Nosek, B. A., Spies, J. R., \& Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. Perspectives on Psychological Science, 7(6), 615-631.
NTHSA. (2016). Visual-Manual NHTSA Driver Distraction Guidelines for Portable and Aftermarket Devices. Washington, DC.
Ophir, E., Nass, C., \& Wagner, A. D. (2009). Cognitive control in media multitaskers. Proceedings of the National Academy of Sciences of the United States of America, 106(37), 15583-15587.
Pea, R., Nass, C., Meheula, L., Rance, M., Kumar, A., Bamford, H., ... Zhou, M. (2012). Media use, face-toface communication, media multitasking, and social well-being among 8 - to 12 -year-old girls. Developmental Psychology, 48(2), 327-336.
Pew Research Center. (2016). Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies.
Ponsford, J., \& Kinsella, G. (1991). The use of a rating scale of attentional behaviour. Neuropsychological Rehabilitation, 1(4), 241-257. http://doi.org/10.1080/09602019108402257
Przybylski, A. K., \& Weinstein, N. (2012). Can you connect with me now? How the presence of mobile communication technology influences face-to-face conversation quality. Journal of Social and Personal Relationships, 30(3), 237-246.
Ralph, B. C. W., Thomson, D. R., Cheyne, J. A., \& Smilek, D. (2014). Media multitasking and failures of attention in everyday life. Psychological Research, 78(5), 661-669.
Thornton, B., Faires, A., Robbins, M., \& Rollins, E. (2014). The mere presence of a cell phone may be distracting implications for attention and task performance. Social Psychology, 45(6), 479-488.
Weller, J. A., Shackleford, C., Dieckmann, N., \& Slovic, P. (2013). Possession Attachment Predicts Cell Phone Use While Driving. Health Psychology, 32(4), 379387.

# The strategic advantages of micro-targeted campaigning: A proof of principle Bayesian Agent-Based Model 

Jens Koed Madsen (jens.madsen@ouce.ox.ac.uk)<br>School of Geography and the Environment, University of Oxford<br>OX1 3QY, South Parks Road, Oxford, United Kingdom<br>Orcid: 0000-0003-2405-8496

Toby D. Pilditch (t.pilditch@ucl.ac.uk)
Department of Experimental Psychology, University College London WC1E 6BT, Gower Street, London, United Kingdom


#### Abstract

Predicting the effect of persuasion campaigns is difficult, as belief changes may cascade through a network. In recent years, political campaigns have adopted micro-targeting strategies that segment voters into fine-grained clusters for more specific targetting. At present, there is little evidence that explores the efficiency of this method. Through an Agent-Based Model, the current paper provides a novel method for exploring predicted effects of strategic persuasion campaigns. The voters in the model are rational and revise their beliefs in the propositions expounded by the politicians in accordance with Bayesian belief updating through a source credibility model. The model provides a proof of concept and shows strategic advantages of micro-targeted campaigning. Despite having only little voter data allowing crude segmentation, the microtargeted campaign consistently beat stochastic campaigns with the same reach. However, given substantially greater reach, a positively perceived stochastic candidate can nullify or beat a strategic persuasion campaigns.


Keywords: Agent-Based Model; Persuasion; Strategic campaigns; Politics; Voting simulation

## Introduction

Persuasion is paramount in political campaigns, and source credibility is a key component of a successful campaign. It influences a range of human cognitive phenomena related to reasoning, argumentation, and judgement and decisionmaking. It influences the reception of persuasive messages (Petty \& Cacioppo, 1984; Chaiken \& Maheswaran, 1994), plays a vital part in the development of children's perception of the world (Harris \& Corriveau, 2011), impacts juror decision making (Lagnado et al., 2012), increases adherence with persuasion strategies (Cialdini, 2007), and influences how people are seen in social situations (Fiske et al., 2007; Cuddy et al., 2011). The specific normative function of source credibility in argumentation, however, is still debated. For example, the dual-process-based Elaboration-Likelihood Model (Petty, 1981) assigns message source to heuristic rather than analytic cues (Briñol \& Petty, 2009) whilst recent Bayesian models integrate credibility in revising beliefs when given evidence from a
source (Bovens \& Hartmann, 2003; Hahn et al., 2012; Harris et al., 2015).

Trustworthiness is an important factor in politics. It increases public policy compliance (Ayres \& Braithwaite, 1992), influences candidate choice (Citirin \& Muste, 1999), increases intention of voting (Householder \& LaMarre, 2014, though not necessarily actual voting, see Dermody et al., 2010), increases societal cooperation (Fukuyama, 1995), and lack of trust may instigate civic participation (see Levi \& Stoker, 2000 for a discussion of this). The current paper expands upon these findings by showing how source credibility influences the convincingness of an argument for a proposed candidate.

In political literature, credibility factors include integrity, competence, fairness, flip-flopping, honest, equitable, and being responsiveness to public needs (Citrin \& Muste, 1999; Levi \& Stoker, 2000). Collating these, Mayer and colleagues identify credibility as ability, benevolence, and integrity (Mayer et al., 1995; Mayer \& Davis, 1999). Empirical exploration of management corroborates such a definition (Colquitt et al., 2011 for a review) while social psychology partitions reliability in two main spheres: warmth and competence (Fiske et al., 2006; 2007; Cuddy et al., 2011). Further studies in management literature differ in whether they identify two (Jarvenpaa et al., 1998) or three credibility traits (Mayer et al., 1995).

The model employed in the paper is in line with the factors identified in the above studies. Specifically, we divide credibility into two factors: expertise (the capability to provide accurate information) and trustworthiness (the willingness to provide accurate information).

## Micro-targeted campaigns

Political campaigns attempt to persuade voters that they should support and vote for a particular candidate or political position. Unsurprisingly, the competitive nature of electoral campaigns has led to the development of strategies regarding belief updating and behaviour changes, in particular through the use of data and voter segmentation (O'Neil, 2016). Typically, an election campaign is divided into two phases: a persuasion phase that focuses on changing the minds of the voters and a "get-out-the-vote" phase that focuses on making sure the voters do indeed turn out for the election. While the former phase typically lasts
for the duration of the entire campaign, the latter is typically implemented in the final 3-4 days (see Green \& Gerber, 2008). In the current paper, we focus on the element of political campaigns concerned with changing the minds of the voters (i.e. the persuasion phase). Specifically, we explore the potential strategy advantages of knowing the voters' attitudes towards the persuader (the politician).

Companies increasingly accrue data on their customers. Given potential access to and purchase of large-scale data sets about voters, recent years have seen the development of specifically targeted political campaigns, known as microtargeted campaigns (MTCs, see Issenberg, 2012). While traditional campaigns use rough voter segmentation such as by gender, income, or place of habitation, individual voter models allow for fine-grained segmentation (e.g., uppermiddle class, Caucasian, suburban, father, Prius-owner, Seattleite; as well as top-five travel destinations, frequented news sites, etc.). Such data allows for highly specified models of the individual voter concerning political leaning, policy priorities, and voting likelihood. The models allow for targeted political adverts that address specific political issues in a way that is tailored to the individual in question.

There is currently little academic research conducted on the effect and strategic administration of micro-targeted campaigns in elections. First, micro-targeted models are a recent element in election campaigns (Issenberg, 2008). As such, most models actually used in campaigns are subject to non-disclosure agreements and are kept by the responsible companies. Second, it is difficult to assess the quality of campaigns; partly due to the aforementioned secrecy regarding the exact models, and partly due to the complexity of campaigns, given the number of free parameters and the uniqueness of each campaign.

The current paper focuses on changes of electorate beliefs and not on the likelihood of voting (at the end of the simulation (i.e. campaign period), all voters vote with a probability of 1). Election campaigns unfold over time, where campaigns can contact voters and attempt to persuade them. As such, the persuasion attempt of the politician is a successive campaign designed to convince the electorate that the voters should support the persuader in question.

While we do not test specific campaign models, the paper provides a proof concerning the potential effect of micro-targeted campaign strategies through Agent-Based Model simulation of interactions between politicians and voters. We stress the exploratory nature of the study, as the model is necessarily simplified. Rather than testing the predictive power of a specific voter model, we explore the strategic potential of MTCs through a Bayesian source credibility model, which has been developed and tested in previous studies. In the following, we present Agent-Based Models as a technique for exploring the development of aggregate patterns (such as changes in beliefs in a population) across time. Aside from testing the potential effectiveness of MTCs, the paper presents Agent-Based Models (ABMs) as a novel method for simulating the predicted effect of persuasion campaigns.

## A Bayesian source credibility model

Bayesian approaches to reasoning and belief revision take point of departure in subjective, probabilistic degrees of beliefs in propositions where Bayes' theorem captures the posterior degree of belief given a prior belief in the hypothesis and some new evidence (Oaksford \& Chater, 2007). The approach has been applied to argumentation theory (Hahn \& Oaksford, 2006; 2007) where findings suggest that Bayesian reasoning may account for crucial elements of human information integration in practical reasoning. Most relevant to the current model, researchers have used Bayesian approaches to describe how humans integrate uncertain information from more or less reliable sources (Bovens \& Hartmann, 2003). The model has been tested empirically (Harris et al., 2015; Madsen, 2016) and enjoys a good fit with observed responses.

Taking point of departure from the Bayesian source credibility model, credibility is defined as a combination of trustworthiness and expertise (Hahn et al., 2009; Harris et al., 2015; see Fig. 1). In order to implement this model and to facilitate communication between persuaders (politicians) and persuadees (voters) and to capture the desired belief updating process, the members of the electorate have subjectively estimated beliefs about the credibility of each persuader.


Fig. 1: A Bayesian source credibility model ${ }^{l}$
Expertise refers to whether or not the persuader is capable of providing accurate and relevant information. For example, a politician may know the legislative framework connected with a policy proposal (thus increasing the chances of providing legislation that is legal and within constitutional law), a doctor may be more qualified to diagnose an illness compared with a layperson and so forth. Conversely, trustworthiness refers to the intention of providing accurate information. Regardless of the expertise of the source, the speaker might wish to misinform, lie, or otherwise deceive the listener. Expertise and trustworthiness are orthogonal and independent in the model (see Fig. 1) such that a person can be inexpert and yet intend to represent her available information as accurately as possible or a person can be highly expert within a field, but wish to misinform the listener. The source credibility model used in the current paper has previously been tested on appeals to political authority, which suggests that the model captures

[^402]part of how voters update their beliefs when politicians publically endorse or critique a policy (Madsen, 2016).

As explained later, the persuaders in the model contact the voters and provide either a positive or a negative statement concerning a hypothesis. To calculate the expected posterior belief in the likelihood of the hypothesis (e.g. the goodness of the candidate), we apply the source credibility model. The equation used to calculate the posterior is an expanded version of Bayes' theorem that incorporates trustworthiness and expertise within the theorem. It is taken from Harris et al. (2015) and relies on advances made in Bovens and Hartmann (2003) and Hahn et al. (2009) on the concept of source reliability (Hahn et al., 2012).

$$
P(H \mid e)=\frac{P(H) P(e \mid H)}{P(H) P(e \mid H)+P(\neg H) P(e \mid \neg H)}
$$

where $\mathrm{p}(\mathrm{h} \mid$ rep $)$ represents the probability that the hypothesis is true (h) given a confirming statement (rep). $\mathrm{P}(\mathrm{h})$ represents the prior belief in the hypothesis, and p (rep $\mid \mathrm{h}$ ) and $\mathrm{p}(\mathrm{rep} \mid \neg \mathrm{h})$ represent the conditional probability that the source would provide a positive statement if indeed the hypothesis was true or false. Trustworthiness and expertise are integrated within $p(r e p \mid h)$ and $p(r e p \mid \neg h)$ through the combination of these conditionals ${ }^{2}$.

To parameterise the model, $\mathrm{p}(\exp )$ and $\mathrm{p}($ trust ) represent prior beliefs in expertise and trustworthiness. Conditional probabilities (see footnote 1) represent the epistemic relationship between model parameters and the likelihood of providing true or good advice. For example, p(rep|h, exp, trust) refers to the likelihood that a speaker declares a hypothesis to be true when the speaker has complete and perfect knowledge of the topic and is unequivocally trustworthy in a world where the hypothesis happens to be true regardless of the statement of the speaker.

The model allows for parameter-free belief revision such that the agent makes use of its estimation of the persuader's source credibility to update its belief when the persuader contacts the voter with a persuasive statement.

## Agent-Based Modelling

Traditional equation-based models typically take point of departure in cognitive functionality in isolation (e.g. belief revision given new information) or in dyads (e.g. prisoners' dilemma). However, when agents can interact and influence each other through time, across space and between multiple agents, behaviour may become dynamic and adaptive. If this happens, patterns may become computationally intractable, making predictions difficult or impossible with isolated or dyadic models, as the system becomes complex (see Parunak et al., 1998). ABMs allow for simulations of interactions between agents and their environment and between multiple agents (Epstein \& Axtell, 1996). In the

[^403]paper, we use this method to simulate a campaign with interactions between politicians and voters. Each round in an ABM is called a tick. Here, each tick represents a campaign day. For the current model, the ABM requires agents and links between agents.

## Agents

Agents are the actors in the simulated model world. The cognitive make-up of each agent may consist of any rules constrain or enable relevant behaviours within the simulated world. By applying the relevant cognitive rules, agents can revise their beliefs about the model world by interacting with the environment. Further, agents can have physical rules such as metabolism, energy consumption, and age. This allows for simulated life-spans in which agents can live, learn, generate progeny, and die. The physical and cognitive rules allow for heterogeneity, as agents may differ in essential characteristics. This allows for dynamic models of heterogeneous populations.

In the ABM presented in this paper, there are three types of agents: voters, strategic persuaders and stochastic persuaders. The persuaders' aim to convince the voters to support them in an election. Politicians engage with voters by providing a statement, supporting one candidate or the other $(\mathrm{H}=0$ or 1$)$. Voters update their belief in the goodness of each candidate on the basis of the prior beliefs $(\mathrm{p}(\mathrm{h}))$ and their perception of the candidate (using the above Bayesian source credibility model).

## Links

While agents have cognitive rules that apply to agents in isolation, ABMs allow for interactivity. Links represent connections between agents that may be encoded with functional capacities. These can be any and all social links that inform and influence behaviour. Links can be direct (e.g., providing information to another agent, fighting with another agent) or indirect (e.g., some agents might prefer to be in the vicinity of other types of agents). In the current model, only direct links are employed, as the persuaders contact voters directly.

In the current model, politicians establish links by seeking out voters. Stochastic candidates engage randomly with voters while MTC candidates only engage with voter that has a positive attitude towards the candidate (using the 'signal factor' described in the following section). There are no links between voters in the current model. Introducing social structure will be a natural development in future work. Indeed, we strongly suspect that MTC candidates would be more efficient in social structures, as they can target 'community leaders' and important social nodes.

## Simulating Micro-targeted campaign strategies

In order to tentatively explore the effect of MTCs, we simulate the span of an election campaign through an ABM in which the politicians (the persuaders) can interact with the voters (the persuadees). Though exploratory in nature, the model has two aims. First, to our knowledge, although
some models have explored opinion chance in politics (e.g. Duggins, 2016), ABMs have not been used to directly explore campaign strategies. The paper consequently provides a new method for exploring the efficiency of persuasion campaign strategies. Second, by implementing a simplified voter and strategy model, the efficiency of minimal voter knowledge is explored. As the strategic potential of MTCs increases given higher voter complexity (e.g. if voters have prioritised political beliefs), the simplified model explores the efficiency of MTCs in situations where they are expected to be least effective. As such, the model explores a conservative modelling scenario. In the following, we present the Agent-Based Model.

Agents The model consists of three types of agents: Voters, Micro-target persuader (MTC), and stochastic persuader (non-MTC). The physical space plays no role in the current model (as the interactions between the persuader and the persuadee may be likened to sending out pamphlets or generating cold-calls). Consequently, voters are randomly distributed across simulation space. All voters were outfitted with the Bayesian source credibility model to inform their belief revision process. To operationalise the model, each voter generates an expertise and trustworthy score for each candidate from a normal distribution (as described later, we manipulate the means in the two simulations, such that mean $=0.5$ or $=0.6, S D=0.25$ ). To fully parameterize the model, agents are given conditional probabilities:

| $\mathbf{H}$ | $\mathbf{H}$ | $\mathbf{H}$ | $\mathbf{H}$ | $\neg \mathbf{H}$ | $\neg \mathbf{H}$ | $\neg \mathbf{H}$ | $\neg \mathbf{H}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{T}, \mathbf{E}$ | $\mathbf{T}, \neg \mathbf{E}$ | $\neg \mathbf{T}, \mathbf{E}$ | $\neg \mathbf{T}, \neg \mathbf{E}$ | $\mathbf{T}, \mathbf{E}$ | $\mathbf{T}, \neg \mathbf{E}$ | $\neg \mathbf{T}, \mathbf{E}$ | $\neg \mathbf{T}, \neg \mathbf{E}$ |
| 80.38 | 58.21 | 34.63 | 18.04 | 22.59 | 42.3 | 59.90 | 71.26 |

Table 1: Conditional probability table
These are taken from Madsen (2016), as this study applies the model to political belief revision. This allows for belief revision given persuader statements (with no free parameters) and further allows for agent heterogeneity, as some agents will rate one candidate highly while another will rate the same candidate poorly.

To provide a 'signaling' factor for the MTC candidate, voters average trustworthiness and expertise scores to generate a 'credibility score'. Each voter generates a prior belief from a normal distribution (mean $=0.5, S D=0.25$, bounded between 0.01 and 1 ), representing a voter's initial (prior) likelihood of voting for either candidate. This introduces the eventual decision (voting) rule: If a voter has p (candidate) $<0.5$, it favours the non-MTC candidate; if p (candidate) $>0.5$, it favours the MTC candidate. The campaign runs for 50 days ( 50 ticks). At the end of the simulation, voters vote for their favoured candidate. There are 10,000 voters in the simulation.

Both persuaders' aim to persuade voters to shift their p (candidate) towards their own position. In order to do so, they establish connection with voters and make opposite claims. In accordance with the source credibility model, the non-MTC candidate represents $p$ (candidate $)=1$ while the MTC candidate represents $p$ (candidate $)=0$. This gives full implementation of the Bayesian source credibility model
where the voter updates the prior belief given representation by a (more or less) credible source. After each connection, the contacted voter takes P (candidate|rep) - i.e. the posterior - as their new value for p (candidate) ${ }^{3}$.

For each tick, the candidate can establish contact with X voters, defined as 'candidate reach'. In the simulations, the MTC candidate has a reach of 20 while we manipulate the reach of the non-MTC candidate to test the efficiency of the MTC strategy. In Fig. 2 below, the reach ratio is the reach of the non-MTC divided by the reach of the MTC.

Central to the model, the MTC and non-MTC campaigns differ in their contact selectivity. The non-MTC is fully stochastic and thereby corresponds to a blind campaign that distributes leaflets or conducts cold-calls with no knowledge of the electorate. The MTC segments voters and will only contact voters with a favourable impression of the credibility of the candidate (voters with a signaling factor > 0.5 ). This selection process does not take into account the voter's prior belief in the candidate. As such, the underlying source credibility factors determine whether the voter is "open" to the candidate's message (i.e. will update in the desired manner). Of the sub-group of (desirable) voters who fit this criterion, a random selection (the amount based on "reach") are selected for contacting. Both campaigns may contact the same voter multiple times during the simulation, but not more than once on a single "day".

In sum, voters entertain prior beliefs about each candidate, rate each candidate for trustworthiness and expertise, and have a signaling factor. When a candidate contacts the voter, the voter updates the belief in p (candidate) in accordance with the Bayesian source credibility model. Candidates are either stochastic (non-MTC) or use the signaling factor to identify favourably disposed voters. Each candidate can reach a fixed number of voters each click. There are 10.000 voters and 2 candidates, and the campaign lasts for 50 days (ticks). At the end of each simulation, voters cast their vote for the candidate they find most favourable.

## Main findings

We conduct two main manipulations. First, we manipulate the mean credibility rating of each candidate by altering voter perception of candidate trust and expertise (mean $=.5$ or mean $=.6, S D=.25$ ), providing a $2 \times 2$ simulation. Second, for each of the credibility combinations, the reach ratio of the non-MTC is between 1-10 (1 represents simulations where the non-MTC and MTC have identical reach; 10 represents simulations where the non-MTC can reach 10 times as many voters per tick). The reach of the MTC candidate is always 20 . Fig. 2 illustrates the percentage of voters who supported the non-MTC on the $y$ axis and the reach ratio of the non-MTC on the x -axis.

[^404]

Fig. 2: Election outcome $(p($ cred $)=.5 ; p($ cred $)=.6)$
The simulations point to two main conclusions. First, a nonMTC with mean credibility rating of .5 is inefficient when run stochastically. As opinions about the candidate are equally divided, a stochastic strategy will necessarily engage with an equal number of supporters and adversaries. As such, blind strategies only work in when the campaign expect general voter estimation of the candidate to be $>.5$. If the candidate is seen as distrustful, a blind campaign will be ineffective or detrimental and will be beaten by campaigns with simple winnowing strategies.

Second, though MTCs provide a distinct advantage in terms of persuasion strategies, the stochastic campaign can beat the strategic campaign through brute force if the average $p$ (cred) of the stochastic candidate is $>0.5$. If the reach of the non-MTC is roughly double, the effect of the MTC is cancelled out. If the reach ratio $>2$, the non-MTC edges out the strategic campaign. This is an interesting finding, as the MTC is effective, but can be beaten. Given the possibility of simulating and calculating a tipping point where a stochastic candidate (with credibility rating > . 5 among the voters) beat strategic campaigns, it is possible to conduct cost-benefit analyses to determine the best available strategy given a limited campaign budget. In general, though highly simplified, the simulations show that it is methodologically possible to estimate the expected effect of a strategic (or stochastic) persuasion campaign by applying cognitive rules to the persuaders and persuadees in a dynamic environment.

## Discussion and future developments

The current model provides an important proof of concept that MTCs have a non-trivial advantage in a limited world where the voters revise their beliefs in the same manner and where the candidates can only advocate their position in a simplified way. We believe the paper provides a novel method for simulating and analysing electoral strategies using Agent-Based Models. However, as a proof of concept, this leaves room for further model developments.

First, real-life voters may exhibit individual differences concerning moral foundation (Haidt, 2012) or reasoning strategies (Lodge \& Taber, 2013). Voters in the current model are cognitively homogenous (though epistemically heterogenous) who revise their beliefs by the same process. Future work could integrate cognitive voter heterogeneity,
which would allow for exploration of strategic choices. Adding personality profiles would make the model more realistic and interesting in terms of testing election strategies for actual elections. Further, real-life campaigners do not have a perfect signal from each voter. Consequently, noise needed to be added to voter signalling.

Second, in the model, voters consider one proposition whereas elections often consist of a multitude of attitudes, beliefs, and desires. The present framework may easily be extended to include multiple policy beliefs, preference rankings, and multiple candidates.

Third, the MTC only considers the favourability of the candidate. Given additional data about complex voters (e.g., policy preference, personality, etc.), a sophisticated MTC may target voters more strategically. For example, an MTC could differentiate between swing and secure voters. Additionally, some voters are more likely than others to vote regardless of their political conviction. This is essential for strategic implementation of get-out-the-vote strategies.

Fourth, in the model, voters cannot communicate with each other. To allow for greater belief diffusion and for a more dynamic simulation of an electorate, it is reasonable to assume voter interaction where voters can share beliefs and persuade each other through their individual networks. Models that explore the role of hierarchy in opinion dynamics would be particularly relevant to explore this function (see e.g., Quattrociocchi et al 2014; Watts \& Dodds, 2007). Some voters might be communal leaders and have more impact than others. Given weighted network structures of the electorate, it would be possible to simulate complex persuasion strategies. This would simulate the relative efficiency of MTCs in highly complex, highly dynamic, and highly adaptive elections.
We predict that models with more complex voter belief systems, individual voter differences, and with interaction between voters will yield much higher benefits to MTCs. That is, we predict a positive correlation between available electorate data and the efficiency of an MTC. Concurrently, we also predict a positive correlation between the complexity of the electorate and the cost of running an MTC, as complex segmentation requires more data and sophisticated models.

By applying a cognitive updating rule in an Agent-Based Model, the paper presents a new methodology for simulating dynamic persuasion campaigns and for estimating their expected effect. We show a strategic advantage of MTCs. In the simulations, non-MTCs require double reach to cancel out this advantage. Despite having only simple voter data that allows for crude segmentation and a very crude selection strategy, the MTC consistently bested stochastic a candidate with the same reach (even when the opponent had a greater average credibility). However, given greater reach, a positively perceived stochastic candidate can beat a strategic candidate.

## References

Ayres, I. \& Braithwaite, J. (1992) Responsive Regulation, Oxford University Press
Bovens, L., \& Hartmann, S. (2003). Bayesian epistemology. Oxford: Oxford University Press.
Briñol, P., \& Petty, R. E. (2009). Source factors in persuasion: A self-validation approach, European Review of Social Psychology, 20, 49-96.
Chaiken, S. \& Maheswaran, D. (1994) Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement, Journal of Personality and Social Psychology 66 (3), 460-473
Cialdini, R. B. (2007) Influence: The Psychology of Persuasion, Collins Business
Citrin, J. \& Muste, C. (1999) Trust in government, in Robinson, J. P.. Shaver, P. R., Wrightsman, L. S. (Eds.), Measures of political attitudes. Measures of social psychological attitudes, Vol. 2. (pp. 465-532), San Diego, CA, US: Academic Press
Cuddy, A. J. C., Glick, P. \& Beninger, A. (2011) The dynamics of warmth and competence judgments, and their outcomes in organizations, Research in Organizational Behavior 31, 73-98
Dermody, J., Hanmer-Lloyd, S. \& Scullion, R. (2010) Young people and voting behaviour: alienated youth and (or) an interested and critical citizenry? European Journal of Marketing 44 (3-4), 421 - 435
Duggins, P. (2016) A Psychologically-Motivated Model of Opinion Chance with Applications to American Politics, arXiv:1406.7770
Epstein, J. \& Axtell, R. (1996) Growing Artificial Societies: Social Science from the Bottom Up, MIT Press
Fiske, Susan T., Cuddy, A. J. C. \& Click, P. (2007) Universal dimensions of social cognition: warmth and competence, Trends in Cognitive Sciences 11 (2), 77-83
Fukuyama, F. (1995) Trust, New York: Basic Books
Green, D. P. \& Gerber, A. S. (2008) Get out the vote: How to increase voter turnout, Brookings Institution Press
Haidt, J. (2012) The Righteous Mind: Why good people are divided by politics and religion; Allen Lane
Hahn, U., \& Oaksford, M. (2006) A normative theory of argument strength, Informal Logic 26, 1-24
Hahn, U., Oaksford, M., \& Harris, A. J. L. (2012). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), Bayesian Argumentation (pp. 1538). Dordrecht: Springer.

Harris, A. J. L., Hahn, U., Madsen, J. K., \& Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach, Cognitive Science 40, 1496-1533
Harris, P. L., \& Corriveau, K. H. (2011). Young children's selective trust in informants, Philosophical Transactions of the Royal Society B, 366, 1179-1187
Householder, E. E. \& LaMarre, H. L. (2014) Facebook Politics: Toward a Process Model for Achieving Political

Source Credibility Through Social Medie, Journal of Information Technology \& Politics 11 (4), 368-382
Issenberg, S. (2012) The Victory Lab: The secret science of winning campaigns, Broadway Books
Jarvenpaa, S. L., Knoll, K., \& Leidner, D. E. (1998) Is anybody out there? Antecedents of trust in global virtual teams, Journal of Management Information Systems 14, 29-64
Lagnado, D., Fenton, N. \& Neil, M. (2012) Legal Idioms: A framework for evidential reasoning, Argumentation \& Computation 1, 1-18
Levi, M. \& Stoker, L. (2000) Political Trust and Trustworthiness, Annual Review of Political Science 3, 475-507
Lodge, M. \& Taber, C. S. (2013) The Rationalizing Voter, Cambridge University Press
Madsen, J. K. (2016) Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates’ opinions, Papafragou, A., Grodner, D., Mirman, D., \& Trueswell, J.C. (Eds.) Proceedings of the 38th Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Science Society, 165-170
Mayer, R. C. \& Davis, J. H. (1999) The effect of the performance appraisal system on trust for management: A field quasi-experiment, Journal of Applied Psychology 84, 123-136
Mayer, R. C., Davis, J. H. \& Schoorman, F. D. (1995) An integrative model of organizational trust, Academy of Management Review 20, 709-734
Oaksford, M. \& Chater, N. (2007) Bayesian Rationality: the probabilistic approach to human reasoning, Oxford University Press, Oxford: UK
O'Neil, C. (2016) Weapons of Math Destruction: How big data increases inequality and threatens democracy, Allen Lane
Parunak, H. v. D., Savit, R. \& Riolo, R. L. (1998) AgentBased Modeling vs. Equation-Based Modeling: A case study and users' guide, (pp. 10-25) Proceedings of multiagent systems and agent-based simulation, Springer
Petty, R. E., \& Cacioppo, J. T. (1981). Attitudes and persuasion: Classic and contemporary approaches. Boulder, CO: Westview Press.
Petty, R. E. \& Cacioppo, J. T. (1984) The Effects of Involvement on Responses to Argument Quantity and Quality: Central and Peripheral Routes to Persuasion, Journal of Personality and Social Psychology 46 (1), pp. 69-81
Quattrociocchi, W., Caldarelli, G. \& Scala, A. (2014) Opinion dynamics on interacting networks: media competition and social influence, Scientific Reports 4:4938, 1-7
Watts, D. J. \& Dodds, P. S. (2007) Influentials, Networks, and Public Opinion Formation, Journal of Consumer Research 34 (4), 441-458

# Growing a Bayesian Conspiracy Theorist: An Agent-Based Model 

Jens Koed Madsen (jens.madsen@ouce.ox.ac.uk)<br>School of Geography and the Environment, University of Oxford OX1 3QY, South Parks Road, Oxford, United Kingdom<br>Orcid: 0000-0003-2405-8496

Richard Bailey (richard.bailey@ouce.ox.ac.uk)<br>School of Geography and the Environment, University of Oxford<br>OX1 3QY, South Parks Road, Oxford, United Kingdom

Toby D. Pilditch (t.pilditch@ucl.ac.uk)
Department of Experimental Psychology, University College London WC1E 6BT, Gower Street, London, United Kingdom


#### Abstract

Conspiracy theories cover topics from politicians to world events. Frequently, proponents of conspiracies hold these beliefs strongly despite available evidence that may challenge or disprove them. Therefore, conspiratorial reasoning has often been described as illegitimate or flawed. In the paper, we explore the possibility of growing a rational (Bayesian) conspiracy theorist through an Agent-Based Model. The agent has reasonable constraints on access to the total information as well its access to the global population. The model shows that network structures are central to maintain objectively mistaken beliefs. Increasing the size of the available network yielded increased confidence in mistaken beliefs and subsequent network pruning, allowing for belief purism. Rather than ameliorating and correcting mistaken beliefs (where agents move toward the correct mean), large networks appear to maintain and strengthen them. As such, large networks may increase the potential for belief polarization, extreme beliefs, and conspiratorial thinking - even amongst Bayesian agents.


Keywords: Conspiratorial thinking; Extreme beliefs; AgentBased Models; Bayesian updating

## Introduction

Truth is the shattered mirror strewn in myriad bits; while each believe his little bit the whole to own
Richard Burton (The Kasîdah of Hâjî Abdû El-Yezdî)
In recent years, scientists, scholars, and commentators have remarked upon the apparent rise of epistemic echo chambers (see e.g., Bakshy et al., 2016) and increasing political polarization. Echo chambers refer to communities, online or otherwise, that interact predominantly with themselves and who rarely, if ever, seek information aside from the information available within the chamber. Whether endogenously created (such as cults) or exogenously created (such as living on an island with no contact to the outside world), the emergence and maintenance of epistemically enclosed systems and their consequences is interesting and worth studying. The current paper explores the possibility of generating, maintaining and strengthening encapsulated belief communities through an Agent-Based Model (Gilbert,
2008) where every agent is rational (here, Bayesian) and where information is potentially available to challenge the viewpoint of the agent.

Specifically, we are interested in exploring the possibility of generating conspiratorial beliefs. That is, beliefs that are maintained despite being objectively false and there being available evidence to challenge or disprove the theory in question. Proponents of such beliefs frequently hold these positions strongly. We explore whether it is possible to strengthen confidence in objectively mistaken beliefs through a rational process given imperfect knowledge about the world. Rather than assuming illegitimate updating processes or special cognitive functionality, the model tests if, in principle, a Bayesian conspiracy theorist can emerge and be maintained. That is, the model explores whether or not individual differences are a necessary requirement for the emergence and maintenance of extreme beliefs.

The Burton quote at the top of the introduction can be seen as foundational for the paper. It suggests that beliefs can be generated and maintained as a fragment of a larger, and often very different, picture. Further, it intimates that humans generate inferences about the world on the back of the evidence available to us at any given time in our lives. This information may come through first-hand experience or through other sources such as parents, peers, media outlets, and experts.

In order to set the scene for the Agent-Based Model, we briefly consider how conspiratorial thinking has previously been approached in the literature.

## Conspiratorial thinking

Conspiracy theories can be loosely defined as beliefs that are held strongly when evidence is broadly available to challenge or entirely refute the theory. Yet, proponents maintain (and might even strengthen) their belief in the theory despite the availability of this evidence. However, in order to adequately simulate emerging conspiracy theories, we need to employ a more stringent definition of conspiratorial thinking.

According to Barkun (2003), conspiracy theories are characterized by three traits. First, conspiracy theories operate under the assumption that nothing happens by accident. From a cognitive perspective, this may be described as causal oversensitivity where the reasoner generates causal links between disparate and supposedly separate pieces of information, leading to over-connection. Second, Barkun argues that for conspiracy theorists, nothing is really what it appears to be on the surface (i.e. the 'real' causal mechanisms between pieces of information is covered up(?) by any official story). This element, too, suggests a cognitive agent who over-weights and overgenerates causal links between independent pieces of information. For example, some proponents of the moon landing conspiracy theory believes the director of A Space Oddysey: 2001, Stanley Kubrick, to be involved in producing faked photography because Kubrick hired crew for 2001 who used to work for NASA (Frederick Orway and harry Lange). Finally, Barkun argues that conspiracy theorists tend to believe things to be highly connected. To this end, Barkun argues that conspiracy theories eventually become enclosed systems that are falsifiable if confronted with additional evidence and therefore "a matter of faith rather than proof". Presumably, this entails that conspiracy theorists might stop seeking new information and instead assume their beliefs to be a priori true. As evident from these definitions, the operationalizing of conspiracy theories usually involves special cognitive make-up and a heuristic process that treat information related to the conspiracy theory as qualitatively different from 'normal' belief updating. The current model explores whether these are valid assumptions.

Indeed, Birchall (2006) describes conspiratorial thinking as illegitimate updating and belief maintenance (as opposed to normative, legitimate reasoning). In general, conspiratorial thinking is typically conceived as an abnormal and potentially fallacious (or illegitimate) reasoning process, which relies heavily on cognitively biased heuristics such as over-generation of causal links, erroneous attribution of motives, and mistaken perception of interconnectivity. Commonly, these conspiratorial thinking accounts assume conspiracies are a product of mistaken or misguided reasoning.

In this paper, we provide a proof of concept that conspiratorial thinking can emerge from Bayesian rational paradigms given access to a subset of evidence and the possibility of interacting with other like-minded agents. As will be argued later, we believe both of these assumptions to be realistic and grounded in psychological findings. As will be explained further in the paper, we show that conspiratorial agents do not require special cognitive abilities or predispositions in order to be supremely confident in their (objectively mistaken) belief. This approach is reminiscent of work conceptualizing supposed reasoning flaws through cognitively reasonable processes. This work includes, but is not limited to, Bayesian accounts of argument fallacies (e.g., Corner et al., 2011; Harris et al.,
2012), a Bayesian model of appeals to authority (e.g. Hahn et al., 209; Harris et al., 2015), and skepticism in climate change (Cook \& Lewandoswky, 2016). Further, Bayesian agents represent a rational process of integrating new information with prior beliefs pertaining to that hypthesis. For this reason, Bayesian agents have been used previously to explore belief diffusion in networks (see e.g. Jern et al., 2009; Olsson, 2013; Denrell \& Le Mens, 2017).

While the current work builds on similar Bayesian accounts of belief updating, we provide a novel contribution to the field by implementing a computational, Agent-Based Model that allows for interaction between agents across time.

For the purpose of this paper, we take conspiratorial thinking to be strongly held beliefs that depart from the objective mean where evidence is available to challenge or refute the theory. Barkun and Birchall argue that these beliefs arise from mistaken or flawed heuristics and/or illegitimate reasoning processes that bias proponents of conspiracy theories toward connectivity, attribution of hidden intentions and over-generation of causal structures. Further, Grimes (2016) argues that conspiratorial beliefs are untenable with larger network structures, as the available information to challenge erroneous views increases. As discussed below, there are some potential challenges for conspiracy accounts that assume special cognitive functions such as oversensitivity toward causal connections as the default cognitive foundation.

## Challenges for traditional accounts

The traditional perspectives on conspiratorial thinking may be challenged on at least two grounds. First, it is potentially problematic to ascribe different cognitive functions to the emergence of conspiratorial thinking for two reasons. For one, it is unclear whether this type of reasoning would permeate all beliefs held by that individual (e.g. would a conspiracy theorist also be prone to over-generate causal structures in billiards or snooker). If it were not systemic, it would (insufficiently) appear to be a post hoc account of a particular belief that happens to exhibit such properties. For another, it would not represent a process account of conspiratorial thinking. Rather, it would assume differences and apply these to arrive at the conclusion. Instead, we explore whether it is possible to generate objectively mistaken beliefs from the same cognitive processes that generate objectively true beliefs. Both of these would remove the expectancy of abnormality on the part of the conspiracy theorist.

Second, traditional accounts tend to focus on the cognitive function of the isolated individual, rather than on systemic belief diffusion as a result of interactions with other people. As discussed in the following, studies and simulations have shown that aggregate behavioral patterns might not be reducible to the components in isolation if the components can interact with each other in meaningful ways (see e.g. Johnson, 2001; Ball, 2005) given complex and dynamic environments (Johnson, 2009). Faced with the problem of
epistemic isolation, we apply Agent-Based Modeling to explore the potential of growing a Bayesian conspiracy theorist without adding special cognitive functions to the agent in question.

## Agent-Based Modeling

In order to circumvent the problems caused by traditional individual-based accounts of cognition, we employ AgentBased Modeling, which allows for simulations of belief networks populated by Bayesian agents. This further allows for introduction of heterogeneity, as will be discussed later (here, initial sampling allows agents to gather and evaluate data individually, which provides heterogeneous priors). Agent-Based Models (ABMs) are computer simulated multi-agent systems that describe the behavior of and interactions between individual agents, who operate in synthetic environments (Gilbert, 2008; Bandini et al., 2009). Agents are encoded on a computational basis, and may implement and explore models of cognitive function. They allow for complex, dynamic and adaptive systems to emerge through interactions between agents and with the environment as well as across time (Miller \& Page, 2007). ABMs may be described in terms of their three fundamental components: Agents, Links, and Patches.

Agents are the nodes representing the active cognitive entities of the system. They can make decisions and make use of information in any way that is formally expressible. These functions include, but are in no way limited to, utility valuations, Bayesian belief inferences, stock market engagement, and so forth. The agents may reproduce (e.g. give birth to a new agent), move around the simulated space, and make new (and potentially more relevant) decisions as they learn more about the environment. In order to engage with the environment, agents will have specified rules for agent-environment interactions such as fishing, purchasing a house, moving around the simulated space and so forth. These behaviors and inferences may yield dynamic and adaptive aggregate behavioral patterns. For example, if all agents harvest simultaneously, Tragedy of the Commons type problems (Ostrom, 2012) can emerge. In the present model, we allow for Bayesian belief updating as the agents encounter new information or talk with other agents via links.

Links represent rules for possible interactions between agents. Links can be any interactivity that can be expressed formally. The interaction may be direct (e.g. communication between two agents or sales structure between agents, see Epstein \& Axtell, 1996) or indirect (e.g. social attraction or repulsion or emotional feedback, see Schelling, 2006; Epstein, 2013). Interactions allow for feedback loops to emerge, which in turn may generate aggregate behavioral patterns that are irreducible to the components in isolation.

Patches represent the simulated environment in which agents exist. They can have any and all properties that are formally describable. If consumable (such as grass for sheep, fish for fishers), they may give the agent energy, money, or other affordances. Patches may be dynamic such
that they might regrow or migrate. Further, patches may facilitate or restrict movement of agents in the simulated space. The patches provide the foundational and potentially dynamic environment in which the agents live and act. In the model we present, the environment restricts interaction between agents if the search potential is low.

Compared with traditional methods, ABMs are capable of simulating dynamic and adaptive decision-making in changeable environments (Miller \& Page, 2007). This allows for agents to self-organize without hard-wiring expected aggregate behavioral patterns such as emergent echo chambers. Rather, ABMs allow for these properties to emerge, or, in the terminology of Epstein and Axtell (1996), to grow. ABMs further allow for agent and environmental heterogeneity (i.e. agents with different cognitive capabilities).

## Growing a Bayesian conspiracy theorist

The aim of the current model is to test a proof of principle that conspiracy theorists can emerge through entirely rational processes without providing any special cognitive functions, heuristic strategies, or access to unique information. In order to do so, we generate an Agent-Based Model where agents can sample information, communicate with one another, and update their beliefs about the world.

Given this initial proof of concept, we simplify the epistemic challenge and consider only one abstract belief. The true probability of the Gaussian distribution from which the agents sample is 0.5 . The standard deviation can be manipulated to represent greater or lesser noise in the information environment. In the present paper, the standard deviation is set to 0.2 . For the sake of understanding, the probability may represent the belief in the fairness of a coin. If the coin is fair, the distribution of tosses is trivially 50-50 between heads and tails. However, if the coin is not fair, the distribution can be skewed in the direction of either heads or tails. Understood in this way, the agents try to understand if they are in a world in which the coin is fair (uncovering, as it were, the true, underlying probabilities) or if they are in a world where the coin is rigged to either side (arriving at an objectively mistaken belief).
If agents are able to generate, maintain and possibly strengthen a mistaken belief in the epistemic state of the belief, the agent will have exhibited conspiratorial traits, as this fulfills the criterion for the definition in the above: a potentially strongly held, yet objectively mistaken belief, availability of information to challenge or refute the theory, and access to that information. The literature review uncovered two central positions that we explore here. One, we explore Grimes' (2016) argument that conspiratorial thinking is untenable in a large network structure. If this is true, we should see a global regression towards the objectively true mean given larger networks (that is, fewer agents who believe they are in a rigged coin world). Two, we explore Barkun and Birchall's arguments that conspiratorial thinking relies on illegitimate reasoning and biased heuristics. As will be described below, the agents in
the model are perfect Bayesian reasoners. If conspiratorial thinking requires special cognitive properties, we should not expect the Bayesian agents to generate strong and mistaken beliefs about the world. The model implements six key elements: generation of prior beliefs, constrained search, network generation, communication between agents, belief updating, and network pruning.

In order to generate a subjective prior belief, agents are born onto the world and sample randomly generated data from a Gaussian distribution ( $\mu=0.5, \sigma=0.2$ ). In a frequentist manner, these are used to calculate a perceived mean and probability density. The sampling represents the worldview of each particular agent before they are able to communicate with other agents.

Having generated a prior belief for each agent (and thus introduced sampling heterogeneity), the model relies on four additional assumptions and mechanisms. First, agents cannot sample all available data in the simulated world. This means that they do not have access to all data sampling that other agents have encountered unless they communicate with the other agent in order to learn the beliefs of that agent. As such, agents do not have perfect and complete knowledge about the world in which they live. We believe this is a reasonable assumption, as humans do not have perfect knowledge in real life. Second, although all other agents are hypothetically available, agents cannot communicate to every other agent in the simulated world. Rather, each agent randomly generates the amount of possible communication links. Like the first assumption, we believe this is a reasonable assumption, as humans in the real world cannot communicate with every other person on the planet, but has to settle for a subset of all living persons.

Third, in order to make the agents rational, they update their beliefs about the world in a Bayesian manner. Bayesian updating represents the rational integration of prior beliefs with new evidence to generate posterior belief in the hypothesis. This approach has been applied to a host of related phenomena such as argumentation (Hahn \& Oaksford, 2006; 2007), source credibility (Bovens and Hartmann, 2003; Harris et al., 2015), and reasoning and decision-making (Oaksford \& Chater, 2007). The integration is formally expressed through Bayes' theorem

$$
P(h \mid e)=\frac{P(h) P(e \mid h)}{P(e)}
$$

where $\mathrm{p}(\mathrm{h} \mid \mathrm{e})$ denotes the posterior belief in the hypothesis (h) given the evidence (e). As such, agents treat each new encounter as a data point to be integrated within their subjective probability density function. Bayesian updating ensures that the agents are fully rational in their belief revision when encountering new evidence.

Finally, several studies on confirmation bias, selectivity bias, and in-group behavior strongly suggest that agents are not entirely stochastic and non-directed in their information search. Taking inspiration from segregation studies (e.g., Schelling, 2006), we introduce a mild preference for people who remotely share their beliefs about the world. The agents
are relatively tolerant and will engage in conversation with any other agent who is within $\pm 1.5$ standard deviations of its own perception of the world. Given Gaussian distributions, this means that the agent will speak to $86.6 \%$ of people within its belief distribution. Thus, they are willing to talk to and integrate information from agents who have different viewpoints than their own. However, if they are confident in their belief, they will engage with less diverging viewpoints, as their probability density narrows. As an analogy, this means that an agent might be willing to discuss political questions with people with different points of view, but would refuse to engage in discussion with people who believe that fair coin-flips are 60-40 rather than $50-50$ in cases where they are absolutely certain about the latter and less certain about the former.
In sum, the agent is born into the world by sampling randomly generated pieces of information related to the hypothesis in question. This informs the mean and standard deviation of their prior. Second, the agents generate networks with other agents within their network radius (which may be limited or encompass the full system). Having set up the model, the agents will communicate freely and honestly (i.e. representing their belief in the hypothesis to the best of their ability), which enables Bayesian belief updating. Agents will maintain communication networks with other agents who are within 1.5 standard deviation of their subjective understanding of the world (i.e. their belief in the hypothesis). If agents within the network fall outside of those boundaries, the agent deactivates the network contact with that particular agent. If agents cannot find any suitable agents within their range, they decrease confidence (simulating negative feedback) and thus expand acceptable search parameters for the following tick. This allows for dynamic network pruning (Ngampruetikorn \& Stephens, 2015).

## Main findings: Limited and extended networks

We implemented the above model in NetLogo (5.2.1) and manipulated the model in terms of the size of the network. For limited networks, agents had a search range of 10 of 100 (as a product of their geographical location). Extended networks, on the other hand, had a search range of 80 of 100. Agents could connect to and sample randomly from other agents within agent search range who fall within their network criteria. Figs 1 a and 1 b show the extent to which search capability influences network generation.


Figures $1 a$ and 1 : Limited and extensive networks
The overall belief structure did not differ significantly between limited and extended networks. Some, but not all agents regressed towards toward the mean while some agents retained their objectively mistaken belief (see Fig. 2a and 2 b , which are histograms where number of believers are
on the y axis and agent belief is on the x axis), we observe differences in belief confidence. As seen in Figs. 3a and 3b, extended networks allowed for interactions with increasingly like-minded agents, which in turn increased belief confidence. This is true both for agents who obtain objectively true and false beliefs. As agents become increasingly confident, their probability density narrows, meaning that they are less willing to engage with agents with differing beliefs. Extended networks allow them to form and maintain contact with agents who share their specific beliefs such that they increase their confidence in that particular view of the world. This means purification of beliefs and purification of networks, i.e. the emergence of epistemic echo chambers.


Figures $2 a$ and $2 b$ : Limited and extended belief structures


Figures $3 a$ and 3b: Limited and extended confidence (0-1)
Overall, the model shows that fully rational agents can maintain and potentially strengthen objectively mistaken beliefs. Further, given a mild preference for interaction with like-minded agents, we observe the rise echo chambers. This effect is strengthened with the size of the network. Rather than making extreme beliefs untenable as predicted by Grimes, we show that large networks, here quantified in terms of the number of reachable agents for any given agent, can engender extreme belief maintenance and belief purism.

## Discussion and concluding remarks

The Agent-Based Model in the paper provides a theoretical proof of concept that a Bayesian agent can become an ardent conspiracy theorist under three main assumptions. One, the agent does not have perfect and full access to all available information that exists in the world, but can only sample a sub-set of that information. This means that the agent does not rely on perfect knowledge of the system. Depending on the practical conceptualization of information accessibility, the agent may have access to very limited or more extended amounts of information. Two, the agent cannot talk to every other person in the world, but can only talk to a sub-set of all existing agents. Similar to assumption one, this means the agent cannot converse with all other agents and learn their subjective access to information. In the current model, information after prior sampling is
gleaned through interactions with other agents. Consequently, by limiting the amount of other agents with whom an agent can engage, the model naturally also limits the access to available information. Principles one and two are concerned with the degree to which the agent can sample information and learn about the world. Three, agents search for and interact with other agents on the basis of their current worldview. They are willing to communicate with most other agents, but avoid other agents with whom they radically disagree about the nature of the world.

## The Rise of Echo Chambers

Together, these three (we believe reasonable) assumptions show that larger networks do not yield belief amelioration (as was postulated by some theoreticians who believed the Internet to facilitate greater communication between people and thereby allow a global regression towards the mean). Rather, the model shows that extended networks, given plausible constraints to exposure, lead to the growth of echo chambers and eventual belief purism, whereby agents increasingly discard those who do not share their specific beliefs about the world.

One might compare this increasing belief purism to development of political ideologies. In a limited network structure (e.g., a small village), the model suggests that leftleaning voters are willing to communicate with other leftleaning voters (and some right-winged voters depending on the mean and probability density function of the specific voter, mutatis mutandis for right-winged voters). However, in an extended network structure (such as a metropolis or Facebook), the model suggests that voters will have access to other voters who have more similar worldviews. This allows for emergence of political echo chambers where extreme voters have access to other extreme voters. From this, greater belief confidence grows and network pruning increases, as belief purism emerges. We therefore expect increases in network structures will facilitate rather than hinder belief extremism and confidence in worldviews.

The model presented in the current paper allows for this dynamic adaption. In the beginning, agents cluster around people with whom they share general beliefs about the world. However, as they increase in confidence, their probability densities narrow, meaning that fewer agents will fit within the $\pm 1.5$ standard deviations of the perceived mean. As the agent becomes increasingly confident in its own reading of the world, it will be decreasingly inclined to engage with agents who entertain different viewpoints. This allows for belief communities to fracture and radical and supremely confident cells to emerge. The emergent echo chambers function as cyclical maintenance of a peculiar belief.

This finding is interesting because larger networks did not yield belief amelioration, but rather belief solidification. It opens up for a novel way to approach and model epistemic communities that maintain strong beliefs despite available data challenging their beliefs (e.g., creationists, climate skeptics, and radicalized or discriminatory beliefs).

## Emergence of reasonably mistaken views

Central to the model, the agents do not have full and perfect knowledge of the world and can only talk to a sub-set of other existing humans. Given the fact that agents update their beliefs in a Bayesian manner, their cognitive system can be described as rational and entirely reasonable. Yet, given incomplete access to data and given the network properties, the model shows that the agents can become entirely confident in objectively mistaken views. As such, we show that extreme beliefs such as conspiracies could emerge through entirely rational processes. While this does not preclude heuristic strategies or special cognitive functions, the model shows that these are not necessary for strongly held mistaken beliefs to emerge. Aside from emerging, mistaken beliefs are also able to survive (and even strengthen) in such an environment rather than being swallowed by mainstream beliefs.

Further, agents had a mild preference for communicating with like-minded agents. Rather than making extreme beliefs untenable, the model suggests that increasing the size of the network intensifies the process of radicalization and augments the confidence even in an objectively mistaken belief. In the age of the Internet, this finding is worth considering seriously and exploring further

In conclusion, we have provided a proof of concept that shows the impact of network structures in generating and maintaining extreme beliefs such as conspiratorial thinking. A Bayesian agent can generate and even increase its confidence in objectively mistaken beliefs.

## References

Ball, P. (2005) Critical Mass: How one things leads to another, Random House, London: UK
Bakhsy, E., Messing, S. \& Adamic, L. A. (2016) Exposure to ideologically diverse news and opinion on Facebook, Science 348 (6239), 1130-1132
Bandini, S., Manzoni, S. \& Vizzari, G. (2009) Agent Based Modeling and Simulation: An Informatics Perspective, Journal of Artificial Societies and Social Simulation 12 (4), 1-16

Barkun, M. (2003) A Cultural of Conspiracy: Apocalyptic Visions in Contemporary America, University of California Press
Birchall, C. (2006) Knowledge Goes Pop: From Conspiracy Theory to Gossip, Berg Publishers, Oxford: UK
Bovens, L., \& Hartmann, S. (2003). Bayesian epistemology. Oxford: Oxford University Press
Cook, J. \& Lewandowsky, S. (2016) Rational irrationality: Modeling climate change belief polarization using Bayesian networks, Topics in Cognitive Sciences 8, 160179
Corner, A., Hahn, U. \& Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. Journal of Memory \& Language, 64, 133-152.

Denrell, J. \& Le Mens, G. (2017) Information Sampling, Belief Synchronization, and Collective Illusions, Management Science 63 (2), 528-547
Epstein, J. (2013) Agent_Zero: Toward Neurocognitive foundations For Generative Social Science, Princeton University Press
Epstein, J. \& Axtell, R. (1996) Growing Artificial Societies: Social Science from the Bottom Up, MIT Press
Gilbert, N. (2008) Agent-Based Models, SAGE Publications
Grimes, D. R. (2016) On the Viability of Conspiratorial Beliefs, PLoS One 11 (1), e0147905
Hahn, U., Harris, A. J. L., \& Corner, A. (2009). Argument content and argument source: An exploration. Informal Logic, 29, 337-367.
Hahn, U., \& Oaksford, M. (2006) A normative theory of argument strength, Informal Logic 26, 1-24
Hahn, U., \& Oaksford, M. (2007) The rationality of informal argumentation: A Bayesian approach to reasoning fallacies, Psychological Review 114, 704-732
Harris, A. J. L., Hahn, U., Madsen, J. K., \& Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach. Cognitive Science 40, 1496-1533
Harris, A., Hsu, A. \& Madsen, J. K. (2012) Because Hitler did it! Quantitative tests of Bayesian argumentation using Ad Hominem, Thinking \& Reasoning 18 (3), 311-343
Jern, A., Chang, K-M. \& Kemp, C. (2009) Bayesian belief polarization, Advances in Neural Information Processing Systems 22 (NIPS 2009)
Johnson, N. (2009) Simply Complexity: A Clear Guide to Complexity Theory, Oneworld Publications
Johnson, S. (2001) Emergence, Penguin Publications
Miller, J. H. \& Page, S. E. (2007) Complex adaptive systems: An introduction to computional models of social life, Princeton University Press
Ngampruetikorn, V. \& Stephens, G. J. (2015) Bias, Belief, and Consensus: Collective opinion formation on fluctuating networks, arXiv 1512.09074 v 1
Oaksford, M. \& Chater, N. (2007) Bayesian Rationality: The probabilistic approach to human reasoning. Oxford, UK: Oxford University Press
Olsson, E. J. (2013) A Bayesian simulation model of group deliberation and polarization, in Zenker, F. (Ed.) Bayesian Argumentation, Synthese Library, Springer
Ostrom, E. (2012) The Future of the Commons: Beyond Market Failure and Government regulation, The Institute of Economic Affairs
Schelling, T. (2006) Micromotives and Macrobehavior, Norton and Company, New York: NY

# The dilution effect: Conversational basis and witness reliability 

Jens Koed Madsen (jens.madsen@ouce.ox.ac.uk)<br>School of Geography and the Environment, University of Oxford<br>OX1 3QY, South Parks Road, Oxford, United Kingdom<br>Orcid: 0000-0003-2405-8496

Ulrike Hahn (u.hahn@bbk.ac.uk)<br>Department of Psychological Sciences, Birkbeck College, University of London<br>WC1E 7HX, Malet Street, London, United Kingdom<br>Marion Vorms (mvorms@gmail.com)<br>Department of Psychological Sciences, Birkbeck College, University of London<br>WC1E 7HX, Malet Street, London, United Kingdom


#### Abstract

The dilution effect occurs when the introduction of nondiagnostic information lessens the impact on reasoning of diagnostic information despite having no relevance to the hypothesis in question. While the effect has been reproduced in several studies, the psychological basis of the effect remains unclear. Some believe it to be conversational while others believe it to be cognitive and social.

The paper tests the conversational basis of the effect by minimising pragmatic, conversational influence. To this end, it makes use of a legal setting with witness testimonies. The studies replicate the dilution effect, which suggests that the basis of the results in the original studies is not conversational. However, the credibility of the source strongly influences whether or not the effect occurs. If reliable sources provide the non-diagnostic information, the effect lessens. Conversely, if unreliable sources provide the non-diagnostic information, we observe a stronger dilution effect.


Keywords: Dilution effect; legal reasoning; source credibility; witness testimonies

## Introduction

Most information that humans gain throughout their lives comes from other sources. It may come from friends and colleagues, from professionals such as weather forecasters or news anchors, or it may come through de-personalised sources such as the Associated Press. However, information comes in various guises. Concerning the evidence itself, information may be highly diagnostic and related to a particular hypothesis at stake in the context or entirely unrelated and non-diagnostic. If, for example, an athlete is tested for doping before a race, the subsequent outcome of the test will be highly relevant in determining whether or not the athlete should be allowed to compete. The colour of the athlete's trousers worn during the drug test, however, should not. In addition, the information may be more or less noisy for a variety of reasons. This noise may be due to degradations in the access to information relating to the hypothesis (such as faulty equipment or poor visibility) or it
may be due to the reliability of the person who delivers the information.

The aim of the current paper is threefold. First, as discussed in the following section, it has been suggested that the dilution effect (see next section) is a conversational rather than a social or a cognitive effect. In the original studies, it is the experimenter himself who presented the participants with diagnostic and non-diagnostic information. If participants believe that the experimenter has chosen the non-diagnostic information for a reason, it may prompt them to try and interpret the information as somehow diagnostic. By removing the experimenter as the source of the diagnostic and non-diagnostic information, we test this possibility. We offer a possible control of this by placing the information in a legal setting and by having witnesses provide the testimonies. Second, as the role of the source of the information has been suggested as an influential element in reasoning, we manipulate the reliability of the source such that the source is either highly reliable or entirely unreliable. Third, from the literature, it is unclear how participants conceptualise non-diagnostic information. In particular, it is unclear whether or not the participants expect the dilution effect to occur if they were put in an observer role. To test this, study 2 allows participants to provide qualitative replies. Here, they are asked to imagine how a jury would react to the information and whether they believe it would make a difference to include the nondiagnostic information with the diagnostic.

## The dilution effect: A conversational explanation?

The dilution effect has been reported in several studies (e.g. Nisbett et al., 1981; Hilton \& Fein, 1981; Krueger \& Rothbart, 1988; Tetlock \& Boettger, 1989, see also Troutman \& Shanteau, 1977). However, aside from a few notable exceptions (e.g. Waller \& Zimbelman, 2003), the effect has received relatively little attention in recent years compared with more prominent cognitive influences on reasoning such as the confirmation bias (e.g. Frost et al.,
2015). In particular, the basis of the effect has remained under-explored.

One question, though, has been raised about the dilution effect, namely whether the effect has a conversational, pragmatic basis rather than a social perceptual basis (see Igou \& Bless, 2005; Kemmelmeier, 2007; Igou, 2007). It is well-known in the field of pragmatics that conversational expectations and extra-linguistic content can influence the interpretation of an utterance (see e.g. Sperber \& Wilson, 1995; Carston, 2002; Katsos, 2008; 2009). If the nondiagnostic information was somehow perceived as relevant given the inclusion by the experimenter, it is plausible that the participants could generate interpretations that make the information more relevant than the experimenter intended.

It is possible that the methodology of the experiments prompts participants to treat all information given to them as relevant, as the experimenter provides it to them. If the participants approach non-diagnostic evidence as potentially diagnostic in some way that they did not understand given the fact that the information was chosen by the experimenter, this may introduce noise into belief revision, which should make judgments less extreme. That is, given an increase in the noise of the data, a participant would be expected to update in a more tempered manner. Kemmelmeier (2007) describes this position (which he criticises) thusly: "The mere fact that the information is provided in the experiment suggests to participants that the experimenter considers this information relevant and wants participants to use it in making their judgments." (p. 49)

The above studies aimed at testing the conversational basis of the dilution effect by trying to manipulate the relevance of the information provided, but kept using the main methodology where the information is provided by the experimenter, and the task had generally to do with social judgment. One way of manipulating the relevance was by explicitly warning participants that the information might not be relevant. For example, Igou and Bless (2005) state, "prior to the sales scenario, half of the participants were informed that some of the presented information might not be relevant to their task". This is a methodological attempt to prepare the participant for the fact that they may encounter irrelevant information.

Kemmelmeier, who argues against the conversational account of the dilution effect, claims that the alleged evidence in favour of the conversational basis is not proving anything. Kemmelmeier concludes:
"Last, there is a very mundane reason to suspect that the dilution effect is not the product of conversational dynamics. The dilution effect occurs as much inside the psychological laboratory as outside of it (see Waller \& Zimbelman, 2003, for a review). Often there are no specific individuals who can be identified as the source of non-diagnostic information, or one even has to assume that one's communication partner is potentially deceptive, as in the case of an accounting audit (Waller \& Zimbelman, 2003). Because the dilution effect occurs regardless of whether non-diagnostic information can be assumed to be part of a meaningful communication, it seems highly questionable that the dilution effect has a conversational basis." (Kemmelmeier, 2007, p. 58)

In order to test the potential influence of the experimenter and to lessen the influence of social context, the current studies are set in a legal setting where the information is presented as a summarised court case concerning a murder in Paris. The existence of identified witnesses (with certain characteristics) attempts to alleviate the methodological problem of the experimenter providing 'irrelevant' information, as witnesses may provide more or less relevant statements during a trial. In order to manipulate the relevance of the statements, we manipulate the witness condition. As discussed in the following section, several studies have shown the influence of source credibility in reasoning tasks.

## The dilution effect and source credibility

As the dilution effect has mainly been explored with the information being provided by the experimenter, little is known about the relationship between the effect and the credibility of the source.

Source credibility has been shown to influence several cognitive phenomena related to reasoning, argumentation, and decision-making. It influences the reception of persuasive messages (Petty \& Cacioppo, 1984; Chaiken \& Maheswaran, 1994; Tormala \& Clarkson, 2007), is integral to the development of children's perception of the world (Harris \& Corriveau, 2011), influences candidate choice (Hetherington, 1999; Citirin \& Muste, 1999), increases adherence with persuasion strategies (Cialdini, 2007), and influences how people judge the quality of evidence from others in social situations (Fiske et al., 2007; Cuddy et al., 2011). The normative function of source credibility in reasoning and argumentation remains contentious. The dual-process-based Elaboration-Likelihood Model (Petty, 1981) describes reliance on the source of the message as a heuristic and shallow cue (Petty \& Cacioppo, 1984; Briñol \& Petty, 2009). Comparatively, Bayesian models integrate credibility in beliefs revision when a source provides information (Bovens \& Hartmann, 2003; Hahn et al., 2012; Harris et al., 2015; Madsen, 2016).

According to the dilution effect, participants who are faced with non-diagnostic information in addition to the diagnostic information provided will become less extreme in their degree of belief in the overall proposition. Given the findings in the literature, we predict that testimonies from reliable witnesses will be seen as more persuasive than testimonies from unreliable witnesses.

## The case study: Murder in Paris ${ }^{1}$

In order to make the experimental setting seem realistic, we made use of simplified version of a court case that happened recently in Paris. In the court case, the defendant, Siem, was accused of assaulting the victim, Tommy, which caused Tommy's fall to the ground. Further, they were told that

[^405]the impact of the ground caused the brain injury, which led to Tommy's death.

The participants were told that they would read an excerpt from a court case in Paris, France. They were further told that the names of the people involved had been changed and that the story had been abbreviated significantly. The participants were then instructed to read the summary of the court case thoroughly as if they were a member of the jury in the trial. Specifically, they were asked to pay attention to what had happened and whether or not the defendant was likely to be guilty or innocent.

## Study 1

Study 1 aims to replicate the dilution effect. To test the potential pragmatic basis of the dilution effect, the study was set in the legal context of a trial with all information provided by witnesses rather than by the experimenter. By using a realistic court case and witness testimonies rather than instructions from the experimenter, the design lessens the likelihood that the experimenter influenced the participants. The dilution effect predicts that participants should decrease their belief in the likelihood of the defendant being guilty when a non-diagnostic testimony was added to the diagnostic witness testimony.

The study was a $2 \times 2$ between-subjects design. To explore the influence of source credibility on belief revision and on the dilution effect, participants in the 'no witness' condition were told that the statements were 'information added to the initial enquiry'. As such, the information was provided with no specific source. In the 'witness' condition, a reliable witness presented the diagnostic testimony while an unreliable witness provided the non-diagnostic testimony ${ }^{2}$.

To test if the dilution effect was replicable, half of participants saw only the diagnostic information while the other half saw the diagnostic and the non-diagnostic information. Diagnostic statements read: "There was a dispute about drugs between Siem and Tommy, and Siem had threatened Tommy several times. Siem was heard several times saying 'he will be dealt with soon, this fucking Caribbean!" Non-diagnostic statements read: "When walking, Siem always took great and long strides. Siem used to wear funny clothes. In particular, he liked to wear bright colours".

Participants: 200 participants were recruited from MTurk (see Paolacci et al., 2010 for validation of MTurk as a tool for data collection in social sciences). All participants had to be native English speakers and aged 18 or above.

Procedure: Having read the background story (see appendix), participants provided their degree of belief in the likelihood that Siem had assaulted Tommy on a gradient scale from $0-1$, with 0 representing complete certainty that Siem did not assault Tommy and 1 representing complete certainty that Siem did assault Tommy. This elicited their

[^406]belief in the likelihood of guilt prior to hearing witness testimonies.

After providing their prior belief in the likelihood of guilt, participants read the testimonies. Having read the testimonies (witness statements or additional information // diagnostic or diagnostic as well as non-diagnostic information), participants were asked to indicate their posterior degree of belief in the likelihood of guilt on an identical sliding scale from 0-1. Diagnostic evidence was presented before non-diagnostic evidence. In order to test the dilution effect, we compare the changes in beliefs from prior to posterior belief between conditions.

Results: As the study was carried out online via MTurk, we eliminated any participants who carried out the study in less than 120 seconds, as the study could not be completed in seriousness in such short time. In total, this eliminated 15 participants, leaving 185 participants.

To test the dilution effect, paired-sample t-tests show a significant difference between prior and posterior degrees of beliefs in both diagnostic groups (No witness condition: $t=$ $3.105, \mathrm{p}=0.003$ (df, 42); Witness condition: $\mathrm{t}=2.890, \mathrm{p}=$ .006 , df (45)) while we observe no difference in degree of belief when non-diagnostic information is provided alongside the diagnostic (No witness condition: $\mathrm{t}=1.839$, p $=0.072$ (df, 48); Witness condition: $\mathrm{t}=.459, \mathrm{p}=.648$, df (46)), see Fig. 1 for means and standard deviations.

| Condition | Prior belief | Posterior belief |
| :---: | :---: | :---: |
| Diagnostic, no witness | $60.67(21.49)$ | $69.93(18.43)$ |
| Non-diagnostic, no witness | $64.36(17.88)$ | $69.18(17.94)$ |
| Diagnostic, witness | $58.97(18.87)$ | $66.32(19.38)$ |
| Non-diagnostic, witness | $62.68(22.24)$ | $61.29(22.27)$ |

Table 1: Prior and posterior beliefs
Tentatively, it looks as if the witness condition yields different patterns in the non-diagnostic condition (as the no witness condition is borderline significant). To test the influence of witnesses, we calculate a change score by subtracting the prior belief from the posterior. Having done this, we run a $2 \times 2$ ANOVA to test the influence of the inclusion of a witness. We find an effect of diagnosticity ( $p$ $=.019, \mathrm{~F}=5.556$ ), but no effect of the witness condition ( p $=.149, \mathrm{~F}=2.105)$.

Testing for influence of gender and age yielded no significant results, as p's were between .103 (influence of age on posterior degree of belief in the likelihood of guilt) and .881 (influence of gender on prior beliefs).

## Study 2

Study 1 suggests that the dilution effect was replicated in an experimental design aimed to lessen the experimenter's role and thereby reduced the potential for conversational effects. However, while the results of study 1 replicated the dilution effect, tentative evidence suggested that the reliability of the witness might have an impact on the relative strength of the effect. For one, the reliable witness always presented
condition the diagnostic information and the unreliable witness always presented the non-diagnostic information.

To test the potential influence of source reliability on the perception of evidence, half of the participants read the diagnostic testimony from the reliable witness and the nondiagnostic testimony from the unreliable witness whilst the other half were presented with the opposite source-message matrix.

While study 1 tested a specific question concerning the conversational basis for the dilution effect, study 2 is more exploratory, as the relationship between source credibility and the dilution effect has, to our knowledge, not been explored in detail (although, see Harkins \& Petty, 1987). As a consequence of the exploratory nature, participants were given the opportunity to provide qualitative feedback.

Importantly, study 2 used a different dependent variable: In study 1, as participants in the previous study were asked to provide their own degree of belief in the likelihood of guilt; in the present study, the participants were asked to provide their personal estimation of how convinced a member of a jury would be if confronted by the diagnostic information in isolation or by the inclusion of the nondiagnostic statement. As such, they were asked to provide an estimation of the strategic potential of including or omitting the non-diagnostic statement. Consequently, all participants read the diagnostic and the non-diagnostic statements. Thus, only two participant groups emerged in the present study: diagnostic (reliable witness) and nondiagnostic (unreliable witness) or diagnostic (unreliable witness) and non-diagnostic (reliable witness).

Participants: 100 participants were recruited from MTurk. All participants had to be native English speakers and be aged 18 or above.

Procedure: Prior belief elicitations were identical to study 1, as participants read the court case and provided their initial estimation of the likelihood of guilt. After the initial case presentation, participants read both the diagnostic and the non-diagnostic statements and were asked to evaluate the degree to which they believed a jury would believe the defendant to be guilty if the diagnostic information was presented in isolation or in conjunction with the nondiagnostic statement. As such, each participant provided one prior degree of belief and two posterior degrees of belief: diagnostic and non-diagnostic.

Results: Initial paired-sample t-tests were conducted between prior and posterior degrees of belief to test the influence of the source on the likelihood that a jury would find the defendant guilty. In accordance with expectations from studies on source credibility in argumentation (e.g. Harris et al., 2015), participants who were presented with diagnostic incriminating evidence from the unreliable source either significantly or borderline significantly decreased their posterior degree of belief in the likelihood of guilt (diagnostic: $\mathrm{t}=2.812$, df (50), $\mathrm{p}=.034$; nondiagnostic: $\mathrm{t}=1.799$, $\mathrm{df}(50), \mathrm{p}=.078$ ). Comparing the diagnostic and non-diagnostic posteriors, we observe no significant difference $(\mathrm{t}=.893$, $\mathrm{df}(50), \mathrm{p}=.376)$. This
suggests that testimonies from an unreliable source might decrease adherence with the proposition despite being diagnostic. It further suggests that no dilution effect was observed when the witness was reliable. See table 2 for means and standard deviations for both conditions.

In the condition where the reliable witness provides the diagnostic evidence, we observe a significant or borderline significant increase in the degree of belief in the likelihood of guilt (diagnostic: $\mathrm{t}=4.848$, df (49), $\mathrm{p}<.001$; nondiagnostic: $\mathrm{t}=1861$, df (49), $\mathrm{p}=.069$ ). While we did not find support for the dilution effect when the reliable witness presented the non-diagnostic information, we observe a significant difference in the condition where unreliable witness presents the non-diagnostic information $(\mathrm{t}=2.983$, df (49), $\mathrm{p}=.004$ ). That is, compared with the condition in which the reliable witness presented diagnostic evidence, the condition where the reliable diagnostic testimony was followed by an unreliable non-diagnostic testimony decreased the overall estimation of guilt.

Comparing the two conditions, this suggests the reliability of the source that provides the non-diagnostic information influences whether the dilution effect occurs or not. As we did not have a clear hypothesis as to the direction of the influence, qualitative replies were also collected. In the following, we examine these replies.

| Condition | Prior belief | Posterior <br> belief <br> (diagnostic) | Posterior <br> belief (non- <br> diagnostic) |
| :---: | :---: | :---: | :---: |
| Unreliable- <br> reliable | 64.19 <br> $(17.18)$ | $56.09(22.31)$ | $58.57(23.66)$ |
| Reliable- <br> unreliable | $57.4(22.25)$ | $71.60(18.33)$ | $63.82(23.82)$ |

Table 2: Prior and posterior beliefs
Qualitative replies By analysing the qualitative responses, we can get a tentative impression of the differences between reliability conditions and between participants themselves. In the unreliable-reliable condition, 25 participants provided qualitative feedback. In the reliable-unreliable condition, 30 participants provided feedback.

Participants in the condition where the unreliable witness presented the diagnostic information did not make specific comments about the persuasive advantage or disadvantage of presenting the non-diagnostic information (despite the fact that it was presented by the reliable witness). Rather, in line with expectations, they provided character-related comments for the unreliable witness and content-related comments for the reliable witness (e.g. "I'm not sure what Mrs. Lanavan's statement had to do with the case. And the fact that Ms. Harry is unstable would reduce her credibility" and "I think the first is incredible due to her personal history and the second's testimony really is irrelevant to the incident").

In the condition where the reliable witness presented the diagnostic information, comments were more mixed. 11 participants directly stated that including the non-diagnostic witness would not make a difference (e.g. "I don't see how

Ms. Harry changes anything. Her testimony doesn't really say anything useful" and "I think the jury would react the same way"). Comparatively, 13 participants stated that it would make a difference to include the unreliable witness (e.g. "The statement by Ms. Harry doesn't prove or disprove anything, but it takes away from the validity of the first witness, IMO", "Tough call- the statement of Ms. Harry would irritate the jury and would lean the jury to the more credible witness", and "I think maybe the prosecution loses some credibility if they put someone on the stand who gives testimony that doesn't seem substantive"). One participant argued that the inclusion would boost the probability of getting Siem convicted ("more witnesses the more weight the testimony will get I imagine, also the woman is more convincing").

Given the limited population size, the above comment should be taken with extreme caution. However, it suggests that participants may entertain two very different ideas of the reasoning of jurors. While the sample is too small for statistical analysis, the participants appear to entertain realistic approximations of their estimations of the reactions of jury members concerning the inclusion of the nondiagnostic information. The 11 participants who stated it would make no difference report no difference between the prior and the posterior.

## Concluding remarks

The paper set out to explore three different aspects of the dilution effect. First, given the debate concerning the basis for the dilution effect (whether it is conversational, cognitive, or social), study 1 used a legal setting to lessen or alleviate the potential influence of the experimenter and present the information as a court case with witness testimonies. Study 1 replicated the dilution effect.

Study 1 suggested that the credibility of the source might influence the strength of the dilution effect. Consequently, , study 2 manipulated the reliability of the witnesses who provided the diagnostic and the non-diagnostic information. Argumentation studies in source credibility suggest that the degree of belief in a proposition can be negatively influenced despite a diagnostic statement in cases where the source is unreliable or distrusted (see Madsen, 2016). In line with these findings, study 2 found that diagnostic statements from an unreliable source decreased participants' degree of belief in the likelihood of guilt while the same statement from a reliable source increased participants' degree of belief. Further, study 2 suggests that the dilution effect does not occur in situations where the non-diagnostic information is provided by a reliable source while we observe a strong dilution effect when an unreliable source presents nondiagnostic information. Future studies should look at the function of and relationship between source credibility and diagnosticity in more detail.

Study 2 gave participants the opportunity to provide qualitative feedback. Of particular interest, we noted a tendency for two strategy approximations to occur when the reliable source presented the diagnostic evidence and the
unreliable source presented the non-diagnostic statement. 11 (of 50 ) participants believed it would make no difference to the minds of a jury while 13 (of 50 ) believed it would have a negative effect. Their posterior belief revisions were in line with these estimations. This suggests that participants might entertain different perceptions of persuasion strategies and of the effect of evidence. However, as the current study is exploratory, we cannot draw any strong conclusions from these reports. We hope that future studies will explore the role of persuasion strategies and the dilution effect in more detail.

## References

Bovens, L., \& Hartmann, S. (2003). Bayesian epistemology. Oxford: Oxford University Press.
Briñol, P., \& Petty, R. E. (2009). Source factors in persuasion: A self-validation approach, European Review of Social Psychology, 20, 49-96.
Carston, R. (2002) Thoughts and Utterances: The Pragmatics of Explicit Communication, Blackwell Publishing
Chaiken, S. \& Maheswaran, D. (1994) Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement, Journal of Personality and Social Psychology 66 (3), 460-473
Cialdini, R. B. (2007) Influence: The Psychology of Persuasion, Collins Business
Citrin, J. \& Muste, C. (1999) Trust in government, in Robinson, J. P.. Shaver, P. R., Wrightsman, L. S. (Eds.), Measures of political attitudes. Measures of social psychological attitudes, Vol. 2. (pp. 465-532), San Diego, CA, US: Academic Press
Cuddy, A. J. C., Glick, P. \& Beninger, A. (2011) The dynamics of warmth and competence judgments, and their outcomes in organizations, Research in Organizational Behavior 31, 73-98
Fiske, Susan T., Cuddy, A. J. C. \& Click, P. (2007) Universal dimensions of social cognition: warmth and competence, Trends in Cognitive Sciences 11 (2), 77-83
Frost, P., Casey, B., Griffin, K., Raymundo, L.,Farrell, C. \& Carrigan, R. (2015) The influence of confirmation bias on memory and source monitoring, The Journal of General Psychology 142 (4), 238-252
Hahn, U., Oaksford, M., \& Harris, A. J. L. (2012). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), Bayesian Argumentation (pp. 1538). Dordrecht: Springer.

Harkins, S. G. \& Petty, R. E. (1987) information utility and the multiple source effect, Journal of Personality and Social Psychology 52 (2) 260-268
Harris, A. J. L., Hahn, U., Madsen, J. K., \& Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach, Cognitive Science 40, 1496-1533

Harris, P. L., \& Corriveau, K. H. (2011). Young children's selective trust in informants, Philosophical Transactions of the Royal Society B, 366, 1179-1187
Hetherington, M. J. (1999) The effect of political trust on the presedential election, 1968-96, American Political Science Review 93 (2), 311-326
Hilton, J. L., \& Fein, S. (1989). The role of typical diagnosticity in stereotype-based judgments, Journal of Personality and Social Psychology 57 (2), 201-211
Igou, E. R. (2007) Additional thoughts on conversational and motivational sources of the dilution effect, Journal of Language and Social Psychology 26 (1), 61-68
Igou, E. R., \& Bless, H. (2005). The conversational basis for the dilution effect. Journal of Language and Social Psychology 24, 25-35
Katsos, N. (2008) The semantics/pragmatics interface from an experimental perspective: the case of scalar implicatures, Synthese 165, 385-401
Katsos, N. (2009) Evaluating under-informative utterances with context-dependent and context-independent scales: experimental and theoretical implications, In Sauerland, U. \& Yatsushiro, K., Experimental Semantics and Pragmatics, Basingstoke: Palgrave Studies in Pragmatics, Language \& Cognition, 51-73
Kemmelmeier, M. (2007) Does the Dilution Effect have a Conversational Bias? Journal of Language and Social Psychology 26, 48-60
Krueger, J., \& Rothbart, M. (1988). Use of categorical information and individuating information in making inferences about personality, Journal of Personality and Social Psychology 55 (2), 187-195
Lagnado, D., Fenton, N. \& Neil, M. (2012) Legal Idioms: A framework for evidential reasoning, Argumentation \& Computation 1, 1-18
Madsen, J. K. (2016) Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions, Papafragou, A., Grodner, D., Mirman, D., \& Trueswell, J.C. (Eds.) Proceedings of the 38th Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Science Society, 165-170
Nisbett, R. E., Zukier, H. \& Lemley, R. E. (1981) The Dilution Effect: Nondiagnostic information weakens the implications of diagnostic information, Cognitive Psychology 13, 248-277
Paolacci, G., Chandler, J., \& Ipeirotis, P. G. (2010) Running experiments on Amazon Mechanical Turk, Judgement and Decision Making 5, 411-419
Petty, R. E. \& Cacioppo, J. T. (1984) The Effects of Involvement on Responses to Argument Quantity and Quality: Central and Peripheral Routes to Persuasion, Journal of Personality and Social Psychology 46 (1), pp. 69-81
Sperber, D. \& Wilson, D. (1995) Relevance: Communication and Cognition, 2nd edition, Blackwell Publishing

Tetlock, P. E. \& Boettger, R. (1989) Accountability: A Social Magnifier of the Dilution Effect, Journal of Personality and Social Psychology 57 (3), 388-398
Tormala, Z. L. \& Jackson, J. J. (2007) Assimilation and Contrast in Persuasion: The Effects of Source Credibility in Multiple Message Situations, Personality and Social Psychology Bulletin 33 (4), 559-571
Troutman, C. M. 6 Shanteau, J. (1977) Inferences based on nondiagnostic information, Organizational Behavior and Human Decision Performance 19, 43-55
Waller, W. S., \& Zimbelman, M. F. (2003). A cognitive footprint in archival data: Generalizing the dilution effect from laboratory to field settings, Organizational Behavior and Human Decision Performance 91, 254268

## Appendix: Background story and witness descriptions

## Background story

On the 31st of December 2010, around 7:30pm, the body of a man was found on the Place de Stalingrad, in Paris. The man was later identified as M. Tommy Tessel, a homeless drug-addict from Martinique. He died a few hours later, in the hospital.

A local police inquiry was conducted. All the people questioned in the neighbourhood initially denied having seen anything directly.
Some of them reported having heard that the victim had fallen after having been punched by a third party. The case was initially treated as an accident. No crime scene inspection was performed; no trace of blood was found.

On the 5th of January 2011 ( 5 days after the event), a person, who wanted to remain anonymous, reported to the police that a drug-addict often hanging around near the Rotonde (the rotunda of the Place de Stalingrad), of Senegalese descent, in his fifties, had punched the victim in the face and the victim had fallen heavily on the ground.

On the 8th of January, a crime investigation was opened. The criminal investigation department asked the local police for the victim's clothes so as to perform a DNA test. But it appeared that they were thrown away on the 5th of January for hygienic reasons.

An autopsy was performed on the deceased. The victim death's was directly caused by the brain injury resulting from the shock of his skull on the ground, probably due to a fall.

Reliable witness: Mrs. Rose Lanavan (55, social worker)
Mrs. Lanavan was employed as a cleaner in a pharmacy for around 20 years, after which she decided to enroll in a training programme for adults to become a social worker. She now works with drugaddicts and homeless people, helping them with any administrative procedures in relation to health, lodging, and judicial issues.

She is unanimously reported as a trustworthy and caring person. She works and lives in the area of Stalingrad, and knows well the people living there.

Unreliable witness: Ms. Edith Harry (26, no occupation)
Ms. Edith Harry is a drug-addict, often lurking in the area. She has tried a rehab several times, but always went back to smoking crack. She is reported to be psychologically unstable (she is reported to suffer from a serious personality disorder - labeled 'paranoid-delusional').

# Influence of using 3D images and 3D-printed objects on spatial reasoning of experts and novices 

Akihiro Maehigashi (mhigashi@cog.human.nagoya-u.ac.jp)<br>Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)<br>Masahiro Oda (moda@mori.m.is.nagoya-u.ac.jp)<br>Graduate School of Informatics, Nagoya University, Japan

Yoshihiko Nakamura (ynakamura@jo.tomakomai-ct.ac.jp)<br>Department of Computer Science and Engineering, National Institute of Technology, Tomakomai College, Japan

Kensaku Mori (kensaku@is.nagoya-u.ac.jp)<br>Information and Communications, Nagoya University, Japan

Tsuyoshi Igami (igami@med.nagoya-u.ac.jp)<br>Graduate School of Medicine, Nagoya University, Japan


#### Abstract

This study focuses on the infuence of a three-dimensional (3D) graphic image and a 3D-printed object on the spatial reasoning of experts and novices in the medical field. The spatial reasoning task of this study required doctors specializing in digestive surgery to infer cross sections of a liver with a 3D image and a 3D-printed object in a situation where liver resection surgery was simulated. The task performance was compared with that of university students who conducted the same task in Maehigashi et al. (2016). The results of the analysis indicated that the doctors showed the same task performance when using the 3D image and the 3D-printed object. However, the university students learned faster and inferred the inside of a liver structure more accurately with the 3D-printed object than with the 3D image, and they performed equally to the professional doctors. Our results are then discussed in relation to previous studies.


Keywords: Spatial reasoning; Spatial mental model; Expertise; External representation; 3D printer

## Introduction

## Spatial reasoning and 3D-printed object

Spatial reasoning refers to the inference of an object's shape and structure and the physical relationship between objects by using spatial information (Byrne \& Johnson-Laird, 1989). Spatial reasoning is ubiquitous in daily activities such as planning routes, inferring a road's slope angle, or even arranging furniture in a room.

When people engage in spatial reasoning, they form spatial mental models in their minds. Spatial mental models are internal representations of the spatial relations among elements, and it is considered that they allow people to do perspective taking, reorientation and spatial inferences (Tversky, 1993). Spatial mental models are strongly influenced by the types of external resources that are referred to for its formation. Tversky (1991) experimentally showed that route searches were more accurate when the route information was displayed on a map rather than text. Moreover, John, Cowen, Smallman, and Oonk (2001) indicated that the understanding of a geometric structure was more accurate with a three-dimensional (3D) graphic image rather than with a
two-dimensional (2D). They explained that using 3D images is more effective because they integrate the multiple perspectives expressed by 2D images into a single viewpoint, provide supplementary depth cues, and display object features that would be invisible in 2D images.

Recently, the prevalence of 3D printers has made it possible for people to replicate objects. 3D printers offer an unprecedented means to express information and are being used in various fields such as education, industrial manufacturing, and medicine. However, very few studies have investigated the influence of 3D-printed objects on spatial reasoning.

Some studies experimentally investigated human understanding of molecular structures using concrete models (Barrett, Stull, Hsu, \& Hegarty, 2015; Stull, Barrett, \& Hegarty, 2013). The results of these experiments demonstrated no difference in task accuracy between the use of 3D images and concrete models. However, in their experiments, task accuracy rate was very high. Therefore, further investigations that consider situations requiring people to understand more complex structures with physical object models are necessary.

Maehigashi et al. (2016) experimentally investigated spatial reasoning of human organ structure using 3D-printed organs. As a result, they found that the understanding of a human organ's structure was more rapid and accurate when examined with a 3D-printed object rather than with a 3D graphic image. Their study indicated the possibility that using 3D-printed objects might reduce both the cognitive load and the cost of information access in forming and manipulating spatial mental models. In addition, based on ethnographical research, Maehigashi et al. (2015) investigated the influence of using a 3D-printed human liver on doctors in real liver resection surgeries. Their results showed that using such objects enhanced the formation of elaborate spatial mental models of a patient's liver. It also enhanced the mental simulation of liver resections and the formation of shared spatial mental models of a patient's liver among doctors.

## Expertise

Experts in various fields use chunking strategy for encoding, storing, and manipulating spatial information. Chase and Simon (1973) showed that chess masters could encode and store multiple positions of chess pieces in a game situation in a single chunk. Also, Busey, Yu, Wyatte, and Vanderkolk (2013) compared the eye movements of experts with those of novices in fingerprint matching. Their results indicated that experts were able to match wider regions of two fingerprints in a single instance than novices, and they could encode and store the various characteristics of fingerprint structures as a single unit. Moreover, an experiment by Hegarty, Keehner, Khooshabeh, and Montello (2009) revealed that fourth year dentistry students inferred the anatomical structure of teeth more accurately than first year students and could encode, store, and manipulate the various characteristics of teeth structures as a single chunk.

Hegarty et al. (2009) proposed that the fourth year dentistry students developed spatial mental models of teeth based on their anatomical knowledge of teeth and their experiences of learning operative skills for dentistry, and such mental models would facilitate chunking strategy. Some studies have shown the similar concepts. For example, Woods (1999) stated that radiologists have developed organized mental matrixes which integrate radiological characteristics, and therefore, they are highly adept in visual management and able to synthesize the characteristics of diseases. Also, Gobet and Simon (1998) demonstrated how chess masters developed mental templates of chess positions based on their prior experiences, enabling them to encode large and multiple quantities of information.

Related to such discussions, some studies have investigated the relationship between an expert's spatial abilities and spatial reasoning. Spatial ability is the capability to mentally store and manipulate spatial representations accurately (Hegarty \& Waller, 2005). An experiment by Hegarty et al. (2009) demonstrated that the spatial abilities of experts influence their performance of spatial reasoning. Conversely, Ackerman (1988) investigated the relationship between expertise and various cognitive abilities and indicated that the development of domain specific knowledge actually decreases the influence of spatial ability on task performance.

The purpose of this study was to investigate the influences of the use of 3D images and 3D-printed objects on the spatial reasoning of both experts and novices. The relationship between the spatial abilities of experts and their performances in spatial reasoning was also examined.

## Experimental task

In this study, we used a spatial reasoning task in a situation where actual liver resection surgery was simulated. The participants inferred the positions of a tumor within a liver and also the veins on cross sections of a liver by referring to its internal structure as displayed in a 3D image and a 3D-printed object.

## Materials

The materials in this study were exactly the same as those used by Maehigashi et al. (2016). Two desks, a primary and a secondary desk, were used in the experiment. The primary desk was set in front of a participant, and the secondary desk was set by the right side of the participant. The primary and secondary desks represented an operating table and tool stand as used in a surgical setting. Three boxes were placed on the primary desk. Each box contained a 3D-printed object of a liver (target) which represented a patient's liver and an answer sheet. Placed on the secondary desk was either a computer displaying a liver's 3D image or a box containing a 3Dprinted object that displayed the inside structure of a liver. Figure 1 shows a 3D image, a 3D-printed object, and a target.


Figure 1: (a) 3D image, (b) 3D-printed object, and (c) target
The 3D image was created by using Pluto, a computeraided diagnosis system developed at Nagoya University's Graduate School of Information Science. It was created based on data from a patient's liver measured by computed tomography (CT) (Figure 1a). In the 3D image, the thickest vein, an inferior vena cava (IVC), and five veins branching from the IVC were represented in blue. A tumor was represented in white. The participants could rotate and zoom in and out of the image by using a mouse.

The 3D-printed object and the three targets were created by using a 3D printer with the same CT liver data as the 3D image (Figure 1b, 1c). The 3D-printed object shows a liver's inside structure. In contrast, the target's surface was colored light gray, and the liver's inside structure was invisible just as a patient's liver would be in real-life surgery. A line was drawn around each of the three targets. Each line was sketched in a different location. Also, on each target the letters " $A$ " and " $B$ " were written to indicate the two separated areas based on the drawn line. Two sets of 3D images, 3Dprinted objects, and three targets were created from different CT liver datum.

## Vein and tumor location tests

The tests used in this study were exactly the same as those used by Maehigashi et al. (2016). The experiment employed a spatial reasoning task. Participants were required to take a vein and tumor location test for each target while referring to either the 3D image or the 3D-printed object. In the vein location test, participants were required to indicate the location of the veins that appeared on the cross section by cutting the target along the drawn line. Specifically, participants were required to mark "O" for the IVC and "X" for the branching
vein on the cross section's outer contour printed on the answer sheet (Figure 2). In the tumor location test, participants were asked to identify the area of the liver, either A or B, where the tumor had occurred.


Figure 2: Vein location test. (a) Contour of cross section of liver, (b) cross section of liver, and (c) participant's answer. (a) shows the outer contour of a liver's cross section printed on the answer sheet. (b) shows an actual cross section of a liver. (c) shows a participant's answer, which provides the number of IVCs, Os, and the branching veins, Xs, (drawn correctly here).

## Experiment

## Method

Participants and Factorial design Twenty-two doctors specializing in digestive surgery participated in this experiment. Their work experience ranged from eight to 22 years ( $M=10.57$ ). The experiment had a single-factor within participants design. The factor was the external representation (image and object).

Procedure The experimental procedure was generally the same as those of Maehigashi et al. (2016). First, the participants took a spatial ability test produced by Guay and McDaniels (1976). It comprised 24 questions that required mental rotations. The participants were required to answer as many questions as possible within three minutes. Next, all of the participants performed practice and experimental tasks with the 3D image in the image condition and the 3D-printed object in the object condition. In the practice task, the 3D image or the 3D-printed object, which represented one IVC and three branching veins, were used. First, the participants observed and learned about the inside structure using either the 3D image or the 3D-printed object for one to three minutes. Following on, they took the vein and tumor location tests for one target, referring to the image or the object.

After the practice, all participants conducted the experimental task. During the learning period, participants observed the inner liver structure for three to five minutes using either the 3D image or the 3D-printed object. When the participants deemed themselves ready after three minutes had passed, or when five minutes passed, the tests began. Participants took out the target and answer sheet from one of the three boxes on the primary desk and attempted to complete the vein and tumor location tests. During the task, participants were allowed to refer to either the 3D image or the 3D-printed object freely.

After the participants completed the tests for one target, they returned it together with the answer sheet back into the box and took a different set from another box. The task was completed when they had finished the tests for all three targets. After the experimental task was completed in one condition, the participants took a five-minute break and performed the practice and experimental tasks in the other condition.

For the image and object conditions, the 3D image and the 3D-printed object created by the different CT liver datum were used. The order of the task conditions and the combinations of CT liver datum were counterbalanced between the participants. Three sets of targets and answer sheets were randomly placed inside the boxes on the primary desk. Participants were instructed to perform the tasks as accurately as possible. Furthermore, removing the target from the primary desk was forbidden during the experiment, as it would be impossible for doctors to remove a patients liver from the operating table in a real-life surgical operation. However, removing the 3D-printed object from the secondary desk was permitted because in a surgical operation, doctors can place a 3D-printed liver right beside a patient's liver to confirm its interior structure (Maehigashi et al., 2015).

## Results

The participants of this study were doctors with anatomical and medical knowledge as well as first hand medical experience. Therefore, the data of this study was treated as an expert's performance data. We compared our data to that of Maehigashi et al. (2016) which examined the exact same tasks under the same conditions on 48 university students who did not possess any anatomical and medical knowledge.

We conducted 2(Expertise: expert and novice) $\times 2$ (External representation: image and object) analysis of variance (ANOVA) on the dependent variables. Since the external representation factor (image and object) was a betweenparticipants factor in the study by Maehigashi et al. (2016), we conducted a two-way between participants ANOVA in our analyses.

First, the learning time was the mean time taken by the participants to observe the inner structure of either the 3D image or the 3D-printed object before attempting the tests in each condition (Figure 3). The results of the analysis showed a significant interaction $(F(1,88)=4.75, p<.05)$. The analysis of the simple main effect showed that in the image condition, the learning time was significantly shorter for the expert condition than for the novice condition $(F(1,88)=14.84, p<$ .001 ). However, in the object condition, there was no significant simple main effect on the expertise factor $(F(1,88)=$ $0.60, p=.44$ ). Also, in the novice condition, the learning time was significantly shorter for the object condition than for the image condition $(F(1,88)=15.70, p<.001)$. However, in the expert condition, there was no significant simple main effect on the external representation factor $(F(1,88)=$ $0.78, p=.38$ ). In addition, there were significant main effects on both the expertise factor $(F(1,88)=10.67, p<.01)$ and the external resource factor $(F(1,88)=11.74, p<.001)$.


Figure 3: Learning time. The error bars indicate the standard error.

Following on, the task completion time was calculated as the mean time from when the first target was pulled out until the third target was returned to the box in each condition (Figure 4). The results of the analysis showed no significant interaction $(F(1,88)=0.10, p=.76)$. There was, however, a significant main effect on the expertise factor as the task completion time was shorter for the novice condition than for the expert condition $(F(1,88)=21.84, p<.001)$. Also, there was no significant main effect on the external resource factor $(F(1,88)=2.08, p=.15)$.


Figure 4: Task completion time. The error bars indicate the standard error.

In the vein location test score, we calculated the mean absolute difference value between the correct number of veins in the stimuli and the number of veins drawn on the answer sheet for the IVC and the branching veins in each condition (Figure 5). If the score is closer to zero, the number of drawn veins is more accurate.

For the IVC, all participants in the expertise condition correctly drew the veins, making the mean absolute difference value zero. On the other hand, for the branching veins, there
was a significant interaction $(F(1,88)=5.23, p<.05)$. The results of the simple main effect analysis showed that in the image condition, there was a significant simple main effect on the expertise factor; in other words, participants in the expert condition drew the number of veins more accurately than those in the novice condition $(F(1,88)=25.45, p<$ .001 ). However, in the object condition, there was no significant simple main effect of the expertise factor $(F(1,88)=$ $3.28, p=.07$ ). Also, in the novice condition, there was a significant simple main effect on the external representation factor, highlighting that more veins were accurately drawn in the object condition than in the image condition $(F(1,88)=9.80, p<.01)$. However, in the expert condition, there was no significant simple main effect on the external representation factor $(F(1,88)=0.01, p=.92)$. In addition, there were significant main effects in both the expertise factor $(F(1,88)=23.50, p<.001)$ and the external resource factor $(F(1,88)=4.58 . p<.05)$.


Figure 5: Absolute difference value. The error bars indicate the standard error.

In each tumor location test, a score of one was assigned if the tumor location was correctly answered. The tumor location test score was the mean total of the test results for the three targets in each condition (Figure 6). In other words, the higher the score, the more accurate is the answer. The results showed a significant interaction $(F(1,88)=$ $4.64, p<.05$ ). The analysis of the simple main effect indicated that in the image condition, the score was significantly higher for the expert condition than for the novice condition $(F(1,88)=5.96, p<.05)$. However, in the object condition, there was no significant simple main effect on the expertise factor $(F(1,88)=0.37, p=.55)$. Also, in the novice condition, the score was significantly higher for the object condition than for the image condition $(F(1,88)=$ $24.27, p<.001$ ). However, in the expert condition, there was no significant simple main effect on the external representation factor $(F(1,88)=3.54, p=.06)$. Additionally, there was a significant main effect on the external resource factor $(F(1,88)=4.64, p<.05)$, but no significant main effect on
the expertise factor $(F(1,88)=1.69, p=.20)$.


Figure 6: Tumor location test score. The error bars indicate the standard error.

Furthermore, correlation analyses were conducted on the relationship between the spatial ability test scores of the experts and their task performance, learning time, task completion time, absolute difference value for branching veins, and tumor location test score, in the image and object conditions. However, there was no significant correlation.

Finally, a t-test was conducted on the test scores for spatial ability in both the expert and novice conditions. The score was higher in the expert condition $(M=10.86)$ than in the novice condition $(M=8.48)(t(68)=2.20, p<.05)$. These results did not confirm the homogeneity of the spatial abilities between the experts and the novices. However, the results of the correlation analyses indicated that the experts did not use the advantage of their spatial ability to conduct the task.

## Discussion

## Accuracy of spatial reasoning

The results of the vein and tumor location tests revealed that the university students inferred a liver's inner structure more accurately with the 3D-printed object than with the 3D image and performed it to a standard equal to that of the professional doctors.

The university students formed spatial mental models of livers probably for the first time. Since the real world offers more depth cues than the virtual 3D environment (Kemeny \& Panerai, 2003), using the 3D image might require more cognitive load in order to form spatial mental models than using the 3D-printed object. The university students who used the 3D image apparently needed to mentally complement or modify spatial information, temporarily storing such information in their memory and mentally resizing it in order to map the information to the target. However, the students with the 3D-printed object were assumedly able to store the spatial information temporarily in their memory as they perceived it
and mapped the information from the 3D-printed object directly to the target without having to internally modify or resize it. Therefore, the university students with the 3D-printed object were assumed to have a smaller cognitive load, and, consequently, make fewer errors from the internal manipulation of spatial information and, therefore, able to show test performances equal to that of doctors.

It is also possible that the university students with the 3Dprinted object experienced lower information accessing costs than those who used the 3D image. Information accessing cost is incurred when acquiring information (Gray, Sims, Fu, \& Schoelles, 2006). Participants with the 3D image had to manipulate a computer mouse in order to acquire the required information. However, participants with the 3D-printed object had only to pick up and physically rotate a 3D-printed object. Thus, accessing information with a 3D-printed object was considered easier and less prone to errors or omissions than working with a 3D image.

However, the doctors showed the same task performance when using the 3D image and the 3D-printed object. It is inferred that the experts obtain spatial mental models developed by their prior knowledge and experiences (Hegarty et al., 2009). The doctors who participated in this study might be in possession of developed rigid spatial mental models of livers, and they were therefore able to modify their mental models based on the information displayed on both the 3D image and the 3D-printed object respectively. As a result, even though the 3D image is not as in-depth as the 3D-printed object, the doctors could still manage to create an equally accurate spatial mental model for the tests by depending on their already developed spatial mental models.

## Learning and task completion time

Analysis of the learning time revealed that the doctors showed the same task performance when using the 3D image and the 3D-printed object. However, the university students with the 3D-printed object finished their period of learning quicker than those with the 3D image and performed equally to the doctors.

As explained above, by using the 3D-printed objects, the university students were assumed to be able to reduce their cognitive load and information accessing cost. Therefore, they might be able to facilitate the formation of spatial mental models and perform equally to the doctors. Also as written above, the doctors were considered to have developed mental models. By modifying these mental models accordingly, they might be able to form spatial mental models with the 3D image as quickly as when they used the 3D-printed objects.

The results of the task completion time indicated that the university students performed the task faster than the doctors either with the 3D image or with the 3D-printed object. Some previous studies also experimentally showed that experts took a longer time to complete tasks than novices (Busey et al., 2013; Krupinski, 1996). One possibility is that since experts could access the related information by recalling and utilizing their existing knowledge, this process might cause
a longer task completion time. Previous studies showed that chess masters focused their eyes on the empty spaces more than novices when pieces on the board were being memorized (Charness, Reingold, Pomplun, \& Stampe, 2001; Reingold, Charness, Pomplun, \& Stampe, 2001). The chess masters were thought to be processing the related information stored in their long term memory. Another possibility is that experts could be more careful than novices. Previous studies showed that fingerprint experts were more skeptical than novices, and it therefore took them longer to match fingerprints (Busey et al., 2013).

## Experts' spatial ability and spatial reasoning performance

When the university students used the 3D image, there was a significant relationship between their spatial abilities and spatial reasoning performance (Maehigashi et al., 2016). In particular, high ability students demonstrated longer learning times and a more accurate inference to the positions of branching veins. On the other hand, in this study, whenever the doctors used the 3D image or the 3D-printed object, there was no relationship between their spatial abilities and spatial reasoning performance.

These results are different to that of Hegarty et al. (2009). The main difference between the previous study and this study can be related to the participants' degrees of expertise. The experts in Hegarty et al. (2009) were fourth year dentistry students. In contrast, the experts in this study were doctors with many years of work experience, and they therefore had many more years of expertise in the specialized field than the experts in Hegarty et al. (2009). Therefore, in this study, the spatial ability of experts did not influence spatial reasoning performance as indicated in Ackerman (1988).

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP15H01614.

## References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. Journal of Experimental Psychology: General, 117, 288-318.
Barrett, T. J., Stull, A. T., Hsu, T. M., \& Hegarty, M. (2015). Constrained interactivity for relating multiple representations in science: When virtual is better than real. Computers and Education, 81, 69-81.
Busey, T., Yu, C., Wyatte, D., \& Vanderkolk, J. (2013). Temporal sequences quantify the contributions of individual fixations in complex perceptual matching tasks. Cognitive Science, 37, 731-756.
Byrne, R. M. J., \& Johnson-Laird, P. N. (1989). Spatial reasoning. Journal of Memory and Language, 28, 564-575.
Charness, N., Reingold, E. M., Pomplun, M., \& Stampe, D. M. (2001). The perceptual aspect of skilled perfor-
mance in chess: Evidence from eye movements. Memory and Cognition, 29, 1146-1152.
Chase, W., \& Simon, H. A. (1973). Perception in chess. Cognitive Science, 4, 55-81.
Gobet, F., \& Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. Memory, 6, 225-255.
Gray, W. D., Sims, C. R., Fu, W.-T., \& Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. Psychological Review, 113, 461-482.
Guay, R., \& McDaniels, E. (1976). The visualization of viewpoints. West Lafayette: The Purdue Research Foundation.
Hegarty, M., Keehner, M., Khooshabeh, P., \& Montello, D. R. (2009). How spatial abilities enhance, and are enhanced by, dental education. Learning and Individual Differences, 19, 61-70.
Hegarty, M., \& Waller, D. (2005). Individual differences in spatial abilities. In P. Shah \& A. Miyake (Eds.), The cambridge handbook of visuospatial thinking (p. 121-169). New York: Cambridge University Press.
John, M. S., Cowen, M. B., Smallman, H. S., \& Oonk, K. M. (2001). The use of 2d and 3d displays for shapeunderstanding versus relative-position tasks. Human Factors, 43, 79-98.
Kemeny, A., \& Panerai, F. (2003). Evaluating perception in driving simulation experiments. Trends in Cognitive Sciences, 7, 31-37.
Krupinski, E. A. (1996). Visual scanning patterns of radiologists searching mammograms. Academic Radiology, 3, 137-144.
Maehigashi, A., Miwa, K., Oda, K., Nakamura, Y., Mori, K., \& Igami, T. (2016). Influence of 3d images and 3d-printed objects on spatial reasoning. Proceedings of the 38th Annual Meeting of the Cognitive Science Society, 414-419.
Maehigashi, A., Miwa, K., Terai, H., Igami, T., Nakamura, Y., \& Mori, K. (2015). Investigation on using 3d printed liver during surgery. Proceedings of the 37th Annual Meeting of the Cognitive Science Society, 1476-1481.
Reingold, E. M., Charness, N., Pomplun, M., \& Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. Psychological Science, 12, 48-55.
Stull, A. T., Barrett, T., \& Hegarty, M. (2013). Usability of concrete and virtual models in chemistry instruction. Computers in Human Behavior, 29, 2546-2556.
Tversky, B. (1991). Spatial mental models. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (p. 109-145). New York: Academic Press.
Tversky, B. (1993). Cognitive maps, cognitive collages, and spatial mental models. In A. U. Frank \& I. Campari (Eds.), Spatial information theory: A theoretical basis for gis (p. 14-24). Berlin: Springer-Verlag.
Woods, B. P. (1999). Visual expertise. Radiology, 211, 1-3.

# Even when people are manipulating algebraic equations, they still associate numerical magnitude with space 

Tyler Marghetis<br>tmarghet@indiana.edu

Robert L. Goldstone<br>rgoldsto@indiana.edu

David Landy<br>dlandy@indiana.edu

Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405


#### Abstract

The development of symbolic algebra transformed civilization. Since algebra is a recent cultural invention, however, algebraic reasoning must build on a foundation of more basic capacities. Past work suggests that spatial representations of number may be part of that foundation, but recent studies have failed to find relations between spatial-numerical associations and higher mathematical skills. One possible explanation of this failure is that spatial representations of number are not activated during complex mathematics. We tested this possibility by collecting dense behavioral recordings while participants manipulated equations. When interacting with an equation's greatest [/least] number, participants' movements were deflected upward [/downward] and rightward [/leftward]. This occurred even when the task was purely algebraic and could thus be solved without attending to magnitude (although the deflection was reduced). This is the first evidence that spatial representations of number are activated during algebra. Algebraic reasoning may require coordinating a variety of spatial processes.


Keywords: algebra; number and space; notations; mousetracking.

## Introduction

The invention of symbolic algebra transformed human civilization. Algebraic notation allows for accomplishments as mundane as buying paint for a new fence and as fantastic as discovering antimatter. But symbolic algebra is a recent cultural invention. Thus, it cannot rely on devoted neural machinery that evolved specifically for that purpose - an innate 'algebra module.' Instead, our capacity for symbolic algebra must be cobbled together from other cognitive capacities. But which?
One proposal is that higher mathematics, including symbolic algebra, builds on a foundation of space (e.g., Lakoff \& Núñez, 2000; Sella et al, 2016). On this proposal, our evolutionarily ancient spatial abilities have been coopted by culture to reason about abstract mathematical entities and relations. Indeed, early spatial abilities are known to predict life-long mathematical performance, from grades in elementary school to the choice of a mathematicsheavy college major.

One crucial aspect of this spatial foundation may be the ability to use space to make sense of number (Hubbard et al, 2005). Spatial representations of number could ground the highly abstract notion of numerical magnitude in the more basic, experiential notion of location. More complex forms of mathematics could then build on this foundation, from algebra to calculus and beyond (Núñez \& Marghetis, 2015). The current study tests this account by examining whether spatial representations are activated during one canonical case of complex mathematical activity: solving equations.

## Mixed evidence for spatial-numerical associations in higher mathematics

There is considerable evidence that spatial representations of number are ubiquitous and automatic, at least during simple numerical tasks (Hubbard et al, 2005; Winter, Marghetis, \& Matlock, 2015). These spatial representations involve both the horizontal and vertical axes. Among literate adults in Western cultures, for instance, processing lesser numbers facilitates subsequent responses on the left, while processing greater numbers facilitates responses on the right (Dehaene et al, 1993). Similarly, when German adults generate random numbers while undergoing upward and downward motion, they produce numbers that are significantly greater when moving upward and lesser when moving downward (Hartmann et al, 2011). In adults, these spatial representations have been shaped considerably by culture. The association between numerical magnitudes and horizontal locations, for instance, is reversed among Palestinians who read both words and numbers from right-to-left (Shaki et al, 2009). When we encounter a number, therefore, we automatically activate spatial representations of its magnitude.
But do these implicit spatial representations play any role in mathematics beyond the domain of simple numbers? The evidence is rather mixed. One point in favor of such a role is that the correlation between early spatial abilities and later school success in mathematics is mediated by the ability to map numbers to a linear path (Gunderson et al, 2012). Causal evidence comes from the finding that training students to map numbers to linear path improves calculation (Siegler \& Ramani, 2009). There is also evidence that spatial representations of number play a role in mathematical communication. Mathematical experts produce gestures that express numbers as locations, even
when their speech does not contain any mention of space (Marghetis \& Núñez, 2013).
On the other hand, there have been a number of failures to find any relation between spatial-numerical associations and higher mathematical ability (e.g., Cipora \& Nuerk, 2013; Cipora, Patro, \& Nuerk, 2015). Indeed, there is little evidence that spatial representations of number are activated at all during mathematical activities that are more complex than simple numerical judgments - judging a number's relative magnitude, or determining whether it is even or odd. In contrast to these simple tasks, real mathematical activity seldom involves single numerals in isolation. Algebra, in particular, is often the first time that students begin to think about numbers, not just in terms of their magnitudes, but also in terms of their structural interrelations. As these more complex notions come to the fore, spatial representations of numerical magnitude may fade into the background.
Moreover, many of these more complex notions may also have a spatial underpinning. During mental calculation, for instance, subtraction and addition are associated with leftward and rightward motion, respectively (Knops et al, 2009; Marghetis, Núñez, and Bergen, 2014); and during algebraic reasoning, space is used to represent the hierarchical syntax of algebraic expressions (Landy \& Goldstone, 2007). If space is playing these other roles, then the association between space and number might reasonably be expected to fade. The neural circuitry responsible for representing spatial location cannot be all things at once. On this account, as space is co-opted for new roles arithmetic, algebraic syntax - its association with numerical magnitude might diminish. Number-space associations may be limited to simple numerical judgments, disappearing as mathematical complexity increases.

## The present study

The current literature, therefore, appears to support conflicting accounts. On one hand, associations between number and space are activated automatically during a variety of simple tasks, and spatial processing more generally has been implicated in higher-level mathematical thinking. On the other hand, individual differences in spatial representations of number do not appear to correlate with mathematical expertise. This presents a puzzle. What is happening to these spatial representations of number as people transition from simple judgments of isolated numerals to more complex mathematical activities?

To resolve this puzzle, we analyzed the spatial dynamics of individuals' manipulations of algebraic equations, using a methodology that we have dubbed Dense Recording of Algebraic Manipulations (DREAM). In this approach, participants manipulate equations using click-and-drag dynamic algebra software; we record the moment-tomoment details of these manipulations, including the precise mouse trajectories used to rearranging equations.

In the current study, algebraic equations were displayed on a computer screen (e.g., $x+3=7$ ), and participants could rearrange these equations by clicking and dragging symbols as if they were physical objects. We also manipulated whether participants performed a task that was focused on magnitude ("Click and drag the greatest/least number") or algebraic structure ("Solve for x"). Throughout, we recorded the fine-grained details of these interactions, including the precise spatial locations at which individual numbers were clicked ${ }^{1}$. By varying the numbers in the equations, we could see whether numerical magnitude had a systematic effect on the location of symbolic manipulations.
We foresaw several possible outcomes. If spatial representations of number play no role in higher mathematics - or play a role only in development - then spatial representations might not be activated at all while equations are manipulated, particularly if the goal is to solve the equation. If, on the other hand, spatial representations of number continue to play a functional role in algebra perhaps by grounding the meaning of otherwise arcane symbolic manipulations - then we might find traces of spatial-numerical associations in the fine details of how equations are manipulated. In particular, greater numbers might be clicked higher or more rightward, while lesser numbers might be clicked lower or more leftward.

## Methods

## Design

Participants manipulated equations with a computer mouse, moving terms as if they were virtual objects. This was implemented with the Graspable Math software (www.graspablemath.com). Think of files and folders on a computer desktop, which can be reorganized and rearranged by clicking and dragging. Graspable Math offers the same functionality but for equations. In response to these manipulations, the software automatically adjusts the equation to maintain its validity. For instance, given the equation ' $x+2=4$,' as the 2 is dragged to the far side of the 4 , the + symbol changes automatically into the - symbol as it crosses the equal sign, so that the final state of the equation would be ' $x=4-2$ ' (Fig. 1a). This allows users to focus on how and why they want to rearrange equations. In addition, clicking on the equals sign flips an equation (e.g., $\mathrm{x}=2 \rightarrow 2=\mathrm{x}$ ), and clicking on an operation performs that operation (e.g., $x=4-2 \rightarrow x=2$; Fig. 1b).

The full system is quite powerful and can be explored online (www.graspablemath.com). The current study used a simplified version that included only the dragging and clicking interactions described above. We recorded where
${ }^{1}$ Technically, participants clicked numerals that denoted numbers, and it was the numerals' denotations that had magnitude. For simplicity of presentation, however, we shall conflate numerals with their denotations and refer to them as numbers.
and when interactions occurred, including $x, y$ coordinates of the mouse cursor. Here we focus on where, exactly, participants clicked on numbers, to investigate whether this spatial behavior was affected by the numbers' magnitude.


Figure 1. Manipulating equations using Graspable Math. (a) As an equation is rearranged, it's updated automatically to remain valid. Here, ' +2 ' is dragged from left to right; the sign is switched as it crosses the equals sign. (b) Operations are triggered by clicking the operator.

On each trial, an algebraic equation appeared on the screen (e.g., ' $x+3=5$ '). Participants performed one of two tasks, assigned between-subjects.
In the Algebra task, participants had to solve for the variable by clicking and dragging to simplify the equation (Fig. 1). For instance, given $x+3=5$, one might start by dragging the 3 to the other side of the equation. Note that this does not require attention to numerical magnitude, only to the algebraic relations between the terms.

In the Magnitude task, participants were presented with the exact same equations, but their task was to find the least number - or the greatest number, depending on the block - and indicate their selection by dragging it to other side of the equation. This click-and-drag response was chosen so that the two tasks involved comparable interactions with identical stimuli.

## Participants

Volunteers participated in exchange for partial course credit ( $N=69, M_{\text {age }}=19$ years, 51 women, 18 men). A target sample size of 68 was determined in advance on the basis of similar studies of number and space (e.g., $n=44$ in Fischer et al, 2010).

## Materials

For both tasks, items consisted of equations in the form $x$ $\pm b=c(N=112)$. Values of $b$ and $c$ ranged from 1 to 9,
excluding 5. The value of $b$ was always different from $c$, so one number was always greater than the other, producing 56 combinations of values for $b$ and $c$. Each combination was used to create two equations: one with addition (e.g., $x+2=$ 3 ) and one with subtraction (e.g., $x-2=3$ ).

## Procedure

Participants gave informed consent, completed a brief tutorial on how to manipulate equations with the mouse, and read task instructions. This was followed by practice trials chosen randomly from the full list of items $(n=4)$. They then completed the experimental trials $(n=224)$. Each item appeared twice, ordered randomly across four blocks. For the Magnitude task, initial target magnitude (greater, lesser) was assigned randomly and switched halfway through.
Each trial began with the appearance of a fixation symbol at the top-center of the screen. Clicking on this fixation symbol triggered the appearance of an equation toward the bottom of the monitor. The equation appeared either on the left or right of the screen and with the variable either on the left or right of the equal sign (i.e., ' $x+2=3$ ' or ' $3=x+2$ '), assigned randomly. Participants were then free to manipulate the equation using the computer mouse. Trials in the Magnitude task ended automatically when a number was dragged across the equal sign and released. In the Algebra task, trials ended automatically when participants had solved for $x$. Participants finished by answering a series of standard demographic questions along with four questions about mathematical experience: Did they study calculus in high school? In college? What was their grade? And what was their SAT score? No other measures were collected.

## Analysis

We focused on where numbers were clicked, specifically the first number manipulated during each trial. Our primary measure was the deflection of these locations, relative to where the participant would click typically (i.e., standardized by participant). A value of zero thus indicated no deflection; negative values, deflections downward or leftward; and positive values, deflections upward or rightward. Analyses used linear mixed-effects models, with centered predictors and the maximal converging effects structure justified by the design (Barr et al, 2013).

## Results

One participant was removed for poor accuracy (72\%). Accuracy was high among remaining participants ( $M=$ $96 \%, 95 \%$ CI $[86 \%, 100 \%]$ ). One additional participant was removed for corrupted data. Before analysis, we removed trials where the participant did not arrive at the correct response ( $4 \%$ of trials), followed by those that were three standard deviations faster or slower than each participant's mean (1.4\% of trials).

## Overall spatial deflection due to numerical magnitude

We first investigated whether numerical magnitude caused systematic spatial deflections in click locations. For each trial, we calculated a measure of overall spatial deflection by summing the deflection along the vertical and horizontal axis (i.e., a signed Manhattan distance). On this measure, positive values indicate deflections that are, overall, congruent with our predictions for greater numerical magnitudes (i.e., rightward and upward), and negative values indicate deflections congruent with predictions for lesser magnitudes (i.e., leftward and downward). This spatial deflection was analyzed with a model that included fixed effects of Relative Magnitude (i..e, whether the selected number was greater or less than the equation's other number), Task (Algebra vs. Magnitude), and their interaction; and random intercepts and slopes for both participants and items.
There was no effect of Task (p > .9). As predicted, interactions with numerals were deflected spatially by their magnitude, $b=.25 \pm .03 \mathrm{SEM}, t=8.8, \mathrm{p}<.0001$. These deflections were congruent with canonical spatial representations of numerical magnitude. When participants manipulated the lesser number in an equation, they clicked a location that was deflected in the congruent left-downward direction ( $M=-0.09$ ); when they manipulated the greater number, they clicked more right-upward $(M=0.11)$.

The size of this spatial deflection, moreover, was moderated by the task, $b=-0.14 \pm 0.06$ SEM, $t=-2.4 p=$ .02. The size of the magnitude-based spatial deflection in the Magnitude task ( $b=0.30 \pm 0.03$ SEM) was significantly larger than in the Algebra task $(b=.18 \pm .04$ SEM $)$, even though the magnitude-based deflection was significant in both tasks (both $p \mathrm{~s}<.0001$ ). Thus, magnitude induced an overall spatial deflection of numeral manipulations, and the size of this deflection was task-dependent.

## Axis-specific spatial deflections

We next investigated whether this task-sensitive spatial deflection was specific to either the vertical or horizontal axis. Along the vertical axis, there was no evidence that responses differed by Task, $b=0.001 \pm .02 \mathrm{SEM}, p>9$. By contrast, a number's relative magnitude had a systematic impact on where it was clicked, $b=.18 \pm .02 \mathrm{SEM}, t=7.7$, $p<.0001$. When the selected number was greater than the other number in the equation, it was clicked 0.18 standard deviations higher than when it was less than the other number. This spatial-numerical deflection was moderated by the task, as revealed by a significant interaction, $b=-0.14 \pm$ $.05 \mathrm{SEM}, t=-2.3, p=.02$. Additional analyses confirmed that a spatial-numerical deflection occurred for both tasks, and differed only in size. In the Magnitude task, greater numbers were clicked higher than lesser numbers, $b=0.23$
$\pm 0.04$ SEM, $p<.0001$. In the Algebra task, greater numbers were still clicked significantly higher, but the deflection was dampened, $b=0.12 \pm .03$ SEM, $p=.0001$. Thus, there was spatial-numerical deflection in both tasks, but the amount of deflection was greater with explicit attention to magnitude.
On the horizontal axis, the effect of Magnitude was smaller but still significant ( $b=0.06 \pm 0.02 \mathrm{SEM}, t=2.9, p$ $<.01$ ). While there was no evidence that this magnitudebased deflection was moderated significantly by the Task ( $b$ $=-0.02 \pm 0.04 \mathrm{SEM}, t=-0.5, \mathrm{p}>.6$ ), additional analysis revealed that the magnitude-based horizontal deflection was only reliable in the Magnitude task, $b=0.07 \pm .03$ SEM, $\mathrm{p}=$ .01. In the Algebra task, by contrast, there was no evidence of a magnitude-based deflection along the horizontal axis, $b$ $=0.05 \pm 0.04$ SEM, $p=.17$.


Figure 2: Magnitude-based spatial deflection while manipulating equations. The vertical axis indicates mean spatial deflection, normalized for each subject (i.e., zscored). Interactions with greater numbers (red squares) were deflected upward and rightward; interactions with lesser numbers (blue circles), downward and leftward. This occurred in both tasks, but it was significantly more pronounced in the Magnitude task. (Error lines = SEM.)

## Discussion

We investigated whether symbol-mediated algebraic reasoning activates spatial representations of number, using dense recordings of algebraic manipulations (DREAM). Manipulations of algebraic equations were deflected upward and rightward when interacting with greater numbers, and downward and leftward when interacting with lesser numbers. The strength of this magnitude-based deflection, however, was moderated by the task. Spatial deflection was greatest when the task required explicit attention to numerical magnitude, and was dampened when the task
required algebraic reasoning. This was true even though the two tasks involved interacting with identical equations using comparable movements. In sum, when manipulating equations, people automatically activate a spatial representation of numerical magnitude, and the strength of this activation depends on the task's mathematical demands.

Spatial deflection along the horizontal axis was less pronounced than along the vertical axis. One explanation of this finding is that algebraic notation uses horizontal spacing for another purpose: to indicate syntactic hierarchy. In algebraic notation, higher-precedence operations are often written with little space between operands (e.g., $3 \cdot x \cdot y$ ) or no space at all (e.g., 3xy), while lower-precedence operations often introduce additional space between operands (e.g., $3+$ $\mathrm{x}+\mathrm{y}$ ). Thus, during equation manipulation, the horizontal axis may be co-opted to represent algebraic structure, dampening horizontal representations of numerical magnitude (Landy \& Goldstone, 2007; Landy, Allen, \& Zednik, 2014). By contrast, on purely numerical or arithmetic tasks, numerical magnitude does deflect hand movements along the horizontal axis: to the left for lesser magnitudes, and to the right for greater magnitudes (Marghetis et al, 2014; Faulkenberry, 2016).

## A new DREAM for studying algebraic reasoning

The study reported here is the first to use a methodology that we have dubbed dense recording of algebraic manipulations (DREAM) to gain insight into the cognitive processes at work during algebraic reasoning. Similar computer mousetracking approaches have been used to study the dynamics of simple numerical judgments (e.g., Faulkenberry, 2016; Song \& Nakayama, 2008) and mental arithmetic (e.g., Marghetis et al, 2014). DREAM extends this mousetracking methodology to a domain where manual interaction with external symbols is not just an artificial feature of the experimental design, but an integral part of the mathematical activity itself. One contribution of this study is to introduce this data-rich paradigm, which we hope can open new avenues of inquiry into mathematical cognition.
Algebraic reasoning is powerful because it transforms difficult conceptual tasks into a series of simple, robust physical manipulations of stable external symbols (Hutchins, 1995). Indeed, it is a canonical example of a cognitive accomplishment that depends on distributing the cognitive load across time and space. This requires coordinating skull-internal processes (perception, planning) with external processes like writing and gesturing. At its core, therefore, the practice of algebra demands the skillful use of hands: writing and erasing equations; using a finger to point to some aspect of an equation. DREAM allows us to analyze this distributed 'manual labor' that is a natural part of algebraic activity.

## Soft-assembling space for mathematics

This is the first evidence that spatial representations of number are activated during algebraic reasoning. Previous research, however, has documented other spatial processes that play a role in algebraic reasoning. The conventions of our algebraic notation use horizontal spacing to indicate syntactic hierarchy: higher-precedence operations are compressed (e.g., xy), while lower-precedence operations introduce more space between symbols (e.g., $x+y$ ). Participants are sensitive to these conventions (Landy \& Goldstone, 2007; Landy et al, 2014). Once participants master the basic syntax of algebra, moreover, they retrain their visual system so that they literally see equations as consisting of visual objects that respect the syntactic hierarchy of algebra (e.g., $x \cdot a+y \cdot b$ is perceived as two objects: 'x • a' and 'y •b'; Marghetis, Landy, \& Goldstone, 2016). The current study adds to this list of spatial processes that are deployed to solve algebraic equations.
This menagerie of spatial processes raises the question of how they are all brought into coordination. We favor an account where these different brain-based spatial resources are soft-assembled: they are brought into coordination in a way that is both transient and situated, responding to the demands of the task and the material environment (Clark, 2008). On this account, the development of mathematical expertise is not merely a process of piling new insights on top of old. Instead, the mathematical expert learns to combine, flexibly, a range of spatial processes, sometimes deploying one representation, other times another.
This account raises just as many questions as it answers. First, what is the time course of these processes? Are they all activated at once, or are they brought online sequentially in a cascade of activations? Marghetis and colleagues (2014), for instance, documented how, when individuals perform exact symbolic arithmetic (e.g., $2+7$ ), they first activate a spatial representation of the magnitude of the first number, then of the arithmetic operation, and finally of the solution. A similar cascade may occur in algebra.
Second, given our limited cognitive resources, how do all these mathematical facets-magnitude, arithmetic, algebraic syntax-become coupled to space without conflicting with each other? The spatial impact of relative magnitude was dampened significantly when the goal was to solve the equation rather than to judge relative magnitude. This suggests that spatial representations of number may fade over time, particularly when it comes to mathematical activities, like algebra, that foreground structural relationships over numerical magnitudes.
This fading may occur on multiple timescales, from the developmental to the momentary. On a developmental timescale, mathematical expertise might involve redeploying spatial resources to represent arithmetic or algebraic relations, pushing aside representations of magnitude. On shorter timescales, the activation of spatial
representations may be task-dependent, as it was in this study, or change from moment-to-moment-for instance, as an individual goes from identifying the symbol they intend to manipulate, to actually moving that symbol. Thus, as attention shifts away from magnitude or as other concepts acquire spatial associations, a symbol may become "semantically bleached" of its spatial-numerical content.

Indeed, the current results leave open the question of whether these spatial-numerical associations play a functional role in algebraic reasoning. Taken to the extreme, our results are consistent with an account wherein, for higher mathematics, spatial representations of number are largely epiphenomenal, playing a diminished role as spatial circuits are re-deployed to represent other aspects of the mathematical content (e.g., hierarchical algebraic structure).

## Conclusion

Are spatial representations of number really as ubiquitous as some have assumed, or are they limited to simple numerical tasks? Using dense behavioral recordings of equation manipulations, we found that numerical magnitude did, indeed, cause deflections that suggest a bottom-to-top and left-to-right spatial representation of number. This occurred even when the task was entirely algebraic, though the deflections were more pronounced when the task did require attending to magnitude. Our capacity for algebraic reasoning depends on a host of skills and processes - many of which are spatial - that must be brought in and out of coordination during situated reasoning. This singular ability would be impossible without the capacity to cobble together such processes both flexibly and dynamically.

## Acknowledgments

We thank C. J. Casper, V. Updike, B. Graves, and S. Malpeddi for data collection, and B. Rogers for comments.

## References

Cipora, K., \& Nuerk, H (2013). Is the SNARC effect related to the level of mathematics? QJEP, 66, 1974-1991.
Cipora, K., Patro, K., Nuerk, H. (2015). Are Spatial Numerical Associations a Cornerstone for Arithmetic Learning? Mind, Brain, \& Education, 9, 190-206.
Clark, A. (2008). Supersizing the Mind. Oxford U. Press.
Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122, 371.
Faulkenberry, T. J. (2016). Testing a direct mapping versus competition account of response dynamics in number comparison. J. of Cognitive Psychology, 28, 825-842.
Gunderson, E., Ramirez, G., Beilock, S., Levine, S. (2012). The relation between spatial skill and early number knowledge. Developmental Psychology, 48, 1229-1241.

Hartmann, M., Grabherr, L., Last, F. (2011). Moving along the mental number line. $J E P: H P P, 38,1416-1427$.
Hubbard, E. M., Piazza, M., Pinel, P., \& Dehaene, S. (2005). Interactions between number and space in parietal cortex. Nature Rev. Neuroscience, 6, 435-448.
Hutchins, E. (1995). Cognition in the Wild. MIT Press.
Knops, A., Thirion, B., Hubbard, E., Michel, V., Dehaene, S. (2009). Recruitment of an area involved in eye movement during mental arithmetic. Science, 324, 1583
Lakoff, G., \& Núñez, R. (2000). Where Mathematics Comes From. New York: Basic Books.
Landy, D, Goldstone, R. (2007). How abstract is symbolic thought? Journal of Experimental Psychology: Learning, Memory, \& Cognition, 33, 720-733.
Landy, D, Allen, C, Zednik, C (2014). A perceptual account of symbolic reasoning. Frontiers in Psychology, 5, 275.
Marghetis, T., \& Núnez, R. (2013). The motion behind the symbols: a vital role for dynamism in the conceptualization of limits and continuity in expert mathematics. Topics in Cognitive Science, 5, 299-316.
Marghetis, T, Núñez, R, Bergen, B (2014) Doing arithmetic by hand: Hand movements during exact arithmetic reveal systematic, dynamic spatial processing. Quarterly Journal of Experimental Psychology, 67, 1579-1596.
Núñez, R \& Marghetis, T (2015). Cognitive linguistics and the concept (s) of number. In R. Cohen-Kadosh \& K. Dowker (eds.), Oxford Handbook of Numerical Cognition. Oxford: Oxford University Press.
Sella, F., Sader, E., Lolliot, S., \& Cohen Kadosh, R. (2016). Basic and advanced numerical performances relate to mathematical expertise but are fully mediated by visuospatial skills.Journal of Experimental Psychology: Learning, Memory, and Cognition, 42, 1458-1472.
Shaki, S., Fischer, M. H., \& Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. $P B R, 16,328-331$.
Siegler, R. S., \& Ramani, G. B. (2009). Playing linear number board games-but not circular ones-improves low-income preschoolers' numerical understanding. Journal of Educational Psychology, 101, 545.
Song, J. J., \& Nakayama, K. (2008). Numeric comparison in a visually-guided manual reaching task. Cognition, 106, 994-1003.
Winter, B., Marghetis, T., \& Matlock, T. (2015). Of magnitudes and metaphors. Cortex, 64, 209-224.
Zorzi, M., Priftis, K., \& Umiltà, C. (2002). Neglect disrupts the mental number line. Nature, 417, 138-139.

# Chunking Ability Shapes Sentence Processing at Multiple Levels of Abstraction 

Stewart M. McCauley (Stewart.McCauley@liverpool.ac.uk)<br>Department of Psychological Sciences, University of Liverpool

Erin S. Isbilen (esi6@cornell.edu)<br>Morten H. Christiansen (christiansen@cornell.edu)<br>Department of Psychology, Cornell University


#### Abstract

Several recent empirical findings have reinforced the notion that a basic learning and memory skill-chunking-plays a fundamental role in language processing. Here, we provide evidence that chunking shapes sentence processing at multiple levels of linguistic abstraction, consistent with a recent theoretical proposal by Christiansen and Chater (2016). Individual differences in chunking ability at two different levels is shown to predict on-line sentence processing in separate ways: i) phonological chunking ability, as assessed by a variation on the non-word repetition task, predicts processing of complex sentences featuring phonological overlap; ii) multiword chunking ability, as assessed by a variation on the serial recall task, is shown to predict reading times for sentences featuring long-distance number agreement with locally distracting number-marked nouns. Together, our findings suggest that individual differences in chunking ability shape language processing at multiple levels of abstraction, consistent with the notion of language acquisition as learning to process.


Keywords: sentence processing; chunking; learning; memory; usage-based approach; language

## Introduction

Language takes place in real time; a fairly uncontroversial observation, yet one with far-reaching consequences that are rarely considered. For instance, a typical English speaker produces between 10 and 15 phonemes per second (Studdert-Kennedy, 1986), yet the ability of the auditory system to process discrete sounds is limited to around 10 per second, beyond which the signal is perceived as a single buzz (Miller \& Taylor, 1948). Moreover, the auditory trace is limited to about 100 ms (Remez et al., 2010). Compounding matters even further, human memory for sequences is limited to between 4 and 7 items (e.g., Cowan, 2001; Miller, 1956). Simply put, the sensory signal is so incredibly short-lived, and our memory for it so very limited, that language would seem to stretch the human capacity for information processing beyond its breaking point. We refer to this as the Now-or-Never bottleneck (Christiansen \& Chater, 2016).

How is language learning and processing possible in the face of this real-time constraint? A key piece of the puzzle, we suggest, lies in chunking: through experience with language, we learn to rapidly recode incoming information into chunks which can then be passed to higher levels of representation.

As an intuitive demonstration of the necessity of chunking, imagine being tasked with recalling a string of letters, presented auditorily: uopfmreepoa ecsgnpl $i r$. After a single presentation of the string, very few listeners would be able to recall a sequence consisting of even half of the letters (cf. Cowan, 2001). However, if exposed to the exact same set of letters but re-ordered slightly, virtually any listener would able to recall the entire sequence with ease: frog m o u s e paperpencil. Clearly, such a feat is possible by virtue of the ability to rapidly chunk the sequence into familiar sub-sequences (frog, mouse, paper, pencil).

According to the proposal of Christiansen and Chater (2016), the Now-or-Never Bottleneck requires language users to perform similar chunking operations on speech and text in order to process and learn from the input. This is necessary both due to the fleeting nature of sensory memory and the speed at which information is encountered during processing. Specifically, language users must perform Chunk-and-Pass processing, whereby input is chunked as rapidly as possible and passed to a higher, more abstract level of representation. Information at higher levels must also be chunked before being passed to still higher, increasingly abstract levels of representation.

Thus, in order to communicate in real-time, language users must chunk at multiple levels of abstraction, ranging from the level of the acoustic signal to the level of phonemes or syllables, to words, to multiword units, and beyond. Indeed, mounting empirical evidence supports the notion of chunking at levels higher than that of the individual word: children and adults appear to store and utilize chunks consisting of multiple words in comprehension and production (e.g., Arnon \& Snider, 2010; Bannard \& Matthews, 2008). Moreover, usage-based (e.g., Tomasello, 2003) and generative (e.g., Culicover \& Jackendoff, 2005) theoretical approaches have highlighted the importance of such units in grammatical development and sentence processing alike.

Chunking has been considered a key learning and memory mechanism in mainstream psychology for over half a century (e.g., Miller, 1956), and has been used to understand specific aspects of language acquisition (e.g., Jones, 2012; Jones, Gobet, Freudenthal, \& Pine, 2014). Nevertheless, few have sought to understand how it may shape more complex linguistic skills, such as sentence processing. McCauley and Christiansen (2015) took an initial step in this direction, showing that individual
differences in low-level chunking abilities were predictive of reading times for sentences involving relative clauses, demonstrating the far-reaching impact of basic chunking skills in shaping complex linguistic behaviors.

The present study seeks to evaluate the predictions of the Chunk-and-Pass framework more closely, by examining individual variation in chunking at two different levels of abstraction. Specifically, whereas chunking has previously been treated as a uniform memory ability, we test the novel theoretical prediction that chunking abilities may be relatively independent at different levels of linguistic abstraction. Participants were first asked to take part in a multiword-based serial recall task (Part 1) designed to yield a measure of chunking at the word level. This was followed by a variation on the non-word repetition task (Part 2), designed to yield a measure of phonological chunking ability. Importantly, due to the memory limitations discussed above, participants must utilize chunking in order to recall more than a few discrete words or phonemes in these tasks (e.g., Cowan, 2001; Miller, 1956). Finally, participants took part in an online self-paced reading task (Part 3). The results show that chunking ability at each level predicts different aspects of sentence processing ability: chunking at the phonological level predicts the extent to which low-level phonological information interferes with or facilitates complex sentence processing, while chunking at the multiword level predicts the role of local information in processing sentences with long-distance dependencies.

## Part 1: Measuring Individual Differences in Word Chunking Ability

The first task sought to gain a measure of individual participants' ability to chunk words into multiword units. To this end, we specifically isolate chunking as a mechanism by employing a classic psychological paradigm: the serial recall task. Serial recall has a long history of use in studies of chunking, dating back to some of the earliest relevant work (e.g., Miller, 1956), as well being used to extensively study individuals' chunking abilities (e.g., Ericsson, Chase, \& Faloon, 1980).

Participants were tasked with recalling strings of 12 individual words, with each string consisting of 4 separate word trigrams extracted from a large corpus of English. Importantly, in order to recall more than a few discrete items (as few as 4 in some accounts; e.g., Cowan, 2001), listeners must chunk the words of the input sequence into larger, multiword units. In this case, we expect them to draw upon linguistic experience with the trigrams in the experimental items.

In addition, we included a baseline performance measure: matched control strings, which featured identical functors to the experimental sequences, along with frequency-matched content words (to avoid semantic overlap effects on recall), presented in random order. Thus, comparing recall for experimental and control trials provides a measure of word chunking ability that reflects language experience while
controlling for such factors as attention, motivation, and-to the extent that it is separable-working memory.

## Method

Participants 42 native English speakers from the Cornell undergraduate population ( 17 females; age: $M=19.8$, $S D=1.2$ ) participated for course credit. Of the original 45 subjects, one was excluded due to audio recording errors, while two subjects failed to complete all three tasks.
Materials Experimental stimuli consisted of word trigrams spanning a range of frequencies, extracted from the American National Corpus (Reppen, Ide \& Suderman, 2005) and the Fisher corpus (Cieri, Graff, Kimball, Miller \& Walker, 2004). The combined corpus contained a total of 39 million words of American English. Each item was compositional (non-idiomatic). Item frequencies, per million words, ranged from 40 to .08 , averaging at .73 .

Each word was synthesized independently using the Festival speech synthesizer (Black, Clark, Richmond, King \& Zen, 2004) and concatenated into larger strings consisting of 12 words ( 4 trigrams). Each trigram was matched as closely as possible for frequency with the others occurring in a sequence.

To provide a non-chunk-based control condition, each item was matched to a sequence of words which contained identical functors but random frequency-matched content words (in order to avoid semantic overlap effects on recall, content words were not re-used). The ordering of the words was then randomized. An example of a matched set of sequences is shown below:

1) have to eat good to know don't like them is really nice
2) years got don't to game have she mean to them far is

The final item set consisted of 20 sequences (10 experimental, 10 control).
Procedure Each trial featured a 12-word sequence presented auditorily. Each word was followed by a 250 ms pause. Immediately upon completion of the string, the participant was prompted to verbally recall as much of the sequence as possible. Responses were recorded digitally and later transcribed by a researcher blind to the conditions as well as the purpose of the study.

The presentation order of the sequences was fully randomized. The entire task took approximately 15 minutes.

## Results and Discussion

Participants recalled significantly more words from experimental strings than the frequency-matched control sequences. The overall recall rate for words occurring in experimental items was $74.0 \%$ ( $S E=2.3 \%$ ), while the recall rate for control sequences was just $39.2 \%$ ( $S E=1.1 \%$ ). The difference between conditions was significant $(t(41)=18.8$, $p<0.0001$ ).

As the purpose of Part 1 was to gain an overall measure of chunk sensitivity, we calculated the difference between conditions individually for each subject ( $M=34.8 \%$,
$S E=1.8 \%$ ), which afforded a measure of word-chunking ability that reflects language experience while controlling for factors such as working memory, attention, and motivation. We refer to this difference measure as the Word Chunk Sensitivity score, and it is used as a predictor of sentence processing ability in Part 3.

In addition to bolstering previous empirical support for compositional (non-idiomatic) multiword sequences as linguistic units in their own right (e.g., Bannard \& Matthews, 2008), Part 1 revealed considerable individual differences across participants in word chunking ability. Recall rates for experimental items ranged from as high as $93.3 \%$ to just $30.4 \%$, with difference scores across the conditions ranging from $50.8 \%$ as low as $3.0 \%$.

## Part 2: Measuring Individual Differences in Phonological Chunking Ability

While the first task sought to gain a measure of individual participants' chunking abilities at the level of words, Part 2 sought to gain a measure of chunking ability at the phonological level. To this end, we re-purposed the standard non-word repetition (NWR) task as a chunking task. NWR has been used extensively to study various aspects of language development. Recent studies, however, have suggested that chunking may better account for NWR performance than more nebulous psychological constructs, such as working-memory (e.g., Jones, 2012; Jones et al., 2014). In one sense, the NWR task can be re-conceptualized as a serial recall task, as in Part 1. Following such work, and in keeping with the Now-or-Never perspective outlined above, we propose that individual differences in chunking ability underlie differences in NWR performance. In turn, NWR-with appropriately constructed stimuli-can serve as an additional dimension along which to measure chunking ability at the level of phonological processing.

Participants engaged in a standard NWR task, with each non-word consisting of 4,5 , or 6 syllables. However, the stimuli were designed such that the same set of syllables occurred in two different non-words, but in different orderings: one ordering yielded an item with high "chunkability," according to corpus statistics, while the other was estimated to be less "chunkable." The two items were then counterbalanced across halves of the task.

## Method

Participants The same 42 subjects from Part 1 participated directly afterwards in this task.
Materials Non-words were generated using an algorithm which took a large list ${ }^{1}$ of English syllables and randomly generated syllable combinations that were evaluated according to distributional statistics at the phoneme level. For the purpose of supplying statistics, the combined corpus used in Part 1 was automatically re-transcribed phonetically using the Festival speech synthesizer (Black et al., 2004).

[^407]For each of three different syllable lengths (4-, 5-, and 6syllables), the algorithm extracted item pairs that differed maximally in sequence likelihood (based on phoneme trigram statistics) across two different sequential orderings of the same set of syllables. In other words, pairs were selected in which one ordering of syllables was highly "chunk-like," while the other ordering of the same syllables was less "chunk-like," according to the phoneme statistics of the corpus. Four sets of non-words (the four in which the pair differed most greatly in terms of sequence likelihood) were selected for each syllable length. An example of a highly "chunk-like" 4-syllable item is krew-ih-tie-zuh, which was matched to the less chunk-like tie-zuh-ih-krew.

Thus, the final set of items included 24 non-words, eight in each of three syllable-length conditions (4-, 5-, and 6syllable), with four being highly "chunk-like" and the other four consisting of alternate orderings of the same syllables which were statistically less "chunk-like."
Procedure The task was split into two blocks, with all NWR item pairs counterbalanced between them. The auditory presentation of each non-word was followed by a 1500 ms pause, after which the participant was prompted to recall the item verbally. As with Part 1, responses were recorded digitally and scored offline. The task took approximately 4 minutes to complete.

Correct responses received a score of 1. Responses involving alteration to a single phoneme (usually a vowel substitution, which could easily stem from differences in regional dialect) received a score of 0.5 . All other responses received scores of 0 .

## Results and Discussion

Participants achieved a mean NWR accuracy rate of $54.1 \%$ ( $S E=2.3 \%$ ). While the overall differences between the high chunk-like $(\mathrm{M}=55.2 \%, \quad \mathrm{SE}=2.5 \%)$ and low chunk-like ( $\mathrm{M}=53.1 \%, \mathrm{SE}=2.5 \%$ ) conditions were in the expected direction, they were subtle, with a mean difference of $2.1 \%$ (non-significant: $t(41)=1.12, p>0.1$ ). However, there was considerable individual variation in the size of this difference across participants ( $S E=1.9 \%$ ), ranging from $29.2 \%$ to less than $0 \%$, at $-16.6 \%$. Therefore, in Part 3, we assess both the overall NWR performance score as well as the difference between the conditions (which we refer to as the Phonological Chunk Sensitivity score) as predictors of sentence processing.

Importantly, neither the overall raw task performance ( $\beta=-0.03, p=0.9$ ) nor the Chunk Sensitivity scores $(\beta=-0.19$, $\mathrm{p}=0.22$ ) from Parts 1 and 2 correlated with one another, consistent with the notion that chunking at each level may have different consequences for sentence processing.

## Part 3: Measuring Individual Differences in Sentence Processing and Chunking

In Part 1, we sought to gain a measure of individual participants' ability to chunk words together, while Part 2 aimed to provide a measure of phonological chunking ability. In Part 3, the same subjects from the first two parts
participated in a self-paced reading task designed to: i) assess on-line sentence processing across two different sentence types which were hypothesized to involve chunking at the word and phonological levels, but to different extents; ii) determine the extent to which chunking ability, as assessed in the first two tasks, predicted processing difficulties for each sentence type.

The first sentence type featured long distance subject-verb number agreement with locally distracting number-marked nouns, exemplified by (1):

1. The key to the cabinets was rusty from many years of disuse.

Previous work (Pearlmutter, Garnsey, \& Bock, 1999) has shown that readers are slower to process the verb when the number of the local noun (cabinets) does not match that of the head noun (key), resulting in the sequence (cabinets was). Reading times are compared to sentences in which the number marking matches, as exemplified by (2):

## 2. The key to the cabinet was rusty from many years of disuse.

In other words, reading times are higher at the verb when the local information is distracting. Following the finding that text-chunking ability predicts decreased difficulty with complex sentences involving long-distance dependencies (McCauley \& Christiansen, 2015), we hypothesized that participants with higher Word Chunk Sensitivity scores (Part 1) would be less susceptible to interference from local information in sentences such as (1). Subjects that are better able to rapidly chunk words together and pass them to higher levels of representation should not only experience decreased computational burden from long-distance dependencies, but should be less affected by locally distracting information.

The second sentence type featured object-relative (OR) clauses, which have been shown to be processed with greater ease by good text chunkers (McCauley \& Christiansen, 2015). However, in the present study we added an element of phonological interference: two pairs of words in each sentence exhibited phonological overlap. Previous work has shown that low-level phonological overlap can interfere with the processing of sentences featuring relative clauses (Acheson \& MacDonald, 2011). An experimental item and its matched control are shown in (3) and (4):
3. The cook that the crook consoles controls the politician. 4. The prince that the crook comforts controls the politician.

In line with the Chunk-and-Pass framework, we predicted that better phonological chunkers, as assessed in Part 2, would be less susceptible to phonological interference, by virtue of their ability to more rapidly chunk and pass phonological information to a higher level of representation.

Thus, participants' resilience to phonological interference was hypothesized to be better predicted by Phonological Chunk Sensitivity (Part 2), while participants' susceptibility to local number mismatch was expected to be better predicted by Word Chunk Sensitivity (Part 1).

## Method

Participants The same 42 subjects from Parts 1 and 2 participated in Part 3 immediately afterwards.
Materials There were two sentence lists-counterbalanced across subjects-each consisting of 9 practice items, 20 experimental items, 20 matched control items, and 68 filler items. There were two experimental conditions, each with 20 items; the first consisted of the OR sentences featuring phonological overlap (the first 20 items from Acheson \& MacDonald, 2011). The second experimental condition consisted of grammatical sentences featuring long-distance number agreement with locally distracting number-marked nouns (the 16 items from Pearlmutter et al., 1999, plus four additional sentences with the same properties).

Each list included, for each condition, 10 of the items in their experimental form and 10 of the items in their control form (without rhymes in the case of the OR sentences; without locally distracting nouns in the case of the number agreement sentences). The lists were counterbalanced such half of the subjects saw the experimental versions of sentences the other half saw in their control form.
Procedure Materials were presented in random order using a self-paced, word-by-word moving window display (Just, Carpenter, \& Woolley, 1982). At the beginning of each trial, a series of dashes appeared (one corresponding to each nonspace character in the sentence). The first press of a marked button caused the first word to appear, while subsequent button presses caused each following word to appear. The previous word would return once more to dashes. Reaction times were recorded for each button press. Following each sentence, subjects answered a yes/no comprehension question using buttons marked "Y" and "N." The task took approximately 10 minutes.

## Results and Discussion

Only trials with correct answers to comprehension questions were analyzed. Accuracy for the number agreement condition was $88.3 \%$; for the object-relatives it was $80.0 \%$. Following Acheson \& MacDonald (2011), raw reaction times over 3000 ms were excluded. Prior to analysis, raw reaction times (RTs) were log-transformed.
Mean RTs for the main verb in the number agreement and phonological overlap sentences were comparable to those in the corresponding original studies (respectively: Pearlmutter et al., 1999; Acheson \& MacDonald, 2011), as was the size of the mean difference between conditions. In the number agreement condition, the verb in experimental items ( $\mathrm{M}=361.1$, $\mathrm{SE}=19.9$ ) was processed more slowly than in controls ( $\mathrm{M}=316.7$, $\mathrm{SE}=13.9$ ), a mean difference of 44 ms (F1[1,41]=12.7, $p<0.001 ; F 2[1,18]=10.2, p<0.01)$. There was a fair amount of individual variation in the difference


Fig. 1: Correlation between Word Chunk Sensitivity (derived from recall scores in Part 1) and the difference in main verb RTs between sentences with locally distracting number information vs. control sentences.
between conditions ( $\mathrm{SD}=79.4$ ).
The critical main verb in OR sentences featuring phonological overlap was processed more slowly ( $\mathrm{M}=605.1$, $\mathrm{SE}=70.6$ ) than in matched controls ( $\mathrm{M}=546.3$, $\mathrm{SE}=42.2$ ), a mean difference of 58.8 which was non-significant ( $F 1[1,41]=1.21, \quad p=0.28 ; \quad F 2[1,18]=0.04, \quad p=0.8 ; \quad$ see discussion). There was, however, considerable individual variation in the difference between conditions ( $\mathrm{SD}=343.7$ ), especially relative to the size of group mean difference.

We were primarily interested in the extent to which differences in RTs between experimental and control sentences could be predicted by the Chunk Sensitivity measures collected in Parts 1 and 2. Below, we analyze these relationships using multiple linear regression, with Word Chunk Sensitivity and Phonological Chunk Sensitivity scores as predictors of RT differences between conditions (recall that the two metrics were not correlated). ${ }^{2}$

For the difference between sentences featuring locally distracting number information and their control counterparts, we found that Word Chunk Sensitivity was a significant predictor of RT difference at the verb ( $\beta=-0.79$, $t=-3.19, p<0.01$ ), while Phonological Chunk Sensitivity and the interaction term did not reach significance. The model for the significant main effect had an $R$ value of 0.42 . The correlation between Word Chunk Sensitivity and the RT difference is depicted in Figure 1. As can be seen, subjects with higher Word Chunk Sensitivity scores appear less susceptible to interference from the locally distracting number information, as reflected by lower differences between verb RTs for experimental vs. control sentences.

With regard to the difference between OR sentences with and without phonological overlap, we found that Phonological Chunk Sensitivity was a significant predictor of RT differences at the main verb ( $\beta=-3.49, t=-2.43$, $p<0.05$ ), while Word Chunk Sensitivity and the interaction

[^408]

Fig. 2: Correlation between Phonological Chunk Sensitivity (derived from repetition scores in Part 2) and the difference in main verb RTs for OR sentences with and without phonological overlap between words.
term did not reach significance. The model for the significant main effect had an $R$ value of 0.36. A scatterplot showing the correlation between Phonological Chunk Sensitivity and the RT difference is shown in Figure 2: better chunking ability resulted in less phonological interference.

Thus, consistent with the predictions of the Chunk-andPass framework, we find evidence for the notion that chunking ability shapes sentence processing differently at two separate levels of abstraction: participants who were more sensitive to word chunk information better processed long-distance dependencies in the face of conflicting local information, while those with higher phonological chunk sensitivity better processed complex sentences with phonological overlap between words. That the two chunk sensitivity measures did not correlate with one another further underscores the notion of chunking taking place at multiple levels of abstraction.

While we failed to find the same effect of phonological overlap on processing as did Acheson and MacDonald (2011), it is likely that our subjects (Cornell undergraduates) had more reading experience than subjects at UW-Madison, and experienced less interference overall. Nonetheless, our measure of phonological chunk sensitivity was sensitive enough to pick up individual differences that predicted sentence processing in the face of phonological interference.

Intriguingly, participants with very high Phonological Chunk Sensitivity appeared to experience an advantage for OR sentences featuring phonological overlap. This raises the possibility that such subjects benefitted from phonologically-based priming of subsequent rhyme words in sentences such as (3). Further work will be necessary to evaluate this possibility.

## General Discussion

In the present study, we show that individual differences in chunking ability predict on-line sentence processing at multiple levels of abstraction: chunking at the phonological level is shown to predict the way phonological information
is used during complex sentence processing, while chunking at the multiword level is shown to predict the ease with which long-distance dependencies are processed in the face of conflicting local syntactic information. In Part 1, we adapted the serial recall task-a paradigm used for over half a century to study memory, including chunking phenomena-in order to gain a measure of individual variation in subjects' ability to chunk word sequences into multiword units. In Part 2, subjects participated in a NWR task with non-words designed to vary according to the ease with which their phonemes could be chunked. The difference in correct repetition rates between highly chunkable and less chunk-able items provided a measure of individual variation in chunking ability at the phonological level. Finally, in Part 3 we showed that chunking at the multiword level was predictive of processing for sentences with long-distance dependencies and distracting local information, while chunking at the phonological level was predictive of complex sentence processing in the presence of phonological overlap between words.

Expanding on the findings of a previous study that showed low-level chunking of sub-lexical letter sequences to predict sentence processing abilities (McCauley \& Christiansen, 2015), the present study supports the notion that chunking not only takes place at multiple levels of abstraction, but that individuals' processing abilities may be differently shaped by chunking at each level. Moreover, chunking at lower levels (e.g., the phonological level) may have serious consequences for processing at higher levels (e.g., sentence processing).

This work is highly relevant to the study of language acquisition. The Now-or-Never bottleneck imposes incremental, on-line processing constraints on language learning, suggesting a key role for chunking. Indeed, a number of recent computational modeling studies have demonstrated that chunking can account for key empirical findings on children's phonological development and word learning abilities (Jones, 2012; Jones et al., 2014), while other work has captured a role for chunking in learning to comprehend and produce sentences (McCauley \& Christiansen, 2011, 2014). There exists a clear need for further developmental behavioral studies-including longitudinal studies-examining individual differences in chunking as they pertain to specific stages of language development as well as more general language learning outcomes.

## Acknowledgments

Thanks to Nick Chater and Gary Jones for helpful discussion, as well as S. Reig, K. Diamond, J. Kolenda, J. Powell, S. Goldberg, and D. Dahabreh for assistance with participant running and recruitment.

## References

Acheson, D.J. \& MacDonald, M.C. (2011). The rhymes that the reader perused confused the meaning: Phonological effects
during on-line sentence comprehension. Journal of Memory and Language, 65, 193-207.
Arnon, I. \& Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62, 67-82.
Bannard, C. \& Matthews, D. (2008). Stored word sequences in language learning. Psychological Science, 19, 241-248.
Christiansen, M.H. \& Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. Behavioral \& Brain Sciences, 39, e62.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24, 87-114.
Culicover, P.W. \& Jackendoff, R. (2005). Simpler syntax. New York: Oxford University Press.
Ericsson, K.A., Chase, W.G., \& Faloon, S. (1980). Acquisition of a memory skill. Science, 208, 1181-1182.
Jones, G. (2012). Why chunking should be considered as an explanation for developmental change before short-term memory capacity and processing speed. Frontiers in Psychology, 3 :167. DOI: 10.3389/fpsyg.2012.00167.
Jones, G., Gobet, F., Freudenthal, D., Watson, S.E. \& Pine, J.M. (2014). Why computational models are better than verbal theories: The case of nonword repetition. Developmental Science, 17, 298-310.
Just, M. A., Carpenter, P. A., \& Woolley, J. D. (1982). Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General, 111, 228-238.
McCauley, S.M. \& Christiansen, M.H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the $33^{r d}$ Annual Conference of the Cognitive Science Society (pp. 1619-1624). Austin, TX: Cognitive Science Society.
McCauley, S.M. \& Christiansen, M.H. (2014). Acquiring formulaic language: A computational model. Mental Lexicon, 9, 419-436.
McCauley, S.M. \& Christiansen, M.H. (2015). Individual differences in chunking ability predict on-line sentence processing. In D.C. Noelle \& R. Dale (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.
Miller, G.A. \& Taylor, W.G. (1948). The perception of repeated bursts of noise. Journal of the Acoustic Society of America, 20, 171-182.
Pearlmutter, N.J., Garnsey, S.M. \& Bock, K. (1999). Agreement processes in sentence comprehension. Journal of Memory and Language, 41, 427-456.
Remez, R.E., Ferro, D.F., Dubowski, K.R., Meer, J., Broder, R.S. \& Davids, M.L. (2010). Is desynchrony tolerance adaptable in the perceptual organization of speech? Attention, Perception, \& Psychophysics, 72, 2054-2058.
Studdert-Kennedy, M. (1986). Some developments in research on language behavior. In N.J. Smelser \& D.R. Gerstein (Eds.), Behavioral and social science: Fifty years of discovery (pp. 208248). Washington, DC: National Academy Press.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, MA: Harvard University Press.

# The Interaction of Worked-Examples/ Self-Explanation Prompts and Time on Algebra Conceptual Knowledge 

Kelly M. McGinn (kelly.mcginn@temple.edu)<br>Temple University, Department of Psychological Studies in Education, 1301 Cecil B. Moore Ave. Philadelphia, PA 19122 USA<br>Julie L. Booth (julie.booth@temple.edu)<br>Temple University, Department of Psychological Studies in Education, 1301 Cecil B. Moore Ave. Philadelphia, PA 19122 USA<br>Laura K. Young (laura.young@temple.edu)<br>Temple University, Department of Psychological Studies in Education, 1301 Cecil B. Moore Ave. Philadelphia, PA 19122 USA


#### Abstract

Success in Algebra I often predicts whether or not a student will pursue higher levels of mathematics and science. However, many students enter algebra holding persistent misconceptions that are difficult to eliminate, thus, hindering their ability to succeed in algebra. One way to address these misconceptions is to implement worked-examples and selfexplanation prompts, which have been shown to improve students' conceptual knowledge. However this effect seems to be greater after a delay. The current study sought to explore such time-related effects on algebra conceptual knowledge. In a year-long random-assignment study, students either studied worked-examples and answered self-explanation prompts ( $\mathrm{n}=$ 132) or solved typical isomorphic problems $(\mathrm{n}=140)$. A three-way mixed ANCOVA (pre-algebra knowledge $x$ condition x time) found a significant condition by time effect. The growth of algebra conceptual knowledge was greater for students studying worked-examples than for those solving typical problems.


Keywords: worked-examples; self-explanation prompts; algebra; conceptual knowledge

## Introduction

Algebra I is often considered to be a gate-keeper course, meaning that a student's success in the course often determines whether he or she will continue on to a higher level mathematics or science course (U.S. Department of Education, 1997). Furthermore, students in the United States tend to struggle mastering algebra concepts, potentially contributing to the lower enrollment of U.S. college students in mathematics and science related majors compared to competing countries.

The newly implemented Common Core State Standards (CCSSI, 2010) stresses the importance of both procedural and conceptual knowledge of mathematics content. However, especially when it comes to algebra, students hold persistent misconceptions, which hinder their ability to master the content. In fact, students often enter Algebra I holding strong misconceptions that may impact their success mastering algebra content (Brown, 1992; Chiu \& Liu, 2004; Kendeou \& van den Broek, 2005). For instance,
misconceptions such as believing that the equals sign is an indicator of operations to be performed (Baroody \& Ginsburg, 1983; Kieran, 1981; Knuth Stephens, McNeils, \& Alibali, 2006), that the negative sign represents only the subtraction operation and does not modify terms (Vlassis, 2004), that subtraction is commutative (Warren, 2003), and that variables cannot take on multiple values (Booth, 1984; Knuth et al., Kuchemann, 1978) are all thought to be critical. Holding such misconceptions have been shown to hinder students' success in problem solving (Booth \& Koedinger, 2008).

A large body of research supports the notion that eliminating mathematics misconceptions is not an easy task. In fact, many students continue to hold these misconceptions after traditional classroom instruction (Booth, Koedinger \& Siegler, 2007; Vlassis, 2004). Often in order to challenge a student's misconception, one must directly draw out and confront the faulty thinking (Donovan \& Bransford, 2005). A combination of worked-examples and self-explanation prompts has been used to do just that.

Worked-examples, which are mathematics problems with worked-out solutions, provide the opportunity to point out common misconceptions to students. Some textbooks offer a small number of worked-examples, often at the beginning of a chapter or section. However, research indicates that interleaved worked-examples, alternating between workedexamples and problems for students to solve, are more beneficial to learning (Clark \& Mayer, 2003; Sweller \& Cooper, 1985).

Furthermore, the benefit of worked-examples can be improved with the inclusion of self-explanation prompts, which are questions that prompt students to explain their reasoning. When students self-explain, they are able to integrate various pieces of knowledge, fill gaps in their own knowledge, and make new knowledge explicit (Chi, 2000; Roy and Chi, 2005). Students at all ability levels who are
prompted to self-explain learn more than those who do not self-explain (Chi, de Leeuw, Chiu \& Lavancher, 1994).

Often if a textbook uses a worked-example, it displays a correct problem solution. However, incorrect workedexamples have also shown benefits to learning. In empirical laboratory students, students who are asked to explain the errors in incorrect solutions, as well as explain effective strategies in correct examples, learn more than students who are asked to only explain correct examples (Durkin \& RittleJohnson, 2009; Siegler \& Chen, 2008).

While the use of worked-examples and self-explanation prompts have been shown to improve learning, often there is a delayed-effect, meaning that the effect is larger on a delayed post-test rather than immediately after the intervention. For instance, Adams and colleagues (2014) found that while students solving isomorphic problems with feedback and students studying incorrect examples did not differ significantly at immediate posttest, students in the incorrect example group scored significantly better on a delayed posttest compared to the problem-solving group. This suggests that the worked-example/self-explanation effect may improve over time.

## Current Study

This study applies previous laboratory research supporting the use of both correct and incorrect worked-examples paired with self-explanation prompts to the classroom. While highly controlled laboratory studies are necessary when developing theories, applied studies are needed in order to investigate the limits of generalization.

We explore the effects of studying worked-examples and answering self-explanation prompts compared to solving typical isomorphic problems on students' algebra conceptual knowledge. We hypothesize that students who study worked examples and answer self-explanation prompts will have less algebra misconceptions and, therefore, will have higher conceptual knowledge compared to those who solve traditional isomeric problems.

Finally, the current study will explore students' conceptual knowledge growth over the course of a full school year, extending the evidence to support a delayed effect by providing longitudinal evidence from repeated interventions.

## Methods

## Participants

Participants included 562 Algebra I students from 28 classrooms ( 12 teachers) from five school districts across the United States. The sample was $49 \%$ female. Students were classified as underrepresented minority (URM; Black, Hispanic, biracial) or non-URM (White, Asian); 65\% of the students were classified as URM. Participants were also
socioeconomically diverse, with $52 \%$ coming from families who qualified for the Free or Reduced Lunch program (FRL).

Due to the restrictions of repeated-measures ANOVA, only students who completed all four quarterly exams were included in the analysis. Due to natural attrition (i.e. students leaving the school or absence on the day of the quarterly exam) the sample was reduced to $272 ; 51 \%$ female, $61 \%$ URM, $50 \%$ FRL.

Classrooms were randomly selected to either complete problem- or example-based worksheets yielding 14 problem-based $(\mathrm{n}=140)$ and 14 example-based ( $\mathrm{n}=132$ ) classrooms. Of the 12 teachers, eight taught one class of each condition; however, two teachers instructed two classes of the problem-based condition and one class of the example-based condition, while two others instructed two example-based classes and one of the problem-based class.

## Procedure

Intervention During the school year, teachers taught the algebra content using their own typically teaching methods; however, they were asked to sporadically assign the 42 study-worksheets at times they deemed appropriate during the year. Teachers did not have to assign the worksheets if they did not cover that material in their curriculum. On average, teachers assigned 27 worksheets (ranging from 15 to 40) throughout the year. There was no significant difference in the number of worksheets assigned between groups, with the problem-based group completing an average of 28 worksheets and the example-based group completing an average of 26 worksheets, $p>.05$. Teachers were given the freedom to assign the worksheets in any order and were told to treat the assignments as they would any other assignment in their class; however they were instructed to have students complete the assignments during the class period, not for homework. Students were allowed to work together if the teacher typically permitted that behavior. Each assignment took about 20 minutes to complete.

The worksheets of both conditions contained four problemsets (with two math problems similar to each other per set). The problem-based worksheets contained four regular problem-sets where students were asked to simply solve each problem, similar to a typical math worksheet. The example-based worksheets replaced one math problem within each set with a worked-example and self-explanation prompt(s). Students in this group were instructed to study the worked-example, answer the self-explanation prompt, and complete the second math problem on their own. Each example-based worksheet contained two correct workedexamples and two incorrect worked-examples. See Figure 1 for sample problem- and example-based problem sets.


Figure 1. Sample problem- and example-based problem sets.
Assessment At the beginning of the school year, all students were given a pre-test assessing their pre-algebra knowledge. Throughout the school year, students were given four quarterly exams. The four exams contained the same 18 items, however teachers were asked to only assign the test items taught to date; therefore, students were not answering items containing content they were not already taught. This exam assessed both procedural and conceptual algebra knowledge. At the conclusion of the year, students were given a post-test, consisting of 10 Algebra I standardizedtest release items.

At the end of the school year, each school provided the researchers with student demographic information, such as gender, ethnicity, and free or reduced lunch qualification. Finally, teachers completed a survey answering questions about their use of the worksheets. The survey contained questions such as "How often did you review the worksheets with the students after completion?"

## Measures

Algebra Conceptual Knowledge The quarterly benchmark exams consisted of 18 items, each of which had multiple parts, yielding a total of 71 sub-items. Of these 71 subitems, 46 measured students' conceptual knowledge of algebra content. We operationally define conceptual knowledge as an understanding of the core features in problems for a given topic (e.g. Booth, 2011). Algebra conceptual knowledge scores were calculated for each quarter by dividing the number of correctly answered items by 46 . This score does not take into account the number of items attempted since each teacher assigned a different number of items each quarter.

Pre-algebra Knowledge The pre-algebra exam was given at the start of the school year before students completed any
study-worksheets. This exam covered content necessary for the success in an algebra course, such as the understanding of equality and difference between coefficient and constant. This exam consisted of 11 items with 71 sub-items. Prealgebra knowledge scores were calculated by dividing the total number of correctly answered items by 71 .

Teacher Reports At the end of the year, teachers were administered a survey about their experience in the study. In one item, they were asked about the frequency with which they reviewed study assignments in class. Teachers responded by selecting one of the following options: $0-20 \%$ of the time, $20-40 \%$ of the time, $40-60 \%$ of the time, $60-$ $80 \%$ of the time or $80-100 \%$ of the time. Teachers' responses were recoded into a $1(0-20 \%)$ to $5(80-100 \%)$ scale.

All measures were scored and coded by two researchers, checking for internal and external consistency.

## Results

The following analysis explores the effects of time, prealgebra knowledge and condition on students' algebra conceptual knowledge. Pre-algebra knowledge was included in the model because student' prior-knowledge is known to greatly influence their future learning. While other outcome measures (i.e. procedural knowledge and standardized test release items) were collected, they are beyond the scope of this study focused on conceptual knowledge growth. The other measures will or are presented in other reports. Finally, URM status and rate of teacher review were included as covariates because differences were found between conditions.

A three-way mixed ANCOVA was run to understand the effects of pre-algebra knowledge, condition, and time on algebra conceptual knowledge. The rate of review and minority status were included as covariates. Using Greenhouse-Geisser estimates, the interaction between condition, pre-algebra knowledge and time was not statistically significant; however, there was a statistically significant two-way interaction between time and all between-subject variables. See Table 1 for results.

Table 1. Greenhouse-Geisser estimates for 3-way ANCOVA for algebra conceptual knowledge.

|  |  |  |  | partial |
| ---: | :---: | :---: | :---: | :---: |
|  | $d f$ | $F$ | $p$ | $\eta^{2}$ |
| Quarter | 2.513 | 10.826 | $<.001$ | .055 |
| Quarter x URM | 2.513 | 4.333 | .008 | .023 |
| Quarter x Review <br> Quarter x Pre- <br> algebra | 2.513 | 17.900 | $<.001$ | .088 |
| Quarter x <br> Condition | 2.513 | 3.991 | .012 | .021 |
| Quarter x <br> Condition x <br> Pre-algebra <br> Residual | 467.459 | 11882 | $<.001$ | .389 |

See Table 2 and Figure 2 for condition by time estimated marginal means. At quarter 1, the example-based group scored slightly lower than the problem-based group; however, by quarter 4, the example-based group outscored the problem-based group.

Table 2. Condition by time estimated marginal means with 95\% confidence intervals.

|  |  |  |  | $95 \%$ CI |  |
| :--- | :--- | :---: | :---: | :---: | :---: |
| Condition | Quarter | Mean | SE | Lower | Upper |
| Problem- | 1 | .179 | .008 | .163 | .195 |
| based | 2 | .321 | .011 | .299 | .343 |
|  | 3 | .435 | .012 | .413 | .458 |
|  | 4 | .516 | .016 | .484 | .548 |
|  |  |  |  |  |  |
| Example- | 1 | .169 | .008 | .153 | .186 |
| bases | 2 | .332 | .011 | .310 | .354 |
|  | 3 | .447 | .012 | .423 | .470 |
|  | 4 | .568 | .017 | .535 | .601 |

Estimated Marginal Means

------ Problem ——— Example

Figure 2. Condition by time estimated marginal means.

## Discussion

Due to the nature of the quarterly exams, it was expected that students would score better over time. As mentioned in the procedure section, the conceptual knowledge portion of the quarterly exam consisted of 46 sub-items. However, students only attempted to answer the items in which they were familiar with. Therefore, students attempted to answer more items as they covered additional content over the course of the school year, leading to potential increased scores over time. However, we were more interested in the interaction between treatment and time. It was hypothesized that there would be differences in the rate of algebra conceptual knowledge growth between the example-based and problem-based groups.

As predicted, this analysis revealed a significant condition by time interaction. At the end of quarter 1, students solving typical algebra problems, in the problem-based group, scored slightly better than students in the example-based condition. However by the end of quarter 2, the opposite occurred. Students studying worked-examples and answering self-explanation prompts scored slightly higher than those in the problem-based group. This gap continued to widen throughout the remainder of the school year. By quarter 4, example-based students scored an average of 5 percentage points higher on the algebra conceptual knowledge test than the problem-based students, which is supported by previous studies finding a delayed effect (i.e. Adams et al., 2014).

The limitations of this study include a sample restricted to those present for all four quarterly exams. In addition, although a within-teacher design controlled for teacher-
related variables, it is possible that there was some contamination across classrooms. For instance, some teachers reported using a few of their own worked-examples with their problem-based classroom. The current analysis was based on linear growth; further studies should consider using a more robust analysis in order to account for possible quadratic or cubic growth curves.

This analysis adds to the current body of research by providing evidence from the classroom to support laboratory findings. It also extends our understanding of the short-term benefits of worked-examples and self-explanation prompts by offering longitudinal data. Our findings emphasize the need to measure learning over longer time intervals.

Based on these findings, it is suggested that teachers interleave worked-examples and self-explanation prompts with traditional algebra problems. In order to receive maximum benefit, students should be exposed to this approach consistently throughout the entire school year, not just in a single instance. Furthermore, such interventions should be interleaved in algebra textbooks, rather than simply displaying a few correct worked-examples at the beginning of a section. Finally, both correct and incorrect worked-examples should be used in the classroom to promote maximum benefit.

As previously noted, success in Algebra $I$ is a known gatekeeper to later mathematics and science success. However, many students enter algebra with persistent misconceptions that obstruct their achievement in algebra. The findings from this study suggest that using workedexamples combined with self-explanation prompts as classroom practice materials can improve student's conceptual knowledge, consequently decreasing their misconceptions. The findings from this study are particularly exciting as they come from a study that took place in actual classrooms and not research laboratories. Due to the setting of the current study, our findings illustrate that even when precision, like that provided in a laboratory, cannot be guaranteed the positive effect of using workedexamples paired with self-explanation prompts is still seen.

## Acknowledgments

Thanks are due to the participating school districts; without their interest this type of applied research would not be possible. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150456 to the Strategic Education Research Partnership (SERP) Institute. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## References

Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., \& Van Velsen, M. (2014).

Using erroneous examples to improve mathematics learning with a web-based tutoring system. Computers in Human Behavior, 36, 401-411.
Baroody, A., \& Ginsburg, H. (1983). The effects of instruction on children's understanding of the equals sign. The Elementary School Journal, 84, 199-212.
Booth, J. L. (2011). Why can't students get the concept of math? Perspectives on Language and Literacy, 37, 31-35.
Booth, J. L., \& Koedinger, K. R. (2008). Key misconceptions in algebraic problem solving. In B. C. Love, K. McRae, \& V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Cognitive Science Society (pp. 571576). Austin, TX: Cognitive Science Society.

Booth, J. L., Koedinger, K. R., \& Siegler, R. S. (2007, August). The effect of prior conceptual knowledge on procedural performance and learning in algebra. Poster presented at the 29th annual meeting of the Cognitive Science Society, Nashville, TN.
Booth, L. R. (1984). Algebra: Children's strategies and errors. Windsor, UK: NFER-Nelson.
Brown, D. E. (1992). Using examples and analogies to remediate misconceptions in physics: Factors influencing conceptual change. Journal of Research in Science Teaching, 29, 17-34.
Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In Glaser, R. (Ed.), Advances in instructional psychology (pp. 161-238). Mahwah, NJ: Lawrence Erlbaum Associates.
Chi, M. T. H., de Leeuw, N., Chiu, M., \& Lavancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.
Chiu, M. H., \& Liu, J. W. (2004). Promoting fourth graders' conceptual change of their understanding of electric current via multiple analogies. Journal of Research in Science Teaching, 42, 429-464.
Clark, R. C., \& Mayer, R. E. (2003). e-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. San Francisco, CA: Jossey-Bass.
Common Core State Standards Initiative (2010). Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.
Donovan, M. S., \& Bransford, J. D. (Eds.). (2005). How students learn: History, mathematics, and science in the classroom. Washington, DC: National Academy Press.
Durkin, K. L. \& Rittle-Johnson, B. (2009, April). Comparison of correct and incorrect examples when learning decimal fractions. Poster presented at the annual meeting of the Society for Research in Child Development, Denver, CO.

Kendeou, P., \& van den Broek, P. (2005). The effects of readers' misconceptions on comprehension of scientific text. Journal of Educational Psychology, 97, 235-245.
Kieran, C. (1981). Concepts associated with the equality symbol. Educational Studies in Mathematics, 12, 317326.

Knuth, E. J., Stephens, A. C., McNeil, N. M., \& Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. Journal for Research in Mathematics Education, 37, 297-312.
Kuchemann, D. (1978). Children's understanding of numerical variables. Mathematics in School, 7, 23-26.
Roy, M., \& Chi, M. T. H. (2005). Self-explanation in a multi-media context. In R. Mayer (Ed.), Cambridge handbook of multimedia learning (pp. 271-286). New York, NY: Cambridge Press.
Siegler, R. S., \& Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. Developmental Science, 11, 433-448.
Sweller, J., \& Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction, 2, 59-89.
U.S. Department of Education. (1997). Mathematics Equals Opportunity. White paper prepared for the U.S. Secretary of Education, Richard W. Riley.
Vlassis, J. (2004). Making sense of the minus sign or becoming flexible in 'negativity.' Learning and Instruction, 14, 469-484.
Warren, E. (2003). The role of arithmetic structure in the transition from arithmetic to algebra. Mathematics Education Research Journal, 15, 122-137.

# Semantic Ambiguity Effects: A Matter of Time? 

Joyse Medeiros (j.medeiros@bcbl.eu)<br>BCBL. Basque Center on Cognition, Brain, and Language<br>UPV. University of the Basque Country<br>Blair C. Armstrong (blair.armstrong@utoronto.ca)<br>Department of Psychology \& Centre for French \& Linguistics at Scarborough, University of Toronto<br>BCBL. Basque Center on Cognition, Brain, and Language


#### Abstract

Are different amounts of semantic processing associated with different semantic ambiguity effects? Could this explain some discrepant ambiguity effects observed between and across tasks? Armstrong and Plaut (2016) provided an initial set of neural network simulations indicating this is indeed the case. However, their empirical findings using a lexical decision task were not clear-cut. Here, we use improved methods and five different experimental manipulations to slow responding---and the presumed amount of semantic processing---to evaluate their account more rigorously. We also expanded the empirical horizon to another language: Spanish. The results are partially consistent with the predictions of the neural network and differ in several important ways from English data. Potential causes of these discrepancies are discussed in relation to theories of ambiguity resolution and cross-linguistic differences.


Keywords: semantic ambiguity; slow vs. fast lexical decision; semantic settling dynamics, neural networks.

Understanding how the meaning of ambiguous words is resolved is critical because the meaning of most words depends on context (e.g., cricket can refer either to a game or to an insect). Developing an account of ambiguity resolution has, however, been challenged by two complications: 1) the complex and often apparently contradictory effects of ambiguity observed between and sometimes even within a given experimental task, discussed below, and 2) the often inconsistent effects observed for polysemes with related senses (e.g., chicken can refer to an animal or its meat) vs. homonyms with unrelated meanings (e.g., cricket) compared to (relatively) unambiguous control words (e.g., chalk).

Recently, Armstrong and Plaut (2016) reported neural network simulations suggesting that many apparently inconsistent effects can be reconciled as a function of (a) how the number and relatedness of a word's meanings are activated over time, (b) the amount of processing that takes place before a response can be generated in a given task (see Figure 1). This semantic settling dynamics (SSD) account posits that early processing is dominated by excitatory/cooperative neural dynamics that would facilitate the processing of polysemes. In contrast, later processing would be dominated by inhibitory/competitive neural dynamics that would impair the processing of homonyms. Thus, "fast" tasks like typical lexical decision, in which
participants must decide whether a letter string forms a word (e.g., cricket) or not (e.g., blicket), would show a polysemy advantage (e.g., e.g., Armstrong \& Plaut, 2016; Beretta, Fiorentino, \& Poeppel, 2005; Rodd, Gaskell, \& Marslen-Wilson, 2002). In contrast, "slow" tasks like typical semantic categorization, in which participants must determine whether a word refers to a member of a particular category (e.g., does cricket refer to a vegetable?), would show a homonymy disadvantage (e.g., Hino, Pexman, \& Lupker., 2006).

The SSD account offers both a contrasting and a complementary explanation to an account positing that different ambiguity effects are due to task-specific configurations of the decision system (Hino et al., 2006). In contrast to the decision system account, the SSD hypothesis stresses how dynamics within semantics can critically shape the ambiguity effects observed in a given task. The decision system should, however, play an important role in determining "when" sufficient evidence has accumulated to generate a response---and thus which portion of semantics is being tapped (for a broader discussion, see Armstrong \& Plaut, 2016).


-     - Homonymous . . . . Polysemous - Unambiguous

Figure 1. Semantic activity as a function of processing time for homonyms, polysemes, and unambiguous controls in the neural network simulation reported by Armstrong and Plaut (2016). Slices A-D highlight how sampling these trajectories at different time points aligns with different behavioural and neural effects reported in the literature, such as typical lexical decision (Slice A) and semantic categorization (Slice C).

Armstrong and Plaut (2016) also put the SSD account to the test in an empirical setting. In their experiments, the overall task (lexical decision) was held constant. They then
manipulated additional properties of the task to slow responses (manipulations of nonword difficulty and/or the brightness of the letters on the screen). Insofar as these slow-downs enabled additional semantic processing to take place, the SSD account predicts this would lead to a shift from a polysemy advantage in the easy/fast conditions (Figure 1, Slice A) and a homonymy disadvantage in the slow/hard conditions (Figure 1, Slice C).

The results were generally---although not perfectly--consistent with these predictions. A polysemy advantage was typically observed in the easy/fast condition, but evidence for this advantage in the harder conditions was more limited. Similarly, there was evidence that a homonymy disadvantage was present in some (but not all) of the hard/slow conditions, but, critically, not in the easy/fast conditions. One possible interpretation of these results is that they are attributable to a slight increase in semantic processing and thus reflect only a small step along the predicted semantic settling dynamics (e.g., Figure 1, Slice A to Slice B, rather than Slice A to Slice C). Additional investigations are needed, however, to better explore this possibility and the validity of the SSD account more broadly.

The present work is a major extension of Armstrong and Plaut's (2016) initial empirical studies. From a theoretical perspective, it follows the abductive reasoning: if a range of different manipulations designed to slow responding all yield the same changes in ambiguity effects, this will provide broad convergent support for the SSD account. Our work also builds upon past work in several important ways: First, for all but one condition, it uses within-participant manipulations to boost statistical power. Second, the experiments were run in Spanish, a language in which it is easier to control for several potential confounding variables (e.g., with few exceptions, each Spanish letter maps to a single sound and vice versa, so matching word lengths in number of letters also matches word lengths in number of phonemes). Doing so also allows for the evaluation of the robustness of particular ambiguity effects and facilitates the development of general as opposed to Anglocentric theories (Share, 2008). Further, recent Spanish homonym meaning frequency norms (Armstrong et al., 2015) allow us to select homonyms with balanced meaning frequencies. This should boost the competitive dynamics assumed to be associated with homonyms during late processing.

## Behavioral Studies of Lexical Decision

We evaluated whether slowing participants' lexical decision responses using several different manipulations reproduced the different semantic ambiguity effects predicted by the SSD account. If these different manipulations produce the anticipated effects, this would support the notion that the time-point at which the response was made---and the corresponding amount of semantic settling---is a critical component of any theory of semantic ambiguity resolution. (Without denying that these dynamics interact and are further shaped by other systems; e.g., the response system.) If the results do not produce the predicted effects, this would
support claims that qualitative differences in the configuration of the response system, as opposed to semantic settling dynamics, explain many discrepant ambiguity effects.

We applied the following manipulations to a standard visual and/or auditory lexical decision task, which we describe in detail subsequently. The first two manipulations relate closely to those in Armstrong \& Plaut (2016) for comparison purposes, whereas the remaining three have never been used in studies of semantic ambiguity.

1. Visual Lexical Decision: Nonword Wordlikeness: "Easy" nonwords with lower bigram frequencies and higher Orthographic Levenshtein distances (OLD; Yarkoni, Balota, \& Yap, 2008) than the word stimuli were used in the baseline; "Hard" nonwords with higher bigram frequencies and lower OLDs than the words were used in the slowed condition. This was the only between-participant manipulation because previous experiments have found carry-over effects when nonword difficulty is blocked within participants (Armstrong, 2012). All other manipulations were within participants and used easy nonwords to avoid potential ceiling effects on how slow lexical decision can be pushed.
2. Visual Lexical Decision: Visual Noise: Standard text was presented in the baseline; visual noise ( 950 3px dots) was superimposed to degrade the text in the slowed condition. This condition is similar to the contrast reduction manipulation in Armstrong \& Plaut (2016).
3. Intermodal Lexical Decision: Visual lexical decision served as the baseline, auditory lexical decision as the slowed condition. This experiment was motivated by different ambiguity effects observed in audio vs. visual lexical decision in Rodd et al. (2002).
4. Auditory Lexical Decision: Auditory Noise: Clear sound recordings were presented in the baseline; noisy recordings---created by replacing $75 \%$ of the auditory signal with signal-correlated noise---were used in the slowed condition.
5. Auditory Lexical Decision: Compression/Expansion: Recordings were played $30 \%$ faster in the baseline and $30 \%$ slower in the slowed condition. The "similarity" time effect in Goldwave ${ }^{\circledR}$ (v6.13) was used to preserve pitch and the naturalness of the vocalization.
Participants. Each experiment was completed by 42 Spanish native speakers (avg. age $=24$ years, $70 \%$ female). All had normal or corrected-to-normal vision and no history of language or psychological disorders. Participants received a monetary payment. Consent was obtained in accordance with the declaration of Helsinki.

Stimuli. Words. The stimuli filled a $2 \times 2$ factorial design that crossed number of unrelated meanings (NoM: one vs. two) with number of related senses (NoS: few [range: 1-5] vs. many [range: 6-14]), similar to past work (Rodd et al., 2002; Armstrong \& Plaut, 2016). NoM and NoS were based on the number of separate entries vs. sub-entries for each word in the Real Academia Española Spanish dictionary (RAE, 2014). For convenience, we will refer to the four
conditions as (relatively) unambiguous words (NoM: 1, NoS: few), homonyms (NoM: 2, NoS: few), polysemes (NoM: 1, NoS: many) and hybrids (NoM: 2, NoS: many).

To maximize the potential for competition between the interpretations of words with two unrelated meanings, we only included homonyms and hybrids with dominant relative meaning frequencies below 0.82 in the Spanish eDom norms (Armstrong et al., 2015). Using the EsPal Spanish word database (Duchon, Perea, Sebastián-Gallés, Martí, \& Carreiras, 2013), the candidate items were also constrained to have no homophones, be between 4 and 10 letters long, have word frequencies between 0.1 and 50 , and have only noun or verb meanings (all had at least one noun meaning). This database also provided information regarding the word's summed bigram frequency, length in phonemes, and length in syllables.

The SOS stimulus optimization software (Armstrong, Watson \& Plaut, 2012) identified 36 items in each cell of the design that were also matched on a range of psycholinguistic covariates (see Table 1). Finally, we collected separate norms for the imageability and familiarity of the words from two groups of 25 native speakers who did not participate in the main experiments.

Nonwords. Candidate nonwords were generated for each of $\sim 80,000$ words sampled from Espal (Duchon et al, 2013) to match the psycholinguistic properties of the experimental words, except for NoM and NoS. Nonwords were generated via the Wuggy nonword generator using the default settings (Keuleers \& Brysbaert, 2010). In total, 144 "easy" nonwords were sampled to have lower bigram frequency and higher OLD than the words, whereas 144 "hard" nonwords were selected to have a higher bigram frequency and lower OLD than the words.

Table 1. Properties of the Word Stimuli

|  | Unambig. | Polyseme | Homonym | Hybrid |
| ---: | :---: | :---: | :---: | :---: |
| Example | tractor | vaina | pinta | pipa |
| \# Meanings | 1 | 1 | 2.1 | 2.4 |
| \# Senses | 3.2 | 9.8 | 3.3 | 9.0 |
| Word Freq. | 5.3 | 5.5 | 5.0 | 6.3 |
| OLD | 1.9 | 1.8 | 1.8 | 1.5 |
| \# Letters | 6.6 | 6.5 | 6.7 | 6.0 |
| \# Phonemes | 6.6 | 6.3 | 6.6 | 5.9 |
| \# Syllables | 2.8 | 2.8 | 2.9 | 2.6 |
| Familiarity | 4.2 | 4.7 | 4.0 | 4.6 |
| Imageability | 4.3 | 5.1 | 4.5 | 4.9 |
| Dom. Freq. | - | - | 0.5 | 0.5 |

Note. Dom. Freq. = Relative Frequency of dominant meaning.
Table 2. Properties of the Word and Nonword Stimuli

|  | Words | Easy Nonwords | Hard Nonwords |
| ---: | :---: | :---: | :---: |
| Bigram Freq. | 1602 | 445 | 2782 |
| OLD | 2.0 | 2.9 | 1.5 |

Audio Recordings. Audio recordings were produced by a male native speaker. Volume was normalized to half the dynamic range. Auditory stimuli were pre-processed using Audacity (Mazzoni, 2013).

Procedure. The experiments were run on a desktop computer with a CRT monitor using Psychopy (Peirce, 2007). Auditory stimuli were presented over headphones.

Each experiment began with 4 practice trials. Participants then completed four blocks of 72 experimental trials, each of which began with 4 unanalyzed warm-up trials. An equal number of words from each cell of the design were presented in each block. The order of the stimuli was pseudorandom, with the constraint that no more than three words or nonwords could be presented in a row.

Each trial began with blank screen for 250 ms , followed by a fixation cross (+) for 750 ms , which was briefly replaced by a blank screen again for 50 ms before the presentation of the word or nonword. In the visual conditions, text was presented in the center of the screen. In the auditory conditions, the recording was played, instead. Response latency was measured from stimulus onset, and the next trial began automatically after a response. A message was displayed if no response was made within 2500 ms . Participants responded by pressing the left and right control keys with their right and left index fingers. Word responses were always made with the dominant hand. The experiment took about 20 minutes to complete.

## Results

Data screening. Participants and items were screened for outliers using the Mahalanobis Distance Statistic and a critical p-value of .001 . This eliminated no more than two participants in each experiment and no more than two words of any type. Trials with latencies < 200 ms or $>2000 \mathrm{~ms}$ were also discarded ( $0.66 \%$ of trials).
Analytical approach. The analyses reported here focused on the critical effects of homonymy and polysemy relative to unambiguous controls, as well as how these variables were affected by the slowing manipulations. We also report exploratory analyses of the hybrids, which should be affected by both cooperative and competitive dynamics.

All of the word data were analyzed with linear mixedeffect models (Bates, Maechler, Bolker \& Walker, 2015) using R (R Core Team, 2016). The models included the key fixed effects of manipulation (with the faster/easier condition used as the baseline) and item type (with separate contrasts between an unambiguous baseline and homonyms, polysemes, and hybrids). To address potential confounds, the models also included fixed effects of imageability, residual familiarity ${ }^{1}$, log-transformed word frequency, OLD, length in letters, and bigram frequency. All of the aforementioned fixed effects were allowed to interact with the effect of manipulation. Further, to reduce autocorrelation effects from the previous trials (Baayen, \& Milin, 2010), the models included fixed effects of stimulus type repetition, previous trial accuracy, previous trial lexicality, previous trial latency, and trial rank. All continuous variables were centered and normalized. Additionally, the models included random intercepts for

[^409]

Figure 2. Correct latency [left] and accuracy [right] for the experiments. H=homonym, U=unambiguous, $\mathrm{P}=\mathrm{Polysemous}$, $\mathrm{Y}=\mathrm{Hybrid}$. Error bars = SEM.
item and participant. Random slopes were omitted because these models did not always converge. Latency was modeled with a Gaussian distribution, whereas accuracy was modeled with a binomial distribution. Effects were considered significant if $\mathrm{p} \leq .05$, and trends are considered marginal if $\mathrm{p} \leq .15$. All tests were two-tailed.

Correct Latency. The latency data are presented in the left panel of Figure 2. Slowing Manipulations. All five manipulations slowed overall response speed (all $p \mathrm{~s} \leq .02$ ). Homonyms. A main effect indicating a homonymy disadvantage was observed in the intermodal and auditory noise manipulations ( $b=24.0, \mathrm{SE}=10.3, t=2.4, p=.02$ and $b=34.0, \mathrm{SE}=15.0, t=2.3, p=.03$, respectively). The homonymy by slowing manipulation interaction produced a significant increase in the homonymy disadvantage in the slower condition of the auditory compression/expansion experiment ( $b=29.1, \mathrm{SE}=13.3, t=2.2, p=.03$ ). A similar marginal trend was observed in the nonword wordlikeness experiment ( $b=13.2$, $\mathrm{SE}=8.2, t=1.6, p=.11$ ). Polysemes. A main effect indicating a polysemy advantage was only detected in the baseline condition of the nonword wordlikeness manipulation $(b=-19.8, \mathrm{SE}=9.4, t=-2.1, p$ $=.04$ ). The polysemy by slowing manipulation interaction indicated that the polysemy advantage marginally decreased in the visual noise experiment $(b=30.7, \mathrm{SE}=16.3, t=1.9$,
$p$ =.06). Hybrids. There were no significant effects involving hybrids in any experiment. Imageability. There was always a marginal or significant facilitatory main effect of imageability (all $p \mathrm{~s} \leq .06$ ). The imageability by slowing manipulation interaction indicated this effect increased marginally in the slowed conditions of the intermodal ( $b=-$ $6.1, \mathrm{SE}=4.0, t=-1.5, p=.12$ ), visual noise ( $b=-9.3, \mathrm{SE}=$ 5.9, $t=-1.6, p=.12$ ), and audio compression/expansion experiments ( $b=-9.2$, $\mathrm{SE}=4.9, t=-1.9, p=.06$ ).
Accuracy. The accuracy data are presented in the right panel of Figure 2. Slowing Manipulations. The slowing manipulation decreased overall accuracy in the visual noise condition ( $b=-2.1, \mathrm{SE}=0.3, z=-8.2, p<.001$ ), whereas it increased overall accuracy in the audio expansion condition ( $b=2.0, \mathrm{SE}=0.3, z=7.0, p<.001$ ). Homonyms. The homonymy by slowing manipulation interaction in the compression/expansion experiment indicated that there was a marginal decrease in homonym accuracy after slowing ( $b$ $=-0.6, \mathrm{SE}=0.4, z=-1.7, p=.10$ ). Polysemes. A marginal main effect indicating a polysemy advantage was observed in the nonword wordlikeness experiment ( $b=0.4, \mathrm{SE}=0.2$, $z=1.7, p$ <.09). There were also marginal polysemy by slowing manipulation interactions in the nonword wordlikeness and audio compression/expansion experiments, indicating that there was decrease in polyseme
accuracy relative to the unambiguous baseline in the slowed conditions ( $b=-0.6, \mathrm{SE}=0.4, z=-1.8, p=.09$ ). Hybrids. As in the latency data, no significant effects involving the hybrids were observed. Imageability. The facilitatory main effect of imageability was always significant (all $p \mathrm{~s} \leq .02$ ), except for in the case of auditory noise (the model did not converge) and in the audio compression/expansion experiment, where the effect was marginal ( $p=.15$ ). There was a marginal interaction between imageability and the slowing manipulation in the visual noise experiment indicating differentially decreased facilitation after slowing ( $b=-0.2, \mathrm{SE}=0.1, z=-1.6, p=.11$ ), whereas in the compression/ extension experiment $(b=0.2, \mathrm{SE}=0.1, z=$ $1.5, p=.14$ ) there was increased facilitation.
Summary. A significant or marginal homonymy disadvantage, or an increased homonymy disadvantage in the slowed condition, was observed in all but the visual noise experiment. A main effect of polysemy was only detected in one experiment and the polysemy advantage marginally decreased in two experiments. Hybrid items were never significantly different from the unambiguous controls, which is likely due, at least in part, to difficulties matching these rare items on other covariates. The facilitatory effect of imageability was significant or marginal in all experiments. The magnitude of these facilitation effects increased marginally in three experiments (intermodal, visual noise, compression /expansion).

## Discussion

The aim of our study was to evaluate whether a range of different manipulations designed to slow responses would lead to different ambiguity effects, as predicted by the SSD account. At first glance, except for speed-accuracy tradeoffs, virtually all of the effects that were significant or marginal were consistent with the SSD account. Additionally, most of non-significant results showed the predicted trends numerically. Thus, this collective body of work does add some additional support to the notion that processing time---and the presumed amount of semantic settling---plays a role in explaining many ambiguity effects. These results also suggest that some broad ambiguity effects transcend different languages.

Additionally, taking a more critical view of the observed effects promises to reveal additional aspects of how and why discrepant ambiguity effects are observed within and between tasks. To begin, our ideal a priori aim was to reproduce a polysemy advantage only in the easiest/fastest tasks (Figure 1, Slice A) and observe a homonymy disadvantage only in the hardest/fastest tasks (Figure 1, Slice C). The overall pattern of results, however, would appear to be more consistent with the easiest task beginning closer to Slice B, where both a weaker homonymy disadvantage and polysemy advantage are predicted. This result is surprising for several reasons. First, Armstrong and Plaut (2016) went to great lengths to make their lexical task as difficult as possible, and yet their results were consistent with earlier processing dynamics (primarily Figure 1, Slice

A-B). Their overall latencies were also approximately 100 ms faster than in the analogous conditions in the present work. The present work did use words with slightly lower frequencies, but it also used considerably easier nonwords, so there is no clear explanation for this large discrepancy. Further, we have conducted an additional experiment with "very easy" nonwords (nonwords with extremely low bigram frequencies and neighborhood sizes) and still not been able to increase overall performance by a substantial degree. These results are also inconsistent with Jager, Green, \& Cleland's (2016) prediction that a polysemy advantage should be strongest for low frequency words because their meanings overlap more.

Another possibility worth considering is that whereas past research has typically struggled to produce a homonymy disadvantage and had more success in obtaining a polysemy advantage, the present work may have experienced the opposite difficulties. This may be due to having used atypically large set of balanced homonyms. This was accomplished by sampling from a database of subjective meaning frequency norms (Armstrong et al., 2015) and may have differentially boosted the power of the homonymy effects. This more powerful manipulation of homonymy may also have coincided with a less powerful manipulation of polysemy based on the recent results of Fraga, Padrón, Perea, \& Comesaña (2016). They found that although the number of senses provided in a subjective meaning norming study and those available in the RAE dictionary (the source of our polysemy counts) correlated highly, only the subjective norms were significant predictors of latencies in lexical decision and naming tasks. Unfortunately, there was insufficient overlap between our items and theirs to corroborate their findings in our own data. However, this recent observation clearly stresses the importance of how polysemy is measured. In English, several studies have used dictionary counts to predict polysemy successfully (e.g., Armstrong et al., 2016; Rodd et al., 2002 both used counts from Wordsmyth; Parks, 1999). Thus, our findings in Spanish suggests that the lexographers administering the RAE dictionary use a different classification scheme for ambiguity, and/or English and Spanish vary in their distributions of polysemes in ways that shape performance to a substantial degree. The latter possibility gains support from the Armstrong et al. (2015) homonym norming study. They observed that despite Spanish and English having similar total numbers of homonyms, Spanish homonyms are much more likely to have a strongly dominant meaning. (This also posed challenges for us finding balanced and well matched hybrids.) Clearly, a more extensive set of polyseme norms with high external validity must be collected in both languages to evaluate these possibilities.

The prior discussion has focused primarily on potential differences in objective or subjective measures of ambiguity. However, is also possible that broader properties of the language and/or of our participants may have contributed to the aforementioned discrepancies. Our use of

Spanish, an orthographically transparent language, may have been advantageous when controlling for orthographic and phonological confounds. However, it may also have allowed for the rapid spreading of activation between orthography and phonology. This could have, in turn, allowed these representations, as opposed to semantics, to be the primary drivers of the response system. Although the significant effects of imageability indicate that semantics did always influence responses, it is possible that semantic effects may have been attenuated such that only the strong effect of homonymy could be detected.

On a related front, the participants tested by Armstrong and Plaut (2016) were all native English speakers in the USA and presumably had limited exposure to other languages. In contrast, the participant population in the Basque Country is bilingual and all participants reported proficiency in one or more other languages that share at least a partially overlapping phonology and/or orthography (e.g., Basque, French, English). Bilingualism in and of itself has been reported to slow responses in some tasks (e.g., Luo, Luk, \& Bialystok, 2010). These results have typically been explained by focusing on dynamics at the (sub)lexical level, however (e.g., in the Bilingual Interactive Activation model; Dijkstra \& van Heuven, 1998). Our results suggest that some of these differences could also be attributable to processing differences at a semantic level. Consistent with this hypothesis, Taler, Zunini, and Kousaiev (2016) found that monolinguals exhibited greater facilitation as a function of increased NoS than bilinguals in a lexical decision task. This was true both in response latency and in EEG measures of the N400, which is known to index semantic processing. Collectively, these results suggest that semantic settling dynamics and ambiguity resolution could be impacted by knowledge of multiple languages. The field would therefore benefit from additional carefully matched experiments across a broad span of languages.

Returning to the initial question that motived our work, does processing time play a critical role in shaping some ambiguity effects? Our results provide partial support that this is, indeed the case. However, the cases in which such support did not materialize are perhaps just as theoretically relevant. These cases highlight how certain core effects in the semantic ambiguity literature may vary as a function of the language in which the test is conducted, and/or as a function of knowledge of a second language. They also point to important methodological issues that remain to be addressed, such as how to classify and compare polysemy across languages. Taken together, the present work therefore serves not only advances our understanding of the semantic settling dynamics in ambiguity resolution. It also highlights the value of cross linguistic comparisons in developing a general as opposed to a language-specific understanding of semantic ambiguity.

## Acknowledgments

This work was funded by the BCBL's Severo Ochoa Center grant SEV-2015-049, CAPES Foundation BEX grant 1692-13-5 to JM,

NSERC, and the University of Toronto. We thank Asier Zarraga for assistance with data processing, and Arthur Samuel for his tool for generating auditory noise.

## References

Armstrong, B. C. (2012). The temporal dynamics of word comprehension and response selection: Computational and behavioral studies. Doctoral dissertation, Psychology Department, Carnegie Mellon University, Pittsburgh.
Armstrong, B. C., \& Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. Language, Cognition and Neuroscience, 31(7), 940966.

Armstrong, B. C., Watson, C. E., \& Plaut, D. C. (2012). SOS! An algorithm and software for the stochastic optimization of stimuli. Behavior Research Methods, 44(3), 675-705.
Armstrong, B. C., Zugarramurdi, C., Cabana, Á., Lisboa, J. V., \& Plaut, D. C. (2015). Relative meaning frequencies for 578 homonyms in two Spanish dialects: A cross-linguistic extension of the English eDom norms. Behavior Research Methods, 1-13.
Baayen, R.H., Milin, P. (2010). Analyzing Reaction Times. International Journal of Psychological Research, 3(2), 12-28.
Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Beretta, A., Fiorentino, R., \& Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. Cognitive Brain Research, 24(1), 57-65.
Dijkstra, T., \& Van Heuven, W. J. (1998). The BIA model and bilingual word recognition. Localist connectionist approaches to human cognition, 189-225., Mahwah, New Jersey: Lawrence Erlbaum Associates.
Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., \& Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. Behavior Research Methods, 45(4), 1246-1258.
Fraga, I., Padrón, I., Perea, M., \& Comesaña, M. (2016). I saw this somewhere else: The Spanish Ambiguous Words (SAW) database. Lingua. 185, 1-10.
GoldWave (v6.13) [Software]. St. John's, NF: GoldWave ${ }^{\circledR}$ Inc.
Hino, Y., Pexman, P. M., \& Lupker, S. J. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding?. Journal of Memory and Language, 55(2), 247-273.
Jager, B., Green, M. J., \& Cleland, A. A. (2016). Polysemy in the mental lexicon: relatedness and frequency affect representational overlap. Language, Cognition and Neuroscience, 31(3), 425-429.
Keuleers, E., \& Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. Behavior Research Methods, 42(3), 627-633.
Luo, L., Luk, G., \& Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. Cognition, 114(1), 29-41.
Mazzoni, D. (2013). Audacity (Version 2.0.5) [Software].
Parks, R. (1999). Wordsmyth English Dictionary-Theasurus.
Peirce, J. W. (2007). PsychoPy-psychophysics software in Python. Journal of Neuroscience Methods, 162(1), 8-13.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Real Academia Española. (2014). Diccionario de la lengua española (23rd ed.). Madrid, Spain.
Rodd, J., Gaskell, G., \& Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. Journal of Memory and Language, 46(2), 245-266.
Share, D. L. (2008). On the Anglocentricities of current reading research and practice: the perils of overreliance on an" outlier" orthography. Psychological Bulletin, 134(4), 584.
Taler, V., Zunini, R. L., \& Kousaie, S. (2016). Effects of Semantic Richness on Lexical Processing in Monolinguals and Bilinguals. Frontiers in Human Neuroscience, 10.
Yarkoni, T., Balota, D., \& Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. Psychonomic Bulletin \& Review, 15(5), 971-979.

# Inferring Intentional Agents From Violation of Randomness 

Yuan Meng (yuan_meng@berkeley.edu)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Fei Xu (fei_xu@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94720 USA


#### Abstract

Humans have a strong "cognitive compulsion" to infer intentional agents from violation of randomness and such an agency-nonrandomness link emerges early in development. In two studies, we directly quantified, formalized, and compared both ends of this link for the first time. In Experiment 1, two groups of participants viewed the same 256 binary sequences (e.g., AABAAABA) and classified each as generated by agents/non-agents or by nonrandom/random processes. We found a strong correlation between two judgments: sequences viewed as more agentive also tended to be judged as less random. In Experiment 2, another two groups were asked to produce sequences that others might appreciate as agentive or nonrandom. Participant-generated sequences in the two conditions had a substantial overlap, indicating common guiding principles of agency and nonrandomness generation. Taken together, the present studies provide evidence for a shared cognitive basis of agency detection and subjective randomness.


Keywords: agency; subjective randomness; agencynonrandomness link; animate-inanimate distinction

## Introduction

We can accept a certain amount of luck in our explanations, but not too much. The question is, how much?

- Richard Dawkins, The Blind Watchmaker (1996)

When we look at something as delicate and orderly as the eye, it is only natural to believe that such a work of art must be designed by someone, an intentional agent with a purpose in mind. In contrast, it takes "a very large leap of the imagination to think the other way around" (Dawkins, 1996, p.7).

This cognitive compulsion to infer agents from order may have given birth to thousands of religions shared by the vast majority of people on Earth (Barrett, 2000; Keil \& Newman, 2015) and it dates back to early childhood (Friedman, 2001) and infancy (Ma \& Xu, 2013; Ma, Berthiaume, Hoch, \& Xu, under revision; Newman, Keil, Kuhlmeier, \& Wynn, 2010). By 10 and 12.5 months of age respectively, infants appreciate that only agents can create regular visual (e.g., YYRYYRYYR, $Y$ and $R$ stand for yellow and red balls; Ma \& Xu, 2013) and auditory (e.g., TTTCTTTTCTTTCTTTC, T and C stand for tambourine and cowbell sounds; Ma et al., under revision) sequences. Similar appreciation has been found in different tasks. For instance, 12-month-olds expect agents (e.g., a ball) but not inanimate objects (e.g., a perceptually similar ball with eyes) to bring order to a disorderly pile of blocks.

Keil and Newman (2015) used these findings to argue that during the first few months of life, infants observe a bulk of ordering and disordering events together with their causes and come to "associate only agents but not non-agents with many kinds of ordered and nonrandom sequences" (p.132).

Notwithstanding counterexamples such as molecular selfassembly or evolution, this association is often true in our universe where entropy tends to increase over time; in contrast, it takes energy, information, and goal-direction to go against the force of nature, all of which strongly indicate agentive causes even to the youngest humans (see Baillargeon, Scott, \& Bian, 2016, for a review). Based on these experiences and intuitive theories, deviation from randomness often indicates nonrandom generation processes behind the scenes (Griffiths \& Tenenbaum, 2001, 2003, 2004, 2007; Sim \& Xu, 2013; Williams \& Griffiths, 2013). When the specifics are unknown, people often default to agents as the "causal placeholders", thinking that they did it somehow (Saxe, Tenenbaum, \& Carey, 2005; Wu, Muentener, \& Schulz, 2015).

If the above line of reasoning holds true, people's judgment of agency should align with their judgment of nonrandomness-the less random something looks, the more agentive it strikes us as. This prediction might be hinted by Kushnir, Xu, and Wellman's (2010) finding that 20-montholds use violation of random sampling to infer the preference of an agent ("She always picked this toy despite its rarity, so she must really like it."). It is possible that nonrandomness not only indicates the psychological states of an already known agent, but also cues its very existence. Imagine if you see a haystack in the middle of a desert (which apparently violates the surrounding vegetation distribution)-you may well conclude that someone must have been there apart from that she is very fond of hay. Alternatively, as Feldman and Tremoulet $(2008)^{1}$ suggested, the intermediate level of nonrandomness is the strongest cue of agency: "Too simple-a simple periodic noise burst, say-and it's an inanimate source, say, a rotating pulsar. Too complex-a totally patternless sequenceand it's just random electromagnetic interference. To seem intelligent it has to be somewhere in between: patterned, but neither perfectly periodic nor completely chaotic" (p.22).

Empirical evidence is needed to test the two possibilities. To the best of our knowledge, there is no research directly examining the relationship between human intuitions about agency and randomness. The closest work to date (Ma \& Xu, 2013; Ma et al., under revision), for instance, lacks a measure of the regularity of the stimuli. Without quantifying

[^410]subjective agency and subjective nonrandomness, however, it is hard to tell whether or not they go hand in hand. In addition, without an independent measure of regularity, one may find herself trapped in circular reasoning-"What is regularity?" "It's the evidence from which people infer intentional agents." "How do people infer intentional agents?" "From regularity." Also, past studies only looked at a small subset of all possible binary sequences of equal length ( 9 digits in Ma \& Xu, 2013; 12 digits in Ma et al., under revision). Chances are that regular sequences in those studies happened to look both regular to the researchers (which is most likely why they were chosen in the first place) and agentive to the infants. We cannot rule out the possibility that there exist a considerable amount of "irregular" sequences that look like the work of an agent, or reversely, "regular" sequences that only call for an inanimate cause. Therefore, to systematically investigate the agency-nonrandomness link, we need to include a wide range of sequences rather than a selected few.

## Linking agency and randomness

To shed light upon the agency-nonrandomness link, we conducted two experiments. In Experiment 1, one group of participants viewed 256 binary sequences of length 8 (e.g., $\Lambda \Lambda \Gamma \Lambda \Gamma \Lambda \Lambda \Lambda)$ and classified the source of each sequence into agentive or non-agentive entities while another group classified the source of each as random or nonrandom processes ${ }^{2}$. Should agency detection be tightly related to subjective randomness, we would expect a high correlation between ratings from the two groups. Sometimes, we not only need to detect other agents, but also wish to be detected by others. For instance, if we are abducted and locked in a truck, flashing the taillight in a "meaningful" way may attract the police and save our lives. Do we actually have such a good intuition about what kind of sequences others may appreciate as agentive? Is it related to our intuition of what others will view as nonrandom? We looked into these questions in Experiment 2 by asking another two groups of participants to generate an 8 -digit binary sequence that they thought might receive the highest agency or nonrandomness score. We examined whether the mean scores of participant-generated sequences were higher than that of all 256 in Experiment 1, as well as the overlap between sequences generated by the two groups.

If people indeed make similar judgments about agency and nonrandomness, then the question is whether they solve the two problems in a similar way. In the last decade, Griffiths and Tenenbaum (2001, 2003, 2004, 2007) formalized the problem of randomness detection as a statistical inference of the data generation process given the data: that is, when judging if a given sequence $X$ is random, people are comparing the probability that $X$ was generated by a random process ( $P($ random $\mid X)$ ) against the probability that $X$ was generated by a regular process $(P($ regular $\mid X))$. The ratio of these two

[^411]probabilities, or "posterior odds", is given by Bayes' theorem (below is the log-odds form):
\[

$$
\begin{equation*}
\log \frac{P(\text { random } \mid X)}{P(\text { regular } \mid X)}=\log \frac{P(X \mid \text { random })}{P(X \mid \text { regular })}+\log \frac{P(\text { random })}{P(\text { regular })} \tag{1}
\end{equation*}
$$

\]

The subjective randomness of sequence $X$ is defined as the only part that depends on $X$-the log-likelihood ratio:

$$
\begin{equation*}
\operatorname{randomness}(X)=\log \frac{P(X \mid \text { random })}{P(X \mid \text { regular })} \tag{2}
\end{equation*}
$$

If $X$ results from flipping a fair coin or the equivalent, then $P(X \mid$ random $)$ is simply $\frac{1}{2}^{l(X)}(l(X)$ is the length of $X)$-the heart of the problem thus becomes evaluating $P(X \mid$ regular $)$. Griffiths and Tenenbaum $(2003,2004)$ specified $P(X \mid$ regular $)$ using a hidden Markov model (HMM) that associates each symbol $x_{i}$ (e.g., H) in $X$ with a hidden state $z_{i}$ (e.g., repeating H). The probability that $X$ is generated by a certain HHM is obtained by summing over the probability that $X$ is generated by each of all possible states $Z$ under this model:

$$
\begin{equation*}
P(X)=\sum_{Z} P(X, Z) \tag{3}
\end{equation*}
$$

Knowing that each $x_{i}$ solely depends on $z_{i}$ and each $z_{i}$ is determined by $z_{i-1}$, we can rewrite Equation 3 as below:

$$
\begin{equation*}
P(X, Z)=P\left(z_{0}\right) \prod_{i=2}^{n} P\left(z_{i} \mid z_{i-1}\right) \prod_{i=1}^{n} P\left(x_{i} \mid z_{i}\right) \tag{4}
\end{equation*}
$$

In this study, the regular generation process is defined by 22 repeating "motifs" of length 1 (repeating H or T ) to 4 (e.g., repeating HTHH, TTHH, etc.). Each symbol in the 22 motifs corresponds to a hidden state, which amounts to 72 . The prior of each motif is $\alpha^{k}$ ( $k$ is the length of a motif) and the probability of continuing with a motif is $\delta$. Using this HMM, we can estimate the subjective randomness of all 256 sequences.

## Experiment 1: Judging sequences

## Method

Participants Seventy-four participants with a United States IP address took part in Experiment 1 on Amazon Mechanical Turk (http://www.mturk.com/, "MTurk") for a payment of $\$ 3.5$. A past acceptance rate equal to or greater than $93 \%$ was required for participation. 40 participants ( 20 women; mean age $=37.15, S D=12.63$, range: $19-70$ years) were randomly assigned to the agency judgment task and 34 (17 women; mean age $=33.98, S D=10.45$, range: $23-67$ years) to the nonrandomness judgment task. Another 36 were excluded for failing one or both instruction check questions.

Stimuli and procedure To begin, participants read a cover story corresponding to their task:

Agency condition. "Welcome to year 3017! Imagine you are a space rescuer whose job is to search for astronauts lost in deep space. These days spaceships are all equipped with a radio transmitter. To call for help, astronauts can use it
to send out sequences made of two types of radio wavesLambda ( $\Lambda$ ) and Gamma ( $\Gamma$ ). However, both types of waves may also be produced by natural phenomena such as celestial body activities-in this case, $\Lambda$ and $\Gamma$ are equally likely to appear. Everyone in space knows that by current technical standard, radio receivers can only pick up 8 waves in a rowthat is, you can only detect sequences that have 8 waves (as mentioned before, each wave is either $\Lambda$ or $\Gamma$ ). If you receive a sequence and think it was produced by natural phenomena, you will ignore it and stay on course. If you think it came from humans, then you will go to them. In this study, you will see about 250 sequences. Your task is to decide whether each sequence was generated by a natural phenomenon or by a human astronaut. Please answer as quickly and accurately as possible!"

Nonrandomness condition. "You're about to see some sequences made of 8 symbols. Each symbol is either Lambda ( $\Lambda$ ) or Gamma ( $\Gamma$ )—which may represent heads or tails, even or odd digits, successes or failures, or other event outcomes. For instance, the sequence $\Lambda Г \Gamma \Lambda$ could stand for "tails, heads, heads, tails", "even, odd, odd, even", etc. Some of these sequences were created by tossing an actual fair coin, which means they are random series-in this case, $\Lambda$ and $\Gamma$ are equally likely to appear. However, other sequences may be generated by nonrandom processes, such as computer programs, successes and losses of a basketball team, and so on. In this study, you will see about 250 sequences. Your task is to decide whether each sequence was generated by a random process or by a nonrandom process. Please answer as quickly and accurately as possible!"

Two quizzes immediately followed to test participants on the sources of sequences as well as the chance of two symbols appearing in the natural phenomenon or the random processs scenario. Images (pixel resolution: $700 \times 525$ ) of 256 binary sequences of length 8 were then displayed on the screen. In the agency judgment task, participants were asked to classify the source of each sequence into "human astronaut" or "natural phenomenon", and in the nonrandomness judgment task, into "nonrandom process" or "random process". The display order was randomized and the relative location of choices counterbalanced between participants. The whole process was self-paced and took an average of 20 minutes.

## Results

An alpha level of .05 was used for all statistical analyses. Each sequence received an agency score (the proportion of participants classifying its source as "human astronaut") as well as a nonrandomness score (the proportion of participants classifying its source as "nonrandom process").

To examine the stability of participants' judgments, we looked at the scores of 128 sequences (e.g., $\Lambda \Lambda Г Г \Lambda \Lambda Г \Lambda$ ) and their complements (e.g., $Г Г \Lambda \Lambda Г Г \Lambda \Gamma)$-in theory, they should be the same. Indeed, the sequence-complement correlation was high in the agency condition, $r(126)=.84,95 \% \mathrm{CI}$ [.78, .89], $p<.001$, adjusted $R^{2}=.71$, and even higher in the nonrandomness condition, $r(126)=.93,95 \% \mathrm{CI}[.91, .95], p$


Figure 1: The agency-nonrandomness link in Experiment 1.
$<.001$, adjusted $R^{2}=.87$, suggesting that participants were making reliable judgments during the task.

Of central interest to Experiment 1 was the correlation between agency and nonrandomness scores. We found a strong positive correlation between the two, $r(254)=.84,95 \%$ CI [.81, .88], $p<.001$, adjusted $R^{2}=.72$ (see Figure 1a).

To see whether detecting agents was a similar problem to detecting deviation from randomness, we tested if the same model (as specified earlier) fit data from both tasks reasonably well. Since our models predicted subjective randomness, we recoded agency scores into non-agency scores (1-agency score) and nonrandomness scores into randomness scores (1 - nonrandomness score). Fitting the non-agency data gave $\delta$ $=.44$ and $\alpha=.11$, with correlation $r(254)=.72, p<.001$, and fitting the randomness data gave $\delta=.45, \alpha=.12$, with correlation $r(254)=.75, p<.001$. Model predictions in the two conditions were strongly correlated, $r(254)=.96,95 \%$ CI [.95, .97],$p<.001$, adjusted $R^{2}=.92$ (see Figure 1b).

## Discussion

In Experiment 1, we found a strong correlation between participants' agency and nonrandomness judgments-that is, sequences viewed as less random were rated as more agentive. A hidden Markov model fit human data in both tasks reasonably well and produced highly overlapping predictions for subjective non-agency and subjective randomness, which suggests that people are solving highly similar problems when detecting agents and violation of randomness.

At the beginning of this paper, we discussed two possible forms the agency-nonrandomness link. Feldman and Tremoulet (2008) suggested that mid-level randomness is most agentive, according to which we should find a reverse $U$ -shaped/U-shaped relation between the degree of agency/nonagency and the degree of randomness. However, this was neither the case in human judgments (Figure 1a) nor model predictions (Figure 1b). Instead, the findings in Experiment 1 provided evidence for a linear agency-nonrandomness link.

## Experiment 2: Generating sequences

Participants A total of 212 participants with a United States IP address who did not participate before took part in Experiment 2 on MTurk for a payment of $\$ 0.5$. A past acceptance rate equal to or greater than $93 \%$ was required for participation. 105 participants ( 38 women; mean age $=35.59$, $S D=10.54$, range: $18-67$ years) were randomly assigned to the agency generation task and 107 ( 34 women, 1 other; mean age $=36.36, S D=11.63$, range: $19-65$ years) to the nonrandomness generation task. Another 78 were excluded for failing one or both instruction check questions or generating sequences that were not binary or of length 8 .

## Method

Stimuli and procedure To begin, participants read a cover story corresponding to their task:

Agency condition. "Welcome to year 3017! Imagine you are an astronaut who is lost in deep space after an accident. These days spaceships are all equipped with a radio transmitter. To call for help, you can use it to send out sequences made of two types of radio waves-Lambda ( $\Lambda$ ) and Gamma $(\Gamma)$. However, both types of waves may also be produced by natural phenomena such as celestial body activities-in this case, $\Lambda$ and $\Gamma$ are equally likely to appear. Everyone in space knows that by current technical standard, radio receivers can only pick up 8 waves in a row-that is, you should only send sequences that have 8 waves (as mentioned before, each wave is either $\Lambda$ or $\Gamma$ ). If space rescuers receive your sequence and thinks it was produced by natural phenomena, they will ignore it and stay on course. If they think it came from humans, then they will come to you.I $n$ order to save yourself, what sequence will you send to space?"

Nonrandomness condition. "You're about to write down a sequence made of 8 symbols. Each symbol is either Lambda ( $\Lambda$ ) or Gamma ( $\Gamma$ )—which may represent heads or tails, even or odd digits, successes or failures, or other event outcomes.

For instance, the sequence $\Lambda \Gamma \Gamma \Lambda$ could stand for 'tails, heads, heads, tails', 'even, odd, odd, even', etc. Sequences like this could be created by tossing an actual fair coin, which means they are random series. However, they could also be generated by nonrandom processes, such as computer programs, successes and losses of a basketball team, and so on. In this study, you will come up with one sequence. Your task is to make it look least random-that is, this sequence should NOT look like the product of a random process; instead, it should look like it's generated by a nonrandom process. In order to fulfill your task, what sequence will you write down?"

Two quizzes immediately followed to test participants on what kind of sequence they should generate and the chance of two symbols appearing in the natural phenomenon or the random processs scenario. Then they entered a sequence in 8 text entry cells (the task requirement was visible). The whole process was self-paced and took about 2-3 minutes.

## Results

Since symbols have no inherent meanings, sequences (e.g., $\Lambda Г \Lambda \Lambda Г \Gamma \Lambda \Gamma)$ and their complements (e.g., $Г \Lambda Г Г \Lambda \Lambda Г \Lambda)$ were coded as the same form (e.g., ABAABBAB). Participants in the agency condition generated 31 unique sequences while those in the nonrandomness condition generated a total of 35 . Table 1 summarized unique sequences generated in both conditions ( 16 overlapping sequences are marked in yellow).

To begin, we looked at whether participants were able to generate good sequences with respect to the task requirement. To do so, we assigned agency and nonrandomness scores in Experiment 1 to participant-generated sequences in Experiment 2 (e.g., the score of ABAABBAB would be the mean score of $\Lambda \Gamma \Lambda \Lambda \Gamma \Gamma \Lambda \Gamma$ and $Г \Lambda Г Г \Lambda \Lambda Г \Lambda)$. In the agency condition, participant-generated sequences had higher scores ( $M$ $=.56, S D=.19)$ compared to all 256 sequences $(M=.43, S D$ $=.14)$, mean difference $=.13,95 \% \mathrm{CI}[.08, .16], t(154.56)=$ $6.58, p<.001, d=.86$. Participant-generated sequences ( $M$ $=.62, S D=.22$ ) also received higher nonrandomness scores than that of the whole set $(M=.50, S D=.23)$, mean difference $=.12,95 \%$ CI [.09, .17], $t(286.11)=5.87, p<.001, d=$ .58. In both the agency and the nonrandomness condition, the frequency of sequences being generated was positively correlated with their scores, $r(103)=.75,95 \%$ CI [.66, .83], $p<$ .001 , adjusted $R^{2}=.56, r(105)=.53,95 \%$ CI $[.38, .65], p<$ .001 , adjusted $R^{2}=.27$, respectively.

To see if the agency-nonrandomness link exists in sequence generation, we examined the overlap of participantgenerated sequences in the two conditions. First, 87 out of 105 "agentive" sequences were also generated in the nonrandomness condition and a similarly large proprotion of "nonrandom" sequences- 83 out of 107 -were found in the agency condition as well. Among the 31 unique agentive sequences and 35 unique nonrandom sequences, 16 were shared by both. The question is whether this overlap was due to chance, a problem that is often faced by bioinformatics scientists when deciding given a genome with $N$ genes, whether one gene list with $a$ genes overlaps with another with $b$ genes

Table 1: Participant-generated sequences in Experiment 2.

| nonrandom | freq. | score | agentive | freq. | score |
| :---: | :---: | :---: | :---: | :---: | :---: |
| ABABABAB | 21 | 0.60 | AAAAAAAA | 40 | 0.76 |
| AAAAAAAA | 17 | 0.90 | ABABABAB | 20 | 0.40 |
| AAAABBBB | 15 | 0.85 | AAAAAAAB | 3 | 0.65 |
| AABBAABB | 14 | 0.56 | AAABBBAA | 3 | 0.50 |
| AAAAAAAB | 3 | 0.85 | AABABBAB | 3 | 0.29 |
| AAABAAAB | 2 | 0.73 | ABBAABBA | 3 | 0.37 |
| AAABAABA | 2 | 0.53 | AAAABBBB | 2 | 0.56 |
| AAABBBAA | 2 | 0.73 | AAABAAAB | 2 | 0.65 |
| AABBABAB | 2 | 0.25 | AAABAABA | 2 | 0.54 |
| AABBBAAB | 2 | 0.40 | AAABBAAA | 2 | 0.65 |
| ABABAABB | 2 | 0.26 | AAABBBAB | 2 | 0.35 |
| ABBABBAB | 2 | 0.49 | AABBAABB | 2 | 0.47 |
| AAAABAAB | 1 | 0.60 | ABAABBAA | 2 | 0.34 |
| AAABABBA | 1 | 0.21 | ABABBABA | 2 | 0.25 |
| AAABBAAA | 1 | 0.73 | AAAAAABB | 1 | 0.60 |
| AAABBABA | 1 | 0.33 | AAAAABAA | 1 | 0.63 |
| AAABBBAB | 1 | 0.48 | AAAABBAA | 1 | 0.57 |
| AABAAABA | 1 | 0.51 | AAABABAB | 1 | 0.32 |
| AABAABAA | 1 | 0.75 | AABAABAA | 1 | 0.56 |
| AABAABBB | 1 | 0.30 | AABBBAAA | 1 | 0.51 |
| AABABBAB | 1 | 0.23 | AABBBABA | 1 | 0.34 |
| AABABBBA | 1 | 0.23 | AABBBBBB | 1 | 0.57 |
| AABBAAAB | 1 | 0.39 | ABAABAAA | 1 | 0.41 |
| AABBABBA | 1 | 0.33 | ABAABAAB | 1 | 0.43 |
| AABBBAAA | 1 | 0.69 | ABAABABB | 1 | 0.28 |
| ABAAABAB | 1 | 0.33 | ABAABBAB | 1 | 0.24 |
| ABAABBAA | 1 | 0.39 | ABABBBAA | 1 | 0.28 |
| ABAABBBA | 1 | 0.24 | ABBAAABA | 1 | 0.29 |
| ABBABAAB | 1 | 0.19 | ABBAAABB | 1 | 0.34 |
| ABBABBAA | 1 | 0.30 | ABBBAABB | 1 | 0.31 |
| ABBABBBA | 1 | 0.40 | ABBBBBBA | 1 | 0.66 |
| ABBBAABA | 1 | 0.20 |  |  |  |
| ABBBAABB | 1 | 0.35 |  |  |  |
| ABBBABAA | 1 | 0.24 |  |  |  |
| AbBbBBBA | 1 | 0.90 |  |  |  |

if they have an intersection of $t$ genes. In our study, the "genome" was all 128 unique sequences while participantgenerated agentive and nonrandom sequences were the two "gene lists". Using Fisher's exact test implemented by the GeneOverlap R package (Version 1.12.0; Shen \& Sinai, 2013), we found that the 16 -sequence overlap between two conditions was unlikely to arise by mere chance, $p<.001$.

## Discussion

Participants in Experiment 2 showed a good sense of what sequences may strike others as agentive or nonrandom: in both the agency and the nonrandomness condition, they generated sequences with higher scores than all 256 sequences. Crucially, we found a statistically meaningful overlap between agentive and nonrandom sequences, indicating that people not only make similar judgments about agency and nonran-
domness, but may also be guide by similar intuitions when generating stimuli that are agentive or nonrandom.

## General Discussion

The present studies provide evidence for a shared cognitive basis of agency detection and subjective randomness. In Experiment 1, participants made similar judgments about agency and nonrandomness: sequences viewed as more agentive also tended to be judged as less random. A hidden Markov model with 72 states and 22 motifs fitted human performance in both tasks equally well and produced identical predictions regarding the degree to which people should view each sequence as agentive or nonrandom. In Experiment 2, participants did a good job generating sequences that others might see as agentive or nonrandom. Sequences in these two conditions had a substantial overlap, indicating common guiding principles of agency and nonrandomness generation.

Our work contributes to a growing body of literature on the perceived link between order and agency (e.g., Barrett, 2000; Friedman, 2001; Ma \& Xu, 2013; Ma et al., under revision; Newman et al., 2010) by directly quantifying, formalizing, and comparing both ends of this link for the first time.

As Williams and Griffiths (2013) pointed out, randomness judgments often boil down to relative frequency (Are two equally likely events equally frequent?) and sequential dependence (Do earlier events influence subsequent events?). However, past researchers studied the link between these two apsects of randomness (or lack thereof) and agency sepa-rately-for instance, Kushnir et al. (2010) focused on the former while Ma and Xu (2013) the latter. In our study, both biased frequency (e.g., AAAAAAAB) and high sequential dependence (e.g., AAABBAAA) lead to high agency ratings, which may help unify past findings. Future work will look at how differently or similarly frequency and dependence are weighted in our nonrandomness and agency judgments.

Although violation of randomness plays an important role in agency detection, we do not claim that it is the only cue to agency or always linked to the latter. Given certain background knowledge or context ${ }^{3}$, "irregular" stimuli may look agentive-a seemingly chaotic drip painting still looks like the work of a human, and "regular" stimuli can appear non-agentive-we would not necessarily think a mechanical watch is alive because it tick-tocks every second. Without such information, however, intentional agents are perhaps the best guess for a nonrandom outcome. Follow-up studies should take a closer look at how rich knowledge may be integrated into the way we reason about agents and randomness.

Another question is whether the current findings generalize to other types of stimuli, such as matrices, numbers, geometric shapes, etc. Answering this question requires us to understand and formalize randomness in these domains, on which there are increasingly more studies in recent years (e.g., Griffiths \& Tenenbaum, 2007; Hsu, Griffiths, \& Schreiber, 2010). We plan to investigate the agency-nonrandomness link us-

[^412]ing new types of stimuli. Even for binary sequences, it is worth looking at if what we found applies to longer sequences where more interesting regularities may emerge, such as the repeating triads in Ma and Xu's (2013) study. Also, as the length extends, it may become difficult to find a global pattern and one may have to focus on local patterns. How will these factor into our agency and nonrandomness judgments?

On the computational level, our study is the first step towards formalizing how people perceive agency. A more precise account requires us to directly estimate the probability distribution of a certain stimulus being generated by an agent, $P(X \mid$ agent $)$, which can be achieved by using a much larger sample size (e.g., the Big Bell Test invited more than 100,000 people from all over the world to generate random responses; see http://thebigbelltest.org/ for details) as well as applying sampling methods such as Markov chain Monte Carlo with People ("MCMCP", Sanborn, Griffiths, \& Shiffrin, 2010).

MCMCP is also able to capture each person's judgment. What looks agentive to some may look inanimate to others; understanding individual differences may allow us to appreciate the complexity of human agency perception and on top of that, explain far-reaching psychological and societal consequences, such as the endorsement of Intelligent Design, the denial of natural selection in favor of creationism, and so on. In regards to other social phenomena, past studies explored the relationship between randomness and perceived efficacy of rituals (Legare \& Souza, 2014), belief in conspiracy theories (Dieguez, Wagner-Egger, \& Gauvrit, 2015), etc.. We wonder whether people's agency intuition plays a similar or a different role in these situations, especially using the more intricate characterization of agency that MCMCP provides.

In his book Scienceblind, Shtulman (2017) argued that many misconceptions of science stick not just because of ideology or the media, but also because they have deep roots in our intuitive theories. As mentioned before, our agencynonrandomness link may be one such root. For science educators, the question at hand is, how malleable is it? Given the right kind and amount of evidence, will we update our belief? For instance, by understanding causal mechanisms by which nonrandomness could arise from non-agentive sources (e.g., a ball rolling down a xylophone produces orderly sounds, but it is not viewed as an agent by adults or even infants, Schachner, Carey, \& Kelemen, 2013), can we weaken or break this link?

## Acknowledgments

We thank Shaun O'Grady for suggesting a critical change in the stimuli and Fred Callaway for helpful discussions.

## References

Baillargeon, R., Scott, R., \& Bian, L. (2016). Psychological reasoning in infancy. Annual Review of Psychology, 67, 159-186.
Barrett, J. L. (2000). Exploring the natural foundations of religion. Trends in Cognitive Sciences, 4(1), 29-34.
Dawkins, R. (1996). The blind watchmaker: Why the evidence of evolution reveals a universe without design. New York, NY: W. W. Norton \& Company.

Dieguez, S., Wagner-Egger, P., \& Gauvrit, N. (2015). Nothing happens by accident, or does it? A low prior for randomness does
not explain belief in conspiracy theories. Psychological Science, 26(11), 1762-1770.
Feldman, J., \& Tremoulet, P. D. (2008). The attribution of mental architecture from motion: Towards a computational theory (Tech. Rep. No. RuCCS TR-87). Department of Psychology, Rutgers University.
Friedman, W. J. (2001). The development of an intuitive understanding of entropy. Child Development, 72(2), 460-473.
Griffiths, T. L., \& Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. In J. D. Moore \& K. Stenning (Eds.), Proceedings of the 23rd annual conference of the cognitive science society (pp. 370-375). Austin, TX: Cognitive Science Society.
Griffiths, T. L., \& Tenenbaum, J. B. (2003). Probability, algorithmic complexity and subjective randomness. In R. Alterman \& D. Kirsh (Eds.), Proceedings of the 25th annual conference of the cognitive science society (pp. 480-485). Austin, TX: Cognitive Science Society.
Griffiths, T. L., \& Tenenbaum, J. B. (2004). From algorithmic to subjective randomness. In S. Thrun, L. K. Saul, \& B. Schölkopf (Eds.), Advances in neural information processing systems (Vol. 16, pp. 953-960). Cambridge, MA: MIT Press.
Griffiths, T. L., \& Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. Cognition, 103(2), 180-226.
Hsu, A., Griffiths, T. L., \& Schreiber, E. (2010). Subjective randomness and natural scene statistics. Psychonomic Bulletin \& Review, 17, 624-629.
Keil, F. C., \& Newman, G. E. (2015). Order, order everywhere, and only an agent to think: The cognitive compulsion to infer intentional agents. Mind and Language, 30(2), 117-139.
Kushnir, T., Xu, F., \& Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. Psychological Science, 21(8), 1134-1140.
Legare, C. H., \& Souza, A. L. (2014). Searching for control: Priming randomness increases the evaluation of ritual efficacy. Cognitive Science, 38(1), 152-161.
Ma, L., Berthiaume, V. G., Hoch, J., \& Xu, F. (under revision). Do Infants Infer Intentional Agents from the Perception of Auditory Regularity?
Ma, L., \& Xu, F. (2013). Preverbal infants infer intentional agents from the perception of regularity. Developmental Psychology, 49(7), 1330-1337.
Newman, G. E., Keil, F. C., Kuhlmeier, V. A., \& Wynn, K. (2010). Early understandings of the link between agents and order. Proceedings of the National Academy of Sciences, 107(40), 1714017145.

Sanborn, A., Griffiths, T. L., \& Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. Cognitive Psychology, 60(2), 63-106.
Saxe, R. J., Tenenbaum, J. B., \& Carey, S. (2005). Secret agents: Inferences about hidden causes by 10 - and 12 -month-old infants. Psychological Science, 16(12), 995-1001.
Schachner, A., Carey, S., \& Kelemen, D. (2013). Inferring the causes of patterned sounds: Were those notes caused by an agent, or an inanimate force? In The 8th biennial meeting of the cognitive development society. Memphis, TN.
Shen, L., \& Sinai, M. (2013). Geneoverlap: Test and visualize gene overlaps [Computer software manual]. Retrieved from http://shenlab-sinai.github.io/shenlab-sinai/ (R package version 1.12.0)
Shtulman, A. (2017). Scienceblind: Why our intuitive theories about the world are so often wrong. New York, NY: Basic Books.
Sim, Z., \& Xu, F. (2013). Infants' early understanding of coincidences. In M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), Proceedings of the 35th annual conference of the cognitive science society (pp. 3402-3407). Austin, TX: Cognitive Science Society.
Williams, J. J., \& Griffiths, T. L. (2013). Why are people bad at detecting randomness? A statistical argument. Journal of Experimental Psychology: Learning, memory, and cognition, 39(5), 1473-1490.
Wu, Y., Muentener, P., \& Schulz, L. E. (2015). The invisible hand: Toddlers connect probabilistic events with agentive causes. Cognitive Science, 1(23), 1-23.

# Does Presentation Format Modulate Adults'Automatic Processing of Proportions? 

Rui Meng (rmeng5@wisc.edu)<br>Department of Educational Psychology, 1025 W. Johnson Street Madison, WI 53706 USA<br>Percival G. Matthews (pmatthews@wisc.edu)<br>Department of Educational Psychology, 1025 W. Johnson Street<br>Madison, WI 53706 USA


#### Abstract

Whereas much is known about how humans categorize and reason based on absolute quantities, research investigating the processing of relative quantities, such as proportions, is comparatively limited. The current study used a Stroop-like paradigm to examine adults' automatic processing of nonsymbolic proportions and how presentation formats modulate this processing. Participants were asked to compare individual components across proportions in six different presentation formats. Congruity between component size and overall proportion affected accuracy of comparison, such that participants were less accurate when proportion (the irrelevant dimension) was incongruent with absolute quantity (the relevant) dimension. Moreover, the congruity effect was modulated by the presentation format. These findings serve as evidence that humans automatically access relative quantity when presented in nonsymbolic formats and provide evidence that the strength of this processing is modulated by the format of presentation.


Keywords: automatic processing; congruity effect; relative quantity; proportions; presentation format

## Introduction

Humans share with many species a non-verbal system to estimate absolute quantity (Dehaene, 1997). The invention of number symbols allows humans to precisely represent absolute quantity instead of mere approximate estimation. However, simple absolute quantification is often not sufficient to guide behavior. We frequently need to relate two quantities to generate a new construct: proportion or ratio. Although much is known about the processing of absolute quantity (either symbolic or nonsymbolic), comparatively little is known about how the brain encodes relative quantity.

To represent relative quantity accurately, humans exploit their symbolic numerical competence by using number fractions. However, children and adults often experience great challenges and difficulties in learning and using fractions (Ni \& Zhou, 2005). Furthermore, research on symbolic fractions suggest that the numerical magnitudes represented by symbolic fractions are not automatically activated (Kallai and Tzelgov, 2009), and that the holistic processing of symbolic fractions depends on the stimuli and task contexts (Meert et al., 2009; Meert et al., 2010; Schneider \& Siegler, 2010). For example, Meert and colleagues (2009) observed that access to the magnitude of
symbolic fractions was affected by the congruity or incongruity between the value of the single components and the value of whole fraction. Schneider and Siegler (2010) found that adults process fraction magnitudes holistically when the task does not allow them to use any shortcut strategies that would enable separate processing of the numerator and denominator magnitudes.

Similar to absolute quantity, which can be judged approximately without symbols, proportion (relative quantity) can also be determined non-verbally. Studies have suggested that even by a young age, humans can understand proportion information when presented nonsymbolically (McCrink \& Wynn, 2007; Jacob, Vallentin, \& Nieder, 2012; Matthews, Lewis, \& Hubbard, 2015). For instance, infants can discriminate between two ratios long before the concept of proportionality is introduced during formal schooling (McCrink \& Wynn, 2007).

Research even suggests that the magnitudes of proportions are automatically activated (Duffy, Huttenlocher, Levine, 2005; Duffy, Huttenlocher, Levine, \& Duffy, 2005; Fabbri et al., 2012; Yang et al., 2015; Matthews \& Lewis 2016). Six-month-olds dishabituated when the relation between a dowel and its container changed but not when the absolute size of both object changed while the relation was held constant (Duffy, Huttenlocher, Levine, \& Duffy, 2005). Moreover, 4-yearolds chose the dowel which had the same dowel-container relation as the original display rather than the one with the same absolute dowel size (Duffy, Huttenlocher, Levine, 2005). Fabbri et al. (2012) found that the magnitude of proportions can be automatically and holistically processed by adults using a congruity manipulation in which the greater numerosity of white dots co-occurred with a lower proportion. Yang et al. (2015) found that proportion interfered with preschool children's area comparison performance.

It is acknowledged that the cognitive processes involved in proportion processing vary depending on the type of proportional relation involved (i.e., part-to-part vs. part-towhole) (Sophian \& Wood, 1997; Spinillo \& Bryant, 1999; Möhring, Newcombe, Levine, \& Frick, 2016), and display types (i.e., continuous, discretized, discrete) (Spinillo \& Bryant, 1999; Jeong, Levine, \& Huttenlocher, 2007; Boyer, Levine, \& Huttenlocher, 2008).

Proportions can be presented as either part-to-part relations or part-to-whole relations. Previous research suggested that part-to-part presentation is easier for 6- to 8-year-old children (Spinillo \& Bryant, 1999). However, other study provided evidence that children performed better for problems involving part-to-whole presentation (Sophian \& Wood, 1997; Möhring, Newcombe, Levine, \& Frick, 2016). There is no evidence yet for if and how adults would perform differently for these two relations presentation.

Proportions can also be displayed as continuous, discrete, or discretized (see Figure 1). Previous studies generally agreed that continuous display encourages perceptual approximate measurement of the intensive quantity, while discrete (and discretized) display would lead to exact counting strategy (e.g., Boyer, Levine, \& Huttenlocher, 2008; DeWolf, Bassok, \& Holyoak, 2015). This was underlined by findings that children showed greater and earlier success in judging proportions displayed as continuous quantities than in judging proportions displayed as discrete quantities even if other variables were controlled to be constant (Spinillo \& Bryant, 1999; Jeong, Levine, \& Huttenlocher, 2007; Boyer, Levine, \& Huttenlocher, 2008). For adults, we do not know yet whether they are still influenced by the display format.

These results suggest that presentation format might influence proportions processing. Fabbri and Yang's results are actually different as for the level of automaticity. It is probably due to the fact that they used different presentation format. Fabbri et al. (2012) used arrays of dots and part-topart proportion judgment, while Yang et al., (2015) asked participants to compare the areas of two sectors that were designed in part-to-whole relation. Therefore, the present study aims to systematically investigate how the level of automaticity will change for different presentation formats.

The current study used a Stroop-like paradigm to examine the processing level of proportions. In a Stroop-like task, participants are asked to make judgments on one dimension while there are other dimensions that may agree or conflict with the one to be judged. Participants' performance can suggest the automatic activation of the irrelevant dimensions. Higher error rate and longer reaction times will be observed for incongruent trials than for congruent trials if the irrelevant dimensions are accessed automatically. For example, people tend to spend more time and make more errors when they are asked to compare the magnitudes of two numbers that have incongruent physical sizes than the pairs that have congruent physical size with corresponding magnitudes (Henik \& Tzelgov, 1982).

In the present study, absolute quantity will be treated as the relevant dimension and relative quantity (proportion) as the irrelevant dimension. For congruent trials, the larger proportion also has the larger components. For incongruent trials, the larger proportion would have the smaller components. If participants' performance is worse in the incongruent condition, it would provide evidence that the representation about proportions is automatically activated.

We aim to investigate whether presentation format of proportions would be automatically activated in different levels and thus have different effects to the absolute quantity comparisons. The size of interference in Strooplike tasks is proposed to be a function of degree of the irrelevant dimension's automaticity (MacLeod \& Dunbar, 1988). Therefore, we focus on the size of interference to see whether the automatic accessing level of proportions will differ.

## Method

## Participants

33 undergraduate students from a large Midwestern university participated for course credit (31 females; ages 18 -22).

## Stimuli

Presentation formats were designed to be all possible combinations of display types and relation types.

Three display types were designed: continuous, discrete, and discretized (see Figure 1). The discrete items were arrays of white and black squares with width of 20 pixels. The discretized items were displays composed of these squares stacked to form line segments, except that they were lined together with 1 pixel distance between them. The continuous items were identical to the discretized displays except that there was no space in between.

We examined both part-to-part and part-to-whole relations. For all three displays, we varied the presentation such that half of the proportions were presented in part-topart relation, and half were presented in part-to-whole relation. The part-to-part relation was defined as the white portion to the black portion; the part-to-whole relation was defined as the white portion to the total portion.

Each proportion was presented in one of the six presentation formats depicted in Figure 1 below.


Figure 1: Example of six presentation formats used in the experiment, all represent proportion of $1 / 3$.

Each stimulus pair consisted of two proportions displayed side by side. The center-to-center distance between the two proportions was 800 pixels.

There were two conditions differing by congruity. In congruent pairs, the stimulus which had the larger white portion also had a larger proportion value of white portion relative to either black portion or total (white plus black) portion. In incongruent pairs, the stimulus which had larger white portion had a smaller proportion value of white portion relative to either black portion or total (white plus black) portion. Table 1 showed all the stimuli used.

Table 1 Stimuli used in the present study

|  | Prop. 1 | Prop. 2 | Absolute <br> Distance | Relative <br> Distance |
| :--- | :--- | :--- | :--- | :--- |
|  | $2 / 8$ | $3 / 9$ | 1 | $1 / 12$ |
|  | $1 / 4$ | $2 / 6$ | 1 | $1 / 12$ |
|  | $1 / 6$ | $2 / 8$ | 1 | $1 / 12$ |
|  | $2 / 3$ | $3 / 4$ | 1 | $1 / 12$ |
|  | $4 / 12$ | $5 / 12$ | 1 | $1 / 12$ |
|  | $7 / 12$ | $8 / 12$ | 1 | $1 / 12$ |
|  | $2 / 12$ | $3 / 9$ | 1 | $1 / 6$ |
|  | $1 / 6$ | $2 / 6$ | 1 | $1 / 6$ |
| congruent | $1 / 12$ | $2 / 8$ | 1 | $1 / 6$ |
|  | $3 / 9$ | $4 / 8$ | 1 | $1 / 6$ |
|  | $4 / 6$ | $5 / 6$ | 1 | $1 / 6$ |
|  | $5 / 10$ | $6 / 9$ | 1 | $1 / 6$ |
|  | $4 / 8$ | $6 / 9$ | 2 | $1 / 6$ |
|  | $2 / 4$ | $4 / 6$ | 2 | $1 / 6$ |
|  | $1 / 6$ | $3 / 9$ | 2 | $1 / 6$ |
|  | $3 / 9$ | $5 / 10$ | 2 | $1 / 6$ |
|  | $2 / 6$ | $4 / 8$ | 2 | $1 / 6$ |
|  | $1 / 3$ | $3 / 6$ | 2 | $1 / 6$ |
|  | $1 / 3$ | $2 / 8$ | 1 | $1 / 12$ |
|  | $2 / 6$ | $3 / 12$ | 1 | $1 / 12$ |
|  | $1 / 4$ | $2 / 12$ | 1 | $1 / 12$ |
|  | $3 / 4$ | $4 / 6$ | 1 | $1 / 12$ |
|  | $5 / 6$ | $6 / 8$ | 1 | $1 / 12$ |
|  | $6 / 9$ | $7 / 12$ | 1 | $1 / 12$ |
|  | $6 / 8$ | $7 / 12$ | 1 | $1 / 6$ |
| $1 / 3$ | $2 / 12$ | 1 | $1 / 6$ |  |
| incongruent | $5 / 6$ | $6 / 9$ | 1 | $1 / 6$ |
|  | $4 / 6$ | $5 / 10$ | 1 | $1 / 6$ |
| $2 / 3$ | $3 / 6$ | 1 | $1 / 6$ |  |
| $2 / 4$ | $3 / 9$ | 1 | $1 / 6$ |  |
| $4 / 6$ | $6 / 12$ | 2 | $1 / 6$ |  |
| $2 / 3$ | $4 / 8$ | 2 | $1 / 6$ |  |
|  | $1 / 2$ | $3 / 9$ | 2 | $1 / 6$ |
| $2 / 4$ | $4 / 12$ | 2 | $1 / 6$ |  |

Note. Prop. 1 means the first proportion value; Prop. 2 means the second proportion value. Absolute Distance means the absolute quantity distance, which is the difference for the white portions of the pair; Relative Distance means the relative quantity distance, which is the difference for the proportion values of white portion relative to either black portion or total (white plus black) portion.

## Procedure

Participants were instructed to select the stimulus which had larger white portion. Participants were asked to press "d"
when they judged the left stimulus had larger white portion and to press " j " when they judged the right stimulus had larger white portion. Both speed and accuracy were emphasized in instructions.

Each trial began with a 500 ms presentation of a fixation cross in the center of the screen, immediately followed by the stimulus pair. The pair stayed on the screen until participants submitted a response or timed out at 3000 ms .

In each block, each of these 34 proportion pairs was presented twice, either with the larger proportion to the left or to the right, giving 68 trials in each block. The stimuli in each block were presented in a random order. There were six different blocks, and the presentation order of these six blocks was counterbalanced, resulting a total of 408 trials.

## Results

Accuracy and mean reaction time (RT) were computed for each condition for each participant and used as the primary outcome variables. Only correct RTs were used in the analysis. We conducted separate repeated-measures ANOVAs using accuracy and RT.

The repeated-measure ANOVA on the accuracy with congruity and presentation format as within-subject factors was calculated. Results revealed that the main effect of congruity was significant, $F(1,32)=15.827, p=0.000$, $\eta_{p}^{2}=0.331$. Participants made more mistakes in the incongruent condition ( $M=77.1 \%, S E=4.6 \%$ ) than in the congruent condition ( $M=92.0 \%, S E=1.3 \%$ ). The main effect of presentation format, however, was not significant, $F(5,160)=0.184, p=0.671, \eta_{p}^{2}=0.006$. This indicated that adults' overall accuracy was not affected by presentation formats. Figure 2 depicted the pattern of accuracy for congruent and incongruent conditions for each presentation format.


Figure 2: Congruity effect on accuracy for each presentation format. "pcn" means part-to-part relation with continuous display; "pdd" means part-to-part relation with discretized display; "pds" means part-to-part relation with discrete display; "wen" means part-to-whole relation with continuous display; "wdd" means part-to-whole relation with discretized display; "wds" means part-to-whole relation with discrete display.

The two-way interaction of presentation format with congruity was significant, $F(5,160)=2.319, p=0.046$, $\eta_{p}^{2}=0.068$. The result indicated that the effect of congruity was modulated by presentation format.

We also analyzed reaction times in the same way as accuracy. Figure 3 displays the pattern of mean correct reaction times across conditions. Only the main effect of presentation format was significant, $F(5,105)=7.575, p=$ $0.000, \eta_{p}^{2}=0.265$. The main effect of congruity and the twoway interaction of congruity and presentation format were not significant, $p \mathrm{~s}>0.05$.


Figure 3: Congruity effect on response times for each presentation format.

## Discussion

The results of current experiment showed that adults made more mistakes making judgments in the incongruent conditions than in congruent conditions. And the size of congruity effect varied by different presentation formats. But response time did not show such a clear pattern as accuracy. Participants seemed to spend about same time comparing congruent and incongruent trials. Overall, the current study suggested that adults can automatically process the magnitudes of proportions even though it was irrelevant and disturbing to the absolute quantity comparison task, and that the congruity effect was modulated by the presentation format.

Even though more and more effort has been made to explore human's understanding of proportions, very little is known about the specific processing level of them. Consistent with previous findings that human have an intuitive understanding of proportion and represent them perceptually (e.g., Jacob, Vallentin, \& Nieder, 2012; Matthews, Lewis, \& Hubbard, 2015), our study demonstrated that proportions can be automatically processed. The observation of the congruity effect confirmed the findings of previous studies that showed the same automatic processing of proportion (e.g., Fabbri et al., 2012; Yang et al., 2015; Matthews \& Lewis 2016). The
study also provided evidence to the fact that humans, at least adults, can process proportion automatically no matter what kind of formats the proportion is presented.

Moreover, based on previous findings that presentation format can influence proportion processing, the present study found that the level of automatic processing of proportion varied for different presentation formats. The size of congruity effect of automatic processing of proportion was modulated by presentation format. Proportions presented as discretized part-to-part display seemed to show the largest difference of accuracy for congruent trials and incongruent trials. This finding was a little bit surprising, because previous studies suggested that continuous display promotes greatest success for proportion processing at least for children (Boyer, Levine, \& Huttenlocher, 2008). It is possible that adults adopt different processing strategies or preference than children. It would be interesting to see whether children shown different congruity effect pattern for these presentation formats. Another possibility is that the task in the current experiment was an implicit and unintentional task for proportion processing, while previous studies showing presentation differences were all explicit and intentional tasks for proportion processing (Sophian \& Wood, 1997; Spinillo \& Bryant, 1999; Jeong, Levine, \& Huttenlocher, 2007; Boyer, Levine, \& Huttenlocher, 2008). Humans might perform differently during two task scenarios. Further studies will be needed to address these issues.

Theorists generally have two different explanations to account for the mechanism of automaticity phenomenon. Some focus on the learned automatic processes and emphasize the learning mechanism (Anderson, 1992). Others believe there are innate automatic processes that humans are born with (Hasher \& Zacks, 1979). The current study cannot tell whether the mechanism of the automatic processing of proportion is natured or nurtured. 5-year-old children have been found to show similar congruity effect for accuracy but not response time in a sector comparison task, which provided some hint that the automatic processing of proportion is not acquired by learning or instruction (Yang et al., 2015). However, more evidence considering culture, education and intelligence, is required to reach final conclusions about the mechanism of automatic processing of proportion.

## Acknowledgments

This research was supported in part by NSF Grant DRL1420211.

## References

Anderson, J. R. (1992). Automaticity and the ACT theory. The American journal of psychology, 165-180.
Boyer, T. W., Levine, S. C., \& Huttenlocher, J. (2008). Development of proportional reasoning: where young children go wrong. Developmental psychology, 44(5), 1478.

Dehaene, S. (1997). The number sense. New York and Cambridge: Oxford University Press.
DeWolf, M., Bassok, M., \& Holyoak, K. J. (2015). Conceptual structure and the procedural affordances of rational numbers: Relational reasoning with fractions and decimals. Journal of Experimental Psychology: General, 144(1), 127.
Duffy, S., Huttenlocher, J., \& Levine, S. (2005). It is all relative: How young children encode extent. Journal of Cognition and Development, 6(1), 51-63.
Duffy, S., Huttenlocher, J., Levine, S., \& Duffy, R. (2005). How infants encode spatial extent. Infancy, 8(1), 81-90.
Fabbri, S., Caviola, S., Tang, J., Zorzi, M., \& Butterworth, B. (2012). The role of numerosity in processing nonsymbolic proportions. The Quarterly Journal of Experimental Psychology, 65(12), 2435-2446.
Hasher, L., \& Zacks, R. T. (1979). Automatic and effortful processes in memory. Journal of experimental psychology: General, 108(3), 356.
Henik, A., \& Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. Memory \& cognition, 10(4), 389-395.
Jacob, S. N., Vallentin, D., \& Nieder, A. (2012). Relating magnitudes: the brain's code for proportions. Trends in cognitive sciences, 16(3), 157-166.
Jeong, Y., Levine, S. C., \& Huttenlocher, J. (2007). The development of proportional reasoning: Effect of continuous versus discrete quantities. Journal of Cognition and Development, 8(2), 237-256.
Kallai, A. Y., \& Tzelgov, J. (2009). A generalized fraction: An entity smaller than one on the mental number line. Journal of Experimental Psychology: Human Perception and Performance, 35(6), 1845.
Lewis, M. R., Matthews, P. G., Hubbard, E. M., \& Matthews, P. G. (2015). Neurocognitive architectures and the nonsymbolic foundations of fractions understanding. Development of mathematical cognition: Neural substrates and genetic influences, 141-160.
MacLeod, C. M., \& Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(1), 126-135.
Matthews, P. G., \& Lewis, M. R. (2016). Fractions We Cannot Ignore: The Nonsymbolic Ratio Congruity Effect. Cognitive Science.
McCrink, K., \& Wynn, K. (2007). Ratio abstraction by 6-month-old infants. Psychological science, 18(8), 740-745.
Meert, G., Grégoire, J., \& Noël, M. P. (2009). Rational numbers: Componential versus holistic representation of fractions in a magnitude comparison task. The Quarterly Journal of Experimental Psychology, 62(8), 1598-1616.
Meert, G., Grégoire, J., \& Noël, M. P. (2010). Comparing $5 / 7$ and $2 / 9$ : Adults can do it by accessing the magnitude of the whole fractions. Acta Psychologica, 135(3), 284292.

Möhring, W., Newcombe, N. S., Levine, S. C., \& Frick, A. (2016). Spatial proportional reasoning is associated with
formal knowledge about fractions. Journal of Cognition and Development, 17(1), 67-84.
Ni, Y., \& Zhou, Y. D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. Educational Psychologist, 40(1), 27-52.
Schneider, M., \& Siegler, R. S. (2010). Representations of the magnitudes of fractions. Journal of Experimental Psychology: Human Perception and Performance, 36(5), 1227.

Sophian, C. \& Wood, A. (1997). Proportional reasoning in young children: The parts and the whole of it. Journal of Educational Psychology, 89(2), 309.
Spinillo, A. G. \& Bryant, P. E. (1999). Proportional reasoning in young children: Part-part comparisons about continuous and discontinuous quantity. Mathematical Cognition, 5(2), 181-197.
Vallentin, D., \& Nieder, A. (2008). Behavioral and prefrontal representation of spatial proportions in the monkey. Current Biology, 18(18), 1420-1425.
Yang, Y., Hu, Q., Wu, D., \& Yang, S. (2015). Children's and adults' automatic processing of proportion in a Stroop-like task. International Journal of Behavioral Development, 39(2), 97-104.

# Utilizing simple cues to informational dependency 

Hugo Mercier<br>CNRS

## Helena Miton

CEU


#### Abstract

Participants can adequately take into account several cues regarding the weight they should grant majority opinions, but that they do not consistently take into account cues regarding whether the members of the majority have formed their opinions independently of each other. We suggest that these conflicting results can be explained by hypothesizing that some cues are evolutionarily valid (i.e. they were present and reliable during human evolution), and others not. Using this framework we derive and test hypotheses about two facets of informational dependency. The first 3 experiments show that participants adequately take into account cues to informational dependency when they are presented in a simple, evolutionarily valid way. Experiments 4 to 7 show that people consistently take into account shared motivation, but not shared cognitive traits, as a source of potential dependency, as predicted by the likely greater importance of differences in motivation during our evolutionary history.


# Children's EEG Indices of Directed Attention during Somatosensory Anticipation: Relations with Executive Function 

Staci Meredith Weiss (SWeiss@ Temple.edu)<br>Department of Psychology, 1701 N. 13th Street<br>Philadelphia, PA USA<br>Peter J. Marshall (PJMarsh@Temple.edu)<br>Department of Psychology, 1701 N. 13th Street<br>Philadelphia, PA USA


#### Abstract

Children's ability to direct attention to salient stimuli is a key aspect of cognitive functioning. Here we examined the magnitude and lateralization of EEG indices during somatosensory anticipation elicited by a left or right directional cue indicating the bodily location of an upcoming tactile stimulus. In 50 children aged 6-8 years, somatosensory anticipation was accompanied by anticipatory negativity and alpha mu rhythm desynchronization at contralateral central electrode sites ( C 3 and C 4 ) overlying the hand area of the somatosensory cortex. Individual differences in these contralateral brain responses during somatosensory anticipation were associated with scores on a flanker task of executive function. The results suggest that processes involved in directing attention in the tactile modality may overlap with those involved in broader executive function abilities.


## Introduction

The ability to direct attention in a focused and efficient manner is crucial to cognitive performance and decisionmaking (Posner and Fan, 2008). Directed attention is the heightened monitoring of spatial location prior to presentation of a target stimulus, and is known to facilitate heightened perception, shorter reaction time and improved inhibition in adults (Rothbart, Posner and Kieras, 2006). Introduction of preparatory cues relevant for upcoming target stimuli allows study of neural activity or behavior during anticipation, prior to subsequent attention and perception of stimuli (Anderson and Ding, 2012). Anticipation is the goal-directed monitoring of sensation in expectation of a stimulus. Within a sensorimotor contingency framework, anticipatory brain responses can be viewed as 'pragmatic' in that they prepare for expected action (Engel et al., 2013), reflecting the reciprocal nature of prior experience, cognition and action (O'Reagen, 2011). In directed attention tasks with children, individual differences in post-stimulus neural activity to target stimuli (Stevens and Bavelier, 2012) are associated with cognitive skills and school achievement, but pre-stimulus anticipatory neural activity is rarely examined as an index of cognitive ability.

Directed attention paradigms investigate how manipulation of endogenous (top-down) attention facilitates subsequent exogenous (bottom-up) stimuli-driven attention, perception, and neural responses to target stimuli. Neural activity in the region that encodes target stimuli features are
modulated not only in response to stimuli, but also during anticipation (Corbetta and Shulman, 20012). This effect is present even when the preparatory cue is presented in a different sensory modality from that of the expected target, allowing temporal and spatial differentiation of anticipatory activity in the target sensory cortex from cue encoding (Zanto and Gazzley, 2009). During anticipatory attention, preparatory cues uniquely engage the intra-parietal sulcus (IPS) to apply a filter on attention, which is not active in subsequent target perception (Corbetta and Shulman, 2002).

Emerging research accounts for neural activity unique to anticipation as filtering the focus of attention in expectation of an upcoming target (Zhang and Ding, 2010). To examine this account, we study how the magnitude of anticipatory neural activity during a somatosensory directed attention task relates to performance on the flanker task, which demands conflict monitoring, or a focus on single target stimuli among distractors competing for attention (Rothbart, Posner and Kieras, 2006). The flanker task requires participants to respond to the direction of a central target arrow amidst congruent or incongruent flanking distractor arrows: incongruent conditions are associated with slower reaction times, explained as resolving the conflict between the target and distractors. The flanker task taps into endogenous attentional abilities measured by executive function (EF). Our understanding of EF is informed by relational-systems theory, with EF defined as the goaldirected regulation of behavior, aligning action with topdown attention (Dick and Overton, 2011). We are concerned with the relations between flanker task performance and individual differences in the neural indices during somatosensory anticipation, elicited in a directed attention paradigm in which a visual cue directs children's attention to the expected spatial location of an upcoming tactile stimulus. Associating neural indices of somatosensory anticipation with EF facilitates study of inter-sensory attention, with potential for linking sensory-specific attentional processes to cognitive skills.

## EEG Indices of Somatosensory Anticipatory Attention

Changes in brain responses recorded through the electroencephalogram (EEG) are reliable indicators of attention orienting, sometimes proving more predictive of relevant behavior than reaction time responses to stimuli (Foxe and Snyder, 2011). Brain responses to directed
attention are lateralized, such that there is a modulation of neural activity contralateral to the direction of the spatial cue (Gazzaley and Nobre, 2012). Attention can be indexed by the modulation of EEG signals in the alpha band (8-14 Hz in adults). Alpha band fluctuations are interpreted as a correlate of underlying attentional states, with the magnitude of change in amplitude sensitive to stimuli salience, strength and individual differences. When monitoring stimuli presented to one visual field, to one hand or to one ear, there is typically a disruption of rhythmic alpha activity known as event-related desynchronization (ERD) in the contralateral sensory cortex (Stevens and Bavelier, 2012). Desynchronization of the alpha rhythm appears to reflect an increase in local field potentials of neurons in the region of interest. Heightened alpha band desynchronization is thought to increase the perceptual salience of upcoming stimuli in the target modality (Foxe and Snyder, 2011). In the tactile modality, anticipatory desynchronization of the alpha-range mu rhythm at central electrodes is an index of somatosensory cortex excitability.

Attention-related changes in the EEG during anticipation can also be indexed via event-related potential (ERP) methods. Relevant here is the contingent negative variation (CNV), a negative-going potential occurring during the anticipatory period between a preparatory cue and a target stimulus (Corbetta and Shulman, 2002). The CNV elicited by a preparatory cue can be considered to reflect endogenous anticipatory directed attention, while later potentials (P1, N2, P3) can be considered to reflect exogenous, stimulus-evoked attention.

There is an emerging literature studying pre-stimulus EEG activity during the anticipation of touch. Detection of weak tactile stimuli was predicted by contralateral power of anticipatory alpha desynchronization (Zhang and Ding, 2010) after a cue indicated upcoming stimulation of the right hand. Anticipatory somatosensory selective attention paradigms often include a cue containing relevant spatial information about upcoming tactile stimulation, to examine the lateralization of brain responses or hemispheric asymmetry. Haegens, Luthur and Jensen (2012) demonstrated anticipatory desynchronization of alpha rhythm in central electrodes contralateral to the direction of the spatial cue, and synchronization of the alpha rhythm in the ipsilateral central electrodes. To our knowledge, there is no existing research on the cognitive mechanisms facilitating the association of pre-stimulus alpha mu desynchronization and post-stimulus enhanced perception of tactile stimuli. There are also no studies on the developmental trajectory of neural indices during somatosensory anticipation.

## Development of Directed Attention and EF

Precursors to attention regulation are apparent in infancy, with gaze fixation sensitive to stimulus features, novelty preferences, and prior learning in newborns (Hood, Willen, Driver, 1998; Sheese et al., 2008). Goldberg, Maurer and

Lewis (2001) found age-related differences in children's target discrimination on visual selective attention paradigms when target and non-target stimuli were presented simultaneously. By age 8 years, performance was comparable to adults in anticipatory visual selective attention tasks. The authors suggested discrepant trajectories in the development of distractibility and anticipation.
There is interest in studying the neural indices elicited by selective attention, when distractors presented simultaneous to the target compete for directed attention. Coch, Sanders and Neville (2005) employed a dichotic listening task to examine neural indices in response to target tones in attended and unattended simultaneous auditory streams in children aged 6-8. They found a slow positive ERP peaking around 150 ms in children as opposed to the typical ERPs found in adults: early sensory potentials, followed by an N2 and P3. The amplitude of the ERP was greater in the attended auditory stream than the unattended auditory stream. Isbell, Wray and Neville (2015) found the positivity of auditory-evoked selective attention potentials related to non-verbal IQ scores, such that only higher-IQs preschoolers exhibited significant differences in amplitude distinguishing target tones in the attended stream from tones in the unattended stream. The nature of the dichotic listening task requires the same skills as the tasks involved in non-verbal IQ assessment. A study of anticipatory attention to visual stimuli in 10-year-old children found that CNV magnitude (evoked by preparatory spatial cues) related to visual short-term memory capacity (Shimi et al., 2014). The measure of EF was response to the targets of 'anticipatory attention', so it is difficult to tease apart the relations among task performance, working memory and neural indices of attention.

Prior studies suggest adult executive function relates to brain responses during attention to visual or auditory stimuli. During a dichotic listening task, the amplitude of the contralateral brain responses during auditory selective attention relates to non-verbal working memory (Giuliano et al., 2014). In contrast, a study of visual selective attention found only ipsilateral brain responses to be related to cognitive skills (Zanto \& Gazzaley, 2009). The authors found that the ipsilateral increase in amplitude accounted for greater variation in working memory, when compared to the magnitude typically associated with selective attention, which is reduced alpha power in the contralateral occipital cortex. The current investigation extends the study of anticipatory attention and cognitive skills to the somatosensory domain by examining how EF relates with EEG indices of directed attention to upcoming tactile stimuli.

## Current Study

Studying pre-stimulus attention in the tactile modality contributes to the basic science of bodily awareness and somatosensation. The fundamental, early-developing nature of somatosensory processing (Marshall \& Meltzoff, 2015) make it a compelling domain of study for examining the
development of top-down attentional processes. Furthermore, perception of tactile pulses appears uniquely associated with neural indices of attention, unlike reaction time in response to tactile pulses (Zhang and Ding, 2010): Electrophysiological data can therefore provide insight into the perception of touch, beyond simultaneous behavioral measures. Our study examines how the somatosensory domain adds to the developmental literature on executive function indices of domain-general attention.

The current study examined the magnitude and lateralization of CNV potentials and sensorimotor alpha mu rhythm modulation, as elicited by a directional cue indicating the bodily location of upcoming tactile stimuli. We hypothesized that individual differences in brain responses during somatosensory anticipation would be associated with scores on a flanker task. This potential association between the cognitive state elicited during anticipation and EF was the key focus of the study.

## Methods

## Participants

Sixty children between the ages six to eight years of age participated in the study ( $M=7.2$ years, $S D=.6 ; 27$ male). Families were recruited from a diverse urban environment using commercially available mailing lists and online advertisements. Families were not invited to participate if their child had any medical or psychological diagnoses, was left-handed, or on any long-term medication. A number of children were excluded from analyses because they did not have a sufficient number of artifact-free trials $(n=6)$ or because the child did not tolerate cap preparation ( $n=4$ ). These 10 excluded children did not statistically differ in scores on the flanker from the remaining sample ( $\mathrm{N}=50$.)

## Procedure

Children were read an assent form outlining the protocol in the presence of their caregiver, who also read the consent forms. Children were then fitted with an EEG cap while seated at a table facing a computer screen, with instructions to stay as still as possible with their hands on their lap, out of sight. Research assistants explained the paradigm as a game that required children to pay close attention to the right or left hand, as indicated by the arrow, and respond to the tactile stimuli by pressing a foot pedal once if they felt one tap or twice if they felt two taps. Each of the 120 trials began with a fixation cross baseline for 1500 ms , followed by an arrow displayed for 500 ms , followed by a response screen which read 'Copy with Your Foot!' (Figure 1).


Figure 1. Protocol consisted of 120 trials: a 1500 ms baseline, then an arrow displayed for 500 ms , then tactile stimuli and response.

Tactile stimulation was delivered using a pneumatic simulator controlled by STIM stimulus presentation software (both device and software from James Long Company), with the compressed air delivered during the arrow display and before the response screen. An inflatable membrane mounted in a plastic casing was placed on the middle fingers of children's left and right hands, held in place by a finger clip. The membrane is inflated by a short burst of compressed air delivered via flexible polyurethane tubing ( 3 m length, 3.2 mm outer diameter). The tactile stimulus feels like a light tap on the finger, lasts around 60 ms , and has a peak force of around 2 N .

The NIH Cognition Toolbox flanker task was then administered. Children completed the flanker task on an iPad by selecting the direction of a central target arrow among 4 flanking distractor arrows, which were either congruent or incongruent in direction to the target arrow. Scores were calculated as number of trials with correct response for incongruent trials weighted by reaction time, such that a higher score indicated better EF abilites.

## EEG Collection and Processing

EEG was recorded using a 32-electrode stretch cap (ANT Neuro, Inc.) from the following sites: Fp1, Fpz, Fp2, F3, F4, Fz, F7, F8, C3, C4, CP1, CP2, T7, T8, P3, P4, Pz, P7, P8, $\mathrm{O} 1, \mathrm{Oz}, \mathrm{O} 2$, and the left and right mastoids. Conducting gel was used and scalp electrode impedances were kept under $25 \mathrm{k} \Omega$. The signal from each site was amplified using optically isolated, high input impedance ( $>1 \mathrm{G} \Omega$ ) custom bioamplifiers (SA Instrumentation) and was digitized using a 16 -bit A/D converter ( $+/-5 \mathrm{~V}$ input range). Bioamplifier gain was 4000 and the hardware filter ( 12 dB /octave rolloff) settings were .1 Hz (high-pass) and 100 Hz (low-pass). EEG analysis was performed using the EEGLAB 13.5.4b toolbox (Delorme and Makeig, 2004) implemented in MATLAB. The signal was collected referenced to the vertex $(\mathrm{Cz})$, and EEG signals were re-referenced offline to an average mastoids reference for further analysis. Independent component analysis (ICA) cleared EEG data of ocular and muscle artifact. The ICA procedure was an automation of the method described by Hoffmann \& Falkenstein (2008). Visual inspection of the EEG signal was then used to reject epochs containing movement artifact. The mean number of artifact-free trials per cue direction was 41 ( $\mathrm{SD}=5.71$ ).

## Results

Analyses focused on electrode sites overlying the hand area (the left and right central electrodes; C3 and C4) of sensorimotor cortex. For all ANOVAs, within-subject effects were adjusted using Greenhouse-Geisser correction factors; pairwise t-test comparisons and multiple linear regressions were reported with p-values adjusted for multiple comparisons.

## Anticipatory Negativity

We extracted mean amplitude during the 300 ms immediately preceding tactile stimulation, within the 500 ms window relevant to anticipatory attention. This window
was selected to study the CNV waveform, accounting for a 200 ms delay after the preparatory cue (Shen et al., 2017).

To study lateralization of anticipatory negativity, an ANOVA compared mean CNV amplitude by electrode (C3 or C4) and cue direction (left or right arrow). There was a significant main effect of cue direction, such that cues directing attention to the right hand elicited an enhanced anticipatory negativity $(\mathrm{F}(1,48)=37.06, \mathrm{p}<.019)$. There was a significant main effect of electrode, such that C3 exhibited enhanced anticipatory negativity $(\mathrm{F}(1,48)=$ $37.06, \mathrm{p}<.001$ ). As expected, there was significant interaction between electrode ( C 3 or C 4 ) and cue direction, $\mathrm{F}(1,48)=15.95, \mathrm{p}<.001$, driven by negativity in the electrode contralateral to the stimuli: pairwise comparisons (adjusted with FDR) reveal amplitude in C3 was lower for right cue than left cue, $\mathrm{p}<.001$, while the opposite trend was found for the amplitude of $\mathrm{C} 4, \mathrm{p}<.061$ (see figure 2).


Figure 2. For each trial, an epoch of 1500 ms was extracted: analyses focused on the 300 ms preceding tactile stimulation during the pre-stimulus period, accounting for the average amplitude during a pre-cue $200-\mathrm{ms}$ baseline. The epochs were then filtered at 30 Hz .
I
Anticipatory ERPs and Flanker. To examine the relations between flanker and anticipatory negativity, regressions were conducted predicting flanker scores from lateralization of CNV amplitude and its interactions with electrode and cue direction. We computed lateralization of CNV amplitude by subtracting the mean amplitude at C 3 from mean amplitude at C 4 , for each participant and each hand. There was a trend of lateralized amplitude predicting flanker score, $\mathrm{t}(1,48)=1.72, \mathrm{p}=.090$. Performance on the flanker task was predicted by a significant interaction between lateralized amplitude and cue direction, $t(1,48)=-2.40, \mathrm{p}$ $=.019$. To further probe this interaction, we performed posthoc regressions predicting flanker score by average amplitude for C3 and C4 during right and left cues. The interaction was driven by marginal relations between flanker and amplitude in the contralateral hemisphere: flanker score related with CNV amplitude over C3 elicited by a right directional cue, $\mathrm{t}(1,49)=-1.72, \mathrm{p}=.094$, and over C 4 elicited by a left directional cue, $\mathrm{t}(1,49)=-1.70, \mathrm{p}=.097$. Amplitude in the ipsilateral hemisphere did not relate with flanker score.

## Time Frequency Analysis

Event-related spectral perturbation (ERSP) analyses were conducted on the alpha frequency band at $7-12 \mathrm{~Hz}$ (Berchicci et al., 2011; Marshall, Bar-Haim, \& Fox, 2002), as is appropriate for children, and baseline corrected for the 500 ms prior to cue onset. Event-related desynchronization (ERD) is as an alpha power decrease relative to the baseline. An ANOVA compared pre-stimulus alpha power by electrode ( C 3 or C 4 ) and cue direction (left or right arrow). There was a main effect of cue direction, such that cues directing attention to the right hand elicited greater desynchronization, $\mathrm{F}(1,48)=37.06, \mathrm{p}<.001$. There was no main effect of electrode position. A significant interaction was observed between electrode and cue direction, $\mathrm{F}(1,48)=15.95, \mathrm{p}<.001$ (see figure 3). The interaction was due to greater ERD in the contralateral hemisphere (see figure 4): the mu rhythm showed greater desynchronization over C3 elicited by the right cue, and over C 4 elicited by the left cue.


Figure 3. For each trial, an epoch of 1500 ms was extracted: spectral power was estimated using Gaussiantapered Morlet wavelets, and changes in power were computed as ERSP focused on the 500 ms pre-stimulus period, relative to a 500 ms baseline $(-1000$ to $-500 \mathrm{~ms})$.

Mean Event-Related Desyncronization(-500-0ms)


Figure 4. Mean ERSP for mu rhythm $(7-12 \mathrm{~Hz})$ in $\mathrm{C} 3 / \mathrm{C} 4$ for left/ right cues from -500 ms to 0 ms . Negative values reflect a reduction in mu power (ERD) relative to a 500 ms pre-cue baseline ( -1000 to -500 ms ).

Anticipatory ERSP and Flanker. To examine the relation of flanker task scores with ERD during somatosensory anticipation, regressions were conducted predicting scores on flanker task by ERD and its interactions with electrode position and cue direction, controlling for within-participant variability. There were no significant main effects or twoway interactions. However, performance on the flanker task was significantly predicted by a three-way interaction between alpha power, electrode position and cue direction, t (46) $=2.33, \mathrm{p}=.012$. Post-hoc regressions (with p-values adjusted for the false discovery rate) further probed the relations, revealing ERD in the hemisphere contralateral to the direction of the cue was inversely related to the scores on the Flanker task: for right cue over C3, $\mathrm{t}(1,49)=-2.65$, $\mathrm{p}=.014$, and for left cue over $\mathrm{C} 4, \mathrm{t}(1,49)=-2.50, \mathrm{p}=.021$. ERD in the ipsilateral hemisphere was not related with flanker score (see Table 1). Refer to Figure 5 for inverse correlations between flanker score and contralateral ERD.

Table 1. Flanker predicted by ERD, Cue, \& Electrode

| Overall Regression | $B$ | Std Error | t value | $p$ |
| :--- | :---: | :---: | ---: | :--- |
| ERD | -3.51 | 2.53 | -1.39 | .16 |
| Electrode | -.627 | 2.49 | -.252 | .80 |
| Cue | 2.44 | 2.45 | .977 | .32 |
| Power*Cue*Electrode | 2.33 | .021 | 2.33 | $.01^{*}$ |
| Post-Hoc Regression | $B$ | Std Error | t value |  |$\quad p$.

*Significant at less than $\mathrm{p}=.05$; p-values of post-hoc corrected for FDR


Figure 5. Correlations of Flanker scores with ERD: significant correlations were found for pre-stimulus ERSP contralateral to cue direction. For right cue over C3, r= .395; for left cue over $\mathrm{C} 4, \mathrm{r}=-.377$. There was no relation between flanker scores and ipsilateral ERSP; right cue over $\mathrm{C} 4, \mathrm{r}=.106$; left over $\mathrm{C} 4, \mathrm{r}=-.021$.

## Discussion

The current study investigated somatosensory anticipation in children, and if anticipatory EEG responses in amplitude and alpha mu rhythm were related with individual differences in flanker scores. Just as seen in adults, we found a preparatory spatial cue directing attention to the bodily location of upcoming tactile stimulation modulated the activity of the alpha mu rhythm. Children's contralateral alpha band activity (in C3 - Right Cue and C4 - Left Cue) during anticipation was inversely associated with performance on a flanker task, while ipsilateral responses had no relation to flanker scores. Anticipatory negativity (i.e., the CNV response) was lateralized, with more negative amplitudes in the hemisphere contralateral to cue direction. The degree of lateralization, computed as mean amplitude at $\mathrm{C} 4-\mathrm{C} 3$, was related to participants' flanker scores, which subsequent analyses demonstrated were driven by contralateral negativity. Children's ability to modulate attention in preparation for tactile stimulation appears to be related to individual differences in EF, as codified by their scores on the flanker task.
Our findings of an association between flanker and neural indices of tactile attention should be interpreted with caution, as our sample is small for a study of individual differences. Studies linking neural indices of somatosensory anticipation and attention with other EF tasks should be conducted, to parallel research in other sensory modalities (Isbell, Wray and Neville, 2015). Other studies of anticipation in adults suggest that ipsilateral EEG activity suppresses responses to distractors (Hagens, Luthur, and Jensen, 2012; Zanto \& Gazzaley, 2009), but our paradigm did not include distractor stimuli; our study observed only a trending alpha power increase in the ipsilateral hemisphere, elicited by stimuli to the left hand. Further investigations can examine whether introducing simultaneous, competing stimuli in the somatosensory domain elicits an ipsilateral increase in alpha mu rhythm, as observed during anticipatory somatosensory selective attention in adults. A paradigm with simultaneous tactile stimulation to both hands would be more parallel to developmental research on the relations of cognitive skills and children's somatosensory selective attention rather than our study of somatosensory directed attention. First, we believed it crucial to replicate modulation in EEG indices during the anticipation of tactile stimuli in children, and further investigation is needed to assess how preparatory cues influence anticipation, attention and perception when distractors are introduced. Our results support that shared processes are involved in sensory-specific directed attention and domain-general EF, but continued study of attention in different modalities will address theories of inter-sensory attentional mechanisms (Gazzaley and Nobre, 2012). Study of brain responses to tactile stimuli uniquely informs models of action-oriented representation, wherein internal (cognitive and physiological) states reflect the environment and prescribe action (Clark, 1998; Engel et al., 2013).

Research on bodily awareness and representation would benefit by studying how the associated neural indices of somatosensation are influenced by attention and related cognitive processes. Children's neural indices during directed attention to bodily sensation appear similar to those in adults, when compared to ERP responses evoked during visual and auditory attention (Stevens and Bavelier, 2012). This could be interpreted as signaling the importance of somatosensory attention in development, although further work is needed to establish whether tactile attention is more predictive than attention in other modalities (e.g. vision). Future studies of the neural indices of somatosensory attention could lay the foundation for interventions that train attention or executive function in children.

## Acknowledgements

The education of the primary author is supported by a PreDoctoral NSF Graduate Fellowship. We thank the members of the Developmental Science Lab for their feedback and assistance in data collection; Guannan, Ashley, Jeb and Val.

## References

Anderson, K. L., \& Ding, M. (2011). Attentional modulation of the somatosensory mu rhythm.
Neuroscience, 180, 165-180.
Berchicci, M., Zhang, T., Romero, L., Peters, A., Annett, R., Teuscher, U., \& Comani, S. (2011). Development of mu rhythm in infants and preschool children. Developmental neuroscience, 33(2), 130-143.
Clark, A. (1998). Being there: Putting brain, body, and world together again. MIT press.
Corbetta, M., \& Shulman, G. L. (2002). Control of goaldirected and stimulus-driven attention in the brain. Nature reviews neuroscience, 3(3), 201-215.
Coch, D., Sanders, L. D., \& Neville, H. J. (2005). An eventrelated potential study of selective auditory attention in children and adults. Journal of cognitive neuroscience, 17(4), 605-622.
Delorme, A., \& Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 134(1), 9-21.
Dick, S., \& Overton, W.F. (2010) Executive Function: Description and Explanation. In Sokol, B. W. (Ed.) Self and social regulation: Social interaction and the development of social understanding and executive functions, 7-34. Oxford University Press.
Engel, A. K., Maye, A., Kurthen, M., \& König, P. (2013). Where's the action? The pragmatic turn in cognitive science. Trends in cognitive sciences, 17(5), 202-209.
Foxe, J. J., \& Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. Frontiers in psychology.
Gazzaley, A., \& Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. Trends in cognitive sciences, 16(2), 129-135.
Haegens, S., Luther, L., \& Jensen, O. (2012). Somatosensory anticipatory alpha activity increases to
suppress distracting input. Journal of Cognitive Neuroscience. 24(3), 677-685.
Hoffmann, S., \& Falkenstein, M. (2008). The correction of eye blink artefacts in the EEG: a comparison of two prominent methods. PLoS One, 3(8), e3004.
Hood, B. M., Willen, J. D., \& Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. Psychological Science, 9(2), 131-134.
Isbell, E., Wray, A. H., \& Neville, H. J. (2015). Individual differences in neural mechanisms of selective auditory attention in preschoolers from lower socioeconomic status backgrounds. Developmental science.
Marshall, P.J., Bar-Haim, Y., \& Fox, N., (2002). Development of the EEG from 5 months to 4 years of age. Clinical Neurophysiology, 113(8), 1199-1208.
Marshall, P. J., \& Meltzoff, A. N. (2015). Body maps in the infant brain. Trends in cognitive sciences, 19(9), 499-505.
O'Regan, J.K. (2011). Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness. Oxford University Press.
Posner, M. I., \& Fan, J. (2008). Attention as an organ system. Topics in integrative neuroscience, 31-61.
Rothbart, M. K., Posner, M. I., \& Kieras, J. (2006). Temperament, Attention, and the Development of SelfRegulation. Chicago.
Sheese, B. E., Rothbart, M. K., Posner, M. I., White, L. K., \& Fraundorf, S. H. (2008). Executive attention and selfregulation in infancy. Infant Behavior and Development, 31(3), 501-510.
Shimi, A., Nobre, A. C., Astle, D., \& Scerif, G. (2014). Orienting attention within visual short-term memory: Development and mechanisms. Child development, 85(2), 578-592.
Stevens, C., \& Bavelier, D. (2012). The role of selective attention on academic foundations. Developmental cognitive neuroscience, 2, S30-S48.
Shen, G., Saby, J. N., Drew, A. R., \& Marshall, P. J. (2017). Exploring Potential Social Influences on Brain Potentials during Tactile Stimulation. Brain Research.
Zanto, T. P., \& Gazzaley, A. (2009). Neural suppression of irrelevant information underlies optimal working memory performance. Journal of Neuroscience, 29(10), 3059-3066.
Zhang, Y., \& Ding, M. (2010). Detection of a weak somatosensory stimulus: Role of the pre-stimulus mu rhythm and its top-down modulation. Journal of cognitive neuroscience, 22(2), 307-322.

# Modeling Unsupervised Event Segmentation: Learning Event Boundaries from Prediction Errors 

Katherine Metcalf (metcalka@indiana.edu)<br>Department of Computer Science, 901 E. 10th St. Bloomington, IN 47408 USA<br>David Leake (leake@indiana.edu)<br>Department of Computer Science, 901 E. 10th St. Bloomington, IN 47408 USA


#### Abstract

Segmenting observations from an input stream is an important capability of human cognition. Evidence suggests that humans refine this ability through experiences with the world. However, few models address the unsupervised development of event segmentation in artificial agents. This paper presents work towards developing a computational model of how an intelligent agent can independently learn to recognize meaningful events in continuous observations. In this model, the agent's segmentation mechanism starts from a simple state and is refined. The agent's interactions with the environment are unsupervised and driven by its expectation failures. Reinforcement learning drives the mechanism that identifies event boundaries by reasoning over a gated-recurrent neural network's expectation failures. The learning task is to reduce prediction error by identifying when one event transitions into another. Our experimental results support that reinforcement learning can enable detecting event boundaries in continuous observations based on a gated-recurrent neural network's prediction error and that this is possible with a simple set of features.


Keywords: Event Cognition; Unsupervised Segmentation; Expectation-based Failures; Reinforcement Learning

## Introduction

The ability to derive meaning from complex observations is a skill that has been recognized as vital for "growing" an intelligent agent from a simple starting state through interactions with a complex environment (Brooks, 1995; Cohen, Oates, Atkin, \& Beal, 1996). Before the agent is able to reason over a world model, it must first develop one. We present an approach in which an agent, exposed to patterns with temporal dependencies, develops a predictive model of its environment. The agent's expectation failures (i.e. prediction errors) are then used as the basis of its event segmentation mechanism. The resulting segments form the foundation of event representations.

The research we present in this paper builds on the work by Reynolds, Zacks, and Braver (2007) to build a computational model of event segmentation. We extend their model by incorporating a reinforcement learning agent to handle the detection of event boundary locations and trigger the subsequent event segmentation. The prediction mechanism is the gated-recurrent neural network (GRNN) model outlined by Reynolds et al. We evaluated several variations on the state representation presented to the reinforcement learning agent. The representations leverage information about the GRNN's
prediction error through time. The first representation evaluated is a simple state representation composed of the ratio of the predictive model's current error to its average error. This simple representation is then expanded to include a measure of input change, the amount of time since the gate was last opened, and the type of event that is expected next. Each state representation we evaluated contained the prediction ratio. We tested the GRNN-RL pair and the state represations on a motion captures dataset representing people executing 13 distinct tasks. Our results support the idea that information about the GRNN's prediction error is sufficient to allow a learned RL policy to appropriately identify event boundaries.

## Motivation

People are able to unconsciously and effortlessly perceive sequences of discrete events from dynamic and continuous sensory input (Radvansky \& Zacks, 2014; Ross \& Baldwin, 2015). People's ability to recognize temporal structure and patterns frequently observed across environmental contexts facilitates their partitioning of continuous activities into discrete events (Elman, 1990; Cleeremans \& McClelland, 1991; Cohen \& Adams, 2001; Reynolds et al., 2007). Therefore, people must learn the sequential dependencies that allow them to reason about sequences of observations as single, individual events. Reasoning about both observed events and their associated spatiotemporal patterns allows humans to reason about the underlying cause of the change in sensory observations (Radvansky \& Zacks, 2014). Evidence suggests that when people use an inferred event (i.e., spatiotemporal pattern) to guide their sensory expectations, they are able to recognize when transitions between events occur because their observations no longer match that of the current, hypothesized spatiotemporal pattern (Braver \& Cohen, 2000; Rougier, Noelle, Braver, Cohen, \& O’Reilly, 2005). We model how agents develop spatiotemporal models and use them to to interpret continuous observations as discrete events.

## Background

The task of this paper is related to previous works such as the Neo project (Cohen et al., 1996). Neo is a simulated infant that implements a computational model of the perceptual analysis by image-schema theory of complex concept formation (Johnson, 1987; Mandler, 1988, 1992; Lakoff \& John-
son, 2008). Neo begins with relatively simple configurations and develops/learns complex concepts through interactions with a simulated, complex world, by analyzing occurrences of discrete, symbolic tokens. We agree on the importance of developing complex agents able to learn via a process of perceptual analysis, representations of objects, states, and activities as its foundation for learning conceptual categories. However, a precondition for systems such as Neo is a process for transforming continuous sensory observations of the world into meaningful, discrete units (i.e., "unitization"). Our work addresses the unitization problems.

Reynolds et al. (2007) propose a relatively simple mechanism for event segmentation. Using its experiences with the world, their mechanism refines and hones the way it segments continuous observations without prior knowledge about the events or about the locations of event boundaries. Their mechanism is an implementation of the first component of Event Segmentation Theory (EST), a theory of event schema/model creation (Zacks, Speer, Swallow, Braver, \& Reynolds, 2007; Kurby \& Zacks, 2008). While previous approaches to event segmentation focused on the degree of change between subsequent observations as the key predictive feature (Newtson, 1976; Gibson, 1979), EST emphasizes the role of prediction failures. The importance of prediction error during event segmentation is based on data suggesting that people attempt to predict what they will observe next (Rao \& Ballard, 1999; Enns \& Lleras, 2008; Niv \& Schoenbaum, 2008).

People maintain working models, dynamic representations that facilitate event comprehension and incorporate predictions about what will be observed next, of the events they are observing (Radvansky \& Zacks, 2014). Evidence suggests that working models are the result of the segmentation and chunking of experience that are triggered by transient increases in prediction error (i.e., expectation failure driven event segmentation). When an event boundary is detected, people update their working model, thus changing their expectations about what will be observed next. However, there is a key limitation in the approach taken by Reynolds et al. when implementing this process, as their system depends on externally set thresholds to determine when one event ends and another begins. The work we present here extends their prediction model by removing externally set thresholds and examining the impact of incorporating higher-level expectations.

## Modeling Prediction Error-based Segmentation

## Reynolds, Braver, and Zack's Segmentation Model

Reynolds et al. (2007) used a gated-recurrent neural network, with the architecture depicted on the left in Figure 1, to model how people might learn sequential dependencies and perceive discrete event categories from continuous observations. The GRNN identified points at which one activity transitioned into another via an expectation failure based heuristic. They selected the GRNN because they considered it the most bio-


Figure 1: GRNN Model Architecture (Reynolds et al., 2007) on the left. Unsupervised Event Segmentation Model Architecture on the right.
logically plausible model available for capturing how people might learn sequential dependencies with the ability to store a representation of the current event in memory based, on contemporary work in behavioral and neuropsychological correlates of event structure and computational studies of sequential domains (for an extensive literature review see Reynolds et al. (2007)). The GRNN adjusted its event representation by triggering a gating mechanism that allowed the event representation to be directly updated based on the GRNN's observations. The gating mechanism controlled the extent to which the event representation was updated by each new observation and, combined with the network's recurrence, allowed the GRNN to maintain representations of the events through time (Elman, 1990; Hochreiter \& Schmidhuber, 1997). In their simulations, the gate was operated either: (1) by ground truth knowledge about the location of event boundaries or (2) by an externally set threshold on the ratio of the model's current sum squared error (SSE) and its average SSE. Their model attempted to predict its next observation; expectation failures were measured as SSE in the model's prediction and the true next observation. Based on the distribution of SSE observed within events versus at event boundaries, the authors concluded that a GRNN with an expectation failure-based gating mechanism is a reasonable approximation of how people might segment sequences of observations into meaningful units.

We built on Reynolds et al.'s (2007) work by extending their GRNN to include a RL agent that learns a policy for controlling the gating mechanism.

## A New Approach to Unsupervised, Self-Regulating Event Segmentation

We incorporated a RL agent that learned a policy for controlling the gating mechanism that Reynolds et al. (2007) created for their final simulation (Simulation IIIB), with the architecture depicted on the right in Figure 1. In Simulation IIIB, the gating mechanism was controlled by a simple mathematical function that evaluated whether the ratio of the models' last observed prediction error relative to the observed average error exceeded a threshold (1.5). Before modifying Simulation IIIB to include the RL agent, we tested our implementation in order to replicate their the experimental results, and we observed the same relations between the within event and
boundary observations. Reproducing the author's results (1) allowed us to evaluate the reproducibility of their work before using it as the basis of our system and (2) allowed us to build our model on one already reviewed and evaluated by the scientific community.

In our model, the gating function from Simulation IIIB was replaced with a RL agent that learned a policy for controlling the gating mechanism. An Expected-Sarsa learning algorithm with a linear function approximator (Sutton \& Barto, 2015) was used to learn the policy for controlling the gating mechanism. Our GRNN and RL combination is only able to build low-level expectations about what it will observe next. However, it could be used as a component in a larger system to build higher-level expectations that could be used to help guide the actions of the RL agent. In our experiments, while our model only directly builds expectations at the lower level, we incorporate information at higher levels of expectation in the the RL agent's state representation. We distinguish between lower and higher level information based on whether or not the information stems from the RL agent's immediately available observations about the state of the GRNN. Lowerlevel information is information that is readily available to the RL agent, whereas higher level information is not immediately available. For instance, the degree to which there is change between two subsequent observations is readily observable, whereas knowledge about the likelihood of a the next event transition being from sitting to standing is not.

In our experiments, the combined GRNN and RL agent was presented with a sequence of frames of motion captures of activities carried out by people (i.e. sitting,standing,jumping,etc.). Each activity constituted a single event and contained some number of frames. The frames are what the GRNN observed. The experiments varied the information presented to the RL agent. The RL agent's state representations consisted of both lower and higher level information about the state of the GRNN and RL agent. The lower-level information included a measure of the GRNN's prediction error (as in Reynolds et al. (2007)), the degree of change in the system's subsequent observations (inspired by Newtson (1976); Gibson (1979)), and the amount of time since the RL agent last updated the event representation. The higher level information included was a representation of the next event the agent expected to observe.

## The Models

The GRNN was constructed with the same parameters used by Reynolds et al.: 54 input units, 100 hidden units, 100 event units, 100 recurrent units, and 54 output units. The input and the hidden units had sigmoidal activation functions. The weights were initialized randomly within the range $[-0.5,0.5]$ and during back-propagation the learning rate was 0.001 . When comparing our implementation of the GRNN to that of Reynolds et al. we trained it to asymptotic performance, roughly 20,020 events, and evaluated it on 900 events. For further specifics about how the GRNN
was configured, please refer to details about Simulation IIIB in Reynolds et al. (2007). The GRNN was trained on 50, 000 events with a perfect gating signal prior to incorporating the RL agent as the gating mechanism.

The RL agent was construction according to an $\varepsilon$-greedy Expected-Sarsa with replacing traces policy learning algorithm. The specific state representations the agent learned to operate over can be seen in the experiments section below. Each of the state representations contained at least one continuous feature, therefore a linear function approximator was used to estimate the value of each state-action pair. Tile coding was used to convert the continuous states into binary feature vectors consisting of 32 layers of tilings with 4 tiles for each feature.

The agent had two possible actions: (1) flip the gate and (2) do not flip the gate. The policies were learned according to -SSE computed from the SSE observed in the GRNN's predictions after each action by the RL agent. We chose this reward, because it allows for unsupervised to control the gating mechanism and it is aligned with event segmentation theory (Radvansky \& Zacks, 2014). Table 1 shows the learning parameters used to learn the policies for each of the state representation experiments.

## Experiments

Experiments evaluated the performance of the RL agent at detecting when the GRNN's event representation should be updated. In each experiment, the RL learning algorithm described above was evaluated according to the quality of the policy it was able to learn given the different state representations. The different state representations incorporated different amounts of low and high-level information. The low-level information described the state of the GRNN and the state of the RL agent. The high-level information described expectations about the next event that would be observed. The RL agent learned over the course of 2,000 episodes. During each episode, the RL agent was exposed to 100 randomly ordered events. For the first 20 events, a perfect gating signal was used before the RL agent began learning. This allowed a reasonable average SSE to be computed before it was used as part of the RL agent's state representation. An overview of the experimental state representations and the learning parameters used by the RL agent to learn a policy for the given state representation can be seen in Table 1.

Each state representation consisted of between 1 and 4 features and always contained a feature describing the GRNN's current prediction error with respect to its historical prediction error, i.e. SSE_Ratio. Each dimension represented different information about the state of the overall system (Table 2):

- SSE_Ratio - the GRNN's current prediction error (i.e. SSE) with respect to a windowed average of the GRNN's historical SSE;
- Obs_Dist - the euclidean distance between two subsequent observations, $X_{t-1}$ and $X_{t}$;

Table 1: State Representations and Learning Parameters.

| State Representation | Learning Parameters |
| :--- | :--- |
| SSE_Ratio | $(\alpha=0.005 ; \gamma=0.95 ; \lambda=0.9)$ |
| SSE_Ratio+Obs_Dist | $(\alpha=0.005 ; \gamma=0.9 ; \lambda=1.00)$ |
| SSE_Ratio+Next_Event | $(\alpha=0.001 ; \gamma=0.9 ; \lambda=0.75)$ |
| SSE_Ratio+Last_Gate | $(\alpha=0.005 ; \gamma=0.95 ; \lambda=0.8)$ |
| SSE_Ratio+Obs_Dist+Last_Gate | $(\alpha=0.005 ; \gamma=0.9 ; \lambda=0.8)$ |
| SSE_Ratio+Obs_Dist+Next_Event | $(\alpha=0.001 ; \gamma=0.85 ; \lambda=0.95)$ |
| SSE_Ratio+Last_Gate+Next_Event | $(\alpha=0.005 ; \gamma=0.95 ; \lambda=0.95)$ |
| SSE_Ratio+Obs_Dist+ | $(\alpha=0.005 ; \gamma=0.9 ; \lambda=0.9)$ |
| Last_Gate+Next_Event |  |

## The Data

The training data set for the GRNN and, subsequently, the RL agent was the motion capture data used by Reynolds et al. (2007), which can be found at http://dcl.wustl.edu/stimuli.html.

The data set contains 3-dimensional motion captures of people performing 13 distinct tasks. Each motion capture lasted 3-4 seconds and contained between 10 and 13 observations. Each motion capture activity was considered to be one event. Each event observation consisted of 18 (x, y, z) points on the body. We preprocessed each observation following Reynolds et al. (2007); the origin of the coordinate frame was transformed such that the points corresponding to person's hip was the origin, all values were scaled to the range $[-1,1]$, and the orientation of each figure was altered such that it was the same across all events.

Following Reynolds et al. (2007), before each training run for the GRNN or the RL agent, the training set was created by randomly ordering the events from the set of 13 events. A new event was randomly selected and added to the training set until the GRNN reached asymptotic performance. This allowed the GRNN and the RL agent to observe each event multiple times and learn a good predictive model for the frames that fell within a given event. The random ordering of the events provided the learning algorithm with a large variety of transition examples. The same process was used to create the training set of the RL agent, but with a stopping condition of the combined GRNN and RL agent having observed a prespecified number of events.

## Results

The results show that it is possible to use reinforcement learning to identify true event boundaries. Furthermore, it is possible to learn a policy for controlling the GRNN gating mechanism without encoding any knowledge within the reward function about where event boundaries actually exist. This is important, because it provides evidence demonstrating that it is possible for an artificial agent to take the first steps towards learning complex concepts using a bottom-up approach. Additionally, it provides evidence that it is possible for an artificial agent to learn on its own without requiring the painstaking process of handcoding thresholds and decision boundaries on the part of a human.

Table 3 shows the results from training the RL agent with the eight different state representations. Dist. describes the average distance between when the RL agent chose to update the event representation and the closest true event boundary. Reward is the total reward received by the agent during the episode. Err. describes the GRNN's average SSE over the course of the episode. Each value in Table 3 is averaged across 50 independent runs.

For each state representation, it was possible to learn a policy by which the RL agent could control the GRNN's gating mechanism. The learning curves for each state representation can be seen in Figure 2. Each of the state representations was

Table 3: Experimental results after 1000 episodes averaged over 50 runs.

| State | Dist. | Reward | Err. |
| :--- | :--- | :--- | :--- |
| SSE_Ratio | 0.2 | -12.0 | 0.15 |
| SSE_Ratio+Obs_Dist | 0.13 | -30.72 | 0.18 |
| SSE_Ratio+Next_Event | 0.52 | -31.44 | 0.18 |
| SSE_Ratio+Last_Gate | 1.22 | -65.78 | 0.38 |
| SSE_Ratio+Obs_Dist+Last_Gate | 1.16 | -42.31 | 0.25 |
| SSE_Ratio+Obs_Dist+Next_Event | 0.76 | -48.69 | 0.28 |
| SSE_Ratio+Last_Gate+Next_Event | 1.1 | -103.07 | 0.3 |
| SSE_Ratio+Obs_Dist+ | 0.32 | -52.56 | 0.5 |
| Last_Gate+Next_Event |  |  |  |

able to achieve an average reward between $0.15-0.5$ over the course of 1000 episodes. When run using a perfect gating function, the GRNN is able to achieve an average prediction error within the same range achieved by the RL agent. It is of note that this improvement in performance over that reported by Reynolds et al. (2007) is due, in part, to advances in deep neural network computing libraries. The state representation SSE_Ratio+Obs_Dist was able to learn a policy best able to maximize its received rewards. This indicates that information about the ratio of the current SSE to the average observed SSE and the degree of change between two subsequent observations are critical features for deciding how and when to update the GRNN's event representation.


Figure 2: GRNN Model Architecture (Reynolds et al., 2007).

The results show that the RL agent was able to learn a policy for each state representation. The agent was able to reach a reasonable average distance from the true event boundary, approximately 1.5 frames for about half of the state representations. We considered the RL agent's ability to identify an event boundary within 0.67 of the true event boundary to be
a reasonable level of performance given that each event lasts for 11 frames on average.

Finding that it is possible for a RL agent to learn when one event ends and another begins using GRNN's the SSE_Ratio alone, while surprising, is encouraging, as an initial step towards learning to unitize continuous observations without the use of higher level information (i.e. Next_Event). It is possible that the SSE_Ratio feature was so powerful on its own because it is correlated with and related to the other lowerlevel features (i.e. Obs_Dist and Last_Gate). However, it is not surprising that the SSE_Ratio+Obs_Dist both resulted in a high performing policy and the best performaning policy given the evidence in the literature suggesting that the degree of change between two subsquent observations plays a large role in segmenting continuous events and detecting boundary points on physical objects (Newtson, 1976; Gibson, 1979) is considered.

The ability to correctly identify event boundaries does not always have a consistent effect on the GRNNs observed prediction error. This finding indicates that it is not the number of boundaries that are correctly identified that is important, but rather which boundaries are correctly identified. Appropriately handling sub-event boundaries could drive down error in the GRNN while causing the RL agent to trigger gates at non-event boundaries, thus increasing the average distance measure. For example, the SSE_Ratio state representation results in an agent that is better able to detect event boundaries than the SSE_Ratio+Next_Event state represention, but the SSE_Ratio+Next_Event results in more rewards and lower average prediction errors in the GRNN.

## Future Work

That the RL agent was able to learn a policy for controlling the gating mechanism based solely on the GRNN's prediction error supports the potential for prediction error to play a primary role in event segmentation. We hypothesize that prediction error is likely to play an important role in other aspects of complex event segmentation and, possibly, event cognition. For example, if a person is unable to predict the event he/she will observe, then the higher level expectation failures might be propagated back to the segmentation gating mechanism and alter how it is identifying event boundaries. The ability of the agent to learn a gate controlling policy given a state representation that includes higher level expectations (i.e. the likelihood that the current event will transition into a standing event), indicates that the combined GRNN and RL agent model should be able to segment continuous observations into discrete events such that the discrete events are maximally predictive based on higher level expectation errors. We intend to study this in future work.

The research in this paper represents a step towards modeling how an intelligent agent can reason about and manipulate its model of the world in order to develop meaningful representations in an unsupervised way. Given that our system can recognize event boundaries, the next step is to develop a
system that is able to recognize sub-events. We believe that extending our approach to learning a policy for segmenting an event into sub-events will depend on two parts: (1) allowing the RL agent to more finely control the amount of influence each observation has on the subsequent event representation and (2) learning prototypical representations of the events. Giving the RL agent more fine-grained control over the influence incoming observations have on the current event representation should allow the agent to account for sub-events within longer, more complex events. Additionally, by learning the sequential dependencies among the event representations, it should be possible to go beyond identifying event boundaries to predicting which event will be observed next. For example, given an observation that a person is currently seated, represented in the form of the GRNN+RL agents event representation and the learned prototypical agent, it should be possible to predict the likelihood that the person will stand up.

## Conclusion

This paper has presented a model for learning to segment continuous observations into event units. Additionally, our model is able to learn to identify boundary points without any prior knowledge. The combined GRNN and RL agent proposed in this paper represents an approach to modeling event segmentation that removes the limitation of externally set thresholds and is able to operate in continuous domains. Experimental results support our conclusion that it is possible to use RL to learn a gate controlling mechanism that is able to accurately identify event boundaries independently without incorporating knowledge about the location of event boundaries in the reward function.

## References

Braver, T. S., \& Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. Control of cognitive processes: Attention and performance XVIII, 713-737.
Brooks, R. A. (1995). Intelligence without reason. The artificial life route to artificial intelligence: Building embodied, situated agents, 25-81.
Cleeremans, A., \& McClelland, J. L. (1991). Learning the structure of event sequences. Journal of Experimental Psychology: General, 120(3), 235.
Cohen, P., \& Adams, N. (2001). An algorithm for segmenting categorical time series into meaningful episodes. In International symposium on intelligent data analysis (pp. 198-207).
Cohen, P., Oates, T., Atkin, M. S., \& Beal, C. R. (1996). Building a baby. In Proceedings of the eighteenth annual conference of the cognitive science society (pp. 518-522).
Elman, J. L. (1990). Finding structure in time. Cognitive science, 14(2), 179-211.
Enns, J. T., \& Lleras, A. (2008). What's next? new evidence for prediction in human vision. Trends in cognitive sciences, 12(9), 327-333.

Gibson, J. J. (1979). The ecological approach to visual perception: classic edition. Psychology Press.
Hochreiter, S., \& Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
Johnson, M. (1987). The body in the mind: The bodily basis of imagination, reason, and meaning. The body in the mind: the bodily basis of imagination, reason and meaning.
Kurby, C. A., \& Zacks, J. M. (2008). Segmentation in the perception and memory of events. Trends in cognitive sciences, 12(2), 72-79.
Lakoff, G., \& Johnson, M. (2008). Metaphors we live by. University of Chicago press.
Mandler, J. M. (1988). How to build a baby: On the development of an accessible representational system. Cognitive Development, 3(2), 113-136.
Mandler, J. M. (1992). How to build a baby: Ii. conceptual primitives. Psychological review, 99(4), 587.
Newtson, D. (1976). Foundations of attribution: The perception of ongoing behavior. New directions in attribution research, 1, 223-247.
Niv, Y., \& Schoenbaum, G. (2008). Dialogues on prediction errors. Trends in cognitive sciences, 12(7), 265-272.
Radvansky, G. A., \& Zacks, J. M. (2014). Event cognition. Oxford University Press.
Rao, R. P., \& Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extraclassical receptive-field effects. Nature neuroscience, 2(1), 79-87.
Reynolds, J. R., Zacks, J. M., \& Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. Cognitive Science, 31(4), 613-643.
Ross, R. A., \& Baldwin, D. A. (2015). Event processing as an executive enterprise. In Emerging trends in the social and behavioral sciences. John Wiley and Sons, Inc.
Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., \& O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. Proceedings of the National Academy of Sciences of the United States of America, 102(20), 7338-7343.
Sutton, R. S., \& Barto, A. G. (2015). Reinforcement learning: An introduction (Vol. 1) (No. 1). MIT press Cambridge.
Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., \& Reynolds, J. R. (2007). Event perception: a mind-brain perspective. Psychological bulletin, 133(2), 273.

# Objections to Computationalism. A Short Survey 

Marcin Miłkowski (mmilkows@ifispan.waw.pl)<br>Instytut Filozofii i Socjologii PAN, ul. Nowy Świat 72, 00-330 Warszawa, Poland


#### Abstract

In this paper, I review the objections against the claim that brains are computers, or, to be precise, informationprocessing mechanisms. By showing that practically all the popular objections are based on uncharitable (or simply incorrect) interpretations of the claim, I argue that the claim is likely to be true, relevant to contemporary cognitive (neuro)science, and non-trivial.


Keywords: computationalism; computational theory of mind; representation; computation; modeling

## Computationalism and Objections

The computational theory of mind, or computationalism, has been fruitful in cognitive research. The main tenet of the computational theory of mind is that the brain is a kind of information-processing mechanism, and that informationprocessing is necessary for cognition; it is non-trivial and is generally accepted in cognitive science. The positive view will not be developed here, in particular the account of physical computation, because it has already been elucidated in book-length accounts (Fresco, 2014; Miłkowski, 2013; Piccinini, 2015). Instead, a review of objections is offered here, as no comprehensive survey is available.

The survey suggests that the majority of objections fail just because they make computationalism a straw man. Some of them, however, have shown that stronger versions of the computational theory of mind are untenable, as well. Historically, they have helped to shape the theory and methodology of computational modeling. In particular, a number of objections show that cognitive systems are not only computers, or that computation is not the sole condition of cognition; no objection, however, establishes that there might be cognition without computation.

## Computer metaphor is just a metaphor

Computational descriptions are sometimes described as $a$ computer metaphor (cf., e.g., Ekman, 2003; Karl, 2012, p. 2101). The use of the term suggests that the proposed description is rough and highly idealized, and cannot be treated literally. However, by using the term, others suggest that no computational model may be treated seriously; all are mere metaphors (Daugman, 1990).

A defender of computationalism might concede this and weaken their position. But the position is also tenable in the stronger version. This is because computer metaphors cannot really be tested and rejected, whereas computational models can. For this reason, in this paper, I will adopt, along with other theorists (Newell \& Simon, 1972, p. 5; Pylyshyn,

1984, pp. xiv-xvi), a stronger version of computationalism, which claims that cognition literally involves computation.

## Software is not in the head

This objection is that there is no simple way to understand the notions of software and hardware as applied to biological brains. But the software/hardware distinction, popular as in the slogan "the mind to the brain is like the software to hardware" (Block, 1995; Piccinini, 2010), need not be applicable to brains at all for computationalism to be true. There are non-program-controllable computers: they do not load programs from external memory to internal memory in order to execute them. A mundane example of such a computer is a logical AND gate. In other words, while it may be interesting to inquire whether there is software in the brain, even if there were none, computationalism could still be true.

## Computers are just for number-crunching

Another intuitive objection, already stated (and defeated) in the 1950 s , is that brains are not engaged in numbercrunching, while computers compute over numbers. But if this is all computers do, then they don't control missiles or send documents to printers. After all, printing is not just number crunching. The objection rests therefore on a mistaken assumption that computers can only compute numerical functions. Computer functions can be defined not only of integer numbers but also of arbitrary symbols (Newell, 1980), and as physical mechanisms, computers can also control other physical processes.

## Computers are abstract entities

Some claim that because symbols in computers are, in some sense, abstract and formal, computers-or at least computer programs-are abstract as well (Barrett, 2015; Barrett, Pollet, \& Stulp, 2014; Lakoff, 1987). In other words, the opponents of computationalism claim that it implies ontological dualism (Searle, 1990). However, computers are physical mechanisms, and they can be broken, set on fire etc. These things may be difficult to accomplish with a collection of abstract entities. Computers are not just symbol-manipulators. They do things, and some of the things computers do are not computational. In this minimal sense, computers are physically embodied, not unlike mammal brains. It is, however, a completely different matter whether the symbols in computers mean anything.

## People are organisms, computers are not

Barrett (2015), among others, also presses the point that people are organisms. It's trivially true but irrelevant:
physical computers are physical, and they may be built in various ways. A computer may be built of DNA strands (Zauner \& Conrad, 1996), so why claim that it's metaphysically impossible to have a biological computer?

## Symbols in computers mean nothing

One of the most powerful objections formulated against the possibility of Artificial Intelligence is associated with John Searle's Chinese Room thought experiment (Searle, 1980). Searle claimed to show that running a computer program is not sufficient for semantic properties to arise, and this was in clear contradiction to what was advanced by proponents of Artificial Intelligence, who assumed that it was sufficient to simulate the syntactic structure of representations for the semantic properties to appear. As John Haugeland quipped: "if you take care of syntax, the semantics will take care of itself" (Haugeland, 1985, p. 106). But Searle replied: one can easily imagine a person with a special set of instructions in English who could manipulate Chinese symbols and answer questions in Chinese without understanding it at all. Hence, understanding is not reducible to syntactic manipulation. While the discussion around this thought experiment is hardly conclusive (Preston \& Bishop, 2002), the problem was soon reformulated by Stevan Harnad (1990) as "the symbol grounding problem" (SGP): How can symbols in computational machines mean anything?

If the SGP makes sense, then one cannot simply assume that symbols in computers mean something just by being parts of computers, or at least they cannot mean anything outside the computer so easily (even if they contain instructional information (Fresco \& Wolf, 2013)). Representational properties do not necessarily exist in physical computational mechanisms (Egan, 1995; Fresco, 2010; Miłkowski, 2013; Piccinini, 2008). So, even if Searle is right and there is no semantics in computers, the brain might still be a computer, as computers need no semantics to be computers. Perhaps something additional to computation is required for semantics.

There is an important connection between the computational theory of mind and the representational account of cognition: they are more attractive when both are embraced. Cognitive science frequently explains cognitive phenomena by referring to semantic properties of mechanisms capable of information-processing (Shagrir, 2010a). Brains are assumed to model reality, and these models can be utilized in computations. While this seems plausible to many, one can remain computationalist without assuming representationalism (the claim that cognition requires cognitive representation). At the same time, a plausible account of cognitive representation cannot be couched merely in computational terms as long as one assumes that the symbol grounding problem makes sense at least for some computers. To make the account plausible, most theorists appeal to notions of teleological function and semantic information (Bickhard, 2008; Cummins \& Roth, 2012; Dretske, 1986; Millikan, 1984), which are not
technical terms of computability theory, neither can they be reduced to such. However, processing of semantic information is still processing of information; hence, computation is necessary for manipulation of cognitive representation.

Computationalism was strongly connected to cognitive representations by the fact that it offered a solution to the problem of what makes meaning causally relevant. Many theorists claim that because the syntax in computer programs is causally relevant (or efficacious), so is the meaning. While the wholesale reduction of meaning to syntax is implausible, the computational theory of mind makes it clear that the answer to the question includes the causal role of the syntax of computational vehicles. Still, the fact that it does not offer a naturalistic account of meaning is not an objection to computationalism itself. That would indeed be too much. At the same time, at least some naturalistic accounts, such as Millikan's and Dretske's, can be used to solve the SGP (see Miłkowski 2013, chap. 4).

## Computers can only represent with all detail

The debate over meaning in computers and animals abounds in red herrings, however. One recent example is Robert Epstein's (2016) popular essay. His most striking mistake is the assumption that computers always represent everything with arbitrary accuracy. Epstein cites the example of how people remember a dollar bill, and assumes that computers would represent it in a photographic manner with all available detail. This is an obvious mistake: representation is useful mostly when it does not convey information about all properties of the represented target. If Epstein is correct, then there are no JPEG files in computers, as they are not accurate, because they are based on lossy compression. Moreover, no assumption of the computational theory of mind says that memory should be understood in terms of the von Neumann architecture, and it is controversial to suggest that it should (Gallistel \& King, 2010).

## People don't process information

Ecological psychologists stress that people do not process information, they just pick it up from the environment (cf. Chemero, 2003; Gibson, 1986). Thus, to understand this, one should make more explicit the meaning of information processing in the computational theory of mind. What kind of information is processed? The information in question need not be semantic, as not all symbols in computers are about something. The minimal notion that could suffice for our purposes is one of structural information: a vehicle can bear structural information in the event that it has at least one degree of freedom, that is, it may vary its state (MacKay, 1969). The number of degrees of freedom, or yesno questions required to exactly describe its current state, is the amount of structural information. As long as there are vehicles with multiple degrees of freedom and they are part of causal processes that cause some other vehicles-just like some models of computation describe these processes
(Miłkowski, 2014)-there is information processing. This is a very broad notion, as all physical causation implies information transfer and processing in this sense (Collier, 1999).

The Gibsonian notion of information pickup requires vehicles of structural information as well. There needs to be some information out there to be picked up, and organisms have to be structured so as to be able to change their state in response to information. Gibsonians could, however, claim that the information is not processed. It is unclear what is meant by this: for example, Chemero seems to imply that processing amounts to adding more and more layers of information, like in Marr's account of vision (Chemero, 2003, p. 584; cf. Marr, 1982). But information processing need not require multiple stages of adding more information. To sum up: the Gibsonian account does not invalidate computationalism at all.

## Consciousness is not computational

Some find (some kinds of) consciousness to be utterly incompatible with computationalism, or at least, unexplainable in purely computational terms (Chalmers, 1996). The argument is probably due to Leibniz's thought experiment in Monadology (Leibniz, 1991). Imagine a brain as huge as a mill, and enter it. Nowhere in the interplay of gears could you find perceptions, or qualitative consciousness. Hence, you cannot explain perception mechanically. Of course, this Leibnizian argument appeals only to some physical features of mechanisms, but some still seem to think that causation has nothing to do with qualitative consciousness.

The argument, if cogent, is applicable more broadly, not just to computationalism; it is supposed to defeat reductive physicalism or materialism. For this reason, this objection might be dismissed as attacking any scientific project that explains consciousness reductively.

Virtually all current theories of consciousness are computational, even the ones that appeal to quantum processes (Hameroff, 2007). For example, Bernard Baars offers a computational account in terms of the global workspace theory (Baars, 1988; cf. also Dennett, 2005), David Rosenthal gives an account in terms of higher-level states (cf. Cleeremans, 2005; Rosenthal, 2005), and Giulio Tononi explains in terms of minimal information integration (Tononi, 2004). Is there any theory of consciousness that is not already computational?

John Searle, however, suggests that only a noncomputational theory of consciousness can succeed. His claim is that consciousness is utterly biological (Searle, 1992). How does this contradict computationalism given that there might be biological computers? Moreover, Searle fails to identify the specific biological powers of brains that make them conscious. He just passes the buck to neuroscience, which often offers computational accounts.

## Computer models ignore time

Proponents of dynamical accounts of cognition stress that Turing machines do not operate in real time. This means that this classical model of computation does not appeal to real time; instead, it operates with the abstract notion of a computation step. There is no continuous time flow, just discrete clock ticks in a Turing Machine (Bickhard \& Terveen, 1995; Wheeler, 2005). This is true. But is this an objection against computationalism?

First, some models of computation appeal to real time (Nagy \& Akl, 2011), so one could use such a formalism. Second, the objection seems to confuse the formal model of computation with its physical realization. Physical computers operate in real time, and not all models of computation are made equal; some will be relevant to the explanation of cognition, and some may only be useful for computability theory. A mechanistically-adequate model of computation that describes all relevant causal processes in the mechanism is required for explanatory purposes (Miłkowski, 2014).

## Brains are not digital computers

Universal Turing machines are crucial to computability theory. One could, however, maintain that brains are not digital computers (Edelman, 1992; Lupyan, 2013).

But computationalism can appeal to models of analog computation (e.g., Siegelmann, 1994), or even more complex kinds of computation (Piccinini \& Bahar, 2013), if required. These models are still understood as computational in computability theory, and some theorists indeed claim that the brain is an analog computer, which is supposed to allow them to compute Turing-incomputable functions. Thus, one cannot dismiss all kinds of computationalism by saying that the brain is not a digital computer. There are analog computers, and an early model of a neural network, Perceptron, was analog (Rosenblatt, 1958). The contention that computers have to be digital is just dogmatic.

## Genuine artificial intelligence is impossible

There are a number of arguments of a form:
People $\psi$.
Computers will never $\psi$.
So, artificial intelligence is impossible (or computationalism is false).

This argument is enthymematic, but the conclusion follows with a third assumption: if artificial intelligence is possible, then computers will $\psi$. The plausibility of the argument varies from case to case, depending on what you fill for $\psi$. For years, it was argued that winning in chess is $\psi$ (Dreyfus, 1979), but it turned out to be false. So, unless there is a formal proof, it's difficult to treat premise 2 seriously.

What could be plausibly substituted for $\psi$ ? There are many properties of biological organisms that simply seem irrelevant to this argument, including exactly the same energy consumption, having proper names, spatiotemporal location, etc. The plausible candidate for substitution is some capacity for information-processing. If there is such a human capacity that computers do not possess, then the argument is indeed cogent.
Only people can see the truth A classical anticomputational argument points to the human ability to recognize the truth of logical statements that cannot be proven by a computer (Lucas, 1961; Penrose, 1989). It is based on the alleged ability of human beings to understand that some statements are true, which is purportedly impossible for machines (this argument is based on the Gödel proof of incompleteness of the first-order predicate calculus with basic arithmetic). The problem is that this human understanding has to be non-contradictory and certain. But Gödel has shown that in general it cannot be decided whether a given system is contradictory or not. So either it's mathematically certain that human understanding of mathematics is non-contradictory, which makes the argument inconsistent as it cannot be mathematically certain because it's undecidable; or the argument just assumes noncontradiction of human understanding, which makes the argument unsound because people make contradictions unknowingly (Krajewski, 2007; Putnam, 1960).

Common sense cannot be formalized Another similar argument points to common sense, which is a particularly difficult capacity. The trouble with implementing common sense on machines is sometimes called (somewhat misleadingly, cf. (Shanahan, 1997)) the frame problem (Dreyfus, 1972, 1979; Wheeler, 2005). Inferential capacities of standard AI programs do not seem to follow the practices known to humans, and that was supposed to hinder progress in such fields as high-quality machine translation (BarHillel, 1964), speech recognition (held to be immoral to fund (Weizenbaum, 1976)), and so on. Even if IBM Watson wins in Jeopardy!, one may still think it's not enough. Admittedly, common sense is a plausible candidate in this argument.

Even if the proponent of computationalism need not require that genuine AI be based on a computer simulation of human cognitive processes, he or she still must show that human common sense can be simulated on a computer. Whether it can or not is still a matter of debate.

## Computers are everywhere

At least some plausible theories of physical implementation of computation lead to the conclusion that all physical entities are computational (this stance is called pancomputationalism, (cf. Müller, 2009)). If this is the case, then the computational theory of mind is indeed trivial, as not only brains are computational, but also cows, black holes, cheese sandwiches etc. are all computers. However, a pancomputationalist may reply by saying that there are
different kinds (and levels) of computation, and brains do not execute all kinds of computation at the same time (Miłkowski, 2007). So not just any computation but some non-trivial kind of computation is specific to brains. Only the kind of pancomputationalism that assumes that everything computes all kinds of functions at the same time is catastrophic, as it makes physical computation indeed trivial (Putnam, 1991; Searle, 1992).

## There are no computers

Another more radical move is to say that computers do not really exist; they are just in the eyes of beholder. According to John Searle, the beholder decides whether a given physical system is computational, and therefore may make this decision for virtually everything. Nothing intrinsically is a computer. But the body of work on physical computation in the last decade or so has been focused on showing why Putnam and Searle were wrong in some sense (Chalmers, 2011; Chrisley, 1994; Copeland, 1996; Miłkowski, 2013; Piccinini, 2015; Scheutz, 1996; Shagrir, 2010b). The contemporary consensus is that computational models can be used to adequately describe causal connections in physical systems, and that these models can also be falsely ascribed. In other words, computational models are not different in kind from any mathematical model used in science. If they are mere subjective metaphors and don't describe reality, then mathematical models in physics are subjective as well (McDermott, 2001).

Intuitively, arguments presented by Searle and Putnam are wrong for a very simple reason: why buy a new computer instead of ascribing new software to the old one? We know that such ascriptions would be extremely cumbersome. Therefore, there must be a flaw in such arguments, and even if the technicalities involved are indeed interesting, they fail to establish a conclusion.

## Conclusion

In this paper, I have listed and summarized a number of arguments against computationalism. The only objection that seems to be plausible at first glance is the one stating that common sense is impossible or extremely difficult to implement on a machine. However, more and more commonsensical capacities are being implemented on machines.

The point is that there's no good reason to think that the brain is not a computer. But it isn't a mere computer: It is physically embedded in its environment and interacts physically with its body, and for that, it also needs a peripheral nervous system (Aranyosi, 2013) and cognitive representations. Yet there's nothing that denies computationalism here. Most criticisms of computationalism therefore fail, and sticking to them is probably a matter of ideology rather than rational debate.

## Acknowledgments

The work on this paper was funded by a National Science Centre (Poland) research grant under the decision DEC2014/14/E/HS1/00803. The author wishes to thank Tomasz Korbak, Martin Hughes, Panu Raatikainen, Błażej Skrzypulec, and anonymous reviewers for their comments.

## References

Aranyosi, I. (2013). The peripheral mind: philosophy of mind and the peripheral nervous system. New York, NY: Oxford University Press.
Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge / New York: Cambridge University Press.
Bar-Hillel, Y. (1964). A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation. In Language and Information (pp. 174-179). Reading, Mass.: Addison-Wesley.
Barrett, L. (2015). Why Brains Are Not Computers, Why Behaviorism Is Not Satanism, and Why Dolphins Are Not Aquatic Apes. The Behavior Analyst, 1-15. https://doi.org/10.1007/s40614-015-0047-0
Barrett, L., Pollet, T. V., \& Stulp, G. (2014). From computers to cultivation: reconceptualizing evolutionary psychology. Frontiers in Psychology, 5, 867-867. https://doi.org/10.3389/fpsyg.2014.00867
Bickhard, M. H. (2008). The interactivist model. Synthese, 166(3), 547-591. https://doi.org/10.1007/s11229-008-9375-x
Bickhard, M. H., \& Terveen, L. (1995). Foundational issues in artificial intelligence and cognitive science: Impasse and solution. North-Holland.
Block, N. (1995). The mind as the software of the brain. In D. Osherson, L. Gleitman, \& S. Kosslyn (Eds.), An Invitation to Cognitive Science. Cambridge, Mass.: MIT Press.
Chalmers, D. J. (1996). The conscious mind: in search of a fundamental theory. New York: Oxford University Press.
Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition. Journal of Cognitive Science, (12), 325-359.
Chemero, A. (2003). Information for perception and information processing. Minds and Machines, 13, 577588.

Chrisley, R. L. (1994). Why everything doesn't realize every computation. Minds and Machines, 4(4), 403-420. https://doi.org/10.1007/BF00974167
Cleeremans, A. (2005). Computational correlates of consciousness. Progress in Brain Research, 150, 81-98. https://doi.org/10.1016/S0079-6123(05)50007-4
Collier, J. D. (1999). Causation is the transfer of information. In H. Sankey (Ed.), Causation, natural laws and explanation (pp. 279-331). Dordrecht: Kluwer.
Copeland, B. J. (1996). What is computation? Synthese, 108(3), 335-359.
Cummins, R., \& Roth, M. (2012). Meaning and Content in Cognitive Science. In R. Schantz (Ed.), Prospects for Meaning (pp. 365-382). Berlin \& New York: de Gruyter.

Daugman, J. (1990). Brain metaphor and brain theory. In E. L. Schwartz (Ed.), Computational Neuroscience (pp. 918). Cambridge, Mass: MIT Press.

Dennett, D. C. (2005). Sweet Dreams. Philosophical Obstacles to a Science of Consciousness. Cambridge, Mass.: MIT Press.
Dretske, F. I. (1986). Misrepresentation. In R. Bogdan (Ed.), Belief: form, content, and function (pp. 17-37). Oxford: Clarendon Press.
Dreyfus, H. (1972). What Computers Can't Do: A Critique of Artificial Reason. New York: Harper \& Row, Publishers.
Dreyfus, H. (1979). What computers still can't do: a critique of artificial reason. Cambridge Mass.: MIT Press.
Edelman, G. M. (1992). Bright air, brilliant fire: on the matter of the mind. New York, N.Y.: BasicBooks.
Egan, F. (1995). Computation and Content. The Philosophical Review, 104(2), 181-181. https://doi.org/10.2307/2185977
Ekman, P. (2003). Emotions revealed: recognizing faces and feelings to improve communication and emotional life. New York: Times Books.
Epstein, R. (2016, May 18). The empty brain. Retrieved December 28, 2016, from https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-acomputer
Fresco, N. (2010). Explaining Computation Without Semantics: Keeping it Simple. Minds and Machines, 20(2), 165-181. https://doi.org/10.1007/s11023-010-9199-6
Fresco, N. (2014). Physical Computation and Cognitive Science. Berlin, Heidelberg: Springer Berlin Heidelberg.
Fresco, N., \& Wolf, M. J. (2013). The instructional information processing account of digital computation. Synthese, 191(7), 1469-1492. https://doi.org/10.1007/s11229-013-0338-5
Gallistel, C. R., \& King, A. P. (2010). Memory and the Computational Brain. Chichester: Wiley-Blackwell.
Gibson, J. J. (1986). The Ecological Approach to Visual Perception. Hove: Psychology Press.
Hameroff, S. R. (2007). The Brain Is Both Neurocomputer and Quantum Computer. Cognitive Science, 31, 10351045.

Harnad, S. (1990). The symbol grounding problem. Physica D, 42, 335-346.
Haugeland, J. (1985). Artificial intelligence: the very idea. Cambridge, Mass.: MIT Press.
Karl, F. (2012). A Free Energy Principle for Biological Systems. Entropy (Basel, Switzerland), 14(11), 21002121. https://doi.org/10.3390/e14112100

Krajewski, S. (2007). On Gödel's Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in any Attempt to "Out-Gödel" the Mechanist. Fundamenta Informaticae, 81(1), 173-181.
Lakoff, G. (1987). Women, fire, and dangerous things: what categories reveal about the mind. Chicago: University of Chicago Press.

Leibniz, G. W. (1991). The monadology. (R. Latta, Trans.). Raleigh, N.C.; Boulder, Colo.: Alex Catalogue; NetLibrary.
Lucas, J. (1961). Minds, Machines and Gödel. Philosophy, 9(3), 219-227.
Lupyan, G. (2013). The difficulties of executing simple algorithms: Why brains make mistakes computers don't. Cognition, 129(3), 615-36. https://doi.org/10.1016/j.cognition.2013.08.015
MacKay, D. M. (1969). Information, mechanism and meaning. Cambridge: M.I.T. Press.
Marr, D. (1982). Vision. A Computational Investigation into the Human Representation and Processing of Visual Information. New York: W. H. Freeman and Company.
McDermott, D. V. (2001). Mind and Mechanism. Cambridge, Mass.: MIT Press.
Miłkowski, M. (2007). Is computationalism trivial? In G. D. Crnkovic \& S. Stuart (Eds.), Computation, Information, Cognition - The Nexus and the Liminal (pp. 236-246). Newcastle: Cambridge Scholars Press.
Miłkowski, M. (2013). Explaining the Computational Mind. Cambridge, Mass.: MIT Press.
Miłkowski, M. (2014). Computational Mechanisms and Models of Computation. Philosophia Scientae, 18(18-3), 215-228. https://doi.org/10.4000/philosophiascientiae. 1019
Millikan, R. G. (1984). Language, thought, and other biological categories: new foundations for realism. Cambridge, Mass.: The MIT Press.
Müller, V. C. (2009). Pancomputationalism: Theory or metaphor? In R. Hagengruber (Ed.), The relevance of philosophy for information science. Berlin: Springer.
Nagy, N., \& Akl, S. (2011). Computations with Uncertain Time Constraints: Effects on Parallelism and Universality. In C. Calude, J. Kari, I. Petre, \& G. Rozenberg (Eds.), Unconventional Computation (Vol. 6714, pp. 152-163). Springer Berlin / Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-21341-0_19
Newell, A. (1980). Physical symbol systems. Cognitive Science: A Multidisciplinary Journal, 4(2), 135-183. https://doi.org/10.1207/s $15516709 \operatorname{cog} 0402 \_2$
Newell, A., \& Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Penrose, R. (1989). The emperor's new mind. London: Oxford University Press.
Piccinini, G. (2008). Computation without Representation. Philosophical Studies, 137(2), 205-241. https://doi.org/10.1007/s1 1098-005-5385-4
Piccinini, G. (2010). The Mind as Neural Software? Understanding Functionalism, Computationalism, and Computational Functionalism. Philosophy and Phenomenological Research, 81(2), 269-311. https://doi.org/10.1111/j.1933-1592.2010.00356.x
Piccinini, G. (2015). Physical computation: a mechanistic account. Oxford: Oxford University Press.

Piccinini, G., \& Bahar, S. (2013). Neural computation and the computational theory of cognition. Cognitive Science, 37(3), 453-88. https://doi.org/10.1111/cogs. 12012
Preston, J., \& Bishop, M. (2002). Views into the Chinese room: new essays on Searle and artificial intelligence. Oxford; New York: Clarendon Press.
Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), Dimensions of Mind. New York University Press.
Putnam, H. (1991). Representation and Reality. Cambridge, Mass.: The MIT Press.
Pylyshyn, Z. W. (1984). Computation and cognition: Toward a foundation for cognitive science. Cambridge, Mass.: MIT Press.
Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408. https://doi.org/10.1037/h0042519
Rosenthal, D. (2005). Consciousness and mind. Oxford, New York: Oxford University Press.
Scheutz, M. (1996). When Physical Systems Realize Functions.... Minds and Machines, 9(2), 1-34. https://doi.org/10.1023/A:1008364332419
Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 1-19. https://doi.org/10.1017/S0140525X00005756
Searle, J. R. (1990). Is the Brain's Mind a Computer Program? Scientific American, (January), 26-31.
Searle, J. R. (1992). The Rediscovery of the Mind. Cambridge, Mass.: MIT Press.
Shagrir, O. (2010a). Brains as analog-model computers. Studies In History and Philosophy of Science Part A, 41(3), 271-279. https://doi.org/10.1016/j.shpsa.2010.07.007
Shagrir, O. (2010b). Towards a Modeling View of Computing. In G. Dodig-Crnkovic \& M. Burgin (Eds.), Information and Computation. Singapore: World Scientific Publishing.
Shanahan, M. (1997). Solving the frame problem: a mathematical investigation of the common sense law of inertia. Cambridge, Mass.: MIT Press.
Siegelmann, H. (1994). Analog computation via neural networks. Theoretical Computer Science, 131(2), 331360. https://doi.org/10.1016/0304-3975(94)90178-3

Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5(1). https://doi.org/10.1186/1471-2202-5-42
Weizenbaum, J. (1976). Computer power and human reason: from judgment to calculation. San Francisco: W.H. Freeman.

Wheeler, M. (2005). Reconstructing the Cognitive World. Cambridge, Mass.: MIT Press.
Zauner, K.-P., \& Conrad, M. (1996). Parallel computing with DNA: Toward the anti-universal machine. In H.-M. Voigt, W. Ebeling, I. Rechenberg, \& H.-P. Schwefel (Eds.), Parallel Problem Solving from Nature - PPSN IV (Vol. 1141, pp. 696-705). Springer Berlin / Heidelberg.

# A comparison between human micro-affordances and computational classification 

Arthur-Henri Michalland (Arthur-Henri.Michalland@univ-montp3.fr, Arthur.Michalland@lirmm.fr)<br>Epsylon Laboratory EA 4556, University Paul Valéry, Site Saint Charles, route de Mende<br>34190 Montpellier, France<br>LIRMM: Laboratory of Informatics, Robotics, Microelectronics of Montpellier, University of Montpellier<br>CNRS, 860 rue de St Priest<br>34095 Montpellier cedex 5<br>Denis Brouillet (Denis.Brouillet@univ-montp3.fr)<br>Epsylon Laboratory EA 4556, University Paul Valéry, Site Saint Charles, route de Mende 34190 Montpellier, France

Philippe Fraisse (Philippe.Fraisse@lirmm.fr)
LIRMM: Laboratory of Informatics, Robotics, Microelectronics of Montpellier, University of Montpellier CNRS, 860 rue de St Priest
34095 Montpellier cedex 5


#### Abstract

This study aimed to assess how specific components of an action could be selected by a simple computational system. We performed an experiment to test associations between grasps (precision or power grip) and several objects. We then ran simulations using a naive bayes classifier to study to what extent it could reproduce participants' choice. This classifier had two learning matrices containing objects' size associated with a grip by means of our experiment. When receiving a new object' size it computed the probability for each grip to be adapted. The highest probability was considered to represent which grip was associated with the object by the classifier. Results show that the classifier can reproduce participants' choice depending on the size of its learning matrices, and can quickly select the right type of grip for a majority of trials, showing that micro-affordances (Ellis \& Tucker, 2000) can be reproduced through naive bayesian classification.


Keywords: affordance; grip; bayesian method; classifier

## Introduction

As Leonard de Vinci said : "movement is principle of life". The way people interact with the world through body movements is indeed a corner stone of psychology, and especially of embodied psychology. As embodied psychology postulates that high-level cognitive processes are bodily rooted, or at least that their result depends on bodily states (Wilson, 2002), movements of the living body is a crucial point to attend. Yet how adapted body movements occur is not well determined and several propositions are made, one of them being particularly attractive for embodied cognitivists: theory of affordances (Gibson, 1979).

Affordance is a concept coined by Gibson (1979) that relies on direct perception. Although it has many interpretations, we will rely on the definition of Chemero (2003) in which affordances represent the relations between an animal's capacities and features of its environment. Abilities of an animal are functional properties, that depends on this animal's history.

This theory highlights the fact that voluntary actions are products of our perception of the situation, our abilities, and what we have learned. Moreover, this theory predicts that
action is part of objects' memories and perception, as it is now established (Brouillet et al., 2015), which is of interest for psychology and for robotics as they permit to gain insight into the perception-action loop (Montesano et al., 2007, 2008).

Yet, the link between perception and affordances needs further investigation, as clues, or features, need to be extracted from, or constructed on the basis of the environment. Such clues would facilitate the link between a rich perception and an adapted movement, and permit on line adaptation.

Our purpose was to test how adapted voluntary movements could be selected, by a very simple computational system, on the basis of clues extracted from perception. To do this we chose to test some specific components of an action: grasping movements (Koester, Schack, \& Westerholz, 2016). A lot of our interactions with the world depend on our ability to grasp things around us in a proper way, for example using a power or a precision grip (i.e. with all fingers of the hand or with the thumb and index, respectively, see Figure 1). These specific components of action (that doesn't include walking, reaching etc...) are termed micro-affordances by Ellis and Tucker (2000). These micro-affordances are supposed to emerge while looking at an object, and to facilitate a specific grasp. We selected object size to be the feature of the environment that could be associated to a specific grasp, in order to create a model that simulates a perceptually based motor activity.

The computational system we used to infer specific grasps rely on bayesian probability (Jones \& Love, 2011; Pearl, 1985). The bayesian approach appears to be promising when studying how humans can interact with the world in presence of uncertainty (Perfors et al., 2011). It can apply to motor planning and control, estimation of context and motor learning (Wolpert \& Ghahramani, 2000; Wolpert, Ghahramani, \& Flanagan, 2001), and can be easily used in its simplest ways (Robert, 2000). This approach rely on conditional probability and allows to determine the probability of a certain event (for example a particular grasp) knowing some information :
past experiences (e.g. earlier grasp in presence of an object) or sensory inputs (e.g. object size) (Naïm et al., 2007).

The particular model we chose is a naive bayes classifier. This model has two learning matrices : one containing the size of objects graspable with a power grip, and one containing the size of objects graspable with a precision grip, size being represented by three parameters $x, y$, and $z$. Once it has computed these matrices, it receives the size of a novel object to be classified as graspable with a power grip or with a precision grip. In order to do so, it selects the most probable grasp, knowing the object size, to be the grasp to produce in presence of this particular object.

This approach of micro-affordances as naive bayesian classification can be of interest for psychologists and roboticians, as it can reduce the size of ontology, or databases, needed for an adapted system, and permits to infer micro-affordances in a very simple way.

In a first part we present the experiment to test microaffordances with human beings and select objects that can be associated with a precision or a power grip. In a second part we explain how the model categorizes objects as being graspable with a precision or a power grip by means of naive bayesian classification, and show the results obtained with this model. We then compare human's and classifier's performances and discuss the possible developments of such applications.

## Selection and association of objects with a precision or a power grip

## Participants

Sixteen students were recruited for a pre-experiment in order to select the objects used in our experiment and simulation. Eighty students, different from the previous ones, were then recruited in order to select the appropriate grasp for each object (seven of them were not taken into account as they changed their grasping for the same objects between trials and differed drastically from the others). All participants freely signed a letter of consent, were right-handed, had normal or corrected to normal vision and over 18 years old, none had problems of motricity.

## Materials

Forty-four pictures of objects were used. Each picture was modified to have the object being centered, vertically oriented, and a half of their real size when displayed on the computer screen.

These images were presented to sixteen students in a preexperiment, with a hand near the object either making a power grip or a precision grip (see Figure 1). Participants had to indicate their level of agreement with the grip being displayed with the object. A high level of agreement with a grip meant that it was a reasonable grip to pick up and use the object.As a result, twenty objects were selected for the experiment, ten being graspable with a power grip and ten with a precision grip.


Figure 1: A hand making a power grip (left picture), and a precision grip (right picture).

## Procedure

All of eighty participants were received one by one in an experimentation room, and sat in front of a computer Lenovo 17.3" with graphics card AMD radeon HD 8500M. They were asked to grab, with their right hand, a device that constrained them to make either a power or a precision grip. They were instructed to look at the computer screen and make the more appropriate grip on the device when seeing an object displayed on the screen. The twenty objects were then displayed randomly. When the twenty objects had been exposed, a second random presentation was made, in order to ensure the grip selected by participants for each object.

## Results

Overall, the grips selected by means of the pre-experiment were respected, as shown in Table 1 and Table 2, and participants showed stable grip for each object. All of which allowed us to classify each object as associated with a precision or a power grip.

Table 1: Percentage of responses for objects associated with a precision grip, a number was attributed to each object for further comparison.

| Objects | N | \% power grip | \%precision grip |
| :--- | :---: | :---: | :---: |
| grain of wheat | 1 | 0.68 | 99.32 |
| tweezers | 2 | 3.42 | 96.58 |
| nut | 3 | 0.68 | 99.32 |
| radish | 4 | 10.96 | 89.04 |
| smart card | 5 | 1.37 | 98.63 |
| screw | 6 | 0.00 | 100.00 |
| paper clip | 7 | 0.00 | 100.00 |
| strawberry | 8 | 6.85 | 93.15 |
| french beans | 9 | 1.37 | 98.63 |
| key | 10 | 2.74 | 97.26 |

## Simulation with a naive bayes classifier

## The naive bayes classifier

The second step of this work was to put the naive bayes classifier to the test. To do so, we had to implement the size of objects used in our experiment. We chose to represent size in

Table 2: Percentage of responses for objects associated with a power grip, a number was attributed to each object for further comparison.

| Objects | N | \% power grip | \%precision grip |
| :--- | :---: | :---: | :---: |
| glass | 11 | 97.26 | 2.74 |
| hair clipper | 12 | 91.10 | 8.90 |
| coconut | 13 | 100.00 | 0.00 |
| apple | 14 | 99.32 | 0.68 |
| corn | 15 | 95.89 | 4.11 |
| computer mouse | 16 | 89.73 | 10.27 |
| board wiper | 17 | 92.47 | 7.53 |
| universal pliers | 18 | 95.21 | 4.79 |
| pepper | 19 | 95.21 | 4.79 |
| deodorant | 20 | 91.78 | 8.22 |

a three dimensional cartesian coordinate system, representing height, width, and depth.


Figure 2: A computer mouse mesured on $x, y$ and $z$.

Table 3: Mean and Variance for objects associated with a precision grip or a power grip.

| Objects | $x$ | Mean (Variance) |  |
| :--- | :---: | :---: | :---: |
|  | $x$ | $y$ | $z$ |
| precision | 1.265 | 0.62 | 4.87 |
|  | $(0.422)$ | $(0.291)$ | $(11.72)$ |
| power | 6.76 | 5.27 | 13.92 |
|  | $(3.83)$ | $(8.58)$ | $(22.94)$ |

We defined a rule to mesure our objects : z axis for the longest axis of the object, $y$ axis for the shortest axis of the object, and x the last one, following the right hand rule (e.g. mesure of a computer mouse in centimeter: $x=6, y=1.65$, $z=11.50$, see Figure 2). These rules were followed in order to satisfy the concept of axis for grasping proposed in Michel (2006), we simplified Michel's studies to reduce the natural axis of prehension of an object to its longest side. Mean and variance of objects associated with a precision grip and objects associated with a power grip are presented in Table 3.

## Procedure

The model received an unknown object to be classified as graspable with a power grip or a precision grip. This ob-
ject, represented by a vector $\left(x_{n}, y_{n}, z_{n}\right)$, was associated by the model to probabilities $P\left(\operatorname{grip}_{i} \mid x_{n}, y_{n}, z_{n}\right)$ for $i=1$ the precision grip $\left(\right.$ grip $\left._{1}=G_{1}\right)$ and $i=2$ the power grip $\left(\right.$ grip $\left._{2}=G_{2}\right)$.

The Bayes' theorem permits to decompose these probabilities :

$$
\begin{equation*}
P\left(\text { grip }_{i} \mid x_{n}, y_{n}, z_{n}\right)=\frac{P\left(\text { grip }_{i}, x_{n}, y_{n}, z_{n}\right)}{P\left(x_{n}, y_{n}, z_{n}\right)} \tag{1}
\end{equation*}
$$

The probability $P\left(\operatorname{grip}_{i}, x_{n}, y_{n}, z_{n}\right)$ can be written as :

$$
\begin{gather*}
P\left(\text { grip }_{i}, x_{n}, y_{n}, z_{n}\right)=P\left(x_{n}, y_{n}, z_{n}, \text { grip }_{i}\right) \\
=P\left(x_{n} \mid y_{n}, z_{n}, \text { grip }_{i}\right) \times P\left(y_{n}, z_{n}, \text { grip }_{i}\right) \\
=P\left(x_{n} \mid y_{n}, z_{n}, \text { grip }_{i}\right) \times P\left(y_{n} \mid z_{n}, \text { grip }_{i}\right) \times P\left(z_{n}, \text { grip }_{i}\right) \\
=P\left(x_{n} \mid y_{n}, z_{n}, \text { grip }_{i}\right) \times P\left(y_{n} \mid z_{n}, \text { grip }_{i}\right) \times P\left(z_{n} \mid \text { grip }_{i}\right) \times P\left(\text { grip }_{i}\right) \tag{2}
\end{gather*}
$$

Here, the naive assumption of conditional independence assumes that given the category $\operatorname{grip}_{i}, x_{n}, y_{n}$ and $z_{n}$ are independent, so that :

$$
\begin{equation*}
P\left(x_{n} \mid y_{n}, z_{n}, \text { grip }_{i}\right)=P\left(x_{n} \mid \text { grip }_{i}\right) \tag{3}
\end{equation*}
$$

and

$$
\begin{equation*}
P\left(y_{n} \mid z_{n}, \text { grip }_{i}\right)=P\left(y_{n} \mid \text { grip }_{i}\right) \tag{4}
\end{equation*}
$$

Thus, using equations (1) (2) (3) and (4)

$$
\begin{align*}
& P\left(\text { grip }_{i} \mid x_{n}, y_{n}, z_{n}\right)= \\
& \frac{P\left(x_{n} \mid \operatorname{grip}_{i}\right) \times P\left(y_{n} \mid \text { grip }_{i}\right) \times P\left(z_{n} \mid \text { grip }_{i}\right) \times P\left(\text { grip }_{i}\right)}{P\left(x_{n}, y_{n}, z_{n}\right)} \tag{5}
\end{align*}
$$

The model then selected the adapted grip for the object $\left(x_{n}, y_{n}, z_{n}\right)$ using :

$$
\begin{equation*}
\operatorname{argmax}\left[P\left(G_{1} \mid x_{n}, y_{n}, z_{n}\right) ; P\left(G_{2} \mid x_{n}, y_{n}, z_{n}\right)\right] \tag{6}
\end{equation*}
$$

In concrete terms the naive bayes classifier had two learning matrices of size $(j, 3), j$ being the number of objects in the learning matrices, represented by their three coordinates $\left(x_{j}, y_{j}, z_{j}\right)$. One matrix included the objects classified as graspable with a precision grip $\left(G_{1}\right)$, the other included the objects classified as graspable with a power grip $\left(G_{2}\right)$.

The following calculations were applied similarly for $G_{1}$ and $G_{2}$, we will only present the calculations for parameter $x$ in $G_{1}$ for the sake of clarity. The classifier computed the probability for an object to be graspable with a precision grip $\left(P\left(G_{1}\right)=\frac{j}{2 j}=\frac{1}{2}\right)$.

And the mean and variance of each parameter $x, y$, and $z$ for a precision grip : $\mu_{G_{1}}(x), \mu_{G_{1}}(y), \mu_{G_{1}}(z)$ and $\sigma_{G_{1}}^{2}(x), \sigma_{G_{1}}^{2}(y), \sigma_{G_{1}}^{2}(z)$; and for a power grip, resulting in $\mu_{G_{2}}(x), \mu_{G_{2}}(y), \mu_{G_{2}}(z)$ and $\sigma_{G_{2}}^{2}(x), \sigma_{G_{2}}^{2}(y), \sigma_{G_{2}}^{2}(z)$.

When a novel object with parameters $\left(x_{n}, y_{n}, z_{n}\right)$ was presented to the model, the classifier had to compute the probabilities $P\left(G_{1} \mid x_{n}, y_{n}, z_{n}\right)$ and $P\left(G_{2} \mid x_{n}, y_{n}, z_{n}\right)$, using (5).

As measurements were on continuous variables, the new parameters were computed given the known parameters
of the model using a gaussian probability density function, in order to calculate $P\left(x_{n} \mid G_{1}\right), P\left(y_{n} \mid G_{1}\right), P\left(z_{n} \mid G_{1}\right)$ and $P\left(x_{n} \mid G_{2}\right), P\left(y_{n} \mid G_{2}\right), P\left(z_{n} \mid G_{2}\right)$ with:

$$
P\left(x_{n} \mid G_{1}\right)=\frac{1}{\sqrt{2 \pi \sigma_{G_{1}}^{2}(x)}} e^{-\frac{\left[x_{n}-\mu_{G_{1}}(x)\right]^{2}}{2 \sigma_{G_{1}}^{2}(x)}}
$$

Then the model selected the highest probability (the appropriate grip), using (6).

As gaussian probability density function could return 0 for the probability of a parameter given a class grip $_{i}$, we distinguished two cases. In the first case only one parameter of the novel object had a probability equal to zero, in this case we did not change anything (we show in discussion why this case is a limit for this type of classification). In the second case two parameters of the novel object, one for each class, had a probability equal to zero (for example $P\left(y_{n} \mid G_{1}\right)=0$ and $P\left(z_{n} \mid G_{2}\right)=0$ ), we changed these probabilities to $\varepsilon$ close to zero , this changed $P\left(G_{1} \mid x_{n}, y_{n}, z_{n}\right)$ and $P\left(G_{2} \mid x_{n}, y_{n}, z_{n}\right)$ to

$$
\begin{equation*}
P\left(G_{1} \mid x_{n}, y_{n}, z_{n}\right)=\lim _{\varepsilon \rightarrow 0} \frac{P\left(x_{n} \mid G_{1}\right) \times \varepsilon \times P\left(z_{n} \mid G_{1}\right) \times P\left(G_{1}\right)}{P\left(x_{n}, y_{n}, z_{n}\right)} \tag{7}
\end{equation*}
$$

and

$$
\begin{align*}
& P\left(G_{2} \mid x_{n}, y_{n}, z_{n}\right)=\lim _{\varepsilon \rightarrow 0} \frac{P\left(x_{n} \mid G_{2}\right) \times P\left(y_{n} \mid G_{2}\right) \times \varepsilon \times P\left(G_{2}\right)}{P\left(x_{n}, y_{n}, z_{n}\right)} \\
& \begin{aligned}
& \text { As } P\left(x_{n}, y_{x}, z_{n}\right)=P\left(x_{n}, y_{n}, z_{n}, G_{1}\right)+P\left(x_{n}, y_{n}, z_{n}, G_{2}\right) \\
& P\left(x_{n}, y_{n}, z_{n}\right)= \\
& P\left(x_{n} \mid G_{1}\right) \times \varepsilon \times P\left(z_{n} \mid G_{1}\right) \times P\left(G_{1}\right)+ \\
& P\left(x_{n} \mid G_{2}\right) \times P\left(y_{n} \mid G_{2}\right) \times \varepsilon \times P\left(G_{2}\right) \\
&=\varepsilon \times\left[P\left(x_{n}, z_{n}, G_{1}\right)+P\left(x_{n}, y_{n}, G_{2}\right)\right]
\end{aligned}
\end{align*}
$$

So that, using (7) and (8):

$$
P\left(G_{1} \mid x_{n}, y_{n}, z_{n}\right)=\frac{P\left(x_{n} \mid G_{1}\right) \times P\left(z_{n} \mid G_{1}\right) \times P\left(G_{1}\right)}{P\left(x_{n}, z_{n}, G_{1}\right)+P\left(x_{n}, y_{n}, G_{2}\right)}
$$

Thus the probability of a grip given the three parameters of the novel object became the probability of a grip given the two parameters of the novel object for which probability was not changed by $\varepsilon$, as the changes operated cancelled each other out.

## Simulation

Simulation was performed using Matlab R2015a with a computer running on Windows 7 with a CPU Intel Core i5-4258U 2.10 GHz .

We aimed at assessing naive bayes classification by analysing classifier's performance with different learning matrices (different learned objects and number of objects
learned). In addition we compared the results of the classifier to the results obtained with human participants.

Simulation ran using $j=1$ to 7 learned objects for each category (we always used the same number of learned objects in the two categories: objects associated with a precision grip and objects associated with a power grip).

Objects that were not used in learning matrices were categorized using the method described earlier.

As learning order did not have any impact on classification, number of trials was defined using the binomial coefficient $\binom{N}{j}$ with $N=10$ the total number of objects in each category and $j$ the number of objects learned in each category. This binomial coefficient gives the number of combination of learned objects without taking into account possibilities of permutation (learning order). The classifier was tested for every possible combination of learning: for each combination of precision grip's learning, we tested all combinations of power grip's learning. This way the results presented in Table 4 and Table 5 show the proportion of correct classification for every object over all possible learnings of our material.

For each object and each $j$ we verified the grip selected by the classifier within each trial. For objects associated with a precision grip by means of our experiment (see Table 1), classification was recorded as right when the classifier calculates a higher probability for precision grip than for power grip. The reverse was made for objects previously associated with a power grip (see Table 2). If probabilities for a precision grip and for a power grip were equal, we considered that classification was incorrect.

## Results

Overall it took 1397.71 seconds ( 23 minutes and 29 seconds) for the program to select learning matrices and make 1837440 classification. The classification of one object took in average $7.61 \times 10^{-1} \mathrm{~ms}$.

When more than one parameter for one class was equal to zero ( 33 cases), or when $P\left(x_{n}, y_{n}, z_{n}\right)$ was considered equal to zero due to very small probabilities ( 82 cases), classification was impossible. These particular numeric cases happened rarely ( 115 objects impossible to classify over 1837440 classified objects).

We computed the percentage of right classification for each object and each $j$ (number of learned objects before classification). The percentages of right classification are shown in Table 4 (the percentage of right classification for objects considered as associated with a precision grip), and Table 5 (the percentage of right classification for objects considered as associated with a power grip).

A few things are to be discussed here. First, we can see that overall the classifier returned the right grip most of the time, in all the conditions ( $92.86 \%$ of right classification).

Secondly we can see that classification was better for objects that were considered associated with a power grip than for the others.

Thirdly, we see that classification performance increased as number of learned objects increased. This is because pa-

Table 4: Percentage of right classification for objects associated with a precision grip and for k objects learned.

| Objects | Number of learned objects |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 85.22 | 95.95 | 99.60 | 100 | 100 | 100 |
| 2 | 76.62 | 86.97 | 95.57 | 99.21 | 100 | 100 |
| 3 | 88.39 | 96.83 | 99.90 | 100 | 100 | 100 |
| 4 | 43.34 | 47.52 | 46.28 | 45.85 | 42.15 | 30.36 |
| 5 | 82.93 | 93.91 | 98.81 | 99.73 | 100 | 100 |
| 6 | 84.07 | 95.37 | 99.43 | 100 | 100 | 100 |
| 7 | 87.49 | 98.33 | 100 | 100 | 100 | 100 |
| 8 | 63.03 | 80.10 | 91.42 | 96.49 | 98.76 | 100 |
| 9 | 52.78 | 56.54 | 63.03 | 67.64 | 73.39 | 80.37 |
| 10 | 86.63 | 94.98 | 99.26 | 100 | 100 | 100 |
| Mean | 75.05 | 84.65 | 89.33 | 90.89 | 91.43 | 91.07 |

Table 5: Percentage of right classification for objects associated with a power grip and for k objects learned.

| Objects | Number of learned objects |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 2 | 3 | 4 | 5 | 6 | 7 |
| 11 | 96.22 | 99.93 | 100 | 100 | 100 | 100 |
| 12 | 86.66 | 96.26 | 99.52 | 100 | 100 | 100 |
| 13 | 96.82 | 99.79 | 100 | 100 | 100 | 100 |
| 14 | 95.54 | 99.56 | 100 | 100 | 100 | 100 |
| 15 | 94.21 | 98.77 | 99.82 | 100 | 100 | 100 |
| 16 | 92.29 | 98.75 | 99.96 | 100 | 100 | 100 |
| 17 | 95.09 | 99.73 | 100 | 100 | 100 | 100 |
| 18 | 86.63 | 94.82 | 98.82 | 99.84 | 100 | 100 |
| 19 | 96.97 | 99.76 | 100 | 100 | 100 | 100 |
| 20 | 94.04 | 99.30 | 100 | 100 | 100 | 100 |
| Mean | 93.45 | 98.67 | 99.81 | 99.98 | 100 | 100 |

rameters $\mu$ and $\sigma$ were more representative of a class (power or precision grip) as number of learned objects increased.

What is counterintuitive is that classification of object number 4 got worse and worse, it is because we put more objects different from object 4 in the precision grip's learning matrice as the simulation went on. Object 4 had its three parameters close to boundaries of the precision grip space (represented by its mean and variance for each parameter $x, y$ and $z)$. Thus, depending on the objects learned, increasing the number of learned objects put object 4 out of the boundaries: the more learned objects associated with a precision grip had parameters close to the parameters of object 4, the more object 4 was classified as part of precision grip's objects. Conversely the more learned objects associated with a precision grip had parameters distant from object 4, the more it was classified as part of power grip's object. Compared to object 4, other precision grip's objects had one of their parameter close to the boundaries of precision grip's space, but not all of their parameters, which made them easier to classify cor-
rectly.
The fact that object 4 was hardly well classified, instead of being a real issue for naive bayesian classification, could be an advantage when comparing the classifier's performance and human classification: in our experiment object 4 reveals the higher percentage of selection for the competing grip (see Table 1).

## Comparison of human and classifier's performance

To compare human's and classifier's performance we used a $\chi^{2}$ test of independence between variable object (object 1 to object 20) and variable responding entity (human participants or naive bayes classifier).

When three, four, five and six objects of each category were put in the classifier's learning matrices, we found that the two variables were independent $\left(\chi^{2}(19)=25.22, p=\right.$ $0.15 ; \chi^{2}(19)=23.06, \quad p=0.23 ; \quad \chi^{2}(19)=21.69, \quad p=$ $0.30 ; \chi^{2}(19)=22.71, p=0.25$, respectively), this meaning that classifier's performance and human grip's choice were not significantly different.

When two objects of each category were put in the classifier's learning matrices, we found that variables object and responding entity were independent, but with a greater difference between human's and classifier's performance $\left(\chi^{2}(19)=29.26, p=0.06\right)$.

Finally, when seven objects of each category were put in the classifier's learning matrices, it appeared that the two variables were not independent anymore $\left(\chi^{2}(19)=33.01, p<\right.$ 0.05 ), human's and classifier's performance became significantly different.

## Discussion

The results we obtained reveal that naive bayesian classification can reproduce the grip's choice made by human participants.

A good association of a novel object and its adapted grip can be accomplished with a reduced database and few parameters. This may permit to determine quickly a subclass of grips belonging to the precision or power grip classes when looking at an object, in other words to detect the possible nested micro-affordances associated with the object (for example a precision grip could comprise several nested microaffordances: a grip with the thumb and the index, a grip with the thumb, the index and the middle finger, with more or less strenght etc...). Quickness of the categorisation in precision or power grip classes could then be an advantage for real-time adaptation.

But some limitations are to be exposed. The calculation of conditional probabilities through gaussian probability density function implies that a parameter could have a zero probability given a certain grip class. This pulled the probability of this grip to zero, while the probability of the competing grip automatically became one, biasing the classification of the object. A second limitation is the ad hoc hypothesis that parameters are independent, which could induce errors for other parameters than the ones we used.

When seven objects of each category were learned by the classifier, the selection made by the classifier and human choice became significantly different probably because classifier's selection only account for a calculation made on the basis of mean and variance of the three parameters representing the objects. This calculation is always the same and as long as enough objects are learned the mean and variance of each class' parameter began to show little variability no matter the learning matrices. This shows that the algorithm used with this classifier produces a rigid classification, and cannot, at some point, reproduce the diversity created both by the complexity of our cerebral structures and the variations of embodiment between different human beings.

Yet this classifier can reproduce, in the majority of cases, human grip's choice in a small amount of time, and with few parameters needed to be taken into account. This shows that micro-affordances could be reproduced in some way with a simple computational system using naive bayesian classification, suggesting that some early stages of the processes linked to human micro-affordance could be performed by some simple probabilistic mechanisms.

Future studies should take more parameters for an object by cutting up the objects in three parts in order to determine the type of grip and the position of the grip on the object (Faria et al., 2014), and introduce action's consequences (Hommel, 2015; Shin, Proctor, \& Capaldi, 2010) through tactilo-kinesthetic parameters (Pfister et al., 2014), like pressure induced by the weight of the object, or muscle tension, in order to permit an efficient grip with a simple classification algorithm. We should also investigate the classifier's performance when an increased number of objects are learned and classified.

## Acknowledgments

We would like to thank Johan Briglia, Jean-Pierre Rivet and Firas Kaddachi for their contributions to this work. We also thank Epsylon's DynaCSE team and LIRMM's IDH team for welcoming us during our thesis, and Paul Valery University for funding this thesis.

## References

Brouillet, D., Vagnot, C., Milhau, A., Brunel, L., Briglia, J., Versace, R., \& Rousset, S. (2015). Sensorymotor properties of past actions bias memory in a recognition task. Psychological Research, 79, 678.
Chemero, A. (2003). An outline of a theory of affordances. Ecological Psychology, 15, 181-195.
Ellis, R., \& Tucker, M. (2000). Micro-affordance : The potentiation of components of action by seen objects. British Journal of Psychology, 91, 451-471.
Faria, D. R., Trindade, P., Lobo, J., \& Dias, J. (2014). Knowledge-based reasoning from human grasp demonstrations for robot grasp synthesis. Robotics and Autonomous Systems, 62, 794-817.
Gibson, J. J. (1979). The ecological approach to visual perception. Boston, MA: Houghton-Mifflin.

Hommel, B. (2015). The theory of event coding (TEC) as embodied-cognition framework. Frontiers in Psychology, 6, 1318.
Jones, M., \& Love, B. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. BEHAVIORAL AND BRAIN SCIENCES, 34, 169-231.
Koester, D., Schack, T., \& Westerholz, J. (2016). Neurophysiology of grasping actions: Evidence from erps. Frontiers in Psychology, 7, 1996.
Michel, C. (2006). Stratégie de saisie pour une main robotisée à caractère anthropomorphique. Doctoral dissertation, Robotics Department, University of Paris 6 Paris.
Montesano, L., Lopes, M., Bernardino, A., \& Santos-Victor, J. (2007). Affordances, development and imitation. In Development and learning, 2007. icdl 2007. ieee 6th international conference on. Piscataway, NJ: IEEE.
Montesano, L., Lopes, M., Bernardino, A., \& Santos-Victor, J. (2008). Learning object affordances: From sensorymotor coordination to imitation. In Ieee transactions on robotics (pp. 15-26). Piscataway, NJ: IEEE.
Naïm, P., Wuillemin, P. H., Leray, P., Pourret, O., \& Becker, A. (2007). Réseaux bayésiens. Paris, France: Eyrolles.

Pearl, J. (1985). Bayesian networks: a model of self-activated memory for evidential reasoning. In Proceedings of the annual conference of the cognitive science society 1985 (pp. 329-334). Irvine, UC: Cognitive Science Society.
Perfors, A., Tenenbaum, J. B., Griffiths, T. L., \& Xu, F. (2011). A tutorial introduction to bayesian models of cognitive development. Cognition, 120, 302-321.
Pfister, R., Janczyk, M., Gressmann, M., Fournier, L. R., \& Kunde, W. (2014). Good vibrations? vibrotactile selfstimulation reveals anticipation of bodyrelated action effects in motor control. Exp Brain Res, 232, 847-854.
Robert, C. (2000). L'analyse statistique bayésienne. Paris, France: Economica.
Shin, Y. K., Proctor, R. W., \& Capaldi, E. J. (2010). A review of contemporary ideomotor theory. Psychological Bulletin, 136, 943-974.
Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin Review, 9, 625-636.
Wolpert, D. M., \& Ghahramani, Z. (2000). Computational principles of movement neuroscience. Nature Neuroscience, 3, 1212-1217.
Wolpert, D. M., Ghahramani, Z., \& Flanagan, J. R. (2001). Perspectives and problems in motor learning. Trends in Cognitive Sciences, 5, 487-494.

# The Role of Schema-Governed Relational Categories in Analogical Inference 

Ricardo A. Minervino (minervinora@gmail.com)<br>IPEHCS-CONICET-UNCo, Av. Argentina 1400, 8300 Neuquén<br>Department of Psychology, University of Comahue.

## Adrián Margni (adrianmargni@gmail.com)

Department of Psychology, University of Comahue, Irigoyen 2000, 8324 Cipolletti.

Máximo Trench (maximo.trench@crub.uncoma.edu.ar)<br>Department of Psychology, University of Comahue, IPEHCS-CONICET-UNCo, Quintral 1250, 8400 Bariloche.


#### Abstract

The standard approach posits that analogical inferences are generated by copying unmapped base relations, substituting base entities by their corresponding target ones, and generating slots for unmapped base entities. Contra this account, results from Experiment 1 revealed that analogical inferences seldom include relations that resemble the base relation from which they were derived. Most of the inferences, however, could be categorized as exemplars of a schema-governed category capable of characterizing the base information to be projected. To gather further precision about the criteria that guide inference generation, in Experiment 2 we showed that analogical inferences tend to match the base information from which they are derived in values of salient dimensions of the relational category to which they belonged. Our results suggest that the relational constructs employed in modeling analogical inference should move beyond one-term multiplace predicates so as to include more complex relational structures.


Keywords: analogy; inference; relational category.

## Introduction

Analogical thinking is a central mechanism in human cognition (Gentner, 2003; Hofstadter \& FARG, 1995; Holyoak \& Thagard, 1995), playing an important role in activities as diverse as categorization, problem solving, scientific discovery, decision making, and argumentation (Gentner, Holyoak, \& Kokinov, 2001). In all these activities, analogy involves establishing a mapping between the compared situations and transferring new knowledge from a more familiar situation (base analog) to a less familiar one (target analog).

Almost all current theories of analogy agree that the alignment that takes place during mapping should satisfy the constraints of one-to-one mapping and parallel connectivity, (e.g., Falkenhainer, Forbus, \& Gentner, 1989; Gentner, 1989, Holyoak \& Thagard, 1989; Hummel \& Holyoak, 1997). While one-to-one mapping requires that each element in one situation maps to at most one element in the other situation, parallel connectivity entails that if two predicates are matched, their arguments must be placed in correspondence according to their roles. The following analogy illustrates these constraints:

> Base analog: John loved Mary and this made John give Mary a perfume

> Target analog: Peter loved Susan

While one-to-one mapping implies that pairing John with Peter should prevent pairing John with Susan, parallel connectivity dictates that if love has been paired with want, John must be placed in correspondence with Peter and Mary with Susan, as agents and patients, respectively, of the previous matched relations.

Theories of analogy also agree in that base assertions that are connected to the collection of mapped elements but which do not have a counterpart in the target will be brought over as candidate inferences. To formulate these inferences, the cognitive system would apply some variant of a copy with substitution and generation mechanism (CWSG; e.g., Falkenhainer et al., 1989; Holyoak, Novick, \& Melz, 1994; Hummel \& Holyoak, 2003). In our example, after projecting the base higher order relation cause to the target, the cognitive system would generate and transfer a "template" proposition like "Peter [give or a similar action"] Susan [something like a perfume]" into the target. This template proposition is generated from "John gave Mary a perfume", via copying give, substituting John by Peter and Mary by Susan (matched during mapping), and generating a slot for the entity perfume.

[^413]Supporting the postulations of the dominant theories of inference generation, there is some evidence that people tend to apply the syntactic constraints of one-to-one mapping (i.e., previous correspondences will dictate substitutions in CWSG) and systematicity (i.e., people are more likely to import an inference from base to target when the fact is causally connected to other matching facts) (e.g., Clement \& Gentner, 1991; Gentner, Ratterman, \& Forbus, 1993; Markman, 1997). There is also evidence that people tend to apply pragmatic criteria to derive their inferences (Spellman \& Holyoak, 1996). The application of these principles would be in the service of guaranteeing some minimal initial plausibility and relevance for the generated inferences (see, e.g., Holyoak et al., 1994).

As a pattern-completion process that takes maximal advantage of the mapping process, CWSG can be regarded as fast and computationally inexpensive. As most theories agree, analogical inference mechanisms should not be required to provide an adequate content to the produced inferences, something that should allegedly rely on post-inference stages of analogical reasoning such as evaluation and adaptation (Holyoak et al., 1994). The question arises as to whether the information conveyed by the templates obtained via CWSG can adequately guide post-inference generation processes in filling them in a semantically appropriate way.

In our example, if we repeat give, there is some probability of generating semantically appropriate inferences from the template "Peter gave [something like a perfume] to Susan" simply by replacing perfume by another exemplar of toiletry. However, while substituting perfume with spa set gives rise to a somewhat adequate inference, substituting perfume with deodorant would be inappropriate. It seems that combining give with an exemplar of toiletry will be adequate only if it gives rise to an instance of say, "manifestations of love". In this sense, the strategy of repeating the base relation and searching for a new exemplar of the base entity categories seems insufficient to guarantee some minimal semantic appropriateness of analogical inferences, to the extent that it requires some kind of "semantic supervision" from more complex category structures that the analogizer should keep in mind during the process. Combining substitutes (similar relations) of the relation give with substitutes of the entity perfume so as to obtain cases of "manifestations of love" would require even more thoughtful control. Just to exemplify, if we replace give by lend, no toiletry seems appropriate to generate a demonstration of love, although we can generate an exemplar of this category via replacing perfume by "his new car" or by "his weekend house at the beach".

The second problem with the CWSG strategy is that it can lead to inconvenient fixations, since many combinations of non similar verbs and objects could result in manifestations of love: write her a poem, prepare her favorite meal, or pick some wildflowers. Searching for cases of manifestations of love without the semantic restrictions imposed by the mechanism of CWSG seems to be a more flexible and productive strategy.

Based on the above considerations, we propose an alternative to CWSG which consists in categorizing the base analog information from which the inference will be derived as an exemplar of a schema-governed relational category (SGC), and searching for new exemplars for this category. Members of SGCs such as murder share a structure that can be instantiated by very different exemplars (Gentner \& Kurtz, 2005; Goldwater, Markman, \& Stilwell, 2011; Markman \& Stilwell, 2001), such as "Fred thrust a knife into Gina’s heart", "Mary had Bob drink poison", or "The offender disconnected the patient's oxygen supply". When the situations in an analogical comparison are exemplars of a SGC, the similarity between the relations and entities of the compared events is no longer necessary to have a good analogy (Minervino, Oberholzer, \& Trench, 2013). The analogical relatedness between "John gave Mary a perfume" and "Peter wrote Susan a poem" is not based on semantic resemblances between give and write or perfume and poem, but rather on the fact that both acts represent exemplars of the SGC "manifestation of love". In this sense, the limitation of CWSG seems to stem from treating analogical inference as an element-by-element pattern replacement guided by isolated similarities, and from not considering the broader meaning of the facts described by propositions. When this broader meaning is taken into account, the cognitive system can do away with element-to-element similarities.

With the aim of determining which of these alternative mechanisms constitutes a better account of how analogical inferences are generated, one of the conditions of Experiment 1 served to document the extent to which analogical inferences produced by participants involve relations that are similar to those of the base analog (as posited by dominant theories), as well as the extent to which they involve facts that pertain to the same schema-governed category as the base effect. In order to confirm that participants' inferences took into account the analogical relation between the target and the source-as opposed to representing plausible consequences of the target analog considered in isolation-, the inferences generated by the abovementioned group were contrasted to the inferences produced by a second group of participants who had to propose likely consequences of the target situations, but without having previously received an analogous source.

## Experiment 1

## Method

Participants and Design Fifty students of psychology at the University of Comahue (mean age $=22.86$ years, $S D=$ 3.42 volunteered to participate in the experiment. They were randomly assigned in equal number to the analogy and the target-only groups. The dependent variables were (1) the similarity between the relation of the base effect and that of the inferred situation, and (2) whether or not the inferred situation belonged to the same SCG as the base effect on which it was inspired.

Materials Ten sets of stimuli were built, each one comprising a base and a target analog. The base analog consisted of a base cause that engenders a base effect. The base causes consisted in three-place predicates in which an agent exerts an action over an object, and directed to a patient. The base effects were predicates in which the former patient exerted another action to another object, but which is directed to the former agent. Participants were tasked with generating a consequence of the target cause that they deemed analogous to that of the base analog. Table 1 displays a sample of the experimental materials.

Table 1: Sample of experimental materials, Experiment 1

| Set \# | Category | Situation |
| :--- | :--- | :--- |
| 1 | Danger | BC: An old man left the door of his <br> kitchen open to his two-year old <br> grandson |
|  |  | BE: The grandson ingested the old <br> man's medicines |
|  | TC: Another old man left the door of <br> his kitchen open to his two-year old <br> grandson |  |
|  | Public <br> welfare | BC: Latvian's low-income population <br> held a manifestation against the <br> government |
|  | BE: The government sent food to the <br> low-income population b |  |
|  | TC: Another low-income population <br> held a manifestation against the <br> government |  |

Note. BC: Base Cause; BE: Base Effect; TC: Target Cause.

Procedure Participants in the analogy condition received a brief written explanation about the potential of analogical comparisons to infer new information about a target situation. The instructions presented the main activity as one in which they were going to receive a first situation comprising a cause and its associated effect, followed by a second situation for which they had to proposed an effect that could be considered analogous to that of the original fact. Participants of the target-only group received a brief written explanation about how people hypothesize effects for
certain facts. The instructions presented the main activity as one in which participants were going to receive a simple situation, with the task of proposing a likely effect. Participants received the stimuli in random order. The experimental stimuli were presented on a computer screen, with participants typing their answers within prespecified fields. The administration took place in groups ranging from two to five participants, with each participant working individually. Participants were allotted a maximum of 30 min to complete the trials at their own pace.

Coding Each of the inferences proposed by participants was analyzed along two key dimensions: (1) the extent to which the action included in an inference was semantically similar to that of the base effect of the corresponding set of materials-a central prediction of the CWSG approach-, and (2) whether o not the inferred fact and its corresponding base effect belonged to the same schema-governed category. To carry out the first analysis, two judges unfamiliar with the purpose of the study received a ten-page table in which the verbs of the critical base effects (one from each set of materials) were matched against the verbs of all the inferences generated by participants for that particular set. Judges were asked to rate the similarity of the verb-pairs using a 5 -point scale ( $1=$ highly dissimilar; $5=$ highly similar). They worked independently of one other, and the scores given by the two judges to each of the verb-pairs were averaged. Judges' scores were found to be reasonably reliable, Cronbach's $\alpha=$.797. While verbs yielding an average score of three or more were classified as "similar", those obtaining an average score of less than three were sorted as "dissimilar". In order to perform the second analysis, two additional judges received each of the inferences proposed by participants preceded by its corresponding target cause and followed by a list of four words or brief descriptors representing SGCs, with the instruction to draw a mark next to any of the descriptions that could be used to categorize the target effect (they could mark as many as they wanted, or leave all of them unchecked in case they considered that none of them applied). For all inferences corresponding to a given set of materials, the list of event categories comprised two SGCs that corresponded to the base effect and two SGCs that did not correspond to the base effect, all presented in random order. For example, for Set 1 (see Table 1), one of the participants generated the inference "The grandson played with the stove". In order to determine whether this inference could be encompassed by the same SGC as the base analog in which it was inspired, judges received the target cause plus the inference at stake, coupled with the following event descriptions: (1) revenge, (2) dangerous situation (3) jealousy reaction, and (4) risky situation. Inferences were scored as sharing a SGC with the base effect in all those cases where the two judges checked at least one of the two "correct" event descriptors (danger and/or risk), regardless of whether they agreed on which of the correct descriptors was checked.

## Results and Discussion

The verbs of the inferences generated by participants of the analogy group resembled those of their corresponding base effect in $31.6 \%$ of the cases. It should be noted, however, that $12.7 \%$ of the inferences generated in response to the target analog alone (i.e., those of the target-only group) involved verbs that resembled those of the base effect received by participants of the analogy group. More fine-grained analyses using chi-square statistics revealed that the rate of utilization of similar verbs by the analogy group differed from the rate of spontaneous utilization of those same verbs by the target-only group in 4 of the 10 sets of materials (see Table 1).

With regards to SGC similarity, judges' analyses showed that while the inferences generated by the analogy group involved the SGC of the base effect in $88 \%$ of the cases, the inferences produced in response to the target analog alone belonged to these same categories in $34.4 \%$ of the cases. Chi- square tests revealed that for all 10 sets of materials the probability of generating an inference that pertains to the SGC of the base effect by participants of the analogy group was higher than the proportion of inferences pertaining to those same SGCs within the target-only condition (See Table 1).

The low proportion of semantically similar relations among the inferences produced by the analogy group suggests that the mechanism of postulating target relations that resemble their counterparts in the base analog, as dictated by CWSG, cannot adequately account for how analogical inferences are derived. In contrast, the fact that the vast majority of the inferences belonged to the same SGC as the causal consequent of the base analog suggests that the dominant mechanism involved in the generation of analogical inferences consists in analyzing the SGCs to which the base effect belongs, and generating further exemplars of such categories.

Table 2. Percentages of inferences exhibiting verb and relational category similarity with the base analog

|  | Similar relations |  | $\chi^{2}$ | Same relational category |  | $\chi^{2}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Set <br> \# | Analogy condition | Target-only condition |  | Analogy condition | Target-only condition |  |
| 1 | 32\% (8) | 16\% (4) | 1.75 | 92\% (23) | 36\% (9) | $17.01^{* *}$ |
| 2 | 36\% (9) | 12\% (3) | $3.95{ }^{*}$ | 88\% (22) | 40\% (10) | 12.5 ** |
| 3 | 32\% (8) | 8\% (2) | 4.5 * | 84\% (21) | 52\% (13) | 5.88* |
| 4 | 28\% (7) | 12\% (3) | 2 | 88\% (22) | 12\% (3) | $28.88{ }^{* *}$ |
| 5 | 20\% (5) | 12\% (3) | 0.6 | 84\% (21) | 24\% (6) | $18.12^{* *}$ |
| 6 | 32\% (8) | 36\% (9) | 0.09 | 92\% (23) | 20\% (5) | 26.3 ** |
| 7 | 16\% (4) | 36\% (9) | 2.6 | 88\% (22) | 56\% (14) | $6.35{ }^{*}$ |
| 8 | 44\% (11) | 0\% (0) | $14.1{ }^{*}$ | 88\% (22) | 36\% (9) | $14.35^{* *}$ |
| 9 | 44\% (11) | 4\% (1) | $10.96{ }^{*}$ | 96\% (24) | 32\% (8) | $22.22^{* *}$ |
| 10 | 32\% (8) | 36\% (9) | 0.09 | 80\% (20) | 28\% (7) | $13.61{ }^{* *}$ |

[^414]Having documented that the majority of inferences were exemplars of a relational category that was readily applicable to the base analog effect, a sensible research question concerned whether inclusion to such relational category suffices as a criterion for generating analogical inferences. As suggested by data obtained by Minervino et al. (2013), a factor that seems to influence the perception of analogical resemblance between exemplars of a SGC has to do with whether the target situation matches the base situation along the most salient dimensions of the relational category to which they belong. Taking the category robbery as an example, the analogability of two exemplars depends on whether they match in central dimensions such its importance, violence or planning. Experiment 2 was aimed at determining whether the observed sensitivity to this constraint generalizes to analogical inference.

As in Experiment 1, participants received a base analog comprising two causally related situations, followed by the presentation of a target situation that was virtually identical to the causal consequent of the base analog and by the task of completing the target situation with a consequence that they deemed analogous to that of the base. The main difference with Experiment 1, however, was that the exemplars of SGCs that were employed as the effects of the base situations were chosen to instantiate either a high or a low value along a central dimension of the relational category to which they belonged. The purpose of the experiment was to assess the extent to which the exemplars of SGCs included in participants' inferences matched the consequent of the base situation in terms of its values along the manipulated dimension.

## Experiment 2

## Method

Participants Twenty-four students of psychology at the University of Comahue (mean age $=21.1$ years, $S D=3.36$ ) volunteered to participate in the study.

Materials and Procedure Ten new sets of materials were built. The sets had the same general structure as those of Experiment 1, with the main difference being that for each base cause we derived two possible base consequences instead of one. These two consequences belonged to the same SGC, but differed from each other in that they scored differently along a central dimension of such category. As an example, the base cause $A$ paleontologist brought fossils to the Trelew Museum was followed either by the base effect The museum commissioned a statue of the paleontologist (an instance of reward with a high value in the dimension "magnitude") or by The museum issued a diploma to the paleontologist (low value in the dimension "magnitude"). Table 3 displays a sample of the experimental materials.
To ensure that participants encoded the base consequences as members of the SGC whose critical dimension was being manipulated, participants were explicitly informed about the specific category to which the base consequence belonged (see Table 3).

Table 3: Sample of experimental materials, Experiment 2

| Set | Category | Base and target situations |
| :---: | :---: | :---: |
| 1 | REWARD | BC : A paleontologist brought fossils to the Trelew Museum $\mathrm{BE}_{\mathrm{hv}}$ : The Museum commissioned a statue of the paleontologist $\mathrm{BE}_{\mathrm{lv}}$ : The Museum issued a diploma to the paleontologist TC: Another paleontologist brought important fossils to the Rawson Museum |
| 2 | ROBBERY | BC: The old lady trusted her house's keys to her nanny <br> $\mathrm{BE}_{\mathrm{hv}}$ : The nanny sold the old lady's jewelry <br> $\mathrm{BE}_{\mathrm{lv}}$ : The nanny took a book from the old lady's house TC: Another old lady trusted her house's keys to her nanny |
| 5 | CONTRI- <br> BUTION | BC : A young man was invited to a barbecue by his friends $\mathrm{BE}_{\mathrm{hv}}$ : He volunteered to pay the meat to his friends <br> $\mathrm{BE}_{\mathrm{lv}}$ : He volunteered to bring matches to his friends <br> TC: Another young man was invited to a barbecue by his friends |

Note. BC: Base Cause; $\mathrm{BE}_{\mathrm{hv}}$ : Base effect with high values on a key dimension of the relational category; $\mathrm{BE}_{\mathrm{lv}}$ : Base effect with low values on such dimension; TC: Target Cause.

Two complementary booklets of materials were built. In each version half of the sets were coupled with consequences instantiating low values along the manipulated dimension of the SGC to which they belonged, and half with consequences embodying high values along such dimensions. The procedure was identical to that on Experiment 1.

Coding Two new judges received each of the inferences generated by participants coupled with the critical dimension that corresponded to that set of materials. They were asked to rate how it fared along such dimension using a 5-point scale ranging from the minimum to the maximum possible levels along the manipulated dimension (e.g., for the reward example, they had to rate the magnitude of the reward from $1=$ very small, to $5=$ huge). The scores given by the two judges to each of the inferences were averaged. Judges' scores were found to be reliable, Cronbach' $\alpha=.823$.

## Results and Discussion

The inferences generated out of base facts ranking high along the manipulated dimensions obtained higher scores than those generated out of base facts displaying lower levels along that dimension $(M=3.18, S D=0.40$ vs. $M=2.18, S D=$ $0.36, t(25)=-10.05, p<.001$.

These results demonstrate that the way in which the base effect fares along a critical dimension of the SGC to which it belongs constrains the way analogical inferences will fare along such dimension. In order to gather a subtler estimate of the strength of this association, judges were also required to score the base effects along the manipulated dimensions. The correlation between the scores assigned to the base effects and those of their associated inferences was strong, $r$ $=.476, n=260, p<.001$. Furthermore, in $71.54 \%$ of the cases the scores of the generated inferences along the manipulated dimension were no farther than one point away from those of the base analogs on which they were inspired.

## General Discussion

A key prediction of the CWSG approach to analogical inference consists in that people will tend to construct their inferences repeating the base relations from which the inferences will be derived or replacing them by similar ones. Against this position, Experiment 1 showed that people do not care much about preserving similarity to base relations but instead focus on generating new exemplars of the SGC applied to the base information from which inferences are derived. In Experiment 2 we collected data favoring a further hypothesis associated to our category-based perspective, namely, that when proposing new exemplars of SGCs people tend to generate cases that fare closer to the base exemplar along critical dimensions of the category to which they pertain.

We have argued that the templates generated by CWSG could sometimes be insufficient to guide the analogizer in generating semantically sensible inferences during the post inference stages of evaluation and adaptation. Returning to the example presented in the Introduction, the chances of generating semantically appropriate inference from the template "Peter [give or a similar action] Susan [something like a perfume] seems rather low. We speculated that while some toiletries could perhaps result in a sensible inference, others do not, and that post-inference generation mechanisms have no semantic basis to distinguish between them.

The standard approach to analogical inference generation could argue that "something like a perfume" should not be interpreted as "an exemplar of toiletry", and that this interpretation is to some extent caricaturizing CWSG, since an intelligent system operating in an analogical mode will not be guided by superficial similarities such as membership to a same category, but would rather interpret it as, say, "give + things that a woman finds romantic". In this sense, the system would promote the search for new exemplars of this ad hoc category (e.g., a teddy, necklace or a bouquet). The problem with this argument is that the very consideration of this ad hoc category supposes the prior conceptualization of the template as a "manifestation of love", something that the analogical machinery has not generated. It is possible that the generalized support that the CWSG approach has received comes in part from the fact that programmers inadvertently read far more understanding than is warranted into the templates produced by this mechanism, as an effect of projecting the SGC that they apply to capture the whole meaning of the template.

The standard approach could also argue that the analogical machinery was not meant to deal with the activity of comprehending the analogs, but rather to start operating once the analogs have been fully comprehended (see, e.g., Morrison \& Dietrich, 1995). In this vein, the analogical engine would receive the fact that John gave Mary a perfume already interpreted as a case of "carrying out a very romantic manifestation of love" (i.e., as a case of this category, and with some specific properties). A problem with this argument is that this conceptualization cannot be captured by a relation-defined as one-term multiplace predicates. The execution of a very romantic demonstration of love would be propositionally represented, stricto sensu, as CARRY OUT [John, ((VERY) ROMANTIC (manifestation of love)), Mary]. The essential information to be transferred is located in an argument represented as a noun (manifestation of love) and its property (VERY ROMANTIC), and not in the one-term predicate outside the brackets (CARRY OUT).

We are far from calling into question the importance of relational aspects in analogical thinking, but we believe it is necessary to discuss and amplify the meaning of "relational", so as to avoid reducing it to one-term multiplace predicates. It should be broadened to include, for example, relational structures as those captured by SGCs. In these structures, relations are only a constituent, being other thematic roles (e.g., agents, patients, objects or instruments) just as important. For example, if an instance of the category "manifestation of love" includes the relation give, the agent's intention should be to awake certain emotions in a person, the patient has to be a candidate for being emotionally affected by the agent at stake, and the object should be pleasant to the patient. The complex interdependency of the constituents of a fact that make it pertain to a SGC makes it proper to talk about these categories as "relational" structures, but the sense of the term is broader than the one employed in computational models of analogy (i.e., a one-term multiplace predicate). The relational character of these categories is also evident in the extremely different situations that can constitute exemplars of a SGC, which could differ even in their relations defined in the traditional way.

## Acknowledgments

We thank Micah Goldwater for suggesting how to operationalize whether participants' inferences belonged to the SCG of the source. This work was supported by the National Agency for Scientific and Technical Research (ANPCyT) under Grants PICT 2352 and PICT 2650, by the University of Comahue under Grants B213, C123 and C134, and by the National Council for Scientific and Technical Research (CONICET) under Grant PIP 0567.

## References

Clement, C. A., \& Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. Cognitive Science, 15, 89-132.
Falkenhainer, B., Forbus, K. D., \& Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial Intelligence, 41, 1-63.

Forbus, K. D., Ferguson, R. W., Lovett, A., \& Gentner, D. (in press). Extending SME to handle large-scale cognitive modeling. Cognitive Science.
Gentner, D. (1989). The mechanisms of analogical transfer. In S. Vosniadou, \& A. Ortony (Eds.), Similarity and Analogical Reasoning. Cambridge, UK: Cambridge University Press.
Gentner, D. (2003). Why we are so smart. In D. Gentner, \& S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought. Cambridge, MA: MIT Press.
Gentner, D., \& Kurtz, K. J. (2005). Learning and using relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, \& P. W. Wolff (Eds.), Categorization Inside and Outside the Laboratory (Vol. 43). Washington, DC: APA.

Gentner, D., Holyoak, K. J., \& Kokinov, B. N. (Eds.) (2001). The analogical mind: Perspectives from cognitive science. Cambridge, MA: MIT Press.
Gentner, D., Rattermann, M. J., \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive Psychology, 25, 431-467.
Goldwater, M. B., Markman, A. B, \& Stilwell, C. H. (2011). The empirical case for role-governed categories. Cognition, 118, 359-376.
Hofstadter D. R., \& the Fluid Analogies Research Group (1995). Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought. New York: Basic Books.
Holyoak, K. J., Novick, L. R., \& Melz, E. R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. J. Holyoak, \& J. A. Barnden (Eds.), Advances in connectionist and neural computation theory (Vol. 2). Norwood, N.J.: Ablex.
Holyoak, K. J., \& Thagard, P. R. (1989). Analogical mapping by constraint satisfaction. Cognitive Science, 13, 295-355.
Holyoak, K. J., \& Thagard, P. R. (1995). Mental leaps: Analogy in creative thought. Cambridge, MA: MIT Press.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.
Markman, A. B. (1997). Constraints on analogical inference. Cognitive Science, 21, 373-418.
Markman, A., \& Stilwell, C. (2001). Role-governed categories. Journal of Experimental \& Theoretical Artificial Intelligence, 4, 329-358.
Minervino, R., Oberholzer, N., \& Trench, M. (2013). Overall similarity overrides element similarity when evaluating the quality of analogies. Journal of Cognitive Science, 14, 287-317.
Morrison, C., \& Dietrich, E. (1995). Structure-mapping vs. High-level perception: The mistaken fight over the explanation of analogy. In J. D. Moore and J. Fain Lehman (Eds.), Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society (pp. 678682). Hillsdale, NJ: Erlbaum.

# An Investigation of Factors that Influence Resource Allocation Decisions 

Percy K. Mistry (pkmistry@uci.edu)<br>Department of Cognitive Sciences, University of California Irvine, Irvine, CA 92697-5100 USA<br>Jennifer S. Trueblood (jennifer.s.trueblood@ vanderbilt.edu)<br>Department of Psychology, Vanderbilt University, 111 21st Ave S Nashville, TN 37240 USA


#### Abstract

We investigate how people allocate a limited set of resources between multiple risky prospects. We found that only a small percentage of decisions followed some form of naive diversification or mean-variance optimization. In general, people were less mean-variance optimal than a naive $1 / \mathrm{N}$ heuristic. Aspects of choice sets, such as domain, skew, and second order stochastic dominance, affected resource allocation decisions in a similar manner to their influence on single choice gambles. Individual traits traditionally linked to risk propensity seem to manifest in terms of the degree to which people are inclined to diversify. Lower risk aversion and higher risk seeking traits are linked to increasing diversification. Risk congruency, the degree to which peoples' self-reported and elicited risk aversion matches, moderates how susceptible people are to cost framing nudges. We find evidence for heterogeneous clusters where people either under-weight or over-weight segregated costs, leading to the same nudge producing opposite behavioral results within two risk incongruent groups.


Keywords: resource allocation; risk tolerance; risky choice; individual differences; nudges

## Introduction

There are many instances where people have to distribute a limited set of resources between multiple choice options. These can be personal (e.g. investment in a set of retirement funds; constructing a stock portfolio; budgeting household expenses) or institutionalized (e.g. capital allocation, bank lending decisions, government budgets) monetary decisions. These could also be non-monetary decisions such as distribution of labor, time, bandwidth, etc. The choice options often vary in terms of their potential costs and benefits, which are probabilistic in nature. Each choice option may thus be represented as a risky prospect which has some probabilistic distribution of outcomes. There is a large amount of literature that examines the decision making process when people have to select only one out of 2 or more risky prospects, that is, where all resources are invested in a single prospect.

There are limited studies however, that extend this to a resource allocation paradigm, namely, how do people distribute a limited set of resources between 2 or more such risky prospects? Some studies suggest that people follow the $1 / \mathrm{N}$ heuristic, which proposes that people tend to naively diversify allocation across the available prospects (Benartzi \& Thaler, 2001; Bardolet, Fox, \& Lovallo, 2011), although this is often the case only for a subset of the people making these decisions. The normative version of this problem is extensively studied in economics - what optimal strategies should people adopt? However there is limited research examining whether people come close to adopting such optimal strategies.

It is important to understand how people deviate from optimality, and to understand what aspects of choice sets influ-
ence the resource allocation process. We highlight that measuring optimality and sensitivity to choice sets is more complex in resource allocation tasks compared to simple choice gambles. For simple choice gambles, a cognitive account will typically entail valuation of different prospects, and specification of a deterministic or probabilistic decision rule to compare these valuations. The decision rule for resource allocation needs to be more complex to allow for allocation weights to be placed for each gamble. It needs to take into account aspects such as choice bracketing - whether the valuation of choices is performed at an aggregate portfolio or segregated choice level. These aspects may influence whether drivers of decision making that explain single choice behavior can also explain resource allocation decisions. Further, choice bracketing may also affect how sensitive people are to cost framing nudges, where outcomes are re-framed into a gross higher outcome, set-off by a corresponding cost element. We report an experimental study on resource allocation and show how manipulation of different design factors affects the allocation behavior. The main questions we ask are (1) do people naively diversify?, (2) how sensitive are people to choice set manipulations?, (3) how well do measures of risk traits explain individual differences in resource allocation behavior?, (4) how sensitive are people to cost framing nudges?and (5) how optimal (or sub-optimal) is allocation behavior?

## Experiment

In this experiment we test people's preferences for distributing a fixed set of resources between multiple risky prospects. 50 undergraduate students from Vanderbilt University participated in the experiment. The cover story for the task was that participants had to play the role of the head of a company that had the opportunity to invest a fixed amount of money (hypothetical $\$ 100,000$ ) into one or more of 4 possible projects. Participants were advised that all projects had the same expected time to completion and their objective was to maximize the return on the invested amount. They were required to invest all the money, but could distribute this in any proportion between the 4 projects, including allocating no resources to one or more projects. Each project had two possible outcomes - success or failure. They were provided with the probability of success $\left(p_{S}\right)$ and failure ( $p_{F}=1-p_{S}$ ) for each project, as well as the percentage returns on their investment depending on whether a project succeeded or failed. A successful project always had a positive return, whereas a failed project resulted in either a lower positive or a negative return. The 4 projects always varied in terms of the variability (standard deviation) of return outcomes. Participants were given


Figure 1: Example interface where participants allocate a fixed set of resources between 4 risky prospects using either text inputs or a moving slider scale.


Figure 2: Example interface of how participants receive feedback for each individual prospect after each trial. The proportion of green to red balls is based on the ratio of probability of success to failure. One of the balls is chosen at random to generate a realized outcome.
an example and a practice trial to familiarize themselves with the interface (see Figure 1). After each trial, participants were provided feedback on the outcome. The outcome was based on the described probability of success and dynamically (randomly) picked by the computer program. The process of realization of the outcome for each project was graphically displayed to the participants. For each project they were shown a box containing $100 \mathrm{X} p_{S}$ green balls and $100 \mathrm{X} p_{F}$ red balls. The computer program randomly traversed the box space and eventually picked one of the balls. A green ball implied success and a red ball implied failure (see Figure 2). The returns on the investment for each project were updated based on these outcomes before moving on to the next trial.

## (A) Between-subjects conditions

Participants were split into 2 groups of 25 students each. The between subjects design entailed different rewards, with the rest of the design factors being identical between the two groups. Group 1 participated for course credit, and group 2 for financial compensation. This between-subjects condition tests whether financial incentives affect resource allocation behavior. There is mixed evidence for this in tasks involv-
ing risky choices (Beattie \& Loomes, 1997). Participants in group 2 received a fixed payout of $\$ 5$ plus an incentive ranging from $\$ 0$ to $\$ 10$ that was linked to their performance on the task. For group 2, at the end of the experiment, one of the trials was randomly selected. The incentive component was calculated as $\$ 5$ plus or minus $\$ 0.10$ times the \%returns achieved on that trial, but limited to the range $\$ 0-\$ 10$. For example, achieving a loss of $20 \%$ resulted in an incentive of $\$ 5-0.1(20)=\$ 3$, and achieving a gain of $20 \%$ resulted in an incentive of $\$ 5+0.1(20)=\$ 7$. This allowed for the incentive to be framed as reductions for losses and increments for gains. The total payout including fixed and incentive components ranged between \$5-\$15.

## (B) Within-subjects factorial design

Each participant completed 36 portfolio choice decisions. There were 12 unique decisions based on a 2 (second order stochastic dominance - present vs absent) X 2 (domain - gains vs mixed) X 3 (skew - none, positive, negative) within-subject factorial design. Each of these 12 decisions was repeated in 3 blocks, with the order randomized within blocks. Although the underlying decision remained equivalent across blocks, the three blocks varied in terms of a cost framing effect. The details of the choice set manipulation are given below:
Second order stochastic dominance (SOSD; 2 levels): In the first level, all prospects in a trial had equal expected value, but the 4 prospects had progressively higher standard deviation. As a result, each prospect had SOSD over the subsequent riskier prospects. In the second level, the prospects were not mean preserving, and riskier prospects (higher standard deviation) also had higher expected values. Thus there was no SOSD. Any behavioral account that is based on a weakly increasing concave utility function, or mean-variance optimization, predicts that prospects with SOSD over other prospects will be a dominated preference. Accordingly, optimal resource allocation under such an assumption would imply allocating $100 \%$ of the resources to the safest prospect. SOSD present choice sets allow a parameter free estimation of deviation from mean-variance optimal allocation. On the other hand, choice sets that do not involve SOSD choices allow measuring the level of risk tolerance within a meanvariance optimization framework.

Domain (DM; 2 levels): In the first level, all-gain domain, all outcomes including project failures resulted in positive returns. In the second, mixed domain, the average returns across prospects on failure were negative. Domain manipulation allows us to test for the effects of asymmetric gain-loss utilities within the portfolio choice framework.

Skew (SK, 3 levels): In the first level, all prospects had zero skew, that is, success and failure were equally likely. In the second, all except the safest prospect had negative skew, that is, failure outcomes were more likely. In the third, all except the safest prospect had positive skew, that is, success outcomes were more likely. Symmonds, Wright, Bach, and

Dolan (2011) showed that risk and skewness are differently encoded in the brain. People have been shown to be relatively averse to negatively skewed gambles (Deck \& Schlesinger, 2010). Manipulation of skew allows us to test whether these effects extend to the portfolio choice paradigm.

Purchase cost framing nudges (PC, 3 levels): In the first level, there are no extraneous purchase costs. In the second and third levels, the outcomes from the first level were translated and re-framed into higher gross outcomes accompanied by an appropriate purchase cost. This re-framing led to prospects that were expected-value-equivalent to the prospects presented in the first level. In the second level, the amount of re-framing was increasingly higher with increasing variability (risk) of prospects outcomes. In the third, the re-framing decreased with increasing variability (risk). The trials were presented in a blocked design with three blocks corresponding to the 3 purchase costs conditions, with the order of the 12 problems in each being randomized. The theory of mental accounting suggests that people may account for re-framed outcomes and corresponding costs in a segregated manner, including under or over weighting the cost component relative to the outcomes. The re-framing thus can act as a nudge, pushing people towards allocation to riskier prospects in the second level and safer in the third level if they underweigh the re-framed costs. The direction of the nudge would be reversed if people over-weigh re-framed costs.

## (C) Testing for risk traits

After the allocation task, participants were required to complete one set of paired lottery choices (Holt \& Laury, 2002), summarized as $H L$, with a higher score indicating greater elicited risk aversion. They also completed the DOSPERT financial risk-taking $(D F)$ subscale (Blais, 2006), with a higher $D F$ score indicating higher self-reported risk seeking behavior, and a locus of control (SL) scale (Rotter, 1966), with higher $S L$ scores indicating a higher self-reported external locus of control. All of these influence risky decision making in single choice tasks.

## (D) Defining the dependent variables

The simplest way of measuring a resource allocation decision is to look at the allocation weights $\left(w_{i}\right)$ for each $\left(i^{\text {th }}\right)$ prospect, where $\Sigma_{i=1}^{N} w_{i}=1$, and $N$ is the total number of choices available. Lopes and Oden (1999) proposed that there are individual differences in whether people approach risky decision making from a perspective of security (protecting low outcomes) or potential (maximizing high outcomes). A simplistic measure of people's security and aspiration levels are measured by the weight allocated to the two extreme prospects - safest $(S)$ and riskiest $(R)$ respectively. In addition, the Herfindahl index $(H)=\sum_{i=1}^{N} w_{i}^{2}$, where $N$ is the total number of prospects in the choice set, measures the degree of diversification (Rhoades, 1993). When all weights are equal, $H$ takes the minimum value of $1 / \mathrm{N}$ (maximum diversification) and when all resources are allocated to a single prospect,
$H=1$. For $\mathrm{N}=4$, values close to 0.25 indicate naive diversification, values close to 0.5 indicate some form of conditional diversification (equal allocation to 2 of 4 prospects), and values close to 1 indicate concentration in a single prospect. These measures $S, R$, and $H$ reflect segregated measures based on attention paid to individual prospects.

Often, the emergent characteristics of the aggregated portfolio are of greater interest than the individual choices. Most normative theories of portfolio choice are based on optimizing some function of the portfolio characteristics. Since a portfolio can be represented as a probability distribution over outcomes, the most common characteristics are derived from the moments of the resulting portfolio. We calculate the expected value $(V)$, and the standard deviation $(D)$ of the aggregate portfolio.

Finally, we test for differences between the 2 cost framing conditions. The framing conditions are setup so that correctly accounting for the costs and translation of outcomes should result in no difference between behavior across the three conditions. However, discounting of the costs framed separately would result in a preference for prospects with a higher degree of framing. In one condition, riskier prospects are subject to higher framing (we denote this condition as $F_{1}$ ), and in the other, safer prospects are subject to higher framing, denoted as $F_{2}$. Discounting the costs would result in higher selection of riskier prospects in the first and safer in the second framing condition. We calculate susceptibility to nudges as, $N=\operatorname{mean}\left[\left(S_{F_{2}}-S_{F_{1}}\right),\left(R_{F_{1}}-R_{F_{2}}\right)\right]$. A value of $N$ close to 0 indicates that people are not susceptible to cost framing nudges. A high positive value indicates that people underweight separately framed costs, and thus are nudged towards options with higher framing (larger translation of outcomes). A high negative value indicates that people over-weight separately framed costs, and thus are nudged towards options with lower framing (smaller translation of outcomes).

## Results

## (1) Is there evidence for naive diversification?

Diversification is directly measured using the Herfindahl index (H). The left panel in Figure 3 shows the distribution of $H$


Figure 3: Distribution of Herfindahl index across participants and trials. The color shading shows the number of unique prospects (1 to 4) selected on each trial.

Table 1: Mean values of dependent behavioral measures by design factor. Differences are tested using a Bayesian repeated measures ANOVA for main effects of design factors. Significant differences, measured by log Bayes factors (LBF) $\geq 2.3$ are highlighted in bold.

|  | SOSD |  | Domain |  | Skew |  |  | Cost framing |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | (Yes) | (No) | (Gain) | (Mixed) | (0) | ( Neg ) | (Pos) | (None) | (Riskier) | (Safer) |
| Herfindahl Index (H) | 0.48 | 0.47 | 0.48 | 0.47 | 0.46 | 0.48 | 0.49 | 0.44 | 0.49 | 0.49 |
| \%Safest (S) | 0.37 | 0.31 | 0.29 | 0.39 | 0.34 | 0.37 | 0.31 | 0.32 | 0.33 | 0.36 |
| \% Riskiest (R) | 0.25 | 0.29 | 0.30 | 0.24 | 0.26 | 0.28 | 0.27 | 0.27 | 0.28 | 0.26 |
| Expected value (V) | 2.50 | 3.98 | 5.84 | 0.64 | 3.23 | 3.25 | 3.24 | 3.27 | 3.26 | 3.19 |
| Standard deviation (D) | 3.83 | 4.42 | 2.06 | 6.19 | 4.38 | 3.60 | 4.39 | 4.23 | 4.32 | 3.83 |
| Susceptibility to nudges ( N ) | 0.01 | 0.04 | 0.00 | 0.05 | 0.02 | 0.03 | 0.02 | - | - | - |
| MPT-error ( $d_{\mathcal{\varepsilon}}$ ) | 0.31 | 0.18 | 0.31 | 0.18 | 0.24 | 0.22 | 0.27 | 0.22 | 0.26 | 0.25 |
| Risk tolerance ( $Q_{\varepsilon}$ ) | - | 63.7 | 19.9 | 107.5 | 64.7 | 52.4 | 73.9 | 67.0 | 76.1 | 48.0 |

across participants and trials. The color shading also shows the distribution of the number of unique prospects selected on any trial. A naive diversification strategy would indicate a value of $\mathrm{H}=0.25$. A large mass of the distribution lies between the range of 0.25 and 0.5 , with further peaks at 0.5 and 1.0 indicating choices where people selected 2 of the 4 prospects equally, or invested all their resources in a single prospect, respectively. The $1 / \mathrm{N}$ heuristic (naive diversification), proposes that people tend to split allocations evenly between available choices. A variant of this strategy called the conditional 1/N heuristic (Huberman \& Jiang, 2006) proposes that people split allocations evenly across a small number of choices rather than the total number of choices available. Using thresholds suggested by Huberman and Jiang (2006), 11\% of the choices can be summarized as single prospect concentration, $7 \%$ as a conditional diversification into 2 prospects, and $4 \%$ as naive diversification into all 4 prospects.

## (2) Sensitivity to choice set manipulations:

The mean values of the dependent behavioral variables grouped by experimental factors (which define the type of choice sets) are summarized in Table 1. We conduct a Bayesian repeated measures ANOVA analysis (JASP-Team, 2016) testing for main effects of these design factors. A log Bayes factor, $\mathrm{LBF} \geq 2.3$ is considered significant, and highlighted in bold in Table 1. There is no evidence that the incentive condition had any effect on $S, R, H, V$, or $D$, hence the remainder of the analysis combined data from the course credit and financial incentive conditions. There is no evidence that the level of diversification as measured by $H$ is affected by the domain, skew, or SOSD manipulations. There is evidence for a main effect of domain (LBF 28.1), SOSD (LBF 12.8), and skew (LBF 3.2) on $S$. Allocation to $S$ is higher in the mixed domain (mean $39 \%$ ) than in the gains domain (mean 29\%), higher in the SOSD (mean 37\%) compared to non-SOSD (mean $31 \%$ ) condition, and higher in negative skew (mean $37 \%$ ) than positive skew (mean $31 \%$ ) conditions. There is evidence for a main effect of domain (LBF 9.6) and SOSD (LBF 4.3) on $R$. Allocation to $R$ is higher in the gains domain (mean 30\%) than in the mixed domain (mean 24\%), and higher in the non-SOSD (mean $29 \%$ ) compared to the SOSD (mean $25 \%$ ) condition. $V$ and $D$ are expected vary
with domain and SOSD by design. There is no evidence for a main effect of skew on $V$, but there is evidence (LBF 6.9) for a main effect of skew on $D$. Participants exhibit the lowest $D$ (mean 3.6) in the negative skew condition and highest $D$ (mean 4.39) in the positive skew condition, indicating a marked preference for lower variability in the negative skew condition.

## (3) Trait-based individual differences

To test if the measured traits influence behavior in the portfolio allocation task, we use a Bayesian ANCOVA analysis treating the between and within subject choice manipulation factors as random effects and testing for the effects of covariates locus of control (SL), risk aversion (HL), and financial risk seeking (DF). We find evidence of an effect of HL (LBF 5.7) and DF (LBF 2.9) on $S$, and an effect of HL (LBF 13.2) on $R$. These indicate that higher risk aversion (higher HL and lower DF scores) are linked to higher allocation to the safest prospect and lower allocation to the riskiest prospect, as might be expected. Testing for effects of the locus of control (SL), we find evidence of an effect on $S$ (LBF 3.7), and on $R$ (LBF 6.2). These indicate that increasing external locus of control is also linked to higher allocation to safest and lower to the riskiest prospect. Directionally, this is in contrast to findings based on risky gambles (Rotter, 1966) which showed that increasing external locus of control was associated with waging more money on riskier bets.

Figure 4 shows the joint density of \% allocations to $R$ and $S$. The color coding in the three panels shows the mean level of trait scores. Areas in the centre indicate diversificationlike behavior. Interestingly, all 3 mean scores are pretty similar for both extreme decisions ( $\mathrm{R}=100 \%$ and $\mathrm{S}=100 \%$ ), but are different for the central areas representing diversification (lower HL scores and higher DF scores). It seems that risk aversion and risk seeking measures are more indicative of how extreme (concentration vs diversification) people are in their allocations, rather than whether they prefer safer or riskier prospects. We find evidence of an effect of HL (LBF 1.4) and DF (LBF 13.4) on $H$. These indicate that higher risk aversion is linked to lower diversification. This behavior is however contrary to the popular notion that diversification leads to reduced risk.


Figure 4: Joint density of $\%$ allocation to safest vs aspirational prospects. The size indicates the $\%$ of all trials. The 3 panels plot the same density data, with the color coding in the 3 panels showing the mean level fo risk trait scores ( $S L, H L$, and $D F$ respectively).

## (4) Sensitivity to cost framing nudges:

Testing for differences between the framing and no framing conditions, we find a significant effect (LBF 10.3) of whether or not there is some cost framing on $H$, so that framing of either type reduces diversification, with mean H increasing from 0.44 in the no framing condition to 0.49 in both the cost framing conditions. One hypothesis is that the introduction of an additional cognitive element induces people to reduce their diversification. This is supported by the observation that in the no framing condition, people selected all 4 prospects on $55 \%$ and either 1 or 2 prospects on $22 \%$ of the trials. Compared to this, in the framing conditions (combined), people selected all 4 prospects on $45 \%$ and either 1 or 2 prospects on $33 \%$ of the trials.

Testing for differences between the two cost framing conditions $F_{1}$ (higher risk framing) and $F_{2}$ (higher safety framing), a Bayesian repeated measures ANOVA shows evidence that there is no effect on $H$ (LBF -2.7), $S$ (LBF -0.4) or $R$ (LBF -1.9) . Similarly, there is no effect of choice set conditions on $N$. This seems to indicate that people are not susceptible to differential cost framing effects and adequately account for segregated costs and translation of outcomes. However, we find that risk traits are significant moderators of susceptibility to nudges. We have two measures of risk aversion, a self-reported $D F$ and elicited $H L$. When these two are congruent, that is, people show both high (low) self-reported and elicited risk aversion, the susceptibility to nudges is lowest, at 0.04 . When these are incongruent, and people selfreport higher risk aversion but elicited preferences show riskseeking behavior, the susceptibility to nudges is higher, at 0.06 , showing a higher discounting of segregated costs. Most interestingly, when these are incongruent, and people selfreport lower risk aversion but elicited preference show higher risk aversion, the susceptibility to nudges is in the reverse direction, at -0.11 . This can be interpreted as an over-weighting of segregated costs, leading to a nudge away from choices that had a higher framing effect. The combination of risk congruency and self-reported risk aversion have a significant (LBF 4.7) effect on susceptibility to nudges, and represents a source of significant heterogeneity.

## (5) How optimal is allocation behavior?

One of the most popular normative theories of resource allocation is modern portfolio theory (MPT), characterized by mean-variance optimization (Markowitz, 1952). It states that people should select weights that optimize the balance between the expected value and standard deviation of the resulting portfolio. The optimization is a function of a risk tolerance factor $Q$, with lower values of $Q$ indicating preference for safer portfolios and high values of $Q$ indicating preference for riskier portfolios. Given a set of prospects, the theory proposes an efficient frontier of possible weight allocations that result in optimization between the desired portfolio mean $(V)$ and variance (standard deviation $D$ ). Given an implicit objective to maximize $V$ and minimize $D$, the frontier represents portfolio choices such that no other combination of weights can result in an increase in $V$ without an increase in $D$, or a decrease in $D$ without a decrease in $V$. Note that the efficient frontier does not depend on risk preference $Q$, but where the selected portfolio lies along the efficient frontier is dependent on the individual preference parameter $Q$. The set of weights $(x)$ on the efficient frontier for a particular value of $Q$ can be found by minimizing the expression : $x^{T} \Sigma x-Q E^{T} x$. Here $E$ is a vector of expected returns on the individual prospects and $\Sigma$ is the covariance matrix for the returns on the prospects.

Actual portfolios constructed by participants may not lie on the efficient frontier. For any observed portfolio allocation we can calculate the minimum distance of the observed portfolio characteristics from the efficient frontier, which gives the smallest distance to optimality. This distance is dependent on the mean and SD values of individual prospects within a choice set. To enable comparison across choice sets and analyze the impact of factors we calculate the ratio of minimum distance to optimality for the observed portfolio to the largest possible minimum distance to optimality for any combination of weights in the choice set. This is denoted as MPT-error, $d_{\varepsilon}$. The risk tolerance value corresponding to the closest point on the efficient frontier is denoted $Q_{\varepsilon}$, and can be inferred to be the risk tolerance level for that choice. Note that in SOSD trials, all prospects have the same EV, and the efficient frontier is a single point that coincides with $100 \%$ allocation to the
safest prospect. Thus, SOSD trials provide a parameter free estimate of $d_{\mathcal{\varepsilon}}$, but do not allow an estimate for $Q_{\varepsilon}$.

Table 1 provides the mean MPT-error and risk tolerance levels by design factor levels. Conducting a Bayesian ANOVA analysis and comparing against a null model that included participants as random effects, we find evidence for the effect of SOSD (LBF $\infty$ ), domain (LBF $\infty$ ), and skew (LBF 2.3) on $d_{\varepsilon}$. We find evidence for the effect of domain (LBF 36.7) and cost framing (LBF 1.1) on $Q_{\varepsilon}$. The distance to optimality $d_{\varepsilon}$ measured as a percentage of the largest possible distance to optimality has a mean of $(31 \% ; 18 \%)$ for SOSD and non-SOSD choice sets; $(31 \% ; 18 \%)$ for gains and mixed domains; and ( $24 \%$; 22\%; 27\%) for no skew, negative skew and positive skew choice sets. As a comparison, $d_{\varepsilon}$ for a $1 / \mathrm{N}$ portfolio would be $(28 \% ; 13 \%)$ for SOSD and nonSOSD sets, $(25 \% ; 16 \%)$ for gains and mixed domains, and $20 \%$ for all skew sets. On an average, the actual allocations that people make are less optimal from a mean-variance optimization standpoint than what a simple $1 / \mathrm{N}$ heuristic would result in. The mean $d_{\varepsilon}$ is $(22 \% ; 26 \% ; 25 \%)$ in the no-framing, and 2 framing conditions. Although the differences are not statistically significant, directionally, framing conditions lead people further away from mean-variance optimality.

The mean inferred risk tolerance $Q_{\varepsilon}$ is 20 in the gains domain and 107 in the mixed domain, indicating that risk tolerance is highly contextual, rather than a stable trait. The mean inferred value of risk tolerance $Q_{\varepsilon}$ is 67 in the no-framing condition, 76 in the higher-riskier-framing condition, and 48 in the higher-safer-framing condition, reflecting sensitivity of risk tolerance to framing effects. Measures of risk traits (SL, HL, DF) do not have any effect on the closest distance to optimality. Evidence is inconclusive (LBF 0.9) for the effect of SL on $Q_{\varepsilon}$.

## Conclusion

The key findings can be summarized as: (1) Only a very small subset of participants follow a naive diversification or $1 / \mathrm{N}$ heuristic. (2) Design factors such as domain, skew and SOSD across options influence the allocation that people make in extreme (safest or riskiest) prospects, directionally similar to the effect that these factors have in single choice gambles. (3) We show that individual traits traditionally linked to risk propensity seem to manifest in terms of the degree to which people are inclined to diversify. Lower risk aversion and higher risk seeking traits are linked to increasing diversification. These traits do not seem to be consistently linked to risk tolerance when measured within the MPT framework, and do not seem to influence the relative levels of risk and safety observed in resource allocation behavior. The results are counter-intuitive to the popular notion that diversification is linked to a reduction of risk. (4) We find that cost framing nudges affect the level of diversification. While the effect of nudges seems insignificant at an overall level, a deeper analysis shows traitbased clusters. We find that risk congruency, whether peo-
ples' elicited and self-reported risk aversion are congruent, is a strong moderator for susceptibility to nudges. We find evidence for heterogeneous clusters where people either underweight or over-weight segregated costs leading to the same nudge producing opposite behavioral results in the two risk incongruent groups. (5) We find that people are not optimal under a mean-variance optimization objective, and that on an average, a $1 / \mathrm{N}$ heuristic is closer to mean-variance optimization than the actual observed behavior.

## References

Bardolet, D., Fox, C. R., \& Lovallo, D. (2011). Corporate capital allocation: A behavioral perspective. Strategic Management Journal, 32.
Beattie, J., \& Loomes, G. (1997). The impact of incentives upon risky choice experiments. Journal of Risk and Uncertainty, 14.
Benartzi, S., \& Thaler, R. H. (2001). Naive diversification strategies in defined contribution saving plans. American economic review.
Blais, A.-R. (2006). A domain-specific risk-taking (dospert) scale for adult populations. Judgment and Decision Making, 1(1), 33-47.
Deck, C., \& Schlesinger, H. (2010). Exploring higher order risk effects. The Review of Economic Studies, 77.
Holt, C. A., \& Laury, S. K. (2002). Risk aversion and incentive effects. American economic review, 92(5).
Huberman, G., \& Jiang, W. (2006). Offering versus choice in 401 (k) plans: Equity exposure and number of funds. The Journal of Finance, 61(2), 763-801.
JASP-Team. (2016). Jasp (version 0.8.0.0). [Computer software].
Lopes, L. L., \& Oden, G. C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and sp/a theory. Journal of mathematical psychology, 43.
Markowitz, H. (1952). Portfolio selection. The journal of finance, 7.
Rhoades, S. A. (1993). The herfindahl-hirschman index. Fed. Res. Bull., 79, 188.
Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological monographs: General and applied, 80.
Symmonds, M., Wright, N. D., Bach, D. R., \& Dolan, R. J. (2011). Deconstructing risk: separable encoding of variance and skewness in the brain. Neuroimage, 58.

# Predicting Individual Differences in Working Memory Training Gain: A Machine Learning Approach 

Shafee Mohammed<br>University of California Irvine, Irvine, California, United States<br>Benjamin Katz<br>University of Michigan<br>Chelsea Parlett<br>University of California, Irvine<br>Martin Buschkuehl<br>MIND Research Institute, Irvine, California, United States<br>Susanne Jaeggi<br>University of California, Irvine


#### Abstract

Working memory (WM) capacity is critically important for the success in school and complex cognitive activities across the lifespan. Training WM skills has shown to lead to improvements in a variety of important cognitive tasks. One's performance on an adaptive and challenging longitudinal WM intervention may serve as an assay of cognitive plasticity. With over 400 participants having completed a minimum of 15 sessions of WM training, we have a rich dataset that allows investigating individual differences and other factors that might determine training outcome using a novel machine learning techniques. Preliminary results suggest that factors such as age, type of n-back, and baseline abilities significantly impact one's ability to improve in training. Other factors such as gender and whether or not training was supervised were not significant. Finally, our model allows prediction of training gain with $78 \%$ accuracy.


# A Longitudinal Study of Differences between Predicted, Actual, and Remembered Personal Change 

Sarah Molouki (smolouki@uchicago.edu)<br>Daniel M. Bartels (bartels@uchicago.edu)<br>Oleg Urminsky (oleg.urminsky@chicagobooth.edu)<br>University of Chicago<br>5807 S. Woodlawn Ave., Chicago, IL 60637


#### Abstract

We investigated people's assessments of their own personal change over time, comparing predicted, actual, and recalled change in personality, values, and performance. On average, participants underestimated the absolute magnitude of their personal change in both prediction and recall. However, people specifically neglected negative future change, resulting in overly optimistic predictions of improvement. In contrast, recall of positive and negative change was relatively more balanced, such that assessments of past improvement were better calibrated on average. Our findings provide insight into how people think about their own identity over time and address disparate theories in the literature regarding predictions of personal stability versus improvement.


Keywords: self-perception; social cognition; future self; past self; identity; time; personal change

## Introduction

Imagine yourself ten years in the future. Will you be nearly the same person you are today, just with grayer hair? Or will you be a significantly changed person with different abilities, values, and personality characteristics? If you will have changed, will you have improved, becoming wiser and kinder, or will you have taken a turn for the worse, ending up lazy and irresponsible?

A large literature on beliefs about personal change suggests that people tend to perceive a trajectory of improvement in their own lives. People claim to possess more desirable characteristics in the present than they did in the past (Wilson \& Ross, 2001), and think they will continue to get even better in the future (Haslam, Bastian, Fox, \& Whelan, 2007; Kanten \& Teigen, 2008). They even expect basic personality traits to improve over most of their lifespan (Krueger \& Heckhausen, 1993), despite some evidence indicating that these traits are mostly unchanging (e.g., Costa \& McCrae, 1989).

Expectations of improvement over time have been associated with people's normative theories about personal identity. For example, Newman, Bloom, \& Knobe (2014) find that when evaluating change in others, people associate improvements, but not declines, with a person's core identity, or "true self." In this view, positive change is a natural part of human development. Other work highlights the finding that people consistently predict greater personal improvement for themselves than for others (e.g., Haslam et al., 2007, Kanten \& Teigen, 2008), suggesting possible self-enhancement motives.

However, separate lines of research suggest that people view their own identity as stable over time. Loewenstein, O'Donoghue, \& Rabin (2003) describe a systematic tendency to overestimate the degree to which one's own preferences will persist into the future. More recently, Quoidbach, Gilbert, \& Wilson (2013) described an "End of History" illusion. In this framework, people reported more change in the past than they anticipated for the future in various domains, possibly due to the difficulty of envisioning new changes in prospect.

It is unclear how to reconcile findings related to belief in self-improvement with research suggesting people underestimate personal change more generally. Do people in fact think that they have stopped changing, or do they believe the biggest improvements are yet to come? Because many studies highlighting people's difficulty in projecting future change have used changes with no obvious direction (e.g., preference change), these have not been directly connected to ideas about improvement or decline. Are these two effects (expectations of improvement and perceptions of stability) in fact conflicting, or do they co-exist? Furthermore, many studies of personal change over time examine either perceptions or actual change, but not both. Are people's predictions and recall of change well-calibrated with their actual change, or do they diverge?

We address these questions by examining both absolute and directional personal change using a repeated measures longitudinal design. Although Quoidbach et al. (2013) found that one sample of people tended to predict less future personal change than the other sample remembered experiencing in the past, the study did not assess actual change in individuals over time. In the current studies, we measure and directly compare predicted, actual, and recalled change within each of our samples of young adults. We use actual change as a baseline to determine whether any observed differences between predicted and remembered change are due to biases in estimating future change, distorted memories of past change, or both. Furthermore, we use measures of personality, values, and ability that allow us to examine change both in magnitude and direction.

Our findings reconcile potentially conflicting viewpoints in the literature by suggesting that although people do underestimate the magnitude of their own change, this finding is moderated by the valence of
change (i.e., whether the change is positive or negative). Averaging across our measures, we find that people are fairly well-calibrated in recall, but specifically underappreciate the potential for future decline. Because people's predictions tend to omit undesirable changes while acknowledging positive change, they simultaneously underestimate absolute change and overestimate (future) improvement on average.

## Methods

We compared predicted, actual, and recalled personal change over time in two panel surveys of young adults, who completed all measures online. Study 1 assessed 155 participants $(60 \%$ female, mean age $=22.1)$ in December 2013 (Time 1) and December 2014 (Time 2), and Study 2 assessed 203 participants ( $73 \%$ female, mean age $=22.6$ ) in May 2016 (Time 1) and September 2016 (Time 2). Study 2 was a conceptual replication that addressed several additional questions raised by the results of Study 1. (Key differences between the two studies are indicated in the "Measures" section). For Study 1, college students (in any year of college) were recruited from across the United States using an online panel. In Study 2, participants were graduating college seniors recruited by the experimenters. Although the time period measured in Study 2 (4 months) was shorter than that in Study 1 (1 year), all participants in Study 2 would be undergoing a major life change (i.e., college graduation) during the study period, making it plausible that significant changes in personality and values could occur between the two assessments.

In both studies, we measured change in personality traits and values, and in Study 2 we also measured change in performance on an objective (knowledge) task. Personality and values are viewed as important psychological determinants of personal identity (Bartels \& Rips, 2010; Chen, Urminsky, \& Bartels, 2016) and have been used in previous research on perceptions of personal change (Quoidbach et al., 2013). We chose these constructs rather than other personal attributes (such as preferences) because change in personality, values, and performance can be measured both directionally (increase vs. decrease) as well as in terms of absolute difference. The performance measure is included to verify that our findings are replicable in a domain where actual change is measured objectively rather than by taking a difference in self-report measures.

Participants provided current measures of personality and values (and performance, in Study 2) at both Time 1 and Time 2. At Time 1, they also predicted what their responses would be at Time 2. At Time 2, they provided their recollection of their responses at Time 1 . Reports of current values were always made before reports of predicted or recalled values.

## Measures

Personality In Study 1, participants completed a 5-item personality assessment that involved reading a short description of each Big Five dimension (i.e., extraversion,
agreeableness, conscientiousness, emotional stability, and openness) and judging how much they thought this trait applied to them. Ratings of current, predicted and remembered personality were reported using a $0-100$ slider scale for each trait. In Study 2, participants again completed this 5 -item personality measure, and also completed a previously validated 10 -item measure with a 7-point response scale (TIPI; Gosling, Rentfrow, \& Swann, 2003; also used in Quoidbach et al., 2013). ${ }^{1}$

Values Values were assessed using a 10 -item version of the Schwartz Value Inventory (Lindeman \& Verkasalo, 2005; also used in Quoidbach et al., 2013), measuring the personal importance of self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence, and universalism. Current, predicted, and remembered value importance ratings were measured on a $9-$ point scale.

Performance In Study 2, participants also answered ten factual multiple-choice questions at each time point (e.g., "How many of the world's tallest buildings are located in the United States?"). Both sets of 10 questions were pretested to ensure that they were of similar difficulty and that there were no floor or ceiling effects. At Time 1, participants reported perceptions of their current and future performance, and at Time 2 they reported their perceived current and past performance.

## Pre-Test to Determine Valence of Change

We used a separate online sample ( $N=100 ; 41 \%$ female, mean age $=34.3$ ) to assess how people generally view increases and decreases in the characteristics of interest. For each personality dimension and value, participants used a 0 100 scale to separately report how they would feel ( $0=$ extremely displeased, $50=$ neither pleased nor displeased, 100=extremely pleased) if the given characteristic were to increase and decrease. For each of the five personality dimensions as well as 9 out of 10 values, the average response was significantly greater than 50 for increases (suggesting people consider increases to be desirable) and significantly less than 50 for decreases (suggesting people consider decreases to be undesirable). ${ }^{2}$ Accordingly, for our basic directional analyses, we treat personality and value increases as improvements, and decreases as declines.

In Study 2 (after all other measures were completed), we also asked individual participants to report whether they viewed each personality or value change as an improvement or decline, in order to better account for individual variation in these beliefs. Coding each individual's changes as improvements and declines based on their own assessments

[^415]rather than the mean assessment from the pretest did not change the overall pattern of our results. However, future research can further explore the effect of individual differences in perceptions of valence of these characteristics, as well as perceived differences in the magnitude of their importance.

## Change Calculation

We quantify change in two different ways, looking at both absolute change and directional change. Absolute change (i.e., the absolute difference between an item at Time 2 and Time 1) represents deviation from the present state, ignoring direction. Averaging absolute change across items captures the distinction between variability (high values indicating large changes in any direction) and stability (low values indicating little change). This approach was taken in some prior work, including Quoidbach et al. (2013).

However, looking only at absolute change neglects the fact (confirmed in our valence pre-test) that increases and decreases generally differ in desirability. For example, an increase in creativity might be considered an improvement, but a decrease of equivalent magnitude might be a decline. Measuring only absolute change obscures this distinction.

We therefore also computed measures of directional change (i.e., directional difference in Time 2 minus Time 1 ratings). Averaging across individual items thereby captures the distinction between overall improvement (high positive values indicating net positive change), overall stasis (nearzero values indicating no net change), and overall decline (high negative values indicating net negative change).

In our analyses of personal change, we compare the following: (i) predicted change (difference between future prediction provided at Time 1 and current rating provided at Time 1), (ii) actual change (difference between current rating provided at Time 2 and current rating provided at Time 1), and (iii) remembered change (difference between current rating provided at Time 2 and past recollection provided at Time 2). Using both absolute and directional measures allows us to examine a) whether participants perceive a smaller absolute magnitude of personal change than they actually undergo over time and b) whether participants overestimate their net improvement over time, for both prediction and recall.

## Results

Across our two studies, we have 6 distinct measures for which we examine change over time. For each of these measures, we conducted both an absolute and a directional comparison of the three types of change (predicted, actual, and recalled). Overall, we found that people predicted future improvement, but systematically neglected the possibility of future decline. This resulted in both an overall underestimation of mean absolute future change and an overestimation of mean future improvement. In contrast, people were more balanced in recall, remembering both positive and negative past change. Thus, although people
underestimated the magnitude of past change, they did not express a directional bias in recall.

Although the overall pattern of our results supports this finding, results across measures and studies were highly variable. We report weighted mean effect sizes (Cohen's $d$ ) across all studies in the text of the paper to summarize our overall findings, and present figures depicting the findings of each individual measure to capture the variability across them.

Absolute Change Across our six domains of measurement, we found significant differences in absolute magnitude between predicted, actual, and recalled change. Overall, participants predicted that they would undergo less personal change in the future than they recalled undergoing in the past (mean $d=-0.18, p<.001$ ), which replicates previous findings (Quoidbach et al., 2013). However, this effect was small in comparison to their larger tendency to underestimate the magnitude of both past (mean $d=-0.51$, $p<.001$ ) and future (mean $d=-0.72, p<.001$ ) change relative to actual change. This general pattern was found in all absolute measures except for the 10 -item personality measure used in Study 2 (see left hand side of Figures 1-4).


Figure 1. Absolute and directional change for 5-item personality measure. Error bars represent $95 \%$ CI.
n.s. nonsignificant, $* p<.05, * * p<.01, * * * p<.001$


Figure 2. Absolute and directional change for 10 -item personality measure (Study 2). Error bars represent $95 \%$ CI. n.s. nonsignificant, $* p<.05, * * p<.01, * * * p<.001$


Figure 3. Absolute and directional change for measure of value importance. Error bars represent $95 \%$ CI.
n.s. nonsignificant, $* p<.05, * * p<.01, * * * p<.001$


Figure 4. Absolute and directional change in performance, measured as number of questions answered correctly. (Study 2). Error bars represent 95\% CI.
n.s. nonsignificant, $+p<.10, * p<.05,{ }^{* *} p<.01,{ }^{* * *} p<$ . 001

Net Directional Change We conducted a directional analysis that accounts for valence, treating net increases in all measures as improvements and net decreases as declines (based on the results of our pre-test). On average across our measures people predicted more positive future change than they subsequently remembered (mean $d=0.30, p<.001$ ). Our analysis also revealed a tendency to predict more positive net future improvement than actually occurred (mean $d=0.22, p<.001$ ). In contrast, people's recall of past improvement was more variable. Although in Study 1, participants recalled greater improvement in personality and values than they had experienced, recalled improvement was either equal to or less than actual improvement for all measures in Study 2 (right hand side of Figures 1-4). Across all measures in Studies 1 and 2, average recall of past directional change was not significantly different from actual change (mean $d=-0.03, p=.377$ ).

Individual-level "Improvers" and "Decliners" How do we explain the fact that on average, participants simultaneously predicted less absolute change and greater improvement than they actually experienced? To better understand the observed mean-level effects, we separated individuals based on whether they exhibited an overall increase versus decrease in their predicted, remembered, and actual change. For most of our measures across both studies 1 and 2, actual directional change across the sample was near zero. However, this was not because people had remained stable (as evidenced by the large absolute change findings); rather, the sample was evenly split into those who had experienced net improvement and those who had experienced net decline of equal magnitude. ${ }^{3}$ In contrast, predicted change was significantly positive for each measure because fewer individuals predicted that they would decline over the study period, and those that did so reported declines of significantly smaller magnitude than the average decline experienced. Figure 5 illustrates this pattern using the personality and value measures from Study 1.


Figure 5. Decomposing directional change in Study 1 into decline and improvement for a) 5-item personality measure, b) values measure. Column width indicates proportion of participants indicating net improvement or decline and height denotes magnitude of improvement or decline. Error bars represent $95 \%$ CI.

[^416]We then performed separate analyses on those who had actually improved versus those who declined. This revealed that the underestimation of absolute magnitude of change at the sample level in fact comes disproportionately from those participants who are underpredicted their actual decline (rather than from those who underpredicted their actual improvement). Although similar patterns are observed for each of our measures, we describe the 5-item personality measure from Study 1 in detail to illustrate the form of this effect. Participants who experienced an actual net decline in the measured traits over the year $(\mathrm{M}=-12.25, \mathrm{SD}=11.65)$ had instead predicted a mean personality improvement of +3.62 ( $\mathrm{SD}=6.67$ ). The difference between these two numbers reflects a significant directional overprediction of improvement by 15.87 scale points, $t(77)=10.87,95 \%$ CI=[12.96,18.78], $d=1.23, \quad p<.001$. Nevertheless, the absolute magnitude of their predicted change was still smaller than the absolute magnitude of their actual change (i.e., 3.62 vs. $12.25 ; t(77)=5.47,95 \% \quad \mathrm{CI}=[5.48,11.77]$, $d=0.62, p<.001$ ). This yields an underprediction of change when future change is defined only in terms of absolute deviation from zero. In contrast, those who experienced an actual net personality improvement $(\mathrm{M}=+13.07, \mathrm{SD}=11.44)$ had predicted an improvement of +7.73 ( $\mathrm{SD}=7.76$ ). This reflects an underestimation of both their directional improvement and their absolute change by 5.34 scale points, $t(76)=4.05,95 \% \mathrm{CI}=[2.72,7.97], d=0.55, p<.001$.

Comparing the size of these prediction errors reveals that actual decliners made significantly larger errors on average than actual improvers (i.e., 15.87 vs. 5.34 ); $t(152)=5.35$, $95 \% \mathrm{CI}=[6.64,14.42], d=0.86, p<.001$. As a result, we observe an overall improvement bias in the sample, given that the errors made by decliners were biased in the direction of positive change. Figure 6 provides a graphical depiction of this phenomenon.


Figure 6: Comparison of actual (x-axis) versus predicted (yaxis) personality change (Study 1). Each point on this graph represents an individual participant from Study 1, with black dots representing those who actually declined over the period and white circles representing those who improved.

The light gray line represents the line $y=x$ (actual improvement or decline=predicted improvement or decline), which is where each point would fall if predictions were completely accurate. The black line is the regression line relating predicted and actual improvement as fitted from the data. The discrepancy between the black and gray lines to the left of the $y$-axis illustrates the overprediction of improvement in those who actually declined, and the discrepancy on the right side of the graph illustrates the (smaller) underprediction of improvement in those who actually improved.

The finding that those who actually declined underpredicted their decline to a greater extent than improvers underpredicted their improvement was observed across all measures: personality $(t(152)=5.35$, $95 \%$ $\mathrm{CI}=[6.64,14.42], d=0.86, p<.001)$ and values $(t(142)=4.53$, 95\% CI=[0.31, 0.80], $d=0.73, p<.001$ ) in Study 1, and both 5 -item $(t(111)=4.02,95 \% \mathrm{CI}=[3.02,8.88], d=.65 ; p<.001)$ and 10 -item $(t(169)=7.82,95 \% \mathrm{CI}=[0.51,0.86], d=1.17 ; p$ <.001) personality measures, values $(t(182)=2.95,95 \%$ $\mathrm{CI}=[.09, .47], d=.43 ; p=.004)$, and performance $(t(152)=$ $2.56,95 \% \mathrm{CI}=[.14,1.09], d=.40, p=.011)$ in Study 2. However, there was no consistent finding in errors related to recall across studies. In Study 1, those who actually declined made larger errors than those who actually improved for measures of both personality and values. In contrast, across measures in Study 2, actual improvers made errors of equivalent or greater size as actual decliners did in recall.

## Discussion

Overall, the results of our two longitudinal studies reveal that people predict smaller absolute change in personality, values, and performance than they actually experience. However, a directional analysis of the data reveals that this discrepancy is specifically driven by a neglect of negative change. Although on average people experience both improvements and declines over time, they incorrectly predict that their future will consist mainly of improvements. Rather than being the end of their personal trajectory, the present moment represents a watershed of a different sort: the moment when people think a somewhat rocky past resolves into a consistent upward climb. Our findings suggest that looking only at the absolute magnitude of change may obscure important aspects of people's beliefs about their own personal change.

Although people do underestimate the magnitude of their future change, our directional measures reveal that this is not generally caused by expectations of stability in personal characteristics. Rather, at the sample level, those who worsened over our study period were disproportionately likely to have neglected the possibility of decline and instead predicted some smaller level of improvement. Thus, the apparent magnitude effect seems to be driven by a general tendency to overestimate improvement, which is consistent with prior research suggesting that people expect the continued development and emergence of positive
personal characteristics (Haslam et al., 2007; Newman et al., 2014). Previous work also suggests that although people are able to distinguish their future expectations from conceptions of their ideal self, predictions are nonetheless likely to be influenced by aspirations (Molouki \& Bartels, 2017). This may be one mechanism underlying the observed overpredictions of improvement.

Although we found that people overpredicted future improvement, on the whole, we did not find this effect in recall of past change. This may be because past recall is more constrained by reality and episodic facts than future prospection (Kane, McGraw, \& Van Boven, 2009), making people more likely to acknowledge that past decline has occurred even if this was counter to their expectations. However, other research suggests that people do revise their perceptions of the past towards a trajectory of improvement (e.g., Wilson \& Ross, 2001). Although we did find that people overestimated past improvement in Study 1, this pattern did not emerge in Study 2. Further research is needed to explore the source of this heterogeneity (and the large heterogeneity across our measures more broadly), by examining effects of contextual factors such as length of time span, intervening life events, dimension of change, and timing of measurement. In particular, a more careful investigation of our personality measures is needed, as some discrepancies were noted between the 5 -item and 10 -item personality scales used.

In addition to providing a reconciliation of previous findings about perceptions of stability versus improvement, the current work makes an important contribution more broadly to a growing body of literature on people's theories of the nature and persistence of personal identity. Existing research in this area has suggested that people endorse normative theories about a fundamentally good essence that forms the core of one's identity and will be revealed over time (Molouki \& Bartels, 2017; Newman et al., 2014; Newman, De Frietas, \& Knobe, 2015; Strohminger, Knobe, \& Newman, in press; Tobia, 2015). The current research explicitly demonstrates that predictions of personal improvement are more pronounced than actual improvement over time. This finding lends empirical support to the idea that predictions may be influenced by normative beliefs that diverge from a purely descriptive account of personal development. Furthermore, we noted new findings about different patterns of prediction error for those who in fact improved versus declined over the study period. Future research can further explore the interactions between specific developmental trajectories and beliefs about personal identity.

## Acknowledgments

This research was supported by funding from the John Templeton Foundation and the James M. Kilts Center for Marketing at the University of Chicago Booth School of Business. We thank Halley Bayer, Stephanie Chen, Ed O'Brien, Jordi Quoidbach, Anne Wilson, and several anonymous reviewers for helpful comments/suggestions.

## References

Bartels, D.M, \& Rips, L.J. (2010). Psychological connectedness and intertemporal choice. Journal of Experimental Psychology: General, 139, 49-69.
Costa, P. T., \& McCrae, R. R. (1989). Personality continuity and the changes of adult life. In M. Storant \& G. R. Vanderbos (Eds.), The Adult Years: Continuity and Change (pp. 45-77). Washington, DC: APA.
Chen, S., Urminsky, O. \& Bartels, D. M. (2016). Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. Psychological Science, 27(10), 1398-1406.
Gosling, S. D., Rentfrow, P. J., \& Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37(6), 504-528.
Haslam, N., Bastian, B., Fox, C., \& Whelan, J. (2007). Beliefs about personality change and continuity. Personality and Individual Differences, 42(8), 1621-1631.
Kane, J., McGraw, A. P., \& Van Boven, L. (2009). Temporally asymmetric constraints on mental simulation: Retrospection is more constrained than prospection. The Handbook of Imagination and Mental Simulation, (pp. 131-149). New York: Psychology Press.
Kanten, A. B., \& Teigen, K. H. (2008). Better than average and better with time: Relative evaluations of self and others in the past, present, and future. European Journal of Social Psychology, 38(2), 34-353.
Krueger, J., \& Heckhausen, J. (1993). Personality development across the adult life span. Journal of Gerontology, 48(3), 100-108.
Lindeman, M. \& Verkasalo, M. (2005). Measuring values with the Short Schwartz's Value Survey. Journal of Personality Assessment, 85(2), 170-178.
Loewenstein, G., O'Donoghue, T., \& Rabin, M. (2003). Projection bias in predicting future utility. The Quarterly Journal of Economics, 118(4), 1209-1248.
Molouki, S., \& Bartels, D. M. (2017). Personal change and the continuity of the self. Cognitive Psychology, 93, 1-17.
Newman, G. E., Bloom, P., \& Knobe, J. (2014). Value judgments and the true self. Personality and Social Psychology Bulletin, 40(2), 203-216.
Newman, G. E., De Frietas, J., \& Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. Cognitive Science, 39(1), 96-125.
Quoidbach, J., Gilbert, D. T., \& Wilson, T. D. (2013). The end of history illusion. Science, 339(6115), 96-98.
Strohminger, N., Newman, G., \& Knobe, J. (in press). The true self: A psychological concept distinct from the self. Perspectives in Psychological Science.
Tobia, K. P. (2015). Personal identity and the Phineas Gage effect. Analysis, 75(3), 396-405.
Wilson, A. E., \& Ross, M. (2001). From chump to champ: people's appraisals of their earlier and present selves. Journal of Personality and Social Psychology, 80(4), 572584.

# Visual Data Exploration: How Expert Astronomers Use Flipbook-Style Visual Approaches to Understand New Data 

Fernanda Monteiro Eliott ${ }^{a}$ (fernanda.m.eliott @ vanderbilt.edu)<br>Keivan Stassun ${ }^{b}$ (keivan.stassun@ vanderbilt.edu)<br>Maithilee Kunda ${ }^{a}$ (mkunda@vanderbilt.edu)<br>${ }^{a}$ Department of Electrical Engineering and Computer Science, ${ }^{b}$ Department of Physics and Astronomy Vanderbilt University, PMB 351679, 2301 Vanderbilt Place, Nashville, TN 37235-1679, USA


#### Abstract

What are the cognitive processes in play when someone uses a visualization tool to interactively explore a new dataset? Here, we focus on one particular type of visualization-the scatter plot-which, despite (or perhaps because of) its simplicity, is still one of the most frequently used plot types in many dataintensive disciplines. We conducted a pilot study to investigate how expert astronomers interact with an unfamiliar dataset using a visualization tool called Filtergraph, which supports rapid and easy visualization of large datasets. We present both qualitative and quantitative results, including observations about the temporal dynamics of visual data exploration as well as interesting behavioral patterns that we saw in our participants, such as users taking "circular walks" through the data at various levels of abstraction.


Keywords: Data exploration; graph understanding; information visualization; scatter plots; visualization software.

## Introduction

When astronomer Henry Norris Russell first introduced the now-called Hertzsprung-Russell (H-R) diagram, he wrote, "The appearance of [the figure] suggests the hypothesis that, if we could put on it some thousands of stars, instead of the 300 now available, ...we would find the points representing them clustered principally close to two lines, one descending sharply along the diagonal...the other ...running almost horizontally. ...These two classes of stars were first noticed by Hertzsprung, who has applied to them the excellent names of giant and dwarf stars" (Russell, 1914, p. 287).

In addition to Russell's obvious desire for more data, his wonderfully vivid description conveys the fundamentally visual nature of this discovery. Indeed, the H-R diagram has been called "perhaps the most spectacularly successful example of a simple scatterplot" in all of science (Spence \& Garrison, 1993, p. 1). Today, like many disciplines, astronomy enjoys volumes of data that Russell could only have imagined. However, an astronomer's expertise to make sense out of data-to recognize which patterns represent actual scientific discovery-remains as vital today as it was in 1914.

Most human sense-making with data involves a visuallymediated interaction between the data and the perceptual/cognitive processes of the user-data visualization. Data visualization can be as short and simple as glancing at a printout of a plot on paper, or as lengthy and complex as spending months analyzing and modeling a large dataset. There is an increasing need for interactive data visualization tools that not only leverage the latest in pattern recognition and data mining algorithms, but also place the cognitive needs of the user
front and center-to assist and augment human capabilities in the discovery process (Honavar, Hill, \& Yelick, 2016).

One vital role for data visualization is the open-ended, open-minded exploration of data that leads to unexpected insight, often manifested at first as a "striking" or "interesting" multivariate plot, such as with the H-R diagram. To take a more recent example, an astronomy research team at Vanderbilt University developed a visualization tool called Filtergraph (see Figure 1) designed to allow people to rapidly and easily explore large datasets of up to a few million points (Burger et al., 2013). Using the Filtergraph software to visualize stellar variability data gathered by the Kepler spacecraft resulted in an unexpected and "visually interesting" scatterplot, which in turn led to the discovery of stellar granulation "flicker" and its utility for stellar and exoplanets research, a significant finding that was published in Nature (Bastien, Stassun, Basri, \& Pepper, 2013).


Figure 1: A screenshot of the Filtergraph data visualization interface (Burger et al., 2013), and also the initial view shown to participants in our pilot study.

Here, we report our preliminary results on a pilot study to investigate how expert astronomers interact with an unfamiliar dataset using scatterplots generated by Filtergraph. We use a novel analysis approach that is different from, but complementary to, existing methods that focus on measuring actions or tasks conducted by the data analyst. Instead, our approach measures interactions in terms of dataset attributes: which variables from a large dataset does an analyst look at, when, and why? We present both qualitative and quantitative results from this study, including observations about the temporal dynamics of visual data exploration as well as interesting behavioral patterns that we saw in our participants, such as users taking "circular walks" through the data at various levels of abstraction.

## Related Work

There is a rich foundation of work in HCI (Human Computer Interaction), visual analytics, and infovis that aims to understand the processes by which people interact with and understand data. Sanderson and Fisher (1994) present a widelyused, general framework for thinking about user interactions with an interactive tool in terms of sequences of actions that represent different user functions, such as connecting ideas or introducing comments.

Specifically in relation to data visualization, Yi and colleagues (2007) identify seven modes of interaction with visualization tools, such as select, explore, and filter, that they believe are important for understanding the visual sense-making process. Brehmer and Munzner (2013) present a task typology that bridges low-level interactions with high level tasks and goals during visualization activities. Pirolli and Card (2005) present a detailed cognitive task analysis of visual sensemaking in the domain of intelligence analysis. ElTayeby and Dou (2016) present methods for studying exploratory data visualization that leverage the automated analysis of rich, quantitative interaction $\log$ data to identify and understand underlying patterns of interaction.

Saraiva and colleagues (2005) conducted a very interesting pilot study specifically focused on open-ended, exploratory data analysis in the domain of bioinformatics. They focus on how a visualization tool can complement a dataset in order to facilitate insights that may lead to a discovery. One important issue they addressed was how to define and measure insight. They defined insight based on the context of their work as well as based on characteristics such as time, hypotheses, correctness, and category. Their results show the influence of the visualization tool itself over the processes of human interpretation and insight.

Mayr and colleagues (2016) looked at how mental models parallel a user's use of external visualization tools. They identify key characteristics pertaining to mental models, such as content, structure, coherence, perspectivity, generalizability, and utility, and they review existing empirical methods for conducting user studies to get at these characteristics.

Finally, there is important work being done to understand data visualization and sensemaking in terms of core cognitive capabilities that people bring to bear on such tasks. Healey and Enns (2012) provide a research survey on cognitive theories of attention and perceptual processing as applicable to data visualization. They provide examples of factors that drive visual attention (such as visual feature hierarchies, memory, and prediction) and also factors that impair attention, such as what happens during change blindness, with observations about how improperly designed visualizations can significantly impact a user's mental models of the data.

Tversky (2003) reflects about humans, actions, and space, emphasizing the importance of differences in how people represent space across different spatial reference frames (e.g., in a navigation task versus a graphical understanding task). She includes discussion of the spatial references frames people
use while processing external visualizations of information.

## Methods

We conducted a pilot study to investigate how expert astronomers interact with an unfamiliar dataset using the Filtergraph visualization tool shown in Figure 1. We chose the domain of astronomy as being representative of today's dataintensive disciplines. In designing our study, we wanted to choose a data exploration task that was open-ended enough to provide a realistic challenge to our participants, but also that had at least some constraints so that we could analyze meaningful differences across participants. We decided to invite astronomers to our lab to participate in one-hour sessions, during which they would be instructed to "explore" an astronomy dataset that they had not previously seen. All necessary IRB approvals were obtained prior to the study.

Dataset. We chose a dataset that we believed would not require complex mathematical operations to make sense of, and also that would not be from too specialized a subfield within astronomy. The dataset we chose is described in Berlind et al. (2006) and contains data describing 90, 893 galaxies, which are individually discriminated through 10 attributes (including the galaxy ID), or group separated through 9 attributes (including the group ID), as shown in Table 1.

Table 1: The 19 attributes in the galaxy dataset used for our pilot study (Berlind et al., 2006), along with letter codes used throughout this paper. The last two shaded items represent arithmetic combination of attributes that we observed participants construct on the fly during the study sessions.

| Attribute | Code | Attribute | Code |
| :--- | :---: | :--- | :---: |
| ID_group | A | DEC_gal | L |
| RA_group | B | Velocity_gal | M |
| DEC_group | C | AbsMag_g_gal | N |
| Velocity_group | D | AbsMag_r_gal | O |
| Ngal_group | E | sersic_gal | P |
| Velocity_dispersion_group | F | fiber_col_gal | Q |
| AbsMag_g_group | G | completeness_gal | R |
| AbsMag_r_group | H | distance_edge_gal | S |
| cen sat_flag gal | I | AbsMag_g gal-AbsMag_r_gal | T |
| ID_gal | J | AbsMag_g_group-AbsMag_r_group | U |
| RA_gal | K |  |  |

Study protocol. We recruited 7 graduate students in astronomy from the Vanderbilt community, ranging in age from 23 to 28 years old, with 2 identifying their gender as female, 4 as male, and 1 as two-spirited. We ran participants in five sessions: two sessions (S1 and S4) each involved two participants conducting data exploration collaboratively, and the other three sessions (S2, S3, and S5) each involved a single participant. Note that as part of our pilot study design, we decided to include both individual and collaborative sessions to better inform our approaches for future studies.

Participants received gift cards as compensation for their time. Each session proceeded as follows. First, the partici-
pant(s) filled out a short demographic questionnaire. Then, we asked participants to sit down at a computer workstation and use Filtergraph to visualize and explore the galaxy data set. We provided a printout listing all dataset attributes and their semantic descriptions. A member of our study team sat with participants during the session, asking open-ended questions to better understand how the interaction was unfolding. At the end of the session, participants were asked to write down their own impressions about the study, and ideas about what software, tools, catalogs, or other data-related affordances would make their life as an astronomer easier.

Table 2: Session details from our pilot study. $M$ gives the number of major observations in each session, and $N$ gives the number of minor observations.

| Session | \# of participants | Duration (seconds) | M | N |
| :---: | :---: | :---: | :---: | :---: |
| S1 | 2 | 3378 | 12 | 67 |
| S2 | 1 | 3160 | 7 | 47 |
| S3 | 1 | 3289 | 20 | 73 |
| S4 | 2 | 4394 | 19 | 107 |
| S5 | 1 | 2921 | - | - |

Filtergraph settings. To constrain the visualizations used by participants, we asked participants not to change the Filtergraph setting that selects scatterplots as the visualization type. Within the scatterplot setting, Filtergraph offers many interactive options for changing how the dataset is visualized. The attributes assigned to X and Y axes can be changed, and a Z axis attribute can optionally be added. Attributes can also be assigned to the color dimension or used to select or filter out portions of the data. Anywhere individual attributes are used, arbitrary mathematical transformations or combinations of attributes can also be assigned. Additionally, there are options for changing the background plot color as well as the size and shape of data points. When the Z axis is in use, there are options to rotate the plot or change its scale.

All the sessions started from the same home screen, shown in Figure 1, with the X -axis set to attribute A , the Y -axis set to attribute B, and the color set to attribute C (see Table 1).

One concern we had was that participants might get bored or fatigued during the session and generate plots only to fill the time, and not through genuine interest and curiosity in exploring the dataset. Thus, the member of our team who sat with participants tried to be friendly and engaging, to help create a positive session environment. (Note that this member of our team comes from a computer science background, not astronomy, and so we do not believe this "social" aspect of the study sessions introduced significant biases in which parts of the dataset the participants would choose to focus on.)

In addition, after 45 minutes, the participants were informed that they could finish their current activity and end the session, if they wanted to, or they could continue to work for as long as they chose. As it turns out, all of our participants chose to stay past the 45 minute mark, with a minimum session duration of about 48 minutes and a maximum dura-
tion of about 73 minutes; see Table 2.
Analysis approach. We gathered data using a combination of note-taking by study personnel, paper forms, and interaction data collected on the computer workstation through screen recordings of each session. To analyze results, we defined the concepts of major observations and minor observations of the dataset, as illustrated in Figures 2 and 3.

Definition: A major observation is a grouping of contiguously viewed scatterplots within which the attributes assigned to X and Y axes remain constant. When a user changes the attribute assigned to either the X or Y axis (or both), a new major observation begins.

Definition: A minor observation is a grouping of contiguously viewed scatterplots that occurs during a major observation, within which one or more individual, highly related scatterplots are viewed. For example, looking at a 3D scatterplot but rotating the plot to see many different views would constitute a single minor observation of the data.

We used screen recordings to identify major and minor observations within each session. We did not count plots that were generated in the course of defining a single set of plotting parameters (i.e., because Filtergraph redraws plots nearly instantly, while the user might still be typing). Also, sometimes participants assigned the exact same parameters to the current plot, not actually changing the plot at all. We did not count these immediately repeated plots either. Finally, for continuously changing plots, which we saw especially when participants were smoothly rotating or moving 3D plots, we only counted the first and last views as separate plots within the same minor observation.

Exclusions. During session S2, an attribute was plotted by mistake; the participant had intended to plot a different attribute instead and only discovered their mistake after 12 minutes. We do not include these "mistake" plots in our current analysis, though certainly these kinds of mistakes will be considered in future work. In addition, session $S 5$ was qualitatively very different from Sessions S1 through S4. In session $S 5$, the participant appeared to pay very little attention to the semantics of the attributes that were being plotted (even when prompted to consider attribute meanings by our study team member). This participant declared to be selecting attributes randomly, and engaging in a primarily perceptual exploration of the dataset. While we find this pattern of interaction in session S5 extremely interesting, we felt that a different approach would be needed to analyze these data, and so we have left the analysis of session S5 for future work.

## Results and Discussion

Here, we discuss preliminary results from our pilot study. While there is certainly work to be done in analyzing the specific, astronomy-related meanings of participants' data visualization choices, including a detailed criterial analysis on their mental models (Mayr et al., 2016), that type of analysis falls outside the scope of the current paper and will be part of future work. For now, we focus on describing the


Figure 2: For session S1, plots showing the first minor observation for each of the 12 major observations.


Figure 3: For session S2, in chronological order, from left to right: plots showing the 28 minor observations within the fourth major observation. Note the "circular walk": the yellow, red, and blue boundaries indicate 3 different plots that were each observed twice. The six plots with a black background are unclear due to the small number of plotted points.


Figure 4: Number of major observations generated during each session for different combinations of attributes.
general behavioral patterns of data visualization that we observed, including the temporal dynamics of participant interactions with the Filtergraph tool. Table 2 gives details of the five sessions that we ran for this study.

Distribution of major observations by attribute. Consider the 17 data attributes contained in the galaxy dataset (see Table 1). Note that two of the original 19 attributes are ID numbers, for galaxies and groups of galaxies, and so do not capture "meaningful" in the same sense as the other 17. The number of possible major observations that could be made from this data, i.e., possible assignments of attributes to X and $Y$ axes, is $(17$ choose 2$)=272$ possibilities. If we allow mathematical transformations or combinations of attributes, then the number of possible major observations is infinite.

Across all sessions, participants viewed a total of only 38 distinct major observations, as shown in Figure 4. This figure also shows which attributes participants assigned to either the

X or Y axes. Interestingly, only 4 major observations were shared by two or more sessions; this shows the high variability in data exploration paths taken by different individuals.

Attribute I was the only one never assigned to either the X or Y axes. It turns out that this attribute is Boolean; it indicates whether the galaxy is the brightest in its group, and so it makes sense to not be chosen for one of the primary axes. Attributes P (a measure of galaxy morphology) and Q (was there a problem while measuring the galaxy) are the other two categorical variables in the dataset; interestingly, these were assigned to axes at various times during one of the sessions.

The BC combination was the only major observation shared across all four sessions. B gives the longitude of the galaxy group center, and C gives the latitude. So, plotting the BC combination produces what is essentially a spatial "map" of the galaxies represented in the dataset.

Temporal dynamics of data exploration. We are partic-


Figure 5: Distribution of major and minor observations over time, for sessions S1 (top) through S4 (bottom). Each point indicates a group of highly related individual plots (or, for example, a single continuous rotation of the same 3D plot). Major observations repeated within or across sessions are marked by colored labels.
ularly interested in understanding the temporal dynamics of open-ended data exploration. How frequently do our participants switch from one major or minor observation to the next?

Figure 5 illustrates the temporal distributions of major and minor observations for all four sessions. By looking at the blue dots in this figure, we can see the most active periods, the ones in which many plots were quickly generated. We can also see periods in which a single minor observation was studied at some length. (Note also that the large gap early in session S2 is due to data we omitted, as described earlier.) The first point of interest is that participants generally chose to "flip through" the dataset at a fairly brisk pace. Part of this could, of course, be due to the participants knowing they had only one hour to complete the study, but we expect the same to hold in more naturalistic settings as well.

The average duration of minor observations was about 61 seconds, though this distribution has a long tail that falls off fairly consistently and extends to the longest minor observations at around 554 seconds. For major observations, the average duration was about 229 seconds, with a maximum of about 1046 seconds. The durations of major observations seem to show a bimodal tendency, with many major observations falling under the 150 second mark, but another large grouping lying in the 150-400 second range.

Interestingly, participants often returned to the same major observation within the same session. This pattern occurs
not at all in session S 2 , occasionally in session S 1 , and quite a lot in sessions S3 and S4. In sessions S2 and S4, we saw similar "circular walk" patterns within some of the major observations; the participant often returned to the same minor observation that they had started with, before moving on to the next minor or major observation. Figure 3 depicts an example of a "circular walk" within session S2.

Within each major observation, it seems as though participants are directing a type of "movement" from the first to the last plot (depicted via minor observations); a story is being told through the movement of data points across the plots. This notion of movement/story strongly brings to mind the idea of flipbooks. For instance, Figure 3 shows a story regarding the relationship between the attributes $F$ and $E$.

Other qualitative observations. To conclude our presentation of preliminary results, we present a few high-level observations about patterns of data exploration in our study.

Starting point: Participants seemed to choose a starting point to anchor themselves in relation to the dataset. Sessions S1 and S2 began by plotting attributes that involve the spatial position of galaxies, perhaps to let participants establish a mental map of the spatial layout of the galaxy dataset. Sessions S3 and S4 began by looking at attributes like the number of galaxies in each group and the velocity with which each group is moving away from us, perhaps establishing an egocentric reference frame of "us versus the galaxies."

Mouse gestures: Participants frequently used the mouse as a communicative or attention-focusing tool, i.e., to gesture at the visualizations. While some of these were directed at our study team member who was observing the session, many of these mouse gestures also occurred when participants were interacting with Filtergraph and thinking about what to do next. Sessions S1 and S2 exhibited many more mouse gestures than did sessions S3 and S4.

Paper aids: Participants in sessions S1 and S2 relied heavily on the paper printout of attribute details throughout the sessions. Participants in S3 and S4, on the other hand, used the printout at the start, but, during the interaction itself, relied more on the list of attributes provided by Filtergraph.

Collaboration/leadership: In both of the sessions with two participants, we observed that one of the participants seemed to lead the exploratory line of thought. But note that leading does not necessarily means commanding the mouse and directly interacting with the workstation. In one session, the "thought" leader was also the one interacting with Filtergraph, but in the other session, the "thought" leader was not the primary tool interaction person.

Collaboration/corrections: We observed that, in the sessions with two participants, one helped the other to quickly correct mistakenly plotted attributes. In contrast, during session S 2 (which involved a single participant), an attribute was plotted by mistake and not discovered for 12 minutes.

## Conclusion and Future Directions

Our findings highlight a few interesting properties of how domain experts explore an unfamiliar dataset, particularly in terms of temporal and dimensional patterns. The next challenge is to understand how, as people follow exploratory paths through a dataset, they build meaningful cognitive representations of what they see, and how they are able to identify encounters with unexpected, significant data patterns.

We anticipate that temporal patterns are especially important from a cognitive perspective because they describe not just moment-to-moment attentional switches but also serve as a way of marking successive stages at which a person's mental model of a dataset is likely changing. Five minutes spent looking at a single plot, versus five minutes spent "flipping" through multiple plots, are both likely to be equally important modes of exploration. The key is figuring out the cognitive purpose served by each.

We are also intrigued by the frequency of "circular walks" in the exploratory paths taken by our participants. Viewing the same plot twice serves no obvious purpose from a purely statistical or data mining perspective. However, in humans, we predict that these "circular walks" actually serve key roles related to memory, attention, salience, etc.

## Acknowledgments

We would like to thank Dr. Andreas Berlind who generously shared the galaxy dataset we used in our study. This research makes use of Filtergraph, an on-line data visualiza-
tion tool developed at Vanderbilt University through the Vanderbilt Initiative in Data-intensive Astrophysics (VIDA).

## References

Bastien, F., Stassun, K. G., Basri, G., \& Pepper, J. (2013). An observational correlation between stellar brightness variations and surface gravity. Nature, 500(7463), 427-430.
Berlind, A., Frieman, J., Weinberg, D., Blanton, M., Warren, M., Abazajian, K., ... others (2006). Percolation galaxy groups and clusters in the sdss redshift survey: identification, catalogs, and the multiplicity function. ApJ Supplement Series, 167(1), 1.
Brehmer, M., \& Munzner, T. (2013). A multi-level typology of abstract visualization tasks. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2376-2385.
Burger, D., Stassun, K., Pepper, J., Siverd, R., Paegert, M., De Lee, N., \& Robinson, W. (2013). Filtergraph: An interactive web application for visualization of astronomy datasets. Astronomy and Computing, 2, 40-45.
ElTayeby, O., \& Dou, W. (2016). A survey on interaction log analysis for evaluating exploratory visualizations. In Proc. of the beyond time and errors on novel eval methods for visualization (pp. 62-69).
Healey, C., \& Enns, J. (2012). Attention and visual memory in visualization and computer graphics. IEEE trans. Visualization \& Comp. Graphics, 18(7), 1170-1188.
Honavar, V., Hill, M., \& Yelick, K. (2016). Accelerating science: A computing research agenda. arXiv preprint arXiv:1604.02006.
Mayr, E., Schreder, G., Smuc, M., \& Windhager, F. (2016). Looking at the representations in our mind: Measuring mental models of information visualizations. In Proc. of the beyond time and errors on novel eval methods for visualization (pp. 96-103).
Pirolli, P., \& Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In Proceedings of international conference on intelligence analysis (Vol. 5, pp. 2-4).
Russell, H. N. (1914, May). Relations Between the Spectra and Other Characteristics of the Stars. Popular Astronomy, 22, 275-294.
Sanderson, P. M., \& Fisher, C. (1994). Exploratory sequential data analysis: Foundations. Human-Computer Interaction, 9(3-4), 251-317.
Saraiya, P., North, C., \& Duca, K. (2005). An insight-based methodology for evaluating bioinformatics visualizations. IEEE trans. visualization \& comp. graphics, 11, 443-456.
Spence, I., \& Garrison, R. (1993). A remarkable scatterplot. The AM STAT, 47(1), 12-19.
Tversky, B. (2003). Structures of mental spaces how people think about space. $E A B, 35(1), 66-80$.
Yi, J., ah Kang, Y., \& Stasko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. IEEE trans. visualization \& comp. graphics, 13(6), 1224-1231.

# The cultural evolution of complex linguistic constructions in artificial sign languages 

Yasamin Motamedi ${ }^{1,2}$ (Yasamin.Motamedi@ed.ac.uk) Marieke Schouwstra ${ }^{2}$ (Marieke.Schouwstra@ed.ac.uk) Jennifer Culbertson ${ }^{2}$ (Jennifer.Culbertson@ed.ac.uk) Kenny Smith ${ }^{2}$ (Kenny.Smith@ed.ac.uk) Simon Kirby ${ }^{2}$ (simon@ling.ed.ac.uk)<br>${ }^{1}$ Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands<br>${ }^{2}$ Centre for Language Evolution, University of Edinburgh, UK


#### Abstract

Though most documented sign languages make use of space to denote relationships between predicate arguments, studies of emerging sign languages suggest that spatial reference does not emerge fully-formed but takes time to develop. We present an artificial sign language learning experiment that expands the cultural evolutionary framework to investigate complex linguistic constructions. Our results demonstrate the gradual emergence of consistent devices to distinguish between sentence arguments, some of which rely on iconic spatial contrasts. These findings mirror data from emerging sign languages and point to the cultural mechanisms that facilitate the evolution of complex linguistic structures.


Keywords: language; cultural evolution; learning; communication; sign language; gesture

## Introduction

Sign languages, as manual-visual linguistic systems, are able to represent relationships between a predicate and its arguments using the space around the signer. Though there are differences in exactly how spatial reference is utilised, spatial modulation is attested across most sign languages (Mathur \& Rathmann, 2012). Signers use indexed locations in space to refer to particular arguments, such that a deictic point to an arbitrary location can pronominally refer to the subject or object of a sentence. This mapping can iconically represent a real-word spatial relationship, such that references to arguments in a sentence reflect their orientation in relation to each other in the real world, but that is often not the case and the relationship between referenced locations is primarily grammatical. It has been suggested that the iconic potential of spatial mappings makes the use of space almost inevitable in sign languages (Aronoff, Meir, \& Sandler, 2005), and the beginnings of spatial reference systems have been attested in several emerging sign languages (Senghas, Coppola, Newport, \& Supalla, 1997; Padden, Meir, Aronoff, \& Sandler, 2010).

However, studies on emerging sign languages also suggest that systematic spatial reference is not a property that emerges immediately in a linguistic system, but takes time to evolve over generations of a language (Padden et al., 2010). Spatial agreement systems are used to represent complex relationships between a predicate and its arguments, and as such pose problems for learners. Furthermore, the ability to represent animate agents using the signer's own body may interfere with the development of abstracted spatial reference. Re-
ferring to multiple participants in space requires abstraction away from the signer's body and may therefore take longer to evolve (Padden et al., 2010; Meir, Padden, Aronoff, \& Sandler, 2007). Finally, the use of space is not the only grammatical tool used to denote who does what to whom, and in fact its use is often restricted to particular classes of verbs.

Using a novel experimental method, we ask how systematic spatial reference emerges in a linguistic system, and how the iconic affordances of the manual modality affect the evolution of such a system. Though emerging sign languages provide valuable natural evidence of the early evolution of languages, the experimental research we present here is able to test the particular factors that influence language with a greater degree of control and precision.

Previous experimental research has experimentally demonstrated the importance of cultural evolutionary processes in the emergence of linguistic structure, namely interaction between language users and transmission to new learners of a language (Kirby, Tamariz, Cornish, \& Smith, 2015; Kirby, Cornish, \& Smith, 2008). The gradual development of spatial agreement systems in naturally emerging sign languages similarly points to the impact of interaction and transmission on the evolution of spatial reference. Therefore, we propose a cultural evolutionary stance on the emergence of spatial agreement, and explore the impact of cultural evolutionary processes in a laboratory study. We present a study that investigates the effect of interaction and transmission on the emergence of signals that participants produce to signal complex events with multiple animate arguments. We place silent gesture (where hearing participants with no knowledge of sign language communicate using gesture; GoldinMeadow, So, Ozyürek, and Mylander (2008); Schouwstra and de Swart (2014); So, Coppola, Licciardello, and GoldinMeadow (2005)) within a cultural evolutionary framework that implements interaction between participants, and transmission with an iterated learning model. Pairs of participants communicate about a set of events using only gesture and the gestures they produce are used to train a new pair of participants, who then use what they have learnt to communicate with each other. We provide an experimental account of the evolution of linguistic structure that is informed by data on natural sign languages. The present study offers a more pre-
cise understanding of the cultural evolutionary mechanisms that facilitate the emergence of linguistic structure, and serves to elucidate how modality-specific factors affect the emergence of systematic structure in language.

## Methods

Participants took part in an artificial sign language learning task where they learnt gestures produced by a previous participant in a training stage, before communicating with a partner during testing, using only gesture.

## Participants

50 participants ( 15 male, 35 female, median age $=22$ ) were recruited from the University of Edinburgh's Careers Hub website, and were compensated $£ 7$ for participation in the experiment, which took between 30 and 50 minutes to complete. Participants were self-reported right-handed native English speakers with no knowledge of any sign language.

## Materials

Participants were asked to communicate events, presented orthographically as pairs of sentences in English. Sentences involved two arguments, Hannah and Sarah, who could either be the agent of the sentence, be the goal or recipient of the sentence, or who might not be present in the target sentence at all (see figure 1a for examples). Sentences were presented in pairs at each trial, and pairs of sentences were grouped into blocks of four, where each block comprised a sentence pair of one of four verb types: plain spatial verbs (e.g. to cycle), spatial locative verbs (e.g. to cycle to), physical transfer verbs (e.g. to kick a ball to), and non-physical transfer verbs (e.g. to help; all verbs are shown in table 1 in the appendix). There were four blocks of four pairs in total, giving a total of 16 sentence pairs, and 32 target sentences (figure 1 gives an example of sentences as they would be shown in pairs and blocks). Two blocks consisted of different-agent pairs, such that if Hannah was the agent in the first sentence of a target pair, Sarah would be the agent in the second sentence of that pair, and vice versa. The two remaining blocks consisted of same-agent sentence pairs, such that either Hannah or Sarah was the agent in both sentences in a target pair (e.g. Hannah is walking to Sarah/Hannah is swimming to Sarah). Order of presentation for target sentences was randomised within sentence pairs and within blocks.

Participants were placed in individual experiment booths; target sentences were presented on-screen using Psychopy (Peirce, 2007) and video recording and streaming between networked computers was enabled via custom software, Videobox (Kirby, 2016).

## Procedure

Participants were organised into 5 transmission chains of 5 generations, where each generation was made up of a pair of participants, who communicated with each other during testing (see figure 1b). Participants in generations 2-5 took part


Figure 1: Examples of sentence pairs and blocks (a) and diagram illustrating diffusion chain structure (b). Pairs of participants interact at each generation (dotted lines); gestures produced by 1 participant in a pair are transmitted as training to the next generation (solid lines with arrows). A single chain is made up of 5 generations and there were 5 chains in total.
first in a training stage, followed by a testing stage. Participants in the first generation of each chain only took part in the testing stage, and were therefore required to innovate gestures at the beginning of each chain.

Training stage Participants in the training stage were trained on gestures produced by a participant in the previous generation of the chain. The training model was selected at random from one of the two participants in the previous generation, and the full set of gestures produced by that participant were used as training data. At each trial, the participant was shown a video of their model gesturing, and was asked to select the pair of target sentences they were trying to communicate. Whilst the video was playing, participants were shown an array of sentence pairs onscreen, from which to make their choice. The array of sentence pairs comprised the target pair, and three distractors. The distractors differed from the target pair on either the agent configuration, the verb, or both. For instance, a target pair that had Hannah as agent in the first sentence and Sarah as agent in the second, would have a distractor pair with the same verb construction, but as a same-agent pair, with Hannah as the agent of both sentences. Another distractor would keep the agent configuration, but would replace the verbs with other verbs from the same category, and a final distractor would present both different verbs and a different agent configuration. Building the arrays in this way required participants to specify who does what to whom, without necessarily having to describe Hannah and Sarah (i.e., the role is important, not the individual). The position of each pair on the screen was randomised at each trial, and participants could make their guess by pressing the $1,2,3$ or 4 key, depending on the position on the screen.

Participants were given feedback about whether their guess was correct or incorrect, and shown the correct answer. Participants completed 16 training trials, one for each sentence pair. Participants completed the training stage individually, without any interaction with their partner.

Testing stage Participants in the testing stage communicated with a partner, taking turns to be director (the person gesturing) and matcher (the person interpreting). As director, a participant was shown a sentence pair on screen and asked to communicate it to their partner. A three-second countdown prepared them for streaming and recording, when they would produce their gesture to camera. The directing participant saw themselves onscreen whilst they gestured, with their image mirrored. An unmirrored image was streamed to their partner in another booth. Either participant could interrupt the video stream when they had finished their gesture or were ready to make a guess. The matcher's task was very similar to the training task; they watched their partner gesture on screen and had to select the correct target pair from an array of four. Once the matcher had made their guess, both participants were given full feedback, about both the correct answer and the matcher's selection. Participants switched roles after every block (every four trials) and directed and matched for the full set of 16 sentence pairs, giving a total of 32 testing trials. Order of the blocks and target sentence pairs was randomised for each participant.

## Results

## Gesture coding

Gestures were coded for the presence of an agent gesture, a goal (or recipient) gesture and a verb gesture. For each argument, the type of gesture was coded, as well as the location and path of the gesture. The goal or recipient of the target sentence was frequently omitted from gestures; as such, we focus on differentiation between agents across sentence pairs.

## Differentiation strategies

We identified three main strategies participants employed to differentiate between agents in target sentences: the lexical strategy, the body strategy and the indexing strategy (exemplified in figure 2). All strategies make use of iconic representation, and both body and indexing strategies make use of the space around the gesturer to disambiguate target sentences.

Lexical strategies Two out of five chains differentiate sentence arguments based on the gesture type, using a 1- and 2handshape to denote Hannah and Sarah in target sentences. Though this begins as a way to simply distinguish the first sentence in a target pair from the second, these handshapes come to represent individual arguments in later generations.

Body strategies A further two chains rely on differences in body orientation to signal differences between agents in target sentence pairs. Participants use an iconic spatial strategy to represent sentence arguments. For instance, in figure 2b, the participant orients their body to the right to represent Hannah,
and to the left to represent Sarah.
Indexing strategy Finally, one chain developed a strategy in which locations in the space around the gesturer were indexed to refer to different sentence arguments. In the example shown in figure 2 c , the participant points to her left to signal the agent, Hannah. In addition, her verb gesture moves between the indexed locations for agent and recipient.

Participants show a difference based on sentence type (whether same- or different-agent), producing different gestures to represent different agents, and showing divergence between sentence context over generations of the experiment. Figure 3 shows the proportion of gesture sequences that are different across two sentences in a target pair. Rows show the proportion of variation across different aspects: agent gesture type, location of the agent gesture, location of the verb gesture and path of the verb gesture. Variation across these aspects corresponds to different strategies. Chains 1 and 5 primarily vary gestures on agent type, using the lexical strategy (e.g. figure 2a). Chains 2 and 4 vary gestures based on agent location, as well as verb path and location, as they are implementing a body strategy, where agent and verb are simultaneously inferred through the participant's use of their own body (e.g. figure 2b). Finally, chain 3 shows the primary difference on the location of agent gestures, using an indexing strategy to place sentence arguments in difference locations (figure 2c).

We analysed the changes in agent distinctions over generations in the experiment, collapsing the measures shown in figure 3 across features to simply account for whether or not participants create a distinction between agents in the two sentences of a target sentence pair, investigating whether participants structure signals in similar ways across strategies. A binomial mixed effects model analysed the fixed effects of generation and sentence type on the proportion of different agent gestures, as well as their interaction. Chain, target and participant were included as random effects with random intercepts, and random slopes of generation and verb type were implemented for chain and target, respectively. The random effects structure for participant was nested within chains. Comparison of the model revealed a significantly better fit over a reduced model without the interaction term $\left(\chi^{2}=11.51, p<\right.$ 0.001 ). The model indicated a significant effect of the different-agent sentence type in comparison to same-agent sentence pairs ( $\beta=2.73, S E=0.46, z=5.91, p<0.001$ ), as well as a significant interaction between generation and sentence type ( $\beta=0.69, S E=0.22, z=3.12, p=0.002$ ), though no significant effect of generation ( $\beta=0.04, S E=0.12, z=$ $0.29, p=0.77$ ). Participants were more likely to produce gestures that differentiate between agents in different-agent contexts compared to same-agent contexts, and this contrast strengthens over generations in chains of participants.

We also analysed the effect of verb type on the distinctions participants made. Spatial reference in signed languages is not used across all verbs, but usually affects specific sets of verbs. As such, it is possible that participants in the experi-
ment create distinctions based on semantic properties of the verbs they encounter. We ran a binomial mixed effects analysis that examined the effect of verb type on differences between agent gestures. The random effects structure described above was also implemented here. The model showed no improvement over the null model ( $\chi^{2}=1.82, p=0.61$ ), indicating that participants do not condition differences between sentence pairs based on verb type.

## Discussion

Our results demonstrate the evolution of systematic agent distinctions, which emerge over generations of interacting participants. In addition, participants frequently use iconic spatial mappings to create those distinctions, which become increasingly contrastive over generations in the transmission chains. These findings are consistent with data from naturally emerging sign languages that suggest the gradual emergence of systematic spatial mappings.

## Differentiation strategies reflect sign language structure

The three main strategies that participants employ in the present experiment all find comparable forms in natural sign languages: specifically, as lexical signs, role-shift, and spatial agreement. The latter two strategies make use of the space around the gesturer to create distinctions between agents in the target sentences. The body strategy involves movement of the participant's body to represent sentence arguments, and exhibits similarities to role-shift found in natural sign languages, and can be used in natural languages to distinguish between sentence arguments (Padden, 1986; Cormier, Fenlon, \& Schembri, 2015).

The indexing strategy, however, is the strategy that most closely resembles sign language verb agreement, such that locations in the space around the signer are indexed to refer to different sentence arguments (Liddell, 2003; Padden, 1986; Lillo-Martin \& Meier, 2011). The use of indexing in chain 3 begins on the axis perpendicular to the participant's body; for example, the participant points to themselves to denote Hannah, and points directly away from their body to denote Sarah. Over generations of the chain, the use of indices is abstracted away from the body and indexes are contrasted parallel to the gesturer's body, such that Hannah might be indexed to the left of the participant, and Sarah might be indexed to the right. This change mirrors development in two young sign languages, Al-Sayyid Bedouin Sign Language (ABSL) and Israeli Sign Language (ISL). In both ABSL and ISL, early generations of signers demonstrate greater preferences for spatial contrasts that centre around the signer's body, on the perpendicular axis (Padden et al., 2010). However, later generations show an increase in the use of the parallel axis, as demonstrated in the present study. Furthermore, participants in our study made no distinction between verb types, consistent with findings from ABSL that showed spatial mappings were not restricted to any class of verbs in early generations
of the language (Padden et al., 2010). Our findings, consistent with natural language data, indicate that systematic spatial reference does not emerge wholesale, despite the iconic affordances of the modality, but takes time to emerge.

## The evolution of complex constructions

The gestures participants produce support a gradual evolution of systematic linguistic structure, including the use of space; participants indicate a difference between sentence arguments from the first generation, but the mechanisms used to create these distinctions are neither consistent nor systematic early on. Instead, participants converge on particular strategies to make distinctions over generations, and participants show increasing divergence between the same- and different-agent sentence contexts. Participants' reliance on iconic, spatial gestures supports silent gesture research showing that hearing participants can use deictic indexing to track referents (So et al., 2005), though the present results further demonstrate how such a system evolves through use. The increased consistency of these systems supports previous iterated learning experiments (Reali \& Griffiths, 2009; Smith \& Wonnacott, 2010) that suggest learning leads to regularisation. Participants also demonstrate the negotiation of a system that is both expressive and learnable; they minimise the number of strategies used to convey differences in target sentences, settling on one strategy to use in the majority of trials, sufficient to express the differences in the meanings they are trying to convey. Consistent with previous experimental research, the systems participants produce maximise simplicity and informativeness (Kirby et al., 2015; Regier, Kemp, \& Kay, 2015). Participants systematically signal differences between agents in target sentences, producing gestures that allow successful communication within their pair.

## The effects of iconicity on emergent structure

All participants rely on iconic representations to communicate target sentences to their partners. In particular, the use of the gesturer's body and the use of space around the gesturer allow for iconic representations of animate agents. The privileged status of the body is attested in natural sign languages (Meir et al., 2007), and use of the body is attested in the development of spatial grammatical devices (Meir et al., 2007; Padden et al., 2010; Kocab, Pyers, \& Senghas, 2014). Further, consistent with research on emerging sign languages and experimental research (Padden et al., 2010; Theisen, Oberlander, \& Kirby, 2010), our results suggest a movement away from iconic reliance on the body (e.g. the axis change in chain 3 ) as gestures become more consistent and regular across the system.

## Conclusion

We have demonstrated the emergence of systematic signals to communicate complex events, through the cultural evolution of communicative signals, via interaction between users and transmission to new users. Participants make use of different representation tools, all of which have analogues in natural


Figure 2: Examples of differentiation strategies used in the experiment. (a) shows an example of the lexical strategy, in which the participant uses 1-and 2-handshapes to denote arguments. (b) shows a participant using body orientation to denote sentence arguments. (c) illustrates the indexing strategy, in which the participant indexes locations to refer to sentence arguments.

Chain 1


Chain 2


Chain 3


Chain 4


$\qquad$ same-agent

Figure 3: Proportion of gestures that differentiate agents in target sentences, based on which aspect of the gesture is varied (agent type, agent location, verb location). Coloured lines show proportions for different-agent (blue circles) and same-agent contexts (green triangles). Columns show the proportions for each chain, at each generation. All chains show differences based on context, though they make distinctions in different ways
sign languages. Our findings support data concerning the evolution of spatial reference in emerging sign languages, which suggest that the phenomenon takes time to emerge and systematise. Using an experimental method, we have been able to observe this gradual evolution in a controlled environment, to test more precisely the mechanisms that drive the emergence of spatial reference. Furthermore, these results shed light on modality-specific effects of iconicity, and their influence on the structure of emerging communication systems.

## Acknowledgments

Yasamin Motamedi was funded by the Carnegie Trust for the Universities of Scotland.

## References

Aronoff, M., Meir, I., \& Sandler, W. (2005). The Paradox of Sign Language Morphology. Language, 81(2), 301-344.
Cormier, K., Fenlon, J., \& Schembri, A. (2015). Indicating verbs in British Sign Language favour motivated use of space. Open Linguistics, 1, 684-707.
Goldin-Meadow, S., So, W. C., Ozyürek, A., \& Mylander, C. (2008). The natural order of events: how speakers of different languages represent events nonverbally. PNAS, 105(27), 9163-8.
Kirby, S. (2016). VideoBox. Edinburgh: University of Edinburgh. Retrieved from http://edin.ac/2haREUz
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. PNAS, 105(31), 10681-6.
Kirby, S., Tamariz, M., Cornish, H., \& Smith, K. (2015). Compression and Communication in the Cultural Evolution of Linguistic Structure linguistic structure. Cognition, 141, 87-102.
Kocab, A., Pyers, J., \& Senghas, A. (2014). Referential Shift In Nicaraguan Sign Language: A Transition From Lexical To Spatial Devices. Frontiers in Psychology, 5.
Liddell, S. K. (2003). Grammar, gesture and meaning in American Sign Language. Cambridge: Cambridge University Press.
Lillo-Martin, D., \& Meier, R. P. (2011). On the linguistic status of 'agreement' in sign languages. Theoretical Linguistics, 37(3-4), 95-141.
Mathur, G., \& Rathmann, C. (2012). Verb agreement. In R. Pfau, M. Steinbach, \& B. Woll (Eds.), Sign language: An international handbook (pp. 136-157). Berlin: De Gruyter Mouton.
Meir, I., Padden, C., Aronoff, M., \& Sandler, W. (2007). Body as subject. Journal of Linguistics, 43(03), 531-563.
Padden, C. (1986). Verbs and role-shifting in American Sign Language. In C. Padden (Ed.), Proceedings of the fourth national symposium on sign language research and teaching (pp. 44-57). Washington DC: The National Association for the Deaf.
Padden, C., Meir, I., Aronoff, M., \& Sandler, W. (2010). The grammar of space in two new sign languages. Sign Lan-
guages: A Cambridge Language Survey. Cambridge University Press, Cambridge, UK, 570-592.
Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. Journal of neuroscience methods, 162(1-2), 8-13. Reali, F., \& Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. Cognition, 111(3), 317328.

Regier, T., Kemp, C., \& Kay, P. (2015). Word meanings across languages support efficient communication. In B. McWhinney \& W. O'Grady (Eds.), The handbook of language emergence (pp. 237-263). Hoboken,NJ: WileyBlackwell.
Schouwstra, M., \& de Swart, H. (2014). The semantic origins of word order. Cognition, 131(3), 431-6.
Senghas, A., Coppola, M., Newport, E. L., \& Supalla, T. (1997). Argument structure in Nicaraguan Sign Language: the emergence of grammatical devices. In E. Hughes, M. Hughes, \& A. Greenhill (Eds.), Proceedings of the boston university conference on language development (pp. 550-561). Boston: Cascadilla Press.
Smith, K., \& Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. Cognition, 116(3), 444-9.
So, W. C., Coppola, M., Licciardello, V., \& Goldin-Meadow, S. (2005). The seeds of spatial grammar in the manual modality. Cognitive science, 29(6), 1029-43.
Theisen, C. A., Oberlander, J., \& Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. Interaction Studies, 11(1), 14-32.

## Appendix

Table 1 shows the verbs in each category used in target sentences in the experiment.

| Verb type | Verbs |
| :---: | :---: |
| plain spatial verbs | to cycle <br> to run <br> to swim <br> to walk |
| spatial locative verbs | to walk to P <br> to run to P <br> to swim to P <br> to walk to P |
| physical transfer verbs | to kick a ball to R <br> to give a book to R <br> to send a letter to R <br> to throw a hat to R |
| non-physical transfer verbs | to help R <br> to phone R <br> to praise R <br> to scold R |

Table 1: Verbs used in target sentences

# Do Accurate Metacognitive Judgments Predict Successful Multimedia Learning? 

Nicholas V. Mudrick (nvmudric@ncsu.edu)<br>North Carolina State University, Department of Psychology, 2310 Stinson Drive, 640 Poe Hall<br>Raleigh, NC 27695 USA<br>Michelle Taub (mtaub@ncsu.edu)<br>North Carolina State University, Department of Psychology, 2310 Stinson Drive, 640 Poe Hall Raleigh, NC 27695 USA<br>Roger Azevedo (razeved@ncsu.edu)<br>North Carolina State University, Department of Psychology, 2310 Stinson Drive, 640 Poe Hall Raleigh, NC 27695 USA


#### Abstract

Successful performance during multimedia learning requires accurate metacognitive judgments. However, little research has investigated the influence of accurate metacognitive judgments for different representations of information (e.g., text and diagram) on performance during multimedia learning. As such, we investigated if participants' metacognitive judgments for text and diagrams (i.e., content evaluations; CEs) were significantly related to increased performance and higher confidence during multimedia learning. Metacognitive judgments and performance measures were collected from 48 undergraduate participants during 18 randomized trials. Results using multilevel modeling indicated that participants' CEs for text-based content were significantly predictive of performance. Results also showed that accurate CEs for diagrams interacted with accurate multiple-choice responses to predict higher retrospective confidence judgments (i.e., higher confidence). Identifying metacognitive judgments predictive of increased performance during multimedia learning has important theoretical, conceptual, and analytical implications.


Keywords: multimedia learning; metacognition; metacognitive judgments; multilevel modeling; performance; science learning

Research indicates learning with multimedia materials (e.g., text and diagram) is more effective than learning through text alone (Butcher, 2014; Mayer, 2014). Successful multimedia learning entails individuals actively and accurately selecting, organizing, and integrating text- and image-based information into a coherent mental model (Mayer, 2014). However, research suggests learners do not always engage in accurate and effective metacognitive monitoring and regulation during learning with multimedia (Azevedo, 2014). Specifically, research has indicated participants often exhibit overconfidence when monitoring their own understanding during multimedia learning (Serra \& Dunlosky, 2010). The multimedia heuristic suggests learners’ own judgments of learning (JOLs; i.e., how well they will remember the information) are largely inflated when compared to their actual performance because individuals perceive multimedia content as being easier to learn than with text alone (Serra \& Dunslosky, 2010).

Research on metacognitive monitoring during multimedia learning has traditionally employed modified metacomprehension paradigms (based on Nelson \& Narens' metamemory framework, 1990), during which participants are asked to make metacognitive judgments (e.g., ease-oflearning [EOL], immediate and delayed JOLs, retrospective confidence judgments [RCJs]) during various stages of multimedia learning (e.g., Burkett \& Azevedo, 2012; Eitel, 2016; Pilegard \& Mayer, 2015). The major assumption of this research is that the timing of metacognitive judgments made during multimedia learning (before learning, during learning, and after learning) will vary in accuracy, selection of cognitive strategies, and subsequent performance, dependent on the specific experimental manipulation (e.g., delayed JOLs are more predictive of performance than EOLs; Burkett \& Azevedo, 2012; Nelson \& Dunlosky, 1991). As this research has identified that most metacognitive judgments for multimedia are often inaccurate (e.g., Serra \& Dunlosky, 2010), much of the literature has focused on ways to improve metacognitive judgments. For example, some research has focused on manipulating the framing of metacognitive judgment prompts to improve judgment accuracy (e.g., Pilegard \& Mayer, 2015; Vössing, Stamov-Roßnagel, \& Heinitz, 2016). Pilegard and Mayer (2015) compared JOLs (i.e., how well do you remember the content) to judgments of understanding (JOUs; i.e., how well do you understand the information) and found JOUs were more predictive of retention and transfer compared to JOLs. These findings suggest that framing metacognitive judgment prompts (e.g., from JOLs to JOUs) significantly impacts the metacognitive processes employed during multimedia learning, potentially indicating there may be other metacognitive judgments participants use that can successfully influence performance. In support of this assertion, research on hypermedia and self-regulated learning (SRL) suggests several other metacognitive processes may be more predictive of multimedia learning outcomes (Greene \& Azevedo, 2009).

Azevedo, Greene, and Moos (2007) developed a classification scheme by which 35 micro-level metacognitive judgments can be evident during successful SRL with hypermedia-based learning environments. One
example of these judgments is a content evaluation (CE). CEs are judgments learners make to assess the relevancy of the content (e.g., multimedia) they are viewing to their current goal (e.g., answering a science question about a human body system; Greene \& Azevedo, 2009). CEs are key metacognitive judgments for successful multimedia learning, such that accurate CEs can direct participants to study more efficiently. For example, if the goal is to answer a science question about the human body system and participants evaluate the text but not the diagram they are viewing to be relevant to their goal, they should invest more effort and time to study the text (as opposed to the diagram), employ the appropriate cognitive strategy (e.g., make an inference), and therefore be more likely to answer the question correctly.

Other research on metacognitive judgments during hypermedia learning has identified the predictive validity of traditional metacomprehension judgments like RCJs. For example, Mengelkamp and Bannert (2010) investigated the stability of participants' RCJs as they learned about operant conditioning with a hypermedia environment. Results indicated that the absolute accuracy (i.e., difference between judgments and performance) was stable throughout the learning session, and relative accuracy (correlation between judgments and performance) was significantly predictive of hypermedia learning outcomes.

Theories of multimedia learning suggest participants cognitively process information from text and diagrams separately and in different ways (Burkett \& Azevedo, 2012; Mayer, 2014). Additionally, researchers have outlined the multimedia effect to indicate that students demonstrate longer periods of recall and higher levels of retention when learning with text and images as opposed to learning only with text (Butcher, 2014). However, evidence suggests learners do not always engage in effective selection, organization, and integration of multiple representations and instead exhibit a bias toward text-based (as opposed to diagram-based) information during multimedia learning (Hegarty \& Just, 1993). Since cognitive processes are different for text and diagrams, it should be expected that metacognitive judgments will also be different.

Accurate metacognitive monitoring and regulation are required during multimedia learning to achieve an increase in learning outcomes (Azevedo, 2014). However, little research has examined the specific processes underlying successful metacognitive monitoring and regulation during multimedia learning. Specifically, few metacognitive judgments have been found to be predictive of successful multimedia learning outcomes (e.g., overconfident JOLs; Serra \& Dunlosky, 2010). We argue that examining other metacognitive judgments (CEs, RCJs) can inform us of monitoring processes that are more indicative of successful learning and performance. In contrast to the limited research on metacognitive judgments during multimedia learning, we focus on different metacognitive judgments and identify how they can contribute to superior learning outcomes.

In this study, we examined participants' text CEs, diagram CEs, multiple-choice responses, and RCJs during multimedia learning to answer the following three questions: (1) Are accurate text and diagram CEs associated with an increase in the likelihood of an accurate multiplechoice response? (2) Is there a significant relationship between text and diagram CE accuracy and RCJs? (3) Is there a significant relationship between the interactions of text and diagram CEs and multiple-choice responses on RCJs?

To address our research questions, we proposed the following hypotheses:

H1: Accurate text and diagram CEs will be significantly associated with an increase in the likelihood of an accurate multiple-choice response.

H2: The relationship between text and diagram CE accuracy and RCJs will be significant.

H3: The relationship between the interactions of text and diagram CEs and multiple-choice responses on RCJs will be significant.

## Method

## Participants

Forty-eight undergraduates ( $69 \%$ female) enrolled at a large mid-Atlantic university participated in this study. Their ages ranged from 18 to $24(M=20.04, S D=1.60)$, and they were compensated up to $\$ 30$ for their participation.

## Experimental Design

This study used a $3 \times 3 \times 2$ within-subjects design (18 trials). Each participant was exposed to three human agent facial expressions: neutral (neutral facial expression), congruent (i.e., joy for facial expressions congruent with the content relevancy), and incongruent (i.e., confusion for facial expressions incongruent with content relevancy). Each participant was also exposed to each type of multimedia content relevancy: fully relevant (text and diagram relevant to the question), text somewhat relevant (but diagram still fully relevant), and diagram somewhat relevant (but text still fully relevant). Additionally, two types of questions were posed: function (regarding the function of a body system) and malfunction (regarding a malfunction of a body system). Based on these manipulations each student completed 18 trials, with different combinations of human agent facial expression, multimedia relevancy type, and question type. For this paper, our analyses focused on metacognitive judgments across the trials and experimental manipulations.

## Materials

The materials used in this study included the following: an informed consent form; a demographic questionnaire; and a researcher-developed, 4-foil, 18-item multiple-choice pretest of basic knowledge of human body systems (e.g., integumentary and nervous systems). Each question on the
pretest specifically related to the content presented in each multimedia science content slide.

Additionally, this study included 18 researcher-developed multimedia science content slides developed with a faculty member in human biology. The relevancy manipulations were created by including information that was related to but not necessary for answering the question.

## MetaTutor Multimedia Learning Environment

The MetaTutor multimedia learning environment is a multimedia-based content presentation tool designed to examine the influence of a human agent's facial expressions on participants' cognitive strategies and metacognitive judgments during learning about human body systems. The environment consists of a human agent capable of facially expressing several emotional states (i.e., neutral, confusion, joy), science questions and corresponding multimedia science content, and metacognitive judgment prompts (EOLs, text and diagram CEs, and RCJs). The multimedia science content consists of three paragraphs (Flesch-Kincaid readability score range: $9.1-12.5 ; M=10.5$ ) and a diagram depicting the concept described in the text.

The environment presents 18 linearly structured, selfpaced trials consisting of metacognitive judgments (e.g., EOLs, CEs, and RCJs), multimedia content presentation, and human agent facial expressions.

The 18 trials have the identical format. In each trial, participants are first presented with a science question and asked to submit an EOL, How easy do you think it will be to learn the information needed to answer this question? Participants made their EOL judgment on a scale from $0 \%$ to $100 \%$, increasing in increments of $20 \%$. Participants were then presented with a content slide containing the text, diagram, science question presented previously, and human agent. After 30 s (to ensure participants had enough time to initially review the material), participants were prompted to assess the relevancy of both the text and diagram, Do you feel the text/diagram on this page is relevant to the question being asked?, by making two CE judgments on a Likerttype scale (ranging from $1-3$ ) on the following statements: The text/diagram is relevant, The text/diagram is somewhat relevant, and The text/diagram is not relevant. Upon making their text and diagram CEs, the human agent expressed a congruent, incongruent, or neutral facial expression based on the relevancy of the content (e.g., a congruent facial expression of joy if the text and diagram were relevant to the question being asked). Following the agent's expression, participants were permitted to reread the text and reinspect the diagram at their own pace. After they re-examined the multimedia content, participants were prompted to answer the science question by choosing the correct response from 4-foil answers. After submitting their answer, participants were prompted to make a RCJ by answering How confident are you that the answer you provided is correct? Participants made their judgment on a scale from $50 \%$ to $100 \%$ increasing in increments of $10 \%$. After submitting their response, participants were required to justify their
answer by typing their response into a text box. Subsequently, participants were asked to make another RCJ based on their justification. This procedure was followed for all 18 trials with each trial randomized across participants.

## Procedure

Once participants entered the lab they were asked to complete an informed consent form. Then the eye tracker was calibrated by the researcher. ${ }^{1}$ Following calibration, participants were asked to complete a computerized demographic questionnaire and an 18-question, 4-foil pretest that assessed their basic science knowledge across the multiple body systems (e.g., urinary, endocrine) presented in the experiment. After the pretest, participants completed the 18 previously described trials. The experimental session lasted approximately 90 min .

## Coding

Text and diagram CE judgments were recorded across the 18 trials (i.e., 18 text +18 diagram $=36$ total CE judgments for each participant). Responses were coded based on their accuracy, such that an accurate CE judgment was given a score of 1 , a partially correct judgment was scored as 0.5 , and an incorrect judgment was scored as 0 . For example, if participants judged the diagram as somewhat relevant and a text as fully relevant during a "diagram somewhat relevant" trial, they were given a score of 1 for each response because the text was still fully relevant to the question being asked, whereas the diagram was only somewhat relevant.

Participants' responses to the 4-foil, multiple-choice questions were coded by correctness. A correct response was coded as 1 and an incorrect response was coded as 0 .

Participants' RCJs were coded on a scale from $50 \%$ to $100 \%$. A score of $50 \%$ indicated participants simply guessed at their answer (indicating they believed they had a 50/50 chance of getting their answer correct), whereas a score of $100 \%$ indicated participants were completely confident in their response.

## Results

## Research Question 1: Are accurate text and diagram CEs associated with an increase in the likelihood of an accurate multiple-choice response?

A fully unconditional model (i.e., with no predictor variables) dichotomous outcomes (i.e., accurate multiplechoice response $=1$, inaccurate $=0$ ), was conducted on multiple-choice accuracy. Results indicated that the average probability of responding to a multiple-choice question correctly was $60 \%$.

A dichotomous outcomes model was conducted on multiple-choice accuracy (i.e., accurate $=1$, inaccurate $=0$ ) with text and diagram CE accuracy as the predictor variables. Results revealed that more accurate text CEs

[^417]$(O R=1.98, t=3.09, p=0.002)$ but not diagram $\mathrm{CEs}(O R=$ $0.98, t=-0.10, p>0.5$ ) were associated with an increase in the likelihood of correctly answering multiple-choice questions. Specifically, as text CE response accuracy increased, there was a $98 \%$ increased chance of responding correctly. That is, if participants were accurate in their text CEs, they were substantially more likely to respond correctly to the multiple-choice questions.

## Research Question 2: Is there a significant relationship between text and diagram CE accuracy and RCJs?

A fully unconditional model conducted on RCJs indicated $29.8 \%$ of the variability was between participants $\left(\tau_{00}=\right.$ 79.61, $z=4.24, p<0.001$ ) and $70.2 \%$ was within participants ( $\sigma^{2}=187.49, z=19.99, p<0.001$ ), justifying further analysis.

An unconstrained multiple level 1 predictor model was run on RCJs using text CE and diagram CE accuracies as the predictor variables. Results revealed that an increase in both text CE accuracy ( $\gamma_{10}=5.70, t=3.95, p<0.001$ ) and diagram CE accuracy ( $\gamma_{20}=6.01, t=4.63, p<0.001$ ) significantly predicted an increase in RCJs. As the accuracies of participants' text and diagram CEs increased, their reported confidence in their performance also increased. This model accounted for $6.2 \%$ of the withinparticipant variance in participants' RCJs.

## Research Question 3: Is there a significant relationship between the interactions of text and diagram CEs and multiple-choice responses on RCJs?

A constrained multiple level 1 predictor model was run on RCJs using text and diagram CE accuracies and their interactions with multiple-choice responses as predictor variables. Results indicated the interaction between text CE accuracy and multiple-choice response accuracy was not significant ( $\gamma_{40}=1.50, t=0.50, p=0.62$ ). However, results did reveal a significant interaction effect between diagram CE accuracy and multiple-choice response ( $\gamma_{40}=-7.21, t=$ $-2.75, p=0.006$ ), such that participants whose diagram CEs were most accurate and who also had more accurate multiple-choice responses also reported more confidence in their answers (see Figure 1). This model accounted for 7.7\% of the within-participant variance in participants' RCJs.

## Discussion

The goal of this study was to examine the relationships between metacognitive judgments and their contributions to increased performance during multimedia learning. Overall, results revealed that when participants made accurate text CEs, they were more likely to respond correctly to multiplechoice questions. Additionally, accurate text and diagram CEs contributed to higher reported confidence in answers. As such, our findings augment current understanding of how different metacognitive judgments, from those


## Diagram Content Evaluation Accuracy

Figure 1: Interaction between diagram CE response accuracy and MC response accuracy on RCJs.
traditionally examined in the multimedia learning literature (e.g., JOLs), can contribute to improved performance and higher confidence.

Results from Research Question 1 indicated accurate text CEs were significantly predictive of an increased chance of responding correctly to multiple-choice questions, whereas diagram CEs were not. These results partially support our hypothesis, demonstrating participants could more accurately assess the relevancy of the text-based (as opposed to diagram-based) material related to answering the science question. Furthermore, these results are consistent with theories of multimedia learning that suggest individuals cognitively process text- and diagram-based material separately (Mayer, 2014; Schnotz, 2014). It is possible that participants not only cognitively process the text and diagrams separately, but also metacognitively monitor the information in text and diagrams separately and with varying levels of accuracy. Given evidence suggesting individuals exhibit a bias toward processing text-based information (at the expense of diagrams; Hegarty \& Just, 1993), in addition to the redundancy of the diagram-based information to the text, participants may have realized the text-based information was sufficient and thus relevant enough to answer the multiple-choice questions correctly.

As hypothesized, results from Research Question 2 demonstrated that text and diagram CEs significantly predicted higher RCJs. Specifically, the more accurate participants' text and diagram CEs were, the more confident they were in their multiple-choice responses. Taken together with the previous finding, these results indicate participants may have relied on their relevancy judgments of both the text and diagram when they made their RCJs (as opposed to answering the question). As such, this finding significantly augments research on metacognitive judgments during multimedia learning by indicating a significant relationship between multiple metacognitive judgments.

Lastly, results from Research Question 3 indicated the interaction between diagram CE accuracy and multiplechoice response accuracy significantly predicted increased RCJs. More specifically, participants who provided more
accurate diagram CEs and responded accurately to multiplechoice questions also reported more confidence in their answers. These results partially support our hypothesis that both text and diagram CEs interact with multiple-choice responses to predict increased RCJs. Additionally, this result is supported by previous literature that suggests a significant relationship between performance and RCJs (e.g., Mengelkamp \& Bannert, 2010). These results also support our assumption that since cognitive processes are different for different representations of information, so too are metacognitive monitoring processes. However, research is limited regarding the metacognitive processes involved when learning with and comprehending diagrams.

Overall, these results suggest that accurately assessing the relevancies of text and diagrams differentially impacts performance and future metacognitive judgments (e.g., accurate CEs related to increased RCJs). Results also indicated that when participants responded to multiplechoice questions, they relied on their metacognitive judgments of the text rather than diagrams. In contrast, participants relied on metacognitive judgments of diagrams and their performance when making RCJs. Previous research has indicated a significant relationship between CEs and performance (e.g., Greene \& Azevedo, 2009). However, unlike previous literature, these results suggest text and diagram CEs differentially impact not only performance, but also reported confidence. Ultimately, these results confirm that other metacognitive judgments for different representations of information can predict greater performance during multimedia learning.

## Limitations

Our study has several limitations. First, as we were primarily interested in the relationship between metacognitive judgments (e.g., CEs, RCJs) and performance across conditions, we did not examine the impact of content relevancy (e.g., fully relevant text and diagram, text less relevant, diagram less relevant) or question type (e.g., function vs. malfunction science question). Furthermore, the information needed to answer the multiple-choice questions correctly was primarily located in the text, which may have influenced participants' CE judgments. Future research should include separate function and malfunction questions based on the information presented in the diagrams. Moreover, we did not examine the accuracies of RCJs as multiple-choice responses were dichotomously coded as correct or incorrect. Future research will include measures of absolute and relative accuracies for RCJs (e.g., Schraw, 2009). Lastly, we can only make limited conclusions regarding the underlying cognitive and metacognitive processes (e.g., multiple fixations on irrelevant diagrams) that contributed to the accuracies of the text and diagram CEs and multiple-choice responses, as multichannel trace data (e.g., eye tracking) were not analyzed. Despite these limitations, this study has several important implications.

## Future Directions and Implications

The results of this study have important implications for future studies examining the influence of metacognitive judgments on performance during multimedia learning. First, future research should include analyses of multichannel trace data (e.g., eye tracking, facial expressions of emotions) that would allow for a more comprehensive depiction of the cognitive, affective, and metacognitive processes that occur when making CEs during multimedia learning (see Azevedo, 2014). Specifically, analyzing eyetracking data can provide a micro-level description of the cognitive processes (e.g., coordination of information sources) contributing to increased performance and accurate text and diagram CEs. For example, does more time spent reading the text contribute to more accurate text CEs? Do specific eye-movement "signatures," as evidenced by scan path analyses, indicate greater integration of multimedia information and subsequently lead to increased performance? Further, examining the influence of participants' affective processes (e.g., emotions) would provide evidence of how they influence cognitive and metacognitive processes. For example, are participants' facial expressions of confusion predictive of decreased CE accuracy? How do participants' facial expressions of frustration influence the quality of their multiple-choice responses? Lastly, as this study was limited to analyzing the accuracy of RCJs, future research should seek to determine how CEs contribute to the accuracy of RCJs. It is possible that participants' CEs were accurate, but they exhibited over- or under-confidence when making their RCJs.

As our results indicated that diagram but not text CEs interacted with multiple-choice responses to predict RCJs, they emphasize the differential impact of multiple representations of information on participants' metacognitive judgments. Future research should examine the specific impact of different representations (e.g., diagrams, graphs, illustrations) on participants' metacognitive judgments to address the gap in the literature and gain better understanding of the metacognitive monitoring processes involved during multimedia learning.

Using a within-subjects design allowed us to examine the differential impact of how accurate metacognitive judgments influenced performance and confidence with reduced error caused by individual differences. Additionally, using multilevel modeling (Raudenbush \& Bryk, 2002) enabled us to accurately assess within-subjects variance without violating traditional statistical assumptions (e.g., independence of observations) that many withinsubjects designs ignore. Despite these benefits, future research should explore other experimental designs that are less controlled (e.g., more naturalistic) to increase the ecological validity of these findings. Due to our sample size, we did not find significant between-subjects variance; future research should replicate these analyses with larger samples to determine individual differences indicative of improved metacognitive judgment accuracy and performance (e.g., prior knowledge of body systems).

Additionally, these results indicate the importance of coordinating multiple sources of information (e.g., text and diagram) and can be used to inform the design of educational training regimens. For example, future research should explore the impact of cognitive (e.g., Bergey, Cromley, \& Newcombe, 2015) and metacognitive (e.g., Azevedo, 2014) instruction that emphasizes how individuals should learn using both text and diagrams. Training can be provided to demonstrate how to accurately judge the relevancy of texts and diagrams, as well as emphasize the importance of accurate metacognitive judgments in relation to increased performance. Furthermore, these results can also inform the design of future intelligent, adaptive multimedia-based learning environments to support and scaffold accurate metacognitive judgments. If participants continuously make inaccurate text CEs, the system can intervene by cueing their attention to the relevant text-based information or by providing additional relevant declarative and conditional knowledge (e.g., how to accurately judge the relevancy of different representations of information).

Lastly, the results from this study suggest accurate metacognitive judgments are required for increased performance and confidence during multimedia learning. Traditionally, metacognitive judgments during multimedia learning have been found to be largely inaccurate. However, our results indicate other metacognitive processes (e.g., CEs) may be more informative of increased performance. For example, future studies could examine the influence of accurate feelings of knowing (i.e., individuals are aware of having read information but are unable to recall it on demand) and how they can contribute to increased performance during multimedia learning. As such, future research examining the influence of other metacognitive judgments will significantly augment our understanding-as well as the contemporary theoretical frameworks of multimedia learning-of the relationship between cognitive and metacognitive processes contributing to increased performance.

## Acknowledgments

The research presented in this paper is supported by funding from the National Science Foundation (DRL 1431532). The authors would like to thank the members of the SMART Lab and the IntelliMedia Group at NCSU for their assistance in this project.

## References

Azevedo, R. (2014). Multimedia learning of metacognitive strategies. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning ( $2^{\text {nd }}$ ed.). New York, NY: Cambridge University Press.
Azevedo, R., Greene, J., \& Moos, D. (2007). The effect of a human agent's external regulation upon college students' hypermedia learning. Metacognition and Learning, 2, 6787.

Bergey, B. W., Cromley, J. G., \& Newcombe, N. S. (2015). Teaching high school biology students to coordinate text
and diagrams: Relations with transfer, effort, and spatial skill. International Journal of Science Education, 37, 2476-2502.
Burkett, C., \& Azevedo, R. (2012). The effect of multimedia discrepancies on metacognitive judgments. Computers in Human Behavior, 28, 1276-1285.
Butcher, K. (2014). The multimedia principle. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning ( $2^{\text {nd }}$ ed.). New York, NY: Cambridge University Press.
Eitel, A. (2016). How repeated studying and testing affects multimedia learning: Evidence for adaption to task demands. Learning and Instruction, 41, 70-84.
Greene, J. A., \& Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. Contemporary Education Psychology, 34, 18-29.
Hegarty, M., \& Just, M. A. (1993). Constructing mental models of machines from text and diagrams. Journal of Memory and Language, 32, 717-742.
Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning ( $2^{\text {nd }}$ ed.). Cambridge, England: Cambridge University Press.
Mengelkamp, C., \& Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. Memory \& Cognition, 38, 441-451.
Nelson, T. O., \& Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." Psychological Science, 2, 267-270.
Nelson, T. O., \& Narens, L. (1990). Metamemory: A theoretical framework and new findings. The Psychology of Learning and Motivation, 26, 125-173.
Pilegard, C., \& Mayer, R. E. (2015). Within-subject and between-subject conceptions of metacomprehension accuracy. Learning and Individual Differences, 41, 5461.

Raudenbush, S. W., \& Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods ( $2^{\text {nd }}$ ed.). Thousand Oaks, CA: SAGE.
Schnotz, W. (2014). The integrated model of text and graphics comprehension. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning (2 ${ }^{\text {nd }}$ ed.). New York, NY: Cambridge University Press.
Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. Metacognition and Learning, 4, 33-45.
Serra, M. J., \& Dunlosky, J. (2010). Metacomprehension judgments reflect the belief diagrams improve learning from text. Memory, 18, 698-711.
Vössing, J., Stamov-Roßnagel, C., \& Heinitz, K. (2016). Images in computer-supported learning: Increasing their benefits for metacomprehension through judgments of learning. Computers in Human Behavior, 57, 221-230.

# Interactive Communicative Inference 

Larry Muhlstein (lmuhlstein@ucsd.edu)<br>University of California San Diego<br>Department of Cognitive Science, 9500 Gilman Drive<br>San Diego, CA 92093 USA


#### Abstract

In the search for an understanding of human communication, researchers often try to isolate listener and speaker roles and study them separately. Others claim that it is the intertwinedness of these roles that makes human communication special. This close relationship between listener and speaker has been characterized by concepts such as common ground, backchanneling, and alignment, but they are only part of the picture. Underlying these processes, there must be a mechanism for making inferences about our interlocutors' understanding of words and gestures that allows us to communicate robustly and efficiently without assuming that we take the same words to have the same meaning. In this paper, I explore this relationship between language and concepts and propose an interactive mechanism that can facilitate these latent conceptual inferences. Finally, I show how this proposal paves the way for a more precise account of the role of interaction in communication.


Keywords: Communication; Coordination; Interaction; Pragmatics; Bayes; Cognitive Linguistics; Inference; Discourse

## Introduction

FRIEDA: "I study cognitive science"
FRED: "Cool! The brain is so interesting!"
FRIEDA: "Uh, cognitive science $\neq$ neuroscience..."
Human communication is fraught with misunderstanding, incorrect assumptions, and uncertainty, yet we still manage to make it work. To handle these impediments, we make ample use of processes such as grounding (H. H. Clark \& Brennan, 1991), alignment (Pickering \& Garrod, 2004), repair, and backchanneling (Schegloff, Jefferson, \& Sacks, 1977), all of which are well described in the scientific literature. Because these processes are thought to be somewhat modular, many linguists-especially computationalists-find it useful to remove these processes from their models and experiments and reasonably assume that they can be added back in later when a more complex and complete theory is desired. While it is reasonable to excise details tangential to the core phenomenon of study, to remove these interactive processes underestimates the degree to which they are embedded in human communication and fails to appreciate how indispensable they are to communicative success.

The misunderstanding at hand emerges from the fact that these processes are not fully understood. On one hand, grounding, alignment, repair, and backchanneling allow us to repair inferential errors and to establish tighter conventions, but each of these processes also presupposes the ability to detect misalignments and miscommunications in the first place. While this might not immediately appear to be much of a problem, there is a complete dearth of empirical literature on such inferences and of any theoretical analysis about
their mechanics. The purpose of this paper is to provide a preliminary analysis of the necessary and sufficient properties of these interactive communicative processes and to argue that any reasonably sophisticated understanding of human communication must build upon an epistemologically-sensitive theory of how we detect and repair misalignments and misinferences in order to communicate robustly and veridically.

## Dead reckoning behavior in models of communication

Contemporary models of human communication tend to operate with strong assumptions about what I will call conceptual alignment-the extent to which interlocutors' mappings from surface structure (words, phrases, and actions) to hidden structure (meanings and concepts) align with one another. Conceptual alignment captures how the correspondence ${ }^{1}$ between a speaker's beliefs about the listener's understanding of a situation and the listener's actual understanding affects their ability to communicate ${ }^{2}$. Typically, these models assume that both interlocutors have complete probabilistic conceptual alignment, and therefore that the listener's belief distribution about the meanings the speaker intends to communicate with her utterances is equivalent to the speaker's probability of producing those utterances given her communicative intentions. With this assumption in place, a rich subset of pragmatic behavior including all sorts of implicatures can be fruitfully investigated, and have been with the advent of Bayesian models of pragmatics such as RSA and its many variants (Degen, Tessler, \& Goodman, 2015; Frank \& Goodman, 2012; Goodman \& Lassiter, 2014; Kao, Wu, Bergen, \& Goodman, 2014).

However, if we were to drop this assumption from these models in favor of a more realistic amount of uncertainty, we would notice some problems. For example, if we allow there to be misalignment between the interlocutors' concepts, we would find that any attempt at communication results in a systematic pattern of errors that are not correctable within the scope of the model. A plausible case of this might be if I said the word "justice" intending it to mean something akin to Rawlsian fairness, where all decisions are made from a position of agnosticism about where one falls in society, but my interlocutor thought I was talking about retributive justice, then everything that I said would be misinterpreted, and this would continue indefinitely in the aformentioned mod-

[^418]els because they don't contain any mechanism for detecting conceptual misalignment (instead they assume alignment and proceed from there). An even simpler case could involve a sense/reference mismatch where I use the word "speaker" to refer to a lecturer, but my interlocutor instead interprets it to refer to a sound production device. While it is likely that the vast difference in meaning would cause me or my interlocutor to notice the misinterpretation, a large class of current formal models of conversation would proceed as if I was being properly interpreted at all times.

It is useful to refer to this type of behavior as dead reckoning communication, as it offers the same perils as nautical navigation without taking periodic measurements of location. A ship without instruments, sailing perpetually into the fog, is inevitably bound to stray far from its destination no matter how precisely it was pointed at the beginning of its journey. This happens because of the pervasive uncertainty about the ship's motion, the effects of navigational actions (trimming the sails, hoisting the spinnaker, turning the rudder, etc.), and the environment-which is always changing. Even if a ship's captain plans out a series of actions in advance in order to get the ship to its intended destination and executes these actions flawlessly (without taking interim measurements about the position, speed, and heading of the ship), there is but a tiny chance that the ship will end up in its intended port. So it goes with communication sans feedback. If a speaker wants to convey some concept or scenario to a reader and she develops a series of communicative actions (a communication plan) and executes it without observing the listener's interpretations, then she risks the listener's gross misinterpretation unless she is willing to put in substantially more detail and effort than is typically prudent or even possible in a conversation. Even for the most closely aligned concepts, communication is bound to stray off course if the participants don't continuously probe the state of the discourse and correct its course when necessary. As H. H. Clark and Wilkes-Gibbs (1986) argued, the constraints of conversation restrict us to brief, ad hoc, ephemeral communicative actions, which limits interlocutors' ability to provide the kinds of lengthy descriptions that might be present (and necessary) in a book. However, conversation also affords interaction. As we will see, this offers additional opportunities for coordinated communication between interlocutors by allowing them to make inferences about each others' interpretations. Just like a navigator's instruments allow her to detect the ship's position and velocity and make informed corrections to its course, a speaker's inferences allow her to probe the state of her partner's understanding and choose her successive communicative actions accordingly.

## Response-based inference

To combat this undesirable dead reckoning behavior, we can look both at the necessary properties for any mechanism to handle pervasive uncertainty and misalignment (given the right abstraction of the problem) and at empirical investiga-

(a) Dead reckoning navigation
(b) Navigation with instruments

Figure 1: Dead reckoning vs. instrument-based navigation on a ship. (a) Dead reckoning cannot handle uncertainty and so the ship ends up far from its intended destination. The gray line represents the planned trajectory, while the black line shows the actual trajectory under dead reckoning. (b) Instrument-based navigation allows the captain to correct the ship's course as it goes leading the ship exactly to its intended destination.
tions about how humans specifically seem to handle it in the case of communication.

## Challenge-response authentication

A particularly good formal analogy for the solution I will soon propose is the cryptographic concept of challengeresponse authentication. Challenge-response authentication was developed as a solution to a problem much like the issue of dead reckoning in communication. Imagine that there are two interlocutors, Alice and Bob, each sitting at one end of a digital communication channel (imagine them sitting at desks in separate buildings connected via the internet.) Bob wants to send Alice a secure message, but first Bob needs to know for sure that he is communicating with Alice and not with a malevolent interloper such as their friend Charlie. To verify Alice's identity, Bob needs Alice to say something that only she would know, such as the secret phrase they agreed upon earlier. Bob receives this phrase from the communication channel and concludes that it must have come from Alice. But Charlie, the conniving and generally clever chap that he is, has tapped the communication channel and observes Alice and Bob's secret phrase. The next day, when Bob wants to communicate with Alice, he hears this same phrase, and again believes that it came from Alice. Unbeknownst to him, Alice was sick and wasn't at the computer terminal she normally uses to communicate with Bob. Instead, the secret phrase came from Charlie, who now proceeds to communicate with Bob as if he was Alice.

A natural question then follows: how can Bob and Alice communicate securely without risk of Charlie impersonating Alice? The answer lies in challenge-response authentication. Bob, having discovered the fatal flaw in his authentication system, comes up with a clever alternative. Instead of agree-
ing with Alice upon a secret phrase, they instead come up with a secret relationship between phrases-i.e. a secret function. The next day, when Alice and Bob try to communicate, Bob sends Alice a message called a challenge. When Alice receives this challenge, she feeds it into the secret function and gets an output that she then sends back to Bob as a response. When Bob receives this response, he feeds his challenge into his own copy of the secret function and compares the result with Alice's response. If they match, then Bob can be sure that he is communicating with Alice and not Charlie. To understand why this works, we can look at Charlie's behavior in this scenario. Like before, Charlie has tapped the communication line and receives both Bob's challenge and Alice's response. The next day, when Charlie tries to pretend that he is Alice, he receives a challenge from Bob. This challenge however, is not the same as the one he saw the day before. This means that the correct response is different as well! Because the secret knowledge is a full function, the observation of a few challenge-response pairs is not sufficient to induce the full function or the correct response to additional challenges. If the secret functions are properly designed, then virtually no amount of observations of the challenge-response pairs will be sufficient for Charlie to induce the secret function. Upon realizing this information, the sullen Charlie decides to leave Alice and Bob alone to search for others with weaker verification algorithms to deceive.

## Utterance-response contingencies

How does this cryptographic mechanism relate to human communication and how does it help us avoid dead reckoning behavior? The problem in human communication is not in establishing the identity of the interlocutor, but rather in verifying the interlocutor's comprehension. If we replace the secret functions from challenge-response authentication with out latent conceptual understandings, we can use the same sort of strategy to verify the similarity or alignment between our communicative intentions and the interlocutor's interpretations. If Alice produces utterance $x$ for Bob, and Bob responds with utterance $y$, then Alice can check Bob's comprehension of $x$ by modeling the plausibility that Bob would have generated response $y$ given various interpretations of $x$. In the simple case, if Bob interprets $x$ exactly as Alice intends, then his response $y$ will be identical to Alice's prediction about his response given her intended interpretation. In the case where there are multiple plausible interpretations, Bob's response $y$ provides Alice with information about how likely Bob is to have made each of the possible interpretations of $x$. We will call this minimal pair of utterance and response an utteranceresponse contingency. Utterance-response contingencies are the basic building blocks of interactive communicative inference, a term I propose to denote the general process of inferring interlocutor beliefs through interaction. This also involves more complex cases where each interlocutor's beliefs about the others' understanding are updated through extended dialogical interaction, which can result in robust alignment processes that make effective discourse possible.

In real discourse, utterance-response contingencies look like this:

GEORGE (UTTERANCE): "Wasn't that a great speech yesterday?"
GEORGIA (RESPONSE): "I mean, it wasn't as horrible as I expected, but I definitely can't say it was good..."
GEORGE (REPAIR): "Oh, did you think I meant Donald's speech? I was talking about the address from the director of the ACLU."

These utterance-response contingencies can come in various forms. They can resolve referential misinterpretations like in the example above, or they can surface more subtle conceptual misalignments where the interlocutors' wordconcept mappings are misaligned, or even when the internal structures of their respective concepts are inconsistent with each other ${ }^{3}$.

The case of misaligned word-concept mappings is illustrated in the opening dialogue between Frieda and Fred.

FRIEDA: "I study cognitive science"
FRED: "Cool! The brain is so interesting!"
FRIEDA: "Uh, cognitive science $\neq$ neuroscience..."
Here we see that Frieda's concept of "cognitive science" is reflective of her being an insider to the field and therefore likely includes associations with each of the "six corners of the hexagon": psychology, computer science, philosophy, linguistics, anthropology, and neuroscience. Fred's concept, however, is closer to how many outsiders think of cognitive science ${ }^{4}-$ as alternative word for neuroscience. Using interactive communicative inference and utterance-response contingencies, Frieda is able to notice this conceptual misalignment and repair it, thus improving the alignment between interlocutors.

## A more systematic formulation

To make this idea a bit more concrete, we can situate it in the context of a more precise way of looking at human communication. Fundamentally, communication is comprised of two kinds of processes: inferences, and actions that facilitate inferences. The idea of an inference is mostly self explanatory. It refers to any kind of inference about the beliefs, communicative intentions, or mental state of your interlocutor. Such beliefs can be about things in the world or more abstract concepts, and so the contents of these inferences are almost unlimited in scope. Inference facilitation, on the other hand, refers to actions that are taken by one participant in the discourse (whom we call the speaker, even though their actions may not be vocally produced) and observed by another participant in the discourse. As these actions are caused (at least in part) by the speaker, they provide the observer of the actions (the listener) with information

[^419]about the speaker and consequently their beliefs, mental state, and communicative intentions. Since the roles of speaker and listener shift dynamically throughout a discourse, we do not associate these terms with particular conversational agents, but with the roles themselves. The listener-in responding to the speaker-therefore temporarily instantiates the role of the speaker, and even the most subtle facial expressions, when produced in a discourse, count as inference-facilitating actions. Besides these two components, there is also the idea of context, which is broadly defined to capture any effect of the discourse topic, the surroundings, or any other processes on the particular instance of communication ${ }^{5}$. The definitions of these communicative concepts are not constrained further than this because to do so would limit their generality as the basic constituents of human communication.

I choose this framework instead of other possibilities such as the pragmatic alternatives framework, classical communication theory, and informal schools of thought such as relevance theory and cognitive linguistics because it affords both precision and generality and it allows us to highlight the basic epistemics of human communication. In this formulation, all communicative actions have the same status. This is because, at the epistemic level, an utterance, a gesture, and an unintentional twist of the lips are all actions taken by a communicator that allow an interlocutor to make various inferences about her communicative intentions and her state of mind. While the contents and causal pathways associated with these actions may differ drastically, these details do not have bearing on the development of a basic framework for communicative analysis, only on specific theories subsequently derived from that framework.

To move towards a formal theory, we can choose a minimal Bayesian framework that captures only the epistemic relationships between the speaker's utterance, the listener's interpretation, the listener's response, and the context. Since all of these, except for the listener's interpretation, are observable to both parties ${ }^{6}$, we can capture their relationships in a causallyderived probabilistic graphical model where each of the nodes is observed except for the node representing the listener's interpretation, as seen in figure 2 (Pearl, 1988). This model captures the fundamental structure of the speaker's inference about the listener's interpretation via utterance-response contingencies, where the details of particular inferences depend on the particular distributions that comprise an instantiation of the model. While the broad epistemic structure of these inferences is simple, the corresponding real-world processes are anything but so. Filling in the details of this model will not be an easy process and will require a large amount of directed scientific experimentation and theoretical analysis. Therefore the goal here is not to develop a full theory of these communicative processes in real humans, but to provide a structured

[^420]

Figure 2: Bayesian network representation of interactive communicative inference. Bolded nodes are directly observable.
framework through which we can understand and investigate them. Bayesian probability is sufficiently abstract to allow us to represent these general inferential structures while allowing the rich human details to be added later.

## Application to known discourse phenomena

To further ground this framework and illustrate its relationship to empirical phenomena in human communication, we will look at three different concepts described in the literature and show how the idea of interactive communicative inference offers each of them a stronger theoretical foundation.

## Common ground

The idea of common ground, first proposed by H. H. Clark and Brennan (1991), suggests that people accumulate a shared repository of knowledge when they interact, and that subsequent interactions are facilitated by this common knowledge. This proposal has received a substantial amount of theoretical analysis, which has lead to a rich account of how humans establish common ground and make use of it in conversation (H. Clark, 1996). There is also a solid foundation of experimental evidence supporting the theory (Brennan \& Clark, 1996; Hanna, Tanenhaus, \& Trueswell, 2003). Perhaps the most notable experiment-known as the "tangrams experiment"-involved a pair of interlocutors who were given a set of cards depicting images of blocky figuresi.e. tangrams-and were tasked with getting the other person to arrange their cards in an order perceivable only to the designated speaker for the round. Complicating this significantly was the fact that the participants were separated by a visual wall and so could only communicate verbally. H. H. Clark and Wilkes-Gibbs (1986) found that the participants quickly established shared conventions for referring to the cards, which resulted in a decrease in the amount of communication required to complete the task as they continued to interact through multiple task rounds.

While this story is compelling both theoretically and empirically, it is missing a mechanistic account of the inference
processes that these interlocutors go through when building and making use of common ground. Dead reckoning models cannot capture this progressive coordination behavior, and so we turn to interactive communicative inference. The basic component of interactive communicative inference, the utterance-response contingency, accounts for how an exchange can allow the speaker to update their understanding of the listener's beliefs and therefore facilitate the listener's comprehension. When this occurs, the speaker can choose to use a referring term that they know the listener will understand, which allows the speaker to communicate more efficiently with the listener via this ad hoc convention. If both communication partners make use of such processes in their interaction, then they will come to a shared understanding of how to use language to communicate particular meanings. Over an extended interaction, such interlocutors can build a common communication system. This predicts the increasing communicative concision throughout an interaction that H. H. Clark and Wilkes-Gibbs (1986) observed. We can even imagine that, if this process involves additional people over multiple interactions, a complete conventionalized communication system should emerge from this process of building shared knowledge via interaction.

## Conversational alignment

Martin Pickering and Simon Garrod have developed a compelling theory of discourse as an interactive alignment process, which has a number of useful relations to the present theory of interactive communicative inference (Pickering \& Garrod, 2006, 2004; Garrod \& Pickering, 2004). Their theory claims that local lexico-syntactic priming in discourse produces a series of cascading effects that causes all participants in a conversation to produce similar surface structures and even to align on semantic content. Using this theory, they argue that the fundamental mechanisms operating in discourse are these alignment processes and construct an account of language processing that does not require a speaker to maintain an explicit and complex model of their interlocutor (Garrod \& Pickering, 2004). While this theory accounts for a wide range of phenomena, it does not offer any account of conceptual alignment phenomena. Conceptual alignment cannot arise from priming-based mechanisms because it is defined as the alignment between latent conceptual structures and surface communicative actions, and this relationship cannot be primed by observing surface structures from an interlocutor's communicative actions. In order to account for the facts that our conceptual structures tend to be relatively aligned within a conventionalized communication system and that we have evidence that they are aligned via interaction, an interactive communicative inference component needs to be added to Pickering and Garrod's theory of discourse.

Their theory also predicts that it should be easier for humans to efficiently communicate meaning through interactive dialog than unidirectional monologue because interaction facilitates the alignment of linguistic representations (Garrod \& Pickering, 2004). This prediction is both consistent with our
informal experience as well as predicted by the interactive communicative inference framework. For a monological description to yield effective communication, it must necessarily contain enough information such that, for all of the likely ways that a listener or reader might misinterpret the meaning, there is additional information that steers them away from these misinterpretations and towards the intended interpretation. In a discourse, however, these counterfactuals do not need to be handled via anticipation and mitigation. Instead, utterance-response contingencies make it so that only actual misinterpretations by the listener need to be ameliorated.

With the addition of the present account of conceptual alignment through interactive communicative inference, we can add a vital component to the conversational alignment story to yield a powerful framework for understanding human communication.

## Backchanneling and repair

Sociolinguistics and conversation analysis research has built an account of discourse expressed in terms of the constituent behaviors of conversations and the types of communications they enable (Sacks, Schegloff, \& Jefferson, 1974). Some of the core concepts in these accounts are the ideas of backchanneling and repair. Backchanneling is invoked in a multichannel model of communication in which interlocutors communicate the majority of their content via a main channel and provide meta-conversational signals in a backchannel (Yngve, 1970). For example, a listener may produce the affirmative "uh huh" in a backchannel to signal to the speaker that the she believes herself to be comprehending and that the speaker should proceed. While these behaviors have been convincingly shown to play a key role in natural conversation, this account leaves open the question of how the listener forms her beliefs about whether or not she is comprehending the speaker. As we have seen, belief of comprehension does not necessarily imply veridical comprehension, because there may be an undetected conceptual misalignment between the speaker and listener. Our account of interactive communicative inference suggests that listeners may be providing these cues to speakers in order to facilitate the speakers' interactive communicative inferences by completing the utteranceresponse contingency. It also suggests that speakers may be comprehending these backchannel signals, not truly as an unequivocal signal to proceed, but as a response that conveys the listener's beliefs about their comprehension. The following dialog illustrates a scenario with a distinction between the affirmative backchannel as anequivocal signal of comprehension and an account of backchanneling as concurrent response. (Tyra and Tyler are roboticists at different universities.)

TYRA (UTTERANCE): "STEVE can walk on two legs now!"
TYLER (BACKCHANNEL): "Oh"
TYRA (REPAIR): "Not Steven the child, STEVE (Self-
Taught-EVacuative-Entity) my robot. It's a ground-
breaking achievement!"

Here Tyra infers Tyler's miscomprehension through his affectless backchannel response to her achievement. She knows that he would be really excited to hear about this research achievement and therefore infers that he must believe her to be talking about a child instead. Tyler mistakenly believes himself to have comprehended Tyra's utterance and produces a backchannel response, but Tyra does not interpret it blindly. Instead, she uses the resultant utterance-response contingency to infer the miscommunication and repair the misunderstanding. As predicted in the classical account, the backchannel provides feedback to Tyra about Tyler's interpretation. However, the response does not reflect Tyler's veridical comprehension, but facilitates a more complex interactive communicative inference.

Miscommunications in conversations are corrected via repairs. These come in many forms, but are often divided into two classes: self-initiated and other-initiated. While the literature provides substantial experimental and ethnographic detail about the role of repairs in conversation, it does not offer an account of how interlocutors infer when a repair needs to be made (Schegloff et al., 1977). The present theory of interactive communicative inference offers a computational mechanism by which miscommunications can be detected in conversation, which allows speakers to repair the miscommunication by correcting their own production or by correcting the listener's interpretation.

## Conclusion

Contemporary models of communication are incomplete. While they offer a detailed understanding of the surface-level phenomena present in discourse, they do not provide a satisfactory explanation of the inferential mechanisms necessary for these phenomena. They are unable to account for the robustness of communication in spite of uncertainty, how people know when to update their beliefs about their interlocutors, and how people establish conventions. I have provided theoretical evidence tied to the empirical to show how all of these types of missing accounts can be derived from the idea of interactive communicative inference.

While researchers have long understood that discourse is not a unidirectional and isolated activity, we have demonstrated that treating it as such, even for the sake of delimiting the domain of a theory or a model, can have the effect of removing an important property of language. When the idea of interactive communicative inference is taken seriously, we can begin to construct a scientific study of communication that can account for how, despite the fact that we can never be in someone else's head to see first-hand what they believe and how they feel, they are able to show us these things simply by engaging us in a cooperative dance of action and interpretation.

## Acknowledgments

I would like to thank Sean Trott for helping motivate me to write this article as well as for his invaluable feedback. I would also like to thank Dorothy Howard for her time helping me polish the article. Finally, I am deeply indebted to the UCSD Department of Cognitive Science and all of its members for their continued support.

## References

Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482.

Clark, H. (1996). Using language. Cambridge, UK: Cambridge University Press.
Clark, H. H., \& Brennan, S. E. (1991). Grounding in communication. Perspectives on socially shared cognition, 127149.

Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22(1), 1-39.
Degen, J., Tessler, M. H., \& Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd..
Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336.
Garrod, S., \& Pickering, M. J. (2004). Why is conversation so easy? Trends in cognitive sciences, 8(1), 8-11.
Goodman, N. D., \& Lassiter, D. (2014). Probabilistic semantics and pragmatics: Uncertainty in language and thought. Handbook of Contemporary Semantic Theory. Wiley-Blackwell.
Hanna, J. E., Tanenhaus, M. K., \& Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language, 49(1), 43-61.
Kao, J. T., Wu, J. Y., Bergen, L., \& Goodman, N. D. (2014). Nonliteral understanding of number words. Proceedings of the National Academy of Sciences, 111(33), 12002-12007.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and brain sciences, 27(02), 169-190.
Pickering, M. J., \& Garrod, S. (2006). Alignment as the basis for successful communication. Research on Language and Computation, 4(2-3), 203-228.
Sacks, H., Schegloff, E. A., \& Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. language, 696-735.
Schegloff, E. A., Jefferson, G., \& Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. Language, 361-382.
Yngve, V. H. (1970). On getting a word in edgewise. In Chicago linguistics society, 6th meeting (pp. 567-578).

# Cognitive and Attentional Process in Insight Problem Solving of the puzzle game "Tangram" 

Yoshiki Nakano (nakano@ed.akita-u.ac.jp)<br>Faculty of Education and Human Studies, Akita University. Tegata Gakuen-machi 1-1. Akita City, 010-8502 JAPAN


#### Abstract

The purpose of this study is to demonstrate a constraint relaxation which is followed by the transition to an appropriate representation in insight problem solving. The puzzle game "Tangram" was used as a new insight problem, in which problem-solvers were presented a silhouette and asked to make the same configuration by arranging 7 pieces. At the beginning, problem-solvers had a constraint allocating the pieces into a geometric shape, but then relaxed this to reach the correct configuration at a later stage of problem solving. Participants' subjective assessments of their confidence to reach the solution predicted neither the constraint relaxation nor the successful problem solving. However, eye-tracking data suggested that the successful problem-solvers tended to search the problem space more widely than the unsuccessful-problem solvers.


Keywords: Tangram; insight problem solving; constraint relaxation; eye tracking.

## Introduction

The Insight problem solving, in general, exhibits a characteristic pattern as a following process (Kaplan \& Simon, 1990; Metcalfe, 1986). At first, an insight problem seems easy to solve, but problem-solvers are caught in an impasse soon after. They get stuck, think that all options have been explored and lose their way. When a sudden and useful idea comes to mind, it often leads problem-solvers rapidly to the solution.

There is an agreement that inappropriate constraints for a solution are the main source of the difficulty to solve an insight problem (Jones, 2003; Orrmerod, MacGregor, \& Chronicle, 2002). These studies suggested that insight requires the relaxation of such inappropriate constraints, and that an impasse can be broken by changing a representation of the problem. A constraint is a tendency of thinking and behavior that is taken in attempting to solve a certain problem. It usually facilitates the process to reach the solution. When a self-imposed constraint, however, is inappropriate to solve a problem, it prompts a critical difficulty to achieve insight, as it activates irrelevant knowledge and causes attempts that cannot contribute to correct solution. This leads problem-solvers into an impasse. In spite of the consensus about the source of difficulty of insight problems, a dynamic process to reach a solution has not been identified clearly. The purpose of this study is to specify an inappropriate constraint which inhibited the insight into the solution in a geometric problem solving, and to provide direct evidence about a critical factor for the


Figure 1: The 7 pieces of Tangram. Top: a configuration when the pieces are divided from a square. Bottom: Separated seven pieces. A problemsolver of Tangram move, rotate or combine them to complete a task silhouette.
successful problem solving. In addition, an eye-tracking technique is adopted to examine whether a proper searching of the problem space can lead successful problem solving or not (Thomas \& Lleras, 2009).

This study also aims to test whether the difficulty of the constraint relaxation can be reflected in an apparent task performance of a problem-solver while independent from subjective awareness. For these purposes, the puzzle game of "Tangram" was used as a new insight problem. Tangram consists of 7 triangular, square or parallelogram pieces (Figure 1). Problem-solvers are presented a task silhouette and required to make the same configuration by arranging these pieces. In Tangram, numerous task silhouettes can be composed by the 7 pieces; for example geometrical shapes, animals or objects. Because each silhouette has an individual configuration of the pieces, problem-solvers often cannot find a correct configuration immediately after the silhouette has been presented. They usually repeat trials with arranging the pieces until the solution is completed.

Tangram has the advantages of allowing researchers to monitor task performance of a problem solving requiring insight. Nakano (2009) recorded protocols of participants and movements of the pieces. His protocol data revealed that the participants who could complete the task silhouette had expressed an "Aha" experience before reaching the correct configuration. Nakano (2009) found that the participants tended to combine the pieces into a geometric shape such as a square or a triangle. This tendency was usually involved combinations of the 2 largest triangles among the 7 pieces. Such the constraints will facilitate the process to reach the solution, in the case that a correct configuration includes geometric combination of the pieces as a square or a triangle. For example, the task silhouette of an arrow has two patterns of the correct configuration, and the both include geometric combination of the 2 largest triangles (bottom in Figure 2). However, when these 2 pieces must be combined into an irregular configuration to complete a task silhouette, the tendency to construct a geometric shape will inhibits the insight into the correct configuration. For example, to complete the silhouette of a lion (top-left in Figure 2) the 2 triangle pieces must be attached by sliding their longer sides in an opposite direction to each other (top-right in Figure 2). Most problem-solvers who tried to complete this silhouette reported that it was difficult to discover this irregular configuration (Nakano, 2009). This finding indicated that the problem-solvers imposed apparent constraints on allocations and combinations of the pieces in the problem solving of Tangram.

The primary purpose of the present experiment is to investigate the correspondence between the constraint relaxation and explicit movements of the pieces. For this purpose, movements of the 7 pieces were recorded and combinations of the 2 largest triangle pieces were identified. These combinations will change as the initial constraint is relaxed and more appropriate representations are constructed. Participants who try to complete the task silhouette of an arrow will not have to relax the constraint to allocate the 2 largest triangle pieces in a geometric pattern. So they will reach the solution easier and faster comparing to the task silhouette of a lion. When they try to complete this task silhouette, they find it impossible to complete the solution only by arranging the pieces into familiar geometric combinations. Then, the participants arrange the pieces into other combinations. To investigate this hypothesis, arrangements of the 2 largest triangle pieces are categorized into three categories, geometric, transitional or irregular combinations. The geometric means the combination of the 2 pieces was a square or a triangle. The irregular combination is that the 2 pieces have a contact each other but no corners of the pieces meet or the edges of the pieces were placed adjacently without combining their corners. The rests of combinations, as that the pieces were arrange in the same direction or in a symmetric configuration, are categorized in the transitional combination.

The second purpose is to examine independence between


Figure 2: Top-left; the task silhouette of a lion. Topright; the correct configuration of the 7 pieces constructing the silhouette of a lion. The two largest triangle pieces (greyed part) are allocated in irregular combination. Bttom-left; the task silhouette of an arrow. Bottom-center and -right; the correct configuration constructing the silhouette of an arrow. There are two patterns of correct configuration. In both, the 2 pieces are allocated in a geometric combination.
the constraint relaxation and subjective awareness in the problem solving of Tangram. In Nakano (2009), participants were required to evaluate how confident they could complete the correct configuration before and during problem solving of Tangram. The subjective confidence decreased over time even in the successful problem-solvers. Thus, the subjective evaluation did not predict performance on the insight problem, as had been suggested by Metcalfe and Wiebe (1987). This finding supports the idea that the transition from one rule to a more appropriate representation for a solution proceeds without subjective awareness.

## Methods

## Participants

Twenty undergraduates ( mean age $=20.8$, age range $=20-$ 26) participated in the experiment. All were naive to Tangram.

## Apparatus

The Tangram was comprised of 7 pieces. A set of the pieces was made by dividing a square plate which was 11.8 cm in width and in height (Figure 1). The task silhouette was a lion or an arrow (Figure 2 left). The size of the both silhouettes were the same as one-fifth of the correct configuration which constructed by the 7 pieces. During the task, participants assessed how confident they were to complete the task at the Graphics Rating Scale (GRS). The GRS involved 6 verbal descriptors and 13 scale marks along a horizontal line. A description "Not at all" was written at the left end of the horizontal line, and "Nearly completed" was at the right end. An index arrow was attached on the GRS board so that the participants could indicate their
positions on the horizontal line taking the descriptions into account. At the beginning of the experiment, the index arrow was located on the center of the 13 scale marks. A digital video camera (Panasonic NV-GS100) recorded movements of the pieces and a location of the index arrow from 65 cm above.

## Procedure.

A participant sat down in front of a work desk on which the 7 pieces of Tangram and the GRS board were situated. The pieces were allocated on the desk as depicted in Figure 1 (below). After the instructions about what the experiment involved, the participant was presented the task silhouette which was printed in black and was required to complete the configuration of it by using all the pieces. The participant indicated how he or she would be able to complete the correct configuration of the silhouette by the index arrow on the GRS board. During the assessment, the participant was allowed to see the pieces and the silhouette but was not permitted to touch or move the pieces. After the assessment, the participant was allowed to move the pieces. Each session lasted for 240 sec , and was followed by the assessment on GRS. Then 1 minute of rest period was given to the participant until the next session was started.

When a participant completed the correct configuration, the problem solving was successfully ended. A sum of the manipulation time until the completion was accounted as a solution time of the "completer". A participant who could not complete the correct configuration until the end of session 5 was counted as a "non-completer". In this case, the total manipulation time was 20 min .

All the participants participated in the two-day experiment. Either the silhouette of the lion (lion task) or the arrow (arrow task) was given to the participant as a task in each day. The order of the task silhouettes was counterbalanced among the participants.

## Recording of eye movements

During process of the problem solving of Tangram, eye movements were captured using Talk Eye Lite (Takei Scientific Instruments Co. Ltd, Japan). Eye tracking was operated in monocular mode on the right eye and at a sampling rate of 33 Hz . The participants were seated on a chair and their head was fixed by using a chin and forehead rest to keep a distance from a surface of a work desk to eye in approximately 40 cm . The surface of the work desk was tilted to about 10 degree angle. The 7 pieces of Tangram were located on the desk surface and were moved within the range of 25 cm in length and 35 cm in width so as not to go outside the participant's eye-field.

## Results

Among the 20 participants, 9 completed the correct configuration of the task silhouette of a lion. The mean solution time was $594.7 \mathrm{sec}(S D=376.0 \mathrm{sec}$, Min $=106 \mathrm{sec}$, Max $=1080 \mathrm{sec}$ ). Three of the 9 completers had finished in session 1 , and 2 of the remaining 6 had finished in session 3. Three had finished in session 4 and 1 had finished in the last session. In the arrow task, 12 participants completed the
correct configuration within 5 sessions. The mean solution time was $483.6 \mathrm{sec}(S D=331.1 \mathrm{sec}$, Min $=58 \mathrm{sec}$, Max $=$ 1030 sec ). Among the 12 participants 4 had finished in session 1 . Two of the remaining 8 completers had finished in session 2. Three of the remaining 6 completers had finished in session 3. Two had finished in session 4 and 1 had finished in the last session. Eight of the 20 participants completed both the lion and the arrow task. The averaged solution time of these "high-achievers" was faster in the arrow task ( 532.8 sec ) than in the lion task ( 644.8 sec ), but statistical analysis revealed that the difference did not reach significant level ( $p>.1$ ). Therefore, contradicting to the expectation, both the ratios of the completer and their solution time did not indicate that there is a difference in difficulty between the two task silhouettes.

Ratings of subjective confidence to the completion were identified from a location of the index arrow on GRS from the video image. The value of 0 was assigned to the mark on left extreme side, and 12 was assigned to the right extreme mark. Six assigned to the central mark which located between the descriptions "possibly, I can complete" and "possibly, I cannot complete". A value of each rating was identified on the basis of relative distance of the index arrow from the marks. Mean ratings were calculated among the completers and the non-completers in each sessions. Figure 3 show the plots of the mean ratings as a function of the sessions for each task silhouette. Because the participants who had completed the task silhouette were not engaged in the assessment anymore, sample numbers included in the plots of the completer decreased as the session proceeded. All mean ratings of non-completers included 11 participants in Figure 3a and 8 in Figure 3b.

To investigate the relationship between participants' confidence and results of the problem solving, they were categorized into a high-achiever who completed the both task silhouettes or into a non-achiever who could not complete the both silhouettes. Thus, 5 participants who only completed the either task were eliminated from this analysis. Table 1 shows mean ratings of confidence in session 1 and in the last session for the high-achievers ( $n=8$ ), excepting one participant who completed the both tasks in session 1 , and for the non-achievers ( $n=7$ ), respectively. For the completers, the last session is when they completed the task, and for the non-completers it was always the fifth session. A $2 \times 2$ analysis of variance (ANOVA) with the variables achievement (high-achiever vs. non-achiever) and time (session vs. last session) showed that the effect of time was significant, $F(1,12)=52.2, p<.01$. In detail, the mean rating in session 1 was higher than that in the last session in the both groups. Supported with this result, the linear regression lines in Figure 3 indicated that participants' subjective assessments of their confidence decreased over time. The main effect of achievement and the interaction between the two variables did not reach statistical significance.

To investigate whether constraint relaxation was reflected on actual manipulation of the pieces, arrangements of the 2


Figure 3: Plots of mean ratings of subjective confidence and linear regression lines in the lion task (a) and the arrow task (b). The plots of Grey circle with a solid line indicate results of the completers and the plots of blank triangle with a dashed line indicate results of the noncompleters. The numbers beside the plots of the completer indicate a number of samples included in each mean value. Error bars indicate $S D$ s.

Table 1: Mean ratings and SDs of a participants' confidence

|  | High-achiever | Non-achiever |
| :--- | :---: | :---: |
| 1st session | $5.79(1.8)$ | $4.82(1.6)$ |
| Last session | $3.36(2.6)$ | $1.79(0.8)$ |

largest triangle pieces were categorized and their duration times were compared. When these 2 triangle pieces were in contact with each other at least on a corner or a side of them, this arrangement was counted as a combination of the pieces. All combinations of the 2 largest triangle pieces were identified on the video image and categorized into a geometric, a transitional or an irregular combination. In the case that the 2 pieces were combined into a triangle or a square, such the allocation was categorized into the geometric combination. In the arrow task, this category includes the correct combination of the 2 pieces. The participants often allocated the 2 pieces such that the short edge of one triangle is placed adjacent to the long edge of the second triangle so that no corners of the sides met, additionally the corner of the long edge of the second triangle did not make contact with the first triangle. As with these examples, an allocation of the 2 pieces in which the edges were placed adjacently without combining their corners was categorized into the irregular combination. In the lion task, this category includes the correct combination of the 2 pieces. All other combinations that were categorized neither into the geometric nor into the irregular combination were categorized into the transitional combination. A measure of the duration of the time in which each combination was made was taken from the video image. The measurement started from the moment at which the 2 pieces were allocated into a certain combination and ended when they were separated or changed into the other arrangement. To investigate a transition of a predominant combination of the pieces, overall manipulation time of each participant was divided into a first and a second half period. A cumulative duration time of each combination was calculated for each participant. Table 2 shows mean percentage of the cumulative duration time of each combination over the total manipulation time.

Three-way ANOVA with the variables completion (completer vs. non-completer), time ( $1^{\text {st }}$ vs. $2^{\text {nd }}$ half) and combination (geometric, transitional or irregular) was computed on percentage of the cumulative duration time. In the lion task, the results showed that there was a significant main effect of time variable, $F(1,18)=5.1, p<.05$, and two-way interaction between completion and time, $F(1,18)$ $=7.9, p<.05$. Post hoc analysis revealed that, in the completers, an averaged percentage of the cumulative duration time over the three combinations was significantly higher in the second half ( $14.9 \%$ ) than in the first half $(6.0 \%), F(1,18)=12.7, p<.01$, but the difference was not significant in the non-completer ( $11.3 \%$ in the $1^{\text {st }}$ half; $10.3 \%$ in the $2^{\text {nd }}$ half). Additionally, two-way interaction between time and combination was also significant, $F(2,36)$ $=7.5, p<.01$. Post hoc analysis revealed that the percentage significantly increased than the first half ( $5.7 \%$ in the $1^{\text {st }}$ half vs. $17.9 \%$ in the $2^{\text {nd }}$ half), $F(1,54)=12.8, p<.01$, while for the transitional combination the increase was marginally significant ( $8.2 \%$ vs. $14.8 \%$ ), $F(1,54)=3.7, p<.06$. In contrast, for the geometric combination, the percentage in the first half was significantly higher than the second half

Table 2: Percentage of the cumulative duration time of each combination over the total manipulation time

| Lion task <br> Geometric | Completer | Non-completer |
| :---: | :---: | :---: |
| $1^{\text {st }}$ half | $9.1 \%$ |  |
| $2^{\text {nd }}$ half | $2.3 \%$ | $15.2 \%$ |
| Transitional |  | $8.1 \%$ |
| $1^{\text {st }}$ half | $6.0 \%$ | $10.4 \%$ |
| $2^{\text {nd }}$ half | $13.1 \%$ | $16.5 \%$ |
| Irregular |  |  |
| $1^{\text {st }}$ half | $3.1 \%$ | $8.3 \%$ |
| $2^{\text {nd }}$ half | $29.3 \%$ | $6.3 \%$ |
| Arrow task | Completer | Non-completer |
| Geometric |  |  |
| $1^{\text {st }}$ half | $29.0 \%$ | $19.2 \%$ |
| $2^{\text {nd }}$ half | $26.8 \%$ | $27.1 \%$ |
| Transitional |  |  |
| $1^{\text {st }}$ half | $1.4 \%$ | $1.7 \%$ |
| $2^{\text {nd }}$ half | $6.1 \%$ | $1.5 \%$ |
| Irregular |  |  |
| $1^{\text {st }}$ half | $5.3 \%$ | $4.9 \%$ |
| $2^{\text {nd }}$ half | $5.3 \%$ | $3.7 \%$ |

$(12.1 \%$ vs. $5.2 \%), F(1,54)=4.1, p<.05$. The three-way interaction reached statistically significant, $F(2,36)=4.9, p$ $<.05$. Post hoc analysis revealed that in the second half the percentage of the irregular combination was higher in the completer than that in the non-completer ( $29.3 \%$ vs. $6.3 \%$ ), $F(1,54)=13.8, p<.01$, but in the first half the difference was not significant ( $3.1 \%$ vs. $8.3 \%$ ). Regarding the completers, the percentage of the irregular combination (29.3\%) was significantly higher than that of the geometric ( $2.3 \%$ ) and the transitional combination (13.1\%) in the second half, all $p s<.01$. This percentage of the completer in the second half was significantly increased than the first half $(3.1 \%), F(1,54)=29.8, p<.01$. In the arrow task, three-way ANOVA showed that the main effect of combination variable was significant, $F(2,36)=11.1, p<.01$. Multiple comparisons revealed that the percentage of the geometric combination averaged over time and completion variables ( $25.2 \%$ ) was significantly higher than that of the transitional ( $2.8 \%$ ) and the irregular combination ( $5.1 \%$ ), all $p \mathrm{~s}<.01$. The other main effect and the interactions did not reach statistical significance.

To investigate whether there is a characteristic attentional shift when problem-solvers reach the solution of Tangram, eye tracking data was recorded during the participants manipulated the 7 pieces. Areas of interests (AOIs) were surface of the each piece and the silhouette which was presented to the participants. When the participants looked
at the range of a single piece or at the silhouette for more than 33 while manipulating the pieces, this duration time was accumulated as a time spent looking at the AOI. The proportion of the time spent looking at the AOI per second during the total manipulation time was calculated for the each piece, and it summed over the 7 pieces for each participant. Table 3 shows the mean proportions of the time spent looking at the pieces and that looking at the silhouette, for the high-achievers who completed both the tasks and for the non-achievers who could not complete the both tasks. Statistical analysis revealed that the proportion of the time spent of the non-achievers looking at the pieces (283.1 $\mathrm{msc} / \mathrm{sec}$ ) was significantly longer than that of the highachievers $(137.4 \mathrm{msc} / \mathrm{sec}), F(1,13)=10.9, p<.01$. There was no difference between these two groups in the proportion of the time spent looking at the silhouette. The result that the time spent looking at the range of the pieces was relatively short might be reflected a wider or more active scanning over the problem space. In order to verify this inference more directly, eye movement distance during the participants manipulated the piece was calculated. The values in the bottom row of Table 3 indicate a mean eye movement distance per second in a visual angle for the high-achiever and for the non-achiever. Statistical test revealed that the eye movement distance of the highachiever was higher than that of the non-achiever but it was marginally significant, $F(1,13)=3.15, p<.10$.

Table 3: Mean proportion of the time spent ( $\mathrm{msc} / \mathrm{sec}$ ) and the eye movement distance (deg/sec)

|  | High-achiever | Non-achiever |
| :---: | :---: | :---: |
| Proportion of time |  |  |
| 7 pieces | $137.4(40.0)$ | $283.1(108.3)$ |
| Silhouette | $25.7(31.4)$ | $13.2(11.9)$ |
| Movement distance | $72.4(47.3)$ | $36.8(25.3)$ |

## Discussion

This research aimed to demonstrate a constraint relaxation which is followed by the transition to an appropriate representation in insight problem solving. For this purposes, Tangram was used as a new tool. To investigate the correspondence between the constraint relaxation and actual manipulation of the pieces of Tangram, combinations of the 2 largest triangle pieces which were the key to completing the two task silhouettes, a lion and an arrow, were analyzed. Nakano (2009) found that problem-solvers of Tangram tended to combine these 2 pieces into a geometric shape. This initial constraint would facilitate reaching an insight for the solution of the arrow task, because the correct configuration of this silhouette could be achieved by arranging the 2 pieces into either a square or a triangle. As expected, in this task both the completers and the non-
completers arranged the 2 pieces predominantly in a geometric combination through the problem solving. While in the lion task, this initial constraint to arrange the pieces into a geometric combination would inhibit the insight to occur, because the 2 pieces should be arranged in an irregular combination to complete the correct configuration of the lion task (top-right in Figure 2). Thus, the participants were expected to achieve better results in the arrow task than in the lion task. Contrary to this prediction, however, neither the percentage of the completers nor the time to completion of the high-achievers who were completed the both tasks were significantly different between the two tasks. This finding supported the view that the relaxation of the initial constraint was not the sole determinant of the insight for reaching the solution (Ormerod et al., 2002). As an evidence of this interpretation, in the lion task, the percentage of the cumulative duration time that the 2 pieces were arranged in the geometric combinations was decreased in the second half comparing with the first half. Importantly, such the decrease was found not only in the completers but also in the non-completers.

The critical difference between the completers and the non-completers was found in the lion task, in that, the completers arranged the 2 pieces in the irregular combination for a longer time in the second half than in the first half. This result for the completer should be a manifestation that the relaxation of the inappropriate initial constraint was followed by the construction of more appropriate representation. In contrast to this steady approach of the completers to reach the solution, the noncompleters could not distinguish which of the three types of combinations would lead them to the solution, even in the second half. Therefore, the critical determinant for reaching an insight to the solution was a clear cut differentiation between the appropriate representation and other alternatives.

The participants' eye movements were measured while they manipulated the pieces, in order to investigate specific feature of attention shift that facilitated the successful problem solving in Tangram. The analysis of the time spent looking at the surface of the pieces indicated the noteworthy difference between the high-achiever and the non-achiever that the latter had a longer time staying attention on the pieces. In contrast to this attentional feature of the nonachiever, the high-achiever had a slightly longer distance of eye movements during the manipulation of the pieces than the non-achiever. These findings suggested that the successful problem-solvers of Tangram tended to search the problem space of the pieces more widely or actively.

The second purpose of this research was to demonstrate the independence between the transition from the initial but inappropriate constraint to the more appropriate representation for the solution and the changing of the confidence to reach the solution. There was a remarkable difference between the results of the present experiment and the series of the researches by Metcalfe (Metcalfe, 1986; Metcalfe and Wiebe, 1987). In these previous researches,
most of the participants rated their confidence lowest level at the starting of the problem solving, and their selfevaluations stayed constant or slightly increased by the floor effect. While in the present experiment, the evaluation to the confidence was relatively high before starting the problem solving of Tangram, and it declined over time. Considering that the findings of Metcalfe was obtained by using insight problems including a spatial task and a linguistic task, Tangram is more likely to give the impression that it is easy to solve than those other insight problems, especially at the beginning of the problem solving. Another notable finding was that the participants who solved the insight problem had evaluated less confidence to reaching the solution (Metcalfe, 1986; Metcalfe and Wiebe, 1987). In other words, the actual achievement of each participant had opposite direction to their subjective confidence. Furthermore, the present experiment indicated that the completer manipulated the pieces so as to approach to the solution steadily but their confidence about it consistently declined. Therefore, this finding demonstrated that the process of deriving an insight and the subjective awareness to the problem solving did not progress independently but in the opposite direction over time.

## Acknowledgments

This research was supported by JSPS KAKENHI Grant-in-Aid for Challenging Exploratory Research (Grant Numbers 26540060).

## References

Jones, G. (2003). Testing two cognitive theories of insight. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 1017-1027.
Kaplan, C. A., \& Simon, H. A. (1990). In search of insight. Cognitive Psychology, 22, 374-419.
MacGregor, J. N., Ormerod, T. C., \& Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 176-201.
Metcalfe, J. (1986). Premonitions of Insight Predict Impending Error. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 623-634.
Metcalfe, J., \& Wieb, D. (1987). Intuition in insight and noninsight problem solving. Memory \& Cognition, 15, 238-246.
Nakano, Y. (2009). The insight problem solving in the puzzle game of "Tangram". Memoirs of Faculty of Education and Human Studies Akita University: Educational Sciences, 64, 65-72.
Ormerod, T. C., MacGregor, J. N., \& Chronicle, E. P. (2002). Dynamics and constraints in insight problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 791-799.
Thomas, L. E., \& Lleras, A. (2009). Covert shifts of attention function as an implicit aid to insight. Cognition, 111, 168-174.

# First Step is to Group Them: Task-Dynamic Model Validation for Human Multiagent Herding in a Less Constrained Task 

Patrick Nalepka (nalepkpk@mail.uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, ML 0376, 47 Corry Blvd, Edwards Center 1 Cincinnati, OH 45221 USA<br>Maurice Lamb (lambmi@ucmail.uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, ML 0376, 47 Corry Blvd, Edwards Center 1<br>Cincinnati, OH 45221 USA

Rachel W. Kallen (rachel.kallen@umail.uc.edu)<br>Center for Cognition, Action and Perception, Department of Psychology, ML 0376, 47 Corry Blvd, Edwards Center 1 Cincinnati, OH 45221 USA

Elliot Saltzman (esaltz@bu.edu)
Department of Physical Therapy and Athletic Training, Sargent College of Health and Rehabilitation Sciences, Boston University, MA, USA

Anthony Chemero (chemeray@ucmail.uc.edu)
Center for Cognition, Action and Perception, Department of Psychology, ML 0376, 47 Corry Blvd, Edwards Center 1 Cincinnati, OH 45221 USA

Michael J. Richardson (michael.richardson@mail.uc.edu)
Center for Cognition, Action and Perception, Department of Psychology, ML 0376, 47 Corry Blvd, Edwards Center 1 Cincinnati, OH 45221 USA


#### Abstract

Biological systems are capable of acting in a shared environment to produce emergent, self-organized behavior that is the result of the constraints imposed by local interactionssuch as bird flocking or ant swarming behavior. These examples present minimal demands for a shared-intention between co-actors, whereas other instances necessitate the formation of a shared goal. In these goal-directed tasks, how much of the observed complexity can be explained by the constraints imposed by both the environment and adherence to the shared task goal? This paper begins to investigate this question by presenting results from a two-person cooperative "shepherding" task first developed in Nalepka et al. (2017) but with fewer constraints. Results provide further evidence that the emergent behavior is the result of the constraints imposed by the task. The included task-dynamic model suggests a general model that can be used to understand multiagent herding behavior in a variety of contexts.


Keywords: joint action, collective herding, task-dynamic modeling

## Introduction

Emergent collective behavior in animal systems can oftentimes be understood by agents whose behaviors are constrained by local information. In non-human systems, such as ant trails, the observed behavior to a food source can be attributable to local interactions between ants and the strength of a deposited pheromone trail (Deneubourg et al., 1989). For humans, the route chosen to go to class in the
winter can be attributable to the paths carved in the snow by previous students (Goldstone \& Roberts, 2006). These examples don't necessitate the formation of a sharedintention (Searle, 1990) as these agents are exploiting their environment to reach their own individual aims.

However, human actors can engage in complex goaldirected behavior such as playing in team sports where the actors are working towards a common shared goal - a jointaction. Work discerning the neurocognitive mechanisms that support the timing and prediction of actions have been proposed to explain how human systems successfully accomplish joint-action tasks (Vesper, et al., 2011). Indeed, suboptimal coordination not only leads to sub-optimal performance, but can have a negative impact on one's selfesteem and one's opinion of a co-actor (Lumsden, Miles, \& Macrae, 2014). Similarly, suboptimal coordination during human-robot interaction (HRI) also leads to poorer performance and a depreciated user experience, with users often attributing poorer performance to a lack of predictability and reciprocal compensation on the part of the robot (Medina, Lorenz, \& Hirche, 2015).

How much of the complexity observed in cooperative action can be attributed to the constraints imposed by the environment, as well as the task goal? An approach to understand the behavioral dynamics that shape and constrain natural human performance is to argue that humans organize themselves as "special-purpose devices" to satisfy the dynamics of a particular task (Saltzman \& Kelso, 1987). For
example, in a reaching task, the body self-organizes so the hand becomes a damped mass-spring that moves towards, say, a mug (a fixed-point attractor). These low-dimensional models, in the case of reaching, can produce straight-line trajectories and deal with perturbations that may occur during the action.

To date, such task-dynamic models, expressed as ordinary differential equations, have been used to understand human path navigation and obstacle avoidance (Fajen \& Warren, 2003; Warren, 2006) and tested in robotic systems, such as a skiing robot (Lahajnar, Koss \& Nemec, 2009). In the joint action literature, virtual agents have also been created to perform oscillatory movements with a human partner (Zhai, et al., 2014; Kostrubiec, et al., 2015) with the movement dynamics of the virtual agent defined by a coupled nonlinear oscillator that produces patterns of coordination consistent with the Haken-Kelso-Bunz (HKB) model of rhythmic coordination (Haken, Kelso, \& Bunz, 1985); namely, stable, or intermittent in-phase $\left(0^{\circ}\right)$ and anti-phase $\left(180^{\circ}\right)$ modes of behavior. Perhaps most noteworthy is recent work by Zhai et al. (2014) and Kostrubiec et al. (2015), who demonstrated the ability for artificial agents incorporating nonlinear oscillatory models to coordinate with humans to reproduce the dynamics observed in human-human pairs, with the added benefit of enabling these nonhuman agents to steer humans to new coordinative modes that are unstable and difficult to master (like a $90^{\circ}$ phase relationship).

Recently, Nalepka et al. (2017) created a virtual shepherding task (Figure 1) to explore and model goaldirected behavior in a multiagent task to understand how stable social behavior emerges in more complex tasks with changing environments. The task required pairs to coordinate their movements in such a way as to corral and contain reactive autonomous spheres (referred to as sheep) to the center of a game field by controlling their player cube (referred to as their sheepdog) with a handheld motiontracking sensor. In the beginning, participants engaged in a behavior termed search and recover (S\&R) which involved moving one's controller towards the farthest sheep so that the sheep would be repelled towards the containment region. Using this strategy, some pairs could meet the success criteria for the task (defined as keeping all sheep within the containment region for a certain proportion of time (see Figure 1). However, a subset of successful pairs transitioned to a new behavioral mode termed coupled oscillatory containment (COC) that was functionally superior to $\mathrm{S} \& \mathrm{R}$. COC was defined by both participants performing oscillatory movements around the containment region to wall-in the sheep.

Interestingly, the COC behavioral mode exhibited similar dynamic stabilities as prototypical interpersonal or visual rhythmic coordination (Schmidt, Carello, \& Turvey, 1990; see Schmidt \& Richardson, 2008 for a review) described by the HKB model above, with pairs naturally exhibiting inphase and anti-phase patterns of COC behavior. Therefore, the shepherding task supplies a functional consequence for coupled rhythmic behavior that can be used to study
interpersonal coordination more generally. Videos illustrating the shepherding task are found at http://www.emadynamics.org/bi-agent-sheep-herdinggame/.


Figure 1: Depiction of task from Nalepka et al. (2017)

## The Shepherding Model

Nalepka et al. (2017) formulated a task-dynamic model of the human behavior observed in the shepherding task (also Richardson et al., 2016); they also successfully validated that this model, embodied in a virtual avatar, can complete the task successfully alongside a human partner (Nalepka et al., 2016). The model defines the task space in terms of a polar coordinate system (see Figure 2). The radial component (whose origin is the center of the containment region) of the system is defined using the following damped mass-spring equation,

$$
\begin{gather*}
\ddot{r}_{i}+b_{r i} \dot{r}_{i}+\varepsilon_{i}\left(r_{i}-\xi_{i}\left(r_{p s(t), i}+\Delta r_{\min , i}\right)\right. \\
\left.-\left(1-\xi_{i}\right)\left(r_{\min , p s(t), i}+\Delta r_{\min , i}\right)\right)=0 \tag{1}
\end{gather*}
$$

where $r_{i}, \dot{r}_{i}$, and $\ddot{r}_{i}$ are the radial position, velocity and acceleration of player $i(i=1,2) ; b_{r i}$ is the radial damping term, $r_{p s(t), i}$ is the radial coordinate of the player $i$ 's radially farthest sheep on their side of the field, $\left(r_{p s(t), i}+\Delta r_{\text {min,i }}\right)$ is the preferred radial target position that the player approaches for this farthest sheep, and $\varepsilon_{i}$ scales the strength of the centrally-directed radial force attracting player $i$ to the targeted sheep. This force is gated by $\xi_{i}$, a Heaviside parameter:

$$
\xi_{i}=\left\{\begin{array}{l}
0, r_{p s(t), i}<r_{\min , p s(t), i}  \tag{2}\\
1, r_{p s(t), i} \geq r_{\min , p s(t), i}
\end{array}\right.
$$

If the radial coordinate of at least one sheep is greater than or equal to $r_{\min , p s(t), i}$, then the player will select the furthest sheep, $r_{p s(t), i}$, and move to ( $r_{p s(t), i}+\Delta r_{\text {min, } i}$ ); otherwise, when $\xi_{i}=0$, the player will move towards $\left(r_{\text {min,ps }(t), i}+\right.$ $\left.\Delta r_{\text {min, } i}\right)$, their preferred distance from the center.

To be consistent with the previous research modeling the dynamics of rhythmic human interlimb and interpersonal coordination captured by the HKB model (Haken et al., 1985), the angular component of the players' movements (centered on the player's sagittal plane on their side of the
field) was modeled using the following modified set of coupled Rayleigh/van der Pol hybrid nonlinear oscillator equations,

$$
\begin{gather*}
\ddot{\theta}_{i}+b_{\theta i} \dot{\theta}_{i}+\beta_{i} \dot{\theta}_{i}^{3}+\gamma_{i} \theta_{i}^{2} \dot{\theta}_{i}+\omega_{i}^{2}\left(\theta_{i}-\xi_{i} \theta_{p s(t), i}\right) \\
=\left(1-\xi_{i}\right)\left(\dot{\theta}_{i}-\dot{\theta}_{j}\right)\left(A_{i}-B_{i}\left(\theta_{i}-\theta_{j}\right)^{2}\right), \tag{3}
\end{gather*}
$$

where $\theta_{i}, \dot{\theta}_{i}$ and $\ddot{\theta}_{i}$ are the angular position, velocity, and acceleration of player $i, . \omega_{i}$ is a player's natural angular oscillation frequency; $b_{\theta i}$ is the angular linear damping term; $\beta_{i} \dot{\theta}_{i}^{3}$ and $\gamma_{i} \theta_{i}^{2} \dot{\theta}_{i}$ are Rayleigh and Van der Pol escapement terms, respectively; and $A_{i}$ and $B_{i}$ are the parameters used in the HKB model to define the relative strength of in-phase and anti-phase coordination patterns. The parameter $b_{\theta}$ is governed by the equation,
$\dot{b}_{\theta i}+\delta_{i}\left(b_{\theta i}-\alpha_{i}\left(r_{p s(t), i}-\left(r_{\min , p s(t), i}+\Delta r_{m i n, i}\right)\right)\right)=0$,
where negative values of $b_{\theta i}$ produce oscillatory behavior, while positive values produce fixed-point behavior. Parameters $\delta_{i}$ and $\alpha_{i}$ govern the dynamics of $b_{\theta i}$ across its range of allowable values.


Figure 2: Depiction of model task space. Player $i=1$ is exhibiting oscillatory behavior, while player $i=2$ exhibits S\&R behavior. The smaller circles illustrate the sheep that must be kept within the containment region (larger circle).

The interplayer system modeled by Eq. $1,3 \& 4$ dictates the behavioral mode player $i$ produces. If $\xi_{i}=1$, the player is uncoupled from their partner $j$ (via the right half of Eq. 3), and moves towards the angular component of the furthest sheep; otherwise, the player will center their angular component to $0^{\circ}$ (here $b_{\theta i}$ will move towards a negative value and begin to produce oscillatory behavior). Parameter $\omega_{i}$ is the rate at which these angular destinations are reached. However, when $\xi_{i}=0$, the player becomes coupled to the
angular component of their partner's movement. This coupling function reproduces both in-phase $\left(0^{\circ}\right)$ and antiphase $\left(180^{\circ}\right)$ stable relative phase relationships, with the relative strength of these two coordination patterns defined by the parameters $A_{i}$ and $B_{i}$.

## The Current Project

The behavioral modes observed in Nalepka et al. (2017) are very like the behavioral modes found in real sheepdog shepherding (Strömbom et al., 2014). However, it is unclear whether the oscillatory behavior seen in Nalepka et al. (2017) emerged from the local interactions of both players and the sheep, or if it was due to participants attuning to pre-defined environmental features of the task. Trajectories observed in Nalepka et al. (2017) tend to trace the outer white circle of the containment region (see Figure 1). In the original experiment, this white circle indicated a failure criteria that ended a trial if all sheep managed to escape. However, anecdotally, some participants asked if they could enter the containment region, opening the possibility that participants perceived certain visually-marked locations to be appropriate (like the white region) and others not (such as the red containment area). Thus, the oscillatory behavior observed may have been a consequence of this perceived task constraint and the circular goal region.

The present work removed these visual landmarks and edited the task to have fewer constraints to test the generality of the developed task-dynamic model. Criteria that would cause a trial to fail prematurely were removed, with participants simply instructed to corral the sheep together, without a target goal region in mind. To keep scoring criteria similar to the original work, a containment region (invisible to participants) moved in accordance with the center of the herd, consistent with Strömbom et al. (2014) who suggested that sheepdogs corral sheep that are furthest from the center of the rest of the herd.

We tested a new set of naive participants to determine whether S\&R and, more importantly, COC behavior would still emerge. We also compared the participant performance to the performance of a slightly modified version of the virtual shepherding model presented above. In short, the model was modified so that the center of the task dynamic space that defined the $(0,0)$ point of the radial distance and polar task axes was dynamically tied to the herd's center of mass (COM), as opposed to being fixed in the center of game field $(0,0)$. At any time $t$, the herd's COM was calculated as the average sheep position in Cartesian coordinates and was subtracted from each game object's $(\mathrm{x}, \mathrm{y})$ positions.

## Method

## Participants

Thirty-eight participants ( $M$ age $=18.82,17-22$ ), recruited as 19 pairs completed the experiment. All participants were undergraduates from the University of Cincinnati and received course credit for participation. For model simulations, 10 artificial pairs were created with the
following parameter values: $b_{r}=10.9987, \varepsilon=98.70672$, $r_{\text {min }, p s(t)}=.062 \mathrm{~m}, \Delta r_{\text {min }}=.061539, \delta=23.08993, \alpha=$ 80.59288, $\beta=.161641, \gamma=7.22282, \omega=7.85, A=-.2$, and $B=.2$. The model was designed to perform COC behavior if all sheep on the player's side of the field was within $r_{m i n, p s(t)}$, as described above. Up to $\pm 1 \frac{r a d}{s^{2}}$ and $\pm 1 \frac{\mathrm{~cm}}{\mathrm{~s}^{2}}$ noise was randomly added to both $\ddot{\theta}_{P i}$ and $\ddot{r}_{P i}$ at a rate of 50 Hz .


Figure 3: Virtual experimental room with example initial sheep arrangement.

## Apparatus and Task

The task was designed using the Unity 3D game engine (version 5.2.1; Unity Technologies, San Francisco, California) and was presented to participants via Oculus Rift DK2 (VR) headsets (Oculus VR, Irvine, California). The virtual environment (Figure 3) was modeled at $1: 1$ scale after the experimental room. The task was presented in the VR headset to appear on a virtual tabletop modeled at $1: 1$ scale to the glass tabletop in the real environment, which acted as the solid physical surface on which participants could move their motion sensors. Participants used wireless Latus motion tracking sensors operating at 96 Hz (Polhemus Ltd, Vermont, USA). Participants moved the sensor along the glass tabletop, and hand movements translated $1: 1$ to movements of the player's cube (sheepdog) in the virtual environment. Participants were given a body in the virtual world, modeled after a crash test dummy of height 1.8 m whose motion was controlled using an inverse kinematic calculator (model and calculator supplied by Root Motion, Tartu, Estonia) based on the real movements of the participant's right hand (via the Latus motion sensor) and head (via the Oculus Rift).

Participants could move their sheepdogs anywhere in twodimensional space within the 1.5 by 0.8 m fenced area of the grass task field. The goal of the task was to jointly find a solution to corral seven wool-covered stimulus spheres (sheep) towards one another so that they turned to a red color. The sheep were programmed to turn red when all sheep were within 10.8 cm of the herd's COM. Note, if the fence was 10.8 cm from the herd center, the sheep did not turn red. This was done intentionally to keep participants from adopting a strategy which involved keeping the sheep cornered. On each trial, sheep appeared randomly within a .50 by .80 m boxed
area, randomly centered either on the center of the game field, or $\pm .50 \mathrm{~m}$ to either side. Forces from a random direction were applied to each sheep at a sample rate of 50 Hz , resulting in Brownian motion dynamics. If a sheep collided with the fence, a repulsive force was applied to move the sheep back towards the center. The sheep also dynamically reacted to the participant-controlled sheepdogs as if threatened, being repelled away from a participant's sheepdog when the sheepdog was within 12 cm of the sheep's game location. When threatened, the sheep would move directly away from the player at a speed proportional to the inverse of the squared distance between the sheep and the player. If the sheep were red for at least $70 \%$ of the last 45 seconds of a two-minute trial, the pair received a point. The experiment ended when the participants scored eight points, or after 45 minutes, whichever came first.

## Procedure

Following informed consent, participants stood on either side of the experimental table where they put on their respective virtual headsets and were given a motion sensor to hold in their right hand. Following calibration, participants were informed about the task goal and the conditions for success and failure. Participants were not told how near the sheep needed to be to turn red. Instead, they were instructed that if the sheep were not red, then they either needed to be closer together, or that the herd was too close to the fence. Participants were told that once the experiment began, they were not allowed to speak with one another until after debriefing. The experimenter remained in the room to enforce the no-talking rule and to answer any questions.

## Results

A preliminary review of participant behavior revealed that all pairs exhibited S\&R behavior and, more importantly, that several pairs discovered and exhibited the same type of COC strategy observed by Nalepka, et al. (2017). Here we present the COC classification criterion utilized, and the performance differences observed between COC trials (from the COCclassified pairs) and S\&R trials (from S\&R-classified pairs). The focus of the analysis presented here was to confirm that (a) COC is a robust emergent mode of behavior and (b) that COC behavior was superior to S\&R behavior. The results from the model simulations followed the same analyses and were employed to determine whether the task-dynamic model could effectively capture the dynamics observed in this less constrained task context. Classification and analyses were conducted on the last 45 seconds of each two-minute trial. This was set because participants were told that performance was measured during this time and that the first 75 seconds was to be used as time to corral the sheep and initiate resultant containment strategy. Because success was defined as keeping the sheep within 10.8 cm of their COM, all data were converted to polar coordinates with the center located at the herd COM. For the purposes of this paper, only successful trials were analyzed.

As defined by Nalepka et al. (2017), a trial was classified as COC if the peak angular oscillatory component was between 0.5 Hz and 2 Hz . More specifically, the classification criterion was as follows,

$$
\begin{equation*}
\varphi_{i, k}=\frac{\omega_{\text {freq }, i, k}-.5}{\left|\omega_{\text {freq }, i, k}-.5\right|} \omega_{\text {power }, i, k}, \tag{5}
\end{equation*}
$$

where $\omega_{\text {freq,i,k }}$ is the peak angular oscillatory component for player $i$, of pair $k$ and $\omega_{\text {power }, i, k}$ is its associated power. The average for both players, $\bar{\varphi}_{k}$, is taken and if the resultant average is positive, the trial is classified as predominately COC. Conversely, negative values resulted in a trial being classified as an S\&R trial. For analyses, the angular component of each dog was detrended and $z$-score normalized and submitted to MATLAB's pwelch function using a $50 \%$ overlap window of 512 samples.

Thirteen of the 19 pairs met the success criteria for the task. One of the remaining six pairs had five of the eight successful trials completed and were included in the analysis, while the remaining five did not have a single successful trial. Successful S\&R trials had an average classification value $\bar{\varphi}_{k}$ $=-0.47$ and successful COC trials had a value $\bar{\varphi}_{k}=0.23$ for human pairs, while it was $\bar{\varphi}_{k}=.22$ for the artificial pairs, which only exhibited successful COC trials. In total, nine pairs exhibited predominately $S \& R$ behavior to complete the task (with two pairs exhibiting one and three trials classified as COC), while the remaining five pairs completed the task with at least four trials classified as $\operatorname{COC}\left(M_{\text {\#trials }}=5.8\right.$ classified as COC). The artificial pairs completed the experiment in less time ( $M=16.8 \mathrm{~min}, S D=1.03$ ) than both $\mathrm{S} \& \mathrm{R}(M=29.75 \mathrm{~min}, S D=8.03)$ and $\mathrm{COC}(M=26.00 \mathrm{~min}$, $S D=6.32$ ) human pairs, $F(2,20)=12.75, p=.001, \eta^{2}=.56$.

The following four variables were examined to characterize performance differences between behavior modes: (1) containment time-the number of seconds the sheep were within 10.8 cm of the herd center while also 10.8 cm away from the nearest fence; (2) average sheep radial distance-the average distance from the herd center; (2) herd travel-the total distance travelled by the herd center; (4) and herd area - the area of the convex hull formed by the set of sheep positions. Only S\&R trials were considered for S\&R pairs and COC trials for both COC pairs and artificial pairs.

Performance differences were found for all variables: containment time, $F(2,21)=68.18, p<.001, \eta^{2}=.87$, average sheep radial distance, $F(2,21)=142.74, p<.001, \eta^{2}=.93$, herd travel, $\left.F(2,21)=140.46, p<.001, \eta^{2}=.93\right)$, and herd area $F(2,21)=11.61, p<.001, \eta^{2}=.53$. Figure 4 provides a summary of the findings. Performance by COC pairs on COC trials were found to be superior to performance by $S \& R$ pairs in all cases: containment time (COC $M=44.1 \mathrm{~s}, S D=1.01$; S\&R $M=39.91 \mathrm{~s}, S D=1.42[p<.001]$ ), average sheep radial distance (COC $M=2.73 \mathrm{~cm}, S D=.41 ; \mathrm{S} \& \mathrm{R} M=3.99 \mathrm{~cm}$, $S D=.20[p<.001])$, herd travel $(\operatorname{COC} M=62.74 \mathrm{~cm}, S D=$ 10.91; S\&R $M=106.90 \mathrm{~cm}, S D=19.16[p<.001])$ and herd area $\left(\mathrm{COC} M=23.87 \mathrm{~cm}^{2}, S D=13.80 ; \mathrm{S} \& \mathrm{R} M=55.66 \mathrm{~cm}^{2}\right.$, $S D=29.32[p=.02])$. Compared to COC human pairs, the
artificial pairs contained the sheep closer to the herd COM ( $M=2.32 \mathrm{~cm}, S D=.04,[p<.01])$ and had less herd travel ( $M=8.92 \mathrm{~cm}, S D=.44,[p<.001])$.


Figure 4: Result Summary Plots.

## Conclusion

Consistent with findings by Nalepka et al. (2016; 2017), the results provide further support that COC behavior is not a consequence of players tracing a visually salient boundary to contain the sheep, but rather emerges naturally from interactions between players within the shepherding task environment. Further, the task-dynamic model developed by Nalepka et al. (2017) can be minimally modified to function in this less constrained herding task space, by tracking the center of the herd. It is important to note that seven pairs completed the experiment without exhibiting any COC behavior. This may be due to the relaxed scoring criteria that allowed for stable S\&R behavior. Increasing task difficulty, like increasing the time needed to contain the sheep, is predicted to cause more pairs to transition to COC behavior, as it was associated with better task performance.

An approach to understand multiagent coordination is to treat human systems as self-organized "special-purpose devices" whose dynamics adhere to the constraints dictated by the task and environment (Saltzman \& Kelso, 1987; Richardson et al., 2016). Task-dynamic models that embody these constraints can be embedded in robot systems to produce human-like behavior. Because the presented model embodies the constraints inherent to herding autonomous agents, the model can be extended to include systems that can
work alongside humans in other herding-like tasks such as fire evacuation and environmental hazard containment. Similarly, these systems can potentially be used to steer novices to discover more optimal modes of behavior - in the shepherding task, but possibly in rehabilitation or educational contexts in the future.

Finally, the shepherding model is symmetrical, but many examples exist where distinct but complementary actions are needed to reach a collective goal - for example a basketball player performing a "pick" while their teammate breaks free to take a shot. Work has been done to incorporate the recent shepherding model by Strömbom et al. (2014) to two virtual sheepdogs, who each are responsible for either collecting or driving the herd (Watanabe and Fujioka, 2017). However, the sheepdogs' roles were rigidly defined and led to greater sheep dispersion compared to single dog performance who could adaptively switch between both modes. Models that allow multiple agents to switch between multiple behavioral modes without interference are still needed to develop systems that can work fluidly alongside humans of various skillsets.

## Acknowledgments

This research was supported by the National Institute of Health (R01-GM105045).

## References

Deneubourg, J. L., Goss, S., Franks, N., \& Pasteels, J. M. (1989). The blind leading the blind: Modeling chemically mediated army ant raid patterns. Journal of Insect Behavior, 2(5), 719-725.
Fajen, B. R. \& Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection. Journal of Experimental Psychology: Human Perception and Performance, 29(2), 343-362.
Goldstone, R. L., \& Roberts, M. E. (2006). Self-organized trail systems in groups of humans. Complexity, 11(6), 4350.

Haken, H., Kelso, J. S., \& Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. Biological Cybernetics, 51(5), 347-356.
Kostrubiec, V., Dumas, G., Zanone, P., \& Kelso, J. A. S. (2015). The virtual teacher (VT) paradigm: Learning new patterns of interpersonal coordination using the human dynamic clamp. PLoS One, 10(11): e014229. doi:10.1371/journal.pone. 0142029.
Lahajnar, L., Kos, A., \& Nemec, B. (2009). Skiing robot design, control, and navigation in unstructured environment. Robotica, 27, 567-577.
Lumsden, J., Miles, L. K., \& Macrae, C. N. (2014). Sync or sink? Interpersonal synchrony impacts self-esteem. Frontiers in Psychology, 5(1064).
Medina, J. R., Lorenz, T., \& Hirche, S. (2015). Synthesizing anticipatory haptic assistance considering human behavior uncertainty. IEEE Transactions on Robotics, 31(1), 180190.

Nalepka, P., Kallen, R. W., Chemero, A., Saltzman, E., \& Richardson, M. J. (2017). Herd those sheep: Emergent
multiagent coordination and behavioral mode switching. Psychological Science.
Nalepka, P., Lamb, M., Kallen, R. W., Shockley, K., Chemero, A., \& Richardson, M. J. (2016). A bio-inspired artificial agent to complete a herding task with novices. In C. Gershenson, T. Froese, J. M. Siqueiros, W. Aguilar, E. J. Izquierdo, \& H. Sayama (Eds.), In Proceedings of the Artificial Life Conference 2016. (pp. 656-663). MIT Press: Cambridge, MA.
Richardson, M. J., Kallen, R. W., Nalepka, P., Harrison, S. J., Lamb, M., Chemero, A., Saltzman, E., \& Schmidt, R. C. (2016). Modeling embedded interpersonal and multiagent coordination. In V. M. Muñoz, O. Gusikhin, \& V. Chang (Eds.), In Proceedings of the $1^{\text {st }}$ International Conference on Complex Information Systems. (pp. 155-164).
Saltzman, E. L., \& Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. Psychological Review, 94(1), 84106.

Schmidt, R. C., Carello, C., \& Turvey, M. T. (1990). Phase transition and critical fluctuations in the visual coordination of rhythmic movements between people. Journal of Experimental Psychology: Human Perception and Performance, 16(2), 227-247
Schmidt, R. C. \& Richardson, M. J. (2008). Dynamics of Interpersonal Coordination. In A. Fuchs \& V. Jirsa (Eds.), Coordination: Neural, Behavioral and Social Dynamics. (pp. 281-308). Heidelberg: Springer-Verlag.
Searle, J. R. (1990). Collective Intentions and Actions. In P. R. Cohen, J. Morgan, \& M. Pollack (Eds.), Intentions in Communication (pp. 402-415). MIT Press.
Strömbom, D., Mann, R. P., Wilson, A. M., Hailes, S., Morton, A. J., Sumpter, D. J. T., \& King, A. J. (2014). Solving the shepherding problem: Heuristics for herding autonomous, interacting agents. Journal of the Royal Society, Interface, 11(100).
Vesper, C., Butterfill, S., Knoblich, G., \& Sebanz, N. (2010). A minimal architecture for joint action. Neural Networks, 23(8-9), 998-1003.
Warren, W. H. (2006). The dynamics of perception and action. Psychological Review, 113(2), 358-389.
Watanabe, H., \& Fujioka, K. (2016). The effective flock control by two sheepdogs. In L. Barolli, F. Xhafa, \& K. Yim (Eds.), In Advances on Broad-band Wireless Computing, Communications and Applications: Proceedings of the $11^{\text {th }}$ International Conference on Broad-Band Wireless Computing, Communications and Applications (BWCAA-2016) November 5-7 2016, Korea (pp. 757-762). Springer International Publishing.
Zhai, C., Alderisio, F., Tsaneva-Atanasova, K., \& di Bernardo, M. (2014). A novel cognitive architecture for a human-like virtual player in the mirror game. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). (pp. 754-759).

# Conflicts Processing among Multiple Frames of Reference: An ERP Study 

Weizhi Nan (nanwz@psych.ac.cn) ${ }^{\text {a,b,c }}$<br>Yanlong Sun (ysun@medicine.tamhsc.edu) ${ }^{\text {c }}$ Hongbin Wang (hwang@medicine.tamhsc.edu) ${ }^{\text {c }}$ Qi Li (liqi@psych.ac.cn) ${ }^{\text {a,b }}$ Xun Liu (liux@psych.ac.cn) ${ }^{\text {a, }}{ }^{\text {b }}$<br>${ }^{\text {a }}$ CAS Key Laboratory of Behavioral Science, Institute of Psychology, 16 Lincui Road, Chaoyang District, Beijing 100101, China<br>${ }^{\mathrm{b}}$ Department of Psychology, University of Chinese Academy of Sciences, China<br>${ }^{\text {c }}$ Center for Biomedical Informatics, Texas A\&M University Health Science Center, 2121 West Holcombe Blvd, Houston, TX 77030, USA


#### Abstract

People rely on various frames of reference (FORs), such as egocentric (EFOR) and intrinsic (IFOR), to represent spatial information. The present study examined electroencephalogram profiles on a two-cannon task, which could regulate the conflict of IFOR-IFOR (red cannon, blue cannon) and IFOR-EFOR (target cannon, observer), to elucidate the brain mechanisms of FOR conflict processing by using event-related potentials (ERPs). Results showed that both of the conflicts occurred in the reaction time (RT) and there was an interaction between them. ERP results showed more negative amplitudes on $\mathrm{N} 2(276-326 \mathrm{~ms}$ ) and P3 (396726 ms ) for IFOR-IFOR conflict of the $180^{\circ}$ cannon angle condition and EFOR-IFOR conflict of the target cannon point-down condition. What's more, there was also an interaction between these two conflicts on the P3 amplitudes ( $561-726 \mathrm{~ms}$ ). In summary, our findings shed new light on the domain-specific conflict monitoring and domain-general executive control for the IFOR-IFOR and EFOR-IFOR conflicts.


Keywords: frame of reference; conflict monitoring; executive control; parallel process; N2; P3;

## Introduction

People adopt multiple frames of reference (FORs) to represent the spatial relationship of objects in a complex environment (Sun \& Wang, 2014). Based on the relationship with the observer, FORs can be classified into three types, egocentric FOR (EFOR), intrinsic FOR (IFOR) and allocentric FOR (AFOR) (Mou \& McNamara, 2002; Tamborello, Sun, \& Wang, 2012). An EFOR-based representation is anchored to the observer, which needs to be updated following the movement of the observer's eye, head, body coordinates (Wang, Johnson, \& Zhang, 2001). In an IFOR-based representation, an object or an object group in the viewing environment but exogenous to the observer is used as the reference point. For example, a car is used as an IFOR anchor in the description "the cat is in front of the car". IFORs remain stable with the observer's movement but have to be updated when the reference object moves. In an AFOR-based representation, the entire environment, such as a room or a city, is taken as the reference point. For a comprehensive review, see (Mou, Fan, McNamara, \& Owen, 2008; Mou \& McNamara, 2002; Sun \& Wang, 2010, 2014; Tamborello et al., 2012; Yamamoto \& Philbeck, 2013).
"Frame of Reference-based Map of Salience" theory (FORMS) states the human brain represents spatial information simultaneously using multiple FORs, all FORs consist of a FOR map of different salience, and human performance is determined by the interaction of all relevant FOR-based representations (Itti \& Koch, 2000; Sun \& Wang, 2010, 2014; Tamborello et al., 2012; Wang et al., 2001; Wang, Sun, Johnson, \& Yuan, 2005).

If different FORs generate different responses for one target, conflict may occur which requires cognitive control to solve it (Chen, Weidner, Weiss, Marshall, \& Fink, 2012; Nan, Li, Sun, Wang, \& Liu, in press; Sun \& Wang, 2014; Tamborello et al., 2012). According to the different kinds of FORs in the map (one EFOR, one AFOR, multiple IFORs), we could categorize the conflict of FORs as four types: EFOR-AFOR, EFOR-IFOR, AFOR-IFOR, and IFOR-IFOR. Plenty of studies has demonstrated that there exists the conflict of EFOR-AFOR (Chen et al., 2012; Conson, Mazzarella, Donnarumma, \& Trojano, 2012; Zhang et al., 2014), EFOR-IFOR (Wang et al., 2005). In addition, previous studies also showed the process of EFOR and AFOR were in parallel. EFOR has high salience and is almost processed automatically that needs little cognitive resource; EFOR is represented and processed in the dorsal visual stream subserving goal-directed actions. AFOR has low salience and needs more cognitive resource to process. AFOR is represented and processed in the ventral visual stream subserving the conscious perception of objects or spatial memory function (Goodale \& Milner, 1992; Zhang et al., 2014).

However, does it also exist a conflict of IFOR-IFOR and AFOR-IFOR? There were rare studies focusing on this question. If yes, how does our brain process and solve all these conflicts of different FORs (IFOR-IFOR, EFOR-IFOR, EFOR-AFOR, AFOR-IFOR)? What's more, in view of the limit cognitive resource, is the process of the multiple IFORs also in parallel as same as the process of EFOR and AFOR, or just in serial? If the process is in serial, only one IFOR could be represented and processed, so we could only observe the EFOR-IFOR conflict. Mou et al. $(2002,2008)$ found that people got higher accuracy for recalling spatial objects' locations represented by IFOR than that represented by EFOR, this means that people might prefer to use IFOR
to represent the environment, so IFOR might not need much cognitive resource. According to this, the process of multiple IFORs might be in parallel, different IFORs could be represented and processed, so we could observe the IFOR-IFOR conflict and the interaction among IFOR-IFOR conflict and the EFOR-IFOR conflict.

Following these questions and hypothesis, we developed a two-cannon task (see Figure 1) which could manipulate the EFOR-IFOR and IFOR-IFOR conflicts (Nan et al., in press; Tamborello et al., 2012). The EFOR-IFOR conflict was examined by the target cannon orientation (congruent condition: target cannon pointed-up, incongruent condition: target cannon pointed-down). The IFOR-IFOR conflict was examined by the cannon angle (congruent condition: $0^{\circ}$ cannon angle, incongruent condition: $180^{\circ}$ cannon angle). The behavioral studies' results showed that the IFOR-IFOR and EFOR-IFOR conflicts (RTs of the incongruent conditions were longer than that of the congruent conditions), and there was an interaction between these two conflicts. The cannon angle effect supported the hypothesis


Figure 1. A schematic illustration of the two-cannon task. At the beginning of each trial, a stimulus with two cannons (one blue and one red) and eight pellets (in either blue or red) was presented on the computer screen for 1000 ms , then the target would flash a yellow ring for 1000 ms , participants were asked to press two buttons in the keyboard to rotate the target cannon (with the same color of the target) to the target in the least distance, as quickly as possible. Cannon angle $\left(0^{\circ}, 180^{\circ}\right)$ was designed to test the conflict of IFOR-IFOR, target cannon orientation (target cannon point-up: the combination of target cannon points up-left, up, and up-right conditions; target cannon point-down: the combination of target cannon points down-left, down, and down-right conditions) was designed to test the conflict of EFOR-IFOR.
that there existed the IFOR-IFOR conflict. The target cannon orientation effect supported the hypothesis that there existed the EFOR-IFOR conflict. The interaction between two conflicts supported the hypothesis that the process of different IFORs was in parallel which the two conflicts would be interactive at the late response-selection stage. In summary, the behavioral results suggested that our brain might use a shared conflict processing mechanism for the IFOR-IFOR and EFOR-IFOR conflicts.

However, how does the conflict processing mechanism work at the neural level? Are they just process by the same conflict processing mechanism or by distinct conflict processing mechanisms? The event-related potential (ERP) has a high temporal resolution at the millisecond scale and could more directly reveal the brain activities of the cognitive process, so it is an excellent index to examine the neural mechanism of the FORs conflict processing (Luck, 2014). For the conflict processing, increasing electroencephalogram (EEG) evidence has demonstrated that the conflict-related N2 component which occurs approximately $250-350 \mathrm{~ms}$ after stimulus presentation is an effective indicator (Folstein \& Van Petten, 2008). The N2 amplitude is thought to index the degree of conflict, with its amplitude increasing as a function of conflict levels (Li et al., 2015). P3 was also typically reported to reflect ERP modulation of conflict process (Frühholz, Godde, Finke, \& Herrmann, 2011; Wang, Li, Zheng, Wang, \& Liu, 2014).

By applied the ERP to the two-cannon task, we expected to find the neural evidence of the conflict processing among different FORs (EFOR-IFOR, IFOR-IFOR), the parallel process of multiple IFORs and clarified the conflict processing among multiple FORs. Our expectation was that, for the behavioral results, we could replicate our previous behavioral studies' results (Nan et al., in press; Tamborello et al., 2012), which was that RT and error rates (ERR) were larger in the incongruent conditions of the EFOR-IFOR and IFOR-IFOR conflicts and there was also an interaction between them. For the ERP results, the N2 and P3 results could help to reveal the shared or distinct conflict processing mechanism of multiple FORs more clearly. We expected that the amplitude of N 2 and P3 would be more negative in the incongruent conditions of the two conflicts and there was also an interaction between them.

## Method

## Participants

Twenty undergraduate students (18-29 years old, average 22.8 years old, 6 women) participated in the present EEG experiment. All participants reported that they had no neurological or psychiatric history. All participants were right-handed and had normal or corrected-to-normal vision. Each participant voluntarily enrolled and signed an informed consent form prior to the experiments. This study was approved by the Institute of Psychology, Chinese Academy of Sciences.

## Procedures

Participants were seated comfortably in a dimly lit and sound-attenuating chamber approximately 80 cm away from a computer screen (resolution, $1024 \times 768$ pixels, vertical refresh rate, 75 Hz ). Stimulus presentation and manual response measurement were controlled by E-Prime 2.0 (Psychological Software Tools, Inc., Pittsburgh, PA, USA).

Each trial began with two cannons surrounding eight colored dots for 1000 ms . Then, the target pellet was marked by a yellow ring for 1000 ms . Participants were instructed to press a button on the keyboard (left-" $z$ " for counter-clockwise or right-"/" for clockwise), as quickly and accurately as possible, to rotate the target cannon (the one with the same color as the target pellet) to shoot the target in the least distance. After the target disappeared, a fixation cross was presented at the center of the screen for $1000-1300 \mathrm{~ms}$.

## EEG Recordings and Offline Processing

The EEG was recorded from 64 scalp sites using $\mathrm{Ag} / \mathrm{AgCl}$ electrodes arranged in an elastic cap according to an extension of the International 10-20 system (NeuroScan Inc., Herndon, VA). Vertical eye movements were recorded by two positioned above and below the left eye. The horizontal electrooculogram was recorded using lateral electrodes from both eyes. Impedances were below $5 \mathrm{k} \Omega$ for all recording sites. EEG signals were amplified using a NeuroScan SymAmps2 amplifier with a band-pass of 0.05 100 Hz and sampled with 500 Hz .

All scalp electrodes were referenced to the left mastoid online and were referenced to the average of the left and right mastoids offline. Each epoch started from 200 ms before the onset of the stimulus and lasted 800 ms , with the first 200 ms as the baseline. Trials with errors or trials that were contaminated with artifacts exceeding $\pm 100 \mu \mathrm{~V}$ were excluded from the analysis. The data were averaged for each condition and then digitally low-pass filtered at 30 Hz (24 dB/octave) with zero phase shift.

## Statistical Analysis

## Behavioral Data Analysis

Repeated-measures ANOVA and paired-sample t-test were performed on reaction times (RTs) of correct responses and error rates (ERs) and evaluated at $p<.05$. Trials with errors or with RT beyond three standard deviations were excluded from the RT analysis. A repeated-measures ANOVA was conducted for cannon angle effect and target cannon orientation effect (Figure. 2 B and Table 2), in which the $2 \times$ 2 factors tested were cannon angle ( $0^{\circ}, 180^{\circ}$ ) and target cannon orientation (up, down). Bonferroni correction was used for pair-wise comparisons.

## ERP Data Analysis

The ERPs of correct responses were averaged for each condition. The time window for N2 and P3 was identified
using the following protocol. First, we detected the peak latencies of all conditions at the midline electrodes ( $\mathrm{Fz}, \mathrm{FCz}$, $\mathrm{Cz}, \mathrm{CPz}$, and Pz ) and calculated the mean of these latencies for $\mathrm{N} 2(301 \mathrm{~ms})$ and P3 ( 561 ms ). For the N2 and P3 components, $50-\mathrm{ms}$ and $330-\mathrm{ms}$ time windows were centered on the mean peak latency, respectively. Therefore, the cannon angle effect and target cannon orientation effect were analyzed within $276-326 \mathrm{~ms}$ on N 2 mean amplitude and within 396-561 and 561-726 ms on P3 mean amplitude after stimulus onset.

Separated repeated-measures ANOVAs were performed on the mean N2 and the two time windows of mean P3, respectively. The factors cannon angle $\left(0^{\circ}\right.$ and $\left.180^{\circ}\right)$, target cannon orientation (up and down) and electrode ( $\mathrm{Fz}, \mathrm{FCz}$, $\mathrm{Cz}, \mathrm{CPz}$, and Pz ) were used to search for cannon angle effect and target cannon orientation effect. The significance level was set at $\alpha<.05$ for all ANOVAs. The mean number of trials retained for each condition are listed in the trial number of Table 1. A two-way analysis of variance (ANOVA) was calculated for the trial numbers of cannon angle and target cannon orientation, results showed there were no significant differences among them. The main effect of cannon angle: $F(1,19)=1.19, p>.05$, the main effect of target cannon orientation: $F(1,19)=1.17, p>.05$, the interaction between them: $F(1,19)=0.18, p>.05$. These analysis results eliminated the potential influence of different signal-noise ratios to statistical comparison.

## Results

## RTs and ERs

Regarding RTs (Figure 2 and Table 1), there was a significant main effect of target cannon orientation, $F(1,19)$ $=256.74, p<.001, \eta_{\mathrm{p}}^{2}=.93$, indicating that the RT in the target cannon point-down condition ( $697 \pm 24 \mathrm{~ms}$ ) was longer than that in the target cannon point-up condition (619 $\pm 22 \mathrm{~ms})$. There was a main effect of cannon angle, $F(1,19)$ $=50.43, p<.001, \eta_{\mathrm{p}}^{2}=.73$, indicating that the RT in the $0^{\circ}$ cannon angle condition ( $558 \pm 24 \mathrm{~ms}$ ) was shorter than that in the $180^{\circ}$ cannon angle condition ( $756 \pm 22 \mathrm{~ms}$ ). The interaction of the two factors was also significant, $F(1,19)=$ $6.48, p<.05, \eta_{\mathrm{p}}^{2}=.25$. Post-hoc analysis showed that target cannon orientation effect in the $180^{\circ}$ condition ( $89 \pm 13 \mathrm{~ms}$ ) was larger than target cannon orientation effect in the $0^{\circ}$ condition $(68 \pm 11 \mathrm{~ms}), t(1,19)=2.55, p<.05$.

ERs showed significant a main effect for target cannon orientation, $F(1,19)=22.54, p<.001, \eta_{\mathrm{p}}^{2}=.54$, indicating that the ER in the target cannon point-down condition (7.7 $\pm$ $1.1 \%$ ) was larger than that in the target cannon point-up condition ( $3.0 \pm 0.7 \%$ ). The main effect of cannon angle and the interaction of the two factors were not significant.

## N2 and P3

Regarding EFPs (see Figure 3), for N2, there was a significant main effect of target cannon orientation $F(1,19)$ $=4.88, p<.05, \eta_{p}^{2}=.20$, with more negative N 2 amplitudes

Table 1 RTs and ERs for target cannon orientation and cannon angle

| Target <br> cannon <br> orientation | Cannon angle |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\mathrm{RT}(\mathrm{ms})$ |  | $0^{\circ}$ | $180^{\circ}$ | $0^{\circ}$ | $180^{\circ}$ | $0^{\circ}$ |
| Down | $592 \pm 25$ | $802 \pm 24$ | $7.1 \pm 1.3$ | $8.2 \pm 1.1$ | $141 \pm 7$ | $147 \pm 5$ |  |
| Up | $524 \pm 23$ | $713 \pm 22$ | $2.5 \pm 0.8$ | $3.6 \pm 0.7$ | $144 \pm 7$ | $150 \pm 6$ |  |



Figure 2. RT of target cannon orientation effect and cannon angle effect. RT of the target cannon point-down condition was longer than that of the target cannon point-down condition; RT of the $180^{\circ}$ cannon angle condition was longer than that of $0^{\circ}$ cannon angle condition. The effect size of target cannon orientation effect in the $180^{\circ}$ cannon angle condition was larger than that in the $0^{\circ}$ cannon angle condition.
to the target cannon point-down condition $(0.27 \pm 0.76 \mu \mathrm{~V})$ than to the target cannon point-up condition $(0.83 \pm 0.76$ $\mu \mathrm{V})$. There was a marginally significant main effect of target cannon orientation, $F(1,19)=3.51, p=.076, \eta_{\mathrm{p}}{ }^{2}$ $=.16$, with more negative N 2 amplitudes to the $180^{\circ}$ cannon angle condition $(0.02 \pm 0.68 \mu \mathrm{~V})$ than to the $0^{\circ}$ cannon angle condition ( $1.07 \pm 0.791 \mu \mathrm{~V}$ ). There was a marginally significant interaction between cannon angle and electrode, $F(4,76)=3.27, p=.057, \eta_{\mathrm{p}}^{2}=.15$, post-hoc analysis showed that the cannon angle effect was significant at FCz and $\mathrm{Cz}, p \mathrm{~s}<.05$, revealed that the N 2 in the $180^{\circ}$ cannon angle condition (FCz: $-0.38 \pm 0.78 \mu \mathrm{~V}, \mathrm{Cz}:-0.18 \pm 0.81 \mu \mathrm{~V}$ ) was more negative than that in the $0^{\circ}$ cannon angle condition (FCz: $1.01 \pm 0.98 \mu \mathrm{~V}, \mathrm{Cz}: 1.21 \pm 1.07 \mu \mathrm{~V}$ ). There were no other significant effects obtained.

For the first time window of P3 (396-561 ms), there was a significant main effect of cannon angle, $F(1,19)=15.39, p$ $<.01, \eta_{\mathrm{p}}^{2}=.45$, with more positive P 3 amplitudes to the $0^{\circ}$ cannon angle condition $(2.07 \pm 0.81 \mu \mathrm{~V})$ than to the $180^{\circ}$ cannon angle condition $(-0.20 \pm 0.65 \mu \mathrm{~V})$. There was a significant main effect of target cannon orientation, $F(1,19)$ $=20.74, p<.001, \eta_{\mathrm{p}}^{2}=.52$, with more positive P3 amplitudes to the target cannon point-up condition (1.47 $\pm$ $0.69 \mu \mathrm{~V}$ ) than to the target cannon point-down condition $(0.40 \pm 0.69 \mu \mathrm{~V})$. There was a significant main effect of
electrode, $F(4,76)=24.01, p<.001, \eta_{\mathrm{p}}^{2}=.56$, with more positive P 3 amplitudes at $\mathrm{Cz}, \mathrm{CPz}$, and Pz compared with Fz and $\mathrm{FCz}(p s<.001)$. There was a significant interaction
between cannon angle and electrode, $F(4,76)=7.42, p<.01$, $\eta_{\mathrm{p}}^{2}=.28$, post-hoc analysis showed that the cannon angle effect was significant at five electrodes, $p \mathrm{~s}<.01$, and the largest difference was at $\mathrm{FCz}(3.04 \pm 0.74 \mu \mathrm{~V})$. There was a significant interaction between target cannon orientation and electrode, $F(4,76)=5.42, p<.01, \eta_{\mathrm{p}}^{2}=.22$, post-hoc analysis showed that the target cannon orientation effect was significant at five electrodes, $p \mathrm{~s}<.01$, and the largest difference was at $\mathrm{FCz}(1.36 \pm 0.30 \mu \mathrm{~V})$. No other significant effects were obtained.

For the second time window of P3 (561-726 ms), there was a significant main effect of cannon angle, $F(1,19)=$ 14.90, $p<.01, \eta_{\mathrm{p}}^{2}=.44$, with more positive P3 amplitudes to the $0^{\circ}$ cannon angle condition $(1.91 \pm 0.64 \mu \mathrm{~V})$ than to the $180^{\circ}$ cannon angle condition $(0.15 \pm 0.64 \mu \mathrm{~V})$. There was a significant main effect of target cannon orientation, $F(1,19)$ $=29.29, p<.001, \eta_{\mathrm{p}}^{2}=.61$, with more positive P3 amplitudes to the target cannon point-up condition (1.68 $\pm 0.59 \mu \mathrm{~V})$ than to the target cannon point-down condition $(0.38 \pm 0.62 \mu \mathrm{~V})$. There was a significant main effect of electrode, $F(4,76)=7.54, p<.01, \eta_{\mathrm{p}}^{2}=.28$, with more positive P 3 amplitudes at $\mathrm{Cz}, \mathrm{CPz}$ compared with Fz and $\mathrm{FCz}(p \mathrm{~s}<.05)$. There was a significant interaction between


Figure 3. Grand-average ERP results. A. N2 activity at FCz and P3 activity at CPz for cannon angle effect and target cannon orientation effect. B. the topography maps of the difference waveforms of cannon angle effect and target cannon orientation effect.


Figure 4. Three kinds of FOR conflict processing models. The three kinds of models are all parallel models. All of them showed the IFORs could be represented and processed in parallel that generates two conflicts of IFOR-IFOR and EFORIFOR. The difference is whether there are specific or shared conflict monitoring module and executive control module. The parallel model (1CM1EC) showed that there was only one conflict monitoring module (CM) and one executive control module (EC) for both conflicts. The parallel model (2CM1EC) showed that there were two conflict monitoring modules for each conflict and only one executive control module for both conflicts. The parallel model (2CM2EC) showed that there were two conflict monitoring modules and two executive control modules for each conflict.
cannon angle and electrode, $F(4,76)=5.96, p<.05, \eta_{\mathrm{p}}{ }^{2}$ $=.24$, post-hoc analysis showed that the cannon angle effect was significant at five electrodes, $p \mathrm{~s}<.01$, and the largest difference was at $\mathrm{FCz}(3.04 \pm 0.74 \mu \mathrm{~V})$. There was a significant interaction between target cannon orientation and electrode, $F(4,76)=16.52, p<.001, \eta_{p}^{2}=.47$, post-hoc analysis showed that the target cannon orientation effect was significant at five electrodes, $p \mathrm{~s}<.05$, and the largest difference was at $\mathrm{FCz}(2.53 \pm 0.60 \mu \mathrm{~V})$. Most interesting, there was a significant interaction between cannon angle and target cannon orientation, $F(1,19)=4.56, p<.05, \eta_{\mathrm{p}}{ }^{2}$ $=.19$, post-hoc analysis showed that target cannon orientation effect in the $180^{\circ}$ cannon angle condition (1.93 $\pm$ $0.40 \mu \mathrm{~V}$ ) increased and was significant compared to that in the $0^{\circ}$ cannon angle condition $(0.67 \pm 0.36 \mu \mathrm{~V})$. No other significant interactions were obtained.

## Discussion

Overall, the findings of the present study suggested that EFOR-IFOR and IFOR-IFOR conflicts had specific neural correlates and the process of IFORs was in parallel.

First, behavioral data showed the conflicts of EFORIFOR and IFOR-IFOR, the interaction between them. Second, the ERP results showed that the independent cannon angle effect and target cannon orientation effect on the N 2 amplitudes, from 276 to 326 ms , which indicated the independent conflict monitoring modules for the conflicts of EFOR-IFOR and IFOR-IFOR. What's more, the two effects interacted on the P3 amplitudes, from 561 to 726 ms , which indicated the shared executive control module for the two conflicts.

Yang, Nan, Li, and Liu (2015) used stimulus-response compatible tasks to collect behavioral and ERP data to support the 2CM1EC model (two domain-specific conflict monitoring modules and one domain-general executive control module for the conflicts of stimulus-stimulus and stimulus-response) of cognitive control for conflict processing, compared to the 1CM1EC model and 2CM2EC
model. For a comprehensive review, see (Li, Nan, Wang, \& Liu, 2014; Li et al., 2015; Liu, Nan, Wang, \& Li, 2013; Liu, Park, Gu, \& Fan, 2010; Wang et al., 2014; Yang et al., 2015).

On the spatial cognition area, we also could hypothesize three kinds of model for the conflict processing among different FORs (1CM1EC, 2CM1EC, and 2CM2EC, see Figure 4). 1CM1EC model shows there is only one general conflict monitoring module and one executive control for conflicts of EFOR-IFOR and IFOR-IFOR. The 2CM1EC model shows there are two specific conflict monitoring modules for conflicts of EFOR-IFOR and IFOR-IFOR, and one general executive control module for the two conflicts. The 2CM2EC model shows there are two specific conflict monitoring modules and two specific executive control modules for the two conflicts.

In our task, the behavioral results showed the cannon angle effect (conflict of IFOR-IFOR), target cannon orientation effect (conflict of EFOR-IFOR) and the interaction between them, which supported that the parallel process of the IFORs and the shared conflict process mechanism at the behavioral level. The ERP results showed the independent cannon angle effect and target cannon orientation effect on the N2 amplitudes, which further suggests there might be two specific conflict monitoring modules for the conflicts of EFOR-IFOR and IFOR-IFOR at the neural level. What's more, we also found the two effects, and the interaction between them on the P3 amplitudes. This suggests there might be only one shared executive control module for these conflicts at the neural level. In summary, our results supported the 2CM1EC model for the cognitive control of spatial conflict processing.

In the current two-cannon task, the AFOR is anchored on the computer screen which has the same direction (point-up) with the EFOR, so we could not separate the conflict of AFOR-IFOR and EFOR-IFOR. In the future, we could try to manipulate the AFOR and observe the interaction of all kinds of conflict which could be more comprehensive understand the spatial conflicts processing of FORs.

## Conclusion

Our task replicated the previous behavioral results well. What's more, the ERP results showed that common and distinct electrophysiological correlates for EFOR-IFOR and IFOR-IFOR conflict processing. On the one hand, EFORIFOR and IFOR-IFOR have domain-specific conflict monitoring modules, as revealed by the independent N 2 component. On the other hand, both of them share a domain-general executive control module, as revealed by the interaction of P3 component. The conflict of IFORIFOR and the interaction of EFOR-IFOR and IFOR-IFOR also suggest that the process of different IFORs is in parallel.'

## Acknowledgments

This work was partially supported by a Natural Science Foundation of China grant (31328013), an Air Force Office of Scientific Research (AFOSR) grant (FA9550-12-1-0457), and an Office of Naval Research (ONR) grant (N00014-16-1-2111).

## References

Chen, Q., Weidner, R., Weiss, P. H., Marshall, J. C., \& Fink, G. R. (2012). Neural interaction between spatial domain and spatial reference frame in parietaloccipital junction. Journal of Cognitive Neuroscience, 24(11), 2223-2236. doi:2210.1162/jocn_a_00260.
Conson, M., Mazzarella, E., Donnarumma, C., \& Trojano, L. (2012). Judging hand laterality from my or your point of view: interactions between motor imagery and visual perspective. Neuroscience Letters, 530(1), 35-40.
Folstein, J. R., \& Van Petten, C. (2008). Influence of cognitive control and mismatch on the N 2 component of the ERP: A review. Psychophysiology, 45(1), 152-170.
Frühholz, S., Godde, B., Finke, M., \& Herrmann, M. (2011). Spatio-temporal brain dynamics in a combined stimulus-stimulus and stimulus-response conflict task. Neuroimage, 54(1), 622-634.
Goodale, M. A., \& Milner, A. D. (1992). Separate visual pathways for perception and action. Trends in neurosciences, 15(1), 20-25.
Itti, L., \& Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res, 40(10-12), 1489-1506.
Li, Q., Nan, W., Wang, K., \& Liu, X. (2014). Independent Processing of Stimulus-Stimulus and StimulusResponse Conflicts. Plos One, 9(2), e89249.
Li, Q., Wang, K., Nan, W., Zheng, Y., Wu, H., Wang, H., \& Liu, X. (2015). Electrophysiological dynamics reveal distinct processing of stimulus-stimulus and stimulus-response conflicts. Psychophysiology, 52(4), 562-571.

Liu, X., Nan, W., Wang, K., \& Li, Q. (2013). Modular Organization of Cognitive Control. Advances in Psychological Science, 21(12), 2091-2102.
Liu, X., Park, Y., Gu, X., \& Fan, J. (2010). Dimensional overlap accounts for independence and integration of stimulus-response compatibility effects. Atten Percept Psychophys, 72(6), 1710-1720. doi:1710.3758/APP.1772.1716.1710.
Luck, S. J. (2014). An introduction to the event-related potential technique: MIT press.
Mou, W., Fan, Y., McNamara, T. P., \& Owen, C. B. (2008). Intrinsic frames of reference and egocentric viewpoints in scene recognition. Cognition, 106(2), 750-769.
Mou, W., \& McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. J Exp Psychol Learn Mem Cogn, 28(1), 162-170.
Nan, W., Li, Q., Sun, Y., Wang, H., \& Liu, X. (in press). Conflict Processing among Multiple Frames of Reference. PsyCh Journal.
Sun, Y., \& Wang, H. (2010). Perception of space by multiple intrinsic frames of reference. Plos One, 5(5), e10442.
Sun, Y., \& Wang, H. (2014). Insight into others’ minds: spatio-temporal representations by intrinsic frame of reference. Frontiers in Human Neuroscience, 8(58), 1-11. doi:10.3389/fnhum.2014.00058.
Tamborello, F. P., 2nd, Sun, Y., \& Wang, H. (2012). Spatial reasoning with multiple intrinsic frames of reference. Experimental Psychology, 59(1), 3-10.
Wang, H., Johnson, T., \& Zhang, J. (2001). The mind's views of space. Paper presented at the In Proceedings of the third international conference on cognitive science, Beijing, China.
Wang, H., Sun, Y., Johnson, T. R., \& Yuan, Y. (2005). Prioritized spatial updating in the intrinsic frame of reference. Spatial Cognition and Computation, 5(1), 89-113. doi:110.1207/s15427633scc15420501_15427634.
Wang, K., Li, Q., Zheng, Y., Wang, H., \& Liu, X. (2014). Temporal and spectral profiles of stimulusstimulus and stimulus-response conflict processing. Neuroimage, 89, 280-288.
Yamamoto, N., \& Philbeck, J. W. (2013). Intrinsic frames of reference in haptic spatial learning. Cognition, 129(2), 447-456.
Yang, G., Nan, W., Li, Q., \& Liu, X. (2015). Behavioral and electrophysiological profiles reveal domainspecific conflict processing. Advances in Computational Psychophysiology, 33-34.
Zhang, M., Tan, X., Shen, L., Wang, A., Geng, S., \& Chen, Q. (2014). Interaction between allocentric and egocentric reference frames in deaf and hearing populations. Neuropsychologia, 54, 68-76.

# Beyond Distributed Cognition: Towards a Taxonomy of Nonreductive Social Cognition 

Zachariah A. Neemeh (zachariah.neemeh@gmail.com)<br>Department of Philosophy, University of Central Florida<br>4000 Central Florida Blvd., Orlando, FL 32816 USA<br>Luis H. Favela (luis.favela@ucf.edu)<br>Department of Philosophy, Cognitive Sciences Program, University of Central Florida 4000 Central Florida Blvd., Orlando, FL 32816 USA


#### Abstract

Studies of social cognition often assume a reductionist, computational-representational conceptual framework. Distributed cognition is one of the few extant conceptual frameworks for a nonreductive understanding of social cognition. This concept's prototypical cases are exclusively of technical-scientific human institutions, including ships, cockpits, and the Hubble Space Telescope. In the first part of the paper, we outline the properties of distributed cognitive systems. We look at the case of wolf (Canis lupus) packs as an instance of distributed cognition in nonhuman systems. Nevertheless, a broad range of social cognitive phenomena across human and animal populations may not fit into this conceptual framework. We present a case study of bird flocks as a counterexample to distributed cognition. We propose "swarm intelligence" as an alternative concept of nonreductive social cognition. This is not to replace distributed cognition as a concept, but to add to and diversify the taxonomy of nonreductive social cognitive systems.


Keywords: social cognition; distributed cognition; swarm intelligence; bird flocks; wolf packs; nonreductive explanations

## Introduction

As a field of research, social cognition grew out of a social psychology that was influenced by the cognitive revolution and its adherence to the computational-representational understanding of mind (CRUM; Gilbert, 1999; Thagard, 2005). CRUM holds that cognition consists of a series of computations performed upon representations. CRUM approaches to social cognition are reductionist in that social phenomena are explained by recourse to the mental or neural mechanisms of individuals. In other areas of the cognitive sciences, nonreductive and systems approaches have been developed as alternatives to CRUM, e.g., ecological psychology (e.g., Gibson, 1979/2015), extended cognition (e.g., Clark \& Chalmers, 1998), embodied cognition (e.g., Rowlands, 2010), radical embodied cognitive science (e.g., Chemero, 2009), and radical embodied cognitive neuroscience (e.g., Favela, 2014).

The primary example of a nonreductive understanding of social cognition is Edwin Hutchins' distributed cognition. Hutchins (1995a) analyzed the cognitive structure of a navy ship's navigation across a network of agents (sailors) and navigational instruments. This study introduced an early
social cognitive concept that did not reduce explanations to events in the brain. In his "cognitive ethnography," the entire system of sailors and nautical instruments constitute a cognitive system. Navigation of the ship is achieved only through the combined efforts of these actors and tools.

While distributed cognition has provided a nonreductive lens by which to understand social cognition, it is rooted in a highly specific prototypical case of social and instrumental organization. We outline the properties of distributed cognitive systems and give an example of a nonhuman social system (wolf packs) in which distributed cognition is operating. We then argue that not all social cognitive systems are cases of distributed cognition. We argue that other forms of social cognition exist, such as swarm intelligence, via a case study of bird flocks.

## Distributed Cognition

Distributed cognition is a nonreductive account of social cognition that includes both agents and tools. Social cognition is not limited to mental events or brain activity in the individual agents who happen to constitute a social network. That is not to say that individual mental or neural events are irrelevant. However, the unit of analysis is the entire organization of agents and tools oriented around specific group tasks, such as a navy ship being navigated (Hutchins, 1995a) or an airplane cockpit being piloted (Hutchins, 1995b).

In Hutchins' studies of distributed cognition, the paradigm cases are of vehicles operated by two or more human agents. An airplane cockpit's cognitive organization consists of two agents (pilot and copilot) and an array of navigational instruments. An orthodox CRUM account of the social cognition of the cockpit would analyze the mental and neural events occurring in the brains of the two pilots. For Hutchins, however, the cognitive phenomena of perception and locomotion of the airplane as a whole occur as a coupled system of the pilots and their instruments.

Giere (2006) offers the Hubble Space Telescope (HST) as another prototypical case of distributed cognition. The HST is not merely a vehicle to be navigated or piloted, but a complicated instrument measuring ultraviolet, visible, and infrared spectra in deep space (Shayler \& Harland, 2016). Its operators are human, but it is orientated and programmed with commands and algorithms rather than piloted or
navigated. The HST itself consists of a vast array of complex, specialized instruments. Significantly, not only does it produce images of deep space objects, but it also provides higher-level scientific outputs such as "authenticated claims about the age of the universe" (Giere, 2006, pp. 712-713). In this respect, the HST is more cognitively complex than a ship or cockpit. The HST collectively produces these scientific outputs as a system of instruments, engineers, and scientists. No one instrument or operator is sufficient to produce any one of these outputs. For example, claims about the age of the universe cannot be substantiated without the HST's spectral analyses of galactic redshifts. Likewise, galactic redshifts cannot be measured without human programmers or the scientists who requested such measurements to, for example, confirm the Hubble constant.

Giere treats the HST as more than a deep space telescope. It is a scientific institution producing empirical claims about the universe. Nevertheless, like Hutchins' navy ship and airplane cockpit, it remains constituted by a network of agents and nonagentic instruments and tools. Kirsh (2006) provides a similar framework for distributed cognition, altering the methodological focus from a systems analysis to one of both individuals and systems, i.e., a "bottom-up topdown model" (p. 250). We provide a formalized list of the components of distributed cognitive systems consistent with Hutchins (1995a, 1995b), Giere (2006), and Kirsh (2006).

## Properties of Distributed Cognitive Systems

Distributed cognitive systems are not explained reductively. They are emergent in the sense that they are not merely the sum of the individual cognition of its components. Crucially, the actors are agentic (Giere, 2006). The sailors of the navy ship, pilots of the cockpit, and engineers and scientists of the HST exhibit significant degrees of agency.

These agents maintain their agency even as members of the system actively participate in its system-wide goals and joint tasks (cf. Amon \& Favela, 2017). Significantly, the cognitive behavior of a distributed cognitive system is not limited to perception and locomotion. Giere's HST system is not merely orientated towards celestial objects to capture images. The HST (the physical HST and its operators) produces falsifiable scientific claims. The following is modified from Amon and Favela (2017).
$S$ is a distributed cognitive system if:
D1. $S$ is emergent.
D2. There is continuous coordination of agents and nonagentic tools as members of $S$.
D3. Each agent maintains a degree of individual agency within $S$.
D4. Each agent actively participates in the overall goal or joint task in which $S$ is engaged.
D5. There is specialization of functions among members of $S$.
D6. The cognitive behavior of $S$ is complex and not limited to perception and locomotion.

In the following, we discuss wolf packs as an example of distributed cognition in the nonhuman animal world. We then present the case of bird flocks as a counterexample of social cognitive systems that are not distributed cognitive systems. We propose a new concept for nonreductive social systems for cases not in lieu of, but along with, distributed cognition.

## Wolf Packs as Distributed Cognitive Systems

Many paradigm cases of distributed cognition are anthropocentric and limited to human technical-scientific institutions (e.g., navy ship, airplane cockpit, and HST). However, social cognition is not limited to humans or such institutions. In some cases, collective animal systems may indeed be described by this anthropocentric concept. Wolf (Canis lupus) packs on the hunt ${ }^{1}$ are one such case of distributed cognitive systems in the animal world.

Wolves hunting in pack formation consist of four to 30 individual members, with hunting efficiency negatively correlated to increasing pack size (Mech, Smith, \& MacNulty, 2015). They are loosely organized around a breeder ("alpha" in older literature) but do not operate by a command structure. Individuals converge upon the prey and, assuming a successful hunt, a single wolf ultimately makes the kill (Tang, Fong, Yang, \& Deb, 2012).

Pack hunting patterns, organization, and coordination are emergent (D1). Several apparent hunting strategies have been noted, including encircling, ambushing, and relay hunting (Mech et al., 2015). Mech (2007) notes that wolves express a degree of mutual comprehension. He concludes from this that wolves communicate hunting strategies. This hypothesis assumes the existence of communication mechanisms that have yet to be discovered. Current evidence provides a more parsimonious account: The observed hunting patterns are wholly explicable in terms of a set of basic perceptual and locomotive procedures operating on the individual level and giving rise to a global structure. This simple process constrains the position of individual wolves so as to be neither too far away from nor too close to other wolves of the pack. Furthermore, individual wolves coordinate their own positions relative to both their prey and relative to the breeder (Muro, Escobedo, Spector, \& Coppinger, 2011). Global hunting patterns, such as encirclement, are not premeditated, directed, or otherwise centrally controlled. They emerge from these basic processes of local interaction (see Figure 1).

Within the pack, there is a continuous coordination of wolves (D2). Simulations by Muro and colleagues (2011) found that emergent hunting patterns can arise from the coordination of spatial positions in real time. Individual wolves coordinate their positions relative to those of both the breeder and the prey (see Figure 1). No advanced

[^421]communication of ideas or intentions is necessary to produce these patterns.


Figure 1: D1. Individual wolves continuously coordinate their movements based on the spatial positions of the breeder (top right) and prey (center). D2. This can result in emergent hunting patterns, such as encirclement (after Tang et al., 2012). Credit: Ahmed Labban.

Each wolf maintains a degree of individual agency within the pack (D3). As noted, the association of the pack is very loose. During the hunt, individual wolves often take individual initiatives, such as cutting off the prey (Mech et al., 2015). Each wolf actively participates in the overall goal or joint task in which the pack is engaged (D4). In this case, the focus is on hunting. The collective task of hunting may be the only activity around which the pack becomes cohesive enough to be considered a distributed cognitive system. Otherwise, a reductive account as is standard in accounts of social cognition may be appropriate.

There is a specialization of functions among the members of the pack (D5). The pack is hierarchical with the breeders at the top of the hierarchy (Mech et al., 2015). During the chase, the breeder serves as one of the two reference points for other wolves. Therefore, there are at least two functional differentiations among the wolves of the pack.

The cognitive behavior of the pack is complex and not limited to perception and locomotion (D6). Hunting is a goal-oriented process of gathering food. The wolf pack does not merely perceive and move towards the prey. It actively seeks to slay it in order to consume it. The hunt itself is a complex and demanding task and each individual actively engages in the task. Certainly this is far less complex of behavior than making falsifiable scientific claims about the age of the universe. Nevertheless, it is beyond the mere perception and locomotion that may characterize the simplest of cognitive systems (cf. Maturana \& Varela, 1980; Thompson, 2007).

## Non-Distributed Social Cognition in Flocks

The prototypical cases of distributed cognition are of very specific types of human institutions. Social cognition is not necessarily circumscribed to such specialized technicalscientific institutions or setups. The social cognition exhibited by wolf packs is appropriately treated as
distributed cognition. ${ }^{2}$ However, not all social systems, human or nonhuman animal, may be compatible with this concept. We present bird flocks as an example of nonreductive, but non-distributed, social cognition.

Flocks of birds vary in size across species and environments. They can range from less than a hundred to many hundreds of thousands of individuals. Quelea quelea flocks, for example, typically range from several to 500 birds but occasionally coalesce to form swarms of biblical proportions with comparable plague-like effects on agriculture (Crook, 1960). They migrate, evade predators, locate food and water, and navigate to roosts.

Flocks do exhibit some of the features of distributed cognition, but not all. Flocks are emergent (D1). Selfpropelled particle (SPP) models, as well as empirical vector analyses captured by computerized cameras, have for the first time made possible the study of cognition as a property of collective systems (Baglietto, Albano, \& Candia, 2013). SPP models show that global flock patterns and behavior can arise from a simple set of procedures governing the local interactions of individual birds (Bialek et al., 2012). In this respect, they are similar to wolf packs.

Bialek and colleagues (2012) model starling (Sturnus vulgaris) flocks by a set of procedures of alignment synchronization. Individual birds each align themselves with several proximal birds, henceforth coordinators (see Figure 2). The number of birds used for alignment synchronization is small, especially relative to the potentially enormous size of the flock itself. An increase in this parameter to a larger set of coordinators increases the entropy of the system, destabilizing the flock and breaking it apart into several smaller flocks (Castellana, Bialek, Cavagna, \& Giardina, 2016).

There is a continuous coordination of birds within the flock (D2). The procedures of alignment synchronization are simple and consist of successive zones of attraction, repulsion, and orientation (Couzin, 2008). If the coordinator is too far away, the bird moves towards it (or is "attracted" to it). If the coordinator is too close, the bird moves away from it (or is "repulsed" by it). If it is neither too close nor too far, it maintains its orientation (cf. Couzin, Krause, James, Ruxton, \& Franks, 2002). Kattas and colleagues (2012) find similar results in homing pigeons (Columba livia domestica) using a different method than Bialek and colleagues (2012), creating a model directly from recorded flight data. These local processes produce an emergent global order of flock movement (Cavagna, Giardina, \& Ginelli, 2013). The apparently wispy and erratic movements characteristic of flocks are effects of the inherent noise arising from imperfect alignments (Cavagna, Duarte Queirós, Giardina, Stefanini, \& Viale, 2013). This global

[^422]order in turn affects individual flight trajectories in local regions of birds.

Individual agency, to the degree to which it may exist to begin with, is not preserved in the flock. This violates D3. Individual birds within the flock, insofar as they constitute the flock, are not agentic in the same sense that a captain is free to abandon their ship. Nor do they express anything akin to the minimal agency of wolves in a pack. Individual wolves operate as a pack by a loose association and often act upon individual initiative (Mech et al., 2015). Birds in a flock, on the other hand, act predictably according to the basic processes of attraction, repulsion, and orientation. While flocks are noisy, the system's noise is constituted by the imperfect coordinating efforts of the birds.


Figure 2: Individual birds within a flock coordinate their movements relative to the positions of a small number of proximal birds. Credit: Ahmed Labban.

In violation of D 4 , each bird does not actively participate in the overall goal or joint task in which the flock is engaged. When a flock evades a hawk, for example, the entire flock does not necessarily perceive the predator. This is especially true of larger flocks of tens of thousands of members or more. To evade the raptor, it is sufficient that a local group within the larger flock perceive and react to it. This local reaction, manifest as a sudden shift in flight paths, creates a ripple effect in the flock as the other birds attempt to realign themselves. In this case, the massive flock is not engaged in the joint task of predator evasion. Rather, a local group is engaged in predator evasion while the majority of the remaining birds are merely continuing to implement the processes of attraction-repulsion-orientation. ${ }^{3}$ This situation contrasts markedly from that of wolves on the hunt, wherein each of the individual wolves perceives the prey and is actively engaged in hunting.

There is no significant specialization of functions among the members of the flock. Within the flock, each bird is more or less functionally isomorphic to the other birds. Hierarchies within flocks exist, but are fluid. "Leader" roles, such as directing migration routes (Mouritsen, 2003), are interchangeable and constantly shifting (Nagy, Ákos, Biro,

[^423]\& Vicsek, 2010). This contrasts with, for example, the captain of a ship or a pack breeder.

The behavior of the flock is limited to perception and locomotion, violating D6. Birds are individually complex and can perform a variety of functions besides moving and perceiving, such as fighting, mating, or raising offspring. Insofar as they constitute a flock, however, they are limited to basic procedures of attraction-repulsion-orientation. ${ }^{4}$ These require only that 1) the bird perceives the coordinators, and 2) the bird adjusts its flight accordingly. The flock as a system likewise only perceives and moves. In the example of predator evasion, the flock (but not necessarily each member thereof) perceives the hawk and changes its flight patterns to evade it. Flocks do not fight, mate, or raise offspring. They are defined only as a perceptual-locomotive social system.

Crook (1960) notes curious synchronized, wave-like movements during feeding and drinking in Quelea quelea. While drinking, the birds collectively alternate their positions, moving forward birds who have not yet drank and moving back birds who have. These "wave-like progression $[\mathrm{s}]$ " (p. 5) are broadly consistent with a system of basic attraction-repulsion-orientation processes. Nevertheless, SPP studies of avian populations have yet to go beyond an analysis of flight patterns and this is a particular research desideratum.

Of the six criteria of distributed cognition, flocks satisfy only two. The concept of "distributed cognition" is too limited to capture the manifold manifestations of social cognitive systems. New concepts are needed to understand complex systems such as bird flocks. We introduce "swarm intelligence" as a concept capable of describing social cognitive systems with characteristics like flocks.

## Swarm Intelligence as a Concept of Social Cognition

We appropriate the term 'swarm intelligence' from computing. In computing, swarm intelligence describes a set of optimization methods with emergent and self-organizing algorithms (Yang \& Karamanoglu, 2013) inspired by bees, ants, wolves, and other collectivist organisms (Beekman, Sword, \& Simpson, 2008). In our usage, "swarm intelligence" refers to a class of rudimentary and nonspecialized social cognitive systems. The prototypical case we offer is the bird flock, although it may also cover cases such as schools of fish, mosquito swarms, and human crowds.
$S$ is a swarm-intelligent system if:
S1. $S$ is emergent.
S2. There is a continuous coordination of individuals as members of $S$.
S3. Individual agency is minimal insofar as the individual constitutes $S$.

[^424]S4. The cognitive behavior of $S$ is limited to perception and locomotion.
S5. Communication or interaction between members of $S$ as they constitute $S$ is minimal.
S6. The organization of $S$ is relatively isomorphic, with no significant specialization of functions.
Flocks are emergent (S1). Overall flock movement in flight arises from the local processes of attraction-repulsionorientation. This global order in turn affects the local movements of individual birds. Comprehending why one bird happens to be caught up in a swirling arm of a murmuration requires a dynamical systems analysis of the entire flock. There is a continuous coordination of individuals as members of the flock (S2). Individual birds orientate themselves continuously in reference to several neighboring birds (see Figure 2).

Individual agency is minimal insofar as the individual constitutes the flock (S3). We noted this earlier as a violation of D3. Regardless of how agentic individual birds may be, they do not exhibit any significant agency within the collective. The cognitive behavior of the flock is limited to perception and locomotion (S4). Flocks are restricted to evasion, migration, and food-locating behaviors. For example, while ants construct elaborate nests as a form of collective shelter, most birds are only capable of creating roosts individually.

Communication or interaction between members of the flock insofar as they constitute the flock is minimal and there is no communication of intentions (S5). While birds have complex modes of communication (e.g. birdsong), they do not directly communicate qua members of a flock. The only information indirectly communicated is relative distance and position. The organization of the flock is relatively isomorphic. There are no significant specializations of functions among members of the flock (S6). Some transient local leaders might guide flight away from predators or towards food sources, but these positions are not enduring.

Overall, swarm intelligence is far simpler than distributed cognition. It does not necessitate any shared goals or joint tasks and its cognitive functions are limited to perception and locomotion. Nevertheless, it may describe systems as varied as bird flocks, mosquito swarms, schools of fish, and human crowds. Indeed, schools of fish and human crowds appear to operate by the same basic processes of attraction, alignment, and repulsion as do flocks of birds (Couzin, 2008; Moussaid et al., 2009).

## Conclusion: Towards a Taxonomy of Nonreductive Social Cognition

The field of social cognition remains dominated by reductionist and CRUM approaches. Since the mid-1990s, studies in distributed cognition have challenged this orthodoxy and demonstrated that social cognition can be understood from a nonreductive and systems perspective. Nevertheless, distributed cognition remains circumscribed in its applicability to different social cognitive systems. Its
paradigm cases are of human technical-scientific institutions. We provide formal criteria for the identification of distributed cognition consistent with Hutchins (1995a, 1995b), Giere (2006), and Kirsh (2006).

Distributed cognition remains useful for describing some systems that are nonhuman or non-technical-scientific, such as wolf packs. We demonstrate point-by-point how wolf packs are cases of distributed cognition. However, this concept is inappropriately applicable to many other social cognitive systems. We offer bird flocks as a counterexample to distributed cognition and propose "swarm intelligence" as an alternative concept of nonreductive social cognition. These two forms of social cognition are not proposed as absolute categories, but rather as relative points of difference between which gradations may exist.

The implications of this are significant. Beyond flocks, swarm intelligence is potentially found among schools of fish, mosquito and other flying insect swarms, and human crowds (cf. Moussaid et al., 2009). Swarm intelligence does not operate in lieu of distributed cognition. Nor is it the only alternative type of nonreductive social cognition. New concepts are needed to establish a more accurate taxonomy of nonreductive social cognition as diverse as the phenomena under investigation. Such a taxonomy allows the many and varied phenomena of social cognition to be recognized as such and studied through an appropriate theoretical lens (cf. O'Donnell et al. 2015 for a misattribution of distributed cognition in part due to such a lack of diversity).

As we have demonstrated, wolf packs or the HST do not operate by the same organizational principles as do bird flocks. This is not surprising given their radically different structures, functions, and components. What is far more surprising is that wolf packs operate, on an abstract level, analogously to certain technical-scientific institutions. Still other types of social cognitive systems await discovery. Ant colonies may not be well described by either distributed cognition or swarm intelligence. For example, unlike birds in a flock, ants have up to 12 modalities of communication (Hölldobler \& Wilson, 1990). Furthermore, they are able to engage in intricate collective projects such as nest building without a central planner. Ultimately, the project of creating a taxonomy of different social cognitive systems will serve to delineate and extend the outer bounds of the concept of "cognition."

## References

Amon, M. J., \& Favela, L. H. (2017). Human-dog system as example of interspecies distributed cognition. Unpublished manuscript.
Baglietto, G., Albano, E. V., \& Candia, J. (2013). Complex network structure of flocks in the Standard Vicsek Model. Journal of Statistical Physics, 153, 270-288.
Beekman, M., Sword, G. A., \& Simpson, S. J. (2008). Biological foundations of swarm intelligence. In C. Blum \& D. Merkle (Eds.), Swarm intelligence: Introduction and applications. Berlin, Germany: Springer.

Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., \& Walczak, A. M. (2012). Statistical mechanics for natural flocks of birds. Proceedings of the National Academy of Sciences, 109 (13), 4786-4791.
Castellana, M., Bialek, W., Cavagna, A., \& Giardina, I. (2016). Entropic effects in a nonequilibrium system: Flocks of birds. Physical Review, 93, 1-12.
Cavagna, A., Duarte Queirós, S. M., Giardina, I., Stefanini, F., \& Viale, M. (2013). Diffusion of individual birds in starling flocks. Proceedings of the Royal Society B, 280, 1-22.
Cavagna, A., Giardina, I., \& Ginelli, F. (2013). Boundary information inflow enhances correlation in flocking. Physical Review Letters, 110 (16), 1-5.
Chemero, A. (2009). Radical embodied cognitive science. Cambridge, MA: MIT Press.
Clark, A., \& Chalmers, D. (1998). The extended mind. Analysis, 58 (1), 7-19.
Couzin, I. D. (2008). Collective cognition in animal groups. Trends in Cognitive Sciences, 13 (1), 36-43.
Couzin, I. D., Krause, J., James, R., Ruxton, G. D., \& Franks, N. R. (2002). Collective memory and spatial sorting in animal groups. Journal of Theoretical Biology, 218, 1-11.
Crook, J. H. (1960). Studies on the social behavior of Quelea q. quelea (Linn.) in French West Africa. Behaviour, 16 (1/2), 1-55.
Favela, L. H. (2014). Radical embodied cognitive neuroscience: Addressing "grand challenges" of the mind sciences. Frontiers in Human Neuroscience, 8, 1-10.
Gibson, J. J. (1979/2015). The ecological approach to visual perception (Classic ed.). New York, NY: Psychology Press.
Giere, R. (2006). The role of agency in distributed cognitive systems. Philosophy of Science, 73, 710-719.
Gilbert, D. (1999). Social cognition. In R. A. Wilson \& F. C. Keil (Eds.), The MIT encyclopedia of the cognitive sciences. Cambridge, MA: MIT Press.
Hölldobler, B., \& Wilson, E. O. (1990). The ants. Cambridge, MA: Springer.
Hutchins, E. (1995a). Cognition in the wild. Cambridge, MA: MIT Press.
Hutchins, E. (1995b). How a cockpit remembers its speeds. Cognitive Science, 19, 265-288.
Kattas, G. D., Xu, X.-K., \& Small, M. (2012). Dynamical modeling of collective behavior from pigeon flight data: Flock cohesion and dispersion. PLOS Computational Biology, 8 (3), 1-15.
Kirsh, D. (2006). Distributed cognition: A methodological note. Pragmatics \& Cognition, 14 (2), 249-262.
Lazarus, J. (1979). Flock size and behaviour in captive redbilled weaverbirds (Quelea quelea): Implications for social facilitation and the functions of flocking. Behaviour, 71 (1/2), 127-145.
Maturana, H., \& Varela, F. (1980). Autopoiesis and cognition: The realization of the living. Dordrecht, Netherlands: D. Reidel Publishing Company.

Mech, L. D. (2007). Possible use of foresight, understanding, and planning by wolves hunting muskoxen. Arctic, 60 (2), 145-149.
Mech, L. D., Smith, D. W., \& MacNulty, D. R. (2015). Wolves on the hunt: The behavior of wolves hunting wild prey. Chicago, IL: University of Chicago Press.
Mouritsen, H. (2003). Spatiotemporal orientation strategies of long-distance migrants. In P. Berthold, E. Gwinner, \& E. Sonnenschein (Eds.), Avian migration. Berlin, Germany: Springer.
Moussaid, M., Garnier, S., Theraulaz, G., \& Helbing, D. (2009). Collective information processing and pattern formation in swarms, flocks, and crowds. Topics in Cognitive Science, 1, 469-497.
Muro, C., Escobedo, R., Spector, L., \& Coppinger, R. P. (2011). Wolf-pack (Canis lupus) hunting strategies emerge from simple rules in computational simulations. Behavioural Processes, 88, 192-197.
Nagy, M., Ákos, Z., Biro, D., \& Vicsek, T. (2010). Hierarchical group dynamics in pigeon flocks. Nature, 464, 890-893.
O'Donnell, S., Bulova, S., DeLeon, S., Khodak, P., Miller, S., \& Sulger, E. (2015). Distributed cognition and social brains: Reductions in mushroom body investment accompanied the origins of sociality in wasps (Hymenoptera: Vespidae). Proceedings of the Royal Society B, 282, 1-4.
Rowlands, M. (2010). The new science of the mind: From extended mind to embodied phenomenology. Cambridge, MA: MIT Press.
Shayler, D. J., \& Harland, D. M. (2016). The Hubble Space Telescope: From concept to success. New York, NY: Springer.
Tang, R., Fong, S., Yang, X.-S., \& Deb, S. (2012). Wolf search algorithm with ephemeral memory. In S. Fong, P. Pichappan, S. Mohammed, P. Hung, \& S. Asghar (Eds.), Seventh International Conference on Digital Information Management (ICDIM 2012). IEEE.
Thagard, P. (2005). Mind: Introduction to cognitive science (2nd ed.). Cambridge, MA: MIT Press.
Thompson, E. (2007). Mind in life: Biology, phenomenology, and the sciences of the mind. Cambridge, MA: Belknap Press.
Yang, X.-S., \& Karamanoglu, M. (2013). Swarm intelligence and bio-inspired computation: An overview. In X.-S. Yang, Z. Cui, R. Xiao, A. H. Gandomi, \& M. Karamanoglu (Eds.), Swarm intelligence and bio-inspired computation. London, United Kingdom: Elsevier.

# Modeling categorical perception with auditory neurons 

Chris Neufeld<br>University of Maryland


#### Abstract

It is well-known that the auditory perception of speech sounds is strongly influenced by the phonetic categories which divide up acoustic space. This paper approaches the problem of modeling categorical perception from the ground up, using a linear model of the tuning properties of auditory neurons - the spectro-temporal receptive field (STRF). An STRF which discriminates voiced from voiceless stops was derived from the TIMIT corpus, and two computer simulations were conducted to investigate its properties. In one simulation, this model neuron was found to exhibit a categorical response to a linear voice-onset-time continuum, closely tracking human behavior. In the second simulation, the STRF was found to exhibit a less categorical, more linear response to a stop-voicing continuum, also in line with human behavior. These two simulations show that perceptual responses to speech, whether non-linear or veridical, can be modeled by the action of auditory neurons.


# Brief Mindfulness Meditation Improves Attention in Novices 

Catherine J. Norris, Ph.D. (cnorris2@swarthmore.edu)<br>Department of Psychology, Swarthmore College, 500 College Avenue Swarthmore, PA 19081 USA<br>Dan Creem (dancreem@gmail.com)<br>Department of Psychology, Swarthmore College, 500 College Avenue Swarthmore, PA 19081 USA


#### Abstract

Past research has found that mindfulness meditation training improves executive attention and that this effect could be driven by more efficient allocation of resources on demanding attentional tasks, such as the Flanker Task. However, it is not clear whether these changes depend on long-term practice. We sought to investigate the effects of a brief, 10 -minute meditation session on attention in novice meditators, compared to a control activity. We also tested moderation by individual differences in Neuroticism. We found that participants randomly assigned to meditate for 10 minutes showed improved performance on incongruent trials on a Flanker task, with no detriment in reaction times, indicating better allocation of resources. Neuroticism moderated this effect, as only those low in Neuroticism showed improved allocation of attentional resources following meditation.


Keywords: mindfulness meditation; attentional network test; executive attention; neuroticism

## Introduction

Mindfulness is often defined as a two component process: (1) attention to present moment experience, coupled with (2) an attitude that is open, non-reactive, and accepting of things as they are (Bishop et al, 2004; Ludwig \& Kabat-Zinn, 2008). Over the past few decades a wealth of research has emerged in both popular science and academic journals on the benefits of mindfulness meditation for cognitive performance (e.g., creativity; Ding, Tang, Tang \& Posner, 2014; attention; Sedlmeier et al., 2012), mental health (Hofmann, Sawyer, Witt, \& Oh, 2010), negative mood (Goyal et al., 2014), treatment for addiction (Brewer et al 2011a; Bowen 2014), prejudice and discrimination (Lueke \& Gibson, 2016), and many other psychological processes. One premise in this area of research is that becoming mindful of an internal state or physiological function, such as one's breath or heartbeat, can hone abilities such as focused attention, working memory, and acceptance. In turn, this is thought to have long-term positive consequences when mindfulness is trained and practiced over an extended period of time.

Much of the past research has focused on the effects of mindfulness meditation training on attentional processes, including alerting, orienting, and executive attention. For example, Maclean and colleagues (2010) found that 3 months of intense meditative training can improve performance on tasks of perceptual discrimination and sustained visual attention. Elliott, Wallace and Giesbrecht
(2014) showed that a weeklong intensive meditation retreat can improve both executive attention and alerting (but not orienting). Jha, Krompinger and Baime (2007), however, found that both an 8 -week MBSR (i.e., mindfulness based stress reduction) course and a 1-month intensive MBSR retreat improved orienting but not executive attention. To the contrary, Tang and colleagues (2007) used a slightly less time-intensive approach and reported that 5 days of 20minute training sessions can improve executive attention and a more recent empirical article also found that mindfulness meditation increased executive attention (Ainsworth et al., 2013). In a review article comparing multiple forms of meditation, Lippelt, Hommel, \& Colzato (2014) conclude that there is good evidence to suggest that focused attention meditation (such as mindfulness) increases sustained attention (Carter et al., 2005; Brefczynski-Lewis et al., 2007). Notably, all of these studies have utilized extensive meditation training, including multiple hour-long sessions administered over an extended period of time.

Clearly, the literature on the effects of meditation on attention is both diverse and in some disarray, with different approaches to meditation (e.g., MBSR retreats vs. daily selfguided sessions) having divergent effects on different forms of attention. In addition, most of these studies have examined the effects of intensive, longitudinal meditative training composed of multiple sessions over weeks and/or years, and often recruit meditation-motivated or experienced individuals as participants.

Mindfulness meditation training takes many forms in the literature, but most often involves either an extended immersive experience (e.g., a 3-month retreat) or repeated daily practice, either in the form of a multi-week course or weeks (or months) of self-guided meditation. Indeed, the vast majority of published work has been on the effects of 8 weeks of training or longer (e.g., Hofmann, Sawyer, Witt, \& Oh, 2010, Brewer et al 2011b). And, although these studies have often documented beneficial outcomes of mindfulness meditation practice, the relevance of such time-consuming, extensive training is debatable for the average individual who might be either unmotivated or unable to dedicate the time and resources necessary to reap such benefits. This relates to the questions of "dose" - once someone begins to practice mindfulness, how soon can they expect to see beneficial effects (e.g., Tang et al., 2015; Zeidan, 2015)? A few recent studies have shown that 3-4 days of training are associated with some beneficial effects (e.g., Zeidan et al.,

2010a, 2010b). The current study focuses on meditationnaïve college students, to examine whether 10 minutes of meditation may have an immediate impact on attention. This approach represents an extreme test of the impact of mindfulness meditation on attention, but also may greatly expand our knowledge of the power of meditation, its boundary conditions, as well as its potential for practice in daily life. In the current studies we focus on whether a single brief audio-guided meditation can have similar benefits for attention in novice meditators.

In an initial attempt to examine the effects of a brief guided meditation on attention in novice meditators, we asked participants to listen to an audio tape (mindfulness meditation vs. control) and then subsequently complete a version of the Flanker task (Eriksen \& Eriksen, 1974; Eriksen, 1995), a measure of executive attentional control. Participants also completed the Big 5 Personality Dimension Inventory to allow for the investigation of moderation by individual differences in Neuroticism.

## Method

Participants. 40 (14 female) undergraduate students between the ages of 17 and $22(\mathrm{M}=19.48, \mathrm{SD}=1.18)$ were recruited from Swarthmore College. Three participants were omitted from final analyses because of scores greater than 3 SDs from the mean (i.e., outliers), leaving a final sample size of 37 ( 12 female; Mage $=19.51, \mathrm{SD}=1.19$ ). Participants were entered into a raffle for one of two \$25 prizes as compensation for completion of the study. All participants gave informed consent, and the study was conducted under the guidance of the Swarthmore College Institutional Review Board.

Procedure. Upon arriving at the laboratory, participants were told that the purpose of the study was to investigate the effects of audio attention on visual acuity. Each participant was seated in front of a desktop computer and was asked to wear headphones and a blindfold, to allow a focus on the audio recording. After the recording, the experimenter returned, removed the blindfold, and provided verbal instructions for the Flanker task. Participants completed twelve practice trials and were given the opportunity to ask questions before beginning the experimental trials. Following the Flanker task participants completed the Big 5 Personality Inventory (John, Donahue, \& Kentle, 1991) and a demographic survey. Finally, the experimenter and participant engaged in a face-to-face funneled debriefing interview.

Experimental Conditions. Participants were randomly assigned to either listen to a 10 -minute guided audio meditation tape (meditation) or a 10 -minute audio control tape (control). The mindfulness meditation tape was developed based on classic mindfulness instructions used in MBSR, and in consultation with several meditation teachers. This tape led participants through a breath-focused mindfulness exercise oriented towards beginners. It included
instructions such as "please set the intention to observe your experience with an accepting attitude," "please notice and begin to follow the natural and spontaneous movement of the breath, not trying to change it in any way," and "stay open and curious about your experience." The control tape was a reading of a National Geographic article about giant sequoias. Importantly, both recordings used the same speaker, speed of speech, number of words, and had similar word frequencies. In addition, both tapes began with instructions on posture within the first few seconds, and included pauses at approximately the same times and for similar durations, throughout.

Flanker Task. The flanker task was delivered using EPrime 2.0 software on a Dell computer with a 22 " LCD monitor (refresh rate $=60 \mathrm{~Hz}$ ). The flanker array consisted of white arrowheads on a black background and was 4.5 cm wide by 1.3 cm high. On average participants sat approximately 70 cm from the screen, producing a visual angle of the array width of 0.026 degrees and of the array height of 0.091 degrees. Each trial consisted of a 500 ms white fixation cross in the center of the black screen, followed by an array of five arrows, which remained on the center of the screen until a response was made. Participants pressed the " f " key with their left hand if the center arrow was facing left, and the " j " key with their right hand the center arrow was facing right. Flanking arrows were either facing in the same direction (i.e., congruent trials) or in the opposite direction (incongruent trials; Figure 1). There were 20 trials in each cell of the 2(direction: left, right) x 2(trial type: congruent, incongruent) design, resulting in a total of 80 trials, presented randomly. Participants were told to respond as quickly and accurately as possible. As soon as a response was made, the next trial began (i.e., there was no intertrial interval).

Big 5 Personality Inventory. After the flanker task, participants completed the Big 5 Personality Inventory (John, Donahue, \& Kentle, 1991), a self-report survey consisting of 44 items designed to measure five personality factors: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Participants indicated the degree to which they agreed or disagreed with each item on a 5-point Likert scale, with endpoints labeled disagree strongly (1) and agree strongly (5). Each item began with the phrase "I see myself as someone who..."; sample items for the Neuroticism subscale items include: "worries a lot" and "is emotionally stable, not easily upset," with the latter reverse-coded.

Demographic Survey. Participants also completed a standard demographic survey in which they reported their age, gender (male, female), and self-reported their race and ethnicity.

Debriefing. Finally, participants completed a funneled debriefing interview in which they were given the
opportunity to report any suspicion about the true purpose of the study, as well as reporting any previous experience with meditation, including duration and frequency of practice.

## Results

We conducted independent samples $t$-tests to examine any differences between groups of participants randomly assigned to listen to the meditation tape versus those randomly assigned to listen to the control tape on variables including: age, gender, race, Big 5 Personality traits, and meditation experience. There were no significant group differences on any of these measures.

Response times. RTs for correct trials only were subjected to a 2(condition: meditation, control) $x$ 2(trial type: congruent, incongruent) general linear model (GLM), with the first factor manipulated between-participants and the second factor manipulated within-participants (collapsing across arrow direction). The main effect of trial type, $F(1,35)=129.32, p<0.001$, indicated that participants were faster to respond on congruent trials $(M=427.05 \mathrm{~ms}, S E=$ 6.84) than on incongruent trials ( $M=466.26, S E=7.53$ ), a replication of past research. No other effects reached traditional levels of significance.

Accuracy. Rates of accurate responding were subjected to a similar 2(condition: meditation, control) x 2(trial type: congruent, incongruent) GLM. The main effect of trial type, $F(1,35)=35.123, p<0.001$, indicated that participants were more accurate on congruent ( $M=0.99, S E=0.003$ ) than on incongruent $(M=0.93, S E=0.01)$ trials, another replication of past research. The main effect of condition was marginally significant, $F(1,35)=3.10, p=.087$, and indicated that participants in the mediation condition were more accurate ( $M=0.97, S E=0.007$ ) than were participants in the control condition $(M=0.95, S E=0.007)$. Importantly, there was a significant interaction between trial type and condition, $F(1,35)=5.24, p=0.028$. Pairwise tests showed that whereas both groups of participants were more accurate on congruent than on incongruent trials ( $p \mathrm{~s}<.05$ ), participants in the meditation condition performed better on incongruent trials $(M=0.95, S E=0.01)$ than did those in the control condition $(M=0.91, S E=0.01), p=.044$ (Figure 2). Participants in the meditation condition $(M=$ $0.99, S E=0.004)$ and the control condition $(M=0.99, S E=$ 0.004 ) performed equally well on congruent trials, $p=.39$.

To further probe the effects of meditation on attention on the Flanker task, we calculated difference scores to capture the overall "Flanker effect" in correct RTs (incongruent - congruent) and accuracy (congruent incongruent), separately. Two independent samples $t$-test conducted on these difference scores showed no difference between meditation and control conditions in correct RTs, $p$ $=.48$, and a significant difference in accuracy, $t(35)=2.29$, $p=0.028$, such that participants in the meditation condition ( $M=0.04, S E=0.06$ ) exhibited a smaller Flanker effect
than those in the control condition $(M=0.08, S E=0.06)$. Thus, participants in the meditation condition showed a decreased Flanker effect in accuracy - reflecting better executive attentional control - as compared to those in the control condition, due to their improved performance on incongruent trials.

Moderation by Neuroticism. First, we conducted an independent samples $t$-test to examine whether random assignment to condition (i.e., meditation versus control tape) affected self-reported Neuroticism. As expected, Neuroticism did not differ between participants assigned to the meditation tape $(M=0.03, S D=0.97)$ and those assigned to the control tape $(M=-0.02, S D=1.06), t(35)=$ $0.14, p=.89$. Thus, Neuroticism scores can be assumed to represent true, stable individual differences rather than an unintended effect of random assignment to condition.

To examine moderation of the effects of meditation on attention by individual differences in Neuroticism, RTs were subjected to a 2(condition: meditation, control) $x$ 2(trial type: congruent, incongruent) x $z$-scored Neuroticism GLM, with the first factor manipulated betweenparticipants, the second factor manipulated withinparticipants, and Neuroticism entered as a continuous between-participants covariate. This analysis allows for the examination of main effects and the interaction between condition and trial time holding Neuroticism constant, as well as investigating the main effect of Neuroticism and its interactions with all other variables. As expected, this analysis merely revealed a main effect of trial type, $F(1,33)$ $=126.30, p<.001$, such that participants were faster on congruent than on incongruent trials even when controlling for neuroticism.

A similar GLM was conducted on accuracy scores. As expected, the main and interaction effects reported above held when controlling for neuroticism, including the condition x trial type interaction, indicating that individuals in the meditation condition were more accurate on incongruent trials than were those in the control condition. However, we also found a condition x trial type x Neuroticism interaction, $F(1,33)=3.72, p=.06$. To better understand this interaction, we examined accuracy estimates at 1 SD above and below the mean Neuroticism score. Those individuals lower in Neuroticism ( -1 SD ) generally exhibited the overall pattern: individuals in the control condition were more accurate to congruent than to incongruent trials ( $p<.001$ ), the two groups did not differ in their accuracies to congruent trials $(p=.30)$, but individuals in the meditation condition performed better on incongruent trials $(M=0.98, S E=.02)$ than did those in the control condition ( $M=0.91, S E=.02$; Figure 3 ). Indeed, meditation improved performance to such a degree that participants in this condition performed as well on incongruent trials ( $M=$ $0.98, S E=.02$ ) as they did on congruent trials $(M=0.99, S E$ $=.01$ ), $p=.78$. Individuals higher in Neuroticism ( +1 SD ), however, showed no effect of meditation: both groups were more accurate on congruent than on incongruent trials ( $p \mathrm{~s}<$
.005) and did not differ in accuracy on either trial type ( $p \mathrm{~s}>$ .75 ; Figure 3).

## Discussion

Results indicate that a brief 10 -minute guided meditation can improve executive attentional control even in naïve, inexperienced meditators. Perhaps more interesting is the fact that this meditation-induced improved performance was only observed for individuals relatively low in Neuroticism; individuals higher in Neuroticism did not exhibit any performance boost following meditation. Neuroticism, which is characterized by anxiety, high negative affect, and worry, may prevent individuals from reaping the benefits of meditation, as it may be difficult for individuals higher in Neuroticism to disregard their negative emotionality and focus on the early stages of mindfulness practice.

Previous research has focused on the reduction of neuroticism, anxiety, and stress due to meditation and less on personality predictors (e.g., Neuroticism) of response to meditation. For example, Williams, Francis, and Durham (1976) found that males who practiced transcendental meditation (i.e., a self-selected group) were more neurotic than the general population, but that they also became less neurotic over the course of a 6 -month period of study and that decreases in Neuroticism were directly associated with frequency of meditation. Similarly, Lane and colleagues (2007) found that Neuroticism moderated treatment effects in a group of individuals who completed training in meditation, such that individuals higher in Neuroticism at baseline showed greater decreases in negative mood, perceived stress, and anxiety over the course of training. Thus, Neuroticism appears to have a positive impact on more long-term consequences of meditation. Less is known about its ability to predict who will benefit from the practice of meditation.

In an early review of the literature, Delmonte (1985) argued that prospective meditators often report higher than average anxiety levels, and that anxiety predicts lower frequency of practice. Ironically, Delmonte (1985) also reported that meditation reliably decreases levels of anxiety over the course of practice. Furthermore, many studies have shown that regular meditators tend to be lower in trait Neuroticism (Leung \& Singhal, 2004). Thus, greater stress and anxiety may drive individuals to engage in meditation, while also negatively impacting the frequency of practice. If they, however, do persist in meditation, these individuals often show decreases in their anxiety and negative affect.

The current study is unique in that we recruited healthy undergraduate students who had not expressed any desire to practice meditation and simply measured Neuroticism; in this way, we did not bias our sample toward particular personality traits. Indeed, we did not tell participants that they were engaging in a guided meditation in order to minimize expectancy effects. And we controlled for frequency of practice by requiring that all participants simply complete one 10 -minute guided meditation in the
laboratory. Thus, our results suggest that trait Neuroticism may negatively impact the efficacy of short, guided meditation; individuals high in anxiety and self-awareness may not be able to relax and follow the instructions presented in guided audio meditation, thus preventing them from reaping the benefits of a meditation intervention. This finding has strong implications for the field, as it suggests that the very population thought to benefit most from meditation (i.e., individuals high in anxiety and Neuroticism) may have difficulty initially engaging in the practice. As meditation becomes more frequently prescribed as part of a holistic treatment for mental health disorders often associated with high Neuroticism, including depression, phobia, and other anxiety disorders, practitioners would benefit from an understanding of the difficulties individuals high in Neuroticism face in both learning and persisting in the practice of mindfulness meditation.

In sum, our results suggest that even in novices, one brief 10-minute audio-guided meditation improves attention. Specifically, on an attentional task performed under time pressure, participants in the meditation condition exhibited a boost in accuracy reflecting increased attentional control. Importantly, this effect was strongest for individuals lower in Neuroticism, indicating that personality may impact the ability to reap the benefits of brief meditation. Although much remains to be studied, the current paper suggests that brief meditation impacts attention even in novice practitioners and therefore has widespread implications.

Figures


Figure 1: Trial schema for the Flanker Task.


Figure 2: The condition $x$ trial type interaction in accuracies. Both groups were more accurate on congruent than on incongruent trials, but individuals in the meditation condition performed better on incongruent trials than did those in the control condition.


Figure 3: The interaction between condition, trial type, and Neuroticism. Meditation was effective in improving performance on incongruent trials for individuals lower in Neuroticism (-1 SD) but not for those higher in Neuroticism (+1 SD).

## Acknowledgments

We would like to thank all of the members of the Norris Social Neuroscience Laboratory at Swarthmore College for their help with data collection, analysis, and feedback. Nikhil Paladugu was instrumental in study design and data collection. Most importantly, we are extremely grateful to Hedy Kober, Ph.D., of Yale University for developing and sharing the meditation and control audio tapes used in both studies.

## References

Ainsworth, B., Eddershaw, R., Meron, D., Baldwin, D. S., \& Garner, M. (2013). The effect of focused attention and open monitoring meditation on attention network
function in healthy volunteers. Psychiatry Research, 210(3), 1226-1231.
Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., ... \& Devins, G. (2004). Mindfulness: A proposed operational definition. Clinical Psychology: Science and Practice, 11(3), 230241.

Bowen, S., Witkiewitz, K., Clifasefi, S. L., Grow, J., Chawla, N., Hsu, S. H., ... \& Larimer, M. E. (2014). Relative efficacy of mindfulness-based relapse prevention, standard relapse prevention, and treatment as usual for substance use disorders: a randomized clinical trial. JAMA Psychiatry, 71(5), 547-556.
Brefczynski-Lewis, J. A., Lutz, A., Schaefer, H. S., Levinson, D. B., \& Davidson, R. J. (2007). Neural correlates of attentional expertise in long-term meditation practitioners. Proceedings of the National Academy of Sciences, 104(27), 11483-11488.
Brewer, J. A., Mallik, S., Babuscio, T. A., Nich, C., Johnson, H. E., Deleone, C. M., ... \& Carroll, K. M. (2011). Mindfulness training for smoking cessation: results from a randomized controlled trial. Drug and Alcohol Dependence, 119(1), 72-80.
Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y. Y., Weber, J., \& Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. Proceedings of the National Academy of Sciences, 108(50), 20254-20259.
Carter, O. L., Presti, D. E., Callistemon, C., Ungerer, Y., Liu, G. B., \& Pettigrew, J. D. (2005). Meditation alters perceptual rivalry in Tibetan Buddhist monks. Current Biology, $15(11)$, R412-R413.
Delmonte, M. M. (1985). Meditation and anxiety reduction: A literature review. Clinical Psychology Review, 5(2), 91-102.
Ding, X., Tang, Y. Y., Tang, R., \& Posner, M. I. (2014). Improving creativity performance by short-term meditation. Behavioral and Brain Functions, 10(1), 1-8.
Elliott, J. C., Wallace, B. A., \& Giesbrecht, B. (2014). A week-long meditation retreat decouples behavioral measures of the alerting and executive attention networks. Frontiers in Human Neuroscience, 8, 1-9.
Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. Visual Cognition, 2(2-3), 101-118.
Eriksen, B. A., \& Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception \& Psychophysics, 16(1), 143-149.
Goyal, M., Singh, S., Sibinga, E. M., Gould, N. F., Rowland-Seymour, A., Sharma, R., ... \& Ranasinghe, P. D. (2014). Meditation programs for psychological stress and well-being: A systematic review and metaanalysis. JAMA Internal Medicine, 174(3), 357-368.
Hofmann, S. G., Sawyer, A. T., Witt, A. A., \& Oh, D. (2010). The effect of mindfulness-based therapy on anxiety and depression: A meta-analytic review.

Journal of Consulting and Clinical Psychology, 78(2), 169-183.
Jha, A. P., Krompinger, J., \& Baime, M. J. (2007). Mindfulness training modifies subsystems of attention. Cognitive, Affective, \& Behavioral Neuroscience, 7(2), 109-119.
John, O. P., Donahue, E. M., \& Kentle, R. L. (1991). The Big Five Inventory--Versions 4a and 54. Berkeley, CA: University of California,Berkeley, Institute of Personality and Social Research.
Lane, J. D., Seskevich, J. E., \& Pieper, C. F. (2007). Brief meditation training can improve perceived stress and negative mood. Alternative Therapies in Health and Medicine, 13(1), 38-44.
Leung, Y., \& Singhal, A. (2004). An examination of the relationship between qigong meditation and personality. Social Behavior and Personality: An International Journal, 32(4), 313-320.
Lippelt, D. P., Hommel, B., \& Colzato, L. S. (2014). Focused attention, open monitoring and loving kindness meditation: effects on attention, conflict monitoring, and creativity-A review. Frontiers in Psychology, 5, 15.

Ludwig, D. S., \& Kabat-Zinn, J. (2008). Mindfulness in medicine. Journal of the American Medical Association, 300(11), 1350-1352.
Lueke, A., \& Gibson, B. (2016). Brief mindfulness meditation reduces discrimination. Psychology of Consciousness: Theory, Research, and Practice, 3(1), 34-44.
MacLean, K. A., Ferrer, E., Aichele, S. R., Bridwell, D. A., Zanesco, A. P., Jacobs, T. L., ... \& Wallace, B. A. (2010). Intensive meditation training improves perceptual discrimination and sustained attention. Psychological Science, 21(6), 829-839.
Sedlmeier, P., Eberth, J., Schwarz, M., Zimmermann, D., Haarig, F., Jaeger, S., \& Kunze, S. (2012). The psychological effects of meditation: A meta-analysis. Psychological Bulletin, 138(6), 1139-1171.
Tang, Y. Y., Hölzel, B. K., \& Posner, M. I. (2015). The neuroscience of mindfulness meditation. Nature Reviews Neuroscience, 16(4), 213-225.
Tang, Y. Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., ... \& Posner, M. I. (2007). Short-term meditation training improves attention and self-regulation. Proceedings of the National Academy of Sciences, 104(43), 1715217156.

Williams, P., Francis, A., \& Durham, R. (1976). Personality and meditation. Perceptual and Motor Skills, 43(3), 787-792.
Zeidan, F. (2015). The neurobiology of mindfulness meditation. In K. W. Brown, J. D. Creswell, \& R. M. Ryan (Eds), Handbook of Mindfulness: Theory, Research, and Practice (pp. 171-189). New York, NY: Guilford.
Zeidan, F., Gordon, N. S., Merchant, J., \& Goolkasian, P. (2010). The effects of brief mindfulness meditation
training on experimentally induced pain. The Journal of Pain, 11(3), 199-209.
Zeidan, F., Johnson, S. K., Diamond, B. J., David, Z., \& Goolkasian, P. (2010). Mindfulness meditation improves cognition: Evidence of brief mental training. Consciousness and Cognition, 19(2), 597-605.

# Who makes use of prior knowledge in a curriculum on proportional reasoning? 

Daniela Nussbaumer (daniela.nussbaumer@hfh.ch)<br>University of Applied Sciences of Special Needs Education, Schaffhauserstrasse 239, 8050 Zurich, Switzerland<br>Christian Thurn (christian.thurn@uni-konstanz.de)<br>Institute for Behavioral Sciences, ETH Zurich, Universitätsstrasse 41, 8092 Zurich, Switzerland<br>Ralph Schumacher (schumacher @ifv.gess.ethz.ch)<br>Institute for Behavioral Sciences, ETH Zurich, Universitätsstrasse 41, 8092 Zurich, Switzerland<br>Elsbeth Stern (stern@ifv.gess.ethz.ch)<br>Institute for Behavioral Sciences, ETH Zurich, Universitätsstrasse 41, 8092 Zurich, Switzerland


#### Abstract

Understanding proportions is a time-intensive process that does not come cheap during late childhood and early adolescence. It is fostered by learning experiences in which students have opportunities to explore, discuss and experiment with situations involving proportions. Children must undergo many informal learning opportunities before they can gain from direct instruction on proportional reasoning. In this study, we aimed to determine whether physics curricula focusing on the concept of density prepares students for learning from a curriculum on proportional reasoning. A $2 \times 2$ design with the factors "physics curricula" (with, without) and "concept used to introduce proportional reasoning" (speed, density) was applied to 253 children from 12 classrooms at the beginning of grade 5 . We expected the "density, with physics curriculum" group to outperform the other three groups. However, only the students who scored in the highest quartile on an intelligence measure gained from the prior knowledge they had acquired through the physics curricula. The results show that curricula on proportional reasoning are worthwhile for all students in early adolescence. However, more capable students can boost their proportional reasoning if they have the chance to acquire prior knowledge through a physics curriculum.


Keywords: proportional reasoning, prior knowledge, STEM

## Theoretical Background

Proportional reasoning involves comparing ratios within or between quantities, and it is based on the formula $\mathrm{a} / \mathrm{b}=\mathrm{c} / \mathrm{d}$. The crucial step is understanding the multiplicative relationship between the quantities, which means knowing that increasing "a" by a certain factor requires either multiplying "c" or "b" with the same factor or dividing "d" by this factor. Most elementary school children erroneously compute differences rather than ratios. Multiplicative proportional reasoning strategies are considered a cornerstone in the cognitive development of adolescents because they are prerequisites for learning more advanced mathematics and for understanding scientific concepts in various formal domains. Moreover, proportional reasoning supports decision making in everyday life, such as when cooking or when evaluating sales.

As stated before, understanding proportions is a timeintensive process during late childhood and early
adolescence. It emerges through repeated and varied experiences and enables mathematical terms and the associated ideas to become connected. The understanding of proportions is fostered by learning experiences in which students have opportunities to explore, discuss and experiment with situations involving proportions. Children must undergo many informal learning opportunities before they can gain from direct instruction on proportional reasoning.

The period from late childhood to adolescence is one of great change, not only in executive control and emotional regulation but also in cognitive competencies. Mastering science and mathematics competencies requires these cognitive tools and skills, which are expected to emerge during elementary school. For example, topics such as fractions, decimals, or ratios, which are at the focus of secondary school mathematics education, presuppose proportional reasoning abilities. These abilities emerge from extending the number concept beyond simple counting (Siegler \& Lortie-Forgues, 2014). Elementary school children's competencies in solving mathematical problems addressing relations and proportions of numbers are highly predictive for secondary school performance, even more so than general cognitive abilities (Stern, 2009; Siegler et al., 2012).

Broadly applicable formal reasoning skills and learning strategies (e.g., proportional, logical and scientific reasoning and metacognitive knowledge) result from an interaction between brain maturation and education. Additionally, proportional reasoning, which is considered a domaingeneral competence, emerges from an interaction of cognitive development (stimulated by brain maturation) and exposure to learning opportunities (Ben-Chaim, Fey, Fitzgerald, Benedetto, \& Miller, 1998). These broadly applicable competencies can be fostered through direct instruction, but most children also acquire them incidentally by abstracting knowledge acquired during elementary school. However, children vary significantly in the ease with which they acquire these skills; these differences are attributed to person characteristics such as intelligence and learning opportunities. Earlier theories of cognitive development focused on universal maturation processes and
assumed that all children reach a formal reasoning stage around puberty (Case, 1993). However, the considerable individual differences found within age groups with regard to formal reasoning tasks demonstrate the importance of domain-specific knowledge. Several studies have detected remarkable individual differences in proportional reasoning: While some eight-year-olds already master multiplicative strategies, some 15 -year-olds still struggle, and adults with little or no standard schooling may never master these skills (Lawson, 1985). Thus, proportional understanding would not suddenly appear if there were no formally or informally acquired knowledge available upon which to build. For example, playing board games in preschool facilitates number-line understanding in elementary school (Siegler, \& Ramani, 2008), and the number-line competencies of elementary school children predict later proportional reasoning and understanding of fractions (Siegler, Thompson, \& Schneider, 2011). These and other longitudinal intervention studies with preschool and elementary school children have identified which learning opportunities support children in developing proportional reasoning competencies.

Research has rarely examined how knowledge about proportional reasoning is represented in a broader network. Represented as a domain-general principle, it should be transferable to isomorphic problems in various contexts. This kind of transferable knowledge is difficult to acquire and requires intensive instruction (Bransford \& Schwartz, 1999). In particular, guided inquiry stands out as an effective means to train the transfer of domain-general principles across situations and time (Chen \& Klahr, 1999, 2008). Similarly, inquiry-based science and math learning has been shown to be a successful means for developing domain-specific content knowledge throughout preschool (Leuchter, Saalbach, \& Hardy, 2014), elementary school (Hardy, Jonen, Möller, \& Stern, 2006), and secondary school.

A longitudinal focus in researching such a complex concept as proportional reasoning would be optimal. However, studies often concentrate on short-term interventions to identify the learner characteristics and instructional factors that affect learning outcomes. These findings do not necessarily capture a generic understanding of proportional reasoning that can be transferred to superficially different but structurally isomorphic problems. Thus, unless learners have acquired expertise, they rarely develop representations of abstract formal structures such as domain-general proportional reasoning (Chi \& VanLehn, 2012).

Embedding a general principle such as proportional reasoning in various contexts can support learners in developing an abstract understanding of general principles that can flexibly be used in novel contexts and situations (Alfieri, Nokes-Malach, \& Schunn, 2013; Gentner, 2010).

## The Current Study

The current study builds on the Swiss MINT Study. (MINT is the acronym for Mathematics, Informatics, Natural Science, and Technology.) In this longitudinal study, elementary school teachers were trained in implementing physics curricula developed by a team of science education experts (https://verlage.westermanngruppe.de/spectra/reihe/ KINTBOX). The inquiry-based curricula included four different basic physics topics: floating \& sinking, air \& atmospheric pressure, sound \& spreading of sound, and stability of bridges. Classes started in third and fourth grade. The curricula were tailored to develop children's domainspecific conceptual content knowledge on these four topics (Möller, \& Jonen, 2005).

In every curriculum, children engaged frequently in experimentation to explore the different basic physics concepts. This inquiry-based approach was accompanied by a strong emphasis on instructional guidance and teacher-led classroom discussion. Teachers who agreed to participate in the study underwent four half-day trainings conducted in small groups. In total, the four curricula encompassed 60 classroom lessons. The floating \& sinking curriculum, for instance, introduced the concepts of water displacement and object density over 15 lessons.

The children were engaged in extensive guided experimentation activities within and across the four curricula, and they encountered many examples of proportional reasoning. In the swimming and floating curriculum, for example, they immersed pieces of different materials with a similar size or similar materials with different sizes into water to examine how these two characteristics influence floating ability. Through this inquiry-based process, children learned about the concept of density. Therefore, through this curriculum, elementary school children gained not only content knowledge but also experience with regard to the domain-general concept of proportions. None of the four curricula involved general, direct remarks about proportional reasoning.

In this study, we want to find out whether children who studied physics curricula that implicitly included proportional concepts better comprehend proportional reasoning in a subsequent teaching unit than those who studied the traditional way. We expected that the manifold guided experimentation activities not only fostered children's domain-specific content knowledge but also helped them understand abstract mathematical concepts such as (the domain-general principle of) proportional reasoning. Thus, we expected a significant main effect of "physics curricula".

We wanted to distinguish between a more general and a more specific effect of prior physics curricula. Therefore, two curricula on proportional reasoning were developed:
one based on the concept of speed, which was not part of the physics curriculum, and one based on the concept of density, which was central in the unit on floating and sinking. The curricula on proportional reasoning were applied either to classes that were part of the previously described Swiss MINT study (and therefore had undergone the physics curricula) or classes that underwent regular science education (which usually does not include physics at all). Thus, we expected a significant interaction between "physics curricula" and "concept used to introduce proportional reasoning".

This led us to the following research questions and hypotheses:

1. Does early science learning affect later mathematical learning (for proportional reasoning)?
We assumed that later math learning is affected positively (but that this effect depends on the problem context chosen for the intervention; see next point). In other words, we assumed that students with early science learning understand proportions better after receiving instruction on proportional reasoning.
2. In what way does early science learning prepare students for future learning? Does it work more generally or more specifically?
For the familiar problem context of density, we expected a greater advantage; for the non-familiar problem context of speed, we expect only marginal group differences. We predicted that students who underwent the physics curricula were able to link the new information to their knowledge about physics.
In short, we predicted that children who underwent the physics curricula and were taught proportional reasoning with density scored highest on a transfer test on proportional reasoning.

## Method

## Participants

Participants included 253 children from 12 classrooms at the beginning of grade 5 (age: $M=10.73$ years, $S D=0.55$ ). Participants per cell of the $2 \times 2$-design are as follows (see below): density/without physics curricula: $n=66$, density/with physics curricula: $n=62$, speed/without physics curricula: $n=66$, speed/with physics curricula: $n=$ 59).

The children in our sample came from different regions of Switzerland. All of them were part of the Swiss MINT Study and either had completed all four physics themes with the aforementioned curricula prior to the curriculum on proportional reasoning or were part of a waiting group that had not yet started with the physics curricula. Whole schools rather than individual teachers volunteered to be a part of the Swiss MINT Study (and it was not the students
who chose a particular school, educational track or curriculum). Nevertheless, it can be assumed that teachers (and school teams) taking part in the current study were STEM oriented and that there were no differences between the "waiting group" stage and the "applying the curricula" stage. Attempts were made to minimize the differences between the student populations by parallelizing the catchment areas of schools at the "waiting group" stage and the "applying the curricula" stage (rural, agglomeration, city and average socioeconomic status of a particular area), as in Switzerland, students are assigned to schools according to their place of residence. Teachers were recruited through a mailing list, and the teachers and classes participated voluntarily during their class time. They received no monetary compensation.

We chose fifth-grade classes for the study as this is a time in which the development of the understanding of proportional reasoning increases, and only towards the end of fifth grade (that is, at the end of our intervention) is proportional reasoning an explicit part of the official study curriculum. Thus, we were able to test whether and to what extent proportional reasoning can develop without formal instruction and to what extent physics experimentation experience additionally boosts this development.

## Procedure

A $2 \times 2$ design with the factors "physics curricula" (with, without) and "concept used to introduce proportional reasoning" (speed, density) was applied. The curriculum on proportional reasoning (both speed and density) consisted of 3 lessons ( 45 minutes each) that were based on the idea of concreteness fading (Goldstone \& Son, 2005). In the speed group, children were faced with two cars that traveled the same distance in different times, while in the density group, children were shown cubes of the same size but different weights. Afterwards, the children were faced with different combinations of time/distance and weight/volume. The dependent variable was a test on proportional reasoning that was applied at the end of the curriculum (subsequently called the transfer test, as this test consisted of untrained word problems). To control for differential effects, a measure of general intelligence was applied.

## Material

Intervention: The intervention was designed in a way that is scientifically proven to be most effective. With the intervention, we followed some basic principles: We tried to make students focus on the underlying mathematical structure of a problem and on multiple solution and representation strategies. We promoted the use of external representations in learning and calculating proportions. We explicitly compared and contrasted the different solutions and representations (see Ziegler and Stern, 2014). Furthermore, we tried to implement self-explanations (Jitendra et al., 2008 on schema-based instruction). To accomplish this, we combined direct instruction with phases
of working alone (vs. in pairs). Furthermore, the lessons were based on the idea of concreteness fading (Goldstone \& Son, 2005).
Problem presentation influences task difficulty and often determines whether a student can solve a problem (see Boyer et al., 2008). Our participants therefore solved and received feedback on two different problem types: comparison problems and missing values problems (see Van Dooren, De Bock, Hessels, Janssens, \& Verschaffel, 2005). The two problem contexts were tightly parallelized, and the same values were used in both of them.
Pre-test, post-test and transfer test (see table 1): In a pretest, prior knowledge about the two problem contexts was assessed, i.e., students' knowledge on density and speed. Additionally, the N 2 subscale of a cognitive ability test (kognitiver Fähigkeitstest, KFT; Heller \& Perleth, 2000) was administered at pre-testing. After the above-described intervention (proportional reasoning introduced in the context of density vs. speed), a post-test on the understanding of proportions (very similar to the pre-testthe same in structure as the tasks solved during the intervention) was administered. Additionally, a transfer test (proportion problems embedded into word problems) had to be solved (see table 1). All participants of the $2 \times 2$ design solved the same test versions. Both tests (post and transfer) took place one to two days after the intervention.
Table 1: Schematic view of the experimental design

| Pre-test | $\begin{aligned} & \text { g } \\ & \frac{0}{0} \\ & 0 \end{aligned}$ | Post-test and transfer test (one to two days after the intervention) |
| :---: | :---: | :---: |
| - Physics knowledge on density <br> - Prior understanding of speed <br> - Cognitive ability test (kognitiver Fähigkeitstest, KFT; Heller \& Perleth, 2000), subscale N 2 |  | - Post-test: knowledge on proportional reasoning (in the same problem context as during the intervention) <br> - Transfer test: knowledge on proportionality in new problem contexts |

## Results

## Pre-Test

Keeping in mind our research question, "Who makes use of prior knowledge in a curriculum on proportional reasoning?" it was important to check whether prior knowledge was actually still available. Indeed, this was the case: The physics knowledge test consisting of the themes floating \& sinking revealed significantly higher knowledge in the group with prior experience with the physics curricula (and no significant difference between the speed vs. density conditions). For the group with prior knowledge, the test can be considered a long-term follow-up from the physics curricula; the students completed the physics curricula up to
two years prior to the actual proportional reasoning curriculum. For the group without prior experience with the physics curricula, the test and the themes were new. For the group with prior physics knowledge, the solution rate was $40 \%$, whereas the solution rate of the group without prior experience was slightly over $20 \%$. Therefore, for the problem context of density, we can build on the differences in knowledge between the groups with and without physics curriculum experience.
For further analyses, we formed subgroups; i.e., we grouped participants into quartiles according to the results on the N2 subscale of the cognitive ability test. The four groups of the 2x2-design did not significantly differ in their level of cognitive abilities, $\mathrm{F}(3,242)=0.33, \mathrm{p}=0.8$. Additionally, when looking at each quartile separately, we find no significant difference between the four groups of the $2 \times 2$ design. Therefore, the distribution of cognitive abilities and quartile groups is comparable between the four groups of the $2 \times 2$ design.
When they had no prior experience with the physics curricula, participants of all quartiles scored similarly low, and their scores closely overlapped. Cognitive abilities were not reflected in the results regarding physics understanding when physics had not yet been formally taught. However, when the physics curricula had been applied, intelligence differences unfolded, with the highest quartile scoring significantly higher ( $47 \%$, right) than the other quartiles (see figure 1 and note that error bars indicate standard errors of the mean). Thus, the cognitive ability test did its share only in the group with prior physics instruction.


Figure 1: Results of the physics test on floating and sinking. Left: group with prior experience with the physics curricula (for this group, the test can be considered a long-term, i.e., up to one year, follow-up test). Right: group without experience with the physics curricula. Depicted are solution rates; error bars indicate standard errors of the mean. Q1-4 refers to grouping participants into quartiles according to their results on the cognitive ability test, with Q4 indicating the quartile with the highest cognitive ability.
The test on prior knowledge about speed revealed no difference between the groups (with/without physics curricula) as a whole and when split into quartiles. "Speed" was not part of the physics curricula. Prior knowledge about speed, however, was positively correlated with participants' results on the cognitive ability test ( $r=.32$ ). Cognitively
abler participants scored higher on the test of prior knowledge about speed. See figure 2 for the differences between the quartiles. The average solution rates were around $50 \%$.


Figure 2: Test of prior knowledge about speed. Left: group with prior experience with the physics curriculum. Right: group without experience with the physics curriculum. Depicted are solution rates; error bars indicate standard errors of the mean. Q1-4 refers to grouping participants according to their results on the cognitive ability test, with Q4 indicating the quartile with the highest cognitive ability. No difference between the groups (with/without physics curriculum) is found. "Speed" was not part of the physics curriculum.

## Post-Test

The post-test was very similar to the pre-test and had the same structure as the tasks solved during the intervention. Therefore, the post-test can be viewed as a manipulation check of the implemented proportional reasoning curriculum. This manipulation check was positive in that participants were able to redo tasks that were administered during the curriculum ( $80 \%$ of participants solved all tasks correctly with no mistakes, and there were no differences between the four groups of the $2 \times 2$ design).

## Transfer-Test

Coming to the core results of this study, overall, no significant interaction was observed between groups (with and without prior physics curricula experience) and context of the intervention (values of the ANOVA comparing the four cells of the $2 \times 2$ design: $F(3,236)=0.94, p=0.42)$ ). Therefore, the main effect of "physics curricula" turned out not to be significant. Additionally, the interaction between "physics curricula" and "concept used to introduce proportional reasoning" turned out not to be significant.

Therefore, the findings corroborate neither the more general nor the more specific hypothesis of a positive influence of prior physics curricula on the learning of proportional reasoning. This encouraged us to more closely examine whether the subgroups possibly gained from the physics curricula with regard to learning proportional reasoning. Indeed, we did find effects: When grouping participants by quartiles according to their cognitive ability measure, we
found the expected positive effect of prior physics learning for students in the highest quartile. Those in the highest quartile who underwent the physics curricula scored higher on the proportional reasoning transfer test than students in the lower three quartiles $(p<.05)$. No such effects were found for students in the highest quartile who did not undergo the physics curricula. In the transfer test, the solution rates were relatively low (mean solution rate just slightly over $20 \%$ ); see figure 3 . Similar distributions were found when we only looked at the density groups without evaluating the speed groups: The highest quartile with prior knowledge scored significantly higher than the three lower quartiles, and for those without prior knowledge, the quartiles did not differ. Taken together, for the most intelligent students, the findings corroborate the more general hypothesis. Therefore, physics curricula can serve as preparation for future learning of proportional reasoning.


Figure 3: Results of the proportional reasoning transfer test by physics curriculum (with/without). Left: group with prior experience with the physics curriculum. Right: group without prior experience with the physics curriculum. The transfer test consisted of entirely novel proportional word problems. Depicted are solution rates; error bars indicate standard errors of the mean. Q1-4 refers to grouping participants according to their results on the cognitive ability test, with Q4 indicating the quartile with the highest cognitive ability.

## Discussion

Our results confirm what has been demonstrated many times: Transfer does not come cheap. The study focused on learning opportunities that foster the emergence of consecutive competencies in related fields. Contrary to our expectation, we did not find a general advantage of physics learning (more precisely, density learning) for learning proportional reasoning taught by referring to density.

However, students scoring in the highest quartile of the intelligence measure were able to make use of the prior knowledge they had acquired during the physics curriculum. We therefore conclude that intelligence differences can unfold students' individual potential in combination with sufficient prior knowledge. If a child has high cognitive ability and encounters many examples of proportional
reasoning situations, he or she will be prepared for a subsequent formal learning situation on proportional reasoning. It is the combination of high intelligence and prior experience or specific prior knowledge that leads to the ability to exploit a learning situation better.

The results show that curricula on proportional reasoning are worthwhile for all students in early adolescence. However, more capable students can boost their proportional reasoning if they have the chance to acquire prior knowledge in a physics curriculum.

Future work could focus on the question of how physics curricula can better support students in understanding proportions. It is possible that introducing the abstract structure of the mathematical concept prior to inquiry-based instruction would have an even greater effect.

## References

Alfieri, L., Nokes-Malach, T. J., \& Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. Educational Psychologist, 48(2), 87-113.
Ben-Chaim, D., Fey, J. T., Fitzgerald, W. M., Benedetto, C., \& Miller, J. (1998). Proportional reasoning among 7th grade students with different curricular experiences. Educational Studies in Mathematics, 36(3), 247-273.
Boyer, T. W., Levine, S. C., \& Huttenlocher, J. (2008). Development of proportional reasoning: where young children go wrong. Developmental psychology, 44(5), 1478.

Bransford, J. D., \& Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. Review of Research in Education, 24, 61-100.
Case, R. (1993). Theories of learning and theories of development. Educational Psychologist, 28(3), 219-233.
Chen, Z., \& Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. Child Development, 70(5), 1098-1120.
Chen, Z., \& Klahr, D. (2008). Remote transfer of scientificreasoning and problem-solving strategies in children. In R. V. Kail (Ed.), Advances in child development and behavior (Vol. 36, pp. 419-470). Amsterdam: Elsevier.
Chi, M. T. H., \& VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. Educational Psychologist, 47(3), 177-188.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Goldstone, R. L., \& Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. The Journal of the Learning Sciences, 14(1), 69-110.
Hardy, I., Jonen, A., Möller, K., \& Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". Journal of Educational Psychology, 98(2), 307-326.

Heller, K. A., \& Perleth, C. (2000). KFT 4-12+ R kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision. Beltz Test. Jitendra, A. K., Star, J. R., Starosta, K., Leh, J. M., Sood, S., Caskie, G., ... \& Mack, T. R. (2009). Improving seventh grade students' learning of ratio and proportion: The role of schema-based instruction. Contemporary Educational Psychology, 34(3), 250-264.
Lawson, A. E. (1985). A review of research on formal reasoning and science teaching. Journal of Research in Science Teaching, 22(7), 569-617.
Leuchter, M., Saalbach, H., \& Hardy, I. (2014). Designing science learning in the first years of schooling. An intervention study with sequenced learning material on the topic of 'floating and sinking'. International Journal of Science Education, 36(10), 1751-1771.
Möller, K., \& Jonen, A. (2005). Die KiNT-Boxen-Kinder lernen Naturwissenschaft und Technik. Klassenkisten für den Sachunterricht, 1. See also: https://verlage. westermanngruppe.de/spectra/artikel/A421/Die-KiNT-
Boxen-Klassenkisten-I-Schwimmen-und-Sinken for the KiNT Box of swimming and floating.
Siegler, R. S., \& Lortie-Forgues, H. (2014). An integrative theory of numerical development. Child Development Perspectives, 8(3), 144-150.
Siegler, R. S., Thompson, C. A., \& Schneider, M. (2011). An integrated theory of whole number and fraction development. Cognitive Psychology, 62, 273-296.
Siegler, R. S., \& Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. Developmental science, 11(5), 655-661.
Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., ... \& Chen, M. (2012). Early predictors of high school mathematics achievement. Psychological science, 23(7), 691-697.
Smith, C. L. (2007). Bootstrapping processes in the development of students' commonsense matter theories: Using analogical mappings, thought experiments, and learning to measure to promote conceptual restructuring. Cognition and Instruction, 25(4), 337-398.
Stern, E. (2009). The development of mathematical competencies: Sources of individual differences and their developmental trajectories. In W. Schneider \& M. Bullock (Eds.), Human development from early childhood to early adulthood: Evidence from the Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC) (pp. 221-236). Mahwah, NJ: Erlbaum.
Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., \& Verschaffel, L. (2005). Not everything is proportional: Effects of age and problem type on propensities for overgeneralization. Cognition and Instruction, 23(1), 5786.

Ziegler, E., \& Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. Learning and Instruction, 33, 131-146.

# The relationship between self-control in intertemporal choices and the control of egocentric during perspective taking 

Garret O'Connell (garret.connell@hu-berlin.de)<br>Berlin School of Mind \& Brain, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany<br>Bhismadev Chakrabarti (b.chakarbarti@reading.ac.uk)<br>School of Psychology and Clinical Language Sciences, University of Reading, Reading RG6 6AL, UK<br>Chun-Ting Hsu (hsuchunting@gmail.com@reading.ac.uk)<br>Department of Psychology, Pennsylvania State University<br>460 Moore Building, University Park, PA 16802

Anastasia Christakou (a.christakou@reading.ac.uk)<br>School of Psychology and Clinical Language Sciences, University of Reading, Reading RG6 6AL, UK<br>Isabel Dziobek (isabel.dziobek@hu-berlin.de)<br>Berlin School of Mind \& Brain, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany


#### Abstract

We infer the thoughts and feelings of other people by taking their perspectives, the accuracy of which depends on abilities to control egocentric bias. Similar processes could arguably be used to understand how we would be affected by future events, such as delayed rewards in intertemporal decisions, by allowing us to accurately take the perspective of future selves. In this paper, we test this idea in two studies. In Study 1, we attempted to lower preferences for delayed rewards to examine if this redced abilities to control egocentric bias in a visual perspective-taking task. In Study 2, we examined the neural overlap in intertemporal decision-making and the control of egocentric bias in a false-belief theory-of-mind task. In both studies, a positive relationship was identified between behavioural and neural markers of egocentric bias control and preferences for delayed rewards. The overall pattern of results suggest the overlap in processes of egocentric bias control and those that determine preferences in intertemporal choices, and demonstrate for the first time the effect of sexual arousal on social cognition in reducing abilities to separate one's own perspective from others'.


Keywords: intertemporal choice, temporal discounting, egocentric bias, perspective-taking, temporoparietal junction

## Introduction

The ability to see the world from different perspectives, imagining the thoughts or feelings that might occur in us and others in a variety of situations, is highly useful. It helps to take the perspectives of others to understand and interact effectively with them. This ability is also useful in an intertemporal context, for instance, when faced with decisions or events with delayed consequences, we typically shift our own perspective into the future to assess how these might impact us later. The relationship between these
capacities for taking the perspectives of others and future selves has previously been speculated (see Jamison \& Wegener, 2010; Mitchell et al., 2011; Buckner \& Carroll, 2007), but it is still unclear if and how they overlap. To investigate this, we previously laid out an empiricallygrounded framework to make plausible hypotheses of how these capacities relate to each other (O'Connell et al., 2015), called the Simulation-based Model of Intertemporal Preferences (SMIP).

The SMIP uses the phenomenon of temporal discounting to illustrate how perspective-taking abilities might underlie the perception of future events. Temporal discounting describes how when making choices between rewards to be received now or larger rewards later, the delayed larger reward decreases in subjective value when its receipt is delayed further in the future, leading to smaller but immediate rewards being preferred, and the rate of which is indexed by the steepness of the "discounting curve".

One way the accuracy of perspective-taking can become compromised is through the false presumption that other people think or feel the same way as we do, an error called egocentric bias, as measured using false-belief theory-ofmind (ToM) tasks. Neuroimaging studies of psychiatric patients and children have found that better control of egocentric bias during ToM judgments corresponds to stronger coinciding neural activity in the right temporoparietal junction (rTPJ) (Gweon et al, 2012; Kana et al., 2009; Dodell-Feder et al, 2013). Further reports indicate the rTPJ is preferentially activated for false beliefs that require egocentric bias control, and not true beliefs (Hartwright et al., 2012; Sommer et al., 2007), specifying
the role of this region in egocentric bias control, and not general capacities of perspective-taking.

The SMIP hypothesizes that intertemporal choices situations are analogous to social situations in which perspective-taking occurs, in how there is a target perspective to be inferred, and an egocentric perspective which needs to be controlled for. The SMIP argues that from one's current egocentric perspective, delayed rewards have to be waited for to be received, and therefore incur costs-of-waiting that diminishes their subjective value. In contrast, from the perspective of one's future self, there is no cost of waiting because they are at the right point in time to receive the reward instantly. Controlling this egocentric bias when taking the perspective of the future self should therefore lead to delayed rewards being preferred more, leading to less steep temporal discounting. The SMIP further explicitly predicts the rTPJ as one of the key nodes for control of egocentric bias during both perspective-taking and intertemporal choices. Note these are only a few of the mechanisms and predictions outlined in the SMIP which are relevant to the current study (for further information see O'Connell et al., 2015).

A recent paper from Soutschek et al. (2016) provides direct support for this neural overlap hypothesized by the SMIP. In two studies, repetitive transcranial magnetic stimulation (TMS) was administered to participants' rTPJ regions to disrupt its function. It was found that both the degree of egocentric bias exhibited by participants in a visual perspective-taking task, and preferences for immediate over delayed reward choices in a temporal discounting task, were both subsequently increased. Furthermore, a positive relationship between egocentric bias and immediate reward preferences was observed across individuals. By demonstrating a relationship between egocentric bias and increased temporal discounting, these two findings can be explained by the SMIP framework.

In this paper, we tested the hypotheses of the SMIP by examining the relationship between egocentric bias control and temporal discounting in two studies: 1) examining the effect on one process when the other is experimentally manipulated, 2) examining their neural overlap using fMRI.

## Study 1: Sexual Arousal Manipulation

In Study 1, instead of brain stimulation as per Soutschek et al. (2016), we attempted to increase the steepness of temporal discounting psychologically by inducing sexual arousal with erotic images, which multiple reports indicate is reliable means of increasing the steepness of temporal discounting (Kim \& Zauberman, 2013; Van den Bergh, Dewitte, \& Warlop, 2008; Wilson \& Daly, 2004). Sexual arousal is thought to lead to steeper temporal discounting by causing a generalized state of desire, in which the immediacy of a reward overshadows its objective value. For a comparison control, equally arousing sports images were used, as per Kim \& Sauberman (2013). The Director task was used to index control of egocentric bias.

## Methods

## Participants and procedure

Heterosexual German speaking males were recruited ( $\mathrm{n}=$ 90, range $19-59$ years, mean 29.3) performed tasks in the following order: 1) temporal discounting (baseline), 2) visual perspective-taking Director task, 3) temporal discounting (manipulation-check), 4) Continuous performance task (CPT). After task 1) the impulsivity manipulation was initiated, where in separate groups participants viewed either erotic or arousal-matched sports images for 8 s , which occurred every trial of the Director task, 6 trials of the temporal discounting task, 13 trials of the CPT, for an average of one image every 14 s .

## Temporal discounting (baseline)

Participants made intertemporal choices between a variable amount of money now $(<100 €)$ or $100 €$ at one of four randomly selected delays (months: 1, 3, 6, 12). The amount of immediate options and indifferences-points were calculated using the double-limits algorithm, and temporal discounting rates ( $k$ ) were estimated by fitting a hyperbolic non-linear model to each participant's indifference-points

## Director task

In the Director task, participants move objects around a set of shelves as instructed by a "director" standing on the opposite side, but to take into account that some objects cannot be seen by the director because they are occluded from their side of the shelf. This requires the participant to control for egocentric information from their own viewpoint when inferring the director's. The Director task was computerized with a real person in the role of director, and the impression of actual shelves was created by positioning the participant's screen back-to-back with the director's. A previously used eye-gaze metric of egocentric bias was used - the average dwell time (sum of 100 ms fixations) to distractor objects, a basic index of how much the distractor object (the correct object from the participant's egocentric viewpoint) was considered as the correct option.

Participants' heads were placed in a chin-rest positioned in front of an Eye-tribe 30 Hz eye-tracker. Trials began with a fixation cross when the experimenter read aloud a scripted instruction, then pressed a key to present the shelves on screen. Instructions in experimental trials referred to distractor objects on dimensions of spatial (e.g. "move the top ball" could refer to a higher hidden ball), size (e.g. "move the large ball" could refer to a larger hidden ball), or semantic (e.g. "move the mouse" could refer to a computer mouse or a hidden toy mouse). In control trials instructions referred to objects without competing referent. 30 trials were performed ( 24 experimental, 6 control), featuring 9 sets of 6 objects, presented in a fixed randomized order.

## Temporal discounting task (manipulation-check)

Participants performed intertemporal choices a second time to check the sexual arousal manipulation. Immediate
reward options were present from ranges $( \pm 10 €)$ around participants' baseline indifference points in 32 trials. Changes in temporal discounting were estimated as valueweighted changes from baseline temporal discounting.

## Continuous Performance Task (CPT)

The CPT measures general attention to test possibility that erotic images affected egocentric bias by merely distracting attention from task goals. Single digit numbers were presented for $100 \mathrm{~ms}(1 \mathrm{~s}$ ISI), and participants had to press one key following a sequence of 3-7 (average every 10 of 130 trials), and another key otherwise.

## Results

## Effects of sexual arousal manipulation

Note tests of the hypothesis were directional and conducted at the 1-tailed level. In the Director task, dwell time to the distractor object was significantly higher following erotic $(\mathrm{M}=317 \mathrm{~ms}, \mathrm{SE}=.18)$ images versus sports images $(\mathrm{M}=270 \mathrm{~ms}, \mathrm{SE}=.17), t(88)=1.9, p=.032$, $\mathrm{d}=.4$ (Figure 1B), an effect not observed in control trials, $t(88)=.043, p=.97$. Although changes from baseline temporal discounting were ordinally higher following erotic $(\mathrm{M}=1.5, \mathrm{SE}=3.3)$ compared to sports $(\mathrm{M}=-.6, \mathrm{SE}=3.6)$ images, this difference was not significant, $t(88)=.44, p=$ .33. There was no significant difference in error rates in the CPT task between the erotic ( $\mathrm{M}=5 \%, \mathrm{SE}=.02$ ) or sports $(\mathrm{M}=2 \%, \mathrm{SE}=.01)$ image groups, $t(88)=1.3, p=.18$.

## Relationship between egocentric bias and temporal discounting

A linear regression was used to model the relationship between baseline temporal discounting rates and egocentric bias, controlling for the influence of group. The model found that after controlling for the group factor $(B=-.524, t$ $=-2.11, p=.038$ ), temporal discounting was a significant positive predictor of egocentric bias $(B=.316, t=2, p=$ .048). Note, data from the second temporal discounting task could not be used for this form of analysis because in this task, choice options were pre-determined from baseline, and not the double-limits algorithm which is required to reliably estimate new temporal discounting rates (Figure 1A).



Figure 1. A: Scatterplot of temporal discounting and egocentric bias (residuals partialling out group effects). B: Group differences in egocentric bias (error-bars: SE).

## Discussion

In support of the hypothesis, it was found that people who exhibited steeper temporal discounting were also more susceptible to egocentric bias. Also in line with this view, egocentric bias was higher following erotic stimuli, which although previously reported to lead to steeper temporal discounting, was not specifically observed here.

Contrary to previous findings, sexual arousal from did not significantly alter temporal discounting here, limiting the extent to which changes in egocentric bias can be attributed to changes in self-control. One explanation for this null finding is that the arousing effects of erotic images had habituated by the second time the temporal discounting task, after already 15 mins of repeated exposure to images. It's worth noting that in pilot data, where temporal discounting was the first task performed, steeper temporal discounting was found following erotic images compared to sports images ( $\mathrm{n}=32, p=.05$ ), giving some confidence that full counterbalancing of tasks in this study would have increased chances of observing effects of sexual arousal on temporal discounting. However, this null effect still warrants caution in interpreting the results.

## Study 2: fMRI

In Study 2, we tested the relationship between egocentric bias and temporal discounting using fMRI. If rTPJ activity during perspective-taking is a correlate of egocentric bias control, as hypothesized by the SMIP and suggested by empirical evidence, then people higher in this marker should prefer delayed rewards more during intertemporal choices. To test this hypothesis, we used a false-belief functional localizer task to extract activity in the rTPJ related to egocentric bias control from each participant. We further tested if activity in this rTPJ cluster was higher when delayed rewards are chosen, as would be expected if egocentric bias control reduces temporal discounting.

## Methods

## Participants and procedure

36 English speaking adults ( 21 female, aged 18-34 years, mean 22.6) performed the following tasks in order: outside scanner, 1) temporal discounting task; inside scanner, 2) temporal discounting task, 3) ToM localizer task.

## Temporal discounting (outside scanner)

Same as temporal discounting task (baseline) in Study 1, except with pounds in place of euros, and using the following delays (months: $1,3,6,9,12,18$ ).

## Temporal discounting (in scanner)

In the scanner, participants were presented with intertemporal choices featuring three delays (months: 6, 9 , 12). As in the temporal discounting task (manipulationcheck) in Study 1, immediate options were estimated from participants' temporal discounting data collected outside
scanner in order to predict how participants would decide and efficiently balance the number of trials in conditions of immediate (IMM) and delayed (DEL) choices, this time from value ranges $\pm £ 5-15$ indifference-points. Trials continued until 32 (balanced across delays) were collected in which each immediate and delayed rewards were chosen (the IMM and DEL conditions). Options were presented together for 5 s , followed by a jittered ITI of 7-15 s.

## fMRI temporal discounting acquisition and analysis

25 out of the 36 participants performed the scanner version of the temporal discounting task. Scanning was conducted using a Siemens 3T Trio MRI scanner with an EPI sequence of TR 3 s , TE $30 \mathrm{~ms}, 2 \mathrm{~mm}^{3}$ voxel size, and 35 interleaved 3 mm slices. Data were preprocessed using a GLM in SPM8, with slice-timing correction, realignment for motion correction, field map unwarping, and sequential co-registration. We contrasted rTPJ activity in DEL > IMM separately for each individual in their specific rTPJ ROIs from the ToM task, by first registering individual rTPJ ROIs to MNI space using FEAT, and the mean contrast values of DEL and IMM conditions were then extracted.

## False-belief ToM localizer task

The false-belief localizer task consisted of 10 stories about other people's beliefs (False Belief) or historical facts (FACT). Each trial started with a blank screen for 12 s , followed by the story for 10 s , and then a question screen for 4 s , which required a "True" or "False" response.

## fMRI ToM localizer data acquisition and analysis

Scanning was conducted with an EPI sequence of TR 2 s , TE $30 \mathrm{~ms}, 2 \mathrm{~mm}^{3}$ voxels, and 37 interleaved 3 mm slices. Using FSL, data were field map unwarped, pre-whitened, motion corrected, slice-time corrected, and high-pass filtered at 128 Hz and smoothed at 8 mm FWHM in native space. False Belief and FACT trials were defined as the 14 s of the story and question screens. Using FEAT, clusters from the False Belief $>$ FACT contrast were identified at height threshold $\mathrm{z}=2.3$, cluster threshold $\mathrm{p}<.05$, minimum size 200 voxels, Gaussian Random Field FWER corrected.

## rTPJ ToM localization procedure

An iterative threshold-adjusting procedure was adapted from Mitchell (2008) to identify individual rTPJ clusters related to false-belief processing. This procedure involved increasing the height activation threshold of the False Belief $>$ FACT contrast in native space in steps of $10-1$, starting from $\mathrm{p}<0.01$ until a cluster in the rTPJ region was identified $25-50$ voxels in size. Percentage signal change in participants' individual clusters, hereafter referred to as $\mathrm{rTP}_{\mathrm{FB}}$. For thoroughness of reporting, this procedure was applied to other regions associated with ToM in the left temporoparietal junction (lTPJ) and precuneus.

## Results <br> Data cleaning and sample selection

rTPJ clusters could not be localized for two participants, and 2 participants were extreme outliers in the DEL vs. IMM in the rTPJ ROI (Tukey's interquartile range), This threshold identified two cases for exclusion.

## Correlation between egocentric bias control rTPJ response and temporal discounting

A significant negative correlation was found between temporal discounting rates $k$ and $\mathrm{rTPJ}_{\mathrm{FB}}, \mathrm{r}=-.32, p=.03$ (Figure 2) ( $\mathrm{lTPJ}_{\mathrm{FB}}, \mathrm{r}=-.36, p=.02$; precuneus ${ }_{\mathrm{FB}}, \mathrm{r}=-.37, p$ $=.01$ ), an effect still significant in the temporal discounting scanning session subsample, $\mathrm{r}=-.36, p=.04$.


Figure 2. Scatterplot of temporal discounting and magnitude of $\mathrm{rTPJ}_{\mathrm{FB}}$ activity.

## fMRI temporal discounting results

A significant difference was found in individual rTPJ ROIs in DEL $>$ IMM, $t=2.06, p=0.025$ (Figure 3B).


Figure 3. A: Overlaid individual rTPJ ROI clusters (crosshairs at peak). B: Differences between choice conditions in individual rTPJ ROIs (error bars: SE).

## Discussion

In Study 2, we found two pieces of evidence in support of the hypothesis that better control of egocentric bias reduces steepness of temporal discounting. First, people exhibiting a higher rTPJ response during false-belief judgments, a putative neural marker of egocentric bias control, had less steep temporal discounting. Second, responding in the same rTPJ cluster involved in egocentric bias control was higher
when delayed rewards were preferred over immediate rewards when making intertemporal choices in the scanner.

Evidently, this interpretation of the results relies on a reverse-inference about the rTPJ's function of egocentric bias control. Neural markers of egocentric bias control have the advantage of being continuous, and more resistant to ceiling effects that standard measures (e.g. false-belief tasks) are often prone to, making them better suited for measuring variability in healthy adults. Increased rTPJ activity has been repeatedly shown to be associated with higher ToM accuracy (Gweon et al., 2012; Kana et al., 2009). Similarly here, despite ceiling effects in accuracy in the current ToM data, a trending positive correlation was noted between this accuracy and magnitude of rTPJ activity ( $\mathrm{r}=.23, p=.09$ ), providing some behavioural support for the function of rTPJ activity claimed here. Some studies with children, who do not show the ceiling effects typically seen in ToM tasks in adults, have shown positive links between accuracy and preferences for delayed rewards (Launay et al., 2015; Marchetti et al., 2014).

The exact function of the rTPJ in perspective-taking remains an open question. Based on reports of the rTPJ's importance in focusing attention on distinctions between self and others (e.g. in terms of preferences; Nicolle et al., 2012), it has been proposed that this region helps avoid the perceptual blurring of self and other perspectives that drives egocentric bias (Brass et al., 2009). More generally, the rTPJ has been claimed to be involved in orienting attention towards task goals (Corbetta \& Shulman, 2002). In any form of choice, it could be argued that a goal to maximize reward outcomes, irrespective of the time of receipt, becomes activated. In intertemporal choices, delayed rewards are larger, and hence, most relevant to this goal. In contrast, immediate rewards could be considered distractions that draw attention away from this goal, shifting focus instead to the immediacy of enjoyment. This view of rTPJ function would account for its observed increased activation when choosing delayed rewards.

## General Discussion

The effects in this paper are modest and mixed, but support the view that temporal discounting is related to abilities of perspective-taking, as hypothesized by the SMIP. To sum up, in both Study 1 and Study 2, individual differences in perspective-taking corresponded to steepness of temporal discounting. In Study 1, an experimental manipulation previously shown to led to steeper temporal discounting was found to increase egocentric bias. However, because of the null effect on temporal discounting itself, the extent to which this egocentric bias can be attributed to processes overlapping with temporal discounting is limited. In Study 2, evidence for the SMIP was found in how individually-localized neural markers of egocentric bias control were higher when delayed rewards were preferred over immediate ones. In all, these findings, provide exploratory first steps in examining potential connections between intrapersonal and interpersonal forms of cognition.

Such a connections has numerous theoretical and practical upshots. Linking the fields of social cognition and intertemproal decision-making would allow the benefits of one (e.g. conceptual frameworks in social cognition, mathematical models in decision-making) be transferred to the other. Practically, it would suggest the potential of temporal discounting as a continuous and fast index of perspective-taking abilities in adults, which are currently required in social cognition research.

To-date, the overwhelming amount of research in intertemporal decision-making has restricted focus to the concept of value, and inputs and outputs to its computation. Increasingly, more concepts about how this value is psychologically represented have come under study, including future prospection (Kwan et al., 2015), and feelings of connectedness with one's future (Urminsky, 2017). The SMIP aims to extend this line of inquiry by attempting to provide a mechanism for how these representations are built, and why they degrade to cause temporal discounting, parsimoniously based on mechanisms that already underlie social abilities.

Sexuality is an integral social setting, but its impact on perspective-taking abilities has not been tested before now. The present finding thus furthers understanding of how an everyday social context modulates mentalizing capacities, which vital information for theories to build sufficiently detailed descriptions of these processes (for similar work, (Galinsky, Magee, Ena Inesi, \& Gruenfeld, 2006; Kanske, Böckler, Trautwein, Parianen Lesemann, \& Singer, 2016; Todd, Forstmann, Burgmer, Brooks, \& Galinsky, 2015).

The current findings, along with those of Soutschek et al. (2016), encourage further investigations of the overlap in intertemporal choice and perspective-taking, but the noted inconsistencies warn that conclusive evidence could be challenging to find. One clear example of this is the difficulty in detecting overlap in capacities that are measured by tasks with vastly different structures and demands. The results also flag sexual contexts as influential to the ability to look past our own perspective to better understand those of others, calling for further investigations into this important but little understood topic. Future work can test the generalizability of these effects in larger samples, using different measures, especially ones that can measure these capacities in more closely matched structures.

## Acknowledgments

GOC received a scholarship from the University of Reading and Humboldt-Universität zu Berlin.

## References

Brass, M., Ruby, P., \& Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364(1528), 2359-2367.
Buckner R. L., Carroll D. C. (2007). Self-projection and the brain. Trends in Cognitive Sciences, 11 49-57.

Corbetta, M., \& Shulman, G. L. (2002). Control of goaldirected and stimulus-driven attention in the brain. Nature Reviews Neuroscience, 3(3), 201-205.
Dodell-Feder, D., Tully, L. M., Lincoln, S. H., \& Hooker, C. I. (2013). The neural basis of theory of mind and its relationship to social functioning and social anhedonia in individuals with schizophrenia. NeuroImage: Clinical, 4, 154-163.
Galinsky, A. D., Magee, J. C., Ena Inesi, M., \& Gruenfeld, D. H. (2006). Power and Perspectives Not Taken. Psychological Science, 17(12), 1068-1074.
Gweon, H., Dodell-Feder, D., Bedny, M., \& Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts: Behavioral and neural development in theory of mind. Child Development, 83, 1853-1868.
Hartwright, C. E., Apperly, I. A., \& Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. NeuroImage, 61(4), 921-930.
Jamison, J., \& Wegener, J. (2010). Multiple selves in intertemporal choice. Journal of Economic Psychology, 31, 832-839.
Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J., \& Just, M. A. (2009). Atypical frontal-posterior synchronization of Theory of Mind regions in autism during mental state attribution. Social Neuroscience, 4, 135-152.
Kanske, P., Böckler, A., Trautwein, F.-M., Parianen Lesemann, F. H., \& Singer, T. (2016). Are strong empathizers better mentalizers? Evidence for independence and interaction between the routes of social cognition. Social Cognitive and Affective Neuroscience, 11(9), 1383-1392.
Kim, B. K., \& Zauberman, G. (2013). Can Victoria's Secret change the future? A subjective time perception account of sexual-cue effects on impatience. Journal of Experimental Psychology: General, 142(2), 328-335.
Launay, J., Pearce, E., Wlodarski, R., van Duijn, M., Carney, J., \& Dunbar, R. I. M. (2015). Higher-order mentalising and executive functioning. Personality and Individual Differences, 86, 6-14.
Marchetti, A., Castelli, I., Sanvito, L., \& Massaro, D. (2014). Is a bird in the hand worth two in the future? Intertemporal choice, attachment and theory of mind in school-aged children. Frontiers in Psychology, 5.
Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. Cerebral Cortex, 18(2), 262-271.
Mitchell J. P., Schirmer J., Ames D. L., Gilbert D. T. (2011). Medial prefrontal cortex predicts intertemporal choice. Journal of Cognitive Neuroscience, 23, 857-866.
Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., \& Behrens, T. E. J. (2012). An agent
independent axis for executed and modeled choice in medial prefrontal cortex. Neuron, 75(6), 1114-1121.
O'Connell, G., Christakou, A., \& Chakrabarti, B. (2015). The role of simulation in intertemporal choices. Frontiers in Neuroscience, 9.
Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., \& Hajak, G. (2007). Neural correlates of true and false belief reasoning. NeuroImage, 35(3), 13781384.

Soutschek, A., Ruff, C. C., Strombach, T., Kalenscher, T., \& Tobler, P. N. (2016). Brain stimulation reveals crucial role of overcoming self-centeredness in self-control. 2, el600992
Todd, A. R., Forstmann, M., Burgmer, P., Brooks, A. W., \& Galinsky, A. D. (2015). Anxious and egocentric: How specific emotions influence perspective taking. Journal of Experimental Psychology: General, 144(2), 374-391.
Wilson, M., \& Daly, M. (2004). Do pretty women inspire men to discount the future? Proceedings of the Royal Society B: Biological Sciences, 271, S177-S179.
Van den Bergh, B., Dewitte, S., \& Warlop, L. (2008). Bikinis instigate generalized impatience in intertemporal choice. Journal of Consumer Research, 35(1), 85-97.

# The dot perspective task revisited: Evidence for directional effects 

Cathleen O'Grady (C.J.O'Grady@sms.ed.ac.uk) ${ }^{\mathbf{a}}$, Thom Scott-Phillips ${ }^{\text {b,c }}$, Suilin Lavelle ${ }^{\text {a }}$, Kenny Smith ${ }^{\text {a }}$<br>${ }^{\text {a }}$ School of Philosophy, Psychology \& Language Sciences, University of Edinburgh<br>${ }^{\mathrm{b}}$ Dept. of Cognitive Science, Central European University, Budapest<br>${ }^{c}$ Dept. of Anthropology, Durham University, UK


#### Abstract

Humans are highly social creatures. Evidence from the dot perspective task suggests that humans automatically track the perspective of other individuals - a disposition that, if true, may help to facilitate social interaction. However, variants of the original dot perspective task suggest the alternative interpretation that the effect in the task is not due to perspective taking. Here, we present a new variant, using improved stimuli to address these issues. Our results replicate previous findings, across both animate and inanimate stimuli, and suggest that the effect is due to directional cueing rather than automatic perspective taking.


Keywords: perspective taking; dot perspective task; automaticity; theory of mind; mindreading

## Introduction

The ability to reason about other individuals' mental states ("mindreading") is thought to be a central component of social cognition in humans (Corballis, 2011; Graziano, 2013; Tomasello, 2008, 2014). In order to explain the social abilities that are best accounted for by mindreading, it seems necessary that certain forms of mindreading are highly efficient (Apperly, 2011; Apperly \& Butterfill, 2009; Butterfill \& Apperly, 2013). Evidence for efficient mindreading comes from various experimental paradigms (Freundlieb, Kovács, \& Sebanz, 2016; Schneider, Slaughter, \& Dux, 2017; Scott \& Baillargeon, 2017), including the dot perspective task (DPT) (Samson, Apperly, Braithwaite, Andrews, \& Bodley Scott, 2010), which suggests that participants rapidly and automatically calculate the perspective of other agents.

However, the interpretation of these results is disputed. Different variants of the DPT (e.g. Cole, Atkinson, Le, \& Smith, 2016; Santiesteban, Catmur, Hopkins, Bird, \& Heyes, 2014) have produced results that may be explained by a simple directional effect, in which attention is directed not exclusively by the gaze of an agent, but rather by any directional stimuli. If the task results are indeed attributable to directional cueing, it would undermine the use of this task as evidence for fast and automatic mindreading. We describe the different variants in the next section, before describing a new variant, using Lego figures, that may be used to address these issues, and the experimental results obtained using it.

## Variants of the Dot Perspective Task

In the dot perspective task, participants observe scenes and answer a simple yes/no question based on the number of dots in the scene. The scenes that participants view feature an on-screen human avatar standing in a room. Arranged on walls around the room are various dots. In some scenes, the dots all appear in front of the avatar, making the avatar's perspective of the dots consistent with the participant's: e.g. if there are two dots on the front wall, the avatar and the participant both see two dots. In other scenes, some of the dots are behind the avatar, making the avatar's perspective inconsistent with the participant's: the avatar might see only one dot, while the participant can see two.

Participants are shown a digit (e.g. " 2 "), followed by one of these scenes, and asked to confirm whether the number of dots matches the pre-scene digit by answering "Yes" or "No." In three different experiments, Samson et al. (2010) found longer reaction times for inconsistent scenes compared to consistent scenes, which they interpreted as evidence for "altercentric interference": the participant had to suppress the avatar's perspective in order to answer the question of whether the digit matched their own perspective, resulting in a delayed response. This suggests that perspective taking, even for an on-screen avatar, is rapid and automatic.

In the first two of these three experiments, participants were asked to judge their own perspective on certain scenes (cued by the word "YOU" appearing before the digit), and the avatar's perspective on others (cued by the word "HE" or "SHE"). Because this may have caused participants to take the avatar's perspective in all scenes, Experiment 3 instructed participants to ignore the stimuli in the middle of the room and judge only their own perspective; the consistency effect persisted.

Santiesteban et al. (2014) argue that the effect of the avatar on reaction times was driven not by perspective taking of the avatar but rather by a directional effect: because the avatar faced one or the other side of the room, the participant's attention might be directed towards stimuli on that side. They repeated the experiment using avatarsized arrows (rather than columns) as controls, finding a consistency effect for both avatars and arrows, both when perspective switched between trials (Experiment 1), and


Figure 1: Example scenes.
A: The main components of each scene. B: Example scene with Sally. C: Example scene with Andrew. D: Example scene with arrow.
when participants were instructed to ignore the stimuli in the centre of the room and judge only their own perspective (Experiment 2). However, because both kinds of stimulus were presented to all participants (i.e. the avatar vs arrow manipulation was within-subjects), it is possible that participants were transferring the "perspective taking" of the avatar over to the arrow.

Cole et al. (2016) note a further problem with this experiment: although arrows and avatars produce a similar effect on reaction times, these effects may in fact be driven by different processes-perspective taking in the case of the avatar, and directional cueing in the case of the arrows. Indeed, Marotta, Lupiáñez, Martella, \& Casagrande (2012) find that, while eye gaze cues participants to a specific location, an arrow provides a more general cue. This suggests that different processes are involved in following the directional cue of an arrow and an avatar.

As an alternative control, Cole et al. (2016) use a set of stimuli that includes a barrier in front of the avatar, as is used in mentalising experiments in non-human animals (Hare, Call, \& Tomasello, 2001). When the barrier "window" is open, allowing the avatar to "see" the dots, they find the expected consistency effect; but they also find the effect when the barrier window is closed, suggesting that the effect is driven simply by the directional effect of the avatar, rather than by mental state attribution.

However, the stimuli used in this experiment do not make it perfectly clear whether or not the barrier is transparent, and the depth and angle of the barrier placement within the room could be ambiguous. Further, the temporary nature of the barriers may create a problem: given that the participant likely assumes that the avatar is a single agent, it is possible that participants infer the agent's knowledge of what is on the other side of the barrier on the basis that they can
sometimes see what is there, and may have done so before the barrier window closed.

Cole et al. (2016) do attempt to deal with these problems. The open or closed barriers were shown in different blocks of trials, and at the beginning of each block, participants were explicitly told whether or not the avatar could see the wall that was blocked by the barrier. However, given the visual ambiguity of the stimuli, it is possible that this kind of explicit knowledge is not taken into account in fast processing, when at a glance the image might be interpretable in different ways.

Using different stimuli and a modified experiment design, we conducted a conceptual replication of Experiment 3 in Samson et al. (2010) and Experiment 2 in Santiesteban et al. (2014). Although our experiment was designed to address details of Yes vs. No responses and arrow vs. avatar stimuli, the design also allowed us to explore the effect of barriers as in Cole et al. (2016), while addressing the problems of ambiguity. Unlike Samson et al. (2010) but following Santiesteban et al. (2014) we used arrows as a directional control for avatars; unlike Santiesteban et al. (2014), we manipulated avatars vs arrows in a between-participants design, rather than within-participants. Our stimuli did not have the same temporal and physical ambiguity as the images used by Cole et al. (2016) (see Figure 1). We used photographs of Lego figures in scenes with unambiguous depth in the third dimension, and solid black barriers were used, preventing any ambiguity in whether or not Lego figures were able to see through them.

A variety of hiding places allowed balls (our equivalent of dots/discs) to be hidden from view of the Lego figures, even when placed in front of them. This allowed us to test the claim that the altercentric effect could be explained by the general directionality of the avatars, rather than perspective taking.

In addition, the use of arrows as control stimuli should indicate whether, as in Marotta et al. (2012), the arrows have a more general directional effect than the avatars. If this were the case, one would expect arrows to cause a reaction time delay only when there are balls placed in the opposite direction to that indicated by the arrow; and the more specific perspective attributed to avatars to cause a reaction time delay in all cases where there are balls not in its field of view (regardless of whether they are hidden behind a barrier in front of, or behind, the avatar).

## Method

## Participants

Sixty participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated $£ 4$ for their participation, which lasted approximately 20 minutes. Thirty participants viewed stimuli with the Lego figures, and thirty viewed control stimuli showing columns with arrows on them. One further participant was excluded from analysis because a postexperiment questionnaire indicated that they had successfully guessed the purpose of the experiment.

## Materials

Participants observed scenes consisting of photographs (Figure 1) of Lego figures (dubbed "Sally" and "Andrew" for ease of reference), a series of barriers created by Lego bricks, and red beads that, at Lego scale, had the appearance of red balls. Control stimuli consisted of Lego columns with the same colours and proportions as Sally and Andrew, with a black arrow on the yellow block, pointing in the same direction as a figure's direction of facing. Each scene featured either Sally or Andrew (each figure could appear on either side of the screen), and between 0 and 4 balls (with a maximum of two balls in any given location).

## Procedure

On each trial, participants were presented with a fixation cross for 750 ms , followed by a digit between 0 and 4 (displayed for 750 ms ), followed by a Lego scene, with the words "Yes" and "No" in the bottom corners of the screen (Yes-side was counterbalanced across participants but remained consistent across trials for a given participant). Participants were instructed to judge whether the picture had the same number of balls as the digit they had been shown with no other comment given about the other elements of the scene - using a two-button button box, pressing the Yesside button for yes and the No-side button for no. Scenes timed out within 2000 ms if no response was given, and moved on to the following trial.

After completing 12 practice trials with correct/incorrect feedback on responses, participants completed 324 trials (36 filler trials with zero balls, and 288 test trials), in random order, divided into four blocks, with a self-paced break between blocks. These 288 trials balanced three different variables: the number of balls in a scene, the consistency
between the Lego avatar's and participant's perspective, and the match between the digit shown and the number of balls in the scene.

There were 72 trials for each number of balls; that is, 72 scenes with one ball, 72 with two balls, and so on. Half of the trials were consistent in perspective: that is, the figure/arrow could "see" (i.e. had unobstructed line of sight to) the same number of balls that the participant could see. The other half were inconsistent, with balls hidden from the figure or arrow by either the central, table-like barrier or the external wall-like barriers, introducing an inconsistency between the participant's perspective and that of the avatar/arrow. The match between the digit shown and the on-screen perspective was balanced (Table 1); the results of analysis of this variable will be reported in a future paper.

Table 1: Match between digit and perspective

|  | Inconsistent |  |  | Consistent |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Avatar sees 2; participant sees 3 |  |  | Avatar sees 2; participant sees 2 |  |
|  |  |  |  |  |  |
| Digit <br> shown | 3 | 2 | 4 | 2 | 3 |
| Correct answer | Yes | No | No | Yes | No |
| Condition | Yes | No- <br> Other | No- <br> None | Yes | No- <br> None |

Post-experiment questionnaires were used to assess whether participants' intuitions about the figures' lines of sight matched those of the experimenters. Pictures showing a variety of scenes with balls in different positions were displayed, and participants were asked to note how many balls the Lego figure could see (regardless of whether they had just completed the avatar or arrow condition of the experiment). All responses to these questionnaires indicated that participants did not expect the Lego figures to be able to see balls hidden by either the central or external barriers, but did expect them to see balls either on the table or at their feet.
The experiment was implemented using PsychoPy (Peirce, 2010).

## Results

We used lme4 (Bates, Maechler, Bolker, \& Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, \& Christensen, 2016) to perform a series of linear mixed effects analyses on reaction time (RT); RT was our only dependent variable given the lack of effect on error rate found in our own data and in previous studies. We removed training trials, trials with zero balls on screen, timed-out trials $(0.69 \%, \mathrm{n}=119)$, and trials where participants made an incorrect response ( $3.12 \%, \mathrm{n}=533$ ). As per Whelan (2008), trials in which the response RT was lower than 100 ms were also removed, on the assumption that these trials could not be genuine responses to the stimuli $(0.01 \%, \mathrm{n}=2)$. No trimming was conducted on higher reaction times, given the imposed cutoff of 2000 ms on all trials. Visual inspection of the reaction time data revealed an obvious deviation from the normal distribution, necessitating a log transform of the data (Baayen \& Milin, 2010).


Figure 2: Mean RTs showing a significant effect of Consistency (error bars show 95\% CI) and no Stimulus x Consistency interaction. The effect of Stimulus is not significant (note that Stimulus, unlike Consistency, is manipulated between-subjects). Y-axis limited for easier comparison with earlier experiments.

## Replication

We first conducted an analysis of the relationship between RT, Consistency and Stimulus. As fixed effects, we entered Consistency and Stimulus (with interaction term) into the model. As random effects, we included random intercepts for participants and images, as well as by-participant and by-image random slopes for the effects of Consistency and Stimulus (without interaction term, to facilitate model convergence).

Following Samson et al. (2010), the model showed a significant effect of Consistency (Figure 2), with consistent trials faster than inconsistent trials $\left(\beta=0.047^{1}, \mathrm{SE}=0.008\right.$, $p<.001$ ). Contra Samson et al. (2010) but consistent with Santiesteban et al. (2014), there was no effect of Stimulus ( $\beta=-0.065, \mathrm{SE}=0.049, p=.187$ ) and no Stimulus x Consistency interaction ( $\beta=0.012$, $\mathrm{SE}=0.01, p=.220$ ). This suggests that an inconsistency in perspective resulted in slower responses, but that this was true for both avatars and arrows. Our between-subjects manipulation of avatars vs arrows ensures that, unlike for Santiesteban et al. (2014), this cannot be explained as a consequence of transfer from avatars to arrows: our participants seeing the arrow stimuli had not seen Lego figures in those positions.


Figure 3: Mean RTs showing a significant effect of Directional Consistency.

## Directional Consistency

Our experimental setup also allowed us to test the hypothesis that the delay is caused not by processing of the altercentric perspective, but rather by preferential attention to objects in the direction of facing/arrow pointing. We predicted, based on Marotta et al. (2012), that the delay would appear only on those trials where balls in front of the avatar are within the avatar's actual field of view, and not on trials where there are balls in front of the avatar, but hidden by obstacles, consistent with the explanation of altercentric interference. We similarly predicted that when the stimulus was an arrow instead of an avatar, the delay would occur on all trials where there are balls within the arrow's field of reference, regardless of barriers between the balls and the arrow.

[^425]To test these predictions, we re-coded the data to classify all trials with balls in front of the avatar/arrow as directionconsistent, and only those trials where a ball appeared behind the avatar/arrow as direction-inconsistent. We then modelled the relationship between this Directional Consistency, Stimulus, and RT (Figure 3). Contrary to our predictions, the results showed that directional-inconsistent trials were slower than directional-consistent trials ( $\beta=0.047, \mathrm{SE}=0.01, p<.001$ ), with no significant effect of Stimulus ( $\beta=-0.058, \mathrm{SE}=0.049, p=.24$ ) and no significant interaction ( $\beta=0.004$, $\mathrm{SE}=0.011, p=.73$ ). This suggests that the consistency effect may be driven by preferential attention to objects within a directional figure's direction of facing/pointing, regardless of the animacy of that figure.

Table 2: Congruence

## 

| Line of Sight | Line of Sight | Line of Sight |
| :---: | :---: | :---: |
| consistent | inconsistent | inconsistent |
| Directional | Directional | Directional |
| consistent | consistent | inconsistent |

However, a further model with both Consistency and Directional Consistency as fixed effects found a significant effect for both variables $(\beta=0.032, \mathrm{SE}=0.011, p=.005$ and $\beta=0.04, \mathrm{SE}=0.008, p<.001$ respectively). In order to explore this, the data was recoded to classify each scene as consistent and/or inconsistent for both definitions of consistency (Table 2). That is, each scene could be (a) line of sight consistent + directional consistent (balls within the avatar's direction of facing and actual field of view); (b) line of sight inconsistent + directional consistent (balls within the avatar's direction of facing, but hidden from the avatar's field of view); or (c) line of sight inconsistent + directional inconsistent (inconsistent based on both direction of facing and field of view).

A model with this variable (Congruence) and Stimulus as fixed effects (with interaction term) found that line of sight consistent + directional consistent trials were faster than both line of sight inconsistent + directional consistent ( $\beta=0.036, \mathrm{SE}=0.012, p=.003$ ) and line of sight inconsistent + directional inconsistent $(\beta=0.077$, $\mathrm{SE}=0.012, p<.001$ ) trials (Figure 4); a re-levelled model showed that line of sight inconsistent + directional consistent was significantly faster than line of sight inconsistent + directional inconsistent $(\beta=0.041$, $\mathrm{SE}=0.01, p<.001$ ). There was no effect of Stimulus or Stimulus x Congruence interaction.


Figure 4: Mean RTs showing a significant effect of Congruence: scenes with consistent perspectives and unobstructed balls are faster than scenes with inconsistent perspectives created by barriers in front of the stimuli; which in turn are faster than scenes with balls hidden both in front of, and behind, the stimuli.


Figure 5: Without the confound of peripheral balls, there is no effect of Line of Sight consistency on RT.

These results would suggest a role for both Consistency and Directional Consistency in affecting reaction times, but there is an important confound: within directionally consistent scenes, line of sight consistent scenes can only have balls in the centre of the screen, while line of sight inconsistent scenes may have balls on the periphery of the screen (the same confound does not apply across directional
consistent vs. directional inconsistent scenes, which may both have peripheral balls). Once data is restricted to only those scenes with balls in the centre of the scene (all of which are directionally consistent), there is no longer an effect of line of sight consistency ( $\beta=-0.008, \mathrm{SE}=0.013$, $p=.525$, Figure 5). This suggests that the consistency effect may be accounted for by the directional hypothesis.

## Conclusion

These results replicate the headline result of Samson et al. (2010) by finding a robust effect of Consistency on reaction times. However, they also replicate the results of Santiesteban et al. (2014) by finding that the Consistency effect appears with inanimate but directional stimuli, even when those stimuli appear in a between-participants design. Additionally, the analysis of Directional Consistency suggests that the effect is driven by a directional cueing effect. These findings cast uncertainty on interpretation of DPT data as evidence for automatic mindreading.

Heyes (2014) argues that evidence for a directional explanation, such as the data we have presented here, is evidence against a mentalising explanation. This dichotomy may be too sharp: directionality and perspective taking are not unrelated. Taking another individual's perspective must entail first following the direction of their gaze; or, in other words, directional effects may be a necessary pre-condition of perspective taking. Our results (and other results too) suggest that directional effects - which are a relevant input into any possible fast and efficient perspective taking - are indeed automatic and efficient. They just do not seem to necessarily lead to perspective taking.

If this speculation is correct, it may be important to distinguish automatic cognitive processes (i.e. those that are mandatory upon the perception of relevant inputs) and spontaneous ones (i.e. those that occur quickly and efficiently as and when needs arise). Our results - and results from other experimental paradigms (e.g. Freundlieb et al., 2016; Schneider et al., 2017) - are consistent with the interpretation that perspective taking is spontaneous but not automatic. Future experimental research could test this possibility directly.

## Acknowledgements

We would like to thank our colleagues Simon Kirby and Jennifer Culbertson, who provided insight and expertise that greatly assisted this research.

## References

Apperly, I. A. (2011). Mindreaders. Psychology Press. Apperly, I. A., \& Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? Psychological Review, 116(4), 953-70.
Baayen, R. H., \& Milin, P. (2010). Analyzing Reaction Times. International Journal of Psychological Research, 3(2), 12-28.

Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Butterfill, S. A., \& Apperly, I. A. (2013). How to Construct a Minimal Theory of Mind. Mind \& Language, 28(5), 606-637.
Cole, G. G., Atkinson, M., Le, A. T. D., \& Smith, D. T. (2016). Do humans spontaneously take the perspective of others? Acta Psychologica, 164, 165168.

Corballis, M. C. (2011). The Recursive Mind: The Origins of Human Language, Thought, and Civilization. Princeton, NJ: Princeton University Press.
Freundlieb, M., Kovács, Á. M., \& Sebanz, N. (2016). When do humans spontaneously adopt another's visuospatial perspective? Journal of Experimental Psychology. Human Perception and Performance, 42(3), 401-12.
Graziano, M. S. A. (2013). Consciousness and the Social Brain. Oxford University Press.
Hare, B., Call, J., \& Tomasello, M. (2001). Do chimpanzees know what conspecifics know? Animal Behaviour, 61(1), 139-151.
Heyes, C. (2014). Submentalizing: I Am Not Really Reading Your Mind. Perspectives on Psychological Science, 9(2), 131-143.
Kuznetsova, A., Brockhoff, P. B., \& Christensen, R. H. B. (2016). ImerTest: Tests in Linear Mixed Effects Models. R package version 2.0-30.
Marotta, A., Lupiáñez, J., Martella, D., \& Casagrande, M. (2012). Eye gaze versus arrows as spatial cues: Two qualitatively different modes of attentional selection. Journal of Experimental Psychology: Human Perception and Performance, 38(2), 326-335.
Peirce, J. (2010). PsychoPy - Psychology software for Python. Integration The Vlsi Journal.
Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., \& Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. Journal of Experimental Psychology: Human Perception and Performance, 36(5), 1255-1266.
Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., \& Heyes, C. (2014). Avatars and arrows: implicit mentalizing or domain-general processing? Journal of Experimental Psychology. Human Perception and Performance, 40(3), 929-37.
Schneider, D., Slaughter, V. P., \& Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. Cognition, 162, 27-31.
Scott, R. M., \& Baillargeon, R. (2017). Early False-Belief Understanding. Trends in Cognitive Sciences, 1-13.
Tomasello, M. (2008). Origins of Human Communication. Communication. MIT Press.
Tomasello, M. (2014). A Natural History of Human Thinking. 2014, 2(1), 1-5.
Whelan, R. (2008). Effective analysis of reaction time data. The Psychological Record, 58, 475-482.

# Creation of Spatial Mental Models with Figural Stimuli: Validation of the Emoji-based Spatial Integration Task 

Polly O’Rourke (porourke@casl.umd.edu)<br>Nick B. Pandža (npandza@umd.edu)<br>Susan Campbell (susanc@umd.edu)<br>Center for Advanced Study of Language<br>University of Maryland, $700552^{\text {nd }}$ Avenue<br>College Park, MD 20742 USA


#### Abstract

The current study examined a new spatial integration (SI) task, based on figural rather than linguistic stimuli, to measure the construct of mental modeling ability. Previous tasks conflated linguistic ability with mental modeling ability by requiring sentence processing, which may have contributed to mixed findings with respect to the relationship between mental model ability and working memory capacity (WMC). The figural spatial integration task produced the canonical continuity effect, such that discontinuous items had lower accuracy than continuous items. Furthermore, WMC and visuospatial ability predicted SI task performance, and both were stronger predictors for the continuous condition. The interactions between predictors and task conditions suggest reliance on heuristics and/or rehearsal during performance of the more difficult discontinuous items.


Keywords: Spatial integration, mental modeling, working memory capacity, spatial manipulation.

## Introduction

Mental models are abstract representations of a situation, derived from a narrative or some other form of input (Ehrlich \& Johnson-Laird, 1982). Successful creation of mental models contributes to logical thinking (e.g., Bell \& Johnson-Laird, 1998; Evans, Handley, Harper, \& JohnsonLaird, 1999) and spatial and temporal reasoning (e.g., Baguley \& Payne, 1999; Carreiras \& Santamaria, 1997; Roberts, 2000). It is also strongly connected to the ability to comprehend written or spoken narratives (Bower \& Morrow, 1990; de Vega, 1995; Radvansky \& Copeland, 2004).

The experimental task most commonly used to assess spatial mental model ability is the Spatial Integration (SI) task (Copeland \& Radvansky, 2007; Radvansky \& Copeland, 2004). In this task, participants are presented with a sequence of three sentences (one at a time), each describing the spatial relation of two of four objects. Immediately following this presentation, participants select from an array the picture which represents the correct spatial arrangement of the four items. There are two conditions referring to how the spatial information is presented in the learning phase: continuous and discontinuous. In the continuous condition, the second screen includes one item from the first screen and the third
screen includes one item from the second, enabling the participant to incrementally construct a mental model. In the discontinuous condition, the second screen and third screen are switched such that the second screen does not contain either of the items in the first screen but the third contains one item from each of the previous screens.

The use of sentence stimuli in this task is, however, problematic. First, task performance may reflect verbal rather than mental model abilities. Second, the processing demand associated with language comprehension may obfuscate the relationship between mental model ability and key underlying cognitive factors, like working memory capacity (WMC; Conway, Kane, Bunting, Hambrick, Wilhelm, \& Engle, 2005). Extant findings support the problematic nature of using sentence stimuli in this task. While several studies have found no relationship between identification accuracy in the SI task and WMC (Radvansky \& Copeland, 2001; 2004; Radvansky, Gibson \& McNerney, 2014), O’Rourke and Bunting (in press) found that when controlling for reading comprehension ability, WMC predicted accuracy in the discontinuous condition. They also found that when participants performed the SI task in their second language, second language proficiency alone predicted performance. As operating in L2 is widely known to absorb available WM resources, this finding and the finding for L1 indicate that variability related to language processing may obfuscate the relationship between WM and mental model creation.

Copeland and Radvansky (2007), in their study of mental model ability in aging adults, a population which generally has reduced WMC (Myerson, Emery, White, \& Hale, 2003), examined both verbal and figural versions of the task. They implemented the SI task with sentence stimuli describing the spatial configurations (Experiment 1), word stimuli appearing in the relevant spatial configurations (Experiment 2), and picture stimuli appearing in the configurations (Experiment 3). Aging adults did very poorly on both continuous and discontinuous conditions in Experiment 1, with performance on the discontinuous condition not differing significantly from chance. WMC (indexed by Operation Span; Turner \& Engle, 1989) predicted identification accuracy in the older participants, but not the young adult group. Performance on the continuous condition was improved in Experiments 2 and 3. Only in Experiment 3, the figural version, did aging adults perform above chance in the discontinuous
condition. WMC predicted performance in both age groups in Experiments 2 and 3. Furthermore, this finding suggests that the cognitive burden of language processing absorbs working memory resources required for successful performance of the SI task.
The goal of the current study was to validate a new figural version of the SI task and to examine the relationship between task performance and working memory capacity. Another potential source of variance in this task is spatial visualization ability, which reflects the ability to represent and manipulate parts of an image (Carroll, 1993). This ability may underpin performance of the figural version in particular as stimuli can be represented visually immediately, without the step of converting word/sentence stimuli into images. Spatial visualization ability will, therefore, also be included as a predictor in the analysis.

## Method

## Participants

A total of 161 ( 96 female) participants between the ages of 18 and $39(\mathrm{M}=20.32, \mathrm{SD}=1.67)$ with normal or corrected-to-normal vision were tested and retained for analysis in the current experiment. They were paid for their participation. Two additional participants were excluded for exhibiting a pattern of not following study directions across multiple tasks.

## Tasks

## Spatial Integration Task

The SI task (adapted from Copeland \& Radvansky, 2007) tests the ability to construct a mental model of the spatial arrangement of four items. In the learning phase, participants are presented with three screens, each containing two of four objects in particular spatial arrangements (see Figures 1 and 2). Items are presented in the continuous (see Figure 1) or discontinuous condition (in which the second and third screens in Figure 1 would be switched; see Figure 2). After the learning phase, participants must select from eight diagrams the one that correctly represents the spatial arrangement of all four objects in relation to one another (see Figures 1 and 2 for correct arrangement for the example item). The three screens in the learning phase are presented for 2 seconds per screen, while the test screen remains available until the participant responds. The task stimuli consisted of 80 emoji downloaded from the Emojione database (emojione.com), representing 20 sets of four semantically related emoji (e.g., fruits, vegetables, vehicles). Each task item was composed of one of the sets of four. Each item appeared once per stimulus list. Two forms of the test were created such that items were matched across conditions; a particular set of emoji appeared in one stimulus list in the continuous condition and in the other stimulus list in the discontinuous condition. As participants must choose the correct answer from eight options, chance performance is about $12 \%$.


Figure 1. Example of continuous item from the SI task with the full spatial arrangement of the four items. Participants can build a partial model immediately.


Figure 2. Example of corresponding discontinuous item from the SI task with the full spatial arrangement of the four items. Participants must wait to integrate the model based on information that comes on the third screen.

## Shapebuilder task

The Shapebuilder (SB) task is a complex visuospatial working memory measure (Atkins et al., 2014). In this task, participants are shown a series of shapes in a $4 \times 4$ grid, and they must recall the shape, color, and location of the series of the shapes in the correct presentation order. There are 26 items and the number of shapes in each item's sequence increases over the course of the task from two shapes (six items), to three shapes (nine items), and finally to four shapes (11 items). Points were earned for each shape in the sequence for which the location, shape, and color were correctly recalled. Partial credit was awarded if the location was correctly recalled. More points were awarded for longer sequences. See Atkins et al. (2014) for full scoring parameters and for WMC factor loadings alongside other working memory measures.

## Paper Folding task

The Paper Folding (PF) task used in the current study to measure spatial visualization ability was a computerized
test adapted from the ETS Kit of Factor-Referenced Tests (Ekstrom et al., 1976). Two forms were created with 20 items each. Items were ordered by increasing difficulty. As participants must choose the correct answer from five options, chance performance is $20 \%$.

## Procedure

The tasks pertinent to this study were administered as part of a larger battery of 18 behavioral tasks and surveys. The 18-task battery was administered in two sessions of three hours each with the opportunity for breaks between each task. Testing took place in a classroom-style computer lab. Written consent was obtained at the beginning of the first testing session. SI and PF were administered in session one and SB was administered in session two.

## Results

Data from six participants in the SB task is missing due to study attrition as they did not return for the second session. As a result, the sample size for analyses including SB is 155. We ran Wilcoxon Signed Rank tests comparing forms, and found no significant differences; therefore forms for SI and PF were collapsed in the correlational analysis. See Table 1 and 2 for descriptive statistics and correlations among measures, respectively.

Table 1. Task performance - Descriptive statistics

| Task | Mean | SD | Min | Max |
| :--- | ---: | ---: | ---: | ---: |
| SI_Con | .67 | .26 | .00 | 1.00 |
| SI_Discon | .52 | .22 | .00 | 1.00 |
| SB | 1711.52 | 466.53 | 755.00 | 2325.00 |
| PF | .74 | .18 | .15 | 1.00 |

Table 2. Correlations among measures. Numbers on the diagonal reflect average internal consistency.

| Task | SI_Con | SI_Discon | SB | PF |
| :--- | :---: | :---: | :---: | :---: |
| SI_Con | $(.75)$ |  |  |  |
| SI_Discon | $.50^{*}$ | $(.56)$ |  |  |
| SB | $.44^{*}$ | $.33^{*}$ | $(.73)$ |  |
| PF | $.60^{*}$ | $.37^{*}$ | $.43^{*}$ | $(.86)$ |
| $* p<001$ |  |  |  |  |

* $p<.001$

We conducted a logistic multilevel model (MLM, or mixed-effects model) on the binary individual trial-level accuracy data in order to generalize across participants and items and account for the fact that particular items were present in both conditions (Baayen et al., 2008; Linck \& Cunnings, 2015). Condition (continuous vs. discontinuous) was included as a fixed effect, nested within-item, as each item appeared in both conditions across the two forms.

The model predicts estimated log-odds (b) of a correct response on the SI task, from which we can derive the
change in odds and probability of accurate performance on the task. The independent variables included to explain variance in subject and item performance were Condition (Continuous, discontinuous), z-scored SB and PF , and the two-way interactions of Condition with SB and PF. Results of this modeling procedure are shown in Table 3, with the model baseline being the discontinuous condition.

Table 3. Logistic MLM results for SI item accuracy

| Fixed Effects | Estimate <br> $(\boldsymbol{b})$ | Odds <br> $(\exp (\boldsymbol{b}))$ | SE | $\boldsymbol{p}$ - <br> value |
| :--- | :---: | :---: | :---: | :---: |
| Intercept | 0.10 | 1.11 | 0.13 | .43 |
| Continuous | 0.18 | 2.17 | 0.08 | $<.001^{*}$ |
| SB | 0.22 | 1.25 | 0.09 | $.01^{*}$ |
| PF | 0.30 | 1.35 | 0.09 | $<.001^{*}$ |
| Con $\times$ SB | 0.16 | 1.18 | 0.10 | $.09^{\wedge}$ |
| Con $\times$ PF | 0.42 | 1.53 | 0.09 | $<.001^{*}$ |
| Random Effects |  | Variance | $\boldsymbol{S D}$ |  |
| Intercepts $\mid$ Subject | 0.47 | 0.69 |  |  |
| Intercepts \| Item | 0.22 | 0.47 |  |  |
| $* p<.05, \wedge p<.10$ |  |  |  |  |

The results show a main effect of Condition, confirming the effect of continuity for the SI task while controlling for the effects of SB and PF.

The results further show visuospatial WMC and spatial visualization (as measured by the SB and PF tasks, respectively) contribute independent variance to SI accuracy on both discontinuous and continuous items. Since SB and PF are z-scored and on the same scale, the sizes of the estimates can be directly compared. The effect for PF is slightly stronger than for SB on discontinuous items ( $b_{\mathrm{PF}}=0.30>b_{\mathrm{SB}}=0.22$ ) and the effect for PF is almost twice the size of the effect for SB on continuous items ( $b_{\text {PF }}$ $+b_{\mathrm{Con} \times \mathrm{PF}}=0.72>b_{\mathrm{SB}}+b_{\mathrm{Con} \times \mathrm{SB}}=0.38$ ).

The effect for SB is positive: as SB scores increase, so do the odds of a correct response on discontinuous items on SI. There is a marginal interaction of continuous $\times \mathrm{SB}$, suggesting that the effect of SB may be even stronger for continuous items than discontinuous (see Figure 3).


Figure 3. Depiction of SI accuracy regressed on Shapebuilder z-scores split by Condition. Shaded area around line represents 1 SE .

The effect for PF is also positive: as PF scores increase, so do the odds of a correct response on discontinuous SI items. Finally, there is a significant interaction of continuous $\times \mathrm{PF}$, indicating that, for each standard deviation increase on PF performance in our sample, the odds of a correct response are higher on continuous items than discontinuous items on the SI task (see Figure 4).


Figure 4. Depiction of SI accuracy regressed on Paper Folding z-scores split by Condition. Shaded area around line represents 1 SE .

## Discussion

The results show that young adults perform similarly on this figural, emoji-based version of the SI task to previously tested text-based versions of the task. Our mean accuracies of $67 \%$ for continuous items and $52 \%$ for discontinuous (both of which are far above chance performance of $12 \%$ ) are consistent with Copeland \& Radvansky (2007)'s findings using the standard, sentence based task ( $68 \%$ for continuous condition and $47 \%$ for discontinuous), and their figure task ( $73 \%$ for continuous and $53 \%$ for discontinuous). The fact that accuracy levels and continuity effects for our figural version of SI mirror extant findings provides evidence that this new version of the task performs similarly to the standard task. One limitation of this study is that we did not compare performance on our figure version to a standard text based version of the SI task.

Our emoji-based SI task extends the findings of Copeland and Radvansky (2007) in regards to the utility of a non-linguistic SI task and has several advantages over their instantiation. In our version of the SI task, the learning phase for each item was experimenter-paced such that participants saw each of the three screens for two seconds. When examined by Copeland, Radvansky and colleagues (Copeland \& Radvansky, 2007; Radvansky \& Copeland, 2001; 2004; Radvansky et al., 2014) the learning phase was self-paced such that participants had as long as they wanted for each training screen and reading times for each screen were dependent variables. The two-second time limit for the present task was ased on pilot data such that it represented a window within which most participants advanced to the next screen. While accuracy was similar to previous results with a self-paced learning phase (Copeland \& Radvansky, 2007), our experimenter-paced version of the task is easier to administer remotely, without a proctor, due to a more consistent task duration.

Another key methodological difference is that our SI task included pictures of real-world objects (e.g., coffee, glass) across different semantic sets (e.g., vegetables, vehicles), allowing greater generalizability of the items than simple geometric shapes (e.g., red square, green star). The more complex nature of the images could also have led to a lower probability of verbal rehearsal strategies, especially since the same four colors were used in every trial in the Copeland and Radvansky (2007) task; however, future research will be needed to test this claim.

Finally, the current study had greater power than Copeland and Radvansky (2007), in that there were more trials ( 20 versus 8 total, 10 versus 4 by condition), more participants ( 161 versus 60 , the latter split between two groups), and all trials were used for more powerful statistical analyses thanks to a multilevel model design.

The examination of the cognitive underpinnings of mental model ability showed that both WMC and spatial visualization ability predict performance on the SI task. Spatial visualization ability emerged as a slightly stronger effect than WMC. Interestingly, both predictors accounted
for more variance in the Continuous condition than the discontinuous condition.

In contrast to other studies using text based versions of the task (Copeland \& Radvansky, 2007, Exp. 1; Radvansky \& Copeland, 2001; 2004; Radvansky et al., 2014), the current study found evidence for WMC as a predictor of performance on the SI task in young adults. As previously noted, lack of effects using the sentence-based SI task could be due in part to the processing demands associated with converting linguistic representations into spatial representations. Furthermore, the interim spatial representations must be maintained while the next sentence is parsed into spatial information. While in figural form, the SI task is demanding on WM resources, eliminating the sentence as a conveyor of spatial information may have resulted in an increase in resources available for mental model creation.

Another possible reason for lack of effect in previous studies, particularly for the discontinuous condition, is that WMC may be a weaker predictor in the discontinuous condition, as shown by the marginal two-way interaction in the current study. This may be due to the "choke" factor whereby high WMC individuals start performing like low WMC individuals when they are under pressure (Sattizahn, Moser, \& Beilock, 2016; Wang \& Shah, 2014), such that WMC no longer predicts performance.
Strategy use may be another factor reducing the effect of individual WMC on performance. Wang and Shah (2014) note that when heuristics are available, people with high and low WM spans may perform similarly. It may be that all participants develop strategies in order to reduce the cognitive effort (Shah \& Oppenheimer, 2008) involved in determining the correct response in the discontinuous condition and, therefore, WMC would no longer predict performance. For example, in the discontinuous condition, after seeing the third screen in the learning phase, a participant may only partially incorporate the spatial arrangement such that he/she knows the positioning of two of the four images (e.g., top two images in the square). This information, though incomplete, may be enough to select the correct answer in the test phase. It may be possible in future iterations of the task to reduce the utility of heuristics via changes to the design. Specifically, the options in the test phase could be modified such that strategy use would be less likely to lead to a correct response.
Another possibility is that participants were more likely to engage in rehearsal in the discontinuous task. Rehearsal is a means of maintaining information in a short-term memory store, without any WM or executive involvement (Conway, Cowan, Bunting, Therriault, \& Minkoff, 2002). As such, if participants were more likely to use rehearsal in order to remember the spatial configurations in the discontinuous condition, then WMC would be a less effective predictor of performance. There are many strategies for preventing rehearsal during cognitive task performance (Cowan, 2008). One example is adding a secondary processing task (e.g., counting backwards) for participants to perform during the period in which information needs to be retained. Given that the retention
interval is fairly brief in this task, it is not clear that a secondary processing task would be effective. Another option might be further reducing the time for which the screens in the learning phase are presented to a less comfortable pace. This adjustment would likely have consequences, however, for the WMC demand. Additional testing is necessary to determine how researchers can prevent rehearsal in this task.

While WMC accounted for a significant amount of variability in SI performance, spatial visualization ability emerged as a stronger predictor. It is, perhaps, unsurprising that spatial visualization ability would predict performance on a spatial reasoning task like SI, particularly our figural version. Spatial visualization ability interacted with task condition such that its utility as a predictor was better in the continuous condition. This pattern is, of course, similar to the pattern observed for WMC, but with spatial visualization ability the effect was significant.

This finding supports the account that in the discontinuous condition, participants were more likely to not create true spatial mental models but rather to use heuristics or rehearsal in order to determine the correct answer at test. The case for rehearsal is particularly strong in that reduced role of spatial visualization ability in the discontinuous condition suggests that participants may not be creating visual representations. If that is the case, then verbal rehearsal is one way to perform the task. While the example of a strategy described above requires some level of visuospatial representation, there may be strategies, other than rehearsal, that do not.

Hitherto unexamined effects of individual variability in spatial visualization ability may have been another factor contributing to the mixed findings in the literature with respect to the contribution of WMC. In the previous textbased versions of the SI task, perhaps individuals with poorer spatial visualization ability had more difficulty transitioning from text-based representations to full visual representations, regardless of WMC, and therefore were unable to create a spatial mental model of the four items.

In conclusion, the current study validated a figure version of the SI task such that results from this task show performance levels and continuity effects consistent with previous studies. Given that this task is not sensitive to individual variability in language processing, or even native language, it can be used as a more pure measure of spatial mental model ability. This conclusion is supported by the finding that WMC predicted task performance in the figure version of the task, and language-related variability may have obscured this relationship in previous studies. Furthermore, we present evidence that spatial visualization ability is a significant predictor of task performance, and that while both WMC and spatial visualization ability predicted performance in both conditions, there was evidence suggesting the effect was stronger in the continuous condition. Future research will determine the source of this interaction.

## References

Atkins, S. M., Sprenger, A. M., Colflesh, G. J., Briner, T. L., Buchanan, J. B., Chavis, S. E., ... \& Harbison, J. I. (2014). Measuring Working Memory Is All Fun and Games. Experimental Psychology, 61, 417-438.
Baayen, H., Davidson, D., and Bates, D. (2008). Mixedeffects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59, 390-412. doi: 10.1016/j.jml.2007.12.005

Baguley, T., \& Payne, S.J. (1999). Recognition memory for sentences from spatial descriptions: A test of the episodic construction trace hypothesis. Memory \& Cognition, 27, 962-973.
Bell, V.A., \& Johnson-Laird, P.N. (1998). A model theory of modal reasoning. Cognitive Science, 22, 25-51.
Bower, G.H., \& Morrow, D.G. (1990). Mental models in narrative comprehension. Science, 247, 44-48.
Carreiras, M., \& Santamaria, C. (1997). Reasoning about relations: Spatial and nonspatial problems. Thinking and Reasoning, 3, 191-208.
Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge, MA: Cambridge University Press.
Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O. \& Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin \& Review, 12(5), 769-786.
Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., \& Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. Intelligence, 30(2), 163-183.
de Vega, M. (1995). Backward updating of mental models during continuous reading of narratives. Journal of Experimental Psychology: Learning, Memory \& Cognition, 21, 373-385.
Ekstrom, R. B., French, J. W., Harman, H. H., \& Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Ehrlich, K., \& Johnson-Laird, P. N. (1982). Spatial descriptions and referential continuity. Journal of Verbal Learning and Verbal Behavior, 21(3), 296-306.
Evans, J.St.B.T., Handley, S.J., Harper, C.N.J., \& JohnsonLaird, P.N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 25, 1495-1513.
Linck, J. A., and Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. Language Learning, 65, 185-207. doi: 10.1111/lang. 12117

Myerson, J., Emery, L., White, D. A., \& Hale, S. (2003). Effects of age, domain, and processing demands on
memory span: Evidence for differential decline. Aging, Neuropsychology, and Cognition, 10(1), 20-27.
O'Rourke, P. \& Bunting, M. (in press). The Cognitive Underpinnings of Mental Model Construction in L1 and L2. Quarterly Journal of Experimental Psychology.
Radvansky, G.A., \& Copeland, D.E. (2001). Working memory and situation model updating. Memory \& Cognition, 28, 1073-1080.
Radvansky, G. A., \& Copeland, D. E. (2004). Working memory span and situation model processing. The American journal of psychology, 191-213.
Radvansky, G.A., \& Dijkstra, K. (2007). Aging and situation model processing. Psychonomic Bulletin \& Review, 14, 1027-1042.
Radvansky, G. A., Gibson, B. S., \& MCNERNEY, M. W. (2014). Working memory, situation models, and synesthesia. The American journal of psychology, 127(3), 325-342.
Roberts, M.J. (2000). Strategies in relational inference. Thinking and Reasoning, 6, 1-26.
Sattizahn, J. R., Moser, J. S., \& Beilock, S. L. (2016). A Closer Look at Who "Chokes Under Pressure". Journal of Applied Research in Memory and Cognition, 5(4), 470-477.
Shah, A. K., \& Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. Psychological Bulletin, 134(2), 207-222.
Turner, M.L. \& Engle, R.W. (1989). Is working memory capacity task independent? Journal of Memory and Language, 28, 127-154.
Wang, Z., \& Shah, P. (2014). The effect of pressure on high-and low-working-memory students: An elaboration of the choking under pressure hypothesis. British Journal of Educational Psychology, 84(2), 226-238.

## Acknowledgements

This research is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government. Nothing in this article is intended to be and shall not be treated or construed as an endorsement or recommendation by the University of Maryland, United States Government, or the author of the product, process, or service that is the subject of this article. Correspondence concerning this article should be addressed to Polly O'Rourke, Center for Advanced Study of Language, University of Maryland, 7005 52 ${ }^{\text {nd }}$ Avenue, College Park, MD 20742 (e-mail: porourke@casl.umd.edu). Special thanks to Meredith Mislevy Hughes and Valerie Karuzis.

# A Domain-Independent Approach of Cognitive Appraisal Augmented by Higher Cognitive Layer of Ethical Reasoning 

Suman Ojha (Suman.Ojha@student.uts.edu.au)<br>Jonathan Vitale (Jonathan.Vitale@uts.edu.au)<br>Mary-Anne Williams (Mary-Anne.Williams@uts.edu.au)<br>University of Technology Sydney - Centre For Artificial Intelligence<br>Innovation and Enterprise Research Lab (The Magic Lab), 15 Broadway, Ultimo NSW 2007, Australia


#### Abstract

According to cognitive appraisal theory, emotion in an individual is the result of how a situation/event is evaluated by the individual. This evaluation has different outcomes among people and it is often suggested to be operationalised by a set of rules or beliefs acquired by the subject throughout development. Unfortunately, this view is particularly detrimental for computational applications of emotion appraisal. In fact, it requires providing a knowledge base that is particularly difficult to establish and manage, especially in systems designed for highly complex scenarios, such as social robots. In addition, according to appraisal theory, an individual might elicit more than one emotion at a time in reaction to an event. Hence, determining which emotional state should be attributed in relationship to a specific event is another critical issue not yet fully addressed by the available literature. In this work, we show that: (i) the cognitive appraisal process can be realised without a complex set of rules; instead, we propose that this process can be operationalised by knowing only the positive or negative perceived effect the event has on the subject, thus facilitating extensibility and integrability of the emotional system; (ii) the final emotional state to attribute in relation to a specific situation is better explained by ethical reasoning mechanisms. These hypotheses are supported by our experimental results. Therefore, this contribution is particularly significant to provide a more simple and generalisable explanation of cognitive appraisal theory and to promote the integration between theories of emotion and ethics studies, currently often neglected by the available literature.


Keywords: Cognitive appraisal theory; computational emotion model; emotion combination; ethics

## Introduction

The attribution of an emotional state to self or others can occur when a complex state of the organism is accompanied by variable degrees of awareness, often referred to as appraisal (Scherer, 2001). Two levels of appraisal can be distinguished (Lambie \& Marcel, 2002): a first-order phenomenological state and a conscious second-order awareness. Both states can be either self-directed (first-person perspective) or worlddirected (third person perspective) (Vitale, Williams, Johnston, \& Boccignone, 2014). The present work will be concerned in discussing the nature of the conscious second-order appraisal process, known as cognitive appraisal process of emotion.

Traditional literature in emotional processing studies suggests that this cognitive appraisal process may underlie the evaluation of a set of variables called appraisal variables (Ortony, Clore, \& Collins, 1990; Lazarus, 1991; Roseman, Spindel, \& Jose, 1990; Scherer, 2001). Appraisal variables can be understood as the criteria used to assess a situation in relation to emotion elicitation process. For example, in
appraisal theory of Ortony et al. (1990), core appraisal variables ${ }^{1}$ considered are desirability - which assesses how desirable an event is, praiseworthiness - which measures how praiseworthy the action of an agent is and appealingness which measures how appealing is the agent to the appraising individual. Appraisal theories suggest that individuals converge to an emotional state depending on the evaluation of these variables. This position is further supported by the majority of existing computational explanations of cognitive appraisal (Dias \& Paiva, 2005; El-Nasr, Yen, \& Ioerger, 2000; Velasquez, 1997). However, the proposed accounts offer limited perspectives addressing only domain specific situations and making use of knowledge shaped as a set of pre-defined rules (Dias \& Paiva, 2005; El-Nasr et al., 2000). Thus, (i) the available literature in cognitive appraisal theory currently does not provide a clear computational explanation for domain-independent cognitive appraisal mechanisms. This is a significant research problem for both cognitive science and computer science research communities; in fact, on one hand, having a computational theory of domain-independent cognitive appraisal mechanisms can assist cognitive science researchers in addressing open research gaps in emotional processing studies, and, on the other hand, this computational account can be more easily integrated in disparate intelligent systems without the need of defining a complex set of domain-dependent rules.

However, this is not the only limitation presented by currently available explanations of cognitive appraisal theory. According to cognitive appraisal theory of emotion, an event can elicit more than one emotions simultaneously with varying intensities (Ortony et al., 1990). Nevertheless, (ii) it is not clear yet what is the best strategy to select an emotional state for attribution following this appraisal process. This is again a significant research problem. In particular, having a mechanism able to determine the final optimal emotional state is a highly desirable feature for intelligent systems interacting with humans, such as social robots (Williams, 2012), since this is a necessary skill for being proficient in emotional intelligence (Mayer \& Salovey, 1993). For example, it has been widely documented that the appraised emotional state of an individual has direct impact on decision making and action selection (Isen \& Means, 1983; Loewenstein \& Lerner,

[^426]2003). Thus, without an appropriate mechanism able to determine the final optimal emotional state, the intelligent system cannot take socially acceptable and ethical actions (Vitale, Williams, \& Johnston, 2014).

This paper aims to present a computational model of emotion processing that adds a higher layer of cognition to appraisal mechanism. The significance of this paper is further increased by this novel approach going beyond the domain of emotion theories and embracing the strengths of ethical theories. Although, the literature includes previous studies suggesting interactions between theories of emotions and ethics (Callahan, 1988; Gaudine \& Thorne, 2001), to our knowledge, there are no computational explanations addressing the interactions between ethics and emotion processing mechanisms (Ojha \& Williams, 2016). Therefore, in this paper we aim to:
(i) Provide a computational model of cognitive appraisal of emotion able to elicit appropriate emotional states without the need of defining pre-determined rules, but rather by using a general domain-independent approach facilitating easy extensibility of the emotionally intelligent systems;
(ii) Provide a novel computational process inspired by ethical theories for the selection of the optimal emotional state among the elicited ones.

The offered outcomes will provide valuable insights to gather a better understanding on how integrating ethical theories in emotion processing mechanisms can improve existing computational models of emotions. This in turn will advance the understanding of the role of cognition in emotion.

## Computational Models of Cognitive Appraisal

Theories from cognitive science and psychology have been implemented in various computational models of cognition. In this section, we will present some of the computational models of emotions implementing cognitive appraisal theory of emotion that are related to our discussion and identify their current limitations.

The models available in literature use evaluation criteria called appraisal variables (Ortony et al., 1990; Lazarus, 1991; Roseman et al., 1990; Scherer, 2001) to assess or evaluate the events for the generation of emotion. The choice of appraisal variables depends on the appraisal theory used and also on the application of the model. One common limitation of the existing accounts is their heavy specificity to the considered application domain and the determination of the elicited emotional states by means of pre-defined rules (Dias \& Paiva, 2005; El-Nasr et al., 2000). This approach likely leads to low extensibility of the system.

One available account is Fuzzy Logic Adaptive Model of Emotions (FLAME), a fuzzy logic based computational model of emotion (El-Nasr et al., 2000) inspired by appraisal theories suggested by Ortony et al. (1990) and Roseman et al. (1990). The main strategy used by FLAME is the evaluation of $i f$-then rules in order to assess the considered appraisal
variables. As we already discussed, this approach leads to a particularly poor extensibility of the system, since adding a new rule would require to consequently revise and adapt the entire knowledge base.

EMotion and Adaptation (EMA) (Gratch \& Marsella, 2004; S. C. Marsella \& Gratch, 2009) borrows the ideas from the cognitive motivational appraisal theory of Lazarus (1991). It stands out from other existing computational models of emotion in that it is able to compute emotions irrespective of the experiment domain. However, this model is not able to achieve this only by using the perceived positivity or negativity of an event like our model, which will be discussed later.

Another related account is Fearnot AffecTIve Mind Architecture (FAtiMA), a computational model of emotion proposed by Dias and Paiva (2005). It is significantly inspired by appraisal theory of Ortony et al. (1990). FAtiMA considerably uses domain specific scenarios built on top of predefined appraisal rules in order to appraise the desired situation without clearly suggesting how to easily generalise the proposed appraisal mechanisms for different domains.

Beside not providing a valid and easy strategy to integrate the suggested computational model in disparate application domains, the available accounts do not offer an effective way to determine the final emotional state in response to an event in a specific situation. This is still an open research problem since most appraisal theories do not explain how this can be achieved (see, for example, Ortony et al. (1990); Scherer (2001)). Some strategies propose to select the emotional state exhibiting (i) highest intensity (Gratch \& Marsella, 2004) or driven by the higher motivational state (i.e. hunger, thirst, pain, and fatigue) (El-Nasr et al., 2000), whereas other strategies propose to (ii) blend the elicited intensities of multiple emotions in order to determine the final emotional state (see Reilly (2006) for more details on the strategy used). In the Evaluation section, we shall discuss why these approaches might not be desirable methods to reach to a final emotional state.

As previously discussed, an emotion processing model developed by using a rule-based approach is unlikely to offer easy extensibility and high integrability in disparate emotionally intelligent systems among different application domains. Thus, in this paper we provide computational mechanisms general enough to be used in different domains without the need of re-implementing or adapting the proposed model, but at the same time able to appraise the appropriate emotional states for the considered situation. In addition, we suggest to use ethical theories to determine the final emotional state among the ones elicited by the cognitive appraisal process. Determining this state is particularly important to drive socially acceptable behaviours.

## Hypotheses

Consider a social interaction between two subjects. In this work we will call sender the subject producing a behavioural
response directed to the other subject, which we will call receiver. Denote with $S_{\text {receiver }}^{(\mathbf{B}, \mathbf{C})}$ a value determining how negative or positive the behaviour $\mathbf{B}$ of the sender is perceived by the receiver in a given context $\mathbf{C}$. $S_{\text {receiver }}^{(\mathbf{B}, \mathbf{C})}$ is a plausible computational representation summarising within a single valanced value the somatovisceral reactions of the body to the given situation $(\mathbf{B}, \mathbf{C})$ following the first-order phenomenological stage of emotional processing (Bechara, Damasio, \& Damasio, 2000). As previously mentioned in the introduction of this paper, this work is not concerned with discussing the implementation of first-order phenomenological processes.

Denote with $\mathcal{C}\left(S_{\text {receiver }}^{(\mathbf{B}, \mathbf{C})}\right)$ a cognitive appraisal process able to appraise the intensities $\mathbf{I}=\left\{i_{e_{1}}, \ldots, i_{e_{n}}\right\}$ of a set of $n$ considered emotional states $\left\{e_{1}, \ldots, e_{n}\right\}$ given the first-order phenomenological reaction of the receiver $S_{\text {receiver }}^{(\mathbf{B}, \mathbf{C})}$. Thus, our first hypothesis is that:
Hypothesis 1 The value $S_{\text {receiver }}^{(\mathbf{B}, \mathbf{C})}$ is a sufficient information to perform a cognitive appraisal process $C$ able to elicit the intensities of the considered emotional states and consequently promoting the selection of a final emotional state resembling human cognitive appraisal.

Importantly, the value $S_{\text {receiver }}^{(\mathbf{B}, \mathbf{C})}$ is completely independent from other pre-existing values $\mathcal{S}$ already available by the system and concerning different behaviours and contexts. In other words, adding a new value $\mathcal{S}$ to our model, thus extending the knowledge of the system, will not require to adapt the pre-existing knowledge and it will not necessitate to modify the parameters of the computational model.

Denote with $\mathcal{E}\left(\mathbf{I}, \theta^{\text {ethics }}\right)$ and with $\mathcal{E}(\mathbf{I})$ two processes able to provide a final emotional state given the set of the elicited emotion intensities I realised by the cognitive appraisal process $\mathcal{C}$. $\mathcal{E}\left(\mathbf{I}, \theta^{\text {ethics }}\right)$ includes parameters operationalising ethical theories, whereas $\mathcal{E}(\mathbf{I})$ uses a generic strategy without considering the ethical dimension of the given situation. Therefore, our second hypothesis is that:
Hypothesis 2 The cognitive appraisal process augmented by ethical reasoning mechanisms $\mathcal{E}\left(\mathbf{I}, \theta^{\text {ethics }}\right)$ converge to more accurate emotional states compared to cognitive appraisal processes augmented by generic reasoning mechanisms $\mathcal{E}(\mathbf{I})$.

In the remainder of this paper we will offer the functional level description of our computational account and experimental results validating our hypotheses.

## Model Implementation

Cognitive Appraisal Process. The process of emotion generation in our computational model ${ }^{2}$ is shown in Figure 1. As mentioned earlier, when an event occurs, its appraisal (evaluation) is done by using a set of variables called appraisal variables. Ortony et al. (1990) state that these appraisal variables are computed based on the goals, standards and attitudes of

[^427]the individual. In the context of our computational model of emotion, if we denote these goals, standards and attitudes as an internal parameter $\theta^{\text {int }}$ and the perceived knowledge of the environment that the system receives when an event occurs as $K^{e n v}$, then a function for computing appraisal variable can be represented as:
\[

$$
\begin{equation*}
v_{i}=\mathcal{V}_{i}\left(K^{e n v}, \theta^{i n t}\right) \tag{1}
\end{equation*}
$$

\]

Which means that the quantitative value of an appraisal variable is the function of the event knowledge gathered from the environment ( $K^{e n v}$ ) and the internal goals, standards and attitudes $\left(\theta^{\text {int }}\right)$. This computation is done by an Appraisal Mechanism component, as shown in Figure 1. Each computed appraisal variable contributes in the generation of one or more emotions (Ortony et al., 1990) and helps in estimating the intensities of the considered emotions ${ }^{3}$.

The majority of available computational models of emotion compute $v_{i}$ by using domain-specific rule-based functions (Dias \& Paiva, 2005; El-Nasr et al., 2000; Velasquez, 1997). Because of this, when the application domain or input parameters ( $K^{e n v}$ ) change in those models, the internal representation of goals, standards and attitudes $\left(\theta^{\text {int }}\right)$ also needs to be changed. In our model, $K^{e n v}$ is modelled as a set of valanced scores $S$ providing an interpretation of the negative or positive connotation of the experienced events. Importantly, the scores $\mathcal{S}$ are completely independent from $\theta^{i n t}$. Thus, extending our model with new knowledge or adapting previous one will not necessitate to modify the model's parameters $\theta^{\text {int }}$. In this paper we will not provide implementation details and we consequently limit our contribution to this functional description, since this is sufficient for the validation of the proposed hypotheses. The detailed mechanism of computation of appraisal variables in our computational model can be found in another paper (Ojha \& Williams, 2017).


Figure 1: General Appraisal Mechanism.

Emotional State Selection. Next crucial step is determining the final emotional state of the system. Our proposition is that when more than one active emotions are generated, then a final emotional state is best determined by a higher cognitive layer of ethical reasoning.

Figure 2 shows more details of the Emotion Combination Mechanism included in Figure 1 and suggests the mechanism to determine the final emotional state for attribution. The emotions $e_{1}, e_{2}, e_{3}, \ldots$. and $e_{n}$ with respective intensities $i_{e_{1}}$,

[^428]

Figure 2: Ethical Emotion Combination Mechanism.
$i_{e_{2}}, i_{e_{3}}, \ldots$. and $i_{e_{n}}$ output from the Affect Generation component are processed by applying the concepts of deontological and consequentialist ethics (Hooker, 1996) in order to determine the final emotional state. Deontological ethics says that one should satisfy owns duties before making a choice of action/decision and consequentialist ethics says that one should consider the consequences to all the relevant parties before making a decision (Hooker, 1996). Functionally, our ethical emotion combination mechanism is shown in 2.

$$
\begin{equation*}
e_{\text {ethical }}=\mathcal{E}\left(\mathbf{I}, \theta^{\text {ethics }}\right) \tag{2}
\end{equation*}
$$

Where, $e_{\text {ethical }} \in\left\{e_{1}, e_{2}, e_{3}, \ldots, e_{n}\right\}$ is the final emotional state. $I$ is the set of emotion intensities and $\theta^{e t h i c s}$ represents ethical standards.

$$
\begin{gather*}
e_{\text {high }}=\mathcal{E}_{\text {high }}(\mathbf{I})  \tag{3}\\
e_{\text {blended }}=\mathcal{E}_{\text {blended }}(\mathbf{I}) \tag{4}
\end{gather*}
$$

Equations 3 and 4 represent the functions computing respectively the final emotional states for highest intensity approach and blended intensity approach, which were introduced earlier. Clearly, these functions only take the intensities of various emotions for the determination of the final emotional state. However, our model reaches to a final emotional state with the help of higher cognitive mechanism of ethical reasoning (as shown in 2).

The emotional responses of our computational model based on: (1) Highest Intensity Approach, (2) Intensity Blending Approach and (3) Ethical Reasoning Approach will be compared with emotion data obtained from human participants in the Evaluation section.

## Evaluation

In order to operationalise our model and to consequently validate our hypotheses, we designed two sets of web-based surveys requiring two tasks: an action scoring task and a mindreading task. In both the experimental conditions we provided a set of stories concerning social exchanges between two individuals, a sender and a receiver, as previously denoted.

Participants covering a broad set of countries were invited on Facebook or through mailing lists to take our surveys. The surveys were completely anonymous. We received a total of

153 responses (male $=82$, female $=71$ ). Importantly, the subjects were randomly attributed to either the action scoring task or the mind-reading task.

## Scenario Design

In order to avoid ad-hoc scenarios facilitating our model, we did not design the scenarios ourselves. Rather, we requested 4 naïve adults, without any knowledge about the objectives of the present research, to cooperate in designing six scenarios under the following conditions:

- The scenario shall include the interactions of two subjects, one of them denoted as sender and the other as receiver;
- A minimum of 5 and a maximum of 10 actions of the sender directed to the receiver describing a plausible social interaction between two persons shall be provided;
- At the beginning, each scenario shall provide the contextual information about the designed situation and the two considered subjects. Moreover, additional contextual information could be provided during the development of the described social exchanges, whenever this information is necessary to contextualise the remaining interactions;
- No contextual information suggesting the potential emotional state of the receiver shall be provided for individual interactions, with the exception of the contextual information provided at the beginning of the scenario.

The result of this process was a set of scenarios used during both the action scoring and the mind-reading tasks mentioned earlier. The scenarios included interactions between (1) two strangers (a male and a female) interacting on a bench of a park, (2) two close friends (both males) meeting at a beach, (3) a husband and a wife having an argument about forgetting the birthday, (4) an elderly woman affected by dementia and her nurse (both females) experiencing a distressful moment, (5) a guy having argument with his brother, and (6) an interaction between a customer of a café and a waiter (both males). In total, the scenarios included 48 social exchanges of the senders directed to the receivers.

## Action Scoring Task

The experimental subjects participating in the action scoring task were asked to guess, for each scenario, how positive or negative each social exchange performed by the sender would be perceived by the receiver in that specific context. The rating was based on 7-point Likert scale: Extremely Negative, Very Negative, Negative, Neither Negative Nor Positive, Positive, Very Positive, Extremely Positive. We numerically evaluated the responses by attributing a weight to each point of the scale (i.e. $-1,-0.66,-0.33,0,0.33,0.66$ and 1 respectively). We averaged the responses obtaining a value $S$ for each of the considered social exchanges $\mathbf{B}$ in the specific context $\mathbf{C}$. In this way we were able to provide the


Figure 3: Results of the experiments. (a) The cumulative rank-distances of the models' predictions from human assessment. (b) The rank-distances of the models' predictions from human assessment.
necessary input knowledge to our system (i.e. a set of numeric scores $S \in[-1,+1]$ ) and to consequently perform cognitive appraisal processes estimating the emotional state of the receivers in each considered scenario. Recall that this process did not require any changes to our computational model, which provides a valid domain-independent approach of cognitive appraisal process.

## Mind-Reading Task

In order to compare the emotional response of our computational model, we asked to the subjects participating in the mind-reading task to guess, for each interaction of the sender, what would have been the chances that the receiver would happen to be in a particular emotional state, based on the just happened interaction and the previously occurred social exchanges and contextual information. Therefore, for each of the eight considered emotional states the rating was based on 6-point Likert scale: Not at all, Very Low, Low, Medium, High and Very High. The additional rating "Not at all" was necessary to allow the participants to express no chances to attribute such emotional state to the receiver. We numerically evaluated the responses by attributing a weight to each point of the scale (i.e. $0.0,0.2,0.4,0.6,0.8$ and 1.0 respectively). Average score given by the participants to various emotions was calculated by performing the weighted average of the ratings.

## Results

Based on the results of the mind-reading task, the emotions for each interaction of each scenario were ranked from 1 to 8 , with the emotion having the highest average score ranked as 1 and the one with lowest score ranked as 8 .

We considered three strategies to computationally predict the final emotional state of each interaction: choosing the emotional state with highest intensity, blending the emotional intensities to determine the final emotional state as described
by Reilly (2006), and our suggested approach based on ethical reasoning. Each of these three strategies followed a common domain-independent cognitive appraisal process, as discussed in the model implementation section. We compared these computational predictions against the gathered human assessments (i.e. emotions ranked 1) by computing their rank-distances, suggesting how close the computational model was compared to human assessment. The results are summarised in Table 1.

Table 1: Descriptive statistics of the gathered results.

|  | Mean | Median | Std |
| :--- | :---: | :---: | :---: |
| High intensity | 2.4167 | 2 | 2.3232 |
| Blended emotion | 2.3125 | 2 | 2.0228 |
| Ethical reasoning | 2.0833 | 1 | 2.3140 |

In order to demonstrate that the proposed common stage of domain-independent cognitive appraisal was able to elicit emotional intensities similarly to human cognitive appraisal process (Hypothesis 1), we analysed the human responses of the mind-reading task. We noticed that for most of the considered interactions some of the emotions resulted in very close averaged scores. Therefore, given $\varepsilon=0.1$, for each interaction we counted the number of emotions having an average score of greater than or equal to the score of highest scored emotion minus $\varepsilon$ for that interaction. $\varepsilon$ was chosen to be equal to half of the score attributed to each point of the Likert scale (i.e. 0.2 ), thus being able to group emotions plausibly ranked with similar likelihood by most of the human assessors. The average number of similarly rated emotional states among all the 48 interactions was 3.2 , thus suggesting that on average human cognitive appraisal promoted 3 comparable emotional states to attribute to the receiver. From Figure 3a it is clear to see that for distances less than 2 ranks to the human assessment (i.e. predictions among the first 3 higher scored emo-
tions) our cognitive appraisal model was able to promote the selection (using all the three considered strategies) of approximately $70 \%$ of the emotional states plausibly attributed by humans participants to the receivers described in the mindreading task scenarios.

In addition, we can also observe that the cognitive process augmented by the proposed ethical reasoning mechanism converges to more accurate emotional states compared to the other investigated strategies (Hypothesis 2). Figure 3b further suggests that the proposed ethical reasoning mechanism reduces average rank-distances from human appraisal. Therefore, the present results support both the proposed hypotheses.

## Conclusion and Future Work

In this paper, we presented our computational model of emotion based on appraisal theory that is able to generate emotions using the expected degree of positivity or negativity associated with an action/event. This allowed our model to be completely independent of the application domain and efficiently appraise a situation for the elicitation of various emotions. In addition, our model adds a higher layer of cognition in the emotion mechanism by integrating an ethical reasoning capability for the determination of the final emotional state when more than one emotions are generated by the model. Experimental results support our first hypothesis proposing that cognitive appraisal is possible without prior domain knowledge and second hypothesis suggesting that ethical reasoning is a better strategy to explain human emotional state attribution process.

Yet, our computational model still has some room for improvement. For example, it is important to consider that people with different personality generate emotions in different ways. Thus, in the future, we aim to use the concept of personality and examine how the difference in personality makes difference in ethical standards and hence in emotion generation.

## References

Bechara, A., Damasio, H., \& Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. Cerebral cortex, 10(3), 295-307.
Callahan, S. (1988). The role of emotion in ethical decisionmaking. Hastings Center Report, 18(3), 9-14.
Dias, J., \& Paiva, A. (2005). Feeling and reasoning: A computational model for emotional characters. In Progress in artificial intelligence (p. 127-140). Springer.
El-Nasr, M. S., Yen, J., \& Ioerger, T. R. (2000). Flame - fuzzy logic adaptive model of emotions. Autonomous Agents and Multi-agent systems, 3(3), 219-257.
Gaudine, A., \& Thorne, L. (2001). Emotion and ethical decision-making in organizations. Journal of Business Ethics, 31(2), 175-187.
Gratch, J., \& Marsella, S. (2004). A domain-independent framework for modeling emotion. Cognitive Systems Research, 5(4), 269-306.

Hooker, J. (1996). Three kinds of ethics.
Isen, A. M., \& Means, B. (1983). The influence of positive affect on decision-making strategy. Social cognition, 2(1), 18-31.
Lambie, J. A., \& Marcel, A. J. (2002). Consciousness and the varieties of emotion experience: a theoretical framework. Psychological review, 109(2), 219.
Lazarus, R. S. (1991). Emotion and adaptation. New York: Oxford University Press.
Loewenstein, G., \& Lerner, J. S. (2003). The role of affect in decision making. Handbook of affective science, 619(642), 3.

Marsella, S., Gratch, J., \& Petta, P. (2010). Computational models of emotion. In A blueprint for an affectively competent agent: Cross-fertilization between emotion psychology, affective neuroscience, and affective computing (p. 2146). Oxford: Oxford University Press.

Marsella, S. C., \& Gratch, J. (2009). Ema: A process model of appraisal dynamics. Cognitive Systems Research, 10(1), 70-90.
Mayer, J. D., \& Salovey, P. (1993). The intelligence of emotional intelligence. Intelligence, 17(4), 433-442.
Ojha, S., \& Williams, M.-A. (2016). Ethically-guided emotional responses for social robots: Should i be angry? In International conference on social robotics (p. 233-242). Springer.
Ojha, S., \& Williams, M.-A. (2017). Emotional appraisal : A computational perspective. In Annual conference on advances in cognitive systems.
Ortony, A., Clore, G. L., \& Collins, A. (1990). The cognitive structure of emotions. Cambridge University Press.
Reilly, W. S. N. (2006). Modeling what happens between emotional antecedents and emotional consequents. In Eighteenth european meeting on cybernetics and systems research (p. 607-612). Vienna, Austria: Austrian Society for Cybernetic Studies.
Roseman, I. J., Spindel, M. S., \& Jose, P. E. (1990). Appraisals of emotion-eliciting events: Testing a theory of discrete emotions. Journal of Personality and Social Psychology, 59(5), 899.
Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. Appraisal processes in emotion: Theory, methods, research, 92, 120.
Velasquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In Aaai/iaai (p. 10-15).
Vitale, J., Williams, M.-A., \& Johnston, B. (2014, October). Socially impaired robots: Human social disorders and robots' socio-emotional intelligence. In 6th international conference on social robotics (p. 350-359).
Vitale, J., Williams, M.-A., Johnston, B., \& Boccignone, G. (2014). Affective facial expression processing via simulation: A probabilistic model. Biologically Inspired Cognitive Architectures, 10, 30-41.
Williams, M.-A. (2012). Robot social intelligence. In Social robotics (pp. 45-55). Springer.

# Tracking the temporal course of counterfactual understanding 

Isabel Orenes<br>UNED<br>\author{ Juan Antonio Garcia Madruga<br><br>UNED, Spain }

Isabel Gómez-Veiga
UNED, Spain
Orlando Espino
Universidad de La Laguna

## Ruth Byrne

TRINITY COLLEGE DUBLIN, UNIVERSITY OF DUBLIN, IRELAND


#### Abstract

This paper explores the dual meaning of counterfactual conditionals, such as 'if there had been gloves, then there would have been scarves', by tracking the temporal course to envisage the possibility corresponding to the conjecture 'there were gloves and there were scarves' and the presupposed facts, 'there were no gloves and there were no scarves'. To test this, we used the visual world paradigm, in which counterfactual and indicative conditionals were heard while four images corresponding to the conjecture, such as an image of gloves and scarves, and the presupposed facts, such as an image of no gloves and no scarves, and two distractors were shown on the screen and eye movements were monitored. We found that people looked at the affirmative image in the indicative conditional, and both types of images (affirmative and negative) in the counterfactual conditional. Results support the dual meaning of counterfactuals.


# Investigating Sensitivity to Shared Information and Personal Experience in Children's Use of Majority Information 

Kay Otsubo (kay.otsubo@mail.utoronto.ca)<br>Department of Psychology, University of Toronto<br>Toronto, ON M5S 3G3 CAN

Andrew Whalen (awhalen@roslin.ed.ac.uk)<br>The Roslin Institute, University of Edinburgh<br>Midlothian, SCT EH25 9RG UK

Daphna Buchsbaum (buchsbaum@psych.utoronto.ca)
Department of Psychology, University of Toronto
Toronto, ON M5S 3G3 CAN


#### Abstract

When learning from others, rather than simply following the majority's opinion, we need to accurately evaluate the quality of the information both the majority and the minority provide, and integrate that information with our own personal experience. This is especially true when the majority's opinion is based on lower quality information, because they shared the same evidence rather than collecting evidence independently. Previous work demonstrated that adults are sensitive to the quality of the majority's information, consistent with the predictions of a Bayesian rational model (Whalen, Griffiths, \& Buchsbaum, in press). In two behavioural experiments, we investigated how preschoolers combine testimony from a majority that conflicts with a minority or with the child's own personal evidence. Unlike adults, children over-relied on the majority when given only testimony. However, when also given their own conflicting evidence, children relied significantly less on the majority and over-relied on their own evidence. These findings help explain why children may follow the majority at times, but in others trust their own judgements.


Keywords: Selective Trust; Conformity Bias; Children; Statistical Dependency; Bayesian Modeling; Social Learning

## Introduction

Learning from others is a valuable strategy to use when encountering uncertainty. Human's use of social information is thought to underlie our ability to live in almost every known environment, and to underpin the evolution of human culture (Boyd \& Richerson, 1985). The information we receive from others is also integrated with the information we gather ourselves through personal experience. Rather than blindly following what others say, it is important we evaluate the information we receive by understanding how others may have formed their opinions, especially when there is disagreement amongst individuals or when their testimony conflicts with our own personal observations.

Evaluating informant testimony is particularly important when the testimony is based on shared information. For example, imagine you are reading four restaurant reviews. In the first two reviews, the reviewers independently visited the restaurant and ordered different dishes, and they both
recommended the restaurant. In the third and fourth reviews, the reviewers went to the restaurant together and shared a single dish, and then both did not recommend the restaurant. Which set of reviews should you trust?

From a pure numbers standpoint, the number of positive and negative reviews is equal. However, the first and second reviews may provide additional information about what an average experience (your experience) at the restaurant would be like. The third and fourth reviews provide less information because the reviewers' shared experience makes their responses statistically dependent on each other. For instance, if the two reviewers shared an unusually salty dish they are both likely to write negative reviews and as such, given the third review, the fourth review provides no new information. Thus, being aware of this statistical dependence will help social learners avoid the mistake of placing trust in a group based on just their number of opinions.

This ability to assess the quality of information being provided is especially important for young children, who are learning much about the world from the testimony of others (e.g., Mills, 2013). Whether young children can use statistical dependency to evaluate testimony quality is an interesting question because much of our social learning occurs during early childhood, when the ability to understand mental states is still developing (Wellman, Cross, \& Watson, 2001). The ability to assess quality of information may require a complex form of "theory of mind" that goes beyond simply copying the majority. To accurately assess the quality of information, children must consider not only the testimony each person gave, but also the unseen information leading to that testimony, and how that information was gathered.

In addition, one's own personal experience can also conflict with what others say, and must be integrated with the testimony received. Imagine that you have a negative experience at that restaurant, and are debating whether to go there again. If there are enough other positive reviews, you may be willing to disregard your own judgement, and give the restaurant another chance. Can children evaluate the quality of their own information relative to a conflicting majority in a similar manner?

In two experiments, we examined how 4- and 5-year-old children evaluated their own private information and the information they received from informants who either shared a piece of evidence or collected evidence independently. Specifically, we investigated whether children can distinguish the quality of information provided by multiple informants and exhibit a sensitivity to shared data (as suggested by Hu et al., 2015) or if they merely conform to the majority (Corriveau, Fusaro, \& Harris, 2009). We then compared children's performance to that of adults' on a similar task (Whalen et al., in press), and the predictions of a Bayesian rational model to understand the extent to which children may conform to the majority despite the amount of information the majority provided.

In Experiment 1, we investigated whether children were sensitive to evidence being shared by a majority group when a single dissenter with independent evidence was present. We found that children were biased towards following a majority opinion and were not sensitive to statistical dependencies between informants. In Experiment 2, we highlighted the source of informants' knowledge by providing children with their own private evidence that conflicted with the majority. We found that given conflicting personal evidence, children no longer followed the majority and instead sided with their own evidence regardless of the quality of the majority's information. Compared to both adults and to the rational model, children were not sensitive to dependency, trusting statistically dependent informants more in Experiment 1, and placing more weight on their own evidence in Experiment 2.

## Background

Previous work by Whalen et al. (in press) demonstrated that adults are sensitive to statistical dependency between informants. Participants correctly rated that an option was more likely when it was endorsed by a majority group with independent evidence than a group with shared evidence (see Figure 1(b) for results). Adults also integrated their own evidence with testimony, appropriately demonstrating no bias towards their own evidence when it conflicted with the majority endorsement. In particular, they endorsed the majority opinion when the group had a higher quality of information than provided by their own personal evidence.

These findings were consistent with a Bayesian model of social learning which captures how an idealized learner might learn from multiple informants with shared information. The model illustrates that conforming to the majority is rational when the majority has a greater quality of information because (like our independent restaurant reviewers) each member contributed additional independent information. Thus, although in some cases adults disregard their own evidence and favour the majority, this may be a product of rationally integrating the two sources of information and assessing their quality, and not a bias towards the majority.

In this paper, we investigated whether 4- and 5-year-old children could assess the information quality provided by a majority when it conflicted with the information of a minority or with the child's own personal evidence. At the age of 4,
children already start to implement strategies in choosing who to listen to by selectively trusting informants, for instance by preferring those who are knowledgeable or accurate (Koenig \& Harris, 2005), or experts in the field (e.g., Kushnir, Vredenburgh, \& Schneider, 2013). However, the current literature is unclear on whether children value conformity or information quality during social learning.

Previous studies argued that children value a consensus even when it conflicts with the child's own perception. For instance, children sometimes followed the majority even when they understood and identified the endorsement of the majority to be incorrect (Corriveau \& Harris, 2010; Haun \& Tomasello, 2011). Using the Asch (1956) paradigm, children were observed to conform especially when answering in public in front of their peers (Haun \& Tomasello, 2011). At the age of four, children are already capable of recognizing a consensus and conforming to them even in ambiguous tasks such as labelling a novel object (Corriveau et al., 2009). These findings then suggest that children may have a bias to conform to a majority, even when the conflicting information comes from their own perception.

On the other hand, some studies have argued that children do exhibit the ability to evaluate the quality of information they receive from multiple informants (Hu et al., 2015). Hu and colleagues (2015) found that, when given testimony from two groups, children preferred the group with the highest quality of knowledge - favouring the group that received direct knowledge via visual perception over those who received indirect knowledge via hearsay. However, when group sizes were not equal, children preferred the group with the most members, even if the members of the larger group had only received hearsay.

Additional work has shown that children avoid a conformity bias if the majority group is proven to be unsuccessful in reaching an apparent goal (Wilks et al., 2014), provide implausible functions for a novel object (Schillaci \& Kelemen, 2014), or have lower expertise than the minority (Burdett et al., 2016). These findings then suggest that, at least in some cases, children have a preference for informants with a greater quality of knowledge, rather than having a preference for the majority per se.

Preschoolers also demonstrate the ability to integrate a single informant's testimony with their own observations. When the two sources conflict, preschoolers acknowledge the confidence and statistical data provided by an informant to assess causal relationships of novel toys (Bridgers et al., 2015), and acknowledge an informant's awareness for appearance-reality when considering their own perception as misleading (Lane et al., 2014). These findings suggest that children can integrate both sources of information, which contrasts with previous theoretic models that emphasized reliance on only social learning (e.g., Rendell, Fogarty, \& Laland, 2010; but see Perreault, Moya, \& Boyd, 2012).

Therefore, whether children can appropriately integrate the quality of informants' knowledge given a majority and conflicting information from either a minority or from personal observation is an open question. However, it is not
always obvious how these different sources of information should normatively be integrated.

## Bayesian Model of Learning from Independent and Dependent Informants

To further understand how an individual can combine the information they receive from testimony and personal evidence, we followed the Bayesian model developed by Whalen et al. (in press) which captures how an idealized learner would integrate information provided by groups with different sources of data - shared or independent - with personal evidence. In this model, a learner collects personal evidence about the state of the world, $e$, and receives testimony from $n$ informants, $t_{1}, \ldots, t_{n}$ who collect their own evidence about the state of the world, $d$. Learners evaluate a potential hypothesis, $h$, using Bayes' rule,

$$
\begin{equation*}
p\left(h \mid e, t_{1}, \ldots, t_{n}\right) \propto p\left(t_{1}, \ldots, t_{n} \mid h\right) p(e \mid h) p(h) \tag{1}
\end{equation*}
$$

where $p\left(h \mid e, t_{1}, \ldots, t_{n}\right)$ is the posterior probability of $h$, the probability that a hypothesis about the state of the world is true given the personal evidence and testimony, while $p(h)$ is the prior probability of $h$, the probability the hypothesis is true before any evidence is given. Finally, $p(e \mid h)$ is the probability of getting that evidence given the hypothesis, and $p\left(t_{1}, \ldots, t_{n} \mid h\right)$ is the probability of getting that testimony.

When multiple informants provide independent testimony, the probability of a series of testimony is equivalent to the product of the probability of each individual testimony:

$$
\begin{equation*}
p\left(t_{1}, \ldots, t_{n} \mid h\right)=\prod_{i=1}^{n} p\left(t_{i} \mid h\right) \tag{2}
\end{equation*}
$$

The testimony of each informant is based on their private data $d_{i}$, so $p\left(t_{i} \mid h\right)$ is obtained by marginalizing over $d_{i}$ :

$$
\begin{equation*}
p\left(t_{i} \mid h\right)=\sum_{d_{i}} p\left(d_{i} \mid h\right) p\left(t_{i} \mid d_{i}\right) \tag{3}
\end{equation*}
$$

where $p\left(t_{i} \mid d_{i}\right)$ is the probability that the informant produces testimony $t_{i}$ after observing $d_{i}$. On the other hand, when multiple informants base their testimony on shared private data, denoted as $d^{\prime}$, the probability of a series of testimony is obtained by marginalizing over the shared private data:

$$
\begin{equation*}
p\left(t_{1}, \ldots, t_{n} \mid h\right)=\sum_{d^{\prime}} p\left(d^{\prime} \mid h\right) \prod_{i} p\left(t_{i} \mid d^{\prime}\right) . \tag{4}
\end{equation*}
$$

In both cases, we assume that informants give testimony in support of a hypothesis proportional to the product of the informant's evidence given the hypothesis and the prior probability, $\quad p\left(t_{i}=h_{i} \mid d_{i}\right) \propto p\left(d_{i} \mid h_{i}\right) p\left(h_{i}\right) \quad$ (for more information, see Whalen et al., in press). This Bayesian model illustrates that, in many cases, conforming to the majority is rational when the majority collects independent
evidence, increasing their quality of information (see Figure 1(a) for example predictions based on our experiment task). In addition, this model accurately predicted the performance of adults in our experimental task suggesting that adults integrate both sources of information rationally (see Figure 1(b) for adult performance).

Following the approach of Whalen et al. (in press), we ran two behavioural experiments that examined how children evaluated the information they were provided by a majority group with shared or independent evidence, along with either a dissenting informant or conflicting private evidence.

## Experiment 1: Dissenting Informant

In Experiment 1, children were shown a video about two jars with differing proportions of red and yellow balls and were asked to guess which jar was being sampled from, given the testimony of three friends who received a ball from the chosen jar. The first two informants endorsed the same jar and made up the majority group while the third informant dissented and endorsed the opposing jar. Children were randomly assigned to either the Shared condition, where the majority shared one ball, or the Independent condition, where majority members each received their own ball.

Our model predicts that a rational learner would choose the jar endorsed by the majority only when each member collected independent evidence, but be at chance when the majority shared one piece of evidence, Figure 1(a). On the other hand, if children have a conformity bias, we should expect them to pick the majority's jar in both conditions.

## Methods

Participants A total of 29 preschoolers (female $=17$, male $=$ 12; mean age $=4$ years 11 months; range $=49-71$ months) were recruited either through local museums or in lab. They were randomly assigned to one of two conditions: The Independent condition $(n=14)$ or the Shared condition ( $n=$ 15). An additional 6 children were excluded due to atypical development (1), provided ambiguous answers (1), did not provide an answer (2), or experimenter error (2).

Procedure In this experiment, children were shown a video on a laptop, where an experimenter introduced two jars comprised of coloured balls - one with mostly yellow balls, a few red balls, and one green ball and one with mostly red balls, a few yellow balls, and one white ball, and introduced her three adult friends. The experimenter explained that she would pour just one of the two jars into her bag and give each of her friends a ball from the bag. Each of her friends would then tell her which of the two jars they thought she picked.

Once the experimenter filled the bag with one of the two jars, she used a cup to randomly scoop a ball from the bag to hand to each of her friends. After looking inside the cup, the informants provided testimony as to which jar they thought the bag was filled from, either the jar with mostly red or mostly yellow balls in it. The first two informants always endorsed the same jar and made up the majority group, while the last informant always chose the opposite jar. The jar

## $\square$ Shared $\square$ Independent



Figure 1: Probability of endorsing the majority opinion in both experiments as predicted by the Bayesian model (a), the average rating of endorsing the majority in both experiments as performed by adults (b) from Whalen et al. (in press), and the percentage endorsing the majority in both experiments as performed by children (c) in the current study.
endorsed by the majority and the actor playing the minority informant were counterbalanced.

In the Independent condition, informants were in the room one at a time and were each given their own randomly sampled ball to view. Each informant stated, for instance, "I looked at the ball and I think that the bag has mostly red balls in it." In the Shared condition, all three informants were present in the room and the first two informants shared a single randomly sampled ball. After providing a testimony, the first informant was asked to pass the same cup to the second informant who then agreed with the first, e.g. "I looked at the ball and thought about what my friend said. I agree with Jessie. I think that the bag has mostly red balls in it." While the third informant received a different random ball and disagreed with the rest, e.g. "I looked at the ball and thought about what my friends said. I disagree with Jessie and Sarah. I think that the bag has mostly yellow balls in it."

Once the video was completed, the on-site experimenter, a different person than the one in the video, reminded the child which jar each informant endorsed and if they saw the same or a different ball as the previous informant, and that all the balls came from just one jar. Finally, she asked a forcedchoice question of which jar the child thought the bag was filled from, either the jar with mostly red or yellow balls in it. The order in which the jars were stated was randomized.

## Results

Each child was given a score of 0 or 1 , with 1 as agreeing with the majority and 0 as disagreeing. Results are shown in Figure 1(c). Overall, children chose the jar endorsed by the majority significantly more often than chance, regardless of how the members of the majority collected their information (binomial test, 21 out of 29 endorsed the majority, $p=0.024$ ).

We analyzed the differences between the conditions using a Fisher's exact test. The difference between the Independent and Shared condition was not significant (10 out of 14 endorsed the majority in the Independent condition, 11 out of 15 in the Shared condition, $p=1$ ). Children chose the majority's jar equally often when the majority had higher quality independent information and lower quality shared information. Finally, compared to adults and to our model,
children appeared to place more weight on the statistically dependent testimony in the Shared condition.

## Discussion

Unlike our model predictions and adults' performance, children were not able to appropriately evaluate the quality of information in an informant's testimony. When two informants received the same ball and gave the same testimony, children over-weighed the majority's shared information relative to the dissenter's independent information. These results support previous findings by Corriveau et al. (2009) who found that children conform to a majority when faced with an ambiguous decision.

Based on the results of Experiment 1, we wanted to identify ways to help children avoid relying on the majority and instead, evaluate which group has the greater quality of information. To do this, we highlighted the independent nature of the minority information by having the child receive private evidence that conflicted with the majority testimony which mimicked many real-world scenarios where our own private experience conflicts with testimony. If the child then has to integrate social learning with personal observation, this may help identify the source of knowledge each individual has and overcome a conformity bias.

## Experiment 2: Own Ball

In Experiment 1, children were making a decision based on testimony alone. However, in most real-world cases, we take in the information that others provide us and evaluate it with our own information. Therefore, in Experiment 2, children no longer saw a minority group, and instead were given their own ball from the bag that conflicted with the testimony. For example, if the informants all endorsed the jar with mostly red balls, the child received a yellow ball from the bag.

As predicted by the Bayesian model, children should choose the jar endorsed by the majority when the members independently collected data and have more information than provided by the child's own single piece of evidence. If, however, the members of the majority shared a single piece of evidence, children should endorse the majority at chance, as the child's own evidence would be as reliable as the
majority's. Similar to Experiment 1, if children present a conformity bias, they will follow the majority regardless of the quality of information provided by the group.

## Methods

Participants A total of 52 preschoolers (female $=24$, male $=$ 28; mean age $=4$ years 11 months; range $=48-71$ months) were recruited through local museums and daycares, or in lab. They were randomly assigned to one of two conditions: The Independent condition ( $n=26$ ) or the Shared condition ( $n=$ 26). An additional 14 children were excluded due to experimenter error (10), previous participation in Experiment 1 (1), inattentiveness (1) and ambiguous answers (2).

Procedure Experiment 2 had the same jars and actors, and similar sampling procedures. However, all three informants endorsed the same jar and made up the majority group for both conditions. As in Experiment 1, children were randomly assigned to either the Independent or Shared condition.

In the Independent condition, each informant received their own distinct randomly sampled ball and provided their testimony in the room one at a time. In the Shared condition, all three informants were present in the room and shared a single randomly sampled ball and provided testimony agreeing with the previous informants.

After the video ended, the on-site experimenter reminded the child which informant endorsed which jar, if they looked at the same or different ball as the previous informants, and that all the balls came from just one jar. In this experiment, all the informants endorsed only one jar. Next, the on-site experimenter brought out an identical bag and stated that it was the same bag from the video containing the same balls. Similar to the experimenter in the video, she used a plastic cup to give the child their own ball from the bag. The on-site experimenter pretended to scoop up a ball at random, but in fact the child always received a ball that was a different colour from the majority testimony. After the child looked inside the cup, the on-site experimenter asked which jar they thought all the balls came from, as in Experiment 1.

## Results

Results are shown in Figure 1(c). Overall, children chose the jar endorsed by the informants below chance regardless of how the informants collected their information (binomial test, 11 out of 52 endorsed the majority, $p<0.001$ ). Similar to Experiment 1, we found that children did not choose the informant's jar more in the Independent condition compared to the Shared condition (5 out of 26 endorsed the majority in the Independent condition, 6 out of 26 in the Shared condition, $p=1$, Fisher's exact test). In both conditions, children weighed their own evidence more, compared to both adults and the predictions of the Bayesian model.

## Discussion

We found that in Experiment 2, children relied heavily on their own evidence and chose the jar consistent with their own ball regardless of whether the majority collected
independent or shared evidence. As in Experiment 1, we found no significant difference between the Independent and Shared conditions, suggesting that children were not sensitive to the statistical dependency. These results support previous findings suggesting that children may rely on their own evidence that they personally collected over the evidence collected by others (Kushnir \& Gopnik, 2005; Kushnir, Wellman, \& Gelman, 2009).

## General Discussion

We investigated how children weighed the value of information they received from multiple individuals and their own personal evidence. We compared the performance of 4and 5-year-old children to the performance of adults on a similar task (Whalen et al. in press) and to the predictions of a Bayesian rational model. Experiment 1 showed that children were not sensitive to the shared information of a group and were instead following the majority. Experiment 2 demonstrated that children would no longer use the strategy of conforming to the majority if they themselves collected conflicting evidence and instead, relied on their own evidence, regardless of the quality of the majority's information. Therefore, compared to adults, children applied a different strategy in the integration of information.

Children's apparent conformity bias is consistent with previous findings that argued that children prefer to rely on the majority (e.g., Corriveau et al., 2009; Haun, Rekers, \& Tomasello, 2012; Haun \& Tomasello, 2011). Children may exhibit this reliance on the majority because it is often a useful and reliable social learning strategy (Haun et al., 2014). After all, the majority made their choices for a reason.

However, in the presence of children's own conflicting evidence, a conformity bias was no longer present. Although previous work has suggested that children may present a conformity bias even in the face of conflicting direct perception (Corriveau \& Harris, 2010; Haun \& Tomasello, 2011), it is important to note that in those studies, a majority of children still favoured their own evidence over the testimony of the majority.

This bias to rely on personal evidence over the evidence collected by others has previously been observed in the causal domain as a self-agency bias, especially when the evidence seemed to be ambiguous or probabilistic (Kushnir et al., 2009). Kushnir and Gopnik (2005) discovered that children would weigh their own causal interventions more heavily than the causal interventions of others. They suggested that children had this bias because they viewed their own actions to be more controlled and reliable and less likely to be confounded than those of other individuals. However, this bias to one's own evidence has not yet been observed in a non-causal domain like that of our current study. Similarly, children in our study might have considered their direct perception of their own ball to be more reliable than the information they received from the informants' testimony.

One possible explanation for the presence of both a bias towards conformity and towards personal evidence in this task is that preschoolers learning how to integrate
information have yet to develop more complex aspects of theory of mind (e.g., Gweon et al., 2012). At this age, children might have had difficulty reasoning about how the informants generated their testimony based on the evidence that they likely received. In other words, 4- and 5-year-old children know that people can have beliefs, but may have difficulty in knowing how these people came to believe something. As a consequence, children might rely on the number of endorsements given rather than on their quality, leading to the appearance of a conformity bias. On the other hand, children were likely confident in what they themselves saw which appeared as a bias towards their own evidence.

Future work should investigate whether children have difficulty inferring the evidence informants likely received based on their testimony, by testing how children respond when they can observe this evidence directly, for instance by presenting it in clear cups. We expect that children would then compare the amount of evidence between the majority and minority group rather than the number of endorsements. If children can identify the statistical dependency when the evidence is visible, they should no longer demonstrate a conformity or a personal evidence bias.

In addition, in ongoing follow-up studies, the child's evidence is presented on-screen within the video rather than performed live to equate saliency. If children still present a bias towards their own evidence, the salience from a live performance as a reason for this bias can be ruled out.

Taken together, our findings suggest that children implement a different social learning mechanism than adults and our Bayesian model. When integrating testimony alone, children over-weighed the quality of information provided by the majority. On the other hand, when the child was given their own conflicting evidence, children under-weighed the quality of information provided by the majority and relied on their own perception. Thus, unlike adults, children require further development in their social learning and perhaps their reasoning of mental states to avoid biases and become sensitive to statistical dependency.

## Acknowledgments

We are grateful for all the families for their participation, and to the Computational Cognitive Development lab for their help with data collection, especially Nafisa Bhuiyan and Elina Kadriu, and Tom Griffiths for helpful discussions. This work was supported by the Natural Science and Engineering Research Council of Canada and the Canada Foundation of Innovation.

## References

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological Monographs: General and Applied, 70(9), 1-70.
Boyd, R., \& Richerson, P. J. (1985). Culture and the evolutionary process. Chicago: University of Chicago Press.
Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., \& Gopnik, A. (2015). Children 's causal inferences from conflicting testimony and observations. Developmental Psychology, 52(1), 9-18.
Burdett, E. R. R., Lucas, A. J., Buchsbaum, D., McGuigan, N.,

Wood, L. A., \& Whiten, A. (2016). Do children copy an expert or a majority? Examining selective learning in instrumental and normative contexts. PLOS ONE, 11(10).
Corriveau, K. H., Fusaro, M., \& Harris, P. L. (2009). Going with the flow: Preschoolers prefer nondissenters as informants. Psychological Science, 20(3), 372-377.
Corriveau, K. H., \& Harris, P. L. (2010). Preschoolers (sometimes) defer to the majority in making simple perceptual judgments. Developmental Psychology, 46(2), 437-445.
Gweon, H., Dodell-Feder, D., Bedny, M., \& Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. Child Development, 83(6), 1853-1868.
Haun, D. B. M., Rekers, Y., \& Tomasello, M. (2012). Majoritybiased transmission in chimpanzees and human children, but not orangutans. Current Biology, 22, 727-731.
Haun, D. B. M., Rekers, Y., \& Tomasello, M. (2014). Children conform to the behavior of peers; other great apes stick with what they know. Psychological Science, 25(12), 2160-2167.
Haun, D. B. M., \& Tomasello, M. (2011). Conformity to peer pressure in preschool children. Child Development, 82(6), 17591767.

Hu, J., Whalen, A., Buchsbaum, D., Griffiths, T., \& Xu, F. (2015). Can children balance the size of a majority with the quality of their information? Proceedings of the 37th Annual Conference of the Cognitive Science Society.
Koenig, M. A., \& Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. Child Development, 76(6), 1261-1277.
Kushnir, T., \& Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. Psychological Science, 16(9), 678-683.
Kushnir, T., Vredenburgh, C., \& Schneider, L. A. (2013). "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. Developmental Psychology, 49(3), 446-453.
Kushnir, T., Wellman, H. M., \& Gelman, S. A. (2009). A selfagency bias in preschoolers' causal inferences. Developmental Psychology, 45(2), 597-603.
Lane, J. D., Harris, P. L., Gelman, S. A., \& Wellman, H. M. (2014). More Than Meets the Eye: Young Children's Trust in Claims That Defy Their Perceptions. Developmental Psychology, 50(3), 865-871.
Mills, C. M. (2013). Knowing When to Doubt: Developing a Critical Stance When Learning From Others. Developmental Psychology, 49(3), 404.
Perreault, C., Moya, C., \& Boyd, R. (2012). A Bayesian approach to the evolution of social learning. Evolution and Human Behavior, 33(5), 449-459.
Rendell, L., Fogarty, L., \& Laland, K. N. (2010). Rogers' paradox recast and resolved: Population structure and the evolution of social learning strategies. Evolution, 64(2), 534-548.
Schillaci, R. S., \& Kelemen, D. (2014). Children's conformity when acquiring novel conventions: The case of artifacts. Journal of Cognition and Development, 15(4), 569-583.
Wellman, H. M., Cross, D., \& Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. Child Development, 72(3), 655-684.
Whalen, A., Griffiths, T. L., \& Buchsbaum, D. (in press). Sensitivity to shared information in social learning. Cognitive Science.
Wilks, M., Collier-Baker, E., \& Nielsen, M. (2014). Preschool children favor copying a successful individual over an unsuccessful group. Developmental Science, 18(6), 1014-1024.

# Do Speaker's Emotions influence their Language Production? Studying the Influence of Disgust and Amusement on Alignment in Interactive Reference 

Charlotte Out (c.out@uvt.nl)<br>Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE, The Netherlands<br>Martijn Goudbeek (m.b.goudbeek@uvt.nl)<br>Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE, The Netherlands<br>Emiel Krahmer (e.j.krahmer@uvt.nl)<br>Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE, The Netherlands


#### Abstract

The influence of emotion on (the early stages of) speech production processes, notably content selection has received little scholarly attention. Goudbeek \& Krahmer (2012) found evidence for alignment at the conceptual level: speakers may start using a dispreferred attribute over a preferred attribute in their referring expressions when they are primed by a prerecorded female voice in a preceding interaction. The current study aimed to assess the role of emotion (using amusement and disgust) in alignment, while simultaneously replicating this finding in a more naturalistic setting involving two human participants in naturalistic dialogue. Our results replicate the findings by Goudbeek \& Krahmer (2012), generalizing their findings to a much more naturalistic setting. In addition, we found that amused, but not disgusted speakers tend to use the preferred attribute more to describe objects to their conversational partner.


Keywords: alignment; egocentricity bias, attentional bias, emotion; amusement; disgust; speech production; referential expressions, psycholinguistics.

## Introduction

Several effects of emotion on various processes in speech have been studied extensively, including effects on articulation and pronoun use. For instance, speakers often signal their emotional state in their prosody, by sobbing, crying or shouting (Bachorowski, 1999; Goudbeek \& Scherer, 2010) and depressed writers have been shown to use more first person singular pronouns (Pajak \& Trzebiński, 2014; Stirman \& Pennebaker, 2001). However, the impact of emotion on other aspects of the speech production processes has received little attention. In this study, we aim to investigate the role of emotion in the earlier processes in speech production, in particular on content selection stage ("deciding what to say") of language production, focusing on referential expressions.

## Emotion and speech production

As far as we know, only a few studies have looked at the relationship between emotional state and content selection. For example, Kempe, Rookes and Swarbrigg (2012) looked at the effect of speaker emotion (positive or negative emotion) on ambiguity avoidance in the production of referring expressions. In their experiment, emotion was
induced by a positive or negative video, accompanied by emotion congruent classical music. After the emotion manipulation, participants were asked to uniquely describe four pictures on the sheet. In the critical trials, two of the four pictures could be described in a linguistically ambiguous way, e.g. as a "bat" which could either be a flying bat or a baseball bat. They found that speakers in a positive state were less likely to disambiguate the second linguistically ambiguous picture, that is, they were more likely to use the word 'bat' for both the flying bat and the baseball bat. These findings suggest that positive emotions might increase ambiguity in referring expressions, which could be the result of an attentional shift in the speaker (Beukeboom and Semin, 2006).
Attentional bias It has been generally accepted that positive emotions (e.g., amusement) broaden attention, whereas negative emotions (e.g., sadness) narrows attention (see Frederickson, 2001). However, Harmon-Jones, Gable, and Price (2013) state that not valence, but the motivational intensity of emotions influence attention: emotions of low motivational intensity (e.g., sadness) broaden cognitive scope and emotions of high motivational intensity (e.g., disgust) narrow cognitive scope. They found that individuals exposed to disgusting pictures (compared to neutral pictures) who did a global-local letter task (Navon, 1977) responded slower to global than to local targets (Gable and Harmon-Jones, 2010), supporting the hypothesis that emotions of high motivational intensity narrow attention and make people focus more on details.
Egocentricity bias Egocentricity of speakers has been known to influence content selection. The egocentricity bias is the tendency of individuals to use their own perspective as reference point to the world (Ross \& Sicoly, 1979). Many authors claim that although individuals are often able to adjust to the perspective of the listener, they initially act egoistically (Epley, Morewedge, and Keysar, 2004; Horton and Keysar, 1996), although some beg to differ (see for example Bezuidenhout, 2013). According to Converse, Lin, Keysar and Epley (2008) and Clore and Hutsinger (2007), individuals in a positive state are less likely to adopt to the perspective of another person than individuals in a negative state, the shift of perspective to the listener might be impaired because positive emotions promote automatic responses.

Although these studies suggest an interesting link between emotion and (the early stages of) speech production, many questions remain. For example, what is the effect of emotion on the language production of speakers and listeners in an interaction?

## Alignment in interaction

In the current study, we focus on the effect of emotion on the amount of alignment between conversational partners in referential expressions. As argued by Garrod and Pickering (2004), one of the ways conversational partners can align to each other is by using the same attributes to refer to an object as their conversational partner. For example, if the other person just referred to an object in terms of its size (the large table), the speaker would be more likely to use size as well in a subsequent reference, because the previous use of size would prime this attribute. However, this is at odds with another tendency that has been reported in the literature, namely that speakers prefer to use certain attributes that are more "absolute" in their meaning over attributes that are less so (e.g. color over size, Martin, 1969; Pechmann, 1989). Inspired by observations such as these, Dale and Reiter (1995) developed the Incremental Algorithm which assumes a fixed preference order of attributes to determine in what order certain attributes are used in the generation of referential expressions. This Incremental Algorithm states that when individuals describe an object, they will first use the most preferred attribute and matching value, e.g., color and then "red", leading to the red chair. When this is not sufficient to single out the target object (e.g., there are multiple red chairs), the speaker will proceed by adding a less preferred attribute, e.g., size, leading to the large red chair. The speaker will continue adding attributes until the listener is able to identify which chair she is talking about. Dale and Reiter's (1995) Incremental Algorithm thus predicts that speakers will never use a dispreferred attribute when a preferred attribute is sufficient for identification. However, Goudbeek and Krahmer (2012) primed speakers with dispreferred attributes (attributes that were used earlier in an interaction) - and investigated whether they would stick to their preferences or align by incorporating the dispreferred attribute that was used by their conversational partner. In their study, participants listened to a pre-recorded female voice referring to one of three furniture objects, using either a preferred (color; "the red chair") or dispreferred (orientation, "the chair seen from the side") attribute. They subsequently indicated which image (the target) matched this description. When they were asked to describe a new target object, they tended to use the same type of attribute that they were primed with before, even when they could also use the preferred attribute to distinguish the target. These results show that speakers may thus use dispreferred attributes over preferred ones when they are primed to do so. In this paper, we study whether the emotions amusement and disgust might inhibit or promote this tendency. In addition, we aim to replicate the findings by Goudbeek and

Krahmer (2012) using a more naturalistic elicitation paradigm.

## The present study

Based on the results from Goudbeek and Krahmer (2012) we predict that speakers will indeed align with their dialogue partners and start using the dispreferred attribute in their referring expressions when primed to do so. With respect to the effect of emotion, previous research (e.g., Kempe et al., 2013, Beukeboom \& Semin, 2006) indicates that the emotional state of a speaker influences the content selection process of language production, and thus potentially the degree to which speakers align with respect to the attributes they use in interaction.

However, mainly the influence of emotions differing in valence (positive vs. negative) on speech production has been studied (see Kempe et al., 2012; Converse et al., 2008) which severely limits our understanding into the role of emotion in speech production. After all, emotions can be differentiated in other ways, which might influence speech production as well.

We induced amusement and disgust, two emotions that differ on multiple appraisals, among which is valence, but also approach/avoidance, potency/control and, possibly, intensity (see Scherer, 2013). Amusement is a positive emotion that occurs when a person experiences something entertaining (e.g., a joke) and feels pleasant (Tong, 2015). Disgust is an emotion that is elicited when a person is confronted with something they deem repulsive, for instance bodily fluids (vomit, pus, urine). We will study the effect of disgust and amusement on alignment in an interactive referential task. Will amused speakers or disgusted speakers align more with their conversational partners, even when they use a dispreferred way to refer to a target?

## Methods

## Participants

A total of 140 Dutch-speaking university students (36 male), participated in the experiment in pairs ( $\mathrm{n}=70$ ).

## Materials

Stimuli Following Goudbeek and Krahmer (2012), we used pictures taken from the TUNA corpus (van Deemter, Gatt, van der Sluis, \& Power, 2012), depicting front-facing furniture items (a fan, a chair, a couch, and a desk) in four different colors (blue, green, red, and grey) and two different sizes (large or small). Participants were asked to uniquely identify the target picture (accompanied by two distractors) to their conversational partner. Previous studies (Gatt et al., 2007; Goudbeek \& Krahmer, 2012) indicate the well-known preference of participants for color in their description of the target picture.

There were three types of trails: color trials, size trials and filler trials. Each participant pair was presented with 60 trials, divided into two blocks, consisting of 20 color trials
and 20 size trials. Additionally, each block included 10 filler trials, all containing large pictures of furniture in greyscale. Four versions were created, containing different orders of trials.

Mood questions To control for participants' mood before the experiment, we asked the participants to rate their mood before they watched the video. They indicated on a 1 to 7 scale how much they experienced each of the following moods: happy/sad, pleasant/unpleasant, satisfied/unsatisfied, content/discontent, cheerful/sullen, in high spirits/lowspirited (Krahmer, van Dorst, \& Ummelen, 2014, based on Mackie and Worth, 1991 and Bohner et al. 1992; English translations of Dutch originals).

Manipulation check To check whether emotion induction was successful, we asked participants after viewing the video to report how much amusement and disgust (and pride, anger, sadness, disgust, surprise and fear) they experienced on a 1 ("not at all") to 7 ("extremely") point Likert scale.

Other-participant questions After the director-matcher task, participants rated on a 7-point Likert scale (ranging from 1: "not at all" till 7: "very") how much they liked the other participant, how empathic they felt towards them, and how much they thought they got along. Finally, they indicated if they knew the participant, choosing either "no", "yes, a little", or "yes, very well".

Videos To counter the possibility of film specific effects, two different disgust-inducing and two different amusement videos were shown. Based on existing literature, we used four videos that were moderate to highly successful in inducing the corresponding emotions amusement and disgust, respectively. The amusement videos were "When Harry met Sally" (1989) and "There's Something About Mary" (1998). The disgust videos were "Trainspotting" (1996) and "Pink Flamingos" (1972). We selected these videos because they have been used effectively in recent work (e.g., Hewig, Hagemann, Seifert, Gollwitzer, Naumann, and Bartussek, 2005; Fajula, Bonin-Guilaume, Jouve, and Blin, 2013; Harlé and Sanfey, 2010; Schaeffer, Nils, Sanchez, and Philippot, 2010; and Rottenberg, Ray, and Gross, 2007).

## Procedure

After the participants had read and signed the consent form, they were sent to separate cubicles and filled in the demographics and answer the mood questions. The participants were informed that they were going to view an (emotional) video and were instructed to pay attention to the video and keep their eyes on the screen, because they would need this information in the video later in the experiment. After viewing the video (the emotion induction), they answered the questions of the manipulation check with respect to their current emotional state. Subsequently, they
went into a new room together with the other participant. To enhance the emotion manipulation, participants discussed the video they viewed with each other for approximately 2 minutes. They were instructed to focus on describing what they saw in the video, and telling the other participant what they thought and felt while viewing the video. They then filled in the mood questionnaire again and go on to perform the director-matcher task together.

Each trial consisted of four turns. First, participant A, the director, described the target picture (framed by a red border on the screen) to participant B , the matcher. Depending on the trial, participant A used (was forced to use) either a preferred or dispreferred attribute to describe the target picture to participant $B$. In the color trials, participant $A$ could only use the preferred attribute color to distinguish the target picture from the distractors. For example, the target picture was a large blue fan, and the distractors were a large red couch and a large red fan. Therefore, participant A had to use color to describe the target picture. In the size trials, participant A could only use the dispreferred attribute size to distinguish the target picture from the distractors. For example, the target picture was a large green desk, and the distractors were a small green desk and a small green fan. Therefore, participant $A$ had to use size to describe the target picture (see Figure 1, square 1).
Second, participant B, the matcher, saw the same pictures on their screen in a different order than participant A. After listening to the description of participant A, they indicated the matching picture by pressing the key of the corresponding number on their keyboard, e.g., " 2 " (see Figure 1, square 2). When participant A knew that participant B had selected an answer, she pressed "Enter" and the participants viewed a new screen.


Figure 1. Example of a size trial in the director-matcher task. Square 1 and 2 depict green pictures. Square 3 and 4 depict a red couch (the target), a blue desk and a grey chair.

Third, the participants switched roles: now participant B was the director and participant A the matcher. In contrast to the previous turn, the combination of pictures on this screen gave participant B the choice to use either the preferred or dispreferred attribute to distinguish the target picture from the distractors. For example, the target picture was a large red couch and the distractors were a small grey chair and a small blue desk. Participant B could either use the preferred attribute ("the red couch") or use the dispreferred attribute ("the large couch") to distinguish the target picture from the distractors (see Figure 1, square 3). In case participant B aligned with participant A , they used color when participant A (i.e., in color trials) used the preferred attribute, and size when participant A used the dispreferred attribute (i.e., in size trials).

Fourth, participant $A$, now the matcher, selected the picture that matched participant B's description by pressing the key of the corresponding number on the keyboard, e.g., "2" (see Figure 1, square 4). When participant B knew that participant A had selected an answer, participant B pressed "Enter", marking the end of the trial. After participant B had pressed "Enter", a new trial appeared and the procedure was repeated. Following the director-matcher task, both participants filled in the questions about the other participant. They got debriefed and received compensation (credits or money).


Figure 2a. Proportion of preferred and dispreferred attributes per Prime (Color or Size) for Amusement


Figure 2b. Proportion of use of preferred and dispreferred attributes per Prime (color or size) for Disgust

## Results

## Manipulation check

First, we tested whether the emotion manipulation was effective. We performed a one-way analysis of variance with Emotion Video (Amusement vs. Disgust) as independent variable and Emotion Scale (amusement vs. disgust) as dependent variable. As expected, we found a significant effect of Emotion on amusement, $F(1,138)=$ $88.89, p<.0001$, and disgust, $F(1,138)=255.47, p<.0001$. The mean scores of the combined videos per emotion indicate that participants who viewed an amusing video reported higher levels of amusement ( $M=4.89, S D=1.38$ ) than disgust ( $M=2.60, S D=1.49$ ). Participants who viewed a disgusting video reported a higher level of disgust ( $M=$ $6.26, S D=1.38)$ than amusement $(M=2.51, S D=1.39)$. This indicates that the emotion manipulation had the desired effect.

## Analyses

To statistically evaluate the effects of emotion, prime, and attribute, we conducted an analysis of variance with the proportion of attribute use as dependent variable and Emotion (Amusement vs. Disgust) as between subject factor, and Prime (Color vs. Size) and Attribute (preferred vs. dispreferred) as within-subject factors. The results of this analysis can be found in Table 1.

A significant main effect was found for Prime, $F(1,68)=$ 47.36, $p<.0001, \eta^{2}=.41$, indicating that the prime indeed influences the selection of attribute. Mean scores (with standard deviations) of the proportion of attribute use as influenced by prime can be found in Table 2.

A significant main effect was found for Attribute, $F(1$, $68)=33.67, p<.001, \eta^{2}=.33$, confirming that the preferred property color $(M=.80, S E=.03)$ is indeed preferred over size $(M=.52, S E=.03)$.
The effect of emotion The interaction between Emotion and Attribute is significant, $F(1,68)=5.01, p=.028, \eta^{2}=$ .07. A one-way analysis of variance with Emotion and Attribute shows a significant effect of Emotion for the use of preferred attribute, $F(1,68)=5.54, p=.022$, regardless of prime. Amused individuals showed a preference for the preferred attribute color ( $M=.86, S D=.19$ ) over the dispreferred attribute size ( $M=.48, S D=.22$ ). Disgusted participants did not show a preference for color ( $M=.73$, $S D=.28)$ or size $(M=.56, S D=.25), F(1,68)=1.91$, $p=.172$. The three-way interaction between Emotion, Prime and Attribute was not significant $(F(1,68)=3.34, p=$ .072), but there was a significant interaction between Emotion and Attribute $(F(1,68)=5.01, p=.022)$. The proportions of preferred and dispreferred attributes as a function of Prime and Emotion are shown in Figure 2a (for Amusement) and Figure 2b (for Disgust).

Table 1. Summary of statistical analysis

|  | $F$ | $p \leq$ | $\eta^{2}$ |
| :--- | :--- | :---: | :---: |
| Emotion | 0.05 | .376 | .01 |
| Prime | 47.36 | .001 | .41 |
| Attribute | 33.67 | .001 | .33 |
| Emotion x Prime | 3.34 | .072 | .05 |
| Emotion x Attribute | 5.01 | .028 | .07 |
| Prime x Attribute | 119.89 | .001 | .64 |
| Emotion x Prime x Attribute | 0.31 | .580 | .01 |

Table 2. Proportions of preferred and dispreferred attributes per Prime (color vs. size)

| Prime | Attribute | $M$ | $S D$ |
| :--- | :--- | :--- | :--- |
| Color | Preferred | 0.85 | 0.21 |
|  | Dispreferred | 0.31 | 0.28 |
| Size | Preferred | 0.75 | 0.31 |
|  | Dispreferred | 0.72 | 0.29 |

## Discussion

The aim of this study was twofold. One, investigating the effect of emotion on alignment in interactive reference production. Two, providing a more naturalistic replication of the results by Goudbeek and Krahmer (2012) by investigating alignment on dispreferred properties in a truly naturalistic version of the interactive alignment paradigm.

Regarding the effect of emotion on attribute use, we found that amused speakers have a stronger preference for the preferred attribute (color) over the dispreferred attribute (size) than disgusted speakers. This finding can be explained by the theory that individuals in a positive state tend to process information more shallow and global than individuals in a negative state (e.g., Beukeboom \& Semin, 2006) therefore using the preferred attribute because it is the default.

Regarding the effect of emotion on alignment, we did not find a statistically significant interaction between emotion, prime and attribute. However, upon inspecting our data, we did observe some interesting trends. We found that amused speakers aligned with their conversational partner regardless of prime. In other words, amused speakers aligned when they were primed with color and when they were primed with size (Figure 2a). This is in line with Harmon-Jones et al. (2013): amusement, an emotion of low emotional intensity, broadens the cognitive scope and therefore, speakers align with their conversational partners, regardless of prime. However, our amused speakers still used the preferred attribute color more (Figure 2a), supporting Clore and Hutsinger (2007), who state that speakers in a positive state find it harder to shift to the perspective of their conversational partner.

Disgusted speakers aligned when they were primed with color and when they were primed with size as well (Figure 2b). However, disgusted speakers primed with size
(opposed to the amused speakers primed with size) use the dispreferred attribute more than the preferred attribute, indicating that disgusted speakers have an even stronger tendency to align than amused individuals. This might be explained by the theory that individuals in a negative state have a narrower scope of attention (Beukeboom \& Semin, 2006). A narrow scope of attention might cause disgusted speakers to focus more on the words of their conversational partner than amused speakers who have a broad focus. The increased attention for the conversational partner in turn results in more alignment, regardless of prime (color or size). However, the result that disgusted speakers align more when they are primed with size than their amused peers can also be explained by the egocentricity bias (see Kempe et al., 2012). If amused speakers are more egocentric than disgusted speakers, they will rely more on their own perspective, using the preferred attribute color more, regardless of prime. This might not be the case for disgusted speakers, who are less self-focused and therefore align with their conversational partner, even when the prime was a dispreferred attribute.

The results of this study are perfectly in line with those of Goudbeek and Krahmer (2012). First, we found that participants generally used the preferred attribute color over the dispreferred attribute size. Second, participants primed with size used the dispreferred attribute size more than when they were primed with color. This is an interesting result, because the paradigm used in the study by Goudbeek \& Krahmer (2012) was much more artificial than the one in our current study. In their experiment, speakers interacted with a computer and were primed by a pre-recorded computerized female voice. In our study, two human participants interacted in pairs in a relatively natural setting: they were asked to interact normally, without restrictions. The participants who primed did this naturally and unconsciously, by being only able to use the preferred or dispreferred attribute to describe the target picture.

The preliminary evidence for the differential effects of amusement and disgust on attribute choice in referential expressions should lead to further explorations of the effect of (various) emotions on language production in human interactions. Future studies could, for example, focus on finding the (inter- and intrapersonal) mechanisms that might underlie the preference of emotional speakers to use either preferred over dispreferred attributes. In our study, amused speakers preferred color much more than disgusted speakers did, which implies that the emotional state of a speaker influences her attribute preferences. These and similar studies should result in a more detailed picture of the underlying mechanisms of the language production of emotional speakers.

## Acknowledgments

The research reported in this article is part of the project "Producing Affective Language: Content Selection, Message Formulation and Computational Modelling" funded by the Netherlands Organization for Scientific

Research (NWO via Grant PR-14-87). We thank Nadine Braun for her support and insights.

## References

Bachorowski, J.A. (1999). Vocal expression and perception of emotion. Current Directions in Psychological Science, 8, 53-57.
Beukeboom, C. J., \& Semin, G. R. (2006). How mood turns on language. Journal of Experimental Social Psychology, 42, 553-566.
Bezuidenhout, A. (2013). Perspective taking in conversation: A defense of speaker non-egocentricity. Journal of Pragmatics, 48, 4-16.
Clore, G. L., \& Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. Trends in cognitive sciences, 11(9), 393-399.
Converse, B. A., Lin, S., Keysar, B., \& Epley, N. (2008). In the mood to get over yourself: Mood affects Theory-ofMind use. Emotion, 8, 725-730.
Dale, R., \& Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. Cognitive Science, 19, 233-263.
van Deemter, K., Gatt, A., van der Sluis, I., \& Power, R. (2012). Generation of referring expressions: Assessing the Incremental Algorithm. Cognitive Science, 36, 799-836.
Epley, N., Morewedge, C. K., \& Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. Journal of Experimental Social Psychology, 40, 760-768.
Fajula, C., Bonin-Guillaume, S., Jouve, E., \& Blin, O. (2013). Emotional reactivity assessment of healthy elderly with an emotion-induced procedure, Experimental Aging Research, 39, 109-124.
Fredrickson, B. L. (2001). The Role of Positive Emotions in Positive Psychology: The Broaden-and-Build Theory of Positive Emotions. The American Psychologist, 56, 218226.

Gable, P. A., \& Harmon-Jones, E. (2010). The blues broaden, but the nasty narrows: Attentional consequences of negative affects low and high in motivational intensity. Psychological Science, 21, 211-215. doi:10.1177/0956797609359622
Garrod, S., \& Pickering, M. J. (2004). Why is conversation so easy? Trends in Cognitive Sciences, 8, 8-11.
Gatt, A., van der Sluis, I., \& van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In Proceedings of the $11^{\text {th }}$ European Workshop on Natural Language Generation, July 17-20, Sloss Dagstuhl, Germany.
Goudbeek, M., \& Scherer, K., (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. Journal of the Acoustical Society of America, 128, 1322-1336.
Goudbeek, M., \& E. Krahmer (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential
overspecification. Topics in Cognitive Science, 4, 269289.

Harlé, K.M., \& Sanfey, A.G. (2010). Effects of approach and withdrawal motivation on interactive economic decisions. Cognition and Emotion, 24, 1456-1465.
Harmon-Jones, E., Gable, P.A., Price, T.F. (2013). Does negative affect always narrow and positive affect always broaden the mind? Considering the influence of motivational intensity on cognitive scope. Current Directions in Psychological Science, 22, 301-307.
Hewig, J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., Bartussek, D. (2005). A revised film set for the induction of basic emotions. Cognition \& Emotion, 19, 1095-1109.
Horton, W. S., \& Keysar, B. (1996). When do speakers take into account common ground?. Cognition, 59, 91-117.
Kempe, V., Rookes, M., \& Swarbrigg, L. (2012). Speaker emotion can affect ambiguity production. Language and Cognitive Processes, 1-12.
Krahmer, E. J., van Dorst, J., \& Ummelen, N. (2004). Mood, persuasion and information presentation: the influence of mood on the effectiveness of persuasive digital documents. Information Design Journal, 12, 4052.

Martin, J.E. (1969). Semantic determinants of preferred adjective order. Journal of Verbal Learning and Verbal Behavior, 8, 697-704.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. Cognitive Psychology, 9, 353-383.
Pajak, K., \& Trzebinski, J. (2014). Escaping the world: linguistic indicators of suicide attemps in poets. Journal of Loss \& Trauma, 19, 389-402.
Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics, 27, 89-110.
Scherer, Klaus R. (2013). The Nature and Dynamics of Relevance and Valence Appraisals: Theoretical Advances and Recent Evidence. Emotion Review, 5, 150-162.
Schaeffer, A., Nils, F., Sanchez, X., \& Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cognition \& Emotion, 24, 1153-1172.
Stirman, S. W., \& Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. Psychosomatic Medicine, 63, 517-522.
Tong, E.M.W. (2015). Differentiation of 13 positive emotions by appraisals. Cognition and Emotion, 29, 484503.

Rottenberg, J., \& Ray, R.D., \& Gross, J.J. (2007). Emotion elicitation using films. In J.A. Coan \& J. J. B. Allen (Eds.), The handbook of emotion elicitation and assessment. London: Oxford University Press.
Ross, Michael; Sicoly, Fiore (1979). Egocentric biases in availability and attribution. Journal of Personality and Social Psychology. 37, 322-336.

# Nudging Problematic Smartphone Use to a Lower Level 

Lauri Palokangas (Lauri.Palokangas@gmail.com)<br>Jyrki Suomala (Jyrki.Suomala@Laurea.fi)<br>Laurea University of Applied Sciences, NeuroLab, Vanha maantie 9<br>02650 Espoo, FINLAND


#### Abstract

Smartphone usage has evolved in people's lives from necessity to habit and in some cases leading to compulsive use and addiction. However, only a little research has been performed on the prevention of Problematic Smartphone Usage (PSU). Behavioral economics has been applied to investigate how smartphone users respond to nudges that try to lower their smartphone usage. Findings revealed that the Total Screen On Time (SOT) decreased when nudging smartphone users with information on their usage behaviors. Intermittent glancing, as well as the median session time increased, and the reduction in SOT was no longer statistically significant in the observation period after the nudges were no longer applied, suggesting relapse in smartphone usage behavior.


Keywords: addiction; behavioral insights; nudge; smartphone

## Introduction

Behavioral economics researchers (Kahneman, 2003) have identified a large number of systematic biases in people's decision-making and judgements. These biases have been regarded as evidence that people do not follow principles of the rationality suggested in neoclassical theory (Samuelson, 1937). Instead, people use a series of heuristics that often lead to systematic errors (Tversky and Kahneman, 1973). Thus, the results of the mainstream views in behavioral economics have a generally low opinion about human rationality.

A new positive approach - nudge - for peoples' decisionmaking have emerged (Thaler and Sunstein, 2008). According to this approach, people could be helped by a nudge to make optimal decisions (Thaler and Sunstein, 2008). By planning the environment based on so-called "choice architects" in order to make people change behavior to make decision makers better off as judged by themselves (Thaler \& Sunstein 2008). One example of the benefit of the nudge and choice architecture is to prompt vaccination receivers to write down the date and time of the appointment to increase vaccination rate (Milkman et al., 2011).

According to the dualistic model people engage two systems of thinking. System 1 is an automatic, effortless and often influenced by habits that cannot be influenced easily, whereas System 2 is effortful, deliberately controlled and associated with conscious thinking operations (Kahneman, 2003). The limited capacity of mental effort results to people preferring the System 1 thinking by applying heuristics. As an outcome, many decisions are based on beliefs of probabilities of possible outcomes (Tversky \& Kahneman, 1973). Nudges build on the proposition of dualistic system. By preferring the effortless processing, "choice architects" can for instance design routinization of medication, thus
creating a habit that is easier to maintain than a medication that is not based on a routine (Ryan \& Wagner, 2003).

## Problematic Smartphone Use

The heuristics suggested from the dualistic system can be theorized to be present in a person's smartphone usage habits. The high daily usage of a smartphone in people's lives have become significant (Montag et al., 2015b; Kim, 2013; Oulasvirta et al., 2012; Lin et al., 2015). Even if $82 \%$ of the respondents say that using their phone during the conversation hurts the setting, $89 \%$ of the people have used a phone themselves during their most recent social gathering (Rainie \& Zickuhr, 2015). People have been shown to frequently return to their uncompelled behavior even if they were willing to change their behavior for better (O'Connell, 1996). Smartphone usage can be seen to have evolved into a habit which can lead to compulsive use and addiction (Lee, Chang, Lin \& Cheng 2014, 373).
Frequent phone use has been connected to the indicators of certain types of addiction. Some studies (Lin et al., 2015; Hong et al., 2012; Lin et al., 2014; Leung, 2008) indicate that the compulsive use of smartphones share the characteristics of drug and alcohol addiction, and internet dependency. Moreover, pathological gambling analyses has been used to classify this type of smartphone addiction (Leung \& Liang, 2016).

Using a mobile device frequently and at excess durations has been shown to lead to various types of symptoms. Using phones in excessive quantities in personal business situations has been shown to lower quality outcomes in negotiations and to give a less trustworthy and less professional impression (Krishnan et al., 2014). In addition, the increased use of smartphones has been shown to lead to reduced concentration levels during school classes and unsafe driving habits (Hong, Chiu \& Huang, 2012). Furthermore, by taking a wireless device even for a short time can increase anxiety (Cheever, Rosen, Carrier \& Chavez, 2014).

Whereas most of studies have focused frequent phone use from addiction point of view, it is hard to find studies that have focused on the prevention of Problematic Smartphone Use (PSU) on healthy test subjects. In order to help lower the smartphone use without coercion or policies, it is important to investigate how PSU can be influenced by using behavior change interventions.

Behavioral economics can be applied to investigate how smartphone users respond to nudges (Thaler and Sunstein, 2008) that try to lower their smartphone usage. A concept of 'nudge' has been introduced in contrast to policies enforcing a desired behavior or to introducing significant economic
incentives. Nudges can be used to design an environment that "alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (Thaler et al., 2008). Although the nudge has been applied in many studies and projects (Johnson \& Goldstein, 2003; Shu et al., 2012), it is hard to find research reports that have focused on applying behavior change interventions to influence smartphone usage.

The nudges used in this research to influence smartphone use were designed based on Michie, van Stralen and West's (2011) Capability, Opportunity-Motivation-Behaviour (COM-B) framework. According to this framework, behavior change involves changing one or more of the capability, opportunity and motivations that relate to the behavior (Michie, Atkins \& West, 2014). Capability refers to knowledge and skills that influence engaging in the activity, opportunity refers to everything outside the test subject that prompts for behavior or makes it possible, and motivation refers to processes that energize and direct behavior (Michie et al., 2014). The first nudge used in this study was designed to influence to the capability component, whereas Motivational and Goal-Attainment nudges were designed to influence to the motivational component in the COM-B framework.

Goal setting combined with a commitment, and feedback concerning the behavior has been shown to lead to behavior change. Where providing information has improved knowledge about the issue, the behavior change has resulted from tailored information, goal setting and feedback. Whether the goal has been set by an external party or the subject themselves, it has not been shown to have influence. (Abrahamse, Steg, Vlek \& Rothengatter, 2007).

The Capability-nudge provided information regarding the phone use. The nudge was designed to be compatible with Hansen \& Jespersen's (2013) definition of Transparent type 1 nudge. In this category, the reflective thinking of a subject is a by-product of the nudge.

Both Motivational-nudge and the Goal-attainment-nudges added influence to the motivational component by providing an optional valentic emoticon based on the progress of the smartphone use. The appearance of the valentic emoticon in a Goal-attainment nudge was shown if the test subject attained a self-defined personal goal in reducing smartphone usage. The valentic emoticon was designed to influence behavior through reflective thinking and to indicate an attainment of a desired behavior. These nudges built on Hansen \& Jespersen's (2013) definition of Transparent type 2 nudge: the emoticon provided feedback to reinforce the commitment mechanism while the test subject maintained a complete freedom of choice, both before engaging with the phone, or after opening the phone and thus becoming subject to the nudge.

Even though excessive smartphone usage can lead to compulsive use and addiction (Lee, Chang, Lin \& Cheng, 2014), little is known how people can voluntarily lower smartphone use. Here we use the COM-B behavioral change
framework to study how smartphone usage can be influenced by nudges.

## Method

## Participants

Total of 201 users were recruited from social media (Twitter, Facebook and LinkedIn) to participate in the research using the following recruitment message: "Are you hooked to your phone? Do you use it way too much? Find it out. Participate in a research. Install Deglancer." The participants were not assessed or selected based on their attitudes towards smartphone usage. The participants were incognito to the researchers throughout the study. The study was initiated by a test subject when installing the application onto their Android smartphone from Google Play store. The users were presented information about the research both before installing the research application, and when the research information sheet was made available to them in the application.

Following the research practices of an earlier research project (Montag et al., 2015b) the data was filtered to include only the participants that completed the full five weeks of research without stopping their phone usage for more than three consecutive days during the research. After discarding corrupted research data and ineligible users, 78 users were included to the data analysis.

Earlier research results (Mueller, van der Heijden, Klein \& Potters, 2011; Altmann \& Traxler, 2014) had shown that the effect of nudges do not significantly correlate with economic or socio-demographic variables. Therefore, sociodemographic background variables are not reported. The ethics committee of the Federation of Universities of Applied Sciences approved the study.

## Procedure

A smartphone application was developed to conduct an intervention study using three different type of nudges. The study was constructed for consecutive five stages, each lasting for 7 days. As soon as the users started the application for the first time, the study initiated. The application registered itself to the service hosted in Google cloud computing infrastructure. In the beginning of the research, the test subjects responded to the Smartphone Addiction Inventory (SPAI) questionnaire in the research application, measuring their attitudes and effects towards smartphone usage (Lin et al., 2014), however, the SPAI data has not been analyzed for this paper. The participants used their personal smartphone for the duration of five weeks during which the interventions were performed and the research data was collected.

The first Baseline stage created a personal baseline of smartphone usage of a participant. During the second, Capability stage, at every unlock of the smartphone, the user was presented a nudge including the following information: the number of minutes that the phone was locked before the unlock event, the number of unlock events so far during the
ongoing day and the total duration that the screen has been turned on during the ongoing day. The purpose of this stage was to test the effect of information to the smartphone use. During the third, Motivational stage, a user was presented with a similar nudge to the second stage. Information in the nudge was preceded with a positively valenced injunctive emoticon if the smartphone user had lowered the smartphone usage and therefore the indicator value had improved: if smartphone was locked for longer than the average sleep time one week earlier, if the number of unlocks up to the current hour of the day was less than the number of unlocks up to the current hour of the day one week earlier, or if the total screen time up to the current hour of the day was less than up to the same hour of the day one week earlier. The purpose of this stage was to test the effect of positively valenced injunctive emoticon judged by an external authority.

In the beginning of the Goal-attainment stage, the user was prompted to select a goal for how much he or she wished to decrease the phone usage this week. If the user did not select a goal, the application used the default goal of $5 \%$ improvement to the previous week. Every time the user unlocked the smartphone, the application calculated if one or more of the indicators had improved more than the target percentage compared to the previous week's information. If the sleep time was at least $5 \%$ longer than the average sleep time in the previous week, if the number of unlocks was at least $5 \%$ less than the number of unlocks up to the same hour of the day in the previous week, or if the total duration of the screen time was at least $5 \%$ less than up to the current hour of the day in the previous week, the indicator was preceded with the same injunctive emoticon that was used in the Motivational stage.

As in the Motivational stage, the nudge in the Goalattainment stage built on the motivation component of the COM-B. However, as the stage included a task to define the percentage of the desired reduction in smartphone use, the goal-setting intended to direct attention and effort to reach the goal defined by the test subject. In order for the test subjects to easily maintain their state of goal attainment, the test subjects would have to reflect their phone usage before engaging with the phone, thus reducing phone usage. The purpose of this stage was to test the effect of goal-attainment, and the effect of injunctive emoticon based on a personally set goal. By prompting the test subject with active decisionmaking regarding the amount to reduce their smartphone use, it was expected that the test subjects would make effort to attain the goal that they had specified themselves.

As the study progressed to the fifth week, all notifications stopped, and application only recorded the user behavior for one week. In the same way with the first stage of the study, the fifth stage did not involve a nudge. The purpose of this stage was to investigate if users relapse to their prior behavior after the nudges are no longer present.

After full five weeks, the application notified the user that the study had been completed. However, the user could continue to use the application, and choose the type of nudge to present at every unlock. The test subjects had a choice to
continue to use the application, or uninstall the application from their smartphone.

## Data Analysis

A total of 606062 events were collected over the 5 -week study period were tested. These events were converted to 2304 observations, each of them representing one day of one test subject, equivalent to the definition of per day per user (pdpu) used in an earlier similar research (Oulasvirta et al. 2012). Five key indicators of smartphone usage were calculated from the research data: Total Screen On Time per day (SOT), Median Screen On Time of each session (Session Time), Total number of phone usage sessions per day (Unlocks), Number of phone usage sessions equal or shorter than 30 seconds in duration, over 10 minutes apart from the previous session (Glances) and Median Screen Off Time between two sessions (Median SFT). Of the 78 participants who completed the research, 58 chose to set their own goal in the goal-setting phase, whereas 20 participants got the default as a goal.

Time series of each key indicator was processed with Hilbert-Huang Transform (HHT). In this so-called sifting process, the time series of each key indicators were broken into intrinsic mode functions (i.e. IMFs) and by sequentially de-composing these intrinsic modes from the original signal, the remaining data represented the trend of the data over the study period. This analysis was compatible with the method used by Lin et al. (2015). Inferential statistics were performed to measure the effect of nudges in Capability, Motivational and Goal-attainment stages compared to the Baseline and to the Observation stages. The differences in key indicators were tested between the stages of the study for each test subject. The inferential statistic tests were done by performing independent-samples t-test using different stages of the study as a grouping variable, and each key indicator as test variable. Finally, a regression equation was calculated to predict a key indicator from another key indicator. The processing of the data was performed with the MatLab software package and the inferential analysis was performed with an SPSS statistical software.

## Results

There was a significant effect of intervention for SOT between Baseline and Capability stages. Between these stages, mean SOT lowered from 3 hours and 40 minutes pdpu to 3 hours and 14 minutes pdpu. The effect of intervention for SOT was also significant between the Baseline stage and the Motivational stage, as well as and between the Baseline and Goal-attainment stages. In the Motivational stage, SOT lowered to approximately 3 hours and 10 minutes, and to 3 hours and 13 minutes in the Goal-attainment stage. The decrease in SOT was not statistically significant when comparing the first stage to the last week, Observation stage, of the study. Table 1 below illustrates differences and statistical significance of SOT by stage.

Table 1: SOT by stage ( $\mathrm{N}=78$ ).

| Stages | Difference | Significance and effect size |
| :--- | :--- | :--- |
| 1 vs. 2 | -26 minutes | $\mathrm{t}(922)=2,888, \mathrm{p}<.01, \mathrm{~d}=0,19$ |
| 1 vs. 3 | -30 minutes | $\mathrm{t}(925)=3,356, \mathrm{p}<.01, \mathrm{~d}=0,22$ |
| 1 vs. 4 | -27 minutes | $\mathrm{t}(916)=2,871, \mathrm{p}<.01, \mathrm{~d}=0,19$ |

The difference in Session Time was significant between the Baseline stage and Observation stage. The mean duration of individual session increased from 51 seconds pdpu to 92 seconds pdpu. The difference was also significant between Capability and Goal-attainment stages, as well as between Capability and Observation stages. This difference was also significant between Motivational and Observation stages. The mean duration increased from 44 seconds in the Capability stage to 49 seconds in Motivational, to 63 seconds in Goal-attainment and finally to 92 seconds in Observation stage. The changes were not significant between adjacent stages. Table 2 below illustrates the changes in Session Time by stage.

Table 2: Session Time by stage ( $\mathrm{N}=78$ ).

| Stages | Difference | Significance and effect size |
| :--- | :--- | :--- |
| 1 vs. 5 | 41 seconds | $\mathrm{t}(913)=-2,466, \mathrm{p}<.05, \mathrm{~d}=-0,16$ |
| 2 vs. 4 | 5 seconds | $\mathrm{t}(920)=-2,298, \mathrm{p}<.05, \mathrm{~d}=-0,15$ |
| 2 vs. 5 | 19 seconds | $\mathrm{t}(917)=-2,945, \mathrm{p}<.01, \mathrm{~d}=-0,19$ |
| 3 vs. 5 | 33 seconds | $\mathrm{t}(920)=-2,674, \mathrm{p}<.01, \mathrm{~d}=-0,18$ |

The difference in Glances was only significant between Capability and Observation stages, $\mathrm{t}(917)=-2,006, \mathrm{p}<.05$. Mean Glances pdpu increased from 41,24 times in Capability to 45,55 times in Observation stage.

Simple linear regression was calculated to predict SOT based on Glances. Poor regression equation was found $(\mathrm{F}(1,2274)=142,124, \mathrm{p}<.000)$ with an R2 of .059 . Also, simple linear regression was calculated to predict Unlocks based on Glances. A significant regression equation was found $(\mathrm{F}(1,2274)=10188,592, \mathrm{p}<.000)$ with an R 2 of .818 .

Median SFT was significantly different between the last two stages when compared to the first three stages. However, due to the HHT being used in the pre-processing stage to address the non-linearity and non-stationarity of the research data, the comparison of the key indicator values using original units of measure might not be accurately depicted. Median SFT values after HHT pre-processing smoothed the data to a negative range without equivalent real world phenomenon. The changes in Median SFT can be characterized so that the difference in Median SFT is not significantly different between stages 1,2 and 3 , but Median SFT is markedly higher in stages 1,2 and 3 compared to stages 4 and 5 . The below Table 3 illustrates the trend of change in Median SFT by stage.

Table 3: Median SFT by stage ( $\mathrm{N}=78$ ).

| Stages | Significance and effect size |
| :--- | :--- |
| 1 vs. 4 | $\mathrm{t}(916)=3,194, \mathrm{p}<.01, \mathrm{~d}=0,21$ |
| 1 vs. 5 | $\mathrm{t}(913)=2,815, \mathrm{p}<.01, \mathrm{~d}=0,19$ |
| 2 vs. 4 | $\mathrm{t}(920)=3,194, \mathrm{p}<.01, \mathrm{~d}=0,21$ |
| 2 vs. 5 | $\mathrm{t}(917)=2,801, \mathrm{p}<.01, \mathrm{~d}=0,18$ |
| 3 vs. 4 | $\mathrm{t}(923)=2,299, \mathrm{p}<.05, \mathrm{~d}=0,15$ |
| 3 vs. 5 | $\mathrm{t}(920)=2,491, \mathrm{p}<.05, \mathrm{~d}=0,16$ |

## Discussion

There was a significant main effect for SOT between the Baseline stage and all of the three stages with the nudges. Consistent with COM-B framework (Michie et al., 2011) this change in SOT could be associated to an individual's aptitude to change their phone usage. Especially in the Capability stage the information pertaining to the user's phone usage was planned to be consistent with the definition of the type 1 transparent nudge (Hansen \& Jespersen, 2013). It can therefore be suggested that the effect of the nudge was significant enough to trigger an automatic reflection of smartphone use.

It is not possible to conclude that one type of nudge has higher significance to smartphone behavior than the other. Statistically significant decrease of SOT between the Baseline stage and both the Capability stage and the Motivational stage suggests that by using nudges that increase capability or motivational components can result to behavior change. However, there was no significant change in any of the key indicators between the different types of interventions.

Locke and Latham (2002) have earlier suggested that "the effects of goal setting are very reliable". Michie, Atkins and West (2014) have also reported that the interventions with "explicit targets and actions plans to feedback" had a higher impact compared to interventions without targets. According to Locke and Latham (2002), failures to replicate the effects of goal settings can be due to many reasons, including for example the lack of feedback, lack of commitment or failure to match the goal to the performance measure. It is possible that the key indicators used in this research do not mediate smartphone usage behavior. Also, by only providing positively valenced feedback about the goal attainment but inhibiting negatively valenced feedback about the failure to attain a goal could explain why this research could not successfully replicate the effects of goal setting.

The level of SOT per day ( 162 minutes) is in line with what Oulasvirta et al. (2012) reports, but it is only 62,3 percent of what Lin et al. (2015) reports as a median daily use time. Lin et al. (2015) report that the recruitment strategy in their study was "based on the potential higher penetration rate of smartphone use". Montag et al. (2015a) have written that substantial part of the sample in Lin et al.'s (2014) study was characterized as being smartphone addicted. The difference in the level of usage compared to Lin et al.'s research results
may suggest that the sample in this research did not include substantial amount of problematic smartphone users or smartphone addicts.

There was a significant difference in Session Time between number of stages as indicated in Table 2. In these comparisons with Observation stage, the Session Time increased from the stage under comparison.

The difference in Glances was only significant between the Capability stage and the Observation stage. Oulasvirta et al. (2012) define intermittent smartphone use as SIRB, short duration isolated, reward-based usage sessions. This definition includes a notion about the type of application: "at least $50 \%$ of the usage session duration is spent interacting with applications that provide the reward values". The definition of Glances is not therefore fully compatible with the definition of SIRBS.

There was no evidence that nudges can reduce the number of Glances. Oulasvirta et al. (2012) have suggested earlier that "checking habits may lead to more use overall". Poor regression equation to predict SOT based on Glances did not support Oulasvirta et al.'s previous findings, however, significant regression equation to predict Unlocks based on Glances would suggest that even though changes in Unlocks were not statistically significant during the research, Unlocks can be expected to increase after the nudges are no longer in effect. Oulasvirta et al. (2012) have earlier concluded that the increased "checking habit" is associated with higher phone usage overall. Oulasvirta et al. (2012) suggest that short sessions act as a "gateway" for other content on the device, and that they can be seen as a proxy for habitual usage.

Median SFT was significantly lower from Motivational stage onwards compared to the baseline. It was not possible to find a report that would have included at least descriptive statistics about the Session Time or Median SFT. In this research, Session Time was 23 seconds, and Median SFT was 198,50 seconds. Due to the lack of prior reported research evidence, these numbers provide little basis for inferential or comparative analysis. Davis (2001) has suggested that procrastination has a role in both the development and maintenance of generalized PIU. However, based on the data from this research it is not evident if a more frequent engagement with the phone is due to the test subjects putting off their responsibilities - as Davis suggests - or due to other reasons.

It can be theorized that the changes in SOT and Median SFT is due to the test subjects reducing their screen time overall even if they engaged with their phone more frequently. In the Goal-attainment stage the nudge was built on Motivational component in COM-B framework, proposing that the explicit goal is associated with lower phone usage. Evidence referred by Klasnja (2009) have proposed that the automatic goal activation can be triggered with presentation of salient information. It is possible that the nudge in the Capability stage had already triggered automatic goal activation, and the differences in nudges between the stages were not significant enough to trigger additional ways of behavior change beyond what was already active from the

Capability stage onwards. This could support an unchanged amount of glances throughout the experiment, although it remains unclear what triggered an increase in the mean Glances in the Observation stage.

Neither Session Time nor the number of unlocks lowered significantly between the Baseline and Capability stage. One possibility is that there was a mere-measurement effect from the beginning of the study and the users made an effort to generally lower the amount of engagement with the phone throughout the study by spending less time with the phone at each unlock. Another possibility is that the users generally reflected their phone usage and did not unlock the phone as often as before. In this case, as soon as they would engage with their phones, they would approximately spend the same amount of time with their phone, but that would happen less often. The changes can, however, be so small that it is not statistically significant for Unlocks or Median SFT. If the latter assumption was true, it would suggest that automatic goal may have triggered users to reflect their phone usage before they engage with their phone. As SOT was significantly or highly significantly lower in all stages of the study compared to the Baseline stage, the observation could be a sign of learning the phone usage behavior resulted by the interventions.

According to these results, a systematic relapse in behavior was seen after the interventions were no longer in effect. Block (2008) has earlier reported that the individuals with internet addiction are resistant to treatment and tend to relapse at a high rate. The findings from this research are compliant with Block's suggestion, although it is not possible to associate the findings from this research to a relapse based on psychiatric reasons.

Even though this research supports both Klasnja et al.'s (2009) as well as Oulasvirta et al.'s (2012) conclusion that interventions can help trigger behavior change, this research does not support the proposition that interventions can help maintain a behavior change.

This research does not provide evidence that the type of nudge explains differences in the number of unlocks or glances per day. It is possible that this is due to the path dependence between the stages of the research and the lack of a control group. Even though there was statistically significant different in the mean Glances between the Capability and the Observation stage, this finding is not supported by current literature. More evidence would be required to prove relapse in smartphone use resulting from the absence of a nudge, by randomizing the order of stages in the research and by introducing a control group.

It can be concluded that nudges can help lower key indicators of smartphone use, however, this might increase intermittent glancing and result to overall increased usage of the phone when the interventions are no longer present.

## Acknowledgments

Authors want to acknowledge two incognito software developers that helped develop application to conduct the research. Authors also want to acknowledge Dr. Sue-Huei

Chen and MD, PhD Yu-Hsuan Lin from the Department of Psychology in National Taiwan University for their assistance regarding Hilbert-Huang Transform, and to permit the use of SPAI questionnaire in the research.

## References

Abrahamse, W., Steg, L. , Vlek, C., \& Rothengatter, T. 2007. The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents. Journal of Environmental Psychology 27(4), 265-276.
Block, J.J. 2008. Issues for DSM-V: Internet Addiction. The American Journal of Psychiatry 165, 306-207.
Cheever, N.A. , Rosen, L.D. , Carrier, L.M. , \& Chavez, A. 2014. Out of sight is not out of mind: The impact of restricting wireless mobile device use on anxiety levels among low, moderate and high users. Computers in Human Behavior 37, 290-297.
Davis, R.A. 2001. A cognitive-behavioral model of pathological Internet use. Computers in Human Behavior 17, 187-195.
Hansen, P.G. \& Jespersen, A. 2013. Nudge and the Manipulation of Choice. A Framework for the Responsible Use of Nudge Approach to Behaviour Change in Public Policy. European Journal of Risk Regulation 2013, 3-28.
Hong, F.-Y. , Chiu, S.-I. , \& Huang, D.-H. 2012. A model of the relationship between psychological characteristics, mobile phone addiction and use of mobile phones by Taiwanese university female students. Computers in Human Behavior 28, 2152-2159.
Johnson, E.J. \& Goldstein, D.G. 2003. Do Defaults Save Lives? Science 302, 1338.
Kahneman, D. 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics. American Economic Review 93, 1449-1475.
Kim, B. 2013. Mobile Consumer Behavior: Myths and Reality. Library Technology Reports 49, 9-14.
Klasnja, P. , Consolvo, S. , McDonald, D.W. , Landay, J.A. , \& Pratt, W. 2009. Using Mobile \& Personal Sensing Technologies to Support Health Behavior Change in Everyday Life: Lessons Learned. AMIA Annual Symposium Proceedings 2009, 338-342.
Krishnan, A. , Kurtzberg, T.R. , \& Naquin, C.E. 2014. The Curse of the Smartphone: Electronic Multitasking in Negotiations. Negotiation Journal 30.
Lee, Y.-K. , Chang, C.-T. , Lin, Y. , \& Cheng, Z.-H. 2014. The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress. Computers in Human Behavior 31, 373-383.
Leung, L. 2008. Linking Psychological Attributes to Addiction and Improper Use of the Mobile Phone Among Adolescents in Hong Kong. Journal of Children and Media 2, 93-113.
Leung, L. \& Liang, J. 2016. Encyclopedia of Mobile Phone Behavior. IGI Global.
Lin, Y.-H. , Chang, L.-R., Lee, Y.-H. , Tseng, H.-W. , Kuo, T.B.J., \& Chen, S.-H. 2014. Development and Validation
of the Smartphone Addiction Inventory (SPAI). PLoS ONE 9, e98312.
Lin, Y.-H. , Lin, Y.-C. , Lee, Y.-H. , Lin, P.-H. , Lin, S.-H. , Chang, L.-R., Tseng, H.-W., Yen, L.-Y. , Yang, C.C.H. , \& Kuo, T.B.J. 2015. Time distortion associated with smartphone addiction: Identifying smartphone addiction via a mobile application (App). Journal of Psychiatric Research 65, 139-145.
Locke, E.A. \& Latham, G.P. 2002. Building a practically useful theory of goal setting and task motivation. A 35-year odyssey. The American Psychologist 57, 705-717.
Michie, S. , Atkins, L. , \& West, L. 2014. The Behaviour Change Wheel Book - A Guide To Designing Interventions. Great Britain: Silverback Publishing.
Michie, S. , van Stralen, M.M. , \& West, R. 2011. The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implementation Science 6, 42.
Milkman, K. Beshears, J., Choi, J., Laibson, D., \& Madrian, B. 2011. Using implementation intentions prompts to enhance influenza vaccination rates. PNAS 108, 1041510420.

Montag, C. , Błaszkiewicz, K. , Lachmann, B. , Sariyska, R. , Andone, I. , Trendafilov, B. , \& Markowetz, A. 2015a. Recorded Behavior as a Valuable Resource for Diagnostics in Mobile Phone Addiction: Evidence from Psychoinformatics. Behavioral Sciences 5, 434-442.
Montag, C. , Błaszkiewicz, K. , Sariyska, R. , Lachmann, B. , Andone, I. , Trendafilov, B. , Eibes, M. , \& Markowetz, A. 2015b. Smartphone usage in the 21 st century: who is active on WhatsApp? BMC Research Notes 8, 331.
Mueller, W., van der Heijden, E., Klein, T. \& Potters, J. 2011. Nudges and Impatience: Evidence from a Large Scale Experiment, Vienna Economics Papers, University of Vienna, Department of Economics.
Oulasvirta, A., Rattenbury, T. , Ma, L. , \& Raita, E. 2012. Habits make smartphone use more pervasive. Personal and Ubiquitous Computing 16, 105-114.
Rainie, L. \& Zickuhr, K. 2015. Americans' Views on Mobile Etiquette. Pew Research Center.
Ryan, G. W., \& Wagner, G. J. (2003). Pill taking "routinization": a critical factor to understanding episodic medication adherence. AIDS Care 15, 795-806.
Shu, L.L. , Mazar, N. , Gino, F. , Ariely, D. , \& Bazerman, M.H. 2012. Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. Proceedings of the National Academy of Sciences 109, 15197-15200.
Tversky, A. \& Kahneman, D. 1973. Availability: A Heuristic for Judging Frequency and Probability. Cognitive Psychology 5, 207-232.
Thaler, R. \& Sunstein, C. 2008. Nudge - Improving Decisions About Health, Wealth, and Happiness. New Haven \& London: Yale University Press.

# Auditory and Visual Contributions to Multisensory Integration 

Jessica L. Parker (parker.1026@osu.edu)<br>Department of Psychology, The Ohio State University at Newark<br>1179 University Dr., Newark, OH 43056, USA<br>Christopher W. Robinson (robinson.777@osu.edu)<br>Department of Psychology, The Ohio State University at Newark<br>1179 University Dr., Newark, OH 43056, USA


#### Abstract

Multisensory integration, or the merging of information from multiple sensory modalities, is important for many everyday tasks. One methodology used for examining this process is the Sound Induced Flash Illusion (SIFI), which presents participants with a number of flashes and either the same number of beeps (congruent) or a different number of beeps (incongruent), and requires the participant to respond by entering how many flashes they saw. The study expands on this research by examining the relative contributions of auditory and visual information on multisensory integration. While congruent and incongruent auditory stimuli affected visual perception (Experiment 1), there was little evidence that visual input affected auditory processing (Experiment 2). These findings support auditory dominance and modality appropriate hypothesis in adult populations and have implications on tasks that require integration across auditory and visual modalities.


Keywords: Crossmodal Processing; Multisensory Integration; Modality Dominance; Sound Induced Flash Illusion

## Introduction

A majority of our daily experiences require people to process multisensory information. As a person walks down the street, for example, they may see a car driving by, hear the engine as it approaches, smell the exhaust and feel the breeze as the car passes. How information from the different sensory modalities is integrated and combined into a unitary percept is considered multisensory integration (Shams, 2000). Using the example above, these multisensory experiences are perceived as one percept (car) instead of having independent experiences. Given the evident impact of multisensory integration in our everyday experiences, it is important to understand the contributions of auditory and visual contribution to multisensory integration and factors that facilitate and inhibit multisensory integration.

Shams, Kamitani, and Shimojo (2000) developed a test of multisensory integration called the Sound Induced Flash Illusion (SIFI), where the number of beeps influences how many flashes people see. In their study, they presented one, two, or three flashes, and in the cross-modal condition, these flashes were paired with one, two or three beeps. Participants were then asked to report how many flashes they saw, regardless of how many beeps they heard. If the number of beeps exceeded the number of flashes, participants tended to overestimate the number of flashes (fission). If the number of
beeps was less than the number of flashes, participants tended to underestimate flashes (fusion). Fission and fusion responses are implications of multisensory integration. This shows that the auditory information is being integrated with the visual information. Another study, using a different procedure and stimuli, tested perception of beeps to see if they could create a flash-induced sound illusion (Andersen, 2004). Under normal intensity levels, the visual flashes had no effect on auditory perception; however, they did find some evidence that flashes influenced beep perception when intensity levels of beeps were weakened to near threshold levels. This finding, in conjunction with Shams et al., may suggest that auditory information has a stronger effect on visual processing than vice versa; however, there were also numerous differences across studies; thus, making it difficult to make strong conclusions.

Why do auditory beeps affect participants' perception of the number of visual flashes? One possible explanation underlying this illusion is auditory dominance (Robinson \& Sloutsky, 2010a). According to this account, auditory and visual stimuli compete for attentional resources; thus, increased attention to one modality might come with a cost delayed or attenuated processing in the other modality. Moreover, because auditory stimuli are dynamic and transient, it may be adaptive to first allocate attention to the auditory modality before the information disappears. Most of the supporting research for auditory dominance comes from the developmental literature, where multisensory presentation attenuates visual processing more than auditory processing (Lewkowicz, 1988a; 1988b; Robinson \& Sloutsky, 2004; 2010b; Sloutsky \& Napolitano, 2003; Sloutsky \& Robinson, 2008). According to this account, the beeps in SIFI may interfere with processing of the visual flashes; whereas, the visual input may have little effect on processing of the beeps. Another possible reason why beeps may affect flash perception is modality appropriateness hypothesis, which states that the modality that is more appropriate for the task is the one that dominates (Welch \& Warren, 1980). Welch and Warren (1980) describe that with information processing, vision is dominant in spatial situations and audition is dominant for temporal judgements. When these two modalities are simultaneously presented and the task has a temporal aspect, studies have shown that audition becomes the dominant modality and can influence vision (Wada, Kitagawa, \& Noguchi, 2003). Thus, both
auditory dominance and modality appropriateness predict that auditory input should have a greater effect on visual perception than vice versa, especially when the task is temporal in nature (but see Tsay, 2013, where visual information affected judgements about musical performance).

Predicting that auditory information will have a greater effect on visual processing conflicts with much of the past research with adults that showed visual dominance, where the simultaneous presentation of auditory and visual information seems to inhibit auditory processing (see Sinnett, Spence, \& Soto-Faraco, 2007, and Spence, Parise, \& Chen, 2012). For example, when adults were required to press one button when they detected a visual stimulus, a different button when they detected an auditory stimulus, and a third button (or both buttons) when both stimuli were presented at the same time, participants often made errors on cross-modal trials by only pressing the visual button (Colavita, 1974). Thus, visual dominance tends to occur when adults are required to make speeded, modality-specific responses to auditory, visual, and crossmodal stimuli (Colavita, 1974; see Sinnett, Spence, \& Soto-Faraco, 2007 for review). One possible explanation for visual dominance is that adults may have a visual response bias to compensate for the fact that visual input is less alerting than auditory (Posner et al., 1974). It is important to note that the current study testing the SIFI is different from some of the modality dominance studies because it requires quantity judgements (how many beeps or flashes), rather than requiring speeded, modality-specific responses to auditory or visual input.

The current study used a modified SIFI task to test both auditory and visual processing and expands previous research in three ways. First, the study expands SIFI research by examining the relative contributions of the auditory and visual modalities in multisensory integration, as opposed to only examining the effects of auditory input on multisensory integration or visual input on multisensory integration. Second, the current study expands SIFI research by using facilitation effects (greater accuracy on cross-modal congruent trials than unimodal trials) as a measure of multisensory integration. Do congruent auditory or visual stimuli increase the accuracy of beep/flash perception? Finally, the current study will contribute to the modality dominance literature by using quantity judgements and accuracy (how many beeps or flashes) rather than speeded, modality-specific responses to auditory and visual information. Experiment 1 examined the effect of beeps on perception of flashes (replicating most SIFI studies), and Experiment 2 tested the effect of flashes on perception of beeps. Based on auditory dominance and modality appropriate hypothesis (Robinson \& Sloutsky, 2010a; Welch \& Warren, 1980), it is expected that congruent and incongruent auditory information will have a stronger effect on visual perception, whereas, it is expected that visual input will have a weak or no effect on auditory perception.

## Experiment 1

## Method

Participants Participants for Experiment 1 included 24 young adults ( 18 to 35 years). Young adults were recruited from the Ohio State University, and received class credit for the Introductory Psychology course in return for their participation. Three participants with uncorrected hearing or vision, as self-reported, were excluded from the data analyses.

Apparatus The experiment was conducted on a 22 " Dell PXL 2230 MW monitor with $1920 \times 1080$ resolution and Dell Optiplex 7040 systems with Intel Core i5 processors. Bose QuietComfort 25 Noise Cancelling headphones were used for auditory stimulus presentation. Stimulus timing and presentation and reaction time/accuracy data was collected using Direct RT software.

Materials The visual stimulus was a white circle $2^{\circ}$ in visual angle in the center of the screen with a black background. Each flash has a 20 ms duration with a 50 ms Inter-Stimulus Interval (ISI) between consecutive flashes. The auditory stimulus was a sine wave presented at 3.5 kHz (no rise or decay ramps). Each beep lasted for 20 ms , and there was a 50 ms ISI in between consecutive beeps. Auditory stimuli were presented via headphones at approximately 50 dB . In the crossmodal condition, the first beep occurred 35 ms before the first flash, or vice versa. The beep first and visual first conditions were randomized among the participants. Figure 1 shows the timing of the stimuli. The stimuli and timing was modeled after the original SIFI study (Shams, Kamitani \& Shimojo, 2000). Based on previous research and on preliminary analyses, the asynchronous timing had no significant effect on the SIFI.


Figure 1. Timeline of a 2 flash $/ 2$ beep stimuli. The dark blocks represent presentation of the stimuli and the grey represent ISI. The numbers above represent the time in ms from the beginning of the stimulus.

Design The experiment consisted of three blocks: visual unimodal, auditory unimodal, and crossmodal. There were five trials for each stimulus in the unimodal visual condition ( 2 flashes, 3 flashes, 4 flashes), and there were five trials for each possible stimulus in the unimodal auditory condition (2 beeps, 3 beeps, 4 beeps). There were also five trials of each possible stimulus in the crossmodal conditions (1 flash/1 beep, 1 flash $/ 2$ beeps, 1 flash $/ 3$ beeps, etc.). See Table 1 for all stimulus frequencies. Of the crossmodal trials, 15 were congruent and 30 were incongruent. Congruent trials had the same number of flashes and beeps, and incongruent had different numbers of flashes and beeps), which provided conflicting information.

Procedure In the unimodal auditory condition, participants heard 2, 3, or 4 beeps, and were asked to report how many beeps they heard. In the unimodal visual condition, they saw 2,3 , or 4 flashes and were asked to report how many they saw. In the crossmodal condition, participants were presented with 2,3 , or 4 beeps and/or 2,3 , or 4 flashes, and they were asked to report only how many flashes they saw. Each condition had a set of instructions before the trials and a conclusion to let the participant know when the condition was over. The order of condition was randomized among the participants, and each trial started within a condition started approximately 1000 ms after responding to the previous trial.

| Unimodal |  |
| :---: | :---: |
| Auditory | Visual |
| 2 Beeps (5) | 2 Flashes (5) |
| 3 Beeps (5) | 3 Flashes (5) |
| 4 Beeps (5) | 4 Flashes (5) |
| Crossmodal |  |
| *2 Flashes/2 Beeps (5) |  |
| *3 Flashes/3 Beeps (5) |  |
| * Flashes/4 Beeps (5) |  |
| 2 Flashes/3 Beeps (5) |  |
| 2 Flashes/4 Beeps (5) |  |
| 3 Flashes/2 Beeps (5) |  |
| 3 Flashes/4 Beeps (5) |  |
| 4 Flashes/2 Beeps (5) |  |
| 4 Flashes/3 Beeps (5) |  |

Table 1: Experiment 1 trial types and frequencies. Note, "*" denotes congruent trials.

## Results and Discussion

On each trial, participants reported how many flashes they perceived. Below we first report overall accuracies and then we report more traditional analyses focusing on actual responses ( 2,3 , or 4 ) and making a distinction between fission trials (more beeps than flashes) and fusion trials (fewer beeps than flashes).
Accuracy Each trial was classified as correct or incorrect. See left side of Figure 2 for means and standard errors of visual responses and the right side of Figure 2 for unimodal auditory accuracy. Analyses in Experiment 1 focus exclusively on visual responses. Using a 3 (number: 2, 3, 4) x 3 (trial type: congruent, incongruent, unimodal baseline) repeated measures ANOVA, a significant effect of condition was found, $F(2,46)=68.31, p<.001, \eta_{p}{ }^{2}=.75$. Accuracies were lower on unimodal trials $(M=.53, S E=.03)$ than congruent trials $(M=.66, S E=.03), t(23)=-3.613, p=.001$, which is consistent with facilitation effects. Interference effects were also found with higher accuracy on unimodal trials than incongruent trials $(M=.21, S E=.09), t(23)=8.42$, $p<.001$. Also, accuracy was also higher on congruent trials than incongruent trials, $t(23)=9.91 p<.001$.

The ANOVA also revealed a significant effect of number, $F(2,46)=17.05, p<.001, \eta_{p}{ }^{2}=.43$, with accuracy decreasing as the number of flashes increased. There was significantly higher accuracy on the 2-flash trials ( $M=.56$, $S E=.03)$ than on the 4 -flash trials $(M=.31, S E=.04), t(23)$ $=4.47, p<.001$. The 3-flash trials $(M=.53, S E=.03)$ also had a higher accuracy than the 4-flash trials, $t(23)=5.02, p$ $<.001$. Finally, the analyses also revealed a trial type x number interaction, $F(4,92)=5.32, p=.001, \eta_{p}{ }^{2}=.188$. As can be seen in Figure 2, cross-modal facilitation effects (congruent > unimodal) was most pronounced when presented with four flashes, and interference effects (unimodal > incongruent) decreased, with the strongest interference on 2-flash trials.


Figure 2. Accuracies across number and trial type. Error Bars denote Standard Errors.

The remaining analyses focus on actual responses ( 2,3 , or 4 ), not accuracies. On fission trials, there were more beeps than flashes, and on fusion trials, there were fewer beeps. Moreover, for fission and fusion cross-modal trials, we could only test two of the three numbers. For example, as can be seen in Table 1, there were no fission trials for 4 flashes and no fusion trials for 2 flashes because there were no trials where we presented five or one auditory stimulus, respectively. Thus, we used two $2 \times 2$ repeated measures ANOVA's to test for fission and fusion effects.

Fission Actual responses were collected on each trial and only fission trials and unimodal trials were submitted to a 2 (trial type: unimodal, fission) x 2 (number: 2, 3) repeated measures ANOVA. See left side of Figure 3 for means and standard errors. A significant effect of trial type was found, $F$ $(1,23)=71.41, p<.001, \eta_{p}^{2}=.76$, suggesting that auditory input affected perception of flashes. In particular, participants reported more flashes on fission trials $(M=3.23, S E=.08)$ than unimodal trials $(M=2.54, S E=.06)$, which was expected since there were more beeps than flashes. There was also a significant effect of number, $F(1,23)=60.15, p<.001$, $\eta_{p}{ }^{2}=.72$. Not surprisingly, participants on 2-flash trials ( $M=$
2.68, $S E=.05$ ) reported fewer flashes than on 3-flash trials ( $M=3.09, S E=.07$ ).

Fusion Actual responses were collected on each trial and only fusion trials and unimodal trials were submitted to a 2 (trial type: unimodal, fusion) x 2 (number: 3, 4) repeated measures ANOVA. See right side of Figure 3 for means and standard errors. A significant effect of trial type was found, $F$ $(1,23)=15.87, p=.001, \eta_{p}^{2}=.41$, which suggests that the number of beeps affected perception of flashes. Participants reported fewer flashes on fusion trials $(M=2.56, S E=.03)$ than on unimodal trials $(M=2.95, S E=.09)$, which was expected since there were fewer beeps than flashes. There was also a significant effect of number of stimuli on response, $F(1,23)=62.16, p<.001, \eta_{p}{ }^{2}=.73$, with participants reporting fewer flashes on 3-flash trials $(M=2.57, S E=.045)$ than 4-flash trials $(M=2.94, S E=.06)$.


Figure 3. Actual responses across number and trial type. The left side of the figure denotes fission trials and the right side denotes fusion trials. Error Bars denote Standard Errors.

## Experiment 2

The purpose of Experiment 2 was to test if the relative contribution of auditory and visual information on multisensory integration was symmetrical or asymmetrical. In cross-modal trials of Experiment 2, participants were asked to report how many beeps they heard. It was hypothesized that the effects would be asymmetrical, with visual input in Experiment 2 having little to no effect on auditory processing.

## Method

Participants, Materials, and Procedure Experiment 2 was identical to Experiment 1, with the exception that we tested effects of flashes on beep perception; thus, the same participants from Experiment 1 were told to respond to report out how many beeps they heard, regardless of how many
flashes they saw. To ensure that they were paying attention to the visual stimuli and did not shut their eyes in the crossmodal condition, a green visual stimulus (small green square) was presented for each possible trial type, and participants were asked to hit the space bar instead of 2,3 , or 4 when they saw the green stimulus. Five participants were removed because they did not detect the green catcher stimulus on at least $75 \%$ of the trials.

## Results and Discussion

Accuracy See left side of Figure 4 for mean accuracies and standard errors on auditory response trials and right side of Figure 4 for unimodal visual responses. Experiment 2 focused exclusively on auditory responses. Using a 3 (number: 2, 3, 4) x 3 (trial type: baseline, congruent, incongruent) repeated measures ANOVA, a significant effect of number of beeps presented was found, $F(2,46)=19.15, p$ $<.001, \eta_{p}{ }^{2}=.45$. Based on the data, there was significantly higher accuracy on the 2-flash trials $(M=.80, S E=.03)$ than on 3-flash trials $(M=.67, S E=.04), t(23)=2.75, p=.011$, and 4-flash trials $(M=.48, S E=.05), t(23)=5.22, p<.001$. The 3-flash trials also had a higher accuracy than 4-flash trials, $t(23)=4.09, p<.001$. In addition, a condition x number interaction was observed, $F(4,92)=3.36, p=.013$, $\eta_{p}{ }^{2}=.13$. No differences were found across trial types for 2and 4-flash trials; however, congruent and incongruent trials both exceeded the baseline on 3-flash trials, $t \mathrm{~s}(23)>-1.94$, ps < . 033 (one-tailed).


Figure 4. Accuracies across number and trial type. Error Bars denote Standard Errors.

Fission Actual responses were collected on each trial and only fission trials and unimodal trials were submitted to a 2 (trial type: unimodal, fission) x 2 (number: 2, 3) repeated measures ANOVA. See left side of Figure 5 for means and standard errors. Using a 2 (condition: unimodal, fission) x 2 (number: 2, 3) repeated measures ANOVA, a significant effect of number of stimuli presented on response was found, $F(1,23)=160.07, p<.001, \eta_{p}{ }^{2}=.87$. The 2-flash trials $(M=$ $2.24, S E=.04$ ) had a significantly lower response than the 3flash trials $(M=2.96, S E=.06), t(23)=-12.65, p<.001$.

There was no effect of trial type, suggesting that flashes did not affect beep perception.

Fusion Actual responses were collected on each trial and only fusion trials and unimodal trials were submitted to a 2 (trial type: unimodal, fusion) x 2 (number: 3, 4) repeated measures ANOVA. See right side of Figure 5 for means and standard errors. A significant effect of number of stimuli presented on response was found, $F(1,23)=60.30, p<.001$, $\eta^{2}=.72$. The 3-flash trials $(M=2.95, S E=.06)$ had a significantly lower response than the 4-flash trials ( $M=3.40$, $S E=.08), t(23)=-7.77, \mathrm{p}<.001$. Again, there was no effect of trial type, suggesting that the flashes did not affect beep perception.


Figure 5. Actual responses across number and trial type. The left side of the figure denotes fission trials and the right side denotes fusion trials. Error Bars denote Standard Errors.

## General Discussion

Many tasks require processing and integration of multisensory information. The primary goal of the current study was to examine relative contributions of auditory and visual information on multisensory integration. In Experiment 1, we hypothesized that auditory information would have a strong effect on visual processing, as seen in Shams et al. (2000). The results of Experiment 1 supported this hypothesis. In particular, when auditory and visual information provided the same information (congruent trials in Figure 2), adults were more accurate at reporting the number of flashes. Moreover, incongruent trials also affected visual perception. Participants overestimated the number of flashes when the flashes were paired with more beeps (fission trials in Figure 3) and underestimated the flashes when paired with fewer beeps (fusion trials in Figure 3). In Experiment 2, it was hypothesized that the visual information would not have as strong of an effect on the auditory processing, based on auditory dominance (Robinson \& Sloutsky, 2010a) and the modality appropriateness hypothesis (Welch \& Warren, 1980). The results of Experiment 2 supported this hypothesis,
as most of the analyses showed that the visual information had no effect on auditory processing.

This expands the SIFI research by observing the effects of both auditory and visual information on multisensory integration. According to our knowledge, previous research has only focused on effects of auditory input on visual processing or vice versa; thus, these studies cannot determine if effects are symmetrical. The current study also used facilitation effects as a measure of multisensory integration. Facilitation effects were observed, and performance on the congruent trials was better than performance on the unimodal trials (baseline). These effects are seen in the visual responding condition with auditory input facilitating visual processing, but were not seen in the auditory responding condition. This asymmetry is consistent with both auditory dominance (Robinson \& Sloutsky, 2010a) and the modality appropriateness hypothesis (Welch \& Warren, 1980).

This study also expands modality dominance literature by measuring quantitative responses, rather than just response times, as seen in visual dominance research. Visual dominance has been observed for the past forty years in adults, showing that visual input often dominates auditory processing when making speeded, modality specific responses (e.g., Colavita, 1974). The findings of the current study were not tied to speeded modality specific responses, but were associated with accuracy of quantitative judgments. The findings support auditory dominance and modality appropriateness hypothesis and show that auditory input has a larger effect on visual processing than vice versa. Future research could take further measures to separate these two findings, as it cannot be distinctly determined whether the results are an effect of auditory dominance or modality appropriateness. Finally, it will be important to examine the role of stimulus intensity on multisensory integration, as changes in unimodal sensitivity may underlie developmental changes in multisensory integration. In particular, increased multisensory integration with age might stem from older adults compensating for weakened unimodal processing (DeLoss, Pierce, \& Anderson, 2013). While Anderson (2004) found that weakening the auditory stimulus to near threshold increased visual effects on multisensory integration, weakening both modalities tends to decrease the SIFI (Parker \& Robinson, in prep), and it is unclear how weakened auditory stimuli affect multisensory integration.

In summary, most of our experiences are multisensory in nature and it is important to understand how auditory and visual information contributes to multisensory integration. Future research needs to examine how this ability changes across the lifespan.

## References

Alais, D., \& Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Current Biology, 14(3), 257-262.

Bahrick, L. E., \& Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. Developmental psychology, 36(2), 190.
Burr, D., Banks, M.S., \& Morrone, M.C. (2009). Auditory dominance over vision in the perception of interval duration. Experimental Brain Research, 198(1), 49-57.
Colavita, F. B. (1974). Human sensory dominance. Perception \& Psychophysics, 16, 409-412.
Colavita, F.B., Tomko, R., \& Weisberg, D. (1976). Visual pre-potency and eye orientation. Bulletin of the Psychonomic Society, 8, 25-26.
Colavita, F.B., \& Weisberg, D. (1979). A further investigation of visual dominance. Attention, Perception \& Psychophysics, 25, 345-347.
Colonius, H., \& Diederich, A. (2006). The race model inequality: Interpreting a geometric measure of the amount of violation. Psychological Review, 113, 148-154.
Conway, C.M., \& Christiansen, M.H., (2005). Modalityconstrained statistical learning of tactile, visual, and auditory sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(1), 24-39.
DeLoss, D., Pierce, R., Andersen, G. (2013). Multisensory Integration, Aging, and the Sound-Induced Flash Illusion. Psychology and Aging, 28, 802-812.
Dennis, N.A., Howard, J.H., \& Howard, D.V. (2006). Implicit sequence learning without motor sequencing in young and old adults. Experimental Brain Research, 175, 153-164.
Egeth, H.E., \& Sager, L.C. (1977). On the locus of visual dominance. Attention, Perception \& Psychophysics, 22, 77-86.
Giard, M.H., \& Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. Journal of Cognitive Neuroscience, 11(5), 473-490.
Lewkowicz, D. J. (1988a). Sensory dominance in infants: 1. Six-month-old infants' response to auditory-visual compounds. Developmental Psychology, 24, 155-171.
Lewkowicz, D. J. (1988b). Sensory dominance in infants: 2. Ten-month-old infants' response to auditory-visual compounds. Developmental Psychology, 24, 172-182
Nissen, M.J., Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19, 1-32.
Parker, J.L., \& Robinson, C.W. (in prep). Aging in Multisensory Integration. Manuscript to be submitted for publication.
Robinson, C. W., \& Sloutsky, V. M. (2013). When audition dominates vision: Evidence from cross-modal statistical learning. Experimental Psychology, 60, 113-121.
Robinson, C. W., \& Sloutsky, V. M. (2010). Development of cross-modal processing. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 135-141.
Robinson, C. W., \& Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. Child Development, 75, 1387-1401.

Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926-1928.
Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., \& Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. Psychological Science, 8, 101-105.
Shams, L., Kamitani, S., Shimojo, S. (2000). Illusions. What you see is what you hear. Nature, 408, 788.
Shams, L., Kamitani, Y., \& Shimojo, S. (2002). Visual illusion induced by sound. Cognitive Brain Research, 14(1), 147-152.
Sinnett, S., Spence, C., \& Soto-Faraco, S. (2007). Visual Fdominance and attention: Revisiting the Colavita effect. Perception \& Psychophysics, 69, 673-686.
Sloutsky, V.M., \& Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. Child Development, 74(3), 822-833.
Song, S., Howard, J.H., \& Howard, D.V. (2008). Perceptual sequence learning in a serial reaction time task. Experimental Brain Research, 189, 145-158.
Spence, C., Parise, C., \& Chen, Y. C. (2012). The Colavita visual dominance effect. In M.M. Murray, \& M.T. Wallace (Eds.), The Neural Bases of Multisensory Processes (pp. 529-556). Boca Raton, FL: CRC Press.
Toro, J. M., Sinnett, S., \& Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. Cognition, 97, B25-B34.
Tsay, C. (2013). Sight over sound in the judgement of music performance. Proceedings of the National Academy of Sciences, 110(36), 14580-14585. Doi;10.1073/pnas. 1221454110
Vadillo, M. A., Konstantinidis, E., \& Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. Psychological Bulletin and Review, 23, 87-102.
Van der Burg, E., Olivers, C.N., Bronkhorst, A. W., \& Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. Journal of Experimental Psychology: Human Perception and Performance, 34, 1053-1065.
Wada, Y., Kitagawa, N., Noguchi, K. (2003). Audio-visual Integration in temporal perception. International Journal of Psychology, 50, 117-124.
Welch, R. B., \& Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. Psychological Bulletin, 88, 638-667.
Younger, B. A., \& Cohen, L. B. (1983). Infant perception of correlations among attributes. Infant Development, 54, 858-860.

# The effect of acute physical activity on children's memory for language 

Carla Pastorino Campos (cap72@cam.ac.uk)<br>Department of Theoretical and Applied Linguistics, 9 West Road Cambridge, CB3 9DP UK

John N. Williams (jnw12@cam.ac.uk)<br>Department of Theoretical and Applied Linguistics, 9 West Road Cambridge, CB3 9DP UK


#### Abstract

Research on the relationship between acute physical activity and cognition in children has often found beneficial effects of exercise on a variety of cognitive abilities. One domain that remains underexplored, however, is the relationship between exercise and long-term memory in children, and in particular whether the general-domain effects observed in previous studies could translate to a school-based learning activity, such as vocabulary learning. To address this issue, this study focused on the possible effects that a bout of moderate, aerobic physical activity could have on the immediate and delayed recall of newly acquired word forms and formmeaning connections of children in a school setting. In line with previous research, the results show a positive effect of exercise, but only for word form recall. This study expands our understanding of the differential effects of exercise on memory, while raising questions regarding the possible moderating influence of gender and memory consolidation.


Keywords: language acquisition, acute physical activity, memory, vocabulary learning, child cognition.

## Introduction

Research on the effects of physical activity has consistently shown that children who lead an active lifestyle also have healthier bodies, and that exercise plays a significant role in the prevention and control of certain diseases, including, but not limited to, obesity and cardiovascular disease (Kesaniemi et al., 2001). The effects of exercising, however, are not purely physiological. Over the last couple of decades, a growing body of research has concentrated on the effects of physical activity on the developing mind, mainly focusing on whether acute or chronic physical activity can influence cognition.

Interest in the effects of exercise on cognition was sparked by the observation that exercising provokes a series of transient cardiorespiratory, hormonal and metabolic changes that affect the brain's function and organization. These changes ultimately influence the way in which humans, and other animals, perform cognitive activities, including those involving executive function, attention, and memory (Coles \& Tomporowski, 2008; Hötting \& Röder, 2013; McMorris, Turner, Hale, \& Sproule, 2016). In recent work, memory consolidation has emerged as a process that could be particularly affected by physical activity (Robertson \& Takacs, 2017). The possible pathways by which exercise-induced arousal may affect long-term memory are still not entirely understood. However, it has
been suggested that exercise-induced upregulation of catecholamines, cortisol or brain-derived neurotrophic factor (BDNF), a protein shown to be involved in neurogenesis and neuroplasticity, could underlie the effects of acute physical activity on memory processes (McMorris et al., 2016; Roig et al., 2016).

Regarding child cognition, few studies have focused on the possible effects of physical activity on children's longterm memory. An exception would be the work of Pesce and colleagues (Pesce, Crova, Cereatti, Casella, \& Bellucci, 2009) who studied the performance of a group of 11 and 12-year-old students in an immediate and delayed ( 12 minutes later) free-recall test. The participants attended several learning sessions, two of which were preceded by a physical education class, one involving high cognitive and social demands (team games) and one with low demands (circuit training). Delayed recall improved for both exercise conditions when compared with the rest condition, whereas immediate recall improved only after the team games. More recently, Etnier, Labban, Piepmeier, David, \& Henning, (2014) observed children's performance on the Rey Auditory Verbal Learning Test - intended to evaluate verbal learning and memory, administered after a period of physical activity or rest. The participants were tested immediately and 24 hours later. Results showed that participants who exercised had better memory retention immediately, but not 24 hours later.

These studies provide interesting insights regarding the effect of exercise on children's memory, yet they focus on general-domain cognitive abilities, tested through tools that do not closely resemble ordinary classroom activities. In both experiments, the participants were tested on their recall of already known words, a task that, though related to different aspects of memory, does not involve learning new linguistic information, such as word forms or form-meaning connections. Learning such linguistic information is at the core of one standard part of school curricula around the world: second/additional language learning. Therefore, it would be of interest to extend the findings of these studies by assessing whether the advantages noticed in these general memory tasks could translate to a particular language learning activity, such as vocabulary learning.

There is some precedent for this type of studies in research done with adults. For example, despite some differences in the experimental designs, benefits of a moderate to intense bout of exercise were found for
vocabulary learning, achieved through the association of translation equivalents (Schmidt-Kassow et al., 2013; Schmidt-Kassow, Kulka, Gunter, Rothermich, \& Kotz, 2010) or through a statistical, associative learning paradigm (Winter et al., 2007). These encouraging results, however, have not yet been observed in children.

Based on the brief review of the literature presented above, the primary purpose of this study was to assess the possible effects of a bout of aerobic, moderate physical activity on children's memory for language learning. In particular, we were interested in observing whether exercising before encoding could lead to a higher rate of recall immediately after learning and after a delay. Acquiring new vocabulary items is a complicated process that involves many different learning activities, from item segmentation to the integration of novel items to existing networks. In this instance, however, we have chosen to focus on only two of these activities: the acquisition of the word's phonological/orthographical form and the linking of the newly acquired form to a meaning. From the results of previous studies and the theorized mechanisms underlying the exercise-cognition effect, we hypothesize that children in the exercise group would exhibit increased recall of both forms and meanings in the delayed tests, after a period of consolidation including sleep.

## Method

## Participants

51 school-aged children (mean age $=9.3,25$ females) were recruited at their school to participate in this experiment. All children were monolingual Spanish speakers, with no knowledge of other languages besides school-level English. None of the children had visual or hearing impairments, nor cardiac or respiratory conditions that would prevent them from exercising at a moderate pace. As reported by their parents or tutors, the children had normal sleeping patterns, with an average sleeping time of 8.5 hours per night.

## Materials

Pseudowords Twenty-four bisyllabic, pronounceable pseudowords were created using legal Spanish consonant/vowel combinations. All pseudowords were four letters long and followed a CVCV pattern. Care was taken to ensure that an equal amount of pseudowords were 'masculine' (ended with 'o') and 'feminine' (ended with ' $a$ ').

Pictures Twenty-four pictures of everyday objects extracted from a subset of Rossion \& Pourtois' (2001) pictorial set, which was based on Snodgrass \& Vanderwart's (1980) original standardized set of visual stimuli but with added detail and colour. For this experiment, we transformed the pictures to greyscale to preserve the enhanced details while avoiding the distraction of colour. Pertinent country-specific normative data was used to assess the concreteness and imagery ratings of the objects depicted (Manoiloff, Artstein,

Canavoso, Fernández, \& Segui, 2010). Following Pesce et al's (2009) procedure, the objects were evaluated by classroom teachers to ensure that they would be recognizable and familiar to the children.

Learning lists Each pseudoword was randomly paired with one of the pictures selected. We divided the pseudoword/picture pairings into two lists of 12 items, that would be presented in different experimental sessions. The order of presentation of both the learning and testing stimuli was randomized for each group.

## Procedure

The experiment used a within-subjects design, counterbalancing the order in which the children participated in each of the experimental conditions, exercise, and no-exercise (control).

The experiment was carried out at the participants' school, during school hours, with the assistance of classroom and physical education teachers. Participants attended six sessions spread over three weeks. The first and fourth sessions started with an intervention stage, where participants either performed a 30 -minute bout of aerobic physical activity (exercise condition) or remained in the classroom doing a passive activity (drawing and colouring) for the same amount of time (control condition). After the intervention, and after a 5 -minute recovery time for the exercise group, the children performed the vocabulary learning activity. In the final stage of the first session, the children were asked first to write down all the words they remembered, regardless of the order of presentation (free recall) and later to write down the words corresponding to the pictures shown in posters by the teachers (cued recall). The free and cued recall tasks constituted the first testing session (immediate test).

This testing phase was repeated on sessions two and three for the pairings learnt in session one, and on sessions five and six for the pairings learnt in session four.

All participants were exposed to both learning lists and took part in both intervention conditions (exercise or control); the order of the interventions was counterbalanced and randomized so that some participants exercised on session one while others exercised on session four.

The experimental procedure is schematized in Figure 1.
Vocabulary learning activity The children were exposed to the lists of 12 pseudoword/picture pairings, presented in printed posters, and shown at a regular interval. The classroom teachers led the presentation, reading the pseudoword aloud and asking the children to repeat it back. This exposure was repeated three times.

Testing tasks Two testing tasks were intended to separately measure children's ability to recall the phonological form as well as the form-meaning connections of the pseudowords learnt in the exposure phase.


Figure 1: Experimental procedure. L1 = word list 1, L2 = word list $2,24 \mathrm{~h}=24$ hours later, $6 \mathrm{~d}=6$ days later, $1 \mathrm{wk}=1$ week later.

The first task, free recall, was designed to assess whether the participants could remember the phonological forms of the newly acquired pseudowords, without prompting from their associated meanings. To perform this task, the teachers instructed the students to write down all the pseudowords they could remember, irrespective of the order of presentation or their ability to associate the words with the drawings.

For the second task, cued recall, the teachers showed the students only the pictures, and the children had to write down which pseudoword corresponded to each picture shown. The pictures were shown at a steady pace, and the children were instructed to leave blank spaces for the pictures they did not remember. This test intended to observe if children could recall the form-meaning connections they had previously learnt.

Physical activity intervention The physical activity intervention consisted of 30 minutes of child-adapted circuit training, focalised on activities that would engage the aerobic system. We chose this type of exercise task since it maintains a major focus on aerobic exercise - as opposed to group games, for example, that may be more cognitively and socially demanding - while still being part of what students normally do in their physical education classes, and thus ecologically valid. The task took place in the school's playground under the supervision of the children's physical education teachers (student-to-teacher ratio $=15: 1$ ).

We used a modified Borg Scale of Perceived Exertion (Borg, 1998) to assess exercise intensity. It was administered to the children in the exercise condition while
they were performing the activities. This scale has been widely used for children in similar contexts and has proved to be reliable not only in assessing intensity but also in determining the nature of the exercise being conducted (aerobic versus anaerobic). By maintaining the general perceived exertion of the group in the second tier (considered "moderate" in our scale), it was assured that the children were performing aerobic exercise at a moderate intensity.

## Results

## Data pre-processing

Ten participants were removed from the final sample, two for reporting learning or psychological conditions that might interfere with the experiment's outcome and eight for not having participated in either of the two experimental conditions (exercise or control). Given that a significant number of children were absent on the day of the very delayed test, for reasons not related to the experiment, we excluded the very delayed test (sessions three and six) in the reported analyses. To keep the design balanced, we additionally removed two participants for not having completed the delayed testing session. All analyses were carried out on the remaining 39 participants (mean age $=$ 9.32, 21 females).

We computed an accuracy score for each testing activity (free or cued recall) by summing all the correct responses given at each testing time. Participant responses were given one point when they matched exactly one of the taught pseudowords (e.g. lofa/lofa), and half a point if the answer had one substitution (e.g. lofa/lifa). For the cued recall task, the responses had to match one of the pseudowords in the taught set in addition to matching the corresponding picture. Partial matches that were placed with the correct picture were awarded half a point.

All statistical analyses were performed using R (R Core Team, 2016) and the ez package (Lawrence, 2016).

## Free recall

A two-way analysis of variance was conducted to evaluate the effects of the experimental condition and testing time on the number of items recalled in the free recall task. Experimental condition (exercise vs. control) and testing time (immediate vs. 24 hours later) were included as factors. No significant effect of experimental condition, $\mathrm{F}(1,38)=$ $1.092, \mathrm{p}=.302$, or testing time, $\mathrm{F}(1,38)=.032, \mathrm{p}=.574$ was found. However, the interaction between exercise intervention and testing time was significant, $\mathrm{F}(1,38)=$ 5.932, $\mathrm{p}=.019, \eta_{\mathrm{g}}{ }^{2}=.006$ ). As shown in Figure 2, it would seem that, whereas there is no difference in the immediate test for the experimental conditions (exercise: $\mathrm{M}=5.97$ [49.7\%], $\mathrm{SD} \pm 2.73$; control: $\mathrm{M}=6.07$ [50\%], $\mathrm{SD} \pm 2.59$ ), participants recalled more items in the delayed test if they had exercised prior to encoding (exercise: $\mathrm{M}=6.26$ [52\%], $\mathrm{SD} \pm 2.71$; control: $\mathrm{M}=5.5$ [45.8\%], $\mathrm{SD} \pm 2.94$ ).


Figure 2. Free recall correct responses per condition and testing time.

Gender differences were also taken into account in the analysis. Figure 3 displays the number of correct responses in each experimental condition, grouped by gender and testing time. From this graph, it is possible to see that, overall, girls seem to remember the same or a larger number of items when compared to boys. Furthermore, the profiles of performance per experimental condition appear to be similar for both genders: when in the exercise condition, the number of accurate responses is either maintained or increases from the immediate to the delayed test, whereas when in the control condition performance decreases. However, while the girls' performance does not seem to be particularly affected by the exercise intervention in either of the tests (exercise immediate: $\mathrm{M}=5.28$ [44\%], $\mathrm{SD} \pm 2.72$; exercise delayed: $\mathrm{M}=5.85$ [48.8\%], $\mathrm{SD} \pm 2.86$; control immediate: $\mathrm{M}=6.04$ [50\%], $\mathrm{SD} \pm 1.8$; control delayed: $\mathrm{M}=$ 5.71 [47.6\%], $\mathrm{SD} \pm 2.93$ ), the boys recalled more items immediately (exercise: $M=5.22$ [43.5\%], $\mathrm{SD} \pm 2.79$; control: $\mathrm{M}=4.33$ [ $36 \%$ ], $\mathrm{SD} \pm 2.95$ ) and 24 hours (exercise: $\mathrm{M}=5.33$ [44.4\%], $\mathrm{SD} \pm 2.63$; control: $\mathrm{M}=4$ [33.3\%], $\mathrm{SD} \pm$ 2.76) when they exercised before encoding.

Given the unequal number of boys and girls in the sample (female $=21$, male $=18$ ), which limits the possibility of comparing the groups, a two-way repeated measures ANOVA, including experimental condition and testing time as factors, was conducted on a reduced dataset comprising only the boys' data. The effect of experimental condition was significant, $\mathrm{F}(1,17)=5.101, \mathrm{p}=.037, \eta_{\mathrm{g}}{ }^{2}=.035$, indicating that the difference between experimental conditions (exercise vs. control) observed in Figure 3, albeit numerically small, may be worth further exploration.

## Cued recall

To assess the effects of experimental condition and testing time, a two-way, repeated-measures ANOVA was performed. As shown in Figure 4, no significant main effects of experimental condition, $\mathrm{F}(1,38)=.026, \mathrm{p}=.872$, or testing time, $\mathrm{F}(1,38)=.638, \mathrm{p}=.429$, were found. The interaction between both independent variables was also
non-significant, $\mathrm{F}(1,38)=.014, \mathrm{p}=.9, \eta_{\mathrm{g}}{ }^{2} \leq .001$. These results indicate that exercise did not influence the cued recall of form-meaning connections in this sample, either immediately (exercise: $\mathrm{M}=5.43$ [45.2\%], $\mathrm{SD} \pm 2.93$; control: $\mathrm{M}=5.51$ [45\%], $\mathrm{SD} \pm 2.52$ ) or 24 hours after encoding (exercise: $\mathrm{M}=5.64$ [47\%], $\mathrm{SD} \pm 2.48$ ); control: M $=5.66[47.2 \%], \mathrm{SD} \pm 3.05)$.


Figure 3. Free recall correct responses by gender, experimental condition and testing time.

A visual inspection of the plotted cued recall data divided by gender showed no indication of a differential effect of exercise. Hence no additional analyses were conducted.


Figure 4. Cued recall correct responses per condition and testing time.

## Discussion

The purpose of this study was to assess whether a bout of aerobic, moderate physical activity performed before encoding could affect children's retention of newly acquired pseudowords and their meanings, immediately after learning and with a delay. In particular, we were interested in observing the possible effects of physical activity in the recall of novel word-forms (pseudowords) and meanings
(pseudoword/picture association). These questions were motivated by the scarcity of studies focusing on exercise and language learning, particularly with child participants.

The free recall test focused on the children's retention of word forms. Data from this task showed that exercising before encoding facilitated recall, but only on the delayed test (24 hours after learning). This outcome may indicate that exercising before learning could help reduce timedependent forgetting on a moderate scale. This finding concurs with previous research outlining the beneficial, rather than detrimental, effects of acute exercise on children's recall (Etnier et al., 2014; Pesce et al., 2009). However, in contrast to these studies, in which increased memory performance was noticed in the immediate test (or after a very short delay), in this experiment the influence of exercise appears to affect the delayed test, performed 24 hours later and after a period of sleep. This may indicate a possible link between acute physical activity and memory consolidation. A relationship between exercise and consolidation has already been postulated for healthy adults (e.g. Coles \& Tomporowski, 2008; Labban \& Etnier, 2011; Robertson \& Takacs, 2017), but would require further exploration to extend it to child memory consolidation. The discrepancy between the results of the present study and previous work could also suggest that, consistent with several meta-analyses (Chang, Labban, Gapin, \& Etnier, 2012; Lambourne \& Tomporowski, 2010; Tomporowski, 2003), the effects of physical activity on cognition could be selective. Since the learning activities used in Etnier and colleagues and Pesce and colleagues' work were focused on the recall of already known linguistic information, and not on the acquisition of entirely novel forms as was the case in this experiment, the tasks rely on different memory processes that may be differentially affected by exercise.
A bout of physical activity before encoding did not affect recall of form-meaning connections. This finding is somewhat unexpected, given that previous research had found an influence of bouts of aerobic physical activity on associative memory in adults (Schmidt-Kassow et al., 2013; Schmidt-Kassow, Kulka, Gunter, Rothermich, \& Kotz, 2010; Winter et al., 2007). Since this experiment required children to learn forms and form-meaning connections simultaneously, it may be that the acquisition of the form was privileged to the detriment of the formation of formmeaning connections. This could have been reinforced by the order of the testing activities - always free recall followed by cued recall - that could have made the recall of forms more prominent. To our knowledge, this is the first study that addresses the effect of a bout of physical activity on the formation of form-meaning connections in children; as such, it only begins to describe the effects observed. Further research could help elucidate whether the effects of exercise seen on standardized associative memory tests could transfer to memory-supported activities that more closely resemble the tasks performed by children in a learning environment.

Another surprising effect was the difference observed in boys and girls. A tentative explanation of this difference, based on the assumption that exercise works as a stressor, would fall in line with research highlighting gender-based differences in brain reaction to stressors, both of an emotional and physical nature (for a review, see Cahill, 2006). Furthermore, because this gender difference appears on the second day, the possibility of the existence of gender differences in offline consolidation, that could be enhanced by exercise, should also be considered. Some evidence points towards greater motor skill memory gains obtained by males after a period of offline consolidation (Dorfberger, Adi-Japha, \& Karni, 2009), an advantage that may be related to differential responses to cortisol levels (Andreano \& Cahill, 2006). Cortisol is a stress hormone shown to be affected by acute exercise interventions, and that plays a role in memory consolidation. It should nonetheless be noted that since not all previous studies on the effects of exercise on cognition have addressed gender differences, and as it was not the primary purpose of this study, it is hard to draw conclusions from this result.
There are some limitations that should be considered in relation to the current study. First, the sample size is relatively small when considering the number of variables addressed. Furthermore, the effects found, although significant, were subtle and can therefore only be considered with caution. Future experiments should also include more information about the sample, including a measure of fitness level and baseline memory performance, as well as some indication of school achievement (e.g. grades). The inclusion of these data could help disentangle variation that might be motivated by external factors not related to the experimental manipulation.
Overall, this study expands our knowledge of the effects of acute, aerobic physical activity on children's cognition, in that it includes a 'learning element' that had been thus far overlooked. The fact that participants had to learn novel word forms, as well as their connection to meaning, after exercising makes this experimental activity more closely related to regular classroom language learning activities, thus providing an initial glimpse into the effects of physical activity in a school environment. The finding that the experimental condition affected free- but not cued-recall in this experiment could suggest a selective effect of exercise on children's memory, but it could also underline the need to utilize more nuanced tests to assess associative memory in this context. Future research on this topic could build upon these findings, by adding more and more sophisticated tests of relational memory, as well as addressing some of the questions that this study has raised, such as the possible moderating effect of gender and the influence of acute physical activity on memory consolidation in children.

## References

Andreano, J. M., \& Cahill, L. (2006). Glucocorticoid release and memory consolidation in men and women. Psychological Science, 17(6), 466-470.

Cahill, L. (2006). Why sex matters for neuroscience. Nature Reviews Neuroscience, 7(6), 477-84.
Chaddock, L., Erickson, K. I., Prakash, R. S., Kim, J. S., Voss, M. W., Vanpatter, M., ... Kramer, A. F. (2010). A neuroimaging investigation of the association between aerobic fitness, hippocampal volume, and memory performance in preadolescent children. Brain Research, 1358, 172-183.
Chaddock, L., Hillman, C. H., Buck, S. M., \& Cohen, N. J. (2011). Aerobic fitness and executive control of relational memory in preadolescent children. Medicine and Science in Sports and Exercise, 43, 344-349.
Chang, Y. K., Labban, J. D., Gapin, J. I., \& Etnier, J. L. (2012). The effects of acute exercise on cognitive performance: A meta-analysis. Brain Research, 1453(250), 87-101.
Coles, K., \& Tomporowski, P. D. (2008). Effects of acute exercise on executive processing, short-term and longterm memory. Journal of Sports Sciences, 26(3), 333344.

Dorfberger, S., Adi-Japha, E., \& Karni, A. (2009). Sex differences in motor performance and motor learning in children and adolescents: An increasing male advantage in motor learning and consolidation phase gains. Behavioural Brain Research, 198(1), 165-171.
Etnier, J. L., Labban, J. D., Piepmeier, A. T., David, M. E., \& Henning, D. A. (2014). Effects of an acute bout of exercise on memory in 6th-grade children. Pediatric Exercise Science, (March 2015), 250-258.
Hötting, K., \& Röder, B. (2013). Beneficial effects of physical exercise on neuroplasticity and cognition. Neuroscience and Biobehavioral Reviews, 37(9), 2243-2257.
Kesaniemi, Y. K., Danforth, E., Jensen, M. D., Kopelman, P. G., Lefèbvre, P., \& Reeder, B. A. (2001). Doseresponse issues concerning physical activity and health: an evidence-based symposium. Medicine and Science in Sports and Exercise, 33(6), S351-S358.
Labban, J. D., \& Etnier, J. L. (2011). Effects of acute exercise on long-term memory. Research Quarterly for Exercise and Sport, 82(4), 712-721.
Lambourne, K., \& Tomporowski, P. (2010). The effect of exercise-induced arousal on cognitive task performance: A meta-regression analysis. Brain Research, 1341, 12-24.
Manoiloff, L., Artstein, M., Canavoso, M. B., Fernández, L., \& Segui, J. (2010). Expanded norms for 400 experimental pictures in an Argentinean Spanishspeaking population. Behavior Research Methods, 42(2), 452-460.
McMorris, T., Turner, A., Hale, B. J., \& Sproule, J. (2016). Beyond the catecholamines hypothesis for an acute exercise-cognition interaction: A neurochemical perspective. In T. McMorris (Ed.), Exercise-Cognition Interaction: Neuroscience Perspectives (pp. 65-103). New York, NY: Academic Press.
Pesce, C., Crova, C., Cereatti, L., Casella, R., \& Bellucci,
M. (2009). Physical activity and mental performance in preadolescents: Effects of acute exercise on freerecall memory. Mental Health and Physical Activity, 2, 16-22.
Robertson, E. M., \& Takacs, A. (2017). Exercising Control Over Memory Consolidation. Trends in Cognitive Sciences, 21(5), 310-312.
Roig, M., Thomas, R., Mang, C. S., Snow, N. J., Ostadan, F., Boyd, L. A., \& Lundbye-Jensen, J. (2016). Timedependent effects of cardiovascular exercise on memory. Exercise and Sport Sciences Reviews, 44(2), 81-88.
Rossion, B., \& Pourtois, G. (2001). Revisiting Snodgrass and Vanderwart's object database: Color and texture improve object recognition. Journal of Vision, (1.3), 413-413.
Schmidt-Kassow, M., Deusser, M., Thiel, C., Otterbein, S., Montag, C., Reuter, M., ... Kaiser, J. (2013). Physical exercise during encoding improves vocabulary learning in young female adults: a neuroendocrinological study. PloS One, 8(5), e64172.
Schmidt-Kassow, M., Kulka, A., Gunter, T. C., Rothermich, K., \& Kotz, S. A. (2010). Exercising during learning improves vocabulary acquisition: behavioral and ERP evidence. Neuroscience Letters, 482(1), 40-44.
Snodgrass, J. G., \& Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6(2), 174-215.
Tomporowski, P. D. (2003). Effects of acute bouts of exercise on cognition. Acta Psychologica, 112, 297324.

Winter, B., Breitenstein, C., Mooren, F. C., Voelker, K., Fobker, M., Lechtermann, A., ... Knecht, S. (2007). High impact running improves learning. Neurobiology of Learning and Memory, 87(4), 597-609.

# Replacing Language: Children Use Non-Linguistic Cues and Comparison in Category Formation 

Margarita Pavlova (margarita.velinova@gmail.com)<br>Department of Cognitive Science and Psychology, New Bulgarian University, 21 Montevideo Street<br>Sofia, Bulgaria


#### Abstract

Language is a powerful instrument for extracting relational information from stimuli. In a label extension task common labels invite comparison processes that help children focus on the more subtle relational similarity and away from the readily available perceptual similarity of the stimuli. The current experiment aims to explore whether non-linguistic representations of category membership are sufficient to invite such abstractions of relational information. Preschool children were asked to extend a category to either a relational or an object match. When given the opportunity to compare two instances of the category, and provided with a nonlinguistic cue children extended the category to the relational match. These results further extend the benefit of comparison in learning, and suggest that language labels are not the only cue children can use in category formation.


Keywords: categorization; cognitive development; relational processing; non-linguistic representations of relations.

## Introduction

Analogical reasoning - the ability to see and use relational similarity between situations and events lies in the core of human cognition (Hofstadter, 2001). It is what makes humans so smart and it is potentially what distinguishes us from other species (Gentner, 2003; 2010). Analogy is central to many cognitive processes including learning, reasoning, decision-making, and categorization (Gentner, 1983; Gentner \& Markman, 1997; Kokinov, 1998) and it promotes conceptual development in children (Christie \& Gentner, 2010; Gentner \& Namy, 1999; Graham, Namy, Gentner, \& Meagher, 2010; Rattermann \& Gentner, 1998). Thus, analogy is a key process of higher-order cognition that benefits learning (Gentner, 2010; Kokinov, 1998).

There is evidence that young children show analogical abilities, if they have knowledge of the relations involved (Gentner, 1983; 1988; Holyoak, Junn, \& Billman, 1984). However, children tend to focus first on object similarities, before they start to notice common relational structure, independent of the objects involved. Gentner defines this phenomenon as the relational shift hypothesis (Gentner, 1988). She asked children to interpret different kinds of metaphors and say how they are alike. When children were asked to interpret the metaphor "A tire is a shoe", 5-6-yearolds replied based on perceptual similarity (e.g. both are black), and $9-10$-year-olds based their answer on the specific roles and functions of the two, thus giving a relational answer [e.g. you can go places with both (Gentner, 1988)]. When asked to perform a mapping task,
children first base their reasoning on the salience of object features (e.g. color, shape, etc.) thus failing to map the relations involved. To have a true relational ability, children must resist the temptation of the object features, and base their reasoning on the more subtle common relations involved (Gentner \& Toupin, 1986).

Rattermann \& Gentner (1998) provided further support for the relational shift hypothesis. They gave children a relational mapping task in which the experimenter and the child each had a set of three objects. The experimenter hid a sticker under one of her objects (e.g., the middle one). The child had to find his sticker in the same place (the middle of his objects). In some of the conditions, objects were crossmapped, i.e. the middle object in the experimenter's set was the same size as the leftmost object in the child's set. In this condition, 3 -year-old children had a difficulty resisting the perceptual similarity and instead of searching under the corresponding relational location, they searched under the identical object in their set. In contrast, 5-year-olds were better able to resist the object matches and give relational matches.

When searching for an interpretation of a given similarity, people (especially young children and novices) first focus on the object commonalities (Christie \& Gentner, 2010; Kotovsky \& Gentner, 1996). However, interpretations based on object attributes are not useful in deriving causal principles. Potential analogs are more difficult to notice because relations are more subtle and require a deeper analysis of the information. However, once found, the analogy is very useful in deriving key principles, since the structure holds true for both the base and the target (or the two situations), regardless of the objects involved in it. There are two factors that contribute to relational reasoning - relational comparison and relational language.

## Relational Language and Learning

The first question is whether providing children with a noun label would help them learn a novel category. Evidence suggests that providing children with a count noun may bias them toward an object-centered interpretation of what kind of members are to be included into the category, since labels invite children to group things of like kinds together (Markman, 1989). Young children often base their reasoning on the more compelling perceptual similarity, and thus their intuition of likeness relies on object similarity. On the other hand, providing a common term could also serve
as an invitation to search for and find a common relation, thus promoting relational abstraction (Gentner, 2010).

When exploring children's understanding of categories, researchers use the word-extension task in which children are taught a new word and are given an example for it and then asked to extend the word to another example, thus showing an understanding of the category membership of the new item. Common labels invite deep reasoning and help children focus on like kinds. For example, in a series of experiments Gentner and Namy (1999) explored children's categorization abilities. The results show that children group items based on shape when they examine them in isolation (i.e. when they see only one standard). However, providing a label enhances the likelihood of children engaging in comparison, prompting children to compare the items bearing the same label.

Christie and Gentner (2010) extended these findings and showed that 3 - and 4 -year-old children successfully learned a novel label and extended category membership to the relational match. They explored how children base their hypotheses for category membership and showed the mutual benefit of comparison and relational label. When children had the opportunity to compare two examples of a given (novel) category, and heard a novel label, they extended the category to a relational match. This research shows the mutual bootstrapping between relational language and analogical abilities in preschool children (Christie \& Gentner, 2010; Gentner, 2010).

Language and analogical comparison interact in the process of learning. This claim is supported by the language as a toolkit view, proposed by Gentner (2003; 2010). According to this view, acquiring a language provides new resources that support cognitive skills, while not replacing prelinguistic abilities. Specifically, this view assumes that structural alignment supports language learning, and that relational language supports structural alignment and reasoning. In addition, Gentner discusses four ways in which language interacts with analogical abilities to foster learning (Gentner, 2003; 2010). First, common labels invite comparison and abstractions, thus prompting children to compare two items that share the same label. Second, a linguistic label helps to preserve the abstraction derived from the comparison and makes it more accessible for future use, thus promoting reification. Next, naming promotes uniform relational encoding, which ensures the encoding of the relations in the same manner on different occasions. Last, the systematic structure of language can invite conceptual structure. Thus, language and analogical comparison interact with analogical abilities to foster learning and development (Gentner, 2010).

## Relational Comparison and Learning

Analogical comparison promotes learning via a structural alignment process that is akin to relational mapping, thus highlighting the common relational structure and rendering it more salient (Gentner, 2010; Gentner \& Markman, 1994; 1997). Similarity comparison process is one of alignment
and mapping of common relational structure, like the structure-mapping process of analogy. A result from carrying out a similarity comparison is that it highlights the relational structure and makes it more salient, thus enabling further abstractions and use. The alignment hypothesis assumes that the process of making a similarity comparison may lead to change in the representation. This change in turn will increase the uniformity of the two representations (Gentner, 1983; Gentner \& Markman, 1997). Thus, alignment makes the relational commonalities more salient and the representations uniform. This typically increases the perceived similarity between the paired items. For example, Gentner and colleagues conducted a series of experiments that investigated the effects of comparison and common labels in children's categorization. The results show that when preschool children saw only one instance of a particular category, they extended the category to the perceptually similar match. In contrast, when children saw two examples simultaneously and were prompted to compare them, they were more likely to extend the category to a new, structurally similar item, than to a perceptually similar one (Christie \& Gentner, 2010; Gentner \& Namy, 1999; Graham, et al., 2010).

Gentner and Namy (1999) sought out to investigate how children form categories. They gave 4 -year-olds a novel label (e.g. a blicket) for a pictured object (e.g. a bicycle) and asked children to find another blicket between two alternatives: a perceptually similar, but taxonomically different object (e.g. eyeglasses) or a perceptually different object from the same category (e.g. a skateboard). When children were presented with only one example of the category (e.g. a bicycle or a tricycle), they tended to choose the perceptual match. However, when they observed two examples (a bicycle and a tricycle), they were more likely to choose the relational match. Interestingly, the obtained results cannot be accounted to a traditional view in which comparison is considered a simple feature overlap. Rather, it seems that comparison selectively highlighted the relational commonalities (Gentner \& Namy, 1999; Gentner, 2010).

Christie and Gentner (2010) followed this procedure and further extended the findings, showing the benefits of comparison in learning new relations. They presented 3- and 4-year-olds with animals in a novel spatial orientation and attached a novel label to it (e.g. a dax). When presented with only one example, or when two examples were provided but children were not prompted to compare them, they extended the category to the object match. However, children who compared the two examples of the category extended it to a relational match.

These results show that analogical comparison is useful in learning new principles, forming new categories, and retaining material better for transfer. One cannot help but see a tendency in all experiments discussed above - the mutual presentation of two (versus one) examples and providing a common label. It would be interesting to see if these two factors work mutually or if they could contribute
to category formation separately. This question is of particular interest in the present study.

## Can language be replaced?

The main question we are asking here is whether language is unique in promoting analogical abstraction. Mutafchieva \& Kokinov (2007) explored the hypothesis that a nonlinguistic representation of specific relations would be beneficial in a relational mapping task. Following the procedure of Rattermann \& Gentner (1998), they used labels (e.g. Daddy, Mummy, and Baby), a train analogy, or a physical representation of the relation pulling in the analogy (e.g. drawbars). The hypothesis was that the drawbars should be sufficient for the child to abstract the relation, and thus solve the mapping task. The results showed no difference between the various types of presentation (i.e. language labels, drawbars, or analogy). Interestingly, there was no evidence that providing labels further benefit performance on a mapping task. The drawbar condition seemed to be successful in promoting relational matches, similar to the labels, and the analogy condition.

Similarly, Gentner and colleagues conducted a series of experiments on the acquisition of relational categories that show that for 4 -year-olds comparison alone without relational language is sufficient to invite relational responding (Gentner, Anggoro, \& Klibanoff, 2011). In addition, Gentner, Namy and colleagues examined the role of comparison and shared names in categorization of novel objects. For example, Graham et al. (2010) gave 4-year-olds novel object sets that consisted of one, or two standards and two test objects to choose from, a texture match and a shape match. The results of the study are quite interesting. When children were presented with one standard, they extended the category based on shared shape, regardless of whether the objects were named. When children were presented with two standards that shared the same texture and the objects were named with the same noun, they extend category based on shared texture. Interestingly, the opportunity to compare in the absence of a shared label, led to an attenuation of the effect of shape, although not to a significant preference of texture. Interestingly, the authors found that adding a common label by itself did not change children's responding, however, it seems to augment the effect of comparison in shifting children toward the texture response (Graham et al., 2010).

One possible explanation of the obtained results is that in the No-word condition the authors provided children with a broad term (e.g. pointing to the standard(s): "This is one!"), thus limiting the possibility that the child could abstract a specific category cue. The term used is too general for children to elicit specific category representations and abstract common relational features between the two standards. On the other hand, the interesting fact that even without labels but with the opportunity to compare, children still could start to resist the perceptual match (e.g. shape), leaves a possibility that another type of non-linguistic representation of the category membership in addition to
comparison could benefit categorization among preschool children.

Continuing this line of research, the present experiment aims to explore the possibility that comparison is sufficient not only to promote abstraction of relational information, but also to aid category formation. The present study suggests that a non-linguistic cue in addition to comparison would successfully promote relational matches in a categorization task, thus showing that language labels are not unique in promoting relational categorization.

## Experiment

The goal of this study is to explore the possibility that children can use non-linguistic cues when categorizing items. If language (a common label) is unique in promoting relational categorization, then it should be sufficient. Moreover, when deprived from the opportunity to benefit from a common label, children should fail to extend category membership based on relational similarity. However, if another mechanism (comparison) is present, and children rely on it during categorization, then it is possible that a non-linguistic cue will provide sufficient ground for children to extend the category based on relational similarity. Specifically, a non-linguistic cue that represents the category membership of two standards (e.g. a sticker) could provide enough ground for children to extend the category membership based on relational similarity. Bearing in mind that young children typically are tempted by the perceptual similarity and often fail to notice the relational similarity between two instances, it is important to investigate the various strategies children use to group the things they encounter into categories.

Following the idea that providing a means for category inclusion (whether the cue is linguistic or not) will further benefit children in their performance, here we test the roles of comparison and various types of cues. Comparison enables children to abstract the interconnected relational structure and focus on the commonalities between the two examples, especially shared relations. This helps children to disambiguate between two working hypothesis: object match vs. relational match. The specific cue provides further support and acts to focus their attention to the underlying common relations. Thus, the highlighted structure will become more salient and more available to new examples of the category.

The specific hypothesis of the study was that children in the Label and Sticker Condition would choose the relational match more often compared to the children in the NoWord Condition. In addition, children in the Label and Sticker Condition will do equally well. This logic follows from the language as a toolkit view discussed earlier (Gentner, 2003; 2010). Building upon this view, I posit that a non-linguistic cue will act in the same way as novel language labels do, prompting children to go beyond the readily available perceptual commonalities, and focus on the deeper relational commonalities.

## Methods

Participants Forty-three 4 -year-olds were recruited from kindergartens in Sofia area ( 28 females, $M_{\text {age }}=52$ months; 15 in Label Condition, 14 in Sticker Condition, and 14 in NoWord Condition). Permission to participate was obtained from their parents prior to the study. Children received a small gift for their participation.

Materials The study uses the materials from the original study ${ }^{1}$. The instructions were translated into Bulgarian. Children were given a word extension task on a triad of pictures that depicted animals. The two standards were labelled with a novel noun, and children were asked to extend the label to one of two alternatives: a relational match (new animals in the same configuration) or an object match (same animal[ $[\mathrm{s}]$ in different configuration).

The stimuli consisted of eight sets of animal pictures, with two standards, an object match and a relational match. Each picture depicted two or three animals configured in a novel spatial relation (e.g., two identical pigs facing each other). The second standard within a given set showed different animals in the same spatial configuration (e.g., two identical fish facing each other). The object match contained an exact animal match from each standard but in a different relational pattern (e.g., a pig and a fish turned back on each other). The relational match was composed of new animals in the same relational configuration as the two standards (e.g., two identical turtles facing each other; Figure 1).

In addition, two training sets depicting shapes were included that aimed to help children become more familiar with the procedure. Children did not receive feedback during the training session and the results from it were not counted in the analyses.

Procedure Children were randomly assigned to one of three between-subjects conditions: Label, Sticker, or NoWord. Materials were presented on laminated paper cards. Children were seated across from an experimenter.

In the Label condition, the experimenter laid the two standards and labeled them with a novel count noun (e.g. pointing to the first card, "This is a blicket." Then, pointing to the second card, "And this is a blicket, too."). Next, the child had to compare the two standards: "Do you see why these two are both blickets?" The experimenter then placed the two alternatives side by side below the standards and asked the child, "Which one of these is also a blicket?" After the child made a choice, the experimenter continued with new standards from a new set. Eight unique novel labels were used, one for each relational pattern.

The NoWord condition began the same way. The experimenter laid the two standards, but instead of labelling them with a novel word, she used the same generic term for all sets: (e.g. "This is one. And this is one, too."). Then, the child was prompted to compare the standards, "Do you see why these two are the same kind of thing?" Last, the two

[^429]alternatives were presented, and the child was asked: "Which one of these is the same kind of thing?" This procedure continued for all eight sets of pictures.
In the Sticker condition, the experimenter laid the two standards and placed a sticker (a small circle) in the top middle of each standard. Then, the child had to compare the standards: "Do you see why these two have stickers?" Then, the two alternatives were presented and the child was asked: "On which one of these should we also put a sticker?" After the child made a choice, the experimenter continued with the next set. Eight different stickers were used for each of the sets.


Figure 1: A sample of the sets in the categorization task

## Results and Discussion

Mean proportion of relational matches in the category task were measured. Two different analyses were used to measure performance. First, a one-way ANOVA was used to calculate differences between the conditions. The analysis revealed a significant effect of condition, $F(2,40)=4,867, p$ $=.013$. Bonferroni post-hoc tests showed that children in the Sticker condition ( $M_{\text {relational }}=0.6, S D=0.35$ ) made significantly more relational matches compared to the Label $\left(M_{\text {relational }}=0.23, S D=0.36\right)$ and the NoWord $\left(M_{\text {relational }}=\right.$ $0.28, S D=0.34$ ) conditions. The Label Condition was not different from the NoWord condition, $p=1.00$.

In the second analysis, the means of each group was compared to chance (50\%). The comparisons revealed that children in the Label and NoWord condition chose object matches significantly more than chance, $t(14)=-2.981, p=$ .01 and $t(13)=-2.48, p=0.028$, respectively. However, the Sticker Condition was not significantly different from chance, $t(13)=1.076, p=.301$.

As predicted, children who received a generic label performed worse than children who received a nonlinguistic cue during category formation. Further, they showed a strong preference toward the object match, selecting it significantly more than chance. However, contrary to prediction, children who heard a novel label also performed worse than the children who received a nonlinguistic cue, choosing the object match more often and significantly above chance level. Importantly, children in
the Sticker condition chose the relational match more often, though not significantly above chance level. The obtained results provide further insights into the strategies that are available for children to use during category formation tasks.

## General Discussion

Children improve dramatically in their analogical abilities over the preschool and early school years. Various factors contribute to the development of analogical abilities in young children. General experience (Rattermann \& Gentner, 1998), maturation of executive functions (Richland, Morrison, \& Holyoak, 2006; Thibaut, French, \& Vezneva, 2010), and processing capacity (Halford, Wilson, \& Phillips, 1998) all contribute to the development of analogical abilities. However, other mechanisms are also crucial to relational ability and learning in general relational language and comparison (Gentner, 2003; 2010; Alfieri, Nokes-Malach, \& Schunn, 2013).

Comparison is a general learning mechanism that provides efficient means for learning. In particular, in relational learning, comparisons provide children with the opportunity to engage in a process that is akin to relational mapping. This means that children are able to notice and abstract the underlying relational structure between the two standards and thus it becomes more salient and more available for new examples. Children acquire relational terms that support relational representation and reasoning (Christie \& Gentner, 2010; Gentner, 2003; 2010; Gentner et al, 2011). Previous research shows that providing two examples with a common label prompts children to focus their attention to the more subtle structural commonalities (Christie \& Gentner, 2010; Gentner \& Namy, 1999; Namy \& Gentner, 2002).

In this experiment, we sought to replicate data from the Christie and Gentner study (2010) and to further the findings with new conditions. We asked 4 -year-olds to compare two examples of a given category and to extend the category to either an object match or a relational match. The results obtained in this study show that when children received a non-linguistic cue (e.g. a sticker) that represented category membership, they extended the category to a relational match. However, when they received either a specific novel label (e.g. a blicket) or a generic word (e.g. one), they extended the category to an object match.

Concerning the linguistic cues, there are two possible explanations for the obtained results. First, it is possible that providing children with a count noun as a category label focused them to pay more attention to the objects involved (Markman, 1989). As mentioned above, children understand that labels refer to like kinds, but their naïve intuition is to assume that the likeness refers to the objects and not to other commonalities. Children often encounter relational nouns (nouns whose meaning is defined by their relation to other entities) in everyday speech when interacting with adults. However, there is some ambiguity between object construal and relational construal. For example, when children hear a
relational noun (e.g. X is an uncle), they typically focus on perceptual features (e.g. old man) than relational features [(e.g. brother of mother); Gentner, 2003; Christie \& Gentner, 2010].

A second possibility is that children in the Label condition had a difficulty to encode the specific labels used. It is possible that the labels we used are phonologically very different from the majority of words in Bulgarian. If children focused on trying to understand the meaning of an awkward word, their capacity to process the relational information for the two standards was limited. In addition, children in both the Label condition and the NoWord condition heard a word representing the category membership, whereas children in the Sticker condition were able to see the sticker at all times during the categorization task. It is possible that when children hear a word it is more difficult to encode and update the cue, but when a cue is always present and readily available there is no need to store it in working memory and thus encoding it is easier. Although such an assumption seems rather unsupported, having in mind the data from previous research on language labels and comparison among preschool children, it is worth investigating further why our experiment failed to replicate the Label condition in the original study.

Following the main idea of the present study - to explore the various strategies children use during relational categorization, further work is needed to see how comparison supports relational abstraction in preschool children independent from language. It is worth mentioning that the experiment presented here is part of a larger ongoing study. We are currently collecting more data that will provide further insights into the specific roles of different types of presentation and cues in category formation in preschool children.

## Acknowledgements

I am grateful to Stella Christie for providing the stimuli used in this experiment and to Dedre Gentner for the stimulating discussions and insightful suggestions. I thank the children who participated in the study, their families, and teachers. I thank Areta Zumerova for her help with data collection.

## References

Alfieri, L., Nokes-Malach, T. J., \& Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. Educational Psychologist, 48(2), 87-113.
Christie, S., \& Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. Journal of Cognition and Development, 11(3), 356-373.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Sceince, 7, 155-170.
Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child development, 59, 47-59.
Gentner, D. (2003). Why we're so smart? In D. Gentner, \& S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought (pp. 195-235). Cambridge, MA: MIT Press.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Gentner, D., \& Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. Psychological Science, 5(3), 152-158.
Gentner, D., \& Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 4556.

Gentner, D., \& Namy, L. L. (1999). Comparison in the development of categories. Cognitive Development, 14, 487-513.
Gentner, D., \& Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. Cognitive Science, 10, 277-300.
Gentner, D., Anggoro, F. K., \& Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. Child Development, 82(4), 1173-1188.
Graham, S. A., Namy, L. L., Gentner, D., \& Meagher, K. (2010). The role of comparison in preschoolers' novel object categorization. Journal of Experimental Child Psychology, 107, 280-290.
Halford, G. S., Wilson, W. H., \& Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences, 21, 803-865.
Hofstadter, D. R. (2001). Epilogue: Analogy at the core of cognition. In D. Gentner, K. J. Holyoak, \& B. N. Kokinov (Eds.), The analogical mind: Perspectives from conitive science (pp. 499-538). MIT Press.
Holyoak, K. J., Junn, E. N., \& Billman, D. O. (1984). Development of analogical problem-solving skill. Child Development, 55(6), 2042-2055.
Kokinov, B. (1998). Analogy is like cognition: dynamic, emergent, and context-sensitive. In K. Holyoak, D. Gentner, \& B. Kokinov (Eds.), Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences (pp. 96105). Sofia: NBU Press.

Kotovsky, L., \& Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.
Markman, E. M. (1989). Categorization and naming in children: Problems of induction. Cambridge, MA: MIT Press.
Mutafchieva, M., \& Kokinov, B. (2007). Can language be replaced? Physical representations of relations instead of language labels in relational mapping: Do they help young children? In D. S. McNamara, \& J. G. Trafton (Ed.), Proceedings of the 29th Annual Meeting of the Cognitive Science Society (pp. 509-514). Austin, TX: Cognitive Science Society.
Namy, L. L., \& Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in
category learning. Journal of Experimental Psychology: General, 131(1), 5-15.
Rattermann, M. J., \& Gentner, D. (1998). The effect of language on similarity: The use of relational labels improves young children's performance in a mapping task. In K. J. Holyoak, D. Gentner, \& B. Kokinov (Eds.), Advances in analogy research: Integration of theory and data form the cognitive, computational, and neural sciences (pp. 274-282). Sofia: New Bulgarian University.
Richland, L. E., Morrison, R. G., \& Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. Journal of Experimental Child Psychology, 94, 249-273.
Thibaut, J.-P., French, R., \& Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive functions. Journal of Experimental Child Psychology, 106, 1-19.

# When is Likely Unlikely: Investigating the Variability of Vagueness 

Kimele Persaud (kimele.persaud@rutgers.edu)*<br>Brian McMahan (brian.mcmahan@ rutgers.edu)**<br>Malihe Alikhani (ma1195@cs.rutgers.edu) **

Kevin Pei (kevin.pei@rutgers.edu) *
Pernille Hemmer (pernille.hemmer@rutgers.edu)*

Matthew Stone (matthew.stone@rutgers.edu)**<br>Department of Psychology, 152 Frelinghuysen Road, Piscataway, NJ 08854 USA*<br>Department of Computer Science, 96 Frelinghuysen Road, Piscataway, NJ 08854 USA ${ }^{* *}$


#### Abstract

An important part of explaining how people communicate is to understand how people relate language to entities in the world. In describing measurements, people prefer to use qualitative words like 'tall' without precise applicability conditions, also known as vague words. The use of vague language varies widely across contexts, individuals, and tasks (single reference vs. comparisons between targets), but despite this variability, is used quite successfully. A potential strategy for using vague language is to leverage the set of alternative descriptors to settle on the best option. To determine whether people use this strategy, we conducted an experiment where participants picked vague words from sets of alternatives to describe either probability or color values. We varied the set of alternatives from which participants could choose. Empirical evidence supports the hypothesis that people use the set of available options to pick vague descriptors. The theoretical implications of this work are discussed.


Keywords: Vagueness, Alternative sets, Probability, Color

## Introduction

An important aspect of human communication is understanding how people use language to describe the world (Quine, 1969). In the simplest case, people use language to refer to concrete real world objects and categories, such as trees or humans. More interestingly, people use words such as blue, tall, and likely to flexibly refer to indefinite ranges of continuous values across different contexts-a phenomenon known as vagueness. Vague words vary in evoking degrees along different kinds of dimensions (Kennedy, 2007; Kennedy \& McNally, 2010). This paper explores people's expectations for different classes of vague words and how they leverage these expectations to communicate effectively.

## The constrained variability of vague language

Although the use of vague language is ubiquitous in everyday talk, it varies dramatically across communicative situations (Budescu \& Wallsten, 1987). To start, different people relate vague words differently to the values they want to describe (Budescu \& Wallsten, 1987; Wallsten, Budescu, Rapoport, Zwick, \& Forsyth, 1986). For example, in comparable contexts, different people may use the same word to describe different values, or different words to describe the same values.

Nevertheless, meanings in context are not random: vague words seem to always denote bounded, convex regions in the appropriate property space (Gärdenfors, 2004). For example, we can appeal to the convexity of color categories to explain patterns of color naming within and across language communities (Jäger, 2010; Regier, Kay, \& Khetarpal, 2007).

At the same time, individuals' use of vague language varies as a function of the context. People use vague words differently depending on the specific objects they need to distinguish in a situation (Van Deemter, 2006). They also use vague words differently depending on the possible alternative descriptions that would be appropriate (Degen, 2015). For instance, the presence of numbers in the set of available options (e.g.['some', 'all', 'not all', '4']), influences the use of the option 'some' (Degen \& Tanenhaus, 2015). Again, there are limits on such effects. For example, absolute terms, such as 'empty', 'flat' and 'straight', are more constrained in how they vary in context than terms that signal open-ended comparisons, such as 'tall' (e.g. Leffel, Xiang, \& Kennedy, 2016).

Finally, vague language varies as a function of speakers' semantic memory, as revealed by the implicit class to which comparisons are drawn (Lassiter, 2009; Wallsten et al., 1986; Kennedy, 2007). In the quintessential example, the word tall is understood very differently when used to describe a basketball player versus a toddler, and differently again when used to describe a skyscraper versus a glass of water (Schmidt, Goodman, Barner, \& Tenenbaum, 2009). Similar semantic effects are seen in the relationships between different cultures, environments, and their color categories (Regier, Carstensen, \& Kemp, 2016; Stickles, 2014).

Despite this constrained variability, people generally communicate successfully with vague words and prefer to use vague language in many tasks (Van Deemter, 2012).

## Characterizing the variability of vagueness

Bayesian cognitive modeling suggests accounting for these effects in terms of expectations derived from semantic memory, the communicative context and patterns of individual variation (Potts, Lassiter, Levy, \& Frank, 2016). Conversely,
it suggests we can also characterize people's semantic and pragmatic representations through analysis of their interpretation of vague language.

In particular, it's natural to suppose that language users coordinate on specific interpretations in context by assuming that speakers have chosen the most informative description from the available alternatives in light of their expectations. Although several models predict that interpretations of vague language will vary in this way (Potts et al., 2016; Wallsten et al., 1986), empirical evidence supporting this hypothesis is thin.

To investigate this hypothesis, we conducted a forced choice experiment where subjects selected one of a small number of linguistic alternatives to describe a value. Critically, as we varied the number of alternative descriptions, we provided a reject option (i.e. 'none of the above') to measure limits in flexibility of vague terms. Despite a large range of related work, no previous studies have addressed this issue explicitly. We compared performance in the forced choice experiment to a free generate task where participants provided their own words to describe values.

We hypothesize that two possible behaviors may arise. When presented with different sets of alternative descriptions, people may flexibly use the same vague word to refer to different ranges of values. In contrast, people may consistently use vague words to describe the same values regardless of the alternative set of words, instead choosing rejection when preferred options are unavailable.

To foreshadow, we find evidence for both patterns, depending on semantic domain. Probability descriptions vary widely depending on the alternative terms presented to subjects, but color terms vary much less. These results suggest that, although individuals representations of the meaning of some vague words are broadly stable, speakers do adjust boundaries within the available range to give the most information in context. ${ }^{1}$ We outline empirical and theoretical consequences of this finding for future work, emphasizing the need to characterize individual differences and contextual variation jointly, as well as the need to explicitly contrast speaker models based on strategic and heuristic choice.

## Experiment

To assess the role of the set of available terms in constraining vague language, we elicit labeling behavior in a task where participants are shown a property value (i.e. probability and hue), and are asked to either choose a corresponding label from a given set of options ( N -AFC), or freely generate a label that corresponds to the presented value. In the $\mathrm{N}-\mathrm{AFC}$ cases, we expect that the distributions will reflect peoples willingness (or lack of willingness) to "stretch" their category assignment of values based on available terms. In the

[^430]

Figure 1: Sample stimuli for the two tasks: Probability (left two panels) and color (right two panels).
free generate cases, we expect that the distributions will reflect people's natural tendencies of assigning terms to values.

## Methods

Participants Three-hundred and sixty individuals from the Amazon Mechanical Turk research pool participated in this study for monetary compensation.

Materials Color. The stimuli for the color condition consisted of 60 equally spaced values sampled from the winHSV240 (hue, saturation, and value) color space. The colors varied along the full range of the hue dimension, while saturation and value were held constant at $90 \% .^{2}$ The set of available vague color words included seven of the eleven universal color terms (red, orange, yellow, green, blue, purple, and pink; Berlin \& Kay, 1969). To create different conditions, we incrementally increased the number of color words available for participants to choose from - starting with three terms and ending with seven terms. We also included a free generate (see Table 1) condition resulting in six conditions for color in total. In addition to the AFCs for each condition, there was also a reject option, indicating that the color value was not described by any of the available color words. The color space was stratified into six regions, so that each participant only saw one stimuli from each region at equal intervals.This design ensured that participants were presented with values that spanned the entire property range.

Probability. The stimuli for the probability conditions also consisted of 60 equally spaced probability values on the range of $0-1$. The vague probability words that could be used to describe the values were selected from a norming phase with a separate set of participants $(\mathrm{N}=32)$. The norming participants were simply asked to provide labels for randomly generated probability values. The six most frequently generated terms were then used here. To match the structure of the color task, the probability task was also comprised of six condi-

[^431]Table 1: List of Available Vague Words by Condition

| Condition | Probability |  | Color |
| :---: | :---: | :---: | :---: |
| 1 | [UL] unlikely, likely | [3-TERM] | red, green, blue |
| 2 | [ULV] unlikely, likely, very unlikely, very likely | [4-TERM] | red, green, blue, yellow |
| 3 | [ULS] unlikely, likely, somewhat unlikely, somewhat likely | [5-TERM] | red, green, blue, yellow, purple |
| 4 | [VS] very unlikely, very likely, somewhat likely, somewhat unlikely | [6-TERM] | red, green, blue, yellow, purple, orange |
| 5 | [ULVS] unlikely, likely, very unlikely, very likely, somewhat unlikely, somewhat likely | [7-TERM] | red, green, blue, yellow, purple, orange, pink |
| 6 | [FG] free generate | [FG] | free generate |

tions ranging from two alternative forced choices (AFC) to six AFCs, and a free generate condition (see Table 1). The option to reject was present in all conditions. The probability space was stratified into six regions, so that each participant only saw one stimuli from each region at equal intervals.
Procedure Participants were told that they would be helping a robot to understand the meaning of vague words by assigning the words to different property values. Participants were first presented with a set of instructions describing the stimuli and the task. For the probability task, they were shown a pie chart with an arrow called a spinner and a shaded region denoting a probability value (Figure 1, top panels). They were informed that their job was to either pick from a given set of words or generate a word that described the likelihood of the spinner landing in the shaded region of the pie. For the color task, participants were presented with a single color patch and were asked to either choose one of the given color words or generate a word to describe the color Figure 1, bottom panels). Each set of instructions was provided immediately before the task that they described. Each participant described 12 unique property values ( 6 probability values and 6 hue values). The conditions and presentation order of values were randomized across participants.

## Results

Probability Results We assessed whether or not participants consistently used the probability words to describe the same probability values across conditions (see Figures 2 and 3, left column), via linear mixed-effects models (lme) using the lme4 package in R (Bates, Mächler, Bolker, \& Walker, 2015). LME models test for significant differences in responses within experimental groups of primary interest (i.e. condition), while accounting for variability that results from factors that are experimentally uncontrolled (e.g. subjects). We followed up the modelling with planned pair-wise comparisons between conditions using a Tukey post hoc analysis, which corrects for family-wise error rates. In the LME models, subjects and stimuli order were always treated as random effects, while condition was treated as fixed. Probability and hue values were treated as the dependent measure and condition was treated as an indicator variable. For
each probability word, we used a single lme model and compared it to the null. We started with a null model of participants and stimuli order and then added condition as a predictor in the alternate models. The null model predicts no difference in the assignment of probability words to values across conditions and the alternate models predict the opposite. Model fit was assessed using a likelihood-ratio test to compare the hypotheses of the null and alternate models. Condition was significant for probability words:likely ( $\beta=66.33, S E=3.01$ ), unlikely $(\beta=21.03, S E=2.84)$, and somewhat likely $(\beta=47.67, S E=4.72)$. Model comparisons for each of these words favored the alternate models (likely: $\chi^{2}(4)=36.77, p<.0001$, unlikely: $\chi^{2}(4)=26.89, p<$ .0001 , and somewhat likely: $\chi^{2}(4)=9.41, p=.02$ ). Planned pairwise comparisons were conducted to identify the conditions where the probability values differed for each word. For readability, we use codes to refer to the specific conditions (See Table 1 for the condition codes and probability terms available in each condition). Results showed a significant difference in the mean values for likely in the UL and ULV conditions ( $p<.001$ ); ULV and ULS conditions ( $p<0.001$ ); and ULS and VS conditions ( $p=0.02$ ). See Figure 4, left panel, for a visualization of the cumulative changes in values for likely across AFC conditions.

Comparisons show that the probability values assigned to unlikely differed in the ULV and FG conditions ( $p<0.01$ ); UL and ULV conditions ( $p=0.01$ ); and ULS and ULV conditions $(p<0.001)$. See Figure 4 , right panel, for a visualization of the cumulative changes in values for unlikely across AFC conditions. A difference in mean values for somewhat likely was observed in the ULVS and VS conditions ( $p<0.01$ ); and a marginal difference inULVS and FG conditions ( $p=.055$ ). We also calculated the percentage of reject option responses in each N -AFC probability condition. In the order of Table 1, the reject option constituted $4 \%, 1 \%$, $1 \%, 1 \%$, and $1 \%$ of the responses. Taken together, the results suggest that not only are the assignment of probability words to values influenced by the set of alternative descriptions that could have been used, but also that the space of probability values do not have a strict partitioning. In other words, a varying number of probability terms can be flexibly used to


Figure 2: Responses from the N -AFC conditions. Labels on the y -axis were the options available to the participants and the x-axis shows the stimuli values. LEFT PANEL: For each plot, the probability terms selected vs. probability values presented to participants. Right PANEL: For each plot, the color terms selected vs. the hue values presented to participants.
describe different values in probability space.

Color Results The most notable difference between the two domains is that the reject option was selected at a much higher rate for color than probability. In the order of Table 1 , the reject option constituted $36 \%, 31 \%, 14 \%, 9 \%$, and $3 \%$ of the responses for each N-AFC color condition. Like in the probability analysis, we used linear mixed-effects models
to assess whether participants consistently used color words to describe hue values across conditions (see Figures 2 and 3, right column). Interestingly, model comparisons only favored the alternate model for the color purple. Again for readability, we use codes to refer to the specific conditions (see Table 1). Planned pairwise comparisons revealed the conditions where hue values for purple were different: 5-TERM and FG conditions ( $p<.01$ ) and 5-TERM and 7-TERM conditions


Figure 3: Responses from the free generate conditions. The y-axis shows the set of labels freely generated more than 5 times and the $x$-axis show the presented values. Probability responses are presented in the left panel and color responses are presented on the right.


Figure 4: Cumulative frequency curves for likely and unlikely across the AFC conditions. The curves show the relative rates of people using likely (on the left) and unlikely (on the right) probability terms given the alternative sets of probability words in each condition.
( $p<.01$ ). Importantly, there were no significant differences in the mean hue values for the remaining colors across conditions (i.e. red, green, blue, yellow, orange, and pink). The results suggest that, unlike some probability words, the assignment of color words to hue values are inflexible and are not influenced by the set of alternative descriptions that were offered to participants. Instead, participants assignments reflect that their preference for color categories already takes into account an alternative set of other color categories. This is further supported by the high rates of the reject option use in the conditions with fewer options for color descriptions.

## Discussion

In this paper, we investigated how people assign vague words to probability and color values as a function of the set of available alternatives. We measured this behavior in two tasks where participants either selected a vague word from a fixed set or freely generated a word to describe values. Results re-
vealed two interestingly opposing behaviors for probability and color.

For probability, words varied in their assignment to probability values when other vague terms were available. For example, likely was assigned to a different set of probability values in the UL condition where only unlikely was available, relative to the ULV condition where very likely and very unlikely were also available or the ULS condition where somewhat likely was available. In contrast, for color, the assignment of vague words to color values was much more rigid. In fact, purple was the only color that varied across conditions. Other words had relatively well-defined categories that did not overlap.

The results suggest that for probability, people are adopting the strategy of using the set of available terms to constrain variability. This is consistent with the well-known framing effect (Tversky \& Kahneman, 1986) where decisions/preferences change as a function of how options are
presented. For color, however, they are not adopting this strategy.

There are two potential reasons why this may be. First, the set of color words used in the task are already constrained to the basic universal color categories. It is possible that people are more flexible when the vague color word is not drawn from the 11 universal terms. For example, teal, which was a freely generated response might shift in its assignment to color values depending on the available options. If blue is present in the set, teal might be selected for more greenish hues, and if green is in the set, teal might be selected for bluish hues. Alternatively, it could be that probability words encode a relative comparison in a way that color words do not (Leffel et al., 2016), and this semantic difference stabilizes the interpretations of color words in context. In other words, color terms come with an intrinsic range of applicability, not just a prototypical or ideal instance of the term.

A possible limitation of this work is in the finding that the distributions for most probability terms are very broad and overlapping (See Figure 2), which might be due to either individual differences or the context provided by the alternative sets manipulation. The current methodology is insufficient to distinguish between these. One way we could assess this would be to build generative statistical models which simulate the behavior of the participants under the two possible stories and compare the simulations to the empirical data. At the same time, this work lays the groundwork for examining future questions such as: how do you represent the applicability of words like likely in ways that explain their constrained variability? And how do speakers combine their sense of what would be a good description with factors like their expectations about how a description will be interpreted?

## Acknowledgements

This work is based upon a project supported by the National Science Foundation grant 1526723, the National Science Foundation Career grant number 1453276, and partially supported by National Science Foundation Graduate Research Fellowship under Grant Number NSF DGE 0937373.

## References

Bates, D., Mächler, M., Bolker, B., \& Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software.
Berlin, B., \& Kay, P. (1969). Basic color terms: Their university and evolution. California UP.
Budescu, D. V., \& Wallsten, T. S. (1987). Subjective estimation of precise and vague uncertainties. Judgmental forecasting, 63-81.
Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. Semantics and Pragmatics, 1-55.
Degen, J., \& Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. Cognitive Science, 667-710.

Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 998-998.
Gärdenfors, P. (2004). Conceptual spaces: The geometry of thought.
Jäger, G. (2010). Natural color categories are convex sets. In Logic, language and meaning. Springer.
Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. Linguistics and philosophy, 1-45.
Kennedy, C., \& McNally, L. (2010). Color, context, and compositionality. Synthese, 79-98.
Lassiter, D. (2009). Vagueness as probabilistic linguistic knowledge. In International workshop on vagueness in communication.
Leffel, T., Xiang, M., \& Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. In (pp. 836-854).
Persaud, K., \& Hemmer, P. (2016). The dynamics of fidelity over the time course of long-term memory. Cognitive Psychology, 1-21.
Potts, C., Lassiter, D., Levy, R., \& Frank, M. C. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. Journal of Semantics, 755-802.
Quine, W. V. (1969). Word and object. Cambridge, Mass.
Regier, T., Carstensen, A., \& Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. PLOS ONE.
Regier, T., Kay, P., \& Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. Proceedings of the National Academy of Sciences, 1436-1441.
Schmidt, L. A., Goodman, N. D., Barner, D., \& Tenenbaum, J. B. (2009). How tall is tall? compositionality, statistics, and gradable adjectives. In Proceedings of the 31st annual conference of the cognitive science society.
Sims, C. R., Ma, Z., Allred, S. R., Lerch, R. A., \& Flombaum, J. I. (2016). Exploring the cost function in color perception and memory: An information-theoretic model of categorical effects in color matching. Proceedings of the 38th Annual Conference of the Cognitive Science Society.
Stickles, R. T., Elise. (2014). The relation of color naming and the environment. In Cogsci.
Tversky, A., \& Kahneman, D. (1986). Rational choice and the framing of decisions. Journal of Business, 251-278.
Van Deemter, K. (2006). Generating referring expressions that involve gradable properties. Comput. Linguist., 195-222.
Van Deemter, K. (2012). Not exactly: In praise of vagueness. Oxford University Press.
Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., \& Forsyth, B. (1986). Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General, 348.

# Counterfactuals, indicative conditionals, and negation under uncertainty: Are there cross-cultural differences? 

Niki Pfeifer (niki.pfeifer@Imu.de)<br>Munich Center for Mathematical Philosophy, LMU Munich, Germany<br>Hiroshi Yama (yama@lit.osaka-cu.ac.jp)<br>Graduate School of Literature and Human Behavioral Sciences, Osaka City University, Japan


#### Abstract

In this paper we study selected argument forms involving counterfactuals and indicative conditionals under uncertainty. We selected argument forms to explore whether people with an Eastern cultural background reason differently about conditionals compared to Westerners, because of the differences in the location of negations. In a $2 \times 2$ between-participants design, 63 Japanese university students were allocated to four groups, crossing indicative conditionals and counterfactuals, and each presented in two random task orders. The data show close agreement between the responses of Easterners and Westerners. The modal responses provide strong support for the hypothesis that conditional probability is the best predictor for counterfactuals and indicative conditionals. Finally, the grand majority of the responses are probabilistically coherent, which endorses the psychological plausibility of choosing coherence-based probability logic as a rationality framework for psychological reasoning research.


Keywords: argument forms; cross-cultural comparison; counterfactuals; indicative conditionals; negation; probability logic; reasoning under uncertainty

## Introduction

In this paper we study selected argument forms involving counterfactuals and indicative conditionals under uncertainty. The aim is to explore potential cross-cultural differences in human reasoning about conditionals and negation under uncertainty between Easterners and Westerners. There are two possible hypotheses: A universal hypothesis and a cultural differences hypothesis. Like universal grammar (Chomsky, 1957), the human mind is conceived as universal across cultures according to mainstream $20^{\text {th }}$ century psychology. Researchers who agree with this hypothesis usually assume that cultural differences are very small since human reasoning has evolved universally (e.g., Mercier \& Sperber, 2011). So far, cross-cultural differences in reasoning involving negations have been described in the classical-logic based (old) paradigm psychology of reasoning literature (see, e.g., Nisbett, Peng, Choi, \& Norenzayan, 2001; Norenzayan, Smith, Kim, \& Nisbett, 2002; Peng \& Nisbett, 1999; Yama, in press). These previous studies demonstrate that Westerners are inclined to engage in rule-based reasoning whereas Easterners are apt to engage in intuitive or dialectical reasoning. In other words, Easterners are more likely to consider contradictory premises dialectically than Westerners. However, Zhang, Galbraith, Yama, Wang, and Manktelow (2015) report that Easterners are not actually more dialectical when they meet contradictory opinions, but they believe due to cultural reasons that dialectical thinking is wiser than Westerners. Because contradictory premises are not used in this ex-
periment, we do not make predictions concerning whether Easterners reason more dialectical or not (see, e.g., Peng \& Nisbett, 1999). Rather, we explore whether the location of negation in the context of conditionals impacts on reasoning and whether our Japanese sample differs from corresponding data of Western samples. If Japanese people see a stronger cultural value in dialectical thinking, it is plausible to assume that they may hesitate to show stronger confidence in the correctness of their judgments. Moreover, the Japanese language differs from European languages in the location of verb and negation. Usually, the verb is placed at the end of a sentence in Japanese. Furthermore, the term "not" is placed after the negated verb. Thus, the word order of a negated sentence is: complement-verb-not. In spite of these differences, cross-cultural studies on logical reasoning which focus on these differences systematically are rare. Our study presents one of the first attempts (see also Yama, in press) to identify cross-cultural differences within the framework of the new probability-based paradigm psychology of reasoning.

Among the various ways of expressing and using counterfactuals (see, e.g. Declerck \& Reed, 2001), we restrict our investigation of counterfactuals to conditionals in subjunctive mood, where the grammatical structure implies that the counterfactual's antecedent $(A)$ is factually false. For instance, consider the utterance of the following counterfactual in the context of a randomly drawn poker card:

> If the drawn card were to show an ace $(A)$, then it would show spades $(C)$.

The grammatical structure of (1) pragmatically entails that the drawn card is not an ace $(\neg A)$, i.e., the antecedent $A$ of (1) is false. By "indicative conditional" we mean an "if-then" statement of the form If A, then $C$, e.g.,

If the drawn card shows an ace, then it shows spades.
Contrary to the counterfactual (1), the indicative conditional (2) does not imply whether the card actually shows an ace or not. While the core meaning of indicative conditionals was equated with the semantics of the material conditional in the classical logic-based paradigm (or "old") psychology of reasoning (see, e.g., Braine \& O’Brien, 1998; JohnsonLaird, 1983; Rips, 1994; Wason \& Johnson-Laird, 1972), our work is located in the new paradigm psychology of reasoning, where conditionals are interpreted as conditional probability assertions (see, e.g., Elqayam, Bonnefon, \& Over, 2016;

Oaksford \& Chater, 2007; Over, 2009; Pfeifer, 2013). Instead of using (fragments of) classical logic, the new paradigm psychology of reasoning uses probability theory as a rationality framework. Probability as a rationality framework is psychologically and philosophically appealing for many reasons (see, e.g., Pfeifer \& Douven, 2014). Let us mention three of them.

First, probability theory allows for managing degrees of belief instead of restricting belief to the two values true and false as in the case of bivalent classical logic. Thus, probability theory provides a much richer framework to study conditionals. It allows for analysing different psychological predictions concerning conditionals: not only in terms of the material conditional $(A \supset C)$ and the conjunction $(A \wedge C)$ as defined in classical logic, but also in terms of the conditional event $(C \mid A)$, as defined in coherence-based probability logic (see, e.g., Coletti \& Scozzafava, 2002; Gilio, Pfeifer, \& Sanfilippo, 2016; Pfeifer \& Kleiter, 2009). Table 1 presents the truth conditions of these three interpretations. Note that the conditional event cannot be expressed in classical bivalent logic. We hypothesise that the degree of belief in a conditional If $A$, then $C$ is interpreted by a suitable conditional probability assertion $(p(C \mid A))$ and neither as the probability of the material conditional $(p(A \supset C))$ nor as the probability of the conjunction $(p(A \wedge C))$. We will test these three interpretations in the following experiment.

Table 1: Truth tables for the material conditional $A \supset C$ interpretation, the conjunction $\wedge$ interpretation and the conditional event interpretation $C \mid A$ of a (counterfactual) conditional If $A$ (were the case), then $C$ (would be the case).

| $A$ | $C$ | $A \supset C$ | $\wedge$ | $C \mid A$ |
| :---: | :---: | :---: | :---: | :---: |
| true | true | true | true | true |
| true | false | false | false | false |
| false | true | true | false | undetermined |
| false | false | true | false | undetermined |

Second, probability logic blocks so-called paradoxes of the material conditional (see, e.g., Pfeifer, 2014). For example, $\neg A$ ("not-A") logically entails $A \supset C$. The paradox arises, when the material conditional is used to formalize a natural language conditional. Then, for example, the conditional "if it rains today, then I'll be a billionaire tomorrow", follows from the premise "it does not rain today": this inference violates common sense but it is logically valid. In probability logic, the inference from $p(\neg A)=x$ to $p(C \mid A)$ is probabilistically non-informative, i.e., if $p(\neg A)=x$, then $0 \leq p(C \mid A) \leq 1$ is coherent; hence, the paradox is blocked (Pfeifer, 2014). Whether an inference is probabilistically informative or not is a binary question: if the best possible coherent probability bounds on the conclusion coincide with the unit interval $[0,1]$, then the argument form is probabilistically non-informative; otherwise, it is probabilistically formative (i.e., the premise set constrains the probability of the conclusion). The the-
oretical prediction that the paradox is probabilistically noninformative also matches experimental data based on samples involving Westerners (Pfeifer \& Kleiter, 2011; Pfeifer \& Tulkki, 2017b). Note that the paradox is not blocked if the conditional probability (conclusion) is replaced by $p(A \supset C)$ or by $p(A \wedge C)$. A subgoal of this paper is to explore how Japanese participants reason about this paradox.

Third, probability allows for retracting conclusions in the light of new evidence while classical logic is monotonic (i.e., adding a premise to a logically valid argument can only increase the set of conclusions). The suppression effect (see, e.g., Byrne, 1989; Stenning \& van Lambalgen, 2005) illustrates peoples' capacity to retract conclusions if new premises are learned. Moreover, experimental data suggests that most people satisfy basic nonmonotonic reasoning postulates of System P (see, e.g. Benferhat, Bonnefon, \& Da Silva Neves, 2005; Pfeifer \& Kleiter, 2005, 2010). The rules of System P describe formally basic principles any system of nonmonotonic reasoning should satisfy (Kraus, Lehmann, \& Magidor, 1990) and different semantics were developed, including probabilistic ones. Probabilistic semantics postulate that conditionals should be represented by conditional probability assertions (see, e.g., Adams, 1975; Gilio, 2002). Interestingly, inference rules which are (in)valid in System P are also (in)valid in standard systems of counterfactual conditionals (like Lewis, 1973). This convergence shows a close relation between conditional probabilities and counterfactuals. Compared to the big number of psychological investigations on indicative conditionals (for overviews see, e.g., Evans \& Over, 2004; Nickerson, 2015), studies on adult reasoning about counterfactuals are surprisingly rare (Over, Hadjichristidis, Evans, Handley, \& Sloman, 2007; Pfeifer \& Stöckle-Schobel, 2015; Pfeifer \& Tulkki, 2017b). Our study sheds new light by adding a cross-cultural perspective on indicative conditionals and counterfactuals.

Table 2: Task names, their abbreviations and formal structures used in the experiment, where $\neg$ denotes negation, $\rightarrow$ is a placeholder for denoting the indicative conditional or the counterfactual, $\supset$ denotes the material conditional, $\therefore$ denotes "Therefore".

| Task name (abbreviation) | Argument form |
| :--- | :--- |
| Aristotle's thesis \#1 (AT1) | it's not the case that: $(\neg A \rightarrow A)$ |
| Aristotle's thesis \#2 (AT2) | it's not the case that: $(A \rightarrow \neg A)$ |
| Negated Reflexivity (NR) | it's not the case that:( $A \rightarrow A)$ |
| From "Every" to "If" (EIn) | Every $S$ is $P \therefore S \rightarrow \neg P$ |
| From "Every" to "If" (EI) | Every $S$ is $P \therefore S \rightarrow P$ |
| Modus Ponens (MP) | $A, A \rightarrow C \therefore C$ |
| Negated MP (NMP) | $A, A \rightarrow C \therefore \neg C$ |
| Paradox (Prdx) | $\neg A \therefore A \rightarrow C$ |

Table 2 lists the task names, their abbreviations, and their underlying logical form used in our experiment. All argu-
ment forms were investigated previously in the literature on Western samples. Each argument form is suitable for indicative and subjunctive formulations. They are carefully selected to distinguish between the material conditional, conjunction and conditional event interpretation of conditionals. Tasks AT1, AT2, and NR (adapted from Pfeifer, 2012) are about negating conditionals. AT1 and AT2 are contingent (i.e., they are neither tautologies nor contradictions) under the material conditional interpretation of conditionals: specifically, $\neg(\neg A \supset A) \equiv \neg(\neg \neg A \vee A) \equiv \neg A$ and $\neg(A \supset \neg A) \equiv \neg(\neg A \vee$ $\neg A) \equiv \neg \neg A \equiv A$. Since we don't know anything about $(\neg) A$, probability logic predicts for AT1: $0 \leq p(\neg(\neg A \supset A)) \leq 1$; likewise, for AT2: $0 \leq p(\neg(A \supset \neg A)) \leq 1$. For the conditional event interpretation, however, both AT1 and AT2 obtain probability one, since in general coherence requires that $p(A \mid \neg A)=p(\neg \mid A)=0$ for any contingent $A$ and since by the narrow scope reading of conditionals, AT1 is represented by $p(\neg A \mid \neg A)$ and AT2 is represented by $p(A \mid A)$ and 1 is the only coherent assessment for the respective conditional probabilities. Note that there are two ways to negate material conditionals, namely the wide scope negation of material conditionals (i.e., $A \supset C$ can be negated by $\neg(A \supset C)$ ) and the narrow scope negation of material conditionals (i.e., $A \supset C$ is negated by negating its consequent $C: A \supset \neg C$ ). Note that if people interpret $\rightarrow$ by $\supset$ but negate the conditional by the narrow scope interpretation of negation of conditionals, the predictions for AT1 and AT2 coincide with the predictions of the conditional probability interpretation of conditionals (since AT1: $p(\neg A \supset \neg A)=p(\neg \neg A \vee \neg A)=p(A \vee \neg A)=1$ and since AT2: $p(A \supset \neg \neg A)=p(\neg A \vee \neg \neg A)=p(\neg A \vee A)=1)$. To disentangle the conditional probability interpretation and the narrow scope negation of the material conditional interpretation, we added the NR task. The NR task, the narrow scope negation of the material conditional interpretation predicts that the whole unit interval is coherent, since the instruction does not reveal any probabilistic information about $\neg A$ and since $(A \supset \neg A) \equiv(\neg A \vee \neg A) \equiv \neg A$, hence $0 \geq p(A \supset$ $\neg A) \leq 1$, while coherence requires that $p(\neg A \mid A)=0$.

Table 3 lists the normative predictions of the different argument forms. Averaging the percentages of responses in three studies reveals that $73 \%$ of the participants in task AT1, $75 \%$ in task AT2, and $80 \%$ of the participants in task NR responded probabilistically coherently according to the conditional probability interpretation (Pfeifer, 2012; Pfeifer \& Stöckle-Schobel, 2015; Pfeifer \& Tulkki, 2017b).

Task EI (resp., task EIn) connects the basic syllogistic sentence type "Every $S$ is $P$ " with associated conditionals (resp., conditionals involving negations) in the indicative and in the counterfactual form. The motivation for these tasks is to shed light on the hypothesised close relations between quantified statements and conditional probability assertions in the literature (see, e.g. Cohen, 2012; Pfeifer \& Sanfilippo, 2017, submitted). Recent data of Westerners suggest, that in task ASP $73 \%$ of the participants respond that the conclusion holds, whereas $88 \%$ of the participants respond that the conclusion
in task ASnP does not hold (Pfeifer \& Tulkki, 2017b), which corresponds to the normative predictions.

We also investigate the well-known MP and its not logically valid but probabilistically informative counterpart NMP. In a sample of Western participants (Pfeifer \& Tulkki, 2017b), $68 \%$ responded correctly, that the conclusion in task MP holds, and $63 \%$ responded correctly that the conclusion in task NMP does not hold (see also Pfeifer \& Kleiter, 2007).

Although tasks EIn, EI, MP, and NMP do not differentiate among the three considered interpretations of the conditionals, these tasks were selected (i) to test whether the responses of the Japanese sample differs from responses of corresponding Western samples and (ii) to investigate whether there are differences in the responses between the two experimental conditions (i.e., indicative versus counterfactual conditionals).

Finally, as mentioned above, we investigate one of the paradoxes of the material conditional. Western data on Task Prdx indicates that most people ( $87 \%$ on the average) understand that this argument form is probabilistically non-informative (Pfeifer \& Kleiter, 2011; Pfeifer \& Tulkki, 2017b).

## Method

## Materials and Design

We used a $2 \times 2$ between-participants design where we crossed task formulations in terms of indicative conditionals versus formulations in terms of counterfactuals. To control for position effects, we used two random orders (generated by random.org). This resulted in four different task booklets.

Each booklet consisted of a brief introduction, of eight tasks, and of questions about the booklets (task difficulty, whether participants took logic or probability classes and whether they like maths). Furthermore, we included usual demographic questions at the end. The logical forms of the eight tasks are explained in Table 2. We instantiated these logical forms into a cover story which was already used in studies on Western samples (see, e.g., Pfeifer \& Kleiter, 2011; Pfeifer \& Tulkki, 2017b). We adapted and translated this cover story for the Japanese sample.

For each task, the participants were asked to imagine the following situation:

> Hanako works in a factory that produces toy blocks. She is responsible for controlling the production. Every toy block has a shape (cylinder, cube or pyramid) and a colour (red, blue or green). For example:
> - Red cylinder, red cube, red pyramid
> - Blue cylinder, blue cube, ...
> - Green cylinder, ...

Then, for example in task AT1 (indicative conditional), the participants were asked to consider the following sentence:

[^432]（もしおもちゃのブロックが立方体ではないならば，そのおも ちゃのブロックは立方体である，というわけではない。）

The instructions continued by the following questions， which prompt answers in a forced choice format：

> Can Hanako infer at all how sure she can be that the sentence in the box holds? (please tick the appropriate box)
> $\square$ NO, Hanako can not infer how sure she can be that the sentence in the box holds.
> $\square$ YES, Hanako can infer how sure she can be that the sentence in the box holds.

The previous question serves to give the opportunity to re－ spond in a non－informative way and thereby avoid conver－ sational implicatures which could bias the participant to re－ spond in an informative manner．Specifically，we aim to in－ vestigate to what extent the participants are able to distinguish probabilistically informative and non－informative argument forms．The next question prompts a qualitative evaluation of the conclusion of argument forms which are perceived to be probabilistically informative：

If you chose＂YES＂，please tick one of the following answers：
$\square$ Hanako can be sure that the sentence in the box holds．
$\square$ Hanako can be sure that the sentence in the box does not hold．

After each target task，the participants were instructed to rate on a scale their subjective confidence in their response．The corresponding AT1 task involving counterfactuals was formu－ lated in exactly the same way with the difference，that the indicative conditional was replaced by a corresponding coun－ terfactual，as follows：

> | It is not the case, that: If the toy block were not a cube, |
| :--- |
| then the toy block would be a cube. |

（もしおもちゃのブロックが立方体ではなかったとすれば，そのお もちゃのブロックは立方体であるだろう，というわけではない。）

Note that AT1 can be conceived as an inference from an empty premise set．For those tasks involving explicit premises（i．e．，in tasks EIn，EI，MP，NMP，and Prdx），we for－ mulated uncertainties in terms of verbal descriptions（＂極め て確実である＂；＂quite sure＂）．For instance，consider task MP：
（A）．．．quite sure that the toy block is a cube．
（B）．．．quite sure that if the toy block is a cube，then it is red．
Our reason for qualitative premise and conclusion probabil－ ities in terms of verbal descriptions of probabilities（instead of quantitative probabilities）was to reduce the psychological complexity of the probabilistic inference．In this study，we were interested in the interpretation of negations and condi－ tionals but not in the numerical propagation of the probabili－ ties from the premises to the conclusion．

## Participants and procedure

63 Osaka City University undergraduate students participated in this study（mean age $20.02(S D=1.05)$ years， 34 females， 21 males， 8 did not disclose their gender）．Their major sub－ jects included various humanistic fields（ 3 commerce， 5 cul－ ture， 1 geography， 5 history， 4 Japanese， 8 law， 5 linguistics， 1 pedagogy， 2 philosophy， 17 psychology， 2 sociology，and 10 other）．Nobody had ever taken logic classes but two partici－ pants had previously taken some probability classes．At the end of the experiment，participants evaluated the set of tasks as rather difficult（mean $2.76(S D=2.11)$ on a scale rang－ ing from 0 （＂very difficult＂）to 10 （＂very easy＂））． $82.54 \%$ reported that they do not like maths．

All participants were tested at the same time during a les－ son in a course on cultural psychology．For reducing the probability for copy－pasting responses，the booklets were dis－ tributed such that the two task orders and the two formula－ tions of the conditionals（indicative vs．counterfactual）alter－ nated systematically．Moreover，the experimenter announced that the task booklets differ before the participants started with filling in their responses．The booklets were formulated in Japanese，the participants＇mother tongue．

## Results and discussion

We performed Fisher＇s exact tests to compare the response frequencies among the four booklets（task order $1 \times$ task or－ der $2 \times$ indicative conditionals $\times$ counterfactuals）and did not observe any significant differences after performing Holm－ Bonferroni corrections for multiple significance tests．Like－ wise，analyses of variance on the participant＇s confidence ratings in the correctness of their responses did not show statistically significant differences among the four booklets． This replicates previous findings in studies which used West－ ern samples．Specifically，studies on probabilistic truth table tasks（Over et al．，2007；Pfeifer \＆Stöckle－Schobel，2015） and on uncertain argument forms（Pfeifer \＆Tulkki，2017b） did not detect significant difference between indicative con－ ditionals and counterfactuals．Thus，our data speak against cross－cultural differences between Easterners and Western－ ers．This calls for further experiments to clarify whether this interesting negative result is due to a too high dissimilarity of our tasks compared to those in other studies on cross－cultural differences．Or，alternatively，whether cross－cultural differ－ ences are not that strong as they are claimed to be（see，e．g．， Zhang et al．，2015）．

Since there were no significant differences in the responses among the four booklets，we pooled the data for the follow－ ing data analysis $(N=63)$ ．Concerning the interpretation of conditionals，we observed high endorsement rates of the con－ ditional probability hypothesis（see Table 3）．This is strong support for the hypothesis that both indicative conditionals and counterfactuals are best modeled by conditional proba－ bility．

Table 4 presents the mean confidence ratings，which shows how sure the participants are that their responses are correct．

Table 3: Percentages ( $n=63$ ) of "holds" (hld), "does not hold" ( $\neg$ hld), and probabilistic non-informativeness responses ( n -inf; see also Table 2). Predictions based on the conditional probability hypothesis of conditionals are in bold. Alternative hypotheses are indicated in parentheses: $\neg_{\supset}$ (resp., $\supset_{\urcorner}$) denotes wide (resp., narrow) scope negation of the material conditional $\supset ; \wedge$ denotes conjunction. If not specified otherwise, predictions coincide.

|  | AT1 | AT2 | NR | EIn |
| ---: | :--- | :--- | :--- | :--- |
| hld: | $\mathbf{6 5 . 0 8}(\supset \neg)$ | $\left.\mathbf{7 6 . 1 9}\left({ }_{\wedge}\right\urcorner\right)$ | 6.35 | 6.45 |
| $\neg$ hld: | 15.87 | 11.11 | $\mathbf{6 3 . 4 9}(\neg \supset)$ | $\mathbf{6 9 . 3 5}$ |
| n-i: | $19.05(\neg \supset)$ | $12.70(\neg))$ | $30.16\left(\neg_{\wedge}\right)$ | 24.20 |
|  | EI | MP | NMP | Prdx |
| hld: | $\mathbf{8 8 . 8 9}$ | $\mathbf{5 3 . 9 7}$ | 9.52 | $0.00\left({ }^{\circ}\right)$ |
| $\neg$ hld: | 6.35 | 3.17 | $\mathbf{5 2 . 3 8}$ | $17.46(\wedge)$ |
| n-inf: | 4.76 | 42.86 | 38.10 | $\mathbf{8 2 . 5 4}$ |

The confidences are relatively high, with an average value of 7.2 on a rating scale from 0 to 10 .

Table 4: Mean $(M)$ and standard deviations $(S D)$ of the participants' confidence ratings ( $n=63$ ) on a scale from 0 ("very sure that my response is not correct") to 10 ("very sure that my response is correct"; see also Table 2).

|  | AT1 | AT2 | NR | EIn | EI | MP | NMP | Prdx |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $M$ | 6.77 | 6.86 | 7.20 | 7.71 | 8.02 | 7.18 | 7.02 | 6.82 |
| $S D$ | 1.99 | 2.06 | 2.37 | 1.99 | 1.97 | 2.10 | 2.08 | 1.93 |

## Concluding remarks

Our data suggest that people form their degree of belief in the counterfactual If A were the case, $C$ would be the case by equating it with the corresponding conditional probability of $C \mid A$. This is consistent with the observation in previous experimental work (with Western participants) that people treat the factual statement as irrelevant when they form their degree of belief in a counterfactual (Pfeifer \& Stöckle-Schobel, 2015; Pfeifer \& Tulkki, 2017b, 2017a). This can be justified and explained by the coherence-based theory of nested conditionals (Gilio \& Sanfilippo, 2013, 2014; Gilio, Over, Pfeifer, \& Sanfilippo, 2017, submitted). Given three events $A, B, C$ with incompatible $A$ and $B$ (i.e., $A \wedge B$ is a logical contradiction) the prevision of the conditional random quantity $((C \mid B) \mid A)$ is equal to $p(C \mid B)$ (Gilio \& Sanfilippo, 2013, Example 1, p. 225). Thus, the counterfactual If $A$ were the case, $C$ would be the case can be modeled by the degree of belief in the conditional random quantity $(C \mid A) \mid \neg A$ which equals to $p(C \mid A)$ (i.e., Prevision $((C \mid A) \mid \neg A)=p(C \mid A)$ ). This is an explanation for why people-as experimentally demonstrated in Western
samples and also in our Japanese sample-respond by corresponding conditional probabilities when asked to give a degree of belief in a counterfactual.

Our data suggest a negative answer to the question whether there are cross-cultural differences between Easterners and Westerners w. r. t. reasoning about indicative conditionals, counterfactuals, and their negations. Further experimental work, e.g., involving causal task material (see, e.g. Over et al., 2007; Pfeifer \& Tulkki, 2017a), is needed to substantiate the hypothesis that conditional probability is the universal key ingredient for psychological theories of indicative conditionals and counterfactuals.

The material conditional interpretation of conditionals was the gold standard to evaluate human reasoning about conditionals in the old paradigm psychology of reasoning. Our data do not support the material conditional interpretation. Rather, our results strongly support the conditional probability interpretation of conditionals, which became prominent in the new paradigm psychology of reasoning and which received strong experimental support in recent years (see, e.g., Elqayam et al., 2016; Fugard, Pfeifer, Mayerhofer, \& Kleiter, 2011; Over, 2009). Even though most of the data was collected on Western samples, but given the theoretical plausibility of the the conditional probability interpretation, we think that this is further suggests that universality in human reasoning.

Finally, we note that adaptation of reasoning styles can be one of the universal adaptive strategies across cultures. The question of which aspects of human reasoning are universal, and in how far they are universal, is important and calls for collaborations of psychologists of reasoning and cultural psychologists.

Acknowledgments Niki Pfeifer is supported by the DFG project PF 740/2-2 (part of SPP1516).

## References

Adams, E. W. (1975). The logic of conditionals. Dordrecht: Reidel.
Benferhat, S., Bonnefon, J.-F., \& Da Silva Neves, R. (2005). An overview of possibilistic handling of default reasoning, with experimental studies. Synthese, 1-2, 53-70.
Braine, M. D. S., \& O’Brien, D. P. (Eds.). (1998). Mental logic. Mahwah: Erlbaum.
Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. Cognition, 31, 61-83.
Chomsky, N. (1957). Syntactic structures. Hague: Mouton.
Cohen, A. (2012). Generics as modals. Recherches linguistiques de Vincennes, 41, 63-82.
Coletti, G., \& Scozzafava, R. (2002). Probabilistic logic in a coherent setting. Dordrecht: Kluwer.
Declerck, R., \& Reed, S. (2001). Conditionals: A comprehensive empirical analysis. Berlin: Mouton de Gruyter.
Elqayam, S., Bonnefon, J.-F., \& Over, D. E. (Eds.). (2016). New paradigm psychology of reasoning. London: Routledge.

Evans, J. St. B. T., \& Over, D. E. (2004). If. Oxford: OUP. Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., \& Kleiter, G. D. (2011). How people interpret conditionals: Shifts towards the conditional event. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(3), 635-648.
Gilio, A. (2002). Probabilistic reasoning under coherence in System P. Annals of Mathematics and Artificial Intelligence, 34, 5-34.
Gilio, A., Over, D. E., Pfeifer, N., \& Sanfilippo, G. (2017). Centering and compound conditionals under coherence. In M. B. Ferraro et al. (Eds.), Soft methods for data science (pp. 253-260). Berlin: Springer.
Gilio, A., Over, D. E., Pfeifer, N., \& Sanfilippo, G. (submitted). Centering with conjoined and iterated conditionals under coherence. https://arxiv.org/abs/1701.07785.
Gilio, A., Pfeifer, N., \& Sanfilippo, G. (2016). Transitivity in coherence-based probability logic. Journal of Applied Logic, 14, 46-64.
Gilio, A., \& Sanfilippo, G. (2013). Conditional random quantities and iterated conditioning in the setting of coherence. In L. C. van der Gaag (Ed.), Ecsqaru 2013 (Vol. 7958, pp. 218-229). Berlin, Heidelberg: Springer.
Gilio, A., \& Sanfilippo, G. (2014). Conditional random quantities and compounds of conditionals. Studia Logica, 102(4), 709-729.
Johnson-Laird, P. N. (1983). Mental models. Cambr.: CUP.
Kraus, S., Lehmann, D., \& Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence, 44, 167-207.
Lewis, D. (1973). Counterfactuals. Oxford: Blackwell.
Mercier, H., \& Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences, 34, 57-111.
Nickerson, R. S. (2015). Conditional reasoning. New York: Oxford University Press.
Nisbett, R. E., Peng, K., Choi, I., \& Norenzayan, A. (2001). Culture and system of thought: Holistic versus analytic cognition. Psychological Review, 108, 291-310.
Norenzayan, A., Smith, E. E., Kim, B. J., \& Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. Cognitive Science, 26, 653-684.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality. Oxford: Oxford University Press.
Over, D. E. (2009). New paradigm psychology of reasoning. Thinking and Reasoning, 15, 431-438.
Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., \& Sloman, S. (2007). The probability of causal conditionals. Cognitive Psychology, 54, 62-97.
Peng, K., \& Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. Am. Psychol., 54, 741-754.
Pfeifer, N. (2012). Experiments on Aristotle's Thesis: Towards an experimental philosophy of conditionals. The Monist, 95(2), 223-240.
Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective. Thinking \& Reasoning,

19(3-4), 329-345.
Pfeifer, N. (2014). Reasoning about uncertain conditionals. Studia Logica, 102(4), 849-866.
Pfeifer, N., \& Douven, I. (2014). Formal epistemology and the new paradigm psychology of reasoning. The Review of Philosophy and Psychology, 5(2), 199-221.
Pfeifer, N., \& Kleiter, G. D. (2005). Coherence and nonmonotonicity in human reasoning. Synthese, 146, 93-109.
Pfeifer, N., \& Kleiter, G. D. (2007). Human reasoning with imprecise probabilities: Modus ponens and Denying the antecedent. In G. De Cooman, J. Vejnarová, \& M. Zaffalon (Eds.), Proc. of ISIPTA (pp. 347-356). Prague: SIPTA.
Pfeifer, N., \& Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. Journal of Applied Logic, 7(2), 206-217.
Pfeifer, N., \& Kleiter, G. D. (2010). The conditional in mental probability logic. In M. Oaksford \& N. Chater (Eds.), Cognition and conditionals (pp. 153-173). Oxford: OUP.
Pfeifer, N., \& Kleiter, G. D. (2011). Uncertain deductive reasoning. In K. Manktelow, D. E. Over, \& S. Elqayam (Eds.), The science of reason (p. 145-166). Hove: Psychol. Press.
Pfeifer, N., \& Sanfilippo, G. (2017). Square of opposition under coherence. In M. B. Ferraro et al. (Eds.), SMPS (pp. 407-414). Berlin: Springer.
Pfeifer, N., \& Sanfilippo, G. (submitted). Probabilistic squares and hexagons of opposition under coherence. https://arxiv.org/abs/1701.07306.
Pfeifer, N., \& Stöckle-Schobel, R. (2015). Uncertain conditionals and counterfactuals in (non-)causal settings. In G. Arienti, B. G. Bara, \& S. G. (Eds.), Proceedings of the EuroAsianPacific joint conference on cognitive science (Vol. 1419, pp. 651-656). Aachen: CEUR. Retrieved from http://ceur-ws.org/Vol-1419/paper0108.pdf
Pfeifer, N., \& Tulkki, L. (2017a). Abductive, causal, and counterfactual conditionals under incomplete probabilistic knowledge. In G. Gunzelmann, A. Howes, T. Tenbrink, \& E. Davelaar (Eds.), Proceedings of CogScil7. Austin: The Cognitive Science Society.
Pfeifer, N., \& Tulkki, L. (2017b). Conditionals, counterfactuals, and rational reasoning. An experimental study on basic principles. Minds and Machines, 27(1), 119-165.
Rips, L. J. (1994). The psychology of proof: Deductive reasoning in human thinking. Cambridge: MIT Press.
Stenning, K., \& van Lambalgen, M. (2005). Semantic interpretation as computation in nonmonotonic logic. Cognitive Science, 29, 919-960.
Wason, P. C., \& Johnson-Laird, P. N. (1972). The psychology of reasoning. Cambridge: Harvard University Press.
Yama, H. (in press). Thinking and reasoning across cultures. In L. J. Ball \& V. A. Thompson (Eds.), International handbook of thinking and reasoning. Hove: Psychology Press.
Zhang, B., Galbraith, N., Yama, H., Wang, L., \& Manktelow, K. I. (2015). Dialectical thinking: A cross-cultural study of Japanese, Chinese, and British students. Journal of Cognitive Psychology, 27, 771-779.

# Abductive, Causal, and Counterfactual Conditionals Under Incomplete Probabilistic Knowledge 

Niki Pfeifer (niki.pfeifer@lmu.de)<br>Munich Center for Mathematical Philosophy, LMU Munich, Germany<br>Leena Tulkki (leena.tulkki@helsinki.fi)<br>Department of Philosophy, History, Culture and Art Studies, University of Helsinki, Finland


#### Abstract

We study abductive, causal, and non-causal conditionals in indicative and counterfactual formulations using probabilistic truth table tasks under incomplete probabilistic knowledge ( $N=80$ ). We frame the task as a probability-logical inference problem. The most frequently observed response type across all conditions was a class of conditional event interpretations of conditionals; it was followed by conjunction interpretations. An interesting minority of participants neglected some of the relevant imprecision involved in the premises when inferring lower or upper probability bounds on the target conditional/counterfactual ("halfway responses"). We discuss the results in the light of coherence-based probability logic and the new paradigm psychology of reasoning.


Keywords: abductive conditionals; causal conditionals; counterfactuals; indicative conditionals; psychological experiment; uncertain argument form; probabilistic truth table task

## Introduction

The probabilistic truth table task was introduced by two independent studies at the beginning of this millennium (Evans, Handley, \& Over, 2003; Oberauer \& Wilhelm, 2003). It serves to investigate how people interpret conditionals under uncertainty. Moreover, it is one of the starting points of the new (probability-based) paradigm psychology of reasoning (see, e.g. Baratgin, Over, \& Politzer, 2014; Elqayam, Bonnefon, \& Over, 2016; Oaksford \& Chater, 2007; Over, 2009; Pfeifer, 2013; Pfeifer \& Douven, 2014), which started to replace the old (classical logic-based) paradigm psychology of reasoning. The probabilistic truth table task was constructed to investigate how people interpret conditionals (i.e., indicative sentences of the form If $A$, then $C$ ). As its name suggests, the task consists of inferring the degree of belief in a conditional based on probabilistic information attached to the truth table cases. What are truth table cases? Let $A$ and $C$ denote two propositions (i.e., sentences for which it makes sense to assign the truth values true or false) like a fair die is rolled and the side of the die shows an even number, respectively. The four truth table cases induced by $A$ and $C$ are: $A \wedge C$, $A \wedge \neg C, \neg A \wedge C$, and $\neg A \wedge \neg C$, where $\wedge$ denotes conjunction ("and") and $\neg$ denotes negation. Classical logic is bivalent, involving only the truth values true and false. Therefore, the conditional defined in classical logic (i.e., the material conditional, see Table 1) is either true or false. The conditional event $C \mid A$ involved in conditional probability ( $p(C \mid A)$ ), however, is not bivalent, as it is void (or undetermined) if its antecedent $A$ is false (see Table 1). Therefore, it cannot be represented by the means of classical logic.

Table 1 presents the truth conditions of the most important psychological interpretations for adult reasoning about indicative conditionals (i.e., conditional event, conjunction, and material conditional). Moreover, it presents the biconditional and biconditional event interpretations, which were reported in developmental psychological studies (see, e.g. Barrouillet, Gauffroy, \& Lecas, 2015).

Psychological evidence for the conditional event interpretation was already observed within the old paradigm psychology of reasoning (see, e.g., Wason \& Johnson-Laird, 1972). The response pattern, which is consistent with the conditional event interpretation was seen as irrational (dubbed "defective truth table"), as it violates the semantics of the material conditional. The material conditional used to be the gold standard of reference for the meaning of indicative conditionals in the old paradigm. However, within the new paradigm psychology of reasoning this response pattern is, of course, perfectly rational (Over \& Baratgin, 2017; Pfeifer \& Tulkki, 2017).

Table 1: Truth tables for the material conditional $A \supset C$, the conjunction $A \wedge C$, the biconditional $A \equiv C$, the biconditional event $C|\mid A$ (i.e., $A \wedge C| A \vee C$ ), and the conditional event $C \mid A$.

| $A$ | $C$ | $A \supset C$ | $A \wedge C$ | $A \equiv C$ | $C \\| A$ | $C \mid A$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| true | true | true | true | true | true | true |
| true | false | false | false | false | false | false |
| false | true | true | false | false | false | void |
| false | false | true | false | true | void | void |

While classical truth table tasks require to respond with truth values, probabilistic truth table tasks require to with respond degrees of belief. Following Pfeifer (2013), we interpret the task as a probability logical inference problem. Specifically, it is formalised as a probability logical argument with assigned degrees of belief in the four truth table cases as its premises and the degree of belief in a conditional as its conclusion. The inference problem consists in propagating the uncertainties of the premises to the conclusion. As an example, consider the conditional probability interpretation of the $(p(C \mid A))$ conditional If $A$, then $C$ as the conclusion. This argument scheme is formalised by:
(A) From $p(A \wedge C)=x_{1}, p(A \wedge \neg C)=x_{2}, p(\neg A \wedge C)=x_{3}$, and $p(\neg A \wedge \neg C)=x_{4}$, infer $p(C \mid A)=x_{1} /\left(x_{1}+x_{2}\right)$.

In argument scheme $(\mathcal{A})$, the fraction $x_{1} /\left(x_{1}+x_{2}\right)$ is the prob-
ability propagation rule for the conditional probability. For different interpretations of the conditional, the probability propagation rules differ. Table 2 presents the corresponding probability propagation rules of the interpretations given in Table 1.

Table 2: Probability propagation rules for the different interpretations of If $A$, then $C$. The premise set $\{p(A \wedge C)=$ $\left.x_{1}, p(A \wedge \neg C)=x_{2}, p(\neg A \wedge C)=x_{3}, p(\neg A \wedge \neg C)=x_{4}\right\}$ entails the respective conclusion (see also Table 1).

| Interpretation | Conclusion |
| :--- | :--- |
| Material conditional | $p(A \supset C)=x_{1}+x_{3}+x_{4}$ |
| Conjunction | $p(A \wedge C)=x_{1}$ |
| Biconditional | $p(A \equiv C)=x_{1}+x_{4}$ |
| Biconditional event | $p\left(C\|\mid A)=x_{1} /\left(x_{1}+x_{2}+x_{3}\right)\right.$ |
| Conditional event | $p(C \mid A)=x_{1} /\left(x_{1}+x_{2}\right)$ |

The main empirical result of classical probabilistic truth table tasks is that the dominant responses are consistent with the conditional event interpretation of conditionals. Moreover, if people solve the task several times, some people shift to the conditional event interpretation during the course of the experiment (see, e.g., Fugard, Pfeifer, Mayerhofer, \& Kleiter, 2011a; Pfeifer \& Stöckle-Schobel, 2015).
A key feature of classical probabilistic truth table tasks is that they present complete probabilistic knowledge w.r.t. the truth table cases: all (point-valued) probabilities $x_{1}, \ldots x_{4}$ are available to the participant (see, e.g., Evans et al., 2003; Oberauer \& Wilhelm, 2003; Fugard et al., 2011a; Pfeifer \& Stöckle-Schobel, 2015). When all probabilities involved in argument scheme $(\mathcal{A})$ are given as point values, for example, it is possible to infer a precise (point-valued) probability of $C \mid A$. Of course, $x_{3}$ and $x_{4}$ are irrelevant for calculating $p(C \mid A)$ in this case. However, as full probabilistic information is usually not available in everyday life, we argue for investigating incomplete probabilistic knowledge. If $x_{1}$ in argument scheme $(\mathcal{A})$ is only available as an imprecise (i.e., interval-valued) probability, i.e., $x_{1}^{\prime} \leq p(A \wedge C) \leq x_{1}^{\prime \prime}$, the probability of the conclusion of $(\mathcal{A})$ is also imprecise, i.e., $x_{1}^{\prime} /\left(x_{1}^{\prime}+x_{2}\right) \leq p(C \mid A) \leq x_{1}^{\prime \prime} /\left(x_{1}^{\prime \prime}+x_{2}\right)$. Table 4 (see below) presents a numerical illustration of the different interpretations of conditionals in the probabilistic truth table task with imprecise premises. Incomplete probabilistic knowledge has not been investigated within the probabilistic truth table task paradigm yet (for an exception, where only indicative conditionals were investigated, see Pfeifer, 2013).

With only a few exceptions (i.e., Over, Hadjichristidis, Evans, Handley, \& Sloman, 2007; Pfeifer \& Stöckle-Schobel, 2015), the important classes of causal conditionals and counterfactuals have not been investigated empirically within the probabilistic truth table task paradigm yet. Causal conditionals are characterized by connecting cause (i.e., the conditional's antecedent) and effect (i.e., the conditional's consequent), like If you take aspirin, your headache will disappear.

Counterfactuals are conditionals in subjunctive mood, where the grammatical structure signals that the antecedent is false. For instance, If you were to take aspirin (A), your headache would disappear $(D)$, which signals that you had not taken aspirin yet $(\neg A)$. This example is a counterfactual version of the above described causal conditional. Of course, there are also non-causal versions of counterfactuals, like If the side of a card were to show an ace, it would show spades.

Traditionally, counterfactuals are interpreted by possible world semantics (most prominently by Lewis and Stalnaker). However, we interpret counterfactuals in terms of coherence based probability logic. In a nutshell, the aforementioned example of a (causal) counterfactual can by interpreted by a (right-)nested conditional, where the antecedent represents the factual statement $\neg A$ and the consequent represents the conditional $D \mid A$. This representation amounts to $(D \mid A) \mid \neg A$, which is a conditional random quantity (because of the pagelimit we refer for the technical details to Gilio \& Sanfilippo, 2013, 2014; Gilio, Over, Pfeifer, \& Sanfilippo, 2017, submitted). It can be proved that the degree of belief in the conditional random quantity $(D \mid A) \mid \neg A$ is also equal to $p(D \mid A)$ (i.e., Prevision $((D \mid A) \mid \neg A)=p(D \mid A)$; see Gilio \& Sanfilippo, 2013, Example 1, p. 225). Therefore, we hypothesize that the participants' degrees of belief in counterfactuals are equal to corresponding conditional probabilities.

Note that the conditional random quantity $(C \mid A) \mid \neg A$ does not coincide with $C \mid(A \wedge \neg A)$, because the Import-Export Principle does not hold (Gilio \& Sanfilippo, 2014). Therefore, as shown by Gilio and Sanfilippo (2014), the counterintuitive consequences of the well-known triviality results (Lewis, 1976) are avoided. By the way, the formula $C \mid(A \wedge$ $\neg A$ ) is unintelligible (in terms of the Ramsey test, you cannot add a contradiction to your stock of beliefs).

To our knowledge, abductive conditionals have not been empirically investigated in the probabilistic truth table task paradigm yet. Abductive conditionals can be conceived as reversed causal conditionals, characterized as follows: the effect is located in the conditional's antecedent and the cause is located in the conditional's consequent. For example, If your headache disappeared, then you took aspirin. Abductive inferences are also known as inferences to the best explanation (for philosophical and psychological overviews on abduction see, e.g., Douven, 2016a; Lombrozo, 2012, respectively). Like indicative and causal conditionals, abductive conditionals can be formulated in indicative and in subjunctive mood.

The aim of the present study is to help to fill the above mentioned research gaps. Specifically, we aim to shed light on the following questions using probabilistic truth table tasks under incomplete probabilistic knowledge: Are there reasoning strategies for inferring lower and upper bounds in the context of incomplete probabilistic knowledge? Is the conditional event interpretation dominant for abductive, causal, and non-causal counterfactuals?


#### Abstract

Method Materials and Design The task material consisted of 18 pen and paper tasks, preceded by 4 examples explaining the answer format. The task sequence consisted of 9 different tasks that were presented twice in the same random order (i.e., task T10 is a repetition of task T1), resulting in the total of 18 tasks. The tasks were designed to test how participants infer about uncertain conditional sentences in four different experimental conditions (see Table 3). All conditions had the same task sequence, with the following variations: For the first two conditions we used a non-causal scenario in both indicative and counterfactual moods. For the other two conditions we used two variations of a causal scenario in counterfactual mood; inference from causes to effects (causal) and inference from effects to causes (abductive). The material was adapted from probabilistic truth table tasks, which involved precise premises (Fugard et al., 2011a; Pfeifer \& Stöckle-Schobel, 2015).

Table 3: Between participant conditions $\mathrm{C} 1-\mathrm{C} 4$ defined by the types and formulations of conditionals, and sample sizes. |  | Type | Formulation | Sample |
| :--- | :--- | :--- | :---: |
| C1 | non-causal | indicative | $\left(n_{1}=20\right)$ |
| C2 | non-causal | counterfactual | $\left(n_{2}=20\right)$ |
| C3 | causal | counterfactual | $\left(n_{3}=20\right)$ |
| C4 | abductive | counterfactual | $\left(n_{4}=20\right)$ |


For the non-causal scenario we used a vignette story about a six-sided die. The story describes that the die was randomly thrown so that the participants did not know which of the sides ended up facing upwards. The sides of the die were illustrated as six squares. Each side had an image of a black or white geometric figure. In tasks T1, T2, T10, and T11 all sides of the die were shown. To introduce incomplete probabilistic knowledge we presented "covered" sides in the rest of the tasks. Covered sides were indicated by a question mark. Here is an example of how we presented the six sides of a die (task T3/T12):


Next, the participants were presented with the question "How sure can you be that the following sentence holds?" (Kuinka varma voit olla siitä, että seuraava lause pitää paikkaansa?). The target sentences were highlighted with a frame to make the scope of the question clear, for example:

[^433]The answer format had two sets of tick boxes in a " $x$ out of $y$ " format for responding interval-valued degrees of belief. The two response boxes were labeled accordingly ("at least" and "at most"); for instance, as follows:


It was explained in the introduction to give point valued responses by marking the same numbers in both response boxes (i.e., lower and upper bounds coincide).

The target sentence in the non-causal tasks was formulated in terms of "If $A$, then $C$ ". In all non-causal tasks the antecedent mentioned a form and the consequent mentioned a color. After completing each task, the participants were asked to rate their confidence in the correctness of their response on a 10 -step rating scale from "fully confident that your answer is incorrect" to "fully confident that your answer is correct".
Apart from the following two differences, the counterfactual task version was identical to the indicative version of the task: (i) we added a factual statement which contradicted the antecedent of the target sentence and (ii) the target sentence was formulated in subjunctive mood. "The form of an upward-facing side of the die is a cube" is an example of a factual statement and the corresponding target sentence is: "if the figure on the upward facing side of the die were a circle, then the figure would be black" (Jos ylöspäin osoittavan kyljen kuvio olisi ympyrä, niin tämä kuvio olisi musta). The suffix -isi in the Finnish original indicates the counterfactual mood.

For the causal and abductive conditions, the tasks were structurally identical to the (non-causal) dice-scenario. However, instead of dice, a vignette story about drugs and their effects created a causal scenario. In the vignette story, six patient reports were shown to the participants. The patient reports were illustrated as six rectangles having a name of a fictional drug and a result of the medication (either "diminishes symptoms" or "no impact on the symptoms"). We used question marks on some patient reports (like in the dice scenario) to introduce incomplete probabilistic knowledge. Here is an example of the patient reports, which contains the same probabilistic information as the above mentioned die-example:


In the causal version of the task material the antecedent denotes the name of a drug and the consequent denotes an effect. In the abductive version the order was reversed: first an effect was presented and then a drug was named. In this way the tasks called for either causal inferences from causes to effects, or abductive inferences from effects to causes. As the material was formulated in counterfactual mood, we added a factual statement to each task, which contradicted the antecedent of the target sentence.

Participants and Procedure Eighty students from the University of Helsinki (Finland) participated in the experiment.

The students were native Finnish speakers with no previous academic training in logic or probability. Each participant was tested individually. The paper and pencil tasks were followed by a short structured interview about how the participants had interpreted the target tasks. Participants were paid $15 €$ for their participation.

## Results and Discussion

We performed Fisher's exact tests to investigate whether the four different versions of the task booklets had an impact on the participants' degrees of belief in the respective target sentences. After p -value corrections for multiple significance tests, we did not observe significant differences between the four conditions and we therefore pooled the data.

Table 5 shows the percentages of responses according to the different interpretations of conditionals. All tasks differentiate between the conditional probability, the conjunction, and the material conditional interpretation. A subset of the tasks differentiates between biconditional and biconditional event interpretations as well. Conditional probability interpretations marked with indices, however, were patterns identified from the data and were not anticipated during the construction of the task material. Therefore, not all tasks differentiate among all interpretations. Table 4 shows the normative answers for each interpretation for the example task discussed in the previous section (see (Die sides)). Both, lower and upper bound responses, had to match the normative lower and upper bounds for the categorization of the responses in Table 5. Since each response box enables 42 different " $X$ out of $Y$ " responses, and since both, lower and upper bound responses needed to match for the classification, the a priori chance for guessing an interpretation was very low (i.e., $\left.1 /\left(42^{2}\right)=0.0006\right)$.


Figure 1: Mean confidence values for tasks T1-T18 by condition. C1-C4 denote the four condition as defined in Table 3.

The task material was designed so that the normative predictions of the three main psychological interpretations of conditionals (i.e., conditional probability, conjunction, and

Table 4: Example of predicted responses where the task consists in inferring the degree of belief in the conditional "If the figure on the upward facing side of the die is a circle, then the figure is black" (i.e., the conclusion), based on the die presented in (Die sides) above (i.e., (Die sides) contains the premises). The index $\bar{l}$ (resp., $\bar{u}$ ) denotes conditional probability responses where the covered sides are ignored for inferring the lower (resp., upper) bound response. These response types are the "halfway lower" and "halfway upper" interpretations, respectively. $\overline{l u}$ denotes conditional probability responses where covered sides are ignored for inferring both bound responses, i.e., the "halfway both interpretation". See also Table 2.

| Interpretation | Predictions |  |
| :--- | :---: | :---: |
|  | at least | at most |
| $p($ black $\mid$ circle $)$ | 1 out of 2 | 2 out of 2 |
| $p(\text { black } \mid \text { circle })_{\bar{l}}$ | 1 out of 1 | 2 out of 2 |
| $p(\text { black \| circle })_{\bar{u}}$ | 1 out of 2 | 1 out of 1 |
| $p(\text { black } \mid \text { circle })_{\bar{u}}$ | 1 out of 1 | 1 out of 1 |
| $p($ circle $\wedge$ black $)$ | 1 out of 6 | 2 out of 6 |
| $p($ circle $\supset$ black $)$ | 5 out of 6 | 6 out of 6 |
| $p($ circle $\equiv$ black $)$ | 3 out of 6 | 4 out of 6 |
| $p($ circle $\\|$ black $)$ | 1 out of 4 | 2 out of 4 |

material conditional) were unique for each task. During the analysis we identified three further response strategies related to the conditional probability interpretation. In what we call halfway lower interpretation (denoted by $p(\cdot \mid \cdot)_{\bar{l}}$ ) the upper bound is the same as in conditional probability, but the lower bound response differs in that the covered sides (i.e., sides marked with question mark) are ignored. Halfway upper interpretation (denoted by $p(\cdot \mid \cdot)_{\bar{u}}$ ) is the same, but in reverse order. In a halfway both interpretation the covered sides are ignored for both bound responses. As these answer strategies are in a sense partial versions of the conditional probability interpretation, we combined their results with conditional probability answers into grouped conditional probability. Notice that the tasks T1,T2, T10 and T11 with full information (i.e., no question marks) have the same value for lower and upper bound answers. Therefore the halfway responses could not be distinguished from the conditional probability answers in these tasks.

Of all 1440 responses, $32.1 \%$ were consistent with standard conditional event responses, $29.9 \%$ were consistent with conjunction responses, and $0.2 \%$ were consistent with material conditional responses. Like the material conditional, also the biconditional and the biconditional event response frequencies play a neglectable rôle in the data $(0 \%-3 \%$ in the four tasks T3, T6, T12, T15 where these interpretations were differentiated). The predominant response strategy in pointvalued tasks (T1, T2, T10 and T11) was consistent with the conditional probability interpretation. In nine out of 14 tasks with incomplete probabilistic information (i.e., tasks involv-

Table 5: Percentages of responses from all four conditions and all 18 tasks (T1-T18; $N=80$ ). "Grouped $p(\cdot \mid \cdot)$ " denotes the sum of all conditional probability responses, including those marked with the indices. The halfway interpretations (indices $\bar{l}$ and $\bar{u}$ ) and the numerical predictions are explained in Table 4. "- -" denotes cases where different conditional probability interpretations cannot be individuated (i.e., in the point value tasks). Similarly, "[- -]" denotes cases where biconditional and biconditional event interpretations cannot be distinguished from the other interpretations. The interpretations are explained in Table 2. "** denotes psychological main interpretations.

| Interpretation | T1 | T2 | T3 | T4 | T5 | T6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $[p(\cdot \mid \cdot)]^{\star}$ | [46] | [55] | [15] | [19] | [24] | [24] |
| $\left[p(\cdot \mid \cdot)_{\bar{l}}\right]$ | [--] | [--] | [5] | [13] | [18] | [11] |
| $\left[p(\cdot \mid \cdot)_{\bar{u}}\right]$ | [--] | [--] | [23] | [10] | [13] | [11] |
| $\left[p(\cdot \mid \cdot)_{\overline{l u}}\right]$ | [--] | [--] | [0] | [3] | [1] | [0] |
| Grouped $p(\cdot \mid \cdot)$ | 46 | 55 | 43 | 44 | 55 | 46 |
| $p(\cdot \wedge \cdot)^{\star}$ | 29 | 28 | 34 | 39 | 34 | 31 |
| $p(\cdot \supset \cdot)^{\star}$ | 1 | 0 | 0 | 0 | 0 | 1 |
| $p(\cdot \equiv \cdot)$ | [--] | [--] | 1 | [--] | [--] | 0 |
| $p(\cdot \\| \cdot)$ | [--] | [--] | 3 | [--] | [--] | 0 |
| Other | 24 | 18 | 20 | 18 | 11 | 21 |
|  | T7 | T8 | T9 | T10 | T11 | T12 |
| $\left[p(\cdot \mid \cdot){ }^{\text { }}\right.$ | [24] | [28] | [26] | [44] | [55] | [25] |
| $\left[p(\cdot \mid \cdot)_{\bar{\prime}}\right]$ | [10] | [15] | [10] | [--] | [--] | [9] |
| $\left[p(\cdot \mid \cdot)_{\bar{u}}\right]$ | [16] | [8] | [10] | [--] | [--] | [23] |
| $\left[p(\cdot \mid \cdot)_{\overline{l u}}\right]$ | [0] | [0] | [0] | [--] | [--] | [0] |
| Grouped $p(\cdot \mid \cdot)$ | 46 | 55 | 43 | 44 | 55 | 46 |
| $p(\cdot \wedge \cdot)^{\star}$ | 34 | 29 | 33 | 26 | 29 | 30 |
| $p(\cdot \supset \cdot)^{\star}$ | 0 | 0 | 0 | 1 | 0 | 0 |
| $p(\cdot \equiv \cdot)$ | [--] | [--] | [--] | [--] | [--] | 0 |
| $p(\cdot \\| \cdot)$ | [--] | [--] | [--] | [--] | [--] | 0 |
| Other | 16 | 21 | 21 | 18 | 18 | 14 |
|  | T13 | T14 | T15 | T16 | T17 | T18 |
| $[p(\cdot \mid \cdot)]^{\star}$ | [34] | [33] | [29] | [26] | [31] | [31] |
| $\left[p(\cdot \mid \cdot)_{\bar{l}}\right]$ | [9] | [13] | [11] | [10] | [18] | [13] |
| $\left[p(\cdot \mid \cdot)_{\bar{u}}\right]$ | [11] | [10] | [11] | [15] | [8] | [11] |
| $\left[p(\cdot \mid \cdot)_{\overline{l u}}\right]$ | [0] | [0] | [1] | [3] | [0] | [0] |
| Grouped $p(\cdot \mid \cdot)$ | 46 | 55 | 43 | 44 | 55 | 46 |
| $p(\cdot \wedge \cdot)^{\star}$ | 28 | 30 | 26 | 29 | 25 | 28 |
| $p(\cdot \supset \cdot)^{\star}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $p(\cdot \equiv \cdot)$ | [--] | [--] | 0 | [--] | [--] | [--] |
| $p(\cdot \\| \cdot)$ | [--] | [--] | 3 | [--] | [--] | [--] |
| Other | 19 | 15 | 19 | 18 | 19 | 18 |

ing "covered" sides or patient reports) the predominant answer strategy was conjunction. We also observed shifts of interpretation towards conditional probability: comparing the first three tasks with incomplete information (i.e., T3-T5) to the last three (i.e., T16-T18), the number of conditional probability answers increased from $19 \%$ to $30 \%$, and conjunction
answers decreased from $35 \%$ to $27 \%$. This replicates shifts of interpretations reported in the literature (Fugard et al., 2011a; Pfeifer, 2013).

However, when all the conditional probability response types are grouped together, the resulting set of response strategies is clearly the predominant one in all tasks. $51.5 \%$ of all answers are consistent with the grouped conditional probability responses. $18.1 \%$ were "other" responses, that is, responses that did not fit the grouped conditional probability, conjunction, biconditional, biconditional event, or material conditional. Thus, in total $81.9 \%$ of the data can be modeled by the investigated hypotheses concerning the interpretation of conditionals.

Compared to a previous study that investigated non-causal indicative conditionals under incomplete probabilistic information (Pfeifer, 2013, i.e., similar tasks as in condition C1), our results show lower level of conditional event responses (compared to the previous $65.6 \%$ ), and higher levels of conjunction responses (compared to the previous $5.6 \%$ ). The material conditional responses were similar (previously $0.3 \%$ ). Pfeifer and Stöckle-Schobel (2015) investigated conditionals under complete probabilistic knowledge and used similar tasks as in our conditions C1, C2 and C3. These authors also reported higher levels of conditional probability answers and lower levels of conjunction responses, while material conditional responses were similarly low. The lower levels of conditional probability responses may be explained by the apparent higher complexity of the tasks used in the present experiment. The tasks are more complex (i) because of the combination of using counterfactuals as target sentences in three of four conditions and (ii) because of imprecise probabilities in the premises (i.e., incomplete probabilistic information).

The tendency to give answers that partially coincide with conditional probability has also been found in a previous study which tested non-causal cases in indicative mood with similar task material as we used for condition C1 (Pfeifer, 2013). In that study our halfway lower-interpretation is referred to as "halfway conditional event strategy". However, in the present study we found two completely new strategies: the halfway upper- and the halfway both-interpretation. The halfway upper-interpretation is particularly interesting, as it explains $12.8 \%$ of the total 1120 responses, slightly more than the halfway lower response strategy (i.e., $11.6 \%$ ). Halfway conditional probability responses might unburden the working memory load by ignoring the covered sides (see also Pfeifer, 2013).

Figure 1 shows the results of the confidence ratings. We performed analyses of variance to investigate impacts of the different conditions. After Holm-Bonferroni corrections we observed significant differences within the three tasks T1, T2 and T 10 . The corresponding p -values were $0.006,0.01$, and 0.008 . In each of these tasks-as well as in all other tasksthe condition C 4 had lower mean confidence values compared to the other conditions. The lower confidence may be because of the apparent higher difficulty of the task material in
condition C 4 for two reasons: first, the target sentence was a counterfactual. Because of the inconsistency between the factual statement and the antecedent, many participants reported counterfactuals as puzzling in the post-test interview. Second, abductive tasks required "backward" inference from effects to causes and are incongruent with the more natural if cause, then effect-direction. In general, backward inferences are known to be harder to draw compared to forward inferences (Evans \& Beck, 1981).

## Concluding Remarks

We investigated how people reason about conditionals under incomplete probabilistic knowledge. The novel features in our test design were comparisons of causal and abductive scenarios, as well as counterfactuals under incomplete probabilistic knowledge. One of our main findings is that the dominant response is consistent with the conditional event interpretation of conditionals among all four groups. Moreover, we discovered two major answer strategies, halfway upper and halfway lower conditional event responses, which can be understood as strategies to unburden the working memory load.

Inferentialist accounts of conditionals propose that there should be an inferential relation between the antecedent and the consequent (see, e.g., Douven, 2016b). Thus, when conditionals with inferential relations (e.g., causal or abductive ones) are compared with conditionals where no apparent inferential relation exists (like in our conditions C1 and C2), one would expect significant differences. Our data, however, do not support this inferentialist hypothesis.

The results of our paper broaden the area of inferences where conditional probability seems to be the best predictor for how people reason. We have shown that coherence-based probability logic provides a formalization of the meaning of counterfactuals and provides a rationality framework for reasoning under complete and incomplete probabilistic knowledge. This suggests that it may also be suitable for subclasses of causal reasoning like abductive reasoning, which is important in the studies on (scientific) explanation and learning.

## Acknowledgments DFG project PF740/2-2 (SPP1516).

## References

Baratgin, J., Over, D. E., \& Politzer, G. (2014). New psychological paradigm for conditionals and general de Finetti tables. Mind \& Language, 29(1), 73-84.
Barrouillet, P., Gauffroy, C., \& Lecas, J.-F. (2015). Probability in reasoning. Cognition, 22-39.
Douven, I. (2016a). Abduction. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Winter 2016 ed.). https://plato.stanford.edu/entries/abduction/.
Douven, I. (2016b). The epistemology of indicative conditionals. Cambridge: CUP.
Elqayam, S., Bonnefon, J.-F., \& Over, D. E. (Eds.). (2016). New paradigm psychology of reasoning. London: Routledge.

Evans, J. St. B. T., \& Beck, M. A. (1981). Directionality and temporal factors in conditional reasoning. Current Psychological Research, 1, 111-120.
Evans, J. St. B. T., Handley, S. J., \& Over, D. E. (2003). Conditionals and conditional probability. Journal of Experimental Psychology: LMC, 29(2), 321-355.
Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., \& Kleiter, G. D. (2011a). How people interpret conditionals. Journal of Experimental Psychology: LMC, 37(3), 635-648.
Gilio, A., Over, D. E., Pfeifer, N., \& Sanfilippo, G. (2017). Centering and compound conditionals under coherence. In M. B. Ferraro et al. (Eds.), Soft methods for data science (pp. 253-260). Berlin: Springer.
Gilio, A., Over, D. E., Pfeifer, N., \& Sanfilippo, G. (submitted). Centering with conjoined and iterated conditionals under coherence. https://arxiv.org/abs/1701.07785.
Gilio, A., \& Sanfilippo, G. (2013). Conditional random quantities and iterated conditioning in the setting of coherence. In Ecsqaru 2013 (pp. 218-229). Berlin: Springer.
Gilio, A., \& Sanfilippo, G. (2014). Conditional random quantities and compounds of conditionals. Studia Logica, 102(4), 709-729.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. Philosophical Review, 85, 297-315.
Lombrozo, T. (2012). Explanation and abductive inference. In Holyoak et al. (Eds.), The Oxford handbook of thinking and reasoning (pp. 260-276). Oxford: OUP.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality. Oxford: OUP.
Oberauer, K., \& Wilhelm, O. (2003). The meaning(s) of conditionals. Journal of Exp. Psy.: LMC, 29, 680-693.
Over, D. E. (2009). New paradigm psychology of reasoning. Thinking and Reasoning, 15, 431-438.
Over, D. E., \& Baratgin, J. (2017). The "defective" truth table. In N. Galbraith, E. Lucas, \& D. E. Over (Eds.), The thinking mind (pp. 15-28). Hove: Psychology Press.
Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., \& Sloman, S. (2007). The probability of causal conditionals. Cognitive Psychology, 54, 62-97.
Pfeifer, N. (2013). The new psychology of reasoning. Thinking \& Reasoning, 19(3-4), 329-345.
Pfeifer, N., \& Douven, I. (2014). Formal epistemology and the new paradigm psychology of reasoning. The Review of Philosophy and Psychology, 5(2), 199-221.
Pfeifer, N., \& Stöckle-Schobel, R. (2015). Uncertain conditionals and counterfactuals in (non-)causal settings. In G. Arienti, B. G. Bara, \& S. G. (Eds.), Proceedings of the EuroAsianPacific conference on cognitive science (pp. 651-656). Aachen: CEUR.
Pfeifer, N., \& Tulkki, L. (2017). Conditionals, counterfactuals, and rational reasoning. An experimental study on basic principles. Minds and Machines, 27(1), 119-165.
Wason, P. C., \& Johnson-Laird, P. N. (1972). The psychology of reasoning. Cambridge: Harvard University Press.

# Pragmatics Influence Children's Use of Majority Information 

Theresa Pham (theresa.pham@mail.utoronto.ca)<br>Department of Psychology, University of Toronto, Toronto, ON M5S 3G3 Canada

Jane C. Hu (janehu@uw.edu)
The Institute for Learning \& Brain Sciences, University of Washington, Seattle, WA 98195-7988 USA

Daphna Buchsbaum (buchsbaum@psych.utoronto.ca)<br>Department of Psychology, University of Toronto, Toronto, ON M5S 3G3 Canada


#### Abstract

Do children always conform to a majority's testimony, or do the pragmatics of that testimony matter? We investigate children's reasoning about mapping a novel word to a referent in an object-labeling task. Across four conditions, we modified the testimony in an object-labeling task, to account for pragmatic principles, so that the majority does and does not provide an explicit opinion about the alternative object chosen by the minority. Four- and 5-year-olds were given a choice between an object endorsed by a three-person majority, or one endorsed by a single minority informant. In the unendorsed condition, informants explicitly unendorsed the unchosen object. In the nothing condition, informants said nothing about the unchosen object. In the ignorance condition, informants explicitly expressed uncertainty about the unchosen object, and in the hidden condition, the chosen object was the only one present at the time of the endorsement. Children were most likely to endorse the majority object in the unendorsed condition, in which the majority explicitly stated that the label applied to only one referent, whereas in the hidden condition, where only one object at a time was present in the discourse, children chose objects endorsed by the majority and the minority equally, with the other two conditions intermediate. This suggests that children might not simply have a conformity bias; rather, they are sensitive to the majority's implied intentions when learning from testimony.


Keywords:social learning; social cognition; consensus; testimony; causal reasoning; pragmatics

## Introduction

Learning from others is especially important for young children who are growing up in a complex social world. One way children gain knowledge from others is by learning from testimony. In particular, there is a growing body of literature showing that, similar to adults, children seem to be influenced by the opinions and behavior of a majority group (e.g., Bernard, Proust, \& Clément, 2015; Burdett et al., 2016; for a recent review see Huan, van Leeuwen, \& Edelson, 2013). For example, children recognize and trust a consensus during word learning. Corriveau, Fusaro \& Harris (2009) found that 3 - and 4-year-old children view a consensus as a reliable source of information when learning novel object labels. Children were more likely to prefer novel labels that were endorsed by the majority, and to selectively trust individuals who were previously part of the majority in a subsequent task. Bernard et al. (2015) found that slightly older children (4- and 5-year-olds) also
exhibited a consensus effect, even after the majority was shown to give unreliable testimony about object labels. Children are also more likely to copy the majority's behavior (Haun, Rekers, \& Tomasello, 2012) and action sequences (Herrmann et al., 2013). As well, children seem to overconform in many situations; majority influence trumps direct source knowledge ( Hu et al., 2015), and sometimes even children's own knowledge (Corriveau \& Harris, 2010), or the knowledge of competent individuals (Burdett et al., 2016).

On the other hand, we also know from previous work that children are rational learners; they selectively learn from other people's testimony and evaluate the information they receive (for a review, see Mills, 2013; Sobel \& Kushnir, 2013; Koenig \& Sabbagh, 2013), suggesting that they might not indiscriminately endorse majority opinions. While having a majority bias in word learning is sensible due to the shared, conventional nature of word meanings, in a less socially constructed domain, such as causal learning, children may be less influenced by the majority group (Hu, Buchsbaum, Griffiths \& Xu, 2013). Similarly, children are less willing to agree with the majority's action when learning about tools if the majority endorses a function that is considered inefficient or implausible (Schillaci \& Kelemen, 2013). Additionally, children selectively learn from informants who display other indicators of reliable knowledge, including a history of providing accurate information (e.g., Pasquini et al., 2007), performing actions successfully (e.g., Wilks, Collier-Baker, \& Nielson, 2014), having expertise in the field (Burdett et al., 2016), and having privileged knowledge (Einav, 2014).

Taken together, the current literature about majority influence in children's social learning suggests that children are rational learners, but that the role of consensus widely impacts their reasoning and social learning more generally. Why then do children conform or not conform to the majority? The mechanism that underlies majority influence is still unclear. A bias to copy the majority simply because it is the majority can often be an effective social learning strategy (Laland, 2004; Perreault, Moya, \& Boyd, 2012) Conforming to the majority is a simple strategy that is often sensible and an indication of reliability (Corriveau et al., 2009). Alternatively, children might not only be attending to the number of informants, but also use pragmatic inferences for learning.


Figure 1. The arrow indicates the toy labeled by an informant as the referent of a novel word. The goal of the participant is to infer whether a novel word (e.g., "modi") means the blue toy, the purple toy, or both.

This second line of reasoning is consistent with the literature on using pragmatic cues to guide learning. Grice (1975) proposed that participants in conversation obey the maxims of cooperative communication-be truthful, informative, relevant, and clear. Specifically, the Maxim of Quantity (be informative) and the Maxim of Relevance (be relevant) are both crucial for motivating our hypothesis. To be informative means to give as much information as needed, and no more. To be relevant means to say things that is pertinent to the given context. Children might assume informants are being informative and relevant with their testimony, influencing what they learn.

Frank and Goodman (2014) showed that during word learning, children are indeed sensitive to speakers' communicative intentions, leading children to make inferences that go beyond explicit testimony. This suggests that children can make use of pragmatic principles in word learning inferences. For instance, Figure 1 illustrates a task similar to Frank and Goodman (2014). If the speaker only calls the toy on the right (marked by the arrow) "a modi," children can infer that 'modi' means the blue toy, and not the purple toy, for example, by assuming that speakers are using language relevantly and informatively.

However, in previous testimony research, the majority's opinion of the minority choice has been left ambiguous, and children's pragmatic reasoning abilities were not considered. For example, when the majority suggest that object X is the referent of a novel label or suggest using strategy X (e.g., Corriveau et al., 2009; Haun et al., 2012), this could pragmatically imply that object Y was not a referent or that they should not use strategy Y, otherwise the speaker would have referred to this option as well, in order to be informative (Frank \& Goodman, 2014). Pragmatic inferences may help children reason: "If the majority labeled object X as a modi and did not comment on object Y , they must believe that only object X is the modi. If the majority wanted me to know that object Y is also a modi, then they would have told me, because they had the opportunity to speak about object Y." Accordingly, the language used could imply that the options are mutually exclusive and only one object is a modi, for example, providing additional evidence against the minority opinion.

Given children's sensitivity to pragmatically implied information, we conducted the present study to investigate
how pragmatics can influence the strength of the majority influence in children. We examined 4- and 5-year old children's preference for the majority in an object labeling task. Specifically, we compared children's tendency to conform when the majority does and does not provide an explicit opinion about the minority's choice.

## Pragmatic knowledge versus consensus

In the present study, we investigate children's reasoning about the mapping of a novel word to a referent in an object-labeling task, when presented with a three-person majority and a conflicting minority informant. There were four testimony conditions-the unendorsed condition, the nothing condition, the ignorance condition, and the hidden condition-that varied in the informativeness of the testimony and the relevance of the object(s) present in the situation. In the most explicit case, the unendorsed condition, the majority endorsed one object and unendorsed the other object, while the minority informant provided the opposite testimony. Here, children learn from declarative testimony that makes the extent of the novel label explicit, and no pragmatic inference is needed. We hypothesize that, since the testimony in this condition explicitly states that the labels are mutually exclusive, the testimony provided by the majority group will outweigh the evidence provided by just one minority informant, and children will be more likely to endorse the majority testimony.

In the nothing condition, the informants endorsed one object and said nothing about the other. This condition was intended to replicate previous work, in which the informants' knowledge or belief about the unchosen object was left ambiguous. We predict that children will favour the majority endorsement, because, as in our example (Figure 1), they will make a pragmatic inference that the speakers are using language informatively, and so the majority must believe that the novel label does not apply to the unchosen object, otherwise they would have referred to the unchosen object using the label as well. Therefore, similar to the unendorsed condition, children in the nothing condition will infer that the labels are mutually exclusive. However, we predict that they will endorse the majority less often than in the unendorsed condition since there is additional ambiguity than when the majority explicitly states their opinion.

In the ignorance condition, the informants endorse one object but express uncertainty in their beliefs about the unchosen object. Since the majority provides information with low certainty about the extension of the novel label, their testimony should carry less weight in determining whether the unchosen object can also be referred to using the novel label. Further, the informants' uncertainty suggests that the label may not be mutually exclusive, and could apply to both objects. Thus, children should be less likely to endorse the majority's testimony, compared to the unendorsed and nothing conditions.

Finally, the language used in the testimony for the hidden condition was exactly the same as in the nothing condition, but only one object-the endorsed toy-was present. The
hidden condition relies on the pragmatic understanding that the speaker is being informative and relevant in their testimony, and is therefore only speaking about object that is present in the discourse context. The result is that, if children make this pragmatic inference-speakers cannot comment on objects that they and their communicative partners do not see-then in the hidden testimony, the majority provides no or the least amount of evidence against the minority's choice. On the other hand, if children are not sensitive to the pragmatics of the testimony, then the results for the hidden condition should be the same as the nothing condition, with children endorsing the majority's choice.

Whereas the current literature supports children's conformity bias as a learning strategy, a pragmatic explanation would suggest that children do not indiscriminately conform to the majority. The overarching aim is to show that children are sensitive to pragmatic principles even if they have a general tendency to trust the majority. That is, the extent of children's conformity depends on the pragmatics of the testimony. Specifically, as the majority's testimony becomes more explicit in their negative judgment of the minority's opinion, children will be more likely to adopt the majority's endorsement over the option endorsed by the minority informant. Thus, we predict that the tendency to endorse the majority will decrease over the conditions: unendorsed condition (most majority endorsement); then, nothing condition; then, ignorance condition; and finally, the hidden condition should exhibit the least majority bias. Alternatively, if children do exhibit a global conformity bias, then they should indiscriminately endorse the majority's opinion regardless of the testimony.

## Methods

Participants. Participants were 112 preschoolers, 49 females and 63 males (mean age $=4$ years 8 months; range $=47-71$ months). An additional 23 were excluded from the study because of experimenter error (9), participant distraction (9), failure to make a choice (2), and failure to remember object label (3). Participants were recruited from the University of Toronto database or from public neighbourhood parks and museums.
Participants were randomly assigned among the four
between-subject conditions: the unendorsed condition ( $\mathrm{n}=$ $28, \mathrm{M}=59$ months, range $=50-71$ months, $32 \%$ female), the nothing condition ( $\mathrm{n}=25, \mathrm{M}=57$ months, range $=48-$ 71 months, $40 \%$ female), the ignorance condition ( $\mathrm{n}=31$, $\mathrm{M}=58$ months, range $=47-71$ months, $52 \%$ female), and the hidden condition ( $\mathrm{n}=28, \mathrm{M}=56$ months, range $=48-$ 67 months, $50 \%$ female).
Materials and Procedure. Children were tested individually. In all conditions, each participant participated in two test trials, a modi trial and a dax trial. Each trial featured two novel objects for a total of four unique objects of differing shape and colour in order to reduce extension.

The trial presented first was counterbalanced across participants. The object pairs and side on which each object was presented were held fixed but the object chosen by the majority and minority was counterbalanced.

To begin each condition, children sat at a table across from the experimenter. The experimenter introduced children to two novel objects and explained that they were unknowledgeable about the labels of the objects. The experimenter suggested that the participant watch a film to learn about the objects' label. Participants then watched a pre-recorded film of four female informants evaluating the objects on a 13 " laptop screen.

A film consisted of four video clips, each featuring a female informant sitting by herself at a table with the same novel objects. Informants wore different colour shirts. In the first three clips, the three-person majority each endorsed one object with the novel label, and in the final clip the one minority informant endorsed the other object with the same novel label, repeated three times so that the frequency with which each participant heard the label used to refer to each object was equal. Each clip concluded with the informant picking up the toy they had endorsed. The identity of the minority informant was counterbalanced across participants. Figure 2 displays schema for the videos shown.

In the unendorsed condition, each majority informant endorsed one object while explicitly unendorsing the other object by saying, "That's a modi (pointing to target toy); that's not a modi (pointing to other toy)." In the ignorance condition, each majority informant endorsed one object while expressing uncertainty about the other object by saying, "That's a modi (pointing to target toy); I don't know


Figure 2. Schema of the videos seen by children. (i) Stimuli placement for the Unendorsed, Ignorance and Nothing condition. Testimony from the Nothing condition. (ii) Stimulus placement and testimony from the Hidden condition.
if that's a modi (pointing to other toy)." In the nothing condition, the majority informants endorsed one object and did not comment on the other object by saying, "That's a modi (pointing to target toy)." In the hidden condition the informant sat at the table with only one object and evaluated that object by saying, "That's a modi (pointing to target toy)." In all conditions the minority informant endorsed the other object with the same novel label three times. For instance, in the unendorsed condition the minority informant said "That's a modi (pointing to other toy); that's not a modi (pointing to target toy). Look at that modi (pointing to other toy); that's not a modi (pointing to target toy). It's a pretty cool modi (pointing to other toy); that's not a modi (pointing to target toy)". The minority scripts in the other conditions followed in the same manner.

Once the film ended, the screen turned black and the objects were brought back. The experimenter then asked the participant to identify the referent of the novel label by asking, e.g., "Can you show me a modi?" Participants' first gestural or vocal response was recorded.

## Results

Table 1: Participant scores by condition

|  | Score |  |  |
| :---: | :--- | :--- | :--- |
| Condition | $\mathbf{0}$ | $\mathbf{1}$ | $\mathbf{2}$ |
| $(1)$ Unendorsed | 0 | 4 | 24 |
| $(2)$ Nothing | 3 | 9 | 13 |
| $(3)$ Ignorance | 4 | 8 | 19 |
| $(4)$ Hidden | 6 | 12 | 10 |



Figure 3. Average number of responses (+/- 1 s.e.) endorsing the majority object. There was a significant effect of condition. Children chose the majority's object most often in the unendorsed condition, and least often in the hidden condition. $* * p<0.01 \quad * * * p<0.001$.

Participants were assigned a score $(0,1$, or 2$)$ based on the number of trials in which they endorsed the majority informants' testimony (see Table 1). ${ }^{1}$ Children's mean responses for all conditions are shown in Figure 3. For all conditions, chance level was a mean score of 1 .

[^434]Children chose the majority's referent to the novel object significantly more often than chance in the unendorsed $t(27)$ $=11.15, \mathrm{~d}=4.29, p<.0001)$; nothing $(t(24)=3.09, \mathrm{~d}=1.26$, $p<0.01)$; and ignorance conditions $(t(30)=3.72, \mathrm{~d}=1.36$, $p<0.001)$. In contrast, in the hidden condition, participants were not more likely to adopt the majority's opinion than the minority's $(t(27)=1.11, \mathrm{~d}=0.39, p=0.24)$.

We also found a significant effect of condition on children's tendency to choose the majority object, onefactor ANOVA $F(3,108)=5.3, \mathrm{MSE}=2.39, p<0.01$. Planned two-sample t-tests for independent samples demonstrated that the unendorsed condition was significantly different compared to all the other conditions: unendorsed vs. nothing: $t(51)=-2.45, \mathrm{~d}=0.67, p<0.01$; unendorsed vs ignorance: $t(57)=-2.19, \mathrm{~d}=0.58, p<0.05$; and, unendorsed vs. hidden: $t(54)=4.31, \mathrm{~d}=1.15, p<$ 0.0001. Similarly, children's performance in the hidden condition was significantly different from the ignorance condition, $t(57)=1.92, \mathrm{~d}=0.50, p<0.05$; and, marginally different from the nothing condition, $t(51)=1.61, \mathrm{~d}=0.44$, $p=0.06$. There was no significant difference in children's performance in the nothing condition compared to the ignorance condition, $t(54)=-0.23, \mathrm{~d}=0.06, p=0.41$.

Follow-up polynomial contrasts indicate asignificant linear trend, $F(1,108)=13.67, p<0.001$, partial $\eta^{2}=0.11$. The linear trend suggests that deference to majority decreases across ordered conditions: unendorsed ( $M=1.82$, $\mathrm{SD}=0.074$ ), nothing ( $\mathrm{M}=1.44, \mathrm{SD}=0.711$ ), ignorance ( M $=1.48, \mathrm{SD}=0.724)$, and hidden $(\mathrm{M}=1.11, \mathrm{SD}=.079)$.

## Discussion

To be rational yet efficient social learners, it would be beneficial for children to learn through explicit instruction as well as pragmatic inferences. This study provides the first empirical evidence that children consider pragmatic inferences when learning from testimony provided by a majority and minority group. We examined the effects of the pragmatic cues in informants' testimony on children's tendency to defer to the majority. Our study found that although children tend to be influenced by the majority, they also weigh informants' opinions using pragmatic cues to assess the meaning of their testimony. By assuming that informants are being cooperative in their communicative intent (e.g., Maxim of Quantity; Maxim of Relevance), children are evaluating the pragmatic implications of the language used and making inferences that go beyond the literal meaning of the testimony.

We found that when learning from explicit, declarative testimony, as was the case in the unendorsed condition, 4and 5- year old children were significantly more likely to endorse the majority object-their tendency to endorse the majority option was almost at ceiling. In this condition, the informants' opinions about both of the novel objects were made explicit in thelanguage of the testimony. They endorsed one object using the novel label (e.g., modi) and provided additional evidence that the unchosen object was not a modi. In this case, following the majority is a sensible
strategy since the explicit endorsement of only one object by three people might outweigh the evidence provided by just one minority informant.

There was also a consensus effect in both the nothing condition-replicating the findings in previous work (e.g., Corriveau et al., 2009) - and the ignorance condition. And yet, in these conditions the majority bias was significantly less than in the unendorsed condition, suggesting that children are aware of the additional ambiguity in these conditions. By contrast, children did not exhibit a majority bias in the hidden condition. The crucial difference between these conditions is the pragmatic inferences made, given the ambiguity of the learning situation.

When the speaker's testimony is ambiguous, as in the nothing condition, children might rely on pragmatic cues to infer the speaker's intent. According to the pragmatic account, a crucial step in the inferential process is the assumption that the speaker, in this case, the informants, is being cooperative with their utterance, and has the goal of being informative. Accordingly, if the speaker had wanted to label both objects then they had the ability to do so, as in the unendorsed and ignorance conditions. The fact that the informants only ever labeled one object in the nothing condition led children to infer that the novel label is only applicable to one object in the given situation, leading to the conclusion that the unchosen object is not a referent of the novel label. However, in the hidden condition, only one object was present, invoking a different pragmatic inference than in the nothing condition-the inference that speakers are only discussing objects relevant to the current situation. Consequently, even though the testimony in these two conditions was identical, children's inferences differed. Children in the hidden condition were no more likely to endorse the majority's testimony than the minority's. This suggests that children's inferences from consensus are influenced by their sensitivity to pragmatic cues embedded in the testimony.

A somewhat unexpected finding was that children were also more likely to choose the majority's object in the ignorance condition. In the ignorance condition, we intended for the uncertainty about the extension of the object label to come from the uncertainty in the speaker's knowledge. Since informants expressed low certainty in their testimony, there should be less evidence against the unchosen object being e.g., a modi, and by association, against the minority. However, children might instead have interpreted the statement of ignorance as a comment about the object rather than about the informants' knowledge. Children might have inferred that the majority was certain about one object having many features of e.g., a modi, but was uncertain about the other object due to its ambiguous or hard to categorize appearance. Future work should disambiguate the type of uncertainty being conveyed.

Together, the results go beyond asking whether or not children have a conformity bias, and explore children's sensitivity to pragmatically implied information. Children's tendency to conform might not simply be driven by a 'copy-
the-majority' strategy (Laland, 2004), since they did not conform to the majority equally across conditions. Instead, the extent to which children prefer members of the majority as informants might vary with the pragmatics of the language used in the testimony. Children may not be overestimating the value of the majority's opinions compared to the minority's testimony. Instead, children are sensitive to the language used by informants, and hence, are selective about situations in which they should go with the consensus view. This finding is consistent with previous work suggesting that children are ableto make sophisticated inferences about implicit, intended meaning in speakers' utterances (e.g., Frank \& Goodman, 2014).

An interesting question is whether our results are specific to the use of an object-labeling task. For instance, children may be more likely to conform to a majority in a labeling task because the meanings of words are socially determined (Hu et al., 2013). In addition, labels may have stronger implications due to other constraints on word learning, such as mutual exclusivity-each object has only one category label (Markman, 1989), and the shape bias-differently shaped objects usually have different category labels (Landau, Smith \& Jones, 1988). For example, if an informant labels Object 1 as a modi, this strongly implies that Object 2, which is very differently shaped, is not a modi. This would provide additional evidence against the minority, by suggesting that both objects cannot be modis.

In our experiment, we were able to modify the testimony so that the strength of the implications regarding the novel label for the object(s) varied by condition. However, if children were exclusively following these types of language learning constraints, they should have assigned a novel label to only one object in all conditions. On the contrary, children's performance followed a linear trend and did not favour the majority label in the hidden condition.

However, future work should examine children's endorsements of majority and minority information in other domains such as causal learning (e.g., Hu et al., 2013). Causal tasks do not rely on the conventions and social construction that make learning labels special. Since causal actions on objects might not be mutually exclusive by nature-one causal action does not necessarily imply that the other actions are ineffective-future work can directly examine how making actions appear mutually exclusive using pragmatic principles can affect children's reasoning. We would be interested in children's reliance on consensus during a causal task, and in turn, how they are making pragmatic inferences about the efficacy of each action.

One further question concerns the operationalization of consensus and the presentation of a majority group versus the minority informant. In this paper, we presented the informants individually, and in a sequential manner with the majority group first, similar to previous work by Haun et al. (2012) and Burdett et al. (2016), and importantly, this was held constant across all of the conditions. However, in both Bernard et al. (2015) and Corriveau et al. (2009), the majority and minority informants were all presented
together, and at the same time rather than one after another. It is possible that this difference in the format of majority presentation might change the pragmatics of the situations, for example if children believe that the dissenter in a simultaneous presentation is attempting to correct the majority's misinformation. Future work could compare simultaneous and sequential presentation of the informants to see how this changes children's inferences.

Finally, it is difficult to quantify our predictions without formalizing our assumptions. A Bayesian model could produce quantitative predictions regarding the ordering of our conditions, and the magnitude of the differences between them. This type of model could examine how a rational learner would balance a majority opinion against the pragmatic implications of their testimony, without a conformity bias, and test those predictions against children's behaviour, building on previous models of learning from testimony (e.g., Buchsbaum et al., 2012) and of making pragmatic inferences (e.g., Frank \& Goodman, 2014).

In sum, this research sheds light on how pragmatic principles can inform children's learning from conflicting majority and minority groups. In conditions where the testimony explicitly stated, or pragmatically implied, that the labels were mutually exclusive, children were more likely to adopt the majority's label than the minority's label. However, when the testimony had weaker implications about the labels of the novel objects, children were not more likely to rely on the consensus view. This suggests that children might not simply have a conformity bias. Instead, children can make sophisticated inferences that go beyond the literal meaning of the testimony. By doing so, they consider both the explicit statements made by informants, as well as the pragmatic inferences implied by the majority opinion in their learning from the social world.

## Acknowledgments

We are grateful for all the families for their participation, and to the Computational Cognitive Development lab for their help with data collection. We would especially like to thank Pingki Mazmunder for help with data collection and analysis.This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canada Foundation for Innovation, and the Connaught Fund.

## References

Bernard, S., Proust, J., \& Clément, F. (2015). Four- to six-year-old children's sensitivity to reliability versus consensus in the endorsement of object labels. Child Development, 86, 11121124.

Buchsbaum, D., Bridgers, S., Whalen, A., Seiver, E., Griffiths, T. L., \& Gopnik, A. (2012). Do I know that you know what you know? Modeling testimony in causal inference. Proceedings of the 34th annual conference of the cognitive science society ( pp . 156-161). Austin, TX: Cognitive Science Society.
Burdett, E. R. R., Lucas, A. J., Buchsbaum, D., McGuigan, N., Wood, L. A., \& Whiten, A. (2016). Do children copy an expert or a majority? Examining selective learning in instrumental and normative Contexts. PLoS One, 11, e0164698.

Corriveau, K. H., Fusaro, M., \& Harris, P. L. (2009). Going with the flow: preschoolers prefer nondissenters as informants. Psychological Science, 20, 372-377.
Corriveau, K.H. \& Harris, P.L. (2010). Preschoolers (sometimes) defer to the majority when making simple perceptual judgments. Developmental Psychology, 26, 437-445
Einav, S. (2014). Does the majority always know best? Young children's flexible trust in majority opinion. PloS one, 9, e104585.
Frank, M. C., \& Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. Cognitive Psychology, 75, 80-96.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and semantics (Vol. 3). New York, NY: Academic Press.
Herrmann, P. A., Legare, C. H., Harris, P. L., Whitehouse, H. (2013). Stick to the script: The effect of witnessing multiple actors on children's imitation. Cognition, 129, 536-543.
Hu, J. C., Buchsbaum, D., Griffiths T. L., \& Xu F. (2013). When does the majority rule? Preschoolers trust in majority informants varies by domain. Proceedings of the 35th annual conference cognitive science society (pp. 2584-2589), Austin, TX: Cognitive Science Society.
Hu, J. C., Whalen, A., Buchsbaum, D., Griffiths, T. L., \& Xu, F. (2015). Can children balance the size of a majority with the quality of their information? Proceedings of the 37th annual conference cognitive science society (pp. 956-961), Austin, TX: Cognitive Science Society.
Haun, D. B. M, van Leeuwen, E. J. C., \&Edelson, M. G. (2013). Majority influence in children and other animals. Developmental Cognitive Neuroscience, 3, 61-71.
Haun, D. B. M., Rekers, Y., \&Tomasello, M. (2012). Majoritybiased transmission in chimpanzees and human children, but not orangutans. Current Biology, 22, 727-731.
Koenig, M. A. \&Sabbagh, M. A. (2013). Selective social learning: New perspectives on learning from others. Developmental Psychology, 49, 399-403.
Laland, K. N. (2004). Social learning strategies. Learning and Behavior,32, 4-14.
Landau, B., Smith, L. B., Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive Development, 3, 299321.

Markman, E. M. (1989). Categorization and naming in children: Programs of induction. MIT Press.
Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. Developmental Psychology, 49, 404-418.
Pasquini, E. S., Corriveau, K. H., Koenig, M., \& Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. Developmental Psychology, 43, 1216-1226.
Perreault, C., Moya, C., \& Boyd, R. (2012). A Bayesian approach to the evolution of social learning. Evolution and Human Behavior, 33(5), 449-459
Schillaci, R. S., \&Kelemen, D. (2014). Children's conformity when acquiring novel conventions: The case of artifacts. Journal of Cognition and Development, 15, 569-583.
Sobel, D. M., \&Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. Psychological Review, 120, 779-797.
Wilks, M., Collier-Baker, E., \& Nielson, M. (2014). Preschool children favor copying a successful individual over an unsuccessful group. Developmental Science, 18, 1014-1024.

# A categorical (fixed point) foundation for cognition: (adjoint) corecursion 

Steven Phillips<br>National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan


#### Abstract

Computationalism has been the pre-eminent framework for models of mind, since the cognitive revolution. However, the plethora of apparently incommensurate approaches seems to undermine hope for a common computational foundation. Category theory provides a mathematically rigorous foundation for computation that includes recursion and corecursion. We show that corecursion unifies various cognitive behaviours for comparison and contrast in a principled and novel way. For instance, Chomsky's merge function is a universal morphism, which has a dual, called comerge. One implication of this work is that corecursion appears to be the rule rather than the (human) exception in contrast to Chomsky's view of recursion.


# Dual-routes and the cost of computing least-costs 

Steven Phillips<br>National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan<br>Yuji Takeda<br>National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan<br>Fumie Sugimoto<br>National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan


#### Abstract

Theories of cognition that posit complementary dual-route processes afford better fits to the data when each route explains a part of the data not explained by the other route. However, such theories must also explain why each route is invoked, lest one can fit any data set with enough routes. One possible explanation is that route selection is based on a least-cost principle: the route that requires fewer cognitive resources (including time) relative to the goal at hand. We investigated this explanation with a dual-route version of visual search, where the target could be identified via opposing (easy or hard forms of) feature and conjunction search conditions. The data support a contextualized version of the least-cost principle in that the cost of computing least-cost also influences route selection: participants assessed alternatives, but only when the cost of that assessment was relatively low.


# The Effect of Economic Scarcity Priming on Perception of Race 

Dariusz Lukasz Pilucik (dpiluc01@mail.bbk.ac.uk)<br>Department of Experimental Psychology, University of Oxford<br>South Parks Rd, Oxford OX1 3UD, Oxford, United Kingdom

Jens Koed Madsen (jens.madsen@ouce.ox.ac.uk)<br>School of Geography and the Environment, University of Oxford<br>OX1 3QY, South Parks Road, Oxford, United Kingdom<br>Orcid: 0000-0003-2405-8496


#### Abstract

Existing research suggests that White individuals are more likely to categorise biracial faces as Black in conditions of resource scarcity. It has been theorised that this effect is due to in-group boundaries becoming more exclusive in scarce conditions. An alternative explanation refers to implicit socioeconomic association between Black individuals and lower level of resources. These two approaches entail different predictions for Black participants performing the categorisation task. If scarcity prompts greater in-group exclusivity, Black participants should, ceteris paribus, categorise more biracial faces as White. If, however, scarcity invokes socio-economic status associations, Black participants should categories biracial faces in the same way as White participants. Experiment 1, explored the effects of priming on White and Black groups. It provided support for the implicit socio-economic association theory. Furthermore, experiment 2 on Asian sample, provided additional support as Asian participants showed the same pattern of response. The paper discusses implications of these findings.


Keywords: intergroup bias; perception of race; boundary formation; resource scarcity; scarcity priming

## Introduction

Substantial evidence suggests that the perception of race can be influenced by environmental factors (e.g. Davis, 1991; Peery \& Bodenhausen, 2008; MacLin \& Malpass, 2001). This effect was examined by Rodeheffer, Hill and Lord (2012), who primed participants with the concepts of resource scarcity, abundance and neutral control. The priming stimuli included slides with captioned pictures showing relevant concepts. Participants were then presented with twenty composite images of biracial faces generated using $50 \%$ content of White and $50 \%$ content of Black individuals' photographs. The task was to categorise faces as either Black or White. Results showed that participants in the scarcity condition were more likely to categorise faces as Black, relative to the abundance or neutral conditions. There was no significant difference between neutral and abundance conditions. Rodeheffer et al. theorised that, in times of economic crisis, boundaries of in-group categorisation become more exclusive where less people are classified as in-group members.

Results from Krosch and Amodio (2014) support these conclusions. Participants were primed with scarcity, negativity, and neutral conditions. Priming consisted of
subliminal presentation of relevant word primes for 20 ms prior to each trial. Participants subsequently categorised faces as either Black or White. Facial stimuli included morphed pictures of mixed race individuals generated using different proportions of original faces of Black and White people at $10 \%$ increments. The dependent variable was the point of subjective equality, defined as the proportion of original Black face content required for the morphed image to be equally likely to be categorised as either Black or White. Results showed that in the scarcity condition, a given image may have significantly lower proportion of Black content to reach the point of subjective equality. In other words, a higher number of images were identified as Black in the scarcity condition. Consistently with previous explanations of Rodeheffer et al. (2012), authors of the study concluded that in the conditions of resource scarcity the boundaries of in-group categorisation become more exclusive, as participants tended to exclude biracial individuals from the White group in the scarcity condition. However, the study had a certain limitation as a quarter of the sample consisted of Asian participants. The facial stimuli included only images of Black and White individuals and no Asian individuals. The experiment was therefore unable to examine any effect of change of ingroup boundaries since there were no in-group images in the stimuli from the point of view of Asian participants. This may indicate that the effect was driven by other factor. The present study explores this possibility.

The paper thus postulates an alternative theoretical explanation relating to the effect of implicit racial bias, based on presumption that Black people constitute a disadvantaged group with lower level of resources (Gilens, 2003). That is, people's sematic network may include an association between scarcity (or lower socio-economic status) and Black individuals. Activation of the concept of scarcity may therefore result in higher cognitive accessibility of the concept of Black individuals. This in turn increases the probability of a given face to be classified as Black following scarcity priming. Extensive previous literature demonstrated a wide range of similar implicit racial biases, which, more importantly, are exhibited by individuals across different ethnicities, including Black people themselves (e.g. Payne, 2006; Correll, Park, Judd \& Wittenbrink, 2002).

We might therefore reconsider the above evidence from Rodeheffer et al. (2012) and Krosch and Amodio (2014). It is possible that the results were driven not by a change in ingroup exclusivity, but by implicit association between lower level of resources and Black individuals. The present study investigates these two theoretical explanations by testing different ethnic groups. According to the initial account, scarcity yields an increase in exclusivity of the in-group categorisation. If this theory is correct, White participants will show a tendency to exclude biracial individuals from the White in-group, increasing the number of faces being judged as Black. Conversely, Black participants should show the opposite pattern of response. According to the alternative explanation, scarcity priming activates the concept of lower level of resources and increases cognitive accessibility of the associated concept of Black individuals. Similar to other implicit race-related biases, this effect can be expected to occur irrespective of the ethnicity of the participant. Therefore, both Black and White participants will show the same pattern of responses, with increased probability of judging ambiguous faces as Black in the scarcity condition, relative to the neutral condition. As neither of the original studies included Black participants, new data is required to decide between the two hypotheses. Experiment 1 directly tests these predictions by analysing responses from Black and White participants. Experiment 2 further tests the hypotheses by analysing responses from Asian participants. Asian people are out-group members to both Black and White faces included in the stimuli. According to the in-group exclusivity theory, they should be unaffected by scarcity priming, as no change in in-group exclusivity can be observed. According to the socioeconomic account, however, they should exhibit similar response patterns, as the White and Black participants from experiment 1.

## Experiment 1: Black and White participants

In the original studies, participants were White. Given the competing hypotheses (in-group/out-group versus implicit biases), we recruited both Black and White participants. If the group hypothesis is correct, we should expect Black participants to exhibit the opposite in-group behaviour as White participants. If, however, the implicit bias hypothesis is right, we should expect both groups of participants to have the same response patterns.

## Participants

Sixty-four people participated in the experiment ( $\mathrm{N}=64$ ); 40 female, 23 male and 1 person classified their gender as "other". The amount of participants was chosen prior to recruiting the participants, and consequently no direct stopping rule was in place. The age of participants ranged from 18 to 55 (mean $=29.84, S D=8.69)$. Participants were recruited opportunistically from Birkbeck, University of London.

## Design

The experiment included two independent variables: priming (two levels: scarcity and neutral), and group (two levels: Black or White). The experimental design was 2 x 2 mixed with group as the between subject variable and priming as the within subject variable. The dependent variable measured the mean response rate across all trials in both conditions (scarcity or neutral) for each group (Black or White). The scope of possible values ranged from 1.0 indicating that $100 \%$ of 45 images presented were identified as Black to 2.0 indicating that $100 \%$ of images were identified as White. The mean score of 1.5 represents that $50 \%$ of the images were classified as White and $50 \%$ as Black.

## Materials

The experiment used Qualtrics survey software and could be conducted online. The priming stimuli for the scarcity condition were obtained from the Rodeheffer et al. (2012) study, as they were proven to be of sufficient quality to produce priming effects. They consisted of captioned pictures showing economic recession, poverty and scarce resources. The stimuli for the neutral condition included three slides showing pictures of clocks, as there is no reason to assume that images of clocks should prime participants in terms of economic scarcity or abundance. Similarly, in order to replicate and test previous findings, the facial stimuli were obtained from the Krosch and Amodio (2014) study. They included morphed pictures generated from photographs of Black and White individuals at $10 \%$ increments from $10 \%$ Black to $90 \%$ Black $^{1}$.

## Procedure

In accordance with ethical approval requirements, participants consented to their involvement and that they could terminate the experiment at any time if they so desired. First, participants saw slides with neutral priming (images of old-fashioned clocks). Having seen the images, participants were presented with 45 morphed facial images (from Krosch \& Amadio, 2014), one picture per slide. Following Rodeheffer et al. (2012), the instruction read: "If you had to choose, would you describe this person as [Black/White]". Participants indicated their response by clicking one of two radio buttons. Following the first set of pictures, participants were presented with the scarcity priming (caption and images of economic scarcity), and then the second set of 45 pictures. The experiment ended with a page requesting the demographic information, including ethnicity, gender and age. The last page included debriefing. The entire procedure lasted for about 15 minutes.

[^435]
## Experiment 1: Results

Only data of participants who classified themselves as Black or White were analysed. The means and standard deviations are presented in Table 1. The lower the mean response rate, the higher is the bias towards categorising faces as Black (cf. the above description of the design). The descriptive statistics show that in neutral conditions the results were close to the expected value of 1.5 . The results in the experimental conditions were lower, which means that more images were classified as Black.

|  | Black | White |
| :---: | :---: | :---: |
| Neutral priming | $1.51(.06)$ | $1.47(.13)$ |
| Scarcity priming | $1.42(.07)$ | $1.43(.14)$ |

Table 1: Scarcity effect on participant groups

A mixed 2-way ANOVA revealed a significant main effect of priming: $F(1,62)=31.57, p<0.001$. Thus, results varied significantly between the priming conditions across both groups such that more faces were categorised as Black in the scarcity condition. The main effect of group was not significant: $F(1,62)=.28, p=.60$, indicating that both groups had similar response rates across the priming conditions. In addition, there was no significant interaction: $\mathrm{F}(1,62)=.135, \mathrm{p}=.14$. This shows that the effect of priming did not vary as a function of the group variable. In order to further test effects of scarcity within each group separately, two 1 -way repeated measures ANOVAs were performed. The test in the Black group confirmed a significant effect of scarcity: $F(1,33)=9.27, p=.003$. Similarly, in the White group, a significant effect was also obtained: $F(1,29)=23.13, p<.001$ (see fig. 1). This confirms that the main effect of priming in the 2-way ANOVA was driven by significant differences in both groups. As illustrated in Figure 1, results of the experiment were consistent with the poverty priming hypotheses.


Fig. 1: Main results, experiment 1

## Experiment 2: Asian participants

Experiment 1 showed that both Black and White participants rated more faces as Black in scarcity conditions. Experiment 2 pursues this hypothesis by testing Asian participants. As these are out-group members for both Black and White faces, they should, according to the in-group/outgroup hypothesis, be unaffected by the priming. If, however, the implicit socio-economic association hypothesis is correct, Asian participants should respond in a similar way as the Black and White participants in experiment 1. Given these results, we predict the latter hypothesis.

## Participants

Thirty-one people of Asian origin were recruited from Mechanical Turk ( 8 female, 23 male$)^{2}$. In order to ensure Asian origin without priming the participants to consider this as an issue for selection, a pre-screening including a number of filler questions unrelated to race and ethnicity as well as information to identify the respondent's ethnicity and race was conducted to pick out participants with relevant ethnicity. Only Asian participants would subsequently allowed to continue with the actual experiment. The age of participants ranged from 18 to 60 (mean $=30.00, S D=9.55$ ).

## Design, materials, and procedure

As in experiment 1 , scarcity and neutral conditions were independent variables. The design was within-subjects, hence all participants were allocated to both conditions. The experiment used the same dependent measure as in the previous experiment.

## Experiment 2: Results

Only data of participants who classified themselves as Asian were analysed (two participants who passed the qualification test, subsequently identified themselves as mixed-race). The mean response rate in the neutral condition was $1.48(S D=.14)$. This result was close to the chance rate of 1.5 . The outcome in the scarcity condition was $1.43(S D=.13)$, revealing that more pictures were classified as Black. A related sample t-test showed that the effect was significant $(t(30)=2.35, p=.013) .{ }^{3}$ The results from experiment 2 are thus in line with findings from experiment 1 and suggest that scarcity priming invoked implicit socio-economic associations rather than increasing in-group/out-group exclusivity.

[^436]
## Discussion

The outcomes of the experiments demonstrate that priming the concept of resource scarcity changes the perception of race. Participants in the scarcity condition were more likely to categorise biracial faces as Black, relative to the neutral condition. This effect has been shown to exist across all groups included in both experiments. The results provide empirical support for the theory of implicit association between the concept of poverty and Black individuals, yielding more faces to be categorised as Black. The results challenge the original explanation presented by Rodeheffer et al. (2012) and Krosch and Amodio (2014). Here, the shift of perceptual threshold of racial categorisation is based on scarcity-related increase of ingroup exclusivity. This was theorised to facilitate resource allocation towards the ingroup. The present results showed that Black, White, and Asian participants had the same pattern of response, which is contrary to the hypotheses derived from their theory. Thus, the present study confirms that the implicit poverty priming theory has higher explanatory power relative to the reduction of the ingroup inclusivity theory.

The results of the study are of high social significance in terms of intergroup relationships, potentially concerning the distribution of wealth and power, and regarding implicit socio-economic associations. The implicit association between poverty and Black individuals may be related to the stereotype of a Black person as poor and chronically welfare dependent. According to Gilens (2003), media portrayal of poverty has become increasingly "racialized" - it shows disproportionately higher number of Black people depicted as poor. He found that almost $60 \%$ of images in American articles on poverty present Black individuals, whereas around $27 \%$ of the poor are Black. This tendency culminated in articles published in 1962 and 1963 during a broad coverage of welfare system abuses which saw $75 \%$ of images representing Black people. This trend reversed in the early 1980s in the times of economic downturn, with the percentage reduced to $33 \%$. These changes indicate that the media increase the overrepresentation in the negative context of welfare system abuse and the "undeserving" poor. Furthermore, they decrease the overrepresentation when poverty can be justified by overall economic decline. Further studies show that attitudes towards the poor are context dependent, e.g. people are more likely to classify poor senior citizens or medical care receivers as deserving public assistance (Smith, 1987; Cook \& Barrett, 1992). Consistently with these notions, Gilens (1999) established that the representation of Black people among the poor varies as a function of positivity of the context. No images of Black individuals were found in articles on senior citizens. Consistently, the overrepresentation in articles on underclass, urban problems and criminality ranged between $85 \%$ and $100 \%$. It can, therefore, be theorised that the disproportionate representation of Black people among the poor in the media contributes to the establishment of the implicit association. In addition, this effect is magnified by
negative context of the media article, which may bias people's beliefs concerning reasons for Black poverty. Indeed, studies demonstrated that news reports on poor Black children produce more personal attributions relative to news reports on poor White children (Iyengar, 1991). In other words, participants were more likely to attribute Black poverty to alleged negative personal characteristics of Black people, while White poverty is more likely to be explained in terms of structural and social factors rather than personal (e.g. economic crisis, unemployment etc.).

Another socially significant consequence of this implicit association relates to the fact that people tend to behave consistently with the content of activated stereotype. In a classic study, Bargh, Chen and Burrows (1996) examined this notion by priming participants with the concept of elderly people. Results showed that participants walked more slowly following this priming, which is part of the stereotype of senior citizens. In addition, priming participants with the concept of Black individuals resulted in a more aggressive behaviour. These phenomena are consistent with previously outlined theoretical explanations of increased cognitive accessibility of concepts associated with currently activated ideas. The study showed that stereotypes not only affect task performance in laboratory settings but also can affect daily behaviour. Results consistent with this notion were obtained by other studies which used the stereotype threat paradigm. Research showed that priming people with racial or gender stereotypes (e.g. lower mathematic ability of Black people and women, while higher ability of Asians) results in stereotype-congruent behaviour (e.g. Steele, 1997; Steele \& Aronson, 1995; Walton \& Cohen, 2003; Steen, 1987). Further evidence demonstrates that the effects of stereotype activation are not limited to academic or IQ tests only, but also extend to economic decision making. For example, women primed with gender stereotypes are less likely to engage in risky financial activities (Powell \& Ansic, 1997; Schubert, Brown, Gysler \& Brachinger, 1999), and they are less likely to pursue careers traditionally regarded as male (Rudman \& Phelan, 2010).
Collectively, these studies suggest that stereotypes can have a long-lasting effect on a range of aspects crucial for life success, such as education and career choice, which translates into wealth and social status. It can be therefore argued that negative racial stereotypes hinder the prospects of Black people, since people have a tendency to unwittingly act in accordance with the content of internalised stereotypes. Concurrently, socio-cultural stereotyping may be activated by relevant cues such as economic scarcity (as the current studies explore). Black people, therefore, experience double social jeopardy effects of implicit racism of White people and damaging effects of the internalised stereotypes. These phenomena contribute to the continuation of Black poverty, which reinforce stereotypes concerning Black people. It can also be hypothesised that biased media portrayal of Black people as poor and dependent on welfare further magnifies this
effect. This means that the media and other organisations, e.g. charities which advertise fundraising campaigns for people in Africa, should be made aware of the negative effects that this continuing bias has. Due to holding social responsibility, the media articles or advertisement should present a balanced and accurate picture of reality. Issues concerning the relationship between socio-cultural stereotypes, group dynamics, socio-economic power, ethnicity, and cognitive function are complex, multifacetted, and inter-dependent. They relate to very real problems in society, and further studies are warranted to explore these delicate aspects, how they function, and how they relate to one another in much more detail.

## Author contributions

D. L. Pilucik developed the idea of the study, rationale and method, conducted data collection and analysis and wrote manuscript draft. J. K. Madsen provided support with developing the study, critical revision, finalised the manuscript, and prepared it for submission. Both authors approved the final version of the manuscript for submission.

## Acknowledgements

The authors of the study wishes to express their gratitude to Amy Krosch and Christopher Rodeheffer for supplying the stimuli materials (see Krosch \& Amodio, 2014; Rodeheffer et al., 2012).

## References

Bargh, J., A., Chen, M., \& Burrows, L. (1996). Automaticity of Social Behaviour. Direct Effects of Trait Construct and Stereotype Activation on Action. Journal of Personality and Social Psychology, 71(2), 230-244.
Cook, F. L., \& Barrett, E. J. (1992). Social Welfare Policy in America: The 1980s as a Crossroads Support for the American Welfare State. New York: Columbia University Press.
Correll, J., Park, B., Judd, C.M., \& Wittenbrink, B. (2002). The police officer's dilemma: Using race to disambiguate potentially threatening individuals. Journal of Personality and Social Psychology, 83, 1314-1329.
Davis, F. J. (1991). Who is Black? One Nation's Definition. University Park: Pennsylvania State University Press.
Gilens, M. (1999). Why Americans Hate Welfare: Race, Poverty, and the Politics of Antipoverty Policy. Chicago: University of Chicago Press.
Gilens, M. (2003). How the Poor Became Black. Race and the politics of welfare reform. Michigan: The University of Michigan Press.
Iyengar, S. (1991). Is Anyone Responsible? How Television Frames Political Issues. Chicago: University of Chicago Press.

Krosch, A. R., \& Amodio, D. M. (2014). Economic scarcity alters the perception of race. Proceedings of The National Academy of Science, 111(25), 9079-9084.
MacLin, O. H., \& Malpass, R. S. (2001). Racial categorization of faces. The ambiguous race face effect. Psychology, Public Policy, and Law, 7(1), 98-118.
Paolacci, G., Chandler, J., \& Ipeirotis, P. G. (2010) Running experiments on Amazon Mechanical Turk, Judgement and Decision Making 5, 411-419
Payne, B. K. (2006). Weapon bias. Split-second decisions and unintended stereotyping. Current Directions in Psychological Science. 15(6), 287-291.
Peery, D., \& Bodenhausen, G. V. (2008). Black + White = Black: Hypodescent in reflexive categorization of racially ambiguous faces. Psychological Science, 19, 973-977.
Powell, M., \& Ansic, D. (1997). Gender differences in risk behaviour in financial decision-making: An experimental analysis. Journal of Economic Psychology, 18(6), 605-628.
Rodeheffer, C. D., Hill, S. E., \& Lord, C. G. (2012). Does this recession make me look Black? The effect of resource scarcity on the categorization of biracial faces. Psychological Science, 23(12), 1476-1478.
Rudman, L. A., \& Phelan, J. E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. Social Psychology, 41(3), 192.
Schubert, R., Brown, M., Gysler, M., \& Brachinger, H. W. (1999). Financial decision-making: are women really more risk-averse? American Economic Review, 89(2) 381-385.
Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter. An analysis of the impact of question wording on response patterns. Public Opinion Quarterly, 51(1), 75-83.
Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. American Psychologist, 52(6), 613-629.
Steele, C. M., \& Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69(5), 797-811.
Steen, L. A. (1987). Mathematics education: A predictor of scientific competitiveness. Science, 237(4812), 251-302.
Walton, G. M., Cohen, G. L. (2003). Stereotype lift. Journal of Experimental Social Psychology, 39(5), 456-467.

# Perception is in the Details: A Predictive Coding Account of the Psychedelic Phenomenon 

Sarit Pink-Hashkes (S.PinkHashkes@student.ru.nl)<br>Radboud University, Donders Institute for Brain, Cognition, and Behaviour, PO Box 9104, 6500 HE Nijmegen, The Netherlands<br>Iris van Rooij (i.vanrooij@donders.ru.nl)<br>Radboud University, Donders Institute for Brain, Cognition, and Behaviour, PO Box 9104, 6500 HE Nijmegen, The Netherlands

Johan Kwisthout (j.kwisthout@donders.ru.nl)
Radboud University, Donders Institute for Brain, Cognition, and Behaviour, PO Box 9104, 6500 HE Nijmegen, The Netherlands


#### Abstract

Psychedelic substances are used for clinical applications (e.g., treatment of addictions, anxiety and depression) as well as an investigative tool in neuroscientific research. Recently it has been proposed that the psychedelic phenomenon stems from the brain reaching an increased entropic state. In this paper, we use the predictive coding framework to formalize the idea of an entropic brain. We propose that the increased entropic state is created when top-down predictions in affected brain areas break up and decompose into many more overly detailed predictions due to hyper activation of $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors in layer V pyramidal neurons. We demonstrate that this novel, unified theoretical account can explain the various and sometimes contradictory effects of psychedelics such as hallucination, heightened sensory input, synesthesia, increased trait of openness, 'ego death' and time dilation by up-regulation of a variety of mechanisms the brain can use to minimize prediction under the constraint of decomposed prediction.


Keywords: predictive coding; psychedelics; level of detail; Bayesian networks, Lysergic acid diethylamide, Psilocybin.

## Introduction

A recent review paper (Nichols, 2016) examines both the current scientific knowledge regarding psychedelics as well as the many positive results in clinical experiments using psychedelics to treat depression and addiction. The brain, under the influence of psychedelics, has been described as 'being in more states than usual' (CarhartHarris et al., 2014), based on an increased activity in a number of specific brain networks such as the default mode network. They suggested that this higher variance of activity allows for enhancement of the repertoire of possible states over time, and introduced the term Entropic Brain to describe this higher entropic state. On a more implementational level, the current consensus is that psychedelics cause their effects by being (partial) agonists of serotonin, i.e., 5 -hydroxytryptamine $2 \mathrm{~A}\left(5-\mathrm{HT}_{2 \mathrm{~A}}\right)$ receptors, with particular importance to those expressed on apical dendrites of neocortical pyramidal cells in layer V . The $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors are excitatory receptors, making the neurons more likely to fire.

In this paper, we combine both these computationallevel and implementational-level insights into a predictive coding account of the effect of psychedelics. We unify notions proposed by Kwisthout, Bekkering, \& van Rooij (2017) regarding the importance of the amount of details or granularity of predictions, and Bastos et al.'s (2012) canonical microcircuits for predictive coding. We propose that the increased entropic state is created when top-down predictions in affected brain areas break up and decompose into many more overly detailed predictions due to hyper activation of $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors in layer V pyramidal neurons. We demonstrate that this novel, unified theoretical account can explain the various and sometimes contradictory cognitive effects of psychedelics such as hallucination, heightened sensory input, synesthesia, increased trait of openness, 'ego death' and time dilation by up-regulation of a variety of mechanisms the brain can use to minimize prediction under the constraint of decomposed predictions.

In the next section we will introduce the main ideas of the predictive coding account. We will then formulize the Entropic Brain hypothesis into a predictive coding account of the psychedelic phenomenon. In the second part of this paper we will show how this formalization can explain the various and sometimes contradictory cognitive effects of psychedelics.

## A Predictive Coding Primer

In his book "The Doors of Perception" (1954), Aldous Huxley described some of his psychedelic experiences, which led him to propose the idea that perception is a door between things that are known and things that are unknown. This idea turned out prescient of the contemporary predictive coding account of brain processing. According to predictive coding, perception is a continuous process of combining the brain’s previous knowledge with new incoming data by using Bayesian updating, so as to best represent the environmental causes of its sensory input. This enables the brain to predict its sensory inputs. Furthermore, the brain is thought to create a hierarchically ordered model (Friston, 2008). For any
pair of levels, the higher-level will have contextdependent hypotheses predicting the bottom-up signals from lower-levels. The hypothesis that generates the best predictions will determine perception. Calculating which hypothesis generates the best predictions is done by calculating the posterior probability of the hypothesis. The posterior probability combines both the likelihood of the bottom-up input and the prior probability of the hypothesis before receiving the input. This can be seen as an advantageous tactic especially under conditions of noisy unreliable bottom-up data, since previous knowledge can be used to come up with the best hypothesis. The predictions stemming from the best hypothesis inhibit the bottom up incoming data 'explaining it away' (Clark, 2016).

Recently, Kwisthout and colleagues proposed a computational-level distinction between the precision of a prediction and the amount of details or granularity of predictions (Kwisthout \& van Rooij, 2015, Kwisthout et al., 2017). This work has shown that more detailed predictions cause higher prediction errors. This work is based on the idea that higher cognitive functions are better described by categorical probability distributions rather than the traditional Gaussian densities (Friston et al., 2015). An important distinction between Gaussian densities and categorical probability distributions is that in the latter the state space granularity (how detailed are the generative models and the predictions that follow from them) is crucial. Whereas the amount of uncertainty (or precision) in a Gaussian density can be adequately described by its variance, a categorical distribution needs both the state space granularity and the entropy of the distribution to describe its precision (Kwisthout \& van Rooij, 2015).

Bastos et al. (2012) have suggested a 'canonical microcircuit' that provides an implementational-level account of the predictive coding in the brain. The idea of such a canonical microcircuit is that a cortical column contains the circuitry necessary to implements a form of approximate Bayesian inference and that these circuits are replicated with minor variations throughout the cortex. This Microcircuit model is based on evidence showing that superficial pyramidal cells have forward connections to higher areas in the brain hierarchy while deep layers, including pyramidal cells in layer V of the cortex, send back propagating signals to lower areas. Bastos et al. present evidence showing that these backwards connections are inhibitory and can plausibly be seen as implementing the top-down 'predictions' as suggested by the predictive coding framework, while forward connections are plausible realizations of the signals representing 'prediction error'. They further suggest that superficial layers of cortex show neuronal synchronization and spike-field coherence predominantly in the gamma frequencies, while deep layers prefer lower (alpha or beta) frequencies. In essence, they claim that
the top down predictions are communicated by lower alpha or beta frequencies while prediction error is communicated by faster gamma frequencies.

Muthukumaraswamy et al. (2013) found, following administration of Psilocybin, a desynchronization of neural activity especially in the slower alpha and beta rhythms, meaning neurons were acting in a more disjoint and separate way, suggesting that the brain was at a higher entropic state. Using dynamic causal modelling they found that this desynchronization is "likely triggered by $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptor-mediated excitation of deep pyramidal cells" (Muthukumaraswamy et al., 2013, p. 15171). While synchronization of post synaptic neuronal groups creating brain wave oscillations are thought to be needed for communication between brain areas and passing of information, the actual information is thought to be found in the a sparse coding of neuron spiking as very specific timings compared to the oscillations (Fries, 2015).

## A Predictive Coding Account of the Psychedelic State

As we have seen, the effects of psychedelics stem from the $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors on pyramidal cells in layer V being activated, lowering the threshold of individual neuronal firing and thus desynchronizing the activity of the neuronal population. We discussed above Bastos et al.'s (2012) view that the information communicated by the synchronous activity of these specific cells is likely to represent the brain's top-down predictions. It is known that within the neocortex, $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors are not distributed equally and different areas have different binding potentials. Higher binding potentials can be found in prefrontal and visual areas while the motor cortex has lower binding potentials (Forutan et al., 2002). Our theory focuses on the dense band of $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors in layer V pyramidal cells.

Based on Kwisthout et al.'s (2017) notion of state space granularity in predictions, we suggest that hyperactivation of the cells in layer V decomposes the broad categorical prediction that is usually calculated by this neuronal population into sub categories, creating a set of higher detailed predictions. These decomposed predictions stemming from prefrontal, parietal and somatosensory cortex are sent backwards to lower layers of the cortical hierarchy. The decomposed higher detailed prediction that has the highest posterior probability now dominates perception. However, under most conditions, no matter which of the higher detailed decomposed predictions best fits the data, it will still fit less data than the 'usual' broad prediction. This will cause a higher level of bottom up prediction error. As we shall see in the second part of the paper, the compensatory mechanisms called to deal with this higher level of prediction error can explain the wide variety of psychedelic effects.


Figure 1. The predictive coding account assumes that the brain generates predictions using a cascading hierarchy of generative models, processing only that part of the inputs that was unpredicted. Under normal circumstances one might predict to observe animals or plants, and interpret the inputs in a likewise manner (left panel). We suggest that after administration of psychedelics these predictions become more decomposed, leading to more fine-grained, very specific predictions, each of which has a fairly low probability. This will in general lead to a higher prediction error and unstable predictions (right panel). Figure adapted from Clark (2016).

To clarify further what a decomposed set of predictions means, imagine a person walking in the forest receiving some sensory input (Figure 1). Under regular conditions the set of her predictions might be P (Animals) $=0.4$, $P$ (Plants) $=0.6$. The relatively low entropy of these predictions can be computed to be $H=-\sum p_{i} \log _{2}\left(1 / p_{i}\right)=$ 0.97 bits. This means there is relatively little uncertainty regarding these possible predictions. Now let us imagine this person is under influence of psychedelics. Under this condition her set of predictions will be decomposed, for instance: $\mathrm{P}($ Birds $)=0.2, \mathrm{P}($ Dogs $)=0.1, \mathrm{P}($ Butterfly $)=$ $0.09, \mathrm{P}($ Elf $)=0.01, \mathrm{P}($ Trees $)=0.3, \mathrm{P}($ Grass $)=0.6$, P (Flowers) $=0.1$. As we can see, the main categorical predictions of 'Animals' and 'Plants' break up, each into more detailed sub categories. These decomposed predictions bring about a higher entropic state, $\mathrm{H}=-\sum \mathrm{p}_{\mathrm{i}}$ $\log _{2}\left(1 / p_{i}\right)=2.49$ bit. In most cases this will result in higher prediction error from lower layers as these decomposed predictions 'explain away' less of the prediction error from lower layers than normal. The 'extra' predictions being activated are likely to be dependent on a subject's personal experiences and history. In general we should expect a flattening of the prediction distribution, and well-established prediction categories that contain many subcategories will be affected more than predictions with fewer subcategories.

## The importance of bottom-up data in this process

A known concept in the psychedelic community is "set and setting". The mind's set can be compared to the brain's predictions while setting considers the environmental data. When precise environmental data combines with decomposed higher detailed predictions the result will be a uniquely clear perception. This type of perception is commonly described by users and can be read in Aldus Huxley's (1954) description of the vividness
of Red Hot Poker flowers he perceived while under the influence of psychedelics. However, due to environmental changes and noise, this clear perception is not likely to stay stable over time. The noisier the bottom-up signal, the more the top-down predictions influence perception (Seth, 2014). Under decomposed predictions, lowering precision of sensory data can result in misclassification of the data. The brain's best explanation for the imprecise 'noisy' data might be one of the sub-threshold predictions that got activated. This will result in a 'hallucination'. Psychedelics are known to both obscure and distort perceptual data as well as add clarity and give the sense of enhanced resolution. These two different sides of the psychedelic state are dependent on the precision of the bottom-up data, i.e., the noisiness of the setting. The more noisy the bottom-up data is, the more likely hallucinations will be.

## Prediction error minimization and the psychedelic state

Under normal conditions the brain can decrease prediction error in several ways (Friston et al., 2012; Kwisthout et al., 2017). It could update predictions; lower prediction error by intervening in the world or it may update the causal model that generated the predictions. In this section, we explore how upregulating these mechanisms, in order to deal with the increased prediction error caused by decomposed predictions, can explain many of the documented psychedelic effects. We will investigate the effects of prediction updating, active inference, changing the weight of predictions, and long-term learning effects.

## Updating the predictions

As we have explained, in the case of decomposed predictions, a smaller amount of sensory inputs will be explained by any specific prediction. This will cause
increased prediction errors. One mechanism the brain might use to minimize prediction error is to change the prediction distribution. However, as the predictions remain decomposed no prediction will be enough to explain away the prediction error for long and so once again the distribution will change and perhaps this time the probability of an otherwise unlikely input becomes the leading prediction and affects perception. This constant revising of the probability distribution will lead to a destabilization of perception. Objects, scenes and even abstract thoughts will 'morph' and change at a rapid speed; however, each percept reflects the best possible prediction at that moment. A room might look bigger or smaller or the prediction of the light condition might change causing colors to morph. This can be the cause of individuals reporting a tendency to see "multiple viewpoints" (Sessa, 2008).

Predictions from other layers of the brain hierarchy that were not affected by activation of the $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors can be upregulated by either increasing their relative strength or lowering their level of detail. This will cause the predictions from these layers to enforce their predictions on more of the incoming data. Google's deep neural network 'deep dream' ${ }^{1}$ (originally created for identifying images) illustrates how this might happen. By allowing different layers of the network to strengthen their predictions these networks were able to produce hallucinatory effects. Strengthening predictions of lower layers (that identified lines) created images with amplified lines, while increasing predictions from higher level abstract layers (e.g., identifying buildings) created images with 'imaginary' buildings being imposed on the original picture. Further proof that this is actually happening in the brain can be seen in the work of Bressloff et al. (2001). Their simulated attenuated low-level predictions of the visual system (V1) and found remarkable resemblance with geometrical hallucinations drawn by people on LSD. This shows that increased predictions from V1 are likely to be behind the specific geometrical visual hallucination. Furthermore, Carhart-Harris et al. (2016) found that increased cerebral blood flow (CBF) in the visual cortex as well as a greatly expanded functional connectivity profile in V1 correlated strongly with subjects' ratings of visual hallucinations. It is impossible to know at the moment whether the increase in CBF is due to increased predictions errors, upregulating of predictions, or both.

## Acting on the Environment

Another mechanism of minimizing prediction error is intervening in the world (i.e., acting on the environment) (Brown, Friston, \& Bestmann, 2011). This changes the actual inputs and sets some of the model's parameters and thus decreases uncertainty. Changing the brain's input can happen both in a passive way, for instance by moving one's eyes, or by actively moving objects in the environment. Since $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptors are not as prevalent

[^437]in the primary motor cortex, top-down prediction from that area wouldn't be as affected and this mechanism is likely to remain intact even under the influence of Psychedelics. This can explain why hallucinations seem to grow stronger while sitting still and can help influence harm reduction policies. By creating motor output, for instance while walking or dancing, the mechanism of active inference (in which motor output minimizes proprioceptive prediction error between the expected and actual position of one's limb, bringing the actual position closer to the expected position; see, e.g., Brown et al., 2011) might enable the brain to lower prediction errors stemming from other parts of the brain too.

## Changing Weight of the Prediction Error

While chemical tolerance to Psychedelics drugs should not exist more than a few days after ingestion (Leshner, 2001) many experienced users will admit that the first few experiences feel stronger than later experiences and increased dosage is needed to reach the same state. This might happen as a result of the brain's attempt to minimize prediction error by lowering the weight of the prediction error or attributing this higher prediction error to 'inherent' noise that does not need to be explained. An example of inherent noise that the brain learns to ignore can be seen in a fair coin toss (Kwisthout et al., 2017). Even if you guess the coin will land on 'heads' and then it actually lands on 'tails' no surprise will follow. The brain has learnt that this type of stochastic noise is inherent to a fair coin toss. The same could happen under extended use of psychedelics. The brain could learn that this state is inherently noisier and lower the weight of the prediction error. We can only postulate that this might happen through affecting the dopamine system which has been implicated in precision weighting of prediction error (Friston et al., 2012).

## Long Term Learning Effects

Within the predictive coding framework the model constructed by the brain is considered to be encoded in the network connectivity. Changes in this connectivity will lead to long term learning. While learning effects in humans after administration of $5-\mathrm{HT}_{2 \mathrm{~A}}$ agonists have not directly been studied in the last decades an interesting study in rabbits has found that agonists at the $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptor including LSD enhanced associative learning at doses that produce cognitive effects in humans (Harvey, 2003). Using the predictive coding framework, depression, addiction and obsessive compulsive disorders have been suggested to stem from overly strong and narrow predictions from certain networks that get 'stuck' (Edwards et al., 2012) and aren't updated based on the bottom-up data. Momentarily decomposing these predictions by $5-\mathrm{HT}_{2 \mathrm{~A}}$ agonists, especially with a combination of supportive bottom-up information coming from a therapeutic setting, might lead to long term model updates. This could be the reason behind the success of
recent clinical trials that have used $5-\mathrm{HT}_{2 \mathrm{~A}}$ agonist to treat these disorders.

A long term model update that psychedelic are known to cause is increasing the trait of 'openness' (MacLean, Johnson, \& Griffiths, 2011). The mechanism we suggest to explain this is as follows. A higher prediction error state caused by administration of $5-\mathrm{HT}_{2 \mathrm{~A}}$ agonists coupled with a positive rewarding setting, leads to surprise becoming a more sought after state. Interest in exploring the unknown and trying new things might grow and people might be 'motivated to enlarge their experience into novel territory' which is what defines the trait of openness (DeYoung et al., 2009).

## Psychedelics research findings reinterpreted

In the following section we will re-interpret previous findings in psychedelics research in light of our theoretical account and see how our account can clarify and shed further light on these results.

Kometer et al. (2006) presented so-called Kanizsa triangles to subjects after administration of psilocybin. These shapes are perceived as complete triangles and circles rather than the complex shapes that they actually are, because of a top-down learnt prediction. Viewing this shape under normal conditions has been shown to evoke a unique change lowering of voltage as measured on the skull $170 \mathrm{~ms}^{2}$ after presentation of this stimulus. Following administration of Psilocybin, Kometer et al. found a decrease in strength of this ERP suggesting a lower strength of these predictions. This is in accordance with the model of decomposed predictions, since decomposed predictions will indeed cause each prediction to be weaker than normal. This same experiment also found desynchronization of alpha band activity which we have discussed previously.

In a behavioral experiment, Spitzer et al. (1996) found increased indirect semantic priming after administration of Psilocybin. They claim their data suggests that Psilocybin leads to an "increased availability of remote associations and thereby may bring cognitive contents to mind that under normal circumstances remain nonactivated" (Spitzer et al., p. 1056). This would indeed be expected if broad categorical 'semantic' predictions are decomposed, activating many more detailed semantic predictions, and allowing for more remote associations to be activated.

Another well-documented effect is known as 'Time Dilation' in which subjective time seems to slow down. A few minutes can subjectively be perceived as taking much longer. Here we postulate that subjective sensation of time is dependent on the amount of prediction error and possibly prediction updates the brain makes in order to minimize prediction error. This idea is based on Ulrich (2006) who discovered that the extent to which a stimulus can be predicted affects time perception, with unexpected

[^438]stimuli perceived as longer. Similarly, Tse et al. (2004) found that a stimulus which stands out as different from all the others in a series appears to last longer than the other stimuli. An increase of prediction updates might cause the subjective feeling that more time has passed. This is similar to the common feeling that the first day of a journey to another country seems longer because it is filled with so many new experiences and so many prediction updates must happen in that day.

The last phenomenon we would like to touch upon is the notion of 'Ego death' many psychedelic users report. Apps \& Tsakiris (2013) describe a predictive coding account of the neural and computational basis of selfrecognition. Here, one's body is recognized as the most likely "me". This probabilistic inference arises through the integration of information from hierarchically organized unimodal systems in higher-level multimodal areas. As we have seen, the brain's attempt to minimize increased prediction error induced by psychedelics breaks down this hierarchical structure which might lead to a total inability to distinguish between environment and self and the unique perception of 'oneness' described by many experiencing 'ego loss'. While Apps \& Tsakiris' account deals with the 'minimal self', we postulate looking at the 'higher ego' as a collection of high-level relatively inflexible predictions regarding the future behaviour of the 'self-organism' in a variety of situations. Following administration of $5-\mathrm{HT}_{2 \mathrm{~A}}$ agonists these predictions will break up based on the subjective pieces of information compromising this category. This relaxation of otherwise rigid predictions about the self might explain positive results for treatment of depression and addiction after administration of psychedelics that have been reported (Nichols, 2016).

## Conclusions

In this paper we presented a computational theory explaining the effects of psychedelics in terms of the predictive coding account of cortical processes. Our theory further explicates the Entropic Brain hypothesis (Carhart-Harris et al., 2014) in terms of predictive coding. We proposed that administration of psychedelics cause the brain to make overly detailed (i.e., decomposed) predictions of the inputs it receives, leading to an increased prediction error. Crucially, while dopamine is considered to modulate precision weighting of prediction errors (Friston et al., 2012), our theory suggests that serotonin might have a role in modulating the granularity ("level of detail") of predictions. Our theory explains how a simple lowering of the excitation threshold of the pyramidal neurons in layer V in prefrontal, parietal and somatosensory cortex (caused by administration of 5$\mathrm{HT}_{2 \mathrm{~A}}$ agonists) in fact decomposes predictions from those areas, causing increased prediction errors from lower levels in the brain hierarchy. The brain's attempts to minimize these increased prediction errors by active inference, prediction updating, modulation of the weights
of prediction errors, or model revision can explain several (and sometimes contradictory) cognitive effects of psychedelics such as hallucination, heightened sensory input, synesthesia, increased trait of openness, 'ego death' and time dilation.

## References

Apps, M.A.J., \& Tsakiris, M. (2013). The free-energy self: A predictive coding account of self-recognition. Neuroscience and Biobehavioral Reviews, 41, 85-97.
Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., \& Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron, 76(4), 695-711.
Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J.R., ... \& Fries, P. (2014). Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron, 85(2), 390-401.
Bressloff, P.C., Cowan, J.D., Golubitsky, M., Thomas, P.J., \& Wiener, M.C. (2001). Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. Philosophical Transactions of the Royal Society B, 356(1407), 299-330.
Brown, H., Friston, K.J., \& Bestmann, S. (2011). Active inference, attention, and motor preparation. Frontiers in Psychology, 2(SEP), 1-10.
Carhart-Harris, R.L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., ... \& Nutt, D.J. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. Frontiers in Human Neuroscience, 8(FEB), 20.
Carhart-Harris, R.L., Muthukumaraswamy, S., Roseman, L., Kaelen, M., Droog, W., \& Nutt, D.J. (2016). Neural correlates of the LSD experience revealed by multimodal neuroimaging. Proceedings of the National Academy of Sciences, 113(17), 4853-4858.
Clark, A. (2016). Surfing Uncertainty. Oxford, UK: Oxford University Press.
DeYoung, C.G., Shamosh, N.A., Green, A.E., Braver, T.S., \& Gray, J.R. (2009). Intellect as distinct from Openness: Differences revealed by fMRI of working memory. Journal of Personality and Social Psychology, 97, 883-892.
Edwards, M.J., Adams, R.A., Brown, H., Pareés, I., \& Friston, K.J. (2012). A Bayesian account of 'hysteria'. Brain, 135(11), 3495-3512.
Fries, P. (2015). Rhythms for cognition: Communication through coherence. Neuron, 88(1), 220-235.
Friston, K.J. (2008). Hierarchical models in the brain. PLoS Computational Biology, 4(11), e1000211.
Friston, K.J., Shiner, T., FitzGerald, T., Galea, J.M., Adams, R., Brown, H., ... \& Bestmann, S. (2012). Dopamine, affordance and active inference. PLoS Computational Biology, 8(1), e1002327.

Friston, K.J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., \& Pezzulo, G. (2015). Active inference and epistemic value. Cognitive Neuroscience, 13, 1-28.
Forutan, F., Estalji, S., Beu, M., Nikolaus, S., Hamacher, K., Coenen, H.H., ... \& Larisch, R. (2002). Distribution of $5 \mathrm{HT}_{2 \mathrm{~A}}$ receptors in the human brain: Comparison of data in vivo and post mortem. Nuklearmedizin-Nuclear Medicine, 41(4), 197.
Harvey, J.A. (2003). Role of the serotonin $5-\mathrm{HT}_{2 \mathrm{~A}}$ receptor in learning. Learning \& Memory, 10(5), 355362.

Huxley, A. (1954). The Doors of Perception. London: Chatto and Windus.
Kometer, M., Schmidt, A., Jäncke, L., \& Vollenweider, F.X. (2013). Activation of serotonin 2A receptors underlies the psilocybin-induced effects on alphaoscillations, N170 visual-evoked potentials, and visual hallucinations. Journal of Neuroscience, 33(25), 10544-10551.
Kwisthout, J. \& van Rooij, I., (2015). Free energy minimization and information gain: The devil is in the details. Cognitive Neuroscience, 6(4), 216-218.
Kwisthout, J,. Bekkering, H., \& van Rooij, I., (2017). To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. Brain and Cognition, 112, 84-91.
Leshner, A.I. (2001). Hallucinogens and Dissociative drugs. NIDA - Research Report.
MacLean, K.A., Johnson, M.W., \& Griffiths R.R. (2011). Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness. Journal of Psychopharmacology, 25(11), 1453-1461.
Muthukumaraswamy, S.D., Carhart-Harris, R.L., Moran, R.J., Brookes, M.J., Williams, T.M., ... \& Nutt, D.J. (2013). Broadband cortical desynchronization underlies the human psychedelic state. Journal of Neuroscience, 33(38), 15171-15183.
Nichols, D. E. (2016). Psychedelics. Pharmacological Reviews, 68(2), 264-355.
Sessa, B. (2008). Is it time to revisit the role of psychedelic drugs in enhancing human creativity? Journal of Psychopharmacology, 22(8). 821-827.
Seth, A.K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. Cognitive Neuroscience, 5(2), 97-118.
Spitzer, M., Thimm, M., Hermle, L., Holzmann, P., Kovar, A., ... \& Schneider F. (1996): Increased activation of indirect semantic associations under psilocybin. Biological Psychiatry, 39, 1055-1057.
Tse, P.U., Intriligator, J., Rivest, J., \& Cavanagh, P. (2004). Attention and the subjective expansion of time. Perception \& Psychophysics, 66, 1171-1189.
Ulrich, R., Nitschke, J., \& Rammsayer, T. (2006). Perceived duration of expected and unexpected stimuli. Psychological research, 70(2), 77-87.

# Repetition improves memory by strengthening existing traces: Evidence from pairedassociate learning under midazolam 

Vencislav Popov (vencislav.popov@gmail.com)<br>Lynne Reder (reder@cmu.edu)<br>Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA


#### Abstract

Here, we examined how repetition under midazolam, a benzodiazepine that prevents the storage of novel associations, affects cued-recall performance of paired-associates. We contrasted word pairs that were initially studied and tested repeatedly without any successful recall prior to the midazolam injection, with other pairs that were studied for the first time after the injection of midazolam. According to our SAC (Source of Activation Confusion) memory model, repetition leads to strengthening existing memory traces rather than creating multiple traces for each repetition. As such, it predicts that repetition under midazolam should benefit only pairs that were originally studied prior to the midazolam injection. This prediction was confirmed. The results suggest that memory traces for pairs studied prior to the midazolam injection were strengthened under midazolam. However, word pairs that had not been studied prior to the injection were not bound in long-term memory because midazolam prevents the formation of new associations.


Keywords: memory strength; paired associate learning; episodic memory; practice; midazolam;

Repetition improves memory performance across the board. Beneficial effects of repetition have been found on most measures of explicit memory such as single-item and paired-associates recognition (e.g., Challis \& Sidhu, 1993; Reder et al., 2000), free recall (e.g., Challis \& Sidhu, 1993; Underwood, 1969), and cued recall tasks (Meltzer \& Constable, 2005; Reder et al., 2007; Reder, Liu, Keinath, \& Popov, 2015). However, despite more than a century of research on repetition effects, there is no consensus about the mechanism through which practice affects memory (Criss \& Koop, 2015; Hintzman, 2010, 2011; Osth \& Dennis, 2015; Pavlik \& Anderson, 2005).

Two major types of theories have been proposed to explain repetition effects. Cumulative-strength models (CSMs) suggest that memory traces differ in strength or familiarity, and this strength increases with repetition and decays with time (e.g., Murdock, Smith, \& Bai, 2001; Pavlik \& Anderson, 2005; Reder et al., 2000; Wickelgren, 1972). In these models, recognition and recall are a function of strength and greater strength leads to better memory performance. In contrast, multiple-trace models (MTMs), usually equated with global matching models (GMMs, Criss \& Koop, 2015), state that each repetition of an item is encoded separately in memory (Bower, 1967; Brown, Neath, \& Chater, 2007; Hintzman, 1984; Lansdale \& Baguley, 2008; Osth \& Dennis, 2015). This leads to redundant memory traces, each of which has some probability of being retrieved during test. Interestingly, while both CSMs and MTMs co-exist in the current literature, several proponents of each class believe that certain
empirical findings have conclusively ruled-out the alternative models (Criss \& Koop, 2015; Hintzman, 2011).

When it comes to CSMs, some researchers have argued that they are incompatible with findings from judgments of frequency (JOF) and judgments of recency (JOR) tasks (Flexser \& Bower, 1974; Hintzman, 2010, 2011 Hintzman \& Block, 1971). Many CSMs in the past have assumed that the estimation of frequency, recency and duration of events is based on a single strength dimension (Hintzman, 2011). As a result, these models predict that, for example, if an event is repeated it should also appear to be more recent and to have lasted longer. That is not the case studies have shown that participants can easily discriminate the frequency, recency and duration of repeatedly studied items (Flexser \& Bower, 1974; Hintzman, 2010, 2010; Hintzman \& Block, 1971).

This line of work seems to provide strong evidence against CSMs, and yet, they are still popular in modeling recognition and recall. Hintzman (2011) refers to this as "the fallacy of cumulative strength", and suggests that the CSMs are still popular because most theorists focus on recognition memory and recall, while ignoring tasks such as JOR and JOF. However, the same criticism can be directed at conclusions from JOR and JOF tasks - the fact that a single strength dimension cannot explain behavioral patterns in such tasks does not mean that repetition effects on recognition and recall memory are not due in part to cumulative strengthening of existing memory traces. It only indicates that memory representations also include rich contextual information, which is an assumption shared by most current dual-process CSMs.

Similarly, when it comes to MTMs and GMMs, other researchers maintain that they cannot account for the divergent patterns of the list-length effect (LLE) and the list-strength effect (LSE) on free recall and recognition (Criss \& Koop, 2015; Shiffrin, Ratcliff, \& Clark, 1990). The LLE shows that increasing the number of different items on a study list decreases free recall, cued recall and recognition performance. Similarly, the LSE shows that increasing the number of repetitions on some items leads to worse free recall for the non-repeated items. However, the LSE generally has no effects on overall performance in recognition tasks. Critics of GMMs have argued that they are fundamentally incompatible with this pattern of results (Criss \& Koop, 2015). This is because GMMs assume that the same mechanism is involved when the number of different items increases and when some items on a list are repeated. Specifically, they both lead to the creation of additional memory traces and to increased global signal variance, which causes interference during retrieval. As such, GMMs supposedly predict that LSE and LLE should always occur together.

However, multiple proponents of GMMs have questioned both the reliability of the pattern of LSE and LLE effects on recognition (Dennis \& Humphreys, 2001; Murdock \& Kahana, 1993) and the inability of GMMs to account for it (Murdock \& Kahana, 1993; Osth \& Dennis, 2014, 2015). Furthermore, despite the fact that most MTMs are also GMMs, a multiple-trace model does not have to depend on global matching for memory decisions, which would make it easier to fit the pattern of LSE and LLE results. For example, even though our Source of Activation Confusion model (SAC; Reder et al., 2000) is a cumulative-strength model, the dual processes that allow it to account for the divergent pattern of LLE and LSE results would allow it to do the same even if each item repetition created a novel episodic trace (Cary \& Reder, 2003; Diana \& Reder, 2005).

In summary, the major problem with contrasting CSMs and MTMs has been that they make similar predictions when it comes to most memory tasks. One way to overcome this would be to attempt to disrupt the mechanism that is responsible for repetition effects in a specific model such as SAC, to make predictions how that will affect behavioral performance and to evaluate how well the model fits the data. To achieve that, we examined how repetition affects cued-recall under midazolam. Midazolam is a benzodiazepine that creates temporary anterograde amnesia by preventing the storage of new associations in LTM (Ghoneim, 2004; Reder et al., 2006), but it does not impair pre-existing memory traces (Ghoneim, 2004) or their strengthening, as evidenced by its limited effect on repetition priming (Hirshman, Passannante, \& Arndt, 2001; Hirshman, Passannante, \& Henzler, 1999).

We compared cued-recall performance for paired-associates that were studied for the first time (control pairs) under midazolam and a subset of the pairs that were studied both before the midazolam injection and re-studied after the injection (practice pairs). The subset of interest were those pairs that had not been recalled on any of the tests that preceded the re-study session under midazolam. Given that midazolam prevents the storage of new associations in LTM, SAC predicts that the recall of control pairs should be at floor levels; performance on practice pairs that were never recalled correctly was an open question. SAC assumes that pairs that had never been correctly recalled might still have subthreshold episodic traces in LTM. If repetition of the pair leads to the strengthening of this sub-threshold episodic trace, as SAC originally assumes, then we would expect greater recall of practice pairs compared to control pairs. If, however, repetition leads only to the creation of additional memory traces, then no advantage should be observed for practice compared to control pairs under midazolam, because midazolam will prevent the storage of the new traces in LTM.

## Method

The data of interest involve a subset of conditions from a larger study previously reported in Reder et al. (2007) and Reder et al. (2006; study 2). For clarity, we will describe the full design.

## Participants

Thirty-one healthy individuals from the Pittsburgh community participated in this experiment. Each participant was screened by a doctor and received $\$ 150$ upon completion.

## Procedure, materials and design

The study took place in two sessions on two separate days. We used a within-subject double-blind cross-over design where the drug condition (saline vs midazolam) was randomly assigned to one of the two days for each participant. Each session consisted of three separate study-test list cycles. The saline/midazolam injection was administered over a 2-min period between Lists 1 and 2. Participants began the study phase of List 2 immediately after the injection.

During each list, participants saw all of the 45 high-frequency word pairs in the following sequence: Study - Test1 and Restudy1 - Test2 and Restudy2. During the initial study phase, each word pair was presented for 3 seconds preceded by a fixation cross for 1 second. After all 45 pairs were studied, participants completed a self-paced cued-recall test for all 45 pairs in a different random presentation order. Test trials began with a 500 ms fixation cross, followed by the presentation of the first word in a pair and a question mark prompting participants to respond. Participants were asked to recall the correct word and type it on a laptop keyboard or press the return key to move to the next trial. Regardless of the accuracy of their response, participants saw the correct answer for 2.5 seconds after each response, which gave them an opportunity to restudy the pairs again. When all 45 pairs were tested and restudied the test-restudy phase was repeated one more time, which concluded the procedure for the first list. This study-testandrestudy1-testand-restudy2 procedure was repeated for two more lists, each of which took approximately 17 min . to complete. On each list, the 45 study pairs were split into 3 conditions with 15 pairs per condition - control pairs, which were unique for each list, practice pairs, which were the same 15 pairs on all 3 lists, and interference pairs, which had the same words on all 3 lists, but the cue words were assigned to different response words on each new list. The order of word pairs in each list, study and test sessions was randomly determined.

## Data analysis and logic for the current study

Only a small subset of conditions was relevant for this study (see Figure 1). Specifically, we looked at cued recall performance on List 2 Test 2 for those control and practice pairs for which participants had failed to recall the response word on all previous tests. The control pairs were unique to each list, and as such the ones we selected were previously studied and tested only once at the beginning of List 2 . The practice pairs were previously studied and tested twice on List 1 as well. We analyzed only those practice pairs which participants failed to recall on all three occasions (L1T1, L1T2 and L2T1).

We focused on the second test of List 2 (L2T2), rather than on the first test on List 2 for the following reason. Even though the injection was administered before the beginning of the second list, practice pairs were restudied immediately after their second test on List 1 (L1T2). Thus, if we observed improved recall for practice pairs on L2T1, it might have been due to the restudy


Figure 1. Visualization of the procedure for the subset of conditions relevant for this study. Small boxes represent the phases of each list (study, cued recall, or restudy), and whether the test was successful. Shown are the subsequent phases only for the relevant trials. The first row shows the procedure for practice pairs; the second shows the procedure for the control pairs. List $1 \& 2$ are shown in separate columns.
session that occurred prior to the injection rather than due to strengthening during the study session of List 2 under midazolam.

In summary, we selected control and practice pairs that showed no evidence of being learned up until L2T1 in either drug condition (midazolam or saline). These pairs were then restudied immediately following that test, and were then tested on L2T2. For control pairs, $75 \%$ qualified for analysis in the saline condition and $99 \%$ in the midazolam condition. For practice pairs, $32 \%$ qualified under saline and $41 \%$ under midaz. We analyzed accuracy on L2T2 as function of pair type (control vs practice) and drug condition (midazolam vs saline) using a logistic mixed effects regression with participants and items as random intercept effects. We compared alternative models with and without each of the main effects and interactions. A Bonferroni correction was applied to all post-hoc tests ( $\mathrm{n}=4$ ).

## Results and discussion

The overall recall for each list collapsed over the two tests in a list are presented in Figure 2 and the results for the subset of trials of interest are presented in Figure 3. Control pairs were recalled less accurately than practice pairs, $\Delta \mathrm{AIC}=-22, \chi^{2}(1)=24.12, \mathrm{p}<$ .001. Word pairs were recalled less accurately under midazolam compared to saline, $\Delta \mathrm{AIC}=-172, \chi^{2}(1)=174.32, \mathrm{p}<.001$. There was a significant interaction between drug condition and type of pair, $\Delta \mathrm{AIC}=-17, \chi^{2}(1)=19.27, \mathrm{p}<.001$. Post-hoc comparisons revealed that practice pairs were recalled significantly more accurately than control pairs in the midazolam condition ( $z=6.72$, $\mathrm{p}<.001$ ), but not in the saline condition $(\mathrm{z}=2.30, \mathrm{p}=.09)$. Finally, both practice and control pairs were recalled more accurately in the saline compared to the midazolam condition ( $z=9.97, p<.001$ and $\mathrm{z}=4.56, \mathrm{p}<.001$, respectively for practice and control pairs). These results are consistent with the view that repetition can strengthen existing memory traces, because only pairs that were initially studied prior to a midazolam injection benefited from additional study under midazolam. These practice pairs were recalled more often than control pairs, which were studied for the first time after the midazolam injection, even though both showed no evidence of learning prior to the final test.

Despite the fact that practice pairs had not been recalled on any of the 3 previous tests, it is possible that an initial association for them was stored during List 1 . It seems reasonable to conclude that even though these associations were inaccessible, they must have been registered in LTM since they were strengthened under midazolam while the control pairs were not. Midazolam is known to block the formation of new associations but there is no evidence that it inhibits strengthening of existing traces. On the contrary, implicit memory is spared under midazolam (Hirshman et al., 2001), and the model presented below already assumes that implicit memory is based on strengthening the same representations involved in familiarity-based recognition (Reder, Park \& Kieffaber, 2009).

An alternative explanation of these data consistent with multiple-trace theories might be that even if no traces were strengthened under midazolam for practice pairs, the pre-existing sub-threshold traces might be spontaneously recovered in a probabilistic way (Brown, Neath \& Chater, 2007). We believe this is unlikely, given that each practice pair analyzed here failed to be recalled on all 3 previous tests. Additionally, we can directly estimate what is the probability of recovery with a multinomial processing tree model (Erdfelder et al., 2009), where at each test there is one of the following possibilities: 1) successful recall of the target due to study/restudy, $r=\mathrm{P}$ (success on test $\mathrm{n} \mid$ fail on test $\mathrm{n}-1$, or when $\mathrm{n}=1$ ), 2) failing to recall a previously recalled target, $f=\mathrm{P}($ failed recall on test $\mathrm{n} \mid$ success on test $\mathrm{n}-1), 3$ ) spontaneous recall of a previously unrecalled target regardless of restudy benefit, $u=\mathrm{P}^{\prime}($ success on test $\mathrm{n} \mid$ fail on test $\mathrm{n}-1)$. We estimated these probabilities from performance on the control pairs on the two tests on List 1 (data not used in the previous analyses):

$$
\begin{array}{lll}
P(\text { success T1 \& success T2) } & =r *(1-f) & =0.42 \\
P(\text { success T1 \& fail T2) } & =r * f & =0.04 \\
P(\text { fail T1 \& success T2) } & =(1-r) *(r+u) & =0.29 \\
P(\text { fai T1 \& fail T2) } & =(1-r) *(1-r-u) & =0.25
\end{array}
$$

Here, successful recall following a failed recall is a combination of reencoding benefit $r$ and a spontaneous recovery $u$. This analysis showed that the probability of successful recall due to (re)encoding was $r=0.46$, the probability of forgetting a previous encoding was $f=0.09$, and the probability of spontaneous
recovery of a forgotten encoding was $u=0.07$ (similar values were obtained if we consider all sequential pairs tests on all three lists for practice pairs in the saline condition, $u=0.089$ ). One-tailed $t$ tests showed that the benefit of restudy under midazolam (Fig 3), was significantly higher for practice pairs, but not for control pairs. Thus, spontaneous recovery of previously forgotten items cannot account for our results.


Figure 2. Cued-recall performance during acquisition, collapsed over the two tests (black points and solid lines) and model fits (white points and dashed lines) in all lists for control and practice pairs as a function of drug condition.


Figure 3. Accuracy on L2T2 for control and practice pairs that were not recalled correctly on any of the previous tests as a function of whether the L2 items were studied under saline or under midazolam. Horizontal red line shows the probability of spontaneous recovery of a forgotten item without strengthening.

To demonstrate that CSMs can fit the data not only verbally, but quantitatively as well, we fit a SAC model (Reder et al., 2000) on a trial-by-trial basis separately for each participant. In general, SAC posits that semantic, episodic and contextual information is represented as a network of interconnected concepts, event and context nodes varying in strength. Each node has an activation value that increases when a node is perceived or when it receives activation from other nodes. This activation decays with time according to a power law to a base-level resting activation that also is strengthened or decays with experience. When new information is studied, two processes occur. First, the current and the resting level activation values of the corresponding preexisting concept nodes are increased. Second, if this is the first occurrence of the
study episode, a new event node is created and it gets associated with the corresponding concept nodes, as well as with the general and specific context nodes. If, however, the study event has occurred previously, the existing event node and its links with the concept and context nodes are strengthened instead.

Retrieval in SAC is based on the activation of the event and concept nodes and the process differs slightly between free recall, cued recall and recognition. During free recall, the general context node and the list node are activated and they spread activation to all episode nodes connected to them. During cued-recall or recognition, the concept node(s) for the cue(s) is also activated and it spreads activation to all episode nodes connected to it. Spreading activation is multiplied by the strength of each association, and divided by the sum total strength of associative links emanating from the sending node. This represents competition for retrieval. Finally, if an episode node's activation passes the retrieval threshold, an item is recalled (free and cued recall) or recollected (recognition). For recognition, if no episode node passes the threshold, the strength of the cue concept node is evaluated. If it passes its retrieval threshold, a familiarity-based response is made.

The majority of parameters in the model were imported from previous studies. Consistent with the fact that midazolam prevents the storage of novel associations in LTM, in the current simulation we manipulated the probability of encoding an episode node. During the first presentation of each word pair there is a certain probability that participants will fail to encode the event node due to inattention, fatigue or insufficient working memory (see Reder et al., 2007). In models of other studies, this value has been constant, but in the current implementation, we allowed it to vary between the saline and the midazolam conditions. The optimal value for the saline condition was estimated from the data ( $p=$ 0.35 ), while the encoding probability for the midazolam condition changed with time elapsed since the injection (see Table 1 for parameter estimates and descriptions, and Table 2 for full model specification). Immediately after the midazolam injection the encoding probability was 0 , reflecting the inability to store new associations at maximum potency, and it gradually increased to half of the encoding probability in the saline condition in 31 minutes (drug halflife for memorial effects, Albrecht et al., 1999).

The overall model fit for all conditions is presented in Figure 2, and the fit for the specific subset of interest is overlaid on Figure 3. Importantly, the model was fit by predicting a single value for each participant - their overall cued-recall performance and by minimizing the RMSE between the predicted and the observed value. Given that the model had no information about the performance in each condition, we obtained a surprisingly good fit for the split by conditions ( 16 summary data points per participant; $\mathrm{RMSE}=0.139, \mathrm{R}^{2}=0.8$ ). The model demonstrates that the beneficial effect of repetition under midazolam can be explained entirely by the strengthening of preexisting memory traces that were previously below the retrieval threshold.

One could question why practice pairs were not recalled better than control pairs in the saline condition, given that they should benefit both from strengthening a pre-existing trace as well as from creating novel associations for pairs that were previously unlearned, while control pairs benefit only from forming new associations. Indeed, while the overall fit of the model was quite
good, the model predicts that there should be a repetition advantage for the subset of analyzed practice pairs even in the saline condition (Figure 3). The behavioral data showed a small effect in that direction, which was not significant after correcting for multiple comparisons $(\mathrm{p}=.09)$. One possibility is that this is due to a selection bias - the practice pairs selected for analysis were those that showed no evidence of learning in three previous tests ( $\sim 32 \%$ ), thus they were generally hard to learn. Note that the control pairs we analyzed were those that had not been recalled on only 1 previous test ( $\sim 76 \%$ ) so they were probably not as difficult to learn. The greater difficulty of the selected practice pairs might have offset the relative repetition benefits under saline (as seen from Fig 2, practice pairs do benefit more under saline than under midazolam). Another possibility is that there were only few observations per cell, and the resulting noise might have obscured the effect in the saline condition. Despite this, the key prediction, namely the comparison between control and practice pairs under midazolam, was quite robust.

In summary, the current study provides evidence that one mechanism through which repetition benefits memory is the strengthening existing memory traces. Despite this result, we do not wish to argue that no additional information beyond strength is stored in memory with each repetition of an event. Based on JOR, JOF, LLE and LSE results reviewed in the introduction, and the results presented here, it is reasonable to conclude that repeated experiences affect memory through a multitude of mechanisms that include both strengthening of previously encoded traces that match in content, as well as storing novel traces to represent the unique features of the repeated experience. What part of that information is accessed likely depends on the nature of the task being performed. While accurate judgments of recency and frequency might require accessing and comparing information across multiple memory traces, recognition and recall can depend on the strength of any one of those traces.

Table 1 SAC model parameters

| Par | Description | Value |
| :---: | :---: | :---: |
| Imported parameters |  |  |
| $A_{\text {boost }}$ | Value added to current activation when an item is perceived | 40 |
| $p_{\text {decay }}$ | Exponential decay constant for current activation | 0.8 |
| $d_{\text {node }}$ | Power-law decay constant for base-level activation | 0.175 |
| $c_{\text {node }}$ | Power-law growth constant for base-level activation | 25 |
| $d_{\text {link }}$ | Power-law decay constant for link strength | 0.12 |
| Clink | Power-law growth constant for link strength | 25 |
| $b_{\text {freq }}$ | Exponent for Kucera and Francis word frequency norms for estimating pre-existing base-level activation | $1 \quad 0.4$ |
| $l_{\text {freq }}$ | Exponent for Kucera and Francis word frequency norms for estimating preexisting link fan | 0.7 |
| Estimated parameters |  |  |
| $P_{\text {baseline }}$ | Baseline probability of encoding an event node for a new word pair | 0.35 |
| $\sigma_{\text {episode }}$ | Standard deviation of the episode node activation | 1 |
| $T_{\text {episode }}$ | Retrieval threshold for episode node activation | $\begin{gathered} 2.9 \pm \\ 1.6^{*} \end{gathered}$ |

[^439]Table 2 SAC model equation

| Equation | Description |
| :---: | :---: |
| $B_{0}=K F^{b_{\text {freq }}}$ | Preexisting base-level activation; a function of Kurcera \& Francis word frequency |
| $B=B_{0}+c_{\text {node }} \sum t_{i}^{-d_{\text {node }}}$ | Current base-level activation is a function of preexisting base-level activation and time since each presentation of a stimulus. $t_{i}$ is the time since the $i$-th presentation |
| $S_{\text {cue,episode }}$ $=c_{l i n k} \sum t_{i}^{-d_{l i n k}}$ | Current strength of the link from the cue to the episode node is a function of time since each presentation of the stimulus. $t_{i}$ is the time since the $i$-th presentation |
| $A_{\text {cue }}=B+A_{\text {boost }}$ | Current activation of the cue is a function of base-level activation and a perceptual boost |
| $A_{\text {input }}=A_{\text {cue }} \frac{S_{\text {cue } e \text { episode }}}{\sum S_{\text {cue }}}$ | The input to an episode node due to spreading activation from the cue is a function of the cue activation level, the strength between the cue and the episode node, and the fan of the cue |
| $A_{\text {episode }}=\ln \left(B+A_{\text {input }}\right)$ | Current activation of the episode node is the natural logarithm of the sum of the base-level activation and the received spreading activation |
| $\begin{aligned} & P_{\text {encoding }} \\ & =P_{\text {baseline }} *(1 \\ & \left.-C \times 2^{-\frac{t_{\text {injection }}}{t_{h l}}}\right) \end{aligned}$ | The probability of encoding the episode node is a function of the baseline probability, whether a the drug was saline $(\mathrm{C}=0)$ or midazolam ( $\mathrm{C}=1$ ), the time since the injection and the half-life of the drug. |
| $\begin{aligned} & P_{\text {retrieval }} \\ & =N\left(A_{\text {episode }} \mid \sigma_{\text {episode }},\right. \\ & \left.\quad \tau_{\text {episode }}\right) \end{aligned}$ | The probability of retrieval of the episode node is the area to the left of the activation value under a standard normal distribution with the threshold as the mean. |

## Acknowledgements

This work was supported by Grants 2-R01-MH52808 and R01MH068243 from the National Institute of Mental Health. VP was supported by a CMU Presidential Fellowship endowed by the R.K.Mellon Foundation and the Hillman Foundation. We are thankful to Xiaonan Liu for comments concerning the selection of trials for the analysis presented herein.

## References

Albrecht, S., Ihmsen, H., Hering, W., Geisslinger, G., Dingemanse, J., Schwilden, H., \& Schüttler, J. (1999). The effect of age on the pharmacokinetics and pharmacodynamics of midazolam. Clinical Pharmacology \& Therapeutics, 65(6), 630-639.
Benjamin, A. S., \& Tullis, J. (2010). What makes distributed practice effective? Cognitive Psychology, 61(3), 228-247.
Bower, G. (1967). A multicomponent theory of the memory trace. Psychology of Learning and Motivation, 1, 229-325.

Brown, G. D., Neath, I., \& Chater, N. (2007). A temporal ratio model of memory. Psychological Review, 114(3), 539.
Cary, M., \& Reder, L. (2003). A dual-process account of the listlength and strength-based mirror effects in recognition. Journal of Memory and Language, 49(2), 231-248.
Challis, B. H., \& Sidhu, R. (1993). Dissociative effect of massed repetition on implicit and explicit measures of memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(1), 115-127.
Criss, A. H., \& Koop, G. J. (2015). Differentiation in episodic memory. Psychology Press Festschrifts. Cognitive Modeling in Perception and Memory: A Festschrift for Richard M. Shiffrin, 112-125.
Dennis, S., \& Humphreys, M. S. (2001). A context noise model of episodic word recognition. Psychological Review, 108(2), 452.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Assfalg, A., Moshagen, M., \& Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. Zeitschrift Für Psychologie/Journal of Psychology, 217(3), 108-124.
Flexser, A. J., \& Bower, G. H. (1974). How frequency affects recency judgments: A model for recency discrimination. Journal of Experimental Psychology, 103(4), 706.
Ghoneim, M. M. (2004). Drugs and Human Memory (Part 1) Clinical, Theoretical, and Methodologic Issues. The Journal of the American Society of Anesthesiologists, 100(4), 987-1002.
Hintzman, D. (1970). Effects of repetition and exposure duration on memory. Journal of Experimental Psychology, 83(3, Pt.1), 435-444.
Hintzman, D. (1984). MINERVA 2: A simulation model of human memory. Behavior Research Methods, Instruments, \&amp Computers, 16(2), 96-101.
Hintzman, D. (2010). How does repetition affect memory? Evidence from judgments of recency. Memory \& Cognition, 38(1), 102-115.
Hintzman, D. (2011). Research Strategy in the Study of Memory: Fads, Fallacies, and the Search for the "Coordinates of Truth." Perspectives on Psychological Science, 6(3), 253-271.
Hintzman, D., \& Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. Journal of Experimental Psychology, 88(3), 297-306.
Hirshman, E., Passannante, A., \& Arndt, J. (2001). Midazolam amnesia and conceptual processing in implicit memory. Journal of Experimental Psychology: General, 130(3), 453.
Lansdale, M., \& Baguley, T. (2008). Dilution as a model of longterm forgetting. Psychological Review, 115(4), 864.
Meltzer, J. A., \& Constable, R. T. (2005). Activation of human hippocampal formation reflects success in both encoding and cued recall of paired associates. NeuroImage, 24(2), 384-397.
Morton, J. (1968). Repeated items and decay in memory. Psychonomic Science, 10(6), 219-220.
Murdock, B. B., \& Kahana, M. J. (1993). Analysis of the liststrength effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(3), 689-697.
Murdock, B., Smith, D., \& Bai, J. (2001). Judgments of Frequency and Recency in a Distributed Memory Model. Journal of Mathematical Psychology, 45(4), 564-602.

Nadel, L., Samsonovich, A., Ryan, L., \& Moscovitch, M. (2000). Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. Hippocampus, 10(4), 352-368.
Osth, A. F., \& Dennis, S. (2014). Associative recognition and the list strength paradigm. Memory \& Cognition, 42(4), 583-594.
Osth, A. F., \& Dennis, S. (2015). Sources of Interference in Item and Associative Recognition Memory. Psychological Review.
Pavlik, P. I., \& Anderson, J. R. (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. Cognitive Science, 29(4), 559-586.
Peterson, L. R., Johnson, S. T., \& Coatney, R. (1969). The effect of repeated occurrences on judgments of recency. Journal of Verbal Learning and Verbal Behavior, 8(5), 591-596.
Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. Cognitive Science, 27(3), 431-452.
Reder, L. (1988). Strategic control of retrieval strategies. Psychology of Learning and Motivation, 22, 227-259.
Reder, L., Liu, X., Keinath, A., \& Popov, V. (2015). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. Psychonomic Bulletin \& Review.
Reder, L., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., \& Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(2), 294.
Reder, L., Oates, J. M., Dickison, D., Anderson, J. R., Gyulai, F., Quinlan, J. J., ... Jefferson, B. F. (2007). Retrograde facilitation under midazolam: The role of general and specific interference. Psychonomic Bulletin \& Review, 14(2), 261-269.
Reder, L., Park, H., \& Kieffaber, P. D. (2009). Memory Systems Do Not Divide on Consciousness: Reinterpreting Memory in Terms of Activation and Binding. Psychological Bulletin, 135(1), 23-49.
Reder, L., Proctor, I., Anderson, J. R., Gyulai, F., Quinlan, J. J., \& Oates, J. M. (2006). Midazolam does not inhibit association formation, just its storage and strengthening. Psychopharmacology, 188(4), 462-471.
Schunn, C. D., Reder, L., Nhouyvanisvong, A., Richards, D. R., \& Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(1), 3.
Shiffrin, R. M., Ratcliff, R., \& Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(2), 179.
Underwood, B. J. (1969). Some correlates of item repetition in free-recall learning. Journal of Verbal Learning and Verbal Behavior, 8(1), 83-94.
Waugh, N. C. (1962). The effect of intralist repetition on free recall. Journal of Verbal Learning and Verbal Behavior, 1(2), 95-99.
Wickelgren, W. (1972). Trace resistance and decay of long-term memory. Journal of Mathematical Psychology, 9(4), 418-455.

# When Less Isn't More: A Real-World Fraction Intervention Study 

Tasha Posid (posid.1@osu.edu)<br>Department of Psychology, The Ohio State University<br>1835 Neil Avenue<br>Columbus, OH 43210 USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)<br>Department of Psychology, The Ohio State University<br>1835 Neil Avenue<br>Columbus, OH 43210 USA


#### Abstract

Although an understanding of fractions is a critical precursor for other mathematical concepts, including proportional reasoning, algebra, and success in STEM fields, surveys of mathematics education in the United States indicate that school-age children lack age-appropriate math skills and proficiency. Thus, understanding the critical precursors of fraction knowledge is important for the development of instructional materials. The aim of the present study was to examine whether instructional format affected children's learning and transfer of fraction concepts, and whether individual variables such as executive function and math knowledge moderated these effects. Six- to 8 -year-old children participated in a longitudinal, pre/post test design, in which they received a fraction-training intervention. Critically, we manipulated the extent to which real-world instruction was grounded in visual vs. symbolic representations. We find that $1^{\text {st }}$ and $2^{\text {nd }}$ graders were able to learn fraction concepts following this intervention, despite having no formal fraction education. The extent to which the instructional stimuli were grounded in visual vs. symbolic representations affected children's proportional reasoning knowledge in a transfer task, and condition effects were moderated my children's working memory and prior math knowledge. This work has implications for instructional design and curriculum development in the classroom.


Keywords: Numerical cognition, fractions, proportional reasoning, education, learning.

## Introduction

An understanding of fractions is a critical precursor for other mathematical concepts, including probability, proportional reasoning, algebra, and much of the STEM fields (Bailey, Hoard, Nugent, \& Geary, 2012; Department of Education, 1997). In fact, early fraction knowledge predicts the acquisition of algebraic knowledge well into middle and high school. However, surveys of mathematics education in the United States indicate that school-age children lack age-appropriate math skills and proficiency (NAEP, 2009; NCES, 2010; Siegler et al., 2012; also see Hurst \& Cordes, 2016). Thus, improving students' math knowledge and reasoning ability about proportions early in a child's education is important. Furthermore, understanding what instructional format may best lead to both the learning and transfer of difficult math concepts (i.e., proportions) should be a fundamental component of instruction and curriculum development.

The aim of the present study was to examine whether the instructional format in which fraction concepts are taught would affect the learning and transfer of novel fraction concepts (Core Curriculum; New Common Core Mathematics Standards, 2000), as well as whether individual variables such as executive function and prior math knowledge would moderate any observed effects. On the one hand, one approach to teaching mathematics to young children involves the use of concrete instantiations, such as vibrant, perceptually-rich visual displays or realworld contextualized examples (e.g., Van de Walle, 2007), presumably because these high-contrast items are attentiongrabbing, motivating, and often found in a child's natural environment (NCTM, 2000). Perceptually rich education materials are abundantly available and often populate children's classrooms in an effort to keep children interested in the materials being taught (Peterson \& McNeil, 2012). Even teachers prefer perceptually rich materials (Peterson \& McNeil, 2012), as they presumably increase children's engagement in the task at hand.

On the other hand, much work suggests a "less is more" approach to teaching children about difficult math concepts. This work suggests that perceptually rich or concrete materials may hinder mathematics concept learning (and perhaps problem solving and computation) because extraneous perceptual information gets integrated into representation of the target concept (e.g., Kaminski \& Sloutsky, 2009, 2013; Kaminski, Sloutsky, \& Heckler, 2009; McNeil \& Fyfe, 2012; Mix, 1999, 2008; Peterson \& McNeil, 2012; Posid \& Cordes, 2014). For example, Posid and Cordes (2014) asked young children (3-6 years) to decide which of two arrays contained a target number of items, where half of the arrays were homogenous in makeup (e.g., all of the same kind of animal) and the other half of the arrays were heterogeneous in make-up (all different animals). They found that children were less accurate when arrays were heterogeneous in make-up, particularly when the task was more difficult (when children were asked to find a larger target numerosity), and that this homogeneity advantage remained present across development (Posid \& Cordes, 2014; also see: Mix, 1999, 2008). Similarly, Kaminski \& Sloutsky (2013) taught young children (kindergarten through second grade) to read bar graphs, while manipulating whether the graphs contained colorful and irrelevant information or monochromatic bars. They
found that the children trained on graphs with irrelevant colorful features often tried to use that extraneous perceptual information incorrectly, decreasing their overall accuracy compared to their peers who were trained on monochromatic bar graphs (Kaminski \& Sloutsky, 2013). Together, these studies suggest that these more perceptually rich concrete instantiations over-communicate information to the learner, compared to their generic counterparts, thereby hindering learning from the relevant mathematical structure or relations at hand.

The impact of perceptual information when learning about proportions or fractions is even less known. In one study, Kaminski and Sloutsky (2009) investigated kindergarteners' ability to identity proportions across two sets of stimuli, which varied in their degree of concreteness of the instantiations. Children in the concrete condition failed to compare novel proportions while children in the generic condition successfully compared novel proportions, following a sparse training and no instruction. These findings suggest that simple proportional relations can be learned following generic instantiations, while concrete instantiations do not promote this same type of learning.

In this vein, young children's learning of fractions may also benefit from more generic instantiations. Specifically, fractions are traditionally introduced as symbols (e.g., 1/2), where visual representations may add extraneous information that could be interpreted ambiguously or incorrectly. For example, additional visual information could (a) add a layer of perceptual richness by conveying concept-irrelevant information and/or distracting information to the learner from the specific math concept to-be-learned, (b) add extraneous conceptual information, such as sharing, when real-world instantiations such as the use of a pizza pie are used, or (c) add a combination of the two. Thus, the present study addresses whether symbolic or visual instantiations in particular provide pre-fraction learners with a better ability to learn about novel fraction concepts. Because little work has addressed whether perceptual or conceptual visual instantiations may be detrimental to the young learner, the present study utilizes minimalistic, black-and-white stimuli (perceptually but not conceptually rich) to begin to address this important research question.

## Overview of the Current Study

The aim of the present study was to examine whether instructional format affected children's learning and transfer of fraction concepts, and whether individual variables would moderate any observed visual vs. symbolic instantiation effects. Critically, we manipulated the extent to which real-world instruction was grounded in visual vs. symbolic representations, while incorporating actual educational practices into the training paradigm (material, context, multi-session lessons). First and second graders participated in a pre/post-test design in which they received fraction instruction over several intervention sessions (Ordinal Comparisons, Addition and Subtraction, Decomposition), followed by a test of transfer (Fraction and Proportional Reasoning).

## Method

## Participants

Seventy-three $1^{\text {st }}$ and $2^{\text {nd }}$ grade children ( $M_{\text {Age }}=6.9$ years) participated in this study. Children were tested in their own elementary school during regular school hours. All children were tested in a quiet room with a single female experimenter.

## Materials

Pre- and Post-Test: Fraction Battery.
The fraction pre- and post-test batteries were identical and consisted of three fraction-knowledge tasks, which asked participants to make judgments about either symbolic or visual fraction information (also see Hurst \& Cordes, 2016; Polinsky, Posid, \& Sloutsky, 2017; Posid \& Sloutsky, 2015, 2017). The first task was an Ordinal Task, in which participants were asked to judge which of two sets was numerically larger and included visual fraction comparison (e.g., $2 / 3$ vs. $1 / 3$, represented as black-andwhite circles divided into three equal parts, with two parts and one part shaded, respectively) and a symbolic fraction comparison (e.g., $2 / 3$ vs. $1 / 3$ ). Children next completed a Matching Task, in which they matched either a symbolic fraction (e.g., $1 / 3$ ) to a visual fraction (e.g., a black-andwhite circle divided into three parts with one part shaded) or vice-versa. Children then completed an Addition and Subtraction Task, in which they were asked to add visual fractions (e.g., black-and-white circles) or symbolic fractions (e.g., $1 / 4+2 / 4$ ). Due to the difficulty level of the Matching Task and Addition/Subtraction Task, four answer choices were offered in a multiple-choice format. For all fraction tasks in the pre-/post-test battery, only fractions $<1$ were used (e.g., $2 / 3$ but not 4/3).

## Fraction Training:

The fraction training intervention consisted of three sessions, which were identical in content and instruction across conditions. Each training session consisted of two parts: one-on-one guided instruction between the child and experimenter followed by a block of practice problems used to measure immediate learning. Importantly, although the content and instruction was consistent across conditions, children participated in one of three training conditions: (1) Visual Only (black-and-white circles only; $n=22$ ), (2) Symbols Only (symbolic fractions only; $n=25$ ), or (3) Visual + Symbols (both black-andwhite and symbolic fractions shown side-by-side; $n=23$ ). The instruction and practice block were exclusively run in the child's randomly assigned condition (see Figure 1).

The first training task was an Ordinal Comparison task, in which children were instructed on how to compare two fractions with either the same denominator (e.g., $1 / 4$ vs. $2 / 4$ ) or same numerator (e.g., $1 / 4$ or $1 / 8$; Figure 1). Critically, the fraction instruction throughout all three training tasks was meant to address two concepts prevalent in the fraction-learning literature to date. First, children were taught to use counting to identify and manipulate the numerators and denominators presented, as children notoriously demonstrate a "whole number bias" when
learning about fractions (e.g., DeWolf \& Vosniadou, 2014; Hurst \& Cordes, 2016; Ni \& Zhou, 2005; Obersteiner, van Dooren, Van Hoof, \& Verschaffel, 2013 Obersteiner, Van Hoof, Verschaffel, \& Van Dooren, 2016; Polinsky et al., 2017). Second, children were taught about fraction magnitude knowledge, specifically as it relates to the partwhole concept (e.g., Siegler, Thompson, \& Schneider, 2011; Stafylidou \& Vosniadou, 2004). This instruction and feedback was followed by a block of practice trials in which children continued to compare fractions, but without scaffolding or feedback from the experimenter.

The second training task was an Addition and Subtraction task, in which children were taught how to systematically add or subtract two fractions. Children were shown a correct strategy for solving this type of fraction problem. This was followed by a practice block, in which children continued to add and subtract fractions, but without the input and feedback from the experimenter.

Finally, children completed a Fraction Decomposition task, in which they were asked to add or decompose a series of fractions (e.g., $1 / 6+2 / 6+1 / 6$; Figure 2). The instruction was similar to that of the addition and subtraction task, with step-by-step instructions on how to identify the fractions, recognize that the denominator (or total number of pieces) was the same using counting, and then add the numerators (or shaded number of pieces) to find the answer. Again, this was followed by a block of practice trials in which children continued to decompose fractions, but without the instruction and feedback of the experimenter.


Figure 1. Examples of the Stimuli used in the Test portion of the Training Sessions.

## Transfer Task:

The transfer task was used to measure children's fraction and proportion knowledge, and consisted of a series of visual questions. Included questions asked children to make judgments related to either probability (e.g., "If you were to roll a dice, what is the probability that that it would land on a 2?" or "If you were to reach into this box of candy without looking, what is the probability that you would randomly pick out a cherry piece?) or fractions
(e.g., "Can you express the triangles as a fraction of the entire set?" or "If you were to reach in to one of two fish tanks, are hoping to pick a fish of a certain color, which fish tank should you reach into?").

All tasks were administered on a Macintosh laptop. These programs were created using RealBasic software, which also recorded participants' reaction time and answers during the tasks.

## Procedure

Pre/Post-Test Fraction Battery:
The procedure consisted of three phases: The Ordinal Task, the Matching Task, and the Addition/Subtraction Task. The Ordinal Task consisted of three blocks: natural number comparisons (warm up), visual fraction comparisons, and symbolic fraction comparisons. All fractions were $<1$ so as to match the visual and symbolic fractions featured in the second and third blocks. Each block in the Ordinal Task consisted of 32 trials, for a total of 96 trials. The Matching Task consisted of three blocks of 12 trials each, for a total of 36 trials. The Addition/Subtraction task consisted of four blocks of 12 trials each, for a total of 48 trials.

## Fraction Training:

The procedure consisted of three training sessions: Ordinal Comparisons, Addition and Subtraction, and Decomposition. The Ordinal Comparisons training session consisted of two phases: Same Denominator comparison training and Same Numerator comparison training. Each instruction phase consisted of two examples, followed by a block of 10 test questions ( 20 total test trials). The Addition and Subtraction training session consisted of two phases: Addition and Subtraction. Each instruction phase contained two examples, followed by a block of 20 test trials ( 40 total test trials). The Decomposition training session consisted of two phases of Decomposition instruction (fractions whose sum was $<1$ and fractions whose sum was $>1$ ), each followed by one blocks of test trials, plus a final block of "intermixed" test trials, for a total of 30 test trials.

## Transfer Task:

The transfer task consisted of seven blocks of proportional reasoning and fraction reasoning picture problems and included: (1) spinner proportions, (2) dice rolling, (3) determining the proportions of shapes in a set, (4) determining the proportion of candy in a bowl, (5) interpreting a pie graph, (6) representing a set of shapes as a fraction, and (7) determining the proportion of fish in a fish tank (for a total of 36 Transfer Task questions). The transfer task was administered approximately 2 weeks after post-test. It was also given in multiple-choice format due to the difficulty of the task itself and for consistency of testing format across sessions.

## Results

The present study examined three outcome variables of interest: (1) Learning at Training: Did children perform above chance on the test trials following each intervention
session? That is, did they learn the information they had just practiced with the experimenter? (2) Pre-Post Test Gains: Did children improve on our Fraction Battery from Pre-Test to Post-Test? (3) Transfer Task: Did children perform above chance on our Transfer Task? Critically, we examined the impact of Training Condition (Visual+Symbol, Visual Only, Symbols Only) on all three of these variables of interest. Finally, we examined the role of individual variables as moderators of Condition effects on our dependent variables.

## Learning at Training

Results indicate that young children were able to learn fraction concepts despite having no formal instruction in the classroom, as indicated by their above-chance performance during each phase of training (vs. chance: Ordinal: $p<.001$, Cohen's $d=2.8$; Addition/Subtraction: $p<.001$, Cohen's $d=2.4$; Decomposition: $p<.001$, Cohen's $d=4.4$; Figure 2). A moderate - but non-significant ( $p$ ' $s<.2$ ) -- trend demonstrated the impact of children's experimental condition across these tasks. That is, children's performance in the Symbols-Only and Visual+Symbols conditions were similar across the three day-of training tasks; however, children's accuracy in the visual-only condition was lower.


Figure 2. Accuracy on the test trials of each Fraction Training Session by Condition. Error bars reflect Standard Error of the Mean.

## Pre- to Post-Test Gains

We also assessed whether children made substantial gains from pre- to post-test within our Fraction Battery. For the purposes of this analysis, difference scores were created for each participant (post - pre) for each portion of the Fraction Battery (Ordinal, Matching, and Addition/Subtraction; pre-test: no effect of Condition: $p>.1$ ), such that each participant had a single difference score for each task representative of any gains made. Then, an average of these difference scores was created to identify children's total gains across the training tasks. A significantly positive difference score (versus zero) would indicate significant gains made by that participant.

Children demonstrated significant gains between pre- and post-test $(t(69)=10.6, p<.001$, Cohen's $d=2.6)$. Notably, condition differences were not observed in any of our three tasks within the Fraction Battery ( $p$ ' $s>.1$ ),
suggesting that, at least for $1^{\text {st }}$ and $2^{\text {nd }}$ graders, any instructional format can promote learning of these difficult math concepts.

## Transfer Task

Performance accuracy on the Transfer Task was calculated for questions pertaining to Proportional Reasoning and Fraction Reasoning. Overall, children performed abovechance on both types of questions in the Transfer Task (Proportional Reasoning: $p<.001$, Cohen's $d>.1$; Fraction Reasoning: $p<.001$, Cohen's $d>2.5$; Figure 3). Condition differences were observed for the Proportional Reasoning portion of the Transfer Task $(F(2,63)=6.5, p=.003)$, such that children in the Visual+Symbol condition outperformed their peers in both the Visual Only and Symbols Only conditions. Of note, children in the Visual Only and Symbols Only conditions performed significantly abovechance, and out-performed a secondary sample of untrained controls whose accuracy did not exceed chance-level $(t(6)=1.7, p=.15$, Cohen's $d=1.4)$. In contrast, no Condition effects were observed for the Fraction Reasoning questions ( $p>.7$ ), mirroring the lack of Condition effects observed in the pre- to post-test gains.

## Individual Variability

A series of regression analyses were run in order to investigate whether individual variables predicted children's performance across our dependent variables. Each regression model tested the following independent variables: pre-fraction knowledge (as assessed through our pre-test Fraction Battery), prior Math Knowledge (assessed through a portion of the Woodcock Johnson and a 3-minute speeded arithmetic test administered at pre-test), Inhibitory Control (assessed through a numerical stroop task administered at pre-test), Working Memory (assessed through a serial ordering task administered at pre-test), and the child's grade in school.

Children's accuracy during the day-of training tasks (composite score) was significantly predicted by their grade in school (Beta $=.224, p=.06$ ), prior math knowledge (Beta $=.573, p<.001$ ), and their pre-test fraction knowledge (Beta=.253, $p=.04$; Model: $R^{2}=.416, p<.001$ ). Children's performance at post-test was significantly predicted by children's prior math knowledge (Beta=.566, $p<.001$; Model: $R^{2}=.401, p<.001$ ), while children's pre- to post-test gains were significantly predicted by their grade in school (Beta $=.382, \quad p=.009$ ), working memory (Beta=.306, $p=.012$ ), and pre-test fraction knowledge (Beta $=.485$, $p=.001$; Model: $R^{2}=.220, p=.014$ ). Additional SEM modeling was conducted to examine whether our significant predictors specifically moderated any effects of Condition on our dependent variables. We find that prior math knowledge does moderate the effects of Condition on children's day-of training accuracy ( $p=.02$ ) and post-test fraction performance ( $p=.002$ ), while working memory moderated children's gains from pre- to post-test ( $p<.001$ ).

Finally, regression and SEM modeling were conducted for children's performance on the Transfer Task. Specifically, children's accuracy on the proportional
reasoning portion of the Transfer Task was significantly predicted by their grade in school (Beta=.227, $p=.059$ ), Condition (Beta $=.222, \mathrm{p}=.064$ ), inhibitory control (Beta $=.250, \quad p=.038$ ), and pre-test fraction knowledge (Beta $=.257$, $p=.086$ ). Again, secondary SEM models indicated that both math knowledge and working memory individually and significantly moderated performance on the Transfer Task. Specifically, children with High working memory did not show Condition differences, while those with low working memory did (Low WM: $F$ (2, $28)=9.0, p=.001$; High WM: $F(2,27)=1.7, p=.206)$. These children benefited most from Visual+Symbol. Like working memory, children with high math knowledge did not show Condition differences on the Transfer Task, whereas children with low math knowledge did. These children benefited most from Visual+Symbol (Low Math: $F(2,26)=4.7, p=.019$; High Math: $F(2,29)=1.7, p=.1)$.


Figure 3. Accuracy on the Transfer Task by Condition. Error bars reflect Standard Error of the Mean.

## Discussion

The aim of the present study was to examine whether instructional format affected children's learning and transfer of fraction concepts, as well as to investigate whether individual variables such as executive function and math knowledge moderated any effects of visual vs. symbolic instantiations. Results indicate two important patterns of performance. First, using real-world instructional stimuli from the current Core Curriculum (Core Curriculum; New Common Core Mathematics Standards, 2000), children as young as $1^{\text {st }}$ and $2^{\text {nd }}$ grade successfully learned new fraction concepts, as indicated by their above-chance performance on day-of training, in their gains observed from pre- to post-test on our Fraction Battery, and their above-chance performance on the Transfer Task. Because an understanding of fractions is an important precursor for other mathematical concepts, including probability, proportional reasoning, algebra, and much of the STEM fields (Bailey, Hoard, Nugent, \& Geary, 2012; Department of Education, 1997), it is critical that elementary school children are involved in a curriculum that employs these critical foundations in fraction education. Although previous surveys of mathematics education in the United States suggest that children lack age-appropriate math skills (NAEP, 2009; NCES, 2010; Siegler et al., 2012; also see Hurst \& Cordes,
2016), the present study suggests that current curriculum is utilizing content that may help close this gap in years to come.

Second, and more importantly, the present study finds that the instructional format in which the to-belearned concepts are presented to children is important. Specifically, those children in the Visual+Symbol condition faired best both during immediate learning within our intervention sessions and in our test of transfer two weeks following post-test. Of note, children in the Visual only condition never out-performed their peers in either of the other two conditions, suggesting that less is not necessarily more when teaching children about new and conceptually challenging fraction concepts. Moreover, children who were low in math knowledge and low in working memory at pre-test benefited most from the Visual+Symbol condition, suggesting that the redundant perceptual information was particularly helpful.

The findings from the present study suggesting that "less" is not "more" when teaching children about new fraction concepts is interesting given much work suggesting that extraneous perceptual information may interfere with children's ability to learn mathematical concepts or make mathematical reasoning judgments (e.g., Kaminski \& Sloutsky, 2009, 2013; Kaminski et al., 2009; McNeil \& Fyfe, 2012; Mix, 1999, 2008; Peterson \& McNeil, 2012; Posid \& Cordes, 2014). This pattern of findings could be accounted for by two explanations. First, perhaps either fractions themselves or novel fraction concepts are a stand-alone category. That is, perhaps "less is more" when children are learning about whole numbers or non-fraction numerical concepts. However, this is unlikely given ample research to suggest that both children and adults demonstrate a whole number bias, especially when learning about or solving fraction problems that are novel or difficult (e.g., see DeWolf \& Vosniadou, 2014; Hurst \& Cordes, 2016; Ni \& Zhou, 2005; Obersteiner et al., 2013; Obersteiner et al., 2016; Polinsky et al., 2017). That is, when solving fraction or proportion problems, children often apply their intuitions about whole numbers to fraction concepts (for example, they might say that $1 / 4+1 / 4=2 / 8$, as they incorrectly assume that you should add the numerators and the denominators, as you would if you were adding whole numbers).

Another explanation for the seemingly divergent pattern of findings observed in the present study comes from the nature of the stimuli used in the study itself. Although condition differences emerged, the stimuli were more perceptually impoverished than those used in previous work reporting "less is more" during mathematical learning and reasoning (Kaminski \& Sloutsky, 2009, 2013; Kaminski et al., 2009; McNeil \& Fyfe, 2012; Mix, 1999, 2008; Peterson \& McNeil, 2012; Posid \& Cordes, 2014). Specifically, the conditions in this study varied by whether the instantiations contained visual vs. symbolic vs. visual + symbolic information. However, the visual stimuli were always black-and-white circles, while the symbolic stimuli were a single monotone color (e.g., black). In contrast, real-world mathematics education
includes much more diverse and vibrant displays, the use of 2D and 3D objects, the use of "interesting" pictures and colors (e.g., a pizza pie to represent a pie graph or visual fraction; Peterson \& McNeil, 2012), and so on. Therefore, perhaps the "perceptually impoverished" framework employed in the present study muted the real-world effects of varying concrete vs. generic instantiations when teaching children about fractions and proportions. Currently, work from our laboratory is exploring this variation to visual vs. symbolic instantiations, through the use of a perceptually rich training paradigm. We are currently exploring whether similar Condition effects and individual moderators will emerge when perceptually rich (e.g., pizza pies rather than black-and-white circles, Sesame Street-like numbers with colors and eyes, etc.) stimuli are used in a similar training intervention.


Figure 4. Perceptually rich instantiations of the stimuli used in the present study.

In conclusion, the present study utilized a realworld fraction training intervention and finds that children can learn fraction information prior to formal education, using instructional material from the current Core Curriculum. Importantly, although all children demonstrated gains following training, those who received redundant perceptual information tended to out-perform their peers following immediate learning and in a transfer test of proportional reasoning. Additionally, children's prior math knowledge and working memory moderated our observed effects, indicating these should be taken into consideration when children are taught novel or difficult fraction concepts. This work has implications for instructional design and curriculum development in the classroom.

## Acknowledgements

We would like to thank Sydney Clark, Abigail Mills, Emily Neer, Naomi Polinksy, Alicia Scimeca, and Emily Thieroff of the OSU Cognitive Development Lab, as well as all of the children, teachers, and schools who participated in this study. Funding: IES grant R305A140214 to VS.

## References

Bailey, D.H., Hoard, M.K., Nugent, L., \& Geary, D.C. (2012). Competence with fractions predicts gains in mathematics achievement. Journal of Experimental Child Psychology, 113, 447455.

Department of Education (1997). Mathematics equals opportunity. White Paper prepared for US Secretary of Education Richard W. Riley.
DeWolf, M., \& Vosniadou, S. (2015). The representation of fraction magnitudes and the whole number bias reconsidered. Learning and Instruction, 37, 39-49.
Hurst, M., and Cordes, S (2016). Rational-number comparison across notation: Fractions, decimals, and whole numbers. Journal of

Experimental Psychology: Human Perception and Performance, 42(2), 281-293.
Kaminski, J.A., \& Sloutsky, V.M. (2009). The effect of concreteness on children's ability to detect common proportion. IN N. Taatgen \& H. van Rijn (Eds.), Proceedings of the XXXI Annual Conference of the Cognitive Science Society (pp. 335-340). Mahwah, NH: Erlbaum.
Kaminski, J.A., \& Sloutsky, V.M. (2013). Extraneous Perceptual Information Interferes with Children's Acquisition of Mathematical Knowledge. Journal of Educational Psychology, 105(2), 351-363.
Kaminski, J.A., Sloutsky, V.M., \& Heckler, A. (2009). Transfer of mathematical knowledge: The portability of generic instantiations. Child Development Perspectives, 3(3), 151-155.
McNeil, N.M., \& Fyfe, E.R. (2012). "Concreteness fading" promotes transfer of mathematical knowledge. Learning and Instruction, 22, 440-448.
Mix, K.S. (1999). Similarity and numerical equivalence: Appearances count. Cognitive Development, 14, 269-297.
Mix, K. S. (2008a). Children's equivalence judgments: Cross-mapping effects. Cognitive Development, 23, 191-203.
Mix, K. S. (2008b). Surface similarity and label knowledge impact early numerical comparisons. British Journal of Developmental Psychology, 26, 13-32.
National Center for Education Statistics (2010). The Nation's Report Card: Mathematics 2009: National Assessment of Educational Progress at Grades 4 and 8. (NCES 2010-451). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Council of Teachers of Mathematics (2000). Principles and standards for school mathematics. Reston, VA.
National Governors Association Center for Best Practices, Council of Chief State School Officers. Common Core State Standards for Mathematics. Retrieved from http://www.corestandars/org/assets/CCSSI_Math\ Standards.pdf.
Ni, Y., \& Zhou, Y.D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. Educational Psychologist, 40(1), 27-52.
Obersteiner, A., Van Dooren, W., Van Hoof, J., \& Verschaffel, L. (2013). The natural number bias and magnitude representation in fraction comparison by expert mathematicians. Learning and Instruction, 28, 64-72.
Obersteiner, A., Van Hoof, J., Verschaffel, L., \& Van Dooren, W. (2016). Who can escape the natural number bias in rational number tasks? A study involving students and experts. British Journal of Psychology, 107, 537-555.
Peterson, L.A., \& McNeil, N.M. (2012). Effects of Perceptually Rich Manipulatives on Preschoolers' Counting Performance: Established Knowledge Counts. Child Development, 84(3), 1020-1033.
Polinsky, N., Posid, T., \& Sloutsky, V.M. (2017). Training aids children's ability to solve a proportional equivalence task: Evidence from an eye-tracking study. Manuscript in preparation.
Posid, T., \& Cordes, S. (2014). Verbal counting moderates perceptual biases found in children's cardinality judgments. Journal of Cognition and Development, 16(4), 621-637. doi: 10.1080/15248372.2014.934372.

Posid, T., \& Sloutsky, V.M. (2016). Kindergarteners and adults learn fraction-rules in a categorization task. Poster presented at the Proceedings of the $38^{\text {th }}$ Annual Conference of the Cognitive Science Society. Philadelphia, PA: Cognitive Science Society.
Posid, T., \& Sloutsky, V.M. (2017). Is less really more: How and when kindergarteners and adults learn fraction-rules in a categorization task. Manuscript in preparation.
Siegler, R.S., Duncan, G.J., Davis-Kean, P.E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M.I., \& Chen, M. (2012). Early predictors of high school mathematics achievement. Psychological Science, 23.
Siegler, R.S., Thompson, C.A., \& Schneider, M. (2011). An integrated theory of whole number and fractions development. Cognitive Psychology, 62(4), 273-296.
Stafylidou, S., \& Vosniadou, S. (2004). The development of students' understanding of the numerical value of fractions. Learning and Instruction, 14(5), 503-518.
Van de Walle, J.A. (2007). Elementary and Middle School Mathematics: Teaching Developmentally. New York: Pearson.

# Is ambiguity detection in haptic imagery possible? Evidence for Enactive imaginings 

Wim Pouw (pouw@fsw.eur.nl) ${ }^{1}$, Asimina Aslanidou (410929aa@student.eur.nl) ${ }^{1}$ Kevin L. Kamermans (388422kk@student.eur.nl) ${ }^{1}$ Fred Paas (paas@fsw.eur.nl) ${ }^{1,2}$<br>1. Department of Psychology, Education and Child Studies, Erasmus University Rotterdam Burgermeester Oudlaan 50, 3062PA Rotterdam, The Netherlands<br>2.Early Start Research Institute, University of Wollongong<br>Wollongong NSW 2522, Australia


#### Abstract

A classic discussion about visual imagery is whether it affords reinterpretation, like discovering two interpretations in the duck/rabbit illustration. Recent findings converge on reinterpretation being possible in visual imagery, suggesting functional equivalence with pictorial representations. However, it is unclear whether such reinterpretations are necessarily a visual-pictorial achievement. To assess this, 68 participants were briefly presented 2-d ambiguous figures. One figure was presented visually, the other via manual touch alone. Afterwards participants mentally rotated the memorized figures as to discover a novel interpretation. A portion $(20.6 \%)$ of the participants detected a novel interpretation in visual imagery, replicating previous research. Strikingly, $23.6 \%$ of participants were able to reinterpret figures they had only felt. That reinterpretation truly involved haptic processes was further supported, as some participants performed co-thought gestures on an imagined figure during retrieval. These results are promising for further development of an Enactivist approach to imagination.


Keywords: visual imagery; haptic imagery; gesture; enactivism; the imagery debate

## Introduction

Early phenomenological observations concerning voluntary visual imagery suggested that nothing new can be discovered in visual imagery that was not present in the intention to imagine to begin with (Sartre, 1940). This fits Descriptivist renderings of visual imagery (Pylyshyn, 2002), where visual imaginings have a fixed mode of presentation, and are "images under a description" (Fodor, 1975, p. 191). This view gained substantive empirical traction (in part) by research showing that when participants memorized an ambiguous figure under a particular percept (e.g., duck), they were unable to discover an alternate novel interpretation (e.g., rabbit) when retrieving the ambiguous figure in visual imagery (Chambers \& Reisberg, 1985; Slezak, 1991). These results were obtained even though the memory of the figure was detailed enough to draw it out, and which allowed for subsequent ambiguity detection when perceiving the drawing.

Subsequent research within the ambiguity detection paradigm showed that the previous studies may have employed too difficult ambiguity examples (e.g., duck/rabbit figure), and failed to properly inform
participants by providing an ambiguity example (Brandimonte \& Gerbino, 1993; Finke, Pinker, and Farah, 1989; Hyman \& Neisser, 1991). One of those studies showed an ambiguity detection rate of $40 \%$ with slightly less complex figures as the classic duck/rabbit figure and providing an ambiguity example (Peterson, Kihlstrom, Rose, \& Glisky, 1992). Some shortcomings of previous studies that found positive findings were resolved by Mast and Kosslyn (2002), who noted that participants might have been alerted by the ambiguity of the figures during perception as they were shown an ambiguous figure before memorizing the target figure. In their study they found that $44 \%$ of the participants were able to detect an alternative interpretation while excluding possible confounds of ambiguity detection in perception rather than imagery.

In a recent study we have expanded upon this research (Kamermans, Pouw, Mast, \& Paas, under review). Participants in Mast and Kosslyn's (2002) study were provided partial visual cues of the ambiguous figure during imagination. Therefore, it could not be fully excluded that some raw sensory information is necessary for ambiguity detection via imagery to occur, allowing for the possibility that ambiguity detection via imagery alone is impossible after all. In our previous study however, we found that ambiguity detection is possible without visual cues as well ( $30 \%$ detection rate), while excluding other possible confounds such as ambiguity detection during perception rather than retrieval in imagery.

The ambiguity detection paradigm has been primarily regarded as being important for The Imagery Debate. This is because, in contrast to the Descriptivist approach, the other contender in The Imagery Debate - the Quasi-pictorial account (Kosslyn, 2002) - explicitly argues for the possibility of ambiguity detection. On such an account, visual imaginations are constituted by internal representations that are experienced as and function like pictorial representations (drawings, diagrams, etc.). As such, analogous to pictures, the representational content of visual images are not intrinsically fixed, rather the mental image preserves the "raw" visual-pictorial information of the previously seen object which is open to reinterpretation.

In summary, there is converging evidence that visual imagination allows for similar re-interpretative feats as pictorial representations (Mast \& Kosslyn, 2002; Kamermans et al., under review). Such evidence has been
mainly interpreted as a win for the Quasi-pictorial account over the Descriptivist account (Mast \& Kosslyn, 2002). Yet, although reinterpretation in mental imaginings might afford similar feats as pictorial representations (e.g., drawings), it need not be the case that imagination functions exactly like visual-pictorial representations. In fact, imagery-based reinterpretation need not be visual-pictorial at all.

## Enactive Imaginings

Indeed, Quasi-pictorial and Descriptivist accounts of mental imagery are not the only game in town (Foglia \& O'Regan, 2016). There are also views that promote an Enactive view of sensory perception and imagination (Hurley, 2002; O'Regan \& Noe, 2001; Thomas, 1999) ${ }^{1}$. Our conception of the Enactive position in relation to imaginings, is that imaginings are something we do rather than an internal state; the success of an imagining is not (primarily) mediated by internally stored knowledge of properties of the object imagined, but rather by a pre-reflective understanding of sensori-motor relations that would hold if the object would be present. The enactive logic is that since perception is an accomplishment of an active embodied system (Gibson, 2014) so too must perceptual imagination be constituted in a practical understanding of the sensori-motor relations that would hold when perceiving some object.

Evidence closely in par with an Enactive view concerns findings that show a functional role for eye-movements in visual imagination (Brandt \& Stark 1997; Laeng \& Teodorescu 2002; Spivey \& Geng, 2001). Note, that the functional role eye-movements seem to have in visual imagination is achieved even though eye-movements themselves do not provide visual information in a classic sense at all, e.g., eye-movements also occur and affect visual imagery when eyes are closed (Spivey, Tyler, Richardson, \& Young, 2000). Note as such that although the Quasi-pictorial account could be invoked here to explain the eye-movements (as the eye-movements interact with supposed internally stored visual information of the object), it is difficult to explain the function of eye-movements which provide no visual information in the classic sense. It is precisely because an Enactive view does not adopt a classic view of perception that it is able to recognize that (pre-reflective knowledge of) bodily movements constitute perception and imagination (O’Regan \& Noe, 2001; Hurley, 2002). Imagination, on such a construal, involves being attuned to sensori-motor potentialities of a particular object imagined (Thomas, 1999; O'Regan \& Noe, 2001). In visual imagery this attunement seems to be achieved in part through reenacting eye movements (a co-constituent of visual perception).

[^440]
## Present study

In the current study participants memorize two figures in succession for 30 seconds. Both of these figures have an alternate interpretation when rotated 180 degrees (i.e., figures are ambiguous). One of these figures is provided visually, the other via touch alone (i.e., haptic perception). It is then assessed whether participants are able to find the alternate interpretation for each figure in their imagination by mentally rotating the memorized figure.

As a further extension of the possibility of ambiguity detection in visual imagery we assess here whether participants can perform visual reinterpretations without being provided with an ambiguity example (cf. Kamermans et al., under review). However, the most important extension relative to previous research that we emphasize in this article, is the assessment of whether ambiguity detection can be performed in haptic imagery as well. That is, similar to visual imagery, can ambiguity be detected upon a mental imagining that is based on an ambiguous figure that was explored via touch?

Importantly, it could be that feeling a figure allows for ambiguity detection in imagery only insofar participants are able to reconstruct visual information based on this haptic perception. If true, ambiguity detection is always performed on visual information. On such a construal it is plausible that ambiguity detection rates of visually perceived figures, is greater than when haptically perceiving figures. After all, if ambiguity detection is a visual achievement, having had direct visual access as opposed to second-hand access (visual reconstruction via haptic perception) would improve the quality of the mental image, and hence improve reinterpretability. Therefore, next to assessing the possibility of ambiguity detection per condition individually, we will assess possible differences in detection rate between the haptically and visually perceived figures that would be predicted if reinterpretation in imagery is strictly a visualpictorial achievement.

Additionally, in light of an Enactive approach to sensory perception and imagination, and our previous theoretical efforts concerning the cognitive function of co-thought gesture (Pouw, de Nooijer, van Gog, Zwaan, \& Paas, 2014), we anticipated to observe co-thought (i.e., silent) handgestures that enacted interaction with the object during haptic imagery. That is, similar to research showing that eye-movements appear to co-constitute visual imagination, so too might manual movements co-constitute haptic imagery processes. Therefore, in the current study we explored manual gestures that occur when retrieving the haptically (as well as visually) memorized figure.

## Method

## Participants \& Design

Sixty-eight participants were tested (61 female, Mage = 20.03 years, $S D a g e=3.36$ years, range 17-37 years) ${ }^{2}$. Recruitment targeted both Dutch and non-Dutch students all of whom received instructions in English. The participants were enrolled in courses taught in English at the Erasmus University Rotterdam. All of the participants took part in the experiment for course credits.

The study had a within-subject design (Visual vs. Haptic Condition; order counterbalanced) using two test figures counterbalanced over condition assignment. Ambiguity detection rate was the main dependent variable. Additionally we assessed co-speech and co-thought gesture occurrences.

## Materials

Test Figures Two ambiguous test figures were cut out from high density foam sheets, similar to Kamermans et al. (under review). The figures were designed by Leo Burnett (2015) for the "Upside Down" campaign retrieved from Google images and further modified by us. Each figure had two readily perceivable interpretations (see figure 1). An alternate interpretation could be discovered by rotating the figure 180 degrees.


Figure 1. Line drawings of the seal/doe, and the penguin/giraffe test figure. As can be seen, one interpretation always showed the body of an animal and the second interpretation the head of a different animal.

Video Recording Performance was recorded using a JVC Everio GZ-MG130 camcorder, to assess gesture occurrence and inspection of the behavioral data when necessary.
Demographics and Control Questions Participants reported their age, sex, and native language. To assess participants' beliefs about the nature of the experiment the following questions were included: "What do you think was the purpose of the current study? (If you have no idea, no answer is necessary)", and "What do you think the researchers are expecting to discover with the current study? (If you have no idea, no answer is necessary)". Finally, the experimenter would explicitly ask participants who reported reinterpretation of one or two figures whether they had noticed the alternate interpretation during memorization or newly discovered it in visual imagery.

[^441]
## Procedure

Participants were tested individually and were told that they took part in a study about visual memory and memory of touch. The experiment consisted of a memorization phase and a testing phase.

In the memorization phase two figures were presented successively (order counterbalanced). Participants were informed that they would be given 30 seconds to inspect the figure and memorize it. In the Visual Condition versus the Haptic Condition participants were only allowed to see or touch the figure, respectively. For the Haptic Condition participants felt the contours of the figure which was presented under a card-board box that prevented visual inspection of the figure. The respective figure assigned to the haptic condition was horizontally attached via velcrotape on a wooden base. The wooden base was shown to participants as to inform them how the figure would be placed on top of it (it was stressed that the wooden base was not the object of inspection). Participants were informed not to move the figure on the wooden base (experimenter ensured that orientation of the figure was not altered, which was assessable via an opening in the card-board box at the experimenter side).

After each presentation of a figure, the participants were asked what they had seen or felt (depending on condition) and the experimenter noted down the response. If participants reported a) two or more distinct interpretations or b) only the interpretation that belonged to the 180 degree alternative orientation, the associated testing phase would be skipped as ambiguity was detected prematurely (a) or signaled (b) during memorization.

After memorizing the two figures, in the subsequent testing phase participants mentally retrieved the memory of each figure consecutively (order of retrieval counterbalanced). They retrieved figures with their eyes closed, as to ensure that visual input during retrieval was consistent between participants. Once participants indicated that they had brought back their memory of the respective figure, the experimenter would inform them that this figure had another interpretation next to the interpretation they already gave. Participants were told that the alternate interpretation could be discovered by rotating the mental image 180 degrees. Participants were given no time restrictions in discovering the alternate interpretation (reaction times were timed however, but were not of special interest).

## Performance and Scoring

An answer was considered correct in case the same interpretation was given by another participant in the respective condition (visual or haptic) of the memorization phase. For example, the answer "cow" for the head (doe) orientation of the seal/doe figure in the haptic condition would be considered correct if another participant had reported the same answer in the memorization phase for the haptic condition and the same orientation. For the main
confirmatory analysis, we chose for checking answers in the congruent modality because the crucial research question is about the continuity between perception and imagination, and this is distorted if both modalities will track properties of objects idiosyncratically.

Note therefore, participants were primarily their own raters in this study, minimizing post hoc experimenter decisions. However, similar to our procedure in Kamermans et al. (under review) we decided that in two cases answers should be counted as correct/incorrect even though these specific interpretations were (not) named by other participants. Namely, we counted as correct "Walrus/Seal" in the haptic condition for the figure seal and the "letter Y" (which partially overlapped with "Y stick") as an incorrect combination in the haptic condition for the figure seal. Note, that these post-hoc choices do not affect interpretation of the results.

## Results

## Interpretations

In Table 1 all the interpretations are given that overlapped between participants in the perception and testing phase (and therefore scored correct).

Table 1: Overlapping (re)interpretations

| Visual Condition <br> Giraffe | Deer, Giraffe <br> (Baby) cow |
| :--- | :--- |
| Doe | Penguin, Bird |
| Penguin | Seal |
| Seal |  |
| Haptic Condition | Tree, Flower |
| Giraffe | Deer, Aeroplane, Bird, Propeller |
| Doe | Penguin, Fish, Human Being |
| Penguin | Sea Lion, Walrus/Seal*, Bird |
| Seal |  |

Note. Overlaps between perception during memorization and mental retrieval of figures between participants. Asterisk pertain to post-hoc decision (1 instance).

## Exclusion

A total of 46 (34\%) out of 136 ambiguity detection trials were excluded as there was either premature ambiguity detection (10/46) or an interpretation was given during memorization that did not match the orientation that the figure was given in (36/46). However, we will exclude each participant data if on any of the two trials within participants premature ambiguity detection was obtained as this signals awareness of ambiguity during the memorization phase, 34 participants were excluded (50\%). The total sample to assess ambiguity detection rates per condition thus consists of participants who were not aware in any of the trials of ambiguity in the figures during memorization, ensuring that ambiguity detection ensued in imagery. Note that such high exclusion rates are common in ambiguity detection research as to maximally control for ambiguity detection during perception rather than imagery (see e.g., Mast \& Kosslyn, 2002).

## Descriptives

Retrieval time Participants in the haptic condition had an average of $23.88(S D=19.51)$ seconds to provide an interpretation or abort attempt. Participants in the visual condition took on average 27.68 ( $S D=21.23$ ) seconds.
Haptic Imagery vs. Visual Imagery As the descriptives show in Table 2, it seems that ambiguity detection is possible both when the figure is memorized and identified visually as well as through touch alone. Furthermore, four participants were able to detect ambiguity in both the haptic and visual conditions, and seven participants only detected ambiguity in the visual $(\mathrm{n}=4)$ or the haptic $(\mathrm{n}=3)$ conditions. Thus data is almost completely symmetrical across conditions and our planned within subjects-test for a binary outcome reflects this, McNemar $p>0.99$.

Table 2: Overall detection rate

|  | Visual | Haptic |
| :--- | :--- | :--- |
| Doe/Seal | $1 / 12(8.3 \%)$ | $6 / 22(27.3 \%)$ |
| Giraffe/Penguin | $6 / 22(27.3 \%)$ | $2 / 12(16.6 \%)$ |
| Total | $\mathbf{7 / 3 4}(\mathbf{2 0 . 6 \%})$ | $\mathbf{8 / 3 4}(\mathbf{2 3 . 6 \%})$ |

Haptic-Visual Imagery: Effects of Crossmodal Scoring However, it could be noted that since each condition (visual vs. haptic) has its own rating system that this could affect/distort our interpretation in significant ways. As such we performed an additional exploratory re-analysis where we counted any interpretation correct if it was named in the visual or the haptic memorization phase (see Table 1). This revealed a McNemar significance test $p=.227$, with a detection rate of $38.2 \%$ for the visual condition and $23.6 \%$ for the haptic condition (table 3). Thus, note that the more lenient scoring system slightly inflated detection rates in the visual condition but not to a degree that the null-hypothesis could be rejected.

Table 3: Overall detection rate

|  | Visual | Haptic |
| :--- | :--- | :--- |
| Doe/Seal | $3 / 12(25 \%)$ | $6 / 22(27.3 \%)$ |
| Giraffe/Penguin | $10 / 22(45.5 \%)$ | $2 / 12(16.6 \%)$ |
| Total | $\mathbf{1 3 / 3 4}(\mathbf{3 8 . 2 \%})$ | $\mathbf{8 / 3 4}(\mathbf{2 3 . 6 \%})$ |

Exploratory: Gesture occurrence During testing we observed that participants adopted spontaneous gestures during the testing phase when retrieving and interpreting the figure even though participants kept their eyes closed during retrieval. Firstly, participants used co-speech gestures when providing an interpretation (often accompanied by a description) of the figure $44 / 68$ ( $64.7 \%$ ) in the haptic condition and slightly less in the visual condition, 36/68 ( $52.9 \%$ ). Perhaps, such co-speech gestures solely fulfill communicative purposes, as the interpretation needed to be communicated to the experimenter. However, we also found that a select few performed gestures in silence (i.e., cothought gestures), as-if feeling the contours of the previously felt figure (3 participants in the haptic condition) and more pointing-and-tracing gestures in the visual
condition ( $n=3$ ), e.g., tracing the contours of the figure on the table. Further note that when premature ambiguity detection is controlled for (sample $=34$ ) we found that $20 \%$ (4/20) of those who gestured in the haptic condition detected ambiguity whereas $28.5 \%$ (4/14) of those did that did not gesture. Furthermore, $23.5 \%$ (4/17) of gesturing participants in the visual condition detected ambiguity as opposed to $17.6 \%$ (3/17) for non-gesturing participants.


Figure 2. Example co-thought gesture in haptic trial

## Discussion

The current results replicate previous research showing that ambiguity detection in visual imagery is possible (Mast \& Kosslyn, 2002; Kamermans et al., under review), i.e., when memorizing a figure under a particular percept this figure can be discovered to have another interpretation by re-inspecting it in visual imagery. This finding further extends this research in that ambiguity detection in visual imagery is possible without showing participants an example of an ambiguous figure after memorization (in contrast to Kamermans et al., under review). Further research could employ more direct comparisons of the effect of providing an ambiguity example or not after memorization on ambiguity detection rates (for a discussion on this see Peterson et al., 1992).

More importantly however, we have shown that ambiguity detection of a figure in mental imagery is possible even when the figure is explored via manual touch alone, as evidenced by the approximate $23.6 \%$ detection rate in the haptic condition ("approximate" barring different coding schemes). The important question is whether this shows that ambiguity detection was (solely) performed via haptic imagery, or whether exploring a figure via touch results in visual images upon which reinterpretation could be performed. This is a question that this study does not address directly.

However, in speculative vain it is striking that we did not find statistically significant differences in a within-subject test on ambiguity detection for visually or haptically explored figures. If ambiguity detection is solely performed in the visual modality of imagination, one would predict that having direct visual access to an ambiguous figure would support the quality of that imagination - hence inflating
ambiguity detection - relative to figures that were only haptically perceived. However, even bearing this in mind it does not exclude that ambiguity detection was not achieved via visual imagery.

On the other hand however, we do find evidence that reinterpretation of imagined haptically explored figures is likely to involve some haptic processes, or at least at times, given that a few participants were manually reenacting manipulating an (imagined) figure. This directly relates to eye-movements (i.e., "ocular reenactments") during visual imagery. These eye-movements reenact perceptual affordances of previously seen figures. In the current example, manual reenactments seem to exploit manual affordances of previously felt figures. In sum, we find it an attractive hypothesis that - at minimum - a cross-modal visual-haptic imagination is performed when reinterpreting previously felt figures. This is in par with a host of studies showing that haptic and visual perception are co-informative (e.g., Lacey, Campbell. Sathian, 2007; Wallraven, Bülthoff, Waterkamp, van Dam, \& Gaissert, 2014).

## A New Imagery Debate?

As mentioned, discovery of a novel re-interpretation of a visual imagining has been regarded as supporting the Quasipictorial account of visual imagery (Mast \& Kosslyn, 2002) and refuting the Descriptivist assumption that visual imagery is necessarily fixed under an original perceptual ascription (Chambers \& Reisberg, 1985; Pylyshyn, 2001). However, on our reading, ambiguity detection can also be accounted for within an Enactive framework, where the achievement reinterpretation requires a skill-full act that is inherently unstable because of the lack of direct access (O'regan \& Noe, 2001; Thomas, 1999). This constraint of imagery as inherently unstable may actually have some explanatory power over the Quasi-pictorial account. Note, for example that only a small portion (not more than $38 \%$ ) of the current sample is able to reassign meaning to a mental imagining ${ }^{3}$. This is striking as there is reason to believe that the current figures are relatively easy and memorized up to a point that participants can accurately draw them, as previous research with shorter memorization times and more complex figures has shown (Chambers \& Reisberg, 1985). Thus, if it were the case that ambiguity detection depends on internally represented pictorial information, reinterpretation rates should be much higher. If however, the achievement of gaining access to a previously seen or felt object via imagining consists in an effortful skill-based employment of a coalition of sensory systems, then it is not surprising that ambiguity detection is as difficult as it is. It does not require a mere retrieval and re-inspection of internally represented visual-pictorial imprints, but an attunement to sensori-motor

[^442]contingencies when interacting with an object. Further note, that the proposed multimodality of imagery does not apply to haptic ambiguity detection alone, as our findings show that even in cases where participants reinterpret a figure in visual imagery, non-communicative gestures are recruited that exploit affordances of the figure as if it were visually present (e.g., by tracing the contour on the table during retrieval with eyes closed).

Before concluding, some shortcomings need to be shortly stressed (as to invite further research). Firstly, no causal relation can be inferred from manual action and haptic imagery at this point. We are currently performing a study that manipulates manual enactment to ascertain its role in (haptic) imagery. Secondly, although results signal that haptic processes may be directly involved in ambiguity detection of previously felt figures, the current design does not exclude the possibility that ambiguity detection is purely a visual achievement. This is because haptic perception might induce visual construals which allow for visual reinterpretation. Additionally, as mentioned by a reviewer, current participants remembered two figures that were alike in their representation of an animal body/head. This likeness might interfere with memorization of both figures, perhaps making comparison between haptic/visual conditions more problematic.

To conclude, we have provided evidence that reinterpretations in mental imagery can be achieved when figures are memorized visually or via manual touch alone. We have argued on the basis of a) the observed manual enactments during imagery, as well as b) the lack of observed differences between visual and haptic ambiguity rates, that ambiguity detection might not be purely visualpictorial. Instead, we have speculated that visual and/or haptic imagery is mediated by a pre-reflective understanding of the sensori-motor relations that would hold were the object present.

## Acknowledgments

This research is funded by the Excellence Initiative grant (Erasmus University Rotterdam) awarded to the Educational Psychology section.

## References

Brandimonte, M. A., \& Gerbino, W. (1993). Mental image reversal and verbal recoding. Memory and Cognition, 21, 23-33. doi: 10.3758/BF03211161
Brandt, S. A., \& Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. Journal of Cognitive Neuroscience, 9, 27-38. doi: 10.1162/jocn.1997.9.1.27
Chambers, D., \& Reisberg, D. (1985). Can mental images be ambiguous? Journal of Experimental Psychology. Human Perception and Performance, 11, 317-328. doi: 10.1037/0096-1523.11.3.317

Finke, R. A., Pinker, S., \& Farah, M. J. (1989). Reinterpreting visual patterns in mental imagery. Cognitive Science, 13, 51-78. doi: 10.1016/0364-0213(89)90011-6
Fodor, J. A. (1975). The language of thought. Cambridge:

Harvard University Press.
Foglia, L., \& O’Regan, K. J. (2016). A new imagery debate:
Enactive and sensorimotor accounts. Review of
Philosophy and Psychology, 7, 181-196. doi:
10.1007/s13164-015-0269-9

Gibson, J. J. (2014). The ecological approach to visual perception: Classic edition. Psychology Press.
Hurley, S. L. (2002). Consciousness in action. Cambridge: Harvard University Press.
Lacey, S., Campbell, C., \& Sathian, K. (2007). Vision and touch: multiple or multisensory representations of objects? Perception, 36(10), 1513-1521.
Laeng, B., \& Teodorescu, D.-S. (2002/3). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. Cognitive Science, 26, 207-231. https://doi.org/10.1016/S0364-0213(01)00065-9
Mast, F. W., \& Kosslyn, S. M. (2002). Visual mental images can be ambiguous: insights from individual differences in spatial transformation abilities. Cognition, 86, 57-70. doi: 10.1016/S0010-0277(02)00137-3
O'Regan, J. K., \& Noë, A. (2001). A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences, 24, 973-1031. doi: 10.1017/s0140525x01000115

Kamermans, K. L., Pouw, W. T. J. L., Mast, F., \& Paas, F. (under review). Ambiguity detection in visual imagery is possible without visual cues.
Peterson, M. A., Kihlstrom, J. F., Rose, P. M., \& Glisky, M. L. (1992). Mental images can be ambiguous: reconstruals and reference-frame reversals. Memory \& Cognition, 20, 107-123. doi: 10.3758/BF03197159
Pouw, W. T. J. L., de Nooijer, J. A., van Gog, T., Zwaan, R. A., \& Paas, F. (2014). Toward a more embedded/extended perspective on the cognitive function of gestures. Frontiers in Psychology, 5, 359. doi: 10.3389/fpsyg.2014.00359

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. Behavioral and Brain Sciences, 25(2), 157-182. doi: 10.1017/S0140525X02000043
Sartre, J. P. (1940). The psychology of imagination. Academic Press: New york.
Spivey, M. J., \& Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. Psychological Research, 65, 235-241. doi: 10.1007/s004260100059
Spivey, M. J., Tyler, M., Richardson, D., \& Young, E. (2000). Eye movements during comprehension of spoken scene descriptions. In Proceedings of the 22nd annual conference of the Cognitive Science Society (pp. 487492).

Thomas, N. J. T. (1999). Are theories of imagery theories of imagination? An active perception approach to conscious mental content. Cognitive Science, 23, 207-245. doi: 10.1207/s15516709cog2302_3

# Categorical vs Coordinate Relationships do not reduce to spatial frequency differences 

Vishaal Prasad (v2prasad@ucsd.edu)<br>UCSD Department of Computer Science, 9500 Gilman Dr. La Jolla, CA 92093 USA<br>Ben Cipollini (ben.cipollini@gmail.com)<br>Classy.org, Diamond View Tower, 350 Tenth Ave<br>San Diego, CA 92101 USA<br>Garrison W. Cottrell (gary@ucsd.edu)<br>UCSD Department of Computer Science, 9500 Gilman Dr. La Jolla, CA 92093 USA


#### Abstract

Categorical and coordinate stimulus processing were hypothesized by Kosslyn (1987) to be lateralized visual tasks, differentiated by task-relevant spatial frequencies. Slotnick et al. (2001) directly tested Kosslyn's hypothesis and concluded that the lateralization presents only when tasks are sufficiently difficult. Our differential encoding model is a three layer neural network that accounts for lateralization in visual processing via the biologically plausible mechanism of differences in connection spread of long-range lateral neural connections (Hsiao, Cipollini, \& Cottrell, 2013). We show that our model accounts for Slotnick's data and that Slotnick's analysis does not convincingly explain their results. Instead, we propose that Kosslyn's initial hypothesis was based on an incorrect assumption: categorical and coordinate stimuli are not solely differentiated by spatial frequencies. The results that our model captures cannot be reproduced by Ivry and Robertson's (1998) Double Filtering by Frequency theory, which is driven solely by lateralized spatial frequency processing.


Keywords: Differential encoding; hemispheric asymmetry; spatial frequency processing; categorical vs. coordinate

## Introduction

The human brain is composed of two largely disconnected hemispheres that communicate via a bridge of neural connections known as the corpus callosum. The level of relative disconnection and redundancy suggests that, for some cognitive processes, it may be advantageous for each hemisphere to specialize and reduce redundancy. This functional specialization, or "lateralization," occurs in many diverse cognitive facilities in humans (Stephan et al., 2003) and non-humans alike (Rogers \& Andrew, 2002). Examples in humans include fine motor skills and language processing, both of which are left hemisphere dominant (Knecht et al., 2000). Of particular interest to us is visual lateralization. Past studies have shown visual lateralization in processing stimuli ranging from frequency gratings to facial recognition (e.g. Ivry \& Robertson, 1998; Sergent, 1985).

Navon's (1977) hierarchical letters are one visual stimulus showing lateralization. These consist of a large, "global" letter (e.g., "T") that is composed of small, "local" letters (e.g., "F"). Sergent (1982) showed an advantage in identifying the local level target (the letter F in the above example)
presented when flashed in the right visual field/ left hemisphere (RVF-LH), and the reverse (the letter T in the above example) in the left visual field/ right hemisphere (LVF-RH). She concluded that the LH performs better with high spatial frequency (HSF) stimuli, whereas the RH does better with low spatial frequency (LSF) stimuli. Kitterle, Christman, and Hellige (1990) directly tested this hypothesis with frequency gratings and showed that the lateralization was driven by task demands, rather than purely by stimulus properties.

Inspired by Sergent's (1982) theory, Ivry and Robertson (1998) proposed their Double Filtering by Frequency (DFF) theory to explain these asymmetric processing results. DFF theory proposes that the hemispheres identically first select the frequency bands relevant to the task, but then are biased so that the left hemisphere preferentially processes HSFs, and the right hemisphere LSFs. DFF theory also accounts for data suggesting that frequency processing differences between the hemispheres are not absolute, but instead are relative to the frequency band relevant for solving the task (Christman, Kitterle, and Hellige, 1991). Finally, their model accounts for the categorical and coordinate spatial relations proposed by Kosslyn (1987) and Kosslyn, Koenig, Barrett, Tang, and Gabrieli (1989), described below. However, there is no neurological basis for the core mechanisms of the DFF theory, nor is there a developmental explanation of how or why this phenomenon would emerge (see Cipollini, 2014 for further discussion).

Kosslyn (1987) and Kosslyn et al. (1989) argued that humans process visual stimuli using two distinct types of spatial relations. Coordinate relations rely on an absolute, metric basis; for example, the statement "the glass of water is 3 inches from my hand" defines a coordinate judgment of one's hand and the glass of water. In contrast, categorical relations rely on abstract, relative terms. The statement "the glass of water is on top of the table" does not tell us exactly where the glass is, only its relative position to a table. In his 1987 paper, Kosslyn observed a RVF-LH advantage for categorical relation judgments and a LVF-RH advantage for coordinate relation judgments in response time. Other work (e.g. Hellige \& Michimata, 1989) provided further support for this hy-
pothesis with more varied types of stimuli (e.g., a bar and dot stimulus).

However, Sergent (1991) found that this lateralization effect presented only when stimuli were degraded, and several analyses have noted that the lateralization only presents in right-handed people (Slotnick, Moo, Tesoro, \& Hart, 2001). Other researchers likewise found weak or inconclusive evidence for lateralization of categorical and coordinate stimuli, especially for the LH advantage on categorical stimuli (Okubo \& Michimata, 2002). Nevertheless, researchers generally agree that a distinction exists, even if it is weaker than originally thought (see Jager \& Postma, 2003 for a review).

In light of these conflicting data, Kosslyn and colleagues now argue that lateralization in categorical and coordinate stimuli results from a differential frequency processing, potentially based on lateralization in neuronal receptive fields (see Baker, Chabris, \& Kosslyn, 1998; Kosslyn, Chabris, Marsolek, \& Koenig, 1992). In support of this hypothesis, Okubo and Michimata (2002) showed that the RH coordinate advantage, but not the LH categorical advantage, was eliminated by contrast balancing, which removes low spatial frequency information without degrading stimuli.

## Slotnick et al. (2001) revisit Categorical / Coordinate

Of particular interest here are the experiments and results of Slotnick et al. (2001). In past experiments, researchers flashed stimuli in one visual field or the other, leaving room for interhemispheric interference. To isolate lateralization effects, Slotnick et al. ran a clinical study on 134 subjects, each of whom had at least one hemisphere temporarily deactivated by an intracarotid injection of sodium amobarbital.

Using the same stimuli as in Kosslyn et al. (1989), Slotnick et al. (2001) sought to reproduce their results in subjects with deactivated hemispheres. In addition, they added a new stimulus type, paired squares, which was designed to resist "categorization" of coordinate tasks, whereby a subject on later trials during an experiment learns a coordinate task (e.g., is the plus more than two inches from the minus) and turns it into a categorical task (Slotnick et al., 2001). This explanation had been proposed to explain the weakening of the RH advantage on coordinate stimuli. The paired squares coordinate stimulus forces the subject to make a direct metric comparison between the two parts of the stimulus.

Unlike the original paper, Slotnick et al. (2001) measured results by classification error, rather than reaction time. Though the results generally aligned with Kosslyn's hypothesis, one coordinate experiment did not show the expected RH dominance, instead showing lateralization opposite of that in the original paper. The authors noted that distances between components of their figures (e.g. the blob and dot), were larger in their experiments than in Kosslyn et al. (1989). Difficulty has been reported to modulate lateralization in other experiments (e.g. Sergent, 1985), and so they posited this made the task too easy to show proper lateralization. Consequently, post-hoc they stratified the tasks by difficulty and found that only when a task is sufficiently difficult does later-
alization arise as expected.

## Differential Encoding Model



Figure 1: Taken from Hsiao, Cipollini, and Cottrell (2013), this diagram shows the autoencoder models with varying connection spreads and symmetric connections. Notice the left hemisphere's hidden units connect to a more spread out set of neurons on average, while maintaining the same number of connections.

Competing with DFF theory is our Differential Encoding (DE) theory (Hsiao, Shahbazi, \& Cottrell, 2008; Hsiao, Cipollini, \& Cottrell, 2013). It is inspired by an anatomical difference in the auditory system's long range lateral connections (LRLCs). On average, a LH neuron connects to neighbors generally farther from itself than the RH neurons do (Galuske, Schlote, Bratzke, \& Singer, 2000). The DE model hypothesizes these LRLCs as the driving factor behind visual lateralization as well. Compared to the DFF theory, the DE model has the advantage of having neurodevelopmental and neuroanatomical plausibility (Cipollini, 2014).

Computationally, the Differential Encoding model is a standard 3-layer neural network which can be thought of as a recurrent neural network unrolled one step in time. The first set of connections is a sparse autoencoder, trained on natural images, to represent how a stimulus might be transformed in the early stages of the brain using low level processing such as Gabor filters. Each neuron in the autoencoder corresponds to a spatial location, and it connects to 5 other neurons generated randomly from a Gaussian centered around the neuron itself. The LH and RH networks vary by the standard deviation, or sigma parameter, of the Gaussian, to mimic the lateralized connection spread of the LRLCs, seen in Figure 1.

Note that this differs from a Gaussian receptive field. A network with Gaussian receptive fields has fixed connections, and the strength of the connections are determined by a Gaussian. In the DE model, the lateral connections themselves are stochastic and determined by randomly sampling from a Gaussian, and the strength of the connections are learned by training.

In Hsiao, Cipollini, and Cottrell (2013), the authors show that the autoencoder in the DE model reconstructs natural images in accordance with the predictions of Sergent (1982). Specifically, the RH model reconstructs low spatial frequency (LSF) components of a stimulus better, whereas the LH model reconstructs HSF components better.

Once trained, each hemisphere's hidden units are then con-


Figure 2: These are the stimuli from Slotnick et al. (2001). Note that paired squares only had a coordinate task, whereas blob/dot and plus/minus have both categorical and coordinate.
nected to a task-specific output unit that is trained by the delta rule to learn some task. In this way, the information represented by the hidden layer is tested as to what tasks it is best at. We have found in many experiments that the LH model is better at tasks that require HSFs, and vice-versa for the RH model (Hsiao, Cipollini, \& Cottrell, 2013).

In addition to the autoencoder properties outlined above, the model has accounted for Sergent's (1982) data, as well as Kitterle, Christman, and Hellige's (1990) data showing task dependence of lateralization (Hsiao, Cipollini, \& Cottrell, 2013). This suggests the DE model has the very frequency encoding properties that Sergent (1982) hypothesized. As Kosslyn and colleagues have suggested (e.g., Baker, Chabris, \& Kosslyn, 1999), the distinction between categorical and coordinate stimuli may stem from lateralized frequency processing. Therefore, we test the network on Slotnick et al. (2001)'s stimuli to further establish the relationship between our model and frequency lateralization, as well attempt to reach parity with the DFF on these stimuli.

## Materials and Methods

Our stimuli mimic those of Slotnick et al. (2001)
The stimuli used in the 2001 study can be seen in Figure 2. There are three types of stimuli: blob/dot, plus/minus, and paired squares. All three stimulus types involve coordinate tasks. Blob/dot requires an evaluation of how far apart the blob and dot are, and plus/minus likewise requires an evaluation of how far apart the plus and minus are. The paired squares task, in contrast, requires judging whether the two sets of paired squares are equidistant or not. The former two stimulus types also have categorical tasks. The blob/dot categorical task requires evaluating whether the dot is on the blob or off of it, and the plus/minus categorical task requires evaluating whether the plus is on the right or the left. There is no categorical task for paired squares.

For the plus/minus and blob/dot coordinate stimuli, "near" configurations were those where the distance between the plus and minus or blob and dot measured smaller than a reference distance of 2 inches; the "far" configurations were larger than 2 inches. In our model, the reference distance was 5.5 and 6 pixels for plus/minus and blob/dot respectively.

Slotnick et al. (2001) hypothesized that these tasks are harder when the distance between stimulus components are close to the reference distance of 2 inches. They defined "hard" configurations as those where the distance between
stimulus components fell within the range of $[1,3]$ in inches, i.e. within one inch from the reference distance of 2 inches. "Easy" configurations were those outside of this range. In our model, "hard" configurations were those where the distance fell within 2 pixels of the reference distance, and "easy" were the other stimuli. Stratifications for paired squares coordinate and the two categorical tasks were less principled and will be explored below.

## Simulation Procedure

The simulation was implemented in MATLAB. All code is open source ${ }^{1}$. Input images were implemented as bitmaps, following the images published in the original paper as best possible. To accomplish this, plus/minus and paired squares images were $34 \times 25$ pixels. Due to the need for increased resolution, blob/dot images were $68 \times 50$ pixels. Due to the varying resolutions, the experiment sizes had different hyperparameters. Of particular interest, the $34 \times 25$ images had a RH and LH standard deviation (sigma) of 4 and 10 pixels respectively; the $68 \times 50$ had 4 and 15 . In both cases, each neuron had five connections per hidden unit, with one hidden unit corresponding to each pixel of the image. Our train and test data were the same, so to prevent overfitting, we used heavy regularization. Specifically, we used a relatively high amount of dropout of 0.7 (Srivastava, Hinton, Krizhevsky, Sutskever, \& Salakhutdinov, 2014) and introduced noise on the input.

The human experiment used 100 LH subjects and 124 RH subjects; however, 54 hemispheres of patients deemed abnormal or otherwise compromised (e.g. those with parietal lobe tumors) were excluded (Slotnick et al., 2001). We followed the same analyses done in the human experiment, and in an attempt to roughly match statistical power, we instantiated each hemisphere in our computational model 100 times. Instantiations with outlier performance were discarded, so there were slightly fewer than 100 instantiations for the final results.

## Results and Discussion

The output of the DE model is a real-valued number between 0 and 1 , where 0 and 1 represent the target labels (e.g. off/on for categorical blob/dot). Error is measured as the sumsquared error (SSE) between the model's output and the true label. In Slotnick et al. (2001), they measure mean percent error in classification. The different measurements means $y$ -

[^443]

Figure 3: Differential Encoding (left) results follow the overall Slotnick et al. (2001) results for task x hemisphere interactions, which can be seen as the slope of each line. Note that since hemispheric performance, not absolute performance, was relevant, $y$-axes were re-scaled to emphasize slope.
axes are not directly comparable, but in this experiment we are concerned with the relative performance of hemispheres on each task. We simply compare the slopes in Figure 3 to see how well the DE model fits the human data.

From Figure 3, it is clear that the DE results do not follow the human data perfectly, but the key concepts are captured. Categorical blob/dot, coordinate blob/dot, and categorical plus/minus all showed the expected LH advantage. Coordinate plus/minus shows no LH advantage ( $F_{1,195}=0.0669$, $p>0.75$ ), and categorical plus/minus has a stronger LH lateralization than it ( $F_{1,391}=12.96, p<0.001$ ). Crucially, the anomalous result in the original paper persists: categorical blob/dot is not more LH-dominant than coordinate blob/dot ( $F_{1,396}=1.200, p>0.25$ ), as it was in Kosslyn et al. (1989).

Paired squares on first glance appears to have reverse lateralization as in the human data. However, paired squares was extremely volatile, and that advantage disappears or reverses spontaneously. Statistically, lateralization was nonsignificant: there are large error bars for paired squares, and the repeated-measures anova of this data reveals this same non-significance ( $F_{1,198}=0.115, p>0.9$ ). Slotnick et al. (2001) reported this stimulus had only marginal statistical significance ( $p<0.1$ ) in their results as well. Therefore, while we plan to investigate this stimulus further, for now, we are less concerned that it did not show LH dominance.

The middle and bottom rows in Figure 3 show the results of the human data and the DE model for easy and hard subsets respectively. Our model replicates the results well. Before further analyzing the results, we take a closer look at Slotnick
et al.'s (2001) difficulty stratification, to understand what relationships across difficulty are crucial to replicate.

## Revisiting Slotnick et al. (2001)

Slotnick et al. (2001) directly measured lateralization in subjects who had a hemisphere temporarily deactivated as part of a presurgical evaluation for treatment of intractable epilepsy. This meant the authors only ran their experiments once. When their data contradicted Kosslyn's results, they ran a post-hoc analysis of the data to explain their results. The crucial takeaway was that lateralization presented only if the task is difficult enough.

We find reasons to doubt Slotnick et al.'s conclusions. First, there are critical inconsistencies in their figures. If the easy and hard instances of a task both lateralize in the same direction, then combining all trials together should as well. Yet, as shown in Figure 3, coordinate plus/minus overall does not lateralize, though its easy and difficult subsets did. This discrepancy is never addressed in their paper.

In addition, most of the stratifications between easy and hard were not built in a principled manner and therefore lack validity. Slotnick et al. (2001) state that the stratification of the paired squares task was an empirical heuristic, as there was no neat way of differentiating easy and hard stimuli. Similarly, they note that no analogous concept of difficulty exists for categorical stimuli, so they just used the same division as their coordinate counterparts.

The coordinate blob/dot results are both internally consistent and well-principled, but the other tasks are not. The overall results of Slotnick et al. (2001), matched by our compu-

| Task | Blob/Dot | Plus/Minus | Paired Squares |
| :---: | :---: | :---: | :---: |
| Coordinate |  |  |  |
| Categorical |  |  | Not Applicable |

Figure 4: Frequency preferences of the DE model for each task. Y-axis is network accuracy, while x-axis marks the center of the frequency window on which a bandpass filter is applied. If spatial frequencies drive categorical and coordinate processing, we should see better performance on categorical tasks at HSFs and vice-versa at LSFs. We do not see this pattern.
tational modeling, are not adequately explained by difficulty stratification. We now look for an alternate explanation.

## Spatial frequency selectivity within the DE model

Kosslyn and colleagues (e.g. Baker, Chabris, \& Kosslyn, 1999) concluded that categorical and coordinate processing lateralized according to preferential frequency processing. As originally hypothesized in Sergent (1982), the LH is thought to outperform the RH in processing HSFs and vice-versa for LSFs. Numerous experiments have shown lateralization as a function of filtering stimuli to specific frequency windows (e.g. Sergent, 1985). The Differential Encoding model has also shown this same differential frequency processing (Hsiao, Cipollini, \& Cottrell, 2013).

To examine whether spatial frequency differences drove results on these five tasks, we tested the model with different bandpass filters for each task. Specifically, all networks trained on the same, unchanged image patches to learn the same features, simulating typical visual experience. However, the perceptron was trained and tested on stimuli run through a bandpass filter window of size four and eight CPI, for $34 \times 25$ and $68 \times 50$ images respectively. This would allow us to empirically deduce which frequencies the network best responded to. Results were agnostic to a host of parameter choices, including sigma, dropout, and bandpass width (within reason), so we believe the results are general to the task, and not specific to anything about our network setup.

Per Kosslyn's hypothesis, we expected to see lateralization in accordance with task type: there should be increased categorical performance on HSFs and likewise for coordinate and LSFs. Figure 4 shows the results. Coordinate paired squares
is almost parabolic with a minimum around 10 CPI . Coordinate plus/ minus is largely agnostic to frequencies, whereas categorical shows bimodal preference, with the global minimum at HSFs. Categorical blob/dot performed equally well at the windows centered from 10 to 18 CPIs, whereas coordinate performed best in the window centered at 10 CPIs and was locally parabolic around that area. Beyond roughly 20 CPI, critical image features are lost and in both cases the networks perform poorly.

There is no unified trend of stimulus type and frequency. Considering the DE model both captures the critical relationships outlined in Slotnick et al. (2001) and accounts for spatial frequency filtering (Hsiao, Cipollini, \& Cottrell, 2013), it appears something besides frequency underlies the categorical and coordinate dichotomy. We plan to explore this further via contrast balancing (Okubo \& Michimata, 2002).

## Conclusion

We show in this paper that the DE model both replicates human data on the categorical and coordinate dichotomy and doesn't behave strictly in accordance with spatial frequencies. This calls into question the hypothesis about spatial frequencies driving coordinate and categorical lateralization. It also provides a point of differentiation between DFF theory and the DE model. Limited to spatial frequency information, DFF would be unable to replicate the conflicting data in Slotnick et al. (2001) as the DE model does.

In addition, we have shown in other experiments that DE models of a larger sigma encode more information at higher spatial frequencies, and vice-versa for smaller sigmas. Furthermore, stimulus size mediates this relationship: as image
size increases, so too does the point where the LH networks outperform the RH networks. We are pursuing this as an explanation for the relative frequency effect, noted in Christman, Kitterle, and Hellige (2001).

The DE model already accounts for faces and the categorical and coordinate results. If we can account for the relative frequency effect, we have superseded the DFF with a model that is biologically grounded and is informative about experiments to run in biology and psychology.

## Acknowledgements

GWC is supported in part by NSF grant SMA 1041755 to the Temporal Dynamics of Learning Center and a gift from HP.

## References

Baker, D. P., Chabris, C. F., \& Kosslyn, S. M. (1999). Encoding categorical and coordinate spatial relations without input-output correlations: New simulation models. Cognitive science, 23, 33-51.
Christman, S., Kitterle, F. L., \& Hellige, J. (1991). Hemispheric asymmetry in the processing of absolute versus relative spatial frequency. Brain and Cognition, 16, 62-73.
Cipollini, B. (2014). Modeling visual lateralization and interhemispheric communication. Unpublished doctoral dissertation, University of California, San Diego.
Gallagher, P., \& Dagenbach, D. (2007). Manipulating noise frequencies alters hemispheric contributions to decision making. Brain and cognition, 64, 42-49.
Galuske, R. A., Schlote, W., Bratzke, H., \& Singer, W. (2000). Interhemispheric asymmetries of the modular structure in human temporal cortex. Science, 289, 19461949.

Hellige, J. B., \& Michimata, C. (1989). Categorization versus distance: Hemispheric differences for processing spatial information. Memory \& Cognition, 17(6), 770-776.
Hsiao, J. H., Cipollini, B., \& Cottrell, G. W. (2013). Hemispheric asymmetry in perception: A differential encoding account. Journal of Cognitive Neuroscience, 25, 9981007.

Hsiao, J. H.-w., Shahbazi, R., \& Cottrell, G. W. (2008). Hemispheric asymmetry in visual perception arises from differential encoding beyond the sensory level. In Proceedings of the Cognitive Science Society (Vol. 30).
Ivry, R. B., \& Robertson, L. C. (1998). The two sides of perception. Cambridge, MA: The MIT Press.
Jager, G., \& Postma, A. (2003). On the hemispheric specialization for categorical and coordinate spatial relations: A review of the current evidence. Neuropsychologia, 41, 504-515.
Kitterle, F. L., Christman, S., \& Hellige, J. B. (1990). Hemispheric differences are found in the identification, but not the detection, of low versus high spatial frequencies. Attention, Perception, \& Psychophysics, 48, 297-306.
Kitterle, F. L., Hellige, J. B., \& Christman, S. (1992). Visual hemispheric asymmetries depend on which spatial frequencies are task relevant. Brain and cognition, 20, 308-314.

Knecht, S., Drger, B., Deppe, M., Bobe, L., Lohmann, H., Flel, A., ... Henningsen, H. (2000). Handedness and hemispheric language dominance in healthy humans. Brain, 123, 2512-2518.
Kosslyn, S. M. (1987). Seeing and imagining in the cerebral hemispheres: A computational approach. Psychological review, 94, 148.
Kosslyn, S. M., Chabris, C. F., Marsolek, C. J., \& Koenig, O. (1992). Categorical versus coordinate spatial relations: computational analyses and computer simulations. Journal of Experimental Psychology: Human Perception and Performance, 18, 562.
Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., \& Gabrieli, J. D. (1989). Evidence for two types of spatial representations: hemispheric specialization for categorical and coordinate relations. Journal of experimental psychology: human perception and performance, 15, 723.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. Cognitive psychology, 9, 353-383.
Okubo, M., \& Michimata, C. (2002). Hemispheric processing of categorical and coordinate spatial relations in the absence of low spatial frequencies. Journal of cognitive neuroscience, 14, 291-297.
Rogers, L. J., \& Andrew, R. (2002). Comparative vertebrate lateralization. Cambridge University Press.
Rumelhart, D. E., Hinton, G. E., \& Williams, R. J. (1986). Learning representations by back-propagating errors. Na ture, 323, 533-536.
Sergent, J. (1982). The cerebral balance of power: confrontation or cooperation? Journal of Experimental Psychology: Human Perception and Performance, 8, 253.
Sergent, J. (1985). Influence of task and input factors on hemispheric involvement in face processing. Journal of Experimental Psychology: Human Perception and Performance, 11, 846.
Sergent, J. (1991). Judgments of relative position and distance on representations of spatial relations. Journal of Experimental Psychology: Human Perception and Performance, 17, 762.
Slotnick, S. D., Moo, L. R., Tesoro, M. A., \& Hart, J. (2001). Hemispheric asymmetry in categorical versus coordinate visuospatial processing revealed by temporary cortical deactivation. Journal of Cognitive Neuroscience, 13(8), 1088-1096.
Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., \& Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.
Stephan, K. E., Marshall, J. C., Friston, K. J., Rowe, J. B., Ritzl, A., Zilles, K., \& Fink, G. R. (2003). Lateralized cognitive processes and lateralized task control in the human brain. Science, 301(5631), 384-386.

# Refuting Overconfidence: <br> Refutation Texts Prevent Detrimental Effects of Misconceptions on Text Comprehension and Metacomprehension Accuracy in the Domain of Statistics 

Anja Prinz (anja.prinz@ezw.uni-freiburg.de)<br>Department of Educational Science, University of Freiburg<br>Rempartstraße 11, 79098 Freiburg, Germany<br>Stefanie Golke (stefanie.golke@ezw.uni-freiburg.de)<br>Department of Educational Science, University of Freiburg<br>Rempartstraße 11, 79098 Freiburg, Germany<br>Jörg Wittwer (joerg.wittwer@ezw.uni-freiburg.de)<br>Department of Educational Science, University of Freiburg<br>Rempartstraße 11, 79098 Freiburg, Germany


#### Abstract

Refutation texts are beneficial for removing misconceptions and supporting comprehension in science. Whether these beneficial effects hold true in the domain of statistics is, however, an open question. Moreover, the role of refutation texts for the accuracy in judging one's own comprehension (metacomprehension accuracy) has received little attention. Therefore, we conducted an experiment in which students with varying levels of statistical misconceptions read either a standard text or a refutation text in statistics, judged their text comprehension, and completed a comprehension test. The results showed that when students read the standard text, having more misconceptions resulted in poorer text comprehension and more inaccurate metacomprehension as indicated by overconfident predictions. In contrast, when students read the refutation text, the number of misconceptions was unrelated to text comprehension and metacomprehension accuracy. Apparently, refutation texts help students to pay attention to inaccuracies in their knowledge and, thereby, can promote self-regulated learning from texts.


Keywords: metacomprehension accuracy; misconceptions; procedural and conceptual understanding; text comprehension

In higher education, statistics has become a central part in many fields of study to enable students to deal with quantitative information (Ben-Zvi \& Garfield, 2004). At the same time, in higher education, students are increasingly expected to engage in self-regulated learning (Cassidy, 2011). For example, in statistics, students often need to advance their knowledge by reading statistics textbooks. However, such learning can be challenging, especially when students have to understand complex statistical concepts, such as covariance, about which they frequently have false ideas in the form of misconceptions. Such statistical misconceptions differ in fundamental ways from the normatively correct conceptions and, thus, can strongly hamper the comprehension and application of statistics (Liu, 2010).

## The Role of Misconceptions for Text Comprehension and Metacomprehension

Text comprehension is a process by which learners actively construct a mental representation of the information provided in a text (Kintsch, 1998). As usually not all possible relations are explicitly stated in a text, the construction of a mental representation requires learners to use their prior knowledge to infer within-text and knowledge relations (McNamara \& Magliano, 2009).

However, when learners possess inaccurate prior knowledge in the form of misconceptions, comprehension can be hampered because misconceptions can trigger false inferences. For example, Kendeou and van den Broek (2005) examined the online processes taking place when learners with misconceptions read texts. Their findings showed that learners with misconceptions used and integrated their prior knowledge with textual information as did learners without misconceptions. Yet, the content of their inferences was contaminated by their misconceptions. This, in turn, resulted in an inappropriate mental text representation after reading. In line with these findings, research in reading and science education shows that misconceptions often hinder memory and comprehension of text (see, e.g., Guzzetti et al., 1993).

When learning from reading texts, it is also important that learners accurately monitor and judge their own comprehension, which is known as metacomprehension accuracy (Dunlosky \& Lipko, 2007). Accurate metacomprehension affects the extent to which learners effectively self-regulate their learning. For example, only a learner who accurately monitors that a text is not yet sufficiently understood to perform well on a comprehension test might decide to further study the material (e.g., Thiede, Anderson, \& Therriault, 2003). Research, however, indicates that learners are often overconfident when monitoring their text comprehension. This is particularly true for learners' predictions, that is, their judgments of comprehension after
they have read a text but before they have taken a comprehension test (e.g., Maki et al., 2005).

Recently, Prinz, Golke, and Wittwer (2017) found that misconceptions produced overconfident predictions. More precisely, their results showed that statistical misconceptions not only impeded the comprehension of a statistics text but also led to inaccurate self-assessments of text comprehension as indicated by overconfident predictions. Apparently, when learners have misconceptions, they are likely to construct a flawed mental text representation. At the same time, when self-assessing their text comprehension, learners tend to focus on the amount of textual information they can retrieve from memory while neglecting whether this information is correct. Consequently, learners with misconceptions are likely to overestimate their actual comprehension (see also Dunlosky, Rawson, \& Middleton, 2005).

In sum, research indicates that learning from standard texts mainly elicits superficial understanding and monitoring when learners hold misconceptions. To overcome these difficulties, it is important that learners become aware of their misconceptions and revise their understanding, a process called conceptual change.

## The Role of Refutations for Text Comprehension and Metacomprehension

Conceptual change occurs when learners modify their existing prior knowledge to include new information (e.g., Chi, 2008). This requires that the existing prior knowledge is identified as inadequate and the new information is understandable, plausible, and useful (Posner et al., 1982). Usually, conceptual change is demanding because a strongly held misconception impedes the recognition of its inconsistency with the correct information provided in a text (Otero \& Kintsch, 1992). A promising instructional approach to inducing conceptual change is the use of refutation texts (Guzzetti et al., 1993). Refutation text passages typically comprise three elements: First, a commonly held misconception is described. Second, a cue that explicitly states that the misconception is in fact inaccurate is presented. Third, the scientifically correct explanation that directly refutes the misconception is provided (Tippett, 2010).

Research has shown that refutation texts in science domains are indeed more beneficial for restructuring incorrect prior knowledge than standard expository texts (e.g., Ariasi \& Mason, 2011; see also Guzzetti et al., 1993; Tippett, 2010). Studies that investigated the processes taking place when reading refutation texts found that refutation texts are effective because misconceptions and correct conceptions are presented in close proximity, thereby increasing the likelihood of simultaneous activation. Only after such coactivation, further conceptual change processes, like the experience of a cognitive conflict, the evaluation of one's current conceptions, and the establishment of coherence in one's knowledge, can take place (e.g., Ariasi \& Mason, 2011; van den Broek \& Kendeou, 2008). Whether refutation texts
prove effective in promoting conceptual change and enhancing comprehension also in statistics is not yet clear. Learning in statistics typically involves the acquisition of both concepts and procedures (Ben-Zvi \& Garfield, 2004). Therefore, it is also an open question whether refutation texts not only benefit the learning of statistical concepts but also the learning of statistical procedures.

With regard to metacomprehension, the role of refutation texts is largely under-researched. An exception is a study conducted by van Loon et al. (2015) that revealed no beneficial effects of refutation texts on monitoring accuracy because learners remained overconfident when predicting their comprehension. However, the texts about misconceptions used in the study were rather short and there was only one comprehension question per text. Therefore, predictions were related exclusively to the comprehension of information about a single misconception. Hence, it is unclear whether refutation texts would promote metacomprehension accuracy when judgments do not focus exclusively on misconceptions. In literature on conceptual change, it has often been theorized that refutation texts increase learners' metacognitive awareness of their own conceptions in relation to the scientific conceptions (e.g., Ariasi \& Mason, 2011). Thus, it seems plausible to assume that refutation texts can support learners in reflecting about their misconceptions, thereby increasing metacomprehension accuracy.

## The Present Study

We investigated to what extent a refutation text compensates for the detrimental impact of misconceptions on text comprehension and metacomprehension accuracy in the domain of statistics. More specifically, we focused on the topic of covariance and examined comprehension and metacomprehension accuracy with respect to both conceptual and procedural aspects of covariance.

The first research question addressed whether the type of text would moderate the effect of misconceptions on text comprehension. We expected that when reading a standard text, more misconceptions would lead to poorer conceptual and procedural text comprehension, whereas this relationship would not be apparent when reading a refutation text.

The second research question concerned whether the type of text would moderate the effect of misconceptions on metacomprehension accuracy. We hypothesized that when reading a standard text, more misconceptions would lead to greater overestimation of conceptual and procedural text comprehension, whereas this detrimental effect would not be apparent when reading a refutation text.

## Method

## Participants and Design

A total of $N=53$ university students ( $M=25.04$ years, $S D=$ $2.42,59 \%$ female) participated in this study. The study had
two independent variables. The first independent variable was categorical and referred to the type of text: Participants read either a standard text or a refutation text about the statistical concept of covariance. The second independent variable was metric and referred to the number of misconceptions about covariance. Dependent variables were text comprehension and metacomprehension accuracy referring to conceptual and procedural aspects of covariance.

## Material

The statistics text about covariance was adapted from a statistics textbook written by Bortz and Schuster (2010). This text existed in two versions: a standard text version and a refutation text version. Both versions addressed conceptual aspects of covariance such as its different directions, explained procedural aspects of covariance such as how it is calculated, provided the formula for computing covariance, and contained three graphs to illustrate positive, negative, and no covariance. In addition, the refutation text contained information challenging four common misconceptions about covariance (i.e., covariance implies causality, covariance is a standardized statistic, covariance is related to the slope of the fit line, and zero covariance proves the absence of any association; see, e.g., Prinz et al., 2017). More precisely, for each misconception, the three typical elements of a refutation text passage were provided: First, the misconception was described. Second, the incorrectness of the misconception was explicitly stated. Third, the scientifically correct explanation was given (Tippett, 2010). In contrast, the standard text only provided the scientifically correct explanation for each misconception. Without the graphs and the formula, the standard text included 515 words and the refutation text included 638 words. We did not equate the length of the two text versions to keep the manipulation unconfounded with other variations (e.g., the inclusion of additional or repetitive information in the standard text; cf., e.g., Diakidoy, Mouskounti, \& Ioannides, 2011).

## Measures

Misconceptions Misconceptions about covariance were assessed by 15 questions, with each question addressing one particular misconception. We collected these misconceptions on the basis of a comprehensive literature review (Prinz et al., 2017). For example, one question referred to the misconception that covariance does not depend on measurement units but represents a standardized statistic:

In a study, sports scientists from a university determined the covariance between the height and the time for a $100-\mathrm{m}$ dash of 20 sprinters. In his calculation, sports scientist A quantified time in seconds. When his colleague, sports scientist $B$, checks again, he quantifies time in milliseconds. Which of the following statements about the covariances calculated by the two sports scientists is correct?
$\square$ The two sports scientists will receive the same covariance because it does not matter if they use different measurement units (misconception).
$\square$ Both calculations will yield no covariance because one cannot calculate covariance from time data (wrong).
$\square$ No statement about the two covariances can be made because it is unknown if the variables time and height are linear (wrong).
$\square$ Sports scientist B will obtain a higher covariance than sports scientist A because milliseconds yield bigger numbers than seconds (correct).
All questions had a single-choice format with four response options. One option represented the correct answer, one option represented the particular misconception, and the two remaining options represented incorrect answers but not a particular misconception. The number of misconceptions was determined by counting how many times participants selected the response option that represented a misconception. Thus, they could record a maximum number of 15 misconceptions.

Text Comprehension Text comprehension referred to both conceptual and procedural comprehension of covariance. Conceptual comprehension was assessed by eight inference questions that had a single-choice format with four response options. Of the eight questions, four questions addressed misconceptions about covariance as already described. These were the four misconceptions that were targeted in the text. For these questions, one response option represented the correct answer, one response option represented the particular misconception, and the two remaining response options represented incorrect answers but not a particular misconception. Another four questions addressed further conceptual attributes of covariance but not specifically misconceptions. For these questions, one response option represented the correct answer and three response options represented incorrect answers but not a particular misconception. The participants received 1 point for each correct answer. Thus, they could achieve a maximum number of 8 points in the conceptual comprehension test.

Procedural comprehension was assessed by four openended questions that required the participants to perform calculations regarding covariance. They received 1 point for each correct answer. Thus, they could achieve a maximum number of 4 points in the procedural comprehension test. Interrater agreement on the procedural comprehension questions was high, Cohen's $\kappa=.98,95 \%$ CI [0.95, 1.00].

To facilitate the interpretation of participants' performance on the conceptual and procedural comprehension questions, we converted the number of conceptual and procedural comprehension questions correct into percent correct.

Metacomprehension Accuracy Before completing the comprehension questions, participants predicted the number of questions they would presumably answer correctly. They made their predictions for the conceptual and procedural
questions separately. Metacomprehension accuracy was calculated by taking the signed difference between participants' judged number of questions correct (converted into percent correct) and their actual number of questions correct (converted into percent correct). Hence, a positive value indicated overconfidence because participants would assume to answer more comprehension questions correctly than they actually did. For example, a value of +.10 means that participants assumed to provide $10 \%$ more correct answers to the questions than they actually did. In contrast, a negative value indicated underconfidence and a value of zero indicated a perfectly accurate judgment.

## Procedure

In the experiment, first, the participants completed the misconceptions test about covariance. Second, they accomplished a reading skills test serving as a filler task to remove the contents of the misconceptions test from working memory. Third, the participants read the statistics text about covariance. They were informed that their conceptual and procedural comprehension of the text would be tested after reading. Fourth, the participants predicted their conceptual and procedural text comprehension. To do so, they were informed about what kind of knowledge the two types of comprehension questions would require. Fifth, they answered the conceptual and procedural comprehension questions.

## Results

To statistically test our hypotheses, we performed multiple regressions. We centered all predictor variables to maintain meaningful estimates of the main effects. In case of a statistically significant interaction effect, we computed simple slopes analyses following the approach suggested by Richter (2007) to investigate the pattern of the interaction. According to this approach, the categorical predictor text type was dummy coded and entered in two complementary regression models to estimate the regression parameters. As before, the metric predictor number of misconceptions was entered in the regression models in centered form. When testing directional hypotheses, we used one-tailed tests. Table 1 displays descriptive statistics for misconceptions and the dependent variables as a function of text type.

Table 1: Descriptive statistics.

| Variable | Refutation text$(n=27)$ |  | Standard text$(n=26)$ |  | Total sample$(N=53)$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | M | SD | M | SD | M | SD |
| NoM | 4.74 | 1.66 | 4.15 | 1.38 | 4.45 | 1.54 |
| CC | . 62 | . 19 | . 65 | . 20 | . 63 | . 19 |
| PC | . 59 | . 32 | . 65 | . 34 | . 62 | . 33 |
| Accuracy CC | . 07 | . 19 | . 07 | . 18 | . 07 | . 19 |
| Accuracy PC | -. 10 | . 37 | -. 14 | . 31 | -. 12 | . 34 |

Note. $\mathrm{NoM}=$ number of misconceptions; $\mathrm{CC}=$ conceptual comprehension; $\mathrm{PC}=$ procedural comprehension.

The refutation text group and the standard text group did not significantly differ from each other with regard to the number of misconceptions, $t(51)=-1.40, p=.167, d=0.39$.

## Text Comprehension

As displayed in Table 2, the multiple regression with conceptual comprehension revealed a marginal significant main effect of misconceptions and a significant interaction effect between text type and misconceptions.

Table 2: Predictors of conceptual comprehension.

| Predictor | $b$ | $S E b$ | $t(49)$ | $p$ | $\Delta R^{2}$ |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Constant | 0.62 | 0.03 | 24.64 | $<.001$ |  |
| Text type | -0.01 | 0.05 | -0.18 | .857 | .01 |
| NoM | -0.03 | 0.02 | -1.88 | .066 | .03 |
| Text type x NoM | 0.09 | 0.03 | 2.73 | .005 | .13 |

Note. $R^{2}=.17, F(3,49)=3.26, p=.029 . \mathrm{NoM}=$ number of misconceptions.

Simple slopes analyses showed that, in the standard text group, the regression coefficient $b$ for number of misconceptions was $-0.08(S E=0.03)$ and significantly different from zero, $t(52)=-3.01, p=.002, \Delta R^{2}=.03$. This means that an increase of one misconception led to a decrease of $8 \%$ in conceptual text comprehension. In contrast, in the refutation text group, the regression coefficient for number of misconceptions was not significant, $b=0.01, S E=0.02, t(52)$ $=0.64, p=.527, \Delta R^{2}=.03$. Thus, there was no significant effect of misconceptions.

As shown in Table 3, the multiple regression with procedural comprehension also revealed a significant interaction effect between text type and misconceptions.

Table 3: Predictors of procedural comprehension.

| Predictor | $b$ | $S E b$ | $t(49)$ | $p$ | $\Delta R^{2}$ |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Constant | 0.61 | 0.05 | 13.38 | $<.001$ |  |
| Text type | -0.04 | 0.09 | -0.45 | .654 | .01 |
| NoM | -0.03 | 0.03 | -1.08 | .287 | .01 |
| Text type x NoM | 0.11 | 0.06 | 1.75 | .043 | .06 |

Note. $R^{2}=.08, F(3,49)=1.38, p=.260 . \mathrm{NoM}=$ number of misconceptions.

Simple slopes analyses showed that, in the standard text group, the regression coefficient $b$ for number of misconceptions was $-0.09(S E=0.05)$ and significantly different from zero, $t(52)=-1.85, p=.036, \Delta R^{2}=.01$. Thus, an increase of one misconception led to a decrease of $9 \%$ in procedural text comprehension. In contrast, in the refutation text group, the regression coefficient for number of misconceptions was not significant, $b=0.02, S E=0.04, t(52)$ $=0.51, p=.611, \Delta R^{2}=.01$, indicating that there was no significant effect of misconceptions.

## Metacomprehension Accuracy

As shown in Table 4, the multiple regression with metacomprehension accuracy of conceptual comprehension revealed a marginal significant interaction effect between text type and misconceptions.

Table 4: Predictors of metacomprehension accuracy of conceptual comprehension.

| Predictor | $b$ | $S E b$ | $t(49)$ | $p$ | $\Delta R^{2}$ |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Constant | 0.08 | 0.03 | 3.05 | .004 |  |
| Text type | -0.01 | 0.05 | -0.14 | .892 | $<.01$ |
| NoM | 0.02 | 0.02 | 0.84 | .406 | .01 |
| Text type x NoM | -0.05 | 0.04 | -1.35 | .092 | .04 |

Note. $R^{2}=.04, F(3,49)=0.73, p=.541$. NoM $=$ number of misconceptions.

Simple slopes analyses showed that, in the standard text group, the regression coefficient $b$ for number of misconceptions was $0.04(S E=0.03)$ and marginally significantly different from zero, $t(52)=1.43, p=.081, \Delta R^{2}$ $=.01$. Hence, an increase of one misconception resulted in $4 \%$ greater overestimation of conceptual text comprehension. In contrast, in the refutation text group, the regression coefficient for number of misconceptions was not significant, $b=-0.01, S E=0.02, t(52)=-0.38, p=.703, \Delta R^{2}=.01$. Thus, there was no significant effect of misconceptions.

As can be seen in Table 5, the multiple regression with metacomprehension accuracy of procedural comprehension revealed no significant main effect or interaction effect.

Table 5: Predictors of metacomprehension accuracy of procedural comprehension.

| Predictor | $b$ | $S E b$ | $t(49)$ | $p$ | $\Delta R^{2}$ |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Constant | -0.11 | 0.05 | -2.28 | .027 |  |
| Text type | 0.01 | 0.10 | 0.13 | .897 | $<.01$ |
| NoM | 0.03 | 0.03 | 1.05 | .300 | .02 |
| Text type x NoM | -0.06 | 0.06 | -0.94 | .177 | .02 |

Note. $R^{2}=.03, F(3,49)=0.56, p=.643 . \mathrm{NoM}=$ number of misconceptions.

## Discussion

A large body of literature demonstrates positive learning effects from reading refutation texts compared with reading standard expository texts. However, little is known about whether refutation texts are also favorable for learning in statistics and for producing accurate self-assessments of comprehension. Thus, the present study is the first to address these questions.

First, the results showed that a refutational statistics text can compensate for the detrimental impact of misconceptions on text comprehension. When students read a standard text about covariance, a higher number of misconceptions about covariance led to poorer conceptual and procedural
comprehension of the text. In contrast, when students read a refutation text about covariance, there was no effect of misconceptions on text comprehension. This result extends prior research by showing that refutation texts can prevent the adverse impact of misconceptions in statistics as well.

Importantly, the beneficial effect of the refutation text was demonstrated for both conceptual and procedural comprehension. Research in mathematics education widely acknowledges the view that conceptual and procedural knowledge are iteratively related to each other, with increases in conceptual knowledge leading to subsequent increases in procedural knowledge and vice versa (e.g., Rittle-Johnson \& Schneider, 2015). Accordingly, in the present study, conceptual and procedural comprehension were quite strongly associated, $r=.46, p=.001$. Therefore, when students read the refutation text, their misconceptions were no longer predictive of both their acquisition of conceptual understanding and procedural skill.

Second, the findings showed that a refutational statistics text can compensate for the detrimental impact of misconceptions on metacomprehension accuracy with regard to conceptual comprehension. When students read a standard text about covariance, a higher number of misconceptions about covariance led to greater overestimation of conceptual comprehension. In contrast, when students read a refutation text about covariance, there was no significant effect of misconceptions on the accuracy with which they judged their conceptual comprehension. In accordance with the interpretation given by Prinz et al. (2017; see also Dunlosky et al., 2005), when reading a standard statistics text, students with a higher number of misconceptions likely constructed a flawed mental text representation. At the same time, when self-assessing their text comprehension, these students might have focused on the amount rather than on the correctness of the textual information they could access from memory. Accordingly, they might have more strongly overestimated their conceptual comprehension. However, when reading a refutational statistics text, the students might have been more inclined to assess the quality of the textual information they could retrieve from memory. This might have been the case because refutation texts promote the coactivation of misconceptions and the scientifically correct conceptions (van den Broek \& Kendeou, 2008), thereby increasing the likelihood of knowledge evaluation and reflection in the context of conceptual change processes. Note, however, that the interaction effect between text type and misconceptions as well as the regression slope of misconceptions in the standard text group only approached the $10 \%$ level of statistical significance. As suggested by power analysis, this likely is the result of insufficient power to detect rather small effects. This is also supported by the findings of Prinz et al. (2017) that revealed a negative effect of misconceptions on metacomprehension accuracy of conceptual comprehension in the case of a standard text when using a sample of 49
participants. Therefore, future research should replicate the findings presented here while using larger sample sizes.

Contrary to expectation, however, the type of text and misconceptions did not affect metacomprehension accuracy of procedural comprehension. It can be assumed that the refutation text failed to coactivate procedural comprehension and, thus, decreased the likelihood that students would closely evaluate this type of comprehension. Yet, online measures such as think-aloud protocols could help to clarify the mechanisms proposed to underlie the effects observed in this study.

In sum, this study showed that refutation texts can compensate for detrimental effects of misconceptions on text comprehension and metacomprehension accuracy in the domain of statistics. Refutation texts appear to promote students to pay attention to inaccuracies in their knowledge, enhancing their self-regulated learning.

## Acknowledgements

This research was supported by grants from the German Federal Ministry of Education and Research (01JA1518A).

## References

Ariasi, N., \& Mason, L. (2011). Uncovering the effect of text structure in learning from a science text: An eye-tracking study. Instructional Science, 39, 581-601.
Ben-Zvi, D., \& Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi \& J. Garfield (Eds.), The challenge of developing statistical literacy, reasoning and thinking. Dordrecht, the Netherlands: Kluwer Academic.
Bortz, J., \& Schuster, C. (2010). Statistik für Human- und Sozialwissenschaftler [Statistics for human and social scientists] (7th ed.). Berlin, Germany: Springer.
Cassidy, S. (2011). Self-regulated learning in higher education: Identifying key component processes. Studies in Higher Education, 36, 989-1000.
Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), International handbook of research on conceptual change. New York, NY: Routledge.
Diakidoy, I.-A. N., Mouskounti, T., \& Ioannides, C. (2011). Comprehension and learning from refutation and expository texts. Reading Research Quarterly, 46, 22-38.
Dunlosky, J., \& Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. Current Directions in Psychological Science, 16, 228-232.
Dunlosky, J., Rawson, K. A., \& Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. Journal of Memory and Language, 52, 551-565.
Guzzetti, B. J., Snyder, T. E., Glass, G. V., \& Gamas, W. S. (1993). Promoting conceptual change in science: A
comparative meta-analysis of instructional interventions from reading education and science education. Reading Research Quarterly, 28, 116-159.
Kendeou, P., \& van den Broek, P. (2005). The effects of readers' misconceptions on comprehension of scientific text. Journal of Educational Psychology, 97, 235-245.
Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge, UK: Cambridge University Press.
Liu, T.-C. (2010). Developing simulation-based computer assisted learning to correct students' statistical misconceptions based on cognitive conflict theory, using "correlation" as an example. Journal of Educational Technology \& Society, 13, 180-192.
Maki, R. H., Shields, M., Wheeler, A. E., \& Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. Journal of Educational Psychology, 97, 723-731.
McNamara, D. S., \& Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), The psychology of learning and motivation: Advances in research and theory. San Diego, CA: Elsevier.
Otero, J., \& Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. Psychological Science, 3, 229-235.
Posner, G. J., Strike, K. A., Hewson, P. W., \& Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education, 66, 211-227.
Prinz, A., Golke, S., \& Wittwer, J. (2017). The double curse of misconceptions: Misconceptions impair text comprehension and metacomprehension accuracy in the domain of statistics. Manuscript submitted for publication.
Richter, T. (2007). How to analyze interactions of metric and categorical predictors: Not with median splits! Zeitschrift für Medienpsychologie, 19, 116-125.
Rittle-Johnson, B., \& Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. C. Kadosh \& A. Dowker (Eds.), The Oxford handbook of numerical cognition. Oxford, UK: Oxford University Press.
Thiede, K. W., Anderson, M. C. M., \& Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. Journal of Educational Psychology, 95, 66-73.
Tippett, C. D. (2010). Refutation text in science education: A review of two decades of research. International Journal of Science and Mathematics Education, 8, 951-970.
Van den Broek, P., \& Kendeou, P. (2008). Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions. Applied Cognitive Psychology, 22, 335-351.
Van Loon, M. H., Dunlosky, J., van Gog, T., van Merriënboer, J. J. G., \& de Bruin, A. B. H. (2015). Refutations in science texts lead to hypercorrection of misconceptions held with high confidence. Contemporary Educational Psychology, 42, 39-48.

# Varieties of Numerical Estimation: A Unified Framework 

Jike Qin (qin.284@osu.edu)<br>Department of Psychology, 1835 Neil Avenue Columbus, OH 43210 USA

Dan Kim (kim.3839@osu.edu)<br>Department of Psychology, 1835 Neil Avenue<br>Columbus, OH 43210 USA

John Opfer (opfer.7@osu.edu)
Department of Psychology, 1835 Neil Avenue Columbus, OH 43210 USA


#### Abstract

There is an ongoing debate over the psychophysical functions that best fit human data from numerical estimation tasks. To test whether one psychophysical function could account for data across diverse tasks, we examined 40 kindergartners, 38 first graders, 40 second graders and 40 adults' estimates using two fully crossed $2 \times 2$ designs, crossing symbol (symbolic, non-symbolic) and boundedness (bounded, unbounded) on free number-line tasks (Experiment 1) and crossing the same factors on anchored tasks (Experiment 2). Across all 8 tasks, $88.84 \%$ of participants provided estimates best fit by a mixed log-linear model, and the weight of the logarithmic component ( $\lambda$ ) decreased with age. After controlling for age, the $\lambda$ significantly predicted arithmetic skills, whereas parameters of other models failed to do so. Results suggest that the logarithmic-to-linear shift theory provides a unified account of numerical estimation and provides uniquely accurate predictions for mathematical proficiency.


Keywords: cognitive development; numerical cognition; number-line estimation; psychophysical function

## Introduction

In this paper, we aimed to address an ongoing debate on the psychophysical functions that link numbers to their magnitude estimates and to provide a unified framework for understanding seemingly-conflicting data from a variety of studies (Barth \& Paladino, 2011; Cohen \& Sarnecka, 2014; Opfer, Thompson, \& Kim, 2016; Siegler \& Opfer, 2003; Slusser, Santiago, \& Barth, 2013). Specifically, we sought to test whether models that fit data from old research methods could accurately predict data from new methods that differed in small increments that were thought to be psychologically meaningful. Finally, we aimed to test whether models that best accounted for numerical magnitude estimates also provided the best predictors of educational outcomes.

The classic theory about developmental change in numerical estimation is that the representation of numerical magnitudes follows a logarithmic-to-linear shift (Siegler \& Opfer, 2003; Siegler, Thompson, \& Opfer, 2009). Because this shift occurs for different numbers at different times (e.g., for $0-100$ number-lines before $0-1000$ number-lines), this change is thought to come from experiences that children have in school with symbolic numbers (Siegler \& Opfer,
2003). This account was originally based on a single version of the number-line task, which is the symbolic bounded free branch in the taxonomy in Figure 1.


Figure 1: Taxonomy of number-line tasks. Branches connected by solid lines were examined in previous studies. Ones connected with dashed lines are new.

Two alternative accounts were recently proposed. One is the proportional-judgment account, claiming participants adopt proportion judgment strategies when estimating numerical magnitudes (Barth \& Paladino, 2011; Slusser et al., 2013). The other is the measurement-skills account, claiming that data from number-line tasks arise from task-specific measurement skills (Cohen \& Blanc-Goldhammer, 2011; Cohen \& Sarnecka, 2014). Like the classic theory, these accounts also depended only on specific sets of number-line tasks (symbolic bounded anchored for proportional-judgment account; symbolic unbounded free for measurement-skills account (see Figure 1).

## 1. Symbolic vs. Non-symbolic

One potentially important variable is whether numerical magnitudes are presented symbolically or non-symbolically. For the log-to-linear shift account, this variable is important because symbolic numbers are presented to children in number-lines in school, on rulers, and in arithmetic lessons. Most studies have focused only on the symbolic magnitude estimates, though with different psychophysical functions being proposed (Barth \& Paladino, 2011; Cohen \& Sarnecka,

2014; Opfer et al., 2016; Siegler \& Opfer, 2003; Slusser et al., 2013).

In contrast, when Dehaene et al. (2008) presented Amazon indigene with non-symbolic numeric magnitudes, they found that a mixed log-linear model (MLLM, see Figure 2A) provided a better fit to number-line estimates than alternatives. Among these alternatives, however, the power models proposed by Slusser et al. (2013) and Cohen \& BlancGoldhammer (2011) were not included.

## 2. Bounded vs. Unbounded

Another potentially important feature of number-line estimation is whether an upper endpoint is provided (bounded) or not (unbounded). Like symbols, the use of a numeric upper bound may make the task easier because it provides an additional reference point against which to estimate the target number. On the other hand, Cohen and his colleagues have claimed that the unbounded task is actually easier because subjects need only to add incremental units, whereas the bounded task requires subtracting from the endpoint at the upper bound. For this reason, they suggested that extensions of cyclic models (CPMs, see Figure 2C) provide best fitting models for estimates in the bounded condition and that scallop power models (SPMs, see Figure 2D) provides best fitting for estimates in the unbounded condition (Cohen \& Blanc-Goldhammer, 2011; Cohen \& Sarnecka, 2014). Though Cohen and colleagues did not include the mixed log-linear model among the alternatives tested, Kim and Opfer (in press) found the MLLM was a better predictor of estimates than CPMs and SPMs for symbolic bounded free and unbounded free number-line tasks.

## 3. Free vs. Anchored

A third potentially important variable is whether subjects are given the numeric magnitude of the half-way point on the number-line (anchored) or not (free). Like the use of an upper bound, Opfer et al. (2016) have argued that the anchored task provides an additional reference point that should increase linearity of estimates. On the other hand, Slusser et al. (2013) have argued that the task reveals changes in proportional reasoning, and they showed that children's symbolic bounded anchored number-line estimates were better fit by one of three adapted cyclical power models (CPMs) (Hollands \& Dyre, 2000) than a simple logarithmic model. Subsequent studies, however, found the MLLM provided a better fit to both symbolic bounded free and symbolic bounded anchored number-line estimates than mixtures of the CPMs (Opfer et al., 2016), which was called MCPM1 (see Figure 2B) in Kim and Opfer (in press)'s study.

## The Current Study

In this study, we manipulated all three variables orthogonally to systematically test the mixed log-linear model against its competitors on 4 previously examined tasks and 4 novel tasks. Thus, we tested all the branches shown in Figure 1,
with symbolic bounded free (SBF), symbolic unbounded free (SUF), non-symbolic bounded free (NBF), non-symbolic unbounded free (NUF) tasks in Experiment 1 and symbolic bounded anchored (SBA), symbolic unbounded anchored (SUA), non-symbolic bounded anchored (NBA), nonsymbolic unbounded anchored (NUA) tasks in Experiment 2. At the end of Experiment 2, we also administrated a battery of math tests, including addition and subtraction, to each subject to determine which model parameters best predicted addition and subtraction proficiency. This issue has educational significance, but it also tests the key cognitive process claim of the measurement-skills account, viz. that unbounded number-line estimates are easier than the bounded ones because they require addition skills rather than subtraction skills.


Figure 2: Illustrations of predicted estimates from the mixed log-linear model (A), the mixed cyclic power model 1 (B), the mixed cyclic power model 2 for bounded condition ( C ) and the mixed scallop power model for unbounded condition (D).

## Experiment 1: Free Numerical Estimation

## Methods

Participants Participants were 40 kindergartners ( $\mathrm{M}=5.98$ years; $47.5 \%$ female), 38 first-graders ( $\mathrm{M}=7.13$ years; $50 \%$ female), 40 second-graders ( $\mathrm{M}=8.09$ years; $57.5 \%$ female) and 40 adults ( $\mathrm{M}=20.1$ years; $50 \%$ female).
Materials and procedure Participants were administered four different number-line tasks using a 2 (symbolic/non-
symbolic) by 2 (bounded/unbounded) fully-crossed design Order of tasks was determined by a balanced Latin square.

In symbolic conditions, participants were presented with 20 number-lines, with a number on each endpoint of the line. The to-be-estimated numerals were evenly sampled from 0 to 30. On each trial, numbers were shown 2 s followed by random-noise mask. In non-symbolic conditions, procedure was similar, except that endpoints of lines and to-beestimated numbers were dot arrays. Sizes of dots were controlled on $50 \%$ of trials, while areas covered by dots were controlled on the other $50 \%$.

In bounded conditions, endpoints of the line were 0 and 30 (symbolic condition) or 0 and 30 dots (non-symbolic condition). In the unbounded condition, endpoints were 0 and 1 (symbolic condition) or 0 and 1 dot (non-symbolic condition). The instructions for the unbounded condition were taken from Cohen and Sarnecka (2014).

## Results

1. Logarithmic-to-linear-shift theory accurately predicted median estimates and individual differences. We first fit median estimates for all four number-line tasks and age groups using MLLM. Across all tasks and age groups (Figure 3), fit of MLLM was very high ( $\mathrm{R}^{2}=.93 \sim 1$ ). Analyses of the weight of logarithmic component ( $\lambda$ ) revealed that with age, estimates changed from logarithmic patterns to linear ones, with $\lambda$ decreasing from kindergartners to adults across all tasks (Figure 3). As expected, $\lambda$ in non-symbolic conditions was higher than in symbolic ones. Also, $\lambda$ in unbounded conditions was higher than in bounded ones regardless of symbolic format, which argues against the view that "the unbounded task requires less mathematical sophistication than the bounded task does" (Cohen and Sarnecka, 2014). To test whether individual performance revealed the same pattern, we computed $\lambda$ for individual participants' data and conducted a mixed ANOVA, with symbolic format and boundedness as within-participant factors and age group as a between-participant factor. Results showed a main effect of symbolic format, $F(1,154)=74.19$, $p<.001$, boundedness, $F(1,154)=86.32, p<.001$, and age group, $F(3,154)=39.08, p<.001$. An interaction between symbolic format and boundedness, $F(1,154)=4.17, p<.05$, indicated that the effect of symbols was greater for the bounded tasks.

To test whether logarithmicity of estimates represented a stable pattern of individual differences, we correlated individual participants' $\lambda$ among all tasks. Results showed that individual participant's $\lambda$ among all the four number line tasks positively correlated (with correlation coefficient . 70 ( $p<.001$ ) between SBF and SUF tasks; .45 ( $p<.001$ ) between SBF and NBF tasks; 35 ( $p<.001$ ) between SBF and NUF tasks; .49 ( $p<.001$ ) between SUF and NBF tasks; .39 between SUF and NUF tasks; and $.54(p<.001)$ between NBF and NUF tasks).


Figure 3: Median estimates on 0-30 free number lines for different age groups.

Table 1: Percent of participants best fit by MLLM for free number line tasks. K, kindergartners; 1, first graders; 2, second graders; A, adults.

|  | MLLM |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
|  | K | 1 | 2 | A | All |
| SBF | 95 | 89 | 83 | 68 | 84 |
| NBF | 95 | 95 | 93 | 90 | 93 |
|  |  |  |  |  |  |
| SUF | 100 | 100 | 100 | 100 | 100 |
| NUF | 100 | 100 | 100 | 100 | 100 |

Table 2: Partial correlation between $\lambda$ in MLLM and math score after controlling for age across the free number line tasks.

| MLLM |  | Partial correlation |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Addition |  | Subtraction |  |
|  | SBF | $\lambda$ | -. 40 | *** | -. 27 | *** |
|  | NBF | $\lambda$ | -. 27 | *** | -. 19 | * |
|  | SUF | $\lambda$ | -. 36 | *** | -. 24 | ** |
|  | NUF | $\lambda$ | -. 17 | * | -. 17 | * |

2. Model comparison. We next compared the fit of MLLM to that of its competitors: MLLM vs MCPM1 and MCPM2 for the bounded conditions and MLLM vs MSPM for the
unbounded ones. The proportion of individual children who were best fit by the mixed log-linear model (MLLM) using AICc was calculated.

As illustrated in Table 1, estimates of $68 \%$ to $100 \%$ of participants were best fit by mixed log-linear model (MLLM) among all four free number line tasks. In the bounded condition, none of the MCPMs was the best fitting model for the majority of any age or task combination. In unbounded condition, we replicated the findings from Kim and Opfer (in press), with MLLM providing a better fit for $100 \%$ of participants' estimates compared to MSPM.
3. Predicting the mathematical performance. We next conducted partial correlation analysis between individual participant's addition and subtraction performance (which were tested in Experiment 2) and the best-fitting parameter values from the models when controlling for age. The addition score was the sum score of simple and complex addition problems, and the subtraction score was the sum score of simple and complex subtraction problems.

As shown in Table 2, the logarithmicity parameter $\lambda$ of the MLLM predicted both addition and subtraction performance across all tasks after controlling for age. In contrast, the correlations among the model parameters of the MLLM competitors were very small, inconsistent, and not expected by the theories that generated the models. Specifically, for bounded conditions, the negative correlation between the absolute value of $\beta \mathrm{s}-1$ of the MCPMs and math performance was found in only a few of number line tasks, with the absolute value of $\beta_{2 \mathrm{CPM}^{-}} 1$ of MCPM1 negatively correlating with addition for the symbolic bounded free number line task $(r(156)=-.18, p<.05)$, the absolute value of $\beta_{\mathrm{SBCM}^{-1}}$ of MCPM2 negatively correlating with addition and subtraction for the non-symbolic bounded free task $(r(156)=-.17, p<.05$ for addition; $r(156)=-.17, p<.05$ for subtraction). Also, only the absolute value of s-1 in MCPM2 negatively correlating with subtraction was found in symbolic bounded free task $(r(156)=-.17, p<.05)$. For the unbounded condition, the negative correlation between the absolute value of $\beta_{\mathrm{MSPM}}-1$ in MSPM and addition was only found in symbolic unbounded free task ( $r(156)=-.18, p<.05$ ). These finding suggests that MLLM uniquely predicts math performance, regardless of tasks or age groups.

## Experiment 2: Anchored Numerical Estimation

## Methods

Participants Participants in Experiment 2 were the same as in Experiment 1.
Materials and procedure Participants received the same 2 (symbolic/non-symbolic) by 2 (bounded/unbounded) number line tasks as in Experiment 1, except that information was given about the location of 15 (or 15 dots) in each of the four tasks. Order of tasks followed a Latin square. After that, 200 arithmetic problems were presented for participants to solve as quickly as possible: simple addition, simple subtraction, complex addition and complex subtraction. For simple
addition problems, each of the addends was a one-digit number and the sum was no more than 10 (e.g., $5+3,2+1$ ). For simple subtraction problems, the difference was less than 10 and both minuend and subtrahend were one-digit numbers (e.g., 9-3, 8-2). For complex addition problems, sums were bigger than 10 but less than 30 , and addends were one- or two-digit numbers (e.g., $4+16,14+15$ ). For complex subtraction problems, differences were bigger than 10 but less than 30 , with the minuend a two-digit number and the subtrahend one- or two-digit numbers (e.g., 16-5, 25-11).

## Results

1. Logarithmic-to-linear-shift theory accurately predicted median estimates and individual differences. We first fit the median estimates for all four number line tasks and age groups using MLLM. As shown in Figure 4, across all tasks and age groups, the fit of MLLM was uniformly high $\left(R^{2}=.93 \sim 1\right)$. Analyses of $\lambda$ revealed that with age, estimates changed from logarithmic patterns to linear ones, with $\lambda$ decreasing from kindergartners to adults (Figure 4). As with Experiment $1, \lambda$ in non-symbolic conditions were higher than in symbolic ones, and $\lambda$ in unbounded conditions were higher than in bounded conditions regardless of symbol. We also computed $\lambda$ for individual participants' data. The mixed ANOVA results again showed a main effect of symbolic format, $\quad F(1,154)=83.17, \quad p<.001$, boundedness, $F(1,154)=21.20, p<.001$, and age group, $F(3,154)=19.63$, $p<.001$.

To test whether the logarithmic-to-linear-shift theory could also capture individual differences, we correlated individual participant's $\lambda$ among tasks. The results showed that individual participant's $\lambda$ among all the four number line tasks positively correlated (with correlation coefficient .81 ( $p<.001$ ) between SBA and SUA tasks; $.61(p<.001)$ between SBA and NBA tasks; . 48 ( $p<.001$ ) between SBA and NUA tasks; . $54(p<.001)$ between SUA and NBA tasks; . 43 ( $p<.001$ ) between SUA and NUA tasks; and .61 ( $p<.001$ ) between NBA and NUA tasks.
2. Model comparison. We next examined whether MLLM is the best model compared to other competitors. According to the previous studies (Cohen \& Sarnecka, 2014; Opfer et al., 2016; Slusser et al., 2013), we compared the fit of MLLM, MCPM1, and MCPM2 on individual data for the bounded condition (which included SBA and NBA tasks). Since the unbounded anchored number-line tasks were new in this study, we compared the fit of all the four models for the unbounded condition (which included SUA and NUA tasks). The proportion of individual children who were best fit by the mixed log-linear model (MLLM) using AICc was calculated.

As illustrated in Table 3, the estimates of $63 \%$ to $100 \%$ of participants were best fit by mixed log-linear model (MLLM) among all four anchored tasks across all age groups. Specifically, in the bounded condition, no matter what types of symbol were given, against to the proportional account and subtraction or division-skill account, none of the MCPMs was the best fitting model for the majority. In the unbounded condition, our results showed that estimates of $65 \%$ to $98 \%$


Figure 4: Median estimates on 0-30 anchored number lines for different age groups.

Table 3: Percent of participants best fit by MLLM for anchored number line tasks. K, kindergartners; 1, first graders; 2, second graders; A, adults.

|  | MLLM |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
|  | K | 1 | 2 | A | All |
| SBA | 85 | 63 | 63 | 68 | 70 |
| NBA | 93 | 97 | 100 | 93 | 96 |
|  |  |  |  |  |  |
| SUA | 88 | 74 | 68 | 65 | 73 |
| NUA | 98 | 97 | 95 | 93 | 96 |

Table 4: Partial correlation between $\lambda$ in MLLM and math score after controlling for age across the anchored number line tasks.

|  |  | Partial correlation |  |  |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :---: | :---: | :---: | :---: |
|  |  |  | Addition |  |  |  |  |  |  | Subtraction |
| MLLM |  |  |  |  |  |  |  |  |  |  |
|  | SBA | $\lambda$ | -.29 | $* * *$ | -.20 | $*$ |  |  |  |  |
|  | NBA | $\lambda$ | -.22 | $* *$ | -.15 |  |  |  |  |  |
|  | SUA | $\lambda$ | -.30 | $* * *$ | -.22 | $* *$ |  |  |  |  |
|  | NUA | $\lambda$ | -.24 | $* *$ | -.20 | $*$ |  |  |  |  |

Note. ${ }^{*} p<.05,{ }^{* *} p<.01,{ }^{* * *} p<.001$
of participants were best fitting by MLLM when compared the fitting of all the four models. All these results suggest the logarithmic-to-linear-shift account for all the anchored numerical magnitude representation, regardless of boundedness or symbolic format.
3. Predicting the mathematical performance. Similar with Experiment 1, we also conducted partial correlation analysis
between individual participants' addition and subtraction performance and the best-fitting parameter values from the models when controlling for age. As shown in Table 4, $\lambda$ in the MLLM predicted both addition and subtraction performance across almost all the anchored number line tasks after controlling for age. However, for bounded condition, the negative correlation between the absolute value of $\beta \mathrm{s}-1$ of the MCPMs and math performance was only found for the symbolic bounded anchored task, with absolute value of $\beta_{2 \text { CPM }^{-1}}$ in MCPM2 negatively correlating with addition and subtraction ( $r(156)=-.26, p<.001$ for addition; $r(156)=-.26$, $p<.001$ for subtraction). The finding suggests that MLLM uniquely predict math performance, regardless of tasks or age groups.

## Discussion

Our experiments indicate that the logarithmic-to-linear shift account provides a unified framework that can account for data coming from a broad array of numerical estimation tasks. Specifically, we found a mixed log-linear model was the best fitting model for the vast majority ( $88.84 \%$ ) of children and adults. This finding held regardless of whether the symbolic format was symbolic or non-symbolic, whether the task was bounded or unbounded, and whether an additional reference was given or not. These results replicate those reported in Opfer et al. (2016) and Kim and Opfer (in press), as well as extending them to 4 novel number-line tasks. Finally, we found that with education, individuals acquire more substantial prior experience with symbolic numbers than non-symbolic dots, with more logarithmic compression shown in non-symbolic than symbolic condition. Also, the additional reference points (either midpoints or bounded endpoints) can increase the linearity of estimates. Thus, the classic number line task (SBF) is not an outlier in eliciting logarithmic pattern of estimates. Of the eight tasks, the highest logarithmicity was observed in kindergartners' estimating non-symbolic unbounded free (NBF) task and the lowest was in SBA.

Our results also showed that the logarithmic weight $(\lambda)$ was not fixed, but depended on the developmental history and prior experiences of the subject, leading to lower $\lambda$ values from kindergartners to adults. These findings met the overarching principle of the logarithmic-to-linear shift theory, which holds that the representation of numerical magnitude will change from the logarithmic pattern to linear one with age and experience (Opfer et al., 2011; Opfer \& Siegler, 2007; Siegler \& Booth, 2004; Siegler \& Opfer, 2003; Thompson \& Opfer, 2008).

Finally, individual differences were stable across the eight tasks: children whose estimates were more logarithmic in one task were also more logarithmic in the other seven tasks, $r(156)=.35 \sim .81, p<.001$. This would not be expected if the eight tasks elicited radically different estimation strategies, and it suggests that the logarithmic-to-linear theory provides an accurate picture for mental representation of all kinds of numerical estimations.

## Implications for alternative accounts

Broadly, our results undercut key claims of the proportionjudgment and measurement-skills accounts. A key claim of the proportion-judgment account is that developmental change involves a change in the degree of bias and use of implicit reference points. In this view, the degree of bias $(\beta)$ was thought to gradually converge on 1 , and more reference points would be utilized by the participants, "from an unbounded power to a one-cycle proportional to a two-cycle proportional version of the model" (Slusser et al., 2013, p.5). If these views were correct, the weights for 0 -cyclic power model ( $w_{1}$ ) and 1-cyclic power model ( $w_{2}$ ) in MCPM1 would be expected to decrease with age and the weight for 2-cyclic power model $\left(w_{3}\right)$ would be expected to increase -at the very least among the bounded tasks in Experiment 1 and 2. However, we found no support for this developmental pattern among any of our eight tasks. Additionally, there was no stable pattern of individual differences in the degree of bias and use of reference points. Given the relatively poor fits of these models, this lack of predictive power might not be surprising, but it does warrant caution about the psychological meaning of the parameter values.

Our results also provide robust evidence against the measurements-skills account. First, according to Cohen and Sarnecka (2014), "the implicit addition needed for the unbounded task is less mathematically sophisticated than the implicit subtraction needed for the bounded task, [therefore] children should perform better on the unbounded task at a younger age" (p. 1643). Against this contention, we found greater accuracy for bounded than unbounded tasks regardless of age, symbolic format, or provision of anchors. Far from being easier, the unbounded tasks were more difficult and actually yielded the highest logarithmicity scores. Even more critically, the parameter values of the models associated with this account (subtraction and scallop bias) were thought to track general subtraction and addition skill. If so, one would expect them to predict subtraction and addition skill when subjects actually performed subtraction and addition. However, we found no evidence that this was the case. Again, given the relatively poor fits of these models, its lack of predictive power should not be surprising.

## Acknowledgments

We would like to thank the administrators, teachers, students, and parents in Columbus for their involvement on the project. Also, we would like to thank Ariel Lindner and Rebecca Zagorsky for help with data collection. Partial support for this project was provided by a grant from the Institute for Educational Sciences (IES R305A160295).

## References

Barth, H. C., \& Paladino, A. M. (2011). The development of numerical estimation: evidence against a representational shift. Developmental Science, 14(1), 125-135.
Booth, J. L., \& Siegler, R. S. (2006). Developmental and
individual differences in pure numerical estimation.
Developmental Psychology, 42(1), 189-201.
Cicchini, G. M., Anobile, G., \& Burr, D. C. (2014).
Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. Proceedings of the National Academy of Sciences of the United States of America, 111(21), 7867-7872.
Cohen, D. J., \& Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. Psychonomic Bulletin \& Review, 18(2), 331-338.
Cohen, D. J., \& Sarnecka, B. W. (2014). Children's numberline estimation shows development of measurement skills (not number representations). Developmental Psychology, 50(6), 1640-1652.
Dehaene, S., Izard, V., Spelke, E., \& Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. Science, 320(5880), 1217-1220.
Hollands, J. G., \& Dyre, B. P. (2000). Bias in proportion judgments: the cyclical power model. Psychological Review, 107(3), 500-524.
Kim, D., \& Opfer, J. (in press). A unified framework for bounded and unbounded numerical estimation. Developmental Psychology.
Opfer, J. E., \& Siegler, R. S. (2007). Representational change and children's numerical estimation. Cognitive Psychology, 55(3), 169-195.
Opfer, J. E., Siegler, R. S., \& Young, C. J. (2011). The powers of noise-fitting: reply to Barth and Paladino. Developmental Science, 14(5), 1194-1204.
Opfer, J. E., Thompson, C. A., \& Kim, D. (2016). Free versus anchored numerical estimation: a unified approach. Cognition, 149, 11-17.
Siegler, R. S., \& Booth, J. L. (2004). Development of numerical estimation in young children. Child Development, 75(2), 428-444.
Siegler, R. S., \& Opfer, J. E. (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. Psychological Science, 14(3), 237-243.
Siegler, R. S., Thompson, C. A., \& Opfer, J. E. (2009). The logarithmic-to-linear shift: one learning sequence, many tasks, many time scales. Mind, Brain, and Education, 3, 143-150.
Slusser, E. B., Santiago, R. T., \& Barth, H. C. (2013). Developmental change in numerical estimation. Journal of Experimental Psychology: General, 142(1), 193-208.
Thompson, C. A., \& Opfer, J. E. (2008). Costs and benefits of representational change: effects of context on age and sex differences in symbolic magnitude estimation. Journal of Experimental Child Psychology, 101(1), 2051.

# Towards a Computational Analogical Theory of Mind 

Irina Rabkina (irabkina@u.northwestern.edu)<br>Clifton McFate (c-mcfate@northwestern.edu)<br>Kenneth D. Forbus (forbus@northwestern.edu)<br>Qualitative Reasoning Group, Northwestern University<br>2133 Sheridan Road, Evanston, IL, 60208, USA<br>Christian Hoyos (choyos@u.northwestern.edu)<br>Department of Psychology, Northwestern University<br>2029 Sheridan Road, Evanston, IL, 60208, USA


#### Abstract

Several theories about Theory of Mind (ToM) have been proposed. The most well-known of these are Theory Theory and Simulation Theory, although alternative and hybrid theories do exist. One such theory, proposed by Bach (2011, 2014), is based on the Structure-Mapping theory of analogy, which has been shown to play a key role in cognitive development. There is evidence that children are more likely to pass false belief tasks when trained using stories that are easy to compare via structural alignment, as opposed to stories that are difficult to compare in this way (Hoyos, Horton \& Gentner, 2015). This paper shows how a computational model based on Bach's account can provide an explanation for the Hoyos et al. training study and proposes directions for future research on human subjects.


Keywords: analogy; theory of mind; false belief; structuremapping; cognitive modeling

## Introduction

The mechanisms behind Theory of Mind (ToM) have been hotly debated for decades. According to one popular theory, Theory Theory, children are little scientists who develop theories about others' beliefs (e.g. Gopnik \& Wellman, 1992). Another theory, Simulation Theory, suggests that children play out scenarios as if they were the agents involved (e.g. Goldman, 1992). Other accounts include hybrid theories (e.g. Bach, 2011), which attempt to combine aspects of Theory Theory and Simulation Theory (see Related Work).

Another important question is how ToM is learned and when. Interestingly, several studies have shown that at least some aspects of ToM can be improved via brief intervention (e.g. Hoyos et al., 2015; Hale \& Tager-Flusberg, 2003; Lohman \& Tomasello, 2003). This paper considers one such study and models how analogical generalization may lead to improved performance on the false belief tasks tested. The model generates testable predictions for future work.

We begin by discussing the theories that underlie our model, the Structure-Mapping Theory of analogy (SMT; Gentner, 1983) and Bach's (2011) structure-mapping account of ToM, along with our computational models of analogical matching and generalization used in the model. We then summarize a ToM training study (Hoyos et al., 2015) and describe how our model explains the performance
improvements provided by training. We close with related work and future directions.

## Background

We base our model on the Structure-Mapping approach to Theory of Mind proposed by Bach (2011, 2014). Because understanding Structure-Mapping Theory (SMT; Gentner, 1983) is essential to understanding this theory and our model, we describe it first. This is followed by a description of Bach's theory. Finally, we describe the computational models of SMT processes that we are using.

## Structure-Mapping Theory

Structure-Mapping (Gentner, 1983) is a theory of analogy and similarity. Under SMT, relational/structural similarity is emphasized over similarity based on features alone. Humans' ability to see these structural similarities across dissimilar cases is a key aspect of higher order cognition, which suggests that structural similarity is used in everyday reasoning. SMT proposes that comparison involves the alignment of elements between two cases, called a base and a target.

Consider a common pedagogical analogy: "A cell is like a city. The city government controls the city. The nucleus controls all the cell's activities. A power station provides electricity. A mitochondrion is like the power station." (Chang \& Forbus, 2015). In this example, the cell acts as a target and the city acts as a base. Structural representations of the two are aligned to form a mapping. SMT predicts that the cell maps to the city, the nucleus maps to the government, control of the cell maps to control of the city, and the mitochondrion maps to the power station (Fig. 1). What about providing electricity? Because of the match between the mitochondrion and the power station, we can infer that the mitochondrion does something like providing electricity. This conclusion is called a candidate inference.

SMT can be extended to include analogical generalization (Kuehne et al., 2000). As a person is exposed to alignable cases, generalizations are formed. For example, we can form a generalization between the city and the cell. This would state that "Something like a city or cell has something like a city government or nucleus that controls it and something like


Fig. 1: A visual representation of SMT. Entities are shown as rectangles. Relationships are in diamonds. Dotted lines show correspondences between the two cases. Dashed lines show the candidate inference.
a power station or a mitochondrion that gives it energy." Eventually, generalizations become abstract schemas that can represent, for example, a single type of event. They can be stored in long term or working memory.

## SMT Theory of Mind

Bach $(2011,2014)$ has proposed that ToM is developed via structure-mapping. He proposes that two forms of base domains are used. The first are abstract schemas built up over time. The second are events from autobiographical memory. This provides a hybrid model: Mappings to the schema domain correspond to theories as described in Theory Theory models, and mappings to the autobiographical domain correspond to simulation. For example, to decide whether a person who arrived 15 minutes late to a flight that was delayed by 10 or a person who arrived 15 minutes late to a flight that left on time would be more upset, a person might retrieve an abstract schema that says "people are very upset when they narrowly miss their goal" or they might simulate how they would feel if they were the person in question by mapping to an autobiographical memory. Bach argues that simulation tends to happen when the general heuristic has not yet been formed, and involves complex combinations of cases (see Bach, 2011 for specifics).

Because we do not attempt to model a complete Theory of Mind in this paper, we assume a simplified version of Bach's theory. Our model focuses on the learning aspect, so we assume that heuristic-like abstractions have not yet been formed. Thus, only concrete autobiographical memories are retrieved from long term memory. Generalizations are formed in working memory, which we propose as a mechanism by which schemas are learned.

## SME and SAGE-WM

The Structure Mapping Engine (SME; Forbus et al., 2016) implements the analogical mapping process of SMT. SME compares a base and target case, both represented in predicate calculus, and computes one or more mappings that align statements and entities. Each mapped expression receives an initial score, which is propagated to its children. Thus, highly
nested expressions have high scores. The score of a mapping is the sum of the scores of its constituents. Thus, mappings between cases that have high structural similarity receive higher scores. Mappings also include candidate inferences that project missing information from one case to the other.

In this model, we deliberately do not model retrieval from long-term memory, to avoid the cost of providing enough distractors to make this challenging, and instead assume that retrieval finds reasonable autobiographical memories. However, we have proposed (Kandaswamy et al., 2014) that analogical generalization also occurs in working memory, what we call interim generalizations. The SAGE-WM model ${ }^{1}$ keeps a list of generalizations and recent examples. Given a new example it uses SME to compute a score between the probe and each generalization in turn, ordered by recency. If the score is over a pre-determined threshold, the probe is assimilated into the generalization. If no generalization is above threshold, the new example is compared to each outlier in turn using SME, again ordered by recency. If any mapping is above threshold, a new generalization is formed. Otherwise the probe becomes a new ungeneralized example.

## Learning Theory of Mind

Several studies have shown that Theory of Mind can be acquired in part using experimental interventions (e.g. Lohman \& Tomasello, 2003; Hale \& Tager-Flusberg, 2003; see Hofmann et al., 2016 for meta-analysis). However, most of these studies involve extended training. On the other hand, there is evidence that ToM can be acquired much more quickly when training examples are highly structurally alignable. In particular, a study by Hoyos et al. (2015) showed that structurally alignable unexpected contents-style stories can improve children's performance on false belief tasks, given just three training examples. In this paper, we examine and model the results of this experiment.

## Modeling Task

In the Hoyos et al. (2015) study, children were first given a false belief pre-test containing one unexpected contents task (UC), one verbal false belief task (VFB), and one unexpected location task (UL). In the UC task, a container (e.g. a cookie box) is shown to have unexpected contents (e.g. grass) and participants are asked to predict what someone who has never seen inside would think the container contains. In VFB participants are told another child holds a false belief (that they think an item is somewhere it is not) and asked to predict where the child will look for the item. Finally, in UL, participants are told a story where one child places an object in a location and leaves the room. Another child then moves the object, and the participants are asked to predict where the first child will look for the object when they return.

Those who passed all three tests were excluded from the study. The remaining children were split into two groups: high alignment and low alignment. Both groups were

[^444]presented with three stories in the style of an UC task, in a repetition-break pattern: the main character in the first two stories held a true belief (e.g. she thought that there was cereal in a cereal box, and there really was cereal inside), while the character in the last held a false belief (e.g. she thought there were crayons in the crayon box, but there were really rocks). The difference was that the stories heard by children in the high alignment condition were very similar, in terms of both structure and linguistic content. The stories heard by children in the low alignment condition, on the other hand, differed on both counts. Following training, all children were tested on the same three tasks (UC, VFB, UL) as before.

Hoyos et al. found that children in both conditions made significant gains from pre- to post-test. Importantly, they found that the children in the high alignment condition made significantly higher gains than those in the low alignment condition. Hoyos et al. concluded that structural alignment aids false belief understanding. Furthermore, they, like Bach (2011, 2014) postulated that analogical comparison is "instrumental in children's understanding of mental states and their relation to the factual world." In this paper, we propose a mechanism for how structural alignment during learning can aid in false belief understanding and forming a complete Analogical Theory of Mind.

## Learning Analogical Theory of Mind

The mean performance increase by children in the high alignment group was 0.75 out of 3 possible, with significant gains made in all three of the false belief tests. Yet few children learned more than one. On the other hand, children in the low alignment condition made an average of 0.23 gains. Only gains in the UC task were significant. Since all of the training examples were variants of UC, it is not surprising that this was the easiest task to learn. However, learning ToM requires the ability to transfer to other tasks, as was the case with children in the high alignment condition. The process of making gains in UL and VFB tasks must, then, be different than the process of only gaining UC.

We argue that analogical comparison in working memory alone leads to gains in the UC. That is, immediate recall of the training examples themselves is sufficient to cause gains. In contrast, a generalization between a training example and an autobiographical memory retrieved from long term memory leads to transfer to the other two tasks, VFB and UL. The violation of expectation generated during training causes the child to probe long term memory for a case of similar surprise. What exactly they find surprising about the training-that something other than what they expected was inside the box, that the character in the story was incorrect in her guess, or something else-affects the case that is retrieved from long term memory. This in turn affects which of UL and VFB the child is able to answer.

## A Computational Model

Our model, like Bach's theory, is based in SMT, using SAGE-WM for reasoning and learning.

## Our Model

A simplified English version of each training and testing example from Hoyos et al. (2015) was semi-automatically encoded using a natural language understanding system (EA NLU; Tomai \& Forbus, 2009). Although syntax was simplified, overall structure and word choices were consistent with the original stories. Figure 2 shows a partial representation of a true belief story. Events are represented in the neo-Davidsonian style: a reified event with role relations connecting it to other constituents. The conjunction of statements about an event participates in causal relations. In English, Figure 2 states that because it is not the case that there is a seeing event in the box by Kim, Kim thinks that there is a containment event wherein the box contains cereal.

During training, the appropriate examples were passed into SAGE-WM in the order that the children in the corresponding condition saw them (true belief, true belief, false belief). The threshold for whether or not a probe was generalized was set to 0.01 . If the incoming example matched to an example already in working memory with a score greater than 0.01 , the model asked whether the match was correct. This corresponds to feedback in the Hoyos et al. (2015) experiment. When told it was correct, the model assimilated the examples into a generalization. Its behavior when told it was incorrect, on the other hand, depended on its calculation of surprise. Surprise occurs when the model encounters an incorrect match whose score is the same order of magnitude as the previous correct match. We propose that this comes out of the repetition break structure of the story order (Hoyos et al., 2015; Loewenstein \& Heath, 2009): the high similarity to the interim generalization leads to a strong expectation of sameness, and the violation leads to a search for recategorization. When surprised, the model probes long term memory for an alternative case to align with.

Figure 3 gives a visual representation of our model. In the high alignment condition (a), the first true belief story is stored in working memory. The second true belief story is then matched to the first, and an interim generalization is formed. When the false belief story comes in, it too matches to the generalization. Due to violated expectations, long term

```
(causes-Underspecified
    (not
        (and
        (inside-UnderspecifiedRegion see85118 box1)
        (perceivedThings see85118
            (InsideOfSpaceRegionFn box1))
        (isa see85118 VisualPerception)
        (doneBy see85118 kim)))
    (opinions kim
        (and
            (containedObject contain84430 cereal84499)
            (containingObject contain84430 box1)
            (isa cereal84499 BreakfastCereal)
        (isa contain84430 ContainingSomething))))
```

Fig. 2: A partial representation of a true belief story. This statement represents the phrase "Kim thinks that the box contains cereal because Kim has never seen inside the box".
memory (dotted line) is probed. Long term memory is a collection of generalized and specific cases that represent memories formed over time. If a case is retrieved, an interim generalization between the match and the false belief case is created and stored in working memory (b).

In the low alignment condition (c), on the other hand, no generalization is formed between the two true belief cases. This leads to them being stored as separate cases in working memory. When the false belief case comes in, it matches to the first true belief case, but no element of surprise is present when the model is corrected. For this reason, long term memory is never probed, and working memory consists of only the three training examples (d). The contents of working memory during testing predict the questions that the child is able to answer.

Testing proceeded as follows: cases were again encoded semi-automatically using EA NLU. These cases were given to the model which retrieved the most similar case from working memory and generated candidate inferences by analogy. The candidate inferences correspond to what the model predicts is missing from the test cases (e.g. what the agents will do). These candidate inferences were manually inspected to determine whether any could result in correctly answering the test questions.


Fig. 3: A visual representation of our model of training in the Hoyos et al. study. (a) shows training in the high alignment condition. (b) is a representation of working memory after high alignment training. (c) and (d) show low alignment training and the consequent working memory, respectively. Cases that are structural matches to the probe are bold.

## Results

The model behaved as predicted. In the high alignment condition, the model generalized the true belief cases with a normalized match score of 0.075 . It then matched the false belief to the generalization with a score of 0.066 , which corresponds to the child incorrectly predicting that the character in the story knows what is in the box. The model was then informed that this match was incorrect. Because the similarity scores it had encountered were within the same order of magnitude, it searched long term memory for another match. It then retrieved one of two memory cases that matched with a normalized score of 0.083 or 0.066 , and created an interim generalization between it and the false belief case. We used stories intended to approximate a memory a child might have (e.g. thinking that a magician put a ball inside of a hat, only to find the hat empty) to model what might plausibly be retrieved. Depending on the case retrieved, the model was then able to answer VFB or UL. Correctness was evaluated based on the candidate inferences generated from the best mapping between the test case and the contents of working memory. For example, to correctly answer "Where is Nora going to look for her ball?" (UL) the mapping must produce a candidate inference stating that there might be a looking event, in which Nora looks for her ball in the appropriate location.

In the low alignment condition, on the other hand, the second true belief case matched to the first with a very low similarity score of 0.0014 , well below threshold. For this reason, the model did not form a generalization between them. When the false belief case was compared, it had a match score of 0.066 with the first true belief case. Similar to the high alignment condition, the model was informed that this was not a correct match.

Because the previous match score was of a different order of magnitude, the model did not look into long term memory, and instead stored the false belief case alongside the two true belief cases. When the UC case came in, the false belief case was retrieved. The mapping generated a candidate inference that would allow the model to properly answer "What does she think is in the box?" This candidate inference stated that not having looked inside the cookie box would cause the agent to believe that it contained something analogous to crayons in the crayon box from the training example. That is, cookies.

Note that this retrieval is due to recency in working memory: the UC test case lacks the explanation present in the training cases about why a person holds a certain belief (e.g. "Kim thinks that the cereal box contains cereal because Kim has never looked inside the box."), so the first true belief case had the same match score. If that case had been retrieved, the model would not have been able to answer UC correctly.

## Discussion

Our model gives one explanation for the results of the Theory of Mind training study presented in Hoyos et al. (2015). It also suggests that an important step in ToM development is generalizing belief-state cases in long term memory. In the
training studies, understanding that the training cases can, and indeed should, be assimilated to long term memory with belief-state interpretation cases is crucial. In other words, children may be accumulating experiences that require reasoning about belief states in long term memory, but these memories remain inert until a surprising event-such the one experienced by the high alignment participants in the Hoyos et al. study-stimulates their retrieval and begins the process of creating schemas that can be used in future ToM reasoning. This predicts that children in the high alignment condition of Hoyos et al. (2015) are more likely to retain what they have learned than the children in the low alignment condition: the children in the high alignment condition were more likely to access those experiences from long term memory and form a generalization with them.

In addition, our model predicts that reversing the order of training examples would cause children in both conditions to fail. In the low alignment case, when the most recent training example is retrieved, children would match the UC task to a true belief scenario, and answer incorrectly. Children in the high alignment case would similarly fall back on retrieval of the most recent case, as they would not experience the surprise caused by the repetition break structure.

Previous studies (e.g. Hale \& Tager-Flusberg, 2003; Lohmann \& Tomasello, 2003) have suggested that experience plays a role in ToM development. Our model provides a concrete explanation for how these experiences might lead to ToM and provides further suggestions for human subject experiments.

## Related Work

## Theories of Theory of Mind

Here, we summarize the best-known ToM theories.
Theory Theory One of the most popular takes on ToM is Theory Theory, which views the child as a scientist with regard to interpreting other people's mental states (e.g. Gopnik \& Wellman, 1994). The child begins with a naïve theory about others, sometimes referred to as a folk psychology, which she modifies and adapts as evidence that supports or refutes the theory is observed. The theory gradually develops from only understanding desire states, to belief states, to how belief and desire states influence each other and behavior (Bartsch \& Wellman, 1995).

Simulation Theory Under the Simulation Theory view, a child mentally simulates events in order to predict others' actions and beliefs (Goldman, 2006), and develops by improvement in simulation abilities (Flavell, 2004). Criticisms of Simulation Theory include that errors made by both children and adults are not consistent with those predicted by Simulation Theory accounts (Saxe, 2005) and that simulation is not sufficient for describing observed developmental patterns (Perner \& Howes, 1992).

Modular Theories Another common account is that ToM can be explained as a single cognitive module. Scholl and Leslie (1999) list six characteristics of modules: they are domain-specific, their behavior is, at least in part non-
voluntary, their processing is fast, their outputs are shallow and highly constrained, they are often located in a particular region of the brain, and their processes may be impairedand selectively impaired-by neural damage. Importantly, according to Scholl and Leslie, modularity theories "intend to capture only the origin of the basic ToM abilities" (1999). In this sense, modularity theories do not necessarily compete with other theories of ToM discussed here.

Hybrid Theories Several hybrid theories have been proposed to bridge the gap between Theory Theory and Simulation Theory. Some, which Bach (2011) calls dividedhybrid models, alternately assign aspects of Theory of Mind to simulation or theorizing, depending on which is better supported by empirical data (e.g. Heal, 1996). This approach, as Bach notes, avoids discussion of acquisition. It is unclear how a child learns to use simulation for some tasks and theory for others, and how simulation and theory develop concurrently. Other hybrid theories, which Bach (2011) calls dynamic-hybrid models, focus on continued development. Bach's model falls under this category. Like other dynamichybrid theories, Bach's allows for development and changes to ToM not only throughout childhood, but into adulthood. This includes switching between theorizing and simulating to complete the same tasks at different points in development. As psychologists continue to find evidence of ToM shifts throughout adulthood (e.g. Hess, 2006), dynamic-hybrid theories become more and more plausible.

## Computation Models of Theory of Mind

Hiatt and Trafton (2010) implemented a model of Theory of Mind using the ACT-R cognitive architecture (Anderson, 2007) that learned to perform the Sally-Ann task. It extracted facts out of the scenario and was asked several false belief questions about what it saw. It was rewarded for answering correctly and punished for answering incorrectly, leading it over time to inhibit true belief responses, producing a learning curve consistent with developmental data. However, unlike our model, the training they used did not follow from an empirical training study. We note that the children in the Hoyos et al. (2015) study were able to learn aspects of false belief after seeing just three examples, only one of which actually was a false belief situation.

Goodman et al. (2006) modeled ToM via two Bayesian networks that respectively represent a naïve and expert theory in a Theory Theory account. They propose the models as competing hypotheses in the Sally-Anne task, and show how, during training, the expert theory becomes preferred over the naïve theory. The need to hand-code both theories in the system's starting endowment makes it more of a computational level model (Marr, 1982), whereas we provide a process-level model of learning. Furthermore, our model is consistent with the evidence from the training study presented by Goodman et al. (2006), which shows that surprise can improve children's ToM performance.

## Future Directions

Our results provide evidence that structure-mapping is indeed a plausible process-level mechanism (Marr, 1982) for ToM and how it is learned. As such, our future work will look toward developing a complete computational Theory of Mind, including both the theory and simulation aspects of Bach's theory, using SAGE as the underlying mechanism.

## Acknowledgments

We thank Alissa Baker-Oglesbee and Dedre Gentner for their helpful comments. This research was supported by the SocioCognitive Architectures for Adaptable Autonomous Systems Program of the Office of Naval Research, N00014-13-10470.

## References

Anderson, J. R. (2007). Using brain imaging to guide the development of a cognitive architecture. Integrated models of cognitive systems, 49-62.
Bach, T. (2011). Structure-mapping: Directions from simulation to theory. Philosophical Psychology, 24(1), 2351.

Bach, T. (2014). A Unified Account of General Learning Mechanisms and Theory- of- Mind Development. Mind \& Language, 29(3), 351-381.
Bartsch, K., \& Wellman, H. M. (1995). Children talk about the mind. Oxford university press.
Chang, M.D. and Forbus, K.D. (2015). Towards Interpretation Strategies for Multimodal Instructional Analogies. Proceedings of the 28th International Workshop on Qualitative Reasoning (QR2015). Minneapolis, MN.
Flavell, J. H. (2004). Theory-of-mind development: Retrospect and prospect. Merrill-Palmer Quarterly, 50(3), 274-290.
Forbus, K. D., Ferguson, R. W., Lovett, A., and Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. Cognitive Science, DOI: 10.1111/cogs.12377, pp 1-50.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Goldman, A. I. (1992). In defense of the simulation theory. Mind \& Language, 7(1- 2), 104-119.
Goldman, A. I. (2006). Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press.
Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ... \& Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In Proceedings of the twenty-eighth annual conference of the cognitive science society (pp. 13821387).

Gopnik, A. \& Wellman H. (1992). Why the childe"s theory of mind really is a theory. Mind and Language, 7, 145-171.

Gopnik, A., \& Wellman, H. (1994). The theory theory. In L. Hirschfeld \& S. Gelman (Eds.), Domain specificity in culture and cognition. NY: Cambridge University Press.
Hale, C. M., \& Tager- Flusberg, H. (2003). The influence of language on theory of mind: A training study. Developmental science, 6(3), 346-359.
Heal, J. (1996) Simulation, Theory, and Content. In Carruthers, P. \& Smith, K. (Eds.), Theories of theories of mind. Cambridge: Cambridge University Press.
Hess, T. (2006). Adaptive aspects of social-cognitive functioning in adulthood: Age-related goal and knowledge influences. Social Cognition, 24(3), 279-309.
Hiatt, L. M., \& Trafton, J. G. (2010, August). A cognitive model of theory of mind. In Proceedings of the 10th International Conference on Cognitive Modeling (pp. 9196).

Hofmann, S. G., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L. A., ... \& Harris, P. L. (2016). Training children's theory-of-mind: A meta-analysis of controlled studies. Cognition, 150, 200-212.
Hoyos, C., Horton, W., \& Gentner, D. (2015). Analogical comparison aids false belief understanding in preschoolers. In CogSci.
Kandaswamy, S., Forbus, K., and Gentner, D. (2014). Modeling Learning via Progressive Alignment using Interim Generalizations. Proceedings of the Cognitive Science Society.
Kuehne, S., Forbus, K., Gentner, D. and Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure-mapping. Proceedings of CogSci 2000, August.
Loewenstein, J., \& Heath, C. (2009). The Repetition-Break plot structure: A cognitive influence on selection in the marketplace of ideas. Cognitive Science, 33, 1-19.
Lohmann, H., \& Tomasello, M. (2003). The role of language in the development of false belief understanding. Child Development, 74(4), 1130-1144.
Marr, D. (1982). Vision. Freeman Publishers.
Perner, J., \& Howes, D. (1992). 'He Thinks He Knows': And More Developmental Evidence Against the Simulation (Role Taking) Theory. Mind \& Language, 7(1- 2), 72-86.
Saxe, R. (2005). Against simulation: the argument from error. Trends in Cognitive Sciences, 9(4), 174-179.
Scholl, B. J., \& Leslie, A. M. (1999). Modularity, development and 'theory of mind'. Mind \& Language, 14(1), 131-153.
Tomai, E. and Forbus, K. (2009). EA NLU: Practical Language Understanding for Cognitive Modeling. Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference. Sanibel Island, Florida.

# A Theory of Resonance: Towards an Ecological Cognitive Architecture 

Vicente Raja<br>University of Cincinnati, Cincinnati, OH, United States


#### Abstract

This paper may be seen as a blueprint for an ecological cognitive architecture. Ecological psychology, I contend, must be complemented with a story about the role of the CNS in perception, action, and cognition. Such a story must be a theory of resonance compatible with the main tenets of ecological psychology. I offer here the two main elements of such a theory: a framework (Anderson's neural reuse) and a methodology (multi-scale fractal DST).


# A model of structure learning, inference, and generation for scene understanding 

David Raposo<br>DeepMind, London, United Kingdom<br>Peter Dayan<br>Gatsby Computational Neuroscience Unit, London, United Kingdom<br>Demis Hassabis<br>DeepMind, London, United Kingdom<br>Peter Battaglia<br>DeepMind, London, United Kingdom


#### Abstract

Humans possess rich knowledge of the structure of the world, including co-occurrences among entities, and covariation among their discrete and continuous features. But how people learn, infer and predict this structure is not well understood. Here we explore everyday scene understanding as a case study of people's structural knowledge and reasoning. We introduce a probabilistic model over scene graphs that can learn the relational structure of objects and their arrangements and support inference and generation. Our model was able to learn the underlying structure of real-world scenes, and use it for inference and compression. In two human psychophysical experiments we found that a corresponding computational cognitive model was able to explain how people learn novel scene distributions and use it for classification and construction. Our work represents the first computational theory of human scene understanding that can account for people's rich capacity for learning and reasoning about structure.


# Challenging the superficial similarities superiority account for analogical retrieval 

Lucas Raynal (lucasraynal@gmail.com)<br>Paragraphe Lab, EA 349, University Cergy-Pontoise<br>Avenue Marcel Paul, 92230 Gennevilliers, FRANCE

Evelyne Clément (evelyne.clement@u-cergy.fr)<br>Paragraphe Lab, EA 349, University Cergy-Pontoise<br>Avenue Marcel Paul, 92230 Gennevilliers, FRANCE

Emmanuel Sander (emmanuel.sander@paris8-univ.fr)<br>Paragraphe Lab, EA 349, University Paris 8, Department of Psychology<br>2 Rue de la liberté, 93526 Saint-Denis Cedex 02, FRANCE


#### Abstract

The predominant view concerning determinants of analogical retrieval is that it is preferentially guided by superficial cues. In order to test the cognitive plausibility of a structural similarities-based retrieval, we constructed a story-recall task in which life-like scenarios shared structural correspondences. In Experiment 1, we showed that such structural similarities induce retrievals when the participant had several source candidate situations sharing superficial similarities with the target cue. Experiment 2 was designed to test whether the encoding was sufficiently oriented on structural similarities to drive retrievals, even if the participants possess only one source candidate situation with superficial matches in memory. The results of the two present experiments lead us to conclude that in some contexts, abstract encoding induces a superiority of structural similarities over superficial ones in retrieval. Further implications for analogical retrieval approaches are discussed.


Keywords: Analogy; analogical retrieval; structural similarity; abstract encoding; story-recall task

## Introduction

Analogy has been identified as a key process to perceive the conceptual structure of a new situation by importing it from a familiar analog representation (Gentner, 1983; Gick \& Holyoak, 1983). Most studies are consensual that the mapping process permitting this transfer is preferentially based on structure rather than surface correspondence. In other words, when comparing two analogs, subjects do not rely on similarities in terms of objects or object attributes, but instead tend to focus on common abstract relations. Figure 1 provides an illustration of a target cue story sharing a structural correspondence with a superficially dissimilar analog source candidate situation ("making a deal to avoid a bad situation"), and a surface correspondence with a superficially similar disanalog source candidate situation belonging to the same semantic domain (raptors, tailfeathers, etc).

Target cue story
Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and
some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.

Superficially dissimilar analog source candidate situation Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.

Superficially similar disanalog source candidate situation Once there was an eagle named Zerdia who donated a few of her tailfeathers to a sportsman so he would promise never to attack eagles. One day Zerdia was nesting high on a rocky cliff when she saw the sportsman coming with a crowsbow. Zerdia flew down to meet the man, but he attacked and felled her with a single bolt. As she fluttered to the ground Zerdia realized that the bolt had her own tailfeathers on it.

Figure 1: Situations sharing different types of similarity in Gentner, Ratterman \& Forbus (1993)

Types of similarity implicated in the retrieval of a source representation in memory are the main issues debated in the literature. The predominant view is that the retrieval of a source candidate situation critically depends on superficial similarities, whereas the influence of structural similarities seems more peripheral (Gentner, 1983; Gentner, Ratterman \& Forbus, 1993; Gick \& Holyoak, 1983; Trench \& Minervino, 2015). In Gentner and Colhoun's (2010) words: "Relational retrieval can be said to be the Achilles' heel of our relational capacity. There is considerable evidence that
similarity-based retrieval, unlike the mapping process, is more influenced by surface similarity than structural similarity." However, opposite assertions have arisen in the literature, attributing a major influence to structural similarities, even overcoming that of surface similarities (Blanchette \& Dunbar, 2000; Dunbar \& Blanchette, 2001; Hofstadter \& Sander, 2013; Kretz \& Krawczyk, 2014). The aim of this paper is to demonstrate that structural matches have a greater influence than superficial ones in the retrievals of life-like situations. This will be shown by creating a competition between source candidate situations sharing either exclusively superficial or structural similarities with the target cue (Figure 2 illustrates this intended feature composition). Before considering the current experiments, we report the major findings stemming from several experimental paradigms.

|  | Analog source <br> candidate situations | Disanalog source <br> candidate situations |
| :---: | :---: | :---: |
| Structural similarity | X |  |
| Superficial similarity |  | X |

Figure 2: Correspondences between target cues and critical source candidate situations intended in Experiment 1 and 2

## Analogical Problem Solving

In the problem solving domain, a common experimental design to study analogical retrieval via analogical transfer is the source-target paradigm (Gick \& Holyoak, 1983; Holyoak \& Koh, 1987). A source problem situation is first proposed with its solution, then an analog target problem is given to be solved. To measure the role of surface correspondence in access, the surface of the source problem is manipulated to either match or not with the target one. The retrieval is considered to have occurred when the participant detects similarities between the target problem and the source problem without further hints from the experimenter, leading to the transfer of the resolution procedure from the source to the target. Results have shown that retrieval is high when the source is both superficially and structurally similar to the target. More precisely, the similarities in terms of the problems' story theme have appeared to be a crucial determinant in access (Ross, 1987). Inversely, superficially dissimilar analog source problems are seldom retrieved by the participants (Gick \& Holyoak, 1983). Studies on problem solving have also demonstrated that structural similarities can, in certain circumstances, have a role in retrieval: when two source analogs are presented jointly to be compared, their critical solution principle can be retrieved when faced with the target (Catrambone \& Holyoak, 1989). Also, retrievals are reported to be more frequent when the two superficially dissimilar analog problems share structural similarities at a less abstract level (Holyoak \& Koh, 1987).

## Story-Recall Paradigm

Another frequently used paradigm is the story-recall task (Gentner et al., 1993; Wharton, Holyoak \& Lange, 1996; Catrambone, 2002). A set of short text stories are presented as source candidates for the retrieval before the introduction of target cue situations which share various similarities with them (Figure 1). Within a problem solving paradigm, the problems generally share the same structure since it is the transfer of an abstract solution being investigated. This is not systematically the case in the story-recall paradigm since the situations are chosen to create a competition between source candidate situations possessing exclusively the surface or exclusively the structure in common with the target cue (Gentner et al., 1993). This paradigm showed superficially similar disanalogs were retrieved significantly more often than superficially dissimilar analogs, leading to the conclusion that superficial similarity is the main factor implicated in access.

Wharton, Holyoak, Downing, Lange, Wickens \& Melz (1994) argued that the minor role attributed to structural matches in story-recall tasks could be due to the fact that only one source candidate, the superficially similar disanalog, shares some semantic features with the target cue. One can note that in real-life conditions, several source candidate situations, corresponding in superficial features with the perceived situation, are generally available in memory. The authors observed that structural similarities may play a certain role in access because when the analog source candidate also shares some superficial similarities, it is better retrieved than the concurrent source candidate possessing only superficial similarities with the target cue. Structural similarities also seem to play a certain role in access when they are implemented without superficial matches: the retrieval of a superficially dissimilar source candidate is higher when it shares structural features with the target cue (Wharton et al., 1996). However, the role of structural similarities in access has only been shown when two source candidates shared the same amount of surface correspondence. Hence, the reviewed works did not demonstrate the superiority of structural similarities in access, since it would require showing that superficially dissimilar analogs are better retrieved than superficially similar disanalogs.

## Production Paradigm

Challenging the ecological nature of traditional experimental conditions (unfamiliar source and target situations, short familiarization time, restrained pool of source candidates), further research focused on the retrieval of situations encoded prior to the experiment in real-life conditions (Blanchette \& Dunbar, 2000; Dunbar \& Blanchette, 2001; Kretz \& Krawczyk, 2014). With this configuration, a high structural overlap was generally observed between the source and the target cue situation. For instance, expert discourse in scientific domains (politics, biology, economics) exhibited predominance for structural analogies, though sometimes also sharing a superficial
correspondence. Those findings were replicated in experimental conditions with a production paradigm, where participants who were allowed to select their own sources of analogy retrieved significantly more semantically distant analogs than superficially similar analogs (Blanchette \& Dunbar, 2000). In other words, those findings not only advocate for the major role of structural similarity but also for the weak impact of superficial similarity when accessing an analog. However, Trench \& Minervino (2015) pointed out that the sources provided by the participants could be invented rather than real memories, and that superficially dissimilar analogs could be more common in memory than superficially similar ones. They tested the potential that while controlling the availability and number of the two types of concurrent source candidates, findings obtained from usual source-target and story-recall paradigms would be replicated when the participant generates analogical retrievals from her or his own experiences. They proposed target situations with either superficial and structural, or solely structural similarities with their memories. In accordance with prior findings, participants more often proposed superficially similar analogs than superficially dissimilar ones. In another study using this paradigm, situations retrieved by management consultants provided with target situations embodying an original negotiation principle manifested only superficial matches (Gentner, Loewenstein, Thompson \& Forbus, 2009). Contrary to previous studies using the production paradigm, this experiment reflects a marginal tendency to access structurally similar source situations. These results, in sharp contrast with the ones obtained by Dunbar and Blanchette (2000), indicate that the natural settings of the encoding condition of one of the analog is not in itself the critical parameter influencing the type of retrieval. Namely, whereas it probably promotes the abstract encoding of this analog situation, the access to its structural matches with the other analog which is still provided by the experimenters is not guaranteed.

## Encoding through abstract concepts

When faced with a target cue situation, the fail to retrieve an analog situation is generally interpreted as a defect of abstract encoding of the two situations (Gick \& Holyoak, 1983; Gentner et al., 2009). If the participant is unable to grasp the structure of the situations, the abstract similarities could not be used as a cue to retrieve. Indeed, only if the participant is incited to compare either two source analogs or two target analogs, he might extract a schema sufficiently abstract, and subsequently perceive this schema in a superficially dissimilar analog, in the context of the retrieval.

However, even if the schema extracted from the situation is not abstract enough so that the similarities introduced between the analogs by the experimenter will be detected, it cannot be claimed that the encoding of a situation is purely concrete and literal (Hofstadter \& Sander, 2013). Whereas a wealth of stimulations is permanently available in our environment, one has to select the properties relevant to make sense of the situations. This cannot be done by processing
every perceptively available superficial piece of information. Instead, the situation's understanding depends on the properties (whether perceptual or abstract) that are compatible with the conceptual structure in construction and on the neglect of those which are not.

We suggest that familiar concepts' evocation during the encoding is a critical point to account for the abstract information raised by perceived situations. This idea is congruent with findings revealing that source analog problems that participants usually fail to retrieve in experiments are better accessed by expert participants (Novick, 1988). When experts have a familiar concept to highlight the abstract properties of these situations, the novice does not have this conceptual door toward the structure. This reveals that the novice participant has not elaborated the specific concept that allows for the encoding of the structure of complex problem situations necessary for the structurallybased retrieval of this type of situations. However, it is very likely that he/she has acquired an expertise in daily-life situations, where one systematically has to deepen the encoding until elementary abstract concepts are established, such as "making a deal", "bad faith", "lie", "authenticity", "prosocial behavior" and so on, to produce adapted behaviors. The activation of these concepts should highlight the structure underlying some daily-life situations in a way that elicit structurally-based retrieval. Whereas most studies have focused on abstract inter-domain analogies that a novice could rarely access, we aim at investigating analogies between situations inspired from social scenarios that can be experienced in different domains of daily-life.

## Experiment 1

In order to demonstrate the superiority of structural similarities over superficial similarities in retrieval, we conducted a first experiment where those two types of similarities were in competition in the retrieval of a source situation candidate. In that way, a story-recall task was used so as to control for the highlighting of the structure underlying both the analog source candidate situation and the target cue situation.

## Method

## Material

Social scenarios inspired from Wharton et al. (1996) involving life-like contexts were used. Although the objects of the analog situations were clearly divergent, they shared very similar role at the required level of abstraction for making sense of the stories. In the analog pair reported in Figure 3 for instance, both stories relate the setting of a social competition between two characters (rival cookers or classmates) having the same goal (turnover or seducing someone) and an unusual way to put an end to it by helping the competitor to enhance her/his critical ability (improve Lorenzo's pizza dough or looking after Diane's appearance). If those situations are not directly taken from the participant's
experience, $\mathrm{s} /$ he still can use her/his general knowledge about social relations to encode such scenarios.

In order to make sure that the two types (superficial or structural) of similarities with the target cue were never confounded in a same source situation, superficially similar source stories structure clearly diverge from the target ones. In this way, contrasting with previous works (e.g. Gentner et al., 1993), disanalogs are not modified versions of a same structure story, but describe structurally different scenarios.

Wharton et al. (1994) noticed that when semantic correspondence was not only present between the source disanalog and the target cue, the artificial saliency of the superficial matches decreased and resulted in a weaker influence on access. In their experiment, the concurrent superficially similar source candidates were the analog and the disanalog. As previously indicated, we aim at isolating the influence of structure and surface similarities by implementing them in different source candidates. Hence, we multiplied the number of source candidates sharing surface features with the target cue by introducing three superficially similar disanalogs. To respect a symmetry between the number of semantically close and distant source candidates, we also introduced three superficially dissimilar source candidate situations (the analog sharing structural correspondence, and two unrelated distractors).

## Target cue story

Luigi holds a pizza truck in a very popular place. Lorenzo, another ambulant pizzaiolo, has placed his truck just beside Luigi's and is detrimental to his turnover. Luigi realizes the dough of Lorenzo's pizzas is bland. Luigi spontaneously gives his personal recipe to Lorenzo so as he can enhance the quality of his product. Since then, his pizza dough is amazingly tasty. The same evening, Lorenzo declares to Luigi that in order to show him how much he found his intention was nice, he will move his truck in another sector, far from this one.

Superficially dissimilar analog source candidate situation
Julie is in love with Victor, her classmate, and she is getting closer to him in order to seduce him. But Diane joins the class in the middle of the year and also has a crush on Victor. Julie remarks that Diane was not very aware of her style and proposes her some relooking advices, showing her fashion photos and taking her for shopping. Diane now looks very cute and chic. Diane is so grateful that she tells Julie that she would stop flirting with Joe.

## Superficially similar disanalog source candidate situation

In a market place, the truck called « At Alessandro \& Fabio's » has various choices of homemade pizzas. The important clientele going there is fond of the authentic atmosphere steaming from this stand held by the two happy

[^445]looking men in Italian traditional suits. However, once they will have left this selling space, the two men will go to another market place after taking care of wearing German traditional clothes to sell sausage specialties. The sign will display «At Hans and Hendrich's ».

Figure 3: Examples of stories used in the Experiment 1

## Procedure and experimental design

The first two pages of a booklet presented the 6 source stories, then a blank page separated them from the last page comporting the target cue situation. Under each source story was a 5 points scale inviting participants to assess the ease they had to imagine the scene while reading it. As recommended by Wharton et al. (1996), this was done to promote a deep treatment of the situations. The dependent variable was the source retrieved during the reading of the target cue situation.

It was indicated that the task took around 10-15 minutes to fulfill but no time limitations was imposed. After they agreed to participate, participants were given the booklet. They were invited to read the instructions available in the first page. The target cue situation was presented on the last page, followed by the solicitation to indicate if they were reminded of one of the previous situations. If it was the case, they had to restitute any element they could remember about it.

## Participants

34 participants ( 25 women and 9 men, mean age 23.8 years) accepted to take part in the experiment in University libraries (Paris 5 and Paris 8). They were all fluent French speakers.

## Results and discussion

Access credit was attributed to the source candidate for which the participant recalled word content. If content word from more than one source was reported, the source containing the more content words in common was credited. If the participant explicitly reported more than one source despite the instruction, his response was excluded. 3 responses were not analyzed for this reason.

Analyses were drawn on a comparison between the superficially and structurally similar source candidates that were retrieved. Structurally similar source candidate situations were much more retrieved (84.61 \%) than superficially similar disanalog source candidate situations (15.39 \%, see Figure 4). This difference reaches high significance $\left(X^{2}(1, N=29)=12.46 ; P<0.001\right)^{1}$.

In real-life condition, one generally has in memory multiple source candidates sharing similarities in terms of superficial objects with the target cue situation. The results reveal that when a pool of semantically similar source candidate situations is available in memory, but those situations do not preserve the structure of the target cue, the
retrieval is preferentially guided by structural matches with sources of a distant semantic domain. However, our results cannot help us identifying whether the structural matches of daily-life scenarios are sufficiently blatant to drive the retrieval when only one concurrent source candidate belongs to the same semantic domain as the target cue. Experiment 2 was designed in order to answer this question.


Figure 4: \% of retrievals of structurally versus superficially similar source candidates in Experiment 1

## Experiment 2

## Method

## Material

Six source candidate situations were proposed before the target cue situation (taken from experiment 1): 4 unrelated stories (distractors), one superficially similar disanalog story (taken from Experiment 1) and a superficially dissimilar analog story (taken from Experiment 1, c.f. Figure 3). Hence, the design was more similar to the one used in traditional recall tasks (Gentner et al., 1993), though it differs in the isolation of structural or superficial similarities in different source candidate stories that are in competition.

## Procedure and experimental design

The procedure and experimental design were the same as in Experiment 1.

## Participants

67 students ( 52 women and 15 men, mean age 20.8 years) accepted to take part in the experiment during a class (University Paris 8).

## Results and discussion

3 responses could not be interpreted since several source candidate situations were reported despite the instruction asking for only one.

[^446]Again, analyses were focused on the comparison between surface and structure similarities-based retrievals. As illustrated in Figure 5, the superficially dissimilar analog source candidate was significantly more retrieved than the superficially similar disanalog source candidate (respectively $81.25 \%$ and $\left.18.75 \%, X^{2}(1, N=62)=22.41, p<0.001\right)^{2}$.

These results share a similar pattern with the ones obtained from Experiment 1. This comparison induces that the presence of multiple situations belonging to the same semantic domain as the target cue was not a determinant factor promoting the superiority of structural similarities in retrieval. Instead, the fact that our stimuli depicted daily-life situations might have been a critical parameter so that the participants may have rely on the abstract structures of the scenarios as retrieval cues.


Figure 5: \% of retrievals of structurally versus superficially similar source candidates in Experiment 2

## Conclusion

In Experiment 1, the superiority of structural similarities in retrieval was observed while the source analog was in competition with several source candidates sharing exclusively surface features with the target cue. As noted by Wharton et al. (1994, 1996), providing participants with only one source candidate sharing objects with the target cue may provoke its retrieval. However, in Experiment 2, providing participants with only one surface matching source candidate in competition with the superficially dissimilar analog did not reduce the proportion of structural similarity-based retrievals. Thus, structural similarity-based retrievals are predominant when the situations experimentally provided are close to the ones that are encountered in daily-life.

Experimental studies have widely converged on the conclusion that retrieval is driven by superficial similarities. However, as the analog situations that have been mainly studied are unfamiliar for the participants, the latter conclusion cannot be applied to the more ecological retrievals that are processed in daily-life. Indeed, in analogical problem
solving, analog problems share a highly abstract resolution principle (Gick \& Holyoak, 1983; Ross, 1987). Alternatively, the nature of retrieval can be better informed when meaningful structural similarities are set between situations potentially encountered by the participants in their familiar environment. Under these conditions, the participant's knowledge allows to encode familiar relations that constitute cues for retrieving former episodes, while surface features may be neglected (Novick, 1988; Hofstadter \& Sander, 2013). Indeed, a filter has to operate in order to identify the relevant properties constitutive of a concept that allows to make sense of the situation.

Some authors have claimed that the commonly observed surface similarity-based retrieval was not so detrimental to our cognition since situations sharing surface generally also share structure (the kind world hypothesis; Gentner \& Medina, 1998; Trench \& Minervino, 2015). It is noticeable indeed that objects in our environment usually interact in regular ways and have typical relations (Bassok, Wu \& Olseth, 1995). For instance, situations involving two pizzaioli in the same place potentially induce very closed relations, such as a competition between them (Figure 3). Two situations sharing both surface and structure can only be more structurally similar than two surface dissimilar situations sharing only structure at a certain level of abstraction. Yet, it has been taken as granted for advocating the superficially driven retrieval view that retrievals of structurally and superficially similar situations were more frequent than only structurally similar ones (Trench \& Minervino, 2015). An attempt to introduce a source candidate sharing only surface and no structure with the target cue has been made in storyrecall paradigms (Gentner, 1993). Yet, a closer look at the stimuli (Figure 1) makes apparent that the superficially similar disanalog source candidate situations systematically still shared some relational features with the target cue (Wharton et al., 1996). Their structures are highly similar (making a deal to avoid a bad situation) until opposite conclusions at the end of the stories (betrayal or respect of the deal). However, a set of objects still can induce a heterogeneous panel of relations (e.g. two pizzaioli: rivalry, friendship, etc), while different types of objects can induce very similar relations (e.g. a loving couple can also induce the relation rivalry for instance, c.f. Figure 3). In our experiments, dissociating object similarities and similarities in terms of familiar relations into different source candidate situations demonstrated that it is not the objects in themselves that drive access, but rather the familiar structural relations that link them.

## References

Blanchette, I., \& Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. Memory \& cognition, 28(1), 108-124.
Catrambone, R., \& Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(6), 1147.

Catrambone, R. (2002). The effects of surface and structural feature matches on the access of story analogs. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(2), 318.
Dunbar, K., \& Blanchette, I. (2001). The in vivo/in vitro approach to cognition: The case of analogy. Trends in cognitive sciences, 5(8), 334-339.
Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. Cognitive Science, 7(2), 155-170.
Gentner, D., \& Colhoun, J. (2010). Analogical processes in human thinking and learning. In Towards a theory of thinking (pp. 35-48). Springer Berlin Heidelberg.
Gentner, D., Rattermann, M. J., \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive psychology, 25(4), 524-575.
Gentner, D., Loewenstein, J., Thompson, L., \& Forbus, K. D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. Cognitive science, 33(8), 1343-1382.
Gentner, D., \& Medina, J. (1998). Similarity and the development of rules. Cognition, 65(2), 263-297.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15(1), 1-38.
Hofstadter, D., \& Sander, E. (2013). Surfaces and essences: Analogy as the fuel and fire of thinking. Basic Books.
Holyoak, K. J., \& Koh, K. (1987). Surface and structural similarity in analogical transfer. Memory \& Cognition, 15(4), 332-340.
Kretz, D. R., \& Krawczyk, D. C. (2007). Expert analogy use in a naturalistic setting. Psychological perspectives on expertise, 108.
Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3), 510.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(4), 629.
Trench, M., \& Minervino, R. A. (2015). The role of surface similarity in analogical retrieval: Bridging the gap between the naturalistic and the experimental traditions. Cognitive science,39(6), 1292-1319.
Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. D., \& Melz, E. R. (1994). Below the surface: Analogical similarity and retrieval competition in reminding. Cognitive Psychology, 26(1), 64-101.
Wharton, C. M., Holyoak, K. J., \& Lange, T. E. (1996). Remote analogical reminding. Memory \& Cognition, 24(5), 629-643.

# Improving a Fundamental Measure of Lexical Association 

## Gabriel Recchia (glr29@cam.ac.uk)

Paul Nulty (pgn26@cam.ac.uk)

Cambridge Centre for Digital Knowledge, CRASSH, Alison Richard Building, 7 West Rd<br>University of Cambridge, Cambridge CB3 9DP, United Kingdom


#### Abstract

Pointwise mutual information (PMI), a simple measure of lexical association, is part of several algorithms used as models of lexical semantic memory. Typically, it is used as a component of more complex distributional models rather than in isolation. We show that when two simple techniques are applied-(1) down-weighting co-occurrences involving lowfrequency words in order to address PMI's so-called "frequency bias," and (2) defining co-occurrences as counts of "events in which instances of word $_{1}$ and word ${ }_{2}$ co-occur in a context" rather than "contexts in which word ${ }_{1}$ and word $_{2}$ co-occur"-then PMI outperforms default parameterizations of word embedding models in terms of how closely it matches human relatedness judgments. We also identify which downweighting techniques are most helpful. The results suggest that simple measures may be capable of modeling certain phenomena in semantic memory, and that complex models which incorporate PMI might be improved with these modifications.


Keywords: semantic spaces; word space models; semantic memory; semantic networks; computational models

## Introduction

Pointwise mutual information (PMI) is a simple measure that plays an important role in many computational models that approximate human judgments of lexical association or semantic relatedness. Such "semantic space" models typically take the form of algorithms that process a corpus of written language, such as Wikipedia or TASA, and construct quantitative representations of the words they encounter on the basis of lexical co-occurrence statistics. The resulting 'lexical representations' (e.g., numerical vectors) are intended to correspond roughly to semantic representations in the human mind, at least at some level of abstraction. Of particular interest is the degree of association that exists between related (and unrelated) words in any such model. This quantity is computed in a manner appropriate to the model at hand, e.g. cosine similarity between two lexical vectors in a vector space model, or Kullback-Leibler divergence between distributions of words over topics in a topic model. Such computationally estimated associations can then be compared to behavioral data that provides evidence of the actual degree to which people perceive particular words to be related, e.g., human judgments of the semantic relatedness of large numbers of word pairs.

Such correlations with behavioral data are frequently used to argue in favor of particular models of human semantic memory (Griffiths, Steyvers, \& Tenenbaum, 2007; Jones \& Mewhort, 2007; Bullinaria \& Levy, 2007), but
lexical associations derived from semantic space models have many other applications as well. For example, a range of semantic space models-including one method that has been recently shown by Levy \& Goldberg (2014) to be implicitly factorizing a matrix of PMI scores-have recently been employed to study associative processing in high-level judgment, modeling phenomena such as the conjunction fallacy and naturalistic judgment problems (Bhatia, 2017). PMI or explicitly PMI-based methods have been used to cluster terms syntactically and semantically (Bullinaria \& Levy, 2007, 2012), recognize synonyms (Turney, 2001), automatically identify clusters that correspond to different senses of a word's meaning (Pantel \& Lin, 2002), extract linguistic collocations from text (Manning \& Schütze, 1999), and identify patterns of relationships between symptoms in dementia (Mitnitski, Richard, \& Rockwood, 2014), among many other applications.

Because of the range of applications to which PMI and PMI-based methods are applied, any modifications that improved PMI's ability to model human semantic judgments would potentially have benefits for the wide range of computational methods in which it is a component. Furthermore, if a slight modification of some neurally plausible algorithm such as PMI was to produce lexical associations that were as good as those produced by state-of-the-art models (in terms of correlation to human data), it would be worth investigating as a possible computational simplification/abstraction of some process actually taking place within human semantic memory. Finally, simple, computationally efficient yet accurate means of estimating lexical associations are useful within the field of artificial intelligence, as they can more readily be scaled up to larger datasets than can methods that take longer to compute. For all of these reasons, simple measures of lexical association are worthy of closer investigation.

PMI is traditionally defined as follows (Church \& Hanks, 1989):

$$
\operatorname{PMI}(x, y)=\log _{2} \frac{P(x, y)}{P(x) P(y)}
$$

This formulation "compares the probability of observing $x$ and $y$ together (the joint probability) with the probabilities of observing $x$ and $y$ independently (chance)" (Church \& Hanks, 1989, p. 77). Estimating these probabilities is commonly done in a straightforward manner: $\mathrm{P}(x, y)$ is estimated by dividing the number of "contexts" (documents, windows of text, etc.) in which $x$ and $y$ co-occur by the total number of contexts in the corpus, and $\mathrm{P}(x)$ is estimated by
dividing the number of contexts containing $x$ by the total number of contexts in the corpus (and likewise for $\mathrm{P}(y)$ ) (Manning \& Shütze, 1999; Turney \& Pantel, 2010).

## Strengths and Weaknesses of PMI

PMI is a component of many different algorithms that have been fit to behavioral data in the psychological literature. For example, a slight variant of it (PPMI, or 'positive PMI,' which differs only in that negative values are set to zero) has been used directly in lexical vector components in models such as 'PPMI Cosines' (Bullinaria \& Levy, 2007, 2012), and as a preprocessing step to be applied to a matrix prior to singular value decomposition or other matrix factorization techniques. Some algorithms that initially seemed to have little to do with PMI are more linked to it than they first appeared. For example, consider the SGNS algorithm of the popular word embedding tool word2vec (Mikolov, Sustskever, Chen, Corrado, \& Dean, 2013), which has been recently used in studies of metaphor perception and associative processing (Agres et al., 2016; Bhatia, 2017), and is responsible for the Google Word2Vec dataset recently described as one of several "data sets with potential relevance for cognitive science" in a recent survey (Goldstone \& Lupyan, 2015, Table 2). Although this algorithm is typically conceived of as a shallow neural network, its core mathematical operations have been shown to be implicitly factorizing the "well-known word-context PMI matrix from the word-similarity literature, shifted by a constant offset" (Levy \& Goldberg, 2014, p. 2177). The same appears to be true of an alternative embedding method known as noise-contrastive estimation (Levy \& Goldberg, 2014). In fact, much of the advantage that "predictionbased ${ }^{1 "}$ models such as word2vec's SGNS initially seemed to hold over more traditional distributional models (Baroni, Dinu, \& Kruszewski, 2014) appears to be due to word2vec's exploitation of 'hyperparameters' -i.e., miscellaneous operations such as smoothing and subsampling (Levy, Goldberg, \& Dagan, 2015). When these more traditional vector space models are enhanced with analogous hyperparameters, they tend to do as well as prediction-based models (Levy et al., 2015).

Given the ubiquity of PMI in computational models of semantic relatedness, it seems that this measure must be capturing something important. Yet the measure is wellknown for its weaknesses. The most fundamental of these is "frequency bias," PMI's tendency to over-weight cooccurrences involving low-frequency words (Levy et al., 2015; Manning \& Shütze, 1999; Turney \& Pantel, 2010). One way to think about the cause of this problem is that although probability estimates are more accurate when they are made on the basis of lots of data (e.g., frequent words) than on sparse data (infrequent words), the formula for PMI

[^447]does not account for this fact. On the contrary, the less frequent the words, the lower the denominator and the larger the result. Thus a chance co-occurrence between two rare words that each occur only once in a large corpus will result in an exceedingly high PMI. Because Zipf's law entails that any corpus will have many more infrequent than frequent lexical types, this problem is pervasive.

## Addressing PMI's Weaknesses

Given the fundamental difficulties inherent in estimating cooccurrence probabilities from infrequent words, various adjustments to PMI have been proposed to mitigate the problem. Here we consider one commonly proposed solution (down-weighting co-occurrences involving lowfrequency words in some way, to counter PMI's tendency to over-weight them), and one solution that we have not previously seen proposed (adjusting how 'co-occurrences' are defined/counted).

Down-weighting. The probabilities in the denominator of the PMI formula naturally down-weight co-occurrences involving frequent words. This is a desired property; without the denominator, the most "associated" words with virtually any term would be "the," "of," and many other words that occur very frequently across the board. As previously mentioned, however, PMI (and PPMI) have the opposite problem, in that the words these measures deem to be most semantically related to a word $w$ "are often extremely rare words, which do not necessarily appear in the respective representations of words that are semantically similar to w" (Levy et al., 2015, p. 213).

As such, several modifications to PMI have been proposed, many of which are enumerated in Table 1. The ultimate goal of all of these is to cause low frequency words to be ranked less highly than in the standard PMI formula. Some further adjustments have been proposed which rely on information other than the co-occurrence counts and frequencies of the words whose association is being calculated. Because these rely on additional information, there is sometimes a fine line between such modifications of PMI and novel distributional models, and they often have additional parameters. Yet other measures, such as $\mathrm{PMI}^{2}$, have been shown to be monotonic transformations of other measures already appearing in Table 1 (Evert, 2005). We confine our comparisons in Study 1 to only the simplest measures, i.e., measures that, when computing the degree of association between words $w_{1}$ and $w_{2}$, rely only upon the corpus-wide counts $f\left(w_{1}\right)$ and $f\left(w_{2}\right)$, and the co-occurrence counts $f\left(w_{1}, w_{2}\right)$.

Counting. It is clear that there is variation in the literature with respect to the manner in which the probabilities involved in PMI are estimated. For example, several researchers report estimating $\mathrm{P}(x)$ the as frequency of $x$ divided by the number of words in the corpus (Church \& Hanks, 1989; Islam \& Inkpen, 2008), while others use the number of documents in which $x$ appears divided by the
number of documents in the corpus (Manning \& Shütze, 1999; Turney, 2001; Turney \& Pantel, 2010). Similarly, many authors mention that they use the "number of cooccurrences" of $x$ and $y$ to estimate $\mathrm{P}(x, y)$, without specifying exactly what counts as a "co-occurrence." A reasonable assumption is that in some cases this is shorthand for "the number of contexts/documents in which $x$ and $y$ co-occur," and indeed this seems to be approach of some authors who spell out their calculations in detail (Manning \& Shütze, 1999; Turney, 2001; Turney \& Pantel, 2010). A more literal interpretation of "number of co-occurrences"-and perhaps the one intended by at least some of the authors who have used this phrase-would be that this refers to the number of co-occurrence events. For example, in the sentence context "Tiger, tiger, burning bright," the word type tiger can be conceived of as cooccurring with bright twice (one co-occurrence for each instance of tiger $)^{2,3}$. We will refer to this method of cooccurrence counting as "event-based counting," as contrasted from the "context-based" method of counting the number of contexts/documents in which $x$ and $y$ appear together. Event-based counting attends to the information available in the corpus at a more fine-grained level than does context-based counting, as it distinguishes between contexts in which word pairs might appear many times and contexts in which they might appear together only a single time. As such, it can be seen as increasing the overall amount of evidence about word associations that go into the estimation of the probabilities.

## Study 1

Down-weighting and event-based counting each have the potential to address PMI's frequency bias-the former by compensating for the fact that rarer words provide weaker evidence, and the latter by bolstering the overall amount of evidence that the measure takes into account. In Study 1, the success of each approach is evaluated individually and in combination. Table 1 provides the formulae for each of the down-weighting methods surveyed in the previous section, with citations provided in footnotes. Some methods, namely $\mathrm{SCI}, \quad \mathrm{SCI}_{\text {sig }}$, and context distribution smoothing, are asymmetric and distinguish between a cue word $x$ and a response word $y$.

In theory, either context-based or event-based counting could be used with any one of these measures. With contextbased counting, $P(x, y)$ is estimated by dividing the total number of contexts in which $x$ and $y$ appear together by a constant factor, namely the total number of contexts in the corpus (Turney \& Pantel, 2010). Analogously, with eventbased counting, it makes sense to divide the number of co-

[^448]occurrence events in which $x$ and $y$ appear together by the total number of co-occurrence events in the corpus $\sum_{i}^{N} \mid$ context $_{i} \mid\left(\mid\right.$ context $\left._{i} \mid-1\right)$. In practice, however, the specific value here is irrelevant, as it merely serves to scale all PMI scores by a constant factor.

Analogously, to estimate the 'global' or 'corpus-wide' probability $\mathrm{P}(x)$ of observing a word, we can either count the total number of contexts in which $x$ appears (contextbased counting), or we can count $x$ 's raw frequency - the total number of times $x$ appears anywhere in the corpus (event-based counting), and divide the result by the relevant constant factor (number of contexts, or number of cooccurrence events).

Some of the measures in Table 1 call for the use of cooccurrence frequencies $f(x, y)$ or global frequencies $(f(x)$, $f(y))$. These are counted as previously described, except that they are not divided by a constant factor.

Table 1: Methods for down-weighting PMI scores.

| Method | Formula |
| :---: | :---: |
| "Discount factor" ${ }^{4}$ | $\left(\frac{f(x, y)}{f(x, y)+1}\right)\left(\frac{\min (f(x), f(y))}{\min (f(x), f(y))+1}\right) p m i$ |
| $\mathrm{SCI}^{5}$ | $\frac{P(x, y)}{P(x) \sqrt{P(y)}}$ |
| $\mathrm{PMI}_{\text {sig }}{ }^{5}$ | $\sqrt{\min (P(x), P(y))}\left(\frac{P(x, y)}{P(x) P(y)}\right)$ |
| $\mathrm{SCl}_{\text {sig }}{ }^{5}$ | $\sqrt{\min (P(x) P(y))}\left(\frac{P(x, y)}{P(x) \sqrt{P(y)}}\right)$ |
| gmean ${ }^{6}$ | $\frac{f(x, y)}{\sqrt{f(x) f(y)}}$ |
| Context distribution smoothing ${ }^{7}$ | $\log \left(\frac{P(x, y)}{P(x) \frac{f(y)^{\alpha}}{\sum_{i} f(i)^{\alpha}}}\right)$ with $\alpha=0.75$ |

## Method

Word pair lists were obtained for all semantic relatedness tasks evaluated in Recchia and Jones (2009), namely the tasks of Miller \& Charles (1991), Resnik (1995), Rubenstein \& Goodenough (1965), and Finkelstein et al. (2002). Because the latter task conflates judgments of semantic similarity (\{car, truck\}) with judgments of semantic relatedness (\{car, road\}), we used the version of this task that had been partitioned into the so-called "WordSim Similarity" and "WordSim Relatedness" subsets (Agirre et al., 2009). Also included was an additional similarity task, SimLex-999 (Hill, Reichart, \& Korhonen, 2014) and two additional relatedness tasks referred to in the literature as

[^449]MEN (Bruni, Boleda, Baroni, \& Tran, 2012) and MTurk (Radinsky, Agichtein, Gabrilovitch, \& Markovitch, 2011).

Raw PMI scores as well as each of the down-weighting metrics in Table 1 were calculated for every word pair in each relatedness and similarity task ${ }^{8}$, using a version of the Westbury Lab Wikipedia Corpus (Shaoul \& Westbury, 2010) with punctuation removed and capital letters converted to lower case. The resulting corpus contained 3,035,070 documents and approximately 1 billion words. Each metric was computed with context-based counting as well as with event-based counting as described in detail on the previous page. Rather than a window size, terms were treated as 'co-occurring' if they appeared in the same document (i.e., Wikipedia article).

Additionally, to get a sense of how these metrics stack up against what are perhaps the most popular distributional models today-the word2vec CBOW and SGNS modelswe trained each word2vec model on the same corpus using the default settings recommended by Google ${ }^{9}$, and used the resulting vectors to estimate semantic relatedness in the standard manner (e.g., computing cosines between 300dimensional vectors). Comparing to distributional models whose parameters have not been optimized for the tasks at hand is in some ways an unfair comparison. Nevertheless, word2vec's 'off-the-shelf' parameters are the ones most frequently employed when word2vec is used in real-world settings. As usual, Spearman rank correlations were computed between each metric and the human judgments provided by each relatedness and similarity task.

## Results

Down-weighting methods. The only down-weighting methods tested that were consistently as good as or better than the standard PMI formula were the discount factor of Pantel \& Lin (worse performance than raw PMI on 1 of the 8 tasks when using context-based counts, 2 tasks when when using event-based counts) and the "context distribution smoothing" of Levy et al. (worse performance than raw PMI on only 1 task, irrespective of counting method employed). All other down-weighting methods exhibited worse performance than raw PMI on over half of all tasks regardless of counting method. Table 2 illustrates Spearman rank correlations between human judgments and these best-performing down-weighting methods using context-based counting, event-based counting, and the two word2vec models.

Counting methods. Restricting ourselves to the downweighting methods that produced reliable improvements, event-based counting resulted in higher correlations to human data than did context-based counting on all tasks except for SimLex-999. Across all tasks, using event-based

[^450]rather than context-based counting increased correlations by an average of 2.7 points for context distribution smoothing, 4.2 points for the discount factor, and 4.9 points for raw PMI scores.

Table 2: Correlations with human judgments of semantic relatedness (tasks 1-5, 7) and similarity (6, 8).

| Task number (see Note below) |  |  |  |  |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :--- | :--- | :--- | :---: |
| Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |  |
| CDS, Context | .68 | .75 | .58 | .83 | .82 | .32 | .64 | .73 |  |
| DF, Context | .63 | .75 | .51 | .85 | .81 | .30 | .57 | .66 |  |
| PMI, Context | .62 | .74 | .50 | .84 | .78 | .30 | .57 | .66 |  |
| CDS, Event | $\underline{.72}$ | $\underline{.81}$ | .58 | $\underline{.87}$ | $\underline{.86}$ | .27 | $\underline{.68}$ | .76 |  |
| DF, Event | .70 | .79 | .55 | .86 | .83 | .29 | .66 | .72 |  |
| PMI, Event | .70 | .79 | .55 | .86 | .82 | .29 | .66 | .72 |  |
| SGNS | .71 | .77 | .64 | .82 | .75 | .30 | .62 | .75 |  |
| CBOW | .67 | .71 | .56 | .73 | .67 | .32 | .47 | .72 |  |

Note. CDS: context distribution smoothing, DF: discount factor; PMI: unmodified PMI; SGNS: word2vec skip-grams with negative sampling; CBOW: word2vec 'continuous bag of words'; "Context" and "Event" refer to the counting method used. Task numbers refer to the judgments of semantic relatedness/similarity compiled by 1 : Bruni et al. (2012); 2: Miller \& Charles (1991); 3: Radinsky et al. (2011); 4: Resnik (1995); 5: Rubenstein \& Goodenough (1965); 6: Hill et al. (2014); 7: WordSim-Relatedness (Agirre et al., 2009); 8: WordSim-Similarity (Agirre et al., 2009). The highest correlation for each task appears in bold.

## Discussion

Down-weighting and event-based smoothing both confer advantages when PMI is used to estimate semantic relatedness judgments. Specifically, the combination of context distribution smoothing (CDS) and event-based counting performed best for all datasets except for two. Each of these was a dataset on which the various versions of PMI all performed poorly. When SimLex-999 was constructed (Hill et al., 2014), respondents were given explicit instructions about the difference between similarity and relatedness, and told to judge similarity only. PMI has no mechanism for distinguishing between related and similar terms, and does not detect relationships between paradigmatically related terms (which tend to be similar) as well as SGNS does. It is not clear why all metrics did well on WordSim-Similarity, but one reason may be that Agirre et al. (2009) did not specifically instruct participants to rate word pairs based on their similarity. Rather, they created WordSim-Similarity with the original judgments from Finkelstein et al. (2002), which had instructions that conflated relatedness and similarity, but they excluded related word pairs that did not share a formal similarity relation (synonymy, antonymy, hyponymy, etc.)

Why does context distribution smoothing work? Given that the $\sum_{i} f(i)^{\alpha}$ term is constant for any fixed value of $\alpha$,
the only thing that really seems to distinguish CDS from the other discounting methods is its use of $\alpha$ (set to .75) in the exponent of $f(y)$. Furthermore, since $\mathrm{P}(y)$ is estimated by dividing $f(y)$ by another constant, context distribution smoothing is closely related to the much more poorly performing SCI metric of Washtell \& Markert (2009), $\frac{P(x, y)}{P(x) \sqrt{P(y)}}$, which merely raises $\mathrm{P}(y)$ to the power of .5 rather than 75 .

Why would there be anything special about .75 ? One possibility is that this value strikes the proper balance between raising $\mathrm{P}(y)$ to the value of 0 (which would ignore the frequency of $\mathrm{P}(y)$ and result in a measure that was highly correlated with $y$ 's frequency), versus raising $\mathrm{P}(y)$ to the value of 1 (yielding PMI, which is known to give outsize values to infrequent words and is thus likely inversely correlated with frequency). In other words, down-weighting may be optimal when it yields a measure that is neither positively nor negatively correlated with word frequency. This possibility is briefly explored in Study 2.

## Study 2

To find the $\alpha$ for which CDS yields a correlation with word frequency as close to zero as possible, $\alpha$ was fit so as to minimize the absolute value of the Spearman rank correlation between word frequency ${ }^{10}$ and CDS. Because there is no reason in this context to modify $\mathrm{P}(\mathrm{y})$ but not $\mathrm{P}(\mathrm{x})$, the same was done for a generalization of the gmean measure, "simple" smoothing, defined simply as $\log \left(\frac{P(x, y)}{P(x)^{\alpha} P(y)^{\alpha}}\right)$. Finally, the value of $\alpha$ that maximized correlations to human data was determined for both measures. Event-based counting was used in all cases due to its superiority over context-based counting in Study 1.

Table 3 illustrates the values of $\alpha$ that minimized the absolute value of the correlation between the measure and word frequency, while Table 4 shows values of $\alpha$ that maximized correlations with human judgments.

Table 3: Values of $\alpha$ that minimized absolute value of correlations with word frequency (Study 2)

Task number (see Note below Table 2)

| Measure | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| CDS | .85 | .77 | 1.0 | .76 | .78 | .74 | .77 | .72 |
| Simple | .91 | .82 | 1.0 | .79 | .85 | .84 | .84 | .82 |

Table 4: Values of $\alpha$ that maximized correlations with human judgments (Study 2)
Task number (see Note below Table 2)

| Measure | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| CDS | .77 | .80 | .52 | .80 | .74 | .97 | .76 | .74 |
| Simple | .85 | .81 | .76 | .81 | .81 | 1.0 | .84 | .87 |

[^451]
## Discussion

For CDS, the values of alpha that minimized the absolute value of the measure's correlation to word frequency (median .77) were not far off from the values of alpha that maximized correlations to human judgments (median .765), with the greatest discrepancies being on those tasks on which CDS did not perform well in Study 1 (\#3 and \#6). The same was true of simple smoothing (medians .84 and .825 , respectively). This suggests that explicitly finding ways to minimize the degree to which lexical measures of association are confounded with word frequency and other covariates could be a promising path toward improving their ability to model human data. Other future directions could include more in-depth exploration of why $\alpha$ so closely corresponds to those values that maximized correlations to human judgements. For example, if an experimental study showed that the same was true of study participants making judgments about the relatedness of words in an artificial language, even when this value was not equal to .75 , this would provide better evidence that the human mind employs some process that makes an explicit correction for lowfrequency events analogous to that proposed by CDS.

It should not be concluded from the results of Studies 1 and 2 that PMI is more effective in isolation than distributional models such as word2vec. It should also be noted that not all datasets are independent. For example, the word pairs in Miller \& Charles (1991) and Resnik (1995) are subsets of Rubenstein \& Goodenough (1965), so it is unsurprising that a measure that does well on one would do well on all three. Even so, the fact that the use of eventbased counting and CDS down-weighting causes PMI to generally outperform word2vec on its default settings suggests that PMI may be a better abstraction of human relatedness judgments than it is commonly understood to be. Furthermore, given that PMI has so many different applications within cognitive science and is a component of so many models of lexical processing, any improvements to this measure have the potential to improve model fits across a wide range of computational studies of cognition.

## Acknowledgments

The authors gratefully acknowledge the support of the Concept Lab and the Cambridge Centre for Digital Knowledge (CCDK) at the University of Cambridge.

## References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M. \& Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pgs. 19-27). Stroudsburg, PA: ACL.
Agres, K. R., McGregor, S., Rataj, K., Purver, M., \& Wiggins, G. A. (2016). Modeling metaphor perception with distributional semantics vector space models. In

Workshop on Computational Creativity, Concept Invention, and General Intelligence. Proceedings of 5th International Workshop, C3GI at ESSLI (pp. 1-14). New York: Springer.
Baroni, M., Dinu, G., \& Kruszewski, G. (2014). Don't count, predict! A systematic comparison of contextcounting vs. context-predicting semantic vectors. Proceedings of the Association for Computational Linguistics (pp. 238-247). Stroudsburg, PA: ACL.
Bhatia, S. (2017). Associative judgment and vector space semantics. Psychological Review, 124(1), 1-20.
Bruni, E., Boleda, G., Baroni, M., \& Tran, N. K. (2012). Distributional semantics in technicolor. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 1 (pp. 136-145). Stroudsburg, PA: ACL.
Bullinaria, J. A., \& Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39(3), 510-526.
Bullinaria, J. A., \& Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stoplists, stemming, and SVD. Behavior Research Methods, 44(3), 890-907.
Church, K. W. \& Hanks, P. (1989). Word association norms, mutual information, and lexicography. Proceedings of the Association for Computational Linguistics (pp. 76-83). Stroudsburg, PA: ACL.
Evert, S. (2005). The statistics of word cooccurrences: Word pairs and collocations. PhD thesis, IMS Stuttgart.
Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., \& Ruppin, E. (2002). Placing search in context: The concept revisited. ACM Transactions on Information Systems, 20, 116-131.
Goldstone, R. L., \& Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. Topics in Cognitive Science, 8(3), 548-568.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114(2), 211-244.
Islam, A. \& Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data, 2(2), 10:1-25.
Jones, M. N., \& Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 114(1), 1-37.
Levy, O., \& Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems (pp. 2177-2185). La Jolla: NIPS Foundation.
Levy, O., Goldberg, Y., \& Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 3, 211-225.

Manning, C. \& Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA: MIT Press.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., \& Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems (pp. 3111-3119). La Jolla: NIPS Foundation.
Miller, G. A., \& Charles, W. G. (1991). Contextual correlates of semantic similarity. Language \& Cognitive Processes, 6, 1-28.
Mitnitski, A., Richard, M., \& Rockwood, K. (2014). Network visualization to discern patterns of relationships between symptoms in dementia. Alzheimer's \& Dementia, 10(4), P752-P753.
Pantel, P., \& Lin, D. (2002). Discovering word senses from text. In Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (pp. 613-619). New York: ACM.
Radinsky, K., Agichtein, E., Gabrilovich, E. \& Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In Proceedings of the 20th international conference on the WWW (pgs. 337346). New York: ACM.

Recchia, G., \& Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. Behavior Research Methods, 41(3), 647-656.
Resnik, P. (1995). Using information content to evaluate semantic similarity. In C. S. Mellish (Ed.), Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 448-453). San Francisco: Morgan Kaufmann.
Rubenstein, H., \& Goodenough, J. (1965). Contextual correlates of synonymy. Communications of the ACM, 8 , 627-633.
Shaoul, C. \& Westbury C. (2010). The Westbury Lab Wikipedia Corpus. Edmonton, AB: University of Alberta. http://www.psych.ualberta.ca/~westburylab/downloads/w estburylab.wikicorp.download.html
Turney, P. D. (2001). Mining the web for synonyms: PMIIR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning (pp. 491502). Berlin: Springer.

Turney, P. D., \& Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37(1), 141-188.
Washtell, J., \& Markert, K. (2009). A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (pp. 628-637). Stroudsburg, PA: ACL.

# Priming the production of implications 

Alice Rees (reesa34@cardiff.ac.uk)<br>Department of Psychology, Tower Building, 70 Park Place<br>Cardiff, CF10 3AT UK<br>Lewis Bott (Bottla@cardiff.ac.uk)<br>Department of Psychology, Tower Building, 70 Park Place<br>Cardiff, CF10 3AT UK


#### Abstract

We present two experiments investigating the production of implicit constructions. Using a confederate scripting paradigm we find that after making an inference participants were more likely to subsequently produce an implicature. This effect occurred at a global and a local level and was unaffected by the perceived role of the conversational partner. Our findings demonstrate that the choice of whether to be implicit is determined by the activation levels of representations specific to implicatures and that inference and implications have overlapping processing representations.


Keywords: Priming; Scalar Implicatures; Speech Production; Inferring

## Producing implications

During conversations speakers have to make a variety of decisions about the message they wish to convey. These decisions include what lexical material to include, what syntactic forms to use, and whether or not to communicate explicitly or implicitly. When communicating implicitly the onus is on the listener to enrich the utterance to reach the speaker's intended meaning.

Research into language production has predominantly focused on speaker's choices of explicit material. That is, their choices of which lexical items to use or which syntactic constructions to use (e.g. Bock \& Levelt, 1994; Levelt, Roelofs, \& Meyer, 1999; Pickering \& Branigan, 1998). In this paper, we focus on the speaker's decision to be implicit or explicit in their speech. Consider the following:

1. A: "Did John eat the cookies?"

B: "He ate some of them."
=> John ate some but not all of the cookies.
2. A: "Have you met Lucy's new boyfriend? He's handsome and intelligent!"
B: "He's handsome."
=>He's not intelligent.
In these exchanges B's utterance conveys more than what is explicitly coded. In (1) B's utterance leads to the inference John ate some but not all of the cookies. This inference arises through the following steps (based on Grice, 1989): (i) speaker A recognises that B could have said "he ate all of them." (ii) Since B did not say this, and assuming B is cooperative, A can infer that "he ate all of them" is not true; (iii) combining what is said with the negation of the
alternative leads to the inference that "John ate some but not all of the cookies." Similar reasoning can be used for (2). Speaker B could have said "Yes, I agree" or "He is handsome and funny." By not saying these A could infer that B thinks Lucy's boyfriend is not handsome. The result of this reasoning process was termed implicatures by Grice (1989).

Since Grice's seminal work, implicatures have been analysed in great detail. For example, there are analyses from the perspective of theoretical semantics (e.g., Chierchia, 2004), acquisition (e.g., Noveck, 2001), clinical disorders (e.g. Chevallier, Wilson, Happé, \& Noveck, 2010), and sentence processing (e.g., Bott \& Noveck, 2004). What all previous work has in common, however, is that they are from the perspective of the listener, and not the speaker. Here we ask how the speaker makes the choice about whether to make an implication ${ }^{1}$.

## Why imply?

Why does a speaker imply when they could be explicit? One reason is that using implicatures is efficient for a speaker. Since articulation is much slower than speech preparation processes, reducing the amount of material to be articulated reduces this articulatory bottleneck and arguably minimises speaker effort while maximising their benefits (Grice, 1989; Horn, 2004; Levinson, 2000; Wheeldon \& Levelt, 1995). Another reason is that implicatures are used out of politeness. Implicatures can be used to maintain face (Bonnefon, Feeney, \& Villejoubert, 2009; Brown \& Levinson, 1987; Feeney \& Bonnefon, 2013; Goffman, 1967; Holtgraves \& Perdew, 2016). In face threatening contexts listeners interpret the use of implicatures as a speaker's attempt at politeness.

Efficiency and politeness provide intuitive explanations for why people use implicatures. However, it is unclear how these socio-pragmatic factors interact with the language processor. One possibility is that social factors modulate the activation of representations specific to the implicature process. That is, there are representations specific to implicatures and the activation of said representations underlie the production and comprehension of implicatures. We present two experiments to investigate this.

[^452]
## Implicit representations

During conversations interlocutors tend to repeat linguistic structures that they have recently heard or produced. This repetition is known as structural priming (e.g. Bock, 1986). Structural priming occurs throughout the language system in production (Brennan \& Clark, 1996; Bock, 1986; Levelt \& Kelter, 1982; Branigan, et al, 2000; 2005) and comprehension (Sturt, Keller, \& Dubey, 2010; Thothathiri \& Snedeker, 2008), and in different languages (Hartsuiker \& Westenberg, 2000; Scheepers, 2003). There is a general consensus that successful priming of a particular structure indicates the presence of a corresponding representation within the language system whereas unsuccessful priming indicates the absence of such a representation (Branigan, Pickering, Liversedge, Stewart, \& Urbach, 1995; Pickering \& Ferreira, 2008).

Priming is not restricted to explicit linguistic forms; Raffray, Pickering, Cai, \& Branigan (2013) found that after encountering a coerced sentence individuals were more likely to subsequently produce a coerced sentence than after a fully-formed sentence. Sentences involved coercion are ambiguous. For example, "The author finished the book" is ambiguous; the verb finish requires a complement that specifies an event. For a comprehender to interpret the sentence they must undertake an enrichment process which coerces the noun into the correct semantic type. Since individuals were more likely to subsequently produce a coerced sentence after comprehending or producing a coerced sentence than after a full-formed sentence Raffray et al. suggested that there are distinct representations corresponding to coerced and full-form sentences i.e. there are distinct representations involved in implicit and explicit language. While the sort of implicit language used by Raffray et al. is very different to Gricean implicatures, their study nonetheless demonstrates that it is possible to prime the choice between using more or less linguistic material.

Further support comes from Bott and Chemla (2016). They showed that that after deriving a Gricean enrichment participants were more likely to subsequently derive an enrichment. This held both within and between enrichment categories. They suggested that there was a mechanism which underlies the derivation of enrichment and after making an enrichment these mechanisms retain some activation which increases the likelihood of making a subsequent enrichment. However, their findings relate to comprehension (and not production), and so do not illustrate how the speaker chooses between an implicit and an explicit construction.

The success of communication can, in part, be ascribed to priming. Representations that are shared between the comprehension and production system reciprocally activate each other so that after comprehending a particular structure the speaker is more likely to use that structure. Consequently interlocutors develop similar representations of linguistic structures and thus become aligned via priming (Pickering \& Garrod, 2004). We propose that the same
occurs for the production of implicit and explicit constructions. There are specific representations which underlie implicit communication and the activation level of these determines whether or not a speaker produces an implicit construction. Thus in a dialogue if one speaker is using implicit constructions it is likely that their conversational partner will also produce implicit constructions since the representations used to comprehend the utterance will have an activation advantage over other representations that were not used. Thus these representations can be primed. We test this using a confederate-scripting priming paradigm adapted from Branigan, Pickering \& Cleland, 2000).

## Experiment 1

A participant and a confederate took turns describing and identifying a referent card from a set of four. These cards consisted of rectangles containing either one or two images (see Figure 1). Cards were displayed on two separate screens (one for the participant and one for the confederate), and neither party could see each other's screen. The referent card was identified to the speaker by being embedded in a bold square, but not to the listener. The task for the speaker was to communicate to the listener which of the cards was the referent card.

The structure of the images in the display were the same on each trial. Figure 1 shows the structure (left panel) and an example trial. The experimental cards were the A and AB cards. Here, one of the images was duplicated (a pencil in the example). This meant that to communicate that the A card was the referent, the speaker could choose between an implicit construction, "The card with the pencil," in which they relied on the listener making an inference, or an explicit construction, "The card with the just the pencil," in which a modifier removed the ambiguity about which image was the referent. Whether the participant (as speaker) chose an explicit or an implicit form was the dependent measure.

There were two forms of priming. The first was a between subjects manipulation in which one group of participants were exposed to predominantly implicit constructions and the other to predominantly explicit constructions. The second was a within-subjects manipulation in which the sequence of trials was designed to prime an implicit or an explicit construction from the participant.

Our hypotheses were as follows. If there are representations corresponding to implicatures, and if they can be activated or deactivated during conversation, we expect more explicit constructions in the explicit global priming condition than in the implicit condition. Similarly, if implicature representations can be activated at a local level, trials in which the confederate uses an implicit construction should be followed by more implicit constructions than trials where the confederate used an explicit construction.


Figure 1. Example trials. Left panel shows the object structure. Right panel shows an example trial.

## Method

Participants. 35 Cardiff University undergraduate students participated for either payment or course credit.
Materials and Design. On each trial the interlocutors were presented with four cards, each containing one or two images. Images were organized in the same structure (see Figure 1). Both interlocutors would see the same set of four cards however, on prime trials the confederate's screen would also display the description to use. The confederate's descriptions always named a single image, e.g. in Experimental trials the confederate would describe the AB card as "The card with the $[\mathrm{B}]$ ". Experimental trials referred to either the A or the AB card and filler trials referred to the C or DE card.

All trials were organized into pairs such that the confederate described a prime trial and the participant described a target trial. For experimental items there were A , and AB primes and targets, thus there were 4 primetarget combinations. There were 8 examples of each combination resulting in 32 experimental pairs. Filler items were 32 pairs of C and DE trials. An additional 8 practice pairs were presented at the start of the experiment to allow participants to get used to the experimental procedure. Consequently there were 32 experimental pairs +32 filler pairs +8 practice pairs $=144$ items in total.
Items were presented in a fixed pseudorandom order. Experimental pair presentation was alternated with filler pairs. To prevent any findings from being attributable to order effects we reversed the presentation order of the pairs to make two separate lists.

The dependent variable was the construction used by participants to describe the card in target trials. Responses which used a single, unmodified referent were coded as implicit and responses that used two referents or a modified single referent were coded as explicit.

Global priming. Global priming tested whether participants would imitate the conversational style of their partner. Since the A item was duplicated across the A and AB cards, describing the A card was potentially ambiguous. The confederate could either use an implicit description "The card with the $[\mathrm{A}]$ ", which required the participant to derive an inference (A and nothing else), or use an explicit description "The card with just the [A]". In the implicit condition the confederate described the A card implicitly, using an unmodified referent. In the explicit condition the confederate always used a modifier.

Local priming. Local priming tested whether we could prime the implicature representations on a trial by trial basis. This was achieved by manipulating the sequence of prime-target trials. There were two prime types (A cards, AB cards) and two target types (A cards, AB cards), thus there were four prime-target sequences: $\mathrm{A}-\mathrm{AB}, \mathrm{A}-\mathrm{A}, \mathrm{AB}$ $>\mathrm{AB}, \mathrm{AB}->\mathrm{A}$. In the implicit condition the confederate's description of A cards required participants to make an implicature and consequently raised the activation levels of the implicature representations. The confederate's descriptions of AB cards, conversely, blocked the implicature (since there was no card equivalent to $B$ and nothing else) and therefore lowered the activation levels of the implicature representations. Consequently A->A sequences should yield higher proportions of implicature production (unmodified single referent descriptions) than $\mathrm{AB}->\mathrm{A}$ sequences. The reverse should hold for $\mathrm{A}-\mathrm{AB}$ descriptions; rates of implicatures should be high following A trials, participants would avoid unmodified single item descriptions for AB cards and instead use a conjunction ("The card with the scooter and skateboard").

## Procedure

The confederate was a female native-English speaking student from the Cardiff University student population. The participant and confederate were sat at opposite sides of a table facing a computer screen. They could not see the other person's screen. The confederate and participant were told that they were "playing a game where they will take turns describing and identifying cards. The same set of cards will be displayed on both computer screens. If you see one with a bold border it is your turn to describe. To make a guess about which card your partner was describing press one of the four keys corresponding to the position of the card on the screen. Do not speak to your partner except to describe the card". Participants were not allowed to describe the position of the card on the screen but, if they needed their partner to, they could ask for their partner to repeat their description.

## Results

Each participant produced 32 target responses. Of the 1120 responses 22 were excluded due to experimenter error. To ensure that participants were paying attention to the confederate's descriptions we looked at key press responses to prime trials. Participants selected the correct card $98 \%$ of the time. Therefore we can be confident that they were paying attention to the confederate.

Data underwent a logit transformation and were analysed using a $2 \times 2 \times 2 \times 2$ Mixed ANOVA. Prime type (A-card and AB -card) and target type (A-card and AB-card) were within-subjects factors and description form (implicit and explicit) and list were between-subjects factors.

Global priming. Participants adopted the conversational style of their partner. When their partner was using implicatures, participants were more likely to also use implicatures $(\mathrm{F}(1,31)=125.11, p<.001)$.


Figure 2. Proportion of implicit responses in implicit and explicit group.

Local priming. We also manipulated which card was described. Whilst there was no effect of prime $(\mathrm{F}(1,31)=$ $1.98, \mathrm{p}=.169)$ or target $(\mathrm{F}(1,31)=1.88, \mathrm{p}=.180)$ there was an interaction between prime and target $(\mathrm{F}(1,31)=$ $8.08, \mathrm{p}=.008$ ). When participants had to describe an A-card target they produced more implications after they had made an inference (A-prime). When participants had to describe an AB target they produced fewer implications after they had made an inference. This is consistent with there being specific representations involved in producing implicatures.


Figure 3. Proportion of implicit responses to A and AB targets by participants in implicit group.

## Discussion

The results suggest that there are representations corresponding to implicatures that can be activated and deactivated during conversation. After comprehending an implicature the representations involved had an activation advantage over other representations that were not used. Consequently these implicature representations were more likely to be used in subsequent speech production. After cancelling an implicature, the implicature representations' activation was suppressed thereby reducing the likelihood of them being used for subsequent production.

## Experiment 2

Experiment 1 used a confederate as the interlocutor. However, we have no way of knowing whether participants believed our deception. Our results could therefore be a
consequence of participants believing that the conversational partner was an experimenter. In Experiment 2 we tested this by manipulating whether the partner was presented as an experimenter or another participant.

There is range of evidence suggesting that the participant could be influenced by the interlocutor's speech characteristics and social status (e.g. Bergen \& Grodner, 2012; Grodner \& Sedivy, 2011; Holtgraves \& Yang, 1990; 1992). For example, Grodner \& Sedivy showed that listeners were less likely to derive an inference when their interlocutor was judged to be an unreliable speaker. We therefore reasoned that our manipulation could have several possible effects on participants' utterances. One was that participants might imitate their partner more in the experimenter condition. Since the partner would be in a position of authority, participants may feel that the best strategy would be to do exactly as the experimenter did. Previous work has shown that imitation is more likely when the partner has higher authority (e.g. Bandura \& Kupers, 1964; McGuigan, 2013). Alternatively, there may be less imitation in the experimenter condition. Since the partner would now be in the participant's social outgroup, there would be less pressure to conform (e.g. Bourgeois \& Hess, 2008; Yabar et al. 2006).

Orthogonal predictions can be reached about the overall levels of implicit language use. Participants might choose to use more implications overall in the experimenter condition. Since the experimenter would generally be in a position of knowledge, there would be little risk of miscommunication by using implications. Alternatively, participants might use fewer implications because if the partner were the experimenter, participants might feel they have to be particularly informative and precise in their responses.

The basic design was exactly the same as Experiment 1. The only difference was that one group of participants were told that the partner was an experimenter and in the other group they were not. In the latter group, there was an experimenter and a confederate, whereas in the former group one experimenter played the role of both experimenter and conversational partner.

## Method

Design and materials were the same as in Experiment 1. Participants. 35 Cardiff University undergraduate students participated for either payment or course credit. Partner manipulation. There were two roles that the conversational partner could play: participant or experimenter. When the conversational partner took the role of a participant the participant was unaware of their partner's involvement in the experiment, just as in Experiment 1. However, when the conversational partner took the role of experimenter, the participant was fully aware of this. The experimenter informed the participant that they would be playing a communication game together and instructed the participant of their task.

## Results

Each participant produced 32 target responses. Of the 1280 responses 49 were removed due to experimenter error.
Partner role. Numerically, participants produced more implicit descriptions when they knew their conversational partner was the experimenter compared to when they thought their partner was another participant (see Fig. 1). Despite the numerical difference this was not statistically significant $(\mathrm{F}(1,36)=1.13, p=30$. $)$. However, experiments investigating social influences often have a larger sample than that of Experiment 2. It is possible that our manipulation was not strong enough, or that our sample size is too small. This is borne out following a Bayesian analysis (Dienes, 2011; 2014; Rouder et al., 2009). Using the default JZS prior we obtained a Bayes Factor of 0.4. This indicates that our data may not be sensitive enough to draw a strong conclusion about the partner role manipulation.


Figure 4. Proportion of implicit descriptions in implicit and explicit group with confederate as participant or experimenter.

There was no interaction between interlocutor role and conversational style $(F(1,36)=.13, p=.73, \mathrm{BF}=0.3)$.
Global priming. We replicated the findings from Experiment 1. Participants in the implicit condition produced more implicit utterances than those in the explicit condition $(\mathrm{F}(1,36)=45.72, p<.001,95 \% \mathrm{CI}=1.97-$ 3.65). This was found irrespective of interlocutor role. The global priming effect was found both when the interlocutor was the experimenter $(\mathrm{F}(1,16)=19.25, p<.001,95 \% \mathrm{CI}=$ $1.53-4.39)$ and the participant $(\mathrm{F}(1,16)=30.06, p<.001$, $95 \% \mathrm{CI}=1.65-3.68$ ).
Local priming. As in Experiment 1 there was no effect of prime $(\mathrm{F}(1,32)=.016, p=.90)$ or target $(\mathrm{F}(1,32)=3.58$, $\mathrm{p}=.068$ ). However, there was an interaction between prime type and target $(\mathrm{F}(1,32)=6.64, p=.015)$. Following an Acard prime participants descriptions of A-card targets were more implicit but when the target was an AB-card descriptions were more explicit.

The general pattern of results was the same when taking each partner role separately. However, there was no significant interaction when the partner was a participant ( F $(1,16)=3.01, p=.10, \mathrm{BF}=0.5)$ or when the partner was an experimenter $(\mathrm{F}(1,16)=4.18, p=.058, \mathrm{BF}=0.4)$. The Bayes Factors give no reason to suggest that these nonsignificant results were anything else but a lack of power.

## Discussion

The main findings from Experiment 1 were replicated: Participants were more likely to produce implicit constructions when their interlocutor was using implicit constructions than when they were using explicit constructions. These effects were shown for local and global priming manipulations.
There appeared to be no influence of the social status of the conversational partner. We found no significant main effects or interactions of the partner manipulation. Global priming effects occurred regardless of the partner role, and local priming effects showed similar patterns in both conditions but were narrowly nonsignificant. Overall, we can conclude that the priming effects we observed in Experiment 1 were not due to particular strategies adopted by participants disbelieving that the partner was another participant.

## General discussion

We presented two experiments demonstrating that the production of implicatures can be primed. After comprehending an implicature participants were more likely to subsequently produce an implicature. This effect was replicated across two studies and was found both within and between participants. Whilst implicatures are an ostensibly pragmatic phenomenon these experiments suggest that there are distinct representations underlying implicatures and it is the activation levels of these representations that are responsible for the production of implicatures.
Previous research has suggested that socio-pragmatic factors influence the decision about whether to use implicit constructions. For example, people might use implicit language to be more polite or to be more efficient (e.g. Holtgraves \& Yang, 1990, 1992; Levinson, 2000). Whilst these factors are likely to be important, our experiments show that they cannot be the only factors involved. In Experiment 1 we did not manipulate any social factors yet participants systematically varied their choice of construction across conditions. In Experiment 2 we manipulated the social status of the conversational partner but found no difference in rates of implicature production as a consequence. Taken together the experiments suggest that there are distinct representations underlying implicatures and their use is not determined entirely by socio-pragmatic factors. Instead, we propose that socio-pragmatic factors may modulate the activation levels of the implicature representations but further work is needed to address this.
Finally, the priming effects we demonstrate speak to the interaction between deriving an inference and producing an implication. Inferring and implying must necessarily use different representations (since one involves comprehension and the other production) but if they were entirely separate we would not have observed priming effects. That deriving an inference primes the production of an implication shows that the representations involved in the two processes overlap. Exactly which representations are used in both, and
which are restricted to the individual processes, is a topic for future research.

## Conclusion

Our study makes three novel contributions. We have shown (1) that people can be primed to produce Gricean implicatures (2) that there are factors other than the sociopragmatic that determine whether a speaker uses an implicit construction (3) that inferring and implying share overlapping mechanisms.

## References

Bandura, A. \& Kupers, C. J. (1964). Transmission of patterns of selfreinforcement through modeling. Journal of Abnormal and Social Psychology, 69, 1-9.
Bergen, L. \& Grodner, D. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 1450-1460.
Bock, J. K., \& Levelt, W. (1994). Language production: Grammatical encoding. In Gernsbacher, M.A (eds) Handbook of psycholinguistics. (pp. 945-84). London, UK: Academic Press.
Bonnefon, J., Feeney, A., \& Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. Cognition, 112, 249-258.
Bott, L. \& Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. Journal of Memory and Language.
Bott, L. \& Noveck, I. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. Journal of Memory and Language, 51, 437-457.
Bourgeois, P. \& Hess, U. (2008). The impact of social context on mimicry. Biological Psychology, 77, 343-352.
Branigan, H., Pickering, M., \& Cleland, A. (2000). Syntactic coordination in dialogue. Cognition, 75, 13-25.
Branigan, H., Pickering, M., Liversedge, S., Stewart, A., \& Urbach, T. (1995). Syntactic priming: Investigating the mental representation of language. Journal of Psycholinguistic Research, 24, 489-506.
Brown, P. \& Levinson, S. (1987). Politeness: Some Universals in Language Usage. Cambridge UK: Cambridge University Press.
Brown-Schmidt, S., Yoon, S.O., \& Ryskin, R. (2015). Chapter threePeople as contexts in conversation. Psychology of Learning and Motivation, 62, 59-99.
Chevallier, C., Wilson, D., Happé, F., \& Noveck, I. (2010). Scalar inferences in Autism Spectrum Disorders. Journal of Autism and Developmental Disorders, 40, 1104-1117.
Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (Ed.), Structures and Beyond. Oxford UK: Oxford University Press.
Dell, G. (1986). A spreading-activation theory of retrieval in sentence production. Psychological review, 93, 283.
Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? Perspectives on Psychological Science, 6, 274-290.
Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in Psychology, 5, 1-17.
Feeney, A. \& Bonnefon, J. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. Journal of Language and Social Psychology, 32, 181-190.
Goffman, E. (1967). On face-work. In Interaction ritual, (pp. 5-45). Chicago, IL: Aldine Publishing Company.
Grice, H. (1989). Studies in the Way of Words. Harvard MA: Harvard University Press.

Grodner, D., \& Sedivy, J. (2011). The effect of speaker specific information of pragmatic inferences. In Gibson, E., \& Pearlmutter, N. (eds) The Processing and Acquisition of Reference (pp. 239272). Cambridge, MA: MIT Press.

Holtgraves, T. (1994). Communication in context: Effects of speaker status on the comprehension of indirect requests. Journal of Experimental Psychology: Learning, Memory, and Cognition, 5, 1205-1218.
Holtgraves, T. \& Perdew, A. (2016). Politeness and the communication of uncertainty. Cognition, 154, 1-10.
Holtgraves, T., \& Yang, J.N. (1990). Politeness as universal: Crosscultural perceptions of request strategies and inferences based on their use. Journal of Personality and Social Psychology, 59, 719729.

Holtgraves, T., \& Yang, J.N., (1992). Interpersonal underpinnings of request strategies: General principles and differences due to culture and gender. Journal of Personality and Social Psychology, 42, 246-256.
Horn, L. (2004). Implicature. In Horn \& Ward (eds.) The Handbook of Pragmatics, 3-28. Oxford, UK: Blackwell.
Horton, W. \& Gerrig, R. (2002). Speakers' experiences and audience design: Knowing when and how to adjust utterances to addressees. Journal of Memory and Language, 47, 589-606.
Levelt, W. J. M. (1989). Speaking: from intention to articulation. Cambridge, MA: MIT Press.
Levelt, W. J. M., Roelofs, A., \& Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22, 1-75.
Levinson, S. (2000). Presumptive Meanings: The Theory of Generalise Conversational Implicature. Cambridge MA: MIT Press.
McGuigan, N. (2013). The influence of model status on the tendency of young children to over-imitate. Journal of Experimental Child Psychology, 116, 962-969.
Metzing, C. \& Brennan, S. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions.
Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. Cognition, 78, 165-188.
Pickering, M. \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioural and Brain Sciences, 27, 169225.

Pickering, M. J., \& Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. Journal of Memory and language, 39, 633-651.
Pickering, M. J., \& Ferreira, V. S. (2008). Structural priming: A critical review. Psychological Bulletin, 134, 427-459.
Raffray, C., Pickering, M., Cai, Z., Branigan, H. (2013). The production of coerced expressions: Evidence from priming. Journal of Memory and Language, 74, 91-106.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., \& Iverson, G. (2009). Bayesian $t$ tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin \& Review, 16(2), 225-37.
Wheeldon, L. \& Levelt, W. (1995). Monitoring the time course of phonological encoding. Journal of Memory and Language, 34, 311-334.
Yabar, Y., Johnston, L., Miles, L. Peace, V. (2006). Implicit behavioural mimicry: Investigating the impact of group membership. Journal of Nonverbal Behavior, 30, 97-113.

# How could a rational analysis model explain? 

Samuli Reijula (samuli.reijula@helsinki.fi)<br>TINT / Social and Moral Philosophy<br>PO BOX 24, 00014 University of Helsinki, Finland


#### Abstract

Rational analysis is an influential but contested account of how probabilistic modeling can be used to construct nonmechanistic but self-standing explanatory models of the mind. In this paper, I disentangle and assess several possible explanatory contributions which could be attributed to rational analysis. Although existing models suffer from evidential problems that question their explanatory power, I argue that rational analysis modeling can complement mechanistic theorizing by providing models of environmental affordances.


Keywords: probabilistic modeling; rational analysis; scientific explanation; mechanism; affordance

## 1. Introduction

During the past two decades, probabilistic modeling has become one of the most visible strands of cognitive modeling alongside connectionism, dynamical systems, and rule-based approaches. Curiously, against the general trend in the psychological sciences where theorizing is increasingly anchored in neuroscience findings, probabilistic modeling of higher cognition has been a characteristically top-down endeavor. Without making any substantial commitments about the underlying cognitive mechanisms, probabilistic modeling has been applied to complex aspects of human cognition, which still largely remain beyond the reach of mechanistic research methods. Models of human memory, categorization, causal learning, concept learning, and conditional inference, to mention a few applications, often show an impressive fit to empirical data, and the novel analyses of cognitive capacities provided by the models appear to have shed new light on the nature of the studied phenomena.

However, how does that shedding light actually occur how do such computational probabilistic models explain? Although probabilistic modeling, in principle, does not rely on any particular method of explanation (and not all models aim to be explanatory), modelers often refer to the idea of rational analysis as the account of how and why their models help us understand the mind (Anderson 1990; Oaksford \& Chater 2007). The striking claim made by rational analysis (RA) modelers is that by treating higher cognitive capacities as forms of inductive inference, we can predict behavior, and explain a lot about human cognition without making any assumptions about the underlying representations and processes. This agnosticism about implementation is typically justified by making reference to a rationality assumption: We know that human agents tend to be well-

[^453]adapted to their environment, and hence a careful analysis of the cognitive task encountered by the mind, coupled with an assumption of the optimality of human behavior in the task, results in a putatively powerful methodology of prediction and explanation.

However, it is a widely-held view in the philosophy of science that explanations, also in the cognitive sciences, should track causal mechanisms, and the way that RA purports to sidestep the evidential and explanatory problems arising from the causal complexity of cognition has given rise to a strongly polarized debate (see, e.g., Jones \& Love 2011). On the one hand, the way that the new mathematical methods in probabilistic modeling can capture the interplay of structure and learning in human thought has led to the emergence of an exciting research paradigm. On the other hand, the proponents of non-mechanistic (or even noncausal) explanation need to show when and how it is that such models genuinely explain rather than only redescribe or merely formally unify various phenomena (see Colombo \& Hartmann 2017). Failing to do that, rational analysis could simply be seen as the last breath of the autonomist dream of studying the mind independently from the brain.

The goal of this paper is to advance the debate by disentangling various explanatory contributions which can be attributed to RA models. By relying on the influential contrastive-counterfactual account of explanation, I distinguish between three possible explanatory contributions: Uncovering (a) constitutive dependencies in cognitive systems (i.e. dependencies between parts and wholes), (b) environment-behavior dependencies, and (c) environmentoptimal behavior dependencies. ${ }^{1}$ I argue that often the option (c) best describes the nature of the new understanding provided by RA models: In many cases, RA models should be interpreted as being explanatory not of human behavior as such, but of environmental affordances. Consequently, well conducted modeling of environmental affordances can complement mechanistic theorizing by providing resources for understanding the possible space of behavior of agents.

## 2. Probabilistic modeling and rational analysis

### 2.1 Procedure of rational analysis

The idea of rational analysis modeling dates back to John Anderson's work on human memory and categorization in The Adaptive Character of Thought (1990). Having already worked on his ACT* cognitive architecture, the new methodology put forward in the book reflected Anderson's
increasing worries that the research methods of the time could not really uncover cognitive mechanisms. Lacking a clear picture of what it is that cognitive mechanisms do (i.e. what the psychological explananda are), the available evidence of cognitive processes and their neural implementation was insufficient to uncover the mechanistic architecture of the human mind (Anderson 1990, pp.23-26). Compared to bottom-up research strategies, rational analysis begins from the other end:

> [...] We can understand a lot about human cognition without considering in detail what is inside the human head. Rather, we can look in detail at what is outside the human head and try to determine what would be optimal behavior given the structure of the environment and the goals of the human. (Anderson 1990, p.3)

According to Anderson, careful mathematical modeling of the environment and task structure combined with an assumption about the optimality of human behavior leads to a new self-standing research strategy for understanding the mind: "As this book is evidence, a rational analysis can stand on its own without any architectural theory" (ibid.). By providing a precise model of what the mind does, rational analysis can constrain the search space for cognitive mechanisms, and, putatively, put the scientific study of the mind on a firm foundation.

This view of the role of computational modeling immediately brings to mind Marr's (1982) account of multilevel theorizing. However, whereas Marr provides no systematic account of how computational-level theories are to be constructed, RA modeling has predominantly proceeded according to the six-step cycle proposed by Anderson (1990, p.29):

1. Specify precisely the goals of the cognitive system
2. Develop a formal model of the environment to which the system is adapted
3. Make minimal assumptions about computational limitations
4. Derive the optimal behavior function, given items 1 through 3
5. Examine the empirical evidence to see whether the predictions of the behavior function are confirmed
6. Repeat, iteratively refining the theory

These steps embody an account of how a large part of probabilistic cognitive modeling is done. However, two further assumptions of RA should be made explicit. First, the derivation of optimal behavior in steps 2-4 typically employs probability calculus (not logic) as the normative baseline theory of rational behavior. Secondly, the link between model predictions (step 4) and observed behavior of humans (step 5) is formed by an assumption about the optimality of the observed behavior (see quoted passage above).

[^454]Below I illustrate this process with an example. However, a comment on the status of the approach in cognitive science is in place: Obviously, not all probabilistic modelers endorse the rational analysis framework (see Brighton \& Gigerenzer 2008; Danks 2015; Frank 2013). Focusing on RA is useful for two reasons, however. The approach is undeniably influential, and its core commitments have been endorsed by a large group of well-known modelers (e.g., Anderson 1990; Oaksford \& Chater 1994; Griffiths \& Tenenbaum 2009). A further advantage of focusing on RA has to do with the fact that often the theoretical commitments of mathematical modelers can be hard to pin down. In some cases, the ambiguities are surely due to the modelers themselves not being clear about where their commitments (about how to understand explanatoriness, optimality, etc.) lie. Rational analysis is a clear account of the conceptual foundations of probabilistic cognitive modeling, and provides a starting point, or at least a foil, for explicating such commitments.

To illustrate the rational analysis process, I now briefly introduce Mike Oaksford and Nick Chater's $(1994,2007)$ analysis of the Wason selection task.

### 2.2 The information gain model

Wason selection task is one of the most famous laboratory experiments discussed in the literature on human rationality. In the original form of the task, participants are given four cards, each of which has a letter on one side and a number on the other. The participants' task is to determine whether the rule "If there is a vowel on one side of the card (p), then there is an even number on the other side (q)" holds. More precisely, they are asked to select those cards which must be turned over to discover whether the rule is true or false. The famous finding from the task and its several replications is that only a small minority (less than $10 \%$ ) select the correct cards (vowel, odd number) corresponding to the falsifying instance. That is, judged in the light of logic, most participants fail to perform in a rational way.

Oaksford and Chater (O\&C) challenge the irrationality claim by arguing that logic-based theories of inference and rationality misrepresent the participants' behavior in the task. O\&C's own information gain model suggests that people's apparently irrational way to test a hypothesis should actually be seen as optimal strategy for uncertainty reduction. The gist of O\&C's reinterpretation is that instead of engaging in deductive reasoning, participants interpret the task as an inductive one. They do not try to falsify the rule, but instead they try to determine which of two hypotheses holds:
a) Independence model, $\mathbf{M}_{\mathbf{I}}: \mathrm{P}(\mathrm{q} \mid \mathrm{p})=\mathrm{P}(\mathrm{q})$ or
b) Dependence model, $\mathbf{M}_{\mathrm{D}}: \mathrm{P}(\mathrm{q} \mid \mathrm{p})$ is high, higher than $\left.\mathrm{P}(\mathrm{q})\right)^{2}$

Being initially equally uncertain about both hypotheses, participants aim to reduce this uncertainty as much as possible by turning as few cards as possible.

The rational analysis proposed by O\&C relies on three core principles:

1) Higher cognition can be modeled as probabilistic (Bayesian) computation.
2) The likelihoods and prior probabilities required by the model can be acquired from the analysis of the environment structure.
3) Behavior of human agents constitutes an optimal response to the task.

The model is constructed roughly as follows. To formalize the idea of uncertainty reduction, O\&C adopt the optimal data selection paradigm, and interpret uncertainty reduction as optimizing expected information gain. Information gain $I_{g}(D)$ from turning over a card $(\mathrm{D})$ is defined as $I\left(M_{i} \mid D\right)$ $I\left(M_{i}\right)$ where the Shannon information ${ }^{3} I\left(M_{i}\right)$ can be derived from the probabilities of the hypotheses before and after observing data, $P\left(M_{i}\right)$ and $P\left(M_{i} \mid D\right)$. The required posteriors can be obtained by the Bayes' rule from the likelihoods $P\left(D \mid M_{i}\right)$ and the prior probabilities of the hypotheses. Reflecting initial ignorance, the priors were set to 0.5 and hence the rest of the crucial model specification is built into the likelihoods, which reflect the nature of the four-card task. Oaksford and Chater (1994) show in detail how the required likelihoods can be read off the contingency tables describing the two hypotheses.

From these derivations, it follows that the base rates of $p$ and $q$ have a central role in determining which behavior is optimal. They describe how frequently positive instances of the antecedent and consequent of the rule appear in the environment. Qualitatively, the expected information gain from each of the four cards turns out to depend on the base rates $\mathrm{P}(p)$ and $\mathrm{P}(q)$ in the following way:

| $\mathrm{P}(\mathrm{q})$ is small | $\rightarrow \mathrm{p}$ card is informative |
| :--- | :--- |
| $\mathrm{P}(\mathrm{p})$ is large | $\rightarrow$ not- q card is informative |
| $\mathrm{P}(\mathrm{p})$ and $\mathrm{P}(\mathrm{q})$ are small | $\rightarrow \mathrm{q}$ card is informative |
|  | (not-p card is not informative) |

How should these base rates, then, be determined? Instead of attempting to somehow measure them in relevant environments for different kinds of rules, O\&C cite various intuitively plausible justifications for their rarity assumption: Relying on the observation that categories in language cut the world quite finely, and that properties that figure in causal relations tend to be rare, the assumption states that, generally, $\mathrm{P}(p)$ and $\mathrm{P}(q)$ are small in most situations. Under rarity, O\&C conclude, the $q$ card is more informative than the not- $q$ card. Hence, the model suggests that the highest expected information gain is achieved by turning over the $p$ and $q$ cards, exactly as the majority of the participants do. In fact, with $\mathrm{P}(p)=0.22, \mathrm{P}(q)=0.27$, the model shows a very good fit to experimental data from the Wason task. Hence, by changing the normative model of rational behavior, O\&C were able to explain away irrationality, and to show that
participants' behavior in experiments is actually close to optimal.

The model has received critical attention in the literature (see Oaksford \& Chater 2009), but it serves our current purposes well. The model specification and the modeling assumptions are conceptually on a par with those in more complex Bayesian models: The complexity often pertains to the number of variables involved, the structure and generation of the hypothesis space, and in many cases advanced numerical methods are needed for solving the model. These mathematical sources of complexity do not, however, change the fundamental conceptual architecture of a model. What is common to all RA models is that none of their main components (hypothesis space, likelihood function, priors) are interpreted in a psychologically realistic way as mental representations (Jones \& Love 2011). Instead, they stand directly for properties of the environment. Furthermore, empirical data about the properties of human cognition is not fed into the model specification to calibrate or to constrain the model. Instead, behavioral data enters only in step 5 of RA (see above) as a means for testing model predictions. In this sense, the information gain model is an illuminating example of the theoretical and conceptual assumptions made in rational analysis modeling.

## 3. What rational analysis models fail to explain

There is no consensus in philosophy (or in the sciences) about what scientific explanation is, or what makes a theory explanatory. However, a shared starting point for many accounts of scientific explanation has been to distinguish explanation from other epistemic activities (e.g., description and prediction) by pointing out that explanations offer information of a specific kind. Explanations show how or why something happened or obtains. According to an influential approach (Woodward 2013), the knowledge that allows one to answer such questions concerns change-relating counterfactual dependencies between the relata in the explanation, the explanans and the explanandum. That is, explanations show how (the state of) some things depend on (the state of) other things.

This contrastive-counterfactual account of explanation suggests that explanatory information has generally the following form:
\{CC $\} y[y$ '] because of $x[x$ '] (variable Y takes the value y instead of $y$ ' because $X$ has the value $x$ instead of $x$ ')

According to the contrastive-counterfactual account, being able to explain means that one is able to correctly answer what-if-things-were-different questions, i.e. questions about how changes in explanantia variables influence the state of the explanandum variable. In addition to being a sufficiently general account of explanation, the contrastivecounterfactual account suits the purposes of this article well, because it does not necessarily tie the notion of explanation

[^455]to that of causation. That is, although the 'because' in $\{\mathrm{CC}\}$ is typically understood as referring to a causal dependency, the account does not rule out the possibility of there being also non-causal explanations (Woodward 2013; Pincock 2015; Rice 2015): If a suitable analysis of invariant dependency in non-causal contexts (e.g., for mathematical dependencies) can be found, the contrastive-counterfactual account can be applied to non-causal explanations as well. Hence, the account of explanation casts the net wide enough to give RA models a fair chance of being explanatory.

A further advantage of treating explanations as answers to questions is that it allows us to sharpen the explananda, i.e. to make more precise the possible explanatory claims arising from RA models. I suggest that there are at least three different kinds of objective dependencies that RA models could be said to track: (1) constitutive dependencies between parts and wholes, (2) environment-behavior dependencies, and (3) environment-optimal-behavior dependencies. In the rest of this section, I argue that often RA models do not have genuine explanatory import with respect to the two first kinds of dependencies. The more promising third option is discussed in section 4.

### 3.1 Constitutive what-ifs

The notion of mechanism has acquired a central position in the philosophical debates on scientific explanation. A clear expression of the mechanistic viewpoint has recently been given in the model-to-mechanism mapping (3M) requirement by Kaplan and Craver (2011). According to the requirement, dynamical and mathematical models in systems- and cognitive neuroscience can be explanatory only if there is a mapping between elements in the model and elements in the mechanism for the phenomenon. As the example discussed above suggests, rational analysis models provide no such mapping. They are agnostic about algorithmic and implementation level details, and intentionally so. They clearly do not track constitutive dependencies. Does this mean they cannot be explanatory?

As Kaplan and Craver themselves admit, their argument ultimately relies on shared norms about explanatoriness in the neuroscience community, and their account of explanation as uncovering multi-level mechanisms reflects these norms. If such norms do not hold among probabilistic cognitive modelers, it is not obvious why they, based on this argument alone, should abide by the 3 M requirement.

The contrastive-counterfactual account suggests a more positive reply to Kaplan and Craver's argument: RA models obviously do not provide information about constitutive and causal dependencies in multi-level mechanisms, but this does not rule out the possibility of them tracking some other kinds of objective dependencies, for example, those holding between relata described in purely computational-level terms. Furthermore, a proponent of RA need not (and should not) claim that adding mechanistic detail never improves a computational explanation. To defend the explanatoriness of RA models, a far weaker claim suffices, one stating that it is possible to learn about objective explanatory dependencies
without always relying on information about cognitive mechanisms.

### 3.2 Environment-behavior what-ifs

A second kind of explanatory question answered by an RA model could be: "How would the behavior of the cognizer change when the cognitive task changes in some particular way?" That is, the model could uncover objective dependencies between properties of the environment and the behavior of cognizers. For example, O\&C's model can be used to derive predictions about what the behavior of the participants in the Wason task would be, were $\mathrm{P}(p)$ and $\mathrm{P}(q)$ to take some range of values.

It is here that the optimality assumption of RA becomes crucial. To predict how human behavior would change in response to changes in the task, without knowing anything about the algorithms and processes producing the behavior, RA relies on the assumption that humans are well-adapted to their environments: If we assume that human behavior is close to optimal across a large variety of environments, the predictions derived from the RA model (step 4 of the RA procedure) should in fact apply to that behavior. Optimality forms the link between the normative theory and observed behavior.

Given that human (ir)rationality has been the topic of a longstanding debate in philosophy and psychology, it is not surprising that the optimality assumption has drawn a lot of criticism (Jones \& Love 2011). Although proponents of RA are correct in arguing that some degree of rationality of target behavior is required for us to even perceive it as intentional action, such modest levels of rationality hardly license the strong optimality assumptions of RA models. Neither do evolutionary arguments provide support for strong optimality claims: Natural selection is a source of design and adaptedness, but not necessarily of globally optimal solutions - merely a local comparative advantage is sufficient for evolutionary solutions to survive.

Being aware of these problems, proponents of RA have avoided evolutionary defenses of the optimality assumption. Instead, they often justify optimality by relying on analogies to behavioral ecology and economics, where similar assumptions are commonly made (Chater et al. 2003). However, such analogies break down due to a crucial dissimilarity between these fields. Unlike in cognitive science, both in biology and economics the rationality claims typically concern aggregate behavior, not that of individual agents. Hence, I do not see how appealing to economics or biology could be a viable way to justify optimality assumptions in RA modeling.

These problems with the general defenses of the optimality assumption suggest that perhaps optimality should be examined more locally. Now, what kind of evidence should be obtained to justify the optimality claim in the case of a particular cognitive task? It seems that to support a claim about there being an objective dependency between environment and behavior, we should gather data about human behavior in a task across a range of parameter values
describing various different environmental states. In other words, if human behavior fits the predictions of the model across a range of conditions, that would appear to be rather strong evidence of optimal performance.

However, the existing RA models rarely make use of such cross-environmental data. First of all, many models do not rely on any actual measurements of environment parameters. Instead, they use plausible-sounding assumptions or analogies. For example, Anderson (1990, ch. 2) relied on data about library borrowings to model usage of memory structures, and Griffiths et al. (2007) use Google PageRank to predict fluency of recall. Models devoid of good quality empirical data should be considered as toy models (at best), incapable of uncovering the actual properties of cognitive environments. Furthermore, Marcus and Davis (2013, table 1) suggest that Bayesian modelers have been selective in choosing the results that they report from experimental tasks, only reporting results where human behavior follows the model predictions and ignoring cases where behavior is not optimal. ${ }^{4}$

That said, in the large literature on the information gain model, predictions from the model have been tested against human performance under different base rates and different framings of the task (e.g., descriptive vs. deontic; Oaksford \& Chater 2007, Ch.6). Although the empirical findings remain inconclusive, such systematic variation of the task parameters should be used to produce evidence of a robust explanatory environment-behavior dependency. ${ }^{5}$

## 4. Rational analysis and the logic of the situation

Finally, let us examine the epistemic value of an RA model if we drop the optimality assumption. Assume that we have a model with a (i) well-specified task structure, (ii) parameter values based on measurements of the environment, and (iii) an empirically informed account of computational costs and cognitive limitations. What such a model could do is it could link combinations of parameter values to best possible behavioral choices in those situations. Is this not a kind of objective change-relating dependency? However, consider what the relata of such a dependency are. The model tells what the optimal behavior would be, given a particular combination of environmental conditions and computational limitations. Such counterfactuals do not say anything about actual human behavior. Instead, they can be seen as increasing our understanding of the environmental affordance, or, the logic of the situation (Popper 1963). ${ }^{6}$

What mathematical models of affordances (of the opportunities that the environment offers for the agent) can help us understand is the possible space of action for cognitive agents. Models of affordances show what a hypothetical rational agent would do in different situations.

[^456]For what kind of purposes could such information be useful? First, were we to design artificial cognitive systems with a particular cognitive task in mind, these systems should approximate the optimal behavior specified by the model. For example, in the selection task, if we are interested in reducing our uncertainty, O\&C's model tells us something non-trivial, i.e. which information sources to examine given the base rates of $p$ and $q$.

Secondly, as in economics, rational models can obviously act as normative baselines to which human behavior can be compared. As Sloman \& Fehrbach (2008) argue, often it is just as interesting to discover that behavior does not conform to the rational norm as to see that it does. Finding out when and how complex systems malfunction is often an efficient way to learn about the underlying processes.

However, in neither one of these cases are RA models used to directly explain human behavior. Instead, the model functions as an inferential aid which helps to chart the possible space of action for agents, when faced with a particular task. Herein lies perhaps the hardest evidential problem faced by rational analysis. How do we know what the mind really does in some situation; where do the functional hypotheses in step 1 of RA come from? For example, how would O\&C defend their Bayesian construal of the selection task against an adamant falsificationist? The currently available empirical evidence can hardly decide the issue: Where $O \& C$ see optimal behavior, the falsificationist sees well-known inferential blunders. Marcus and Davis (2013) have argued that similar problems of model selection plague several other RA models as well.

The difficulty seems to come down to the fact that the cognitive tasks and the affordances available to an organism depend on its "life space" - not the physically objective world in its totality, but reality filtered through the organism's needs, drives and perceptual apparatus. Therefore, we should not think that the researcher's intuitions are necessarily a reliable guide to what the tasks faced by different aspects of the human cognition really are. Ad-hocness in task specification, in turn, raises serious worries about the relevance of $R A$ modeling: Constructing detailed mathematical models of potential affordances is of little interest unless such affordances can be shown to be ones actually offering themselves to the human mind.

This worry suggests that the six-step rational analysis modeling cycle introduced in section 2.1 should not proceed independently from knowledge originating from mechanistic research: As both the connectionist rivals of RA and proponents of multi-level mechanistic explanation have argued, functional hypotheses (step 1 of RA) in cognitive science must be formulated in an iterative process between bottom-up and top-down research strategies (see McClelland et al. 2010; Bechtel \& Richardson 2010). In particular, knowledge of perceptual capacities and embodiment

[^457](informing step 2), as well as of the computational constraints of organisms (step 3) mostly originate from the bottom-up research on the mind-brain, and this knowledge should be allowed to constrain RA models. In this sense, Anderson's and O\&C's claims about the self-standing explanatory role of RA are not vindicated by my analysis.

However, neither can bottom-up research strategies stand on their own, or at least they fail to reach high enough. The discussions on mechanistic explanation often have a reductionist bias, and understanding the environments within which cognitive mechanisms function has not received sufficient attention. Here RA models can complement mechanistic theories of cognition by providing precise mathematical models of the task and the environment. For example, as Chater et al. (2003) point out, a correctly formulated rational analysis can show why it is that some simple heuristic can be successful in solving a computationally complex task.

## 5. Conclusions

I have argued that given a sufficiently broad account of scientific explanation, there are several possible ways in which probabilistic modeling could increase our understanding of the mind. However, the strictly computational-level approach embodied in the six-step formula of rational analysis has led to theorizing which often fails to reliably uncover genuine explanatory dependencies. The shortcomings of RA are evidential in nature: The data, and the way it is used in model construction, often cannot support the counterfactual inferences needed explaining human behavior.

My new proposal about the epistemic role of RA models is that they can be understood as models of environmental affordances. Interpreted in this way, the models do not actually provide information about the mind works, or even hypotheses about actual cognitive functions (cf. Marr 1982; Zednik \& Jäkel 2014). Instead, they help to chart the possible cognitive space of action for an organism. The nature of the explanatory contribution of such information is best worked out as a part of a non-reductionist mechanistic research programme.

## References

Anderson, J. (1990). The Adaptive Character of Thought. Hillsdale, JN: Lawrence Erlbaum Associates.
Bechtel, W., \& Richardson, R. (2010). Discovering Complexity. The MIT Press.
Brighton, H., \& Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: harmony or dissonance? In Chater \& Oaksford (eds.) The Probabilistic Mind. Oxford University Press.
Chater, N., et al. (2003). Fast, frugal, and rational: How rational norms explain behavior. Organizational Behavior and Human Decision Processes, 63-86.
Chater, N., \& Oaksford, M. (eds.) (2007). The Probabilistic Mind. Oxford: Oxford University Press.

Colombo, M., \& Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. British Journal for the Philosophy of Science, 68, 451-484.
Danks, D. (2015). Unifying the Mind. MIT Press.
Frank, M. (2013). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). Cognition, 128, 417-423.
Goodman, N. Frank, M., Griffiths, T., Tenenbaum, J., Battaglia, P., \& Hamrick, J. (2015). Relevant and robust: A response to Marcus and Davis. Psychological Science, 26, 539-541.
Griffiths, T., Steyvers, M., \& Firl, A. (2007). Google and the mind. Psychological Science, 1069-1076.
Griffiths, T., \& Tenenbaum, J. (2006). Optimal predictions in everyday cognition. Psychological Science, 767-773.

- (2009). Theory-based causal induction. Psychological Review, 661-716.
Jones, M., \& Love, B. (2011). Bayesian fundamentalism or enlightenment? Behavioral and Brain Sciences, 34, 169231.

Kaplan, D., \& Craver, C. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. Philosophy of Science, 78, 601627.

Marcus, G., \& Davis, E. (2013). How robust are probabilistic models of higher-level cognition? Psychological Science, 24, 2351-2360.
Marr, D. (1982/2010). Vision. W.H. Freeman/MIT Press.
Oaksford, M., \& Chater, N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review, 101, 608-631.

- (2007). Bayesian Rationality: The probabilistic approach to human reasoning. Oxford University Press.
- (2009). Precis of Bayesian Rationality. Behavioral and Brain Sciences, 69-120.
Pincock, C. (2015). Abstract explanations in science. British Journal for the Philosophy of Science 66, 857-882.
Popper, K. (1963). Models, instruments, and truth. Manuscript. Karl Popper Collection at the Hoover Institution Archives at Stanford University.
Rice, C. (2015). Moving beyond causes: Optimality models and scientific explanation. Noûs 49, 589-615.
Sloman, S., \& Fehrbach, P. (2008). The value of rational analysis: as assessment of causal reasoning and learning. In The Probabilistic Mind.
Woodward, J. (2013). Mechanistic explanation: Its scope and limits. Proceedings of the Aristotelian Society Supplementary Volume, lxxxvii: 39-65.
Zednik, C., \& Jäkel, F. (2014). How does Bayesian reverseengineering work? In P. Bello et al. (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society, 666-671.


# Randomness in binary sequences: Conceptualizing and connecting two recent developments 

Stian Reimers (stian.reimers@city.ac.uk)<br>Department of Psychology, City, University of London<br>Northampton Square, London EC1V 0HB, United Kingdom


#### Abstract

Recent theoretical research has shown that the assumptions that both laypeople and researchers make about random sequences can be erroneous. One strand of research showed that the probability of non-occurrence of streaks of repeated outcomes (e.g., HHHHHH) is much higher than that for a more irregular sequence (e.g., HTTHTH) in short series of coin flips. This tallies with human judgments of their likelihood of occurrence, which have conventionally been characterized as inaccurate and heuristic-driven. Another strand of research has shown that patterns of hits and misses in games like basketball, traditionally seen as evidence for the absence of a hot-hand effect, actually support the presence of the effect. I argue that a useful way of conceptualizing these two distinct phenomena is in terms of the distribution of different sequences of outcomes over time: Specifically, that streaks of a repeated outcome cluster whereas less regular patterns are more evenly distributed.


Keywords: randomness; rationality; hot hand fallacy, gambler's fallacy.

## Introduction

One of the more important things that organisms must do to prosper is to identify, extract, and act on patterns in the environment. At a perceptual level, detecting potential threats in a noisy and ambiguous environment is crucial for survival. At a higher level, the ability to detect patterns of events over time or space, such as the presence of absence of prey in different locations at times, the changes in temperature or weather, and so on, allows an organism to predict the future state of the world, and adapt behavior accordingly. In a more contemporary environment, anyone able to detect behavioral patterns in markets, organizations or individuals would be able to exploit that knowledge to their benefit.
In order to detect patterns, an organism has to separate signal from noise. As such, one would expect organisms to accurately represent the absence of a signal, that is, randomness. A poor representation of what random patterns look like would make it harder to spot the times when patterns contain information.
As such, it is surprising that across a wide range of research procedures, people are systematically poor at representing randomness (for reviews see, Nickerson, 2002, 2004; Bar-Hillel \& Wagenaar, 1991; Falk \& Konold, 1997; Rapaport \& Budescu, 1992). For example, people underestimate the frequency of 'streaks' or 'runs' of a particular outcome (such as getting five heads in a row when flipping a coin repeatedly), and treat such streaks
when they appear as evidence for non-randomness. Related, people rate sequences of binary outcomes containing negative serial dependency (that is, an alternation rate between outcomes of greater than .5 ), as being more random than truly random sequences.

One reason why people may be poor is that many properties of random sequences are counterintuitive. For example, relative wait times for different sequences of binary outcomes violate transitivity (see, e.g., Nickerson, 2007).

In this paper I focus on sequences of binary, Bernoulli i.i.d. events such as coin flips (which could come down H or T ), and behaviors which may be modelled by them, such as basketball shots (which could come down as a hit $-\mathrm{X}-$ or a miss - O).

## Representativeness and probability of occurrence

One of the most influential studies to demonstrate an apparent bias in perception of randomness was that of Kahneman and Tversky (1972). In their studies, they asked participants about the relative frequency of occurrence of different birth orders of girls (G) and boys (B) in families with six children in a hypothetical city. They found that participants estimated that there would be many fewer examples of a precise sequence of BGBBBB relative to a precise sequence of GBGBBG. (Of course all precise sequences of birth orders are equiprobable). To examine whether this finding was a result of just the relative frequency of B and G , Kahneman and Tversky also compared estimates of the relative frequency of BBBGGG and GBGBBG, finding that the former was seen as significantly less probable than the latter. Thus, both the relative frequency of outcomes, and the order in which outcomes occur appear to be important in judging the probability of occurrence.

Traditionally, findings of this nature have been explained in terms of heuristics and biases, specifically a misapplication of a representativeness heuristic (but see, e.g., Gigerenzer, 1996; Ayton \& Fischer, 2004): People believe that the properties of short sequences of random outcomes should be representative of those seen in longer sequences (e.g., equal proportions of outcomes, an absence of structure or compressibility), and sequences that share those properties are deemed more probable.

However, recently Hahn and Warren (2009) observed that in situations where one looks for patterns of outcomes in a
finite sequence of, for example, coin flips, different sequences have different probabilities of occurrence.

To give a concrete example (used by Hahn and Warren, 2009), compare the probability of non-occurrence of a HHHH vs HHHT in a sequence of 20 coin flips. The streak of a repeated outcome $(\mathrm{HHHH})$ is around twice as likely not to occur relative to HHHT. The argument made by the authors is that if people use previous experience of merely the occurrence (at least once) or absence of a particular string to judge the probability of occurrence in the future, then they would be quite accurate to say that HHHH would be less likely to occur in a sequence of 20 coin flips than HHHT.
This was also extended to account for the gambler's fallacy: If experience dictates that HHHH is less likely to occur than HHHT, then an individual who sees HHH and is asked to be on whether the next observation is H or T , would with some justification bet on $T$.
Although there is ongoing discussion about the extent to which or circumstances under which Hahn and Warren's theory predicts judgments (Reimers, Donkin \& Le Pelley, 2017), the observation that different strings of outcomes have different probabilities has meant that researchers have needed to reconsider what normative baselines for randomness judgement should be, and potentially turn what previously appeared to be a clear bias into a slight misapplication of a genuine property of the environment.

## The hot-hand-fallacy fallacy

A second challenge to researchers' assumptions about normative baselines has been seen with the hot hand effect.
The hot hand effect is a phenomenon - accepted as selfevidence by many sports participants and spectators - that players go through periods when their performance varies consistently over time, having streaks when they are 'hot', and during that period of time their performance is consistently better than usual, as measured by, for example, the proportion of baskets or putts they manage to sink. If the hot hand were real, it would mean that probability of success had positive autocorrelation: Following a streak of hits, a person would be more likely to score another hit.

Despite popular belief in the hot hand phenomenon, the effect has until recently been seen as a fallacy. Gilovich, Valone and Tversky (1985) examined the performance of professional and amateur basketball players, and argued that there was no evidence for a hot hand effect. They operationalized a hot hand effect in basketball shooting as a difference between the probability of getting a hit (scoring from a free throw) after a streak of $k$ consecutive hits (X) and the probability of getting a hit after a streak of $k$ successive misses ( O ), for example, $\mathrm{p}(\mathrm{X} \mid \mathrm{XXX})>$ $\mathrm{p}(\mathrm{X} \mid \mathrm{OOO})$. The logic, which appears superficially entirely reasonable, was that if the probabilities of a hit after $k$ hits and a hit after $k$ misses were identical in a well-power study, then that provided evidence for the absence of a hot-hand effect. Across several studies, they found no difference in
probabilities, so concluded that the hot-hand effect was a fallacy.

Recently this conclusion has been challenged. Miller and Sanjurjo (2016) note problems with measures traditionally used to support the absence of a hot hand effect (see Rinott and Bar-Hillel, 2015 for less technical overview of an earlier version). Specifically, they prove that if one were to calculate the strength of the hot hand effect for players individually along the lines of calculating $\mathrm{p}(\mathrm{X} \mid \mathrm{XXX})$ / $[p(X \mid X X X)+p(X \mid O O O)]$, and then take the average across individual, that average would be less than .50 . So if a wellpowered study shows a mean proportion of around .50 , then rather than being evidence against a hot hand effect, it is in fact substantial evidence for such an effect.

Miller and Sanjurjo (2016) prove the counterintuitive finding that that in any finite binary sequence, the mean proportion of streaks of length $k$ that are followed by a repetition of the same outcome is on average lower than the proportion of streaks of length $k$ that are followed by the opposite outcome. They note that for $k=1$, the effect is entirely driven by a sampling-without-replacement effect, such that in, say, a short sequence of coin flips where the number of heads and tails is expected to be identical, choosing to look at outcomes following a H removes a H from the sample, meaning that the probability of all other observations, including the next one being a T is slightly greater than .5 .

More relevant for this discussion is the effect where $k>1$. Here, Miller and Sanjurjo note that the effect is driven much more by the extent to which sequences of outcomes can overlap with each other (or show autocorrelation, in the terminology of Guibas \& Odlyzko, 1981). They note that some sequences of outcomes can overlap with themselves: For example the sequence HHH can partially overlap with itself such that in a series of five coin flips, it is possible to observe three overlapping instances: HHHHH; conversely, the sequence HTT cannot overlap at all, and so can only occur once in a series of five coin flips. They note that because overall the expected number of occurrence of HHH and HTT must be identical, HTT must be observed in a greater number of series of five coin flips to compensate for the fact that HHH can occur multiple times within a single series. As such, they prove that

## Variance of occurrences in short sequences

Here, in contrast to Miller and Sanjurjo's (2016) formal proof, a stochastic approach to this issue is taken, in part to make the relationship between the findings of Hahn and Warren (2009) and Miller and Sanjurjo (2016), and in part to attempt to show how the varying distribution of observations of different sequences of outcomes in a longer series of binary outcomes can account for both findings.

This is not the first attempt to relate these two phenomena. In recent iterations of their working paper, Miller and Sanjurjo have attempted to account for the gambler's fallacy as well as the hot hand fallacy, by assuming a degree of insensitivity to sample size. Sun and


Figure 1: Raster plot of the occurrence of strings of HHHH or HHHT across 1,000 simulated coin flips. The horizontal dimension gives the flip from the first on the far left to to the last on the far right.

Wang (2010) note that different forms of waiting time for sequences of outcomes vary differently with outcome. Thus, the mean inter-observation gap is the same for all sequences of a single length, whereas the expected waiting time from first flip of a coin is much greater for some sequences of outcome (such as HHHH) than others (HHHT), and that the variances in these two forms of waiting time vary substantially.

The argument presented here is based on the observation that the variance of the number of trials between observations of a given sequence of outcomes varies. Specifically, the observations made by Hahn and Warren, and those made by Miller and Sanjurjo are both consequences of the same property of random sequences, specifically that within any finite sequence of equiprobable binary outcomes, the distribution of frequency-ofoccurrence for 'streaks' (i.e. repetitions of the same outcome, like HHHH ) is much wider than that for nonstreaks (like HHHT).

To compare the distribution of two sequences of outcome HHHH and HHHT, across 1,000 simulated coin flips, see Figure 1. As both Hahn and Warren (2009) and Miller and Sanjurjo (2016) note, although the total number of occurrences of HHHH and HHHT is approximately equal, HHHH tend to cluster more than HHHT, with several overlapping occurrences together, and then large gaps between them. One way of explaining this it is that we know that overall frequency of HHHH and HHHT must be on average identical. However, immediately after flipping HHHH , there is a $50 \%$ chance of flipping another head, giving another instance of HHHH , and then a $50 \%$ chance of another, and so on. This leads to clusters of consecutive overlapping instances of HH H . Conversely, after flipping HHHT, it takes a minimum of four more flips to get HHHT again. This means that HHHT cannot cluster in the same way.

The consequence is that for shorter sequences of, say, 100 random binary outcomes, the frequency of HHHT will be fairly consistent, whereas the frequency of HHHH will be much more variable. This can be seen in Figure 2, in a simulation of 10,000 runs of 100 coin flips. Here, the string HHHT appears between 3 and 9 times on $95 \%$ of runs of 100 flips. HHHH only appears between 3 and 9 times on $67 \%$ of runs.

Hahn and Warren's theory explains the fact that people seem to think HHHH is less likely to occur than HHHT, by
looking at the difference in the probability of nonoccurrence of a string (or conversely the probability of its occurring at least once). Although they use shorter runs for their examples, the same pattern is observed: In Figure 2, the string HHHH is much more likely not to occur than HHHT is. In fact, although it is hard to see from the graph, the probability of HHHH's non-occurrence is around 100 times that of HHHT. This is - of course - a consequence of the fact that the mean of the frequency-of-occurrence distribution for HHHH is the same as that for HHHT, but the variance is much greater. Hahn and Warren suggest that when making judgments, people, whose experience is limited to short runs of outcomes, might attend to whether a string occurs or not, but not attend to the number of times it occurred. This means that they will see HHHT occurring in a lot more runs than they will HHHH , and will rate it more probable.

Conventional analysis of Gilovich et al.'s hot hand data used the logic that in the absence of a hot hand effect, the average proportion of players' shooting successes would be the same following three successes as following three failures. Miller and Sanjurjo note that this is not the case. The observation I make here is that this is a direct consequence of the distribution of frequency-of-occurrence in 100 binary outcomes being much wider for streaks than non-streaks is that the proportion of XXXX from \{XXXX, $\mathrm{XXXO}\}$ (or, by symmetry $\{\mathrm{XXXX}, \mathrm{OOOX}\}$ ) is less than . 5 .

To give a concrete example, if every day your grocer randomly gives you either 3,4 or 5 apples, and either 2,3 , 4 , 5 , or 6 oranges, and each day you work out what proportion of the fruit you were given is apples, you will find that, averaging across many days, the proportion of apples is greater than .5 , even though the total number of apples and oranges you receive is on average identical.
(If this is not obvious, consider the case where the grocer always gives you 4 apples, and also randomly gives you either 0 or 8 oranges. Half the time you leave with a bag that contains $100 \%$ apples; half the time you leave with a bag that contains $33 \%$ apples, so overall, the proportion of fruit in your bag that is apples averages $67 \%$. However, the overall number of apples and oranges you receive will be the same.)

Thus, in general, if one draws a sample from two distributions which have the same mean but different variances, and then looks at the proportion of the combined outcome that comes from each distribution, the expected


Figure 2: Distribution of frequency-of-occurrence for two different strings of outcomes in 10,000 simulated sequences of 100 coin flips
proportion from the lower variance distribution will be greater than that for the higher variance distribution.
This phenomenon can be seen more generally in Figure 3, which takes a normal approximation of the frequency-ofoccurrence distributions shown in Figure 2, with equal


Figure 3: Simulated samples drawn from distributions with a common mean, Color indicates the mean proportion of the sum of the two samples that comes from the y sample
means, capping at 0 , and varying the SD of the two frequencies-of-occurrence. The color indicates the mean proportion of outcome y , averaged across 50,000 simulated trials of each of 100 random binary outcomes. Where SDs are equal, then of course $p(y)=p(x)=.5$. Where $S D(y)>$ $\mathrm{SD}(\mathrm{x}), \mathrm{p}(\mathrm{y})<\mathrm{p}(\mathrm{x})$, and vice versa. A circle indicates the approximate point where $\mathrm{SD}(\mathrm{x})=\mathrm{SD}(\mathrm{HHHH})$ and $\mathrm{SD}(\mathrm{y})=$ SD(HHHT).

Replacing HHHH and HHHT with XXXX and OOOX, it is clear that, as Miller and Sanjurjo (2016) note, it is not correct to assume that, in the absence of a hot hand effect, the expected proportion of successes following $k$ successes, averaged across a large set of players, should be .5 . Rather, it is significantly lower, as a direct consequence of the distribution of frequency-of-occurrence for XXXX being broader than that for OOOX.

## Conclusion

The argument presented here is that both Hahn and Warren (2009) and Miller and Sanjurjo's (2016) findings can be explained the same way: In sequences of random binary outcomes, streaks of the same outcome (whether heads, HHHH, or successes, XXXX) cluster more than nonstreaks (HHHT, OOOX); this leads to a broader distribution of frequency-of-occurrence of streaks in finite sequences of random binary events relative to non-streaks. This both increases the chance of the non-occurrence of a streak (which H\&W argue makes people think justifiably that HHHH is less likely to occur than HHHT and other nonstreaks) and reduces the average proportion of XXXX among observations of $\{\mathrm{XXXX}$ and OOOX\} (which Miller and Sanjurjo convincingly argue means that evidence for a hot hand effect has been overlooked).

There are potentially interesting implications from these observations for the kinds of cognitive representation that would mediate the biases seen here. For example, an agent that counted the total number of occurrences of different strings of outcomes would see that the number of occurrences of, say, HHHH and HHHT were identical, so should rate them as equally probable. An agent that discarded all information about the frequency of occurrence of a string and recorded only whether or not it was observed (at least once) in a particular set of connected outcomes would of course perceive HHHT as more common than HHHH. Similarly, an agent that, rather than counting the number of occurrences of a string, instead encoded only the relative frequency of different strings, as a proportion of the total number of observations across occasions, would also conclude that HHHT was more frequently observed than НННH.
(Of course the overlapping of streaks described above may account for the biases seen here in more superficial ways. Chater (2014) argues that cognitive segmentation processes may differentially mask the frequency of occurrence of different strings. For example, a sequence of TTHHHHHHTT might be parsimoniously chunked as two tails - six heads - two tails, underplaying the three overlapping occurrences of HHHH within the sequence.)

Overall it seems clear that an examination of the distribution of frequency-of-occurrence for different strings of binary outcomes, allows one to create a parsimonious and intuitive account for two important recent theoretical observations, both of which have implications for the study of rationality.

## Acknowledgments

Thanks to Peter Ayton, Adam Harris, and Mike Le Pelley, for helpful discussions on this topic.

## References

Ayton, P., \& Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: two faces of subjective randomness? Memory \& Cognition, 32, 1369-78.
Bar-Hillel, M., \& Wagenaar, W. A. (1991). The perception of randomness. Advances in Applied Mathematics, 12, 428-454.
Chater, N. (2014). Cognitive science as an interface between rational and mechanistic explanation. Topics in Cognitive Science, 2, 331-337.
Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. Psychological Review, 103, 592-596.
Gilovich, T., Vallone, R., \& Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17, 295-314.
Falk, R., \& Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. Psychological Review, 104, 301-318.

Hahn, U., \& Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. Psychological Review, 116, 454-461.
Kahneman, D., \& Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.
Miller, J. B. \& Sanjurjo, A., (2016) Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers. IGIER Working Paper No. 552. Retrieved February 2, 2017 from: https://ssrn.com/abstract=262735. doi: 10.2139/ssrn. 2627354
Nickerson, R. S. (2002). The production and perception of randomness. Psychological Review, 109, 330-357. doi:10.1037//0033-295X.109.2.330
Nickerson, R. S. (2004). Cognition and chance: The psychology of probabilistic reasoning. Mahwah, NJ: Erlbaum
Nickerson, R. S. (2007). Penney Ante: Counterintuitive probabilities in coin tossing. The UMAP Journal, 28, 503-532
Rapoport, A., \& Budescu, D. V. (1997). Randomization in individual choice behavior. Psychological Review, 104, 603-617.
Reimers, S., Donkin, C., \& Le Pelley, M. E., (2017). Perceptions of randomness in binary sequences: Normative, heuristic, or both? Manuscript in preparation.
Rinott, Y., \& Bar-Hillel, M. (2015). Comments on a 'hot hand' paper by Miller and Sanjurjo (2015). $\mathrm{http}: / /$ ssrn.com/abstract=2642450. doi:10.2139/ssrn. 2642450
Sun, Y. \& Wang, H. (2010) Gamblers fallacy, hot hand belief, and the time of patterns. Judgement and Decision Making, 5, 124-132.

# Using punctuation as a marker of sincerity and affective convergence during texting 

Katherine Reynolds (kreyno20@kent.edu)<br>Speech Pathology \& Audiology, Kent State University<br>Kent, OH 44223 USA<br>Blair Casarotto (casarotto@wisc.edu)<br>Communication Sciences and Disorders, University of Wisconsin-Madison<br>Madison, WI 53706 USA<br>Sarah Noviski (snovisk@bgsu.edu)<br>Communication Sciences and Disorders, Bowling Green State University<br>Bowling Green, OH 43403 USA<br>Jennifer M. Roche (jroche3@kent.edu)<br>Speech Pathology \& Audiology, Kent State University<br>Kent, OH 44223 USA


#### Abstract

Face-to-face communication is a rich, natural form of communication that incorporates multimodal behavioral cues belying meaning and intention. However, computer-mediated communication (e.g., texting) removes many of the multimodal cues in face-to-face communication (e.g., vocal prosody). Recent research has suggested that punctuation might mimic vocal prosody in text (Gunraj et al., 2016), but there is no clear indication of what the overall effects may be. Therefore, the current study investigates the use of punctuation to express intonation. We first replicate Gunraj and his colleagues by showing that a single word ending in a period promotes the appraisal of negative affect. Interestingly, we extend this research by demonstrating that intonational punctuation has the potential to increase social distance, which our preliminary results suggest may occur through processes of emotional contagion and interactive alignment.


Keywords: pragmatics; texting; emotion contagion; interactive communication

## Introduction

With an increased reliance on technology during communication, mistaking the tone of a message may be common. Though mistakes are possible, numerous interesting studies suggest that some aspects of text may help interlocutors interpret conversational (i.e., affective) tone (e.g., lexical choices, punctuation, character features, emoticons; Byron \& Baldridge, 2007; Gunraj et al., 2016; Niederhoffer \& Pennebaker, 2002; Riordin \& Kreuz, 2010).

While these studies have touched on many features of text, less is known regarding the impact of pragmatic tone as indicated by punctuation (i.e., intonational punctuation), on lexical style matching during texting conversations. In the current study, we evaluate the effect of a sender's punctuation use as a paralinguistic cue to emotional tone, and its effect on a receiver's tendency to align their tone and texting preferences with the sender.

## Conveying Tone of Voice in Text

Face-to-face (FFC) communication benefits from the use of multimodal nonverbal cues like eye gaze, vocal prosody, and shared attention (e.g., Banziger \& Scherer, 2005; Burgoon et al., 1995; Knapp et al., 2013). One common area that most strongly affects text-based communication (TBC) is its unimodality. This unimodality is typically confined to linguistic, typographical, and/or grammatical cues, making TBC less rich in non-verbal cues found in FFC (e.g., Byron \& Baldridge, 2007; Kruger et al., 2005). However, a large body of research demonstrates that language users adapt to and make use of the communication medium to convey richer information about their message. For example, interlocutors make use of vocal spelling ("yeaaaaaaah"), non-standard spelling ("ermahgerd"), emoticons, and manipulated grammatical markers ("..."); all of these typographic cues have the potential to indicate tone of voice (Harris \& Paradice, 2007; Riordan \& Kreuz, 2010).

More recently, Gunraj et al. (2016) found that the use of a one-word response followed by a single period in TBC (relative to written communication) was perceived by participants to be rude and insincere. In fact, the use of typographical cues is adaptive, as it may aid in the decoding of an ambiguous message (e.g., when information is missing; Byron \& Baldridge, 2005; Derks et al., 2008; Harris \& Paradice, 2007; Lo, 2008; Riordan et al., 2014; Walther \& D'Addario, 2001). Though conventional multimodal nonverbal cues are absent from TBC, interlocutors have the potential to interpret a sender's intentions beyond the literal meaning.

## Faster but Ambiguous Messages

In a texting context, language users are constrained by the texting medium, in which their messages are typically fast, agrammatical, and largely ambiguous (for review see

Ling \& Baron, 2007). When an interlocutor lacks the relevant cues, one might rely more heavily on their own representations of the world (i.e., egocentric perspective). When this information is decoded incorrectly or the message is overly ambiguous, miscommunication has the potential to follow shortly behind.

Ambiguity has the potential to make receivers (listeners) of a message uncomfortable with uncertainty. However, listeners may handle this uncertainty by making predictions about a speaker's (sender's) intentions (Uncertainty Reduction Theory; Berger \& Bradac, 1992). Interlocutors essentially make predictions to decode ambiguous information by relying on nonverbal cues (e.g., paralinguistic cues) presented by a communication partner (Byron \& Baldridge, 2007). From this account, the receiver evaluates typographic cues (e.g., capitalization, emoticons, etc.) in the message to determine the sender's intentions.

Subsequently, the receiver attempts to synthesize known personality traits with the typographic cues to derive the correct intention. For example, when vocal tone is absent from the sender, a receiver reading a text message may construct an egocentric "tone" representation, similar to that of representing verbally produced "tone" in a face-to-face conversation -- as a means to reduce ambiguity, promote successful communication and decrease social distance (Byrne, 1971; Chartrand \& Bargh, 1999; Dennis \& Valacich, 1999). This leads one to question how a receiver's representation of tone is integrated in a deprived texting environment. More specifically, do interlocutors use similar communicative mechanisms to decrease social distance, when the cards are stacked against them?

## Convergence in Interpersonal Communication

A large amount of research has shown that lexical style matching during communication has the potential to decrease social distance (Byrne, 1971; Bernieri \& Rosenthal, 1991; Chartrand \& Bargh, 1999; Niederhoffer \& Pennebaker, 2002). However, recent research has shown that alignment need not always occur to promote positive communication outcomes (Fusaroli et al, 2012). In fact, these researchers show that though interlocutors have a proclivity to converge linguistically, dyads that adapted and diverged linguistically improved performance on a task i.e., more is not always better (Giles \& Coupland, 1991). It is possible, that under emotional contexts, one would be more likely to converge linguistically and para-linguistically during positive interactions and strategically diverge in negative interactions to decrease social distance (Paxton \& Dale, 2013). Currently, the utility of behavioral convergence and divergence during communication as a potential means to promote pragmatic interpretation, has been less well established in an emotional texting context.

When texting, the receiver of the message automatically colors the message with their interpretation of the sender's tone. If the sender produces emotionally valenced non-verbal cues, then the receiver might produce a similarly valenced response. Hatfield et al. (1994), for example, showed that interlocutors emotions tended to align when interacting with a social other. Additionally, it has also been argued that interlocutors prime each other (at various levels of the interaction), to promote effort saving communication behaviors (phonological to pragmatic priming: Dijksterhuis \& Bargh, 2001; Garrod \& Pickering, 2004). But what if the very nature of the intonational punctuation cue intends to increase social distance? Does the receiver of such a message diverge emotionally or with intonational punctuation? Or will the receiver follow suit and respond reciprocally (e.g., Theory of Reciprocity; McCroskey \& Richmond, 2000), further breaking down the conversation and increasing social distance?

The purpose of the current study is to determine the effect of intonational punctuation on a texting conversation. There are three main goals: (1) As a replication of previous studies, do interlocutors process punctuation for affective intent?, (2) Given intonational punctuation matching, will intonational convergence occur equally across positive and negatively valenced communication? and (3) What are the effects of punctuation alignment on the perception of conversational sincerity? We expect to see intonational punctuation impacting the perceived sincerity, and thus impacting the reciprocal use of punctuation to match conversational tone. We evaluate convergence through a distributional analysis, rather than time series analysis (Paxton et al., 2016). This allowed for the preliminary establishment of alignment, at the global level.

## Method

## Participants

Twenty Kent State University undergraduate students (females $=11$, mean ${ }_{\text {age }}=19.5 \mathrm{yrs}, \min _{\text {age }}=19 \mathrm{yrs}$, $\max _{\text {age }}=$ $23 \mathrm{yrs}, \mathrm{sd}=1.30$ ) participated in return for a $\$ 5$ gift card. One participant was excluded, because the program crashed half-way through the experiment. All participants were native speakers of English, with normal to normal-corrected vision and no diagnosed speech or hearing impairments.

## Materials

All stimulus presentations and data collection were controlled by a Matlab, Psychtoolbox-3 program, that recorded computer-mouse clicks. All participants were seated in front of a 21 -inch iMac computer screen in a sound-attenuated booth.

## Stimuli

The pseudo-confederate (fake texting partner) texting conversation was a neutral conversation about living in Kent, OH. Thirty-two pseudo-confederate texting response bubbles were created and paired with 5 alternative forced choice participant response bubbles. The participant response bubbles ranged from positive to neutral to negative (similar to a Likert scale), with the top most response option representing positive, the middle being neutral, and the bottom most option being most negative (e.g., Fig. 1). More specifically, the participant response bubbles contained one of the punctuation types, with the exclamation being most positive, no punctuation, ellipsis, and question mark being neutral, and the period being most negative (see Fig. 1).

These five types of punctuation were also used in the pseudo-confederate text messages: question mark, ellipsis, period, no punctuation, and exclamation. No punctuation, question marks, and ellipses acted as fillers and reflected typical uses of typographical markers in texting behavior (two-thirds of the trials; see Riordin \& Kreuz, 2010). The rest of the trials were primarily made up of exclamations (positive condition) or periods (negative condition).

## Design \& Procedures

At the beginning of the task, participants were told they would be having a texting conversation with another person, but instead of being able to freely type their responses, they would be given five response options to choose from (see Fig. 1). The participant was also asked to imagine that the conversation he or she was having was with a person he or she knew well, to promote ecological validity.


Figure 1: This is an example of a negative (left) and positive (right) pseudo-confederate text bubble, with the participant 5 alternative forced-choice response bubbles.

During the course of the interaction, the participant was presented with an image of an iPhone (Apple, Inc.), with a pseudo-confederate text bubble. They were then asked to read the text message, then to click the "next" button to transition into the participant response screen. On this
screen, they were able to see the pseudo-confederate text message bubble, with the 5 forced choice response options to the right of the iPhone (Fig. 1). After the response was made, the participant was brought to another screen that displayed a continuous rating scale, that required the participant to rate his or her perception of the conversation's sincerity (Fig. 2). Participants were randomly assigned to one of two between subjects conditions: Pseudo-confederate Valence Condition (positive: $\mathrm{n}=9$ or negative: $\mathrm{n}=10$ ).


Figure 2: The continuous rating scale participants used to rate the sincerity of the conversation, after each trial ("Please rate how sincerity of the conversation.")

## Measures

Participant Valence Selection Participants had five responses to choose from, that were ordered from positive (exclamation), to neutral (no punctuation, ellipsis, question mark), to negative (period). The response option chosen corresponded with an ordinal scale (exclamation (1) = most positive to period (5) $=$ most negative). Though the difference was not large, the participants in the positive condition chose response options on the more positive side of the Likert scale (mean: 1.99), relative to the negative conversation (mean: 2.23; $\mathfrak{t}=2.212, p<.05$ - independent samples t -test), with a larger number indicating a more negative valence. In order to reflect a more continuous measure of emotional contagion, a running average of the Likert scale ratings was calculated to reflect fluctuations in valence over the course of the interaction.

Intonational Punctuation Alignment Each pseudoconfederate response had one form of punctuation -- with each participant response bubble containing one of the five punctuation types. Depending on the response option chosen, the response was recoded as matching or mismatching the pseudo-confederate's use of punctuation. That is, if the pseudo-confederate used a period, what was the likelihood of the participant selecting a response option containing a period, too? Participants in the positive condition matched the pseudo-confederates choice of punctuation approximately $48 \%$ of the time, but only $28 \%$ of the time in the negative condition $(\mathrm{t}=2.913, p<.001-$ independent samples $t$-test).

Conversation Sincerity Participants were also asked to rate the sincerity of the conversation. This was done by clicking along a continuous rating scale (see Fig 2), and was measured based on pixels. The pixel rating scale ranged
from approximately $400-1200 \mathrm{px}$, with smaller numbers being related to more sincerity, and higher numbers related to lack of sincerity. Overall, positive conversations were rated as significantly more sincere (mean: 613.22px) than negative conversations (mean: 778.13px; $\mathrm{t}=3.848, p<$ .001- independent samples t-test).

## Results

Outcomes are reported from linear mixed-effects models built using lme4 package in R (R Core Development Team, 2008). A mixed effects regression was used to predict receiver perceived sincerity by Pseudo-confederate Valence Condition (positive or negative), Participant Valence Selection (Likert-like response options), and punctuation alignment (match or mismatch). The models implemented maximal random effect structures to achieve model convergence, with participants and trial set as random intercepts. All categorical variables were dummy coded.

Pseudo-Confederate Valence (Manipulation Check) As a manipulation check, to determine if the conversations would be perceived as positive or negative by our participants, the pseudo-confederate responses were presented to four female participants (mean age $=20.5$ ) who did not participate in the pseudo-texting conversation. These participants were asked to rate each pseudo-confederate response bubble on a continuous rating scale for positive/negative valence. The rating scale spanned 400-1200 pixels, with lower numbers representing more positive valence, and higher numbers representing negative valence. The positive pseudo-confederate messages received an average pixel rating of 642.11 px , while the negative pseudo-confederate messages were rated at approximately 850.77 px ( $\beta=$ -208.66, $\mathrm{se}=28.03, \mathrm{t}=-7.37, p<.001$; similar to Fig 2.).

Intonational Punctuation Replication To determine if pseudo-confederate punctuation was interpreted as affective in nature (question 1), conversational sincerity was assessed between the conversational valence conditions (dummy codes: positive $=0$, negative $=1$ ). Results revealed that the negative condition was rated as significantly more insincere than the positive condition $(\beta=208.66$, $\mathrm{se}=25.53, \mathrm{t}=8.17$, $p<.001$ ), replicating Gunraj et al. (2016).

Intonational Punctuation Style Matching To determine if participants aligned with the pseudo-confederate's use of punctuation (question 2), intonational punctuation was evaluated between the two conversational conditions (dummy codes: positive $=0$, negative $=1$ ). Results from this mixed effects model revealed that the negative condition was rated as significantly more insincere than the positive condition ( $B=-1.05$, se $=0.20, \mathrm{t}=-5.22, p<.001$; logistic model, family set as binomial; Jaeger, 2008). This suggests
that interlocutors might be more likely to align under positive, than negative contexts.

Conversational Appraisal In the above analyses, we show that intonational punctuation and style matching occurs differently depending on the conversational valence. The last analyses (question 3) intended to determine whether or not the appraisal of the interaction (sincerity) should be affected by the pseudo-confederate's tone, similarity (to the pseudo-confederate) of their intonational punctuation, and the valence of the response selection.

The results indicated a number of main effects and an interaction. Specifically, there was a main effect of Pseudo-confederate Valence Condition, indicating that participants in the positive condition rated the conversation as significantly more sincere than participants in the negative condition ( $\beta=124.675$, $\mathrm{se}=52.747, \mathrm{t}=2.364, p<$ .05). This suggests that participants were sensitive to pseudo-confederate affective tone.


Figure 3: This figure represents the relationship between punctuation alignment and sincerity, as a function of pseudo-confederate valence.

There was also main effect of Participant Valence Selection (i.e., Likert-like responses; $\beta=97.348$, se $=$ 43.442, $\mathrm{t}=2.241, p<.05$ ) but no interaction between Pseudo-confederate Valence Condition x Participant Valence Selection. This suggests that as participants rated the interaction as less sincere, they chose more negative response options (e.g, emotional contagion).

Lastly, though there was no main effect of intonational punctuation style matching ( $p=.352$ ), there was a significant Pseudo-confederate Valence Condition $x$ punctuation alignment interaction $(\beta=49.841$, $\mathrm{se}=18.424$, $\mathrm{t}=2.705, \mathrm{p}<.01$; see Fig. 3). This indicated that as participants assessed the sincerity of each turn, the more sincere the turn seemed, the more likely the participant would match the pseudo-confederate's punctuation.

Alternatively, the more insincere the conversation, participants were less likely to align their punctuation, with the caveat that highly insincere ratings increased negative punctuation alignment.

## Discussion

Texting provides a wealth of communicative benefits. However, texting necessarily requires a fast, reduced, and often ambiguous delivery of information (Ling \& Baron, 2007). In the face of ambiguity, the receiver of these messages must use their own representations of the world to interpret tone correctly, especially when the tone is not as explicit as vocal tone because of its dual meaning (i.e., grammatical and adapted pragmatic cues; Byron \& Baldridge, 2005; Derks et al., 2008; Harris \& Paradice, 2007; Lo, 2008; Riordan et al., 2014; Walther \& D'Addario, 2001). Therefore, it is not always easy for a receiver of an emotionally valenced text message to correctly interpret tone. In the current study, we specifically evaluate the role of positive and negatively valenced punctuation and its effect on a receiver's ability to interpret tone.

Consistent with Gunraj et al. (2016), the results indicate that participants are in fact sensitive to punctuation as an effective cue to conversational tone. Additionally, the participants perception of the pseudo-confederate's sincerity shaped the participants' responses -- with a more insincere pseudo-confederate receiving responses that were more negatively valenced, though dampened. And finally, we see the convergence and divergence of punctuation use by the participant for pragmatic effect.

Though the current results provide interesting insight into the pragmatic nature of intonational punctuation, the current study is not without limitation. The main limitation of the current study is the lack of ecological validity, which may have affected patterns of responding. For example, participants in the positive condition may have attempted to decrease social distance by using the same communicative mechanisms engaged in FFC (behavioral entrainment; Byrne, 1971; Bernieri \& Rosenthal, 1991; Chartrand \& Bargh, 1999; Niederhoffer \& Pennebaker, 2002). Specifically, participants in the positive condition exhibited emotional contagion and intonational punctuation alignment, which in other domains has been suggested to promote interpersonal liking and communicative smoothness (Chartrand \& Bargh, 1999). Even though the participants were instructed to imagine they were texting with someone they knew well, our participants may have aligned less and been unsure of how to interpret the one word responses ending with a period (negative), because they lacked relevant history with their texting partner (Bernieri \& Rosenthal, 1991).

Alternatively, participants may have defaulted to rules of social engagement, in that it is typically socially
unacceptable to have contentious interactions with strangers (Morand et al., 2003). Therefore, participants may have been more likely to disengage synchronous mechanisms to defuse the negative interaction. However, the more contentious (i.e., more insincere the conversation seemed) the more likely the participant converged their text response with the pseudo-confederate. One should approach this interpretation with caution, because it is difficult to claim intentionality because the interaction was with an assumed stranger, and behavioral frequencies were assessed, and not via time course analysis. These results, nonetheless, are consistent with negative FFC interactions, in which decreases in convergence have been found (Abney, Paxton, Kello, \& Dale, 2014; Paxton \& Dale, 2013).

Finally, we did not look specifically at the time course of synchrony across the interaction. This was mostly due to the low sample size and the preliminary nature of the current study. We first wanted to show that emotional valence has the potential to differentially impact communication in a texting context. Punctuation matters just as much as vocal prosody, because of the pragmatic implications of the cue. Therefore, the next step of this project will be to expand this paradigm, by collecting more data so we might be able to explore the temporal dynamics of emotional valence during text based communication.

## Conclusions

We are relying more and more on digital forms of communication, with even the most prominent political leaders communicating through short, fast text-based responses via social media. In the current study, we provide preliminary insight into the cognitive mechanisms (e.g., emotional contagion) that drive the interpretation of intentionality. Texters (college-aged) not only use typographic variation to indicate pragmatic meaning, but also use it to infer intentions. Additionally, the choice to use certain typographical cues may push the valence of a conversation in a more positive or negative direction. Therefore, one should be aware that one's use of punctuation has pragmatic implications over and above grammatical form, in texting. We see that texting follows similar rules as FFC. In that, language is naturally ambiguous, but in a texting context ambiguity is a critical feature of communicating. Additionally, the texter will follow the valence of their communicative partner as a means to increase and decrease social distance. Therefore, failure to use the appropriate (texting) affective cues could lead to higher rates of miscommunication, misunderstandings, and generally hard feelings.

## Acknowledgments

A special thanks goes to the MADI lab research assistants that worked so hard to collect data.

## References

Abney, D. H., Paxton, A., Dale, R., \& Kello, C. T. (2014). Complexity matching in dyadic conversation. Journal of Experimental Psychology: General, 143(6), 2304.
Bänziger, T., \& Scherer, K. R. (2005). The role of intonation in emotional expressions. Speech communication, 46(3), 252-267.
Berger, C. R., \& Bradac, J. J. (1982). Language and social knowledge: Uncertainty in interpersonal relations (Vol. 2). Hodder Education.

Bernieri, F. J., \& Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony.
Burgoon, J. K., Le Poire, B. A., \& Rosenthal, R. (1995). Effects of pre interaction expectancies and target communication on perceiver reciprocity and compensation in dyadic interaction. Journal of experimental social psychology, 31(4), 287-321.
Byrne, D. E. (1971). The attraction paradigm. Academic Pr.
Byron, K., \& Baldridge, D. C. (2005). Toward a model of nonverbal cues and emotion in email. In Academy of Management Proceedings, 2005(1), B1-B6.
Chartrand, T. L., \& Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. Journal of personality and social psychology, 76(6), 893.
Dennis, A. R., \& Valacich, J. S. (1999, January). Rethinking media richness: Towards a theory of media synchronicity. In Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on (pp. 10-pp). IEEE.
Derks, D., Fischer, A. H., \& Bos, A. E. (2008). The role of emotion in computer-mediated communication: A review. Computers in Human Behavior, 24(3), 766-785.
Dijksterhuis, A., \& Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. Advances in experimental social psychology, 33, 1-40.
Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., \& Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. Psychological science, 0956797612436816.
Garrod, S., \& Pickering, M. J. (2004). Why is conversation so easy?. Trends in cognitive sciences, 8(1), 8-11.
Giles, H., Coupland, J., \& Coupland, N. (1991). Contexts of accommodation: Developments in applied sociolinguistics. Cambridge University Press.
Gunraj, D. N., Drumm-Hewitt, A. M., Dashow, E. M., Upadhyay, S. S. N., \& Klin, C. M. (2016). Texting insincerely: The role of the period in text messaging. Computers in Human Behavior, 55, 1067-1075.

Harris, R. B., \& Paradice, D. (2007). An investigation of the computer-mediated communication of emotions. Journal of Applied Sciences Research, 3(12), 2081-2090.
Hatfield, E., Cacioppo, J. T., \& Rapson, R. L. (1994). Emotional contagion. Cambridge university press.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434-446.
Knapp, M. L., Hall, J. A., \& Horgan, T. G. (2013). Nonverbal communication in human interaction. Cengage Learning.
Kruger, J., Epley, N., Parker, J., \& Ng, Z. W. (2005). Egocentrism over e-mail: can we communicate as well as we think?. Journal of personality and social psychology, 89(6), 925.
Ling, R., \& Baron, N. S. (2007). Text messaging and IM linguistic comparison of American college data. Journal of Language and Social Psychology, 26(3), 291-298.
Lo, S. (2008). The nonverbal communication functions of emoticons in computer mediated communication. CyberPsychology \& Behavior, 11, 595-597.
McCroskey, J. \& Richmond, V. (2000). Applying reciprocity and accommodation theories to supervisor/subordinate communication. Journal of Applied Communication Research, 28(3), 278-289.
Morand, D. A., \& Ocker, R. J. (2003, January). Politeness theory and computer-mediated communication: A sociolinguistic approach to analyzing relational messages. In System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on (pp. 10-pp). IEEE.
Niederhoffer, K. G., \& Pennebaker, J. W. (2002). Linguistic style matching in social interaction. Journal of Language and Social Psychology, 21(4), 337-360.
Paxton, A., \& Dale, R. (2013). Argument disrupts interpersonal synchrony. The Quarterly Journal of Experimental Psychology, 66(11), 2092-2102.
Paxton, A., Dale, R., \& Richardson, D. C. (2016). 18 Social coordination of verbal and nonverbal behaviours. Interpersonal Coordination and Performance in Social Systems, 259.
Riordan, M. A., \& Kreuz, R. J. (2010). Emotion encoding and interpretation in computer-mediated communication: Reasons for use. Computers in human behavior, 26(6), 1667-1673.
Riordan, M. A., Kreuz, R. J., \& Olney, A. M. (2014). Alignment is a function of conversational dynamics. Journal of Language and Social Psychology, 33(5), 465-481.
Walther, J. B., \& D'Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. Social science computer review, 19(3), 324-34.

# Influencing Network Graph Perception and Judgment: Effects of Direct Connections, Base Rates, and Visual Layout Proximity on Social Network Analysis 

Rebecca E. Rhodes ${ }^{1}$ (rebecca.rhodes@jhuapl.edu), Isaiah Harbison ${ }^{1}$ (james.harbison@jhuapl.edu), Nathan Bos ${ }^{1}$ (nathan.bos@jhuapl.edu), Celeste Lyn Paul ${ }^{2}$ (clpaul@tycho.ncsc.mil), Clay Fink ${ }^{1}$ (clay.fink@jhuapl.edu), Anthony Johnson ${ }^{1}$ (anthony.johnson@jhuapl.edu)<br>1. Johns Hopkins University Applied Physics Laboratory, Laurel, MD<br>2. U.S. Department of Defense, USA


#### Abstract

Social network graphs are often used to help inform judgments in a variety of domains, such as public health, law enforcement, and political science. Across two studies, we examined how graph features influenced probabilistic judgments in graph-based social network analysis and identified multiple heuristics that participants used to inform these judgments. Study 1 demonstrated that participants' judgments were influenced by information about direct connections, base rates, and layout proximity, and participants' self-reported strategies also reflected use of this information. Study 2 replicated findings from Study 1 and provided additional insight into the hierarchical ordering of these strategies and the decision process underlying judgments from social network graphs.


Keywords: social network analysis; graph comprehension; data visualization; judgment and decision making

## Introduction

## Social Network Analysis

Social network analysis is an analytical method for understanding data that depicts relationships between entities, such as communications between people or the flow of information through a community. Social network analysis (SNA) can be used to inform judgments such as who the most connected or influential person is in a community. The applications for SNA are widespread. The CDC has used social networks to map the spread of infectious diseases (Cook, 2007), and intelligence professionals have used SNA to map potential terrorist suspects. For example, Krebs (2002) was able to map a network of suspected $9 / 11$ terrorists after the attacks based on publicly available information in the news and provide insight into the terrorist organization.

Social networks are typically depicted using node-link diagrams, in which nodes (depicted with circles) represent objects, such as people, and edges between the nodes (depicted with lines) represent a connection, such as a friendship or a recent communication. Although there are quantitative measures of graph structures, such as network centrality, SNA often involves some degree of visual interpretation. In many applications of SNA, such as tactical decision making in military organizations, users of network graphs have limited time to make their decisions and may not have the background to supplement their visual
interpretation with more objective, quantitative measures. Thus, clearly communicating the important information in network graphs is essential. In practice, however, these network visualizations are often complex with hundreds or thousands of nodes, and relatively little is known about the cognitive strategies people use to make probabilistic judgments from these graphs. In the present study, we aimed to identify graph features and heuristics that influenced probabilistic judgments in scenarios that mirrored real-world SNA tasks.

## Graph Perception

Much research has applied perceptual and cognitive theories to improve graph comprehension and readability. Purchase, Cohen, and James (1997) first demonstrated that several aesthetic qualities could improve comprehension, such as maximizing the symmetry of the graph, minimizing the number of intersecting edges, and minimizing the number of bends in a series of segments. Purchase (2000) found that graph layouts that minimized intersecting edges significantly improved participants' abilities to answer questions about the relationships in the graph. Similarly, Ware, Purchase, Colpoys, and McGill (2002) found that cutting unnecessary edges significantly improved cognitive performance.

Other perceptual features in graphs have been studied as well. For example, McGrath, Blythe, \& Krackhardt (1997) found that nodes that were positioned closer to the center of the graph were perceived as more prominent, suggesting a bias towards layout centrality. Additionally, nodes were more likely to be considered to be in the same group if they were positioned closer to each other, and nodes positioned directly between two nodes (as opposed to at an angle) made them more likely to be considered a "bridge" between the two nodes.

Gestalt principles may help explain why some aesthetic features can influence graph interpretation by exploiting people's tendency to identify patterns (Novick \& Bassok, 2005). The Gestalt principle of good continuation suggests that paths are more easily recognized when they contain less jaggedness, suggesting that the fewer the bends between two nodes, the more readily the path connecting them will be observed. The principle of proximity suggests that spatial proximity between objects implies logical groupings. Other cognitive heuristics may also influence graph interpretation.

For example, people sometimes ignore base rates when making probabilistic judgments. The extent to which this occurs when making judgments from network graphs, which may depict base rate information in a visual way, is unclear (but see Micallef, Dragicevic, \& Fekete, 2012).

## Pilot Study

We conducted a pilot study with 30 participants recruited from Amazon Mechanical Turk (AMT) to identify graph features people attend to when making judgments about the likelihood that a particular node belonged to a defined category. We planned to use the results of the pilot study to inform our decision about graph features to manipulate in our experiments ${ }^{1}$. Participants were asked to play the role of a data analyst at a law enforcement agency and use nodelink graphs to estimate the likelihood that they would further investigate a person of interest (POI) in a community that has known drug users (Figure 1). They were also asked to describe the strategies they used to make their judgments. The POI was indicated as a black node, known drug users as blue nodes, and everyone else as yellow nodes. Lines were drawn between nodes to indicate recent communications. Two graph features were varied, 1) the number of direct connections the POI had, and 2) the number of drug users in the whole graph (base rate). The graph visual was generated using a force-directed layout, a popular way of visualizing graph data. Participant responses revealed two common strategies. The first was based on the number of drug users within the POI's direct connections (what we call the "ratio" strategy, $N=10$ ). The second strategy was based on the spatial proximity of drug user nodes to the POI node in the visual layout, regardless of whether they were actually closely connected to the POI ("proximity" strategy, $N=8$ ). We expected to see instances of the ratio strategy, but not necessarily the proximity strategy, since the position of nodes in a force-directed layout is a function of connections rather than nodes. In general, layout position in a graph can be misleading, and two nodes visually close to each other but not connected may not necessarily have a strong relationship. Three participants in the pilot study also mentioned relying on the total number of drug users in the graph. This strategy was expected to be more frequent considering that base rates were explicitly manipulated across graphs. However, participants apparently found base rates less informative than ratio and proximity features for the pilot set of graphs.

## Overview of Experiments

We conducted two experiments to understand how the graph features identified in the pilot study influence probabilistic judgments. In each study, we manipulated the three features identified during our pilot study: 1) ratio, or the number of connections the POI has to salient nodes, 2)

[^458]base rate of salient nodes, and 3) proximity of salient nodes to the POI.


Figure 1: Example network graph provided in pilot study.

## Study 1

## Participants and Procedure

Participants. We recruited 30 participants from AMT. Participants were required to have a "Masters" qualification, indicating that they had repeatedly demonstrated good performance in the AMT marketplace. Most participants reported having no experience ( $72.41 \%$ ) or only slight experience ( $17.24 \%$ ) with SNA. One participant was excluded from analysis for failing an attention check.

Scenario. Similar to the pilot study, participants were asked to take on the role of a data analyst at a law enforcement agency and use graphs to identify potential drug users in a community. They were provided a set of 36 graphs that described communications in a small community over the past month. Participants were told that links between nodes represented non-family relationships with others, such as friends and coworkers. No information was provided on how the graphs were constructed. For each graph, participants made a judgment of how likely they would be to investigate the POI and described how they used the graph to make their judgment.

Materials. The graphs were small real-world graphs generated from the Polbooks dataset of co-purchased political books (V. Krebs, unpublished, http://www.orgnet.com/). Graphs were rendered with the default graph plotting parameters in the R iGraph package, which uses the Fruchterman-Rheingold force-directed algorithm. Each graph consisted of 104 nodes. As in the pilot study, the POI was indicated as a black node, known drug users as blue nodes, and everyone else as yellow nodes. Direct connections to the POI were highlighted in the graphs to aid interpretation. Emphasizing direct connections in this way may have led to more emphasis on the ratio strategy, but highlighting connections was deemed important to help participants distinguish between graph connectedness from layout proximity (i.e. visual closeness).

Conditions. Three graph conditions were manipulated within-subjects: 1) number of drug users in the POI's N direct neighbors ( $0 / \mathrm{N}, 1 / \mathrm{N}$, Half/N); 2) overall base rate of known drug users in entire graph ( $5 \%$ or $50 \%$ of all nodes); and 3 ) proximity placement of known drug user nodes (near or far from POI). Note that proximity was the physical location of drug user nodes within the visual graph layout and not the graph connectedness. Graph connectedness was controlled by removing edges that connected proximal drug user nodes to the POI's direct neighbors (Figure 2).


Figure 2: Graph feature conditions. Note that in the Proximity: Near graph, the proximal blue nodes are not connected to the POI or the POI's direct neighbors.

Likelihood Judgments. For each graph, participants were asked to rate how likely they would be to investigate the POI on a scale of 1 to 7 ( $1=$ Extremely Unlikely, $7=$ Extremely Likely).

Strategy Use. After giving their likelihood rating, participants were also asked to describe any specific graph information they used in making their rating. Responses were open-ended text responses and were analyzed using a structured coding approach, described in Table 1, with three coders who coded $75 \%$ of the responses independently and $25 \%$ overlapping. Inter-rater reliability on the $25 \%$ co-coded responses was high (Cohen's Kappa > .75).

Table 1: Coding scheme for open-ended responses.

| Strategy | Description |
| :---: | :---: |
| Ratio | Mentions number of users in POI |
| connections |  |
| Base Rate | Mentions number of users in graph |
| Proximity | Mentions visual closeness of POI to users |

## Results and Discussion

Likelihood Judgments. We modeled likelihood judgments with a linear multilevel model with base rate, proximity, and ratio conditions treated as random effects. Graph feature conditions were nested within participant. Multilevel models are able to estimate the variance associated with each random effect; thus, the model can account for the within-individual variability for different graph feature conditions. In addition to modeling random effects, multilevel models can also simultaneously estimate fixed effects, which in this case represent the average effect for the whole sample.

We sequentially entered each graph feature into the model as a fixed effect to examine its effect on likelihood ratings. Base rate, ratio, and proximity each significantly improved the fit of the model $\left(\chi^{2}(1)=34.51, p<.001, \chi^{2}(2)=148.62\right.$, $p<.001$, and $\chi^{2}(1)=45.07, p<.001$, respectively). Specifically, higher base rates increased the likelihood of investigating the POI $(b=0.97, t(1011)=7.54, p<.001)$, higher ratios increased the likelihood of investigating ( $b=$ $1.24, t(1011)=7.18, p<.001$ for $1 / \mathrm{N}$ vs. $0 / \mathrm{N}$ and $b=3.86$, $t(1011)=22.34, p<.001$ for Half/N vs. $1 / \mathrm{N})$, and closer proximities also increased the likelihood of investigating ( $b$ $=1.02, t(1011)=9.58, p<.001)$. These results suggest that participants considered all three graph features when making their judgments. The observation that participants' suspicion levels of the POI increased with proximity to known drug users was significant, because it suggests that participants believed proximity to be a meaningful indicator.

We also found a significant interaction between base rate and proximity $\left(\chi^{2}(1)=146.00, p<.001\right)$. When proximity was far, there was no difference between the base rate conditions, but when proximity was near, higher base rates significantly increased the likelihood of investigating ( $b=$ $1.70, t(1006)=13.96, p<.001$; Figure 3). In other words, base rate information only affected judgments to the extent that it placed more users near the POI.

There were also interactions between the base rate and ratio conditions $\left(\chi^{2}(2)=112.90, p<.001\right)$ as well as between the ratio and proximity conditions $\left(\chi^{2}(2)=81.89, p\right.$ $<.001$ ), as depicted in Figure 3. Increasing the base rate had a stronger effect in the $0 / \mathrm{N}$ and $1 / \mathrm{N}$ conditions compared to the Half/ N condition $(b=1.70, t(1006)=11.40, p<.001$ and $b=0.94, t(1006)=6.32, p<.001$ for $0 / \mathrm{N}$ vs. Half/N and $1 / \mathrm{N}$ vs. Half/ N , respectively). Likewise, proximity had a stronger effect in the $0 / \mathrm{N}$ and $1 / \mathrm{N}$ conditions as well $(b=$ $1.36, t(1006)=9.09, p<.001$ and $b=0.87, t(1006)=5.85$, $p<.001$, respectively). These interactions reveal that, in the highest ratio condition, base rate and proximity information had little effect. One interpretation of this result is that when ratio was high, participants did not feel a need to look at other information in the graph. In contrast, when the ratio was $0 / \mathrm{N}$ or $1 / \mathrm{N}$, base rate and proximity information both influenced judgments, such that higher base rates and closer proximity increased likelihoods of investigating.

Strategy Use. Analysis of participants' self-reported strategies validated our findings from the pilot study. Participants mentioned using the ratio strategy for the majority of graphs (85.34\%). Proximity was the next most frequently mentioned strategy ( $22.41 \%$ ), followed by base rate a small percentage of the time (12.64\%).


Figure 3: Interactions between graph feature conditions.

## Study 2

Study 2 sought to replicate the findings of Study 1 with a broader range of stimuli. We introduced two additional scenarios, public opinion and disease propagation. Network graphs can be used to represent relational data from a variety of domains, and assumptions about the data may change interpretations of the graph. For example, base rates may be utilized more when the activities represented in the graphs are perceived to be more mobile or contagious. We speculated that a network graph representing the spread of an infectious disease could lead people to use base rate strategies more often than a network graph representing the spread of drug use. Study 2 also sought to better understand the hierarchical ordering of cognitive strategies identified in the previous studies by constructing a decision tree that could predict the conditions under which particular strategies will be used.

## Participants and Procedure

Participants. We recruited 196 participants from AMT. Participants were limited to U.S. residents and had at least a $90 \%$ approval rating for previous HITs. We excluded 11 participants who either failed attention checks or had participated in a previous study. Most participants said they had no prior experience ( $64.86 \%$ ) or only slight prior experience ( $28.65 \%$ ) with SNA.

Scenario and Materials. Participants were provided the same set of 36 graphs from Study 1 and were asked to make
likelihood judgments about the POI in each graph.
Conditions. The same three graph conditions from Study 1 were manipulated within-subjects. We also introduced different scenarios as a between-subjects manipulation for the source of the graph data. Participants were randomly assigned to one of three data scenarios: drug use, political opinion, and infectious disease (Table 2).

Table 2: Description of data scenarios.

| Scenario | Brief Description |
| :--- | :--- |
| Drug | In a community in which some percentage of <br> people are known users, how likely will the <br> POI become a drug user in the next six <br> months? |
| Opinion | In a community in which some percentage of <br> people are known proponents of a new <br> proposal, how likely will the POI become a <br> proponent in the next six months? |
| DiseaseIn a community in which some percentage of <br> people are infected with a new disease <br> transmitted through social contact, how likely <br> will the POI become infected in the next six <br> months? |  |

Strategy Use. In contrast to Study 1, in which participants described their analysis strategy in an open-text form, participants in Study 2 rated the strategies from Study 1 (ratio, base rate, and proximity) as well as distractors (e.g., central or peripheral location of POI, presence of a cluster of nodes) as to how important they were to their likelihood judgment. Ratings were on a scale of 1 (Not at all important) to 5 (Extremely important). The list order of strategies was randomized at each presentation.

## Results and Discussion

Likelihood Judgments. We used the same model from Study 1 but added scenario to the model as a fixed effect. There was a significant main effect of scenario on likelihood judgments ( $\chi^{2}(2)=26.40, p<.001$ ); specifically, participants in the disease condition gave higher likelihood ratings on average than participants in the drugs or opinion conditions $(b=0.58, t(182)=4.65, p<.001$ and $b=0.63, t(182)=$ $4.71, p<.001$ for disease vs. drugs and disease vs. opinion, respectively). As in Study 1, there were also main effects of base rate $\left(\chi^{2}(1)=288.97, p<.001\right)$, proximity $\left(\chi^{2}(1)=\right.$ 291.26, $p<.001$ ), and ratio $\left(\chi^{2}(2)=671.77, p<.001\right)$. As depicted in Figure 4, higher base rates led to higher likelihood ratings $(b=1.21, t(6471)=24.06, p<.001)$, higher ratios led to higher ratings $(b=0.62, t(6471)=8.92$, $p<.001$ for $1 / \mathrm{N}$ vs. $0 / \mathrm{N}$, and $b=2.16, t(6471)=31.39, p<$ .001 for Half/N vs. $1 / \mathrm{N}$ ), and closer proximities also led to higher ratings $(b=0.99, t(6471)=25.00, p<.001)$.

We again found two-way interactions between base rate and proximity $\left(\chi^{2}(1)=645.45, p<.001\right)$, base rate and ratio $\left(\chi^{2}(2)=424.43, p<.001\right)$, and proximity and ratio $\left(\chi^{2}(2)=\right.$
501.50, $p<.001$ ). Consistent with Study 1, increasing the base rate only mattered when proximity was high ( $b=1.36$, $t(6466)=28.31, p<.001)$; increasing the base rate had stronger effects when ratios were $0 / \mathrm{N}$ or $1 / \mathrm{N}$ compared to Half/N $(b=1.28, t(6466)=21.70, p<.001$ for $0 / \mathrm{N}$ vs. Half $/ \mathrm{N}$, and $b=0.79, t(6466)=13.43$ for $1 / \mathrm{N}$ vs. Half/N); and closer proximity also had stronger effects when ratios were $0 / \mathrm{N}$ or $1 / \mathrm{N}(b=1.30, t(6466)=22.16, p<.001$ for $0 / \mathrm{N}$ vs. Half $/ \mathrm{N}$, and $b=0.36, t(6466)=6.19, p<.001$ for $1 / \mathrm{N}$ vs. Half/N). In other words, when the ratio of users in the POI's connections was very high (Half/N), other graph features had less influence on judgments.


Figure 4: Main effects of data scenario and graph features.
As depicted in Figure 4, main effects of graph features were found for each scenario we tested. The main interaction we were interested in was between scenario and base rate, since we expected that base rates would be utilized more when the scenario described more contagious activities, such as the spread of disease. There was not a significant interaction between scenario and base rate, although significant interactions did emerge between scenario and the proximity and ratio graph features $\left(\chi^{2}(2)=\right.$ $6.31, p=.04$ and $\chi^{2}(4)=40.21, p<.001$, respectively $)$. However, Figure 4 suggests that differences were small and, overall, graph features affected judgments in a consistent way across scenarios.

Strategy Use. Consistent with Study 1, participants' ratings of strategy importance reflected a tendency to rely on the ratio strategy more than other strategies. A multilevel model with strategy treated as a random effect and nested within participant revealed that participants rated proximity and ratio information as more important to their judgment than base rate information $(b=0.24, t(368)=3.43, p<.001$ and $b=0.78, t(368)=11.32, p<.001$, respectively). Participants
rated ratio information as more important than proximity information, $b=0.54, t(368)=7.89, p<.001$.

Decision Tree. Given the multiple interactions between graph features, we modeled the influence of graph features on likelihood judgments with a decision tree using the rpart package in R. Human judgment is often based on 'fast and frugal' heuristics (Gigerenzer, \& Goldstein, 1996), which can be modeled by decision trees. Decision trees identify a series of binary decisions to maximize prediction accuracy of an outcome variable.

The model in Figure 5 shows the decision tree for the Study 2 data. The first decision point splits the data based on ratio. If the ratio was high (many salient nodes in the POI's direct connections), the model estimated that the likelihood rating was high (5.8), and no other variables were considered (far right side of Figure 5).


Figure 5: Decision tree predicting likelihood of investigating the POI. $n$ represents number of judgments collected from participants.

The left side of the first decision point indicates that when there were zero or one salient node(s) in the POI's direct connections, the model next split the data based on the base rate. If the base rate was low (left side of Figure 5), the model then made another split based on ratio, and proximity was not used. If the base rate was high, the model then used proximity. If proximity was far, the model used ratio information again to make the final classification.

## General Discussion

In this research we identified three graph features that influenced judgments about a specific person of interest (POI) in a social network graph: 1) number of salient nodes within the POI's direct connections, 2) base rate of salient nodes in the graph, and 3) proximity of salient nodes to the POI. Across two experiments, we demonstrated the influence of these features on probabilistic graph-based
judgments by manipulating their presence in a series of graphs. Additionally, through our analysis of participants' self-reported strategies and strategy importance ratings, we demonstrated that participants consciously use these graph features in their judgments. Participants reported primarily using the ratio strategy to make their judgments, followed by the proximity and base rate strategies. No matter which data scenario was used (e.g., drug use, disease, public opinion), base rate, proximity, and ratio manipulations influenced judgments in similar ways, with only slight differences across scenarios.

## Strategy Use

Participants consistently used the ratio strategy to make likelihood judgments. The three-level manipulation of this variable $(0 / \mathrm{N}, 1 / \mathrm{N}, \mathrm{Half} / \mathrm{N})$ was the strongest determinant of likelihood judgments and the ratio strategy was the most frequently self-reported strategy. Furthermore, a decision tree analysis suggests that when the number of connections was high, participants made 'fast and frugal' decisions without using the other strategies.

In other cases, participants made use of base rate and layout proximity. The decision tree analysis suggests that each of these graph features was considered by participants when making likelihood judgments, and manipulating each of these variables led to significant main effects on judgments. However, the relative importance of these two strategies is less clear. Participants were more likely to mention using the proximity strategy than the base rate strategy in Study 1, and in Study 2 participants rated the proximity strategy as more important than the base rate strategy. Yet the decision tree analysis revealed that base rate was a decision point before proximity was considered, suggesting that base rate may be more important than proximity to participants' judgments. These results suggest that there may be a disconnect between participants' selfreported strategies and the strategies revealed by their actual judgments. One interpretation of this finding could be that participants were mistaken in how much they considered each graph feature. Future research could further explore the relative importance of these strategies by testing different response formats, scenarios, and graphs.

## Is Use of Proximity an Error?

An important question about these results is whether the use of proximity should be considered a reasoning error. It is true that in force-directed layouts, which was the layout algorithm used to generate the graphs in these studies, layout distance does often have a relationship with graph distance. In other words, the physical distance between two nodes is somewhat related to the number of edges that separate those nodes. However, it could be misleading to rely only on layout proximity to make the kinds of judgments in these studies for two reasons. First, the extent to which proximity provides meaningful information about the relationship between two nodes depends critically on the layout algorithm used for graph construction. Second, in
many cases, users viewing a graph will not have knowledge of the algorithm used to construct the graph (as was the case in this study), leaving the meaningfulness of proximity unclear. Thus, although use of spatial proximity as a factor in judgment is not necessarily wrong in itself, overuse of this heuristic could lead to misinterpretation in some cases. Understanding how novice audiences interpret proximity could help inform the design of layouts and the use of graphs by analysts, particularly for graphs whose layout algorithms are independent of spatial distance.

## Future Directions

The present studies were carried out with participants who lacked a background in SNA. Future studies plan to examine how novices and experts differ in their use of graph information. We also plan to further assess the validity of proximity information by testing different graph layouts and examining correlations between path and spatial distance.

## Acknowledgments

This work was funded by the Department of Defense under Contract No.: H98230-14-D-0037 through the Computer and Analytic Sciences research group.

## References

Cook, V. J., Sun, S. J., Tapia, J., Muth, S. Q., Anguello, D. F., Lewis, B. L., Rothenberg, R. B., McElroy, P. D., \& the Network Analysis Project Team (2007). Transmission network analysis in tuberculosis contact investigations. The Journal of Infectious Diseases, 196, 1517-1527.
Gigerenzer, G. \& Goldstein, D.G. (1996). Reasoning the Fast and Frugal Way: Models of Bounded Rationality. Psychological Review, 103:4, 650-669.
Krebs, V. E. (2002). Mapping networks of terrorist cells. Connections, 24(3), 43-52.
McGrath, C., Blythe, J., \& Krackhardt, D. (1997). The effect of spatial arrangement on judgments and errors in interpreting graphs. Social Networks, 9, 223-242.
Micallef, L., Dragicevic, P., \& Fekete, J.D. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. IEEE Transactions on Visualization \& Computer Graphics, 18(12), 2536-2545.
Novick, L. R., \& Bassok, M. (2005). Problem solving. In K. J. Holyoak \& R. G. Morrison (Eds.), Cambridge Handbook of Thinking and Reasoning, pp. 321-349. New York: Cambridge University Press.
Purchase, H. C., Cohen, R. F., \& James, M. I. (1997). An experimental study of the basis for graph drawing algorithms. Journal of Experimental Algorithmics, 2(4).
Purchase, H. C. (2000). Effective information visualization: A study of graph drawing aesthetics and algorithms. Interacting with Computers, 13, 147-162.
Ware, C., Purchase, H., Colpoys, L., \& McGill, M. (2002). Cognitive measurements of graph aesthetics. Information Visualization, 1, 103-110.

# Explicit Predictions for IIIness Statistics 

Talia Robbins (talia.robbins@rutgers.edu) Pernille Hemmer (pernille.hemmer@rutgers.edu)<br>Department of Psychology, 152 Frelinghuysen Road<br>Piscataway NJ, 08901 USA


#### Abstract

People's predictions for real-world events have been shown to be well-calibrated to the true environmental statistics (e.g. Griffiths and Tenenbaum 2006). Previous work, however, has focused on predictions for these events by aggregating across observers, making a single estimate for the total duration given a current duration. Here, we focus on assessing predictions for both the mean and form of distributions in the domain of illness duration prediction at the individual level. We assess understanding for both acute illnesses for which people might have experience, as well as chronic conditions for which people are less likely to have knowledge. Our data suggests that for common acute illnesses people can accurately estimate both the mean and form of the distribution. For less common acute illnesses and chronic illnesses, people have a tendency to overestimate the mean duration, but still accurately predict the distribution form.


Keywords: Prediction; Judgment; Health; Cognition

## Introduction

Imagine that you have the flu and need to decide whether you will be better in time to travel to a conference this weekend. You are now faced with predicting how long you will be sick. For this inference, you will need to use your knowledge of real-world statistics, including both the mean duration and most likely form of the duration distribution.

People have been shown to make optimal predictions for the duration of many real-world events (Griffiths \& Tenenbaum, 2006). In these domains, people's beliefs about the underlying distribution of quantities (e.g. cake baking times are captured by a bimodal distribution) have been shown to be accurate in the aggregate. These findings have been extended to people's ability to make predictions for illness duration (Robbins and Hemmer, in revision). People were able to make predictions that were consistent with both the mean and form of illness distributions for common acute illnesses (e.g. common cold and seasonal flu), but systematically overestimated the duration of chronic illnesses (illnesses with which they had significantly less experience). This suggests they had knowledge of the correct form of the underlying illness duration distributions.

One limitation of the procedures used in previous experiments (e.g. Griffiths \& Tenenbaum, 2006) is that each participant made only one prediction about a total duration given its current duration. As such, data was aggregated over participants to assess the fit of participant data to the true duration distributions. As Griffiths and Tenenbaum (2006) explain, this gives a guide to peoples' implicit beliefs about the distributions. As such, these experiments do not allow for an assessment of whether people have knowledge of the correct form of the underlying illness distribution at
the individual level. Accordingly, these studies could be illustrating the wisdom of the crowds effect, whereby aggregating over many individual judgments from a group of people leads to a response that is closer to the ground truth than that of a smaller group (Surowiecki, 2004).

To our knowledge, no previous work has assessed the correspondence between people's beliefs and the statistics of the environment-specifically illness statistics-at the individual level. Therefore, in the current study, we assessed whether people understood the true statistics for the durations of different illnesses by asking them directly what they thought the mean and correct form of illness duration distributions were. This allowed us to evaluate whether people have an internal model for real-world statistics that they can consciously access and use to make predictions.

Understanding illness duration is critical for illness identification. For instance, imagine you have a cough and high fever, and thinking you have the flu you try to estimate how long you will be sick. One thing you will draw on is your understanding of the real-world distribution of durations for different illnesses. If your symptoms begin to fade after three days, this may confirm your suspicion that you have the flu, since this is within the normal distribution for the flu. However, if you are still sick after 10 days, you might begin to believe you have a different illness such as the common cold, because you know that 10 days is reasonable within the distribution of duration for the common cold. This estimation requires an understanding of the entire distribution of illness duration, rather than just the mean or some conditional duration. With only the mean of the distribution, you would not know how much variation in duration is normal, or at which point a particular illness is unlikely given the duration of your symptoms.

Illness further provides an interesting example for prediction because people have different levels of experience for different illnesses-e.g. common illnesses such as the cold, or less common illnesses such as bacterial meningitis. Experience may also differ between acute (e.g. cold) and chronic (e.g. asthma) illnesses. An acute illness is defined as one which can be cured with treatment, while a chronic illness is defined as one that can be managed but not cured. Differing levels of experience between chronic and acute illnesses may influence the accuracy of a person's prior beliefs, and different priors might be appropriate for different illnesses, given personal experience.

The observer's prior beliefs play an important role, as optimal predictions are assumed to follow Bayesian principles. Bayes rule gives a principled account of how people should update their prior beliefs given evidence from the world. Each time a person experiences an illness, they
should update their prior probability distributions for the duration of that illness. This would result in illnesses that are experienced more often having very accurate prior distributions. For illnesses that are less commonly experienced, people might adjust their prior beliefs to those of illnesses for which they have more knowledge of the correct form of the distribution, when making inferences. While people might use evidence from other sources when updating their priors, evidence that is personally experienced is better integrated than information acquired in other ways (Sallnas, Rassmus-Grohn, \& Sjostrom, 2000).

In this paper, we sought to assess whether the correspondence of people's beliefs to Bayesian optimal predictions in the aggregate (e.g. Griffiths \& Tenenbaum, 2006; Robbins \& Hemmer, in revision) extended to estimations for the mean and correct form of illness duration distributions at the individual level. We further sought to apply this approach to a domain with direct implications for real world problems-specifically patient health. In Experiment 1, we simply asked participants to predict the mean duration of each of nine illnesses.

In Experiment 2, we sought to assess whether people could make estimations of the correct form of illness distributions. To do this, we gave participants four distribution options-each fit to the true clinical data for that illness-and asked them to select the distribution form that best described that illness. Because each of the distribution options was fit to the clinical data, consistent selection of the correct distribution would clearly illustrate that there is a correspondence between people's internal model and the true statistics of the environment. This suggests that they have a cognitive representation of the form of the distribution of durations that they can consciously access.

## Experiment 1: Mean Estimation

## Methods

Participants Ninety-Nine Mechanical-Turk workers from the United States participated in exchange for $\$ 1$.
Materials We selected nine illnesses-five acute and four chronic (see Table 1) -intended to span a range of durations and familiarity. Familiarity was determined based on prevalence statistics for the number of people diagnosed with that illness each year (see Table 1). Table 1 also includes the source of the clinical data used for the illness duration distributions.

We first needed to determine the mean and correct form of the nine illness distributions. Illness durations have been found to be well modeled by a type of distribution known as a survival function, which includes Gamma, Exponential, and Weibull. The Erlang distribution is a special case of the Gamma distribution, where $\alpha$ must be an integer, which is often used to model illness duration and illness stages in transmission models of infectious disease, and to infer parameters from clinical data (Krylova \& Earn, 2013). For this reason, we assume Erlang is the correct distribution for the nine illnesses. See Figure 1 for the clinical duration


Figure 1: Histograms of clinical data for nine illnesses with best fitting Erlang distributions. Grey bars show the frequency of each illness duration, black lines show the Erlang fit to clinical data. $\mu$ gives the distribution mean.
distributions for the nine illnesses used in this experiment, with corresponding Erlang distribution fits. The clinical data provides a ground truth for both the mean and form of distributions to compare to participant responses (see Table 1 for clinical data sources).
Procedure The procedure was identical to that of Griffiths and Tenenbaum (2006), with the important difference that we did not condition on the current unit of time. As a consequence of the units of time available in the experiment, there may have been an anchoring effect, which is when people are systematically influenced by starting points regardless of whether they are informative (e.g. Chapman \& Johnson, 1999). By not providing the units of measure, we eliminate any possible anchoring effect. Our current procedure provides a truer picture of people's ability to estimate the mean because they are not given a frame of reference. Participants simply made a prediction about the total duration of each of the nine illnesses. The question read: "Given that you meet someone with illness $X$, what do you think will be the total duration of their illness?" Participants responded by typing in a number and selecting a unit of time from a dropdown menu presented on the computer screen. The experiment was performed using the Qualtrics interface. The order of presentation was randomized.

Participants were also asked to categorize each illness
Table 1: Sources for Clinical Data (in order of prevalence)
Illness Source of Clinical Data
(Prevalence $/ 10,000$ )
Acute (in order of prevalence)
Bacterial Meningitis (.14) Kilpi \& Anttila (1991)
Mononucleosis (5) Cameron et al. (2006)
Appendicitis (9)
Seasonal Flu (1250)
Singh et al. (2014)
Common Cold (2360)
Kohno et al. (2010)
Chronic (in order of prevalence)
COPD (4.5) Shavelle (2009)
Asthma (800) American Lung
Type II Diabetes (860)
Association (2012)
/statistics/duration/fig1.htm
Chronic Heart Disease (1130) Proudfit et al. (1983)
using one of five labels: "Lasts a short time, will go away completely even without treatment", "Can vary in length, requires immediate treatment, but can be cured", "Is long term, requires treatment, but can eventually be cured", "Lasts the rest of a person's lifetime, treatment can only manage symptoms, it cannot be cured, but does not necessarily cause death", "Varies in length, treatment can only manage symptoms, cannot be cured, eventually causes death". Participants were also asked several basic demographic questions (e.g. age and experience with the nine illnesses) which are not analyzed here.

## Results

Given that participants could respond with any unit of time, we first normalized participant responses to the unit of time for the clinical distributions. Responses were then filtered for outliers. Data was excluded in the following way: unreasonably large responses (defined as those 3 standard deviations greater than the mean response for a given illness) and participants who had more than two data points excluded based on the above criteria. The responses analyzed were 85 for appendicitis, 90 for the seasonal flu, 90 for the common cold, 87 for bacterial meningitis, 77 for mononucleosis, 90 for COPD, 90 for chronic heart disease, 90 for type II diabetes, and 90 for asthma.

First, we examined people's ability to characterize the durations of acute and chronic illnesses. Chronic illnesses are lifelong, which is a critical difference from acute illnesses which are curable. To determine whether participants had basic knowledge of the illnesses they were making estimations about, we examined their responses to questions asking to characterize each illness. For the common acute illnesses-common cold and seasonal flu$92 \%$ of participants correctly responded that the illnesses were short term and curable. For the less common acute illnesses-appendicitis and bacterial meningitis- $81 \%$ and $66 \%$ of participants respectively labeled these illnesses as short term. For the four chronic illnesses $74 \%-84 \%$ of participants correctly responded that these illnesses were lifelong. This clearly shows that people understand the chronicity of the chronic and common acute illnesses.

We first evaluated the accuracy of participant's mean


Figure 2: Red bars show the percentage of participants that were X number of standard deviations from the mean. Positive numbers indicate estimations above the mean, and negative numbers indicate estimations below the mean.
responses (see Table 2). A qualitative evaluation of the data illustrates that participant responses were close to the true mean for more prevalent acute illnesses (i.e. common cold and seasonal flu), and that participants overestimated the duration of chronic illnesses, similar to the pattern found by Robbins \& Hemmer (in revision).

In order to evaluate whether participant responses were accurate relative to the true mean of the empirical illness distributions, we used a two one-sided t-test approach (e.g. Limentani et al., 2005). We used this approach as it allows us to test for practical equivalence (e.g. Rogers, Howard, \& Vessey, 1993). A one-sample t-test might find a significant difference between a population mean of seven days and a participant response mean of eight days. While this difference is significant, it places too rigid a standard for our purposes, leading to an inaccurate conclusion that participants do not understand the mean of that illness. For this reason, we set a criterion considering accuracy to be within one standard deviation of the mean of the empirical illness distributions (standard deviations for each illness are displayed in Figure 2). We then conducted a t-test on either end of this threshold to determine if participant responses were significantly greater than the lower threshold, and significantly less than the upper threshold.

We found that for mononucleosis and the common cold,

Table 2: True and estimated illness durations

| Illness | True Duration | Participant Response | \% using unit of time (correct unit is bolded) |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  |  |  | Hours | Days | Weeks | Months |
| Years |  |  |  |  |  |  |
| Acute |  |  |  |  |  |  |
| Appendicitis | 39 hours | $471.6(\mathrm{SD}=969.5)$ hours | $\mathbf{8 . 4}$ | 32.6 | 39.0 | 12.6 |
| Seasonal Flu | 3.9 days | $8.9(\mathrm{SD}=4.5)$ days | 2.1 | $\mathbf{3 7 . 9}$ | 56.8 | 3.2 |
| Common Cold | 5.1 days | $6.3(\mathrm{SD}=3.2)$ days | 1 | $\mathbf{6 5 . 3}$ | 33.6 | 0 |
| Bacterial Meningitis | 5.5 days | $37.3(\mathrm{SD}=44.0)$ days | 2 | $\mathbf{1 0 . 5}$ | 45.3 | 36.8 |
| Mononucleosis | 10 weeks | $9.3(\mathrm{SD}=13.0)$ weeks | 1 | 9.5 | $\mathbf{3 5 . 8}$ | 32.6 |
| Chronic |  |  |  |  |  | 21.1 |
| COPD* | 6 years | $36.6(\mathrm{SD}=22.0)$ years | 0 | 1 | 0 | 5.3 |
| Type II Diabetes | 12 years | $36.0(\mathrm{SD}=22.5)$ years | 0 | 1 | 0 | 5.3 |
| Chronic Heart Disease | 13 years | $26.4(\mathrm{SD}=20.0)$ years | 0 | 1 | 2.1 | 2.1 |
| Asthma | 15 years | $42.5(\mathrm{SD}=25.7)$ years | 4.2 | 1.1 | 0 | 2.1 |

[^459]responses were within the one standard deviation of the true mean-meaning the estimates were practically equivalent to the true mean (upper threshold: Mononucleosis: $t(76)=-6.1$ $p<.01$, Common cold: $t(89)=-6.9$. $p<.0$; lower threshold: Mononucleosis: $t(76)=5.1$. $p<.01$, Common cold: $t(89)=13.4 . p<.01)$. For the other seven illnesses, responses were found to be greater than the lower end of the threshold, but not less than the higher end of the threshold, suggesting a pattern of overestimation, (Appendicitis: $t(84)=4.3 . p<.01$, Seasonal flu: $t(89)=13.0, p<.01$, Bacterial meningitis: $t(86)=7.1, \quad p<.01$, COPD: $t(89)=14.9, p<.01$, Type II diabetes: $t(89)=20.4, \quad p<.01$, Chronic heart disease: $t(89)=20.0 . p<.01$, Asthma: $t(89)=13.2 . p<.01)$.

Given that participants were not within the one standard deviation threshold for seven illnesses, we wanted to further examine how misaligned they were for each illness. Therefore, we calculated the percentage of participants at each standard deviation from the mean (see Figure 2). For the common cold and mononucleosis, the majority of participants (approx. 80\%) were within one standard deviation, as illustrated in the TOST. For the seasonal flu more than $70 \%$ of participants were within four standard deviations of the mean, which may seem like a large deviation from the correct response, however it is also important to note that the standard deviations varied greatly between illnesses. For the seasonal flu, the standard deviation was only 1.73 days, meaning that more than $70 \%$ of participants responded within 6.8 days of the true mean. Conversely, for the least prevalent acute illnesses, appendicitis and bacterial meningitis, only $34 \%$ and $38 \%$ of participants respectively were within four standard deviations of the true mean, with some participants being up to 80 standard deviations away (for appendicitis this corresponded to 1416 hours or 59 days). This illustrates that participants had lower agreement, and less accurate mean estimations for these illnesses.

For the chronic conditions, fewer participants were within four standard deviations of the mean, with $31 \%$ for COPD, $100 \%$ for type II diabetes, $61 \%$ for chronic heart disease, and $47 \%$ for asthma. Participant responses were all within four standard deviations of the mean for type II diabetes because the standard deviation is 24 years.

We then examined whether the absence of a time anchor influenced the unit of time participants used to respond (see Table 2). For the acute illnesses, multiple units of time can be used to express the same value; i.e., a one week long illness can be characterized as seven days or one week. For seasonal flu and common cold, more than $80 \%$ of participants responded with either the clinical (days) or the adjacent and reasonable (weeks) unit of time. For mononucleosis, approximately $66 \%$ of participants used the clinical or adjacent unit of time. For the least prevalent acute illnesses-appendicitis and bacterial meningitisparticipants used the clinical or adjacent unit of time only $40 \%$ and $55 \%$ of the time. For the four chronic illnesses, $92 \%$ to $95 \%$ of participants chose the clinical unit of time. The results suggest that participants could reliably use the
clinical unit of time when estimating durations of prevalent acute illnesses and chronic illnesses.

## Experiment 2: Distribution Form Estimation

## Methods

Participants Forty Mechanical-Turk workers from the United States participated in exchange for \$2. The participants had not participated in Experiment 1.
Materials The same nine illnesses from Experiment 1 were used. We selected four distributions as response options in the distributional form task: Erlang, Gaussian (a.k.a. Normal), Uniform, and Bimodal. These distributions were chosen as they can reasonably describe illness durations. The Erlang, which was always the correct answer, was chosen because illness distributions have been found to be well modeled by this distribution and provide a good fit for all the clinical distributions. Normal was chosen because the bell-curve is ubiquitous, and in some cases is very close to the Erlang distribution. This allows us to evaluate how well participants can discriminate very similar distributions. Bimodal was chosen because for chronic illnesses it might be reasonable to assume that there is one group of people who die immediately, and another group that lives with the illness for a longer time. Lastly, uniform was chosen because simple Bayesian prediction models assume a single uninformative (or uniform) prior (e.g. Gott, 1993). Selecting the uniform form of the distribution might suggest observers using a heuristic insensitive to prior beliefs.

Distributions were presented to participants as histograms of the average total duration of an illness. For each illness, the presented histograms were created by producing the best fit to the true clinical data for that illness for each of the four distributions. In this way, participants' choice of distribution


Figure 3: Screenshot of experimental interface for sample question (seasonal flu). Distribution types, top left to bottom right, are: Uniform, Normal, Erlang, and Bimodal.
would be based solely on distribution form. The histograms were presented with descriptive captions. The captions for each distribution form were consistent for all illnesses. Captions described several critical points on the graph using frequencies out of 100 (see Figure 3). The descriptions for each distribution form were matched to illustrate the same number of points on the histogram. Four naïve raters evaluated the relationship between the descriptors and the histograms and in all cases found them to be well-matched and easily understood. The experiment was presented using the Qualtrics interface.
Procedure Participants were first shown instructions on how to read graphs in our task. They then completed a training task, with two training sessions of four trials each. For each trial, participants were shown one histogram (illustrating one of the four distributions types used throughout this experiment) and asked to match it to one of four captions. The training trials were designed to illustrate duration without referencing illnesses. One set depicted the amount of time it takes for a person to turn into a zombie after being bitten, and the second set depicted the number of licks it takes to get to the center of a tootsie pop.

After the training task, participants were asked to choose the appropriate histogram from the four distribution options for each of the nine illnesses (presented one at a time) by selecting it with a radio button. Both question and choice order were randomized.

## Results

Data were excluded if participants answered two or more questions incorrectly in each of the two four trial trainingsets. This removed two participants' data from analysis.

First, we assessed the proportion of trials for which participants chose the clinical distribution (Erlang). Participants chose Erlang $42 \%$ of the time, which was significantly greater than chance ( $25 \%$ ), based on a onesided Binomial test ( $p<.01$ ). It was also chosen significantly more often than any of the other distributions: Normal $X^{2}(1, \mathrm{~N}=342)=11.8, \quad p<.01$, Uniform $\quad X^{2}(1, \mathrm{~N}=342)=93.9$, $p<.01$, and Bimodal $X^{2}(1, \mathrm{~N}=342)=48.0, p<.01$.

While participants selected Erlang with the greatest frequency overall, we were further interested in how frequently they chose the correct response for each individual illness. We performed a one-sided Binomial test and found that for six of nine illnesses, participants performed better than chance (i.e. significantly more than $25 \%$ of participants chose the Erlang distribution): seasonal flu ( $53 \%, p<.01$ ), common cold ( $50 \%, p<.01$ ), bacterial meningitis ( $42 \%, p=.016$ ), mononucleosis ( $42 \%$, $p=.016$ ), COPD ( $45 \%, p<.01$ ), type II diabetes ( $47 \%$, $p<.01$ ). Participants did not select any of the other distributions at a level higher than chance. See Figure 4 for the proportion of participants that chose each distribution option for the nine illnesses.

Lastly, we performed a chi squared test to determine whether participants selected the Erlang distribution significantly more often than the other distribution choices.

Participants chose Erlang more often than Uniform for eight out of nine illnesses: seasonal flu $X^{2}(1, \mathrm{~N}=38)=18.0, p<.01$, common cold $\left.X^{2}(1, \mathrm{~N}=38)=19.0, p<.01\right)$, bacterial meningitis $X^{2}(1, \mathrm{~N}=38)=9.8, p<.01$, mononucleosis $X^{2}(1, \mathrm{~N}=38)=17.0$, $p<.01, \operatorname{COPD} X^{2}(1, \mathrm{~N}=38)=13.3, p<.01$, chronic heart disease $\quad X^{2}(1, \mathrm{~N}=38)=7.9, \quad p<.01$, type II diabetes $X^{2}(1, \mathrm{~N}=38)=8.8, p<.01$, and asthma $X^{2}(1, \mathrm{~N}=38)=4.7, p=.03$.

Erlang was chosen significantly more than Bimodal for five of nine illnesses: seasonal flu $X^{2}(1, \mathrm{~N}=38)=9.7, p<.01$, common cold $X^{2}(1, \mathrm{~N}=38)=4.5, p=.03$, bacterial meningitis $X^{2}(1, \mathrm{~N}=38)=8.0, p<.01, \operatorname{COPD} X^{2}(1, \mathrm{~N}=38)=9.2, p<.01$, and type II diabetes $X^{2}(1, \mathrm{~N}=38)=23.6, p<.01$.

Participants chose Erlang significantly more than Normal for two out of nine illnesses: seasonal flu $X^{2}(1, \mathrm{~N}=38)=8.1$, $p<.01$, and common cold $X^{2}(1, \mathrm{~N}=38)=8.4, p<.01$. As shown above, Erlang was chosen significantly more often than any other distribution for both seasonal flu and common cold.


Figure 4: Red bars show the percentage of participants that chose a distribution choice.

## General Discussion

We evaluated people's ability to estimate the mean and correct form of duration distributions at the individual level within the domain of health. Examining people's representations of illness duration statistics is important, because it allows us to understand the correspondence between people's beliefs and the statistics of the environment-in this case-illness statistics. In addition, these experiments shed light on people's internal representations of real world statistics.

Our most interesting finding is that participants appeared to have knowledge of the correct form of the underlying illness distribution, choosing the assumed clinical distribution (Erlang) more frequently than any other distribution. When broken down by illness, they chose the clinical distribution more frequently for the most prevalent acute illnesses: seasonal flu and common cold.

While participants often inferred the form to be the normal distribution, this may be explained by the similarity of many of the normal fits to the Erlang fits. This occurred because the normal distributions were truncated by a lower duration bound of zero. We deliberately included the Normal distribution because of the potential confusability with the clinical distribution. As such, the fact that participants still chose the clinical distribution as the correct form overall, suggests they have strong beliefs about the
form of illness duration distributions and that these correspond to the environmental statistics. It is important to note that research has illustrated that people often fail at graphical interpretation (e.g. Gerteis et al., 2007), which makes participant performance in this task impressive.

When examining participants' estimates for the mean, we found that for more prevalent acute illnesses (i.e., common cold and seasonal flu), they were able to accurately estimate the mean duration. We also found a pattern of overestimation for chronic illnesses and less-prevalent acute illnesses which was similar to the pattern of overestimation found by Robbins and Hemmer (in revision).

The pattern of overestimation for chronic illnesses might be explained by people applying a probabilistic model of life expectancy to their understanding of the distribution form for illness durations. Because they have little experience with chronic illnesses, and they understand that chronic illnesses are life-long, their overestimation might be due to a strategy of applying parameters from the true distribution of lifespans (adjusted slightly to account for decreased life-expectancy with a chronic illness) to their knowledge that illnesses follow the form of an Erlang distribution. Their ability to select the appropriate distribution form for these illnesses suggests that they can use knowledge of the form of other illness distributions even if they do not have enough experience to set the parameters accurately. This overestimation might also be adaptive in terms of planning for the future. For chronic illnesses, it may be safer to assume a longer duration to plan sufficiently for the future, i.e., retirement savings.

A logical next step for this work would be to ask participants to independently generate distributions, rather than asking them to select from a limited number of options. Goldstein \& Rothschild (2014) have shown that participants can generate these distributions when presented with data, which suggests that this method could be used to evaluate peoples' internal representations of real-world statistics.

Our results illustrate that people hold accurate representations for both the form and mean of duration distributions of prevalent acute illnesses. Significantly, the most prevalent acute illnesses-the com mon cold and seasonal flu-are also the ones for which participants consistently demonstrate knowledge of the correct distribution form, and accurately predict the mean at the individual level. This suggests a prior belief that is better calibrated to the true environmental statistics for illnesses participants have experience with. Taken together, the data suggests that people have an internal representation of illness statistics that they can consciously accessindicating that people can not only combine illness experiences with rational statistical updating, but also have accurate knowledge of these prior distributions.

## Acknowledgments

This research was supported by the National Science Foundation grant \#1453276.

## References

American Lung Association (2012). Trends in Asthma

Morbidity and Mortality.
Cameron, B., Bharadwaj, M., Burrows, J., Fazou, C., Wakefield, D., Hickle, I., ... Lloyd, A. (2006). Prolonged illness after infectious mononucleosis is associated with altered immunity but not with increased viral load. Journal of Infectious Disease, 193, 664-671.
Chapman, G.B., \& Johnson, E.J. (1999). Anchoring, activation, and the construction of values. Organizational Behavior and Human Decision Processes, 79, 115-153.
Gerteis, M., Gerteis, J.S., Newman, D., \& Koepke, C. (2007). Health Care Financing Review, 28, 31-45.

Goldstein, D. G. \& Rothschild, D. (2014). Lay understanding of probability distributions. Judgment and Decision Making, 9, 1-14.
Gott, J.R. (1993). Implications of the Copernican principle for our future prospects. Nature, 363, 315-319.
Griffiths, T.L., \& Tenenbaum, J.B. (2006). Optimal predictions in everyday cognition. Psychological Science, 17, 767-773.
Gwaltney, J. (1967). Rhinovirus infections in an industrial population: II. Characteristics of illness and antibody response. JAMA, 202, 494-500.
Kohno, S., Kida, H., Mizuguchi, M., \& Shimada, J. (2010). Efficacy and Safety of Intravenous Peramivir for Treatment of Seasonal Influenza Virus Infection. Antimicrobial Agents and Chemotherapy, 54, 4568-4574.
Kilpi, T., \& Anttila, M. (1991). Severity of childhood bacterial meningitis and duration of illness before diagnosis. Lancet, 338, 406.
Limentani, G.B., Ring, M.C., Ye, F., Bergquist, M.L., McSorely, E. O. (2005) Beyond the t-test: Statistical equivalence testing. Analytical Chemistry, 221A-226A.
Proudfit, W. J., Bruschke, A. V. G., MacMillan, J. P., Williams, G. W. \& Sones, M. S. (1983). Fifteen year survival study of patients with obstructive coronary artery disease. Circulation, 68, 986-997.
Robbins, T. \& Hemmer, P. (in revision). Optimal predictions in illness cognition.
Rogers, J.L., Howard, K.I., \& Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. Psychon B Rev, 113, 553-565.
Rothberg, M.B., Haessler, S.D., \& Brown, R.B (2008). Complications of viral influenza. AM J Med 121, 258-264.
Sallnäs, E.L., Rassmus-Grön, K., Sjöström, C. (2000). Supporting presence in collaborative environments by haptic force feedback. Journal ACM Transactions on Computer-Human Interaction, 7, 461-476.
Surowiecki, J. (2004). The Wisdom of Crowds. New York, NY: W. W. Norton \& Company, Inc.
Shavelle, R.M., Paculdo, D.R., Kush, S.J., Mannino, D. M., Strauss, D. J. (2009). Life expectancy and years of life lost in chronic obstructive pulmonary disease: Findings from the NHANES III Follow-up Study. International Journal of Chronic Obstructive Pulmonary Disease, 137:148.
Singh, M., Kadian, Y.S., Rattan, K.N. \& Jangra, B. (2014). Complicated appendicitis: Analysis of risk factors in children.African Journal of Paediatric Surgery, 11,109-13.

# Information theoretic factors in marking linguistic focus: A laboratory-language approach 

Gareth Roberts (gareth.roberts@ling.upenn.edu)<br>University of Pennsylvania, Department of Linguistics, 3401-C Walnut Street, Philadelphia, PA 19104 USA

Jon Scott Stevens (stevens.400@osu.edu)
The Ohio State University, Department of Linguistics, 118G Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43201 USA


#### Abstract

We present an experimental study investigating the role of information-theoretic factors in determining patterns of redundancy and focus in language and other communication systems. Pairs of participants played a simple communication game using a non-linguistic visual medium to send messages to each other. We manipulated noise, effort, and time pressures and measured message length, redundancy, and accuracy. Participants behaved as predicted based on an informationtheoretic model, with message length and redundancy varying according to circumstance, but accuracy remaining constant.


Keywords: communication; focus; information theory; language; redundancy; signaling game

## Introduction

Questions invite answers. However, the invitation is not entirely open: There are important constraints on the form that an answer can take. Most obviously, it must be relevant (Grice, 1975; Wilson \& Sperber, 2004). In response to the question of who invented the printing press, for instance, "Мy sister loves cheesesteak" is not an acceptable answer unless some double meaning is understood. Perhaps less obviously, the length and syntactic completeness of an answer are also constrained. Fragment answers-that is, answers that are not whole sentences-are permitted and may even be preferred in some circumstances. At the same time, prosodic and morphosyntactic mechanisms (e.g., pitch fluctuations; cleft constructions such as "it was ... who") may be employed to add redundancy to certain elements in the sentence. Patterns of message length and redundancy should not be assumed to be random. In this paper we present a set of communication game experiments investigating how they vary according to information-theoretic factors.

As an example of how such patterns vary in English, consider the three answers to the printing press question that are given in (1). While (1a) is an acceptable answer, (1b) and (1c) are just as good, and would even be preferable in many contexts.
(1) Who invented the printing press?
a. Gutenberg invented the printing press.
b. Gutenberg did.
c. Gutenberg.

As evidenced by the acceptability of (1a), fragment answers are not obligatory. Indeed, whole-sentence answers
may help establish that the respondent heard and understood the question correctly; in some cases they may also serve to aid parsing of the what Schmitz (2008) calls the "critical" element (in this case Gutenberg). An answer like "The printing press was invented by Gutenberg", for instance, makes very clear through syntactic and prosodic means when precisely the listener should expect to get the answer. An utterance that does contain such unnecessary material is constrained in important ways; in particular, focus must be marked on the critical element. This is realized in English and many other languages by a pitch accent on that element (i.e., the element is distinguished from other parts of the sentence through variation in pitch, length and intensity; for discussion of this welldocumented phenomenon, see Ladd, 1996; Rooth, 1992). It is important to note that this focus is constrained by the grammar of the language and is more or less obligatorydistinguishing the wrong element, distinguishing the right element by the wrong means, or distinguishing no element at all, is likely to be perceived as odd. Example (2) illustrates this pattern, with small caps denoting pitch accent.
(2) Who invented the printing press?
a. $\checkmark$ Gutenberg invented the printing press.
b. $\checkmark$ Gutenberg did.
c. $X$ Gutenberg invented the PRINTING PRESS
d. $X$ GUTENBERG invented the PRINTING PRESS
e. $X$ Gutenberg invented the printing press

It should be noted that similar patterns pertain even when no explicit question has been asked. For example, a restaurant worker who notices a patron wandering around might say to them something like "Over there", "They're OVER THERE", or "The toilets are OVER THERE", even if the patron does not actually ask for directions.

Stevens (2016), following Schmitz (2008) and Bergen and Goodman (2015), argues for a theory of focus based partly in information theory (Shannon \& Weaver, 1949), whereby focus and redundancy are presented as a solution to noise, the random deletion/alteration of parts of a signal. Given that elements in messages may be lost as a result of mishearings, attention failures, and the like, focus is a means of emphasizing those elements that the speaker considers most important to transmit accurately. This can be seen as a process of adding
redundancy to important parts of the linguistic signal, compensating for the effects of noise. At the same time, depending on pressures of time and effort, non-focused elements (the parts of the signal that are easily inferred from context) can be reduced or even elided.

Stevens (2016) followed Rooth (1992) and many others in arguing that the inferrabillity of semantic material from the discourse context follows from the set of alternatives that is evoked by that context. For example, "Who invented the printing press?" evokes a set of possible answers \{x invented the printing press $\}$ which contains propositions like "Gutenberg invented the printing press", "Edison invented the printing press", etc. These alternatives all share overlapping semantic material, namely "invented the printing press" which is inferrable and thus less important in terms of information transmission. The material in need of protection from noise, in other words, is the material that does not overlap among all contextually available alternatives. We will therefore refer to redundancy on this material as non-overlapping redundancy, and redundancy on inferrable material as overlapping redundancy.

## Predictions of an information-theoretic model

If patterns of focus and redundancy in language are indeed a reflex of general information-theoretic concerns, then we expect that analogs to it will arise in any communication system that shares the goal of signaling the selection of one object from among a set of alternatives. For such systems we can thus make the following predictions. First, overall message length should vary according to time and effort costs: Messages should be longer if effort costs are low and time is not pressing. This might seem trivial, but the key point is that messages will not simply be minimally short in all cases. Because redundancy can be useful, messages will in fact be longer than they strictly need to be if time and effort constraints allow. Second, longer messages should differ from shorter messages not only with respect to length, but also with respect to the proportion of overlapping redundancy; that is, overlapping redundancy is a greater luxury than nonoverlapping redundancy, and should be dispensed with more readily. Third, the distribution of effort in a message should take noise into account. In particular, non-overlapping redundancy should be higher when noise is higher, both in an absolute sense (more redundancy overall) and in a relative sense (more non-overlapping redundancy than overlapping redundancy). Fourth, unless noise and time pressures become so great that they simply prevent accurate communication, we should expect communicative accuracy to remain relatively constant, regardless of the variation between messages predicted above. This is because that variation is (according to this account) designed to help maintain accuracy under different conditions.

There is evidence from natural language that speakers are sensitive to time and effort pressures. Corpus analysis suggests, for instance, that syntactic reduction is used to optimize information density (Jaeger, 2010), while work in experimen-
tal pragmatics demonstrates that referring expressions get shorter over time as interlocutors establish common ground (Krauss \& Weinheimer, 1964, Clark, 1996).

In general, however, it is hard to fully test the informationtheoretic basis of focus using natural-language data alone. There is more than one reason for this. A common difficulty in testing predictions in natural language is that the relevant explanatory factors are hard to manipulate (cf. Roberts, in press). To a small extent this is true here: While it is relatively straightforward to introduce noise and time pressures, for example, it is harder to make the act of producing natural-language utterances more effortful without imposing awkward physical constraints on participants. However, this is not the most serious obstacle in this case-in fact there is a history of research in phonetics in which awkward physical constraints are put on participants (e.g., Lindblom, 1990). The more serious problem is that, in established natural languages, there are constraints on such phenomena as pitch accent, focus, and utterance length that are encoded in the grammar. This means that while the patterns of interest might have arisen initially as a result of the information-theoretic factors discussed, we should expect them to an important extent to be "fossilized" as part of the language, incorporated into relatively constrained grammatical constructions, and less sensitive to contingent factors. Even if the information-theoretic account is right, therefore, behavior in established languages may be somewhat weak evidence for it; better evidence would be furnished by observing the establishment of a novel communication system. Furthermore, as stated above, we should expect our claims and predictions to apply across communication systems, linguistic or otherwise. It would therefore be advantageous for our purposes to test predictions in some non-linguistic communicative medium. Fortunately, the last decade or so has seen the development of a line of research in which participants communicate in the lab using "laboratory languages"-either artificial languages taught to the participants, or novel communication systems developed collaboratively over the course of the experiment (Galantucci, Garrod, \& Roberts, 2012, Roberts, in press). This approach allows researchers to investigate principles of communicative behavior while abstracting away as far as possible from the established natural languages that the participants bring with them into the lab. A common kind of study involves the use of visual communication systems. For instance, Garrod, Fay, Lee, Oberlander, and MacLeod (2007) had participants play a Pictionary-like game, while Galantucci (2005) made participants communicate by drawing on a pad that simulated continuous vertical motion, preventing the use of most established conventions. This kind of approach is particularly useful for investigating the challenges involved in establishing reference and constructing a new system from scratch. However, we were interested in the pressures acting on a system in which reference is in principle relatively straightforward, as is the case for a native speaker of a natural language, but where there are no established grammatical conventions as
in language. For this reason we had participants play a very simple communication game in which they had to fill in cells in a grid to convey a line that was drawn over the grid (Figure 1). Under ideal conditions, this task is rather trivial. We made conditions less ideal by manipulating noise, time pressure, and the effort required to produce a signal. We then measured the redundancy in the signals that were produced and the success rate in interpreting them.

## Method

## Participants

One hundred and twenty University of Pennsylvania undergraduate students participated in pairs for course credit or $\$ 5$.

## Procedure

In each trial a pair of participants played a simple cooperative signaling game. Each sat in a separate cubicle with a computer; neither participant could see or hear the other. The game consisted of a series of turns in each of which one player was nominated as Sender and the other as Receiver; players alternated roles, with the Sender in the first turn being selected at random. At the start of a turn the Receiver saw a white screen with the message "You are waiting on a message transmission from the other player". The Sender saw a screen as in Figure 1 On the left were two $7 \times 7$ grids, over each of which a different line figure was drawn. Every line figure consisted of a continuous line drawn between the center points of eleven contiguous grid cells. One of the two line figures was selected in green, while the other was white $?^{1}$

On the right of the screen was an empty $7 \times 7$ grid, slightly larger than the other two. Beneath the two leftmost grids there was also a button marked Send. The Sender's task was to communicate to the Receiver which of the two grids was selected in green by clicking on cells in the rightmost grid. At the moment of the Sender's first click a timer would start. Once the timer stopped (after either 5 or 30 seconds, depending on condition; see Section Experimental conditions) the grids would disappear and be replaced by the message "You are waiting on a guess from the other player." Clicking the Send button would have the same result. Once the Sender's turn had come to an end in one of those two ways, the Receiver's screen would change to display the two line figures that had been displayed to the Sender (in a random order, but each in the same orientation as for the Sender) as well as a third $7 \times 7$ grid in which certain cells might be colored black, and a button marked $O K$ (Figure 22. The black cells would always be cells that the Sender had clicked; however, not all cells that the Sender had clicked would necessarily be sent. The means of deciding which cells would be sent depended on the condition (Section Experimental conditions). The Receiver's task was to select which of the two line figures they

[^460]thought the Sender was trying to communicate. Both players were then told whether the Receiver chose correctly. Then a new turn began. There were 48 turns in total, which were preceded by two practice turns. In half the turns (the Overlap turns) the two line figures overlapped by five squares (as in Figures 1 and (2). In the other half (the Filler turns) there was no such overlap, such that any cell through which a line figure passed would serve to distinguish it from its competitor.

The only differences between the practice turns and the other turns were that the players' success in the practice turns did not count toward their final score, and that, following the practice turns, they were told to ask questions if they had any. If players scored over $80 \%$ in the non-practice turns, they were rewarded with $\$ 2$ each.


Figure 1: Sender's screen.


Figure 2: Receiver's screen.

## Experimental conditions

There were six between-subjects conditions in total (Table 1 ). Conditions differed from each other with respect to the time available to the sender (either 5 seconds or 30 seconds) and with respect to the means with which the sender could ensure that a message be sent. In the Effort conditions clicking a cell a specific number of times would guarantee that it was sent

Table 1: Experimental conditions

| Time limit | Effort | Noise |
| :---: | :--- | :--- |
| 30 seconds | High effort | High noise |
| 30 seconds | Low effort | Low noise |
| 5 seconds | Low effort (5s) | Low noise (5s) |

to the Receiver. The Sender had to click fifteen times on each cell in the High effort condition and five times in the Low effort conditions. Once a cell had been clicked the requisite number of times, it turned black. In the Effort conditions, a cell that turned black for the Sender was sure of being seen by the Receiver, but a cell that had not been clicked enough remained white and was sure not to be sent. The same was not true in the Noise conditions. In these conditions, each click on a cell would make it appear darker. Any cell that had been clicked once had a chance of being sent. Clicking more on the cell not only made the cell darker, but increased that chance (and the darkness of the cell increased proportionally to the probability that the cell would be sent). The probability of a given cell being sent was calculated as $1-(1-d)^{n}$, where $d$ is a decay parameter between 0 and 1 , and $n$ equals the number of times the Sender clicked on the cell in question. Two values were used for the decay parameter. In the High noise condition, it was set at 0.1 . In the Low noise conditions it was set at 0.4 . This meant that it would take many more clicks in the High noise condition than in the Low noise conditions to feel confident that the cell would be sent. For instance, 4 clicks in the latter condition would result in an $87 \%$ chance of the cell being sent, but a $34 \%$ chance in the former. In all conditions the number of cells clicked on can be taken to correspond roughly to utterance length, while in the Noise conditions, cell darkness can be taken to be analogous to greater effort in marking prominence.

The length of time available for the Sender to click on cells was varied, being set at either 5 or 30 seconds. For the Low effort and Low noise conditions, there was both a 5 -second condition and a 30 -second condition. For the High effort and High noise conditions, however, we ran only 30 -second conditions; this is because 5 seconds was not enough time for participants to send more than one cell in the High effort condition, or to have a good chance of doing so in the High noise condition, meaning that the results would be rather trivial and most of our measures could not be calculated.

## Measures

We measured the following variables.

Message length. This was calculated by counting how many squares had been clicked on (for Noise conditions) or turned black (for Effort conditions). Since one cell would be sufficient to distinguish between any two line figures (assuming it arrived), any message consisting of more than
one cell can be considered to contain redundancy.
Click effort. This was calculated by counting the number of clicks made by the Sender.

Overlapping redundancy. This was calculated by counting how many cells the Sender clicked on (for Noise conditions) or turned black (for Effort conditions) that overlapped with both line figures, and would therefore not differentiate the target figure from its competitor. (Note that this could be measured only for the Overlap turns, and not the Filler turns.).

Accuracy. This was calculated by counting the number of turns in which the Receiver chose the correct line figure.

## Results

Data reported in the following sections come from Overlap turns only, for the sake of consistency between the measures. Unless otherwise stated, all models reported are mixed effects linear regression models with random intercepts for both subject and item, using the Satterthwaite approximation of degrees of freedom to obtain a $p$-value from a t -value.

## Variation in message length

We predicted that overall message length should vary according to time and effort costs, with messages being longer than strictly necessary if time and effort constraints allowed. This was supported by the data. One cell would have been sufficient in every condition to signal which line figure to choose; however, messages were longer than this in every condition (Figures 3 and 4). In all but the High noise and High effort conditions, message length also remained relatively constant throughout a game, though it was lower in the Low noise (5s) than in the Low noise condition ( $\beta=-4.83$, $S E=$ $0.75, t=-6.48, p<0.001$ ) and in the Low effort ( 5 s ) condition than in the Low effort condition ( $\beta=-7.64, S E=$ $0.65, t=-11.83, p<0.001$ ), suggesting that the shorter message lengths in the 5 s conditions were due to time constraints. In both the High noise condition and the High effort condition, message length began high and fell over the course of the game in a rather linear fashion, converging in the High effort condition with the Low effort (5s) condition. This suggests that effort came to exercise increasing pressure as time went on. It is likely also the case that it partly reflects participants' growing familiarity with the game, although if this were the main explanation, one would expect a sharper fall by the middle of the game rather than a linear decline. Message length in the noise conditions did not quite converge (i.e., the slope was less steep), likely because less effort was necessary in this condition.

## Message length and overlapping redundancy

We predicted that shorter messages should differ from longer messages with respect to the proportion of overlapping redundancy. This was confirmed. The proportion of clicks devoted


Figure 3: Message length in effort conditions.


Figure 4: Message length in noise conditions.
to overlapping redundancy (as opposed to non-overlapping redundancy) correlated with click rate ; $r=0.29, p<0.001$ ), but the correlation was stronger $(r=0.56)$ in the High effort and High noise conditions alone, where there was greater pressure on participants (Figure 5).

## Distribution of effort

We predicted that the distribution of effort would take noise into account and that non-overlapping redundancy would be higher when noise is higher, both in an absolute sense and relative to overlapping redundancy. This was confirmed by results. First, click effort was lowest in the Low noise (5s) condition, but higher in the High noise condition than in the Low noise condition ( $\beta=58.57, S E=7.56, t=7.75, p<0.001$ ), in both of which rounds lasted 30 seconds (Figure 6. Sec-


Figure 5: Correlation between click effort and overlapping redundancy (High effort and High noise conditions only).


Figure 6: Click effort in noise conditions.


Figure 7: Proportion of clicks on overlapping redundant cells in Noise conditions.
ond, for the proportion of overlapping redundancy, there was a significant interaction of turn and condition for the High noise and Low noise conditions ( $\beta=0.006, S E=0.002, t=$ 3.07, $p<0.01$ ): Although all the noise conditions began at roughly the same place (Figure 7), the proportion of overlapping redundant cells clicked in the Low noise condition remained relatively constant, but decreased on the High noise condition. There is also a main effect of condition if the interaction terms are excluded from the model $(\beta=0.11, S E=$ $0.04, t=2.81, p<0.01)$. The proportion of clicks on overlapping redundant cells was significantly lower in the Low noise ( 5 s ) than in the High noise condition ( $\beta=-0.10, S E=$ $0.04, t=-2.44, p<0.05)$.

## Accuracy rates

We predicted that accuracy would remain similar between conditions, regardless of differences in message length, click effort, and redundancy. This was broadly true. Overall mean accuracy was very high ( $97 \%$ ), and did not differ significantly between conditions, with one exception (Figure 8): It was lower in the Low noise (5s) condition ( $\beta=-0.07, S E=$ $0.02, t=-2.81, p<0.01)$. This likely represents an underestimation of noise by participants; this was the only condition in which senders had little time to send a signal, but could not be sure before sending it how much the receiver would see.

## Discussion

We modeled the emergence of a system similar to linguistic focus by having participants play a simple non-linguistic


Figure 8: Accuracy rates by condition. Error bars show 95\% confidence intervals.
communication game, in which we manipulated noise, effort, and time pressures. We made four predictions consistent with an information-theoretic account of focus: that message length should vary according to time and effort costs, that longer messages should differ primarily with respect to redundant material that is shared with alternatives, that the distribution of effort in the message should take noise into account, and that accuracy should be stable in spite of variation in other measures. These predictions were confirmed. The patterns we observed also resemble patterns in natural language. While message length varied considerably between conditions, accuracy was maintained, largely due to distribution of effort being skewed toward protecting nonoverlapping material from noise. On the one hand this is consistent with the natural-language examples given in the introduction to this paper. On the other hand, it is also consistent with findings from experimental pragmatics studies in which participants repeatedly refer to a set of unfamiliar shapes; referring expressions are reliably shorter at later stages of such interactions (Krauss \& Weinheimer, 1964, Clark, 1996). This is typically explained in terms of the establishment of common ground-participants develop a shared perspective on the objects in question. On the surface, this does not obviously apply so well to our study, but the fundamental mechanism is the same. The point in both cases is that message length varies as a result of time and effort pressures on the one hand, and the sender (or speaker)'s confidence that the message will be understood by the receiver, on the other.

Overall, our results lend support to the view that linguistic focus may have emerged as a response to informationtheoretic pressures. We also consider that the experimental approach we have taken may prove fruitful for future work.

## Acknowledgments

We are grateful to Robin Clark for insightful comments on the development of this project, to Lisa Raithel for preparing the software, and to members of the Cultural Evolution of Language Lab for assistance in running the experiment. This work was funded in part by the American Council of Learned Societies (ACLS).

## References

Bergen, L., \& Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. Topics in Cognitive Science, 7(2), 336-350.
Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.
Galantucci, B. (2005). An experimental study of the emergence of human communication systems. Cognitive Science, 29(5), 737-67.
Galantucci, B., Garrod, S., \& Roberts, G. (2012). Experimental Semiotics. Language and Linguistics Compass, 6(8), 477-493.
Garrod, S., Fay, N., Lee, J., Oberlander, J., \& MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? Cognitive Science, 31(6), 961-87.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. L. Morgan (Eds.), Syntax and semantics, vol. 3: Speech acts (pp. 41-58). New York: Academic Press.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61(1), 23-62.
Krauss, R. M., \& Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction-a preliminary study. Psychonomic Science, l(5), 113-114.
Ladd, D. R. (1996). Intonational phonology. New York: Cambridge University Press.
Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H\&H theory. In W. J. Hardcastle \& A. Marchal (Eds.), Speech production and speech modelling (Vol. 55, pp. 403-439). Dordrecht: Kluwer Academic Publishers.
Roberts, G. (in press). The linguistic Drosophila: Experiments in language change. Linguistics Vanguard.
Rooth, M. (1992). A theory of Focus interpretation. Natural Language Semantics, 1, 75-116.
Schmitz, H.-C. (2008). Accentuation and interpretation. New York: Palgrave MacMillan.
Shannon, C. E., \& Weaver, W. (1949). The mathematical theory of communication. Champaign, IL: University of Illinois Press.
Stevens, J. (2016). Focus games. Linguistics and Philosophy, 39(5), 395-441.
Wilson, D., \& Sperber, D. (2004). Relevance theory. In L. R. Horn \& G. Ward (Eds.), Handbook of pragmatics (pp. 607-632). Oxford: Wiley Blackwell.

# Manual Response Dynamics Reflect Rapid Integration of Intonational Information during Reference Resolution 

Timo B. Roettger \& Mathias Stoeber<br>timo.roettger@uni-koeln.de, m.stoeber@uni-koeln.de<br>Department of Linguistics - Phonetics, University of Cologne, Herbert-Lewin-Straße 6<br>D-50931 Cologne, Germany


#### Abstract

Intonation plays an integral role in comprehending spoken language. It encodes post-lexical pragmatic functions such as sentence modality and discourse contexts. The present experiment investigates how and when listeners integrate intonational information to anticipate reference resolution. While most work on the real-time processes of intonation-based intention recognition has utilized eye tracking, the present study uses the mouse tracking paradigm, a valuable complementary method to investigate the time course of speech processing. Participants had to choose an interpretation based on pre-recorded instructions containing different intonation contours. Recordings of the $\mathrm{x}, \mathrm{y}$ coordinates of participants' computer mouse movements reveal that listeners integrate intonational information rapidly as soon as they become available and anticipate potential referential interpretations early on.


Keywords: intonation, reference resolution, mouse tracking

## Introduction

During the perception of an unfolding speech signal, listeners use acoustic information to guide their interpretation of what a speaker intends to communicate. This process can take place long before disambiguating lexical information becomes available, allowing the listener to make rapid inferences about what a speaker intends to say, even if these inferences are based on partial information.

Intonation plays an integral role in this interpretation process. Among other things, intonation is commonly used to express discourse relations such as givenness and contrastiveness (e.g. Ladd 2008). Intonational acoustic events such as pitch accents have been shown to consistently encode the discourse status of referents. For example, in German or English, a high rising pitch accent generally signals new information, while for example deaccentuation (i.e. the absence of a pitch accent) tends to signal given information (e.g. Fery \& Kügler 2008, Cangemi et al. 2015 inter alia).

While much work has been done on how intonational events encode discourse relations, there is only little work on how and when listeners integrate this acoustic information with the relevant discourse information.

To fully understand intonation-based intention recognition, it is necessary to use experimental techniques that measure the real-time integration of intonational information to resolve temporally ambiguous interpretations. While eye tracking experiments have advanced our knowledge about these processes tremendously (e.g. Dahan et al. 2002, Weber et al. 2006, Watson et al. 2006, Kurumada et al. 2014), it has been pointed out that the nature of oculomotor patterns constitutes a limitation of the eye-tracking paradigm (e.g. Spivey et al.

2005, Dale et al. 2007). Eye-movement data is characterised by ballistic "jumps" of the eye. Only by averaging over many trials can a pseudo-continuous trajectory be calculated, which is then interpretable as evidence for a continuous comprehension process.

This potential methodological shortcoming can be overcome by measuring another form of movement behaviour: the movements of hands. Over the last decade, it has been demonstrated that continuous nonlinear trajectories recorded from the streaming of $x, y$ coordinates of computer mouse movements can serve as an informative indicator of cognitive processes (e.g. Spivey et al. 2005, Magnuson 2005). Even though mouse tracking has been applied to diverse phenomena in cognitive science, its usefulness for speech processing research has been somewhat neglected. This paper provides evidence that mouse tracking is suitable to unravel real-time dynamics of speech processing beyond lexical and phonemic processing (see also Tomlinson \& Bott 2013 and Warren 2017).

## Real Time Integration of Intonation

Several studies have demonstrated that comprehenders can rapidly integrate intonational information to map an utterance containing referential expressions onto intended referents. These studies have focused on the discourse status of referents, i.e. whether an item has or has not already been mentioned or is explicitly contrasted to another referent.

Dahan et al. (2002) utilised the visual world paradigm in which specific items and geometrical shapes were distributed in a grid. Upon hearing specific auditory instructions listeners had to move the objects above or below the shapes. In one of their experiments, subjects heard a trigger sentence such as "Put the candle below the triangle." The object "candle" and the location "below the triangle" were thus introduced to the listener as given information. After the trigger sentence, listeners heard the critical instruction, either referring to the given object ("candle") or a lexical competitor which shares its word onset with the given object (here "candy"). When the target word was deaccented, listeners' eye movements revealed significantly more fixations to the already given object before the lexical disambiguation was available. Conversely, when the target word was accented, there were more fixations to the competitor.

Weber et al. (2006) extended these findings by showing that the presence or absence of a contrastive pitch accent on a modifying adjective allows listeners' anticipation of contrastivity of the noun (see also Watson et al. 2006).

Kurumada et al. (2014) examined the time course of the construction "It looks like an X" pronounced with either a high pitch accent on the final noun followed by a low boundary tone (e.g. it looks like a ZEBRA), or a contrastive high rising pitch accent on the verb and a rising boundary tone, a contour that can support contrastive inference, (e.g. it LOOKS like a zebra (but it is not)). They found that listeners integrate the pitch accent information on the verb to anticipate the status of the target referent.

These reported effects are consistent with the hypothesis that listeners integrate intonational information rapidly as soon as it becomes available and anticipate potential referential interpretations early on.

The contributions of the present paper to this literature are twofold: On the one hand, we aim to replicate previous findings showing that listeners take up intonational information rapidly to anticipate pragmatic interpretations using mouse tracking. On the other hand, we aim to proof that continuous response tracking can provide valuable insights into the realtime comprehension of utterance-long speech signals.

## The Present Study

The present study investigates intonation-based intention recognition in German using the mouse tracking paradigm. It is hypothesised that listeners integrate intonational information to anticipate referential ambiguity early on and that this anticipation is reflected in the dynamics of their mouse trajectories during response selection. In line with standards of reproducible research, all materials (including audio and visual stimuli), scripts, and raw data are available here: https://osf.io/n79x3

## Methodology

Participants had to choose visually presented response alternatives corresponding to pre-recorded speech files in a twoalternative forced choice design. Stimuli differed with respect to the available discourse context and the intonationally encoded information status of referents, enabling us to investigate the real-time integration of intonational information during reference resolution.
Experimental Set-up Participants were seated in front of a MacBook Pro 3.1 GHz Intel Core i7 with a display resolution of $1280 \times 800$. They controlled the experiment via a Logitech B100 corded USB Mouse. Cursor acceleration was made linear and cursor speed was slowed down using the CursorSense® application (version 1.32).

Participants and Procedure Ten native speakers of German (five male, five female) with an average age of 30.3 years $(\mathrm{SD}=4.9)$ participated in this experiment. All of them had normal or corrected-to-normal vision.

Participants were told about two different fantasy creatures which were introduced as 'wuggies'. These wuggies were displayed as having picked up certain real world objects such as a pear or a violin. The two wuggies differed in colour
("blau" 'blue' vs. "gelb" 'yellow’) and there were 10 different objects that the wuggies could pick up (bee, chicken, fork, marble, pants, pear, rose, saw, vase, violin).

On each trial, participants were exposed to a question screen that either did or did not provide a specific discourse context, followed by a response screen in which participants had to choose visually presented response alternatives depending on an auditorily presented sentence.

On the question screen, participants either heard nothing or they heard a question such as (1):
(1) "Hat der gelbe Wuggy die Geige aufgesammelt?" 'Did the yellow wuggy pick up the violin?'

The question provided a discourse context with certain elements being activated as given for the participant (here: the yellow wuggy and the violin). The question screen was visible for 2500 ms .

Following the question screen, participants saw two visually presented response alternatives, each depicting a wuggy carrying an object. After 1000 ms , a yellow circle appeared at the bottom centre of the screen. Participants were instructed to click on the yellow circle to initiate playback of an audio recording. The audio recording was a statement specifying which wuggy has picked up which object, e.g., "Der gelbe Wuggy hat die Geige aufgesammelt." 'The yellow wuggy has picked up the violin.' Participants were instructed to move their mouse immediately upwards after clicking the initiation button and choose the respective response alternative as soon as they could.

After each response selection, the screen was left blank for a 1000 ms inter-stimulus interval.

Prior to the beginning of the experimental trials, participants were given 36 practice trials to familiarise themselves with the paradigm.

Speech Material There were two sets of acoustic stimuli: questions providing a discourse context presented on the question screen and statements triggering participants' responses on the response screen. There were twenty different questions for all possible combinations of wuggies and objects (two wuggies $\times$ ten objects).

Likewise, there were twenty different statements, which were produced with four different intonation contours. Based on the question, i.e. the discourse context, and the visual scene at hand, statements differed with regard to the information status of the relevant constituents of the sentence: The question in (1) ("Hat der gelbe Wuggy die Geige aufgesammelt?") asks for confirmation that the proposition (including the identity of the subject and object) is true. Now consider the following answers (2-4):
(2) Der gelbe Wuggy hat die Geige aufgesammelt. 'The yellow wuggy has collected the violin.'
(3) Der gelbe Wuggy hat die Birne aufgesammelt. 'The yellow wuggy has collected the pear.'
(4) Der blaue Wuggy hat die Geige aufgesammelt. 'The blue wuggy has collected the violin.'


Figure 1: Representative waveform and f 0 contour for a statement produced with a rising accent on "Wuggy" and a falling accent on "Geige", a typical contour for broad focus. Accented words are highlighted with grey boxes.


Figure 2: Representative waveform and f0 contour for a statement produced with a rising accent on the auxiliary "hat", typically used to indicate verum focus. The accented word is highlighted with a grey box.

Dependent on the discourse context (here: whether there is question or not, and the question being asked), the answers in (2-4) are realised with different intonation contours (Fery \& Kügler 2008, Cangemi et al. 2015). If there is no discourse context available, both the subject and the object are new information in (2) (often referred to as broad focus) which can be prosodically encoded by specific pitch accents on both constituents. A common contour in these cases is a rising accent on the subject, followed by a high stretch of f 0 and a high or falling accent on the object (cf. Figure 1).

Alternatively, if there is a relevant discourse context such as the question in (1), the utterance in (2) can prosodically emphasise that the proposition of the question is true. This can be indicated, for example, by verum focus, which manifests itself here in the form of a high rising accent on the auxiliary ("hat", cf. Figure 2).

In contrast, the answers in (3) and (4) correct the proposition of the question. In (3), "Birne" is explicitly contrasted with "Geige", typically expressed by an intonation contour with a high rising accent on "Birne" (cf. Figure 3). In (4), "blaue Wuggy" contrasts with "gelbe Wuggy" in the question. In this context, a high rising accent on "blaue" and no


Figure 3: Representative waveform and f 0 contour for a statement produced with a rising accent on the referent "Birne", typically used to indicate contrastive focus. The accented word is highlighted with a grey box.


Figure 4: Representative waveform and f 0 contour for a statement produced with a rising accent on the subject modifier "blaue", typically used to indicate contrastive focus. The accented word is highlighted with a grey box.
accent on "Geige" is typically found (cf. Figure 4).
All acoustic stimuli were produced by a trained phonetician in a sound-attenuated booth at the Institute of Phonetics in Cologne with a headset microphone (AKG C420) using 48 $\mathrm{kHz} / 16$ bit sampling. The average stimulus duration of the trigger sentences was 1993 ms .

Visual Stimuli The pictures of the fantasy creatures were taken from a hand drawn set developed and used by van de Vijver \& Baer-Henney (2014). The pictures of objects were taken from the BOSS corpus (Brodeur et al. 2010).

Response alternatives of critical trials differed visually by the identity of the referent only (e.g. yellow wuggy carrying a pear vs. yellow wuggy carrying a violin). In addition to the critical trials, we included the same number of filler trials, in which response alternatives differed visually by the colour of the wuggy only (e.g. yellow wuggy carrying a pear vs. blue wuggy carrying a pear), or by both the colour of the wuggy and the identity of the object (e.g. yellow wuggy carrying a pear vs. blue wuggy carrying a violin). These visual contrasts were introduced to ensure that participants do not simply
learn to anticipate certain combinations of questions and visual contrasts, disregarding the acoustic information.

Stimuli Presentation and Predictions There were four different experimental conditions: In the broad focus condition, participants did not receive a question and had to respond to a broad focus statement (cf. Figure 1). Since participants had no discourse context available, they had to rely on lexical information only. It is expected that the mouse movements during reference resolution do not change until the lexical information becomes available (the onset of "Geige" in example 2).

In the other three conditions, participants received a question and were thus able to integrate the given discourse context with the intonational information encoding the information status of the referents. Participants saw a already mentioned, given object (here: "Geige") and a new object (here: "Birne").

In the object focus condition, the pitch accent on the object indicates the contrastive nature of the object. The available pitch accent information becomes available simultaneously with the lexical information, i.e. the rise in pitch starts at the onset of the word (cf. Figure 3). Assuming that the pitch accent information primarily cues contrastivity, we do not expect listeners to anticipate the referent, i.e. the broad and object condition should not differ.

In the verb focus condition, the pitch accent on the verb indicates verum focus, i.e. signalling that the proposition of the question is true implying that the statement contains the already mentioned object (here: "Geige"). As soon as the intonational information on "hat" becomes available, listeners are expected to integrate this information, enabling reference resolution before the lexical information becomes available.

In the subject focus condition, the pitch accent on the subject modifier indicates the contrastive nature of the subject. This information enables an early inference towards the given nature of the object which only occurs later in the utterance, making reference resolution possible very early on.

Left/right placement of target vs. distractor response alternatives was counterbalanced within participants.

Analysis The x, y screen coordinates of the computer mouse were sampled at 100 Hz using the mousetrap plugin (Kieslich \& Henninger 2016) implemented in the open source experimental software OpenSesame (Mathôt et al. 2012). Trajectories were processed with the package mousetrap (Kieslich et al. 2017) using the statistical software R (2016).

There was a total of 80 target trials, for a grand total of 800 trajectories across participants ( 200 per condition). Overall, $4.36 \%$ of trials with incorrect responses and $0.45 \%$ of trials with initiation times greater than 500 ms were discarded. Additionally, $1.67 \%$ of trials were excluded due to movement behaviour that violated instructions (loops, reaching the top of the screen before response selection).

For each of the remaining trials, we computed two measurements based on time- and space-normalised trajectories: First, we collected overall reaction times (RT) measured from
the initiation click up until reaching the target response. This serves as a latency baseline. Second, we measured the area under the curve (AUC) operationalised by the geometric area between the observed trajectory and an idealised straight-line trajectory drawn from the start and end points (Freeman \& Ambady 2010). A greater AUC is indicative of greater response competition between target and competitor during response selection.

We analysed data using hierarchical linear models using R and the package lme4 (Bates et al. 2015), afex (Singmann et al. 2016), and lmerTest (Kuznetsova, Brockhoff, \& Christensen 2016). Discourse condition (broad, object, verb, subject) was included as a fixed effect. Participants were specified as by-condition random slopes and referents were specified as random intercepts.

## Results and Discussion

Inspection of time- and space-normalised horizontal trajectories over time (cf. Figure 5) suggests that trajectories were characterised by initially gravitating toward the midpoint between response alternatives (horizontal cursor position $=0$ ) before eventually curving towards the target response (horizontal cursor position $=-1$ ).

Focus conditions elicited similarly-shaped trajectories that mainly differed with respect to their temporal characteristics. Not surprisingly, conditions differed in their overall response latency, measured from clicking the initiation circle to reaching the target response area $\left(\chi^{2}(3)=19.6, \mathrm{p}=0.0002\right)$ with the broad condition being the overall slowest ( $\beta=1578 \mathrm{~ms}$, $\mathrm{SE}=40.7$ ) followed by the object condition $(\beta=1457 \mathrm{~ms}$, $\mathrm{SE}=49.5$ ), the verb condition ( $\beta=1367 \mathrm{~ms}, \mathrm{SE}=51.9$ ), and the subject condition ( $\beta=1121 \mathrm{~ms}, \mathrm{SE}=93.8$ ) (cf. Figure 6, Table 1). Pairwise comparisons reveal significant differences between all four conditions. The earlier the relevant intonational cue in the acoustic signal, the faster listeners selected a response. Moreover, the difference between broad condition and object condition suggests that the integration of discourse context and intonation facilitated reference resolution (object) in contrast to cases without available discourse context (broad).

These overall response latencies were neatly reflected in early moments of direction change: The broad condition started curving towards the target response around 300 ms after lexical disambiguation (dashed line in Figure 5). This time-lag can be interpreted as the time it takes for listeners' movements to be affected by the relevant acoustic cue of the lexical item.

The subject condition elicited response trajectories that deviated towards the target response very early in the signal, after having heard the contrastive pitch accent on the subject modifier. As opposed to that, the verb condition started curving towards the target shortly after the acoustic onset of the referent suggesting that integrating intonational information of the verum focus led to an immediate anticipation of the target referent before lexical disambiguation has taken place.


Figure 5: Horizontal cursor position of mouse trajectories plotted over time for broad, object, verb, and subject condition. Dashed line indicates the averaged acoustic onset of the critical lexical item.


Figure 6: Violin plots of overall response latency (RT) of response selection.


Figure 7: Violin plots of area under the curve (AUC) values of response selection.

Crucially, the object condition started its curvature towards the target less than 200 ms after the point of lexical disambiguation. Taking the time lag of the broad condition into account, it becomes clear that even the object condition elicited trajectories that started curving towards the target response before lexical disambiguation had taken place. Listeners' anticipation in the object conditions seems puzzling. Both the lexical cue (onset of disambiguating phones) and the intonational cue (onset of the rising pitch movement) become available in the signal at the same time, i.e. the onset of the referential expression. The question arises as to how listeners anticipate the referent in the object condition. We propose two
possible answers to this question: On the one hand, linguistic functions are expressed by multiple acoustic cues distributed throughout the signal. Listeners might have picked up acoustic evidence indicating the contrastive nature of the referent before the pitch accent information had become available. On the other hand, within the microcosms of the experiment, listeners might have been able to anticipate the referent based on the absence of contradicting information. In other words, listeners did not hear a pitch accent on either the subject modifier nor the auxiliary, leading them to the conclusion that the object must be contrastive.

Overall, the observed patterns suggest that the integration of intonational information and discourse context facilitated reference resolution due to successful anticipation.

Table 1: Descriptive and inferential summary statistics for RT and AUC for each focus condition.

| Condition | RT |  |  | AUC |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | mean | est | SE | mean | est | SE |
| Broad | 1587 | 1578 | 40.7 | 0.39 | 0.39 | 0.02 |
| Object | 1466 | 1457 | 49.5 | 0.4 | 0.4 | 0.02 |
| Verb | 1367 | 1363 | 51.9 | 0.37 | 0.37 | 0.01 |
| Subjet | 1116 | 1121 | 93.8 | 0.31 | 0.31 | 0.03 |

Beyond these temporal operationalisations, results for area under the curve measurements (AUC) indicated that conditions differed in overall attraction of trajectories towards the competitor ( $\left.\chi^{2}(3)=12.4, \mathrm{p}=0.006\right)$ with the object condition exhibiting the greatest $\operatorname{AUC}(\beta=0.4, \mathrm{SE}=0.02)$, followed by the broad condition ( $\beta=0.39, \mathrm{SE}=0.02$ ), the verb condition ( $\beta=0.37, \mathrm{SE}=0.02$ ), and the subject condition ( $\beta=0.31$, $\mathrm{SE}=0.03$ ) (cf. Figure 7, Table 1). Not surprisingly, the earlier the relevant intonational cue in the acoustic signal, the less curvature towards the competitor was found. Importantly, albeit highly correlated, AUC and RT were not a direct mirror image of each other. While the broad condition is clearly the slowest, it was not the condition with the greatest AUC indicating that these measures reflect two different aspects of the response selection process: Overall latency and response competition.

## General discussion

The present study investigated intonation-based intention recognition in German using the mouse tracking paradigm. Listeners were exposed to different discourse contexts and different intonational patterns encoding the discourse status of referents. Analyses of continuous computer mouse movements during response selection suggest that listeners integrated intonational information rapidly as soon as it became available and anticipated potential referential interpretations early on. These insights are not new, of course. The present study merely replicates well-known results from studying oculomotor patterns with eye tracking (e.g. Dahan et al. 2002, Weber et al. 2006, Watson et al. 2006, Kurumada et al. 2014). Using the mouse tracking method, we showed that when listeners received discourse
relevant intonational information, their hand motions began to curve towards the target response before lexical disambiguation had taken place.

While the literature has mainly looked at intonational processing of rather clear mappings of intonational form and pragmatic interpretation, i.e. the presence vs. absence of a prominent pitch accent indicating contrastiveness, it remains to be seen, how these results generalise to scenarios in which listeners are exposed to more variable intonational information. Intonational categories have been shown to be characterised by a tremendous amount of variability (e.g. Grice et al., in press, inter alia), exhibiting no one-to-one mapping of form and function. Future research will need to answer the question as to how listeners accommodate to this degree of uncertainty in intonation-based intention recognition. The present study serves as a proof of concept that such questions can be conveniently studied using the mouse tracking paradigm (see also Tomlinson \& Bott 2013, and Warren 2017). We hope that our results spark more interest for this low-cost and pragmatically flexible experimental paradigm for research on speech perception within domains that go beyond phonemic and lexical processing. Mouse tracking proofs to be a fertile method to unravel the real-time dynamics of speech processing such as intonation-based intention recognition.

## References

Bates, D. Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1).
Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., \& Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. PloS ONE, 5(5), e10773.
Cangemi, F., Grice, M., \& Krüger, M. (2015). Listener-specific perception of speaker-specific production in intonation. In Fuchs, S., Pape, D., Petrone, C., \& Perrier, P. (eds.), Individual Differences in Speech Production and Perception (pp. 123-145), Frankfurt am Main: Peter Lang.
CursorSense (2016). (computer software, version 1.3.2). Plentycom Systems. Retrieved from http://plentycom.jp/en/cursorsense/download.php
Dahan, D., Tanenhaus, M.K., \& Chambers, C.G. (2002). Accent and reference resolution in spoken-language comprehension. Journal of Memory and Language, 47(2), 292314.

Dale, R., Kehoe, C., \& Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. Memory \& Cognition, 35(1), 15-28.
Féry, C., \& Kügler, F. (2008) Pitch accent scaling on given, new and focused constituents in German. Journal of Phonetics, 36, 680-703.
Freeman, J. B., \& Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behavior Research Methods, 42(1), 226-241

Grice, M., Niemann, H., Ritter, S., Roettger, T. B. (in press). Integrating the discreteness and continuity of intonational categories. Journal of Phonetics. https://doi.org/10.1016/j.wocn.2017.03.003
Kieslich, P. J., \& Henninger, F. (2016). Mousetrap: Mousetracking plugins for OpenSesame (Version 1.2.1). doi: 10.5281/zenodo. 163404

Kieslich, P. J., Wulff, D. U., Henninger, F., and Haslbeck, J. M. B. (2017). mousetrap: Process and Analyze MouseTracking Data. R package version 3.0.0. https://CRAN.Rproject.org/package=mousetrap
Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., \& Tanenhaus, M.K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. Cognition, 133(2), 335-342.
Kuznetsova, A., Brockhoff. P. B., \& Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-33. https://CRAN.R-project.org/package $=$ lmerTest
Ladd, D.R. (2008). Intonational phonology. Cambridge: CUP.
Magnuson, J.S. (2005). Moving hand reveals dynamics of thought. Proceedings of the National Academy of Sciences, 102, 9995-9996.
Mathôt, S., Schreij, D., \& Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. Behavior Research Methods, 44(2), 314-324.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Singmann, H., Bolker, B., Westfall, J., \& Aust, F. (2016). afex: Analysis of Factorial Experiments. R package version 0.16-1. https://CRAN.R-project.org/package=afex

Spivey, M.J., Grosjean, M., \& Knoblich, G. (2005). Continuous attraction toward phonological competitors. Proceedings of the National Academy of Sciences, 102, 1039310398.

Tomlinson, J. \& Bott, L. (2013). How intonation contrains pragmatic inference. In Proceedings of the 35th Annual Conference of the Cognitive Science Society, 3569-3575.
van de Vijver, R., \& Baer-Henney, D. (2014). Developing biases. Frontiers in psychology, 5, 1-8.
Warren, P. (2017). The interpretation of prosodic variability in the context of accompanying sociophonetic cues. Laboratory Phonology: Journal of the Association for Laboratory Phonology, 8(1), 11.
Watson, D., Tanenhaus, M.K., \& Gunlogson, C. (2008). Interpreting pitch accents in on-line comprehension: $\mathrm{H}^{*}$ vs L+H*. Cognitive Science, 32, 1232-1244.
Weber, A., Braun, B., \& Crocker, M.W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. Language and Speech, 49, 367-392.

# Beyond candidate inferences: People treat analogies as probabilistic truths 

Brad Rogers (bwrogers@indiana.edu)<br>David Landy (dlandy@indiana.edu)<br>Department of Psychological and Brain Sciences, 1101 East 10th Street<br>Bloomington, IN 47405 USA


#### Abstract

People use analogies for many cognitive purposes such as building mental models, making inspired guesses, and extracting relational structure. Here we examine whether and how analogies may have more direct influence on knowledge: Do people treat analogies as probabilistically true explanations for uncertain propositions?

We report an experiment that explores how a suggested analogy can influence people's confidence in inferences. Participants made predictions while simultaneously evaluating a suggested analogy and observed evidence. In two conditions, the evidence is either consistent with or in conflict with propositions based on the suggested analogy. We analyze the responses statistically and in a psychologically plausible Bayesian network model. We find that analogies are used for more than just generating candidate inferences. They act as probabilistic truths that affect the integration of evidence and confidence in both the target and source domains. People readily treat analogies not as a one-way projection from source to target, but as a mutually informative connection.


Keywords: Analogy, Bayesian Network, Computation, Confidence, Explanation, Inference, Reasoning

## Introduction

A teacher proposes to her class that atmospheric carbon concentration is like the water level in a bathtub (Kunzig, 2009). This science classroom analogy suggests many possible inferences about the atmosphere (and maybe bathtubs) that students can test by collecting evidence. Perhaps the atmospheric carbon level rises or falls based on the difference between carbon "faucet" and "drain" rates. Or maybe once carbon levels hit the upper limit, carbon dioxide will spill over into outer space. These new inferences might be true, or not. But what about the analogy itself? Is it the sort of thing that can be true or false? Does it depend on the inference? If it can be true, what kind of evidence would support it?

Analogy is used in a wide range of uncertain contexts such as contentious negotiation (Loewenstein et al., 1999), ambiguous accounting determinations (Magro \& Nutter, 2012), scientific discovery (Gentner, 2002; Hesse, 1966), thinking about astronomical distances (Resnick et al., 2012), and war declaration decisions (Khong, 1992). We use analogies when knowledge is scarce. But does analogy act like other explanations? Can we combine analogy with observed evidence? Like explanations, do we believe in them more when they successfully predict or explain our observations? Currently there is no account for how we
integrate analogy and observed evidence when grappling to understand uncertain situations. Even more, there is no psychological account that explicitly affords epistemic value to analogy.

In this paper we examine and affirm the hypothesis that people treat analogies as probabilistic truths. Analogies can be treated as true or false, and people integrate analogies with evidence much like they do for causal explanations.

## Candidate Inferences Hypothesis

Analogy is often called "the weakest form of evidence." Indeed, one account is that an analogy does not provide any evidence at all to favor its suggested candidate inferences since the act of constructing an analogy does not involve collecting new observations. Proponents of this account suggest that an analogy might render propositions more plausible, but not more probable (Bartha, 2010). Popular theories of analogical inference (Doumas et al., 2008; Gentner, 1989; Hummel \& Holyoak, 1997; Lu et al., 2012) largely invoke this candidate inferences account: ${ }^{1}$

Analogical reasoning in uncertain contexts begins with a well-described source domain and a target domain that requires an inference. A speculative analogy is made from the source to the target, which establishes a structural map between the two situations. With some luck, the source domain might contain useful correspondences to unknown elements of the target, producing candidate inferences that can only be validated by observed evidence in the target domain.

Some computational models treat analogy as a weighted score (e.g., ACME, SME), but this score is typically taken to reflect coherence (Holyoak \& Thagard, 1989; Thagard, 1989) or structural consistency (Gentner, 1983) and has not been extended to estimate the truth or rationality of the analogy. The correspondence identified by the analogy is not something that could be true or false. Rather, it is considered an artifact of our thinking about possible target inferences that should only guide our pursuit of evidence.

[^461]
## Analogy as Evidence Hypothesis

In this paper, we explore a stronger account of analogy with an expanded epistemic role. On this account, analogies not only introduce plausible inferences; they create a probabilistic connection between source and target that establishes and conveys inferential confidence.

The basic intuition is suggested by Peirce's notion of abduction (1935) and Harman's notion of inference to the best explanation (1965). Inference to the best suggests that people have confidence in the explanations that make their observations the least surprising. When an analogy suggests an inference in the target domain, observed evidence for that inference should increase the likelihood of the analogy itself when the analogy is taken as an explanation for the target inference. Conversely, if the target inference turns out to be false, the analogy becomes suspect. Our account builds on this insight to propose that people treat analogies and source knowledge as raising the conditional probability of target inferences.

Some readers may find it easy to consider that analogies act as a kind of theory whose truth can be supported or refuted by evidence. Indeed, some philosophical investigations have proposed statistical bases for analogical rationality (Harrod, 1956; Mill, 1882), and a recent study has found that people are sensitive to these rational statistics (Rogers \& Landy, 2016). But this epistemological view of analogy has not been dominant in the literature. Still, we are only interested here in the psychological question of whether people treat analogies as a probabilistic truth, rather than the normative question of whether they ought to.

## Approach

We conducted an experiment that asked participants to rate their confidence in competing explanations in two domains that may or may not be related. We provided observed evidence in one domain whose coherence with competing explanations was manipulated across subjects. A statistical analysis estimates primary effects to determine whether observed evidence influences reported confidence in the analogy. A Bayesian network model was used to compare responses with a psychologically plausible instantiation of analogy that integrates with observed evidence.

## Experiment

Participants were presented a fictional narrative situation describing two novel scientific phenomena, including simple visual representations. ${ }^{2}$ Mutually exclusive

[^462]explanations are provided for the phenomena. A suggested correspondence between the phenomena is described as leading scientists to develop an explanation and experiment. After receiving the stimulus and a test condition statement, participants rate their confidence in each of the explanations and the analogy.

For 20 years, biological scientists have fought over the relation between FCS, exachrome, and nuwen in the human hippocampus. Some scientists believe that exachrome is produced in cell nuclei, and that exachrome drives up production of FCS. They think that nuwen doesn't matter for FCS production. The more exachrome, the more FCS. Other scientists argue that it is nuwen produced in the cell nuclei that drives production of FCS, and that exachrome is an irrelevant byproduct. Both of these production pathways (the exachrome pathway and the nuwen pathway) are chemically plausible; which is correct is a matter of current debate. It is quite unlikely that both are correct. The following image summarizes the debate:


Rat hippocampi are much less well understood than human hippocampi. Rat hippocampi do exhibit FCS, but they do not contain exachrome or nuwen. Recently, one scientist (who was not attached to either of the other groups mentioned before) has suggested that FCS might be produced in rats in a way similar to that of humans. She identified two chemicals--called endochrome and oowen--that are similar to exachrome and nuwen, and that are produced in the rat hippocampus.

In other non-biological contexts, nuwen is sometimes used to predict properties of oowen, and exachrome is sometimes used to predict properties of endochrome. The following image summarizes the possibilities suggested by this scientist:


Rat Hippocampus

Recently, on the basis of the suggested links between exachrome and FCS in humans, and between human and rat FCS production, the scientist and her colleagues tested a novel hypothesis using rat hippocampi. The scientists injected the rats with a hormone that stimulates the production of endochrome. Several days later, they examined the level of FCS in the rat brain, predicting that it would show an increase.

Figure 1. Stimulus narrative presented to all participants

## Participants

We recruited $\mathrm{N}=300$ adults living in the US from Amazon's Mechanical Turk where participants can volunteer to complete short studies and other tasks in return for 35 cents.

## Design

Each participant was presented the same narrative (Fig. 1). Additionally, they were presented a single statement regarding the outcome of the scientific experiment implemented on the rat brain. This statement varied between subjects for three balanced conditions:

- As it turned out, increasing endochrome led to a large increase in the level of FCS in the rat hippocampus. (Confirm)
- The experiment results haven't been released yet, so we don't know how it worked out. (Neutral)
- As it turned out, increasing endochrome did not increase the level of FCS in the rat hippocampus at all. (Disconfirm)


## Procedure

Following presentation of the narrative and conditional statement, participants were asked to indicate their confidence for each explanation using a 7-point Likert scale:

- Exachrome causes the production of FCS in human hippocampus.
- Nuwen causes the production of FCS in human hippocampus.
- Endochrome causes the production of FCS in rat hippocampus.
- Oowen causes the production of FCS in rat hippocampus.

They were also asked to indicate their confidence that the situations are analogous:

- The production of FCS in rat hippocampus works similarly to that of human hippocampus.

Two balanced question orders were used. No response differences were observed on the basis of question order, so the factor was removed from subsequent analysis. All conditions contained a simple attention check. About $1 / 4$ of participants failed the attention check and were removed from the analysis.

## Statistical Analysis \& Results

We analyzed the participant responses by regressing each response item against the between-subject condition statements with each condition coded as a dummy variable. Since the assumptions violated linearity, we used resampling with 10,000 replications to evaluate statistical significance. For comparison, we also calculated Cohen's d to corroborate the significance of the observed effect sizes.

As expected, participant confidence in this explanation increased for the confirmation condition and decreased for the disconfirmation condition ( $\mathrm{p}<0.0001, \mathrm{~d}=2.5$ ). Since the procedure asserted that the two explanations about the rat hippocampus were unlikely to be simultaneously true, we predicted that the competing explanation would follow the opposite pattern. Indeed, when Endochrome $\Rightarrow$ FCS was
supported, confidence ratings for the Oowen $\Rightarrow$ FCS explanation decreased ( $\mathrm{p}<0.0001, \mathrm{~d}=2.5$ ).

Participants confidence ratings in the source domain explanations were also influenced by the observed evidence in the target domain. Confidence in the corresponding source explanations about the human hippocampus changed in the direction consistent with the correspondence structure of the analogy. For the Endochrome $\Rightarrow$ FCS explanation confidence increased with positive evidence and decreased with negative predictions ( $\mathrm{p}<0.0001, \mathrm{~d}=0.75$ ). Confidence in the competing source explanation Nuwen $\Rightarrow$ FCS was inversely affected ( $p<0.0001, d=-0.80$ ). Finally, successful predictions made participants more confident in the idea that the two domains were analogous ( $\mathrm{p}<0.0001, \mathrm{~d}=0.50$ ).

Participant responses strongly supported our hypothesis people treat the analogy as evidence for the inferences is suggests. New successful predictions made on the basis of a mapping from the source to the target increased confidence in the commonality of the domains, as well as in the untested scientific explanation that generated them.

## Bayesian Network Model \& Results

We further analyzed the data using a Bayesian network model (Pearl, 2009) to estimate the influence of the suggested analogy on the response item confidence statements in a way constrained by a plausible causal structure. In the model, each causal explanation is represented as a single node and assigned a prior baseline probability. Since it was stated in the stimulus that the two explanations within a domain were unlikely to be simultaneously true, the model places a negative correlation between the explanations. Without an analogy, the source and target domains (i.e., human and rat hippocampus, respectively) have no causal linkage. On the other hand, if there is a known analogy that is taken as certain, strong causal linkages are present from the source to the target domain.


Figure 2. Bayesian network structure without analogy and with certain analogy

With an uncertain analogy, though, the structure itself becomes probabilistic. To capture this, we take the model a step further by representing the analogy itself as a single node. In this way, we can gauge the evidentiary influence of
the analogy and participants' confidence in it using their confidence ratings. If the domains were sufficiently complex that multiple mappings were possible, it might be necessary to include structural evaluations in the model such as rankings from a model of structural correspondence (Landy \& Hummel, 2010). But in this case, the mapping from source situation to target situation is plainly obvious and can be treated as a single node.

Although the distinction is often drawn between superficial and deep analogies, how people consider the truth of an analogy has not been investigated to the best of our knowledge. As a starting place, the analogy was modeled as a Boolean variable-true or false. Participant confidence in the analogy was estimated by a Beta-distribution..

Now the probability of a target domain explanation prior to observing the experimental results depends on both the probability of the truth of the source domain explanation and the probability of the truth of the analogy. If the analogy is true, then what is true or false in the source domain is also true or false in the target domain. However, if the analogy is false, the truth of the target explanation is independent of the source domain knowledge. In other words, an analogy guarantees correspondence, but a failed analogy does not guarantee non-correspondence. This approach effectively introduces a probabilistic switch between the no analogy and certain analogy network structures.

The prior probabilities of the Bayesian network were fit without including the evidence obtained by the experimental results (i.e., the test condition statement). So each individual is taken to have an estimate of the prior probability of each of the source explanations, the analogy, and the target explanations. The prior probabilities provide an associated estimate of participant confidence that the experimental results will be confirmed or disconfirmed.


Figure 3. Bayesian network structure with uncertain analogy and evidence from experimental results (i.e., test condition)

The participant data was fit using a hierarchical model. Participants were assumed to have been randomly selected from a population having a single distribution of subjective priors for each node. The priors were estimated using Dirichlet distributions for the domain explanation
probabilities ${ }^{3}$ and Beta distributions for the analogy and the evidence probabilities. The model had 14 population-level free parameters, fit to 1500 participant responses. Participants from the neutral condition were assumed to respond based on these prior parameter estimates without any additional evidence. Participants from the evidence condition were modeled by updating the Bayes net given the appropriate evidentiary outcome, and these posterior estimates were fit to the responses.


Figure 4. Posterior predictive distributions by condition for each explanation compared with response distributions

We solicited confidence ratings using a Likert scale rather than explicit probability estimates. So a final step in the model was to translate posterior probabilities from the

[^463]Bayesian network into Likert response values. We treated Likert values as ordered and evenly distributed from 0 to 1. Responses were then treated as beta-distributed among these values, with mean at the subjective probability. This allowed variance from the specific posterior subjective probabilities, minimized degrees of freedom in the model, and afforded a limited flexibility in translating posterior probabilities into Likert scale responses. The model was fit in Stan via R: 1,000 posterior samples proved sufficient for model convergence with population-level $\widehat{R}$ values all less than 1.1 (Bates et al., 2015).

Figure 4 indicates posterior predictions of each Bayesian network node overlay with the fit participant Likert responses for each condition. The major patterns in the data were generally well-captured by the model, suggesting that people were integrating evidence from the target prediction success into their confidence in the analogy, and were doing so in a manner that approaches rational behavior. Predictions matched the direction of the observed effects for all five model nodes. If the analogy were rejected by participants, we would expect no differences between conditions in the responses about the analogy and about the source explanations.

|  | Participant <br> (Likert fit) |  | Model Predicted 95\% HPI <br> Lower bound |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  |  | 0.14 |  | 0.13 |  |
| Upper bound |  |  |  |  |  |

Figure 5. Difference between confirm and disconfirm participant confidence ratings versus model predictions

The model fit can be evaluated by comparing differences between the distribution of participant responses and the simulated posterior predictions of the population (i.e., 1,000 samples of the posterior for each of 300 participants). Although the model matched the direction of the empirical results in every case, the outcome of this analysis revealed a systematic bias (discussed later) that could not be accounted for by this computational approach.

## Discussion

What does it mean to be confident in an analogy? What does it mean for an analogy to be assigned a probability value at all? This is an important open question. Analogies are rarely exact correspondences. Useful analogies are sometimes even known from the outset to be poor, such as "atoms are like solar systems." Alternatively, models and simulations in the social sciences are often presented as valid simplifications of complex phenomena. It seems, then, that we can be confident in an analogy's validity even when we do not believe the correspondence to be exact. This paper takes a first step toward answering these open questions by establishing a basic fact: people do treat analogies as probabilistic truths and integrate them with evidence.

## Implications for analogical inference

If analogies just generate candidate inferences, then people's confidence in explanations in one domain would be unaffected by observations in another. In contrast, we found that analogical mappings do raise posterior estimates of the likelihood of candidate inferences. Moreover, when uncertainty exists in the source domain knowledge, confirmed analogical inference in the target domain raises confidence in the corresponding source knowledge. This effect suggests that people treat analogies not as a one-way projection from source to target, but mutually informative.

Results show that the effective confidence of the analogy itself is influenced by the success of its inferences suggesting that people evaluate the analogy on more than its degree of structural correspondence. The analogy seems to have a causal property that can be integrated with and influenced by observed evidence. To that point, no evidence was ever presented in the source domain that could arbitrate between the proposed explanations, so evidence confirming a target domain inference could not possibly strengthen the structural correspondence between the domains. And yet, if the new information confirmed inferences made by the analogy, differences by condition in participant confidence ratings suggest they credited the analogy for the success.

It is worth noting that while the candidate inferences account is implied in many extant studies of analogy, the authors of those studies may not wish to explicitly commit to it. For the most part, we believe that the role of evidence in influencing the value of the analogy has been deferred rather than denied. We see these results as extending rather than negating extant approaches.

## Deviations from rationality

Although the observed confidence differences are quantitatively close to the model predictions, the observed differences are not completely compatible with rational allocation of probabilities under the assumption that analogies act as evidence for their inferences. Participants attributed success or failure of the analogy more to the veracity of the source explanations and less to the analogy than would be expected by the model structure. In other words, we expected confidence in the analogy to justify shifts in confidence in the source domain explanations. But the observed shifts in the source domain outpaced participant reported confidence in the analogy. One possible explanation is that participants may have interpreted the analogical statement more broadly than intended, so that the possibility of any related dissimilarity would reduce their confidence in the analogical statement. Another possibility is that people use different cognitive processes to rate confidence in analogical statements than they use to rate domain-specific statements. If true, then it may be necessary to apply a simple transformation to reported analogical confidence when modeling analogy in a Bayesian network.

## Limitations of the present study

One limitation of the experiment is that the relationship between mammal brains is not only a near analogy, it is also a biological homology. Rats and humans evolved from a common ancestor, so similarities between them may reflect properties of their ancestor rather than attribution of evidence to the analogy per se. Indeed, scientists regularly use animal models to predict properties of human beings on this basis. Because the inference of the experiment may have a biological explanation, shifts in confidence may reflect an alternate process of inference about the cause rather than about the analogy. In subsequent experiments using more distant domains-such as suggesting a link between ion behavior in "super-cooled glass" and macroeconomic decisions by nations-we find consistent, but less pronounced effects to those presented here.

## Future study

Even though we can conclude that people are willing to treat an uncertain analogy effectively as a probabilistic truth, it is not clear what cognitive processes underlie this effect. Two alternate hypotheses are:

1. People may treat the analogy as a kind of theory whose truth can only be supported by evidence in the source and target domains. This is the most straightforward interpretation of the experiment and the approach taken by the ERIC model of explanatory reasoning under uncertainty (Landy \& Hummel, 2010).
2. Success of an inference may imply a stronger structural correspondence than is actually observed. Confirming evidence for an inference in one domain may improve an implied estimate of unobserved, but still predictive, structural correspondence (Rogers \& Landy, 2016).

More investigation is needed to distinguish between these possibilities. We still await a fully integrated account of reasoning across correspondences among structures about which people have probabilistic beliefs.

Acknowledgment - This material is based upon support of BR by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1342962.

## References

Bartha, P. (2010). By Parallel Reasoning. Oxford University Press. Bates, D., Mächler, M., Bolker, B., \& Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. Journal of Statistics Software, 67(1).
Doumas, L. A. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological Review, 115(1).
Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In J. Shrager \& P. Langley (Eds.), Computational Models of Scientific Discovery and Theory Formation (pp. 157-196). Morgan Kaufmann Publishers, Inc.
Gelman, A., Carlin, J. B., Stern, H. S., \& Rubin, D. B. (2013). Bayesian data analysis. CRC Press.
Gentner, D. (1983). Structure-Mapping: A Theoretical Framework
for Analogy. Cognitive Science, 7(2).
Gentner, D. (1989). The mechanisms of analogical learning. In Similarity and Analogical Reasoning. Cambridge University Press.
Gentner, D. (2002). Analogy in Scientific Discovery: The Case of Johannes Kepler. In Model-Based Reasoning (pp. 21-39). Springer US.
Gentner, D., \& Colhoun, J. (2010). Analogical Processes in Human Thinking and Learning. In Towards a Theory of Thinking (pp. 35-48). Springer Berlin Heidelberg.
Gentner, D., \& Markman, A. (1997). Structure Mapping in Analogy and Similarity. American Psychologist, 52(1).
Harman, G. H. (1965). The Inference to the Best Explanation. The Philosophical Review, 74(1).
Harrod, R. (1956). Foundations of Inductive Logic. Harcourt, Brace \& Co.
Hesse, M. (1966). Models and Analogies in Science. Notre Dame Press.
Holyoak, K. J., \& Hummel, J. E. (2000). The Proper Treatment of Symbols in a Connectionist Architecture. In Cognitive dynamics: Conceptual change in humans and machines. MIT Press.
Holyoak, K. J., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive Science: A Multidisciplinary Journal, 13.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104(3).
Khong, Y. F. (1992). Analogies at War. Princeton University Press.
Kunzig, R. (2009). The Carbon Bathtub. National Geographic.
Landy, D., \& Hummel, J. E. (2010). Explanatory Reasoning for Inductive Confidence. In Proceedings of the 32nd Annual Conference of the Cognitive Science Society.
Lee, H. S., \& Holyoak, K. J. (2008). The role of causal models in analogical inference. Journal of Experimental Psychology. Learning, Memory, and Cognition, 34(5).
Loewenstein, J., Thompson, L., \& Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. Psychonomic Bulletin \& Review, 6(4).
Lu, H., Chen, D., \& Holyoak, K. J. (2012). Bayesian analogy with relational transformations. Psychological Review, 119.
Magro, A. M., \& Nutter, S. E. (2012). Evaluating the Strength of Evidence: How Experience Affects the Use of Analogical Reasoning and Configural Information Processing in Tax. The Accounting Review, 87(1).
Mill, J. S. (1882). A System of Logic: Ratiocinative and Inductive.
Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.
Peirce, C. S. (1935). Collected Papers of Charles Sanders Peirce, vols 1-6. (C. Hartshorne \& P. Weiss, Eds.). Harvard University Press.
Resnick, I., Shipley, T. F., Newcombe, N., Massey, C., \& Wills, T. (2012). Examining the Representation and Understanding of Large Magnitudes Using the Hierarchical Alignment model of Analogical Reasoning. In Proceedings of the 34th Annual Conference of the Cognitive Science Society.
Rogers, B., \& Landy, D. (2016). Investigating Rational Analogy in the Spirit of John Stuart Mill: Bayesian Analysis of Confidence about Inferences across Aligned Simple Systems. In Proceedings of the 38th Annual Conference of the Cognitive Science Society.
Thagard, P. (1989). Explanatory coherence. Behavioral and Brain Sciences, 12.

# A Model-based Approach for Assessing Attentional Biases in People with Depressive Symptoms 

Isa Rutten (isa.rutten@kuleuven.be)<br>Quantitative Psychology and Individual Differences Research Group, Tiensestraat 102, University of Leuven, 3000 Leuven, Belgium (" same affiliation for 2 other authors)<br>Wouter Voorspoels (wouter.voorspoels@kuleuven.be) Laboratory of Experimental Psychology, (")<br>Ernst H.W. Koster (ernst.koster@ugent.be)<br>Department of Experimental Clinical and Health Psychology, Henri Dunantlaan 2, Ghent University, B-9000 Ghent, Belgium<br>Wolf Vanpaemel (wolf.vanpaemel@kuleuven.be)<br>Quantitative Psychology and Individual Differences Research Group, (")


#### Abstract

Biased attention is assumed to play an important role in the etiology and maintenance of depression and depressive symptoms. In this paper, we used data from a categorization task and an associated model to assess the attentional bias of people with varying levels of depressive symptoms. Attentional bias was operationalized as the parameter estimate in a prototype model of categorization. For estimation, we used a Bayesian hierarchical mixture approach. We expected to find a positive correlation between depressive symptoms and an $A B$ for negative material and a negative correlation between depressive symptoms and a bias toward positive material. Despite good model fit, Bayesian regression analyses revealed weak or moderate evidence in favor of the null model assuming no association between attentional preferences and depressive symptoms, both for negative and positive material.


Keywords: psychology; cognitive science ; attention; concepts and categories; Bayesian modeling; mood disorder

## Introduction

Biased attention takes an important position in cognitive theories explaining the etiology and maintenance of psychological disorders. In particular, depression has been theorized to be linked to biased attention for negative information (Beck, 1976). Multiple methods have been developed to assess attentional biases (AB). The most popular approaches include reaction time assessments, in which response latencies for negative, positive, and neutral material are compared, and translated into AB indices. Eye tracking techniques, comparing fixation durations on negative, positive, and neutral stimuli, are also a popular approach. Caution is recommended however, in both approaches. When relying on reaction time assessments, one has to consider the ambiguity and lack of reliability often associated with response latencies (Rodebaugh et al., 2016). Eye tracking techniques appear to yield good reliability estimates, but only if certain conditions are met (Rodebaugh
et al., 2016). Moreover, as these techniques focus exclusively on overt attentional processes, they are less informative with regard to attentional resources that are allocated covertly, without saccadic eye movements. Perhaps not surprisingly, results obtained with these existing approaches are not consistent (Peckham, McHugh, \& Otto, 2010).

The goal of this paper is to explore novel methodologies for assessing attentional biases associated with depressive symptoms. In particular, inspired by Viken, Treat, Nosofsky, McFall, and Palmeri (2002), we tested the applicability of a categorization approach to assess AB's in the context of depression (see also Kruschke \& Vanpaemel, 2015). Participants were presented pictures of human faces varying in emotional expression (emotional stimulus dimension) and hair color (neutral stimulus dimension). They were asked to classify these stimuli in two different categories, each represented by a prototype stimulus of that category. The two prototypes reflected extreme levels of the stimulus dimensions and were each other's opposites. For example, prototype A had a light hair color, and a very sad facial expression, versus prototype B with a dark hair color and a slightly sad facial expression. In this way, participants could either choose to focus on the hair color (or neutral) dimension or on the facial expression (or emotional) dimension to classify the pictures. These data, taking the form of classification counts of each category, per stimulus, were then used to estimate participant-specific parameters in a prototype model (e.g., Nosofsky, 1987). One of these parameters corresponds to the attentional weight (AW) for one stimulus dimension, reflecting the relative attentional preference for that stimulus dimensions compared to the other stimulus dimension.

We expected to find a positive correlation between depressive symptoms and an AB for negative material and a negative correlation between depressive symptoms and a bias toward positive material (Peckham et al., 2010). In addition
to depressive symptoms, we assessed anxiety symptoms, as depression and anxiety are known for their comorbidity, and our aim was to isolate the relation of $A B$ to depressive symptoms. Brooding or depressive rumination, and a negative mood, were also investigated in this study, since both variables have already been related to an attentional bias for negative material (Bradley, Mogg, \& Lee, 1997; Koster, De Lissnyder, Derakshan, \& De Raedt, 2011). Finally, the personality trait, neuroticism was assessed, in order to explore whether this important risk factor for depression, could be related to an AB for negative information.

## Method

Every participant received two versions of a classification task, each version was administered in two (withinparticipant) conditions. In all versions and conditions, participants were asked to classify facial stimuli according to two prototype stimuli. The stimuli were made up of only two different stimulus dimensions: brightness of hair color and intensity of emotional expression. In the happy condition, the emotional expression ranged from a slightly to a very happy facial expression, whereas in the sad condition, it ranged from slightly to very sad.

We report all data exclusions, all included questionnaires or measures, and all study conditions.

## Participants

A total of 309 first-year psychology students participated in this study in exchange for course credits ( 262 women, mean age $=18.53, \mathrm{SD}=1.90$, with a range from 17 to 39 ). The sample size was determined by the number of participants showing up during the two weeks of data collection, available for all first-year students of the Psychology department of the University of Leuven (Belgium).

## Materials

Self-report Measures The Center for Epidemiologic Studies - Depression Scale (CES-D; Radloff, 1977) was used to assess depressive symptoms (score: $0-60$ ). The Hospital Anxiety and Depression Scale (HADS; Spinhoven et al., 1997), was included to assess comorbid depression (score: 0 - 21) and anxiety symptoms (score: 0-21). We also included the Rumination Response Scale (RRS; Nolen-Hoeksema, \& Morrow, 1991) to assess brooding (score: 5-20), and the Ten Item Personality Inventory (TIPI; Gosling, Rentfrow, Swann, 2003) to assess neuroticism (score: 1-7). Finally, the current mood of participants was measured with a 5 -point Likert scale ("How do you feel at this moment? 1 (very unhappy) 2 (slightly unhappy) -3 (neutral) -4 (slightly happy) -5 (very happy"). Dutch versions of these questionnaires were administered.

Stimuli The stimuli were adopted from the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, \&

Öhman, 1998). Applying Fotomorph 13.9 (Softland SRL) and GIMP (2008), the pictures were adjusted in such way they only systematically differed from each other on two dimensions: The intensity of facial expression and the brightness of the hair color. On the basis of the modified stimuli, a negative and positive stimulus set were created. In the sad condition, the intensity of the facial expression ranged between very sad and slightly sad, whereas in the happy condition, the intensity ranged from very happy to slightly happy. By creating five different levels of emotional expression (mild intensity - strong intensity) and five different levels of brightness of hair color (light - dark) and combining all possible levels, we obtained a stimulus set consisting of 25 different pictures for each condition. Prototypes were extreme on both dimensions (for example, a stimulus having a very sad emotional expression of level 5 , and very dark hair color of level 5). Figure 1 presents example stimuli.

Task In both conditions, participants were asked to classify all pictures of the condition-specific stimulus sets into category A or B. Category A was represented by one of the four prototypes, and category B by the inverse prototype, within the same emotional condition. For example, in the sad condition, if prototype A had level 5 of sadness (very sad) and level 5 of hair color (very dark), then prototype B had level 1 of sadness (slightly sad), and level 1 of hair color (very light). The prototypes stayed on participants' computer screen during the entire task. In each trial, participants had to classify one picture into one of the two categories. No feedback was provided, so participants could freely choose how to classify the stimuli. In making the classification decisions, participants thus could either base their classifications on the pictures' hair color or facial expression, reflecting their attentional focus on these dimensions.


Figure 1. The two upper pictures represent two possible prototypes. The stimuli at the bottom are random examples of the negative stimulus set, to be classified in category A or B. Adapted from "The Karolinska Directed Emotional Faces" - KDEF", by D. Lundqvist et al., 1998, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9. Copyright 2015 by D. Lundqvist.

For each condition, there were two versions of the task, differing in the prototype pairs used. Version 1 consisted of prototype A1, with level 5 of sadness, and level 5 of hair color, and prototype B1, with level 1 of sadness, and level 1 of hair color. Version 2 consisted of prototype A2 with level 1 of sadness, and level 5 of hair color, and prototype B2 with level 5 of sadness and level 1 of hair color.

## Procedure

Each participant ran through both conditions, and in each condition, they performed two versions of the same task. The order of the conditions, and task versions within the conditions, was counterbalanced between participants.

In each task, participants categorized two blocks in which all 23 non-prototype stimuli were presented in a random order. Thus, within each task version, each stimulus was classified twice. After completing the categorization tasks, participants were asked to indicate their current mood state, and to fill out the CES-D, HADS, RRS, and TIPI.

## Model

The categorization data were analyzed using Nosofsky's (1987) weighted prototype classification model:

$$
\mathrm{P}(\mathrm{~A} \mid \mathrm{i})=\frac{\eta_{\mathrm{iA}}}{\eta_{\mathrm{iA}}+\eta_{\mathrm{iB}}}
$$

The model assumes that the probability of classifying stimulus $i$ in category $A, P(A \mid i)$, is driven by the perceived similarity between stimulus $i$ and prototype $A, \eta_{i A}$, divided by the overall perceived similarity between stimulus i and prototype A and prototype B. The perceived similarity between stimulus $i$ and prototype $A$ is assumed to be an exponential decay function determined by a sensitivity parameter $c$, and the weighted distance between the stimulus and the prototype, $d_{i A}$ :

$$
\eta_{i A}=e^{-c d_{i A}}
$$

The sensitivity parameter $c$ reflects how clearly the stimuli could be discriminated from each other. The weighted distance between the stimuli and prototype $A, d_{i A}$, was computed as follows:

$$
d_{i A}=w_{a}\left|x_{i a}-x_{A a}\right|+\left(1-w_{a}\right)\left|x_{i h}-x_{A h}\right|
$$

where $x_{i a}$ represents the coordinate of stimulus $i$ on dimension affect, and $x_{i n}$ the coordinate of stimulus $i$ on dimension hair, in psychological space. $x_{A a}$ is the coordinate of prototype $A$ on dimension affect, and $x_{A h}$ the coordinate of prototype $A$ on dimension hair. $w_{a}$ is the AW for the affect dimension, and $1-w_{a}$ is the AW for the hair dimension.

The coordinates of the stimuli and prototypes were obtained in a separate study in which a different group of 32 participants rated the two dimensions of all stimuli on a 10-
point Likert scale, ranging from 0 (light hair color/mild facial expression) to 9 (dark hair color/ intense facial expression). For both the sad and happy conditions, the split-half reliabilities of respectively the affect and hair color dimensions were .96 and .99 . The coordinates of the stimuli were calculated by taking the mean rating on each dimension for every stimulus.

The prototype model was extended hierarchically and with a mixture component in a Bayesian framework (see e.g., Bartlema, Lee, Wetzels, \& Vanpaemel, 2014). The hierarchical extension was used to accommodate continuous, qualitative individual differences, and to shrink extreme values, whereas the mixture component was included to accommodate discrete, quantitative individual differences. In particular, it allowed differentiating between three groups of participants: The first group consists of people whose behavior was captured better by a guessing model, which assumed that the response probability for each stimulus was .5. Identifying these participants avoids contamination of our parameter estimates by participants for whom the prototype model was not sufficiently appropriate, that is, participants who appeared to be guessing (see, e.g., Voorspoels, Rutten, Bartlema, Tuerlinckx \& Vanpaemel, in press, and Zeigenfuse \& Lee, 2010 for a similar approach).

Among the participants assigned to the prototype group, we allowed two subgroups: the 'affect group' and the 'neutral/hair group' with the 'affect group' having a higher group-level AW for the affect dimension, as compared to the 'hair group'. In particular, we restricted the mean AW for the affect dimension in the 'affect group' to be higher than the mean AW for affect in the 'hair group'.

## Results

## Model Analyses

The model was implemented in JAGS (Plummer, 2011). We ran 3 chains with 36000 iterations each, after discarding 4000 iterations for burn in. We performed separate analyses for the sad and happy conditions. The data from the two task versions (different prototypes) within each condition were jointly modelled in order to obtain a single AW in each condition. The model analyses identified two clearly distinguishable groups (a group focusing on affect and a group focusing on hair color) in both conditions.

In the sad condition, the 'affect group' contained 141 participants, whose AW was larger for the affect dimension (the group-level posterior had a mean $w_{-}$sadness $=.88$ ), compared to the 'neutral/hair group', containing 163 participants (mean $w_{-}$sadness $=.21$ ).

In the happy condition, the 'affect group' contained (again) 141 participants, whose AW was larger for the affect dimension (the group-level mean attention to happiness was: $w_{-}$happiness $=.87$ ), compared to the 'neutral/hair group' (mean $w \_$happiness $=.19$ ), consisting of 167 participants.

Also, a number of participants were not distinguishable from guessers. In the sad condition, we identified five guessers, with one of them being also, the only, guesser in the happy condition.

To evaluate the fit of the model, we inspected the grouplevel posterior predictive for all categorization tasks. To give an idea of how well the model performs, Figure 2 presents the posterior predictive for one of the two task versions in the sad condition. Each panel shows a schematic representation of the stimulus space (five by five stimuli, represented by the squares), with the corner stimuli being prototypes. The panels depict the latent groups. In each square, the background color provides an indication of the model's posterior prediction for the corresponding stimulus and group, as a gradient between the top left prototype (orange) and lower right prototype (blue). The stronger the color matches the prototype, the more firmly the model predicts classification in the corresponding category. In each square, the circle color is an indication of the average observed classification count of the corresponding stimulus, across all participants in the group. Thus, matching colors between background and circle provide insight in the match between the model's posterior predictions and the observed data.


Figure 2. Posterior predictive check, see text for details.

Inspection of Figure 2 reveals that the model successfully captures the patterns within each latent group. Also, the predicted and observed stimulus groupings are sensible considering the AWs applied in each group, with the affect group categorizing the stimuli according to the affect dimension, and the hair group according to the hair color dimension.

## Regression Analyses

After excluding five participants who were assigned to the guessing group, multiple regression analyses were performed to investigate the relationship between attentional preferences, as operationalized by the individual-level estimates of $w_{a}$, and depressive symptoms, anxiety symptoms, brooding, neuroticism, and current mood. All variables showed sufficient variability (see Table 1 for the descriptive statistics). Variance inflation factors (VIF; Hair, Anderson, Tatham, Black, 1995) indicated only a low degree of multicollinearity in our model (largest VIF was 3.22, well below the threshold of 10 ).

Table 1: Descriptive statistics.

|  | SD | mean | range |
| :--- | :---: | :---: | :---: |
| CES-D | 10.06 | 17.83 | $0-52$ |
| HADS_D | 3.63 | 4.41 | $0-18$ |
| HADS_A | 3.97 | 7.07 | $0-20$ |
| TIPI_E | 1.41 | 4.36 | $1-7$ |
| RRS | 3.52 | 10.47 | $5-20$ |
| Mood | 0.60 | 3.11 | $1-5$ |
| w_sadness | 0.37 | 0.51 | $0.04-0.99$ |
| w_happiness | 0.37 | 0.49 | $0.05-0.98$ |

Note. w_sadness is $w_{a}$ in the sad condition, whereas $\mathrm{w}_{-}$happiness is $w_{a}$ in the happy condition.

The regression analyses were performed using the BayesFactor package (Morey \& Rouder, 2013). A BF compares the evidence for the null model with the evidence for the alternative model. Given the null model in the numerator, a $\mathrm{BF}>1$ indicates evidence in favor of the null model, whereas a $\mathrm{BF}<1$ indicates evidence in favor of the alternative model.

To test the effect of each predictor (e.g., depressive symptoms), we compared a restricted model containing all predictors, except for the predictor of interest (null model), against a full model containing all predictors (alternative model). As a sensitivity analysis, both medium and ultrawide scale factors were used to calculate the Bayes factors. As can be seen in Table 2, the Bayes factors showed evidence in favor of the null model for all predictors in both conditions (with the exception of current mood in the sad condition when a medium scale factor was used). The strength of the evidence depends on the exact choice of scale factor. When the scale factor is medium, the evidence in favor of the null
model is weak, and thus these results are best interpreted as being inconclusive. With increasing scale factor, the evidence in favor of the null grows stronger, but it is never very strong. Overall, these analyses suggest that no strong evidence for meaningful associations could be found between the symptom and traits scores and the AWs. The near zero partial correlation coefficients (using the ppcor package; Seongho, 2015), confirm this picture, as can be found in Table 2.

Table 2: Overview of the BFs (> 1 indicates support for the null model), and the corresponding partial correlation coefficients.

|  |  | 3 | sad | happy |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | BFm | BFu | PC | BFm | BFu | PC |
| CES-D | 1.89 | 3.23 | -.00 | 2.59 | 4.60 | .02 |
| HADS_D | 2.50 | 4.38 | .03 | 2.84 | 5.07 | .03 |
| HADS_A | 2.75 | 4.87 | .03 | 3.05 | 5.49 | .02 |
| TIPI_E | 2.82 | 4.99 | -.02 | 3.04 | 5.48 | -.02 |
| RRS | 2.32 | 4.05 | -.03 | 2.53 | 4.48 | .04 |
| mood | 0.87 | 1.39 | -.09 | 1.17 | 1.93 | -.08 |

Note. $\mathrm{BFm}=$ Bayes factor with medium r scale, $\mathrm{BFu}=$ Bayes factor with ultrawide r scale. $\mathrm{PC}=$ partial correlation coefficient.

## Discussion

Applying a categorization approach to the assessment of attentional biases in people with varying levels of depressive symptoms revealed weak to moderate evidence for the absence of an association between severity of depressive symptoms and an attentional bias to sadness or happiness. Similar BFs were observed for the other predictors of interest: anxiety symptoms, brooding, neuroticism, and current mood. In the light of the small Bayes factors, especially when using a medium scale factor, we cannot make strong statements about rejecting the alternative model, or accepting the null model.

These findings are in line with previous inconsistencies in results obtained with the frequentist approaches investigating attentional biases in the context of depression (Peckham et al., 2010). In some studies p values above .05, whereas in other studies, p values below .05 were found, without a clear explanation as to when to expect significant results and when not.

An important limitation of the current study that could explain the results, is the recruited sample. A high functioning student sample was recruited. Though we could observe a reasonable amount of variability in depression scores, the sample may have been too healthy to detect attentional biases related to depressive symptomatology. A next step is to apply this approach to data obtained in a sample of more severely depressed participants.

We believe the modelling approach demonstrated here, has a number of advantages that might prove useful in helping to
solve the elusiveness of attentional biases in the context of depression. First, attentional preferences were extracted from a model, that excluded people whose behavior could better be predicted by a guessing model, instead of the prototype model. This means that data resulting from random behavior were filtered out. Second, by considering a specific parameter in a model to conceptualize attentional biases, other processes that might influence participants' behavior in the task, such as people's discrimination abilities ( $c$ parameter), were factored out. Third, assessing attentional processes by considering their impact on categorization behavior could be quite insightful, given that classification decisions reflect the way in which people organize and structure their world. Attentional bias indices obtained by analyzing categorization behavior can give us an idea about how strongly attentional preferences influence the way in which people perceive and organize their environment.

## Acknowledgments

The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06).

## References

Bartlema, A., Lee, M., Wetzels, R., \& Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. Journal of Mathematical Psychology, 59, 132-150. http://dx.doi.org/10.1016/j.jmp.2013.12.002
Beck, A. T. (1976). Cognitive therapy and the emotional disorders. New York: International University Press.
Bradley, B. P., Mogg, K., \& Lee, S. C. (1997). Attentional biases for negative information in induced and naturally occurring dysphoria. Behaviour Research and Therapy, 35(10), 911-927.
http://dx.doi.org/10.1016/S0005-7967(97)00053-3
Fotomorph 13.9. Softland SRL. http://fotomorph.soft112.com/
GIMP, G. (2008). Image manipulation program. User Manual, Edge-Detect Filters, Sobel, The GIMP Documentation Team.
Gosling, S. D., Rentfrow, P. J., \& Swann, W. B. Jr., (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37, 504-528. doi:10.1016/S0092-6566(03)00046-1
Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. \& Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan.
Koster, E. H., De Lissnyder, E., Derakshan, N., \& De Raedt, R. (2011). Understanding depressive rumination from a cognitive science perspective: The impaired disengagement hypothesis. Clinical Psychology Review,

31(1),
http://dx.doi.org/10.1016/j.cpr.2010.08.005
Kruschke, J. K., \& Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In J. Busemeyer, J. Townsend, Z. J. Wang, \& A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology (pp. 279-299). Oxford: Oxford University Press.
Lundqvist, D., Flykt, A., \& Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.
Morey, R. D. \& Rouder, J. N. (2013). Bayes Factor: Computation of Bayes factors for common designs. R package version 0.9.4. Retrieved from http://CRAN.Rproject.org/package=BayesFactor
Nolen-Hoeksema, S., \& Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta Earthquake. Journal of Personality and Social Psychology, 61(1), 115121. http://dx.doi.org/10.1037/0022-3514.61.1.115

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(1), 87-108.
http://dx.doi.org/10.1037/0278-7393.13.1.87
Peckham, A. D., McHugh, R. K., \& Otto, M. W. (2010). A meta-analysis of the magnitude of biased attention in depression. Depression and Anxiety, 27(12), 1135-1142. doi: 10.1002/da. 20755
Plummer, M. (2011). JAGS Version 3.1. 0 user manual. International Agency for Research on Cancer, 2.
Radloff, L. S. (1977). Center for Epidemiological Studies Depression Scale. Applied Psychological Measurement 1, 385-401.
Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., ... \& Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. Journal of Abnormal Psychology, 125(6), 840-851.
http://dx.doi.org/10.1037/abn0000184
Seongho K. (2015). ppcor: Partial and semi-partial (part) correlation. R package version 1.1. https://CRAN.Rproject.org/package=ppcor
Spinhoven, P. H., Ormel, J., Sloekers, P. P. A., Kempen, G. I. J. M., Speckens, A. E. M., \& Van Hemert, A. M. (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. Psychological Medicine, 27(2), 363-370.
Viken, R. J., Treat, T. A., Nosofsky, R. M., McFall, R. M., \& Palmeri, T. J. (2002). Modeling individual differences in perceptual and attentional processes related to bulimic symptoms. Journal of Abnormal Psychology, 111(4), 598. http://dx.doi.org/10.1037/0021-843X.111.4.598

Voorspoels, W., Rutten, I., Bartlema, A., Tuerlinckx, F., \& Vanpaemel, W. (in press). Sensitivity to the prototype in children with high-functioning autism spectrum disorder: An example of Bayesian cognitive psychometrics. Psychonomic Bulletin \& Review.
Zeigenfuse, M.D., \& Lee, M.D. (2010). A general latentassignment approach for modeling psychological contaminants. Journal of Mathematical Psychology, 54, 352-362. doi: 10.1016/j.jmp.2010.04.001

# Is it fair? Textual effects on the salience of moral foundations 

Eyal Sagi (esagi@stfrancis.edu)<br>University of St. Francis<br>500 Wilcox St. Joliet, IL, 60435


#### Abstract

Many of the important decisions we make have moral implications. Moral Foundations Theory (Haidt \& Joseph, 2004) identifies 5 distinct styles of moral reasoning that may be applied to such decisions. This paper explores how reading text that emphasizes one of these styles might affect our reasoning. After participants read a series of tweets that emphasized the Fairness/Cheating foundation they exhibited an increased reliance on this style compared to when they read tweets emphasizing the Care/Harm foundation. This affected participants' answers to a questionnaire designed to measure the perceived importance of the different foundations, as well as in their rating of the foundations evident in other tweets. Interestingly, this effect was short lived and was not observed for the Care/Harm foundation. These results suggest that exposure to the moral reasoning of others might temporarily influence what moral arguments we are likely to accept and employ.


Keywords: Framing; Moral Foundation Theory; Moral Cognition; Priming; Text

## Introduction

Many of the important decisions we make have moral implications. But what factors might affect these decisions? In this paper, I examine the effect that encountering moral arguments might have on subsequent reasoning about moral issues. More specifically, I will argue that moral reasoning is subject to priming effects, where being confronted with a particular style of moral reasoning will result in increased salience for that style of reasoning.

## Moral Foundations Theory

While psychological research on morality encompasses a wide range of theoretical approaches (e.g., Gray, Young, \& Waytz, 2012; Malle, Guglielmo, \& Monroe, 2014; Rai \& Fiske, 2011; Young \& Saxe, 2011), in this paper I am interested in comparing different styles of moral reasoning and will therefore focus on Moral Foundations Theory (Graham et al., 2013; Haidt \& Joseph, 2004). Moral Foundations Theory identifies five different types of moral intuitions or concerns: Harm, Fairness, Loyalty, Authority, and Purity. Each of these moral concerns accounts for a different style of reasoning about moral dilemmas.

For instance, consider a person who believes that climate change is a problem because it endangers the lives of people and animals. This person is primarily concerned with the harm that climate change could cause to living beings. In contrast, another person might argue that climate change is problem because of its complexity and global reach, making it the obligation of nations to adhere to guidelines set by
international treaties. That person is using a type of argument that emerges from reasoning about authority. Critically, when analyzing any argument, it is important to remember that such moral concerns are not exclusive, and that a single argument can exhibit traits from several different concerns.

## Priming Moral Reasoning

Research based on Moral Foundations Theory has demonstrated that sensitivity to the different moral concerns varies across cultures (Graham, Haidt, \& Nosek, 2009), as well as based on ideological beliefs (Graham et al., 2009; Koleva, Graham, Iyer, Ditto, \& Haidt, 2012). Much of this research implicitly assumed that these styles of reasoning are stable and related to personality traits and beliefs. However, many stable traits in psychology provide a baseline for behavior that is affected by contextual and situational factors, such as priming.

The study presented here is designed to test whether such factors can also affect the salience of individual foundations. Specifically, I hypothesize that exposure to moral ideas and beliefs will result in temporary changes to the salience of the foundations that are at the core of these ideas.

For example, if an individual is presented with a text that relies on reasoning based on fairness, this individual might then become sensitized to the Fairness/Cheating foundation and be more likely to consider it as an important aspect of other, more ambiguous lines of reasoning. Likewise, reading a text about an individual that is harmed by a callous individual is likely to predispose the reader to identify harm as a more relevant consideration in subsequent texts that they otherwise would have.

To test this prediction, participants will be presented with a series of tweets that endorse either the Care/Harm foundation or the Fairness/Cheating foundation. Following this presentation, they will be asked to complete tasks that are designed to measure their sensitivity to these concerns. If moral reasoning is subject to priming effects, it is expected that participants who were presented with tweets endorsing the Fairness/Cheating foundation would find issues of fairness to be more relevant and important. In contrast, participants who read tweets that highlight Care/Harm should show heightened concern for that foundation.

## Method

## Participants

Thirty-six native English speakers from the University of St. Francis participated in the study in exchange for course credit.


Figure 1 - Mean scores on both parts of the MFQ30, by priming condition. The prime is administered after part 1 and before part 2. Error bars represent standard of error of the mean.

## Materials

## Moral Foundations Questionnaire

One of the frequently used tools for assessing an individual's level of concern for each of the 5 foundations is the 30 item Moral Foundations Questionnaire (MFQ30; Graham et al., 2011).

This questionnaire is composed of 2 parts: The first part asks participants to rate the relevance (on a 6-point scale) of various considerations to whether an act is right or wrong (e.g. "Whether someone suffered emotionally"). The second part asks participants to rate their agreement (on a 6-point scale) to various statements (e.g. "Chastity is an important and valuable virtue"). Each part is comprised of 16 items, 3 items corresponding to each of the foundations and 1 catch item.

It is important to note that while the two parts are measuring the same underlying concepts, they are using different approaches and therefore the scores on one part of the MFQ are not directly comparable to scores on the other. Nevertheless, a high score on a particular foundation in the first part can be taken as an indication of high concern for that foundation, and is therefore predictive of the score on the second part.

In this study, I used the first part of the MFQ30 to establish a baseline profile of the participants and the second part (administered after the care or fairness prime) to test for a priming effect.

## Tweets

In addition to the Moral Foundations Questionnaire, this study made use of several sets of tweets. These tweets were chosen from a corpus of over 700,000 tweets about the U.S.

Federal Shutdown of 2013 (cf. Dehghani et al., 2016; Sagi \& Dehghani, 2014b) ${ }^{1}$. Tweets were selected based on ratings of moral language computed statistically based on the Moral Foundations Dictionary (Graham et al., 2009) following the method described in Sagi and Dehghani (2014a).

The first set of primes, used as the prime in the Care condition, were uniformly high on the foundation of Care/Harm and low on the other 4 foundations. Likewise, a second set of primes served as the prime in the Fairness/Cheating condition. These primes were high on fairness and low on the other 4 foundations. Each of these lists comprised of 14 tweets, 7 tweets from liberal users and 7 from conservatives (see Appendix A).

In addition, a list of 25 tweets was selected such that each foundation was represented by 5 tweets. As before, for a foundation to be so represented, the tweet had to rate high on that foundation and low on all other foundations. This list of tweets was used for the rating task.

## Procedure

Participants first completed the first half of the 30 item Moral Foundations Questionnaire (MFQ30; Graham et al., 2011). Next, they rated their agreement, on a scale of 1 to 6 , to a series of 14 tweets that emphasized either the Fairness/Cheating foundation (Fairness condition) or the Care/Harm foundation (Care condition). After rating the primes, they completed the second half of the MFQ30.

Finally, each of the 5 moral foundations was described to the participants using the text from the website moralfoundations.org and they were asked to rate, on a scale of 1 to 6 , the relevance of each of the foundations to 25 tweets. Of the 25 tweets, 5 were primarily associated with each of the foundations. The tweets were presented in a

[^464]

Figure 2 - Mean ratings of tweets on the foundations of Care/Harm and Fairness/Cheating by experimental condition and order. Each block represents 5 tweets, in order of presentation. Error bars represent standard of error of the mean.
random order and its reverse, counterbalanced across participants. The ordering of the tweets (i.e., whether presented in the original random order or the reversed order) did not affect any of the analyses.

## Results

## Moral Foundations Questionnaire

Figure 1 presents the mean scores on both parts of the questionnaire. Since the prime is only presented after participants complete the first part, no differences are predicted in it. Agreement with the primes did not significantly differ based on condition (Care: $M=3.69, \mathrm{SD}=$ 0.68; Fairness: $M=3.61, S D=0.63 ; t(34)=-0.36$, n.s.)

Participants' responses to the second part of the MFQ30 were analyzed using a separate general linear model for each of the foundations, with the foundation score in the first part of the MFQ30 and the prime condition (Fairness vs. Care) as independent variables ${ }^{2}$. Participants' scores on the second part of the questionnaire were correlated to their scores on the first part, at least marginally (after correcting for multiple tests), for all but the Loyalty/Betrayal foundation $(F(1,32)=$ 1.87, $p=.18 ; F(1,32)>6, p<.05, r^{2}>.19$ for all other foundations). Only the Fairness foundation showed the predicted interaction between the score on the first part of the MFQ30 and condition $\left(F(1,32)=10.00, p<.01, r^{2}=.34\right.$; Fairness condition: $M=4.30, \mathrm{SD}=0.78$; Care condition: $M$ $=4.04, \mathrm{SD}=0.98 ; F<1$ for all other foundations).

[^465]Moreover, the scores on the fairness questions of the second part of the MFQ30 of participants in the Care condition correlated with their fairness score on the first part $(r(16)=.73, p<.001)$ while those of participants who read tweets evoking fairness did not $(r(16)=-0.11$, n.s.). This suggests that following the fairness primes participants concern for fairness was uniformly high - the prime essentially set all participants to the same level of concern on fairness. In contrast, similar correlations on the harm scores of the MFQ30 did not differ significantly (care condition: $r(16)=.42, p=.08$; fairness condition: $r(16)=.63, p<.01)$. These results suggest that the fairness prime successfully increased the salience of the Fairness/Cheating foundation, the care prime did not increase the salience of the Care/Harm foundation.

## Ratings of Tweets

To simplify the analysis of the ratings and avoid repeated tests, the analysis of the 25 rated tweets used a single model that contrasted the ratings of harm and fairness (although a similar, post-hoc, model using all 5 foundations yielded qualitatively similar results). This model included the participant and the tweet as random variables, and the condition as well as the foundation being rated as independent variables. To test for the possibility that this effect diminishes over time, the 25 tweets were divided into 5 blocks of 5 tweets based on order of presentation and this variable was included in the model (model $r^{2}=.37$ ). As
models that include this variable show no effect of agreement and are otherwise unchanged.
predicted, participants rated tweets as higher in fairness if they were previously exposed to tweets that exhibited fairness-based reasoning and vice versa $(F(1,1734)=5.46, p$ <.05). However, this effect quickly diminished as is evident by its interaction with the order of presentation $(F(1,1734)=$ $3.88, p<.05$; see Figure 2).

## Discussion

The results of the present study demonstrate that reading texts that evoke principles of moral reasoning can affect judgments and decisions made later. The effects observed in this paper are therefore best considered to be a type of priming effects. Since priming effects are, for the most part, short lived, the rapid decay of the effect in the second part of the study is also easily explained. However, it is likely that, because the second part of the study overtly asked participant to consider all five styles of moral reasoning, it accelerated this decay and that in a more natural setting the effect might last longer.

Perhaps more interesting is the fact that while reading tweets involving fairness and cheating resulted in a priming effect, reading tweets that favored the foundation of Care/Harm did not. One possible explanation is that while the federal shutdown readily appealed to the foundation of Fairness/Cheating, its appeal to Care/Harm is less direct and evident. This is reflected in the tweets - although Care/Harm was a dominant foundation in the corpus for liberals, considerations of fairness dominated the overall debate (cf. Sagi \& Dehghani, 2014b). It is possible that rather than simply evoking a moral foundation, a consistent and/or clear moral position might be required for a text to affect the moral reasoning of its reader.

More generally, there are numerous studies that demonstrate how the use of language can affect reasoning, both in the lab (e.g., Tversky \& Kahneman, 1981), and outside of it (e.g., Goodwin, 1994). Moreover, it is possible to use language to measure and trace the history of such frames (Sagi, Diermeier, \& Kaufmann, 2013).

In a similar vein, there is evidence that situational factors affect an individual's moral reasoning. The bystander effect, where individuals are less likely to render assistance when there are many other bystanders than when there are few, is a prominent example of such an effect (Darley \& Latane, 1968).

This study combines these two well-known effects and demonstrates that this type of framing can provide a context in which moral reasoning takes place. More interestingly, it is possible that repeated exposure to particular styles of reasoning might have a cumulative effect and eventually lead to the salience of the relevant foundation being permanently increased (or, perhaps, decreased, depending on the circumstances of exposure). This type of effect might be at the root of the development of moral beliefs and might provide insight into how and why such beliefs change.

Moreover, even temporary effects might have important implications. For example, the language used to draft jury instructions might influence the verdict one way if it highlights fairness and another if it highlights care. Likewise,
during negotiations, it is possible that a particular choice of language and reasoning by one side can serve to focus the negotiation in a particular direction, influencing all parties towards emphasizing the importance of a specific concern.

## References

Darley, J. M., \& Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8(4, Pt.1), 377-383. https://doi.org/10.1037/h0025589
Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., ... Graham, J. (2016). Purity homophily in social networks. Journal of Experimental Psychology: General, 145(3),

366-375. https://doi.org/10.1037/xge0000139
Goodwin, C. (1994). Professional vision. American Anthropologist, 96(3), 606-633.
Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., \& Ditto, P. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. Advances in Experimental Social Psychology, 47, 55-130. https://doi.org/10.1177/0963721412456842
Graham, J., Haidt, J., \& Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96(5), 1029-1046. https://doi.org/10.1037/a0015141
Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., \& Ditto, P. H. (2011). Mapping the Moral Domain. Journal of Personality and Social Psychology, 101(2), 366-385. https://doi.org/10.1037/a0021847
Gray, K., Young, L., \& Waytz, A. (2012). Mind Perception Is the Essence of Morality. Psychological Inquiry, 23(2), 101-124. https://doi.org/10.1080/1047840X.2012.651387
Haidt, J., \& Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. Daedalus, 133(4), 55-66.
Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., \& Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. Journal of Research in Personality, 46(2), 184-194.
Malle, B. F., Guglielmo, S., \& Monroe, A. E. (2014). A Theory of Blame. Psychological Inquiry, 25(2), 147-186. https://doi.org/10.1080/1047840X.2014.877340
Rai, T. S., \& Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. Psychological Review, 118(1), 57-75. https://doi.org/10.1037/a0021867
Sagi, E., \& Dehghani, M. (2014a). Measuring Moral Rhetoric in Text. Social Science Computer Review, 32(2), 132-144. https://doi.org/10.1177/0894439313506837
Sagi, E., \& Dehghani, M. (2014b). Moral Rhetoric in Twitter: A Case Study of the US Federal Shutdown of 2013. In P. Bello, M. Guarini, M. McShane, \& B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 1347-1352).
Sagi, E., Diermeier, D., \& Kaufmann, S. (2013). Identifying Issue Frames in Text. PloS One, 8(7), e69185.

Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453-458. https://doi.org/10.1126/science. 7455683
Young, L., \& Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. Cognition, 120(2), 202-214.
https://doi.org/10.1016/j.cognition.2011.04.005

## Appendix A - Primes

## Care/Harm Primes

> Dr. Seuss's \#Congress Who Stole \#SNAP: Kids seniors face health risks \#GovernmentShutdown. \#PublicPolicy

New \#Obama Doctrine: Protect oil, allies, the homeland from terrorists weapons of mass destruction.
\#governmentshutdown Day 9. Private charity pays Military death benefits instead of \#Pentagon. What do you think?

As \#soldiers we're told \#WWII, \#Vietnam, \#Iraq, were all to protect \#Democracy, yet the latest attack comes from The \#TeaParty \#Shutdown
\#Shutdown: \#Obama Keeps Military \#Golf_Courses \#Open, \#Closes Military \#Grocery_Stores |

Mother of fallen \#soldier denied death benefits: \#Criminal to treat our \#soldiers this way \#congress. via @todayshow

If \#obama can treat our military and vets like he does imagine how he's going to be with civilians and our healthcare. Disgust
\#Obama is trying his hardest to create pain - Military death benefits denied to families of fallen troops

I don't care about the shutdown...PAY the families of our fallen heroes!!! \#shutdown \#governmentshutdown \#Military \#veterans \#Obama

I wish \#Congress cared as much about war vets benefits as they do about the war vets memorial. \#hypocrite \#pander \#teaparty

Sickening that the families of our fallen heroes denied benefits by shutdown. Time to stop the madness.\#shutdown

Outrageous not paying death benefits to families of our fallen servicemen! This SOB \#Obama looking for a civil war to become dictator!

The D-Day memorial in Normandy, France has been closed, upsetting tourists and veterans. via @WSJ \#shutdown

Refugees Waiting Overseas Are in Limbo as U.S. Shutdown Continues \#refugees \#shutdown \#resettlement \#newcomers \#USA

## Fairness/Cheating Primes

Liberal \#Congress members claim that the law must apply equally to all...well, except them. \#Obamacare \#Dems \#GOP

Libs scream \#obamacare = law of the land. Weird cuz theyre VERY WILLING 2 ignore immigration borders, ya kno another LAW OF THE LAND

A bunch of liberals looked really stupid tonight, talking about \#obamacare. They're still ignorant of the law. \#tcot \#election2014 \#pathetic

Fighting Republican hysteria with calm analysis on the ACA. \#p2 \#toppage \#dems \#liberals \#progressives \#healthcare

Hey \#GOP! \#OBAMACARE website overloaded huh? Looks like Americans want an alternative you elephant sized asses!

I think it's hilarious T-Party called ACA \#Obamacare as a negative slur. The more popular it gets, the bigger my SMILE gets POTUS's too!
Hey \#GOP look up the 14th amendment! If u love the Constitution Founding Fathers so much, then ADHERE to the law of the land. \#JustVote

Liberal \#Congress members claim that the law must apply equally to all...well, except them. \#Obamacare

Equal under the law; all laws enforced equally - its pretty simple for everyone to understand except Obama \#TeaParty \#tcot \#tngop \#gop \#ccot

Y did the unions get exempt from \#Obamacare I thought it was the law of the land doesn't it apply to everyone like every other law \#tcot

Funny how libs like @tamaraholder are all about \#obamacare being the law but other social issues like upholding the sanctity of marriage..

LIberals progressives say that \#obamacare is the law of the land, but they ignore illegals breaking the law of the land!
also calling progressives 'liberals' (not saying that someone IS liberal, but calling them 'liberals') is \#GOP branding.

Smart Libs know Repubs hate that \#Obama wins. He beat them twice in elections, SCOTUS upheld \#ACA. It just kills em. 2BAD!

# Estimating Causal Power between Binary Cause and Continuous Outcome 

Motoyuki Saito (motoyuki.saito28@gmail.com)<br>Department of Psychology, University of California, Los Angeles<br>Los Angeles, CA 90095, USA<br>Department of Psychological Science, Kwansei Gakuin University<br>Nishinomiya, Hyogo 662-8501, JAPAN<br>Patricia, W. Cheng (cheng@lifesci.ucla.edu)<br>Department of Psychology, University of California, Los Angeles<br>Los Angeles, CA 90095, USA


#### Abstract

Previous studies of causal learning heavily focused on binary outcomes; little is known about causal learning with continuous outcomes. The present paper proposes a qualitative extension of the causal power theory to the situation where a binary cause influences a continuous effect, and induces causal power under various ceiling situations with the continuous outcomes. To test the predictions, we systematically manipulated the type of outcome (continuous vs. percentage vs. binary) and the contingency information. The experiment shows that people estimate causal strength based on the linear-sum rule for continuous outcomes and the noisy-OR rule for binary outcomes. In the partial ceiling situation where causal power is partially inferred but not precisely estimated, the distribution of participants' judgments was bimodal with one mode at the minimum value and the other at the maximum value, suggesting some participants made conservative estimates while others made optimistic estimates. These results are generally consistent with the predictions of the causal power theory. Theoretical implications and future directions are discussed.


Keywords: causal reasoning; causal inference; causal power; continuous variable; integration rules.

## Introduction

The ability to learn causal relations is essential for explaining past events, controlling the present environment, and predicting future outcomes. Decision making based on causal knowledge enables us to achieve desired outcomes and to avoid undesired consequences. When there are two causes of a desired outcome, we should consider which cause has a high causal strength for producing the outcome. To estimate causal strength, we need to consider not only the states of the effect in the presence of the cause but also that in the absence of the cause (Rescorla, 1968). When a teacher thinks about the effect of active encouragement on students' homework performance, for example, he or she has to check whether the student finishes the homework both in the presence and absence of encouragement. It has been recognized that both children and adults readily form representations of causal networks (see Holyoak \& Cheng, 2011 for a review).

As causal relations are unobservable, they must be induced from observable events, and covariation among observable events serves as a fundamental cue to learn
causal relations (Hume, 1739/2000). For binary variables, covariation is represented as patterns of presence and absence. A measure of contingency is described by $\Delta P$ (Jenkins \& Ward, 1965):

$$
\begin{equation*}
\Delta P=P(E=1 \mid C=1)-P(E=1 \mid C=0) \tag{1}
\end{equation*}
$$

where $P(E=1 \mid C=1)$ is the probability of effect $E$ given the presence of candidate cause $C$, and $P(E=1 \mid C=0)$ is the probability of $E$ given the absence of $C$. Values of $\Delta P$ range from -1 to +1 . Positive $\Delta P$ values indicate a generative causal relation; negative $\Delta P$ values indicate a preventive causal relation.

Because $\Delta P$ is a measure of associative strength, it does not address issues in causation such as confounding and ceiling effects. For example, although it is impossible to judge the causal effect when the outcome always occurs regardless of the presence or absence of the cause (i.e., $P(E$ $=1 \mid C=1)=P(E=1 \mid C=0)=1)$, the $\Delta P$ model indicates that there is no causal relation (i.e., $\Delta P=0$ ). To model causal strength, Cheng (1997) proposed the power PC theory and derived generative causal power as an estimate:

$$
\begin{equation*}
w_{\mathrm{c}}=\Delta P /[1-P(E=1 \mid C=0)] \tag{2}
\end{equation*}
$$

Causal power $w_{c}$ is a function not only of contingency but also of the base rates of the effect. When the effect is always present, generative causal power is undefined, therefore explaining the (generative) ceiling effect. Buehner, Cheng, and Clifford (2003) systematically manipulated covariation information and demonstrated that judgments were well described by the causal power. Causal power is interpreted in the framework of causal Bayes nets by Glymour (2001) and of causal Bayesian models by Tenenbaum \& Griffiths (2001; Griffiths \& Tenenbaum, 2005).

Lu, Rojas, Beckers, and Yuille (2016) proposed a Bayesian theory of sequential causal learning. Their theory assumes that people select a different integration rule according to the type of outcome variable. On one hand, the noisy-OR rule is appropriate for a binary outcome and is consistent with Equation (2) in the causal power theory (Cheng, 1997). The rule assumes that two causes influence an outcome independently. It states that the effectiveness of two causes, both present, is the sum of the causal power of
each minus their product (i.e., $P(E=1 \mid A=1, B=1)=w_{\mathrm{A}}+$ $w_{\mathrm{B}}-w_{\mathrm{A}} \times w_{\mathrm{B}}$ ). On the other hand, the linear-sum rule is appropriate for a continuous outcome and is widely used in associative learning models such as the R-W model (Rescorla \& Wagner, 1972). This rule simply calculates the sum of the influence of each cause (i.e., $P(E=1 \mid A=1, B=$ 1) $=w_{\mathrm{A}}+w_{\mathrm{B}}$ ). Lu et al. (2016) presented a sequential Bayesian model that explains previous findings on outcomeadditivity in variations of the blocking paradigm.

Several empirical studies have provided supporting evidence for the use of the linear-sum rule for a continuous outcome. Rashid and Buehner (2013) systematically manipulated the quantity of continuous outcomes and tested which integration rules people use. The results appear inconsistent in that participants use the linear-sum rule for a generative cause and noisy-OR rule for a preventive cause. They suggested that the use of the linear-sum rule might be due to the absence of an upper limit for the quantity of the continuous outcome in their cover story. Prevention has a natural lower limit, the outcome quantity equal to 0 , sharing that property with binary outcomes. Saito (2015) manipulated means and standard deviations in causal learning with continuous outcomes and found that judgments are largely explained by difference in the means, but not by difference in the standard deviations. White (2015) examined causal judgments of interventions in temporal sequences of a continuous outcome variable in single individuals and reported that most of the results were explained by the difference between the mean outcome value for the pre-intervention time periods and that for the post-intervention time periods. These results suggest that people use the linear-sum rule for continuous outcomes. However, these studies do not reveal whether people use a different integration rule depending on the type of outcome variable since they did not compare judgments for continuous outcomes with those for the binary outcomes. In addition, it remains unknown how people estimate causal strength under various ceiling situation with the continuous outcomes. Since integration rules are core parts of the models of causal learning, it is important to investigate how people choose an integration rule.

In this paper, we extend the causal power theory qualitatively to address continuous outcomes and derive predictions under various ceiling effects. For our purposes, we treat cardinal outcomes as a special case of continuous outcomes. We also report a study investigating whether people choose the appropriate integration rules according to the type of outcome variables and whether their judgments correspond to causal-power predictions.

## Estimating causal power with continuous outcomes

The reasoner's goal is to induce the unobservable causal power of a candidate cause from observable events (Cheng, 1997). Consider a situation where a continuous effect $E$ may be produced by a binary background cause $B$ and/or a binary candidate cause C. Assume that:
(1) $B$ and $C$ influence $E$ independently,
(2) $B$ could increase $E$ but not reduce it,
(3) The causal powers of $B$ and $C$ are independent of the frequency of occurrences of $B$ and $C$, and
(4) $E$ does not change unless it is influenced.

These assumptions are similar to those with binary cause and effect (cf. Cheng, 1997; Pearl, 1998).

The joint influence of background cause $B$ and candidate cause $C$ on the continuous outcome $E$ is given by the linearsum rule (cf. Lu et al., 2016). According to this integration rule, the influences of multiple causes are integrated by simple addition. Since the outcome can take on different values, expected value and conditional expected value are used to describe its state. The expected value of the continuous outcome is calculated as follows:

$$
\begin{equation*}
E[e]=P(b) \cdot w_{\mathrm{b}}+P(c) \cdot w_{\mathrm{c}} \tag{3}
\end{equation*}
$$

In this equation, $P(b)$ and $P(c)$ denote the probabilities of occurrences of the background cause and candidate cause. Variables $w_{\mathrm{b}}$ and $w_{\mathrm{c}}$ are causal powers of the background cause and candidate cause. Although two different integration rules are used for a binary outcome (i.e., noisyOR rule for generative cause; noisy-AND-NOT rule for preventive cause), there is no distinction between generative and preventive causes in case of a continuous outcome.

When the cause is present (i.e., $P(c)=1$ ), the conditional expected value given the presence of the cause is

$$
\begin{equation*}
E[e \mid c]=P(b \mid c) \cdot w_{\mathrm{b}}+w_{\mathrm{c}} \tag{4}
\end{equation*}
$$

Similarly, the conditional expected value given the absence of the cause (i.e., $P(c)=0$ ) is

$$
\begin{equation*}
E[e \mid \neg c]=P(b \mid \neg c) \cdot w_{\mathrm{b}} \tag{5}
\end{equation*}
$$

Subtracting Equation 5 from Equation 4 yields the difference in conditional expected values (i.e., $\Delta E=E[e \mid c]$ $-E[e \mid \neg c]$ ). The difference in conditional expected values is:

$$
\begin{equation*}
\Delta E=P(b \mid c) \cdot w_{\mathrm{b}}+w_{\mathrm{c}}-P(b \mid \neg c) \cdot w_{\mathrm{b}} \tag{6}
\end{equation*}
$$

If we assume that the background cause and the candidate cause occur independently, two conditional probabilities equal to one another (i.e., $P(b \mid c)=P(b \mid \neg c)=P(b)$ ). Therefore, the causal power of the candidate cause $w_{c}$ is represented as follows:

$$
\begin{equation*}
w_{\mathrm{c}}=\Delta E-\{P(b \mid c)-P(b \mid \neg c)\} w_{\mathrm{b}}=\Delta E \tag{7}
\end{equation*}
$$

Within the range of outcome values greater than the minimum and less than the maximum, predicted values of the causal power are simply the differences in conditional expected values.

Predictions of the value of $w_{c}$ vary depending on the value of the continuous outcome. To illustrate these predictions,
consider a situation where a teacher investigates the effect of active encouragement on students' homework performance and gives 100 homework problems to each student. For example, a student had finished 25 out of 100 previous homework problems assigned; after the encouragement, the student finished 75 out of 100 new homework problems assigned. The causal power $w_{\mathrm{c}}$ is the difference in performance before and after the encouragement (i.e., $\Delta E=E[e \mid c]-E[e \mid \neg c]=75-25=50$ ).

However, this is not the case where one of the outcome values reaches the upper limit. We hypothesize that depending on the reasoner's assumption about the counterfactual value of the outcome if there were no upper limit, $w_{\mathrm{c}}$ has a range of possible values. We replace $E[e \mid c]$ in Equation (4) with the assumed counterfactual value $E^{\prime}[e \mid c]$. Suppose a student had finished 50 out of 100 previous homework problems assigned and then finished 100 out of 100 new homework problems assigned. It is inferred that the causal power is equal to or larger than 50 , but not precisely determined. Thus, the prediction of the causal power theory is an interval. Whereas some cautious reasoners might estimate the minimum value in the interval (50 in this case, resulting from $E^{\prime}[e \mid c]=100$ ), other reasoners might estimate a higher value in the interval (e.g., 100 , resulting from $E^{\prime}[e \mid c]=150$ ). When both outcomes are at the maximum value (e.g., a student finished 100 out of 100 homework problems regardless of the encouragement), the interval spans the entire range from 0 on and causal power becomes undefined. We call the former the partial ceiling situation and the latter the full ceiling situation. The difference between the partial and full ceiling situations is a unique feature in causal learning with continuous outcomes that have maximum values. The predictions of the causal power theory are shown in Table 1.

The purpose of the present study is to investigate whether people use proper integration rules according to the type of outcome variable and whether people differentiate between the partial and full ceiling situations. In addition to the conditions with continuous outcomes and binary outcomes, we added a condition with percentage outcomes. This is

Table 1: Design and predictions of the experiment.

| Continuous \& Percentage |  |  |  |  | Binary |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $E[e \mid c]$ | $E[e \mid c c]$ | $\Delta E$ | causal <br> power |  | $P(e \mid c)$ | $P(e \mid c c)$ | $\Delta P$ | causal <br> power |
| $100(0)$ | $0(0)$ | 100 | $\geq 100^{*}$ |  | $1.00(0)$ | $0.00(0)$ | 1.00 | 1.00 |
| $100(0)$ | $25(4.3)$ | 75 | $\geq 75$ |  | $1.00(0)$ | $0.25(0.4)$ | 0.75 | 1.00 |
| $75(4.3)$ | $0(0)$ | 75 | 75 |  | $0.75(0.4)$ | $0.00(0)$ | 0.75 | 0.75 |
| $100(0)$ | $50(5)$ | 50 | $\geq 50$ |  | $1.00(0)$ | $0.50(0.5)$ | 0.50 | 1.00 |
| $75(4.3)$ | $25(4.3)$ | 50 | 50 |  | $0.75(0.4)$ | $0.25(0.4)$ | 0.50 | 0.67 |
| $50(5)$ | $0(0)$ | 50 | 50 |  | $0.50(0.5)$ | $0.00(0)$ | 0.50 | 0.50 |
| $100(0)$ | $75(4.3)$ | 25 | $\geq 25$ |  | $1.00(0)$ | $0.75(0.4)$ | 0.25 | 1.00 |
| $75(4.3)$ | $50(5)$ | 25 | 25 |  | $0.75(0.4)$ | $0.50(0.5)$ | 0.25 | 0.50 |
| $50(5)$ | $25(4.3)$ | 25 | 25 |  | $0.50(0.5)$ | $0.25(0.4)$ | 0.25 | 0.33 |
| $25(4.3)$ | $0(0)$ | 25 | 25 |  | $0.25(0.4)$ | $0.00(0)$ | 0.25 | 0.25 |
| $100(0)$ | $100(0)$ | 0 | NA |  | $1.00(0)$ | $1.00(0)$ | 0.00 | NA |
| $75(4.3)$ | $75(4.3)$ | 0 | 0 |  | $0.75(0.4)$ | $0.75(0.4)$ | 0.00 | 0.00 |
| $50(5)$ | $50(5)$ | 0 | 0 |  | $0.50(0.5)$ | $0.50(0.5)$ | 0.00 | 0.00 |
| $25(4.3)$ | $25(4.3)$ | 0 | 0 |  | $0.25(0.4)$ | $0.25(0.4)$ | 0.00 | 0.00 |
| $0(0)$ | $0(0)$ | 0 | 0 |  | $0.00(0)$ | $0.00(0)$ | 0.00 | 0.00 |

Note. Numbers in parentheses are standard deviations. The causal power theory predicts " $\geq 100$ " for the continuous group and " 100 " for the percentage group.
because the upper limit for percentage outcomes has a clear maximum of 100 , unlike that for continuous outcomes.

## Method

## Participants

A total of 136 participants were recruited from Amazon Mechanical Turk (http://www.mturk.com/). An additional 35 participants were tested but excluded for failing to pass the comprehension question (see below for details). All were native English speakers and residing in the US.

## Experimental design

Participants were randomly assigned to one of three groups differing on the type of outcome (continuous, percentage, or binary). For all groups, the candidate cause was a binary variable (i.e., presence or absence of encouragement). Exclusion by the comprehension question resulted in unequal group sizes ( 56 participants in the continuous group, 43 in the percentage group, and 37 in the binary group). In addition to manipulating type of outcome, contingency information was manipulated within-subject (see Table 1). In the continuous and percentage groups, there were 15 contingency conditions resulting from the combination of five levels ( $100,75,50,25,0$ ) of conditional expected values in the presence and absence of the cause. The difference between $E[E \mid C=1]$ and $E[E \mid C=0]$ for each condition yielded five levels of nonnegative values in the outcome magnitude $(\Delta E=E[E \mid C=1]-E[E \mid C=0]=100$, $75,50,25,0)$. Similarly, the binary group had five levels of nonnegative values in the difference (i.e., $\Delta P=P(E=1 \mid C=$ 1) $-P(E=1 \mid C=0)=1.00, .75, .50, .25, .00)$. Participants in each group completed the causal learning task for all contingency conditions. The order of the contingency conditions was randomized across participants.

## Procedure

Instructions Participants were asked to read the instructions carefully and answer each question thoughtfully. The exact instructions in the continuous group were as follows (italicized sentences differed across groups):

A math teacher wants to investigate the effect of active encouragement on students' homework performance. Students are given 100 math homework problems of similar difficulty. The teacher randomly assigns some students to receive encouragement and assigns other students to receive no encouragement.

Imagine that you are a teaching assistant for the class. You are responsible for checking whether or not a student receives encouragement and how many out of the 100 homework problems the student finishes (0-100).

You will see several sets of student records. Each set contains the records of students from a school ordered in a random sequence. Each record describes
a student's homework performance before and after the experiment. After observing the records of sixteen students from a school, you will be asked to judge how much the encouragement increases performance at that school.

For the continuous group, the effect was a continuous variable (i.e., number of finished homework problems). The same instructions were used in the percentage group with one exception: the outcome observation was described as "what percentage of the homework problems the student finishes $(0-100 \%)$." In the binary group, both cause and effect were binary variables. Specifically, the instructions stated the outcome observation as "whether or not the student finishes the homework problems."

After reading the instructions, participants were asked to answer the comprehension question that checks the understanding of random assignment. The exact question was (italicized sentences differed across groups):

Before you begin viewing the records of the students' homework performance, consider the following situation. Suppose we conduct a study, and find that: the average number of the homework problems students in the experimental group (those who received encouragement) finish is 65 . Likewise, the average number of the homework problems students in the control group (those who did not receive encouragement) finish is 65 as well. Recall that the students are were randomly assigned to one or the other group. Can the homework performance in the experimental group be attributed to encouragement?

Participants were required to provide a "yes" or "no" answer and to justify their answer briefly. This question was intended to exclude participants who did not read the instructions properly and to encourage the assumption that the influence of background causes (i.e., causes other than encouragement) on homework performance was constant across the two groups (cf. Buehner et al., 2003). Similar questions were used in the percentage and binary groups with the corresponding modifications of the descriptions in terms of percentages. Participants received no feedback on their answers to this question.
Learning phase The learning phase consisted of 16 trials that presented information about the cause and effect in a pre-post design. For the continuous group, participants were requested to observe whether a student receives encouragement (present or absent) and how many out of the 100 homework problems the student finishes (0-100) before and after encouragement. On each trial, homework performance before the encouragement for a student was described with the illustration and text (e.g., "A student (ID: 12345) at this school finished 25 out of 100 previous homework problems assigned"). Student ID was a five-digit random number and designed to show that each trial described a different student. The states of the
encouragement were provided with the sentence (e.g., "The student received encouragement" or "The student did not receive encouragement"). The other two groups followed an identical procedure, except that the outcomes were expressed in percentage terms for the percentage group (e.g., "The student finished $75 \%$ of new homework problems") and as present or absent in the binary group (e.g., "The student finished the new homework problems"). The interstimulus interval was $1000-\mathrm{ms}$, and the button to proceed to next trial was presented $500-\mathrm{ms}$ after the presentation of all the information. Each trial was separated by a $500-\mathrm{ms}$ blank screen. Participants were required to learn causal strength of the encouragement on homework performance through trials.

There were 16 trials for each contingency condition in Table 1. Encouragement was present on 8 trials and was absent on 8 trials. For the continuous group, the outcomes were normally distributed with the variance set to be ten times that in the binary group (see standard deviations in parentheses in Table 1). The order of trials was randomized within-subject. To familiarize participants with the procedure, practice trials were presented prior to the learning phase.
Test phase After the 16 learning trials, participants were asked to estimate the causal strength of the candidate cause in a counterfactual question. In the continuous group, the question was "Suppose the next student (ID: 23456) at this school finished 0 out of 100 previous homework problems assigned. If the student now receives encouragement, how many out of 100 new homework problems will the student finish?" The responses were made on a rating scale ranging from 0 to 100 . Our scale limits the maximum strength to 100 so that responses can be compared across groups. Similar questions were used in the percentage and binary groups with modifications of the descriptions corresponding to the outcome type (e.g., for the binary group, "If these 100 students now receive encouragement, how many of them will finish their new homework problems?"). In addition, participants were also asked to report confidence in their judgment on a scale ranging from 0 (not confident at all) to 100 (extremely confident). After their judgments, participants completed the next contingency condition. To encourage the independence of judgments in each condition, participants received the following instructions: "Recall that the schools have students from very different socioeconomic backgrounds, and encouragement may have different effects on the students from school to school. Please evaluate each school separately."

## Results

Participants who failed to pass the comprehension question were excluded from our analysis below. This procedure reduced noise, but did not alter the general pattern of the results. Since the causal power theory makes different predictions for the non-ceiling and ceiling situations, separate analyses were conducted. Figure 1 shows the mean ratings of causal strength in non-ceiling situations. Overall, participants clearly differentiated between continuous and

Continuous


Percentage


## Binary



Figure 1: Mean ratings of causal strength in each contingency condition. Judgments with the same level of $\Delta E$ or $\Delta P$ in the non-ceiling situations are connected by lines. Judgments in the partial and full ceiling situations are represented by black symbols.
percentage outcomes on one hand and binary outcomes on the other. In the continuous group, judgments generally corresponded to the difference between the conditional expected values (i.e., $\Delta E=E[E \mid C=1]-E[E \mid C=0]$ ). Similar results were obtained in the percentage group, but the trend was much more evident. In contrast, judgments in the binary group were affected by both the difference between conditional probabilities $\Delta P$ and the base rates of the effect $P(E=1 \mid C=0)$. These descriptive analyses were confirmed by statistical analyses.

A two-way mixed ANOVA with the type of outcome (continuous vs. percentage vs. binary) as between-subjects factor and the contingency condition (11 contingency conditions except for the ceiling situations) as withinsubject factor resulted in a significant two-way interaction, $F(20,1330)=6.45, M S E=210.3, p<.001, \eta_{\mathrm{G}}^{2}=.069$. To explore the results in greater detail, we analyzed the effect of the type of outcome for each $\Delta E$ and $\Delta P$ condition. In the $\Delta E=50$ and $\Delta P=.50$ conditions, a two-way mixed ANOVA revealed a significant interaction between the type of outcome and contingency condition, $F(2,133)=5.54$, $M S E=116.1, p=.005, \eta_{\mathrm{G}}^{2}=.036$. As expected, judgments varied as a function of the base rate of the effect in the binary group, $F(1,36)=10.95, M S E=206.5, p=.002, \eta_{\mathrm{G}}^{2}$ $=.118$, but not in the continuous and percentage groups, $F \mathrm{~s}$ $<1$. The interaction was also significant in the $\Delta E=25$ and $\Delta P=.25$ conditions, $F(4,266)=7.76, M S E=153.6, p$ $<.001, \eta_{\mathrm{G}}^{2}=.057$, and in the $\Delta E=0$ and $\Delta P=.00$ conditions, $F(6,399)=2.33, M S E=242.7, p=.032, \eta_{\mathrm{G}}^{2}$ $=.015$. Although the incremental pattern of the results in the binary group in the $\Delta P=.00$ condition was inconsistent with the predictions of the causal power theory, it may be explained by misperception of contingency for sequential trials due to working memory limitations. This outcomedensity effect is consistent with causal-power predictions given the misperceptions (Cheng, 1997). It is worth noting that a similar but smaller trend was found in the continuous group, but not in the percentage group. This might be because the continuous group needs an assumption of equal
upper limits to compare outcomes whereas the percentage group does not.

Figure 2 depicts distributions of individual judgments in the partial and full ceiling situations. In the partial ceiling situation (i.e., $100-25,100-50,100-75$ conditions), a range of causal power is inferred (e.g., equal to or larger than 75 in the 100-25 condition). The distribution of participants' judgments appears bimodal with one mode at the minimum value of the interval and the other at the maximum value of the interval given our scale. These results indicate that some participants made conservative estimates while others made optimistic estimates. Dip tests confirmed the bimodality both in the continuous group ( $D=0.08, p=.013$ in the 100 25 condition, $D=0.13, p<.001$ in the 100-50 condition, $D$ $=0.14, p<.001$ in the $100-75$ condition) and percentage group ( $D=0.13, p<.001$ in the $100-25$ condition, $D=0.16$, $p<.001$ in the $100-50$ condition, $D=0.16, p<.001$ in the 100-75 condition). In contrast, the bimodality was not observed in the binary group, and the mode of the distribution corresponded to the point estimate of causal power (i.e., $w_{\mathrm{c}}=1$ ).

In the full ceiling situation where the causal power cannot be estimated (i.e., 100-100 condition), the distributions of the continuous and percentage groups appear bimodal while that of the binary group appear trimodal. This might be because participants had no option to answer "I don't know" in our materials.

## Discussion

The present study qualitatively extended the causal power theory to deal with the continuous outcomes and tested whether people differentiate between continuous and binary outcomes. The results showed that people estimate causal strength based on the linear-sum rule for continuous outcomes and the noisy-OR rule for binary outcomes. In the partial ceiling situation where the estimation of causal power has a range, the distribution of participants' ratings was bimodal with one mode at the minimum value and the other at the maximum value, suggesting some participants made conservative estimates while others made optimistic


Figure 2: Distributions of individual judgments in the partial and full ceiling situations.
estimates. These results are generally consistent with the predictions of the causal power theory.

The present study has theoretical implications for understanding how people estimate causal power. Whereas covariation models (e.g., $\Delta P$ model, Jenkins \& Ward, 1965) and associative models (e.g., R-W model, Rescorla \& Wagner, 1972) adopt one integration rule, Bayesian models generally assume multiple integration rules (Griffiths \& Tenenbaum, 2005, 2009; Lu et al., 2008, 2016). Our results demonstrate that people choose the proper integration rule according to the type of outcome, supporting the Bayesian models. Notably, this finding implies that people assume the invariance of causal power as a default, consistent with the proposal that causal invariance plays a key role in the construction of generalizable causal knowledge (Cheng \& Lu , in press). The two integration rules respectively represent the invariance of causal power for the two outcome variable types. Another theoretically important aspect is the bimodal distributions in the judgments in the partial ceiling situations. Computational models generally predict averaged results. The observed bimodality suggests that models incorporating different conservatism values and/or priors may explain individual differences in the partial ceiling situations. Further investigations will shed more light on the question of how people estimate causal power.

## References

Buehner, M. J., Cheng, P. W., \& Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 1119-1140.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104, 367-405.
Cheng, P. W., \& Lu, H. (in press). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing the invariance of causal power. In M. R. Waldmann (Ed). The Oxford Handbook of

Causal Reasoning. Oxford, England: Oxford University Press.
Glymour, C. (2001). The mind's arrows: Bayes nets and graphical causal models in psychology. Cambridge, MA: MIT Press.
Griffiths, T. L., \& Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive Psychology, 51, 334-384.
Griffiths, T. L., \& Tenenbaum, J. B. (2009). Theory-based causal induction. Psychological Review, 116, 661-716.
Hume, D. (1739/2000). A treatise of human nature. Oxford: Oxford University Press.
Holyoak, K, J., \& Cheng, P, W. (2011). Causal learning and inference as a rational process: The new synthesis. Annual Review of Psychology, 62, 135-163.
Jenkins, H., \& Ward, W. (1965). Judgment of contingency between responses and outcomes. Psychological Monographs, 7, 1-17.
Lu, H., Rojas, R. R., Beckers, T., \& Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. Cognitive Science, 40, 404-439.
Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., \& Holyoak, K. J. (2008). Bayesian generic priors for causal learning. Psychological Review, 115, 955-984.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Francisco, CA: Morgan Kaufmann Publishers.
Rashid, A. A. A., \& Buehner, M. J. (2013). Causal reasoning with continuous outcomes. In M. Knauff, M., Pauen, N., Sebanz, \& I. Wachsmuth (Eds.) Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 115-120). Austin TX: Cognitive Science Society.
Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. Journal of Comparative and Physiological Psychology, 66, 1-5.
Rescorla, R. A., \& Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black \& W. F. Prokasy (Eds.), Classical conditioning II: Current research and theory (pp.64-99). New York: Appleton-Century-Crofts.
Saito, M. (2015). How people estimate effect sizes: The role of means and standard deviations. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, \& P. P. Maglio (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp2075-2079) . Austin TX: Cognitive Science Society.
Tenenbaum, J. B., \& Griffiths, T. L. (2001). Structure learning in human causal induction. In T. K. Leen, T. G. Dietterich, \& V. Tresp (Eds.), Advances in neural information processing systems 13 (pp. 59-65). Cambridge, MA: MIT Press.
White, P. A. (2015). Causal judgements about temporal sequences of events in single individuals. The Quarterly Journal of Experimental Psychology, 68, 2149-2174.

# Spatial Training and Mathematics: The Moderating Effect of Handedness 

Giovanni Sala (giovanni.sala@liv.ac.uk)<br>Department of Psychological Sciences, Bedford Street South, University of Liverpool, L69 7ZA, United Kingdom<br>Martina Bolognese (martina.bolognese@gmail.com)<br>Department of Life Sciences, Via Festa del Perdono, University of Milan, 20122, Italy<br>Fernand Gobet (fernand.gobet@liv.ac.uk)<br>Department of Psychological Sciences, Bedford Street South, University of Liverpool, L69 7ZA, United Kingdom


#### Abstract

The positive relationship between spatial ability and mathematical skills is a classical result in developmental and cognitive psychology. Given this correlational relationship, researchers have tried to establish whether spatial training can increase mathematical ability. Such research has provided mixed results. In this study, we analysed the effects of two types of spatial training and handedness on primary school children's arithmetical ability. The participants were pre-tested on a test of arithmetic and assigned to one of three groups: (a) one hour of mental rotation and translation training, (b) one hour of mental translation training only, or (c) a no-contact group. The results showed no significant difference between training groups and a significant interaction between training group and category of handedness. Interestingly, only extremely right-handed children in the mental rotation and translation group seemed to benefit from the training. These outcomes suggest that any spatial training needs to include mental rotation activities to be effective, and that the relationship between spatial training and achievement mathematics appears to be moderated by handedness.


Keywords: mathematics; mental rotation; spatial ability, handedness; STEM.

## Introduction

Concerns have been raised about young people's low achievements in mathematics, both in Europe (Greg, 2009) and the United States (Hanushek, Peterson, \& Woessmann, 2012; Richland, Stigler, \& Holyoak, 2012). Students' insufficient mathematical ability has serious implications, as the likelihood of graduating in Science, Technology, Engineering, and Mathematics (STEM) subjects is limited by one's mathematical ability. The job market increasingly demands workforce with STEM expertise and requires increasingly higher competencies, making competition fiercer worldwide (Halpern et al., 2007).
Students' attainment in mathematics is thus a matter of crucial practical importance. For this reason, an impressive amount of research has been devoted to pinpointing the cognitive correlates of mathematical ability (e.g., Deary, Strand, Smith, \& Fernandes, 2007; Lubinski, 2010; Peng, Namkung, Barnes, \& Sun, 2016; Rohde \& Thompson, 2007;

Wai, Lubinski, \& Benbow, 2009) and finding effective methods to improve students' mathematical skills.

These methods not only include traditional school interventions (for a review, see Hattie, 2009), but also cognitive-training based treatments. Examples of such treatments to foster students' attainment in mathematics and other academic and cognitive skills include working memory training (Sala \& Gobet, 2017a), chess instruction (Gobet \& Campitelli, 2006; Sala, Foley, \& Gobet, 2017; Sala \& Gobet, in press-a; Sala \& Gobet, 2016; Sala, Gobet, Trinchero, \& Ventura, 2016; Sala, Gorini, \& Pravettoni, 2015; Trinchero \& Sala, 2016), and music training (Sala \& Gobet, 2017b). The results show either minimal overall effects on academic achievement and overall cognitive ability (music and working memory training) or medium effects possibly due to placebo effects (chess). These results are in line with Thorndike and Woodworth's (1901) common element theory according to which far transfer - i.e., the generalization of a set of skills across domains only loosely related - rarely occurs (Gobet, 2016; Sala \& Gobet, 2017c; Sala \& Gobet, in press-b).

## Spatial Training

Another, relatively understudied, type of intervention to enhance mathematical ability is spatial training. Spatial training includes activities such as 2D and 3D mental rotation, spatial reasoning and visualizations (Sorby, 2011). However, given the difficulty of far transfer to take place, why should spatial training increase mathematical ability?

## The Relationship between Spatial and Mathematical Abilities

Problem solving in mathematics and STEM disciplines largely relies on spatial ability (Stieff \& Uttal, 2015). Mechanical physics and engineering deal with movement and interaction between elements in a geometrical space. Mathematicians work with functions represented in 2D and 3D space. More generally, several branches of mathematics - necessary to master disciplines such as physics and engineering - require the manipulation of spatial relationships (e.g. geometry, calculus, topology).

The tight relation between spatial ability and mathematical ability has been established empirically. These two separate constructs are highly correlated to each other (Mix et al., 2016). Spatial abilities - such as mental rotation ability (Mix et al., 2016; Wai, Lubinski, \& Benbow, 2009) - are thus strong predictors of achievement in mathematics, both in children (Lauer \& Lourenco, 2016) and in undergraduate and doctorate students (Wai et al., 2009). Thus, several researchers have suggested that training spatial ability causes improvement in mathematics achievement.

## Spatial Training to Train Spatial and STEM Abilities: The Empirical Evidence

Before asking whether spatial training leads to improving mathematical skills such as arithmetic or geometry, one has to verify whether spatial ability can be trained. A metaanalysis carried out by Uttal et al. (2013) suggests that this is the case. Spatial training appears to transfer both to the trained tasks and other spatial tasks not directly trained. ${ }^{1}$ Crucially, from a practical point of view, spatial ability seems to be malleable enough to be significantly boosted by a shortterm training (Uttal et al., 2013).

The evidence supporting the effectiveness of spatial training at improving performance on spatial tasks appears to be quite solid. Regrettably, it is not possible to reach the same conclusion for non-spatial tasks. The research on spatial training to improve STEM achievement has provided promising results, but the number of studies is still relatively limited.
In His, Linn, and Bell (1997), a group of undergraduates improved their attainment in an engineering course after attending a voluntary spatial training (3D orthographic projections). However, the fact that the sample was selfselected casts serious doubts upon the reliability of the outcome. More recently, Sorby (2009) reported that a group of undergraduates in engineering with low spatial ability improved their course grades after spatial training (Sorby, 2011), whereas a control group with no training did not show any amelioration. These positive findings were replicated two years later (Sorby, Casey, Veurink, \& Dulaney, 2013). Less clear were the results in Miller and Halpern's (2013) study. They did find a moderate positive effect after delivering spatial training, but only in items related to Newtonian mechanics. No benefits occurred in other courses.

The studies mentioned above dealt with university students. Cheng and Mix (2014) focused on the effects of short-term ( 40 minutes) spatial training on children's basic arithmetical ability. The training consisted of 40 minutes of mental rotation and mental translation exercises suitable for children (Ehrlich, Levine, \& Goldin-Meadow, 2006). The treatment group showed a small improvement (approximatively $d=0.20$ ) in the test of arithmetic, limited

[^466]to one particular type of items (missing-term problems). A study by Hawes, Moss, Caswell, and Poliszczuk (2015) found no significant effects of mental rotation training on a group of primary school children's arithmetical ability. Finally, Xu and LeFevre (2016) reported no transfer from spatial training to a number line task in a sample of kindergarten children.

## The Potential Moderating Role of Handedness

Several researchers have argued that the relation between mathematical and spatial ability may be moderated by handedness (Casey, Pezaris, \& Nuttall, 1992). Handedness is believed to affect achievement in mathematics because it represents the degree of dominance and development of the right hemisphere, which is involved in cognitive tasks such as spatial reasoning (Ganley \& Vasilyeva, 2011) and mental rotation ability (O'Boyle et al., 2005). Some non-righthanders (i.e., left-handed and ambidextrous people) have a more developed right hemisphere than right-handers (Gutwinski et al., 2011). Such a condition may explain why non-right-handers excel in domains where spatial ability is required. For example, non-right-handers are present among chess players in significantly greater ratio than the general population (Gobet \& Campitelli, 2007). The same pattern has been found in artists (Preti \& Vellante, 2007).

Whether non-right-handers are better than right-handers in mathematics is still a matter of debate (e.g., Benbow, 1986; Cheyne, Roberts, Crow, Leask, \& García-Fiñana, 2010; McManus, 2002). However, it appears that among righthanders, those who show a consistent preference for using the right hand (hereafter, extreme right-handers) underperform in mathematics (e.g., Annett \& Manning, 1989; Cheyne et al., 2010; Peters, 1991). The possible explanation relies again on the degree of development of the right hemisphere in comparison to the left hemisphere. According to Annett (2002), a strong dominance of the left hemisphere may lead to both being extremely right-handed and suffering from some deficits in spatial ability and, hence, in mathematics. In line with this idea, in a recent large study (total $N=2,314$ ), extreme right-handers obtained a poorer score on a variety of tests of mathematics compared to moderate right-handers (Sala, Signorelli, Barsuola, Bolognese, \& Gobet, submitted).

## The Present Study

In this study, we replicated and extended Cheng and Mix's (2014) study. There were two crucial additions. First, we tested whether the effects of training (if any) on mathematical ability interact with handedness. We expected extreme righthanders to perform more poorly on the pre-test of arithmetic than the moderate right-handers and non-right-handers. Most importantly, given that the extreme right-handers are believed to have a lower mathematical ability because of a
variety of tasks may stem from some general ability at performing spatial tasks (e.g., better strategies). In any case, this important theoretical issue is beyond the aims of this article.
spatial deficit, we also expected them to benefit most from the spatial training task. Second, along with the treatment and no-contact groups, another treatment group practicing only mental translation was included. The rationale was to understand whether mental translation training alone (i.e., no mental rotation) could positively influence attainment in mathematics (see below for details).

## Method

## Participants

A total of 159 first, second, and third graders in nine classes of a primary school in northern Italy took part in this experiment. The mean age of the participants was 7.61 years ( $\mathrm{SD}=0.89$ ). Parental consent was asked and obtained for all the participants.

## Materials

The participants were administered (a) the Edinburgh Handedness Inventory (EHI), ${ }^{2}$ (b) a spatial ability task (mental rotation and translation) suitable for children (score range $0-16$; Ehrlich et al., 2006), and (c) a test of arithmetic, designed by the experimenters (score range $0-27$; Cronbach Alpha = .96).

EHI is a multiple-item questionnaire that provides a continuous measure of handedness ( $h$ ), which is calculated using the formula $h=\frac{R-L}{R+L}$, where R and L indicate the number of preferences for the right and left hand, respectively. The range of values is between -1 , for extreme left-handedness, and +1 , for extreme right-handedness. The participants were categorized according to their $h$-values (Casey, 1995):
a) Extreme right-handers: $h \geq .90(N=48)$.
b) Moderate right-handers: $.40<h<.90(N=81)$.
c) Non-right-handers: $h \leq .40(N=30)$.

The test of mental rotation and translation ability consists of 16 items. The participant is shown four whole pictures and two parts of a flat shape. The participant has to mentally put the two pieces together and choose one of the four whole pictures (Figure 1).

In the test of arithmetic, finally, the participants solved simple mathematical equations (e.g., $3+4=$ ?) and missingterm problems (e.g., $3+$ ? $=7$ ).

## Design

All the nine classes were pre-tested in arithmetic, spatial ability (rotation/translation), and EHI. A week later, the nine classes were randomly assigned to three groups:
a) Three classes (one first-, one second-, one thirdgrade) attended 60 minutes of mental rotation and translation exercises. This training consisted of 16

[^467]rotation items and 16 translation items. Analogously to the testing session, the participants were asked to choose one of the pictures. Finally, children were given the two parts of the picture on separate pieces of cardstock and requested to confirm or change the choice after putting them together (Figure 2).
b) Three classes (one first-, one second-, one thirdgrade) attended 60 minutes of 32 translation exercises only (translation group; Figure 3). The training procedure was analogous to that in the previous group.
c) Three classes (one first-, one second-, one thirdgrade) did not carry any activity (no-contact group).


Figure 1. Two examples of the items used in the spatial test (translation and rotation).


Figure 2. An example of rotation item used in the fulltraining group.


Figure 3. An example of translation item used in both the training groups.

Finally, the three groups were post-tested in arithmetic and spatial ability immediately after the end of the training.

## Results

## Preliminary Analyses

The three groups did not differ in terms of age ( $p=.970$ ) or pre-test arithmetic test scores $(p=.391)$. As expected, the category of handedness had a significant effect on the pre-test scores in arithmetic $(F(2,156)=6.50, p=.002)$. Extreme right-handers were outperformed by both moderate righthanders ( $p<.001$ ) and non-right-handers ( $p=.048$ ).

The pre-post test correlations for arithmetical ability and spatial ability were $r=.94$ and $r=.60$, respectively (both $p \mathrm{~s}$ $<.001$ ).

Finally, an ANCOVA (pre-test scores as the covariate) confirmed that the spatial training had a significant effect on the score of spatial ability. In fact, the two training groups (full and translation) outperformed the no-contact group ( $p=$ .030 and $p=.004$, respectively).

## Main Analysis: Scores in Arithmetical Ability

The pre- test and post-test scores in arithmetical ability are summarized in Table 1.

Table 1. Scores in arithmetical ability the three groups.

| Group | $N$ | Pre-test | Post-test |
| :--- | ---: | ---: | ---: |
| Full-training | 56 | $16.73(9.20)$ | $18.04(7.98)$ |
| Translation | 53 | $18.81(7.02)$ | $18.79(6.79)$ |
| No-contact | 50 | $17.18(8.32)$ | $18.40(8.17)$ |

Note. Standard deviations are shown in brackets.
An ANCOVA (Table 2) was run to analyse the effects of the independent variables (group and category of handedness) on the results of the post-test of arithmetical ability, using pretest score and age as covariates. The results showed no
significant effect of age (in years, $p=.176$ ), category of handedness ( $h$-cat; $p=.846$ ), or group ( $p=.491$ ). As expected, a significant effect of the pre-test scores was found ( $p<.001$ ). Interestingly, a significant interaction between group and category of handedness was reported ( $p=.013$ ).

Table 2. The ANCOVA model of the scores in arithmetic

| Variable | $D f$ | $F$-value | $p$-value |
| :--- | ---: | ---: | ---: |
| Group | 2 | 0.72 | .491 |
| $h$-cat | 2 | 0.17 | .846 |
| Age | 1 | 1.85 | .176 |
| Pre-test scores | 1 | 333.69 | $\mathbf{. 0 0 0}$ |
| Group* $h$-cat | 4 | 3.27 | $\mathbf{. 0 1 3}$ |

The extreme right-handers in the full-training group showed the greatest mean improvement in the test of arithmetic compared to the other extreme right-handers. The pre- and post-test mean scores are summarized in Table 3.

Table 3. Extreme right-handers' scores in the three groups.

| Group | $N$ | Pre-test | Post-test |
| :--- | ---: | ---: | ---: |
| Full-training | 14 | $10.71(8.32)$ | $13.57(7.80)$ |
| Translation | 17 | $18.41(7.98)$ | $18.47(7.54)$ |
| No-contact | 17 | $12.88(8.35)$ | $14.76(8.88)$ |

Note. Standard deviations are shown in brackets.

## Discussion

The results of this experiment show no significant impact of the one-hour spatial training on children's arithmetical ability. In fact, most of the variance in the post-test of arithmetical ability is explained by pre-test scores $\left(r=.94 ; r^{2}\right.$ $\times 100=88 \%$ ). This outcome is in accordance with previous experimental studies (e.g., Cheng \& Mix, 2014; Hawes et al., 2015; Xu \& LeFevre, 2016) examining the effects of spatial training on arithmetical ability. In a wider perspective, our results are consistent with substantial research on far transfer (Burgoyne et al., 2016; Sala et al., 2017; Sala \& Gobet, 2016, 2017a, 2017b).

However, the significant role played by handedness, predicted by Annett (2002), sheds light on the potential benefits of spatial training on arithmetical ability. The extreme right-handers in the full-training group reported the best improvement in mathematical ability compared to both the whole groups and the sub-samples of extreme righthanders. This pattern of results suggests that short-term mental spatial training may be effective for a particular subsample of underachievers in arithmetic (i.e., extreme right-handers), as long as mental rotation activities are included.

## Recommendations for Future Research

This study highlights the possible benefits of mental rotation training for extreme right-handers' arithmetical ability. In order to confirm (or disconfirm) our results, future investigations should replicate and extend the design of the
current study. First, even if the total sample was large ( $N=$ 159), the subgroup of extreme right-handers consisted of only a few tens of individuals $(N=48)$ distributed across three groups. Given the importance of that subgroup for the main hypothesis of this study, future experiments should include more participants (e.g., as twice as many) to increase the statistical power, and hence the reliability, of the analysis and outcomes. Second, the future investigations should systematically manipulate the duration of training and administer both immediate and delayed post-test assessments. This way, it would be possible to evaluate whether the effects of spatial training on extreme righthanders' on mathematical ability increase with the duration of training and last after its end. Third, we collected only one measure of mathematical ability (i.e., arithmetic). The use of multivariate measures of mathematical ability and spatial ability would contribute to establishing whether spatial training benefits for extreme right-handers goes beyond basic arithmetic ability.

## Acknowledgments

The authors gratefully thank Yi-Ling Cheng and Kelly Mix for the assistance in designing the experiment, and the principal and all the teachers in the project for their support.

## References

Annett, M. (2002). Handedness and brain asymmetry: The right shift theory. Hove, East Sussex, UK: Psychology Press.
Annett, M., \& Manning, M. (1989). The disadvantage of dextrality for intelligence. British Journal of Psychology, 80, 213-226.
Benbow, C. P. (1986). Physiological correlates of extreme intellectual precocity. Neuropsychologia, 24, 719-725.
Burgoyne, A. P., Sala, G., Gobet, F., Macnamara, B. N., Campitelli, G., \& Hambrick, D. Z. (2016). The relationship between cognitive ability and chess skill: A comprehensive meta-analysis. Intelligence, 59, 72-83.
Casey, M. B. (1995). Empirical support for Annett's conception of the heterozygotic advantage. Cahiers de Psychologie Cognitive, 14, 520-528.
Casey, M. B., Pezaris, E., \& Nuttall, R. L. (1992). Spatial ability as a predictor of math achievement: The importance of sex and handedness patterns. Neuropsychologia, 30, 3545.

Cheng, Y. L., \& Mix, K. S. (2014). Spatial training improves children's mathematics ability. Journal of Cognition and Development, 15, 2-11.
Cheyne, C. P., Roberts, N., Crow, T. J., Leask, S. J., \& García-Fiñana, M. (2010). The effect of handedness on academic ability: A multivariate linear mixed model approach. Laterality, 15, 451-464.
Deary, I. J., Strand, S., Smith, P., \& Fernandes, C. (2007). Intelligence and educational achievement. Intelligence, 35, 13-21.

Ehrlich, S., Levine, S., \& Goldin-Meadow, S. (2006). The importance of gestures in children's spatial reasoning. Developmental Psychology, 42, 1259-1268.
Ganley, C. M., \& Vasilyeva, M. (2011). Sex differences in the relation between math performance, spatial skills, and attitudes. Journal of Applied and Developmental Psychology, 32, 235-242.
Gobet, F. (2016). Understanding expertise: A multidisciplinary approach. London: Palgrave/Macmillan.
Gobet, F., \& Campitelli, G. (2006). Educational benefits of chess instruction. A critical review. In T. Redman (Ed.), Chess and education. Selected essays from the Koltanowski Conference (pp. 124-143). Dallas, TX: University of Texas at Dallas.
Gobet, F., \& Campitelli, G. (2007). The role of domainspecific practice, handedness and starting age in chess. Developmental Psychology, 43, 159-172.
Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. Journal of Education Policy, 24, 23-37.
Groen, M. A., Whitehouse, A. J. O., Badcock, N. A., \& Bishop, D. V. M. (2013). Associations between handedness and cerebral lateralisation for language: A comparison of three measures in children. PLoS ONE 8: e64876.
Gutwinski, S., Löscher, A., Mahler, L., Kalbitzer, J., Heinz, A., \& Bermpohl, F. (2011). Understanding lefthandedness. Deutsches Arzteblatt International, 108, 849853.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., \& Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. Psychological Science in the Public Interest, 8, 1-51.
Hanushek, E. A., Peterson, P. E., \& Woessmann, L. (2012). Achievement growth: International and US state trends in student performance. Harvard's Program on Education Policy and Governance.
Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. New York: Routledge.
Hawes, Z., Moss, J., Caswell, B., \& Poliszczuk, D. (2015). Effects of mental rotation training on children's spatial and mathematics performance: A randomized controlled study. Trends in Neuroscience and Education, 4, 60-68.
Hsi, S., Linn, M. C., \& Bell, J. E. (1997). The role of spatial reasoning in engineering and the design of spatial instruction. Journal of Engineering Education, 86, 151158.

Lauer J. E., \& Lourenco, S. F. (2016). Spatial processing in infancy predicts both spatial and mathematical aptitude in childhood. Psychological Science, 27, 1-8.
Lubinski, D. (2010). Spatial ability and STEM: A sleeping giant for talent identification and development. Personality and Individual Differences, 49, 344-351.
McManus, I. C. (2002). Right hand, left hand. London: Orion Books Ltd.
Miller, D. I., \& Halpern, D. F. (2013). Can spatial training improve long-term outcomes for gifted STEM
undergraduates? Learning and Individual Differences, 26, 141-152.
Mix, K. S., Levine, S. C., Cheng, Y. L., Young, C., Hambrick, D. Z., Ping, R., \& Konstantopoulos, S. (2016). Separate but correlated: The latent structure of space and mathematics across development. Journal of Experimental Psychology: General, 145, 1206-1227.
O’Boyle, M. W., Cunnington, R., Silk, T. J., Vaughan, D., Jackson, G., Syngeniotis, A., \&, Egan, G. F. (2005). Mathematically gifted male adolescents activate a unique brain network during mental rotation. Cognitive Brain Research, 25, 583-587.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. Neuropsychologia, 9, 97-113.
Peng, P., Namkung, J., Barnes, M., \& Sun, C. Y. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. Journal of Educational Psychology, 108, 455-473.
Peters, M. (1991). Sex, handedness, mathematical ability, and biological causation. Canadian Journal of Psychology, 45, 415-419.
Preti, A., \& Vellante, M. (2007). Creativity and psychopathology: Higher rates of psychosis proneness and nonright-handedness among creative artists compared to same age and gender peers. Journal of Nervous \& Mental Disease, 195, 837-845.
Richland, L. E., Stigler, J. W., \& Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. Educational Psychologist, 47, 189-203.
Rohde, T. E., \& Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. Intelligence, 35, 83-92.
Sala, G., Burgoyne, A. P., Macnamara, B. N., Hambrick, D. Z., Campitelli, G., \& Gobet, F., (2017). Checking the "academic selection" argument. Chess players outperform non-chess players in cognitive skills related to intelligence: A meta-analysis. Intelligence, 61, 130-139.
Sala, G., Foley, J., \& Gobet, F. (2017). The effect of chess instruction on pupils' cognitive and academic skills: State of the art and theoretical challenges. Frontiers in Psychology, 8:238.
Sala, G., \& Gobet, F. (in press-a). Does chess instruction improve mathematical problem-solving ability? Two experimental studies with an active control group. Learning \& Behavior.
Sala, G., \& Gobet, F. (in press-b). Does far transfer exist? Negative evidence from chess, music, and working memory training. Current Directions in Psychological Science.
Sala, G., Gobet, F. (2016). Do the benefits of chess instruction transfer to academic and cognitive skills? A meta-analysis. Educational Research Review, 18, 46-57.
Sala, G., \& Gobet, F. (2017a). Working memory training in typically developing children: A meta-analysis of the
available evidence. Developmental Psychology, 53, 671685.

Sala, G., \& Gobet, F. (2017b). When the music's over. Does music skill transfer to children's and young adolescents' cognitive and academic skills? A meta-analysis. Educational Research Review, 20, 55-67.
Sala, G., \& Gobet, F. (2017c). Experts' memory superiority for domain-specific random material generalizes across fields of expertise: A meta-analysis. Memory \& Cognition, 45, 183-193.
Sala, G., Gobet, F., Trinchero, R., \& Ventura, S. (2016). Does chess instruction enhance mathematical ability in children? A three-group design to control for placebo effects. Proceedings of the $38^{\text {th }}$ Annual Meeting of the Cognitive Science Society.
Sala, G., Gorini, A., \& Pravettoni, G. (2015). Mathematical problem-solving abilities and chess. SAGE Open, 5.
Sala, G., Signorelli, M., Barsuola, G., Bolognese, M., \& Gobet, F. (submitted). The relationship between handedness and mathematics is non-linear and is moderated by gender, age, and type of task.
Shipstead, Z., Redick, T. S., \& Engle, R. W. (2012). Is working memory training effective? Psychological Bulletin, 138, 628-654.
Sorby, S. A. (2009). Education research in developing 3-D spatial skills for engineering students. International Journal of Science Education, 31, 459-480.
Sorby, S. A. (2011). Developing spatial thinking. Independence: Cengage.
Sorby, S. A., Casey, B., Veurink, N., \& Dulaney, A. (2013). The role of spatial training in improving spatial and calculus performance in engineering students. Learning and Individual Differences, 26, 20-29.
Stieff, M., \& Uttal, D. (2015). How much can spatial training improve STEM achievement? Educational Psychology Review, 27, 607-615.
Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., \& Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. Psychological Bulletin, 139, 352-402.
Trinchero, R., \& Sala, G. (2016). Chess training and mathematical problem-solving: The role of teaching heuristics in transfer of learning. Eurasia Journal of Mathematics, Science \& Technology Education, 12, 655668.

Wai, J., Lubinski, D., \& Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. Journal of Educational Psychology, 101, 817-835.
Xu, C., \& LeFevre, J. A. (2016). Training young children on sequential relations among numbers and spatial decomposition: Differential transfer to number line and mental transformation tasks. Developmental Psychology, 52, 854-866.

# Dissolving the Grounding Problem: How the Pen is Mightier than the Sword 

Nancy Salay<br>Queen's University


#### Abstract

The computational metaphor for mind is still the central guiding idea in cognitive science despite many insightful and well-founded rejections of it. There is good reason for its staying power: when we are at our cognitive best, we reason about our world with our concepts. But the challengers are right, I argue, in insisting that no reductive account of that capacity is forthcoming. Here I describe an externalist account that grounds representations in organism-level engagement with its environment, not in its neural activity.


# The Causal Frame Problem: An Algorithmic Perspective 

Ardavan S. Nobandegani ${ }^{1,2} \quad$ Ioannis N. Psaromiligkos ${ }^{1}$<br>\{ardavan.salehinobandegani@mail.mcgill.ca, yannis@ece.mcgill.ca\}<br>${ }^{1}$ Department of Electrical \& Computer Engineering, McGill University<br>${ }^{2}$ Department of Psychology, McGill University


#### Abstract

The Frame Problem (FP) is a puzzle in philosophy of mind and epistemology, articulated by the Stanford Encyclopedia of Philosophy as follows: "How do we account for our apparent ability to make decisions on the basis only of what is relevant to an ongoing situation without having explicitly to consider all that is not relevant?" In this work, we focus on the causal variant of the FP, the Causal Frame Problem (CFP). Assuming that a reasoner's mental causal model can be (implicitly) represented by a causal Bayes net, we first introduce a notion called Potential Level (PL). PL, in essence, encodes the relative position of a node with respect to its neighbors in a causal Bayes net. Drawing on the psychological literature on causal judgment, we substantiate the claim that PL may bear on how time is encoded in the mind. Using PL, we propose an inference framework, called the PL-based Inference Framework (PLIF), which permits a boundedly-rational approach to the CFP, formally articulated at Marr's algorithmic level of analysis. We show that our proposed framework, PLIF, is consistent with several findings in the causal judgment literature, and that PL and PLIF make a number of predictions, some of which are already supported by existing findings.


Keywords: Frame Problem; Time and Causality; Bounded Rationality; Algorithmic Level of Analysis; Rational Process Models

## 1 Introduction

At the core of any decision-making or reasoning task, resides an innocent-looking yet challenging question: Given an inconceivably large body of knowledge available to the reasoner, what constitutes the relevant for the task and what the irrelevant? The question, as it is posed, echoes the wellknown Frame Problem (FP) in epistemology and philosophy of mind, articulated by Glymour (1987) as follows: "Given an enormous amount of stuff, and some task to be done using some of the stuff, what is the relevant stuff for the task?" Fodor (1987) comments: "The frame problem goes very deep; it goes as deep as the analysis of rationality."

The question posed above perfectly captures what is really at the core of the FP, yet, it may suggest an unsatisfying approach to the FP at the algorithmic level of analysis (Marr, 1982). Indeed, the question may suggest the following twostep methodology: In the first step, out of all the body of knowledge available to the reasoner (termed, the model), she has to identify what is relevant to the task (termed, the relevant submodel); it is only then that she advances to the second step by performing reasoning or inference on the identified submodel. There is something fundamentally wrong with this methodology (which we term, sequential approach to reasoning) which bears on the following understanding: The relevant submodel, i.e., the portion of the reasoner's knowledge deemed relevant to the task, oftentimes is so enormous (or even infinitely large) that the reasoner-inevitably bounded
in time and computational resources-would never get to the second step, had she adhered to such a methodology. In other words, in line with the notion of bounded rationality (Simon, 1957), a boundedly-rational reasoner must have the option, if need be, to merely consult a fraction of the potentially largeif not infinitely so-relevant submodel.

Icard and Goodman (2015) elegantly promote this insight when they write: "Somehow the mind must focus in on some "submodel" of the "full" model (including all possibly relevant variables) that suffices for the task at hand and is not too costly to use." ${ }^{1}$ They then ask the following question: "what kind of simpler model should a reasoner consult for a given task?" This is an inspiring question hinting to an interesting line of inquiry as to how to formally articulate a boundedlyrational approach to the FP, at Marr's (1982) algorithmic level of analysis.

In this work, we focus on the causal variant of the FP, the Causal Frame Problem (CFP), stated as follows: Upon being presented with a causal query, how does the reasoner manage to attend to her causal knowledge relevant to the derivation of the query while rightfully dismissing the irrelevant? We adopt Causal Bayesian Networks (CBNs) (Pearl, 1988; Gopnik et al., 2004, inter alia) as a normative model to represent how the reasoner's internal causal model of the world is structured (i.e., reasoner's mental model). First, we introduce the notion of Potential Level (PL). PL, in essence, encodes the relative position of a node (representing a propositional variable or a concept) with respect to its neighbors in a CBN. Drawing on the psychological literature on causal judgment, we substantiate the claim that PL may bear on how time is encoded in the mind. Equipped with PL, we embark on investigating the CFP at Marr's algorithmic level of analysis. We propose an inference framework, termed PL-based Inference Framework (PLIF), which aims at empowering the boundedly-rational reasoner to consult (or retrieve ${ }^{2}$ ) parts of the underlying CBN deemed relevant for the derivation of the posed query (the relevant submodel) in a local, bottom$u p$ fashion until the submodel is fully retrieved. PLIF allows the reasoner to carry out inference at intermediate stages of the retrieval process over the thus-far retrieved parts, thereby obtaining lower and upper bounds on the posed causal query.

[^468]We show, in the Discussion section, that our proposed framework, PLIF, is consistent with several findings in the causal judgment literature, and that PL and PLIF make a number of predictions, some of which are already supported by the findings in the psychology literature.

In their work, Icard and Goodman (2015) articulate a boundedly-rational approach to the CFP at Marr's computational level of analysis, which, as they point out, is from a "god's eye" point of view. In sharp contrast, our proposed framework PLIF is not from a "god's eye" point of view and hence could be regarded, potentially, as a psychologically plausible proposal at Marr's algorithmic level of analysis as to how the mind both retrieves and, at the same time, carries out inference over the retrieved submodel to derive bounds on a causal query. We term this concurrent approach to reasoning, as opposed to the flawed sequential approach stated earlier. ${ }^{3}$ The retrieval process progresses in a local, bottomup fashion, hence the submodel is retrieved incrementally, in a nested manner. ${ }^{4}$ Our analysis (Sec. 4.1) confirms Icard and Goodman's (2015) insight that even in the extreme case of having an infinitely large relevant submodel, the portion of which the reasoner has to consult so as to obtain a "sufficiently good" answer to a query could indeed be very small.

## 2 Potential Level and Time

Before proceeding further, let us introduce some preliminary notations. Random Variables (RVs) are denoted by lowercase bold-faced letters, e.g., $\mathbf{x}$, and their realizations by nonbold lower-case letters, e.g., $x$. Likewise, sets of RVs are denoted by upper-case bold-faced letters, e.g., $\mathbf{X}$, and their corresponding realizations by upper-case non-bold letters, e.g., $X . \operatorname{Val}(\cdot)$ denotes the set of possible values a random quantity can take on. Random quantities are assumed to be discrete unless stated otherwise. The joint probability distribution over $\mathbf{x}_{1}, \cdots, \mathbf{x}_{n}$ is denoted by $\mathbb{P}\left(\mathbf{x}_{1}, \cdots, \mathbf{x}_{n}\right)$. We will use the notation $\mathbf{x}_{1: n}$ to denote the sequence of $n \operatorname{RVs} \mathbf{x}_{1}, \cdots, \mathbf{x}_{n}$, hence $\mathbb{P}\left(\mathbf{x}_{1}, \cdots, \mathbf{x}_{n}\right)=\mathbb{P}\left(\mathbf{x}_{1: n}\right)$. The terms "node" and "variable" will be used interchangeably. To simplify presentation, we adopt the following notation: We denote the probability $\mathbb{P}(\mathbf{x}=x)$ by $\mathbb{P}(x)$ for some $\mathrm{RV} \mathbf{x}$ and its realization $x \in \operatorname{Val}(\mathbf{x})$. For conditional probabilities, we will use the notation $\mathbb{P}(x \mid y)$ instead of $\mathbb{P}(\mathbf{x}=x \mid \mathbf{y}=y)$. Likewise, $\mathbb{P}(X \mid Y)=\mathbb{P}(\mathbf{X}=X \mid \mathbf{Y}=Y)$ for $X \in \operatorname{Val}(\mathbf{X})$ and $Y \in \operatorname{Val}(\mathbf{Y})$. A generic conditional independence relationship is denoted by $(\mathbf{A} \Perp \mathbf{B} \mid \mathbf{C})$ where $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ represent three mutually disjoint sets of variables belonging to a CBN. Furthermore, throughout the paper, we assume that $\varepsilon$ is some negligibly small positive real-valued quantity. Whenever we subtract $\varepsilon$ from a quantity, we simply imply a quantity less than but arbitrarily close to the original quantity. The rationale behind adopting such a notation will become clearer in Sec. 4.

[^469]Before formally introducing the notion of PL, we articulate in simple terms what the idea behind PL is. PL simply induces a chronological order on the nodes of a CBN, allowing the reasoner to encode the timing between cause and effect. ${ }^{5}$ As we will see, PL plays an important role in guiding the retrieval process used in our proposed framework. Next, PL is formally defined, followed by two clarifying examples.

Def. 1. (Potential Level (PL)) Let $\operatorname{par}(\mathbf{x})$ and $\operatorname{child}(\mathbf{x})$ denote, respectively, the sets of parents (i.e., immediate causes) and children (i.e., immediate effects) of $\mathbf{x}$. Also let $T_{0} \in \mathbb{R} \cup\{-\infty\}$. The PL of $\mathbf{x}$, denoted by $p_{l}(\mathbf{x})$, is defined as follows: (i) If $\operatorname{par}(\mathbf{x})=\varnothing, p_{l}(\mathbf{x})=T_{0}$, and (ii) If $\operatorname{par}(\mathbf{x}) \neq \varnothing, p_{l}(\mathbf{x})$ is a real-valued quantity selected from the interval $\left(\max _{\mathbf{y} \in \operatorname{par}(\mathbf{x})} p_{l}(\mathbf{y}), \min _{\mathbf{z} \in \operatorname{child}(\mathbf{x})} p_{l}(\mathbf{z})\right)$ such that $p_{l}(\mathbf{x})-\max _{\mathbf{y} \in \operatorname{par}(\mathbf{x})} p_{l}(\mathbf{y})$ indicates the amount of time which elapses between intervening simultaneously on all the RVs in $\operatorname{par}(\mathbf{x})$ (i.e., $\operatorname{do}\left(\operatorname{par}(\mathbf{x})=\operatorname{par}_{x}\right)$ ) and $\mathbf{x}$ taking its value $x$ in accord with the distribution $\mathbb{P}\left(x \mid \operatorname{par}_{x}\right)$. If $\operatorname{child}(\mathbf{x})=\varnothing$, substitute the upper bound of the given interval by $+\infty$.

Parameter $T_{0}$ symbolizes the origin of time, as perceived by the reasoner. $T_{0}=0$ is a natural choice, unless the reasoner believes that time continues unboundedly into the past, in which case $T_{0}=-\infty$. The next two examples further clarify the idea behind PL. In both examples we assume $T_{0}=0$.

(a)

(b)

Figure 1: Relation between PL and time. Three hollow dots signify that the depicted CBNs extend into the past and future.

For the first example, let us consider the CBN depicted in Fig. 1(a) containing the RVs $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ with $p_{l}(\mathbf{x})=$ $4, p_{l}(\mathbf{y})=4.7$, and $p_{l}(\mathbf{z})=5$. According to Def. 1, the given PLs can be construed in terms of the relative time between the occurrence of cause and effect as articulated next. Upon intervening on $\mathbf{x}$ (i.e., $d o(\mathbf{x}=x)$ ), after the elapse of $p_{l}(\mathbf{y})-p_{l}(\mathbf{x})=0.7$ units of time, the RV $\mathbf{y}$ takes its value $y$ in accord with the distribution $\mathbb{P}(y \mid x)$. Likewise, upon intervening on $\mathbf{y}$ (i.e., $d o(\mathbf{y}=y)$ ), after the elapse of $p_{l}(\mathbf{z})-p_{l}(\mathbf{y})=$ 0.3 units of time, $\mathbf{z}$ takes its value $z$ according to $\mathbb{P}(z \mid y)$.

For the second example, consider the CBN depicted in Fig. 1(b) containing the RVs $\mathbf{x}, \mathbf{y}, \mathbf{z}$, and $\mathbf{t}$ with $p_{l}(\mathbf{x})=4, p_{l}(\mathbf{y})=$ 4.7, $p_{l}(\mathbf{z})=5$, and $p_{l}(\mathbf{t})=5.6$. Upon intervening on $\mathbf{x}$ (i.e., $d o(\mathbf{x}=x))$ the following happens: (i) after the elapse of $p_{l}(\mathbf{y})-p_{l}(\mathbf{x})=0.7$ units of time, $\mathbf{y}$ takes its value $y$ according to $\mathbb{P}(y \mid x)$, and (ii) after the elapse of $p_{l}(\mathbf{z})-p_{l}(\mathbf{x})=1$ unit of

[^470]time, $\mathbf{z}$ takes its value $z$ according to $\mathbb{P}(z \mid x)$. Also, upon intervening simultaneously on RVs $\mathbf{y}, \mathbf{z}$ (i.e., $d o(\mathbf{y}=y, \mathbf{z}=z)$ ), after the elapse of $p_{l}(\mathbf{t})-\max _{\mathbf{r} \in \operatorname{par}(\mathbf{t})} p_{l}(\mathbf{r})=0.6$ units of time, $\mathbf{t}$ takes its value $t$ according to $\mathbb{P}(t \mid y, z)$.

In sum, the notion of PL bears on the underlying time-grid upon which a CBN is constructed, and adheres to Hume's principle of temporal precedence of cause to effect (Hume, 1748/1975). A growing body of work in psychology literature corroborates Hume's centuries-old insight, suggesting that the timing and temporal order between events strongly influences how humans induce causal structure over them (Bramley, Gerstenberg, \& Lagnado, 2014; Lagnado \& Sloman, 2006). The introduced notion of PL is based on the following hypothesis: When learning the underlying causal structure of a domain, humans may as well encode the temporal patterns (or some estimates thereof) on which they rely to infer the causal structure. This hypothesis is supported by recent findings suggesting that people have expectations about the delay length between cause and effect (Greville \& Buehner, 2010; Buehner \& May, 2004; Schlottmann, 1999). It is worth noting that we could have defined PL in terms of relative expected time between cause and effect, rather than relative absolute time. Under such an interpretation, the time which elapses between the intervention on a cause and the occurrence of its effect would be modeled by a probability distribution, and PL would be defined in terms of the expected value of that distribution. Our proposed framework, PLIF, is indifferent as to whether PL should be construed in terms of absolute or expected time. Greville and Buehner (2010) show that causal relations with fixed temporal intervals are consistently judged as stronger compared to those with variable temporal intervals. This finding, therefore, seems to suggest that people expect, to a greater extent, fixed temporal intervals between cause and effect, rather than variable ones-an interpretation which, at least to a first approximation, favors construing PL in terms of relative absolute time (see Def. 1). ${ }^{6}$

## 3 Informative Example

To develop our intuition, and before formally articulating our proposed framework, let us present a simple yet informative example which demonstrates: (i) how the retrieval process can be carried out in a local, bottom-up fashion, allowing for retrieving the relevant submodel incrementally, and (ii) how adopting PL allows the reasoner to obtain bounds on a given causal query at intermediate stages of the retrieval process.

Let us assume that the posed causal query is $\mathbb{P}(x \mid y)$ where $\mathbf{x}, \mathbf{y}$ are two RVs in the CBN depicted in Fig. 2(a) with PLs $p_{l}(\mathbf{x}), p_{l}(\mathbf{y})$, and let $p_{l}(\mathbf{x})>p_{l}(\mathbf{y})$. The relevant information for the derivation of the posed query (i.e., the relevant submodel) is depicted in Fig. 2(e).

[^471]

Figure 2: Example. Query variables are shown in orange.

Starting from the target RV $\mathbf{x}$ in the original CBN (Fig. 2(a)) and moving one step backwards, ${ }^{7} \mathbf{t}_{1}$ is reached (Fig. 2(b)). Since $p_{l}(\mathbf{y})<p_{l}\left(\mathbf{t}_{1}\right)$, $\mathbf{y}$ must be a non-descendant of $\mathbf{t}_{1}$, and therefore, of $\mathbf{x}$. Hence, conditioning on $\mathbf{t}_{1} d$-separates $\mathbf{x}$ from $\mathbf{y}$ (Pearl, 1988), yielding $\left(\mathbf{x} \Perp \mathbf{y} \mid \mathbf{t}_{1}\right)$. Thus $\mathbb{P}(x \mid y)=$ $\sum_{t_{1} \in \mathbf{V a l}\left(t_{1}\right)} \mathbb{P}\left(x \mid y, t_{1}\right) \mathbb{P}\left(t_{1} \mid y\right)=\sum_{t_{1} \in \mathbf{V a l}\left(t_{1}\right)} \mathbb{P}\left(x \mid t_{1}\right) \mathbb{P}\left(t_{1} \mid y\right)$ implying: $\min _{t_{1} \in \operatorname{Val}\left(\mathbf{t}_{1}\right)} \mathbb{P}\left(x \mid t_{1}\right) \leq \mathbb{P}(x \mid y) \leq \max _{t_{1} \in \operatorname{Val}\left(\mathbf{t}_{1}\right)} \mathbb{P}\left(x \mid t_{1}\right)$. It is crucial to note that the given bounds can be computed using the information thus-far retrieved, i.e., the information encoded in the submodel shown in Fig. 2(b). Taking a step backwards from $\mathbf{t}_{1}, \mathbf{t}_{2}$ is reached (Fig. 2(c)). Using a similar line of reasoning to the one presented for $\mathbf{t}_{1}$, having $p_{l}(\mathbf{y})<p_{l}\left(\mathbf{t}_{2}\right)$ ensures $\left(\mathbf{x} \Perp \mathbf{y} \mid \mathbf{t}_{2}\right)$. Therefore, the following bounds on the posed query can be derived, which, crucially, can be computed using the information thus-far retrieved: $\min _{t_{2} \in \operatorname{Val}\left(\mathbf{t}_{2}\right)} \mathbb{P}\left(x \mid t_{2}\right) \leq \mathbb{P}(x \mid y) \leq \max _{t_{2} \in \operatorname{Val}\left(\mathbf{t}_{2}\right)} \mathbb{P}\left(x \mid t_{2}\right)$. It is straightforward to show that the bounds derived in terms of $\mathbf{t}_{2}$ are equally tight or tighter than the bounds derived in terms of $\mathbf{t}_{1}$. Finally, taking one step backward from $\mathbf{t}_{2}, \mathbf{y}$ is reached (Fig. 2(d)) and the exact value for $\mathbb{P}(x \mid y)$ can be derived, again using the submodel thus-far retrieved (Fig. 2(d)).

We are now ready to present our proposed framework.

## 4 PL-based Inference Framework (PLIF)

In this section, we intend to elaborate on how, equipped with the notion of PL, a generic causal query of the form ${ }^{8}$ $\mathbb{P}(\mathbf{O}=O \mid \mathbf{E}=E)$ can be derived where $\mathbf{O}$ and $\mathbf{E}$ denote, respectively, the disjoint sets of target (or objective) and observed (or evidence) variables. In other words, we intend to formalize how inference over a CBN whose nodes are endowed with PL as an attribute should be carried out. Before we present the main result, a few definitions are in order.

Def. 2. (Critical Potential Level (CPL)) The target variable with the least PL is denoted by $\mathbf{o}^{*}$ and its PL is referred to as the CPL. More formally, $p_{l}^{*}: \triangleq \min _{\mathbf{0} \in \mathbf{O}} p_{l}(\mathbf{0})$ and

[^472]$\mathbf{o}^{*}: \triangleq \arg \min _{\mathbf{0} \in \mathbf{O}} p_{l}(\mathbf{0})$. E.g., for the setting given in Fig. 2(a), $\mathbf{o}^{*}=\mathbf{x}$, and $p_{l}^{*}=p_{l}(\mathbf{x})$. Viewed through the lens of time, $\mathbf{o}^{*}$ is the furthest target variable into the past, with PL $p_{l}^{*}$.

There are two possibilities: (a) $p_{l}^{*}>T_{0}$, or (b) $p_{l}^{*}=T_{0}$, with $T_{0}$ denoting the origin of time; cf. Sec. 2. In the sequel, we assume that (a) holds. ${ }^{9}$

Def. 3. (Inference Threshold (IT) and IT Root Set (ITRS)) To any real-valued quantity, $\mathcal{T}$, corresponds a unique set, $\mathbf{R}_{\mathcal{T}}$, obtained as follows: Start at every variable $\mathbf{x} \in \mathbf{O} \cup \mathbf{E}$ with PL $\geq \mathcal{T}$ and backtrack along all paths terminating at $\mathbf{x}$. Backtracking along each path stops as soon as a node with PL less than $\mathcal{T}$ is encountered. Such nodes, together, compose the set $\mathbf{R}_{\mathcal{T}}$. It follows that: $\max _{\mathbf{t} \in \mathbf{R}_{\mathcal{T}}} p_{l}(\mathbf{t})<\mathcal{T}$. $\mathcal{T}$ and $\mathbf{R}_{\mathcal{T}}$ are termed, respectively, Inference Threshold (IT) and the IT Root Set (IT-RS) for $\mathcal{T}$.

For example, the set of variables circled at the stages depicted in Figs. 2(b-d) are the IT-RSs for $\mathcal{T}=p_{l}(\mathbf{x})-\varepsilon$, $\mathcal{T}=p_{l}\left(\mathbf{t}_{1}\right)-\varepsilon$, and $\mathcal{T}=p_{l}\left(\mathbf{t}_{2}\right)-\varepsilon$, respectively. Note that instead of saying $\mathcal{T}=p_{l}(\mathbf{x})-\varepsilon$ we could have said: for any $\mathcal{T} \in\left(p_{l}\left(\mathbf{t}_{1}\right), p_{l}(\mathbf{x})\right)$. However, expressing ITs in terms of $\varepsilon$ liberates us from having to express them in terms of intervals, thereby simplifying the exposition. We would like to emphasize that the adopted notation should not be construed as implying that the assignment of values to ITs is such a sensitive task that everything would have collapsed, had IT not been chosen in such a fine-tuned manner. To recap, in simple terms, $\mathcal{T}$ bears on how far into the past a reasoner is consulting her mental model in the process of answering a query, and $\mathbf{R}_{\mathcal{T}}$ characterizes the furthest-into-the-past concepts entertained by the reasoner in that process.

Next, we formally present the main idea behind PLIF, followed by its interpretation in simple terms.

Lemma 1. Let $\mathbb{P}(O \mid E)$ denote the posed causal query, with $\boldsymbol{O}$ and $\boldsymbol{E}$ denoting, respectively, the disjoint sets of target and observed variables. For any chosen IT $\mathcal{T}<$ $p_{l}^{*}$ and its corresponding $\boldsymbol{R}_{\mathcal{T}}$, define $\boldsymbol{S}: \triangleq \boldsymbol{R}_{\mathcal{T}} \backslash \boldsymbol{E}$. Then the following holds: $\quad \min _{S \in \operatorname{Val}(\boldsymbol{S})} \mathbb{P}(O \mid S, E) \leq \mathbb{P}(O \mid E) \leq$ $\max _{S \in \operatorname{Val}(\boldsymbol{S})} \mathbb{P}(O \mid S, E)$. Crucially, the provided bounds can be computed using the information encoded in the submodel retrieved in the very process of obtaining the $\boldsymbol{R}_{\mathcal{T}}$.

The message of Lemma 1 is quite simple: For any chosen inference threshold $\mathcal{T}$ which is further into the past than $\mathbf{0}^{*}$, Lemma 1 ensures that the reasoner can condition on $\mathbf{S}$ and obtain the reported lower and upper bounds on the query by using only the information encoded in the retrieved submodel.

It is natural to ask under what conditions the exact value to the posed query can be derived using the thus-far retrieved submodel, i.e., the submodel obtained during the identification of $\mathbf{R}_{\mathcal{T}}$. The following remark bears on that. ${ }^{10}$

Remark 1. If for IT $\mathcal{T}, \mathbf{R}_{\mathcal{T}}$ satisfies either: (i) $\mathbf{R}_{\mathcal{T}} \subseteq \mathbf{E}$, or (ii) for all $\mathbf{r} \in \mathbf{R}_{\mathcal{T}}, p_{l}(\mathbf{r})=T_{0}$, and $\min _{\mathbf{e} \in \mathbf{E}} p_{l}(\mathbf{e})>\mathcal{T}$, or

[^473](iii) the lower and upper bound given in Lemma 1 are identical, then the exact value of the posed query can be derived using the submodel retrieved in the process of obtaining $\mathbf{R}_{\mathcal{T}}$. Fig. 2(d) shows a setting wherein (i) and (iii) are both met.

### 4.1 Case Study

Next, we intend to cast the Hidden Markov Model (HMM) studied in (Icard \& Goodman, 2015, p. 2) into our framework. The setting is shown in Fig. 3(left). We adhere to the


Figure 3: Left: The infinite-sized HMM discussed in (Icard \& Goodman, 2015) with parameterization: $\mathbb{P}\left(x_{t+1} \mid x_{t}\right)=$ $\mathbb{P}\left(\bar{x}_{t+1} \mid \bar{x}_{t}\right)=0.9$, and $\mathbb{P}\left(y_{t} \mid x_{t}\right)=\mathbb{P}\left(\bar{y}_{t} \mid \bar{x}_{t}\right)=0.8$. Right: Applying PLIF on the HMM shown in left. Vertical and horizontal axes denote, respectively, the value of the posed query $\mathbb{P}\left(x_{t+1} \mid y_{-\infty: t}\right)$ and the adopted IT $\mathcal{T}$. The vertical bars depict the intervals within which the query lies due to Lemma 1. The dotted curves-which connect the lower and upper bounds of the intervals-show how the intervals shrink as IT $\mathcal{T}$ decreases.
same parameterization and query adopted therein. All RVs in this section are binary, taking on values from the set $\{0,1\}$; $\mathbf{x}=x$ indicates the event wherein $\mathbf{x}$ takes the value 1 , and $\mathbf{x}=\bar{x}$ implies the event wherein $\mathbf{x}$ takes the value 0 . We assume $p_{l}\left(\mathbf{x}_{t+i}\right)=i-2 .{ }^{11}$ We should note that the assignment of the PLs for the variables in $\left\{\mathbf{y}_{t-i}\right\}_{i=0}^{+\infty}$ does not affect the presented results in any way. The query of interest is $\mathbb{P}\left(x_{t+1} \mid y_{-\infty: t}\right)$. Notice that after performing three steps of the sort discussed in the example presented in Sec. 3 (for the IT $\mathcal{T}=-3-\varepsilon)$, the lower bound on the posed query exceeds 0.5 (shown by the red dashed line in Fig. 3(right)). This observation has the following intriguing implication. Assume, for the sake of argument, that we were presented with the following Maximum A-Posterior (MAP) inference problem: Upon observing all the variables in $\left\{\mathbf{y}_{t-i}\right\}_{i=0}^{+\infty}$ taking on the value 1, what would be the most likely state for the variable $\mathbf{x}_{t+1}$ ? Interestingly, we would be able to answer this MAP inference problem simply after three backward moves (corresponding to the IT $\mathcal{T}=-3-\varepsilon$ ). In Fig. 3(right), the intervals within

[^474]which the posed query falls (due to Lemma 1) in terms of the adopted IT $\mathcal{T}$ are depicted.

Our analysis confirms Icard and Goodman's (2015) insight that even in the extreme case of having infinite-sized relevant submodel (Fig. 3(left)), the portion of which the reasoner has to consult so as to obtain a "sufficiently good" answer to the posed query could happen to be very small (Fig. 3(right)).

## 5 Discussion

To our knowledge, PLIF is the first inference framework proposed that capitalizes on time to constrain the scope of causal reasoning over CBNs, where the term scope refers to the portion of a CBN on which inference is carried out. PLIF does not restrict itself to any particular inference scheme. The claim of PLIF is that inference should be confined within and carried out over retrieved submodels of the kind suggested by Lemma 1 so as to obtain the reported bounds therein. In this light, PLIF can accommodate any inference scheme, including Belief Propagation (BP), and sample-based inference methods using Markov Chain Monte Carlo (MCMC), as two prominent classes of inference schemes. MCMC-based methods have been successful in simulating important aspects of a wide range of cognitive phenomena and accounting for many cognitive biases; cf. (Sanborn \& Chater, 2016). Also, work in theoretical neuroscience has suggested mechanisms for how BP and MCMC-based methods could be realized in neural circuits; cf. (Gershman \& Beck, 2016; Lochmann \& Deneve, 2011). For example, to cast BP into PLIF amounts to restricting BP's message-passing within submodels of the kind suggested by Lemma 1. In other words, assuming that BP is to be adopted as the inference scheme, upon being presented with a causal query, an IT according to Lemma 1 will be selected-at the meta-level-by the reasoner and the corresponding submodel, as suggested by Lemma 1, will be retrieved, over which inference will be carried out using BP. This will lead to obtaining lower and upper bounds on the query, as reported in Lemma 1. If time permits, the reasoner builds up incrementally on the thus-far retrieved submodel so as to obtain tighter bounds on the query. ${ }^{12}$ MCMC-based inference methods can be cast into PLIF in a similar fashion.

The problem of what parts of a CBN are relevant and what are irrelevant for a given query, according to (Geiger, Verma, \& Pearl, 1989), was first addressed by Shachter (1988). The approaches proposed for identifying the relevant submodel for a given query fall into two broad categories (cf. Mahoney \& Laskey, 1998, and references therein): (i) top-down approaches, and (ii) bottom-up approaches. Top-down approaches start with the full knowledge of the underlying CBN and, depending on the posed query, gradually prune the irrelevant parts of the CBN. In this respect, top-down approaches are inevitably from "god's eye" point of view-a characteristic which undermines their cognitive-plausibility. Bottom-

[^475]up approaches, on the other hand, incrementally construct a submodel (by moving backwards from the query variables), using which the posed query can be computed. It is crucial to note that bottom-up approaches cannot stop at intermediate steps during the backward move and run inference on the thus-far constructed submodel without running the risk of compromising some of the (in)dependence relations structurally encoded in the CBN, which would yield erroneous inferences. This observation is due to the fact that there exists no local signal revealing how the thus-far retrieved nodes are positioned relative to each other and to the to-be-retrieved nodes-a shortcoming circumvented in the case of PLIF by introducing PL. It is worth reiterating again that PLIF subscribes to what we call the concurrent approach to reasoning (as opposed to the flawed sequential approach mentioned earlier), whereby retrieval and inference take place in tandem. The HMM example analyzed in Sec. 4.1, with infinitely large relevant submodel, stresses the importance and shows the efficacy of the concurrent approach.

Work on causal judgment provides support for the socalled alternative neglect, according to which subjects tend to neglect alternative causes to a much greater extent in predictive reasoning than in diagnostic reasoning (Fernbach \& Rehder, 2013; Fernbach, Darlow, \& Sloman, 2011). Alternative neglect, therefore, implies that subjects would tend to ignore parts of the relevant submodel while constructing it. Recent findings, however, seem to cast doubt on alternative neglect (Cummins, 2014; Meder, Mayrhofer, \& Waldmann, 2014). Meder et al. (2014), Experiment 1 demonstrates that subjects appropriately take into account alternative causes in predictive reasoning. Also, Cummins (2014) substantiates a two-part explanation of alternative neglect according to which: (i) subjects interpret predictive queries as requests to estimate the probability of the effect when only the focal cause is present, an interpretation which renders alternative causes irrelevant, and (ii) the influence of inhabitory causes (i.e., disablers) on predictive judgment is underestimated, and this underestimation is incorrectly interpreted as neglecting of alternative causes. Cummins (2014), Experiment 2 shows that when predictive inference is queried in a manner that more accurately expresses the meaning of noisyOR Bayes net (i.e., the normative model adopted by Fernbach et al. (2011)) likelihood estimates approached normative estimates. Cummins (2014), Experiment 4 shows that the impact of disablers on predictive judgments is far greater than that of alternative causes, while having little impact on diagnostic judgments. PLIF commits to the retrieval of enablers as well as disablers. As mentioned earlier, PLIF abstracts away from the inference scheme operating on the retrieved submodel, and, hence, leaves it to the inference scheme to decide how the retrieved enablers and disablers should be weighted and subsequently integrated. In this light, PLIF is consistent with the results of Experiment 4 in (Cummins, 2014).

In an attempt to explain violations of screening-off reported in the literature, Park and Sloman (2013) find strong
support for the contradiction hypothesis followed by the mediating mechanism hypothesis, and finally conclude that people do conform to screening-off once the causal structure they are using is correctly specified. PLIF is consistent with these accounts, as it adheres to the assumption that reasoners carry out inference on their internal causal model (including all possible mediating variables and disablers), not the potentially incomplete one presented in the cover story; see also (Rehder \& Waldmann, 2017; Sloman \& Lagnado, 2015).

Experiment 5 in (Cummins, 2014), consistent with (Fernbach \& Rehder, 2013), shows that causal judgments are strongly influenced by memory retrieval/activation processes, and that both number of disablers and order of disabler retrieval matter in causal judgments. These findings suggest that the CFP and memory retrieval/activation are intimately linked. In that light, next, we intend to elaborate on the rationale behind adopting the term "retrieve" and using it interchangeably with the term "consult" throughout the paper; this is where we relate PLIF to the concepts of Long Term Memory (LTM) and Working Memory (WM) in psychology and neurophysiology. Next, we elaborate on how PLIF could be interpreted through the lenses of two influential models of WM, namely, Baddeley and Hitch's (1974) Multi-component model of WM (M-WM) and Ericsson and Kintsch's (1995) Long-term Working Memory (LTWM) model. The M-WM postulates that "long-term information is downloaded into a separate temporary store, rather than simply activated in LTM," a mechanism which permits WM to "manipulate and create new representations, rather than simply activating old memories" (Baddeley, 2003). Interpreting PLIF through the lens of the M-WM model amounts to the value for IT being chosen (and, if time permits, updated so as to obtain tighter bounds) by the central executive in the M-WM and the submodel being incrementally "retrieved" from LTM into MWM's episodic buffer. Interpreting PLIF through the lens of the LTWM model amounts to having no retrieval from LTM into WM and the submodel suggested by Lemma 1 being merely "activated in LTM" and, in that sense, being simply "consulted" in LTM. In sum, PLIF is compatible with both of the narratives provided by the M-WM and LTWM models.

A number of predictions follow from PL and PLIF. For instance, PLIF makes the following prediction: Prompted with a predictive or a diagnostic query (i.e., $\mathbb{P}(\mathbf{e} \mid \mathbf{c})$ and $\mathbb{P}(\mathbf{c} \mid \mathbf{e})$, respectively), subjects should not retrieve any of the effects of e. Introspectively, this prediction seems plausible, and can be tested, using a similar approach to (Cummins, 2014; De Neys, Schaeken, \& d'Ydewalle, 2003), by asking subjects to "think aloud" while engaged in predictive or diagnostic reasoning. Also, PL yields the following prediction: Upon intervening on cause $\mathbf{c}$, subjects should be sensitive to when effect $\mathbf{e}$ will occur, even in settings where they are not particularly instructed to attend to such temporal patterns. Recent findings suggesting that people have expectations about the delay length between cause and effect already provide some supporting evidence for this prediction (Greville \& Buehner,

2010; Buehner \& May, 2004).
There is a growing acknowledgment in the literature that, not only time and causality are intimately linked, but that they mutually constrain each other in human cognition; cf. (Buehner, 2014). In line with this view, we see our work also as an attempt to formally articulate how time could guide and constrain causal reasoning. While many questions remain open, we hope to have made some progress towards better understanding of the CFP at the algorithmic level.

Acknowledgments. We are grateful to Thomas Icard for valuable discussions. We would also like to thank Marcel Montrey and Peter Helfer for helpful comments on an earlier draft of this work. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant RGPIN 262017.

## References

Baddeley, A. (2003). Working memory: looking back and looking forward. Nature Review Neuroscience, 4(10), 829-839.
Baddeley, A. D., \& Hitch, G. (1974). Working memory. Psychology of Learning and Motivation, 8, 47-89.
Bramley, N. R., Gerstenberg, T., \& Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In Proceedings of the 36th annual conference of the cognitive science society (pp. 236-241).
Buehner, M. J. (2014). Time and causality: editorial. Frontiers in Psychology, 5, 228.
Buehner, M. J., \& May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. Quarterly Journal of Experimental Psychology Section B, 57(2), 179-191.
Cummins, D. D. (2014). The impact of disablers on predictive inference. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(6), 1638.
De Neys, W., Schaeken, W., \& d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. Memory \& Cognition, 31(4), 581-595.
Ericsson, K. A., \& Kintsch, W. (1995). Long-term working memory. Psychological Review, 102(2), 211.
Fernbach, P. M., Darlow, A., \& Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. Journal of Experimental Psychology: General, 140(2), 168-185.
Fernbach, P. M., \& Rehder, B. (2013). Cognitive shortcuts in causal inference. Argument \& Computation, 4(1), 64-88.
Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres.
Geiger, D., Verma, T., \& Pearl, J. (1989). d-separation: from theorems to algorithms. Fifth Workshop on Uncertainty in Artificial Intelligence, pp. 118-125.
Gershman, S. J., \& Beck, J. M. (2016). Complex probabilistic inference: From cognition to neural computation. In Computational Models of Brain and Behavior, ed A. Moustafa.
Glymour, C. (1987). Android epistemology and the frame problem, The Robot's Dilemma: The Frame Problem in AI. pp. 65-75.
Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., \& Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. Psychological Review, 111(1), 3-32.
Greville, W. J., \& Buehner, M. J. (2010). Temporal predictability facilitates causal learning. Journal of Experimental Psychology: General, 139(4), 756.
Hume, D. (1748/1975). An inquiry concerning human understanding. Oxford University Press.
Icard, T. F., \& Goodman, N. D. (2015). A resource-rational approach to the causal frame problem. Proc. of the 37th Annual Meeting of the Cognitive Science Society.
Lagnado, D. A., \& Sloman, S. A. (2006). Time as a guide to cause. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(3), 451-60.
Lochmann, T., \& Deneve, S. (2011). Neural processing as causal inference. Current Opinion in Neurobiology, 21(5), 774-781.
Mahoney, S. M., \& Laskey, K. B. (1998). Constructing situation specific belief networks. Proc. of the Conference on Uncertainty in Artificial Intelligence, pp. 370378.

Marr, D. (1982). Vision: a computational approach.
Meder, B., Mayrhofer, R., \& Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. Psychological Review, 121(3), 277.
Park, J., \& Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the markov property in causal reasoning. Cognitive Psychology, 67(4), 186-216.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.
Rehder, B., \& Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. Memory \& Cognition, 45, 245-260.
Sanborn, A. N., \& Chater, N. (2016). Bayesian brains without probabilities. Trends in Cognitive Sciences, 20(12), 883-893.
Schlottmann, A. (1999). Seeing it happen and knowing how it works: how children understand the relation between perceptual causality and underlying mechanism. Dev Psychol, 35(1), 303.
Shachter, R. D. (1988). Probabilistic inference and influence diagrams. Operations Research, 36(4), 589-604.
Simon, H. A. (1957). Models of man. Wiley.
Sloman, S. A., \& Lagnado, D. (2015). Causality in thought. Annual Review of Psychology, 66, 223-247.

# Experimental and Computational Investigation of the Effect of Caffeine on Human Time Perception 

Remya Sankar ${ }^{1}$ (remya.s@research.iiit.ac.in)<br>Anuj K. Shukla ${ }^{1}$ (anuj.shukla@research.iiit.ac.in)<br>Raju S. Bapi ${ }^{1,2}$ (raju.bapi@iiit.ac.in)<br>${ }^{1}$ Cognitive Science Lab, International Institute of Information Technology, Hyderabad, India<br>${ }^{2}$ School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India


#### Abstract

Perception of time is an active process that takes place continually. However, we are yet to learn its exact mechanisms conclusively. The temporal bisection task is ideal to investigate the circuitry underlying time perception. Caffeine, a commonly used stimulant, has been known to play a role in modulation of time perception. The objective of this article is to explore the role of caffeine, a neuromodulator, in the perception of time in human beings by conducting suitable experiments. The experiment shows that an expansion of time is perceived by subjects after caffeine ingestion and that caffeine has an accelerating effect on our time perception system. Additionally, we present a preliminary 2 -step decision model that fits the results of the experiment and potentially gives insights into the mechanisms of caffeine. We conclude by pointing out future directions towards a more biologically realistic computational model.


Keywords: Caffeine; Timing; Perception; Temporal bisection; Computational modeling; Decision making

## Introduction

Time perception is essential for human survival. It is a multilayered process which covers a wide range of timescales, from microsecond estimation to the maintaining of the circadian rhythm. However, there is still a lack of consensus on the mechanisms behind temporal perception.

Researchers have proposed several qualitative and quantitative models to explain the data obtained in various temporal judgement experiments (Jeffress, 1948; Machado, Malheiro, \& Erlhagen, 2009; Oprisan \& Buhusi, 2011), with Internal Clock Theory being one of the most widely accepted ones. It suggests that our perception of time highly relies on the clock speed. Akin to an internal clock, the theory of scalar expectancy postulates that a group of oscillating neurons would work as a pacemaker and help in the judgement of durations (Gibbon, 1977; Gibbon, Church, \& Meck, 1984; Wearden, 1991). As described in the information processing model in Figure 1, a pacemaker oscillates at a mean frequency and produces regular clock pulses, which are gated to an accumulator in working memory via a switch. The accumulator records and stores the number of pulses from the onset of the stimulus. A comparator decides if the current record in the working memory is close enough to the reference memory and responds accordingly. If the response is reinforced, the time value recorded in the working memory is stored in the permanent reference memory for reinforced values.

In the previous findings, it has been argued that the nuclei involved in the circadian rhythm in the brain partici-
pate in our perception of time (Cheng, Meck, \& Williams, 2006). Findings from psychopharmacological studies also suggest that caffeine and other psychoactive drugs affect these nuclei (Dunlap, 1999). Further, literature also shows that dopaminergic drugs influence the speed of internal pacemaker (Buhusi \& Meck, 2002). In general, caffeine has been known to have effects on other cognitive processes like vigilance, attention, memory and other cognitive functions (McLellan, Caldwell, \& Lieberman, 2016). There have been a wide range of studies that have investigated the role of caffeine in time perception. However, the findings are still largely inconclusive (Hussain \& Cole, 2015; Favila \& Kuhl, 2014; Borota et al., 2014).


Figure 1: Information-processing model for Scalar Expectancy Theory (adapted from SET, Gibbon (Gibbon et al., 1984)).

Given the above theoretical basis, we hypothesized that the administration of caffeine would cause a difference in perception of duration by influencing the speed of the internal pacemaker. The aim of the current experiment is to explore the role of caffeine, a neuromodulator, on time judgement via suitable experiments and to design a computational model, based on decision-making, to investigate the possible mechanisms underlying the perception of duration.

The temporal bisection task was initially used in 1977 to study temporal discrimination in rats (Church \& Deluty, 1977). As the task requires several time-dependent cognitive functions, such as the comparison of durations, it is
an ideal technique to study perception and processing of time (Wearden, 1991; Allan \& Gibbon, 1991). We have thus chosen the temporal bisection task as the paradigm to investigate the modulation of judgement of duration by caffeine.

## Experiment

## Participants

The study sample consisted of 24 adults ( 8 females and 16 males, mean age $=21$ years, $\mathrm{SD}=0.89$ ), who were students of the International Institute of Information Technology, Hyderabad, India. All participants gave informed consent prior to the experiment. A Python script was used to randomly assign each subject to either the control group (0mg caffeine) or the experimental group ( 200 mg caffeine). All participants were right-handed and had normal or corrected-to-normal vision.

## Materials and apparatus

Each participant was tested individually in a quiet room in the institute. The experiment was presented on a Macintosh laptop, which controlled the presentation of the experimental stimuli and recorded the participants' responses with Psychopy (Peirce, 2009). The participants were asked to convey their response using the ' $S$ ' and ' $L$ ' keys on the keyboard, for 'short' and 'long', respectively. The stimuli used for representation of duration in the bisection task were a white rectangle (during the training phase) and a white triangle (during the testing phase) on a black background, presented in the center of the screen. During the training phase, post-response feedback was presented as white text on a black background. The feedback was presented for 2 s in the center of the laptop screen (Droit-Volet, Brunot, \& Niedenthal, 2004).

The participants were administered plain or caffeinated milk orally, in the control or experimental group, respectively. The participants in the group were administered a moderate dose of 200 mg caffeine since it has been observed that caffeine enhances performance in several cognitive tasks with minimal side effects, in doses up to approximately 300 mg (Lieberman, Tharion, Shukitt-Hale, Speckman, \& Tulley, 2002).

Peak plasma levels of caffeine are found in the body about 30 minutes after ingestion (Blanchard \& Sawers, 1983), following which the effects are felt substantially for approximately 30 minutes (Barry, Clarke, Johnstone, \& Rushby, 2008). Hence, the experiment was conducted 20 to 25 minutes after the administration of plain or ceffeinated milk. One session of the experiment lasted for a duration of about 35 minutes.

## Experimental Procedure

The temporal bisection task comprises duration judgement between two reference durations. The task involves subjects classifying various probe durations as either 'short' or 'long'. The conscious realm of time perception occurs in the range of seconds and minutes (Mauk \& Buonomano, 2004). Effects of emotion, age, etc. on time perception have been studied via
temporal bisection tasks in this time range (Droit-Volet et al., 2004). In order to see the effects of caffeine on this time range and to investigate conscious time estimation, we have chosen 400 ms (short standard) and 1600 ms (long standard), as the reference durations for our experiment.

The temporal bisection task consisted of two phases: training and testing. The training phase, in itself, was composed of three sections. In the first section, participants were presented with the short(S) and long(L) standard stimulus durations. Each standard was initially presented five times each and the subjects were asked to observe carefully. In section two of the training phase, the participants were presented with 5 trials each for $S$ and $L$, in randomised order. In each trial, after the presentation of stimulus, the participants were asked to decide if the given stimulus was short or long by pressing the ' $S$ ' or 'L' keys, respectively. On responding, they were presented with a feedback, informing them if their response was accurate or not. The feedback message lasted on the screen for a duration of 2 s . In section three of the training phase, the participants were again presented with 5 trials each of $S$ and $L$ in randomised order, similar to section 2 and asked to respond if they perceived the duration as short or long. However, this time, they were provided with no feedback. The inter-stimulus interval (ISI) in this phase was fixed at 1s. After the completion of section three, participants with a score higher than 7 correct trials out of 10 (in section three) were selected for the testing phase. The testing phase was conducted after a break of 1 minute.


Figure 2: Schematic diagram of the testing phase.

In the testing phase, as depicted in Figure 2, the participants were presented with 5 intermediate probe durations, $600 \mathrm{~ms}, 800 \mathrm{~ms}, 1000 \mathrm{~ms}, 1200 \mathrm{~ms}, 1400 \mathrm{~ms}$ in addition to the $400 \mathrm{~ms}(\mathrm{~S})$ and $1600 \mathrm{~ms}(\mathrm{~L})$ reference durations, in a randomised order and asked to respond if they perceived the given duration as short or long (Droit-Volet et al., 2004). The inter-stimulus interval was randomly chosen between 0.8 s and 1.2 s . No accuracy feedback was presented in the testing phase. Each block consisted of 35 trials, i.e. each probe duration occurred 5 times in a block in a random order. After each block, the participants were asked to take a 30s break. The participants were presented with 10 such blocks.

## Results and Analysis: Experiment

For each participant, the proportion of 'long' responses was calculated for each probe duration. In Figure 3, the proportion of 'long' responses from all participants in both groups has been plotted against the probe durations. The point of subjective equality (PSE) is the stimulus duration for which a subject recorded a response of 'long' with a $50 \%$ probability. The PSE was calculated for each participant by fitting a Weibull curve to the plot of proportion of 'long' responses vs probe duration. The PSEs for the 12 participants in the control group ( $M=1.137 s, S D=0.118$ ) was found to be generally higher than in the experimental group ( $M=0.929 s, S D=0.140$ ). An independent samples t-test revealed that there was a statistically significant difference between the two groups, $t(22)=3.76, p=0.001$.


Figure 3: Proportion of 'long' responses as a function of stimulus duration, by subjects from the control and experimental groups. Also, the PSEs of both groups are depicted (not to scale).

## Discussion: Experiment

An independent samples t-test showed that the PSE for the participants in the two groups remarkably varied from each other. This allows us to conclude that the administration of caffeine leads to a notable change in the perception of duration. The mean PSE of the 12 participants in the control group is higher than the mean PSE of the participants in the experimental group. Moreover, as shown in Figure 3, a clear shift in the PSE of the experimental group towards the shorter reference duration is observed. This shift in PSE implies that for a given stimulus duration, there is a higher probability that a participant responds 'long' in the experimental group than in the control group. In other words, a given probe duration is perceived as longer by participants under the influence of caffeine. These results lead us to conclude that caffeine produces a perception of expansion of time in humans.

The scalar expectancy theory postulates that a pacemaker
sends pulses at a mean frequency from the onset of stimulus, gated by a switch, to the accumulator. In the working memory, a comparator judges if the number of pulses accumulated is closer to a reference memory value of the short or the long standard and responds accordingly (Gibbon et al., 1984). In accordance with this information processing model, we can infer that caffeine could influence discrimination of temporal durations in one or more of the following ways.

- By increasing the frequency with which pulses are generated by the pacemaker. This would lead to a higher number of pulses getting accumulated for a given duration, due to which the comparator would associate it to be closer to the long reference duration.
- By causing distortion in the memory of the reference durations. Since the recall in the long-term memory has a higher variance, the interactions between these distorted representations of the reference durations would lead the working memory to make inaccurate comparisons, which could in turn result in increased 'long' responses.


## Computational Model

The dataset obtained from the above experiment was modeled using a simple decision model that fits the data, although biologically infeasible. This computational model is a twostep Gaussian model which has only two free parameters and is capable of reproducing the characteristics of the empirical data. The memories of the short and long reference durations in the temporal bisection task are modeled using scalar Gaussian distribution. The Gaussian helps depict the inherent noise in human memory (Kopec \& Brody, 2010).

## Description of the Model

The model proposed by us comprises of two steps. In step I, the model determines if the given stimulus is one of the reference durations, in which case, it responds accordingly, or is an intermediate duration, in which case it moves to the second step in order to make a decision. In step II, the model computes the difference between the stimulus and its memories of both the reference durations, and responds according to the one which is lesser in magnitude (Kopec \& Brody, 2010). Each of the two steps is explained in detail below.

## Model Step I

The memory of each of the two reference durations is modeled as a Gaussian distribution over durations, with a mean equal to the reference duration, and a standard deviation proportional to the reference duration. This proportion, referred to as the coefficient of variation, is randomly chosen from a suitable range of values (discussed in "Results and Analysis: Computational Model" section). The height of the Gaussian distribution of a particular reference duration at a given stimulus duration is taken as the probability of the stimulus being labelled as that reference duration( pL and pS ). As a participant can potentially classify the reference duration correctly
with $100 \%$ accuracy, discounting human error, the Gaussian distributions range from 0 to 1 (Kopec \& Brody, 2010). We take the probability of a stimulus duration being labeled as 'intermediate' ( pI ), to be the sum of the probabilities of the two reference memory distributions at that stimulus duration subtracted from 1, i.e. $p I=1-(p L+p S)$.

If the probabilities of the 2 reference memory density distributions at the stimulus duration are approximately equal, then the model responds either 'short' or 'long' with an equal probability. Otherwise, a choice is made if a given stimulus is long, short or intermediate depending upon the values of their respective probability distributions, $p L, p S$ and $p I$. If the stimulus is determined to be either the short or the long reference duration, then the model responds 'short' or 'long', respectively. If the stimulus is deemed to be 'intermediate', the model proceeds to step 2.

## Model Step II

The model computes if the stimulus duration ' $s$ ' is closer to either reference duration stored in memory and responds accordingly. The scalar Gaussian distributions for the short(TS) and long(TL) standards are used to model the reference duration values pulled from memory. One value ts is drawn from the TS distribution, and one value tl is drawn from the TL distribution. In order to model the shift in PSE brought about due to caffeine administration, the model is explicitly biased in this step towards responding 'short' or 'long', depending upon whether it's simulating the control group or the experimental group. The bias factor, B, is randomly picked from a certain optimal range (discussed below) depending upon the group. If $a b s(t s-s) * B<a b s(t l-s)$, then the subject responds 'short', and otherwise, the subject responds 'long'.

## Results and Analysis: Computational Model

The model contains only two free parameters, the coefficient of variation (CV) of the two probability distributions used to model the reference memories, and $B$, an intrinsic bias factor influencing the decision process. The values for these parameters were chosen by testing the parameter space over a range of values (CV range : $0.18-0.27$, resolution : 0.01 ; Brange : $0.6-1.4$, resolution : 0.1). The data generated by a certain value of CV and B was evaluated on the basis of an independent samples t -test between the data generated by the model and the empirical dataset collected from the control group. It can be observed from the Figure 4 that the following range of values are optimal for the 2 parameters:-

- For experimental group, $C V: 0.23-0.26$ and $B: 1.0-1.4$
- For control group, $C V: 0.17-0.22$ and $B: 0.6-1.0$

The lesser the p-value, the more significant the difference between the simulated data and the data collected from the control group. Thus, Figure 4 shows that for low values of CV and B , the data generated by the model is significantly similar to the empirical data for the control group.


Figure 4: Bubble chart of p-values of independent samples t -test between the empirical data of the control group and the simulated data at varying CV and B values.

The final model generates a dataset over 12 runs consisting of 350 trials each, simulating 12 subjects each for both the control and experimental groups. At the beginning of each run, the values for CV and B are randomly chosen from the optimal range for the concerned group.

To analyse the data generated by the model, the proportion of 'long' responses was calculated for each probe duration for every run. The point of subjective equality (PSE) was calculated for each run by fitting a Weibull curve to the plot of proportion of 'long' responses vs probe durations. The PSEs for the 200mg caffeine group ( $M=0.907, S D=0.022$ ), similar to the empirical data, was found to be lower than the PSEs for the control group ( $M=1.014, S D=0.047$ ). An independent samples t-test between the data showed a statistically significant difference between the two groups, $t(22)=6.736, p<0.001$.

## Discussion: Computational Model

From the statistics regarding the PSEs for the 2 groups, we can see that for corresponding groups, the simulated dataset as well as the human dataset gives similar mean PSE values. The standard deviation of the PSEs generated by the model is considerably lesser than the same in the human dataset. The higher variance in the human dataset might be due to human error, fatigue and slight inconsistencies in perception of time by different participants. The model tries to incorporate this variance between the PSEs for different subjects, by randomly picking a CV and B value, for each run, from the optimal range for the concerned group. Yet, this does not give rise to sufficient variation in the generated data as compared to the human dataset. However, the mean PSE is accurately simulated for both groups.

The independent samples t-test and the Weibull fit between the generated datasets for the experimental and control groups show that the model closely mimics human temporal judg-


Figure 5: Effect of setting the parameter CV on the size of decision regions $\mathbf{A}$. When CV is set to optimal values found for the 0mg group, the central region expands $\mathbf{B}$. When CV is set to optimal values found for the 200 mg group, the central region shrinks and there is more overlap of the Gaussians corresponding to the reference durations.
ment.
As mentioned in the previous section, high values of the parameters, $C V$ and $B$, were found to be suitable to model the experimental group, while low values of CV and B were suitable to simulate the control group. These parameter ranges might lend us some insight into the mechanism of caffeine action, as discussed below.

In the model, the value for B increases or decreases the distance of the stimulus from the short standard, depending upon whether it is high or low. If the distance of the stimulus from the short standard is higher than the distance from the long standard, the model would respond 'long'. This explains why a high value of B is suitable for modeling the experimental group and vice versa. We can, therefore, infer that this range of values for ' $B$ ' can be indicative of the frequency of the pacemaker. A rise in the rate at which pulses are generated would lead to more pulses being accumulated for a given duration and could lead to a perceived expansion of time.

In addition to the range of B-values, we also find that high values of CV are suited to modeling the experimental group. This can be explained as follows (see Figure 5). In step 1 of the model, the probability distribution for the 'long' reference duration has a larger standard deviation, as it has the same coefficient of variation as the 'short' reference duration, despite
having a larger mean. This implies that for a given stimulus, if the decision is made in step 1 itself, there is a higher probability that the response be 'long'.

The decision to proceed to step 2, is dependent on the value of pI , i.e., the probability that the stimulus is judged as 'intermediate' in step 1. In the experimental group, as the model uses higher CV values, for a given stimulus, the values of pS and pL would be higher than the values for the same in the control group, where the model uses lower CV values. This would cause a decrease in pI for a given stimulus in the experimental group's simulation. Therefore, the probability of the decision being made in step 1 increases, implying that there is a higher probability of the response being 'long' as compared to the control group, as explained in the previous paragraph.

A higher value of CV, while mean remains fixed, implies a larger standard deviation (SD). As the experimental group is being modeled accurately with a higher range of CV values as compared to the control group, the width of the Gaussian distributions used to model the reference durations is higher in the experimental group. This change in the width implies that caffeine might have the potential to cause distortion in the memory of durations. This leads us to infer that caffeine mechanism possibly works via the memory pathway rather than an attentional pathway, as the latter would require a leaner spread of the probability distribution. Despite investigative experiments, there is no general consensus on the nature of acute effects of caffeine on memory (McLellan et al., 2016). However, our model indicates an increase in uncertainty in the reference memory caused by caffeine.

## Limitations and Future Work

One major shortcoming of the model suggested by us is that it is purely a decision model and does not take into account the neural circuitry mediating time perception in humans. The model is pitched at an abstract level and in order to obtain biologically rooted insights, there is a need for a more realistic model.

Substantial evidence has been found that indicates that the basal ganglia and its dopaminergic pathways control time perception to an extent. For instance, it has been observed that PD patients, when administered medication that brings the dopamine concentration back to normal, are capable of performing time estimation accurately, unlike when offmedication (Jones, Malone, Dirnberger, Edwards, \& Jahanshahi, 2008). Furthermore, time perception studies can help in the early detection of such diseases that affect dopamine production and will also increase our understanding of the pathways and the brain areas that may be involved in such diseases.

The fundamental circuitry behind caffeine's action has been established to be the antagonism of adenosine receptors in the central nervous system which leads to interaction with dopamine receptors (Davis et al., 2003; Ferré, 2016). Caffeine blocks A2A receptors in the striatum and promotes a direct excitatory potentiation of D2 receptors. This leads
to an increase in the stimulation of psychomotor activity by dopamine (Ferré, 2016). Hence, we think that a model can be conceptualised which shows caffeine indirectly affecting time processing in the basal ganglia by modulating dopamine.

Alternatively, it has been observed that a cortical neuronal network, without the use of any kind of pacemaker, may have the potential to track duration by storing recent information (Mauk \& Buonomano, 2004). This can be used as inspiration for another biologically feasible model. Furthermore, there is indication in literature that a reinforcement learning based model of interval timing might be able to explain several behavioural as well as neural phenomena (Gershman, Moustafa, \& Ludvig, 2014). These are few methods that can be explored to further investigate the effect of caffeine on time perception.

## Acknowledgments

This work was partially supported by the Department of Science and Technology (DST), Government of India under Indo-French CEFIPRA Grant for the project Basal Ganglia at Large (No. DST-INRIA 2013-02/Basal Ganglia dated 13-09-2014) grants awarded to RSB.

## References

Allan, L. G., \& Gibbon, J. (1991). Human bisection at the geometric mean. Learning and Motivation, 22(1), 39-58.
Barry, R. J., Clarke, A. R., Johnstone, S. J., \& Rushby, J. A. (2008). Timing of caffeine's impact on autonomic and central nervous system measures: clarification of arousal effects. Biological psychology, 77(3), 304-316.
Blanchard, J., \& Sawers, S. (1983). The absolute bioavailability of caffeine in man. European journal of clinical pharmacology, 24(1), 93-98.
Borota, D., Murray, E., Keceli, G., Chang, A., Watabe, J. M., Ly, M., ... Yassa, M. A. (2014). Post-study caffeine administration enhances memory consolidation in humans. Nature neuroscience, 17(2), 201-203.
Buhusi, C. V., \& Meck, W. H. (2002). Differential effects of methamphetamine and haloperidol on the control of an internal clock. Behavioral neuroscience, 116(2), 291.
Cheng, R.-K., Meck, W. H., \& Williams, C. L. (2006). $\alpha 7$ nicotinic acetylcholine receptors and temporal memory: synergistic effects of combining prenatal choline and nicotine on reinforcement-induced resetting of an interval clock. Learning \& Memory, 13(2), 127-134.
Church, R. M., \& Deluty, M. Z. (1977). Bisection of temporal intervals. Journal of Experimental Psychology: Animal Behavior Processes, 3(3), 216.
Davis, J. M., Zhao, Z., Stock, H. S., Mehl, K. A., Buggy, J., \& Hand, G. A. (2003). Central nervous system effects of caffeine and adenosine on fatigue. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 284(2), R399-R404.
Droit-Volet, S., Brunot, S., \& Niedenthal, P. (2004). Brief report perception of the duration of emotional events. Cognition and Emotion, 18(6), 849-858.

Dunlap, J. C. (1999). Molecular bases for circadian clocks. Cell, 96(2), 271-290.
Favila, S. E., \& Kuhl, B. A. (2014). Stimulating memory consolidation. Nature neuroscience, 17(2), 151-152.
Ferré, S. (2016). Mechanisms of the psychostimulant effects of caffeine: implications for substance use disorders. Psychopharmacology, 233(10), 1963-1979.
Gershman, S. J., Moustafa, A. A., \& Ludvig, E. A. (2014). Time representation in reinforcement learning models of the basal ganglia.
Gibbon, J. (1977). Scalar expectancy theory and weber's law in animal timing. Psychological review, 84(3), 279.
Gibbon, J., Church, R. M., \& Meck, W. H. (1984). Scalar timing in memory. Annals of the New York Academy of sciences, 423(1), 52-77.
Hussain, S. J., \& Cole, K. J. (2015). No enhancement of 24hour visuomotor skill retention by post-practice caffeine administration. PloS one, 10(6), e0129543.
Jeffress, L. A. (1948). A place theory of sound localization. Journal of comparative and physiological psychology, 41(1), 35.
Jones, C. R., Malone, T. J., Dirnberger, G., Edwards, M., \& Jahanshahi, M. (2008). Basal ganglia, dopamine and temporal processing: performance on three timing tasks on and off medication in parkinsons disease. Brain and cognition, 68(1), 30-41.
Kopec, C. D., \& Brody, C. D. (2010). Human performance on the temporal bisection task. Brain and cognition, 74(3), 262-272.
Lieberman, H. R., Tharion, W. J., Shukitt-Hale, B., Speckman, K. L., \& Tulley, R. (2002). Effects of caffeine, sleep loss, and stress on cognitive performance and mood during us navy seal training. Psychopharmacology, 164(3), 250261.

Machado, A., Malheiro, M. T., \& Erlhagen, W. (2009). Learning to time: A perspective. Journal of the Experimental Analysis of Behavior, 92(3), 423-458.
Mauk, M. D., \& Buonomano, D. V. (2004). The neural basis of temporal processing. Annu. Rev. Neurosci., 27, 307340.

McLellan, T. M., Caldwell, J. A., \& Lieberman, H. R. (2016). A review of caffeines effects on cognitive, physical and occupational performance. Neuroscience \& Biobehavioral Reviews, 71, 294-312.
Oprisan, S. A., \& Buhusi, C. V. (2011). Modeling pharmacological clock and memory patterns of interval timing in a striatal beat-frequency model with realistic, noisy neurons. Frontiers in integrative neuroscience, 5, 52.
Peirce, J. W. (2009). Generating stimuli for neuroscience using psychopy.
Wearden, J. (1991). Human performance on an analogue of an interval bisection task. The Quarterly Journal of Experimental Psychology, 43(1), 59-81.

# Peculiarity doesn't trump ordinarity: On recognition memory for exceptions to the category rule 

Olivera Savic (savic.7@osu.edu)<br>Department of Psychology, 1835 Neil Avenue, Columbus, OH 43210, United States

Vladimir Sloutsky (sloutsky.1@osu.edu)<br>Department of Psychology, 1835 Neil Avenue, Columbus, OH 43210, United States


#### Abstract

While exceptions to a regularity might be rare, categories that have exceptions are not. Previous studies on learning categories that have exceptions suggested special status of exceptional items in memory (e.g. Palmeri \& Nosofsky, 1995, Sakamoto and Love, 2004). However, this might be true only for a special kind of exceptions - those that call for forming complex binding structures, and could be learned only if they are fully memorized. In the two experiments in this study, we show that memory for exceptions is not better than memory for regular category members (Experiment 1). On the contrary, both children and adults had better memory for the features of regular items (Experiment 2). In addition, adults, but not 4-year-olds, showed better memory for the rule than for probabilistic features. The overall results challenge the idea of the special status of exceptions in memory.


Keywords: rule-plus-exception; differential memory; category structure

## Introduction

An exception is a case to which a rule or a general statement does not apply. Tomatoes are an exception to the category of vegetables, penguins are an exception to the category of birds, bats to the category of mammals, and the verb "cut" (since it does not change its form) is an exception to the rule of tense formation. While exceptions may be of various kinds, what is common for all of them is that they violate our expectations about how something (or someone) should behave, what it should look like, where it should belong, as well as other expectations that are based on our previous knowledge. Therefore, understanding of how we learn about and how we represent those rare, deviant cases is an interesting problem for the theories of category learning.

A common assumption underlying models that aimed to explain how we learn and represent exceptions, is that exceptions have a privileged memory status. Work in the schema literature contrasting memory for schema-consistent (i.e., in accord with expectations) and schema-inconsistent information, demonstrated that schema-inconsistent information is remembered better (for meta-analysis see, Rojahn \& Pettigrew, 1992). It has been argued that the schema-inconsistent memory advantage (i.e. tendency to false alarm to schema-consistent information) may be a specific case of a general advantage for distinctive
information. Similar to the "von Restorff effect", where there is a recall advantage for a single word in uppercase in a list of lowercase words (von Restorff, 1933), it is expected that once expectations about an event or category structure are formed, the deviant item should attract more attention and thus have stronger memory trace.

Another account that aimed to explain inconsistent information memory advantage focused on the difference in the depth of processing. Since deviant items may be more difficult to process than regular items (Fabiani \& Donchin, 1995; Graesser, 1981), they tend to receive more study time (Stern, Marrs, Millar, \& Cole, 1984), and this leads to better memory. When study time is limited, there should be no advantage, or the pattern may even be reversed (Metcalfe, 2002; Thiede \& Dunlosky, 1999).

Studies in category learning also support the claim of better memory for exceptions. In an old-new recognition task, Palmeri and Nosofsky (1995) tested participants' memory for two newly learned categories. In both categories, the majority of items could be categorized based on a simple (singledimension) rule, but there was one exception item which respected the rule of the contrasting category. The main result of their study was that participants showed superior recognition memory for those items that were exceptions to the rule. These findings were in accordance with the prediction of the RULEX (rule-plus-exception) model of classification learning (Nosofsky, Palmeri and McKinley, 1994). According to this model, people tend to form simple logical rules to define categories, and if not all the members of the category follow the formed rule, those occasional exceptions are stored in memory. Thus, regular members of the category and exceptions are supposed to be learned using two independent mechanisms. Based on RULEX, the role of memory processes in categorization of the regular category items should be minimal, which stands in high contrast with purely memory based representation of the exception.

Similar to RULEX, SUSTAIN (Love, Medin, Gureckis, 2004) model assumes formation of specialized representations (clusters) for exceptions that violate initially formed representation (cluster), and predicts that differential storage of exceptions makes them more distinctive in memory. The main difference between the two models lies in flexibility. SUSTAIN emphasizes the need for a flexible search of a given category structure, and allows for clusters
to be of different nature (e.g. rules, prototypes, attractors), all depending on the (sub)structure of the category and the task goals. That way, in addition to successfully predicting memory advantage for exceptions, SUSTAIN is sensitive to effects of structure saliency (e.g. frequency effects), familiarity effects (differentiating between old and new rulefollowing items) or unsupervised learning, which are all problematic for RULEX to account for (Sakamoto \& Love, 2004).

Despite described differences, both RULEX and SUSTAIN are in accord with the previous categorization and schema literature regarding (a) memory advantage for ruleviolating exceptions, (b) deeper (at a greater detail) processing of exceptions compared to regular category members (Loftus \& Mackworth, 1978) and (c) attribution of memory advantage for exceptions to differential attention during encoding (von Restorff, 1933).

## Nature of the exceptions

It is important to note here that previous studies (e.g. Palmeri \& Nosofsky, 1995, Sakamoto and Love, 2004; Davis, Love, \& Preston, 2012) focused primarily on a specific type of exceptions - exceptions that violate prior knowledge expectations by respecting the contrasting category rule. Those exceptions could not be learned by relying on a rule, nor by relying on the similarity with the other category members. In order to be successfully categorized they required forming complex binding structures. Thus, it is unclear whether the better memory for exceptions results from exceptions being rare and violating the prior knowledge expectation (in this case, the category rule), or because of their peculiar structure.

## Developmental differences

All hypothesized solutions for learning and representing exceptions were formulated with an adult in mind. Little is known about how exceptions may be learned and represented early in development.

Both RULEX and SUSTAIN assume engagement of selective attention and (to different extent) optimization of memory resources during category learning. However, previous studies suggest that in contrast to adults and older children, who optimize their attention to category and task relevant dimensions, young children tend to allocate attention to both relevant and irrelevant information (Sloutsky, 2010; Deng \& Sloutsky, 2016). The developmental differences in attention allocation during category learning (i.e., selective vs. distributed) have important consequences on what is remembered about categories. While selective attention results in better memory for information that is particularly useful for distinguishing the categories (e.g. rule features), distributed attention results in all information, relevant and relevant, being remembered equally well (Deng \& Sloutsky, 2016). Thus, if difference in recognition memory for regular category members and exceptions arise from optimization and selectivity, as previously suggested, no difference in memory for regular and exceptional items should be expected for young children.

## Current study

In the two experiments reported here, we tested the generalizability of the assumption of memory advantage for exceptions. In Experiment 1, we tested the claim of memory advantage for exceptions in situation of learning categories that have exceptions that violate previous expectation since they look more like the members of the other category and they also violate the category rule. However, in contrast to the exceptions used in the previous studies that respect contrasting category rule, they have a new rule on the deterministic dimension, which is on its own sufficient for successful categorization. In Experiment 2, the structure of the regular category members remained the same, but the nature of the exceptions was changed. In Experiment 2, exceptions were items that had all features new. Since exceptions in our study are individuals, for this latter kind of exceptions, different kinds of rules could be formed, since each feature is fully predictive.

Although the exceptions used here are very different from the ones used in the previous studies, they retain all the characteristics that are assumed to contribute to their special status in memory. They are rare, they may be studied for unlimited amount of time, and, most importantly, they violate the expectations based on the knowledge of regular items, both in terms of rule and appearance. On the other hand, they could be categorized equally successfully by employing different learning mechanisms and forming different representations, which makes them advantageous in comparison to the types of exceptions that could be learned by memorizing only.

Additionally, we examined developmental differences in learning and representing exceptions. Two age groups participated in the experiments: four-year-olds and adults. As previously described, a developmental study is particularly interesting since it will allow test of differences in the memory status of exceptions under regimes of distributed end selective attention category learning.

## Experiment 1

## Method

## Participants

Participants were 27 four-year-old children (Mage $=54.6$ months, range $48.5-59.9$ months, 14 girls) and 36 adults.

Data of one additional child and one adult participant were excluded due to the failure to discriminate between old (High-Match) items and items that had 5 of 6 features completely new (All-new-P) (A' not different from chance level of 0.5 , one sample $t$-test $p s>.05$ ).

All four-year-olds that took part in the experiments reported in this paper were recruited from preschools located in middle-class suburbs of Columbus. In order to take a part in the study they had to be between 48 and 60 months old. They were tested during their regular school hours in a quiet room in their preschool.

All adults that participated in the experiments reported here were The Ohio State University undergraduate students. They were tested in a quiet room in the laboratory located on campus and they received course credits for their participation.

## Stimuli

Stimuli were artificial dinosaur-like creatures created using Spore Creature Creator and Gimp (Figure 1). These creatures were accompanied by two novel category labels: Lulu and Momo.


Figure 1. Examples of stimuli used in the study where hands are the rule feature.

## The category structure

The categories of Lulus and Momos were dense 7dimensional categories with 1 non-diagnostic dimension, 5 probabilistic dimensions and 1 deterministic dimension (Table 1). Non-diagnostic dimension varied independently and gave no information about the category membership. Probabilistic dimensions varied between categories and within-category, with significantly higher between-category in comparison to within-category variance. Hence, probabilistic dimensions were predictive when taken together, since they reflected the overall similarity between category items. Deterministic dimension was fully predictive.

The neck length (short/long) was always the nondiagnostic dimension. The other 6 dimensions were: antennas, mouth, belly, wings, hands and feet. All dimensions were binary. The choice of deterministic feature (belly or hands) was balanced across the participants. Table 1 presents the structure of training and test items.

During the training, only High-Match and Exception items were presented. High-Match items always respected the category rule and had most of the probabilistic features of their own category (4 of 5). Exceptions were designed so they look more like the other category members (they had probabilistic features of the other category) but they also had a new rule feature.

In the test session, in addition to items presented during the training (High-Match and Exceptions), there were additional

4 types of items. Those new items were based on High-Match and Exception items, but either had one probabilistic feature new (One-new-P, E-One-new-P), or all probabilistic features new (All-new-P, E-All-new-P).

## Design and procedure

For adults, all instructions and questions were written on the screen and they responded by pressing designated keys on a computer keyboard. For four-year-olds, all instructions and questions were read by a trained experimenter who collected their verbal responses using a computer keyboard.

Table 1: The abstract category structures used in Experiment 1 and Experiment 2.

|  | Momo |  | Lulu |  |
| :---: | :---: | :---: | :---: | :---: |
| Experiment 1 |  |  |  |  |
|  | Probabilistic | Rule | Probabilistic | Rule |
| High Match | 10000 | 0 | $\begin{array}{llllll}0 & 1 & 1 & 1 & 1\end{array}$ | 1 |
| New-D | $1 \begin{array}{lllll}1 & 0 & 0 & 0 & 0\end{array}$ | N | $\begin{array}{lllll}0 & 1 & 1 & 1\end{array}$ | N |
| One-new-P | $0 \begin{array}{lllll}0 & 0 & 0 & 0\end{array}$ | 0 | 1 N 11 | 1 |
| All-new-P | N N N N N | 0 | N N N N N | 1 |
| Exception | 111111 | 2 | 0 0 0 0 0 0 | 3 |
| E-One-new-P | $1 \begin{array}{lllll}\text { N } & 1 & 1\end{array}$ | 2 | 0 N 0000 | 3 |
| E-All-new-P | N N N N N | 2 | N N N N N | 3 |

Experiment 2

|  | Probabilistic |  |  |  |  | Rule | Probabilistic |  |  |  |  |  | ule |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| High Match |  | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |  |  |
| New-D |  | O | 0 | - | - | N | 0 | 1 | 1 | 1 | 1 |  | N |
| One-new-P | 0 | N | 0 | 0 | 0 | 0 | 1 | N | 1 | 1 | 1 |  |  |
| All-new-P |  | N N | N | N | N | 0 | N | N | N | N | N |  |  |
| Exception | 4 | 44 | 4 | 44 | 44 | 4 | 5 | 5 | 5 | 5 | 5 |  | 5 |
| E-New-D | 4 | 4 | 4 | 44 | 44 | N | 5 | 5 | 5 | 5 | 5 |  | N |
| E-One-new-P | 4 | N | 4 | 44 | 44 | 4 | 5 | N | 5 | 5 | 5 |  | 5 |
| E-All-new-P | N | N N | N | N N | N N | 4 | N | N | N | N | N |  | 5 |

## Instructions

After the cover story about the two dinosaur families, Momos and Lulus, was read, participants were presented with the prototypes of Momos and Lulus. The prototypes were presented together, on the same screen. Participants were told that that is how Momos and Lulus usually look like and each of the six features of the two creatures was introduced, using the sentence frame: "Momos/Lulus usually have antennas like these" and pointing to the named feature (Figure 2). Training

During the training participants were presented with the exemplars of Lulus and Momos and they were asked to classify them. Items were presented individually in the center of a white background screen, accompanied by the question "Is this is a Momo or a Lulu?" Two buttons, labeled Momo and Lulu, were presented on the same screen. Participants responded by pressing one of the buttons if adults, or giving verbal answers if children. After they made a response, corrective feedback was provided. Feedback had two
elements. First, the button of the correct response was presented (that is, the correct answer button stayed on the screen) and second, feedback sentence was presented. If participant gave the correct answer, she received message "That's right! That's Momo (Lulu)!" If participant made a wrong choice, message "Oops! That's Momo (Lulu)!" was presented.

During the training session, 70 items were presented: 60 High Match items and 10 Exceptions. In the first block only High Match items were presented (20), while in the second and the third block participants saw both High Match (20) and Exception (4) items in random order. At the end of the third block, we presented additional 2 Exception items. The logic behind the dynamics described was that in order for Exceptions to be seen as Exceptions they should be presented after a representation of High-Match items was formed, and they needed to be less frequent. Order of presentation was randomized for each participant.
Memory test
Memory test was introduced immediately after the training session. Items were presented individually, followed by the question „Did you see exactly the same creature in the first part of the game?" and two buttons labeled „old" and „new". After participant gave a response, the next trial was presented. There was no feedback.

Memory test was given in one block. It had 64 trials in total, 8 trials of each item type, presented in random order.

## Results

## Training performance

For both age groups, average accuracy in categorizing HighMatch items was above the chance (one-sample t-tests against chance yield $t \mathrm{~s}>2.70$, both $\mathrm{ps}<.05$, two-tailed).

Both groups of participants misclassified Exceptions. The average proportion of accurately classified items was .41 for four-year-olds and .33 for adults, based on performance on all 10 items presented during the training session (both bellow the chance, $t \mathrm{~s}>2.38$, $p \mathrm{~s}<.05$ ). Since Exceptions had probabilistic features of the other category High-Match items, participants based their responses on the overall similarity of exceptions to High-Match items.

## Recognition memory

In order to estimate participants' recognition memory, we calculated A' scores (Snodgrass, Levy-Berger, \& Haydon, 1985). A' is a non-parametric analogue of the d' statistic (Brophy, 1986) and it is a measure of discriminability. No discrimination (chance performance) is indicated by value of 0.5 . With better discrimination the A' score increases.

Both age groups demonstrated high recognition accuracy (old - All-new-P). Average memory sensitivity scores were well above chance (both $p s<.001$ ).

To examine the hypothesized differences in memory for regular items and exceptions, A-prime scores were subjected to a two-way (Age by Type) ANOVA.

For the overall memory (old - All-new-P) the analysis indicated that there was no significant main effect of item type ( $p>.05$ ), whereas the main effect of age was significant,
$F(1,120)=8.87, p<.01, \eta=.07$ with 4-year-olds' performance being significantly lower than adults' performance (Figure 2).

The pattern was the same for the memory for probabilistic features. Again, adults have shown better memory than 4-year-olds $(F(1,120)=8.61, p<.01, \eta=.07)$ and there was no difference in memory for regular items and exceptions.

Note here that the lack of difference in memory for regular and exceptional items may be due to poor learning. This problem is resolved in Experiment 2.


Figure 2. Memory sensitivity scores (A-prime) for overall memory (OLD - All-new-P) across age groups and two item types in Experiment 1. The dashed line represents the point of no sensitivity. Error bars represent the standard errors of the mean.

## Experiment 2

## Method

## Participants

Thirty-three four-year-olds (Mage $=52.6$ months, range 44.0 - 53.6 months, 12 girls) and 39 adults took part in Experiment 3. Three additional 4-year-olds and one adult were excluded based on the same criteria used in Experiment 1(A' not different from chance level of 0.5 , one sample $t$-test, all $p s$ >.05). Participants were recruitment from the same participants' pool as in Experiment 1.

## Stimuli

The structure of the High-Match items and accompanying test items was the same as in the Experiment 1. The same stimuli set was used. Exceptions were different. They had all features new. Exceptions were individuals, thus there was one exception of Momo and one of Lulu category.

In addition to the stimuli types used in the Experiment 1, there were 2 additional item types: New-D and New-D Exceptions (See Table 1).
Design and procedure
Design and procedure of Experiment 2 respected the one described for Experiment 1 in every respect. Memory test had

88 trials. The experiment took approximately 15-20 minutes for adults and 25-30 minutes for children.

## Results

## Training performance

Performance on High-Match items was above chance for both age groups ( $t s>4.78$, both $p s<.001$, two-tailed). Both children and adults learned to categorize exceptions (ts (36, $23)=3.57,6.64, p s<.001)$.

## Recognition memory

Overall memory sensitivity was high for both age groups and well above chance (both $p s<.001$ ) (Figure 3). Differences in overall memory (old - All-new-P) were tested in a 2 (Age: 4-year-olds, adults) x 2 (Type: regular, exception) ANOVA. Participants had better memory for regular items (A' scores), regardless of their age $(F(1,140)=4.88, p<.05, \eta=.03)$.


Figure 3. Memory sensitivity scores (A-prime) for overall memory (OLD - All-new-P) across age groups and two item types in Experiment 2. The dashed line represents the point of no sensitivity. Error bars represent the standard errors of the mean.

In order to test for differences in memory for ( P and D ) features, 2 (Age: 4 -year-olds, adults) x 2 (Type: regular, exception) x 2 (Feature: P, D) ANOVA was conducted on A' scores. The analysis revealed significant main effects of age $(F(1,280)=5.13, p<.05, \eta=.02)$ and type $(F(1,280)=$ 13.26, $p<.001, \eta=.05$ ), and a significant age by feature interaction $(F(1,120)=6.05, p<.05, \eta=.02)$ on A' scores. As expected based on the previous studies (Deng \& Sloutsky, 2016), adults, but not 4 -year-olds, have shown differential memory - specifically better memory for rule than probabilistic feature. Both age groups had better memory for features of High-Match items, than those of exceptions (Figure 4).

## Discussion

Results presented in this paper challenge assumptions of the models of classification learning like RULEX and SUSTAIN. In two experiments reported here we have shown
that both children and adults have better memory for features of regular items than features of exception. These findings have at least two important implications. First, they show that regular category members are not processed minimally, as it is suggested by models which assume high level of optimization of attention. Secondly, they also show that exceptional items are not represented fully, that is, they are not necessarily memorized. Not only that RULEX model cannot account for these findings, but it predicts completely the opposite pattern. Although SUSTAIN would fit the data better (especially good memory for regular items), the finding of better memory for exceptions runs counter to its assumptions.


Figure 4. Memory sensitivity scores (A-prime) for probabilistic and deterministic features of two item types in Experiment 2 (Panel A: four-year-olds; Panel B: adults). The
dashed line represents the point of no sensitivity. Error bars represent the standard errors of the mean.

The representation of an exception depends on its nature. When exceptions violate categories defined by rules by respecting contrasting category rule and can be learned only if there is binding of features, there is better memory for exceptions than regular items, as shown in the previous studies (e.g. Palmeri \& Nosofsky, 1995, Sakamoto and Love, 2004). However, when the nature of exceptions, and the nature of a category they belong to, allows for more flexible approach, participants tend to optimize. In the case of the exceptions used in this study, they could be equally successfully categorized based on different representations, some of which could simply contain memory for one of the item's features. However, despite their easy-to-learn structure, if special status of exceptions is to be attributed to the fact they violate previous knowledge expectations, they are schema-inconsistent and rare, exceptions in our study would also be processed with more attention, more deeply and they would have stronger memory trace. However, this was not the case.

Contrary to the predictions based on schema literature, that participants are more prone to notice missing features or new features in schema-inconsistent items than schema-consistent items (Friedman, 1979; Goodman, 1980), our participants were more sensitive when we changed regular items' features.

In addition to the difference in memory for regular category members and exceptions, developmental differences in memory were also found. While adults had better memory for rule, than for probabilistic features, 4-yearolds didn't show differential memory. This pattern is in accordance with previous studies (Deng and Sloutsky, 2016). However, contrary to our predictions, there were no developmental differences in memory status of exceptions. Both age groups had better memory for regular category members, despite differences in attention allocation (selective vs. distributed).

Taken together, findings of this study suggest new directions for models of category learning and memory, by providing new evidence on attention and memory optimization during category learning.

## Acknowledgments

We would like to thank Sophia Deng for her comments on the design of the study. This research was supported by NIH grants R01HD078545 and P01HD080679 to VMS.

## References

Davis, T., Love, B.C., \& Preston, A.R. (2012). Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members. Cerebral Cortex, 22, 260-273.

Deng, S., \& Sloutsky, V. (2016) Selective attention, diffused attention, and the development of categorization. Cognitive Psychology, 91, 24-62.
Fabiani, M., \& Donchin, E. (1995). Encoding processes and memory organization: A model of the von Restorff effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 224-240.
Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. Journal of Experimental Psychology: General, 108, 316355.

Goodman, G. S. (1980). Picture memory: How the action schema affects retention. Cognitive Psychology, 12, 473495.

Graesser, A. C. (1981). Prose comprehension beyond the word. New York: Springer-Verlag.
Loftus, G. R., \& Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Perception and Performance, 4, 565-572.
Love, B. C., Medin, D. L., \& Gureckis, T. M. (2004). SUSTAIN: A network model of human category learning. Psychological Review, 111, 309-332.
Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? Journal of Experimental Psychology: General, 131, 349-363.
Nosofsky, R. M., Palmeri, T. J., \& McKinley, S. C. (1994). Rule-plusexception model of classification learning. Psychological Review, 101, 53-79.
Palmeri, T. J., \& Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 548-568.
Rojahn, K., \& Pettigrew, T. F. (1992). Memory for schemarelevant information: A meta-analytic resolution. British Journal of Social Psychology, 31(2), 81-109.
Sakamoto, Y., \& Love, B. C. (2004). Schematic influences on category learning and recognition memory. Journal of Experimental Psychology: General, 133, 534-553.
Sloutsky, V. M. (2010) From perceptual categories to concepts: What develops? Cognitive Science, 34, 12441286.

Stern, L. D., Marrs, S., Millar, M. G., \& Cole, E. (1984). Processing time and recall of inconsistent and consistent behaviors of individuals and groups. Journal of Personality and Social Psychology, 47, 253-262.
Thiede, K. W., \& Dunlosky, J. (1999). Toward a general model of selfregulated study: An analysis of selection of items for study and selfpaced study time. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1024-1037.
von Restorff, H. (1933). Analyse von vorgängen im spurenfeld: I. Über die wirkung von bereichsbildungen im spurenfeld [Analysis of processes in the memory trace: I. On the effect of group formations on the memory trace]. Psychologische Forschung, 18, 299-342.

# Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior 

K.J. Savinelli, Gregory Scontras, and Lisa Pearl<br>\{ksavinel, g.scontras, lpearl\} @uci.edu<br>University of California, Irvine


#### Abstract

Investigations of scope ambiguity resolution suggest that child behavior differs from adult behavior, with children struggling to access inverse scope interpretations. For example, children often fail to accept Every horse didn't succeed to mean not all the horses succeeded. Current accounts of children's scope behavior involve both pragmatic and processing factors. Inspired by these accounts, we use the Rational Speech Act framework to articulate a formal model that yields a more precise, explanatory, and predictive description of the observed developmental behavior.


Keywords: Rational Speech Act model, pragmatics, processing, language acquisition, ambiguity resolution, scope

## Introduction

If someone says "Every horse didn't jump over the fence," do you think any horses made it over the fence? If you think not, then you've interpreted this utterance as something like (1a). In contrast, if you think it's possible some horses made it, you've interpreted this utterance as something like (1b). These two different interpretations are possible because the utterance is scopally ambiguous. That is, it contains two scope operators: a quantifier (every $=\forall$ ) and a negation ( $n^{\prime} t=\neg$ ). Either element can take scope over the other (indicated as $\gg$ in (1)), and so yield two different interpretations.
(1) Every horse didn't jump over the fence.
a. $\quad \forall \gg \neg$ (surface scope):

None of the horses jumped over the fence.
b. $\quad \neg \gg \forall$ (inverse scope):

Not all of the horses jumped over the fence.
While adults can access both interpretations given appropriate context, 5 -year-old children typically struggle to obtain the inverse scope in (1b) (Musolino, 1998; Lidz \& Musolino, 2002; Musolino \& Lidz, 2006; Musolino, 2006; Viau et al., 2010; Tieu, 2015). For example, in a context where two out of three horses did in fact jump over the fence, only the inverse scope interpretation in (1b) is true. Adults charitably interpret the ambiguous utterance in a way that makes it a true statement (i.e., with the inverse scope in a two-out-of-three scenario), but 5-year-olds stick with the surface interpretation in (1a), which is false. Why does children's behavior differ from adults' in this context?

Previous accounts of children's scope interpretation behavior have recognized that both processing and pragmatic factors may contribute to non-adult-like behavior. Musolino $(1998,2006)$ observed that the surface scope interpretation in (1a) may be easier to process because the scope relationship in the semantics (i.e., $\forall$ scopes over $\neg$ ) aligns with the linear
order of these elements in the utterance (i.e., Every precedes $n ' t)$. In contrast, for the inverse scope interpretation in (1b), this isomorphism does not hold, with the scope relationship (i.e., $\neg$ scopes over $\forall$ ) opposite the linear order of the elements in the utterance. Musolino hypothesized that this lack of isomorphism would make the inverse scope interpretation more difficult to access. In line with this prediction, Conroy et al. (2008) found that when adults are time-restricted, they favor the surface scope interpretation. We thus see a potential role for processing factors in children's inability to access the inverse scope. Perhaps children, with their still-developing processing abilities, can't allocate sufficient processing resources to reliably access the inverse scope interpretation.

In addition to this processing factor, Gualmini et al. (2008) noted that discourse properties, such as what children consider the question under discussion (QUD), may impact their scope interpretation behavior. Formal theories of pragmatics suggest that all discourse transpires with respect to some QUD, whether implicit or explicit; utterances in the discourse need to (at least partially) answer the QUD to be pragmatically felicitous (Roberts, 2012). Gualmini and colleagues (Hulsey et al., 2004; Gualmini et al., 2008) suggest that children are very sensitive to this requirement. In particular, children may be able to access the inverse scope interpretation but nonetheless choose the surface scope interpretation because it better answers the perceived QUD in the contrived experimental setups. So, children's observed behavior would derive from a still-developing ability to manage the contextual information available and correctly infer the intended QUD.

Thus, children's developing processing and pragmatic abilities may both be a source of the observed non-adult-like behavior (Viau et al., 2010), though current experimental studies have struggled to clearly isolate the influence of each type of factor. To this end, we formally articulate the mechanism of scope ambiguity resolution using the Bayesian Rational Speech Act (RSA) computational modeling framework (Frank \& Goodman, 2012; Goodman \& Frank, 2016) in order to identify the separate contributions of processing and pragmatic factors.

We first summarize key experimental results from the literature on child scope ambiguity resolution, noting three core variables (one processing, two pragmatic) that affect children's scope disambiguation behavior. We also highlight the nature of the task children are being asked to engage in, which we then formally articulate using an RSA model that specifies the role of each of these three variables. Our results suggest that pragmatic factors play a larger role than processing fac-
tors in explaining children's non-adult-like scope ambiguity resolution behavior, and the computational modeling framework allows us to understand exactly why that's so. These results additionally suggest targeted future behavioral experiments to verify the impact of the specific pragmatic factors we identify. More generally, our model yields a more precise, explanatory, and predictive description of the observed developmental scope ambiguity resolution behavior.

## Background: Experimental results

Children's ability to access the inverse scope interpretation has been shown to be sensitive to manipulations of experimental context. The methodology typically used to assess children's scope disambiguation is the Truth Value Judgment Task (TVJT; Crain \& McKee, 1985). In the basic TVJT, children are presented with a background story about the actorsfor example, horses engaging in some activities. After this background story, children watch as the horses attempt to complete an action, such as jump over a fence. The critical not-all result state meant to prompt the inverse scope interpretation is illustrated in Figure 1, where the white horse fails to jump over the fence.


Figure 1: Sample not-all scenario from Musolino and Lidz (2006): 2 of 3 horses succeed at jumping over the fence.

In this scenario, the surface scope interpretation of the scopally-ambiguous utterance Every horse didn't jump over the fence (i.e., none of the horses jumped over the fence) is false, and the inverse scope interpretation (i.e., not all of the horses jumped over the fence) is true. A puppet then says the scopally-ambiguous utterance, and the child is asked to state if the puppet is right. That is, the child is asked whether $\mathrm{s} / \mathrm{he}$ would endorse the puppet's utterance as a true description of the scenario. Typically, children refuse to endorse the puppet's utterance, saying that the puppet is wrong. This behavior has been interpreted as children failing to access the inverse scope interpretation that would make the utterance true.

Interestingly, various alterations to the TVJT setup have yielded more adult-like behavior in children, namely greater rates of endorsing the puppet's ambiguous utterance in not-all scenarios. Musolino and Lidz (2006) observed that negation in an utterance might require certain felicity conditions to be met. In particular, negated utterances require a preceding $a f$ firmative context to contrast with (Wason, 1965). Musolino
and Lidz augmented the basic TVJT to include an additional contrast condition in which the puppet precedes its negative scopally-ambiguous utterance with a contrasting affirmative clause. This additional clause describes a previous successful story action (i.e., early-success), such as Every horse jumped over the log, but every horse didn't jump over the fence. This early-success contrast manipulation increased children's willingness to accept the scopally-ambiguous utterance in the not-all scenario: Children in the baseline condition endorsed the puppet's statement just $15 \%$ of the time, while children in the early-success affirmative context condition endorsed the puppet's statement $60 \%$ of the time. Viau et al. (2010) later replicated this increase in utterance endorsement using only an early-success story context. That is, the utterance endorsement rate was maintained by an early-success story context alone, and children didn't need an explicit contrast clause in the test utterance.

Notably, the early-success affirmative context manipulation potentially changes several aspects of the experimental context. First, it can shift participants' expectations about successful outcomes in the experimental world. This shift then potentially increases the salience of a QUD targeting this success, such as "Did all the horses succeed?" (all?). Recognizing this QUD's potential significance, Gualmini (2004) attempted to manipulate the experimental context so it favored the all? QUD. With all? as the salient QUD, children's endorsement of a scopally-ambiguous utterance that perfectly answers all? in the critical not-all scenario increased to $90 \%$. Even for a scopally-ambiguous utterance that does not answer the all? QUD, children's endorsement rate was at $50 \%$-markedly higher than the $15 \%$ baseline from the original study by Musolino and Lidz (2006). This finding highlights that privileging the all? QUD increases children's utterance endorsement in these scenarios.

A third potential impact of the affirmative context manipulation involves scope access. By altering the experimental world expectations and/or QUD to increase access to the inverse scope, the inverse scope interpretation may also become more accessible for later use. Viau et al. (2010) term this structural priming. Children who are better able to access the inverse scope are then more likely to endorse the scopallyambiguous utterance in subsequent not-all scenarios. Viau et al. investigated structural priming explicitly by attempting to directly alter the accessibility of the inverse scope interpretation. In one modified TVJT, they attempted to prime the access of the inverse scope interpretation, and in another modified TVJT, they attempted to directly prime the inverse scope's logical structure (e.g., $\neg \gg \forall$ ).

The first structural priming manipulation was implemented via the now-familiar affirmative context (i.e., pragmatic) manipulation. For the first three trials, the prior experimental context indicated successful outcomes and the effect was that children endorsed the scopally-ambiguous utterance $50 \%$ of the time. Crucially, the subsequent three trials removed the supportive affirmative context manipulation-yet children
continued to not only endorse the scopally-ambiguous utterance, but to endorse it more than they had before ( $80 \%$ ). Viau et al. (2010) attribute this result to a priming effect of the inverse interpretation from the first three trials. Interestingly, the increase in utterance endorsement could be due to priming multiple factors that are products of the affirmative context manipulation: (i) the expectations about successful outcomes in the experimental world, (ii) the salience of the all? QUD, or (iii) the ease of access to the inverse scope interpretation.

The second structural priming manipulation removed the affirmative context story in the first three trials. In its place, children were asked whether they would endorse a scopallyunambiguous utterance (e.g., "Not every horse jumped over the fence") whose interpretation had logical operators in the same order as the inverse scope interpretation of the scopallyambiguous utterance (e.g., $\neg \gg \forall$ ). Children endorsed this utterance $80 \%$ of the time. In the subsequent three trials, children were asked if they would endorse the scopallyambiguous utterance in the same experimental scenario-and their endorsement rate remained at $80 \%$. Viau et al. (2010) interpret this effect as priming of the relevant logical form: The inverse scope was easier to access in the scopally-ambiguous utterance because it was so recently accessed in the unambiguous utterances. The authors argue that this priming effect proceeded in the absence of manipulations to the pragmatic context, yet even here, there may still be pragmatic factors at work. The unambiguous utterance accomplishes three things: (i) it provides an instance of the $\neg \gg \forall$ configuration, (ii) it provides information about successful outcomes, and (iii) it suggests the all? QUD, answering it with no. Thus, in this attempt to prime the inverse logical form, the authors may have also altered expectations about the pragmatic context of the experiment, as related to the successful outcomes and relevant QUDs.

These experimental studies highlight at least three core factors (two pragmatic, one processing) that underlie children's utterance endorsement behavior in the TVJT: (i) pragmatic: expectations about the experimental world (e.g., how likely successful outcomes are), (ii) pragmatic: expectations about the QUD (e.g., whether all outcomes were successful), and (iii) processing: the accessibility of the inverse scope (i.e., the ease by which the logical form is accessed). These experimental studies have also supported different theoretical proposals for the source of children's differences. The proposals split on whether they attribute the differences solely to an inability to manage contextual information (i.e., pragmatic factors; Gualmini, 2008) or whether processing deficits also significantly contribute (i.e., difficulty accessing inverse scope; Viau et al., 2010). Importantly, it is not obvious from any of the existing experimental manipulations how to separate the independent contributions of these components. To capture and independently manipulate the contributions of each of the pragmatic and processing factors, we formalize their role in the interpretation of scopally-ambiguous utterances, using tools from probabilistic modeling.

## The model

We model ambiguity resolution within the Bayesian Rational Speech Act (RSA) framework (Goodman \& Frank, 2016). This framework views language understanding as a social reasoning process. A pragmatic listener $L_{1}$ interprets an utterance by reasoning about a cooperative speaker $S_{1}$ who is trying to inform a literal listener $L_{0}$ about the world. Our model is a "lifted-variable" extension wherein the ambiguous utterance's literal semantics is parameterized by interpretation-fixing variables (e.g., the relative scope of the quantificational elements; Lassiter \& Goodman, 2013). Hearing an ambiguous utterance, a pragmatic listener reasons jointly about the true state of the world (e.g., how many horses jumped over the fence), the scope interpretation that the speaker had in mind (i.e., surface vs. inverse), as well as the likely QUD that the utterance addresses (e.g., all?).

To connect our model's predictions with the available TVJT data, we follow Degen and Goodman (2014) and Tessler and Goodman (2016), modeling participants’ TVJT behavior as the (relative) endorsement of a pragmatic speaker $S_{2}$ for an utterance about an observed situation. That is, we model whether a speaker would endorse the scopallyambiguous utterance as a description of the observed state, or whether the speaker would prefer to say nothing at all. The pragmatic speaker $S_{2}$ makes this decision by reasoning about the probability that a pragmatic listener $L_{1}$ (who is reasoning about a speaker $S_{1}$ reasoning about a literal listener $L_{0}$ ) would arrive at the correct world state after hearing the utterance.

We take world states $w \in W$ to correspond to the number of successful outcomes, for example, the horses that successfully jumped over the fence $(W=\{0,1,2,3\})$. We assume a simple truth-functional semantics where an utterance $u$ denotes a mapping from world states to truth values (Bool $=\{$ true, false $\}$ ). We parameterize this truth function so that it depends on the scope interpretation $i \in I=$ \{inverse, surface\}, $[u]_{]^{i}}^{i}: W \rightarrow$ Bool. We consider two alternative utterances $u \in U$ : the null utterance (i.e., saying nothing at all, and so choosing not to endorse the utterance) and the scopally ambiguous utterance amb (e.g., "Every horse didn't jump over the fence"). So, $U=\{$ null, amb $\}$. The utterance semantics appears in (2), where the parameterization only impacts the truth value for utterance amb (since that's when multiple interpretations are available). If inverse is active, this corresponds to the not all reading, and so is true as long as not all (i.e., $w \neq 3$ ) outcomes were successful. If surface is active, this corresponds to the none reading, and so is only true in world state 0 .
(2) Utterance semantics $\llbracket u \rrbracket^{i}$ :
a. $[$ null $]]^{i}=$ true
b. $\quad[\mathrm{amb}]^{i}=$ if $i=$ inverse [inverse],
else [【surface]】
where:

$$
\llbracket \text { inverse }]=\lambda w . w \neq 3
$$

$$
\llbracket \text { surface }]=\lambda \mathrm{w} \cdot \mathrm{w}=0
$$

We consider three QUDs $q \in Q$ : (i) "How many horses made it over?" (how-many?), (ii) "Did all the horses make it over?" (all?), and (iii) "Did none of the horses make it over?" (none?). The QUDs serve as projections from the inferred world state to the relevant dimension of meaning, $q: W \rightarrow X$ (Kao, Wu, et al., 2014; Kao, Bergen, \& Goodman, 2014). In practice, the QUDs establish partitions on the possible world states, as shown in (3): how-many? is an identity function on world states, all? returns true only if all three outcomes were successful, and none? returns true only if none of the outcomes were successful.

QUD semantics $\llbracket q \rrbracket$ :
a. $\quad[$ how-many? $]]=\lambda \mathrm{w} . \mathrm{w}$
b. $\quad[$ all? $]]=\lambda \mathrm{w} . \mathrm{w}=3$
c. $[$ none? $]=\lambda w . w=0$

The literal listener $L_{0}$ has prior uncertainty about the true state, $P(w)$, and updates beliefs about $w$ conditioned on the the literal semantics. That is, $L_{0}$ restricts prior beliefs to those worlds that $\left[[u]^{i}\right.$ maps to true. The function $\delta_{\llbracket u]^{i}(w)}$ maps the Boolean truth value to a probability, 1 or 0.

$$
P_{L_{0}}(w \mid u, i) \propto \delta_{[u]^{i}(w)} \cdot P(w)
$$

To capture the notion that communication proceeds relative to a specific QUD $q, L_{0}$ must infer not only the true world state $w$, but also the value of the QUD applied to that world state, $[\lfloor q](w)=x$.

$$
P_{L_{0}}(x \mid u, i, q) \propto \sum_{w} \delta_{x=\llbracket q \rrbracket(w)} \cdot P_{L_{0}}(w \mid u, i)
$$

The speaker $S_{1}$ chooses an utterance $u$ in proportion to its utility in communicating about the true state of the world $w$ with respect to the QUD $q, \llbracket q \rrbracket \rrbracket(w)=x$. Thus, the speaker maximizes the probability that $L_{0}$ arrives at the intended $x$ from $u$. This selection is implemented via a softmax function (exp) and free parameter $\alpha$, which controls how rational the speaker will be in utterance selection.

$$
P_{S_{1}}(u \mid w, i, q) \propto \exp \left(\alpha \cdot \log \left(L_{0}(x \mid u, i, q)\right)\right)
$$

Utterance interpretation happens at the level of the pragmatic listener $L_{1}$, who interprets an utterance $u$ to jointly infer the world state $w$, the interpretation $i$, and the QUD $q$. We therefore model ambiguity resolution as pragmatic inference over an under-specified utterance semantics (i.e., the interpretation variable $i$ ). To perform this inference, $L_{1}$ inverts the $S_{1}$ model by Bayes rule, and so the joint probability of $w, i$, and $q$ is proportional to the likelihood of $S_{1}$ producing utterance $u$ given world state $w$, interpretation $i$, and QUD $q$, as well as the priors on $w, i$, and $q$.

$$
P_{L_{1}}(w, i, q \mid u) \propto P_{S_{1}}(u \mid w, i, q) \cdot P(w) \cdot P(i) \cdot P(q)
$$

To model the utterance endorsement implicit in TVJT, we need one more level of inference. The pragmatic speaker $S_{2}$ observes the true world state $w$ and selects $u$ by inverting the
$L_{1}$ model, thus maximizing the probability that a pragmatic listener would arrive at $w$ from $u$ by summing over possible interpretations $i$ and QUDs $q$ that accompany world $w$.

$$
P_{S_{2}}(u \mid w) \propto \exp \left(\log \sum_{i, q} P_{L_{1}}(w, i, q \mid u)\right)
$$

To generate model predictions, we must fix various model parameters. The $S_{1}$ speaker rationality parameter $\alpha>0$ is set to 2.5. The priors $P(w)$ and $P(q)$ correspond to expectations for the discourse context (i.e., likely world states or QUDs). In the default case, we set these priors to be uniform over their possible values: $P(w=0)=P(w=1)=P(w=2)=P(w=3)=\frac{1}{4}$; $P($ how-many $?)=P($ all $?)=P($ none $)=\frac{1}{3}$. The interpretation prior $P(i)$ corresponds to how easy it is to access the inverse scope interpretation. Experimental literature on scope ambiguity resolution suggests that speakers more readily access the surface interpretation (Anderson, 2004; Conroy et al., 2008). We model this tendency by setting these default values: $P($ surface $)=0.7$ and $P($ inverse $)=0.3$. Importantly, to better understand children's utterance endorsement behavior with scopally-ambiguous utterances, we can independently manipulate the values of the priors on $W, Q$, and $I$, and observe their impact on utterance endorsement.

## Results

To test how pragmatic and processing factors contribute to non-adult-like utterance endorsement in the TVJT, we systematically manipulate the relevant priors to favor specific parameter values, shown in Figure 2.

For the world state prior (Figure 2, left), we systematically favor specific world states by setting their prior probability to 0.9 ; if a world state is not favored, it receives a prior probability of $0.1 / 3=0.033$. Holding the QUD and scope priors at their default values, we see a marked increase in endorsement of the ambiguous utterance in the not-all scenario as beliefs about horse success increase. Utterance endorsement is at its lowest ( 0.25 ) when prior knowledge suggests that horses are particularly unlikely to succeed at jumping; utterance endorsement is at its highest (i.e., most adult-like: 0.90) when we believe horses are very likely to succeed.

Just as with the world state prior, we can systematically manipulate the QUD prior (Figure 2, center). Favored QUDs receive a prior probability of 0.9 ; other QUDs receive a prior probability of 0.05 . Holding the other priors at their default values, we see an increase in utterance endorsement from the none? (did no horses succeed?; 0.28) to how-many? (how many horses succeeded?; 0.37) to all? (did all horses succeed?; 0.64) QUDs. The model predicts the most adult-like behavior when the QUD concerns whether all the horses succeeded.

Finally, for the binary scope prior (Figure 2, right), we systematically manipulate the prior probability of inverse from 0.1 to 0.9 . Holding the other priors at their default values, we see a monotonic increase in utterance endorsement as the probability of inverse increases. At its most adult-like, the


Figure 2: Model predictions for ambiguous utterance endorsement (e.g., Every horse didn't jump over the fence) in a not-all scenario (e.g., two-out-of-three horses jump over the fence). Lower endorsement probability corresponds to less adult-like (i.e., more child-like) behavior. For the pragmatic variables (world state, QUD), the favored parameter value receives most of the prior probability weight $(P($ favored $)=0.9)$. For the processing variable (scope), the prior corresponds to how strongly the inverse scope is favored.
model predicts an endorsement probability of 0.57 when the prior probability of inverse is at its highest (0.9)—at its lowest (0.1), endorsement only drops to 0.4 .

To summarize, the world state and QUD priors have a more dramatic impact on utterance endorsement than the scope prior. There are two main reasons for this. First, for the world state prior, when expectations favor success (i.e., $w=3$ ), the ambiguous utterance is maximally informative regardless of the scope interpretation it receives: amb communicates to a listener that prior expectations do not hold (i.e., None/Not all of the horses succeeded goes against the expectation that all three horses would succeed). So, amb is particularly useful for communicating about the a priori unlikely not-all world states that appear in the experimental scenarios. Second, for the QUD manipulation, when all? is favored, either interpretation of amb fully resolves the QUD: whenever amb is true (i.e., whether None or Not all the horses succeeded), it is not the case that all the horses succeeded. A pragmatic speaker recognizes the utility of amb as an answer to all? in a not-all world state, irrespective of the intended scope interpretation.


Figure 3: Model predictions for ambiguous utterance endorsement when optimal world state $(w=3)$ and optimal QUD (all?) are favored $(P($ favored $)=.9)$.

So far, we have considered independent manipulations to the factors of interest. Figure 3 shows the interaction of all three factors for utterance endorsement when $w=3$ and all?
are favored. Here we see the additive effects of the world state and QUD priors; together, they lead to near-total endorsement of the ambiguous utterance. We also see more clearly the relatively small contribution of the scope prior, where changing the prior probability of inverse from 0.1 to 0.9 leads to just a 0.01 increase in endorsement probability. Thus, we see how the priors on the pragmatic factors overwhelm the processing scope prior. When the optimal (i.e., optimal for endorsement) QUD and world state are favored, even when inverse is highly inaccessible (i.e., $P$ (inverse) $=0.1$ ), we still predict massive utterance endorsement (0.99).

## Discussion

Our model of ambiguity resolution qualitatively captures the changes in children's utterance endorsement from the experimental literature; our results suggest that when it comes to understanding non-adult-like behavior in the TVJT, there is a stronger role for the pragmatics of context management (as realized in priors on world state and QUD) than for grammatical processing (as realized in the prior on scope interpretations), although there is likely a role for both. So, the observed failure of children to endorse scopally-ambiguous utterances in not-all scenarios likely stems more from children's beliefs about the world of the experiment (e.g., whether horses are a priori likely to succeed) and about the topic of conversation (e.g., whether the conversational goal is to determine if all the horses succeeded), than their inability to grammatically derive the inverse scope interpretation in real time. Indeed, our model predicts the highest rates of utterance endorsement to occur when resolving the scope ambiguity is irrelevant for communicating successfully about the not-all world-that is, when expectations favor total success (i.e., $w=3$ ), or when the QUD asks if all? of the horses succeeded. In either case, both scope interpretations serve to inform a listener, either that the a priori likely $w=3$ isn't true or that the answer to the all? QUD is no.

These results also underscore the need for well-defined
mapping hypotheses from observed experimental behavior to the psychological processes they inform, particularly for the sophisticated reasoning that occurs in tasks like the TVJT. In our brief review of the experimental literature, we were careful to point out alternative interpretations of the various experimental manipulations and their potential, unintended pragmatic consequences. At the very least, we hope to have demonstrated that utterance endorsement is not simple. A TVJT participant must reason recursively about the potential informativity of the utterance, attending to knowledge about the world of the experiment and the likely topic of conversation. That children stumble when attempting to perform these complex recursive inferences isn't so surprising. We suggest that a plausible source of differences in child and adult behavior on the TVJT is children's inability to successfully manage pragmatic information. We therefore propose to move the discussion away from the fragility of accessing inverse scope in children as a grammatical processing deficit and toward the complexity of behavior that scope interpretations require.

In addition to formalizing the pragmatics of ambiguity resolution in context, our results also motivate future experimental investigations that explicitly measure children's (and adults') expectations about the world and topic of conversation. We saw how past experiments did not completely deconfound the relevant factors. Perhaps the most straightforward way of testing these factors' effects is to measure the prior knowledge that participants bring to bear in the TVJT. These explicit measurements of pragmatic context can then form the basis of future modeling studies in this framework that could quantitatively match the behavioral results.

More generally, our results provide the foundation for more complete theories of the developmental process underlying scope ambiguity resolution. Children's relative lack of experience managing world and conversational knowledge likely contributes to their sensitivity to the experimental context. In short, five-year-olds may know the right interpretation, but they're still figuring out whether it's the best answer in the context of the experimental conversation.

## References

Anderson, C. (2004). The structure and real-time comprehension of quantifier scope ambiguity (Unpublished doctoral dissertation). Northwestern University.
Conroy, A., Fults, S., Musolino, J., \& Lidz, J. (2008). Surface scope as a default: The effect of time in resolving quantifier scope ambiguity. In Poster presented at the 21st cuny conference on sentence processing, march (Vol. 13).
Crain, S., \& McKee, C. (1985). The acquisition of structural restrictions on anaphora. In Proceedings of nels (Vol. 15, pp. 94-110).
Degen, J., \& Goodman, N. D. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In Proceedings of the 36th annual conference of the cognitive science society (pp. 397-402).

Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336, 998998.

Goodman, N. D., \& Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences, 20, 818-829.
Gualmini, A. (2004). Some knowledge children don't lack. Linguistics, 957-982.
Gualmini, A. (2008). The rise and fall of isomorphism. Lingиа, 118(8), 1158-1176.
Gualmini, A., Hulsey, S., Hacquard, V., \& Fox, D. (2008). The question-answer requirement for scope assignment. Natural language semantics, 16(3), 205-237.
Hulsey, S., Hacquard, V., Fox, D., \& Gualmini, A. (2004). The question-answer requirement and scope assignment. MIT working papers in Linguistics, 48, 71-90.
Kao, J. T., Bergen, L., \& Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In Proceedings of the 36th annual meeting of the cognitive science society (pp. 719-724).
Kao, J. T., Wu, J. Y., Bergen, L., \& Goodman, N. D. (2014). Nonliteral understanding of number words. Proceedings of the National Academy of Sciences, 111(33), 12002-12007.
Lassiter, D., \& Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Semantics and Linguistic Theory (SALT) 23 (pp. 587-610).
Lidz, J., \& Musolino, J. (2002). Children's command of quantification. Cognition, 84(2), 113-154.
Musolino, J. (1998). Universal grammar and the acquisition of semantic knowledge: An experimental investigation into the acquisition of quantifiernegation interaction in english (Doctoral dissertation). University of Maryland, College Park.
Musolino, J. (2006). Structure and meaning in the acquisition of scope. In Semantics in acquisition (pp. 141-166). Springer.
Musolino, J., \& Lidz, J. (2006). Why children aren't universally successful with quantification. Linguistics, 44, 817852.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. Semantics and Pragmatics, 5(6), 1-69.
Tessler, M. H., \& Goodman, N. D. (2016). A pragmatic theory of generic language. (http://arxiv.org/abs/1608.02926)
Tieu, L. (2015). Isomorphism for all (but not both): Floating as a means to investigate scope. Language Acquisi-tion(ahead-of-print), 1-16.
Viau, J., Lidz, J., \& Musolino, J. (2010). Priming of abstract logical representations in 4-year-olds. Language Acquisition, 17(1-2), 26-50.
Wason, P. C. (1965). The contexts of plausible denial. Journal of verbal learning and verbal behavior, 4(1), 7-11.

# Surprisingly: Marker of Surprise Readings or Intensifier? 

Anthea Schöller and Michael Franke<br>(anthea.schoeller@uni-tuebingen.de, mchfranke@gmail.com)<br>Department of Linguistics, Wilhelmstraße 19<br>72074 Tübingen, Germany


#### Abstract

We investigate the influence of the adverb surprisingly on the meaning of the quantity words few and many, which themselves have been associated with a reading expressing surprise. To learn about the meaning contribution of "surprise", we compare surprisingly with the intensifier incredibly and a compared to phrase explicitly marking surprise. Based on an empirical measure of subjects' expectations about everyday events, a Bayesian model uses data from a sentence judgment task to infer likely levels of surprise associated with the different constructions of interest.


Keywords: intensifier, surprise, computational modeling, few, many, surprisingly

## Introduction

A long tradition in psychology has acknowledged the role of prior expectations in the use of vague and context-dependent expressions like tall, heavy, few and many (e.g. Clark, 1991; Sanford, Moxey, \& Paterson, 1994). Fernando and Kamp (1996) spell out a semantic theory which makes the truth conditions of few and many dependent on prior expectations. Socalled "cardinal surprise readings" convey that a cardinality is smaller or greater than what is expected for the situation:
(1) For a man from the US, Chris saw few/many movies last year. $\rightsquigarrow$ Chris saw less/more movies than expected for a US male.

Such a surprise-based account raises interesting questions. First, how can expected cardinalities be distinguished from surprising ones? Fernando and Kamp (1996) stipulate that the lexical meanings of few and many comprise contextually-stable thresholds $\theta_{\text {few }}$ and $\theta_{\text {many }}$ which operate on a contextually-variable representation of a priori expectations. Second, if sentences with few and many express that a cardinality is surprising anyway, are they different from sentences in which the surprise element is overtly marked? The surprise reading can be made salient by a compared to phrase (2) or by modifying few and many with an adverb like surprisingly (3).
(2) Compared to what you would expect for a man from the US, Chris saw few / many movies last year.
(3) For a man from the US, Chris saw surprisingly few / many movies last year.

If surprise were the single factor which determines truth conditions of the cardinal surprise reading, we should not find a meaning difference between (1) and the overtly marked surprise reading in (2) and (3). Alternatively, it could be hypothesized that surprisingly in (3) acts not as a marker of surprise but as an intensifier, yielding a higher $\theta_{\text {surpr. many }}$ than $\theta_{\text {many }}$
and a lower $\theta_{\text {surpr. few }}$ than $\theta_{\text {few }}$. The pragmatic theory of intensifiers by Bennett and Goodman (2015) would predict that surprisingly has very similar effects to incredibly (see below).

We set out to experimentally test the influence of the modifiers surprisingly, incredibly and compared to on the threshold values predicted for Fernando and Kamp (1996)'s surprise readings of few and many. We employ linear mixed effects regression to compare judgment data and a computational model to infer said thresholds from our data.

## A Surprise-based Semantics for few and many

Partee (1989) characterized cardinal many as describing cardinalities which are at least $x_{\min }$, where $x_{\min }$ is a large number, and few as describing cardinalities which are at most $x_{\text {max }}$, where $x_{\text {max }}$ is a small number, see (4).
(4) Cardinal reading of "Few/Many As are B"

$$
\text { Few: }|A \cap B| \leq x_{\max } \quad \text { Many: }|A \cap B| \geq x_{\min }
$$

One concrete proposal of how $x_{\text {min }}$ and $x_{\text {max }}$ might be identified is presented by Fernando and Kamp (1996). The "cardinal surprise reading" of few and many in sentences like (1) is an intentional comparison between the actual number of movies that Chris saw last year and a probabilistic belief $P_{E}$ about the expected number of watched movies in some contextually provided comparison class. The for-phrase in (1) triggers a comparison class of US males. The prior expectation $P_{E}$ is highly context-dependent. In contrast, $\theta_{\text {few }}$ and $\theta_{\text {many }}$ are context-independent. They are fixed thresholds on the cumulative distribution of $P_{E}$. Truth conditions of the surprise reading of sentences like (1) are given in (5).

$$
\begin{array}{ll}
\text { a. } & \llbracket \text { Few As are } \mathrm{B} \rrbracket=1 \text { iff }|A \cap B| \leq x_{\text {max }} \text { where }  \tag{5}\\
& x_{\max }=\max \left\{n \in \mathbb{N} \mid P_{E}(|A \cap B| \leq n)<\theta_{\text {few }}\right\} \\
\text { b. } & \llbracket \text { Many As are } \mathrm{B} \rrbracket=1 \text { iff }|A \cap B| \geq x_{\text {min }} \text { where } \\
& x_{\text {min }}=\min \left\{n \in \mathbb{N} \mid P_{E}(|A \cap B| \leq n)>\theta_{\text {many }}\right\}
\end{array}
$$

From (5b), entities which have properties $A$ and $B$ can be described as "many" if their cardinality is at least $x_{\min } . x_{\text {min }}$ is the lowest number for which the cumulative density mass of prior expectation $P_{E}$ about the number of As with property B is higher than the threshold $\theta_{\text {many }}$. In other words, "Many As are B" is a true description of cardinalities which are surprisingly high with respect to the contextually given $P_{E}$ and the context-independent threshold $\theta_{\text {many }}$ on $P_{E}$.

To illustrate, consider the example in Figure 1 for the many-sentence in (1). Prior expectations $P_{E}$ could look like in Figure 1a: they would assign a probability to any natural number $n$, indicating how likely we think it is that Chris saw $n$


Figure 1: Illustration of a surprise-based semantics
movies last year. Figure 1 b shows the cumulative distribution of the distribution in Figure 1a. If $\theta_{\text {many }}$ was fixed to, say, 0.8 , then the semantics would identify $x_{\min }$ to be 23 . Accordingly, for this $P_{E}$, the many-sentence in (1) would be false for any $n<23$ and true for any $n \geq 23$. Schöller and Franke (2015) present evidence for the fixed threshold hypothesis by identifying fixed values for $\theta_{\text {few }}$ and $\theta_{\text {many }}$, which correctly predict the applicability of few and many in different contexts, given experimentally measured prior expectations.

## Surprisingly: Intensifier or Marker of Surprise?

Two views are prima facie plausible for the meaning contribution of the adverb surprisingly. On the one hand, surprisingly can be taken to intensify the meaning of few and many just like other intensifiers like incredibly or very do. As a result, surprisingly many might be associated with a threshold $\theta_{\text {surpr. many }}$ higher than $\theta_{\text {many. }}$. The contrasting view is to classify surprisingly as a marker of the surprise reading, which overtly marks that truth-conditions must draw on a threshold on a measure of surprise. This view is supported by the semantic literature which suggests that "being surprisingly tall comes to mean taller than expected" (Nouwen, 2011, 154).

Note that our hypotheses for surprisingly apply to sentences with a salient cardinal surprise reading and a restricted comparison class. To discriminate between the two views on surprisingly, we deduce two experimentally testable hypotheses. Another two auxiliary hypotheses are tested alongside to complement our understanding of modified few and many, see Table 1. In what follows, we spell out these general hypotheses in terms of their predictions about the threshold values $\theta_{\text {few }}$ and $\theta_{\text {many }}$ as assumed by Fernando and Kamp (1996) and test them with a computational model which infers these threshold values on the basis of experimental data.
Salient surprise reading. We cannot exclude that few and many may also denote a small or large cardinality, independent of prior expectations. To test the auxiliary assumption that the most salient readings of our experimental test sentences (see Appendix) are cardinal surprise readings given the comparison class for which we measure subjects' prior expectations (see below), we contrast sentences with bare few and many with sentences modified by the compared to phrase in (2) which makes the relevant expectations overt. It is necessary to test this because if few and many did not have the intended surprise reading, differences between few/many and surprisingly few/many could be due to different readings and possibly different threshold values associated with them. Alongside few and many's intrinsic surprise reading, we test another related assumption: the for- phrase used to mark the comparison class triggers the same prior expectations $P_{E}$ as the compared to phrase which openly addresses expectations, see (7).
Marker of surprise. If the function of surprisingly is to mark a cardinal surprise reading, thresholds are the same as for unmodified few/many, where these cardinal surprise readings are most salient anyway (see above). Furthermore, sentences with surprisingly should not be different from sentences with compared to, as in (2).
Intensifier. Modification by surprisingly raises the threshold of many and makes it applicable to a smaller range of cardinalities, resulting in a stronger statement than the alternative with bare many. Few's threshold decreases.
Bennett \& Goodman. The intensifier hypothesis is in line with work by Bennett and Goodman (2015) who explain the

|  | intensifier | marker of surprise | salient surprise reading |
| :--- | :---: | :---: | :---: |
| predictions | many $\leq$ surprisingly many | many $=$ surprisingly many | many $=$ compared to... many |
|  | few $\geq$ surprisingly few | few $=$ surprisingly few | few $=$ compared to... few |
|  | surprisingly $=$ incredibly | surprisingly $=$ compared to | few |
| results | few: $\times$ many: $\checkmark$ | few: $\checkmark \quad$ many: $\times$ | few: $\checkmark$ many: $\checkmark$ |

Table 1: Hypotheses and results
strength of an intensifying degree adverb as "pragmatic inference based on differing cost [(their length and frequency)] rather than differing semantics" (p. 1). However, they do not test surprisingly. From the adverbs tested by Bennett and Goodman (2015), incredibly comes closest to surprisingly, as they have the same number of syllables and the most similar frequency in an updated version of the corpus Bennett and Goodman (2015) used, the Google Web 1 T 5grams corpus $(4,987,059$ occurrences as compared to $4,373,670$ occurrences of surprisingly). Following Bennett and Goodman (2015), we hypothesize that the thresholds of surprisingly few/many are roughly the same as for incredibly few/many.

## Experiments

To test the hypotheses in Table 1, two experiments were conducted to gather acceptability ratings of sentences with (modified) few and many and to measure participants' prior expectations. Prior expectations will be input to the computational model, which is presented in the next section.

## Experiment 1: Prior elicitation

Design. To get an empirical estimate of participants' prior expectations, we used a binned histogram task. Participants saw descriptions of a context as in (6a) and a question as in (6b). Subjects were presented with 15 intervals, whose ranges were determined by a pre-test (in which we asked for the most likely, lowest and highest possible cardinality). Subjects rated the likelihood that the true value lies in each of the intervals, by adjusting a slider labeled from "extremely unlikely" to "extremely likely." For example, they would adjust a slider each for the probability that Chris saw 0-2, 3-5, ..., $39-41$ or more than 42 movies last year.

## (6) Prior elicitation example

a. Background: Chris is a man from the US.
b. How many movies do you think he saw last year?

Participants. 80 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

Materials \& Procedure. After reading instructions, each subject saw all of the 14 experimental items (see Appendix), one after another. For each item, the 15 intervals were presented horizontally on the screen and paired with a vertical slider. Participants had to adjust or at least click on each slider before being able to proceed.

Results. Two participants were excluded for not being native speakers of English. For each item, each participant's ratings were normalized and these normalized ratings were then averaged across participants. We understand these probability distributions $P_{E}$, see Figure 2, as approximations of the beliefs held in the population (Franke et al., 2016).

## Experiment 2: Judgment task

Design. In a binary judgment task we measured acceptance of sentences with few and many with and without modifiers (surprisingly, incredibly or compared to). Participants were presented with a context which introduced a situation and an interval as in (7a). The interval was randomly chosen from 8 of the 15 intervals from the prior elicitation task (see Appendix). We presented only four low intervals for few and four high intervals for many to avoid a large number of combinations. The context was described by a statement as in (7b) which contained either few or many. We elicited data of four groups of participants which each saw a different modifier.

## (7) Production study example

a. Context: Chris is a man from the US who saw [0-2 | 6-8 | $\ldots \mid 42$ or more] movies last year.
b. Statement: [For | Compared to what you would expect for] a man from the US, Chris saw [-| surprisingly | incredibly] [few | many] movies last year.

Materials \& Procedure. Each participant was randomly assigned to one modifier condition (unmodified, compared to construction, surprisingly, incredibly). After reading a short explanation of the task, each subject saw all of the 14 contexts from the Appendix one after another in random order. Sentences with unmodified few and many or incredibly or surprisingly were introduced by a for-phrase which made the intended comparison class overt. The fourth group saw a compared to phrase which additionally made expectations salient. For each context, a quantity word and one of its four associated intervals were assigned randomly. Participants had to click on one of two radio buttons labeled with TRUE or FALSE before being able to proceed to the next item.
Participants. We recruited 787 participants with US-IP addresses via Amazon's Mechanical Turk, among them 301 participants in the unmodified condition and 162 participants


Figure 2: Empirically measured prior expectations. Error bars are estimated 95\% confidence intervals.


Figure 3: Proportion of TRUE answers from Experiment 2.
each in the other three conditions. The unmodified condition had more participants because it was part of a previous experiment in which we presented 8 of 15 intervals for both few and many. For the analysis only data from those intervals presented in the other three conditions was used.
Results. Data was excluded of 25 participants who reported not to be native speakers of English or to not having understood the task. Figure 3 shows the proportion of TRUE answers.

For each of the quantity words few and many we specified a linear mixed effects regression model predicting the proportional acceptance of statements as in (7b). During a guided search through the model space, we started out with a model containing only the random effect ITEM and added fixed effects if this significantly increased the model's fit to the data
(measured by AIC).
For many, the final model includes the fixed effects INTERVAL and MODIFIER and their interaction. Significantly more participants accepted the statements for higher intervals $(\beta=0.02, S E=0.007, p<0.01)$. The modification of many by surprisingly leads to a lower acceptance ( $\beta=$ $-0.59, S E=0.12, p<0.001$ ) than of sentences with unmodified many. This suggests that surprisingly intensifies the meaning of many. The same is the case for sentences with incredibly, which were also rated lower than unmodified many ( $\beta=-0.53, S E=0.12, p<0.001$ ). There is no difference between sentences with a compared to phrase and unmodified many ( $\beta=-0.17, S E=0.12, p<0.15$ ), which suggests that many receives a surprise reading in both cases. Surprisingly and compared to are rated significantly different
( $\beta=-0.42, S E=0.12, p<0.001$ ), but there is no difference between surprisingly and incredibly. Furthermore, there is a significant interaction between INTERVAL and MODIFIER for surprisingly $(\beta=0.03, S E=0.01, p<0.001)$ and incredibly ( $\beta=0.02, S E=0.01, p<0.01$ ).

For few, the final model, obtained by the same procedure, includes the fixed effects interval and modifier. The proportion of participants accepting the statement is significantly lower for higher numbers $(\beta=-0.12, S E=0.004, p<$ 0.001 ). Among the modifiers only incredibly is significantly different from bare few $(\beta=-0.05, S E=0.02, p<0.05)$; for surprisingly and compared to this is not the case. No significant difference between surprisingly and compared to is found, but incredibly is rated significantly lower than surprisingly ( $\beta=-0.05, S E=0.02, p<0.05$ ).

These results are expected under the "salient surprise reading" hypothesis. While surprisingly seems to behave like an intensifier for many, for few it seems to redundantly mark surprise.

## Computational Model

The regression models reported above include a random effect for items but do not constrain these to reflect prior expectations. Moreover, regression models do not predict judgments as a function of thresholds on expectations. It is therefore insightful to complement regression modeling with an explicit theory-driven model of a possible datagenerating process. We use the computational model of Schöller and Franke (2015) for this purpose. The model takes empirically measured prior expectations as input and treats $\theta_{[i] f e w}$ and $\theta_{[i] m a n y}$ for each modifier condition $i$ (unmodified, surprisingly, incredibly, compared to) as latent parameters, whose values will be estimated to fit experimental data. The model specifies a likelihood function $P$ (Observation $\left.\theta_{[\mathrm{i}] \text { many }}, \theta_{[\mathrm{i}] \text { few }}\right)$ which assigns to values of latent parameters a probability of seeing a particular experimental observation. Bayesian inference is one way to infer plausible threshold values, given the likelihood function and a prior:
$P\left(\theta_{[\mathrm{i}] \text { many }}, \theta_{[\mathrm{i}] \text { few }} \mid \mathrm{O}\right) \propto P\left(\theta_{[\mathrm{i}] \text { many }}, \theta_{[\mathrm{i}] \mathrm{few}}\right) P\left(\mathrm{O} \mid \theta_{[\mathrm{i}] \text { many }}, \theta_{[\mathrm{i}] \mathrm{few}}\right)$
Our goal, then, is to see for each modifier which pairs of threshold values $\theta_{[i] m a n y}$ and $\theta_{[i] f e w}$ are likely given the data. We estimate the a posteriori credible threshold values and compare how similar they are across conditions. We focus on many in the exposition, but the case for few is parallel. Straightforwardly, (5) translates into a probabilistic rule $P$ ("[modifier $i]$ many" $\left.\mid n, P_{E} ; \theta_{[\mathrm{i}] \text { many }}\right)=\delta_{n \geq x_{\text {mini, }, ~}}$, where $x_{\text {min, }}$ is derived from $P_{E}$, as in (5), based on $\theta_{[i] m a n y}$. This is a degenerate probabilistic rule because it maps the applicability of "many" to 0 and 1 only. To allow for noise, we look at a parameterized, smoothed-out version.
$P\left("[i]\right.$ many" $\left.\mid n, P_{E} ; \theta_{[\mathrm{i}] \text { many }}, \sigma_{j}\right)=\sum_{k=0}^{n} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \mathcal{N}\left(y ; x_{\text {min, } \mathrm{i}}, \sigma_{j}\right) \mathrm{d} y$


Figure 4: Estimated $95 \%$ credible intervals for $\theta_{\text {few }, \mathrm{i}} \& \theta_{\text {many,i }}$

Here, $\sigma_{j}$ is another free model parameter that regulates the steepness of the curve, and $\mathcal{N}\left(y ; x_{\min , \mathrm{i}}, \sigma_{j}\right)$ is the probability density of $y$ under a normal distribution with mean $x_{\min , \mathrm{i}}$ and standard deviation $\sigma_{j}$. This rule predicts noisy acceptability ratings under a surprise-based semantics where the amount of noise is controlled by $\sigma_{j}$, see Figure 1c. Noise can be caused by uncertainty about the exact shape of $P_{E}$ and the amount of uncertainty differs across contexts. This is why we allow an individual value of $\sigma_{j}$ for each context $j$. Furthermore, we assume that the parameter values $\theta_{[i] m a n y}, \theta_{[i] f e w}$ and $\sigma_{j}$ are independent of each other and that they have uniform priors over an interval that is large enough to accommodate a range of plausible values without weighting them.

$$
\begin{aligned}
P\left(\theta_{[i] m a n y}, \theta_{[\mathrm{i}] \text { few }}, \sigma_{j}\right)= & \operatorname{Uniform}_{[0 ; 1]}\left(\theta_{[\mathrm{i}] \text { many }}\right) . \\
& \text { Uniform }_{[0 ; 1]}\left(\theta_{[i] \mathrm{few}}\right) \cdot \operatorname{Uniform}_{[0 ; 10]}\left(\sigma_{j}\right)
\end{aligned}
$$

To approximate the joint posterior distribution, we used MCMC sampling, as implemented in JAGS (Plummer, 2003). We collected 10,000 samples from 2 MCMC chains after a burn-in of 10,000 . This ensured convergence, as measured by $\hat{R}$ (Gelman \& Rubin, 1992). Figure 4 shows the estimated $95 \%$ credible intervals for the marginalized posteriors over thresholds per modifier. Where intervals (clearly) do not overlap, there is reason to believe that thresholds differ. For example, $\theta_{\text {surpr.many }} \in[0.863,0.903]$ tells us that surprisingly many describes cardinalities which are higher than at least $86 \%$ of the cumulative density mass of $P_{E}$. This threshold is higher than bare many's, $\theta_{\text {many }} \in[0.657,0.701]$. Taken together, the model predicts that surprisingly many is restricted to describe higher cardinalities than unmodified many.

## Discussion and Conclusions

Table 1 summarizes the results from regression and theorydriven modeling. The data supports the "salient surprise reading" hypothesis assumed by Fernando and Kamp (1996) and suggests that an expectation-based reading is the canonical interpretation of cardinal few and many in our test sentences. There is no difference between unmodified sentences and sentences in which expectations are made salient by a compared to phrase.

For surprisingly, the picture is less clear. Sentences with many provide support for the "intensifier" hypothesis. Speakers prefer it for higher cardinalities than those which render unmodified many or sentences with a compared to construction true. Furthermore, we do not find a difference with incredibly. When combined with few, however, surprisingly does not appear to be an intensifier. Sentences with few, surprisingly few and compared to are rated equally, speaking in favor of a "marker of surprise" hypothesis. For the comparison between surprisingly and incredibly, we get conflicting results from the regression and the theory-driven model. The regression analysis finds that incredibly few is rated lower than surprisingly few, but the computational model identifies an overlap in the estimated credible intervals. However, we want to once more stress that we are here comparing conclusions based on models which are decidedly different. Whereas the computational model is theory-driven and includes experimentally measured prior expectations, the regression model only looks at numerical differences in the ratings. Ultimately, we believe in the computational model.

Keeping in mind that few only applies to small cardinalities, the lack of a difference could also be due to a floor effect. This is where future research should tie in. Few should be presented in contexts like book or facebook, in which large cardinalities are plausible and few can operate away from 0 . Additionally, the presented intervals should be more finegrained. A follow-up study as well as further discussion of the semantic differences between few and many are presented in Schöller (2017).

## Acknowledgments

MF and AS's work was supported by the Priority Program XPrag.de (DFG Schwerpunktprogramm 1727). MF's work was supported by the Institutional Strategy of the University of Tübingen (DFG, ZUK 63).

## Experimental material

1. book - A friend's favorite book has been published only recently (and has [0-40, 81-120, 161-200, 241-280, 321-360, 401-$440,481-520,560$ or more] pages).
2. bus - Vehicle No. 102 is a school bus (which has seats for [0-4, $10-14,20-24,30-34,40-44,50-54,60-64,70$ or more] passengers).
3. calls - Lisa is a woman from the US (who made [0-4, 10-14, $20-24,30-34,40-44,50-54,60-64,70$ or more] phone calls last week).
4. class - Erin is a first grade student in primary school. (There are [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] children in Erin's class.)
5. coffee - Andy is man from the US (who drank [0-1, 4-5, 8-9, 12-13, 16-17, 20-21, 24-25, 28 or more] cups of coffee last week).
6. cook - Tony is a man from the US (who cooked himself [0-3, 8-11, 16-19, 24-27, 32-35, 40-43, 48-51, 56 or more] meals at home last month).
7. facebook - Judith is a woman from the US (who has [0-69, 140-209, 280-349, 420-489, 560-629, 700-769, 840-909, 980 or more] Facebook friends).
8. friends - Lelia is a woman from the US (who has [0-1, 4-5, 8-9, 12-13, 16-17, 20-21, 24-25, 28 or more] friends).
9. hair - Betty is a woman from the US (who washed her hair [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] times last month).
10. movie - Chris is a man from the US (who saw [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] movies last year).
11. poem - A friend wants to read you her favorite poem (which has $[0-3,8-11,16-19,24-27,32-35,40-43,48-51,56$ or more] lines).
12. restaurants - Sarah is a woman from the US (who went to [0-3, 8-11, 16-19, 24-27, 32-35, 40-43, 48-51, 56 or more] restaurants last year).
13. shoes - Melanie is a woman from the US (who owns [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] pairs of shoes). - intervals:
14. tshirts - Liam is a man from the US (who has [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] T-shirts).

## References

Bennett, E., \& Goodman, N. D. (2015). Extremely costly intensifiers are stronger than quite costly ones. In Proceedings of CogSci (pp. 226-231).
Clark, H. H. (1991). Words, the world, and their possibilities. In G. R. Lockhead \& J. R. Pomerantz (Eds.), The perception of structure: Essays in honor of Wendell R. Garner (pp. 263-277). American Psychological Association.
Fernando, T., \& Kamp, H. (1996). Expecting many. In T. Galloway \& J. Spence (Eds.), Linguistic society of america SALT (pp. 53-68). Ithaca, NY: Cornell University.
Franke, M., Dablander, F., Schöller, A., Bennett, E. D., Degen, J., Tessler, M. H., ... Goodman, N. D. (2016). What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data. In Proceedings of CogSci (pp. 2669-2674).
Gelman, A., \& Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical science, 457-472.
Nouwen, R. (2011). Degree modifiers and monotonicity. In Vagueness and language use (pp. 146-164). Springer.
Partee, B. (1989). Many quantifiers. In J. Powers \& K. de Jong (Eds.), $5^{\text {th }}$ eastern states conference on linguistics (escol) (pp. 383-402).
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, \& A. Zeileis (Eds.), Proceedings of the 3 rd international workshop on distributed statistical computing.
Sanford, A. J., Moxey, L. M., \& Paterson, K. (1994). Psychological studies of quantifiers. Journal of Semantics, 11(3), 153-170. doi: 10.1093/jos/11.3.153
Schöller, A. (2017). How many are many? Exploring contextdependence with probabilistic computational models. (Unpublished doctoral dissertation)
Schöller, A., \& Franke, M. (2015). Semantic values as latent parameters: Surprising few \& many. In S. D’Antonio, M. Moroney, \& C. R. Little (Eds.), Proceedings of SALT (Vol. 25, pp. 143-162). doi: 10.3765/salt.v25i0.3058

# Developing cognitive flexibility in solving arithmetic word problems 

Calliste Scheibling-Sève (calliste.scheibling.seve@gmail.com)<br>Department of Psychology, University of Paris 8 ,<br>2 Rue de la Liberté, 93526 Saint-Denis Cedex 02, Paris, France<br>Emmanuel Sander (Emmanuel.sander@univ-paris8.fr)<br>Department of Psychology, University of Paris 8,<br>2 Rue de la Liberté, 93526 Saint-Denis Cedex 02, Paris, France

Elena Pasquinelli (elena.pasquinelli@fondation-lamap.com)
Fondation La main à la pâte
43 rue de Rennes, 75006 Paris, France


#### Abstract

In problem solving situation, cognitive flexibility appears to be a major skill. Fostering cognitive flexibility is therefore a specific stake in mathematics education. This research introduces a learning method to develop mathematical concepts when solving word arithmetic problems. The study was conducted with 8 classes $\left(4^{\text {th }}-5^{\text {th }}\right.$ Grades) from highpriority education schools in the Paris area following this protocol: pre-tests, 5 learning sessions for experimental and control groups, post-tests. During learning sessions, students studied arithmetic word problems that can be solved in two different ways: an expansion strategy and a factorization one. The experimental teaching method, based on a recategorization principle, allowed experimental students to improve more than the control students in ability to use the factorization strategy even in contexts where it is the less intuitive and to consider the two successful strategies. Educational entailments of our finding are discussed.


Keywords: cognitive flexibility, evidence-based education, categorization, learning method, word arithmetical problem

## Introduction

In mathematics, proposing flexible and adaptive representations and strategies reflects higher problem solving skills (Heinze, Star \& Verschaffel 2009). We proposed to study not only strategies in algebra problems but also the related representations derived from word problem. Indeed when solving word problems, novices intuitively induce a superficial structure, triggering a misleading categorization of the situation (Chi, 2008). The present study aimed to improve pupil's cognitive flexibility in problems solving in order to develop an expert categorization on problems that reflects a better mastering of the underlying mathematical notions. Students are encouraged to reelaborate the notion's representation, which leads to recategorize it. Based on recategorization principle, this method is applied on problems involving the distributive property. The distributive property problems admit two solving strategies whose preferential use depends on the representation built by the solver.

## An induced representation

Phrasing of mathematical word problems can influence the induced representation of the problem by students (Vergnaud, 1982; Hudson, 1983; De Corte, Verschaffel \& De Win, 1985). But in addition to linguistic features, semantic effects also rely on semantic relations or scenario depicted in the problem: when solving a word problem, students build a mathematical representation based on semantic relations inferred from real-word objects (Bassok \& Olseth, 1995, Bassok, Chase \& Martin, 1998). For instance, a problem involving apples and baskets is likely to evoke the asymmetric "contain" relation. So students align this semantic relation with structurally analogous mathematical relations: apples and baskets support the semantic relation contain (content, container) and thus the mathematical relations of division (dividend, divisor). This spontaneous encoding of problem situations results from the properties and relations of the entities or objects depicted in a problem. Semantic alignment, namely alignment between the semantic and mathematical relations, influences the difficulty of mathematical problems.
A way to study this effect of semantic content on the spontaneous encoding is to use problems solvable by two strategies. Indeed the semantic context can influence the encoding of the problem and thus lead to a preference for one of the two strategies. Several studies showed that the variable involved in the problem impacts the problem's representation built by the solver (Bassok et al., 1995; Vicente, Orrantia \&Verschaffel, 2007; Gamo \& Sander, 2010). In Gamo et al.'s study, 4th and 5th grade students had to solve isomorphic problems involving one of the three following variables - the number-of- elements, price, and age -. This type of problem ("Antoine took painting courses at the art school for 8 years and stopped when he was 17 years old. Jean began at the same age as Antoine and took the course for 2 years less. At what age did Jean stop?") can be solved by two strategies: a "complementation" strategy (in three steps: $17-8=9 ; 8-2=6 ; 9+6=15$ ) and a 'matching'" strategy (in one step: $17-2=15$ ). But the variable involved in the problem fosters one of the two representations of the problem: (a) a part-whole schema that
underlies unordered units which triggers the computation of the difference between the part and the whole given in the first half of the problem; and (b) a comparison schema that underlies ordered units. Number-of-elements problems are spontaneously encoded according to the part-whole schema and lead to the complementation strategies whereas the age problems foster the matching strategy but not exclusively.

## Cognitive flexibility in problem solving

Even after instruction, non-relevant representations remain: experts do not systematically use the most efficient strategy when solving arithmetic problems (Star \& Newton, 2009) even though they master it. Therefore, arithmetic problem solving raises the question of the influence of problem representations on the possibility to choose flexibly the most efficient strategy. Cognitive flexibility seems to be critical while solving problems (Clement, 2006). Indeed, it refers to the ability to select adaptively among multiple representations of an object, perspectives or strategies in order to adjust to the demands of a situation (Cragg \& Chevalier, 2012; Diamond, 2013). Through the problem of 'water-jug measuring problems' (Luchins, 1942), Clement (2006) proposed the concept of representational flexibility in problem solving: following an impasse situation, individuals recode the situational properties and adopt a new representation that leads to transfer the right strategy. Hence, cognitive flexibility is related both to abstraction and transfer. Cognitive flexibility can therefore be measured through the mastering of multiple strategies and of their appropriate use (Rittle-Johnson \& Star, 2007; Star \& Seifert, 2006). Students with high flexibility in problem solving are more likely to adapt existing strategies when faced with unfamiliar transfer problems and to better understand domain concepts (Hiebert \& Wearne, 1996; Rittle-Johnson \& Star, 2007).

## Recategorization in problem solving

Studies on cognitive flexibility in mathematics focused either on interpretation of the situation or on strategies (Heinze et al., 2009). Being able to adopt a multiplicity of categorization makes it possible to change point of view according to the needs of the situation. For example, a physicist who sees a glass falling down does not need to categorize a glass as a body under the law of gravitation and on which forces are exerted. Categorizing a glass only as "an object made of a fragile material" is sufficient to act in the appropriate way, namely to catch up the glass. Thus, the more an individual diversifies his repertoire of categorization, the more he is able to adopt different perspectives. By articulating different points of view on the same situation, the individual can embrace its complexity (Hofstadter \& Sander, 2013).

In the present study, we proposed to focus on recategorization as a mechanism to recode a representation and transfer the adapted strategy to a new context. Evidences from social psychology showed that if an individual seems to be inconsistent with his/her category
membership, perceivers would integrate other information and recategorize the individual in the newly applied category (Gawronski \& Creighton, 2013). When it comes to solving problems, a same situation or entity can be categorized at different levels of abstraction in multiple ways. The categorization adopted has been identified as an indicator of expertise. For example, in physics, novices categorize problems according to the objects used (problems of pulley or inclined planes), while experts categorize problems according to the physical principle (e.g. Newton's third law) (Chi, Feltovich, and Glaser, 1981, Chi et al., 1989). Unlike the experts, the novices therefore construct their categories mainly on the basis of superficial information, such as specific objects (Schoenfeld \& Herrmann, 1982). Experts rely on a greater number of categories and levels of categorization than novices to represent a situation. Since novices use some superficial cues, they can make negative transfers, using an irrelevant strategy by analogy with a problem that share the same superficial traits (vocabulary, object, theme) (Chen, 2002). Teaching to recategorize in a relevant manner seems to be a lever to develop students' ability to transfer strategies.

Thus, training cognitive flexibility is a challenge for developing learning method. Teaching experiments in mathematics mainly studied number calculations: multidigit addition and subtraction (Carpenter et al., 1997), decomposition (Klein, Beishuizen, \& Treffers, 1998) and linear equations (Rittle-Johnson \& Star, 2007). Evidences have therefore been obtained for the algorithmic aspects but are sparser when it comes to word problems. One method consists in comparing two strategies for the same problem (Brissiaud, 1994). Gamo et al. (2010) proposed a training method in order to develop mathematical concepts through the semantic recoding of the word problem. The principle is to recode the semantically induced structure into a more apt mathematical structure. By recoding the problem, the students adopts a new point of view, which leads them to develop a representation of the problem that corresponds to the mathematical structure and succeed to use more expert solving strategies. In Gamo et al.'s study, students in Grade 4 and 5 had to solve problems sharing the same deep structure but being spontaneously categorized as problems solvable by complementation strategy and not by the most efficient one (the matching strategy). During the training session, students compared the problems to stress the common structure. At post-test, students improved their use of matching strategy. Whereas interpretation initially realized at a level of abstraction based on the semantic structure, students acquired an additional degree of abstraction based on the mathematical structure, after semantic recoding. Comparison fosters a more abstract representation of the problems.

## The present study: a training method based on recategorization

Because of the lack of understanding of abstract ideas, when they are not operationalized (Willingham, 2009), and
the difficulties of transferring solving strategies (Ross, 1984), the learning method to develop students' cognitive flexibility is applied in a specific school context: word arithmetic problems on the distributive property. This type of problems -listed in the French curricula in $4^{\text {th }}$ and $5^{\text {th }}$ grade- has the methodological interest to be solvable by two strategies. Moreover, the nature of the variable involved in the problems favor one of the strategies (expansion or factorization) (Sander, 2008; Moreau \& Coquin-Viennot, 2003).

The main goal of this method is to allow students to overcome the spontaneous encoding of problem situations. In order to develop learning methods favoring abstraction, while taking into account the difficulties of transfer, the training of recategorization was conducted through a semantic analysis and was supported by an explicit method built with students. This method consisted in allowing students to switch between two conceivable points of view on the same situation. Then they were prompted to choose their own one. This choice of point of view by students is related to a reflexive level and is consistent with previous work (Siegler, 1999; Blöte, 2001) that stressed the flexible use of strategies and encouraged students to think about the value of different procedures for solving a given problem. The different steps of the experimental training are detailed in Method, Training sessions.

## Hypothesis

We therefore hypothesized that the experimental training method should favor students' cognitive flexibility on a mathematical concept- the distributive property- involved in arithmetic word problems.

Students should be able to adopt the two points of view on the problem and use the two strategies (expansion and factorization) without depending on the semantic context. For the training problems, no significant difference in factorization and dual strategies use between the two groups should be observed, since they are both trained to solve this kind of problems (Hypothesis 1). Yet, the experimental method based on recategorization should favor far-transfer compared to the traditional method. Thus the experimental group should propose significantly more factorization and dual strategies than the control group at the post-test and higher progression should be observed for the experimental group for the non-trained problems (Hypothesis 2).

## Method

## Participants

The experiment was conducted with eight classes from four elementary schools located in high-priority education network in Paris region during regular classwork school hours. 142 students took part in the study: 74 were fifth graders and 68 were fourth graders (mean age $=10$ years and 3 months, $\mathrm{SD}=6$ months, 78 boys, 64 girls).

The experimental group included 66 students ( 375 th graders from 2 classes and $294^{\text {th }}$ graders from 2 classes. The control group included 76 students ( $375^{\text {th }}$ graders from

2 classes and 39 were 4th graders from 2 classes).

## Design

The experiment included three phases: pretest, training sessions, post-tests. The pre- and posttest were strictly identical. Both the experimental group (EG) and the control group (CG) followed training sessions taught by the experimenter. Within each group, training sessions were identical in their duration, organization and problems statements.

## Material

## Pre and post-tests

The material was composed by 8 isomorphic distributive word problems (Table 1) and 5 filler problems. Indeed, filler problems were proposed between distributive problems, in order to make the structural similarities between the distributive problems less salient for the students.

Each distributive problem describes a situation involving one factor and three summands. The final question whose structure is "How much/many ... in all?" was placed at the end of the text. Two main solving strategies make it possible to reach the solution: Expansion strategy (sum of each part multiplied by the factor: $4 \times 6+$ $4 \times 7+4 \times 8$ ) and Factorization strategy (sum of the parts, then multiplied by the factor: $(6+7+8) \mathrm{x} 4)$. We selected four different variables (Numbers, Duration, Price, Weight). For each variable, a statement whose summands are said specific categories and a problem whose summands are called general categories were proposed (Table 1). The summands of the specific category problem are grouped at a base level, while those of the general category problem at a more abstract level than the base level (in the context of a treat cone, 3 balloons, 8 cookies, 4 figurines or 7 lollipops, 8 8 candies, 3 chocolates). So summing them as a whole is easier for specific than general category, and could influence the strategy.

The distributive problems were constructed by controlling the familiarity of the vocabulary and the numerical values at stake. The numerical values had different features in common in order to limit numbers' effects. Factor value was between 4 and 8 ( 5 was excluded since the associated multiplication table is easier). The three summands were between 2 and 8 ( 5 is excluded). Their sum lied between 11 and 21. And the result was inferior to 100 (between 72 and 98 ) in order to control the level of calculation difficulty.
A pedagogical advisor of the French National Education was involved in the conception of all sessions, in order to assure ecological material and ecological pedagogical acts. Therefore, 8 booklets were constructed by controlling the numerical values, the order of presentation of the problems variable and problems versions. On each page of the booklet, the problem was presented in written form with two sections in order to propose two strategies. The instruction for solving the problem with two strategies was both orally given by the experimenter and written on each page:

| Variables | Duration | Numbers | Weight | Price |
| :---: | :---: | :---: | :---: | :---: |
| Context | X has made a list of purchase for y years. Each year, X's purchases are: | X wants to prepare a treat cone per child. There are y children. Making a treat cone requires the following items: | X wants to fill his/Her pencil case with items that weight y grams each. <br> In the pencil case, there are: | $X$ is at the checkout of a supermarket. He bought some items. For each item, he took y in his basket: |
| Specific Categories | Printers <br> Computers <br> Scanners | Lollipops <br> Candies <br> Chocolates | Pens <br> Gums <br> Markers | Buns Cakes Pies |
| General Categories | Microscopes Desks Hamsters | Balloons <br> Cookies <br> Figurines | Shells, Key-rings Candies | Ice Plants Plates |
|  | In all, how many purchases has X bought? | In all, how many objects does X need? | In all, how much does these items weight? | In all, how much did X spend? |

Table 1: The 8 problems at pre and post-tests

Write all your calculations and the result in the following section:

Do you see any other method to come up with the same result? If yes, write it down while writing all calculations you performed to find the result.

Tests lasted 45 minutes. Students were given 3 minutes to solve each problem. Students were instructed that they could ask the experimenter or the teacher to read aloud the problem in order to bypass reading difficulties, and that they had to write down all calculations. When the time was over, students had to turn the page and begin the new problem when the experimenter gave the instruction. They could not modify their answer once they turned the page.

## Training sessions

Training sessions took place in 5 sessions over 5 weeks ( 45 minute session each week) for each class (Table 2).

Usual French textbooks inspired the pedagogical method used by the control group for Grades 4 and 5 (Vive Les Maths and Companion Math). The experimental method was built for this study. The two approaches did only differ in their treatment of the problem. In the control group, students learnt to select relevant information in statements and choose operations. And the experimental group looked for the semantic relations (the sum of the parts forms a whole) and chose the point of view it wished to adopt. For instance, one of the training problem was the following: « $A$ team of 5 athletes participated in a relay: each athlete ran on a loop of 8 km , then on a straight line of 2 km and finally on a loop of 6 km . How many kilometers has the team traveled? »
To find the number of kilometers, two strategies are possible. Experimental group learnt to choose between:

- adopting the point of view of each part of the relay (loop and line): each loop/line is a separate part and we realize an expansion strategy: $5 \times 8+5 \times 2+5 \times 6=80 \mathrm{~km}$
- adopting the more abstract point of view of the relay: the different parts (loops and line) form the relay, we carry out a factorization strategy: $(8+2+6) \times 5=80 \mathrm{~km}$

Thus whereas categorizing each addend as a part leads to an expansion strategy, categorizing them as a whole leads to a factorization strategy.

Regarding all other aspects, session organization was similar between the two groups: students began by exercising on the slate, in order to engage them in the task.

Then the students had to answer on an exercise sheet, whose support was also projected on the blackboard. Finally, students ended the session by answering the question "What did I learn today? " and then a general conclusion was proposed by the experimenter and was written by the students. The distributive problems studied in the sessions 3, 4 and 5 were identical between the two groups and involved only two types of variables: Number-of-Elements and Distance.

## Coding and scoring

For pre and post-test, problems were analyzed under two criteria:

- the use of factorization strategy (correct reasoning and calculation) to solve the problem as a first or second strategy
- the use of double strategy (reasoning and correct calculation)
Then a global improvement score was calculated. At pre and post-test, each problem was coded as 1 when it was solved by an expansion, by 2 when it was solved by a factorization, by 4 when it was solved by dual-strategies and 0 if otherwise. Then the difference between post and pre-test was computed. Each student got therefore an improvement score, reflecting his/her progress between the pre and posttest.


## Results

Hypothesis 1 stated that the frequency of factorization and dual-strategies by students was not expected to be different between the two groups for the training problem (Number-of-elements) due to the effect of training in both groups. The improvement score was 0.34 for the control group and 0.37 for the experimental group ( $\mathrm{p}>0.5$ ) for factorization strategy and 0.30 for dual-strategy for each group $(\mathrm{p}=1)$.

|  | Experimental Group | Control Group |
| :---: | :---: | :---: |
| Session 1 | A problem: way of seing a situation, parts and whole | A problem: a question, useful data, operations |
| Session 2 | Multiplication and commutativity by semantic recoding | Multiplication and commutativity by repeated addition |
| Session 3 | Distributivity: semantic relations and choice in point of view | Distributivity: useful data and choice in operations |
| Session 4 | Dual strategies <br> by change in point of view | Dual strategies <br> by equivalency of procedures |
| Session 5 | Distributivity problems: choice in points of view | Distributivity problems: useful data, operations |

Table 2: The training sessions

|  | PreTest |  | PostTest |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Factorization | Dual Strategy | Factorization | Dual Strategy |
| CG | $0.14(0.24)$ | $0.05(0.14)$ | $0.47(0.40)$ | $0.36(0.40)$ |
| EG | $0.18(0.27)$ | $0.07(0.18)$ | $0.63(0.38)$ | $0.47(0.43)$ |

Table 3: Means (and SD) of factorization and dual strategy frequency by students at pre and post-test in use of in the experimental group and control group

## Discussion

This absence of difference between the two groups for the trained problems reflects the similarity in terms of learning method between the two groups: they both learned to use dual-strategies (expansion and factorization) for the distributive problems. The control group focused on relevant information in statements, digits and operations whereas the experimental group focused on semantic relations, words and point of view. At the pretest, repeated measure ANOVAs with group as the between factor and problem variable as the within-subjects factor showed that there was no difference regarding the use of factorization ( $\mathrm{F}<1, \mathrm{~ns}$ ) or dual strategies, ( $\mathrm{F}<1$, ns). Both groups showed the same pattern of choice of strategies for each type of problem.

Hypothesis 2 stated that a better transfer should appear in the experimental group for non-training problems. At the posttest, reapeated measure ANOVAs with group as the between subjects factor and problem variable as the withinsubjects factor were performed: the experimental group was significantly superior to the group control for the use of factorization $(\mathrm{M}=0.63$ vs $\mathrm{M}=0.47 \mathrm{~F}(1,140)=6.15, \mathrm{p}=0.01)$ and we observed a marginal trend for the use of dual strategies $(\mathrm{M}=0.47$ vs $\mathrm{M}=0.36, \mathrm{~F}(1,140)=2.72, \mathrm{p}=0.1)$ (Figure 1 and Table 3).

Then we analyzed the improvement score. The global improvement score raised 1.46 for the experimental group and the control group's one raised 1.10. A repeated measure ANOVA was performed. The difference in improvement between the groups got a statistically significant trend $(\mathrm{F}(1,140)=3.5, \quad \mathrm{p}=0.06)$. For problems from general category, we observed an improvement score by 1.59 for the experimental group compared to 1.10 for the control group. A repeated measure ANOVA was performed (Table 4). Therefore, the improvement score for the more abstract problems (general category) was significantly higher for the experimental group $(\mathrm{F}(1,140)=6.12, \mathrm{p}=0.014)$.

Figure 1: Mean in factorization by students at Post-Test


The learning method based on both the resolution strategy comparison and the explicit analysis of semantic relations during classroom activities showed its success in promoting transfer. The experimental group was more successful in transferring the factorization strategy and dual strategies to non-trained problems in the post-test. The Price, Duration and Weight variables in post-test problems had not been trained during the learning sessions. The progression in terms of factorizing strategy and dual strategy suggests that the experimental group became less dependent on the choice of the variable than the control group. That means that the experimental group shows a greater ease in independence to context. They were successful switching from the spontaneous representation influenced by the variables of the problem to flexible representation based on the mathematical structure. Since progress for trained problems are similar between groups, the added value of the recategorization method lies in the success of far transfer.

The use of isomorphic problems made it possible to identify more precisely the robustness of the transfer effects from the learning method. Indeed as we studied the nontrained problems, the greater progression for the experimental group shows that the training was not superficial. This transfer reflects a semantic change by students that could adopt a double point of view on the problem.

In addition, our findings support the work of Vicente et al. (2007) who pointed out that the difficulty for students lies in developing the conceptual relations between the entities of the problem. Thus, in their study, the success rate of problems whose rewording shed light on "part-whole" relationships was higher than problems with additional information about the problem's situation. The properties and relations of the entities or objects depicted in a problem are therefore key in the choice of strategies. In our study, we did not use a conceptual rewording that underlines the underlying semantic relations but the experimental method consisted in orienting students to establish these relations because their categorizations of the elements of the problems were based on them.

## Conclusion

The students from the training group became less dependent from semantic context. Their choice of strategy was less constrained by the nature of the variable. The substantial transfer of the non-preferred strategy (factorization) illustrates the ability to adopt a new point of view on the situation. Thus students were able to change their encoding based on spontaneous representations to an encoding based on conceptual relations. To adopt this flexible and multiple
points of view on a problem, the training method based on recategorization seems to be promising. In addition to improve semantic analysis, students were encouraged to adopt a reflexive attitude thanks to the notion of point of view. Thus students developed their cognitive flexibility: developing flexible strategies with the ability to transfer them to new problems. Yet studying the extent of this transfer could be the goal of further research. The teaching method appears to be a useful framework to identify if cognitive flexibility is domain-general or domain-specific. Indeed, fostering cognitive flexibility takes part of a broader goal, namely promoting conceptual development.

## Acknowledgments

This research was supported in part by a National Research Agency Grant ANR-06-APPR-015. We thank the pedagogical advisor Arbya Eichi.

## References

Bassok, M., \& Olseth, K. L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(6)
Bassok, M., Chase, V. M., \& Martin S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. Cognitive psychology, vol. 35, no2, 99-134.
Blöte, A., Van der Burg, E., \& Klein, A. S. (2001). Students' flexibility in solving two-digit addition and subtraction problems. Journal of Educational Psychology, 93(3), 627.
Brissiaud, R. (1994). Teaching and development: solving 'missing addend' problems using substraction. European Journal of Psychology of Education, 9, 343-365
Chen, Z. (2002). Analogical Problem Solving: A Hierarchical Analysis of Procedural Similarity. Journal of Experimental Psychology, 28, 81-98
Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., \& Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145-182.
Chi, M. T., Feltovich, P. J., \& Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive science, 5(2), 121-152.
Chi, M. T. (2008). Three types of conceptual change: International handbook of research on conceptual change, 61-82.
Clément, E (2006). Approche de la flexibilité cognitive en résolution de problème. L'Année Psychologique, 106
Cragg, L., \& Chevalier, N. (2012). The processes underlying flexibility in childhood. The Quarterly Journal of Experimental Psychology, 65(2), 209-232.
DeCorte, E. Verschaffel, L., \& De Win, L. (1985) Influence of rewording verbal problems on children's problem representations and solutions. Journal of Educational Psychology, 77. 460-470
Diamond, A. (2013). Executive functions. Annual review of psychology, 64, 135-168.
Gamo, S., Sander, E., \& Richard, J-F. (2010). Transfer of strategies by semantic recoding in arithmetic problem
solving. Learning and Instruction, 20, 400-410.
Gick, M. L., \& Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychology, 12, 306-355.
Heinze, A., Star, J. R., \& Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education.
Hiebert, J., \& Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. Cognition and instruction, 14(3), 251-283.
Hofstadter, D., \& Sander, E.(2013). Surfaces and EssencesNew York, Basic Books.
Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. Child Development, 54(1), 84-90.
Klein, A. S., Beishuizen, M., \& Treffers, A. (1998). The empty number line in Dutch second grades: Realistic versus gradual program design. Journal for Research in Mathematics Education, 443-464.
Lautrey, J., Rémi-Giraud, S., Sander, E., \& Tiberghien, A. (2008). Les connaissances naïves. Armand Colin.

Lusk, Cynthia M., and Charles M. Judd. "Political expertise and the structural mediators of candidate evaluations." Journal of Experimental Social Psychology 24.2 (1988)
Richland, L. E., \& Simms, N. (2015). Analogy, higher order thinking, and education. Wiley Interdisciplinary Reviews: Cognitive Science, 6(2), 177-192.
Rittle-Johnson, B., \& Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. Journal of Educational Psychology, 99(3), 561.
Schoenfeld, A. H., \& Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. Journal of Experimental Psychology. Learning, Memory and Cognition, 8, 484.
Siegler, R. S. (1999). Strategic development. Trends in Cognitive Sciences, 3(11), 430-435.
Star, J. R., \& Newton, K. J. (2009). The nature and development of experts' strategy flexibility for solving equations. $Z D M, 41(5), 557-567$.
Star, J. R., \& Seifert, C. (2006). The development of flexibility in equation solving. Contemporary Educational Psychology, 31(3), 280-300.
Star, J. R., \& Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. Learning and Instruction, 18(6), 565-579.
Vergnaud, G. (1982). A classification of cognitive tasks and operations of 1195 thought involved in addition and subtraction problems. In T. P. Carpenter, 1196 J. M. Moser, \& T. A. Romberg (Eds.), Addition and subtraction: A 1197 cognitive perspective (pp. 39e59). Hillsdale, NJ: Erlbaum.
Vicente, S., Orrantia, J., and Verschaffel. L., Influence of situational and conceptual rewording on word problem solving. British journal of educational psychology 77.4 (2007): 829-848.

Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? American Educator, 8-19.

# Aging of the Exploring Mind: Older Adults Deviate more from Optimality in Complex Choice Environments 

Job J. Schepens<br>(job.schepens@fu-berlin.de)<br>Center for Cognitive Neuroscience Berlin<br>Freie Universität, Berlin 14195, Germany<br>Ralph Hertwig, Wouter van den Bos<br>(\{hertwig, vandenbos\}@mpib-berlin.mpg.de)<br>Center for Adaptive Rationality<br>Max Planck Institute for Human Development, Berlin 14195, Germany


#### Abstract

Older adults (OA) need to make many important and difficult decisions. Often, there are too many options available to explore exhaustively, creating the ubiquitous tradeoff between exploration and exploitation. How do OA make these complex tradeoffs? We investigated age-related shifts in solving exploration-exploitation tradeoffs depending on the complexity of the choice environment. Participants played four and eight option bandit problems with numbers of gambles and average rewards available on the screen. OA reliably performed worse in a more complex choice environment and were also more deviant from an optimality model (Thompson sampling), which keeps track of uncertainty beyond just the mean or last reward. OA seem to process important information in more complex choice environments sub-optimally, suggesting limited representations of future rewards. This interpretation fits to multiple contexts in the complex cognitive aging literature, in particular to the context of challenges in the maintenance of goal-directed learning.


## Introduction

In today's aging societies, more and more older adults (OA) are making cognitively demanding decisions about work, finances, their health, etc. Many such decisions benefit from thinking about future goals because the options available create explore-exploit tradeoffs. How do OA usually respond to these cognitive challenges in increasingly complex choice environments?

Decision makers generally have access to a number of learning mechanisms, habitual experience-based learning, and goal-directed learning. Goal-directed learning depends on some internal model, so that learning can be adapted flexibly, for example like when managing a research project. Habitual learning has been related to a dorsolateral striatal to sensorimotor cortex control loop while goal-directed learning has been related to a dorsomedial striatal to ventromedial and lateral prefrontal cortex control loop (Daw \& O'Doherty, 2014). Importantly, goal-directed learning is impaired in OA and this impairment has been associated to lower activation in prefrontal cortex areas (Eppinger \& Bruckner, 2015; Eppinger, Walter, Heekeren, \& Li, 2013). OA rely relatively more often on experience-based learning, which may arise from white matter integrity changes in the ventromedial and lateral prefrontal cortex (Chowdhury et al., 2013; Eppinger et al., 2013; Samanez-Larkin \& Knutson, 2015).

It is unclear how such changes in learning mechanisms in OA depend on the relative complexity of a task. Such a dependency would be likely, however, from the perspective of ecological rationality (Mata et al., 2012), which focusses on adaptation effects between the mind and the environment. For example, when OA need to explore among many options in order to choose between them later, OA rely on more minimal exploration strategies than YA (Frey, Mata, \& Hertwig, 2015). Here, we study such age-related performance changes in explore-exploit tradeoffs with varying cognitive demands. Analyzing effects of the complexity of choice environments this way could help to better understand the effects of task demands on age-related changes in learning mechanisms.

We used typical N -armed bandit problems to study changes in learning mechanisms across choice environments. Participants made inferences about risky options by sampling information from a four and eight option choice environment. Rewards were consequential, ensuring that participants needed to trade-off exploration and exploitation. Participants had to find options that give them the most money while having to minimize sampling from low reward options. N -armed bandit problems are well studied and afford detailed analysis of information processing in terms of continuation and switching behavior. Theoretically, expectations of future reward should rise with adequate, but not excessive, exploration. Such "smart" exploration requires one to have a good representation of the task and its structure, which typically weighs already observed rewards by the degree to which an option has been explored. This would thus involve an internal model of the task contingency between average payoff and average payoff uncertainty (see also Worthy, Cooper, Byrne, Gorlick, \& Maddox, 2014). Good performance in the task thus depends on learning mechanisms that use adequate future reward representations while performance anomalies will involve inadequate future reward representations.

We hypothesized that OA achieve lower performance and arrive slower at the higher reward options, depending on the number of options. If OA focus more on reward in current states, their rewards in future states should suffer. Such shortterm planning would more closely resemble experience-based learning rather than goal-directed learning. OA not arriving


Figure 1: Example eight-option choice profiles. Green indicates rewards and red indicates no rewards on a trial.
at the higher reward options at all would show a lack of an explore-exploit trade-off. We assumed that problems with a larger number of options are relatively more cognitively demanding because of a larger search space, which increases the amount of necessary information processing and representation.

Next, we describe methods and results from six kinds of data analyses: choice proportions statistics, choice proportions over trials, regret over trials, comparisons to an optimality model, comparisons to a fitted optimality model, and one-step ahead predictions of a fitted optimality model. We end with a discussion.

## Methods

Participants 32 older adults ( $\mathrm{OA}, M_{\text {age }}=70.5,65-74,38 \%$ female) and 29 younger adults (YA, $M_{\text {age }}=24.3,19-30$, 45\% female) participated in this study. All participants were healthy, right-handed, native German speakers with normal or corrected to normal vision, and without a history of psychiatric or neurological disorders. There were no group differences in gender proportion, educational level, and socioeconomic status. Compensation amounted to about 10 Euro per hour, plus on average 2 Euro performance-dependent bonus. Participants were recruited using advertisements.

Task The task of the participants was to maximize the sum of rewards in a total of 16 alternating four and eight-armed bandit problems. Rewards could be earned by selecting pictures of casino-style gambling machines presented on a computer screen using a keyboard or mouse. The gambling machines provided random rewards ( 1 or 0 ) with a hidden probability that was specific to each machine. The rewards were displayed on the respective bandit after each play. Participants had 100 trials for every problem to explore the hidden reward probabilities and to exploit those machines that give rewards most often. Remaining trials were displayed on the screen. Also, every bandit showed the number of plays so
far and the probability of a reward based on the observed rewards so far. This information is sufficient to make an optimal choice at any point in time, reducing the role of working memory. Of course, participants still need to figure out how they want to trade off exploration and exploitation. $89 \%$ of YA and $70 \%$ of OA ( $\mathrm{p}=.14$, test of equal proportions) indicated in a post-task questionnaire that "the extra information regarding the options" was helpful.

Procedure Participants were instructed 1) to maximize the sum of rewards, 2) how the task looked and worked, 3) that each trial is independently generated, and 4) that the best gambling machine in every individual problem had $p_{\text {opt }}=.6$ (to help comparability across problems). All participants had taken part in an unrelated fMRI study on risk-taking preference several weeks beforehand. Ethics approval was granted by the Institutional Review Board of the Max Planck Institute for Human Development.

Design The experiment made use of a repeated withinsubject condition (four vs eight options), and a betweensubject condition (age group). We chose the other hidden probabilities in steps of .075 below .6. Reliably finding the better options thus required a significant part of exploration out of the 100 available trials. See also Figure 1 for example choice profiles and the unique hidden probabilities. All participants saw the same randomly generated rewards for all 16 problems. This allowed comparison of the problem difficulty across participant groups as well as a reduction of an unnecessary source for variance in performance while keeping the probabilistic character of the task intact. Four different problem orders were generated and counterbalanced across participants. Two different orders started with four options and two different orders started with eight options. Between problems, performance was displayed on the screen and a keypress was required to continue with the next problem. Participants in both groups took about half an hour to


Figure 2: Predictions from a mixed effects model with the hidden reward probabilities of participant's choices as dependent variable and interaction effects between age, number of options, trial, and gender.
finish the experiment. The minimum response time was set to 200 milliseconds.

## Results

We first investigated age-related differences in task performance. Proportions of choices for each option revealed that OA chose the option with the highest hidden probability about $5 \%$ less often than YA did in both four and eight option conditions (four option 95\% HDI: .003-.093; eight option 95\% HDI: . 018 - .096; Bayesian ANOVA with logit function and broad prior), see Figure 4 for the differences for all options. We also tested a linear mixed effects model with the hidden probability of every chosen bandit as dependent variable and with participant ID, problem ID, and bandit position ID as random effects. We used Satterthwaite's approximations of p-values ( ${ }^{* * *}$ indicating $\mathrm{p}<.001$ ). We found negative interaction effects for OA in eight options $\left(\mathrm{B}=-.021^{* * *}\right)$ and for OA in eight options over trials ( $\mathrm{B}=-.011^{* * *}$ ). Together, these indicated a lower performance for OA in eight options, as well as an increasingly lower performance over trials. We also found a positive interaction effect for both YA and OA in eight options over trials ( $\mathrm{B}=.021^{* * *}$ ), as participants could improve relatively more over time for eight options. There still was a main negative effect of eight options ( $\mathrm{B}=-.092^{* * *}$ ) and a main positive effect of trial number ( B $=-.009 * * *)$. No significant difference or decrease over trials for OA remained, so the age effect is captured only by the higher-level interactions. Furthermore, we also controlled for gender effects, which indicated that male OA performed better over trials $\left(\mathrm{B}=.010^{* * *}\right)$ and that male OA performed better with 8 options $\left(\mathrm{B}=.024^{* * *}\right)$, and that males generally performed better $\left(\mathrm{B}=.003^{* *}\right)$. For visualizing these high level interactions, we generated predictions from this model using the package merTools, see Figure 2 Note that the visualization does not show raw data and that the differences in intercepts and slopes for the lines displayed should be inter-


Figure 3: Predictions from a mixed effects model with switching as dependent variable and three-way interaction effects between age, number of options, trial, reward, and the hidden probabilities of participant's choices.
preted in the light of all data included in the model. Besides performance, we used a similar statistical analysis to test age differences in switching probability over time. The resulting logistic regression model included age, number of options, trial number, the hidden probability, the reward for the participant's choices, and all three-way interactions. Together, the estimated effects on switching indicated that OA switch less often (**), OA switch away less often after sampling from an option with a relativey low hidden probability ( ${ }^{* * *) \text {, espe- }}$ cially in eight options over trials $\left({ }^{* *}\right)$. Beta's were not easily comparable for this model. We again generated predictions from this model to visualize these switching patterns, see Figure 3

Second, we examined development of age-differences in choice proportions over trials, see Figures 6 a and 6 b. For every trial, the solid lines represent the average number of times that participants chose an option. The local instabilities in the trajectories may result from individual differences and variation across the several problems. On average, the third best option stops overlapping with the second best option after about 25 trials for YA. For OA, the same separation exists between the second and third option after twice as many trials, see the right panel of Figure 6 In the eight option condition, YA separate between the better three options and the worse five options after about 50 trials. OA do this after about 75 trials. These 2-2 and 3-5 separations could reflect the participants' psychologically most salient explore-exploit representations. Together, the choice trajectories show that already from the beginning onwards, YA choose more often from the better options.

Third, we analyzed another measure of performance to compare performance across all of the options at once. We


Figure 4: Boxplots of variation in average choice proportion across participants for both choice environments.
choose to measure regret (a common measure in machine learning) as it generalizes over the specific outcomes of the random number generation process. Regret can be computed as $R_{T}=\sum_{i=1}^{100}\left(p_{\text {opt }}-p_{B(i)}\right)$, where $p_{\text {opt }}=.6$ and $p_{B(i)}$ is the hidden probability of a reward for the chosen bandit. It follows that randomly behaving agents get a total regret of 11.25 points for four options and 26.25 points for eight options. Overall, the age effect on regret was large ( $\mathrm{p}<.01$, Cohen's d .707) for eight options ( $M_{O A}=19.87, S E=.79, M_{Y A}=16.84$, $S E .75$ ) and medium ( $\mathrm{p}<.05$, Cohen's d .550) for four options $\left(M_{O A}=9.16, S E=.32, M_{Y A}=8.17, S E .33\right)$. These agerelated differences varied slightly across the unique problems, which only differed by random number generation, see Figure 5. We also investigated how regret differences emerged using the shapes of the exploration-exploitation trade-offs over trials within the choice profiles. We observed a slowing increase in regret over time in general but increasing age-related differences for both conditions, see Figure 7. Age-differences became significant after trial 24 in eight options and 23 trials in four options. It seems that exploration in OA happens less effectively. Regret was significantly ( $\mathrm{p}<.05$, t.test) better compared to a random agent (four options: YA after trial 17, OA 32; eight options: YA after trial 15, OA 16).

Fourth, we wanted to know how participant performance differed from optimality. We used Thompson sampling as an optimality model (Thompson, 1933), but we observed that differences in regret for similar algorithms are small in the context of the present task. Thompson sampling uses an inverse cumulative distribution function (also known as percentage point function or quantile function) that is used to choose the bandit with the highest certainty that the hidden probability of a bandit is smaller or equal than some randomly generated value. This way, the algorithm minimizes uncertainty that there exists a better option by making sure that the probability of choosing a certain bandit is proportional to the probability of it being the best bandit. By taking uncertainty into account, the algorithm affords a way of more rapidly adapting its decision if not only the mean of a certain


Figure 5: Variation in performance across the 16 different problems in the task.
bandit gets overtaken by another mean, but the whole posterior probability distribution. Conceptually, the algorithm keeps track of beliefs about the hidden probabilities of the bandits and then updates these beliefs each time after seeing an outcome. The algorithm is initialized by setting a uniform prior for all options. The algorithm then plays option x proportional to the probability of it being the best. Finally, it updates its priors using the newly collected information. See also Table 1 Regret as computed from applying Thompson sampling 29 times to the same games as participants played was significantly worse compared to participants (four options: YA after trial 14, OA 11; eight options: YA after trial 16, OA 16). Expected regret was 6.5 points for four options and 11.0 points for eight options, which is considerably better than YA performed on average ( $170 \%$ larger than the gap between YA and OA for four options and $193 \%$ for eight options). Interestingly, 5 out of $32 \mathrm{OA}(16 \%)$ and 9 out of 29 YA ( $31 \%$ ) achieved a median regret score within $10 \%$ of Thompson sampling for four options, while this was 1 (3\%) and 4 (14\%) for eight options. Some individuals were thus able to achieve regret scores similar to Thompson sampling.

Table 1: Thompson sampling in r pseudocode, with n being the number of bandits, $x$ a randomly generated probability, and qbeta for looking up quantiles from the Beta distribution.

| Step | Computation |
| :--- | :--- |
| Init | wins $=\operatorname{rep}(0, n)$ |
|  | pulls $=\operatorname{rep}(0, n)$ |
| Choose | softmax $(q, \theta)$ |
|  | $q=\max (q$ beta $(x, \alpha, \beta))$ |
| Update | wins $=$ wins + reward |
|  | pulls $=$ pulls +1 |
|  | $\alpha=1+$ wins |
|  | $\beta=1+$ pulls - wins |

Fifth, we wanted to know how well a fitted optimality model predicted participant's decisions. Thompson sampling


Figure 6: Observed choice proportions and one-step ahead predictions for fitted Thompson sampling. The color of the lines corresponds to the value of the hidden probabilities, where blue colors represent lower probabilities. Local polynomial regression was used as a moving window to smooth the trajectories (using a neighborhood of $40 \%$ of all points where neighboring points are also being weighted by their distance tricubically)
was fitted to individual games by scaling predicted choices using a softmax function with a fitted inverse temperature parameter $\theta$, ranging from 0.003 to 30 (higher $\theta$ values produced more randomness). OA deviated more from fitted Thompson sampling than YA did ( $\mathrm{p}<.05$, Wilcoxon tests, Cohen's d = . 57 for four options and Cohen's d $=.53$ for eight options). $\theta$ was also significantly lower for YA than for OA ( $p<.05$, Wilcoxon test, Cohen's d $=.96$ for four and 2.68 for eight options), indicating more randomness and worse matches to predictions of Thompson sampling in OA than in YA. OA and YA both significantly decreased their median $\theta$ for eight options compared to the four option condition ( $\mathrm{p}<.05$, Wilcoxon tests), see Figure $8 \theta$ for OA significantly varied more in both conditions than for YA ( $\mathrm{p}<.01$ for both conditions, Wilcoxon tests) and average variation across games was significantly lower in four options for YA, but for OA this was similar in both conditions ( $\mathrm{p}<.01$ vs. $\mathrm{p}=.4$, Wilcoxon tests). In all, OA adults were more random and less homogeneous, possibly indicating more strategy changes.

Finally, we compared the shapes of the mean observed exploration-exploitation trade-off trajectories to shapes from one-step-ahead predictions across all trials. These predictions are plotted in Figures 6 a and 6 b using dashed lines. We used the median $\theta$ of every participant for both conditions as data scaling parameter. The predictions from this fitted Thompson sampling model resulted in accurately ordered trajectories for both groups and for both conditions: The orderings of solid and dashed lines were identical for all four graphs for most trials, except in the first few trials. The latter may indicate more rapid exploration in Thompson sampling and that both YA and OA explore less rapidly, with OA taking the longest.

## Discussion and Conclusions

We aimed to identify changes in the ways OA and YA make goal-directed choices depending on the complexity of the
choice environment. We found a large age-related effect on performance in a typical eight-armed bandit task and a smaller effect in a four-armed bandit task. YA also deviated less from optimality than OA did. Choice trajectories showed that age effects were already observable in the early exploration stage, suggesting that OA explore longer or less efficiently. Theoretically, the early stages require fast exploration using not only average rewards but also their associated uncertainty. This was illustrated here using Thompson sampling, which is a kind of randomized probability matching algorithm. Participants diversified their choices similar to Thompson sampling, in line with previous work (Konstantinidis, Ashby, \& Gonzalez, 2015; Speekenbrink \& Konstantinidis, 2015). Furthermore, OA had higher and more variable inverse temperature parameter estimates across choice environments, indicating more randomness in OA. OA thus rely on less effective learning strategies that consider important information less effectively, in particular in the more complex environments.

Why would OA fail to represent important information like uncertainty or a specific task model? The role of working memory influences should be minimal as this is not strictly necessary to perform well in the task. General "slowing", gender effects, and more cautious risk taking, all of which could favor exploitation of short-term rewards, also mark cognitive aging. We did indeed observe gender interactions, age-differences in reaction times, and in standard neuropsychological test results (working memory, fluid intelligence, and risk-taking). However, as performance is mainly determined by a cognitively costly explore-exploit tradeoff and adequate future reward representations, our findings specifically point towards underreliance on goal-directed learning.

A logical next step is to assess if fits of simple learning strategies can indeed better accommodate OA. Specifically, the exploration phase seems to happen sub-optimally in OA


Figure 7: Age-differences in the increase in regret at every trial with standard deviations displayed around the means (standard errors were too narrow to visualize). Regret increased quickly first and increased slower later on, but slower for YA.
and in a more varied way. Favoring short-term rewards could be a sign of a learning mechanism that sub-optimally represents future rewards. More varied or reduced processing of important information such as uncertainty would be able to account successfully for the observed age-related changes. Furthermore, if the task indeed probes OA to rely less on goal-directed learning, we may also expect differences in connectivity to prefrontal regions (pending analyses). In all, OA may be using less effective learning strategies the more demanding the choice environment becomes. Identifying such task-dependent differences is typically neglected in neuro-computational models of decision-making. In the context of cognitive aging, this may be useful for empowering aging decision makers to navigate cognitively demanding choice environments.

## Acknowledgments

JS is supported by European Commission Horizon 2020 MSCA No 708507. Robert Lorenz, Jann Wscher, Dania Esch, and Lukas Nagel supported with participant recruitment, organization, and data collection. Materials, data, and analysis can be found on https://github.com/ jobschepens/Bandits/.

## References

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dayan, P., Huys, Q., Dzel, E., \& Dolan, R. J. (2013). Dopamine restores reward prediction errors in old age. Nature Neuroscience, 16, 648-653.
Daw, N. D., \& O'Doherty, J. P. (2014). Multiple Systems for Value Learning. In P. W. G. Fehr (Ed.), Neuroeconomics (Second Edition) (pp. 393-410). San Diego: Academic Press.
Eppinger, B., \& Bruckner, R. (2015). Towards a Mechanistic Understanding of Age-Related Changes in Learning and


Figure 8: Histogram and medians of median deviance per individual across games from fitted Thompson sampling.

Decision Making: A Neuro-Computational Approach. In T. M. Hess \& J. S. E. Lckenhoff (Eds.), Aging and Decision Making (pp. 61-77). San Diego: Academic Press.
Eppinger, B., Walter, M., Heekeren, H. R., \& Li, S.-C. (2013). Of goals and habits: Age-related and individual differences in goal-directed decision-making. Frontiers in Neuroscience, 7.
Frey, R., Mata, R., \& Hertwig, R. (2015). The role of cognitive abilities in decisions from experience: Age differences emerge as a function of choice set size. Cognition.
Konstantinidis, E., Ashby, N. J., \& Gonzalez, C. (2015). Exploring Complexity in Decisions from Experience: Same Minds, Same Strategy. In 37th annual meeting of the Cognitive Science Society (CogSci 2015) (pp. 23-25).
Mata, R., Pachur, T., Von Helversen, B., Hertwig, R., Rieskamp, J., \& Schooler, L. (2012). Ecological rationality: A framework for understanding and aiding the aging decision maker. Decision Neuroscience, 6, 19.
Mata, R., Schooler, L. J., \& Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. Psychology and Aging, 22, 796-810.
Samanez-Larkin, G. R., \& Knutson, B. (2015). Decision making in the ageing brain: Changes in affective and motivational circuits. Nature Reviews Neuroscience. 16(5), 278-289.
Speekenbrink, M., \& Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. Topics in Cognitive Science, 7, 351-367.
Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 285-294.
Worthy, D. A., Cooper, J. A., Byrne, K. A., Gorlick, M. A., \& Maddox, W. T. (2014). State-based versus reward-based motivation in younger and older adults. Cognitive, Affective, \& Behavioral Neuroscience, 14, 1208-1220.

# The impact of sleep on the formation and consolidation of spatial survey knowledge 

Wiebke Schick<br>Cognitive Neuroscience, Department of Biology University of Tübingen, Tübingen, Germany<br>Julia Holzmann<br>Cognitive Neuroscience, Department of Biology University of Tübingen, Tübingen, Germany<br>Hanspeter Mallot<br>Cognitive Neuroscience, Department of Biology University of Tübingen, Tübingen, Germany


#### Abstract

A supporting effect of sleep on memory consolidation was reported for different contents. Here, we investigated the influence of sleep on the transformation of previously learned route and place knowledge into survey knowledge, a more abstract representation. The results support the assumption of both a consolidating as well as transforming effect: the wayfinding performance in the test session - namely the usage of unfamiliar shortcuts - suggests a consolidating effect of sleep.


# The roles of item repetition and position in infant sequence learning 

Christina Schonberg (ccs185@g.ucla.edu)<br>Department of Psychology, 502 Portola Plaza<br>Los Angeles, CA 90095 USA

Gary F. Marcus (gary.marcus@nyu.edu)
Department of Psychology, 6 Washington Place
New York, NY 10003 USA

Scott P. Johnson (scott.johnson@ucla.edu)<br>Department of Psychology, 502 Portola Plaza Los Angeles, CA 90095 USA


#### Abstract

We examined mechanisms underlying infants' ability to detect, extract, and generalize sequential patterns, focusing on how saliency and consistency of distributional information guide infant learning of the most "likely" pattern in audiovisual sequences. In Experiment 1, we asked if 11- and 14-month-old infants could learn a "repetition anywhere" rule (e.g., $\mathrm{ABBC}, \mathrm{AABC}, \mathrm{ABCC}$ ). In Experiment 2 we asked if 11 - and 14-month-olds could generalize a "medial repetition" rule when its position is consistent in sequence, and in Experiment 3 we asked if 11 -month-olds could identify a nonadjacent dependency occurring at edge positions. Infants were first habituated to 4-item sequences (shapes + syllables) containing repetition- and/or position-based structure, and were then tested with "familiar" structure instantiated across new items or combinations of items vs. "novel" (random) sequences. We found that 11 -month-olds failed to learn the repetition rule both when the structure appeared in initial, medial, or final position (Experiment 1) and when it was restricted to the medial position (Experiment 2). Fourteen-month-olds learned repetition rules under both conditions. Finally, in Experiment 3 11-month-olds succeeded in learning a nonadjacent dependency in sequences identical to those used to test repetition learning in Experiment 2. Our results suggest that infants at 11 months, like adults, are relatively insensitive to patterns in the middle of sequences.


Keywords: infant learning; rule learning; sequence learning

## Introduction

In the present paper, we examine mechanisms underlying infants' ability to detect, extract, and generalize sequential patterns. Sequence learning is essential for processes ranging from the acquisition of language to everyday activities such as preparing for bed, learning to count, learning to read, and getting ready for school. Insights into development of sequence learning in infancy, therefore, are vital for theories of developmental and cognitive function across a variety of domains.

What kinds of learning mechanisms are available to infants, and what are the limits of these mechanisms? Our particular focus is on two means of knowledge acquisition, "statistical learning" and "rule learning": the extent to which infants can use transitional probability information among items to extract units from an unbroken stream of stimuli
(e.g., Saffran, Aslin, \& Newport, 1996) or the extent to which infants can distinguish simple reduplicative patterns from one another (e.g., Gerken, 2006; Marcus, Vijayan, Rao, \& Vishton, 1999), respectively. Sequence learning is guided by multiple mechanisms (Arciuli, 2017; Krogh, Vlach, \& Johnson, 2013; Thiessen, Kronstein, \& Hufnagle, 2013), and its development in infancy can be better understood by investigations of the salience and consistency of statistical and rule-governed structures (Aslin \& Newport, 2012, 2014). Some structures, such as identity relations or positions of items in order, might serve as "perceptual primitives," processed by specialized mechanisms to detect and remember specific features in patterned sequences (Endress, Nespor, \& Mehler, 2009).
The Saffran et al. (1996) transitional probability task, a well-known example of statistical learning, presented 8-month-olds an auditory stimulus consisting of four unique strings (e.g., tupiro, golabu, bidaku, and padoti) presented in random order as a continuous, unsegmented stream for 2 minutes. Infants then heard isolated strings in repetition (e.g., tupiro, tupiro, tupiro...) alternating with "part-word" strings composed of parts of two of the familiar words (e.g., rogola, rogola...) from a speaker located either to the left or right. Infants exhibited a postfamiliarization novelty preference for the part-words relative to the words, implying that they detected the differences in transitional probability across word boundaries in the input sequence.

Rule learning in sequential patterns was assessed by Marcus et al. (1999), who exposed 7-month-olds to strings that followed either an "ABA" pattern (e.g., gah tee gah) or an "ABB" pattern (e.g., gah tee tee). After 2 minutes of exposure, the infants heard the same (familiar) pattern instantiated by different phonemes (e.g., woh fei woh, dee koh dee), and a second (novel) pattern on alternating trials and showed a preference for the novel, a result that extended to a test of ABB vs. AAB. Because transitional probabilities between test-string syllables were zero, performance could not have been based on statistical learning.

Studies of infant rule learning have produced mixed results with respect to the learnability of a simple repetition rule (adjacent, as in AAB or ABB , or nonadjacent, as in ABA). Overall, the findings of these experiments seem to
differ based on how familiar the infants may be with the stimuli: 7-month-olds successfully learn ABA vs. ABB , $A B B$ vs. $A A B$, and $A A B$ vs. $A B B$ patterns when the stimuli are auditory (Marcus et al., 1999, 2007) or familiar visual stimuli (e.g., faces and animals; Bulf et al., 2015; Saffran et al., 2007). However, when stimuli consist of sequences of colored shapes, learning seems to be more difficult, perhaps because learning visual sequential input is constrained by limits in visual working memory (Johnson et al., 2009). In this case, 8 -month-olds learned a late repetition rule (adjacent repetition in the final edge position) when tested vs. nonadjacent repetition ( ABB vs. ABA ), but failed to learn late vs. early repetition ( ABB vs. AAB ), early vs. nonadjacent repetition (AAB vs. ABA ), and nonadjacent vs. late repetition (ABA vs. ABB ). Eleven-month-olds learned all these rules except nonadjacent vs. late repetition.

Adults' learning of repetition-based structure also appears to be constrained by position (Endress, Scholl, \& Mehler; 2005): Adults discriminated seven-syllable sequences from sequences of new items based on differences in internal vs. edge repetitions (e.g., ABCDDEF vs. ABCDEFF), but could only generalize when given edge repetitions. In summary, repetition structures in edge positions appear to be reliably learned by both infants and adults, but learnability of internal repetitions remains unknown for infants because the structures tested in previous structural learning studies only involved repetitions located at either the initial or final edge of the sequence ( $\mathrm{AAB}, \mathrm{ABA}$, and ABB ).

In Experiment 1, we asked if 11- and 14-month-olds can detect, extract, and generalize a "repetition anywhere" rule (i.e., $\mathrm{ABBC}, \mathrm{AABC}, \mathrm{ABCC}$ ). If infants detect repetition of items during a learning phase, they may subsequently recognize repetitions of new items (in new sequences), which we take as evidence for generalization. However, it may be that consistent position information is a key part of repetition learning at this age, as it appears to be for adults; in this case, variability in the position of the repetition might pose difficulty in its identification and recall.

In Experiment 2, we used a sequence with two possible underlying patterns to examine how consistency and salience contribute to sequence learning. In this experiment, we asked whether 11 and 14-month-olds could generalize a "medial repetition" rule when its position is consistent in sequence, but not at an edge. In Experiment 3, using the same sequence types as Experiment 2, we asked whether 11-month-olds could identify a nonadjacent dependency occurring at initial and final edge positions that may be more salient than the medial position of the repetition.

In all experiments we used an intermodal presentation method in which looming shapes were accompanied by spoken syllables, a method known to facilitate rule learning, relative to visual or auditory only, in 5- and 7-month-olds (Frank, Slemmer, Marcus, \& Johnson, 2009; Thiessen, 2012). Because infants as young as newborns look longer at randomly-ordered shape sequences vs. sequences with statistical structure (Addyman \& Mareschal, 2013; Bulf, Johnson, \& Valenza, 2011; Kirkham, Slemmer, \& Johnson,
2002), we reasoned that longer looking at novel vs. familiar sequences (i.e., random vs. structured, respectively) in the current studies would reflect learning and/or generalization of structural and/or statistical structure during habituation.

## Experiment 1

## Method

Participants Twenty 11-month-olds ( $M_{\text {age }}=11.25$ months; $S D=.297 ; 8$ girls $)$ and 2014 -month-olds $\left(M_{\text {age }}=14.20\right.$ months; $S D=.313 ; 9$ girls) participated. An additional ten 11-month-olds were tested but excluded for failure to habituate (7), fussiness (2), or preterm birth (1). an additional twelve 14 -month-olds were tested but excluded for failure to habituate (8) or fussiness (4).
Materials and Apparatus Visual stimuli consisted of 18 colored shapes (see Figure 1). Auditory stimuli consisted of an inventory of 18 spoken syllables produced with a speech synthesizer and identical to those used in Marcus et al. (1999) (e.g., bah, dee, doo, gei, jai, jah, kei, poh).


Figure 1: Schematic depiction of example habituation and test sequences for Experiment 1.

Shape-syllable pairings were determined randomly (see Figure 1). Sequences were assembled from a randomly chosen set of nine (out of the total 18) shape-syllable combinations (hereafter called "items" for simplicity), so that three items composed each four-item sequence by repeating one of the three items, either the first, second, or third (determined randomly) to yield a repetition in the initial, medial, or final position. Items always appeared in the same order within each habituation sequence.
Procedure Stimuli were presented using Macromedia Director on a Macintosh computer and a 53 cm color screen. In a separate room, the experimenter used closed-circuit video to view the infant and record his or her looking times during the experiment; the experimenter was blind to what was being presented on the screen. Before each habituation trial, a visual attention-getter appeared in the center of the screen to draw the infant's attention. Each shape was presented on a black background and increased in size from 4 cm to 24 cm high (2.4-14.6 ${ }^{\circ}$ visual angle) over a period of 667 ms ; onset of syllables ( M duration 338 ms ) was coincident with the appearance of each shape. Thus each sequence was 2 s long, and sequences were separated by a 667 ms black screen. The 4 -item sequences were randomly displayed one after another with no immediate repetition of any specific sequence. When the mean looking time over
four consecutive trials fell to less than $50 \%$ compared to the mean looking time for the first four habituation trials (i.e., habituation) infants viewed the test sequences.

At test, infants viewed "familiar" and "novel" four-item sequences drawn from the remaining nine in the total inventory that were not shown during habituation. Familiar test sequences followed the same constraints as those described previously for the habituation sequences. Novel test sequences were composed of the same nine items, presented in sequences of four, and random ordering of items with no constraints except no repeated items in any single sequence. Infants viewed six alternating familiar and novel trials presented in pairs (i.e., three test trial blocks), and viewing order was counterbalanced such that half the infants viewed a familiar trial first (followed by a novel trial) and half the infants viewed a novel trial first (followed by a familiar trial). Preliminary analyses examining sex differences in performance revealed no reliable effects in any of the experiments in this report (all $p \mathrm{~s}>.05$ ).

## Results and Discussion

A 2 (age group) $\times 2$ (trial type - novel or familiar) $\times 2$ (order - novel or familiar first) x 3 (test trial block) mixed ANOVA on posthabituation looking times revealed a main effect of test trial block, $F(2,72)=4.79, p=.011, \eta^{2}{ }_{\mathrm{p}}=.12$, the result of a decline in looking across trials, and an age group x trial type interaction, $F(1,36)=8.03, p=.008, \eta^{2}{ }_{p}$ $=.182$. There were no other significant effects. Follow-up ttests indicated that 11-month-olds did not look differently to novel and familiar test stimuli, $t(19)=-.916, p=.371$, ns, but 14 -month-olds looked longer at novel vs. familiar sequences, $t(19)=3.06, p=.006$. Thus 11 -month-olds provided no evidence for learning a "repetition anywhere" rule, whereas 14 -month-olds appeared to do so.

## Experiment 2

Because 11-month-olds showed no evidence of learning repetitions in variable locations in Experiment 1, in Experiment 2 we examined the role of positional consistency in sequence learning. Here, we tested 11 and 14-month-olds' learning of a medial repetition rule. Habituation sequences comprised two different patterns that could be extracted: a medial repetition rule (a changing identity in the AxxC pattern), and a nonadjacent dependency (between A and C in the AxxC pattern). Experiment 2 specifically tested rule learning, and as in Experiment 1, we reasoned that this learning would be reflected in longer looking during novel vs. familiar test trials.

## Method

Participants Twenty 11-month-olds (Mage $=11.15$ months, $\mathrm{SD}=.34 ; 14$ girls ) and twenty 14 -month-olds ( $M_{\text {age }}=14.14$ months, $S D=.39$; 9 girls) participated in Experiment 2. Six additional 11-month-olds were tested but excluded for failure to habituate (4) or fussiness (2). Eight additional 14-month-olds were tested but excluded due to failure to habituate (2), technical error (2), or fussiness (4).

Materials and Apparatus The item stimuli and presentation apparatus were the same as in Experiment 1. Habituation sequences contained both a medial repetition and a nonadjacent dependency between the first and fourth items of the sequence (see Figure 2). Sequences were assembled from a randomly chosen set of ten from the inventory of 18. Again, three items composed each fouritem sequence, but the second item was always repeated, instantiating a medial repetition rule. Four items were selected (from the ten) for first and fourth positions in two unique sequences, and three items were selected for the medial positions in each sequence (e.g., ABBC, DEEF, AGGC, DHHF, etc.). The two sequences were presented in alternation during habituation.

The test sequences were constructed such that familiar sequences tested the generalization of the medial repetition rule with new exemplars that did not use consistent shapes/syllables in the first and fourth positions across sequences. Familiar sequences were composed of items drawn from the entire shape inventory, with the constraints that the second item always repeated, and the first and fourth items in sequence could not be one of the four items that occupied those positions in the habituation sequences. The novel sequences followed the same constraints described in Experiment 1.


Figure 2: Schematic depiction of example habituation and test sequences for Experiments 2 and 3.

Procedure The procedure was the same as in Experiment 1. Infants were habituated to sequences that contained a medial repetition and a nonadjacent dependency (described above), and an infant-controlled habituation paradigm was used. At test, infants saw six trials that alternated between familiar (i.e., contained a medial repetition) and novel, viewing three of each trial type in total. Infants were randomly assigned to see either a familiar trial or a novel trial first.

## Results and Discussion

A 2 (age group) $\times 2$ (trial type) $\times 3$ (test trial block) mixed ANOVA revealed a significant 3-way interaction, $F(2,76)=$
$5.71, \mathrm{p}=.022, \eta_{\mathrm{p}}^{2}=.13$, which stemmed from a relatively precipitous decline in looking across novel trials by older infants but not younger infants. More importantly, there was an age x trial type interaction, $F(1,38)=7.42, p .=010, \eta^{2}{ }_{\mathrm{p}}$ $=.16$, due to differences in looking at novel and familiar test sequences. Although this analysis yielded a significant main effect of trial type, $F(1,38)=4.36, p=.044, \eta_{\mathrm{p}}^{2}=.10$, this effect was driven by 14-month-olds' longer looking to novel test trials, as 11 -month-olds looked equally to novel and familiar test trials. There was also a significant main effect of test trial block, $F(2,76)=25.11, p<.001, \eta^{2}{ }_{\mathrm{p}}=.40$, due to a decline in looking times across trials.
14 Month Olds A 2 (trial type) x 2 (order) x 3 (test trial block) mixed ANOVA revealed a significant main effect of trial type, $F(1,18)=8.20, p=.010, \eta_{\mathrm{p}}^{2}=.313$, due to longer looking overall at the novel test sequence. There was also a reliable main effect of test trial block, $F(2,36)=22.37, p<$ $.001, \eta_{\mathrm{p}}^{2}=.55$, due to a decline in looking across trials. These main effects were qualified by a significant trial type x test trial block interaction, $F(2,36)=10.70, p=.004, \eta_{\mathrm{p}}^{2}=$ .37 and a significant trial type x order x test trial block interaction, $F(2,36)=9.25, p=.007, \eta_{\mathrm{p}}{ }^{2}=.34$. There were no other significant effects. The two higher-order interactions were a result of longer looking during the first trial block toward the novel sequence by infants in the novel-first order, relative to infants in the familiar-first order, $t(18)=2.29, p=.034$; comparisons across the second and third trial blocks were $n s, p s>.16$.
11 Month Olds A 2 (trial type) x 2 (order) x 3 (test trial block) mixed ANOVA yielded a main effect of test trial block, $F(2,36)=5.50, p=.031, \eta_{\mathrm{p}}^{2}=.23$, the result of a decline in looking across trials, and a significant trial type $x$ order interaction, $F(1,18)=6.44, p=.021, \eta_{\mathrm{p}}{ }^{2}=.23$, due to a (nonsignificant) tendency for infants in both order conditions to look longer at the trial type that was presented first. There were no other significant effects; the trial type effect was ns at $p=.364$. These results support the conclusion that the 14 -month-olds generalized the medial repetition rule, whereas the 11 -month-olds did not.

## Experiment 3

Because 11-month-olds failed to learn the "medial repetition" rule in Experiment 2, Experiment 3 addressed the possibility that learning structures in 4 -item sequences is too difficult for 11-month-olds, perhaps due to limits in visual working memory. We used the same habituation sequences as Experiment 2 but instead tested for statistical learning (specifically, the nonadjacent dependency between A and C in the AxxC pattern). As in previous experiments, we reasoned that this learning would be reflected in longer looking during novel vs. familiar test trials.

## Method

Participants Twenty 11 -month-olds $\left(M_{\text {age }}=11.16\right.$ months, $S D=.32 ; 6$ girls) participated in Experiment 3. An additional nine infants were tested but excluded due to failure to habituate (8) or fussiness (1).

Materials and Apparatus The item stimuli and presentation apparatus were the same as in Experiments 1 and 2.
Procedure The procedure was the same as in Experiment 2, except for the structure of the test sequences. The familiar test trials maintained the relation between the first and fourth items across two sequences, using the exact same first and fourth items from habituation, but had no repetitions (i.e., each familiar sequence was composed of four unique items; see Figure 2). The novel test trials followed constraints described previously.

## Results and Discussion

A 2 (trial type) x 2 (order) x 3 (test trial block) mixed ANOVA yielded a reliable main effect of trial type, $F(1,18)$ $=7.86, p=.012, \eta_{\mathrm{p}}^{2}=.30$, due to longer looking overall at novel vs. familiar test sequences (see Figure 4). There was also a main effect of trial block, $F(2,36)=7.82, p=.012$, $\eta_{\mathrm{p}}^{2}=.30$, due to a decline in looking across trials, and a significant interaction between trial block and trial type, $F(2,36)=9.63, p=.006, \eta_{\mathrm{p}}^{2}=.35$. Infants looked more toward the novel sequence than the familiar in the first block, $t(19)=2.75, p=.013$, and in the second block, $t(19)$ $=2.80, p=.011$, but not in the third block, $t(19)=.22$, $n s$. The overall presence of a novelty preference suggests that infants abstracted the nonadjacent dependency pattern during habituation.

We compared performance of the 11-month-olds in Experiments 2 and 3 with a 2 (experiment) x 2 (trial type) x 3 (test trial block) mixed ANOVA. This analysis revealed a significant main effect of test trial block, $F(2,76)=11.44, p$ $<.001, \eta_{\mathrm{p}}^{2}=.23$, due to a decline in looking times across trials. More importantly, there was an experiment x trial type interaction, $F(1,38)=6.83, p=.013, \eta_{\mathrm{p}}^{2}=.15$, due to differences in looking at novel and familiar test sequences, as noted earlier. Taken together, therefore, the results of Experiments 2 and 3 provide evidence that when both statistical and rule-bound information was available in habituation sequences, 11 -month-olds detected the presence or absence of nonadjacent dependencies (ordinal positions of initial and final shapes in sequence; i.e., statistical information) across habituation and test, but not a medial repetition (i.e., rule-bound information).

## General Discussion

In a departure from past studies showing that 11-month-olds learn an adjacent repetition rule when the repetition appears in the initial or final positions in sequence (Johnson et al., 2009), we discovered that 11-month-olds failed to learn this rule when the repetition appeared in any position (initial, medial, or final, Experiment 1), or when it was restricted to the medial position (Experiment 2). Fourteen-month-olds, however, appeared to learn repetition rules under both conditions. Finally, in Experiment 3, 11-month-olds succeeded in learning a nonadjacent dependency in sequences identical to those used in Experiment 2. We conclude that 11 -month-old infants do not seem to
recognize repetitions when they appear in multiple positions in sequence, or in a consistent middle position, although items at edge positions in sequence appear to be distinctly salient. Similar findings for adults were reported by Endress et al. (2005, 2010).

## Repetition and Position as Perceptual Primitives

Infant sequence learning is constrained by saliency and consistency of information well as general limits in attention and memory (Aslin \& Newport, 2012, 2014). Some of these constraints are specific to modality (e.g., speech cues; Johnson \& Jusczyk, 2001; Thiessen \& Saffran, 2003) or the experimental setting (e.g., gaze or action cues; Baldwin, Andersson, Saffran, \& Meyer, 2008), but others, such as repetition and position, are domain-general and may operate similarly across contexts. Evidence from the 11-month-olds in Experiment 2 is consistent with findings from adults, who generalized a repetition to new items when it appeared in final position, but not a medial position, in 7-item syllable sequences (Endress et al., 2005). These studies suggest that item position, most notably final position, is more salient than item repetition, relatively speaking; these findings are consistent with the well-documented serial position curve (Ebbinghaus, 1885) and the recency effect in memory (Baddeley \& Hitch, 1974)

Because any particular set of items in a group potentially supports an infinite number of possible structures and generalizations thereof, a learner must determine the most likely pattern given a limited amount of experience with it. One way in which this problem may be constrained is by a "gradient of generalization:" the most consistent information across a distribution produces the best learning (Aslin \& Newport, 2014; Gerken, 2006). In Experiments 2 and 3 , information for medial repetition and nonadjacent dependency was available, yet 11-month-olds learned only the statistical information. Notably, 14 -month-olds appeared to learn a repetition rule both when it was restricted to the medial position (Experiment 2) and when it was free to appear in initial, medial, or final position (Experiment 1), implying that important developments in structural learning consist of the "separation" of perceptual primitives such that they become less interdependent and perhaps more salient on their own.

## Infant Sequence Learning in Context

Rule learning and generalization for shorter sequences can be observed in infants as young as 4 months (Dawson \& Gerken, 2009). Rule learning in 5-month-olds from 3-item shape-syllable sequences was also reported, using a similar design (Frank et al., 2009). Studies that tested for statistical learning reported that 3-month-olds appeared to recognize violations of serial order in 3-item shape-sound sequences (Lewkowicz, 2008), and 5-month-olds segmented shape sequences from differences in transitional probability (Marcovitch \& Lewkowicz, 2009; Slone \& Johnson, 2015). (To our knowledge, there is no published evidence for rule or statistical learning in auditory or visual sequences prior to

4 months.) By 8 months, infants seem to use a "chunking" mechanism to segment shape sequences when tested for learning of "illusory" sequences or "embedded" units in streams of looming shapes (Slone \& Johnson, 2016; cf. Endress \& Mehler, 2009; Giroux \& Rey, 2009).

Finally, consider the findings (from Experiments 2 and 3) that 11-month-olds extracted statistical patterns, but not rules, from identical sequences. In a previous test of multiple pattern learning, adults listened to speech streams that could be interpreted in terms of rules or statistical relations (Endress \& Bonatti, 2007). With briefer listening times, participants learned the rules, but did not identify the statistical structure without substantially longer exposure durations. This result led to the claim that there is a fastworking mechanism for extracting rule-bound patterns, and a second slower mechanism that requires additional time to learn associations among items. Yet the infants we observed appeared to learn statistical relations, but not rules, during a relatively brief period of habituation. The reasons for this effect are unclear. Recently, 8-month-old infants were found to learn different statistical structures (transitional probabilities and "chunks" of items) as a function of exposure time (Slone \& Johnson, 2016), and it may be that 11-month-olds would learn rules in the current stimulus set if they accumulated more looking times than allowed for by the infant-controlled habituation method. Nor is it clear from the current studies or the larger literature whether, in general, rule learning systems might come "on line" earlier during development than statistical learning systems, or vice-versa. These questions await future study.

## Conclusions

Perceptual primitives may be best thought of as helping to support learning by attracting learners' attention and memory resources to likely structures in the environment. Yet they do not seem to attract attention automatically, as do some sensory primitives such as motion or high contrast, nor are they automatically committed to memory, for either infants or adults. Rather, evidence to date suggests that infants at birth can discriminate certain rules and statistical patterns when compared to unstructured input, but learning and generalization of rules develop across the first year after birth and beyond. On this account, perceptual primitives such as repetition and position serve as building blocks upon which more complex structures can be built.

## Acknowledgments

This research was supported by NIH Grant R01-HD73535. We thank the research staff at the NYU Infant Perception Lab and the UCLA Baby Lab, and our infant participants and their families.

## References

Addyman, C., \& Mareschal, D. (2013). Local redundancy governs infants' spontaneous orienting to visual-temporal sequences. Child Development, 84, 1137-1144.

Arciuli, J. (2017). The multi-component nature of statistical learning. Philosophical Transactions of the Royal Society B, 372, 20160058.
Aslin, R. N., \& Newport, E. L. (2012). Statistical learning: from acquiring specific items to forming general rules. Current Directions in Psychological Science, 21, 170176.

Aslin, R. N. \& Newport, E. L. (2014). Distributional language learning: mechanisms and models of category formation. Language Learning, 64: Cognitive Neuroscience Supplement 2, 86-105.
Baddeley, A. D., \& Hitch, G. (1974). Working memory. Psychology of Learning and Motivation, 8, 47-89.
Baldwin, D., Andersson, A., Saffran, J., \& Meyer, M. (2008). Segmenting dynamic human action via statistical structure. Cognition, 106, 1382-1407.
Bulf, H., Brenna, V., Valenza, E., Johnson, S. P., \& Turati, C. (2015). Many faces, one rule: The role of perceptual expertise in infants' sequential rule learning. Frontiers in Psychology, 6: 1595.
Bulf, H., Johnson, S. P., and Valenza, E. (2011). Visual statistical learning in the newborn infant. Cognition, 121, 127-132.
Dawson, C., \& Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7 -month-olds do not. Cognition, 111, 378-382.
Ebbinghaus, H. (1885). Uber das Gedachtnis (On memory). Leipzig: Duncker \& Humblot.
Endress, A., \& Bonatti, L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. Cognition, 105, 247-299.
Endress, A. \& Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. Journal of Memory and Language, 60, 351-367.
Endress, A. D., Nespor, M., \& Mehler, J. (2009). Perceptual and memory constraints on language acquisition. Trends in Cognitive Sciences, 13, 348-353.
Endress, A. D., Scholl, B. J., \& Mehler, J. (2005). The role of salience in the extraction of algebraic rules. Journal of Experimental Psychology: General, 134, 406-419.
Frank, M. C., Slemmer, J. A., Marcus, G. F., \& Johnson, S. P. (2009). Information from multiple modalities helps five-month-olds learn abstract rules. Developmental Science, 12, 504-509.
Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. Cognition, 98, B67-B74.
Gervain, J., Macagno, F., Cogoi, S., Peña, M., \& Mehler, J. (2008). The neonate brain detects speech structure. Proceedings of the National Academy of Sciences (USA), 105, 14222-14227.
Gervain, J., Berent, I., \& Werker, J. F. (2012). Binding at birth: the newborn brain detects identity relations and sequential position in speech. Journal of Cognitive Neuroscience, 24, 564-574.

Giroux, I. \& Rey, A. (2009). Lexical and sublexical units in speech perception. Cognitive Science, 33, 260-272.
Johnson, E. K., \& Jusczyk, P. W. (2001). Word segmentation by 8 -month-olds: When speech cues count more than statistics. Journal of Memory and Language, 44, 548-567.
Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus. G., Rabagliati, H., \& Slemmer, J. A. (2009). Abstract rule learning for visual sequences in 8 -and 11-month-olds. Infancy, 14, 2-18.
Kirkham, N. Z., Slemmer, J. A., \& Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition, 83, B35B42.
Krogh, L., Vlach, H. A., \& Johnson, S. P. (2012). Statistical learning across development: Flexible yet constrained. Frontiers in Psychology, 3, 598.
Marcovitch, S., \& Lewkowicz, D. L. (2009). Sequence learning in infancy: The independent contributions of conditional probability and pair frequency information. Developmental Science, 12, 1020-1025.
Marcus, G. F. (2000). Pabiku and ga ti ga: Two mechanisms children could use to learn about language and the world. Current Directions in Psychological Science. 9, 145-147.
Marcus, G., Fernandes, K., \& Johnson, S. (2007). Infant rule learning facilitated by speech. Psychological Science, 18, 387-391.
Marcus, G. F., Vijayan, S., Rao, S. B., \& Vishton, P. M. (1999). Rule learning by seven-month-old infants. Science, 283, 77-80.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-oldinfants. Science, 274, 1926-1928.
Saffran, J. R., Pollack, S. D., Seibel, R. L., \& Shkolnik, A. (2007). Dog is a dog is a dog: infant rule learning is not specific to language. Cognition, 105, 669-680.
Slone, L. K. and Johnson, S. P. (2015). Infants' statistical learning: 2-and 5-month-olds'segmentation of continuous visual sequences. Journal of Experimental Child Psychology, 133, 47-56.
Slone, L. K., \& Johnson, S. P. (2016). When learning goes beyond statistics: infants represent visual sequences in terms of chunks. Manuscript submitted for publication.
Thiessen, E. D. (2012). Effects of inter- and intra-modal redundancy on infants' rule learning. Language Learning and Development, 8, 197-214.
Thiessen, E. D., Kronstein, A. T., and Hufnagle, D. G. (2013). The extraction and integration framework: a twoprocess account of statistical learning. Psychological Bulletin, 139, 792-814.
Thiessen, E. D., \& Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7to 9-month-old infants. Developmental Psychology, 39, 706-716

# Silent gesture and noun phrase universals 

Marieke Schouwstra, Simon Kirby, Jennifer Culbertson<br>Centre for Language Evolution<br>University of Edinburgh


#### Abstract

In this paper we investigate a hypothesized cognitive bias for isomorphic mappings between conceptual structure and linear order in the noun phrase. This bias has been proposed as a possible explanation for a striking asymmetry in the typology of the noun phrase-linear orders which place the adjective closest to the noun, then the numeral, then the demonstrative, are over-represented in the world's languages. Previous experimental work has provided evidence that an isomorphism bias affects English-speaking learners' inferences about the relative order of modifiers in an artificial language. Here, we use the silent gesture paradigm to explore whether the isomorphism bias influences spontaneous gestures innovated by participants in a modality with which they have relatively little prior experience. We find that gesture string order largely conforms to the same striking pattern found in noun phrase typology, supporting the role of the isomorphism bias in shaping the emergence of language (and language-like) systems.


Keywords: silent gesture; noun phrase; word order; linguistic universals; cognitive biases

## Introduction

Linguists studying word order have long noticed striking differences in frequency among possible word order patterns. Explaining why certain patterns are more common than others is a source of ongoing debate. On the one hand, these typological differences may reflect evolved properties of human cognition. On the other hand, they may be the result of a complex interplay between various non-cognitive factors: genetic and areal relationships between languages, social or cultural pressures, and accidents of history (Evans \& Levinson, 2009; Dunn, Greenhill, Levinson, \& Gray, 2011; Piantadosi \& Gibson, 2014; Ladd, Roberts, \& Dediu, 2015).

In this paper we will investigate a well-known pattern in language typology relating to the structure of the noun phrase, using a experimental methodology that has not yet been applied in this domain, in which participants must improvise gestures to communicate pictures or scenes. What we find is a clear preference for gesture orders which conform to a structural template that is found in the majority of languages. These orders do not in general follow the typical linear order of noun phrases in their native language, English. Most prominently, they often produce adjective gestures following the noun. We argue that the results of our experiment reflect the underlying conceptual structure of the noun phrase, suggesting a cognitive explanation for the typological pattern.

## Universal 20 and the isomorphism bias

Greenberg (1963) formulated a number of typological 'universals', based on the relative frequency of syntactic patterns in 30 different languages. Universal 20 concerns the Noun Phrase, in particular the order of the noun and its modifiers.

It states that when any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or the exact opposite (Greenberg, 1963).

In other words, the three most common NP orders according to Greenberg are Dem-Num-Adj-N ('these five large houses'), N-Dem-Num-Adj ('houses these five large'), and N-Adj-Num-Dem ('houses large five these'). More recent analyses, based on a larger set of languages have found that of the three orders, N -Dem-Num-Adj does occur, but is far less frequent than the other two (Cinque, 2005; Dryer, 2009).

To explain the difference in frequency between Greenberg's two post-nominal orders ( $\mathrm{N}-\mathrm{Adj}$-Num-Dem and N-Dem-Num-Adj), Culbertson and Adger (2014) appeal to the notion of isomorphism, present in some form in a number of theoretical accounts of this universal (Cinque, 2005; Rijkhoff, 2004). In general, isomorphism refers to a transparent relationship between meaning and structure. To see how this applies to the noun phrase, consider the distinct semantic contributions of the different modifier types (Culbertson \& Adger, 2014). Adjectives modify properties that are inherent to the noun, numerals group together these smaller units, and demonstratives connect these grouped units to the external discourse.

In a complex noun phrase, the adjective is thus conceptually closest to the noun, followed by the numeral, and finally the demonstrative. These relations determine constituency, and can be seen, for example, in semantic scope. A numeral (like 'five) takes scope over the noun+adjective unit (like 'large houses'); the meaning of the numeral applies to the noun as well as to the adjective. Similarly, a demonstrative takes scope over a noun+adjective+numeral unit (such as in 'these five large houses') to connect it to the discourse. ${ }^{1}$ These conceptual relations are illustrated in Figure 1.


Figure 1: The conceptual structure of the noun and its modifiers: the adjective modifies the meaning of the noun most closely; the numeral takes scope over this unit; the demonstrative is conceptual most distant, taking widest scope.

[^476]Note that the conceptual structure does not fully determine linear order, rather there are several possible ways to map structure to order, all of which preserve the underlying relations between elements. For instance, Dem-Adj-N-Num, Dem-N-Adj-Num, and Adj-N-Num-Dem can all be 'read off' the structure in Figure 1 directly, without perturbing the constituency relations. On the other hand, Adj-Dem-N-Num cannot be read off the structure directly, the only way to get this order is to move Adj outside of its unit with N . There are in fact eight ways of forming a structure-preserving string, these are the isomorphic orders. They make up one third of the 24 possible ways of ordering Dem, Num, Adj and N.

Returning to the two post-nominal orders mentioned above, N-Adj-Num-Dem and N-Dem-Num-Adj, we can now see that isomorphism is a possible explanation for the frequency asymmetry between them: the more frequent N -Adj-Num-Dem is isomorphic, while the infrequent N -Dem-Num-Adj is not. More generally, isomorphic orders tend to be more frequent than non-isomorphic ones, as shown in Figure 2.


Figure 2: Frequency of NP structures: isomorphic orders are all among the most common in this sample (Cysouw, 2010).

To investigate whether a bias for isomorphism plays a role in language learning, Culbertson and Adger (2014) conducted a series of experiments in which participants were trained on an artificial language with simple noun phrases consisting of a noun plus a single modifier (either an adjective, a numeral or a demonstrative). Participants were then tested on complex noun phrases with more than one modifier (i.e., an adjective and a demonstrative), which they had not seen before. For instance, participants who learned N-Adj and N-Dem strings in the training phase were then prompted to construct a phrase containing all three elements. They could either choose N -Adj-Dem, which is isomorphic, or N -Dem-Adj, which shares the modifier order of English (Dem-Adj) but is
non-isomorphic. Participants chose isomorphic structures in the majority of cases, suggesting that relative order of modifiers was inferred based on an underlying assumption of isomorphism rather than surface similarity to English.

## Silent gesture: evidence from improvisation in the lab

The findings of Culbertson and Adger (2014) are limited in the extent to which they provide evidence for an isomorphism bias reflecting a general property of human cognition, since participants may have learned (through learning their native language) at a more abstract level that surface order is isomorphic to conceptual structure. Recent work using the 'silent gesture' paradigm offers a potential method for tapping into biases in word order while bypassing the effects of prior linguistic knowledge. In silent gesture experiments, participants with no knowledge of sign language are asked to convey information using only their hands and no speech. Existing work using this method has mainly focused on basic word order (sequences expressing information about who did what to whom). This research has found that when participants use silent gesture to describe simple transitive events, they do not rely on the dominant order of their native language (GoldinMeadow, So, Özyürek, \& Mylander, 2008), but instead take the semantic properties of the event into account (Gibson et al., 2013; Hall, Mayberry, \& Ferreira, 2013; Schouwstra \& de Swart, 2014; Schouwstra, 2016).

## Experiment 1

To investigate the isomorphism bias in a modality distinct from participants' previous language experience, we conducted an experiment in which adult participants used silent gesture to describe pictures of objects modified in various ways. We hypothesise that the ordering of these gestures will conform to isomorphism, even when they do not reflect the linear order of English Noun Phrases.

## Materials, participants, and procedure

We created a stimulus set consisting of images of groups of 4 or 5 (Num) shapes, which were either squares or triangles (N), either striped or spotted (Adj), and appeared in a proximal or a distal (Dem) location. Locations were represented by two iPads on which the images could appear-one closer to the participant, the other further away. Figure 3 provides example images, and Figure 4 shows the position of the iPads relative to the participant. The set of 8 different images, presented on two different iPads, together formed 16 total items.

Participants ( $\mathrm{N}=20$, native speakers of English, no experience with sign languages) were seated across the table from the experimenter, with the two iPads in front of them, as in Figure 4. They were filmed using a Logitech webcam connected to a MacBook Air. Before starting the silent gesture part of the experiment, participants were shown the full set of stimuli as printed pictures. Subsequently, they were asked to describe each stimulus item using only their hands, so that someone watching the recording would be able to work out


Figure 3: Example stimuli: 'five striped squares' and 'four spotted triangles'.


Figure 4: Experiment set-up. Note that the two iPads were placed on the table, in front of the participant: one close to the participant, and one further away.
which of the images appeared on which screen. The stimuli were presented twice in random order for each participant, with a brief break after the first run of the stimuli ( 32 trials in total).

## Results

The videos were coded by identifying which of the gestures indicated information associated with $\mathrm{N}, \mathrm{Adj}$, Num, or Dem. ${ }^{2}$ Occasionally, participants specified the spatial layout of the figures. This information was invariably provided in addition to other gestures referring to the number or the object, and was ignored for our coding purposes. In addition, some gestures included combinations of two elements (e.g., N with Adj ). No relative ordering information can be determined for combined elements, therefore these were excluded from relevant analyses. We focused on two measures in analyzing this data: how were modifiers ordered relative to the noun (were they pre- or post-nominal), and how were modifiers ordered relative to each other (were they isomorphic given the position of the noun, or not). To code for isomorphism we looked at each modifier pair that appeared on the same side of the noun. For instance, a string N -Num-Adj-Dem would be coded as non-isomorphic for Num-Adj, but isomorphic for Dem-Num and Dem-Adj. Modifier pairs that were on different sides of the noun were excluded from this part of the analysis as these do not provide information about isomorphism. Finally, overall isomorphism for each full gesture string was

[^477]coded according to whether any isomorphism violations were present. For this overall measure (contrary to what we did for the modifier pairs) we did include strings that had modifiers on different sides of the noun. Strings that excluded any of the modifiers (such as $\mathrm{N}-\mathrm{Adj}-\mathrm{Num}$ ) were excluded.


Figure 5: Results of Experiment 1: Proportion of modifiers placed post-nominally, by modifier type.


Figure 6: Results of Experiment 1: Proportion of modifier pairs conforming to isomorphic order, by modifier pair, in pre- and post-nominal position, plus overall scope isomorphism (last column). The dotted line represents chance level. Note that for overall isomorphism, chance level is at 0.33 (8 isomorphic orders/24 possible orders).

The typical order of English noun phrases is Dem-Num-Adj-N. Analysis of overall rates of pre- vs. post-nominal placement for each type of modifier reveal that participants' gestures deviate from pre-nominal order most obviously when it comes to the placement of Adj. Indeed, as shown in Figure 5, there was a strong preference for post-nominal placement. By contrast for Dem, participants preferred a pre-nominal position. Num fell in between. This was confirmed by one sample $t$-tests comparing average placement by participant to chance for each modifier type ( t $=-0.2717, \mathrm{df}=18, \mathrm{p}=0.79$ for Num , and $\mathrm{t}=-6.9561$, $\mathrm{df}=$ $15, \mathrm{p}<.001$ for Dem; because Adj was used post-nominally without exceptions, a t-test cannot be performed on that
data). ${ }^{3}$.
Turning to isomorphism, as Figure 6 shows, participants had a strong tendency to provide gesture strings which conformed to isomorphic ordering for all modifier pairs. Importantly, this was the case both for pre-nominal and postnominal pairs. A one sample t-test confirms that proportions of isomorphic word orders are significantly different from chance $(\mathrm{t}=65.1549, \mathrm{df}=15, \mathrm{p}<.001)$.

## Discussion

Participants' overwhelming preference for isomorphic order in silent gesture strings provides support for a link between cognitive biases and this typological tendency. A further surprising finding is that participants very consistently produced post-nominal adjectives. This pattern is not dominant in the native language of the participants (English), although it is in fact more common typologically (Dryer, 2009).

However, two properties of our stimulus items suggest the possibility that our results may not generalize to other Noun Phrase types. First, the most common gesture order used by participants was Dem-Num-N-Adj. This corresponds to the physical layout of the information in each item: the iPad was the outermost, largest element, the numeral information was in some sense the next largest part of the scene, then the object shape itself, and finally, the adjective information (stripes or spots) which was always inside the object (Figure 3). If participants were starting from the outermost information and proceeding in, then, our stimuli could have set participants up to place the adjective after the noun.

Further, as mentioned above, some responses contained gestures which combined information for two of the elements present. For example, participants sometimes conveyed information about the numeral and the adjective simultaneously, for example by repeatedly drawing spots or stripes (Adj) four or five (Num) times. Such combinations were much more likely to involve the adjective, making it impossible to determine either isomorphism (for Adj combined with other modifiers), or position relative to the noun (for Adj combined with N ) in a number of cases.

We therefore conducted an additional experiment using stimuli in which the adjective is not depicted inside the object, and which discourage the use of combined gestures.

## Experiment 2

To address the concerns pointed out above, and to further investigate the prevalence of post-nominal adjectives, we conducted a second silent gesture experiment, using different stimuli, as described below.

## Materials, participants, and procedure

Our stimuli consisted of line drawings of groups of 4 or 5 (Num) objects, which were either toothbrushes or pencils ${ }^{4}$ (N),

[^478]either big or small (Adj), and appeared on a proximal or a distal (Dem) location. The adjectives 'big' and 'small' were chosen on the basis of their visual properties: when depicted, the adjective information is not visually inside the object. Moreover, we expected that these adjectives would lead to fewer combined gestures, particularly with the noun (since both likely require difference handshapes). ${ }^{5}$


Figure 7: Example stimuli: 'five large pencils' and 'four small pencils'.

The procedure of the experiment was identical to that of experiment 1 , except for the number of trials: participants ( $\mathrm{N}=20$, native speakers of English, no experience with sign languages) described each of the stimuli once ( 16 trials in total).


Figure 8: Results of Experiment 2: Proportion of modifiers placed post-nominally, by modifier type.

## Results

As in Experiment 1, we coded the videos by identifying which portion of the gesture string indicated $\mathrm{N}, \mathrm{Adj}$, Num, or Dem. From this we determined whether the modifiers were placed pre- or post-nominally, and obtained isomorphism scores for modifier-pairs as well as full strings. The results show that although the proportions follow the pattern found in experiment 1 (note, however, that the proportion of postnominal adjectives is no longer significantly greater than 0.5 ; $\mathrm{t}=0.5138, \mathrm{df}=18, \mathrm{p}=0.61$ ). Additionally, there was again

[^479]an overall preference to produce isomorphic structures: a one-sample T-test confirms that the proportion of isomorphic strings differs from chance $(\mathrm{t}=5.7149, \mathrm{df}=16, \mathrm{p}<0.001) .{ }^{6}$ However, this tendency was less deterministic than in Experiment 1.

Zooming in on the scores for different modifier pairs, participants are less likely to produce isomorphic order for Adj-Num combinations, in contrast to Experiment 1. When these two modifiers were placed post-nominally, they were no longer isomorphic (see Figure 9; $\mathrm{t}=0.9077, \mathrm{df}=10, \mathrm{p}=$ 0.39).


Figure 9: Results of Experiment 2: Proportion isomorphic orders, by modifier pair, plus overall scope isomorphism (last column). The dotted line represents chance level. Note that for overall isomorphism, chance level is at 0.33 ( 8 isomorphic orders/24 possible orders).

## General discussion

Isomorphism is a hypothesized cognitive principle proposed to explain the way Noun Phrases tend to be ordered in languages of the world. Languages which obey isomorphism are much more frequent than those which don't. However, given the many other factors likely to influence typological distributions, evidence which explicitly links isomorphism to a cognitive bias is needed. Previous experimental studies confirmed that isomorphism appears to play a role in the kinds of inferences people make when learning an artificial language with word order that differs from their native language (Culbertson \& Adger, 2014). However, the linguistic systems individual speakers of a given language know have already solved the problem of going from a multidimensional conceptual structure (see Figure 1 above) to a linear representation. In this paper, we investigated what happens when people start 'from scratch' and improvise utterances in the absence of a conventional system, by conveying information presented as images using only their hands and no speech.

[^480]This silent gesture paradigm has been shown to be a fruitful way of investigating what happens when people are forced to communicate without being able to use existing word order conventions. However, previous work using this paradigm focuses mainly on the order of major sentence constituents. The structure of the noun phrase has never before been studied using silent gesture.

Our experiments showed that the gestures improvised by participants to describe pictures with N, Adj, Num, and Dem information, are ordered in a way that is isomorphic to the underlying conceptual structure-adjective property closer to the noun than numerosity, and distal/proximal location furthest away. This result, combined with the fact that participants did not simply use English NP order for their gestures, supports the claim that a bias for isomorphism affects linear order independently of prior linguistic experience. This general bias therefore plays a plausible role in explaining the frequency distribution of NP orders across languages.

The clearest difference between the gesture orders participants provided and their native language experience is in the placement of the adjective. Experiment 1 showed an extremely strong preference to place the Adjective postnominally (unlike in English), and in Experiment 2, though the placement was more variable, there was still no overall preference for pre-nominal adjectives. As mentioned above, post-nominal adjectives are in fact more common cross-linguistically, and this may represent a second active bias. Interestingly, Adj and Num gestures in Experiment 2 showed no isomorphism bias, ${ }^{7}$ whereas they did in Experiment 1. One possible explanation for this may lie in the nature of the adjectives used in our two experiments. Adjectives describing a texture were used in Experiment 1, while adjectives for size were used in Experiment 2. Size adjectives are gradable (Kennedy, 2007), a property which affects the role of contextual information in their interpretation: for gradable adjectives, context is needed to determine what counts, e.g., as 'large' or 'small'. This closer connection to the context may make the adjectives in the second experiment less conceptually tied to the noun. Perhaps relatedly, size adjectives are general argued to scope higher, relative to other adjectives (Kemmerer, 2000), including 'striped' and 'spotted' (as reflected in their order: e.g., 'small striped triangle' sounds more natural than 'striped small triangle'). An nonisomorphic order of a numeral and a wider-scoping adjectives may thus be a weaker violation of isomorphism compared to a lower-scoping adjective. Additional investigating with a wider range of adjectives is needed to justify these claims. Note however, that they are related to a similar finding in Culbertson and Adger (2014), in which non-isomorphic orders were more likely to be chosen for structurally less distant modifier pairs (i.e., Adj with Num, compared to Adj with Dem).

[^481]
## Conclusion

The experiments reported here show that when people improvise gestures corresponding to simple pictures of objects with different properties, numerosities, and locations, they order their gestures in a way that corresponds to the underlying conceptual structure of these elements. This same structure is respected by the majority of languages in the way they order elements in the Noun Phrase (nouns, adjectives, numerals, and demonstratives), suggesting that a cognitive bias for isomorphism between meaning and linear order might shape linguistic systems in this domain.

Experiments that use improvised silent gesture, like the ones presented here, provides a window into the evolution of linguistic systems. The method gives us an experimental analog to real world situations in which language rules spontaneously emerge; for example, homesign (Goldin-Meadow \& Brentari, in press), emerging sign languages (Meir et al., 2017) and early stages in spontaneous second (spoken) language acquisition by adults (Schouwstra, 2016). Accordingly, we believe that a fruitful line for future research will be an investigation of the structure of the noun phrase in these systems, providing an invaluable naturalistic complement to laboratory experiments.

## Acknowledgments

MS was funded by the British Academy.

## References

Adger, D. (2003). Core syntax. Oxford: Oxford University Press.
Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. Linguistic Inquiry, 36(3), 315-332.
Culbertson, J., \& Adger, D. (2014). Language learners privilege structured meaning over surface frequency. Proceedings of the National Academy of Sciences, 111(16), 58425847.

Cysouw, M. (2010). Dealing with diversity: Towards an explanation of NP-internal word order frequencies. Linguistic Typology, 14(2), 253-287.
Dryer, M. (2009). On the order of demonstrative, numeral, adjective and noun: An alternative to Cinque. (Talk presented at Theoretical approaches to disharmonic word orders, Newcastle University, May-June 2009.)
Dunn, M., Greenhill, S. J., Levinson, S. C., \& Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. Nature, 473(7345), 79-82.
Evans, N., \& Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and brain sciences, 32(05), 429448.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., \& Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. Psychological science, 0956797612463705.

Goldin-Meadow, S., \& Brentari, D. (in press). Gesture, sign and language: The coming of age of sign language and gesture studies. Behavioral and Brain Sciences, 1-82.
Goldin-Meadow, S., So, W. C., Özyürek, A., \& Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. PNAS, 105(27), 9163-9168.
Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), Universals of language (pp. 73113). MIT press.

Hall, M. L., Mayberry, R. I., \& Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. Cognition, 129(1), 1-17.
Kemmerer, D. (2000). Selective impairment of knowledge underlying prenominal adjective order: Evidence for the autonomy of grammatical semantics. Journal of neurolinguistics, 13(1), 57-82.
Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. Linguistics and philosophy, 30(1), 1-45.
Ladd, D. R., Roberts, S. G., \& Dediu, D. (2015). Correlational studies in typological and historical linguistics. Annu. Rev. Linguist., 1(1), 221-241.
Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., .. . Sandler, W. (2017). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. Cognition, 158, 189-207.
Partee, B. H. (1987). Noun phrase interpretation and typeshifting principles. In J. Groenendijk, D. de Jongh, \& M. Stokhof (Eds.), Studies in discourse representation theory and the theory of generalized quantifiers (p. 115-143). Dordrecht: Foris.
Piantadosi, S. T., \& Gibson, E. (2014). Quantitative standards for absolute linguistic universals. Cognitive Science, 38(4), 736-756.
Rijkhoff, J. (2004). The noun phrase. Oxford: Oxford University Press.
Schouwstra, M. (2016). Temporal structure in emerging language: From natural data to silent gesture. Cognitive Science, 41(S4), 928-940. Retrieved from http://dx.doi.org/10.1111/cogs. 12441 doi: 10.1111/cogs. 12441

Schouwstra, M., \& de Swart, H. (2014). The semantic origins of word order. Cognition, 131(3), 431-436.
Sutton-Spence, R., \& Woll, B. (1999). The linguistics of British Sign Language: an introduction. Cambridge University Press.

# Discourse continuity promotes children's learning of new objects labels 

Jessica F. Schwab (jschwab@princeton.edu)<br>Casey Lew-Williams (caseylw@princeton.edu)<br>Department of Psychology, Princeton University, Peretsman-Scully Hall<br>Princeton, NJ 08540 USA


#### Abstract

The present study examined the influence of continuity of reference (i.e., discourse continuity) on children's learning of new objects labels. Four-year-old children were taught three new label/objects pairs, where the speaker's references to objects were either continuous (i.e., clusters of utterances referred to the same object) or discontinuous (i.e., no two sequential sentences referred to the same object). In two experiments, children learned new word/object mappings more successfully when object labels were accompanied by continuous references to the same object. This research reveals how discourse cues support children's encoding of new words, and in doing so, advances our understanding of the specific features of parents' language input that facilitate children's language development.


Keywords: discourse continuity; word learning; childdirected speech

## Introduction

Children are adept at analyzing the complexities of their language input in order to learn new words, but there is also substantial variability in their learning. In order to better understand these differences, researchers have examined various features of caregivers' input shown to influence vocabulary growth, including social cues (such as eye gaze and pointing; e.g., Booth, McGregor, \& Rohlfing, 2008; Brooks \& Meltzoff, 2008), structural cues (such as repetition and utterance length; Brent \& Siskind, 2001; Lew-Williams, Pelucchi, \& Saffran, 2011; Schwab \& LewWilliams, 2016), visual cues (such as the size of labeled objects in the visual field or their perceptual salience; Pereira, Smith, \& Yu, 2014; Pruden, Hirsh-Pasek, Golinkoff, \& Hennon, 2006), and auditory cues (such as intonation and pitch; e.g., Ma, Golinkoff, Houston, \& HirshPasek, 2011; Singh, Nestor, Parikh, \& Yull, 2009). Here we focus on a contextual cue of parents' speech that may also facilitate children's vocabulary development: the content or structure of the discourse exchange. Specifically, discourse continuity, or the clustering of utterances that reference the same topic, may promote children's word learning (e.g., Frank, Tenenbaum, \& Fernald, 2013). Recent research on this topic suggests that discourse continuity does promote children's in-the-moment disambiguation of word-referent mappings in noisy referential contexts (Horowitz \& Frank, 2015), but it is not yet clear whether discourse continuity also contributes to children's encoding of new words in less ambiguous contexts, i.e., when caregivers hold and talk about an object in front of children, as is common in natural communication. Thus, the present study tests whether
discourse continuity influences the learning of multiple new object labels in 4 -year-old children.

Previous research has revealed that young children are sensitive to various aspects of the discourse context and structure. For example, 24-month-olds have been shown to understand that adults pay attention to - and talk about novel aspects of an interaction (Akhtar, Carpenter, \& Tomasello, 1996). That is, children are able to learn a new word when an adult labels an object that is novel to the discourse context from only the adult's own point of view. Relatedly, two-year-olds have been shown to use speakers' speech disfluencies to predict their intended referents during object labeling (Kidd, White, \& Aslin, 2011). Finally, crosslinguistic research has revealed that children who hear more consistent referential patterns within discourse specifically, regarding the use of either null, pronominal, or lexical verb arguments - tend to produce more consistent patterns earlier, compared to children exposed to inconsistent discourse patterns (Guerriero, Oshima-Takane \& Kuriyama, 2006). Nevertheless, there is little research to date that specifically looks at children's ability to take advantage of discourse continuity, or the idea that neighboring utterances are likely to refer to the same topic (e.g., Frank, Tenenbaum, \& Fernald, 2013; Hoff-Ginsberg, 1994; Ochs \& Shieffelin, 1983). For example, if a child simply hears, "I rode a camel!", he or she might come to the incorrect conclusion that a camel is some sort of automated vehicle. If instead the child hears, "I took a trip to the desert. I rode a camel! He was so sweet and let me pet him," he or she might use the topic continuity between camel and other words in the discourse in order to discern the meaning of camel (i.e., an animal living in the desert), as well as to encode its meaning more concretely and accurately.

Most existing research on the topic of discourse continuity and children's language learning has examined the use of discourse continuity in child-caregiver interactions (Frank, Tenenbaum, \& Fernald, 2013; Rohde \& Frank, 2014). Rohde and Frank (2014) analyzed discourse continuity in parents' interactions with their children using three different methods: raw annotations of speakers' referent, the output of a computational model, and judgments made by human coders. Across the three methods, the researchers determined that many topicsignaling cues - such as pronoun use and sentence-final reference - found in adult discourse are also present in child-directed speech. They conclude that the function of these cues in child-directed speech may be to help children acquire additional referential information from their input, particularly when individual utterances are ambiguous. Hoff
(2010) revealed that children produce topic-continuing discourse themselves, particularly during certain languagerich activities such as reading. Other work suggests that speakers' discourse continuity might be relevant for supporting a key component of children's language development: the learning of new words. Frank, Tenenbaum, \& Fernald (2013) found that caregivers’ references to objects in a child-parent play session were more continuous (or "clumpy") than would be expected by chance. Moreover, computational modeling work has shown some evidence of the importance of discourse continuity for word learning. In their word-learning model, Luong, Frank, and Johnson (2013) set speakers' intended referent to be continuous across utterances. This discourse information, combined with social cues, led to some improvements in the model's word learning performance. Together, these studies suggest that discourse continuity exists in adult-child interactions and provides helpful cues to word learning, yet they are unable to conclusively determine whether or not discourse continuity improves children's word-learning abilities.

Erika Hoff (2003) began to answer this question - of whether continuity of discourse promotes children's word learning - by looking at topic-continuing replies, i.e., caregivers' utterances that continue a topic previously introduced by the child. Hoff found that the amount that mothers used topic-continuing replies predicted their children's vocabulary growth ten weeks later, suggesting that continuity in mother-child interactions may indeed promote children's language learning. Horowitz and Frank (2015) went further by testing whether children are able to use a speaker's discourse continuity as a strategy for determining object reference in ambiguous word learning situations. In their study, children ages 2-6 years completed a novel word-learning task, where the only cue to reference was the placement of a labeling event within the discourse structure of the interaction. Specifically, children heard an object label (with no associated gestural cues to the referent) flanked by descriptions of either toy A or toy B (which were accompanied by gestural cues). If children are able to use discourse continuity as a cue to reference, they should be able to determine the object/label pairing if the labeling event occurs between two descriptions of the same object (either toy A or toy B), i.e., if the labeling episode is discourse continuous. If the labeling event occurs between two descriptions of different objects (toy A and toy B), the label/object pair should be indeterminable. The results revealed that children were in fact only able to successfully determine the referent when labels were discourse continuous. Moreover, children only started showing successful disambiguation by age $3-4$, and showed even better learning through ages 5 and 6, suggesting that children's ability to use discourse information in determining object reference might develop over the course of childhood.

Discourse continuity clearly seems to be helpful for disambiguation, i.e., determining reference in uncertain
situations. However, it has not yet been determined whether discourse continuity, in addition to helping determine an accurate word/object mapping in the moment, is also helpful for children's encoding of a new word that is clearly the focus of attention. Not only is the latter common in caregiver-child interactions (e.g., Pereira, Smith, \& Yu, 2013), but so is caregivers' tendency to refer to a string of objects in sequence. As speakers rapidly shift focus from one object to the next in conversation, it is possible that providing context for each labeling episode through topic continuity helps children successfully encode and remember new object labels. In the present paper, we test this prediction in 4 -year-olds by teaching them three new words, either with or without discourse continuity. If discourse continuity does in fact promote children's word learning, we predicted that children who heard clusters of continuous reference to objects would show better learning of object labels (defined as proportion of object/label mappings correctly identified in the test phase) compared to children who heard object references distributed over the course of the learning phase.

## Experiment 1

In Experiment 1, we tested the extent to which continuity of reference influences children's learning of three new word/object pairs. In the Continuous condition, clusters of three utterances included one labeling utterance directed toward a particular object, accompanied by two additional utterances describing - but not explicitly labeling - the same object. In the Discontinuous condition, children heard the same labels for each object and the same object-directed utterances as in the Continuous condition, but the discourse was not continuous (e.g., a label for Object A might be immediately followed by commentary about features of Object B). Each label or object-directed utterance was unambiguous, i.e., it was accompanied by the speaker gazing toward and grasping the object. At test, children were presented with a two-alternative forced choice reaching task in order to measure their knowledge of each object label. If discourse continuity does in fact promote children's word-learning abilities, children should show more successful learning of correct object/label mappings in the Continuous compared to the Discontinuous condition.

## Method

Participants Participants were 40 4-year-old children ( $M=46.41$ months, $S D=3.71$, Range $=42.1-53.63$ ). Twentythree participants were male, and all participants came from monolingual English-speaking homes. Children had no history of pervasive developmental delays. Twenty children were randomly assigned to each of two experimental conditions: a Continuous or Discontinuous condition, described in detail below. Three additional participants were tested but not included due to fussiness/refusal to cooperate ( $n=2$ ) or taking an extended break halfway through test trials ( $n=1$ )


Figure 1. Schematic depicting sample trials in the learning phase for the Continuous and Discontinuous conditions in Experiment 1. Between each trial, the speaker rested both hands in her lap and smiled at the participant.

Stimuli and Design Three novel words-gazzer, cheem, and tobu-corresponded to one of three novel objects, each characterized by a different color, texture, and shape (see Figure 1). Half of participants were exposed to one set of word/object pairings, and half were exposed to a second, counterbalanced set of pairings.

In the Continuous condition, blocks of three adjacent trials in the learning phase referred to the same object. Either the first or second trial was a labeling trial, while the other two trials provided identifying visual information about the object (e.g., "This is a gazzer. / This is small and green. / This feels really spiky."). There were two blocks of trials for each novel word/object pair. Each object was referred to six times total ( 2 x per object label).

The Discontinuous condition consisted of the same exact trials as the Continuous condition, but trials within each block of the learning phase were pseudo-randomly ordered such that no two adjacent utterances referred to the same object (see Figure 1). Thus, participants heard the same number of total references to each object and the same number of object labels as in the Continuous condition, but discourse continuity was absent.

Procedure During the experiment, an experimenter sat across from the participant at a table and told him or her, "We're going to play a game together! Just watch and pay attention because I'm going to ask you some questions about these things later. Are you ready? Here we go!"

During the learning phase, the experimenter placed all three objects in a line directly in front of her on the table (in one of two counterbalanced orders). On each of 18 learning trials (approximately four seconds in duration), the experimenter began with her hands in her lap. Then she 1) smiled at the participant, 2) looked down at an object, 3) grabbed the object, raised it slightly, and tilted it up, 4) looked back at the participant and said a labeling or objectdirected sentence about the object, 5) looked back at the object and set it back down, and 6) put her hands back in her lap. Two counterbalanced trial orders were used for each condition across participants.

The test phase began immediately after the learning phase. The experimenter removed all three objects from the table and told the participant that she was now going to ask some questions. Next, the experimenter took two objects at a time, placed them in an uncovered basket, and put the basket on the table. Without looking down at the objects, the experimenter slid the basket toward the participant. Then she asked the participant to choose one of the objects and hand it to the experimenter (e.g., "Which one is the cheem? Can you give me the cheem?"). During each test trial, the experimenter maintained eye contact with the participant. If a child initially touched more than one object, the object that was finally handed to the experimenter was recorded as his or her choice. There were 12 test trials total (four trials per object/label pairing). Two counterbalanced test orders were used across participants. Across conditions, participants saw the same pairs of two novel objects, positioned on the left and right sides of the basket.

Finally, children's vocabulary was assessed using the Peabody Picture Vocabulary Test (PPVT) (Dunn \& Dunn, 2007). The PPVT is a standardized measure to assess children's receptive vocabulary by asking them to identify familiar words from a flipbook of pictures. Children were also rewarded with stickers following the test phase and again during and after the administration of the PPVT.

## Results and Discussion

Word learning was measured in terms of the proportion of word/object pairs that children correctly identified in the test phase. A two-tailed independent samples $t$-test showed that learning was significantly greater in the Continuous condition ( $M=.88, S E=.03$ ) compared to the Discontinuous condition ( $M=.77, S E=.04 ; t(34.27)=2.05, p<.05, d=.65$; see Figure 2). Additionally, between conditions, there was no significant difference in children's mean age (Continuous: $M=46.96$ months, $S D=4.13$; Discontinuous: $M=45.86$ months, $S D=3.26 ; \quad t(36.05)=.94, \quad p=.35)$ or mean standardized PPVT score (Continuous: $M=118.63$, $S D=11.96 ; \quad$ Discontinuous: $\quad M=114.21, \quad S D=15.05$; $t(34.26)=1.0, p=.32)$. Interestingly, however, learning was significantly greater than chance for both the Continuous
$(t(19)=12.28, \quad p<.001)$ and Discontinuous conditions $(t(19)=6.20, p<.001)$, suggesting that children are able to successfully learn the novel words even without discourse continuity. However, continuity of reference does seem to provide an additional word-learning boost.

Because we tested children ranging from 3.5 to 4.5 years of age, we examined a possible interaction between age and discourse continuity on children's word learning. A $2 \times 2$ factorial Analysis of Variance (ANOVA) with age and condition (Continuous or Discontinuous) as betweensubjects factors revealed a significant main effect of condition $(F(1,36)=4.31, p<.05)$, but no significant main effect of age $(F(1,36)=2.04, p=.16)$ and no significant condition x age interaction $(F(1,36)=.80, p=.38)$. Thus, across the 3.5 - to 4.5 -year range, discourse continuity supported children's word learning equivalently.

These results suggest that discourse continuity promotes word learning for 3.5- to 4.5-year-old children. Importantly, however, the "discourse" in our experiment provided relevant visual information about each object, such as its color or texture (in a similar manner to Horowitz \& Frank, 2015). Thus, it remains unclear whether topic continuity in this experiment facilitated learning due to the informative discourse that accompanied object labels, or whether simply having continuity of object reference drove the more successful learning in the Continuous condition. In particular, we wanted to determine whether continuity of uninformative discourse - i.e., discourse that provided relatively neutral information about the objects (e.g., "This is good and neat") - would promote children's word learning in a similar manner. In Experiment 2, we sought to answer this question by replicating Experiment 1, but using uninformative instead of informative discourse.

## Experiment 2

Experiment 2 sought to replicate the results of Experiment 1 using uninformative discourse. In each condition (Continuous/Uninformative and Discontinuous/ Uninformative), object labels were the same as in Experiment 1, but accompanying discourse provided no relevant information about each object. If the relevant contextual cue boosting children's performance in this task is continuity of reference more generally - and not the pairing of object labels with important visual information in the discourse - we again expected children to show more successful learning of object/label mappings in the Continuous/Uninformative condition compared to the Discontinuous/Uninformative condition.

## Method

Participants Participants were 40 4-year-old children ( $M=46.37$ months, $S D=3.36$, Range $=42.27-53.13$ ). Sixteen participants were male, and all participants came from monolingual English-speaking homes. Children had no history of pervasive developmental delays. Twenty children were randomly assigned to one of two experimental conditions: a Continuous/Uninformative or Discontinuous/

Uninformative condition, described in detail below. Two additional participants were tested but not included due to experimenter error $(n=1)$ or being bilingual (less than $85 \%$ English exposure) ( $n=1$ ).

Stimuli and Design The words and objects used were identical to Experiment 1. The Continuous/Uninformative condition was identical to the Continuous condition from Experiment 1, except that object-directed utterances provided no identifying visual information about each object (e.g., "This is a gazzer. / This is good and neat. / This is nice and cute."). Object-directed utterances in the Continuous/Uninformative condition were matched to sentences in the Continuous condition from Experiment 1 in total number of syllables. The Discontinuous/Uninformative condition consisted of the exact same trials as the Continuous/Uninformative condition, but trials within each block of the learning phase were ordered such that no two adjacent utterances referred to the same object. Thus, as in Experiment 1, participants in this condition heard the same number of total references to each object and the same number of object labels as the Continuous conditions, but there was no discourse continuity.

Procedure The procedures for the learning phase, test phase, and administration of the PPVT in Experiment 2 were identical to the procedures in Experiment 1.

## Results and Discussion

Again, word learning was measured in terms of the proportion of word/object pairs that children correctly identified in the test phase. A two-tailed independent samples $t$-test showed that learning was significantly greater in the Continuous/Uninformative condition ( $M=.88, S E=.03$ ) compared to the Discontinuous/Uninformative condition ( $M=.72, S E=.04 ; t(34.64)=2.93, p=.006, d=.92$; see Figure 2). Additionally, between conditions, there was no significant difference in children's average age (Continuous: $M=46.27$ months, $S D=3.20$; Discontinuous: $M=46.47$ months, $S D=3.60 ; \quad t(37.50)=-.18, \quad p=.86)$ or average standardized vocabulary score (Continuous: $M=119.5$, $S D=13.13$; $\quad$ Discontinuous: $\quad M=114.72, \quad S D=12.62$; $t(35.83)=1.14, p=.26)$.

Because we tested children ranging from 3.5 to 4.5 years of age, however, we again examined a possible interaction between age and discourse continuity on children's word learning. A $2 \times 2$ factorial Analysis of Variance (ANOVA) with age and condition (Continuous or Discontinuous) as between-subjects factors revealed a significant main effect of condition $(F(1,36)=8.27, p=.007)$, but no significant main effect of age $(F(1,36)=.01, p=.92)$ and no significant condition x age interaction $(F(1,36)=.62, p=.44)$. Similar to Experiment 1, across the 3.5- to 4.5-year range, discourse continuity promoted children's word learning equivalently. Additionally, similar to Experiment 1, there was significant learning compared to chance for both conditions (Continuous: $\quad t(19)=11.83, \quad p<.001 ; \quad$ Discontinuous:
$t(19)=4.95, p<.001)$, again suggesting that continuity of reference supports word learning in the absence of discourse continuity.

Comparing across Experiment 1 and Experiment 2, there was no significant difference in learning between the Continuous and Continuous/Uninformative conditions $(t(37.95)<.01, p>.99, d<.01)$, or between the Discontinuous and Discontinuous/Uninformative conditions $(t(37.99)=-.81$, $p=.42, d=.26$ ). A $2 \times 2$ mixed analysis of variance (ANOVA) with experiment ( 1 or 2 ) as a between-subjects factor and condition (Continuous or Discontinuous) as a withinsubjects factor revealed a significant main effect of condition $\left(F(1,76)=12.47, p<.001, \eta_{p}^{2}=.14\right)$, but no significant main effect of Experiment $(F(1,76)=.44, p=.51$, $\eta_{p}^{2}=.005$ ), and no significant condition x experiment interaction $\quad\left(F(1,76)=.44, \quad p=.51, \quad \eta_{p}^{2}=.005\right)$. Thus, Experiment 2 successfully replicated the results of Experiment 1 in an uninformative discourse context. Together, these results suggests that continuity of reference generally - and not just continuity of informative discourse - seems to promote children's word learning.


Figure 2: Mean proportion object/label mappings correctly identified in Experiment 1 (Continuous vs. Discontinuous conditions) and Experiment 2 (Continuous/Uninformative vs. Discontinuous/Uninformative conditions). Error bars show +/- 1 SEM across participants.

## General Discussion

In two experiments, we show that continuity of reference promotes 4 -year-old children's learning of new object labels. Moreover, the speaker's discourse does not need to provide informative content in order to promote children's word learning - simply having continuity of reference in child-directed speech seems to be sufficient to support learning. Thus, not only does discourse continuity help children determine ambiguous word/object mappings in the moment (Horowitz \& Frank, 2015), but also, it helps children encode multiple new object labels in the context of rapidly shifting adult-child interactions.

A great deal of recent research has focused on children's ability to track statistical co-occurrences in language in order to learn word-referent mappings (e.g., Smith \& Yu,
2008), but fewer studies have focused on children's ability to use information about the structure of discourse in order to learn new words. Because children have been shown to be adept at tracking object-label regularities over time, in some contexts these kinds of contextual cues may not be necessary. More likely, however, discourse cues, in addition to socio-pragmatic cues, help children encode information about word/object co-occurrences over time, presumably by increasing their salience. Relatedly, Pereira, Smith, and Yu (2013) have suggested that there are optimal visual moments for learning new word/object pairs. That is, when objects appear in a clean, stable view in front of a child while it is being labeled, that child is more likely to learn the object's label. Here, continuity of reference may provide a similarly optimal contextual moment for learning a new word/object pair, where each word and referent are clearly linked within the discourse, allowing children to attend to their features or potential functions.

The present results are convergent with findings showing that repetition of words across neighboring utterances is helpful for learning (e.g., Onnis, Waterfall, \& Edelman, 2008; Schwab \& Lew-Williams, 2016). In particular, previous research has shown that repetition of object labels in blocks of successive utterances promotes two-year-olds' encoding of new word/object pairings. Here, at least with older preschool-age children, simply referencing one object for several sentences in a row - without repeating the object label itself - may enable the learner to better encode a word/object pairing. It is possible that repetition of object labels themselves - compared to continuity of reference more generally - promote word learning differentially along the developmental continuum. For example, previous work suggests that children's ability to exploit discourse continuity to disambiguate moments of reference increases as children age, with children under 3 years not showing the ability to take advantage of discourse cues in this context (Horowitz \& Frank, 2015). In a similar manner, the need for caregivers to repeat object labels in neighboring sentences may decrease over time as children increase their proficiency in inferring information from the discourse, i.e., become better at learning from discourse continuity. Future research should aim to directly examine differences in the influence of partial repetition and discourse continuity on children's learning across a wider age range, as well as relate children's learning abilities to differences in caregivers' naturalistic use of these cues in the home.

Finally, it is not yet clear from the present results whether children's increased learning in the Continuous conditions is a facilitation or interference effect. Specifically, it may be the case that continuity of reference promotes learning, or that discontinuity in object reference interferes with learning because of rapid shifts in attention to different objects. We are currently pursuing follow-up studies to determine whether visual continuity is sufficient to support children's word learning in this experimental context, or whether visual discontinuity interferes with learning. If children learn words similarly regardless of continuous or
discontinuous visual exposure, this would suggest that continuity in a speakers' discourse in particular seems to promote children's word learning.

Overall, the present experiments reveal that discourse continuity promotes 4 -year-old children's learning of new object labels, and this seems to be true regardless of the content or informativity of the discourse. Previous research on discourse continuity has found that natural child-directed discourse tends to be "clumpy" (Frank, Tenenbaum, \& Fernald, 2013), and continuity of discourse helps children disambiguate between possible referents in the moment (Horowitz \& Frank, 2015). The present work goes further by suggesting that clusters of reference to particular objects can help children more successfully encode new words in the context of hearing sequential label/object pairings, as speakers rapidly shift focus from one object to the next. This research has implications for our understanding of how differences in caregivers' language input can influence children's vocabulary development.

## Acknowledgments

Thank you to the participating families and to Jessica
Quinter for all of her help with testing participants and coding videos. We also thank Eva Fourakis, Carolyn Mazzei, and the rest of the Princeton Baby Lab for their administrative assistance. This work was supported by grants to C.L.W. from the National Institute of Child Health and Human Development (R03HD079779), the American Speech-Language-Hearing Foundation, and the American Hearing Research Foundation.

## References

Akhtar, Carpenter, \& Tomasello (1996). The role of discourse novelty in early word learning. Child Development, 67(2), 635-645.
Booth, A. E., McGregor, K. K., \& Rohlfing, K. J. (2008). Socio-pragmatics and attention: Contributions to gesturally guided word learning in toddlers. Language Learning and Development, 4(3), 179-202.
Brent, M. R., \& Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. Cognition, 81(2), B33-B44.
Brooks, R., \& Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. Journal of Child Language, 35(1), 207.

Dunn, L.M., \& D.M. Dunn (2007). Peabody Picture Vocabulary Test-Fourth Edition. Bloomington, MN: NCS Pearson, Inc.
Frank, M. C., Tenenbaum, J. B., \& Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. Language Learning and Development, 9(1), 1-24.
Guerriero, A. S., Oshima-Takane, Y., \& Kuriyama, Y. (2006). The development of referential choice in English
and Japanese: a discourse-pragmatic perspective. Journal of Child Language, 33(4), 823-857.
Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. Child Development, 74(5), 1368-1378.
Hoff, E. (2010). Context effects on young children's language use: The influence of conversational setting and partner. First Language, 30(3-4), 461-472.
Hoff-Ginsberg, E. (1994). Influences of mother and child on maternal talkativeness. Discourse Processes, 18(1), 105117.

Horowitz, A. C., \& Frank, M. C. (2015). Young children's developing sensitivity to discourse continuity as a cue for inferring reference. Journal of Experimental Child Psychology, 129, 84-97.
Kidd, C., White, K. S., \& Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. Developmental Science, 14(4), 925-934.
Lew-Williams, C., Pelucchi, B., \& Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. Developmental Science, 14(6), 1323-1329.
Luong, M. T., Frank, M. C., \& Johnson, M. (2013). Parsing entire discourses as very long strings: Capturing topic continuity in grounded language learning. Transactions of the Association for Computational Linguistics, 1, 315326.

Ma, W., Golinkoff, R. M., Houston, D. M., \& Hirsh-Pasek, K. (2011). Word learning in infant-and adult-directed speech. Language Learning and Development, 7(3), 185201.

Ochs, E. \& Schieffelin, B. (1995). The impact of language socialization on grammatical development. In P. Fletcher \& B. MacWhinney (Eds.), The Handbook of Child Language (pp. 73-94). Oxford: Blackwell.
Onnis, L., Waterfall, H. R., \& Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. Cognition, 109(3), 423-430.
Pereira, A. F., Smith, L. B., \& Yu, C. (2014). A bottom-up view of toddler word learning. Psychonomic Bulletin \& Review, 21(1), 178-185.
Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., \& Hennon, E.A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. Child Development, 77(2), 266-280.
Rohde, H., \& Frank, M. C. (2014). Markers of topical discourse in child-directed speech. Cognitive Science, 38(8), 1634-1661.
Schwab, J.F., \& Lew-Williams, C. (2016). Repetition across successive sentences facilitates young children's word learning. Developmental Psychology, 52(6), 879-886.
Singh, L., Nestor, S., Parikh, C., \& Yull, A. (2009). Influences of infant-directed speech on early word recognition. Infancy, 14(6), 654-666.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.

# The Facilitatory Effect of Referent Gaze on Cognitive Load in Language Processing 

## Mirjana Sekicki (mirjana@coli.uni-saarland.de)

Maria Staudte (masta@coli.uni-saarland.de)<br>Department of Language Science and Technology; Saarland Univeristy, Germany


#### Abstract

This paper considers prediction in language processing by examining the role of the visual context, and specifically, the role of speaker referent gaze on cognitive load. We inspect the anticipatory visual attention during sentence processing together with the cognitive load induced at the points of the gaze cue, and the linguistic referent. Employing a novel measurement of cognitive load - the Index of Cognitive Activity (Marshall, 2000) allowed us to simultaneously consider both anticipatory eye-movements and cognitive load. Our results show that the gaze cue is being followed, and considered as a relevant piece of information, which subsequently reduces the cognitive load on the linguistic referent. In addition, we found that considering the gaze cue is in itself not costly, unless it cues an object mismatching with the previous linguistic context.


Keywords: Gaze; Cognitive Load; Index of Cognitive Activity; Prediction; Eye-tracking

## Introduction

A series of investigations in the visual world paradigm (VWP) have shown how listeners simultaneously combine linguistic and visual cues to predict upcoming linguistic input (for a review of the VWP see Huettig, Rommers, \& Meyer, 2011). Based on the idea that prediction is a unifying principle of the human mind, a large body of psycholinguistic research, in the VWP, as well as employing EEG, has been examining the role of prediction in language processing (see Huettig, 2015; Huettig \& Mani, 2016). Anticipatory eye-movements collected in the VWP have reliably shown that people predict upcoming referents based on the previous linguistic material (e.g. Altmann \& Kamide, 1999; Kamide, Altmann, \& Haywood, 2003), as well as based on the visually presented events (Knoeferle, Crocker, Scheepers, \& Pickering, 2005).

Our present work examines the influence of the visual modality on processing linguistic information by specifically investigating speaker gaze, as an inseparable part of the visual context in situated communication, and its influence on prediction making. We hypothesized that speaker gaze to the upcoming referent helps constrain the set of possible targets and thus, by increasing the predictability of the cued object reduces the cognitive load induced by its linguistic referent. In addition, we examined whether any cost reduction on the referent would be accompanied by a cost increase on the gaze cue, effectively spreading the cognitive load across the two modalities, namely, the gaze cue and the language.

The Gaze Cue Gaze has been shown to play an important role in situated communication. Listeners inspect objects they anticipate will be mentioned next (Altmann \& Kamide, 1999), and they fixate the mentioned object 200-300 ms after the speaker started referring to it (e.g. Tanenhaus, SpiveyKnowlton, Eberhard, \& Sedivy, 1995). Speakers also fixate
the relevant object $800-1000 \mathrm{~ms}$ before mentioning it (e.g. Griffin \& Bock, 2000). But, how does speaker's gaze add to the listener's prediction?

Previous research on gaze and language has established the gaze cue to be utilized and proven helpful while processing linguistic material (e.g. Hanna \& Brennan, 2007) even when the speaker is a robot (Staudte \& Crocker, 2011) or a virtual agent (Staudte, Crocker, Heloir, \& Kipp, 2014). The conclusions about the effects of gaze following are drawn on the basis of participants' eye-movements and responses to the task at hand. The eye-movement data give insight into the shifts of visual attention indicating whether the gaze cue was considered. In addition, the effect of considering the gaze cue on the comprehension of linguistic material is assessed by reaction times, comprehension and production tasks. Importantly however, no direct effect of speaker gaze on cognitive effort required for language processing has, to our knowledge, been examined yet. This paper sets out to investigate just that by combining eye-movement data with an online pupillary measure of cognitive load, which enabled us to measure this cost / benefit directly, and independent of a specific task.

We examined both the process of creating and (dis)confirming predictions in one and the same setting. To this end we employed a novel measurement of cognitive load - the Index of Cognitive Activity (ICA; Marshall, 2000) which allows for an experimental design that combines eyemovements, shedding light on the predictive processes, while simultaneously measuring cognitive load at different stages of sentence processing. The results revealed that referent gaze is followed and that it indeed adds to the predictability of the upcoming referent such that the spoken reference induces less cognitive load. Note that whether this effect could also be induced by arrows or other visual pointers is considered irrelevant at this point.
The Index of Cognitive Activity (ICA) The two experiments presented in this paper are conducted in the VWP. We considered the eye-movements and the cognitive load reflected in pupil size. While the traditional eye-movement analysis helps reveal any patterns of anticipation of potential target objects, the ICA allowed us to simultaneously also measure cognitive load both on the gaze cue and on the referent noun, i.e. the cognitive load induced by creating and (dis)confirming one's predictions.

Pupil dilation happens in consequence of either changes in light or cognitive activity. Two groups of muscles are responsible for pupil size: circular muscles that make the pupil contract, and radial muscles that make it dilate. Pupil size
changes due to light and due to cognitive activity employ different activation and inhibition processes, the dilation due to cognitive load being shorter and more abrupt (Beatty, 1982).

The ICA measurement disentangles the two types of change in pupil size by performing a wavelet analysis on the pupil dilation record and removing the large oscillations, while considering only the quick pupil jitter that is related to cognitive activity (Marshall, 2000). Such events of small abrupt changes in pupil size are referred to as the ICA events.

The ICA has been shown to reflect changes in cognitive load in a variety of different studies since its appearance (e.g. Marshall, 2002, 2007). However, only recently has it been examined with cognitive load induced by linguistic processing (Demberg \& Sayeed, 2016). Demberg and Sayeed present a series of seven experiments showing that the ICA indexes linguistic processing difficulty for both reading and auditory presentation of linguistic stimulus. In addition, the ICA proved to be robust with respect to eye-movements making it a valid measure of processing difficulty in the VWP.

Hence, employing the ICA in present experiments allowed for simultaneous assessment of both visual attention and cognitive load.

Current Questions and Predictions Two studies were set out to examine whether gaze is considered as part of the context determining the predictability of the subsequent referent. We examined if the gaze cue actually helps reduce the cognitive load of a linguistic referent online and whether cognitive load is in fact spread across gaze and spoken reference such that the gaze cue itself then induces higher cognitive load.

Experiment 1 made use of the gaze cue that was always fitting (the previous linguistic context) and congruent (cuing the object to be referred to linguistically). We manipulated the existence of the gaze cue in order to answer the following research questions:
a) Does the gaze cue influence the predictability of a linguistic referent?
b) If so, how does it influence the cognitive load induced by the referent?
c) Can we measure cognitive load on the gaze cue itself?

Experiment 2 also made use of congruent gaze, while manipulating the fit of the referent (thus, also the fit of the gaze cue) with the previous linguistic context. This was done in order to answer the following research questions:
a) Does the gaze cue help reduce the cognitive load on the linguistic referent even when they both do not fit the previous linguistic context?
b) Does the gaze cue to a mismatching object itself induce higher cognitive load?

We expected mismatching gaze to be surprising and thus, more costly, which would as its consequence have a reduction in cognitive load on the corresponding linguistic referent.

## Experiment 1

This study aimed to examine whether the online measure of cognitive load also supports previous findings that the gaze cue


Figure 1: A trial timeline example (from Exp.1) - referent gaze condition (left); and no-gaze condition (right).
is actively considered in language processing, by quantifying how its existence modifies the cost induced by the linguistic referent. In addition, we were interested in measuring the potential cost of gaze perception.

## Method

The study made use of $2 \times 2 \times 2$ mixed factorial design. The independent variable Gaze was a between subjects variable, i.e. half the participants were presented with the version of the experiment where all items included the gaze cue, while the rest saw the version with items never having the gaze cue. Fillers balanced the gaze conditions to the ratio of $1: 1$. In addition, four linguistic conditions were created with the two within subjects variables, Constraint and Plausibility. Constraint was manipulated by verb restrictiveness (spill vs. order), and Plausibility by noun fit (spill: water vs. ice-cream).
Participants 64 students of Saarland University took part in this study ( 45 women) and were monetarily reimbursed for their partaking. Their age ranged from 18 to 34 years old ( $M$ $=24.16$ ). Participants were all native speakers of the German language with normal or corrected to normal vision.

Materials and Design Each participant was presented with 20 items and 30 fillers, both consisting of visual and auditorily presented linguistic stimulus. In addition, visual displays included a face-like object forming the gaze cue.

Note that the gaze cue used in our studies, since being an always congruent visual pointer, is arguably not different from an arrow. This is true, but, currently irrelevant, since the differences between a visual pointer and a gaze cue are potentially to be expected in cases of manipulated congruence.

We made use of simple German sentences (Subject - Verb - Adverb - Object) that included a restrictive (spill) and a nonrestrictive (order) verb and two object nouns (water vs. icecream) of differing semantic fit in relation to the restrictive verb. The chosen nouns were controlled for frequency and two pretests have shown that in the context of order, water was more predicted (cloze probability of $13.67 \%$; plausibility rating of 1.12 on a 7 -point Likert scale ${ }^{1}$ ), than ice-cream (cloze

[^482]

Figure 2: (Exp.1) Proportion of fixations in the four linguistic conditions without the gaze cue (left) and with ref. gaze (right). The verb onset, gaze and noun onset are shown averaged across trials, and aligned to the 100 ms bins within which they fall.
probability of $0.16 \%$; plausibility rating of 2.76 ). The same adverb, neutral in meaning (gleich - ENG: soon), was used for all items and served the purpose of a spillover region.

In addition, visual displays with four concrete objects ${ }^{2}$ were presented. Two of the four objects fit the category introduced by the restrictive verb (spill: water, ice-cream), while all four fit the non-restrictive verb (order: water, ice-cream, suitcase, coat). The referent noun was always fitting with the previous linguistic context and the gaze cue was always congruent, that is, cuing the object that is about to be mentioned. The main manipulation of the study was the presence (vs. absence) of the gaze cue which was presented before the target object was referred to verbally.

Figure 1 illustrates a trial timeline. The visual scene with open eyes was presented 1000 ms prior to sentence onset. The gaze cue (or closed eyes) was introduced 300 ms after the verb, i.e. from adverb onset to sentence end. Finally, the eyes would look straight for another 1000 ms .

Fillers Fillers included the same visual setting, but differed in the structure and complexity of the linguistic stimulus and the number of objects that fit the verb category. 30 fillers were used, 25 of which had the opposite and 5 the same gaze condition as the items (ratio of 1:1 for gaze and no-gaze). 19 fillers were followed by simple yes/no comprehension questions that were answered on a key-press. The questions were related exclusively to the linguistic content. This was done in order not to inspire extensive inspection of the visual scene, but rather so that participants consider it only optionally and freely in addition to the linguistic information.

Procedure An EyeLink II head-mounted eye-tracker (SR Research, Ltd; Mississauga, Ont., Canada) was used to track

[^483]both eyes at a sampling rate of $250 \mathrm{~Hz} .{ }^{3}$ Participants were instructed to listen carefully to the sentences while looking freely at the presented objects. They would advance the experiment on a button press after each trial. Other two buttons were used to answer the comprehension questions. The experimental session was preceded by a three-trial practice session. The experiment lasted for approximately 15 minutes.

## Results

First, in order to gain insight into the patterns of prediction and visual attention we consider the proportion of fixations to the presented objects throughout a trial. Second, we analyse new inspections. Consecutive fixations to the same interest area are considered as one inspection. Since we are interested in the shift of attention inspired by a relevant stimulus, we analyse new inspections, i.e. the first inspection to an interest area that started after the linguistic or visual point of interest (as done in e.g. Staudte \& Crocker, 2011). We consider new inspections from verb onset (showing linguistic predictions) and from gaze cue onset (showing if the gaze cue influenced visual attention). Finally, the ICA events are extracted from the pupil jitter, summed over a duration of a relevant time-window and statistically analysed. For the ICA analysis, we considered the gaze time-window and the referent noun window.

Variable Coding and Data Analysis In their VWP experiment, Demberg and Sayeed (2016) establish a time-window taken 600-1200 ms from the onset of the critical word to be an appropriate window size and timing for the analysis of the ICA events. Since our critical words differ in length across items, we correct this potential confound by taking a time-window that starts from the middle of a word ${ }^{4}$, and con-

[^484]sider the following 600 ms . In addition, we analyse the gaze window: 600 ms from the gaze cue onset.

The ICA events are extracted for both eyes separately. Since there is no clear theoretical reason why differences should be expected for the two eyes, we combine the two datasets by summing the ICA events for corresponding time-windows and conduct the analyses on the combined data.

All independent variables were contrast coded for the statistical analysis. New inspections, a binary dependent variable required the use of generalized mixed effects models of binomial type. On the other hand, the analysis of the ICA, a count variable, required the use of generalized mixed effects models with Poisson distribution. All models included a maximal converging random structure for both Item and Subject. The analyses were conducted in R programming environment ( R Core Team, 2013) and using the lme4 package.

Proportion of Fixations Figure 2 illustrates the proportions of fixations to all presented objects during a trial. The first dashed line presents verb onset; second line - gaze onset (not relevant for no-gaze); the third line - referent noun onset. It is apparent that the restrictive verb (spill) shifts the focus of visual attention to one particular object (water). The less fitting object (ice-cream) is considered only upon being referred to linguistically (no-gaze), or earlier, at the point of the gaze cue (referent gaze), confirming that the visual attention is not only influenced by linguistic content but also by the gaze cue.
New Inspections We conducted a statistical analysis of new inspections to an object (water, ice-cream, distractors) from verb onset (to verb offset). In addition, we consider the new inspections to both water and ice-cream, from gaze onset (to adverb offset).

Considering the verb window, three identical models were run for the inspections to the three relevant objects. ${ }^{5}$ Looks to water: A main effect of Constraint ( $\beta=-0.361, S E=0.124$, $z=-2.904, p=0.004$ ) suggests that more new inspections to water occurred upon hearing spill (vs. order). Looks to icecream: No effect of Constraint ( $p=0.406$ ) suggests ice-cream was looked at with no significant difference in the contexts of both verbs. Looks to the distractors: A marginal effect of Constraint ( $\beta=0.153, S E=0.081, z=1.89, p=0.059$ ) suggests that there were somewhat more new inspections to the two distractors in the non-restrictive context of order.

Considering the gaze window, we analysed new inspections to both water and ice-cream together as TargetInspections and considered the effects of gaze on the looks to these two objects. ${ }^{6}$ We find a main effect of Gaze $(\beta=0.427, S E=$ $0.141, z=3.022, p=0.003$ ) confirming that the objects were more readily looked at with the gaze cueing them compared to the absence of gaze, i.e. that the gaze cue caused an immediate shift in visual attention.

[^485]

Figure 3: (Exp. 1) ICA events at the four time-windows of a sentence presented for the no-gaze (above) and the ref. gaze (below) conditions separately. (95\% CI error bars)

The Index of Cognitive Activity The analysis of the gaze time window did not yield significant results.

Considering the referent noun window ${ }^{7}$, we found a main effect of Gaze ( $\beta=-0.116, S E=0.051, z=-2.26, p=0.024$ ), suggesting that the presence of the gaze cue led to the reduction of cognitive load on the subsequent referent. Moreover, we found a significant Constraint:Plausibility interaction ( $\beta=-$ $0.184, S E=0.042, z=-4.37, p<0.001$ ), as well as a main effect of Plausibility ( $\beta=0.057, S E=0.029, z=2.01, p=$ 0.045). Further comparisons show a main effect of Plausibility in the subset of spill ( $\beta=0.152, S E=0.035, z=4.37, p$ $<0.001$ ), suggesting that spill water induced less cognitive load than spill ice-cream. No such effect was found in the non-constraining subset ( $p=0.249$ ), suggesting no difference between order water and order ice-cream. Figure 3 illustrates these findings (note: Adverb - gaze window; Object - referent window). Finally, to rule out an effect of experiment part found in the second study, experiment Half was included in the fixed effects structure. We found no Half:Gaze interaction, but a main effect of experiment Half ( $\beta=-0.047, S E=0.013$, $z=-3.57, p<0.001$ ), since the second part of the experiment induced less cognitive load.

## Discussion

The results show that the gaze cue inspired fixations to the cued object even when it was not predicted by the verb. Moreover, the presence of gaze led to the reduction of cognitive load on the referent noun in all conditions, while preserving the preference for the item that best matched the verb. Interestingly,

[^486]the existence of the gaze cue did not in itself induce additional cost on cognitive load.

Hence, we saw that the gaze cue influences predictability of the linguistic referent and subsequently reduces the cognitive load it induces. Interestingly, on the cue itself, no differences in cognitive load were induced either by its mere existence, or by whether it was cuing an object already anticipated based on the linguistic context.

## Experiment 2

The second study aimed at examining, firstly, whether the gaze cue helps reduce cognitive load on the linguistic referent even when they are both mismatching with the previous linguistic context, and, secondly, whether the cue to such an object is in itself more costly, since unexpected.

## Method

Participants 36 students of Saarland University ( 23 female) took part in the study and were monetarily reimbursed for their partaking. ${ }^{8}$ Their age ranged from 18 to 34 years $(M=23.36)$. Two students were excluded from the analysis due to technical issues, and two because their mother tongue was established to be Luxemburgish. Thus, 32 participants, German native speakers, were included in the analysis.

Materials and Design This experiment made use of $2 \times 2$ experimental design, combining Gaze (no-gaze/referent gaze) and referent noun Fit (fitting/mismatching). Only restrictive verbs were used (spill), combined with either a thematically fitting (water) or mismatching referent noun (sausage). 20 items were created, half of which were anomalous ${ }^{9}$. Note that the gaze cue was, again, always congruent (cuing the object subsequently referred to linguistically). When the referent noun did not fit the previous linguistic context, that made the gaze cue to the object in question mismatching as well. The same procedure was implemented as in Experiment 1.

Fillers The experiment included presenting 75 trials in total, 55 of which were fillers. $20 \%$ of the total number of sentences were anomalous ( 10 items, 5 fillers). The gaze cue was present in $2 / 3$ of all trials ( 10 items, 40 fillers). Only $16 \%$ of all trials included an anomalous gaze cue, i.e. gaze that was cueing a mismatching object ( 5 items, 3 fillers).

## Results

The same measures and analyses were conducted as in Experiment 1 , except for the new inspections analysis where only the gaze window was considered, due to the differing experimental design of Experiment 2.

Proportion of Fixations Figure 4 shows the proportion of fixations to all presented objects during a trial. As previously, the first dashed line presents verb onset; the second line -

[^487]

Figure 4: (Exp. 2) Proportion of fixations to presented objects in the four experimental conditions.
gaze onset (not relevant for no-gaze conditions); and the third line - referent noun onset. We see that the verb shifts the focus of visual attention to one particular object (water). The mismatching object (sausage) is considered only upon being referred to linguistically (no gaze), or earlier, at the point of the gaze cue (referent gaze). Thus, we observe the same pattern as in Experiment 1, namely, of the gaze cue shifting visual attention, on a par with the linguistic information.
New Inspections Considering the gaze window we combined the new inspections to both water and sausage together as TargetInspections and examined the effect of gaze on the looks to these two objects. ${ }^{10}$ A main effect of Gaze ( $\beta=$ $0.381, S E=0.191, z=1.995, p=0.046$ ) confirmed that the gaze was followed, as found in Experiment 1.

The Index of Cognitive Activity We first analysed the referent window ${ }^{11}$ and found a main effect of Fit ( $\beta=0.223$, SE $=0.043, z=5.21, p<0.001$ ), suggesting that the anomalous spill sausage required more cognitive load than spill water. Considering the effect of the gaze cue on the cost of the referent, a significant Gaze:Half interaction was observed ( $\beta=$ $-0.126, S E=0.062, z=-2.05, p=0.040$ ). Further analysis showed a marginal main effect of Gaze in the second half of the experiment $(\beta=-0.091, S E=0.047, z=-1.93, p=0.054)$, suggesting that the referent gaze reduced the cognitive load on the referent noun in both linguistic conditions. No such effect was found in the first half of the experiment ( $p=0.392$ ).

Since gaze affected referent processing in each experimental half differently, we considered cognitive load on the cue itself (gaze window) for each half separately. ${ }^{12}$ The first part of the experiment revealed a Gaze:Fit interaction $(\beta=0.179, S E=$

[^488]

Figure 5: (Exp.2) ICA events at the four time-windows of a sentence in the first (above) and the second (below) half of the experiment. ( $95 \%$ CI error bars)
$0.084, z=2.13, p=0.033)$. In the second half, a main effect of Gaze ( $\beta=0.099, S E=0.045, z=2.20, p=0.028$ ) suggests that the referent gaze induced higher cognitive load than the nogaze condition. The results are illustrated in Figure 5 (note: Adverb - gaze window; Object - referent window).

## Discussion

The eye-movements data showed evidence of gaze following even when it was unpredicted, or worse, also mismatching. Regarding the cost of processing the referent noun, initially, the existence of gaze did not have an effect; but the cognitive load induced by the mismatching cue itself was higher than that induced by both fitting gaze and no gaze cue. However, in the second half of the experiment, cognitive load on the referent noun was marginally reduced due to the helpful gaze cue; while the cue itself (to both fitting and mismatching object) now induced higher cognitive load. This suggests that participants gradually adapted to and started relying on the surprising gaze cue (increasing load on the cue) and started making use of its informativity (lowering load on the noun).

## Conclusions

Referent gaze is actively considered in the process of prediction making, shifting the visual attention to the cued object, and leading to the reduction of cognitive load on its linguistic referent. This holds even when the cue (and the corresponding referent) is mismatching with the verb. Gaze perception proved not to be costly unless mismatching with the verb.

Both studies included conditions with unpredicted but congruent gaze cue (Exp.1: order water, order ice-cream; Exp.2: spill sausage). Such a condition induced a higher processing cost on the gaze cue in Exp. 2, but not in Exp. 1. We argue that in Exp. 1 gaze is still processed naturally, as a gaze cue, due to its overall fit, while in Exp. 2, due to the mismatch with the verb, the cue became more salient, treated as a visual
pointer, regardless if cuing a fitting or mismatching object. We interpret this as evidence against a spread of cognitive load between gaze and linguistic reference.

In sum, the gaze cue is exploited to predict the upcoming referent such that it can be processed with less effort. If verb selectional features direct visual attention to a particular object and the gaze cue (alternatively) introduces a different object, this creates a shift in visual attention but, does not negate the existing preference. Cognitive load is reduced as an effect of the gaze cue, but only when the cue is established as informative and reliable, and regardless of its contextual fit.

## Acknowledgments

The reported research is funded by the Deutsche Forschungsgemeinschaft (DFG) under grant SFB1102, and by the Cluster of Excellence Multimodal Computing and Interaction (MMCI). We thank Christine Ankener, Laura Fraedrich and Luigi Palumbo for help and support throughout the process of working towards this paper.

## References

Altmann, G., \& Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73, 247-264.
Beatty, J. (1982). Task-evoked pupilary responses, processing load, and the structure of processing resources. Psychological bulletin, 91(2), 276.
Demberg, V., \& Sayeed, A. (2016). The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. PLoS ONE, 11, e0146194.
Griffin, Z. M., \& Bock, K. (2000). What the eyes say about speaking. Psychological Science, 11, 274-279.
Hanna, J., \& Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. Journal of Memory and Language, 57, 596-615.
Huettig, F. (2015). Four central questions about prediction in language processing. Brain Research, 1626, 118-135.
Huettig, F., \& Mani, N. (2016). Is prediction necessary to understand language? probably not. Language, Cognition and Neuroscience, 31(1), 19-31.
Huettig, F., Rommers, J., \& Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. Acta Psychologica, 137, 151-171.
Kamide, Y., Altmann, G., \& Haywood, S. (2003). Prediction and thematic information in incremental sentence processing: evidence from anticipatory eye movements. J. Mem. Lang., 49, 133-156.
Knoeferle, P., Crocker, M., Scheepers, C., \& Pickering, M. (2005). The influence of immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. Cognition, 95(1), 95-127.
Marshall, S. P. (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. US Patent, 6,090,051.
Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In proceedings of the 7th conference on human factors and power plants (pp. 7-5-7-9). IEEE.
Marshall, S. P. (2007). Identifying cognitive state from eye metrics. Aviation, space, and environmental medicine, 78, B165-B175.
R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/
Staudte, M., \& Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. Cognition, 120, 268-291.
Staudte, M., Crocker, M. W., Heloir, A., \& Kipp, M. (2014). The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts. Cognition, 133, 317-328.
Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., \& Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.

# Gestural Hesitation Reveals Children's Competence on Multimodal Communication: Emergence of Disguised Adaptor 

Kazuki Sekine (kazuki.sekine@mpi.nl)<br>Radboud University/ Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6526 XD Nijmegen, the Netherlands


#### Abstract

Speakers sometimes modify their gestures during the process of production into adaptors such as hair touching or eye scratching. Such disguised adaptors are evidence that the speaker can monitor their gestures. In this study, we investigated when and how disguised adaptors are first produced by children. Sixty elementary school children participated in this study. There were ten from each school year (from 7 to 12 years of age). They were instructed to remember a cartoon and retell its story to their parents. The results showed that children did not produce disguised adaptors until the age of 8. The disguised adaptors accompany fluent speech until the children are 10 years old and accompany dysfluent speech until they reach 11 or 12 years of age. These results suggest that children start to monitor their gestures when they are 9 or 10 years old. Cultural influences and cognitive changes were considered as factors to influence emergence of disguised adaptors.


Keywords: co-speech gestures; disguised adaptors; elementary school children; speech dysfluency.

## Introduction

Researchers have examined the development of gestures in children in terms of when the frequency or repertoire increases and how the relationship between a gesture and speech changes with age. The present study focused on spontaneous suppression of gesture production during speech. Analyzing when and how children try to not produce gestures would provide insight into when children become aware that their gestures are socially communicative.

The present study focuses on co-speech gestures (hereafter simply referred to as 'gestures') that spontaneously co-occur during speech and that have no standard of well-formedness, unlike sign language, but are created idiosyncratically on the fly (McNeill, 1992). A gesture typically has three phases: preparation, stroke, and retraction. In the preparation phase, the hand moves from a position of rest. The stroke phase is the central part of a gesture that conveys substantial information. The relevant meaning represented by this phase is usually expressed in the concurrent speech. Sometimes, a hold phase, where the hand is held in mid-air at the same position, occurs before and/or after the stroke phase.

As children prefer to use speech as a means of communication, the frequencies of gestures that appear around the first word and are used alone without speech, such as deictic gestures and symbolic gestures, decreases (Volterra \& Iverson, 1995). In their place, gestures that cooccur with meaningful words, called "co-speech gestures", appear around the period. Goldin-Meadow \& Butcher
(2003) observed that gestures begin synchronizing with speech both semantically and temporally in the transitional phase to the two-word speech period, at about 18 months of age. Given that gestures are not often used solely but are coproduced with speech, gestures and speech seem to form an integrated system during this period (Goldin-Meadow \& Butcher, 2003).
Previous research has shown that speech and co-speech gesture develop hand-in-hand even after two-word period. Mayberry \& Nicoladis (2000) observed that longitudinally bilingual children between 2- and 3-and-half years old produce more gestures when they speak either language that allows them produce a longer utterance, as measured by the mean number of morphemes (Mayberry \& Nicoladis, 2000). They concluded that gestures are closely related to morphosyntax level. This conclusion is indirectly supported by Fujii's (1999) study showing that the frequency of gestures does not correlate with vocabulary in the preschool period. McNeill (1992) observed that by the end of the preschool period, the frequency of gestures ascends to near adult levels. Once children start having formal education, they gradually develop the ability to create a coherent narrative by using language devices such as anaphora expressions, substitutions, ellipses and connectives (Wigglesworth, 1990). Research has shown that as children acquire spoken referential expressions for making coherent discourse, they also use gestures to mark introduced or maintained referents in the narrative (McNeill, 19992), and the number of gestures consistently increases during the elementary school period (Colletta et al., 2014; Sekine \& Furuyama, 2010). Thus, these previous studies indicate that gesture and speech develop hand-in-hand across the development. However, it is not clear if children start monitoring their gestures or become aware of gestures as information resources that their listeners can make use of.
Studies on self-repair in speech have asserted that to correct one's speech, the speaker has to monitor his or her speech process continuously, and thus, correction of speech errors is considered to reflect the speaker's ability to monitor speech (Karmiloff-Smith, 1986). Based on this assumption, it seems that determining when children start to correct their gestures would provide insight into understanding when they start monitoring their gestures. However, unlike speech in which the message is delivered by aligning linearly linguistic elements that exist independently, a gesture conveys a meaning globally at once, and any decomposition into elements is dependent on the whole (McNeill, 1992). Because of the linearity in speech, it is easier to understand where speech errors occurred and how the speaker corrected them than errors in
gestures. In contrast, gestures are mostly continuous and some of their parameters change simultaneously. This makes it difficult to determine whether and where a speaker has corrected a gesture. Considering these differences in semiotic characteristics, this study focused on a specific type of gestural correction; i.e., disguised adaptors.

Adaptors are movements that help persons to satisfy personal needs, manage emotions, and adapt to their environment such as touching one's hair or adjusting one's glasses (Ekman \& Friesen, 1969). At times it is observed that a speaker stops making a gesture in the middle of production and switches it to an adaptor to hide the gesture. Such movements may be able to say socially preferable in the situation. In this study, this kind of behavior is termed a disguised adaptor, which is defined as a gesture that is altered into an adaptor before or during the stroke phase of the gesture.

It can be observed in daily conversation that when a speaker is asked a question by the listener in the middle of her gesture and speech production, she stops producing the gesture, and puts her hand on her head or eye to scratch as if it feels itchy, like shown in Figure 1. Figure 1 indicates a scene in which the speaker on the right was retelling an episode of cartoon that she had watched to the listener on the left ((1) in Figure 1). When she was describing the figure of the drainpipe with a gesture, the listener started asking her whether one character was in the birdcage (2). At that moment, she stopped her gesture in the middle, and put her right hand on her head and scratched it until the listener finished the question (3).


Speaker (on the right side):
[soko made (1)ikitakutte, sono mado no (2)tokoro nikou'(he) (1)wants to go there, and (2)at the place where the window is, like this-'
Listener (on the left side):
[torikago (3)no naka ni haitteiruno?]
'is (he) (3)inside the cage?'
Figure 1: Halt of a gesture by listener's question.
Here, and in subsequent examples, the square brackets represent the start and end points of the motion of the speaker's hands, boldface marks the stroke phrase of the gesture phrase, underline indicates a motionless hold phase, and double underline represents the duration in which a disguised adaptor, such as touching the body or clothes. ' $\%$ ' indicates a smacking sound, '*' represents self interruption, and ' $\because$ ' in speech indicates an elongated phonation. The numbers used in figure, correspond to the numbers in the transcription, which in turn indicate the places where
gestures occurred. In the transcription, the first chunk is the original Japanese speech and the second chunk is the English translation.

Similar behaviors to the one in Figure 1 can be observed in other situations, for example when one raises his hand to catch a taxi, but misses it, or in which one is waving to her friend, and then quickly becomes aware of mistaken identity. In these situations, they often switch the hand movements to self-contact behaviors such as scratching head or eyes as if it is meant to be. Interestingly, literature on Tourette's syndrome, a chronic neurological disorder characterized by multiple involuntary movements and uncontrollable vocalizations called tics, has documented the correction of tic movements made by patients. Sottofattori and Nicolai (2007) observed that the patients can modify tics or odd movements into other movements like gestures in a natural conversation. For example, when a tic affects the right or left arm, the patient tries to move the forefinger straight with the tic as if it were a deictic gesture. All these cases suggest that when we are aware of mistake of hand movement or when we are interrupted in executing our hand movements, we often change the hand movement to a more socially acceptable movement such as scratching a body part or producing gestures.

In the light of development of gestures in children, it is important to examine whether typically developing children also are able to modify body movement into a more socially acceptable one. Because if children perform this kind of correction, it implies that they can monitor their gestures. If they already know that gestures can be seen as information resource by their listeners and they can detect errors in their gestures or speech, they may try to modify their gestures into more socially acceptable movement, such as scratching the neck.

Suppose that gesture and speech interacts in the production process (McNeill, 1992), and then gestural correction may affect or be affected by speech errors or dysfluencies. Thus, the relationship between disguised adaptors and speech errors should also be investigated. Karmiloff-Smith (1986) found that the percentage of speech repairs denoting sensitivity to the linguistic system, such as the determiner functions of articles, adjectives and possessives, increased during the elementary school period. Given these findings, it is predicated that the number of disguised adaptors also increases during this period as children acquire an ability to monitor their expressions. In addition, if disguised adaptors are due to difficulties in retrieving words, they would co-occur with filled pauses, unfilled pauses, or speech errors, rather than with intact linguistic elements, because previous studies have shown that adaptors tend to occur with speech dysfluencies while the speaker is retrieving words (Fujii, 1997; Pine, Bird, \& Kirk, 2007). However, if a disguised adaptor does not cooccur with speech dysfluencies, it might occur influenced by other factors such as cultural or cognitive factors. For instance, cognitive capacity or cultural pressure to produce or inhibit gestures may control the emergence of disguised
adaptors. Under this hypothesis, this study examined when disguised adaptors emerge during the elementary school period and how this emergence is related to speech.

## Method

## Participants

Sixty elementary school children and their parents participated in this study. There were ten children from each grade, 1 st to 6 th grade. Half were boys. In this study, each grade is referred to by their mean age, 7 - to 12 -years-olds (7-year-olds, $\mathrm{M}=7 ; 0$, Range 6;9-7;4. 8 -year-olds, $\mathrm{M}=7 ; 11$, range $7 ; 9-8 ; 10,9$-year-olds. $\mathrm{M}=9 ; 4, \mathrm{R}=9 ; 0-9 ; 6,10$-yearolds, $\mathrm{M}=10 ; 0, \mathrm{R}=9 ; 7-10 ; 5$, 11-year-olds, $\mathrm{M}=11 ; 4, \mathrm{R}=10 ; 9$ $11 ; 11, \quad 12$-year-olds, $\mathrm{M}=12 ; 0, \mathrm{R}=11 ; 5-12 ; 10$ ). All the participants were native monolingual Japanese speakers from middle-class families, and the children attended public or private elementary schools in Tokyo, Japan.

## Material and Apparatus

Each child watched a seven-minute animated color cartoon of the Tweety and Sylvester series, titled 'Canary Row' (Warner Brothers, Inc.). This cartoon was displayed on a 14-inch color computer monitor (Panasonic CF-F8). A miniDV camcorder (Sony HDR-HC9) was used to record the children's gesture and speech.

## Procedure

The experiments were conducted in a quiet room in the participant's home or a local community center. Each child was instructed to remember the cartoon story shown on the computer monitor and retell it to his or her parent as a listener in as much detail as possible. The parent was allowed to respond to the child by nodding their head or by using back channels during the child's narration. The whole session was recorded using the mini-DV camcorder on a tripod.

## Coding

Speech data. All narratives were verbatim transcribed. From the transcriptions, the mean number of clauses was then calculated. A clause was loosely defined as a combination of a noun phrase and a verb phrase). The mean number of unfilled pauses, which was defined as periods of silence longer than 200 msec . (Beattie, 1983), filled pauses, such as 'unttoo' (umm) or 'eeetto' (ehhh), and speech errors, including repetitions, replacements or false starts, were measured in order to ascertain the relationship between speech fluency and production of disguised adaptors.
Gesture data. First, co-speech gestures were identified. Hand movements were classified as gestures only when they had an identifiable beginning and a clear end and were synchronized with speech. After identifying which movements were gestures, the total number of gestures was counted. Next, disguised adaptors were identified. We coded a hand movement as a disguised adaptor if two criteria were met; 1) A gesture was altered into adaptor
before or during the stroke phase without any pause, and 2) the direction of the gesture suddenly changed when it is altered into adaptor. To analyze the temporal relationship between disguised adaptors and speech fluency, their combinations were categorized into the following six types: DA (disguised adaptor), FP (filled, pause), IS (intact speech that was completely articulated, such as Noun or Verb phrases), PR (the preparation phase of a gesture), SE (speech error), and UP (unfilled pause). The orders were as follows:

## 1. PR $\rightarrow$ IS + DA

After the preparation phase started, a disguised adaptor occurred with intact speech.
2. $\mathrm{PR} \rightarrow \mathrm{FP}+\mathrm{DA}$

After the preparation phase started, a disguised adaptor occurred with a filled pause.
3. PR $\rightarrow$ UP + DA

After the preparation phase started, a disguised adaptor occurred with an unfilled pause.
4. $\mathrm{PR} \rightarrow \mathrm{FS}+\mathrm{DA}$

After the preparation phase started, a disguised adaptor occurred with a filled pause.
5. PR $\rightarrow$ FS (FS) $\rightarrow$ IS + DA

After the preparation phase started, a false start or a number of false starts occurred. Then, a disguised adaptor occurred with intact speech.
6. $\mathrm{PR} \rightarrow \mathrm{FP}$ and/or FS $\rightarrow \mathrm{FP}+\mathrm{DA}$

After the preparation phase started, a false start and/or filled pause occurred. Then, a disguised adaptor occurred with a filled pause.

## Results

Number of gestures, clauses, pauses, and speech errors To calculate the frequency of gestures and the proportion of unfilled pauses for each age, the total number of gestures and the total amount of time spent for unfilled pauses were divided by the total speaking time. To calculate the frequency of filled pauses and speech errors, the total number of filled pauses and speech errors were each divided by the mean number of clauses (Table 1). After performing an angular transformation on the proportion of unfilled pauses in the speaking time, an ANOVA was conducted on each index (Table 1). A main effect of age group was found for the gesture frequency, $F(5,54)=3.86, p=.005$, the proportion of unfilled pauses in the speaking time, $F(5,54)$ $=2.44 p=.046$, and the proportion of speech errors in clauses, $F(5,54)=2.49, p=.042$. A post hoc comparison (Tukey, $p<.05$ ) showed that 12 -year-olds produced gestures more frequently than $7-, 9$ - and 10 -year-olds did and that the proportion of unfilled pauses during the speaking time was significantly greater for 7-year-olds ( $47 \%$ ) than for 12 -year-olds ( $33 \%$ ). There was no significant age-group difference in the total number of clauses or frequency of speech errors. These results indicate that the proportion of unfilled pauses gradually decreases during the elementary school period, whereas the frequencies of
gestures and speech errors increase in the late elementary school period.

Table 1: Number of Clauses, Gestures, Pauses, and Speech Errors.

| Age | 7 | 8 | 9 | 10 | 11 | 12 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| -Frequency of gestures | $0.12(0.06)$ | $0.14(0.09)$ | $0.10(0.09)$ | $0.09(0.06)$ | $0.19(0.13)$ | $0.27(0.17)$ |
| per second | $62.4(29.93)$ | $69(20.42)$ | $77.8(24.68)$ | $77.3(15.58)$ | $91.9(20.90)$ | $78.6(19.47)$ |
| -Total number of clauses | $6.47(0.19)$ | $0.40(0.11)$ | $0.38(0.08)$ | $0.34(0.06)$ | $0.36(0.08)$ | $0.33(0.08)$ |
| -Proportion of unfilled | $0.38(0.25)$ | $0.29(0.17)$ | $0.26(0.07)$ | $0.39(0.31)$ | $0.31(0.18)$ | $0.25(0.10)$ |
| pauses in speaking time | $0.20(0.10)$ | $0.21(0.12)$ | $0.38(0.19)$ | $0.28(0.11)$ | $0.36(0.20)$ |  |
| -Filled pauses per clause | $0.3)$ |  |  |  |  |  |
| -Speech errors per clause | $0.26(0.15)$ |  |  |  |  |  |

## Number of children who produced a disguised adaptor

In total, 22 disguised adaptors were observed. The absolute number of disguised adaptors produced by each age group was three times for 9 -year-olds, three times for 10 -yearolds, twelve times for 11-year-olds, and four times for 12-year-olds. Three 9 -year-olds, three 10 -year-olds, five 11 -year-olds, and two 12 -year-olds produced a disguised adaptor at least once during their narrations. There were no 7 - or 8-year-olds who produced disguised adaptors. A Fisher's exact test was used to examine the relationship between the age of the group and the number of children who produced disguised adaptors. There was a significant association between them (Fisher's exact test, $p=.03$ ). A residual analysis was conducted to find out where the significant differences among age groups were. The analysis indicated that 11-year-olds produced disguised adaptors more often than the other age groups did.

ntto: saisho: sono neko ga: tweety o: sagashi ni: it-te: [de mi-ta mi-ta nekotan] ttsut-te: sorede [(1)\% (2)sono ] (2.07) (3)sagashi ni it-ta tokoro wa: inu: toka neko wa dame tte iu omise de
'well, at first, the cat went to look for Tweety and (he) said "(he) saw, saw the pussy cat," then, well the place where (he) went to look for the cat was a shop where dogs and cats were not allowed'

Figure 2: A 11-year-old girl telling a story.
Figure 2 shows a typical case of a disguised adaptor in which an 11-year-old girl described the first of eight scenes, in which Sylvester the Cat goes into a building to catch Tweety Bird who lives there. Regarding the scene where Tweety said that he saw the cat, "mita mita neko tan ttutte sorede", the girl raises her right arm and puts her fingers into an O-shape ((1) in Figure 2). Given that she uses the right side of the space and a hand shape as if it is holding an object while talking about where Sylvester went to find

Tweety, the right hand movement seems to be part of the preparation phase to depict the building that Tweety is in.

However, while saying 'sono', which functions as a filled pause and also means an article "the" in English, she stopped production of the gesture's stroke and modified it into a disguised adaptor of rubbing her right eye with her right hand ((2) in Figure 2). After that, an unfilled pause, which lasted about 2 seconds, occurred until the next word started. During this pause, she retracted the disguised adaptor and described in speech the place where the cat went to find the bird. This case can be interpreted as being that she first tried to depict the building that the cat goes into by using both speech and gestures. However, because she could not remember the proper name for it, she abandoned the gesture in the middle of the production and modified it into a disguised adaptor.

Temporal relationship between speech and disguised adaptor
Each disguised adaptor could be categorized into one of six types (Table 2). The most frequent type was the cooccurrence of a disguised adaptor and intact speech (Type 1), followed by a combination of a disguised adaptor and intact speech after a false start(s) (Type 5) and synchronization of a disguised adaptor and a filled pause after a speech error and/or filled pause (Type 6).

Table 2: Number of cases categorised into temporal relationships

| Types ${ }^{1)}$ | Absolute number (\%) |
| :--- | :---: |
| 1. PREP $\rightarrow$ Intact speech + DA | $7(32)$ |
| 2. PREP $\rightarrow$ Filled pause + DA | $3(14)$ |
| 3. PREP $\rightarrow$ Unfilled pause + DA | $2(9)$ |
| 4. PREP $\rightarrow$ Speech error + DA | $1(5)$ |
| 5. PREP $\rightarrow$ Speech error ( $\rightarrow$ Speech | $5(23)$ |
| error) $\rightarrow$ Intact speech + DA | $4(18)$ |
| 6. PREP $\rightarrow$ Filled pause and/or |  |
| Speech error $\rightarrow$ Filled pause + DA <br> Total | $22(100)$ |

1) Abbreviations and notations: DA (disguised adaptor), PREP (preparation phase), $\rightarrow$ (order of temporal direction), + (synchronization of elements)

To see whether disguised adaptors were related to word searches, the six types were further classified into two
groups in terms of whether the disguised adaptor occurred with a speech error, (un)filled pause, or intact speech element. The former group is termed the dysfluent speech combination and includes Types 2, 3, 4, and 6, whereas the latter group is named the fluent speech combination and includes Types 1 and 5 . The number of children who produced each combination was counted for each age group. Three 9 -year-olds, three 10 -year-olds, three 11 -years-olds, and one 12-year-old produced fluent speech combinations. Five 11-year-olds and one 12-year-old produced dysfluent speech combinations. No children aged 9 or 10 produced dysfluent speech combinations. A Fisher's exact test was used to examine the relationship between age group and each combination. A significant association was found only in the dysfluent speech combination (Fisher's exact test, $p=$ .002). In contrast to the fluent speech combinations produced by 9 - to 12 -years-olds, dysfluent speech combinations were not produced by 9 - and 10 -year-olds, and all of 11-year-olds produced dysfluent speech combinations.

Figure 3 show a 9-year-old boy who produced a fluent speech combination. The figure shows him explaining a scene in which Sylvester got inside Tweety's apartment, but was struck by the bird's owner with an umbrella. While he was talking about the umbrella, he raised his right hand and shaped his hand as if it were holding an umbrella ((1) and (2) in Figure 3). However, without finishing the preparation as a gesture, he moved his hand to his eye to scratch it as a disguised adaptor. Unlike the girl in Figure 1, who produced a dysfluent speech combination, his speech did not contain obvious speech errors or pauses. Thus, his disguised adaptor may have been caused by other factors besides a word search (this point is taken up in the discussion section).

sorede ouchi no naka ni hait-ta $n$ dakedo sono kainu[shi no obaachan ga] dete[(1)ki-te (2)kasa de: (3)\# (4)nagurare-te: tsugi wa are ]
'and (he) got inside the house, the grandma, the owner comes out and (he) is struck with an umbrella, and the next is, umm'

Figure 3: A 9-year-old boy telling a story.

## Discussion

The present study investigated disguised adaptors as an index of a child's ability to monitor his or her own gestures by focusing on the relationship between disguised adaptors and speech flow. The results showed that the gesture frequency and the proportion of speech errors increase with age, especially in the late elementary school period, whereas
the proportion of unfilled pauses decreases with age. The increase in speech errors suggests that children tend to dedicate much effort to planning coherent narratives especially from the age of 10 . Considering pauses may reflect cognitive processes underlying speech planning including word search, syntax, conceptual and articulation planning (Schönpflug, 2008), it is considered that children gradually acquire the ability to plan speech quickly during their elementary school years.
In this study, disguised adaptors were produced by children who were more than 9 -years-old. None of the $7-$ and 8 -year-olds used disguised adaptors at all. This result indicates that children become aware of their gestures as an informational resource for listeners from the age of 9 onwards. In other words, they acquire the ability to monitor their gestures from the age of 9 .
The analysis of the temporal relationship between a disguised adaptor and speech fluency suggests that disguised adaptors are caused by not only speech dysfluency but also other factors. Children from 9- to 12 -years-old produce fluent speech combinations, but only the higher graders produced dysfluent speech combinations. This implies that disguised adaptors of 11- and 12-year-olds are partly caused by the act of searching for an adequate word or planning a sentence. In these cases, children may notice that they have to stop speaking to retrieve a word or re-plan a sentence, and accordingly they modify their gestures to a disguised adaptor in the middle of gesture production. Based on these results, I will discuss why disguised adaptors appear around the age of 9 years in terms of cultural and cognitive factors

## Cultural influence

Previous studies on the gestures of elementary-school-age children suggest a cultural influence. Some studies have reported that the frequency of gestures consistently increases during the elementary school years. These trends appear across cultures, although most of the studies were on children in Indo-European language cultures (e.g., Colletta et al., 2014). However, this study on Japanese children showed that the frequency of gestures decreases temporarily in the middle grades compared with in the lower grades or higher grades. This difference may come from their educational environment. In Japan, sometimes pupils are implicitly and explicitly warned by their teacher to avoid fidgeting or moving their hands when the teacher or another child is speaking or sometimes even when they themselves speak. In fact, speakers in Asia sometimes learn not to gesticulate (Neu, 1990). These findings suggest that Japanese children as young as 9 -years-old attempt to embody the rule about hand movement during communication. This may be related to why disguised adaptors produced by 9 - and 10 -year-olds do not synchronize dysfluences. Because they seem to start noting that their hand movements can be read by someone as symbols, even when they do not have a problem with word search, they may try to suppress their hand movements.

## Cognitive change

As a factor influencing the emergence of disguised adaptors from the age of 9 or 10 , one may consider cognitive changes occurring during this period. Piaget \& Inhelder (1969) suggested that the period is considered to be the concrete operational stage at which abstract logical thought is first applied to the physical world. Karmiloff-Smith (1986), who investigated the development of metalinguisitc awareness in 4 - and 12 -year-olds, found that many children from 9 years onwards explicitly have metalinguistic awareness. Thus, children in the middle grades of elementary school seem to develop metacognitive knowledge to notice that there are underlying rules or mechanisms in the physical world and human communication. At same time, they may also become aware that gestures are informational resources for the listener. Because children in this period are sensitive to such rules, they start monitoring their expressions to check whether the message in the expression is adequate given the communicative context. This awareness seems to result in the emergence of disguised adaptors and an increase of speech errors in children in the later years of elementary school. Ito and Tahara (1985) found that 10-year-olds had poorer usage of the postpositional particle $W A$ in comparison with other age groups. The speculated that because children in this period are just beginning to notice and attempt to grasp the multifunctional nature of language devices, their performance seems to decline temporarily. This suggests that the ages of 9 and 10 can be seen as the transitional period in which Japanese children begin noticing the communicative function of gesture and linguistic system, and monitoring them.

Examining when children suppress gestures contributes to an understanding of children's gestural development. Just as certain self repairs in speech that are spontaneously made by children during narratives reflects metalinguisitc awareness that they have acquired (Karmiloff-Smith, 1986), the emergence of disguised adaptors implies that children have an awareness of gestures. This study showed that although the production of gestures may be mostly an unconscious process (Goldin-Meadow \& Butcher, 2003), the speaker can notice that she is producing a gesture after she raises her hand for a gesture, and that this awareness begins at about 9 years of age. Future task is exmaning whther disguised adaptors are rubost phenomonen by collectiong more data from other age groups and from other cultuore groups.

## References

Beattie, G. (1983). Talk: An Analysis of Speech and Nonverbal Behaviour in Conversation. Open University Press: Milton Keynes.
Colletta, J. M., Guidetti, M., Caprci, O., Cristilli, C., Demir, O. E., Kunene-Nicolas, R. N., \& Levine, S. (2014). Effects of age and language on co-speech gesture production: an investigation of French, American, and Italian children's narratives. Journal of Child Language, 42, 122-145.

Ekman, P., \& Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding, Semiotica, 1, 49-98.
Fujii, M. (1997). Effects of self-touching behavior on the performance of lexical Retrieval. Japanese Journal of Psychology, 68, 187-196.
Fujii, M. (1999). Effects of gestures in communication: Developmental investigation of speech and gesture. Japanese Journal of Educational Psychology, 47, 87-96.
Goldin-Meadow, S., \& Butcher, C. (2003). Pointing toward two-word speech in young children, in: S. Kita (Ed.), Pointing: Where Language, Culture, and Cognition Meet (pp.84-107).Mahwah, N.J., Lawrence Erlbaum Associates.
Ito, T., \& Tahara, S. (1985). A psycholinguistic approach to the acquisition of multifunctionality in Japanese particles wa and ga. Descriptive and Applied Linguistics, 18, 121-131.
Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. Cognition, 23, 95-147.
Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. Key (Ed.), The relationship of verbal and nonverbal communication (pp. 207-227). The Hague, the Netherlands: Mouton.
Mayberry, R. I., \& Nicoladis, E. (2000). Gesture reflects language development: Evidence from bilingual children. Current Directions in Psychological Science, 9, 192-196.
McNeill, D. (1992). Hand and Mind, Chicago: University of Chicago Press.
Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence, in R. C. Scarcella, E. S., Andersen \& S. D. Krashen (Eds), Developing Communicative Competence in a Second Language (pp.121-138). New York: Newbury House.
Piaget, J. \& Inhelder, B. (1969). The Psychology of the Child. New York: Basic Books.
Pine, K. J., Bird, H., \& Kirk, E. (2007). The effects of prohibiting gestures on children's lexical retrieval ability. Developmental Science, 10, 747-754.
Schönpflug, U. (2008). Pauses in elementary school children's verbatim and gist free recall of a story. Cognitive Development, 23, 385-394.
Sekine, K. \& Furuyama, N. (2010). Developmental change of discourse cohesion in speech and gestures among Japanese elementary school children. Rivista di psicolinguistica applicata, 10, 97-116.
Sottofattori, S. M., \& Nicola, F. (2007). Self-Manipulation and the Tourette's syndrome, in Proceedings of the 3rd Conference of the International Society for Gesture Studies, pp.46-47. Chicago, America, June.
Volterra, V., \& Iverson, J.M. (1995). When do modality factors affect the course of language acquisition? in: E, Karen, J. S. Reilly (Eds.), Language, Gesture, and Space (pp. 371-391). Hillsdale, N.J., Lawrence Erlbaum Associates.
Wigglesworth, G. (1990). Children's narrative acquisition: A study of some aspects of reference and anaphora. First Language, 10, 105-125.

# Adaptability and Neural Reuse in Minimally Cognitive Agents 

Matthew Setzler (msetzler@indiana.edu)<br>Cognitive Science Program, Indiana University Bloomington, IN 47408 USA

Eduardo J. Izquierdo (edizquie@indiana.edu)<br>Cognitive Science Program, School of Informatics and Computing, Indiana University Bloomington, IN 47408 USA


#### Abstract

Cognitive agents are continuously faced with new problems. To facilitate adaptation, emerging theories of neural reuse propose that evolution might often favor re-purposing existing brain structures for new functions. This paper presents a novel approach to the study of neural reuse based on the evolution of simulated agents in an object-categorization task. We artificially evolve populations of dynamic neural networks to perform two variants of a categorization task that alternate over evolutionary time. We find that populations become increasingly adaptive over repeated exposures to the tasks. Analysis of evolved networks reveals two types of equally-fit solutions: one that is specialized to a given task variant and does not adapt to changes easily; and another that is more general, in that it can adapt to the other task with minimal change to its structure. Interestingly, we find that populations exposed to alternating tasks spontaneously locate the latter type of structures.


Keywords: neural networks; minimally-cognitive behaviors; neural reuse; artificial evolution; evolvability

## Introduction

A central goal of cognition is adaptation. Cognitive agents are continuously faced with new problems and it is in their best interest to reuse pre-existing solutions from prior problems wherever possible. Many lines of research in cognitive science are motivated by this broad notion of adaptation. An emerging class of theories suggest neural reuse is a fundamental organizing principle of adaptation in the brain (Anderson, 2007; Gallese, 2008; Dehaene, 2005; Hurley, 2008). According to these theories, neural circuits established for one purpose are often put to different use without losing entirely their original functions. Different versions of this theory have been proposed, some that operate over the developmental timescale (Dehaene, 2005) and others over evolutionary timescales (Anderson, 2007). These theories are supported by a wealth of empirical evidence (see Anderson, 2010, for more details). However, a grounded account of how reuse might occur remains poorly understood. Computational models in this area have tended to focus on disembodied cognitive architectures (Braylan, Hollenbeck, Meyerson, \& Miikkulainen, 2016), largely detached from the increasingly accepted viewpoint that cognition is distributed over a dynamically coupled brain-body-environment system (Chiel \& Beer, 1997).

The work presented here takes a different approach to studying adaptation and neural reuse over the evolutionary timescale. We study adaptation in a computational model of dynamic embodied agents evolved to perform two variants
of a visual discrimination task. Specifically, we artificially evolve populations of dynamic neural networks to perform two closely-related variants of a visual discrimination task that alternate over evolutionary time. To the extent that adaptability is achieved, we should observe populations adapting quicker to changes in the task requirement over the course of an evolutionary run. The degree to which neural reuse occurs will be evident in the evolved neural structures, which are readily analyzable after evolution.

Our approach extends beyond the scope of other work in two important ways. First, studying reuse in a deliberately simple computational model gives us a tractable and fully transparent system which can be analyzed to yield a grounded account of how adaptability and neural reuse play out in an embodied and situated context. Secondly, unlike other computational approaches which employ evolutionary frameworks specifically designed to facilitate neural reuse (Stanley \& Miikulainen, 2002), we use a standard genetic algorithm, thus minimizing assumptions about the form of neural reuse by seeing how it might occur spontaneously as a result of implicit pressures imposed by changing task requirements.

In this paper, we aim to show adaption via neural reuse over the evolutionary timescale in an idealized computational model of dynamic, embodied agents performing two variants of a previously studied visual discrimination task (Beer, 1996, 2003). Insofar as model-agents exhibit hallmarks of cognition, this is of interest in and of itself. Additionally, insight gained in this idealized computational setting will provide a proof of principle for how reuse might occur in naturalistic (embodied/situated) settings, which will help inform the conceptual and analytical tools used to study this phenomenon in living systems.

## Methods

## Agent, Environment, and Task

In previous work (Beer, 1996, 2003), model agents were evolved that could "visually" discriminate between objects of different shapes, catching some while avoiding others. All details of the agent, environment, and task have been adapted from these previous studies.

The agent lives on the floor of a $275 \times 400$ rectangle (Figure 1A). It has a circular body with a diameter of 30 and an "eye" which consists of 7 vision rays evenly distributed over an angle of $\pi / 6$. These rays extend out from the agents body with a maximum range of 220 . There are two kinds of objects


Figure 1: Agent and environment, adapted from (Beer, 1996). [A] Agent moves horizontally while an object falls towards it from above. Object can be one of two shapes: circle or line segment. Agent's sensory apparatus consists of an array of seven distance sensors (dashed lines). [B] Neural architecture. Distance sensors (black) project to a layer of fully interconnected interneurons (gray, recurrent connections not shown), which in turn project to the two motor neurons (light gray).
in the world: circles (with a radius of 15 ) and horizontal line segments (of length 30). These objects fall from a height of 275 at some initial horizontal offset with respect to the agent. Objects fall with a constant vertical velocity of -3 and no horizontal motion. If an object intersects a ray within this range, an external input is fed to a corresponding sensory neuron. The value of the input is inversely proportional to the distance at which the intersection occurs, normalized from 0 to 10.

As depicted in Figure 1B, the agent's nervous system is a 3-layer continuous time recurrent neural network (CTRNN) with the following state equation (Beer, 1996):

$$
\begin{equation*}
\tau_{i} \frac{d y_{i}}{d t}=-y_{i}+\sum_{j=1}^{N} w_{j i} \sigma\left(y_{j}+\theta_{j}\right)+I_{i} \tag{1}
\end{equation*}
$$

where $y$ is the state of each neuron, $\tau$ is its time constant, $w_{j i}$ is the strength of the connection from the $j^{t h}$ to the $i^{t h}$ neuron, $\theta$ is a bias term, $\sigma(x)=1 /\left(1+e^{-x}\right)$ is the standard logistic activation function, and $I$ represents an external input (e.g., from a sensor). The top layer consists of 7 sensory neurons which are stimulated by the agent's vision rays as described above. These project down to a middle layer of 5 fully interconnected neurons, which in turn feed into two motor neurons. The difference in output between the motor neurons results in an instantaneous horizontal velocity which moves the agent in one direction or the other. The network is bilaterally symmetric in terms of connection weights, gains and biases, with the additional stipulations that all sensory neurons shared the same gain and biases. This makes for a total of 47 parameters. States were initialized to 0 and circuits were integrated using the forward Euler method with an integration step size of 0.1 .

The task of the agent is to visually discriminate between objects of different shapes. These experiments were designed
to produce evolved examples of categorical perception (Beer, 2003). For the purpose of our study, there are two variants of the object discrimination task. In Task A, agents must distinguish between circles and line objects by reliably moving towards the former and away from the latter. In Task B, the agents must do the opposite: move towards line objects and avoid circles. The tasks were chosen because of how closely related they are to one another.

## Evolutionary Swapping

A real-valued genetic algorithm was used to evolve CTRNN parameters: connection weights, biases, time constants, and gain. Agents were encoded as 47-dimensional vectors of real numbers varying from $[-1,1]$. Each vector element linearly mapped to a parameter of the CTRNN: interneuron and motorneuron biases $\in[-5,5]$, sensory neuron biases $\in[-4,-2]$, time-constants ranged $\in[1,2]$, and connection weights $\in[-5,5]$. Parents were selected with a rank based mechanism, with an enforced elitist fraction of 0.04 . Offspring were generated from uniform crossover of two parents (probability 0.5 ). A Gaussian distributed mutation vector was applied to each parent $\left(\mu=0, \sigma^{2}=0.01\right)$.

Fitness was calculated across 24 trials with objects dropped at horizontal offsets uniformly distributed over the range of $+/-50$. The performance measure to be maximized was: $\sum_{i=1}^{24} p_{i} / 24$, where $p_{i}=1-\left|d_{i}\right|$ for the objects that need to be caught and $p_{i}=\left|d_{i}\right|$ for the objects that need to be avoided, and $d_{i}$ is the horizontal distance between the centers of the object and the agent when their vertical separation goes to zero on the $i$ th trial. The distance $d_{i}$ is clipped to a maximum value of 45 and normalized to run between 0 and 1 .

The two tasks have inverted fitness rewards. In Task A, agents are required to catch circles and avoid line segments; in Task B agents are required to catch line segments and avoid circles. Therefore, high fitness with respect to one guarantees proportionately low fitness on the other.

We performed evolutionary runs under changing conditions. In this case, a population was initially evolved for a given task variant until the best agent reached a fitness threshold of $95 \%$. Once the fitness threshold was reached, the task was changed to the opposite task. Populations were continuously evolved in this alternating manner for 2500 generations. In addition to these, runs were performed for Task A and Task $B$ in isolation under the same evolutionary conditions.

## Results

The following sections report results obtained from 40 evolutionary swapping runs. We focus first on understanding one exemplary evolutionary run where adaptation occurs via neural reuse. That is, we show that over the course of an evolutionary run the population locates a general-purpose neural structure that enables it to quickly respond to changing task requirements. We then analyze this increase in adaptability in terms of population dynamics in parameter space and reuse of neural structure in the agents. In the last section, we look
at how well these findings generalize to the full ensemble of evolutionary runs.

## Increased adaptation over time

As most work on evolving minimally cognitive behaviors focuses on one task, we set out to test whether evidence for adaptation could be observed in populations under changing conditions. That is, does learning one task help a population adapt to new tasks over time?

Figure 2 depicts the fitness trajectory of the best agent in the population in each generation for one of the evolutionary runs. Altogether, this population achieved a total of 98 swaps in 2500 generations. The first 4 are labeled for the sake of discussion; A1 indicates the first time the population achieves fitness threshold for Task A, B1 indicates the first time population reaches threshold for Task B, etc. This trajectory shows evidence of adaptation. As can be seen, the population starts off close to random behavior (0.5) and rapidly improves performance at the first task Task A: circlecatching, line-avoidance). As soon as the best individual in the population surpasses the established fitness threshold, the task is swapped (A1), at which point fitness of the population suddenly drops. Recall that fitness measures of each task are inversions of one another, so we should expect a population with high fitness with respect to one task to have a sudden drop in fitness when tasks are swapped. The population quickly begins to make improvements in performance at the new task (task B: circle-avoidance, line-catching). As soon as the threshold is met for this task (B1), the task is swapped back to Task A, and the fitness drops again. Interestingly enough, the fitness does not drop anywhere as low as what we would expect from an agent that performs the opposite behavior, or even a random behavior. Thereafter, the population evolved to very quickly re-adapt to the new task after each new swap (Figure 2B).

The fitness trajectory of this evolutionary run shows evidence of adaptation. The trajectory is characterized by periods of ascent leading to a peak and sudden drop offs that occur due to task swaps. The time-to-adapt corresponds to the number of generations taken to reach the fitness threshold for the new task after a swap. In this study, we consider adaptability as the extent to which populations improve their time-to-adapt to changes in the fitness landscape over time. Therefore, a population shows evidence of adaptability when the time-to-adapt decreases over the course of an evolutionary run. In this example, we observe that the first time the population evolves for Task A and Task B it took around 500 generations each. After that, the time-to-adapt dropped sharply to between 1 and 30 generations. In other words, the population improved substantially its time-to-adapt by the third task swap, and maintained this ability for the duration of the evolutionary run.

## Spontaneous meta-fitness selection

How did such a dramatic increase in adaptation efficiency occur? We hypothesize that the population increases its adap-


Figure 2: Increased adaptation over time in an evolutionary run. [A] Fitness trajectory of the best individual in the population over time. Dashed vertical lines mark task swaps. Blackdotted horizontal line represents the $95 \%$ fitness threshold. The first 4 task swaps are labeled A1, B1, A2 and B2 respectively. Panel [B] shows the inset in panel [A].
tation efficiency by locating a region of meta-fitness, which we define as a region in parameter space where individuals with high fitness on both tasks exist in close proximity to one another. In other words, the proposal is that adaptability over evolutionary timescales involves a spontaneous selection for regions of parameter space that support multiple tasks over regions of parameter space that are specialized for only one of the tasks. In order to test this hypothesis, a principal component analysis (PCA) was performed on the 47-dimensional set of all genotypes in populations at points A1, B1, A2 and B2. Figure 3 shows these populations in a reduced 2D space, where the first dimension captures $73.1 \%$ of the overall variance and the second captures $1.7 \%$.

The movement of the population in parameter space during evolution shows evidence of meta-fitness selection. Figure 3 shows the structure of the evolutionary dynamics over the course of the first four task swaps. The population starts off at random in parameter space and they find a region that is adapted to Task A for the first time (A1). The task swaps to Task B, and the population moves from A1 to B1, where they find a region that is adapted to Task B, also for the first time. When Task A is presented once again, the population finds a region of parameter space that is relatively nearby (A2). The same occurs when the task swaps back to B (B2). The number of generations to adapt to each task, therefore, corresponds to the relative distances that the population moved in


Figure 3: Spontaneous meta-fitness selection. [A] Projection of populations at A1, B1, A2 and B2 in reduced 2D space obtained through PCA. [B] Fitness distributions of populations seeded around the best individual at A1 (blue) and the best individual at A2 (green). Populations were obtained by applying Gaussian perturbations within the range of $[-0.2,0.2]$ to the seeded values in each dimension.
parameter space. This suggests that the fitness landscape contained two qualitatively different kinds of adaptive regions. Although individuals in populations A1 and A2 were both equally adapted to Task A, individuals in A1 were isolated in parameter space, while individuals in A2 were in close proximity to regions of high fitness for the opposite task. What is interesting about this result is that we did not include direct selection pressure towards these regions; instead populations were guided towards them spontaneously over repeated exposure to different tasks.

What is the fundamental difference between these two types of adaptive regions? To address this question we sampled random populations in each region and evaluated their fitness distributions; results are presented in Figure 3B. Recall that each agent achieved equally high fitness with respect to Task A, but the best individual at A2 was found to be located in a region of meta-fitness whereas the best individual at A1 was not. As expected, both populations contain equally high-performing circuits. However, the fitness distribution of agents sampled around A1 (blue) is highly skewed towards high fitness values, as would be expected for a typical fitness peak. In contrast, the fitness distribution of agents sampled around A2 (green) is significantly flatter, indicating that it encompasses a greater diversity of phenotypes. This result reveals a crucial difference between meta-fitness regions and otherwise equally-fit fitness peaks.


Figure 4: Similar behavioral strategy and neural structure for opposite tasks. Behavior (left) and structure (right) of best individuals from populations A2 and B2. Left panels: Horizontal position of agents over time as object falls for all 24 trials (circle trials purple, line trials cyan). Right panels: Neurons represented by disks (opacity depicts bias parameter, where a large negative bias is white and a large positive bias is black) and connections represented by lines (excitatory grey, inhibitory black; thickness proportional to strength).

## Neural reuse in meta-fitness regions

Meta-fitness regions in parameter space have interesting implications for neural reuse. Due to linearity in the genotypephenotype map, close proximity of genomes in parameter space translates to a high degree of similarity between neural structures in corresponding phenotypes. As a result, we observe that best-agents from the same meta-fitness region share the same essential structure despite being evolved to perform different tasks. This is illustrated in Figure 4, which presents graphic depictions of the neural structures of the best agents taken from populations A2 and B2. The two structures are nearly identical, despite the fact that they ultimately support different task variants. In other words, structure evolved to perform Task A in A2 was largely reused in the subsequent evolution for Task B.

In addition to neural reuse, we observe that these two individuals utilize the same behavioral strategy for performing each task (Figure 4). In each case, the agent scans back and forth over the mid-line several times before either centering or heading away. Although not shown, this strategy is qualitatively different from the behaviors of best agents performed in isolated runs.

## Ensemble Analysis

How reliably does evolutionary swapping result in increased adaptation of populations? Of the 40 runs, 17 never reached the fitness threshold in the allotted 2500 generations. We consider these to be null cases. Of the remaining 23 runs, 9 showed evidence of increased adaptation over time. It is unclear whether or not the runs that did not demonstrate signs


Figure 5: Increased adaptability via meta-fitness. Distances between best-genomes from swapping runs (black) versus distances between best-genomes produced by successful isolated runs (grey).
of increased adaptability would have had they been allowed to run past the arbitrary 2500 generation cutoff. Regardless, the remainder of the section is concerned with the 9 runs that showed evidence of increased adaptability.

In the case study, increased adaptation was due to the evolving population locating an area of meta-fitness. Is this also the case for the rest of the populations that showed increased adaptability? We computed the set of Euclidean distances separating best genomes in successive epochs for all swapping runs and compared this with the set of distances separating best genomes coming from successful isolated runs. As depicted in Figure 5, best-agents produced in the same swapping run are usually quite closer to one another in parameter space than random opposite-task best-genomes sampled from isolated runs. This suggests that in general, adaptability is achieved via meta-fitness regions.

Given that populations in general became more adaptive by locating meta-fitness regions, it is of interest to ask whether each run located the same region or different meta-fitness regions were located by different runs. Figure 6 displays distributions of within-run distances (Euclidean distances separating the best agents of all target populations occurring within given runs) and inter-run distances (Euclidean distances separating best agents of all target populations occurring in different runs). If all swapping runs located the same meta-fitness region we would expect these two distributions to be equivalent. That within-run distances are substantially less than inter-run distances suggests that different swapping runs indeed located different meta-fitness regions.

As we observed in the case study, a region of meta-fitness in parameter space corresponds to a general-purpose neural structure. In the case study, we also observed that best-agents from the same meta-fitness region can perform opposite task variants with the same behavioral strategy. Figure 7 illustrates the generalization of these ideas to multiple meta-fitness regions. It depicts three pairs of best-agents from separate evolutionary runs. In each pair both agents have nearly-identical neural structures and qualitatively similar behavioral strategies despite performing opposite tasks. If we contrast pair


Figure 6: Different runs locate different meta-fitness regions. Distances between best-genomes of the same evolutionary run (black) versus distances between best-genomes of different evolutionary runs (grey).

A with pair B we see that they each perform the tasks with the same qualitative behavioral strategy, despite the fact that they have different neural structures. This demonstrates how a given behavioral strategy can be supported by qualitatively different neural structures. In contrast, pair C has a qualitatively different behavioral strategy from both A and B.

## Discussion

We set out to test whether evidence for adaptation and neural reuse could be observed in populations of embodied agents required to perform two variants of an object discrimination task that changed over evolutionary time. Evidence of adaptation was indeed observed: over the course of an evolutionary run populations gained the ability to more quickly adapt to changes in task requirement. In general, populations achieved this increase in evolutionary efficiency by evolving generalpurpose neural structures, which could be reused to support either task variant with minimal structural modifications. Below we discuss these results in light of continued work in neural reuse and evolvability.

Despite a growing body of empirical and theoretical support for neural reuse (Anderson, 2010), a grounded understanding of how neural reuse works in organisms remains elusive. There have been some computational efforts to this effect (Hurley, 2008; Anderson, 2010), but the present work contributes in two important ways. First, whereas other modeling work incorporates elements of reuse into the model, we see reuse spontaneously arising in our system as a result of the evolutionary pressures imposed by the varying task requirements. Secondly, in studying dynamic embodied agents, we examine neural reuse in light of the increasingly accepted viewpoint that cognition is distributed over a brain-body-environment system (Chiel \& Beer, 1997). In previous work, it was shown that the model-agents studied here offload some cognitive load to their bodies and the environment (Beer, 2003). It is likely that successful neural reuse observed in these simulations relies on features of embodiment. Ongoing work on a deeper, more thorough analysis of agents exhibiting reuse promises to bring some of these issues to light.


Figure 7: Different behavioral strategies and general-purpose neural structures are located in different evolutionary runs. Example solutions from three different evolutionary runs are shown in panels $[\mathrm{A}],[\mathrm{B}]$ and $[\mathrm{C}]$. For each evolutionary run, two solutions are shown: the best agent for line-catching and circle avoidance (i); the best agent for the opposite task obtained successively (ii). Graphics follow same coloring conventions as Fig. 4.

Work on lifetime learning and adaptation in cognitive systems parallels ongoing research on evolvability in evolutionary systems. Evolvability refers to the ability of a population to adapt to changes in its fitness landscape, and has been a hot topic in theoretical biology and evolutionary computation (Pigliucci, 2008). In order for an evolutionary system to support evolvability there must be redundancy in the fitness landscape; there must be many genomes which have high fitness in a given environment, and some of these must be more adaptable to future environmental changes than others. Typically investigations of evolvability examine the role of developmental processes indirectly linking genotype to phenotype as a mechanism for such redundancy (Pigliucci, 2008). In the present work, however, we observe signatures of evolvability in a system without any developmental scenario. While our system has a direct genotype to phenotype map, the complexity of agents introduces sufficient redundancy in the phenotype to behavior map (qualitatively different neural structures can produce the same basic behavior) and behavior to fitness map (qualitatively different behaviors can yield the same fitness) to allow for increased adaptation.

The work presented above demonstrates that evolving embodied agents for time-varying cognitive tasks is fertile ground for exploration of cognitive adaptation in general, and neural reuse more specifically. There are many ways in which efforts here can be built upon. First, given that neural reuse is most concerned with brain circuits supporting vastly different cognitive functions (Anderson, 2010), we would like to expand the repertoire of tasks to include some that are qualitatively different. As our understanding of neural reuse advances, we can extend our approach to increasingly complex interesting tasks. Secondly, we would like to analyze in more depth the operation of agents that exhibit reuse. Past work has shown that the mathematical tools of dynamical systems theory and information theory can be used to provide rich accounts of cognition exhibited in these systems (Beer \& Williams, 2015). Ultimately, we would like to provide a grounded understanding of neural reuse in embodied, situated agents capable of diverse cognitive phenomena.

## References

Anderson, M. (2007). Evolution of cognitive function via redeployment of brain areas. The Neuroscientist, 13, 1321.

Anderson, M. (2010). Neural reuse: a fundamental organizational principle of the brain. Behav Brain Sci., 11(4), 245-266.
Beer, R. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In From animals to animats 4 (pp. 421-429). MIT Press.
Beer, R. (2003). The dynamics of active categorical perception in an evolved model agent. Adaptive Behavior, 33(4), 209-243.
Beer, R., \& Williams, P. (2015). Information processing and dynamics in minimally cognitive agents. Cognitive Science, 39(1), 1-38.
Braylan, A., Hollenbeck, M., Meyerson, E., \& Miikkulainen, R. (2016). Reuse of neural modules for general video game playing. AAAI.
Chiel, H., \& Beer, R. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. Trends in Neuroscience, 20(12), 553-557.
Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. In From monkey brain to human brain. MIT Press.
Gallese, V. (2008). Mirror neurons and the social nature of language: The neural exploitation hypothesis. Social Neuroscience, 3(3-4), 317-333.
Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. Behavioral and Brain Sciences, 31(1), 1-58.
Pigliucci, M. (2008). Is evolvability evolvable? Nature Reviews Genetics, 9, 75-82.
Stanley, K., \& Miikulainen, R. (2002). Evolving neural networks through augmenting topologies. Evol. Comput, 10(2), 99-127.

# The role of talker similarity in the perceptual learning of $\mathbf{L} 2$ tone categories 

Jing Shao (jing.shao@polyu.edu.hk)<br>Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University Hung Hom, Hong Kong

Joanna Chor Yan Mak (joanna.mak@connect.polyu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University Hung Hom, Hong Kong

Caicai Zhang (caicai.zhang@polyu.edu.hk)<br>Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University Hung Hom, Hong Kong<br>Shenzhen Institutes of Advanced Technology, The Chinese Academy of Sciences<br>Shenzhen, China


#### Abstract

Different hypotheses were proposed concerning the role of talker variability in lexical learning. It remains unclear whether new phonetic categories are acquired as episodic memory traces with talkers' voice information preserved or as abstract categories. The current study investigated the role of voice similarity in perceptual learning of Cantonese tones. Six high-variability training sessions were given to 12 Mandarin speakers. Voice similarity was controlled in the training and pre-and posttests. Results indicate that the training positively transferred to both similar and dissimilar talkers. However, in the pretest, the performance was not significantly different between similar and dissimilar voices, whereas significant better performance was found in the similar voices in the posttest. These results suggest that learners retained speakers' information in the learning process and made use of such information for future perception. This implies that lexical tones are probably encoded episodically in the mental representation of Mandarin L2 learners.


Keywords: Talker similarity, high variability training, Cantonese lexical tones, Mandarin leaners of Cantonese, mental representation.

## Introduction

When presented with a word auditorily, a listener receives a lot of information. For instance, the phonological form and meaning of the word, as well as the auditory information of the speaker. There existed a great amount of acoustic differences between speakers resulted from the differences in the shape and length of vocal tract, articulatory dynamics and native dialects (Goldinger, 1998). A long-documented problem for theories is how the speech perception and spoken word recognition achieve perceptual constancy in spite of the highly variable speech signals (Bradlow, Nygaard and Pisoni, 1999).

Two different accounts had been put forward concerning how the speaker's information is encoded in the memory system during auditory lexical learning, which are called abstractionist approach (Pisoni, 1997) and episodic theory (Nygaard, Sommers, \& Pisoni, 1994). According to the speaker normalization hypothesis (Joos, 1948, as cited in

Goldinger, 1998), acoustic variances produced by different speakers were considered as redundant information or noise that was quickly forgotten after lexical access and only lexical-semantic information was encoded in long-term memory. This allows listeners to understand new speakers instantly by only following the lexical-semantic content of speech and disregarding the superficial details such as speaker's information (Goldinger, 1998). In this abstractionist approach, a speaker-specific representation which contained all the lexical and speaker's information was first modified to a relatively speaker-neutral abstract representation prior to the encoding of lexical-semantic information (Johnson, 1997). The encoded lexical-semantic information was stored as an abstract representation, which then formed a template or prototype that could be used to match with the incoming speech signals to allow lexical retrieval (Goldinger, 1998). For instance, Peterson and Barney (1952) found that listeners could correctly perceive the target vowel despite the great variation in vowel formant frequencies produced by men, women and children. Evidence was also found in the perceptual normalization in consonants and prosody (Johnson, 2005), but normalization of such categories requires further intrinsic and extrinsic cues (Johnson, 2005; Moore, 1996; Strand \& Johnson, 1996; Zhang \& Chen, 2016). In a more recent study, it was also found that newly learned words could be sufficiently lexicalized, and be abstract with respect to talker voices (Kapnoula and McMurray, 2016).

On the other hand, a growing body of research suggested another direction. Episodic theory proposed that perceptual details including speaker's information were retained in memory during lexical learning and were integrated into perception later (Goldinger, 1996; 1998). All experienced instances were defined in a perceptual category and no abstract categories or prototypes were created. In accordance with this theory, Goldinger, Pisoni and Logan (1991) found that listeners made use of speaker variability to recall individual items in multiple-speaker word list, achieving higher accuracy than that in single-speaker word list. This result suggested that speaker variability in
multiple-speaker word list facilitated word encoding and retrieval. Moreover, Bradlow, Nygaard and Pisoni (1999) presented a word list to a group of participants auditorily and asked them to judge whether they had heard the words before. They found that listeners made use of speakers' information as well as speaking rate and amplitude information in the tasks which also supported the episodic theory. In addition, Nygaard and Pisoni (1998) found that speech from familiar voices was more intelligible than from unfamiliar voices, which suggested speakers' information were related to linguistic processing at both word and sentence levels (see Souza et al., 2013 for similar findings). Altogether, these studies suggested that the talker information was stored as episodes in the long-term memory.

The two hypotheses mentioned above hold different views on the role of talker information in language processing/acquisition. Perceptual training in second language (L2) learners offers a scenario to test whether the new phonetic categories are learned as episodic memory traces with speaker's voice information preserved or as abstract categories, which is the aim of the current study.

High variability perceptual training (HVPT hereafter) was developed from the low variability training in which only one speaker was involved and very limited phonetic context was provided (Strange \& Dittman, 1984). HVPT exposes subjects to a wider range of stimuli, including sounds produced by several speakers, in multiple phonetic contexts, and at multiple syllable positions (Bradlow, 2008). Several studies have adopted this approach to improve non-native speakers' identification of consonants (Bradlow et al., 1997; Bradlow et al., 1999;), vowels (Iverson \& Evans, 2009; Iverson, Pinet \& Evans, 2012), as well as lexical tones (Wang, Jongman \& Sereno, 1999; Wang, Spence, Jongman \& Sereno, 2003; Wang, 2013). These findings confirmed that HVPT was very effective in facilitating the acquisition of non-native phonetic contrasts.

Although HVPT was found to be effective in the generalization to novel stimuli and speakers, previous studies did not control the speaker voice similarity between training and pre-/posttests. The current study aims to control the voice similarity in the training sessions and pre-and posttest, and to compare the generalization effect to novel speakers whose voice was either similar or dissimilar to those speakers during training. Via this study, the question of whether L2 speech sounds are encoded as an abstract representation or an episodic representation during L2 perceptual learning will be investigated.

The current study focuses on the L2 lexical tone learning. Like consonants and vowels, lexical tones are important in differentiating word meanings in tonal languages. Mandarin and Cantonese are both tonal languages where pitch patterns of a syllable are crucial to its lexical meaning. Tone contours are commonly shown by numbers representing the pitch register according to a scale of five, 1 being the lowest and 5 being the highest. Usually two numbers, e.g., 55, indicate the pitch at the beginning and end of a syllable respectively (Chao, 1930). Mandarin has four lexical tones:

Tone 1: high level (55); Tone 2: rising (35); Tone 3: fallingrising (214); and Tone 4: falling (51) (Norman, 1988). Cantonese has six regular tones, which are classified according to their register and contour (Bauer \& Benedict, 1997). The six distinctive tones are: Tone 1: high level (55); Tone 2: high rising (25); Tone 3: mid level (33); Tone 4: mid-low falling (21); Tone 5: mid-low rising (23); and Tone 6: mid-low level (22). These six distinctive tones were included as the stimuli of the current study.

We aim to investigate the nature of mental representation of Cantonese tones in Mandarin L2 learners by controlling voice similarity of novel speakers. This allows us to examine the two hypotheses mentioned above. If Cantonese tones were encoded as an abstract representation, Mandarin listeners would ignore speaker variability and therefore generalize to novel speakers no matter whether their voices are similar or dissimilar to those in perceptual training. If Cantonese tones were encoded as an episodic representation, listeners would make use of speakers' information in lexical retrieval and therefore demonstrate better generalization to novel speakers whose voices are similar than those whose voices are dissimilar to the speakers in the perceptual training.

## Method

## General design

A pretest-training-posttest design was employed to assess the subjects' initial ability and the effects of training. Pretests and posttests consisted of two main parts: tone category identification and discrimination. The identification and discrimination tasks were designed to test whether identification training is effective and transfers to new talkers whose voice were similar and dissimilar to the talkers used in the training sessions.

## Participants

Student participants were recruited in Hong Kong Polytechnic University (PolyU). 45 students responded to an online self-report questionnaire. 19 participants were selected based on the following criteria: (1) resided in Hong Kong for less than five months prior to the pretest session, (2) speaks Mandarin as mother tongue and did not speak any Southern dialect including Hakka and Southern Min dialect, (3) did not receive professional musical training. Seven participants withdrew from the study before the posttraining test. A total of 12 participants completed the whole program.

## Talkers and Stimuli

## Stimuli

The stimuli were 60 words contrasting six Cantonese tones (high level tone (T1)-/55/, high rising tone (T2)-/25/, mid level tone (T3)-/33/, extra low level/low falling tone (T4)$/ 21 /$, low rising tone (T5)-/23/, low level tone (T6)-/22/) on ten base syllables (/jan/, /ji/, /jau/, /jiu/, /fan/, /fu/, /ngaa/, $/ \mathrm{si} /$, /se/ and /wai/), all are meaningful in Cantonese. Each
monosyllabic target word was embedded in a carrier phrase context＂呢個係＿lei1 go3 hai6［target word］＂（this is ［target word］）．Six female and six male native Cantonese speakers recorded the stimuli．Each speaker recorded six repetitions of each target word．One token for each target word was chosen by the experimenters according to its intelligibility and tone accuracy．The carrier phrase was normalized in duration to 877 ms （mean value of all carrier phrases），and the target word was normalized to 631 ms （mean value of all target words）．The mean intensity was scaled to 70 dB using Praat．

## Voice Similarity Judgment

The voice similarity among the 12 talkers was rated by another 12 native Cantonese speakers who were blind to the purpose of the current study．One speaker for each gender （F01 and M01）were chosen as references．The other five talkers（F02，F03，F04，F05 and F06；M02，M03，M04，M05 and M06）in each gender were compared against the reference speaker in term of voice similarity．In the similarity judgment experiment，raters were asked to listen to the target words embedded in the context＂呢個係．．．lei1 go3 hai6．．．＂［This is ．．．］spoken by the reference speaker and one other speaker of the same gender．They were asked to rate the voice similarity of the speakers on a scale of 1 （very dissimilar）to 9 （very similar）．The 60 target words were included and each trial was repeated twice．

The similarity score was averaged across raters．Table 1 shows the similarity score of each talker．Three talkers who received highest similarity scores in each gender group were used in the training sessions（F02，F04 and F06；M03，M04 and M06）．The talkers with lowest similarity rating score in each gender（F03 and M05）were included in the pre－and posttests as the speakers with dissimilar voices．

Table 1：Result of similarity judgment test averaged across 12 raters．

| Female <br> talkers | Similarity <br> score |  | Male <br> talkers | Similarity <br> score |
| :---: | :---: | :---: | :---: | :---: |
| F02 | $\mathbf{7 . 4 4}$ |  | M04 | $\mathbf{7 . 4 9}$ |
| F04 | $\mathbf{7 . 2 7}$ |  | M06 | $\mathbf{7 . 7 3}$ |
| F06 | $\mathbf{6 . 0 7}$ |  | M03 | $\mathbf{7 . 3 8}$ |
| F05 | 5.90 |  | M02 | 5.91 |
| F03 | 4.70 |  | M05 | 5.30 |

## Procedure

The training programme consisted of a pretest，training，and posttest phase．All sessions were conducted in a soundproof room in PolyU．

## Training

There were six sessions of HVPT（i．e．，tone identification with feedback）．The entire course of training for each subject was completed over 1－2 weeks，and each session lasted about 1 hour．There was a different talker each session，as is typical of HVPT procedures．Female and male
stimuli were trained alternatively．Moreover，our design was different from previous studies in that the six talkers in the training sessions（ 3 female and 3 male）were similar to two of the talkers who were presented in the pre－and posttest． Although the talkers used in the training phase were judged to have similar voices，it is still a HVPT，as the subjects were exposed to a wide range of stimuli，produced by several speakers，and the tones were carried by multiple syllables．


Figure 1．An example of identification choice in the training sessions．

The 60 stimuli were repeated twice，which gave 120 basic trials in each training session．On each trail，participants were presented with a target word with the context＂呢個係 ．．．leil go3 hai6．．．＂［This is ．．．］．Six choices with minimal contrast of tones were presented on the computer screen（see an example of＂ngaa＂in Figure 1）．As can be seen in Figure 1，visual demonstration of each of the six tones were provided．The subjects were then asked to indicate the tone of the target word with the number keys 1－6 corresponding to the tones one to six．After each response，a feedback screen was presented．The feedback information includes： （1）if the response made by the participant was correct in blue text or incorrect in red text，（2）the cumulative accuracy of the current training session and（3）the correct response， its corresponding tone contour and its corresponding word written in traditional Chinese．If the response made was correct，the participant was proceeded to the next trial，if the response was incorrect，the incorrect trial was repeated until a correct response was indicated．

## Pre－and Posttest

The pretest and posttest consisted of identification and discrimination tasks．In both identification and discrimination task，there were four talkers separated in four blocks．Two talkers（female F01 and male M01）were similar to those used in the training sessions，while two talkers were of dissimilar voice to the talkers used in the training（F03 and M05）．

Tone identification．In the identification task，the 60 stimuli words were presented with the context＂呢個係．．． lei1 go3 hai6．．．＂［This is ．．．］．Each trial was repeated twice and presented randomly to the participants．Six choices with minimal contrast of tones were presented on the computer screen with the tone，its corresponding tone contour and its corresponding character in traditional Chinese as shown in Figure 1．Participants were asked to indicate the tone of the
target word by pressing the number keys 1-6 corresponding to the tones one to six. No feedback was given.

Category Discrimination. In the discrimination task, the syllable /ji/ was selected as the stimuli. Fifteen different tone pairs were presented in both directions of comparison (T1-T2, T2-T1, T1-T3, T3-T1, T1-T4, T4-T1 etc.) and then repeated twice, making up 60 "different" pairs. Six same tone pairs (T1-T1, T2-T2, T3-T3 etc.) were repeated 10 times to make up 60 "same" trials in order to balance the number of "same" and "different" trials, which gave a total of 120 trials in the discrimination test. Participants were asked to discriminate whether the two syllables were of the same or different tones by pressing the left (same) or right (different) arrow. No feedback was given.

For both discrimination and identification tasks, a short practice session was given before the first set of stimuli. Participants were allowed to take a break every 20 trials in both tasks. The responses of participants were recorded and coded. Response time was also collected.

## Results

In both discrimination and identification tasks, for the accuracy analysis, mixed-effects logistic regression models were conducted, with the response to each trial as the input, training (pretest, posttest) and voice similarity (similar, dissimilar) as two fixed effects, and subjects as a random effect. For the response time analysis, two-way repeated measures ANOVAs were conducted, with the response time as the dependent variable and training (pretest, posttest) and voice similarity (similar, dissimilar) as independent variables. Figures 2 and 3 showed the mean accuracy of all participants achieved in identification and discrimination tasks respectively, in pretest and posttest for similar and dissimilar scenarios. Figure 4 and 5 showed the mean response times of all participants in identification and discrimination task.

For the identification accuracy, mixed-effects logistic regression model revealed significant main effect of training $(\chi 2(1)=315.81, p<0.001)$, and significant two-way interaction between training and voice similarity $(\chi 2(2)=$ 5.185, $p<0.05$ ), while the effect of voice similarity alone was insignificant $(\chi 2(1)=2.267, p=0.132)$. Post-hoc tests showed that the accuracy in the posttest was significantly higher than the pretest in both similar $(z=-13.936, p<$ 0.001 ) and dissimilar voices scenarios $(z=-10.962, p<$ 0.001 ). Within the pretest, there was no significant difference between similar and dissimilar voices $(z=-0.389$, $p=0.697$ ). However, in posttest, the accuracy in the similar voices scenario was significantly higher than that in dissimilar voices scenario $(z=2.713, p<0.01)$.

For the response time in the identification task, there were significant main effects of training $(F(1,11)=20.471, p=$ $0.001)$, voice similarity $(F(1,11)=25.73, p<0.001)$, as well as significant two-way interactions between training and voice similarity $(F(1,11)=8.552, p=0.014)$. Independent sample $t$-tests were then conducted within pretest and posttest to test the effects of speaker similarity.

The results showed that in the pretest, response time in the similar voice scenario was marginally significantly longer than the dissimilar voice scenario $(t(22)=1.899, p=0.071)$, but the difference was not significant in the posttest $(t(22)=$ $0.563, p=0.579$ ), implying that training improved the response time in the similar voice condition. Within the similar speaker condition, the response time in the posttest was significantly shorter than the pretest $(t(22)=2.708, p=$ 0.013 , suggesting the effects of training. While in the dissimilar voice condition, there was no significant difference between the pretest and posttest $(t(22)=1.479, p$ $=0.153$ ), indicating that training had very little impact on the response time in the dissimilar voice scenario.


Figure 2: Mean accuracy in the identification task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.


Figure 3: Mean accuracy in the discrimination task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.

In discrimination tasks, mixed-effects logistic regression model revealed a significant two-way interaction between training and voice similarity $(\chi 2(2)=4.362, p<0.05)$, while the effects of training $(\chi 2(1)=2.875, p=0.090)$ and voice similarity $(\chi 2(1)=1.110, p=0.292)$ were not significant. Post hoc tests showed that accuracy in pretest was significantly higher than that in posttest ( $z=2.665, p<0.01$ ) with dissimilar voices. However, no significant difference
was found between pretest and posttest in the similar voices scenario ( $z=-0.333, p=0.740$ ). The accuracy in similar voices scenario was significantly higher than that in dissimilar voices scenario within posttest $(z=2.197, p<$ 0.05 ), while the difference in accuracy with similar and dissimilar voices was found to be insignificant within pretest ( $z=-0.809, p=0.419$ ).
For the response time in the discrimination task, no effects were significant.


Figure 4: Mean response time in the identification task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.


Figure 5: Mean response time in the discrimination task averaged in the pre- and post-test, under similar talker scenario and dissimilar talker scenario.

## Discussion

The current study provided HVPT to Mandarin speakers who had minimal exposure to Cantonese prior to the study. The generalization to novel speakers who had similar and dissimilar voices to the training stimuli was investigated to explore the role of talker voice similarity in the learning of Cantonese tones and henceforth to shed light on the nature of mental representation of Cantonese tones in L2 learners. If Cantonese tones were encoded as an abstract representation, no significant difference would be found in the generalization to similar and dissimilar novel voices. On the other hand, if Cantonese tones were encoded as an
episodic representation, generalization to similar voices would be significantly better than to dissimilar voices.

We found that in the identification task, the effect of training was transferred successfully to both the similar and dissimilar voice scenarios. This finding fell in line with previous studies in that exposing the listeners to multiple talkers and phonetic contexts would enhance the phonemic categorization of the trained speech sounds. Miller, Zhang \& Nelson (2016) investigated whether adult listeners who became deaf postlingually and had cochlear implant (CI) could benefit from multiple-talker category identification training. They found that the perception performance was significantly improved for the CI listeners for the familiar talkers (i.e., talkers used in the training sessions) and also generalized to the unfamiliar talkers (i.e., talkers not included in the training sessions). There was also evidence that talker variation aids young infants' phonotactic learning (Seidl, Onishi \& Cristia, 2014). Together with these previous studies, our study provided extra evidence that the multiple-talker training was highly successful in learning to categorize speech sounds.

In both discrimination and identification posttests, the effect of voice similarity alone was significant. However, significantly better performance was found with similar voices than dissimilar voices in posttest but not in pretest and hence significant interaction between training and voice similarity was revealed. These findings echoed with previous studies that the speech from familiar voices was easier to identify than unfamiliar voices (Nygaard and Pisoni, 1998; Souza et al., 2013), supporting the hypothesis that Cantonese tones were encoded episodically in the mental representation of Mandarin L2 speakers. L2 learners retained perceptual details, speakers' voice characteristics, when encoding the tonal information into their mental representation. These perceptual details were integrated into later perception when the learner encountered with novel speakers. It is likely that the acoustic/phonetic representations are stored during the training stage, which facilitates the identification of tones in the similar voice context. In contrast, no matching representation is available in the dissimilar voice scenario, and thus the identity of the tone has to be construed from scratch.

As mentioned in the result, a significant training effect was found in identification tasks but not discrimination tasks. It is probably due to ceiling performance in discrimination tasks, for the reason that speakers without knowledge of Cantonese tones could also discriminate different tones merely by relying on psychoacoustic differences of the stimuli (Qin \& Mok, 2011). In addition, since Mandarin is a tonal language, participants already had some tonal categories in their mental representation within their L1, although it is not as complex as Cantonese categories. Therefore, participants could make use of psychoacoustic differences and their L1 knowledge to tell apart perceptually different tones in discrimination tasks even they had no knowledge of Cantonese tones. Moreover, since only the syllable / $\mathrm{ji} /$ was used in the discrimination
task, lack of syllable variability also reduced the cognitive loading of the tasks.

It should be noted that our study focused on L2 learners with limited Cantonese exposure. The episodic encoding of lexical tones might paly a role in early stage of learning, when a new speaker' voice counts as a distinct learning episode. It is unclear whether the episodic representation of L2 phonetic categories will change in late stages of learning. Future studies may include the experienced L2 learners to test the scope of the episodic learning.

## Conclusion

In the present study, perceptual trainings were given to native Mandarin speakers who had minimal exposure to Cantonese, and the voice similarity among the talkers in training and test phases was controlled. We found that the HVPT was highly effective, for the training effects generalized to both similar and dissimilar voices. However, the degree of generalization was significantly different between similar and dissimilar voices, which supported the hypothesis that learners retained speaker information during learning and made use of such information in future tone perception. This implies that newly learnt tones are encoded episodically in the mental representation of L2 learners. Future studies may explore the performance on the individual tones, so as to reveal the relationship between L1 and L2.

## Acknowledgments

This paper was supported partly by grants to the last author from the Research Grants Council of Hong Kong (ECS: 25603916), and the PolyU Start-up Fund for New Recruits (Project account no.: 1-ZE4Y). We thank Kelly Kwong for the help with data collection.

## References

Bauer, R. S., \& Benedict, P. K. (1997). Modern Cantonese phonology. Berlin: Mouton de Gruyter.
Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., \& Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /1/: Long-term retention of learning in perception and production. Perception \& Psychophysics, 61(5), 977-985.
Bradlow, A. R., Nygaard, L. C., \& Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Perception \& psychophysics, 61(2), 206-219.
Chao, Y.R. (1930). A system of tone letters. Le Maitre Phonetique. 45: 24-27.
Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(5), 1166-1183.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105(2), 251-279.

Goldinger, S. D., Pisoni, D. B., \& Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17(1), 152-162.
Kapnoula, E. C., \& McMurray, B. (2016). Newly learned word forms are abstract and integrated immediately after acquisition. Psychonomic bulletin \& review, 23(2), 491499.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson \& J. W. Mullennix (eds.), Talker Variability in Speech Processing (pp. 145-66). San Diego: Academic Press.
Johnson, K. (2005). Speaker Normalization in Speech Perception. In Pisoni, D. B., \& Remez, R. E. (eds.), The handbook of speech perception (pp.363-389). Malden, MA: Blackwell Pub.
Miller, S. E., Zhang, Y., \& Nelson, P. B. (2016). Efficacy of Multiple-Talker Phonetic Identification Training in Postlingually Deafened Cochlear Implant Listeners. Journal of Speech, Language, and Hearing Research, 59(1), 90-98.
Moore, C. (1996). Speaker and rate normalization in the perception of lexical tone by Mandarin and English listeners. PhD Dissertation, Cornell University, Ithaca, NY.
Norman, J. (1988). Chinese. Cambridge University Press.
Nygaard, L. C., \& Pisoni, D. B. (1998). Talker-specific learning in speech perception. Perception \& Psychophysics, 60(3), 355-376.
Peterson, G. E., \& Barney, H. L. (1952). Control methods used in a study of the vowels. The Journal of the acoustical society of America, 24(2), 175-184.
Qin, Z. \& Mok, P. P. K. (2011). Discrimination of Cantonese tones by Mandarin, English and French speakers. In The Psycholinguistic Representation of Tone, 2011 (pp. 50-53). Hong Kong: Causal Production.
Seidl, A., Onishi, K. H., \& Cristia, A. (2014). Talker variation aids young infants' phonotactic learning. Language Learning and Development, 10(4), 297-307.
Strand, E. A. \& Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (ed.), Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld (pp. 14-26). Berlin: Mouton de Gruyter.
Souza, P., Gehani, N., Wright, R., \& McCloy, D. (2013). The advantage of knowing the talker. Journal of the American Academy of Audiology, 24(8), 689-700.
Wang, Y., Spence, M. M., Jongman, A., \& Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. The Journal of the Acoustical Society of America J. Acoust. Soc. Am., 106(6), 3649.
Zhang, C. \& Chen, S. (2016). Toward an Integrative Model of Talker Normalization. Journal of Experimental Psychology: Human Perception and Performance.

# A Spiking Neural Bayesian Model of Life Span Inference 

Sugandha Sharma (s72sharm@uwaterloo.ca)<br>Aaron R. Voelker (arvoelke@uwaterloo.ca)<br>Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Centre for Theoretical Neuroscience, University of Waterloo<br>Waterloo, ON, Canada, N2L 3G1


#### Abstract

In this paper, we present a spiking neural model of life span inference. Through this model, we explore the biological plausibility of performing Bayesian computations in the brain. Specifically, we address the issue of representing probability distributions using neural circuits and combining them in meaningful ways to perform inference. We show that applying these methods to the life span inference task matches human performance on this task better than an ideal Bayesian model due to the use of neuron tuning curves. We also describe potential ways in which humans might be generating the priors needed for this inference. This provides an initial step towards better understanding how Bayesian computations may be implemented in a biologically plausible neural network.


Keywords: Neural Engineering Framework; biologically plausible inference; neural bayesian model; expectation maximization

## Introduction

Computations performed by the nervous system are subject to uncertainty because of the influence of sensory, cellular, and synaptic noise. At the level of cognition, the models that the brain uses to interact with its environment must necessarily cope with missing and imperfect information about the world. Behavioral studies have confirmed that humans often account for uncertainty in a way that is nearly optimal in the Bayesian sense (i.e., "Bayes optimal") (Ma, Beck, Latham, \& Pouget, 2006)). This implies that (1) neural circuits must, at least implicitly, represent probability distributions, and (2) neural circuits must be able to effectively compute with these probability distributions in order to perform Bayesian inference near-optimally.

Probabilistic models based on Bayes' rule have been widely used for understanding human cognition including inference, parameter and structure learning (Jacobs \& Kruschke, 2011), and word learning (Xu \& Tenenbaum, 2007). However, most Bayesian models lack biological plausibility because it is unclear how these computations might be realized in the brain. In particular, these models rely on sophisticated computations including high-dimensional integration, precise multiplication, and large-scale structure representation, without the use of spiking neuron models to implement these necessary computations.

A biologically plausible Bayesian approach can provide us insights into the working of the brain at multiple levels of analysis (Eliasmith, 2013). Moreover, it can also help in making more accurate normative predictions about how the perceptual system combines prior knowledge with sensory observations, enabling more accurate interpretations of data from psychological experiments (Doya, 2007). And finally, it
can point the way towards approximate Bayesian algorithms that are efficiently implemented in a neural substrate.

Griffiths, Chater, Norris, and Pouget (2012) conclude that different theoretical frameworks, such as Bayesian modeling and connectionism, have different insights to offer about human cognition, distributed across different levels of analysis. Here we make an initial attempt towards integrating these frameworks. We explore the biological plausibility of Bayesian inference by implementing a neural model of a life span prediction task using the Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003). We answer questions about how probability distributions can be represented in a connectionist framework using spiking neurons, and how the neural representations of these probability distributions can be used in meaningful ways. The next section describes the life span prediction task which we use.

## Life span prediction task

Griffiths and Tenenbaum (2006) evaluate how cognitive judgments compare with optimal statistical inference by asking people to predict human life spans. A group of 208 people were asked to predict human life spans, after being presented by a question in survey format as given below:
"Insurance agencies employ actuaries to make predictions about people's life spans - the age at which they will die based upon demographic information. If you were assessing an insurance case for an 18-year-old man, what would you predict for his life span?"

The responses were recorded and compared with the predictions made by a Bayesian model.

## Bayesian model

If $t_{\text {total }}$ indicates the total amount of time the person will live and $t$ indicates his current age, the task is to estimate $t_{\text {total }}$ from $t$. The Bayesian model computes a probability distribution over $t_{\text {total }}$ given $t$, by applying Bayes' rule:

$$
\begin{equation*}
p\left(t_{\text {total }} \mid t\right)=p\left(t \mid t_{\text {total }}\right) p\left(t_{\text {total }}\right) / p(t) \tag{1}
\end{equation*}
$$

where:

$$
\begin{equation*}
p(t)=\int_{0}^{\infty} p\left(t \mid t_{t o t a l}\right) p\left(t_{\text {total }}\right) d t_{\text {total }} \tag{2}
\end{equation*}
$$

We assume that the maximum age is 120 years. Thus, when calculating $p(t)$ in practice, the integral may be computed from 0 to 120.

Prior Griffiths and Tenenbaum (2006) use publicly available real-world data to identify the true prior distribution $p\left(t_{\text {total }}\right)$ over life spans (shown in Figure 1A).

Likelihood The likelihood $p\left(t \mid t_{\text {total }}\right)$ is the probability of encountering a person at age $t$ given that their total life span is $t_{\text {total }}$. Griffiths and Tenenbaum (2006) assume for simplicity that we are equally likely to meet a person at any point in his or her life. As a result, this probability is uniform, $p\left(t \mid t_{\text {total }}\right)=$ $1 / t_{\text {total }}$, for all $t<t_{\text {total }}$ (and 0 for $t \geq t_{t o t a l}$ ).

Prediction function Combining the prior with the likelihood according to Equation 1 yields a probability distribution $p\left(t_{\text {total }} \mid t\right)$ over all possible life spans $t_{\text {total }}$ for a person encountered at age $t$. As is standard in Bayesian prediction, Griffiths and Tenenbaum (2006) use the median of this distribution-the point at which it is equally likely that the true life span is either longer or shorter-as the estimate for $t_{t o t a l}$. This identifies a prediction function that specifies a predicted value of $t_{\text {total }}$ for each observed value of $t$.

Results Results obtained by Griffiths and Tenenbaum (2006) through this Bayesian model are shown in Figure 1B.


Figure 1: (A) Empirical distribution of the total life span $t_{\text {total }}$. (B) Participants' predicted values of $t_{\text {total }}$ for a single observed sample $t$. Black dots show the participants' median predictions for $t_{\text {total }}$. Solid line shows the optimal Bayesian predictions based on the empirical prior distribution shown in A. Dotted lines show predictions based on a fixed uninformative prior. Note: the fit between the human predictions (black dots) and Bayesian predictions (solid line) looks spot on in this figure due to the compressed y-axis, but Figure 4b shows a zoomed version revealing that this is not the case. Adapted from Griffiths and Tenenbaum (2006).

## Neural Engineering Framework

The Neural Engineering Framework (NEF) is based on three principles-representation, transformation, and dynamics-
which are used to construct large-scale neural models. The first two principles are described in the following sections. The principle of representation also describes how probability distributions can be represented using spiking neurons. The third principle is not required for this paper, and its details can be found elsewhere (Eliasmith \& Anderson, 2003).

## Principle 1 - Representation

In the NEF, information is represented as time-varying vectors of real numbers by populations of neurons. We say that a population of neurons has activities $a_{i}(\mathbf{x})$, which encode an $n$ dimensional stimulus vector, $\mathbf{x}=\left[x_{1}, x_{2}, \ldots, x_{n}\right]$, by defining the encoding:

$$
\begin{equation*}
a_{i}(\mathbf{x})=G_{i}\left[J_{i}(\mathbf{x})\right], \tag{3}
\end{equation*}
$$

where $G_{i}[\cdot]$ is the nonlinear transfer function describing the neuron's spiking response, and $J_{i}(\mathbf{x})$ is the current entering the soma of the neuron. For the purpose of our model, we have chosen $G_{i}[\cdot]$ to be the leaky integrate-and-fire (LIF) neuron model. The soma current is defined by:

$$
\begin{equation*}
J_{i}(\mathbf{x})=\alpha_{i}\left\langle\mathbf{e}_{i}, \mathbf{x}\right\rangle_{n}+J_{i}^{\text {bias }} \tag{4}
\end{equation*}
$$

where $J_{i}(\mathbf{x})$ is the current in the soma, $\alpha_{i}$ is a gain and conversion factor, $\mathbf{x}$ is the stimulus vector to be encoded, $\mathbf{e}_{i}$ is the encoding vector which corresponds to the "preferred stimulus" of the neuron-consistent with the standard idea of a preferred direction vector (Schwartz, Kettner, \& Georgopoulos, 1988)—and $J_{i}^{\text {bias }}$ is a bias current that accounts for background activity. The notation $\langle\cdot, \cdot\rangle_{n}$ indicates an $n$ dimensional dot-product.

Given this encoding, the original stimulus vector can be estimated by decoding those activities as follows:

$$
\begin{equation*}
\hat{\mathbf{x}}=\sum_{i} a_{i}(\mathbf{x}) \mathbf{d}_{i} . \tag{5}
\end{equation*}
$$

The decoding vectors $\mathbf{d}_{i}$ (also known as "representational decoders") are typically found in the NEF by least-squares optimization, which we use here (Eliasmith \& Anderson, 2003). Thus, the decoders resulting from this optimization complete the definition of a population code over a set of neurons $i$ for the representation of $\mathbf{x}$. The code is defined by the combination of nonlinear encoding in Eq. 3 and weighted linear decoding in Eq. 5.
Temporal representation The population code does not explicitly address the issue of how information is encoded over time. To do so, we can begin by considering the temporal code for each neuron in isolation by taking the neural activities to be filtered spike trains as shown in Eq. 6:

$$
\begin{equation*}
a_{i}(t)=\sum_{m} h_{i}(t) * \delta\left(t-t_{m}\right)=\sum_{m} h_{i}\left(t-t_{m}\right), \tag{6}
\end{equation*}
$$

where $\delta_{i}(\cdot)$ are the spikes at time $t_{m}$ for a given neuron $i$, generated by $G_{i}[\cdot]$ and $h_{i}(t)$ are the linear decoding filters. We can compute the optimal filters for decoding using the NEF, however to make our model biologically plausible, we
have chosen these filters $\left(h_{i}(t)\right)$ to be the postsynaptic currents (PSCs) induced in subsequent neuron by the arrival of a spike. Eliasmith and Anderson (2003) have shown that this assumption causes minimal information loss which can be further reduced by increasing the population size.

This temporal code can be combined with the population code defined before (Eqs. 3, 4, 5), to provide a general population temporal code for vectors. The encoding and decoding equations for such a code are given by Eq. 7 and Eq. 8:

$$
\begin{align*}
\delta\left(t-t_{i m}\right) & =G_{i}\left[\alpha_{i}\left\langle\mathbf{e}_{i}, \mathbf{x}\right\rangle_{n}+J_{i}^{\text {bias }}\right]  \tag{7}\\
\hat{\mathbf{x}} & =\sum_{i, m} h_{i}\left(t-t_{m}\right) \mathbf{d}_{i} . \tag{8}
\end{align*}
$$

Representing probability distributions Probability distributions are essentially functions of some parameters. Having described how to represent vectors using the NEF, we consider the relationship between vector and function representation. For any representation, we need to specify the domain of that representation. In case of vectors, the domain is the subspace of the vector space that is represented by the neurons (e.g., the $\mathbf{x}$ vector). We define the relevant function domain by parameterizing the set of represented functions by an $n$ dimensional vector of coefficients $\mathbf{k}=\left[k_{1}, k_{2}, \ldots, k_{n}\right]$. These define any function of interest over a fixed set of basis functions $\phi(v)$ as follows:

$$
\begin{equation*}
\mathbf{x}(v ; \mathbf{k})=\sum_{j=1}^{n} k_{j} \phi_{j}(v), \quad \text { for } k \sim p(\mathbf{k}) . \tag{9}
\end{equation*}
$$

Thus we define a particular probability distribution $p(\mathbf{k})$ by limiting the space spanned by the basis $\phi(v)$ to some subspace of interest depending on the application. This is also the domain over which the optimization to find the decoders in Eq. 5 is performed.

Next, we define population encoding and decoding analogous to that in Eqs 3 and 5 for functions:

$$
\begin{align*}
a_{i}(\mathbf{x}(v ; \mathbf{k}))=a_{i}(\mathbf{k}) & =G_{i}\left[\alpha_{i}\left\langle\mathbf{e}_{i}(v), \mathbf{x}(v ; \mathbf{k})\right\rangle_{n}+J_{i}^{\text {bias }}\right]  \tag{10}\\
\hat{\mathbf{x}}(v ; \mathbf{k}) & =\sum_{i} a_{i}(\mathbf{k}) \mathbf{d}_{i}(v), \tag{11}
\end{align*}
$$

where $\mathbf{e}_{i}(v)$ and $\mathbf{d}_{i}(v)$ are the encoding and decoding functions of the neurons. We project these functions onto the same basis $\phi(v)$ used to identify the function space. For simplicity, we assume that $\phi(v)$ is an orthonormal basis - an analogous derivation for a bi-orthonormal set can be found elsewhere (Eliasmith \& Martens, 2011). Hence, we get the following encoding and decoding functions:

$$
\begin{align*}
& \mathbf{e}_{i}(v)=\sum_{j=1}^{n} e_{i j} \phi_{j}(v)  \tag{12}\\
& \mathbf{d}_{i}(v)=\sum_{j=1}^{n} d_{i j} \phi_{j}(v), \tag{13}
\end{align*}
$$

where $e_{i j}$ and $d_{i j}$ identify the $n$ coefficients that represent
the encoding and decoding functions in $\phi(v)$ basis for each neuron. We now substitute these into Eq 10 :

$$
\begin{align*}
a_{i}(\mathbf{x}(v ; \mathbf{k})) & =G_{i}\left[\alpha_{i}\left(\sum_{m, n} k_{n} \phi_{n}(v) e_{i m} \phi_{m}(v)\right)+J_{i}^{\text {bias }}\right] \\
& =G_{i}\left[\alpha_{i}\left(\sum_{m, n} k_{n} e_{i m} \delta_{m n}\right)+J_{i}^{\text {bias }}\right]  \tag{14}\\
& =G_{i}\left[\alpha_{i}\left(\sum_{n} k_{n} e_{i n}\right)+J_{i}^{\text {bias }}\right] \\
& =G_{i}\left[\alpha_{i}\left\langle\mathbf{e}_{i}, \mathbf{k}\right\rangle_{n}+J_{i}^{\text {bias }}\right] .
\end{align*}
$$

This way, function encoding is expressed as vector encoding identical to Eq. 7. Similarly, function decoding can also be expressed as vector decoding as follows:

$$
\begin{equation*}
\hat{\mathbf{k}}=\sum_{i} a_{i}(\mathbf{k}) \mathbf{d}_{i} . \tag{15}
\end{equation*}
$$

To summarize, we have shown that it is mathematically equivalent to talk in terms of (finite-dimensional) function spaces or (finite-dimensional) vector spaces. Since probability distributions are most generally functions, we can approximate them as high-dimensional vectors over a fixed set of basis functions using the NEF.

## Principle 2 - Transformation

Transformations of neural representations are functions of the vector variables represented by neural populations.

To perform a transformation $f(\mathbf{x})$ in the NEF, instead of finding the representational decoders $\mathbf{d}_{i}$ to extract the originally encoded variable $\mathbf{x}$, we can re-weight the decoding to specify some function $f(\mathbf{x})$ other than identity. In other words, we can find the decoders $\mathbf{d}_{i}^{f(\mathbf{x})}$ (also known as "transformational decoders") by using least-squares optimization to minimize the difference between the decoded estimate of $f(\mathbf{x})$ and the actual $f(\mathbf{x})$, which results in the transformation:

$$
\begin{equation*}
\hat{\mathbf{x}}=\sum_{i} a_{i}(\mathbf{x}) \mathbf{d}_{i}^{f(\mathbf{x})} \tag{16}
\end{equation*}
$$

Both linear and nonlinear functions of the encoded vector variable can be computed in this manner (Eliasmith \& Anderson, 2003). In the NEF, connection weights between neurons can be defined in terms of encoders and decoders as: $\omega_{i j}=\alpha_{j} \mathbf{e}_{j} \mathbf{d}_{i}^{f(\mathbf{x})}$, where $i$ indexes the presynaptic population, $j$ indexes the postsynaptic population, and $\mathbf{d}_{i}^{f(\mathbf{x})}$ are representational or transformational decoders.

## Neural model of life span prediction

Figure 2 shows the architecture of the neural model for life span inference built using the NEF. All neural ensembles (populations of neurons; symbolically represented by five circles) are 20 dimensional and contain 200 LIF neurons each, except the Normalized Posterior ensemble which is 120 dimensional and contains 800 LIF neurons. The


Figure 2: A schematic diagram of the neural model. Here "Likelihood" and "Prior" contain 200 neurons each, "Product" network contains 4000 neurons and "Normalized Posterior" contains 800 neurons.
product network computes an element-wise product of its inputs. Though multiplication is nonlinear, it has a wellcharacterized implementation in neurons that does not require nonlinear interactions, and can be implemented accurately with the NEF (Gosmann, 2015). The product network makes use of this characterization. It has 40 neural ensembles of 100 neurons each for a total of 4,000 neurons. The entire model contains 5,200 neurons.

To represent the probability distributions (prior and likelihood) needed to perform the task, we define a basis $\phi_{20}(v)$ to span the space of each distribution. To compute the basis we sample from a family of 120 dimensional distributions and do Singular Value Decomposition to obtain a 20 dimensional basis. This basis is used to determine the encoders (as given by Eq. 12) used in the NEF simulation. The same basis is used for the optimization to find the neuron decoders (as given by Eq. 13) that are needed to perform the desired computations. Similar to the encoding and decoding functions, the 120 dimensional prior and likelihood functions are also projected to the 20 dimensional space through weights over the basis. Refer to the supplemental material for details.

The likelihood_input and prior_input are nodes that provide the named 20 dimensional inputs to the neural ensembles Likelihood and Prior respectively. The product network receives input from these ensembles and computes the posterior distribution (in the 20 dimensional space). The output connection from product network to Normalized Posterior reconstructs the posterior back to 120 dimensional space and computes the normalization function using principle 2 of the NEF. Thus, the Normalized Posterior ensemble represents the normalized posterior distribution. Next we approximate the median of this distribution on the connection between the Normalized Posterior ensemble and the Prediction node (again using principle 2). We read out the model prediction from the Prediction node.

Figure 3 shows the inference results obtained from the spiking neural network run in the Nengo (Bekolay et al., 2014) software package. Model predictions are plotted for current ages $(t)$ from 1 to 100 . The difference between the results in Direct mode and Neuron mode is due to the limited number of neurons in the Normalized Posterior ensemble. As the number of neurons in this ensemble increases, the results


Figure 3: Inference results from neural model ( $95 \%$ confidence intervals), compared to humans and Direct mode our model with computations in low-dimensional ( 20 dimensional basis) space, but without neurons.
approach the Direct mode results ( 800 neurons provide the best fit to human data). Thus, neural results match the human data better due to the approximate representation of the normalized posterior by the neurons in the Normalized Posterior ensemble. The tuning curves of the neurons in this ensemble were fit to a function space consisting of a family of distributions which have three parameters (similar to the parameters in the prior) and also depend on the current age $(t)$ (similar to the likelihood function). The three parameters: $a$ - the skewness parameter was varied from -7 to -4 , scale - used to scale the distribution was varied from 26 to 29 and $l o c$ - used to shift the distribution was varied between 49 to 101 . The current age $(t)$ was varied in the range of $+/-5$ for a given age in a trial except ages below 5 for which the range was taken to be from $[1,10]$. This provides the function space that was used to sample the encoders for Normalized Posterior ensemble.

We use the Kolmogorov-Smirnov (K-S) test to examine the goodness of fit of the neural model predictions relative to the Griffiths and Tenenbaum (2006) model. The data used for the K-S test are shown in Figure 4b. The dissimilarity of the Griffiths and Tenenbaum (2006) model relative to human predictions is 9.628 , while that of the neural model is 1.959 , indicating the much closer fit of the neural model to the human data. Figure 4a shows a comparison between the Griffiths and Tenenbaum (2006) model, the computational model (our replication of their model), and direct mode (our model with computations in a compressed 20 dimensional space, but without neurons). Since the results obtained from the direct mode are the same as the computational model, the low dimensional embedding is not losing any information. However, we expect some error due to this constraint for more complex priors (though we have not explored the minimum dimensionality for this prior).

Overall, our results suggest that the closer fit of the neural data can be solely attributed to fitting the neuron tuning


Figure 4: (a) Results from Griffiths and Tenenbaum (2006) model (only data corresponding to human data), Computational model i.e., our replication of their model, and Direct mode i.e., our model with computations in low-dimensional space, but without neurons. (b) Kolmogorov-Smirnov (K-S) test results. Dissimilarity relative to human predictions - Griffiths and Tenenbaum (2006) model: 9.628, neural model: 1.959. Neural model and human data are median predictions. Note: Griffiths and Tenenbaum (2006) model data and human data were obtained from Figure 1B through a web plot digitizer.
curves in the Normalized Posterior ensemble, where 800 neurons provide the best match to human performance. Since the low-dimensional neural implementation can be made to match the human data, this is some evidence in support of the hypothesis that human brains represent low-dimensional state spaces (low-dimensional parameterizations of highdimensional distributions fit using neural tuning curves).

## Generalized life span inference

In our neural model, we use the prior obtained empirically by Griffiths and Tenenbaum (2006). However, our neural modeling methods can further be used to explore how this prior might be learned in the human brain. Here, we lay some theoretical ground work for addressing this question, while building the complete neural model remains for future work.

We assume that priors that humans have about life spans are a result of their experiences encountering people of different ages in their daily lives. Thus the prior will be inferred from the data that comes from daily experience. We further assume that the prior is parameterized by some unknown hyperparameters ( $\alpha$ ) which are to be estimated from the observed ages of $n$ distinct people, given by $\mathbf{X}=\left\{x_{1}, \ldots, x_{n}\right\}$. Here, each random variable $x_{i}$ corresponds to a separate $t$ from the previous model. Likewise, we model each element of $\mathbf{X}$ as being drawn independently from each element of $\mathbf{Z}=\left\{z_{1}, \ldots, z_{n}\right\}$ corresponding to the (unknown or hidden) life spans of these same $n$ people. Each random variable $z_{i}$ corresponds to a separate $t_{\text {total }}$ from the previous model, which in turn is drawn from the unknown prior. We now describe two standard methods for determining a prior by obtaining an estimate $\hat{\alpha}$ of the hyperparameters.

If we do not know the actual prior, then the optimal solution can be found by trying them all. That is, we directly find the hyperparameters $\hat{\alpha}$ that maximize the marginal like-
lihood of the observed data, $L(\alpha ; \mathbf{X})$ (or equivalently the loglikelihood for numerical stability):

$$
\begin{equation*}
\hat{\alpha}=\operatorname{argmax}_{\alpha} L(\alpha ; \mathbf{X})=\operatorname{argmax}_{\alpha} \sum_{i=1}^{n} \log \sum_{z_{i}} p\left(x_{i}, z_{i} \mid \alpha\right) \tag{17}
\end{equation*}
$$

In general, however, the procedure described above is intractable, since it requires that we iterate over all combinations of $\alpha$ and $\mathbf{Z}$. This motivates near-optimal iterative procedures such as the widely-used expectation maximization algorithm (EM; Dempster, Laird, \& Rubin, 1977). Below we work out the details of the EM procedure for the case where the hyperparameters are $\alpha=\left(\mu, \sigma^{2}\right)$, i.e., the prior is assumed to be normally distributed with unknown moments. We begin by simplifying the expectation function using independence and other known facts about the model:

$$
\begin{align*}
Q\left(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^{(t)}\right) & =E_{\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\alpha}^{(t)}}[\log L(\boldsymbol{\alpha} ; \mathbf{X}, \mathbf{Z})] \\
& =\sum_{i=1}^{n} E_{z_{i} \mid x_{i}, \boldsymbol{\alpha}^{(t)}}\left[\log L\left(\alpha ; x_{i}, z_{i}\right)\right]  \tag{18}\\
& =\sum_{i=1}^{n} \sum_{z_{i}} T\left(x_{i}, z_{i}\right) \log \left(p\left(z_{i} \mid \alpha\right) / z_{i}\right)
\end{align*}
$$

where we have defined $T\left(x_{i}, z_{i}\right):=p\left(z_{i} \mid x_{i}, \alpha^{(t)}\right)$ to be some fixed function with respect to $\alpha^{(t)}$. Next, we simplify the log expression using our model of the prior:

$$
\begin{align*}
& \log \left(p\left(z_{i} \mid \alpha\right) / z_{i}\right)=\log \left(\frac{1}{\sqrt{2 \sigma^{2} \pi}} e^{-\frac{\left(z_{i}-\mu\right)^{2}}{2 \sigma^{2}}}\right)-\log z_{i}  \tag{19}\\
& =-\frac{1}{2}\left(\left(z_{i}-\mu\right)^{2} / \sigma^{2}+\log \sigma^{2}+\log (2 \pi)+2 \log z_{i}\right)
\end{align*}
$$

and then differentiate this with respect to $\mu$ :

$$
\begin{equation*}
\frac{\partial \log \left(p\left(z_{i} \mid \alpha\right) / z_{i}\right)}{\partial \mu}=\left(z_{i}-\mu\right) \sigma^{2} \tag{20}
\end{equation*}
$$

and with respect to $\sigma^{2}$ :

$$
\begin{equation*}
\frac{\partial \log \left(p\left(z_{i} \mid \alpha\right) / z_{i}\right)}{\partial \sigma^{2}}=\frac{1}{2}\left(\left(z_{i}-\mu\right)^{2}-\sigma^{2}\right) / \sigma^{4} . \tag{21}
\end{equation*}
$$

By linearity of differentiation, we then know that the derivatives of $Q(\cdot)$ are zero when:

$$
\begin{align*}
\frac{\partial Q\left(\alpha \mid \alpha^{(t)}\right)}{d \mu} & =\sum_{i=1}^{n} \sum_{z_{i}} T\left(x_{i}, z_{i}\right)\left(z_{i}-\mu\right) \sigma^{2}=0 \\
\Longleftrightarrow \quad \mu & =\frac{\sum_{i=1}^{n} \sum_{z_{i}} z_{i} T\left(x_{i}, z_{i}\right)}{\sum_{i=1}^{n} \sum_{z_{i}} T\left(x_{i}, z_{i}\right)}, \quad \text { and similarly: }  \tag{22}\\
\frac{\partial Q\left(\alpha \mid \alpha^{(t)}\right)}{d \sigma^{2}} & =\sum_{i=1}^{n} \sum_{z_{i}} T\left(x_{i}, z_{i}\right) \frac{1}{2}\left(\left(z_{i}-\mu\right)^{2}-\sigma^{2}\right) / \sigma^{4}=0 \\
\Longleftrightarrow \quad \sigma^{2} & =\frac{\sum_{i=1}^{n} \sum_{z_{i}}\left(z_{i}-\mu\right)^{2} T\left(x_{i}, z_{i}\right)}{\sum_{i=1}^{n} \sum_{z_{i}} T\left(x_{i}, z_{i}\right)} \tag{23}
\end{align*}
$$

Finally, by the generalized Bayes' rule, we know:

$$
T\left(x_{i}, z_{i}\right)=p\left(z_{i} \mid x_{i}, \boldsymbol{\alpha}^{(t)}\right)=\frac{p\left(z_{i}, x_{i} \mid \boldsymbol{\alpha}^{(t)}\right)}{\sum_{z_{i}} p\left(z_{i}, x_{i} \mid \boldsymbol{\alpha}^{(t)}\right)},
$$

which we may compute via Eq. 1. We also note that since $T(\cdot)$ is a probability density function over $z_{i}$, that:

$$
\sum_{i=1}^{n} \sum_{z_{i}} T\left(x_{i}, z_{i}\right)=\sum_{i=1}^{n} 1=n
$$

Therefore, each EM iteration must make the update $\alpha^{(t+1)}=$ $\left(\mu^{(t+1)}, \sigma^{(t+1)}\right)$, where:

$$
\begin{align*}
\mu^{(t+1)} & =\frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{z_{i}} z_{i} p\left(z_{i}, x_{i} \mid \alpha^{(t)}\right)}{\sum_{z_{i}} p\left(z_{i}, x_{i} \mid \alpha^{(t)}\right)}  \tag{24}\\
\boldsymbol{\sigma}^{(t+1)} & =\sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{z_{i}}\left(z_{i}-\mu^{(t+1)}\right)^{2} p\left(z_{i}, x_{i} \mid \boldsymbol{\alpha}^{(t)}\right)}{\sum_{z_{i}} p\left(z_{i}, x_{i} \mid \boldsymbol{\alpha}^{(t)}\right)}} .
\end{align*}
$$

This converges to some locally optimal estimate of the hyperparameters. For initial $\alpha^{(0)}$ chosen sufficiently close to global optimum $\hat{\alpha}$ given by Eq. 17, this converges to the optimum.

This provides a tractable procedure for updating the prior. In particular, we begin with some initial guess at the hyperparameters, and then update them iteratively to better explain the observed data. In practice only a few iterations are required (results not shown). Once we have an estimate of the hyperparameters $(\hat{\alpha})$, we then know the prior $p\left(t_{\text {total }} \mid \hat{\alpha}\right)$. This prior can be used directly by the previously described model to provide a good prediction. In fact, it is possible to run both the prior optimization and inference at the same time, and both will become progressively more accurate over time.

## Conclusions

We have presented a spiking neural network able to effectively perform Bayesian inference in a manner that more accurately matches human behavior than an ideal Bayesian
computation. We constructed the network using the NEF to map function spaces into vector spaces and approximate the necessary computations. We suggested a means of estimating the prior for the life span task that can be implemented using these same methods.
Notes Supplemental material (scripts and derivations) can be found at https://github.com/ctn-waterloo/cogsci17-infer.

## Acknowledgments

This work was supported by CFI and OIT infrastructure funding, the Canada Research Chairs program, NSERC Discovery grant 261453, ONR grant N000141310419, AFOSR grant FA8655-13-1-3084, OGS, and NSERC CGS-D.

## References

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., ... Eliasmith, C. (2014). Nengo: a Python tool for building large-scale functional brain models. Frontiers in neuroinformatics, 7, 48.
Dempster, A. P., Laird, N. M., \& Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.
Doya, K. (2007). Bayesian brain: Probabilistic approaches to neural coding. MIT press.
Eliasmith, C. (2013). How to build a brain: A neural architecture for biological cognition. Oxford University Press.
Eliasmith, C., \& Anderson, C. H. (2003). Neural engineering: Computation, representation, and dynamics in neurobiological systems. MIT press.
Eliasmith, C., \& Martens, J. (2011). Normalization for probabilistic inference with neurons. Biological cybernetics, 104(4), 251-262.
Gosmann, J. (2015). Precise multiplications with the NEF [Technical Report]. University of Waterloo, Waterloo, Ontario, Canada. Retrieved from http://dx.doi.org/10.5281/zenodo. 35680
Griffiths, T. L., Chater, N., Norris, D., \& Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on bowers and davis (2012).
Griffiths, T. L., \& Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. Psychological science, 17(9), 767-773.
Jacobs, R. A., \& Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. Wiley Interdisciplinary Reviews: Cognitive Science, 2(1), 8-21.
Ma, W. J., Beck, J. M., Latham, P. E., \& Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nature neuroscience, $9(11), 1432-1438$.
Schwartz, A. B., Kettner, R. E., \& Georgopoulos, A. P. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. I. Relations between single cell discharge and direction of movement. The Journal of Neuroscience, 8(8), 2913-2927.
Xu, F., \& Tenenbaum, J. B. (2007). Word learning as Bayesian inference. Psychological review, 114(2), 245.

# The Role of Causality in Temporal Binding: Evidence for an Intentional Boost 

Daniel A. Shiloh, Peter A. White \& Marc J. Buehner ([ShilohDA], [WhitePA], [BuehnerM]@Cardiff.ac.uk)<br>School of Psychology, Cardiff University, Tower Building, 70 Park Place<br>Cardiff, CF10 3AT, UK


#### Abstract

Temporal binding refers to the subjective contraction in time between an action and its consequence. Since it was reported in 2002 the effect has generated much interest, although a consensus regarding the mechanisms behind it remains elusive. While multiple theoretical accounts have been proposed, a key point of contention remains whether the effect is the result of the perception of intentionality or causality. We deployed a new apparatus to compare intentional to mechanical causation. Thirty participants reported the interval between two events in self-causal, mechanical-causal and non-causal conditions. The results of a Bayesian analysis pointed to smaller temporal estimates in the self-causal condition compared with the mechanical-causal condition, in addition to smaller estimates in the mechanicalcausal condition compared with the non-causal condition. The evidence presented here suggests that causality alone may be sufficient for temporal binding to occur, but that this effect is boosted by the presence of intentional action.


## Introduction

Temporal binding refers to the mutual attraction (in subjective time) between a causal action and its consequence, relative to two unrelated events. In a seminal paper Haggard et al (2002) found evidence for delayed awareness of the time of action and early awareness of the time of its consequence. Subsequent research has replicated this effect with a variety of paradigms, including interval estimation (Humphreys \& Buehner, 2009), stimulus anticipation (Buehner \& Humphreys, 2009) and the method of constant stimuli (Nolden, Haering \& Kiesel, 2012). The use of various interval estimation methods has demonstrated that, in addition to shifts in the perceived time of events, intentional binding also manifests as a shortening of the overall perceived interval between an action and its consequence.
Haggard et al. (2002) originally referred to the effect as intentional binding and proposed that it reflects "a general linkage through time between representation of action and effect" (p. 384), and that the subjective shortening of the interval between them may contribute to our sense of agency and motor learning through forward models. While multiple accounts of the mechanisms behind temporal binding have been proposed since then (Buehner, 2015; Moore \& Obhi, 2012; Eagleman \& Holcombe, 2002), the role of intentionality has been central to much of the work on the subject. More recently, studies have increasingly made use of temporal binding as an implicit measure of sense of agency, for example in studies of mindfulness (Jo,

Whittmann, Hinterberger \& Schmidt, 2015; Lush, Parkinson \& Dienes, 2016) narcissism (Hascalovitz \& Obhi, 2015) and Schizophrenia (Voss et al, 2010).

The focus on intentionality in the literature notwithstanding, Buehner and Humphreys (2009) have argued that temporal binding is instead driven by awareness of causality, and should be termed "causal binding": Temporal binding reflects a bi-directional interpretation of David Hume's (1739/1888) assertion that temporally contiguous events are more likely to be perceived as causally related. Specifically, because human time perception is inherently noisy and uncertain, it is subject to top-down modulation. From a Bayesian perspective it thus follows that if contiguous event pairs are likely to be causally related, then event pairings that are known to be causally linked are also likely to have occurred contiguously. Time and causality thus mutually constrain each other in subjective experience.

While there is a general consensus that a causal relationship is necessary for temporal binding to occur, there is less agreement on whether causality on its own is also sufficient (Moore \& Obhi, 2012). However, Buehner (2015) reported mutual attraction in subjective time between voluntary actions and their outcomes (i.e. the typical binding effect) as well as between involuntary, induced, causal actions and their outcomes. Furthermore, Buehner (2012) also demonstrated temporal binding between a nonbiological mechanical action (a robot arm pressing a key) and its outcome (an LED flash). While both studies revealed evidence for temporal binding in the presence of causality alone (thus demonstrating its sufficiency to result in binding), they also found a more pronounced effect when the cause was an intentional action. Thus, while causal binding appears to be rooted in causality, it seems to be subject to an intentional boost.

A limiting factor in this earlier research is that it always deployed key-presses as intentional causal actions, meaning that participants had access to precise proprioceptive feedback about the successful completion of the causal action, as well as the precise time of the start of the causal interval (i.e. the moment the key was depressed). In contrast, this type of feedback was not available in control conditions. We set out to maximize the perceptual similarity between experimental conditions. Specifically, we replaced key-presses with a continuous upwards movement made by the participant, and created a mechanical causal as well as a control condition that matched the perceptual experience. In all three conditions, participants were able to rely purely on
visual information to determine the onset of a two-event sequence, and we eliminated any tactile or auditory feedback.

Participants took part in three conditions: self-causal, mechanical-causal and non-causal control. On each trial, participants had to reproduce the interval between two sequential events (which were causally linked in the two causal conditions). Both the self-causal and mechanicalcausal conditions made use of a laser pointed at a light sensor. Upon detecting the laser beam (event 1) the light sensor responded by switching on a red LED after a randomized delay (event 2 ). In the self-causal condition participants allowed the laser to reach the light sensor manually by moving a wooden paddle out of its way, whereas in the mechanical-causal condition this was done mechanically, without input from the participant. In the noncausal control condition the laser was replaced with a small red LED which was positioned where the laser beam could be seen in the other two conditions. In this condition, the two event sequence consisted of deactivating of the small LED (simulating the perceptual experience of the laser hitting the light sensor), followed by the switching on of the red LED as in the other two conditions. This sequence was controlled by a computer.

According to the intentional binding account, temporal estimates for the two -event sequences should be smaller in the self-causal condition than in the other two (self-causal < mechanical-causal $=$ non-causal control); according to the causal binding account, temporal estimates should be larger in the non-causal control condition than in the self-causal and mechanical-causal conditions (self-causal = mechanical-causal $<$ non-causal control). Finally, if temporal binding is rooted in causality, but subject to an intentional boost temporal estimates should be lowest in the self-causal condition, followed by the mechanical-causal condition, with both being shorter than the non-causal control condition (self-causal $<$ mechanical-causal $<$ noncausal control).

## Methods

## Participants

Thirty Cardiff University students and staff (2 male, age range 18-33) participated in exchange for a payment of $£ 3$ or course credits. Participants were recruited through Cardiff University's electronic Experiment Management System and electronic noticeboard. Participants were asked to report (in writing) whether they felt they knew the purpose of the experiment prior to debriefing. Of the thirty, only three responded 'yes', and none correctly understood the purpose of the experiment.

## Apparatus

A schematic diagram of the apparatus can be seen in Figure 1. The apparatus was situated on top of a desk and placed on a platform at a height of 9.8 cm , with a gap 18.8 cm in length. The light sensor was positioned opposite the laser module,
both at a height of 14.5 cm . Between the laser and light sensor a wheel ( 21.5 cm diameter) was placed with a round 1 cm diameter hole positioned in the location through which the laser beam passed. The wheel was attached to a motor which allowed it to spin clockwise at a speed of approximately one revolution per four seconds.


Figure 1: a schematic diagram of the apparatus. $1=$ light sensor (containing a Raspberry Pi computer) connected to a red LED bulb; $2=$ wheel with 1 cm diameter perforation; $3=$ laser module; $4=$ a box housing a geared motor able to spin the wheel clockwise.

The light sensor module consisted of a $7 \times 7 \times 10 \mathrm{~cm}$ box housing a raspberry pi computer, with a 10 mm LED bulb mounted at its top and the light sensor on its front (facing the laser module). A separate, portable, 5 mm red LED bulb was also connected to the computer, but only visible to participants during the non-causal condition (see design and procedure). For the self-causal condition (see design and procedure), a rectangular wooden paddle ( 6 cm in width and 14 cm in height, with handle at its centre) with a 1 cm diameter hole was used in place of the wheel.

Participants were placed at a chin rest behind the laser module. Participant responses were recorded using a computer mouse on a separate computer. Finally, a debrief questionnaire was used to measure perceived causality using a 9-point Likert Scale (see appendix 1). For each condition, participants were presented with the question "in the condition where [condition description] did it seem like [first event] was causing [second event] ( $1=$ definitely yes, $5=$ not sure, $9=$ definitely no)?" These scores were inverted for analysis.

## Design and Procedure

After completing a consent form, participants were given safety instructions and were allowed to adjust the height of their seat. Instructions were presented verbally at the beginning of the experiment and before each trial. Throughout the experiment participants kept their head in the chin rest, ensuring that the light sensor, wheel and laser beam were visible. Participants were instructed to fixate their gaze on the laser point during the self-causal and mechanical-causal conditions, and on the 5 mm diameter LED bulb during the non-causal control condition.

Participants worked through the three conditions, with order of conditions counterbalanced between participants. Each condition consisted of 40 trials, during which
participants observed a critical two-event sequence lasting for an interval between 200 - 400ms (randomized, described below) and were asked to reproduce this interval by holding down the left mouse key for their perceived duration. Prior to each experimental block, participants worked through as many practice trials as they needed (minimum: three, regardless of performance) to understand the task. Task comprehension was assessed by the experimenter by observing the participants performing the task to ensure that participants were performing the correct movement (if any) and reporting time intervals after each trial. Probing questions were used to ensure that participants were reporting the correct time intervals and that they did not have any further questions.

The conditions were as follows (see Figure 2 for a photographs of each experimental condition):

Self-causal: Participants performed an intentional action that generated a causal consequence after a short delay. The wheel was placed with the hole aligned to the laser beam and light sensor and remained stationary throughout (i.e. the laser beam could pass through to the light sensor, when allowed through by the participant. The light sensor responded to the laser beam by switching on the 10 mm red LED at the top of the housing after a randomised delay of $200-400 \mathrm{~ms}$, and switching off after a randomised interval of $200-400 \mathrm{~ms}$, if the beam was no longer received. All randomised delays used in the experiment were drawn from a uniform distribution. Participants were told that the sensor responds to the beam after a delay, and this was demonstrated by the experimenter prior to the practice trials by using hand movements to either block the laser or allow it through. Participants were not told any additional information about these delays. Participants were instructed to place the paddle at the bottom of the apparatus, with the hole beneath the laser beam, such that the paddle blocked the beam. Participants were instructed to keep the paddle positioned adjacent to the wheel and move it upwards in front of the laser beam in each trial, such that the laser would pass through the hole. This was done to keep this condition as perceptually similar as possible to the mechanical-causal condition (see below). Participants were instructed to reproduce the time interval between the laser beam reaching the light sensor and the LED lighting up before placing the paddle back for the next trial.

Mechanical-causal: The wheel rotated continuously at a speed of approximately 4 seconds per revolution and blocked the laser beam from reaching the sensor, except when the hole came in line with it (once every 4 seconds). The light sensor was switched on and functioned in the same way as in the self-causal condition. This was demonstrated prior to the practice trials; the experimenter demonstrated that when the laser beam was blocked the light sensor did not respond at all, regardless of the position of the wheel, and that the light sensor always responded after the laser passed through the hole in the wheel. Participants were instructed to reproduce the interval between the laser reaching the sensor and the LED lighting
up as in the self-causal condition. Note that in both the selfcausal and mechanical-causal conditions, the critical causal event 1 (the laser reaching the light sensor) coincided with the perceptual experience of the laser spot (temporarily) being no longer visible against the paddle or wheel.

Non-causal control: Participants reproduced the interval between two sequential LED flashes. The wheel was positioned in the same way as in the self-causal condition. The laser module was switched off, and the 5 mm LED was placed in the hole in the wheel. At the beginning of each trial, the 5 mm LED switched on for one second before switching off, followed by the 10 mm LED at the top of the housing switching on for $200-400 \mathrm{~ms}$. Following this, participants were asked to reproduce the time interval between the 5 mm LED switching off and the 10 mm LED switching on. Participants were not told any information about the causal relationship between the two lights, but only that they turned on and off in a regular sequence. In order that the switching off of the first light would be equally predictable as the laser passing through the wheel in the mechanical-causal condition participants were informed that the first light will switch off after exactly one second on each trial. This sequence repeated automatically for the duration of the condition, with an overall trial length matching the duration of a single wheel revolution. Participants were instructed to fixate their gaze on the 5 mm LED bulb throughout.

At the end of the experiment participants were asked to fill in the debrief questionnaire, where they were asked to report whether they believed the first event in the interval they were judging caused the second event to occur, per condition. These causal ratings were taken as a manipulation check, to ensure participants correctly perceived the causal structure of the self-causal and mechanical-causal conditions (the laser beam causing the light sensor to respond) and the non-causal control condition (both lights shared a common cause). Following this participants were debriefed as to the purpose of this experiment.


Figure 2: Photographs of all experimental conditions from the participants' perspective. Self-causal condition (left): the paddle is set with the hole below the laser beam at the beginning of a trial. Mechanical-causal condition (centre): the wheel is rotating clockwise and the laser beam is obstructed. Non-causal control (right): the laser beam is
replaced with a red LED bulb positioned where the laser point can be seen in the other two conditions.

## Results

## Exclusions

Three participants were excluded for failing to follow instructions (consistently making multiple estimates per trial, or making estimates during, rather than between, trials). One further participant was excluded due to a technical error. For all other participants, individual trials for which there were two estimates and estimates which overlapped with the time of the event being judged were removed from analysis ( 8 participants with excluded trials, mean average 4.88 exclusions out of 120 trials).

## Causal estimates

A Friedman's ANOVA was used due to the ordinal nature of the causal scores. We found a significant main effect of condition on causal scores $\left(\mathrm{X}^{2}(2)=15.58, \mathrm{p}<.001\right)$. Posthoc testing using a Bonferroni correction found significantly lower scores for the non-causal control condition (median $=$ 6) compared with the self-causal condition (median $=8, \mathrm{p}<$ .05 ) and the mechanical-causal condition (median $=8, \mathrm{p}<$ .05). No significant difference was found between the selfcausal and mechanical-causal conditions ( $\mathrm{p}>.05$ ).

## Temporal estimates

Transformation A preliminary analysis of the data found significant variability in the range of reproductions between participants (see Table 1 for pre-transformation data). Additionally, a Shapiro-Wilk test found significant deviations from the normal distribution in two of the three conditions ( $\mathrm{p}<.05$ ). In order to reduce the influence of individual differences and reduce the positive skew of the data, temporal reproductions were converted to z-scores. To do this, each participant's grand mean was subtracted from each of their interval estimates. The difference from the mean of each score was divided by the standard deviation of all estimates (per participant). The mean z-score per condition for each participant was used for the temporal estimates analysis. Following transformation, the assumption of normality was met for all conditions ( $\mathrm{p}>$ $.05)$. The mean z-scores can be seen in Figure 3.

Table 1: Descriptive Statistics for raw Temporal Reproductions

| Condition | Mean | Standard deviation |
| :--- | :--- | :--- |
| Self-causal | 380.26 | 197.49 |
| Mechanical-causal | 406.8 | 134.81 |
| Non-causal Control | 501.14 | 315.39 |



Figure 3: Mean z scores of Temporal Reproductions by condition. Error bars represent the $95 \%$ confidence interval.

Analysis A one-way ANOVA found a significant main effect of condition on the z score-transformed estimates, $F(2,50)=4.46, p<.05, \eta^{2}=.15, M S E=1.57$. Planned simple contrasts were used to investigate the differences between the mechanical-causal condition and both other conditions. A significant difference was found between the mechanical-causal and self-causal conditions ( $p=.048$ ), but not between the mechanical-causal and non-causal control ( $p=.23$ ). The frequentist analysis, therefore, appears to favour the intentional binding account.

A Bayesian analysis was carried out using the BayesFactor package for R statistics (Morey, Rouder, Jamil \& Morey, 2015). A Bayesian repeated-measures ANOVA (see Rouder, Morey, Speckman \& Province, 2012 for details) found a Bayes factor of 17.23 for the unconstrained model (self-causal $\neq$ mechanical-causal $\neq$ non-causal control), indicating that the data observed is over 17 times more likely under the unconstrained model compared with the null model (intercept only). We also analysed three further models, as predicted by the intentional binding account (self-causal $<$ mechanical-causal $=$ non-causal control), the causal binding account (self-causal = mechanical-causal < non-causal control) and the 'intentional boost' account (self-causal < mechanical-causal < noncausal control). The highest Bayes factor was found for the model predicted by the intentional boost account $\left(\mathrm{BF}_{10}=\right.$ 91.82), and as such it is the preferred model compared with the models predicted by the intentional binding account $\left(\mathrm{BF}_{10}=44.57\right)$ and causal binding account $\left(\mathrm{BF}_{10}=16.22\right.$; denominator $=$ intercept only model for all Bayes factors).

## Discussion

We set out to investigate whether the perception of causality is sufficient for temporal binding to occur. We compared temporal estimates across three conditions: self-causal,
mechanical-causal and non-causal control. In contrast to the previous work on this topic, the first and second events in each sequence were equally predictable and perceived in the same modality (visual), thus eliminating possible confounding variables.

The results of the Bayesian analysis suggest that the most plausible model underlying our data is one of causal binding with an 'intentional boost'. It is noteworthy, however, that evidence of causal binding is only present in the Bayesian analysis, and cannot be seen in the frequentist planned contrasts. This apparent discrepancy may be the result of effect size; in this case there may have been a causal binding effect which was too small to be detectable under frequentist statistics, but still contributed the intentional boost model being the preferred model in the Bayesian analysis.

Our manipulation check (causal ratings) revealed higher-than-expected perceived causality between the two lights in the non-causal control condition. Although participants reported significantly weaker causal impressions in the noncausal control condition, the median score was 6 (on a 1-9 scale), indicating that 13 of the 26 participants included in the analysis perceived some causal relationship between the two lights. The causal binding view therefore would predict a reduced binding effect in those participants, due to reduced distinctiveness of the causal compared to the control conditions. .Therefore, while the manipulation has been successful in that the majority of participants reported weaker causal links between the two lights than the laser and light sensor, this may not have been sufficiently consistent across the entire sample to result in an effect size detectable by a frequentist analysis.

Although two previous studies have reported an intentional boost to causal binding (Buehner, 2015; 2012), it is still unclear how causal and intentional binding relate to each other. The causal ratings obtained here appear to rule out the possibility that participants perceived stronger causal relationships between the two events when the cause was self-initiated, so the intentional boost cannot be attributed to enhanced causal impressions following self-initiated vs mechanical causal actions. Instead, our findings suggest three possibilities. The first is that there may be two, separate causal and intentional binding effects, of differing strengths and with different roots, acting independently. This appears unlikely, however, in light of previous research suggesting that temporal binding does not occur in the absence of causality (Moore, Langado, Deal \& Haggard, 2009; Buehner \& Humphreys, 2009). Such findings indicate that the temporal binding effect is inextricably linked to perceived causality; specifically, that causality is necessary for the binding to occur.

An alternative explanation may be that causal binding and the intentional boost are a product of the same Bayesian processes, specifically, Bayesian cue integration. Humans appear to integrate information from multiple sensory cues in a manner that is statistically optimal: Ernst \& Banks (2002) found that when judging the height of a stimulus based on visual and haptic information, participants attached
more weight to the cue that has lower variance. Moore, Wegner \& Haggard (2009) suggested that Bayesian cue integration may also govern our sense of agency. Just as multiple sensory cues contribute to our judgment of physical properties such as size or shape, the perception of intentionality results from multiple cues, both internal (e.g. forward model predictions) and external (sensory cues). Applying this rationale to temporal binding, one could argue that the size of the effect depends on the noisiness of perceptual cues - specifically, the (perceived) times of the action and its consequence. One would expect that if prior expectation of the time of an effect is determined to some extent by the time of its cause, increased certainty in the time of the cause would lead to an increased weight being attached to it. Specifically, the shift in the perceived time of the effect (towards the cause) would be greater, the more certain one was about the time of the cause. While both the manual and mechanical actions were equally predictable in this experiment, participants had additional internal cues to the onset of their own action than to an observed mechanical event. It would be expected that the combination of these cues and visual feedback would lead to a more reliable and less variable percept of the time of the self-cause relative to the time of the machine-cause. As the expected time of the second event must be determined by the time of the first event (the cause is predictive of the effect), a more reliable percept of the time of the cause would lead to a more reliable prior for the time of the second event, which would be weighted more heavily against new cues to the time of the second event. This in turn would result in a greater backwards shift of the perceived time of the effect in the self-causal compared to the mechanical-causal condition. In line with this idea, Zhao et al. (2016) found greater temporal binding when participants had tactile cues to the time of the cause (key-press), compared with an action without tactile feedback (key release).

However, greater certainty in the time of the first action would also mean that it is less liable to be biased by the time of the second event. Altogether, Bayesian cue integration would thus be expected to lead to a lesser forward shift in the perceived time of the self- compared to the mechanicalcause, but a greater shift in the perceived time of the consequence towards a self- compared to a mechanicalcause. Thus, Bayesian cue integration based purely on the noisiness of the temporal cues leads to two distinct shifts in event perception working in opposite directions. We would argue that in addition to perceptual cues, the Bayesian integration process also takes into account expectations of cause-effect contiguity, in line with Hume's (1888) principles of causal inference. Crucially, that causes are predictive of their effect forms a key part of how we perceive causality; we perceive one event as causing another when it comes before it, when the second event occurs close in time to the first event and when the second event is contingent upon the first. The cause is thus more predictive of the effect than vice versa. These assumptions, however, do not necessarily accompany the perception of events
which are not causally related (e.g. a non-causal sequence of events). Furthermore, previous research (e.g. Haggard et al., 2002) has shown that outcome binding (i.e. shifts in the awareness of an outcome towards its cause) is typically greater than action binding (shifts in the awareness of a causal action towards its consequence). Therefore, the combined influence of the nature of causal relationships and the perceptual differences between self-action and mechanical actions may explain the presence of both causal binding and an intentional boost in our findings, within a single model.

More research is needed to determine what underlies the intentional boost to causal binding. The majority of research investigating the effect of agency on temporal binding to date has failed to take account of the potential role of causality: This is evidenced by failure to include adequate non-causal control conditions (e.g. Caspar, Christensen, Cleeremans \& Haggard, 2015; Zhao et al, 2016), or to obtain causal ratings as manipulation checks (e.g. Haggard et al, 2002). The present findings demonstrate the importance of control conditions in temporal binding, without which a reduced magnitude of temporal binding is indistinguishable from the absence of temporal binding. In particular, if there is indeed an intentional boost to causal binding, this calls for a reinterpretation of research suggesting there is no temporal binding effect in the absence of intentional action.

## Acknowledgements

We thank Dennis Simmonds and Stephen Michael for constructing the apparatus used in this experiment.

## References

Buehner, M. J. (2012). Understanding the past, predicting the future causation, not intentional action, is the root of temporal binding. Psychological Science, 23, 1490-1497
Buehner, M. J. (2015). Awareness of voluntary and involuntary causal actions and their outcomes. Psychology of Consciousness: Theory, Research, and Practice, 2(3), 237.
Buehner, M. J., \& Humphreys, G. R. (2009). Causal binding of actions to their effects. Psychological Science, 20(10), 1221-1228.
Caspar, E. A., Christensen, J. F., Cleeremans, A., \& Haggard, P. (2016). Coercion changes the sense of agency in the human brain. Current biology, 26(5), 585-592.
Eagleman, D. M., \& Holcombe, A. O. (2002). Causality and the perception of time. Trends in cognitive sciences, 6(8), 323-325.
Ernst, M. O., \& Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429-433.
Haggard, P., Clark, S., \& Kalogeras, J. (2002). Voluntary action and conscious awareness. Nature neuroscience, 5(4), 382-385.

Hascalovitz, A. C., \& Obhi, S. S. (2015). Personality and intentional binding: an exploratory study using the narcissistic personality inventory. Frontiers in human neuroscience, 9.
Hume, D. (1888). A Treatise of Human Nature: Reprinted from the Original Edition in Three Volumes. S. L. A. Selby-Bigge (Ed.). Clarendon Press.
Humphreys, G. R., \& Buehner, M. J. (2009). Magnitude estimation reveals temporal binding at super-second intervals. Journal of Experimental Psychology: Human Perception and Performance, 35(5), 1542.
Jo, H. G., Wittmann, M., Hinterberger, T., \& Schmidt, S. (2014). Brain Correlates of Intentional Binding: An EEG Study in Mindfulness Meditators. Procedia-Social and Behavioral Sciences, 126, 240.
Libet, B., Gleason, C. A., Wright, E. W., \& Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-
potential). Brain, 106(3), 623-642.
Lush, P., Parkinson, J., \& Dienes, Z. (2016). Illusory Temporal Binding in Meditators. Mindfulness, 1-7.
Moore, J., \& Haggard, P. (2008). Awareness of action: Inference and prediction. Consciousness and cognition, 17(1), 136-144.
Moore, J. W., Lagnado, D., Deal, D. C., \& Haggard, P. (2009). Feelings of control: contingency determines experience of action. Cognition, 110(2), 279-283.
Moore, J. W., \& Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. Consciousness and cognition, 21(1), 546-561.
Moore, J. W., Wegner, D. M., \& Haggard, P. (2009). Modulating the sense of agency with external cues. Consciousness and cognition, 18(4), 1056-1064.
Morey, R. D., Rouder, J. N., Jamil, T., \& Morey, M. R. D. (2015). Package 'BayesFactor'. URL http://cran. rproject. org/web/packages/BayesFactor/BayesFactor. pdf (accessed 10.06. 15).
Nolden, S., Haering, C., \& Kiesel, A. (2012). Assessing intentional binding with the method of constant stimuli. Consciousness and cognition, 21(3), 1176-1185.
Rouder, J. N., Morey, R. D., Speckman, P. L., \& Province, J. M. (2012). Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology, 56(5), 356374.

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., \& Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. Brain, 133(10), 3104-3112.
Zhao, K., Hu, L., Qu, F., Cui, Q., Piao, Q., Xu, H., ... \& Fu, X. (2016). Voluntary action and tactile sensory feedback in the intentional binding effect. Experimental brain research, 1-10.

# Familiarity-matching in decision making: Experimental studies on cognitive processes and analyses of its ecological rationality 

Masaru Shirasuna (m.shirasuna1392@gmail.com)<br>Graduate School of Arts and Sciences, The University of Tokyo<br>3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan<br>Hidehito Honda (hitohonda.02@gmail.com)<br>Graduate School of Arts and Sciences, The University of Tokyo<br>3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan

Toshihiko Matsuka (matsuka@chiba-u.jp)<br>Department of Cognitive and Information Science, Chiba University<br>1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan


#### Abstract

Previous studies have shown that individuals often make inferences based on heuristics using recognition, fluency, or familiarity. In the present study, we propose a new heuristic called familiarity-matching, which predicts that when a decision maker is familiar (or unfamiliar) with an object in a question sentence, $s$ /he will choose the more (or less) familiar object from the two alternatives. We examined inference processes and ecological rationality regarding familiarity-matching through three studies including behavioral experiments and ecological analyses. Results showed that participants often used familiaritymatching in solving difficult binary choice problems, and that familiarity-matching could be applied in an ecologically rational manner in real-world situations. A new perspective on human cognitive processes is discussed in this study.


Keywords: binary choice task; heuristic; familiarity; familiarity-matching; ecological rationality

## Introduction

When making decisions, individuals often use simple inference strategies such as heuristics. In the field of heuristics research, many researchers initially focused on cognitive biases involved in heuristics (e.g., Tversky \& Kahneman, 1973, 1974, 1983). In contrast, recent studies have discussed the adaptive aspect of heuristics (e.g., Gigerenzer \& Goldstein, 1996; Gigerenzer \& Todd, 1999; Goldstein \& Gigerenzer, 2002). Some studies investigated human inference cues or inference strategies using binary choice tasks (e.g., Goldstein \& Gigerenzer, 2002; Hertwig, Herzog, Schooler \& Reimer, 2008; Honda, Abe, Matsuka \& Yamagishi, 2011; Honda, Matsuka \& Ueda, in press). These studies showed that subjective memory experiences, such as recognition, fluency, or familiarity of an object could be valid inference cues. For example, in the binary choice task, "Which city has a larger population, Tokyo or Chiba?" when a decision maker recognizes (or is more fluent or familiar with) Tokyo and does not recognize (or is less
fluent or familiar with) Chiba, s/he tends to choose the recognized (or the more familiar or fluent) city -as the one with a larger population size. An interesting observation is that, in many cases, this simple inference can often lead to correct inferences. Thus, a simple heuristic using subjective memory experiences can be ecologically rational (e.g., Goldstein \& Gigerenzer, 2002; Hertwig et al., 2008; Schooler \& Hertwig, 2005; Honda et al., in press).

## Choice of an object based on similarity of familiarity: Familiarity-matching

So far, previous studies have investigated the effects of subjective memory experience for finding correct alternatives in a binary choice task. However, if the familiarity of an object in alternatives can serve as a valid inference cue, it is possible that the same holds true for the familiarity of an object in a question sentence. For example, if we consider the binary choice task, "Which country is Hameln in, Germany or Liechtenstein?" A decision maker may infer it as "I have heard the name 'Hameln' and I am familiar with this city. Further, I am more familiar with Germany than Liechtenstein; therefore, Hameln should be in Germany!" In this case, the decision maker chose the more familiar alternative because the familiarity of the chosen alternative was similar to that of the object in the question sentence. Likewise, in the task, "Which country is Schellenberg in, Germany or Liechtenstein?" A decision maker may infer it as "I have never heard the name 'Schellenberg' and I am unfamiliar with the city. Further, I am less familiar with Liechtenstein than Germany; therefore, Schellenberg should be in Liechtenstein!" In this case, the decision maker chose the less familiar alternative because the two objects were similarly unfamiliar. A decision maker may thus use an inference strategy like "matching familiarity" between an object in the question sentence and another object in the alternatives. That is, a decision maker makes inferences based on similarity of familiarity between
objects. Similarity judgments are closely connected to decision making and similarities between the familiarity of an object in a question sentence and that of an object in alternatives may become an important cue for making decisions. In fact, a recent study (Hiatt \& Trafton, in press) has shown that familiarity can be one of the most important cues in similarity judgments.

Based on these considerations, we propose a new heuristic termed as familiarity-matching: If an object in a question sentence is familiar (or unfamiliar) for a decision maker, then $s /$ he will choose the more (or less) familiar object from the two alternatives in a binary choice task. The goal of this study was to examine if cognitive processes in binary choice can be explained in terms of familiaritymatching and to investigate its ecological rationality. In the following sections, we shall report on three studies. In Study 1, we conducted a behavioral experiment and examined if familiarity-matching could adequately explain inference processes. In Study 2, we examined the ecological rationality of familiarity-matching. Finally, in Study 3, we analyzed the real-world environment in terms of familiarity.

## Study 1: Examination of inference processes

The purpose of Study 1 was to investigate if individuals tend to rely on familiarity-matching in a binary choice task.

## Method

Participants Japanese under graduate students $(N=31)$ participated in this study.
Tasks, materials, and procedure We conducted the binary choice task and the measurement of familiarity.

In the binary choice task, participants answered 100 binary choice questions. All question sentences had the following format: "X is a city in, country A1 or A2?" (e.g., "Sikasso is a city in, Mali or Switzerland?"). The order of the 100 questions was randomized (see Appendix for the procedure to generate the questions). For each question, participants were also asked to rate the difficulty level in answering the question using a visual analog scale (VAS). The scale consisted of a horizontal line labeled "very easy" on the left end and "very difficult" on the right end. Participants' responses were recorded over a range of 101points (i.e., from $0=$ "very easy" to $100=$ "very difficult").

In the measurement of familiarity, participants were asked to indicate how familiar they were with each object presented in the binary choice task (i.e., 20 countries and 100 cities) using a VAS. Participants' responses were recorded over a range of 101 -points (i.e., from $0=$ "do not know at all" on the left end of the scale to $100=$ "know much" on the right end of the scale).

We conducted the above two tasks using a questionnaire. Participants completed the binary choice task followed by the measurement of familiarity.

## Results

Hereafter, the familiarity ratings for the object in the question sentence and for the two objects presented as
alternatives are expressed as "Fam(Q)," "Fam(A1)," and "Fam(A2)," respectively. In the following analyses, we excluded the questions in which $\operatorname{Fam}(\mathrm{A} 1)$ was identical to Fam(A2).
Can familiarity-matching predict inference patterns? First, we analyzed the accordance rate of observed inferences with familiarity-matching for each participant. For example, when $\operatorname{Fam}(Q)=45, \operatorname{Fam}(A 1)=30$, and $\operatorname{Fam}(\mathrm{A} 2)=80$, familiarity-matching predicts that the participant would choose A1. Figure 1 shows the accordance rate for each participant. In 29 out of the 31 participants, accordance rates were above chance level. The mean accordance rate was .88 . These results indicate that the observed choices were predicted accurately by familiarity-matching.


Figure 1: Accordance rate of observed inferences with familiarity-matching (individual data). The red line denotes chance level (.50) and the dotted line shows the mean accordance rate (.88).

Does the difficulty of a problem affect the use of familiarity-matching? Previous studies have shown that individuals do not always use heuristics but tend to rely on them for solving a difficult problem (e.g., Kahneman \& Frederick, 2005; Honda et al., in press). Therefore, we examined if experiencing difficulty in a problem affected the use of familiarity-matching. We defined a dichotomized difficulty rating, high or low difficulty, based on the difficulty ratings being above or below the median for each participant. Hereafter, a problem assigned a rating above the median is expressed as "difficult problem" and a problem assigned a rating below the median as "easy problem". We examined the use of familiarity-matching for both types of problems.

Some researchers have debated that accordance rates are not always a good indicator for examining if individuals "truly" use heuristics (e.g., Hilbig \& Richter, 2011). Thus, we used Discrimination Index (DI) (Hilbig \& Pohl, 2008) as
an indicator of the blind usage of familiarity-matching by the participants. DI was calculated using the following equation:

$$
\text { DI }=(\text { Hit })-(\text { False Alarm })
$$

where (Hit) and (False Alarm) denote the proportion in which the accordance of a heuristic results in a correct or false inference, respectively. Since DI is defined as the difference between (Hit) and (False Alarm), DI ranges from -1 to +1 . It is assumed that when a decision maker always follows a heuristic (i.e., s/he blindly uses a heuristic), DI should reach zero, as s/he uses the heuristic irrespective of its correctness, suggesting that $\mathrm{s} /$ he does not take advantage of specific knowledge relevant to the inference problem.

For each participant, we calculated DI for the two problem types. Figure 2 shows the distributions of DI for the two cases. We found that DI for the difficult problem was generally lower than DI for the easy problem. We also found that the mean DI for the difficult problem was not significantly deviated from zero (Mean $=.07, t(30)=1.23, p$ $=.23$, Median $=.06$ ), while the mean DI for the easy problem was significantly deviated from zero (Mean $=.41$, $t(30)=6.47, p<.001$, Median $=.39$ ). These results implied that individuals used memory-based simple heuristics when they experienced difficulty in solving inference problems, which was consistent with the previous finding in Honda et al. (in press).


Figure 2: DI (Discrimination Index) for the difficult problem (left) and for the easy problem (right).

## Discussion

In this behavioral experiment, the accordance rate of the prediction by familiarity-matching was sufficiently high (mean accordance rate $=.88$ ), showing that familiaritymatching predicted inference patterns effectively. Furthermore, our findings implied that participants used familiarity-matching when they experienced difficulty in problems. These results suggest that individuals take
advantage of the familiarity of objects in both question sentences and alternatives as a cue when making inferences.

In the behavioral experiment, the materials used were selected by experimenters to serve as stimuli for the binary choice task. Therefore, the question of using familiaritymatching in a binary choice task as a valid inference strategy remains open for evaluation. Thus, we examined the ecological rationality of familiarity-matching.

## Study 2: Analysis of ecological rationality

The purpose of Study 2 was to examine if familiaritymatching could serve as an ecologically rational strategy. In this study, we measured individuals' familiarity of objects and then examined whether familiarity-matching was generally a valid inference strategy in a binary choice task.

## Method

Participants Japanese under graduate students $(N=39)$ participated in the task. None of them had participated in Study 1.
Materials, tasks, and procedure We used the 50 countries with the highest population in the world and their 50 capitals as materials. We investigated the participants' familiarity with each of the 100 objects (i.e., 50 cities and 50 countries). We conducted the measurement of familiarity which was similar to the method used in Study 1.
Analysis of the validity of familiarity-matching Familiarity ratings in Study 2 were converted into z-scores for each participant and the following analyses were conducted.

We analyzed the validity of familiarity-matching in the binary choice task using the familiarity ratings for the 50 countries and their capitals collected from participants. Specifically, we calculated the accuracy rate (i.e., validity) of familiarity-matching using the following procedure:

1. A hypothetical binary choice task such as " X is a city in, A1 or A2?" was conducted and each problem "was inferred" based on $\operatorname{Fam}(\mathrm{Q})$, $\operatorname{Fam}(\mathrm{A} 1)$ and $\operatorname{Fam}(\mathrm{A} 2)$.
2. For each question, if the absolute difference between $\operatorname{Fam}(\mathrm{Q})$ and $\operatorname{Fam}(\mathrm{A} 1)$ was less than that between $\operatorname{Fam}(\mathrm{Q})$ and $\operatorname{Fam}(\mathrm{A} 2)$, then A1 was selected, and vice versa (i.e., in the same manner as the prediction by familiarity-matching in Study 1).
3. We applied the above two steps to all possible combinations ( 50 cities * 49 alternative pairs) using the familiarity ratings provided by each participant, and then calculated his/her accuracy rate.

## Results and discussion

Figure 3 shows participant accuracy rates calculated as described above $(N=39)$. The horizontal and vertical axes shows the participants (individual data) and the accuracy rate, respectively, while the red line in the graph indicates chance level (.50). Participant accuracy rates (Mean = .67) exceeded the chance level for all participants. Therefore, it is suggested that participants can accurately "make correct
inferences" by matching more familiar objects in the binary choice task, even in the real-world environment.


Figure 3: Accuracy rate of familiarity-matching by individual familiarity ratings. The red line denotes chance level (.50) and the dotted line shows the mean accuracy rate (.67).

## Study 3: Analysis of the real world

According to previous studies (e.g., Goldstein \& Gigerenzer, 2002; Honda et al., in press), if a city or country appears more frequently in the real-world environment (e.g., mentioned in media), then individuals will be more familiar with it because they are more likely to see or hear the name. Therefore, it can be estimated that familiarity-matching can be applied effectively in the binary choice task when the frequency of appearance of the name of a city is correlated with that of the name of a country in the real-world environment.

In this section, we investigated this issue using the following procedure. As an index of the frequency of appearance in the real-world environment, we used the logtransformation of the mean number of hits for each object in two online databases of Japanese newspapers ${ }^{1}$. When we searched for objects in both databases, we traced back from the oldest to latest date as possible (see footnote) on national news. We converted the log-transformed index into z-scores, which were used in this analysis.

First, we calculated the correlation coefficient between the number of hits for the 50 cities and that for the 50 countries using z-scores. This correlation coefficient was 86 ( $p<.001 ; 95 \% C I: .77 \sim .92$; Figure 4). Therefore, it was found that the frequency of appearance of a city name in the media was highly correlated with that of the country name that corresponded to the city.

[^489]

Figure 4: Correlation between the number of hits for cities and that for countries (log-transformed $z$-scores).

Although the correlation between participants' familiarity with an object and the number of appearances in the media has already been reported in previous studies (e.g., Goldstein \& Gigerenzer, 2002; Schooler \& Hertwig, 2005), we confirmed that we could replicate such a correlation in the present study. Subsequently, using $z$-scores, we calculated the correlation coefficient between participants' familiarity with each object (z-scores of the mean familiarity ratings for 39 participants in Study 2) and the number of hits for each object. This correlation coefficient was .88 ( $p<.001$; 95\% CI: . $84 \sim .92$; Figure 5). Therefore, it was found that the more often an object appeared in the media, the more familiar with the object individuals were, which was consistent with previous studies.


Figure 5: Correlation between familiarity of objects (zscores) and the number of hits for these objects (logtransformed z-scores).

The combined results of Study 2 and Study 3 suggest that familiarity-matching can be valid even in the real-world environment, as the frequency of appearance of a city name in the media is highly correlated with that of the corresponding country name, and individuals' familiarity with an object is also highly correlated with its frequency of appearance in the media. In other words, since a more frequently appearing object in the environment is likely to
be more familiar for individuals, inferences based on similarity of familiarity can be valid in a binary choice task. Therefore, familiarity-matching can be applied as an ecologically rational strategy.

## General discussion

In the present study, we proposed a new heuristic, familiarity-matching, which predicts that if an object presented in a question sentence is familiar for a decision maker, then s/he is likely to choose the more familiar object presented as alternatives in a binary choice task. The results of Study 1 showed that familiarity-matching could predict individuals' inference patterns effectively. In particular, the results implied that individuals used familiarity-matching when they experienced difficulty in inference problems. In addition, the results of Study 2 and Study 3 showed that familiarity-matching could be an ecologically valid strategy in the binary choice task, because of the high correlations between the frequency of appearance of a city name and that of a country name, and between the frequency of appearance of an object and individuals' familiarity with it.

So far, only the use of "familiarity" in making inferences has been primarily examined. Generally, in a binary choice task, "familiarity" of an object can be an informative cue (e.g., Honda et al., 2011, in press). In a binary choice task, "unfamiliarity," contrary to "familiarity," is often considered uninformative in making inferences. The present findings, however, indicate that individuals can also use "unfamiliarity" as an informative inference cue. Familiaritymatching can be applied for both, a familiar object and an unfamiliar object, in a question sentence. According to familiarity-matching, when presented with an unfamiliar object, a decision maker will infer the following: "The correct answer will also be the unfamiliar object." In this situation, the "unfamiliarity" can become an informative cue. Perhaps, in cases where a decision maker uses "unfamiliarity" as an inference cue, the cognitive processes may differ from those involved in a situation where $s / h e$ uses "familiarity" as an inference cue. The present study did not examine this issue, which, therefore, should be investigated in the future.

However, familiarity-matching still has some limitations in the practical aspect. The definition of the familiaritymatching is limited to a binary choice task. We think that familiarity-matching can be applied to a multiple-choice task, because a decision maker has only to "match" familiarity of an object in a question sentence with that of an object in alternatives, no matter how many alternatives the task contains. However, it is not clear how familiaritymatching can be extended to other more complex tasks, so we may also need to investigate this issue.

To the best of our knowledge, the present study was the first study to examine the effect of familiarity of objects in a question sentence. We believe that focusing on the relationship between objects presented as both, a main theme ("question sentence," in this study) and a supplement
("alternatives," in this study), has revealed a new perspective on human inferences.

## Acknowledgment

This research was supported by Grants-in-Aid for Scientific Research (A, 16H01725; B, 25280049, 16H02835) and Young Scientists (B, 16K16070), and Grants-in-Aid for Scientific Research on Innovative Areas (16H06569).

## References

Gigerenzer, G., \& Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 103, 650-669.
Gigerenzer, G., \& Todd, P. M. (1999). Simple heuristics that make us smart. New York: Oxford University Press.
Goldstein, D. G., \& Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. Psychological Review, 109(1), 75-90.
Hertwig, R., Herzog, S., Schooler, L., \& Reimer, T. (2008). Fluency Heuristic: A Model of How the Mind Exploits a By-Product of Information Retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 5, 1191-1206.
Hiatt, L. M., \& Trafton, J. G. (in press). Familiarity, Priming, and Perception in Similarity Judgments. Cognitive Science.
Hilbig, B. E. \& Pohl, R. F. (2008). Recognizing Users of the Recognition Heuristic. Journal of Research in Personality, 42, 1641-1645..
Hilbig, B. E. (2010). Precise models deserve precise measures: A methodological Dissection. Judgment and Decision Making, 5(4), 272-284.
Hilbig, B. E., Erdfelder, E., \& Pohl, R. F. (2010). Onereason decision making unveiled: A measurement model of the recognition heuristic. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(1), 123-134.
Hilbig, B. E., Erdfelder, E., \& Pohl, R. F. (2011). Fluent, Fast, and Frugal? A Formal Model Evaluation of the Interplay Between Memory, Fluency, and Comparative Judgments. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(4), 827-839.
Hilbig, B. E., \& Richter, T. (2011). Homo heuristicus outnumbered: Comment on Gigerenzer and Brighton (2009). Topics in Cognitive Science, 3, 187-196. doi:10.1111/j.1756-8765.2010.01123.x
Honda, H., Abe, K., Matsuka, T., \& Yamagishi, K. (2011). The role of familiarity in binary choice inferences. Memory and Cognition, 39 (5), 851-863.
Honda, H., Matsuka, T., \& Ueda, K. (in press). Memorybased simple heuristics as attribute substitution: Competitive tests of binary choice inference models. Cognitive Science.
Kahneman, D., \& Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak \& R. G. Morrison (Eds.),

Cambridge handbook of thinking and reasoning (pp. 267293). New York: Cambridge University Press.

Schooler, L. J., \& Hertwig, R. (2005). How forgetting aids heuristic inference. Psychological Review, 112, 610-628.
Tversky, A., \& Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability. Cognitive Psychology, 5, 207-232.
Tversky, A. \& Kahneman, D. (1974). Judgments under uncertainty: Heuristics and biases. Science, 185, 11241131.

Tversky, A. \& Kahneman, D. (1983). Extensional versus intuitive reasoning; The conjunction fallacy in probability judgement. Psychological Review, 90, 293-315.

## Appendix. Binary choice task

The binary choice task (Study 1) was generated by following the four steps listed as under:

1. For "objects presented as alternatives," we selected 20 countries (more than 2 countries from 5 regions: Asia, Europe, Africa, North America, and South America) and randomly assigned these 20 countries to 2 groups: "Alternative A1" and "Alternative A2" (each group consisted of different 10 countries).

| Alternative A1 | Alternative A2 |
| :--- | :--- |
| America | Canada |
| Sweden | Bolivia |
| Mexico | Italia |
| Columbia | Ukraine |
| Holland | Switzerland |
| Egypt | Iran |
| Turkey | Spain |
| Saudi Arabia | Kazakhstan |
| Australia | New Zealand |
| Mali | Morocco |

2. Using the groups, "Alternative A1" and "Alternative A2" described above, we created $10 * 10=100$ pairs as alternatives for the binary choice task.

|  | Alternative A1 | Alternative A2 |
| :--- | :--- | :---: |
| 1 | Holland | Iran |
| 2 | Australia | Bolivia |
| 3 | Columbia | Kazakhstan |
| 4 | Saudi Arabia | Morocco |
| $\ldots$ |  |  |
| 98 | Turkey | New Zealand |
| 99 | Mexico | Switzerland |
| 100 | America | Ukraine |

3. From each country ("objects presented as alternatives"), we selected 5 cities (total of $20 * 5=$ 100 cities) using the following criteria:
(I) Out of the 5 cities, we selected the 2 cities with the largest population size in the country.
(II) For the remaining 3 cities, we selected cities which satisfied following one (or more) of the following criteria: "is the high population size", "its
name is included in that of a historical treaty, conference, or a similar historical event," "has a world heritage site," or "has hosted the Olympic or the Paralympic Games."

| Cities | Countries |  |  |
| :--- | :--- | :---: | :---: |
| New York | America |  |  |
| Washington D.C. | America |  |  |
| Portsmouth | America |  |  |
| San Francisco | America |  |  |
| Bretton Woods | America |  |  |
| Rapallo | Italia |  |  |
| Trent | Italia |  |  |
| Roma | Italia |  |  |
| Milano | Italia |  |  |
| Genova | Italia |  |  |
| Teheran | Iran |  |  |
|  |  |  | Mali |
| Sikasso | Mexico |  |  |
| Puebla | Mexico |  |  |
| Tlatelolco | Mexico |  |  |
| Guadalajara | Mexico |  |  |
| Monterrey | Mexico |  |  |
| Villahermosa | Morocco |  |  |
| Rabat | Morocco |  |  |
| Marrakesh | Morocco |  |  |
| Tangier | Morocco |  |  |
| Fes | Morocco |  |  |
| Casablanca |  |  |  |

Note: We provided criterion (II) for two reasons: First, if all alternatives consisted of top cities in terms of population, participants might be more likely to know the answer (i.e., to use knowledge-based inference cues instead of heuristics), as it seemed that larger cities were comparatively more famous. Second, we wanted to create objects presented as question sentences (i.e., cities) that would be the only familiar element for participants (i.e., when participants are only familiar with a city, they often do not know the country it belongs to). However, even if a city satisfied criterion (I) or (II), we excluded cities whose names included the name of the country (e.g., Mexico City) or were located in several countries (e.g., Melbourne is located not only in Australia but also in America).
4. In order to make one of the two alternatives (from step 2) a correct answer, we placed a city (from step 3 ) in " $X$ " in each question sentence (" $X$ is a city in,").

|  | Alternative A1 | Alternative <br> A2 |
| :--- | :--- | :--- |
| 1. Ramsar is a city in, | Holland | Iran |
| 2. La Paz is a city in, | Australia | Bolivia |
| 3. Bogota is a city in, | Columbia | Kazakhstan |
| 4. Rabat is a city in, | Saudi Arabia | Morocco |
|  |  |  |
| 98. Ankara is a city in, |  |  |
| 99. Villahermosa is a city in, | Mexico | New Zealand |
| 100. Bretton Woods is a city in, | America | Uwitzerland |

Note: Sentences in the actual questionnaire were written in Japanese.

# Bridging a Conceptual Divide: How Peer Collaboration Facilitates Science Learning 

Andrew Shtulman (shtulman@oxy.edu)<br>Andrew Young (ayoung2@oxy.edu)<br>Department of Psychology, Occidental College<br>1600 Campus Road, Los Angeles, CA 90041


#### Abstract

Collaboration is generally an effective means of learning new information, but is collaboration productive in domains where collaborators may hold qualitatively different conceptions of the domain's causal structure? We explored this question in the domain of evolutionary biology, where previous research has shown that most individuals construe evolution as the uniform transformation of an entire population (akin to metamorphosis) rather than the selective survival and reproduction of a subset of the population. College undergraduates $(n=44)$ completed an assessment of their evolutionary reasoning by themselves (pretest), with a partner (dyad test), and several weeks later (posttest). Collaboration proved ineffective for the higherscoring partner in each dyad, as their scores generally remained unchanged from pretest to dyad test to posttest, but it proved effective for the lower-scoring partner. Not only did lowerscoring partners increase their score from pretest to dyad test, but they maintained higher scores at posttest as well. Followup analyses revealed that participants' posttest scores were predicted by their partners' pretest scores but only for lowerscoring partners, and the relation was negative: the smaller the difference between pretest score, the greater the gain from pretest to posttest for lower-scoring partners. These findings indicate that collaboration in domains characterized by conceptual change is possible, but that learning from such collaboration is asymmetric (i.e., individuals with low levels of understanding benefit more than their partners do) and unequal (i.e., individuals with low levels of understanding benefit more if their partner's understanding is only moderately higher). Thus, bridging the gap between a novice's view of a conceptually complex domain and an expert's view appears to require instruction more aligned with the former than the latter.


Keywords: collaboration, conceptual development, science learning, intuitive theories, evolutionary reasoning

## Introduction

Some ideas are more difficult to learn than others. Ideas that can be encoded in terms of preexisting concepts, like the name of an unfamiliar animal or the function of an unfamiliar artifact, are much easier to learn than ideas that require new concepts for their encoding, like the reason the seasons change or the reason projectiles fall to the ground in a parabolic path. Learning the latter requires conceptual change, or knowledge restructuring at the level of individual concepts (Carey, 2009; Chi, 1992). Conceptual change is an intrinsic part of science learning. Most domains of science entail entities, properties, and mechanisms that defy our intuitive knowledge of how the world works and can only be represented if that knowledge is reorganized and restructured (Nersessian, 1998; Vosniadou, 1994).

Conceptual change is empirically distinguishable from other forms of knowledge acquisition in that it results in systematic failures of teaching and learning. In domains
requiring conceptual change, individuals who have yet to undergo that change exhibit misconceptions about the domain's content that are internally coherent and developmentally widespread. These misconceptions are robust in the face of counterevidence or counterinstruction, and they create impasses in communication between those who have achieved conceptual change and those who have not (for reviews, see Carey, 2009; Shtulman, 2017).

Consider the domain of evolutionary biology-the domain of choice in the present study. Evolution results from differential survival and differential reproduction within a population; the traits possessed by the most reproductively successful individuals spread through the population over time. Most people, however, view evolution as the uniform transformation of an entire population, where every organism is guaranteed to have offspring more adapted to the environment than it was at birth (Bishop \& Anderson, 1990; Shtulman, 2006). This view is grounded in the commonsense assumption that all members of a species share the same inner nature, or essence, which determines their outward appearance and behavior (Gelman, 2003; Shtulman \& Schulz, 2008). Evolution is thus seen as a kind of crossgenerational metamorphosis; selection plays no role in the process. This essentialist view of evolution has been documented in people of varying ages (Berti, Toneatti, \& Rosati, 2010; Shtulman, Neal, \& Lindquist, 2016) and educational backgrounds (Coley \& Tanner, 2015; Gregory \& Ellis, 2009), and it characterizes how a person reasons about several aspects of evolution, including variation, inheritance, adaptation, domestication, speciation, and extinction (Shtulman \& Calabi, 2013).

The focus of the current study is a particular hallmark of conceptual change: impasses in communication between those who have achieved conceptual change and those who have not. Such impasses have been observed in conversations between children and adults (e.g., Carey, 1985), in conversations between science students and science teachers (e.g., Wiser \& Amin, 2001), and in conversations between scientists working within different theoretical paradigms (e.g., Kuhn, 1977). Such impasses are often encountered in the context of learning-e.g., a child learning about the properties of living things from a parent or a student learning about the properties of thermal systems from a teacher-but it is unclear how they affect learning. Achieving conceptual change requires overcoming the conceptual gap responsible for the impasse, but how and with whom?

Conceptual impasses in communication are particularly important to study in light of the finding that learning is often facilitated through collaboration. For many types of inductive problems, individuals who collaborate on those problems are
more likely to solve them-and learn from them-than individuals who work alone (Gauvain \& Rogoff, 1989; Laughlin, Vanderstoep, \& Hollingshead, 1991; Leman, Skipper, Watling, \& Rutland, 2016). Collaboration is effective for several reasons. It opens partners' eyes to ideas they would not have generated on their own, highlighting alternative approaches to the same problem (Schwarz, Neuman, \& Biezuner, 2000; Young, Alibali, \& Kalish, 2012) or alternative explanations for the same phenomenon (Ames \& Murray, 1982; Howe, 2009). It forces collaborating partners to articulate their reasons for endorsing a particular hypothesis or favoring a particular solution strategy and defend those reasons with evidence (Okada \& Simon, 1997; Teasley, 1995). And it introduces social incentives for completing the task at hand, increasing partners' motivation to persist in the face of unexpected obstacles (Butler \& Walton, 2013).

Given the pedagogical benefits of collaboration, we sought to determine whether collaboration is useful-or even possible-in domains requiring conceptual change. The answer to this question has both practical and theoretical implications. From a practical point of view, educators who instruct students on topics requiring conceptual change (e.g., evolution, microbiology, mechanics, thermodynamics, fractions) would benefit from knowing whether collaboration is an effective instructional strategy or a dead end. From a theoretical point of view, models of conceptual change would be further informed by research clarifying which kinds of input foster conceptual change and which do not. Parent-child conversation, for instance, is a form of collaboration that may help foster conceptual change (Gunderson \& Levine, 2011; Jipson \& Callanan, 2003), but it is unclear how beneficial this activity is relative to other domain-specific activities (e.g., refutation-based instruction, inquiry-based instruction, informal exploration).

Previous studies have found that collaboration in conceptually complex domains can be successful. For instance, Asterhan and Schwarz (2007) found that undergraduates who collaborated on devising evolutionary explanations for two instances of biological adaptation (mosquitos developing resistance to an insecticide, cheetahs acquiring the ability to run faster than any other mammal) provided more sophisticated (selection-based) explanations for biological adaptation from pre-collaboration to postcollaboration. Likewise, Loyens, Jones, Mikkers, and van Gog (2015) found that undergraduates who collaborated on determining the paths traced by three projectiles (a child jumping from a swing, an object falling on someone's head, a coyote falling from a cliff) drew more accurate motion paths from pre-collaboration to post-collaboration. In both the domain of evolution and the domain of motion, learning a correct, scientific view of the domain is difficult to achieve; direct (lecture-based) instruction on these topics has typically proven unsuccessful (see Shtulman, 2017, for a review).

These studies demonstrate that collaboration can facilitate learning in domains characterized by conceptual change, but they are limited in that they explored only one aspect of those
domains-explanations for adaptation in the study by Asterhan and Schwarz (2007) and trajectories of projectile motion in the study by Loyens et al. (2015). The present study explored whether collaboration is effective for learning several facets of a conceptually complex domain-namely, the phenomena of variation, inheritance, adaptation, domestication, speciation, and extinction within the domain of evolutionary biology. We chose this domain because its content is notoriously difficult to understand and because individuals who are asked to reason about domain-relevant problems would most likely hold different levels of understanding. Collaboration under these circumstances thus provides a stringent test of whether, and how, collaboration can facilitate conceptual change.

## Method

## Participants

The participants were 44 college undergraduates, recruited from introductory psychology and cognitive science courses and compensated either with extra credit in those courses or with a small stipend. They were taken from a larger dataset of 174 participants, assigned to one of 87 dyads. All participants in the larger dataset were invited to complete a posttest (for a $\$ 12$ Amazon gift card) but only 44 did. Those 44 came from 36 different dyads. Approximately half were the higher-scoring partner in their dyad $(n=25)$ and half were the lower-scoring partner $(n=19)$. In other words, approximately half collaborated with a partner who demonstrated a higher level of understanding prior to the collaboration, and half collaborated with a partner who demonstrated a lower level of understanding.

The 44 participants who completed a posttest earned similar pretest scores to those who did not complete a posttest. That is, the 19 low scorers who completed a posttest scored similarly to the 68 who did not ( $M=9.4$ vs. $M=9.7$, $t(85)=0.22, p=0.83$ ), and the 25 high scorers who completed a posttest scored similarly to the 62 who did not ( $M=17.1$ vs. $M=15.9, t(85)=1.13, p=0.26$ ). While there may be motivational differences between those who opted to complete a posttest and those who did not, there were no reliable knowledge differences between the groups (relative to participants' classification as a low scorer or a high scorer).

## Materials

Participants were assessed on their understanding of evolution using an instrument developed by Shtulman (2006). The assessment consisted of six sections, each devoted to a different biological phenomenon (inheritance, variation, adaptation, domestication, speciation, and extinction). Participants' understanding of the phenomenon was assessed with five questions or tasks designed to elicit either an essentialist interpretation or a selection-based interpretation.

With respect to inheritance, for instance, participants were asked to make predictions about parent-offspring resemblance with questions like the following: "Imagine that
biologists discover a new species of woodpecker that lives in isolation on a secluded island. These woodpeckers have, on average, a one-inch beak and their only food source is a treedwelling insect that lives, on average, one-and-a-half inches under the tree bark. Compared to its parents, the offspring of any two woodpeckers should develop: (a) a longer beak, (b) a shorter beak, or (c) either a longer beak or a shorter beak; neither is more likely." The correct response is (c), because offspring vary randomly from their parents, but most people select (a), reasoning that offspring will inherit whatever traits will help them survive-traits conferred by an underlying essence that adaptively changes in response to the species' current needs.

As another illustration, consider this task designed to probe participants' understanding of within-species variation: "During the 19th century, England's native moth species, Biston betularia, evolved darker coloration in response to the pollution produced by the Industrial Revolution. Imagine that biologists gathered a random sample of Biston betularia once every 25 years from 1800 to 1900 . What range of coloration would you expect to find at each point in time?" Participants were given a five-by-five matrix of moth outlines and instructed to shade the moths to reflect what the moths might look like at $1800,1825,1850,1875$, and 1900. The two most common response patterns are depicted in Figure 1. The pattern on the left depicts a mutation for darker coloration spreading through the population over time and is consistent with a selection-based view of evolution. The pattern on the right depicts the uniform transformation of the population, such that variation occurs between generations but not within generations, and is consistent with an essentialist view.


Figure 1: A selection-based response pattern (left) and an essentialist response pattern (right) on the moth-shading task of the evolution comprehension assessment.

The full battery of questions can be found in the Appendix of Shtulman (2006), along with criteria for scoring each question or task. Participants were assigned 1 point for every correct. selection-based response and 0 points for every incorrect, essentialist response. Responses too vague to be counted as selection-based, were also assigned 0 points. Participants' scores thus ranged from 0 to 5 per section and from 0 to 30 for the assessment as a whole.

## Procedure

The evolution comprehension assessment was administered on a computer. It took between 30 and 45 minutes to complete, and participants completed it twice by themselves (pretest and posttest) and once with a partner (dyad test). Participants were tested in pairs in a room in the Psychology Department. They completed the pretest by themselves, and they completed the dyad test together immediately following the pretest. Participants typically did not know their dyad partner, and they were given no instruction on how to coordinate their responses. They were simply asked to complete the survey as a pair, on a single computer. Their conversations were recorded and transcribed at a later date. (Data from the conversations are not reported here, for lack of space).

The posttest was administered one semester (i.e., half a year) after the dyad test. The average delay between dyad test and posttest was 7.6 months, and the delay for the high scorers was equivalent to the delay for the low scorers ( $M=$ 6.8 vs. $M=8.8, t(42)=1.13, p=.138$ ). Effects of collaboration detectable after half a year arguably represent long-term changes in understanding, as participants' memory for the episodic details of the collaboration session would likely have faded.

## Results

For the analyses below, we used linear mixed models (LMMs) with random-effects structures specified according to the procedure recommended by Bates and colleagues (Bates et al., 2015). We used likelihood ratio-test (LRT) comparisons and 95\% confidence intervals (CI) for inference.

## Evolution Scores at Pretest, Dyad Test, and Posttest

Overall assessment scores at pretest, dyad test, and posttest are shown in Figure 2 (left panel). Dyads generated higher scores than individuals at pretest, $\beta=2.16,95 \% \mathrm{CI}[.27$, 4.05]. However, individuals' posttest scores were similar to their pretest scores, $\beta=.91,95 \%$ CI $[-.98,2.80]$.

## Low- versus High-Scoring Partners

An individual's ability to profit from collaboration likely depends on both their own prior knowledge and their partner's prior knowledge. For example, individuals paired with partners that demonstrated greater conceptual knowledge of evolution at pretest might have a greater opportunity to learn from collaboration than those paired with partners with less conceptual knowledge. To explore this possibility, we categorized participants in terms of whether they were the lower or higher scoring partner in their respective dyads at pretest. Figure 2 shows pretest, dyad test, and posttest scores for lower-scoring partners (middle panel) and higher-scoring partners (right panel). There was an interaction between scoring status and test, LRT $\chi^{2}(2)=$ $20.13, p<.001$. For high-scoring partners, pretest, dyad test, and posttest scores were similar. However, for low-scoring partners, dyad tests were greater than pretests, $\beta=6.53,95 \%$


Figure 2. Mean evolution score at pretest, dyad test, and posttest for all participants, low scoring participants, and high scoring participants. Error bars represent $\pm$ SE.

CI [3.92, 9.13], and posttests were greater than pretests, $\beta=$ $3.37,95 \%$ CI [1.87, 7.08].

## Low- versus High-Scoring Partners by Section

Only 5 of the 44 participants demonstrated consistently greater or poorer pretest performance than their partner across all six sections of the assessment. Categorizing participants as low- or high-scoring by section potentially provides a more nuanced view of performance than overall pretest scores. Figure 3 shows lower- and higher-scoring partners' pretest, dyad test, and posttest scores for each of the six sections. A three way interaction between scoring status, test, and section suggested variation in low and high scorers’ learning across section, LRT $\chi^{2}(10)=19.10, p=.0317$.

Scores for Inheritance and Speciation demonstrated little pretest to posttest change for both low and high scorers. Most consistent with the overall assessment, lower scorers demonstrated pretest to posttest gains for Domestication, $\beta=$ $1.50,95 \%$ CI $[.54,2.46]$, and Extinction, $\beta=.76,95 \%$ CI [$.02,1.55]$, whereas high scorers had similar pretest and posttest scores. Low scorers again demonstrated pretest to posttest gains for the Adaptation, $\beta=1.71,95 \%$ CI $[.84$, 2.57], and Variation, $\beta=1.43,95 \%$ CI [.47, 2.38]. However, high scorers surprisingly demonstrated pretest to posttest losses for Adaptation, $\beta=-1.50$, $95 \%$ CI $[-2.26,-.74]$, and Variation, $\beta=-1.21,95 \%$ CI [-2.03, -.39$]$.

## Predicting Posttest Scores

For low scorers, collaborating with a more advanced partner yielded pretest to posttest improvement in 4 out of 6 sections. For high scorers, collaborating with a less advanced partner yielded pretest to posttest decline in 2 out of 6 sections. These results suggest that participants' posttest performance was influenced both by their own understanding of the domain and by their partner's understanding.

To explore this possibility further, we fit an LMM on posttest scores (by section) with participant pretest scores, partner pretest scores, scoring status (high vs. low within a participant's respective dyad), and their interactions as fixed effects. Participant pretest score was a positive predictor of posttest score, $\beta=.38,95 \% \mathrm{CI}[.08, .67]$, and did not interact
with scoring status. In contrast, partner pretest score interacted with scoring status, LRT $\chi^{2}(1)=3.86, p=.049$. Partner pretest scores were not predictive of posttest scores for high scorers, $\beta=.08,95 \% \mathrm{CI}[-.15, .31]$. However, partner pretest scores were negatively related to posttest scores for low scorers, $\beta=-.22,95 \%$ CI $[-.43, .00]$. Thus, it appears that low scorers learned more from partners with slightly greater knowledge than themselves at pretest compared to partners with much greater knowledge.

## Discussion

Collaboration is an effective and efficient means of devising new hypotheses (Okada \& Simon, 1997) and learning new problem-solving strategies (Schwarz et al., 2000), but is collaboration possible in domains where individuals are known to hold vastly discrepant views of the domain's causal structure? The answer appears to be yes. Individuals who collaborated on tasks within the domain of evolutionary biology-a domain characterized by qualitatively different theories of what evolution is and how evolution works (Shtulman, 2006)-demonstrated a higher level of understanding together than they did individually. This finding was far from guaranteed given the content of the task. When partners disagreed, their disagreements typically reflected fundamental differences in their understanding of the task domain. Resolving those disagreements entailed more than just recognizing who knew the answer. It entailed recognizing which of two answers-a selection-based answer and an essentialist answer-was more plausible or justifiable.

Collaboration not only facilitated more accurate responding, it also facilitated learning, though the effects were nuanced. Individuals who entered the collaboration with lower levels of understanding demonstrated increased understanding at posttest (several months later), whereas individuals who entered the collaboration with higher levels of understanding demonstrated no gains at posttest.

The learning exhibited by less-knowledgeable partners was generally robust across different sections of the assessment, as was the stasis exhibited by more-knowledgeable partners. That said, there were some sections on which the lessknowledgeable partners exhibited no gains from pretest to


Figure 3. Mean pretest, dyad test, and posttest scores for low scoring and high scoring partners by section. Error bars represent $\pm$ SE.
posttest and some sections on which the more-knowledgeable partners exhibited losses from pretest to posttest. Collaboration in conceptually complex domains may thus hinder learning for some individuals in some contexts. Still, the net benefits of collaboration were positive, which is a surprising finding given that (a) collaborators often had to communicate across a conceptual divide and (b) the collaboration itself consisted solely of discussion. There were no opportunities to generate evidence or test hypotheses, which suggests that such activities may not be necessary for learning in cases where the primary intellectual challenge is just interpreting what one's partner is saying.

Perhaps the most provocative finding was that, among participants who learned from collaboration (the low scorers), those who learned the most collaborated with partners who had moderately higher levels of understanding. Individuals who collaborated with partners with substantially higher levels of understanding benefited less, at least by posttest. This finding, though tentative, may have resulted from differential impasses in communication; the greater the discrepancy between partners' understanding of the domain, the more likely they encountered impasses in communication and the more strained their collaboration may have become. Consider, for instance, the following conversation between a participant who earned a pretest score of 19 (P1) and one who earned a pretest score of 2 (P2) about the woodpecker question presented above:

P1: Alright, for the first one I put either a shorter or longer beak because it says compared to its parents, and compared to its parents it pretty much has the same beak because it has the same genes.
P2: Okay. Hmm. I put longer beak ... because, yeah, they have to eventually evolve into the thing, but I can see what you are saying about, like, it wouldn't take one generation.

P1: Well ... the next generation would end up with a longer beak, but this one particular woodpecker would have the same [beak] as its parents, if you understand what I'm saying. The generations would get longer beaks because the ones with the shorter beaks will be killed off. [But] no matter what, the offspring are gonna have beaks pretty much the same as its parents.
P2: Okay, I see what you're saying. Yeah, I guess I just assumed that they would interbreed or they would have a woodpecker from a different... Okay, I see what you're saying.

P2 claims to understand what P1 is saying, but P2's attempts to resolve the discrepancy - by acknowledging P2's answer as correct on the assumption that "it wouldn't take one generation" or that birds with different beak lengths did not "interbreed"-do not actually address P2's point that evolutionary change occurs at the population level, not the individual level. This type of impasse may be more common in conversations between partners with discrepant levels of understanding than in conversations between partners with similar levels of understanding, though confirmation of this pattern awaits further analysis of the conversational data.

The finding that participants benefited most from collaborating with individuals who were only moderately more knowledgeable about the domain helps answer the question of how individuals are able to communicate across a gap in conceptual understanding. Communication may be possible only if the gap is not too wide; wider gaps may lead to irreconcilable differences in how partners perceive or analyze the problems at hand. We plan to test this idea directly by analyzing the dynamics of participants' conversations in relation to their score differences from pretest to dyad test and from pretest to posttest. Previous research on how domain experts converse with domain novices suggests that the experts supply novices with
specialized knowledge, in the moment, by adjusting how they label or how they describe objects of shared attention (Clark \& Schaefer, 1989; Isaacs \& Clark, 1987). However, such studies have involved domains in which the difference between a novice's knowledge and an expert's knowledge is quantitative rather than qualitative (e.g., knowledge of New York City landmarks). It remains an open question how domain novices and domain experts are able to bridge differences in knowledge, through discourse patterns, when that knowledge entails conceptual change.

## References

Ames, G. J., \& Murray, F. B. (1982). When two wrongs make a right. Developmental Psychology, 18, 894-897.
Asterhan, C. S., \& Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. Journal of Educational Psychology, 99, 626-639.
Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1-48.
Berti, A. E., Toneatti, L., \& Rosati, V. (2010). Children's conceptions about the origin of species: A study of Italian children's conceptions with and without instruction. The Journal of the Learning Sciences, 19, 506-538.
Bishop, B. \& Anderson, C.A. (1990). Student conceptions of natural selection and its role in evolution. Journal of Research in Science Teaching, 27, 415-427.
Butler, L. P., \& Walton, G. M. (2013). The opportunity to collaborate increases preschoolers' motivation for challenging tasks. Journal of Experimental Child Psychology, 116, 953-961.
Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: MIT Press.
Carey, S. (2009). The origin of concepts. Oxford, UK: Oxford University Press.
Chi, M. (1992). Conceptual change within and across ontological categories. In R. Giere (Ed.), Cognitive models of science (129-186). Minneapolis: U of Minnesota Press.
Clark, H. H., \& Schaefer, E. F. (1989). Contributing to discourse. Cognitive Science, 13, 259-294.
Coley, J. D., \& Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. CBE: Life Sciences Education, 14, 1-19.
Gauvain, M., \& Rogoff, B. (1989). Collaborative problem solving and children's planning skills. Developmental Psychology, 25, 139-151.
Gelman, S. A. (2003). The essential child. Oxford, UK: Oxford University Press.
Gregory, T. R., \& Ellis, C. A. (2009). Conceptions of evolution among science graduate students. BioScience, 59, 792-799.
Gunderson, E. A., \& Levine, S. C. (2011). Some types of parent number talk count more than others: Relations between parents' input and children's cardinal-number knowledge. Developmental Science, 14, 1021-1032.

Howe, C. (2009). Collaborative group work in middle childhood. Human Development, 52, 215-239.
Isaacs, E. A., \& Clark, H. H. (1987). References in conversation between experts and novices. Journal of Experimental Psychology, 116, 26-37.
Jipson, J. L., \& Callanan, M. A. (2003). Mother-child conversation and children's understanding of biological and nonbiological changes in size. Child Development, 74, 629-644.
Kuhn, T. S. (1977). The essential tension: Selected studies in scientific tradition and change. Chicago: University of Chicago Press.
Laughlin, P. R., Vanderstoep, S. W., \& Hollingshead, A. B. (1991). Collective versus individual induction. Journal of Personality and Social Psychology, 61, 50-67.
Leman, P. J., Skipper, Y., Watling, D., \& Rutland, A. (2016). Conceptual change in science is facilitated through peer collaboration for boys but not for girls. Child Development, 87, 176-183.
Loyens, S. M., Jones, S. H., Mikkers, J., \& van Gog, T. (2015). Problem-based learning as a facilitator of conceptual change. Learning and Instruction, 38, 34-42.
Nersessian, N. J. (1989). Conceptual change in science and in science education. Synthese, 80, 163-183.
Okada, T., \& Simon, H. A. (1997). Collaborative discovery in a scientific domain. Cognitive Science, 21, 109-146.
Schwarz, B. B., Neuman, Y., \& Biezuner, S. (2000). Two wrongs may make a right... If they argue together! Cognition and Instruction, 18, 461-494.
Shtulman, A. (2006). Qualitative differences between naive and scientific theories of evolution. Cognitive Psychology, 52, 170-194.
Shtulman, A. (2017). Scienceblind: Why our intuitive theories about the world are so often wrong. New York: Basic Books.
Shtulman, A., \& Calabi, P. (2013). Tuition vs. intuition: Effects of instruction on naïve theories of evolution. Merrill-Palmer Quarterly, 59, 141-167.
Shtulman, A., Neal, C., \& Lindquist, G. (2016). Children's ability to learn evolutionary explanations for biological adaptation. Early Education and Development, 27, 12221236.

Shtulman, A., \& Schulz, L. (2008). The relationship between essentialist beliefs and evolutionary reasoning. Cognitive Science, 32, 1049-1062.
Teasley, S. D. (1995). The role of talk in children's peer collaborations. Developmental Psychology, 31, 207-220.
Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. Learning and Instruction, 4, 45-69.
Wiser, M., \& Amin, T. (2001). "Is heat hot?" Inducing conceptual change by integrating everyday and scientific perspectives on thermal phenomena. Learning and Instruction, 11, 331-355.
Young, A. G., Alibali, M. W., \& Kalish, C. W. (2012). Disagreement and causal learning: Others' hypotheses affect children's evaluations of evidence. Developmental Psychology, 48, 1242-1253.

# A Data Driven Approach for Making Analogies 

Mei Si (sim@rpi.edu)<br>Department of Cognitive Science, Rensselaer Polytechnic Institute Troy, NY 12180 USA<br>Craig Carlson (carlsc2@rpi.edu)<br>Department of Computer Science, Rensselaer Polytechnic Institute<br>Troy, NY 12180 USA


#### Abstract

Making analogies is an important way for people to explain and understand new concepts. Though making analogies is natural for human beings, it is not a trivial task for a dialogue agent. Making analogies requires the agent to establish a correspondence between concepts in two different domains. In this work, we explore a data-driven approach for making analogies automatically. Our proposed approach works with data represented as a flat graphical structure, which can either be designed manually or extracted from Internet data. For a given concept from the base domain, our analogy agent can automatically suggest a corresponding concept from the target domain, and a set of mappings between the relationships each concept has as supporting evidence. We demonstrate the working of this algorithm by both reproducing a classical example of analogy inference and making analogies in new domains generated from DBPedia data.


Keywords: creativity; analogy; intelligent agents

## Introduction

This work proposes a data-driven approach for dialogue agents to make analogies between concepts. Analogies describe the comparative relationships between two sets of concepts, i.e. concepts A and B are related in a similar way to how concepts C and D are related. Analogies are widely used in writings and dialogues for explaining new concepts or for making the narration more vivid and more interesting. Typically, one set of concepts is more familiar to the audience than the other. Analogies can, therefore, help the audience understand concepts in unfamiliar domains.

Though making analogies is natural for human beings, it is not a trivial task for dialogue agents. There are at least two challenges associated with this task. One is how to find out and represent what people know about a domain. The other is the computational complexity of establishing mappings between two domains. Both challenges become more significant when the domains the agent tries to make analogies with are not defined explicitly. For example, it is much harder to represent what people know about music genres than linear algebra. There is both more uncertainty and more information in the first case. In addition, there may be multiple good mappings between the concepts in the two domains. For example, one's life can both be
mapped to a tree or a road depending on the purpose of making the analogy.

Many cognitive theories have been proposed for explaining how people form analogies (Keane, 2012; Kubose, Holyoak, \& Hummel, 2002; Larkey \& Love, 2003). Structure-Mapping Theory (SMT) is one of most influential theories for analogies and has been supported by a number of empirical studies using human subjects (Falkenhainer, Forbus, \& Gentner, 1989; Gentner, 1983; Gentner \& Smith, 2012). According to SMT, an analogical mapping is created by establishing a structural alignment of relationships between two sets of concepts (in two different domains). The closer the structural match is, the more optimal the inferred analogy will be.

One of the main challenges of implementing SMT is its computational complexity. Many researchers have pointed out that the computational time of establishing the mapping is intractable. Heuristics and alternative theories have been developed to form analogies and cut down the computational time. Holyoak and Thagard's Multiconstraint Theory reduces analogy inference to a constraint satisfaction problem (1989). (Forbus \& Oblinger, 1990; Grootswagers, 2013; van Rooij, 2008; Wareham, Evans, \& van Rooij, 2011) have all worked on creating heuristics for speeding up the structural mapping process.

Another challenge comes from applying SMT or other similar theories to dialogue agents. They typically require a hierarchical relationship structure in the data. For example, the analogy between the solar system and the Rutherford model is a classic example used in computational models of analogy. Figure 1 is taken from (Falkenhainer, Forbus, and Gentner, 1989) for illustrating the solar system domain. For representing this domain, SMT prefers to know not only the relationships between the concepts, e.g. the planet revolves around the sun, the sun's mass is greater than the planet's mass, and the sun attracts the planet but also the relationships among relationships, i.e. the latter two relationships are the cause for the first one. When designing virtual characters with automatically generated or crowd-sourced dialogue content, we often do not have such hierarchical information. The alternative is to design content solely by hand, which creates a huge authoring burden. This challenge is particularly significant when we study
analogy inference not only for understanding human cognition but also for procedurally generating dialogues for virtual characters.


Figure 1: Solar System

In this work, we experiment with loosening up the constraints on input data and using a flat graphical structure for representing the agent's knowledge, i.e. our proposed algorithm only needs to know the relationships between each pair of concepts. Instead of mapping the structures of the relationships, we seek to map the type of relationships from one domain to another. This algorithm is completely data driven; there is no manually designed mapping rule. Our algorithm generates comparable results with SMT when being applied to a classical analogy inference example. We also demonstrate applying the algorithm to larger domains that were automatically generated by crawling data from DBpedia (Bizer, Lehmann, Kobilarov, Auer, Becker, Cyganiak, \& Hellmann, 2009).

The results from the analogy-making module will be integrated into an automated narrative agent we developed for making presentations using data gathered through crowdsourcing or from the Internet (Si, Battad, \& Carlson, 2016). The success of this project will contribute greatly to creating interesting dialogues and computational creativity. The analogy-making module is self-contained, and the details of the presentation agent are skipped in this paper. In the next sections, we will first describe our input data's format and example domains. Then, we will present our analogy-making algorithm, and results generated by this algorithm, followed by discussions and future work.

## Example Domains and Knowledge Representation

We want to use a knowledge representation that is both compatible with structured data, such as the results from querying DBpedia, and is intuitive enough for nontechnical authors to manually design and edit the knowledge base. We use a XML format that encodes knowledge as a directed graph. Each concept is represented
as a node with a unique ID. The nodes are linked to each other by their relationships, and thus form a directed graph.

We will demonstrate the application of our algorithm using two examples. The first one makes analogies between the solar System and the Rutherford model. This example has been discussed extensively by Gentner et al. (see (Falkenhainer, Forbus, \& Gentner, 1989; Gentner, 1983) for more detailed descriptions of the example.) Figure 2 shows the solar system represented as a knowledge graph in our system. Because we don't use hierarchical relationships in our data representation, the higher-level relationships, such as "And" and "Cause" in Figure 1 are lost. However, the relationships between each pair of concepts, such as "Attracts" and "Revolve" are kept. We created a new relationship "More massive" for representing the sun's mass is greater than the planet. Our representation does not use attributes and functions. For attributes, we converted them into a relationship the concept has with another concept, e.g. the sun has a relationship with a concept called Yellow. Currently, we don't have a corresponding encoding for SMT's concept of function in our system.


Figure 2: Solar System without Hierarchical Relationship Structure

The first example only contains about a dozen concepts. For examining how well our algorithm scales up, we created a second set of example domains which are much larger. One domain is about music genres, and the other is about programming languages. In this work, we used Wikipedia data as the base of knowledge. The two domains are generated by crawling for information from DBpedia using a tool we developed in the lab. The tool uses one or more DBpedia entries as the starting points and iteratively expanding the graph by including neighbors of the entries that are already in the graph.

Each entry in DBpedia is converted to a node in our knowledge graph and represents a unique concept. The type of link between them in DBpedia becomes the relationship link in our data. These domains are significantly larger than the ones in the first example. The music genres domain contains 999 nodes and 6418 relationships. The programming language domain contains 2589 nodes and 9952 relationships. Figure 3 shows part of the data from the music domain.


Figure 3: Music Genre Data

## Proposed Approach

In this work, our main objective is to provide a dialogue agent or a game character a tool for conducting richer and more interesting dialogues, or for making explanations for a new concept to the user. We hope to help create dialogues that are creative and innovative. Therefore, different from most existing work on analogy inference, we do not necessarily need to find the best analogy we can make given the two domains. Instead, we want to be able to make analogies that are interesting and explainable. Furthermore, the computation needs to complete in a reasonable amount of time.

Our proposed algorithm follows the same philosophy as SMT in that we want to find mappings between concepts and relationships that are supported by mappings between other concepts and relationships. In other words, we want all of the mappings to be consistent with each other. Our algorithm seeks to achieve these goals while working with large and uniformly structured data.

More specifically, instead of trying to map a relationship structure, we seek to map relationship types from one domain to another. These mappings are supported by the similarities in the concepts being linked to, and the relationships related to those concepts. Because our data is large and not manually designed, there may not be a single mapping that is better than all the alternatives. Instead, there may be multiple good candidates. Therefore, instead of looking for the best mapping for all possible hypotheses between the concepts and the relationships in the base and destination domains, we seek to find the best analogy we can make just about a single concept.

Algorithms 1-4 contain the pseudo code for our proposed algorithm. On a high level, it works in two steps: 1) computes a unique index for each concept and each relationship type. This index can be used for comparing the similarities between two concepts or two relationship types; 2) generates and evaluates the hypotheses of mapping a concept in the base domain to a concept in the target domain.

```
Algorithm 1 Index_Relationship_Type (domain):
loss, gain, same, diff, index = {} # empty dictionar-
ies
# n: concept; r: relationship; d: destination concept
of r
for each n in domain do
    for each r,d of n do
        # compare n's relationships with d's relation-
    ships
        Loss[r] += n.relationship - d.relationship
        gain[r] += d.relationship - n.relationship
        same[r] += Common(n.relationship, d.relationship)
        diff[r] += Difference(n.relationship,
                            d.relationship)
    end for
end for
for each r in domain do
    index[r] = (Jaccard_index(Loss[r], gain[r]),
        Jaccard_index(Loss[r], same[r]),
        Jaccard_index(Loss[r], diff[r]),
        Jaccard_index(gain[r], same[r]),
        Jaccard_index(gain[r], diff[r]),
        Jaccard_index(same[r], diff[r]))
end for
return index
```

Algorithm 1 creates a vector of size 6 for describing each relationship type in a domain. Inspired by the structural mapping process in SMT, here we argue two relationship types are similar if they are always used in similar contexts. Because we don't have the relational structure for providing a context, we operationally defined the context as the origin and the destination concepts linked by the relationship, and we judge the similarity of these two concepts by looking at the differences between the relationships they have and what they share in common. For example, for the relationship "Hotter than" in Figure 2, n is Sun and d is Planet. The loss set, in this case, equals to ["More massive", "Is"]. It contains all the relationships the Sun has, but the Planet does not have. If "Hotter than" also links other concepts in the knowledge base, the loss set will be appended every time this relationship is used. The gain set contains all the relationships the destination concept has, but the origin concept doesn't. The same set contains all the relationships the origin and the destination concepts have in common, and diff contains all the relationships that are either in loss or gain. Currently, we are only using the measurements that represent the results of basic set operations, i.e. complement, intersection, symmetric and difference. As part of our future work, we will be looking for other measurements that can help with differentiating the relationship types.

For each relationship type, Algorithm 1 aggregates the results from every time it is used in the domain. The second for-loop converts the information in the four sets, i.e. loss, gain, same and diff into a one-dimensional vector by calculating the Jaccard indices between them. We used Jaccard index because it can provide a numerical measurement of the similarities between two sets.

```
Algorithm 2 Get_Node_Index (n,rtype_index):
\# rtype_index: the relationship indexes computed by
Algorithm 1
\# n: concept; r: relationship
tmpv \(=(0,0,0,0,0,0)\) \# a zero vector
for each \(r\) of \(n\) do
    tmpv += rtype_index[ \(r\) ]
end for
return Normalize(tmpv)
```

Based on the relationship indices computed by Algorithm 1, Algorithm 2 returns an index for a concept. Similarly, this index will be used for computing the difference between two concepts in the knowledge network. We used a simple heuristic here: a concept's index is decided by the sum of the index values of all the relationships it has. This value is then normalized to a unit vector.

Finally, Algorithm 3 generates and tests the matching hypotheses, and Algorithm 4 creates a one-to-one mapping between all the relationships a concept $n$ has in the base domain to the relationships in the target domain. With this mapping, it is straightforward to find the concept in the target domain that has the most mapping relationships with $n$.

For establishing the mapping, Algorithm 3 generates all the possible hypotheses of mapping the relationships and destinations ( $r l, d l$ ) associated with $n$ to another pair of relationship and destination $(r 2, d 2)$ in the destination domain. For evaluating the quality of this mapping, Algorithm 3 looks at both how different the two relationships ( $r 1$ and $r 2$ ) are -- rdiff, and how different the two destinations ( $d 1$ and $d 2$ ) are $--n d i f f$. The difference here is given by the cosine similarity between the two vectors. These two difference values are combined. The smaller the overall difference is, the stronger the mapping is. The variable "hypotheses" contains the list of all the hypotheses and their strengths.

```
Algorithm 3 Generate_Hypotheses( \(n, B, T\) ):
\# B: base domain
\# \(T\) : destination domain
hypotheses = [] \# hypotheses for mapping
rtype_index = Index_Relationship_Type( \(B, T\) )
\# index for source node
svec = Get_Node_Index(n, rtype_index)
for each node \(t\) in \(T\) do
    \# index for candidate node
    cvec \(=\) Get_Node_Index(t, rtype_index)
    for each \(r 2\), \(d 2\) of \(t\) do
        for each \(r 1\), \(d 1\) of \(n\) do
            rdiff \(=\) Cosine_Similarity(rtype_index[r1],
                                    rtype_index[r2])
            d1vec = Get_Node_Index(d1, rtype_index)
            diff1 = svec - divec
            d2vec \(=\) Get_Node_Index(d2, rtype_index)
            diff2 \(=\) cvec - d2vec
            ndiff \(=\) Cosine_Similarity(diff1, diff2)
            normalized_score \(=(\) rdiff + ndiff) \(/ 2\)
            hypotheses.Append(normalized_score, r1, d1, r2, d2)
        end for
    end for
```

end for
return hypotheses

Similar to SMT, we want the mappings to be unambiguous. We used a greedy algorithm to resolve the conflicts in the hypotheses. In case there are hypotheses for both mapping $(r 1, d 1)$ to $(r 2, d 2)$, and to $(r 3, d 3)$ in the destination domain, we simply accept the best -- the mapping that has the highest score -- hypotheses first, and reject any subsequent mappings that intend to revise an existing one (Algorithm 4).

```
Algorithm 4 Map_Relationships(hypotheses):
map = {}
# sort the hypotheses based on normalized_score
hypotheses.Sort_Descending()
for each h in hypotheses do
    # ensure a one-to-one mapping
    if both h.r1 and h.r2 are not mapped then
        # map r1 in base to r2 in destination
        map[r1] = r2
    end if
end for
return map
```


## Example Results and Discussion

The proposed algorithm has been applied to making analogies in the two example scenarios described in the Example Domains and Knowledge Representation section.

## The Solar System and the Rutherford Model

For making analogies between the solar system and the Rutherford model, we obtained perfect results. Our algorithm correctly generated the mapping between the Nucleus and the Sun, and between the Electron and the Planet. Our algorithm does not produce mapping relationship structure for supporting the analogy. Instead, it produces matching pairs of relationships and destination concepts. Tables 1 and 2 list the evidence for these two mappings.

Table 1: Mappings between Nucleus and Sun

| Nucleus | Sun |
| :--- | :--- |
| (Attracts, Electron) | (Attracts, Planet) |
| (Distance, Electron) | (Distance, Planet) |
| (Has, Electric charge) | (Has, Mass) |
| (More massive than, Electron) | (Hotter than, Planet) |

In Table 1, all the mappings except the last one are straightforward. We checked the intermediate results. The last mapping was an artifact. (Hotter than, Planet) and (More massive than, Planet) received the same score, and the system did not know how to break the tie. All the mappings in Table 2 are consistent with the original example. We are quite encouraged to get this result without the need of using data with hierarchical relationships. We believe the flat concept-relationship structure we designed in Fig-
ure 2 is friendlier to both human designers and automated programs that convert data from other sources.

Table 2: Mappings between Electron and Planet

| Electron | Planet |
| :--- | :--- |
| (Attracts, Nucleus) | (Attracts, Sun) |
| (Distance, Nucleus) | (Distance, Sun) |
| (Has, Electric charge) | (Has, Mass) |
| (Revolves around, Nucleus) | (Revolves around, Sun) |

## Music Genres and Programming Languages

Making analogies between these two domains generated some interesting results, and inspired us with directions for future work. These domains are much larger than the solar system and the Rutherford model. In our evaluation, the typical running time is less than a second on a Lenovo T430 laptop. We will discuss two pieces of example results below.

Table 3: Mapping Relationships between Punk rock and

| LPC |  |
| :--- | :--- |
| Punk Rock | LPC |
| Music fusion genre | Influenced |
| Stylistic origin | Influenced by |
| Instrument | Paradigm |

In the first example, the system mapped the music genre Punk Rock to the programming language LPC. Because of space limitation, in Table 3 we only list the matching relationship types the system provided for this analogy. Two of these mappings are quite reasonable. By mapping "Stylistic origin" to "Influenced by", the system provided us supporting evidence such as "the stylistic origin of Punk Rock is Garage Rock, Glam Rock, and Surf Music, just like LPC is influenced by Lisp, Perl, and C." By mapping "Music fusion genre" to "Influenced", the system provided corresponding supporting evidence "Celtic Punk is a music fusion genre of Punk Rock, just like LPC influenced Pike." Intuitively, these two examples make sense. Music genres that are influenced by other music genres have their styles originating from those genres. The inverse works as well; if genre A is a music fusion genre for genre B , then A influenced B. The system was able to equate these relation types without any explicit help.

The mapping from "Instrument" to "Paradigm" isn't as a clear cut as the other mappings. The evidence provided is "the relationship between Punk rock and Bass guitar or Electric guitar is Instrument, just like the relationship between LPC and Procedural programming and Functional programming is Paradigm." This assertion isn't inherently wrong. However, to a human observer this mapping may not seem intuitive enough.

Interestingly, we also asked the system to make an analogy about the programming language Python, and the system responded with Hardcore Punk. Table 4 provides the matching relationships for this analogy.

Most of the relationship mappings in Table 4 are reasonable. For example, we can say "Python influenced F Sharp, Ruby, and Swift just like Black Metal, Thrash Metal, and Industrial Metal are derivatives of Hardcore Punk."

Table 4: Mapping Relationships between Python and Hardcore Punk

| Python | Hardcore Punk |
| :--- | :--- |
| Influenced | Derivative |
| Influenced by | Stylistic origin |
| Operating system | Instrument |
| Paradigm | Format |

The most interesting part of this example is the assertion that Python is influenced by Perl in the same way as the stylistic origin of Hardcore Punk is Punk Rock (and hence Hardcore Punk is influenced by Punk Rock). This goes against the previous analogy of Punk Rock being comparable to LPC, since Python is not influenced by LPC. We think this shows the weakness of our approach. Without the hierarchical relationship information which in fact provides a global structure of the data, our algorithm does not do a good job in creating analogies that are globally consistent. However, the analogies are still locally consistent for a given topic because of Algorithm 4.

Another thing to note is the difference in mapping between Instrument and Paradigm. In Table 3, "Instrument" is mapped to "Paradigm", but in Table 4, "Instrument" is mapped to "Operating system." LPC does not have a relationship of the type "Operating system", so no mapping could have been made. Table 4 indicates that "Instrument" is more analogous to "Operating system" than to "Paradigm." As mentioned before, our system cannot enforce global consistency yet. Realistically, however, it's hard to say which is truly correct in this case. A similar phenomenon can be observed with "Influenced" and "Music fusion genre" in Table 3. This time in Table 4, "Influenced" is mapped to "Derivative" because the match is better, not because Hardcore Punk lacks that relation type. In Table 4, mappings from "Influenced" to "Music fusion genre" are ignored because a one-to-one mapping of relation types is enforced by the algorithm. Currently, one-to-one mappings must be enforced in order for coherent analogies to be made. However, there are cases when using many-to-one mappings is more suitable. This is especially true when using crowd-sourced data or data from the Internet where sometimes the only real difference in relationship type is semantics (e.g. "Instrument" / "Instruments").

## Future Work

We have planned future work both in the direction of improving our algorithms for finding better mappings and discovering more creative uses of the algorithms.

First of all, we want to address the issue of the agent sometimes creating conflicted mappings between two pairs of concepts. When working with a large data set, exclusively checking all the possible conflictions would be very time-consuming. Instead, we plan to develop a greedy solution. When the agent needs to make a new analogy, it will assume all the relationship mappings it used to support its previous analogies are already true. This way, instead of asking two separate questions of "what is LPC like" and "what is Python like", we are asking the system "If LPC is like Punk Rock, what Python would be like?"

Secondly, we are looking for better ways for indexing the relationships and the concepts. Right now, the semantic information of the relationship types is rarely used. Algorithm 1 only looks at whether they are different or not. We are considering using other semantic tools for helping us to get a direct measure of how close two relationship types are, and even how close two concept descriptions are. This would solve the aforementioned problem caused by the one-to-one mapping restriction. Another consideration in the indexing process is the fact that when dealing with human authored content, there is no guarantee that different contributors will use the same relation type in the same way. Such inconsistencies could throw off the results of Algorithm 1, leading to bad analogies.

Thirdly, many benchmarks have been created for analogy inference, such as (Holyoak \& Thagard, 1989). Most of the benchmarks' formats are compatible with SMT and the algorithms derived from it. We will be looking into ways of evaluating our algorithm using a standard benchmark.

Finally, we believe this work has great potential of contributing to creating rich and vivid virtual characters, interesting and interactive stories, and computational creativity. We are interested in finding new and innovative applications of our proposed algorithms in addition to making analogies for a single concept. In particular, we are interested in exploring how these algorithms can be used in creating digital stories. As one of our next steps, we plan to experiment with using this algorithm to learn how a person tells a story or how a good story is constructed and then apply the learning results for telling new stories using data from a different domain.

## Acknowledgments

This work was supported in part by the Cognitive and Immersive Systems Laboratory -- a collaboration between the IBM Research and RPI in the IBM Cognitive Horizons Network.

## References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., \& Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. Web Semantics: science, services and agents on the world wide web, 7(3), 154-165.
Falkenhainer, B., Forbus, K., \& Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial Intelligence, 41, 1-63.
Forbus, K., \& Oblinger, D. (1990). Making SME greedy and pragmatic. Proceedings of the 12th Annual Conference of the Cognitive Science Society (pp. 61-68).
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7 (2), 155170.

Gentner, D., \& Smith, L. (2012). Analogical reasoning. In V. Ramachandran (Ed.), Encyclopedia of human behavior (2nd ed.) (pp. 130-136). Elsevier; Oxford, UK.
Grootswagers, T. (2013). Having your cake and eating it too: Towards a fast and optimal method for analogy derivation. Master dissertation, Radboud University, The Netherlands.
Holyoak, K., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive Science, 13 (3), 295-355.
Kline, P. J. (1983). Computing the similarity of structured objects by means of a heuristic search for correspondences. Doctoral dissertation, University of Michigan.
Kubose, T. T., Holyoak, K. J., \& Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. Journal of memory and language, 47(3), 407435.

Larkey, L. B., \& Love, B. C. (2003). CAB: Connectionist analogy builder. Cognitive Science, 27(5), 781-794.
Si, M., Battad, Z. \& Carlson, C. (2016) Intertwined storylines with anchor points. Proceedings of the 9th International Conference on Interactive Digital Storytelling (ICIDS) (pp 247-257), Los Angeles, CA.
van Rooij, I. (2008). The tractable cognition thesis. Cognitive science, 32(6), 939-984.
Wareham, T., Evans, P., \& van Rooij, I. (2011). What does (and doesn't) make analogical problem solving easy? A complexity-theoretic perspective. The Journal of Problem Solving, 3 (2), 30-71.

# Word-object associations are non-selective in infants and young children 

Sia Ming Yean (khyx5smy@nottingham.edu.my)<br>School of Psychology, The University of Nottingham Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor Darul Ehsan, Malaysia.

Julien Mayor (julien.mayor@psykologi.uio.no)<br>Department of Psychology, University of Oslo, Forskningsveien 3A, 0317 Oslo, Norway.


#### Abstract

For decades, theories of early word learning have assumed that infants are equipped with learning biases that help them learn words at a fast pace. One of these biases, called Mutual Exclusivity, suggests that infants reject second labels for name-known objects. Our first two experiments, with children and with infants, suggest that novelty preference during Mutual Exclusivity tasks should not be taken as evidence that associations between novel labels and name-known objects have not taken place. A third experiment, supplemented with computational modeling, ruled out cascaded activation patterns as alternative explanations and, instead, confirmed that word-object associations are non-selective throughout infancy and childhood.


Keywords: Mutual exclusivity; early word learning; crosssituational statistical learning

## Introduction

Children learn words at a fascinating pace (Bloom, 2000). Researchers have suggested that infants are equipped with language learning biases that help them learn words efficiently (Markman, 1990). One such word learning bias, called Mutual Exclusivity (ME; Markman \& Wachtel, 1988), suggests that each object has only one label. Markman and Wachtel (1988) found that children selected a novel object significantly more than a name-known object when hearing a novel label. The associations formed between novel labels and novel objects through ME have been shown to be retained, as Mather and Plunkett (2011) found that children were able to match novel labels with the matching novel objects they were exposed to during a learning phase. This provided evidence that ME can indeed be used to learn words. Yet, it did not address the question of whether infants reject additional labels for name-known objects or not.

In parallel, Smith and Yu (2008) found that when only name-unknown objects are present, infants retain multiple associations between labels and objects. According to their cross-situational statistical learning account (CSL), different word-object associations are being retained and their strengths evolve along with the presentation of labels and objects. Ultimately, a hierarchy of word-object associations is established through the differing numbers of cooccurrences between words and objects. The strongest word-object associations can be seen as providing a basis
towards establishing robust patterns of word learning. This framework suggests that, contrary to ME, children are capable of forming more than one association between objects and words.

Further evidence brought a nuanced view to strict ME accounts. Learners were found to be able to overcome ME and performed above chance when forming two-to-one mappings in CSL-type experiments (Yurovsky and Yu, 2008). Kachergis, Yu and Shiffrin (2009) and Poepsel and Weiss (2014) found that although children performed better in one-to-one mapping, they readily violated ME if there was strong evidence that a new mapping was required.

Yet, Trueswell, Medina, Hafri, Gleitman (2013) found that children do not store all word-object associations. Instead, they make an initial guess and evaluate the validity of this guess in subsequent trials. If the guess is proven to be correct, then the association is strengthened. Otherwise, children will make another guess while discarding previously-made associations. This strategy was coined as a Propose-but-Verify hypothesis (PbV). In a more recent study, Stevens, Gleitman, Trueswell, and Yang (2016) refined their PbV hypothesis, and suggest that children store previous associations as references for future trials.

In our present contribution, we ask whether infants and young children accept second labels for name-known objects. To this end, we adapted a classic ME task, in which novel labels are being uttered while a name-known object and a novel object are being displayed. Strict ME accounts would suggest that children will map novel labels to novel objects, and that they will reject the formation of an association between novel labels and name-known objects.

In our adaptation, the learning phase featured two sets of ME training trials: in each set a novel label was uttered in the presence of a name-known object and of a novel object. During the testing phase, both novel objects used during training were shown together. If children were able to retain the associations learned during ME practice, they would be able to map each novel label to the matching novel object. This testing phase was aimed at verifying that ME can indeed be used to learn words, thus replicating Mather and Plunkett (2011). In another testing block, the two nameknown objects used during ME were shown together, while playing one of the novel labels used during training. According to a strict interpretation of ME, children should not display preference for either objects, since they should
have inhibited the formation of an association from the novel word to the name-known objects. In contrast, if children displayed a preference for the name-known object of the matching set, this would suggest that infants do not reject additional labels for name-known objects, and that they are non-selective when forming word-object associations. In other words, children would map the novel word with all objects present in the scene.

Thereafter, we will present three experiments aiming at refining our understanding of the formation of early wordobject associations.

## Experiment 1

## Methods

Participants 174 children were recruited in Nottingham (UK) from which only data from English monolinguals ( $N=$ 148) was analysed, as bilinguals are expected to differ from monolinguals in ME-related tasks (e.g., Byers-Heinlein \& Werker, 2009; Davidson \& Tell, 2005; Houston-Price, Caloghiris, \& Raviglione, 2010). The participants' age ranged from 4 years to 12 years ( $M=7.36$ years, $S D=$ 2.06). Among them, 65 were male and 83 were female.

Stimuli The visual stimuli (pictures of $640 \times 480$ pixel) were obtained from Frank, Sugarman, Horowitz, Lewis, \& Yurovsky (2016). The novel labels were "dax" and "modi" and both novel words and familiar words were embedded in the carrier sentence "Find the _ !". Auditory labels were recorded by a female native English-speaker in an infantdirected manner.

Procedure The experiment was carried out on an iPad. The participants had to first complete a warm-up task, where they were instructed to tap on five dots appearing in random places, followed by five smiley faces presented on the screen of the iPad. After the warm-up task, the experiment was started. There were three experimental blocks, namely a Mutual Exclusivity (ME) training, a Word Learning test and a Selectivity test (see Figure 1).


Figure 1: Example of different block types in Experiment 1.

In ME training, two sets of stimuli were used. Each of the sets consisted of one pair of images, one name-known image and one novel image, along with a novel label ("dax" and "carrot" for the first set, "modi" and "cat" for the other set). In each ME trial, one pair of images was displayed while the novel label was being played (embedded in the carrier sentence "find the __!"). We defined the target to be the novel object. Participants had the opportunity to learn two novel labels for the two novel images via ME. Each pair was repeated four times.

In the Word Learning test, both novel images (the "targets" in ME training) were displayed side-by-side while one of the corresponding novel labels used during training was played. These trials were repeated four times, such that both novel labels were uttered twice. The aim was to test whether the participant had formed an association between a novel label and the corresponding novel image; a prerequisite for word learning.

In the Selectivity test, the two name-known images (the "distractors" in ME training) were displayed side-by-side while one of the novel labels used for training was played. Target selection (i.e., tapping on the name-known image from the matching set) would provide evidence that an association between the name-known image and a second label was not inhibited during ME training. This block also consisted of four trials. The order of presentation of the Word Learning test block and the Selectivity test block was randomised across participants.

## Results and Discussion of Experiment 1

As children were required to select an image out of two presented to them, binomial tests were run to measure the proportion of accurate responses as compared to chance (.50) in the different blocks. The proportion of accurate responses for the different blocks may be seen in Table 1.

Table 1: Observed proportion of accurate responses in different blocks

| Auditory <br> stimuli | ME <br> training | Word <br> learning test | Selectivity <br> test |
| :--- | :--- | :--- | :--- |
| "dax" | $.99^{* * *}$ | $.98^{* * *}$ | $.60^{*}$ |
| "modi" | $.98^{* * *}$ | $.96^{* * *}$ | $.69^{* *}$ |
| $* p<.05 . * * p<.01 . * * * p<.001$ (1-tailed). |  |  |  |

The results obtained provided evidence for the occurrence of ME during training, where the children selected the novel image upon hearing a novel word, thus replicating classical ME results (e.g., Golinkoff, Hirsh-Pasek, Bailey, \& Wenger, 1992; Mather \& Plunkett, 2011).

Children also performed above chance in the Word Learning test, indicating that they were able to form an association between the novel images and the novel words. This also replicates previous studies (e.g., Mather \& Plunkett, 2011; Mervis \& Bertrand, 1994;), which suggest that ME can indeed be used to learn words.

Crucially, it was also found that children were able to form associations between novel words and name-known objects, suggesting that word-object associations are nonselective.

Age was not found to correlate significantly in the Selectivity test ( $r=.113, p=.171$ ), suggesting that younger infants going through a rapid expansion in their vocabulary may also be able to associate novel labels to name-known objects.

## Experiment 2

Experiment 2 aimed at establishing whether infants are also non-selective when forming word-object associations. The design of Experiment 2 was similar to the previous experiment, with the main difference that infants were tested using an eye-tracker.

## Methods

Participants Forty-three 21-month-old infants ( $M=21.01$, $S D=0.57$ ) participated in the experiment ( 25 boys). Nine additional children were tested but excluded because of fussiness (4), failed calibration (4) or software problem (1). All participants were French native speakers recruited in the canton of Geneva, Switzerland.

Stimuli Four novel labels were created for the experiment: "pogalle", "pizelle", "nidoupe" and "loutade". The novel labels were all defined to be of feminine gender, so that no disambiguation could be applied before the onset of the word itself. In addition to the novel labels, eight familiar words were used throughout the experiment. All words were embedded in the sentence "Regarde la __!" (Look at the !). All auditory stimuli were recorded by an enthusiastic French native speaker of Switzerland in a child-directed manner. Visual stimuli were photographs of objects on a light grey background which were extracted from the NOUN database (Horst \& Hout, 2016).

Procedure The procedure of Experiment 2 was similar to Experiment 1 with the exception of two changes, listed thereafter. First, Experiment 2 was conducted with an eyetracker. Thus, looking preference was analysed as opposed to target selection in Experiment 1. Second, another set of images and labels was used.

Infants sat on their caregiver's laps, in front of a flatscreen, approximately 70 cm from the screen. An SMI RED500 eye-tracker recorded infants' fixations at a sampling rate of 500 Hz . The experiment started with a 5point calibration and validation sequence. Upon successful calibration, the experiment started. The calibration procedure was repeated for infants who failed to go through all 5 points or if the validation revealed substantial deviations.

Analysis method Due to data loss in eye-tracking studies, measures typically used in intermodal preferential looking (IPL) paradigms, Preferential Target Looking (PTL) and Longest Look (LL) measures are not appropriate (Wass,

Forssman, \& Leppänen, 2014). A novel analysis was thus introduced similar to the one described by Maris and Oostenveld (2007). The approach is of the model-fitting type, whereby one does not merely compute the maximum likelihood estimates for a set of parameters; the models can also be tested whether or not they are significantly different from each other. The likelihood ratio test then provides the means for comparing the likelihood of the data under the hypothesis that infants are biased towards the target, against the likelihood of the data under the more restricted hypothesis (or null hypothesis) that infants do not have a preference for either the target or the distractor.
A binomial analysis is performed for every time step of every test trial in the post-naming phase, from 367 ms after the onset of the target word (accounting for the time it takes infants to fixate the target object, see Swingley \& Aslin, 2000) for 1500 ms . For each time step the number of infants looking at the target is counted, as well as the number of infants looking at the distractor. The binomial test can reveal if an excess of infants looking at either the distractor or the target at that moment is likely to result from a biased looking behavior (e.g., that infants tend to look more at the target) or if an observed imbalance in the number of infants looking to the target and the distractor can be attributed to mere random variations or noise.

## Results and Discussion of Experiment 2

Results Figure 2 depicts the looking preference of the 21-month-old infants when they were presented during the test phase with the two novel objects used during training (in black) and the two name-known objects used during training (in red). Vertical bars indicate the time steps for which an individual binomial test rejects the hypothesis that infants are not biased towards any object. In both test situations, hypotheses were rejected in favor of a bias towards the target.

Infants display a preference for the target and the loglikelihood that each point belongs to the distribution of unbiased simulated infants can be computed, as it corresponds to the negative square of the Mahalanobis distance (Mahalanobis, 1936). The log-likelihood $L$ that the 21-month infants is unbiased equals $L=1.891$ for the Word Learning test and $L=-44.765$ for the Selectivity test. The statistical relevance of this hypothesis can only be made from the comparison to another model; the biased model.

Log-likelihoods that the 21-month-olds belong to the distribution of simulated infants can be computed for each different bias, and one can estimate the maximum loglikelihood estimate, associated with the optimal bias; e.g., that accounts best for the data.

The maximum log-likelihood equals $L=-0.010$ for a bias towards the target when both novel objects are used during the Word Learning test and equals $L=-0.003$ for a bias towards the target for selectivity test.

Finally, the likelihood ratio test is applied. In the Word Learning test, the 21-month-old infants approach significance $(p=.053)$. On the Selectivity test, the
likelihood ration test shows that infants are significantly biased towards the target ( $p<.001$ ). In other words, they have formed a second association between a novel label and a name-known object; 21-month-old infants also seem to be non-selective when forming word-object associations.


Figure 2: Proportion of infant target looking in the Word Learning test (in black) and in the Selectivity test (in red). Vertical lines highlight significant preference for the target according to a binomial test.

## Experiment 3

In Experiment 1 and 2, children were found to be able to form associations between novel labels and name-known images. Yet, one cannot rule out that infants have formed an encyclopaedic mapping; i.e., they rely on the co-occurrence between the novel object and the name-known object to form an association between both objects. In turn, they may exploit this object-object association to select the nameknown object when hearing the novel label, through cascaded activation from the novel label to novel object (through ME) and from the novel object to the familiar object (through the encyclopaedic mapping). To test if the association between the novel word and the name-known object is lexical or encyclopaedic, a new test, referred to as the Encyclopaedic test, was created. In this test, both novel images used during training were presented side-by-side, but the labels of the name-known distractors were played instead (see Figure 3). As the novel images were never displayed in the presence of these familiar labels, we would expect children's performance to be at chance level.

## Methods

Participants 150 children were recruited at The University of Nottingham. Only data from the 124 English monolingual children ( 55 male, 69 female) was retained for analysis. Children were 3 to 12 years of age ( $M=7.31$ years, $S D=2.28$ ).

Stimuli Visual stimuli were obtained from the same source as Experiment 1. The novel labels used were "pifo" and "dofa" paired with images of a ball and a cup, respectively.

Procedure The procedure was the same as Experiment 1 except for the sequence of the tasks. In Experiment 3, there were four blocks, namely the ME training, the Selectivity test, the new Encyclopaedic test and the Word Learning test. There were four trials in both the Selectivity test and the Encyclopaedic test but only two trials in Word Learning test. The order of the Selectivity test and the Encyclopaedic test was counterbalanced whereas the Word Learning test was always administered at the end of the experiment.

In the Encyclopaedic test, novel images used during ME training were displayed side-by-side while the names of the corresponding familiar images were played (see Fig. 3).


Figure 3: Example of the Encyclopaedic test.

## Results and Discussion of Experiment 3

Binomial tests were run to measure the proportion of accurate responses as compared to chance (.50) in the different blocks. The results once more supported the presence of ME during training (all $p$ 's $<.001$ ) and that word-object associations are retained in the Word Learning test (all $p$ 's $<.001$ ).

However, and contrary to our hypothesis, children selected the target in the Encyclopaedic test (significantly above chance for "cup" ( $p=.001$ ), and approaching significance for "ball", $p=.063$ ), suggesting that infants formed an encyclopaedic mapping between all items, rather than forming multiple associations at a lexical level between a name-known object and multiple labels. The only explanation is that children display evidence of cascaded activation, from the familiar name to the corresponding name-known object, and through the novel object it was paired with.

Does cascaded activation also explain results concerning the Selectivity test? Maybe children are strictly following ME; upon hearing the novel label, they activate the representation of the matching novel object (learned during ME) and select the name-known object that co-occurred with the novel object? How can we distinguish between the cascading and the non-selectivity explanations?

Let us look at correlations between the different experimental blocks. A strict ME account would suggest that lexical associations can only be formed between the novel label and the novel object during ME training. A
stronger ME effect would translate into stronger associations between novel labels and novel objects. Through cascaded activation, novel label to novel object (ME) and from novel object to familiar object (through an encyclopaedic mapping), a strict ME account would suggest a positive correlation between performances in the ME training phase and performance in the Selectivity test.

In contrast, the hypothesis that word-object associations are non-selective (following the arguments of CSL) would predict that the novel label would also be associated with the name-known object. Accordingly, higher accuracy in the ME training block would suggest that children should spend more time on the novel object than at the name-known object, in turn leading to the formation of stronger associations between novel labels and novel objects but weaker associations between novel labels and name-known objects. Thus, if the associations between novel words and novel objects are non-selective, and that such associations take place at a lexical level, the accuracy score in ME training should correlate negatively with the accuracy score in the Selectivity test.

In our data, the accuracy score of ME training correlated positively with the Word Learning test ( $r=.232, p=.009$ ), while correlations between ME and the Encyclopaedic test were not significant ( $r=-.075, p=.406$ ), thus far consistent with both explanations. However, ME training results correlated negatively with performance during the Selectivity test ( $r=-.218, p=.015$ ). The experimental results support the hypothesis that word-object associations are non-selective, and that the formation of association between novel labels and name-known objects take place at a lexical level. Next, we further scrutinise the above reasoning by constructing two simple computational models of Experiment 3.

## Computational Models of Experiment 3

Two models were constructed, in order to compare a strict ME account with a non-selective account of the CSL type. In both models, associations are modulated based on cooccurrence of items presented simultaneously; between objects and labels, as well as between both objects. In each trial, previous association strengths define relative looking time towards each object via the application of Luce's forced choice rule (with a separation parameter of $k=8$ ). Looking time, in turn, modulate the magnitude of the association strength update, in a Hebb-like update rule (with a learning rate of 0.1 ). The associations between both objects are obtained by computing the product of the relative preference associated with each object. Similarly, indirect associations such as cascaded activation from the novel label to the novel object and the familiar object are computed through the product of the association strength between the novel label and the novel object and the association strength between the novel object and the familiar object.

The order of stimuli was identical in the model and for participants. 100 individual models were created for each
hypothesis, and had a mean novelty-preference of 0.8 , and Gaussian random variations of a standard deviation of 0.05 . In the strict ME model, associations between novel labels and name-known objects were inhibited whereas in the nonselective model, such associations were permitted. Correlations between preference for the novel object during ME training and preference for the target object in each test block were computed.

As predicted, modelling results showed that a strict ME account sustains positive correlations between ME training and the Word Learning test $(r=.89, p<.001)$ and between ME and the Selectivity test ( $r=.68, p<.001$ ). In contrast, the non-selective model displays a positive correlation between ME training and the Word Learning test $(r=.89, p$ $<.001$ ) but a negative correlation between ME and the Selectivity test ( $r=-.33, p<.001$ ). The modelling results, along with the empirical results, provide additional evidence that children are non-selective when forming word-object associations.

## General Discussion

While Mutual Exclusivity assumes that children associate only one word to one object, Cross-situation Statistical Learning accounts suggest that children maintain a hierarchy of word-object pairings established on the basis of co-occurrence of object and labels. These two theories have typically been tested using different experimental designs; while ME studies generally feature both novel and familiar objects, CSL studies typically present only novel objects. Our approach aims at testing both theories by using a Selectivity test, so as to examine whether children accept additional associations between a name-known object and a novel label.

In all three studies, we found that children tend to map the novel label to the novel object, replicating classic ME experiments, that infants tend to look at (or select) novel objects upon hearing a novel word.

We also replicated findings that associations between novel objects and novel labels are momentarily retained, thus providing a necessary basis towards the consolidation of word-object associations, and towards word learning (e.g., see Mather \& Plunkett, 2011).

Yet, we argue that evidence of ME - a preference for a novel object when hearing a novel label - should not be taken as proof that associations between novel labels and name-known objects are suppressed. Experiment 1, with school-aged children, and Experiment 2, with 21-month-old infants, provided converging evidence that word-object associations are non-selective: children may well display evidence of ME during training, but they also show evidence that associations between novel labels and nameknown objects are maintained during test blocks, and not inhibited as suggested by strict ME accounts.

Experiment 3 aimed at ruling out alternative explanations for the results observed in the first two studies. The introduction of an additional test block, the Encyclopaedic test, suggested that children displayed cascaded activation
patterns. Yet, correlation analyses between the ME training phase and the different test blocks suggested that the pattern of results can be best explained if children are non-selective when forming word-object associations. Strict ME accounts would predict a positive correlation between ME training and selectivity test as children have to rely on cascaded activation to identify the target. In contrast, non-selective learning accounts would suggest that children displaying stronger novelty preference during ME training would display weaker associations between the novel label and the name-known object, thus leading to a negative correlation between ME training and the Selectivity test. The latter pattern of results was observed experimentally.

The finding that infants are non-selective in their formation of word-object associations sits well with other recent findings that infants are flexible in their interpretation of the meaning of novel words (Ramscar, Dye, \& Klein, 2013) and that infants engage into cross-situational statistical learning with (multiple) objects in their visual field (Yu \& Smith, 2011; Yurovsky, Smith, \& Yu, 2013).

The very first stages of word learning taking place during ambiguous naming situations, such as in ME experiments or cross-situational statistical learning situations, do seem to be principled by low-level associationist mechanisms whereby multiple word-object pairings are being built. The hierarchy of word-object associations can then evolve across situations so that ultimately only relevant word-object mappings are retained. In sum, infants and children appear to be flexible when learning words and readily entertain the possibility that objects can have multiple names.

## References

Bloom, P. (2000) How Children Learn the Meanings of Words, MIT Press
Byers-Heinlein, K., \& Werker, J. F. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. Developmental science, 12 (5), 815-823.
Davidson, D., \& Tell, D. (2005). Monolingual and bilingual children's use of mutual exclusivity in the naming of whole objects. Journal of Experimental Child Psychology, 92(1), 25-45.
Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis. M. L., \& Yurovsky, D. (2016). Using tablets to collect data from young children, Journal of Cognition and Development, 17(1), 1-17.
Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., \& Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. Developmental Psychology, 28(1), 99-108.
Horst, J. S., \& Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. Behavior research methods, 48(4), 1393-1409.
Houston-Price, C., Caloghiris, Z., \& Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. Infancy, 15(2), 125-150.

Kachergis, G., Yu, C., \& Shiffrin, R. M. (2010). Crosssituational statistical learning: Implicit or intentional. Proceedings of CogSci 32 (pp. 2362-2367). Austin, TX: Cognitive Science Society.
Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta), 2, 49-55.
Maris, E., \& Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of neuroscience methods, 164 (1), 177-190.
Markman, E. (1990). Constraints children place on word meanings. Cognitive Science, 14, 57-77.
Markman, E., \& Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive psychology, 20(2), 121-157.
Mather, E., \& Plunkett, K. (2011). Mutual exclusivity and phonological novelty constrain word learning at 16 months. Journal of Child Language, 38(5), 933-950.
Mervis, C. B., \& Bertrand, J. (1994). Acquisition of the Novel Name-Nameless Category (N3C) Principle. Child Development, 65(6), 1646-1662.
Poepsel, T. J., \& Weiss, D. J. (2014). Context influences conscious appraisal of cross situational statistical learning. Frontiers in Psychology, 5(691), DOI: 10.3389/fpsyg.2014.00691

Ramscar, M., Dye, M., \& Klein, J. (2013). Children value informativity over logic in word learning. Psychological science, 24 (6), 1017-1023.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Stevens, J. S., Gleitman, L. R., Trueswell, J. C., \& Yang, C. (2016). The pursuit of word meanings. Cognitive Science, 1-39, DOI: 10.1111/cogs. 12416.
Swingley, D., \& Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. Cognition, 76, 147-166.
Trueswell, J. C., Medina, T. N., Hafri, A., \& Gleitman, L. R. (2013). Propose but verify: Fast mapping meets crosssituational word learning. Cognitive psychology, 66 (1), 126-156, DOI: 10.1016/j.cogpsych.2012.10.001.
Wass, S. V., Forssman, L., \& Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. Infancy, 19 (5), 427-460.
Yu, C., \& Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. Developmental science, 14(2), 165-180.
Yurovsky, D., Smith, L. B., \& Yu, C. (2013). Statistical word learning at scale: the baby's view is better. Developmental science, 16 (6), 959-966.
Yurovsky, D., \& Yu, C. (2008). Mutual exclusivity in crosssituational statistical learning. In Proceedings of the 30th annual conference of the cognitive science society (pp. 715-720). Austin, TX: Cognitive Science Society.

# The tortoise wins only when the race is long: How the task environment changes the behavior of Tetris models 

Catherine Sibert<br>Rensselear Polytechnic Institute, Troy, New York, United States<br>Wayne Gray<br>Rensselaer Polytechnic Institute, Latham, New York, United States


#### Abstract

Tetris can be viewed as a highly complex decision making task, and used as a paradigm for studying human expertise. We hypothesized that models capable of playing Tetris for a long time are doing so by adopting slow but steady strategies to accumulate points, while human players are much more prone to using high-risk, high-reward strategies that earn more points in a shorter time frame. This work used the MindModeling.org computational cognitive modeling platform to develop the best models capable of playing long term games and short term games, and then compared the performance of the two. The best long term model adopted the slow and steady strategy, while the best short term model displayed the higher-risk, higher-reward strategy that more closely matches behavior observed in human players. Models that "trained long" but "played short" did worse than those that both trained and played "short."


# Information density of encodings: The role of syntactic variation in comprehension 

Les Sikos (sikos@coli.uni-saarland.de), Clayton Greenberg (claytong@coli.uni-saarland.de), Heiner Drenhaus (drenhaus@coli.uni-saarland.de) and Matthew W. Crocker (crocker@coli.uni-sb.de)<br>Department of Language Science and Technology, Saarland University, 66123 Saarbrücken, Germany


#### Abstract

The Uniform Information Density (UID) hypothesis links production strategies with comprehension processes, predicting that speakers will utilize flexibility in encoding in order to increase uniformity in the rate of information transmission, as measured by surprisal (Jaeger, 2010). Evidence in support of UID comes primarily from studies focusing on word-level effects, e.g. demonstrating that surprisal predicts the omission/inclusion of optional words. Here we investigate whether comprehenders are sensitive to the information density of alternative encodings that are more syntactically complex. We manipulated the syntactic encoding of complex noun phrases in German via meaningpreserving pre-nominal and post-nominal modification in contexts that were either predictive or non-predictive. We then used the G-maze reading task to measure online comprehension during self-paced reading. Results were consistent with the UID hypothesis. In predictive contexts, post-nominal encodings elicited a more uniform distribution of processing effort. Conversely, in non-predictive contexts, more uniform effort was found for pre-nominal encodings.


Keywords: Language comprehension; surprisal; uniform information density hypothesis; G-maze; self-paced reading.

## Introduction

Levy and Jaeger's (2007) Uniform Information Density hypothesis postulates that speakers adjust their lexical and syntactic realization of a message for the benefit of comprehenders. Specifically, they suggest that there is an overarching preference to produce message encodings that distribute information as evenly as possible across the linguistic signal. This account fundamentally links encoding and decoding processes, asserting that language producers will exploit the flexibility in encoding so as to increase uniformity in the rate of information transmission, as measured by surprisal (Hale, 2001; Levy, 2008). As such, the UID hypothesis can be viewed as part of a rational theory of communication - from an information theoretic perspective - in which encoding strategies take into account resource limitations of the comprehender.

There is robust empirical evidence that surprisal accounts for cognitive load during comprehension - at least at the level of individual words in a sentence (Drieghe, Rayner \& Pollatsek, 2005; Kliegl, Grabner, Rolfs \& Engbert, 2004; Rayner, Aschby, Pollatsek \& Reichle, 2004; Rayner \& Well, 1996). However, there exists little direct online evidence that comprehenders are indeed sensitive to the surprisal and density profiles of alternative syntactic encodings - a critical assumption underlying the UID hypothesis. Furthermore, current support for UID in production is limited to relatively local encoding choices,
such as that-deletion (Jaeger, 2010; Levy \& Jaeger, 2007), contraction (Frank \& Jaeger, 2008), and the use of single word equivalents that vary in word length (chimpanzee vs. chimp; Mahowald, Fedorenko, Piantadosi \& Gibson, 2013). Although the above studies provide important and compelling support for the notion that UID modulates aspects of syntactic encoding, the generality of the findings is limited by the observation that all the phenomena are instances of highly local syntactic reduction.

The goal of the current study is to investigate whether comprehenders are sensitive to the information density of more complex alternative syntactic encodings. Consider the following examples:
(1) The journalist published...

Predictive context
The man evaluated...
Non-predictive context
a. ...[the carefully written essay].
Obj $\mathrm{NP}_{\text {adj }}$
b. ...[the essay that was carefully written].
Obj $\mathrm{NP}_{\text {rel }}$

Each object noun phrase (NP) above arguably expresses the same message, ${ }^{1}$ however (a) uses a pre-nominal adjective phrase while (b) uses a post-nominal relative clause. While the head noun (essay) is more expected in the predictive (1) than non-predictive (2) contexts, the expectation for the adjective carefully presumably does not differ across contexts. One potential encoding strategy for increasing the uniformity of information density would be to produce lowsurprisal words early in the sentence, as this may facilitate the processing of subsequent less predictable words. For instance, in the non-predictive context, the UID hypothesis predicts a processing advantage for the pre-nominal encoding because carefully written should reduce the surprisal of essay. In the predictive context, on the other hand, the pre-nominal encoding may result in a trough in information density at essay, as it is highly expected (and thus not very informative) following the verb and modifiers. In this case, the post-nominal relative clause may distribute the informational content more uniformly. The UID hypothesis therefore, predicts a greater benefit for the relative clause encoding in more predictive contexts.

We tested the above predictions using a self-paced reading design to measure online differences in cognitive load during the critical object NP.

[^490]
## Experiment

Our primary goal was to test whether comprehenders are sensitive during online processing to differences in the information density of alternative syntactically-complex encodings that nevertheless convey a similar message. The materials crossed two factors (context $\times$ encoding), as illustrated in Table 1. Because this manipulation distributes information within the critical region differently across conditions, we assessed cognitive load using a variation of self-paced reading that is less susceptible to spill-over effects than standard forms of self-paced reading. The grammaticality maze task (G-maze; Forster, Guerrera \& Elliot, 2009) can precisely identify the word at which processing time differences emerge during online comprehension (Witzel, Witzel \& Forster, 2010) and is therefore well-suited for our purposes. In this task sentences are presented word by word as a sequence of forced choices between two alternatives, only one of which continues the sentence grammatically. If the participant successfully navigates the "maze" by choosing the correct word from each pair, the selected words form a coherent sentence (Figure 1).

## Methods

Participants Twenty-seven native German speakers (age $M$ $=24, S D=2.6$ ) with normal or corrected to normal vision were recruited from the Saarland University community and were compensated $8 €$ for their participation. Participants that did not successfully navigate at least $70 \%$ of mazes in all experimental conditions were excluded $(\mathrm{n}=3)$.

Materials Forty-eight sets of sentences were constructed in German by crossing context (predictive, non-predictive) and syntactic encoding of the object NP (pre-nominal modification, post-nominal modification), resulting in four conditions per item (Table 1). In order to create the context manipulation, the same object noun (e.g., Essay, "essay") was used in all conditions, but the object was designed to be more expected in predictive contexts than non-predictive contexts. This was accomplished by choosing different subject-verb combinations for each context. Subject-verb
combinations were neutral with respect to the object noun in the non-predictive context (e.g., Der Mann bewertete, "The man evaluated"), but were semantically associated with the object in the predictive contexts (e.g., Der Journalist veröffentlichte, "The journalist published"). Importantly, however, highly expected object nouns (e.g., Artikel, "article") were avoided in order to increase the possibility of detecting surprisal differences between predictive/prenominal and predictive/post-nominal conditions.

The information density of object NPs was manipulated via pre- and post-nominal modification, affecting both the linear ordering and length (in words) of the message. Prenominal modifiers (e.g., sorgfältig verfassten, "carefully written") were shorter, containing 2 to 4 words, but positioned the head noun at the end of the NP. Post-nominal modifiers used a relative clause construction (e.g., der sorgfältig verfasst worden war, "that was carefully written") and were therefore longer, ranging from 4 to 6 words, and constrained the head noun to the beginning of the NP. To avoid having any words within the critical object NP region be sentence-final, all items ended with an adverbial phrase.


Figure 1: Example trial structure of G-maze task. Sentences (in German) were presented word by word as a sequence of forced choices between two alternatives, only one of which continued the sentence grammatically.

Table 1: Example stimulus item in four conditions with approximate English translations. The critical region of interest was the object NP. RTs were analyzed separately for the object noun (bold) and modification region (underlined).

| Context | Encoding | Example |
| :--- | :--- | :--- |
| Predictive | Post-nominal | Der Journalist veröffentlichte den Essay, der sorgfältig verfasst worden war, unter Einbeziehung des größeren Kontextes. <br> "The journalist published the essay that was carefully written, taking into account the larger context." |
| Predictive | Pre-nominal | Der Journalist veröffentlichte den sorgfältig verfassten Essay unter Einbeziehung des größeren Kontextes. <br> "The journalist published the carefully written essay, taking into account the larger context." |
| Non-predictive | Post-nominal | Der Mann bewertete den Essay, der sorgfältig verfasst worden war, unter Einbeziehung des größeren Kontextes. <br> "The man evaluated the essay that was carefully written, taking into account the larger context." |
| Non-predictive | Pre-nominal | Der Mann bewertete den sorgfälig verfassten Essay unter Einbeziehung des größeren Kontextes. <br> "The man evaluated the carefully written essay, taking into account the larger context." |

Four counterbalanced lists were constructed from these materials according to a Latin Square design such that each list contained 12 items in each condition, but no item appeared more than once in the same list. An additional 48 sentences with the same structures as above, but containing highly predictable object nouns, were constructed as fillers (e.g., Der Schneider zerschnitt den stark gemusterten Stoff am Mittwoch., "The tailor cut the heavily patterned fabric on Wednesday."). Half of the filler sentences contained prenominal modification of the object noun and the other half contained post-nominal modification. No object nouns were repeated across experimental or filler items.

Cloze probability and contextual constraint An offline Cloze completion study was conducted to confirm that object nouns were more expected following predictive than non-predictive contexts, but were not highly expected in either context. A separate group of 58 native German speakers (age $M=22.0, S D=2.9$ ) were presented with sentence fragments from the 48 experimental items described above. Fragments contained only the contexts, followed by a blank (e.g., Der Mann bewertete__ ; Der Journalist veröffentlichte___). Predictive and non-predictive contexts were counter-balanced across two lists. Participants were asked to fill in the blank with the first determinernoun combination that came to mind. Cloze probabilities were computed as the percentage of participants who provided the experimental object noun for a particular item. As expected, object nouns had low cloze probabilities in both contexts but were reliably more expected following predictive (cloze $=0.06, S D=0.18$ ) than non-predictive contexts (cloze $=0.00, S D=0.01), t(47)=-2.32, p<.05$.

The percentage of the most frequently occurring response to each sentence fragment in the cloze test was also used to assess the contextual constraint of predictive and nonpredictive contexts. As expected, the mean constraint of predictive contexts was reliably greater (51.3\%) than that of the non-predictive contexts $(21.3 \%), t(47)=-8.46, p<.001$.

Surprisal profiles To compare our response time results against a more theoretical notion of predictability, we computed surprisals for all experimental stimuli using an interpolated modified Kneser-Ney 5-gram language model trained on a 2017-01-01 dump of German Wikipedia. To obtain the corpus, we filtered the original XML dump using the tool WikiExtractor, split the corpus into sentences using the NLTK sentence splitter for German, and preprocessed the resulting dataset. ${ }^{2}$ After replacing all types occurring fewer than 15 times with <unk>, the vocabulary size was $833,734 .{ }^{3}$ We split the corpus into training, development,

[^491]and test sections with the ratio $8: 1: 1$. The resulting training section contained $666,561,150$ tokens. The model was trained using the SRILM toolkit (Stolcke, 2002) and achieved perplexities of 25 on the training section, 201 on the test section, and 1583 on our stimuli. ${ }^{4}$

Procedure Participants were randomly assigned to a stimulus list (6 per list). The G-maze task was implemented in E-prime (Schneider, Eschman, \& Zuccolotto, 2002). Each trial began with two crosses $(+)$ that remained on screen for 1000 ms , indicating where subsequent word pairs would appear. Each word in the sentence (except the first word) was then presented together with a foil word, ${ }^{5}$ which was not a grammatical continuation of the sentence. The first word in every sentence was paired with " $x-x-x$ ". The presentation side (left, right) was randomized such that the correct word appeared equally often on each side. Any punctuation (i.e., comma, period) that appeared with a word also appeared with its foil. Participants were instructed to choose as quickly and as accurately as possible the word that best continued the sentence. Participants indicated their selection by pressing the left or right button on a button box and the amount of time required for selecting the grammatical continuation was recorded as the response time (RT). If the correct word was chosen, the next pair of words appeared automatically. However, if a foil word was selected, negative feedback (Inkorrekt, "Incorrect") was displayed and the trial was aborted. Once the end of a sentence was reached, positive feedback (Korrect, "Correct") was given. Participants initiated each new trial by button press.

To confirm that participants read the sentences for meaning, a Yes/No comprehension question appeared after $1 / 3$ of the items. Half of the questions asked about the subject noun and half about the object noun. The correct answer was Yes for $50 \%$ of questions. Participants used the button box to respond. No feedback was given.

In order to familiarize participants with the task, five practice items (three with comprehension questions) were presented before the experiment began. Participants took approximately 40 minutes to complete the experiment.

[^492]
## Results and Discussion

Completed mazes Overall performance on the G-maze task was high, with participants successfully navigating $85.6 \%$ ( $S D=0.08$ ) of experimental and filler items to completion. However, because the critical region of interest was the object NP, the RT analyses reported below were conducted on all experimental items that were completed through at least the end of the critical region $(M=0.90, S D=0.06)$.

Comprehension question accuracy Performance on the comprehension questions was near ceiling ( $M=0.97, S D=$ 0.04), confirming that participants were reading the sentences for meaning during the G-maze task.

Response time RTs were analyzed with linear mixed effects models with participants and items as crossed, independent, random effects. All models included maximal random effects structures (Barr, Levy, Scheepers \& Tily, 2013). Analyses were conducted using the lmer function (lme4 library, version 1.1-10; Bates \& Sarkar, 2007) in the statistics software package $R$, version 3.2.2 ( R Development Core Team, 2013). Fixed effects were evaluated via likelihood ratio tests implemented in lmerTest (Kuznetsova, Brockhoff \& Christensen, 2015), where denominator $d f$ was estimated using the Satterthwaite method. We report estimates, standard errors, $t$ and $p$ values associated with likelihood ratio tests for significant results only.

All raw RTs that were abnormally low (below 200 ms ) or abnormally high (above 5000 ms ) were excluded ( $0.3 \%$ ), and outliers exceeding 3 standard deviations by participant were then trimmed ( $1.8 \%$ ). The remaining RTs were adjusted for word length (Ferreira \& Clifton, 1986) and punctuation using a linear mixed effects regression model with fixed effects for word length, punctuation (i.e., whether or not a comma or period was presented with the word), and their interaction. The residuals of this model, lengthadjusted RTs, served as the dependent variable in the analyses reported below. ${ }^{6}$ Because the number of words used to modify object nouns varied across items and conditions, we computed the length-adjusted RT for the modification region by averaging across modifier words.

The upper panel of Figure 2 presents the mean lengthadjusted word-by-word RTs for each condition. Differences first emerge at the subject noun, where RTs were slower for predictive than non-predictive contexts. This is not surprising as these words (e.g., journalist) are less frequent than their non-predictive counterparts (e.g., man). More relevant to the research question, all four conditions diverge within the critical object NP region (Table 2).

Object noun analysis. Length-adjusted RTs for object nouns were regressed onto a model including fixed-effect factors for context (predictive, non-predictive), encoding (pre-nominal, post-nominal), and their interaction. In order to control for task adaptation, a main effect of stimulus order was also included.

[^493]Figure 3 (left panel) shows that object nouns were read more quickly in predictive than non-predictive conditions, $\beta$ $=-161.01, S E=29.59, t(44.22)=-5.44, p<.001$. This finding replicates previous work demonstrating that expected linguistic material is easier to process than unexpected material.

Within the non-predictive conditions, pre-nominal modification clearly facilitated the processing of unexpected object nouns. Length-adjusted RTs for object nouns were faster for pre-nominal modification than for post-nominal modification, $\beta=-124.86, S E=30.42, t(23.07)=-4.104, p$ $<.001$. This result is consistent with the UID hypothesis, which predicts a processing advantage for the pre-nominal encoding: pre-modification reduces the surprisal of the unexpected word and results in a more uniform distribution of processing effort across the linguistic signal.

Within the predictive conditions, length-adjusted RTs for object nouns were also faster for pre-nominal modification than for post-nominal modification, $\beta=-51.00, S E=18.67$, $t(29.26)=-2.73, p<.05$. However, the facilitation effect was weaker for predictive conditions, resulting in a context $\times$ encoding interaction, $\beta=73.67, S E=33.02, t(53.56)=$ $2.23, p<.05$. The UID hypothesis predicts a trough in information density for words that are both highly expected and pre-modified. Figure 2 (upper panel) is compatible with this prediction. RTs drop steeply in the predictive / prenominal condition at the object noun. Note that this is true despite the fact that object nouns were selected to be lowcloze. Thus, as shown in Figure 3, the post-nominal condition distributes the informational content more uniformly, resulting in a smoother RT profile.

Modification region analysis. Length-adjusted RTs for the modification region were analyzed using the same mixed effects model as above. Figure 3 (right panel) shows that encoding influenced the processing of the modification region in a way that was complementary to its effect on object nouns (see also Table 2). Pre-nominal modification was read more slowly than post-nominal modification in both contexts, $\beta=106.04, S E=15.02, t(75.75)=7.06, p<$ .001. However the magnitude of this effect was greater in the non-predictive context, reflected in a context $\times$ encoding interaction, $\beta=-53.44, S E=18.69, t(51.68)=-2.86, p<.01$.

Table 2: Mean length-adjusted RT (ms) by condition for object noun (upper panel) and modification region (lower panel). Standard deviation in parentheses.

| Object Noun |  |  |  |
| ---: | ---: | ---: | ---: |
|  | Pre-nominal | Post-nominal | Mean |
| Predictive | $-54(77)$ | $-5(54)$ | $-30(50)$ |
| Non-predictive | $27(70)$ | $56(98)$ | $92(46)$ |
| Mean | $-13(56)$ | $76(62)$ |  |
|  |  |  |  |
| Modification Region |  |  | Pre-nominal |
| Post-nominal | Mean |  |  |
| Predictive | $54(46)$ | $2(39)$ | $28(31)$ |
| Non-predictive | $113(53)$ | $10(30)$ | $61(34)$ |
| Mean | $84(30)$ | $6(22)$ |  |



Figure 2: Upper panel: Mean length-adjusted word-by-word RTs. Lower panel: Surprisal profiles as determined by a language model trained on the German Wikipedia corpus. RTs and surprisals for the modification region were calculated by averaging across all modifiers words. Error bars represent one standard error of the mean.


Figure 3: Mean length-adjusted RTs for object nouns (left panel) and the modification region (right panel). Error bars represent one standard error of the mean.

Surprisal profiles. The lower panel of Figure 2 shows the surprisal profiles produced by the language model. Surprisals at sentence positions 1-4 patterned closely with RTs, reflected in a positive correlation within this region ( $r$ $=0.36$ ). However, the surprisal pattern differed somewhat from the pattern of RTs during the critical object NP region
( $r=-0.06$ ). There are at least two plausible explanations for this divergence. First, the predictable contexts may not have made our atypical (i.e., low cloze) object nouns statistically more predictable, given our German Wikipedia corpus. For instance, verpackte ("boxed", a verb in the predictive context) and Geschenk ("gift", the corresponding object noun) were both present in the corpus but never co-occurred in the same sentence. We assessed this possibility and found that, on average, subject nouns in predictive contexts did not substantially increase the predictability of object nouns above the general case. Verbs, however, did so by a factor of 6 . Second, the dependencies that existed in the training corpus may not have been fully captured by the language model. To test this possibility, we calculated the mean gram size used for surprisal predictions in the object NP region ( $M=1.86$ ). This finding indicates that despite being trained on 5 -grams, the model predictions in this region were based predominantly on more local statistics (i.e., bigrams), effectively modeling only the non-predictive conditions. ${ }^{7}$

Despite these caveats, the results broadly confirm our assumptions about the distribution of surprisal across prenominal and post-nominal encodings of the critical object

[^494]NP: according to the language model, the pre-nominal encodings had more uniform information densities. To capture the behavior found for reading times in the predictive conditions, either a closer domain match between training corpus and stimuli would be required, or a language model architecture that is less sensitive to word position.

## General Discussion

The UID hypothesis links production strategies with comprehension processes and predicts that speakers utilize flexibility in encoding to distribute information as evenly as possible across the linguistic signal (Jaeger, 2010; Levy \& Jaeger, 2007). While prior evidence for UID comes primarily from word-level effects (Frank \& Jaeger, 2008; Jaeger, 2010; Levy \& Jaeger, 2007; Mahowald et al., 2013), a critical assumption underlying the UID hypothesis is that comprehenders should also be sensitive to the information density of alternative syntactically-complex encodings. To our knowledge, the current study is the first to investigate this important and challenging question.

We manipulated the syntactic encoding of complex noun phrases via meaning-preserving pre-nominal and postnominal modification in contexts that were either predictive or non-predictive. The results were consistent with the UID hypothesis. In predictive contexts, post-nominal encodings elicited a more uniform distribution of processing effort than pre-nominal encodings. This makes sense because the head noun is already expected in such contexts, thus prenominal modification could lead to a trough in information density at the noun. Conversely, in non-predictive contexts, a more uniform RT profile was found for pre-nominal encodings, where pre-modification served to reduce the surprisal of the unexpected head noun. This pattern of comprehension results provides indirect support for UID as a rational strategy for producers to adopt. An important question for further investigation is whether speakers are indeed attentive to such factors when making their encoding decisions.

## Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 Information Density and Linguistic Encoding.

## References

Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bates, D., \& Sarkar, D. (2007). Ime4: Linear mixed-effects models using S4 classes. 'R' package. Version 0.9975-12. http://CRAN. R-project.org
Drieghe, D., Rayner, K., \& Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. Journal of Experimental Psychology: Human Perception and Performance, 31(5): 954-969.

Ferreira, F., \& Clifton, C. (1986). The independence of syntactic processing. Journal of Memory and Language, 25(3), 348-368.
Forster, K. I., Guerrera, C., \& Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. Behavior Research Methods, 41, 163-171.
Frank, A. \& Jaeger, F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In Proceedings of the Annual Conference of the Cognitive Science Society, (Vol. 30, No. 30).
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, (pp. 1-8). Association for Computational Linguistics.
Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. Cognitive Psychology, 61(1): 23-62.
Kliegl, R., Grabner, E., Rolfs, M., \& Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. European Journal of Cognitive Psychology, 16: 262-284.
Kuznetsova, A., Brockhoff, P. B., \& Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models (Version 2.0-30) http://cran.r-project.org/web/package= lmerTest
Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3): 1126-1177.
Levy, R. \& Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In Schöllkopf, B., Platt, J., and Hoffman, T., editors, Advances in Neural Information Processing Systems, pp. 849-856, Cambridge, MA. MIT Press.
Mahowald, K., Fedorenko, E., Piantadosi, S. T., \& Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. Cognition, 126(2): 313-318.
R Core Team (2015). R: A language and environment for statistical computing. $R$ Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org
Rayner, K., Aschby, J., Pollatsek, A., \& Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the ez reader model. Journal of Experimental Psychology: Human Perception and Performance, 30(12).
Rayner, K. \& Well, A. (1996). Effects of contextual constraint on eye movements in reading: A further examination. Psychonomic Bulletin \& Review, 3(4): 504509.

Schneider, W., Eschman, A., \& Zuccolotto, A. (2002). EPrime: User's guide. Psychology Software Incorporated.
Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. In Interspeech (Vol. 2002, p. 2002).
Witzel, N., Witzel, J., \& Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. Journal of Psycholinguistic Research, 41(2), 105-128.

# Bridging Visual Working Memory Research from Infancy through Adulthood with Dynamic Neural Field Modeling 

Vanessa R. Simmering (simmering@ wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street, Madison, WI 53706 USA<br>Sammy Perone (sammy.perone@wsu.edu)<br>Department of Human Development, Johnson Tower 516, Pullman, WA 99163 USA


#### Abstract

Theories that span tasks and developmental periods require explaining how a single cognitive system can flexibly adapt across contexts yet show stable age-related improvement. Here we present a computational model that embodies a unified theory of visuospatial cognitive development. We use this model to bridge between previously disconnected domains, as diverse as infant habituation and visual working memory capacity in adults. We illustrate how the same real-time and developmental processes can account for behavior across tasks and age groups. We conclude with a discussion of the implications of a unified theory for understanding cognition and development more broadly, with an eye toward early intervention.


Keywords: visual cognition; working memory; development; infancy; neural field model

## A Unified Theory of Visuospatial Cognition

A central challenge in cognitive science is to create theories that generalize across tasks and developmental periods. Computational models provide a concrete tool to confront this challenge. We illustrate this using a dynamic neural field (DNF) model of visuospatial cognition. Our goal is to explicitly connect the processes that operate across the range of behavioral tasks used from infancy through adulthood to measure different aspects of visuospatial cognition. By demonstrating that such disparate tasks and phenomena arise through common underlying processes, we can construct a broader theory to contrast with prior theories that have been proposed to account for only a single task and age group.

The unified theory of visuospatial working memory development we espouse is implemented in a single model architecture with a single developmental mechanism to explain change from infancy to adulthood. Here we show how this model can bridge between previously disconnected domains and developmental periods. We begin by describing the basic dynamics of DNFs. Next, we present the specific architecture we have applied across tasks and age groups. We then illustrate how this model can be used to account for visual and spatial working memory processes across tasks and development. Finally, we close by discussing the implications of our unified theory for understanding cognitive development and interventions that strengthen cognition in at-risk or developmentally delayed populations.

## Dynamic Neural Fields

DNFs belong to a larger class of bi-stable attractor networks
(Amari, 1977; Wilson \& Cowan, 1972) and simulate neural population dynamics to represent a continuous dimension, such as space or color (Schöner, Spencer, \& the DFT Research Group, 2015). DNFs have a functional topographic organization such that neighboring nodes within a field representing similar features (e.g., shades of blue in color, neighboring locations in space). In DNFs, a stimulus input excites selectively-tuned neurons which then interact through local excitatory and lateral inhibitory connections to create a localized "peak" of activation (illustrated in Fig. 1 below).

A peak in a DNF represents a real-time neuronal estimate of the stimulus. With relatively weak local excitation and lateral inhibition, peaks are only present when supported by input, that is, when the stimulus is present in the environment - we refer to this as an encoding state. With stronger connectivity, peaks can persist after a stimulus disappears (i.e., input is removed), which we refer to as a working memory state. Peaks leave excitatory memory traces, a simple Hebbian-type history of activation, that facilitate the re-formation of peaks at similar values (e.g., color, location) at future points in time. For example, when presented with a blue stimulus, the model will produce a peak that estimates the specific hue. The peak will leave a memory trace that facilitates the formation of a peak for the color blue at a future point in time. We will show that this feature of DNFs has implications for behavior in working memory tasks.

Multiple DNFs can be coupled together to create more complex neural architectures that simulate neurocognitive processes of encoding, maintenance, comparison, and recognition (described further below). To use such models to understand behavior, these neural architectures can be coupled to behavioral systems to generate the particular behavioral dynamics of interest - below we describe systems to simulate looking behavior, same/different judgments, or pointing/recall responses. In the next section, we describe a three-layer architecture that we have used to simulate performance in visuospatial working memory tasks from infancy to adulthood.

## A Three-Layer Dynamic Neural Field Model

Figure 1 shows the three-layer model (reviewed by Johnson \& Simmering, 2015; referred to here as the "dynamic model" for simplicity) used for the simulations we describe. This instantiation of the model consists of a fixation and visualcognitive system. The fixation system consists of a collection of nodes that represent looking to left, right, center, and away
locations in a virtual world. The nodes compete in a winner-take-all fashion. The winner (left node in Fig.1) opens a perceptual gate and the stimulus at that location (green and red) is input to the visual-cognitive system (see green arrow from virtual world to visual-cognitive system). The visualcognitive system consists of a perceptual field (PF), which receives input from the fixation system representing the color of the stimulus. This input creates peaks representing the stimulus; connectivity in this field is set to the encoding state described above. Activation in PF supports continued fixation through reciprocal connectivity (see green arrow from PF to virtual world) and also feeds into a working memory (WM) field (see green arrow from PF to WM). These fields interact through a shared field of inhibitory nodes (Inhib). When WM activity for an item is strong, WM sends strong activity to Inhib (see red arrow from WM to Inhib). This, in turn, suppresses activity for that item in PF (see red arrow from Inhib to WM). In addition to this threelayer (PF-Inhib-WM) architecture, the dynamic model includes memory trace (MT) layers associated with PF and


Figure 1. Three-layer dynamic neural field architecture, coupled to a fixation system viewing colors in a virtual world. Green versus red arrows indicate excitatory versus inhibitory connections; horizontal dashed lines indicate the zero threshold in each field. See text for further description.

WM that accumulate activation over a longer time scale (for simplicity, only $\mathrm{MT}_{\mathrm{WM}}$ is shown in Fig.1), serving the Hebbian function described above.

Figure 1 illustrates how the dynamic model simulates encoding, maintenance, and comparison of items in WM. A critical insight gained from applying this model across tasks and development is how the same real-time processes underlying these cognitive functions can produce a range of seemingly unrelated behavioral signatures (e.g., habituation, perseveration, novelty preferences, capacity limits, dimensional attention), as described below. At the moment represented in Figure 1, WM is maintaining the colors light green and orange, which has inhibited encoding in PF (see inhibitory troughs at sites tuned to light green and orange), which released fixation from the right location (see position in virtual world). This inhibition of encoding by WM is the mechanism of recognition in the model. After fixation was released from the right location in the simulation shown in Figure 1, the model fixated the left location and is encoding a dark green and red stimulus there. This is the mechanism of novelty detection in the model - fixating and encoding items not held in WM. Notice that the model has MTs associated with the light green and orange items. This will enable the model to form robust WM peaks for those colors in the future, which can support recognition of those items as familiar.

This simulation shows the dynamic model equipped with a fixation system that looks at multiple locations, which simulates looking behavior (Perone, Simmering, \& Spencer, 2011). To simulate the behaviors required by different visuospatial working memory tasks, the model can be used to generate continuous recall responses (e.g., pointing to a remembered location or color) based on peak positions (e.g., Spencer, Smith, \& Thelen, 2001) or equipped with a same/different response system (Johnson, Spencer, Luck, \& Schöner, 2009). Critically, however, each of these different behavioral responses is driven by the same underlying cognitive processes embodied in the three-layer architecture.

In the following sections, we synthesize recent applications of the dynamic model to provide a unified explanation of visuospatial cognitive processes across previously disconnected domains and development: habituation and visual recognition during infancy, and VWM capacity limits from infancy to early childhood and adulthood.

## Common Processes Underlying Visual Working Memory from Infancy to Adulthood

In this section, we describe how this model can account for behavior and development in three domains, highlighting that a single developmental mechanism produces change in all three domains. We begin by describing how the model links infant looking at a single location to WM formation in the habituation paradigm. Next, we show how the same model looks to multiple location in a visual recognition context, the visual paired comparison task. After that, we show how the same model can once again be adapted to explain visual working memory capacity limits in children and adults.

## Habituation

Infant looking paradigms form the foundation of our understanding of the origins of human cognition. Habituation of looking behavior has been used for half a century to probe perceptual, memory, and cognitive processes during infancy. In a typical habituation task, infants are presented with a single stimulus (e.g., blue circle). Initially, they exhibit high levels of looking which decreases with repeated presentation. They typically renew looking when presented with a novel stimulus (e.g., red star). Prior theories have not considered looking behavior a central part of the learning process (Cohen, 1972) but rather as an output. However, there is evidence that how infants distribute their looking through time structures what they learn (Jankowski, Rose, \& Feldman, 2001). To explore the interplay between looking and learning, Perone and Spencer (2013) used the dynamic model with a simple fixation system that stochastically oscillated between looking at a single stimulus and looking away. As described above, when the fixation node was looking to the stimulus, it opened a perceptual gate that sent input PF. Strong activation in PF supported continued looking and led to the formation of memory representations in WM and MT. When the WM representation grew robust across presentations, encoding in PF was inhibited (as described above), and looking was released. Thus, the dynamic model showed habituation in looking time, just as infants do.

Perone and Spencer (2013) tested whether the dynamic model could account for the developmental changes infants show in habituation tasks: faster habituation rates and the ability to make finer-grained distinctions with age. To simulate development in the visual-cognitive system of the dynamic model, they implemented the Spatial Precision Hypothesis (SPH). The SPH posits that the strength of excitatory and inhibitory connectivity within and between layers increases over development (see Simmering \& Schutte, 2015, for review). Implementing the SPH in the context of the habituation task led to faster, more stable WM formation and more robust novelty detection. This led to quick habituation and improved discrimination with age, just as infants show. The dynamic model's performance highlighted the link between visual exploration and learning. For example, spontaneous long looks helped WM form, which led to fast habituation. Conversely, spontaneous short looks led to slowed memory formation. This provides an explanation for how individual differences in visual exploration can structure learning. This highlights the importance of simulating real-time behavior in a model to understand how the cognitive system functions and develops.

## Visual Paired Comparison

The visual paired comparison (VPC) paradigm is commonly used to study visual recognition and categorization processes during infancy. VPC differs from the habituation paradigm in a critical way: it introduces competition. Infants are presented with pairs of stimuli and can freely look back and forth between them. This context yields a rich set of looking
measures, including shift rate (gaze switches per second of looking), look duration (mean duration of each look), peak look (longest look), and preferences (proportion of looking to one item greater than chance). Infants' recognition memory is assessed via pairing a previously seen, familiar item with a novel item. A preference for the novel item is evidence of both (1) recognition of the familiar item and (2) discrimination between the familiar and novel items (as illustrated in Fig.1). With age, infants exhibit faster shift rates, shorter look durations, shorter peak looks, more finegrained discrimination, and stronger novelty preferences. These looking behaviors develop more slowly in at-risk populations, such as preterm infants (e.g., Rose, Feldman, \& Jankowski, 2001).

The dynamic model can adapted to VPC by equipping it with a fixation system that looks at left and right locations (see Fig.1), compared to the single item/fixation location used for habituation. The dynamics of the visual-cognitive system are otherwise identical to the model simulations of habituation from Perone and Spencer (2013). Perone and Spencer (2014) asked whether this same model and developmental mechanism could account for the range of behavioral changes infants show over development in VPC. They probed this by testing infants' looking behavior and discrimination abilities between 5 and 10 months of age, then simulating the paradigm in the dynamic model. They found that infants exhibited faster shift rates, shorter look durations, and shorter peak looks with age. They also found that infants were able to make discriminations along a continuous metrically organized dimension by 7 months of age. The model exhibited the same behavioral pattern over development for precisely the same reasons as it did in the habituation paradigm: faster, more robust memory formation.

Perone and Spencer (2014) also analyzed individual differences. In particular, individual differences in looking during the learning phase of VPC predicted their discrimination abilities during the testing phase. This pattern was found in the dynamic model's performance as well. But where did these individual differences come from? There were no parameter changes to simulate "individuals" in the model; rather, the individual differences in patterns of performance were emergent. The structure of looking behavior that builds memory representations and supports discrimination in the dynamic model emerged autonomously. This parallels the insight gained from the simulations of habituation: infants' exploratory behavior in the task influenced the formation of memory, which in turn shaped their subsequent looking behavior. Although the processes at work in the habituation and VPC are generally considered similar, what infants remember in each paradigm is different (Oakes \& Ribar, 2005). This is the first theory to formally account for how the same learning process unfolds in both contexts.

## Capacity Limits over Development

One of the hallmarks of WM is its limited capacity. Visual working memory (VWM ) in particular is limited to only three or four items in adults (Luck, 2008). The majority of
work characterizing VWM capacity limits have focused on children and adults, with the change detection task being a common approach, shown in Figure 2A. In this task, a small number of simple items (e.g., colored squares) is shown briefly, followed by a brief blank delay, then a test array in which either all of the items remained the same or one has changed. Capacity estimates from this task (using a formula proposed by Pashler, 1988) have shown a gradual increase from early childhood through adolescence (Simmering, 2016; Simmering \& Perone, 2013). Studies with infants, however, present seemingly contradictory results, with estimates of capacity reaching adult-like levels within the first year of life (e.g., Oakes, Ross-Sheehy, \& Luck, 2006; Ross-Sheehy, Oakes, \& Luck, 2003). One way to address this apparent discrepancy across tasks and age groups is through the dynamic model framework presented here.


Figure 2. Sample trials from two tasks used to estimate VWM capacity: (A) change detection used with children and adults and (B) change-preference used with infants.

The task used to estimate capacity during infancy, shown in Figure 2B, is a variant of VPC called the change preference paradigm. Infants are presented with the same number of colored squares is presented on each of two displays. The squares briefly appear and disappear repeatedly throughout each trial; across these presentations, the colors in the "nochange" display remain the same; on the other "change" display, one color changes following each blank delay. Infants' fixation is tabulated over the course of the trial, and compared to chance (i.e., equal looking to both displays). Capacity is estimated from the highest set size (i.e., number of colors per display) at which infants show a reliable preference for the "change" display. The rationale behind this interpretation is that if infants can remember the colors within a display, the "change" display will appear novel and therefore support a looking preference. Ross-Sheehy et al. (2003) estimated capacity to be only one item at 6 months, but three to four items at 10 months. Oakes et al. (2006) then showed that the capacity increase from one to three items occurred between 6 and 7 months of age.

How can the change preference task yield a VWM capacity of 3-4 items at 10 months but the change preference task only yield a capacity of 1-2 items at 3 years? Perone et al. (2011) situated the dynamic model in the change preference
paradigm and showed that the SPH could account for the agerelated changes in capacity estimates during infancy. One intriguing finding from these simulations was that a robust preference in the model did not depend on holding all of the items in memory: that is, a preference for set size three did not necessarily reflect that three items could be held in WM. Perhaps this means that the items required to be remembered to yield different estimates differs across tasks. Simmering (2016) probed this possibility by situating the dynamic model in both the change preference and change detection task. In order for the dynamic model to simulate performance in the change detection task, it must be equipped to give the "same" or "different" responses required by the task. This type of response system can be implemented by building from the mechanisms of recognition and familiarity inherent in the model's visual-cognitive system (see Fig.1): peaks in WM indicate familiar items whereas peaks in PF indicate novel items. Thus, a simple system in which activation from WM projects to a "same" decision node, and activation from PF projects to a "different" decision node, can use these signals to general a discrete response on each trial (see Johnson \& Simmering, 2015, for further discussion).

Model simulations revealed that the two tasks used to estimate capacity showed different relationships between the underlying memory representations and the behavioral measures used to estimate it (Simmering, 2016). In particular, while simulations of the infant task suggested that behavioral estimates may over-estimate the number of items held in memory (Perone et al., 2011), simulations of adults' performance in change detection indicated it under-estimated the number of items held in memory (Johnson, Simmering, \& Buss, 2014). Simmering (2016) bridged these results from infancy and adulthood by testing young children in both types of capacity tasks, then directly comparing performance across tasks and simulating results within a unified model. Simulations showed that developmental changes in both tasks could be accounted for within the same model through strengthening connectivity. Furthermore, although the tasks yielded different estimates of capacity between 3 and 5 years of age - at least six items in the looking task versus only two to three items in the change detection task - the common underlying processes were evident in correlations across tasks. Motivated by the common processes that support the detection of novelty across the two tasks in the model, Simmering (2016) found that children's preference scores in the looking task were significantly correlated with their hit rates (i.e., proportion correct on change trials) in change detection. This relationship across tasks was not evident from considering only the capacity estimates from each task, but rather depended on a systematic understanding of how cognition and behavior relate.

## Spatial Cognition and Development

The preceding sections showcased the use of the same model and developmental mechanism to adapt across contexts and development for visual (featural) memory processes. In this section, we illustrate that the same model
can be adapted to account for performance in four spatial cognition tasks- (1) perseverative reaching in the Piagetian A-not-B task, (2) A-not-B-type biases in a sandbox task, (3) reference-related biases in spatial recall and (4) similar biases in position discrimination. Moreover, we show that developmental change across all of these domains was explained with the SPH.

## A-not-B Tasks

Beginning with the A-not-B task (cf. Thelen, Schöner, Scheier, \& Smith, 2001), weaker connectivity to simulate early infancy (8-10 months) led to perseverative reaching because the peak representing the second (B) location was not strong enough to overcome the history of reaches at the first location (A). With stronger connectivity to capture development, this peak could be maintained accurately through the delay to support accurate reaching as seen in older infants (10-12 months).A similar interaction between memory for the current target and prior reaches can be seen in older children's performance in a sandbox version of the A-not-B task (e.g., Schutte, Spencer, \& Schöner, 2003; Spencer, Smith, \& Thelen, 2001). Children between the ages of 2 and 6 years show recall responses that are biased toward previously-remembered locations, with a developmental change in the spatial spread of this influence - younger children's performance is biased over larger separations than older children's (Schutte et al., 2003). This metric change in the influence of reaching history in the task has been simulated first through changes in the spread of activation within the three-layer architecture (Schutte et al., 2003) and later through changes in only the strength of connectivity (Simmering et al., 2008; see Simmering \& Schutte, 2015, for further discussion). By showing that the SPH can account for changes in infants' performance in the canonical A-not-B task as well as the metric changes during early childhood, this model architecture brought together previously disconnected age groups and tasks into a unified framework.

## Spatial Recall \& Discrimination

Young children's performance in spatial recall tasks also shows influences of the spatial structure of the space, in addition to prior history of reaches. Specifically, young children recall locations as closer to the midline symmetry axis of the task space, whereas older children and adults recall locations further from midline and the edges of the task space (see Spencer, Simmering, Schutte, \& Schöner, 2007, for review). These effects can be simulated in the same three-layer architecture through the coordination between perceptual and memory processes in the sandbox or "spaceship" tasks. As connectivity strengthens over development, the representation of information in both perception and memory lead to the transition in bias (Schutte \& Spencer, 2009, 2010). These processes operate continuously through time, and can be detected even in the brief delay of position discrimination tasks, linking together previously disconnected areas of research (Simmering \& Spencer, 2008).

## Implications and Future Directions

We presented a unified theory of working memory development that spans an impressive number of domains and periods of development. Importantly, this was only possible by using a concrete tool - a computational model to tackle the difficult challenge of explaining cognition across domains and development. There are a number of implications of this work. First, our theory indicates that cognitive processes are not domain specific. Instead, the same general visual-cognitive system can account for multiple neurocognitive processes by organizing itself differently in different contexts with different behavioral demands. These include infant habituation (looking), visual working memory capacity (same/different judgments), and spatial recall (position estimation). Second, our theory indicates that the developmental mechanisms that drive change across domains are not unique. We showed that the SPH could account for changes in performance across multiple domains and radically different periods of development. Last, our account raises the intriguing possibility that we can target basic visual-cognitive processes to strengthen early in development, which may have an impact across many domains and over a long period of time.

One long-term goal of employing such a computational framework is to make further connections across age groups and domains, and to provide a mechanistic account of how behavior emerges in specific task contexts. Such examples can already be found in the domains of executive function (e.g., Buss \& Spencer, 2014; Perone, Molitar, Buss, Spencer, \& Samuelson, 2015) and word learning (e.g., Samuelson, Schutte, \& Horst, 2009; Samuelson, Smith, Perry, \& Spencer, 2011). By connecting the same real-time processes of encoding, maintaining, and comparing visual inputs with the longer time-scale of learning in contexts that connect to verbal labels, we can test how far relatively simple cognitive mechanisms can go toward explaining complex behaviors (cf. Smith, Jones, \& Landau, 1996).

## Acknowledgments

Both authors contributed equally to this work. Thanks to Aaron T. Buss, Christian Faubel, Jeffrey S. Johnson, John Lipinski, Yulia Sandamirskaya, Sebastian Schneegans, Gregor Schöner, Anne R. Schutte, and John P. Spencer for helpful discussion and support during the development of this line of research.

## References

Amari, S. (1977). Dynamics of pattern formation in lateralinhibition type neural fields. Biological Cybernetics, 27, 77-87.
Buss, A. T., \& Spencer, J. P. (2014). The emergent executive: A dynamic field theory of the development of executive function. Monographs of the Society for Research in Child Development, 79, 1-103.
Cohen, L. B. (1972). A two process model of infant visual
attention. Paper presented at the Merrill Palmer Conference on Research and Teaching of Infancy Development.
Jankowski, J. J., Rose, S. A., \& Feldman, J. F. (2001). Modifying the distribution of attention in infants. Child Development, 72, 339-351.
Johnson, J. S., \& Simmering, V. R. (2015). Integrating perception and working memory in a three-layer dynamic field architecture. In G. Schöner, J. P. Spencer, \& the DFT Research Group (Eds.), Dynamic thinking: A primer on dynamic field theory. New York, NY: Oxford University Press.
Johnson, J. S., Simmering, V. R., \& Buss, A. T. (2014). Beyond slots and resources: Grounding cognitive concepts in neural dynamics. Attention, Perception, \& Psychophysics, 76, 1630-1654.
Johnson, J. S., Spencer, J. P., Luck, S. J., \& Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. Psychological Science, 20, 568-577.
Luck, S. J. (2008). Visual short-term memory. In Visual Memory. New York: Oxford University Press.
Oakes, L. M., \& Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. Infancy, 7, 85-98.
Oakes, L. M., Ross-Sheehy, S., \& Luck, S. J. (2006). Rapid development of feature binding in visual short-term memory. Psychological Science, 17, 781-787.
Pashler, H. (1988). Familiarity and visual change detection. Perception and Psychophysics, 44, 369-378.
Perone, S., Molitar, S., Buss, A. T., Spencer, J. P., \& Samuelson, L. K. (2015). Enhancing the executive functions of 3-year-old children performing the dimensional change card sort task. Child Development, 86(3), 812-827.
Perone, S., Simmering, V. R., \& Spencer, J. P. (2011). Stronger neural dynamics capture changes in infants' visual working memory capacity over development. Developmental Science, 14, 1379-1392.
Perone, S., \& Spencer, J. P. (2013). Autonomy in action: Linking the act of looking to memory formation in infancy via dynamic neural fields. Cognitive Science, 37, 1-60.
Perone, S., \& Spencer, J. P. (2014). The co-development of looking dynamics and discrimination performance. Developmental Psychology, 50, 837-852.
Rose, S. A., Feldman, J. F., \& Jankowski, J. J. (2001). Attention and recognition memory in the 1st year of life: A longitudinal study of preterm and full-term infants. Developmental Psychology, 37, 135-151.
Ross-Sheehy, S., Oakes, L. M., \& Luck, S. J. (2003). The development of visual short-term memory capacity in infants. Child Development, 74, 1807-1822.
Samuelson, L. K., Schutte, A. R., \& Horst, J. S. (2009). The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization. Cognition, 110, 322-345.
Samuelson, L. K., Smith, L. B., Perry, L. K., \& Spencer, J. P. (2011). Grounding word learning in space. PLoS One, 6(12), e28095.

Schöner, G., Spencer, J. P., \& the DFT Research Group. (2015). Dynamic thinking: A primer on dynamic field theory. New York, NY: Oxford University Press.
Schutte, A. R., \& Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: capturing a qualitative developmental transition in spatial working memory. Journal of Experimental Psychology: Human Perception and Performance, 35, 1698-1725.
Schutte, A. R., \& Spencer, J. P. (2010). Filling the gap on developmental change: Tests of a dynamic field theory of spatial cognition. Journal of Cognition and Development, 11, 328-355.
Schutte, A. R., Spencer, J. P., \& Schöner, G. (2003). Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. Child Development, 74, 1393-1417.
Simmering, V. R. (2016). Working memory capacity in context: Modeling dynamic processes of behavior, memory, and development. Monographs of the Society for Research in Child Development, 81, 7-148.
Simmering, V. R., \& Perone, S. (2013). Working memory capacity as a dynamic process. Frontiers in Developmental Psychology, 3, 567.
Simmering, V. R., \& Schutte, A. R. (2015). Developmental dynamics: The spatial precision hypothesis. In G. Schöner, J. P. Spencer, \& the DFT Research Group (Eds.), Dynamic thinking: A primer on dynamic field theory. New York, NY: Oxford University Press.
Simmering, V. R., Schutte, A. R., \& Spencer, J. P. (2008). Generalizing the dynamic field theory of spatial cognition across real and developmental time scales. Brain Research, 1202, 68-86.
Simmering, V. R., \& Spencer, J. P. (2008). Generality with specificity: The dynamic field theory generalizes across tasks and time scales. Developmental Science, 11, 541555.

Smith, L. B., Jones, S. S., \& Landau, B. (1996). Naming in young children: A dumb attentional mechanism?. Cognition, 60(2), 143-171.
Spencer, J. P., Simmering, V. R., Schutte, A. R., \& Schöner, G. (2007). What does theoretical neuroscience have to offer the study of behavioral development? Insights from a dynamic field theory of spatial cognition. In J. M. Plumert \& J. P. Spencer (Eds.), The emerging spatial mind. New York, NY: Oxford University Press.
Spencer, J. P., Smith, L. B., \& Thelen, E. (2001). Tests of a dynamic systems account of the A-not-B error: The influence of prior experience on the spatial memory abilities of 2-year-olds. Child Development, 72, 1327-1346.
Thelen, E., Schöner, G., Scheier, C., \& Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. Behavioral \& Brain Sciences, 24, 1-86.
Wilson, H. R., \& Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. Biophysical Journal, 12, 1-24.

# Pupil Dilation and Cognitive Reflection as Predictors of Performance on the Iowa Gambling Task 

Boban Simonovic (b.simonovic@derby.ac.uk)<br>Centre for Psychological Research, University of Derby<br>Kedleston Road, Derby, DE22 1GB<br>Edward J. N. Stupple (e.j.n.stupple@derby.ac.uk)<br>Centre for Psychological Research, University of Derby<br>Kedleston Road, Derby, DE22 1GB<br>Maggie Gale (m.gale@derby.ac.uk)<br>Centre for Psychological Research, University of Derby<br>Kedleston Road, Derby, DE22 1GB<br>David Sheffield (d.sheffield@derby.ac.uk)<br>Centre for Psychological Research, University of Derby<br>Kedleston Road, Derby, DE22 1GB


#### Abstract

Risky decisions involve cognitive and emotional factors. As the primary test for the Somatic Marker Hypothesis (SMH), the Iowa Gambling Task (IGT) examines these factors. Skin conductance shows anticipatory physiological responses on the IGT supporting SMH. Pupil dilation offers an alternative physiological marker. Predictive effects of anticipatory pupillary responses to positive and negative decks on IGT performance were examined in an extended IGT. The extended Cognitive Reflection Test (CRT) examined the relationship between reflective thinking and IGT performance. Data demonstrated correlations between reflective thinking and performance from the second block onwards and that task learning continued into the additional blocks - performance was not optimized even in the final block. Regression analysis showed both anticipatory pupil dilation for disadvantageous and advantageous decks, and reflective thinking were strong predictors of IGT performance. While both emotional and reflective processes are implicated in IGT performance, analytic cognition is more important than traditionally acknowledged.


Keywords: Pupil dilation; Iowa Gambling Task; Cognitive Refection; Somatic Marker Hypothesis; Dual-process Theory.

## Introduction

Learning and decision making in uncertain situations is an important activity, and it can be challenging to find an optimal decision even for simple choices. Decisions can be driven by the desire to maximize expected utility (Quartz, 2009), but information management regarding reward utility is frequently uncertain. Cognitive and emotional influences on risky decision making were traditionally regarded as separate in nature, with emotional factors typically seen as a hindrance. However, more recent evidence indicates that there is an interplay between the two, such that cognitive functions may serve as moderators for emotion-based learning (e.g. Brevers, Bechara, Cleeremans, \& Noel, 2013; Simonovic, Stupple, Gale \& Sheffield, 2016).

Damasio (1994) developed Somatic Marker Hypothesis (SMH) arguing that emotional processes play a central role in risky decision-making. SMH postulates that decisions are guided by subjective 'gut feelings' (e.g. bodily representations) about the inherent goodness or badness of future choices. These somatic markers direct individuals towards alternatives that have been positive previously or guide them away from the negative options. Particularly in uncertain conditions, response options are marked with an emotional signal, and only those options that are marked as favorable are cognitively processed (Damasio, 1994; Bechara \& Damasio 2005). Somatic markers operate covertly, indicate arousal anticipation and are regarded as physiological markers of emotion-based learning (Bechara \& Damasio, 2005; Critchley et al., 2001).
A further theoretical framework for investigating risky decision making is Dual Process which proposes that there are two types of cognitive process: unconscious, emotional gut-feelings (Type 1) that contrast with explicit, effortful, analytic processes (Type 2) (e.g. Kahneman. 2003). This proposal has been linked with SMH; for example Type 1 processes include a range of intuitive processes such as emotional responses or gut feelings (Glockner \& Witteman 2010) that can be measured through physiological techniques. There is also evidence of a role for cool reflective processing (Brevers, Bechara, Cleeremans, \& Noel, 2013; Simonovic et al., 2016) which maps onto Type 2 processing.
The primary paradigm in evaluation of emotion-based learning is the Iowa Gambling Task (IGT, Bechara, Damasio, Damasio, \& Anderson, 1994). The IGT offers a means of testing decision preferences and performance and has become an important experimental tool in evaluation of emotion-based learning and decision making. It has been argued that IGT resembles real life decision making as it involves uncertainty and monitoring of rewards and punishments (Bechara \& Damasio, 2005). Participants are
required to choose cards from four decks ( $\mathrm{A}, \mathrm{B}, \mathrm{C}$, and D ), all of which differ in frequencies of financial rewards and punishments. Advantageous decks (C and D) offer moderate rewards and small punishments whereas disadvantageous decks (A and B) offer larger rewards but substantial penalties, which result in an overall loss.
During the IGT, participants need to learn from experience about the 'goodness' or 'badness' of decks based on the feedback of learned contingencies. Thus, while participants experience deck reward properties they also assign affective values to the decks which implicitly influence decision making. The standard IGT consists of five blocks of 20 trials and healthy participants are considered to reach ceiling performance in the final block as the disadvantageous selections have been extinguished. Optimal IGT performance rests therefore on monitoring emotional responses and impulse inhibition related to the rewards and punishments (e.g., Bechara \& Damasio, 2005).
An important finding for SMH is that anticipatory somatic markers of emotions occur before decisions are made, indicating that covert anticipatory emotions can guide decision making (e.g. Bechara, Damasio, Tranel \& Damasio, 1997). Indeed, there is evidence demonstrating anticipatory Skin Conductance Responses (aSCR) to rewards and punishments under uncertain conditions (e.g. Bechara et al., 1997; Wagar \& Dixon 2006). Furthermore, interpretation of these aSCR's highlight the primary role of emotions in guiding decision making performance (e.g. Bechara \& Damasio, 2005).

In contrast, there is also evidence that reflective evaluation of affective choices guides future decisionmaking and occurs relatively early in the decision-making processes (e.g. Bowman, Evans, \& Turnbull, 2005 Brevers et al., 2013; Simonovic et al., 2016). This evidence is consistent with an interplay between Type 1 and Type 2 processes in determining the outcome of the decision making process (cf. Kahneman, 2003). Indeed, Brevers et al. (2013) argued that anticipation of long-term consequences in uncertain condition rely on two neural systems: a 'cool' and a 'hot' systems. The 'hot' system is impulsive, laden with affective 'gut feelings' akin to intuition, while the 'cool' system is reflective and includes analytic aspects. Learning and optimal decisions depend on the integration of both systems whereby, a 'cool' reflective process can be critical in monitoring or inhibiting 'hot' processes. It has also been argued that cool reflective processes should not play a role until the deck contingencies become explicit and so there should not be a role for Type 2 processing in the early blocks. Simonovic et al., (2016), however, demonstrated that reflective processes play a role earlier than previously predicted (cf. Schiebener, Zamarian, Delazer and Brand 2011).
Evidence suggest that aSCRs represent a good example of anticipatory somatic markers or 'hot processing'. They are, however, imperfectly represented because the SCR is not sufficiently sensitive in discriminating between negative and positive valence (Dunn et al., 2006). Faster measures of
emotion feedback (e.g. heart rate and blood pressure with an electrocardiogram or pupil dilation using eye trackers) are warranted (e.g. Bradley, Codispoti, Cuthbert \& Lang, 2001). Indeed, studies that use faster physiological measurement, (e.g. eye-tracking methodology) can better capture surprised responses to unexpected stimuli (e.g. Lavin, San Martin, \& Jubal, 2014).
Recent studies have also shown pupil dilation can measure surprise such as when feedback does not meet expectation (Preuschoff, Hart, \& Einhauser, 2011), when negative feedback occurs during the gambling task (Satterthwaite et al., 2007), and as evidence of learning (Lavin et al., 2014). Moreover, there is evidence linking pupillary responses to Locus Coeruleus (LC) - norepinephrine (NE) activity in the brain stem in anticipation of a reward, suggesting memory enhancement (Tully \& Bolshakov, 2010), and consolidation of behavioural decisions (Bouret \& Sara, 2005). Some evidence indicates greater pupillary responses before selecting negative options (e.g. Bierman, 2004), or after experiencing unexpected losses (Satterthwaite et al., 2007), thus indicating that anticipatory pupillary responses can be related to negative outcomes. However, Lavin et al. (2014) argue that pupillary responses are associated with positive feedback. Thus, although anticipatory pupillary responses serve as affective physiological markers and may offer a measure of the somatic markers that moderate learning in uncertain conditions their interpretation is also not necessarily straightforward.

To our knowledge, only one study has utilised eyetracking methodology during the IGT performance in a healthy population. Lavin et al. (2014) tested IGT performance and measured pupil dilation in a sample of 10 participants and demonstrated changes in pupil dilation due to learned uncertainty. Their results suggest that changes in pupil dilation reflect learned uncertainty about future feedback conditions, thus indicating differential processing of unexpected feedback. However, a non-standard version of the IGT was used and did not differentiate between disadvantageous and advantageous deck selection. In the present study, we extend Lavin et al.'s (2014) findings, with a larger sample and an alternative approach to measuring anticipatory pupil dilation.

Our focus was on the period during the IGT where participants had hypothetically developed somatic markers but that these were not yet sufficient to extinguish particular card selection. On this basis we measured pupil dilation in the 500 ms prior to the final selection from each deck and hypothesized that there should be somatic markers indicating negative anticipation for disadvantageous decks and positive anticipation for advantageous decks. We measured anticipatory pupillary responses for the advantageous $(\mathrm{C}+\mathrm{D})$ and the disadvantageous $(\mathrm{A}+\mathrm{B})$ final options. If anticipatory somatic markers play a role in IGT performance then these should be evident prior to the final selection of each type of card.

Moreover, we included a direct measure of deliberative thinking to replicate previous findings demonstrating that
the CRT was highly predictive of IGT performance (Simonovic et al., 2016). We used the extended seven-item Cognitive Reflection Test (CRT ${ }^{1}$ ), developed to measure the ability to resist and override intuitive responses by engaging analytic ability (Toplak, West, \& Stanovich, 2014), this is a more comprehensive measure than the original three item CRT used in the Simonovic et al. (2016) study.

It was predicted that the CRT and last anticipatory pupillary responses for advantageous and disadvantageous deck picks would predict IGT performance. It was also predicted that the correlations observed by the Simonovic et al., (2016) between CRT score and disadvantageous card selections across blocks would be replicated (such that strong correlations would be found in blocks $2-4$, but no correlation would be observed in the early trials and the correlation would be reduced in the final blocks). Finally, the standard analysis of IGT performance across blocks was extended to test whether performance reached ceiling levels in the fifth block (the final block in the standard IGT) or whether performance continued to improve.

## Method

## Design

Predictor variables were: the seven-item CRT (Toplak et al., 2014) and pupillary responses averaged across the 500 ms prior to the final selection for both advantageous (C $+\mathrm{D})$ and disadvantageous $(\mathrm{A}+\mathrm{B})$ decks. The CRT was used as a measure of analytic thinking. The dependent variable was the IGT score. Performance across blocks was also examined for completeness with $(\mathrm{C}+\mathrm{D})-(\mathrm{A}+\mathrm{B})$ as the dependent variable for performance in each block

## Participants

Sixty-nine ${ }^{2}$, healthy students from the University of Derby, aged 19-29 years, received course credit for participation. Research was conducted in accordance with stipulations of the local ethics committee. Participants had normal or corrected to normal vision.

## Materials and Procedure

Participants completed Bechara et al.'s (1994) computerised version of $\mathrm{IGT}^{3}$. Scoring was derived by deducting 'good' card picks ( $\mathrm{C}+\mathrm{D}$ ) from total 'bad' picks $(\mathrm{A}+\mathrm{B})$. A positive score indicates a more optimal decisionmaking strategy.

[^495]The seven-item CRT (Toplak, West \& Stanovich, 2014) score was the total number of correct answers. Higher CRT scores indicated higher reflective ability. Cronbach's alpha was $\alpha=$. 66 .
Eye movements were recorded with the Eye-gaze binocular system Tobii-X2-30 (Inquisit 4 milliseconds plugins), with a remote binocular sampling rate of 30 Hz and an accuracy of about $0.45^{\circ}$. The X2 Eye Tracker is a stand-alone eye tracker, and it was attached to a laptop (Dell, Precision M6700, 2.70Ghz). Participants were seated approximately 70 cm from the laptop monitor. The Tobii measured 184 mm ( $7.2^{\prime}$ ') in length and enabled tracking at close distances (up to $36^{\circ}$ gaze angle). The eye-tracker used both bright and dark pupil illumination setups to calculate the optimal gaze position. Blinking periods were filtered and replaced via linear interpolation (e.g. Siegle, Steinhauer Carter, Ramel, \& Thase, 2003). The anticipatory pupil dilation (aPD) diameter was defined as the mean pupillary response generated 500 ms before card selection. A 500 ms time frame was identified a priori as a period where fixation occurs, and direction of the information search can be determined (e.g. Horstmann. Ahlgrimm, \& Glockner, 2009).

## Analytic Strategy

Initial analyses focused on participants' performance per block by using repeated measures ANOVA. Next, correlations between CRT scores and selection of disadvantageous cards for each block were calculated. Finally, regression analysis was used to examine the independent contributions of CRT scores and pupil dilations on IGT performance. Analysis was conducted using IBM SPSS 24 for Windows.

## Results

Performance across blocks was tested using a Greenhouse-Geisser adjusted repeated measures ANOVA. There was a main effect of Block condition, $\mathrm{F}(3.86,239.12)$ $=25.21, \mathrm{p}<.001, \eta_{\mathrm{p}}{ }^{2}=.29$. Bonferroni adjusted post hoc tests demonstrated that performance improved significantly through the blocks of trials (excluding Block 6). Notably the nonstandard additional blocks 6 and 7 continued to show changes in performance relative to earlier blocks such that performance dipped in Block 6 but Block 7 was significantly better than all but Block 5. Means and standard deviations are shown in Table 1.

Table 1: Mean (SD) IGT Performance as a function of Trial Block.

| Trial Block | IGT Performance |
| :--- | :--- |
|  |  |
| Block 1 | $-3.65(6.29)$ |
| Block 2 | $-.016(7.71)$ |
| Block 3 | $2.76(8.79)$ |
| Block 4 | $4.19(10.45)$ |
| Block 5 | $7.38(9.75)$ |
| Block 6 | $5.86(11.24)$ |
| Block 7 | $9.35(9.63)$ |
| Total | $26.27(46.70)$ |

Table 2: Correlations between Disadvantageous card selections and CRT score as a function of Trial Block

| Trial Block | Correlation |
| :--- | :--- |
| Block 1 | $\mathrm{r}=-.18, \mathrm{p}=.150$ |
| Block 2 | $\mathrm{r}=-.41, \mathrm{p}=.001$ |
| Block 3 | $\mathrm{r}=-.74, \mathrm{p}<.001$ |
| Block 4 | $\mathrm{r}=-.81, \mathrm{p}<.001$ |
| Block 5 | $\mathrm{r}=-.70, \mathrm{p}<.001$ |
| Block 6 | $\mathrm{r}=-.71, \mathrm{p}<.001$ |
| Block 7 | $\mathrm{r}=-.67, \mathrm{p}<.001$ |
| Total | $\mathrm{r}=-.89, \mathrm{p}<.001$ |

Correlations between CRT score and selection of disadvantageous cards across blocks were conducted (see Table 2). These demonstrated a significant negative relationship between CRT score and disadvantageous card selections in all but the first block of trials.
A multiple regression (Enter method) tested the relative predictive strength of last anticipatory pupillary responses for disadvantageous ( $\mathrm{A}+\mathrm{B}$ ) (mean, $\mathrm{SD}=3.02,0.36 \mathrm{~mm}$ ) and advantageous $(\mathrm{C}+\mathrm{D})$ (mean, $\mathrm{SD}=3.00,0.38 \mathrm{~mm}$ ) deck picks and CRT scores (mean, $\mathrm{SD}=2.13,1.76$ ) for performance on the IGT. Data indicated that the three predictors combined reliably accounted for $35 \%$ of the variability in IGT scores. The standardized beta for disadvantageous cards showed a negative correlation with pupil dilation while the advantageous cards showed a positive correlation. This indicated that increased pupil dilation on the last pick of a disadvantageous card predicted poorer overall performance in contrast with increased pupil dilation for advantageous cards which was associated with better overall performance. The CRT score was the strongest predictor with higher scores on the CRT predicting better card selections.

Table 3: Multiple Regression Analysis of CRT, Final Anticipatory Pupil Dilation for Disadvantageous (AB)

Decks, Last Pupil Dilation for Advantageous (CD) decks as predictors (standardized betas) of IGT performance

| Predictors |  |
| :--- | :--- |
| Model 'Enter.' | $\mathrm{R}^{2}=.38, \mathrm{R}^{2}{ }_{2 d j}=.35$ |
|  | $\mathrm{~F}(3,58)=12.03, \mathrm{p}=.001$ |

$\begin{array}{ll}\text { CRT scores } & \beta=.56, \mathrm{p}<.001 \\ \text { Last aPD }(\mathrm{A}+\mathrm{B}) & \beta=-.46, \mathrm{p}=.05 \\ \text { Last aPD }(\mathrm{C}+\mathrm{D}) & \beta=.52, \mathrm{p}=.03\end{array}$
Durbin Watson $=1.93$, VIF $=1.042 ; 4.965 ; 4.992$

## Discussion

Consistent with our predictions anticipatory pupillary responses and reflective thinking were reliable independent predictors of IGT performance. Importantly, pupillary responses differ according to the nature of the deck and incrementally predict performance in addition to cognitive reflection. Specifically, increased pupil dilation on the last pick of disadvantageous cards predicted poorer overall performance, whereas increased pupil dilation for the last pick of advantageous cards was associated with better overall performance. This is important because it indicates that differing somatic markers may develop for advantageous and disadvantageous decks and that these predict task performance alongside cognitive reflection.
Correlations between CRT scores and IGT broadly replicated the findings from the control group in Simonovic et al., (2016) but with stronger correlations and evidence that reflective processing is implicated even earlier in the task. Finally, block by performance analysis demonstrated that IGT performance did not reach ceiling at block 5 and significantly improved in block 7 after a (non-significant) dip in block 6, albeit it was not greater than block 5.

While our data indicate that participants' last aPD responses predict IGT performance, these somatic markers require some deciphering. Pupil dilation can be interpreted in various ways with anticipated threat, anticipated reward and general cognitive effort all potentially resulting in dilated pupils. Our data showing increased pupil dilation for advantageous deck is consistent with participants anticipating a positive outcome rather than a threat. However, it is possible that an increased level of cognitive effort may be in play (which would also be consistent with the correlations with cognitive reflection). Irrespective of the precise interpretation these data demonstrate a role for somatic markers in performance on the IGT, but allow for the possibility that these somatic markers are of cognitive effort as well as an indicator of emotional learning.

Pupillary responses data for the IGT contributes to the understanding of SMH and support a role for anticipatory physiological mechanisms in successful performance on risky decision-making tasks. These somatic markers inform explicit knowledge and facilitate learning of deck contingencies (e.g. Bechara \& Damasio, 2005). However,
our findings are incongruent with the proposition that IGT performance is primarily dependent on the development of somatic markers (Bechara et al., 1997; Wagar \& Dixon, 2006) and are instead compatible with the dual-process model where 'cool' reflective processes inhibit impulses that interfere with long-term goals. This is consistent with the proposition that integrating reflective and emotional processes is necessary to explain IGT performance and suggests that the ability to reflect on gut feelings about decisions may improve performance (Schiebener et al., 2011; Simonovic et al., 2016).

The CRT was shown to be a stronger predictor of IGT performance than the pupil dilation measures, with higher scorers clearly outperforming lower scorers. This is clear evidence that Type 2 reflective processing plays a salient role in the task and supports the view of Brevers et el. (2013) that the IGT is best understood within a dual process framework. Toplak et al.'s (2014) extended seven item version of the CRT was used and the more comprehensive nature of this measure along with the greater variability may explain the stronger correlations and greater proportion of variance explained than in Simonovic et al. (2016). The evidence from the correlations between CRT and performance across the blocks replicated findings from Simonovic et al. (2016). These data indicate a consistent role for analytic ability in determining IGT performance from the second block onwards. This is inconsistent with the view that the learning on the task is implicit until the contingencies are well established and is instead indicative of a role for explicit monitoring of deck contingencies even in the early blocks.

These CRT data nonetheless need to be interpreted with caution. There is debate as to whether the CRT is a measure of cognitive miserliness or a more general measure of analytic thinking or numerical ability as it has been correlated with both working memory (e.g., Stupple et al., 2013) and risk neutrality (Oechssler, Roider, \& Schmitz, 2009), which could impact on performance or task strategy.

Most of our participants began the task by exploring disadvantageous decks $(\mathrm{A}+\mathrm{B})$. Hence it could be argued that a reduced pupillary reaction for disadvantageous cards in relation to IGT performance occur because participants had 'unlearned' the initial preferences for big reward. This is consistent with a suggestion that during the IGT performance reversal learning, needs to be implemented to suppress learned preferences that are no longer beneficial (Dunn, Dalgleish, \& Lawrence, 2006). Lavin et al. (2014) suggest that successful performance on IGT depends on positive feedback (based on the money gain) and highlighted pupillary responses to unexpected punishments on positive decks. This proposal is consistent with as both are indicative of anticipatory effects, as both are indicative of anticipatory effects, however, the differing methodologies of the current study and Lavin et al.'s make direct comparisons difficult.

Since the SCR has a relatively slow time course, it is possible that a distinct somatic marker cannot be
distinguished by conventional SCR measurements (Newell \& Shanks, 2014). The use of an eye-tracker allows a distinction between somatic reactions on different options before a decision has been made. This is particularly important because the anticipatory SCR captured during the IGT performance may represent part of a broader response such as attentional bias, implicit learning and a risk-taking.

Steingroever et al. (2013), called for greater scrutiny of IGT performance in healthy populations to bolster the ecological validity of IGT and demonstrate that IGT scores measure real-life decision making. The validation of IGT performance in healthy population is of great importance for such a widely used clinical tool; our analyses add to this literature on healthy populations. The prominent role of cognitive reflection in IGT performance leads us to urge caution in its application in diagnosing emotional deficits in populations who may lack the working memory capacity to perform well on the CRT and, by implication the IGT.

The measures used in the present study are relatively narrow and further applications of the pupil dilation methodologies are necessary to more fully explore the utility of this measure in investigating the IGT and the SMH more broadly; in particular, extending analysis across the task to examine how pupillary responses relate to IGT performance curves may be illuminating. Moreover, alternative eye-tracking measures such as fixations on particular decks of cards offer strong potential in investigating the locus of explicit attention as learning progresses on the task.

In conclusion, our data demonstrated that a combination of anticipatory pupil dilation and reflective thinking predicted IGT performance, such that both emotional and reflective processes are implicated in IGT performance. That is, anticipatory pupil dilation may serve as learning markers particularly for individuals with higher levels of cognitive reflection. Analytic cognition, moreover, plays a more salient role than traditionally acknowledged.

## Acknowledgments

This research is part of Boban Simonovic's PhD studies undertaken at the University of Derby.

## References

Bagneux V., Font H., Bollon T. (2013). Incidental emotions associated with uncertainty appraisals impair decisions in the Iowa Gambling Task. Motivation and Emotion, 37, 818-827.
Bechara, A., \& Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. Games and Economic Behaviour, 52, 336-372.
Bechara, A., Damasio, A. R., Damasio, H., \& Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. Cognition, 50,7-15.
Bechara, A., Damasio, H., Tranel, D., \& Damasio, A. (1997). Deciding advantageously before knowing the advantageous strategy. Science, 275, 1293-1996.

Bierman, D. (2004). The use of an implicit grammar task and eye measurements to study the somatic marker hypothesis. Retrieved on the $26^{\text {th }}$ of January 2016 from: http://uniamsterdam.nl/D.J.Bierman/PUBS/2004/bial_210_final.pdf
Bouret, S., \& Sara, S. (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. Trends in Neuroscience 28, 574-582.
Bowman, C. H., Evans, C., \&Turnbull, O. H. (2005). Artificial time constraints on the Iowa Gambling Task:
The effects on behavioral performance and subjective experience. Brain and Cognition, 57, 21-5.
Brevers, D., Bechara, A., Cleeremans, A., \& Noel, X. (2013). Iowa Gambling Task (IGT): Twenty years after-Gambling d disorder and IGT. Frontiers in Psychology, 4, 665.
Critchley, H.D., Mathias, C. J., Dolan, R.J., 2001. Neural activity in the human brain relating to uncertainty and arousal during anticipation. Neuron, 29, 537-545.
Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. New York: Putnam Publishing
Dunn, B. D., Dalgleish, T., \& Lawrence, A. D. (2006). The somatic marker hypothesis: a critical evaluation. Neuroscience and Biobehavioral Reviews, 30, 239-271.
Einhäuser, W., Koch, C., \& Carter, O. L. (2010). Pupil dilation betrays the timing of decisions. Frontiers in Human Neuroscience, 4, 18.
Elliot, R., Friston, K. J., \& Dolan, R. J. (2000). Dissociable neural responses in human reward systems. Journal of Neuroscience, 20, 6159-6165.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19,25-42.
Glöckner, A. \& Witteman, C. (2010) Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. Thinking \& Reasoning, 16,1-25.
Horstmann, N., Ahlgrimm, A. \& Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. Judgment and Decision Making, 4, 335-354.
Jepma, M., \& Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. Journal of Cognitive Neuroscience, 23, 1587-1596.
Kahneman, D. (2003). A perspective on judgment and choice. Mapping bounded rationality. American Psychologist, 58, 697-720.
Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X. \& Milham, M. P. (2006). Distinct neural mechanisms of risk and ambiguity: a meta-analysis of decision-making. Neuroimage, 32 (1), 477-484.

Lavin, C., San Martin, R., \& Jubal, E. R. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. Frontiers in Behavioral Neuroscience, 7, 218.

Montague, P. R., Hyman, S. E., Cohen, J. D. (2004). Computational roles for dopamine in behavioral control. Nature, 411, 760-767.

Newell, B. R., \& Shanks, D. R., (2014). Unconscious influences on decision making: A critical review. Behavioral and Brain Sciences, 37, 1-19.
Oechssler, J., Roider, A., \& Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. Journal of Economic Behavior \& Organization, 72(1), 147-152.
Preuschoff, K., T Hart, B. M., \& Einhauser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. Frontier Neuroscience, 5, 115.
Quartz, S. R. (2009). Reason, emotion and decisionmaking: risk and reward computation with feeling. Trends in Cognitive Sciences, 13 (5), 209-215.
Satterthwaite, T.D., Green, L., Myerson, J., Parker, J. Eamaratham, M., \& Buckner, R. L. (2007). Dissociable but inter-realted systems of cognitive control and reward during decision making: evidence for pupillometry and event- related fMRI. Neuroimage, 37, 1017-1031.
Schiebener, J., Zamarian, L., Delazer, M., \& Brand, M.
(2011). Executive functions, categorization of probabilities, and learning from feedback: What does really matter for decision making un- der explicit risk conditions? Journal of Clinical and Experimental Neuropsychology, 33, 1025-1039.
Siegle, G. J., Steinhauer, S. R., Carter, C. S., Ramel, W., \& Thase, M. E. (2003). Do the seconds turn into hours? Relationships between sustained pupil dilation in response to emotional information and self-reported rumination. Cognitive Therapy and Research, 27, 365-382.
Simonovic, B., Stupple, E. J. N., Gale, M., Sheffield, D. (2016). Stress and risky decision making: cognitive reflection, emotional learning or both. Journal of Behavioral Decision Making. DOI: 10.1002/bdm. 1980
Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., \& Wagenmakers, E. J. (2013). Performance of healthy participants on the Iowa Gambling Task. Psychological Assessment, 25, 180-193.
Stupple, E. J. N., Gale, M., \& Richmond, C. (2013). Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test. In M. Knauff, M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Cognitive Science Society: Austin, TX: Cogntive Science Society.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2014). Assessing miserly information processing: an expansion of the Cognitive Reflection Test. Thinking and Reasoning 20, 147-168.
Tully, K., \& Bolshakov, V. Y. (2010). Emotional enhancement of memory: how norepinephrine enables synaptic plasticity. Molecular brain, 3, 15.
Wagar, B., \& Dixon, M. (2006). Affective guidance in the Iowa Gambling Task. Cognitive, Affective, Behavioral Neuroscience, 6, 277-290.

# Talking Through Your Arse: Sensing Conversation with Seat Covers 

Sophie Skach (s.skach@qmul.ac.uk)<br>School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS<br>Patrick G. T. Healey (p.healey@qmul.ac.uk)<br>School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS<br>\section*{Rebecca Stewart (rebecca.stewart@qmul.ac.uk)}<br>School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS


#### Abstract

People move in characteristic ways during conversation and these movements correlate with their level of particpation. For example, speakers normally gesture significantly more than listeners. These visible, overt movements are normally analysed using full body video or motion capture. Here we explore the potential of a 'minimal' approach to sensing these participatory movements in part of the natural environment of everyday interactions; chair seat covers. Using custom built fabric sensors we test whether we can detect people's involvement in a conversation using only pressure changes on the seats they are sitting in. We show that even from this impoverished data we can distinguish between talking, backchanneling and laughter; each state is associated with distinctive patterns of pressure change across the surface of the chair. We speculate on the possible applications of this new, unintrusive form of social sensing for architecture, performance and augmented human interaction.


Keywords: human interaction; dialogue; non-verbal communication; social sensing; smart textiles; posture analysis; fabric sensors;

## Introduction

People make a variety of distinctive body movements during conversation. The most commonly studied of these are the gestures that speakers produce while talking. These include gestures that contribute to the content of what is said, such as iconics, metaphorics and pantomimes (McNeill, 1992; de Ruiter, 2000), as well as gestures that help to orchestrate the interaction such as beat gestures and gestures that can hold or hand over the turn to someone else (Bavelas, Chovil, Lawrie, \& Wade, 1992; Healey \& Battersby, 2009). Listener's body movements are also organised in characteristic ways. Most obviously through the production of concurrent feedback or 'backchannels' (Yngve, 1970). Although these are often produced as non-interruptive verbal acknowledgements such as a brief "aha" or "mmhm" people also frequently backchannel by nodding in response to an ongoing turn. Listeners are also distinguished from speakers by their relative lack of hand movement although they move their hands much more when a speaker requests clarification or makes repairs to their turn (Healey, Plant, Howes, \& Lavelle, 2015).

The significance of this non-verbal choreography is illustrated by how much we can infer about an interaction from
the observation of body movements alone. We can often tell just by looking at who is talking to whom, who -if anyone- is listening, who is likely to speak next, whether the interaction is hostile or friendly and so on (Kendon, 1990). These inferences from non-verbal performances can be striking; people appear to be able to make reliable estimates of the quality of someone's teaching over a whole semester from a single 5 second video of body movements alone (Ambady \& Rosenthal, 1992).

Research on non-verbal communication has tended to focus on these relatively large scale overt body movements; they are the easiest signals for participants to perceive and respond to and the most tractable for analysis (although see e.g. Ekman \& Friesen 1969). Research typically takes advantage of video and, more recently, motion capture equipment to capture and analyse these movements (e.g. Healey \& Battersby 2009; Vinciarelli et al. 2009. The rapid development of new sensor technologies and their application to social signal processing has opened an intriguing new space of possibilities for detecting patterns of interaction (Vinciarelli, Pantic, \& Bourlard, 2009). For example, it is possible to detect people's levels of interest, stress and intoxication in conversation using the speech signal alone i.e. without knowing anything about the content of what is said (B. W. Schuller \& Rigoll, 2009; B. Schuller et al., 2013). In contrast to relatively intrusive technologies such as video or automatic speech recognition, this approach makes it possible to create anonymised 'minimal' forms of social sensing by using textile technology (see also e.g. Rekimoto 2001).

Here we explore the potential of this approach for one of the most commonly used parts of the physical environment for social interaction; chairs. Even the shape and position of unoccupied, uninstrumented chairs can indicate a great deal about interaction; chairs around a small table suggest something very different from chairs in rows (see also Anderson 1996). Moreover, chair covers are often made out of stretch and soft fabric that makes the textile surface itself a potentially promising sensing material. Using metallic yarn gives a fabric conductive properties so that it can be turned into a
pressure sensitive surface. Different possibilities of using textile surfaces as sensing materials or interfaces for electronic devices have been explored in recent years, for example by turning a jacket into an interface to a mobile phone (Poupyrev et al., 2016) or by measuring biomechanical data for healthcare applications (e.g. Pacelli et al. 2006). Here we apply a similar fabrication technique to chair covers to address the basic question of whether it is possible to detect patterns of conversational interaction from movements on the chair surface alone.

## Sensing Chairs

Informal observation suggests that people frequently change the position of the torso, lower body, and feet during seated conversations. These movements necessarily cause pressure changes on the surface of the chair and are therefore potentially detectable by measuring changes in resistance. Previous work has investigated the use of chairs to classify postures through pressure sensors, creating pressure maps of both, static and dynamic postures - posture identification versus continuous tracking (Tan, Slivovsky, \& Pentland, 2001; Slivovsky \& Tan, 2000). A commercially available pressure measurement system, BPMS (Body Pressure Measurement System) by Tekscan ${ }^{1}$ has been used in some of these research projects (e.g. D'Mello et al. 2007 and Arnrich et al. 2010), which consists of a plastic mat with 64 integrated pressure sensors that allow for the creation of detailed pressure maps. The main applications for these sensing systems have been in the analysis of posture to improve seating comfort (e.g. Milivojevich et al. 2000), designs for objects involved in rehabilitation (e.g., wheelchairs) and Human-ComputerInteraction. For example, presenting chairs as novel haptic interfaces for computer games (Tan, Slivovsky, \& Pentland, 2001), or as a system to measure people's cognitive states in various situations Arnrich et al. (2010), including measuring a car driver's fatigue (Furugori, Yoshizawa, Iname, \& Miura, 2003) or identifying drivers (Riener \& Ferscha, 2008), as well as measuring boredom in students (D'Mello, Chipman, \& Graesser, 2007). However, this approach has not previously been applied to sensing aspects of social interaction. With this study, we explore what information about social behaviour can be retrieved from pressure sensor data on a chair.

## Methods

Drawing on informal observations of people's leg and torso movements in meetings we decided on a configuration of eight sensors that were integrated in the chair cover and distributed in a symmetric arrangement; four in the seat of the chair and four on the back (see Figure 1), dividing the chair into four key areas to be sensed in order to determine postures: shoulders (at the top of the back rest), waist (lower back), buttocks and thighs. These observations also laid the basis upon which initial hypotheses about different states in a conversation were built.

[^496]

Figure 1: Reverse side of the chair cover showing sensors.

## Sensor Development

The textile sensors were made from conductive fabric and resistive foam, hand sewn into soft sensor patches that were manually attached to the backside of a chair cover (which was made of jersey knit fabric). The conductive fabric, SaniSilver, was purchased from LessEMF and woven with a silver yarn showing on one side of the fabric and a cotton yarn visible on the other. The sensors are constructed such that two swatches of conductive fabric are facing the resistive foam on both sides. When pressure is applied to the conductive fabric on either side of the foam, the foam compresses and reduces the resistance between the two fabric swatches. This change in resistance is measured by the microcontroller. The sensors have the advantage of behaving like an ordinary fabric that could also be used in other wearable applications, such as garments (since, through the use of cotton fibre, the fabric retained a soft touch and remained comfortable to wear).

## Data Collection

A microcontroller (a Teensy 3.2) collected the pressure data from the sensors and stored it on micro SD cards. The sampling frequency of the sensors was 4 readings per second. Using these piezoresistive sensors, the unit of measure is Ohm $(\Omega)$. Since the aim was to investigate postural behaviour in a situation of social interaction, three chair covers were manufactured, each housing one micro controller that were placed underneath the chair. Wires were hooked into the conductive fabric and connected the sensors with the Teensy (to ground and to an assigned analog pin providing 3.3 volts to run the programme, which read analog output values from the sensors).

## Participants

Participants were recruited in groups of three friends or colleagues to ensure they all had some initial level of familiarity
with each other. We conducted 9 trials in total, collecting data from 27 participants, of which were 11 female and 16 male and between the age of 20 and 40.

## Procedure

The experiment was carried out in the Human Interaction Lab at Queen Mary University of London. Groups of three participants were asked to resolve a moral dilemma: the balloon task. This is a fictional scenario describing three people in a hot air balloon that is about to crash, if not one of the passengers jumps to their certain death. The task is then to come to an agreement of who to throw off. The participants were told that the aim of the experiment is to investigate collaborative interaction. They were seated at a round table and asked to discuss options and come to an agreement on how to resolve this dilemma. We aimed to record 15 to 20 minutes of conversation, so if not having come to an agreement after this time, participants were given the option to stop the conversation or carry on (vice versa, if they came to an agreement faster, alternative scenarios were provided to encourage further discussion). Due to the materiality of the sensors, the presence of the sensor patches was not noticed by the participants, so that the experience wasn't different to sitting on a common chair.

## Data Analysis

The interactions were captured on two cameras placed in different corners in the room. Lapel microphones were used to facilitate speaker-specific analysis of the audio for transcription. The data from the video recordings was annotated using Elan (Brugman, Russel, \& Nijmegen, 2004). Coding focused on three key behaviours with: speaking, laughter and backchannels. When determining speaking modes, periods of overt speech were coded, regardless of postural and gestural changes, or nodding. But focusing on postural movement overall, it was noticed that often, a postural or gestural change was performed immediately prior to speaking. This makes the start of an utterance ambiguous. For the purposes of this study, the beginning of utterances was defined as the onset of speaking. For laughter, responsive as well as speakers concurrent laughter was noted. Therefore, laughter is annotated for both, speakers and listeners. Backchannels were coded for all continuous verbal particles of response, as well as repair initiations. An overview of the coding scheme for these behavioural cues can be seen in Table 1.

Table 1: Coding scheme used in Elan.

| Tiers per participant | Social behaviour |
| :--- | :--- |
| speaking | verbal utterance |
| laughter | responsive and concurrent |
| backchannel | responsive, repair initiation |

Following this coding scheme, all elements that mark listening modes are created from the gaps of the annotations
for laughter, speaking and backchannels. This means that within the listening mode, any gross and subtle body movement, as well as nodding or any other conversational action is included. With the aim of distinguishing speakers from listeners, this level of detail in annotations is sufficient, although the sensitivity of the sensors allows for richer and more finegrained distinctions.

## Results

The data from all eight sensors were analysed in a General Linear Model Multivariate Regression using SPSS v.24. Talking, Laughing and Backchanneling were included as binary predictors coded as 1 or 0 for presence / absence of each behaviour. All two and three-way interactions of these three factors were included in the model. Participants were also included as a main effect to ensure individual variation was accounted for.
Since the relative changes for each participants were calculated, changes of weight had no effect on the outcome of the analysis.


Figure 2: Estimated means of all participants for TALK: thighs: left(187.513), $\operatorname{right(209.379);~butt:~left(137.721),~}$ right(175.910); waist: left(231.599), right(345.421); shoulders: left(137.810), right(195.288)

Multivariate Tests (Pillai's Trace) show all three dialogue factors reliably predict the outputs of the pressure sensors (Talk: $F_{(8,82933)}=9.68, p<0.00$; Backchannel $F_{(8,82933)}=$ $10.2, p<0.00$; Laugh: $F_{(8,82933)}=6.95, p<0.00 ;$ ). The effects are very small with Partical Eta Squared of 0.001 and observed power for Alpha $=0.05$ of 1 . The contribution of individual variation is, by contrast, much larger: Participant: $F_{(8,82933)}=6.95, p<0.00$, Partial Eta Squared $=0.71$ ).

Analysis of the contributions of each sensor show that different patterns of pressure changes across the chair are associated with the different dialogue states. The sensors most sensitive to talking were in the seat of the chair and correspond


Figure 3: Estimated means of all participants for LAUGHTER: thighs: left(118.642), $\operatorname{right}(209.379)$; butt: left(178.079), right(187.614); waist: left(229.081), $\operatorname{right}(342.121)$; shoulders: left(143.385), right(179.532)
to increased pressure from the thighs and reduced pressure from the buttocks. In contrast to this laughter corresponded to reduced pressure in the thighs and increased pressure in the buttocks with no significant changes detected in the seat back. The pattern of pressure changes for the relatively brief backchannels were distributed across both the seat and back of the chair and corresponded to increased pressure across thighs, buttocks and waist but a reduction across the shoulders. The estimated means for changes at each sensor are illustrated in Figures 2, 3 and 4 (numbers based on modified population marginal mean).

## Discussion

The results show that it is possible, in principle, to detect significant aspects of social interaction from quite limited, indirect and noisy data. The small movements detected by pressure sensors embedded in chair seats are small-scale and almost completely invisible correlates of the gross body movements that typically distinguish speakers from hearers and laughter from silence. Interestingly, even the relatively small nodding movements of the head associated with backchannels appear to create a distinguishable pressure signature on a chair.

This is the first attempt to detect significant conversational states from simple 'homemade' pressure sensors and the signal to noise ratio is low. Individual variations in movement in particular account for far more of the variance than differences in dialogue state. Further work to optimise the size and position of the sensors would doubtless improve the quality of the sensing. It is also likely that other approaches, such as training person-specific classifiers and machine learning mechanisms, would improve the accuracy and robustness of


Figure 4: Estimated means of all participants for BACKCHANNELS: thighs: left(176.345), right(199.648); butt: left(172.195), right(189.949); waist: left(246.298), $\operatorname{right}(351.819)$; shoulders: left(114.193), right(189.709)
the approach although this would also undermine the advantages of anonymity. The demonstration that even relatively crude sensors can detect minimal changes in posture, suggests that future work should explore the possibility of capturing more complex social behaviour, especially relational questions such as whether interactions are, for example: convivial or combative; autocratic or egalitarian, or whether it is possible to characterise regularities in multiparty interaction (see e.g. (Abney, Paxton, Dale, \& Kello, 2014)).

What could this form of sensing be used to do? The principle opportunities for application are in any situations where there is value in the ability to unintrusively gather information about general patterns of social interaction including levels of interest and engagement. One example is architecture where the ability to sense a building's energy performance and patterns of air flow is highly valued but currently has no social counterpart. We speculate that the ability to make simple, systematic assessments of a building's 'social performance' by instrumenting the chairs in a building could also have a significant positive impact on domestic and workplace design. A second example is in the evaluation of audience responses (e.g. continuous audience response measure, CARM, which is used by broadcast hosts to evaluate their programs). The deployment of such a sensor network in an auditorium, meeting room or a classroom could help to assess levels of engagement of students and other audiences. In addition, there are possibly applications to augmented human interaction where, for example, live feedback about how much people are dominating (or not) a conversation can have significant effects on the conduct of the interaction (Donath, 2002). If nothing else these results shed some light on Stephen Fry's (1984) advice that when delivering Shakespeare one should
"always gather from the buttocks".

## Summary

This paper presents a new sensing system using textile pressure sensors that are designed to be integrated in a chair cover and that are able to reliably distinguish speakers from listeners and detect laughter and backchannels. These fabric sensors provide a non-intrusive way to measure conversational engagement. Data about pressure changes on the seat and back rest alone make it possible to differentiate various behavioural states in a seated conversation. The ability to extract such patterns of social interaction from sensing pressure changes could replace other, more complex motion detection systems and mitigate privacy concerns, since the data collection is anonymous involves no audio or video data and does not capture any of the content of the conversation.

## Acknowledgments

We thank the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology, who funded this research. In particular, we also thank Adan Benito Temprano for his support, as well as the Digital Catapult Centre.

## References

Abney, D. H., Paxton, A., Dale, R., \& Kello, C. T. (2014). Complexity matching in dyadic conversation. Journal of Experimental Psychology: General, 143(6), 2304.
Ambady, N., \& Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. American Psychological Association.
Anderson, R. J. (1996). A security policy model for clinical information systems. In Security and privacy, 1996. proceedings., 1996 ieee symposium on (pp. 30-43).
Arnrich, B., Setz, C., La Marca, R., Tröster, G., \& Ehlert, U. (2010). What does your chair know about your stress level? IEEE Transactions on Information Technology in Biomedicine, 14(2), 207-214.
Bavelas, J. B., Chovil, N., Lawrie, D. A., \& Wade, A. (1992). Interactive gestures. Discourse processes, 15(4), 469-489.
Brugman, H., Russel, A., \& Nijmegen, X. (2004). Annotating multi-media/multi-modal resources with elan. In Lrec.
de Ruiter, J. P. (2000). 14 the production of gesture and speech. Language and gesture, 2, 284.
D'Mello, S. S., Chipman, P., \& Graesser, A. (2007). Posture as a predictor of learner's affective engagement. In Proceedings of the cognitive science society (Vol. 29).
Donath, J. (2002). A semantic approach to visualizing online conversations. Communications of the ACM, 45(4), 4549.

Ekman, P., \& Friesen, W. V. (1969). Nonverbal leakage and clues to deception. Psychiatry, 32(1), 88-106.
Furugori, S., Yoshizawa, N., Iname, C., \& Miura, Y. (2003). Measurement of driver's fatigue based on driver's postural change. In Sice 2003 annual conference (Vol. 1, pp. 264269).

Healey, P. G., \& Battersby, S. A. (2009). The interactional geometry of a three-way conversation. In Proceedings of the 31st annual conference of the cognitive science society (pp. 785-790).
Healey, P. G., Plant, N., Howes, C., \& Lavelle, M. (2015). When words fail: collaborative gestures during clarification dialogues. In 2015 aaai spring symposium series: Turn-taking and coordination in human-machine interaction.
Kendon, A. (1990). Spatial organization in social encounters: The f-formation system. Conducting interaction: Patterns of behavior in focused encounters, 209-238.
McNeill, D. (1992). Hand and mind: What gestures reveal about thought. University of Chicago press.
Milivojevich, A., Stanciu, R., Russ, A., Blair, G., \& Van Heumen, J. (2000). Investigating psychometric and body pressure distribution responses to automotive seating comfort (Tech. Rep.). SAE Technical Paper.
Pacelli, M., Loriga, G., Taccini, N., \& Paradiso, R. (2006). Sensing fabrics for monitoring physiological and biomechanical variables: E-textile solutions. Proceedings of the 3rd IEEE-EMBS International Summer School and Symposium on Medical Devices and Biosensors, ISSS-MDBS 2006, 1-4. doi: 10.1109/ISSMDBS.2006.360082
Poupyrev, I., Gong, N.-W., Fukuhara, S., Karagozler, M. E., Schwesig, C., \& Robinson, K. E. (2016). Project jacquard: Interactive digital textiles at scale. In Proceedings of the 2016 chi conference on human factors in computing systems (pp. 4216-4227).
Rekimoto, J. (2001). Gesturewrist and Gesturepad: Unobtrusive Wearable Interaction Devices. ISWC '01 Proceedings of the 5th IEEE International Symposium on Wearable Computers, 21-27.
Riener, A., \& Ferscha, A. (2008). Supporting implicit human-to-vehicle interaction: Driver identification from sitting postures. In The first annual international symposium on vehicular computing systems (isvcs 2008) (p. 10).
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., \& Narayanan, S. (2013). Paralinguistics in speech and languagestate-of-the-art and the challenge. Computer Speech \& Language, 27(1), 4-39.
Schuller, B. W., \& Rigoll, G. (2009). Recognising interest in conversational speech-comparing bag of frames and suprasegmental features. In Interspeech (pp. 1999-2002).
Slivovsky, L. A., \& Tan, H. Z. (2000). A real-time sitting posture tracking system.
Tan, H. Z., Slivovsky, L. A., \& Pentland, A. (2001). A sensing chair using pressure distribution sensors. IEEE/ASME Transactions On Mechatronics, 6(3), 261-268.
Vinciarelli, A., Pantic, M., \& Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. Image and vision computing, 27(12), 1743-1759.
Yngve, V. H. (1970). On getting a word in edgewise. In Chicago linguistics society, 6th meeting (pp. 567-578).

# Individual Differences in Spontaneous Analogical Problem-Solving: The Reflective Mind Account 

Slavi Todorov Slavov (s.t.slavov5@gmail.com)<br>Department of Cognitive Science and Psychology, Montevideo Str. 21, Sofia, Bulgaria<br>Penka Hristova (phristova@cogs.nbu.bg)<br>Department of Cognitive Science and Psychology, Montevideo Str. 21, Sofia, Bulgaria


#### Abstract

Analogical problem-solving involves transfer of knowledge that has been obtained from a source analog and successfully applying it in the solution of a structurally similar target problem. What is usually found in the so-called hint/no-hint paradigm is that spontaneous solution to a problem is hard to achieve. This leaves the possibility for individual differences. This study searched for and found a positive correlation to exist between scores on the Cognitive Reflection Test and spontaneously solved analogical problems which, although a weak one, possibly accounts for the differences that exist between people who need a hint to solve an analogical problem, and people that do not need a hint.


Keywords: Analogy; Analogical problem-solving; Reflective Mind thinking; Cognitive Reflection Test

## Introduction

Imagine you are presented with a problem - an oil well is on fire and is consuming large amounts of petrol every minute. You know you have enough foam to put out the fire, but if you use the one large hose that is available to shoot it at the well, the fire can be extinguished, but the pressure would also destroy the machines around the well that facilitate the oil extraction, which would be an expensive cost. If you use one of the several smaller hoses that are also available, the machines will be spared, but the fire would not be extinguished. How can this problem be resolved? Now, imagine that without any external hint to relate the problem to anything, you recall a story about an exhibition designer, who has to figure out a way to illuminate a replica of a ship, that is positioned in the center of transparent tank filled with water and fish that are sensitive to light. If the designer illuminates the replica with a powerful spotlight, the fish will be disturbed, but if she uses a low-powered spotlight, the ship would not be illuminated enough. So she decides to use several low-powered spotlights to illuminate the replica from several directions, which will not disturb the fish, but the focused light would be enough to illuminate the ship. In fact, these two superficially dissimilar problems are analogous - the solution to the fire problem is to shoot the foam using many small hoses from different directions so as to spare the machines, but also to provide enough foam to extinguish the fire. How many people would spontaneously think of using the solution of the problem they know to solve the analogous one? Probably not many, given that the stories appear to be different on the surface. The successful
solver would probably need to be able to reflect on what he is processing, to suppress the irrelevant information, and set his priorities in accordance to the task at hand.
The paradigm that is used in studying analogical problemsolving requires a relevant analog known to the solver to be available, as well as the target problem that is presented to be sufficiently novel and challenging in order for the analogy to be useful (Gick \& Holyoak, 1983). The framework that is generally required for the solver to represent the analogical relationships involves first of all a story describing the problem and how it is solved to be read and understood. Once the information is represented, it can be used to generate solution to the target problem by mapping the similar relations of the two systems, employing a top-down reasoning, forming expectations, and finally using the mapping in order to generate the solution to the target problem (Gick \& Holyoak, 1980).

## The "retrieval gap" in analogical problem-solving

In order to analogously solve the problem, participants must retrieve the correct analogical relationships. The role of retrieval is usually investigated in the so-called hint/no-hint paradigm (Novick \& Holyoak, 1991). By giving a hint to the solver in one of the two experimental conditions, they are informed that the two stories are connected and the solution to one of them can be used in solving the other. If they are not given a hint in the no-hint condition, solving the target problem would indicate spontaneous analogical transfer. What is usually found is that about $75 \%$ of the people solve the Radiation problem ${ }^{1}$ using the correct analogical solution when the appropriate analog story ${ }^{2}$ had

[^497]been previously presented and they are given explicit hint to use that story for the solution (Gick \& Holyoak, 1983). However, if no base story analog is presented, only less than $10 \%$ of the participants manage to find the correct solution. What Gick and Holyoak (1983) have found when their participants read the Attack-Dispersion story as a base, disguised to be remembered for a subsequent recall, is that $30 \%$ of them arrived at the correct solution of the Radiation problem presented subsequently, without receiving any hint, i.e. spontaneously. This apparent difference in the difficulty of mapping and retrieving the correspondences of an analog are referred to as "retrieval gap" (Holyoak, 2012), and can be considered in terms of at least 3 explanations:

Structural and surface similarity Problem-solving using analogs is very much dependent on the level of structural and surface similarity between the two stories in terms of the level of facilitation of retrieval (Blanchette \& Dunbar, 2000; Holyoak \& Koh, 1987). More specifically, if the superficial features of the base story are more similar to the ones of the target problem, spontaneous retrieval of convergence solution to the Radiation problem is as high as $90 \%$, compared to about $20 \%$ if the surface features were dissimilar (Blanchette \& Dunbar, 2000). It is suggested that because an analog that is from a remote domain, it does not share many of the salient surface features of the target, which might block the spontaneous retrieval of relevant analogs, unless the solver is able to focus on aspects that are causally related for the target (Holyoak \& Koh, 1987).

The experimental paradigm Blanchette and Dunbar (2000) show in their experiments the importance of the experimental setting in which the participants reason analogically. In the so-called "reception paradigm", the participants are given base and target problems and are required to identify the relations between them. As Blanchette and Dunbar's (2000) experiment shows, this type of setting constraints the participants and prompts them to make more analogies based on superficial similarity. On the other hand, an experimental setting organized in a "production paradigm" involves participants being given the target problem and being asked to generate possible source stories, arguably resulting in analogy generation based on deep structural features. Two of the experiments in Blanchette and Dunbar's study (2000) involved analogical reasoning using production paradigm. The results clearly indicated production of more analogies that were structurally similar. Their third experiment used arguments from the previous two experiments as stimuli, but the task was arranged in a reception paradigm. The results showed domination of retrieval of superficially similar stories. The findings are explained in terms of different type of encoding in the different types of tasks. In "reception paradigm" tasks, the initial presentation of the problem is usually guised as a comprehension evaluation or measuring recall,

[^498]which arguably causes the encoding to be more superficial. Furthermore, the base representation building may not necessarily include the relevant relations for the subsequent analogical problem solving. The "production paradigm", on the other hand, involves the participants in deeper structural encoding of the problem from the beginning, possibly resulting in more structurally similar analogies.

Possibility for individual differences The previously mentioned source Radiation problem, when learned in a different context, enables problem solvers to spontaneously produce the correct analogous solution to the superficially similar Lightbulb problem for $81 \%$ of the participants even several days after the presentation of the base problem (Holyoak \& Koh, 1987). The results for the spontaneously solved problems are discussed in terms of the possible demand characteristics of the task, or in other words that the participants might suspect the two stories to somehow be related due to them being present in the same experiment (Gick and Holyoak, 1983). This might suggest the possibility of individual differences to be present, specifically that some people might be sensitive to events occurring in the same context and interpret them as connected. Day and Goldstone (2011) discuss the possibility of individual differences in intelligence or the level of engagement in the experiment to be responsible, at least to some extent, for the difference between the transfer and the reported understanding of the analogy itself. Another possibility for individual differences in spontaneous analogical problem solving can be drawn from the so called Reflective Mind (Stanovich, 2012). According to the Tripartite model (Stanovich, 2012), the Reflective Mind is able to initiate the suppression of the initial response, due to its higher cognitive level control, that is carried out by the Algorithmic mind ${ }^{3}$. The Reflective mind is tested in the socalled typical performance situations, in which participants solve tasks without overt instructions to maximize their success. Spontaneous problem solving resembles a typical performance situation, since participants in the no-hint condition are not explicitly instructed to find and use the analogy with previous problems. Moreover, the mechanisms of Reflective mind such as cognitive decoupling operation, allows a suppression of the initial response that is provided by the Autonomous mind and creating a secondary representation of the world that could be manipulated until the correct solution is reached and then applied in reality. Just like Day and Goldstone (2011) have argued that some individual differences due to intelligence (i.e. Algorithmic Mind) may explain the superior problem solving performance of some individuals, we argue that differences regarding the Reflective Mind can also be expected in analogical problem solving. Indeed, spontaneous analogies are especially interesting case for individual differences

[^499]stemming from the Reflective mind. On the one hand, the ability to create a secondary representation of the problem that may hold and manipulate the base and target problems seems to guarantee a successful analogical problem solving. On the other hand, Holyoak (2012) has argued that the difficulties which people experience in spontaneous analogical problem solving in particular indicate that analogical mapping requires a Type 2 processing.

## Correlation between Reflective mind thinking and spontaneous analogical problem solving

The main goal of this study is exploratory - given that not much is known about individual differences in spontaneous problem solving, especially with regard to the Reflective mind, the aim would be to find a correlation between these two variables. More specifically, possible individual differences might be expected in the no-hint condition, where spontaneous analogy-making depends on the correct identification of the structural similarities in the two problems, as well as the appropriate mapping, which might be reasonable to expect from people with higher rational dispositions who are arguably better at prioritizing goals and performing well without overt instructions what exactly is expected of them.

## Method

## Design

This is a correlational study, aiming to research whether a positive correlation exists between scores on the Cognitive Reflection Test and the analogical problems that are solved spontaneously. For the purposes of the research, a reception paradigm was used. The research has been approved by the ethical commission at the New Bulgarian University.

## Stimuli

Analogical stories The stimuli for the analogical problemsolving task consisted of six problems: three bases and three targets. The problems were selected so that they can be structurally identical, but superficially dissimilar.

- Red Adair \& Aquarium problems

The first set of analogous stories consisted of the Red Adair problem (Kurtz \& Loewenstein, 2007) and the Aquarium problem (Catrambone \& Holyoak, 1989). The former described a problem, in which an oil well that is burning has to be extinguished. If a big hose is used to shoot the foam into the well, the machines in the well that facilitate petrol extraction will be destroyed, even though the fire will be put out. But if one of the many smaller hoses is used, the machines will be spared, but the fire will not be extinguished. For the Aquarium problem, a replica of a ship had to be illuminated for an exhibition, without disturbing the fish swimming around it, which were sensitive to light. If one powerful spotlight was used, the fish would be disturbed and the replica illuminated, and if one less powerful spotlight was used - the fish would not be
disturbed, but the replica would not be illuminated. The solution for both problems involved "convergence of forces", or using small amounts of force from different directions (small hoses to put out the fire and low-powered spotlights to illuminate the ship). The Red Adair problem was modified so as to obtain full structural similarity with the Aquarium, by making the using of large force from one direction causing damage to peripheral elements (machines for petrol extraction in Red Adair and the fish in Aquarium).

- Garden and Marching band problems

The second set of stories were the Garden problem and the Marching band problem (Novick \& Holyoak, 1991). These were mathematical problems, involving finding how many plants a family can have in their garden, given that they had chosen the exact number of plants, which could be divided into 10,4 , and 5 kinds of plants, but there would be space for 2 more plants. Only when they divide them in 6 , they fit in without remainder. The Marching band described musicians marching in rows of 12,8 , and 3 , but having one musician march alone. Only when they march in rows of 5, there is nobody left out. The successful solution procedure for both problems is to find the lowest common multiple of the given three divisors that leave a constant remainder, then to generate multiples of that number, add the remainder to each of them, and finally find from this set the number that is divisible to the fourth number without a remainder.

- Orange and Tribe problems

The third pair of stories consisted of the story about the sisters, who were quarreling because each of them wanted one orange for herself. The problem was resolved when the mother found out that one of the sisters wanted to use the peel of one orange for baking, and the other wanted to eat the fruit, so each of them took the respective part of the whole orange (adapted from Fisher, Ury, \& Patton, 2011). An analog to this story was created, which was about two clans from the same tribe, that have recently captured an island, and each of the clans wants the whole island for themselves. So the chief of the tribe steps in and finds out that one of the clans wants the island for its territory, and the other one wants it because the people on the island pertain to their clan. The solution, then, is to divide the people from the territory, so that each side can be satisfied.
Cognitive Reflection Test The extended version of Cognitive Reflection Test (CRT) (Toplak, West \& Stanovich, 2014) was used as a measure of Reflective Mind. CRT was introduced by Frederick (2005) and measures cognitive reflection - a concept defined as "the ability or disposition to resist reporting the first response that comes to mind". Toplak et al (2014) have expanded the CRT to a total of seven questions in a study assessing people's tendency to process information miserly. Each of the seven questions presented a problem, which had an intuitive, but wrong answer immediately coming into mind, and requiring the suppression of that answer and searching for the correct one. For example, a problem describing that a bat and a ball cost 1.10 dollars in total, and the bat costs a dollar more
than the ball, asks how much the ball costs. An intuitive answer would be 10 cents, but the correct one is 5 cents.

## Participants

A total of sixty-seven participants took part in the study (18 males). All of them were native Bulgarians. They participated either for partial fulfilment for a course credit, or voluntarily. Forty-seven of the participants were students at the New Bulgarian University. The participants' age ranged from 18 to 53 years $(M=25.18, S D=8.21)$.

## Procedure

The procedure consisted in participants signing an informed written consent, and solving all six problems and the CRT individually in a single 45-50 minute session. First, the three base problems were presented one by one, with participants having 5 minutes to solve for each problem. If the participants failed to produce the correct solution, it was given to them. Then the CRT was given, with 10 minutes time to complete it. Finally, the three remaining target problems were given one at a time.

In order to control which of the target problems were solved spontaneously, the participants were given 5 minutes per problem, and if they did not produce the correct solution, they were given a hint to use one of the previously solved problems and additional 2 minutes were allowed. If again there was no correct solution, a second hint was given to use the specific analogous base problem to solve the current one, again allowing for additional 2 minutes.

The analogical problems were chosen in such a way, so as to be symmetrical, as well as structurally identical. Due to this fact, the analogical pairs were alternated with respect to being either a base or a target, with the Red Adair problem appearing half of the times as base, half of the times as target. The same applied for all six problems. The presentation of the base and target stories was balanced, with each of the stories appearing first, second or third as a base and first, second and third as a target equal amount of times. The full randomization resulted in 72 possible presentations of the problems without repetition of the presentation order. Thus, each participant was given a unique sequence of problems arrangements, with 67 out of the 72 randomized possibilities being used in the study.

## Results and discussion

## Analysis of the analogical problems

Several types of analyses were made on the obtained data. Firstly, the time to solve the base and target problems was calculated. The mean time to solve all three base problems was $198.29 \mathrm{sec}\left(\mathrm{s}^{4}=115.44\right)$, whereas the target problems were solved faster for an average of $161.88 \mathrm{sec}(\mathrm{s}=129.77)$. That difference was significant $(\mathrm{t}(66)=3.76$, $\mathrm{p}=.00)$, indicating that some facilitation due to analogical transfer may have taken place. The tasks in each analogical pair

[^500]were randomly assigned to the base or target position, thus any differences between the base and target task cannot account for the observed faster solutions of the target compared to the base problem. Moreover, only response time for correctly solved, but not for unsolved targets was faster than the base solution time: $\mathrm{F}(1,66)=38.32, \mathrm{p}=.00$ (Figure 1). Solved targets were worked out faster than solved base problems. Unsolved problems took up approximately the same amount of time, irrespective of the base-target role they have played in a given analogy. Therefore, analogy, rather than task order, may explain the obtained facilitation in solving the target tasks.


Figure 1: Time needed to solve successfully or not a base and a target problem (in seconds).

Huge differences in both response time and accuracy, however, were observed between the individual problem pairs which share analogous relational structure. The analogical pair Garden and Marching band (noted G and M, respectively) were correctly solved as bases for average of $183.43 \mathrm{sec}(\mathrm{s}=44.17)$, which took the longest amount time to be solved out of the three pairs. The Orange and Tribe pair (noted O and T, respectively) took 70.52 sec on average ( $\mathrm{s}=73.22$ ) or was fastest of the three problems to be successfully solved as bases, and the Red Adair and Aquarium problems (noted R and A, respectively) took on average $146.00 \mathrm{sec}(\mathrm{s}=90.25)$ to be solved correctly as bases (see Figure 2).


Figure 2: Time needed to correctly solve a problem from an analogical pair as a base and as a target (in seconds).
One-Way ANOVA yielded statistically significant difference with respect time to solve the bases $(\mathrm{F}(2,104)=$ $14.40, \mathrm{p}=.00$ ). Specifically, according to a Fisher LSD post
hoc test the pair O and T was solved correctly faster as a base compared to $G$ and $M(p=.001)$ and faster than $R$ and A ( $\mathrm{p}=.00$ ). The pair $G$ and $M$ did not differ from $R$ and $A(p$ $>$.05). Target problems from the analogical pair G and M were solved for $187.75 \mathrm{sec}(\mathrm{s}=72.62)$, which again was the longest amount of time out of the three problems. O and T were correctly solved for 67.18 ( $\mathrm{s}=91.88$ ), and R and $\mathrm{A}-$ for $62.18 \mathrm{sec}(s=64.43)$ (see Figure 2). There was again significant difference between time needed to solve targets from each pair $(\mathrm{F}(2,117)=4.82, \mathrm{p}=.01)$. Fisher LSD posthoc test showed that the G-M pair was correctly solved as a target for the slowest amount of time compared to O-T (p $=.004)$ and R-A ( $\mathrm{p}=.002$ ), while there was no significant difference between R-A and O-T pairs ( $\mathrm{p}>.05$ ).

Additionally, the number of spontaneously solved target analogies was calculated for each pair. The analogical pair G-M was solved spontaneously only 4 times, or by $5.97 \%$ of the people. For the pair O-T the number was 55 ( $82.01 \%$ ), and for the pair R-A it was 61 ( $91.05 \%$ ) (Figure 3). The difference between the number of spontaneously solved problems from the pair G-M was significant from that of O$\mathrm{T}\left(\chi^{2}(1, \mathrm{~N}=59)=78.76, \mathrm{p}=.00\right)$ and also from R-A $\left(\chi^{2}(1\right.$, $\mathrm{N}=65)=97.07, \mathrm{p}=.00)$. The difference between $\mathrm{O}-\mathrm{T}$ and R-A pairs was not significant. Likewise, participants reported less often that they have been aware of the analogy between the problems in the G-M, compared to the other analogous pairs: $\mathrm{F}(2,200)=39.72, \mathrm{p}=.00$.


Figure 3. Relative frequency of solved target problems for each analogical pair

In sum, the superficially dissimilar analogous problems used in that study were quite different with respect to solution time and accuracy. Some of the target problems were solved faster and more accurately (i.e. O-T and R-A) than others (i.e. G-M). The target problem itself can hardly explain that discrepancy, since both tasks in each pair were randomly assigned as base and target for each participant. The order of the three base and the three target tasks was also randomized across participants.
In this specific case, the G-M pair consisted of mathematical problems that, although analogical, might be impeding the correct mapping or retrieval that is necessary for correct solution just because of the difficulty of the problem itself. Given that mathematical expertise has been found to be an important predictor of analogical transfer (Novick \&

Holyoak, 1991), it could be reasonable to expect that for this specific analogical pair, some additional factors might have operated by impeding the transfer. The retrieval gap (Holyoak, 2012), however, seems to be wider for some analogous problems, but not for others, probably depending on the specific expertise of participants, as suggested in our study, where most participants had background in humanities ${ }^{5}$ and failed to solve the G-M problem that requires mathematical skills (Novick \& Holyoak, 1991).

## Correlational analyses: who solves problems by means of spontaneous analogies

A correlational analysis was conducted between the variables scores on the Cognitive Reflection Test and the number of analogical problems that were solved correctly without an explicit hint (i.e. spontaneously). Importantly, a Kolmogorov-Smirnov test was applied to test for normality the two variables. Both of them were not normally distributed ( $\mathrm{p}=.000$ ), which required utilizing a nonparametric correlational test, such as Spearman's rank order correlation. There was a significant positive correlation obtained between the two variables $\left(r_{s}(67)=.25, \mathrm{p}=.045\right)$. The results indicate that a high score on the CRT tends to go together with higher number of spontaneously solved analogical problems. A scatterplot summarizes the results (Figure 4).

## Correlation Chart



Figure 4: Scatter plot with jitter, showing the correlation between scores on CRT and spontaneously solved analogical problems. The x -axis represents the score on CRT, the $y$-axis represents the number of spontaneously solved analogical target problems.

## Discussion

This study demonstrates that a positive correlation exists between the score on the Cognitive Reflection Test and the number of analogical problems that are solved spontaneously. Given the rationale of the hypothesis, this result can be explained in terms of individual differences with respect to the Reflective mind (Stanovich, 2012) at least partially accounting for the analogical problems that

[^501]are solved without a hint. Although the correlation is weak, the results indicate that the goals and hypothesis of the research are in the right direction.
Generally, what was found in this investigation was that people solve different analogical problems with a different amount of speed and also different degree of success. The finding that the analogical pair $G-M$ was solved less than the other ones and for more amount of time might point to the idea that the nature of the problems themselves might play a role in how easy or how fast the solution is extracted from the base problem in order to be applied to the target one. A possible explanation remains to be looked for in expertise in solving mathematical problems (Novick \& Holyoak, 1991) In addition, the nature of the CRT itself could be questioned as to the extent it requires a certain level of expertise. Thomson and Oppenheimer (2016) have developed an alternate version of the CRT which addresses the criticisms to the original form - that it relies on mathematical sophistication to produce the correct answer.
The weak correlation that was found between scores on CRT and the spontaneously solved analogical problems needs to be compared to other similar correlations of CRT and cognitive abilities. For example, Toplak, West, \& Stanovich (2011) show significant correlations to exist between CRT and syllogistic reasoning tasks ( $\mathrm{r}=.36$ ), heuristic-and-biases tasks ( $\mathrm{r}=.42$ ), executive functions measures (. 17 to .34 ) and thinking dispositions measures (. 18 to .19). Thus, the current study seems comparable to others with respect the strength of association between CRT and tasks involving reasoning measurement.
It should be noted, however, that correlations between CRT and cognitive ability measured by Wechsler Abbreviated Scale of Intelligence has been found to exist ( $\mathrm{r}=.32$ ), suggesting some overlap between the two (Toplak et al., 2011). A possibility to search for a partial explanation of spontaneously solved analogical problems in the cognitive ability of intelligence, thus, cannot be fully overruled.

## Conclusion

The reported correlation between the Reflective Mind measure and spontaneous analogical problem solving adds a new explanation for the retrieval gap in analogical reminding. Low superficial similarity and non-compatible relational structures between the base and target problems may explain the difficulties that participants robustly demonstrate in psychological labs when analogical problem solving abilities are tested by the means of the reception paradigm. Nevertheless, generally $20 \%$ of participants find the analogous solution (Holyoak, 2012), despite the mentioned difficulties that the reception paradigm seems to impose on them. The reported correlation indicates that among the key abilities within the profile of the successful problem solver is the reflective reasoning. It presumably enables the motivated search for possible connections between the tasks, and possibly a re-representation of the relevant relations, if needed for the purposes of the analogical problem solving. Therefore, spontaneous analogy
making may benefit from the reflective reasoning, since it most probably transforms the task into an explicit task for searching the analogy, or at least boosts the motivation to cope with the task.

## Acknowledgements

This research was supported financially by the European Office for Aerospace Research and Development under grant FA9550-15-1-0510 (Anticipating Future by AnalogyMaking).

## References

Blanchette, I., \& Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. Memory \& cognition, 28(1), 1-33.
Catrambone, R., \& Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. JEP: Learning, Memory, and Cognition, 15(6), 1147-1156.
Day, S. B., \& Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. JEP: Learning, Memory, and Cognition, 37(3), 551-567.
Fisher, R., Ury, W. L., \& Patton, B. (2011). Getting to yes: Negotiating agreement without giving in. Penguin.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic perspectives, 25-42.
Gick, M. L., \& Holyoak, K. J. (1980). Analogical problem solving. Cognitive psychology, 12(3), 306-355.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive psychology, 15(1), 1-38.
Holyoak, K.J. (2012). Analogy and relational reasoning. The Oxford handbook of thinking and reasoning, 234-259.
Holyoak, K. J., \& Koh, K. (1987). Surface and structural similarity in analogical transfer. Memory \& Cognition, 15(4), 332-340.
Kurtz, K. J., \& Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. Memory \& Cognition, 35(2), 334-341.
Novick, L. R., \& Holyoak, K. J. (1991). Mathematical problem solving by analogy. JEP: Learning, Memory, and Cognition, 17(3), 398-415.
Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. The Oxford handbook of thinking and reasoning, 343-365.
Thomson, K. S., \& Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. Judgment and Decision Making, 11(1), 99-113.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory \& Cognition, 39(7), 1275-1289.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking \& Reasoning, 20(2), 147-168.

# Gender or Community: What Drives STEM Interest Among Middle School Students? 

Evans Smalley (evans.smalley@wright.edu)<br>Department of Math and Statistics, Math/Microbiology Bldg 103<br>3640 Colonel Glenn Hwy, Dayton, OH 45435-0001<br>Jennifer Kaminski (jennifer.kaminski@wright.edu)<br>Department of Math and Statistics, Math/Microbiology Bldg 103<br>3640 Colonel Glenn Hwy, Dayton, OH 45435-0001


#### Abstract

Much attention has been given to increasing women's and girls' interest and participation in STEM fields. One way of increasing STEM interest is to target STEM-gender stereotypes by presenting students with female stereotypedisconfirming exemplars. However, the exemplar approach has had mixed effectiveness in adolescent populations. The present study examined middle school students' interest in STEM fields and their communal goal interest. Participants were given different interventions that either presented a female exemplar of a scientist or leveraged communal goals. Results found no gender differences in STEM interest, but a correlation between communal goal endorsement and a belief that STEM careers are compatible with communal goals. The intervention that leveraged communal goals effectively increased STEM interest for some students, while the exemplar interventions were ineffective. These findings suggest that with regard to STEM interest in early adolescence, endorsement of communal goals may be a more influential factor than gender category membership.


Keywords: Science, Technology Engendering and Math, Gender, Communal Goals, Adolescents

## Introduction

As the world becomes more reliant on technology, there is a growing interest in how to increase participation in science, technology, engineering and math (STEM) fields. Of particular concern is the underrepresentation of women in STEM fields. For example, in 2012, only $20 \%$ of bachelors' degrees in physics, engineering, and computer science were earned by women (National Science Foundation, 2015). The gender disparity is more pronounced at higher levels. For example, women earned $40 \%$ of bachelor's degrees in mathematics and statistics but only $20 \%$ of master's and doctoral degrees in these fields (NSF, 2015).
What explains the gender differences in levels of STEM interest and participation, and when do these differences emerge? An interesting part of the situation is the fact that there are no substantial gender differences in early mathematics and science ability (see Hill, Corbett, \& St. Rose, 2010 for discussion), but there are differences in attitudes toward mathematics and science (e,g, Ganley \&

Lubienski, 2016; Hyde, Fennema, \& Lamon, 1990). As early as elementary and middle school, girls tend to perceive themselves as less competent in mathematics than do boys (Ganley, \& Lubienski, 2016; Herbert \& Stiptik, 2005). In middle school and high school, boys are more likely to state that they are interested in math and science than girls (Cunningham, Mulvaney, \& Sparks, 2015; Hill, Corbett, \& St. Rose, 2010).

One factor that may contribute to students' attitudes toward mathematics and science is the pervasive stereotypes that scientists and mathematicians are men (Eccles, 1987; Fennema, 1985; Nosek, Banaji, \& Greenwald, 2002). Young children in the United States tend to draw men when asked to depict a scientist or mathematician (Chambers, 1983; Steele, 2003). There is evidence that gender and the strength of gender identity are correlated with preference for mathematics, mathematics identity, and math-gender stereotypes (Nosek, Banaji, \& Greenwald, 2002).

Because it is possible that gender-STEM stereotypes negatively impact women and girls participation in STEM, many interventions have been developed to break common stereotypic misconceptions by presenting stereotypedisconfirming exemplars to students. For example, reading about the successes of women in STEM and non-STEM fields has been shown to boost women's performance on math examinations (McIntyre, Paulson \& Lord, 2003).

The proposed mechanism behind these exemplars is that encountering a particular stereotype-defying example will encourage participants to expand their notion of membership in the stereotyped category. In this instance, presenting examples of women who are successful in STEM fields would prompt a change in the perceived membership of the category of "scientist" or "mathematician". By extension, presenting a highly feminine exemplar as successful in a stereotypically male domain like STEM may be effective because it forces the perceiver to include all feminine characteristics in their category definition of who may be successful in STEM fields and discourages subtyping.

Although the power of stereotypes and representation biases cannot be denied, the stereotype explanation leaves questions about gender differences in STEM participation. Women and girls continue to select out of STEM fields, even as women continue to gain representation in other high
achieving, stereotypically male disciplines like law and medicine (Snyder, Dillow, \& Hoffman 2009; Wang, Eccles \& Kenny, 2013). This suggests that other factors such as motivation and personality characteristics may be contributing to the gender disparity in STEM fields. In addition, there may be practical problems with designing STEM participation interventions based simply on expanding students' perception of STEM "membership" because these theories assume that students will see the exemplar as similar to themselves. How would characteristics such as race, socio-economic background, geographic region, and culture be appropriately included in exemplars to appeal to all students?

Indeed, despite their success with adult women, exemplars often generate mixed effects, especially among adolescents. Exposure to highly feminine exemplars actually weakens future goals to take optional math and science among adolescent girls who are disidentified with STEM fields and decreases adolescent girl's sense of efficacy and short-term expectations of success in math and science (Betz \& Sekaquaptewa, 2012). This evidence suggests that using highly feminine exemplars to influence the future career plans of adolescent girls does not target the cleanest and most effective mechanism of STEM interest and identification.

These counterintuitive effects may stem from the complicated relationship adolescent girls have with gender stereotypes. During adolescence, especially early adolescence, girls' self-perceived math ability begins to decline relative to boys' (Wigfeild, Eccles, Mac Iver, Reuman \& Midgely, 1991). In light of this intricate relationship between budding identity development and harsh cultural stereotypes, it is not surprising that interventions designed to leverage mechanisms tied to gender stereotypes yield adverse or mixed results.

Considering the baggage accompanying gender-STEM stereotypes during adolescence, it is important to explore other theories and mechanisms that could shed light on this phenomenon. One such theory is the goal congruity perspective. The goal congruity perspective posits that women highly value communal goals like intimacy, working together and helping others. This valuation is at odds with stereotypes about STEM fields portraying careers involving those fields as individualistic and isolating. In fact, compared to other high achieving careers, STEM careers are perceived as actually hindering the path to attain goals like helping others (Diekman, Brown, Johnson \& Clark, 2010). These stereotypes work to portray STEM careers as incompatible with communal goals, leading to disinterest in pursuing math and science domains for those individuals that value communal goals. Thus, the mechanism behind the goal congruity perspective attempts to increase the degree to which goal affordance beliefs, or beliefs about what actions or pursuits will best facilitate the attainment of specific goals, align STEM careers with communal goals. In this way it might be possible to increase positivity toward STEM careers among individuals
who value communal goals (Diekman, Clark, Johnston, Brown \& Steinberg, 2011).

Indeed, this idea maps well on to the difference in female representation within STEM fields. According to 2013 census data, women make up $61 \%$ of social scientists, a field that has high communal goals stereotypicallity compared to only $27 \%$ and $13 \%$ of computer workers and engineers respectively, fields with low communal goal steryotypicallity (Landivar, 2013).

## The current study

This study seeks to examine STEM interest in an adolescent population and to test several interventions designed to increase STEM interest. Specifically, we will consider the role of gender and communal goal endorsement. The use of exemplars, along with manipulations of female stereotypicality has yielded mixed results in this population, while the communal goal approach remains untested. It is predicted that adolescents will endorse communal goals and that the intervention targeting communal goals will yield the greatest increase in STEM interest, in particular a future desire to purse science and math, compared to exemplar focused interventions.

## Method

## Participants

Ninety-four seventh grade students (41 male, 53 female) at middle school in the Midwestern United States were recruited for this study. The majority ( $93 \%$ ) of the sample self identified as White. Parental consent as well as participant consent was obtained.

## Materials

The experiment included three phases: (1) measure of communal goal interest and pretest of STEM interests (2) an intervention designed to increase STEM interests and (3) posttest of STEM interest. Participants were randomly assigned to one of three conditions (feminine exemplar, neutral exemplar, or communal goals) that varied the intervention.
Interventions. Participants were exposed to either a stereotypically feminine exemplar (feminine condition), a female but stereotypically neutral exemplar (neutral condition), or a group exemplar designed to speak to communal goals (communal condition). The intervention text appears in Appendix A exactly as it was presented to participants. All exemplars were identical except for the necessary manipulations.
The material presented to participants consisted of a short paragraph written form the perspective of the exemplar detailing their hobbies and what they enjoy about their job. A picture followed this short paragraph. In the communal condition the picture included an ethnically diverse group of men and women, and in the feminine and neutral conditions the picture included an ethnically ambiguous woman.

Exemplars were portrayed as a "profile of a scientist" (feminine and neutral conditions) or the "profile of a team of scientists" (communal condition). This profile was sourced from a website that aims to connect students to female mentors working in STEM fields and can be found in full in Appendix A.
The manipulations of feminine stereotypicality (feminine and neutral conditions) were taken from Clark, Fuesting, \& Diekman, 2016. This manipulation consisted of hobbies that were independently rated as highly feminine (knitting, watching romantic comedies and yoga) or neutral on feminine stereotypicality (running, watching nature documentaries and photography). The hobbies for the communal condition were simply "working and spending time together."

Following the intervention manipulation, participants were asked to complete a writing assignment. In this task, they first read a list of daily tasks the exemplar would perform. Then they were asked to write about what they think an average day would be like if they were a scientist doing a similar job (Appendix A). This writing exercise was timed; all participants were instructed to write for five minutes. The daily tasks were manipulated to reflect or not reflect communal goals similar to the manipulation in Clark, Fuesting, \& Diekman, 2016. For the feminine and neutral conditions the daily tasks consisted mainly of solitary work and problem solving (ex: "Look up and read about past research to help you develop new experiments."). In the communal condition the daily tasks reflected working with others (ex: "Brainstorm with your fellow researchers about past research to help you develop new experiments.").

Communal Goals Scale. The nine-question communal goals scale (Clark, Fuesting, \& Diekman, 2016) was used to assess communal goal identification. Participants were asked to rate the importance to them of nine goals, like helping others, intimacy, relationships with others and working with people, on a seven-point Likert scale.

Measures of STEM Interest. Three measures were used to assess STEM interest: a measure of future goals (Betz \& Sekaquaptewa, 2012), a measure of STEM positivity and a measure of STEM goal attainment (Clark, Fuesting, \& Diekman, 2016). The measure of future goals asked participants to rate their likelihood of taking science and math classes in high school and college on a seven point Likert scale. The measure of positivity asked participants to rate how positive they feel towards a STEM career and how much they would enjoy being successful in a STEM career on a seven point Likert scale. The STEM goal attainment measure contains two questions that ask participants to rate on a seven point Likert scale the extent to which a career in STEM fields would "fulfill goals like intimacy, working with people and helping people" and to what extent such a career would fulfill their own goals of intimacy, working with people and helping people.

## Procedure

Students took part in this study as an extension of their daily math class. Participants were first administered consent forms. Participants then completed a pen and paper pre-test questionnaire consisting of the communal goals scale, the two question future goals scale, the two question STEM positivity scale, and the two question STEM goal attainment scale. Participants were also asked to list their top three favorite academic subjects.

Participants were then exposed to the exemplar intervention materials that aligned with their assigned condition (feminine, neutral or communal). Participants were then given a uniform amount of time ( 5 minutes) to complete the writing assignment corresponding to their condition: the feminine and neutral condition received the non-communal manipulation and the communal condition received the communal manipulation.

Following this writing assignment participants were administered a post-test consisting of the future goals scale, the STEM positivity scale and the goal attainment scale. Participants were then thanked and debriefed.

## Results

## Effects of gender and communal goal endorsement.

The communal goals scale, the future goals scale, the STEM positivity scale and the goal attainment scale were each assessed on a 1 (low) to 7 (high) likert scale. These four scales were averaged to yield mean scores for each participant on each measure. Eleven participants were removed from analysis. One participant was removed due to incomplete responses and ten were removed as outliers due to reporting a communal goals score or a measure of STEM interest score that was greater than 2.5 standard deviations from the mean of the entire sample.

Overall, participants reported high communal goals scores and pre-test STEM interest scores (Table 1). Eighty percent of participants reported an average communal goal score of 5 or above and only two participants reported an average communal goal score below 4. There were no gender differences on ratings of communal goals or measures of STEM interest, ANOVA with gender as a factor $F(1,80)$ s $<$ 1.52, $\mathrm{ps}>.22$.

To consider whether participants' communal goal endorsement affected STEM interest, correlations between participants' communal goals ratings and their STEM interest scores were examined. Pearson correlation found that communal goals scores were positively correlated with STEM goal attainment scores $r(82)=.35, p<.01$. Communal goals scores were not correlated with future goals scores $r(82)=.18, p>.05$ or STEM positivity scores $r(82)=.10, p>.05$.

Table 1. Mean Ratings of Communal Goal Endorsement and STEM Interest Measures. Standard deviations are in parentheses.

|  | STEM Interest Measures |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | Communal <br> Goal <br> Endorsement | Future <br> Goals | Positivity | Goal <br> Attainment |
| Boys | $5.67(.73)$ | $5.94(.86)$ | $5.57(1.5)$ | $5.77(.96)$ |
| Girls | $5.86(.68)$ | $6.13(.88)$ | $5.40(1.3)$ | $5.80(.93)$ |

## Effectiveness of the Interventions.

Because communal goals scores were positively correlated with goal attainment scores, it is possible that the effects of the interventions may be different for participants with high and low communal goals scores. To consider the effect of communal goal endorsement, a median split was performed based on communal goal ratings, $($ median $=5.78$ ) which resulted in $46 \%$ of participants below the median and $54 \%$ or participants at or above the median.

To examine the effectiveness of the three different interventions and possible interactions with communal goal endorsement, a 2(gender) X 2(communal goals: high, low) X 3(condition: feminine, neutral, communal) repeated measures ANOVA was conducted with pretest and posttest STEM interest (time) as the within repeated measures. There was a significant interaction of time and gender on future goals, $F(1,68)=4.04, p<.05, \eta_{p}{ }^{2}=.06$. Overall, girls' post-intervention future goals scores decreased ( $M=$ 6.13, $S D=.88$ to $M=6.03, S D=.87$ ) while boys' ratings increased $(M=5.93, S D=.86$ to $M=6.19, S D=.71)$. A significant interaction between time and communal goals emerged for goal attainment, $F(1,68)=9.05, p<.01, \eta_{p}{ }^{2}=$ .11. Goal attainment scores decreased $(M=5.96, S D=.86$; $M=5.43, S D=1.2$ ) for participants with high communal goal ratings and increased for participants with low communal goal ratings $(M=5.59, S D=.99 ; M=6.00, S D=$ .91).

There were no significant interactions with time, condition, or gender on any of the measures of STEM interest $F(2,70)$ s $<2.46, p \mathrm{~s}>.08$. However, a significant interaction between time, condition and communal goals on future goals scores was observed $F(2,69)=5.879, p<.01, \eta_{p}{ }^{2}=$ .14. This result suggests that condition produced different changes to STEM interest ratings (in particular future goals) as a function of participants' communal goal score.

To explore the interaction between time, condition, and communal goals and further investigate the effectiveness of each intervention on future goals scores, separate ANOVAs were conducted on future goals with communal goals scores as a covariate on each intervention condition. The communal condition evidenced a significant interaction between time (pretest and posttest STEM interest) and communal goals $F(2,25)=6.44, p<.01, \eta_{p}{ }^{2}=.46$. No significant effects of time, communal goals or interactions between time and communal goals were observed in the feminine or neutral conditions, $F(2,27)$ s $<2.17, p \mathrm{~s}>.08$. This suggests that only participants in the communal goal condition had significant changes in their STEM future goal ratings. In the communal condition, future goal scores for participants with low communal goal scores increased
significantly by $15 \%$, paired sample t -test $t(11)=2.40, p<$ .05. However, none of the changes in the other conditions were statistically different than $0, t \mathrm{~s}<=1.74, p \mathrm{~s}>.11$. Figure 1 presents the percent increase in future goals scores across the three conditions, split by high and low communal goals ratings.


Figure 1: The percent increase in STEM Future Goal scores by experimental intervention condition and split into high and low communal goals groups. Error bars represent standard error of the mean.

## Discussion

The goal of the present research was to examine adolescents' STEM interest and the relationships to gender as well as communal goal endorsement. Several interventions designed to increase STEM interest were tested. Previous research has suggested that girls are less interested in STEM fields than boys are, and that exposing students to examples of female scientists will increase STEM interest. However, we found no gender differences on any measure of STEM interest. Further, the interventions presenting a highly stereotypically feminine exemplar or a female but stereotypically neutral exemplar did not improve STEM interest among female or male participants.

While the interventions involving female exemplars did not increase STEM interest, the communal goals intervention did. This intervention was designed to increase recognition of the communal aspects of work in STEM fields. Participants with low communal goal scores who received this intervention increased their future goals scores, suggesting they were more likely to take optional science and math classes in high school and college. This increase in STEM interest was only observed in the low communal goals category possibly because participants in the high communal goals category reported future goals scores at ceiling on both the pre-test and post-test measures.

It was found that this sample of adolescents evidenced high average communal goal endorsement overall.

Approximately $80 \%$ of participants averaged a score at or above 5 on a seven-point scale. Communal goal endorsement was also positively correlated with goal attainment scores, or the belief that a career in STEM fields would fulfill both communal goals in general, and specifically the participants' communal goals.

These findings lend support to the goal congruity hypothesis of STEM interest. A simple, 10-minute intervention emphasizing the role of communal goals in STEM fields was able to significantly and positively affect one measure of STEM related future goals of students who did not score at ceiling on the communal goals scale. In comparison, more traditional interventions centered on exemplars and the emphasis of femininity produced no significant change in any measure of STEM interest.

Of course this is not to say that the power and influence of gender stereotypes should be discounted, certainly they do play a role in the decision to pursue a STEM career (McIntyre, Paulson \& Lord, 2003; Nosek, Banaji, \& Greenwald, 2002). However, our data suggests that during early adolescence valuing communal goals, like working with and helping others, may possibly be a more influential factor than gender category membership on STEM interest. Simply attributing STEM interest to the influence of gender stereotyping may be somewhat of an oversimplification. Thus, our data suggests that emphasizing communal goals in this population may be a more effective and practical way to increase STEM interest than more traditional interventions that emphasize stereotype-disconfirming exemplars and femininity.

There are limitations to this study. The sample was taken from one school district and was not racially diverse. In addition, our participants appeared to have high communal goal desires causing possible ceiling effects, which might have obscured other findings. Future studies should explore this finding with a larger, more diverse sample and a modified communal goals scale to attempt to combat ceiling effects.

## References

American Association of University Women. (2010). Why so few? Women in science, technology, engineering, and mathematics. Washington, DC. AAUW.
Betz, D. E., \& Sekaquaptewa, D. (2012). My fair physicist? Feminine math and science role models demotivate young girls. Social Psychological And Personality Science, 3(6), 738-746.
Chambers, D. W. (1983). Stereotypic images of the scientist: The draw-a-scientist test. Science Education, 67, 255-265.
Clark, E. K., Fuesting, M. A., \& Diekman, A. B. (2016). Enhancing interest in science: Exemplars as cues to communal affordances of science. Journal Of Applied Social Psychology, 46(11), 641-654.

Cunningham, B.C., Mulvaney, K. H. ,\& Sparks, D., (2015). Gender Differences in Science, Technology, Engineering, and Mathematics (STEM) Interest, Credits Earned, and NAEP Performance in the $12^{\text {th }}$ Grade. Retrieved from https://nces.ed.gov/pubs2015/2015075.pdf
Diekman, A. B., Brown, E. R., Johnston, A. M., \& Clark, E. K. (2010). Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers. Psychological Science, 21(8), 1051-1057.
Diekman, A. B., Clark, E. K., Johnston, A. M., Brown, E. R., \& Steinberg, M. (2011). Malleability in communal goals and beliefs influences attraction to stem careers: Evidence for a goal congruity perspective. Journal Of Personality And Social Psychology, 101(5), 902-918.
Eccles, J. S. (1987). Gender roles and women's achievement-related decisions. Psychology of Women Quarterly, 11, 135-172.
Fennema, E. (1985). Attribution theory and achievement in mathematics. In S. R. Yussen (Ed.), The development of reflection. New York: Academic Press.
Ganley, C. M., \& Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. Learning and individual differences, 47, 182-193.
Herbert, J. \& Stipek, D. (2005). The emergence of gender difference in children's perceptions of their academic competence. Journal of Applied Developmental Psychology, 26(3), 276-295.
Hill, C., Corbett, C., \& St. Rose, A. (2010). Why so few? Women in science, technology, engineering, and mathematics. Washington, DC, AAUW. Retrieved from https://www.aauw.org/files/2013/02/Why-So-Few-
Women-in-Science-Technology-Engineering-and-
Mathematics.pdf
Hyde, J.S., Fennema, E., \& Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. Psychological Bulletin, 107(2), 139-155.
Landivar, 1., C. (2013). Disparities in STEM employment by sex, race and hispanic origin. American Community Survey Reports, ACS-24, U.S. Census Bureau, Washington, DC
McIntyre, R. B., Paulson, R. M., \& Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. Journal of Experimental Social Psychology, 39, 83-90.
National Science Foundation, National Center for Science and Engineering Statistics. (2015). Women, Minorities, and persons with disabilities in science and engineering: 2015 Special Report NSF 15-311. Arlington, VA. Retrieved from http://nsf.gov/statistics/wmpd/
Nosek, B. A., Banaji, M. R., \& Greenwald, A. G. (2002). Math $=$ male, me $=$ female, therefore math $\neq$ me. Journal Of Personality And Social Psychology, 83(1), 44-59.
Snyder, T.D., Dillow, S.A., \& Hoffman, C.M. (2009). Digest of edu- cation statistics, 2008 (NCES 2009-020).

Washington, DC: U.S. Department of Education, National Center for Education Statis- tics, Institute of Education Sciences.
Steele, J. (2003). Children's gender stereotypes about math: The role of stereotype stratification. Journal of Applied Social Psychology, 33, 2587-2606.
Wang, M., Eccles, J. S., \& Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. Psychological Science, 24(5), 770-775.
Wigfield, A., Eccles, J.S., Mac Iver, D., Reuman, D.A., and Midgley, C. (1991). Transitions during early adolescence: Changes in children's domain-specific selfperceptions and general self-esteem across the transition to junior high school. Developmental Psychology, 27, 552565.

## Acknowledgements

This research was supported by a grant from the Institute of Education Sciences, U.S. Department of Education
(\#R305A140214).

## Appendix A: Stimuli

## Feminine and Neutral Conditions.

Profile of a Scientist:
Name: Lisa Johnson
Biography:
Lisa is an electrical engineer. She enjoys yoga, watching romantic comedies and knitting (In Neutral Condition: running, watching nature documentaries and photography) Lisa Says:

I am Electrical Engineer at the NASA Glenn Research Center. I work on Power System Development for future NASA missions. Specifically, I support the development of a power processing unit for high power, high voltage electric propulsion applications and the development of a flywheel energy storage system.

Flywheel energy storage, or storing energy mechanically in a rotating wheel, offers an alternative to traditional chemical energy storage systems, such as batteries, for future missions.

Every day I work to solve problems and learn new things. I figure out how to implement new technologies all the time and I go home everyday knowing that I made a difference!
Writing Activity:
Listed below are some typical daily tasks that a scientist like Lisa would perform. Imagine you are also a scientist doing work in an environment similar to Lisa's. What do you think your average day would be like? Please write a few sentences describing what you think it would be like if you worked as a scientist like Lisa.

Daily Tasks of a Scientist:

- Check a database for updates on ongoing experiments.
- Look up and read about past research to help you develop new experiments.
- Watch videos of other scientists presenting their recent findings
- Update a lab notebook with information about the progress and status of your experiments
- Work out data analysis problems by yourself
- Make a PowerPoint presentation of your recent experimental results to email to a supervisor


## Communal Condition.

Profile of a team of Scientists:
Name: NASA Power Team
Biography:
The NASA Power Team is made up of six electrical engineers. They enjoy working and spending time together The Power Team Says:

We are a team of six Electrical Engineers at the NASA Glenn Research Center. We work closely together on Power System Development for future NASA missions. Specifically, we support the development of a power processing unit for high power, high voltage electric propulsion applications and the development of a flywheel energy storage system.

Flywheel energy storage, or storing energy mechanically in a rotating wheel, offers an alternative to traditional chemical energy storage systems, such as batteries, for future missions. This technology also has the potential to help people here on Earth in many ways.

Every day we work together to solve problems and learn new things. Together, we figure out how to implement new technologies all the time. Each of us goes home everyday knowing we made a difference!

## Writing Activity:

Listed below are some typical daily tasks that scientists like The NASA Power Team would perform. Imagine you are also a scientist doing work in an environment similar to the NASA Power Team. What do you think your average day would be like? Please write a few sentences describing what you think it would be like if you worked as a scientist like the NASA Power Team.

Daily Tasks of a Scientist:

- Talk with team members about updates on ongoing experiments.
- Brainstorm with your fellow researchers about past research to help you develop new experiments.
- Attend presentations from other scientists about their recent findings
- Update your team coordinator with information about the progress and status of your experiments
- Work out data analysis problems with your other members of your lab team.
- Present your recent experimental results to your supervisor with your team members.


# 20-month-olds Use Social-Group Membership to Make Inductive Inferences 

Megan A. Smith (msmith57@ucmerced.edu)<br>Rose M. Scott (rscott@ucmerced.edu)<br>Psychological Sciences, University of California Merced


#### Abstract

Previous research suggests that preschool children expect members of social groups to share stable, inherent characteristics (e.g., Waxman, 2013). Here we explored the origins of these social-group based inferences by examining whether infants generalize food preferences across members of an arbitrary social group. Experiment 1 demonstrated that infants expected two individuals to share food preferences when they belonged to the same social group, but not when they belonged to two different social groups. Experiment 2 replicated and extended these findings to social groups that were labeled with adjectives instead of nouns. These results suggest that by 20 months of age, infants use social-group membership to make inductive inferences about the behavior of group members.


Keywords: social groups; inductive inference; psychological reasoning; social cognition

## Introduction

Categorization is vital to human cognition. Category representations provide an efficient way of organizing our knowledge about the world, and they enable generalization of prior knowledge to novel entities and situations (Gelman, 1988; Medin, ojalehto, Waxman, \& Bang, 2015). Upon identifying that a novel entity belongs to a familiar category (e.g., dog), one can infer that it likely possesses common properties of that category (e.g., it is alive, wags its tail, etc.). Categories also aid in reasoning about kinds of people: Adults tend to assume social categories (e.g., doctors, women) capture fundamental, inherent similarities amongst collections of individuals and thus use prior knowledge about a social category to make inductive inferences about the physical, psychological, and behavioral properties of novel group members (e.g., Agerström, Björklund, Carlsson, \& Rooth, 2012).

The tendency to use social-group membership to make inductive inferences about category members is well established by the preschool years. (e.g., Bigler, Jones, \& Lobliner, 1997; Birnbaum, Deeb, Segall, Ben-Eliyahu, \& Diesendruck, 2010; Diesendruck \& HaLevi, 2006; Waxman, 2013). For example, 5-year-old children expect that members of the same social category will prefer the same activities (Diesendruck \& HaLevi, 2006) and preschoolers expect that members of the same, but not the opposite, sex will prefer the same toys (Martin, Eisenbud, \& Rose, 1995).

When and how does this tendency to make social-group based inferences emerge? As early as 3 months, infants notice visual and auditory features that are associated with social-group membership (e.g., Bar-Haim, Ziv, Lamy, \& Hodes, 2006; Howard, Henderson, Carrazza, \& Woodward, 2015; Shutts, Kinzler, McKee, \& Spelke, 2009). Evidence
for this comes primarily from tasks that assess whether infants demonstrate preferences for individuals who are similar to themselves. For example, by 3 months infants living in primarily own-race environments prefer to attend to own-race over other-race faces (Bar-Haim et al., 2006). By 19 months, infants prefer to accept toys and foods endorsed by a speaker of their native language over a speaker of a foreign language (Kinzler, Dupoux, \& Spelke, 2007; Shutts et al., 2009) and are more likely to imitate actions produced by a native-language speaker (Howard et al., 2015). By 11.5 months, infants also attend to food preferences and clothing as potential markers of group membership (Mahajan \& Wynn, 2012).

There is also some evidence that infants spontaneously categorize individuals into social groups instead of merely detecting features correlated with group membership (Powell \& Spelke, 2013; Liberman, Kinzler, \& Woodward, 2014; Liberman, Woodward, Sullivan, \& Kinzler, 2016; Rhodes, Hetherington, Brink, \& Wellman, 2015). These findings come from "third-party" tasks in which the infant is not a member of the groups in question. For example, Liberman et al. (2016) examined whether 14 -month-old infants expect individuals who affiliate to share food preferences. In a violation-of-expectation task, infants were first introduced to two actors who either affiliated with one another by smiling and saying "Hi", or disengaged from one another by turning away, crossing their arms, and saying "Hmph." Next, Actor-2 watched as Actor-1 ate one of two foods and emoted positively. In the test trial, Actor-2 ate the same food and emoted negatively, actively disagreeing with Actor-1's preference. Infants who saw the actors affiliate expected them to prefer the same food, and looked longer if they disagreed. In contrast, infants did not expect actors that had previously disengaged to share food preferences. Together with the results of several additional conditions, these findings suggested that infants can use social relationships to predict and interpret the behavior of agents.

However, it remains unclear whether infants expect that members of a social category will share stable, inherent characteristics. This is because in prior studies, such as the one just described, the target character always acted in the presence of its group members. Infants' responses may therefore have reflected an expectation that the target character would conform to social pressures or imitate group members, rather than expectations about the inherent properties and tendencies of the individual. The present research thus asked whether infants use social group membership to make inductive inferences about the properties of an individual, even when that individual is acting in the absence of other group members.

To address this question, we examined whether 20-month-old infants expected members of a social group to share food preferences. Several decades of research suggest that by this age, infants attribute preferences to agents (e.g., Woodward, 1998). Moreover, by 18 months infants assume that an agent's preference is specific to that individual unless given indication otherwise (Egyed, Király, \& Gergely, 2013). This allowed us to test whether social-group membership overrides this default assumption. We focused specifically on food preferences because foods are culturally relevant and thus likely to be shared amongst members of a social group (Cashdan, 1998; Rozin \& Siegal, 2003).

Infants were tested in a violation-of-expectation task involving arbitrary social groups, Topids and Brinkos. Arbitrary social groups were used in order to ensure infants had equal amounts of experience with the social groups being tested. The groups were identified using both shared appearance and noun labels because previous research suggests that infants might not form social categories based on physical appearance alone (e.g., Powell \& Spelke, 2013). In the familiarization trials, infants saw a member of one of the social groups (a Topid) demonstrate a preference for one of two novel foods. In the test trial, infants saw a single agent from either the same group (another Topid) or a different group (a Brinko) choose between the two foods. If infants use social group membership to make inductive inferences about food preferences, then they should expect members of the same social group to pick the same foods, and should look longer if the Topid picks a different food instead. In contrast, infants should not use the preferences of one social group to make inferences about the preferences of an individual from a different social group. Infants should thus have no expectations about what the Brinko should choose and look equally regardless of whether she chooses the same or a different food.

## Experiment 1

## Method

Participants 36 healthy term infants participated (18 female; ages 18 months, 10 days to 21 months, 18 days, $M=$ 20 months, 8 days). Another 12 infants were tested but excluded because they were fussy (8), because of parental interference (2), or because their test looking time was over 2.5 SD away from the mean of their condition (2). Half the infants were randomly assigned to the same-group condition ( $M=20$ months, 16 days) and half to the different-group condition ( $M=20$ months, 0 days).

Stimuli Stimuli consisted of digitized high-definition video recordings of actors performing a series of actions. All infants saw four familiarization trials and one test trial. A separate video was played for each trial. Each trial consisted of an initial phase followed by a final phase. The duration of the initial phase was fixed and identical for all participants. The duration of the final phase was infant-controlled. All trials are described from the infants' perspective.


Figure 1: Familiarization trial 1 of the same-group condition of Experiment 1.

Same-group familiarization trials At the start of the first familiarization trial, three female actors sat around a table (Figure 1). Two of the actors (Topid-A, Topid-B) wore bright pink turtlenecks and decorated yellow visors while the third (Brinko-A) wore a plaid shirt and a propeller hat.

All actors began the trial with their heads down. During the $10-\mathrm{s}$ initial phase of the trial, the actors looked at one another and labeled themselves: Topid-A said "Hi, I'm a Topid," Topid-B said, "Hi, I'm a Topid too," and Brinko-A said, "Hi, I'm a Brinko." As each actor labeled herself, she looked back and forth between the other two actors. When not labeling themselves, the actors looked at the speaking actor as she spoke. After all actors had labeled themselves, the actors looked down and paused. The infants viewed this paused scene until the trial ended (see Apparatus and procedure section for trial-ending criteria).

The infants then received three familiarization trials in which Topid-A demonstrated her preference for one of two foods. On each trial, Topid-A sat behind a table. In front of her were two white plates ( 18 cm in diameter) placed 25 cm apart. The plate on the right held purple pasta and the plate on the left held blue cereal. During the $10-\mathrm{s}$ initial phase of the trial, Topid-A selected one of the foods (counterbalanced across infants) and ate it while saying, "Mmm!" She then looked down at the center of the table between the two plates and paused until the trial ended. Topid-A selected the same food on all three trials, demonstrating that she preferred it to the other food.


Figure 2: Test events shown in the same-group condition of Experiment 1.

Same-group test trials The infants received either a samefood or different-food test trial (Figure 2). For ease of description, the test trials are described from the perspective of the infants who saw Topid-A choose blue cereal in the familiarization trials.

At the start of the trial, none of the actors were present. The plates of blue cereal and purple pasta again sat on the table. During the $10-\mathrm{s}$ initial phase of the trial, Topid-B entered from the left, sat down, and then selected a piece of blue cereal (same-food event) or purple pasta (different-food event), raised it to her mouth, and ate the food. After eating the food, she proceeded to say, "Mmm!" and smile, displaying positive affect and indicating she enjoyed eating the food. She then looked down at the center of the table between the two foods and paused until the trial ended.


Figure 3: Familiarization trial 1 of the different-group condition of Experiment 1

Different-group familiarization and test trials The procedure for the different-group condition was identical to that of the same-group condition with one exception: the actor who played Topid-B in the same-group condition now played Brinko-B throughout the experiment. In the first familiarization trial, she wore the same costume as BrinkoA and labeled herself as a Brinko (Figure 3). In familiarization trials 2-4, infants saw Topid-A establish her food preference, as in the same-group condition. In the test trial, the actor wore a Brinko costume, but her actions were otherwise identical to those she performed in the samegroup condition. The infants in both conditions thus saw the exact same actor in the test trial. All that differed was which costume she wore and whether she had previously labeled herself as a Brinko or a Topid. Any observed differences in looking times across conditions could therefore not be due to a preference for a particular individual.

Apparatus and procedure The infants sat on their parent's lap 91.5 cm in front of a large television screen $(68.5 \mathrm{~cm} \mathrm{x}$ 122 cm ). The room was dimly lit. A camera hidden at the base of the television (centered, 89 cm above the floor) recorded the infant's face during the experiment. Parents were instructed to close their eyes or look down to avoid biasing their infant's responses.

The television was connected to a Macintosh computer located to the left of the infant behind a sound-dampening
room divider. This computer controlled the presentation of the experimental stimuli using custom software written in Python (Peirce, 2007). The software selected the correct version of each trial based on the infant's condition and presented the video in the center of the television screen (each video measured $64 \mathrm{~cm} \times 37 \mathrm{~cm}$ on screen). The software also controlled the duration of each trial. An experimenter observed the infant on a monitor and pressed a button on the keyboard whenever the infant attended to the video. The software separately computed looking times for the fixed-duration and infant-controlled portions of each trial; looking times during the infant-controlled portion of the trial were used to determine when each trial ended. In between trials, an attention-getter (a yellow smiley face measuring $28 \mathrm{~cm} \times 20 \mathrm{~cm}$ ) was displayed on the screen for 4 seconds and a brief tone was played to attract the infant's attention back to the television screen.

At the start of the experiment, the attention-getter was presented in the center of the television screen. When the infant attended to the screen, the experimenter initiated the presentation of the stimuli on the television screen. The infants first viewed four familiarization trials appropriate for their condition. Each familiarization trial ended when the infant either (1) looked away for 2 consecutive seconds after having looked for at least 4 cumulative seconds or (2) looked for 60 cumulative seconds without looking away for at least 2 consecutive seconds.

Finally, the infants viewed the test trial that was appropriate for their condition; half the infants in each condition saw the same-food trial and half saw the differentfood trial. This trial ended when the infant either (1) looked away for .5 consecutive seconds after having looked for at least 4 cumulative seconds or (2) looked for 30 cumulative seconds without looking away for at least .5 consecutive seconds.

Coding and analysis In order to present events with trial duration contingent on the infant's attention, online coding was conducted by the experimenter (blind to condition and test trial), as described above. All infants were then coded offline from silent video by a trained coder who was naïve to the condition and test trial that the infant received; the looking times resulting from this coding were used in all analyses. For each trial, the coder indicated the infant's direction of gaze (at the stimuli or away) for each frame of the video. Another trained coder who was naïve to the infant's condition and test trial coded all sessions, and these two coder's agreed on the child's direction of gaze for $96 \%$ of video frames. Trials in which agreement was less than $90 \%(15 / 180)$ were resolved by a third naïve coder.

The infants were highly attentive during the initial phase of the familiarization trials: averaged across the four familiarization trials, the infants attended for $96 \%$ of the initial phase. The infants were also highly attentive during the initial phase of the test trial, attending for $97 \%$ of the initial phase.

Preliminary analyses of the test data indicated no
significant interactions of condition and event with sex or which food Topid-A preferred (blue vs. purple), all $F$ s $<$ 1.55 , all $p \mathrm{~s}>.22$. The data were therefore collapsed across these factors in subsequent analyses. In order to control for baseline differences in attention, all analyses were run with average looking time during the final phases of the familiarization trials as a covariate.


Figure 4: Mean looking time (sec) of the infants during the test trial of Experiments 1 and 2 as a function of condition and event. Error bars represent standard errors, and asterisks indicate a significant difference between events with a condition ( $p<.05$ ).

## Results and Discussion

The infants' looking times during the test trial (see Figure 4) were analyzed using an analysis of covariance (ANCOVA) with condition (same-group, different-group) and event (same-food, different-food) as between-subjects factors. There was a main effect of event, $F(1,31)=14.20, p=.001$, $\eta_{p}{ }^{2}=.31$, indicating that the infants who saw the differentfood event looked longer than those who saw the same-food event in the test trial. However, this effect was qualified by a significant interaction of condition and event, $F(1,31)=$ 9.03, $p=.005, \eta_{p}{ }^{2}=.23$. There was no main effect of condition, $F<1$. Planned simple effect comparisons revealed that in the same-group condition, the infants who received the different-food event $(M=17.18, S D=4.14)$ looked reliably longer than those who received the samefood event $(M=7.94, S D=2.29), F(1,31)=23.00, p<$ .001 , Cohen's $d=2.76$. In the different-group condition, the infants looked about equally whether they received the same-food event ( $M=10.81, S D=4.16$ ) or the differentfood event ( $M=12.15, S D=5.80$ ), $F<1$.

As predicted, the infants in the same-group condition looked reliably longer if they received the different-food event than if they received the same-food event. This suggests that the infants expected members of the same
social group to share food preferences, and they looked longer if members of the same social group had different food preferences. In contrast, the infants in the differentgroup condition looked equally regardless of whether members of different social groups picked the same or different foods.

However, a possible alternative explanation of these results is that when Topid-A and Topid-B wore the same outfit, the infants were unable to discriminate between them. If so, then the infants in the same-group condition might have thought that the agent in the test trial was the same agent that they had seen in the familiarization trials and hence looked longer at the different-food event because that agent appeared to suddenly change food preferences. To address this possibility, an additional group of 12 infants were tested in an actor-discrimination condition (procedure adapted from Buresh \& Woodward, 2007). Infants first saw the same familiarization trials as in the same-group condition. Infants then viewed two test trials in which either Topid-A (old-actor event) or Topid-B (new-actor event) entered and ate the same food that Topid-A had chosen during the familiarization trials (order of test events counterbalanced across infants). The infants' looking times during the test trials were analyzed using an analysis of variance (ANOVA) with test event (old-actor event, newactor event) as a within-subjects factor and event order (oldactor first, new-actor first) as a between-subjects factor. The analysis yielded a main effect of event, $F(1,10)=9.08, p=$ $.013, \eta_{p}{ }^{2}=.48$, indicating that the infants looked longer at the new-actor test event ( $M=11.44, S D=4.58$ ) than the old-actor event $(M=7.34 \mathrm{~S} D=2.24)$. No other effects were significant, all $F \mathrm{~s}<1$. If the infants had not noticed the change in actor, they would have looked equally to the newactor and old-actor events. However, the infants found the new-actor event novel, suggesting that infants were able to discriminate between the two actors.

Together, these findings provide additional evidence that infants can reason about members of a group that they themselves do not belong to (Powell \& Spelke, 2013; Liberman et al., 2016) and add to these prior findings by demonstrating that infants expect members of social groups to share inherent properties.

## Experiment 2

Experiment 2 had two goals. The primary goal was to replicate the positive findings from the same-group condition in Experiment 1 . Our secondary goal was to explore whether the noun labels in Experiment 1 were necessary for children to establish social groups with inductive potential. To investigate this question, infants were assigned to an adjective condition that was identical to the same-group condition of Experiment 1 except that the actors labeled themselves with adjectives instead of nouns. If infants require noun labels to identify social groups with inductive potential, then when the social groups are labeled with adjectives, infants will no longer generalize preferences across members of a social group. If, however,
infants can use adjectives to identify social groups with inductive potential, then infants will expect members of the same social category to pick the same foods, replicating the results of the same-group condition of Experiment 1.

## Method

Participants 30 healthy term infants participated (16 female; ages 18 months, 3 days to 21 months, 29 days, $M=$ 19 months, 16 days). Another 5 infants were tested but excluded because they were fussy (3), or because of parental interference (2). Eighteen infants were randomly assigned to the same-group condition ( $M=19$ months, 27 days) and 12 to the different-group condition ( $M=19$ months, 0 day).

Stimuli, Apparatus, and Procedure The stimuli, apparatus, and procedure were identical to Experiment 1 except that in the first familiarization trial the actors labeled themselves with adjectives instead of nouns (e.g., "Hi! I'm Topish.").

Coding and analysis As in Experiment 1, all infants were coded offline by a coder naïve to the condition and test trial that the infant received. An additional naïve coder coded all participants; agreement between the two coders was $97 \%$. Trials in which agreement was less than $90 \%(4 / 150)$ were resolved by a third naïve coder

Infants were highly attentive during the initial phase of all familiarization trials; averaged across all four trials, infants attended for $98 \%$ of the initial phase. Infants were also highly attentive during the initial phase of the test trial, attending for $98 \%$ of the initial phase.

Preliminary analyses of the test data indicated no significant interactions of condition and event with sex, or which food Topid-A preferred (blue vs. purple), all $F \mathrm{~s}<1$. The data were therefore collapsed across these factors in subsequent analyses.

## Results and Discussion

Infants' looking times during the test trial were analyzed using an ANCOVA with condition (same-group, differentgroup) and event (same-food, different-food) as betweensubjects factors, and infants' average looking times during the final phases of the familiarization trials as a covariate. Results revealed a significant interaction between condition and event $F(1,25)=4.75, p=.039, \eta_{p}{ }^{2}=.16$. There were no main effects of event or condition, both $F \mathrm{~s}<1$. Planned simple effect comparisons revealed that in the same-group condition, the infants who received the different-food event ( $M=19.90, S D=8.40$ ) looked reliably longer than those who received the same-food event ( $M=10.44, S D=3.73$ ), $F(1,25)=8.59, p=.007, d=1.46$. In the different-group condition, the infants looked about equally whether they received the same-food event $(M=14.54, S D=5.43)$ or the different-food event ( $M=12.38, S D=4.79$ ), $F<1$. These results suggest that, similar to the results of Experiment 1, infants in Experiment 2 expected members of a social group to share food preferences and looked longer when members
of the same social group had different preferences. In contrast, infants had no expectation about whether members of different social groups would share food preferences.

To investigate whether infants' looking time patterns were similar across experiments, infants' looking times to the test trial were analyzed using an ANCOVA with Experiment (1, 2), condition (same-group, different-group), and event (same-food, different-food) as between-subjects factors, and infants' average looking times during the final phases of the familiarization trials as a covariate. Results revealed a main effect of event, $F(1,57)=12.33, p<.001$, $\eta_{p}{ }^{2}=.18$. This effect was qualified by a significant interaction of condition and event, $F(1,57)=13.93, p=<$ $.001, \eta_{p}{ }^{2}=.20$. No other effects were significant, all $F \mathrm{~s}<$ $2.3, p \mathrm{~s}>.14$. The absence of any main effects or interactions involving Experiment suggests that regardless of whether the social groups were labeled with nouns or adjectives, infants expected members of the same social group to prefer the same foods.

## General Discussion

By preschool, children expect members of a social group to share characteristics and thus use social-group membership to draw inductive inferences about the properties of novel individuals. The current studies examined the origins of this social-group based reasoning in infancy. Specifically, we examined whether infants expect members of a social group to share preferences. Infants were introduced to members of two arbitrary social groups, Topids and Brinkos, and learned that a particular Topid preferred one of two foods. Infants later expected another Topid to prefer the same food and looked longer if she did not. However, infants had no expectations about whether members of different social groups (i.e. a Topid and a Brinko) would share preferences. Infants held similar expectations regardless of whether the group members labeled themselves with nouns (Topid, Brinko) or adjectives (Topish, Brinkish).

These findings expand our understanding of infants' social-group based reasoning in several key ways. First, these studies provide additional evidence that infants can categorize individuals as members of a social group, even if they themselves are not members of that group (e.g., Liberman et al., 2016; Powell \& Spelke, 2013). Second, these studies expand on prior work by providing the first empirical evidence that infants as young as 20 months use social-group membership to make inductive inferences about the likely behavior of group members, even when other group members are not present. In our experiments, only one agent was present in the test trial, and that agent did not see which food the other agent had selected during the familiarization trials. Thus, infants' expectations regarding the agent's behavior in the test trial are unlikely to have been based on social pressures or imitation. Even without the presence of other group members, infants expected members of a social group to share stable, inherent properties.

Together, the current studies begin to shed light on the
circumstances under which infants treat social group members as alike, and the age at which these expectations emerge. Future studies should further explore the precise nature of infants' social-group based inferences. For instance, we have discussed our results in terms of shared preferences: when Topid-A selects blue cereal over purple pasta, infants interpret this as signaling that Topids prefer blue cereal and hence expect another Topid to share this preference and also select blue cereal. However, perhaps infants were instead reasoning about shared avoidance of the food that Topid-A did not select - Topids do not eat purple pasta - and hence expected another Topid to avoid that food as well. Both shared preferences for and shared avoidance of specific foods exist across cultures (i.e. some cultural groups have a strong preferences for pork products, whereas other groups prohibit consuming pork). Future research should examine whether infants expect group members to like the same foods, avoid the same foods, or both. Additionally, future research should examine whether infants were reasoning specifically about foods, or whether infants could also have been reasoning about other features of the event (i.e. reaching for a particular color). Such studies will help clarify the characteristics that infants expect social-group members to share.

## References

Agerström, J., Björklund, F., Carlsson, R., \& Rooth, D. O. (2012). Warm and competent Hassan $=$ cold and incompetent Eric: A harsh equation of real-life hiring discrimination. Basic and Applied Social Psychology, 34, 359-366.
Bar-Haim, Y., Ziv, T., Lamy, D., \& Hodes, R. M. (2006). Nature and nurture in own-race face processing. Psychological Science, 17, 159-163.
Bigler, R. S., Jones, L. C., \& Lobliner, D. B. (1997). Social categorization and the formation of intergroup attitudes in children. Child Development, 68, 530-543.
Birnbaum, D., Deeb, I., Segall, G., Ben-Eliyahu, A., \& Diesendruck, G. (2010). The development of social essentialism: The case of Israeli children's inferences about Jews and Arabs. Child Development, 81, 757-777.
Buresh, J. S., \& Woodward, A. L. (2007). Infants track action goals within and across agents. Cognition, 104, 287-314.
Cashdan, E. (1998). Adaptiveness of food learning and food aversions in children. Social Science Information, 37, 613-632.
Diesendruck, G., \& HaLevi, H. (2006). The role of language, appearance, and culture in children's social category-based induction. Child Development, 77, 539553.

Egyed, K., Király, I., \& Gergely, G. (2013). Communicating shared knowledge in infancy. Psychological Science, 24, 1348-1353.
Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. Cognitive Psychology, 20, 65-95.

Howard, L. H., Henderson, A. M., Carrazza, C., \& Woodward, A. L. (2015). Infants' and young children's imitation of linguistic in-group and out-group informants. Child Development, 86, 259-275.
Kinzler, K.D., Dupoux, E., \& Spelke, E.S. (2007). The native language of social cognition. The Proceedings of the National Academy of Sciences of the United States of America, 104, 12577-12580.
Liberman, Z., Kinzler, K. D., \& Woodward, A. L. (2014). Friends or foes: Infants use shared evaluations to infer others' social relationships. Journal of Experimental Psychology: General, 143, 966-971.
Liberman, Z., Woodward, A. L., Sullivan, K. R., \& Kinzler, K. D. (2016). Early emerging system for reasoning about the social nature of food. Proceedings of the National Academy of Sciences, 113(34), 9480-9485.
Mahajan, N., \& Wynn, K. (2012). Origins of "us" versus "them": Prelinguistic infants prefer similar others. Cognition, 124, 227-233.
Martin, C. L., Eisenbud, L., \& Rose, H. (1995). Children's gender-based reasoning about toys. Child Development, 66, 1453-1471.
Medin, D., ojalehto, b., Waxman, S., \& Bang, M. (2015). Relations: Language, epistemologies, categories and concepts. In E. Margolis \& S. Laurence (Eds.), The Conceptual Mind: New Directions in the Study of Concepts (pp. 349-378). Cambridge: MIT Press.
Peirce, J. W. (2007). PsychoPy-psychophysics software in Python. Journal of Neuroscience Methods, 162(1), 8-13.
Powell, L. J., \& Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. Proceedings of the National Academy of Sciences, 110, E3965-E3972.
Rhodes, M., Hetherington, C., Brink, K., \& Wellman, H. M. (2015). Infants' use of social partnerships to predict behavior. Developmental Science, 18, 909-916.
Rozin, P., \& Siegal, M. (2003). Vegemite as a marker of national identity. Gastronomica, 3, 63-67.
Shutts, K., Kinzler, K.D., McKee, C., \& Spelke, E.S. (2009). Social information guides infants' selection of foods. Journal of Cognition and Development, 10, 1-17.
Waxman, S. R. (2013). Building a better bridge. In M. Banaji \& S. Gelman (Eds.), Navigating the Social World: What Infants, Children, and Other Species can Teach Us (pp. 292-296). Cambridge: Oxford University Press.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. Cognition, 69, 1-34.

# Thinking inside the box: Motion prediction in contained spaces uses simulation 

Kevin A Smith (k2smith@mit.edu), ${ }^{1}$ Filipe de A B Peres (filipe.peres@columbia.edu), ${ }^{2}$ Edward Vul (evul@ucsd.edu), ${ }^{3}$ Joshua B Tenenbaum (jbt@mit.edu) ${ }^{1}$<br>${ }^{1}$ Department of Brain and Cognitive Sciences, MIT, Cambridge, MA<br>${ }^{2}$ Columbia University, New York, NY<br>${ }^{3}$ Department of Psychology, University of California, San Diego, La Jolla, CA


#### Abstract

Theories of the mental processes people use to perform physical reasoning often differ on whether they are based on simulation or on logical reasoning. Here we test how these different processes might combine in a motion-prediction task that can be solved either by simulation or by reasoning about the topology of the scene. Participants were asked to predict which of two goals a computerized ball would reach first, but in some of these scenes the ball was 'contained' in the same space as one goal but was topologically separated from the other. Even in these contained scenes, participants responded faster when they received motion information that would speed up simulation but not affect topological parsing. This suggests that simulation contributes to predicting short-range motion, even when alternate strategies are available.


Keywords: intuitive physics; simulation; topology; containers

## Introduction

A long-standing debate in physical reasoning is whether people use simulation (Shepard \& Metzler, 1971; Battaglia, Hamrick, \& Tenenbaum, 2013) or symbolic reasoning (Forbus, 1983; Davis, Marcus, \& Chen, 2013) to understand the world, and whether the mental representations that support physical reasoning are continuous or based on a qualitative analysis of the scene. Recently, models based on continuous simulation have had success explaining a range of human behaviors such as stability judgments (Battaglia et al., 2013), motion prediction (Smith \& Vul, 2013), and causality judgments (Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2012). However, others have noted that simulation is an overly cumbersome process for many instances of physical reasoning where simple logical rules could suffice - e.g., if a ball is placed in a box and shaken, it seems easier to notice the topological relationship of containment and use the rule "An object in a closed container remains in the container" than to simulate the exact trajectory of the ball (Davis \& Marcus, 2015).

In this paper we study how people make physical predictions in cases where either simulation or a logical parsing of the scene can provide an answer. As in the example of Davis and Marcus (2015), we focus on the topological relationship of containment: an object is contained in a portion of space if there exist other objects that surround it and prevent it from leaving that space. We choose to study containment for three reasons. First, Davis et al. (2013) demonstrate that topological containment can be parsed using simple rules of firstorder logic for a rapid understanding of the scene. Second, people automatically and unconsciously process certain types


Figure 1: Diagram of continuous simulation (top) versus reasoning about kinematics using a qualitative scene parsing (bottom). Noisy simulation traces potential trajectories of objects through a continuous representation of the world. Qualitative physical reasoning segments the scene by topological regions and defines motion trajectories through a graph based on connections between those regions (bottom figure from Forbus, 1983).
of containment (Strickland \& Scholl, 2015). And finally, in previous work we found that we could explain motion prediction using a model of noisy physical simulation across a wide variety of scenes, but that people made predictions more rapidly than would be expected under the simulation model in the handful of scenes where an object was topologically contained to make one outcome impossible (Smith, Dechter, Tenenbaum, \& Vul, 2013). Together, this suggested that predicting motion in contained spaces would be a good candidate task for finding traces of both simulation and logical reasoning about scene topology.

Here we test whether and when topological reasoning about a scene occurs before simulation, versus when simulation supports prediction. We use a similar paradigm to Smith et al. (2013), in which participants observe a ball bouncing around a computerized screen and predict which of two 'goals' the ball will reach first. To make predictions using simulation, people would need to form a representation of the scene then step the motion of the ball forwards in time until it reaches a goal, but do not need to recognize the combined spatial relationship of objects: e.g., that a set of walls delineates one part of the space from another. Thus the time it takes to form simulations and make a prediction should be proportional to the path length the ball travels (Moulton \&

Kosslyn, 2009), and so the amount of time to produce a response using simulation should be affected by the motion of the ball: if the ball is moving towards a goal, the relatively shorter path should require less time to simulate, whereas motion away from the goal would produce longer simulation paths and proportionally longer response times. Conversely, to predict that the ball will reach one of the two goals because it cannot reach the other requires representing the scene, logically reasoning about whether the walls form distinct regions, then deciding whether the ball rests within a region with one but not both of the goals. However, this reasoning process does not require moving the ball forward in time, and so should be insensitive to any differences in motion information.

In this experiment, we therefore asked participants to make a single prediction and measured their response times. We further varied the motion information provided: the ball could move towards the goal, away from the goal, or have no observed motion. A facilitation effect in which motion towards the goal produces faster responses is evidence that people are at least in part relying on simulation, while absence of this effect points towards topological processing. Finally, we vary the way in which the ball can be contained in order to test for the limits of topological processing.

## Experiment

## Procedure

We recruited 100 participants from Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016) to take part in this experiment. The experiment lasted $\sim 10-12$ minutes, for which participants were compensated $\$ 1.20$.

On each trial of the experiment, participants would observe a scene with a ball that could move in a straight line but bounce off of walls (such as the one in Figure 2) and were asked to predict whether the ball would reach the green goal or the red goal first. The colors of the goals were randomly switched to avoid any color biases, and responses were adjusted in switched-color trials for consistency of analysis. In two thirds of the trials, participants would observe the ball in motion for 500 ms ; in the remaining third of trials, participants would observe no motion but were instructed that the ball would move in a direction not known to them until after they made their prediction.


Figure 2: Diagram of experimental trials (Left: non-topological, Right: topological). Participants would either observe the ball in motion or a static ball, and would be asked to indicate whether they believed the ball would reach the green or red goal first. The arrow was not displayed but indicates the direction of motion.

Participants registered which goal they believed the ball would reach by pushing either the ' $z$ ' or ' $m$ ' buttons on the keyboard. The mapping between the key and goal color was randomized across participants to control for any directional effects. To ensure that participants observed the full motion path and to equate for processing time, the prediction could not be registered until either the 500 ms of motion had stopped in the motion condition, or after the response buttons flashed after 500 ms in the no motion condition.

After participants registered their response, the ball would travel along its trajectory until it reached a goal, and participants would be assigned a score between 0 and 100 points, with faster reaction times (up to 300 ms ) earning more points. If participants made an incorrect prediction, they always lost 10 points. If a participant did not respond for 2500 ms , the trial would end and the participant would be awarded no points. These points were used to incentivize rapid responses and as motivation, but did not affect compensation.

On each trial we recorded both reaction time (between when a response could be indicated and when the button was pressed) ${ }^{1}$ and the goal the participant predicted the ball would reach.

## Stimuli

Participants observed 120 trials throughout the experiment. Of these trials, 96 were 'non-topological' trials that were randomly constructed such that the ball would reach a goal within 15 seconds, but were not hand designed with topological relationships. These uncontained trials were used to ensure that participants did not develop a deliberative, top-down strategy of judging topology.

The remaining 24 trials each participant saw were crafted to investigate one of four different dimensions of topological processing. There were six trial templates for each of the four dimensions (for a total of 24 templates), and each of these templates was adjusted to create three levels of containment across that dimension - each dimension started from the most contained, most simple, or smallest (level 1) and progressed to the most open, most complex, or largest (level 3). Thus there were 72 different topological trials used in the experiment, but each participant saw only one of the three levels formed from each template to avoid carry-over from similar trials. The dimensions of topological differences are described below.

Size This dimension was used to test whether topological parsing was performed by exploring the enclosed space at a constant rate, or whether topology is processed based on the configuration of the scene. If it is performed at a constant rate, then larger scenes with the same configuration should take longer to parse as topologically contained. We crafted these stimuli such that the smallest scene had dimensions that were $50 \%$ of the largest scene, while the middle scene had

[^502]

Figure 3: Sample trials from each of the topological dimensions along the three levels. The size trials (upper left) varied from small to large containers. The porousness trials (lower left) varied the size of gaps in the wall of the container. The stopper trials (upper right) varied the distance of the goal from making a seal with the rest of the container. The complexity trials (lower right) varied the internal and external structure of the containers. Balls in the "towards" condition moved in the direction indicated by the line, while balls in the "away" condition moved in the opposite direction. In all cases the ball ended at the green goal.
dimensions that were $75 \%$ as large (see Figure 3, upper left).
Porousness This dimension was used to test whether an object is topologically parsed as a container only if it is fully sealed, or whether containment is relative to the object inside. If topological processing is based solely on the container, then any gaps should prevent it from being seen as a container, whereas if it is calculated relative to the object inside, then it should still be observed as a container if gaps in its walls are smaller than the object it is holding. In the three levels along this dimension, one container was fully sealed with no breaks in its walls, one had gaps that were smaller than the ball, and the final level had gaps large enough for the ball to fit through (see Figure 3, lower left).
Stopper Forbus (1983) suggests that scene and motion descriptions also take into account what sorts of motions are allowable within the scene. If this is the case, then topological processing might also be affected by whether the path an object takes to exit a container is implausible or impossible. We therefore tested whether participants would consider the ball to be contained if there were almost no conceivable physical paths that would allow it to escape, even if there exist simple paths outside that do not account for physical motion. Level 1 along this dimension was produced with one goal forming a seal with the rest of the container. Level 2 moved the goal away from the container so that the ball could fit through, but in a physically implausible way. Level 3 moved the goal even further so that it is easily possible that the ball could exit the container without hitting the goal under plausible kinematic motion (see Figure 3, upper right).

Complexity This dimension was used to test how topology and simulation interact - even in situations where the
ball is fully contained will people use simulation if parsing the boundaries of the container is too difficult? Along this dimension, the levels included a simple configuration (e.g., the screen is split into two parts by a single wall), a moderately complex configuration with more internal and external structure, and very complex configuration (see Figure 3, lower right).

In all of the levels of all of the trials, participants would see either motion that is in the general direction of the goal within the container, motion in the opposite direction away from the goal, or no motion. To ensure each trial was novel, each participant only saw one type of motion for each trial, counterbalanced across participants. These motion conditions were tested because differences in velocity information should affect simulation but not topological judgments. If people are using simulation, we would expect that motion towards the goal should speed processing as compared to motion away, since each simulation will have a shorter distance to travel and fewer bounces before reaching the goal (Hamrick, Smith, Griffiths, \& Vul, 2015). Similarly, if people predict the outcome of no motion trials by simulating paths the ball could take in any direction, because most potential paths would be longer than the paths created with motion towards the goal, predictions should also be slowed in this case. If containment is judged by parsing the topology without using information about velocities, then changing the type of motion information provided should not change the speed of this mental process. We can therefore test for the presence of simulation by the presence of faster reaction times in the "towards" motion condition compared to the "away" or "no motion" conditions.


Figure 4: Geometric means of reaction times across topological dimensions and motion conditions. Bars indicate $95 \%$ confidence intervals bootstrapped from 500 samples. In all cases reaction times in the "towards" condition were faster than those in the other two conditions, indicating a use of simulation.

## Results

To ensure that we do not use data from participants who were not paying attention, we eliminated responses from trials where participants minimized or otherwise hid their browser screen ( $0.3 \%$ of trials) or where participants did not indicate a prediction in the allotted time ( $1.6 \%$ of trials). Finally, because participants observed the scene for 500 ms before making a response, some responses could be anticipatory. To prevent these measurements from skewing the data, we removed all responses under 10 ms ( $0.7 \%$ of all trials). For the purpose of all analyses, reaction times were log-transformed to account for long tails (Whelan, 2008) but transformed back for reporting and display.

We first test for overall differences in speed of processing across all topological dimensions, levels, and motion directions. ${ }^{2}$ This analysis suggests that motion direction plays a pivotal role in explaining reaction times $(F(2,2152)=$ $61.8, p \approx 0$ ), with "towards" ( $321 \mathrm{~ms}, 95 \% \mathrm{CI}$ : $[288,357]$ ) being faster than "away" ( $394 \mathrm{~ms}, 95 \%$ CI: [354, 439]), which in turn is faster than "no motion" ( $487 \mathrm{~ms}, 95 \%$ CI: [437, 543]) over all trial types.

The dimension of topology also affects reaction time $\left(F(3,58)=14.2, p=7.8 * 10^{-8}\right)$, with the complex trials ( $481 \mathrm{~ms}, 95 \%$ CI: $[424,546]$ ) being slower than the size trials ( $359 \mathrm{~ms}, 95 \% \mathrm{CI}:[316,407]$ ), the porous trials $(373 \mathrm{~ms}, 95 \%$ CI: [329, 423]), and the stopper trials ( $377 \mathrm{~ms}, 95 \% \mathrm{CI}$ : [333, 428]).

Finally, the level of topology had an overall effect $(F(2,58)=5.9, p=0.0046)$. Although the specific way in which trials changed with differences in level was not the same across topological dimensions, they were all ordered such that the first level was expected to produce the fastest predictions and the third the slowest. Here the simplest / most contained trials were the fastest ( $359 \mathrm{~ms}, 95 \% \mathrm{CI}$ : [319, 404]), followed by the intermediate trials (393ms, $95 \% \mathrm{CI}$ :

[^503][349, 443]), followed by the most complex / least contained (436ms, $95 \%$ CI: [387, 490]).

Nonetheless, there was no statistically reliable effect of any interaction (all Fs $<1.5$, all ps $>0.13$ ). This suggests that the amount of speed-up from observing motion does not change with the type of topological trial, which in turn suggests that simulation is used across all topological trials.

To more directly test for the use of simulation across dimensions of topology, we can compare how fast people respond in the "towards" condition as compared to the "away" and "no motion" conditions. We calculated a simulation facilitation index as the ratio of the reaction times in the "towards" condition versus the average of the other two conditions. If this index is less than one, then we have evidence that participants were using simulation to make predictions in that condition. As can be seen in Table 1, across every condition the simulation facilitation index is numerically less than one, and in most conditions ( 9 of 12) the $95 \%$ confidence intervals do not include one either.

These simulation facilitation effects are not driven by a small set of outlier trials. Across all topological trials, reaction times of participants in the "towards" condition were faster than those in the "away" condition in 52 of 72 trials (binomial test, $p=0.0002$ ), and faster than those in the "no motion" condition in 61 of 72 trials (binomial test, $p=1.6 * 10^{-9}$ ).

We also consider whether the facilitation in the "towards" condition is truly facilitation, or whether this effect is observed because "away" motion slows down processing: in some cases the ball was moving away from the correct goal and towards the incorrect goal, and simple directional motion towards a goal might speed reactions for that goal and slow reactions for the incongruent goal. However, participants were still faster in the "away" motion condition than the "no motion" condition both in average reaction time and in 49 of 72 trials (binomial test, $p=0.003$ ), so the differences in reaction time cannot be explained simply as a slowdown due to motion towards the incorrect goal.

Table 1: Simulation facilitation index for each of the topological dimensions and levels. Numbers in brackets indicate $95 \%$ confidence intervals. In all cases, there is a simulation facilitation advantage, and in all but three conditions the confidence intervals are below one. This suggests that simulation is used across all topological conditions, including the conditions with simple containers.

|  | Level 1 | Level 2 | Level 3 |
| ---: | :--- | :--- | :--- |
| Size | $0.764[0.593,0.983]$ | $0.82[0.635,1.06]$ | $0.716[0.556,0.921]$ |
| Porousness | $0.714[0.554,0.92]$ | $0.716[0.556,0.922]$ | $0.869[0.676,1.12]$ |
| Stopper | $0.592[0.46,0.761]$ | $0.704[0.548,0.904]$ | $0.681[0.529,0.877]$ |
| Complexity | $0.845[0.655,1.09]$ | $0.7[0.543,0.903]$ | $0.713[0.554,0.917]$ |

Although motion towards the goal speeds reaction times across all trial types, we can test whether participants were simply cued to respond faster in general when the ball is moving towards the goal, or whether there is evidence that these judgments are based on simulation. If people are using simulation, then as the ball travels further to reach the goal we would expect mental simulations to also travel a longer path and thus take more time to produce (Moulton \& Kosslyn, 2009). We therefore expect that reaction times should increase roughly in line with the time it takes the ball to actually reach the goal. ${ }^{3}$ As can be seen in Figure 5, there is a relationship between the actual travel time of the ball and participants' reaction time on that trial across all of the topological trials $(r=0.29, t(142)=3.5, p=0.00056)$, but we do not have evidence that this relationship differs between the "towards" and "away" conditions $(F(1,141)=0.76, p=0.38) .{ }^{4}$ This relationship suggests that simulation was in general used to produce motion predictions in this task regardless of the direction of motion.

Finally, we considered two alternate explanations that might give rise to this pattern of data by chance. First, if participants were 'guessing' more in the towards motion trials, we might expect them to respond faster but be less accurate. Second, if participants changed the speed with which they responded over time, this could be a potential confound in our analyses. However, neither of these alternate explanations hold.

If participants are using a different speed-accuracy tradeoff across motion types, we might expect that the reduction in speed is counterbalanced by higher accuracy. Among the topological trials, participants were numerically most accurate in the 'no motion' condition ( $89.0 \%$ ), followed by the 'towards' condition ( $86.7 \%$ ), then the 'away' condition ( $81.3 \%$ ). While there is an overall effect of motion direction on accuracy $\left(\chi^{2}(2)=20, p=4.5 * 10^{-5}\right)$, this is driven by the 'away' condition being less accurate than the other two (vs. 'towards', $z=2.6, p=0.024$; vs. 'no motion', $z=3.28, p=0.003$ ) rather than by a difference between the

[^504]

Figure 5: Comparison of the time it takes the ball to actually reach the end goal versus the geometric mean of participants' reaction time in that trial $(r=0.29)$. Topological trials are linked between 'towards' and 'away' motion with dashed lines.
'towards' and 'no motion' conditions $(z=0.74, p=0.74)$. Furthermore, a speed-accuracy trade-off cannot explain why people are both slower and less accurate in the 'away' condition than they are in the 'towards' motion condition. ${ }^{5}$

We also tested for changes in the speed of response throughout the experiment. There was a minuscule effect of trial order on response speed that was not statistically reliable (each additional trial was $0.08 \%$ slower, $95 \%$ confidence interval $=[-0.02 \%, 0.18 \%], F(1,151)=2.7, p=0.10)$. Therefore changes in response times over the experiment cannot explain the difference in reaction times across the different motion conditions.

## Discussion

In this study we tested for situations where reasoning about topological containment preempts physical simulation across a wide variety of trials where both topological relationships and simulations could be used. We found that participants were using simulation across all types of topological trials, including the most simple cases of containment.

But why do we find evidence of simulation when a simple topological analysis alone would suffice? We consider five

[^505]possibilities.
First, the simulation facilitation effect could arise from a mixture of individuals, some of whom use simulation and others who use topological processing. Due to the small number of topological trials each participant saw in this experiment, we cannot precisely measure whether each participant individually had a simulation facilitation effect, only that this effect is found on average. Further research is required to investigate individual differences in the use of these two processes.

Second, because the majority of the trials could not be solved with topological reasoning, if people must choose between either simulation or topological reasoning, simulation would be the more general choice. Thus if there are cognitive costs for switching between different processing strategies, participants might constantly use simulation. Future work will study whether people continue to use simulation even when it is not as frequently required.

Third, Davis and Marcus (2015) suggest that "simulation is effective for physical reasoning when the task is prediction, when complete information is available, ... and when the range of spatial or temporal scale involved is moderate" - exactly the conditions of this experiment. Perhaps simulation is automatically activated in tasks that fit this description but not in others, and we happened to use a task that relied on simulation. This might also imply that the "no motion" trials involved a separate, logic-based process as opposed to the motion trials with complete information. Indeed there is a numerical pattern in these results that would support this interpretation: in Figure 4 the "away" reaction times are always slightly slower than the "towards" reaction times, but the difference between "towards" and "no motion" is more variable across conditions. Although there was not statistical evidence for such a difference, this pattern would be consistent with people using a separate process that requires a longer and more variable amount of time in cases where no motion was observed.

Fourth, people may be using simulation to gain information about containment. Liang, Zhao, Zhu, and Zhu (2015) explain human ratings of how well one object will be contained by another by simulating how often the first object will stay inside the second when dropped into it. This might suggest that for simple tasks our perception of containment is statistical (one would not expect this object to ever leave the container) rather than logical (the topology of the container entails the object inside will not leave).

Finally, making predictions may involve multiple processes running in parallel, including both simulation and topological parsing. In many of the topological trials participants observed - especially the "towards" trials - the ball did not have to travel far to reach the goal. If both simulation and topological reasoning are active at the same time, these might be the cases where simulation provides an answer quickly and wins out over topological processing. In Figure 5, the relationship between the time the ball actually takes to reach the goal and reaction time becomes more vari-
able and flatter as the travel time takes longer. These longer trials might be cases where simulation fails to provide an answer before less continuous processes can, and so we do not see the same sort of relationship between path length and reaction time. Intriguingly, this relationship is reduced even for the non-topological trials that last this long, suggesting perhaps that simulation can only look a short time into the future, after which point we use more qualitative scene representations that could support either qualitative simulation or logical reasoning.

Although simulation appears to be active in simple tasks that require predicting the motion of objects, fully explaining human physical reasoning will require a better understanding of how simulation interacts and trades off with more qualitative methods of conceptualizing the world.

## Acknowledgments

We thank the three anonymous reviewers for their comments on this manuscript. This work was supported by the Center for Brains, Minds \& Machines, funded by NSF STC award CCF-1231216.

## References

Battaglia, P., Hamrick, J., \& Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45), 18327-18332.
Davis, E., \& Marcus, G. (2015). The scope and limits of simulation in cognitive models. arXiv preprint arXiv:1506.04956.
Davis, E., Marcus, G., \& Chen, A. (2013). Reasoning from radically incomplete information: The case of containers. In Proceedings of the Second Annual Conference on Advances in Cognitive Systems (p. 273-288).

Forbus, K. D. (1983). Qualitative reasoning about space and motion. In D. Gentner \& A. Stevens (Eds.), Mental models (p. 53-73). New Jersey: LEA Associates, Inc.
Gerstenberg, T., Goodman, N., Lagnado, D. A., \& Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In Proceedings of the 34th Annual Meeting of the Cognitive Science Societyroceedings of the 34th.
Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. Behavior research methods, 48(3), 829-842.
Hamrick, J., Smith, K. A., Griffiths, T. L., \& Vul, E. (2015). Think again? Optimal mental simulation tracks problem difficulty. In D. C. Noelle et al. (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Liang, W., Zhao, Y., Zhu, Y., \& Zhu, S.-C. (2015). Evaluating human cognition of containing relations with physical simulation. In D. C. Noelle et al. (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Moulton, S. T., \& Kosslyn, S. M. (2009). Imagining predictions: mental imagery as mental emulation. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1521), 1273-1280.
Shepard, R. N., \& Metzler, J. (1971). Mental rotation of threedimensional objects. Science, 171(3972), 701-703.
Smith, K. A., Dechter, E., Tenenbaum, J. B., \& Vul, E. (2013). Physical predictions over time. In M. Knauff, M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Smith, K. A., \& Vul, E. (2013). Sources of uncertainty in intuitive physics. Topics in Cognitive Science, 5(1), 185-199.
Strickland, B., \& Scholl, B. J. (2015). Visual perception involves eventtype representations: The case of containment versus occlusion. Journal of Experimental Psychology: General, 144(3), 570-580.
Whelan, R. (2008). Effective analysis of reaction time data. The Psychological Record, 58, 475-482.

# Promoting Spontaneous Analogical Transfer: The Role of Category Status 

Sean Snoddy (ssnoddy1@binghamton.edu)<br>Kenneth J. Kurtz (kkurtz@binghamton.edu)<br>Department of Psychology, Binghamton University (SUNY) Binghamton, NY 13902 USA


#### Abstract

Analogical comparison promotes spontaneous transfer by encouraging a more abstract representation that may be easier to retrieve. The category status hypothesis states that: if knowledge is represented as a relational category, it is easier to activate as a result of categorizing (as opposed to cue-based reminding). To investigate these two pathways to analogical transfer, participants were assigned to different study conditions: 1) standard comparison of two analogs; 2) standard comparison followed by a second comparison of two new analogs; or 3) a guided category-building task based on sequential summarization. Category-building showed a reliably higher rate of spontaneous transfer during an analogical problem solving task than standard comparison (numerically higher than double-comparison). Another experiment measured spontaneous remindings to cues on the basis of matching structure. Category-building showed a reliable advantage over both comparison conditions. This supports categorization as a novel pathway to spontaneous transfer by enhancing retrieval of structurally similar information.


Keywords: concepts and categories; analogy; problem solving; comparison; transfer

## General Introduction

People are able to transfer prior knowledge to solve problems in a superficially dissimilar context. Gick and Holyoak (1980) demonstrated that individuals who encoded a base passage, which described how a general captured a fortress by dividing an army into small groups of soldiers that simultaneously attacked the fortress from various angles, were able to transfer the passage's solution to solve an isomorphic target problem about how a doctor could destroy a tumor with a ray of radiation. The prevailing cognitive account (henceforth abstraction account) explains knowledge transfer across domains (e.g., military strategy to medical treatment) in terms of analogy. According to the abstraction account, transfer involves a mapping process where distinct superficial information between analogs is filtered out (e.g., general and doctor), and similar relationships between analogs are placed into correspondence (e.g., simultaneous application) (Gentner, 1983; Gick \& Holyoak, 1980). This mapping process allows for candidate inferences from the structure of the base to fill in missing predicates of the target problem, which allows for a solution to be devised (e.g., lower intensity rays simultaneously converging on a tumor from multiple locations) (Gentner, 1983; Gick \& Holyoak, 1980).

Despite the capacity for analogical transfer, individuals often fail to spontaneously transfer knowledge from a single
base analog to solve a problem in a different domain without an explicit hint about the base's relevance (Gick \& Holyoak, 1980). This failure is known as the paradox of similarity-based retrieval: superficially similar information to the target problem (e.g., other medical problems) is favored during retrieval, even though inferences require structural overlap between the base and target (e.g., the concept of convergence) (Gentner, Ratterman \& Forbus, 1993; Holyoak \& Koh, 1987; Ross, 1987). With this paradox in mind, the key to understanding how spontaneous transfer occurs is to determine what promotes structurebased retrieval.

Some key findings of the abstraction account are that comparison is an effective way to learn (Alfieri, NokesMalach \& Schunn, 2013) and to promote spontaneous transfer (Gentner, Loewenstein \& Thompson, 2003; Gick \& Holyoak, 1983). During comparison, cases with matching structure are presented side-by-side and participants are prompted to consider the similarities between them. This facilitates a mapping process to occur during encoding that is similar to the one that occurs during transfer, which highlights commonalities between cases and promotes the formation of an abstract schema via filtering out surfacelevel mismatches (Markman \& Gentner, 2000). These abstract schemas are more accessible in memory than representations of specific cases due to a lack of superficial mismatches with targets (Forbus, Gentner \& Law, 1995; Gick \& Holyoak, 1983). The heightened accessibility of abstract schemas facilitates retrieval of structurally similar matches during the memory search triggered by an opportunity to spontaneously transfer knowledge.

While the abstraction account focuses on the role of schema abstraction in spontaneous transfer, the type of materials used in these studies can also be viewed as embodying relational categories (Gentner \& Kurtz, 2005). Categorization may provide another pathway to structurebased retrieval. The category status hypothesis predicts that when knowledge is represented in the form of a category, it is fluidly accessed and applied (Kurtz \& Honke, submitted). There are three important aspects that contribute to the development of category status. First, category intension is knowledge of the category defining structure. This may be similar to an abstract schema, and could be conferred through the comparison process (Goldwater \& Schalk, 2016). Second, category extension is knowledge of specific members and non-members of a category as well as how to differentiate between them. Third, categorization confers experience in bi-directional mapping between generic knowledge of the category and specific cases. Thus, the key
claim of the category status hypothesis is as follows: when knowledge is represented as a psychological category, construal of a target stimulus as a category member facilitates direct activation of the category-defining concept in semantic memory. The key difference is that the category status hypothesis predicts direct activation of structurally relevant matches as opposed to a filtering out of superficial mismatches followed by an evaluation for any structural similarity.

Initial support for the category status hypothesis comes from Kurtz and Honke (submitted) who had participants learn a relational principle either through category construction or a single comparison opportunity. In the category construction task, participants formed two categories out of three examples of a principle and three alignably different examples. This was compared to the standard version of the comparison task that received two cases presented side-by-side and provided a similarity rating as well as an explanation of the similarities between cases. Category construction led to a higher rate of spontaneous transfer than the comparison task, which suggested that promoting category status may provide a novel pathway to spontaneous transfer (Kurtz \& Honke, submitted).

The goals of the present work are: 1) to further evaluate the plausibility of the category status hypothesis as an alternative account of spontaneous transfer, and 2) to further explore the abstraction account. To address the first goal, a novel category building task based on the sequential summarization of cases was used to promote category status. This task is often used as a control to comparison, and does not confer the same level of abstraction-based transfer benefits (Catrambone \& Holyoak, 1989; Gentner et al., 2003; Rittle-Johnson \& Star, 2007). If sequential summarization can be combined with additional supports that promote category status, then it should become an effective way to promote transfer. The categorization supports that were integrated with the core summarization task were: 1) summarization of category-membershiprelevant aspects of multiple cases, 2) identification of each case with a shared category label, and 3) a description of the category after encountering all cases. If these supports contribute to the development of category status, a categorization-based summarization task should become an effective way to promote transfer.

The second goal of the present work sought to further understand the abstraction account. Prior analogical comparison research has largely focused on the effects of a single comparison opportunity on transfer success. The effect of an additional comparison opportunity was explored, which should improve schema abstraction by providing additional surface-level mismatches to filter out. The additional comparison opportunity also serves as a control for case exposure in the category-building condition.

## Experiment 1

Experiment 1 used the analogical transfer paradigm (Gick \& Holyoak, 1980, 1983) to assess the impact of categorization-
based summarization (category-building), the standard version of the comparison task (single comparison) (cf. Catrambone \& Holyoak, 1989; Gentner et al., 2003; Gick \& Holyoak, 1983), a standard comparison task that is repeated a second time with novel cases (double comparison), and a baseline condition on spontaneous transfer performance. Spontaneous transfer success is contingent upon being able to both spontaneously access and retrieve relevant knowledge from memory as well as apply that knowledge to devise a solution (Gick \& Holyoak, 1983). The use of both spontaneous and hint-aided transfer assessments allows for the differentiation of the relative impact that each study condition has on application ability and retrieval.

Both the abstraction account and category status hypothesis make different predictions about what type of task will improve the retrieval process that underlies successful spontaneous transfer. The predictions for spontaneous transfer are as follows: 1) all study conditions will promote transfer (i.e., result in a higher rate of transfer than baseline), 2) the category-building condition will result in a higher rate of transfer than both comparison conditions, and 3) double comparison will result in a higher rate of transfer than single comparison. Neither account makes explicit predictions about application ability, so hint-aided transfer performance is exploratory.

## Method

Participants A total of 355 undergraduate students from Binghamton University participated for course credit. Hintaided transfer data from seven participants were excluded due to a failure to complete the assessment in the allotted time.

Materials and Design The materials consisted of both study cases and the transfer problem. All materials demonstrated the principle "problem-as-a-solution": when a large-scale event causes a large amount of damage, the event can be mitigated by repeatedly causing it on a small scale and incurring minor damage each time. The study cases were all from the domain of natural disasters. The passages were a single paragraph that consisted of a description of the problem followed by the solution.

The transfer problem involved the prevention of cybercrime. It contained a similar description of the problem as the study cases. However, the solution was replaced with an open-ended question about how the threat of computer hackers could be minimized. The general formatting of the transfer problem was different from the study tasks to make the separate phases appear unrelated. The transfer problem was presented twice. It was first presented under the guise of a new experiment about problem solving (spontaneous transfer). The same problem was presented again with a hint for participants to use their knowledge from the study phase (hint-aided transfer).

Procedure Prior to the experiment, all participants were informed that they would take part in multiple experiments,
then were randomly assigned to one of the study tasks. Participants in the single comparison condition received two study cases, which were referred to as solved problems, presented side-by-side. Participants were informed that there were important ways that the solved problems were alike and were asked to consider the similarities between them. After reading the cases, participants provided a similarity rating on a five-point scale that ranged from 'not at all similar' to 'very similar'. Participants were then asked to describe the similarities and differences they considered when making their similarity judgement. The doublecomparison task was similar to single comparison, with the only difference being a second comparison opportunity for two novel cases. During this second comparison opportunity, participants were asked to consider the important ways in which all four of the solved problems were alike, provide a similarity rating for the two additional cases, and describe the similarities used to create the rating.

In the category-building condition, participants were given instructions that explicitly stated the task involved learning about a category, and then were sequentially presented with all four cases. Upon presentation of each case, participants were provided with the label to identify the case as a member of the category (e.g., "Here is an example of the Tongo category") and were asked to summarize the relevant information to category membership. After completion of the final summarization task, participants were asked to provide a description of the category they had just learned.

Following the study task, participants were told that they were beginning a new experiment on problem solving, and proceeded to the spontaneous transfer assessment. Since the baseline condition was meant to establish chance production of transfer solutions to the problem, participants in that condition only received this assessment. After completion of the spontaneous transfer assessment, participants were then given another chance to solve the problem with an explicit hint to use their knowledge from the study phase. Both of these assessments presented participants with the transfer problem, and asked them to devise a solution. For each transfer opportunity, participants were allowed to provide multiple solutions to the problem. Each of the proposed solutions was scored by the first author blind to condition. If at least one of a participant's proposed solutions demonstrated the target principle, that participant was coded as a transfer success.

## Results and Discussion

Hint-aided Transfer Hint-aided transfer performance reflects participants' ability to retain and apply the knowledge from the learning tasks. To evaluate hint-aided transfer performance, a logistic regression model (R Core Team, 2016) was built to predict hint-aided success with condition. There were no significant differences between category-building and single comparison $(\beta=0.274, S E=$ 0.313, Wald $Z=0.872, p=.383$ ) or double-comparison $(\beta=$ $-0.174, S E=0.318$, Wald $Z=-0.547, p=.585)$.

Additionally, there were no significant differences between double-comparison and single-comparison $(\beta=0.447, S E=$ 0.313 , Wald $Z=1.426, p=0.154$ ) (see Table 1).

Table 1: Hint-aided transfer performance.

| Learning Task | \% Transfer (N) |  | 95\% C.I. | N |
| :--- | :---: | :---: | :---: | :---: |
| Single-Comparison | $55 \%$ | $(47)$ | $45 \%-66 \%$ | 85 |
| Double-Comparison | $66 \%$ | $(58)$ | $55 \%-75 \%$ | 88 |
| Category-Building | $62 \%$ | (52) | $51 \%-72 \%$ | 84 |

Spontaneous Transfer Spontaneous transfer success was modelled using a logistic regression with condition as the predictor. Both the category-building ( $\beta=1.68, \mathrm{SE}=0.433$, Wald $Z=3.88, p<.001$ ) and double-comparison ( $\beta=1.1$, $S E=0.449$, Wald $Z=2.453, p=.014$ ) conditions resulted in a higher rate of transfer than baseline. However, the single comparison condition was not significantly different from baseline transfer performance $(\beta=0.407, S E=0.491$, Wald $Z=0.828, p=.408$ ) (see Table 2).

The category-building condition led to a significantly higher rate of spontaneous transfer than single comparison ( $\beta=1.274, S E=0.393$, Wald $Z=3.239, p<.01$ ). However, there was no significant difference between categorybuilding and double comparison ( $\beta=0.579, S E=0.339$, Wald $Z=0.171, p=.088$ ). There was no significant difference between double and single comparison ( $\beta=$ $0.695, S E=0.411$, Wald $Z=1.692, p=.091$ ) (see Table 2). To account for slight numeric differences in hint-aided transfer, a more conservative analysis was done that included only participants with hint-aided transfer success to clearly reflect differences in the retrieval process. The same pattern of results was observed.

Table 2: Spontaneous transfer performance.

| Learning Task | \% Transfer (N) |  | 95\% C.I. | N |
| :--- | ---: | :--- | :---: | :---: |
| Baseline | $9 \%$ | $(8)$ | $4 \%-17 \%$ | 91 |
| Single-Comparison | $13 \%$ | $(11)$ | $7 \%-21 \%$ | 87 |
| Double-Comparison | $22 \%$ | $(20)$ | $15 \%-32 \%$ | 89 |
| Category-Building | $34 \%$ | $(30)$ | $25 \%-45 \%$ | 88 |

Contrary to prior research, which demonstrated summarization was a less effective way to promote transfer than comparison (Catrambone \& Holyoak, 1989; Gentner et al., 2003), the present work found that combining summarization with categorization supports (categorybuilding condition) is an effective way to promote transfer. The category-building led to a higher rate of spontaneous transfer than single comparison, which provides support for the category status hypothesis as a viable account of transfer. The lack of a significant difference between category-building and double comparison provides limited support, since categorization supports in the categorybuilding task led to higher performance than is typically expected of a summarization task (i.e., it was not
significantly lower than double comparison). Further, the lack of differences on the hint-aided transfer assessment suggests that the category-building task's spontaneous transfer advantage cannot be attributed to a differential ability to apply knowledge, but instead results from improved structure-based retrieval. These findings support the category status hypothesis as an alternative account of transfer.

The interpretation of the comparison conditions is less clear. Prior work has demonstrated that single comparison is an effective way to promote transfer (Catrambone \& Holyoak, 1989; Gick \& Holyoak, 1983), so the lack of an advantage over baseline is puzzling. The present study used a novel stimulus set, and the observed transfer performance is appreciably lower than has been reported with the convergence materials (Catrambone \& Holyoak, 1989; Gick \& Holyoak, 1983). This stimulus set may be more difficult than commonly used materials, and a single comparison opportunity might require additional support to remain effective under more difficult circumstances.

There was no significant difference between single and double comparison on spontaneous transfer performance, which suggests that an additional comparison opportunity might not appreciably improve abstraction. If the present principle is more difficult than previous materials, an additional comparison opportunity may also be lacking in the support needed to remain effective. While singlecomparison did not promote transfer above baseline levels of performance, double-comparison did. This suggests that there may be some small benefit to engaging in the second comparison opportunity.

## Experiment 2

Experiment 2 was conducted to conceptually replicate the main findings and clarify some of the outstanding questions of Experiment 1. The same study conditions were used with the exception that a baseline condition was not included. Instead of a problem-solving assessment, a spontaneous reminding task was used. Participants were given a series of cue passages that were superficially distinct from, but contained matching structure with the study materials, and were asked what each cue reminds them of. Since participants were given completed cases as a cue, the entire relational structure of the principle guides the memory search. This is in contrast to the problem-solving transfer assessment that provides only the problem statement as a cue to initiate the memory search. Under less demanding retrieval circumstances, a double-comparison advantage might be accrued. Other modifications were made in an attempt to support the comparison conditions. Both problem-as-a-solution and convergence (Gick \& Holyoak, 1983) materials were used to test if the problem-as-asolution principle was more difficult to retrieve than convergence, since the difficulty of the principle in the first experiment may have been a barrier to comparison success. The instructions for the comparison conditions were modified to increase the symbolic juxtaposition - invitation
to compare through shared labels (Gentner, 2005) - of the cases in another attempt to enhance the comparison task.

The main prediction was that category-building will lead to more structure-based remindings than either comparison condition. Since the procedure was made less demanding in an attempt to promote comparison performance, doublecomparison was predicted to have a higher rate of reminding success than single-comparison. Given the overall low rate of transfer in the first experiment, it was predicted that cues for the convergence principle will result in a higher rate of successful remindings than the cues for the problem-as-asolution principle.

## Method

Participants A total of 104 undergraduate students from Binghamton University participated for course credit. Data from three participants were excluded due to a failure to complete the experiment in the allotted time, and another participant was excluded for failing to follow instructions.

Materials and Design The study materials consisted of both the problem-as-a-solution and convergence (Gick \& Holyoak, 1983) principles. The study cases for problem-as-a-solution were the same as in Experiment 1. The convergence cases used were as follows: The General, The Commander, Red Adair, and The Fire Chief (Gick \& Holyoak, 1983). These cases were rewritten to be comparable in length and grammatical structure to the problem-as-a-solution materials. The order of principles remained constant across participants; problem-as-asolution occurred first and convergence occurred second.

The reminding assessment consisted of six cue cases. Two cues were used that demonstrated the problem-as-asolution principle from Experiment 1. The transfer problem used in Experiment 1 was rewritten to include the solution and the other cue involved police infiltrating black markets. The Radiation Problem (Gick \& Holyoak, 1983) and The Aquarium (Catrambone \& Holyoak, 1989) were rewritten as reminding cues for the convergence principle. Two distractor cases - The Wine Merchant (Gick \& Holyoak, 1980) and The Birthday Party (Gick \& Holyoak, 1983) were also included in the reminding assessment in an attempt to disguise the true purpose of the assessment. The order of the cues was constant across participants: distractor, problem-as-a-solution, convergence, distractor, problem-as-a-solution, convergence.

Procedure The study task procedure was similar to the first experiment, with only a few differences. Participants in both comparison conditions received the same task from the first experiment, then repeated it a second time for the convergence materials. In addition, the principles were referred to as separate 'series' to connote that they reflected different principles. In double comparison, the instructions were modified to clearly connote that the first four passages shared important commonalties, and the second four passages also shared important commonalities. The
category-building condition repeated the task for the convergence principle after completion of the task for the problem-as-a-solution principle. The only other difference in the category-building condition was that the category label was replaced with "Conaway Scenario" for the first principle and "Rummel Scenario" for the second principle to clearly reflect the change in principles.

After the study phase all participants were given the reminding packet, which was introduced as a new experiment. Participants were told they would be shown a set of passages, and were supposed to write down anything that each passage reminded them of in as much detail as possible. Participants were then presented with each reminding cue sequentially and made their response. Reminding performance was scored by the first author and an undergraduate research assistant. A successful reminding on the basis of shared structure met at least one of the three following criteria: 1) used the category label or referred to solved problems, 2) referenced one of the cases from the study task, or 3) described the principle from the study task. Any remindings of another cue from within the assessment were considered non-scoring. Both raters agreed on scores for $99.8 \%$ of the reminding responses, all disagreements were resolved through discussion.

## Results and Discussion



Figure 1: Proportion of structural remindings to target cues by study task and principle. Error bars reflect 95\% binomial confidence intervals (Dorai-Raj, 2014).

Reminding performance was modelled trial-wise via a mixed-effects logistic regression (Bates et al., 2015) with the main effect of interest as a predictor and participant included as a random intercept. The predictions concern participants' responses to only the target cues that had shared structure with the study materials, so only those cues are considered. Category-building led to a higher rate of successful remindings than double-comparison ( $\beta=3.436$, $S E=1.671$, Wald $Z=2.057, p=.0399$ ) and singlecomparison ( $\beta=6.178, S E=1.649$, Wald $Z=3.746, p<$ .001). Double-comparison led to a significantly higher rate of successful remindings than single comparison $(\beta=2.741$,
$S E=1.307$, Wald $Z=2.1, p=.036$ ). Collapsing across condition, convergence cues led to a significantly higher rate of successful remindings than problem-as-a-solution cues $(\beta=1.5, S E=.622$, Wald $Z=2.41, p=.016$ ) (see Figure 1).

The category-building task led to a higher rate of structurally based remindings to target cues than either comparison condition. This suggests that category-building promotes the spontaneous access and retrieval of relevant structural matches from memory, and that this is driving the spontaneous transfer differences observed in the first experiment. These results provide further support for the category status hypothesis and a successful replication of the main finding in Experiment 1.

These results also address some of the outstanding questions from the first experiment. In contrast to the previous findings, double comparison had a significantly higher rate of structural remindings to target cues than a single comparison opportunity. This suggests that additional comparison opportunities can enhance the retrieval of structurally relevant information from memory. However, we cannot identify which changes were responsible for the observed improvements. Additionally, convergence cues led to a higher rate of successful structural remindings than problem-as-a-solution cues, which may suggest that the convergence materials result in higher rates of transfer than other materials. However, the convergence study materials were always presented after problem-as-a-solution, so future work should explore if this is the result of a practice effect.

## General Discussion

When a sequential summarization task was given additional categorization-based supports (the category-building condition), it led to better spontaneous transfer performance than the standard version of the comparison task (single comparison), but did not significantly differ from a task that controlled for case exposure (double comparison). The second experiment replicated the advantage of categorybuilding over single comparison, and found that categorybuilding led to a higher rate of structure-based remindings than double comparison. This supports the conclusion that the spontaneous analogical transfer gains in the categorybuilding condition were due to an increase in retrieval based on matching structure. Taken together, both experiments provide additional support for the category status hypothesis as a viable account of spontaneous transfer.

It is possible that the use of category labels in the category-building condition might confer symbolic juxtaposition (Gentner, 2005), which may allow for abstraction to occur in the absence of the temporal and spatial juxtaposition that is present during comparison. However, this explanation is unlikely the sole factor driving the results. The comparison conditions referred to cases as 'solved problems' to control for the use of a category label, and referred to cases from each principle as a coherent 'series'. The comparison conditions' controls for the presence of a label likely conferred some degree of
symbolic juxtaposition. If category-building benefits were due to symbolic juxtaposition promoting abstraction, it seems unlikely that a condition which has temporal, spatial, and symbolic juxtaposition (comparison conditions) would perform significantly worse.

The benefits of an additional comparison opportunity are less clear. During the analogical transfer assessment, no advantage of an additional comparison opportunity was accrued. However, in the reminding assessment, an extra comparison opportunity led to a higher rate of retrieval on the basis of shared structure. The benefits of an additional comparison opportunity may occur only under less demanding circumstances, such as being given the full structure as a retrieval cue and not needing to apply the knowledge to solve a problem.

The category status hypothesis can provide an alternative perspective about the success of double-comparison in the second experiment that is not mutually exclusive with the abstraction account. The second instance of comparison provides a chance to build extensional knowledge of the category, since participants are told that the cases are related and participants are given a chance to test hypotheses about why. Further, since some abstraction has likely occurred during the first comparison, the second comparison may afford the opportunity for a bi-directional mapping between generic knowledge of the principle and concrete knowledge of the cases. Future work should further explore the conditions required for additional comparison opportunities to promote spontaneous transfer as well as the role of the category status hypothesis in improving comparison.

There are two possibilities about why category-building and double-comparison led to better retrieval of structural information. First, promoting category status could engage the use of a different type of retrieval process. This retrieval process might occur through the mechanism of categorization as opposed to cue-based reminding. Alternatively, conferring category status might enhance or alter the cue-based reminding process described in Forbus et al. (1995). Future work should try to uncover the mechanism by which these two pathways to improved structure-based retrieval operate.

## Acknowledgments

We would like to thank Daria E. Slaby and Ian T. Goodman for their help scoring data, Garrett Honke for creation of the problem-as-a-solution stimulus set, and members of the Learning and Representation in Cognition lab.

## References

Alfieri, L., Nokes-Malach, T. J., \& Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. Educational Psychologist, 48, 87-113.
Bates, D., Maechler, M., Bolker, B., Walker, S., (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1-48.
Catrambone, R., \& Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1147-1156.
Dorai-Raj, S. (2014). binom: Binomial confidence intervals for several parameterizations. R package version 1.1-1. https://CRAN.R-project.org/package=binom
Forbus, K. D., Gentner, D., \& Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. Cognitive Science, 19, 141-205.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe \& D. H. Rakison, (Eds.), Building object categories in developmental time. (pp. 245-275). Hillsdale, NJ: Erlbaum.
Gentner, D., \& Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. Goldstone, B. Love, A. Markman, \& P. Wolff (Eds.), Categorization inside and outside the laboratory (pp.151-175). Washington, DC: American Psychological Association.
Gentner, D., Loewenstein, J., \& Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95, 393408.

Gentner, D., Rattermann, M. J., \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive Psychology, 25, 524-575.
Gick, M. L., \& Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychology, 12, 306-355.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.
Goldwater, M. B., \& Schalk, L. (2016, March 7). Relational categories as a bridge between cognitive and educational research. Psychological Bulletin, 142, 729-757.
Holyoak, K. J., \& Koh, K. (1987) Surface and structural similarity in analogical transfer. Memory and Cognition, 15, 332-340.
Kurtz, K. J., \& Honke, G. (2017). Sorting out the problem of inert knowledge: Using category construction to promote spontaneous transfer. Manuscript submitted for publication.
Markman, A.B., \& Gentner, D. (2002) Structure-mapping in the comparison process. American Journal of Psychology, 113, 501-538.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rittle-Johnson, B., \& Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. Journal of Educational Psychology, 99, 561574.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 629-639.

# Geometry-based Affordances 

Sterling Somers (sterling@sterlingsomers.com)<br>Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive<br>Ottawa, On., Canada


#### Abstract

A representational approach to ecological psychology is presented. This paper identifies a computational-level commonality in ecological psychology research related to passability of apertures. It is argued that a cognitive mechanism capable of comparing the geometric properties of an environment and the geometric properties of the agent can be used to support judgments for action in space.


Keywords: affordances; ecological psychology; spatial representation.

## Introduction

Two of the most central proposals in ecological psychology are the concept of affordances and the theory of direct perception. Gibson describes an affordance as properties that objects offer to animals that have the capacity to perceive it (Gibson, 1986). This position, shared with Michaels and Carello (Michaels \& Carello, 1981), maintains that the semantics of an action, by which I mean how an agent knows what actions can be performed given the objects in its environment, are properties of that object. Because the action semantics are encoded in the environment, they are claimed to be directly perceived.

Chemero \& Turvey (2007) divide ecological psychologists into two camps: Gibsonian and representationalist. Gibsonians maintain that affordances are directly perceived, while representationalists (e.g. Vera \& Simon, 1993) maintain that affordances (the actions an object affords) are inferred. This paper presents a representationalist position that is inspired by affordance research. However, the representations proposed in both theory and model are non-static and do not include semantically-laden representations of the environment. An example of a semantically-laden representation, with respect to action, is to label a feature of the environment as a 'doorway', such that doorways are features of the environment that can be passed-through.

The representationalist approach presented here is not necessarily inconsistent with a Gibsonian approach. The aim of the theory presented is to leverage the appropriate framework to make use of the computational cognitive architecture, ACT-R (Anderson \& Lebiere, 1998), in order to identify a plausible set of information processing steps involved in an aperture-passage affordance.

## Gibsonian Positions

There are two main theoretical positions in favor of direct perception. The first, Gibson's own theory, has already been described above. Although I will provide no argument against this position here, I agree with Chemero (2003), that

Gibson's position represents a non-standard ontology, in which the environment is not simply made up of physical properties but also semantic properties. I will assume that this alternative ontology is sufficient to reject Gibson's position for the purposes of this paper. For an argument against Gibson's position and related affordance-as-property positions, I refer the reader to Chemero (2003).

The second Gibsonian position is that action semantics are emergent properties which arise from the interaction between an animal and its environment (Chemero, 2003; Chemero \& Turvey, 2007; Stoffregen, 2003). While Chemero (2003) differs slightly in the terms he uses (relations instead of properties, to avoid certain philosophical problems), neither author's version addresses how the emergent properties or semantic-laden relations arise.

The theory presented here, the theory of geometric affordances, is inspired by research on aperture passage but also attempts to be commensurate with traditional representationalist views popular in the cognitive sciences. Briefly, this paper proposes that one of the mechanisms which can inform action properties (such as passability) is a geometric comparison between the features of the environment and current or possible future geometric properties of the body.

The aim of this paper is to illustrate, by way of example, how a represenationalist approach, which posits cognitive mechanisms, leaves open the possibility to develop unifying theories about different experimental findings within the affordance literature. The research reviewed in the following section is ecological psychology research regarding the affordance of passability of an aperture. The purpose of this brief review is to illustrate how a representationalist approach can posit a cognitive mechanism that compares the geometric properties of an environment with the geometric properties of the agent. I term these affordances, geometry-based affordances. I maintain that geometry-based affordances are only one class of possibly many types of affordances.

## Aperture Passability Research

Research into the passability of apertures, such as doorways, has shown that there is a body-size/aperture-width ratio at which apertures are judged to be passable (Fath \& Fajen, 2011; Higuchi, Seya, \& Imanaka, 2012; Wagman \& Taylor, 2005; Warren \& Whang, 1987). With different degrees of commitment, the central aim of that research is to show that a passability affordance can be directly perceived.

In Warren and Whang (1987), for example, they show that people judge apertures as passable only when the ratio
between the aperture width and body width is greater than 1 . Warren and Whang argue that we perceive the width of apertures in units of body width. The theory they propose is that we see the width of the doorway in units of eye-height. Since eye-height is in a constant proportion to shoulder width, we are effectively perceiving in units of shoulderwidth. Instead of seeing a doorway, estimating its width, estimating body width, and making a determination regarding passability; we simply perceive apertures as either passable or impassible.

In a related study by Fath and Fajen (2011), participants view simulated environments while wearing a headset. In a set of experiments, Fath and Fajen modified the visual properties available to the participants. For example, they eliminate the ground plane, making the estimate of eyeheight implausible. They argue that the visual properties related to body-width-scaled units are not the only properties that can be used to make passability judgments. They propose that visual information related to head-sway and stride length (both while walking towards the aperture) can be calibrated to body-width and used in lieu of eyeheight, to directly perceive passability.

Other studies such as those by Higuchi, Seya, and Imanaka (2012) and Wagman and Taylor (2005) have participants holding objects. Chang, Wade, and Stoffergen (2009), furthermore, studied passability for people grouped in dyads. Higuchi, Takada, and Matsuur ( 2004), finally, studied passability for novel wheelchair users.

When taken together, it is not clear whether a direct perception account can extend to situations such as dyads. Judging aperture passage for yourself plus another individual seems to require the building of a representation of the total width of yourself and your compatriot. Especially considering the methodology in Chang et al. (2009), where participants are paired with different people during the course of the experiment. Because the optical information from the environment does not change, then another source of information seems modulate judgment. Regardless of the source, it seems as though two pieces of information are used to make a passability judgment: optical information and some-as-yet-to-be-determined source. It is unclear how the ecological psychologist can maintain that the judgment is direct.

A second, perhaps more important, aspect to consider is the methodology used in, for example, Warren and Whang (1987) and Higuchi et al. (2012). The experiments in these studies include a methodology where participants walk through the apertures of various sizes, rotating their shoulders as needed. In Warren and Whang there is also a condition where participants judge whether they can pass through apertures, without rotating their shoulders. In all of these cases what the participant seems to be doing is making a judgment about passibility with respect to a future configuration of their body. Judgment in these cases does not seem to be based upon their current body width but, rather, the width of their body after they have rotated their shoulders. If that is the case, then it is not clear that
passability can be directly perceived in these cases. Instead, it seems as if the passability judgment is based upon a representation of the future state of the body. Such a representation can be plausibly drawn from memory or the result of a simulation.

## A Representation-Based Theory

The theory being presented is an information processing theory about the steps involved in passability judgments of the kind exhibited in previous aperture passage research. At a functional level the theory of geometric affordances posits a geometric comparison process that compares the geometric properties (width, depth, height) of an aperture against a current or stored body posture. The geometric comparison is used both when judging whether an aperture is passable as well as a top-down metric to control shoulder rotation during passage.

The information relevant to body postures is derived from body schemas. Although previous affordances research have rejected the notion of body schemas (e.g. Carello, Grosofsky, Reichel, Solomon, \& Turvey, 1989) there reasonable evidence for their existence and their role in motor planning and performance.

## Evidence for Body-Schemas

Schwoebel and Coslett (2005) identify three types of body representations: body schemas that represent the positions of the body parts and is used to plan motor movements, the body structure that is a topological map of body part locations, and a body image which is a lexical-semantic representation of the body detailing body part names, their function, and their relationship to related artifacts. Since body schemas are central to the information processing theory being proposed, a brief summary of evidence for the existence of body schemas will be presented in this subsection.

Neural evidence provides support for the functional role of body schemas as real-time representations of the body. Firing-rates in parietal area 5 of primates supports the idea of encoding arm postures both when the arm is occluded and when a realistic, fake arm is visible, suggesting both somatosensory and visual input is used to create body schemas (Graziano, Cooke, \& Taylor, 2000).

There is strong evidence for the use of body schemas in motor simulation. For example, the hand laterality paradigm has been used to study the link between imagined movement and actual movement (e.g. Parsons, 1987). There are two aspects of the laterality paradigm. The first involves making judgements of laterality (left vs. right) and the second involves simulating arm orientations. The reaction time for both tasks was relative to orientation differences between the participant's arm and the target arm. Simulated movements were strongly correlated with actual movements.

There is also evidence of a physiological overlap between imagined and actual movements (Decety, 1996; Lotze et al., 1999). The fMRI work by Lotze et al. (1999) also supports
the view that the main difference between imagined and actual motor movements is inhibitory signals from cortical motor areas to cerebral regions, inhibiting actual movements. Sirigu et al. (1996) also show that only patients with parietal damage do not show a correlation between the times for actual and imagined finger movements.

Finally, Schwoebel, Coslett and Buxbaum (Coslett, Buxbaum, \& Schwoebel, 2008; Schwoebel, Coslett, \& Buxbaum, 2001) provide evidence for Forward Modeling. Forward Models have been theorized to be used to develop representations of body schemas based partially from efferent copies of planned motor movements. What is particularly interesting in their work is the dissociation exhibited by a patient (JD) between body schemas due to purposeful movement and body schemas for passive movement. JD had accurate reach and pointing ability when moving her hand to a target in both occluded-hand and nonoccluded hand conditions, suggesting that she had an accurate representation of the position of her arm and hand. However, JD's pointing and reaching ability were impaired when her arm was moved by an experimenter (passive movement), suggesting that, in those cases, she did not build an accurate representation of her arm posture. As pointed out by Schwoebel and Cosslett (2005), this dissociation suggests that JD had an intact ability to generate posture representations from an internal model based on predicated movements (Forward Model).

The above findings in combination imply that humans have representations of the biomechanical constraints of our bodies. If reaction times for imagined movements mimics reaction times for actual movement, then this suggests that the simulated movement has similar kinematic and biomechanical properties as real movements. The fact that there is a strong neurological overlap between simulation areas and areas responsible for actual movements suggests that motor movements are encoded in the same format for simulation as they are for actual movements. It can be inferred that some form of biomechanical representation has to exist to support biomechanically-accurate simulations. This offers compelling evidence that the biomechanical constraints of the body are also likely represented (in order to support simulation). It stands to reason that simulation can produce predictions of body posture in simulated motor planning in much the same was as forward modeling does for active motor behavior.

The theory being forwarded here assumes that we store body schemas of biomechanical constraints. This would be useful for motor planning because it would reduce the complexity of choosing a goal posture. For example, shoulder rotation would require only three representations: a body schema for relaxed, non-rotated posture; and a body schema for full rotation to the left; and full rotation to the right. Although the body is capable of rotating any angle between constraints, it would be costly to store them all. Instead, biomechanical constraints can provide sufficient conditions for an action (fully rotated shoulders might be sufficient for passing through an aperture), which is suitable
for planning. Online motor control during action could then be used to control and produce only the necessary motor movements to carry out the action for a particular circumstance.

## Information Processing Theory

It is useful to divide the processes proposed in this theory into two phases: the judgment phase and the performance phase. In the judgment phase, we first determine if we can pass through an aperture at all. The performance phase occurs once we have judged an aperture as passable and begin to walk through it. The performance phase can be subdivided further into three sub-phases: rotation initiation, rotation, and rotation termination. The following section will outline how body schemas are used in the passability judgment.

## Judgment Phase

Although it is discussed very little in the previous aperturepassage literature, before we ever attempt to pass through an aperture, we must first make a judgment of whether passage is at all plausible. Anecdotally, this must be the case because we simple do not often find ourselves trying to squeeze through apertures smaller than our bodies. This process has to be more complex than the direct perception theory proposes because passage cannot be judged purely on current posture. That is, optical information tuned to a nonrotated posture can only inform passibility judgments where no postural change is required. However, in order to judge passability in the condition where some degree of shoulder rotation is required, the optical information would somehow have to be tuned to a future state of the shoulders. It is unclear how a direct perception approach could account for this.

Geometric affordance theory proposes that a positive passability judgment results from two possible cases. In the first case, body geometry is estimated from a body schema of the current body posture. This information can then be used top-down in a visual search to find apertures of an appropriate size. If the vision system is able to return a feature in the environment that meets those constraints, the returned apertures are considered passable. In this case, the agent can simply walk through the aperture. If no environmental feature is returned by the vision system, the second case proceeds. Note that the representations used in this phase are non-static: they are current (based on current body posture) and can include other sources of information including visual or proprioceptive (such as estimates of body size while carrying objects, or in a dyad).

In the second case, a potential series of memory requests are made for stored body schemas that closely match the current body posture (e.g. standing) and current action capabilities (e.g. supportive of walking action) but are relaxed on an increasing number of postural details (e.g. no need to match with respect to the upper-half of the body). In the case of a simple doorway-like aperture, a reasonable memory request would be for a posture that affords walking
(e.g. a standing posture) but allows for variation in torso posture (such as shoulder rotation). As discussed in the previous subsection, storing only the biomechanical constraints vastly reduces the search space for a suitable posture. If a suitable schema is returned, the geometric properties of that schema are used to filter visual results in the same manner described above for the for the first case. In the second case, the retrieved body schema functions as a goal state for the motor system during the rotation phase. That is, the motor system will try to achieve the posture at the biomechanical constraint (e.g. shoulders fully rotated) regardless if that posture is necessary for the desired action.

## Rotation Phase

Another aspect of aperture passage with no known discussion in the affordances literate is the need for some trigger that starts the rotation. One possibility is that the agent plans to rotate at some specific point and initiates rotation upon arrival. A second possibility, and the one explored here, is that there is a bottom-up environment trigger that is responsible for initiating the rotation. The theory proposed here is that the visual system performs bottom-up obstacle avoidance and that the presence of the edges of the aperture triggers the rotation. When the edges of the aperture are within a multiple of the agent's rotation radius, the vision system pushes information into the visual buffer, and the agent can respond by carrying out the motor plan.

Recall that during the judgment phase a stored body schema memory may be recalled and used as a goal state for the motor system to achieve the affordance. In Warren and Whang's (1987) first experiment there is a multi-second delay between what I am describing as the judgment phase and the rotation phase (while the participants walks to the aperture). It is proposed that once a body schema is retrieved it is maintained in working memory. When the presence of the obstacle (aperture edges) is pushed in to the visual buffer, combined with the presence of a body schema in memory, the agent can then carry out the motor rotation plan. Note that for shoulder rotation, the goal state will be a biomechanical constraint, e.g. fully-rotated shoulders. However, we know from Warren and Whang (and intuitively) that we do not rotate our shoulders to maximum rotation every time we rotate. Instead the theory assumes that rotation completion is controlled by a vision-action loop in the dorsal visual stream (Milner \& Goodale, 2008).

## Rotation Completion Monitoring

In their Two Visual Streams Hypothesis, Milner and Goodale propose a functional distinction between the dorsal visual stream and the ventral visual stream (Goodale \& Milner, 1992; Milner \& Goodale, 2008). They propose that the ventral stream composes what they call vision-forperception and that the dorsal stream composes what they call vision-for-action. While the ventral stream is used for planning action and carrying out unpracticed action, the dorsal stream is used for moment-to-moment visual
updating of actions that are comparatively more automatized.
The theory proposes that a moment-to-moment visual updating can occur through rapid repetition of the original top-down visual filter process described above (i.e. the current body schema is used in a top-down visual search to determine if there are any environmental features that meet those constraints). This moment-to-moment visual updating continues until (in this case) the shoulders have rotated enough to produce a match between the body-width of the agent and the width of the aperture. Although a biomechanical constraint was originally retrieved in the judgment phase, the agent need not always rotate the shoulders maximally. This process ends once the shoulders have rotated sufficiently to pass through the aperture. In other words, the goal state of the motor system was to fully rotate the shoulders, but a moment-to-moment visual update limits the total rotation by comparing the geometric properties of the current body schema (rotated shoulder in this example) to the geometric properties of the aperture. If an aperture is found as a result of the visual search, that means an aperture with sufficient geometric constraints has been found (for whatever posture the body is currently in). In this way, there can be a limited number of stored biomechanical constraints but a large variance in intermediate postural change (a large variance in shoulder rotation). Note these processing steps are the exact same steps used in the judgment phase.

## Computational Model Support

A computational model of the shoulder rotation experiments in Warren and Whang (1987) and in Higuchi et al. (2012) was developed as an initial test of the overall theory. The model was modeled in an extension to Python ACT-R called ACT-R 3D (Somers, 2016). At a high-level, the model follows the information processing description described above. Importantly, with respect to affordance research, the model is not semantically informed about the aperture in its environment.

It would not be atypical for an ACT-R model to be semantically informed. It is fairly customary for a model to use what is termed a 'visual icon' with a chunk identifying to the programmer what visual information the agent is 'seeing.' Although semantic information is not contained in the visual icon, it would not be atypical for a production to be pre-programmed to respond to the contents of the chunk in the visual icon.

The visual system in ACT-R 3D is slightly less informed. There is, in the agent's 3D environment, nothing labeled as an aperture. In fact, an aperture is negative space between environment features (such as walls) and cannot in fact be labelled in ACT-R 3D. Although the walls in the 3D environment are labelled, the agent has no access to those labels.

Instead, the agent has a goal to walk forward and in order to carry out this goal, it looks for obstacles. Upon finding an obstacle (the wall), the agent then uses a top-down visual
search for features that might be passable in the manner described in previous sections. Put simply, the agent does a visual search for empty space in front of it that meets the geometric constraints of the agent's body (or an achievable body posture). In this way the agent does not in fact represent the aperture as an aperture. Importantly, this also means that agents of different sizes will make different passability judgments.

The task the model must perform is to walk through an aperture, rotating the shoulders as needed, or avoid walking to apertures that it thinks it cannot pass through. As described above, if the agent does perform shoulder rotation, a moment-to-moment visual update occurs to determine if the agent should stop rotating. A single model is used for both small and large agents in slow and fast walking conditions, walking through apertures of various sizes, modeling experiments in Warren and Whang (1987); as well agents holding bars of various lengths and walking through apertures of various sizes in Higuchi et al. (2012).

The measure of fit to Warren and Whang was with respect to total rotation which is influenced partially by the number of agents who decide to pass through an aperture of a given size, rotation speed, and walking speed. There were four conditions to fit: 2 (size: small vs. large) x 2 (speed: slow vs. fast); with Pearson correlations ranging from 0.91 and 0.98 . The same model was then given bars of different sizes and performed the experimental conditions given in Higuchi et al. (2012). Although the fit was not as good in this case, as it showed a strong over-rotation in one condition; the fit was still reasonable, especially with the exclusion of the results for the over-rotated condition. The measure was rotation angle as well as the safety margin made between the end of the bar and the edge of the aperture, producing a Pearson's correlation of 0.84 for absolute rotation and 0.89 with respect to safety margin.

The success of the model is encouraging, given that the accuracy of the results are dependent on the timing involved, which is a product of the information processing steps (in the form of productions) that the agent carries out.

## Discussion

There was a number of difficulties pointed out in the first section that affordances based upon direct perception has to contend with. This section will address those difficulties but will also describe an interesting fallout from using an affordance-based approach.

## Addressing Difficulties with Direct Perception

The first difficulty pointed out in the first section that direct perception has to contend with is a person-plus-other system. In cases like these, there is no invariant property of the body that can act as units to directly perceive: there are measures beyond the body that affect the judgment. The theory presented in this paper also has to be extended to account for situations like these. When an agent is part of a person-plus-other system, the theory proposes that the agent could combine representations, including body schemas, to
make a estimation of the total geometric properties. Currently neither the theory nor the model define processes for including accompanying objects (in the bar experiments, the agent has special access to the dimensions of the bar). However, the advantage with the model is that there is a clear question that can be incorporated into a unified theory in the future.

In the same manner the model (and theory) also assumes that the geometric properties of the environment can be suitably perceived. The details of this process are not yet modeled, however, we can assume that aspects such as eyeheight, head-sway, and stride-length, can all be combined to creates a representation of the aperture width. In that respect, the model would be very much in line with findings from the aperture-passage literature.

The model can also help answer questions about representational content. The model presented here is part of a series of models that address whether A/S ratio or spatial margin (between edges of the agent and edges of the aperture) might be used as a metric for aperture passage. The model presented here implement an analog of spatial margin to judge the fit between aperture width and body width, supporting Higuchi et al. (2012).

The model also helps explain over rotation evident in Warren and Whang (1987) and Higuchi et al. (2012). Because the processes during the rotation completion monitoring affect timing, they also introduce a degree of variance in the rotation. The model does not rotate perfectly each time and exhibits similar over rotation to human performance.

## Extensions

The proposed processing description given above could easily be extended to include other affordances as well. For example, Stefanucci and Geuss (2010) researched aperture passage that required a ducking action. There is no principled reason why the same model could not be used to model those experiments as well. Since the problem is largely geometric, followed by a postural change, there is no principled reason why that postural change could not be for a ducking action. The same process could also be used for any situation that requires a postural change in order to accommodate the size of the body.

Secondly, not all affordances are purely geometric but could involve a geometric comparison process. Grasping, for example, has a number of elements, one of which could involve a judgment of whether the target object would fit in a grasp.

## Conclusion

The term 'affordance', though convenient, does not come without certain theoretical baggage. The aim of this paper is not to dismiss or discredit ecological psychology or the notion of direct perception but, rather, to compliment it with an information processing description. The term, 'representation', need not carry the kind of baggage that it may have historically. The representations used in the model
are, for the most part, not static and semantically-laden. For example, each agent learns their own body schemas before experiments by performing 'exercises', storing and updating new representations for biomechanical constraints. Furthermore, the environment is not labelled in any way. Agents in the simulation have to determine what apertures are passible individually.

Adapting affordance research to a representationalist framework opens some doors for research. This work is mainly philosophical, arguing for the need to unify research in a way that is falsifiable. The theory here presented relies on a cognitive mechanism capable of comparing the geometric properties of an environment with the geometric properties of an agent or agent-plus-object systems. This high-level presentation of the theory does, admittedly, offer very little detail about the working of the mechanism but does so in the hope of inciting research into the area.

## References

Anderson, J. R., \& Lebiere, C. (1998). The Atomic Components of Thought. Mahwah, NJ: Lawrence Erlbaum Associates Ltd.
Carello, C., Grosofsky, A., Reichel, F. D., Solomon, H. Y., \& Turvey, M. T. (1989). Visually Perceiving What is Reachable. Ecological Psychology, 1(1), 27-54.
Chang, C.-H., Wade, M. G., \& Stoffregen, T. A. (2009). Perceiving affordances for aperture passage in an environment-person-person system. Journal of Motor Behavior, 41(6), 495-500. h
Chemero, A. (2003). An Outline of a Theory of Affordances. Ecological Psychology, 15(2), 181-195.
Chemero, A., \& Turvey, M. T. (2007). Gibsonian Affordances for Roboticists. Adaptive Behavior, 15(4), 473-480.
Coslett, H. B., Buxbaum, L. J., \& Schwoebel, J. (2008). Accurate reaching after active but not passive movements of the hand: Evidence for forward modeling. Behavioural Neurology, 19(3), 117-125.
Decety, J. (1996). The neurophysiological basis of motor imagery. Behavioural Brain Research, 77(1-2), 4552.

Fath, A. J., \& Fajen, B. R. (2011). Static and dynamic visual information about the size and passability of an aperture. Perception, 40(8), 887-904.
Gibson, J. J. (1986). The ecological approach to visual perception. Hillsdale, NJ: Erlb.
Goodale, M. A., \& Milner, A. D. (1992). Separate visual pathways for perception and action. Trends in Neurosciences, 15(1), 20-5.
Graziano, M. S., Cooke, D. F., \& Taylor, C. S. (2000). Coding the location of the arm by sight. Science (New York, N.Y.), 290(5497), 1782-1786.
Higuchi, T., Seya, Y., \& Imanaka, K. (2012). Rule for Scaling Shoulder Rotation Angles while Walking through Apertures. PLoS ONE, 7(10), 1-8.
Higuchi, T., Takada, H., Matsuura, Y., \& Imanaka, K. (2004). Visual estimation of spatial requirements for
locomotion in novice wheelchair users. Journal of Experimental Psychology. Applied, 10(1), 55-66.
Lotze, M., Montoya, P., Erb, M., Hülsmann, E., Flor, H., Klose, U., ... Grodd, W. (1999). Activation of Cortical and Cerebellar Motor Areas during Executed and Imagined Hand Movements: An fMRI Study. Journal of Cognitive Neuroscience, 11(5), 491-501.
Michaels, C. F., \& Carello, C. (1981). Direct Perception. (J. J. Jenkins, W. Mischel, \& W. W. Hartup, Eds.). Englewood Cliffs, NJ: Prentice-Hall.
Milner, A. D., \& Goodale, M. A. (2008). Two visual systems re-viewed. Neuropsychologia, 46(3), 774785.

Parsons, L. M. (1987). Imagined spatial transformation of one's body. Journal of Experimental Psychology. General, 116(2), 172-191.
Schwoebel, J., \& Coslett, H. B. (2005). Evidence for Multiple, Distinct Representations of the Human Body. Cognitive Neuroscience, 17(4), 543-553.
Schwoebel, J., Coslett, H. B., \& Buxbaum, L. J. (2001). Compensatory coding of body part location in autotopagnosia: Evidence for extrinsic egocentric coding. Cognitive Neuropsychology, 18(4), 363-381.
Sirigu, A., Duhamel, J.-R., Cohen, L., Pillon, B., Dubois, B., \& Agid, Y. (1996). The Mental Representation of Hand Movements After Parietal Cortex Damage. Science, 273(5281), 1564-1568.
Somers, S. (2016). ACT-R 3D: A 3D Simulation Environment for Python ACT-R. In D. Reitter \& F. E. Ritter (Eds.), 14th International Conference on Cognitive Modeling (pp. 107-112). University Park, PA.
Stefanucci, J. K., \& Geuss, M. N. (2010). Duck! Scaling the height of a horizontal barrier to body height. Attention, Perception \& Psychophysics, 72(5), 13381349.

Stoffregen, T. A. (2003). Affordances as Properties of the Animal-Environment System. Ecological Psychology, 15(2), 115-134.
Vera, A. H., \& Simon, H. A. (1993). Situated Action: A Symbolic Interpretation. Cognitive Science, 17(1), 748.

Wagman, J. B., \& Taylor, K. R. (2005). Perceiving Affordances for Aperture Crossing for the Person-Plus-Object System. Ecological Psychology, 17(2), 105-130.
Warren, W. H., \& Whang, S. (1987). Visual guidance of walking through apertures: body-scaled information for affordances. Journal of Experimental Psychology. Human Perception and Performance, 13(3), 371-83.

# Acquiring pitch associations across modalities: the role of experience 

Laura Speed<br>Radboud University, Nijmegen, Netherlands<br>Ilja Croijmans<br>Radboud University, Nijmegen, The Netherlands<br>Sarah Dolscheid<br>University of Cologne, Cologne, NRW, Germany<br>Asifa Majid<br>Radboud University, The Netherlands, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands.


#### Abstract

When interpreting our perceptual world, information from multiple perceptual modalities is often associated. Such crossmodal associations can arise from innate structural connections in the brain, statistical correlations in the environment, or through language. In a large group of participants across a wide age range and language background, we tested crossmodal associations between pitch and 7 dimensions in comparison modalities. We found evidence supporting the existence of all 7 types of associations, but the strength of association varied by dimension. Pitch-angularity and pitch-weight judgments were the most robust associations. In general, strength of associations increased with age, with significant associations occurring in the oldest age group (age 19+), consistent with experiential accounts of crossmodal associations.


# Theory of Mind and Valuation during Cooperation 

Michael Spezio<br>Scripps College, Claremont, California, USA<br>Tessa Rusch<br>Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf Hamburg, Germany<br>Yuqing Lei<br>Psychology \& Neuroscience, Scripps College, 1030 Columbia Avenue Claremont, CA 91711<br>Vanessa Hayes<br>Neuroscience, Scripps College, 1030 Columbia Avenue Claremont, CA 91711<br>Jan Glaescher<br>Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf Hamburg, Germany


#### Abstract

Societal progress requires humans to excel at cooperation over time. To sustain successful cooperation, people coordinate, especially about who is in the best position at any given moment to make the best decision or to take the best action for the team as a whole. We used a novel cooperation task between involving dynamic assignment of Teacher and a Learner under conditions of uncertainty both about reward and about who is the expert at any given time. The task is similar to Theory of Mind tasks but actually gives the participants a stake in the outcome. We found evidence for effortful representation of the preferences of others, and that successful prediction fosters cooperative success. Neural components and putative sources signaled changes in the role of expert in the task. Further, the task design allows novel applications of computational models to the cognitive dynamics and associated neural systems for cooperation.


# Fake News and False Corroboration: Interactivity in Rumor Networks 

Michael J. Spivey (spivey@ucmerced.edu)<br>Cognitive and Information Sciences<br>University of California, Merced<br>Merced, CA 95340 USA


#### Abstract

Rumors inundate every social network. Some of them are true, but many of them are false. On rare occasions, a false rumor is exposed as the lie that it is. But more commonly, false rumors have a habit of obtaining apparent verification, by corroboration from what seems to be a second independent source. However, in complex social networks, the connectivity is such that a putative second source is almost never actually independent of the original source. In the present work, rumor network simulations demonstrate how remarkably easy it is for a node in the network to be fooled into thinking it has received independent verification of a false rumor, when in fact that "second source" can be traced back to the original source. By developing a theoretical understanding of the circumstances under which the spread of false rumors, "alternative facts," and fake news can be controlled, perhaps the field can help prevent them from ruining elections and ruining entire nations.


Keywords: Networks, Social Networks, Interaction

## Introduction

The interactivity that exists among the subsystems that form a cognitive system has powerful and lasting consequences. In the human brain, the interactivity among the neural subsystems that form the language comprehension network is what allows phonetics to influence syntactic processing (Farmer, Christiansen, \& Monaghan, 2006) and semantics to influence speech perception (Gow \& Olson, 2015; Spivey, 2016). In the human brain, the interactivity among the neural subsystems that form the visual perception network is what allows depth perception to influence motion discrimination (Trueswell \& Hayhoe, 1995) and attention to influence visual perception (Gandhi, Heeger, \& Boynton, 1999; Spivey \& Spirn, 2000). In the human brain, the interactivity between the language comprehension network and the visual perception network is what allows visual context to influence spoken word recognition (Allopenna, Magnuson \& Tanenhaus, 1998; Spivey-Knowlton, 1996), and linguistic input to influence visual perception (Lupyan \& Spivey, 2010; Lupyan \& Ward, 2013). These examples form just a tiny subset of the many consequences of interactivity in the human brain.

Outside the human brain, interactivity in a social network has powerful consequences for group behavior. When two people cooperate on a shared task, or even just have a conversation, they often exhibit real-time motor coordination in their postural sway (Shockley, Santana, \&

Fowler, 2003; M. Richardson, Marsh \& Schmidt, 2005), their eye movements (D. Richardson, Dale \& Kirkham, 2007), their gestures (Paxton \& Dale, 2013), and their language use (Louwerse, Dale, Bard, \& Jeuniaux, 2012). It has even been shown that behavioral and neural responses of two participants cooperating on a task exhibit the signatures of competition between the two subtasks, even though each person is in charge of only one of those subtasks (Sebanz, Knoblich, Prinz, \& Wascher, 2006). Essentially, each person is doing some of the thinking for the other person. When these mechanisms of coordination are optimized between two people, they can even perform a joint perceptual task at a level that is better than either of them alone (Fusaroli et al., 2012).

When people share information with each other, they tend to self-organize into a larger cognitive system (Goldstone \& Gureckis, 2009). Much like how cognition may be an emergent property of billions of neurons interacting with one another in a brain (Kello, Beltz, Holden \& Van Orden, 2007), group cognition may also be an emergent property of multiple people interacting with one another in a shared context (Thiener, Allen, \& Goldstone, 2010). Due to the continuous fluid flow of information throughout the network, every node (be it a neuron or person) is richly interdependent with every other node, at least indirectly. Not only can positive influences spread throughout such a network, as when two brains show improved performance on a shared perceptual task (Fusaroli et al., 2012), but negative influences can also spread throughout the network and infect nearly every component. Network simulations of rumor-spreading have recently begun to analyze this process of false information infecting a social network (Roshani and Naimi, 2012).

Traditional studies of rumor transmission tended to focus on linear sequential transfer of a rumor, and how the content can often become accidentally modified after several transmissions (Allport \& Postman, 1947). Sometimes this is referred to as the "telephone game." However, more recent studies of rumor transmission have used network theory to examine how non-linear transmission of rumors happens in complex social networks that are richly interconnected (Del Vicario et al., 2016). For example, when the network has islands of homogeneity, tight-knit like-minded enclaves that connect mostly just to their own group, these subnetworks can become "echo chambers" that reinforce false narratives and conspiracy theories within their walls. Alternatively, when the
connectivity of a social network is scale-free (neither random nor homogenous) - much like the brain's connectivity (Kello, 2013; Sporns, 2010) - then almost any rumor can be expected to spread throughout the entire network, irrespective of whether it is true or false (Nekovee, Moreno, Bianconi, Marsili, 2007). What has not been explored yet in this small cottage industry of research is how easily a false rumor can obtain independent verification via an apparent second source, even when that "second source" actually has the original source as its origin.

When an interactive system (be it a brain or a group of people) spends any amount of time sending signals back and forth among its subcomponents, it quickly becomes difficult to trace the source of a signal and determine whether a given signal is afferent (recently coming from an external source) or efferent (better described as generated endogenously). Under these circumstances, following the trail of a rumor in a social network is extremely difficult. The journalistic practice of "corroborating the story" can become quite complicated. A common method of fact-checking is to find a second source for the same story. If the second source is independent of the first source, and says essentially the same thing, then it adds veracity to the report. Even naïve experimental participants tend to use this tactic (Kim et al., 2008). However, in an interconnected network of people sharing information, almost no one is actually independent of anyone else. Frequently, an apparent second source, which gets used as verification of the rumor, actually acquired its information indirectly from the original source.

One concrete real-world example of such false corroboration is the U.S. Pentagon's case for Saddam Hussein stockpiling weapons of mass destruction (WMD) at the beginning of the $21^{\text {st }}$ century. It has now been wellestablished that U.S. leaders were proactively seeking justification for a pre-existing plan to invade Iraq and depose its leader (Dreyfuss \& Vest, 2004; Ryan, 2006). It turned out to be all too easy for information gatherers to fool themselves into thinking they had corroborated reports of WMD, when in fact the corroboration was actually a duplicate of the original false rumor. The CIA, British intelligence services, and the New York Times all collected reports of WMD in Iraq, and carefully sought independent verification. Each of these entities received fallacious reports from the same Iraqi defector, codenamed "Curveball" by the CIA. And what's more, each of them used un-sourced reports from one another as corroboration of their own report. What they each did not realize at the time was that the "second source" to corroborate their report from Curveball was actually just someone else's report from Curveball (Bamford, 2005; Prados, 2004).

False rumors, "alternative facts," and fake news have become an everyday occurrence recently, where too many people obtain their news reports on social network sites and blogs, where "news" is provided that has not been vetted by policies of ethical journalism. For example, in January of 2016, journalist and author, Fareed Zakaria, was "trolled" on the internet with a fake report of him calling for "jihad
rape of white women to depopulate the white race." Some people believed this false rumor so strongly that they made threats on Zakaria's life, and frightening phone calls to his daughters in the middle of the night (Zakaria, 2016).

Similarly, in the fall of 2016, fake news reports were disseminated widely on Facebook about presidential candidate Hillary Clinton being involved in a child sextrafficking ring based at a particular pizza shop in Washington, D. C. One man believed that false rumor so strongly that he felt compelled to travel across state lines to visit that pizza shop with an assault rifle in his hands and fire a shot to let them know he was there to save the children. The U. S. Department of National Intelligence has recently determined that many such fake news stories about Hillary Clinton were fabricated and disseminated via social networks specifically with the intent of influencing the results of the 2016 U. S. election (DNI Report, 2017).

In Del Vicario et al.'s (2016) computational analysis of conspiracy theories on the internet, they concluded that, "many mechanisms cause false information to gain acceptance, which in turn generates false beliefs that, once adopted by an individual, are highly resistant to correction." In the following rumor network simulations, the results suggest that false corroboration may be one of those many mechanisms.

## Random Rumor-Net Simulations

In this first group of simulations, a 100 -node network was constructed and given random placement of bidirectional connections, excluding self-connections. In one set of 100 simulations, the network was given $10 \%$ connectivity, such that each node on average was connected to about 10 of the possible 99 other nodes (i.e., average node degree $=10$ ). The average clustering coefficient for this network (which shows how interconnected each node's friends are) was .10. Another set of 100 simulations used a clustering coefficient of .33 , and Figure 1 shows an example degree distribution from one of those networks. Another set of 100 simulations used a clustering coefficient of .5 , and a fourth set used a clustering coefficient of .67 .


Figure 1: Degree distribution from a 100 -node random network in which, on average, most nodes are connected to about 33 other nodes.

To begin a simulation, node \#1 was infected with a rumor by flipping its state from zero to 1.0 . This is the one-and-only origination of the rumor in this network. It could be true or false, but for the purpose of testing its evolution into "fake news," the rumor is treated as false. For every instance of transmitting the rumor, a randomly chosen infected node would select randomly among its connections to spread the rumor with one other node. After spreading, that bidirectional connection was erased in the network to prevent it from being used again in the future. (The simulation assumes that if the same rumor were shared again between the same two people, it would not count as a transmission.) For that very first transmission, this obviously involved node \#1 sharing the rumor with one of the nodes connected to it. At which point there would then be two nodes that have been exposed to the rumor. Then one of those nodes was randomly selected to spread the rumor again. After 100 transmissions of the false rumor, some of the nodes had still never been exposed, some had been exposed once, and some had heard the rumor from two or more different connections. This latter case counts as people who had heard the rumor corroborated by what would seem to be a second source. However, the simulation actually has only one source of the rumor: node \#1. For example, node \#1 might spread the rumor to node \#47, who then spreads the rumor to node $\# 23$. Next, node \#47 might share the rumor again, this time with node $\# 87$, who shares it with node $\# 18$, who then shares the rumor with node \#23. In that scenario, node \#23 could easily be fooled into believing that it had received independent corroboration (from node \#18) of the rumor it first heard from node \#47.

In this first group of simulations, the number of nodes that received this false corroboration was recorded for low-, medium-, high- and very high-connectivity networks (i.e., clustering coefficients of $.1, .33, .5$, and .67). Interestingly, after 100 transmissions of the rumor, there were no differences across these four different types of random networks (results averaged across the 100 simulations in each case). In all simulations, irrespective of how densely interconnected the network was, around 26 of the 100 nodes had heard the rumor from two or more sources (Table 1). This insensitivity to network density is likely due to the fact that a rumor-spreader is randomly selected each time (among nodes that know the rumor), and its relative likelihood of spreading the rumor to a knowing node or an unknowing node is unchanged by how well-connected it is.

Table 1: Random networks with different numbers of connections show about the same number of nodes hearing false corroboration of the rumor ( $2+$ times).

| Avg Node <br> Degree | Clustering <br> Coefficient | Never <br> Heard | Heard <br> Once | Heard 2+ <br> times |
| :---: | :---: | :---: | :---: | :---: |
| 10 | .10 | 33.6 | 40.1 | 26.3 |
| 33 | .33 | 34.6 | 38.8 | 26.6 |
| 50 | .50 | 34.8 | 38.4 | 26.8 |
| 67 | .67 | 34.9 | 38.4 | 26.7 |

With 200 nodes and 200 rumor transmissions (or 500 nodes and 500 rumor transmissions), again about onequarter of the nodes obtain false corroboration - irrespective of how densely or sparsely connected the network is. With half as many transmissions as there are nodes, about $10 \%$ of the nodes obtain false corroboration. And with twice as many transmissions as nodes, about $60 \%$ of the nodes obtain false corroboration. Based on these initial simulations, it appears that false corroboration of a rumor may be remarkably easy to obtain in a social network.

## Scale-Free Rumor-Net Simulations

Most real-world networks, including social networks, are not at all random in their connectivity. Instead, social networks tend to have a scale-free pattern of connectivity, meaning that most nodes have a smallish number of connections (node degree), while a few nodes have a very large number of connections. Using a version of Barabasi and Albert's (1999) preferential attachment process, a group of scale-free rumor networks were designed that show a power-law in their degree distribution (Figure 2).


Figure 2: (A) Degree distribution from a 100 -node scale-free network where the mean number of connections per node is 33 , but most nodes have $<25$ connections and a few nodes have $>75$ connections. (B) On log-log coordinates, the degree distribution forms a relatively straight line with a slope of -1.3 .

By contrast to a scale-free network, in a random network the proportion of connections each node has generally corresponds to the clustering coefficient as well. That is, if each node in a random network has about $10 \%$ of the possible connections, then the clustering coefficient (showing what proportion of each node's friends are connected to each other) will also tend to be around .10 . However, in a scale-free network, the clustering coefficient (.62, in Figure 2) tends to be substantially higher than the average proportion of connections the nodes have (.33, in Figure 2). That is, in a scale-free network, most nodes have relatively few friends, but a sizeable proportion of those friends know each other.

In these next simulations, a hundred 100 -node scalefree networks were designed that had an average of 10 connections per node, along with another hundred networks that had an average of 17 connections per node, then another hundred with 25 , and another hundred with 33 connections per node. (In a scale-free network, when the average number of connections approaches $50 \%$ of the possible connections, its degree distribution can become bimodal and no longer adheres to a scale-free power law. Therefore, the highest node degree used here was 33.)

Each rumor-spreading simulation with these scale-free networks was carried out in a fashion similar to those with the random networks, except that the first rumor-infected node could not be an arbitrary choice because some nodes were substantially more connected than others. To test the limiting case, the least-connected node in each scale-free network was selected as the first node to spread the rumor. After that starting point, 100 transmissions of the rumor took place exactly as it did with the random networks.

Table 2: Scale-free networks with different numbers of connections show about the same number of nodes hearing false corroboration of the rumor ( $2+$ times).

| Node <br> Degree | Cluster <br> Coeff. | loglog <br> slope | Never <br> Heard | Heard <br> Once | Heard <br> $2+$ times |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 10 | .22 | -0.71 | 40.1 | 34.6 | 25.3 |
| 17 | .39 | -1.05 | 40.4 | 34.3 | 25.3 |
| 25 | .56 | -1.25 | 40.8 | 33.9 | 25.3 |
| 33 | .68 | -1.27 | 40.8 | 33.6 | 25.6 |

In these scale-free rumor networks, a slightly larger proportion of the people never hear the rumor (about 40\%) compared to that in the random networks (about $35 \%$ ). However, remarkably, approximately the same number of false corroborations is observed $(\sim 25)$ as that seen with the random networks (compare Tables $1 \& 2$ ). As was tested with the random networks, this $25 \%$ false corroboration rate replicates for scale-free networks with 200 nodes and 200 rumor-transmissions. When there are 3-4 times as many transmissions as nodes, almost every node will have heard the rumor, and about $3 / 4$ of them will have heard it more than once (irrespective of network density). Not surprisingly, in these scale-free networks, it is usually the well-connected nodes that first obtain these false corroborations.

## When a False Rumor Becomes Fake News

Based on all these simulations, when there are as many rumor-transmissions as there are nodes, then almost $2 / 3$ of them will hear the rumor, and about $1 / 4$ of them will obtain a false corroboration of the rumor - even though it never actually had any independent secondary source. This is true for both random rumor networks and for scale-free rumor networks. However, when one of the people in the network is a reporter for a news agency, who will broadcast the story to everyone if they obtain apparent corroboration, then it turns out that the type of connectivity does, in fact, matter. If one assumes that the reporter is among the most widelyconnected people in the network, then the different degree distributions for random networks and for scale-free networks (Figures 1 and 2) make for substantially different reporters. In a random network, the most-connected node (i.e., the reporter) will have a number of connections that is greatly influenced by the density of the network's connectivity (its average node degree). However, in a scalefree network, the most-connected node is often connected to $>85 \%$ of the other nodes, irrespective of the average node degree. Therefore, a reporter in a random network will only occasionally obtain a false corroboration, and thus publish the story (Table 3). However, in a scale-free network, a reporter (who is massively well-connected) will almost always obtain false corroboration, and therefore publish the rumor (Table 4). If that rumor is false, then its publication qualifies as fake news.

Table 3: In random rumor-nets, false corroboration sometimes leads to the publication of fake news.

| Node Degree | Clustering <br> Coefficient | Reporter-node <br> Publishes Fake News |
| :---: | :---: | :---: |
| 10 | .10 | $58 \%$ |
| 33 | .33 | $42 \%$ |
| 50 | .50 | $38 \%$ |
| 67 | .67 | $35 \%$ |

Table 4: In scale-free rumor-nets, false corroboration almost always leads to the publication of fake news.

| Node <br> Degree | Clustering <br> Coefficient | loglog <br> slope | Reporter-node <br> Publishes Fake News |
| :---: | :---: | :---: | :---: |
| 10 | .22 | -.071 | $92 \%$ |
| 17 | .39 | -1.05 | $93 \%$ |
| 25 | .56 | -1.25 | $93 \%$ |
| 33 | .68 | -1.27 | $87 \%$ |

Surprisingly, with random networks, denser connectivity leads to a reduced likelihood of the reporternode obtaining false corroboration and publishing the rumor. Upon closer examination, this makes sense given the parameters of the simulation. In a random network with a small average node degree (sparse connectivity), whenever a rumor-infected node is about to spread the rumor, it has a small number of friends to choose among. If
one of them happens to be the reporter, which is somewhat likely since the reporter is the most connected node, then the reporter might hear the rumor. And if that happens a second time, then a (false) corroboration has taken place, and the story gets broadcasted. By contrast, in a random network with a large average node degree (dense connectivity), whenever a rumor-infected node is about to spread the rumor, it has a large number of friends to choose among. One of them is probably the reporter, but a random selection of to whom the rumor will be spread leaves the reporter with a slim chance. In many of these random rumor-net simulations, the reporter never even heard the rumor once.

The situation is very different in a scale-free network. In a scale-free rumor-net, most nodes have fewer connections than they would in a comparable random network. Therefore, when a rumor-infected node is randomly selected to spread the rumor, it is usually one that has a smallish number of friends to choose among, and one of them is almost certainly the well-connected reporter (see also Doerr, Fouz, \& Friedrich, 2012). Thus, almost every time the rumor is transmitted, the reporter has a reasonable chance of being its recipient. As a result, the reporter-node in such a network is highly likely to hear the rumor, and also highly likely to obtain a false corroboration of this rumor, even though the rumor actually has only one source.

## Conclusion

Interactivity in a network is usually a good thing. Ambiguities or uncertainties present in one part of the network will often be resolved by strongly biasing information present in another part of the network (e.g., Kawamoto, 1993; MacDonald, Pearlmutter, and Seidenberg, 1994; McRae, Spivey-Knowlton, \& Tanenhaus, 1998). However, when that strongly biasing information is objectively false, the interactivity within a network can compromise its ability to align itself with reality.

The present network simulations do not specifically distinguish between objectively false rumors and true rumors, but a recent analysis of 330 rumor threads on Twitter does. For a false rumor, the time between rumor onset and debunking can be as much as seven times longer than the time between rumor onset and verification for a true rumor (Zubiaga, Liakata, Procter, Hoi, \& Tolmie, 2016). That is, it takes much longer to debunk a false rumor than it does to verify a true rumor. Therefore, if a longstanding uncertain rumor has not been verified as true, then the odds are steadily increasing every day that it is a false rumor (that just hasn't been debunked yet). Most true rumors get verified very quickly.

However, the nature of this verification process comes into question when considering the present rumor simulations. If the apparent verification comes in the form of a seemingly independent source that corroborates the original rumor, it may be illusory. The interactivity inherent in social networks can all too easily make a false corroboration (i.e., an echo from the echo chamber) appear as genuine independent corroboration.

One potential solution to this problem is for reporters to make better efforts at tracing the lineage of a report, so that two reports from the same source might be identified as such. A more reliable solution would be for journalism practices to avoid using secondary-source corroboration on its own as sufficient evidence to disseminate a story. These rumor network simulations demonstrate that it is simply too easy to obtain such corroboration in a fraudulent manner. Instead, the criterion for publication of a story might ought to include evidence that cannot easily be faked, such as photos, video, audio recordings, and documents whose source can be reliably determined. For example, if the report is that a public figure made sexist comments, or mocked a disabled person, or told the public a brazen lie, simply relying on two seemingly-independent sources to publish such a story may be insufficient. If the comments or mocking are evident in a video clip of the public figure, or if the lie is present in a verifiably-sourced tweet from the public figure, then those pieces of evidence should be what are repeatedly disseminated in reporting the story. Reports without such concrete evidence should be taken with a grain of salt, or perhaps not published in the first place.

It has been proven time and time again in everyday life, as well as in high-stakes politics, that the dissemination of false rumors can ruin lives, ruin elections, and even ruin entire nations. Understanding the mechanisms that allow, and exacerbate, the spread of misinformation in a social network of any kind may help with efforts to curtail and minimize the damage that can be done. The present simulations of a false rumor spreading throughout a network show convincingly that, even in a sparsely connected network, the "apparent corroboration" of a story often comes from a source whose own source can be traced back to the originator of the story, and thus should not actually count as independent corroboration. To quote Fareed Zakaria, "No matter how passionate people are, no matter how cleverly they can blog or tweet or troll, no matter how viral things get, lies are still lies."

## References

Allopenna, P. D., Magnuson, J. S., \& Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38(4), 419-439.
Allport, G. W. \& Postman, L. (1947). The psychology of rumor. NY: Holt \& Co.
Bamford, J. (2005). A pretext for war: 9/11, Iraq, and the abuse of America's intelligence agencies. Anchor Publishing.
Barabási, A. L., \& Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.
Del Vicario, M., et al. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554-559.
DNI report (2017). Background to "Assessing Russian Activities and Intentions in Recent US Elections ': The

Analytic Process and Cyber Incident Attribution.
https://www.dni.gov/ files/documents/ICA_2017_01.pdf
Doerr, B., Fouz, M., \& Friedrich, T. (2012). Why rumors spread so quickly in social networks. Communications of the $A C M, 55(6), 70-75$.
Dreyfuss, R. \& Vest, J. (2004, February). The lie factory. Mother Jones, pp. 34-41.
Farmer, T. A., Christiansen, M. H., \& Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. Proceedings of the National Academy of Sciences, 103(32), 12203-12208.
Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., \& Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. Psychological Science, 23(8), 931-939.
Gandhi, S. P., Heeger, D. J., \& Boynton, G. M. (1999). Spatial attention affects brain activity in human primary visual cortex. Proceedings of the National Academy of Sciences, 96(6), 3314-3319.
Goldstone, R. L., \& Gureckis, T. M. (2009). Collective behavior. Topics in Cognitive Science, 1(3), 412-438.
Gow Jr, D. W., \& Olson, B. B. (2015). Sentential influences on acoustic-phonetic processing: A Granger causality analysis of multimodal imaging data. Language, Cognition and Neuroscience, 31(7), 841-855.
Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. Journal of Memory and Language, 32(4), 474-516.
Kello, C. T. (2013). Critical branching neural networks. Psychological Review, 120(1), 230-245.
Kello, C. T., Beltz, B., Holden, J., \& Van Orden, G. (2007). The emergent coordination of cognitive function. Journal of Experimental Psychology: General, 136(4), 551-568.
Kim, N. S., Yopchick, J. E. E., \& De Kwaadsteniet, L. (2008). Causal diversity effects in information seeking. Psychonomic Bulletin \& Review, 15(1), 81-88.
Louwerse, M. M., Dale, R., Bard, E. G., \& Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. Cognitive Science, 36(8), 1404-1426.
Lupyan, G., \& Spivey, M. J. (2010). Making the invisible visible: Verbal but not visual cues enhance visual detection. PLoS One, 5(7), e11452.
Lupyan, G., \& Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. Proceedings of the National Academy of Sciences, 110(35), 14196-14201.
MacDonald, M. C., Pearlmutter, N. J., \& Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, 101(4), 676-703.
McRae, K., Spivey-Knowlton, M., \& Tanenhaus, M. (1998). Modeling the effects of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language, 37, 283-312.
Nekovee, M., Moreno, Y., Bianconi, G., \& Marsili, M. (2007). Theory of rumour spreading in complex social
networks. Physica A: Statistical Mechanics and its Applications, 374(1), 457-470.
Paxton, A., \& Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. Behavior Research Methods, 45(2), 329-343.
Prados, J. (2004). Hoodwinked: The documents that reveal how Bush sold us a war. New Press.
Richardson, D. C., Dale, R., \& Kirkham, N. Z. (2007). The art of conversation is coordination common ground and the coupling of eye movements during dialogue. Psychological Science, 18(5), 407-413.
Richardson, M. J., Marsh, K. L., \& Schmidt, R. C. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. Journal of Experimental Psychology: Human Perception and Performance, 31(1), 62-79.
Roshani, F., \& Naimi, Y. (2012). Effects of degree-biased transmission rate and nonlinear infectivity on rumor spreading in complex social networks. Physical Review E, 85(3), 036109.
Ryan, M. (2006). Filling in the 'unknowns': Hypothesisbased intelligence and the Rumsfeld commission. Intelligence and National Security, 21(2), 286-315.
Sebanz, N., Knoblich, G., Prinz, W., \& Wascher, E. (2006). Twin peaks: An ERP study of action planning and control in coacting individuals. Journal of Cognitive Neuroscience, 18(5), 859-870.
Shockley, K., Santana, M. V., \& Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance, 29(2), 326-332.
Spivey, M. J. (2016). Semantics influences speech perception. Language, Cognition, and Neuroscience, 31, 856-859.
Spivey, M. J., \& Spirn, M. J. (2000). Selective visual attention modulates the direct tilt aftereffect. Perception \& Psychophysics, 62(8), 1525-1533.
Spivey-Knowlton, M. J. (1996). Integration of visual and linguistic information: Human data and model simulations. PhD Dissertation, U. Rochester.
Sporns, O. (2010). Networks of the Brain. MIT press.
Sunstein, C. R. (2016). How Facebook makes us dumber. Bloomberg View, Jan. 8.
Theiner, G., Allen, C., \& Goldstone, R. L. (2010). Recognizing group cognition. Cognitive Systems Research, 11(4), 378-395.
Trueswell, J. C., \& Hayhoe, M. M. (1993). Surface segmentation mechanisms and motion perception. Vision Research, 33(3), 313-328.
Zakaria, F. (2016). Bile, venom and lies: How I was trolled on the internet. The Washington Post, Jan. 14.
Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., \& Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS One, 11(3), e0150989.

# Adapting to a listener with incomplete lexical semantics 

Sadhwi Srinivas (sadhwi@gmail.com)<br>Cognitive Science Department, Johns Hopkins University Baltimore, MD 21218 USA

Barbara Landau (landau@jhu.edu)<br>Cognitive Science Department, Johns Hopkins University<br>Baltimore, MD 21218 USA

Colin Wilson (colin.wilson@jhu.edu)
Cognitive Science Department, Johns Hopkins University
Baltimore, MD 21218 USA


#### Abstract

Speakers involved in a communicative exchange construct an internal model of their addressees and draw upon the model to craft utterances that are likely to be understood. In many realworld situations (e.g., when talking to a non-expert, non-native speaker, or a child), this process of audience design involves identifying gaps in the lexical-semantic knowledge of the listener and selecting alternative expressions. We examine speaker adaptation to a listener with incomplete lexical knowledge in the spatial domain, specifically a failure to comprehend the basic terms left/right. Experimental and modeling results provide evidence of rapid adaptation that is modulated by the availability of alternative spatial terms. We consider how our approach relates to recent work in computational pragmatics, and suggest that adaptation to the lexical knowledge of the addressee is an important but relatively understudied topic for future research.


Keywords: language adaptation; audience design; spatial language; lexical semantics; computational pragmatics

## Introduction

Speakers choose referring expressions on the basis of several factors, including their beliefs about the linguistic and conceptual knowledge of addressees (e.g., Pate \& Goldwater, 2015; Brennan \& Clark, 1996). For example, speakers tend to avoid or supplement proper names that, in their judgment, listeners do not know (e.g., Isaacs \& Clark, 1987; Fussell \& Krauss, 1992; Wu \& Keysar, 2007; Kutlak et al., 2016). This is part of a more general pattern of audience design (Clark \& Murphy, 1982) in which speakers construct internal models of specific addresses and use these models to facilitate communication. Another example of audience design is the formation of partner-specific conceptual pacts during a conversation. Speakers also show some ability to adapt their descriptions to a listener's viewing perspective when it is different from their own, presumably making it easier for the listener to identify intended referents in a scene (e.g., Schober, 2009). These adjustments to the needs of particular listeners in specific circumstances are analogous to the wellknown Lombard effect, in which people tend to talk louder in the presence of ambient noise (Lombard, 1911), suggesting that audience design is a fundamental phenomenon that occurs at many linguistic and conceptual levels.

Empirical and computational work on audience design has largely adopted the assumption that discourse participants share knowledge of basic vocabulary items. For example, the Rational Speech Acts framework (Frank \& Goodman, 2012) assumes that speakers and addressees have the same literal meanings for lexical expressions, and derives pragmatic usage from literal semantics through iterated probabilistic inference. However, the assumption of common word knowledge is not completely valid for many real-world scenarios. In the same way that experts addressing novices should avoid overly technical jargon, speech tailored to nonnative or child listeners must regularly work around basic lexical-semantic gaps.

Recently, Ferrara et al. (2016) closely investigated the linguistic choices that parents made when communicating spatial information to their 3-4yo. children. The language used by parents to describe the location of items in simple spatial arrays differed significantly from that of adults addressing other adults in the same task. Most notably, parents avoided the horizontal axis terms left/right-terms that are known to be acquired relatively late by children in general, and that were not reliably understood by many of the particular children in the study-while they used many other spatial terms (e.g., the vertical axis terms above/below) in essentially the same way that adults do when taking to one another. These findings support the claim that parents have well-tuned internal models of their children's lexicalsemantic knowledge and can design utterances for them by circumventing their lexical gaps.

In this study, we investigated whether adaptation to gaps in spatial language would occur in a minimal communicative setting. The parents in Ferrara et al.'s study had developed internal models of their children through extensive interaction with them. Here we sought to determine whether calibration to the addressee's lexical knowledge could develop much more rapidly, perhaps after only a few instances of communicative breakdown. Furthermore, parents presumably have access to a variety of top-down and bottom-up cues to gaps in children's spatial lexicons (e.g., general experience with child spatial language, instances where children explicitly ask for clarification of spatial descriptions). Here we sharply restricted the interaction
among interlocutors, providing only ambiguous, bottom-up cues to the addressee's knowledge of the spatial lexicon.

Following Ferrara et al. (2016), our study took the form of a referential communication task. Participants provided spatial descriptions to a listener who had either full knowledge of lexical-semantics or full knowledge except for a gap in the horizontal terms left/right. This is the same gap observed in young children, and indeed comprehension of left/right can be demanding even for typical adults (Sholl \& Egeth, 1981). If audience design is operative in this setting, participants should more often supplement or employ alternatives to the basic horizontal terms when addressing the listener who does not accurately comprehend them.

We further investigated whether and how accommodation of the listener's lexical gap was modulated by the stimulus array. For example, in describing the location of the object marked by the arrow in Fig 1A, alternatives other than left/right may not be obvious to the speaker. Contrast this with the arrangements of Fig 1B and 1C, in each of which a non-horizontal spatial relation is available to identify the target (i.e., inside the box, below a triangle). For arrangements such as Fig 1C, speakers might even prefer to refer to the vertical axis independently of any particular addressee (e.g., Logan, 1995; Ferrara et al., 2016).


Figure 1: Examples of array types used in the spatial communication task

Finally, we were interested in whether the adaptation found experimentally could be accounted for with a simple inhibitory mechanism: one that penalizes the use of terms that have resulted in communication errors, leaving unchanged all other aspects of the speaker's system for generating referring expressions. We formalize this mechanism in a high-level computational model of referential communication, show that it matches the detailed pattern of spatial language with minimal free parameters, and discuss how it relates to and extends previous work in computational pragmatics. The detailed empirical findings presented here contribute to the understanding of the form and limitations of lexical audience design, and our model delineates a way by which previous computational models may be augmented to account for lexical differences among interlocutors.

## Spatial communication experiment

Our experiment involved communication between speakers (the participants) and a simulated listener. In each trial, the
speaker described the spatial location of a target object in a visual array (see examples in Fig 1). The listener responded by selecting one of the objects, and the participant was then shown the listener's choice alongside the original display. Apart from this minimal communicative interaction, participants were provided no information about or feedback from the listener.

There were two conditions that differed by listener type. In the Full knowledge condition, the listener comprehended all English spatial expressions without error. In the Partial knowledge condition, the listener was identical except that comprehension of left/right and minor variants of those terms was at chance. We varied the spatial arrangements across the stimulus arrays to elicit a range of linguistic expressions and, most importantly, to provide varying opportunities for adaptation in the Partial condition. We were further interested in whether adaptation would involve primarily avoiding left/right or supplementing those terms with other spatial information, as well as in the coarse-grained time course of adaptation. Rapid avoidance of the lexical-semantic gap would provide evidence of an adaptation mechanism that is quite sensitive to bottom-up feedback and that inhibits expressions that have resulted in communication errors.

## Participants

This experiment was part of a series conducted online using Amazon's Mechanical Turk service. There were 48 participants $\left(\mathrm{M}_{\text {age }}=33.9\right.$ years, 25 males $), 24$ in each condition (Full vs. Partial knowledge). Individuals received a small monetary compensation for participating.

## Materials

The stimuli consisted of 32 spatial arrays similar to those of Ferrara et al. (2016). Target objects, marked by red arrows in the display, could not be uniquely identified by intrinsic properties such as shape and color. For example, yellow circle would not be a uniquely identifying description of the target in Fig 1C, but yellow circle on the bottom or yellow circle below a triangle would both be sufficient.

In all stimuli, the target could be identified by its position on the horizontal axis (e.g., yellow circle on the right). This maximized the potential contrast between participants in the Full condition, who could in principle describe all targets with left/right, and those in the Partial condition, who would have to employ other terms to communicate successfully.

The arrays varied in the alternative spatial descriptions that they afforded. In the Horiz type (10 items), the target could be identified only by its relative position on the horizontal axis (Fig 1A). Under the assumption that left/right are generally the most accessible or preferred terms for horizontal position, these arrays would be expected to provide the greatest challenge for adaptation.

All other array types contained alternative spatial relations that could be used in identifying descriptions: proximity of the target to another object; containment of the target within a bounding box (see Fig 1B); internal vertical or horizontal orientation of the target (e.g., a pencil pointing up or down);
or a vertical relation between the target and another object. We grouped the proximity, containment, and vertical/ horizontal orientation arrays into a single type called Horiz+Other (18 total items). For such arrays, adaptation could involve using horizontal terms other than left/right or, perhaps more simply, reference to the other spatial relation (e.g., circle in the box instead of circle on the left).

The Horiz+Vert array type (4 items; see Fig 1C) was separated from the others on the basis of extensive evidence that reference to the vertical axis is both linguistically and non-linguistically privileged relative to the horizontal axis (e.g., Logan, 1995; Carlson-Radvansky \& Logan, 1997; Fitneva \& Song, 2009). Speakers could prefer vertical terms for these arrays, not because of any specific beliefs about the experimental addressee, but as a reflection of this general privilege (as was observed in Ferrara et al., 2016). Adaptation would then be difficult to measure for these arrays, as left/right may be independently dispreferred by participants in both the Full and Partial knowledge conditions.


Figure 2: Proportion of responses using left/right only (i.e., without any additional spatial expression) produced by participants and predicted by the adaptation model.

Errors bars show standard error of the mean.

## Procedure

In each trial, the participant was presented with a single spatial array and asked to describe the target so that it could be uniquely identified by the listener. Participant descriptions were typed into a text box; no restrictions were placed on the terms used (i.e., responses were completely free). Following a brief delay, participants were provided feedback about the object selected by the listener (marked with a blue arrow in a copy of the array): this was either the target or the distractor object of the same shape and color.

In the Full knowledge condition, the simulated listener selected the target whenever the description contained at least one word from a large list of spatial expressions (determined through pilot testing). While participants in this condition could in principle have communicated "successfully" by using spatial terms that did not pick out the target, as we
report below the vast majority of responses were felicitous and identifying. In the rare instance that a description contained no spatial term, the Full knowledge listener selected the target or the distractor with equal probability.

The simulated listener in the Partial knowledge condition behaved identically except when the description contained left/right (or variants such as leftmost/rightmost) without any other spatial terms. For such responses, the listener randomly selected the target or matched distractor with equal probability (i.e., the listener responded as if no spatial term was present in the description).

Three practice trials at the beginning of the experiment emphasized the goal of providing complete, unambiguous descriptions (e.g., yellow circle at the bottom instead of simply circle or yellow circle for Fig 1C). Each stimulus array was repeated 4 times over the course of the experiment, resulting in 128 trials per participant. The order in which the arrays appeared was pseudo-randomized by participant, with the constraint that there were an equal number of Horiz trials in each half of the experiment ( 20 per half). This ensured that a participant who consistently used left/right for this array type would receive, on average, at least ten instances of negative feedback during the first half of the experiment.

## Results

Descriptions tended to be brief, containing a single spatial expression (average number of words per utterance: 2.83 in the Full condition, 3.13 in the Partial condition). Two manual coders determined that more than $95 \%$ of the responses across both conditions were sufficient to uniquely identify the target (given complete knowledge of spatial terms). The rare insufficient responses were produced sporadically across participants (i.e., not concentrated on any particular speaker).

Statistical analysis was performed on the sufficient descriptions. For these, communication errors could occur only in the Partial condition and only when left/right was used without any other spatial term. Accordingly, we focused on the way in which the rate of 'left/right-only' descriptions (i.e., left/right alone or combined with only shape or color terms) varied across conditions and array types (see Fig 2).

A mixed-effects logistic regression was performed with left/right-only as the dependent variable and fixed factors of condition (Full vs. Partial), array type (Horiz, Horiz+Other, Horiz+Vert), and three experimental 'phases' (with approximately one-third of the total trials in each phase). Phase was included as a rough estimate of the time course of adaptation, which we informally gauged to occur quite rapidly (i.e., after only a few instances of negative feedback). All fixed effects were weighted sum-coded, and the model included random intercepts for participants and stimuli. In light of the number of comparisons involved and the relative novelty of our research question and design, a conservative level for significance was chosen $(p<.01)$.

As anticipated earlier, the left/right-only rate was higher overall for the Horiz array type ( $\beta=2.24, p=.006$ ), with the rate for the Horiz+Other type effectively equal to the mean across all types $(\beta=0.05, p>.8)$. Note that this implies a
much lower rate for the Horiz+Vert arrays (as expected from the vertical bias discussed earlier). There were no significant main effects of experimental condition or of phase. However, condition and array type interacted significantly, reflecting the fact that the difference in left/right-only rate for Horiz vs. Horiz+Other was much larger in the Partial condition (Condition $\times$ Horiz: $\beta=-0.28, p<.01$; Condition $\times$ Horiz+Other: $\beta=0.39, p=.01$ ). In the Full condition, the rate of left/right-only descriptions was high for both Horiz and Horiz+Other. The Partial condition showed adaptation to the listener for the Horiz+Other array type that did not fully generalize to the Horiz arrays.

Finally, condition and phase interacted significantly (Condition $\times$ Phase1: $\beta=-0.5, p<.01$ ). This accords with the numerical finding that left/right-only rate increased slightly across phases in the Full condition (phase1: 69\%, phase2: $71 \%$, phase3: 73\%) but decreased across the first two phases in the Partial condition (phase1: 57\%, phase2: $47 \%$, phase3: $42 \%$ ). The small increase in the Full condition may be due to the sufficiency of left/right for all arrays: implicit selfpriming may have elevated the frequency of these terms, or participants may have explicitly realized that there was little need to generate alternative expressions. The decline after the first phase in the Partial condition, and indeed the difference in the first phase across the two conditions, indicates that adaptation to the listener occurred rapidly (i.e., within approximately 43 trials, after an average of 12 instances of unsuccessful communication).

In summary, we observed fine-grained adaptation to the addressee with a lexical-semantic gap in the spatial domain. There was little opportunity to find adaptation in the case of Horiz+Vert arrays, given the general bias to use vertical expressions. But for Horiz+Other arrays, participants in the Partial knowledge condition began to avoid left/right-only descriptions, responding to listener errors and performing differently than participants in the Full knowledge condition, within the first third of the experiment. Adaptation in the Partial condition was significantly lower for the Horiz arrays, which did not provide an alternative spatial relation that could be used to identify the target.

The preceding analysis does not reveal whether participants in the Partial condition attempted to avoid using left/right altogether, or continued to use the problematic terms but supplemented them with additional spatial expressions. For the Horiz+Other arrays, we found that avoidance was the primary strategy. For the Horiz+Vert arrays, we observed more of a tendency to produce redundant descriptions (e.g., top right or bottom left). However, this tendency was observed in both the Full and Partial conditions, suggesting that it may reflect lexicalization o spatial collocations rather than any adaptation strategy.

A major remaining issue is why the Partial condition participants found it relatively difficult to avoid (or supplement) left/right for the Horiz arrays. Replicating Ferrara et al. (2016), we found that when participants did switch to alternative expressions these were mostly ordinal terms such as first/second/last. Ordinals would in fact have
been sufficient to identify the target in all of the array types, making their relative infrequency as alternatives to left/right all the more striking. One hypothesis is that, when exposed to errors on left/right, participants (implicitly) concluded that the listener has imperfect understanding of reference to the horizontal axis in general. If correct, this hypothesis would imply that adaptation occurred at the level of spatial relations or axes rather than at the level of spatial expressions.

An alternative hypothesis is that ordinals-in contrast to proximity, containment, etc. terms-are strongly dispreferred relative to left/right for the purpose of describing our stimulus items. Under this hypothesis, the difficulty of adaptation for the Horiz arrays in the Partial condition should be mirrored by avoidance of ordinals in the Full condition. More generally, the probability of switching from left/right to another spatial expression in the Partial condition may closely track independently-established relative frequencies of terms used to describe our arrays. We formalized this hypothesis in the computational model of spatial language use and adaptation developed below.

## Computational model of adaptation

The model has two main components: baseline (or preadaptation) preferences for spatial term usage, and a mechanism for modifying the preferences in response to errors made by the listener. Our goal in this paper is not to explain the baseline preferences, but rather to estimate them from empirical usage frequencies. The estimates take the form of numerical (dis)preferences (or 'weights') assigned to various spatial and non-spatial term types (or 'attributes'). Once the baseline has been established, we show that a single inhibition parameter (i.e., a penalty for using left/right) suffices to closely match the detailed adaptation pattern of the experiment. A uniform penalty for left/right has different effects across the array types because the viable alternative attributes for each type vary independently in their weights.

## Baseline preferences

The baseline model assigns probabilities to a large set of array-specific sufficient descriptions. Each description contains one or more binary-coded attributes indicating the presence of spatial and other terms. Specifically, the spatial attributes we considered are horizontal (horiz: left/right), vertical (vert: above/below/up/down/top/bottom), proximity (prox: close to/next to/near/far/beside), containment (cont: inside /outside/within), vertical orientation (v.o.: pointing up/down, facing up/down), horizontal orientation (h.o.: facing towards/ away from), and ordinal (ord: first/second/last). The non-spatial attributes are shape (circle/pencil) and color (yellow/green) and. This coding abstracts away from minor syntactic permutations (e.g., circle to the right vs. right circle) and lexical variation (e.g., the circle on the right vs. rightmost circle). The set of sufficient descriptions for each array was formed by considering the spatial relations that could be used to identify the target and fully crossing these with one another and all
possible shape and color combinations (e.g., Table 1 lists the relations that are relevant for the Horiz+Other arrays).

Relative frequencies of the sufficient descriptions for each array were determined, in part, from the results of the Full knowledge condition above. However, because all targets in that condition could be successfully identified with left/right, it is plausible that this data overestimates the frequency of horiz (e.g., due to participant self-priming). More generally, we were concerned that Full condition data may provide a somewhat skewed estimate of the relative accessibility of different sufficient descriptions. Therefore, we conducted an additional experiment in which each participant provided up to five descriptions of an array. This experiment was performed by 19 undergraduates at the Johns Hopkins University, each completing 32 trials (one per array) for a small amount of course credit. The total frequency of a description for a given array was equal to the sum of its frequencies in the Full condition and in this experiment.

A conditional log-linear (or maximum entropy) probability distribution over descriptions was defined by assigning a weight $w_{i}$ to each binary attribute $f_{i}$ (e.g., Jurafsky \& Martin, 2009). The conditioning information was the array, which determines the set of alternative sufficient descriptions. Weights were tied across array types and fit by maximum likelihood to the array-specific description frequencies. The resulting weights were as follows: horiz (-1.14), vert ( -0.43 ), prox (-2.73), cont (-2.3), v.o. (-2.57), h.o. (-5.78), ord (3.06), shape (1.0) and color ( -0.68 ), where a higher weight indicates a greater preference for the attribute. Note that the model assumes independence of attributes, an idealization that we show to be largely effective but which is not inherent to the maximum entropy formalism.

## Modeling adaptation

Prior to experience with the listener, participants in the Partial condition should have the baseline attribute weights. After failed instances of communication with left/right, the weights could in principle be modified in various ways (e.g., by large changes after single errors, or much more gradually over the course of the entire experiment). Given the rapid adaptation found in the experiment, and in order to restrict the number of free parameters, we implemented adaptation as a single array- and speaker- independent decrease in the weight of horiz subsequent to the first listener error.

The error-driven penalty against horiz was fit by maximum likelihood to the Partial data, with the weights of all other attributes fixed at their baseline values. The best-fitting penalty $(\approx-0.70)$ was sufficient to make alternative spatial expressions more probable than left/right in the Horiz+Other arrays. However, left/right remained the most probable expression for Horiz arrays. Because the penalty was uniform across all array types, this and other asymmetries must reflect the relative frequencies of alternative spatial expressions in the baseline data. In this sense, the model derives the nuanced pattern of audience design in the Partial condition from independently-established usage patterns and a minimal assumption about the mechanism of adaptation (see Fig 2).

## Detailed results

We examined the predictions of the model in more detail for the four subtypes of Horiz+Other arrays. Using the weights and horiz penalty above, we generated predicted frequencies of the sufficient descriptions for each subtype by sampling responses for 24 simulated participants.

Collapsed over shape and color, the predicted frequencies of the various spatial attributes were highly correlated with the actual frequencies across the subtypes ( $r=0.96$ ). In particular, for the Horiz+Other arrays offering containment, proximity and vertical orientation as alternatives to the horizontal relation, the experiment revealed that participants adapted by switching to these alternatives in the Partial condition, thereby increasing the frequencies of these features over the horizontal. However, in the arrays where horizontal orientation of the target was available as an alternative, participants continued to use left/right for identifying the target. Table 1 shows that the model captured this difference and other variations in attribute frequency.

Table 1: Proportion of responses containing each relevant spatial attribute (in bold) produced by participants in the Partial knowledge condition, and predicted by the adapted model (in parentheses), for the Horiz+Other array subtypes.

| Proximity |  | containment |  |
| :--- | ---: | :--- | :--- |
| horiz | $.48 \pm .08(.53)$ | horiz | $.44 \pm .08(.39)$ |
| ord | $.20 \pm .08(.21)$ | ord | $.13 \pm .06(.16)$ |
| prox | $.21 \pm .06(.20)$ | cont | $.26 \pm .06(.34)$ |
| vertical orientation |  | horizontal orientation |  |
| horiz | $.33 \pm .09(.45)$ | horiz | $.73 \pm .08(.69)$ |
| ord | $.17 \pm .07(.19)$ | ord | $.23 \pm .08(.25)$ |
| v.o. | $.36 \pm .09(.28)$ | h.o. | .00 |

## Model limitations

While highly successful relative to our original goals, the model contains a number of simplifications that could be addressed in future iterations. The attribute weights and horiz penalty were assumed to be identical for all participants (and trials), but there may be substantial individual (or even triallevel) variation in preferences for referential descriptions. The assumption of independent attribute weights was for the most part viable, but some form of interaction is required to account for frequent vert+horiz collocations (e.g., top right). Our focus was on spatial expressions, but use of shape and color terms would also be of interest, especially when these are used redundantly. The assumption that the inhibition of horiz applies after the first communication error, rather than coming into effect more gradually, was also an idealization.

Finally, no attempt was made to predict the baseline preferences for spatial terms or attributes (e.g., the strong preference for left/right over first/last in the Full knowledge condition). This raises the more general question of what cognitive representations and processes lead speakers to select particular utterances from a set of sufficient referential
descriptions, only some aspects of which are due to audience design.

## General discussion

In this paper, we experimentally tested whether speakers adapt their language to listeners with a lexical-semantic gap. Such situations may arise commonly, both when experts talk to novices and when adult speakers of a language address second language learners or children. Inspired by previous work with children, we focused on the case in which the listener commands all spatial terms other than left/right.

We found that participants were able to rapidly identify the listener's lexical gap, and to avoid it in cases where other alternatives were readily available. Specifically, when the target object could be identified with another spatial relation, participants mostly switched to using that relation. However, adaptation occurred to a lesser extent when the target could only be identified by its horizontal relation. This pattern of results suggests that spatial language elicited in the experiment was shaped by audience design, but that other factors prevented complete adaptation to the listener.

We formalized those factors with a computational model that assigns probabilities to sufficient descriptions with independent attribute weights. The weights were fit to utterances from the Full knowledge condition, supplemented by data in which participants provided multiple descriptions of each array. This model may reflect the endpoint of iterative pragmatic reasoning, as in the RSA framework (Frank \& Goodman, 2012), but is closer in practice to the approach of Monroe \& Potts (2015), who remedy limitations of that framework by setting attribute weights empirically. Adaptation was then modeled in a simple form, as an errordriven inhibition of left/right that applied uniformly to all array types (and participants). Despite its simplicity, the model correctly predicted the different types and degrees of adaptation observed across arrays in the experiment.

While some previous models have addressed adaptation from bottom-up information about the listener (e.g., Janarthanam et al., 2010), none have considered gaps in basic lexical knowledge. Indeed, much work in theoretical and computational pragmatics assumes a generic addressee with the same lexical semantics as the speaker. The model developed here could be applied to other cases in which listeners have idiosyncratic gaps in technical or non-technical vocabulary. Adaptation to the lexical knowledge of the listener is an important aspect of cooperative communication.

## References

Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482-1493.
Carlson-Radvansky, L. A., \& Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. Journal of Memory and Language, 37(3), 411-437.

Clark, H. H., \& Murphy, G. L. (1982). Audience design in meaning and reference. Advances in Psychology, 9, 287299.

Ferrara, K., Silva, M., Wilson, C., \& Landau, B. (2016). Spatial language and the Embedded Listener Model in parents’ input to children. Cognitive Science, 40, 18771910.

Fitneva, S. A., \& Song, Y. (2009). The comprehension of "left" and "right" in a referential communication task. In Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 2687-2691). Austin, TX: Cognitive Science Society.
Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084), 998-998.
Fussell, S. R., \& Krauss, R. M. (1992). Coordination of knowledge in communication: effects of speakers' assumptions about what others know. Journal of Personality and Social Psychology, 62(3), 378-391.
Isaacs, E. A., \& Clark, H. H. (1987). References in conversation between experts and novices. Journal of Experimental Psychology: General, 116(1), 26-37.
Janarthanam, S., \& Lemon, O. (2010, September). Adaptive referring expression generation in spoken dialogue systems: Evaluation with real users. In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 124-131). Association for Computational Linguistics.
Jurafsky, D., \& Martin, J. (2009). Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River: Pearson Prentice Hall.
Kutlak, R., van Deemter, K., \& Mellish, C. (2016). Production of referring expressions for an unknown audience: A computational model of communal common ground. Frontiers in Psychology, 7: 1275.
Logan, G. D. (1995). Linguistic and conceptual control of visual spatial attention. Cognitive Psychology, 28(2), 103174.

Lombard, É. (1911). Le signe de l'elevation de la voix. Annales Maladies de l'Oreille et du Larynx, 37, 101-119.
Monroe, W., \& Potts, C. (2015). Learning in the rational speech acts model. arXiv preprint arXiv:1510.06807.
Pate, J. K., \& Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. Journal of Memory and Language, 78, 1-17.
Schober, M. F. (2009). Spatial dialogue between partners with mismatched abilities. In K.R. Coventry, T. Tenbrink, \& J.A. Bateman (Eds.), Spatial Language and Dialogue (pp. 23-39). Oxford: Oxford University Press.
Sholl, M. J., \& Egeth, H. E. (1981). Right-left confusion in the adult: A verbal labeling effect. Memory \& Cognition, 9(4), 339-350.
Wu, S., \& Keysar, B. (2007). The effect of information overlap on communication effectiveness. Cognitive Science, 31(1), 169-181.

# Children's use of lexical flexibility to structure new noun categories 

Mahesh Srinivasan (srinivasan@berkeley.edu) \& Catherine Berner (catherineberner@berkeley.edu)<br>Department of Psychology, University of California, Berkeley<br>Berkeley, CA 94704 USA<br>Hugh Rabagliati (hugh.rabagliati@ed.ac.uk)<br>Department of Psychology, The University of Edinburgh<br>Edinburgh, EH8 9JZ UK


#### Abstract

Because most common words have multiple meanings, children are often learning new senses of existing words, rather than entirely new words. Here, we explore whether children can use their knowledge of an existing word sense to constrain their interpretation of a new word meaning. Across two studies, we teach 3- and 4 -year-olds and adults novel words for materials, and manipulate whether those words are also used flexibly, to label objects made from those materials. We find that participants of all ages assign markedly different interpretations to the object labels when they have a prior, material meaning: Rather than extending them to other objects of similar shapes, they extend them on the basis of shared material, thus overriding the well-documented shape bias. These findings suggest that language learners can use a word's prior meaning to learn about the structure of its new meaning.


Keywords: polysemy; lexical flexibility; word extension; word learning; shape bias

## Introduction

A great deal of evidence in language development suggests that children constrain their guesses about the referents and extensions of new words through a variety of heuristics and biases (Clark, 1990; Markman, 1990). For instance, if children are told that a novel object is called a dax, they typically infer that dax will also refer to objects that are similarly shaped (Landau, Smith \& Jones, 1988). This socalled shape bias arises early in acquisition, and is thought to play an important role in lexical development.

Of course, the shape bias is not sufficient for acquiring adult-like meanings for words, since the extension of many words meanings goes beyond shape. A single word can often apply to multiple items that vary in shape, but that share an underlying essence (e.g., natural kind terms like bird; e.g., Gelman \& Markman, 1986), intended function (artifact terms like chair; e.g., Bloom, 1996), or substance (e.g., for mass nouns like bread). Consistent with this, in order to develop adult-like meanings for words, children are thought to draw on a variety of different cues - such as animacy, background knowledge, or functional affordances - to override the shape bias, and structure their new word meanings (Booth \& Waxman, 2002; Jones \& Smith, 1988; Kemler Nelson, 1995).

While previous studies have investigated how children learn the meanings of entirely new words, they have yet to address the fact that many of the new word meanings that
children have to acquire are not associated with novel word forms, but instead with word forms they are already familiar with. This is because most words are not unambiguous, but are instead flexible: Most common words are polysemous and denote a variety of different senses of meaning (Nerlich, Todd, Herman \& Clarke, 2003). The word glass, for instance, refers to both a transparent material and a drinking vessel made from that material. This phenomenon, which we refer to as lexical flexibility, is common both within languages and across languages. Further, lexical flexibility follows systematic patterns in English and in other languages: Multiple English words, for example, can label materials and objects made from those materials (glass, tin, etc.), animals and their meat (chicken, lamb, etc.), tools and functional uses of those tools (hammer, saw, etc.), and more (see Srinivasan \& Rabagliati, 2015).

In the present studies, we investigate how children's biases about word meanings interact with lexical flexibility: How does a word's first-acquired meaning influence children's guesses about the extension of new, additional senses of the word? In particular, are additional word senses learned in isolation, or does a word's first meaning bias the application of constraints like the shape bias?

Consistent with the idea that the extension of new word senses is constrained by knowledge of other word senses, many historically-derived senses of words appear to be partially influenced by their historically-primary sense. This can be observed with the word glass, whose extended drinking vessel sense is defined by a combination of shape, function and material (and whose meaning thus differs from words like cup, which are defined by shape and function alone). Lending support to the idea that children's interpretation of new word senses could be constrained by their understanding of existing, already-learned senses, a recent body of work in language development indicates that even preschool-aged children understand the semantic relations that license lexical flexibility in their language. For example, children expect words to label animals and their meat, but not other thematically-related items (Srinivasan \& Snedeker, 2014), and generalize patterns of lexical flexibility to new words (e.g., such that they expect new words to label tools and their functional uses; Srinivasan, Al-Mughairy, Foushee \& Barner, 2017).

Given these facts, it seems plausible that if children know one sense of a word, then they might use that knowledge to constrain their guesses about how that word should be
extended. This could cause them to override the shape bias in some cases, and extend new word senses according to other criteria. Some preliminary evidence for this idea comes from a study by Yoshida and Smith (2003), who showed that the shape bias was increased if a novel object was labeled with a familiar name that was strongly associated with a characteristic shape (e.g., ball), compared to a familiar name that was associated with a substance. However, subjects in this previous study were learning novel exemplars for existing words - as opposed to novel senses of existing words - thus leaving open the role of lexical flexibility in structuring new semantic categories.

In the present studies, children and adults were first taught a novel name for a material. They were then presented a novel object made from the same material, and either learned that the object name was the same as the material name (both were called gup) or was labeled using a new, distinct word (the material was called zev and the object called $g u p$ ). Our studies tested whether this manipulation of lexical flexibility affected participants' guesses about the extension of the object name - e.g., by making them more likely to privilege shared material as basis for extension in the flexibility conditions - using a forced-choice task (Study $1)$ and a more open-ended sorting task (Study 2).

## Experiment 1

Adults and 3- and 4 -year-olds participated one of three conditions: In the flexibility condition, the novel material and novel object were given the same label, and in the unambiguous condition they were given different labels. A final material $v s$. object condition tested whether participants learn distinct material and object senses of a flexible word, or instead a single vague meaning encompassing both objects and materials.

## Methods

## Participants

This study included 100 3- and 4-year-olds from the Berkeley area (Range: 3;0-4;11; Mean age $=4 ; 0$ ), split roughly evenly among the three conditions. 48 adults were also recruited from the UC Berkeley campus community, with 16 participating in each condition. English was the primary language spoken by all participants. Children were tested in lab, and at local preschools and museums; Adults were tested in lab or at designated locations on the UC campus. Children were given a small gift for participating, and adults were given either course credit or a small gift. 16 additional children participated but were excluded for failing catch trials administered at the end of the task (described below; $n=12$ ), parental interference ( $n=3$ ), or experimenter error $(\mathrm{n}=1)$. Three adults were also excluded due to experimenter error. All participants were tested individually by a female experimenter.

Warm-up trials Participants completed three warm-up trials to ensure that they understood the task. The stimuli consisted of three sets of toy animals: two identical animals
and one contrasting animal (e.g. two bears and a horse). In each warm-up trial, the experimenter placed one of the duplicate animals on the table, and named it (e.g. "Here is a bear!"), and then placed the remaining two animals and asked the participant to point to the other matching animal (e.g. "I want another bear. Can you point to a bear?").

Test Trials Participants completed four test trials. The trials varied depending on which of the three conditions the participant was in.

The stimuli consisted of four sets of novel objects. Each set included (1) a jar of small pieces of a novel material and a wooden spoon, (2) a standard object made out of the novel material, (3) a material-match test object that was made out the novel material, but was of a different shape than the standard, and (4) a shape-match test object that was the same shape as the standard, but was made out of a different material.

## Material

This stuff is called
"gup" [flexibility] / "zev" [unambiguous]


Standard Object
This thing is called "a gup."


Figure 1: Example test trial from the polysemy and unambiguous conditions of Experiment 1. The novel material and standard object were given the same word in the polysemy condition and different novel words in the unambiguous condition.

In each test trial (Fig. 1), the experimenter brought out a jar of novel material and a wooden spoon. The experimenter labeled the material with a novel word, using mass syntax (e.g. "This stuff is called gup. This stuff is called gup. I have half a jar of gup here.") and then stirred the material with the spoon and scooped some of it out of the jar to emphasize that it was a material. The name given to the material varied depending on the condition: The material was labeled with the same novel word (e.g. gup) that was later used to label the standard object in the flexibility condition, and was given a different name (e.g. zev) in the uanmbiguous condition (Fig. 1).

Next, the experimenter brought out the standard object and named it (e.g. "Now look at this thing! This thing is called a gup."), and illustrated that it was an object by using count syntax and attributing a vague function to it ("I have two gups and I use them in my garage."). Then, the experimenter brought out the two test objects - the material-
match object and the shape-match object (in the flexibility and unambiguous conditions) - and asked the participant to extend the label for the standard object to one of the two test objects, using count noun syntax ("I want another gup. Can you point to a gup?"; Fig. 1). We expected that if participants use a prior sense of the word to constrain their interpretation of a new word sense, they should be more likely to override a shape bias-which typically arises when a count noun labels a rigid object (Landau et al., 1988)-and choose the material-match object in the flexibility condition.

The material vs. object condition was conducted to test whether participants who learned that the material and object were given the same word (e.g., when both were labeled $g u p$ ) in fact learned two distinct senses of the word (as opposed to a single word that can label both materials and objects). To test this, at test participants were asked to choose between the material-match object and a pile of the material. We reasoned that if subjects had learned a novel object sense of the critical word, they would choose the material-match more often than the pile of material, since the request at test employed count syntax (Can you point to a gup?), and thus a request for an individual.

Catch trials Finally, participants completed three catch trials at the end of the task to ensure that they had sustained their attention throughout the study. In these trials, the experimenter labeled a novel object with a novel word, and then asked participants to point to which of two subsequent objects could be labeled by the word. One of the choice objects was identical, and the other differed in shape and material. Participants who failed to correctly respond on at least two out of the three catch trials were excluded.

## Results

Our results are consistent with the idea that the meaning of one word sense guides children's guesses about subsequent senses. As indicated in Figure 2, children in the flexibility condition extended the name of the standard object to the material-match object ( $70 \%$ of trials; $S E=4 \%$ ), and were more likely to do so than children in the unambiguous condition ( $27 \%$ of trials; $S E=4 \%$ ), as revealed by a linear model ( $\beta=-1.86, S E=0.27, z=-6.79, p<.001$ ). Thus, while children in the unambiguous condition exhibited the robust shape bias documented in prior work (Landau, Smith, \& Jones, 1988), children in the flexibility condition overcame this bias and extended the new word sense according to material, rather than shape. ${ }^{1}$ This tendency to select the material-match object could not be explained by a failure to learn distinct senses of the flexible word (i.e., to learn both a material and object sense of the word), as children in the material vs. object control condition reliably selected the material-match object over the pile of material on $83 \%$ of trials ( $S E=3 \%$ ) of the time. This suggests that

[^506]children understood that the experimenter was requesting an object using the object sense of the novel word at test, as opposed to simply re-using the material word they had first been trained on.


Figure 2: Percentage of trials in Experiment 1 in which children chose the material-match object across conditions. Error bars show +/- 1 SE. Dashed line shows $50 \%$ mark.


Figure 3: Percentage of trials that adults chose the material-match object across the conditions of Experiment 1. Error bars show $+/-1$ SE. Dashed line shows $50 \%$ mark.

Adults (Fig. 3) showed a similar pattern of word extension choices to children in the flexibility condition, and extended the name of the standard object to the material-match object on $89 \%$ of trials ( $S E=4 \%$ ), significantly more than they did in the unambiguous condition ( $2 \%$ of trials, $S E=2 \% ; \beta=-6.24, S E=1.08, z=$ $-5.76, p<.001$ ). Unexpectedly, however, adults did not show the same pattern of choices as children in the material vs. object condition, and only chose the material-match object on $56 \%$ of trials ( $S E=6 \%$ ). This surprising result leaves open whether adults differentiated between the two senses of the word as clearly as children.

## Discussion

Experiment 1 showed that lexical flexibility allows children to override semantic heuristics like the shape bias: When a label for an object had previously also been used to label a material, then children's guesses about the further extension
of that object label were less reliant on shape, compared to a condition in which different labels were given to the object and material. Importantly, the additional material versus object control condition provided evidence that children did not simply conflate the "object" and "material" senses of this novel word into a single vague meaning: When asked to choose "a gup" from an object and a pile of material, children consistently chose the object. Thus suggests that children understood that while one sense of the word gup referred to a kind of material, another sense of the word (identified with count syntax) referred to an object. Surprisingly, adults behaved a chance in this condition, and we return to this result in the General Discussion.

What, then, is the status of the shape bias when children learn an additional sense under these conditions? In particular, did the participants believe that shape was entirely irrelevant to the meaning of the second sense, or did they simply privilege material when they were forced to choose between an item that matched in material (but not shape) and an item that matched in shape (but not material)? One possibility left open by the results of Experiment 1 is whether children in the flexibility condition might have chosen to extend the novel label for the object only to other items that matched in both material and shape, had they not been forced to choose between a material match and shape match (a limitation of the 2-alternative-forced-choice task used in Experiment 1).

## Experiment 2

Here, we employed a more open-ended task, giving participants more choice in how they determined the extension of the newly-learned words. 4-year-olds and adults were taught a label for a novel material and a label for a novel standard object, just as in the flexibility and unambiguous conditions of Experiment 1. Then, participants were shown an array of new objects that varied in shape, material, and size from the standard object, and were asked to classify which of these additional objects could be labeled by the same word as the word for the standard object (Fig. 4). As in Experiment 1, we varied whether the newlylearned object and newly-learned material shared a label. Using this method, we were interested in whether participants in the flexibility condition would restrict word extension to only items of the same shape and material as the standard.

## Methods

## Participants

Participants were recruited from the Berkeley area as described in Experiment 1. Experiment 2 included 32 4-year-olds (Range: 4;0-4;11; Mean age $=4 ; 6$ ), divided evenly between the flexibility and unambiguous conditions. 33 adults also participated (17 in the flexibility condition; 16 in the unambiguous condition). English was the primary language spoken by all participants. Three additional children participated, but were excluded for failing the
initial warm-up trials ( $\mathrm{n}=2$ ), or due to parental interference ( $\mathrm{n}=1$ ). All participants were tested individually by a female experimenter either at a children's museum or at designated locations on the UC Berkeley campus.

## Materials and procedure

Warm-up trials Participants completed three warm-up trials. Participants who failed on two or more of these trials were excluded. The stimuli consisted of three sets of toy animals. Each set included three animals from a single category and two animals from contrasting categories (e.g. three horses, a cat, and a fish). In each trial, the experimenter brought out a toy animal and told the participant what is was (e.g. "Here is a horse!"). The experimenter then put the animal into a plastic box and told the participant that the box was for the target animals (e.g. horses). The experimenter then placed the animal on the table with the other four animals and asked the participant to sort all of the animals (e.g. horses) into the box and all the animals that were not in that category, into a plastic bowl.

## Material

This stuff is called "kiv" [flexibility] /
"lof" [unambiguous]


## Standard Object

This thing is
called "a kiv"


## Sorting Test

Can you put all of the "kivs" into this box and all of the other things into this bowl?


Figure 4: An example test trial from Experiment 2. Participants sorted five objects as belonging to the target category or not. The novel material and standard object either received the same label (flexibility condition) or different labels (unambiguous condition).

Test Trials In each of the four test trials (Fig. 4), participants were first introduced to a novel material (but instead of pieces of a solid material, novel non-solid materials were used). As before, participants were then shown a standard object that was made out of the same novel material. Then, participants were asked to sort a set of five objects (four test objects plus the standard object itself) as either belonging to the target category or not. The four test objects varied in whether they matched the standard object in material and shape. In total, participants were asked to sort: (1) a +Material/-Shape Object that was made
out the same material, but was a different shape than the standard, (2) a -Material/+Shape Object that was the same shape as the standard, but was made out of a different material, (3) a +Material/+Shape object that shared the same material and shape as the standard, but was smaller, (4) a -Material/-Shape object that contrasted with the standard in both shape and material, and finally (5) the Standard Object itself.

In each test trial, the experimenter took out a jar of novel material. The experimenter told the participant the material's name (e.g. "This stuff is called kiv.") and then stirred the material with the spoon and then scooped and/or stretched the material, and took some material out of the jar (This was to underscore fact that the novel material was indeed a material). The name of the material varied depending on the condition, as before: In the flexibility condition, the material was labeled with the same novel word that was later used to label the standard object, and in the unambiguous condition was given a different name.

Next, the experimenter brought out the standard object and named it, using count syntax (e.g. "This thing is called a kiv."). The experimenter then brought out the four test objects and said "Some of these are kivs and some are not kivs." and then asked the participant "Can you put all of the kivs into this box and all of the other things into this bowl?"

## Results

The results from the more open-ended task of Experiment 2 paralleled those of Experiment 1. In particular, participants were more likely to privilege material in their extensions in the flexibility than in the unambiguous condition. Consistent with this, children in the flexibility condition were more likely to sort the + Material/-Shape object $(61 \%, S E=6 \%)$ as a member of the target category than children in the unambiguous condition ( $14 \%, S E=4 \% ; \beta=-2.25, S E=$ $0.44, z=-5.11, p<.001)$. In contrast, children in the unambiguous condition were more likely to show a shape bias, and sort the - Material/+Shape object $(88 \%, S E=4 \%)$ as a member of the target category than children in the flexibility condition $(38 \%, S E=6 \% ; \beta=2.46, S E=0.46, z$ $=5.37, p<.001)$. Meanwhile, children in both the flexibility and unambiguous conditions almost always sorted the Standard Object (Flexibility: 98\%, $S E=2 \%$; Unambiguous: $100 \%$ ) and +Material/+Shape Object (Flexibility: 98\%, $S E=$ $2 \%$; Unambiguous: $100 \%$ ) as members of the target category, and almost never sorted the -Material/-Shape Object as a category member (Flexibility: $2 \%, S E=2 \%$; Unambiguous: 0\%).

To examine whether individual children were internally consistent in their sorting, we coded the data in terms of their categorization strategies. Strategies were defined using a $75 \%$ cut-off: Participants who used the same strategy for 3 or 4 of the test trials were classified as having that categorization strategy, but were otherwise coded as other. As indicated in Figure 5, children in the flexibility condition more often sorted objects using a material-based strategy (i.e., sorting all three of the objects that matched in
material as being part of the target category), than children in the unambiguous condition. In contrast, a shape-based strategy (i.e., sorting all three of the objects that matched in shape as part of the target category) was more prevalent in the unambiguous condition.


Figure 5: Categorization strategies of children in the flexibility condition (left) and unambiguous condition (right).

Adults showed a similar pattern of choices to children. Participants in the flexibility condition were more likely to sort the + Material/-Shape object $(74 \%, S E=5 \%)$, as a member of the target category than adults in the unambiguous condition $(5 \%, S E=3 \% ; \beta=-4.03, S E=$ $0.65, z=-6.19, p<.001)$. In contrast, adults in the unambiguous condition were more likely to show a shape bias, and sort the -Material/+Shape object $(81 \%, S E=5 \%)$ as a member of the target category than adults in the flexibility condition ( $19 \%, S E=5 \% ; \beta=2.91, S E=0.44, z$ $=6.54, p<.001$ ). No adults in either condition sorted the -Material/-Shape as a member of the target category, and all but one participant sorted the Standard object and +Material/+Shape object as a member of the target category.


Figure 6: Categorization strategies of adults in the flexibility condition (left) and unambiguous condition (right).

Finally, similar to the children, most adult participants in the flexibility condition sorted objects using a materialbased strategy, while most in the unambiguous condition used a shape-based strategy. Only three adults (1 in the flexibility condition, 2 in the unambiguous condition) employed a material- and shape-based strategy, constraining the target category to only the standard object and the + Material $/+$ Shape object. Note that this third, more conservative strategy was never used by children.

## Discussion

While Experiment 1 showed how lexical flexibility can reduce reliance on the shape bias, Experiment 2 explored the nature of that reduction. We found that, in the presence of lexical flexibility, both children and adults tended to extend new meanings based on a single feature - material rather than through a combination of multiple features, such as shape and material (only 1 out of 16 adults in the flexibility condition used the shape \& material strategy), even though the task of Experiment 2 was open-ended enough to allow this strategy to emerge.

This reliance on material when extending the object labels in some ways conflicts with how flexible material words are used in languages. For instance, although the object senses of words like glass and tin are defined partially by material (a wooden box is not a "tin"), they label specific kinds of objects that are defined by shape and function (not all artifacts made of tin can be called "tins"). One reason that children and adults in the flexibility condition may not have used shape in their strategies is because we did not provide specific information about the functions of the standard objects; Such information could constrain hypotheses about the likely shapes - and functional affordances - of kind members (Kemler Nelson, 1995).

## General Discussion

How do children make inferences about the structure of new lexical categories? Guided by the fact that most common words are polysemous, our studies explored whether children's understanding of one meaning of a word would affect how they interpreted a subsequent meaning for that word. Two experiments demonstrated that, after children and adults learned that a substance name could also be used to label an object, they were less likely to extend that name according to shape, and instead privileged material.

These findings are consistent with a recent proposal (Srinivasan \& Rabagliati, 2015) that lexical flexibility plays an important functional role in language development, in facilitating the acquisition of the lexicon. By this account, it may be more difficult for children to learn an unambiguous lexicon in which each meaning has its own word, compared to one in which words label multiple meanings in predictable ways. Consistent with this idea, the present studies show that children's knowledge of an initial word sense can facilitate their learning of a second word sense.

Our findings also raise a host of interesting questions for future research. Some questions concern the precise meanings that children learned in our task. For instance, when children learned that gup could label both a material and an object, did they actually learn separate and conventionalized senses? Or did they simply realize that it was possible to "coerce" the meaning of gup from a material sense to a portioned object sense (cf. ordering two coffees, Frisson \& Frazier, 2005). Experiment 2 provides some evidence for the latter account, as children (and adults) were willing to extend the newly learned object name to any other object made of the material, regardless of its shape. This might also help make sense of why adults were at chance in
choosing between the material-match object and pile of material in Experiment 1: Adults may have thought that the flexible "object" label could apply to any individual made of the material, and may have been more flexible than children in construing a pile of material as an individual.

Although we have only explored our hypothesis in the context of materials and objects, our results hold implications for how lexical flexibility might shape conceptual development more broadly. While prior research has focused on how children use labels as 'invitations' to group items into common categories, our findings show that children understand that labels can pick out items from distinct, but related categories. In particular, by attending to lexical flexibility, children could use naming practices to draw inductive inferences about the structure of the world. For instance, just as hammers are used for hammering and shovels for shoveling children could reason that something called a dax that supports daxing is probably designed for that function, and that all daxes should support this function.

## Acknowledgments

Thank you to Nicole Hsu, Megan Merrick, and Sophie Waldman for their help with data collection. This research is supported by NSF SBE-16302040 to MS.

## References

Bloom, P. (1996). Intention, history, and artifact concepts. Cognition, 60(1), 1-29.
Clark, E. V. (1990). On the pragmatics of contrast. Journal of child language, 17(02), 417-431.
Frisson, S., \& Frazier, L. (2005). Carving up word meaning: Portioning and grinding. Journal of Memory and Language, 53(2), 277-291.
Gelman, S. A., \& Markman, E. M. (1986). Categories and induction in young children. Cognition, 23(3), 183-209.
Jones, S. S., \& Smith, L. B. (1998). How children name objects with shoes. $\operatorname{Cog}$ Dev, 13(3), 323-334.
Landau, B., Smith, L. B., \& Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive development, 3(3), 299-321.
Nerlich, B., Todd, Z., Herman, V., \& Clarke, D. D. (Eds.). (2003). Polysemy: Flexible patterns of meaning in mind and language (Vol. 142). Walter de Gruyter.
Srinivasan, M., Al-Mughairy, S., Foushee, R., \& Barner, D. (2017). Learning language from within: Children use semantic generalizations to infer word meanings. Cognition, 159, 11-24.
Srinivasan, M., \& Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. Lingua, 157, 124-152.
Srinivasan, M. \& Snedeker, J. (2014). Polysemy and the taxonomic constraint: Children's representations of words with taxonomically-different meanings. Language Learning and Development, 10(2), 97-128.
Yoshida, H., \& Smith, L. B. (2003). Known and novel noun extensions: Attention at two levels of abstraction. Child Development, 74(2), 564-577.

# Rationalizing subjective probability distortions 

Nisheeth Srivastava

IIT Kanpur

## Ed Vul

UCSD Psychology


#### Abstract

You cannot know the contents of a memory until after you have actually retrieved it. This paper considers the implications of this straightforward observation upon the psychological process of preference construction. We show that this constraint renders observers with random access memory susceptible to tail risks. We show that this difficulty can be rectified by permitting observers to weight memory retrieval for such observations, that outcome utility cannot be used for this purpose, but information-theoretic surprise can serve as a useful proxy for it. Using two novel experiments, we present evidence in support of our account. With the first, we show that humans find surprising experiences easier to remember. With the second, we show that surprising experiences in the past have a greater influence on future decisions than is statistically warranted. This twofold demonstration substantiates a psychologically plausible account for the origin of subjective probability distortions.


# Memory of relative magnitude judgments informs absolute identification 

Nisheeth Srivastava

IIT Kanpur


#### Abstract

The question of whether people store absolute magnitude information or relative local comparisons of magnitudes has remained unanswered despite persistent efforts over the last three decades to resolve it. Absolute identification is one of the most rigorous experimental benchmarks for evaluating theories of magnitude representation. We characterize difficulties with both absolute and relative accounts of magnitude representation and propose an alternative account that potentially resolves these difficulties. We postulate that people store neither long-term internal referents for stimuli, not binary comparisons of size between successive stimuli. Rather, they obtain probabilistic judgments of size differences between successive stimuli and encode these for future use, within the course of identification trials. We set up a Bayesian ideal observer model for the identification task using this representation of magnitude and propose a memory-sampling based approximation for solving it. Simulations suggest that the model adequately captures human behavior patterns in absolute identification.


# Exploration and Skill Acquisition in a Major Online Game 

Tom Stafford<br>Department of Psychology, University of Sheffield<br>t.stafford@sheffield.ac.uk

Sam Devlin<br>Digital Creativity Labs, University of York sam.devlin@york.ac.uk

Rafet Sifa<br>Fraunhofer IAIS, Germany rafet.sifa@iais.fraunhofer.de

Anders Drachen<br>Digital Creativity Labs, University of York andersdrachen@gmail.com


#### Abstract

Using data from a major commercial online game, Destiny, we track the development of player skill across time. From over 20,000 player record we identify 3475 players who have played on 50 or more days. Our focus is on how variability in elements of play affect subsequent skill development. After validating the persistent influence of differences in initial performance between players, we test how practice spacing, social play, play mode variability and a direct measure of game-world exploration affect learning rate. These latter two factors do not affect learning rate. Players who space their practice more learn faster, in line with our expectations, whereas players who coordinate more with other players learn slower, which contradicts our initial hypothesis. We conclude that not all forms of practice variety expedite skill acquisition. Online game telemetry is a rich domain for exploring theories of optimal skill acquisition.


Keywords: learning; games; skill acquisition; expertise; game analytics

## Introduction

Computer games afford a rich data set for the investigation of skill acquisition. Players invest tens, hundreds or even thousands of hours on individual games, and unlike offline domains of expertise - details of every action during practice can be unobtrusively recorded. The present analysis uses data from the online shooter video game Destiny, which has over 30 million active users as of 2016 (Nunneley, 2016). Using data on players' performance we trace their skill acquisition over time and relate it to their practice habits. Specifically we are interested in how variability in practice relates to learning.

The power law of learning is justly well-known in cognitive science (A. Newell \& Rosenbloom, 1981; Ritter \& Schooler, 2001), both for being a dependable regularity in skill acquisition data (Rosenbaum et al., 2001) and for expressing a truth we know from personal experience: when we first begin learning something new progress is often rapid, but later it slows or stalls. Nonetheless the presentation of Power Law learning curves based on averages masks the variability that occurs both within and between individuals (Gallistel et al., 2004; Gray \& Lindstedt, 2016). This is important for two reasons. Firstly because individual variability is interesting as an outcome. We wish to know why some individuals learn more rapidly, and achieve greater eventual levels of performance (and why some individuals are hindered in
their learning). Secondly, variability is interesting as a driver of learning. Previously it has been suggested that greater initial variability in practice may drive higher subsequent performance (Stafford et al., 2012; Stafford \& Dewar, 2014), a result which accords with computational accounts of how learning must balance exploration and exploitation of options (Sutton \& Barto, 1998; Humphries et al., 2012).

In addition to looking at how a skill is practised, there are also results which suggest an effect on skill acquisition of when a skill is practised - the issue of practice spacing (Stafford \& Haasnoot, 2017; Delaney et al., 2010; Cepeda et al., 2008) - as well as an effect of variability in how different components are practised (Magill \& Hall, 1990). From this perspective, variability is as much an engine of learning as consistency (Schmidt, 1975; Van Rossum, 1990; K. M. Newell \& McDonald, 1992; Ranganathan \& Newell, 2010). This raises the question of exactly which kinds of variability, and in what quantities, support optimal skill acquisition.

Previous work has looked at skill learning in a simple online game (Stafford \& Dewar, 2014; Stafford \& Haasnoot, 2017), with the emphasis that even a simple online game contains many fundamental cognitive processes perceptual, decision making and action implementation. Others have looked at skill development in more complex games (Thompson et al., 2017, 2013), and here we use the opportunity to analyse data from one such game, Destiny, to explore issues of how playing style, and particularly variability within play, affects skill development.

Destiny is a science-fiction themed, massively multiplayer, online game where players need to defend the Earth from various alien threats, taking on the role of Guardians. Players journey to different planets, complete missions, daily events, and perform a variety of different tasks to build up their characters. Destiny is a hybrid digital game that blends features from a number of traditional game genres including role-playing games and massively multi-player online games but which is first and foremost a shooter (Tammasia et al., 2016). The main components of the gameplay is focused on tactical single-player or small-team combat against players or artificial agents (Drachen et al., 2016).

Thousands of behavioural or performance-based metrics are tracked and stored by Bungie, the developer of Destiny, which in aggregate provides a detailed record of the behavioural history of Destiny players.

The metrics that can be calculated based on such datasets varies, and previous research in game analytics and other domains have seen such behavioural data being used for a variety of purposes (Tammasia et al., 2016; Rattinger et al., 2016; Drachen et al., 2016). For Destiny, a number of these metrics are of key interest in relation to evaluation of player skill and skill evolution.

- Playtime: Playtime in the current context simply refers to the amount of time a player spends playing the game per day, across either a single or all characters.
- Kills, Assists, Deaths: the shooter-heavy gameplay of Destiny means that traditional skill indicators from shooter games such as Kill/Death Ratio (KDR) form an important means for evaluating player skill.
For Destiny, a variant of KDR, the Kill-Assists/Death Ratio (KADR) is also used. An assist is a common term in esports signifying that a player helped another player take down a specific enemy (or in other ways help another player), without scoring the killing shot/hit on that enemy.
KADR is thus a more nuanced aggregate measure of performance than KDR. We use KADR-KDR as a measure of a players' propensity for 'social play'.
- Combat Rating: The Combat Rating (CR) is a game metric designed to reflect a players' overall skill. How CR is based on the TrueSkill system (Herbrich \& Graepel, 2006), a Bayesian model used for player/team ranking. TrueSkill and CR both serve a similar functionality to ELO (Charness, 2005). While the algorithm is confidential, it broadly works by initialising a player at CR 100. If the player is part of a team that wins a match, their CR goes up, more if there is a large difference in the CR between the two teams. Conversely, if they lose, the CR goes down, again in relation to the gap in CR between the two teams. CR is used by the Destiny matchmaking system to configure players into teams and balancing opponents. This means that players will be playing with and against players with similar CR (i.e. they are matched against players of simialr skill-levels).
- Grimoire Score: A Grimoire in Destiny is a record of a players experience - new cards are awarded the first time a specific action is taken or challenge overcome. In essence, the Grimoire score is an expression of the degree to which a player has explored the world and content of Destiny.

Working with very large datasets introduces some new opportunities for the cognitive scientist (Goldstone \&

Lupyan, 2016; Stafford \& Haasnoot, 2017). Observational studies, however large, necessarily have reduced power of causal inference compared to experimental studies. Large numbers mean that the data can be 'sliced' to explore if and how potential effects play out throughout the population, as well as allowing matching of individuals on various properties which might confound any effect. With enough data any observable difference can be 'statistically significant'. In experimental studies effort is expended in achieving enough power to make convincing inferences. With large data set it is more important to invest effort in exploring possible confounds and putting observable differences in the context of other effects via calculation of effect sizes.

Our hypothesis is that early variability will be associated with faster skill acquisition. This assumes that players have a tendency to under-explore the space of possible actions, and so, due to this reliance on habit, will be learning sub-optimally. We will test this hypothesis against different indices of variability in early practice: spacing of play, social play, world knowledge (grimoire score), and distribution of play across game modes (event entropy). These metrics are defined further below.

## Data and method

Our data comprise low level daily metrics indicating performance and meta information for over 20,000 randomly selected Destiny players. The behavioral telemetry was provided by Bungie.

For each player the data consists of a unique player ID and character IDs for each character the player has. A player is allowed to have at most three characters per account. For each character, the dataset contains daily aggregate player behavior such as number of deaths, completed objectives, weapon usage and average life span, and importantly playtime, each across the six game modes - or ways to play the game.

Our analytic strategy is first to split the data into a development (8682 players) and validation set (12861 players). All exploratory analysis was finalised on the development set, before being run on the validation set to produce the figures presented here. All conclusions presented are unaffected by the minor differences between the development and validation set results. This affords us some protection against discovering false patterns in our data that result from researcher degrees of freedom in analysis. It is inappropriate to make inferences from hypothesis testing p-values for an exploratory analysis such as this, but we report them for completeness where we have done standard analyses. Our main focus is on measures of effect size and confidence estimates around those measures.

Analysis scripts, player summary data, and full reports of both development and validation set results are available at https://osf.io/c59n9/. For commercial con-
fidentiality reasons the full raw dataset is not available at the point of writing.

## Analysis

First, we seek to confirm that players improve with practice. Following the method of (Stafford \& Dewar, 2014), we first select only players who play some minimum number of games (50). This produces a data set of 3475 longer term players (in the validation set; 1984 in the development set) and then divide by ranking all players according to the average of their three all time best ratings (in terms of CR). Figure 1 shows the average score per game for those who scored in the top third, middle third and lowest third of the high score table. This shows that the learning curve exists for averaged data, and that - in line with (Stafford \& Dewar, 2014) - players who end up with the highest scores begin the game with performance already above that of lower scorers (compare (Stafford \& Dewar, 2014) Figure 2).


Figure 1: Average performance rating as a function of game number and ranking based on players' highest three ratings. Error bars show +/- 1 standard error.

Note that our learning curves show performance, rather than speed, on the x -axis, and so are inverted relative to the classic 'Power Law of Learning'. None the less they reflect the expected decelerating function of learning with practice amount. Our investigation of other factors must take account this fundamental pattern in how player performance changes over time, as well as the stratification that we observe between players of differing initial performance. To do this, we fit a linear regression for each player's performance against game number. Because this regression produces a slope and an intercept, we are able to subsequently analyse player differences in both level of initial performance and subsequent change in performance (i.e. rate of learning). Henceforth when we refer to "learning rate" we mean the slope of this regression for each player. In order to explore which variables might be related to player learning rate we first visualise players split on some candidate variables against mean combatRating against prac-
tice amount (game number).

## Variation in practice timing - spacing

In order to compare practice timing, we calculate the time range over which players recorded their first 25 days of play (obviously this has a minimum of 25 days, and no theoretical maximum). This range correlates positively (Pearson's $\mathrm{r}=0.18,99 \%$ confidence interval $0.14,0.22$ ) with learning rate and negatively (Pearson's $\mathrm{r}=-0.09$, $99 \%$ confidence interval $-0.14,-0.05$ ) with initial performance.

In order to visualise the effect of greater or less spacing, we select players in the top quartile for spacing their first 25 games ('spacers') and those in the bottom quartile for spacing their first 25 games ('groupers') and plot the average performance against game for the two groups. This is shown in Figure 2.


Figure 2: Average score as a function of game number and high and low spacers. Error bars show +/- 1 standard error.

## Variation in practice type

Playing style - social play For each player we have a game by game measure of their 'assists', which are kills made by teammates which they are near to. Variation on this measure allows us to rate players according to a propensity for social play, i.e. a higher rate of assists will reflect a player who coordinates their actions with their team.

This measure correlates negatively (Pearson's $\mathrm{r}=$ $-0.16,99 \%$ confidence interval $-0.20,-0.12$ ) with learning rate and positively (Pearson's $\mathrm{r}=0.50,99 \%$ confidence interval $-0.47,0.54$ ) with initial performance.

Figure 3 shows the learning curve for players split on the average of their assists over their first 25 games, as an index of players' propensity for "social play". Those in the top quartile of the distribution of assists we term 'social players'. Those in the bottom quartile we term 'lone wolves'.


Figure 3: Social play and skill acquisition. Error bars show +/- 1 standard error.

World knowledge - Grimoire score For each player we are able to see the complete history of their Destiny playing, including how many games they play in total. Each player also has a 'grimoire' score, which is a count of the items they have encountered in the world. Obviously this is higher for players who have played more games, but there is considerable between-player variability, suggesting that some players focus on exploring the world, completing actions and collecting items, whereas others aren't focused on this aspect of the game. In order to compare grimoire scores between players who have complete different numbers of games, we calculate a normalised $(Z)$ score for each player based on the distribution of grimoire scores among players who have completed the same number of games.

This measure does not correlate with learning rate (Pearson's $\mathrm{r}=0.04,99 \%$ confidence interval $0.00,0.09$ ) and correlates positively, but weakly (Pearson's $\mathrm{r}=0.13$, $99 \%$ confidence interval $0.10,0.18$ ) with initial performance.

Figure 4 shows the average score, in terms of CR, against game for players whose grimoire Z scores are in the top and bottom quartiles of the distribution.

Playing style - mode entropy The play modes in Destiny are:

- Strikes: 3 player cooperative events.
- Raid: 6 player cooperative missions, requiring high level skills to complete.
- Story: the main single-player game mode, which can be played cooperatively by up to 3 players.
- Patrol: a single-player exploration mode.
- $P v P$ : all the player-vs-player ( PvP ) game modes of Destiny

Note that due to the aggregation into daily sets, it is possible for players to have played multiple sessions of Destiny within the same 24 hour cycle. Because Destiny


Figure 4: Average score by games for players with high and low grimoire count. Error bars show +/- 1 standard error.
has six different main game modes, it is of interest to evaluate how a player spends his or her time across those game modes. In order to quantify the measure of heterogeneity in terms of how a player splits their time between game modes, we use Shannon's entropy [see e.g. (Lessne, 2014; Algoet \& Cover, 1988)] which is defined as:

$$
\begin{equation*}
H=-\sum_{i} p_{i} \log _{2}\left(p_{i}\right) \tag{1}
\end{equation*}
$$

where $p_{i}$ denotes the probability of the player's activity across game modes $i$. For game mode $p_{i}$ is calculated as the amount of time spent in specific game mode $i$ divided by the total time spent playing all game modes that day.

Event entropy over the first 25 games for each player does not correlate with learning rate (Pearson's $\mathrm{r}=$ $-0.02,99 \%$ confidence interval $-0.06,0.03$ ) and correlates positively (Pearson's $r=0.22$, $99 \%$ confidence interval $0.17,0.26$ ) with initial performance.

Figure 5 shows performance against game for those in the top and bottom quartiles for event entropy calculated over the first 25 games.
Statistical model Hitherto, we have explored our data using visualisation of different groups and reported bivariate correlations. By entering all factors into a regression model we can check whether how all factors combine to explain variation in the learning rate. This is an essential complement to the visualisation. It allows us to confirm that patterns visible in the data are statistically significant. As well as the four measures described above - spacing, social play, grimoire score and event entropy - we include maximum numbers of games a player plays as a measure of overall motivation. The results of the regression of our five factors against the learning rate are shown in shown in Table 1.


Figure 5: Play mode entropy and skill acquisition. Error bars show +/- 1 standard error.

Table 1: Regression of player behaviours on player learning rate

| Factor | $B$ | $T$ | $p$ |
| :--- | :--- | :--- | :--- |
| Games played | 0.044 | 1.99 | 0.0465 |
| Spacing | 0.199 | 10.90 | 0.0001 |
| Assists | -0.172 | 10.04 | 0.0001 |
| Grimoire | 0.003 | 0.16 | 0.872 |
| Event entropy | 0.011 | 0.62 | 0.537 |
| $\mathrm{R}^{2}=0.063, F(5,3287)=44.47, p<0.0001$ |  |  |  |

Note that only spacing and assists, our measure of social play, are significant. Figure 6 shows the standardised regression coefficients (beta weights) when our five factors are used to predict learning rate (slope) and for the initial performance (intercept) of individual learning functions.

## Discussion

Using a complex online game we show that changes in player's performance can be tracked and related to aspects of how they play. We validate the separation of learning curves by initial performance shown by


Figure 6: Beta weights for players factors used to predict slope (learning rate) and intercept of individuals' learning functions. Standard error bars shown.)
(Stafford \& Dewar, 2014). As with that previous result, players who achieve the eventual highest levels of performance also perform better on their first game. Further, the difference between those with high and low initial performance only grows as more practice is completed.

Two other factors influence rate of learning - spacing, and social play. The effect of spacing matches that found in experimental studies of skill acquisition, as well as previous analysis of a different, simpler, game (Stafford \& Haasnoot, 2017). The differences between players who space their practice and those who don't is striking, such that the high-spacing players, on average, perform less well initially, but because they learn at a faster rate their average rating exceeds that of the low-spacing players by game 50 . The effect of social play was not predicted: those who play more socially, as measured by their assist rate, learn slower - perhaps because the demands of team coordination distract from skill honing. Two other direct measures of exploration are not found to relate to rate of learning, in contrast to earlier results (Stafford et al., 2012; Stafford \& Dewar, 2014). This suggest that curiosity alone is not sufficient to enhance skill acquisition.

Destiny, and online games in general, represent a rich test-bed for theories of skill acquisition. Games are played for reasons of intrinsic motivation and so represent an important contrast to lab studies which are completed for financial incentives or as part of a course requirement. In addition they represent an opportunity to collect large data sets, which allow confidence in the estimates of the effects of the factors analysed. Overcoming statistical uncertainty allows researchers to move swiftly to wrestling with interpretative uncertainty.

In this case, although variability in practice timing - spacing - enhances skill acquisition, we failed to demonstrate that our measures of other kinds of practice variability can be related to enhanced skill acquisition. This leaves open the possibility that the exploration which supports skill acquisition is not captured by our measures, or that more complex skills, such as Destiny playing, require an equal match of habitual practice and exploratory variability.

## References

Algoet, P. H., \& Cover, T. M. (1988). A sandwich proof of the shannon-mcmillan-breiman theorem. Annals of Probability, 899-909.
Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., \& Pashler, H. (2008). Spacing effects in learning a temporal ridgeline of optimal retention. Psychological science, 19(11), 1095-1102.
Charness, M. K. R. R. E. V. E., N.; Tuffiash. (2005). The role of deliberate practice in chess expertise. Applied Cognitive Psychology, 151-165.

Delaney, P. F., Verkoeijen, P. P., \& Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. Psychology of learning and motivation, 53, 63-147.
Drachen, A., Green, J., Gray, C., Harik, E., Lu, P., Sifa, R., \& Klabjan, D. (2016). Guns and guardians: Comparative cluster analysis and behavioral profiling in destiny. In Proc. of the ieee cig.
Gallistel, C. R., Fairhurst, S., \& Balsam, P. (2004). The learning curve: implications of a quantitative analysis. Proceedings of the national academy of Sciences of the united States of america, 101(36), 13124-13131.
Goldstone, R. L., \& Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. Topics in cognitive science, 8(3), 548568.

Gray, W. D., \& Lindstedt, J. K. (2016). Plateaus, dips, and leaps: Where to look for inventions and discoveries during skilled performance. Cognitive Science.
Herbrich, R., \& Graepel, T. (2006). Trueskill: A bayesian skill rating system. Microsoft Research Tech. Rep. MSR-TR-2006-80.
Humphries, M. D., Khamassi, M., \& Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. Frontiers in neuroscience, 6, 9.
Lessne, A. (2014). Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. Mathematical Structures in Computer Science.
Magill, R. A., \& Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. Human movement science, 9(3), 241-289.
Newell, A., \& Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.), Cognitive skills and their acquisition (pp. 1-55). Lawrence Erlbaum.
Newell, K. M., \& McDonald, P. V. (1992). Searching for solutions to the coordination function: Learning as exploratory behavior. In G. E. Stelmach \& J. Requin (Eds.), Tutorials in motor behavior (Vol. 2, pp. 517532). North-Holland.

Nunneley, S. (2016). Activision q1: Destiny has nearly 30 m players; 55 m monthly players for all ip. VG24/7. Retrieved from www.vg247.com/2016/05/05/activision-q1 -destiny-has-nearly-30m-players-online -player-community-hits-55m/
Ranganathan, R., \& Newell, K. M. (2010). Motor learning through induced variability at the task goal and execution redundancy levels. Journal of motor behavior, 42(5), 307-316.

Rattinger, A., Wallner, G., Drachen, A., Pirker, J., \& Sifa, R. (2016). Integrating and inspecting combined behavioral profiling and social network models in destiny. In Springer lncs vol. 9926 (p. 77-89).
Ritter, F. E., \& Schooler, L. J. (2001). The learning curve. In N. Smelser \& P. Baltes (Eds.), International encyclopedia of the social \& behavioral sciences (pp. 8602-8605). New York: Elsevier.
Rosenbaum, D. A., Carlson, R. A., \& Gilmore, R. O. (2001). Acquisition of intellectual and perceptualmotor skills. Annual review of psychology, 52(1), 453470.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. Psychological review, 82(4), 225260.

Stafford, T., \& Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. Psychological science, 25(2), 511-518.
Stafford, T., \& Haasnoot, E. (2017). Confirming and quantifying sleep consolidation in skill learning: a field study using an online game. Topics in Cognitive Science.
Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., . . . Redgrave, P. (2012). A novel task for the investigation of action acquisition. PloS one, 7(6), e37749.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge University Press.
Tammasia, M., Raffe, W., Sifa, R., Drachen, A., Zambetta, F., \& Hitchens, M. (2016). Predicting player churn in destiny: A hidden markov models approach to predicting player departure in a major online game. In Proc. of the ieee cig.
Thompson, J. J., Blair, M. R., Chen, L., \& Henrey, A. J. (2013). Video game telemetry as a critical tool in the study of complex skill learning. PloS one, 8(9), e75129.
Thompson, J. J., McColeman, C. M., Stepanova, E. R., \& Blair, M. R. (2017). Using video game telemetry data to research motor chunking, action latencies, and complex cognitive-motor skill learning. Topics in Cognitive Science.
Van Rossum, J. H. (1990). Schmidt's schema theory: The empirical base of the variability of practice hypothesis: A critical analysis. Human Movement Science, 9(3), 387-435.

## Acknowledgements

The authors would like to extend their sincere gratitude to Bungie for making telemetry from Destiny available. Thanks also to four anonymous reviewers.

# Biases and labeling in iterative pragmatic reasoning 

Jon Scott Stevens (stevens.400@osu.edu)<br>Department of Linguistics, The Ohio State University<br>Columbus, OH 43210 USA


#### Abstract

This paper presents a series of reference game experiments (Frank and Goodman, 2012) and fits the results to a number of Bayesian computational models in order to explore the role of linguistic and perceptual bias in iterative pragmatic reasoning. We first discuss the modeling choices made by Franke and Jäger (2016) and others who have used similar frameworks to model reference game tasks. We introduce a space of different plausible Bayesian models based on this work, and compare models' fit to new experimental data to replicate the basic findings of Franke and Jäger (2016) regarding the strong role for perceptual salience (e.g., the primacy of color over shape as a differentiating property for possible referents) and linguistic category (e.g., a preference for nouns over adjectives) in pragmatic reference resolution. We then uncover an additional possible effect of what we call labeling, whereby a hearer may simply ignore non-salient, non-differentiating semantic properties, in a manner similar to how an incremental algorithm (Reiter and Dale, 1992) might ignore certain semantic properties when generating referring expressions. Keywords: Iterative pragmatic reasoning; probabilistic pragmatics; reference games; computational modeling; perceptual bias; reference resolution


## Introduction

When someone says, "hold up your finger," you are most likely inclined, without much thought, to hold up your index finger. This may seem unsurprising, as your index finger is particularly salient for a number of reasons. But, as Franke and Degen (2016) point out, further reflection raises the question of why the thumb-which is technically a finger, and which we might expect to be even more salient than the index finger-is never a candidate for reference. The thumb is a prime example of a pragmatic 'blocking' effect: though it is indeed a finger, the existence of the more specific word "thumb" tends to block it from reference by the word "finger". Hence, there is a tension between salience and pragmatic blocking in resolving the referent of "your finger".

This paper presents an exploration of this kind of tension using reference games (Frank and Goodman, 2012; Franke and Degen, 2016, and others). Reference games are communicative tasks where subjects are asked to either produce or interpret short utterances, which are potentially ambiguous in the context, to describe shapes on a screen. Reference games are used as a test of models of iterative pragmatic reasoning, whereby certain potential referents of an utterance are blocked by the existence of a better, more informative alternative utterance available to the speaker.

We further probe work begun in Franke and Jäger (2016) and Stevens (2016) by setting up reference games that favor strong biases toward particular visually salient referents. We test a range of different variants of the Rational Speech Act (RSA) model of Frank and Goodman (2012) on our results. We come to two conclusions:

- We replicate the basic findings of Franke and Jäger (2016), while improving their implementation of RSA by reducing the number of free parameters required from four to one.
- We examine variation between items and uncover a possible effect of what we call labeling-an independently motivated mechanism for assigning possibly incomplete semantic labels to potential referents based on salient preferred properties. We show that by introducing labeling into the model, the fit between model predictions and empirical results is improved.
Before diving into these results, we review prior work on reference games, RSA models and bias in reference resolution.


## Prior work

A recent movement toward probabilistic pragmatics-the use of Bayesian, game-theoretic and other similar methods to model how non-literal meaning is conveyed by utterances in context- has been accompanied by an emphasis on using computational models of pragmatic reasoning to explain empirical results (see Franke and Jäger, 2016, for a summary). This includes the rational speech act (RSA) model (Frank and Goodman, 2012; Franke and Jäger, 2016; Bergen et al., 2016, among many others) and its variants, as well as game-theoretic and decision-theoretic models (see e.g. Franke, 2009; Stevens, 2016). These frameworks all tell a similar story at their core: pragmatic phenomena are largely a byproduct of iterated reasoning of the form, 'I expect that she expects that I will say $\phi$ in context $C$,' or some variant.

Reference games A reference game task (Frank and Goodman, 2012) is a simple experiment which is designed to elicit iterative pragmatic reasoning behavior. A speaker and a hearer are presented with an array of colored and/or patterned shapes like the one seen in Fig.1. The speaker is assigned one of the shapes and is tasked with choosing a single word to convey to the hearer which shape she has been assigned. For Fig. 1 the choices would be "circle," "triangle," "blue" and "red." The hearer receives one of these words from the speaker and tries to guess correctly what was meant. A simple game-theoretic model of Gricean pragmatic reasoning, such as Franke's (2009) iterated best response (IBR) model, makes categorical predictions about the interpretation of ambiguous-in-context words ("triangle" and "blue" in this case). Quite simply, the hearer should assume that the speaker would have used an unambiguous word if she could have, i.e. "red" for the red triangle and "circle" for the blue circle, which leads to the conclusion that either "blue" or "triangle" alone should be taken to refer to the blue triangle. But such categorical models


Figure 1: Simple three-image setup for a reference game task.
are typically meant to be normative, and do not aim to reflect the probabilistic nature of how people actually behave. The RSA approach, which builds a Bayesian probabilistic component into a bounded IBR-style reasoning model (Frank and Goodman, 2012; Franke and Jäger, 2016) allows for computational models that more closely match experimental results.

Rational speech acts The rational speech act (RSA) approach to modeling pragmatic reasoning computes the probability of a hearer choosing a referent $r$ given a description $d$ via Bayes' rule assuming the speaker chose their utterance rationally, which in this case means the speaker attempted to maximize the chance of successful communication. We begin with a function encoding likelihood of referential success of a description $d$ given intended referent $r$ assuming a naive hearer-a hearer who randomly selects a referent consistent with $d$ 's denotation. Let's call this function $\mathcal{H}_{0}$.

$$
\begin{equation*}
\mathcal{H}_{0}(r \mid d)=\frac{1}{|\llbracket d \rrbracket|} \text { if } r \in \llbracket d \rrbracket \text {, else } 0 \tag{1}
\end{equation*}
$$

The probability of a rational speaker producing description $d$ to describe intended referent $r$ is taken to be a function of $\mathcal{H}_{0}$, namely a soft max function, which has the effect of approximately maximizing $\mathcal{H}_{0}$ by introducing a rationality parameter, $\lambda$. The rationality parameter encodes the degree to which speakers behave as perfect reasoners. As the value of $\lambda$ increases, this production probability-which we'll call $S_{1}$ asymptotically approaches an $\arg$ max of $\mathcal{H}_{0}$. Similarly to Franke and Jäger (2016), we also posit that a bias function, $\beta(d, r)$ is added to $\mathcal{H}_{0}$, which encodes possible prior bias toward certain types of descriptions over others (e.g., a preference for nouns over adjectives, empirically determined for our models). We'll return to the exact nature of the bias function in the next section.

$$
\begin{equation*}
S_{1}(d \mid r)=\frac{e^{\lambda \mathcal{H}_{0}(r \mid d)+\beta(d, r)}}{\sum_{d^{\prime}} e^{\lambda \mathcal{H}_{0}\left(r \mid d^{\prime}\right)+\beta\left(d^{\prime}, r\right)}} \tag{2}
\end{equation*}
$$

Finally, the production probability $\mathcal{S}_{1}(d \mid r)$ can be plugged in to Bayes' rule, where $P(r)$ is the prior probability of referent $r$ being referred to, to obtain a pragmatically motivated probability function for the hearer, which we will call $\mathcal{H}_{2}$.

$$
\begin{equation*}
\mathcal{H}_{2}(r \mid d)=\frac{S_{1}(d \mid r) \times P(r)}{\sum_{r^{\prime}} S_{1}\left(d \mid r^{\prime}\right) \times P\left(r^{\prime}\right)} \tag{3}
\end{equation*}
$$

We will use $S_{1}$ to make predictions about production probability and $\mathcal{H}_{2}$ to make predictions about interpretation probability. We use empirically determined values of $P(r)$.

Franke and Degen (2016) also implement a variant of RSA that starts the iteration with the speaker instead of the hearer. That variant begins with a 'literal speaker', which we could call $S_{0}$, who randomly selects from among appropriate descriptions. Then $\mathcal{H}_{1}$ selects referents that maximize the probability of having been referred to by $\mathcal{S}_{0}$ 's utterance, and then a pragmatic speaker $S_{2}$ chooses descriptions via Bayes' rule taking $\mathcal{H}_{1}$ into account. We implement this variant as well.

Biases and salience Cognitively oriented pragmatic models like RSA must take into account the prior biases that interlocutors bring to the table. Two such biases factor into Franke and Jäger's model of reference games. Firstly, the authors use data from a prior elicitation task to show that hearers have a prior bias toward picking referents that are more visually salient. For example, there is expected to be a bias toward the red triangle in Fig. 1 due to the pop-out effect that its unique color creates. Secondly, the authors use production experiment data to show that there is a prior bias toward using shape nouns rather than color-denoting adjectives to describe an intended object. These biases are built into their probabilistic model, the former being encoded in the prior probability distribution over speaker intentions, and the latter being encoded as the bias parameter $\beta$ which boosts production probability for shape terms. This allows a closer fit to experimental results when compared to more purely Gricean models.

Investigating perceptual bias in the visual domain can shed light on the role of salience in iterative pragmatic reasoning more generally, given that parallels have been found between visual salience and e.g., the use of definite referring expressions (Duan et al., 2013). In this study we find evidence that visual salience affects how hearers assign their own internal semantic labels to the potential referents in a scene. Namely, behavior on certain experimental items suggests that hearers selectively consider properties of potential referents (i.e., the object's color, shape, pattern, etc.) which serve to differentiate them from their competitors. More specifically, we suggest that hearers can generate sets of expected linguistic descriptions for each object using something like an incremental algorithm (Reiter and Dale, 1992; Krahmer and Van Deemter, 2012), which has been used to generate referring expressions in a psychologically plausible way. This algorithm is informally sketched in Fig.2. To illustrate, consider the picture in Fig.1. Imagine that the most salient property type is COLOR. To label the red triangle, the algorithm takes its value for COLOR-'red'-and checks whether there is at least one member of the distractor set (the blue triangle and the blue circle) which is not red. There is, and so 'red' gets added to the label set, and both of the non-red items are removed from the distractor set. This leaves an empty distractor set, and so the algorithm halts on the singleton set of labels, \{'red'\}. The same algorithm will generate $\{$ 'blue', 'triangle'\} for the

- Given an object $O$ in a set of objects $\Omega$, let $L$ be $O$ 's label-a set of semantic properties (e.g., \{'red', 'triangle'\}) to describe $O$. Let $P^{*}$ be an ordered sequence of property types which are ordered by salience (e.g. 〈COLOR, SHAPE〉, if color is more salient than shape). Let $D$ be the set of distractors, i.e., $\Omega \backslash O$
- Initialize $L$ to $\}$
- For $P$ in $P^{*}$ :

1. Let $V$ be the value that $O$ has for property $P$
2. Let $\Omega_{\neg V}$ be the set of objects that have a different value for $P$ than $V$
3. If $D \cap \Omega_{\neg V} \neq\{ \}$, then add $V$ to $L$ and remove all members of $\Omega_{\neg V}$ from $D$
4. If $D=\{ \}$, return $L$

Figure 2: An informal presentation of an incremental algorithm for generating salient and informative semantic labels for referents.
blue triangle and \{'blue', 'circle'\} for the blue circle. The red triangle is simply labeled as the red thing, because 'red' is a preferred property that uniquely differentiates it, while the other two shapes are labeled according to both color and shape.

## Computational models

We implement a variety of models centered around the RSA implementation of Franke and Jäger (2016), though we reduce the number of free parameters from four to one. The first reduction comes from the choice to use a single value of the rationality parameter $\lambda$ to predict both speaker and hearer behavior, where Franke and Jäger (2016) fit two $\lambda$ values separately. The second reduction comes from the use of empirical data to determine values of $\beta(d, r)$-the observed bias toward nouns in production-where Franke and Jäger (2016) use a pair of fixed values that were tweaked for best performance. Using a speaker norming task, as described in the next section, we obtain a prior probability of noun vs. adjective for each experimental item we want to model. We then set $\beta(d, r)$ to be proportional to this prior probability.

$$
\begin{equation*}
\beta(d, r)=\frac{P(d)}{\sum_{d^{\prime} \mid r \in \llbracket d^{\prime} \rrbracket} P\left(d^{\prime}\right)} \text { if } r \in \llbracket d \rrbracket \text {, else } 0 \tag{4}
\end{equation*}
$$

We now have a single-parameter RSA model that will make predictions about both production and interpretation rates in a reference game task, taking biases into account.

We implement a number of variants of this model to allow us to assess some of the modeling choices we and others have made. In particular, we want to answer the following three questions about our modeling choices:

1. Bias vs. no bias: Do we really need the $\beta(d, r)$ term?
2. Naive hearer vs. literal speaker: Should we really start with a naive hearer $\mathcal{H}_{0}$, as opposed to with a literal speaker $\mathcal{S}_{0}$ à la the variant in Franke and Degen (2016)?
3. Uniform level-0 prior vs. empirical level-0 prior: Should the naive hearer and/or literal speaker select randomly from

Uniform level $0 \quad$ Empirical level 0

| $S_{1} / \mathcal{H}_{2}$, no bias | F\&G (2012) |  |
| :--- | :---: | :--- |
| $\mathcal{S}_{1} / \mathcal{H}_{2}, \mathcal{S}_{1}$ bias | $\mathrm{F} \& \mathrm{~J}(2016)$ |  |
| $\mathcal{S}_{2} / \mathcal{H}_{1}$, no bias | $\mathrm{F} \& \mathrm{D}(2016)$ |  |
| $\mathcal{S}_{2} / \mathcal{H}_{1}, \mathcal{H}_{1}$ bias |  |  |
|  |  |  |

Table 1: Eight possible model variants based on the three questions posed, where three of the cells are occupied by examples of a model of that type-Frank and Goodman (2012), Franke and Jäger (2016) and Franke and Degen (2016).
among semantically appropriate actions, as opposed to selecting proportionally to the empirically determined prior?
There are a total of $2^{3}=8$ possible combinations of yes/no answers to these three questions, each corresponding to a different model variant. The variant we have described, based on Franke and Jäger (2016), is the "yes/yes/yes" model. That means that bias is implemented, we follow the trajectory $\mathcal{H}_{0} \rightarrow \mathcal{S}_{1} \rightarrow \mathcal{H}_{2}$ rather than starting with a non-rational speaker, and $\mathcal{H}_{0}$ chooses a random semantically compatible meaning, ignoring the prior probability $P(r)$. Table 1 lays out the possibilities and points to examples of a few of the models from the RSA literature.

We implemented all models with integer $\lambda$ values between one and ten ${ }^{1}$ and used root mean square error (RMSE) as a measure of overall difference between model predictions and experimentally determined values.

## Experiments

Participants and materials We conducted four experiments via Amazon Mechanical Turk, two experiments to determine prior probabilities for referents and descriptions, and two reference game tasks, one where the Turker played the part of the speaker and one where they played the part of the hearer. For each experiment, 100 Turkers were assigned to one of two lists containing nine experimental items, for a total of 18 items. Each item was an array of three images similar to Fig.1, where one image was distinguished from the other two by its shape, another image was distinguished by another salient attribute, and the third image was not distinguished along any dimension. The order of item presentation was randomized, as was the order in which the shapes were presented on the screen. Items fell into one of three categories based on which salient distinguishing attribute was used, with 6 items in each category:

1. Color: Red vs. blue, as in Fig. 1
2. Pattern: Striped vs. solid (one striped and two solid)
3. Size: One shape bigger than the other two

Native language was assessed as part of a post-task questionnaire. Subjects were paid $\$ 0.70$ for about 5 minutes of their time. If any subject's responses were incomplete, or if the

[^507]subject was not a native speaker of English, the data from that subject was excluded from analysis.

Experiment 1: Eliciting hearer priors Following Frank and Goodman (2012) and others, we use empirically determined values for the prior probabilities in our model. The prior probability $P(r)$ of choosing a referent $r$ is taken to be a measure of the salience or 'newsworthiness' of a referent, i.e., a general measure of how likely $r$ is to be talked about. To elicit this experimentally, we asked subjects to select an image to describe, giving them no guidance on which images to select, and then type a description of it. The point of this experiment was not what the descriptions were, but rather which shapes they chose to talk about. We took this as a proxy for the salience of the referent, and thus its prior probability of being referred to. We asked them to type descriptions as a secondary task in order to situate the shape selection within a natural communicative setting. We used data from 97 subjects after exclusions.

For the color items we obtained similar results to Franke and Jäger (2016), where the red shape was picked much more often (probability 0.5 ) than either of the blue shapes, and where the distinguished blue shape (e.g., the circle in Fig.1) was picked more often (0.33) than the non-distinguished blue shape (0.17). For the size items, we found that the distinguished smaller shape had a high prior probability ( 0.5 vs . 0.26 and 0.24 for the large and small competitors, respectively), and for the pattern items, the priors were closer to equal for the striped and distinguished solid shapes ( 0.36 and 0.40 , respectively), and lowest for the non-distinguished shape (0.24).

Experiment 2: Eliciting speaker priors To empirically determine whether and to what extent speakers are biased toward nouns like 'circle' over adjectives like 'red', we ran an experiment just like Experiment 1, but with two important differences: (i) subjects were assigned one of the three images to describe, rather than being asked to pick one themselves (image assignments were counterbalanced across lists so that shape-distinguished, attribute-distinguished and nondistinguished items were equally represented), and (ii) subjects were told to limit their descriptions to a single word, in order to bring the task more in line with a reference game task. To discourage any kind of pragmatic reasoning, subjects were asked to use the 'first word that came to mind' and not to overthink it. We analyzed data from 84 subjects after exclusions, and only looked at items where either a shape-denoting noun or relevant attribute-denoting adjective was used (very few did not fall into this category). The words were input as free text, and thus we hand-tokenized the responses to account for spelling mistakes and superficial lexical differences (e.g., 'big' vs. 'large'). We found an overwhelming prior bias toward nouns.Overall, shape terms were used two-thirds of the time. There is evidence that this task successfully elicited prior linguistic biases and limited the amount of pragmatic

|  | Word |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| Image | $\mathrm{ATTR}_{D}$ | $\mathrm{ATTR}_{N}$ | $\mathrm{SHAPE}_{N}$ | $\mathrm{SHAPE}_{D}$ |
| $\operatorname{ATTR}_{D} \mathrm{SHAPE}_{N}$ | $.75 / \mathbf{1}$ | $.00 / \mathbf{0 0}$ | $.25 / . \mathbf{5 2}$ | $.00 / \mathbf{0}$ |
| $\operatorname{ATTR}_{N} \operatorname{SHAPE}_{N}$ | $.00 / \mathbf{0}$ | $.29 / . \mathbf{7 4}$ | $.71 / . \mathbf{4 8}$ | $.00 / \mathbf{0}$ |
| $\operatorname{ATTR}_{N} \operatorname{SHAPE}_{D}$ | $.00 / \mathbf{0}$ | $.03 / . \mathbf{2 6}$ | $.00 / \mathbf{0 0}$ | $.97 / \mathbf{1}$ |

Table 2: Production of $d$ given $r$ (on the left, sum horizontally to 1 ) / selection of $r$ given $d$ (on the right in bold, sum vertically to 1 ) in Experiments 3 and 4. Subscripts $D$ and $N$ mean 'distinguishing' and 'non-distinguishing', respectively, and ATTR stands for 'attribute'.
reasoning being used to determine descriptions. For example, for the items where an attribute term would uniquely distinguish the intended referent, shape terms nonetheless comprised $60 \%$ of responses, more than double the shape-term response rate for Experiment 3, which was designed to elicit pragmatic reasoning.

Experiment 3: Reference game, speaker role Experiments 3 and 4 instantiate the canonical reference game task described in the second section. Experiment 3 asks subjects to play the role of the speaker in a reference game. Similarly to Experiment 2, subjects are assigned one of the three images and asked to give a one-word description. But for this experiment, they are explicitly told to select from a list of the relevant words (e.g., "red", "blue", "triangle", "circle"). And unlike Experiment 2, the task is framed as a game. Subjects are told they are sending a message, and to assume that a "receiver" will receive these descriptions and make a guess as to which image was assigned. The goal, they are told, is for the receiver to guess correctly as often as possible. Data from 79 subjects was used.

Experiment 4: Reference game, hearer role Experiment 4 asks subjects to play the role of the hearer, or the "receiver". For each item, a single word is displayed at the top of the screen, which the subjects are told has been carefully selected and sent to them by a sender who wants them to correctly guess an image from a one-word description. Word selection was counterbalanced across both lists. The subjects were required to select a single image for each item. Data from 91 subjects was used.

Results Our reference game experiment results are in line with other reference game results in the literature, and are summarized in Table 2. Like Franke and Jäger (2016), we find that the expected propensity toward interpreting ambiguous shape and attribute words (like "triangle" and "blue" in Fig.1) as referring to the non-distinguished shape (like the blue triangle) is dampened for the shape words, likely reflecting hearer knowledge of speakers' prior noun bias, where the prior noun bias makes a shape term like "triangle" a less reliable signal that the non-distinguished referent is intended.

|  | Uniform level 0 | Empirical level 0 |
| :---: | :---: | :---: |
| $S_{1} / \mathcal{H}_{2}$, no bias | $.12 / .20$ | $.14 / .23$ |
| $S_{1} / \mathcal{H}_{2}, S_{1}$ bias | $\mathbf{. 0 6} / \mathbf{1 8}$ | $.09 / .22$ |
| $\mathcal{S}_{2} / \mathcal{H}_{1}$, no bias | $.23 / .20$ | $.23 / .20$ |
| $S_{2} / \mathcal{H}_{1}, \mathcal{H}_{1}$ bias | $.25 / .22$ | $.24 / .23$ |
|  |  |  |

Table 3: RMSE for speaker predictions (left) / hearer predictions (right). Best-case $\lambda$ value used for each reported RMSE. Best model results are in bold.

Table 3 shows how our model variants line up with the empirical results in terms of the root mean square error (RMSE), which is a measure of overall difference between predicted and observed values obtained by calculating the mean of the square of the difference between each predicted vs. observed value and taking the square root. We used the difference in predicted vs. observed subject means for each experimental item (i.e., each array of images) to determine RMSE. Not only is our refinement of Franke and Jäger (2016) the best model to predict these data, but our best-case value of the $\lambda$ parameters $(\lambda=4)$ is the best-case value for both the speaker and hearer model independently. That is to say, we would not do a lot better by allowing for separate $\lambda$ values for speaker and hearer. We take this to be a nice replication of the basic finding of Franke and Jäger (2016), obtained using only one free parameter that was only broadly tweaked. ${ }^{2}$

The numbers in Table 2 are somewhat closely replicated, with every value being within three percentage points of the real value. A plot of predicted vs. actual results from Table 2 is given in Fig.3. However, the numbers in Table 2 are averaged over all items, and tell us nothing about the range of variation of responses for different kinds of images. RMSE gives us an overall assessment of error taking into account error at the level of each individual item. What the RMSE values in Table 3 tell us is that the speaker model fits considerably better than the hearer model.

Why is the hearer model so noisy? Given the proximity of predicted to actual results on average in Fig.3, the source of the noisiness must be coming from differences between item types. An item-level investigation of the source of the higher-than-expected RMSE will lead us to posit that when there are highly perceptually salient options, as in these experiments, hearers are inclined to label their options in a way that is similar to the output of an incremental algorithm (Fig.2).

## Labeling

We now break down by-item behavior further, looking not only at whether the image array was shape- color- or sizedistinguished, but also at which word was sent to the hearer. We find that the predicted qualitative pattern-that ambiguous descriptions (and only ambiguous descriptions) should prompt a plurality of guesses of the non-distinguished

[^508]

Figure 3: Hearer predictions vs. observed values, averaged over subjects and items.
image-holds in all but two cases. These two cases are depicted in Fig.4, where we see a deviation for (i) color items when the hearer is sent an ambiguous color term and (ii) size items when the hearer is sent an ambiguous shape term. There is a pattern to these deviations. The pattern is that we see a shift away from the non-distinguished image only in cases where the semantically ruled out referent (e.g., the red thing if the description is "blue") has a high prior (in both cases, $\sim 50 \%$ ). Let's break down what this means for the three item types. First, when the hearer receives "blue" for an item like Fig.1, we find higher-than-expected selection of the unique shape (the circle in Fig.1). This is the item type for which the attribute-distinguished image (the red triangle) is maximally salient according to Experiment 1. Second, when a hearer receives an ambiguous shape term for a size-distinguished item, we find higher-than-expected selection of the uniquely large referent. This is the item type for which the shapedistinguished image is maximally salient according Experiment 1. Finally, the pattern-distinguished items fall entirely in line with what we expect, and those are the items where the priors for shape- and attribute-distinguished images are much closer to each other.

Qualitatively speaking, we would expect this if the referents were labeled according to salient distinguishing properties along the lines of Fig.2, a well-established algorithm for generating referring expressions, which we adapt for generating hearer-internal labels for possible referents. Consider Fig. 1 one more time: for the $\sim 50 \%$ of subjects in Experiment 1 who chose the red triangle, we can assume that COLOR would be their primary salient property type for purposes of Fig.2. This would generate the labels \{ 'blue circle', 'blue triangle', 'red'\}. Assuming these same priors for Experiment 4 (as we have been) we could posit that on $\sim 50 \%$ of trials, the subject has this same labeling. In that case, upon hearing the description "blue", the subject would be at chance between the two blue shapes, because under this labeling, the speaker could have used 'triangle' to uniquely describe the blue triangle and 'circle' to uniquely describe the blue circle, leaving no principled way to interpret "blue" other than to guess.

Labeling could explain the qualitative deviations, and even though the numbers are not perfect, it does indeed improve model fit to add a labeling component to the model. We can do this by substituting a new $\mathcal{S}_{1}$ function $\mathcal{S}_{1}^{\prime}$ into the $\mathcal{H}_{2}$ equa-


Figure 4: Predicted (left) and actual (right) referent selection for two combinations of item type and description type.
tion which takes labeling into account. Letting $\mathcal{L}$ be the set of labels for each possible referent, $S_{1}^{\prime}$ can be defined as follows:

$$
\begin{equation*}
\mathcal{S}_{1}^{\prime}(d \mid r)=\sum_{\mathcal{L}} P(\mathcal{L}) \times \frac{e^{\lambda \mathcal{H}_{0}(r \mid d, \mathcal{L})+\beta(d, r)}}{\sum_{d^{\prime}} e^{\lambda \mathcal{H}_{0}\left(r \mid d^{\prime}, \mathcal{L}\right)+\beta\left(d^{\prime}, r\right)}} \tag{5}
\end{equation*}
$$

We introduce no new free parameters if we simply take $P(\mathcal{L})$ to be the prior probability of the shape-distinguished referent for the $\mathcal{L}$ obtained when shape is primary, the prior of the attribute-distinguished referent for the $\mathcal{L}$ obtained when attribute is primary, and the prior of the non-distinguished shape for the 'full' $\mathcal{L}$, which omits no information. Doing this, we can reduce the RMSE from 0.18 to 0.15 , and could perhaps reduce it further if we could independently assess how primary salient properties are chosen. Using multinomial choice probabilities to determine log likelihood, we can show that the data from Experiment 4 are significantly more likely under the model with labeling. ${ }^{3}$

## Conclusion

Using a variety of models and experimental items and tasks, we have replicated existing results regarding behavior in reference games, and potentially found a new one, an effect of labeling under conditions where certain referents have highly salient properties. That it might matter how people internally label possible referents is not really a new idea, and is in fact in line with game-theoretic literature on coordination (see e.g., Sugden, 1995). But it provides somewhat of a paradox. On the one hand, this and other studies find that speakers exhibit a bias toward noun descriptions in reference games, across the board, and yet it seems as if hearers are assigning labels to potential referents that in some cases would lead them to expect the opposite (e.g., to expect "red" to describe the red triangle in Fig.1). Thus further work is warranted to probe whether such a mismatch between speaker behavior and hearer expectations is generally observable.

[^509]Further work must also be done to probe the details of exactly how labeling works in reference games, and what the implications are for iterative pragmatic reasoning more generally. For example, it remains to be seen whether labeling should be seen as part of a rational process of pragmatic reasoning, or as something that competes with it, as Stevens (2016) would suggest. Finally, future work will use online measures to probe the mechanisms that give rise to the probabilities in our models. This would take us beyond computational-level models, using such models only as a starting point to guide us toward a more fine-grained understanding of this behavior (see Yang, to appear).

## Acknowledgments

This work was funded by the American Council of Learned Societies (ACLS). Thanks to Marie-Catherine de Marneffe, Micha Elsner and the OSU psycholinguistics lab group.

## References

Bergen, L., Levy, R., and Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. Semantics and Pragmatics, 9. Advance online publication.
Duan, M., Elsner, M., and de Marneffe, M.-C. (2013). Visual and linguistic predictors for the definiteness of referring expressions. In Proceedings of the 17th SemDial Workshop, Amsterdam.
Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084):998.
Franke, M. (2009). Signal to Act: Game Theory in Pragmatics. PhD thesis, Universiteit van Amsterdam.
Franke, M. and Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. PloS one, 11(5).
Franke, M. and Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. Zeitschrift für Sprachwissenschaft, 35(1).
Krahmer, E. and Van Deemter, K. (2012). Computational generation of referring expressions: A survey. Computational Linguistics, 38(1):173-218.
Reiter, E. and Dale, R. (1992). A fast algorithm for the generation of referring expressions. In Proceedings of the 14th conference on Computational Linguistics, Volume 1, pages 232-238. ACL.
Stevens, J. (2016). When do we think strategically? Zeitschrift für Sprachwissenschaft, 35(1).
Sugden, R. (1995). A theory of focal points. The Economic Journal, 105:533-550.
Yang, C. (To appear). Rage against the machine: Evaluation metrics in the 21st century. Language Acquisition.

# Actions that Modify Schedules of Reinforcement 

Mac Strelioff (mstrelio@uci.edu)<br>Mimi Liljeholm (m.liljeholm@uci.edu)<br>Department of Cognitive Sciences, University of California Irvine<br>Irvine, CA 92697-5100 USA


#### Abstract

Many everyday activities involve the use of one action to modify the effects of another: When driving, shifting gears modifies the influence of pressing the gas pedal on acceleration; when cooking, the rate of adding a particular ingredient modifies the influence of stirring on viscosity. Here, we investigate a general ability to learn how to use actions to control schedules of reinforcement. In Experiment 1, participants quickly discovered the optimal rate of responding on an action that controlled the rate of reward contingent on performing a different action. In Experiment 2, when the modifying action was itself rewarded, participants failed to discover the optimal rate. Implications for formal theories of instrumental behavior are discussed.


Keywords: Schedules of reinforcement; reward learning; instrumental contingencies.

## Introduction

Since the early $20^{\text {th }}$ century, researchers have investigated the influence of various reward schedules on the rate and selection of instrumental responses. For example, ratio schedules, in which reward delivery depends on the number of responses since the last reward, produce higher rates of responding than do interval schedules, in which reward delivery depends on the time elapsed since the last reward (Fester \& Skinner, 1957). When two or more action alternatives are available, that which yields the greatest, most immediate, or most certain reward is, all other things being equal, generally that most frequently selected (e.g., Rachlin et al., 1991). However, in the real world, many responses serve only to modulate the effects of other actions: The rate and pattern of pressing strings on a guitar does not itself yield music, but profoundly impacts the sounds produced by strumming. Here, we assess a domain-general capacity for learning about actions that control schedules of reinforcement on other actions.

Formally, the relationship between a particular action and its outcome has been modeled as a complex associative structure (Dickinson \& Balleine, 1993), as the difference between probabilities of reward given the presence versus absence of the action (Hammond, 1980), as the probability and subjective utility of the outcome given the action (Savage, 1954), or as a cached value assigned to the action based on its reinforcement history (Watkins, 1989). What these diverse approaches have in common is that they address the identity and/or
latency of a single action at a time, ignoring situations in which multiple actions are performed in concert and potentially interact. In our paradigm, an intermediate rate of responding on one action maximizes the reward contingent on performing a different, concurrently available, action.

## Experiment 1

## Methods

Participants Thirty undergraduates at the University of California, Irvine ( 22 females; mean age $=20 \pm 2.17$ ) participated in the study for course credit. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

Task \& Procedure The task is illustrated in Figure 1. We used a free operant paradigm in which participants were allowed to respond at will on either or both of two concurrently available actions, graphically represented on the computer screen, by pressing the corresponding keys on the computer keyboard. Whenever a response was made a selection square appeared around the chosen action for 300 ms . If the response was rewarded, an image of a quarter appeared center screen for 500 ms and a count of the cumulative monetary earning, continuously displayed above the quarter image location, would increment by $+\$ 0.25$. The task was comprised of ten 2-minute blocks separated by brief rest periods. All monetary earnings were fictitious.

In the "Modify" group ( $\mathrm{n}=15$ ), the rate of responding on a "modifying" action influenced the probability that the concurrently available "modified" action would produce a reward. When the modifying action was performed at an "optimal" rate of 1.25 to 2.75 presses per second, the probability of reward given a response on the modified action was 0.9 . When response rates on the modifying action were outside of the 1.25 to 2.75 range, the probability of reward given the modified action was 0 . The modifying action did not itself produce any reward. Response rates on the modifying action were tracked using a differential equation that increased by an impulse of 1 at the time of a response and decayed each impulse at a linear rate of 0.2 per second, so that each impulse from a response decayed to zero after 5 seconds.

Specifically, for an impulse $\left(a_{i}\right)$, which was 1 if an action were taken during the current iteration of the program and 0 otherwise, a decay rate of 0.2 , a counter for the number of responses that occurred within the last 5 seconds $\left(N_{5}\right)$ and the difference in time between the current iteration of the program and the previous iteration $(d t)$, the response rate variable $(R)$ was updated on each iteration $i$ by:

$$
R_{i} \leftarrow R_{i-1}+a_{i}-0.2 N_{5} d t
$$

This method adjusts more quickly to changes in response rate than the commonly used approach of dividing the number of responses in a time window by the length of the window (e.g., Soto et al., 2006). The probability of reward on the modified action was set to 0.9 whenever the response rate variable, $R$, was in the optimal, 1.25 to 2.75 , range and 0.0 otherwise.

Note that the optimal rate of responding on the modifying action was intermediate; this was done to rule out the contribution of systematic biases of either very high or very low responding. On the other hand, an intermediate rate might represent an average towards which most responders converge in free operant tasks. To address this possibility, a second, "Yoked", group was included ( $\mathrm{n}=15$ ), in which the rate of responding on the modifying action had no influence, while the probability of reward on the modified action was yoked to that of a participant in the Modify group. We predicted that, by the end of the session, participants in the Modify group would respond on the modifying action at a rate falling within the optimal range, while those in yoked group would not.


Figure 1: Task Illustration, see text for details.

## Results

We divided the number of responses on the modifying action in each 5 -second bin of task performance with 5 (i.e., responses per second), and computed the distance of this response rate from the bounds of the optimal range, for the first and last 5 seconds of the task. We then used a mixed analysis of variance (ANOVA) with "group" as the between-subject factor and "bin" as the within-subject factor to assess a change in optimal responding between the first and last bins. There was no main effect of bin, $\mathrm{F}(1,28)=3.19, p=0.09$, but a main effect of group, $\mathrm{F}(1,28)=6.53, p<0.05$, and, critically, a
bin-by-group interaction, $\mathrm{F}(1,28)=5.78, \quad p<0.05$. Planned comparisons revealed that while the two groups did not differ with respect to optimal responding on the modifying action in the first bin, $\mathrm{t}(28)=0.13, p=0.89$, by the last bin, participants in the Modify group were significantly closer to the optimal response rate than were participants in the Yoked group, $\mathrm{t}(28)=3.61$, $p<0.01$. As can be seen in Figure 2, while the mean deviation from the optimal rate significantly decreased from the first to the last bin in the Modify group, $\mathrm{t}(14)=2.69, p<0.05$, they remained unchanged across bins in the Yoked group, $\mathrm{t}(14)=0.05, p=0.63$. The apparent absence of a change in optimal responding by Yoked participants reflects a tendency to either increase or decrease responding on the modifying action across blocks, resulting in no net change for the group; in contrast participants in the Modify group coherently converged towards the optimal rate.


Figure 2: Mean deviation of response rates on the modifying action from the optimal range in the first and last 5 seconds of task performance, for subjects in the Modify (black) and Yoked (blue) groups, and for the single group of Experiment 2, in which the modifying action was rewarded (red). Error bars=SEM.

We also assessed performance in terms of the proportion of bins with optimal response rates, early and late in the task. Bins were scored as optimal if the windowed (5 seconds) response rate was in the optimal range of 1.25 to 2.75 responses per second. For each subject, we assessed the number of optimal 5 -second bins in the first and last 30 seconds of the task. (We used 30 seconds, rather than the full 2 -minute blocks, to ensure that the index of early learning did not include already asymptotic performance.) The results using this metric were consistent with those described above: The groups did not differ in the first 30 -second block, $\mathrm{t}(28)=1.12$, $p=0.24$, but by the last 30 -second block, the mean proportion of optimal bins was significantly greater for the Modify group than for the Yoked group, $\mathrm{t}(28)=6.93$, $p<0.01$. Indeed, while the proportion of optimal bins increased significantly from the first to the last block in
the Modify group, $\mathrm{t}(14)=3.60, p<0.01$, it decreased, albeit with marginal significance, $\mathfrak{t}(14)=2.09, p=0.06$, in the Yoked group. The mean proportion of optimal bins in each 30 -second block throughout the task is shown in Figure 3.


Figure 3: Mean proportion of bins with an optimal response rate on the modifying action in each 30 -second block of the task for the Modify (black) and Yoked (blue) groups of Experiment 1, and for the Reward group of Experiment 2 (red). Shading=SEM.
With respect to the modified action, response rates were, overall, higher than those on the modifying action, for both the Modify, $\mathrm{t}(14)=3.10, p<0.05$, and Yoked, $\mathrm{t}(14)=6.76, p<0.05$, groups. This likely reflects the fact that, while the probability of reward given the modified action was either a function of (Modify) or independent of (Yoked) responding on the modifying action, the actual delivery of reward was contingent only on performing the modified action.

## Experiment 2

A well-studied phenomenon closely related to our query is that of "melioration" - a tendency to select an action alternative that produces a greater immediate pay-off, but that, when selected repeatedly, lowers the overall rate of reward (Herrnstein, 1991). Such tendencies are commonly attributed to impulsivity (Herrnstein, 1991; Otto, Markman, \& Love, 2012), but have also been described as rational choices under uncertainty (Gureckis \& Love, 2009a, 2009b; Sims et al., 2013). Other related paradigms, such as delay discounting (Ainslie, 1975; Johnson \& Bickel, 2002) and differential reinforcement of low response rates (Wilson \& Keller, 1953; Carter \& MacGrady, 1966), have convincingly demonstrated the interfering influence of salient reward on rational decision-making (Ainslie, 1975; Van den Broek, Bradshaw, \& Szabadi, 1987).

In Experiment 2, we assess whether the lure of an immediate reward results in a failure to suppress responding on the modifying action, thus interfering
with the ability to control the schedule of reinforcement on the modified action.

## Methods

Participants Fifteen undergraduates at the University of California, Irvine ( 10 females; mean age $=19.7 \pm 1.1$ ) participated in the study for course credit. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

Task \& Procedure Participants performed a task that was identical to that of the Modify group in Experiment 1, with one exception: In addition to modulating the schedule of reinforcement on the modified action, the modifying action was itself rewarded by $\$ 0.25$, with a probability of 0.2 . Note that, since this reward probability is much lower than the conditional, 0.9 , probability of reward on the modified action, maintaining an optimal, intermediate, response rate on the modifying action dramatically increases the average reward rate.

## Results

We computed the same measures of optimal responding as those used in Experiment 1. Comparing the first and last 5 seconds of performance, there was no change in the deviation of response rates on the modifying action from the bounds of the optimal rate, $\mathrm{t}(14)=0.48, p=0.64$ (see Figure 2). Likewise, the proportion of optimal bins did not differ between the first and last 30 -second blocks of the task, $\mathrm{t}(14)=0.00, p=1.00$. In the absence of random assignment, we refrain from making any statistical comparisons between the results of this experiment and those obtained in Experiment 1. Nonetheless, it is worth noting that, as illustrated in Figures 2 and 3, when the modifying action was itself rewarded, the rate of responding on the modifying action was apparently closer to that in the Yoked group than in the Modify group. Finally, although, overall, response rates were again higher on the modified than the modifying action, unlike for the groups in Experiment 1, this difference was only marginally significant, $\mathrm{t}(14)=2.09, p=0.06$, presumably reflecting the fact that, in Experiment 2, reward delivery was potentially contingent on performing either action.

## General Discussion

In two experiments, we assessed the discovery and performance of an action that controlled the schedule of reinforcement on another, concurrently available, action. In Experiment 1, participants quickly discovered and implemented an optimal, intermediate, response rate on a modifying action that, while not producing any rewards itself, modulated the reward contingent on a distinct, concurrently available, action. Response rates
in a yoked control group confirmed that convergence to the optimal rate was due to the influence of the modifying action on the reward schedule of the modified action. In Experiment 2, consistent with a large literature on the failure to suppress inappropriate responding in the face of immediate reward (Ainslie, 1975; Carter \& MacGrady, 1966; Van den Broek et al., 1987; Wilson \& Keller, 1953), reinforcement of the modifying action apparently prevented discovery of the optimal response rate. The focus in the existing literature on the disruptive effects of immediate reward has largely overshadowed the question raised here of whether, and how, agents learn about actions that modify schedules of reinforcement. Our results suggest that, in the absence of interfering or competing reward contingencies, increasing levels of instrumental control can be achieved by incorporating information about dependencies between actions.

In a model-free reinforcement learning account of free operant responding, Niv et al. (2007) proposed that, for each decision, the agent selects both the latency and the identity of the to-be-executed action, based on the relative degree to which that action increases the average reward rate. Although it is possible that participants in the Modify group of Experiment 1 similarly learned about the modifying action based on its reinforcement history, several aspects of our task depart from the specification of Niv et al. (2007). Most notably, participants in our task would have to include a representation of the modified action in their state space when updating the value of the modifying action - that is, assess the value of a particular latency of the modifying action given that the modified action is simultaneously ${ }^{1}$ or proximally performed - since the modifying action is never itself rewarded. Likewise, the value of the modified action has to be specified conditional on the performance of the modifying action, since the probability of reward on the former is zero whenever responding on the latter falls outside the optimal range. It is of course possible to specify a model-free learner that has enough conditionals built into its state-representation to identify the combination of responding on modifying and modified actions that maximizes reward ${ }^{2}$.

An alternative, model-based, approach is for the agent to create a graphical probabilistic model representing

[^510]the dependencies between actions, states and rewards (e.g., Acuna \& Schrater, 2010). Although initially ignorant of the nature of these dependencies, a Bayesian reinforcement learner generates beliefs over a set of possible dependency structures and updates those beliefs, after each observation, using Bayesian inference. For example, a learner in our task might postulate two possible worlds: one in which the latency to respond on an action can modulate the probability of reward given that same, or some other, action, and one in which response latencies have no influence on schedules of reinforcement. The former possibility must of course be further partitioned into several putative structures, each with a particular set of links (e.g., an action modifying its own probability of reward vs. that of a different action) and associated parameters. The learner then updates the belief distribution over structures based on sequences of actions, latencies and rewards.

Critically, the approach sketched in the previous paragraph, to address model-based inferences regarding action dependencies, can also be used to explain some of the most basic aspects of instrumental behavior, such as the distinction between interval and ratio schedules Recall that, whereas on interval schedules a response is rewarded based on the amount of time elapsed since the last reward, on ratio schedules a response is rewarded based on the number of responses since the last reward. These qualitatively different schedules produce distinct response profiles (Fester \& Skinner, 1957), suggesting some, implicit or explicit, discrimination by the agent. Notably, the interval schedule can be conceptualized as a case in which the rate of performing an action modifies the schedule of reinforcement, rather than just the rate of reward: Specifically, on a given interval schedule, any response rate greater than "one per the required interval" will decrease the probability of reward conditional on that action. Other wellestablished schedules, such as differential reinforcement of high or low responding (Van den Broek, et al., 1987) can also be characterized as actions modifying schedules of reinforcement, as can the "seeking" component of seeking-taking schedules (Balleine, Garner, Gonzalez, \& Dickinson, 1995). Thus, the framework proposed here potentially applies to a wide range of instrumental phenomena.

At the neural level, model-free and model-based RL approaches have been mapped to dissociable neural substrates, with the ventral striatum, posterior putamen and premotor cortex being implicated in model-free responding (Glascher et al., 2010; Lee et al., 2014; Tricomi et al., 2009; de Wit et al., 2012), and the caudate, ventromedial prefrontal cortex and inferior parietal lobule in model-based computations (de Wit et al., 2012; Liljeholm et al., 2011, 2013, 2015; Lee et al., 2014). It should be noted, however, that, with some exceptions (e.g., Liljeholm et al., 2013), the work
identifying such dissociations has focused on relatively simple model-based processes, such as the encoding of individual action-outcome contingencies or sensitivity to changes in an outcomes utility. In contrast, the model-based learner postulated here engages in complex reasoning regarding how actions may be used to control action-outcome relationships. Such processes may warrant the involvement of brain regions known to support relational and inductive reasoning, including the rostolateral and dorsolateral prefrontal cortex (e.g., Krawczyk et al., 2011).

Finally, an important point regarding action dependencies such as those addressed here is how they relate to the actual representations of actions. In our task, the instructions and materials clearly defined and distinguished between action alternatives (see Figure 1), so that there could be little doubt about how many, and exactly what, actions were available. It is interesting to consider how inferences and performance might have differed had the grouping of elements into discrete action alternatives been more ambiguous. One possibility is that increasing ambiguity would afford a more rapid acquisition of relevant dependencies (Pezzulo, Rigoli, \& Friston, 2015) and, further, that those inferred dependencies might serve to configure action elements into more clearly delineated action representations based on reinforcement learning principles (e.g. Reynolds, \& O’Reilly, 2009).

In conclusion, we have demonstrated a domaingeneral ability to learn about, and take advantage of, an action that modifies the schedule of reinforcement on a different action. We have also sketched a model that, by making inferences about dependencies between response latencies and conditional reward probabilities, might account for behavior across a wide range of instrumental schedules. Future work will focus on extensions of our experimental paradigm, further development of formal accounts, and investigations of mediating neural substrates.

## References

Acuna, D., \& Schrater, P. (2010). Structure learning in human sequential decision-making. PLoS Computational Biology, 6, 1-8.
Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. Psychological bulletin, 82(4), 463.
Balleine, B. W., Garner, C., Gonzalez, F., \& Dickinson, A. (1995). Motivational control of heterogeneous instrumental chains. Journal of Experimental Psychology: Animal Behavior Processes, 21(3), 203.
Carter, D. E. \& MacGrady, G. J. (1966). Acquisition of a temporal discrimination by human subjects. Psychonomic Science, 5, 309-310.
de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., van de Vijver, I., \& Ridderinkhof, K. R. (2012).
Corticostriatal connectivity underlies individual
differences in the balance between habitual and goaldirected action control. Journal of Neuroscience, 32(35), 12066-12075.
Dickinson, A., \& Balleine, B. (1993). Actions and responses: The dual psychology of behaviour. In N. Eilan, R. A. McCarthy, \& B. Brewer (Eds.), Spatial representation: Problems in philosophy and psychology (pp. 277-293). Malden: Blackwell Publishing
Fester, C. B., \& Skinner, B. F. (1957). Schedules of reinforcement. New York, NY: Appleton-CenturyCroft.
Gläscher, J., Daw, N., Dayan, P., \& O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron, 66(4), 585-595.
Gureckis, T. M., \& Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. Journal of Mathematical Psychology, 53(3), 180-193.
Gureckis, T. M., \& Love, B. C. (2009b). Short term gains, long term pains: How cues about state aid learning in dynamic environments. Cognition, 113(3), 293-313.
Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. Journal of the experimental analysis of behavior, 34(3), 297-304.
Herrnstein, R. J. (1991). Experiments on stable suboptimality in individual behavior. The American Economic Review, 81(2), 360-364.
Hogarth, L., Dickinson, A., Wright, A., Kouvaraki, M. \& Duka, T. (2007). The role of drug expectancy in the control of human drug seeking. Journal of Experimental Psychology-Animal Behavior Processes, 33(4), 484-496.
Johnson, M. W., \& Bickel, W. K. (2002).
Within-subject comparison of real and hypothetical money rewards in delay discounting. Journal of the experimental analysis of behavior, 77(2), 129-146.
Krawczyk, Daniel C., M. Michelle McClelland, and Colin M. Donovan. "A hierarchy for relational reasoning in the prefrontal cortex." Cortex 47.5 (2011): 588-597.

Lee, S. W., Shimojo, S., \& O’Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning.Neuron, 81(3), 687-699.
Liljeholm, M., Tricomi, E., O'Doherty, J. P., \& Balleine, B. W. (2011). Neural correlates of instrumental contingency learning: differential effects of actionreward conjunction and disjunction. Journal of Neuroscience, 31(7), 2474-2480.
Liljeholm, M., Wang, S., Zhang, J., \& O'Doherty, J. P. (2013). Neural correlates of the divergence of
instrumental probability distributions. Journal of Neuroscience, 33(30), 12519-12527.
Liljeholm, M., Dunne, S., \& O'doherty, J. P. (2015). Differentiating neural systems mediating the acquisition vs. expression of goal-directed and habitual behavioral control. European Journal of Neuroscience, 41(10), 1358-1371.
Niv, Y., Daw, N. D., Daphna, J., \& Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology, 191(3), 507520.

Otto, A. R., Markman, A. B., \& Love, B. C. (2012). Taking more, now: The optimality of impulsive choice hinges on environment structure. Social Psychological and Personal-ity Science, 3, 131-138.
Pezzulo, G., Rigoli, F., \& Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. Progress in Neurobiology, 134, 17-35.
Rachlin, H., Raineri, A., \& Cross, D. (1991). Subjective probability and delay. Journal of the experimental analysis of behavior, 55(2), 233-244.
Reynolds, J. R., \& O'Reilly, R. C. (2009). Developing PFC representations using reinforcement learning. Cognition, 113(3), 281-292.
Savage, Leonard J. (1954). The Foundations of Statistics. New York, Wiley.
Sims, C. R., Neth, H., Jacobs, R. A., \& Gray, W. D. (2013). Melioration as rational choice: Sequential decision making in uncertain environments. Psychological Review, 120(1), 139-154.
Soto, P. L., McDowell, J. J., \& Dallery, J. (2006). Feedback functions, optimization, and the relation of response rate to reinforcement rate. Journal of the Experimental Analysis of Behavior, 85, 57-71.
Tricomi, E., Balleine, B. W., \& O’Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. European Journal of Neuroscience, 29(11), 2225-2232.
Van den Broek, M. D., Bradshaw, C. M., \& Szabadi, E. (1987). Behaviour of 'impulsive'and 'nonimpulsive'humans in a temporal differentiation schedule of reinforcement. Personality and Individual Differences, 8(2), 233-239.
Watkins, C. J. C. H. (1989). Learning from delayed rewards (Doctoral dissertation, University of Cambridge).
Wilson, M. P. \& Keller, F. S. (1953). On the selective reinforcement of spaced responding. Journal of Comparative and Physiological Psychology, 46, 190193.

# Shaping the Dynamics of Category Learning in Infants and Adults by Varying Learning Context 

Jelena Sucevic (jelena.sucevic@psy.ox.ac.uk)<br>Department of Experimental Psychology, University of Oxford, 9 South Parks Road Oxford, United Kingdom<br>Nadja Althaus (nadja.althaus@psy.ox.ac.uk)<br>School of Psychology, University of East Anglia, Norwich Research Park<br>Norwich, United Kingdom<br>Department of Experimental Psychology, University of Oxford, 9 South Parks Road Oxford, United Kingdom<br>Kim Plunkett (kim.plunkett@psy.ox.ac.uk)<br>Department of Experimental Psychology, University of Oxford, 9 South Parks Road Oxford, United Kingdom


#### Abstract

During the first year of life, infants develop a remarkable ability to group objects based on their similarities and differences. This ability of category formation represents one of the main mechanisms underlying the organisation of the semantic system. Early categories are formed spontaneously, in a non-supervised fashion and this type of category acquisition remains present even when more sophisticated forms of supervised category learning emerge. Even though there are various models of categorisation mechanisms across the lifespan, there is a gap in the research investigating implicit categorisation at different stages of cognitive development. Therefore, the aim of the current study was to compare processes of spontaneous concept formation in infants and adults using an experimental paradigm based on novelty preference. We discovered that both infants and adults show evidence of category learning (Experiment 1), though with different amount of training being needed to achieve the task. Adults successfully categorised objects already after a single block of training. Infants reached a level comparable to that of adults after twice the amount of training. As these tasks inevitably pose different cognitive and sensory demands to the two groups, in Experiments 2 and 3 we explored how varying parameters of the learning context affect dynamics of category formation. Decreasing memory demands of the task resulted in an acceleration of infants' category formation (Experiment 2), whereas posing memory load in an implicit category learning task decelerated adults' dynamics of category formation (Experiment $3)$.


Keywords: categorisation, learning context, non-supervised category acquisition, novelty preference, cognitive load, memory demands, infants, adults, eye tracking

## Introduction

The ability to group objects based on their similarities and differences represents one of the main mechanisms underlying the organisation of the semantic system. Concepts "embody much of our knowledge of the world telling us what things there are and what properties they have" (Murphy, 2002, p. 1). Therefore, categorisation
ability is considered to be critical for the organisation and stability of cognition (Mareshal \& Quinn, 2001). The ability to detect regularities in the environment and form categories emerges early in development. At the age of 3-4 months, infants already demonstrate the ability to differentiate categories of dogs and cats (Eimas \& Quinn, 1994), but also to form abstract perceptual categories (Bomba \& Siqueland, 1983). This ability becomes even more refined around ten months of age when infants become able to shape categories based on statistical regularities of category members (Younger \& Cohen, 1986).

In the domain of infant research, novelty preference is a standard method employed to explore categorisation processes. Typically, a familiarisation phase consisting of a set of training items (for instance members of two categories), is followed with a test phase where infants are presented with two novel items, one belonging to the familiarised category and one coming from a different category. Under the assumption of novelty preference, i.e. that infants look longer at the object that is perceived as less familiar, differences in looking times are interpreted as an index of category formation (e.g. Eimas \& Quinn, 1994). For instance, in a study exploring how infants form categories based on correlational feature structure, two test items are presented, one of them depicting an average of all presented items and the second object representing a subcategory average of one of two categories that could be formed (e.g. Plunkett et al., 2008). Infants who formed two categories demonstrated preference for the novel out-ofcategory overall average.

Despite the fact that processes of implicit, non-supervised category formation have been extensively studied in adults as well as in different patient populations (Reed, Squire, Patalano, Smith, \& Jonides, 1999), there have been few attempts to directly compare category learning processes in infants and adults. One recent computational model of categorisation offered an integrative account for infant and adult category learning (the SUSTAIN model, Gureckis \&

Love, 2004). Initially developed as a model of adult categorisation, the model proposes that mechanisms underlying infant and adult categorisation are not substantially different and assumes a continuous trajectory of conceptual development. Two explanations are offered to account for developmental differences - memory limitations and stimulus encoding limitations (Gureckis \& Love, 2004). In order to empirically test these assumptions, in the present study we developed an experimental paradigm for adults providing similar learning conditions as employed in the infant study. We aimed to address the question whether there are shared mechanisms of spontaneous conceptual organisation across the lifespan.

To parallel the visual familiarisation procedure used with infants, we designed a task to explore implicit category learning in adults where preferential looking was used as an index of category learning, and which provided similar learning conditions as those encountered in the infant study. As infants were merely presented with a set of objects, adults did not receive any explicit training or feedback on category formation. The adult task was also designed to tap into implicit, unsupervised category learning. As in the infant experiment, we presented adults with a series of objects as part of the familiarisation phase, followed by a test phase in which two test objects were presented and looking preferences were measured. Several studies using visual paired-comparison procedure with adults have shown that novelty preference can be used as an index of visual recognition in adults (Richmond, Colombo \& Hayne, 2007). As the magnitude of novelty preference increases with familiarisation time in object recognition (Richmond et al., 2007), we propose that the same effect can be interpreted as an index of category formation.

Even though the same type of experimental task was used with both infants and adults, there are inevitable differences in the demands to the two groups of participants this task poses. Task difficulty and memory demands might lead to differences in performance. One recent study has demonstrated the importance of learning conditions in altering categorisation in adult participants. Carvalho and Goldstone (2014) showed that category structure influences how efficiently category representations will be formed, and that this effect is tied to the way in which category members are presented. The authors conclude that there is categoryspecific attention allocation - simultaneous presentation promotes attention to commonalities among objects, while sequential presentation emphasizes differences among objects. These studies suggest that categorisation cannot be seen only as an extraction of abstract rules or computation of feature statistics, but emphasize that the dynamics of learning have an important role in category formation.

After discovering that adults and infants both showed categorisation, but at different rates (Experiment 1), we conducted Experiments 2 and 3 to investigate further which factors are relevant for categorsation in these cases. In order to explore how the context of learning affects category learning, we varied the task difficulty and investigated its
effects on categorisation in infants and adults. Our hypothesis was that decreasing the task demands will accelerate category formation in infants (Experiment 2), whereas adding an additional cognitive load to the task will delay category formation even in adults (Experiment 3).

## Experiment 1

The aim of the first experiment was to compare implicit category learning processes in infants and adults. In addition, we were interested in exploring the effects of the amount of training on forming categories based on statistical regularities, i.e. features correlations. Thus, three training blocks were interleaved with three blocks in which category formation was tested.

## Participants

Thirty-two 10-month-old infants took part in this study (two participants were excluded due to fussiness and refusal to look at the screen). Participants were recruited at the local maternity ward and all were full-term babies with no known health conditions. All participants came from homes where English was the only language spoken.

In addition, 24 adults took part in the experiment (mean age $=23.67$ years $(\mathrm{SD}=3.08)$ ). Two participants were excluded from the analyses (one due to calibration failure and one due to eye-tracker track loss).

## Stimuli

A set of novel objects was designed for the purposes of this study. Coloured and textured 3D looking objects represented novel creatures (called Sukis). As illustrated in Figure 1, each Suki consisted of four features: body, antennae, hands and legs. Each feature varied systematically on a scale of seven dimensions (body shape, number of antennae, hand size, length of legs), (see table 1). A set of 24 Sukis was designed in a way to resemble the structure of objects used in several categorisation studies (Younger \& Cohen, 1986; Plunkett et al., 2008; Mather \& Plunkett, 2011). Values of one feature were predictive for values on other dimensions, thus inviting participants to form two categories (defined as the narrow condition in Plunkett et al., 2008). However, the range of potential dimensions each feature can take was extended, thus instead of a range of 5 dimensions used in above mentioned studies, we introduced 7 potential variations of each feature.


Figure 1. Examples of the Sukis: Subcategory average objects

The reason for increasing the variability of the stimuli set is to have the possibility to create test items made of completely novel dimensions that have not been presented in any instance during the familiarisation phase. Four additional Sukis were designed to be presented as test items: an overall average object (consisting of mean values on each dimension, i.e. 4444), two subcategory averages (2222 and 6666). In addition, completely novel, out-of-category objects which comprised of the same features as all objects, but organized in a completely different manner were presented in the final trials of the test phase. All objects were depicted against a 5\% grey background.

Table 1: Stimulus structure (first familiarisation set)

| Stimulus | Category | Antenna | Hand | Body | Legs |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 1 | 1 | 1 | 3 | 3 |
| 2 | 1 | 1 | 3 | 1 | 3 |
| 3 | 1 | 3 | 1 | 3 | 1 |
| 4 | 1 | 3 | 3 | 1 | 1 |
| 5 | 2 | 5 | 5 | 7 | 7 |
| 6 | 2 | 5 | 7 | 5 | 7 |
| 7 | 2 | 7 | 5 | 7 | 5 |
| 8 | 2 | 7 | 7 | 5 | 5 |

## Procedure

Infants After written consent was obtained from a carer, an infant was seated on a carer's lap approximately 50 cm from a $1920 \times 1080$ inch screen in a sound-proof experimental booth. The parent was asked to keep their eyes closed for the duration of the experiment. Data was recorded using a Tobii TX300 Eye Tracker with a 120 Hz sampling frequency and four point calibration. The study was run with a custom Matlab stimuli presentation software PresentMate based on the Psychophysics Toolbox. Infants' behaviour was monitored via a centrally-located camera above the screen. Trials were initiated by the experimenter when the infant was attending the screen. Each familiarisation block consisted of eight trials. Each trial started with a presentation of an animated star in the central location of the screen accompanied by a chiming sound for the duration of 2000 ms . Following this, one stimulus ( 500 x 500 pixels) was presented in the central location for 6000 ms . As a previous study has shown that the order in which stimuli are presented may affect category formation (Mather \& Plunkett, 2011), we calculated mean Euclidean distance (as an average of seven distances between consecutive objects) for all possible stimuli sequences (40320 sequences) and selected sequences that fall within the range between the 40 th and 60 th percentile ( 8112 sequences). Then, for each participant a particular sequence from this pool was randomly selected. Three test blocks were interleaved with learning blocks. In each test, after an attention getter was presented for 2000 ms , two test objects were presented simultaneously for 10000 ms . The first two
trials were category formation test trials in which the overall category average (object 4444) and a subcategory average object ( 2222 or 6666 ) were presented. The positions of the two objects were counterbalanced across the two trials. The third test trial was always a novelty preference test in which one of the learning items from the previous learning phase was presented along with the novel, previously unseen out-of-category object. The purpose of this trial was to check whether infants were engaging in the task and expressing the expected novelty preference. The choice of the subcategory average object ( 2222 or 6666 ) presented in a particular test block was balanced across test blocks. The third test block was identical to the first test block for half of the participants, whereas others saw identical items as in the second test block. Which subcategory average object was presented first was counterbalanced across participants.

Adults Participants were instructed they would take part in a free viewing task so their only task would be to look at the objects presented on the screen. After written consent was obtained, participants were seated in front of the eye-tracker and their eye-movements were recorded using 120 Hz tracking frequency. Upon completion of the experiment, none of the participants reported they were aware what the purpose of the experiment was. The experimental design was kept as similar as possible to the infant version. Participants were presented with 3 blocks of training, each consisting of 8 trials. Each trial started with a centrally presented fixation cross for 500 ms followed by a presentation of a training items for the duration of 2000 ms . Training blocks were interleaved with test blocks. Analogous to the infant version, each test consisted of two categorisation test trials and a novelty preference trial. After the fixation cross was presented for 500 ms , test trials were presented on screen for 3000 ms .

## Results: Infants

Category Formation Test For category formation test trials, preference scores were calculated for all trials by dividing looking at the overall average object by total looking time to the overall and modal object. A repeatedmeasures ANOVA with factors Block (1, 2 and 3 ) and Test (1 and 2) showed no significant effects (all $\mathrm{ps}>0.05$ ) ${ }^{1}$. Planned comparisons against chance were performed for each test. Infants expressed a preference for the overall average object in the second trial of the second test block $(\mathrm{t}(25)=1.99, \mathrm{p}<0.05)$, (Figure 2).

Novelty Preference Test To validate that infants' behaviour was driven by novelty preference, infants were presented with the novelty preference test after each category formation test block. Infants' looking to the novel object was divided by the total looking time and a one-way

[^511]ANOVA with Block as a within-subjects factor revealed no effect of Block. Planned comparisons showed that infants' preference for the novel object differed from chance only in the second block $(\mathrm{t}(26)=2.58, \mathrm{p}<0.05)$.


Figure 2: Looking preferences in test

## Results: Adults

Category Formation Test Looking preference scores were calculated in the same way as for infants. The proportion of looking toward the overall average object was divided by the total looking time to both overall and modal objects. A repeated measures ANOVA with the within-subject factors Block and Trial revealed a significant main effect of Block ( $\mathrm{F}(2,44)=3.07, \mathrm{p}=0.05)$. Planned comparisons against chance revealed that preference towards overall average was significant only in the first block $(\mathrm{t}(22)=2.07, \mathrm{p}=0.05)$, whereas in the remaining does not significantly differ from chance ( $\mathrm{p}>0.05$; Table 2).

Novelty Preference Test Participants exhibited preference for the out-of-category object in all three novelty preferences tests (no difference was found between 3 novelty preference trials).

Table 2: Mean looking preferences on test (Experiment 1, Adults, SDs provided in brackets)

| Block | Test | Novelty |
| :---: | :---: | :---: |
| 1 | $0.55(0.17)^{*}$ | $0.67(0.22)^{*}$ |
| 2 | $0.51(0.13)$ | $0.67(0.18)^{*}$ |
| 3 | $0.47(0.17)$ | $0.61(0.20)^{*}$ |

## Experiment 1: Discussion

Results of Experiment 1 have shown that both infants and adults show evidence of category formation based on feature correlations in a free viewing task. In addition, this experiment revealed that paradigms based on novelty preference might be a useful tool in studying non-supervised category learning in adults. The results also showed that infants were slower in forming categories requiring a greater amount of familiarization to demonstrate similar level of
performance as adults. The observed differences in the performance might be due to various factors related to developmental differences. We hypothesize that one of the main factors driving these differences are memory demands. In order to compare an item presented during familiarisation with a previously presented one, information about the former items needs to be kept active in the working memory for comparison with the currently presented item. This might result in slower category formation in infants as it takes more resources to perform in this task due to limited memory abilities.

In order to test this hypothesis, we conducted the second experiment in which we decreased memory demands of the task by presenting familiarisation items in pairs. We hypothesized that, if memory load is reduced, infants will be faster in extracting category information.

## Experiment 2

## Participants

Twenty-eight 10 -months old infants took part in this study (two infants were excluded from the later analyses due to failing to reach minimum amount of looking time during familiarisation).

## Stimuli and procedure

The stimuli set used in this experiment was identical to the one in Experiment 1. As opposed to sequential presentation in the first experiment, familiarisation items were presented in pairs. A total of four trials were presented in each learning block. Following the presentation of an attention getter, two objects were presented simultaneously for 12 000 ms . Pairs of objects were selected based on sequences used in the sequential condition. Namely, each sequence used in Experiment 1 had a corresponding paired sequence in Experiment 2. It is important to note that despite the difference in the number of trials in each learning block, the total duration of learning blocks was identical across the two experiments.

## Results

Category Formation Test For category formation test trials, preference scores were calculated for all trials by dividing looking at the overall average object by total looking time to the overall and sub-category average object. A repeated measures ANOVA with factors Block (1, 2 and 3 ) and Test (1 and 2) revealed a main effect of Block ( $\mathrm{F}(2$, 36 ) $=3.15, \mathrm{p}<0.05$ ) (only participants who contributed to all trials were included in this analysis, $\mathrm{N}=19$ ). As there was no main effect of Trial, we averaged performance in the two test of the same block and performed planned comparisons. Performance in each block was compared against chance and we found that preference for the overall average object was significantly above the chance in the first block
$(\mathrm{t}(25)=2.91, \mathrm{p}<0.01)$, whereas in the remaining two blocks preference did not significantly differ from chance (mean preference values and dispersion are depicted in Table 3).

Novelty Preference Test To validate that infants' behaviour in test trials was driven by novelty preference, following each category formation test block, infants were presented with the novelty preference test. A one-way ANOVA with Block as a within-subjects factor revealed a near significant effect of Block (Greenhouse-Geiser $\mathrm{F}(2,38)=0.76$, $\mathrm{p}=0.06$ ). Planned comparisons revealed that infants' preference for the novel object differed from chance only in the first block (Wilcox signed rank test: $\mathrm{V}=242, \mathrm{p}<0.05$ ).

Table 3: Mean looking preferences on test (Experiment 2,

| SDs provided in brackets). |  |  |
| :--- | :--- | :--- |
| Block | Test | Novelty |
| 1 | $0.57(0.12)^{*}$ | $0.63(0.26)^{*}$ |
| 2 | $0.46(0.18)$ | $0.48(0.29)$ |
| 3 | $0.49(0.14)$ | $0.60(0.27)$ |

## Experiment 2: Discussion

Experiment 2 revealed that the dynamics of infants' category formation can be shaped by varying the parameters of the learning context. Decreasing memory demands in the task leads to a boost in extracting category relevant information. Infants' faster learning was also resembled in the fact that they demonstrated novelty preference already in the first test. Diminished novelty preference in the following tests and well as a larger attrition rate suggests that infants learned faster and then disengaged from the task. We conducted Experiment 3 to investigate whether increasing memory demands would lead to a decrease in the speed of category formation in adult participants. For the purpose of investigating implicit category learning in adults we adapted the N-back task, which is typically used in studies of working memory. Infants in the sequential condition were presented with one object at a time and had to mentally compare objects that were presented. Thus, we used a 1back version of the task and investigated whether participants spontaneously form categories under higher cognitive load conditions. If this incidental categorisation occurs, we expect that participants will judge betweencategories pairs faster than they would judge withincategory comparisons. Having perceptual similarity and semantic distance between to-be-compared items controlled, we predict that differences in the discrimination speed might reflect processes of categorisation.

## Experiment 3

## Participants

Twenty-four participants, students at Oxford University took part in this study (mean age $=23$ years $(\mathrm{SD}=2.54)$ ).

All participants were right handed and had normal or corrected-to-normal vision. Prior to taking part in the experiment, all participants signed an informed consent and upon experiment completion received course credits for their participation.

## Stimuli and Procedure

An identical stimuli set to the one in Experiments 1 and 2 was used in this study. The experiment consisted of seven blocks. Each block had two parts. A total of 36 learning trials were presented in the first part, and 10 test trials in the second part of each block. The order of presentation was pseudorandomised within each block. In order to balance the number of "identical" and "different" responses, participants were instructed to give a response only for probed trials, where a red dot would appear in the centre of the screen. In the learning part, there were 16 "different" comparisons (and an equal number of "identical"), half of which were within category, whereas the other half crossed the category boundary. In the test part, there were four "identical" and four "different" comparisons. While visual similarity and semantic distance (expressed through Euclidean distance) between the two compared objects was identical, some pairs crossed the category boundary (33335555), one pair was on the boundary (4444-6666) and some pairs were within the same category (1111-3333, 55557777).

## Results

In order to explore whether there is a difference in reaction times for between and within category judgments during the course of the experiment, growth curve analysis (Mirman, 2014) was used. Mean reaction times in each test block are presented in the Figure 3. The reaction time in mismatch trials were modeled using a linear growth curve model with a fixed effect of mismatch type (within-category and between-categories comparison) on the intercept and slope terms and random effects of participants on the intercept and slope to model individual differences in initial speed and rate of change. The fixed effect was added to the base model and its' effect on model fit was evaluated using model comparisons. All analyses were carried out in R version 3.3.1 using the lme4 package (version 1.1-12). There was a significant effect of mismatch type on the intercept $\left(\chi^{2}(1)=4.13, p<0.05\right)$ suggesting that participants responded faster when comparing items from different categories as opposed to performing within-category comparisons. This result suggests that participants organised items into two separate categories which resulted in making members of the same category look more similar than items belonging to different categories, even though perceptual similarity for both types of comparisons was identical.


Figure 3. Mean reaction time across 7 test blocks

## General Discussion

Taken together, the results of the experiments reported in this paper demonstrate how changing parameters of the learning context affect the dynamics of category learning in infants and adults. Experiment 1 provided evidence that both infants and adults can form categories in a free-viewing task, though it takes a different amount of exposure to succeed. Infants showed evidence of category formation after two blocks of familiarisation, whereas adults reached a similar level already after one training block. In addition, this experiment suggests that experimental paradigms based on novelty preference, the standard approach in infant research, can be used to explore non-supervised category learning in adult population as well.

Findings obtained in the second and the third experiment suggest that changing the task difficulty can accelerate (Experiment 2) or decelerate (Experiment 3) the process of extracting category relevant information for infants and adults, respectively. That the task structure can modulate infants' learning is further confirmed by the significant interaction between Experiment (1 and 2) and Block (1, 2 and 3 ) in a combined analysis $(F(2,66)=4.518, \mathrm{p}<0.05)$. The finding that paired presentation leads to faster category formation in infants is consistent with existing literature suggesting positive effects of comparison on learning and memory (Oakes \& Ribar, 2005). In addition, the results of Experiment 3 with adults also suggest that increasing the load impedes category formation. Initially developed to explain cross-modal effects of labels on categorisation, the perceptual load hypothesis can also offer a way of interpreting the obtained results (Plunkett, 2010). This hypothesis assumes that extraction of statistical information during category formation is also dependent on the perceptual load required to process individual stimuli and not exclusively on the feature correlations alone. Paired presentation may represent an optimal amount of available information for category formation in infants. Alternatively, it might be the case that the invitation to compare stimuli is contributing to the modulations of category learning. Future research needs to explore does manipulating load in other ways would result in a similar modulations of learning timecourse.

## References

Bomba, P. C., \& Siqueland, E. R. (1983). The nature and structure of infant form categories. Journal of Experimental Child Psychology, 35, 294-328.
Carvalho, P. \& Goldstone, R. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. Memory and Cognition, 42, 481495.

Eimas, P. D., \& Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. Child Development, 65, 903-917.
Gureckis, T. \& Love, B. C. (2004). Common Mechanisms in infant and adult category learning. Infancy, 5, 173-198.
Mareshal, D. \& Quinn, P. (2001). Categorization in Infancy. Trends in Cognitive Science. 5(10), 443-450.
Mather, E. and Plunkett, K. (2011). Same items, different order: Effects of temporal variability on infant categorization. Cognition, 119(3), 438-447.
Mirman, D. (2014). Growth curve analysis and visualization using $r$. Taylor \& Francis.
Murphy, G. L. (2002). The Big Book of Concepts. Cambridge, MA: MIT Press.
Oakes L. M., Ribar R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. Infancy 7, 8598.

Plunkett, K. (2010). The role of auditory stimuli in infant categorization. In L. M. Oakes, C. H. Cashon, M. Casasola, \& D. H. Rakison (Eds.), Infant perception and cognition: Recent advances, emerging theories, and future directions(pp. 203-221). New York, NY: Oxford University Press.
Plunkett, K., Hu, J.-F., \& Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. Cognition, 106, 665-681.
Reed, J., Squire, L., Patalano, A., Smith, E. \& Jonides, J. (1999). Learning about categories that are defined by object-like stimuli despite impaired declarative memory. Behavioral Neuroscience, 1, 113, 073507044.

Richmond, J., Colombo, M., \& Hayne, H. (2007). Interpreting performance in the visual pairedcomparison task. Journal of Experimental Psychology: Learning, Memory and Cognition, 33, 823-831.
$R$ Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Younger, B. A., \& Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes.Child Development, 57, 803-815.

# Early produced signs are iconic: Evidence from Turkish Sign Language 

Beyza Sümer (bsumer@ku.edu.tr)<br>Koç University, Turkey<br>\author{ Clara Grabitz (c.grabitz@student.ru.nl)<br><br>Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands }

Aylin Küntay (akuntay@ku.edu.tr)<br>Koç University, Turkey


#### Abstract

Motivated form-meaning mappings are pervasive in sign languages, and iconicity has recently been shown to facilitate sign learning from early on. This study investigated the role of iconicity for language acquisition in Turkish Sign Language (TID). Participants were 43 signing children (aged 10 to 45 months) of deaf parents. Sign production ability was recorded using the adapted version of MacArthur Bates Communicative Developmental Inventory (CDI) consisting of 500 items for TID. Iconicity and familiarity ratings for a subset of 104 signs were available. Our results revealed that the iconicity of a sign was positively correlated with the percentage of children producing a sign and that iconicity significantly predicted the percentage of children producing a sign, independent of familiarity or phonological complexity. Our results are consistent with previous findings on sign language acquisition and provide further support for the facilitating effect of iconic form-meaning mappings in sign learning.


Keywords: Iconicity, language acquisition, sign language

## Introduction

Arbitrariness, lack of a motivated link between a linguistic form and its meaning, has long been considered as a design feature of human language (de Saussure, 1915, 1983; Hockett, 1960). However, recent evidence has shown that iconicity, resemblance between form and its referent, is a more pervasive feature of language than previously thought (e.g., Perniss, Thompson, \& Vigliocco, 2010; Dingemanse, Blasi, Lupyan, Christiansen, \& Monaghan, 2015). For example, ideophones are used to express a wide range of features such as manner of movement, color, shape, size of an object, or emotional and psychological states (e.g., Japanese words 'korokoro' to refer to a light object rolling repeatedly and 'gorogoro' to a heavy object rolling repeatedly) or onomatopoeic words, which use the sound of a word to depict the sound of its referent (e.g., "moo" to refer to a cow) (e.g., Imai \& Kita, 2014). Compared to spoken languages, the use of visual-spatial modality makes iconicity a more prominent feature in sign languages (e.g., Taub, 2001). Signers can, for example, use a curved handshape to refer to a cup or use the signing space in front of them to show the location of different objects in relation with each other.
Pervasive existence of iconic forms in languages has intrigued many researchers about its role in language development. Accumulating evidence has shown that iconicity has a facilitating effect for early language
development. Imai, Kita, Nagumo, \& Okada (2008) reported an advantage for 3-year-old Japanese acquiring children in learning action words when these words are sound-symbolic compared to those which are arbitrarily linked to the sound of the action. In another study, 2.5-year-old children showed a tendency to match the words with rounded vowels to rounded shapes and words with unrounded vowels to pointed shapes (Mauer, Pathman, \& Mondloch, 2006). Onomatopoeic words constitute a substantial portion of early acquired vocabulary by German speaking children (Laing, 2014). Also, early acquired words in English and Spanish were rated more iconic than the ones acquired later (Perry, Perlman, \& Lupyan, 2015; Massaro \& Perlman, 2017).
The potential effects of iconicity have also been explored in the context of sign language acquisition research although in fewer studies compared to spoken language research. To understand the role of iconic nature of linguistic forms in sign languages, Orlansky \& Bonvillian (1984) analyzed whether the first signs learned by signing children were more iconic, but their initial analyses did not find an overrepresentation of such signs in the first 10 words and beyond. In their data, only about a third of words were iconic, which was in line with the overall proportion of iconic signs in American sign Language (ASL). In contrast, Lloyd, Loeding, \& Doherty (1985) reanalyzed the data based on a broader definition of iconicity and found an over-representation of iconic signs in early acquired signs. Since then it has become clear that iconicity is a more complex property rather than being a holistic concept, and it is now common to rate the iconicity of signs on a scalar scale of 1 (not at all iconic) to 7 (highly iconic) (e.g., Vinson, Cormier, Denmark, Schembri, \& Vigliocco, 2008 for BSL norms). However, iconicity of signs can also differ qualitatively in meaningful ways: for example, signs can represent actions or perceptual qualities of their referents (see Thompson et al., 2011; Ortega, Sümer, \& Özyürek, 2016). Under this new view of iconicity, recent studies with signing children have shown that iconicity has a predictive power in early sign learning, i.e., iconic signs are acquired earlier than non-iconic signs (Thompson, Vinson, Woll, \& Vigliocco, 2012; Caselli \& Pyers, in press). Thompson and colleagues (2012) explored the relationship between iconicity and sign language acquisition using a mixed cross-sectional, longitudinal design on productive and comprehensive vocabulary size of British Sign Language (BSL) acquiring deaf children
of deaf parents administering the MacArthur Bates Communicative Developmental Inventory (CDI). The CDI is a parental report in which parents indicate vocabulary learning by ticking items on a list of words if their child produces or understands them (Fenson, Dale, Reznick, Bates, Thal, \& Pethick, 1994). They collected data from 31 deaf children aged between 8 and 30 months and found that iconicity facilitates sign learning from early development on, but particularly in children older than 21 months. They argue that the advantage seen in older children could be due to more cognitive capacity or more environmental experience which children can use to understand and establish such iconic links between meaning and form.
It has also been proposed that other properties of a sign may be as or even more important than iconicity such as lexical frequency (i.e., familiarity), neighborhood density (i.e., the number of lexical items that are phonologically related to a target), or phonological complexity (Caselli \& Pyers, in press; Thompson et al., 2012). Phonological complexity in sign language is based on motoric subunits and seems to be a crucial factor because children learning a sign language show phonological reductions and substitutions of features that are marked less - similar to children learning a spoken language (e.g., Meier et al., 2008). To what extent these factors are influential on early sign learning is also controversial since Thompson et al. (2012)'s results indicate iconicity to be a more powerful predictor than other factors such as familiarity or phonological complexity, while Caselli and Pyers (in press) argue that neighborhood density and lexical frequency (familiarity) are also as strong contributors as iconicity in early sign development.
Studies with signing and speaking children provide converging evidence on the facilitating role of iconic form-meaning mappings in early lexical development regardless of the language modality. Investigating the effects of iconicity in sign language acquisition provides further evidence for this "modality-free" role of iconic forms in this domain of language development and contributes to our understanding on to what extent general principles of conceptual development influence the language development in signing children - a less studied population compared to speaking children.

## The Present Study

The aim of the current study was to add to the growing literature on the role of iconicity in sign language acquisition by studying children acquiring TID, which has been studied less than many Western sign languages such as ASL or BSL. Analyzing data from other sign languages is crucial, because conflicting views on the role of iconicity in sign language acquisition still exist (e.g. Thompson et al., 2011). Based on Thompson et al. (2012) and Caselli \& Pyers' (in press) findings, we assumed that the visual-spatial modality of sign language, which is rich in iconic form-meaning mappings, would modulate language acquisition and that iconicity could present a potential advantage in early
language development for TID acquiring children. We hypothesized that the iconicity of a sign would be correlated with the percentage of children producing the sign. We further expected iconicity to be a significant predictor of the percentage of children producing a sign, even if after controlling for familiarity and sign complexity.

## Method

## Design

The experiment was realized with a correlational design. Simultaneous multiple regression was used to test whether different sign characteristics (iconicity, familiarity and phonological complexity) were good predictors for the percentage of children producing a sign.

## Participants

Data were collected for 27 deaf children of deaf parents and 16 hearing children of deaf parents (i.e., CODAs), thus for a total of 43 children (female 28) born to deaf families and exposed to TID from birth. Where possible, parents participated in data collection at two separate time points, with a 3-month interval, thus increasing our data set to a total of 57 . Although CDI typically tracks vocabulary development in children between 8-36 months of age, the age of the children in the current study ranges between $10-45$ months ( $M=25.98, S D=10.23$ ) to increase sample size. The majority of children were from families with a middle or upper socio economic status, meaning that at least one parent works in a paid job and completed high-school education.

## Procedure

Data were elicited from a version of the MacArthur Bates Communicative Developmental Inventory (CDI) adapted for TID. In this version, three source tests were taken into consideration to account for modality and cultural specific issues: ASL CDI (Anderson \& Reilly, 2002), BSL CDI (Woolfe, Herman, Roy, \& Woll, 2009) and Turkish CDI (TIGE, Aksu-Koç, Küntay, Acarlar, Maviş, Sofu, Topbaş, \& Turan, 2009). As a result, TID CDI consists of 500 items grouped into 18 categories such as "animals", "toys" and "actions".
Although previous CDIs, both for sign and spoken languages, were administered in a pen-and-paper format, considering low reading abilities of deaf people (e.g., Bloomquist Traxler, 2000), we presented TID CDI in a web-based format where parents themselves logged onto the system to see videos of signs one after each other and decided whether their children produced the sign in the video. This procedure was preceded by a training session in which they saw instruction videos in TID and asked any clarification questions to a deaf assistant, who was also online during entire training session. Only sign production data was collected since a small pilot study with a group of deaf parents showed that it was often confusing to differentiate comprehension versus production of a sign for them. They also expressed that they were less sure about their answers regarding
comprehension since they cannot see comprehension but only production of the signs.
As part of another study (Taşçı \& Sumer, in prep), 4 deaf signers of TID $\left(\mathrm{M}_{\text {age }}=32.3\right)$ were shown a total of 328 signs on a computer screen and asked how iconic they think these signs are. Here, the iconicity was defined as the similarity between the linguistic form (i.e., sign) and the entity that it refers to, including both the perceptual and/or action-based properties. In another session, 5 deaf signers ( $\mathrm{M}_{\mathrm{age}}=33.4$ ), were asked how familiar they think these signs are. Both iconicity and familiarity ratings were on a scale of 1 (not at all iconic/familiar) to 7 (highly iconic/familiar) (e.g. see Vinson et al., 2008 for norms in BSL for comparison). We additionally included phonological complexity ratings following Mann, Marshall, Mason, \& Morgan (2010), in which three main phonological parameters of signs (i.e., handshape, location, movement) were assigned a complexity value. For example, unmarked handshapes in TID, as determined by Kubus (2008), were rated less complex than other handshapes.

## Results

We excluded data points with unrealistic productive sign scores aged 10 to 20 months, if they were outside the Mean plus Standard Deviation found for ASL norms (Anderson \& Reilly, 2002) for the child's age range. These sign scores can be attributed to a misunderstanding during data collection. Exceptionally high sign scores were not excluded for children aged older than 20 months, as high variability is a key component of language acquisition. Thus, we included a total of 51 data points in our analyses. The productive sign score for the subsample of 104 signs was $(M=51.27, S D=32$ ) (Table $1)$.

Table 1. General descriptive statistics after excluding outliers ( $N=51$ )

| Age | Total Productive <br> sign score | Subset Productive <br> sign score |
| :---: | :---: | :---: |
| $10-45$ | $1-500$ | $0-104$ |
| $M=27.45$ | $M=215.49$ | $M=51.27$ |
| $S D=9.74$ | $S D=154.56$ | $S D=32$ |

$N=$ Sample size, $M=$ Mean, $S D=$ Standard deviation
Each sign was on average produced by $44 \%$ of children ( $M=0.44, S D=0.18$ ). Spearman's correlation between age and subset productive sign score was significant ( $r_{s}$ $=.54, p<.001$ ).


Figure 1. Spearman's correlation between subset productive sign score and age in months. Linear trend lines included.

Iconicity, familiarity and complexity scores of 104 signs were available and used for further analyses (Table 2).

Table 2. Sign ratings and descriptive statistics for the subset of 104 signs

| Iconicit <br> $\mathbf{y}$ | Familiarit <br> $\mathbf{y}$ | Complexit <br> $\mathbf{y}$ | Mean Age <br> $\mathbf{o f}$ <br> Productio <br> $\mathbf{n}$ | PerPro <br> $\mathbf{d}$ |
| :---: | :---: | :---: | :---: | :---: |
| $1-7$ | $2-7$ | $0-2$ | $4-27$ | $.11-.84$ |
| $M=5.15$ | $M=6.16$ | $M=.64$ | $M=15.13$ | $M=.44$ |
| $S D=$ <br> 1.89 | $S D=.91$ | $S D=.67$ | $S D=5.84$ | $S D=.18$ |

$M=$ Mean, $S D=$ Standard deviation, Mean Age of Production = mean age of children capable of producing a sign, PerProd $=$ percent of children capable of producing a sign

Spearman's correlations were carried out to clarify the relationship between the main variables in the study (Table 3). Mean Age of Production and Percentage Producing were highly positively correlated and therefore only percentage of production was used for further regression analyses. Iconicity $r_{s}=.38, p<.001$ and familiarity ratings $r_{s}=.32, p=.001$ were both significantly positively correlated with Mean Age of Production and Percentage Producing. However, iconicity and familiarity ratings were not correlated. Phonological complexity was not correlated with Mean Age of Production and Percentage Producing and was also unrelated to iconicity and familiarity ratings.

Table 3. Spearman's correlations between the main variables in the study

|  | PerProd | Icon | Fam | Complexity |
| :--- | :--- | :--- | :--- | :--- |
| Mean Age of <br> Production | $1^{* *}$ | $.39^{*}$ | $.32^{*}$ | -.13 |
|  |  | $*$ | $*$ |  |
| Percentage <br> Producing |  | $.38^{*}$ | $.32^{*}$ | -.13 |
| Iconicity |  |  | .13 | -.18 |
| Familiarity |  |  |  |  |
| Complexity |  |  |  |  |
| $\leq .005$, PerProd = Percentage of children capable of |  |  |  |  |
|  |  |  |  |  |



Figure 2. Spearman's correlation between iconicity ratings and percentage of children producing a sign. Linear trend line included. $r_{s}=.38, p<.001$.


Figure 3. Spearman's correlation between familiarity ratings and percentage of children producing a sign. Linear trend line included. $r_{s}=.32, p=.001$.

The data were entered into simultaneous multiple regression analysis, using the percentage of children producing a sign as dependent variable and iconicity, familiarity and complexity as predictors. The results for the model indicate that the predictors explained $17 \% R^{2}=$ $.17, F(3,100)=7.84, p<.001$. Both iconicity $\beta=.31$, $t(100)=3.38, p=.001$ and familiarity $\beta=.25, t(100)=$ $2.78, p=.006$ significantly predicted the percentage of children producing a sign. Phonological complexity was not a significant predictor. Adjusted R Squared values were used in the analysis.

## Discussion and Conclusion

We investigated the role of iconicity in the acquisition of Turkish Sign Language (TID) by signing children of deaf parents and found that the iconicity of a sign was positively correlated with the percentage of children producing a sign. In addition to iconicity, familiarity, but not phonological complexity, seems to be influential in early sign learning. We thus provide further evidence regarding the facilitating role of iconicity in early sign learning by signing children.

Our results converge with what previous studies with signing children have found so far (Thompson et al., 2012; Caselli \& Pyers, in press). There is robust evidence showing that early acquired signs are iconic, which suggests that resemblance between form and meaning in sign languages bootstraps word learning in sign languages. Moreover, analyzing parental input to BSL signing children aged between $25-51$ months, Perniss, Lu , Morgan, \& Vigliocco (2017) suggest that deaf parents exploit iconicity while communicating with their children. These studies also show that iconicity seems to be more advantageous for sign language acquiring children when they are at around 30 -months of age although this age group is called "older" in Thompson et al. (2012) and "younger" in Caselli \& Pyers (in press). This seems to be related to increasing cognitive skills or more experience with environment that enables establishing the link between linguistic form and meaning (Thompson et al., 2012).
Further evidence from spoken languages, which are less rich in iconic forms than sign languages (Taub, 2001), has been presented about the facilitating role of iconicity, as well (e.g., Imai \& Kita, 2014; Imai et al., 2008; Laing, 2014; Perry et al., 2015; Massaro \& Perlman, 2017). The effect of iconicity in early word acquisition in spoken languages seems to be more prominent earlier compared to what studies with signing children report. Studying expressive and receptive vocabulary development in the first four years of English acquiring children (6-47 months of age), Massaro and Perlman (2017) show that iconicity is more prevalent early in acquisition and decreases with increasing age and vocabulary size. There might be a difference in the role of iconicity throughout development due to different modalities of sign and spoken languages: sign languages are rich in iconic forms and signing children are more likely to encounter iconic forms - not only at the lexical level but also at the level of morphology (e.g., classifiers) and syntax (e.g., expressing spatial relations in signing space) than speaking children whose lexicon gets enriched with less iconic words (more arbitrary forms) as they get older. Therefore, iconicity seems to help children in their early word learning, but its role might change as children acquiring languages in different modalities advance in their language development.
Our results regarding other factors than iconicity such as familiarity and phonological complexity are partially in line with Thompson et al. (2012), who found iconicity to be a stronger predictor of early sign acquisition than others. The current study, on the other hand, reveals the role of familiarity as important as iconicity, which is in line with Caselli \& Pyers (in press). The findings regarding the role of phonological complexity do not suggest that it predicts early sign learning - as opposed to Caselli \& Pyers (in press). The difference might come from different definition of complexity since Caselli \& Pyers (in press) focused on neighborhood density (the number of lexical items that are phonologically related to a target) rather than a complexity rating system as used in the current study. Thompson et al. (2012) also
observed an effect of phonological complexity, but this effect was restricted to younger children (11-20 months of age) only while Caselli \& Pyers (in press) found the neighborhood density effect across all age groups (8-35 months of age). This might be still a result of different approaches taken to the analysis of phonological complexity in different studies.
However, one needs to be careful when interpreting our findings as the sample size was small and we had to exclude multiple outliers for the main analyses. Furthermore, only production scores were collected. Since comprehension scores are less prone to phonological and motor constraints, drawing conclusions from production scores only may underestimate the role of iconicity, especially for younger children who produce substitution errors while producing signs (e.g., Lu, Jones, \& Morgan, 2016).
It is also important to note that iconicity and familiarity ratings were mostly available for nouns which could have further skewed our sample as some of the first words were "come" or "kiss". Perry et al (2015)'s results from English and Spanish suggest that adjectives are rated as more iconic than nouns and function words, and verbs as more iconic than nouns and function words in English. Perniss et al. (2017), however, found that signs for objects and actions are rated more iconic than those for properties (e.g., blue, fast). This might be the result of modality difference between sign and spoken languages and underlines the importance of including different lexical categories in such an analysis.
Additionally, the current study is clearly limited by the correlational approach taken. With sufficient resources a Bayesian modelling approach similar to Thompson et al. (2012) or a mixed-effect logistic regression modelling approach will be more powerful as one can simultaneously account for child-specific and itemspecific variability while controlling for factors such as familiarity or phonological complexity. The current data set could be used as a basis for further analyses, but it will also be beneficial to collect iconicity and familiarity ratings for more signs, in particular for signs that are action related and represent a wider range of familiarity ratings.
Finally, the present study is clearly limited by the use of parental reports - spontaneous production sessions that target sign and speech output and/or recordings of the children that are scored will be useful to further qualify results, particularly in regard to underlying mechanisms and driving forces. Innovative approaches towards testing are needed, such as Perniss et al. (2017) who analyzed child directed signing using only the parents in an experimental setting and showed that child-directed signing exploits iconicity, especially when referents are not present. Such studies will further qualify the input that signing children receive and might be decisive in determining the real importance of iconicity for language acquisition.
Our study represents a further step on the way exploring iconicity in relation to sign language acquisition. While we do not agree with the notion that only cognitive
development drives language acquisition based on our results, iconicity cannot explain all aspects of early sign language acquisition. Acquisition of these signs is likely to be driven by contextual factors such as use of frequency (with both adults and children) or neighborhood density (Caselli \& Pyers, in press).
In summary, language acquisition is likely to be facilitated by iconicity. Considering the potential benefit of meaningfully motivated form-meaning for language acquisition in general, both iconicity and arbitrariness should be re-evaluated as general properties of a language (Perniss et al., 2010), although more studies are needed to further support this claim and its relevance for all languages.

## Acknowledgments

This work has been supported by Tubitak 2236 (Cofunded Brain Circulation Scheme) awarded to the first author. We thank deaf families who kindly agreed to participate in the study and our deaf assistants Sevinç Akın for her help with data collecting.

## References

Aksu-Koç, A., Küntay, A., Acarlar, F., Maviş, İ., Sofu, H., Topbaş, S., \& Turan, F. (2009). Türkçe'de Erken Sözcük ve Dilbilgisi Gelişimini Ölçme ve Değerlendirme Çalışması Türkçe İletişim Gelişimi Envanterleri: TİGE-I ve TİGE-II. TÜBİTAK'a sunulmuş rapor, Proje No: 107K058. (The assessment and evaluation of early lexical and grammatical deveopment in Turkish: The Turkish Communicative Development Inventories, TIGE-I and TIGE-II. Final report of Project No: 107 K 058 , submitted to Turkish Scientific and Technological Research Foundation.)
Anderson, D. \& Reilly, J. S. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. Journal of Deaf Studies and Deaf Education, 7, 83-106.
Bloomquist Traxler, C. (2000). The Standford Achievement Test, $9^{\text {th }}$ Edition: National norming and performance standards for deaf and hard-of-hearing students. Journal of Deaf Studies and Deaf Education, 5(4), 337-348.
Caselli, N. \& Pyers, J. (in press). The road to language learning is not entirely iconic: Neighborhood density, iconicity, and frequency facilitate sign language acquisition. Psychological Science.
de Saussure, F. (1983). Course in general linguistics (R. Harris, Trans,; C. Bally \& A. Sechehaye, Eds.). La Salle, IL: Open Court. (Original work published in 1915).

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen M. H., \& Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. Trends in Cognitive Sciences, 19, 603-615.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., \& Pethick, S. J. (1994). Variability in early communicative development. Monographs of the Society for Research in Child Development, 59, 1-173.

Hockett, C. F. (1960). The origin of speech. Scientific American, 203, 88-96.
Imai, M., Kita, S., Nagumo, M., \& Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109, 54-65.
Imai, M. \& Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. Philosophical Transactions of the Royal Society B, 369, 20130298.
Kubus, O. (2008). "An analysis of Turkish Sign language (TID) Phonology and morphology". Master' Thesis, Middle East Technical University.
Laing, C. E. (2014). A phonological analysis of onomatopoeia in early word production. First Language, 34, 387-405.
Lloyd, L., Loeding, B., \& Doherty, J. (1985). Role of iconicity in sign acquisition: A response to Orlansky and Bonvillian (1984). Journal of Speech and Hearing Disorders, 50, 299-301.
Lu, J., Jones, A., \& Morgan, G. (2016). The impact of input quality on early sign development in native and non-native language learners. Journal of Child Language, 1, 1-16.
Mann, W., Marshall, C. R., Mason, K., \& Morgan, G. (2010). The acquisition of sign language: The impact of phonetic complexity on phonology. Language Learning and Development, 6, 60-86.
Massaro, D. W. \& Perlman, M. (2017). Quantifying iconicity's contribution during language acquisition: Implications for vocabulary learning. Frontiers in Communication, 2:4. doi: 10.3389/fcomm.2017.00004.

Mauer, D., Pathman, T., \& Mondloch, C. J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. Developmental Science, 9(3), 31622.

Perniss, P., Thompson, R., \& Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. Frontiers in psychology, 1, 227.
Perniss, P., Lu, J. C., Morgan, G. \& Vigliocco, G. (2017). Mapping language to the world: the role of iconicity in sign language input. Developmental Science, e 12551.
Perry, L. K., Perlman, M., \& Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. PLoS ONE, 10, e0137147.
Orlansky, M. D., \& Bonvillian, J. D. (1984). The role of iconicity in early sign language acquisition. Journal of Speech and Hearing Disorders, 49(3), 287-292.
Ortega, G., Sumer, B., \& Ozyurek, A. (2016). Type of iconicity matters in the vocabulary development of signing children. Developmental Psychology. Advance online publication. doi:10.1037/dev0000161.
Taub, S. F. (2001). Language from the body: Iconicity and metaphor in American Sign Language. Cambridge University Press.
Tasci, S. \& Sumer, B. (in prep). Comparison of iconicity judgments of Deaf signers and hearing non-signers.

Thompson, R. L. (2011). Iconicity in language processing and acquisition: What signed languages reveal. Language and Linguistics Compass, 5(9), 603616.

Thompson, R. L., Vinson, D. P., Woll, B., \& Vigliocco, G. (2012). The road to language learning is iconic: Evidence from British Sign Language. Psychological science, 23(12), 1443-1448.
Vinson, D. P., Cormier, K., Denmark, T., Schembri, A., \& Vigliocco, G. (2008). The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity. Behavior Research Methods, 40, 1079-1087.
Woolfe, T., Herman, R., Roy, P., \& Woll, B. (2009). Early vocabulary development in deaf native signers: a British Sign Language adaptation of the communicative development inventories. Journal of Child Psychology and Psychiatry, 51(3), 322-31.

# A Minimal Neural Network Model of The Gambler's Fallacy 

Yanlong Sun (ysun@tamhsc.edu)

Hongbin Wang (hwang@tamhsc.edu)<br>Center for Biomedical Informatics, Texas A\&M University Health Science Center Houston, TX 77030 USA


#### Abstract

The gambler's fallacy has been a notorious showcase of human irrationality in probabilistic reasoning. Recent studies suggest the neural basis of this fallacy might have originated from the predictive learning by neuron populations over the latent temporal structures of random sequences, particularly due to the statistics of pattern times and the precedence odds between patterns. Here we present a biologically-motivated minimal neural network model with only eight neurons. Through unsupervised training, the model naturally develops a bias toward alternation patterns over repetition patterns, even when both patterns are equally likely presented to the model. Our analyses suggest that the way the neocortex integrates information over time makes the neuron populations not only sensitive to the frequency signals but also relational structures embedded over time. Moreover, we offer an explanation for how higher-level cognitive biases may have an early start at the level of sensory processing.


Keywords: gambler's fallacy; alternation bias; waiting time; temporal integration; predictive learning.

## Introduction

The gambler's fallacy-a belief that chance is a self-correcting process where a deviation in one direction would induce a deviation in the opposite direction-has been a notorious showcase of human irrationality in probabilistic reasoning (Tversky \& Kahneman, 1974). For decades, this fallacy is thought to have originated from a cognitive bias called the "representativeness heuristic", which is attributed to the belief of the "law of small numbers" that small samples are highly representative of the populations from which they are drawn (Gilovich, Vallone, \& Tversky, 1985; Tversky \& Kahneman, 1974).

Recent development in neuroscience and computational models suggests that the human mind develops structured probabilistic representations about the world and performs nearoptimal Bayesian inferences (Pouget, Beck, Ma, \& Latham, 2013; Tenenbaum, Kemp, Griffiths, \& Goodman, 2011). For example, representativeness has been defined with a Bayesian belief-updating structure in which different hypotheses are evaluated based on different sets of the input data (Gigerenzer \& Hoffrage, 1995; Griffiths \& Tenenbaum, 2001). However, it remains elusive how the structured hypothesis space has originated in the first place, and how cognitive biases can arise from normative probabilistic models.

On the topic of randomness perception, there has been a growing speculation that people's intuition about random process, also known as the subjective randomness, is biased by the statistical structures in the learning environment (Budescu, 1987; Falk \& Konold, 1997; Hahn \& Warren, 2009; Lopes \& Oden, 1987; Nickerson, 2002; Oppenheimer \& Monin, 2009; Oskarsson, Van Boven, McClelland, \& Hastie, 2009; Sun,

Tweney, \& Wang, 2010). Particularly, we have argued that without a predefined hypothesis structure, biases underpinning the gambler's fallacy can emerge by simply capturing the temporal relations between patterns as a random process unfolds over time (Sun \& Wang, 2010a, 2010b, 2012, 2015). With a biologically realistic simple recurrent model that learns to reencode sequential binary data through unsupervised learning, we show that dissociation of random patterns can naturally emerge as the consequence of inhibitory competition between overlapped representations (Sun et al., 2015). Our findings indicate that cognitive biases in overt behavior can emerge early and locally at the level of sensory processing, and neurons' sensitivity to the temporal structures in the learning environment is the key in bridging the gap between neurons and behavior.

In the following, we first introduce some basic normative measures on the time of random patterns. Then, based on the neural model we reported earlier, we present a minimal neural network model with only eight units. We will show that this minimal model can mostly replicate our previous findings and provide new insights regarding the neural encodings of sequential patterns.

## Temporal Distance between Patterns

In sequences generated by a random process, there can be fundamentally different types of statistical structures regarding how often a pattern occurs and when a pattern is to occur. Our previous works have been focusing on the distinction between the mean time statistic that measures how often the pattern occurs in a global sequence, and the waiting time statistic that measures when a pattern will first occur since the beginning of the observation. Here we introduce a more compact yet more comprehensive framework that incorporates not only both types of statistics for individual patterns but also the statistics depicting the relational structures between different patterns.

To compute the temporal distances between different patterns with different initial states, we use the first-order dependent Markov chains parameterized by the probability of alternation $\left(p_{A}\right)$ between consecutive trials and the corresponding generating functions (Figure 1). ${ }^{1}$ Define $E\left[T_{j \mid i}\right]$ as the expected number of transitions from the initial state $i$ until the

[^512]

Figure 1: Markov chains for generating the waiting times $E\left[T_{\mathrm{HH} \mid \varnothing]}\right]$ and $E\left[T_{\mathrm{HT} \mid \varnothing}\right]$ given the initial empty state $\varnothing$ (Figure $\mathbf{A}$ ), and the additional times $E\left[T_{H \mathrm{H} \mid \mathrm{H}}\right], E\left[T_{\mathrm{HT} \mid \mathrm{H}}\right]$, and $E\left[T_{\mathrm{T} \mid \mathrm{H}}\right]$, given the same initial state $H$ (Figure B). In each chain, states $S_{j \mid i}$ represent all possible sequences that start from the pattern $i(i=\varnothing$ means starting anew) and end with the first arrival of the pattern $j$. States $M_{k}$ represent all possible sequences that end with the pattern $k$ but do not contain the expected pattern $j$. Transitions between nonempty states are labeled as either repetition (R) or alternation (A). Figure $\mathbf{C}$ : Pairwise precedence odds between patterns when the probability of alternation $p_{A}=1 / 2$, for example, the odds are 3 to 1 that one is to first encounter TH than to first encounter HH .
first arrival of the pattern $j$. When the initial state is empty $i=\varnothing$ (i.e., the counting process starts anew), $E\left[T_{j \mid \varnothing}\right]$ is referred to as the waiting time of pattern $j$. For example, from Figure 1A, the waiting times for the patterns HH and HT are respectively,

$$
\begin{align*}
& E\left[T_{\mathrm{HH} \mid \varnothing]}=1+\frac{1}{2 p_{A}}+\frac{2}{1-p_{A}},\right. \\
& E\left[T_{\mathrm{HT} \mid \varnothing]}=1+\frac{1}{2 p_{A}}+\frac{1}{p_{A}} .\right. \tag{1}
\end{align*}
$$

When $p_{A}=1 / 2$ (namely, independent Bernoulli trials with a fair coin where repetitions and alternations are equally likely), we have $E\left[T_{\mathrm{HH} \mid \varnothing}\right]=6$ and $E\left[T_{\mathrm{HT} \mid \varnothing \varnothing}\right]=4$.

When the initial state is not empty, $E\left[T_{j \mid j}\right]$ is referred to as the additional time for pattern $j$ given the initial state $i$. For example, from Figure 1B, we have

$$
\begin{equation*}
E\left[T_{\mathrm{HH} \mid \mathrm{H}}\right]=\frac{2}{1-p_{A}}, \quad E\left[T_{\mathrm{HT} \mid \mathrm{H}]}\right]=E\left[T_{\mathrm{T} \mid \mathrm{H}}\right]=\frac{1}{p_{A}} . \tag{2}
\end{equation*}
$$

At $p_{A}=1 / 2$, we have $E\left[T_{\mathrm{HH} \mid \mathrm{H}}\right]=4$ and $E\left[T_{\mathrm{HT} \mid \mathrm{H}}\right]=E\left[T_{\mathrm{T} \mid \mathrm{H}}\right]=2$.

When the initial state $i$ is exactly the desired pattern $j$, $E\left[T_{j \mid}\right]$ denotes the expected number of transitions between any two consecutive occurrences of the pattern $j$, and is referred to as the mean time of pattern $j$. Since the first-order Markov chain is memoryless between consecutive transitions, we have relations such as $E\left[T_{\mathrm{HH} \mid \mathrm{HH}]}\right]=E\left[T_{\mathrm{HH} \mid \mathrm{H}}\right]$, and $E\left[T_{\mathrm{HT} \mid \mathrm{HT}}\right]=E\left[T_{\mathrm{HT} \mid \mathrm{T}}\right]$. Therefore, the mean times for the patterns HH and HT are respectively,

$$
\begin{equation*}
E\left[T_{\mathrm{HH} \mid \mathrm{H}}\right]=\frac{2}{1-p_{A}}, \quad E\left[T_{\mathrm{HT} \mid \mathrm{T}]}\right]=\frac{2}{p_{A}} . \tag{3}
\end{equation*}
$$

At $p_{A}=1 / 2$, we have $E\left[T_{\mathrm{HH} \mid \mathrm{H}]}\right]=E\left[T_{\mathrm{HT} \mid \mathrm{T}}\right]=4$. The inverse of mean time is frequency. For example, $E\left[T_{H H \mid H}\right]=4$ means that we expect to see the pattern HH once in every 4 tosses of an fair coin.

Among these different measures, the most striking distinction is that at $p_{A}=1 / 2$, we have $E\left[T_{\mathrm{HH}} \mathrm{H}\right]=4$ but $E\left[T_{\mathrm{HT} \mid \mathrm{H}}\right]=2$, in spite of the fact that given an H , the next digit is equally likely to be either an H or a T . This is because a reoccurrence of the pattern HH can "reuse" the ending elements from its previous occurrence, but a reoccurrence of the pattern HT must always start anew. This statistical property of faster transition times when starting anew is known as new better than used (NBU) (Ross, 2007). Essentially, the NBU property is due to the overlap between a pattern and a shifted copy of itself or between different patterns. For example, as shown in Figure 1B, pattern HH overlaps with its shifted copy by one element H , but pattern HT does not. Then, towards the destination state $S_{\mathrm{HH} \mid \mathrm{H}}$, anytime things go astray (i.e., the process ends in state $M_{\mathrm{T}}$ ), the waiting for HH has to start all over. In contrast, the waiting for HT is always on average two flip away from the state $M_{\mathrm{H}}$. As a result, the transition to $S_{\mathrm{HH} \mid \mathrm{H}}$ is "delayed" than the transition to $S_{\mathrm{HT} \mid \mathrm{H}}$. This overlap also explains the pairwise precedence relation shown in Figure 1C. For example, in the competition between the patterns HH and TH , the former reuses the last element of the latter but the latter starts anew. As a result, if we toss a fair coin repeatedly (i.e., $p_{A}=1 / 2$ ), the odds are 3 to 1 that we first encounter TH than first encounter HH .

## A Neural Model of Temporal Integration

The NBU property of pattern times has fundamental implications in neural encoding of pattern events. As shown in Figure 1, different measures of waiting time, additional time, mean time and pairwise precedence odds are all due to the overlap between temporal patterns. Recent developments in neuroscience and computational models suggest that neural encodings of events and values are always overlapped, and it is the encoding of neural populations that give rise to higher-level and more abstract representations (Adolphs, 2015; Dehaene \& Brannon, 2010; Pouget et al., 2013). In the domain of temporal integration, probabilistic encoding must consider the overlap between representations at different times, namely, recurrent processing (Elman, 1990). Then, we would immediately conjecture that via merely encoding the random sequences that unfold over time, populations of neurons would naturally capture the temporal structures depicted by the pattern times statistics.


Figure 2: A neural network model of temporal integration (figures adopted from Sun et al., 2015). Figure A: A two-unit input layer scans a sequence of binary digits one digit at a time ("online" input at time $t$ ), while its temporal context representation keeps a copy of the previous input ("context" at time $t-1$ ). A 100 -unit internal prediction layer attempts to predict the next input, while its temporal context representation keeps a copy of the model's prediction at time $t-1$. Figure B: After unsupervised training, the model shows fewer repetition detectors than alternation detectors ( $R: A$ ratio $<1$ ) when the actual probability of alternation is greater than $3 / 7$.

We have recently reported a biologically-motivated neural model that behaves consistently in accord with the pattern time statistics (Sun et al., 2015). The architecture of the model and the main result are shown in Figure 2. The model employs a recently developed neural algorithm for temporal integration (O'Reilly, Wyatte, \& Rohrlich, 2014). At the sensory level, a 2-unit input layer scans non-overlapping signals of heads (H) versus tails (T) one digit at a time from sequences generated by the first-order dependent Markov trials. Then, a 100-unit internal prediction layer attempts to predict the next input, with the benefit of a prior temporal context representation. The bidirectional activation dynamics between the input layer and the internal prediction layer allow us to use a single input layer for both providing inputs and receiving predictions.

Through unsupervised learning, the model was trained with binary sequences generated at various levels of probability of alternation $\left(p_{A}\right)$. After training, the model was tested with a sequence generated at the same $p_{A}$ level. By activation-based receptive field analysis, we decoded the representations on the internal prediction layer and classified its units as either repetition detectors (whose activations are significantly correlated with the input pattern either HH or TT), or alternation detectors (activations correlated with either HT or TH). We then counted the numbers of detectors and used the $R / A$ ratio (repetition over alternation) to measure the model's performance (Figure 2B).

Most interestingly, at $p_{A}=1 / 2$ (i.e., flipping a fair coin independently), despite the same training frequency of the patterns (i.e., the same mean time, see, Equation 3), the model consistently produced fewer repetition detectors than alternation detectors at a ratio of $R / A \approx .70$. We then used this $R / A$ ratio to compute the subjective probability of alternation, $p_{A}^{\prime}$,
as the model's internal representation of its experienced $p_{A}$,

$$
p_{A}^{\prime}=\frac{A}{R+A}=\frac{1}{1+R / A} \approx 0.59
$$

This $p_{A}^{\prime}$ value was consistent with the value from empirical findings. From a comprehensive review of previous studies (Falk \& Konold, 1997), a unanimous finding was that people perceived or generated random sequences with a $p_{A}^{\prime}$ value around $0.58 \sim 0.63$. Moreover, we found that this $p_{A}^{\prime}$ value directly produces the besting-fitting bias-gain parameter in an existing Bayesian model for subjective randomness of longer patterns (Goodfellow, 1938; Griffiths \& Tenenbaum, 2001).

## A Minimal Neural Network Model

The model presented in Figure 2 has a prediction layer of 100 units, and its temporal context representation is equivalent to another 100 units as in a recurrent neural network. Then, an immediate question is, how many neurons are required to produce the minimal effect of the alternation bias? Apparently, to differentiate repetition versus alternation patterns, we need at least two types of detectors. However, we also notice that if patterns are aggregated too "early", namely, combining HH with TT and combining HT with TH before counting each of the four detectors, the alternation bias would be "washed out" (see the supplementary material by Sun et al., 2015). In addition, the pairwise precedence odds shown in Figure 1C indicates that to differentiate all patterns of length two, we need at least four detector neurons.


Figure 3: An eight-unit neural network model of temporal integration. A two-unit input layer scans a sequence of binary digits one digit at a time ("online" input at time $t$ ), while its temporal context representation keeps a copy of the previous input ("context" at time $t-1$ ). The prediction layer has four units for detecting each of the four binary patterns of length two. The status of each detector is determined by the projection weights from the input units. For example, detectors HH and HT receive the same projection weights from the context input units, but detector HH receives a stronger weight from the online input unit H , and detector HT receives a stronger weight from the online input unit T .

Figure 3 shows the structure of an eight-unit model for temporal integration. The model is called "minimal" as it uses the least number of neurons to produce the minimal effect of the alternation bias in the gambler's fallacy. Its input layer is identical to the bigger model in Figure 2, with two units for scanning the "online" input at time $t$ and two units for keeping a copy of the "context" input at time $t-1$. However, its
prediction layer has only four units without explicit temporal context representation. Also different from the bigger model where the initial status of detectors was set by random weights, the detectors in the eight-unit model are initially set by distinctive projection weights from the input units. For example, detectors HH and HT receive the same projection weights from the context input units, but detector HH receives a stronger weight from the online input unit H , and detector HT receives a stronger weight from the online input unit T . In other words, given the same initial state H , detector HH tends to predict a repetition and detector HT tends to predict an alternation.

Crucially, the controlled weights allow a more precisely controlled experiment by eliminating variations produced by random weights. In the bigger model, different detectors would "naturally" emerge by a random initialization of weights. However, this may produce a disparity in the number of detector types at the initial stage (e.g., more HT detectors than HH detectors), and such a disparity has to be accounted for by averaging multiple simulations with different random initializations. This disparity is eliminated in the eight-unit model, such that the model as a whole is initially unbiased toward any of the four patterns. (We have implemented different sets of controlled weights, e.g., 0.2 versus 0.8 , or 0.4 versus 0.6 , and find that they all produce the same results.)


Figure 4: Pattern dissociation at different levels of the probability of alternation $p_{A}$ after training. The repetition detectors HH and TT only showed higher activations when the model was trained with sequences generated by $p_{A}<1 / 3$. Box plots represent distribution quantile.

The eight-unit model was trained and tested in the same way as the bigger model, and the only difference is the analyses of the test results. Instead of counting the number of detectors, we directly measure the activations of each detector given different input patterns. Figure 4 shows the main result. We first notice that after being trained with truly random sequences (i.e., $p_{A}=1 / 2$ in independent fair coin tossing), the aver-
aged activations of detectors were significantly lower when the current inputs were repetition patterns (HH or TT) than alternation patterns (HT or TH). That is, in spite of the initially unbiased representations of all patterns and the equal training frequency (i.e., the mean time is the same for all patterns at $p_{A}=1 / 2$ ), repetition patterns eventually were significantly under-represented than alternation patterns.


Figure 5: The updating trajectories of projection weights from context input units to detector units HH and TH during the training phase. The initial weight values are marked by the circles at the first trial. For example, at $p_{A}=1 / 2$, the unit initially designated as the HH detector became an TH detector after approximately 200 trials, as all its weight values from the context units switched to the opposite side of 0.5 , whereas the unit initially designated as the TH detector remained stable. The same trend was also observed between TT and HT detectors. The projection weights from the online input units remained about the same thus are not plotted here.

To locate the source of the alternation bias, we found that during the training phase, the projection weights from the context input units to each detector underwent a dramatic remapping (Figure 5), whereas the projection weights from the online input units remained about the same.

Specifically, at $p_{A}=1 / 2$, the detector unit HH initially received a weight of 0.7 from the context input unit H and a weight of 0.3 from the context input unit T (hence its initially designated detecting status). After about 200 trials, these two weights switched to the opposite sides of 0.5 , effectively "switching" the HH detector into a TH detector (but not a HT detector). Similarly, the TT detector switched to an HT detector (but not a TH detector). The directions of these switches corresponded exactly to the pairwise precedence relationship depicted in Figure 1C. At $p_{A}=2 / 3$, the switch was even more obvious. At $p_{A}=1 / 6$, it was the alternation detectors' turn to
be under-represented and switched to the repetition detectors. Finally at $p_{A}=1 / 3$, all projection weights from the context units approached then stabilized around 0.5 , so that the model eventually learned to be indifferent to the contextual information, resulting in unbiased activations for all patterns as shown in Figure 4.

Moreover, we also tested models with only a subset of particular detectors (e.g., HH versus TH only). When $p_{A}=1 / 2$, regardless of the initial pattern preference set by the projection weights, the model would eventually react indifferently to all patterns. This result indicates that in order to capture different pattern time statistics or the pairwise precedence odds between patterns of length two (Figure 1), the inhibitory competition between at least four types of detectors is required. When we tested models with only online input units, the model showed the same indifference at $p_{A}=1 / 2$. This indicates that predictive learning, namely, predicting what will happen next based on the historical context, is critical in producing the alternation bias. Together, these observations confirmed our hypothesis that this eight-unit model is a minimal model to produce the alternation bias in the gambler's fallacy.

In comparison, the alternation bias exhibited by the minimal eight-unit model in Figure 3 is in the same direction as that exhibited by the bigger model in Figure 2. However, the equilibrium point (the $p_{A}$ level where the model was indifferent to all patterns) was different (compare Figure 2 with Figure 4). Specifically, the equilibrium point was $p_{A}=1 / 3$ for the eightunit model but $p_{A}=3 / 7$ for the bigger model. This indicates that the minimal model was more sensitive to the waiting time (delay) than to the mean time (frequency) of pattern occurrences, because by Equations 1 and 3, all patterns have the same waiting time but different mean times at $p_{A}=1 / 3$,

$$
E\left[T_{\mathrm{HH} \mid \varnothing}\right]=E\left[T_{\mathrm{HT} \mid \varnothing}\right]=11 / 2, \quad E\left[T_{\mathrm{HH} \mid \mathrm{H}}\right]=3, \quad E\left[T_{\mathrm{HT} \mid \mathrm{T}}\right]=6 .
$$

In contrast, the bigger model was more "balanced" toward both statistics, because at $p_{A}=3 / 7$,

$$
E\left[T_{\mathrm{HH} \mid \varnothing}\right]+E\left[T_{\mathrm{HH} \mid \mathrm{H}}\right]=E\left[T_{\mathrm{HT} \mid \varnothing}\right]+E\left[T_{\mathrm{HT} \mid \mathrm{T}}\right]=55 / 6
$$

One particular reason for such difference is that the competition would be stronger among fewer detectors due to the homeostatic mechanism implemented in the network. This mechanism keeps individual neurons from firing too much or too little over time, which is essentially a normalization mechanism in self-organizing learning at long time scales (Bienenstock, Cooper, \& Munro, 1982; Cooper, 2000; Hebb, 1949; O’Reilly, Munakata, Frank, Hazy, \& Contributors, 2012). As a consequence, the bigger model with more neurons would be more likely to maintain diversity in the specialization of neurons thus its equilibrium point could be determined by both waiting time and mean time statistics.

## Conclusion

Overall, our results from both models suggest that pattern dissociation can naturally emerge from temporal reconstructions of the input data. Particularly with the minimal eight-neuron
model, detector neurons "reoriented attention" to the past information (i.e., remapping the projection weights from the context units), and the driving force behind such reorientation was more of the waiting time rather than of the mean time of patterns. As for the model with more neurons we reported earlier, the specialization of neurons would be more diversified thus would enable the model to develop sensitivity to both types of pattern time statistics.

The observation that both models exhibited the alternation bias is consistent with the representativeness bias underpinning the gambler's fallacy (Gilovich et al., 1985; Tversky \& Kahneman, 1974). For example, Figure 4 show that at $p_{A}=1 / 2$, the model had higher activations for alternation patterns than repetition patterns, in spite of the sequential independence of events. Critically, such bias emerged through unsupervised training without any pre-defined hypothesis structures, since both models were initially symmetrically structured, and were not provided with any prior knowledge on how different $p_{A}$ levels would affect the occurrences of different patterns.

Given the simplicity of our models, one far-reaching implication is that cognitive biases and structured abstractions can emerge early and locally at the level of sensory processing. Nevertheless, it should be noted that our models only address purely bottom-up learning mechanisms without implementing any top-down learning or higher-level representations such as beliefs or goals. For the early and locally developed biases to be maintained and utilized in later and global processes, higher-level representations and top-down structures must also be involved through a hierarchical structure of abstractions (Munakata et al., 2011; Tenenbaum et al., 2011).

Lastly, probabilistic thinking has to consider the consequence of time (Buchanan, 2013; Hawkins \& Blakeslee, 2004). Our findings suggest that rich semantics in the learning environment can be extracted by neuron populations in predictive learning through temporal integration. This learning over time would lead to the structured hypothesis spaces such as those required by Bayesian inference thus provide essential building blocks in bridging the gap between neural computations and overt behavior.

## Acknowledgments

This work was supported by the Air Force Office of Scientific Research (AFOSR) grant number FA9550-12-1-0457, the Office of Naval Research (ONR) grant number N00014-16-12111.

## References

Adolphs, R. (2015). The unsolved problems of neuroscience. Trends in Cognitive Sciences, 19(4), 173-175. doi: 10.1016/j.tics.2015.01.007
Bienenstock, E. L., Cooper, L. N., \& Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. The Journal of Neuroscience, 2(1), 32-48.
Buchanan, M. (2013). Gamble with time. Nature Physics, 9(1), 3. doi: 10.1038/nphys2520

Budescu, D. V. (1987). A Markov model for generation of random binary sequences. Journal of Experimental Psychology: Human Perception and Performance, 13(1), 25-39. doi: 10.1037/0096-1523.13.1.25
Cooper, L. N. (2000). Memories and memory: A physicist's approach to the brain. International Journal of Modern Physics A, 15(26), 4069-4082. doi: 10.1142/S0217751X0000272X

Dehaene, S., \& Brannon, E. M. (2010). Space, time, and number: A Kantian research program. Trends in Cognitive Sciences, 14(12), 517-519. doi: 10.1016/j.tics.2010.09.009
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211. doi: 10.1207/s15516709 $\operatorname{cog} 1402 \_1$
Falk, R., \& Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. Psychological Review, 104(2), 301-318. doi: 10.1037/0033-295x.104.2.301
Gigerenzer, G., \& Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102(4), 684-704. doi: 10.1037/0033295x.102.4.684
Gilovich, T., Vallone, R., \& Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17(3), 295-314. doi: 10.1016/0010-0285(85)90010-6
Goodfellow, L. D. (1938). A psychological interpretation of the results of the zenith radio experiments in telepathy. Journal of Experimental Psychology, 23(6), 601-632. doi: 10.1037/h0058392

Griffiths, T. L., \& Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. In J. D. Moore \& K. Stenning (Eds.), Proceedings of the 23 rd annual conference of the cognitive science society (pp. 370-375). Mahwah, NJ: Lawrence Erlbaum Associates.
Hahn, U., \& Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. Psychological Review, 116(2), 454-461. doi: 10.1037/a0015241
Hawkins, J., \& Blakeslee, S. (2004). On intelligence. New York: Henry Holt.
Hebb, D. O. (1949). The organization of behavior. New York: Wiley.
Lopes, L. L., \& Oden, G. C. (1987). Distinguishing between random and nonrandom events. Journal of Experimental Psychology: Learning Memory and Cognition, 13(3), 392400. doi: 10.1037/0278-7393.13.3.392

Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., \& OReilly, R. C. (2011). A unified framework for inhibitory control. Trends in Cognitive Sciences, 15(10), 453-459. doi: 10.1016/j.tics.2011.07.011
Nickerson, R. S. (2002). The production and perception of randomness. Psychological Review, 109(2), 330-357. doi: 10.1037//0033-295X.109.2.330

Oppenheimer, D. M., \& Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. Judgment and Decision Making, 4(5), 326-334.

O’Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., \& Contributors. (2012). Computational cognitive neuroscience. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu.
O'Reilly, R. C., Wyatte, D., \& Rohrlich, J. (2014). Learning through time in the thalamocortical loops. Preprint at: http://arxiv.org/abs/1407.3432.
Oskarsson, A. T., Van Boven, L., McClelland, G. H., \& Hastie, R. (2009). What's next? Judging sequences of binary events. Psychological Bulletin, 135(2), 262-285. doi: 10.1037/a0014821

Pouget, A., Beck, J. M., Ma, W. J., \& Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. Nature Neuroscience, 16(9), 1170-1178. doi: 10.1038/nn. 3495
Ross, S. M. (2007). Introduction to probability models (9th ed.). San Diego, CA: Academic Press.
Sun, Y., OReilly, R. C., Bhattacharyya, R., Smith, J. W., Liu, X., \& Wang, H. (2015). Latent structure in random sequences drives neural learning toward a rational bias. Proceedings of the National Academy of Sciences, 112(12), 3788-3792. doi: 10.1073/pnas. 1422036112
Sun, Y., Tweney, R. D., \& Wang, H. (2010). Occurrence and nonoccurrence of random sequences: Comment on Hahn and Warren (2009). Psychological Review, 117(2), 697-703. doi: 10.1037/a0018994
Sun, Y., \& Wang, H. (2010a). Gambler's fallacy, hot hand belief, and time of patterns. Judgment and Decision Making, 5(2), 124-132.
Sun, Y., \& Wang, H. (2010b). Perception of randomness: On the time of streaks. Cognitive Psychology, 61(4), 333-342. doi: 10.1016/j.cogpsych.2010.07.001
Sun, Y., \& Wang, H. (2012). Perception of randomness: Subjective probability of alternation. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (pp. 1024-1029). Austin, TX: Cognitive Science Society.
Sun, Y., \& Wang, H. (2015). Generating functions in neural learning of sequential structures. In D. C. Noelle et al. (Eds.), Proceedings of the 37th annual conference of the cognitive science society (pp. 2302-2307). Austin, TX: Cognitive Science Society.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279-1285. doi: 10.1126/science. 1192788

Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 11241131. doi: 10.1126/science.185.4157.1124

# Temporal variability in moral value judgement 

Alexandra Surdina (a.surdina@warwick.ac.uk)<br>Department of Psychology, University of Warwick Coventry CV4 7AL, United Kingdom

Adam Sanborn (a.n.sanborn@warwick.ac.uk)<br>Department of Psychology, University of Warwick Coventry CV4 7AL, United Kingdom


#### Abstract

Moral judgments are known to change in response to changes in external conditions. But how variable are moral judgments over time in the absence of environmental variation? The moral domain has been described in terms of five moral foundations, categories that appear to capture moral judgment across cultures. We examined the temporal consistency of repeated responses to the moral foundations questionnaire over short time periods, fitted a set of mixed effects models to the data and compared them. We found correlations between changes in participant responses for different foundations over time, suggesting a structure with at least two underlying stochastic processes: one for moral judgments involving harm and fairness, and another for moral judgments related to loyalty, authority, and purity.


Keywords: morality, moral foundations theory, consistency, variability

## Introduction

Morality is a vital part of who we are. A person's moral beliefs are tied into their identity (Aquino \& Reed II, 2002; Aquino et al., 2009) - humans believe that if their moral values changed, they would change (Heiphetz et al., 2016). Are people's intuitions about this correct? Are our moral values consistent over time?

Since moral beliefs tend to be associated with a person's sense of identity, we should expect people's underlying moral values to largely endure over short time periods. Yet, there have been many recent explorations of moral inconsistency. These have included manipulations of two types - manipulations of response timing, or manipulations by exposure to new information or decisions. In terms of timing, we now know that time-limited decisions appear to be more altruistic (Rand et al., 2012) and that choices can be influenced by forcing decisions at a specific point in time (Pärnamets et al., 2015), indicating that the actual decision outcome is time-sensitive. Regarding information or decisions, dishonest behaviour increases future dishonesty (Garrett et al., 2016; Engelmann \& Fehr, 2016). A morally good action makes a subsequent morally bad action more appealing and vice versa, effects known as moral cleansing and moral licensing (Merritt et al., 2010; Sachdeva et al., 2009). Exposure to a moral dilemma leads to belief revision in moral decisions that persists for multiple hours (Horne et al., 2015).

The fact that changes in external circumstances can influence the outcomes of moral decisions is hardly surprising assuming morality evolved as an adaptive strategy (Machery \& Mallon, 2010). Likewise, viewing moral judgment as a decision process, we would expect the effects of changed response timing on general decisionmaking (McClelland, 1979; Usher \& McClelland, 2001) to transfer into the moral domain. But in the absense of such manipulations, are our moral judgments fundamentally noisy? Outside of the moral domain, there is evidence in decision making research that people's decisions vary stochastically even in cases where external conditions remain constant (Mosteller \& Nogee, 1951). We are interested in exploring whether there is a corresponding moral variability beyond the actual decision process: are our moral values different from moment to moment, even in the absence of new information or manipulations of response timing?

Moral Foundations Theory (MFT) provides a way to look at this. It is based on a dominant model of morality, the social intuitionist model, according to which moral choices are made primarily intuitively and then justified post hoc (Haidt, 2001). MFT maps out the moral domain in terms of six fundamental hidden parameters that appear to capture an individual's moral judgment (Graham et al., 2009), enabling us to distinguish between conservative and liberal political profiles on the basis of an agent's foundation weights. This idea that there are foundational categories that guide intuitive moral judgement has the potential to explain people's tendency to disagree on moral issues, and predict future moral judgement based on the individual scores. If we can find a systematic structure in the stochastic changes of different foundation scores beyond merely a layer of noise, this would point towards moral variability, rather than just motor variability or variability in how the response scale is used.

In line with the aforementioned results indicating temporal consistency, moral foundation scores appear stable over longer time periods; Graham et al. (2011) tested participants again after approximately a month and found that their moral foundation scores exhibited test-retest reliability. Yet, effects such as moral licensing and moral cleansing - where the outcome of an individual's moral decision influences subsequent moral de-
cisions, even decisions made by others in their ingroup (Kouchaki, 2011), over the course of single experimental sessions and thus shorter timescales - suggest the possibility of an interaction between moral foundations. Moreover, the list of known moral foundations is likely incomplete - a view shared by moral foundations theorists (Haidt \& Joseph, 2011).

Viewing moral decisions as a sampling process from a distribution that represents an agent's moral values, we can use the framework provided by MFT to investigate hidden parameters which predict an individual's moral variability. Conversely, observing within-subject variability over time can help us understand to which extent individual moral variability reflects between-individual variability that has been used to support the existence of MFT (Graham et al., 2011). Are we all sometimes a little bit more conservative and sometimes a little bit more liberal in our moral judgments and values?

In this paper, we aim to discuss the extent to which randomness plays a role in moral judgment over time by collecting responses to the moral foundations questionnaire delivered repeatedly. We subsequently fit a set of models to the data and compare them. If the variability we observe stems merely from randomness in the decision process, we expect variation in individual responses to be explained by a single noise-generating process. We find evidence for at least two separate stochastic processes associated with different sets of moral foundations, indicating the existence of inherent variability in moral values.

## Method

## Participants

The participant pool consisted of 80 psychology undergraduate students (mean age 19 years, $90 \%$ female). 14 participants were excluded from the analysis due to wrong responses on the two 'catch' trials, as done by Graham et al. (2011).

## Materials

The original moral foundations questionnaire (MFQ30) asks participants to respond using a $1-6$ scale; to enhance precision and avoid subjects simply recalling previous answers, the participants in our task had to use a slider bar to indicate their responses instead:


In addition, our version of the questionnaire contained four further questions (see Figure 1). Those were chosen so as not to correspond in any obvious way to the five foundations measured in the MFQ30, nor to the recent addition of the liberty foundation (Graham et al., 2012; Haidt, 2012). We added these questions because we wanted the same number of presumably neutral trials as the number of foundation-related questions - the

MFQ30 includes six question for each foundation but only two neutral 'catch' items.


Figure 1: Moral foundations questionnaire. Questions added by us are marked with $\left(^{*}\right)$.

## Procedure

The questionnaire was presented six times in randomised order, with a word search task before the last two trials. In each trial, one of the two question types was displayed (see Figure 1, left and right side, respectively), along with one of the statements for that question type. Randomisation was implemented so that each statement was shown to the participant exactly once in each block: The set of questions within each block was shuffled, and presented within the block in randomised order, so no regular pattern in the order of foundations would occur.

After four blocks, a word search task ${ }^{1}$ was shown for 6 minutes to provide a timed break ${ }^{2}$ : Participants had to find and mark words in a 18 x 18 letter square filled with a selection of words and random letters, based on the WordFind.js library (Scheidel, 2012). With the exception of the timed word search task, participants provided responses at their own pace. The experiment took approximately $20-25$ minutes to complete.

## Results

Since participant responses are indicated using slider bars, foundation scores change between the blocks (participants will be unable to recall the exact position of the slider for previous trials). But beyond the expected variation resulting from differences in participant's slider operation accuracy, is there a relationship between these changes in different moral foundation scores?

## Means

As found by Graham et al. (2011), we anticipated and found our psychology undergraduate subject pool in the UK to remain largely at the liberal end of the U.S. political spectrum. Welch's t-test shows that the differences between the means for harm and fairness $(p=.16)$ and for loyalty and authority $(\mathrm{p}=.44)$ are not significant. All other pairs of means indeed differ significantly ( $\mathrm{p}<.001$ ). In particular, the first two foundation means differ significantly from the last three, with higher subject scores for harm $(M=72.9, S D=24.6)$ and fairness $(M=70.3$, $\mathrm{SD}=23.6$ ) and lower scores for loyalty $(\mathrm{M}=51.1, \mathrm{SD}=$ 27.1), authority ( $\mathrm{M}=48.8, \mathrm{SD}=26.1$ ) and purity ( $\mathrm{M}=$ 42.1, $\mathrm{SD}=28.8$ ). The between-subject standard deviation is notably larger than the within-subject standard deviation (see Figure 3), supporting the MFT framework for examining between-subject differences.

A within-subjects $\mathrm{ANOVA}^{3}$ showed a main effect of both foundation $(\mathrm{F}(5,325)=72.67, \mathrm{p}<.001)$ and block $(\mathrm{F}(5,325)=6.26, \mathrm{p}<.001)$ on average slider bar values, as well as an interaction between foundation and block $(\mathrm{F}(25,1625)=1.764, \mathrm{p}=0.011)$. But we are mainly interested in changes in the absence of new information, and Figure 5 suggests that the very first block in which the whole questionnaire was new to the participant qualitatively differs from the others. Excluding the first

[^513]

Figure 2: Spider plot of means for each foundation and block.


Figure 3: Average slider value for each response, and the average of within-subject standard deviations. The catch trials and the baseline level are marked in blue.
block from the analysis indeed makes the effects of block $(\mathrm{F}(4,260)=6.26, \mathrm{p}=.47)$ and the interaction effect between foundation and block $(F(20,1300)=1.201, \mathrm{p}=.24)$ in the ANOVA above no longer significant: While moral foundation scores differ between Block 1 and the other blocks, for the later blocks alone, this is no longer true.

## Variability over time

We also expected that within-foundation variance, i.e. the variance between participant responses to the sets of questions for each respective foundation, would decrease over time: As time passes, people's certainty which choice they will make will increase as they get more familiar with the questionnaire. Moreover, we thought we might be able to observe a shift towards more extreme values for each question over time - as people become increasingly familiar with the set of questions they will
encounter, there would be less need to for caution about new options which are more or less morally upsetting than the previous maximum or minimum, respectively.

We computed residual slider values by subtracting the within-subject mean for each foundation from the slider values for each trial. The two hypotheses above can be rephrased as: The slider residual variance for each participant and block decreases as a reflection of the increase in certainty; and the average absolute residual value increases over time as a result of the decision drifting towards the extremes.


Figure 4: Changes over time, by foundation. The colours represent the different blocks.

In fact, we found no significant effect of block number on foundation variance: Again, an ANOVA only yields significant results for the variance hypothesis $(\mathrm{F}(5,325)=18.71, \mathrm{p}<.001)$ and the absolute residual hypothesis $(\mathrm{F}(5,325)=47.4, \mathrm{p}<.001)$ if we are taking the very first block into account - here, a slight decrease after the first block can be spotted (see Figure 5). If we are looking at only the other blocks, we do not find any significant change in the variance $(\mathrm{F}(4,260)=1.90, \mathrm{p}=.11)$, nor in the absolute slider residual $(\mathrm{F}(4,260)=1.22, \mathrm{p}=.30)$.

## Between-foundation variability

One hypothesis is that changes in moral foundations that are opposed with respect to their representation on the political spectrum, such as harm and purity, will balance each other out - that is, they are negatively correlated (Fig. 6a). Each person may have a constant morality 'budget', and thus an increase in a moral foundation score will inevitably be accompanied by a decrease in others. This would imply that people's position on the liberal-conservative spectrum might not be fixed. Another hypothesis is that changes in opposing moral foundations are positively correlated (Fig. 6b). This would for instance be the case if people's moral profile was


Figure 5: Absolute value and within-subject standard deviation of slider residual over time
indeed fixed, and the sampled moral foundation scores are scaled by a time-dependent factor. Alternatively, changes in different moral foundations may not be correlated at all.


Figure 6: Relative changes in foundation scores

We did not find evidence for any of these relationships between participant scores for different foundations. On the contrary, the changes in foundation scores over time were not particularly large.

To test for interactions between changes in different foundation scores, we modelled the data using mixed effects models with a full covariance matrix and a diagonal covariance matrix, respectively. We created dummy coded variables for each foundation. Since we did not detect any notable change in the means after the first block, we now focused on the variability and removed the influence of the means entirely by modelling slider residuals: we calculated the mean slider value for each question for each participant, and subtracted it from the raw slider values. We used residuals for each question rather than for each foundation score because of the differences in responses to the different questions within each foundation (see Figure 4). Furthermore, we excluded the first block in which all information had been newly introduced from the analysis. We fitted two models to the data: First, a model including a full covariance matrix and thus allowing for interactions between the different foundations, and second, a model with a diagonal covariance matrix reflecting the assumption that sampling occurs for each foundation individually.

As a baseline model, we used a model assuming a random slider residual for each participant and block, sam-
pled from the same distribution for each foundation (random noise model). The models for the slider residual $y_{i j k l}$ of Participant $i$ in Block $j$ for a question or statement $l$ relating to Foundation $k$ are:

$$
\begin{equation*}
y_{i j k l}=u_{i j}+u_{i j k}+\varepsilon_{i j k l}, \tag{M1-M3}
\end{equation*}
$$

with $u_{i j} \sim \mathcal{N}(0, \sigma)$, and

$$
\begin{align*}
& u_{i j k}=0  \tag{M1}\\
& \left(\begin{array}{l}
u_{i j 1} \\
u_{i j 2} \\
u_{i j 3} \\
u_{i j 4} \\
u_{i j 5}
\end{array}\right) \sim \mathcal{N}\left(\mathbf{0},\left[\begin{array}{lllll}
\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\
\sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\
\sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55}
\end{array}\right]\right)  \tag{M2}\\
& \left(\begin{array}{l}
u_{i j 1} \\
u_{i j 2} \\
u_{i j 3} \\
u_{i j 4} \\
u_{i j 5}
\end{array}\right) \sim \mathcal{N}\left(\mathbf{0},\left[\begin{array}{ccccc}
\sigma_{11} & 0 & 0 & 0 & 0 \\
0 & \sigma_{22} & 0 & 0 & 0 \\
0 & 0 & \sigma_{33} & 0 & 0 \\
0 & 0 & 0 & \sigma_{44} & 0 \\
0 & 0 & 0 & 0 & \sigma_{55}
\end{array}\right]\right) \tag{M3}
\end{align*}
$$

The model M2 $\left(\chi^{2}(10)=26.98, p=.003\right)$ differs significantly from the baseline model M1. Comparing the models M1-M3 to each other suggests that M2 $(\mathrm{BIC}=71058)$ has a lower BIC value than M1 $(\mathrm{BIC}=70960)$ and M3 ( $\mathrm{BIC}=70993$ ). M2, which has a full covariance matrix, shows an interesting pattern of dependencies between the different foundation types:

| Foundation | Harm | Fair | Loya | Auth |
| :--- | :--- | :--- | :--- | :--- |
| Fair | 1 |  |  |  |
| Loya | -0.86 | -0.86 |  |  |
| Auth | -0.95 | -0.95 | 0.95 |  |
| Puri | -0.7 | -0.7 | 0.64 | 0.80 |

Responses for harm and fairness appear to be positively correlated with each other and negatively correlated with responses for the other foundations, and vice versa. This would be less surprising if it was merely capturing a between-participant relationship between foundation scores. Note however that this model describes the slider residuals which add up to zero for each foundation and participant - yet, this result suggests that participants who drag the slider bar a bit further to the right for harm-related questions than in the last block will do a similar thing with the fairness-question slider, but the opposite with sliders on loyalty, authority and purity trials.
Is there some overlap between which property of morality harm and fairness on the one hand and loyalty, authority and purity on the other hand are measuring? Since the mean foundation scores for harm and fairness, and the scores for loyalty, authority, and purity seem similar to each other (see Figure 2), we introduced alternative models that only distinguish between these two groups instead of the individual foundations.

To find out if we could confirm the five-dimensional moral foundations structure, we fitted a set of linear
mixed effects models to the data. As an alternative, we dummy-coded two foundation types (the 'individualising' foundations harm and fairness and the 'binding' foundations loyalty, authority, and purity (Graham et al., 2009)). Again, we fitted a full covariance model and a diagonal covariance model to the data, adding the two models below to our list of candidate models. They are describing the slider residual $y_{i j m l}$ of Participant $i$ in Block $j$ for a question $l$ of Foundation type $m$ :

$$
\begin{equation*}
y_{i j m l}=u_{i j}+u_{i j m}+\varepsilon_{i j m l}, \tag{M4-M5}
\end{equation*}
$$

with

$$
\begin{align*}
& \binom{u_{i j 1}}{u_{i j 2}} \sim \mathcal{N}\left(\mathbf{0},\left[\begin{array}{ll}
\sigma_{11} & \sigma_{12} \\
\sigma_{21} & \sigma_{22}
\end{array}\right]\right)  \tag{M4}\\
& \binom{u_{i j 1}}{u_{i j 2}} \sim \mathcal{N}\left(\mathbf{0},\left[\begin{array}{cc}
\sigma_{11} & 0 \\
0 & \sigma_{22}
\end{array}\right]\right) \tag{M5}
\end{align*}
$$

We find that out of these, the model M5 differs significantly from the baseline model $\left(\chi^{2}(2)=27.85, p<.001\right)$. Comparing M4 (BIC=70960) and M5 (BIC=70951) to the models above suggests that M5 is preferable to M2 and M4. Thus, it appears that from a model comparison perspective, the main distinction in the moral foundation framework lies in the two different foundation types rather than the individual foundations, and that at this level of description, between-foundation correlations do not play a prominent role.

## Discussion

We found that people showed moral variability even in the absence of new information or time pressure. This moral variability is distinguishable from response variability because we found two random processes that were associated with different sets of moral foundations. The evidence for MFT is based on an analysis of betweenindividual responses to the MFQ (Graham et al., 2011), and much of this may actually be due to the withinindividual variability that we have found. This withinindividual variability may also be what allows timing interventions to have an effect (Pärnamets et al., 2015), and might potentially even allow to influence the outcomes of value-related decisions (such as election results).

While for our dataset a simpler two-type model was preferable to the more complex model including five moral foundations, we hesitate to draw general conclusions about the number of moral foundations due to the small size and relative cultural homogeneity of our subject pool. Yet, our brief glimpse at candidates for additional foundations suggests the possibility of a wider underlying structure of which MFT has captured but a part.
A common criticism of MFT is that the known moral foundations are unlikely to capture moral judgment in its entirety (Suhler \& Churchland, 2011). We had expected
our added questions to be rated similarly irrelevant to morality as the more conservative moral foundations in our liberal subject pool. Somewhat surprisingly, the responses to our added, 'neutral' foundation appear to be less neutral overall. We chose the four additional statements in the neutral foundation because we suspected that they might turn out to be morally relevant. Figure 3 suggests that questions 2 and 5 in particular (see Figures 1 and 4) indeed resonate with our participants' values. While the act of lying may arguably be related to the purity scale, it is remarkably more morally relevant than any of the purity questions. This particularly utilitarian view on having children also appears to lie outside of the given scales.

Interesting open questions remain that reach beyond refining and expanding MFT. While we observe a range of scores for different moral foundations, we do not yet understand the actual decision process: How are different moral values integrated in a decision between options that are morally relevant for more than one moral foundation, or options that are uncertain? Which impact does moral variability have on the kinds of moral decisions we face every day?

## Acknowledgments

This research was supported by a Leverhulme Trust Doctoral Scholarship within the Bridges programme at the University of Warwick.

## References

Aquino, K., Freeman, D., Reed II, A., Lim, V. K., \& Felps, W. (2009). Testing a social-cognitive model of moral behavior: the interactive influence of situations and moral identity centrality. Journal of personality and social psychology, 97(1), 123.
Aquino, K., \& Reed II, A. (2002). The self-importance of moral identity. Journal of personality and social psychology, 83(6), 1423.
Engelmann, J. B., \& Fehr, E. (2016). The slippery slope of dishonesty. Nature Neuroscience, 19(12), 15431544.

Garrett, N., Lazzaro, S. C., Ariely, D., \& Sharot, T. (2016). The brain adapts to dishonesty. Nature Neuroscience.
Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., \& Ditto, P. H. (2012). Moral foundations theory: The pragmatic validity of moral pluralism. Advances in Experimental Social Psychology, Forthcoming.
Graham, J., Haidt, J., \& Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of personality and social psychology, 96(5), 1029.
Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., \& Ditto, P. H. (2011). Mapping the moral domain.

Journal of personality and social psychology, 101(2), 366.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review, $108(4), 814$.
Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. New York: Pantheon.
Haidt, J., \& Joseph, C. (2011). How moral foundations theory succeeded in building on sand: A response to Suhler and Churchland. Journal of Cognitive Neuroscience, 23(9), 2117-2122.
Heiphetz, L., Strohminger, N., \& Young, L. L. (2016). The role of moral beliefs, memories, and preferences in representations of identity. Cognitive science.
Horne, Z., Powell, D., \& Hummel, J. (2015). A single counterexample leads to moral belief revision. Cognitive science, 39(8), 1950-1964.
Kouchaki, M. (2011). Vicarious moral licensing: the influence of others' past moral actions on moral behavior. Journal of personality and social psychology, 101(4), 702.
Machery, E., \& Mallon, R. (2010). Evolution of morality. The moral psychology handbook, 3-46.
McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. Psychological review, 86(4), 287.
Merritt, A. C., Effron, D. A., \& Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. Social and personality psychology compass, 4(5), 344357.

Mosteller, F., \& Nogee, P. (1951). An experimental measurement of utility. Journal of Political Economy, 59(5), 371-404.
Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., \& Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. Proceedings of the National Academy of Sciences, 112(13), 4170-4175.
Rand, D. G., Greene, J. D., \& Nowak, M. A. (2012). Spontaneous giving and calculated greed. Nature, 489(7416), 427-430.
Sachdeva, S., Iliev, R., \& Medin, D. L. (2009). Sinning saints and saintly sinners the paradox of moral selfregulation. Psychological science, 20(4), 523-528.
Scheidel, B. (2012). Wordfind. https://github.com/ bunkat/wordfind.
Suhler, C. L., \& Churchland, P. (2011). Can innate, modular "foundations" explain morality? challenges for haidt's moral foundations theory. Journal of cognitive neuroscience, 23(9), 2103-2116.
Usher, M., \& McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. Psychological review, 108(3), 550.

# Enforced pointing gesture can indicate invisible objects behind a wall 

Hajime Takahashi (16rmd19@ms.dendeai.ac.jp)<br>Graduate School of Tokyo Denki University, Ishizaka. Hatoyama Saitama, 350-0311 Japan

Tetsuya Yasuda (t-yasuda@jumonji-u.ac.jp)<br>Jumonji University, 2-1-28. Sugasawa Niiza<br>Saitama, 352-8510 Japan

Harumi Kobayashi* (h-koba@mail.dendai.ac.jp)<br>Graduate School of Tokyo Denki University, Ishizaka. Hatoyama<br>Saitama, 350-0311 Japan


#### Abstract

The pointing gesture is regarded as indicating an object or location in the environment. People sometimes point to invisible objects, but the inferential mechanism is not known. This study examined comprehension of pointing with a bent index finger at an invisible object behind a wall. The experimenter pointed at an object using either typical pointing or "enforced pointing" behind a wall that was either opaque or transparent. In enforced pointing, the experimenter moved his arm in an arc movement. The participants guessed which object was being denoted. The wall was also either relatively high or relatively low. When the participants looked at typical pointing, they thought that objects both in front of the wall and behind the wall were being denoted. However, when they looked at enforced pointing, they more frequently thought that objects behind the wall were denoted. People seemed to use pragmatic knowledge on this "enforced" pointing gesture.


Keywords: gesture; declarative pointing; common ground; non-linguistic information

## Introduction

The pointing gesture usually indicates direction. If any object or location exists in the indicated direction, such object or location is interpreted as denoted. Clark (2003) discussed use of attention-getting gestures of pointing and placing. He noted that pointing at a referent and placing a referent are both useful ways to convey information about referents. In pointing, a person directs the addressee's attention to the referent object; for example, a customer may point at a package of medicine that is difficult for him or her to reach but is easy for the clerk. In placing, a person puts a referent object in the area of an addressee's attention; for example, a customer may place a package of medicine on the checkout counter where a clerk waits. These communications are possible without saying any words. To communicate smoothly, people must share mutual understanding of pointing at referents and placing referents in different situations. Clark (1996) proposed that people use "common ground" as implicit mutual knowledge in human communication. Previous research has focused mostly on language and verbally describable information included in common ground. Non-verbal information such as gestures must also be comprehended using common
ground as to how people use gestures in different situations; however, usage of gestures as common ground has not yet been thoroughly explored. Pointing gestures are often used with demonstratives (Coventry et al., 2008; 2014). Pointing gestures are also used to examine children's inferential ability (Doherty et al., 2004; Kobayashi, 2007). In these studies, the addresser can easily share information about visible objects using visual joint attention and common ground.

Kita (2003) discussed that the semiotic processes-that is, interpretation of a pointing gesture and its referent-and intended meaning of the overall action must be analyzed in interpretation of a pointing gesture. The referent of a pointing gesture can be ambiguous in many situations and in many ways. People must make correct inferences about the observed pointing gesture. Tomasello (2008) discussed that a customer points at an empty glass, and a bartender understands the request of the customer ("Please fill the glass"). In other situations-for example, a client and a glass designer - the client is pointing at an empty glass may be interested in or selects the design of the glass.

Goodwin (2003) suggests that an "activity framework" specifies which features of the environment are relevant for the ongoing activity and, hence, are likely to be the referent of a pointing gesture. Goodwin also suggests that different forms of pointing may correlate with particular types of referents.

Pointing is usually used for visible objects when an indicator and an observer can jointly attend to the same object or location. People sometimes point to invisible objects, but the inferential mechanism is not known. Kobayashi and Yasuda (2015) examined how people interpret the experimenter's pointing with a bent index finger at an invisible object behind a panel. The experimenter pointed at bottles that were placed either in front of the panel or behind the panel using a straight index finger or a bent index finger, and the participants guessed which object was being indicated. In the with-obstacle condition, bent pointing tended to be interpreted as referring to all the objects, both in front and behind the panel. However, this tendency was not observed with straight index-finger pointing. Thus, when the participants looked at
straight pointing, they thought that only the objects in front of the panel were being indicated. However, when they looked at bent pointing, they thought that objects both in front of and behind the wall were being indicated. The significance level of the shape of the pointing finger's effect was substantially large ( $\eta^{2}=0.804$ ) and the results suggest that people think bent pointing can refer to objects behind walls, but straight pointing cannot. The study suggests that people have "common ground" in their interpretation of different types of pointing.

This study examined comprehension of pointing with a bent index finger at an invisible object behind a wall. The present study examined participants' interpretation of the experimenter's pointing at an object using either typical pointing or "enforced pointing" behind a wall that was either opaque or transparent. In enforced pointing, the experimenter moved his arm in arc movement. The participants guessed which object was being denoted. The wall was also either relatively high $(28 \mathrm{~cm})$ or relatively low ( 14 cm ). We examined the effect of pointing movement because this iconic movement may suggest the indicator's intention of "overriding the wall." We examined the transparency of the wall because, if the wall is transparent, "overriding" intention may look ambiguous. The reason is that objects behind the transparent wall can be indicated without such effortful movement. However, if the wall is opaque, the "overriding" intention may be naturally understood. We also examined the height of the wall. The enforced pointing movement was more easily understood when a relatively high ( 28 cm ) wall was used than when a relatively low ( 14 cm ) wall was used.

## Methods

## Participants

Sixteen Japanese undergraduate university students ( $M$ age $=22.0$ years; $S D=0.816 ; 1$ female) participated. The experiment was conducted in accordance with Tokyo Denki University's code of ethics.

## Experimental setting

Fig. 1 shows the experimental setup. On the table, there were 4 small bottles ( $\mathrm{W}: 3 \mathrm{~cm} \times \mathrm{H}: 7 \mathrm{~cm}$ ) designated $1,2,3$, and 4 with a different color on each bottle. Bottles were placed 10 cm apart. Bottle \#1 was placed 75 cm away from the edge of the side of the table where the experimenter sat. The experimenter sat on one side of the table and wore a black sun visor during the experiment so that participants could not see the experimenter's gaze direction. The participant sat at the table at a right angle to the experimenter. The panel ( $\mathrm{W}: 14 \mathrm{~cm} \times \mathrm{H}: 28 \mathrm{~cm}$ ) was placed in the middle of the table between bottles \#2 and \#3. Participants were randomly assigned to all conditions.

The experimental conditions consisted of two types each of pointing (2: typical vs. enforced), visibility (2: visible vs. invisible), and position ( 2 : lengthwise vs. widthwise).

Regarding the pointing condition, "typical pointing" was when the experimenter pointed at the referent with his arm extended horizontally and his index finger kept bent (Fig. 2); "enforced pointing" was when the experimenter pointed at the referent with his arm moving in arc and his index finger kept bent (Fig. 3).


Fig. 1: Experimental setup in "invisible" with "lengthwise" conditions. Objects were placed 10 cm apart.


Fig. 2: Flow of typical pointing


Fig. 3: Flow of enforced pointing
Regarding the visibility condition, "visible" meant that a transparent panel was used; the experimenter and the participant could see all bottles. "Invisible" meant that a black opaque panel was used; the experimenter could not directly look at bottles \#3 and \#4. The participant could see all bottles. In the obstacle position condition, the height of the "lengthwise" obstacle was 28.0 cm , and the width was 14.0 cm . The height of the "widthwise" obstacle was 14.0 cm , and the width was 28.0 cm . For example, in the "visible" with "lengthwise" obstacle condition, there was a
small transparent lengthwise panel on the table (Fig. 4a). In the "invisible" with "widthwise" obstacle condition, there was a small black opaque widthwise panel on the table (Fig. 4b).


Fig. 4: Each type of obstacle; "a" is used in "visible" with "lengthwise" conditions and "b" is used in "invisible" with "widthwise" conditions.

## Procedure

First, the experimenter and the participant looked at all the bottles placed on the table. Then, the participant sat on the experimenter's chair and looked at the table. Then, the experimenter put the panel between bottles \#2 and \#3, and the participant again looked at the table. Thus, the participant experienced the experimenter's view in both visible and invisible obstacle conditions.

In the typical pointing with invisible widthwise obstacle condition, the experimenter put the black-opaque widthwise panel between bottles \#2 and \#3 and said to the participant, "I cannot see bottles \#3 and \#4. Now, I will point at one of the four bottles." Then, the experimenter pointed at bottle \#3 using the typical index finger. Next, the experimenter pointed at bottle \#3 continuously and said, "Now, I am pointing at something. What is the color of the bottle you would guess I am pointing at?" The participant responded orally using the bottle color. The bottle corresponded to the distance from the edge of the table: Bottle \#1's distance was 75 cm ; \#2, $85 \mathrm{~cm} ; \# 3,95 \mathrm{~cm}$; and \#4, 105 cm . All bottles were of different colors, and bottle positions were randomized. In addition, the experimenter wore a sun visor so that the participant could not see the experimenter's eye gaze.

In the enforced pointing condition, the procedure was the same as with the typical pointing except that enforced pointing was used. In the visible condition, the procedure was the same as with the invisible condition except that the visible condition was used. In each block, the experimenter pointed at bottle \#3, and there were 24 trials in all. Overall, the order of color bottles of these blocks was counterbalanced between the participants.

The experimenter's pointing was trained to show the same pointing gesture in either the typical pointing or the enforced pointing in the aspects of speed of movement. In
addition, the angle of his forefinger maintained the same shape (Fig. 2 and 3).

## Results

Fig. 5 shows the participant's responses when the experimenter pointed at the object in each pointing and each visibility condition. A 2 (Pointing: typical, enforced) $\times 2$ (Visibility: visible, invisible) $\times 2$ (Position: lengthwise, widthwise) three-way ANOVA was performed with the number of the bottle that the participant responded as the dependent measure. There was a marginally significant main effect of Pointing, $F(1,15)=3.479, p=.08, \eta^{2}=0.050$. There was also a significant interaction of Pointing $\times$ Visibility, $F(1,15)=5.497, p<.05, \eta^{2}=0.015$.

To explore the significant Pointing $\times$ Visibility interaction, the simple main effects of Pointing within each Visibility and Visibility within each Pointing were analyzed. The simple main effect test revealed that there was a significant difference between typical pointing and enforced pointing in the invisible obstacle, $F(1,15)=7.120, p<.05, \eta^{2}=0.106$. There was also a marginally significant difference between the visible and invisible obstacles in typical pointing, $F(1,15)=3.407, p=.08, \eta^{2}=0.029$. Other effects were not significant (all $p>.05$ ).

When enforced pointing was used $(M=2.958, S E=$ 0.166 ), participants interpreted a farther bottle as being indicated than when the typical pointing was used ( $M=$ $2.427, S E=0.175$ ) in the invisible obstacle condition.


Fig. 5 Participants' responses when the experimenter pointed at bottle \#3 using each type of pointing in each visibility and position condition. * denotes a significant difference. Error bars denote the standard errors of the means.

## Discussion

This study examined comprehension of pointing with an enforced movement of a bent index finger at an invisible object behind a wall. The study examined participants' interpretation of an experimenter's pointing using either typical pointing or "enforced pointing" at an object behind a wall that was either opaque or transparent. In enforced pointing, the experimenter moved his arm in an arc movement. The participants guessed which object was being denoted. The wall was also either relatively high ( 28 cm ) or relatively low ( 14 cm ).

When the participants looked at typical pointing, they thought objects both in front of the wall and behind the wall were being denoted. However, when they looked at enforce pointing with an opaque wall, they more frequently thought objects behind the wall were denoted. The height of the wall did not have any effect in this experiment.

Participants interpreted that enforced pointing could "override" the wall if the wall was opaque. They might think enforced pointing suggested an overriding trajectory (Fig. 6) to point to an invisible object behind the wall. However, for the objects behind the transparent wall, enforced pointing was not necessary. Thus, this enforced movement was sufficiently informative for participants to interpret the indicator's intention to "override" the wall. That seems to be the reason why the height of the wall had no effect.

The study suggested that people use an indicator's arm movement and the features of the environment to comprehend the referent of pointing. The result suggests that we have "common ground" in terms of interpretation of different types of pointing. Furthermore, we think the linguistic-cognitive framework presented by Relevance Theory (Sperber \& Wilson, 1995; Wilson \& Sperber, 2004) may be applied to our result. Discussing Relevance Theory from a biological perspective, Scott-Phillips (2010) stated that humans' and other animals' every signal carries a presumption of its own optimal pertinence. Here, non-verbal signals such as human gestures can be processed as relevant signals in addition to utterances.

It can be said that the study suffers from reduction of pointing situations. The study investigated only the interpretation of enforced pointing in the controlled experiment. It is necessary to study the production in addition to the interpretation of enforced pointing in a more real human environment.


Fig 6: Pointing trajectories

## Acknowledgments

We would like to thank all participants who participated in our experiment. This study was partially supported by the Japan Society for the Promotion of Science (JSPS): Grant-in-Aid for Young Scientists (B) \# 26870549 (T.Y.) and JSPS: Grant-in-Aid for Scientific Research (C) \# 16K04318 (H.K.).

## References

Clark, H. H. (1996). Using language. Cambridge UK: Cambridge University Press.
Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), Pointing. Where language, culture, and cognition meet (pp. 243-268). Hillsdale NJ: Erlbaum.
Clark, H. H., Schreuder, R., \& Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. Journal of Verbal Learning and Verbal Behavior, 22, 245258.

Coventry, K. R., Griffiths, D., \& Hamilton, C. (2014). Spatial demonstratives and perceptual space: Describing and remembering object location. Cognitive Psychology, 69, 46-70
Coventry, K. R., Valdés, B., Castillo, A., \& GuijarroFuentes, P. (2008). Language within your reach. Near-far perceptual space and spatial demonstratives. Cognition, 108, 889-895.
Doherty, M. J., Anderson, J. R., \& Howieson, L. (2009). The rapid development of explicit gaze judgment ability at 3 years. Journal of Experimental Child Psychology, 104, 296-312.

Goodwin, C. (2003). Pointing as situated practice. In S. Kita (Ed.), Pointing: Where language, culture, and cognition meet (pp. 217-241). Hillsdale NJ: Erlbaum.
Kita, S. (Ed.). (2003). Pointing: Where language, culture, and cognition meet. Hillsdale NJ: Erlbaum.
Kobayashi, H. (2007). The effect of touching object parts on learning novel object part names among young children and adults. Studies in Language Sciences, 6, 61-76.
Kobayashi H., \& Yasuda T. (September, 2015). Pointing to an invisible object behind a wall: Comprehension of pointing with a bent index finger. Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science (pp. 477-481), Torino, Italy.
Scott-Phillips, T. C. (2010). The evolution of relevance. Cognitive Science, 34(4), 583-601.
Tomasello, M. (2008). Origins of human communication. Cambridge UK: MIT Press.
Tomasello, M. (2014). A natural history of human thinking. Harvard University Press.

# Similarities Between Objects in Analogies Framed by Schema-Governed Categories 

Lucía Micaela Tavernini (MICAELATAVERNINI@ Gmail.Com)<br>IPEHCS-CONICET-UNCo. Av. Argentina 1400, 8300 Neuquén<br>Department of Psychology, University of Comahue

Máximo Trench (MAXIMO.TRENCH@Crub.Uncoma.Edu.Ar)
IPEHCS-CONICET-UNCo, Quintral 1250, 8400 Bariloche. Department of Psychology, University of Comahue

Valeria Olguín (VALITAO@Yahoo.Es)<br>Department of Psychology, University of Comahue, Irigoyen 2000, 8324 Cipolletti

Ricardo A. Minervino (MINERVINOR@Jetband.Com.Ar)<br>IPEHCS-CONICET-UNCo. Av. Argentina 1400, 8300 Neuquén<br>Department of Psychology, University of Comahue


#### Abstract

The present study was aimed at assessing the effect of object similarities on participants' evaluations of analogical quality. Results from an experimental condition in which the relations involved in the compared situations were explicitly highlighted, showed that general object similarities (membership to same category) positively affected the evaluations of analogical quality. In contrast, no such effect was found under another experimental condition in which the analogical comparisons between the same situations were framed by a schema-governed category. An analysis of participants' justifications revealed that the object similarities that were taken into account under this second condition were related to central dimensions of the schema-governed category that was used to frame the analogies. We explain these findings within the category assignment approach developed by Minervino et al., and discuss the implications of this alternative perspective of analogical reasoning for the role of similarities between entities playing several thematic roles.


Keywords: analogy; schema-governed category; semantic similarity.

## Introduction

Drawing an analogy consists in recognizing that two situations are comparable because they share a common relational structure despite their superficial differences. Frequently, the purpose of analogical comparisons is to transfer knowledge from one of those situations (the base analog) to the other (the target analog) to enhance its comprehension (Gentner, 1983; Holyoak, 1984).

The structure-mapping theory (Gentner, 1983, 1989; Gentner \& Markman, 1997) and the multiconstraint theory (Holyoak \& Thagard, 1989; Hummel \& Holyoak, 1997) have dominated the discussion about analogical reasoning since the 1980s. Due to the commonalities between these two theories we will refer to them as the standard approach. According to this approach, analogical reasoning involves, among other subprocesses, a mapping between base and target elements, and an evaluation of the quality of the
analogy. Both theories agree in that the quality of a match will be considered higher or lower to the extent that the alignment satisfies the structural constraints of one-to-one mapping and parallel connectivity (Gentner, 1989; Holyoak \& Thagard, 1989). Nevertheless, the mentioned theories differ with respect to the role of semantics in the evaluation of an analogy. Structure-mapping theory posits that relational similarity is the only factor that counts when evaluating an analogy, with object similarities playing no role at all (Gentner, 1983, 1989). On the other hand, the multiconstraint theory contends that semantic similarity between objects is also taken into account during quality evaluations (Holyoak, 1984; Holyoak \& Thagard, 1989). The mechanisms proposed by these theories for estimating semantic similarity tend to resort to general knowledge like the one represented in networks (Minervino, Oberholzer, \& Trench, 2013). The scarce available evidence (Gentner, Ratterman \& Forbus, 1993; Gentner \& Kurtz, 2006) is consistent with structure mapping theory: it shows a clear relational focus, with only a minimal influence of object similarities on judgments of analogical relatedness.

To illustrate the implications of the relational focus in a more concrete way, suppose that people were faced with a task of rating the quality of the analogies that hold between the top and bottom scenes of Figure 1. Dismissing the similarity between the objects computer mouse and laptop as two members of the entity category computer equipment, and the semantic differences between a computer mouse and a pair of socks (two objects lacking a common entity category), people would probably rate Facts A and B as equally analogous to the Key fact, because they are three cases of giving, and what counts in quality evaluations are relational similarities.

In the present research we present an alternative approach to standard theories of analogy, with implications on the role of object similarities during analogical comparisons. While we tend to agree with the multiconstraint theory in granting a role to object similarities, we posit that the computation of object properties that takes place during


Figure 1. One example of an analogy in which the base (key) and target situations (A and B) have similar relations, but hold general differences among their objects.
evaluations of analogical quality does not involve assessing the proximity between base and a target objects within a hierarchically organized network of concepts. The proposals of the standard theories may be suitable for certain kind of analogies like those used by Gentner and Kurtz (2006). For example, the analogical relatedness between John bought the candy and John took the lollipop could be decided considering the similarity between buy and take in isolation. However, standard proposals might not be appropriate in those cases in which the interaction between the propositional elements composing the analogs gives rise to exemplars of a schema-governed relational category.

Exemplars of a given schema-governed category like aggression share a common schematic structure (Gentner \& Kurtz, 2005; Goldwater, Markman \& Stilwell, 2011; Markman \& Stilwell, 2001), which could be instantiated by a wide range of exemplars, like the preschooler threw a pen to his partner, the soccer player broke a leg to his rival and the girl sank her brother's head into a bucket. The alternative approach of analogical reasoning that we are presenting here supposes that two situations would be considered analogous if they could be assigned to a common schema-governed category (Minervino et al., 2013; Oberholzer, Trench \& Minervino, 2011). Frequently, in real communicational scenarios a certain schema-governed category is explicitly introduced previous to the acts of comparing and evaluating the analogability of two situations. For example, a man could tell to a friend: "Hey, I have seen cheating-like behaviors in my wife: The other day she hid her cellphone from me". To express comprehension, the friend could reply by telling an analogous situation: "The same thing happened to me! My wife hid her credit
card records from me". On the other hand, in absence of a framing schema-governed category, the comparison could be alternatively focused in the lower-level actions implied in the compared situations (e.g., the two situations could be considered analogous just as cases of hide). While standard theories of analogies seem appropriate to explain the lowerlevel type of analogical processing, the assignment approach is focused on a higher-level type of analogical reasoning guided by schema-governed categories.

With respect to the evaluation of analogies, the assignment approach posits that relational similarity is not necessary nor sufficient. Note in above examples that the preschooler threw a pen to his partner and the soccer player broke a leg to his rival could be considered analogous as cases of aggression despite not sharing similar relations (i.e., throw is not semantically similar to brake). Furthermore, hiding the cellphone from the husband could be considered as not analogous to hiding the pie from the husband besides sharing the relation hide, due to the fact that they could not be assigned to a common schemagoverned category (e.g., cheating-like behavior).

With respect to the role of objects in analogical relatedness judgments, the assignment approach posits that object similarities are taken into account to the extent that they could modify the value that the facts under comparison displayed on one dimension of the schema-governed category to which the analogs belong. For example, in Figure 1, the assignment approach predicts that if the comparison is framed by the schema-governed category awarding, people would rate Fact B as more analogous to the Key fact than Fact A because of the closeness of those first two situations in the value that they exhibit along the importance of the awarding. Object properties that count in evaluating the analogical relatedness of the Key fact with respect to Facts A and B are those related to the importance of the awarding, for example, the price of each object. Thus, as a pair of socks has a similar price than a computer mouse, the quality of the analogy between the Key fact and the Fact B would be better rated than the quality of the analogy between the Key fact and the Fact A, in which the award consists of a more expensive laptop. This treatment of semantics differs from the one given by standard theories of analogy, because it does not consider similarities and differences between objects like those provided by general knowledge networks (e.g., the similarity between a computer mouse and a laptop as members of the category computer equipment). Instead, the mechanism proposed by the assignment's approach for computing semantic similarities takes into account those aspects of objects that become relevant as an effect of the schema-governed category that is framing the comparison. For example, the price of the objects becomes relevant because it affects the values that the compared situations exhibit in the dimension of importance of the awarding.

The objective of this study was to assess whether the object similarities in properties that affect the value of the compared situations in a certain dimension of the framing
schema-governed category have an effect on quality evaluations of the analogy, as opposed to general object similarities, operationalized as membership to same category ${ }^{1}$. With this purpose, we designed an experiment in which two groups of participants received triplet of images like the ones displayed in Figure 1, which were framed either by a relation or by a schema-governed category. While one of the target objects maintained general similarities with respect to the base object in terms of their membership to a same category (e.g., computer mouse and laptop), the object of the second target did not hold general similarities with the base object (e.g., computer mouse and pair of socks). Nevertheless, the second target object could be considered more similar than the first one with respect to the base object in terms of a property (e.g., the price of the awards) that becomes relevant under a schema-governed category framing (e.g., awarding). In light of existing evidence regarding the negligible role of objects during standard analogical comparisons, we predicted that participants whose analogies were framed by the central relation (i.e., action represented by a verb) of the compared situations would rate both targets as being equally analogous to the base situation. However, the above considerations about the likely role of schema-governed categories during analogical comparisons led us to predict that framing the analogies under a common schema-governed category would lead participants to give higher scores to target analogs whose objects were not taxonomically related to those of the base situation, but which allowed a matching with respect to the base situation in terms of their value along a relevant dimension of the framing category.

## Method

## Participants and design

Forty students of Psychology at the University of Comahue (mean age $=21.4$ years, $S D=2.08$ ) volunteered to participate in the experiment, and were randomly assigned to one of two experimental conditions. The experiment has a $2 \times 2$ repeated measures design, with the independent variable framing of the analogies (same relation vs. same schema-governed category) as a between subject factor, and general object similarity between base and target (presence vs. absence), as a within subject factor. The dependent variable was the quality evaluation of analogies.

## Procedure

Participants of the two conditions of the experiment received a brief instructional text. It anticipated that they would receive several sets of three drawings in which a first scene (Key fact) would be followed by two analogous

[^514]scenes (Fact $A$ and $B$, respectively), that is, by two "situations that are analogous to the first one in essential aspects, despite being different in others". Participants were told that they should consider whether the Fact A or B seems to them as more analogous to the Key fact, or if both could be seen as equally analogous to the Key fact. The instructional material also anticipated that each set of scenes would be preceded by a brief description of the actions involved in the three drawings of the set, and followed by a prompt to provide ratings for the extent to which Facts A and $B$ could be considered analogous to the Key fact, and a verbal justification for their scoring. After reading these instructions, participants were provided with a set of practice materials. They did not receive any feedback about the content of their responses during this practice or during the experiment itself, but they did receive feedback about how to carry out the tasks. Upon reading the instructions and solving the practice trial on a computer display, all of the participants were asked to analyze each of the subsequent triplets of visual scenes at their own pace, typing their responses into the spaces provided in the Word file, and advancing to the next screen once they had finished the tasks of each set. Participants of both groups received six critical sets of scenes interleaved by six filler sets designed to prevent participants from grasping the response pattern that was favored by the manipulation in each condition. Each trial, which appeared on a different screen of the computer file, displayed a first visual scene (the base analog, named Key fact) on the uppermost section of the screen, followed by two alternative visual scenes (the targets, named respectively Fact A and Fact B) placed one next to the other (see Figure 1). Participants in the similar relation framing condition ( $\mathrm{n}=20$ ) received all triplets preceded by a brief statement stressing the verb that applied across the scenes (e.g., "These are three instances of giving"). In contrast, participants in the similar schema-governed category framing condition $(\mathrm{n}=20)$ received all triplets preceded by a schema-governed category that could be applied to all scenes (e.g., "These are three instances of awardings").

## First task: Evaluation of the quality of the analogies

 Using a 5-point Likert scale ( $1=$ barely analogous; $5=$ completely analogous), participants were asked to rate (1) the extent to which the Key fact could be considered analogous to the Fact A, and (2) the extent to which the Key fact could be considered analogous to the Fact B.Second task: Verbal justifications of the quality scores Participants were encouraged to write down a verbal justification for the previously assigned scores. For this task, they were presented with three possible kinds of unfulfilled answers followed by a blank space: (1a) "I have assigned a higher score to the comparison between Key fact/Fact A than to the comparison between Key fact/Fact $B$ because...", (1b) "I have assigned a higher score to the comparison between Key fact/Fact B than to the comparison between Key fact/Fact A because..." or (1c) "I have assigned the same scores to both comparisons because...".

## Materials

Six critical sets of drawings were built, each one comprising a base situation and two target situations (see Table 1). The characters and the actions they were performing were identical across the three scenes of each set, and could be framed either as instantiations of a common relation (e.g., two cases of giving) or as instantiations of a common schema-governed category (e.g., two instances of awarding). In contrast to the characters and their actions, which were kept constant across the three scenes of each set, the objects involved in such actions were varied. The object of one of the targets was similar in general aspects to its corresponding base object (e.g., a computer mouse and a laptop are two exemplars of the entity category computer equipment), but could not be equated to the base object along a central dimension of the schema-governed category that could be potentially applied to both scenes (e.g., if the base and the first target were framed under the schema-governed category awarding, then the importance of winning a laptop is not comparable to that of winning a computer mouse). The objects involved in the second target were chosen to display the opposite pattern: they did not maintain general object similarities with respect to the base object, but could nevertheless be equated to the base object in certain properties that become relevant under a schema-governed category framing that could be potentially applied to both scenes (e.g., while computer mouse and a pair of socks are not similar, they constitute awards of comparable importance). The order of presentation of the critical sets and the right/left position of their corresponding targets were counterbalanced.

Table 1: Experimental materials

| Set <br> \# | Framing SGC/ <br> Relation | Description of the pictorial situations |
| :---: | :---: | :---: |
| 1 | Physical exercise/ Relocate | A man is relocating big desks (BA)/ tiny stools (TA1)/ truck wheels (TA2) inside a room |
| 2 | Bragging/ Point | A man is pointing a sports car (BA)/ bicycle (TA1)/big house (TA2) to a woman |
| 3 | Smuggling/ Load | A man is loading led TV's (BA)/ fans (TA1)/ paintings (TA2) into a truck |
| 4 | Celebration/ Open | A woman is opening a bottle of champagne (BA)/ soda (TA1) / cake box (TA2) |
| 5 | Awarding/ Give | A man is giving a computer mouse (BA) / laptop (TA1)/ pair of socks (TA2) to a woman |
| 6 | Electricity consumption/ Plug | A woman is plugging in a table lamp (BA)/ floodlight (TA1) / radio (TA2) |

Note. SGC: Schema-governed Category; BA: Base Analog; TA1: Target Analog 1; TA2: Target Analog 2.

## Results

A $2 \times 2$ ANOVA with Framing condition (same relation vs same schema governed category) as between-subjects factor and General object similarity (presence vs absence) as within-subjects factor was conducted to assess how these variables impacted the perceived quality of analogies. Main effects were neither found for condition, $F(1,38)=1.01$, $M S e=0.76, p=.31$, nor for general object similarity, $F(1,38)=3.93, M S e=0.90, p=.055)$. However, there was a significant interaction between the framing condition and general object similarity, $F(1,38)=146.32, M S e=33.58, p$ < . 01 (see Figure 2). Post-hoc Tukey HSD tests revealed that in the same relation condition the ratings of the quality of the analogy were significantly higher for items with general object similarity ( $M=3.77, S D=0.697$ ) than for items without general object similarity $(M=2.68, S D=$ $0.739, p<.01$ ). Against structure mapping, these results showed that general object similarities affected the perceived quality of analogies framed by a relation. The opposite pattern of results was observed across the same schema-governed framing condition, in which the items without general object similarity but with similar dimensional value obtained significantly higher ratings ( $M=$ 3.78, $S D=0.549$ ) than items with general object similarities ( $M=2.27, S D=0.797, p<.01$ ).

A further qualitative analysis of the verbal justifications was performed in order to explore the principles underlying the evaluations of the quality of analogies. Two independent judges, both cognitive psychologists, were instructed to classify the principles applied by participants in their justifications into one of three categories: related to general object similarities, related to a verb, or related to a dimension of the framing schema-governed category. Judges should classify the principle as related to general object similarities if participants' justifications referred object similarities and differences that were not related to the verb or to the schema-governed category used to frame each triplet of images (i.e., that could be identified and conceptualized independently of the specific framing verb or schema-governed category). Judges were told that responses of this type may include similarities and differences between intrinsic properties of objects, their functions or their taxonomic membership, and were provided with examples (e.g., a computer mouse has the same color than a laptop). The principle should be classified as related to a verb if participants mentioned similarities and differences in object properties that affect dimensions of the framing relations (e.g., a computer mouse is easier to give than a laptop). Judges should classify the principle as related to a dimension of the framing schemagoverned category if participants mentioned similarities and differences between object properties that are related to a dimension of the framing schema-governed category (e.g., receiving a laptop is a more important awarding compared to receiving a computer mouse). Judges agreed in $86 \%$ of the cases, and cases of disagreement were resolved by discussion. This qualitative analysis showed that in the


Figure 2. Interaction between framing condition and general object similarity in quality evaluations.
condition framed by a relation, participants' justifications followed a principle related to general object similarities in $75 \%$ of the cases, and a principle based on dimensions of the verbs in $25 \%$ of the cases. In contrast, in the condition framed by a schema-governed category, participants used the principle related to general object similarities only in $7 \%$ of the cases, whereas in the remaining $93 \%$ of the cases they applied the principle related to a dimension of the framing schema-governed category. None of the responses of the schema-governed framing condition referred to a principle related to the verb. This analysis provides complementary evidence that general object similarities are taken into account in the same relation framing condition, but had almost no influence over quality evaluations of the schema-governed framing condition. Moreover, justifications' analysis confirmed that object similarities that count in quality evaluations of analogies framed by a schema-governed category are those related to object properties that bear on the degree to which the facts under comparison match along a relevant dimension of the framing category.

## Discussion

The present study showed that when matched objects belong to the same category, this similarity positively influences the perceived quality of a "standard" analogy (i.e., an analogy not framed by a schema-governed category). One possible explanation of the inconsistency between this finding and previous evidence (e.g., Gentner \& Kurtz, 2006; Gentner et al., 1993) could be that in our study participants were asked to justify their ratings, and there is some evidence that justifying ratings of analogical relatedness can lead to poorer discrimination between superficial and structural aspects of analogies (Sieck, Quinn \& Schooler, 1999). In any case, the central finding of the present study was that when an analogy is framed in terms of a schema-
governed category, the object properties that matter when assessing analogical comparisons are those that affect the value of the compared situations along certain dimensions of the framing schema-governed category, as opposed to the type of object similarities considered by the standard approach.

A question that may arise from our study refers to whether the framing of analogical comparisons under schema-governed categories (as in our schema-governed condition) represents a frequent or a rare occurrence in daily real-life scenarios. An example widely discussed by the multiconstraint theory indeed suggests that schemagoverned analogies are rather frequent, and that the conceptualization of similarities and differences between objects has clear implications for the generation of analogical inferences. In the context of the Vietnam/Persian Gulf analogy, Holyoak and Thagard (1995) mention that the contrast between the jungle of Vietnam and the desert sands of Kuwait was key to predicting whether the army of Saddam could be defeated by air strikes: As opposed to the aptness of the Vietnamese jungle for concealing the army, the desert sands are not of great help. As this example clearly illustrates, the comparison is framed under the schema-governed category war, and under this framing certain properties of the object ground become relevantproperties that one would not analyze in the absence of such particular framing. It is our intuition that many analogical comparisons that take place in everyday contexts are contextually framed in terms of schema-governed categories: Is the economic crisis of 2008 analogous to that of 1930?, Are the terrorist attacks perpetrated by Muslim fanatics analogous to those of radical independists of Ireland?, Is the populism led by Donald Trump analogous to the one led by Cristina Fernández de Kirchner?

The discrimination between analogies that are processed under a schema-governed category and those that are not is relevant to designing experimental materials as well as to interpret results of existing studies on analogical thinking. Just to illustrate, Figure 3 displays one of the sets of pictures employed by Markman and Gentner (1996, p. 242) to determine whether alignable differences count more than non-alignable differences during similarity judgments. If forced to decide which of the alternative target scenes is more analogous to the base situation, readers would probably choose the one on the left, as did participants of such study. As posited by the authors, what explains this preference is the fact that while the replacement of the target object by a bird represents an alignable difference in the context of the rightmost scene, such replacement represents a non alignable difference within the left scene. Alternatively, one could sensibly argue that while the left scene has to do with the physical ability of aiming at a target, the right scene represents a case of the schema-governed category zoocide. The processing of analogies under schema-governed categories does not always require the external provision of a suitable schema-governed category. As shown by Minervino et al. (2013), schema-governed categories could be naturally activated in reasoners by the


Figure 3. Sample of materials used to determine whether alignable differences or nonalignable differences count more during similarity evaluations. Adapted from
"Commonalities and differences in similarity comparisons", by A.B. Markman \& D. Gentner, 1996, Memory \& Cognition, 2, p. 242
situations themselves. In fact, during the construction of the stimuli employed in the present study we found it hard to devise situations for which the schema-governed categories that were explicitly presented in the schema-governed category framing condition would not be automatically evoked by participants in the same relation condition.

While the present study shows that a schema-governed category framing leads people to highlight object properties (e.g., price of the award) that are relevant to analogical quality evaluations, a question that may arise is whether just activating those properties without activating the whole schema will yield the same results. An experiment that could shed light on this would involve a comparison between the schema-governed framing condition of the present experiment and another identical condition in which the woman (i.e., the patient of the awarding) was replaced by a billionaire. While the consideration of the price of the objects under the first condition allowed participants to rate the Key fact as more analogous to the Fact A than to the Fact B, we hypothesize that under the other condition people would rate both facts A and B as almost equally analogous to the Key fact. This would evidence that in the schema-governed category framing condition people are not just taking into account object properties, but are considering the interaction between object and patient properties (e.g., object's price and richness of the patient). Thus, it seems likely that people take into account the interaction between the fillers of every thematic-role (e.g., object, patient, agent, and instrument) and their relevant properties to judge the analogability of the compared situations. Future studies should assess the adequacy of these and other predictions that stem from the category assignment approach.

## Acknowledgments

This work was supported by the National Agency for Scientific and Technical Research (ANPCyT) under Grants PICT 2352 and PICT 2650, by the University of Comahue under Grants C123, C134 and B213, and by the National Council for Scientific and Technical Research (CONICET) under Grant PIP 0567.

## References

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Gentner, D. (1989). The mechanisms of analogical transfer. In S. Vosniadou \& A. Ortony (Eds.), Similarity and Analogical Reasoning. Cambridge, UK: Cambridge University Press.
Gentner, D., \& Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman \& P. W. Wolff (Eds.), Categorization inside and outside the lab. Washington, DC: APA.
Gentner, D., \& Kurtz, K. (2006). Relations, objects, and the composition of analogies. Cognitive Science, 30, 609-642.
Gentner, D., \& Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 45-56.
Gentner, D., Rattermann, M. J., \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive Psychology, 25, 431-467.
Goldwater, M.B., Markman, A.B., \& Stilwell, C.H. (2011). The empirical case for role-governed categories. Cognition, 118, 359-376.
Holyoak, K. J. (1984). Analogical thinking and human intelligence. En R. J. Sternberg (Ed.) Advances in the psychology of human intelligence. Hillsdale, N.J.: Erlbaum.
Holyoak, K. J., \& Thagard, P. R. (1989). Analogical mapping by constraint satisfaction. Cognitive Science, 13, 295-355.
Holyoak, K. J., \& Thagard, P. R. (1995). Mental leaps: Analogy in creative thought. Cambridge,MA:TheMITPress.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.
Markman, A. B., \& Gentner, D. (1996). Commonalities and differences in similarity comparisons. Memory \& Cognition, 2, 235-249.
Markman, A. B., \& Stilwell, C. H. (2001). Role-governed categories. Journal of Experimental and Theoretical Intelligence, 13, 329-358.
Minervino, R., Oberholzer, N., \& Trench, M. (2013). Global Similarity Overrides Element Similarity when Evaluating the Quality of Analogies. Journal of Cognitive Science 14, 287-317.
Oberholzer, N., Trench, M., \& Minervino, R. (2011). When lighting a candle becomes a superstition: Analogical recategorization through the application of relational categories. $33^{\text {rd }}$ Annual Meeting of the Cognitive Science Society (pp. 568-573). Boston, Massachusetts.
Sieck, W. R., Quinn, C. N., \& Schooler, J. W. (1999). Justification effects on the judgment of analogy. Memory \& cognition, $5,844-85$.

# What can Hand Movements Tell us about Audience Engagement? 

Lida Theodorou (l.theodorou@qmul.ac.uk)<br>School of Electronic Engineering and Computer Science, Mile End Road London, E1 4NS<br>Patrick Healey (p.healey@qmul.ac.uk)<br>School of Electronic Engineering and Computer Science, Mile End Road London, E1 4NS


#### Abstract

Conventional seated audiences have relatively restricted opportunities for response. Perhaps the most salient is applause but they use their hands to make other visible movements: to fix hair, adjust glasses, scratch ears. The question we address here is whether these apparently incidental movements may provide systematic clues about an audience's level of engagement with a performance. We investigate this in the context of contemporary dance performances by analysing audience hand movements in four performances at the London Contemporary Dance School. Hand movements were tracked using a reflective wristband worn by each audience member. A blob detection algorithm applied to the video recording examined whether changes in hand movement are associated with audience arousal levels to the performance. The results show that hands move least during the most preferred and most during the least preferred dance pieces. We conclude that still hands are a signal of higher levels of engagement.


Keywords: Audience; Engagement; Blob Detection; Hand movement; Handedness; Contemporary Dance.

## Introduction: Interacting with Audiences

In many live performances audiences are separated from performers; seated in the dark observing the performance. The primary conventional opportunity for members of an audience to express their satisfaction or dissatisfaction about a performance is through applause and/or cheering. Nonetheless, audiences have notoriously recruited other means of signalling their ongoing responses including the organised and carefully timed use of apparently innocent activities such as coughing (Wagener, 2012; Broth, 2011).

Our programmatic hypothesis is that audiences' ongoing responses are part of a bi-directional system of audience-performer communication that distinguishes live from recorded performance. A key motivation for this hypothesis is that performers themselves distinguish between "good" or "bad" audiences for the same performance and between moments of engagement or "lift" and moments of boredom in an audience (Healey, Oxley, Schober, \& Welton, 2009). The question this raises is what could performers be detecting in these situations that informs their dynamic sense of how well a performance is going. Here, we consider an especially restrictive case: contemporary dance. In a typical performance the audience will be in the dark, the performers behind bright lights and loud music accompanying the dancing. Audience behaviours are restricted by conventions on the types of response that it is considered acceptable to display. Performers are constrained by the need to concentrate on the physical movements required for their performance. In
contrast to live genres such as Street Performance, Stand-up Comedy or Drama there few, if any, opportunities for direct eye contact or verbal exchanges with the audience. Almost the only available channel of communication between audience and performers is body movements.

One overt physical response that is visually salient and potentially detectable by dancers is audience hand movements. Casual observation of a dance audience reveals a wide range of ongoing hand-movements by audience members involving an apparently diverse set of activities: scratching, adjusting hair, adjusting glasses, support the chin and drinking amongst others. The question addressed here is whether these movements may provide a signal of audience engagement and thereby form part of a feedback cycle between the performers and their audience.

## Measuring Audience Responses

Understanding and sensing audience responses can provide an evaluation tool to help performance directors understand how their work is received. Performance unfolds in time, making the collection of data more problematic for researchers (Schubert, Vincs, \& Stevens, 2009). However, there are a growing number of studies in dance research that use sensing technologies to examine dancer positions in time (Calvo-Merino, Glaser, Grèzes, Passingham, \& Haggard, 2005) although very little research has focused specifically on audiences (Katevas, Healey, \& Harris, 2015; Gardair, Healey, \& Welton, 2011; Vincs, Stevens, \& Schubert, 2010; Healey et al., 2009).

The most obvious way that one can measure satisfaction in audiences is from the levels of applause. Mann, Faria, Sumpter, and Krause (2013) used a mathematical model to quantify the role of social contagion in the starting and stopping of applause during a presentation. They found that the rate at which new individuals start clapping is proportional to how many people are already clapping. However, this is a measure of response after the end of the presentation rather than a concurrent response. An alternative approach asks participants to make explicit self-reports of their responses during a performance. Vincs et al. (2010) took this approach with a 'portable Audience Response Facility' (pARF), a PDA that records participant's ratings of engagement during a dance work. They found that periods of high self-reported engagement often follow choreographic surprises, and that periods of low engagement tend to be associated with more pre-
dictable dance structures (Vincs et al., 2010; Vincs, Schubert, \& Stevens, 2009). Using a post-performance methodology, Stevens et al. (2009), explored the reactions of 472 audience members as they watched contemporary dance by using an Audience Response Tool (ART) that collects responses using qualitative and quantitative questionnaires. Some of the items probed participants on their experience and enjoyment of the performance included visual and aural cues, dancer characteristics, movement, choreography, novelty, spatial elements, intellectual and emotional stimulation. Unfortunately, the act of asking participants to evaluate the performance after its end has the disadvantage of the "peak-end" effect, which shows that a measure taken immediately after an experience is strongly influenced by the peak emotion and by the emotion experienced at the end of the experience.

Two non-intrusive approaches have demonstrated connections between the movements of performers and audiences as an element of live feedback between performers and audience. Healey et al. (2009) pioneered the use of motion capture techniques in this context by exploring the intercorrelations between patterns of head movement between a seminar speaker and their audience during a seminar. The results of their study indicate that head movements of the performer are reliably triggered by head movement of audience members. Using a detailed ethnographic approach, Gardair et al. (2011) examined audience dynamics in a study of street performance in Covent Garden. She explored how passers-by notice when a street performance is happening, by first becoming watchers and then transformed into audience members. Gardair argued that people's body orientations show the spaces that people most often interact and also that people use their body torque to express their engagement levels with the performance.

## What Do Hand Movements Signal?

Although hand movements are visually salient they have a wide variety of potential causes. It is especially challenging to interpret naturally occurring audience hand behaviour without convergent verbal feedback. Most of the gesture literature focuses on explicitly designed co-speech gestures. Audience hand behaviour includes hand to face gestures or selftouch gestures (STG) that appear to lack overt, intentional design and may be performed with little or no awareness (Harrigan, Kues, Steffen, \& Rosenthal, 1987). One important class of non-speech hand movements that are relevant for audiences are self-touch gestures (STG's). According to Harrigan et al. (1987), 55\% of STG are applied to head or face, $8 \%$ are applied to the legs and $2 \%$ of STG are directed to the trunk. They are thus likely to be visible at a distance.

Studies have shown that there is an increase in selftouching behaviour in stressful and fearful situations (Butzen, Bissonnette, \& McBrayer, 2005; Heaven \& McBrayer, 2000) although Ekman and Friesen (1972) suggested that STGs may also occur when a person is relaxed. Butzen et al. (2005) found a significant increase of STG in response to a video about chiggers compared to another kind of video. In a study from Heaven and McBrayer (2000) the participants listened
to texts about leeches and canaries and then had to answer several questions. Although there were no differences between the two listening conditions there was an increase in STG for the leeches text during the answering period. Rogels, Roelen, and Van Meen (1990) found that children between 3 and 6 years showed more self-touch gestures while talking about a cartoon they had just seen than while watching the cartoon. Other studies (Grunwald, Weiss, Mueller, \& Rall, 2014) hypothesise that there is a relationship between the frequency of STG and arousal. Barroso and Feld (1986) investigated this by testing the occurrence of self-touch gestures performed with one or both hands as a function of four different auditory attention tasks. They found that with increasing complexity and attentional demands both one and two handed self-touch gestures increased. Handedness also appears to play a role. Kimura (1972) showed right handed participants perform STG's equally often with the left and the right hand. However, there is evidence that people use their right and left hand for different reasons while talking. Kipp and Martin (2009) found an association of handedness with the emotional dimensions of arousal. In particular, they found that the right hand is used more when experiencing anger and the left hand when experiencing relaxed and positive feelings. According to Roether, Omlor, and Giese (2010) the body seems asymmetric in its emotional expressivity. The left side uses higher energy and higher amplitude for emotional movements than the right side.

Hand behaviour and boredom is another relationship that might help us interpret audience hand movements. According to Kroes (2005) people experiencing boredom tend to relax their body muscles. Bull (1978) claims that there are specific head positions that characterise boredom such as drops head, turns head and head lean. Bored people also tend to use their hands to support their head or perform self-touching gestures (rubbing or clutching face). However, Kroes (Kroes, 2005) notes that this hand behaviour is a sign of low arousal but it might not be a sign of dissatisfaction. According to BianchiBerthouze, Kim, and Patel (2007), boredom is mostly associated with the decrease of body movement. However, contradictory findings show that increase of movement was associated with frustration, loss of interest and boredom (Kapoor, Burleson, \& Picard, 2007), this suggests that boredom can also correlate with episodes of high movement. In summary, we believe that the claims in the literature about movements and boredom are not entirely consistent and seem to depend a lot on the social context of the activity.

Overall, the interpretation of hand gestures is problematic however, it appears that STG are implicated in the regulation of emotional and cognitive processes. Based on the literature presented above we believe that in the context of contemporary dance audience hand movement might give us information about audience engagement to the performance. We try to investigate this by first testing the general hand behaviour patterns that evoke in an audience and then examine whether these patterns affect the audience engagement. This paper
presents some initial results relevant to audience hand motion patterns by testing the following hypotheses. The first two hypotheses examine whether hands are an important sign of audience engagement that might be detectable by the dancers.

Hypothesis 1 (H1): Audience hand movements provide a potentially salient signal of response to performers.

Hypothesis 2 (H2): If hand movements have a special status as a response signal they should diverge from other movements.

In addition, hypothesis 3 and 4 examine the relationship of hand speed with engagement and/or boredom. In contrast to the association of boredom with low movement described above, we believe that in the context of a dance performance applies the opposite. Based on this we will test the following hypotheses:

Hypothesis 3 (H3): Less movement of hands on face signals engagement.

Hypothesis 4 (H4): More movement of hands on face signals disengagement.

These hypotheses will be tested using the methodology described below.

## Performances by London Contemporary Dance School

The study presented in this paper took place at "ThePlace" theater in London where four contemporary dance pieces performed by dancers of the London Contemporary Dance School(LCD). As part of our second study on audience responses, we filmed audiences and dancers during the performance. The performance lasted for 1 hour and 40 minutes and consisted of 4,20 minute dance pieces (part 1 to part 4) and a 20 minute interval between the second and the third piece (see figure 1). Each dance was performed by LCD postgraduate students and directed by commissioned professional choreographers.


Figure 1: Performances Part 1 to Part 4 (from left to right) performed by LCD.

## Methods

## Equipment set-up

In order to be able to capture a big enough sample of the audience, we used two Basler ace (1280x1024px resolution) night vision cameras. An infrared light (IR) was attached on top of each camera so as to be able to film the audience during the dark periods of the performance. Both cameras and IR lights were placed on the theatre truss on top of the stage pointing towards the part of the audience to be filmed (See figure 2). For the filming of the dancers a JVC professional camera was
hanged from the rig facing the stage. For a synchronised double GEV camera recording we used Gecko software made by Vision Experts operated on a Windows 10 pc . Gecko gave us more data accuracy since it provides a timestamp on each frame and is able to capture 45 frames per second.


Figure 2: Plan drawing of "ThePlace's" theatre showing equipment setup.

## Hand tracking: Reflective Wristbands

Apart from filming the audience and the dancers, the methodological aim of this study was to extract the hand (wrist) movements of each audience member automatically. In order to do this, we created bracelets made of 5 mm reflective rope. A small plastic bag with two reflective bracelets together with instructions of how to wear them was placed on the arm of each theatre seat (See figure 3). Each audience member had to wear one wristband on each hand. The bracelets were only visible in the video recordings because of the IR lights shooting directly on them. This was the cheapest and easiest solution we could find to do an automatic tracking and record continuous wrist movements.


Figure 3: Front and back side of ziplock plastic bag with two reflective wristbands and how to wear instructions (left). Blob detection algorithm running on audience footage (right).

## Hand and body data extraction

A blob detection algorithm made by the blobscanner processing library was applied on audience footage to track the reflective wristbands on each frame (See right image on figure 3). Blob detection is a computer vision method that is able to detect similar regions in a digital image, such as those with the same brightness or colour, compared to surrounding regions. In our case a blob is a region of white pixels (reflective
wristbands) in the image. By applying this method on the videos, we extracted the position $x, y$ of each blob in the image or in other words the right and left wrist locations of each audience member in each frame of the performance.

However, due to the complexity of the human hand movement and the limitations of the algorithm it was impossible to get a stable continuous detection of the hands since the algorithm was unable to attach and keep the correct blob on each hand throughout the frames. Due to this limitation, the data was extracted first, the hand positions $\mathrm{x}, \mathrm{y}$ were played back on top of the footage and some manual work was done to help the algorithm pinpoint the correct blob for each hand for the whole duration of the video. By doing this we ended up with a datafile with the $x, y$ positions of left and right hand of each audience member during each dance piece. In order to be able to test the significance of the hands in the performance, we compare their behaviour with that of the rest of the body. To measure the general upper-body movement of the audience, we used an optical flow algorithm. In particular we applied the algorithm on each person separately removing the optical flow vectors of each person's hands. This gave us the upper-body movement of each person without the movement of the hands (See images on figure 4).


Figure 4: Captured areas for optical flow.

## Results

The informal observation of the video footage showed that there were very few overt responses in the audience; the most salient identifiable movements were those of bringing the hands up to the face. Overall, we extracted hand motion data from 27 audience members ( 18 females and 9 males). In order to test if audience hands produce enough movement to be detectable by the dancers (H1), we measure the duration of hands being up and down as well as the hand speed during the performance. The results indicate that people have their hands on their faces for about half of the performance ( $42 \%$ of the time have their hands up compared to $58 \%$ that they have them down) while the hands are moving faster when they are up compared to when they are down, in a resting position. We examine separately the case of hands to the face by calculating the number of times each hand is moving (fix hair, adjust glasses, scratch ears) or not (hands on chin or supporting head). We found that overall the number of times the hands are moving is approximately the same with the times the hands are still ( $48 \%$ moving, $52 \%$ still).

Therefore, it appears that the audience performs enough hand movements for the performers to detect. In order to examine if hand movement could provide us signals of audience affective state (H2), we compare its similarity to the movement of the upper-body. Figure 5 shows audience body movement and hand movement for each part averaged every 1 minute. From these two graphs, it is apparent that body and hands behave differently throughout the performance while overall there is a decrease of movement from part 1 to part 3 followed by a sharp increase in the movement of the hands at the end of part 4. From these two plots, it is apparent that audience body movement is low in part 2 compared to the other 3 parts while hand movement seems to be lowest in part 3. In summary, these findings provide us with some evidence that hand movement might be a significant audience response that might be detectable by the performers and can potentially give us information regarding audience engagement to the performance.


Figure 5: Upper-body speed (left) and hand speed (right) for each part of the performance averaged every 1 minute.

Focusing separately on each part of the performance, we next examine audience engagement levels by testing how long people keep their hands up or down and how do hands behave when they are up to the face (H3, H4). The left plot on figure 6 shows the amount of time hands are up or down in each part. It is apparent that people keep their hands on their face for longer as the performance progresses from part 1 to part 3 , although that duration decreases slightly in part 4 . The right plot on figure 6 shows the number of times hands are up (moving/not moving) for each part of the performance. From this plot, it appears that the mobility of the hands decreases as the performance progresses. Looking at the two plots, we see differences between the 4 parts of the performance. In part 1 , hands seem to be down for longer while when they are up their mobility is relatively high. Overall, part 1 presents more moving than still hands. In part 3, people have their hands up for longer, however most of the time hands are still when they are up. This means that in part 3 we have an increase in the number of still hands on face. Finally, parts 2 and 4 are somewhere in between with the only difference that in part 2 hands
are more likely to be found down while the opposite is true for part 4. In summary, it seems that there is an increase in the number of hands being on face while hands are getting stiller as the performance progresses. This result fits with the movement of the body that also decreases from part 1 to part 3. At the end of part 4 both body and hand speed increase. We compared these results with audience preference levels for each part through an online survey sent to participants with a range of familiarities to dance. The main aim of the survey was to ask people to watch the footage of the 4 dance pieces and rank them in order of preference. The order of the performances was randomised for each participant. We collected answers from 21 people ( 18 females and 3 males), 8 of which were professionally connected to dance and were watching dance more that 4 times a year. The rest were dance enthusiasts that were going to dance performances 3 to 4 times a year. The results of the survey indicate that the 2nd part of the performance is the most preferred, 3rd part comes next while 1st and 4th are the least preferred in that order. In comparison to the previously mentioned results, we observe that during the most preferred parts (part 2,3) there are more hands still on face while in the least preferred there is a higher mobility of the hands that gets more distinct at the second half of part 4, very end of the performance.


Figure 6: Duration of hands being up and down for each part of the performance (left). Number of time hands are up (Moving and Not moving) for each part of the performance (right).

An unexpected finding of this study is the different behaviour between audiences left and right hand. Overall, during the performance the left hand moves slightly faster than the right while the right hand is more likely to be found up on the face compared to the left. The first plot on figure 7 shows the mean speed of the left and right hand for each part of the performance. In parts 1 and 2 the left hand seems to move faster compare to the right while in the Part 3 and 4 is the other way around, the speed of the right hand is slightly higher compared to the left. Finally, the second plot on figure 7 shows the average number of times left and the right hand were found up for each part of the performance. This plot indicates that there is a difference between the number of times people use their left and right hand which is getting
progressively bigger from part 1 to part 4.


Figure 7: Mean hand speed of left and right hand for each part of the performance (left). Mean number of times left and right hand were up during each part (right).

## Discussion

The results described above provide us with some initial clues to the importance of overt audience reactions to contemporary dance. Like Theodorou, Healey, and Smeraldi (2016), these results show that overall, audiences have their hands up to their faces for about half the performance while the speed of hand movements varies a lot throughout. This suggests that there is audience hand movements that are both frequent and potentially detectable to the dancers. Previous studies have shown that audience faces tend to be expressionless during dance performances and so hands might be the part of the body provide signals of satisfaction or dissatisfaction (Theodorou et al., 2016). Combined with the preferences expressed in the survey, the results show that the most preferred performances are the ones that the audience moves least while during the least preferred performances hand movement increases and people perform more self-touching gestures. We interpret this finding as suggesting that people become restless and this leads to more spontaneous self-touching gestures. These observations suggest that it is actually the lack of movement that is a key signal of how engaged people are in the performance and fidgeting and spontaneous self-touching relate more to audiences boredom or nervousness (Healey, Theodorou, \& Woods, Forthcoming 2017). However, the ratings collected by the online survey can only capture overall preference levels, rather than the momentary engagement of the audience. This is something that needs to be explored in future work by, for example, showing people shorter videos from different parts of a performance instead of judging the dance piece as a whole.

An unexpected finding of the study was the systematically different behaviour of the left and right hand throughout the performance. In particular, the results indicate that overall the left hand moves faster compared to the right while the right hand is more likely to be found up. This finding is opposite to what we found in our first study which showed that people have their left hand up more times and for longer compared to the right hand. These different hand responses may indicate that people have a left-right asymmetry in their expres-
siveness when watching dance. Kipp and Martin (2009) have proposed that there is an association of gesture handedness with the emotional dimensions of pleasure and arousal. In the future we plan to examine this association further, and test the aforementioned questions of boredom and engagement using a more controlled methodology with recruited audience members.

## Acknowledgments

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) as part of the Centre for Doctoral Training in Media and Arts Technology. The authors extend a personal thank to Peter Laycock producer of "The Place" for letting us film the performances, as well as Marco Cifre,Sam Wood and Petros Koutsolampros for helping with all the technical requirements.

## References

Barroso, F., \& Feld, J. K. (1986). Self-touching and attentional processes: The role of task difficulty, selection stage, and sex differences. Journal of Nonverbal Behavior, 10(1), 51-64.
Bianchi-Berthouze, N., Kim, W. W., \& Patel, D. (2007). Does body movement engage you more in digital game play? and why? In International conference on affective computing and intelligent interaction (pp. 102-113).
Broth, M. (2011, June). The Theatre Performance as Interaction between Actors and Their Audience. Nottingham French Studies, 50(2), 113-133.
Bull, P. (1978). The interpretation of posture through an alternative methodology to role play. British Journal of Clinical Psychology, 17(1), 1-6.
Butzen, N. D., Bissonnette, V., \& McBrayer, D. (2005). Effects of modeling and topic stimulus on self-referent touching. Perceptual and motor skills, 413-420.
Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., \& Haggard, P. (2005, August). Action observation and acquired motor skills: an FMRI study with expert dancers. Cerebral cortex (New York, N.Y. : 1991), 15(8), 1243-9.
Ekman, P., \& Friesen, W. (1972). Hand Movements. Journal of Communication.
Gardair, C., Healey, P. G., \& Welton, M. (2011). Performing places. Proceedings of the 8th ACM conference on Creativity and cognition $-C \& C$ ' $11,51$.
Grunwald, M., Weiss, T., Mueller, S., \& Rall, L. (2014, April). EEG changes caused by spontaneous facial selftouch may represent emotion regulating processes and working memory maintenance. Brain research, 1557, 11126.

Harrigan, J. a., Kues, J. R., Steffen, J. J., \& Rosenthal, R. (1987, December). Self-Touching and Impressions of Others. Personality and Social Psychology Bulletin, 13(4), 497-512.
Healey, P. G. T., Oxley, R., Schober, M., \& Welton, M. (2009). Engaging Audiences. , 1-2.

Healey, P. G. T., Theodorou, L., \& Woods, P. (Forthcoming 2017). Stillness and motion: Two hypotheses about audience engagement.
Heaven, L., \& McBrayer, D. (2000). External motivators of self-touching behavior. Perceptual and motor skills(1981), 338-342.
Kapoor, A., Burleson, W., \& Picard, R. W. (2007). Automatic prediction of frustration. Int. J. Human-Computer Studies, 65, 724-736.
Katevas, K., Healey, P. G. T., \& Harris, M. T. (2015, August). Robot Comedy Lab: experimenting with the social dynamics of live performance. Frontiers in Psychology, 6(August), 1-9.
Kimura, D. (1972). Manual activity during speaking - 1. Right-handers*. Neuropsychologia, 45-50.
Kipp, M., \& Martin, J.-C. (2009, September). Gesture and emotion: Can basic gestural form features discriminate emotions? 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 1-8.
Kroes, S. (2005). Detecting boredom in meetings. University of Twente, 1-5.
Mann, R. P., Faria, J., Sumpter, D. J. T., \& Krause, J. (2013). The dynamics of audience applause The dynamics of audience applause. (June).
Roether, C. L., Omlor, L., \& Giese, M. A. (2010). Dynamics of Visual Motion Processing., 313-340.
Rogels, P. L. J. M., Roelen, E., \& Van Meen, J. (1990). The function of self-touchings, posture shifts, and motor discharges in children from 3 to 6 years of age. Perceptual and motor skills, 1169-1178.
Schubert, E., Vincs, K., \& Stevens, C. (2009). A Quantitative Approach to Analysing Reliability of Engagement Responses to Dance.
Stevens, C. J., Schubert, E., Morris, R. H., Frear, M., Chen, J., Healey, S., ... Hansen, S. (2009, September). Cognition and the temporal arts: Investigating audience response to dance using PDAs that record continuous data during live performance. International Journal of Human-Computer Studies, 67(9), 800-813.
Theodorou, L., Healey, P. G., \& Smeraldi, F. (2016). Exploring audience behaviour during contemporary dance performances. In Proceedings of the 3rd international symposium on movement and computing (p. 7).
Vincs, K., Schubert, E., \& Stevens, C. (2009). Measuring Responses to Dance. Dance Dialogues: Conversations across cultures, artforms and practices.
Vincs, K., Stevens, C., \& Schubert, E. (2010). Effects of observer experience on continuous measures of engagement with a contemporary dance work. Proceedings of the 9th Conference of the Australasian Society for Cognitive Science, 357-361.
Wagener, A. (2012). Why do people (not) cough in concerts? The economics of concert etiquette.

# Generalizing relations during analogical problem solving in preschool children: does blocked or interleaved training improve performance? 

Jean-Pierre Thibaut<br>jean-pierre.thibaut@u-bourgogne.fr<br>LEAD-CNRS, UMR 5022, University Bourgogne Franche-Comté, Pôle AAFE - Esplanade Erasme<br>21000 DIJON. FRANCE<br>Micah B. Goldwater<br>micah.goldwater@sydney.edu.au<br>The University of Sydney, School of Psychology, Brennan MacCallum Building (A18), NSW 2006 Australia


#### Abstract

Analogical reasoning, the mapping of structured relations across conceptual domains, is commonly recognized as essential to human cognition, but young children often perform poorly in the classical $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? analogical reasoning task. Particularly, young children have trouble when the objects in the task are not strongly associated with each other, and/or when there are strong associative lures among the potential answers. Here, we examine whether successive trials that repeat the same relation needed to solve the analogy can help overcome some of the challenges with weakly associated items. In the first of two experiments, our results were mixed. In the second, we simplified the design, and were able to more clearly show a benefit of repeating relations across consecutively solved problems.


Keywords: Analogical reasoning; development.

## Introduction

Analogical reasoning lies at the core of human cognition (Holyoak, 2012; Hofstadter \& Sander, 2013). It refers to the transfer of a structured set of relations from a source domain to a target domain, which can often generate insights into how to solve novel problems and generate new ideas (e.g., Gick \& Holyoak, 1980; Lindsey, Wood \& Markman, 2008). A typical way to research and assess analogical thinking ability is the $\mathrm{A}: \mathrm{B}:: \mathrm{C}: \mathrm{D}$ analogy (e.g., dog:doghouse::bird: ? solution "Nest", in which the "lives in" relation must be abstracted).

Many experiments have been devoted to the study of ontogenetic changes in analogical reasoning ability (Gentner, 1988; Holyoak, Junn, \& Billman, 1984; Richland, Morrison, \& Holyoak, 2006; Thibaut, French, \& Vezneva, 2010b). Children's analogical reasoning capacities improve as their knowledge of the involved relations, or their abilities to resist irrelevant information increase (e.g., Goswami, 1992). Several models have been proposed in order to explain these changes. They fall roughly into two subclasses: models that try to explain development of analogical reasoning by emphasizing the increase of structured knowledge about the world (Goswami, 1992) and models that emphasize the maturation of control processes, such as working memory or response inhibition (Halford, Wilson, \& Phillips, 1998; Richland et al., 2006).

Richland et al. (2006) and Thibaut and colleagues (Thibaut, French, \& Vezneva, 2008, 2010b; Thibaut, French, Vezneva, Gérard, \& Glady, 2011) posited that while
knowledge of relations is necessary for analogical reasoning, it is insufficient. They claimed that cognitive control processes are also critical for strategically inhibiting irrelevant information and responding consistently with the task main goal. Thibaut et al. interpreted their results as showing that younger children's difficulties with analogy making arose because of insufficiently developed control processes, specifically inhibition. In one experiment involving semantic $\mathrm{A}: \mathrm{B}:: \mathrm{C}$ : ? analogies with four possible responses Thibaut, French, and Vezneva, (2010b) compared weak and strong analogies (i.e., analogies in which the items of the $A: B$ and $C: D$ pairs were weakly, or strongly, associated). Results revealed poorer results in weak (e.g., shirt:suitcase::toy:box) analogies than in strong ones, especially when the number of distractor items was high (i.e., three vs. one). Importantly, the authors controlled to ensure that the children knew the semantic relations within the pair (i.e., the semantic relations between A and B, and between $C$ and $D$ ). Thus, children's failure to map the $A: B$ pair on the potential C:D target pair could not be explained by a lack of knowledge. They showed that a greater number of distractors led to poorer performance in the case of weak analogies. They suggested that for strongly associated A:B and $\mathrm{C}: \mathrm{D}$ item pairs, children were not interfered with by the semantic distractors. In contrast, when the problem involved weakly associated items, mapping the $\mathrm{A}: \mathrm{B}$ pair onto the $\mathrm{C}: \mathrm{D}$ pair requires more than simply accessing the obvious semantic dimensions of the items.
The authors characterized analogy-making as a search through a space of features and potential relations. The number of relations holding between any $A: B$ pair is potentially large because, depending on the context, any number of different relations might be relevant (French, 1995). As mentioned above, the structure of the search space and the presence or absence of competing nonanalogical solutions have an effect on the search, especially for young children, who have greater difficulty handling the cognitive load associated with a more elaborate search of the space of possible solutions.

The notion of "searching in a semantic space" was directly investigated in an eye-tracking study by Thibaut and French (2016; Thibaut, French, Missault, Gérard, \& Glady, 2011). The authors used an eye-tracker because cognitive monitoring is difficult to assess with the sole
performance measures (i.e., error measures and reaction times) that are usually used in the literature (e.g., Rattermann \& Gentner, 1998; Richland et al., 2006).

In a $\mathrm{A}: \mathrm{B}:: \mathrm{C}: \mathrm{D}$ format, they found key differences between adults and children in the temporal organization of their respective search profiles. First, adults focused on the A and B pair at the beginning of the trial, paying less or no attention to C and to stimuli in the solution set. Later they focused on C and the Target, which they compared with the semantically related distractor. By contrast, children devoted more time on C on which they actively focused during the entire trial and was used as an anchor stimulus, compared to A and B for adults, and that the Target and the semantic distractor were focused on earlier by children than by adults. These results suggest that children might fail in analogical reasoning tasks because they do not pay sufficient attention to A and B or do not include them in their search for the "one that goes with C".

This analysis led us to the central prediction of the present paper. We started with the general hypothesis that young children find it hard to follow the instructions, that is, to integrate A and B in their exploration of C and the solution set. However, as Thibaut et al (2010b) showed, strongly associated items constrain the search space sufficiently for children to readily map the relevant relation, with no ill effects of distractors. Here we examine whether presenting successive trials that require the same relation to solve the analogy will improves its use. That is, perhaps the semantic search does not need to start from scratch with weakly associated items if the relevant relation was still active in memory when the next item with the same relation is introduced. This account would predict that a blocked presentation of trials, which presents blocks of successive trial with same relation, would outperform an interleaved presentation format which would alternate between relations on successive trials (see Rohrer \& Pashler, 2010). An advantage for blocked presentation would also be consistent with structural alignment accounts that emphasize that comparing pairs of objects bound by the same relation should help abstract that relation and generalize it to more disparate sets of objects because the process of comparison itself serves to shift attention to common relations and away from superficial differences between objects (Gentner, 2010). That is, multiple strongly associated pairs objects bound by the same relation should help children recognize when that relation applies to more weakly associated objects.

On the other hand, models of memory and discrimination learning predict benefits from interleaved presentation (Rohrer \& Pashler, 2010). That is, spacing out instances of the same relation may elicit deeper processing each time it is retrieved, strengthening its memory compared to sequential presentations, which may reduce the attention paid to the repeated item (e.g., Greene, 1989). Additionally, spacing has been shown to aid not just memory, but generalization in young children (Vlach, Ankowsky, \& Sandhofer, 2012) potentially because at each new instance,
only the most relevant information is re-activated (i.e, the information common to both initial and later items), while irrelevant information is forgotten. Further, interleaving may build on the advantages of temporally spacing out examples by also filling in the temporal gaps with problems that rely on distinct structural relations. This interleaving of different relations can often aid learning and problem solving by setting up useful contrasts, sharpening the understanding of each relation.
In two experiments, we directly test the prediction that a blocked presentation will improve performance, shared by a semantic-search account and a structural-alignment account, against the prediction that interleaved presentation will improve performance, made by memory and discrimination learning accounts.

## Experiment 1

In the first experiment, we aimed to test (1) whether accuracy using a relation improved over multiple trials wherein that specific relation solved the analogy (2) whether Weak trials specifically benefitted from following Strong trials and 3. whether these benefits depended on either a Blocked or Interleaved presentation.

## Methods

## Participants

Subjects were 474 -5-year-old preschool children ( $\mathrm{M}=56$ months; range, 49 to 64 months). Their participation to the experiment was submitted to informed consent of their parents.

The subjects were equally divided into two groups: Blocked group ( $\mathrm{N}=24 ; \mathrm{M}=55$ months; range, 49-62 months) and Interleaved group ( $\mathrm{N}=23$; $\mathrm{M}=57$ months; range, $50-64$ months). Participants were randomly assigned to the blocked or the interleaved condition.

## Materials

The experiment consisted of 2 practice trials, 12 learning trials and 4 posttest trials, which occurred with a minute delay after the learning trials. (See Table 1 for the list of trials). Analogies were of the $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? format and were composed of 7 items (colored pictures; see Figure 1). The problem consisted of the A:B pair (the source), the C item (the target), and an empty rectangle. The solution set was composed of four stimuli: the analogical answer, two distractors that were semantically related to the C item, and 1 distractor that was not semantically related to C. Positions of the different alternatives were counterbalanced across trials. There were two types of analogies, called "strong" and "weak". Strong and weak analogies were defined in terms of the semantic association strength within each pair of pictures defining an analogy, that is between A and B, and between C and the analogical target. It was determined by university students. They were asked to rate to what extent each item of the pair made them think of the other one. It was stressed that the task was to rate how strongly
the two items were associated in their mind. The ratings were on a 1-to-7 scale. The strongly associated trials were composed of strongly associated A-B and C- T(arget) pairs, and the weakly associated trials were composed of weakly associated A-B and C-T(arget) pairs. The mean C-Target association strength was 3.53 ( $\mathrm{SD}=1.11$ ) for the weak analogies, and $4.89(\mathrm{SD}=1.44)$. for the strong.

In both learning conditions, there were 4 relations (tool for, produces, contains, becomes). For each relation, there were 3 learning trials, composed of 2 strong analogies and 1 weak analogy. In half of the trials, the two strong analogies were introduced before the weak trial whereas the reverse was true for the other half of trials. In the blocked condition, the three trials for one relation came in a row (either weak, strong 1, strong2, or strong1, strong2, weak, with two relations starting with a strong analogy and two relations with a weak analogy) whereas in the interleaved condition, each of the 4 relations was displayed once in a row, followed by another exemplar of the four relations. Two relations out of 4 started with a weak analogy and two relations by a strong. This was done in such a way that in both conditions the same "weak-strong-strong" or "strong-strong-weak" sequence were introduced for each relation. The 4 posttest trials were weak trials, one per relation. There were four versions of the blocked condition and four versions of the interleaved condition to counterbalance the order of the presentation of the relations, exemplars within the relations (e.g., which weak exemplar was in the post-test and which in the learning phase), which relations had strong exemplars first and which had weak.

The trials were presented to the children on a screen through a PowerPoint file.

## Procedure

Children were individually tested in their school, in a quiet room. In both the blocked and the interleaved conditions, the 7 items defining one trial were displayed simultaneously. There were two practice trials. In the first practice trial, the task was explained to children as follows: "Let me explain how it works. At first, you have to find why these two pictures [showing A and B] go well together. So, why do you think [A] goes with [B]? OK! You see this one [showing C]? It is alone. What you have to do is to find one picture in these four images [showing the four answer options] that goes well with this one [C] in the same way as this one [B] goes with [A] so the two pairs of pictures go together. Which picture goes up there [showing the empty slot] with [C] like [B] with [A]? The child gave an answer and justified his/her choice. Then, the experimenter rephrased the entire trial, explaining and emphasizing why "A and B" and "C and Target" go together for the same reason. During the second practice trial, they were asked to do the same. When children did not attend to the A:B pair while explaining their choice, they were asked to do so, and care was taken to ensure that they understood the instructions during the training trials. In the relational learning phase, they were asked to do the same thing that
was explained to them during the experiment trials and to justify their answer afterward. No feedback was given for the relational learning trials. The experiment was then interrupted for one minute. The experimenter and the child talked freely. Then, the four test trials, one per relation, were introduced as novel trials.

## Contains- Weak 1



Contains- Strong 1


analogical

semantic

semantic

unrelated

Figure 1: Two examples of analogies used for the "contains" relation, one "weak" and one "strong". Analogical: Analogical answer; Semantic: Distractor semantically related to the C item; Unrelated: distractor semantically unrelated to C.

At the end of the experiment, children's understanding of the semantic relation between A and B and between C and Target was assessed. They were shown the A:B pairs and were asked why the two items of each pair went together. The same was true for the C:Target pairs (see Thibaut et al., 2011, for more details).

## Results

We first removed all the trials in which children could not identify one of the two semantic relations, either $A: B$ or C:D. As a result, 49 trials out of 752 trials (the majority of them for "car producing exhausts) were removed from
subsequent analysis. We first ran a two-way ANOVA on the proportions of correct answers for weak analogies with Position of weak (Before strong, After strong, At test) as a within factor and Presentation (Blocked, Interleaved) as a between-subject factor. It revealed a significant main effect of Position of weak, F $(2,90)=3.50, \mathrm{p}<.05, \eta^{2}=.07$. See Table 1. A Tukey HSD test revealed that the weak analogies were marginally significantly better understood when they were introduced after strong items rather before strong ( $\mathrm{p}=$ $.056)$ or at test $(\mathrm{p}=.075)$ with $24 \%, 39 \%$ and $25 \%$ of correct answers for the weak before the strong items, after the strong and at test, respectively. There was a decline for the weak items at test, which failed to reach significance, p $=.074$, compared to the second weak item). There was no main effect of Presentation and no interaction. A three-way ANOVA was conducted on the percentage of correct answers for strong stimuli, with Type of trials (Before weak, After weak), Position (First, Second strong) as within factors, and Presentation (Blocked, Interleaved) as a between factor. It revealed a main effect of Position, F (1, $45)=6.00, \mathrm{p}<.05, \eta^{2}=.12$ ), with the second strong higher than the first strong ( $\mathrm{M}=49 \%$ and $60 \%$, for the first and second respectively). There was no main effect of Position and no significant effect of Presentation, and no interaction between any of the factors.

Table 1: E1 Means (SD's) for the proportion of accurate responses for each trial type

|  | Blocked | Interleaved |
| :--- | :--- | :--- |
| Strong 1 | $.49(.22)$ | $.49(.25)$ |
| Strong 2 | $.55(.26)$ | $.65(.20)$ |
| Weak, before Strong | $.27(.36)$ | $.22(.29)$ |
| Weak, after Strong | $.35(.40)$ | $.43(.43)$ |
| Weak, after delay | $.23(.21)$ | $.27(.20)$ |

## Discussion

Results were mixed. On the one hand, for both Strong and Weak items, there were main effects of Position, suggesting that repeating relations improves performance. While the numerical increase of Weak trials following a Strong appears to be the root of the main effect of Position for those trials, post-hoc tests specifically looking at a benefit of a Weak trial following a Strong did not find a significant advantage contrasting it with either one of the other positions alone. In addition, there was no significant difference between a Blocked and an Interleaved mode of presentation, and not a significant effect specifically for Weak trials following Strong ones. Further, even after a delay of just one minute, there was quite low performance for weak items at test, perhaps due to that interruption. Indeed participants engaged in an informal discussion with the experimenter and this might have contributed to decrease their attention.

## Experiment 2

Because of the mixed results, perhaps due to a lack of power, we simplified the design for the second experiment. First the teaching/test phase distinction was abolished. The same four relations were used with four trials each without any delay between trials. Another simplification was that the two strong analogies were always introduced before the weak analogies. The idea was to test whether weak analogies, that are more difficult than the strong analogies, would get more positive influence from strong analogies in the Blocked or in the Interleaved condition. The design of E1, with some Weak trials appearing after a delay and others before the Strong trials, may have prevented any potential benefit that a blocked presentation could provide. This simplified design will have a greater potential to detect any effect of Presentation. In addition, any effect of Presentation would show that experience with a specific relation improves performance over an above a general order effect which may simply reflect more general improvement at performing the task. The same hypotheses as in Experiment 1 apply here.

## Methods

## Participants

Subjects were 57 4-5-year-old preschool children ( $M=55.4$ months; range, 49 to 63 months). Their participation to the experiment was submitted to informed consent of their parents.

The subjects were equally divided into two groups: Blocked group ( $\mathrm{N}=29 ; \mathrm{M}=54$ months; range, 49 to 62 months) and Interleaved group ( $\mathrm{N}=28 ; \mathrm{M}=56$ months; range, 50 to 63 months). Participants were randomly assigned to the blocked or the interleaved condition.

## Materials

The same set of analogies as in Experiment 1 was used, except for three pictures that were replaced in this novel version. In the two presentation conditions, the two strong analogies were always introduced before the two weak analogies. In the Blocked, the four analogies (trials) illustrating one relation (e;g., contains) were introduced before the four analogies depicting the next relation (e.g., tools for) were introduced. In each case, the two strong trials were introduced before the two weak trials. In the interleaved case, one strong analogy from each of the four relations were first introduced. It was followed by the second strong analogy of each of the fours relations which in turn was followed by the first and the second weak analogies. There were four versions of the blocked condition and four versions of the interleaved condition in which the order of presentation of the first and second strong, and of the first of second weak was modified.

## Procedure

The same procedure as in Experiment 1 was used here, except that there was no test phase. The experiment started with two practice trials, and was followed by the 16 learning
trials. There was no feedback in the learning trials.

## Results

We ran a three-way ANOVA, with Presentation (Blocked, Interleaved) as a between factor, Analogy type (Strong, Weak), and Item Position (First, Second) as within subject factors. It revealed that strong analogies were significantly better understood than weak analogies ( 52 vs $30 \%$ ), $\mathrm{F}(1,55$ ) $=43.45, \mathrm{p}<.0001, \eta^{2}=.44$, and that the blocked presentation gave better results than the interleaved presentation, $\mathrm{F}(1,55)=8.22, \mathrm{p}<.01, \eta^{2}=.13(\mathrm{M}=46 \%$ vs $36 \%$ ). The key association strength $x$ presentation interaction was also significant, $\mathrm{F}(1,55)=4.22, \mathrm{p}<.05, \eta^{2}$ $=.07$,showing that the difference between strong and weak analogies was larger in the interleaved case than in the blocked case. However, to examine effects on the Weak Trials specifically, we conducted a Presentation (Blocked, Interleaved) X Position (First, Second) mixed-effects ANOVA, which showed a main effect of Presentation, with Blocked $(M=38 \%)$ eliciting higher accuracy than Interleaved $(\mathrm{M}=21 \%)$. $\mathrm{F}(1,55)=12.24, \mathrm{p}<.005, \eta^{2}=.18$ (no other effects approached significance). See Figure 2.


Figure 2: E2 Means and standard errors for proportion of accurate responses.

## Discussion

Overall, the experiment showed that the Blocked condition led to better results than the Interleaved condition: seeing the four trials illustrating a given relation in a row led to better results than seeing the same relation in an interleaved way. Importantly, Presentation interacted with Association strength because the difference between Blocked and Interleaved conditions was concentrated in the Weak trials when analyzed alone. That is, examining Figure 2 clearly shows that Presentation order had little effect on the Strong trials, but did effect the Weak trials. When Weak trials directly followed Strong in the Blocked condition, this elicited more accurate use of the repeated relation.

Taking the two experiments together, a clear picture starts to emerge. Both experiments showed effects of the position of individual trials, generally supporting the idea that the use of a relation in one analogy problem can constrain the semantic search in a subsequent problem. The more simple design of E2 revealed an overall advantage of blocked presentation, supporting a semantic-search or structural alignment account, and suggests that the design of E1 was not sensitive to this advantage. In addition, while potentially an overall order effect could explain the effect of position in E1, in E2 the Blocked advantage is independent of an overall order effect, as the trials were matched in terms of the number of trials preceding them.

In Experiment 2, children in the blocked condition can first discover or build the relation using the two strong trials, then apply it to the following weak items that appear immediately after, without being interfered by the other relations. This limits memory decline between the strong and the weak trials. In the interleaved case, the larger decline between strong and weak items suggests that the interval between the weak and strong items was too important to allow a strong-to-weak generalization. Or the relations interfered one with the others.

While we showed preliminary support for the Blocked advantage further research is needed to confirm this advantage, and to clarify what kind of effects, if any, exists for sequence and presentation on Strong trials. For example, in E1 the second Strong trial was performed at a higher rate than the first, but during E2, it was the reverse! Both experiments were properly counter-balanced, so this difference is not due to item-differences. Additionally, if the Blocked advantage is further confirmed, then further research needs to test how blocking trials supports analogical problem solving. At the moment it is unclear whether this potential advantage is rooted in aiding the strategic retrieval of the relevant relational representation without changing that representation (as deficits in strategic semantic retrieval often explains poor reasoning performance in young children, Whitaker et al., 2017), or whether the successive trials allow for the children to abstract the relational commonalities, creating a representation less tied to the specifics of highly associated objects. Regardless, at this point, there is little evidence that the kinds of memory processes that specifically produce spaced and interleaved advantages in other domains seem to not play a crucial role in strengthening the use of relational representations in analogical problem solving. Interleaved schedules seem most helpful when the primary challenge concerns refining representations to aid discrimination. Our data suggests that this kind of "relational fine-tuning" or discriminating among similar relations is not a primary cause for children's poor performance.

Here their effect, if any, was in favor of the Blocked trials. However, they may be crucial in helping to establish longer-lasting relational representations, given the post-test results from E1 showing the advantages from Blocked presentation may be quite short-lived. One way to test for
lasting effects of the blocked presentation, would be to add the same set of weak trials at the end of the experiment and compare how participants behave in that case. A lasting difference between the two conditions in favor of the blocked trials would be strong argument in favor of this condition. Understanding how relational representations can be robust to temporal delays is a crucial direction for future research.

## Acknowledgement

This research has been supported by a Région de Bourgogne Franche-Comté Grant Pari-2015 and Australian Research Council Grant DP150104267.

## References

Blaye, A., \& Chevalier, N. (2011). The role of goal representation in preschoolers' flexibility and inhibition. Journal of experimental child psychology, 108(3), 469-83.
Evans, T. (1968). A program for the solution of geometricanalogy intelligence test questions. In M. Minsky (Ed.), Semantic information processing (pp. 271-353). Cambridge, MA: MIT Press.
Forbus, K. D., Usher, J., \& Lovett, A. (2008). CogSketch: Open-domain sketch understanding for cognitive science research and for education. Proceedings of the fifth eurographics workshop on sketch-based interfaces and modeling.
French, R. M. (1995). The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making. Cambridge, MA: The MIT Press.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155-170.
Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child Development, 59(1), 47-59.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34(5), 752-775.
Goswami, U. (1992). Analogical Reasoning in Children. Hillsdale, NJ: Lawrence Erlbaum Associates.
Greene, R.L. (1989). Spacing effects in memory: Evidence for a two-process account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(3), 371-377.
Gruber, O., \& Goschke, T. (2004). Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes. Acta Psychologica, 115(2-3), 10521.

Halford, G. S., Wilson, W. H., \& Phillips, S. (1998). Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. The Behavioral and Brain Sciences, 21(6), 803-64.
Hofstadter, D., \& Sander, E. (2013). Surfaces and essences: Analogy as the fuel and fire of thinking. Basic Books.

Holyoak, K. J. (2012). Analogy and relational reasoning. The Oxford handbook of thinking and reasoning, 234-259.
Holyoak, K. J., Junn, E. N., \& Billman, D. O. (1984). Development of analogical problem-solving skill. Child Development, 55(6), 2042-2055.
Linsey, J. S., Wood, K. L., \& Markman, A. B. (2008). Modality and representation in analogy. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 22(02), 85-100.
Lovett, A., Tomai, E., Forbus, K. D., \& Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. Cognitive science, 33(7), 1192-231.
Richland, L. E., Morrison, R., \& Holyoak, K. J. (2006). Children's Development of Analogical Reasoning: Insights from Scene Analogy Problems. Journal of Experimental Child Psychology, 94(3), 249-273.
Rohrer, D., \& Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. Educational Researcher, 39(5), 406-412.
Thibaut, J. P., \& French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. Cognitive Development, 38, 10-26.
Thibaut, J.-P., French, R. M., Missault, A., Gérard, Y., \& Glady, Y. (2011). In the Eyes of the Beholder: What EyeTracking Reveals About Analogy-Making Strategies in Children and Adults. Proceedings of the Thirty-third Annual Meeting of the Cognitive Science Society (pp. 453-458).
Thibaut, J. P., French, R., \& Vezneva, M. (2008). Analogymaking in children: the importance of processing constraints. In B. C. Love, K. McRae, \& V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 475-480). Austin, TX: Cognitive Science Society
Thibaut, J.-P., French, R. M., \& Vezneva, M. (2010). Cognitive load and semantic analogies: Searching semantic space. Psychonomic Bulletin \& Review, 17(4), 569-74.
Thibaut, J.-P., French, R. M., Vezneva, M., Gérard, Y., \& Glady, Y. (2011). Semantic analogies by young children: testing the role of inhibition. In B. Kokinov, A. Karmiloff-Smith, \& N. J. Nersessian (Eds.), European Perspectives on Cognitive Science. New Bulgarian University Press.
Vlach, H. A., Ankowski, A. A., \& Sandhofer, C. M. (2012). At the same time or apart in time? The role of presentation timing and retrieval dynamics in generalization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(1), 246
Whitaker, K. J., Vendetti, M. S., Wendelken, C., \& Bunge, S. A. (2017). Neuroscientific insights into the development of analogical reasoning. Developmental Science.

# Generalizing novel names in comparison settings: Role of conceptual distance during learning and at test 

Jean-Pierre Thibaut (jean-pierre.thibaut@u-bourgogne.fr)<br>LEAD-CNRS, UMR 5022, Université de Bourgogne Franche-Comté, Pôle AAFE - Esplanade Erasme, 21065 Dijon, France

Arnaud Witt (arnaud.witt@u-bourgogne.fr)<br>LEAD-CNRS, UMR 5022, Université de Bourgogne Franche-Comté, Pôle AAFE - Esplanade Erasme, 21065 Dijon, France


#### Abstract

In a comparison setting (two stimuli), we tested 4- and 6-yearold children's generalization of novel names for objects. We manipulated the semantic distance between the two learning items (e.g., two bracelets versus a bracelet and a watch), and the semantic distance between the learning items and the test items (e.g., a pendant versus a bow tie). We tested whether smaller semantic distance between learning items would lead to more taxonomic (vs. perceptual) choices at test, than broader semantic distance during learning, especially in the case of distant test stimuli. Results revealed main effects of learning distance, of generalization distance and that only children aged 6 years benefited from broader semantic during learning at test. Four year-old children failed to generalize to far test stimuli even with semantically distant learning items. We discuss how conceptual distance during learning differentially affects generalization performance across age groups.


Keywords: Comparison, Distinctiveness, Conceptual development, Executive Functions.

## Introduction

When learning novel concepts for objects, children have to capture which object dimensions are important to define the corresponding concept and neglect idiosyncratic aspects, particularly irrelevant perceptual dimensions (e.g., Murphy, 2002, for a discussion). Indeed, in many cases perceptual similarities (e.g., objects from different categories displaying the same texture and/or the same color) or differences are more salient than variations along the relevant features.

Do children spontaneously generalize novel names according to perceptual similarities such as shape similarities or do they use deeper core knowledge? On one hand, there is large evidence showing that children's early words refer to deep conceptual properties. This has been shown in triad selection tasks, in which young children are shown a standard object and are later asked to choose between a categorically related object and a thematic match. -"This [standard] is a dax. Can you find another dax?". Children usually select the categorically related object (Markman, 1989, see Imai, Gentner, \& Uchida 1994, for evidence that children generalize on the basis of shape in this paradigm). On the other hand, there is evidence that, early in development, children often generalize object novel names to perceptually similar objects, especially "shape-
similar objects" (Landau, Smith, \& Jones, 1988; Smith, Jones, \& Landau, 1992).
In many cases, perceptual similarities between the standard and the test object are conceptually irrelevant. It has been shown that ignoring these salient irrelevant perceptual similarities can be challenging for young children (e.g., Augier \& Thibaut, 2013; Gentner \& Namy, 1999). Hence, understanding what situations promote nonobvious over salient properties is a crucial issue for cognitive science and concept learning. There is now considerable evidence that comparison learning situations promote generalization based on deeper conceptual properties than classical learning situations in which children are provided with only one learning exemplar.
A large body of research in both children and adults shows that comparison can highlight nonobvious shared properties. For example, Gentner and Namy (1999) used pictures of objects from familiar taxonomic categories (e.g. fruits) to teach a novel name and tested 4 -year-olds novel names extensions to other referents. In a one-standard condition (e.g. an apple introduced as a blicket) children preferred to extend the new label to a perceptually similar object (e.g. a balloon) rather than to a taxonomically-related-but-perceptually-dissimilar object (e.g. a banana). This preference was reversed when children had the opportunity to compare two standards (e.g. an apple and an orange, introduced as blickets).
The benefits of comparison have been demonstrated for object names (e.g., Gentner \& Namy, 1999; Augier \& Thibaut, 2013), adjectives (e.g., Waxman \& Klibanoff, 2000), action verbs (e.g., Childers \& Paik, 2009), names for parts (Gentner, Loewenstein, \& Hung, 2007), relational nouns (Gentner, Anggoro, \& Klibanoff, 2011; Thibaut \& Witt, 2015) and perceptual categories (e.g., Hammer, Diesendruck, Weinshall, \& Hochstein, 2009). Augier and Thibaut (2013) also obtained this positive effect of comparison with unfamiliar objects. Four- and 6-year-olds were randomly assigned to a no-comparison (one object) condition or to a comparison condition (two objects). In both conditions, the same two posttest objects were used. In the no-comparison condition, the standard - training - item had the same texture but not the same shape as one of two test objects. The standard also shared its shape but not its texture with the other test object. In the comparison condition, they pitted an unfamiliar test object displaying a perceptually non-salient dimension (texture) that was shared
with both training items against another unfamiliar object that shared a perceptually salient dimension (shape) with one of the two training standards only. In the no-comparison condition, a majority of children extended the new label to the same-shape test object. By contrast, in the comparison condition, a majority of children extended the novel name to the same-texture match rather than to a stimulus that had the same shape as one of the two standards. Taken together, these results show that comparison situations are a powerful tool for conceptually-based rather than superficially based novel name learning in children.
However, a growing body of research shows that comparisons generate cognitive costs. Recent studies on semantic analogies (Richland, Morrison, \& Holyoak, 2006; Thibaut, French, \& Vezneva, 2010a) or perceptual analogies (Thibaut, French, \& Vezneva, 2010b) (see also Richland \& Burchinal, 2013) support this cognitive costs hypothesis. In these studies, irrelevant perceptual features or semantic distractors explained part of children's performance. The hypothesis was that these experimental conditions required inhibition and flexibility. Thus, finding out nonobvious relevant relations requires inhibiting superficial irrelevant dimensions and integrating more difficult dimensions.
In this context, Augier and Thibaut (2013) manipulated the number of items to-be-compared in 4-year-old and 6-year-old children. According to the authors, introducing more evidence in favor of the target dimension (texture) also means more comparisons and more information to integrate, generating more executive costs. They included age as a factor. They hypothesized that the younger group might not benefit from increasing the number of items in the same way. Results showed that both groups benefited from the two-standard condition. However, only the older group benefited from an increased number of standards (four standards versus two standards).
In the same executive control framework, Thibaut and Witt (2015) studied relational categories with 42-month-old children. Relational categories are defined by relations between objects rather than by the intrinsic properties of the objects involved in these relations (e.g., neighbor). In this experiment, they used relational categories such as "the knife is the dax for the apple" (Gentner, Anggoro, \& Klibanoff, 2011). They manipulated the number of pairs of pictures of objects used in the training phase to illustrate a relational category ( 2,3 or 4 pairs such as an apple and a knife for "the knife is the dax for the apple") and the distance between the domains depicting the relation. For example, a knife with an apple and another knife with an orange come from close domains whereas a knife with an apple and a log with a saw come from more remote conceptual domains. In the transfer phase, results revealed that three learning pairs were better than two or four and that learning pairs from remote domains were led to better generalization than learning pairs from close domains. These results suggest that increasing the quantity of relevant information might interfere with young children's ability to abstract relevant dimensions in this type of task. More
generally, they suggest that there is an optimal number of information that can be integrated in such comparison situations. It is likely that this optimal number increases with age. The distance between domain effect suggests that a broader conceptual distance between learning exemplars helped participants abstracting the relevant relation between objects. A smaller distance between domains might have led participants to constrain the semantic domains around very similar entities (e.g., fruits) and similar operators (e.g., knifes).

## Goals of the present experiment

We examined the effect of learning and transfer distance in a comparison of real objects task (e.g., two apples, or two fruits, see Gentner \& Namy, 1999). Most former studies with real object categories contrasted no-comparison and comparison conditions. We will focus on comparison conditions and study in which condition(s) comparison leads to better conceptual generalization in a novel name learning task. A closer look at the stimuli in former studies reveals that the objects in the learning pairs come from semantic domains the semantic distance of which is not well controlled for. The same is true for the conceptually related transfer item. In other words, the distance between semantic domains in the learning items, and the distance between the learning items and the transfer items (i.e. the conceptually related target) has not been controlled as independent variables. However, it can be argued that the "width" of learning and generalization depends on the learning exemplars. There is a large body of literature showing to what extent generalization depends on the nature of the training items (Son, Landy, \& Goldstone, 2008), on the one side, and factors affecting the generalization width on the other side (e.g., Klahr \& Chen, 2011). Thus, knowing at which distance children generalize is a main issue in the study of the ontogeny of categories, subordinate, basic, and superordinate categories.
In the following experiment, we manipulated the semantic distances between both the learning items and the test items (in the generalization phase). Further, we compared two age groups (4- and 6- year-olds) in order to study cognitive resources might interact with these distances. Indeed, children of different ages might not benefit from comparison situations in the same way as a function of the distance between learning instances and the distance between learning and transfer instances. For example, it might well be that both age groups would generalize similarly in the close learning and close generalization case, whereas younger participants might encounter more difficulties to capture conceptual similarities in the case of more distant learning items and or to apply them to more distant domains.

## Methods

Participants One hundred French speaking preschoolers were tested individually in a quiet room at their school. Two age groups were recruited. The younger group was
composed of 48 children (mean age $=4$ years, 9 months; SD $=6.7$ months; range: $50-65$ months) and the older group was composed of 52 children (mean age $=6$ years, 8 months, $\mathrm{SD}=3.8$ months, range: $74-87$ months) were randomly assigned to one of the two experimental conditions with 52 (close learning items) or 48 (far learning items) children per condition. Informed consent was obtained from their school and their parents.

Design Four and six-year-old children were compared. This factor was crossed with learning distance (Close vs. Far learning, between subject factor) and Generalization (Close vs. Far generalization, within subject factor).

Materials Seven sets of six objects were created for each distance condition (close or far) (See Table 1). Each picture was displayed on a 8 cm by 8 cm piece of cardboard. Each set corresponded to one category of objects (e.g., clothing accessories, food, tools, etc.). The learning pair was composed of one learning object and either a close training object (close learning condition), or a more distant training object (far learning condition) (see Figure 1). Thus, we


Figure 1: Example of a stimulus set and instructions adapted for the four experimental conditions crossing the learning distance (Close vs. Far learning) and Generalization (Close vs. Far generalization) factors.
manipulated the conceptual distance between the two training objects that were compared in each learning condition (Close or Far) in our comparison paradigm. For each object category (e.g., clothing accessories), the close learning objects were composed of perceptually and semantically close items (e.g., a bracelet - a curb chain), while the far pairs were composed of perceptually similar but conceptually more distant items (e.g., a bracelet - a watch) (see Table 1). The two test pictures consisted of two objects in both the close and the far generalization conditions: an item that was perceptually similar but semantically unrelated to the two training items (e.g., a tire in our bracelet case) and a taxonomic choice. As a function of the generalization condition, close or far, the taxonomic
choice was semantically close or more distant to the learning items (e.g., a jewel pendant in the close generalization case, or a bow tie in the far generalization case). Figure 1 depicts the objects used to instantiate the close and far learning distance and the close and far generalization conditions for the "clothing accessories" object category.

Table 1: List of items for the close and far conditions

|  | Learning |  |  | Generalization |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Initial item | close item | far item | percetive | close taxonomic | far taxonomic |
| SET 1 |  |  |  |  |  |  |
| "clothing accessories" | bracelet | curb chain | watch | tire | pendant | bow tie |
| "tools" | hamer1 | hamer2 | axe | ostrich head | pair of pincers | chainsaw |
| "clothing" | sock1 | sock2 | jeans | pipe | sweater | hat |
| "food" | apple1 | apple2 | cherry | bulb | banana | beefsteak |
| "animals" | ladybird1 | ladybird2 | beetle | ball | butterfly | duck |
| "music player" | guitar1 | guitar2 | cello | bottle | keybord | hi-fi |
| "game/toy" | ball1 | ball2 | cuddly toy | orange | Lego | video game |
| Set2 |  |  |  |  |  |  |
| "food" | pear1 | pear2 | strawberry | candle | pineapple | fry |
| "food2" | pumpkin1 | pumpkin2 | tomato | ball | cucumber | grilled chicken |
| "house tools/appliances | "house |  |  |  |  |  |
| "animals" | snake 1 | snake2 | lizard | rope | alligator | bird |
| "vehicle" | bike1 | bike2 | scooter | glasses | rollerblade | boat |
| "office items" | pencil1 | pencil2 | ruler | candy cane | scissors | laptop |
| "clothing accesorries" | knit cap1 | knit cap2 | hardhat | turtle | crown | boots |

Independent similarity ratings from 54 students confirmed that the close learning object condition were conceptually closer one to the other than the objects composing the far learning pairs, $t(26)=3.98, p<.001$, and that close generalization stimuli were semantically more similar to the two learning stimuli than far generalization stimuli, $t(26)=$ $6.86, p<.001$. For the purpose of our experiment it is crucial that semantically related generalization items are perceptually less similar to the learning items than the perceptually similar lures. Perceptual similarity ratings revealed that the perceptual choices were perceptually more similar to the learning items than the semantically related choices (taxonomic choices) in both the close and the far conditions, $t(26)=14.03, p<.001$ and $t(26)=18.49, p<$ .001. Importantly, we also performed perceptual similarity and conceptual similarity ratings between the close learning stimuli (e.g., two apples) and the far learning stimuli (e.g., an apple and a cherry) on the one side and the taxonomically related generalization item. They showed that overall the generalization stimuli were equally distant to both types of learning items. This was true for both types of generalization items: close generalization items, perceptual distance, $t(26)=.70, p=.46$, semantic distance, $t(26)=$ $1.21, p=.41$; far generalization: perceptual distance, $t(26)$ $=.24, p=.65$, semantic distance, $t(26)=.43, p=.89$. This
is central because we want to avoid that performance differences between close and far generalization items to be due to perceptual but also semantic similarity differences between learning items. We included semantic similarity differences in order to keep only taxonomic distance influence. For example, if we get a difference between close and far generalization items (e.g. between jewel pendent and bow tie) we do not want it to be due to other semantic information (e.g., the fact that the jewel pendent would be more thematically related to bracelet than the bow tie) than the taxonomic distance.
Each learning pair was randomly associated with one out of 14 two-syllable novel names (e.g., youma, buxi, dajo, zatu, sepon, xanto, vira, etc.) (see procedure).

## Procedure

The experiment started with two practice trials. They were followed by fourteen experimental trials presented in a random order. Each standard learning stimulus was introduced with a novel count noun (Landau, Smith \& Jones, 1988) (e.g. "this is a buxi" and "this is a buxi TOO" for the other standard). A puppet named Yoshi was used in order to make the task more attractive for children and to make the use of non-words to refer to known objects more meaningful with the following instructions: "In this game we are going to learn the language of Yoshi. Yoshi is living far away from here". The objects were presented sequentially and were left in view during the entire trial. The two learning stimuli were presented in a row and their location was determined randomly. The forced-choice test phase was identical in all conditions. The two test objects (i.e., the perceptual and the taxonomic matches) were introduced and the child was asked to point to the one which was also a member of the category (e.g., "Show me which one of these two is ALSO a buxi").

## Results

We performed a 2 Age (4 vs. 6-year-olds) x 2 Training distance (Close vs. Far distance) x 2 Generalization (Close vs. Far) ANOVA on the percentage of taxonomic choices (see Figure 2). Age and Training distance were betweensubject factors and Generalization a within-subject factor. There were significantly more taxonomic choices in the Far training condition ( $M=64.4 \%$; $S D=22.74$ ) than in the Close training condition ( $M=51.3 \% ; S D=28.73$ ), $F(1,96)$ $=6.06, p=.016, \eta_{P}^{2}=.06$. The main effect of Age was not significant, $F(1,96)=1.14, p=.29, \eta_{P}^{2}=.01$ (4-years: $M=$ $54.9 \%$; 6-years: $M=60 \%$ ) and Training distance did not interact with Age, $F<1$. In addition, children performed better in the Close Generalization condition ( $M=64.79 \%$; $S D=29.69$ ) than in the Far Generalization condition ( $M=$ $51.4 \% ; S D=28.42), F(1,96)=29.79, p<.001, \eta_{P}^{2}=.24$, but the Generalization effect did not interact neither with Age, $F<1$, nor with Training distance, $F<1$, and the triple Generalization x Age x Training distance interaction effect did not reach significance, $F<1$.

When comparisons with chance were run, student-t tests for independent groups with Bonferroni correction for multiple comparisons revealed that children performed above chance in the Close generalization condition, in the Far learning condition only, (respectively, 4 years: $t(23)=$ $3.49, p=.002$; 6 years: $t(25)=6.82, p<.001)$ but not when the learning items were conceptually close (respectively, 4 years: $t(23)=1.09, p=.28 ; 6$ years: $t(23)=1.04, p=.30)$. Interestingly, in the Far Generalization condition, only the 6-year-old children performed significantly above chance, $t(25)=3.29, p<.0125$, while the performance of the younger children did not differ from chance, $t(23)=.63, p=$ . 53.


Figure 2: Percentage of Taxonomic choices as a function of the conceptual distance between during learning (close vs.
far learning distance) and at test (close vs. far generalization). The error bars correspond to one standard error and the dashed line represents chance level (50\%).

Taken together these findings show a clear impact of conceptual distance in our comparison framework, the Far learning condition giving more taxonomic choices than the Close learning condition. Our results also show that the Close generalization is easier than the Far generalization condition. Even though our results revealed no main effect of Age effect and interaction of this factor with learning distance, comparisons with chance confirm the beneficial role of conceptual distance between the learning items and suggested that in Far Generalization contexts only the older children may benefit of comparison between conceptually distant learning items.

## Discussion

As mentioned in the introduction, there is a large body of studies showing that young children generalize novel names according to shape when only one standard stimulus is introduced in the learning phase. Our study capitalized on the idea that comparison situations during lexical learning favor deeper generalizations based on less obvious features that will, as a result, favor taxonomic generalization. However, Augier and Thibaut (2013) showed that
comparison situations generated cognitive costs that might prevent younger children from using all the available information. This result suggested that the effect of comparison on generalization depends on the ease of processing dimensional similarities and differences. Our rationale was that the deep commonalities in close learning items can easily be accessed because of many conceptual commonalities. However, this situation might have provided little information regarding conceptual similarities subtending generalizations to broader categories. On the contrary, comparisons between distant learning items may be more difficult to unify conceptually because conceptual similarities would be more difficult to abstract. By contrast, comparisons might provide more abstract knowledge supporting broader generalizations. Because younger children might encounter more difficulties to capture conceptual similarities in the case of distant learning items and or to apply them to distant domains, we hypothesized that conceptual distance during learning and at test might differentially impact benefits of comparison across groups of age. World knowledge might also contribute to the difference between age groups, since older children have more knowledge regarding the objects than younger children.
Our results showed that both age groups benefited from broader inter-item conceptual distance during the learning phase since they perform better in the far learning case than in the close one. However, close generalization was better than the far generalization. Taken together, these two results suggest that broader learning range lead to better close generalization. The fact that only the older children performed above chance in the far generalization condition in the far learning case suggests a development from, first, a better performance in the case of broad learning distance to, second, a progressively better performance in the generalization width. This last result suggests that if all age groups were able to benefit from conceptual distance during the learning phase, the benefit is probably qualitatively different across age groups. We think that far learning allowed both groups to defocus from perceptual similarities and to access basic conceptual similarities ("is a jewel"), while far learning would help older children to abstract superordinate properties ("is a clothing accessory"), making the former able to perform correct taxonomic choices only in the close generalization condition ("the pendant is a jewel too"), while the latter were able to generalize in the close as well as in the far condition (" the pendant and the bow tie are clothing accessories too").
Importantly, this pattern of results backs up the classical result in developmental psychology that superordinate categories are more difficult to learn than basic level categories (Mervis \& Rosch, 1981; Murphy, 2002) and decomposes the sources of this difficulty. Children might have more difficulties to generalize to broad categories rather than to abstract from broad conceptual distances during learning. Here, we used a perceptual lure. It should be interesting to study how participants abstract categories
that are less grounded in perceptually similar instances of the same category.
To conclude, our study suggests that making appropriate use of comparison might entirely depends on tiny differences along the properties of what is compared and on executive capacity to process them. This finding has important implication about the role comparison plays in learning. Indeed, the executive constraints on comparison processing might explain under which conditions comparison can or cannot be fruitful.

## Acknowledgments

The authors wish to thank the Conseil Regional de Bourgogne for its financial support (PARI Grant), the University of Bourgogne Franche-Comté (BQR Grant) Pauline Bonnet, Marion Mallot, for collecting and analyzing the data.

## References

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. Child Neuropsychology, 8(2), 71-82.
Augier, L., \& Thibaut, J.P (2013). The benefits and costs of comparisons in a novel object categorization task: Interactions with development. Psychonomic Bulletin \& Review.
Christie, S., \& Gentner, D. (2010). Where Hypotheses Come From: Learning New Relations by Structural Alignment. Journal of Cognition and Development, 11(3), 356-373.
Gentner, D, \& Colhoun, J. (2010). Analogical processes in human thinking and learning. Towards a theory of thinking.
Gentner, D, \& Namy, L. L. (1999). Comparison in the development of categories. Cognitive Development, 14(4), 487-513.
Gentner, D., Anggoro, F. K., \& Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. Child development, 82(4), 1173-88.
Gentner, D., Loewenstein, J., \& Hung, B. (2007). Comparison facilitates children's learning of names for parts. Journal of Cognition and Development, 8(3), 285307.

Goldstone, R., Day, S., \& Son, J. (2010). Comparison. Towards a Theory of Thinking.
Graham, S., Namy, L., Gentner, D., \& Meagher, K. (2010). The role of comparison in preschoolers' novel object categorization. Journal of Experimental Child, 107, 280290.

Hammer, R., Diesendruck, G., Weinshall, D., \& Hochstein, S. (2009). The development of category learning strategies: What makes the difference? Cognition, 112(1), 105-119.
Imai, M., Gentner, D., \& Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. Cognitive Development, 9(1), 45-75.

Klahr, D., \& Chen, Z. (2011). Finding one's place in transfer space. Child Development Perspectives, 5(3), 196-204.
Landau, B., Smith, L. B., \& Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive development, 3(3), 299-321.
Landau, B., Smith, L., \& Jones, S. (1998). Object shape, object function, and object name. Journal of Memory and Language, 38(1), 1-27.
Markman, E. M. (1989). Categorization and naming in children: Problems of induction. Mit Press.
Mervis, C. B., \& Rosch, E. (1981). Categorization of natural objects. Annual review of psychology, 32(1), 89-115.
Murphy, G. (2002). The big book of concepts. Cambridge, Mass: MIT press.
Richland, L. E., \& Burchinal, M. R. (2013). Early executive function predicts reasoning development. Psychological science, 24(1), 87-92.
Richland, L. E., Morrison, R. G., \& Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. Journal of experimental child psychology, 94(3), 249-273.
Smith, L. B., Jones, S. S., \& Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. Developmental Psychology, 28, 273-286.
Son, J. Y., Smith, L. B., \& Goldstone, R. L. (2008). Simplicity and generalization: Short-cutting abstraction in children's object categorizations. Cognition, 108(3), 626638.

Thibaut, J.P., French, R.M., \& Vezneva, M. (2010a). Cognitive load and semantic analogies: Searching semantic space. Psychonomic bulletin \& review, 17(4), 569-574.
Thibaut, J.P, French, R.M., \& Vezneva, M. (2010b). The development of analogy making in children: Cognitive load and executive functions. Journal of experimental child psychology, 106(1), 1-19.
Thibaut, J. P., \& Witt, A. (2015). Young children's learning of relational categories: multiple comparisons and their cognitive constraints. Frontiers in psychology, 6.
Waxman, S. R., \& Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. Developmental Psychology, 36(5), 571-581.

# Solving the Puzzle to Reach the Summit: Using Metaphor to Gauge Public Perceptions of Science 

Paul H. Thibodeau (pthibode@oberlin.edu)<br>Department of Psychology, Oberlin College<br>120 W. Lorain St., Oberlin, OH 44074, USA

Stephen J. Flusberg (stephen.flusberg@purchase.edu)<br>Department of Psychology, SUNY Purchase College<br>735 Anderson Hill Road, Purchase, NY 10577, USA

Kevin J. Holmes (kevin.holmes@coloradocollege.edu)<br>Department of Psychology, Colorado College<br>14 E. Cache La Poudre St., Colorado Springs, CO 80903, USA


#### Abstract

Skepticism towards science has risen sharply in recent years. Cognitive scientists can help address this issue by illuminating how people conceptualize the scientific process, paving the way for improved communication with the public. We recruited a large sample of lay Americans, as well as academics in the sciences and humanities, to answer a series of questions assessing their views about science. Because metaphors have been identified as useful tools for communicating about complex domains, we asked participants to choose which of two metaphors-working on a puzzle or scaling a mountain-best captured their beliefs about the scientific process. Results revealed substantial variation in perceptions of science across groups, and we highlight the ways in which scientists seem to conceptualize science differently from non-scientists. Importantly, metaphor preference was associated with particular patterns of thinking, though not always in our originally hypothesized direction. We discuss the implications of these findings.


Keywords: metaphor, science, concepts, public perceptions

## Introduction

Scientific research requires a variety of skills and involves a range of tasks and experiences. It can be like piecing together a puzzle, in which a diverse set of empirical findings are connected to fill in the details of big-picture scientific theories. And it can be like climbing a mountain, in which careful planning and steadfast persistence are necessary to move projects forward. Both of these metaphors-working on a puzzle and scaling a mountaincapture aspects of the scientific process. In this paper, we use these two metaphors to gauge how people think about science. We recruited a large sample of the general public, as well as academics in the sciences and humanities, to answer a series of questions assessing their views about science, scientific practices, and the priorities of working scientists. They also selected which of the two metaphors (puzzle or mountain) better represented their beliefs about science, and we explored associations between this choice and broader patterns of thinking about scientific practice.

Understanding how people conceptualize science is important given the widespread anxiety about everything from climate change to vaccines to genetically modified
organisms (Achenbach, 2015). Indeed, recent populist political movements have been accompanied by an increasing distrust in science and data. In a recent US poll, almost half of participants (and over two-thirds of President Trump's base) said they did not trust the economic data being reported by government agencies (Ryssdal, 2016). There is also concern in the scientific community about the quality of science education and the lack of public investment in science: "An overwhelming majority of scientists [over 75\%] see the public's limited scientific knowledge as a problem for science," according to another recent poll (Funk \& Rainie, 2015).

The lack of support for science represents a direct threat to addressing important real-world situations like climate change and, potentially, the stability of societal institutions at large (Otto, 2016). As one recent Washington Post article concluded, "This is how a democracy crumbles: not with a bang, but with data trutherism" (Rampell, 2016).

Recent work in the cognitive sciences has explored how to improve communication with the public for specific scientific issues like climate change and public health (Flusberg, Matlock, \& Thibodeau, 2017; Thibodeau, Perko, \& Flusberg, 2015). Metaphors have been identified as useful communication and explanatory tools, as they can help people make sense of complex issues by relating them to more familiar domains, leveraging the schematic knowledge people already have in order to reason about new and complicated subjects (Thibodeau, Crow, \& Flusberg, 2016). To date, however, little research has investigated the role of metaphor in thinking about the scientific process itself (but see Thibodeau, 2016, and Harwood, Reiff, \& Phillipson, 2005).

In addition to addressing practical concerns about public perceptions of science, therefore, the present work also has theoretical implications, as cognitive scientists (particularly cognitive linguists) often treat metaphor as a window into how people think (e.g., Fairclough, 2013; Lakoff \& Johnson, 1980). That is, puzzle and mountain metaphors for science seem to have different entailments, which may suggest that people who talk about science as a puzzle think about the scientific process differently than people who talk about science as a mountain. On the other hand, some have
questioned this approach because of the assumptions that are made about the nature of thinking simply from observing patterns of language use (Keysar \& Bly, 1995; McGlone, 2011; Murphy, 1996, 1997).

One of the few studies to address this issue in the context of reasoning about science involved structured interviews with scientists aimed at identifying (a) key characteristics of scientific inquiry and (b) metaphors that scientists use to conceptualize these issues (Harwood et al., 2005). The goal was to improve science education by encouraging science teachers to use metaphors in the classroom more deliberately. The results indicated that scientists' descriptions of the scientific process emphasized five key characteristics: open-mindedness, putting yourself in your work, utilizing resources, problem solving, and making connections. These characteristics were then matched to conceptual metaphors that the scientists used in the interviews. For example, a puzzle metaphor was often used to emphasize how scientists seek to make connections; an artist metaphor was used to stress the importance of being open-minded; a gardening metaphor was used to talk about immersing oneself in their work.

One possibility is that these metaphors can encourage non-scientists to think about the scientific process in a way that is more consistent with how scientists think about scientific inquiry (Harwood et al., 2005). However, an alternative possibility is that the meaning of these metaphors will be different for scientists and non-scientists. That is, the knowledge and experience that people have with science may influence how they interpret metaphors for the scientific process. A recent investigation into how different groups of people understand militaristic metaphors in biology like invasive species suggests support for the latter possibility (Larson, Nerlich, \& Wallis, 2005): a metaphor that means one thing to scientists can mean something else to non-scientists.

Therefore, the present study represents an important empirical step in comparing how different groups of people interpret metaphors for the scientific process. Do people actually think about science in a way that is consistent with the metaphor they would use to talk about it?

## Experiment

## Methods

Participants A sample of 518 people representing the general public was recruited from Amazon's Mechanical Turk $\left(60 \%\right.$ female; $\left.M_{\text {age }}=35\right)$. A second sample of academics was recruited from the faculty listings of college and university websites in the United States, drawn from a list of top research and liberal arts institutions. We created a list of 2,000 academics, roughly half representing the sciences (i.e. faculty working in Physics, Chemistry, or Biology Departments), and half representing the humanities (i.e. faculty working in English, History, or Philosophy Departments). An email asking for voluntary participation in the survey, yielded responses from 156 academics ( 93 from the sciences and 63 from the humanities).

Although we were primarily interested in comparing how scientists and the general public think about science, we were also interested in understanding why these groups may think differently. Including the group of academics from the humanities helps to address this question. Like scientists, this sample is highly educated, familiar with working on projects that can take long periods of time, and conduct their work in a college or university setting. On the other hand, like participants from the general public, this sample may not be as familiar with the day-to-day experience of conducting scientific work. As a result, including the humanists allows us to investigate why the general public might hold views about science that are different from scientists. For instance, are differences related to more general factors like education level or related to factors more directly tied to being immersed in scientific work?

Materials \& Design All participants were asked to choose between two metaphors for science. The instructions for this judgment read, "We are interested in how people think about science. Which of the following metaphors best captures how you view the process of working on a scientific project?" The order of the two options-Working on a puzzle or Scaling a mountain-was counterbalanced. ${ }^{1}$

These two metaphors for science were chosen because of their use in prior work (Thibodeau, 2016), and because they are commonly used by scientists to talk about the scientific process (Harwood et al., 2005). Thibodeau (2016), for example, found that metaphorically framing a scientist's work as a puzzle led people to value "testing completely novel theories" over "using methods that are simple for others to follow," whereas framing the scientist's work as a climbing a mountain led people to value using simple methods over testing novel theories.

All participants were also asked to rank six aspects of the scientific process in order of importance, three of which were designed to be more consistent with the entailments of the puzzle metaphor and three of which were designed to be more consistent with the entailments of the mountain climbing metaphor (see Table 1). The relationship between the entailments and metaphors was based on prior work (Harwood et al., 2005; Thibodeau, 2016), and experimenter intuition; one goal of the study is to test whether different groups of people have similar intuitions about the relationship between the entailments and metaphors. The order of the statements was randomized across participants. ${ }^{2}$

[^515]We compare how the three populations (General Public; Scientists; Humanists) rank the statements overall, and we test whether people who prefer the puzzle metaphor for science rank the puzzle-congruent statements as more important than the mountain-congruent statements (and vice versa for the mountain metaphor).

Table 1. Tasks related to the scientific process that were ranked by participants.

| Actions and Behaviors Related to Science | Metaphor |
| :--- | :--- |
| 1. Find creative ways to study important research questions | Puzzle |
| 2. Seek insight from diverse sources | Puzzle |
| 3. Find connections between seemingly unrelated ideas | Puzzle |
| 4. Make a detailed research plan | Mountain |
| 5. Persist in the face of setbacks | Mountain |
| 6. Revise theories in light of new data or counter-evidence | Mountain |

Finally, participants were asked to complete the 40 -item Scientific Attitude Inventory II (SAI II), which is designed to measure attitudes related to science along six dimensions (Moore \& Foy, 1997; see Table 2). This instrument has been used to measure perceptions of science among students and the general public, and to predict who is likely to pursue a career in a STEM field (e.g., Bathgate, Schunn, \& Correnti, 2014; Moore \& Foy, 1997).

As with the rank order task, we use responses from the SAI II to compare how the three populations think about science and to test whether certain dimensions of the scale map onto the view that science is like working on puzzle versus scaling a mountain.

Table 2. Six dimensions of the SAI II and example items..

| Dimension | Example Item |
| :--- | :--- |
| Theory | Scientific ideas can be changed. |
| Limited | Scientists cannot always find the answers to their questions. |
| Empirical | Scientific questions are answered by observing things. |
| Goal | Ideas are the important result of science. |
| Public | Every citizen should understand science. |
| Interest | I would enjoy studying science. |

At the end of the survey, participants from the general public were asked background and demographic questions, including their gender, age, education level, math/science training, political ideology (0, Very liberal, to 100, Very conservative), and personality (the Big Five personality dimensions; Gosling, Rentfrow, \& Swann, 2003). Academics were asked to identify as working in the humanities or sciences.

## Results

Metaphor Preference Among participants sampled from the general public, $89 \%$ preferred the puzzle metaphor for science, $\chi^{2}(1)=315.09, p<.001$. The puzzle metaphor was also preferred by $89 \%$ of academics, $\chi^{2}(1)=92.31, p<.001$, although scientists (84\%) were marginally more likely to choose the mountain metaphor compared to humanists (95\%), $\chi^{2}(1)=3.71, p=.054$.

For the sample recruited from the general public, we tested whether any of the individual difference measures (i.e. gender, age, education level, math/science training,
political ideology, personality) predicted participants' choice of metaphor. The only reliable predictor of participants' choice was age: older participants were especially likely to endorse the puzzle metaphor for science, $B=.40, S E=.17, p=.016$.

Ranking Priorities We first compare how the three populations ranked the six statements about science, focusing on contrasts between scientists and (a) humanists and (b) participants from the general public. Then we test for a relationship between participants' preferred metaphor and their rankings of the statements. For this second analysis, we excluded the humanists because of the small number who preferred the mountain metaphor $(n=3)$.

First, a mixed-effects linear model was fit to the rankings with statement (1-6) treated as a within-subjects effect and sample (public, humanists, scientists) treated as a betweensubjects effect (Bates, Maechler, Bolker, \& Walker, 2014). The model revealed that the samples gave different rankings to the statements, $\chi^{2}(5)=209.91, p<.001$.


Figure 1. Mean ranking of six statements about science by sample: of the general public, academics working in the humanities, and academics working in the sciences.

Figure 1 illustrates how participants from the three groups ranked the statements overall. One pattern to note is that scientists tended to show more agreement on how the statements were ranked (Kendall's $W=.25$ ) than participants from the general public ( $W=.07$ ) or humanists ( $W=.21$ ). For instance, "planning" was ranked as the most important aspect of science by the general public, with $53 \%$ of participants from this sample ranking it first or second. In contrast, "creativity" was ranked as the most important aspect of science by scientists, with $76 \%$ of participants from this sample ranking it first or second. On the other end of the spectrum, "finding connections" was ranked as the least important aspect of science by the general public ( $46 \%$ ranked it fifth or sixth), while "planning" was ranked as the least important aspect of science by scientists ( $66 \%$ ranked it fifth or sixth). This suggests that scientists, as a group, have a more consistent conception of the scientific process than the general public; the humanists showed an intermediate level of consistency.


Figure 2. Comparing rankings of scientists to those of people from the general public and humanists with a measure of effect size (Cohen's $d$ ). Bars that extend to the right indicate that scientists ranked the statement as more important (purple: compared to participants from the general public; orange: compared to humanists); bars that extend to the left indicate that scientists ranked the statement as less important. Stars indicate a statistically significant difference between scientists and the comparison group.

Figure 2 shows how scientists ranked the statements compared to the other two groups of participants by plotting a measure of effect size (Cohen's $d$ ). As shown, scientists tended to place less emphasis on planning than humanists, $t(154)=3.69, p<.001$, or participants from the general public, $t(609)=10.21, p<.001$. On the other hand, scientists tended to place more emphasis on persistence than humanists, $t(154)=2.82, p=.005$, or participants from the general public, $t(609)=2.95, p=.003$. The two groups of academics ranked the other four statements similarly. Scientists and the sample from the general public ranked three of the remaining four statements differently: scientists placed less emphasis on seeking diverse sources of insight, $t(609)=2.49, p=.013$, but more emphasis on finding connections between seemingly unrelated ideas, $t(609)=$ 2.21, $p=.027$, and creativity, $t(609)=6.72, p<.001$; these two groups placed similar emphasis on revising theories in light of new data.

A second analysis tested whether people who considered science to be more like a puzzle ranked puzzle-congruent statements as more important than people who considered science to be more like mountain climbing. A mixed-effects linear model was fit to the data with predictors for statement type (puzzle- or mountain-congruent), metaphor choice (puzzle or mountain), and sample (scientists versus general public), which revealed an interaction between the statement type and metaphor choice, $\chi^{2}(1)=4.07, p=.044$. Contrary
to our initial hypothesis, preference for the puzzle metaphor was associated with prioritizing the mountain-congruent statements (and vice versa). This pattern was consistent across both groups of participants (i.e. there was no 3 -way interaction between statement type, metaphor choice, and sample), $\chi^{2}(1)=1.39, p=.239$ (see Figure 3). Of note, the analysis also revealed an interaction between sample and statement type, $\chi^{2}(1)=38.44, p<.001$, such that scientists tended to rank the puzzle-congruent statements as more important, regardless of preferred metaphor, than participants from the general public.


Figure 3. Mean ranking of statements by type (puzzle- or mountain-congruent) and metaphor chosen (puzzle or mountain) for the two samples analyzed (from the general public and of scientists).

People who preferred the mountain metaphor were particularly likely to rank "finding connections" ( $M=3.69$, $S D=1.52$ ) as more important than people who preferred the puzzle metaphor $(M=4.15, S D=1.53)$, whereas people who preferred the puzzle metaphor were particularly likely to rank "planning" ( $M=3.09, S D=1.91$ ) as more important than people who preferred the mountain metaphor ( $M=$ $3.44, S D=2.00$ ).

This finding is consistent with the view that the two metaphors capture structured ways of thinking about the scientific process that are different from one another. People who reported thinking science was a puzzle ranked the statements in a systematically different way than people who reported thinking science was a mountain. However, the finding is inconsistent with how we had mapped the entailments of the metaphors onto the statements, suggesting that the intuitions of language researchers may differ from how metaphoric language is used and understood in the real world.

It is also possible that behaviors on the two taskschoosing a metaphor and ranking the statements-were complementary. People may have had a sense of the limitations of their preferred metaphor, which they expressed in the rank-order task (or vice versa). For instance, a participant may believe that finding connections and planning are both vital to the scientific process. Such a belief may lead this participant to choose the puzzle metaphor as more appropriate (because it captures the value of finding connections) and also to rank planning highly (since it is captured less well by the puzzle metaphor). In
other words, participants may consider many aspects of science to be important, not just those that are consistent with the entailments of a single metaphor. This may lead them to express a preference for one metaphor, as required by our forced-choice task, and then to emphasize inconsistent entailments in the rank-order task.

Scientific Attitude Inventory A similar set of analyses was applied to data from the SAI II. First, we found differences in the extent to which participants endorsed the six dimensions measured by the survey, $\chi^{2}(5)=1637.2, p<$ .001: participants agreed most strongly with statements about the necessity of adopting an empirical mindset, followed by statements about the importance of public outreach, about an interest in doing scientific work, that the scope of science is limited to observable phenomena, and finally, that the end goal of science is ideas (rather than a tangible product like technology).


Figure 4. Comparing ratings of scientists to those of people from the general public and humanists with a measure of effect size (Cohen's $d$ ). Bars that extend to the right indicate that scientists rated the dimension as more important (purple: compared to participants from the general public; orange: compared to humanists); bars that extend to the left indicate that scientists rated the dimension as less important. Stars indicate a statistically significant difference between scientists and the comparison group.

Second, the analysis revealed differences in how the three samples rated the statements, $\chi^{2}(2)=196.57, p<.001$. Overall, scientists tended to endorse the statements more strongly than participants from the general public, $B=.45$, $S E=.03, p<.001$, and humanists, $B=.17, S E=.04, p<$ .001. Humanists endorsed the statements more strongly than participants from the general public, $B=.28, S E=.04, p<$ . 001 .

Third, the analysis revealed an interaction between ratings of the dimensions and the three samples, $\chi^{2}(10)=199.57, p$ $<.001$. As shown in Figure 4, scientists endorsed all six dimensions more strongly than participants from the general public, $p \mathrm{~s}<.001$. Compared to humanists, scientists more strongly endorsed having an empirical mindset, the view that public support is important, and were more likely to say they enjoyed doing scientific work, $p \mathrm{~s}<.01$, whereas humanists were more likely to view science as being limited to the study of natural phenomena, $p<.001$. The two groups of academics expressed similar views about scientific theorizing and about the end-goal of scientific work.


Figure 5. Differences in ratings by metaphor chosen for people from the general public and scientists, illustrated by a measure of effect size (Cohen's $d$ ). Bars that extend to the right indicate higher ratings among people who preferred the puzzle metaphor; bars that extend to the left indicate higher ratings among people who preferred the mountain metaphor.

Finally, we tested for a relationship between the metaphor participants preferred and ratings of the dimensions (excluding data from humanists). As illustrated in Figure 5, among scientists, preference for the mountain metaphor was associated more strongly with the view that scientific study is limited to natural phenomena, $t(91)=2.72, p=.009$; among the general public, preference for the puzzle metaphor was associated with a more empirical mindset, $t(516)=2.35, p=.019$, the view that public support is important, $t(516)=2.53, p=.012$, and a stronger interest in doing scientific work, $t(516)=3.46, p<.001$. These results provide further evidence that the two metaphors capture different ways of thinking about the scientific process, but also suggest that what exactly is captured by the metaphors is different for scientists and non-scientists.

## Discussion

Skepticism towards scientific research can be found among the general public as well as politicians on both sides of the political aisle, raising significant concerns about the current quality of science education and communication. As an initial step towards addressing this critical issue, we aimed to illuminate how people think about the scientific process itself, contrasting the beliefs of scientists with those of academics in the humanities and members of the general public, and exploring the role of metaphor in representing broad conceptual viewpoints.

We found several notable similarities and differences in how scientists and non-scientists conceptualized the scientific process. Of particular interest, scientists tended to prioritize persistence more than the two samples of nonscientists, who tended to prioritize planning more than the scientists. This suggests that being immersed in scientific work makes salient the determination needed to complete research projects. Simply hearing about scientific findings in the classroom or the news, on the other hand, may make it seem like scientists spend most of their time planning. In line with this distinction, scientists were more likely than non-scientists to think science was like scaling a mountain, although all three groups showed an overall preference for the puzzle metaphor.

Individuals who preferred the puzzle metaphor tended, counter-intuitively, to value statements about science that were designed to be congruent with the mountain metaphor (and vice versa). Preference for the puzzle metaphor was also associated with a more empirical mindset, the view that public support is important for scientific progress, and an interest in doing scientific work-but only among the general public, not among scientists. These findings imply that metaphors for science will be interpreted differently depending on one's scientific knowledge and expertise. The findings also highlight the importance of identifying the systems of knowledge associated with metaphor use rather than merely assuming them (Keysar \& Bly, 1995; McGlone, 2011; Murphy, 1996, 1997).

Future research in this area should explore additional metaphors for scientific inquiry. For instance, one scientist in the study suggested that science was more like mapmaking or exploring than it was like working on a puzzle or scaling a mountain. Future work should also investigate whether these metaphors can causally influence how people think about the scientific process.

## References

Achenbach, J. (2015). Why do many reasonable people doubt science? National Geographic, 14 .
Bates, D., Maechler, M., Bolker, B., \& Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4.
Bathgate, M. E., Schunn, C. D., \& Correnti, R. (2014). Children's motivation toward science across contexts, manner of interaction, and topic. Science Education, 98, 189-215.

Fairclough, N. (2013). Critical discourse analysis: The critical study of language. New York: Routledge.
Flusberg, S. J., Matlock, T., \& Thibodeau, P. H. (2017). Metaphors for the war (or race) against climate change. Environmental Communication. doi:10.1080/17524032.2017.1289111
Funk, C., \& Rainie, L. (2015). Public and scientists’ views on science and society. Pew Research Center. Retrieved January 23, 2017: http://www.pewinternet.org/2015/01/ 29/public-and-scientists-views-on-science-and-society/
Gosling, S. D., Rentfrow, P. J., \& Swann Jr., W. B. (2003). A very brief measure of the Big-5 personality domains. Journal of Research in Personality, 37, 504-528.
Harwood, W. S., Reiff, R. R., \& Phillipson, T. (2005). Putting the puzzle together: Scientists metaphors for scientific inquiry. Science Educator, 14, 25-30.
Keysar, B., \& Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? Journal of Memory and Language, 34, 89-109.
Larson, B. M., Nerlich, B., \& Wallis, P. (2005). Metaphors and biorisks: The war on infectious diseases and invasive species. Science Communication, 26, 243-268.
McGlone, M. S. (2011). Hyperbole, homunculi, and hindsight bias: An alternative evaluation of conceptual metaphor theory. Discourse Processes, 48, 563-574.
Moore, R. W., \& Foy, R. L. H. (1997). The scientific attitude inventory: A revision (SAI II). Journal of Research in Science Teaching, 34, 327-336.
Murphy, G. L. (1996). On metaphoric representation. Cognition, 60, 173-204.
Murphy, G. L. (1997). Reasons to doubt the present evidence for metaphoric representation. Cognition, 62, 99-108.
Otto, S. L. (2016). The war on science: Who's waging it, why it matters, what we can do about it. Minneapolis: Milkweed Editions.
Rampell, C. (2016). When facts don't matter, how can democracy survive? The Washington Post. Retrieved January 23, 2017: https://www.washingtonpost.com /opinions/when-the-facts-dont-matter-how-can-demo cracy-survive/2016/10/17/560ff302-94a5-11e6-9b7c57290af48a49_story.html?utm_term=.35e1a60b4a74
Ryssdal, K. (2016). Poll finds Americans' economic anxiety reaches new high. Marketplace. Retrieved January 23, 2017: http://www.marketplace.org/2016/10/13/econom y/americans-economic-anxiety-has-reached-new-high
Thibodeau, P. H. (2016). Extended metaphors are the home runs of persuasion: Don't fumble the phrase. Metaphor and Symbol, 31, 53-72.
Thibodeau, P. H., Crow, L., \& Flusberg, S. J. (2016). The metaphor police: A case study of the role of metaphor in explanation. Psychonomic Bulletin \& Review. doi:10.3758/s13423-016-1192-5
Thibodeau, P. H., Perko, V. L., \& Flusberg, S. J. (2015). The relationship between narrative classification of obesity and support for public policy interventions. Social Science \& Medicine, 141, 27-35.

# Metaphors, Roles, and Controls in Framing Studies 

Paul H. Thibodeau (pthibode@oberlin.edu)<br>Oberlin College, Department of Psychology<br>120 W. Lorain St., Oberlin, OH 44074, USA

Stephen J. Flusberg (stephen.flusberg@purchase.edu)
SUNY Purchase College, Department of Psychology
735 Anderson Hill Road, Purchase, NY 10577, USA


#### Abstract

Metaphors have been shown to be effective explanatory and communicative tools, shaping how people think and reason about complex domains. To date, however, most studies have addressed only coarse-grained effects of metaphor framing, leaving many questions unanswered about the relative power of metaphor compared to more literal linguistic framing devices. We addressed this issue in a large, pre-registered framing study, comparing the effects of describing the role of police officers as (a) metaphorical guardians of a community (b) literal protectors of a community, and (c) a no-label control. We found no main effect of framing condition, suggesting that positively valenced metaphors may exert little influence on their own in this domain. However, we did observe an interaction between condition and political ideology, such that the guardian metaphor was especially effective at improving attitudes towards police officers for liberals, whose initial approval ratings were relatively low.


Keywords: metaphor, framing, attitudes, policing

## Introduction

"Evolutionary sequences," wrote the popular biologist Steven Jay Gould (1977, p. 61), "are not rungs on a ladder, but our retrospective reconstruction of a circuitous path running like a labyrinth, branch to branch, from the base of the bush to a lineage now surviving at its top." Metaphorical explanations like this are common, and research has established that they can be effective as well: framing a discussion or explanation with metaphor has been shown to shape how people understand and reason about a range of complex issues (Flubserg, Matlock, \& Thibodeau, 2017; Sopory \& Dillard, 2002; Thibodeau, 2016; Thibodeau \& Boroditsky, 2011; Thibodeau, Crow, \& Flusberg, 2016).

In a recent study, for example, Thibodeau, Crow, and Flusberg (2016) sought to test the explanatory power of metaphor in the context of people's understanding of-and attitudes towards-law enforcement. Our primary research question was whether or not people would spontaneously use the structure of a metaphorical source domain (guardian or warrior) to reason about a target domain (policing).

One way that we tested this question was by having participants read that police officers are either guardians or warriors of the community before reporting on their attitudes toward policing and the criminal justice system. We found that participants who had read that police officers are guardians expressed more positive attitudes about policing and the criminal justice system, overall, compared to people who had read that police officers are warriors.

This effect may be the result of the emotional tone that is set by the metaphors: we found that guardian, in the context of policing, conveys a more positive emotional valence than warrior.

In addition, we found that the metaphorical explanations selectively affected certain attitudes toward policing and the criminal justice system more than others. Specifically, people who read that police officers are guardians expressed a more favorable "attitude toward police practices" than people who read that police officers are warriors, but the metaphorical explanation had no effect on participants' views about the "difficulty of being a police officer." This was consistent with the results of an initial norming study, where a separate group of participants made an explicit judgment about which metaphor-guardian or warrior-was more appropriate for the current state of policing (rather than being exposed to just one of the metaphors). That is, the norming study found that people who came into the study with the view that police officers are more like guardians expressed more a favorable view of police practices compared to people who considered police officers to be warriors. However, people who considered police officers to be guardians expressed similar beliefs about the difficulty of being a police officer as people who considered police officers to be warriors. Taken together, these findings suggest that the metaphors instantiate different schematic knowledge structures for policing and the criminal justice system-and that they capture and convey more than an emotional tone.

A second way that we tested our research question was by having participants list a synonym either to "guardian" or "warrior" before reporting on their attitudes toward the police and criminal justice system. We found no effect of these lexical primes: people who listed a synonym for "guardian" expressed similar attitudes toward policing and the criminal justice system as people who listed a synonym for "warrior." Participants expressed more moderate attitudes in these conditions compared to the conditions in which a metaphor was used to explain the role of police officers (i.e. less positive than participants who read that police officers are guardians but more positive than participants who read that police officers are warriors).

We interpreted these results as showing (a) that metaphorically framing police officers as guardians activates a different mental model of policing (with a different affective profile) than metaphorically framing police officers as warriors, and (b) that simply seeing the
word "guardian" or "warrior" is insufficient to activate this mental model. In other words, people have prior knowledge about what it means to be a guardian (and warrior). This knowledge influences how people think about policing when police officers are explicitly described as guardians (or warriors)-but not when people are asked questions about policing after simply seeing the word "guardian" (or "warrior").

A natural follow up question to this study might be: which metaphor has a bigger effect on how people think about policing? Intuitively, it may seem like there is an easy way to address this question: by running a condition that does not include a metaphorical explanation. One might expect that describing police officers as guardians would lead to a more positive view of policing compared to a "neutral control" condition, and that describing police officers as warriors would lead to a more negative view of policing compared to a "neutral control" condition. In addition, one might be tempted to infer that the metaphor condition that is more different from the "neutral control" condition is having a bigger effect on people (cf. Reijnierse, Burgers, \& Steen, 2015; Steen, Reijnierse, \& Burgers, 2014).

However, as we have argued before (see Thibodeau, In press; Thibodeau \& Boroditsky, 2015), there are many differences between metaphor frames and "neutral control" conditions that make such comparisons difficult to interpret. For example, it is not clear that there is a suitable nonmetaphorical counterpart to the "guardian" and "warrior" metaphors for policing (e.g., a term like "protector" could be used, since it is less metaphorical than the two metaphorical frames; but its meaning seems more similar to "guardian" than "warrior," making it a poor candidate to serve a "neutral control" condition along side the two metaphorical conditions). Comparing the two metaphorical frames to a condition that omitted a nominal descriptor altogether would confound a variety of factors between the two metaphor conditions and the "neutral control" condition, including the valence, tone, and word frequency of the language used to describe policing-not just the metaphoricity of the conditions.

We do, however, think that there are research questions that warrant a comparison between metaphorical frames and non-metaphorical counterparts. Here, we consider such a case. Namely, does a guardian metaphor lead people to adopt a more favorable view of policing than a comparable literal description of the role police officers play in the community? Addressing this question is important when considering the potential practical applications of research on the persuasive power of metaphor. Therefore, we compared three conditions in the present study, building on the work described in Thibodeau, Crow, \& Flusberg (2016). Before reporting their attitudes towards law enforcement and the criminal justice system, participants read one of the following framing prompts:
a. Police officers are the guardians of modern communities. They are strong men and women who serve a vital role in society.
b. Police officers are the protectors of modern communities. They are strong men and women who serve a vital role in society.
c. [No label control: participants simply answered the targeted questions about police officers in this condition]

We chose "protector" as a non-metaphorical counterpart to guardian because it was the word most frequently used to explain what it means for police officers to be guardians in the original study (Thibodeau, Crow, \& Flusberg, in press). In the context of this more applied question, we did not include a condition that described police officers as warriors, since such a description would be expected to elicit comparatively negative views of police officers. Instead, we compared the effects of a metaphor frame (and a matched literal frame) to a "neutral control" condition because we were interested in whether and to what extent describing police officers as guardians leads people to express a more favorable view of policing.

We were also interested in a mechanistic question about the role of metaphor in explanatory discourse, which we addressed by comparing the guardian and protector conditions. Are metaphors more persuasive than literal counterparts? In a meta-analysis, Sopory and Dillard (2002) found that metaphors are about $6 \%$ more persuasive than literal language, which they attributed to the power of metaphors to organize the way people think about a target domain.

The design of the current study provides a novel context for testing this claim. One possibility is that the guardian metaphor may call to mind a more coherent and favorable mental model of policing than the non-metaphorical counterpart, protector, and lead people to the most positive view of police officers and the criminal justice systemmore positive than the protector and "neutral control" conditions.

An alternative possibility, though, is that the literal counterpart to guardian (protector) serves a similar organizational function in describing the role of police officers in the community. That is, depending on the complexity of the target domain and intended meaning of the metaphor, there may be issues for which a literal frame is as effective as a metaphorical one in shaping thought. Support for this possibility would be found if people express similarly positive views of the police in these two conditions-both of which should lead people to a more positive view of policing than the "neutral control" condition. Such a finding would contribute to the literature by identifying an important boundary condition on metaphor framing effects (cf. Steen, Reijnierse, \& Burgers, 2014).

In addition to the framing experiment, we also conducted a norming study to assess the perceived metaphoricity and emotional valence of three possible descriptions of police officers: (a) guardians of modern communities, (b) protectors of modern communities, and (c) warriors of modern communities. One critical assumption that was
made in our prior work was that the terms guardian and warrior were actually interpreted as metaphors, and not, for example, as literal descriptions of the role of police officers ${ }^{1}$. The norming study allows us to test this assumption. We expected that the guardian and warrior descriptions would be rated as more metaphorical than protector. The norming study also allowed us to quantify the emotional tone of the three descriptions. We expected that the guardian and protector descriptions would be rated as conveying a more positive emotional valence than warrior. Both studies were pre-registered on the Open Science Framework: osf.io/eb853.

## Norming Study

## Methods

Participants We recruited 100 participants for the norming study from Amazon's Mechanical Turk. After excluding participants who failed to finish the study or provide a valid completion code, we were left with data from 88 participants for analysis ( $51 \%$ male; $M_{\text {age }}=33$ ).

Materials and Design Participants were asked to rate the metaphoricity (1, Not at all metaphorical, to 5, Very metaphorical) and valence (1, Very negative, to 5, Very positive) of three statements on 5-point scales (Thibodeau \& Boroditsky, 2015).
a. Police officers are guardians.
b. Police officers are warriors.
c. Police officers are protectors.

These statements were presented on the same screen; the order of the statements was randomized across participants.

Afterward, participants were asked background and demographic questions, including their gender, age, education level, political ideology ( 0 , Very liberal, to 100 , Very conservative), and political affiliation (Democrat, Independent, Republican, Other). They also completed the attitudes towards policing measure described in the experiment below, although we did not analyze responses to these questions for participants in the norming study.

## Results and Discussion

A repeated measures ANOVA revealed differences in the rated metaphoricity of the three statements, $F(2,174)=$ 27.59, $p<.001, \eta^{2}=.24$. Warrior $(M=3.56, S D=1.33)$ was rated as more metaphorical than guardian $(M=3.02$, $S D=1.21), t(87)=3.48, p<.001$, or protector $(M=2.38$, $S D=1.28), t(87)=6.35, p<.001$; guardian was rated as more metaphorical than protector, $t(87)=4.82, p<.001$ (see Figure 1).

[^516]On the one hand, this pattern of results confirms our intuition that the terms warrior and guardian are perceived as more metaphorical than the term protector in the context of describing the role of police officers. On the other hand, we did not predict that the term warrior would be viewed as more metaphorical than guardian.

One possibility is that people consider the guardian metaphor to be more apt (cf. Glucksberg, 2001), which affects judgments of metaphoricity (Thibodeau, Sikos, \& Durgin, 2015). In our original study, $82 \%$ of participants thought police officers should strive to be guardians (rather than warriors) of their communities (Thibodeau, Crow, \& Flusberg, 2016). Talking about police officers in a way that is inconsistent with a preferred mental model of policing (i.e. in a way that is less apt) may lead people to think the description is more metaphorical. In any case, the critical difference in metaphoricity for the present study pertains to the contrast between guardian and protector: as expected, people interpreted guardian to be more metaphorical than protector.

A second repeated measures ANOVA revealed differences in the rated valence of the three statements, $F(2$, $174)=51.01, p<.001, \eta^{2}=.37$. Warrior $(M=3.38, S D=$ 1.21) was rated as more negative than both guardian ( $M=$ 4.32, $S D=0.80), t(87)=7.22, p<.001$, and protector $(M=$ 4.49, $S D=0.82$ ), $t(87)=8.04, p<.001$; protector was rated as more positively valenced than guardian, $t(87)=2.19, p=$ .031 (see Figure 1).

Again, this pattern largely conforms to our predictions: guardian and protector both express a positive view of policing compared to warrior. Although protector was judged to be more positive than guardian, this difference was fairly small.


Figure 1. Ratings of the metaphoricity and valence of three descriptions of police officers. Error bars denote standard errors of the means.

## Experiment

In the experiment, we tested whether describing police officers as guardians, compared to protectors (and to a condition that lacked a label for police officers), leads to
more positive attitudes toward policing. That is, previous work has suggested that metaphorical language is more persuasive than literal language, owing to the organizational role that metaphors play in discussions of complex issues (Sopory \& Dillard, 2002). The norming study suggests that guardian and protector differ substantially in the extent to which they are metaphorical, but only slightly in the emotional tone that they convey (in favor of the nonmetaphorical label).

If the guardian label leads people to express to more positive attitudes towards policing than protector, the experiment would provide further evidence of the persuasive value of metaphor (over and above comparable literal language). If people express similar attitudes in the guardian and protector conditions, on the other hand, it would suggest that, in some cases, non-metaphorical language can serve a similar organizational function as metaphorical language.

## Methods

Participants We recruited 600 participants to participate in the experiment on Amazon's Mechanical Turk. After excluding participants who failed to finish the study or provide a valid completion code, we were left with data from 592 participants for analysis ( $49 \%$ male; $M_{\text {age }}=34$ ).

Materials and Design Participants were randomly assigned to one of three conditions. In one, police officers were described as guardians, "Police officers are the guardians of modern communities-strong men and women who serve a vital role in society." In the second, police officers were described as protectors, "Police officers are the protectors of modern communities-strong men and women who serve a vital role in society." A comparable sentence about police officers was omitted from the third condition. In other words, there was no description of police officers in the third condition; this group simply answered the follow-up questions about policing. Participants in all three groups were instructed, "Although most people agree that police are necessary for maintaining law and order, there is disagreement about a variety of issues related to policing. On the following screen, you will be asked several questions about your view of police officers and the criminal justice system. Please answer candidly; your responses are anonymous."

Then participants were asked eight questions about policing and the criminal justice system. Consistent with Thibodeau, Crow, and Flusberg (2016), three of the questions were asked on a 7-point scale: "Police officers have a _ job" (from very easy to very difficult), "Police officers are _ at maintaining law and order" (from very ineffective to very effective), and "How would you describe the criminal justice system in the U.S.?" (from very far from the ideal to very near to the ideal). The other five questions included two response options, asking about whether participants thought police treated citizens equally (yes/no), whether they thought the police were more fair or unfair,
more honest or deceitful, more selfish or selfless, and whether participants felt safe or unsafe around police officers. Responses to all eight of these questions were combined into a single measure of participants' attitudes toward policing, using principal components analysis (see Thibodeau, Crow, \& Flusberg, 2016).

Finally, participants completed the same demographics questions as participants in the Norming Study.

## Results

A between-subjects ANOVA with predictors for condition (guardian, protector, none) revealed no effect of the descriptions on participants' attitudes toward policing, $F(2,589)=0.18, p=.837$. That is, neither the guardian $(M$ $=3.12, S D=1.28)$ nor the protector $(M=3.14, S D=1.45)$ labels for police officers led people to a more positive attitude toward policing compared to a description that lacked a label $(M=3.20, S D=1.44)$. And the two treatment conditions (guardian vs. protector) did not differ from one another.

Given the lack of support for our primary prediction, we considered alternative hypotheses that could be explored in the data. One salient possibility highlights the role of peoples' prior beliefs (e.g., Hardisty, Johnson, \& Weber, 2009; Johnson \& Taylor, 1981; Thibodeau \& Boroditsky, 2011; Thibodeau \& Flusberg, 2017) in combination with a mechanistic claim about how metaphors are processed-by serving as peripheral or heuristic cues, rather than through a process of conscious deliberation and rationalization (cf. Chaiken, Wood, \& Eagly, 1996; Petty \& Cacioppo, 1986).

That is, prior work has found that framing manipulations are, not surprisingly, more impactful on people who have room to be persuaded about an issue (i.e. are not already at ceiling). For instance, Hardisty, Johnson, \& Weber (2009) found that Democrats would support a program designed to decrease the level of carbon dioxide in the environment, regardless of whether it was described as an "offset" program or a "tax." Since Democrats tended to support this type of environmental action, their attitudes were relatively consistent, regardless of how it was framed (i.e. a ceiling effect). Republicans, on the other hand, showed lower support for the program overall, affording more opportunity for attitude change. In turn, Hardisty et al. (2009) found that Republicans were more likely to endorse the program when it was framed as an "offset" than a "tax."

Since political conservativeness tends to be associated with more positive attitudes toward policing (Gerber \& Jackson, 2017), this suggests that the framing manipulation-describing police officers as guardians or protectors-may have a more pronounced effect among politically liberal participants.

One reason to think that the guardian frame will be more persuasive among liberal participants than the protector frame is that metaphors have been argued to exert a persuasive influence through an indirect route. People who are skeptical about the increasing tendency for violence among police officers may perceive the term protector as an
overt attempt to change the way they think about police practices, making them resistant to the persuasive appeal. In contrast, the term guardian may not register as a persuasive message and, thus, bypass this sort of counter-arguing among participants (cf. Chaiken, Wood, \& Eagly, 1996; Petty \& Cacioppo, 1986).

To examine these possibilities, we conducted a second analysis on the data in which political ideology (a continuous variable ranging from 0 , Very liberal to 100 , Very conservative) was included as a covariate. To conduct this analysis, we first tested for an expected positive relationship between political ideology and attitudes toward policing. We found a strong positive relationship, $F(1,590)$ $=79.54, p<.001:$ the more politically conservative the participant, the more positive their view of the police, $B=$ $.34, S E=.04, p<.001$. We then tested for an interaction between political conservativeness and condition (guardian, protector, none), which was significant, $F(2,586)=3.06, p$ $=.048$. Of note, the relationship between political conservativeness and condition did not differ for a contrast between the "neutral control" condition and the protector condition, $F(1,397)=1.53, p=.217$. However, the relationship between political conservativeness and condition did differ when contrasting the "neutral control" condition to the guardian condition, $F(1,391)=6.52, p=$ .011 (see Figure 2).


Political Ideology
Figure 2. Attitudes toward policing by political ideology (left = very liberal; right $=$ very conservative) and frame (none, protector, guardian). Lowess smoothing ( $f=2 / 3$ ) was applied to the lines to facilitate presentation of the trends.

Specifically, among the most liberal participants (i.e., those who reported a score less than 33 on the continuum of political ideology that ranged from 0 , very liberal, to 100 , very conservative; $n=256$ ), attitudes toward policing were more positive in the guardian condition ( $M=2.93, S D=$
1.06) than in the protector condition ( $M=2.54, S D=1.44$ ), $t(160)=1.98, p=.049$; but no different from the condition that lacked a label $(M=2.63, S D=1.52), t(178)=1.54, p=$ .126. There were no differences between conditions for participants whose political ideology was in the middle of the ideological spectrum (i.e. between 33 and $66 ; n=229$ ), $p \mathrm{~s}>.1$, and no differences between conditions for participants whose political ideology was at the conservative end of the spectrum (i.e. $>66 ; n=107$ ), $p \mathrm{~s}>.1$.

## Discussion

In response to mounting tensions between law enforcement and civilians, former president Barack Obama commissioned a task force on $21^{\text {st }}$ century policing, which released its final report in 2015 (Ramsey \& Robinson, 2015). The report suggested that to increase trust between police officers and the communities they serve, "Law enforcement culture should embrace a guardian-rather than a warrior-mindset." In a previous set of experiments, we used this real-world example as a case study to explore the power of explanatory metaphors, demonstrating that describing police offers as guardians did in fact lead people to express more positive attitudes towards law enforcement than describing them as warriors (Thibodeau, Crow, \& Flusberg, 2016). Because of the real-world applications of this line of research, however, there are additional questions that warrant empirical investigation.

In the present study, we took a preliminary step in this direction by asking whether metaphorical framing provides any additional persuasive power over and above a more literal linguistic descriptor. In an initial norming study, we confirmed that describing police officers as guardians of a community was perceived as more metaphorical than describing them as protectors of a community. In our main experiment, we contrasted the effects of framing the role of law enforcement using these two terms, along with a "neutral" control condition that included no framing device.

Our initial analysis revealed that participants in these three conditions did not differ overall with respect to their attitudes towards policing. This could suggest that there is little advantage to using metaphorical framing compared to more literal language (or even to no frame whatsoever) in a practical attempt to improve attitudes towards policing in the United States. This also offers some support for the view that in our original study, it was the more negatively valenced warrior metaphor that was "doing the work," so to speak, in shifting attitudes towards policing (i.e., in a negative direction). This would be consistent with a large body of work in psychology that suggests people are typically more sensitive to negative information (or losses) than positive information (or gains; Baumeister et al., 2001).

However, we also considered an alternative possibility: that individual differences in prior attitudes towards policing (e.g., due to ideological commitments) might interact with our framing manipulation in a principled fashion. Previous research has shown that framing effects are most effective when they target people who are not at
ceiling (or floor) on an issue already (i.e., who have room to be persuaded; Hardisty, Johnson, \& Weber, 2009; Thibodeau \& Boroditsky, 2011; Thibodeau \& Flusberg, 2017). In the present case, we reasoned that because ideologically conservative participants would have come into the study with very positive views of policing already (Gerber \& Jackson, 2017), they might be less persuaded by a positive metaphorical frame compared to more liberal participants. To test this possibility, we included a continuous measure of political ideology as a covariate in an exploratory analysis.

The results of this analysis supported our revised hypothesis: for the most liberal participants, framing police officers as guardians of the community led to more positive attitudes compared to framing them as protectors of the community. This is consistent with previous work demonstrating a principled interaction between metaphor framing and prior beliefs (Hardisty, Johnson, \& Weber, 2009; Thibodeau \& Boroditsky, 2011; Thibodeau \& Flusberg, 2017), and lends support to the view that metaphors may provide an additional persuasive punch compared to more literal language (at least under certain conditions; Sopory \& Dillard, 2002).

Taken together, these findings paint a more nuanced picture of the relationship between, and consequences of, metaphorical versus literal framing, at least in domain of attitudes towards policing. To be sure, more research in this vein is required, especially considering the practical applications of this sort of work, and the assumptions that often accompany reasoning about metaphor and thought (Lakoff \& Johnson, 1980; Ramsay \& Robinson, 2015). We suggest researchers and public policy communicators interested in these issues should aim for more large-scale, pre-registered, and nuanced empirical studies of framing effects.

## References

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., \& Vohs, K. D. (2001). Bad is stronger than good. Review of general psychology, 5(4), 323-370.
Chaiken, S., Wood, W., \& Eagly, A. H. (1996). Principles of persuasion. In E. T. Higgins \& A. W. Kruglanski (Eds.), Social psychology: Handbook of basic principles (pp. 702-742). New York, NY: Guilford.
Flusberg, S. J., Matlock, T., \& Thibodeau, P. H. (2017). Metaphors for the war (or race) against climate change. Environmental Communication.
Gerber, M. M., \& Jackson, J. (2017). Justifying violence: legitimacy, ideology and public support for police use of force. Psychology, Crime \& Law, 23, 79-95.
Glucksberg, S. (2001). Understanding figurative language: From metaphor to idioms. Oxford: Oxford University Press.
Hardisty, D. J., Johnson, E. J., \& Weber, E. U. (2009). A dirty word or a dirty world? Attribute framing, political affiliation, and query theory. Psychological Science, 21, 86-92.

Johnson, J. T., \& Taylor, S. E. (1981). The effect of metaphor on political attitudes. Basic and Applied Social Psychology, 2, 305-316.
Gould (1977). Ever Since Darwin. W. W. Norton \& Company, Inc. New York, NY.
Lakoff, G., \& Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.
Petty, R. E., \& Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology (pp. 123205). New York, NY: Academic Press.

Ramsey, C., \& Robinson, L. (2015). Final report. The President's task force on 21st century policing. Retrieved Oct 28, 2016 https://cops.usdoj.gov/pdf/ taskforce/taskforce_finalreport.pdf
Reijnierse, W. G., Burgers, C., Krennmayr, T., \& Steen, G. J. (2015). How viruses and beasts affect our opinions (or not): The role of extendedness in metaphorical framing. Metaphor and the Social World, 5, 245-263.
Rumelhart, D. E. (1979). Some problems with the notion of literal meanings. In Ortonoy, A. (ed). Metaphor and Thought. Cambridge: Cambridge University Press.
Sopory, P., \& Dillard, J. P. (2002). The persuasive effects of metaphor: A meta-analysis. Human Communication Research, 28, 382-419.
Steen, G. J., Reijnierse, W. G., Burgers, C. (2014). When Do Natural Language Metaphors Influence Reasoning? A Follow-Up Study to Thibodeau and Boroditsky (2013). PloS ONE, 9(12): e113536.

Thibodeau, P. H. (In press). The Function of Metaphor Framing, Deliberate or Otherwise, in a Social World. Metaphor and the Social World.
Thibodeau, P. H. (2016). Extended metaphors are the home runs of persuasion: Don't fumble the phrase. Metaphor and Symbol, 31, 2, 53-72.
Thibodeau, P. H., \& Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. PLoS ONE, 6(2), e16782.
Thibodeau, P. H., \& Boroditsky, L. (2015) Measuring Effects of Metaphor in a Dynamic Opinion Landscape. PLoS ONE, 10(7): e0133939.
Thibodeau, P. H., Crow, L. \& Flusberg, S. J. (2016). The metaphor police: A case study of the role of metaphor in explanation. Psychonomic Bulletin \& Review. doi: 10.3758/s13423-016-1192-5

Thibodeau, P. H., \& Flusberg, S. J. (2017). Metaphorical accounting: How framing the federal budget like a household's affects voting intentions. Cognitive Science. doi: 10.1111/cogs. 12475
Thibodeau, P. H., Perko, V. L., \& Flusberg, S. J. (2015). The relationship between narrative classification of obesity and support for public policy interventions. Social Science \& Medicine, 141, 27-35.
Thibodeau, P. H., Sikos, L., \& Durgin, F. H. (2016). What Do We Learn from Rating Metaphors? Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA.

# Part-whole categorization is culture-specific 

Timothy J. Tilbe (timothyt @buffalo.edu)<br>Department of Linguistics, University at Buffalo, 609 Baldy Hall<br>Buffalo, NY 14260 USA

Jürgen Bohnemeyer (jb77@buffalo.edu)<br>Department of Linguistics, University at Buffalo, 609 Baldy Hall Buffalo, NY 14260 USA

Gabriela Pérez Báez (PerezBaezG@si.edu)<br>Department of Anthropology, National Museum of Natural History, P.O. Box 37012 Smithsonian Institution, Washington D.C., 20013-7012 USA<br>Dedre Gentner (gentner@northwestern.edu)<br>Department of Psychology, Northwestern University, Swift Hall 102, 2029 Sheridan Road<br>Evanston, IL 60208 USA


#### Abstract

We present two experiments on the role of culture in the categorization of object part-whole structures. A triadic categorization task pitted shape against function as factors driving similarity judgments on selected parts of different types of objects. Speakers of American English were significantly more likely than speakers of two indigenous languages of Mexico, Tseltal Maya and Isthmus Zapotec, to choose categorization by function, even when familiarity of the various stimulus objects was factored in. In the second study, members of the two indigenous groups matched parts of a doll to parts of novel objects of unfamiliar shape. The Tseltal participants were significantly more likely to match according to a shape-analytical algorithm rather than global analogy, consistent with predictions based on prevalent strategies in verbal part labeling in the two languages. We conclude that while cognition of object parts undoubtedly has a strong biological basis, there are also robust cultural effects.


Keywords: object mereology; meronymy; shape perception; function; cross-cultural research

## Introduction

The ability to categorize objects crucially involves identifying their parts. In general, mereology - the conceptualization of parts, and of how they relate to each other and to the wholes they form - is deeply involved in how humans make sense of their physical world.

A strong case can be made that the segmentation of physical objects into parts has a basis in shape recognition (Biederman 1987; Marr 1982; Palmer 1977; Tversky \& Hemenway 1984; inter alia) and is thus likely biologically grounded. At the same time, function plays a key role in the categorization of both body parts of living things and object parts of artifacts (e.g., Croft \& Cruse 2004: 153-156; Rose \& Schaffer 2015; Svorou 1994: 78-79, 91-92; Tversky 1989). Ears and lids can come in a great many distinct shapes; what unites these diverse manifestations is the function they play in the whole of which they are a part. But
the attribution of functions depends on knowledge, beliefs, and assumptions that are at least to a large extent learned. Function dependence thus creates an opening for cultural effects in cognition of parts.

In addition, there is some evidence suggesting that the role of shape, and geometry more generally, in mereological cognition may be to some extent subject to cultural variation as well. This evidence comes from meronymy, the nomenclature for object parts. In-depth studies of the meronymy of non-Indo-European languages are few and far between. But the few available reports present evidence of striking differences vis-à-vis the terminology familiar from English and other European languages.

Our interest in this complex was aroused by descriptions of the meronymies of two indigenous languages of southern Mexico, Ayoquesco Zapotec (MacLaury 1989) and Tseltal Maya (Levinson 1994). Both languages belong to the Mesoamerican linguistic and cultural area, the members of which have been in contact with one another for millennia (Campbell, Kaufman \& Smith-Stark 1986).

A feature that the accounts of MacLaury (1989) and Levinson (1994) converge on is a core set of meronyms that are assigned both to body parts of humans and animals and to the parts of inanimate objects (though not generally to plants). They claim that labeling of object parts with general-purpose body part terms is pervasive, and based largely on shape and geometry. In contrast, function-based meronyms for inanimate objects are largely absent.

Where the two systems appear to diverge is in the strategies used to assign the generalized meronyms to object parts. MacLaury (1989) describes a strategy strictly based on a global analogical mapping from the human body in canonical erect position to the object in its actual orientation at the time the assignment pertains to. (See Figure 1A for the part labelings that this account predicts for a novel object.) The parts that are named in this manner, the 'head,' 'face,' 'sides,' 'back,' and 'buttocks,' have fixed spatial
relationships to each other in any object. For instance, if one knows which part of the object is called its 'face,' and the vertical axis is determined, one can correctly predict the locations of its 'head,' 'sides,' 'back,' and 'buttocks.' The orientation of the object with respect to gravity is crucial; for example, the topmost part is the 'head,' no matter what the structure of that part happens to be, and changing the orientation of the object causes the labels to be reassigned.

In contrast, Levinson (1994) discusses the process by which meronyms are assigned to objects in Tseltal as an algorithm that takes a visual segmentation of the object as its input. (See Figure 1B for an example this account's predicted labelings for a novel object.) Axes of generalized cones are identified for the main volume and any secondary volumes of the object, as well as axes of symmetry. A sense of direction is assigned to each axis, and meronyms are assigned to the ends of the axes, in some instances taking the shape of the part into account. So, for instance, the default meronym for the head (in the vector sense) of the main axis is the word for 'head,' but if that region of the object is pointy, 'nose' is used, and if it is a negative space, 'mouth' is used. This system does not take into account the orientation of the object with regard to gravity, and because the axes are in certain ways independent of each other, meronyms do not occur in a fixed spatial schema. ${ }^{1}$

The question we wish to address in this paper is whether these putatively distinctive properties of meronymy are restricted to language, or whether they are associated with deeper cognitive differences - between Mesoamericans and English speakers on the one hand, and between Zapotec and Tseltal speakers on the other. We present two studies. Experiment 1 explores the respective role of shape and function in object part categorization, comparing data from speakers of Tseltal, Zapotec, and American English in a three-population design. Based on the available descriptions of verbal behavior, we predict function to play a greater role in the mereological categorizations of Americans than in those of either Tseltal or Zapotec participants.

Experiment 2 compares Tseltal and Zapotec participants in terms of their preference for categorizing the parts of unfamiliar objects by comparing them globally to the human body vs. by doing so based on the shape-analytical algorithm even when it is not licensed by a global mapping. If the differences in part categorization strategies go beyond language, we predict that the Tseltal participants should be more likely to prefer mappings that are at odds with global analogies.

## Experiment 1: shape vs. function

Speakers of English, Tseltal Maya, and Isthmus Zapotec compared images (with one exception, photographs were used) of part-whole configurations. Isthmus Zapotec is

[^517]

Figure 1. A: Meronyms predicted by the global mapping account. B: Meronyms predicted by the algorithmic account.
closely related to Ayoquesco Zapotec as described by MacLaury (1989); ongoing field research by the third author suggests that MacLaury's analysis of Ayoquesco meronymy applies to Isthmus Zapotec as well as far as the predictions of the present study are concerned, though with additional complications (Pérez Báez 2011). Each trial involved a triad of images. The participants selected the configuration that was least like the other two. The triads were composed so as to trade off functional against shape-based similarity. Since function attribution likely depends on the participants' knowledge of the object, a norming study was carried out to assess their familiarity with the stimuli.

## Method

Participants. 27 participants of each population were recruited at field sites in La Ventosa, Oaxaca, Mexico (Isthmus Zapotec - 16 women, 11 men; 14 young adult, 9 middle-aged, 4 elderly) and Tenejapa, Chiapas, Mexico (Tseltal Maya - 18 women, 7 men; 16 young adults, 8 middle-aged, and 1 elderly, with 2 participants' demographic information missing), the University at Buffalo and in Raynham, Massachusetts (English - 19 women, 8 men; 17 young adults, 6 middle-aged, and 4 elderly). Recruitment was conducted by word of mouth and at the University at Buffalo by flyer. Participants completed the tasks in about 30 minutes and were compensated 100 Pesos (approximately $\$ 5$ ) and $\$ 10$, respectively.

Materials. Each participant was given 12 test trials interspersed with 17 fillers, for a total of 29 trials, preceded by one practice trial of the same design as the fillers. The three pictures for each trial were printed on a single sheet, and the sheets placed into a binder so that one triad could be displayed at a time. The placement of the images on each page was pseudo-randomized, in order to reduce any possible bias toward choosing the picture in any particular one of the three positions.

Each trial consists of showing a participant three pictures of artifacts or plants that are presumably familiar to all three populations, with certain parts highlighted in red, and asking them to choose the one whose highlighted part they judge to
be most different from the other two. The experimental triads are designed to pit shape against function, in that there is a pivot object part and two alternates: the pivot shares its shape with one alternate, and its function with the other. An experimental triad is shown in Figure 2.

The filler trials use the same visual layout as the experimental trials, and the action the participant is expected to perform is the same. However, instead of a single pivot, there are two pictures in which the indicated parts of the object are similar in both shape and function, while the remaining picture's indicated part is the odd one out in terms of both shape and function. Therefore, a shapebased strategy and a function-based strategy would tend to produce the same response. This provides a check on the participants' attention to and comprehension of the task.

Procedure. The participants were instructed in their native languages by the first author and, in the case of the Tseltal and Zapotec participants, by bilingual assistants with experience in linguistic field research, to pick out the partwhole configuration in each triad they considered least like the others. The following standardized instructions were used: "In this game, I am going to show you some drawings and photos of various things - three at a time. And if you are not sure of what some of the objects are, please ask me. One part of each thing is red. Two of the parts are more similar, and the other is different. I want you to look at those parts, and find the different one. You should compare only the parts, not the whole objects. When you decide which is different, circle it using this marker. For example, let's look at these three [the practice triad]. Here there's a dog, with its leg red. And here is a cat, with its head red. And a pig, with its leg red. So, which part is different? Correct, the head. And so, do you have any questions before beginning? ${ }^{2}$

Norming. As part of the follow-up task 'Shape-Function Norming,' participants rated the familiarity of each individual picture used in the Shape-Function Triads. This serves the purpose of checking their interpretation of the pictures and providing additional factors for statistically modeling the experiment's results. Participants rated each of the three pictures in each experimental trial for familiarity, using a five-point Likert scale in their native language.

## Results

Participant exclusions. The responses of two Zapotec participants were excluded from the analysis because their performance on the filler trials was below the preestablished $80 \%$ threshold, indicating that they did not sufficiently comprehend the task.

Trial exclusions. Trials were excluded from the analysis in two situations. In one of these, the participant chose the

[^518]

Figure 2. Example of a test triad for Experiment 1
pivot as the odd one out in the trial in question, which is not interpretable as classifying by shape or function (11 trials in the Zapotec data set, 12 in the Tseltal data set, and 8 in the English data set). In the other situation, they did not give an answer to that trial at all ( 1 trial in the Zapotec data, 0 in the Tseltal data, and 0 in the English data). Approximately 4\% of the data points were missing or excluded.

Analysis. Figure 3 shows the breakdown of responses.
The responses to the norming scale, interpreted as numerals, were centered so that -2 corresponded to "completely unfamiliar" and 2 corresponded to "completely familiar." (The standard deviations of the ratings for each object ranged from 0.37 to 1.26 for the Zapotec speakers, and from 0.23 to 1.49 for the Tseltal speakers.) Within each language population, the mean of these familiarity scores was calculated for each picture in order to generalize across participants. Any picture whose average rating was equal to or greater than 1 was counted as "familiar," and any whose average rating was less than 1 was counted as "unfamiliar." The overall familiarity of each triad was coded by the number of familiar and unfamiliar pictures it contained, from "A" for three familiar pictures through "D" for none, and this code was treated as an ordinal factor in the regression models of the responses.

A binomial logistic mixed-effects regression model was fitted with population (identified by language) and code of triad familiarity as fixed factors, and random intercepts for participant and trial. Tseltal and Zapotec participants proved


Figure 3. Experiment 1 responses by population and type. Error bars represent 95\% prediction intervals.
significantly different from American participants at the $\mathrm{p}<.001$ level and from one another at the $\mathrm{p}<.05$ level. No significant effect of familiarity was obtained.

## Discussion

As predicted, the Mesoamerican participants were more likely to categorize parts by shape than the American participants. Shape in fact strongly dominated the Tseltal and Zapotec speakers' categorizations (at $65.4 \%$ and $76.8 \%$ of responses), whereas function strongly dominated among the English speakers $(67.7 \%){ }^{3}$ Familiarity did not appear to significantly affect these ratings, suggesting that the selection of the stimuli from among objects of everyday interactions for all three cultures was successful. It is possible, however, that the similarity judgments were influenced by subvocal use of meronyms.

## Experiment 2: global vs. shape-analytical mapping

The purpose of this experiment was to test for population differences in mereological cognition, by obtaining responses of nonverbal mapping from body parts to the parts of novel objects. This was done with the researcher indicating various parts of a humanoid doll, and asking the participants to find as many corresponding parts as they could on the Novel Objects. In order to test for effects of subvocal rehearsal, this task had two conditions: In one condition, the participants did a verbal elicitation task before the experiment, and in the other condition, they did not.

## Method

Participants. 44 Tseltal speakers ( 29 women, 16 men; mean age 39.7, SD 12.6) and 45 Isthmus Zapotec speakers ( 33 women, 12 men; mean age 33.2, SD 13.6) were recruited and tested in La Ventosa, Oaxaca, Mexico (Isthmus Zapotec) and San Cristóbal de las Casas, Chiapas, Mexico (Tseltal), relying again on word of mouth. The task took about an hour to complete and the participants were compensated 100 Pesos (approximately \$5).

Materials. The tasks used as stimuli a set of six solid plastic forms, part of the MesoSpace Novel Objects stimuli. ${ }^{4}$ These are abstract forms that, as far as possible, bear little resemblance to any item familiar to the participants, so that the objects are not biased toward any particular meronym assignment strategy, and the participants have to fall back on their general principles for construing mereological structure. Figure 1 features an example. For the practice

[^519]trial, instead of a Novel Object, a blobby humanoid figure made of Sculpey modeling clay was used. A plastic action figure doll representing a young adult male, fairly realistic in proportions, was used to represent a target body part in each trial - either the head, face, side (flank), back, or buttocks. These five parts were used because both MacLaury (1989) and Levinson (1994) had identified the meronyms for them as belonging to the languages' most productive meronymic systems, and these Zapotec and Tseltal terms are rough translation equivalents with regard to the human body. The participants used bits of Play-Doh to mark the parts of the Novel Objects that they judged as corresponding with these target parts. This doll was in a standing position in each trial, which, if an orientationdependent mapping strategy is used, would favor the choice of the uppermost part of the Novel Object as corresponding with the doll's head, the part(s) of the Novel Object to the participant's left or right as corresponding to the doll's side, etc.

Procedure. Conditions: In order to detect possible subvocal rehearsal effects, half of the participants carried out a verbal labeling task prior to the experiment; the other half did not do this task. Participants were randomly assigned to one condition or the other.
The verbal elicitation task was administered as follows: the participant was handed each of the Novel Objects, one at a time (in an opaque bag to avoid imposing any orientation on the object), asked to take it out and inspect it from all sides, and then prompted to delineate and label its parts. The experimenter recorded the delineated areas on twodimensional images of the objects.
Setup: In each trial, the experimenter set one of the Novel Objects on the tabletop directly in front of participant, sticking it into a base of Play-Doh if it would not stay in the desired orientation unsupported. Each Novel Object was therefore relatively in front of the participant and absolutely to the north (the participants were seated to face north). Trials: In each trial, one of the target parts of the doll was manually indicated by the experimenter. In view of the importance of vertical orientation in Zapotec meronymy, the trials were administered in two orientation variants: "aligned" (that is, the gravitationally-defined vertical axis coincides with the Novel Object's algorithmically-defined 'model axis', i.e., the axis from which the central volume of the object is generated), or "unaligned" (these axes are orthogonal). This yielded a total of 60 test trials: 5 doll parts x 6 Novel Objects x 2 orientations. In addition, there was one practice trial employing instead of a Novel Object a blobby humanoid figure made of Sculpey, designed to abstractly resemble the doll.

Instructions: The participants were instructed that when the experimenter indicated a part on the doll by delineating it with a finger, they should mark with the Play-Doh as many parts on the Novel Object as they thought corresponded to the doll's part, whether this resulted in no part being marked, just one part, or multiple parts. The
participants were instructed in their native languages by bilingual research assistants working with the first author as follows: "I'm going to give you some objects, one at a time. This first object is an example. You can turn it around to see how it is. And I'm going to show you some part of this doll, and you should decide if the object has a part of the same kind. It's possible that it doesn't have any. In that case, simply tell me that it isn't there. It's also possible that it has a part like that, or more than one at a time. In that case, take a bit of this Play-Doh and stick it to the part or parts that are similar. And when it's finished, I want you to lift the object and turn it slowly to show what you have done."

Recording: The responses, in terms of the landing sites on the Novel Objects the participants marked with Play-Doh, were recorded by the first author verbally in English on a coding sheet. The sessions were videotaped in their entirety.

## Results

Participant exclusions. One Zapotec participant's responses had to be excluded because the participant appeared unable to grasp the instructions.

Trial exclusions. 16 of the 60 test trials were excluded from the analysis because the algorithm Levinson (1994) proposed for meronymic labeling in Tseltal predicted that the particular Novel Object lacked a corresponding part, and the Tseltal participants nevertheless in almost all cases identified some part of it despite having been given the option not to select a mapping. These responses therefore could not be evaluated for whether they fulfilled the predictions of the algorithm. One Tseltal trial was not completed. The analysis was thus performed on 1,936 trials with Zapotec participants and 1,979 trials with Tseltal participants.

Coding. In order to code the responses for whether they fit the global mapping account, simplifying assumptions were adopted. When the 'head' of the doll had been indicated, the global prediction was considered fulfilled if and only if the Play-Doh was placed somewhere on the upper region of the object (as defined by the vector of gravity). Globally, the 'buttocks' had to be on the lower region, the 'face' on the region toward the participant, the 'back' on the region away from them, and the 'sides' on the regions to their relative left or right. These regions were interpreted as both surfaces and volumes. So, for example, a placement that was both on the top surface of the object, and also displaced from the center of that surface in the direction toward the participant, would satisfy the prediction for 'head' by virtue of being on the upper part of the object, and would also satisfy the prediction for 'face' by virtue of being on a volume part that is toward the participant. The volume interpretation of parts was also followed for the algorithmic predictions.

Response types. Four response types were distinguished, based on whether the proposed match was predicted solely by global mapping ('Global only'), solely by Levinson's
(1994) shape-analytical algorithm ('Algorithm only'), by both, or by neither. From MacLaury's (1989) global mapping account, two key predictions were derived: the 'face' of any object faces toward the observer, and the 'head' of any object points up against the pull of gravity. Since Levinson's (1994) algorithm is orientationindependent, it was assumed that given a part of the doll and a Novel Object, the intrinsic location of the matched part on the Novel Object will be constant across varied orientations.

Analysis. Figure 4 shows the breakdown of the four response types across the two populations.

Responses in "Doll to Novel Objects Mapping"


Figure 1. Experiment 2 responses by population and type.
Using Begg-Gray approximation of multinomial logistic regression (Begg \& Gray 1984), four binomial logistic mixed effects regression models were fitted, one for each response category, with population identified in terms of language, condition, alignment, and trial as fixed factors, and random intercepts for the stimulus doll part, the Novel Object, and the participant. The Algorithm-only model showed significant effects of population and alignment at the $\mathrm{p}<.001$ level. The Global-only model showed effects of alignment and trial at the $\mathrm{p}<.001$ level. There was a significant interaction between Zapotec and alignment at the $\mathrm{p}<.05$ level. The Both model yielded a significant effect of population at the $\mathrm{p}<.01$ level and alignment and trial at the $\mathrm{p}<.001$ level. The Neither model yielded no significant effects. None of the models produced a significant effect of condition.

## Discussion

As predicted, the Zapotec participants were significantly more likely to propose matches that agreed with global analogical mapping, but violated Levinson's (1994) shapeanalytical algorithm. Also in line with predictions, the orientation of the Novel Object had a significant effect on the Zapotec participants' matches, but not on those proposed by the Tseltal participants. Also as predicted, the "aligned" trials favored responses in the 'both' category. There was no effect of condition; we take this to suggest that subvocal rehearsal played no major part in the results. Subvocal rehearsal would have predicted that the two populations should have performed significantly more
different from one another in the Verbal-priming condition, contrary to fact. ${ }^{5}$

## General discussion

In both experiments, language proved a significant predictor of nonverbal mereological categorization: English speakers significantly preferred categorizing parts in terms of function, whereas Tseltal and Zapotec speakers were significantly more likely to categorize parts by shape (Experiment 1). And Zapotec speakers proved significantly more likely than Tseltal speakers to adhere to global analogy in mapping the parts of the human body to those of inanimate objects of unfamiliar shape, and were also significantly more likely to factor the orientation of the objects into their matches (Experiment 2).

We cannot exclude the possibility of subvocal rehearsal effects in Experiment 1. It is possible that the participants used their native languages for guidance in deciding between function-based and shape-based categorization. Future research will have to determine to what extent our results are truly representative of the nonverbal cognition of these groups. However, our findings are in line with previous research suggesting that geometry, as opposed to function, plays a relatively greater role among Mesoamericans compared to Westerners (Lucy \& Gaskins 2001). Meanwhile, in Experiment 2, we plausibly ruled out a significant contribution from language as a direct resource, suggesting robust differences in nonverbal cognition.

The findings presented here are also in line with Whorfian interpretations according to which language use may habituate speech communities to particular biases in mereological cognition and serve as a conduit of their cultural transmission (Bohnemeyer et al 2015). Here, too, we must defer to future research for ascertaining whether language merely reflects mereological cognition or is a causal factor in it.

## Conclusions

We have provided evidence of the existence of significant cross-cultural differences in the categorization of the mereology of physical objects. This is hardly surprising, as the categorization of objects and their parts clearly depends in part on acquired knowledge. What is surprising, in our view, is that research into such cultural effects is still in its infancy. We hope to have made a small contribution towards rectifying this.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1354277, 'Doctoral Dissertation Research: Parts and Wholes in Mesoamerican Language and Cognition,' awarded to the first author.

[^520]We gratefully acknowledge the contributions of JeanPierre Koenig, Peter Pfordresher, Gilles Polian, Reyna López López, Rosaura López Cartas, Antonia Sántiz Girón, María Gómez Sánchez, and John Pate.

## References

Begg, C. B., \& Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika, 11-18.
Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. Psychological review, 94(2), 115.
Bohnemeyer, J., Donelson, K. T., Moore, R. E., Benedicto, E., Eggleston, A., O’Meara, C. K., Pérez Báez, G., Capistrán Garza, A., Hernández Green, N., Hernández Gómez, M., Herrera Castro, S., Palancar, E., Polian, G., \& Romero Méndez, R. (2015). The Contact Diffusion of Linguistic Practices. Language Dynamics and Change, 5(2), 169-201.
Campbell, L., Kaufman, T., \& Smith-Stark, T. C. (1986). Meso-America as a linguistic area. Language, 530-570.
Croft, W., \& Cruse, D. A. (2004). Cognitive linguistics. Cambridge University Press.
Levinson, S. C. (1994). Vision, shape, and linguistic description: Tzeltal body-part terminology and object description. Linguistics, 32(4-5), 791-856.
Lucy, J. \& Gaskins, S. (2001). "Grammatical categories and the development of classification preferences: a comparative approach." In S. C. Levinson \& M. Bowerman (eds.), Language acquisition and conceptual development. Cambridge UP.
MacLaury, R. E. (1989). Zapotec body-part locatives: Prototypes and metaphoric extensions. International Journal of American Linguistics, 55(2), 119-154.
Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman.
Palmer, S. E. (1977). Hierarchical structure in perceptual representation. Cognitive psychology, 9(4), 441-474.
Pérez Báez, G. (2011). Semantics of body part terms in Juchiteco locative descriptions. Language Sciences, 33(6), 943-960.
Rose, D., \& Schaffer, J. (2015). Folk mereology is teleological. Noûs.
Svorou, S. (1994). The grammar of space (Vol. 25). John Benjamins Publishing.
Tversky, B. (1989). Parts, partonomies, and taxonomies. Developmental Psychology, 25(6), 983-995.
Tversky, B., \& Hemenway, K. (1984). Objects, parts, and categories. Journal of Experimental Psychology: General, 113, 169-193.

# The Stroop Effect From a Mixture of Reading Processes: A Fixed-Point Analysis 

## Gabriel Tillman (gabriel.tillman@newcastle.edu.au)

Zachary Howard (zachary.howard@newcastle.edu.au)
Paul Garret (paul.garret@newcastle.edu.au)
Ami Eidels (ami.eidels@newcastle.edu.au)
School of Psychology, University of Newcastle
NSW, 2308, Australia


#### Abstract

For the last 80 years, the Stroop task has been used to test theories of attention and cognitive control and it has been applied in many clinical settings. Most theories posit that the overwhelming power of written words overcomes strict instructions to focus on print color and ignore the word. Recent evidence suggests that trials in the Stroop task could in fact be a mixture of reading trials and non-reading trials. Here we conduct a critical test of this mixture hypothesis, where a mixture of processes should satisfy the fixed-point property (Falmagne, 1968).


Keywords: Stroop Effect; Mixture Model; Fixed-Point Analysis

The Stroop effect is one of the most replicated experimental effects in cognitive psychology (see MacLeod, 1991, for a review). The effect has been used to investigate cognitive control, and has also been applied in many clinical settings (Strauss, Sherman, \& Spreen, 2006, p. 477). The task involves naming the print color of a word, where the word itself is typically the name of a color (e.g., the word GREEN printed in red print requires a response of 'red'; Stroop, 1935). People are faster at naming the print color when it matches the word (congruent stimuli, e.g., RED in red) compared to when the word and print color do not match (incongruent stimuli, GREEN in red).

A common measure of the Stroop effect is the difference in mean response time (RT) between congruent and incongruent trials. In the Stroop task, participants are instructed to name the color and ignore the word, yet it seems people cannot help but read the word (e.g., Cohen, Dunbar, \& McClelland, 1990; Melara \& Algom, 2003), which gives rise to faster responses on congruent trials than incongruent trials on average. Although reading must happen on some trials for an effect to be observed, it is not clear whether reading occurs on every trial, or to the same extent across trials.

Eidels, Ryan, Williams, and Algom (2014) compared the Stroop effect obtained from a standard Stroop task to the effect obtained from a novel forced-reading task. In the standard task, participants were asked to classify
the print color of color-words irrespective of the content of the word. In the forced-reading task participants were asked to classify the print color of color-words (e.g., RED, GREEN), but withhold their response when presented with non-color-words (BED, GREED). To conform with the instructions, participants were forced to read every word presented. Consequently, the forcedreading Stroop task yielded a Stroop effect derived from fully processed words on every trial. The researchers found that the magnitude of the Stroop effect in the forced-reading task was larger than in the standard task, suggesting that the standard Stroop effect results from reading on only a portion of trials (see also Tillman, Eidels, \& Finkbeiner, 2016).

One possible account for these results is that on any particular trial of the standard task a participant might only be processing the word to a limited extent, or not processing the word at all. A simple, formal way of explaining how different processes are mixed to yield some observed distribution of RTs is a probability-mixture model (Eidels et al., 2014; Tillman et al., 2016). Under this model, the empirical RT distributions observed in either the congruent or incongruent conditions of the standard task are a binary mixture of two unobserved distributions: one distribution of reading trials and one distribution of non-reading trials. A given trial is a sample drawn from the reading distribution (with probability $p$ ) or the non-reading distribution (with probability $1-p)$. The forced-reading task increases the probability of reading to $(p=1)$.

This mixture-of-reading-processes hypothesis can be tested in a number of ways. One method assumes that a mixture of two different RT distributions should result in a bimodal observed distribution, and applies Hartigan's dip test to assess the bimodality. The test assumes the null hypothesis of unimodality over the alternative hypothesis of multimodality. If the dip statistic is greater than the 95th percentile of the reference distribution, then the null hypothesis is rejected and the observed distribution is considered bimodal (Hartigan \& Hartigan, 1985). Another method is to fit both a one-component
and a two-component Gaussian mixture model to the observed data and compare both models using model selection techniques, such as AIC (Akaike, 1974).

In simulation studies, researchers have found that the Hartigan's dip test correctly identifies bi-model distributions only $65 \%$ of the time and the AIC model selection method falsely identifies bimodality $80 \%$ of the time (Freeman \& Dale, 2013). In general, bimodality is difficult to detect in empirical data and requires the underlying distributions to be well separated and variability to be low (Williams, Eidels, \& Townsend, 2014). However, recent software and computational advances may facilitate a more robust approach to this problem. In this paper, we test the hypothesis that the standard Stroop effect results from a mixture of reading processes by using a mathematical property of probability-mixture distributions, the fixed-point property (Falmagne, 1968).

## The Fixed-Point Property

A set of mixture distributions, which are all based on the combination of two base distributions, will all intersect at a common coordinate - the fixed-point property (Falmagne, 1968, see Figure 1). Although this mathematical property could be a powerful means of identifying mixture models, researchers in the past have not commonly employed the fixed-point property test for two reasons (van Maanen, de Jong, \& van Rijn, 2014). Firstly, estimating the probability density function (PDF) of the observed RT distribution from noisy data is not trivial. Secondly, it has been difficult to provide statistical evidence for the presence of the fixedpoint property, which requires providing evidence for the null hypothesis.

We address the first issue by using the Epanechnikov kernel density function (Epanechnikov, 1969), which has been shown to approximate the PDF of RT distributions well (Silverman, 1986; Turner \& Sederberg, 2014). To select a bandwidth for the kernel we use Silverman's "rule of thumb" (Silverman, 1986, p. 48, eq (3.31)). The default software libraries in R ( R Development Core Team, 2016) allow for easy use of both the Epanechnikov kernel and Silverman's "rule of thumb". We address the second issue by using Bayesian methods to assess the degree to which there is no difference between a particular crossing point in all mixture distributions. Bayesian hypothesis testing, or Bayes factors, quantify evidence in favor of either the null hypothesis or the alternative hypothesis as a ratio. For example, when $B F_{10}=5$ the observed data are 5 times more likely under the alternative hypothesis than under the null hypothesis. When $B F_{10}=$ .2 the observed data are 5 times more likely under the null hypothesis than under the alternative hypothesis.

There is some precedent for using a fixed-point analysis to test mixture models in RT data (Brown, Lehmann, \& Poboka, 2006; van Maanen et al., 2014; van Maanen, 2016). For RT distributions, when the observed RTs are


Figure 1: Illustration of the fixed-point property in Stroop distributions. The 'No Reading' distribution consists of $0 \%$ reading trials and the Forced distribution consists of $100 \%$ reading trials. The Standard distribution is a mixture of trials from both distributions. All three distributions will intersect at a common point, which is labeled "Fixed-point property" in the figure.
made up of a mixture of unobserved distributions, there will be one RT for which the probability of providing a response at that particular time is equal for all mixtures (see Figure 1 again).

Here we test whether RT distributions in the Stroop task satisfy this fixed-point property. In the Stroop task participants are requested to classify the print color of color words and ignore the words' meaning. The ubiquitous Stroop effect implies they fail to exclusively focus on color and succumb to the overwhelming (perhaps automatic) attraction of reading. Previous evidence suggests that participants may not always process word meaning to the same extent (Eidels et al., 2014; Tillman et al., 2016). They may process words on some trials and not on others, in a way commensurate with a binary mixture model.

To test the mixture model we presented participants with three experimental conditions, each intended to induce a different level of reading (gauged by the probability p): a color naming task involving rectangles (probability of reading, $p=0$ ), forced reading Stroop task in which each word must be read to its full extent, on each and every trial $(p=1)$, and a standard Stroop task, where participants may involuntarily read the words on some proportion of the trials $(0<p<1)$. A probabilitymixture account of reading in the Stroop task predicts that RT distributions of the three conditions will cross each other at a common point (the 'fixed-point property').

## Method

## Participants

Twenty two students (19 females and 3 males) from the University of Newcastle (mean age $=22.41$ and SD age $=$ 4.74) participated in the study. Participants had a proficiency in English and normal or corrected to normal vision with intact color vision. Each participant completed the standard, forced, and color naming Stroop tasks and participants were reimbursed $\$ 15$ per session.

## Apparatus

Each task was carried out on Dell computers running Windows XP with 17 " Diamond View color monitors. Contrast and brightness were set to 80 and 50 , respectively. We used the Tektronix J17 Lumacolor digital photometer and J1800 series sensor heads to calibrate color clarity across all testing stations. The software 'Presentation' was used to run the experiment and record data. Participants responded using a Cedrus response pad, model RB-830. The response keys on the response pad were marked with color stickers corresponding to the red, green or blue response.

## Stimuli

For the color naming task, the stimuli were color filled rectangles in the center of the screen. For the standard and forced task, the stimuli were were 12 words that were printed in either the color red, green, or blue. The 12 words were RED, GREEN, and BLUE and three variants for each of these words. The variants differed from the color words by one letter and if substituting one letter resulted in a non-word, two letters were changed instead. The variants were GREED, GRAIN, QUEEN, RENT, ROD, BED, BASE, BLUR, and GLUE.

The variants made up the neutral stimuli for the standard and forced-reading task. The neutral stimuli were matched to the color stimuli on length, neighborhood frequency, and phonetics using the software N-Watch (Davis, 2005) and based on the CELEX word frequency database. All words were written in uppercase bold Arial font, with no words exceeding 2.55 cm , or 4 visual degrees when the participant was seated 60 cm from the screen.

Red, green and blue print colors of the words and rectangles had $R G B$ values of $R=220, G=0, B=0$ for red, $\mathrm{R}=0, \mathrm{G}=0, \mathrm{~B}=240$ for blue, and $\mathrm{R}=0, \mathrm{G}=170, \mathrm{~B}=0$ for green. The stimuli made up three conditions in the standard and forced-reading task. The congruent condition consisted of stimuli that had the print color and word match (RED in red, GREEN in green). The incongruent condition consisted of stimuli that had the print color and word mismatch (RED in green, GREEN in red). All non-color words were classified as neutral trials.

## Procedure

Each participant completed three sessions on separate days. Each session involved the standard, forcedreading, or color naming task. The former two took about an hour to complete and consisted of 10 experimental blocks with 1 minute breaks between each. The color task took 20 mins and consisted of one experimental block. The order of task presentation and position of the response buttons was counterbalanced across participants.

Each task was completed in a dark room with a desk lamp as the light source. At the beginning of each session, participants were shown 9 example trials that demonstrated the correct response. They also completed two practice blocks that consisted of 24 trials, with feedback for correct and incorrect responses in the first block.

In the color naming task, participants were instructed to respond to the print color of the rectangles by pressing the corresponding button on the response pad. In the standard task, participants were instructed to ignore the word and respond to the print color of the word. In the forced-reading task, participants were instructed to respond to the print color of words, but withhold responses to neutral words (e.g., BED, GREED, RENT).

On each trial, a fixation cross appeared in the center of the screen for 500 ms , followed by a blank screen for 500 ms . Following this, either a rectangle printed in color (color naming task) or a word printed in color was presented for 500 ms in a random position within 40 pixels distance from the center. The spatial uncertainty prevented participants from using spatial cues to respond. Participants were required to respond within 2500 ms after stimulus presentation before the trial timed out.

The color naming task involved 50 trials of blue, red, and green rectangle presentations, making for 150 trials in total per participant. For the standard and forced task, each of the ten experimental blocks consisted of 18 congruent trials, 36 incongruent trials, and 54 neutral trials. In the forced task, this allowed for half the trials to contain no response, which controls for participants predicting a non-response trial. Each combination of congruent and incongruent stimuli were presented 6 times per block. The order of stimulus presentation was randomized within each block. The RT was recorded in milliseconds.

## Results

The probability mixture account makes two testable predictions. First the fixed cross point, where all three Stroop distributions will have a single RT with equal probability of providing a response at that time - we test this in the following section. The mixture account also predicts that the observed (mixture) distribution will be bound between the faster non-reading distribution and the slower forced-reading distribution. This is exactly
what we observed in our data. The color naming task had a mean RT of 436 ms ( $\mathrm{SD}=132 \mathrm{~ms}$ ). In the congruent condition, the standard and forced-reading tasks had mean RTs of $470 \mathrm{~ms}(\mathrm{SD}=150 \mathrm{~ms})$ and 700 ms (SD $=208 \mathrm{~ms}$ ), respectively. While in the incongruent condition, these tasks had mean RTs of 495 ms ( $\mathrm{SD}=172 \mathrm{~ms}$ ) and 832 ms ( $\mathrm{SD}=233 \mathrm{~ms}$ ), respectively.

We also used Bayesian paired samples $t$-tests to evaluate the evidence for differences between the color naming, standard, and forced-reading mean RTs in the congruent and incongruent conditions. In the congruent condition, participants were slower in the standard task than the color naming task $\left(B F_{10}=8.8 \times 10^{660}\right)$ and were slower in the forced-reading compared to the standard task $\left(B F_{10}=5 \times 10^{559}\right)$. In the incongruent condition, participants were slower in the standard task than the color naming task $\left(B F_{10}=1.2 \times 10^{1295}\right)$ and were slower in the forced-reading task compared to the standard $\left(B F_{10}=6.9 \times 10^{2770}\right)$.

## Fixed-Point Analysis

The analysis was carried out using the 'fp' package (van Maanen et al., 2014) in R (R Development Core Team, 2016) - but we used the Epanechnikov kernel instead of the default Gaussian kernel as recommended by Silverman (1986, p. 43).

The analysis involved calculating the probability density of each RT distribution in each task. For example, focusing only on the congruent condition (and later similarly focusing on the incongruent condition) we estimated the RT distribution for the color naming task, the standard task, and the forced reading task, which by design have a mixture proportion of $p=0,0<p<$ 1 , and $p=1$, respectively. We then found the crossing point of each pair of distributions (i.e., forced-standard, forced-color naming, standard-color naming). The fixedpoint property holds if all pairs cross at the same point along the x and y axis (see Figure 1).

We tested whether the fixed-point property holds for the sample of participants in our study. We calculated the crossing points per pair of mixture proportion tasks for each of the participants for both the congruent and incongruent distributions, but the color naming distribution was the same for both the congruent and incongruent comparison. We then subjected these crossing points to Bayesian analysis of variance (ANOVA). The Bayesian analysis was conducted using the Bayes Factor package (Morey, Rouder, \& Jamil, 2014; Rouder, Morey, Speckman, \& Province, 2012) in R. The Bayes factor from the ANOVA provides evidence for or against the fixed-point property.

We calculated the Bayes factor as the ratio of the evidence for the null hypothesis over the alternative. The null hypothesis posits that there is no difference in crossing points between all distributions in question, and thus suggests that the fixed-point property is satisfied. In line
with Kass and Raftery (1995) we consider a Bayes factor greater than 3 as positive evidence in favor of the null (fixed cross point) and against the alternative hypothesis that there is a difference between crossing points.

The RT distributions for the congruent and incongruent trials are presented in Figure 2. For the congruent condition, the Bayes factor ANOVA revealed that the null model was preferred to the alternative model by a Bayes factor of 1.25 . The data provide equivocal evidence in favor of both the null and alternative hypothesis for the congruent Stroop distributions. For the incongruent condition, the Bayes factor ANOVA revealed that the null model was preferred to the alternative model by a Bayes factor of 2.27 . The data provides evidence in favor of the hypothesis that there is no differences between crossing points, but the evidence is inconclusive.

## General Discussion

In the Stroop task, slower responses on incongruent trials relative to congruent or even neutral trials implies participants read the words despite instructions to focus on color and ignore the words' meaning. Recent evidence suggests participants may read on some proportion of the trials and not on others (Eidels et al., 2014; Tillman et al., 2016). When the observed RT on a single trial is sampled from a non-reading distribution, color naming will not be slowed down by the incongruent word. When the observed RT on a single trial is sampled from a reading distribution, the speed of color naming will be slowed down by an incongruent word, therefore, contributing to a Stroop effect. The magnitude of an observed Stroop effect reflects the proportion of trials in which the participant has read on - the greater the proportion, the larger the effect. To statistically test for this mixture of reading processes in the Stroop task, we ran a fixed-point property analysis on Stroop RT distributions with different reading proportions. We found some evidence for a mixture of distributions in the incongruent condition, but the results of the analysis were not conclusive.

The fixed-point property analysis is one method for testing for a mixture of processes, but it requires the strong assumption that there is a pure mixture of reading and non-reading processes. That is, the approach assumes that the only difference between the three tasks is the proportion of reading trials. This assumption may be compromised by other contaminant processes across the tasks. For example, the Stroop effect can be effected by attentional resources (Kahneman \& Chajczyk, 1983), practice (MacLeod \& Dunbar, 1988), dimensional discriminability and experimental correlation (DishonBerkovits \& Algom, 2000), target set size (La Heij \& Vermeij, 1987), and the number of colored letters in the stimulus word (Besner, Stolz, \& Boutilier, 1997). Further, there are differences in stimuli (words vs rectangles) across tasks. Whilst our results are inconclusive with re-


Figure 2: Overall RT density for congruent and incongruent Stroop distributions.
gards to identifying a mixture process, they certainly do not preclude the mixture hypothesis as being a viable explanation for the Stroop effect.

Our study also reflects the difficulties in distinguishing between single-process and dual-process mental phenomena, which is an issue that besets cognitive psychology (e.g., Yap, Balota, Cortese, \& Watson, 2006; Wixted, 2007; Freeman \& Dale, 2013). Nonetheless, the mixture model of Stroop has clinical, empirical, and theoretical implications. If the Stroop effect distribution is derived from a reading distribution and a non-reading distribution, and the combination of these distributions makes up the observed distribution, then clinical applications of the Stroop task need to consider this mixture of reading processes. For instance, differences in Stroop effect magnitude may not only reflect differences in attentional control, but could simply reflect a difference in the proportion of reading across trials. Empirically, future work could account for the proportion of reading trials by employing the benchmark forced-reading task along with the standard task. Finally, theories of Stroop (e.g., Cohen et al., 1990; Melara \& Algom, 2003) will need to consider what mechanism allows for a Stroop effect to only arise on some proportion of trials but not others. Given these implications, we hope to see more robust testing of the mixture-of-reading-processes hypothesis outlined here.

## Acknowledgments

We would like to thank Katie Berka for data collection and Scott Brown for helpful comments related to the fixed-point analysis.

## References

Akaike, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716-723.
Besner, D., Stolz, J. A., \& Boutilier, C. (1997). The stroop effect and the myth of automaticity. Psychonomic Bulletin \& Review, 4 (2), 221-225.
Brown, S. D., Lehmann, C., \& Poboka, D. (2006). A critical test of the failure-to-engage theory of task switching. Psychonomic bulletin Eamp; review, 13(1), 152159.

Cohen, J. D., Dunbar, K., \& McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. Psychological review, 97(3), 332-361.
Davis, C. J. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. Behavior research methods, 37(1), 65-70.
Dishon-Berkovits, M., \& Algom, D. (2000). The stroop effect: It is not the robust phenomenon that you have thought it to be. Memory \& Cognition, 28(8), 14371449.

Eidels, A., Ryan, K., Williams, P., \& Algom, D. (2014). Depth of processing in the stroop task: Evidence from
a novel forced-reading condition. Experimental Psychology, 61 (5), 385-393.
Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability 6 Its Applications, 14 (1), 153-158.
Falmagne, J. C. (1968). Note on a simple fixed-point property of binary mixtures. British Journal of Mathematical and Statistical Psychology.
Freeman, J. B., \& Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. Behavior research methods, 45(1), 83-97.
Hartigan, J. A., \& Hartigan, P. (1985). The dip test of unimodality. The Annals of Statistics, 70-84.
Kahneman, D., \& Chajczyk, D. (1983). Tests of the automaticity of reading: dilution of stroop effects by color-irrelevant stimuli. Journal of Experimental Psychology: Human Perception and Performance, 9(4), 497-509.
Kass, R. E., \& Raftery, A. E. (1995). Bayes factors. Journal of the american statistical association, 90 (430), 773-795.
La Heij, W., \& Vermeij, M. (1987). Reading versus naming: The effect of target set size on contextual interference and facilitation. Perception $\mathcal{B}$ psychophysics, 41 (4), 355-366.
MacLeod, C. (1991). Half a century of research on the stroop effect: An integrative review. Psychological Bulletin, 109, 163-203.
MacLeod, C., \& Dunbar, K. (1988). Training and strooplike interference: evidence for a continuum of automaticity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14 (1), 126-135.
Melara, R. D., \& Algom, D. (2003). Driven by information: a tectonic theory of stroop effects. Psychological review, 110 (3), 422-471.
Morey, R., Rouder, J., \& Jamil, T. (2014). Bayesfactor: Computation of bayes factors for common designs. $R$ package version 0.9, 8.
R Development Core Team. (2016). The r project for statistical computing [Computer software manual]. Vienna, Austria.
Rouder, J. N., Morey, R. D., Speckman, P. L., \& Province, J. M. (2012). Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology, 56, 356-374.
Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.
Strauss, E., Sherman, E. M., \& Spreen, O. (2006). A compendium of neuropsychological tests: Administration, norms, and commentary. Oxford University Press, USA.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18(6), 643-662.
Tillman, G., Eidels, A., \& Finkbeiner, M. (2016). A
reach-to-touch investigation on the nature of reading in the stroop task. Attention, Perception, \& Psychophysics, 78(8), 1-11.
Turner, B. M., \& Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. Psychonomic bulletin \& review, 21(2), 227-250.
van Maanen, L. (2016). Is there evidence for a mixture of processes in speed-accuracy trade-off behavior? Topics in cognitive science, 8(1), 279-290.
van Maanen, L., de Jong, R., \& van Rijn, H. (2014). How to assess the existence of competing strategies in cognitive tasks: a primer on the fixed-point property. PloS one, 9(8), e106113.
Williams, P., Eidels, A., \& Townsend, J. T. (2014). The resurrection of tweedledum and tweedledee: Bimodality cannot distinguish serial and parallel processes. Psychonomic bulletin E review, 21 (5), 1165-1173.
Wixted, J. T. (2007). Dual-process theory and signaldetection theory of recognition memory. Psychological review, 114 (1), 152.
Yap, M. J., Balota, D. A., Cortese, M. J., \& Watson, J. M. (2006). Single- versus dual-process models of lexical decision performance: Insights from response time distributional analysis. Journal of Experimental Psychology: Human Perception and Performance, 32(6), 1324-1344.

# Interleaving area problems in the $4^{\text {th }}$ grade classroom: What is the role of context and practice? 

Rachael Todaro (rtodaro@kent.edu) \& Bradley J. Morris (bmorri20@kent.edu)<br>Department of Educational Psychology, 405 White Hall<br>Kent State University, Kent, OH 44242 USA


#### Abstract

Typical mathematics instruction involves blocked practice across a set of conceptually similar problems. Interleaving, or practice across a set of conceptually dissimilar problems, improves learning and transfer by repeatedly reloading information and increasing discrimination of problem features. Similarly, comparing problems across different contexts highlights relevant and irrelevant knowledge. Our experiment is the first to investigate the relative effects of interleaving geometry problems and interleaving contexts. Thirty-three fourth-grade students received the same practice problems but were randomly assigned to one of three conditions: interleaved by math skill, interleaved by context, and interleaved by math skill and by context (i.e., hyperinterleaved). Afterward, each participant was exposed to tests assessing declarative and procedural knowledge. The results suggest that interleaving math skill within varying contexts enhances the acquisition of mathematical procedures.


Keywords: interleaving, cognitive development, mathematics instruction

## Introduction

Mathematics has been subject to a broad array of interventions and techniques that could potentially improve learning, retention, and transfer of knowledge to novel contexts. One promising intervention is known as interleaved practice, in which exposures to concepts (e.g., math skills) are followed by exposures to dissimilar concepts (Rohrer, 2012). Another promising technique is presenting concepts in multiple contexts, which supports generalization (Vlach, Sandhofer, \& Kornell, 2008). The purpose of the present study was to examine the effects of interleaving geometrical problems across two different, yet familiar, contexts. We investigated the effects of interleaving context, math skill, or context and math skill by presenting this information in blocks (i.e., the same format across a series of examples). The main hypothesis of the present study was that interleaving across both context and math skill (hereafter hyper-interleaving) produces an additive effect, which would increase learning, retention, and transfer beyond other conditions because such presentation highlights differences between examples and supports greater discrimination among math skills.

## Mechanisms underlying Interleaving

Presenting math problems in an interleaved fashion improves performance outcomes because this type of presentation supports two fundamental mechanisms for
successful learning: discrimination training and repeated reloading. Discrimination training involves comparing and contrasting problem features, which may leading to a higher likelihood of increased learning of concepts and procedures as well as an increased ability to transfer solution strategies to novel problems (Kang \& Pashler, 2012; Rittle-Johnson \& Star, 2007). Repeated reloading occurs when a student revisits the same type of problem and supports effortful recall of the information, which increases the likelihood for successful encoding (Bjork \& Bjork, 2011).

Presenting interleaved problems typically consists of two components. The first component is the presentation of conceptually dissimilar problem types during the practice session (e.g., a problem about the area of a square following a problem about the area of a triangle). The second component is the distribution of those problems across multiple practice sessions. That is, the student returns to practice the interleaved problems on more than one instance (Rohrer et al, 2014). Presenting interleaved problems supports comparisons and contrasts between members of different categories (e.g., perimeter of squares vs. triangles). In this manner, comparing and contrasting perceptual and conceptual information not only promotes learning regarding how to perform each procedure, but trains the learner to discriminate which solution strategy is appropriate for each problem example (i.e., discriminative contrast; Birnbaum, Kornell, Bjork, \& Bjork, 2013; Kornell \& Bjork, 2008).

Initially, interleaved practice appeared to owe most of its effectiveness to spacing. Spacing is inherent in interleaved practice because there is time between each opportunity to practice concepts. This is distinct from massed practice as it allows students to forget irrelevant information between learning events, which increases the potency of encoding on subsequent presentations (Bjork and Allen, 1970; Cuddy and Jacoby, 1982). However, interleaving does not solely rely on the benefits and efficacy of spacing. Given the same amount of temporal space between each exposure, interleaved presentation produces greater gains in learning than blocked presentation (Kang \& Pashler, 2012). More recent literature emphasizes the importance of repeated reloading (Bjork \& Bjork, 2011), which suggests that accurate memory retrieval is enhanced when a learner must repeatedly reload specific concepts from long-term to shortterm memory. In fact, interleaved practice may provide its benefit more from repeated reloading rather than to the amount of time between successful reloads (e.g. temporal spacing; Kang \& Pashler, 2012).

## Learning with Contexts

Learning is context-dependent (Willingham, 2009). Placing math problems in familiar contexts may not only be an effective presentation method but an additive one as it activates domain knowledge that may facilitate learning and problem solving by providing a framework in which the student can make sense of the concept (Willingham, 2009; Rittle-Johnson \& Star, 2007). Context also offers the learner cues to solve a problem because it draws his/her attention to the right details and improves learning in memory retrieval, problem solving (Godden \& Baddeley, 1975), and reasoning tasks (Cheng \& Holyoak, 1985). The main finding across studies regarding context is that when learning and testing contexts are the same, there is an improvement in performance. These findings also suggest that if a student learns within a single context, she may fail to retrieve information outside the context (i.e., context dependency, Godden \& Baddeley, 1975; Vlach et al., 2011). Distributing learning across multiple events that use multiple contexts can reduce context dependency (Rothkopf, Fisher \& Billington, 1982; Smith, 1982). Additionally, learning across multiple contexts results in a greater number and variance of salient cues during learning, which may increase the likelihood of recall (Smith, 1982; Vlach et al., 2011). When multiple contexts are presented across multiple learning events and these contexts are similar, the shared contextual support leads to greater learning than does providing a single cue from one context (Thiessen \& Saffran, 2003; Vlach et al., 2011).

While ours is the first study to interleave both math skill and context, Rau, Aleven, and Rummel (2013) investigated how interleaving specific dimensions of math problems may affect performance. By either interleaving fraction problems (i.e., dividing fractions) or their graphical representations (i.e., pie chart, number line), they found that interleaving problems was most beneficial to learning whereas interleaving the representations was not. In a follow-up study, Rau, Aleven, Rummel, and Pardos (2014) found that interleaving both dimensions significantly benefited learning over interleaving problems alone.

The present study investigated the relative contributions of interleaved sequencing across math skills and context on declarative knowledge (i.e., knowledge of facts), procedural knowledge (knowledge of how to choose and carry out a procedure), and transfer assessments (i.e., ability to apply knowledge in novel contexts). In addition, we investigated whether these elements produce an additive effect in which their combination produces greater learning gains than either element presented individually. To test these effects, we created a $2 \times 2$ factorial design as follows: math skill interleaved (interleaved math skill/blocked context), context interleaved (blocked math skill/interleaved context), hyperinterleaved (interleaved math skill/interleaved context). The authors chose to omit the fully condition (blocked math skill/blocked context) since it does not address the current research question, as both dimensions are blocked. As mentioned in the introduction, we
hypothesized that the hyperinterleaved group would perform better than the math skill interleaved and context interleaved group on all assessments post-intervention (i.e., posttest, delayed posttest, transfer, and delayed transfer) due to increased discrimination training across contexts and procedures. We also predicted that the math skill interleaved group would perform better than the context interleaved condition since relevant problem features are highlighted.

## Method

## Participants

Thirty-seven children ( 15 girls) ranging from nine to ten years of age were recruited from an elementary school in Northeast Ohio and completed up to six separate sessions. Each child was enrolled in the $4^{\text {th }}$ grade during the 20162017 school year. Four students completed the first session but were not present for either the second or third session and their data were excluded from all analyses. The remaining 33 students completed both sessions of the intervention and were subject to analyses.

## Task

Participants were taught to define and solve for area of four, two-dimensional geometrical shapes: square, triangle, rectangle, and parallelogram. Although various problems resulted in the same numerical solution during pretest, intervention, post-test, transfer, delayed posttest and delayed transfer test, no problem appeared more than once. In each problem, the participants were given the necessary features of each shape to solve the problems successfully (e.g., for a triangle, the base, height, and length of the sides were given). Throughout all sessions of the experiment, the participants needed to select and carryout the appropriate math skill (e.g., solving for the area of each shape).

## Design \& Procedure

The current study included five sessions: a pretest, first and second intervention, a one-day delayed posttest and transfer test, and a 30-day delayed posttest and transfer test. Each of the first four sessions occurred across four consecutive days, one session per day. The fifth session occurred 30 days after the fourth session ended. All sessions occurred in the $4^{\text {th }}$ graders' classroom. Each participant was randomly assigned to one of three conditions: math skill interleaved, context interleaved, or hyper-interleaved. Problems in the math skill interleaved condition were presented such that the area problems were interleaved but blocked by context. Problems in the context interleaved condition were presented such that problem contexts were interleaved but the math skill was blocked. Problems in the hyperinterleaved condition were presented such that both context and math skill were interleaved. See Table 1. The math skills used in the study were formulas in which to solve area (e.g., area of a square). The contexts used in the current study were "indoor maintenance" (How much carpet is needed to cover a bedroom?) and "outdoor renovations" (How many feet of
the playground is covered by mulch). The first author administered all procedures described below.

## Coding Rubric

The coding rubric was created for the purpose of representing the performance expectations for all assessments. The rubric was separated into two component parts (i.e., declarative and procedural knowledge) and provided clear descriptions of criteria needed to satisfy each one. The declarative knowledge component assessed adequate shape drawing as well as accurate defining of geometrical terms (i.e., area). The rubric included acceptable answers (i.e., "space inside of a shape") ranging from one to two points and unacceptable answers (i.e., area around the shape"), which were worth zero points. For the procedural knowledge component, the following criteria were included: two points for the application of the correct procedure, one point for correct answer, and one point for correct unit notation. Inter-rater reliability was strong for the declarative and procedural components, with Cohen's kappa $=.90$ and .96 , respectively.

Table 1: Groups were presented problems interleaved by math skill, context, or by both math skill and context.

| Group | Sequence of Math <br> Skill | Sequence of <br> Context |
| :---: | :---: | :---: |
| Math Skill <br> Interleaved | Interleaved | Blocked |
| Context <br> Interleaved <br> Hyperinterleaved | Blocked | Interleaved |

## Session 1: Pretest

During the first session, participants were administered a pre-test in order to assess prior declarative knowledge regarding the concept of area. One question was asked regarding the definition of area. Students earned up to two points if they included key terms listed on the scoring rubric. They were also asked to draw each shape, which was worth one point. Participants were also asked to define the shapes with the possibility of earning up to two points for each shape's definition. Additionally, within the pretest the participants were given two area problems for each shape to solve, which assessed prior procedural knowledge. Each procedural problem was worth four points. As per the scoring rubric, the participants needed to demonstrate the correct procedure, correct answer, and correct notation of units (e.g., $\mathrm{ft}^{2}$ ). In total, participants were able to earn up to 46 points on the pretest. The participants were given 30 minutes to complete the pretest. The pretest problems did
not contain a context but a box below asking for the area. The participants were expected to show their work inside the box.

## Session 2: Intervention Phase 1

During the second session, students were given a supplemental packet with the definitions of concepts (e.g., height, base, area, etc.) and worked examples (Sweller, 1988) of area problems across all four shapes.

Shortly afterward, students were given a brief lesson about the area of each shape encouraging the students to identify key words such as "cover" for area, show their work in the specified boxes below each figure, and to use their supplemental packet to follow along in the lesson. The lesson lasted between 10-15 minutes.

After the lesson, participants were administered the first phase of the training packet. The entire training packet consisted of 24 area problems to be divided between three sessions of training. Two area problems of each shape were included in each training packet. The first phase of the training packet consisted of a mixture of 8 area problems across either one or both contexts depending on which condition the participants pertained. If the participants had questions, they were directed to the supplemental packet and were advised to focus on the cue words in order to solve for area of the four various shapes. Students were given minimal feedback for the duration of the intervention. The first phase of the intervention lasted approximately 30 minutes.

## Session 3: Intervention Phase 2

The second intervention session was similar to the first. The students received the second set of eight area problems of four different shapes across one or both contexts. The participants were given $25-30$ minutes to complete the second session of the intervention.

## Session 4: Posttest and Transfer Test

The posttest was similar to the pretest but differed in numerical measurements across all four shapes in order to avoid practice effects. The first section required that the students define the concepts of area along with drawing and defining the four different shapes. Again, this section assessed declarative knowledge and was worth a possible 14 points. The participants then completed four area problems of each shape for a total of eight posttest problems. Six questions that contained novel contexts were included in the posttest, in order to assess for transfer of procedural knowledge. The purpose of including these problems was to investigate whether any of the presentation conditions would lead to increased transfer to novel contexts. The transfer problems dealt with contexts distinct from those seen previously in the experiment (baseball fields, cooking, distances riding a bike, etc.). Like the pretest, participants could earn up to four points on each posttest and transfer procedural problem provided that they demonstrate the correct procedure, answer, and unit notation. In total,
participants were able to earn up to 46 points on the posttest and 24 points on transfer problems.

## Session 5: 30-day Delayed Posttest and Transfer Test

The delayed post-test and the delayed transfer test were similar to the original post-test and transfer test. The first section of the delayed posttest was identical to the one-day posttest: the participants defined, drew, and described the concepts and shapes listed. The problems on both the delayed post-test and delayed transfer test only differed slightly on the numerical measurements of the shapes and novel figures in order to avoid practice effects. In total, participants were able to earn up to 46 points on the delayed posttest and 24 points on delayed transfer problems.

Table 2: Means and Standard Deviations $(\mathrm{N}=33)$

| Test | Math Skill <br> Interleaved <br> $(n=12)$ | Context <br> Interleaved <br> $(n=10)$ | Hyper- <br> interleaved <br> $(n=11)$ |
| :--- | :---: | :---: | :---: |
| Pretest |  |  |  |
| Declarative | $5.1(3.0)$ | $5.5(2.8)$ | $5.7(2.8)$ |
| Procedural | $5.1(4.6)$ | $1.7(1.8)$ | $4.4(4.5)$ |
|  |  |  |  |
| Posttest |  |  |  |
| Declarative | $6.3(1.6)$ | $7.4(2.2)$ | $7.1(2.2)$ |
| Procedural | $18.2(8.2)$ | $9.3(5.9)$ | $20.0(9.3)$ |
| Transfer | $12.3(6.9)$ | $9.2(5.8)$ | $15.8(9.0)$ |
|  |  |  |  |
| Delayed |  |  |  |
| Posttest | $7.8(1.9)$ | $7.0(2.3)$ | $7.3(2.7)$ |
| Declarative | $19.4(11.8)$ | $14.3(5.8)$ | $23.2(5.2)$ |
| Procedural | $12.9(7.0)$ | $11.1(5.1)$ | $14.8(7.0)$ |
| Transfer |  |  |  |

## Results

Table 2 displays the means and standard deviations for the pretest, posttest, transfer, delayed posttest, and delayed transfer test separated by declarative items, problems, and transfer problems. It is clear in the table that there was a great deal of variability within each group on each of the problem and transfer test scores.

Three repeated measures ANOVAs were conducted to examine the differences in number of points scored as a function of (1) 3 Interleaving Types (between; Math Skill Interleaved, Context Interleaved, and Hyperinterleaved x 3 Procedural Problems (within; Pre, Post, Delayed); (2) 3 Interleaving Types x 2 Transfer Problems (within; Transfer \& Delayed Transfer); (3) 3 Interleaving Types x 3 Declarative Knowledge Problems (within; Pre, Post, Delayed).

## Interleaving x Declarative Knowledge

In this analysis we used the three interleaving types as a between-subjects variable and pretest, posttest, and delayed posttest as a within subject variable. This was to examine whether or not interleaving types influenced declarative knowledge. The test of within subjects effects indicated a main effect of test, $F(2,54)=7.746, M S e=27.033, p<$ $.001, \eta_{p}^{2}=.223$. Within subjects contrasts indicated that participants performed better on the posttest declarative questions compared to the pretest declarative questions, $F(1$, $27)=6.133, M S e=45.633, p=.020, \eta_{p}^{2}=.185$. Although the mean for delayed posttest declarative questions was higher than posttest, this difference was not statistically significant, $F(1,27)=3.381, M S e=12.033, p=.077, \eta_{p}{ }^{2}=$ .111. There was not a significant interaction between test and group from posttest to delayed posttest declarative questions, $F(2,27)=2.931, M S e=10.433, p=.07, \eta_{p}{ }^{2}=$ .178. Tests of between-subjects effects determined that there was not a significant effect of group, $F<1$.


Figure 3: Mean scores of each group on declarative
knowledge. Error bars indicate standard errors of the means.

## Interleaving x Procedural Problems

In this analysis we used the 3 interleaving types as a between subjects variable and pretest, posttest, and delayed posttest as a within subject variable. This was to answer the question of whether or not the different interleaving types impacted the problem solving accuracy of pre, post, and delayed posttest problems. As is evident in Figure 1 regarding procedural problem accuracy, the test of within subjects effects indicated that there was a main effect of test, $F(2,54)=53.46, M S e=2235.54, p<.001, \eta_{p}{ }^{2}=.66$. Within subjects contrast indicated that students performed better on the posttest problems compared to the pretest problems, $F(1$, 27) $=85.19, M S e=4538.70, p<.001, \eta_{p}{ }^{2}=.76$. Although the mean for the delayed posttest problems was higher than the posttest problems, it was not statistically significant, $F(1,27)=3.89, M S e=407.01, p=.117, \eta_{p}^{2}=.089$. The interaction between test and group was approaching statistical significance, $F(1,27)=3.157, M S e=168.175, p$ $=.059, \eta_{p}{ }^{2}=.190$.

The test of between-subjects effects determined that there was a significant main effect of group, $F(1,27)=5.209$, $M S e=136.608, p=.012, \eta_{p}{ }^{2}=.278$. A Bonferroni test of multiple comparisons revealed that participants in the hyperinterleaved condition performed significantly better
from posttest to delayed posttest procedural problems when compared to the context interleaved condition, $p=.011$, whereas the math skill interleaved group did not, $p=.108$. The hyperinterleaved condition did perform better than the math skill interleaved condition on solving procedural problems from post- to delayed post test, however, these findings were not statistically significant, $p>.088$.


Figure 1: Mean scores of each group on procedural knowledge. Error bars indicate standard errors of the means.

## Interleaving x Transfer

In this analysis we used the 3 interleaving types as a between subjects variable and transfer and delayed transfer test as a within subject variable. This was to answer the question of whether or not the different interleaving types impacted the participants' problem accuracy of transfer and delayed transfer test. The test of within subjects effects did not find a main effect of test, $F(1,28)<1, M S e=1.492, p=$ $.805, \eta_{p}{ }^{2}=.040$. Although represented in Figure 2, the context and math skill interleaved groups encountered a rise in performance from transfer to delayed transfer problems, within subjects contrasts revealed that the interaction effects were not significant, $F(1,28)<1, M S e=13.981, p=.566$, $\eta_{p}{ }^{2}=.040$. While the hyperinterleaved group performed best on both transfer tests, tests of between-subjects effects determined that there was not a significant effect of group, $F(1,28)=2.25, M S e=127.801, p=.124, \eta_{p}^{2}=.138$.


Figure 2: Mean scores of each group on transfer of procedural knowledge to novel contexts. Error bars indicate standard errors of the means.

Overall, all groups demonstrated significant learning of procedural problems from pretest to posttest. In fact, there was a main effect of group on the posttest procedural
problems in which the hyperinterleaved group performed significantly better than the context interleaved group and the math skill interleaved group did not. However, the result pattern could be due to lower pretest scores for the context interleaved group.

To further examine the effect of group on assessments, change scores were computed and a one-way ANOVA was conducted. The ANOVA demonstrated that change scores from pretest to posttest of the hyperinterleaved group significantly differed from the context interleaved group, $F(2,28)=3.301, M S e=169.563, p=.05$. Bonferroni tests of multiple comparisons indicated this difference was not statistically significant, $p=.07$. Change scores from pretest to delayed posttest and from posttest to delayed posttest indicated no significant differences between groups, $p \mathrm{~s}>$ .05. Furthermore, there were no differences between groups in changes scores from transfer to delayed transfer, $p>.05$.

## Discussion

Our results demonstrate that when math skill was interleaved (i.e., in the math skill interleaved and hyperinterleaved groups), procedural performance on posttest was significantly better than when math skill was blocked (i.e., context interleaved group). These findings provide additional support for interleaved practice as a technique that enhances memory by increasing the number of repeated reloads and by promoting discriminative contrast among problems. Recall that the context interleaved group blocked the math skill problems and interleaved the contexts. Blocking math skills does not allow the learner space between problems to reload relevant information. Additionally, blocking these problems does not allow the learner to discriminate between the features of other shapes in order to highlight key elements within the problem in order to apply the appropriate procedure.

The lack of statistical differences between the math skill interleaved and the hyperinterleaved group on posttest may be due to the lack of variation between contexts. Recall that practicing problems in multiple, varying contexts typically reduces context dependency, which supports generalization to novel situations. In the current study, the contexts may not have been different enough to decrease the level of context dependency. And, a stronger effect of hyperinterleaving may have been observed if more than two contexts were included in this study. Alternatively, the spacing of practice across two separate sessions of intervention may have equalized the effects of interleaving math skill and hyperinterleaving.

The results of our experiment align with those of Rau et al. (2013; 2014), suggesting two notions. One, math skill is the problem dimension that benefits most from an interleaved practice sequence. Two, interleaving both math skill and another dimension (i.e., context) may enhance learning when compared to interleaving math skill alone. It is important to note that the hyperinterleaved group demonstrated increased learning on all post, delayed posttest, and transfer procedural problems. One explanation
for these findings may be that interleaving context while also interleaving math skill may require more effort on the part of the learner during practice, resulting in enhanced memory. Another explanation is that shuffling familiar contexts during practice may facilitate the application of knowledge outside of the context in which it was learned. This experiment provided an important contribution in understanding the benefits of more effortful interleaved practice when learning new skills and transferring them to novel contexts.

Although the fourth-grade sample size in the current study was small, the apparent trend of interleaving math skills within different contexts that led to better performance seems to be promising. The results suggest that early learning of math skills such as solving for area may benefit from all types of practice, especially when spaced over time. For future research, it may be beneficial to examine the effect of context versus no context in interleaving experiments that evaluate retention and transfer of declarative and procedural knowledge.

## Conclusion

Our experiment demonstrated the potential educational benefits of hyperinterleaving math skill with contexts. The results of the current study suggest that along with the advantages that interleaving area problems offers, shuffling contexts throughout this practice may also contribute to better generalization of these skills. Our study is one of the first to examine the effectiveness of combining interleaved practice with another common instructional technique. Placing examples in a familiar context is often used in classroom settings to make learning tasks recognizable for young learners and this study has provided unique insight on how this technique interacts with that of interleaved practice. More research is necessary to understand how interleaved practice interacts with context and other effective learning techniques, especially within the classroom environment.

## Acknowledgements

A special thanks goes to the fourth graders at Willyard Elementary, Caitlin Fortson, Jennifer Roche, Christopher Was, Tracy Arner, and Colleen Dragovich.

## References

Bjork, R.A., Allen, T.W. (1970). The spacing effect: consolidation or differential encoding? Journal of Verbal Learning and Verbal Behavior, 9 567-572.
Bjork, E.L., Bjork, R.A. (2011). Making things hard on yourself but in a good way: Creating desirable difficulties to enhance learning. In Gernsbacher, M.A., Pew, R.W., Hough, L.M., \& Pomerantz, J.R. (Eds.), Psychology and the real world: Essays illustrating fundamental contributions to society (pp. 56-64). New York, NY: Worth Publishers.
Cheng, P.W., \& Holyoak, K.J. (1985). Pragmatic reasoning schemas, Cognitive Psychology, 17, 391-416.

Cuddy, L.J., Jacoby, L.L. (1982). When forgetting helps memory: an analysis of repetition effects. Journal of Verbal Learning and Verbal Behavior, 21, 451-467.
Godden, D.R., \& Baddeley, A.D. (1975). Contextdependent memory in two natural environments: On land and under water. British Journal of Psychology, 66, 325331.

Kang, S.H.K., Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. Applied Cognitive Psychology, 26, 97-103.
Kornell, N., Bjork, R.A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? Psycholgical Science, 19, 585-592.
Rau, M.A., Aleven, V., Rummel, N. (2013). Interleaved practice in multidimensional learning tasks: Which dimension should we interleave? Learning and Instruction, 21, 98-114.
Rau, M.A, Rau, M.A., Aleven, V., Rummel, N., Pardos, Z. How should intelligent tutoring systems sequence multiple graphical representations? A mutli-methods study. International Journal of Artificial Intelligence in Education, 24, 125-161.
Rittle-Johnson, B., Star, J. (2007) Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. Journal of Educational Psychology, 99, 561-574.
Rohrer, D. (2012) Interleaving helps students distinguish among similar concepts. Educational Psychology Review, 24, 355-367.
Rohrer, D., Dedrick, R.F., Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. Psychonomics Bulletin and Review, 21, 1323-1330.
Rothkopf, E., Fisher, D., \& Billington, M. (1982). Effects of spatial context during acquisition on the recall of attributive information. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 126138.

Smith, S.M. (1982). Enhancement of recall using multiple environmental contexts during learning. Memory \& Cognition, 10, 405-412.
Sweller, J. (1988). Cognitive load during problem solving, Cognitive Science, 12, 257-285.
Theissen, E.D., \& Saffran, J.R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7-to-9 month old infants. Developmental Psychology, 39, 484-494.
Vlach, H.A., Sandhofer, C.M., \& Kornell, N. (2008). The spacing effect in children's memory and category induction. Cognition, 109, 163-167.
Vlach, H.A., Sandhofer, C.M. (2011). Developmental differences in children's context-dependent word learning. Journal of Experimental Child Psychology, 108, 394-401.
Willingham, D.T. (2009). Why Don't Students Like School? San Fransisco: Jossey-Bass.

# Scarcity impairs online detection and prospective memory 

Brandon M. Tomm (brandon.tomm@psych.ubc.ca)<br>Department of Psychology, University of British Columbia<br>2136 West Mall, Vancouver, BC V6T 1Z4, Canada<br>Jiaying Zhao (jiayingz@psych.ubc.ca)<br>Department of Psychology, and Institute for Resources, Environment and Sustainability<br>University of British Columbia, 2136 West Mall<br>Vancouver, BC V6T 1Z4, Canada


#### Abstract

Operating under limited resources poses significant demands on the cognitive system. Here we demonstrate that people under time scarcity failed to detect time-saving cues as they occur in the environment (Experiment 1a). These time-saving cues, if noticed, would have saved time for the time-poor participants. Moreover, the visuospatial proximity of the time-saving cues to the focal task determined successful detection, suggesting that scarcity altered the spatial scope of attention (Experiment 1b \& 1c). People under time scarcity were also more likely to forget previous instructions to execute future actions (Experiment 2). These instructions, if remembered and followed, would have saved time for the time-poor participants. Failures of online detection and prospective memory are problematic because they cause neglect and forgetting of beneficial information, perpetuating the condition of scarcity. The current study provides a new cognitive account for the counterproductive behaviors in the poor, and relevant implications for interventions.


Keywords: scarcity, attention, perception, memory, recall

## Introduction

The condition of scarcity is widespread and manifests in many domains. For example, four billion people experience severe water scarcity during at least part of each year (Mekonnen \& Hoekstra, 2016), and more than $10 \%$ of the world population live with less than US\$1.90 per day (World Bank, 2016). A growing body of evidence has revealed how scarcity fundamentally shapes the way people perceive the environment and behave accordingly (Mani, Mullainathan, Shafir, \& Zhao, 2013; Mullainathan \& Shafir, 2013; Shah, Mullainathan, \& Shafir, 2012; Shah, Shafir, \& Mullainathan, 2015; Tomm \& Zhao, 2016).

Since the cognitive system is limited in attentional and working memory capacity (Baddeley, 1992; Luck \& Vogel, 1997; Miller, 1956; Pashler, Johnston, \& Ruthruff, 2001; Rock \& Gutman, 1981), scarcity induces a trade-off of attentional and cognitive resources dedicated on the focal task and other tasks that also require attention (Tomm \& Zhao, 2016). This corroborates with past research showing that engagement with complex tasks can cause a failure to notice highly salient events (Simons \& Chabris, 1999), even at the expense of personal safety (see Strayer, Drews, \& Johnson, 2003).

In the current study, we investigate how time scarcity affects the online detection of information, and how time
scarcity affects prospective memory performance. Our study is motivated by past work showing that people only start to increase their efforts to accomplish their goals when a deadline becomes salient (Gersick, 1988). Further, time pressure causes fewer attributes to be considered when choosing between alternatives (Wright, 1974). Given these findings, we propose that time scarcity may enhance attentional focus on the task at hand, while inducing neglect of other information in the environment, even if the information is beneficial. In two experiments, we examined the attentional and memory consequences of time scarcity.

## Experiment 1a

The goal of the first experiment was to investigate how time scarcity affects the online detection of information in the environment. We hypothesize that time scarcity draws attention to the focal task, while inducing neglect of other useful information in the environment.

## Participants

Undergraduate students ( $N=90$ ) were recruited from the Human Subject Pool at the Department of Psychology at the University of British Columbia (UBC), and participated in the experiment in exchange for course credit. All participants provided informed consent to participate. All experiments reported here were approved by the UBC Behavioral Research Ethics Board. We conducted a power analysis using G*Power (Faul, Erdfelder, Lang, \& Buchner, 2007), which showed that given an effect size of 0.6 (based on our prior work, Tomm \& Zhao, 2016), a minimum of 90 participants would be required to have $80 \%$ power to detect the effect in our design.

## Stimuli and Procedure

In the experiment, each participant was asked to solve a series of puzzles on the computer. The puzzles were a total of 50 trials of the Raven's Progressive Matrices (Raven, 2000). Each matrix appeared at the centre of a computer screen. The bottom right corner of the matrix was missing, and participants had to find the right piece that fits with the general pattern in the matrix. Each participant was asked to correctly solve the matrices to earn as many points as possible. In each trial, participants were shown one Raven's
matrix, with the numbered pieces appearing below. The response keys appeared in a vertical list on the left side of the screen. In the top-left corner of the screen, the question number and time remaining were displayed (see Figure 1). Participants were not told of the total number of trials until starting the first trial. To solve each matrix, participants pressed a number key corresponding to the correct piece.


Figure 1. Trial screen for Experiment 1a.
To manipulate time scarcity, participants were randomly assigned with either a rich time budget (they had 40 minutes in total to solve the matrices; the time-rich condition, $N=$ 45 ), or a poor time budget (they only had 10 minutes in total to solve the matrices; the time-poor condition, $N=45$ ). Without explicit instruction or prompting, a time-saving cue appeared in the lower right part of the screen during the experiment. Specifically, on even-numbered trials starting from trial \#24, the cue appeared on the screen stating: "This question is not worth any points. Press 'A' to skip." (see Figure 1) Thus, 14 of the 50 trials were allowed to be skipped without any loss of points. The cue appeared at the same as the matrix for those trials, and remained on the screen for 5000 ms , and then disappeared. These trials presented an opportunity to skip the question in order to save time. Participants were not told anything about the cue. We wanted to see if they were able to detect this message during the experiment and skipped the even-numbered questions from trial \#24.

## Results and Discussion

Participants in the time-poor condition almost unanimously used their entire time budget ( 10 minutes) while participants in the time-rich condition used less than half of their time budget ( 16 minutes). Given this constraint, the time-poor participants spent less time on the task overall compared to time-rich participants $[t(88)=6.51, p<.001, d=1.37]$ (Figure 2a). The time-poor participants completed fewer trials than the time-rich participants $[t(88)=4.71, p<.001$, $d=.99]$ (Figure 2b).

Notably, there was marginal difference in accuracy on the Raven's Progressive Matrices between the time-poor and the time-rich participants $[t(88)=1.69, p=.09, d=.36]$
(Figure 2c). When accounting for the total amount of time spent on the task, the time-poor participants scored higher accuracy per minute than time-rich participants $[t(88)=$ 8.09, $p<.001, d=1.71]$. This result suggests that time scarcity can cause a greater focus on the task at hand, enhancing task performance within the time limit.


Figure 2. Results for Experiment 1a. Error bars represent $\pm 1$ SEM. $* p<.05$, $* * * p<.001$. Note: accuracy was computed for all trials excluding skipped trials.

Examining the number of questions skipped, we found that there was no significant difference in the average number of questions skipped between the time-poor and the time-rich participants $[t(88)=1.23, p=.22, d=.26]$ (Figure 2d). However, only $26.7 \%$ of the participants in the timepoor condition skipped at least once, and there were more time-rich participants (48.9\%) who skipped at least once $\left[X^{2}(1,90)=4.72, p=.03\right]$ (Figure 2e). This result suggests that time scarcity caused a failure to use the time-saving cue appearing on the bottom of the screen.

To control for the total number of trials completed, we calculated skip efficiency as the number of questions skipped divided by the number of possible questions that could be skipped. There was no difference in skip efficiency between the time-poor and the time-rich participants $[t(88)$ $=.91, p=.36, d=.19]$ (Figure 2f).

Among those who skipped at least once, there was no difference in the number of questions skipped between the time-poor and the time-rich participants $[t(31)=.89, p=$ $.38, d=.34$ ] (Figure 2 g ). This means that if the participant noticed the cue at least once, they were able to skip the same number of questions, regardless of scarcity.

To measure retrospective recall of the time-saving cues, we asked participants after completing the task during debriefing to report whether they saw any messages appearing on the screen during the task. We found that the time-poor participants were less likely to report seeing the cues than the time-rich participants $\left[X^{2}(1,84)=3.81, p=\right.$ .05] (Figure 2h).

These results showed that fewer participants under time scarcity skipped the questions at least once, and reported seeing the cues, compared to time-rich participants. This suggests that time scarcity may narrow attention to the central task, while inducing a neglect of peripheral, even beneficial information in the environment. An alternative explanation is inattentional blindness, suggesting that the time-poor participants were less able to attend to salient but task-irrelevant information, than the time-rich participants. To tease these two accounts apart, we conducted the next experiment, probing whether scarcity alters the spatial scope of attention, or the ability to notice salient stimulus. Specifically, we manipulated the location of the time-saving cue, and examined the likelihood of skipping questions as a function of the spatial location of the cue under scarcity.

## Experiment 1b

In this experiment, we reduced the spatial distance between the time-saving cue and the matrix (i.e., the focal task) by moving the cue closer to the center of the screen, and investigated how the spatial proximity of the time-saving cue to the focal task impacted its detection.

## Participants, Stimuli, and Procedure

Participants $(N=87)$ were recruited from the Human Subject Pool at UBC, and participated in the experiment in exchange for course credit. The stimuli and the procedure
were exactly the same as those in Experiment 1a, except one important change: the time-saving cue (i.e., the message to skip even-numbered questions after trial \#24) now appeared directly underneath the Raven's Progressive Matrix after trial \#24 for even-numbered questions (Figure 3).

If the neglect of the time-saving cue in Experiment 1a was due to the spatial narrowing of attention under scarcity, we would predict that the time-poor participants would be more likely to notice the cue, because it was not close to the central task. On the other hand, if the neglect of the timesaving cue was due to inattentional blindness, moving the cue closer to the central task would not affect performance.


Figure 3. Trial screen for Experiment 1b, where the time-saving cue appeared right below the matrix.

## Results and Discussion

Since in Experiment 1a, time scarcity influenced the number of participants who skipped at least once, we examined the same measure here again. We found that now there was no statistical difference in the percent of participants who skipped at least once $\left[X^{2}(1,90)=.71, p=\right.$ .40] (Figure 4a). Comparing Figure 4 a to Figure 2e, the time-rich participants were not influenced by the change in the position of the cue, but the poor seemed to benefit from the closer proximity of the cue to the central task. This suggests that if the cue falls within the spatial scope of attention, the time-poor participants could still take advantage of the cue.


Figure 4. Results for Experiment 1b.

During debriefing, the time-poor participants were marginally less likely to report seeing any messages during the task compared to the time-rich participants $\left[X^{2}(1,88)=\right.$ 3.78, $p=.05$ ] (Figure 4b). Compared to the time-poor participants in Experiment 1a ( $34 \%$ reported noticing the cue), the closer proximity seemed to provide a large benefit to the time-poor participants in Experiment 1b ( $48 \%$ reported noticing the cue). These results support the account that scarcity narrows spatial attention to the focal task.

## Experiment 1c

To further explore the boundary condition of the spatial narrowing effect of scarcity, in this experiment we moved the time-saving cue farther away from the focal task, and examined how likely participants were to notice the cue.

## Participants, Stimuli, and Procedure

Participants $(N=86)$ were recruited from the Human Subject Pool at UBC, and participated in the experiment in exchange for course credit. The stimuli and the procedure were identical to those of Experiment 1a, but this time the time-saving cue appeared in the bottom right corner of the screen (Figure 5), which was even farther away from the focal task than in Experiment 1a.


Figure 5. Trial screen for Experiment 1c, where the time-saving cue appeared far from the matrix, on the bottom right corner of the screen.

## Results and Discussion

We found that participants in both conditions failed to take advantage of cue. There was no statistical difference in the percent of participants who skipped at least once $\left[X^{2}(1,90)=\right.$ $1.54, p=.21]$ (Figure 6a). During debriefing, there was no difference in the likelihood to report seeing any messages during the task between the participants in both conditions $\left[X^{2}(1,87)=2.70, p=.10\right]$ (Figure 6b). In fact, there was a floor effect in both the time-poor and the time-rich participants in skipping the questions or noticing the cue. This suggests that when the cue was spatially far away from the focal task, participants could not notice the cue, regardless of scarcity.


Figure 6. Results for Experiment 1c.

## Experiment 2

Experiments 1a-c showed that time scarcity narrowed attention on the focal task, resulting in the neglect of a timesaving cue which appeared in the peripheral during the experiment. However, in daily life, we do not always have cues in the external environment as reminders for certain actions. Instead, we need to rely on internal cues from memory that need to be activated at the right time to direct actions. For example, in order to pick up groceries on the way home from work, we must remember to turn at the right intersection in order to go to the grocery store. This depends on prospective memory, which is the ability to remember to execute future actions based on previous instructions. Cues for prospective memory are internal, and must be present in mind in order to cue behavior at the right time (Graf, Uttl, \& Dixon, 2002; Loftus, 1971). In this experiment, we examined how time scarcity affects prospective memory performance.

## Participants

Participants $(N=90)$ were recruited from the Human Subject Pool at UBC and completed the study in exchange for course credit.

## Stimuli and Procedure

Participants were asked to solve the same set of 50 Raven's Progressive Matrices used in Experiments 1a-c. As before, participants were randomly assigned either a small time budget ( 5 minutes; the time-poor condition), or a large time budget ( 20 minutes; the time-rich condition). A critical difference in this experiment was that the time-saving cue never appeared in the experiment. Rather, all participants were explicitly instructed at the start of the experiment the following: "Even-numbered questions from number twentyfour on are not worth any points. You can skip these questions without losing any points." This instruction was presented on paper to participants to read, and the experimenter also read through these instructions with each participant to maximize the comprehension of the instruction. As before, the question number and remaining time appeared in the top-left corner of the screen, and the keys available for the participants to press were listed on the
left side of the screen. Note that now the "A (skip)" key is listed among the available keys and was listed for every single question (Figure 7). There were no visual cues during the experiment to remind participants which questions they were allowed to skip. Thus, participants needed to remember to use the opportunity to skip when the applicable questions were reached.


Figure 7. Trial screen for Experiment 2.

## Results and Discussion

Participants in the time-poor condition almost unanimously exhausted their time budgets, while participants in the timerich condition usually completed the experiment with some time to spare (Figure 8a). The time-poor participants spent less time solving the Raven's Matrices than the time-rich participants $[t(88)=13.33, p<.001, d=2.81]$. They also completed significantly fewer trials than the time-rich participants $[t(88)=10.14, p<.001, d=2.14]$ (Figure 8b), and were significantly less accurate $[t(88)=2.29, p=.02, d$ $=.48]$ (Figure 8c). When accounting for the total amount of time spent on the task, the time-poor participants scored higher accuracy per minute than time-rich participants [ $t(88)$ $=9.53, p<.001, d=2.01]$, suggesting that time scarcity enhancing performance on the focal task.

The time-poor participants on average skipped fewer questions than the time-rich participants $[t(88)=2.52, p=$ $.01, d=.53$ ] (Figure 8d). However, this result is likely driven, at least in part, by the considerably smaller number of questions completed by the time-poor participants. Similarly, we found that fewer time-poor participants skipped at least once compared to the time-rich participants $\left[X^{2}(1,90)=10.08, p<.01\right]$ (Figure 8e), but this could be due to the smaller number of possible skips experienced by the time-poor participants. Thus, we examined the skip efficiency defined as the number of questions skipped divided by the number of possible questions that could be skipped experienced by the participant. We found that the time-poor participants were less likely to skip than time-rich participants (two time-poor participants were excluded from this analysis due to failing to reach trial number twentyfour) $[t(86)=2.01, p=.05, d=.43]$ (Figure 8f). This finding suggests that time scarcity impairs prospective
memory performance. We should note that among participants who skipped at least once, there was no difference in the number of questions skipped between the time-poor and the time-rich participants $[t(40)=.59, p=$ $.56, d=.19]$ (Figure 8 g ), or in skip efficiency $[t(40)=.76, p$ $=.45, d=.26]$ (Figure 8h).


Figure 8. Results for Experiment 2. Error bars represent $\pm 1$ SEM. *p<.05, **p $<.01, * * * p<.001$.

## General Discussion

The goal of the present study was to examine how time scarcity impacts attention and prospective memory. We found that people under time scarcity were less likely to take advantage of a time-saving cue that appeared peripheral to the focal task (Experiment 1a), but nonetheless performed well on the focal task under the time constraint. This suggests that people under time scarcity are ironically less likely to notice opportunities to save time. This effect could be explained by a narrowing of spatial attention to the focal task (Experiments $1 \mathrm{~b} \& 1 \mathrm{c}$ ). In the absence of an external cue, participants under time scarcity were less likely to remember to skip questions in the future (Experiment 2), suggesting that they failed to retrieve a cue from memory to execute actions at the right time.

These findings were particularly problematic for people under time scarcity because the attentional neglect of timesaving opportunities or the failure to remember to save time could be detrimental, perpetuating the condition of scarcity and creating a vicious cycle of scarcity. These cognitive impairments could explain a range of counter-productive behaviors observed in the low-income individuals, such as forgetting to follow instructions, or not signing up for public benefit programs. In addition, prospective memory errors can be seen by others as an indication of incompetence of the poor (Graf, 2012). The present findings instead attribute the memory failures not to the poor individuals themselves but to the condition of scarcity. The current study provides useful implications for designing policies and programs to mitigate the impact of scarcity, such as the use of reminders, automatic enrolment, or setting the right default, to reduce the attentional and memory burdens in the poor.

## Acknowledgements

We thank Ivy Tam, Justin Reed, Sherry Guo, Mark Ng Shun, Lauren Duda, and Michelle Lee for assistance with data collection. We also thank Alan Kingstone, Peter Graf, Rebecca Todd, and the members in the Zhao Lab for helpful comments. This work was supported by Sloan Foundation Grant (2014-6-16), the Canada Research Chairs program (to JZ), and the Leaders Opportunity Fund from the Canadian Foundation for Innovation (F14-05370 to JZ).

## References

Baddeley, A. D., \& Hitch, G. J. (1974). Working memory. The Psychology of Learning and Motivation, 8, 47-89.
Faul, F., Erdfelder, E., Lang, A. G., \& Buchner, A. (2007). $\mathrm{G}^{*}$ Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191.
Gersick, C. J. (1988). Time and transition in work teams: Toward a new model of group development. Academy of Management Journal, 31(1), 9-41.
Graf, P. (2012). Prospective memory: Faulty brain, flaky person. Canadian Psychology/Psychologie Canadienne, 53(1), 7.

Graf, P., Uttl, B., \& Dixon, R. (2002). Prospective and retrospective memory in adulthood. Lifespan Development of Human Memory, 257-282.
Loftus, E. F. (1971). Memory for intentions: The effect of presence of a cue and interpolated activity. Psychonomic Science, 23(4), 315-316.
Luck, S. J., \& Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390, 279-281.
Mani, A., Mullainathan, S., Shafir, E., \& Zhao, J. (2013). Poverty impedes cognitive function. Science, 341(6149), 976-980.
Mekonnen, M. M., \& Hoekstra, A. Y. (2016). Four billion people facing severe water scarcity. Science Advances, 2(2), e1500323.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81.
Mullainathan, S., \& Shafir, E. (2013). Scarcity: Why having too little means so much. Henry Holt and Company, New York.
Pashler, H., Johnston, J. C., \& Ruthruff, E. (2001). Attention and performance. Annual Review of Psychology, 52, 629-651.
Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. Cognitive Psychology, 41(1), 1-48.
Rock, I., \& Gutman, D. (1981). The effect of inattention on form perception. Journal of Experimental Psychology: Human Perception and Performance, 7, 275.
Shah, A. K., Mullainathan, S., \& Shafir, E. (2012). Some consequences of having too little. Science, 338(6107), 682-685.
Shah, A. K., Shafir, E., \& Mullainathan, S. (2015). Scarcity frames value. Psychological Science, 26(4), 402-412.
Simons, D. J., \& Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception, 28, 1059-1074.
Strayer, D. L., Drews, F. A., \& Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. Journal of Experimental Psychology: Applied, 9, 23-32.
Tomm, B. M., \& Zhao, J. (2016). Scarcity captures attention and induces neglect: Eyetracking and behavioral evidence. In A. Papafragou, D. Grodner, D. Mirman, \& J.C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 11991204). Austin, TX: Cognitive Science Society.

World Bank (2016). Poverty and Shared Prosperity 2016: Taking on Inequality. Washington, DC: World Bank. doi:10.1596/978-1-4648-0958-3.
Wright, P. (1974). The harassed decision maker: Time pressures, distractions, and the use of evidence. Journal of Applied Psychology, 59(5), 555-561.

# Specificity and entropy reduction in situated referential processing 

Elli N. Tourtouri (elli@coli.uni-saarland.de)<br>Francesca Delogu (delogu@coli.uni-saarland.de)<br>Matthew W. Crocker (crocker@coli.uni-saarland.de)<br>Department of Language Science \& Technology<br>Saarland University, Saarbrücken, Germany


#### Abstract

In situated communication, reference to an entity in the shared visual context can be established using either an expression that conveys precise (minimally specified) or redundant (over-specified) information. There is, however, a long-lasting debate in psycholinguistics concerning whether the latter hinders referential processing. We present evidence from an eye tracking experiment recording fixations as well as the Index of Cognitive Activity - a novel measure of cognitive workload - supporting the view that overspecifications facilitate processing. We further present original evidence that, above and beyond the effect of specificity, referring expressions that uniformly reduce referential entropy also benefit processing.


Keywords: referential processing; over-specification; visual entropy reduction; eye tracking; Index of Cognitive Activity

## Introduction

Grice's maxims of Quantity (Grice, 1975) stipulate that speakers' utterances be minimally informative, avoiding redundancy. In visually situated communication, this predicts utterances should provide strictly the information necessary for the identification of a referenced object. For example, in the context of a blue and a green ball, the adjective "blue" is necessary to unambiguously establish reference. When there is only one ball, however, the adjective becomes superfluous. Such over-specifications expressions that convey more information than minimally required - are, however, produced by adult speakers at an estimated rate of $10-60 \%$ (see Engelhardt, Bailey \& Ferreira, 2006, and references therein).

Even though Grice arguably did not intend to make any implications about the cognitive processes associated with the violation of his maxims (cf. Geurts \& Rubio-Fernández, 2015), over the past few decades, psycholinguistic research has tried to test their empirical validity. It remains under debate, however, whether or not over-specifications are detrimental to referential processing. A number of studies have suggested that over-specifications impair listeners' online processing and lead to slower and less accurate identification of the target (e.g., Engelhardt, Bailey \& Ferreira, 2006; Engelhardt, Demiral \& Ferreira, 2011; Davies \& Katsos, 2013), while others find evidence that they are as good as minimal descriptions or may even facilitate processing (e.g., Arts, Maes, Noordman \& Jansen, 2011; Tourtouri, Delogu \& Crocker, 2015).

In an ERP experiment, Tourtouri, Delogu and Crocker (2015) presented participants with visual scenes of 6 objects
and audio instructions to locate a target, like "Find the yellow bowl" (in German). The experiment manipulated the specificity of the referring expression by combining the same instruction with different visual displays that rendered it minimally or over-specified. An attenuated N400 effect was found on the noun for over- compared to minimallyspecified references. This finding was interpreted as evidence that over-specifications are in fact beneficial to referential processing, at least when in the presence of visual displays where the over-specified adjective identified exactly one object. That is, at "yellow" the bowl was the only object that fit the description. Interestingly, both color and pattern adjectives were used to identify the target, and the effect was present for both types of adjectives, suggesting that any facilitation of over-specification is not merely due to the perceptual salience of color. It can be argued, however, that the reduced N 400 for overspecifications may just reflect the predictability of the noun as determined by the information on the visual scene in combination with the linguistic input up to the adjective. Therefore, it still remains unanswered whether overspecification has a general facilitatory effect, even when displays allow the adjective to select a second object, which fits a minimally specified continuation of the referring expression, i.e., it is part of a contrast pair.

A similar question was addressed by Sedivy, Tanenhaus, Chambers and Carlson (1999) in a series of experiments that tested (among other things) whether intersective adjectives such as color are interpreted contrastively. Participants' eye movements were tracked while they heard instructions to manipulate objects in a workspace in front of them. The visual scenes consisted of four objects, two that formed a contrast pair differing in color, e.g. a yellow and a pink comb, and two singletons: one sharing the color feature with an object from the pair, e.g. a yellow bowl, and a distractor object of different color. The critical instruction mentioned either of the two objects with the shared feature, and was always heard second, following the instruction referring to an object from the contrast pair. An effect of referent type was found, such that if the target was part of the contrast pair it was looked at faster than if it was not. This result was taken to indicate that initially, before the noun was heard, listeners assigned a contrastive meaning to intersective adjectives, consistent with Grice's maxim of Quantity. We believe, however, that this may not be the case, especially since listeners' attention was already focused on the contrast pair, as the immediately preceding instruction always made reference to one of the contrasting objects (e.g., the pink


Figure 1. Sample visual stimuli for a color item, combined with the instruction "Find the blue ball". The four resulting conditions were A. Minimally specified - High reduction (MS-HR), B. Minimally specified - Low reduction (MS-LR), C. Over-specified - High reduction (OS-HR), D. Over-specified - Low reduction (OS-LR)
comb). Similar results were obtained in the subsequent experiments, where the critical instruction was heard first, but with the use of scalar adjectives, which inherently invoke comparisons between entities.

The current study seeks to determine whether and how over-specifications affect processing of pre-nominally modified referring expressions, when the visual context enables both a minimally and an over-specified reading of intersective adjectives, such as color and pattern. That is, how is referential processing influenced when the adjective is redundant (as in the bottom displays of Fig.1) as opposed to when it is required to uniquely identify the target (as in the top displays of Fig.1)? Furthermore, as the instruction sentence unfolds over time, incoming words incrementally restrict the set of referential candidates. Therefore, in situated communication, the information conveyed by a linguistic unit is determined by the extent to which it reduces the number of potential referents, in addition to the linguistic information of each word, as determined by its probability and preceding context (Shannon, 1948; Crocker, Demberg \& Teich, 2016). In other words, the information on the word "blue" in the sentence "Find the blue ball" is not defined only in terms of its probability to occur in this (linguistic and visual) context, but also by the amount of uncertainty about the target (referential entropy) it reduces. For example, in the left-hand displays of Figure 1 "blue" reduces referential entropy by 1.58 bits, while in the righthand displays it only reduces it by 0.58 bits. The noun, then, eliminates the remaining entropy, reducing it by 1 bit in the
former and by 2 bits in the latter case, resulting in a uniform and a less uniform entropy reduction profile, respectively. This study also touches on whether, above and beyond any effects of specificity, the rate at which the linguistic input reduces referential entropy also influences processing. To examine these questions, we recorded participants' fixations as they viewed displays such as the ones in Figure 1, while listening to instructions like "Find the blue ball" in German, and present results from inspection probabilities to the objects of interest and the Index of Cognitive Activity (Marshall, 2000) per region.

## The Index of Cognitive Activity

It is well established that fluctuations of the pupil size index cognitive effort in a variety of tasks, including language processing (e.g., Just \& Carpenter, 1993). However, changes in the lighting conditions of the environment are also responsible for pupil dilation. The Index of Cognitive Activity (Marshall, 2000) is a measure of cognitive workload that separates variation in pupil size due to cognitive effort and due to light reflex, while also accounting for random noise. The small and rapid pupil dilations that remain are associated with higher cognitive workload (Marshall, 2002). Demberg and Sayeed (2016) showed, for example, that the ICA is sensitive to linguistic manipulations such as ungrammaticality, with conditions related to higher processing demands resulting in higher ICA values. They also demonstrated that ICA is particularly suitable for the Visual World Paradigm, since it is robust to
the change of fixation positions and can thus complement the standard visual attention metrics in order to assess cognitive effort during linguistic processing.

## Experiment

We used a $2 \times 2$ design crossing Specificity (Minimally specified vs Over-specified) and Entropy Reduction (Uniform vs Non-uniform). Based on findings that overspecifications are commonly used by adult speakers during production (cf. Engelhardt et al., 2006; Pechmann, 1989; Rubio-Fernández, 2016; Tarenskeen, Broersma \& Geurts 2015), we hypothesized that over-specification would not impede referential processing, as rational speakers would unlikely use them so frequently if they did. We, therefore, expected that over-specified expressions (OS) would be as easy as, or easier than their minimally-specified (MS) counterparts (as found in the ERP study by Tourtouri et al., 2015). As for the entropy reduction manipulation, we generally expected a greater processing advantage in the uniform reduction (UR) compared to the non-uniform reduction (NR) conditions, as has been proposed for the related measure of surprisal (UID, Jaeger, 2010). Finally, we expected that the two factors should interact, namely that processing would be particularly benefited when the expression was OS and the redundant adjective contributed to the uniform reduction of entropy.

## Method

Participants Twenty-four students from Saarland University (mean age 25, 7 male) participated in the experiment for monetary compensation. They were all native speakers of German with normal or corrected-tonormal vision and normal color perception.

Materials Pictures of 30 common use objects (e.g., balls, mugs, etc.) differing in color (blue, green and red) and pattern (checkered, dotted and striped) were employed to create the visual stimuli. Both color and pattern were used as distinguishing features, because they are intrinsic to the object, as opposed to scalar adjectives such as size that trigger comparisons to other entities on the display. This ensured that any looks to objects in contrast pairs would be driven due to the manipulation and not because of the adjective type. Furthermore, pattern was the mentioned property in half of the trials, in order to make sure that any effect of over-specification would not be merely due to color salience, but would be attributable to the experimental manipulation. Color hue and brightness were adjusted using GIMP (Version 2.8.10). Naming agreement was tested for the object pictures in an offline picture naming study to ensure that they were identifiable in all colors and patterns, and that the names used in the experiment matched participants' own naming preferences. Twenty-four independent participants were presented with the object images in all colors and patterns (distributed over 8 lists), and were asked to name them while always mentioning their
colors and patterns. Only objects with a naming agreement of $80 \%$ or higher were employed in the visual stimuli.

A set of 120 experimental items was created, each item comprising one spoken instruction (with either color or pattern as the target feature) and four visual scenes (essentially four versions of the same scene). The target color, pattern and position were counterbalanced throughout the experiment. Displays for experimental items accommodated all four conditions for both target features, so that nothing would reveal the target before the instruction was heard. To this end, one visual scene contained 6 objects (two pairs of same-type objects and two singletons) in two colors and two patterns, such that the pairs made up the two MS and the singletons the two OS referents for both target features, as shown in Figure 1. Furthermore, displays never
 ensuring that disambiguation of the target would always occur on noun onset. For the same reason only same-gender objects were used per display, as German marks determiners for gender.

In total, 660 visual displays were created, of which 480 were used in experimental items ( $120 \times 4$ conditions), and 180 in fillers. Twelve of the fillers served as practice trials in a familiarization phase. Fillers differed from experimental items in multiple aspects. First, they differed in terms of their display structure, with almost half of the fillers depicting 4 objects ( 3 of the same type and one singleton) and the rest containing 6 objects. Six-object fillers either showed 2 contrast pairs and 2 singletons, where reference was always made to the contrast pair that was not relevant in the experimental items (e.g., the two rucksacks in Fig.1), or they showed a set of 3 same-type objects, a contrast pair and a singleton. The 3 -object sets made a second modifier also required for target identification, thus adding more variation not only to the display types that participants viewed, but also to the referential entropy reduction possibilities. Secondly, fillers differed in terms of the specificity of their instructions, that could be minimally, over-, or under-specified (US), while care was taken so that throughout the entire experiment, participants would hear MS expressions to a greater extent than OS - as is the case in everyday language use - as well as a small portion of US. A set of fillers without pre-nominal modification was also used, that were essentially the minimally specified versions of the OS items, thus assuring that participants would not always expect to hear an adjective and that they would not get overly used to reference being redundant.

In experimental items, displays were paired with audio instructions containing a pre-nominally modified referring expression like "Finde den blauen Ball" (Find the blue ball) in Figure 1, that identified the target by mentioning either its color or its pattern. In fillers, instructions had zero, one or two modifiers. For the latter the order of mention of color and pattern adjectives was counterbalanced. Audio stimuli were recorded with neutral intonation by a young, female speaker of German, in a soundproof booth using Cubase AI5. Speech was continuous and no artificial pauses were
inserted in between words. Sentences were then cut and annotated for adjective and noun onsets using Praat (Version 5.3). Mean word duration was 481.3 ms ( $\mathrm{SD}=32$ ) for the adjectives and $557.2 \mathrm{~ms}(\mathrm{SD}=75.7)$ for the nouns.

Stimuli were distributed over 4 lists using the Latin Square design, and were pseudo-randomized for each participant. At least one filler appeared between consecutive experimental items, and items of the same condition did not appear more than two times in a row. Each participant saw 288 stimuli split in 4 blocks, which allowed for breaks in between blocks. Before the experiment started, a short practice session of 12 filler trials familiarized participants with the task. The experiment was implemented and run using E-prime 2.0 (Psychology Software Tools, Inc.).

Procedure An SMI RED500 eye tracker (SensoMotoric Instruments) attached to the bottom of a 25 inch Dell monitor was used to track participants' eye movements at a rate of 250 Hz . After they gave informed consent and read the instructions, participants were seated at a distance of approximately 60 cm in front of the monitor using a chinrest to minimize head movements. They then completed a familiarization phase, during which the experimenter gave them feedback after each trial, ensuring that the task was clear before the experiment begun. Calibration was performed at the beginning of each block.

A trial started with a fixation cross appearing in the middle of the display for a period controlled by the experimenter. The objects then appeared while the cross was still on screen for another 500 ms , and 1500 ms later the audio instruction started. The objects stayed on the screen for another 500 ms after the audio offset, and a prompt screen to the task appeared asking participants to indicate which side of the screen the target entity was on, or whether it was not possible to tell (US fillers) by pressing the corresponding button on a response pad in front of them. Displays were presented at a $1680 \times 1050$ resolution. One experimental session lasted on average 40 min , depending on whether calibration had to be repeated.

Analysis For the analyses of both measures we considered the regions of the Adjective ("blauen"), and Noun ("Ball"). For the analysis of fixations, we compared inspection probabilities to areas of interest (AOI) across conditions. First, fixations shorter than 80 ms were pooled with the immediately preceding or following fixation, if the distance between them was smaller than 12 pixels, otherwise they were excluded from the analysis. Subsequently, fixations to an AOI within a region, before a saccade outside the area was made, were counted as one inspection. For each AOI and region, we coded trials that contained at least one inspection to the AOI as 1 , and trials that did not as 0 . Therefore, mean values represent inspection probabilities per AOI and region.

As information about the target became incrementally available, different objects and different comparisons were interesting per region. In particular, at the Adjective, the
only available information about the target was its distinguishing property, so the specificity manipulation is still irrelevant (it is still unknown whether the target is minimally or over-specified). We, therefore, compared inspections to the singleton and contrasting objects that bore the target property (cf. the blue ball \& mitten in A\&B, Fig.1) between UR and NR (collapsing across Specificity). Finally, at the Noun, when the target is revealed, both factors become relevant, so we contrasted inspection probabilities to the target (the blue ball: MS in A\&B, OS in C\&D, Fig.1) and to the competitor (the blue mitten: OS in A\&B, MS in C\&D, Fig.1) across conditions.

To calculate the ICA we used BeGaze ${ }^{\text {TM }}$ with the ICA Module (SensoMotoric Instruments) and Workload ${ }^{\text {RT }}$ (EyeTracking, Inc.). Since the ICA values that the BeGaze ${ }^{\mathrm{TM}}$ software outputs are too coarse-grained for the type of effects we expect, we used the ICA Coefficients to compute ICA values per 100 ms (see Demberg \& Sayeed, 2016 for more details). Data points with a pupil diameter smaller than 2.5 SD of that participant were eliminated, and a mean ICA value for both eyes was calculated. As fixation positions are not relevant for the ICA, we were interested only in differences between UR and NR (collapsing across Specificity) for the Adjective, and across conditions for the Noun. We compared mean ICA values across conditions within a window of 600 ms starting from the middle of each region. We analyzed inspection probabilities and ICA values using generalized linear mixed effects models (lme4 package, R Version 3.3.2) with random intercepts for participants and items, as well as random slopes for the predictors of interest. For the analysis of the Adjective, Reduction (UR vs NR) was the predictor of interest. For the analysis of the Noun, the models included the effects of Specificity (OS vs MS), Reduction (UR vs NR), and Target Feature (Color vs Pattern), and their interaction. When the maximal models did not converge, we simplified the random effects structure as suggested by Barr, Levy, Scheepers, and Tily (2013).

## Results

## Adjective

Singletons bearing the target feature (cf. the mitten in A\&B, and ball in C\&D) were inspected equally frequently in UR and NR (Coeff. $=.083, S E=2.317, Z=.829, p>.05)$. Contrast objects (cf. the blue ball in A\&B), on the other hand, were more frequently inspected in UR than in NR (Coeff. $=.329, S E=.0628, Z=3.107, p=.001) .{ }^{1}$ The ICA values (see Fig.2) did not differ significantly between UR and NR (Coeff. $=-.031, S E=.0249, Z=-1.236, p>.05$ ).

[^521]

Figure 2. Mean ICA values per condition and region. Error bars represent $95 \%$ CI.

## Noun

Analyses of inspection probabilities to the target (the blue ball) and competitor (the blue mitten) objects, including target feature as a predictor, produced a significant effect of feature (Coeff. $=.205, S E=.104, Z=1.971, p=.048$ ). We followed up this effect with separate analyses for color and pattern items. For target inspections in color items (Fig.3), we found a main effect of Entropy Reduction with more inspections in UR than in NR (Coeff. $=-.241, S E=.122, Z$ $=-1.971, p=.048$ ), as well as a marginally significant effect of Specificity (Coeff. $=.237, S E=.127, Z=1.877, p=.06$ ), such that the target was inspected more frequently in OS than in MS. The analysis of inspections to the competitor resulted in a main effect of Entropy Reduction with the competitor receiving more inspections in UR than in NR (Coeff. $=.39, S E=.151, Z=2.583, p=.009$ ), and no effect of Specificity ( $p>.05$ ). For pattern items, none of the comparisons produced significant results (all $p>.228$ ). Interestingly, the ICA analysis produced main effects of both Entropy Reduction and Specificity for both color and pattern items (see Fig.2), such that ICA was higher for NR vs UR (Coeff. $=-.07, S E=.024, Z=-2.96, p=.003$ ), and for MS vs OS (Coeff. $=.087, S E=.025, Z=3.42, p<.001)$.

## General Discussion

We investigated the effects of Specificity on situated language processing comparing listeners' inspection patterns and cognitive effort when exposed to minimally or over-specified reference. In accordance with previous research (cf. Arts et al., 2011; Tourtouri et al., 2015) we found a facilitation for OS vs MS on the noun, with the target object receiving more inspections when the referring expression included a redundant vs a contrastive adjective. However, this effect was observed only for color items, raising the question whether what facilitates processing is in fact color salience as opposed to over-specificity in general. The answer is provided by ICA, a novel measure of cognitive workload based on the count of rapid pupil dilations, which we used to directly assess the cognitive effort expended in each condition, showing that in both
color and pattern items OS was indeed easier to process than MS. This discrepancy between the two measures seems to suggest that, while pattern is more difficult to perceive than color, its mention is nevertheless as beneficial to visual search as that of color. Further research is necessary to determine the relation between visual attention as measured by inspection probabilities, and cogntive load as measured by the ICA.

We further examined if and how processing is influenced by a more or less uniform reduction of referential entropy, i.e., of the size of the referential search space. Specifically, we contrasted conditions where the pre-nominal adjective reduced entropy from by 1.58 bits (UR) with cases in which entropy was decreased by only 0.58 bits (NR), to establish whether what determines efficient entropy reduction is determined by the more or less uniform decrease of entropy over the referential expression. Our results seem to provide evidence that processing is facilitated by the uniform reduction of referential entropy, though not in the predicted region. That is, in the Adjective there were no differences for either inspections (to singleton objects) or for ICA values between UR and NR. On the Noun, however, we found indications that the greater reduction of entropy at the first step contributing to a more uniform entropy reduction profile was preferred, since in (both MS and OS) UR conditions the target object collected more inspections (though only in color items), and, perhaps more interestingly, ICA values were lower than in NR (for all types of items). Importantly, this finding demonstrates that the ICA is sensitive to visual search difficulty, capturing differences in the cognitive effort expended with different rates of visual entropy decrease. With respect to our research question, visual search, and therefore referential processing, appears to be more efficient when the remaining set of possible referents at the final step is rather small, as is the case in UR. Interestingly, however, there is no penalty for this increased entropy reduction on the adjective.

We acknowledge that these results are open to alternative interpretations. For example, the absence of an Entropy Reduction effect on the adjective may be due to our operationalization of Uniform and Non-uniform Reduction, and not because the entropy reduction rate only affects the


Figure 3. Inspection probabilities for the noun region in color items. Error bars represent 95\% CI.
final step. That is, reducing referential entropy by 1.58 bits vs 0.58 bits may not be sufficient to induce a differential cost on the adjective. So, if the difference between the remaining entropy in the two conditions was enhanced, by using a larger referent set and going from e.g. 12 to 2 versus to 8 potential referents in Uniform and Non-uniform Reduction, respectively, might serve to amplify a reductionrelated cost on the adjective in the Uniform condition. Relevant to this issue, it would be interesting to compare processing of OS as implemented in this experiment, with their MS counterparts, i.e. without modification of the noun (e.g., "Finde den Ball"), as the latter is not only a case of rapid reduction of entropy, but is also MS. Any facilitation for OS under this comparison should suggest that processing ease for OS is due to the insertion of an intermediate step in reducing visual entropy that makes reduction more uniform. A final possible explanation that is worth pursuing, as it is related to the nature of the ICA measurements, is that ICA may not be sensitive to such modulations of entropy reduction. In other words, ICA may only be able to capture whether visual search has been demanding or not. Future research is of course required to tackle these questions.

In sum, we present eye-tracking evidence confirming that the use of redundant noun modifiers (over-specification) facilitates referential processing. In addition, we showed that listeners rapidly exploit incoming information about the target to reduce the referential search space in situated comprehension. Greater reduction in referential entropy on the adjective - while not associated with any increase in cognitive load in that region - results in an overall more uniform entropy reduction profile and in reduced cognitive effort when processing the noun. This result leads us to conclude that efficient processing is determined by both the degree of specificity of the reference, as well as to the distribution of entropy reduction across the utterance.

## Acknowledgments

We thank Mindaugas Mozuraitis and Maria Staudte for assistance with the eye tracking analyses, and Mirjana Sekicki, Vera Demberg and Ben Peters for help with the ICA analyses. This research was supported by SFB/CRC1102 "Information Density and Linguistic Encoding" awarded by the German Research Foundation (DFG).

## References

Arts, A., Maes, A., Noordman, L., \& Jansen, C. (2011). Overspecification facilitates object identification. Journal of Pragmatics, 43, 361-374.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.
Crocker, Demberg, \& Teich (2016). Information Density and Linguistic Encoding (IDeaL). Künstliche Intelligenz, 30, 77-81.

Davies, C., \& Katsos, N. (2013). Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al. (2006). Journal of Pragmatics, 49(1), 78106.

Demberg, V., \& Sayeed A. (2016). The Frequency of Rapid Pupil Dilations as a Measure of Linguistic Processing Difficulty. PLoS ONE, 11(1). doi:10.1371/journal.pone. 0146194
Engelhardt, P. E., Bailey, K., \& Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity. Journal of Memory and Language, 54, 554-573.
Engelhardt P. E., Demiral, Ş. B., \& Ferreira, F. (2011). Over-specified referring expression impair comprehension: An ERP study. Brain and Cognition, 77, 304-314.
Geurts, B., \& Rubio-Fernández P. (2015). Pragmatics and Processing, Ratio, 28(4), 446-469.
Grice, P. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and Semantics: Speech Acts (Vol.III). New York: Academic Press.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61, 23-62.
Just, M. A., \& Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. Canadian journal of experimental psychology. 47(2), 310-339.
Marshall, S. P. (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. US Patent $6,090,051$.
Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants, (pp. $7.5-7.9$ ). New York: IEEE.
Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics 27, 89-110.
Rubio-Fernández P. (2016). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. Frontiers in Psychology 7:153. doi: 10.3389/fpsyg.2016.00153

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., \& Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual interpretation. Cognition, 71, 109-147.
Shannon, C. E. (1948), A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379423. doi:10.1002/j.1538-7305.1948.tb01338.x

Tarenskeen S., Broersma M., \& Geurts B. (2015). Overspecification of color, pattern, and size: salience, absoluteness, and consistency. Frontiers in Psychology 6:1703. doi: $10.3389 /$ fpsyg. 2015.01703
Tourtouri, E. N., Delogu, F., \& Crocker, M. W. (2015). ERP indices of situated reference in visual contexts. Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 2422-2427). Austin, TX: Cognitive Science Society.

# A Cognitive Model of Social Influence 

J. Gregory Trafton<br>(greg.trafton@nrl.navy.mil)<br>Naval Research Laboratory<br>Washington, DC

J. Malcolm McCurry<br>(malcolm.mccurry.ctr@nrl.navy.mil)<br>Harris<br>Washington, DC

Kevin Zish<br>(kzish@gmu.edu)<br>George Mason University<br>Fairfax, VA

Laura M. Hiatt<br>(laura.hiatt@nrl.navy.mil)<br>Naval Research Laboratory<br>Washington, DC

Sangeet Khemlani<br>(sangeet.khemlani@nrl.navy.mil)<br>Naval Research Laboratory<br>Washington, DC


#### Abstract

We describe two different cognitive process models of a well known experiment on social influence (Salganik, Dodds, \& Watts, 2006). One model, the social influence model, reproduced the choices that participants took by modeling both the cognitive processes the participant engaged in and the social influences that the participant saw. The second model, the pure cognitive model, used only cognitive capabilities and did not model any social influences that the participant saw. Somewhat surprisingly, the two models showed no difference in quality of fit (the pure cognitive model actually fit slightly better than the social influence model), suggesting that social influence models should take cognitive functions into account in their theories.


Keywords: Cognitive models, social influence, cognitive architectures

## Introduction

People are routinely influenced by other people: this is the crux of social influence. There are many factors that can impact social influence, including the popularity of others (Cialdini, Reno, \& Kallgren, 1990; Latane, 1981), authority or expertise (Cialdini et al., 1990; Milgram, 1963), and culture (Milgram, 1963). While each of these factors can have a large impact in different situations, a fourth factor, visibility -- seeing what others have done or are doing -seems to be among the most important (Cialdini et al., 1990).

Most models of social influence describe the effect in terms of social constructs (e.g., conformity, peer pressure, etc.) and/or networks of people (e.g., families or friends), and that cognition has a relatively minor explanatory role.

For example, MacCoun (2012, in press) has proposed a very successful model of social influence called the uBOP (unidirectional burden of proof). The model itself is a mathematical model in the form

$$
\mathrm{p}=\mathrm{m} /(1+\exp [\mathrm{c}(\mathrm{~S} / \mathrm{T}-\mathrm{b})])
$$

where p is the probability that an individual chooses an option, $m$ is a ceiling parameter, $S$ is the number of people advocating one option, T is the number of advocates advocating a second option, $b$ denotes where an individual is more likely to adopt the group's decision and c reflects the difficulty to make a decision (steepness). This model can successfully characterize the classic Milgram (1969) and Asch (1951) studies with few changes in parameters (MacCoun, 2012; MacCoun, in press).

Social network models have also been used to model social influences. Social networks consist of nodes (people) who are linked through some form of interdependency (family, friends, beliefs, etc.). Social networks have been very successful at differentiating the effects of social ties from other external influences and have been applied to explain phenomena as diverse as smoking (Christakis \& Fowler, 2008) and obesity (Christakis \& Fowler, 2007). Social network models typically use graph theoretic or network models or statistical models (e.g., structural modeling or autoregression).

Cognition in all of these models serves, at best, a purely functional role: people perceive others' actions or remember actions that others have performed, but the theoretical power comes from social or network constructs. The uBOP model can be used to describe both individuals and groups, but has nothing that could be considered a cognitive process. Social network models describe relationships and membership rather than an individual's cognitive activities. The fact that there are few (if any) cognitive processes in these models is, perhaps, not surprising: most of the existing models are not process models. We believe that cognition is a large component of most social behavior and will explore this issue by developing a cognitive process model of a wellknown social influences study (Salganik et al., 2006; Salganik \& Watts, 2009).

Salganik and colleagues investigated the effects of social influence in a cultural market with a novel paradigm
(Salganik et al., 2006; Salganik \& Watts, 2009). Salganik et al. created an artificial music market where participants could listen and download previously unknown songs. Salganik and colleagues created independent instantiations, or worlds, where the markets could grow without influence from other worlds; this was a between subjects manipulation. Individuals in each world could only be influenced by individuals in their own world. This approach allowed the authors to explore how social influences develop over time in different situations. Across two experiments, they looked at two conditions, an independent world and a social influence world. Participants in the independent world made decisions about what songs to listen to based only on the names of the bands and the song, while in the social influence worlds, participants could also see how many times each song had been downloaded by previous participants. In their first experiment, they found a modest social influences effect, but in their second experiment they found a very strong social influences effect. We describe and model the second experiment.

## Method (Salganik et al., 2006)

A complete description of the experiment can be found in Salganik et al. (2006) and Salganik and Watts (2009).

## Participants

There were 7192 participants recruited from a music website (Bolt). There were approximately 700 participants in each of eight social influence worlds and approximately 1400 participants in the independent world. ${ }^{1}$ Participants logged onto the website and various times over 83 days.

## Setup and Procedure

48 Songs were presented in a single column and sorted by the number of downloads for the social influence worlds and in a random order for the independent world. The display was updated as every participant downloaded songs. Additionally, the social influence world displays contained information about the number of downloads each song had received; this information was dynamically updated as the experiment progressed; see Figure 1.

[^522]

Figure 1: A screenshot of the social influence world, taken from Salganik and Watts (2009).

Participants were able to click any song on the list to listen to it. While the song was playing, participants were asked to rate the song on a $1-5$ scale where 1 was "I hate it" and 5 was "I love it." After rating the song, participants could download the song and then could go back to the primary display so they could listen, rate, and download more songs if they chose. Participants were able to download as many of the 48 songs as they wished, but they had to listen to them and rate them before they could download each song.

Each of the social worlds began in a random state, so each social world could evolve based on the participants' behavior in that specific world.

## Measures

There were several different variables that the authors coded in the data. The number of songs each participant listened to and the number of downloads that were made was recorded. The popularity of individual songs was also recorded. One of the most informative variables that was recorded was how often participants listened to a song at a specific rank (regardless of what song it was). When each participant examined songs, each song had a specific market rank (the song with the most downloads had a market rank of \#1).

## Results and Discussion

Salganik et al. reported that participants listened to an average of 3.6 songs and downloaded an average of 1.4 songs. Figure 2 shows the probability that a participant would listen to a song based on its rank market share. Note that all the social influence worlds will be combined for this and all further analyses, as reported in Salganik et al.


Figure 2: The probability that a participant would listen to a song based on its rank market share for both the independent and social influence conditions in Salganik et al.'s experiment.

As Figure 2 suggests, the independent condition was mostly flat, with no strong effect of either social influence or of song quality. However, the social influence conditions showed a very strong effect of social influence: the top ranked song had a $45 \%$ chance of being listened to, while the average song had only a $5 \%$ chance of being listened to. There is also an interesting "hipster" effect when the song that was ranked last got listened to a great deal more than average.

Salganik et al. suggest that these results "confirm that the popularity of the songs affected participants' choices and generally led them to listen to the more popular songs-a result that is consistent with the large literature on social influence and conformity" (Salganik \& Dobbs, 2009, p. 447). They also show that while specific songs were considered better than others, the social influence condition had a substantial effect on the success of the songs.

As suggested earlier, most models of social influence are not cognitive process models. So we developed a cognitive process model of the individuals in this experiment in order to examine the effect of cognition and social influence. By developing a process model, we were able to create two slightly different models: a social influence model and a pure cognitive model. By developing the two models, we will be able to determine how much better the social influence model fits the data beyond the pure cognitive model and thus determine the importance of social influence over basic cognitive factors. The models were developed using the ACT-R architecture.

## Architecture and Model Description

ACT-R is a hybrid symbolic/sub-symbolic production-based system (Anderson et al., 2004; Anderson, 2007). ACT-R consists of a number of modules, buffers, and a central pattern matcher. Modules in ACT-R contain a relatively specific cognitive faculty usually associated with a specific region of the brain. For each module, there are one or more buffers that communicate directly with that module as an interface to the rest of ACT-R. At any point in time, there may be at most one item in any individual buffer; thus, the module's job is to decide what and when to put a symbolic object into a buffer. The pattern matcher uses the contents of the buffer to match specific productions.

ACT-R uses if-then rules (productions) that will fire when their preconditions are met by matching the contents of the buffers. If there is more than one production that can fire, the one with the highest utility (production strength) will fire. Each production can change either internal state (e.g., buffer contents) or perform an action (e.g., click on a button).

ACT-R interfaces with the outside world through the visual module, the aural module, the motor module, and the vocal module. The architecture supports other faculties through intentional, imaginal, temporal and declarative modules.

## High Level Description of the Social Influences Model

There are three components to each model: search, consideration, and decision-making. Each component has different productions that instantiate the specific goal. There model is a pure performance model: there is no learning in the model.

Search ACT-R has a theory about visual attention (Byrne \& Anderson, 1998), which this model follows. In brief, the model searches for an unattended song, then moves its visual attention to that song and then encodes the information about the song. The model determines which song to search for in one of four ways:
(1) The model begins at the top of the display to search for an unattended song. This is typical ACT-R behavior for searching.
(2) The model finds a random song and attends to it.
(3) The model starts at the bottom of the display to search for an unattended song. This is the "hipster" component of the model.
(4) Stop searching completely and finish.

All four of these rules are in competition any time the model has a goal to look for a song. Note that if we were in a different culture where reading occurred in a different direction, the model would need to take those preferences into account. Also note that sometimes the model will have a goal to search for a new song and then give up; participants also began the study and stopped the experiment before listening to any songs.

Consideration After a song has been attended to and encoded, the model next determines whether that song should be listened to or not. It has three options:
(1) The model decides that the song "looks interesting" so it decides to listen to it. We assume that people have some preference for the name of the band or the name of the song; this is a simple version of that preference process.
(2) The model decides that the song "looks terrible" so it decides not to listen to it. Again, this is a simple way to model the preferences that people have.
(3) The song is listened to based on its rank. The probability is a very simple $1 /$ rank. There are other, more sophisticated versions of selection based on group behavior (Mullen, 1983), but this simple version suffices for this model. Note that this is where social influence occurs in this model.
All three of these rules are in competition any time the model has a goal for considering whether to listen to the song. If the model decides not to listen to a song, it searches for another song.

Decision-making If the model does listen to a song, it must next decide whether to download it. The decision to download is very straightforward: there is a $50 \%$ chance the model will decide it should download the song. If it does download the song, the world is updated; if it does not download the song the world is not updated in any way, but the model then searches for another song.

## A series of sample experimental model runs

For the following example, three models are run in the same world; the social influences model is being run in the social influences condition. We assume that each model corresponds to a single simulated participant. The world is updated based on what each model does in the world, and the world is displayed appropriately based on what others have done.

The first thing that the model does in an experiment is to search for a song. The first model in the experiment stops searching and no updating occurs.

The second model in the experiment starts at the top of the list of songs; the first song on the list is unattended, so it encodes it and considers whether to listen to it. The model decides that it will listen to it and then must decide if it should download it. There is a $50 \%$ chance the model will download it, which it does in this case. This song now becomes the most popular with rank 1 and for future participants it will show as the top song on the display. The model then searches for another song, again decides to search from the top and finds the second song, which is the top unattended song. The second song does not look interesting, so the model does not listen to it (and thus does not download it, either). The model next searches for another song, but then stops searching and this model is finished.

The third model sees the previously downloaded song in the first slot. The model, however, chooses a random song from the list and decides to listen and download it. The model next starts at the bottom of the display and looks for an unattended song. The model will listen to this song based on its rank, which is currently 48; so it has approximately a $2 \%$ chance of listening to the song. Luckily, for this run the model will listen to it, so that song now is tied for rank 1 , and all future models will evaluate appropriately. After 2000 model runs, the simulation is stopped and the simulated world is reported.

The run just described was based on the social influences model. In this model, social influence occurs during the consideration stage. ${ }^{2}$ The pure cognitive model was identical to the social influences model, except it did not have pay attention to any social influence. Without social influences, the model simply considers a song based on whether it is "interesting" or "terrible."

For all models, we kept most of the ACT-R parameter defaults. The parameters that were changed include a production noise parameter (.4, which is within a normal range for this parameter) to provide some stocasticity and the aforementioned $50 \%$ probability for downloading a specific song. Parameter fits were run using the social influence model and those same parameters were also used for the pure cognitive model.

## Model fit

First, it is possible to examine how many listens and downloads each model performed and compare them to the experimental data. On average, there were 3.6 listens per participant in the experimental data; both models made 3.2 listens. Comparably, there were an average of 1.4 downloads per participant; both models had 1.6.

Figure 3 shows the fit of the independent condition; both models provide the same results. As Figure 3 suggests, the fit is quite good, with the model data overlapping a great deal with the experimental data. Calculating an $\mathrm{R}^{2}$ is uninformative because both the data and the model are flat. RMSD for this model is .02 , which demonstrates quite a good fit.

[^523]

Figure 3: Data and model of the independent condition.
Figure 4 shows the fit for the social influence model and Figure 5 shows the fit for the pure cognitive model. As is evident from Figures 4 and 5, they both show quite a good fit; Table 1 shows the quantitative fit statistics. The social influence model had a very strong fit in both $\mathrm{R}^{2}$ and RMSE. Somewhat surprisingly, however, the pure cognitive model had a slightly better $\mathrm{R}^{2}$ fit and a comparable RMSE fit. We can conclude from these analyses that the social influence model does not fit better than the pure cognitive model.

## Data/Model Fit

Salganik et al., 2006


Figure 4: Experimental data from the social influence condition and the social influence model.


Figure 5: Experimental data from the social influence condition and the pure cognitive model..

| Model | $\mathbf{R}^{\mathbf{2}}$ | RMSE |
| :--- | :--- | :--- |
| Social Influence | .88 | .023 |
| Pure cognitive model | .92 | .028 |

Table 1: Fit metrics for both the social influence model and the no social influence model. Both models were compared to the social influence condition of Salganik et al. (2006).

## General Discussion

We described a process model of a well-known experiment on social influence (Salganik et al., 2006; Salganik \& Watts, 2009). The experiment showed that when people had access to what others had done, it greatly influenced their behavior, consistent with current theories on social influence (Cialdini et al., 1990; MacCoun, 2012). We built two slightly different cognitive process models that perform the perceptual and cognitive steps in the experiment. Both the social influence model and the pure cognitive model fit the data extremely well. However, somewhat surprisingly, the pure cognitive model fit the experimental data slightly better than the social influence model. We interpret these results as showing that for this experiment, the effect of social influence is very small: a pure cognitive model was able to fit the data at least as well (if not slightly better) than the social influence model.

It was a bit surprising that the pure cognitive model and the social influence model shared so much overlap: this is almost assuredly one of the reasons for the similarity in the two models. This should not come as a big surprise, however: this type of task of searching and selecting objects on a computer screen is a classic cognitive task that
has been investigated both experimentally and theoretically many times.

It could be argued that during the search phase, the scanning down rule is also a social influence rule since participants knew that songs were ranked from top to bottom in order of the number of downloads. However, we would argue that scanning from the top to the bottom of a list is more a cognitive and cultural function than a social influence function. Many other researchers have shown that people in the US search for objects approximately top-down and left-to-right on computer interfaces (Byrne, Anderson, Douglass, \& Matessa, 1999; Norman, 1991; Schunn \& Anderson, 1999).

Note that we are not saying that people are not influenced by social influence. There are many experiments and models that show the importance of social influence. For example, Cialdini et al. (1990) found that when there was evidence that other people had littered, individuals were more likely to litter than when there was evidence that people had not littered. Many other classic experiments have shown the importance of social influence (Asch, 1951; Milgram, 1963)

The model presented here does, however, highlight the importance of cognitive processes in explaining at least some social influence effects. We believe that providing a process level description of cognitive and social behavior will lead to a better understanding of how social influences impact people's behavior. Specificially, we can isolate those processes that may result from cognitive aspects of the task from those processes that result from social influence.

## Acknowledgments

This work was supported by the Office of Naval Research to JGT. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies of the U. S. Navy. We thank Sunny Khemlani, Laura Hiatt, Alan Schultz, and Joe Kreke for their advice and comments on a previous draft.

## References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., \& Qin, Y. (2004). An integrated theory of mind. Psychological Review, 111(4), 1036-1060.
Anderson, J. R. (2007). How Can the Human Mind Occur in the Physical Universe? Oxford University Press, USA.
Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H . Guetzkow (Ed.), Group leadership and men (pp. 177190). Pittsburgh, PA: Carnegie Press.

Byrne, M. D., \& Anderson, J. R. (1998). Perception and action. In J. R. Anderson \& C. Lebiere (Eds.), Atomic Components of thought (pp. 167-200). Mahwah, NJ:: Lawrence Erlbaum.

Byrne, M. D., Anderson, J. R., Douglass, S., \& Matessa, M. (1999). Eye tracking the visual search of click-down menus. Proceedings from Proceedings of the computer human interaction conference.
Christakis, N. A., \& Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network Over 32 Years. New England journal of medicine.
Christakis, N. A., \& Fowler, J. H. (2008). The Collective Dynamics of Smoking in a Large Social Network. New England journal of medicine.
Cialdini, R. B., Reno, R. R., \& Kallgren, C. A. (1990). A focus theory of normative conduct: Recyling the concept of norms to reduce littering in public places. Journal of personality and Social Psychology, 58, 1015-1026.
Latane, B. (1981). The psychology of social impact. American psychologist, 36, 343-356.
MacCoun, R. J. (2012). The burden of social proof: Shared thresholds and social influence. Psychological review, 119, 345-372.
MacCoun, R. J. (in press). Computational models of social influence and collective behavior. In Computational Models in social psychology. New York: Psychology Press.
Milgram, S. (1963). Behavioral study of obedience. The Journal of abnormal and social psychology, 67(4), 371.

Mullen, B. (1983). Operationalizing the effect of the group on the individual: A self- attention perspective. Journal of Experimental Social Psychology, 19, 295322.

Norman, K. L. (1991). The psychology of menu selection: designing cognitive control of the human/computer interface. Norwood: Ablex.
Salganik, M. J., Dodds, P. S., \& Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311, 854-856.
Salganik, M. J., \& Watts, D. J. (2009). Web-based experiments for the study of collective social dynamics in cultural markets. Topics in Cognitive Science, 1, 439-468.
Schunn, C. D., \& Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. Cognitive Science, 23(3), 337-370.

# Why Teach How Things Work? Tracking the Evolution of Children's Intuitions about Complexity 

Emmanuel Trouche (Emmanuel.Trouche@yale.edu)<br>Department of Psychology, Yale University,<br>New Haven, CT 06520 USA<br>Aaron Chuey (Aaron.Chuey@yale.edu)<br>Department of Psychology, Yale University, New Haven, CT 06520 USA

Kristi L. Lockhart (Kristi.Lockhart@yale.edu)<br>Department of Psychology, Yale University, New Haven, CT 06520 USA

Frank C. Keil (Frank.Keil@yale.edu)<br>Department of Psychology, Yale University, New Haven, CT 06520 USA


#### Abstract

Mechanistic information can be characterized as the interacting causal components underlying a phenomenon - in short, how something works. Children and adults are notoriously poor at learning, remembering, and applying mechanistic information, so it comes as no surprise that the wisdom of teaching mechanism has come under increasing scrutiny in science education. However, while a rich memory for mechanistic details may be out of the average student's grasp, we argue that exposure to mechanism does not leave students empty-handed. Instead, it refines their intuitions about science and the world in significant ways. For the current study, we focused on one kind of intuition in particular: beliefs about causal complexity. Children ages 611 rated the complexity of a heart and a lock and were then given either mechanistic or non-mechanistic information about them. Afterwards, they were asked if their intuitions about complexity had changed and if so by how much. Three weeks later, children were asked again about their intuitions about complexity. Crucially, children who were given mechanistic information demonstrated a significantly greater shift in their assessments of complexity for both the heart and door lock compared to their counterparts who were given non-mechanistic information. This contradicts the notion that mechanism provides learners with few benefits while also demonstrating how mechanism can be a powerful force in shaping children's intuitions.


Keywords: causal mechanisms; explanation; complexity intuitions; meta-knowledge; cognitive development

## Introduction

Humans possess cognitive systems that enable them to grasp causal relations around them. As early as eight months, children are able to predict outcomes of novel causal events (Sobel \& Kirkham, 2006) and less than a year later, they are capable of making successful causal interventions (Gopnik et al., 2004; Gopnik, Sobel, Schulz,
\& Glymour, 2001), even for causal relationships defined by abstract relational properties (Walker \& Gopnik, 2014).

However, given the rapid growth of human knowledge and technology in the modern era, many of the known causal relations in the world are becoming increasingly complex and inaccessible (Arbersman, 2016). Except for relevant experts, this burgeoning set of causal patterns presents a challenge for how laypeople grasp the causal structure of the world around them. Most adults, for example, are unable to give even a basic explanation of the mechanisms underlying everyday objects like a door lock or a clock, let alone more complicated objects like a car engine or a computer. Similarly, we have a surprisingly poor understanding of the mechanisms underlying the functioning of living things and even how our own bodies work. To make matters worse, laypeople often believe they possess detailed mechanistic knowledge about the world despite having next to none, a phenomenon known as the Illusion of Explanatory Depth (IOED) (Alter, Oppenheimer, \& Zemla, 2010; Rozenblit \& Keil, 2002). These major gaps in adult causal knowledge occur not just in recall but also in recognition. For example, many adults fail to recognize the difference between a schematic of a functional bicycle and one that is completely inoperable (Lawson, 2006). In children, this illusion is present to an even greater degree (Mills \& Keil, 2004).

Our cognitive mechanisms dealing with causality seem to fall short in keeping vivid representations of somewhat complex causal patterns. Despite this tendency to forget mechanistic information, humans however show a certain curiosity about how things work. Children ask for mechanistic information about things they encounter, usually phrased as ubiquitous "why" and "how" questions, starting around three years of age (Callanan \& Oakes, 1992), and they are often relentless in their questioning until
they receive a mechanism-oriented response (Chouinard, 2007; Frazier, Gelman, \& Wellman, 2009). This preference for mechanistic explanation persists into adulthood (e.g., Ahn et al., 1995; Johnson \& Ahn, 2015). In short, despite our poor ability to learn, retain, or recall mechanistic information, people of all ages often show considerable interest, and at times an outright preference, for mechanistic information.

## Mechanisms in the Classroom

A natural reaction to the massive decay in our retention of mechanism is to downplay the need to learn it at all and refocus science education on topics such as the nature of science, epistemological stances, and methodology (Osborne et al, 2003). After all, even if students enjoy learning information about how things work, they fail to retain it; shouldn't we focus on teaching them things they can actually remember? Without denying the importance of topics like epistemology, we argue there are insidious costs in failing to expose children to rich mechanistic details. At root, the benefits of mechanism lie not with the details learned, but with the higher order intuitions acquired and sharpened as a result of teaching mechanism.

More precisely, we argue the bulk of cognitive gain from exposure to mechanism occurs at the "meta-knowledge" level (Kominsky, Zamm \& Keil, in press). Even if we have no idea how a car engine actually works, we do have some intuitions about the underlying mechanisms: for example, we might think it involves metal and plastic components as opposed to organic parts, that it is extremely complex and difficult to learn about, and more crucially, we may also be able to tell apart experts from laypeople when hearing them talk about the mechanism. Thus, even if we do not know the details of how an object works, we often have surprisingly accurate intuitions about how much "stuff" is in a mechanism, how complicated a mechanism is, and whose expertise we can rely on.

Indeed, despite the decay of knowledge about the mechanism itself, some kind of mechanistic information seems to persist: mechanistic information influences causal reasoning (Ahn et al., 1995; Schlottmann, 1999), and mechanism may constrain Bayesian causal learning by reshaping priors about what causal links exist or how strong they are (Griffiths \& Tenenbaum, 2005, 2009). Thus, some aspects of mechanistic information are preserved, but are neither detailed nor complete (e.g., DiSessa, Gillespie, \& Esterly, 2004; Straatemeier, van der Maas, \& Jansen, 2008; Vosniadou, 2002). If most individuals do not retain deep, integrated understandings of "how things work", what kind of mechanistic knowledge does persist?

The current study pursues the beginnings of an answer to this broad question by focusing on intuitions about complexity. Intuition about causal complexity is a good candidate for a kind of meta-knowledge that would persist after memory decay. For instance, one may have a strong feeling that the mechanisms underlying the human ear or a
clock are highly complex without being able to give any accurate descriptions of the mechanism itself.

To summarize, we argue that mechanism instruction is essential to STEM learning as long as we acknowledge what is actually retained over time and what is not. Expecting children to retain fine-grained mechanistic details is simply an unrealistic goal. Instead, focusing mechanistic exposure on building richer meta-mechanistic knowledge establishes both achievable and useful goals. In particular, we argue that exposure to mechanism is a necessary pathway to other forms of more enduring representations such as intuitions about causal complexity, the focus of the current study.

## Experiment

## Stimuli

We chose a heart and a door lock as stimuli because they look quite simple from the outside while having a somewhat rich causal mechanism on the inside. In order to control for a potential reaction of surprise to hidden complexity in the mechanistic condition, the verbal information provided in the non-mechanistic condition included, both for the heart and the door lock, some surprising facts such as "people think that the heart is red but actually the heart itself is dark brown".

The number of words of verbal information presented to the children was matched between the mechanistic (212 words) and non-mechanistic condition (208 words). We also created the text stimuli so that the non-mechanistic information was more superficially complex than the mechanistic information (e.g., Flesh Reading Ease: 79.97 for the non-mechanistic text and 91.53 for the mechanistic text).

## Predictions

We had no predictions about the age of children in this study but included this variable in our analysis as exploratory.

H1) Exposure to mechanistic information should shift children's intuitions of complexity.

H2) The children who underestimate complexity should move towards higher complexity judgments once provided with mechanistic information.

H3) This shift towards higher complexity should still be observable three weeks later.

## Methods

Participants We recruited 144 children from an elementary school in the New Haven, CT area. Our sample was somewhat atypical in two ways: the elementary school is situated in a low SES neighborhood and the elementary school has a strong focus on science (classified as a STEM school). Our sample consisted of 20 Kindergarteners, 15 first graders, 15 second graders, 41 third graders, 34 fourth graders and 18 fifth graders. The experimenters interviewed child participants individually for about ten minutes in a
quiet spot in the school. All participants were rewarded with a small toy.

Training Phase Children were told they would be presented with pictures of things and would be asked if they thought the thing is simple or complicated. Participants were explicitly told both at the very beginning and just after the training phase "there are no right or wrong answers, I just want to know what you think".

The two training questions were designed to introduce children to our complexity scale and to have a clear criterion of exclusion. The training consisted in showing children black and white drawings of an hourglass and grandfather clock. In order to prevent children's intuitions about complexity being driven only by the ease of use - as previous pilot experiments suggested - children were told that "both these two things are easy to use but the way they work is really different", followed by a short justification of why the hourglass is simple ("It's just sand going down") and why a grandfather clock is more complicated ("It has many gears and pendulums inside that all work together to move the hands on the clock").

Children were then presented with black and white drawings of a bicycle and a motorcycle. For each, they were asked whether they thought it was simple or complicated. Depending of their first answer, they were asked if they thought it was "really" simple/complicated or "kind of" simple/complicated. Children's answers on each entity can thus be coded on a 2 point scale (simple or complicated) or on a 4 point scale: really simple ( 0 ), kind of simple (1), kind of complicated (2) and really complicated (3).

Test Phase Just after the training phase, children were presented with a black and white drawing of a heart or a door lock (order of presentation was counterbalanced). They were asked whether they thought the entity was simple or complicated in the same way as in the training phase. All participants were then told that some information about the entity would be presented to them. Half the participants were randomly assigned to the mechanistic condition and the other half were assigned to the non-mechanistic condition. Participants in the mechanistic condition were presented with pictures of the inside of a heart and of a door lock, along with some verbal information about how it works. Participants in the non-mechanistic condition were presented with pictures illustrating some facts about the heart or door locks along with corresponding verbal information matched in length to the mechanistic condition.

After hearing the information, all participants were reminded of their initial complexity judgment and asked if they wanted to keep their answer or if they thought a heart / a lock was more complicated/simple than they had previously thought. Depending on their answer, they were then asked if it is much more complicated/simple than what they though before or a little bit more complicated/simple than what they though before. Their shifts in complexity judgments were coded with a 5 point scale: much more
simple (-2), a little more simple (-1), still what I think (0), a little bit more complicated $(+1)$ and much more complicated $(+2)$. The exact same procedure was repeated with the second entity.

Retention Test, Just after the complexity judgments, all children were asked a series of questions about the information (mechanistic or non-mechanistic) that had been given to them. This test was used to assess the children's retention of the information that was just presented to them in order to measure retention scores on the exact same test three weeks later. Questions were either "yes - no" questions ( 0.5 point); questions about quantity or colors (1 point) or open-ended (2 points). In the mechanistic condition, scores could range from 0 to 12.5 ; in the nonmechanistic condition, scores ranged from 0 to 18.5 . All the scores were normalized to range between 0 and 1 . Based on the Yes/No questions, we calculated the chance level for each condition: 0.10 in the mechanistic condition and 0.04 in the non-mechanistic condition.

Exclusion The 24 children (17\%) who, during the training phase judged a bicycle as more complex than a motorcycle were excluded from the analysis. In addition, since $50 \%$ (9) of the kindergarteners failed to judge a motorcycle as more complex than a bike, all the kindergarteners were excluded from the analyses. The following analyses apply to a sample of 111 children.

## Results

Analysis Children were grouped into three age groups: $1^{\text {st }}$ and $2^{\text {nd }}$ graders $(N=26), 3^{\text {rd }}$ graders $(N=37), 4^{\text {th }}$ and $5^{\text {th }}$ graders ( $N=48$ ).

Our analyses focus on three dependent variables, the absolute value of shift, the raw value - direction -- of shift, and the direction of long-term shift three weeks after the initial measure. We looked at the effect of three independent variables. Two variables were linked to our hypotheses, condition (mechanistic and non-mechanistic) and initial rating (which we grouped in two levels - simple or complicated - instead of the 4 measured ones - kind of / really - in order to increase our statistical power). The third variable was age group ( $1^{\text {st }}$ and $2^{\text {nd }}$ graders, $3^{\text {rd }}$ graders, and $4^{\text {th }}$ and $5^{\text {th }}$ graders), which was exploratory. Only the interactions with condition were tested.

Initial complexity judgments were significantly higher for the heart $(M=1.7, S D=1.14)$ than for the lock $(M=0.96$, $S D=1.03$ ), paired t-test $t(110)=4.79, p<.001$. In the following analyses the two entities were analyzed separately.

Absolute value of shift For both the heart and the lock, we performed a 2 (condition) x 2 (initial rating) x 3 (age groups) fully between-subjects Analysis of Variance (ANOVA) with the absolute value of shift as DV. For both entities, the ANOVAs revealed a main effect of condition $\left(\right.$ Heart: $F(1,103)=4.13, p=.04, \eta_{\mathrm{p}}^{2}=.04$; Lock: $F(1,103)$
$\left.=17.78, p<.001, \eta_{p}{ }^{2}=.14\right)$. Those main effects of condition came from a larger absolute value of shift for children in the mechanistic condition both for the heart ( $M=$ $1.14, S D=0.67$ versus $M=0.87, S D=0.75$ ) and the lock ( $M=1.22, S D=0.60$ versus $M=0.69, S D=0.77$ ).

A main effect of initial ratings was also found for both entities (Heart: $\left(F(1,103)=4.10, p=.05, \eta_{\mathrm{p}}{ }^{2}=.04\right.$; Lock: $\left.F(1,103)=4.07, p=.05, \quad \eta_{\mathrm{p}}^{2}=.03\right)$ corresponding to a larger absolute value of shift for participants initially judging an entity as simple. No other main effects or interactions were significant for either entity.

Direction of shift For each entity, we performed a 2 (condition) x 2 (initial rating) x 3 (age groups) fully between-subjects ANOVA with the shift in complexity judgment measured just after children were exposed to some information as a DV (from -2 to +2 ) (see Fig. 1).

For the heart, a significant interaction between the initial rating and the condition was found $(F(1,103)=5.56, p=$ $.02, \eta_{\mathrm{p}}^{2}=.05$ ). To further explore this interaction, two posthoc two sample t-tests with Bonferroni adjusted alpha levels of .025 per test (.05/2) showed that for the children who initially judged the heart as simple, being exposed to mechanistic information resulted in an average shift towards higher complexity ( $M=0.70, S D=1.36$ ) compared to children exposed to non-mechanistic information ( $M=-$ $0.21, S D=1.23 ; t(39.7)=2.26, p=.02)$. As for children who initially judged the heart as complex, there was no difference in complexity shift between the mechanistic ( $M=$ $-0.09, S D=0.14$ ) and non-mechanistic condition ( $M=0.14$, $S D=1.12 ; t(66.4)=-0.83, p=.41)$.

For the lock, the same analysis did not show any significant effect or interaction.

Analyses of High Initial Retention Participants Median scores were calculated on the retention task for each condition, entity, and crucially for each of the three age groups in order to avoid having mostly older children in the high retention group. In order to explore the possibility that some participants were not paying sufficient attention to the task or had difficulty understanding the material presented to them, children scoring lower than the median in each of the groups were dropped from the sample. All the following analyses are similar to the analyses presented before but includes only the high retention half of our sample (for the heart, $N=35$ in the mechanistic condition and $N=29$ in the non-mechanistic condition; for the lock $N=39$ and $N=31$ respectively).

In terms of absolute value of shift, results for the high retention group were similar to those of the entire sample with a significantly larger shift in the mechanistic condition (main effect of condition. heart: $F(1,56)=5.5, p=.02, \eta_{p}{ }^{2}$ $=.08$; lock: $\left.F(1,62)=6.1, p=.01, \eta_{p}{ }^{2}=.09\right)$. As before, a main effect of initial judgment was found only for the heart $\left(F(1,56)=4.2, p=.05, \eta_{p}{ }^{2}=.06\right)$ but not for the lock $\left(F(1,62)=2.6, p=.11, \eta_{p}{ }^{2}=.04\right)$ ), likely due to a lack of power.

Despite having our sample size cut in half, results on the direction of the shift and on the final ratings were even more in line with our initial hypotheses.

With respect to the direction of the shift for the heart, the ANOVA revealed a significant interaction between condition and initial rating (Heart: $F(1,56)=9.92, p=.01$, $\eta_{p}{ }^{2}=.10$ ). For the lock, as opposed to the same analysis with the entire population, the interaction between condition and initial rating was at trend level $(F(1,62)=3.97, p=.09$, $\eta_{p}{ }^{2}=.04$ ). A main effect of initial judgment was also found for the lock $\left(F(1,62)=4.5, p=.04, \eta_{p}{ }^{2}=.06\right)$ but not for the heart $\left(F(1,56)=2.2, p=.14, \eta_{p}^{2}=.04\right)$. As displayed in Figure 1, one pattern seems similar for both entities: children who start by judging the entity as simple in the mechanistic condition tend to increase their complexity rating to a greater extent compared to both children in the non-mechanistic condition as well as children who start by judging the entity as complicated.


Figure 1: Average value of shift (y-axis) with standard errors bars for the high initial retention group, in the mechanistic (grey) and non-mechanistic condition (white) as a function of initial complexity judgment ( $x$-axis). Results for the heart and lock are presented on the left and right panels respectively.

Shift in complexity three weeks later We again asked children about their intuitions of complexity three weeks later using the same methodology. Twelve children (11\%) had changed schools or were absent during the times we were testing. Therefore, the following analyses apply to a sample of 99 children.

For each entity, we performed the same 2 (condition) x 3 (age group) x 2 (initial rating) with the shift between children's initial rating and their rating three weeks later as a dependent variable (long-term shift).

For both the heart and the lock, we found main effects of initial rating (heart: $F(1,91)=44.9, p<.001, \eta_{p}{ }^{2}=.31$; lock: $\left.F(1,91)=33.6, p<.001, \eta_{p}{ }^{2}=.27\right)$. A main effect of
condition was found at trend level for the heart $(F(1,91)=$ $3.0, p=.08, \eta_{p}{ }^{2}=.02$ ), but not for the lock.

When performing the same analysis on the high initial retention group, ANOVAs continued to show a main effect of initial rating for both entities (heart: $F(1,52)=31.7, p<$ $.001, \eta_{p}{ }^{2}=.33$; lock: $\left.F(1,56)=31.2, p<.001, \eta_{p}^{2}=.34\right)$. This was the only significant effect for the lock. For the heart, there was also a significant interaction between condition and initial rating for the heart $(F(1,52)=6.42, p=$ $\left..01, \eta_{p}^{2}=.07\right)$ as well as a main effect of condition at trend level $\left(\left(F(1,52)=3.46, p=.07, \eta_{p}{ }^{2}=.04\right)\right.$. As displayed in Figure 2, the interaction between condition and initial rating was driven by children initially judging the heart as simple, who moved towards higher complexity in the mechanistic condition ( $M=1.46, S D=0.97$ ) compared to the nonmechanistic condition ( $M=0.25, S D=0.88$; Post-hoc t-test with Bonferroni adjusted alpha levels of .025 per test $(.05 / 2), t(16.0)=2.93, p=.009)$. For children initially judging the heart as complex, there was no significant effect of condition (mechanistic: $M=-0.75, S D=1.16$; nonmechanistic: $M=-0.44, S D=1.04 ; t(36)=-0.85, p=.40)$.


Figure 2: For the high retention group, average value of long term shift (y-axis) with standard errors bars in the mechanistic (grey) and non-mechanistic condition (white) as a function of initial complexity judgment (x-axis). Results for the heart and lock are presented on the left and right panels respectively.

Retention Tests Three weeks later, children's retention scores had significantly dropped significantly by 0.20 in the non-mechanistic condition (from 0.64 to 0.44 , paired t -test: $t(95)=9.30, p<.001)$ and by 0.11 in the mechanistic condition (from 0.44 to $0.33, t(101)=4.67, p<.001$ )

When dividing our population between high versus low initial retention, the low initial retention group did not show any decay in the mechanistic condition (from 0.19 to 0.25 ). By contrast, the high retention group had a decay of 0.20 (from 0.57 to $0.37, t(64)=7.33, p<.001$ ). In the nonmechanistic condition, both groups showed significant
decay (low retention group had a decay of 0.17 , from 0.48 to $0.31, t(41)=5.84, p<.001$; high retention group had a decay of 0.24 , from 0.76 to $0.53, t(53)=7.40, p<.001)$.

## Discussion

Our first hypothesis H1 is well supported by the data with an absolute value of shift significantly larger in the mechanistic condition than in the non-mechanistic condition for the two entities. In short, mechanistic information influences children's intuitions about complexity more than non-mechanistic information.

Hypothesis H2 is supported with respect to the heart: both when analyzing the full population and the high retention group, children in the mechanistic condition who initially judged the heart as simple moved toward higher complexity ratings more than children in the non-mechanistic condition. With respect to the lock, the same hypothesis was only supported by the high retention group. This pattern suggests that the influence of mechanistic information prompts more than unpredictable shifts in children's intuitions. The influence of retention group also suggests that the quality of mechanistic exposure has a discernable impact on children's ultimate intuitions. The finding that the low retention group slightly increased their retention score three weeks later likely means they were near floor from the start, indicating they had encoded and understood little to no mechanistic details presented to them.

For both entities, we also found significant main effects of initial rating in the following direction: children initially judging an entity as simple move toward higher complexity and children initially judging an entity as complex tend to decrease their complexity judgments. This pattern raises a question: to what extent were the shifts simple regressions to the mean? At least in the cases of the heart, a tendency towards the mean is not the only significant influence on children's complexity judgments, even if the influence of mechanistic information works in the same direction. Indeed, mechanistic information shifted initially low complexity judgments higher than the overall tendency to regress towards the mean can explain.

Hypothesis H3 was only weakly supported in the case of the heart, with a main effect of condition at trend level. However, this modest effect fits with our hypothesis, since it is driven by children in the mechanistic condition initially judging the heart as simple who showed, three weeks later, a greater increase in their complexity judgments compared to the non-mechanistic condition. The size of this effect illustrates the challenge of trying to shift children's longterm intuitions about the world with less than 10 minutes of instruction.

Participants' SES, background, and attendance at a school having a strong science and engineering focus may also have diminished the hypothesized effects in two ways: first, low SES children often face increased attentional challenges in school settings (Mezzacappa, 2004; NICHD Early Child Care Research Network, 2003). These challenges may also help explain why our main hypotheses were supported more
by the high retention half of our sample. Second, the STEM focus of the school may have diminished the strength of the main effect of mechanism on the size of the shift by giving children more previous exposure to mechanism than is typical, in turn providing less room for intuitions about complexity to shift in a mere 5-10 minute span.

## Conclusion

Our results have shown that even very short "mechanistic interventions" can lead to immediate and sizable changes in children's intuitions about complexity. Crucially, when those changes happened, they were still observable three weeks later. These results suggest that teaching mechanism early in school can directly influence students' intuitions about science and the world more broadly, even in the long term when the details are long forgotten.

## Acknowledgments

The authors would like to thank Grace Nathman, the Principal of the Quinnipiac Real World Math STEM Elementary School, as well as the teachers and all the members of the school community; with a special thanks to science teacher Stephanie White who helped us at every step of the study.

The authors also thank the National Science Foundation which funded this research (proposal DRL 1561143 to F. Keil).

Finally the authors thank all the members of the Yale Cognition and Development Lab for their helpful feedback.

## References

Ahn, W. K., Kalish, C. W., Medin, D. L., \& Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. Cognition, 54(3), 299-352.

Alter, A. L., Oppenheimer, D. M., \& Zemla, J. C. (2010). Missing the trees for the forest: a construal level account of the illusion of explanatory depth. Journal of Personality and Social Psychology, 99(3), 436.

Arbesman, S. (2016). Overcomplicated: Technology at the Limits of Comprehension. Penguin, New York.

Callanan, M. A., \& Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. Cognitive Development, 7(2), 213-233.

Chouinard, M. M., Harris, P. L., \& Maratsos, M. P. (2007). Children's questions: A mechanism for cognitive development. Monographs of the Society for Research in Child Development, i-129.

Frazier, B. N., Gelman, S. A., \& Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. Child Development, 80(6), 15921611.

Gillespie, N. M., \& Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. Cognitive Science, 28(6), 843-900.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., \& Danks, D. (2004). A theory of causal
learning in children: causal maps and Bayes nets. Psychological Review, 111(1), 3.

Gopnik, A., Sobel, D. M., Schulz, L. E., \& Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. Developmental Psychology, 37(5), 620.

Griffiths, T. L., \& Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive Psychology, 51(4), 334-384.

Griffiths, T. L., \& Tenenbaum, J. B. (2009). Theorybased causal induction. Psychological Review, 116(4), 661.

Johnson, S. G., \& Ahn, W. K. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. Cognitive Science, 39(7), 1468-1503.

Kominsky, J. F., Zamm, A. P., \& Keil, F. C. (In press). Knowing when help is needed: A developing sense of causal complexity.

Lee, V. E., \& Burkam, D. T. (2002). Inequality at the starting gate: Social background differences in achievement as children begin school. Washington, DC: Economic Policy Institute.

Mezzacappa, E. (2004). Alerting, orienting, and executive attention: Developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. Child Development, 75(5), 1373-1386.

National Institute of Child Health and Human Development Early Child Care Research Network (2003). Do children's attention processes mediate the link between family predictors and school readiness? Developmental Psychology, 39, 581-593

Osborne, J., Simon, S., \& Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. International Journal of Science Education, 25(9), 1049-1079.

Rozenblit, L., \& Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26(5), 521-562.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: how children understand the relation between perceptual causality and underlying mechanism. Developmental Psychology, 35(1), 303.

Sobel, D. M., \& Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. Developmental Psychology, 42(6), 1103.

Straatemeier, M., van der Maas, H. L., \& Jansen, B. R. (2008). Children's knowledge of the earth: A new methodological and statistical approach. Journal of Experimental Child Psychology, 100(4), 276-296.

Vosniadou, S. (2002). Mental models in conceptual development. In Model-based Reasoning (pp. 353-368). Springer US.

Walker, C. M., \& Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. Psychological Science, 25(1), 161-169.

# The Impact of Presentation Order on the Attraction Effect in Decision-making 

Jennifer S. Trueblood (jennifer.s.trueblood@vanderbilt.edu)<br>Aneesha Dasari (aneesha.dasari@vanderbilt.edu)

Department of Psychology, Vanderbilt University, $11121^{\text {st }}$ Ave S
Nashville, TN 37240 USA


#### Abstract

The attraction effect in decision-making is a famous example of how preferences are influenced by the availability of other options. One emerging hypothesis for the effect is that biases in attention influence preferences. In the past, these ideas have been explored indirectly through computational modeling and eye tracking. In the present paper, we directly manipulate attention through presentation order, presenting choice options sequentially. Our results show that presentation order has a large impact on the effect - some presentation orders enhance the effect and other orders reverse the effect. To understand these results, we fit a dynamic model, called the Multiattribute Linear Ballistic Accumulator model, to the choice and response time data. Modeling results reveal that presentation order influences the allocation of attention on the positive and negative differences between options. In sum, our results show that attention has a direct impact on the attraction effect.


Keywords: preferential choice; context effects; order effects; response time modeling; Bayesian parameter estimation

## Introduction

Everyday we make hundreds of choices. Some are seemingly trivial -- what cereal should I eat for breakfast? Others have long lasting implications -- what stock should I invest in? Despite their obvious differences, these two decisions have one important thing in common; both can be sensitive to context. That is, our preferences for existing alternatives can be altered by the introduction of new alternatives.

Context effects -- preference changes depending on the availability of other options -- have attracted a great deal of attention among consumer researchers studying high-level decision tasks. In recent work, context effects have also been shown in low-level domains such as perception (Trueblood, Brown, Heathcote, \& Busemeyer, 2013). This suggests that context effects are a general feature of human choice behavior and calls for a common theoretical explanation that applies across paradigms. One emerging hypothesis is that context effects occur because of biases in attention. When comparing options, one might pay attention to some features more than others and this in turn influences preferences. This idea has been explored using dynamic models that implement attention-weighting mechanisms such as Multi-alternative Decision Field Theory (MDFT, Roe, Busemeyer, \& Townsend, 2001), the Leaky Competing Accumulator model (LCA, Usher \& McClelland, 2004), and the Multiattribute Linear Ballistic Accumulator model (MLBA; Trueblood, Brown, and

Heathcote, 2014). In addition, Noguchi and Stewart (2014) used eye tracking to examine the role of visual attention in context effects. Their results suggest that alternatives are compared in pairs and specific patterns of gaze transitions are correlated with context effects. Further, recent work in economics has proposed that context effects might arise due to "rational inattention" (Woodford, 2012). The basic idea is that attention is a scarce resource and places constraints on the amount of information individuals can process during a decision. Taken together, this set of results strongly suggests attention is crucial to context effects. No previous work has directly manipulated attention in the attraction effect.

In the present paper, we directly manipulate attention by presenting choice options sequentially. Studies of context effects typically involve choices among three alternatives where one option is identified as the "target", one option is the "competitor", and the third option is the "decoy". For example, suppose there are two options ( X and Y ) in a choice, which are almost equally attractive. If an alternative D is introduced that is similar to alternative X , but inferior, it makes X more attractive. This is known as the attraction effect (Huber, Payne, \& Puto, 1982). In this example, X is the target, Y is the competitor, and D is the decoy. In all past studies of the attraction effect, the alternatives $\mathrm{X}, \mathrm{Y}$, and D were presented simultaneously and were visible to participants until they made a choice. In the current experiment, we presented the options $\mathrm{X}, \mathrm{Y}$, and D one at a time, thus manipulating what participants saw first, second, and last. Our goal is to understand if changes in attention (as manipulated by presentation order) influence final choices.

Our experiment uses a perceptual version of the attraction effect where participants judge which of three rectangles has the largest area, with height and width as the attributes. The experiment uses the same rectangle stimuli as Trueblood et al. (2013). Using a perceptual version of the attraction effect has a number of advantages including the ability to collect sufficient choice and response time data for computational modeling. In addition, the rectangle attraction task is well established in the literature and the results have been replicated in adults (Farmer, Warren, El-Deredy \& Howes, 2016), children (Zhen \& Yu, 2016), and non-human primates (Parrish, Evans, \& Beran, 2015).

## Experiment

## Participants

Fifty undergraduate students from Vanderbilt University voluntarily participated in this computer-based experiment
in the laboratory at the time of their choosing and received course credit for their participation.

## Methods

Participants were told that they would be shown three rectangles presented one at a time, and that they would have to choose the rectangle that they believed to have the largest area by pressing one of the three indicated keys. There was no value tied to the choice of rectangle (i.e., representation of an earned dollar amount), and participants did not receive feedback for their decisions.

Each rectangle stimulus had various dimensions of height and width, which both acted as attribute dimensions. The dimensions of the rectangles were set by numbers of pixels. The target, competitor, and decoy rectangles were determined by the following procedure. First a set of horizontally oriented rectangles, denoted H , were chosen out of a bivariate normal distribution with a mean of 50 pixels for height and a mean of 80 pixels for width with a variance of 2 pixels. This noise allowed for variation in the rectangles across trials. A second set of rectangles, denoted V , were defined in terms of H , but were oriented vertically. Specifically, the height of the V rectangles was defined as the width of the H rectangles plus a random number selected from the interval $[-2,2]$. The width of the V rectangles was then calculated so that the V and H rectangles had equal area. In half of the trials, the target rectangle $\mathrm{T}_{\mathrm{H}}$ was defined using the H rectangles and in the other half of the trials the target $\mathrm{T}_{\mathrm{V}}$ was defined using the V rectangles. Thus, the orientation of the target (i.e., horizontal or vertical orientation) was counterbalanced so that half of the trials consisted of the horizontally longer target, $\mathrm{T}_{\mathrm{H}}$, and half of the experimental trials consisted of the vertically longer target, $\mathrm{T}_{\mathrm{V}}$. The competitor rectangles, C , were defined in the opposite manner of the target rectangles so that they were given the same area but were oriented opposite to the target (i.e. vertically if the target was horizontal and horizontally if the target was vertical). The decoys used in this experiment were "range" decoys, options that are a little weaker than the target on the target's weakest attribute. Let $\mathrm{D}_{\mathrm{H}}$ denote a horizontally oriented decoy similar to $T_{H}$ and $D_{V}$ denote a vertically oriented decoy similar to $\mathrm{T}_{\mathrm{V}}$. A range decoy $\mathrm{D}_{\mathrm{H}}$ has the same width as $T_{H}$ but a shorter height since height is the shortest (weakest) dimension of a horizontally oriented target. Likewise, the $\mathrm{D}_{\mathrm{V}}$ decoy has the same height as $\mathrm{T}_{\mathrm{V}}$ but a shorter width since width is the shortest (weakest) dimension of a vertically oriented target (see Figure 1 for a schematic of the choice options). The shortest dimension of each decoy was defined as the shortest dimension of the corresponding target minus a random number selected from the interval [7,9].

Each experimental trial began with a fixation cross appearing at the center of a white screen for 0.250 ms . This was followed by the appearance of the numbers " 1 ", " 2 ", and " 3 " from left to right on the screen to indicate that one rectangle will appear above each number. Participants made their choice by pressing the corresponding " 1 ", " 2 ", or " 3 "
key at the top of the keyboard. Black rectangles were shown one at a time. Each rectangle was shown above one of the numbers for 1.0 second before disappearing. The location of each rectangle was randomized across trials. The order of appearance for the rectangles was randomized in a controlled manner, such that experimental trials of each order appeared an equal number of times.


Figure 1: Schematic of the choice options in the rectangle attraction effect task. The options H (horizontally oriented rectangles), V (vertically oriented rectangles), and decoys $D_{H}$ and $D_{V}$ are plotted in a two-dimensional attribute space defined by the logarithm of height and width. The dotted line indicates options that should objectively be indifferent because they have the same area. In the attraction effect, preference for H and V can be affected by the presence of either $D_{H}$ or $D_{V}$.

Each participant completed 720 randomized trials that were divided into eight blocks of 90 trials each. Within the 90 trials of each block, there were 30 filler trials and 10 trials for each of the possible six orders the rectangles could be presented. These six orders are as follows: TCD, CTD, TDC, CDT, DTC, and DCT. Within these orders, there were two variations, one where $T_{H}$ served as the target and one where $\mathrm{T}_{\mathrm{V}}$ served as the target to minimize effect based on orientation of the target rectangle. The 30 filler trials were meant to serve as an estimate of accuracy for participants. Each filler trial had a clearly larger rectangle that would allow participants to make a correct choice.

## Results

One participant's data were removed due to computer error. Overall, the mean accuracy of participants' performance on filler trials was $66.62 \%$ correct with two participants falling two standard deviations below average. However, these participants' data were not removed. This data is analyzed using the relative choice share for the target, or RST, which is defined as the number of target options selected divided by the total number of target plus competitor options selected (i.e., $\mathrm{T} /(\mathrm{C}+\mathrm{T})$, Berkowitsch, Scheibehenne, Rieskamp, 2014). For the results described below we
collapsed across the two different orientations of the target. The attraction effect was still observed with an average of $51.10 \%$ target chosen, significantly different from $50 \%$ target chosen, the theoretical RST if the decoy had no influence $(\mathrm{t}(48)=2.77, \mathrm{p}=0.008)$.

Although the attraction effect was not observed for each presentation order, each presented order of rectangles were also significantly different from the $50 \%$ theoretical RST, refer to Table 1. Figure 2 shows a bar graph of the RST values for each of these orders. A one-way ANOVA showed a significant main effect of order $(F(5,288)=18.27, p<$ 0.001 ). In particular, the orders CTD, CDT, and DTC showed the attraction effect, with RSTs significantly higher than $50 \%$, and the orders TCD, TDC, and DCT had RSTs significantly lower than $50 \%$ (a reverse attraction effect).

Table 1. The RST value as a percentage, the $t$-value, and the p-value for each order.

| Order | RST (\%) | $t(48)$ | $p$-value |
| :--- | :--- | :--- | :--- |
| TCD | 41.99 | -4.75853 | $<0.001$ |
| CTD | 57.62 | 4.7803 | $<0.001$ |
| TDC | 42.71 | -3.0831 | 0.003 |
| CDT | 60.25 | 4.8931 | $<0.001$ |
| DTC | 57.81 | 3.7921 | $<0.001$ |
| DCT | 45.93 | -2.1742 | 0.035 |



Figure 2: Results show the RST for each presentation order of the rectangles as well as collapsed across all orders (combined). The dotted line at 0.5 indicates equal preference for the target and competitor. Bars above the dotted line show the standard attraction effect. Bars below the dotted line show a reversed attraction. Error bars show the standard error of the mean.

## Modeling

In order to better understand how presentation order influences choices in the attraction effect, we fit the MLBA model (Trueblood et al., 2014) to the choice and response time (RT) data. MLBA is a dynamic model that explains why context effects occur in multi-alternative choice. This model explains how preferences are constructed through a dynamic process of comparing the different features of available options. Context effects occur because of differences in the amount of attention given to specific comparisons (for example, if two options are difficult to discriminate on a particular feature, an individual pays more attention to that feature).

## Model Details

MLBA is an extension of the Linear Ballistic Accumulator (LBA) model developed by Brown and Heathcote (2008). The LBA accounts for choice and RTs using independent accumulators that race toward a threshold. The accumulators are linear and accumulate information deterministically. At the beginning of each trial, each accumulator starts at a randomly determined amount of evidence drawn from a uniform distribution on the interval $[0, A]$. The accumulators increase at speeds defined by a set of drift rates, until one of the accumulators reaches the threshold $b$. The option associated with the accumulator that reaches the threshold first is selected. On each trial, the drift rates are drawn from normal distributions with different means and the same standard deviation, $s=1$. The model also has a non-decision time parameter $T_{0}$ that accounts for encoding and motor response times. The MLBA model adds to the LBA model by specifying how drift rates arise from the evaluation of choice options.

Consider three alternatives (indexed as $i, j, k$ ) that have two attributes, $P$ and $Q$, where $P_{i}$ and $Q_{i}$ denote the value of option $i$ on the two attributes. The mean drift rate $d_{i}$ for option $i$ is defined as: $d_{i}=\gamma V_{i j}+\gamma V_{i k}+I_{0}$. The term $V_{i j}$ represents a comparison between options $i$ and $j$. Likewise, $V_{i k}$ represents a comparison between options $i$ and $k$. The term $I_{0}$ is a positive constant to ensure that at least one of the three mean drift rates is positive, avoiding non-termination in the LBA model. For our purposes, we can fix $I_{0}=1$. The parameter $\gamma$ is a scaling parameter that ensures that drift rates are in the appropriate range for the LBA model.

In the valuation function $V_{i j}$, option $i$ is the focal option and option $j$ is evaluated relative to it. Let $\left(u_{P i}, u_{Q i}\right)$ and $\left(u_{P j}\right.$, $u_{Q j}$ ) be the subjective values for options $i$ and $j$. In our experiment, the attribute dimensions, $P$ and $Q$, are the height and width of the rectangles in pixels. A pair of options were experimentally defined as indifferent in they have equal area, for example, $P_{i} \times Q_{i}=P_{j} \times Q_{j}$. We define the subjective values simply as the logarithm of the number of pixels for each dimension (e.g., $u_{P i}=\log \left(P_{i}\right)$ ). Please see Trueblood et al. (2014) for other possible mappings from objective to subjective values. The valuation function $V_{i j}$ is defined by the difference in the subjective values of the options:

$$
V_{i j}=w_{P i j}\left(u_{P i}-u_{P j}\right)+w_{Q i j}\left(u_{Q i}-u_{Q j}\right)
$$

where the weights $w_{P i j}$ and $w_{Q i j}$ reflect the amount of attention given to a particular comparison.

Based on research showing that visual attention (e.g., fixation duration) increases with decreasing discriminability of items (Gould, 1967, 1973), we hypothesize that attention weights are larger when attribute values are more similar and smaller when they are more distinct. Using Shepard's (1987) law of generalization, we define the attention weights as

$$
\begin{gathered}
w_{P i j}=\exp \left(-\lambda_{+}\left|u_{P i}-u_{P j}\right|\right) \text { if } u_{P i} \geq u_{P j} \\
w_{P i j}=\exp \left(-\lambda_{-}\left|u_{P i}-u_{P j}\right|\right) \text { if } u_{P i}<u_{P j} \\
w_{Q i j}=\exp \left(-\beta \lambda_{+}\left|u_{Q i}-u_{Q j}\right|\right) \text { if } u_{Q i} \geq u_{Q j} \\
w_{Q i j}=\exp \left(-\beta \lambda_{-}\left|u_{Q i}-u_{Q j}\right|\right) \text { if } u_{Q i}<u_{Q j}
\end{gathered}
$$

where $\lambda_{+}$and $\lambda_{-}$are free parameters that allow for attention to be asymmetric. That is, the attention weights are different when comparing positive differences in attribute values (i.e., the parameter $\lambda_{+}$) and negative differences in attribute values (i.e., the parameter $\lambda_{\text {_ }}$ ). This follows from work showing that similarity judgments often violate symmetry (Tversky, 1977) as well as modifications to Shepard's law that allow for such violations (Nosofsky, 1991). The parameter $\beta$ is a bias parameter that allows for attributes to be weighted differently. For example, in consumer choice, the attribute of price might receive more weight than the attribute quality. With rectangles, Holmberg and Holmberg (1969) suggested an "elongation effect" where height plays a more important role in area judgment than width.

In summary, the MLBA has the following free parameters: accumulator start-point $A$, threshold $b$, nondecision time $T_{0}$, drift rate scaling $\gamma$, positive attention parameter $\lambda_{+}$, negative attention parameter $\lambda_{-}$, and bias $\beta$.

## Hierarchical Bayesian Parameter Estimation

We fit the MLBA model with hierarchical Bayesian parameter estimation methods using DE-MCMC (Turner, Sederberg, Brown, \& Steyvers, 2013). We note that, as far as we are aware, this is the first time the MLBA (or any dynamic model of context effects) has been fit to both choice data and the full distribution of RT data. In the past, dynamic models of context effects have only been evaluated by qualitative measures or when quantitative fitting was performed, only choice data was used. Thus, we see the present work as a significant methodological step forward in the evaluation of dynamic models of context effects.

In our experiment, there are six order conditions: TCD, CTD, TDC, CDT, DTC, and DCT. We hypothesized that order would influence attention and thus we had separate attention parameters $\lambda_{+}$and $\lambda_{-}$for each condition. We also fit six $\gamma$ scaling parameters, one for each condition. We allowed for different scaling parameters across conditions to accommodate the different attention weights, which directly
impact the magnitude of the drift rates. The remaining parameters were assumed to be the same across conditions.

In our model, we had both group-level (or hyper parameters) and individual-level parameters. The individual-level parameters were drawn from normal distributions defined by the hyper parameters. Let $\mu_{x}$ and $\sigma_{x}$ represent the hyper mean and standard deviation of the group-level normal distribution for parameter $x$. The priors for the hyper means were the following: $\mu_{b} \sim \mathrm{~N}(1,0.5), \mu_{A} \sim$ $\mathrm{N}(1,0.5), \mu_{T 0} \sim \mathrm{~N}(0.25,0.25), \gamma \sim \mathrm{N}(5,1.5), \lambda_{+} \sim \mathrm{N}(0.5,1.5)$, $\lambda_{-} \sim \mathrm{N}(0.5,1.5), \beta \sim \mathrm{N}(1,1.5)$. The priors for all of the hyper standard deviations $\sigma_{x}$ were defined as $\operatorname{Gamma}(1,1)$ distributions expect for the standard deviation for nondecision time, which was $\operatorname{Gamma}(1,0.5)$. We ran 24 MCMC chains for 2500 iterations with a burn-in of 500 iterations. All chains converged.

## Results

To assess the fit of the model, we calculated the correlation between choice and mean RT data with model predictions. The model predictions were calculated by using the mean of the posterior distributions of the individual parameters. The correlation between the choice data and model predictions was $0.886(p<0.001)$. The correlation between the mean RT data and the model predictions was 0.588 ( $\mathrm{p}<0.001$ ). Thus, the model does a good job at capturing general trends in the data.
We also examined how well the model accounted for the average choice data for each condition. Figure 3 shows the mean choice proportions for each option in the 12 different choice sets used in the experiment. The 12 choice sets arise from the two possible placements of the decoy $\left(D_{H}\right.$ or $\left.D_{V}\right)$ in each of the six order conditions. The model predictions were calculated using the mean of the posterior distributions of the individual parameters.

To understand how presentation order influences choices in the attraction effect, we examined the values of the attention weights for the six conditions (see Table 2). Specifically, we examined the posterior means of the grouplevel attention weight parameters ( $\lambda_{+}$and $\lambda_{-}$) for each condition. We did not see any obvious trends in the attention weights when we examined them individually. However, the ratio of the positive weight to the negative weight revealed an interesting pattern. In the conditions that exhibited the standard attraction effect (i.e., CTD, CDT, and DTC), the ratio was smaller than the conditions that exhibited an inverse attraction effect (i.e., TCD , TDC, DCT ). This suggests that presentation order influences the amount of attention given to positive and negative differences in attribute values. When the attraction effect is observed, more attention is placed on negative differences as compared to when the reverse attraction effect occurs.

## Discussion

Our goal in the present paper was to explore the role of attention in the attraction effect through direct manipulation. We manipulated attention through presentation order,


Figure 3. Observed choice proportions and model predictions for 12 choice sets in the rectangle attraction effect task. Each choice set consists of three options (the target, competitor, and decoy). There are two choice sets for each order condition due to the two possible placements of the decoy (either near the horizontally oriented rectangle or the vertically oriented rectangle). The model predictions are shown in light gray and observed choice proportions in dark gray.
presenting the options sequentially rather than simultaneously. The sequential presentation of the options had a large impact on choices - some presentation orders enhanced the attraction effect whereas other presentation orders reversed the attraction effect. To better understand why presentation order impacted choices, we used computational modeling. We fit the MLBA model to choice and response time data. Model fits revealed differences in the attention weights for different presentation orders. For the presentation orders that showed a standard attraction effect, there was increased attention on negative differences as compared to the presentation orders that showed a reverse attraction effect.

Recently, researchers have discovered large individual differences in context effects (Liew, Howe, \& Little, 2016; Trueblood, Brown, \& Heathcote, 2015). Some individuals show the standard effects, but others do not. For some individuals, the effects are even reversed. This has lead to the conclusion that context effects are fragile (Trueblood et al., 2015). This raises two important questions: (1) Why are
the effects fragile? and (2) What underlies individual differences in the effects? The present work provides one possible explanation. The effects are fragile because they result from biases in attention. Small shifts in attention can have dramatic influences on choice. It is possible that individual differences in the effects arise because of individual differences in attention.

Table 2: Posterior means of the group-level attention weight parameters for the six order conditions.

| Condition | $\lambda_{+}$ | $\lambda_{-}$ | $\lambda_{+} / \lambda_{-}$ |
| :--- | :--- | :--- | :--- |
| TCD | 1.50 | 4.25 | 2.83 |
| CTD | 2.05 | 4.51 | 2.20 |
| TDC | 0.92 | 2.06 | 2.24 |
| CDT | 1.39 | 2.52 | 1.81 |
| DTC | 1.86 | 3.43 | 1.84 |
| DCT | 1.11 | 2.48 | 2.23 |

The results of our experiment also pose a challenge to a recent rational model of context effects that claims the effects are a consequence of expected value maximization given noisy observations (Howes, Warren, Farmer, ElDeredy, \& Lewis, 2016). In our experiments, simply changing the presentation order of the same set of options has a dramatic influence on choices. It is unclear how a rational model could account for the influence of presentation order on the effects.

In sum, we have demonstrated that presentation order, which influences attention, can both strengthen and weaken the attraction effect. The MLBA model suggests that presentation order changes the allocation of attention between positive and negative differences between options. These findings provide an explanation for individual differences in context effects and also pose a challenge to recent rational models of the effects.

## Acknowledgments

AD was supported by an NSF Research Experience for Undergraduates students associated with grant SES1556325. JST was supported by NSF grant SES-1556325.

## References

Berkowitsch, N. A., Scheibehenne, B., \& Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. Journal of Experimental Psychology: General, 143(3), 1331-1348.
Brown, S. D., \& Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. Cognitive psychology, 57(3), 153-178.
Farmer, G. D., Warren, P. A., El-Deredy, W., \& Howes, A. (2016). The Effect of Expected Value on Attraction Effect Preference Reversals. Journal of Behavioral Decision Making.
Gould, J. D. (1967). Pattern recognition and eye-movement parameters. Perception \& Psychophysics, 2, 399-407.
Gould, J. D. (1973). Eye movements during visual search and memory search. Journal of Experimental Psychology, 98, 184-195.
Holmberg, L., \& Holmberg, I. (1969). The perception of the area of rectangles as a function of the ratio between height and width. Psychological Research Bulletin, 9, 1-6.
Howes, A., Warren, P. A., Farmer, G., El-Deredy, W., \& Lewis, R. L. (2016). Why contextual preference reversals maximize expected value. Psychological review, 123(4), 368.

Huber, J., Payne, J. W., \& Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. Journal of Consumer Research, 9(1), 90-98.
Liew, S. X., Howe, P. D., \& Little, D. R. (2016). The appropriacy of averaging in the study of context effects. Psychonomic bulletin \& review, 23(5), 16391646.

Noguchi, T., \& Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are
repeatedly compared in pairs on single dimensions. Cognition, 132(1), 44-56.
Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. Cognitive Psychology, 23, 94-140.
Parrish, Audrey E., Theodore A. Evans, and Michael J. Beran. "Rhesus macaques (Macaca mulatta) exhibit the decoy effect in a perceptual discrimination task." Attention, Perception, \& Psychophysics 77, no. 5 (2015): 1715-1725.

Roe, R. M., Busemeyer, J. R., \& Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. Psychological review, 108(2), 370-392.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.
Trueblood, J. S., Brown, S. D., Heathcote, A., \& Busemeyer, J. R. (2013). Not just for consumers context effects are fundamental to decision making. Psychological science, 24(6), 901-908.
Trueblood, J. S., Brown, S. D., \& Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. Psychological review, 121(2), 179-205.
Trueblood, J. S., Brown, S. D., \& Heathcote, A. (2015). The fragile nature of contextual preference reversals: Reply to Tsetsos, Chater, and Usher (2015). Psychological review, 122(4), 848-853.
Turner, B. M., Sederberg, P. B., Brown, S. D., \& Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. Psychological Methods, 18(3), 368-384.
Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.
Usher, M., \& McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. Psychological Review, 111, 757-769.
Woodford M (2012) Prospect theory as efficient perceptual distortion. The American Economic Review, 102, 41-46.
Zhen, S., \& Yu, R. (2016). The development of the asymmetrically dominated decoy effect in young children. Scientific reports, 6.

# Interpretation and Processing Time of Generalized Quantifiers: Why your Mental Space Matters 

Alice Ping Ping Tse (tse@tf.uni-freiburg.de) and Marco Ragni (ragni@tf.uni-freiburg.de)<br>Institut für Informatik, Technische Fakultät, Albert-Ludwigs-Universität Freiburg<br>Georges-Köhler-Allee, Geb. 052, 79110 Freiburg im Breisgau, Germany


#### Abstract

Classical quantifiers (e.g., "all", "some" and "none") have been extensively studied in logic and psychology. In contrast, generalized quantifiers (e.g., "most") allow for fine-grained statements about quantities. The discrepancy in the underlying mental representation and its interpretation among interpreters can affect language use and reasoning. We investigated the effect of quantifier type, quantification space (set size) and monotonicity on processing difficulty (in response time, RT) and response diversity of 77 generalized quantifiers. Shannon entropy was employed to measure response diversity. Our findings indicate: (i) Set size is a significant factor of response diversity, which implies that the underlying space is relevant for the interpretation. (ii) Quantifiers possess a rather static underlying representation within and across tasks within a participant. (iii) Quantifier type and monotonicity can affect response diversity; while the response diversity can predict RT. (iv) In reasoning, the number of generalized quantifiers versus classical quantifiers in a syllogism is a factor of response diversity. Diversity in the interpretation of generalized quantifiers may be a cause of human's deviation from logical responses.


Keywords: generalized quantifiers; syllogism; total set size; monotonicity, individual differences

## Introduction

"Quantifiers" can have two definitions: In logic, a quantifier acts as the binder to denote the relationships between sets. In natural language, a quantifier is a determiner or pronoun indicative of quantity or amount. In daily English, it limits and modifies the quantity of the noun it is attached to. They map categories to types. Hence, they are the basis for many fundamental concepts in different fields, especially logic, linguistics and psychology. In first-order logic, there are only two basic quantifiers: the universal "for all, $\forall$ " and the existential quantifier "there exist (or for some), $\exists$ ", which denote quantities. In Aristotelian logic (Austin, et al., 1971; Westerståhl, 2011), there are three quantifiers, "all/every", "some" (also for "some... not"), and "no". However, the aforementioned first-order quantifiers are too restricted in daily language use. Generalized quantifiers (also known as the second-order predicates or binary quantifiers) are in a wider use in language, for example, when the exact amount is not available (which is quite usual in daily situations) or to emphasize a rather qualitative property of the amount (e.g., "more than half", "most", "a few"). Generalized quantifiers (or just quantifiers) include words and phrases like 'most', 'many', 'few', 'a few', 'some', 'more than half', 'commonly', 'typically' and cardinal numbers (e.g., more than one, and exact numbers such as two, a hundred).

Since the first articles by Barwise and Cooper (1981) in the field of linguistics and Lindström (1966) in the field of logic, an increasing number of research articles have focused on generalized quantifiers. The interpretation of generalized quantifiers can be affected by factors like the quantification space - its total set size (Newstead, Pollard, \& Riezebos, 1987), word frequency (Chase, 1969), monotonicity (Szymanik \& Zajenkowski, 2013), common belief and background knowledge (Newstead \& Collis, 1987; Moxey \& Sanford, 1993), and context and working memory (Zajenkowski, Szymanik, \& Garraffa, 2014). Some psychological studies (e.g., Newstead et al., 1987; Ragni, Eichhorn, Bock, Kern-Isberner, \& Tse, 2017) have demonstrated that many quantifiers do not have a precise true/false cutoff for the quantity they represent, on a scale from 0 to 100 . Even more, the minimum and maximum values of human's subjective valuation responses to individual quantifier can vary a lot (Ragni et al., 2017). This may hint at a fuzzy underlying space of quantifiers among people, in terms of using and interpreting quantifiers (Budescu \& Wallsten, 1985).

Generalized quantifiers have been recently employed in studies of syllogisms ${ }^{1}$, such as the Probability Heuristics Model (PHM; Oaksford \& Chater, 1994). A recent study (Ragni, Singmann, \& Steinlein, 2014) has extended three syllogistic reasoning theories (Matching Hypothesis, Mental Model Theory and Preferred Mental Models) to generalized quantifiers. However, only the two quantifiers - "most" and "few" were included. The interpretation of generalized quantifiers plays a role in most reasoning theories, especially regarding the set relationship. For example, in mental model theory, it is the basis for the construction of mental models. Endorsement of invalid conclusion can be resulted when reasoners commit the illicit conversion fallacy of interpreting "All As are Bs" as equivalent to "All Bs are As" and make a mistake in the initial mental model construction. This is indeed a very common error in syllogistic tasks. While for PHM, it is relevant to the selection of the preferred quantifier in the conclusion.

Knowing the factors affecting the underlying representation of generalized quantifiers is essential for cognitive theories for reasoning. One example is that "most As are

[^524]Bs" is "equivalent" to "most Bs are As" if A and B have the same set size. However, if B has a much larger set size or cardinality than A, then the proposition does not hold after the switch (also known as illicit conversion). About $50 \%$ of the participants in the experiments of a previous study chose "most" as the conclusion quantifier for a syllogism with "most" as the quantifier for both premises ${ }^{2}$ (Chater \& Oaksford, 1999). It is interesting that half of the participants considered A, B and C as having the same cardinality while the other half of the participants may not. About $15 \%$ of the participants in experiment 1 and $20 \%$ of the participants in experiment 2 chose "no valid conclusion" as their responses. It is very possible that these participants may be aware of the fact that the differences in total set sizes of $A, B$ and $C$ can lead to different conclusions for the syllogism.

Factors affecting the variety in the interpretation of quantifiers and underlying space have to be controlled in studies employing these quantifiers. Besides, are there individual differences in the underlying representation of a quantifier? Does the underlying space affect the response diversity? The answers may provide insight for the questions why human participants do not always draw the same or logical conclusions but some particular irrational conclusions are preferred and why more response diversity was found for more difficult reasoning problems (e.g., Khemlani \& Johnson-Laird, 2012). Also, would the degree of vagueness/uncertainty of a quantifier cause more individual differences? What are the factors of the response diversity in the interpretation of generalized quantifiers? Will the degree of uncertainty cause a larger processing difficult which can be reflected by a longer processing time? And what are the factors of processing difficulty of generalized quantifiers? This analysis investigates these questions regarding the underlying space and processing time (difficulty) of generalized quantifiers. More precisely, we focus on three levels of tasks: 1. Spontaneous valuation of the quantity or frequency the quantifier represents in the Subjective Valuation Task. 2. A Truth Judgement Task in which participants were asked to judge if a quantified statement holds true for a picture. 3. Finally, a syllogistic reasoning task with generalized quantifiers that participants were asked to reason and derive a conclusion from two premises of quantified statements. Please refer to the Method section for details about the three tasks.

Two measures of response diversity were employed in this study, namely the standard deviation and Shannon entropy. In information theory, Shannon entropy calculates the expected value of the information transmitted in a message (Shannon \& Weaver, 1949), as a function of the probability of the occurrence of each possible message. For each quantifier, the entropy in the Truth Judgement Task, was calculated by the aggregated normalized probabilities of the truth responses for each of the pictures/scenarios presented (see Method for the details) by the Shannon equation: $-\sum \mathrm{p}_{\mathrm{i}} \log _{2} \mathrm{p}_{\mathrm{i}}$, where $p_{i}$ is the probability of a truth response. There will be

[^525]several "small" probabilities if the response is more diverse. Conversely, if the responses are condensed to a few values, the probabilities of these selected values will be high. The smaller the probability, the larger the entropy value calculated by the equation. And thus, a larger Shannon entropy value indicates more discrepancy in the responses. It was used to measure the response diversity in syllogistic reasoning in a meta-analysis study (Khemlani \& Johnson-Laird, 2012). Standard deviations of the responses in the Subjective Valuation Task and Truth Judgement Task were calculated as well to check if the two measures of diversity agree with each other. For answering the question regarding underlying space, Newstead et al. (1987) found that the amount of entities represented by certain quantifiers (e.g., "some") could be affected by the assumed total set size of the experimental scenario. However, does this hold for all generalized quantifiers? Besides the extra-linguistic factor of total set size, several properties of the quantifier itself can also affect the diversity of human responses (interpretation) and response time. They include, among others, quantifier type, monotonicity, and word frequency. We will elaborate two aspects below.

## Quantifier Types

There are many different ways of classifying quantifiers (e.g., logical quantifiers versus different types of binary quantifiers; simple versus complex quantifiers). In this study, the quantifiers were classified in the sense of natural language, namely frequency versus quantity quantifiers. Many studies have been conducted for quantity quantifier, while there are only few for frequency ones. For instance, Newstead and Collis (1987) studied the context effect in the interpretation of ten frequency quantifiers. In contrast to some previous findings for quantity quantifiers (Chase, 1969; Newstead \& Griggs, 1984; Newstead et al., 1987), no significant set size effect or effect due to the presence of other quantifiers were found. This supported that processing of quantifiers of different types may be different due to their specific properties.

## Monotonicity

A generalized quantifier $\mathrm{Q}_{\mathrm{up}}$ is upward monotone/entailing (or monotone increasing) if and only if for all $M$ and all A , $\mathrm{B} \subseteq \mathrm{B}^{\prime} \subseteq \mathrm{M}, Q_{M}(\mathrm{~A}, \mathrm{~B})$ implies $Q_{M}\left(\mathrm{~A}, \mathrm{~B}^{\prime}\right)$. That means $\mathrm{Q}_{\mathrm{up}}$ license the inference from subsets to supersets. Similarly, a $\mathrm{Q}_{\text {down }}$ is downward monotone/entailing if and only if for all $M$ and all $\mathrm{A}, \mathrm{B} \subseteq \mathrm{B}^{\prime} \subseteq \mathrm{M}, Q_{M}\left(\mathrm{~A}, \mathrm{~B}^{\prime}\right)$ implies $Q_{M}(\mathrm{~A}, \mathrm{~B})$. Contrastive to $\mathrm{Q}_{\mathrm{up}}, \mathrm{Q}_{\text {down }}$ license the inference from supersets to subsets. For example, "Some men are Germans implies some men are Europeans". With the fact that "Germans" is within the set of "Europeans", "some" is an upward monotone quantifier. Similarly, "No men are birds" implies "No men are eagles". With the fact that eagles are birds, "no" is downward monotone. There are nonmonotone quantifiers also, e.g., "exactly three". For example, "exactly three men are Germans" does not imply "exactly three men are Europeans" or vice versa. According to
the definition, many natural language quantifiers are (either upward or downward) monotone, including the three Aristotle quantifiers, "all", "some" and "no". Barwise (1981) suggested that monotone quantifiers are easier to process than non-monotone ones.

## Aims of the Study and Research Questions

We aimed to examine human's interpretation of a large number of generalized quantifiers and factors affecting the response diversity and processing time of these quantifiers to facilitate further studies of generalized quantifiers in different fields. As mentioned before, the significant properties have to be controlled in studies involving these quantifiers in order to eliminate some confounding factors. The analyses focus on three factors, namely total set size, quantifier type (quantity quantifier and frequency quantifier), and monotonicity (upward, downward and non-monotone), according to two domains, namely degree of variation in underlying representation space (among interpreters, in terms of response diversity) and processing difficulty (in terms of processing time) of the generalized quantifiers.

## Research Question 1: Factors of Response Diversity

Unlike "All", "No" or "Seven" (numerical quantifiers), the amount (or proportion) represented by most generalized quantifiers can be rather fuzzy. Humans do not agree with each other regarding the representation space of individual quantifier. What are the factors affecting the differences in the underlying representation space of quantifiers? In other words, are total set size, quantifier type and monotonicity the factors affecting response diversity? Research question 1 (RQ 1) was examined by analyzing the standard deviation (SD) and Shannon entropy measures in both the Subjective Valuation and Truth Judgement Tasks. We hypothesized that the smaller set size condition, quantity quantifiers and upward monotone quantifiers may exhibit smaller response diversities, i.e., smaller standard deviation and entropy measure values.

## Research Question 2: Processing Time

Does greater degree of fuzziness cause a longer processing time (in terms of response time)? Besides, is the monotonicity a factor of processing time as well? Szymanik and Zajenkowski (2013) found a significant interaction effect of monotonicity and the truth value of the quantified statement in a verification task of four quantifiers but failed to find a significant main effect of monotonicity. We extended the study with more quantifiers of different quantifier types. Word frequency was included as a covariant because it has a general effect in word recognition ${ }^{3}$. Quantifiers of higher

[^526]word frequency are expected to be processed faster. We hypothesized that the entropy measures and word frequency are significant predictors of RT; and the quantifier type may affect the RT as well.

## Method

## Participants

104 native English speakers ( $\mathrm{M}=40.8$ years; range $=21-75$ years; 63 females) participated in the online experiment on Amazon Mechanical Turk. We controlled for one participant from a given computer. They received a nominal fee.

## Materials, Design and Procedure

A search for common quantity and frequency quantifiers was performed in Google with the keywords "quantifiers", "frequency quantifiers", "frequency adverbs", "determiners", "how often", "how many", and "how much". 77 generalized quantifiers were selected ${ }^{4}$, with 34 frequency quantifiers (frequency adverbs); and 14,13 and 16 quantity quantifiers which can be used with countable, uncountable and both countable and uncountable nouns (type-both) respectively. Each participant had to perform two tasks:

A Subjective Valuation Task Participants were asked to provide a subjective value of the amount the quantifier represents. They had to move a slidebar to indicate their responses in terms of percentages (from $0 \%$ to $100 \%$ ). Each quantifier was evaluated once.

A Truth Judgement Task Participants had to evaluate the validity of a quantified statement presented above a picture. For the effect of total set size, participants were randomly assigned to either the smaller-set group or larger-set group. The number of participants in each group was counterbalanced. For countable and type-both quantifiers, pictures of 10 circles or 100 circles were displayed, with 0 to all of them colored black (see Fig. 1). While for uncountable and type-both quantifiers, pictures of a heap of sand or desert (composed of 10 heaps of sand) were presented with 0 to $100 \%$ of the sand or desert colored brown (see Fig. 2).


Figure 1: Pictures for countable space in the Truth Judgement Task. Participants received the left (10 circles) or right picture ( 100 circles) and had to evaluate whether a quantified statement like "Some circles are black" (presented at the left hand upper corner) is a true description of the picture or not.

[^527]For frequency quantifiers, timelines of a week with a coffee cup icon for 0 to 7 days or a monthly schedule with an icon of football for 0-31 days were presented. Each quantifier was tested 2 times for each participant in two blocks. Pictures displayed were counterbalanced within participants in the sense that if the participant received less than or equal to half positive situation (e.g., 3 out of 10 circles colored black) in the first block, he/she would receive more than or equal to half positive situation in the second block (e.g., 7 out of 10 circles are black) and vice versa. The same manipulation was applied to all the three scenarios (circles, sand/desert, and timeline/calendar). Type-both quantifiers were tested four times for each participant as they were presented twice in both the circle and sand/desert scenarios. The possible picture options were counterbalanced among participants. For countable and type-both quantifiers, the statement was in the form of "Quantifier (of the) circles are black" or "These circles are Quantifier black" (for "commonly" and "typically"). For the sand/desert situation, the statement was "Quantifier sand is colored brown". The statement was in the form of "Tim Quantifier drinks coffee" or "Tim drinks coffee Quantifier" for the weekly timeline scenario; and "Tim Quantifier plays soccer" or "Tim plays soccer Quantifier" for the monthly calendar scenario. Participants were asked to judge as accurately and quickly as possible whether the statement was a truth description of the picture. Participants always performed the Subjective Valuation Task first.


Figure 2: Pictures for an uncountable space in the Truth Judgement Task. Participants had to judge whether a statement like "Most sand is colored brown" is a true description of the picture or not.

## Results

## The Underlying Representation Space

For the first research question, the diversity in the responses was evaluated by the Shannon entropy and standard deviation measures of the responses in both the Truth Judgement Task and Subjective Valuation Task, as the indices. The standard deviation and entropy measures of the responses were calculated according to the two different set size conditions in the Truth Judgement Task (SD1, Entropy1; and SD2, Entropy2). SD1 and Entropy1 are the standard deviation and entropy measure of the smaller set size pictures (10 circles, 1 heap of sand and weekly timeline). SD2 and Entropy 2 are the standard deviation and entropy measure of the larger set size condition ( 100 circles, desert and monthly schedule). The Spearman's rank correlations between the three SDs and entropy measures were tested both within and
across the two tasks. Significant correlations were found except for SD2 with the entropy in the Subjective Valuation Task and Entropy1, see table 1 for the results. For the effect of set size on the response diversity, significant differences were found between both SD1 and SD2; and Entropy1 and Entropy2, $\mathfrak{t}(76)=-6.142, p<.001$ and $\mathfrak{t}(76)=-7.268, p<$ .001, respectively, with SD2 and Entropy2 being significantly larger. The two measures (SD and entropy) provided similar results, as the entropy measures were more reliable indices for the response diversity (according to the positive correlations across tasks), we used the entropy measures for the following analyses for the sake of simplicity.
Regarding the property of monotonicity, the quantifiers were classified into 28 upward monotone, 23 downward monotone, 11 monotone and 12 non-monotone quantifiers. The effects of the three quantifier properties on the two entropy measures (for underlying space) in the Truth Judgement Task were then tested. The 2 (quantifier type: frequency and quantity) x 4 monotonicity (monotonicity: upward, downward, monotone and non-monotone) MANOVA, with word frequency as a covariate, for the two entropy measures showed that the quantifier type had a significant multivariate effect on the two entropy measures, $\mathrm{F}(2$, $64)=179.820, p<.001$, Wilk's $\lambda=.151, \eta_{\mathrm{p}}{ }^{2}=.849$; as well as monotonicity and word frequency, $\mathrm{F}(6,128)=5.568, p$ < .001 , Wilk's $\lambda=.629, \eta_{\mathrm{p}}^{2}=.207$ and $\mathrm{F}(2,64)=10.130, p<$ .001 , Wilk's $\lambda=.760, \eta_{\mathrm{p}}^{2}=.240$, respectively. The interaction effect of quantifier type and monotonicity was not significant.

The following post-hoc tests were performed according to quantifier type (frequency versus quantity) and monotonicity (upward and downward). The t -tests showed that the two entropy measures (Entropy1 and Entropy2) were reliably different for frequency quantifiers, $\mathrm{t}(33)=-23.426, p<$ .001 , but not for quantity quantifiers. Regarding the monotonicity, the two entropy measures were reliably different for upward and downward monotone quantifiers, Entropyl: $\mathrm{t}(49)=2.328, p=.024$; Entropy2: $\mathrm{t}(49)=3.198, p=.002$.

Differences between the response diversity indices for the first half and second half of the tasks were also examined. Regarding Entropy1 and Entropy2, both t-tests were not significant, Entropy1: $\mathrm{t}(76)=1.236, p=.220$, Entropy2: $\mathrm{t}(76)=.455, p=.650$. For the Subjective Valuation Task, there was no difference between the first and second half of the task neither, $\mathrm{t}(73)=-.067, p=.946$.

## Processing Time

We filtered out the response times which exceed average RT +/- 2 SD according to individual participant. Firstly, a stepwise regression was performed to test if the word frequency and the three entropy measures significantly predicted the response time. The results of the regression analysis showed that the two entropy measures in the Truth Judgement Task explained $38.6 \%$ of the response time (adjusted $\mathrm{R}^{2}=.386$, $\mathrm{F}(2,74)=24.928, p<.001$, Entropy1: $\beta=.975, p<.001$, Entropy2: $\beta=-.603, p<.001$. As the response times of the two quantifier types were significantly different,

Table 1: Results of Spearman's rank correlation of the standard deviation (SD) and entropy measures of the responses in the Subjective Valuation Task (SD and entropy) and Truth Judgement Task (SD1 and SD2; and Entropy1 and Entropy2).

$\mathrm{t}(75)=7.188, p<.001$, the regression was repeated according to the two quantifier types. For frequency quantifiers, word frequency was the only significant predictor, adjusted $\mathrm{R}^{2}=.157, \mathrm{~F}(1,32)=7.128, p=.012 ; \beta=-.427, p=.012$. While for quantity quantifiers, entropy in the Subjective Valuation Task was the only significant predictor of the response time, adjusted $\mathrm{R}^{2}=.136, \mathrm{~F}(1,41)=7.609, p=$ $.009 ; \beta=.396, \quad p=.009$. The effect of monotonicity on response time was not significant. Do generalized quantifiers affect the response diversity in syllogistic reasoning? Syllogisms are chosen as quantifiers are the essence of syllogistic reasoning and so their effect may be most visible.

## Entropy in Reasoning with Generalized Quantifiers: Additional Empirical Support

We reanalyzed the data from Ragni et al. (2014) with the entropy measure for response diversity. Twenty-five native English speakers participated in the online experiment on Amazon Mechanical Turk. Each participant had to solve 40 syllogistic problems with at least one of "most" and "few" being the quantifier of one of the two premises. Participants had to choose the conclusion quantifier of the syllogism among the four classical Aristotle quantifiers and the two generalized quantifiers "most" and "few" (i.e., all, no, some, some...not, most and few), to the question "what follows?" after reading the two premises. The conclusion direction presented (a-c or c-a) was counterbalanced. 20 problems were tested for each conclusion direction. "Most" and "few" appeared in the first premise respectively in 6 of the syllogisms, with the second premise being one of the six quantifiers ( $6 \times 2=12$ problems). For the 8 remaining syllogisms, "most" and "few" appeared in the second premise, with the first premise being one of the four Aristotle quantifiers.

The entropy measure of the responses for each syllogism was calculated and an ANOVA and a t-test were performed. The 2 (conclusion direction: a-c vs. c-a) x 2 (position of the generalized quantifier: first premise vs. second premise) ANOVA showed a significant main factor of the position, $\mathrm{F}(2,39)=4.738, p=.015, \eta_{\mathrm{p}}^{2}=.218$, but both conclusion direction and the interaction effect were not significant. Post-hoc analysis showed that syllogisms with generalized quantifier in the first premise had a significantly higher entropy, $\mathrm{t}(30)=2.174, p=.038$ (2-tailed). The number of
generalized quantifier affects the entropy as well. If both premises contained generalized quantifiers, the entropy was marginally smaller, Independent Samples Test: $t(38)=$ $1.957, p=.058$ (2-tailed) ${ }^{5}$. The marginal result might due to the fact that only 8 syllogisms have two generalized quantifiers but 32 problems have only one generalized quantifiers.

## General Discussion

While extensive research in psychology of reasoning and logic has dealt with the four classical quantifiers ("all", "some", "some...not", and "none"), few cognitive reasoning theories for syllogisms have been extended to generalized quantifiers - and often to "most" and "few" only. Different quantifiers possess different specific properties which affect their interpretation (especially in terms of interpretation diversity) in daily language. For example, for universal quantifiers like "All" and "No", most participants would select $100 \%$ and $0 \%$, respectively, in the Subjective Valuation Task, with few selecting values within $95 \%$ to $99 \%$ and $0 \%$ to $5 \%$, respectively. In contrast, the more "fuzzy" generalized quantifiers elicit a greater diversity in the responses. For example, for "some", we got 6 responses for $20 \%$ and $35 \%$, 5 responses for $25 \%$, and 10 responses for $45 \%$ (among 104 responses). In total, 47 different percentages (out of 101 possible choices) were selected as the responses in the Subjective Valuation Task. It seems that the right tool to analyze the interpretation diversity is missing. We argue that Shannon entropy, which was developed for communication, is an excellent method which can be employed to measure the response diversity of generalized quantifiers.

Using Shannon's entropy to measure response diversity was introduced in this study as it is a binary-based element which fits the dichotomous experimental design of the Truth Judgement Task. It shows reliable correlated results with the classical standard deviation measure within and across tasks. Shannon entropy seems a better measure for response diversity across tasks than the SD. Our results show that the total set size, quantifier types (frequency versus quantity) and monotonicity can affect the interpretation diversity of a quantifier; while the interpretation diversity (in terms of

[^528]entropy measures) can in turn affect the response time. In accordance with the findings of Szymanik and Zajenkowski (2013), we did not find a reliable effect of the monotonicity of quantifiers on RT. One can speculate that the difference in processing difficulty applies to cardinal quantifiers only.

The smaller set size condition has a smaller entropy value as hypothesized, in contrast to the frequency quantifiers and downward monotone quantifiers. Further studies are required to explain this finding. Quantity quantifiers have a slower RT in general and this might be affected by the larger discrepancy in the underlying representation space, which hints a fuzzier underlying representation of quantity quantifiers among participants. However, our results suggest that quantifiers possess a rather static underlying mental representation space within participants, not changing within or across tasks, as there is no difference for the response diversity measures between the first and second half of the tasks.

Despite our finding of total set size being a factor in the interpretation of generalized quantifiers, it is still possible for human to interpret quantifiers without the knowledge of total set size (Van Tiel \& Geurts, 2013). But we can speculate that participants usually represent the underlying set by a default mental model for the respective quantifiers.

Our study shows that total set size, quantifier type, and monotonicity (and word frequency) are all contributing to the possible diversity in the use or reasoning of generalized quantifiers. Based on these factors, natural extensions of theories which already assume models of different sizes and are analogous representations of the state of affairs (like the mental model theory) to incorporate the proposed results is possible. Extension to generalized quantifiers is increasingly important for cognitive reasoning theories to avoid a selfcentered focus, which renders them ultimately useless for explaining or predicting complex everyday communication. Large-scale studies of these generalized quantifiers in reasoning tasks can test if the diversity in the interpretation of these quantifiers is the factor of the response diversity in reasoning tasks. It is possible that differences in the interpretation of the quantifier contribute to the deviation from logical responses, other than reasoning/heuristic processes. Controlling the above significant factors is important for studies involving quantifiers, to avoid hidden experimental confounds. Also, for theories with predictions on response time, it is possible that interpretation diversity is a significant factor. Further studies on this hypothesis are necessary.

## Acknowledgement

This work is supported by DFG-Projects under grant numbers RA 1934/2-1, RA 1934/3-1, and RA 1934/4-1.

## References

Austin, J. L., Strawson, P. F., Grice, H. P., Chomsky, N., Katz, J. J., Goodman, N., et al. (1971). The philosophy of language (Vol. 39). London: Oxford University Press.
Barwise, J., \& Cooper, R. (1981). Generalized Quantifiers and Natural Language. Linguistics and Philosophy, 4(2), 159-219.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., \& Böhl, A. (2011). The word frequency effect. Experimental psychology, 58(5), 412-424.
Budescu, D. V., \& Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. Organizational Behavior and Human Decision Processes, 36(3), 391-405.
Chase, C. I. (1969). Often is where you find it. American Psychologist, 24(11), 1043.
Chater, N., \& Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. Cognitive psychology, 38(2), 191-258.
Khemlani, S., \& Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. Psychological Bulletin, 138(3), 427-457.
Lindström, P. (1966). First order predicate logic with generalized quantifiers. Theoria, 32(3), 186-195.
Moxey, L. M., \& Sanford, A. J. (1993). Prior expectations and the interpretation of natural language. European Journal of Cognitive Psychology, 5(1), 73-91.
Newstead, S. E., Pollard, P., \& Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. Applied ergonomics, 18(3), 178-182.
Newstead, S., \& Collis, J. M. (1987). Context and the interpretation of quantifiers of frequency. Ergonomics, 30(10), 1447-1462.
O'Malley, S., \& Besner, D. (2008). Reading aloud: Qualitative differences in the relation between stimulus quality and word frequency as a function of context. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6), 1400-1411.
Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., \& Tse, A. P. P. (2017). Formal Nonmonotonic Theories and Properties of Human Defeasible Reasoning. Minds \& Machines, 27(1), 37-77.
Ragni, M., Singmann, H., \& Steinlein, E. M. (2014). Theory Comparison for Generalized Quantifiers. Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 1330-1335). Austin, TX: Cognitive Science Society.
Segui, J., Mehler, J., Frauenfelder, U., \& Morton, J. (1982). The word frequency effect and lexical access. Neuropsychologia, 20(6), 615-627.
Shannon, C. E., \& Weaver, W. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.
Szymanik, J., \& Zajenkowski, M. (2013). Monotonicity has only a relative effect on the complexity of quantifier verification. Proceedings of the 19th Amsterdam Colloquium, (pp. 219-225). Amsterdam.
Van Tiel, B., \& Geurts, B. (2013). Truth and typicality in the interpretation of quantifiers. Proceedings of Sinn und Bedeutung 18, (pp. 451-468). Basque Country.
Zajenkowski, M., Szymanik, J., \& Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. Journal of psycholinguistic research, 43(6), 839-853.

# Right hemisphere lateralization and holistic processing do not always go together: An ERP investigation of a training study 

Ricky Van-yip Tso (rvytso@eduhk.hk)<br>1. Department of Psychology; and 2. Psychological Assessment and Clinical Research Unit, Centre for Psychological Health; The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong<br>Hangyu Chen (ariachan@connect.hku.hk), Yui Andrew Yeung (andrewy@connect.hku.hk)<br>Terry Kit-fong Au (terryau@hku.hk), \& Janet Hui-wen Hsiao (jhsiao@hku.hk)<br>Department of Psychology, The Jockey Club Tower, Centennial Campus, The University of Hong Kong, Pokfulam Road, Hong Kong


#### Abstract

Holistic processing (HP) and right-hemispheric lateralization both mark expertise in visual object recognition such as face and sub-ordinate object perception. However, counterexamples have been found recently: Experiences of selective attention to parts such as writing experiences in Chinese characters reduced HP while increased right hemisphere lateralization. We investigated the association between HP and brain activities measured by event-related potentials (ERP) in participants trained to recognize artificially-created scripts using either whole-word or grapheme-to-phoneme approaches. Stronger N170 activities were found in both hemispheres in both training approaches. Though the type of training approaches induced opposite directions in correlations between HP and the ERP signals in the right hemisphere: In the whole-word condition, the HP effect increased with stronger right-hemispheric N170 activities; while the direction of this correlation was reversed in the grapheme-to-phoneme condition. This demonstrates that HP and right hemispheric lateralization are separate processes that are associated with different perceptual mechanisms.


Keywords: holistic processing, hemisphere lateralization, ERP, EEG, perceptual expertise

## Introduction

## Holistic processing and right hemisphere lateralization

Holistic processing (HP) has consistently been reported to be a perceptual marker of visual expertise in face and subordinate-level visual object recognition (Bukach et al., 2006; c.f. Mckone, Kanwisher, \& Duchaine, 2007). For example, Gauthier, Williams, Tarr, and Tanaka (1998) trained participants to recognize "Greebles"-novel artificial objects-and found a positive relationship between HP and performance in within-category object recognition. Similarly, when participants were trained to individualize "Ziggerins" (an artificial object type), they showed an increase in HP (Wong, Palmeri, \& Gauthier, 2009).

HP in face perception can be demonstrated with the composite face illusion induced by the composite paradigm: Two identical top halves of two faces are more likely judged as different when the two bottom-half faces are from different faces, (see Rossion, 2013, for a review). The composite illusion suggests that all facial parts are obligatorily attended to, which results in the failure of selectively attention to parts (Richler, Wong, \& Gauthier,
2011). This paradigm demonstrates one of the three types of configural processing according to Maurer et al. (2002).

Hemispheric asymmetry may be another expertise marker for object recognition. Neuroimaging studies generally showed stronger activation in the right occipitotemporal area for face recognition (Rossion, Hanseeuw, \& Dricot, 2012). Complementing this finding, Gauthier and Tarr (2002) found that as participants were trained to recognize individual Greebles, increase in HP was correlated with activation changes in the right occipitotemporal regions. Because of the concurrence of robust HP in face and objection recognition with stronger right-hemisphere (RH) activations, HP is suggested to be a property of RH visual processing (Ramon \& Rossion, 2012). It is also consistent with the holistic-analytic dichotomy proposed in the hemispheric asymmetry literature (Cooper \& Wojan, 2000).

However, recent studies suggest that HP and RH lateralization do not necessarily go together. For example, in Chinese character perception, Hsiao and Cottrell (2009) found that while expert readers showed a reduced HP as compared with novice readers, the left-side bias effect, which is suggested to be an indication of RH lateralization, was shown only in experts ${ }^{1}$. Tso et al. (2014) reported an inverted U-shape development pattern in HP of Chinese characters: as compared with novices, Chinese readers with limited writing experiences showed increased HP, whereas Chinese readers skilled in writing Chinese characters showed reduced. This result suggests that HP is modulated by sensorimotor experiences while RH lateralization is not.

## Theories and model of hemispheric processing

The RH has long been suggested to preferentially execute whole-based/configural/coarse/global processing while partbased/analytic/fine/local processing is more involved in the left hemisphere (LH) (e.g., Sergent, 1982). Ivry and Robertson (1998) proposed the Double Filtering by Frequency (DFF) theory, which suggests that visual information is processed in the brain by frequency-based

[^529]representations at two stages: at the first stage, attention processes select a task-relevant frequency range; at the second stage, high spatial frequency (HSF) information is amplified in the LH while low spatial frequency (LSF) information is amplified in the RH. The DFF theory is able to account for hemispheric asymmetry in processing local (HSF) and global (LSF) information. For example, using Navon's hierarchical patterns (1977; Fig. 1), Sergent (1982) found a left-visual field (LVF)/RH advantage in judgements made based on global information and a right-visual field (RVF)/LH advantage in judgements made based on local information. Similarly, a LVF/RH advantage was found in identifying LSF gratings while a RVF/LH advantage for HSF gratings (e.g., Christman, Kitterie, \& Hellige, 1991). These results suggest that the LH is more tuned to processing local/HSF information while the RH more tuned to processing global/LSF information.


Fig. 1. Hierarchical letter patterns. The pattern on the left shows the global form ' $L$ ' consisting of local elements ' $H$ '. The one on the right shows the global form ' H ' and local elements ' $L$ '.

In visual word recognition, a stronger LH lateralization is typically observed for alphabetic than logographic scripts. Hsiao and Lam (2014) showed that this effect could be accounted for by a computational implementation of the DFF theory: the decomposition of words into graphemes for grapheme-phoneme mapping requires more HSF/LH processing than logographic reading. Hsiao and Cheung (2011) and Hsiao and Galmar (2016) examined the relationship between HP and RH lateralization in visual recognition using the same model (with triangular symbols consisting of 3 English letters and faces respectively). They found a positive correlation between HP and RH lateralization when the recognition task relied purely on the distanced between features (i.e., the second order relationship, a type of configural processing; Maurer et al., 2002), while this correlation became negative when the recognition task relied purely on the identity/features of local components. These results suggest that HP and RH lateralization are separate processes modulated by different recognition requirements. Since the recognition of words in alphabetic languages relies more on the identity of local components for grapheme-phoneme conversion than that in logographic languages, it is possible that alphabetic and logographic reading will result in different relationships between HP and RH lateralization.

## ERP component N170

In EEG studies, the ERP component N170, peaking between 150 and 200 ms after the onset of visual stimulus presentation, was found to be associated with perceptual
expertise effects (e.g. ,Maurer, Zevin, \& McClandiliss, 2008). Consistent with neuroimaging and behavioural research on hemispheric asymmetry in visual object recognition, EEG/ERP studies also showed reliable hemispheric asymmetries of visual expertise effects in N170, such as a larger N170 response in the RH for faces (e.g., Scott \& Nelson, 2007), and a larger N170 response in the LH for words (e.g. Maurer, Brandeis, \& McCandliss, 2005). Thus, the N170 responses towards visual stimuli, which are suggested to reflect occipito-temporal activities in visual object recognition, can be considered an electrophysiological indication of hemispheric asymmetry in visual object processing (e.g., Maurer et al., 2008).

## The present study

Here we aim to examine how different visual object recognition requirements modulate the relationship between HP and RH lateralization. We specifically contrast the difference between visual word recognition in alphabetic and logographic languages, the two major types of scripts currently in use. To do this, we trained participants to recognize artificially-created characters and examined the perceptual and electrophysiological changes. Participants learned to recognize the same set of characters under which the decoding method was manipulated to be using either whole-word (logographic) or grapheme-to-phoneme (alphabetic) approaches. Any difference in the perceptual or hemispheric lateralization changes occurring after the training should mainly come from the difference in the decoding methods (logographic vs. alphabetic). According to the previous studies (e.g., Hsiao \& Galmar, 2016), the requirement of grapheme-phoneme conversion in learning to read the characters alphabetically may induce a negative correlation between HP and RH lateralization, whereas a positive correlation may be observed when learning to read the same characters logographically. This is the first training study to investigate HP and its association with hemispheric lateralization of reading alphabetic and logographic scripts.

## Methods

## Participants

54 college students aged 18 to 26 with no prior knowledge to Korean Hanguls were recruited: 18 of which spoke English as a native language and 34 were Cantonese-English bilinguals who spoke Cantonese as a native language. 25 of them were females. They were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971) with normal or corrected to normal vision. Half of them were randomly assigned to the logographic condition while half of them were assigned to the alphabetic condition, with native language and gender matched between the conditions.

## Materials

A total of 30 artificial components were created to make 80 Artificial Korean-like Characters (AKC). The AKCs were of a top-bottom configuration with two top components and one bottom component in each character-this arrangement
simulated the top-heavy configuration of faces as well as a structure of Chinese characters. In the Alphabetic condition, each component in an AKC corresponded to a phoneme. Each AKC mapped onto a syllable with its combination of components following a consonant-vowel-consonant (CVC) phonological rule. In the Logographic condition, each AKC was randomly assigned a syllable pronunciation that appeared in the Alphabetic condition (Fig. 2).


Fig. 2. Examples of (a) AKC components and (b) an AKC

## Training Phrase

Each participant learned all 80 AKCs during 3 learning sessions in 3 consecutive days. Each learning session consisted of two blocks with 40 AKCs learned in each block. Two learning blocks in each learning session allowed participants to be exposed to all 80 AKCs per day. In the Logographic condition, each AKC was shown as a whole character for four times in each trial, accompanied by its pronunciation read by a female voice in each display. Each of the first three displays lasted for 500 ms , with the fourth display stayed on the screen for the participants to familiarize with for 5 seconds. In the Alphabetic condition, each AKC was also shown as a whole character for four times in each trial. A different component was highlighted in each of the first 3 displays, accompanied by the pronunciation of the component's phoneme read in a female voice in each display, for 500 ms . The last display of the AKC was accompanied by the pronunciation of the whole AKC and stayed on the screen for 5 seconds.

To monitor and encourage learning progress, after each learning session, participants completed a forced-choice quiz. In each trial, two AKCs were displayed on the screen accompanied by a syllable sound. Participants chose the AKC that matched the sound by pressing the corresponding buttons on a response box. There were a total number of 160 trials with each AKC-sound pair appearing twice. A feedback on the correctness with the accumulated percentage of correct responses was given immediately at the end of each trial. At the end of the last training session, participants in both the Alphabetic and Logographic condition developed over $80 \%$ accuracy in the quiz.

## Post-test and Pretest

Participants performed a complete composite task and a sequential matching task with EEG recording with AKC stimuli once before and once after the training
Complete Composite Task. We employed the complete composite paradigm to examine HP of AKCs, adopting the procedures from Hsiao and Cottrell (2009). Eighty pairs of AKCs taught in the training were selected. 20 pairs were presented in each of the four conditions: same-congruent, different-congruent, same-incongruent, and differentincongruent trials (Fig. 3a). In the congruent trials, the
attended halves and the irrelevant halves led to the same response (i.e. both the attended part and the irrelevant part were the same or different). In the incongruent trials, the attended halves and the irrelevant halves led to different responses: In same incongruent trials, the attended halves were the same while the irrelevant halves were different; whereas in different incongruent trials, the attended halves were different while the irrelevant halves were the same ${ }^{2}$.

Each trial started with a fixation cross for 1000 ms , followed by a cue indicating the part that participants should attend to (either top or bottom) for 1000 ms . A pair of AKCs-one above and the other one below the initial fixation respectively, about five degrees of visual angel away from each other-appeared for 500 ms , followed by a mask. Participants were instructed to judge whether the attended halves of the two AKCs were the same or not as quickly and as accurately as possible by pressing the corresponding buttons on the response box (Fig. 3). Accuracy of each trial was recorded.


Fig. 3. (a) Illustration of stimulus pairs in the complete composite paradigm; the attended components are circled in red. (b) Trial sequences, the red line shows the splitting point between top and bottom halves.

Each AKC was approximately $1.5 \mathrm{~cm} \times 1.5 \mathrm{~cm}$ in size on the screen, spanned about 1.6 degree of visual angle at a viewing distance of 55 cm . The participants' discrimination sensitivity A' was measured as:

$$
A^{\prime}=0.5+\left[\operatorname{sign}(H-F) \frac{(H-F)^{2}+|H-F|}{4 \max (H, F)-4 H F}\right]
$$

where H and F are the hit and false alarm rate respectively. We used $A^{\prime}$ to measure sensitivity due to its bias-free nonparametric property, as $d^{\prime}$ may be affected by response biases when normality and equal standard deviations are not assumed (Stanislaw \& Todorov, 1999). We measured Holistic $A^{\prime}$ as a normalized measure of HP that takes into consideration the individual baseline performance differences (Singer \& Sheinberg, 2006): the greater the magnitude, the stronger the degree of holistic processing.

$$
\text { Holistic } A^{\prime}=\frac{A^{\prime} \text { in congruent condition }-A^{\prime} \text { in incongruent condition }}{A^{\prime} \text { in congruent condition }+A^{\prime} \text { in incongruent condition }}
$$

[^530]EEG recording and analysis A sequential matching task was used to measure ERPs in response to the presentation of AKCs. The task consisted of 240 trials, separated into 6 blocks. Each trial started with a central fixation for 500 ms , followed by an AKC appearing at the screen center for 150 ms . The screen then turned blank for 1000 ms . A second character then appeared at the center and remained until participants made a response judging whether the two characters were the same or different. Each character subtended a visual angle of around 1.7 degree. Participants were instructed not to blink during a trial until they saw the letter ' B ' on the screen. These trials were conducted using E-prime v2.0 (Psychology Software Tools, Pittsburgh, PA).

EEGs were recorded using a 64-channel ANT system (Electro-cap International). EEG activities were sampled at 512 Hz . The data analysis was performed using EEGLAB (Delarme \& Makeig, 2004) and ERPLAB (Lopez-Calderon \& Luck, 2014). Only trials with correct responses were included in the EEG analysis. Bin-based epochs were extracted from -200 ms to 800 ms of the stimulus onset. The time window 140 to $200 \mathrm{~ms}(170 \pm 30 \mathrm{~ms})$ was chosen based on the grand average data of the participants in the Alphabetic and Logographic training conditions for identifying N170 peak amplitudes. PO7 electrode in the LH and its symmetrical electrode PO8 in the RH were selected for analysis as these electrodes were where the peak amplitude was found within the selected time window (see Hsiao et al., 2007 Yoncheva et al., 2010).

## Results <br> Holistic processing (HP)

Repeated-measures ANOVA was used to investigate HP effects (time: Pretest vs. Post-test x condition: alphabetic vs. logographic). For Holistic $A^{\prime}$, there was a marginal effect of time, $F(1,47)=2.868, p=.097, \eta_{p}{ }^{2}=.057$ : HP decreased as the result of training. There was no main effect of condition or an interaction between time and condition. ${ }^{3}$ (Fig. 4).


Fig. 4. Holistic A' in the pretest and post-test in the Alphabetic and Logographic conditions

## EEG neural correlates

Mixed ANOVA was used for analyzing N170 peak amplitude data (Time: Pretest vs. Post-test x Hemisphere:

[^531]Left vs Right x Condition: Alphabetic vs. Logographic). A significant main effect of time was observed, $F(1,53)=$ $7.457, \mathrm{p}=.009, \eta_{p}{ }^{2}=.123$, showing that N170 amplitude was increased after training. There was a marginal main effect of hemisphere, $F(1,53)=3.064, \mathrm{p}=.086, \eta_{p}{ }^{2}=.055$, and a marginal effect of condition, $F(1,53)=3.678, p=$ $.061, \eta_{p}{ }^{2}=.065$. No significant interaction effect was observed.


Fig. 5. N170 responses ( $\mu \mathrm{V}$ ) in (a) PO7 (left hemisphere) and (b) PO8 (right hemisphere) in the pretest and post-test, averaged across all participants.

## Pearson's correlation and moderation analysis

Correlation analyses between Holistic $A^{\prime}$ and N170 amplitude at PO7 (LH) and PO8 (RH) were performed separately for the Alphabetic and Logographic conditions to examine the relationship between HP and RH lateralization. In the Alphabetic condition, Holistic $A^{\prime}$ in the post-test was correlated positively with the N170 amplitude at PO8 (RH) in the post-test, $r^{2}=.435, p<0.05$, as well as the N170 amplitude change between the pretest and the post-test at PO8, $r^{2}=.506, p<0.05$. In contrast, these correlations were negative in the logographic condition, $r^{2}=-.483, p<0.05$, and, $r^{2}=-.409, p<0.05$, respectively. See Fig. 6


Fig. 6. The correlation between Holistic A' and PO8 N170 Amplitude ( $\mu \mathrm{V}$ ).

To further understand the differences in the direction of the correlations between Holistic A' and N170 amplitude at PO8 in the post-test, a moderation analysis was conducted. In the first step, training condition (Logographic vs. Alphabetic) and N170 amplitude at PO8 were entered in the regression analysis. In the second step, the interaction term between training condition (Logographic vs. Alphabetic) and N170 amplitude at PO8 was entered, and it explained a significant increase in variance in Holistic $A^{\prime}, \Delta R^{2}=0.203$, $\mathrm{F}(1,48)=11.593, p=.001$. Thus, training condition significantly moderated the correlations between Holistic A'
and N170 amplitude at PO8 in the post-test: While a more negative/larger N170 amplitude at PO8 correlated with a weaker HP effect in the alphabetic condition, a more negative/larger N170 amplitude at PO8 correlated with a stronger HP effect in the logographic condition (Fig. 5). This suggests a stronger N170 activity correlated with a weaker HP effect in the alphabetic condition, while it is vice versa in the logographic condition in the RH. However, we did not find significant correlations between HP and N170 at PO7 in either condition.

## Discussions

In the current study, we aimed to examine how learning to read words alphabetically or logographically modulates the relationship between holistic processing (HP) and right hemisphere ( RH ) lateralization in the perception of visual words. Previous computational modeling studies have suggested that in visual object recognition, when the recognition task relies purely on the distances among local components (second order relationships, a type of configural processing; Maurer et al., 2002), there was a positive correlation between HP and RH lateralization. In contrast, when the recognition task relies purely on the identity of local components, this correlation becomes negative (Hsiao \& Cheung, 2011; Hsiao \& Galmar, 2016). This result is consistent with the face recognition and perceptual expertise literature, which typically shows an increase in HP coincided with RH lateralization, especially when the task involved processing of configural information (e.g., Gauthier \& Tarr, 2002; Ramon \& Rossion, 2012). It is also consistent with the literature on expert Chinese character processing: decreased HP due to writing experience, which required selective attention to local components, was correlated with increased left side bias/RH processing (Tso et al., 2014). Here we tested this modeling prediction through a training study, in which we measured changes in HP and ERP N170 amplitude as the result of learning to read artificial Korean-like characters (AKCs) either alphabetically or logographically.

Our study revealed that training to read AKCs in either the Alphabetic or the Logographic conditions increased N170 amplitude in both hemispheres at electrodes PO8 and PO7. This result is consistent with the perceptual expertise literature, which typically showed an increased N170 amplitude as the result of the expertise (e.g. Maurer et al., 2008; Tanaka \& Curren, 2001).

More importantly, in the post-test, we found that the HP effect of AKCs correlated with N170 amplitude in the RH differently between the 2 conditions: while the correlation analysis showed the stronger the HP effect, the more negative the N170 amplitude at PO8 in the Logographic condition, the direction of this correlation was reversed in the Alphabetic condition. It seems that different learning approaches to recognizing a written script moderates the direction of the correlation between HP and neural activities in the right occipital temporal regions. This effect is consistent with the modeling data based on the DFF theory
(Hsiao \& Galmar, 2016; Hsiao \& Cheung, 2011). In the Logographic condition, participants may have used a wholeword recognition approach, which led to increased HP, as well as a higher sensitivity to the distances the components, a type of configural processing (Maurer et al., 2002). This type of configural processing has been shown to involve RH lateralization (Scott \& Nelson, 2006). Thus, in this condition, RH lateralization and holistic processing are positively correlated with each other. In contrast, in the Alphabetic condition, the requirement of graphemephoneme conversion during learning may have encouraged local featural/high spatial frequency processing for identifying local component, which is typically leftlateralized (Ivry \& Robertson, 1998). In addition to identifying local components, word recognition in the Alphabetic condition also required recognizing components in a particular sequence/configuration, or more specifically, the first order relationship among features (Maurer et al., 2002). This processing may require integration of information among components, leading to increase in HP. Thus, in the Alphabetic condition, HP was negatively correlated with RH lateralization, since the increase in HP due to the use of configural information for relative positions of components (i.e., the first order relationship among features) may coincide with decreased reliance on RH global processing. Future work will examine these possibilities.

Consistent with the modeling data, the current results suggest that HP (as measured in the composite paradigm) and RH lateralization do not always go together in visual object recognition. It depends on the requirements of the recognition task. Consistent with this finding, in an fMRI study, Harris and Aquirre (2010) showed that neurons in the right occipito-temporal region (fusiform face area, FFA) could flexibly represented two facial features either conjointly (suggesting HP) or separately, depending on the recognition task requirements. Note however that our current results regarding the relationship between HP and RH lateralization is limited to the HP as measured in the composite paradigm. In the literature, HP effects have been demonstrated using different paradigms, such as the partwhole task (Farah, Wilson, Drain, \& Tanaka, 1998) in addition to the composite paradigm. HP effects demonstrated using different paradigms likely involve different underlying mechanisms (Richler et al., 2012). Future work will examine whether similar relationships between HP and RH lateralization can also be observed using other HP paradigms.

In conclusion, this is the first training study to report on the changes in both HP and hemispheric lateralization in learning to read an artificial script under different decoding methods (i.e., logographic vs. alphabetic). Different learning approaches induced opposite directions of correlations between HP and RH activities: Learning a script alphabetically induced a negative correlation between HP and RH lateralization, while that induced by learning a logographic script was positive. It seems that HP and RH
lateralization do not always go together, depending on the decoding strategy in visual object recognition, or more specifically, the type of configural information used in the recognition processes.

## Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (HKU 745210H \& HKU 758412H to J. Hsiao). We are also grateful to Remy Schmitz, Joey Ma, Micaela Tang, Sam Choi, Connie Chu, Jocelyn Kwok and Wendy Zhang for their input in the experiment preparation. This study has also appeared in the PhD thesis of the first author.

## References

Bukach, C. M., Gauthier, I., \& Tarr, J. M. (2006). Beyond faces and modularity: The power of an expertise framework. TICS, 10, 156-166.
Brady, N., Campbell, M., \& Flaherty, M. (2005). Perceptual asymmetries are preserved in memory for highly familiar faces of self and friend. Brain Cog, 58, 334-342.
Christman, S., Kitterle, F. L., Hellige, J. (1991). Hemispheric asymmetry in the processing of absolute versus relative spatial frequency. Brain Cog, 16, 62-73.
Cooper E. E., Wojan T. J. (2000). Differences in the coding of spatial relations in face identification and basic-level object recognition. J. Exp. Psychol. Learn. Mem. Cogn. 26, 470-488.
Gauthier, I., \& Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: bridging brain activity and behavior. Journal of Experimental Psychology: Human Perception and Performance, 28, 431.
Gauthier, I., Williams, P., Tarr, M. J., \& Tanaka, J. (1998). Training "Greeble" experts: Aframework for studying expert object recognition processes. Vision Research, 38,
Harris, A., \& Aguirre, G. K. (2010). Neural tuning for face wholes and parts in human fusiform gyrus revealed by FMRI adaptation. J. Neurophysiol., 104(1), 336.
Hsiao, J. H., \& Cheung, K. C. F. (2011). Computational exploration of the relationship between holistic processing and right hemisphere lateralization in featural and configural recognition tasks. Paper presented at CogSci2011.
Hsiao, J. H., \& Cottrell, G. (2009). Not all visual expertise in holistic, but it may be leftist: The case of Chinese character recognition. Psychol Sci, 20, 455-463.
Hsiao, J. H., \& Galmar, B. (2016). Holistic processing as measured in the composite task does not always go with right hemisphere processing in face perception. Neurocomputing, 182, 165-177.
Hsiao, J. H., \& Lam, S. M. (2013). The modulation of visual and task characteristics of a writing system on hemispheric lateralization in visual word recognition - A computational exploration. Cognitive Sci, 37, 861-890.
Ivry, R. B., \& Robertson, L. C. (1998). The Two Sides of Perception. Cambridge: MIT Press.
Maurer, D., Le Grand, R., \& Mondloch, C. (2002). The many faces of configural processing. TICS, 6, 255-260.

Maurer, U., Brandeis, D., \& McCandliss, B. (2005). Fast, visual specialization for reading in English revealed by the topography of the N170 ERP response. Behavioral \& Brain Functions, 1, 13.
Maurer, U., Zevin, J.D., McCandliss, B.D. (2008) Leftlateralized N170 effects of visual expertise in reading: evidence from Japanese syllabic and logographic scripts. Journal of Cognitive Neuroscience, 20, 1878-1891.
McKone, E., Kanwisher, N., \& Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? Trends in Cognitive Sciences, 11, 8-15.
Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia, 9, 97-113.
Ramon, M. \& Rossion, B. (2012). Hemisphere-dependent holistic processing of familiar faces. Brain $\operatorname{Cog}, 78,7-13$.
Richler, J. J., Cheung, O. S., \& Gauthier, I. (2011). Beliefs alter holistic face processing ... if response bias is not taken into account. Journal of Vision, 11, 1-13.
Richler, J. J., Palmeri, T. J., \& Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. Frontiers in psychology, 3:553.
Richler, J. J., Wong, Y. K., \& Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. Curr Dir Psychol Sci, 20(2), 129-134.
Rossion, B., Hanseeuw, B., \& Dricot, L. (2012). Defining face perception areas in the human brain: A large scale factorial fMRI face localizer analysis. Brain and Cognition 79(2), 138-157.
Scott, S. L., \& Nelson, C. A. (2006). Featural and configural face processing in adults and infants: A behavioral and electrophysiological investigation. Perception, 35, 11071128.

Sergent, J. (1982). The Cerebral Balance of Power: Confrontation or Cooperation? J Exp Psychol Human, 8, 253-272.
Singer, J.M., \& Sheinberg, D.L. (2006). Holistic processing unites face parts across time. Vision Res, 46, 1838-1847.
Stanislaw, H., \& Todorov, N. (1999). Calculation of signal detection theory measures. Behav Res Meth Ins C, 31, 137-149.
Tanaka, J. W., \& Curran, T. (2001). A neural basis for expert object recognition. Psychol Sci, 12, 43-47.
Tanaka, J. W., \& Farah, M. J. (1993). Parts and wholes in face recognition. Q J Exp Psychol, 46A, 225-245.
Tso, R. V. Y., Au, T. K., \& Hsiao, J. H. (2014). Perceptual expertise: Can sensorimotor experience change holistic processing and left side bias? Psychol Sci, 25, 1757-1767.
Wong, A. C., Palmeri, T. J., \& Gauthier, I. (2009). Conditions for face-like expertise with objects: Becoming a ziggerin expert-but which type? Psychol Sci, 20, 11081117.

Yoncheva, Y.N. , Blau, V.C, Maurer, U. \& McCandliss, B.D. (2010), Attentional Focus During Learning Impacts N170 ERP Responses to an Artificial Script. Developmental Neuropsychology, 35, 423-445.

# 'He's pregnant": simulating the confusing case of gender pronoun errors in L2 English 

Chara Tsoukala (c.tsoukala@let.ru.nl)<br>Centre for Language Studies<br>Radboud University<br>Stefan L. Frank (s.frank@let.ru.nl)<br>Centre for Language Studies<br>Radboud University

Mirjam Broersma (m.broersma@let.ru.nl)
Centre for Language Studies
Radboud University


#### Abstract

Even advanced Spanish speakers of second language English tend to confuse the pronouns 'he' and 'she', often without even noticing their mistake (Lahoz, 1991). A study by AntónMéndez (2010) has indicated that a possible reason for this error is the fact that Spanish is a pro-drop language. In order to test this hypothesis, we used an extension of Dual-path (Chang, 2002), a computational cognitive model of sentence production, to simulate two models of bilingual speech production of second language English. One model had Spanish (ES) as a native language, whereas the other learned a Spanish-like language that used the pronoun at all times (non-pro-drop Spanish, NPD_ES). When tested on L2 English sentences, the bilingual pro-drop Spanish model produced significantly more gender pronoun errors, confirming that pronoun dropping could indeed be responsible for the gender confusion in natural language use as well.


Keywords: L2 pronoun errors, language transfer, Dual-path model, bilingual sentence production

## Introduction

Second language (L2) speech errors have been employed in the past as a means to understand bilingual speech production as well as the acquisition process of a foreign language (Antón-Méndez, 2010; Poulisse, 1999). Certain L2 errors are observed more often due to discrepancies between the first language (L1) and the L2. For example, if the expression of a message in the L 2 requires the inclusion of a specific feature that would not be necessary in the L1, then speakers of these two languages may produce a speech error in their L2 due to L1 transfer (Odlin, 1989). In this study, we focus on a gender-related L2 pronoun error that has been observed among native speakers of Spanish and Italian; namely, errors involving the third person singular nominative pronouns 'he' and 'she'. Even advanced Spanish speakers of L2 English occasionally confuse the two pronouns, referring to an actress as 'he' or a father as 'she', often without even noticing their mistake (Lahoz, 1991). At first, this phenomenon seems surprising because the Spanish language does have two equivalent pronouns ('él' for 'he' and 'ella' for 'she'), and also a very strong separation between the two genders, even more so than in English. For instance, depending on the suffix a word can be feminine or masculine (e.g., maestro - teacher
[masculine], maestra - teacher [feminine]; niño - child [masculine, a.k.a. boy], niña - child [feminine, a.k.a. girl]). This means that the gender mistakes that Spanish speakers make in English cannot be attributed to the lack of familiarity with the distinction. Furthermore, the challenge the English pronoun system poses for native speakers of Spanish could not be due to its inherent difficulty, as this would mean that most non-native speakers of English, regardless of their L1, would produce the same mistake. As Lahoz (1991) noted, a low proficiency level of the native Spanish speakers is not a reason either. This was also demonstrated in the experiments of Antón-Méndez (2010), where the participants showed an intermediate to upper intermediate knowledge of English. Finally, note that the gender mismatch error cannot be classified as a syntactic error; the produced sentence is grammatically correct, but it conveys the wrong meaning.

A hypothesis which has been put forward (Lahoz, 1991; Antón-Méndez, 2010) regarding the cause of errors in the use of English pronouns is the pro-drop status of the Spanish but not the English language. In pro-drop languages, nominative personal pronouns are often omitted (1b) because the number and person information is conveyed in the conjugated verb (Davidson, 1996), whereas in English the omission of the pronoun would result in an ungrammatical sentence (2b).
1.(a) Él/ella tiene un perro (Spanish)
(b) tiene un perro
2.(a) $\mathrm{He} /$ she has a dog (English)
(b) * has a dog

It is hard to imagine, however, how the pro-drop feature of the L1 might result in a gender pronoun error ("He's walking", when referring to a woman) instead of an omission ("Is walking"), which would be the case in a direct language transfer.

As a matter of fact, native speakers of Spanish have been noted to produce another gender-related pronoun error in English, this time regarding possessive pronouns ('his', 'her'). Due to the high frequency of this type of errors, a lot more
emphasis has been given to the misuse of these pronouns than the subject pronouns (White, Muñoz, \& Collins, 2007; Anton-Mendez, 2011). The reason that English possessive pronouns pose a challenge for native speakers of Spanish is most likely that in Romance languages the possessive pronoun agrees in gender and number with the possessum, namely the noun that follows, whereas in Germanic languages such as English the possessive pronoun refers to the antecedent. For example:
i. His daughters are on vacation.
[his: 3rd person masculine singular]
ii. Sus hijas están de vacaciones.
[sus: 3rd person feminine plural]
Due to the different information encoding Spanish speakers of English may occasionally make gender mistakes such as "He called her mother", where 'her' refers to the antecedent ('he') and not a different female person. This is because 'mother' is female, and a Spanish speaker would use that gender information to construct the possessive pronoun in Spanish. The resulting error in English is, of course, confusing, as a speaker of English would not guess that 'her' in this case refers to the same subject ('he'). The gender error in the case of L2 English possessive pronouns seems clearly due to L1 transfer, because the properties of Spanish are directly applied to English. In the case of the subject pronoun gender errors, on the other hand, it is not evident that the pro-drop feature of one language would lead to a gender error in L2. The present study addresses only the latter type of errors.

Antón-Méndez (2010) has investigated the hypothesis that the pro-drop feature of Spanish is responsible for the gender pronoun errors in L2 English ("pro-drop hypothesis"). She conducted an experiment eliciting semi-spontaneous speech in English, where she compared native Spanish and native French speakers of L2 English with respect to the pronoun errors they produced. French was chosen as it is a Romance language that is similar to Spanish in several aspects, but which, in contrast to the Spanish language, is not a pro-drop language. Each test group consisted of 20 participants who were comparable in terms of education, age of English acquisition, frequency of use and proficiency. The participants were shown 43 illustrations and were asked questions designed to elicit pronoun production. The subjects were instructed to respond freely, and the pronoun errors they produced were recorded. The types of reported errors fall in the following categories: person errors (e.g., 'I' instead of 'you'), number errors (e.g., 'I' instead of 'we'), gender errors ('he' instead of 'she' and vice-versa), animacy errors (e.g., 'he' instead of 'it'), omission errors (e.g., 'is swimming'), insertion errors (e.g., 'the boy he played' instead of 'the boy played') and other errors (e.g., 'it' instead of 'there' in 'there is').

Spanish speakers of L2 English indeed made significantly more gender errors $(4.30 \%)$ compared to other types of pro-
noun errors and to the French group $(0.68 \%)^{1}$. The pronoun errors recorded were not due to erroneous transfer of the Spanish L1 grammar, as the Spanish speakers made no omission errors ('is swimming'); thus, in none of the items of Antón-Méndez's experiment did the subjects omit a pronoun, which would have been the case in a grammatical transfer. Importantly, even though there were slightly more 'he' than 'she' errors (he: $5.68 \%$, she: $2.98 \%$ ), the difference is not statistically significant. Therefore, the Spanish speakers were not using a default pronoun (e.g., always 'he' instead of 'she'). The use of 'he' as the default pronoun would have suggested that another factor might underlie the error, for instance, the difficulty that the English phonology poses for speakers of Spanish. The Spanish phonology does not contain the phonemes $/ \mathrm{J} /$ in 'she' and $/ \mathrm{h} /$ in 'he', therefore one explanation for the gender pronoun issue could be at the phonological level. In the present study we focused only on the pro-drop feature, not because we disregard the potential role of the phonology, but because we wanted to investigate whether the pro-drop feature has the capacity of causing this type of gender errors in L2.

In order to focus on the pro-drop feature, we simulated bilingual sentence production using computational cognitive modeling. The pro-drop feature is not the sole difference between the French and Spanish languages, and one could argue that the differences in the error patterns between the two groups could have been partially attributed to confounding factors, for instance, to a different L2 English teaching system in Spain and France.

Using computational modeling we can remove all possible confounds and therefore minimize the variance by focusing only on the phenomenon of interest, which in this case is the pro-drop feature and its possible effect on L2 English pronouns. For this reason, we modified Dual-path (Chang, 2002), a computational cognitive model of sentence production, to account for bilingualism. We then compared L2 English speech production of simulated native speakers of Spanish (ES) on the one hand, to L2 production of simulated native speakers of a Spanish-like language ('non-pro-drop Spanish', NPD_ES) on the other hand. The latter contained all the features of the Spanish language (lexicon, allowed structures) except the pro-drop feature; therefore, pronouns needed to be used at all times. All input languages (ES, NPD_ES and EN) were artificially generated and based on the Spanish and English language, using a subset of their lexica and syntactic structures. If the bilingual Spanish-English (ES-EN) Dualpath model produces significantly more subject pronoun errors in English than its Spanish-like non-pro-drop equivalent (NPD_ES-EN), it will be clear that the pro-drop feature of the Spanish language is the reason for this particular L2 error in the simulation, as the two simulated languages differ only in their pro-dropness. If this is the case, we will have confirmed

[^532]that the pro-drop feature has the capacity to lead to gender pronoun errors in L2 English.

## Method

In order to simulate Spanish speakers of L2 English, we developed two bilingual models using a modified version of Dual-path which is a connectionist model based on the Simple Recurrent Network (SRN; Elman, 1990) achitecture (Chang, 2002).

## Bilingual Dual-path model

Dual-path (Figure 1) learns to convert a message into a sentence by predicting the sentence word by word ("next word prediction"). It has two pathways (hence the name) that influence the production of each word; the meaning system which learns concepts, roles and event semantics, and the sequencing system which is an SRN that learns to abstract syntactic patterns. Both paths influence the word output layer. The sequencing system consists of one recurrent hidden layer (of 30 units in our simulations) and two "compress" layers (of 12 units each) that are placed between the input word, the hidden layer and the output word.

The meaning system learns to map the input word onto a concept, which is linked to a specific thematic role (that is given for each sentence through fixed connections). The fixed connections allow the separation between concepts and roles, which, in turn, enables the model to generalize and to produce words in novel places. The thematic role is connected to the hidden layer, and so is the "event-semantics" layer. The hidden layer spreads the activation to the next thematic role (in the meaning path, and the "compress" unit in the syntactic path), which is in turn linked to a specific predicted concept that is used as input to the output word layer, along with the "compress" unit.

In the original model, all layers use the tanh activation function, except the output layer that uses softmax. In the modified version of the model, we also employed softmax for the predicted role layer. This led to a stricter selection of the upcoming thematic role which helped overcome a difficulty that the model had with learning the correct articles regarding gender and definiteness (e.g., 'a' vs 'the'). Furthermore, our version has a "target language" layer in the meaning path that is used as an additional input to the hidden layer, along with the "event-semantics" layer. The "target language" denotes the intended spoken language and helps the model handle more than one language. The modified model can be found at https://github.com/xtsoukala/dual_path.

## Input languages

Message Dual-path is trained using randomly generated sentences paired with their meaning (Chang, Dell, \& Bock, 2006). The meaning (message) contained information regarding four thematic roles (AGENT, PATIENT, ACTION, RECIPIENT). A concept (e.g., 'WOMAN' for the English word 'woman' or Spanish word 'mujer') was assigned to each thematic role depending on the meaning that needed
to be expressed (e.g., in the sentence "the woman run -s" the message would include AGENT=WOMAN, DEF). Furthermore, the message contained event-semantic information (denoted as ' $E$ '), which gave information regarding the tense (PRESENT or PAST) and aspect (SIMPLE or PROGRESSIVE). The message contained information about the target language (ES or EN) as well. This information was given at the beginning of the sentence along with the roles and the event-semantics, so that the model knew whether it was supposed to produce an English or Spanish sentence.

Structures The allowed structures for all languages were the following (where ' $S$ ', the subject, is omitted in the prodrop case):

1. (S)V: (Subject) - Verb, e.g., "He runs"
2. (S)VO: (Subject) - Verb - Object, e.g., "She kicked the ball"
3. (S)VIODO: (Subject) - Verb - Indirect Object - Direct Object, e.g., "He gave the girl a book"
4. (S)VDOIO: (Subject) - Verb - Indirect Object - Direct Object, e.g., "He gave a book to the girl"

The sentences in English and in non-pro-drop Spanish always started with a pronoun, and the sentences in pro-drop Spanish never started with a pronoun but always with a verb.

Lexicon The total lexicon consisted of 34 nouns: 11 male ('man', 'boy', 'father', 'brother', 'dog', 'hombre', 'niño', 'padre', 'hermano', 'perro', 'maestro'), 11 female ('woman', 'girl', 'mother', 'sister', 'cat', 'mujer', 'niña', 'madre', 'hermana', 'gata', 'enfermera') and 12 inanimate ('ball', 'stick', 'toy', 'kite', 'key', 'bag', 'pelota', 'palo', 'juguete', 'cometa', 'llave', 'bolso'), 24 verbs (e.g., 'give', 'show', 'walk', 'throw', 'present', 'dar', 'lanzar', 'presentar', 'nadar', 'caminar') and 26 function words (e.g., articles ('a', 'the', 'un', 'una', 'el', 'la'), pronouns ('he', 'she', 'él', 'ella') and auxiliary verbs ('is', 'was', 'está', 'estaba')).

The model treats the verb lemma ('give') and the suffix ('-s') as two different units. Note that syntactic information (such as 'verb', 'noun') is not given explicitly, but is learned by the model during training through the syntactic path. The syntactic gender was also learned implicitly during training through the article of Noun Phrases (NP) and pronouns. Semantic gender (e.g., 'ACTRESS, F', 'ACTOR, M') was not included in the model.

Thematic roles could be expressed using either an NP with definite (DEF) or indefinite (INDEF) articles (e.g., 'the woman', 'a woman') or the pronoun (PRON) equivalent ('she').


Figure 1: Bilingual Dual-path model

Example The following message would be the same across languages:

AGENT=WOMAN, PRON;
ACTION=GIVE;
PATIENT=INDEF, KEY;

## RECIPIENT=DEF, GIRL;

E=SIMPLE, PRESENT, AGENT, PATIENT, RECIPIENT
and it would be expressed linguistically in the following manner for the three languages:

1. she give -s the girl a key. [EN]
2. d -a a la niña una llave . [ES]
3. ella d -a a la niña una llave . [NPD_ES]

## Training

The two models were trained on 2000 randomly generated sentences (training set) and tested on 500 unseen sentences (test set). The models contained almost identical sets, with the only difference that the NPD_ES model expressed the subject pronoun at all times, whereas the ES model never did and always started with a verb. For each model we ran 100 simulations using the same input, but different random initial weights per simulation, as the input and the weights are the only non-deterministic parts of the model. The models were trained for 20 epochs, where 1 epoch corresponds to a full iteration of the training set ( 2000 sentences). At the beginning of each epoch, the training set was shuffled. In order to simulate late L2 acquisition, we first trained the models for 20 epochs using Spanish input only, and then used the fully trained weights as initial weights for the bilingual models. The bilingual input consisted of newly generated (2000 training and 500 test) sentences, this time using $50 \%$ (prodrop or non-pro-drop according to the model) Spanish and $50 \%$ English. We excluded from the analysis 7 simulations that did not manage to learn at least $75 \%$ of the test set by


Figure 2: Performance on the training and test sets over the training period ( 20 epochs) averaged over 93 simulations for the two bilingual models. Performance is measured in percentage of correctly produced Spanish and English sentences.
the end of the training in one of the two models, leading to a total of 93 simulations. Both bilingual models were able to perform equally well by the end of the training, reaching $99.69 \%$ correct for ES-EN and $99.70 \%$ correct for NPD_ESEN (Figure 2) on the test set that contained English and Spanish sentences.

## Results

In order to assess the performance of the two bilingual models on L2 pronouns, we focused only on the English sentences ( $50 \%$ of the test set). If a pronoun error was detected and the sentence was grammatical, it was classified as a gender pronoun error. We compared the performance of the two bilingual models with regard to the gender pronoun error production. If the models had a comparable performance we would not be able to confirm that the pro-drop feature has the capacity to lead to gender pronoun errors in L2 English. If, on the other hand, the NPD_ES model made fewer gender pronoun


Figure 3: Production of English gender errors (in log scale) averaged over 93 simulations for the two bilingual models. Note that the gaps in the non-pro-drop model are because $\log (0)$, for $0 \%$ error rate, is not defined, and therefore not plotted.
errors than the ES model it would indicate that the pro-drop feature is a possible explanation.

The non-pro-drop Spanish-English (NPD_ES-EN) bilingual model (Figure 3) produced almost no gender pronoun errors (maximum percentage: $0.11 \%$ ) whereas the bilingual model based on pro-drop Spanish (ES-EN) initially produced $9.75 \%$ pronoun errors, gradually dropping to $0.05 \%$.

Crucially, the ES-EN model never reached $0 \%$ (minimum error rate: $0.02 \%$ ) whereas the NPD_ES-EN model did. Following visual inspection, we ran a $z$-test for proportions from epoch 5 onwards to test for a difference in error rate between the models. The difference is significant $(z=7 ; \mathrm{p}<.001)$.

## Discussion

Our simulations showed that a bilingual model with L1 prodrop Spanish and L2 English produced significantly more gender pronoun errors than a similar model with L1 non-prodrop Spanish. These sentences were grammatically correct: the only error they contained was a pronoun with incorrect gender. Given that the only difference between the two L1s was the pro-drop feature, we have demonstrated that the prodrop nature of Spanish can indeed cause the gender pronoun error as observed in L1 Spanish speakers of L2 English.

Why the pro-drop feature does not lead to a direct language transfer ("is walking") in either the model or humans remains to be investigated, as the current simulations and results do not explain how pro-dropness in L1 could lead to gender errors in L2. Nevertheless, having a computational model that simulates the gender pronoun errors in L2 English can point us in the right direction. Our hypothesis for the occurrence of the gender error is that the gender information is not as crucial for the message planning, at least in the subject position, of a pro-drop language, and is therefore weaker or omitted,
even when producing sentences in a non-pro-drop L2.
It is important to point out that the Dual-path model does not contain a phonological level (Garrett, 1988). One might have thought that the reason Spanish speakers confuse the words 'he' and 'she' is because of the difficulty the English phonology poses for native speakers of Spanish. However, our simulations have produced gender errors without having any phonological representations. This does not mean that phonology could not play a role, but rather that it is not the only possible explanation.

It is also crucial to note two simplifying assumptions in these simulations. First, as mentioned in the Method section, the input for all three languages (EN, ES, NPD_ES) was artificially generated and it only represented a subset of the actual languages. In general, using natural input would be preferable as it would increase the validity and naturalness of the results. However, the benefit of miniature languages that are typically used in cognitive modeling is that they can be easily manipulated. For instance, in the simulations described here we were able to add and remove the pro-drop feature at will, leaving everything else the same, and thus to isolate this important feature from confounding factors.

Second, a crucial simplifying assumption in the miniature language is the absence of full NP subjects. We therefore repeated the simulations using new input for all languages, this time including $50 \%$ pronouns at the subject position and $50 \%$ noun phrases. Preliminary simulations show no gender errors in either model, which means that further research is needed using more natural language input, starting with a more naturalistic proportion of pronouns and NPs in the subject position based on English and Spanish corpora.

## Conclusion

Computational modeling can be used to validate or generate linguistic hypotheses while focusing on specific factors of interest and minimizing the variance. In this study, we have addressed the question as to whether the pro-drop feature of the Spanish language has the capacity to cause the gender pronoun errors that Spanish speakers of L2 English have been shown to produce (Lahoz, 1991; Antón-Méndez, 2010). The reported simulations showed that the model with L1 pro-drop Spanish produced more gender pronoun errors in L2 English than the model with L1 non-pro-drop Spanish, which is a necessary but not sufficient condition for the pro-drop hypothesis.

## Acknowledgements

The work presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

## References

Antón-Méndez, I. (2010). Gender bender: Gender errors in L2 pronoun production. Journal of Psycholinguistic Research, 39, 119-139.

Anton-Mendez, I. (2011). Whose? L2-English speakers' possessive pronoun gender errors. Bilingualism: Language and Cognition, 14, 318-331.
Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. Cognitive Science, 26, 609651.

Chang, F., Dell, G. S., \& Bock, K. (2006). Becoming syntactic. Psychological Review, 113, 234.
Davidson, B. (1996). 'Pragmatic weight' and Spanish subject pronouns: The pragmatic and discourse uses of 'tú' and 'yo' in spoken Madrid Spanish. Journal of Pragmatics, 26, 543-565.
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.
Garrett, M. F. (1988). Processes in language production. Linguistics: The Cambridge survey, 3, 69-96.
Lahoz, C. M. (1991). Why are he and she a problem for Spanish learners of English? Revista Española de Lingüústica Aplicada, 129-136.
Odlin, T. (1989). Language Transfer: Cross-Linguistic Influence in Language Learning. Cambridge: CUP.
Poulisse, N. (1999). Slips of the tongue: Speech errors in first and second language production (Vol. 20). Amsterdam: John Benjamins.
White, J., Muñoz, C., \& Collins, L. (2007). The his/her challenge: Making progress in a 'regular' L2 programme. Language Awareness, 16, 278-299.

# Deconstructing Transitional Probabilities: Bigram Frequency and Diversity in Lexical Decision 

Russell Turk (russell.turk2012@my.ntu.ac.uk) ${ }^{1}$<br>Gary Jones (gary.jones@ntu.ac.uk) ${ }^{1}$<br>Duncan Guest (Duncan.guest@ntu.ac.uk) ${ }^{1}$<br>Angela Young (angela.young@ntu.ac.uk) ${ }^{1}$<br>Mark Andrews (mark.andrews@ntu.ac.uk) ${ }^{1}$<br>${ }^{1}$ Department of Psychology, Nottingham Trent University, 50 Shakespeare Street<br>Nottingham, NG1 4FQ


#### Abstract

Statistical learning paradigms traditionally use transitional probabilities as a measure of statistical distribution within a language. The current study suggests that alternative metrics may exist that can account for differences in language processing ability. Two primed lexical decision tasks are used to examine the effects of bigram frequency and diversity on speed and accuracy of word recognition. It is demonstrated that both frequency and diversity contribute to word recognition performance; findings and theoretical implications are discussed.


Keywords: Statistical learning; lexical decision; language

## Introduction

Humans are superlative learners capable of identifying and tracking patterns in their environment, both implicitly and explicitly. This ability has been investigated using both implicit and, more recently, statistical learning paradigms (Perruchet \& Pacton, 2006) across a number of different domains including shapes (Kirkham, Slemmer, \& Johnson, 2002), music (Daikoku, Yatomi, \& Yumoto, 2014; Koelsh et al., 2016; Saffran et al., 1999), tactile stimuli (Conway \& Christianson, 2005) and, most prominently, language acquisition (Newport \& Aslin, 2004; Saffran, Aslin, \& Newport, 1996; Thiessen, \& Erickson, 2013; Vouloumanos, 2008) highlighting the ability of learners, ranging from infant (Saffran et al., 1996) to adult (Koelsh et al., 2016), to track the transitional probabilities (TPs) within a given set of stimuli.

Over the past two decades a plethora of researchers have investigated this phenomenon and have found transitional probabilities to be a robust indicator of performance across a number of different tasks and languages (e.g. Liu \& Kager, 2011; Toro, Sinnett, Soto-Faraco, 2005). This has led to the acceptance of TPs as the standard metric of co-occurrence within natural (and artificial) languages. However, if we consider that the TP of any given stimulus stems from an
interaction between the frequency of sequence XY and the number of potential candidates for Y then we are presented with two alternative metrics of statistical distribution. These, in turn, can be used to investigate the types of statistics which learners can attend to.

When applied to words in natural language these metrics can be termed Bigram Frequency, which is equal to the total number occurrences for a given sequence of two words within a language or representative selection thereof; and Bigram Diversity which can be defined as the number of items that potentially follow word X in the sequence XY .

It is logical to presume that both bigram frequency and diversity would be predictive of performance in languagerelated tasks. Evidence from Freudenthal et al. (2015) demonstrates that a frequency-based chunking mechanism can successfully reduce output errors in children's speech. This suggests that learners can track not only the TPs of the bigrams but also the frequency with which they occur. No evidence yet exists for a diversity-driven account of language proficiency. Nonetheless, it is recognised that predictability is an important facet of language processing (Bates \& MacWhinney, 1987; Glenberg \& Gallese, 2012; Goldberg, Casenhiser, \& Sethuraman, 2005; Pickering \& Garrod, 2004, 2007; Van Berkum et al., 2005); it follows therefore that a larger number of potential competitors for stimulus Y would serve to reduce predictability and thereby prove detrimental to response fluency.

Historically statistical learning paradigms such as those developed by Jusczyk and Aslin (1995, also Saffran, Aslin \& Newport, 1996) have exposed learners to artificial languages with carefully built-in TPs. This allows for admirable control of the input at the expense of both diversity and complexity. It has been argued that these languages are too simplistic to assess the extent to which learners are able to process distributional statistics within natural language (Frank et al, 2010; Johnson \& Tyler, 2010). To highlight this point, Saffran et al. (1996) reported
inter-syllable TPs of 1.0 in their seminal study whereas naturally occurring TPs are often considerably lower (the bigram little baby has a TP of less than 0.002).

Thus, the true test of statistical learning theories is their application to a more naturalistic dataset, one which retains the complexity and diversity of natural language whilst allowing for the accurate tracking of distributional cues; natural language corpora represent such datasets. The British National Corpus (BNC) is a collection of contemporary natural language which comprises approximately 100 -million words of written and spoken British English drawn from a variety of sources ranging from telephone calls to academic journals. By analysing the distributional statistics within the BNC it is possible to present learners with verisimilar but also quantifiable samples of natural language.

This raises another issue however, in that learners already have a great deal of experience interacting with natural languages. This makes traditional methods of testing such as those used by Saffran et al. (1996, also Frank et al., 2010; Jusczyk \& Aslin, 1995) unsuitable for natural language stimuli. Thus, two solutions are immediately apparent; the use of unfamiliar or non-native languages or an alternate method of assessment. Non-native languages would seem to be the ideal solution except that the complexity of these languages means that learners require either long periods of familiarisation or simplified samples in order to obtain actionable data. It is therefore favourable to introduce an alternate measure of language proficiency whilst retaining the complexity of the language and avoiding a lengthy familiarisation process.

The current study seeks to address this issue by assessing language proficiency using a primed lexical decision task (LDT) where the first word of a bigram acts as the prime and the second word the target. It is predicted that, using bigram frequency and diversity as statistical primes, response time for stimuli Y will be predicted by the strength of its association with prime X. Based on this prediction two hypotheses are proposed:

H1: Response times on a LDT will be quicker when primed with high frequency bigrams compared to low frequency or non-bigrams, and

H2: Response times will also be quicker when primed with low diversity bigrams compared to high diversity or non-legal bigrams

## Method

## Participants

Thirty-one participants ( 25 females) aged between 18 and 41 years ( $M=20.77, S D=4.17$ ) were recruited from Nottingham, UK. All participants reported English as their first language and were screened for language difficulties. Participants took part in both experiments and received research credits in exchange for their participation where applicable. An a priori power analysis showed that a sample of at least twenty-four participants was necessary to achieve statistical power of above .8.

## Experiment One

## Design

Experiment one used a LDT to assess the extent to which bigram frequency affects word recognition. The aim of the experiment was to identify any statistical priming effect that may result from high frequency word pairs within natural language.

## Materials

Three 30 -item lists were generated using bigrams found within the BNC in addition to one 90 -item non-word list which was created using entries from the ARC Non-word database (Rastle, Harrington, \& Coltheart, 2002). The BNC contains only samples of British English which increases its validity as a natural language representation for a UK sample.

Bigrams were extracted from the BNC by using a python script to parse the .xml version of the corpus into word pairs before writing them to a database and tallying the number of occurrences. This resulted in a list of $12,293,349$ unique bigrams. A further script was used to remove any bigrams with a frequency of less than 0.1 per million since these were considered too infrequent to provide meaningful data. The remaining corpus was then filtered to exclude any bigrams containing acronyms, initialisations, contractions, hyphenations, non-standard or non-English words, names, numbers expressed as digits, or words with fewer than three letters.

Table 1: Diagnostic means and standard deviations for target words

|  | Bigram Type | Log(Frequency) | Concreteness | Letters | Phonemes |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Experiment one | High Frequency | $3.20(3.86)$ | $3.10(1.09)$ | $5.01(1.48)$ | $4.10(1.24)$ |
|  | Low Frequency | $3.21(3.87)$ | $3.11(1.41)$ | $4.92(1.41)$ | $3.97(1.18)$ |
|  | Non-Bigrams | $3.20(3.86)$ | $3.10(1.08)$ | $5.00(1.48)$ | $4.05(1.27)$ |
| Experiment two | High Diversity | $2.08(0.27)$ | $2.85(0.98)$ | $5.43(1.17)$ | $4.33(1.15)$ |
|  | Low Diversity | $2.01(0.49)$ | $3.96(1.00)$ | $5.00(0.88)$ | $4.00(0.96)$ |
|  | No Diversity | $2.18(0.02)$ | $3.21(1.03)$ | $5.06(1.26)$ | $4.20(1.19)$ |

Stimuli lists were organised according to the frequency with which the bigrams occur within the BNC; the three lists contained bigrams of high ( $>100$ occurrences) or low frequency ( $<20$ occurrences), or bigrams consisting of words that do not appear together in the BNC. A number of metrics were obtained for each of the bigrams including word frequency (http://ucrel.lancs.ac.uk/bncfreq/flists.html), concreteness (Brysbaert, Warriner, \& Kuperman, 2014), number of letters, and number of phonemes. Due to the nature of the sample exact matching across conditions was impossible without compromising the number of available bigrams, however word lists were balanced so as to not differ significantly on any of these characteristics (each $p>$ 0.05 ); list diagnostics are presented in table 1 . Individual word frequencies were log-transformed. Examples of stimuli can be seen in Table 2.

Table 2: Example stimuli for experiment one

| Bigram Type | Example Stimuli (prime target) |
| :--- | :--- |
| High Frequency | recent times; last night; other hand |
| Low Frequency | craggy face; local access; time across |
| Non-bigrams | oval hipster; meet gone; chilli call |

## Procedure

Participants were presented with letter strings and were asked to indicate whether the string constituted a real English word by pressing either ' $z$ ' or ' $m$ ' on a standard QWERTY keyboard; key mapping was systematically varied so that half of all participants used ' $z$ ' to indicate a word and ' m ' to indicate a non-word whilst half responded with ' $m$ ' for words and ' $z$ ' for non-words. Strings were presented for a maximum of 1500 ms and were immediately preceded by a 75 ms prime. All prime-target pairs mapped exactly onto bigrams from the stimuli lists whereby the first word of the bigram acted as a prime and the second word as the target. A fixation point was presented in the centre of the screen for 500 ms prior to each trial. Prime-Target pairs were presented in two blocks each containing fifteen lowfrequency bigrams, fifteen high-frequency bigrams, fifteen non-bigrams, and forty-five non-word trials. The order of presentation for both blocks and trials was randomised for each participant.

## Analysis and Results

All participants scored more than $80 \%$ on the LDT. Data was then trimmed to exclude incorrect responses as well as those made faster than 200 ms , slower than 1500 ms (Perea et al., 2016), or more extreme than three standard deviations from the participant's mean (Madan et al., 2016), following this procedure $2.29 \%$ of correct trials were removed across participants.

All response time data were log-transformed; data was then analysed categorically using a repeated-measures analysis of variance to identify any differences in response
time between the high and low frequency bigrams ( $\mathrm{M}=$ 6.310, $\mathrm{SD}=.080$ ), non-bigrams ( $\mathrm{M} 6.507, \mathrm{SD}=.101$ ) and non-words ( $\mathrm{M}=6.548, \mathrm{SD}=.130$ ).

Bigram frequency had a significant effect on response time, $\mathrm{F}(3,28)=53.759, \mathrm{p}<.001, \eta_{\mathrm{p}}^{2}=.852$. Post hoc pairwise comparison using Bonferroni correction show that words in the non-bigram condition were recognised more slowly than those in both the high (p < .001) and low (p < .001) bigram frequency conditions. There was no difference between high and low frequency bigrams ( $\mathrm{p}=.305$ ). Nonwords were recognised more slowly than words in the high frequency ( $\mathrm{p}<.001$ ), low frequency ( $\mathrm{p}<0.001$ ), and nonbigram conditions ( p . .038). Figure 1 illustrates these differences.


Figure 1: Non-transformed group means for bigram frequency, bars depict standard error

A further repeated-measures analysis of variance was also conducted to assess any differences in response accuracy between the four conditions. Response accuracy also shows an effect of bigram frequency, $\mathrm{F}(3,28)=6.796, \mathrm{p}=.001, \eta_{\mathrm{p}}{ }^{2}$ $=.421$. Post hoc analyses using Bonferroni correction show that participants responded less accurately to words from the non-bigram condition than those in the high ( $\mathrm{p}=.005$ ) or low ( $\mathrm{p}=.002$ ) frequency conditions. All other comparisons were non-significant (each p>.062). Figure 2 shows means and standard error for accuracy.


Figure 2: Proportion of correct responses by group, bars depict standard error.

## Experiment Two

## Design

Experiment two used a LDT to assess the extent to which bigram diversity affects word recognition. The aim of the experiment was to identify any statistical priming effect that may result from the predictability of the second word in a bigram given the diversity of the first.

## Materials

Stimuli were obtained and processed using an identical procedure to experiment one with the exception that the word lists were organised according to high ( $>100$ potential followers) or low ( $<2$ potential followers) diversity or bigrams consisting of primes that do not have followers within the BNC. Word lists were balanced in the same way as the first experiment, each $p>0.06$ with the exception that the low diversity list differed significantly from both the high and no diversity list on concreteness (high: $\mathrm{p}<0.01$, no: $\mathrm{p}<0.01$ ); this is due to the relative scarcity of low diversity bigrams within the BNC and the theoretical decision to prioritise controlling individual word frequency since this represents the largest predictor of word recognition performance (Brysbaert \& New, 2009; Ferrand et al., 2010; Keuleers, Diependaele, \& Brysbaert, 2010; Keuleers et al., 2012; Yap \& Balota, 2009). List diagnostics are presented in table 1. Individual word frequencies were logtransformed. Example stimuli can be seen in Table 3; none of the bigrams were repeated across the two experiments.

Table 3: Example stimuli for experiment two

| Bigram Type | Example Stimuli |
| :--- | :--- |
| High Diversity | that place; with number; this ancient |
| Low Diversity | revolve around; beady eyes; gilded cage |
| No-Diversity | yonder month; ribbed final; orate red |

## Procedure

The experimental procedure was identical to that used in the first experiment.

## Analysis and Results

All participants scored more than $80 \%$ on the LDT. Data was trimmed in the same way as the first experiment and a total of $2.04 \%$ of correct trials were removed. Response time data was log-transformed.

Data was analysed categorically using a repeatedmeasures analysis of variance to identify any differences in response time between the high ( $\mathrm{M}=6.375, \mathrm{SD}=.054$ ), low $(M=6.395, S D=.059)$ and no diversity $(M=6.422$, $\mathrm{SD}=.581)$ bigrams as well as non-words $(\mathrm{M}=6.548, \mathrm{SD}=$ .130); means and standard error can be seen in Figure 3.


Figure 3: Non-transformed group means for bigram diversity, bars depict standard error

Bigram diversity had a significant effect on response time, $\mathrm{F}(3,28)=35.932, \mathrm{p}<.001, \eta_{\mathrm{p}}^{2}=.794$. Post hoc pairwise comparison using Bonferroni correction shows that nonwords were recognised more slowly than those in high (p < .001 ), low ( $\mathrm{p}<.001$ ) and no diversity ( $\mathrm{p}<.001$ ) conditions. Words in the no diversity condition were also recognised more slowly than those in both the high ( $\mathrm{p}=.007$ ) and low ( $\mathrm{p}=.011$ ) diversity conditions; there was no significant difference between high and low diversity bigrams ( $\mathrm{p}=$ .261).

Bigram diversity had no effect on response accuracy, $\mathrm{F}(3,28)=1.486, \mathrm{p}=.208$.

## Comparison with Transitional Probability

Transitional probabilities were calculated for all bigrams using the formula:

$$
\mathrm{P}(\mathrm{Y} \mid \mathrm{X})=\frac{\mathrm{P}(\mathrm{Y}, \mathrm{X})}{\mathrm{P}(\mathrm{X})}
$$

Where Y is the target stimulus and X is the initial word of a given bigram pair.

An item analysis was then run using a multiple linear regression with data from both experiments to assess the relationship between response time (log-transformed) on a LDT and the three key variables bigram frequency, bigram diversity, and transitional probability, $\mathrm{F}(3,168)=2.937, p=$ .035. Individual coefficients (see Table 4.) indicate that bigram frequency represents the strongest predictor of word recognition performance; neither bigram diversity or TP were significant predictors of response time.

Table 4: Coefficients and $p$-values

|  | Beta | Stand. <br> Beta | P |
| :--- | :--- | :---: | :--- |
| Bigram frequency | $-2.51 \mathrm{e}-5$ | -.189 | $<.016$ |
| Bigram diversity | $-1.21 \mathrm{e}-5$ | -.060 | .456 |
| Trans. Probability | -.038 | -.107 | $<.169$ |

## Discussion

The current study aimed to assess whether bigram frequency and bigram diversity would have an effect when used as primes in a LDT. Findings from the categorical analyses suggest a binary interaction between bigram frequency and response time where naturally occurring bigrams are recognised significantly more quickly than illegal bigrams or non-words. The same is also true for bigram diversity.

This suggests that any amount of exposure to a language is beneficial regardless of the frequency or diversity of individual structures within the input. This is an interesting effect which may have been overlooked by previous studies that have focussed on TPs since the methodologies
employed tend to focus on recognition of familiar versus unfamiliar strings. It could be argued however that the bigram frequencies presented in the current study, although highly infrequent, do not accurately represent the extremes of low frequency within the BNC. It is therefore suggested that further investigation needs to access frequencies of less than 0.1 per million in order to identify the absolute minimum amount of exposure required to elicit statistical priming effects.

Comparison of the key predictors also suggests that bigram frequency outperforms TPs as a predictor of response time in a statistically primed LDT. This can be attributed to the lower computational costs associated with tracking bigram frequency compared to the calculation of TPs. To the authors knowledge, the current study is the first to assess statistical learning using a LDT. These findings should therefore be interpreted with caution until they can be demonstrated in alternative paradigms.

It is proposed that the findings presented are evidence for the use of metrics other than TP in statistical learning paradigms, particularly when applied to natural language where TPs tend to be very small. A case can also be made that LDTs are a viable paradigm for the investigation of statistical effects in natural language where traditional recognition tasks may not be appropriate.

Crucially, they suggest that theories of statistical learning can deal with the scale-up in variety and complexity that comes from moving between artificial and natural languages. This begins to address one of the most fundamental criticisms of statistical learning theory.

When interpreting the data presented herein it would be prudent to consider that, although the BNC constitutes a multifarious selection of British English it does not encapsulate the entirety of written and spoken language. It is therefore posited that any findings presented be considered as representative rather than absolute in their accuracy. Future investigation should include the analysis of alternate corpora in order to ensure that any results are not artefactual in nature.

It is recognised that neither bigram frequency or diversity represent a complete account of statistic learning, nor is it suggested that learners utilise these metrics in place of TPs. Rather, it is posited that bigram frequency and, to a lesser extent, diversity constitute 'another brick in the wall' which may one day lead to a comprehensive understanding of how humans process language.

In conclusion, the current study demonstrates that individuals are capable of using bigram frequency and diversity to respond to statistical primes in a lexical decision task and that these metrics may be comparable to transitional probabilities when applied to natural language.

## Acknowledgments

Data cited herein have been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

## References

Bates, E., \& MacWhinney, B. (1987). Competition, variation, and learning. In B. MacWhinney (Ed.), Mechanisms of language acquisition. Mahwah, NJ: Erlbaum.
Brysbaert, M., \& New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41, 977-990.
Brysbaert, M., Warriner, A. B., \& Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior research methods, 46, 904-911.
Conway, C. M., \& Christiansen, M. H. (2005). Modalityconstrained statistical learning of tactile, visual, and auditory sequences. Journal of experimental psychology. Learning, memory, and cognition, 31, 24-39.
Daikoku, T., Yatomi, Y., \& Yumoto, M. (2015). Statistical learning of music- and language-like sequences and tolerance for spectral shifts. Neurobiology of Learning and Memory, 118, 8-19
Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., \& Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. Behavior Research Methods, 42, 488-496.
Freudenthal, D., Pine, J. M., Jones, G., \& Gobet, F. (2016). Developmentally plausible learning of word categories from distributional statistics. In Proceedings of the 38th Annual Conference of the Cognitive Science Society, 674679.

Glenberg, A. M., \& Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. Cortex, 48, 905-922.
Goldberg, A. E., Casenhiser, D. M., \& Sethuraman, N. (2005). The role of prediction in construction-learning. Journal of Child Language, 32, 407-426.
Jusczyk, P. W., \& Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29, 1-23.
Keuleers, E., Diependaele, K. \& Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch monoand disyllabic words and nonwords. Frontiers in Psychology 1:174.
Keuleers, E., Lacey, P., Rastle, K., \& Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. Behavior Research Methods, 44, 287-304.
Kirkham, N., Slemmer, J., \& Johnson, S. (2002). Visual statistical learning in infancy: Evidence for a domain general mechanism, Cognition, 83, B35-B42
Koelsh, S., Busch, T., Jentschke, S., \& Rohrmeier, M. (2016). Under the hood of statistical learning: A statistical

MMN reflects the magnitude of transitional probabilities in auditory sequences, Scientific Reports, 6.
Liu, L., \& Kager, R. (2011). How do statistical learning and perceptual reorganization alter dutch infants perception to lexical tones. In Proceedings of the 17th International Congress of Phonetic Sciences, 1270-1273.
Madan, C. R., Shafer, A. T., Chan, M., \& Singhal, A. (2016). Shock and awe: Distinct effects of taboo words on lexical decision and free recall, The Quarterly Journal of Experimental Psychology, 1-18
Newport, E. L., \& Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. Cognitive Psychology, 48, 127-162.
Perea, M., Marcet, A., Vergara-Martínez, M., \& Gomez, P. (2016). On the Dissociation of Word/Nonword Repetition Effects in Lexical Decision: An Evidence Accumulation Account. Frontiers in psychology, 7.
Perruchet, P., Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches, Trends in Cognitive Sciences, 10, 233-238.
Pickering, M. J., \& Garrod, S. (2007). Do people use language production to make predictions during comprehension? Trends in Cognitive Sciences, 11, 105110.

Rastle, K., Harrington, J., \& Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. Quarterly Journal of Experimental Psychology, 55A, 1339-1362.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science (New York, N.Y.).
Saffran, J. R., Johnson, E. K., Aslin, R. N., \& Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70, 27-52.
Thiessen, E. D., \& Erickson, L. C. (2013). Discovering words in fluent speech: The contribution of two kinds of statistical information. Frontiers in Psychology, 3
Toro, J. M., Sinnett, S., \& Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. Cognition, 97.
Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., \& Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 443.
Yap, M. J., \& Balota, D. A. (2009). Visual word recognition of multisyllabic words. Journal of Memory \& Language, 60, 502-529.

# Extraneous visual noise facilitates word learning 

# Katherine E. Twomey (k.twomey@lancaster.ac.uk), Lizhi Ma (malizhi110@hotmail.com) Gert Westermann (g.westermann@lancaster.ac.uk) 

Department of Psychology, Lancaster University<br>Lancaster, LA1 4YF UK


#### Abstract

Variability is important to learning; however, whether it supports or hinders language acquisition is unclear. 3D object studies suggest that children learn words better when target objects vary, however storybook studies indicate that contextual variability impairs learning. We tested a dynamic systems account in which background variability should boost learning by speeding the emergence of new behaviors. Two groups of two-year-old children saw arrays of one novel and two known objects on a screen, and heard a novel or known label. Stimuli were identical across conditions, with the exception that in the constant condition objects appeared on a white background, and in the variable condition backgrounds were colored. Only children in the variable condition showed evidence of word learning, suggesting that extraneous variability supports learning by decontextualizing representations, and indicating that adding low-level entropy to the developmental system can trigger a change in behavior.


Keywords: word learning; language acquisition; variability; memory decontextualization; dynamic systems

Children's early word learning has long fascinated researchers. When a child hears the new word spaceship, linking it with a new toy flying machine rather than their toy dog - or indeed the flying machine's wings, its color, the way it moves, and so on - seems to pose little problem. Given that the space of potential referents is theoretically infinite (Quine, 1960), this ability to quickly map a novel word to a novel object is impressive; indeed, robust referent selection has been observed in children as young as 18 months (Carey \& Bartlett, 1978; Halberda, 2006; HoustonPrice, Plunkett \& Harris, 2005; Markman \& Wachtel, 1988).

However, there is mounting evidence that a single episode of referent selection is not sufficient for full word learning; rather, children learn word-object associations incrementally, forming in-the-moment mappings between labels and objects and strengthening memories of these mappings across repeated encounters via cross situational learning (Horst \& Samuelson, 2008; Smith \& Yu, 2008; Yurovsky, Fricker, Yu, \& Smith, 2014). Clearly, then, memory and language are linked from very early in development (Taylor, Liu, \& Herbert, 2016): learning a new word depends critically on children's ability to form and retain word-object associations. Consequently, the field has recently focused on the multiple factors that affect children's ability to retain word-object mappings, demonstrating that referent selection and word learning are flexible, even fragile processes which depend heavily on the temporal and visual availability of information in the learning environment, for example repetition, competition,
and timing (e.g., Arias-Trejo \& Plunkett, 2010; Horst, Scott, \& Pollard, 2010; Mather \& Plunkett, 2009).

Developmental research has demonstrated that variability of to-be-learned items is a key influencing factor in early learning. For example, visual variability encountered across stimuli facilitates categorization in 6- to 7-month-old infants (Quinn \& Bhatt, 2010), and phonological variability in affect or speaker has been shown to support early word recognition (Rost \& McMurray, 2009). Recent work has revealed a similar effect of variability on word learning: when shown a novel 3D object category with exemplars that varied in color, 30-month-old children learned category labels, but did not when exemplars were identical, or varied in shape and color simultaneously (Twomey, Ranson, \& Horst, 2014). Thus, while some target variability supports word learning, too much variability appears to disrupt it.

In addition to target variability there is good theoretical reason to expect extraneous, non-target variability - entropy - to support word learning. Evidence from adult problemsolving studies suggests that introducing entropy to a task facilitates learning. For example, adults solving a series of gear system problems presented on a computer screen learned a short-cut solution faster when the task contained entropy in the form of variability in spatial location of the stimuli than when stimuli were presented in a consistent spatial location (Stephen, Dixon, \& Isenhower, 2009). On dynamic systems theories of cognition and development, cognitive structure emerges from the dynamic interactions of multiple, coupled components including the learner's body, learning history and in-the-moment characteristics of the task (Thelen \& Smith, 1996). Cognitive structure is instantiated as a stable state ("attractor") in the behavior of this complex system. Dynamic systems of this type exhibit "phase shifts" from one attractor to another, resulting in qualitative and quantitative changes in the system's behavior. Because phase shifts result in behavioral change, from the dynamic systems perspective, they index learning. As Stephen et al. (2009) demonstrate, extraneous entropy during learning destabilizes attractor states, speeding the onset of a phase shift. During development, then, non-target variability should speed up learning by helping new cognitive structure emerge via a shift from one behavioral state to another.

Despite this strong theoretical prediction, evidence for an effect of non-target variability in early learning is mixed. The categorization literature suggests that non-target variability helps learning. For example, Goldenberg \& Johnson (2015) presented 16- to 20-month-old infants with a looking time task. Children saw novel category exemplars
on backgrounds which (a) repeated, (b) varied randomly, or (c) varied within interleaved blocks. Only infants who saw backgrounds which varied in interleaved blocks correctly generalized category labels at test. In contrast, the word learning literature suggests that lack of contextual variability supports word learning: when learning words from a storybook, repeating the context in which 3-year-old children encounter novel words by reading from the same book repeatedly boosts word retention relative to teaching children the same novel words from multiple different books (Horst, Parsons \& Bryan, 2011; Williams \& Horst, 2014). More broadly, however, the prediction from dynamic systems theory that additional entropy should boost word learning in a single task has yet to be explicitly tested. Critically, if background variability helps children learn words, this would provide evidence for continuity in the low-level mechanisms driving learning, from toddlerhood to adulthood. The current study addressed this gap by presenting children with a word learning task in which objects appeared either on a white background or on multiple colored backgrounds. We selected two-year-old children in line with previous research which demonstrates this age group's success in similar looking-based referent selection tasks (Bion, Borovsky \& Fernald, 2013). On the dynamic systems account, children in the variable color condition should show stronger retention of label-object associations than children in the constant color condition.

## Method

## Participants

Thirty typically developing, monolingual English-learning two-year-old children ( 14 girls, $M=22.77$ months, $S D=$ 1.87 months; range $=20.0-26.0$ months) with a mean productive vocabulary of 176.04 words ( $S D=117.50$ words, range $=4-413$ words) and no family history of colorblindness participated. Half of the children were randomly assigned to the constant color condition, and half to the variable color condition. Children's ages and productive vocabularies were the same in either condition ( $p \mathrm{~s}>.30$ ). Data from six children were excluded due to fussiness (1), parental interference (3), bilingualism (1), and an eye tracker sample rate of under $25 \%$ (1). Parents were reimbursed for travel expenses and children received a small gift for participating.

## Stimuli

Each child saw a warm-up, referent selection and test phase. Critically, stimuli for each phase were identical across conditions with the exception that during warm-up and referent selection in the variable color condition objects appeared on colored backgrounds, and in the constant color condition backgrounds were always white. Children also saw engagement and attention-getting stimuli. Overall, warm-up, referent selection and retention stimuli were videos containing 2D photographic images of known and/or novel objects (depicted in Fig. 1).

|  | Known 1 | Known 2 | Novel | Novel label |
| :---: | :---: | :---: | :---: | :---: |
| Set <br> 1 |  |  | zorch |  |
| Set <br> 2 |  |  | tife |  |
| Set |  |  |  |  |
| 3 |  |  |  |  |

Fig 1. Object depicted in the current study.
Known objects were an apple, a ball, a banana, a car, a cup and a fork, and were selected because their labels are familiar to children of this age group (Fenson et al., 1993). Novel objects were a purple, green and black foam rocket (labeled zorch), a spherical yellow object with multiple flexible legs capped with pink and green balls (labeled tife), and a blue kazoo with raised orange spots (labeled blick), selected from an online database of objects unfamiliar to children of this age (NOUN Database; Horst \& Hout, 2015). Each trial consisted of a single video of three objects. Videos were created in Microsoft Powerpoint 2010, and converted to .avi format using Microsoft Windows Live Movie Maker 2011. Each video was accompanied by embedded audio consisting of the same female speaker saying Can you find the [label]? Look at the [label]! Where's the [label]?, as well as sound effects to keep children engaged in the task. Known labels were the appropriate English labels for those objects, and novel labels were blick (kazoo), tife (legs/balls) and zorch (rocket), selected as plausible but unfamiliar English object names. Auditory stimuli commenced 5 s after the start of each trial. First label onsets occurred from $0.78-0.90 \mathrm{~s}$ after the beginning of the auditory stimulus and offsets from 1.27 to 1.58 s ; second label onsets from $2.20-2.48 \mathrm{~s}$ and offsets from $2.65-3.25 \mathrm{~s}$; and third label onsets from $3.54 \mathrm{~s}-4.21$ s and offsets from 4.22 s to 5.19 s .

Engagement. Engagement stimuli consisted of a 7 s video of a female experimenter on a white background, smiling and saying Hello! Let's play a game! Can you find what I'm looking for? in child-directed speech.
Warm-up. Warm-up stimuli were 16 s videos, each depicting a set of three of the known objects, designed to familiarize children with the task. In the first 0.5 s , a small colored rectangle appeared in the middle of a black screen and spun in an anticlockwise circle, expanding until it filled the whole screen, at which point it became the background on which the objects would appear. In the constant color condition, the background on each of the three warm-up trials was white. In the variable color condition, the background was blue, green, pink, purple or red. In the next 2 s the three objects appeared in the top left-hand corner of the screen and bounced diagonally downwards accompanied by a boing sound, coming to a rest in the center of the screen and remaining there for 9.5 s , during which time the
target object was labeled three times (e.g., Can you find the apple? Look at the apple! Where's the apple). During the next 3 s the target object rotated accompanied by a twinkling sound, followed by ostensive auditory feedback (e.g., There's the apple!). In the final 1 s the objects bounced diagonally towards the bottom right hand corner and offscreen, accompanied by the sound of children cheering.
Referent selection. Referent selection trials were 13 s long, and identical to warm-up trials with the exception that children saw one novel and two known stimuli, and there was no ostensive feedback phase. Background colors were either white (constant color) or pseudorandom (variable color), as in the warm-up trials. Object location was pseudorandomized.
Retention. Retention trials were 9.5 s long and proceeded in an identical manner to referent selection trials except that the background was always gray and appeared immediately (i.e., there was no 0.5 s period where the background appeared) and all three objects were novel. Each object was labeled on two trials.

## Procedure and Design

Before the experiment began the experimenter showed caregivers pictures of the known and novel objects to ensure they were appropriately known and novel to the child. All children were familiar with the known objects and unfamiliar with the novel objects. Caregivers were asked to complete a UK adaptation (Hamilton, Plunkett, \& Schafer, 2000) of the MacArthur-Bates Communicative Development Inventory (Fenson et al., 1994), a vocabulary inventory commonly used to score toddlers' receptive and productive vocabulary. Caregivers completed the vocabulary inventory either before or after the experiment, depending on the child's level of engagement.
The eyetracking session took place in a quiet, dimly-lit room. Children sat on their caregiver's lap 50-70 cm in front of a $21.5 " 1920 \times 1080$ computer screen. Beneath the screen a Tobii X120 eyetracker recorded the child's gaze location at 17 ms intervals, and a video camera above the screen recorded the caregiver and child throughout the procedure. Caregivers were instructed not to interact with their child or look at the screen during the task to avoid biasing their child's behavior, and were asked to sit at a $90^{\circ}$ angle from their child to ensure the eyetracker tracked the child's eyes only.

The eyetracker was first calibrated using a five-point infant calibration procedure available in Tobii Studio. Immediately following calibration, children saw the engagement stimulus once.

Warm-up. The three warm-up trials immediately followed the engagement stimulus. The warm-up phase in each condition was identical with the exception that in the constant variable color condition, backgrounds were multiple, uniform colors, while in the constant color condition, backgrounds were white. Which objects appeared, which served as targets, and left-right positioning
of objects were pseudorandomized across children such that no object appeared on more than two successive trials.
Referent selection. Fifteen referent selection trials immediately followed the warm-up phase. An example referent selection phase for the variable color condition is depicted in Fig. 1. Again, the corresponding warm-up phase in the constant color condition was identical with the exception that backgrounds were white. Referent selection trials were presented in three blocks of five trials for each set. Sets were kept constant across trials to maximize children's retention of novel labels (Axelsson \& Horst, 2014); thus, one child might see a block of five repetitions of the apple + fork + zorch set, followed by the banana + cup + tife set, and finally the car + ball + blick set, with block order Latin square counterbalanced across children. In each referent selection block children were asked to look at a known object on two trials and a novel object on three trials. Known/novel trial order and background color (variable color condition only) was pseudorandomized such that no more than two of the same trial type appeared in succession.


Fig. 2. Example referent selection phase.
During referent selection an attention getting stimulus appeared six times pseudorandomly such that it was always succeeded by at least one referent selection trial, and consisted of a 3 s video of the speaker saying What's next? Finally, after the referent selection phase, children saw a 5 s "Well done" video of the speaker saying Well done! All finished! See you soon!
Break. Following referent selection, children took a fiveminute break. During this time they either remained on their caregiver's lap and watched an age-appropriate animation or moved to a seating area in the same room and colored pictures from a book.
Warm-up. After the break children saw a further warm-up trial, presented on a gray background.
Test. Three memory recativtaion and three retention trials immediately followed the warm-up trial, each depicting the three novel objects seen during referent selection. Trial order and object location were pseudorandomized.
Coding and data cleaning. Left, middle and right AOIs were square and centered on each object's stationary position after they had bounced into the screen. Unreliable/offscreen and non-AOI looks were discarded,
resulting in a final dataset of 115,762 referent selection and 61,247 test gaze samples. Individual gaze samples were numerically coded ( $1=$ target look, $0=$ non-target look $)$, creating a raw looking time measure, which was further collapsed into 100 ms time bins for statistical tractability. All subsequent analyses use this target looking measure, and are standardized from the offset of the first label plus 233 ms (Swingley, Pinto, \& Fernald, 1999) to 6733 ms postlabeling.

## Results

Because the focus of the current paper is the effect of extraneous variability on children's word learning, and due to space constraints, we present here the results from the test phase. Analyses of looking during referent selection are reported separately and discussed in detail in Twomey, Ma \& Westermann (under review); overall, however, we found chance level looking and no difference between conditions. At test, each novel object served as a target on one memory reactivation trial and one retention trial. Fig. 3 depicts looking times during the memory reactivation trials and shows little difference in target looking in the two conditions. This conclusion was supported by a linear mixed effects model with main effects of time bin (treated as continuous) and condition and their interaction, with byparticipant random slopes and intercepts for condition and by-item random intercepts to rule out item effects (Barr, Levy, Scheepers \& Tily, 2013). As in referent selection, there was a small but robust increase in looking with time (beta $=0.0019, S E=0.00063, t=2.99, \chi^{2}(1)=10.49, p=$ .0012). However, condition had no independent effect on looking times, and did not interact with time bin (main effect of condition: beta $=0.043, S E=0.00098, t=0.66$, $\chi^{2}(1)=0.12, p=.73$; time bin x condition interaction: beta $=$ $\left.-0.0080, S E=0.00098, t=-0.81, \chi^{2}(1)=0.67, p=.41\right)$.

Data from the three retention trials show a markedly different pattern, however. As Fig. 4 illustrates, children in the variable color condition looked at the target at abovechance levels immediately following labeling and again at around 4000 ms , suggesting that encountering variable colored backgrounds during referent selection facilitated their retention of the novel label-object mappings. A mixed effects model with the same fixed effects structure as above and by-participant and by-item random intercepts and slopes for condition revealed that target looking decreased over time (time bin: beta $=-0.029, S E=0.00078, t=-3.65, \chi^{2}(1)$ $=32.55, p<.001$ ). This effect was constant for children in either condition (time bin x condition: beta $=-0.0010, S E=$ $\left.0.0011, t=-0.87, \chi^{2}(1)=0.75, p=.39\right)$. Critically, however, proportion target looking was greater for children in the variable color condition than in the constant color condition (beta $\left.=-0.26, S E=0.090, t=2.85, \chi^{2}(1)=5.41, p=.020\right)$.


Fig. 3. Proportion target looking during memory reactivation trials. Error bars represent 95\% CIs. Where bins are marked with a point, looking is significantly above chance ( $0.33 ; p<.05$, one-sample, two-tailed $t$-tests; ditto for Fig. 4).


Fig. 4. Proportion target looking during retention trials.

## Discussion

The current study explored whether extraneous variability would boost young children's word learning. We trained two groups of two-year-old children with novel label-object associations via multiple referent selection trials. Stimuli presented to both groups were identical except that half the children saw arrays of novel objects displayed on a white background (constant color condition), and half saw objects on multiple colored backgrounds (variable color condition). Analyses of test trials revealed a clear effect of background variability: while children did not appear to correctly identify previously-seen novel objects during the memory reactivation trials, on retention trials children who had seen variable backgrounds during referent selection looked for longer at target objects than did children who had seen constant colored backgrounds, and did so at levels greater would be expected by chance. Thus, infants who had seen objects on variable backgrounds learned and retained the novel object-label mappings, but infants who had seen the objects on a constant background did not. These results offer converging evidence that following reactivation of memory traces, background variability facilitates learning, raising several interesting issues (see also Twomey et al., under review).
The importance of memory reactivation Children in the variable color condition looked at target objects at chance levels on the three memory reactivation trials, but at levels
greater than expected by chance on the subsequent three retention trials. Typically in word learning studies children see only a single retention trial for each object. Our results suggest that null findings in these studies could be due to a lack of recall ability rather than a lack of learning. These data indicate that including memory reactivation trials in future studies could help shed light on whether children are failing to learn, or failing to recall. Establishing the locus of memory reactivation in the word learning field is therefore critical for a thorough understanding of the delicate memory processes underlying early language acquisition.

Decontextualization in early learning The fact that only children in the variable color condition retained novel labelobject associations may seem unexpected in light of recent work in word learning indicating that consistency in context supports, not impairs, word learning (e.g., Axelsson \& Horst, 2014). Given these results, why should what seems to be a more challenging task (i.e., variable color versus constant color backgrounds) lead to better learning? In fact, our results are in line with a wealth of adult literature demonstrating that background variability supports recall (e.g., Godden \& Baddeley, 1975). More recent work has explored the effect of context on adults' category learning; for example, Finch, Carvalho and Goldstone (2016) showed that variable backgrounds led to better retention of previously-seen exemplars of a bird category.

These results are attributed to a decontextualisation mechanism. When memories are formed after a single encounter, both context and target are encoded. On subsequent encounters, if the context stays the same, it remains part of the representation. These context-dependent memories are harder to recall when the context changes. Godden and Baddeley (1975) describe a classic example of this effect, showing that divers who had learned word lists either on dry land or underwater were better at recalling words learned underwater when tested underwater, and better at recalling words learned on dry land when tested on dry land. When an item is encountered in multiple different environments, however, the representation becomes decontextualized: the context becomes less important to the representation. If an item with a decontextualized representation is encountered in a new environment, then, it is easier to recall than if the representation were contextdependent.

The same mechanisms that explain these adult data can account for children's word learning in the current study. During referent selection, children in the constant color condition learned context-dependent representations, while children in the variable color condition learned decontextualized representations. At test children encountered objects on a gray screen - and critically, neither group had seen objects presented on a gray screen until this point. Thus, recall was possible for children in the variable color condition, who were able to generalize their decontextualized memory traces to the new test context. This raises the question of why contextual consistency in existing studies supports word learning - the opposite of
the current findings. It is possible that different types of context have qualitatively different effects. Here, in line with Stephen et al. (2009), "context" was low-level, extraneous variability. In contrast, the contexts in the existing literature were rich and salient: in the storybook studies, books were constructed from photograph-like images, resulting in a complex visual scene that varied from page to page. In addition, the sentence contexts in which novel words appeared also varied (Horst et al., 2011). Similarly, in the referent selection work, "context" consisted of the competitor objects presented alongside the targets, which were considerably more complex than a simple block of color (Axelsson \& Horst, 2014). Thus, it may be that in rich learning environments, restricting complexity supports learning (Radesky \& Christakis, 2016), while in simpler learning environments, increasing complexity by adding background noise helps learning.

This decontextualization account provides a mechanism by which added variability can support learning, as predicted by the dynamic systems account. Importantly, decontextualization is one among many potential mechanisms by which learning under the dynamic systems account may be shaped. As noted above, this theory predicts that background entropy should facilitate learning by speeding up the emergence of new stable behavioral states (Stephen et al., 2009). However, the dynamic systems account also suggests that other types of variability should support learning, raising the intriguing possibility for future work that entropy introduced in a different modality, for example sound or spatial location, could also support word learning. Work is underway to test these predictions. Overall, however, on either the specific decontextualization account or the broader dynamic systems approach, the current work extends a well-established phenomenon in adult cognition to children with a new task, pointing to a view of development as a continuous process driven by domain-general mechanisms.

## Acknowledgements

Data are available on request from the authors. The support of the Economic and Social Research Council is gratefully acknowledged (LuCiD: ES/L008955/1; Future Research Leaders grant to KT: ES/N01703X/1). GW was further supported by a British Academy/Leverhulme Trust Senior Research Fellowship (SF150163). We thank the parents and children who made this work possible.

## References

Arias-Trejo, N., \& Plunkett, K. (2010). The effects of perceptual similarity and category membership on early word-referent identification. Journal of Experimental Child Psychology, 105(1-2), 63-80.
Axelsson, E. L., Churchley, K., \& Horst, J. S. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. Frontiers in Psychology, 3.

Axelsson, E. L., \& Horst, J. S. (2014). Contextual repetition facilitates word learning via fast mapping. Acta Psychologica, 152, 95-99.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bion, R. A., Borovsky, A., \& Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel wordobject mappings in relation to vocabulary learning at 18 , 24, and 30months. Cognition, 126(1), 39-53.
Carey, S., \& Bartlett, E. (1978). Acquiring a single new word. Papers and Reports on Child Language Development, 15, 17-29.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., \& Pethick, S. J. (1994). Variability in early communicative development. Monographs of the Society for Research in Child Development, 59(5), R5-+.
Finch, D., Carvalho, P., \& Goldstone, R. L. (2016). Variability in category learning: The effect of context change and item variation on knowledge generalization. In Papafragou, A., Grodner, D., Mirman, D., \& Trueswell, J.C. (Eds.) (2016). Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX.: Cognitive Science Society.
Godden, D. R., \& Baddeley, A. D. (1975). Contextdependent memory in two Natural environments: On land and underwater. British Journal of Psychology, 66(3), 325-331.
Goldenberg, E. R., \& Johnson, S. P. (2015). Category generalization in a new context: The role of visual attention. Infant Behavior and Development, 38, 49-56.
Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. Cognitive Psychology, 53(4), 310-44.
Hamilton, A., Plunkett, K., \& Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. Journal of Child Language, 27(3), 689-705.
Hildreth, K., \& Rovee-Collier, C. (1999). Decreases in the response latency to priming over the first year of life. Developmental Psychobiology, 35(4), 276-289.
Horst, J. S., \& Hout, M. C. (2015). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. Behavior Research Methods, 1-17.
Horst, J. S., Parsons, K. L., \& Bryan, N. M. (2011). Get the story straight: contextual repetition promotes word learning from storybooks. Frontiers in Psychology, 2.
Horst, J. S., \& Samuelson, L. K. (2008). Fast mapping but poor retention by 24 -month-old infants. Infancy, 13(2), 128-157.
Horst, J. S., Scott, E. J., \& Pollard, J. P. (2010). The role of competition in word learning via referent selection. Developmental Science, 13(5), 706-713.

Houston-Price, C., Plunkett, K., \& Harris, P. (2005). "Word-learning wizardry" at $1 ; 6$. Journal of Child Language, 32(1), 175-189.
Hsu, V. C., Rovee-Collier, C., Hill, D. L., Grodkiewicz, J., \& Joh, A. S. (2005). Effects of priming duration on retention over the first $11 / 2$ years of life. Developmental Psychobiology, 47(1), 43-54.
Markman, E. M., \& Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. Cognitive Psychology, 20(2), 121-157.
Mather, E., \& Plunkett, K. (2009). Learing words over time: The role of stimulus repetition in mutual exclusivity. Infancy, 14(1), 60-76.
Morgan, K., \& Hayne, H. (2006). Age-related changes in memory reactivation by 1 -and 2 -year-old human infants. Developmental Psychobiology, 48(1), 48-57.
Quine, W. V. O. (1960). Word and object: An inquiry into the linguistic mechanisms of objective reference. Cambridge: MIT Press.
Quinn, P. C., \& Bhatt, R. S. (2010). Learning perceptual organization in infancy: The effect of simultaneous versus sequential variability experience. Perception, 39(6), 795806.

Radesky, J. S., \& Christakis, D. A. (2016). Keeping children's attention: The problem with bells and whistles. JAMA Pediatrics, 170(2), 112-113.
Rost, G. C., \& McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. Developmental Science, 12(2), 339-349.
Smith, L. B., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Stephen, D. G., Dixon, J. A., \& Isenhower, R. W. (2009). Dynamics of representational change: Entropy, action, and cognition. Journal of Experimental Psychology: Human Perception and Performance, 35(6), 1811.
Swingley, D., Pinto, J. P., \& Fernald, A. (1999). Continuous processing in word recognition at 24 months. Cognition, 71(2), 73-108.
Taylor, G., Liu, H., \& Herbert, J. S. (2016). The role of verbal labels on flexible memory retrieval at 12 -months of age. Infant Behavior and Development, 45, Part A, 11-17.
Thelen, E., \& Smith, L. B. (1996). A Dynamic Systems Approach to the Development of Cognition and Action. Cambridge, Mass.: MIT Press.
Twomey, Ma \& Westermann (under review). All the right noises: Background variability helps early word learning. Infancy.
Twomey, K. E., Ranson, S. L., \& Horst, J. S. (2014). That's more like it: Multiple exemplars facilitate word learning. Infant and Child Development, 23(2), 105-122.
Yurovsky, D., Fricker, D. C., Yu, C., \& Smith, L. B. (2014). The role of partial knowledge in statistical word learning. Psychonomic Bulletin \& Review, 21(1), 1-22.

# Comparison strategies in the change detection task are influenced by task demands. 

Rob Udale<br>University of Bristol<br>Simon Farrell<br>University of Western Australia<br>Chris Kent<br>University of Bristol


#### Abstract

Current models of visual working memory (VWM) assume that comparing memory with the environment obligatorily involves a spatial comparison process. Can changing task demands determine whether a spatial or non-spatial comparison processes is employed? Study displays of three colored shapes were presented, followed by test displays of three coloured shapes. Participants decided whether a feature changed between displays. Task-irrelevant changes to the probed item's locations or feature bindings reduced memory performance, suggesting that participants employed spatially guided comparison process. This finding occurred irrespective of whether participants decided about the whole display, or only a single cued item within the display. When task-irrelevant feature changes occurred amongst uncued items, performance was unaffected by irrelevant changes in location or feature bindings. These results suggest that participants can flexibly shift comparison strategy in response to changing task demands. These findings have implications for models of VWM, which assume obligatory location-based comparisons in VWM.


# Weight matters: The role of physical weight in non-physical language across age and culture 

Tomer D. Ullman (tomeru@mit.edu)<br>Brain and Cognitive Sciences, MIT<br>Santiago Alonso-Diaz (salonsod@ur.rochester.edu)<br>Brain and Cognitive Sciences, University of Rochester

Stephen Ferrigno (sferrigno@ rcbi.rochester.edu)<br>Brain and Cognitive Sciences, University of Rochester

Sarina Zahid (szahid@u.rochester.edu)<br>Brain and Cognitive Sciences, University of Rochester

Celeste Kidd (celestekidd@gmail.com)
Brain and Cognitive Sciences, University of Rochester


#### Abstract

Languages commonly use physical properties to discuss distinctly non-physical states and events in the world (e.g., "I'm not a huge fan of licorice"). Here, we investigate the degree to which associations between physical properties and abstract concepts are culturally specific constructs. To do this, we tested three distinct populations-US adults, US children, and adults from an indigenous group in the lowlands of Bolivia, the Tsimane'-on their associations between the physical concept of weight and a variety of abstract attributes (e.g., importance, emotional state, moral worth). We find a strong relationship between the associations of US and Tsimane' adults, but little-to-no relationship between US children and either adult population. These results suggest that the property of weight plays a similar role in everyday thought across cultures, but that it takes time to develop. Further, we found that these associations could not be recovered from a simple semantic embedding analysis, suggesting that the cross-culturally shared connections between physical and abstract attributes may be learned through more complex experiences than language alone.


Marty: Are you trying to tell me that my mother has got the hots for me?
Doc: Precisely!
Marty: Whoa, this is heavy.
Doc: There's that word again: "heavy." Why are things so heavy in the future? Is there a problem with the Earth's gravitational pull?
(Back to the Future. Dir. Robert Zemeckis)

## Introduction

Physical notions weigh in on everyday conversation. We say a person forced herself to meet a deadline, as though she is pushing a cart uphill. We say a deadline is fast approaching, as though an actual train hurtling towards our location in time.

Our concepts of force, causality, space, and substance seem to shape how we talk about the world (Talmy, 1988; Pinker, 2007). Certainly some abstract thoughts rely on a universal understanding of the physical world. If a friend describes writing a paper as 'I'm banging my head against the wall', we can understand they are frustrated, and not literally writing a paper on the effects of head-banging (Figure 1, right). In the reverse direction, our language also shapes how we think about basic concepts, such space and time (e.g. Boroditsky, 2001, Núñez \& Sweetser, 2006). And some thoughts are culture- and language-dependent in their meaning. If a friend says writing feels like carrying the day out in a basket, the thrust would not be universally recognized (Figure 1, left).


Figure 1: Details from Flemish Proverbs by Pieter Bruegel the Elder, 1559. Left: Carrying the day out in a basket, i.e. wasting time Right: Banging one's head against the wall.

The purpose of this paper is not to untangle the knot of development, culture, language, and physical concepts. Rather, we mean to pick up one strand of thought as it relates to an understanding of a physical quality, and to tug on it gently. In particular, we consider the concept of weight (heavy and light), which has received less attention in terms of its impact on thought, compared to concepts like force, space, and time 1 . Intuitively, we seem to associate weight with worth: In 2015, a technical teardown of Beats headphones found that $30 \%$ of their weight was accounted for by metal objects that add no function, but make them feel 'solid and valuable' Einstein, 2015). When a character in Romeo and Juliet exclaims 'O heavy day!', we recognize that as an expression of dismay at an unfortunate event. But is all this because of the quirks of the English language, and our own WEIRD makeup (Henrich et al. 2010), or something deeper about the way weight ties in with concepts such as worth and sadness?

[^533]To examine this question, we asked three groups (US adults, US children, and adults from the indigenous Tsimane' of Bolivia) to pick which of two differently weighted, visually identical boxes was better described by various attributes ( external, internal, mental and non-mental). We reasoned that if the three groups show a systematic bias for some attributes within the group, but no relationship is found between the judgments of these three groups, then the use of a weight concept in our everyday thinking outside a strictly physical context is more likely to be a cultural construct. If the three groups all show similar judgments, then this is evidence in favor of an early shared conceptual organization. If the Western adults show similar judgments to Tsimane' adults, but are not similar to children, this would suggest a shared conceptual organization, but one that takes time to develop. A final option is that all groups will show random behavior, failing to associate any attribute with the boxes in a systematic way, which would be evidence of certain poor decisions about experiment design or a problem with the fundamental research question. The authors were agnostic about the most likely outcome out of the ones just listed.

## Experiment 1: US Adults

## Participants

Participants ( $N=100$, 42 female, median age 32.0 years) were recruited through Amazon's Mechanical Turk service (Crump et al. 2013) and paid a monetary sum for their participation, equivalent to $\$ 9$ per hour. Participants were restricted to those living in the United States.

## Materials and methods

Participants were presented with an image of identical boxes marked A and B (see Figure 2, top), and asked to imagine that there were two boxes before them, as in the image. Participants were asked to imagine lifting up the boxes and discovering that one of the boxes is much heavier (the identity of the heavy box was randomized across participants).

Participants read 12 descriptions in succession, choosing the box that best fit the description. For each description, participants were reminded which box was heavier, and then given a prompt as follows: "Which box is [attribute]?", where the attribute varied from one question to the next. Participants indicated their response using a radio button. The 12 attribute adjectives were presented in random order, chosen from a list of 24 possible attributes that reflect inner traits (e.g., good/bad), external qualities (pretty/ugly), emotions (sad/happy), and external evaluation (cheap/expensive, important/unimportant). For a full list, see Table 1.

Each participant saw only 12 attributes, rather than the full list of 24 , to prevent cognitive fatigue. Participants always saw only one of a possible antonym pair. In total, this meant there were 50 individual ratings per attribute. Following the attribute questions, the participants supplied basic demographic information, and were invited to share any comments they may have.

| Important* / Unimportant* [Not Important] |
| :--- |
| Valuable* / Cheap* ; Old / Young |
| Serious* / Funny* ; Sad* / Happy* |
| Ugly* / Pretty*; Interesting* / Boring* |
| Mean / Nice ; Smart* / Stupid* [Not Smart] |
| Good* / Evil* [Bad] ; Angry / Calm |
| Brave / Coward [Scared] |

Table 1: The 24 attributes applied to the boxes in Experiments 1 and 2, grouped into antonyms. Attributes in [parentheses] indicate a child-friendly replacement for the preceding word, used in Experiment 2. Asterisks indicate words used in Experiment 3.


Figure 2: (i) Illustration of stimuli shown to participants in Experiment 1, for a specific attribute (ii) The boxes used in Experiment 2 with children, and in Experiment 3 with Tsimane' adults.

## Results and analysis

Participants' ratings for each attribute were converted into the following measure: $\frac{\# \text { participants that chose light }}{\# \text { total participants }}$. This weight choice fraction (WCF) goes from 0.0 (all participants chose the heavy box for this attribute) to 1.0 (all participants chose the light box). The results are shown in Figure 3, with bars indicating 95\% bootstrapped confidence intervals (CI, 1000 samples per attribute) around the WCF measure.

Of the 24 attributes, 15 had WCFs with CIs that do not overlap 0.5 , indicating that participants considered these attributes as statistically significantly associated with heavy or light. The same result is obtained when using a two-tail binomial test at the $p=0.05$ level. Such a result is highly unlikely to occur by chance: Using an additional bootstrap analysis that repeats the same procedure from the previous paragraph (counting the number attributes with WCF CIs that do not overlap 0.5), the median (and mean) expected number of attributes with a measure that does not overlap chance is 2. Also, the empirical distribution of participants' WCF is statistically significantly different from a distribution drawn from a random sample that assumes the same participant numbers, but with answers based on an unbiased coin flip (Kolmogorov-Smirnov two-sided test for 2 samples, $K S=0.28, p<0.05$ ) .

The attributes significantly associated with heavy and light seem partially in line with intuition ${ }^{2}$ Heavier boxes are more likely to be seen as valuable, important, and interesting, as opposed to the cheap, unimportant, and boring lighter boxes. This is consistent with the marketing-driven decision by Beats to add superfluous weight to their headphones. This association makes sense given that more weight may imply more "stuff", which could generally be considered more desirable ${ }^{3}$. Participants also associated more personality-type traits with the boxes, in a way that is not accounted for by a simple positive-negative spectrum. Heavy boxes are more good and brave, but also mean and angry. Lighter boxes are more cowardly, but also more pretty. Presumably participants were able to anthropomorphize the boxes to some degree, seeing them as agents. For example, on this analysis a light agent is more likely to run away, and is likely to be younger. However, this does not account for the full pattern of results, such as seeing heavier boxes as more "good" and less "evil".

This pattern also cannot be recovered from a semantic embedding analysis. The analysis worked as follows: We embedded the attributes from Table 1, as well as the words heavy and light, in a high-dimensional semantic vector space, which was constructed using the co-occurrence statistics of several hundred-thousand words in a large corpus Pennington et al. 2014). Specifically, we used 100-dimensional GloVe word vectors pre-trained on the Wikipedia 2014 + Gigaword 5 datasets. Such semantic embeddings have proved useful for measuring similarity between words, in the service of machine-learning applications such as sense-making, translation, and question answering (see for example Vedantam et al., 2015, Wolf et al., 2014, Yu et al., 2015). Intuitively, a shorter euclidean distance or larger cosine similarity between two points in this space indicate a larger degree of similarity between the words that those points represent. After embed-

[^534]| Heavy $0.0 \quad 0.25 \quad 0.5$ | $0.75 \quad 1.0$ Light |
| :---: | :---: | :---: | :---: | :---: |
| Important |  |
| Valuable |  |

Figure 3: Results of Experiment 1: Participant responses per attribute are converted into WCF measure running from 0 (all participants chose the heavy box) to 1 (all participants chose the light box). Bars indicate $95 \%$ bootstrapped confidence intervals around the mean of this measure. Colors indicate the degree to which an attribute is associated with heavy (blue) or light (red). Beige indicates WCFs with CIs overlapping 0.5, indicating a random response or equal association.
ding our terms, we measured the relative euclidean distance between the attributes and the terms heavy and light (that is, distance (heavy, attribute $i_{i}$ ) - distance (light, attribute $i_{i}$ )). We found no correlation between participants' response and this distance. A similar analysis with cosine similarity also found no such relation. This suggests that while useful, basic semantic embedding does not necessarily capture association that is based in physical properties.

While the pattern shown by US adults is interesting on its own, the original driving motivation was comparing this pattern to children and non-US cultures. With that, we turn to children.

## Experiment 2: US Children

## Participants

Fifty individuals were recruited from the Rochester Kid Lab participant pool ( 28 female, Median 4.0 years, range $3-\ell^{4}$ ).

## Materials and Methods

Participants were tested in a designated room in the Rochester Kid Lab. Parents gave their informed consent, and generally did not accompany their children during the test, unless requested, or children expressed shyness. Parents who accompanied their children were explicitly advised not to encourage responses from their child. Families were compensated for their time and child participants were also given a small gift (a shirt or toy).

In the testing room, participants were asked to sit next to a table, where two boxes were laid out. The boxes were $3 \times 3 \times 3$ inches, made of wood, and covered in blue fabric with a gold pattern (see Figure 2, bottom). The boxes were hollow, and inside one of them was a 200 gram metal weight, along with padding to prevent the weight from bouncing and rattling when the box was handled. The locations of the boxes with respect to a participant were randomized across children.

Participants were first asked if they noticed a difference in the boxes, based on visual appearance. Participants were then asked to hold the boxes, and to indicate if there was a difference (which they were able to verbally verify).

The participant continued to hold the boxes in each hand, as they answered the following question: "Which box do you think is [attribute]?", for a randomized set of 12 attributes taken from 24 attributes similar to Experiment 1 (and see Table 1). The study took a maximum of 10 minutes. Participants answered verbally or with a gesture, with the experimenter noting their response. Participants were also asked to explain their answers, but their reasons were scantly supplied and proved inconsistent across children ${ }^{5}$ Again, to prevent cognitive fatigue, participants were asked to judge 12 attributes rather than the full 24 , with each participant seeing only one of a possible pair of antonyms. Thus there were 25 individual ratings per attribute.

## Results and analysis

Participants' ratings for each attribute were again converted into the WCF measure used in Experiment 1 (with 0.0 indicating all participants chose heavy, and 1.0 indicating all chose light). In this case, however, only 3 attributes were

[^535]different from chance, using a two-tailed binomial test at the $p=0.05$ level). The empirical distribution of children's WCFs was also not statistically significant from a distribution drawn from a random sample, one that assumes the same participant numbers but with answers based on the flip of an unbiased coin (Kolmogorov-Smirnov two-sided test for 2 samples, $K S=0.21, p=0.22$ ). It is possible to conclude that children did not understand the task, either because of lowlevel explanations like inappropriate materials and framing, or because physical weight does not play a similar association role in their general thought as it does for US adults.

When correlating with the responses of adults from Experiment 1 , we find there is a weak correlation $\left(r_{s}=0.4, p=0.05\right.$, and see Figure 4. Taken in a positive light, this may indicate a fledgling understanding after all of the full adult association between the attributes used and physical weight. Still, this relationship is statistically tenuous. A median-split by age does not show a difference between younger and older children.

Is it possible US adults exhibit a culture-specific pattern of association with physical weight, one that requires years to acquire? In the last experiment, we consider a non-WEIRD adult population, an indigenous people of Bolivia.


Figure 4: Comparison of adult responses from Experiment 1 with child responses from Experiment 2. The x and y axis both use the same weight-fraction measure, going from heavy to light. The shaded area indicates a $95 \%$ bootstrapped confidence interval on the linear regression model fit.

## Experiment 3: Tsimane’ Adults

The Tsimane are a native people of lowland Bolivia, consisting of several thousand individuals, who live in mostly small
communities in the northeastern department of Beni. Traditional Tsimane' are farming-foragers who subsist off hunting, fishing, and some farming and trade. Members of the Tsimane' have highly variable education levels, and own few artifacts (Huanca, 2006, Reyes-García, 2001). As members of a relatively isolated non-industrial society, Tsimane' have been the topic of several previous studies, from market behavior (Reyes-García, 2001) to counting (Piantadosi et al., 2014), to color concepts (Cibelli et al. 2016), to notions of fairness (Jara-Ettinger et al. 2016).

## Participants

Our final sample included fifty-five individuals ( 33 female, median age $=28.0$ years, range 17-65) from twelve Tsimane' communities.

## Materials and methods

Experiments took place in a community classroom, with a translator reading from a script, and a separate transcriber recording responses. The experiments were conducted in Tsimane', translated from a Spanish script. The translation was confirmed by a second Spanish-Tsimane' translator. Other people were present in the room, but could not see participant responses. Participants were compensated with gift bags equivalent to roughly $\$ 10$ per hour. Participants completed other tasks in addition to the one in this study, with a total testing time of approximately one hour.


Figure 5: Comparing US adults from Experiment 1 with Tsimane' adults from Experiment 3. The x and y axis both use the same WCF measure, going from heavy to light. Shaded area indicates a $95 \%$ bootstrapped confidence interval.

Participants were presented with two identically marked
boxes on a table. These were same boxes used in Experiment 2, and weighted as in Experiment 2 (see Figure 2 bottom). Participants were instructed to pick each box up before any questions were asked. Participants were then asked: "Which box is [attribute]?", and were instructed to point to a box. This was then repeated until all adjectives were covered. The order of the adjectives was randomized, as was the particular adjective from a given pair was randomized. Participants were allowed to pick up the boxes at any point. As participants saw only one word out of a possible pair, there were on average 21 individual ratings per attribute. In general, participants in this experiment went over a subset of 16 of the 24 adjectives in Table 1, due to translation difficulties.

## Results and analysis

As in Experiments 1 and 2, participants' ratings for each attribute were transformed into a WCF measure. Four of the 16 attributes were different from chance, using a two-tailed binomial test at the $p=0.05$ level). In addition, the empirical distribution of Tsimane' WCFs is statistically significant from that drawn from a random sample that assumes the same participant numbers, but with answers sampled from an unbiased coin flip (Kolmogorov-Smirnov two-sided test for 2 samples, $K S=0.36, p<0.05$ ).

We also correlated Tsimane' responses with those of US adults in Experiment 1. We found a significant correlation ( $r_{s}=0.8, p<0.001$, and see Figure 5]. As a final comparison, we correlated Tsimane' responses with those of US children in Experiment 2, and found no significant correlation ( $r_{s}=$ $0.0, p=0.94)$. We next consider the general pattern of results.

## Discussion

Thoughts weigh nothing, but they can weigh heavily on us. A man might feel lighthearted after dispensing with a heavy obligation. We can take matters lightly, but we should not take them too lightly.

Thoughts, obligations, and matters don't actually weigh anything, but we feel their press on us. Our language of thought cleaves the world into concepts that behave like objects with physical properties, located in space and acted on by force Pinker (2007). Conversely, our mental concepts can color our perception of the physical. In this paper we considered the particular physical notion of weight, and its relation to different non-physical qualities such as value, interest and seriousness.

We examined people's associations between weight and these different qualities in Western adults and children, and in members of the non-industrial Tsimane' society. We found a strong relation between the answers given by Tsimane' and Western adults, a tenuous relation between Western adults and children, and no relation between Tsimane' adults and Western children. Taken together, these findings indicate that weight acts a similar cross-cultural role in everyday thought, but that it takes time to fully get its act together. So, it may be language and culture-independent to think of important matters as physically weighing more, for example. However, a
fuller treatment would require relating the attributes to other measures beyond weight, such as imageability and affect.

Different alternative explanations can be put forward for why children provided responses that were inconsistent with one another. First, it is possible that children simply have not had the life experiences required to form strong, systematic associations between abstract attributes and physical properties like weight. Alternatively, it is possible they cannot anthropomorphize the boxes. This seems unlikely, as children can engage in pretend play with inanimate objects, but attributing metarepresentations may have required a more active signaling of the task as pretend play Lillard (1993). It may be that young children lack the basic physical skills associated with telling a heavy object from a light object and predicting their different behaviors, but previous research shows most of the basic intuitions are in place by the lower end of our age range, with young children predicting the effects of different masses interacting, and taking weight into account when planning actions (Baillargeon, 2004, Upshaw \& Sommerville, 2015). Under these alternative explanations that posit children could have experienced confusion about the task, we would typically expect certain behavioral indicators of this state, such as failures, resistance or delays in providing responses. However, in our sample we observed no such indicators. Children were generally swift and willing to select a particular box for each attribute about which we inquired.

This study does not give a definitive final answer to questions of culture, development, and the constructs of thought, but it does shed light on a piece of the puzzle, in the form of weight. Also, it is not hard to think of other physical properties that our methodology could stretch to accommodate, roughly speaking.

## Acknowledgments

We are grateful to Holly Palmeri for her support recruiting and scheduling children in the Rochester Kid Lab, to Tomas Huanca for logistical help with visiting the Tsimane' villages, to Ted Gibson and Julian Jara-Ettinger for provoding support in Bolivia, and to Dino Nate Aez, Robertina Nate Aez, and Salomon Hiza Nate for help translating. We would also like to thank Steve Piantadosi for doing some of the heavy lifting. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

## References

Baillargeon, R. (2004). Infants' physical world. Current Directions in Psychological Science, 13, 89-94.
Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. Cognitive psychology, 43(1), 1-22.
Carey, S. (1999). Knowledge acquisition: Enrichment or conceptual change. Concepts: core readings, 459-487.
Carey, S. (2009). The Origin of Concepts. Oxford University Press.
Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., \& Regier, T. (2016). The sapir-whorf hypothesis and probabilistic inference: Evidence from the domain of color. PloS one, 11(7), e0158725.

Crump, M. J. C., McDonnell, J. V., \& Gureckis, T. M. (2013, jan). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS one, 8(3), e57410.
Einstein, B. (2015). How its made series: Beats by dre. Retrieved 2015-06-15, from https://blog.bolt.io/how -it-s-made-series-beats-by-dre-154aae384b36
Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., \& Tenenbaum, J. B. (2016). Inferring mass in complex physical scenes via probabilistic simulation. Cognition, 1,2 .
Henrich, J., Heine, S. J., \& Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3), 61-83.
Huanca, T. (2006). Tsimane'oral tradition, landscape, and identity in tropical forest. Tomás Huanca L.
Jara-Ettinger, J., Gibson, E., Kidd, C., \& Piantadosi, S. (2016). Native amazonian children forego egalitarianism in merit-based tasks when they learn to count. Developmental Science, 19(6), 1104-1110.
Lillard, A. S. (1993). Pretend play skills and the child's theory of mind. Child development, 64(2), 348-371.
Núñez, R. E., \& Sweetser, E. (2006). With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. Cognitive science, 30(3), 401-450.
Pennington, J., Socher, R., \& Manning, C. (2014). Glove: Global vectors for word representation. $E M N L P, 12,1532-$ 1543.

Piantadosi, S. T., Jara-Ettinger, J., \& Gibson, E. (2014). Children's learning of number words in an indigenous farmingforaging group. Developmental Science, 17(4), 553-563.
Pinker, S. (2007). The stuff of thought: Language as a window into human nature. Penguin.
Reyes-García, V. (2001). Indigenous people, ethnobotanical knowledge, and market economy. a case study of the tsimaneamerindians in lowland bolivia.
Talmy, L. (1988). Force dynamics in language and cognition. Cognitive science, 12(1), 49-100.
Upshaw, M. B., \& Sommerville, J. A. (2015). Twelve-monthold infants anticipatorily plan their actions according to expected object weight in a novel motor context. Frontiers in public health, 3.
Vedantam, R., Lin, X., Batra, T., Lawrence Zitnick, C., \& Parikh, D. (2015). Learning common sense through visual abstraction. In Proceedings of the ieee international conference on computer vision (pp. 2542-2550).
Wolf, L., Hanani, Y., Bar, K., \& Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. International Journal of Computational Linguistics and Applications, 5(1), 27-44.
Yu, L., Park, E., Berg, A. C., \& Berg, T. L. (2015). Visual madlibs: Fill in the blank description generation and question answering. In Proceedings of the ieee international conference on computer vision (pp. 2461-2469).

# Learning in the Wild: Real-World Experiences Shape Children's Knowledge Organization 

Layla Unger (LUnger@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue<br>Pittsburgh, PA 15214 USA

Anna V. Fisher (Fisher49@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue<br>Pittsburgh, PA 15214 USA


#### Abstract

The organization of knowledge according to relations between concepts is critically involved in many cognitive processes, including memory and reasoning. However, the role of learning in shaping knowledge organization has received little direct investigation. Therefore, the present study investigated whether informal learning experiences can drive rapid, substantial changes in knowledge organization in children by measuring the effects of a week-long Zoo summer camp versus a control camp on the degree to which 4- to 9 -year-old children's knowledge about animals was organized according to taxonomic relations. Although taxonomic organization did not differ at pre-test, only Zoo camp children showed increases in taxonomic organization at post-test. These findings provide novel evidence that informal, real-life learning experiences can drive rapid knowledge organization change.


Keywords: Cognitive Development; Semantic Knowledge; Semantic Development

## Introduction

Knowledge is not merely a mentally stored body of information, but rather an interconnected network of concepts linked by meaningful relations (e.g., McClelland \& Rogers, 2003). This organization of knowledge according to meaningful relations between concepts plays a critical role in many cognitive processes, including memory, reasoning, learning, and visual attention (Bjorklund \& Jacobs, 1985; Bower, Clark, Lesgold, \& Winzenz, 1969; Chi, Feltovich, \& Glaser, 1981; Moores, Laiti, \& Chelazzi, 2003; Pinkham, Kaefer, \& Neuman, 2014). Therefore, a key facet of understanding cognition is understanding the development of knowledge organization.

Current conceptual development accounts suggest that learning experiences drive changes in knowledge organization, though the posited nature of such learning processes and the degree to which learning is emphasized in contrast with early (possibly innate) conceptual biases varies across accounts (e.g., Carey, 1985; Fisher, 2015; Gelman, 2003; Tenenbaum, Kemp, Griffiths, \& Goodman, 2011). However, the role of learning in organizing knowledge can only be indirectly inferred from prior research, because no past studies have directly investigated learning experiences that may drive knowledge organization changes. This lack of direct evidence leaves open several key questions about learning-driven knowledge organization change. First, it is
unknown whether such changes result from the protracted accumulation of learning experiences over long-term developmental time-scales (e.g., months and years), or whether they can result from experiences over relatively brief developmental time-scales (e.g., days or weeks). Second, the role of formal education versus day-to-day informal learning experiences in driving knowledge organization change remains poorly understood.

Indirect evidence for the contribution of learning experiences to knowledge organization comes from cognitive development, expertise, and learning science research. Within the cognitive development field, the emergence of knowledge organization has been a focus since the field's inception (Inhelder \& Piaget, 1964). Specifically, numerous studies have attempted to characterize the developmental trajectory of knowledge organization by measuring the degree to which children of different ages possess knowledge of different types of relations between concepts. The relations that have received greatest interest to date include: Similarity based on shared perceptual features (e.g., shape or color); taxonomic relations based on membership in the same, stable category (e.g., mammal); and thematic relations based on cooccurrence in the environment (e.g., dog and bone) (Blaye, Bernard-Peyron, Paour, \& Bonthoux, 2006; Nguyen, 2007; Unger, Fisher, Nugent, Ventura, \& MacLellan, 2016).

Evidence from some cognitive development studies has suggested that children apprehend and reason on the basis of multiple types of relations from an early age (e.g., Gelman \& Coley, 1990; Nguyen, 2007). For example, Nguyen and colleagues used match-to-sample tasks to demonstrate that young children can match target items to both taxonomically and thematically related items versus unrelated items from age two (Nguyen, 2007; Nguyen \& Murphy, 2003). These findings suggest apprehension of multiple relations from early childhood, and thus de-emphasize learning-driven changes in knowledge organization throughout development. However, this conclusion remains controversial. For instance, studies using the same paradigm demonstrated that young children could explain perceptual and thematic, but not taxonomic matches (Sell, 1992; Tversky, 1985). Similarly, Blaye et al. (2006) demonstrated that five-year-old children relied on thematic and perceptual relations to complete an ostensibly taxonomic organization task, whereas older children increasingly used true taxonomic knowledge. These
findings suggest that (1) Young children can make decisions that are consistent with knowledge of a given relation without actually possessing such knowledge, and (2) Relations knowledge develops significantly beyond early childhood.

A handful of recent studies provide direct evidence that knowledge organization indeed evolves gradually across development (Fisher, Godwin, \& Matlen, 2015; Fisher, Godwin, Matlen, \& Unger, 2014; Unger et al., 2016). These studies were designed to capture gradual knowledge organization changes by using a paradigm in which children make graded spatial judgments of the degree to which items are related (Goldstone, 1994). The results of these studies reveal that: (1) Overall, children increasingly differentiate related from unrelated items between ages four and seven, and (2) The influence of specific relations (i.e., thematic and taxonomic) on knowledge organization increases gradually across childhood. In contrast with the perspective suggesting early apprehension of different types of relations, it has been argued that this continuous evolution of knowledge organization implies a key role for learning experiences that transpire throughout childhood. However, these studies provide no direct insight into learning-driven knowledge organization change itself, including whether learning experiences must accumulate over months and years, and whether they must take place in formal education settings.

Further indirect evidence for the effects of learning on knowledge organization comes from expertise studies, which show that expertise in a specialized domain is associated not merely with knowledge of more concepts than novices possess, but the organization of concepts according to relations that are meaningful in the domain (e.g., Chi et al., 1981; Gobbo \& Chi, 1986; Medin, Lynch, Coley, \& Atran, 1997). However, this research does not illuminate whether knowledge organization changes can transpire without extensive learning experiences. Moreover, some researchers have argued that the effects of learning that produce expertize in the formalized knowledge domains studied to date are qualitatively different from the effects of learning in everyday domains (Gelman, 2003). Therefore, the role of everyday learning experiences on driving the acquisition of relations knowledge remains open for debate.

Finally, several learning science studies have measured the effects of brief learning experiences such as educational field trips on knowledge in everyday domains such as animals (e.g., DeWitt \& Storksdieck, 2008; Farmer, Knapp, \& Benton, 2007; Prokop, Tuncer, \& Kvasničák, 2007). However, these studies almost exclusively measured children's knowledge of individual concepts (e.g., "Kangaroo rats have giant feet", Gottfried, 1980, p. 172). Accordingly, this research does not illuminate whether such informal and relatively brief learning experiences can organize children's knowledge according to meaningful relations.

## Present Study

The present study aimed to bridge prior cognitive development, expertise, and learning sciences research by measuring the influence of real-world learning experiences
on knowledge organization. To capture the effects of learning that appear in prior research to transpire over months or years (Fisher et al., 2015; Fisher et al., 2014; Unger et al., 2016), we measured the effects of a concentrated, immersive experience in a domain. Specifically, we measured the effects of a week-long, zoo-based summer camp versus a control, school-affiliated summer camp on 4- to 9 -year old children's knowledge of biological taxonomic relations for a set of animals. We focused on the animal domain because it is familiar to children from an early age, and appears to undergo significant organizational changes with development (Unger et al., 2016). Therefore, this venue provides an ideal opportunity to investigate real-world experience-driven changes in knowledge organization.

## Methods

## Study Sites

Zoo Camp. Zoo camp consisted of lessons, interactions with animals, zoo tours, games, and crafts. Activities for each day were designed around a specific theme, such as "creatures of the night". Each year, the zoo camp organizers choose different themes for each age group. The themes for the age groups spanning our sample (4-5, 6-7, and 8-9 years of age) over the two summers during which testing took place are listed in Table 1 . The majority of themes were not designed to teach biological taxonomic relations, with the exception of themes for children in the 8-9 age group in Year 2 and one instance of a "Reptiles" theme for children in the 4-5 age group in Year 2. These exceptions did not influence our results (see Results). To illustrate how themes shaped camp activities, the activities chosen for the "Extreme Families" theme were as follows. Children took part in two lessons: One about "Extreme Parents" that do or do not protect their offspring (e.g., chickens versus sharks), and another about benefits to animals that live in family groups (e.g., elephant). Children visited in person, completed crafts and played games related to a subset of animals described in lessons.

Control Camp. The Control Camp was a school-affiliated summer camp that did not provide immersive experiences

Table 1: Curriculum themes for each age group.

|  | Year 1 | Year 2 |
| :--- | :--- | :--- |
| 4-5 | Domestic Animals | Animal Locomotion |
|  | Super Senses | Reptiles |
|  | Tropical Treasures | Aquatic Animal Diets |
|  | Savanna Survival | Savanna Animal Patterns |
| 6-7 | Animal Babies | Rainforest |
|  | How Animals Learn | African Savanna |
|  | Animal Families | Ocean |
|  | Aquarium Animals | Islands |
| 8-9 | Extreme Families | Mammals |
|  | Extreme Senses | Birds |
|  | Extreme Architects | Reptiles |
|  | Animal All-Stars | Amphibians |

with animals. At camp, children engaged in outdoor play, dance, crafts, games, and cooking. Additionally, children went on a field trip each week (e.g., to a baseball game), but did not visit the Zoo during this study.

## Participants

Participants were 4 to 9 -year-old children enrolled in the Zoo or the Control Camp located in the same Northeastern US city. The initial sample included 33 Zoo Camp ( 19 females) and 32 Control Camp ( 17 females) children. Of this sample, data from six Zoo Camp children and one Control Camp child were not included in analyses of performance on one of the two outcome measures due to a camera malfunction (see the Scoring section below). Although random assignment to a camp was not possible, children enrolled in the two camps performed equivalently on measures of taxonomic knowledge at pre-test (see Results), and were approximately matched for age (Zoo Camp: $M_{\text {age }}=6.89$ years, $S D=1.43$; Control Camp: $M_{\text {age }}=6.23$ years, $\left.S D=1.21 ; t(57)=1.9, p=.06\right)$.

## Design

The study was a quasi-experiment in which children recruited from Zoo and Control camps participated in both Pre- and Post-Test sessions at the beginning and end of a week of camp. To ensure sufficient number of participants in the Zoo camp condition, data in this condition were collected in the summer of 2015 (Year 1) and 2016 (Year 2), and collapsed across years for analysis.

## Stimuli

The animal stimuli were selected from the zoo camp curricula, such that knowledge in a given age group was assessed for animals about which Zoo Camp children in that age group learned. Accordingly, we developed a separate stimulus set for each age group in each year. Each set consisted of 15 animals, with an equal number of items in each of three biological taxonomic categories: Mammals, birds, and reptiles. To represent the animal stimuli, we used line drawings chosen to minimize perceptual similarity between animals in the same taxonomic category (see examples in Figure 1A-B).

## Materials and Procedures

Participants completed pre- and post-test sessions on Monday and Friday morning of the same week that took place during a "before care" period prior to the start of camp activities. In both Years 1 and 2, these sessions included a Spatial Arrangement Method (SpAM) task (Goldstone, 1994), and in Year 2, children additionally completed a Match-to-Sample task (Figure 1). These tasks have complementary advantages: The match-to-sample task provides a straightforward assessment of taxonomic reasoning that is well-established in developmental research (e.g., Fisher, 2011; Smiley \& Brown, 1979; Waxman \& Namy, 1997), whereas SpAM yields a more graded measure of taxonomic relations knowledge and the degree to which it changes with training. A recent


Figure 1: Schematic depiction of the SpAM task (A) and the match-to-sample task (B).
longitudinal study (Fisher et al., 2014) provided evidence that the two measures converge on the same underlying construct.

SpAM Task. Participants were seated at a game board consisting of a $10 \times 10$ grid, and were told that they were going to play a game in which their job was to help a fictional character, Zibbo, organize his favorite animals on the board (Figure 1-A). The experimenter then showed participants a stimulus sheet that depicted all 15 animal stimuli selected for the child's age group, named each animal, and removed the sheet from view. Next, the experimenter told the participant that they would organize the animals using cards that each depicted an animal on the game board, such that "animals that are the same kind of animal go close together, and different kinds of animals go farther apart". The experimenter placed one of the 15 animal cards on a central game board square, then named and presented each of the remaining cards for the participant one-by-one. Cards were presented in a predetermined, pseudo-random order in which no more than two animals from the same taxonomic group appeared consecutively. Participants were allowed to move cards that they had placed earlier in the task. Finally, the experimenter photographed the board to record the locations of the cards.

Match-to-Sample Task. The animal stimuli for a given age group were arranged into six triads that each consisted of a Target, a Taxonomic Match that belonged to the same taxonomic category as the Target and a Lure that belonged to a different category (Figure 1-B). Of the six triads, two had mammal Targets, two had bird Targets, and two had reptile Targets. Triads were designed to eliminate non-taxonomic cues to Taxonomic Matches, such as visual similarity or shared habitat. For example, a triad might consist of a type of flightless bird such as penguin as the Target, a bird capable of flight such as owl as the Taxonomic Match, and a non-bird such as polar bear as the Lure.

For each triad, the experimenter asked participants to choose whether the Taxonomic Match or the Lure was "same kind of animal" as the Target. The experimenter pointed to and labeled the animals while providing these instructions (e.g., "Which one is the same kind of animal as the penguin, the owl or the polar bear?).

## Results

Of the Zoo Camp sample, children in both Year 1 and Year 2 ( $N=27$ ) completed the SpAM task, whereas only children in Year 2 ( $N=16$ ) completed the Match-to-Sample task. All 32 children in the Control Camp sample were tested during Year

2, and completed both tasks (although one participant's SpAM data were excluded due to camera malfunction).

## Scoring

SpAM Task. The photographs taken following each arrangement trial were scored by treating the $10 \times 10$ board as a coordinate plane, identifying the coordinates of each card, and calculating the Euclidean distance between the coordinates. The range of possible distances was 1 (adjacent cards) to 12.73 (cards on diagonally opposite corners of the board). These distances are taken as a measure of the degree to which participants judge a given pair of animals to be the "same kind of animal", where shorter distances indicate stronger judgments that the pair are of the same kind.

We used these distance data to calculate a Difference Score for each participant at both Pre- and Post-test that captured the degree to which participants placed taxonomically related animals closer together than unrelated animals by subtracting distances between pairs of animals from the same taxonomic category from distances between taxonomically unrelated pairs. Accordingly, larger Difference Scores reflected stronger judgments that taxonomically related versus unrelated animals were of the "same kind".

Match-to-Sample Task. We calculated an Accuracy score for each participant in which we calculated the proportion of times they chose the Taxonomic Match.

## Pre-Test Performance

We first assessed whether participants in both camps performed comparably on the two measures of taxonomic relations knowledge at Pre-Test. Note that the range of Difference Scores on the SpAM task at Pre-Test was -. 65 to 5.25 (chance=0), and of Accuracy scores on the Match-toSample task was .17 to .83 (chance=.5). The results of independent samples t-tests indicated that there was no significant difference at Pre-Test between the performance of participants in the two camps on either measure (SpAM: $M_{\text {zoo }}=.80, M_{\text {control }}=.61, t(56)=.57, p=.57$; Match-to-Sample: $\left.M_{\text {zoo }}=.54, M_{\text {control }}=.58, t(46)=.69, p=.49\right)$. Performance at PreTest on the SpAM task was above chance in both camps (both $t \mathrm{~s}>2.86$, both $p \mathrm{~s}<.01$ ), whereas performance on the Match-toSample task was above chance in the Control Camp only (Zoo: $t(15)=.94, p=.362$, Control: $t(31)=2.61, p=.014$ ).

## Effects of Zoo versus Control Camp

These analyses measured the effects of Zoo Camp versus Control Camp on changes from Pre- to Post-Test in taxonomically organized knowledge: I.e., the degree to which participants' SpAM Difference Scores indicated that they made "same kind" judgments based on taxonomic relations, and the degree to which participants chose the Taxonomic Match on the Match-to-Sample Task.

For each measure, we assessed whether Zoo camp participants manifested greater improvements from Pre- to Post-Test than participants in the Control camp using two analyses. First, we used paired t-tests to compare Pre- versus


Figure 2. Change scores for Zoo and Control Camp participants in SpAM (left) and Match-to-Sample (right) tasks. Error bars represent standard errors of the mean.

Post-Test Difference and Accuracy Scores for participants in each camp separately, and found that whereas Zoo camp participants performed significantly better across both measures at Post- than Pre-test (SpAM: $M_{\text {pre }}=.80, M_{\text {post }}=1.30$, $t(26)=3.01, \quad p=.006$, Cohen's $d=.34$; Match-to-Sample: $M_{\text {pre }}=.54, M_{\text {post }}=.73, t(15)=3.74, p=.002$, Cohen's $\left.d=1.02\right)$, Control camp participants' performance did not improve from pre- to post-test (SpAM: $M_{\text {pre }}=.61, M_{\text {post }}=.59, t(30)=.19$, $p=.85$, Match-to-Sample: $M_{\text {pre }}=.58, M_{\text {post }}=.56, t(31)=.61$, $p=.55$ ).

Second, to compare performance of participants at both camps directly, we calculated a Change Score for each participant in which we subtracted Pre- from Post-Test scores (such that larger Change Scores indicated larger improvements). We then used independent samples $t$-tests to compare SpAM and Match-to-Sample Change Scores between the camps, and observed that across both measures, Change Scores for Zoo Camp participants were larger than those for Control Camp participants (SpAM: $t(56)=2.61$, $p=.011$, Cohen's $d=.68$, Match-to-Sample: $t(46)=3.48$, $p=.001$, Cohen's $d=1.06$; Figure 2).

## Effects of Camp across Age Range

To test whether the effects of attending Zoo versus the Control camp varied with age, we measured the correlation between age and Change Score for each task in each camp. Age was not correlated with Change Score for either task in Control camp participants ( $r_{\text {match-to-sample }}=-.033, r_{\mathrm{SpAM}}=-.002$, $p s>.86$ ), whereas in Zoo camp participants, age was significantly correlated with Match-to-Sample task Change Score ( $r=.56, p=.024$ ) and marginally correlated with SpAM task Change Score ( $r=.33, p=.09$ ) (Figure 3). Moreover, these correlations were not merely due to older children having more taxonomic relations knowledge to start out with: In Zoo camp children, pre-test performance on the SpAM task was not correlated with Change Score ( $r=.11, p=.597$ ) and pre-test performance on the Match-to-Sample task was marginally negatively correlated with Change Score ( $r=-.49, p=.052$ ).

## Influence of Taxonomic Themes

Finally, to test whether these results were driven by the handful of Zoo Camp themes designed to teach taxonomic relations, we re-ran all analyses excluding all data from


Figure 3: Correlations with age of Change Scores in SpAM and Match-to-Sample tasks for each camp shown with best-fit lines.
children in the 8-9 age group in Year 2 who experienced a week of taxonomically-oriented themes ( $N=2$ ), and trials involving reptiles from children in the $4-5$ age group in Year 2 who experienced a reptile-oriented theme ( $N=4$ ). All results reported above remained unchanged: Significant outcomes remained for Pre- to post-test Zoo Camp comparisons, Zoo vs. Control Camp Change Score comparisons, and correlation between age and Match-to-Sample task Change Score in Zoo Camp (all $p \mathrm{~s}<.05$ ).

## Discussion

This study aimed to capture learning-driven changes in knowledge organization in action by measuring the effects of concentrated, real-world learning experiences on the organization of children's knowledge about animals. Specifically, we measured the effects of a week-long Zoobased summer camp on children's knowledge of biological taxonomic relations between animals. We observed that across two converging measures, taxonomic relations increasingly influenced knowledge organization in Zoo but not Control Camp children. These effects transpired despite equivalent Pre-Test performance, suggesting that the difference between camps at Post-Test cannot be attributed to greater prior taxonomic relations knowledge in Zoo Camp children. Moreover, the difference between camps remained
even when data from Zoo Camp children who received explicit taxonomic instruction were removed from analyses. Finally, the results provided evidence that the degree to which Zoo Camp experiences improved taxonomic relations knowledge was associated with age, suggesting that older children learn relations more effectively than younger children (a possibility we discuss further below). Taken together, these findings provide the first direct evidence that learning experiences need not accumulate over lengthy periods of time or take place in formal education settings to shape knowledge organization. Instead, an immersive but relatively brief learning experience in an informal setting can promote significant knowledge organization changes.

## Open Questions

The evidence for learning-driven knowledge organization change presented here highlights the importance of examining the mechanisms by which experience shapes knowledge organization. For example, although the present study was not designed to arbitrate between accounts of conceptual development that place different emphases on early conceptual biases versus domain-general processes and learning mechanisms, our findings are inconsistent with the perspectives emphasizing early conceptual biases towards perceiving entities as organized into taxonomic categories (e.g., Gelman, 2003; Keil, 2007; Wellman \& Gelman, 1992). By the same token, these findings support a key role for learning throughout development in shaping the organization of semantic knowledge (e.g., Fisher et al., 2015; McClelland \& Rogers, 2003; Tenenbaum et al., 2011). Research following on from the present study could further arbitrate between these accounts, particularly with respect to illuminating the nature of learning mechanisms posited to shape knowledge organization given environmental input.

Finally, our results provide some evidence that the magnitude of learning-driven knowledge organization changes increased with age. One characteristic of the learner that may improve with age is prior knowledge organization (Unger et al., 2016). Although taxonomic knowledge at pretest for the specific animals tested in this study was not correlated with learning-driven improvements, it is possible that older children's knowledge of animals in general was better organized than younger children's knowledge. Consequently, it may have been easier for older versus younger children to integrate new information into existing knowledge structures. Future research that investigates the relationship between such learner characteristics and learning-driven knowledge organization changes could illuminate how learning from experience improves with age.

## Conclusions

This study demonstrated that immersive learning experiences at a zoo summer camp produced changes in organization of children's knowledge about animals. These findings build upon research in several domains, including cognitive development, expertise, and the learning sciences by providing the first direct evidence for learning-driven
changes in knowledge organization. Future research should further investigate the mechanisms by which learning drives the development of knowledge organization.

## Acknowledgements

This work was supported by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education, Institute of Education Sciences (R305B040063) and by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition Scholar Award (220020401) to the second author. We thank Vladimir Sloutsky for input on an earlier version of this paper. We also thank Anna Vande Velde, Kristen Boyle, Camille Warner, Ashli-Ann Douglas, Lindsay Gorby, participating children, parents, and camps for their vital contributions to this research.

## References

Bjorklund, D. F., \& Jacobs, J. W. (1985). Associative and categorical processes in children's memory: The role of automaticity in the development of organization in free recall. Journal of experimental child psychology, 39, 599617.

Blaye, A., Bernard-Peyron, V., Paour, J.-L., \& Bonthoux, F. (2006). Categorical flexibility in children: Distinguishing response flexibility from conceptual flexibility. European Journal of Developmental Psychology, 3, 163-188.
Bower, G. H., Clark, M. C., Lesgold, A. M., \& Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. Journal of Verbal Learning and Verbal Behavior, 8, 323-343.
Carey, S. (1985). Conceptual change in childhood. Cambridge, Massachusetts: MIT Press.
Chi, M. T., Feltovich, P. J., \& Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive science, 5, 121-152.
DeWitt, J., \& Storksdieck, M. (2008). A short review of school field trips: Key findings from the past and implications for the future. Visitor Studies, 11, 181-197.
Farmer, J., Knapp, D., \& Benton, G. M. (2007). An elementary school environmental education field trip: Long-term effects on ecological and environmental knowledge and attitude development. The journal of environmental education, 38, 33-42.
Fisher, A. V. (2015). Development of inductive generalization. Child Development Perspectives, 9, 172177.

Fisher, A. V., Godwin, K. E., \& Matlen, B. J. (2015). Development of inductive generalization with familiar categories. Psychonomic Bulletin \& Review1-25.
Fisher, A. V., Godwin, K. E., Matlen, B. J., \& Unger, L. (2014). Development of Category-Based Induction and Semantic Knowledge. Child development.
Gelman, S. A. (2003). The essential child: Origins of essentialism in everyday thought: Oxford University Press, USA.

Gelman, S. A., \& Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. Developmental Psychology, 26, 796.
Gobbo, C., \& Chi, M. (1986). How knowledge is structured and used by expert and novice children. Cognitive Development, 1, 221-237.
Goldstone, R. (1994). An efficient method for obtaining similarity data. Behavior Research Methods, Instruments, \& Computers, 26, 381-386.
Gottfried, J. (1980). Do children learn on school field trips? Curator: The Museum Journal, 23, 165-174.
Inhelder, B., \& Piaget, J. (1964). The early growth of logic in the child. New York: Norton.
Keil, F. C. (2007). Biology and beyond: Domain specificity in a broader developmental context. Human Development, 50, 31-38.
McClelland, J. L., \& Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. Nature Reviews Neuroscience, 4, 310-322.
Medin, D. L., Lynch, E. B., Coley, J. D., \& Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? Cognitive psychology, 32, 49-96.
Moores, E., Laiti, L., \& Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. Nature neuroscience, 6, 182-189.
Nguyen, S. P. (2007). Cross-classification and category representation in children's concepts. Developmental Psychology, 43, 719-731.
Nguyen, S. P., \& Murphy, G. L. (2003). An Apple is More Than Just a Fruit: Cross-Classification in Children's Concepts. Child development, 74, 1783-1806.
Pinkham, A. M., Kaefer, T., \& Neuman, S. B. (2014). Taxonomies Support Preschoolers' Knowledge Acquisition from Storybooks. Child Development Research, 2014.
Prokop, P., Tuncer, G., \& Kvasničák, R. (2007). Short-term effects of field programme on students' knowledge and attitude toward biology: a Slovak experience. Journal of Science Education and Technology, 16, 247-255.
Sell, M. A. (1992). The development of children's knowledge structures: Events, slots, and taxonomies. Journal of Child Language, 19, 659-676.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. science, 331, 1279-1285.
Tversky, B. (1985). Development of taxonomic organization of named and pictured categories. Developmental Psychology, 21, 1111-1119.
Unger, L., Fisher, A. V., Nugent, R., Ventura, S. L., \& MacLellan, C. J. (2016). Developmental Changes in Semantic Knowledge Organization. Journal of experimental child psychology, 146, 202-222.
Wellman, H. M., \& Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. Annual review of psychology, 43, 337-375.

# The Effects of Social Task Setting on Time Perception 

Kerem Alp Usal (e199523@metu.edu.tr)<br>Annette Hohenberger (hohenber@metu.edu.tr)<br>Department of Cognitive Sciences, METU Informatics Institute, Universiteler Mahallesi, Dumlupinar Bulvarı, No:1, 06800, Ankara, Turkey


#### Abstract

This study investigates the effects of the social setting on prospective time estimation, how time is perceived when a task is performed (i) alone, (ii) with a collaborative, or (iii) with a competitive partner. $N=90$ participants were tested ( 30 in each condition). Participants performed a concurrent Simon task for three different durations ( 15,30 and 45 seconds) which was followed by a time reproduction phase. Results revealed a main effect of social condition. Reproduction ratios in dual conditions were smaller than in the single condition and also smaller in the competitive condition compared to the cooperative condition. The results provide first evidence that social condition affects time estimation: time "flies" when we work together, in particular when we compete with a partner, showing that cognitive and social processes are heavily intertwined.


Keywords: time cognition; time perception; joint task; joint action; social Simon effect; social cognition; prospective time estimation

## Introduction

The passage of time has always captured the curiosity of humans. As archaeological studies revealed sundials were in use some 3500 years ago (Vodolazshkaya, 2014). However, measuring the passage of time with clocks is not the same as the "feeling" of how much time has passed. Therefore, it has been suggested that humans have internal and possibly innate mechanisms for keeping track of time and these mechanisms have been studied and explained with internal clock models which facilitate the understanding of how cognitive factors can affect time estimation (Droit-Volet, 2013).

This study brings together two lines of study in cognitive science: time perception and joint action. Time perception is a basic cognitive ability implied in a wide variety of experimental tasks and daily activities (Grondin, 2010). Forming joint attention and performing joint action is another cognitive ability which has recently been the focus of several studies showing that people's performance in any task is heavily affected by joint attention and joint action (Sebanz, Bekkering, \& Knoblich, 2006; Sebanz, Knoblich, \& Prinz, 2003; Vesper et al., 2011). Also, studies in the literature suggest that time perception might be under the influence of a person's mood at that time (Droit-Volet \& Meck, 2007).

The purpose of this study is to provide an experimental research paradigm linking the social aspect of the task setting to participants' prospective time estimation during that task, in order to investigate the effects of the social setting of the task on time perception.

## Time Perception

There are dedicated and intrinsic models for time processing (Ivry \& Schlerf, 2008). The dedicated models are modular, such as the Attentional Gate Model (Block \& Zakay, 2006) or the cerebellar timing hypothesis (Ivry et al., 2002). On the other hand, intrinsic models suggest that time perception is distributed in various neural networks instead of a certain part of the brain (Reutimann et al., 2004).

The Attentional Gate Model contains a pacemaker which emits pulses continuously on a certain rate, and it can only be affected by arousal on a small scale. These pulses flow through an attentional gate, which is regulated by an executive function that determines whether attentional resources should be directed to the task at hand or to the keeping of time, which might be affected by diverting attentional resources to another task. A switch between the attentional gate and the counter starts or stops the connection, and the counter system keeps track of how many pulses have passed since the beginning of the event and stores that information in memory. Later, the number of pulses are retrieved from memory to represent how much time has passed during the given event. Then the decision on the amount of time that has passed is based on the latest information from the counter and the beginning of the counting of the pulses. The additional attentional mechanism for the explanation of mistakes in time estimation seen in humans, especially when there are other attention-demanding tasks in parallel with time estimation (Block \& Zakay, 2006).

The temporal paradigm used in this study is prospective duration judgment, also called "experienced duration" (Block, 2014). In this paradigm, participants are aware that they will perform a time reproduction after some experienced duration (Zakay \& Block, 2004). Participants use their attentional resources to keep track of time while they are performing a secondary task during that interval.

In accordance with the Attentional Gate Model (Block \& Zakay, 2006), the amount of attention devoted to keeping track of time decreases in more demanding secondary tasks, e.g., executive tasks, compared to easier tasks, which results in an underestimation of the actual duration of the interval (Duzcu \& Hohenberger, 2014). As the amount of cognitive load increases, the ratio of reproduced duration to actual duration decreases, which means that participants tend to underestimate time (Block, Hancock, \& Zakay, 2010). This finding is explained by the Attentional Gate Model as follows: the attentional gate is down, because the participant is focusing on the difficult task at hand, and therefore more pulses of the pacemaker are missed.

## The Simon Task

The Simon task is a spatial compatibility task first described in a paper by Simon \& Rudell (1967). The Simon task is a two-choice reaction task and stimulus has relevant (e.g. color) and irrelevant (e.g. location) dimensions. Participants are instructed to respond according to the relevant dimension of the stimulus and not to the irrelevant dimension. The Simon task consists of congruent trials in which the irrelevant dimension is spatially compatible and incongruent trials in which the irrelevant dimension is not spatially compatible.

The first true Simon effect was shown in another study by Simon \& Small (1969). The Simon effect is based on the universal tendency to respond faster when stimulus and response location overlap, i.e. the congruent condition (Hommel, 2011).

## Joint Action

People frequently perform an action together, which is called joint action. The social Simon task is a joint action paradigm in which participants share a Simon task and respond only to half of the stimuli, e.g., blue or red stimuli, occurring on either side of the monitor, respectively. Interestingly, it has been shown that an individual's actions in the social Simon task are represented in the other person's mind and have an impact on their actions. Therefore, the social Simon task results in the same findings as the individual Simon task, i.e., people respond faster to a stimulus on their side ("congruent" condition) as compared to a stimulus on the opposite side ("incongruent" condition) even if the social Simon task does not necessitate a spatial reference as in the individual Simon task. This construction of a mental representation of each other results in an increase in the amount of cognitive load (Sebanz, Knoblich \& Prinz, 2003).

Previous studies have shown that the increase in the amount of cognitive load results in underestimation of time (Block, Hancock \& Zakay, 2010) which is in accordance with the Attentional Gate Model.

The present study brings together time perception and the social Simon task in a single study. It will broaden our understanding of how human time perception is affected by the social setting and the nature of this setting.

## Hypotheses

Our hypothesis is that subjects' time perception during a task is affected by the social setting of the task. In line with the Attentional Gate Model (Block \& Zakay, 2006), we argue that joint settings require more attentional resources than the single setting, since participants co-represent their partners' task, thus leaving less resources for time estimation. Due to social facilitation and attention demands, we expect that subjects will perceive time as proceeding faster during a joint-action task than in a single person task. Furthermore, the nature of the social setting - whether cooperative or competitive - may affect time perception. If subjects experience competitive settings as even more attention-
demanding they may perceive time as proceeding even faster during a competitive joint action task than a cooperative one.

## Method

## Participants

A total of 90 participants ( 42 males, mean age: 25.90, $\mathrm{SD}=5.234$ ) were tested in three different groups. The Single Task group ( $\mathrm{n}=30,14$ males, mean age: $26.03, \mathrm{SD}=6.206$ ) were tested alone whereas the Cooperative Task group ( $\mathrm{n}=30$, 14 males, mean age: $25.03, \mathrm{SD}=5.442$ ) and the Competitive Task group ( $\mathrm{n}=30,14$ males, mean age: $26.63, \mathrm{SD}=3.819$ ) were tested in dyads. Dyads always consisted of participants from the same gender. Participants were recruited through email invitation. They were undergraduate or graduate students from various METU departments. All participants were right-handed and had normal or corrected-to-normal vision. Before the study, ethics approval has been obtained from METU Human Studies Ethical Committee. All participants volunteered to join the study and no monetary reward was offered for participation or performance, since it might affect time perception (Failing \& Theeuwes, 2016).

## The Simon Task

A Simon task was performed for three different duration lengths ( 15,30 and 45 seconds) which was followed by a time reproduction phase. In the single condition, participants were tested alone and they did all Simon tasks and following time reproductions themselves. In the joint conditions (Cooperative and Competitive) participants performed a social Simon task in which each participant was assigned to a specific stimulus color and response button. The participant on the left side was instructed to use only the ' $z$ ' button and respond only to red stimuli whereas the participant on the right side was instructed to use only the '. ' button and respond only to blue stimuli. These buttons were chosen because on a Turkish Q-style keyboard they are the furthest apart horizontally. All participants used their right hand index finger to respond in order to achieve the same setting between dyads, since literature in the field suggests that the position of hands during a social Simon task might affect performance (Liepelt, 2014; Welsh, 2009). Stimuli occurred on the left and right side of the screen, randomly.

Participants in all conditions were told that they would receive points for their correct responses in the Simon task. In the single condition they were told that their points would be compared with other participants individually, in the cooperative condition they were told that their points would be calculated as a team and compared with other teams, and in the competitive condition they were told that each participant's points would be compared with the other participant in the dyad.

## Time Reproduction Task

Before the reproduction phase begun, participants were informed through a message on the screen that they were going to see a big square in the middle of the screen,
indicating the time reproduction phase has begun. Participants used the same button for the time reproduction phase, depending on the color of the big square, i.e. ' $z$ ' for red and '.' for blue.

They were instructed to wait as long as they thought the previous trial has lasted and then press the button to indicate the end of the duration. A message on the screen warned the participants before each time reproduction phase, which stayed on the screen for 2 seconds and the time reproduction has begun automatically afterwards. In the cooperative and competitive conditions, participants were instructed that if the big square was in the color they were responsible for, they were assigned to do the time reproduction. In other words, if it was red, the left participant did time reproduction and if it was blue, the right participant did time reproduction. The order of the color was random and balanced between subjects. Participants in all conditions performed a total of 18 time reproductions (6 of each duration length).

## The Questionnaire

After the test, participants were presented with a short questionnaire. The first 5 questions were presented to participants in all social conditions and they were regarding their mood and self-assessment during the trials. The second part of the questionnaire, which consisted of questions 6 to 9 , were only presented to participants in the joint conditions and were regarding partner-assessment and social warmth. There was also a $10^{\text {th }}$ question which was different amongst the two social conditions. The participants in the cooperative condition were asked to evaluate the quality of their cooperation whereas the participants in the competitive condition were asked the quality of their competition.

## Statistical Analysis

Collected data was analyzed in three different sections: Time Reproduction, the Simon Task and the Questionnaire. For the time reproduction, $3 \times 3$ mixed ANOVAs with social task setting (individual, cooperative, competitive) as a betweensubjects variable and duration (short, medium, long) as a within-subjects variable were conducted on three dependent measures: Duration Ratio (Reproduced Duration/Objective Duration), Absolute Error/Actual Duration and Coefficient of Variation (SD/Mean).

For the Simon task, response times for compatible $v s$ incompatible trials were analyzed as a dependent measure.

For the analysis of the questionnaire, presented options were given values from 1 to 5 , with the most negative option being 1 and the most positive option being 5 . The first 5 questions, which were presented to all participants, were analyzed with a One-way ANOVA for the 3 task settings (Single, Cooperative, Competitive). The second part, which consisted of questions 6-10, were only presented to the participants in dual task settings. A One-way ANOVA for the 2 task settings (Cooperative, Competitive) was carried out for each question.

## Results

## Time Reproduction

The first analysis was performed on Duration Ratios (Reproduced Duration/Objective Duration). The main effect of duration was statistically significant $(F(2,174)=174.64$, $p<.001, \eta_{p}{ }^{2}=.67$ ). Simple contrasts revealed that reproduction ratios were smaller for long durations ( $M=.48, S E=.014$ ) as compared to moderate $(M=.53, S E=.015) \quad(F(1,87)=49.93$, $p<.001, \eta_{p}{ }^{2}=.37$ ) and short durations ( $M=.65, S E=.015$ ) $\left(F(1,87)=225.26, p<.001, \eta_{p}{ }^{2}=.72\right)$, indicating that long durations were underestimated more than moderate and short durations. There was a main effect of task setting ( $F(2,87)=14.59, p<.001, \eta_{p}{ }^{2}=.25$ ). Helmert contrasts revealed a significant difference when the single task setting was compared to both dual task settings $(F(1,88)=18.30, p<.001$, $\eta_{p}{ }^{2}=.17$ ). The reproduction ratios in the dual task settings were smaller $(M=.51, S E=.022)$ than in the single task setting ( $M=.64, \quad S E=.027$ ), indicating that duration was underestimated more by the participants in the dual task settings as compared to the single task setting. Also, the difference between the cooperative task setting compared to the competitive task setting was significant $(F(1,58)=11.42$, $p=.001, \eta_{p}{ }^{2}=.16$ ). Reproduction ratios were smaller, hence durations were more underestimated in the competitive task setting ( $M=.46, S E=.023$ ) compared to the cooperative task setting ( $M=.56, S E=.019$ ) (see Figure 1).


Figure 1: Mean Ratio of Reproduced/Objective Duration across duration lengths for all task settings. (Error bars represent $S E$ and the numbers above the bars show the values of absolute time durations)

The analysis of the absolute errors showed that the main effect of duration was significant $(F(2,174)=157.77, p<.001$, $\left.\eta_{p}{ }^{2}=.64\right)$. Error ratios were higher, indicating that the inaccuracy of participants time estimation was higher in the long duration $(M=.52, S E=.13)$ than the short $(M=.36$, $S E=.12$ ) and the medium duration ( $M=.47, S E=.14$ ). The setting of the task had a significant effect on accuracy $\left(F(2,87)=15.38, p<.001, \eta_{p}{ }^{2}=.26\right)$. The first Helmert contrast revealed that participants in both dual task settings showed higher error ratios $\left(F(1,88)=18.56, p<.001, \eta_{p}{ }^{2}=.17\right)$, hence
were less accurate ( $M=.49, S E=.018$ ) compared to the single task setting $(M=.38, S E=.022)$. Moreover, as the second Helmert contrast revealed $\left(F(1,58)=11.42, p=.001, \eta_{p}{ }^{2}=.16\right)$, error ratios were higher, hence accuracy was lower in the competitive task setting $(M=.54, S E=.016)$ than the cooperative task setting ( $M=.44, S E=.021$ ). The effect of the interaction between duration and task setting was not significant $\left(F(2,87)=1.45, p>.05, \eta_{p}^{2}=.03\right)$ (see Figure 2).


Figure 2: Mean Values of Absolute Error/Objective Duration across duration lengths for all task settings. (Error bars represent $S E$ and the numbers above the bars show the values of absolute errors)

The third analysis was carried out on the Coefficient of Variation, which is calculated by dividing the standard deviation of reproduced durations by the mean reproduced durations. The CV is regarded as a very important variable in Scalar Expectancy Theory because a stable CV is a sign of the scalar invariance of subjective estimation of time across different duration lengths (Church \& Meck, 2003). The effects of duration $\left(F(2,174)=1.58, p>.05, \eta_{p}{ }^{2}=.02\right)$ as well as task setting on the CV were not significant $(F(2,87)=2.7$, $p>.05, \eta_{p}^{2}=.06$ ), indicating scalar invariance, as expected.

## The Simon Task

The analysis of the Simon task revealed that congruency had a significant effect $\left(F(1,87)=101.03, p<.001, \eta_{p}{ }^{2}=.54\right)$. Response times were significantly shorter in the congruent condition ( $M=525.66, S E=2.77$ ) in comparison to the incongruent condition ( $M=533.21, S E=2.79$ ) (see Figure 3). This difference amounts to the "Simon effect". Task setting did not have a significant effect on overall response times ( $\left.F(2,87)=1.53, p>.05, \eta_{p}^{2}=.03\right)$ : participants' reaction speed was similar in single ( $M=535.39, S E=4.29$ ), cooperative ( $M=529.35, S E=5.79$ ) and competitive ( $M=523.57, S E=4.18$ ) task settings. The interaction effect between congruency and task setting was not significant $(F(2,87)=2.24, p>.05$, $\left.\eta_{p}{ }^{2}=.05\right)$. Participants in all task settings were faster in the congruent condition than in the incongruent condition. Overall, these results revealed that the Simon effect was not affected by the various task settings, indicating that the primary time estimation task did not interfere with the secondary, concurrent task.


Figure 3: Mean Values of Response Time for congruent and incongruent trials across task settings. (Error bars represent $S E$ and the numbers above the bars show mean response times)

In order to assess whether the side at which the participant was seated had any effect on the Simon task, a 2 (Congruency: Congruent, Incongruent) x 2 (Participant's Side: Left, Right) Mixed ANOVA was conducted on response times. Participant's side was a between-subject factor and congruency a within-subject factor. This analysis revealed that congruency had a significant effect $\left(F(1,58)=62.47, p<.001, \eta_{p}{ }^{2}=.52\right)$. Participants' response times were significantly lower in the congruent condition ( $M=522.32, S E=3.58$ ) in comparison to the incongruent condition ( $M=530.59, S E=3.60$ ). Participant's side did not have a significant effect on overall response times ( $F(1,58)=0.27, p=.869, \eta_{p}^{2}=.00$ ), i.e., participants' reaction speed was similar on both the left ( $M=527.05, S E=5.17$ ) and the right side ( $M=525.87, S E=4.99$ ). The interaction effect between congruency and side was also not significant $\left(F(1,58)=1.00, p=.321, \eta_{p}^{2}=.02\right)$. Participants on both sides were faster in the congruent condition than in the incongruent condition, which shows that the Simon Effect was observed in participants on both sides (see Figure 4).


Figure 4: Mean Values of Response Time for participant's side across congruency. (Error bars show $S E$ and the numbers above show the values of mean response times)

## The Questionnaire

The analysis of the first five questions revealed that main effects were not significant for enjoyment/boredom during the trial $(F(2,87)=.242, p=.785)$, excitement $(F(2,87)=.079$, $p=.925$ ), pressure $(F(2,87)=.706, p=.496)$, self-assessment for the Simon task $(F(2,87)=1.375, p=.258)$ and selfassessment for the time reproduction task $(F(2,87)=1.457$, $p=.239$ ).

The analysis of the second part of the questionnaire revealed that main effects were not significant for questions 6 to 9: partner-assessment for the Simon task $(F(1,58)=0$, $p=1$ ), partner-assessment for the time reproduction task $(F(1,58)=.887, \quad p=.350)$, friendliness towards partner $(F(1,58)=.267, p=.526)$ and social warmth $(F(1,58)=0, p=1)$.

The results of the $10^{\text {th }}$ question on the quality of their cooperation/competition, revealed a significant main effect $(F(1,58)=10.401, p=.002)$. Participants in the Cooperative task setting assessed their cooperation with a higher value ( $M=3.83, S D=.87$ ) than participants in the Competitive task setting assessed their competition ( $M=2.90, S D=1.32$ ). This means that cooperative dyads reported to feel more as a team, compared to competitive dyads which reported to feel more as rivals.

## Discussion

The results of this study show that there is a strong relation between the social setting of a concurrent executive task and the subjectively perceived duration. Participants estimated the actual duration of the task to be shorter in the joint task settings compared to the single task setting. Also, the nature of the joint action had an impact on the amount of this underestimation, as participants in the competitive task setting underestimated time more in comparison to the participants in the cooperative task setting. These findings are in accordance with previous studies (Dolk et al., 2011; Ford \& Aberdein, 2015; Vesper et al., 2011; Vlainic et al., 2010) showing that joint-action tasks affect cognitive performance. In these studies, the effect concerned their behavior in the Simon task, where a social Simon effect occurred. In our task, however, the social Simon effect is not in the focus of our attention. We were primarily interested in the effect of joint action on the primary task, i.e., the time perception task.

The underestimation of the actual duration can be explained with the Attentional Gate Model (Block \& Zakay, 2006). Previous studies (Sebanz, Bekkering, \& Knoblich, 2006; Sebanz, Knoblich \& Prinz, 2003) have shown that when two or more people are performing a task together, they need to create a mental representation of their partner's part of the task, which requires attentional resources to be shifted towards this demanding task. Additionally, participants in the joint task settings had higher cognitive load due to inhibiting their response when the stimulus on the screen was the color of their partner and it was a no-go trial for them, whereas participants in the individual task setting always had a gotrial since they responded to both colors, and only had to keep track of which button to respond. This means that participants in the joint task settings also had an increase in cognitive load
caused by task switching. Furthermore, participants in the joint task settings had to monitor their partner's responses as well, since their score contributed to the outcome in dual conditions.

Since cognitive load is high and attention is focused on both the executive task and the mental representation of the partner in dual task conditions, the Attentional Gate is low, i.e., little attention is left to keep track of time, which results in a shorter experienced duration. The attention-depleting effect of executive tasks and the underestimation caused by it is well documented in the literature (Block, Hancock \& Zakay, 2010; Duzcu \& Hohenberger, 2014). Here, we have shown that also the social task setting affects this attentional mechanism.

Another possible explanation for the decrease in time estimation observed in the social task settings in comparison to the single task setting might be the effect of the "switch" part of the Attentional Gate Model which determines when attending to the passage of time starts and ends. It might be that when the participant is not acting herself but the partner is acting, these parts are "cut out" of her time experience by the closing of the switch. The switch would only open again when it's the subject's turn again. However, it is not possible to explain the difference between cooperative and competitive task groups with this explanation whereas the difference in cognitive load can explain both results.

Previous studies in the field (Decety et al., 2004; Ruissen \& de Bruijn, 2016) showed that, although both cooperation and competition result in self-other integration, participants in the competitive condition also spend attentional resources on keeping track of the differences between themselves and the other participant in the dyad. Participants in our study had to manage different cognitive loads according to the social condition: Cooperative dyads only needed to follow their cumulative scores, but participants in the competitive condition needed to follow their performance and their partner's performance as separate information, in order to predict who was more successful. This results in a higher cognitive load and thus more severe underestimation of time.

The literature (Droit-Volet \& Gil, 2016; Droit-Volet \& Meck, 2007) suggests that mood has a certain effect on time estimation. However, our questionnaire did not reveal any difference in participants' mood during the experiment, despite the significant contrast in their time estimation. This result suggests that the underestimations were caused by the depletion of attentional resources rather than by the effect of mood on the pacemaker.

Our results have also revealed a significant congruency effect in the Simon task, individual and social, which is in line with the vast literature on the Simon task (Hommel, 2011). The results also indicated that there was no difference in reaction times between participants who were seated on the right side and the left side, which shows that seating did not have any effect on participants' performance.

## Conclusion

The results of this study provide first evidence that social condition affects time estimation: people perceive time to flow faster when they are performing a task with someone in comparison to when they are alone, and even faster when the nature of the social condition is competitive rather than cooperative. This finding can be applied to daily life in education and at the workplace, by supporting joint action over individual work. Our findings also add to the growing literature on "joint action" (Sebanz, Bekkering, \& Knoblich, 2006), showing that there is a strong link between cognitive and social processes. This study has methodological implications in terms of promoting the use of joint settings in cognitive science.

For future studies, experiments that feature another task with similar cognitive load but no social setting would provide information in order to distinguish between the effect of cognitive load and the effect of sociality on time perception. Also, different social manipulations on the same task can provide further explanation whether the difference in time perception is the result of the attentional gate, the switch or the arousal.

## References

Block, R. A. (2014). Cognitive models of psychological time. Psychology Press.
Block, R. A., \& Zakay, D. (2006). Prospective remembering involves time estimation and memory processes. In J. Glicksohn \& M. S. Myslobodsky (Eds.), Timing the future: The case for a time-based prospective memory (pp. 25-49). London: World Scientific.
Block, R. A., Hancock, P. A., \& Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. Acta psychologica, 134(3), 330-343.
Church, R. M., \& Meck, W. H. (2003). A concise introduction to scalar timing theory. Functional and neural mechanisms of interval timing, 3-22.
Decety, J., Jackson, P.L., Sommerville, J.A., Chaminade, T., \& Meltzoff, A.N. (2004). The neural bases of cooperation and competition: an fMRI investigation. Neuroimage, 23(2), 744-751.
Dolk T., Hommel B., Colzato L.S., Schütz-Bosbach S., Prinz W., \& Liepelt R. (2011). How "social" is the social Simon effect? Front. Psychology, 2:84, 1-9.
Droit-Volet, S., \& Gil, S. (2016). The emotional body and time perception. Cognition and Emotion, 30(4), 687-699.
Droit-Volet, S., \& Meck, W.H. (2007). How emotions colour our perception of time. Trends in Cognitive Sciences, 11, 504-513.
Droit-Volet, S. (2013). Time perception, emotions and mood disorders. Journal of Physiology-Paris, 107(4), 255-264.
Duzcu, H. \& Hohenberger, A. (2014). Prospective duration judgments: The role of temporality and executive demands of the concurrent task. Acta Psychologica, 147, 34-41.
Failing, M., \& Theeuwes, J. (2016). Reward alters the perception of time. Cognition, 148, 19-26.

Ford, R. M. \& Aberdein, B. (2015). Exploring social influences on the Simon task: Empathy and friendship. Front. Psychol. 6:962.
Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. Attention, Perception, \& Psychophysics, 72, 561582.

Hommel, B. (2011). The Simon effect as tool and heuristic. Acta Psychologica, 136, 189-202.
Ivry, R. B., \& Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. Trends in cognitive sciences, 12(7), 273-280.
Ivry, R. B., Spencer, R. M., Zelaznik, H. N., \& Diedrichsen, J. (2002). The cerebellum and event timing. Annals of the New York Academy of Sciences, 978(1), 302-317.
Liepelt, R. (2014). Interacting hands: The role of attention for the joint Simon effect. Frontiers in psychology, 5.
Reutimann, J., Yakovlev, V., Fusi, S., \& Senn, W. (2004). Climbing neuronal activity as an event-based cortical representation of time. Journal of Neuroscience, 24(13), 3295-3303.
Ruissen, M. I., \& de Bruijn, E. R. (2016). Competitive game play attenuates self-other integration during joint task performance. Frontiers in Psychology, 7.
Sebanz N., Knoblich G., \& Prinz W. (2003) Representing others' actions: Just like one's own? Cognition, 88, B11-23.
Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: Bodies and minds moving together. Trends in Cognitive Sciences, 10(2), 70-76.
Simon, J. R., \& Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. Journal of Applied Psychology, 51, 300-304.
Simon, J. R., \& Small, A. M., Jr. (1969). Processing auditory information: Interference from an irrelevant cue. Journal of Applied Psychology, 53, 433-435.
Vesper C., van der Wel R., Knoblich G., \& Sebanz N. (2011). Making oneself predictable: reduced temporal variability facilitates joint action coordination. Exp Brain Res, 211:517.
Vodolazhskaya, L. N. (2014). Reconstruction of ancient Egyptian sundials. Archaeoastronomy and Ancient Technologies, 2(2).
Vlainic E., Liepelt R., Colzato L.S., Prinz W., \& Hommel B. (2010). The virtual co-actor: the social Simon effect does not rely on online feedback from the other. Front. Psychology, 1:208.
Welsh, T. N. (2009). When $1+1=1$ : The unification of independent actors revealed through joint Simon effects in crossed and uncrossed effector conditions. Human Movement Science, 28(6), 726-737.
Zakay, D., \& Block, R. A. (2004). Prospective and retrospective duration judgments: An executive-control process. Acta Neurobiologiae Experimentalist, 64, 319-32.

# Instruction type and believability influence on metareasoning in a base rate task 

Pavle Valerjev (valerjev@unizd.hr)<br>Department of Psychology, University of Zadar, Obala kralja Petra Krešimira IV 2<br>Zadar, 23000 Croatia<br>Marin Dujmović (marin.dujmovic@gmail.com)<br>Department of Psychology, University of Zadar, Obala kralja Petra Krešimira IV 2<br>Zadar, 23000 Croatia


#### Abstract

Task dependent conflict has been shown to reduce metacognitive judgements of confidence and prolong response times in various reasoning tasks. For this study a modified version of the base rate task was used to induce conflict while measuring response times and judgements of confidence. The aim of this experiment was to determine the influence of different instruction conditions (reasoning according to belief or according to mathematical probability) on fluency and metacognitive judgements. As expected, participants experienced higher levels of conflict when reasoning according to mathematical probability even though conflict effects were present in both conditions. Additionally, higher believability items mitigated conflict influence while reasoning in accordance with belief and increased it when reasoning in accordance with mathematical probability. These results enrich the growing field of metareasoning research and are discussed as such.


Keywords: metacognition; metareasoning; base rate neglect; conflict monitoring; judgement of confidence

## Introduction

The dual process approach in the field of reasoning is based on the hypothesis of two distinct types of processing. Type 1 processing is heuristic based, fast and comes at a low cognitive resource cost, while Type 2 processing is analytic based, slower, less intuitive, and comes at a high resource cost (Evans, 2007). However, this is just a simplified representation of what is a large set of theories. Since this approach encompasses a large number of theories and models we recommend an excellent review by Evans (2012) for a better understanding of all the complexities in this field. Within this framework a cognitive bias can be defined as a dominant Type 1 response when it is not appropriate and results in a normatively incorrect answer. One of the most commonly studied biases is the belief bias in which a believable response is dominant and more acceptable regardless of correctness (Evans, Barston, \& Pollard, 1983). Tasks which elicit the belief bias have mostly been used in studies of formal types of reasoning such as syllogistic logical reasoning. However, there are other tasks in which participants are led to reason according to their belief. One such task is a modified version of Kahneman and Tversky's (1973) classic base rate task.

The modified base rate task is interesting to researchers because it is suitable for introducing conflict between processes that give rise to different responses. To better understand this task, examine a simple example.

Person X is popular.
Person X is chosen at random from a group consisting of 875 postmen and 125 actors.

The question: Which is more probable?

1. Person X is a postman
2. Person $X$ is an actor

In the example above there are two sources of information on which a person can base his or her response. The mathematical ratio (the base rate) of postmen to actors would indicate that the first answer is more probable. However, common belief based on personal, everyday experience would indicate actors are more likely to be popular than postmen. This strong association between the trait of popularity and the group of actors leads most participants to choose answer number two (e.g. Obrecht \& Chesney, 2016; Pennycook, Fugelsang \& Koehler, 2015). By changing the base rate, congruent versions of this task can be constructed and studied in comparison to conflict ones.

Our research is based on the model proposed by Pennycook and his colleagues (2015). They propose that multiple processes generate initial responses to a particular task or problem (Type 1 processes). The initial responses may be congruent or in conflict. Conflict monitoring then detects (or fails to detect) incongruent results of these processes which leads to Type 2 processing (or simple Type 1 processing if no conflict is detected). Type 2 processing then includes both rationalization of the dominant initial response, and what is usually considered pure Type 2 processing - cognitive decoupling (choosing an alternative response). It is important to note that rationalization and decoupling are not necessarily processed on a conscious level. The dominant response is the one with the highest weight, which would explain the difference in influence and persistence of those responses even when instructed to reason by different criteria. In the example given above the belief-based response would represent the dominant initial response, while the mathematical probability response would be the second initial response. This model provides clear hypotheses on the influence of conflict on response times, levels of confidence, required cognitive load and others.

As indicated, conflict monitoring and detection are key processes which link Type 1 and Type 2 processes
according to the proposed model. For the purposes of this study inducing conflict was of the greatest importance. Conflict monitoring and detection have been identified as important processes in a variety of reasoning tasks (De Neys, 2014; De Neys \& Glumicic, 2008). Conflict in base rate tasks is detected even when participants choose the stereotypical answer because it can be observed in prolonged response times (Pennycook et al., 2015). Since conflict detection in these tasks is present regardless of participant response it is interesting how this process effects other psychological constructs and processes.

Metacognition is broadly concerned with knowledge about, the monitoring and evaluation of other mental processes (Nelson \& Narens, 1990). Recent research in the field has been focused on possible sources of metacognitive judgements in various reasoning tasks (Markovits, Thompson, \& Brisson, 2015; Thompson, Evans, \& Campbell, 2013; Thompson \& Johnson, 2014; Thompson et al., 2013). Building on older research concerning metamemory Ackerman and Thompson (2015) lay out a framework of metareasoning. They outline a number of different metareasoning judgements. For example, the judgement of solvability refers to the probability a particular task is solvable based on type of task, prior knowledge and experience of the participant. Of the basic metareasoning judgments we focus on the final judgements of confidence. These judgements represent retrospective confidence that a final solution to a problem or task is correct. Typically, metareasoning studies make use of a variety of tasks which include the possibility of both heuristic and analytical responses. The interpretation of results in these studies then naturally takes into account the dual-process approach to reasoning. Participants usually express more confidence for heuristic-based responses (Thompson, Evans, \& Campbell, 2013). It has also been established that in these types of tasks participants tend to be overconfident and their judgements are not dependent on accuracy (Thompson et al., 2013). Therefore, metacognitive judgements in these tasks are formed based on other cues such as fluency (the speed and ease of generating responses) and the presence of conflicting answers (Thompson et al., 2013). Conflict in base rate tasks has been shown to decrease final judgements of confidence (Pennycook Trippas, Handley, \& Thompson 2014) and prolong response times (Pennycook et al., 2015).

Manipulating the content of tasks is only one way to influence reasoning processes. One of the more interesting manipulations is varying instruction types, giving more weight to a particular type of reasoning or to specific content. Explicit instruction to reason according to logic increases accuracy (Evans, Handley, Neilens, \& Over, 2010) as well as confidence ratings (Trippas, Thompson, \& Handley, 2016), depending on task difficulty. Within the described model of dual processing, instruction type modifies the weight of initially generated responses therefore influencing the level of experienced conflict.

The focus of our study was to determine how simple effects (congruence, instruction type, believability) interact
and effect response times and confidence levels. We predicted participants would experience a greater level of conflict when base rates and stereotypes do not point toward the same answer. This effect should be stronger when instructed to reason based on mathematical probability compared to reasoning based on belief. The level of experienced conflict is expected to increase the likelihood of Type 2 reasoning and manifest as prolonged response times and lowered judgements of confidence. We predicted the addition of different believability levels of stereotypes would influence judgements and response times differently depending on the type of instruction. When reasoning according to belief high believability should mitigate the influence of conflict while increasing it when reasoning according to mathematical probability.

## Method

## Participants and design

The sample ( $\mathrm{N}=38$ ) was recruited among undergraduate psychology students. The experiment was a 2(instruction type) $\times 2$ (congruence) $\times 2$ (believability) repeated measures design.

There were two different instructions. For half of the items participants were instructed to answer according to their belief. The exact instruction was "For the following items respond as fast as you can according to what you know to be probable from everyday experience". For the other half of the items they were instructed to respond in accordance with mathematical probability. The exact instruction was "For the following items respond as fast as you can according to mathematical probability".

The second independent variable varied the level of congruence. For half of the items there was no conflict between belief and mathematics based probability while for the other half there was.

Finally we varied the level of believability of the content by creating two types of items as displayed in the materials section.

In order to control order effects half of the participants first completed the block based on belief instructions and then the one based on mathematical probability, while the other half had the reversed order. Items within each block were randomized for each participant.

## Materials

Previous research with base rate tasks regularly used extreme base rate ratios to achieve the conflict effect, 995/5 and higher (De Neys \& Glumicic, 2008; De Neys, Vartanian, \& Goel., 2008; Obrecht \& Chesney, 2016; Pennycook et al, 2014). We chose to use less extreme and random ratios in order to control for a possible effect of repetitiveness on participant decisions. We implemented clear rules for ratio selection. The highest allowed ratio was $900 / 100$, while the lowest was set at $850 / 150$. Within this spread ratios were generated randomly by a computer algorithm.

To vary believability we followed two different approaches. For half of the items, attributes associated with the person were characteristic of one group, but not exclusive to it. E.g. Person $A$ is physically attractive. Person A is chosen from a group of supermodels or a group of secretaries. Obviously, secretaries can be attractive, but the attribute is an integral part of the concept supermodel. For the other half, we increased the difference in the believability of the attribute for the two groups. For one group the attribute was still integral, but for the other it was highly uncharacteristic, the opposite of what people would expect. E.g. Person B is courageous. Person B is chosen from a group of firefighters or a group of deserters. Courage is an integral part of what people think of firefighting, while cowardice is strongly associated with deserting.

Conflict is achieved when base rate ratios and belief based probability do not point towards the same choice. By adhering to the before mentioned guidelines, forty main and four practice items were selected from a larger pool of constructed items based on researcher scores. Base rate ratios were assigned randomly to the items which were then randomly assigned to experimental conditions. Examples for the four possible combinations of believability and congruence can be seen in Table 1. When combined with the two different instruction types this forms a total of eight experimental conditions (five items per condition).

Table 1: Examples of item types.

|  | Attribute | Subgroups |
| :--- | :--- | :--- |
| Lower <br> believability/Congruent | Elegant | 854 ice skaters, |
| Lower |  | 146 teachers |
| believability/Conflict | Creative | 866 waiters, |
| Higher |  | 134 painters |
| believability/Congruent | Comical | 880 comedians, |
| Higher |  | 120 morticians |
| believability/Conflict | Honest | 842 smugglers, |

The order in which subgroups appeared was randomized among items to avoid habitual responses from our participants.

## Procedure

The experiment was designed in E-Prime v2.0.10.356 and conducted in the Laboratory for Experimental Psychology. Before the main experiment participants underwent practice to associate themselves with the way in which they were required to react. For the main experiment participants were told an attribute describing a person would be presented for a few seconds after which they would receive information about the groups from which the person was randomly selected. Finally, they were presented with a choice and were required to answer from which group they thought the
person was probably chosen based on one of the instruction criteria (belief in one block, and mathematical probability in the other). Confidence judgements were made on a six point scale, with each point representing a percentage of confidence. Scale value 1 represented $50 \%$ confidence (guessing) with each successive value representing an increase of $10 \%$ with the scale value 6 representing $100 \%$ confidence (complete confidence). An example of the single trial procedure can be seen in Figure 1.


Figure 1: Example of a single trial.

## Results

Prior to analysis response time data was processed to eliminate outliers by removing responses outside of the +/3 standard deviation range. Outliers made up of $2.2 \%$ of all responses. Response times were averaged for items within each experimental condition ( 5 items per condition) for final analysis. Confidence ratings were also averaged to get the final confidence judgements for each condition. Before the main analysis two $2 \times 2 \times 2 \times 2$ mixed analyses of variance were conducted with an additional variable of block order (instruction order) to determine whether the order in which participants completed the experiment influenced response times and confidence judgements. In both analyses the main effect of block order and interactions which include the effect were not significant. The results of these analyses show that the order in which participants completed the task had no influence on response times and confidence judgements.

A 2 (instruction condition) x 2 (congruence) x 2 (believability) repeated measures analysis of variance was conducted on response time data. Results of the analysis can be seen in Table 2. A strong main effect of congruence showed response times were significantly shorter for congruent compared to conflict trials. Higher believability in general led to slower responses, but this was due to a strong effect in conflict situations when participants were instructed to reason according to probability (see the threeway interaction interpretation). The main finding of this study is reflected in the significant two-way interaction between instruction type and congruence which is shown in

Figure 2 (for all figures error bars represent 95\% confidence interval for the mean).

Table 2: ANOVA results for response times.

| Effect | $F(1,37)$ | $\eta_{\mathrm{p}}{ }^{2}$ |
| :--- | :---: | :---: |
| Instruction | .05 | .00 |
| Congruence | $11.68^{* *}$ | .24 |
| Believability | $7.07 *$ | .16 |
| Instruction by Congruence | $7.70^{* *}$ | .17 |
| Instruction by Believability | 2.32 | .06 |
| Congruence by Believability | 2.69 | .07 |
| Three-way interaction | $11.34^{* *}$ | .23 |
| ${ }^{* p<.05 \cdot * * p<.01}$ |  |  |



Figure 2: Response times as a function of conflict and instruction type.

The influence of conflict on response times was significantly lower when participants responded in accordance with belief (mean difference between congruent and conflict responses $M_{\text {diff }}=76 \mathrm{~ms}$ ) than when they responded in accordance with mathematical probability ( $M_{\text {diff }}=230.36 \mathrm{~ms}$ ). The three-way interaction effect reflects the different influence of believability depending on the instruction. Higher believability mitigated the influence of conflict when participants responded according to belief, and increased it when they responded in accordance with mathematical probability.

The same analysis was conducted for confidence judgements for which results can be seen in Table 3. The analysis showed a similar pattern of results. Participants expressed lower levels of confidence for conflict compared to congruent items, which is reflected in a large main effect of congruence. Believability, in general, slightly increased confidence ratings but was present in more complex interaction effects. Again, the main finding of this study is best observed by considering the significant two-way interaction between instruction type and congruence in Figure 3. Conflict lowered confidence judgements for mathematics based reasoning (mean difference between
congruent and conflict responses $M_{\text {diff }}=5.24 \%$ ) more than for reasoning based on belief ( $M_{\text {diff }}=2.38 \%$ ).

Table 3: ANOVA results for confidence judgements.

| Effect | $F(1,37)$ | $\eta_{\mathrm{p}}{ }^{2}$ |
| :--- | :---: | :---: |
| Instruction | .21 | .00 |
| Congruence | $20.49^{* *}$ | .36 |
| Believability | $6.54^{*}$ | .15 |
| Instruction by Congruence | $5.63^{*}$ | .13 |
| Instruction by Believability | .36 | .00 |
| Congruence by Believability | 2.30 | .06 |
| Three-way interaction | $4.77^{*}$ | .11 |
| ${ }^{p}<.05 ; * * p<.01$ |  |  |



Figure 3: Confidence judgements as a function of conflict and instruction type.

In the three-way interaction believability increased confidence ratings and mitigated conflict influence when participants responded in accordance with belief, and lowered confidence ratings for conflict items when they responded in accordance with mathematical probability.

Next, we analyzed accuracy depending on instruction type. For the belief based instruction the stereotypical response could be considered as correct, and for the probability instruction the opposite. Accuracy was high for the two belief conditions and the congruent probability condition (above 86\%), but low for conflict trials in the probability condition $(46.84 \%)$. Since this data was not distributed normally, we tested differences using the Wilcox matched pairs test. Instruction to reason according to mathematical probability significantly lowered stereotypical responses in conflict trials $(Z=4.94, p<.01)$, the same was true for the belief instruction condition ( $Z=2.65, p<.01$ ). To evaluate confidence judgements we calculated differences between confidence levels and accuracy for each participant (for this and similar procedures see Koriat, Lichtenstein, \& Fischoff, 1980). An instruction by congruence ANOVA showed a significant interaction effect $(F(1,34)=34.81$, $p<.01, \eta_{\mathrm{p}}{ }^{2}=.51$ ). Results showed the difference between
confidence and accuracy was largest for conflict trials in the probability instruction condition (Figure 4).


Figure 4: Differences between confidence and accuracy as a function of instruction type and congruence

Additionally, analyses of variance were conducted only for responses that were correct depending on the instruction. A total of 25 participants made up this dataset while the rest did not have correct responses for all of the experimental conditions. For both response times and judgements of confidence, results followed a very similar pattern to the analysis of the full dataset. Once again participants were considerably slower for conflict items $(F(1,23)=12.71$, $p<.01, \eta_{\mathrm{p}}{ }^{2}=.36$ ), and slightly slower for more believable items $\left(F(1,23)=9.63, p<.01, \eta_{\mathrm{p}}^{2}=.29\right)$. The instruction type by congruence interaction $\left(F(1,23)=9.54, p<.01, \eta_{\mathrm{p}}^{2}\right.$ $=.29$ ) was again the most interesting result. Conflict had a greater influence on response times when reasoning according to probability than when reasoning according to belief. The three-way interaction was no longer significant but showed the same pattern of results. For judgements of confidence the analysis showed significantly lower levels of confidence in conflict compared to congruent conditions $\left(F(1,23)=19.60, p<.01, \eta_{p}^{2}=.46\right)$. The instruction by congruence interaction remained significant $(F(1,23)=$ 5.62, $p<.05, \eta_{\mathrm{p}}{ }^{2}=.20$ ) and showed conflict lowered confidence judgements to a larger extent than when reasoning according to probability. Additionally, instruction by congruence $\left(F(1,23)=12.37, p<.01, \eta_{\mathrm{p}}{ }^{2}=.35\right)$ and congruence by believability $\left(F(1,23)=4.92, p<.05, \eta_{\mathrm{p}}^{2}=\right.$ .18) interactions showed participants were less confident for high believability items when reasoning according to probability and that conflict had a larger influence on confidence for higher believability items. The three-way interaction was no longer significant, but showed the same pattern of results as the analysis of the total response data.

Finally, we calculated an item-level correlation between response times and confidence ratings. Results ( $r(38)=-.56$, $p<.01$ ) showed that participants gave higher judgements of confidence for items they responded to faster.

## Discussion and conclusions

According to the proposed dual-process model by Pennycook et al. (2015), initial responses are generated by

Type 1 processes in reasoning tasks. If there is a conflict between the initially generated responses, and it is successfully detected, Type 2 processing resolves the conflict in two possible ways. One outcome is the acceptance of a dominant initial response (rationalization), and the other is choosing an alternative response (cognitive decoupling). Because of the expected dominance of belief based responses, we predicted that induced conflict would have a greater influence when instructed to reason based on mathematical probability compared to reasoning based on everyday belief. According to the prediction, this greater influence would initiate Type 2 processes to a larger extent, which would manifest in prolonged response times and lower confidence judgements in the mathematical instruction condition. Both three-way ANOVAs (Tables 2 and 3 ) prove this prediction to be correct. The expected strong main effect of congruence was significant, which is the usual result in this type of research (Pennycook et al., 2015; Pennycook et al., 2014; Thompson \& Johnson, 2014; Thompson et al., 2013). The main findings show that participants responded slower in conflict trials when reasoning according to mathematical probability. Conflict influence was less prominent when reasoning in accordance with everyday belief. This pattern of results is evident for both response times and confidence judgements (Figures 2 and 3). We hypothesize that stereotypical responses have a greater weight during initial response generation (Type 1 processing), which leads to stronger interference of belief on probability based reasoning than vice versa.

Our additional experimental manipulation of stereotype believability resulted in significant three-way interaction effects (Tables 2 and 3 ). When reasoning according to belief, higher believability mitigates the impact of conflict on response times and confidence levels. On the other hand, higher levels of believability increase the influence of conflict when reasoning based on mathematical probability. This result may represent further proof for the existence of differently weighted initial responses among which belief based responses are very prominent.

Participants expressed a higher level of confidence for items which had shorter response times indicating response fluency is a strong cue in the formation of metacognitive judgements. This finding was obtained in recent studies using different thinking and reasoning tasks (see Thompson et al., 2013).

Furthermore, when instructed to reason according to belief, conflict decreased stereotypical responses, but to a far lesser degree than when instructed to reason according to mathematical probability. It is important to note that even when instructed to reason according to probability participants chose belief based responses in over $50 \%$ of conflict trials. This further strengthens the conclusion that everyday belief dominates reasoning in this specific task. Within the framework proposed by Pennnycook et al. (2015), this would indicate instruction to reason according to probability influenced the relative importance of belief and probability information, but did not fully override the
initial dominance of belief based responses. Participants were overconfident only for conflict trials in the probability instruction condition (Figure 4). This was probably due to the fact that in the other three conditions the dominant belief based answers were correct, while in this one that was not the case.

When we analyzed only correct responses the same pattern emerged. As the model predicts, participants were slower and less confident in conflict trials and the conflict had a larger effect when reasoning according to probability. Based on these results, we can conclude that emphasizing a particular way of reasoning can have an effect on the relationship between conflict, response fluency and metareasoning judgements.

When these results are considered together we can conclude everyday belief has a stronger interference on mathematics based reasoning in this type of task than vice versa. Since a main effect of instruction (reasoning type) was not observed, we can speculate the two processes run in parallel, but that the result of the belief based process has a higher weight.

The results may have practical implications, particularly in educational settings. Many tasks require students to ignore intuitive modes of reasoning in favor of analytical thinking, and it is in those types of tasks where results such as found by this study could be applied to increase efficiency.

To conclude, the results of this study confirm the strong influence of conflict on response times and confidence levels in reasoning tasks. The study expands on previous research by introducing further complexity into established relationships between processes. Explicit instructions in combination with different levels of believability moderate the influence of conflict on fluency and confidence judgements. Results may indicate parallel processing of multiple, differently weighted processes, but more sophisticated research is required to explore the findings further.

## Acknowledgments

This research was supported by Grant 4139 from the Croatian Science Foundation, and the University of Rijeka Research grant 13.04.1.3.11.

## References

Ackerman, R., \& Thompson, V.A. (2015). Meta-reasoning: What can we learn from meta-memory? In A. Feeney \& V.A. Thompson (Eds.), Reasoning as Memory (pp. 164182), New York: Psychology Press.

Evans, J.St.B.T. (2007). On the resolution of conflict in dual process theories of reasoning. Thinking \& Reasoning, 13(4), 321-339.
Evans, J.St.B.T. (2012). Questions and challenges for the new psychology of reasoning. Thinking \& Reasoning, 18(1), 5-31.

Evans, J. S. B. T., Barston, J. L., \& Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. Memory \& Cognition, 11, 295-306.
Evans, J.St.B.T., Handley, S.J., Neilsen, H., \& Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. The Quarterly Journal of Experimental Psychology, 63(5), 892-909.
De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. Thinking \& Reasoning, 20(2), 169-187.
De Neys, W., \& Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. Cognition, 106, 12481299.

De Neys, W., Vartanian, O., \& Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. Psychological Science, 19(5), 483-489.
Kahneman, D., \& Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), 237-251.
Koriat, A., Lichtenstein, S., \& Fischoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6(2), 107-118.
Markovits, H., Thompson, V.A., \& Brisson, J. (2015). Metacognition and abstract reasoning. Memory \& Cognition, 43, 681-693.
Nelson, T.O., \& Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Brower. (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory (Vol. 26, pp. 125-173). San Diego: Academic Press.
Obrecht, N.A., \& Chesney, D.L. (2016). Prompting deliberation increases base-rate use. Judgment and Decision Making, 11(1), 1-6.
Pennycook, G., Fugelsang, J.A. \& Koehler, D.J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive Psychology, 80, 34-72.
Pennycook, G., Trippas, D., Handley, S.J., \& Thompson, V.A. (2014). Base rates: Both neglected and intuitive. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(2), 544-554.
Thompson, V.A., Evans, J.St.B.T., \& Campbell, J.I.D. (2013). Matching bias on the selection task: It's fast and feels good. Thinking \& Reasoning, 19(3), 431-452.
Thompson, V.A., \& Johnson, S.C. (2014). Conflict, metacognition, and analytic thinking. Thinking \& Reasoning, 20(2), 215-244.
Thompson, V.A., Turner, J.A.P., Pennycook, G., Ball, L.J., Brack, H., Ophir, Y., \& Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. Cognition, 128, 237251.

Trippas, D., Thompson, V.A., \& Handley, S.J. (2016). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. Memory \& Cognition. Advance online publication. DOI: 10.3758/s13421-016-0680-1.

# Text, images and diagrams as information providers 

Sara Verbrugge (Sara.Verbrugge@kuleuven.be)<br>Aurélie Stas (aureliestas8@hotmail.com)<br>Koen Jaspaert (Koen.Jaspaert@arts.kuleuven.be)<br>Department of Linguistics, University of Leuven, Blijde Inkomststraat 21, 3000 Leuven, Belgium


#### Abstract

We studied the effect of adjunct displays on recall in an expository text (based on McCrudden, Schraw, Lehman, \& Poliquin, 2007) in order to find out which means of display aided pupils in the last years of secondary school to recall information. We included four conditions in the experiment: text only, text and causal diagram, text and images and causal diagram only. Participants were checked for their recall of main ideas and causal sequences. Recall for main ideas did not vary significantly across conditions. Contrary to McCrudden et al. (2007), our results for the causal sequences revealed that participants who studied a causal diagram only could recall more steps from causal sequences than participants in any of the other conditions. We will interpret the findings in the light of the literature on redundancy effects, dual coding theory, and the causal explication hypothesis.


Keywords: Causal diagram; Text comprehension; Causal relationships; Visual/spatial display

## Introduction

We set up an experiment studying the benefits of adding causal diagrams and illustrations to expository texts for recall of information. We will discuss the use of adjunct displays (i.e. visual representations that complement text with the purpose of making the information present more easily understandable and better recalled) in previous research. After that, we will go into more detail on the specific advantages of causal diagrams for text interpretation.

## Benefits of adjunct displays

Adding so called "adjunct displays" i.e. powerpoint presentations, images, causal diagrams, etc. to a spoken or written text does not always entail the expected benefits for the reader or listener. When these kinds of adjunct displays are added to text, they restate the same information of the text in a new form (such as causal diagrams) or add new chunks of information (such as images). The audience's preexisting knowledge about the topic and the amount of overlap between the two sources of information are two variables that need to be considered. When taking these nuances into account, adding visual displays can promote recall, provided they are implemented in the right way.

Vekiri (2002) presented an overview of the advantages of different graphical displays for recall and suggested that graphics make a valuable contribution to learning only when readers can interpret and integrate the information without
extra processing demands. The effects of adding graphics are mediated by variables such as the participants' preexisting knowledge about the subject of study and visuospatial ability. Also, graphic displays need to be adjusted to task demands. E.g. when an important part of text comprehension is understanding cause-effect relations, the diagram should explicate these relations and their interactions (Mayer \& Gallini, 1990). When graphs are used in a right way, they improve understanding. When this is not the case, they often interfere with learning because they impose extra processing demands reducing working memory capacity for learning activities.

Because of the way diagrams organize text and explicate relations, they make it easier to draw inferences about the relations that are present (Robinson \& Kiewra, 1995). Precisely because links are provided between information bits, relations that might otherwise have remained implicit are now more easily computed (Larkin \& Simon, 1987). See also the causal explication hypothesis (Graesser \& Bertus, 1998; McCrudden et al., 2007, p. 372): adding causal diagrams improves text comprehension because "they provide an explicit visual representation of a text's causal structure that helps the reader understand the text's causal structure". Analyzing causal relationships between text fragments is cognitively demanding. Presenting a causal diagram can help readers to construct a mental model of the text and hence leave more room for inference drawing and consequently a deeper comprehension of the text.

McCrudden et al. (2009, p. 81) present four reasons why the use of adjunct displays boosts understanding of causal relationships. First, a saliency issue: the relevant causal steps are brought to the fore. Their second reason refers to the reduction of cognitive effort. Third, a holistic understanding of the causal structure of the text is more easily achieved. A last reason is concerned with spreading the information via different channels: "it is possible that an adjunct display distributes the information across verbal and spatial working memory stores, functionally increasing its capacity" (McCrudden et al., 2009, p. 81).

Apart from causal diagrams, images are also often added to text. The visual "argument hypothesis" and "picture superiority effect" state that adding images is effective because they are more easily processed than text. In that respect, reference is often made to the "dual coding theory" (Paivio, 1990) that attributes these benefits to the assumption that there are two different codes for visual information (stemming from pictures or words). The dual coding theory postulates that there are two distinct cognitive
systems for information processing and retrieval. "Imagens" stores nonverbal information in mental images, whereas "logogens" stores verbal information in word-like codes (see also Vekiri, 2002). The dual coding theory had been called into question by Johnson-Laird (1998) claiming that information from both verbal and image cues is represented in a single amodal form (see also Vekiri, 2002). These challenges have received recent support from neurological studies. Specifically, recent neural evidence has shown (Shinkareva, Malave, Mason, Mitchell, \& Just, 2011) that patterns of brain activity when thinking about concrete objects are independent of the stimulus presentation format (i.e. words or pictures). Neural states could be identified that are common to pictorial and verbal input referring to objects categories. Not only could they find this effect within participants, but regions of brain activity could also be predicted across participants. "The category of a noun that a person reads or thinks about can be identified solely on the basis of activation patterns obtained from other individuals" (Shinkareva et al. 2011, p. 2422). This evidence at least suggests that the dual coding theory and the picture superiority effect need rethinking.

## Influence of causal diagrams on comprehension

In this section, we address the question whether causal diagrams bring specific advantages for text interpretation. Several studies with undergraduate students have revealed the beneficial nature of adding causal diagrams to texts.

McCrudden et al. (2007) present participants with texts accompanied or unaccompanied by a causal diagram (i.e. a type of visual display that explicitly represents cause-effect relationships, McCrudden 2007, p. 367) summarizing the main ideas and causal sequences of the text. Their results showed that participants who studied both text and causal diagram understood better the five causal sequences in the text. In a second experiment, in which participants studied either the text or the causal diagram, no differences could be found for recall of main ideas or causal sequences between the two conditions. They conclude that causal diagrams are not merely redundant with text.

McCrudden, Schraw, and Lehman (2009) set up two experiments in which participants read a text, either followed by the task to study a causal diagram, to study a list, or to reread the text. After that, participants were asked a range of questions focusing on the steps in the causal sequences and the transitive relationships between causes and effects. Participants in the reread condition performed worse than participants in the other conditions. On the basis of similar results in the two experiments they conducted, McCrudden, Schraw, and Lehman (2009, p. 80) conclude that explicating the steps in a causal chain improves comprehension of the text and learning of causal relationships.

In another study McCrudden, Magliano and Schraw (2011) investigated the online reading processes of participants while they were reading a text or studying a diagram. In their first experiment participants either studied
the causal diagram before reading the text or did not study the causal diagram before reading the text. Participants who studied a diagram before reading the text had faster overall reading times than participants who did not. Moreover, participants in the diagram condition recalled more information than participants in the no-diagram condition. In a second experiment, participants in the no-diagram condition had to read the text twice in order to make the experiment similar for the two groups, and in order to preclude the fact that the effect from the first experiment was due to the repetition of content (2011, p. 78). Again they conclude that "exposure to the diagram had a beneficial effect on comprehension over and above simple repetition" (2011, p. 81). Reading time data showed that reading times of the text of participants who first studied a causal diagram were longer than those of participants who simply reread the text. After that they compared text reading times of participants in the diagram condition with text reading times of participants (of the first reading of the text) in the reread condition. No differences could be found between them. In a third experiment participants were asked to "think aloud" while studying the texts or diagrams. It was shown that presence of a diagram improved the participants' recognition of causal relations.

## Experiment: the contributions of text, images and causal diagrams for recall

We investigate the effects of images and visual displays on recall of information from expository science texts. As in McCrudden et al. (2007) we included conditions with text only, text and diagram, and diagram only. The diagram consists of key terms of the text; relations and sequences are expressed by means of arrows.

Our study diverged from the one by McCrudden et al. (2007) in five important respects. First, we added a condition with text and images. We wanted to find out whether visual displays aided recall of ideas and causal sequences and if they did, which kind of visual display was most beneficial for a thorough understanding and recall. Do causal diagrams entail a bigger advantage because they explicate causal relations? Or do images contribute equally to the comprehension process because they visually represent the content? Second, we did not include a reread condition in our experiment because we think this condition is not straightforwardly comparable to the other conditions. Participants who were asked to reread the text probably invested less effort than participants who received slightly new materials summarizing the main ideas of the text they had just read. In our opinion, rereading a text that is exactly the same as the one you have just read is bound to go a lot faster. Third, we opted for a direct comparison between all four conditions in one experiment. Fourth, participants were given less time to study the causal diagram than in the study by McCrudden et al. (2007). Less information needs to be processed and we wanted to find out whether similar results could be achieved with less studying time. Fifth, we decided
to recruit pupils from secondary schools because the lion's share of research has been carried out with university students and elementary school children, and little attention has been paid to pupils in the last years of secondary school (see also: Mason, Tornatora, \& Pluchino, 2013). Because pupils in this age category are studying many different subjects, are getting acquainted with studying larger chunks of information and need to make important decisions about their future study and career orientations, they are a very interesting group to look into in more detail.

In the experiment participants were allocated to one of four conditions: study a text combined with images, a text combined with a causal diagram, only the causal diagram, or only the text. Participants were tested for their recall of main ideas and for their recall of steps from a causal chain. As in McCrudden et al. (2007) we expect no differences for recall of the main ideas, because their understanding does not rely on participants' ability to link information. We do expect differences for recall of causal sequences because these require participants to build a mental model of the text with causal links, which burdens cognitive resources (Graesser \& Bertus, 1998). If the participants in the text only condition outperform the other conditions, this provides evidence that accompanying materials are redundant. If on the other hand, participants in the condition of text accompanied by causal diagram or causal diagram alone outperform the other conditions, it provides evidence in favour of the causal explication hypothesis: explicating causal relations in a diagram helps processing. If the condition of text combined with images is better than the rest, this supports the visual argument hypothesis.

On the basis of the literature, we hypothesize that participants will perform better in the conditions where text is accompanied by a visual aid, be it a causal diagram or an image.

## Method

Participants 91 pupils from the fifth and sixth grade secondary education (mean age $=17.05, S D=0.85$ ) took part in the experiment. The participants' mother tongue was Dutch.

Design The materials were manipulated between subjects. Participants were randomly attributed to one of four conditions: text and images, text and diagram, text only, or diagram only.

Materials The materials varied in the four conditions. Basis in all cases was an explanatory scientific text about space travel. In the 'text only' condition participants read a text of 1171 words. The text was taken from McCrudden et al. (2007) ${ }^{1}$ and translated into Dutch. A few minor

[^536]adaptations were made in order to make the text more easily readable in Dutch. In the 'text and images' condition the same text was used, but now accompanied by six images. Every image highlighted a particular step from one of the five causal sequences. Three images illustrated a step from the causal sequence (osteoporosis, an excess of bodily fluids in the upper regions of the body, contradictory signals in the brain about the body's orientation). The other three images illustrated a consequence or danger of the causal sequence (kidney stones, muscle loss, heart shrink). In the 'text and diagram' condition, the text was accompanied by a causal diagram. Text and diagram were presented simultaneously. The causal diagram was taken from McCrudden et al. (2007) and translated into Dutch. The causal diagram consists of five cause-consequence relations originating from a common cause: lack of gravity during space travel. In the 'diagram only' condition participants only studied the diagram. In the 'text and diagram' condition and 'diagram only' condition, the materials were preceded by a few lines instructing the participants how to read and interpret the diagram.

Test materials consisted of a questionnaire consisting of two parts. The first part contained three questions. The first asked about participants' prior knowledge about space travel ( $1=$ knew nothing $-6=$ already knew everything). The second question asked participants to write a short text about their willingness to become a space traveler after having read about the dangers of space traveling. This question was inserted in order to create a time buffer between reading the text and answering recall questions. The third question enquired about the "main ideas": participants were asked to name parts of the body that were affected by space travel, and the associated risks (e.g. lack of gravity influenced bone structure, which augmented the risk of kidney stones). The second part of the questionnaire focused on the causal sequences. A particular risk was given (e.g. space travel may cause kidney stones) and participants were asked to list as many causal steps as they could remember. They were invited to explain why the risk existed and to name as many causes and consequences as possible.

Procedure The experiment was conducted in a large study room. Participants in the 'text only', 'text and images', and 'text and diagram' conditions were given 8 minutes to study the materials, participants in the 'diagram only' condition were given 4 minutes to study the materials. (Time needed was determined on the basis of a small pretest.) The 8 minutes group was told after four minutes that four minutes remained. All groups were informed one minute before the end that one minute remained. The materials were collected
site for NASA: When Space Makes You Dizzy (Phillips \& Hullander, 2002) and Mixed up in Space (Phillips \& Hullander, 2001), and one Web-page from an educational Web site for the National Space Biomedical Research Institute: Human Physiology in Space (Lujan \& White, 2002)." (McCrudden et al., 2007, p. 373)
and the first part of the questionnaires was handed out. When participants had finished the first part of the questionnaire, they raised their hands, handed in the first part and received the second part of the questionnaire. No time limits were imposed to fill in the questionnaires. The experiment took the participants approximately 45 minutes.

## Results

Preliminary questions: On the question how well acquainted participants were with the subject of space travel on a scale from 1 (=knew nothing) to 6 (=knew everything already), an average of 2.3 was reached ( 23 participants chose 1 on the scale, 34 chose 2,19 chose 3,12 chose 4,3 chose 5 , none chose 6 ). Correlations were checked between overall recall of main ideas and prior knowledge and between overall recall of causal sequences and prior knowledge but no significant correlations were found.

Main ideas: An ANOVA was conducted with text type (causal diagram only - text only - text and diagram - text and images) as the independent variable and main ideas as the dependent variable. No main effect of text type could be found: $F(3,87)=.739, p=.53$. Recall of main ideas did not differ significantly whether participants read the 'text only', 'text and diagram', 'text and images' or the 'causal diagram only'.

Causal sequences: As for recall of the causal sequences, we conducted an ANOVA with text type as between subjects variable and causal sequence as within subjects variable (repeated measures). To allow for comparisons across causal steps, scores were converted to proportions. I.e. per causal sequence, the number of causal steps remembered was divided by the total number of causal steps for that particular sequence.

There was a main effect of text type $F(3,87)=3.93$, $p=.011$. Causal diagram $(M=0.52)$ led to greater recall than any of the other text types (text only $M=0.34$ - text and diagram $M=0.39$ - text and images $M=0.33$ ). Planned comparisons revealed that 'diagram only' differed significantly from 'text and diagram' $F(1,87)=4.91$, $p=.0015$, which was the second best condition for recall.

There was a main effect of specific topic of the causal sequences: $\quad F(4,348)=18.80, \quad p=.00001$. Planned comparisons revealed that the causal sequence of 'muscle loss' was significantly more recalled than the other four causal sequences ('muscle loss' $M=0.57$ - kidney stones $M=0.43 ; F(1,87)=10.66, p=.002)$. No difference was found between the causal sequence 'kidney stones' and 'motionsickness' ( $M=0.42$ ), but these two causal sequences were better recalled than the causal sequence 'infections' (infections $M=0.31$ - kidney stones; $F(1,87)=4.61, p=.035$ ), which was in turn more recalled than the causal sequence about 'heart shrink' (heart shrink $M=0.25$ - infections $F(1$, 87) $=3.97, p=.049$ ).

The interaction between text type and causal sequence almost reached significance level: $F(12,348)=1.66, p=.07$. The causal sequence that was easiest to remember (muscle loss) did not differ between the different text types. For the more difficult causal sequences, planned comparisons revealed that the participants in the "diagram only" condition always scored best (motion-sickness $F(1,87)=6.96, p=.009$; kidney stones $F(1,87)=9.32, p=.003$ infections $F(1,87)=8.25, \quad p=.005$, and heart shrink $F(1,87)=7.59, p=.007)$.

## Discussion

As in McCrudden et al. (2007) no differences were found between the groups for recall of the main ideas.

The results for the causal sequences show that participants had higher recall rates when they studied a causal diagram only than any of the other types of materials (text only - text and diagram - text and images). No differences were found between the other conditions. Moreover, participants in the 'diagram only' condition only had four minutes to study the diagram, whereas the participants in the other conditions could study the materials for eight minutes. The fact that participants were able to achieve better comprehension in a shorter period of time, is surprising. This result runs counter to the results obtained in McCrudden et al. (2007) where no differences in recall of causal sequences between diagram and text could be found.

A straightforward explanation is the fact that the diagram only contained information that was relevant for the questions participants would receive (cf. the saliency issue discussed in the introduction). The other conditions with text also contained secondary information. A more in-depth explanation for these results might reside in the fact that participants need to invest more processing effort in understanding the diagram and this leads to deeper processing and hence a better retention of the studied materials in the brain. Ainsworth and Loizou (2003) and Moore and Scevak (1997) show that students make more inferences in diagrams than in running text. McCrudden, Magliano, and Schraw (2011) show that studying a diagram led to higher recall of information compared to text reading.

This experiment provides evidence that studying a diagram leads to a better retention in memory. When participants study texts, they may opt for the easy way out: the text reads smoothly, deep processing is not necessary for superficial text comprehension. Additional evidence for this claim can be found in the fact that, similar to the results of McCrudden et al. (2007), the more complex the causal sequence, the better participants could recall the causal sequences when they had studied a diagram compared to the other conditions. Difficult causal sequences may not have been understood profoundly enough in the text condition, whereas the diagram condition made sure the more difficult causal sequences were studied thoroughly. This fits in with the finding by Cromley et al. (2010, p. 69) that "students used a significantly higher proportion of inferences and high-level strategies and a significantly lower proportion of
low-level strategies in diagrams than in text. (...) diagrams seemed to promote more high-level, integrative activity and seemed to discourage low-level superficial strategies."

An explanation why the 'text and diagram' condition did not have higher recall rates than the 'text only' condition, is to be found in the fact that students often skip diagrams or skim only parts of the diagram (see e.g. Schmidt-Weigand et al., 2010). Schmidt-Weigand et al. (2010) also showed that the time participants took to inspect visualizations of the text was considerably lower than the time participants took to read the text. Participants might find the information in the text sufficient and find the diagrams superfluous. When their expectations of relevance were met, they did not take the effort to study the diagrams in detail. This could be overcome in future research by presenting the diagram before the text.

Another surprising finding is that our results fail to support the so called "picture superiority effect" since the condition in which participants studied text and images had the lowest averages of all conditions (but not significantly lower than text only and text and diagram). It has been suggested that when words are processed meaningfully, memory for them may be comparable to that of pictures (Weldon \& Coyote, 1996, p. 671). Processing of information (words, pictures, ...) is optimized according to task requirements (Job \& Tenconi, 2002). Once that purpose is defined, "the processing of information seems to proceed to levels that satisfy the task requirements" (Miller, 2011, p. 719). So in our experiment, adding pictures did not lead to better recall because the information present in the text was sufficient to the participants. Adding pictures did not entail any additional benefits. Apparently participants thought that the information was processed satisfactorily for the current purposes. While Levin \& Mayer, 1993 and Marcus et al., 1996 (as cited in Pike et al. 2010) provided evidence that the use of illustrations reduces the demands on working memory and hence leaves more resources for higher order processing of the text, no evidence for this could be found in our results.

## General Discussion

Here we will situate our results with regard to the causal explication, verbal redundancy, and picture superiority hypothesis. We will also discuss the implications for devising course materials.

## Causal explication hypothesis

McCrudden et al. (2007) provided evidence in favour of the causal explication hypothesis, when text and causal diagram were presented together, because participants in the condition text and causal diagram outperformed the ones in the text only condition. On the basis of our results, we can say that our results tilt toward a redundancy effect when two sources of information are presented together. Even though it is claimed in McCrudden et al. (2007) that causal diagrams are not merely redundant with texts, this seems to be the case in our study. When adding a causal diagram to
text, recall rates drop to levels similar to those of text only. This may be due to the fact that participants only had a look at the diagrams after having already processed the text and hence devoted less attention to it than when they first would have had a look at the causal diagram instead of vice versa.

However, when a causal diagram is presented on its own, recall for causal sequences improves. These results underpin the causal explication hypothesis. When implicit causal relations are made explicit in a causal diagram, recall significantly improves. Whether the effect is due to the fact that only the relevant information for the recall questions was summarized in the causal diagrams or whether it is actually due to their structure or both factors combined, will have to be left unresolved for the time being. However, there can be little doubt as to the role causal diagrams play in alleviating the strains put on our cognitive capacity. The resources made available in this way can then be devoted to storing the information and the links between bits of information more firmly in memory.

## Verbal redundancy

The effect of redundant information has been studied on various levels. Many studies have been conducted investigating "verbal redundancy". When similar information is simultaneously given via different channels (i.e. spoken and written information), comprehension and recall are not necessarily better than when the information is only given in one form. These studies have often been carried out in multimedia environments. The overall conclusion is that when redundant information is given, learning becomes impaired. Rey and Buchwald (2011), Sweller (2005a, 2005b) have shown that offering redundant material often interferes with learning rather than facilitating it (see Sweller, 2005b, p. 159). This redundancy effect is attributed to the fact that working memory capacity is burdened excessively with integrating identical information received via different sources. The results of the experiment are in accordance with the aforementioned studies. When the information is twice presented (causal diagram and text, or text and pictures) learning does not improve compared to the text only condition. So the fact that the information is given in two versions does not hinder learning but neither does it improve learning.

## Picture superiority

As we have argued in the discussion, we fail to find evidence for the picture superiority effect. When pictures are added to text, recall does not improve. The positive effect of a causal diagram cannot be interpreted in terms of picture superiority, because causal diagrams are not pictures, but they are small summaries of causal information in the text.

## Suggestions for study materials

In conclusion, we can say that causal diagrams turn out to be a very convenient study aid, when a deeper understanding of relations is aimed for. It is even the case
that causal diagrams appear to be the principal matter to study for students instead of texts when recall of causal relations is at stake. Participants who studied the causal diagram alone, outperformed the participants in any of the other conditions. So, we can recommend students to draw causal diagrams of the materials they have to study and authors of school and college books to add causal diagrams whenever possible summarizing the main ideas and the relations between them. Making the structure of the text more insightful is a vital characteristic of causal diagrams.

In short, our findings hints towards the need to elucidate course material with diagrams in order to boost memory.

## References

Ainsworth, S. E., \& Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. Cognitive Science, 27(4), 669-681.
Cromley, J. G., Snyder-Hogan, L. E., \& Luciw-Dubas, U. A. (2010). Cognitive activities in complex science text and diagrams. Contemporary Educational Psychology, 35(1), 59-74.
Graesser, A. C., \& Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. Scientific Studies of Reading, 2(3), 247-269.
Job, R., \& Tenconi, E. L. (2002). Naming pictures at no cost: Asymmetries in picture and word conditional naming. Psychonomic Bulletin and Review, 9, 790-794.
Johnson-Laird, P. N. (1998). Computer and the Mind: An Introduction to Cognitive Science. Harvard University Press.
Larkin, J. \& Simon, H. (1987) Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99.
Levin, J. R., and Mayer, R. E. (1993). Understanding illustrations in text. In B.K. Britton, A. Woodward, \& M. Brinkley (Eds.), Learning from Textbooks (pp. 95-113), Erlbaum, Hillsdale, NJ.
Marcus, N., Cooper, M., and Sweller, J. (1996). Understand instructions. Journal of Educational Psychology, 88, 4963.

Mason, L., Tornatora, M., \& Pluchino, P. (2013). Do fourth graders integrate text and picture in processing and learning from an illustrated science text? Evidence from eye-movement patterns. Computers \& Education, 60, 95109.

Mayer, R. E., \& Gallini, J. K. (1990). When is an illustration worth ten thousand words? Journal of Educational Psychology, 82(4), 715-726.
McCrudden, M., Magliano, J., \& Schraw, G. (2011). The effects of diagrams on online reading processes and memory. Discourse Processes, 48(2), 69-92.
McCrudden, M., Schraw, G., Lehman, S., \& Poliquin, A. (2007). The effects of causal diagrams on text learning. Contemporary Educational Psychology, 32, 367-388.

McCrudden, M., Schraw, G., \& Lehman, S. (2009). The use of adjunct displays to facilitate comprehension of causal relationships in expository text. Instructional Science, 37(1), 65-86.
Miller, P. (2011). The processing of pictures and written words: A perceptual and conceptual perspective. Psychology, 2, 713-720.
Moore, P. J., \& Scevak, J. J. (1997). Learning from texts and visual aids: A developmental perspective. Journal of Research in Reading, 20(3), 205-223.
Paivio, A. (1986). Mental representations: A dual-coding approach. New York: Oxford University Press.
Pike M. M., Barnes M. A., \& Barron R. W. (2010). The role of illustrations in children's inferential comprehension. Journal of Experimental Child Psychology, 105, 243-255.
Rey, G. D., \& Buchwald, F. (2011). The expertise reversal effect: Cognitive load and motivational explanations. Journal of Experimental Psychology: Applied, 17, 33-48.
Robinson, D. H., \& Kiewra, K. A. (1995). Visual argument: Graphic organizers are superior to outlines in improving learning from text. Journal of Educational Psychology, 87, 455-467.
Schmidt-Weigand, F., Kohnert, A., \& Glowalla, U. (2010). A closer look at split visual attention in system- and selfpaced instruction in multimedia learning. Learning and Instruction, 20, 100-110.
Shinkareva, S., Malave, V., Mason, R., Mitchell, T., \& Just, M. (2011). Commonality of neural representations of words and pictures. Neuroimage, 54 (3), 2418-2425.
Sweller, J. (2005a). Implications of cognitive load theory for multimedia learning. In R.E. Mayer (Ed.), The Cambridge Handbook of Multimedia Learning (pp. 19-30). New York: Cambridge University Press.
Sweller, J. (2005b). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), Cambridge handbook of multimedia learning (pp. 159-168). New York: Cambridge University Press.
Vekiri, I. (2002). What Is the Value of Graphical Displays in Learning? Educational Psychology Review 14 (3), 261312.

Weldon, M. S., \& Coyote, K. C. (1996). Failure to find the picture superiority effect in implicit conceptual memory tests. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(3), 670-686.

# The Role of Linguistic Information in Learning Abstract Words: Evidence from Children with Specific Language Impairment (SLI) 

Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)<br>Experimental Psychology, University College London<br>Gower Street, London WC1, UK

Marta Ponari (m.ponari@kent.ac.uk)
School of Psychology, University of Kent
Canterbury CT2 9NP, UK

Courtenay F. Norbury (c.norbury@ucl.ac.uk)<br>Language and Communication, University College London<br>Gower Street, London WC1, UK


#### Abstract

Accounts of abstract word learning suggest that learning these words relies primarily on access to linguistic cues, such as the statistical co-occurrence of words with similar semantic properties. Thus, children with language impairment (LI), who by definition have impoverished access to linguistic context, should have disproportionate impairments in abstract word knowledge. Here, we compared verbal definitions and lexical decisions to both abstract and concrete words of children with LI (ages 8 to 13) and both age-matched and vocabulary-matched typically developing (TD) peers. Relative to age-matched peers, children with LI had significant deficits in both tasks. Crucially, however, there was not greater impairment of abstract words. We conclude that that linguistic knowledge is not a sine qua non to learning abstract words and concepts and other mechanisms, which are not specifically impaired in LI, are at play.


Keywords: abstract concepts; semantic representation; distributional semantics; lexical decision; specific language impairment (SLI); vocabulary development.

## Introduction

Children learn thousands of words quickly and efficiently, often without any formal training, and even in impoverished environments. Learning words is hard because even when the referent is present in the physical environment, rarely it is isolated in the visual scene (Medina, Snedeker, Trueswell \& Gleitman, 2011). To make the situation worse, referents are not always present in the physical environment, either because they are spatially and/or temporally displaced (e.g., talk about past or future events), or because they are abstract and have no tangible referent. Most theories of vocabulary acquisition focus on the mechanisms by which words referring to concrete concepts (i.e. objects, actions and other events that can be experienced with our senses and through our own actions) can be learnt; it is less clear how a child learns abstract concepts, which are not perceivable by the senses.

It has been argued that children learn the meaning of concrete words such as "cat" or "run" by observing the statistical contingencies between the words and the objects,
people and actions occurring in the physical environment (e.g., Yu and Smith, 2007). In addition, such contingencies could be enhanced by the use of social communicative cues, such as eye-gaze, or pointing, through which caregiver directs attention to the correct referent (Baldwin, 1991) or by infants actively isolating intended referents from the visual background by picking them up (Morse, Benitez, Belpaeme, Cangelosi \& Smith, 2015).

Word meanings, however, can also be learnt from the linguistic context in which the words occur (Firth, 1957). Recent work has demonstrated how models of semantic memory, based on co-occurrences of words in text (also called Distributional Semantic Models), can predict a variety of semantic effects in adults and children (e.g., Andrews, Vigliocco \& Vinson, 2009; Bruni, Tran \& Baroni, 2014; Landauer \& Dumais, 1997; Griffiths, Steyvers \& Tenenbaum, 2007). This strategy could complement, at least for concrete words, other social-cognitive strategies. For abstract words, distributional information may provide a powerful, if not the most important, mechanism for learning (e.g., Andrews et al., 2009; Johns \& Jones, 2012).

In line with distributional semantic models, abstract words are acquired late in development (Kousta, Vigliocco, Vinson, Andrews \& Del Campo, 2011; Ponari, Norbury \& Vigliocco, in press; Schwanenflugel, 1991). Early studies of children's language production (Brown, 1957, reported in Schwanenflugel, 1991) suggested that $75 \%$ of the words most frequently produced by school-aged children (6-12 years of age) were concrete; in contrast, only $28 \%$ of the words used most commonly by adults were concrete. Schwanenflugel (1991) further reported that, while 6-yearold children have already mastered the majority of concrete words most frequently used by adults, it is not until adolescence that children have mastered the majority of abstract words used by adults. These facts align well with the idea that a sufficient amount of linguistic input is necessary to extract meaning for abstract words.

Thus, if the ability to learn meaning from co-occurrences in the input is critical for learning abstract concepts and words, abstract words should be especially challenging for
children with developmental language impairments (LI). Language impairment is a neurodevelopmental disorder affecting approximately $7.5 \%$ of children at school entry (Norbury et al. 2016, Tomblin et al. 1997). Children with LI have language abilities significantly below expectations for age in the absence of obvious social, sensory or neurodevelopmental explanations. Children with LI typically present with severe deficits in morphosyntax and other aspects of grammar (Rice, 2013), accompanied by vocabulary that is reduced in both breadth and depth relative to typically-developing peers (McGregor, Oleson, Bahnsen \& Duff, 2013). Unfortunately there is a dearth of empirical investigation into the acquisition of abstract words by children with LI.

Here, we investigate implicit and explicit knowledge of abstract and concrete words in children with Language Impairment (LI). Target words were selected at different age of acquisition bands and controlled for variables that are known to affect adult processing, including frequency, number of letters and valence. Lexical decision was used to test implicit knowledge, while verbal definitions were used to test explicit knowledge.

## Methods

## Participants

Eighteen children with an existing diagnosis of Language Impairment (LI; 14 males; mean age $=10.03, \mathrm{SD}=1.76$ ) were recruited from schools in Southeast England. Children in the TD groups were selected from a pool of 73 children who completed both tasks and were matched to the children with LI on age and gender ( $\mathrm{TD}_{\text {age }} ; \mathrm{n}=18,14$ males; mean age $=10.34, \mathrm{SD}=1.44$ ) or by raw scores on the British Picture Vocabulary Scale (BPVS; Dunn, Dunn, Whetton, \& Burley, 1997) ( $\mathrm{TD}_{\text {vocab }} ; \mathrm{n}=18,14$ males; mean age $=8.16$, $\mathrm{SD}=2.12$ ). TD children were recruited from local schools and did not have any reported special educational needs, or history of language delay. Children's non-verbal cognitive abilities were assessed using the Matrix Reasoning test of the Wechsler Abbreviated Scale of Intelligence (WASI, Wechsler, 1999). LI children were also administered the Recall Sentences subtest of the Clinical Evaluation of Language Fundamentals: Core Language Scales (CELF; Semel, Wiig, \& Secord, 2006), see Table 1. The same children participated in both tasks.

## Materials

Thirty-six abstract and 36 concrete words were selected from a pool of 3,505 words for which normative data on a range of lexical variables could be obtained. These variables included: Age of Acquisition (AoA; Kuperman, Stadthagen-Gonzalez \& Brysbaert, 2012), concreteness, familiarity (Coltheart, 1981), valence (Warriner, Kuperman \& Brysbaert, 2013), and frequency (Balota et al., 2007). AoA ratings were used to ensure the items selected were appropriate for our participants' age: words were divided into Age of Acquisition bands (1: words acquired at 4-5
years of age; 2: words acquired at 6-7 years; 3: words acquired at 8-9 years; 4: words acquired at 10-11 years). Within each AoA band, triplets of negative (valence ratings $<4.0$ ), positive (valence ratings $>6.0$ ) and neutral (valence ratings of 4.5-5.5) words matched on length (number of letters), concreteness, log frequency and familiarity were created. Triplets of abstract words were then paired to concrete triplets matching for average length, frequency and familiarity. Among these 72 words, 24 ( 12 abstract and 12 concrete) were shared between the two tasks; 24 ( 12 abstract and 12 concrete) were used for the definitions task only, and the remaining 24 were used for the lexical decision task only. Additionally, for the lexical decision task, forty-eight pronounceable non-words were created by changing one phoneme from 48 words matched to the experimental words on length, AoA, valence and concreteness. All words and non-words were recorded by a native English speaker using Audacity v. 1.2.2.

|  |  | Age-matched |  |  | Vocabulary-matched |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | LI | $\mathrm{TD}_{\text {age }}$ | t-test | p | $\mathrm{TD}_{\text {yocab }}$ | t-test | p |
| Age | $\begin{aligned} & 10.40 \\ & (1.83) \end{aligned}$ | $\begin{gathered} 10.33 \\ (1.44) \end{gathered}$ | 0.127 | . 899 | $\begin{gathered} 8.16 \\ (2.12) \end{gathered}$ | 3.383 | . 002 |
| Matrix reasoning | $\begin{gathered} 40.33 \\ (10.67) \end{gathered}$ | $\begin{aligned} & 49.22 \\ & (9.37) \end{aligned}$ | 2.656 | . 012 | $\begin{aligned} & 51.41 \\ & (8.31) \end{aligned}$ | 3.413 | . 002 |
| BPVS | $\begin{aligned} & 108.72 \\ & (25.03) \end{aligned}$ | $\begin{aligned} & 129.66 \\ & (14.74) \end{aligned}$ | 3.059 | . 004 | $\begin{gathered} 109 \\ (24.25) \end{gathered}$ | 0.034 | . 973 |
| CELF recall sentence | $\begin{gathered} 4.83 \\ (4.23) \end{gathered}$ | NA | - | - | NA | - | - |

## Procedure

All children were tested in a quiet room in their school and received stickers for participation. Stimuli were presented acoustically using E-Prime version 2.0 software (Psychology Software Tools, Pittsburgh, PA) running on a Dell Latitude E6320 laptop with a touchscreen display. Children were presented with short computer games in which they were asked to help a cartoon alien learn English. The Definitions task was always presented before the Lexical Decision task, in a single session.
Definitions. After wearing the headset (which included a microphone), children were presented with a practice trial (the concrete noun: "rose") before the experiment. They were encouraged to provide an accurate and comprehensive definition, including as much information as they could on the meaning of each word. Each trial included the presentation of the alien in the center of the computer screen, along with the auditory presentation of an English word. Children's responses were audio-recorded using EPrime and then scored off-line. The 48 words were presented in four blocks of 12-items arranged in AoA blocks (block 1: words acquired at 4-5; block 2: words acquired at 6-7; block 3: words acquired at 8-9; block 4: words acquired at 10-11); words within each block were presented in random order. The task ended when the subject was unable
to define three words within a single AoA block or responded to all 48 words.

Scoring of definitions We used a 0 to 4 scale. Four points were awarded if an answer showed complete semantic understanding of the word; 3 points if an answer showed a good understanding of the word (e.g., one or more features of the concept); 2 points if the answer provided correct but generic information that doesn't help to identify the element in an unequivocal way (e.g., giraffe = animal; anger $=$ a feeling); 1 point if the answer was not wrong, but poor in content (e.g., evening $=$ is when we dine); 0 point if the answer was wrong; no answer was given; or the concept was repeated (e.g. Photo $=$ to take a photo). Scoring was performed by two independent researchers who were blind to the identity or diagnosis of the children. Interclass correlation coefficient (ICC) was computed to determine the level of agreement between the two scorers, yielding a high degree of reliability, ICC $=.86$ ( $95 \% \mathrm{CI}$ : . $845-.879$ ), $p<$ .001. A third independent researcher moderated instances in which the scores differed by more than 1 point ( $12.6 \%$ of all definitions), and the instances in which only one scorer awarded a score of 0 ; all other scores were averaged.
Lexical decision. Children were presented with six practice trials (three non-words and three words that were not used in the experiment). In each trial, a cartoon alien was presented in the middle of the screen for 1000 ms , followed by the auditory presentation of either a real English word or a nonword. Immediately after the offset of the word (average stimulus duration $=830 \mathrm{~ms}$ ), two touch screen buttons appeared at the bottom left (a red thumbs-down icon) or the bottom right (a green thumbs-up icon) of the screen (see Figure 1).


Figure 1 - Lexical decision task. Trial timeline.
Children were instructed to press the green button when they heard a word they knew, or the red button if they heard a "funny, made-up" word. After the six practice items, participants completed all 96 items ( 24 abstract and 24 concrete words, plus 48 non-words) presented in a randomised order.
Data analysis Separate mixed-design ANOVAs with concreteness (abstract, concrete) as within-subject factor and
group (LI, TD) as between-subject factor were used to analyse average rating (in the definition task) and accuracy (in the lexical decision task), for both age-matched groups and vocabulary-matched groups.

We further assessed the individual performance of children with LI on abstract and concrete words against the difference in those conditions exhibited by matched TD controls, using the Revised Standardized Difference Test (RSDT) (Crawford and Garthwaite, 2005a, 2005b). This test was developed in neuropsychology research to test for dissociation between patient performance on two or more tasks. Here, the two concreteness sets (abstract, concrete) are treated as the two different 'tasks', and the difference in performance between them is evaluated against TD averages. The RSDT controls for Type I error rates when there are correlations between the tasks under study; we entered simple correlations between abstract and concrete raw scores from the TD groups.

## Results

## Definitions

Only $13.4 \%$ of our TD children could provide any definition for words of AoA block 4 (words acquired at 1011); therefore, we excluded block 4 from further analyses, thus reducing the total number of items to 36 words (18 abstract and 18 concrete).
LI vs TD ${ }_{\text {age. }}$ Average accuracy ratings for definitions provided by children with LI and matched $\mathrm{TD}_{\text {age }}$ children for concrete and abstract words are shown on Figure 2 (topleft). A mixed ANOVA yielded a significant main effect of concreteness, $\mathrm{F}(1,34)=9.277, \mathrm{p}=.004, \eta_{p}^{2}=.214$, with concrete words (1.31) attracting more complete and accurate definitions than abstract words ( 0.80 ). The main effect of group was also significant, $\mathrm{F}(1,34)=20.314, \mathrm{p}<.001, \eta_{p}^{2}$ $=.374$; definitions provided by children with LI $(0.80)$ were rated as significantly poorer in quality than those given by their age-matched TD peers (1.63). However, the group $\times$ concreteness interaction was not significant ( $\mathrm{p}=.427$ ), indicating that poor quality definitions were provided for both abstract and concrete words by children with LI.
LI vs TD vocab. One $\mathrm{TD}_{\text {vocab }}$ child did not complete the task and his definitions were excluded along with data from the matched child with LI. Average ratings of definitions provided by LI and matched $\mathrm{TD}_{\text {vocab }}$ children ( $\mathrm{n}=17$ per group) for concrete and abstract words are shown on Figure 2 (top-right). Analyses demonstrate a significant main effect of concreteness, $\mathrm{F}(1,32)=21.687, \mathrm{p}<.001, \eta_{p}^{2}=.404$, with concrete words (1.31) eliciting more accurate and detailed definitions than abstract words (0.86). Importantly, the group $\times$ concreteness interaction were not significant (all $\mathrm{p}>.170$ ).
Individual LIs vs control group comparisons. Individual performance of LI children against matched $\mathrm{TD}_{\text {age }}$ and $\mathrm{TD}_{\text {vocab }}$ groups is shown on Figure 2 (bottom). For all children with the LI group, the abstract vs concrete
comparison was not significantly different from the pattern shown by both age-matched and vocabulary-matched peers.


Figure 2 - Top: Average score (on a 0-4 scale) of definitions to abstract and concrete words, comparing performance of LI with $\mathrm{TD}_{\text {age }}\left(\mathrm{N}=18\right.$; left), and with $\mathrm{TD}_{\text {vocab }}(\mathrm{N}=17$; right) children. Error bars indicate standard error of the mean. Bottom: Proportion of errors for individual LI children and the $\mathrm{TD}_{\text {age }}$ and $\mathrm{TD}_{\text {vocab }}$ groups in defining abstract and concrete words. ${ }^{\wedge}$ Child LI4 was not included in the comparison with the $\mathrm{TD}_{\text {vocab }}$ group. Error bars for the TD groups data indicate standard error of the mean.

Lexical decision. In order to ensure children attention and compliance to task instructions, the examiner controlled stimulus presentation and did not ask the children to respond quickly, but rather as accurately as possible. Reaction times are therefore not reliable and our analyses are limited to accuracy (proportion of correct responses).
Pre-processing. We checked the children's overall performance with words and non-words to determine whether some of the children showed a bias toward either answering "word" or "non-word". We computed the response bias (or criterion, c), calculated by multiplying the sum of the normalised hit rate (correctly identifying a word) and the normalised false alarm rate (incorrectly claiming that a non-word was a word) by -0.5 (e.g., Fox, 2004). The average criterion bias was $-0.002(\mathrm{SD}=0.33)$ for TD children, and $-0.02(\mathrm{SD}=0.50)$ for children with LI. Children who showed a criterion bias higher than 1.5 standard deviations above their group mean (indicating a strong bias toward "non-word" responses) or lower than 1.5 standard deviations below their group mean (indicating a strong bias toward "word" responses) were excluded from further analyses. Using these criteria, 3 children were excluded from the LI group (LI9: $c=-0.97$; LI12: $c=-0.74$; LI17: $c=-0.97$ ); to maintain the matching between the LI and TD groups, we also excluded the corresponding TD children.
$\mathbf{L I}_{\text {age }} \mathbf{v s} \mathbf{T D}_{\text {age. }}$. Proportion of correct responses of the 14 LI and matched $\mathrm{TD}_{\text {age }}$ children for concrete and abstract words is shown on Figure 3 (top-left). There was no main effect of concreteness, $F(1,26)=1.203, p=.283$; but the main effect of group was significant, $\mathrm{F}(1,26)=7.971, \mathrm{p}=.009, \eta_{p}^{2}=$ .235 , indicating that $\mathrm{TD}_{\text {age }}$ children (.85) were more accurate overall than children with $\mathrm{LI}_{\text {age }}$ (.72). Crucially, the group $\times$ concreteness interaction was not significant.
$\mathbf{L I}_{\text {vocab }}$ vs $\mathbf{T D}_{\text {vocab. }}$. Two $\mathrm{TD}_{\text {vocab }}$ children did not complete the task; therefore they were excluded along with their matched LI partner. This left 12 children per group for the LI - $\mathrm{TD}_{\text {vocab }}$ comparison. The proportion of correct responses of LI and $\mathrm{TD}_{\text {vocab }}$ children for concrete and abstract words is shown on Figure 2 (top-right). In this analysis, there were no significant main effects of concreteness, $F(1,22)=1.234, p=.279$, or valence, $F(2$, 44) $=.376, p=.689$. Crucially, the main effect of group and the group $\times$ concreteness interaction were also not significant (all $\mathrm{p}>.330$ ).


Figure 3 - Top: Proportion of correct responses to abstract and concrete words, comparing performance of LI with $\mathrm{TD}_{\text {age }}\left(\mathrm{N}=14\right.$; left), and LI with $\mathrm{TD}_{\text {vocab }}(\mathrm{N}=12$; right) children. Bottom: Proportion of correct responses of individual LI children and the $\mathrm{TD}_{\text {age }}$ and $\mathrm{TD}_{\text {vocab }}$ groups for recognition of abstract and concrete words. Children LI3 and LI7 were not included in the comparison with the $\mathrm{TD}_{\text {vocab }}$ group. The asterisk indicate one child who showed a greater difference between abstract and concrete words when compared against $\mathrm{TD}_{\text {age }}$ data ( $\mathrm{p}<0.05$, two-tailed). Error bars indicate standard error of the mean.
Individual LIs vs control group comparisons. Individual performance of LI children against matched $\mathrm{TD}_{\text {age }}$ and $\mathrm{TD}_{\text {vocab }}$ groups is shown on Figure 3 (bottom). In general, the discrepancy between abstract and concrete words was not significantly different from the discrepancy pattern shown by both age-matched and vocabulary-matched TD children. Only one child with LI (LI8) showed a
significantly larger difference between abstract and concrete words when compared to $\mathrm{TD}_{\text {age }}$ peers, $\mathrm{t}(13)=3.342, \mathrm{p}=$ .005. This difference reflected an advantage for concrete (.79) over abstract (.42) words. In all other children, the abstract vs concrete comparison was not significantly different from either $\mathrm{TD}_{\text {age }}$ or $\mathrm{TD}_{\text {vocab }}$ matched controls.

## Discussion

We compared performance of children with LI to that of age-matched or vocabulary-matched TD peers on two tasks: the first, a verbal definitions task, provided an explicit measure of children's semantic knowledge of abstract and concrete words. The second, a lexical decision task, did not require linguistic output and served as implicit measure of word processing. Both tasks used concrete and abstract words that were matched on a number of variables known to affect word processing in adults, such as frequency, valence, age of acquisition and length.

In the definition task, we found a significant effect of concreteness, indicating that concrete words were easier to define by both children with LI and their age- and vocabulary-matched peers. This may be because to define abstract words, children need to retrieve other abstract words and these latter may be more difficult (not just because of their abstractness but also because they may be longer, less familiar etc) than the concrete words they need to retrieve for defining the concrete stimuli. Importantly, we found that children with LI were significantly worse than their age-matched peers in defining all words, both concrete and abstract. When compared with younger TD children matched for receptive vocabulary, no difference was found between the two groups. There were no significant interactions between concreteness and group, indicating that even when LI children are worse overall than TD peers in defining words, they do not have disproportionate difficulties defining abstract words. By analysing the performance of individual LI children against the difference between abstract and concrete words shown by the two TD comparison groups, we further demonstrated that this is a finding consistent across the whole sample. No individual child with LI showed a greater impairment defining abstract words relative to concrete words.

In the lexical decision task, there was no significant effect of concreteness, which is consistent with findings in the adult literature that, when all lexical variables that have been shown to favour concrete words (such as length and familiarity) are tightly controlled, the concreteness advantage disappears (see Kousta et al., 2011). Critically, we again found that children with LI were significantly less accurate overall in making decisions about words relative to their age-matched TD peers, but there was no interaction between group and concreteness. In other words, even on this implicit task, children with LI were not disproportionately impaired in their processing of abstract words compared to concrete words. The case-series analyses comparing individual LI children with their age- or vocabulary-matched controls once again confirmed that
even at an individual level, children with LI responded to abstract and concrete words in a similar manner to that of TD children, for all but one child (LI8).

These findings challenge any theory that posits linguistic competence as a necessary prerequisite for acquiring abstract words. Children with LI do not have the same vocabulary competence as typically developing children (McGregor et al., 2013), moreover it has been shown that they do not take advantage of correlational information to the same extent as their typically developing peers (Evans, Saffran \& Robe-Torres, 2009). For all these reasons, learning abstract words should present an almost insurmountable challenge for them. However, children with LI in the current study, despite their language limitations, did not show any evidence of disproportionate deficits in abstract word knowledge.

Distributional Semantics models offer a powerful mechanistic account of how word meanings can be acquired from language. On the basis of the linguistic contexts in which a word is used, children could make inferences about word meaning (e.g., Landauer \& Dumais, 1997, Griffiths et al., 2007; Andrews, Vinson \& Vigliocco, 2009). Such a mechanism would be at play for both concrete and abstract words, although it could play a greater role for abstract words. Our results indicate that whereas these mechanisms can be at play, there is no evidence for them to have a different role for concrete and abstract words. Our results may also be considered to be problematic (especially if replicated with argument bearing verbs) for the "syntactic bootstrapping hypothesis (e.g., Gleitman et al., 2005), according to which phrasal and syntactic information is used to constrain possible word meanings. Gleitman et al. (2005) specifically discuss how this information may be especially important in learning verbs (which are more abstract than nouns). Under the plausible assumption that our children with LI have a history of problems in processing sentencelevel linguistic structure, our results suggest that such a strategy may not be the only manner in which children learn abstract vocabulary.

Thus, other mechanisms are also at play. Ponari, Norbury, and Vigliocco (in press) presented initial evidence that learning abstract words could be based on multiple strategies and, at least in the earlier stages of acquisition, take advantage of the strong association between abstractness and emotional valence (Kousta et al., 2011). Emotional valence could support the establishment of the distinction between concrete and abstract domains of knowledge because, while concrete words would refer to observable entities and actions that we can experience with our senses and act upon, abstract words would refer to internal states of self and others that trigger embodied emotional reactions and experiences. These emotional reactions could come about from interactions with caregivers in which children associate words heard with emotions expressed by the caregivers or by the child themselves. Such a view posits that communicative social interaction would play a central role in language acquisition,
along the lines proposed by recent social-cognitive theories of lexical development (e.g., Tomasello, 2000). To this end, it is important to note that children with LI do not show evidence of fundamental socio-cognitive deficits. It is interesting to note that emotion, however, does not seem to have a privileged role abstract vocabulary after the age of 9 (cf. Ponari et al., 2017). It is likely that by this age, strategies grounded in basic socio-cognitive processes (e.g., ability to make inferences on others intentions) or emotional experience become insufficient to differentiate among the meanings of an increasingly larger number of abstract words in the child's vocabulary and language-based strategies may become more important.

## References

Andrews, M., Vigliocco, G., \& Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. Psychological Review, 116(3), 463-98.
Baldwin, D. A. (1991). Infants' Contribution to the Achievement of Joint Reference on JSTOR. Child Development, 62(5), 875890.

Balota, D. a, Yap, M. J., Cortese, M. J., Hutchison, K. a, Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39(3), 445-59.
Bruni, E., Tran, N. K., \& Baroni, M. (2014). Multimodal Distributional Semantics. Journal of Artificial Intelligence Research, 49, 1-47.
Coltheart, M. (1981). The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4), 497-505.
Crawford, J. R., \& Garthwaite, P. H. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. Neuropsychology, 19(3), 318-331.
Dunn, L. M., Dunn, L. M., Whetton, C., \& Burley, J. (1997). British Picture Vocabulary Scale, 2nd edition. Windsor, UK: NFER-Nelson.
Evans, J., Saffran, J. \& Robbe-Torres, K. (2009). Statistical learning in children with specific language impairment. Journal of Speech, Language and Hearing Research, 52(2), 321-335.
Firth, J. R. (1957). Papers in linguistics, 1934-1951. Oxford University Press.
Fox, J. R. (2004). A Signal Detection Analysis of Audio/Video Redundancy Effects in Television News Video. Communication Research, 31(5), 524-536.
Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., \& Trueswell, J. C. (2005). Hard words. Language Learning and Development, l(1), 23-64.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114(2), 21144.

Johns, B. T., \& Jones, M. N. (2012). Perceptual inference through global lexical similarity. Topics in Cognitive Science, 4(1), 103-20.
Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., \& Del Campo, E. (2011). The representation of abstract words: why emotion matters. Journal of Experimental Psychology. General, 140(1), 14-34.

Kuperman, V., Stadthagen-Gonzalez, H., \& Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, 44(4), 978-90.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.
McGregor, K. K., Oleson, J., Bahnsen, A., \& Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. International Journal of Language \& Communication Disorders, 48(3), 307-319.
Medina, T. N., Snedeker, J., Trueswell, J. C., \& Gleitman, L. R. (2011). How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences, 108(22), 9014-9019.
Morse, A. F., Benitez, V. L., Belpaeme, T., Cangelosi, A., \& Smith, L. B. (2015). Posture Affects How Robots and Infants Map Words to Objects. PLOS ONE, 1O(3), e0116012.
Norbury, C.F. et al. (2016). The impact of non-verbal IQ on prevalence and clinical presentation of language disorder: evidence from a population study. Journal of Child Psychology and Psychiatry, 57(1), 65-73.
Ponari, M, Norbury, C. \& Vigliocco, G. (2017). Acquisition of abstract concepts is influenced by emotional valence. Developmental Science, DOI: 10.1111/desc. 12549
Rice, M. L. (2013). Language growth and genetics of specific language impairment. International Journal of SpeechLanguage Pathology, 15(3), 223-33.
Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), The psychology of word meanings (pp. 223-250). Hillsdale, NJ, England: Lawrence Erlbaum Associates.
Semel, E., Wiig, E. H., \& Secord, W. A. (2006). Clinical Evaluation of Language Fundamentals (4th ed.). San Antonio, TX: PsychCorp.
Tomasello, M. (2000). The Social-Pragmatic Theory Of Word Learning. Pragmatics.
Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., \& O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. Journal of Speech, Language, and Hearing Research: JSLHR, 40(6), 1245-60.
Warriner, A. B., Kuperman, V., \& Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45, 1191-207.
Wechsler, D. (1999). Wechsler Abbreviated Scale of Intelligence (WASI) - Subject Baseline. New York, NY: The Psychological Corporation: Harcourt Brace \& Company.
Yu, C., \& Smith, L. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. Psychological Science, 18(5), 414-420.

# Facial Motor Information is Sufficient for Identity Recognition 

Jonathan Vitale, Benjamin Johnston, \& Mary-Anne Williams<br>\{jonathan.vitale, benjamin.johnston, mary-anne.williams\}@uts.edu.au<br>University of Technology Sydney - Centre For Artificial Intelligence<br>Innovation and Enterprise Research Lab (The Magic Lab) - 15 Broadway, Ultimo NSW 2007 - Australia


#### Abstract

The face is a central communication channel providing information about the identities of our interaction partners and their potential mental states expressed by motor configurations. Although it is well known that infants ability to recognise people follows a developmental process, it is still an open question how face identity recognition skills can develop and, in particular, how facial expression and identity processing potentially interact during this developmental process. We propose that by acquiring information of the facial motor configuration observed from face stimuli encountered throughout development would be sufficient to develop a face-space representation. This representation encodes the observed face stimuli as points of a multidimensional psychological space able to assist facial identity and expression recognition. We validate our hypothesis through computational simulations and we suggest potential implications of this understanding with respect to the available findings in face processing.


Keywords: face perception; face processing; face-space; face identity processing; face expression processing; mirroring

## Introduction

Face processing capabilities are of paramount importance for the development of social skills (Grossmann, 2015).

Developmental studies suggest that newborns can match observed facial motor configurations via overt imitative behaviour (Meltzoff \& Moore, 1983, 1992) or covert inner simulation mechanisms (Simpson, Murray, Paukner, \& Ferrari, 2014; Gallese \& Caruana, 2016), even well before the development of early cognitive capabilities (but see Oostenbroek et al., 2016 and Simpson et al., 2016 for a recent discussion on the topic). Hence, it has been suggested that facial expression recognition may be mediated by early neural mechanisms mapping sensory information of the observed facial configuration into a proprioceptive motor format (Gallese \& Caruana, 2016; Iacoboni, 2009) and therefore assisting imitatory mechanisms (Simpson et al., 2014).

On the contrary, face identity processing capabilities follow a developmental process (Grossmann \& Vaish, 2009). Currently, facial identity processing development is not yet well understood. For example, we do not know yet where in the face processing hierarchy representations of invariant (i.e. identity features of the face) and dynamic (i.e. motor features of the face) features interact (Simion \& Di Giorgio, 2015).

According to the 'face-space' framework (Valentine, 1991; Valentine, Lewis, \& Hills, 2015), facial representations are encoded in a multidimensional psychological space. The dimensions of this space are assumed to encode properties of the facial signals that better discriminate one face from another. The distance between two representations underlies their dissimilarity from a psychological perspective. This
framework was initially designed to only account for coding identity-related features, such as sex, distinctiveness, age and attractiveness (Valentine, 1991). Nevertheless, dynamic aspects of faces, such as facial expressions, were neglected. Recently, we developed a computational tool building on top of the face-space framework (Vitale, Williams, \& Jonhston, 2016) and able to exhibit interesting features in agreement with modern understanding in face processing studies. In particular, we demonstrated that this novel face-space can represent both invariant and dynamic features of face stimuli under a shared representation facilitating the recognition of both facial expression and identity exhibited by novel face stimuli (Vitale et al., 2016).

In this paper we offer a new understanding of this facespace, suggesting that facial identity processing capabilities can plausibly develop by interpreting the motor configuration of observed face stimuli.

In particular, from a functional level of analysis, we aim to demonstrate that assuming the existence of an early or innate system $\mathscr{M} \operatorname{otor}\left(x_{i}\right) \Rightarrow \mathscr{E}\left(x_{i}\right)$ able to map perceptual information of the observed face stimulus $x_{i}$ onto a motor interpretation of the exhibited facial expression $\mathscr{E}\left(x_{i}\right)$, it is possible to develop another system Cognitive $\left(X_{\text {new }}\right) \Rightarrow$ $\left\{\mathscr{E}\left(X_{\text {new }}\right), \mathscr{I}\left(X_{\text {new }}\right)\right\}$ assisting the discrimination of facial expressions $\mathscr{E}\left(X_{\text {new }}\right)$ and identities $\mathscr{I}\left(X_{\text {new }}\right)$ exhibited by newly encountered face stimuli $X_{\text {new }}$. Therefore, this paper aims to provide computational evidence supporting the following hypothesis:

Hypothesis: It is possible to generalise the face-space framework to realise a twofold multidimensional space structure able to facilitate facial expression and identity processing capabilities by only interpreting the motor configuration exhibited by the face stimuli encountered during the developmental process.

This work is a significant contribution able to provide a plausible explanation unifying traditional and modern findings in face processing studies, as we will discuss in the remainder of this paper.

## Previous Findings

Recently, we provided a novel understanding of the facespace framework (Vitale et al., 2016). The face-space framework is a widely used tool in face perception and processing research able to explain many of the phenomena underlying facial identity discrimination in both human experimental settings (Lee, Byatt, \& Rhodes, 2000; Rhodes, Jaquet, et


Figure 1: The dual face-space presents a twofold structure: on one side it allows observations with similar motor configurations to lie within close spatial locations $(\uparrow)$, whereas at the same time "repulsing" observations of similar identities away $(\dashv \vdash)$; on the other side, it happens exactly the viceversa. This facilitates respectively facial expression and identity recognition, under common multidimensional codings.
al., 2011) and computational simulations (A. J. Calder, Burton, Miller, Young, \& Akamatsu, 2001). This framework is so important in face studies that it is "virtually impossible to explain the interactions between the computational and cognitive approaches to understanding face recognition without reference to this model. It serves as the glue that binds the theoretical and computational aspects of the problem together" (A. Calder, 2011, page 17).

According to Valentine's face-space, faces are points of a multidimensional space based on their perceived properties. This structure can plausibly account for coding identityrelated features. Unfortunately, dynamic aspects of the face, such as its motor configuration, were neglected in the traditional face-space account. This is a significant limitation, preventing the analysis of the interactions happening between facial expression and facial identity processing.

Therefore, to fill this gap, we introduced a novel hypothesis: the duality hypothesis. This hypothesis suggests that the face-space can plausibly exhibit a twofold structure integrating both dynamic and invariant features of the face into shared codings, although preserving some separation among them to facilitate both facial expression and identity recognition (see Figure 1 for a visual example). We named this understanding with dual face-space and we validated the hypothesis, from a computational perspective, through a mathematical presentation and quantitative results.

## The Dual Face-Space

Given a set of face stimuli shaped as column vectors of a matrix $X$, these stimuli have dimension $\mathscr{D}$ equal to the total number of pixels representing each face stimulus. By submitting the matrix $X$ to a Principal Component Analysis (PCA) (Turk \& Pentland, 1991) it is possible to obtain a mapping matrix $V_{p c a}$ able to map the $\mathscr{D}$-dimensional face stimuli $X$ into compressed d-dimensional representations $\bar{X}$. This process preserves most of the information carried by the face stimuli, but it compresses them in representations having di-


Figure 2: An example of face-space development resulting by applying the mapping function in Equation 2. Face samples belonging to the same identity are on average perceptually closer to each other, thus being a bias for the classification of facial expressions.
mension $d \ll \mathscr{D}$ and it ensures desirable properties in subsequent stages of the model (e.g. positive definiteness, see Vitale et al., 2016):

$$
\begin{equation*}
\bar{X}=V_{p c a}^{\top} X \tag{1}
\end{equation*}
$$

It is important to note that in this paper we do not aim to test the classification performance of the proposed model against other computational models of face recognition, but rather the plausibility of the proposed hypothesis in providing a new understanding of the mechanisms potentially underlying human face processing skills. Therefore, in our studies we used the pixels intensities of static images as input to our models to provide a simplified linear understanding of our theory and related argument. Importantly, the input $\bar{X}$ can be any vector of features extracted by the given face stimuli and able to encode perceptual information of the observed stimuli. Therefore, a viable non-linear alternative of our model can be obtained by pre-processing the input face stimuli $X$ by using an unsupervised deep neural network model trained to preserve invariant and dynamic features of the face in a more compressed and smart representation (Le et al., 2013), instead of the proposed linear PCA. Finally, temporal dynamics can be included by pre-processing a set of consecutive stimuli instead of static images, or by using other techniques improving temporal coherence in the resulting pre-processed representation (Mobahi, Collobert, \& Weston, 2009). These computational pre-processing stages resemble early processing of human visual cortex and are therefore suitable examples for potential future extensions of our theory and related model.

In our previous work (Vitale et al., 2016), we showed that it is possible to implement the dual face-space by solving the following objective function:

$$
\begin{equation*}
V^{\star}=\underset{V \in \mathbf{R}^{d \times d}}{\arg \min } \frac{\operatorname{Tr}\left(V^{\top} \bar{X}\left(I_{N}-W^{\mathscr{E}}\right) \bar{X}^{\top} V\right)}{\operatorname{Tr}\left(V^{\top} \bar{X}\left(I_{N}-W^{\mathscr{I}}\right) \bar{X}^{\top} V\right)} \tag{2}
\end{equation*}
$$

where $W^{\mathscr{E}}$ and $W^{\mathscr{I}}$ are two weight matrices setting desired topological constraints on the face-space via the resulting objective mapping matrix $V^{\star}$. It is possible to obtain the weight
matrix $W^{\mathscr{E}}$ by knowing the facial expressions exhibited by the training samples and, when this matrix is used in Equation 2, it encourages pairs of samples associated with the same facial expression to be in nearby locations in the resulting facespace:

$$
W_{i j}^{\mathscr{E}}= \begin{cases}\frac{1}{n_{\mathscr{E}}}, & \text { if } \mathscr{E}\left(x_{i}\right)=\mathscr{E}\left(x_{j}\right)  \tag{3}\\ 0, & \text { otherwise }\end{cases}
$$

In Equation 3, $n_{\mathscr{E}_{i}}$ is the number of samples in $X$ belonging to the facial expression class $\mathscr{E}\left(x_{i}\right)$ of the face stimulus $x_{i}$ in the column $i$ of matrix $X$.

It is possible to realise the weight matrix $W^{\mathscr{I}}$ by knowing the identities exhibited by the training samples and, when this matrix is used in Equation 2, it promotes repulsive forces between pairs of samples belonging to the same identity, thus reducing misclassification of facial expressions due to the identity bias (Sariyanidi, Gunes, \& Cavallaro, 2015):

$$
W_{i j}^{\mathscr{J}}= \begin{cases}\frac{1}{n_{\mathscr{I}}}, & \text { if } \mathscr{I}\left(x_{i}\right)=\mathscr{I}\left(x_{j}\right)  \tag{4}\\ 0, & \text { otherwise. }\end{cases}
$$

In Equation 4, $n_{\mathscr{I}_{i}}$ is the number of samples in $X$ belonging to the identity class $\mathscr{I}\left(x_{i}\right)$ of the face stimulus $x_{i}$ in the column $i$ of matrix $X$. Figure 1 and Figure 2 show examples of the rationale behind the constraints set by the suggested weight matrices in Equation 2.

Finally, given a generic matrix $M$ and the following permutation function:

$$
\tilde{M}=\sigma(M)=\left(\begin{array}{ccccc}
m^{1} & m^{2} & m^{3} & \ldots & m^{d}  \tag{5}\\
m^{d} & m^{d-1} & m^{d-2} & \ldots & m^{1}
\end{array}\right)
$$

permutating each column vector $m^{i}$ with $i \in[1, \ldots, d]$ of the matrix $M$ in the inverse order ${ }^{1}$ we demonstrated that Equation 2 is sufficient to provide multidimensional representations able to facilitate both facial identity and expression recognition (Vitale et al., 2016).

In fact, given $V^{\star}$ as the optimal solution of the objective function in Equation 2, we demonstrated that the mapping matrix $\tilde{V}^{\star}=\sigma\left(V^{\star}\right)$ is the optimal solution of another objective function promoting facial identity discrimination obtained by inverting Equation 2. The mapping matrix $\tilde{V}^{\star}$ is dual to the mapping matrix $V^{\star}$, since it shares the same components (i.e. column vectors) of $V^{\star}$ but sorted in the opposite order. Therefore, the objective function in Equation 2 realises common codings able to facilitate on one hand facial expression classification $\left(V^{\star}\right)$, and on the other hand facial identity discrimination $\left(\tilde{V}^{\star}\right)$.

## The $\Delta$ Face-Space

To validate our hypothesis, we suggest to approximate the weight matrix $W^{\mathscr{I}}$ with another weight matrix $W^{\Delta}$ implemented without necessarily knowing the identity classes of

[^537]the training face stimuli. In this way the weight matrix $W^{\mathscr{I}}$ in Equation 2 can be replaced by the matrix $W^{\Delta}$, thus realising the following objective function:
\[

$$
\begin{equation*}
V^{\Delta \star}=\underset{V \in \mathbf{R}^{d \times d}}{\arg \min } \frac{\operatorname{Tr}\left(V^{\top} \bar{X}\left(I_{N}-W^{\mathscr{E}}\right) \bar{X}^{\top} V\right)}{\operatorname{Tr}\left(V^{\top} \bar{X}\left(I_{N}-W^{\Delta}\right) \bar{X}^{\top} V\right)} \tag{6}
\end{equation*}
$$

\]

The optimal solution of the objective function in Equation 6 is the mapping matrix $V^{\Delta \star}$. Thus, given a mapping matrix $V_{p c a}$ gathered by submitting the training data $X$ to a PCA, as previously described, it is possible to obtain the final mapping matrix $V_{\text {overall }}^{\Delta}$ realising the $\Delta$ face-space as following:

$$
\begin{equation*}
V_{o v e r a l l}^{\Delta}=V_{p c a} V^{\Delta \star} \tag{7}
\end{equation*}
$$

The mapping matrix $V_{\text {overall }}^{\Delta}$ is able to realise facespace representations facilitating facial expression recognition, whereas the mapping matrix $\tilde{V}_{\text {overall }}^{\Delta}=\sigma\left(V_{\text {overall }}^{\Delta}\right)$, having the same component of $V_{\text {overall }}^{\Delta}$ but permutated in the inverse order, realises representations able to facilitate facial identity discrimination, although without the need of knowing the identities exhibited by the training samples, as suggested by our hypothesis.

## Defining the New Weight Matrix

The purpose of the weight matrix $W^{\mathscr{I}}$ in Equation 2 is to avoid that two face stimuli sharing the same identity, but exhibiting different facial expressions, would get projected to nearby locations of the face-space promoting their misclassification in the same facial expression class (see Figure 2). This misclassification can easily happen since face stimuli of the same identity share most of their perceptual features, and, on average, they are close-by in the perceptual space (Sariyanidi et al., 2015; Turk \& Pentland, 1991). This property exhibited by face stimuli can be used to our advantage to realise the desired weight matrix $W^{\Delta}$.

For each of the $N$ training face stimuli $x_{i}$, shaped as column vectors $i \in[1, \ldots, N]$ of the matrix $X$, we denote with $\Delta_{x_{i}}$ the set containing the perceptual distances $\delta\left(x_{i}, x_{j}\right)$ between the face stimuli $x_{i}$ and the face stimulus $x_{j} \in X$ with $i \neq j$ and exhibiting a different facial expression from the one exhibited by $x_{i}$ :

$$
\begin{equation*}
\Delta_{x_{i}}=\left\{\delta\left(x_{i}, x_{j}\right) \mid x_{j} \in X \wedge x_{i} \neq x_{j} \wedge \mathscr{E}\left(x_{j}\right) \neq \mathscr{E}\left(x_{i}\right)\right\} \tag{8}
\end{equation*}
$$

Since face stimuli of the same identity are perceptually close, their respective distances would be, at least on average, well below their distances from face stimuli with different identities. Then, given the mean $\mu_{\Delta_{x_{i}}}$ and standard deviation $\sigma_{\Delta_{x_{i}}}$ of the distances included in the set $\Delta_{x_{i}}$ it is possible to compute the set $\mathscr{I}_{i} \approx$ described as follow:

$$
\begin{equation*}
\mathscr{I}_{i}^{\approx}=\left\{x_{j} \mid \delta\left(x_{i}, x_{j}\right)<\mu_{\Delta_{x_{i}}}-\beta \sigma_{\Delta_{x_{i}}}\right\} \tag{9}
\end{equation*}
$$

where $\beta$ is a parameter suggesting how many standard deviations below the mean distance would be set the maximum threshold. In this work, $\beta$ was set equal to 2.5 after empirical tests with face stimuli gathered from different datasets
available in face recognition literature. The resulting set $\mathscr{F}_{i} \approx$ includes most of the training samples sharing the same identity of the sample $x_{i}$.

Therefore, the weight matrix $W^{\Delta}$ can be realised as follow:

$$
W_{i j}^{\Delta}= \begin{cases}\frac{1}{n_{\cup i j}}, & \text { if } x_{j} \in \mathscr{I}_{i} \approx x_{i} \in \mathscr{I}_{j} \approx  \tag{10}\\ 0, & \text { otherwise }\end{cases}
$$

where $n_{\cup_{i j}}$ is the number of unique samples in the set $\mathscr{I}_{i} \approx \cup$ $\mathscr{I} \approx$. The realised weight matrix $W^{\Delta}$ is clearly symmetric and the associated Laplacian behaves as a block centring matrix, thus promoting a norm-based space (for in-depth details and mathematical proof refer to Vitale et al. (2016)). The objective function in Equation 6 can be solved through the iterative algorithm proposed by Ngo, Bellalij, and Saad (2012), similarly to our previous contribution (Vitale et al., 2016).

## Experiments

In this paper, we will evaluate the proposed model using the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist, Flykt, \& Öhman, 1998), similarly to our previous contribution. The dataset contains static images of 70 subjects- 35 female and 35 male-exhibiting seven different prototypical facial expressions of basic emotions (anger, disgust, fear, happiness, neutral, sadness and surprise). The pictures are taken in various face orientations and in two different sessions (A and B).

We used the frontal pictures taken in session A . We extracted the facial region from the images and reduced their resolution to $80 \times 80$ pixels. Eyes and mouth were at approximately the same position. Illumination variations were reduced by applying a simple equalisation process to the images (using the histeq function available in Matlab software).

We first pre-processed the data by submitting the pixels of the images in input to a PCA as explained previously. We retained the components able to explain $95 \%$ of the variance of the original data resulting in 200 components.

## Procedure

The present experiment tests the ability of the new $\Delta$ facespace, implemented without knowing the identity labels of the training stimuli, to support subsequent processes of identity and facial expression recognition.

In both the two conditions (i.e. facial expression and identity recognition) we used repeated random iterations of the dataset's samples (in this work 35 iterations for both the tasks). In each iteration 25 subjects were randomly selected as the test set among the 70 possible subjects to simulate unfamiliar identities. For each of the 25 selected subjects were randomly chosen 2 facial expressions as probes for the identity recognition task, and the remaining 5 facial expressions as test samples, leading to a total of 125 test samples for each iteration. The images of the other 45 subjects, together with the 50 selected probes, were used as the training set of the current iteration, leading to 365 training samples for each iteration.

With each training data we estimated the mapping matrix $V_{\text {overall }}^{\Delta}$ of the $\Delta$ face-space proposed in this chapter as per Equations 6 and 7. Then, each test sample was mapped onto the $\Delta$ face-space, thus obtaining the encodings $Y^{\Delta \mathscr{E}}=$ $V_{\text {overall }}^{\Delta \top} X$ and $Y^{\Delta \mathscr{\mathscr { I }}}=\tilde{Y}^{\Delta \mathscr{E}}=\tilde{V}_{\text {overall }}^{\Delta \top} X$, respectively used during the expression and identity recognition tasks for the $\Delta$ face-space condition.

For each iteration, we compared the performance of the $\Delta$ face-space against a baseline approach. The baseline approach used all the pixels of the face stimuli to match similar facial expressions or identities. This is a fair methodology considering we pre-processed raw pixels data with a simple PCA. In our previous contribution (Vitale et al., 2016) we showed that the baseline and PCA performance are not differing. Thus, we used this approach as our baseline to demonstrate that matching the expressions and identities of the considered dataset samples in the perceptual space was not a trivial task and that our psychological face-space can indeed facilitate facial expression and identity recognition.

The classification was performed using the nearest neighbour algorithm. For each sample, $x_{i}$, used by the baseline approach, and $y_{i}^{\Delta}$, used by the face-space model, we computed the Euclidean distances from the centroids of each class in the corresponding space, and we selected the class associated with the centroid closer to the sample.

For each test sample during each iteration, the baseline approach provided a single prediction. Instead, our face-space model can use the first $k=[1, \ldots, d]$ components of the mapping matrix $V_{\text {overall }}^{\Delta}$ to map the face stimuli in face-space representations and perform recognition tasks. Thus, our model provided $d$ predictions for each test sample during each iteration. To gather a single prediction, we selected the most frequent class (mode) predicted by the face-space model for each test sample during each iteration, as per a majority voting approach. For each iteration, we then computed the overall recognition rate for the baseline approach and the $\Delta$ facespace in both facial expression and identity recognition conditions. This process led to 35 samples for each considered approach and task.

## Results

The distribution of the sampled recognition rates was first assessed for normality using a D'Agostino's K-squared test (D'Agostino \& Pearson, 1973) finding that the samples from both facial expression and identity tasks followed a normal distribution ( $p$-values respectively 0.8571 and 0.1382 ). Thus, the effect between the baseline approach and our face-space model were evaluated by a Student's t-test (Keppel, 1991) at a significant level of $\alpha=0.01$. The effect size was assessed by computing Cohen's $d$ (Cohen, 1977).

The results for facial expression and identity recognition are shown in Figure 3a and Figure 3b respectively. From the plots, it is possible to see that the novel $\Delta$ face-space can facilitate both facial expression and identity recognition.

In addition, the t -tests rejected the null hypothesis in both facial expression ( $p$-value $=6.5 \mathrm{e}-19$ ) and facial identity ( $p$ -


Figure 3: Comparative analysis of the performance. (a,b) The recognition rates of the baseline approach and our face-space model respectively during facial expression and facial identity recognition tasks.
value $=1.6 \mathrm{e}-6)$ recognition tasks. The computed effect size suggested a large effect for both the two tasks ( $d=3.03$ for facial expression recognition and $d=0.98$ facial identity recognition). The statistics reached high powers (both $>0.98$ ).

## Potential Implications of the Hypothesis

Although we validated our hypothesis through computational simulations and it is not our aim to suggest that human brain implements the proposed face-space in this way, in this section we will discuss how these results can be of major importance for cognitive science community, at least by focusing at a functional level of analysis.

Modern literature in face perception studies widely suggest interactions between invariant and dynamic features of face stimuli. For instance, it has been shown that women and younger individuals appear to increase cues associated with happiness, whereas men and older people those of anger (Becker, Kenrick, Neuberg, Blackwell, \& Smith, 2007) and studies in face processing broadly suggest that face stimuli can be plausibly represented in multidimensional norm-based spaces (Rhodes \& Jeffery, 2006; Rhodes, Leopold, Calder, \& Rhodes, 2011) and that invariant and dynamic codings of these spaces interact (A. J. Calder et al., 2001).

Interestingly, the proposed hypothesis well integrates with traditional understandings in face studies suggesting distinct routes processing invariant and dynamic features of the face, while still supporting more recent findings suggesting that representations of invariant and dynamic facial features partially overlap (Pell \& Richards, 2013). In fact, Haxby, Hoffman, and Gobbini (2000) suggest that changeable aspects of the face (i.e. eye gaze, expression and lip movement) are processed in the Superior Temporal Sulcus (STS), whereas invariant aspects of the face necessary to classify the exhibited identity are processed in a distinct brain area, the Lateral Fusiform Gyrus (LFG). The STS presents neural connections with the amygdala and other brain areas usually associated
with emotional processing capabilities (Adolphs, 2002) and interactions were observed between the STS and the LFG (Haxby et al., 2000). Recent neuroscience studies suggest that the STS is also related to mirroring mechanisms and imitative capabilities (Buxbaum, Shapiro, \& Coslett, 2014) and Molenberghs, Brander, Mattingley, and Cunnington (2010) provided evidence suggesting that the role of the STS in imitation is not only to passively register observed biological motion, but rather to actively represent sensory-motor correspondences between one's actions and the actions of others. Therefore, the STS, assisted by putative emotional brain areas like the amygdala, can plausibly provide information necessary to interpret the observed facial expression, as suggested in this paper with the assumed system $\mathcal{M}$ otor. This information, in turn, can be then used by the LFG to develop facial identity recognition capabilities, as proposed by the psychological face-space discussed in this paper.

## Conclusions

We provided a new understanding of the face-space framework proposed by Valentine (1991) and able to realise a twofold structure encoding invariant and dynamic features of the face under shared codings and consequently facilitating facial expression and identity recognition capabilities. This face-space can develop by only interpreting motor behaviour exhibited by face stimuli encountered during development. We demonstrated the validity of our claim by providing compelling computational evidence and we discussed the potential implications of this new theoretical understanding in face perception and processing studies. Future works aim in extending the model with non-linear techniques and possibly include temporal features, while at the same time testing the theory by collecting human data from perceptual experiments.

## References

Adolphs, R. (2002). Neural systems for recognizing emotion.

Current Opinion in Neurobiology, 12(2), 169-177.
Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K., \& Smith, D. M. (2007). The confounded nature of angry men and happy women. Journal of Personality and Social Psychology, 92(2), 179.
Buxbaum, L. J., Shapiro, A. D., \& Coslett, H. B. (2014). Critical brain regions for tool-related and imitative actions: a componential analysis. Brain.
Calder, A. (2011). Oxford handbook of face perception. Oxford University Press.
Calder, A. J., Burton, A. M., Miller, P., Young, A. W., \& Akamatsu, S. (2001). A principal component analysis of facial expressions. Vision Research, 41(9), 1179-1208.
Cohen, J. (1977). Statistical power analysis for the behavioral sciences (revised ed.). New York: Academic Press.
D’Agostino, R., \& Pearson, E. (1973). Tests for departure from normality. empirical results for the distributions of b 2 and b1. Biometrika, 60(3), 613-622.
Gallese, V., \& Caruana, F. (2016). Embodied simulation: beyond the expression/experience dualism of emotions. Trends in Cognitive Sciences.
Grossmann, T. (2015). The development of social brain functions in infancy. Psychological Bulletin, 141(6), 1266.
Grossmann, T., \& Vaish, A. (2009). Reading faces in infancy: developing a multi-level analysis of social stimulus. In T. Striano \& V. Reid (Eds.), Social cognition: Development, neuroscience and autism. Oxford, UK: Blackwell Publishing.
Haxby, J. V., Hoffman, E. A., \& Gobbini, M. I. (2000). The distributed human neural system for face perception. Trends in Cognitive Sciences, 4(6), 223-233.
Iacoboni, M. (2009). Do adolescents simulate? developmental studies of the human mirror neuron system. In T. Striano \& V. Reid (Eds.), Social cognition: Development, neuroscience and autism. Oxford, UK: Blackwell Publishing.
Keppel, G. (1991). Design and analysis: A researcher's handbook. Prentice-Hall, Inc.
Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., ... Ng, A. Y. (2013). Building high-level features using large scale unsupervised learning. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8595-8598).
Lee, K., Byatt, G., \& Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. Psychological Science, 11(5), 379-385.
Lundqvist, D., Flykt, A., \& Öhman, A. (1998). The karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, 91-630.
Meltzoff, A. N., \& Moore, M. K. (1983). Newborn infants imitate adult facial gestures. Child Development, 702-709.
Meltzoff, A. N., \& Moore, M. K. (1992). Early imitation within a functional framework: The importance of person identity, movement, and development. Infant Behavior and Development, 15(4), 479-505.

Mobahi, H., Collobert, R., \& Weston, J. (2009). Deep learning from temporal coherence in video. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 737-744).
Molenberghs, P., Brander, C., Mattingley, J. B., \& Cunnington, R. (2010). The role of the superior temporal sulcus and the mirror neuron system in imitation. Human Brain Mapping, 31(9), 1316-1326.
Ngo, T. T., Bellalij, M., \& Saad, Y. (2012). The trace ratio optimization problem. SIAM review, 54(3), 545-569.
Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., ... Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. Current Biology, 26(10), 1334-1338.
Pell, P. J., \& Richards, A. (2013). Overlapping facial expression representations are identity-dependent. Vision Research, 79, 1-7.
Rhodes, G., Jaquet, E., Jeffery, L., Evangelista, E., Keane, J., \& Calder, A. J. (2011). Sex-specific norms code face identity. Journal of Vision, 11(1), 1.
Rhodes, G., \& Jeffery, L. (2006). Adaptive norm-based coding of facial identity. Vision Research, 46(18), 2977-2987.
Rhodes, G., Leopold, D. A., Calder, A., \& Rhodes, G. (2011). Adaptive norm-based coding of face identity. The Oxford Handbook of Face Perception, 263-286.
Sariyanidi, E., Gunes, H., \& Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6), 1113-1133.
Simion, F., \& Di Giorgio, E. (2015). Face perception and processing in early infancy: inborn predispositions and developmental changes. Frontiers in Psychology, 6.
Simpson, E. A., Maylott, S. E., Heimann, M., Subiaul, F., Paukner, A., Suomi, S. J., \& Ferrari, P. F. (2016). Commentary on "Animal studies help clarify misunderstandings about neonatal imitation" by Keven and Akins.
Simpson, E. A., Murray, L., Paukner, A., \& Ferrari, P. F. (2014). The mirror neuron system as revealed through neonatal imitation: presence from birth, predictive power and evidence of plasticity. Philosophical Transactions of the Royal Society B, 369(1644).
Turk, M., \& Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1), 71-86.
Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. The Quarterly Journal of Experimental Psychology, 43(2), 161-204.
Valentine, T., Lewis, M. B., \& Hills, P. J. (2015). Facespace: A unifying concept in face recognition research. The Quarterly Journal of Experimental Psychology, 1-24.
Vitale, J., Williams, M.-A., \& Jonhston, B. (2016, August). The face-space duality hypothesis: a computational model. In 38th Annual Meeting of the Cognitive Science Society (p. 514-519).

# Forgetting My Memories by Listening to Yours: The Impact of Perspective-Taking on Socially-Triggered Context-Based Prediction Error 

Madalina Vlasceanu<br>Princeton University, Princeton, New Jersey, United States<br>Rae Drach<br>State University of New York Albany<br>Alin Coman<br>Princeton University, Princeton, New Jersey, USA


#### Abstract

The mind is a prediction machine. In most situations in which it finds itself, it has expectations as to what might happen. But when people's expectations are invalidated by experience, the memories that gave rise to these expectations are suppressed. The present research explores the effect of these prediction errors on listener's memories during social interaction. We reasoned that listening to a speaker recounting experiences similar to one's own would trigger prediction errors on the part of the listener that would result in the suppression of his/her memories. Study 1 offers evidence for the effect of socially triggered context based prediction errors on listener's mnemonic representations. Study 2 replicates these findings and shows that this effect is sensitive to a perspective-taking manipulation. Taken together, these findings provide evidence for a yet unrecognized phenomenon by which our conversations shape the memories that we come to hold.


# Rank Aggregation and Belief Revision Dynamics 

Igor Volzhanin (ivolzh01@mail.bbk.ac.uk), Ulrike Hahn (u.hahn@bbk.ac.uk), Dell Zhang (dell.z@ieee.org)<br>Birkbeck, University of London<br>London, WC1E 7HX UK

Stephan Hartmann (s.hartmann@lmu.de)<br>Munich Center for Mathematical Philosophy, LMU München<br>Geschwister-Scholl-Platz 1, D-80539 München, Germany


#### Abstract

In this paper we compare several popular rank aggregation methods by accuracy of finding the true (correct) ranked list. Our research reveals that under most common circumstances simple methods such as the average or majority actually tend to outperform computationally-intensive distance-based methods. We then conduct a study to compare how actual people aggregate ranks in a group setting. Our finding is that individuals tend to adopt the group mean in a third of all revisions, making it the most popular strategy for belief revision.


Keywords: rank aggregation; distance measure; probabilistic model

## Introduction

The problem of rank aggregation, where ranked lists from a diverse set of "judges" are combined into a single "consensus" ranked list, is an active research area in computer science. Particularly, rank aggregation has found successful applications in meta-search (Dwork, Kumar, Naor, \& Sivakumar, 2001; Renda \& Straccia, 2003; Fernández, Vallet, \& Castells, 2006), crowd-sourcing (Niu et al., 2015), and recommender systems (Baltrunas, Makcinskas, \& Ricci, 2010).

Although extensive studies have already been conducted on this topic by computer scientists, these largely concern only the algorithmic issues, i.e., how to produce the "optimal" ranked list, without questioning the very concept of "optimal". Typically, a distance measure is chosen, and the ranked list with the minimum total distance to all the given ranked lists is presumed to be the best one (Dwork et al., 2001). In this paper, we challenge such a view and address the problem from the perspective of cognitive science. Just as importantly, much of the previous research has been theoretical in nature and no empirical work has been conducted to determine how humans actually aggregate ranks. To that end, we went beyond the theoretical models described in section 1 and conducted a group study to better understand real-world rank belief revision. To the best of our knowledge, there has been no similar work to date.

## Modeling

In the first instance, we developed a theoretical simulation to test the accuracy of various rank aggregation methods. The simulation can be thought of in terms of the most preferred order in which to display results of a web search.

Given a set of $m$ items (e.g., web pages), we consider $n$ ranked list of them, $\left\{r_{1}, \ldots, r_{n}\right\}$, each of which is given by a
judge (e.g., search engine). One, and only one, of the possible ranking orders (permutations) $r_{*}$ is deemed to be true (correct).

Each judge is characterised by his "competence" which is defined as the probability of providing the true list.

Our simulation takes the various generated lists and aggregated them into a single list using one of the rules outlined further down in this section.

Unlike in some previous work, such as Fernández et al., for each item in a list we know only its rank position, vis-avis other items, and not any other numeric properties. It is often impossible or unrealistic to obtain the scores of individual items and only their relative positioning to each other is available (Dwork et al., 2001; Renda \& Straccia, 2003). More importantly, a wealth of psychological research suggests that, in many domains, humans represent faithfully only ranking order information and more detailed information is unhelpful (Stewart, Chater, \& Brown, 2006)

For the sake of simplicity our modeling considers that each judge will produce a complete list and no ties are possible. So when ranking items, they will rank all of the choices and will rank them relative to each other in such a way that each item will occupy a unique position. Furthermore, every judge in our model has the same level of competence $c \in[0,1]$. Finally, when a certain rule produces multiple lists that are equally optimal, one of them is selected at random. This work could be generalized straightforwardly in the future by relaxing these constraints.

Following rank aggregation methods have been proposed in previous studies and are widely used in practice, so we will use them in our comparison:

- majority: the consensus list is just the ranked list that appears most frequently.
- average: the consensus list is generated by ranking the items according to their average rank positions, which is essentially same as the Borda's count (Dwork et al., 2001).
- Spearman: the consensus list is the one with the minimum sum of Spearman's footrule to all the given ranked list. Spearman's footrule is defined as the total number of displacements needed to achieve parity between two lists.
- Kendall: the consensus list is the one with the minimum sum of Kendall's tau to all the given ranked list. Kendall's
tau is defined as the total number of inversions required to achieve parity between two lists.
- Kemeny-Snell: the consensus list is the one with the minimum sum of Kemeny-Snell distance to all the given ranked list. The Kemeny-Snell (KS) distance is similar to Kendall's tau, but more robust when dealing with ties.

While the first two methods are simple and easy to compute, the other three that are based on distance measures and have a high computational complexity. It has been shown that finding the optimally ranked list based on Kendall's tau (known as the Kemeney optimal aggregation) is an NP hard problem with just four full lists(Dwork et al., 2001).

Our research question is then: "which rank aggregation method is most accurate?" Here by accuracy, we mean how often the consensus list returned by a rank aggregation method is indeed the true list.

## Computer Simulations

We prepared a simulation in $R$, which samples a number of judges and uses different aggregation methods to determine the list reflective of the group of judges. The generated consensus lists are then compared with the true list to calculate accuracy, which we used as our "performance" measure for the aggregation method. This procedure is repeated across pools of judges of different sizes. In order to smooth out effects of randomness, we performed bootstrapping at each number of judges and took the average value. Therefore each set of judges was simulated several times, before adding additional judges.

Our simulation had a number of parameters that could be altered:

- list size: number of unique items in a list
- competence level: individual probability of picking the correct list
- aggregation method: methods of aggregation described above
- number of runs: each run increased the number of judges in a group by one
- number of simulations: a number of repeats of the same simulation with the same conditions to smooth out any noise due to randomness

We began with a list size of 4 . With no ties there are 24 possible permutations. In the simulation $k$ groups consisting of $n$ number of judges would draw a single list from the full list of permutations. Using one of the aggregation methods, a single list would be selected for each group as the aggregate product, and then compared to the true list. Each group of judges would be re-sampled a number of times to boostrap the results to get a smoother result. Thus, scores reported below are the average results sampled over multiple trials for the same group.


Figure 1: The comparison of aggregation methods in the linear-decay error model.

Error Model One important consideration in the study was the underlying error model that governed a judge's probability of picking the wrong list among all possible permutations. Each judge had a competence measure which reflected the probability of picking the true list. The rest of the probability was distributed among the remaining possible choices. Assuming that judges know anything about the domain in question, the probability of picking a wrong list is likely to be an inverse function of the distance from that list to the true list. Without loss of generality, we used the Kemeny-Snell distance measure $d(\cdot, \cdot)$ to determine the probability of a given list being selected as follows.

$$
\operatorname{Pr}\left[r_{i}\right]= \begin{cases}c & \text { if } r_{i}=r_{*}  \tag{1}\\ (1-c) \frac{1 / d\left(r_{i}, r_{*}\right)}{\sum_{j \neq *}\left(1 / d\left(r_{j}, r_{*}\right)\right)} & \text { otherwise }\end{cases}
$$

In effect, lists that are closer to the true list, would be more likely to be drawn than the lists further away.

We wanted to see relative performance of the various aggregation methods, as the number of judges increased. For all results, we maintained a constant competence level $c=0.1$, which meant a $10 \%$ chance for a judge to pick the true list $r_{*}$. We selected the simulation range from 5 to 100 judges.

## Results

After running several different simulations we produced a number of interesting and insightful results. We present our findings in a series of figures that illustrate the relative performance of the different aggregation methods (see figures $1,2,3$ ).

The majority rule performs significantly worse than the alternatives and does not increase in accuracy as the number of


Figure 2: The comparison of aggregation methods in the fastest-decay error model.
judges increases. On the other hand, the other four methods perform similarly to each other and their accuracy increases as the number of judges goes up, as can be observed in figure 1. It is important to note that the Kemeny-Snell aggregation method does not perform significantly better than the other distance-based methods, despite the fact that the underlying error model is based on the Kemeny-Snell distance! Furthermore, average, which is a very simple method (both computationally and cognitively), performs at least on par with the distance-based methods.

A minor comment regarding high competence is due at this point. When the competence level is above a threshold, e.g., 0.2 , we see a quick rise towards perfect accuracy of all methods, which is not particularly interesting, or informative. Therefore, we kept the competence level low and tried to understand how robust different rank aggregation methods would be under the more challenging condition of lower individual competence.

The above linear-decay error model as described in Eq. (1) is just one way of converting the underlying KS-distance to the targeted true list into a probability of erroneous list selection. Actually any monotonic decaying transformations such as an exponential decay - of those distances could be utilised to pick the non-true lists. To generalise our results we considered two extreme cases of monotonic decay functions of distance: at the one end (fastest-decay), the selection probability drops so rapidly as a function of distance that only the closest lists stand a chance of being selected; at the other end (none-decay), the selection function is flat and the lists of all distances are equally likely to be selected. We have examined both of these extreme cases.

First, let us consider the case where only the ranked lists


Figure 3: The comparison of aggregation methods in the none-decay error model.
closest to the true list had a non-zero selection probability (with each list at that distance equally likely to be picked).

From the results of the simulation we see that the majority method plummets almost immediately towards zero accuracy. This is due to the fact that the competence level, i.e., the probability of picking the true list ( $10 \%$ ), is significantly lower than the probability of picking any of those closest lists (which is $30 \%$ in this example as there are three closest lists in total).

Interestingly, the average method appears to outperform the other methods, and quickly moves towards perfect accuracy as the number of judges increases. There appears to be little difference among the other distance-based methods and they all behave similarly to the average method.

Second, we consider the case where a judge is equally likely to pick any of the wrong lists, regardless of their distance to the true list.

The results of this simulation stand in stark contrast to the other two simulations. The majority method performs significantly better and improves with the number of judges, which is exactly reverse of what was observed in the earlier simulations.

Although the observation was initially quite surprising, the explanation is fairly intuitive. Since the probability of picking the true list is $10 \%$, the remaining probability would be distributed evenly over 23 other possible permutations, which leads to only $3.9 \%$ per permutation. Therefore, the ranked list occurred most frequently is almost guaranteed to be the true list, and the majority method would always perform the best.

Just as importantly, the other aggregation methods appear to falter at this stage. Although there is some improvement along with the increase in the number of judges, the
accuracy stays well below 0.5 , even for groups with 100 judges. Notably, the average method performed the worst, while the Spearman method performed the best among the three distance-based methods.

## Discussion

A few key insights emerge from our modeling efforts. The first and most important one is that there appears to be little benefit of using computationally-expensive distance-based methods to conduct rank aggregation. Secondly, there is clear robustness of adopting the group mean. Accuracy is constantly increases in most scenarios and the method itself is simple enough to calculate and act upon.

The one research question that remains open, however, is what real human subjects would do given a similar task. While it may appear that taking the group mean is advantageous from the accuracy point of view, it is also more difficult to determine than simply adopting the majority opinion for example. To test, this we designed a study that looked at individual rank revision in a group setting.

## Experiment - Rank Revision

This experiment was set up to test what rules, if any, individuals use to revise their beliefs in light of new information. Unlike similar studies on the topic which have mostly looked at absolute answers and estimates, we were interested in applying this in the context of rank revision. In other words, our interest was to understand better how participants revise ranked orders when presented with information from their peers. From the modeling exercises above we knew that adopting the group mean is the most beneficial strategy a person can take in most situations, however, we could not locate any research that corroborated this in an empirical study.

## Method

Participants Participants for this study were volunteers from the University of London community. Participants were paid 5 for taking part in the study. There were 19 participants who took part, which created three panels of five participants and one panel with four participants ( $n=19$ ). Each group of participants took part at the same time and were hosted in the same room. No particular exclusion criteria were used and participants were free to self select which of the time slots worked best for them to attend the study. It did not appear that any participants knew each other prior to the study.

Materials \& Procedure Participants were seated in a computer lab, spaced apart in a way that prevented them from seeing each others' screens. Each participant had a computer in front of them that contained a NetLogo interface that was connected in a network to other computers in the room. See Figure 4 for a sample interface that each participant saw.

Initially, participants were read basic instructions regarding the task. The task involved each participant to rank four cities from the largest to smallest by population size. Each city was presented in a text box and contained a number along with


Figure 4: NetLogo participant interface


Figure 5: NetLogo participant interface


Figure 6: Zoomed in participant view
the name of the city (see example in figure 4). In the drop down box 'City A' they were instructed to put the number of the city they believed to be the largest, 'City B' were to contain the second largest, and so on. After all four boxed were filled, participants had to submit their answers and wait for everyone else in the room to finish. Once, all answers were submitted, participants could see how everyone else had ranked the cities. At this point, everyone had an opportunity to revise their answers in light of additional information (see figure 5 and zoomed in view in figure 6). They repeated this process three times for each question, resulting in four rounds - initial round, plus three revision rounds.

In total, each participant answered 21 questions. There is an initial practice question which participants did in a directed manner, followed by 20 other questions, which were done independently and free from any additional instructions. Each question contained a different set of cities and in different order, but the task was the same. There was only a single experimental condition and all participants were treated the same; they were shown the same set of questions, in the same order.


Figure 7: Number of revisions per Participant

## Results

In the first instance we were interested in individual belief revision. We analyzed how often individuals changed their answers and what rules they have have used to do so.

Individual Revision Discounting the first question, there were 60 opportunities for revision for each participant (20 questions * 3 revision rounds). On average participants changed their answers 10.3 (SD 7.51) times over the course of the simulation, or about $10 \%$ of the time. With some participants changed their answers as little as once, and others changed almost a third of their answers. In total there were 196 revision for all participants. See figure 7 for a visual representation of the number of revisions per participant.

Overall, most revisions occurred in the first round, where almost as many revisions occurred as the subsequent two rounds. Table 1 breaks down revisions by round.

Revisions occurred unevenly between questions. Seven questions had between 13 and 15 revisions, while remaining 13 questions had between five and nine revisions.
The number of revisions made by participants was rather low, but the overall profile of the changes, i.e. mostly in the first round and more for some questions than others, is consistent with some of the other studies in the field of decisionmaking.

Table 1: Round Revisions

| Revision | Revision | Revision |
| :--- | :--- | :--- |
| Round 1 | Round 2 | Round 3 |
| 96 | 58 | 42 |

Models of Revision We fitted several models presented in the first part of the paper trying to predict individual belief revision rules that induced the change (such as mean, median and majority models). We decided to restrict our fitting to two models in particular: mean and majority. As these
models had very interesting properties and were most likely to be available and calculable to participants. Since ranked lists were presented near each other identifying the majority list, or calculating the mean list was a conceivable task that a participant could engage in prior to revising their beliefs.
In order to test whether participants actually behaved in a way predicted by a model, we generated an answer that a participant would pick if they were guided by a model and then compared the predicted answer with the actual answer in a binary fashion. We used two models: mean - using simple Borda count - and majority lists.

Table 2 demonstrates that there were significantly more revisions that moved towards the mean than majority. In fact, of the 196 total revisions, 62 , or $31 \%$ were revisions that adopted the group mean, and 44 or $22 \%$ that adopted the majority list. On average, the mean model was adopted 3.26 times per participant, while majority model was adopted 2.32 times. The rest of the revisions were not accounted by these two models and were being guided by unknown rules.

Naturally, there were instances where both models predicted the same list and the above numbers include revisions where the mean and majority lists coincide. There were 35 revisions where both models predicted the same result.

When removed from the total revision count for each model, there were 27 revisions that adopted the group mean and only 9 revisions that adopted the majority list. This provides strong evidence to suggest that participants in our study adopted the group mean much more readily than the majority list.

Table 2: Model Revision

| Model | Total | Model Only | Average | Revision \% |
| :--- | :--- | :--- | :--- | :--- |
| Mean | 62 | 27 | 3.263 | 31 |
| Majority | 44 | 9 | 2.316 | 22 |

Toward a Model of Rank Belief Revision Our findings suggest that human participants are 3 times more likely to adopt the group mean over the majority list in cases where the two do not coincide. This suggests that computational models that emphasize mean ranks may be closer to the way humans make revisions given additional information in a ranked format.

We did not test other, more complex models on the study dataset. Therefore, it is difficult to say at this point whether adopting the group mean is the most preferred strategy. It should also be noted that revisions represented only $11 \%$ of all choices made by participants and most of the time participants did not change their answers and were not influenced by additional information. However, where revisions did occur, in a third of all cases it was towards the group mean, which is a significant finding. Future models that seek to replicate human behaviour should take these findings into account when constructing more human-like models.

## Conclusions

Our research outlined a basic error model as well as two limit cases. In all three scenarios, distance-based methods did not produce significantly better results, suggesting that the problem of rank aggregation could be satisfactorily solved by simpler methods such as taking the average or majority.

As the performances of the two simple methods are diametrically opposite, which method should be used depends on the underlying error distribution in a population. Conversely, if one is able to measure accuracy, the performances of various rank aggregation methods can actually inform us the underlying error distribution and allow us to make inferences about the cognitive process of ranking.

In order to expand on our findings, we conducted a lab experiment where we tested actual belief revision in a group setting. Our findings suggest that when revising their answers, participants most often adopted the group mean, suggesting that human cognition gravitates towards this method of revision. This is significant, in light of the fact that adopting group mean is both computationally less strenuous and quite advantageous in most situations. This suggests that human cognition is adaptive in this sense, using a strategy that our modeling shows to be robust in most cases.

## References

Baltrunas, L., Makcinskas, T., \& Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In Proceedings of the 2010 ACM conference on recommender systems (recsys) (pp. 119-126). Barcelona, Spain.
Dwork, C., Kumar, R., Naor, M., \& Sivakumar, D. (2001). Rank aggregation methods for the web. In Proceedings of the 10th international conference on world wide web ( $w w w$ ) (pp. 613-622). Hong Kong, China.
Fernández, M., Vallet, D., \& Castells, P. (2006). Probabilistic score normalization for rank aggregation. In Proceedings of the 28th european conference on IR research (ecir) (pp. 553-556). London, UK.
Niu, S., Lan, Y., Guo, J., Cheng, X., Yu, L., \& Long, G. (2015). Listwise approach for rank aggregation in crowdsourcing. In Proceedings of the 8th ACM international conference on web search and data mining (wsdm) (pp. 253262). Shanghai, China.

Renda, M. E., \& Straccia, U. (2003). Web metasearch: Rank vs. score based rank aggregation methods. In Proceedings of the 2003 ACM symposium on applied computing (sac) (pp. 841-846). Melbourne, FL, USA.
Stewart, N., Chater, N., \& Brown, G. D. (2006). Decision by sampling. Cognitive Psychology, 53(1), 1-26.

# Overcoming the Tragedy of Personnel Evaluation? <br> Momme von Sydow ${ }^{1,2}$ (momme.von-sydow@urz.uni-münchen.de) Niels Braus ${ }^{2}$ (n.braus@stud.uni-heidelberg.de) Ulrike Hahn ${ }^{1,3}$ (u.hahn@bbk.ac.uk) <br> ${ }^{1}$ University of Munich (LMU), Munich Center for Mathematical Philosophy (MCMP), Ludwigstr. 31, D-80539 München, Germany <br> ${ }^{2}$ University of Heidelberg, Department of Psychology, Hauptstr. 47-51, D-69117 Heidelberg, Germany ${ }^{3}$ Birkbeck College, University of London, Malet Street, London WC1E 7HX, U.K. 


#### Abstract

Human beings are essentially - by nature or second nature members of groups. They contribute to these groups not just as isolated individuals but also through their interaction with others. Consequently, personnel evaluation in companies and organizations requires assessing not only evaluating individual performance but also the overall direct and indirect effect one has on a team. Others' work may be improved or hampered by the presence of a particular employee. We investigate Two-level Personnel-Evaluation Tasks (T-PETs) with information on individual and group earnings, where an individual focus may lead to evaluate the overall best employee as being the worst. We have previously found a Tragedy of Personnel Evaluation where focus on direct individual impact did have such systematic effect. In two experiments, one on team size, the other on kinds of information provided, we explore the boundary conditions of this effect and suggest how it may be overcome.


Keywords: Tragedy of Personnel Evaluation; Rationality of Personnel Decisions; Inner-Individual Dilemma; Social Psychology; Personnel Evaluation; Personnel Selection; Bounded Decision Making; Causal Induction

## Introduction

The success of teams in organizations or companies not only relies on the direct performance of individuals, but often also on interactions between team members (Mathieu, Maynard, Rapp, \& Gilson, 2008; Memmert, Plessner, Hüttermann, Froese, Peterhänsel, \& Unkelbach, 2015). Individuals may, for instance, help or hinder each other. The vital role of prosocial or altruistic behaviours for teams in organisations and companies (George \& Bettenhausen, 1990; Li, Kirkman, \& Porter, 2014; Nielsen, Hrivnak, \& Shaw, 2009; Organ, 1997; Podsakoff, Whiting, Podsakoff, \& Mishra, 2010) and for functioning societies on the whole is being increasingly acknowledged (Engel, 2011; Hendrich et al., 2005; Gollwitzer, Rothmund, Pfeiffer, \& Ensenbach, 2009; Post, 2005; cf. Melis et al., 2016). In Organizational Psychology several types of contextual performance (Organ, 1997; van Scotter \& Motowidlo, 1996) and prosocial behavior (Brief \& Motowidlo, 1986; Li, Kirkman, \& Porter, 2014) have been distinguished. Researchers have also pointed out that not only is prosocial behavior crucial for the success of organizations, but that people are actually sometimes rewarded for it (Organ, 1997; Scotter, Cross, \& Motowidlo, 2000; Grant \& Patil, 2012, 562).

In previous work, we began investigating participants' behaviour as hypothetical human-resource managers evaluating employees working in different configurations
each shift (von Sydow \& Braus, 2016). We employed Twolevel Personnel-Evaluation Tasks (T-PETs) that, across several rounds ('shifts'), provide information on how both individuals and teams contribute to a store's earnings. Crucially, the individual and team information suggest opposite rankings of the employees' contributions. By design, the presence of a so-called 'altruist', someone who positively affects the performance of the others, was most positively correlated with the overall team performance ( $r=.99$ ), even though the altruist individually performed the worst. We focused on the example of one employee strongly affecting the whole group, as this case is influential in biological models of altruism that assume an unconditional advantage to all group members in the presence of an altruist (Sober \& Wilson, 1999; Wilson \& Wilson, 2007; but Nowak \& Sigmund, 2005). The participants' task involved evaluating employees (Personnel Evaluation) and selecting the best team (Personnel Selection). Participants saw only one group (one shop) and the teams were assembled by selecting 4 out of 5 employees (thus 5 team configurations were possible).

Results from von Sydow \& Braus (2016) suggest what they called a "Tragedy of Personnel Selection". After 40 shifts, repeated measurement, and no time-constraint for analyzing the data of a shift, participants systematically judged the overall best employee to be the worst (on the different hidden-profile problem, Mojzisch, Grouneva, \& Schulz-Hardt, 2010). Recently we also explored negative interactions (egoist detection) in T-PETs (subm.). This tragedy is reminiscent of the well-known "Tragedy of the Commons", a notion for the often tragic outcomes of social dilemma situations (such as public-good games). Note however, that the 'T-PETs' do not strictly involve social dilemma, as the participant manager has the explicit goal of choosing the best team for the company. It is only what has been called an inner-individual dilemma (von Sydow, 2015) between two levels of goal descriptions, since it is irrational to optimize more specific goals at the expense of lowering overall utility. Since positive (and negative) interactions with other employees are ubiquitous, and number-based evaluations are important in HR-management, these results suggest such tragedy may well be found in everyday life.

Here we present two new experiments exploring the generalizability or boundary-conditions of the tragedy. Experiment 1 varies team sizes and begins exploring the idea of multiple groups in parallel. Experiment 2 investigates longer learning periods and whether with forced focus on the group level people are able to detect the altruist.

## Experiment 1

## Design

In the T-PETs we provided information on both individual and overall group level earnings. Experiment 1 investigates the extent to which the Tragedy of Personnel Evaluation depends on group size (number of workers: $3,4,5$ versus 7) and the number of groups (one versus two) (Figure 1, Table 1). From Conditions 1 to 4 , group size increases. In a shift, all workers apart from one are working. Condition 5 has the same group size as Condition 4 but is characterized by a group-comparison scenario, where the 6 employees are split in two groups with three employees each.

Table 1: Numbers of workers and groups, and their mean earnings (normal workers, NW; altruist, A) in the five conditions.

|  | C1 | C2 | C3 | C4 | C5 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Number |  |  |  |  |  |
| Groups | 1 | 1 | 1 | 1 | 2 |
| All Workers | 3 | 4 | 5 | 7 | 7 |
| Shown Workers | 2 | 3 | 4 | 6 | $3 / 3$ |
|  | Mean Earnings ( $€$ ) |  |  |  |  |
| NW with A | 3000 | 3000 | 3000 | 3000 | 3000 |
| NW without A | 2000 | 2000 | 2000 | 2000 | 2000 |
| Altruist (A) | 1500 | 1500 | 1500 | 1500 | 1500 |
| Group with A | 4500 | 7500 | 10500 | 16500 | $7500 /$ |
|  |  |  |  |  | 6000 |
| Group without A | 4000 | 6000 | 8000 | 10000 | $6000 /$ |
|  |  |  |  |  | 6000 |

## Method

Participants 221 Participants from the US began the experiment via MTURK. 158 participants finished it, passing all selection criteria (time spent on first page and correct answer out of four, rephrasing the instructions). The participants obtained a reward of $2 \$ .46 \%$ were male, the mean age was 35 years; $53 \%$ mentioned having a Bachelor's or Master's degree, and $39 \%$ a high school degree as highest level of education. Participants were randomly assigned to one of the five conditions.
Procedure and material The payoff structure for the individual employees remains constant over the five conditions, whereas the differing group sizes led to differing earnings of the groups (Table 1). We adopted the repeated measurement design with 18 shifts for each of the 4 rounds, and at the end of a round a rating followed by a selection task. The total number of shifts was 72. (This number is higher than the 40 trials investigated in previous studies.) In C3, for instance, the selection task was to select a team of 4 from 5 available workers that would be best for the company (Figure 1). In C5 people could select 6 workers
from 7 for both groups. In the last round, we additionally asked for the employee with the greatest and lowest utility, and assigned a Need-For-Cognition-Task (Cacioppo, Petty, \& Kao, 1984), a working memory task, a Commentary, and demographics.


Figure 1: Illustration of the materials. In the T-PETs the overview information for each round contains, in the first row the photos of the employee (in random order); in the second row their individual earnings; and in the third row the overall group/team earnings.

## Results



Figure 2: Average ratings (with SE) for the normal workers (N1-N6) and the altruist worker (A) of Conditions 1, 2, 3, 4 and 5 (Panels A to E).

Figure 2 shows that the rating for the altruist in all conditions remained clearly below the other ratings. In an ANOVA of the altruist ratings only, Condition is a significant between-subject factor $(F(4,153)=3.57, p<.01)$ and Phase (the four test phases) a significant within-subject factor (Pillai-Spur-Test, PST, $F(3,151)=3.92, p<.05)$, with no further significant interaction effects ( $p=.32$ ). This outcome seems in line with the prediction that one obtains the best results for the condition with the fewest workers
(C1), despite the altruist increased overall effect on the mean group earnings in larger groups, and for the group condition (C5). Bonferroni-corrected post-hoc comparisons show significant differences between C 1 and C 4 ( $p<.01$ ) and between C 4 and C 5 ( $p<.05$ ). However, the ratings remain predominantly based on individual comparisons in all conditions.


Figure 3: The proportion of 'managers' choosing a team and excluding the altruist $(A)$ or a normal worker $(N)$ in the test phases (Panel A to D) of the personnel selection task. The dark shading represent selections based on individual earnings, the light ones correspond to overall earnings.

Figure 3 shows that in all conditions the largest proportion of participants tends to expulse the overall most useful worker, the altruist, from the team. This is clearly the case in Phase 1 (always, $p<.01$ ). The proportion of other choices overall increased over time $\left(\chi^{2}(1,316)=13.15, p<\right.$ .001). Nonetheless altruist expulsion remained statistically above chance in Phase 4 (apart from C1, $p=.10$; all other $p$ $<.01$ ). For the variable which worker is deemed to have the least utility for the company (not presented here), in Round 4 , the choice of the altruist even remained dominant for all conditions ( $p<.01$ ).


Figure 4: Average altruist-detection rate for Conditions 1, 2, $3,4,5$ in the personnel selection task (Panel A) and the rating task (proportion of altruist rated larger than all other workers; Panel B).

Figure 4A shows the increase of group-level answers particularly for only a few employees (C1). The group condition C5 has a relatively high start but does not
increase. However, Figure 4B suggests that there is also an increase for C 5 if one considers the stricter criterion of rating the altruist to be higher than all other workers.

Figure 5A shows at least descriptively that in the highest utility task the altruist is positively singled out relatively frequently in C 1 and C 5 (but note the different numbers of workers). Additionally, participants' comments were deemed 'insightful' if they detected possible differences between an individual's direct and overall earnings (Figure 5B). Note that this measure is not directly affected by the number of answer-options. There was a reliably higher number of insightful comments in C 1 than $\mathrm{C} 2, \mathrm{C} 3$ and C 4 , but not higher than in C 5 .



Figure 5: A) Proportion of Participants selecting the altruist to be of highest utility. B) Insight rate shown in Comments.

Overall, Experiment 1 shows that the tragedy is quite stable over group size. However, it also suggests that although the altruist's summative effect increased with the group size, the best participant performance was in the condition with the lowest number of employees (C1). The results also only showed subtle advantages of introducing different groups on the altruist detection rate ( C 4 vs C 5 ).

## Experiment 2

Experiment 2 investigates conditions where people are forced to focus on the group level only, to see whether all participants realize that the altruist performs better on this level. We thus investigate whether the tragedy of personnel evaluation is due to an inability to see complex effects on a group level (despite information concerning this level). Additionally, and in contrast to Experiment 1, we distinguish different mean individual earnings of normal workers to check the extent to which people distinguish even slighter performance differences on the individual level.

## Design

The experiment had a mixed 2 (information: global-only versus local-and-global) $\times 2$ (earnings of normal workers: homogeneous versus heterogeneous) between-subjects design, with a within-subjects factor of four test phases (Table 2). In each test phase both evaluation and personnel selection tasks were assigned. Additionally, in the last round, highest and lowest utility tasks and other tests were completed, as well as a Need-For-Cognition (NFC) test.
In the local-and-global conditions, participants were provided with almost the same overview information as in Condition 3 of Experiment 1 (only the Altruist individually contributed 1600 instead of 1600). In each round,
information was given on both the direct earnings of the four workers on a shift and the overall earnings of the shift. The overall earnings involved not only the direct effects of individuals but also their indirect effects. In the global-only conditions, only the overall payoffs of a group (shift) were presented, without showing individual contributions.

Table 2: The four conditions, also showing the overall versus direct impact of a worker on group-earnings

| Condition | C1 C2 | C3 C4 |
| :---: | :---: | :---: |
| Information | Local <br> and <br> global Global <br> only | Local <br> and <br> global Global <br> only |
| Earnings NW | Homogeneous | Heterogeneous |
| Overall impact | $\begin{aligned} & \text { A >> NW1 = NW2 } \\ & =\text { NW3 = NW4 } \end{aligned}$ | $\begin{aligned} & \hline \text { A >> NW1 > NW2 } \\ & >\text { NW3 > NW4 } \end{aligned}$ |
| Direct impact | $\begin{aligned} & \text { NW } 1=\text { NW2 }= \\ & \text { NW3 }=\text { NW } 4 \text { A } \end{aligned}$ | $\begin{aligned} & \text { NW1 > NW2 > } \\ & \text { NW3 > NW4 > A } \end{aligned}$ |

Note: NW = normal worker; A = altruist.
The homogeneous and heterogeneous conditions correspond to either identical or different individual impact of the normal workers (see Table 2). The group earnings remained identical in both kinds of conditions. The 'altruist' (A) always has the most positive impact on the overall earnings. NW earnings (€) without $A$ were 2000 (homogeneous); 1400, 1800, 2200, 2600 (heterogenous); with A, 3000 (homogeneous); 2400, 2800, 3200, 3600 (heterogeneous), but the altruist had the lowest direct (individual) impact, 1600.

## Method

Participants As in Experiment 1, relatively strict selection criteria for participants were used to ensure high data quality. After passing a first criterion (time spent on the first page), 150 people properly started the task and 7 people failed the second criterion (correct rephrasing of the task; four options). Of the remaining 143 volunteers, 122 finished the experiment, and only their data was analysed. Participants were recruited from MTURK: 57\% were male, $42 \%$ female; mean age was 33 , and $68 \%$ had a Bachelor's or Master's degree (with $32 \%$ a high school degree). They received $\$ 2$ for participation.

Procedure and material We used almost the identical materials and procedure as in Experiment 1, C3. The experiment had 80 rounds, with four test phases administered after Rounds 20, 40, 60, and 80. In all four test phases, participants completed both a personnel-evaluation task and a per-sonnel-selection task. In the final test phase, we additionally administered a highest-/lowest-utility task, a ranking task, a Kimchi-Palmer item, an attention-test item, and an 18-item Need-For-Cognition Test (Cacioppo, Petty, \& Kao, 1984).
In the global-and-local conditions, the overview information presented in each round corresponds to C3 in Figure 1. In the global-only conditions, the second line of this panel (presenting the individual earnings of each employee) was omitted.

## Results

Figure 10 shows the mean ratings for the workers' contributions to company earnings. An overall ANOVA with Workers (5 workers) and Phases (4 phases) as within-subjects factors, and Conditions as between-subject factor, yielded a highly significant effect of Conditions $\times$ Workers (Pillai-Spur Test, PST, $F(12,306)=22.5, p<.001)$. This corresponds to the predicted change of rank of the altruist's ratings in the global-only versus global-and-local conditions. Additionally, the factors Workers, Phase $\times$ Condition as well as Phase $\times$ Worker approached significance $(\mathrm{PST}, F(4,100)=17.3, p<.001 ; \operatorname{PST}, F(9$, $309)=1.69, p=.09$; PST, $\mathrm{F}(12,92)=1.51, p=.13)$. Changes over the phases were not significant.


Figure 6. Average ratings (with SE) in Experiment 4 for the four normal $(N)$ and altruist $(A)$ workers in test phases P1 to P4 of Conditions 1, 2, 3, and 4 (Panels A to D).

In the homogeneous local-and-global condition (C1, Panel A), the altruist was again evaluated as the worst, despite being most strongly correlated with high overall earnings. In an ANOVA for Phase 4, the within-subject factor Workers was clearly significant $(\operatorname{PST}, F(4,23)=10.1, p<.001)$, and contrasts confirmed that all normal workers were rated higher than the altruist (all, $p<.05$ ). In the heterogeneous Condition 3, participants were well able to differentiate between normal workers with different individual performance. A corresponding ANOVA showed a general effect of Workers (PST, $F(4,24)=38.8, p<.001)$ and significant contrasts between the workers in the predicted order, $\mathrm{N} 1>\mathrm{N} 2>\mathrm{N} 3>\mathrm{N} 4>\mathrm{A}($ each, $p<.001)$.
In the global-only conditions ( C 2 and C 4 ), in which people were to base their ratings of a worker's utility on the teams' overall earnings only, they clearly detected that, of all workers, the altruist correlated most demonstrably with high overall team earnings. Participants grasped this surprisingly early. An ANOVA for C2 (test phase 4) shows
significant results for the factor Workers (PST, $F(4,22)=$ 23.04, $p<.001$ ), and pairwise contrasts show that the altruist is rated higher than all normal workers (always with $p<.001$ ). In Condition 4, the order of the average ratings of the altruist and the normal workers was likewise reversed (relative to Condition 3). In an ANOVA a significant effect of the factor Workers was found (PST, $F(4,26)=15.5, p<$ .001 ); and contrasts show that the altruist was rated significantly higher than even the normal worker, who was rated highest ( $p<.001$ ). One ANOVA without the altruist reached significance $(\mathrm{PST}, F(4,27)=3.62, p<.05)$, but only one Bonferroni-corrected post hoc comparison between normal workers (the one expected to differ most: NW1NW4) led to significant results ( $p<.05$ ). In sum, despite clearly detecting that the altruist has a larger effect on overall output in the global-only conditions, participants show a reduced ability to distinguish between the normal workers.

Figure 6 shows the proportion of 'managers' choosing a worker to have the "highest" (Panel A) or "lowest" (Panel B) "total utility for the company" in the final test phase.


Figure 7. Percentage of participants choosing a normal worker ( $N$ ) or the altruist ( $A$ ) as of highest (Panel A) or lowest (Panel B) overall utility for the company in Phase 4 in Conditions C1, C2, C3, and C4. The choices corresponding to individual earnings are marked in black; those corresponding to overall earnings in dark gray.

Figure 7 presents the team selections in the personnel selection task. In Condition 1, we replicated a strong tendency to select a team without the overall best member, the altruist (from five possible configurations). Even in Phase 4 , after 80 rounds, $70 \%$ of the participants selected this team, $\chi^{2}(1, N=30)=46.88, p<.001$. Its reduction was not reliable, $\chi^{2}(1, N=60)=.80, p=.37$. By contrast, Condition 2 shows that participants provided with global-information-only were highly capable of quickly detecting that the altruist should be part of the team (Phase $1, \chi^{2}(1, N$ $=28)=7.00, p<.001)$. Also the contrast between Conditions 1 and 2 was highly significant (Phase 4$), \chi^{2}(1, N$ $=58)=27.15, p<.001$. In the heterogeneous global-andlocal Condition 3, selections began with a high proportion of no-altruist team-choices in Phase 1 (Figure 12), $\chi^{2}(1, N=$ $32)=67.57, p<.001$. In Phase 4, these individual-related selections, which exclude $A$, are likewise found to be above chance $(56 \%), \chi^{2}(1, N=32)=31.01, p<.001$; but now the group-related selections are above chance as well (excluding N4, with $34 \%) ; \chi^{2}(1, N=32)=4.13, p<.05$. By contrast, in

Condition 4 (a global-only condition), even in Phase 1 the optimal team-related selection (with-altruist team excluding N4) was the most frequently selected ( $43 \%$ ), $\chi^{2}(1, N=32)$ $=11.28, p<.001$; and the no-altruist team, conversely, was selected below chance ( $3 \%$ ), $\chi^{2}(1, N=32)=5.70, p<.05$. In Phase 4, the selection of the no-altruist team was still selected with low relative frequency ( $3 \%$ ), and the optimal team by $59 \%$ of participants.


Figure 8. The results of the personnel selection task in the four test phases show the proportion of 'managers' choosing a particular team, thus excluding either worker $N 1, N 2, N 3$, $N 4$, or the altruist $A$. In the global-and-local conditions, C 1 and C3, the black columns correspond to the predicted selections based on individual performance only. In the global-only conditions, C 2 and C 4 , no individual-level information was available. In all conditions, the dark gray columns represent the optimal selection(s) based on overall performance of teams.

In the local-and-global conditions we coded comments as insightful that showed understanding of the differences between an individual and a group level. After 80 learning rounds, at least $38 \%$ of the participants in these conditions were classified as providing comments with insight ( $33 \%$ in C1 and $43 \%$ in C3). Of these participants, $87 \%$ selected the altruist personnel selection task (in Phase 4), whereas from the participants not demonstrating insight only $3 \%$ made this selection. Finally, from the additional tests, only the Need-For-Cognition Scale (2.6 vs. 10.5), $t(60)=1.93, p=$ .03 (one-tailed) correlated with insightful comments.

## General Discussion

Experiment 1 shows that the tragedy of personnel selection is very stable across different team sizes. Even in the smallest team, most participants in the role of 'managers' evaluated the most useful worker for the group to be the worst. However, Experiment 1 suggests that small team sizes mitigates this problem, and a minority in this condition saw the difference between individual and overall contribution of an employee. Experiment 2 shows that
people are in principle well able to detect the strong correlation between presence of the altruist and high team performance (with $r=.99$ ) very early on from the grouplevel information. However, in other conditions most make no use of this ability and seem to ignore the overall payoff, focusing only on workers' direct individual contribution.

More generally, the findings may be due to people's problems dealing with decisions involving a Simpson's Paradox (Fiedler et al., 2003; Sydow et al., 2016; Waldmann \& Hagmayer, 2001). If people do not merely optimize in a standard decision-theoretic way (here by simply choosing the team with the highest past performance), and instead, as we suggest, aim for a deeper understanding by identifying clear causal or logical patterns between events (e.g., Funke, 2001; Hagmayer \& Meder, 2013; Osman, 2010; Sloman \& Hagmayer, 2006; von Sydow, 2016; Waldmann \& Hagmayer, 2001), this may yield the disadvantage of overlooking small correlations, pathways, exogeneities or interactions (Novick \& Cheng, 2004), even if they may add up, tragically, to be the predominant effect of a scenario.

## Acknowledgments

This work was partly supported by the grant Sy111/2-1 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program New Frameworks of Rationality (SPP 1516). We are grateful to Maria Waltmann and Martha Cunningham for help and corrections. The first author is grateful to Daniel Holt, Klaus Fielder and Michael Waldmann for discussing related issues. Furthermore, we would like to thank anonymous reviewers for helpful comments and suggestions.

## References

Brief, A. P. \& Motowidlo, St. J. (1986), Prosocial Organizational Behavior. Academy of Management Review, 11(4), 710-725.
Fiedler, K., Walther, E., Freytag, P., \& Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. Personality and Social Psychology Bulletin, 29, 14-27.
Grant, A. M. \& Patil, S. V. (2012). Challenging the norm of selfinterest. Minority influence and transitions to helping norms in work units. Academy of Management Review, 37(4), 547-588.
Engel, C. (2011). Dictator Games: A Meta Study. Experimental Economics, 14, 583-610.
Funke, J. (2001). Dynamic systems as tools for analyzing human judgment. Thinking and Reasoning, 7, 69-89.
Gollwitzer, M., Rothmund, T., Pfeiffer, A., \& Ensenbach, C. (2009). Why and when justice sensitivity leads to pro- and antisocial behavior. Journal of Research in Personality, 43(6), 999-1005.
Hagmayer, Y., \& Meder, B. (2013). Repeated causal decision making. Journal Of Experimental Psychology: Learning, Memory, And Cognition, 39, 33-50. doi: 10.1037/a0028643
Hardin, G. (1968). The Tragedy of the Commons. Science, 162 (3859), 1243-1248. doi:10.1126/science.162.3859.1243

Li, N., Kirkman, B. L., \& Porter, C. O. L. H. (2014). Toward a Model of Work Team Altruism. Academy of Management Review, 39(4),541-565. http://dx.doi.org/10.5465/amr.2011.0160

Mathieu, J., Maynard, M. T., Rapp, T., Gilson, L. (2008). Team effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future. Journal of Management, 34(3), 410-476. DOI: 10.1177/0149206308316061
Melis, A. P., Hare, B., Tomasello, M. (2006). Chimpanzees Recruit the Best Collaborators. Science, 311, 1297-1300.
Memmert, D., Plessner, H., Hüttermann, S., Froese, G., Peterhänsel, C., \& Unkelbach, C. (2015). Collective fit increases team performances: Extending regulatory fit from individuals to dyadic teams. Journal of Applied Social Psychology, 45, 274281. doi: 10.1111/jasp. 12294

Mojzisch, A., Grouneva, L., \& Schulz-Hardt, S. (2010). Biased evaluation of information during discussion: Disentangling the effects of preference consistency, social validation, and ownership of information. European Journal of Social Psychology, 40(6), 946-956. doi:10.1002/ejsp. 660
Nielsen, T. M., Hrivnak, G. A., \& Shaw, M. (2009). Organizational Citizenship Behaviour and Performance. A MetaAnalysis of Group-Level Research. Small Group Research, 40(5), 555-577.10.1177/1046496409339630
Novick, L. R., \& Cheng, P. W. (2004). Assessing Interactive Causal Influence. Psychological Review, 111, 455-485.
Nowak, M. A., \& Sigmund, K. (2005). Evolution of indirect reciprocity. Nature, 437/27, 1291-1296.
Organ, D. W. (1997). Organizational Citizenship Behaviour: It's Construct Clean-Up Time. Human Performance, 10(2), 85-97.
Osman, M. (2010) Controlling Uncertainty: A Review of Human Behavior in Complex Dynamic Environments. Psychological Bulletin, 136(1), 65-86.
Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., \& Mishra, P. (2010). Effects of organizational citizenship behaviors on selection decisions in employment interviews. Journal of Applied Psychology, 96 (2), 310-326.
Schulz-Hardt, S., \& Mojzisch, A. (2012). How to achieve synergy in group decision making: Lessons to be learned from the hidden profile paradigm. European Review of Social Psychology, 23(1), 305-343. doi:10.1080/10463283.2012.744440.
Sloman, St. \& Hagmayer, Y. (2006). The Psycho-Logic of Choice. Trends in Cognitive Science, 10(9), 407-411.
Sober, E., \& Wilson, D. (1999). Unto Others: The Evolution of Unselfish Behavior. Harvard University Press.
Van Scotter, J. R., Motowidlo, S. J., \& Cross, T. C. (2000). Effects of task performance and contextual performance on systemic rewards. Journal of Applied Psychology, 85(4), 526-535.
von Sydow, M. (2015). The Tragedy of Inner-Individual Dilemmas. In D. Noelle, et al. (Eds.), Proceedings of the ThirtySeventh Annual Conference of the Cognitive Science Society (pp. 2517-2522). Austin, TX: Cognitive Science Society.
von Sydow, M. (2016). Towards a Pattern-Based Logic of Probability Judgements and Logical Inclusion "Fallacies". Thinking \& Reasoning, 22(3), 297-335. doi:10.1080/13546783.
von Sydow, M., \& Braus, N. (2016). On the Tragedy of Personnel Evaluation. In A. Papafragou, et a. (Eds.), Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society (pp. 105-110). Austin, TX: Cognitive Science Society
von Sydow, M., Hagmayer, Y., \& Meder, B. (2016). Transitive reasoning distorts induction in causal chains. Memory \& Cognition, 44(3), 469-487. doi:10.3758/s13421-015-0568-5
Waldmann, M. R., \& Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. Cognition, 82, 27-58. doi: 10.1016/S0010-0277(01)00141-X
Wilson, D. S., \& Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. Quarterly Review of Biology, 82(4), 2007, 327-348. doi: 10.1086/522809

# Altruist vs Egoist Detection and Individual vs Group Selection in Personnel Management 

Momme von Sydow ${ }^{1,2}$ (momme.von-sydow@urz.uni-muenchen.de)<br>Niels Braus ${ }^{2}$ (n.braus@stud.uni-heidelberg.de)<br>${ }^{1}$ University of Munich (LMU), MCMP, Ludwigstr. 31, D-80539 München, Germany<br>${ }^{2}$ University of Heidelberg, Department of Psychology, Hauptstr. 47-51, D-69117 Heidelberg, Germany


#### Abstract

In the Wason-Selection Task debate it has been suggested that people may be able to detect cheaters but not co-operators or altruists. This position has been challenged. Here we focus on a scenario that is more ecologically valid with regard to different strategies for detecting workers who negatively interact with others (here 'egoists') and positive interactors (here 'altruist'). The results on altruist detection in two-level personnel evaluation tasks (T-PETs), with information on individual and team performance, suggested a disregard of the team performance and a resulting "Tragedy of Personnel Evaluation". Experiment 1 transfers the idea of altruist detection in a personnel evaluation and personnel selection task (von Sydow \& Braus, 2016) to egoist detection and explores whether there are analogous problems for egoist detection. Experiment 2 explores egoist and altruist detection in more realistic settings where individual and group-selection may affect our sampling of the interactor.


Keywords: Altruist/Egoist Detection; Wason Selection Task; Personal Selection Task; Tragedy of Personnel Selection; Group Selection; Learning Correlations; Decision Making

## Introduction

In an influential debate on hypothesis-testing (Wason Selection Tasks, WST), it has been suggested that in socialcontract situations people are adapted for cheater detection but not for co-operator or altruist detection (Cosmides, 1989). These proposals have contributed to differentiating between checking deontic rules and testing descriptive hypotheses (Oaksford \& Chater, 1994; Beller, 2001; von Sydow, 2006). Despite evidence for subclasses within the deontic domain (Fiddick, Cosmides, \& Tooby, 2000), other research shows that reasoning with standard deontic rules (including social contracts) seems to be a quite systematic faculty resembling deontic logic (Beller, 2001; Bucciarelli \& Johnson-Laird, 2005; von Sydow, 2006) and depends on the goals pursued (von Sydow, 2006; Rand, Dreber, Ellingsen, Fudenberg, \& Nowak, 2009; Sperber \& Girotto, 2002). However, the WST-paradigm has also been criticised as being too specific to address issues of real-life cooperation (Sperber \& Girotto, 2002).

Von Sydow \& Braus (2016) explored participants’ ability as personnel managers to detect how employees positively interacted with others' performance. Research in organizational and social psychology has acknowledged the importance of teams beyond mere individual contributions (Mathieu, Maynard, Rapp, \& Gilson, 2008; Memmert, Plessner, Hüttermann, Froese, Peterhänsel, \& Unkelbach, 2015) and the crucial role of prosocial or altruistic extra-role or role behaviours in teams (e.g., Li, Kirkman, \& Porter,
2014). Our tasks were more complex than WST and used two-level personnel selection tasks (T-PETs; in von Sydow \& Braus, 2016). In these T-PETs, participants obtained information about employees' performance on the direct individual level and on the overall group level. The presence of the altruist correlated consistently, reliably, and strongly with the teams' overall performance ( $r=.99$ ). Nonetheless, people tended to evaluate the altruist to be worst for the team, mostly based only on the individual information, and tended to ostracise him or her in selection tasks. This led us to suggest a potential "tragedy of personnel selection".

To explore the controversial asymmetry between altruist and egoist selection discussed in the WST literature in a more complex setting, and to explore the generality of the Tragedy of Personal Selection, Experiment 1 investigates TPETs not for altruist detection, but for egoist detection. Experiment 2 compares egoist and altruist detection in a single experiment and explores further potential factors of group vs. individual selection. This is broadly in line with the increasing influence of multi-level modelling in biology (Wilson \& Wilson, 2007), and personnel psychology (Polyhart, 2012). Furthermore, we explore the resultant effect of sampling (Fiedler, 2008). We will suggest that group selection could lead to greater altruist and egoist detection, but that this does not necessarily imply deeper understanding. Thus we suggest that in such a perhaps ecologically more valid scenario several further factors come into play over and beyond a mere potential difference between egoist and altruist detection.

## Experiment 1

The first experiment explores whether the Tragedy of Personnel Evaluation is unique to altruist detection, or whether there is an analogous phenomenon for egoist detection as well. Here 'egoists' have the highest individual earnings in the team while in fact most negatively affecting the team's overall performance. In this study participants were again acting as personnel managers, repeatedly making personnel evaluations and selections.

Table 1 shows the average earnings of the negative interactor, the egoist ( $E$ ), and, depending on the latter's presence or absence, the average earnings of the normal workers in the four conditions. The conditions vary homogeneous and heterogeneous earnings for the normal workers ( C 1 , normal worker with homogeneous earnings condition; C 4 , most heterogeneous earnings of the normal workers) to investigate participants' sensitivity to small differences in their impact on the individual level.

## Design

Table 1: Mean earnings of normal workers (NW: N1 to N4) and of 'egoist' worker ( $E$ ), overall earnings with or without the egoist in the four conditions ( C 1 to C 4 ), and resulting predictions

| resulting predictions |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| C 1 |  |  |  |  |
| C 2 | Predictions |  |  |  |
| Indivi | E>N1=N2 | E>N1>N2 | E>N1=N2 | E>N1>N2 |
| -dual | =N3=N4 | =N3=N4 | >N3=N4 | $>$ N3>N4 |
| Over- | N1=N2= | N1>N2= | N1=N2> | N1>N2> |
| all | N3=N4>E | N3=N4>E | N3=N4>E | N3>N4>E |


| Mean of earnings without egoist |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| N1 | 3000 | 3300 | 3400 | 2600 |
| N2 | 3000 | 2900 | 2400 | 3200 |
| N3 | 3000 | 2900 | 2600 | 2800 |
| N4 | 3000 | 2900 | 2600 | 2400 |
| Mean of earnings with egoist |  |  |  |  |
| N1 | 2000 | 2300 | 2400 | 2600 |
| N2 | 2000 | 1900 | 2400 | 2200 |
| N3 | 2000 | 1900 | 1600 | 1800 |
| N4 | 2000 | 1900 | 1600 | 1400 |
| E | 3400 | 3400 | 3400 | 3400 |


| Mean of overall earnings of a group |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| With- <br> out E | 12000 | 12000 | 12000 | 12000 |
| With <br> E | 9400 | 9400 | 9400 | 9400 |

## Method

Participants 161 participants from MTURK passed a first participation-criterion (time spent on the first page $>20 \mathrm{sec}$. and < 6 min .) and began the task. 120 participants finished the experiment and were included in the analysis ( $52 \%$ male; mean age 33), most of them with a high school or even a university degree (59\% Bachelor's or Master's; $38 \%$ high school). The volunteers obtained rewards of $\$ 1$. Participants were randomly assigned to one of four conditions (cf. Table 1).
Material and procedure The crucial difference to prior work was that we replaced altruist detection by egoist detection (Table 1). Apart from changed individual and overall earnings, the scenario, T-PET procedure, and dependent variables (von Sydow \& Braus, 2016). But we were now concerned with individually best performing 'egoists', whose presence correlated consistently and most negatively with the team's overall performance. As in the altruist detection task, participants in each round obtained overview information in tables about workers' individual earnings, together with their photographs and information
about overall earnings of the team. The presentation order of the pictures was randomized. Again there were only five workers, with four workers per shift - thus only five possible team configurations. There were 40 rounds and four test phases, one after every ten rounds. The first three test phases included rating-tasks and a team selection task only; in the final test phase we asked participants additionally to choose the employee of highest and lowest utility, and to comment on the task and their decision.

## Results

In all conditions, the average ratings (Figure 1) resemble more closely the predictions based on individual rather than overall team-contributions (cf. Table 1). An ANOVA with the between-subjects factor Conditions and the withinsubjects factors Workers and Phases (in a multivariate Pillai-Spur Test, PST) showed significant effects of Workers, $F(4,110)=95.9, p<.001$, Workers $\times$ Conditions, $F(12,336)=21.9, p<.001$, and Phases, $F(3,111)=3.41, p$ $<.05$, and a marginally significant effect of Phase $\times$ Person, Workers $\times$ Conditions, $F(12,102)=1.78, p=.06$.


Figure 1: Average ratings (with SE) in Experiment 1 for the four normal workers $(N)$ and egoist worker $(E)$ in the test phases P1 to P4 of Conditions 1, 2, 3, and 4
(Panels A to D)
In Condition 1, the egoist is characterized by higher ratings than the other four workers throughout all phases; main effect of Worker: $F(4,26)=4.19, p<.01$. In Condition 2, there was again only a reliable effect of Workers, PST, $F(4,24)=25.2, p<.001$. Bonferronicorrected post hoc comparisons showed no differences between the egoist and normal worker 1 (E, N1), predicted to be the highest normal worker ( $p=1.00$ ); but N 1 , as predicted, had a higher rating than the other normal workers, which was also the case for the egoist (all, $p<.001$ ). In Condition 3, again only the factor Workers was significant
(PST, $F(4,22)=24.1, p<.001)$. Bonferroni-corrected post hoc comparisons showed that the egoist was not rated higher than the normal workers predicted to be highest ( N 1 , N 2 ) (both $p=1.00$ ), but that the worker in this group, as well as the egoist, reliably differed from the workers in the second group of normal workers (N3, N4; all $p<.001$ ). Condition 4 again showed an overall effect only for the factor Worker (PST, $F(4,28)=58.6, p<.001$ ), and in corrected post hoc comparisons significant effects even of all five workers in the order predicted by the individual earnings (all $p<.01$ ).

The results of the Personnel Selection Task (Figure 2) show that the majority selected teams with optimal earnings on the individual level (black; individual-related selections). Only a few selected the team without the egoist (from five possible teams), even though this team had the best overall performance (dark gray; team-related selections). The remaining selections (light gray) selected the egoist for the team, along with other, individually non-optimal workers. With regard to temporal changes, there is an apparent increase in the proportion of team-related selections (dark gray) from Phases $1(9 \%)$ to $4(23 \%) ; \chi^{2}(1, N=240)=8.00$, $p<.01$. But even in the final test phase, Phase 4, the individual-related selections over all conditions occurred more frequently than the team-related ones, $\chi^{2}(1, N=120)=$ $34.7, p<.001$.
The highest-utility task (Figure 3, Panel A) reveals relatively frequent 'egoist'-judgments (black). In all conditions these judgments were clearly above chance level $\left(\chi^{2}(1, N=31)=63.9, p<.001 ; \chi^{2}(1, N=30)=44.5, p<\right.$ $.001 ; \chi^{2}(1, N=27)=15.8, p<.001 ; \chi^{2}(1, N=32)=47.5, p$ $<.001$ ). Considering the team-related judgments (dark gray), they are also above chance level relative to the other ones (light gray), $\chi^{2}(1, N=31)=49.0, p<.001(\mathrm{C} 2$ to C 4$)$.


Figure 2. Results of the personnel selection task in the four test phases of Experiment 1, showing the proportion of 'managers' choosing a team of four out of five, thus excluding worker $N 1, N 2, N 3, N 4$, or the egoist worker $E$. Individual-related optimal selections are marked in black,
with team-related optimal selections in dark gray and other selections in light gray.

In the lowest-utility task (Fig. 3B), a similar pattern can be recognized. The individual-related selections (black) were chosen more often than chance level, $\chi^{2}(1, N=120)=$ $300.8, p<.001$. In the three conditions ( C 2 to C 4 ), where one can contrast the team-related judgments (egoist has the lowest utility; dark gray) with judgments that were neither individually nor on group-level optimal (light gray), the team-related judgments overall occurred reliably more often than expected by chance (exact bin. test, $N=12, p<.001$ ).
As to the comments, $21 \%$ participants mentioned explicitly that the individual and overall group-level contributions of a worker differ, or that there are interactions between participants. These insightful comments were highly associated with group-level selections and ratings.


Figure 3. Percentage of 'managers' choosing either a normal worker $(N)$ or the egoist worker $(E)$ as of the highest (Panel A) or lowest (Panel B) utility for the company (Conditions C1, C2, C3, C4). The individual-related choices are marked in black, the team-related ones in dark gray and the neither-individual-nor-team-related ones in light gray.

## Discussion

The results of Experiment 1 show that egoist detection seems to be affected by similar problems as altruist detection (von Sydow \& Braus, 2016). When the negative interactor individually contributed the highest earnings but led overall to the lowest group earnings, the majority of participants nonetheless rated the egoist as most valuable. Moreover, in the personnel selection task they even systematically chose the egoist for the team, even though the latter consistently performed the worst. This was the case even though participants were sensitive to relatively small individual differences. Thus the results suggest a kind of tragedy of personnel selection as well with regard to 'egoists' or negative interactors.
In comparison to the results on altruist detection by von Sydow \& Braus (2016), which seem comparable in population and method, there is perhaps a slight advantage of egoist over altruist detection. In particular, there were significantly more insightful comments for egoist than for altruist detection, $\chi^{2}(1, N=240)=7.53, p<.01$. This suggests that the content of egoist versus altruist detection leads to different interaction detection rates. However, the overall findings point rather to the similarity between altruist and
egoist detection. There also seems to be a Tragedy of Personnel Evaluation with regard to egoist detection.

## Experiment 2

This experiment explores the potential effect of personnel selection on the individual versus group level and resulting distortions linked to sampling only particular information (thus favoring learning of specific relationships or not). We combine these issues with an investigation of egoist and altruist detection in a single experiment. We propose the hypothesis that group in contrast to individual selection may at least behaviourally increase selections corresponding to the group-level performance. However, deeper insight is predicted to depend on sampling effects that may affect both individual and group selection. To test the effect of sampling, the participants were given the opportunity to influence the material by their selections. Sampling may imply a greater inclusion of the interactor, which will lead to a better learning of the interactor's impact either on the individual or on the group level and hence to a solution of the Tragedy of Personnel Selection. In contrast, if sampling leads to exclusion of the interactor, a correct selection needs not to improve the understanding of the interactor's impact.

## Design

Experiment 2 investigates how individual versus group selection and the presence of an 'altruist' versus an 'egoist' influence personnel selection and evaluations (Table 2). Apart from this, the conditions used almost identical scenarios and the interactors had comparable effects on the group level. The egoist is characterized by the highest and the altruist by the lowest individual earning. In contrast to the individual earnings, the presence of the egoist leads to lower overall earnings of the group, whereas that of the altruist yields greater overall earnings for the group.

Table 2: Selection and interactor conditions and mean of earnings for normal workers (NW), altruist and overall for
four conditions.

| Cour conditions. |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| Selection type | Individual | Individual | Group | Group |
| Interactor type | Altruist | Egoist | Altruist | Egoist |
| Mean of earnings |  |  |  |  |
| NW with <br> altruist/egoist | 3000 | 1500 | 3000 | 1500 |
| NW without <br> altruist/egoist | 2000 | 2500 | 2000 | 2500 |
| Altruist/egoist | 1500 | 3000 | 1500 | 3000 |
|  | Mean of overall earnings of a group |  |  |  |
| With <br> altruist/egoist | 7500 | 6000 | 7500 | 6000 |
| Without <br> altruist/egoist | 6000 | 7500 | 6000 | 7500 |

## Method

Participants 182 participants from MTURK began the task. 119 participants passed the strict selection criteria (time spent on first page; correct rephrasing the instructions), finished the task, and were included in the analysis ( $51 \%$ male; mean age 35); most of them had a degree ( $65 \%$ Bachelor's or Master's; 35\% high school). The volunteers obtained rewards of $\$ 1$. Participants were randomly assigned to one of four conditions (cf. Table 2).

Material and procedure The material and procedure build on previous T-PETs but vary in some aspects. The data is again shown for each day or shift, containing individual and overall earnings. Whereas previous T-PETs involved one group (with different member configurations) only, here we also presented two groups, each with 3 employees. During the experiment, participants received data on a total of 10 employees and 81 days (rounds). Moreover, in contrast to former T-PETs the participants’ judgments in personnel selection tasks influenced the materials presented to the participants. This task was repeated 11 times. On the first day of each of the 10 test phases (and in a final test phase), the altruist/egoist and 5 randomly assigned normal workers were presented. Based on the selection in the test phase, participants selected 3 out of 6 employees (individual conditions) or one of two groups (group conditions). This selected group or these individuals were excluded from the following 7 days (rounds). From the 7 employees left, we created the two groups in the successive 7 days, so that each one of the employees left was excluded for one day. Although we added some noise ( $\mathrm{SD}=600 €$ ) to the individual earnings, the presence of the interactor in such trials still strongly correlated with a higher overall outcome ( $r=.95-.98$ ). In every second test phase (starting with the second phase), the participants additionally had to rate the 10 employees before selection. On the last day, at test phase 11, the participants had to complete a rating and selection task (here without consequences) as well as the utility-task (cf. Experiment 1) and a Need-For-Cognition task (NFC; Cacioppo, Petty \& Kao, 1984); and finally to comment on the task.

## Predictions

- Hypothesis 1 (H1): Participants may base their selections (partly) on the observed performance in the test phases, with the selection focusing them either on the group or the individual level. This would entail more selections optimal on the group level in the group condition than in the individual conditions: $\mathrm{C} 3=\mathrm{C} 4>$ $\mathrm{C} 2=\mathrm{C} 1$.
- However, participants in line with H1 will tend to exclude the interactor in C 4 (group selection) and C 1 (individual selection) and tend to include the interactor in C3 (group selection) and C2 (individual selection). Based on these expected sampling effects, one may derive the following hypothesis:
o Hypothesis 2 (H2): People in the inclusion conditions will learn more about the individual earnings, which vice-versa would lower judgments in line with adequate group-level predictions in the inclusion conditions: $\mathrm{C} 1=\mathrm{C} 4>\mathrm{C} 2=\mathrm{C} 3$.
o Hypothesis 3 (H3): Alternatively (or additionally), in the inclusion conditions people may start to realize the major impact of the interactor on group earnings, with inverse implications for group-level results: $\mathrm{C} 1=\mathrm{C} 4<\mathrm{C} 2=\mathrm{C} 3$.
H2 and H3 may both apply and cancel each other out. Alternatively one may find different effects over time (first H 2 then H 3 ). We predict that the hypotheses will have a major impact on different dependent variables: H1 may particularly affect the used selection task which may be dominated by the currently shown data. H2 or H3 may dominate the overall results in the rating and utility tasks. With regard to insight in the two-level nature of the task in the comments seems normally to be limited by realizing the group level effects; thus insight should follow the pattern of H3.


## Results

As expected, different clusters of dependent variables referring to performance effects H1 (here selection tasks) and to effects of understanding, H2 versus H3 (evaluation tasks) reveal differential results: Considering the personnel selection task, compared to individual selection (Figure 5, Panel A), the group selection as expected (H1) appears to lead to greater (overall optimal) exclusion of the egoist and inclusion of the altruist. However, the results descriptively show optimal group-level answers in the rank order of $\mathrm{C} 4>$ $\mathrm{C} 3 \gg \mathrm{C} 1>\mathrm{C} 2$. This suggests a main effect in line with H1 and an additional effect, even for this selection variable, in line with H2. Inferentially, the aggregated correct selections of each participant over the 11 selections revealed a strong effect of Selection type $\left(F(1,115)=119.60, p<.001, \eta^{2}=.52\right)$ and a slight interaction effect of Selection and Interactor types, $F(1,115)=9.17, \quad p=.003, \eta^{2}=.07$. As a high correlation between the accuracy of selection and amount of earnings suggests, the earnings replicate the strong positive effect of Selection type, $F(1,115)=128.18, p<.001, \eta^{2}=.52$, Interactor type: $F(1,115)=11.10, p=.001, \eta^{2}=.09$, and Phase: $F(6,687)=2.29, p<.05, \eta^{2}=.02$.

The rating task (Figure 4) seems roughly to reflect the individual pay-off structure affected by sampling (H2). The altruist has lower ratings than the other workers, and the egoist is rated more positively. Comparing the ratings of the interactor to all normal workers, the interactor differs highly significantly across all 5 measurements (R1-R5), R1: $F(1,116)=1345.68, p<.001$; R2: $F(1,116)=1369.64, p<$ .001 ; R3: $F(1,116)=1540.53, p<.001$; R4: $F(1,116)=$ 2028.63, $p<.001$; R5: $F(1,116)=2731.81, p<.001$.

The highest-utility and lowest-utility tasks (Figure 5, Panel B) seem to mirror mainly the individual earnings of the employees (high rate of egoists in the highest utility task, and of altruist in the lowest utility task). Thus the Interactor type affects the accuracy of this task (highest:
$\chi^{2}(1,119)=11,84, \quad p=.001$; lowest: $\quad \chi^{2}(1,119)=19,02$, $p<.001$ ); but the selection type had no impact.


Figure 4: Average ratings (with SE) for the normal workers (N1-9) and the altruist/egoist worker ( $A, E$ ) of Conditions C1, C2, C3 and C4 (Panels A to D).

Insightful comments (Panel C), in line with H3, showed in C 2 and C 3 marginally significantly higher correct comments than in C 1 and $\mathrm{C} 4, \chi^{2}(1,119)=3,55, p=.06$. Participants, commenting correctly, revealed greater aggregated correct selections $(t(51)=-2,28, p<.01)$ and ratings $(t(56)=-$ $2,80, p<.01$ ), and tended to choose the person with the highest or lowest overall utility more frequently, highest: $\chi^{2}(1,119)=15,77, p<.001$; lowest: $\chi^{2}(1,119)=3,05, p=$ .08. Participants with insightful comments can be characterized by higher NFC-Scores: $t(117)=-2,65, p<.01$.


Figure 5: Average correct answer rate for Conditions 1, 2, 3 and 4 (C1-C4) of the Selection, Utility- and Comment Tasks (Panel A to C).

## Discussion

Experiment 2 shows intricate influences of sampling on the egoist and altruist detection in group and individual
selection scenarios. Group selection, at least on a direct performance level in the selection tasks, leads to greater overall optimal selections compared to individual selection (H1). In addition, group selection increases economic outcomes. Considering the accuracy of personnel evaluation, the results create a more complex pattern. Egoist versus altruist detection does not have a great impact; neither does Individual vs. Group-Selection. Sampling processes seem to matter and can have simultaneous opposed effects. The ratings and the highest/lowest utility-tasks show that gaining more information about the interactor leads to stronger individual-based understanding (H2). Insight in the comments nonetheless revealed, as predicted, that gathering information about the interactor also increased the detection of group level effects (H3). In line with this finding, insightful comments were associated with the NFC-Score.

## General Discussion

Experiment 1 shows that egoist detection, as with similar altruist-detection tasks, may systematically lead to judging the egoist as best for a company although he clearly correlates strongly with negative overall team performance. Although the results suggest a slight advantage of egoist over altruist detection, both show basically similar results, with participants in both scenarios falling prey to a Tragedy of Personnel Selection. In Experiment 2, instead of egoist versus altruist detection, other factors such as group versus individual selection (with group selection improving performance) and sampling processes (in different ways affecting understanding on individual and group levels) more strongly influenced participants' judgments.

With regard to the personnel selection literature (e.g., Polyhart, 2012; Li, Kirkman, \& Porter, 2014), the results warn us against the generality of the suggested Tragedy of Personnel Selection shown here to affect not only altruist but also egoist detection. With regard to the Wason Selection Task debate (e.g., von Sydow, 2016; Sperber \& Girotto, 2002), we found no large differences between altruist and egoistic detection in the T-PETs (however, a small one). The results more generally pose the question as to how far the difficulties detecting the strongest overall correlation of variables' presence with overall outcome points to problems linked to Simpson's Paradox (Fiedler et al., 2003; Sydow et al., 2016; Waldmann \& Hagmayer, 2001); also whether it is a negative side effect of people's constructing detailed logical or causal models over and above optimizing observed utilities (e.g., Funke, 2001; Hagmayer \& Meder, 2013; Osman, 2010; Sloman \& Hagmayer, 2006; von Sydow, 2016; Waldmann \& Hagmayer, 2001), with the disadvantage that one tends to neglect small correlations, pathways, exogeneities or interactions, even if they tragically dominate a scenario.

## Acknowledgments

This work was partly supported by the Grant Sy111/2-1 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program New Frameworks of Rationality
(SPP 1516). We are grateful to Maria Waltmann, who helped prepare the material, to Martha Cunningham for corrections, and to Klaus Fiedler, Daniel Holt, and Max Ihmels for discussing related issues.

## References

Beller, S. (2001). A model theory of deontic reasoning about social norms. In J. D. Moore \& K. Stenning (Eds.), Proceedings of the 23rd Annual Conference of the Cognitive Science Society (pp. 63-68). Mahwah, NJ: Lawrence Erlbaum.
Bucciarelli, M., \& Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. Cognitive Psychology, 50, 159-193.Cacioppo, J. T., Petty, R. E., \& Kao, C. F., (1984). The Efficient Assessment of Need for Cognition. Journal of Personality Assessment, 48, 306-307.
Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. Cognition, 31(3), 187-276.
Fiddick, L., Cosmides, L., \& Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. Cognition, 77, 1-79.
Fiedler, K. (2008). The Ultimate Sampling Dilemma in Experience-Based Decision Making. Journal of Experimental Psychology, Learning, Memory, \& Cognition, 34(1), 186-203.
Li, N., Kirkman, B. L., \& Porter, C. O. L. H. (2014). Toward a Model of Work Team Altruism. Academy of Management Review, 39(4),541565.

Mathieu, J., Maynard, M. T., Rapp, T., Gilson, L. (2008). Team effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future. Journal of Management, 34(3), 410-476. doi: 10.1177/0149206308316061
Memmert, D., Plessner, H., Hüttermann, S., Froese, G., Peterhänsel, C., \& Unkelbach, C. (2015). Collective fit increases team performances: Extending regulatory fit from individuals to dyadic teams. Journal of Applied Social Psychology, 45, 274-281.
Oaksford, M., \& Chater, N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review, 101, 608-631.
Osman, M. (2010). Controlling Uncertainty: A Review of Human Behavior in Complex Dynamic Environments. Psychological Bulletin, 136(1), 65-86.
Ployhart, R. E. (2012). Multilevel selection and the paradox of sustained competitive advantage. In N. Schmitt \& N. (Ed) Schmitt (Eds.), The Oxford handbook of personnel assessment and selection. (pp. 667685). New York, NY, US: Oxford University Press.

Rand, D., Dreber, A., Ellingsen, T., Fudenberg, D., \& Nowak, M. (2009). Positive interactions promote public cooperation. Science, 325(5945), 1272-1275.
Sperber, D. \& Girotto, V. (2002). Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides, and Tooby. Cognition, 85, 277290.

Van Scotter, J. R., \& Motowidlo, S. J. (1996). Interpersonal Facilitation and Job Dediction as Separate Facets of Contextual Performance. Journal of Applied Psychology, 81(5), 525-531.
von Sydow, M., \& Braus, N. (2016). On the Tragedy of Personnel Evaluation. In A. Papafragou, et al. (Eds.), Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society (pp. 105-110). Austin, TX: Cognitive Science Society.
von Sydow, M. (2006). Towards a Flexible Bayesian and Deontic Logic of Testing Descriptive and Prescriptive Rules. Dissertation, Georg-August-Universität Göttingen.
von Sydow, M. (2016). Towards a Pattern-Based Logic of Probability Judgements and Logical Inclusion "Fallacies". Thinking \& Reasoning, 22(3), 297-335. doi:10.1080/13546783
Waldmann, M. R., \& Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. Cognition, 82, 27-58. doi: 10.1016/S0010-0277(01)00141-X
Wilson, D. S., \& Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. Quarterly Review of Biology, 82(4), 2007, 327-348. doi: 10.1086/522809

# Rational and Semi-Rational Explanations of the Conjunction Fallacy: A Polycausal Approach 

Momme von Sydow (momme.von-sydow@urz.uni-münchen.de)<br>University of Munich (LMU), Munich Center for Mathematical Philosophy (MCMP),<br>Ludwigstr. 31, D-80539 München, Germany


#### Abstract

Conjunction fallacies (CF) have not only been a major obstacle in justifying the rationality of a Bayesian theory of belief update; they have also inspired a variety of theories on probability judgment and logical predication. Here we provide an overview of Bayesian logic (BL) as rational formulation of a pattern-based class of conjunction fallacies. BL is described here as a generalization of Bayesian Occam's razor. BL captures the idea that probabilities are sometimes used not extensionally but intensionally, determining the probabilistic adequacy of ideal logical patterns. It is emphasized that BL is a class of models that depend on representations and the meanings of logical connectives. We discuss open questions and limits of BL. We also briefly discuss whether other theories of the CF may be good supplementary theories of CFs (and predication) as well, if linked to functional explanations.


Keywords: probability judgments; biases; conjunction fallacy; inclusion fallacy; inductive logics; intensional logics; Bayesian logics; predication; strong sampling; categories; Lockean Thesis; rationality debate; Bayesian Occam's razor

## Extensional vs. Intensional Probabilities

## Extension vs. Intension



Figure 1: (A) Extensions as elements and intensions demarcated by set boundaries. (B) Characterization of extensional and (C) intensional probabilities (cf. McKay, 2003, 28).

Although less known to the psychologist than to the philosopher or logician, the notions of extension and intension ( $\neq$ "intention") have a tradition going back to Leibniz, Carnap, and Stegmüller; with several analogous terms proposed by others, such as 'meaning' and 'denotation' (Russell) or 'Sinn' and 'Bedeutung' (Frege). Extension refers to the elements of a set, and intension to the meaning, which may be symbolized by the area determined by the set boundaries (Figure 1A). Correspondingly, a set can be described extensionally by specifying its elements or intensionally by specifying one or several defining features.

## Extensional narrow norms of predication

Extensional approaches have long dominated set theory (Zermelo-Fraenkel set theory), logic (propositional logic),
and probability theory (Kolmogorov's axiomatization). According to these extensional approaches, two sets are identical if they have the same elements (the same extension). Such an approach entails that any logically stronger (more specific) proposition implies any more general proposition (e.g., $A \wedge B \Rightarrow A \vee B$ ); and that any more specific hypothesis can never be more probable than a more general one.

Extensional logic and extensional probabilities have been proposed to provide universal criteria of rational predication. Predication attributes a predicate or logical combination of predicates to a subject. First, valid (assertive) predication of general logical relationships has traditionally often been linked to a logical truth-table definition of connectives (Frege, Russell, Whitehead, and Wittgenstein) (Table 1). Accordingly, if a conjunction $A \wedge B$ is true in a universe of discourse $(X)$, no cases in $X$ fall outside of the corresponding set (the intersection) and the truth of the conjunction implies, for instance, the truth of the affirmation $A$ as well as of the disjunction $A \vee B$ (Table 1).

Table 1: Truth tables of some dyadic logical connectives: conjunctions, exclusive disjunctions, affirmations, and inclusive disjunctions

| A | B | A $\wedge B$ | $A><B$ | $A$ | $A \vee B$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $T$ | T | T | F | T | T |
| T | F | F | T | T | T |
| F | T | F | T | F | T |
| F | F | F | F | F | F |

Assertive, contingent sentences, such as "Members of Species $X$ are aggressive ( $A$ ) AND curious ( $B$ )" (von Sydow $\&$ Fiedler, 2012) (in predicate logic: $\forall \mathrm{x} A(x) \wedge B(x)$ ), can thus be falsified by a single observation (Popper, 1934; but see Oaksford \& Chater, 2007). One problem in using (extensional) logic as adequacy criterion of contingent general predications is that they often involve exceptions (von Sydow, 2013). The problem of exception refers to the phenomenon that we seem to employ general sentences, like ravens are black, even if there are known exceptions, such as albino ravens. One solution to this problem is to replace the logical criterion of adequate predication by a high-probability criterion $(P$ (Assertion) $>\varphi>.5$ ) (Schurz, 2001; cf. Adams, 1986; Oaksford \& Chater, 2007, Pfeiffer, 2013; BL makes use of this idea, but for intensional probabilities).

Second, Kahneman \& Tversky (1983) argued that any deviation from the probabilistic extensional conjunction rule, $P(A \wedge B) \leq P(A)$, involves a conjunction 'fallacy', even in the context of predication. It is argued that extensional logic and extensional probability are narrow norms (cf. Gigerenzer, 1996; Fiedler \& von Sydow, 2013) if applied as
adequacy criterion for rational predication (von Sydow, 2011, 2016). One problem of a direct application of an extensional probability criterion to predication is sample size. That is, (extensional) relative frequencies do not distinguish between $1 / 1$ and 1000/1000 confirmative ravens.

However, the main problem with extensional probability is that predicates referring to subsets can never yield a higher probability than those referring to supersets. However, it should be possible do deem more specific hypotheses to be more adequate; otherwise one could never prefer a more specific hypothesis. And it seems absurd, if, for instance, " $X$ are aggressive AND curious" $(A \wedge B)$ could never be more probable and hence more adequate than " $X$ are aggressive OR curious or both" $(A \vee B)$. Likewise, the tautology (or 'verum') " $X$ are aggressive or not, and they are curious or not" (ATB) by definition (with a maximal extensional probability of 1) could never be less adequate than a more suitable, specific hypothesis, independent of empirical evidence. Therefore a universal extensional high-probability criterion fails the requirement to be empirically informative.

## Intensional Probabilities of Bayesian Logic

## Bayesian Occam's Razor

The basic idea of a Bayesian Occam's razor already offers a partial solution to the problem of inclusion (Jeffreys \& Berger, 1992; Tenenbaum \& Griffiths, 2001; McKay, 2003; cf. Navarro et al., 2012). If a consequential region of hypotheses H1 is a subset of a consequential region of a hypothesis H2 (Figure 1C), even a subjective Bayesian account may be extensional in the sense of requiring that the more specific hypothesis can never be more probable than a more general one (this is sometimes called 'weak sampling'). However, if one treats the nested hypotheses nonetheless as alternative explanatory patterns (hypotheses) whose consequence regions may each have produced the data, the size of the consequential regions matters (sometimes called 'strong sampling'). In this case, data coherent with the specific (and also the general) hypothesis (cf. Figure 1C) is more likely to occur based on the more specific hypothesis: $P($ data $\mid \mathrm{H} 1)>$ $P$ (data|H2). If one additionally assigns an equal prior to these alternative hypotheses, $P(\mathrm{H} 1)=P(\mathrm{H} 2)$, one can indeed get a higher posterior for the more specific hypothesis, $P(\mathrm{H} 1)>P(\mathrm{H} 2)$. For extensional probabilities, by contrast, a more specific hypothesis could never obtain a higher probability than a more general one (inclusion rule). We even treat this rule as the defining feature of extensional probabilities. Since Bayesian Occam's razor (strong sampling) violates this rule, we may thus call it an 'intensional' probability. For intensional probabilities, actually both the size of the extension and that of the intension matter.

However, such a basic application of Bayesian Occam's razor does not allow for exceptions. A hypothesis is still falsified by a single disconfirmatory instance. But predications about contingent facts often allow for exceptions (von Sydow, 2013b). Otherwise one still could never prefer a more specific predication over a tautology. This problem of
applying basic Bayesian Occam's razor (without noise) to real predications presumably explains why the predominant view holds that Bayesian accounts cannot rationally account for CFs (Fisk, 1996; Gigerenzer, 1998; Neace et al., 2008).

## BL as Generalized Bayesian Occam's Razor

Bayesian logic (BL, von Sydow, 2011, 2016; cf. von Sydow, 2013; von Sydow \& Fiedler, 2012) addresses the problem of exception together with the problem of inclusion. Thus BL can be understood as using the 'natural' implications of Bayesian Occam's Razor together with the assumption that people are not interested in deterministic logical hypotheses but rather in noisy-logical hypotheses that are still similar to the deterministic hypotheses.


Figure 2: BL model sketch in 5 steps (cf. main text)
Step 1 of the model (Figure 2) turns deterministic truth tables into ideal (and still in some sense deterministic) probability tables (PTs). In the PTs above dark shadings represent a high probability (up to 1 ) and light shadings a low probability (down to 0 ). Step 1 only constructs PTs with noise level $r=0$. Here, cells of a PT that correspond to logically false truth values $(F)$ are assigned the value 0 . Logically confirmatory cells ( $T$ ) in these ideal representations, however, are assumed to be equi-probable (Johnson-Laird, Legrenzi, Girotto, Legrenzi, \& Caverni, 1999). Step 1 (combined with Steps 3 and 4) already provides a basic Occam's razor solution. However, here, connectives are still falsified by single disconfirmatory events.

Step 2 addresses the problem of exceptions and constructs ideal noisy PTs (with $r>0$ ) by adding noise to each cell of the ideal patterns from Step 1 and then renormalizing the PTs. There may be alternatives for modelling noise/acceptance levels (e.g., von Sydow, 2014), but as long as they yield similar results and address the problems of inclusion and exception simultaneously this seems a predominantly technical issue, and they should be seen as variants of the same computation model class of BL. Note that the PTs here are still ideal explanatory hypotheses composed of four cell probabilities adding up to 1 and a (second order) probability representing the belief in this PThypothesis (also adding up to 1 , but now over all PThypotheses). We normally assume a flat prior distribution
over all PTs (implying a flat prior distribution for both connectives and noise levels).

Step 3. The likelihood of each explanatory $2 \times 2$ PT (for $i \times j$ PTs, with $i$ modelled connectives and $j$ modelled equidistant noise levels) given the data, $P(\mathrm{PT} \mid D)$, can be calculated by a multinomial distribution. The data are i.i.d. observations in a $2 \times 2$ contingency matrix ( $A$ vs. non $-A, B$ vs. non- $B$ ).

Step 4 uses Bayes' theorem to derive the posterior probabilities $P_{I}(P T \mid D)$ from the likelihoods $P(D \mid P T)$ and priors.

Step 5 sums up the posterior probabilities of all PTs created based on a particular logical hypothesis (over all $j$ noise levels). This results in (intensional and noise-tolerant) posterior probabilities for each of the $i$ connectives.


Figure 3. Intensional probabilities of logical hypotheses for 11 noise levels for a given $2 \times 2$ data contingency matrix with the cells a, b, c, d, here [7, 3, 2, 0], and a flat prior.

Figure 3 shows resulting intensional probabilities for all modeled connectives at different noise levels (Step 4). If one assumes no additional weighting for particular noiselevels (Step 5) the marginal probabilities provide the intensional posteriors for the connectives. In the example, $P_{i}(A \wedge B), P_{i}(A)$, and $P_{i}(A \vee B)$ have the highest overall probability. The intensional probabilities entail $P_{i}(A \wedge B)>$ $P_{i}(B)$, although the extensional conjunction rule requires $P_{E}(A \wedge B)>P_{E}(B)$. ' $A$ or $B$ (or both)' $(A \vee B)$ is (intensionally) most probable at low levels of noise, but ' $A$ and $B$ ' is most likely at higher levels of noise. Thus noise priors - for instance a belief in deterministic relationships - may change the intensional probability assessment.

## Findings Corroborating Standard Dyadic BL

Bayesian Logic, in its outlined main version, is an inductive logic providing intensional probabilities of dyadic logical connectives. The connectives relate two dichotomous events, and the model input are priors or frequencies (or equivalent Dirichlet-distributed, degrees of belief). BL provides a rational reconstruction of a class of pattern-based conjunction fallacies in line with Bayesian updating (cf. Hartmann, \& Meijs, 2012, for another rational Bayesian model of CFs, based on source reliability). BL in a way
detects the noisy-logical pattern that is most 'similar' to the data (actually $P(P T \mid D)$ ).
Some comments may be appropriate: First, if participants must rank the probability of two nested logical connectives (e.g., a conjunction and one of its conjuncts; Tversky \& Kahneman, 1983), using intensional probabilities, $P_{I}$, instead of extensional ones, $P_{E}$, is not fallacious; both are probabilities. We here continue to speak of conjunction 'fallacies' only for reasons of convenience. If the previous argument is correct, it is even reasonable, in the context of predication and looking for the most adequate connective, to apply intensional probability, since it serves a reasonable function (providing an empirically informative probabilistic adequacy criterion for predication). Second, the intensional probabilities only supplement extensional probabilities. BL is likewise based on the standard extensional axioms of probability (Kolmogorov's axioms), but applies these axioms not on the level of extensions, but on that of probabilities of alternative logical hypotheses. (Similarly, in Bayes nets one may apply hypotheses probabilities to graphs without invalidating the underlying joint probability matrix. ${ }^{1}$ ) Third, BL is formulated not only as a normative but also as a descriptive (computational-level) theory of probability judgments concerning logical predications. However, the claim is, of course, not that people have a deliberate analytic understanding of BL. Their judgments may be roughly reasonable, as our perception system makes reasonable inferences without requiring conscious calculations. Moreover, people may merely be using something similar to intensional probabilities; and it needs to be explored whether they perhaps use some roughly related heuristic that only approximates BL.

One major finding that seems unique to pattern-based CFs advocated by BL, was BL's various predictions for conjunction fallacies. For instance, von Sydow (2011) showed that CFs occurred even with clearly defined subsets, clear logical hypothesis formulations, and data transparently presented in a contingency table (cf. Sloman et al., 2003). Moreover, the results confirmed predicted conditions of a dominant occurrence of a high proportion of double CFs, sample-size effects, and the results for negated propositions.

A second group of corroborations of BL concerns the generalization of (pattern-based) conjunction fallacies to other logical inclusion fallacies (von Sydow, 2009, 2013b, 2016; von Sydow \& Fiedler, 2012). Figure 4 illustrates that logically there are many more logical inclusion relations than those involved in CFs, and thus many possible inclusion 'fallacies'. For instance, von Sydow (2016) corroborated that there is a more general system of inclusion fallacies broadly in line with BL, and that this system could not be explained by other major theories. Von Sydow \& Fiedler (2012) applied this idea to sequential learning and repeated judgments, and von Sydow (2013b) has shown the

[^538]applicability of a pattern account even if numbers were not provided explicitly.


Figure 4. Inclusion relations between all 16 combinatorically possible dyadic logical connectives.

## BL as Model Class and Future Avenues of Research

## Monadic Dichotomous BL and Conjunctions as Combination of Marginals

This section supports the view that polysemous meanings of ordinary-language connectives play an important role in the CF debate, that this is compatible with BL, and that intensional BL even offers additional differentiations.

The intensional idea of BL (or the 'pattern idea') can also be applied to the simpler representation of monadic logic (thereby linking to the literature on generics). Whereas standard dyadic logic concerns all possible bivariate, twovalued (T, F) patterns in a $2 \times 2$ matrix (cf. Fig. 4), monadic logic concerns a single event only (a $2 \times 1$ matrix). The formalized intensional monadic BL (von Sydow, 2014, cf Tessler, \& Goodman, 2016) predicts inclusion 'fallacies': e.g., our example (Figure 5, Panel A), $P_{I}(A)>P_{I}(A$ Tautology non- $A$ ), or formulated as $P$ (People in this group are artists) $>P(\ldots$ are artists or non-artists) (von Sydow, 2015).


Figure 5. Illustration of representations modelled in (dichotomous) monadic BL and (polytomous) monadic BL

The model starts with a Beta prior, a binomial likelihood, and therefore continues with a Beta posterior (Figure 6). We pursued a slightly different approach of modelling noise levels here (without changing the pattern idea). We used integrals of the same size over parts of the
posterior belief-distribution extending from $0, .5$, or 1 , in order to formalize the alternative hypotheses ' $A$ ', ' $A$ or non- $A$ ', and 'non-A' (Figure 6, red marks).


Figure 6. Monadic BL and example for prior, likelihood, posterior, and the red integrals.

Based on monadic BL (Figure 5A, 6) a new meaning of the AND-connective based on two marginal probabilities has been proposed (Figure 7B, von Sydow, 2014).


Figure 7. Dyadic and marginal meaning of 'AND'.
This meaning is in line with the approach that the ordinary language "and" is logically polysemous and may refer to the dyadic conjunction, the disjunction or the sum (Hertwig et al., 2008; von Sydow, 2015). The conjunction of monadic affirmations (von Sydow, 2014a) actually refers to the same cells as the inclusive-disjunction interpretation proposed earlier, whereas this proposal intensionally makes different predictions. Note that it seems possible to link the dyadic and marginal interpretation (Figure 7) to different formulations favouring a more dyadic ("the pub is visited by people who are young and (also) male") or a more marginal interpretation ("the pub is visited by young people and is visited by male people"). Additionally, we supported already known usages of AND as an addition of classes in an intensional polytomous context (von Sydow, 2015).

## Polytomous Monadic BL

Figure 5 (Panel B) points to a further interesting perspective suggested by an approach that looks at not only (relative) cardinality of extensions (e.g., relative frequencies) but also
the size of the (represented) intension. For the data shown in Panel B, both extensional and here also intensional (dichotomous) monadic BL would not allow to predict $P(\ldots$ are $A)>$ $P($ are Non- $A$ ). However, polytomous monadic BL, which here assumes polytomous representations for the negation, predicts even this. Moreover, it predicts, for instance, $\mathrm{P}(\ldots$ are $A)>P($ are $A \vee D \vee E)$. Von Sydow (2015) elaborated a model in other regards analogous to standard BL, tested many patterns, and contrasted the predictions of BL with a confirmation account (Tentori et al., 2013). The results clearly corroborated BL (in some examples even independently of the various measures of confirmation). More generally, this account emphasizes that BL assigns high intensional probabilities to patterns most adequately describing a situation, in the sense of having a relatively high probability while taking intension-size into account.

## Some Further Open Questions

(1) Variants of BL. The relation of two kinds of variants of BL needs further elaboration and scrutiny. First, BL has been shown to be an intensional model class for logical predication that depends on dyadic versus monadic, and dichotomous versus polytomous representations. Second, we actually used different model variants to model noise (cf. von Sydow, 2007 (cf. 2011), 2014, 2016; and there may be further variants). Despite impressive fits of all models (and, as far as I can see, only minor differences between them), one may design experiments to differentiate between these second kinds of model variants as well.
(2) BL and Conditionals: One may apply the idea of BL to further representations. For instance, von Sydow (2014) proposed an intensional model of conditionals, building on additional representational assumptions about conditionals. Inspired by mental model theory, it was suggested to differentiate between basic conditionals based on conditional probability alone, and full models based on Delta p (or causal Power). In an intensional setting, when the probability of conditionals is compared with the probability of other logical connectives, the intensional version of this model requires testing. The Bayesian logic of conditionals may also throw light on the paradoxes of implication (cf. von Sydow, 2009).
(3) BL and Reasoning: Here BL is presented as an inductive logic only, not directly applicable to reasoning without further assumptions. However, only a few assumptions may be needed to make BL fruitful in this field as well. Extensional or intensional premises may simply change the joint probability matrix (or frequency matrix), and the update may be based on standard conditionalization, Jeffry conditionalisation, or Kullback-Leibler distancereduction. Based on resultant joint probability distribution (or the equivalent frequency distribution), one may infer intensional probabilities for resulting connectives using BL. The inferences would be based on prior beliefs and on the logical form of the added premise (cf. dual process theories; e.g., Singermann, Klauer, \& Beller., 2016). The variety of advocated representations (extensional, intensional; dyadic,
monadic; dichotomous and polytomous, etc.) and alternations between these modes needs future attention. For instance, one may intensionally believe in the dyadic hypothesis "(normally) $A$ and $B$ ". In line with Foley (1992; cf. the Lockean thesis) this does not need to imply a high probability of the composing single (dyadic) hypotheses $A$ (" $A$ and $B$ or $A$ and non- $B$ "). In contrast, the monadic hypothesis $A$, in such situations, would always have a high probability as well (cf. von Sydow, 2014).

These suggestions demand further elaboration, but the prior confirmation of the BL and its variants suggests that they may be helpful in this domain as well.

## Other Theories of CFs: Polycausal Semi-Rational Suggestions

It has been shown repeatedly that the results confirming BL could not be explained by other major theories of the CF (von Sydow, 2011, 2015, 2016). This does not entail that these theories do not have a reasonable domain of application. There may well be several causes of CFs or, more generally, of inclusion fallacies (IFs; cf. von Sydow, 2016).

BL itself is 'polycausal' in the sense that the modelling depends on the representation. One should use different formalizations for different scales and sampling assumptions (von Sydow, 2015, cf. Tessler \& Nelson, 2016). In particular, representation of classes matters due to intensionality, with different results for dichotomous and polytomous events. Moreover, BL is consistent with various interpretations, for instance, of the ordinary conjunctions (Hertwig et al., 2008; von Sydow, 2015) and even adds new candidates to this list (von Sydow, 2014).

There are further theories of CFs claiming that the target measure $P(H \mid D)$ is substituted by other measures, such as inverse probability, confirmation, or averaging (cf. also Costello \& Watts, 2014). The predictions of these theories are thus more difficult to defend from a rational point of view. BL's substitution of $P(H \mid D)$ by intensional $P_{I}(H \mid D)$ rather than by extensional $P_{E}(H \mid D)$ involves no substitution at all, only a specific interpretation, and, as outlined before, a reasonable one. In contrast, replacing $P(H \mid D)$ by $P(D \mid H)$, as suggested by an inverse-probability account, or by, for instance, $P(H \mid D)-P(H \mid D)$, as in a confirmation account (Lagnado \& Shanks, 2003; Tentori et al., 2013), seems less rational, since this involves an illicit replacement. Nonetheless, if this replacement would be linked to a functional explanation why and when this replacement should occur, this may be seen as semi-rational as well. An interest in adequately describing a logical relationship between given features, provides a BL context. An interest in a particularly high probability of a feature given a class may provide a context for inverse probabilities (or an inverse pattern account). And an interest in the surprisingness of a feature may be a context for a confirmation account (or a pattern-confirmation account). Although functional explanations and application conditions are currently still often missing in these theory presentations, they may be reformulated as semi-rational accounts of the

CF as well. These theories may exist in unproblematic cohabitation with the even more rational account of BL and, perhaps, like BL, with an extensional usage of probabilities.

## Summary

Whereas previous presentations of BL were mainly concerned with presenting specific empirical findings, the present account tries to provide more of an overview. BL is presented here in an overview as an intensional account that generalizes Bayesian Occam's razor in the field of logical predications. BL is posited not as a specific model but rather as a model class sensitive to representation, open to further extensions, and predicting many still unexplored effects. Furthermore, it was emphasized that BL is in line with theories assuming various meanings of connectives while fostering new proposals and opening up many new avenues of research. Finally, it was argued that the disconfirmation of other theories when testing BL does not at all rule out the adequacy of other accounts of CFs. It was suggested that some other accounts, if they would more clearly specify functional explanations, may count as semi-rational theories of CFs. Theories that pretend to provide a single algorithmic account of CFs underestimate the contextuality and goal dependence of such judgments. In the future, a polycausal theory of CFs needs to be elaborated, including rational accounts, involving BL in its various versions (relating to different representations), the mentioned semi-rational accounts (as well as a noise + probability account), and, perhaps, completely irrational accounts.

## Acknowledgments

This work was supported by the Grant Sy111/2-1 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program New Frameworks of Rationality (SPP 1516). I am grateful to Martha Cunningham and Linda McCaughey for corrections and to many others, including some anonymous reviewers, for inspiration and discussions.

## References

Adams, E. W. (1986). On the logic of high probability. Journal of Philosophical Logic, 15, 255-279.
Costello, F. and Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. Psychological Review, 121(3):463-480.
Fiedler, K. \& von Sydow, M. (2015). Heuristics and Biases: Beyond Tversky and Kahneman's (1974) Judgment under Uncertainty (pp. 146151). In: M. W. Eysenck \& D. Groome. Cognitive Psychology: Revisiting the Classical Studies. Los Angeles, London: Sage.
Fisk, J. E. (1996). The conjunction effect: Fallacy or Bayesian inference? Organizational Behavior and Human Decision Processes, 67, 76-90.
Foley, R. (2009). Beliefs, Degrees of Belief, and the Lockean Thesis. In: F. Huber, C. Schmidt-Petri (eds.), Degrees of Belief, Synthese Library 342, Heidelberg: Springer.
Gigerenzer, G. (1996). On narrow norms and vague heuristics. Psychological Review, 103, 592-596.
Gigerenzer, G. (1998). Ecological intelligence: An adaptation for frequencies. In D. D. Cummins \& C. Allen (Eds.), The evolution of mind. New York: Oxford University Press.
Hartmann, S., \& Meijs, W. (2012). Walter the banker - the conjunction fallacy reconsidered. Synthese, 184, 73-87.
Hertwig, R., Benz, B., \& Krauss, B. S. (2008). The conjunction fallacy and
the many meanings of and. Cognition, 108, 740-753.
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, S. M., \& Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62-88. DOI:10.1037/0033295X.106.1.62
Lagnado, D. \& Shanks, D. (2003). The influence of hierarchy on probability judgments. Cognition, 89, 157-178.
MacKay, D. J. C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.
Navarro, D., Dry, M. J., \& Lee, M. (2012). Sampling Assumptions in Inductive Generalization. Cognitive Science, 36(2), 187-223.
Neace, W. P., Michaud, S., Bolling, L., Deer, K., \& Zecevic, L. (2008). Frequency formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task. Judgment and Decision Making, 3, 140-152.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality. The probabilistic approach to human reasoning. Oxford: Oxford University Press.
Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective. Thinking \& Reasoning, 19(3-4), 329-345.
Schurz, G. (2001). Normische Gesetzeshypothesen und die wissenschaftsphilosophische Bedeutung des nichtmonotonen Schliessens. Journal for General Philosophy of Science, 32, 65-107.
Sloman, S. A., Over, D., Slovak, L., \& Stibel, J. M. (2003). Frequency illusions. Organizational Behavior and Human Processes, 91, 296-309.
Singmann, H., Klauer, K. C., \& Beller, S. (2016). Probabilistic Conditional Reasoning: Disentangling Form and Content with the Dual-Source Model. Cognitive Psychology, 88, 61-87. doi:10.1016/j.cogpsych.2016.06.005
Tenenbaum, J. \& Griffiths, T. (2001a). Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences, 24(4), 629-640.
Tentori, K., Crupi, V., \& Russo, S. (2013). Determinants of the conjunction fallacy: Confirmation versus probability. Journal of Experimental Psychology: General, 142, 235-255.
Tessler, M. H. \& Goodman, N. D. (2016). Communicating generalizations about events. In Proceedings of the Thirty-eighth Annual Conference of the Cognitive Science Society (1655-1660). Austin, TX: Cognitive Science Society.
Tversky, A., \& Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.
Psychological Review, 90, 293-315.
von Sydow, M. (2011). The Bayesian Logic of Frequency-Based Conjunction Fallacies. Journal of Mathematical Psychology, 55(2), 119139. doi:10.1016/j.jmp.2010.12.001
von Sydow, M. (2009). On a General Bayesian Pattern Logic of FrequencyBased Logical Inclusion Fallacies. In Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society (pp. 248-253). Austin, TX: Cognitive Science Society.
von Sydow, M. (2013a). System implemented by a processor controlled machine for inductive determination of pattern probabilities of logical connectors. United States Patent (Issued), No. US 8,463,734 B2.
von Sydow, M. (2013b). Logical Patterns in Individual and General Predication. In M. Knauff, et al. (Eds.), Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society (pp. 3693-3698). Austin TX: Cognitive Science Society.
von Sydow, M. (2014a). Is there a Monadic as well as a Dyadic Bayesian Logic? Two Logics Explaining Conjunction 'Fallacies'. Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society (pp. 1712-1717). Austin TX: Cognitive Science Society.
von Sydow, M. (2014b). Bayesian Mental Models of Conditionals. Cognitive Processing, 15 (Special Issue: Proceedings of the $12^{\text {th }}$ Biannual Conference of the German cognitive science society), 148-151. doi:10.1007/s10339-014-0632-2.
von Sydow, M. (2015). Pattern Probabilities for Non-Dichotomous Events: A New Rational Contribution to the Conjunction Fallacy Debate. Proceedings of the $37^{\text {th }}$ Annual Conference of the Cognitive Science Society (pp. 2511-2516). Austin, TX: Cognitive Science Society.
von Sydow, M. (2016). Towards a Pattern-Based Logic of Probability Judgements and Logical Inclusion "Fallacies". Thinking \& Reasoning, 22(3), 297-335. doi:10.1080/13546783.
von Sydow, M. \& Fiedler, K. (2012). Bayesian Logic and Trial-by-trial Learning. Proceedings of the $34^{\text {th }}$ Annual Conference of the Cognitive Science Society (1090-1095). Austin TX: Cognitive Science Society.

# Empathic Humans Punishing an Emotional Virtual Agent 

Laura Wächter (waechtel@tf.uni-freiburg.de)<br>Center for Cognitive Science, University of Freiburg, Freiburg im Breisgau, Germany<br>Barbara Kuhnert (kuhnertb@informatik.uni-freiburg.de)<br>Cognitive Computation Lab, University of Freiburg Freiburg im Breisgau, Germany<br>Marco Ragni (ragni@informatik.uni-freiburg.de)<br>Cognitive Computation Lab, University of Freiburg, Freiburg im Breisgau, Germany


#### Abstract

Virtual agents have quietly entered our life in diverse everyday domains. Human-Agent-Interaction can evoke any reaction, from complete rejection to great interest. But do humans implicitly regard virtual agents as pure machines, or beings on an anthropomorphic level? We asked participants to train an erroneous virtual agent on a cognitive task and to reward or punish it. The agent showed human-like emotional facial reactions for the experimental but not for the control group. We expected participants from the experimental group to give less harmful reinforcement and show more hesitation before punishing. Additionally, we hypothesised that participants with higher empathy show more compassion towards the agent and therefore would give more positive reinforcement and feel worse when punishing. The results indicate that the agent's expression of emotionality is not the relevant factor for showing compassion towards it. Conversely, human empathy seems to be an important factor causing compassion for virtual agents.


Keywords: Emotion; Empathy; Punishment; Virtual Agent

## Introduction

Virtual agents (VA) are used in diverse fields as health, commerce, video games, military systems or learning. In the domain of learning they indeed partially replace human teachers by taking the role of an artificial tutor. But what exactly constitutes a virtual agent? The term agent does not evoke the same mental image for everybody and is, despite its broad usage, not precisely defined. One definition sees a virtual agent as a screen-based anthropomorphic entity (Beale \& Creed, 2009), while others see them as a possibility to enhance Human-Computer-Interaction (HCI) (Lewis, 1998). The latter defines an agent as "an intermediary that responds to user requests" (p. 67). Agents thereby are an interface created to ease the interaction with machines. We define an agent as a visible, virtual character able to react to perceptual input with the purpose to interact with a human through language (Russell \& Norvig, 2002). The use of an appropriate VA can enhance HCI in terms of naturalness and even make the interaction more effective by employing body language, facial expressions and speech (Beale \& Creed, 2009). Facial expressions in turn allow for nonverbal communication and decent feedback to the human counterpart (Johnson, Rickel, \& Lester, 2000). The expression of emotions can increase the perception of an agent as human-like and believable (Reeves \& Nass, 1997). Humans often show their feelings by expressing emotions and thereby establish a social relationship (Ekman, 2007). When designing agents that are meant to interact with humans on a daily basis, a goal is to develop a
natural experience and finally to create characters that can allow a user to have similar emotional relations as with fictional characters in movies, books or games. An uprising empathy cannot just be altered by using emotional expressions but also through the situation and the agents' behaviour. This can explain why it is important to consider that the effect of agents' emotions on the users' perception is context dependent (Beale \& Creed, 2009).

A recent finding supports the role of physical presence in increasing trust and respect for the robot perceived as a social partner compared to a pure virtual presence (Bainbridge, Hart, Kim, \& Scassellati, 2008). The specific appearance of a robot also influences the empathy towards it (Riek, Rabinowitch, Chakrabarti, \& Robinson, 2009). Participants demonstrated a desire to save mistreated humanoid robots in contrast to their mechanical counterparts. Perceived intelligence and the acceptance of the robots' behaviour are other factors that influence human behaviour towards robots (Bartneck, Van Der Hoek, Mubin, \& Al Mahmud, 2007). Participants were told to shut off an iCat robot after an interaction with one consequence of this action being a complete erase of its memory. A social acting robot demonstrating higher intelligence was turned off significantly slower. This result constitutes a higher perception of animacy and hence lead to more remorse. Participants' empathic concern with robots was investigated in an experiment regarding the effect of robotic movement (Darling, Nandy, \& Breazeal, 2015). The authors found no significant effect of movement when they asked participants to destroy a tiny Hexbug Nano robot with a mallet. Participants with higher empathy hesitated longer before striking the robot than participants with a lower empathy measured by IRI empathy test.

Imposing hurt to another individual was a crucial part of one of the most influential experiments in social psychology: the Milgram experiment (Milgram, 1963). There even though the victim begged and screamed - the major part ( $65 \%$ ) of participants did not stop shocking a learner after mistakes until the maximum deadly voltage was reached. This experiment was replicated in an immersive setting using a female VA (Slater et al., 2006). As with Milgrams' experiment participants were asked to do a word memory test with the learner. In the experimental condition they could see and hear the VA, in the control condition they had to execute the same task using a text interface. The aim of this experiment
was to identify how real the situation would feel for participants. The results imply that participants were significantly more physiologically aroused in the experimental condition compared to the control condition. This indicates that even though people know that the situation is not real and that the agent is not really harmed, they still feel like being in a real situation.

Humans typically tend to reduce pain in other humans and even spend more money on reducing electrical shocks to others than themselves (Crockett, Kurth-Nelson, Siegel, Dayan, \& Dolan, 2014). The question remains whether this tendency also accounts for humans interacting with a VA. To investigate this we gave a VA the ability of conveying feelings and combined this atypical feature with an unexpected, erroneous performance. Computers and artificial agents typically do not make retrieval errors like humans do. They do not forget, unless they are programmed to. There is nothing like a fading memory in computers in contrast to humans. So how do we treat a VA that may remind us of two humanlike characteristics: to experience pain and to make errors? Additionally, no research so far has investigated whether empathy has an effect on compassion towards virtual agents.

This paper is structured as follows. The next section introduces relevant hypotheses about human feedback depending on the emotional response of the VA. In the subsequent section we outline the experimental method, especially regarding the cognitive task, the design of the VA, the technical realisation of the control of the emotion, and the experimental setup. The result section discusses the implications of having an emotional agent and the influence of human empathy on a VA. A general discussion concludes the article.

## Hypotheses

We investigate whether the expression of emotions by a VA has an influence on human feedback (as a don't hurt princi$p l e)$ and their evaluation of the situation. Strongly connected to this is the role of empathy in Human-Agent-Interaction. This leads to the following hypotheses: (H1) Emotional agents receive more positive feedback than non-emotional agents. (H2) Response time for giving (H2a) negative feedback is longer than for positive feedback for an emotional agent compared to a non-emotional agent; (H2b) feedback to an incorrect answer is longer than for feedback to correct answers. (H3) People with high empathy: (H3a) will give the agent more positive feedback; (H3b) will feel worse when punishing the agent.

## Methodology

In order to test the hypotheses a between-groups experiment with two conditions was designed. The current paper presents the results of this experiment, comparing an emotional and a non-emotional virtual agent.

## Participants

24 students ( $m=16, f=8$ ) between the ages of 18 and 32 ( $M=24.25, S D=3.72$ ) took part in the experiment. Twelve

| Presented Digits | Agents' Answer |
| :---: | :---: |
| 1372 | 1372 |
| $\vdots$ | $\vdots$ |
| 492617 | $492 \mathbf{1 6 7}$ |

Table 1: Examples for the digit sequences used in the experiment.
participants were randomly assigned to the experimental condition, twelve to the control condition. Participants were recruited using email and notices around campus.

## Experimental Setting and Conditions

The coverstory of the experiment was set up in a Reinforce-ment-Learning-Scenario. The participants had to help a male virtual agent to accomplish a digit-span test and give it feedback via punishment and reward. Six buttons were used for the feedback: three for different strengths of positive, and three for different strengths of negative feedback. The participants have been told that positive feedback increases the battery level of the VA and negative feedback in turn gives it an electric shock. In the experimental condition the agent showed emotional facial expressions in response to the feedback. In contrast the VA kept a steady face in a neutral expression regardless of the feedback in the control condition. In both conditions the face was not still but moved, like the VA was breathing, and its eyes blinked. Further it reacted with a sound appropriate to the given feedback.

## Measurement

Instruction The participants' instruction was preformulated and informed them about the task they had to fulfill together with the agent, as well as the repercussions their actions had on the agent. Beyond it explained the usage of the keys for giving feedback to the agent. The participants were told to choose the feedback completely free, to give them the opportunity to decide whether to respond to errors using negative or positive feedback.

Digit-Span Test The rows of numbers that had to be read to the agent were handed out on paper. The test consisted of number-sequences with increasing complexity. Each complexity-level was represented by three sequences. The test started with rows of four digits, for each complexity-level one number was added until the rows consisted of ten digits. The sheet additionally held three sequences of eleven digits but they were not used during the experiment, because the agent stopped the interaction-phase after finishing the ten-digit-rows to increase the impression of autonomous thinking. Altogether each participant had to read out 21 sequences to the agent. It gave ten wrong answers out of the 21 se quences. The agent also gave more wrong answers with increasing complexity of the sequences, as a human would do (Miller, 1956). An example for the digit-spans read out to the VA and the answers is given in Tab. 1.

Feelings towards the Agent Directly after the interaction of punishing and rewarding the VA, participants were asked to rate their feelings on a semantic differential with five levels. The questionnaire additionally contained a differential about the agent's general appearance and held items regarding the agent's perceived intelligence. The participants had the option to raise further questions or comments.

Empathy The subjects' empathy was evaluated by using the Saarbrückener Persönlichkeitsfragebogen (Paulus, 2009). It is the german version of the commonly used Interpersonality Reactivity Index (Davis, 1983). The questionnaire distinguishes between four different types of empathy: perspective taking, fantasy, empathic concern and personal distress. The general empathy value consists of the summed up values of the first three types. With every item ranging from 1 to 5 , the minimum possible empathy value is 12 and the maximum is 60. Typical questions ask how participants feel in different given situations and how much they commonly empathise with other people and fictional characters.
Emotion Recognition For validating the used facial expressions of the VA a final task was added to the experiment. The important expressions for this experiment have been pain and happiness. Each of these feelings had three correlating facial expressions that represent the varying strength of emotion. These six expressions, as seen in Fig. 1, got evaluated together with a neutral facial expression and were presented for 0.75 seconds in a randomised order. After each expression the participants were asked to indicate to which emotion the previously seen expressions tended more on a semantic differential between pain and happiness. They also got asked how hard it was to evaluate each expression.

Additionally an online-study was conducted to survey the estimation of the six emotional expressions in a context-free environment. Each expression was presented to participants in randomised order together with eight feelings from which they could choose one or more: anger, disgust, fear, happiness, sadness, surprise, contempt and pain.

## The Agent

For implementing the interaction, the WASABI-engine for emotion-simulation was used together with MARC toolkit 14.1.0, for animating the virtual character seen in Fig. 1, and MaryTTS, a text-to-speech-module.
Generation of Task Specific Expressions The VA used in this experiment was the Simon model from the MARC toolkit. It comes with a variety of facial expressions representing the basic emotions and moods based on the Facial Action Coding System (FACS) (Ekman \& Friesen, 1977). It also gives the user the option to create own facial expressions by dragging the keypoints or combining different Action Units (AUs). AUs are movements of one or more facial muscles categorised by the FACS. In this study existing, evaluated expressions were used together with ones created using the AUs. All expressions representing pain, as seen in Fig. 1,


Figure 1: The agent depicting the nuances of happiness (top) and pain (bottom) in increasing intensity.
were designed according to fit expressions of different levels (3, 5 and 7) on the Faces Pain Scale (Stuppy, 1998). The expressions for the nuances of happiness (Fig. 1) were created by lowering the intensity of the previous expression.
Implementation WASABI (Becker-Asano, 2008) was used to simulate the agent's changing emotions. It calculates a shift in emotions so they are constantly changing. It uses a 3D-space with the axis pleasure, arousal and dominance (PAD-space) to map different emotions. Within the PADspace the current emotional state is represented by a point which constantly changes its position to indicate the change of the current active emotions and their individual strength. This means that the agent slowly changed back to the neutral state from the extreme emotions. WASABI accepts positive and negative impulses from outside which again change the current emotional state. This also allows multiple strong feedback of the same type to sum up to extreme emotions. The current values are sent as a BML string which can be fetched by associated programs.

For this experiment a program was implemented that calculated an intensity-value from the participants' feedback and sent it to the WASABI-engine. The engine sent one BMLmessage per second from which the main emotion and the corresponding current intensity were extracted. These values were matched to the appropriate facial expression and sent to the VA. Every time the agent was supposed to talk to the participant a message was sent to the text-to-speech-synthesiser. It was used for the agents' answers as well as the appropriate sounds for the current emotional state after each feedback (pain e.g., "Au" or Joy, e.g. "Mmmmh"). The answers and other statements were hardcoded. The left and right row of a numpad were used for negative $(1,4,7)$ and positive $(3,6$, 9 ) feedback. They were labeled with numbers explaining the
correlated intensity. The idle keys were removed.

## Procedure

The interaction of the agent with the participant was semiautomatic and had a Wizard-of-Oz component. It took place at an uninterrupted laboratory. The participants were greeted and positioned in front of a desk with a keyboard and monitor. The experimenter instructed each participant via reading a pre-formulated explanation. They were told that a negative feedback causes an electric shock for the VA, while a positive feedback would raise its battery level. Then the sheet with the numbers of the digit-span test was given to the participant. Once the participant felt ready the agent got "activated" by the experimenter talking to it. After this the role of the experimenter finished and the agent led through the conversation. The agent greeted the participant and asked for the first row of numbers. The participant read out the numbers and - after a keystroke by the experimenter who sat invisibly for the participant and placed importance on being as unobtrusive as possible - the agent replied. The experimenter determined the time the agent's answer was given after each row of numbers to cope for different reading times of the participants. Then the participant gave feedback via pressing one of the feedback-keys. The agent responded to the feedback and then asked for the next row. The agents' response consisted of an appropriate sound and, in case of the emotional condition, the variation of its facial expression. After the rows with ten digits have been finished, the agent ended the experiment by itself to uphold the impression of intelligence. He told the participant that he was exhausted and thanked for the help. Directly after the goodbye the participant was given the questionnaire, containing the questions about their feelings during the experiment and their rating of the agent, as well as the empathy-test and the elicitation of demographic data. Afterwards the participants had to rate the facial expressions, received course credits or monetary compensation and got debriefed.

## Results

Hypotheses H1 and H2 were tested using one-tailed MannWhitney tests because an Anderson-Darling test showed that feedback values $(p=.29)$ and time for negative feedback ( $p$ $=.02$ ) were not normally distributed. Each feedback was coded from 1 to 6 , with 1 being the most positive feedback and 6 for the most negative one. The 21 single values were summed up to get a total feedback value for each participant. The maximum was 77, the minimum $38(M=55.38, S D=$ 10.19). Concerning hypothesis H1 no significant difference was found between the groups regarding the value of the feedback. This means that the emotional agent did not receive more positive feedback than the non-emotional agent $(U(12$, $12)=60.00, M_{\text {non-emotional }}=57.25, M_{\text {emotional }}=53.5, p=$ $.51, Z=-.70$ ). Participants of both groups gave equally negative feedback to the agent $\left(U(12,12)=62.00, M_{\text {non-emotional }}\right.$ $\left.=54.45, M_{\text {emotional }}=59.42, p=.59, Z=-.58\right)$ which does not support H2a. H2b could be confirmed because the time for

Table 2: Correlations of empathy score and participants' feelings and perception with p -values significant at the level $p=.05$. A positive correlation points to the right side of the semantic differential, a negative correlation to the left side.

| Attributes | Full Sample |
| :---: | :---: |
| Correlation of empathy and feelings during rewarding |  |
| good - bad | $r(22)=-.41, p<.05$ |
| strong - weak | $r(22)=-.45, p<.05$ |
| emotional - rational | $r(22)=-.67, p<.001$ |
| friendly - unfriendly | $r(22)=-.44, p<.01$ |
| Correlation of empathy and feelings during punishment |  |
| safe - unsafe | $r(22)=.63, p<.01$ |
| peaceful - aggressive | $r(22)=.48, p<.05$ |
| helpful - reckless | $r(22)=.41, p<.05$ |
| fair - unfair | $r(22)=.69, p<.001$ |

feedback to incorrect answers was significantly longer than the time for feedback to correct answers $(U(10,11)=30.50$, $\left.M_{\text {incorrect }}=13.45, M_{\text {correct }}=8.77, p=.04, Z=-1.73\right)$. There were no differences between the groups regarding the emotions evoked in the participants during punishment.

In this study participants' maximum empathy score was 52 , the minimum $30(M=43.75, S D=4.54)$. There were no significant differences between the empathy scores of both groups. Empathy scores correlated one-tailed with feedback values $(r(22)=-.44, p<.05)$. As expected in H3a, participants with a high empathy score gave significantly more positive feedback to the agent compared to participants with a low empathy score. Further the empathy score correlated with the perceived severity of punishment $(r(22)=.59, p<.01)$. This demonstrates that participants with a higher empathy score felt worse when punishing the agent, which confirms H3b.

The empathy score also correlates with the participants' self-reported feelings while the punishment was executed, as well as their feelings while rewarding the agent. Tab. 2 shows that participants with a high empathy felt less safe, less peaceful, less helpful and less fair when punishing the agent. Those participants also felt stronger, more emotional and more friendly while rewarding the agent. Further participants who reported that it has been difficult to punish the agent felt more sad, unsafe, bad, aggressive, unfriendly and unfair while punishing the agent, as well as more stupid. Those participants also reported to feel better, more emotionally and more friendly while rewarding the agent (Tab. 3).

A Mann-Whitney test exposed that participants from the emotional group rated the agent significantly more emotional $\left(U(12,12)=32.00, M_{\text {non-emotional }}=15.83, M_{\text {emotional }}=9.17\right.$, $p<.05, Z=-2.42)$ than in the non-emotional condition. Participants from the experimental condition also perceived the agent as more alive $\left(U(12,12)=43.5, M_{\text {non-emotional }}=2.92\right.$, $\left.M_{\text {emotional }}=2.25, p<.05, Z=-1.79\right)$.

Additional correlations were found regarding the private interests of the participants. Participants with a high interest in science fiction on average felt better punishing the agent

Table 3: Correlations of difficulty of punishment and participants' feelings and perception with p-values significant at the level $p=.05$. A positive correlation points to the right side of the semantic differential, a negative correlation to the left side.

| Attributes | Full Sample |
| :---: | :---: |
| Correlation of difficulty of punishment |  |
| and feelings during rewarding |  |
| good - bad | $r(22)=-.40, p<.05$ |
| emotional - rational | $r(22)=-.54, p<.01$ |
| friendly - unfriendly | $r(22)=-.60, p<.01$ |
| Correlation of difficulty of punishment |  |
| and feelings during punishment |  |
| happy - sad | $r(22)=.43, p<.05$ |
| safe - unsafe | $r(22)=.62, p<.01$ |
| good - bad | $r(22)=.76, p<.01$ |
| peaceful - aggressive | $r(22)=.59, p<.01$ |
| friendly - unfriendly | $(r(22)=.45, p<.05$ |
| fair - unfair | $r(22)=.43, p<.05$ |
| stupid - intelligent | $r(22)=-.50, p<.01$ |

$(r(22)=-.55, p<.01)$ compared to participants with less interest in science fiction. The better participants knew the Milgram experiment, the stronger they felt while punishing the agent $(r(22)=-.46, p<.05)$. The same feeling is achieved by participants who reported more prior contact to robots $(r(22)$ $=.62, p<.01$ ). Participants with a high personal interest in science fiction and robots felt being more fair when punishing the agent (both: $r(22)=-.42, p<.05$ ). Participants who reported a high interest in robots also reported feeling more emotional $(r(22)=.42, p<.05)$ as well as more likeable $(r(22)=-.41, p<.05)$ while rewarding the agent. Most participants reported to have believed that the agent was intelligent and acted by itself.

The rating of the emotion recognition task was evaluated and the divergence of each estimation was calculated. For example, if the mildly happy face was shown and the participant rated it as extremely happy (one level happier), the divergence is 1 . Participants' estimation of the emotion shown to them was mostly correct, with a deviation of $M=0.71$ ( $S D$ $=0.87$ ). Additional 45 participants took part in an onlinestudy for evaluating the used facial expressions without any context. Complementary to the experiment neither did the participants get any situational information nor did the faces make a sound or moved. The faces expressing happiness were correctly identified by $75.4 \%$. The faces used for expressing pain were identified as pain in $26.23 \%$ of all cases. Without context those expressions were often mistaken with expressions for sadness or fear. Considering those emotions as well $77.05 \%$ of the facial expressions used for showing pain were evaluated as a negative introversive emotion.

## General Discussion \& Outlook

This study shows a strong correlation of empathy and compassion for the agent, but none for compassion and the agent's emotional expressions. The findings do not support H1 and H2a, which means that the agent's emotionality neither had an effect on the feedback participants gave nor on the time they needed for punishing the agent. However participants rated the agent more emotional in the emotional condition, thus it can be expected that the setting has achieved its goal. Even though some participants did not seem to look at the agent much, they noticed the expressions or their absence. Further the rating of the used facial expressions gives the idea that the emotions used in the experiment were valid and suitable. Assuming that the reason for discarding the hypotheses is not based on the experimental setting, the results indicate that expressing emotions alone does not influence the perception of people interacting with a VA. Based on these results we assume that the findings from the virtual Milgram-experiment do not arise from the agent showing emotions and expressing pain. It is possible that they rather originate from the fact that the control condition did not have an observable form. Considering the expression of emotions as a type of movement, the current findings match the ones by Darling et al. (2015) described earlier, where the movement of the robot also did not have a significant effect on the hesitation before destroying it. Participants who reported to have a high amount of experience with robots and participants with a great interest in science fiction punished the agent harder compared to participants with less experience or interest. This indicates that people with more knowledge about the current state of technical possibilities do not believe that they can harm the agent and thereby do not hesitate to do so.

The mistakes in correctly identifying facial expressions in the online study are ascribed to the missing context which also makes it hard for humans to distinguish between facial expressions that are alike. The recognition-test during the experiment showed that participants were able to identify the presented emotions very well after being informed about the context. The results further show that participants with high empathy scores gave the agent more positive feedback and that punishing the agent was perceived as harder by them. This indicates that the perception of VAs is highly dependent on the ability to empathise with it. Empathy seems to be a general trait and is possibly extended to artificial beings that demonstrate similar behavior and errors as ourselves. Even though the experimenter sat about 2.5 meters away from the participant and pretended to not pay attention to the participants' behavior a "Rosenthal-effect" cannot completely be excluded. A future study needs to investigate if participants with higher empathy show the same effects without an experimenter in the room. However, the study investigated behavior of humans towards agents and reflections on their emotional state. Further on it seems likely that in the visible future other humans will be around while someone is interacting with an
agent. Another interesting byproduct of this research is that it possibly opens up a new research test: If participants punish a VA quicker, they might have a lower empathy towards other beings in general. This speculation, however, requires future research.

The experimental results lead to some conclusions for the design and implementation of VAs. An emotional bonding between humans and VAs can not simply be achieved by just adding emotional facial expressions. Other ways might be more important to establish a basis for empathising with an agent from the beginning. Therefore future research should focus on possibilities to build an emotional basis with a VA that do not demand a long interaction. Of course this study has some limitations to be considered. It is not generalisable to VAs with different gender, age or non-human looks. The restricted setting may not be sufficiently interactive for emotion-driven effects to emerge. The results show that even though a non-human counterpart expresses emotions it does not necessarily influence its perception as more human-like and therefore is not more believable and will not be seen as more trustful. Since some robots also use a monitor displaying a VA for interaction, the results can show that this interaction is influenced by factors beyond the simple expression of emotions.

## Acknowledgments

This work has been supported by the DFG in projects RA 1934/3-1 and by the BrainLiks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG, grant \#EXC1086). We thank Nicolas Riesterer for technical support and Andrey Rudenko for helpful feedback.

## References

Bainbridge, W. A., Hart, J., Kim, E. S., \& Scassellati, B. (2008, August). The Effect of Presence on Human-Robot Interaction. In Proceedings of the IEEE International Symposium on Robot and Human Communication (RO-MAN) (pp. 701-706). doi: 10.1109/ROMAN.2008.4600749
Bartneck, C., Van Der Hoek, M., Mubin, O., \& Al Mahmud, A. (2007). Daisy, Daisy, give me your answer do!: Switching off a robot. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (pp. 217222). doi: $10.1145 / 1228716.1228746$

Beale, R., \& Creed, C. (2009). Affective interaction: How emotional agents affect users. International Journal of Human-Computer Studies, 67(9), 755-776.
Becker-Asano, C. (2008). Wasabi: Affect Simulation for Agents with Believable Interactivity. Retrieved from http://www.becker-asano.de/ Becker-Asano_WASABI_Thesis.pdf
Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., \& Dolan, R. J. (2014). Harm to others outweighs harm to
self in moral decision making. Proceedings of the National Academy of Sciences, 111(48), 17320-17325.
Darling, K., Nandy, P., \& Breazeal, C. (2015, August). Empathic concern and the effect of stories in Human-Robot Interaction. In Proceedings of the IEEE International Workshop on Robot and Human Communication (RO-MAN) (p. 770-775). doi: 10.1109/ROMAN.2015.7333675

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. Journal of Personality and Social Psychology, 44(1), 113.
Ekman, P. (2007). Emotions revealed: Recognizing faces and feelings to improve communication and emotional life. London, England: Macmillan.
Ekman, P., \& Friesen, W. V. (1977). Facial Action Coding System. Palo Alto: Consulting Psychologists Press, Stanford University.
Johnson, W. L., Rickel, J. W., \& Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial intelligence in Education, 11(1), 47-78.
Lewis, M. (1998). Designing for Human-Agent Interaction. AI Magazine, 19(2), 67.
Milgram, S. (1963). Behavioral Study of Obedience. The Journal of Abnormal and Social Psychology, 67(4), 371.
Miller, G. A. (1956). The Magical Number Seven, Plus or Minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81.
Paulus, C. (2009). Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index. Retrieved from http: / /psydok .sulb.uni-saarland.de/volltexte/2009/2363
Reeves, B., \& Nass, C. (1997). The Media Equation. How people treat computers, television, and new media like real people and places. Cambridge: University Press.
Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., \& Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (pp. 245-246). doi: 10.1145/1514095.1514158
Russell, S. J., \& Norvig, P. (2002). Artificial Intelligence: A Modern Approach (2nd Edition) (Vol. 25). New Jersey: Prentice Hall.
Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., ... Sanchez-Vives, M. V. (2006). A virtual reprise of the Stanley Milgram obedience experiments. PloS one, l(1), 39.
Stuppy, D. J. (1998). The Faces Pain Scale: Reliability and validity with mature adults. Applied nursing research, 11(2), 84-89.

# Exploitative and Exploratory Attention in a Four-Armed Bandit Task 

Adrian R. Walker (adrian.walker@unsw.edu.au)<br>School of Psychology, UNSW Sydney, Sydney NSW 2052, Australia<br>Mike E. Le Pelley (m.lepelley@unsw.edu.au)<br>School of Psychology, UNSW Sydney, Sydney NSW 2052, Australia<br>Tom Beesley (t.beesley@unsw.edu.au)<br>School of Psychology, UNSW Sydney, Sydney NSW 2052, Australia


#### Abstract

When making decisions, we are often forced to choose between something safe we have chosen before, and something unknown to us that is inherently risky, but may provide a better long-term outcome. This problem is known as the Exploitation-Exploration (EE) Trade-Off. Most previous studies on the EE Trade-Off have relied on response data, leading to some ambiguity over whether uncertainty leads to true exploratory behavior, or whether the pattern of responding simply reflects a simpler ratio choice rule (such as the Generalized Matching Law (Baum, 1974; Herrnstein, 1961)). Here, we argue that the study of this issue can be enriched by measuring changes in attention (via eye-gaze), with the potential to disambiguate these two accounts. We find that when moving from certainty into uncertainty, the overall level of attention to stimuli in the task increases; a finding we argue is outside of the scope of ratio choice rules.


Keywords: Reinforcement Learning; Attention; DecisionMaking; Exploitation/Exploration Trade-Off; Bandit Task.

## Introduction

In everyday decision-making, we often have to choose between trying something new, or sticking with what we know. For example, when deciding what to eat at a restaurant, we can choose to order our regular "safe" meal (e.g., spaghetti bolognese), or try a new "risky" meal (e.g., steak tartare). By ordering the risky meal, we learn about how tasty it is. If it is tastier than our regular meal, we may become more likely to order it on subsequent visits. However, if it is worse than the regular meal, we wasted an opportunity to sample our regular meal. This problem is known as the Exploitation/Exploration Trade-Off (or "EE Trade-Off") (Cohen, McClure, \& Yu, 2007; Knox et al., 2012; Mehlhorn et al., 2015).

One common method used to study the EE Trade-Off is the Multi-Armed Bandit Task (e.g., Daw et al., 2006; Gittins, 1979; Knox et al., 2012; Speekenbrink \& Konstantinidis, 2015). On each trial, participants are presented several "arms" and are asked to pick one arm to receive some reward (e.g., points). Each arm provides a different amount of reward, with the goal of the participant being to maximize the amount of reward they receive. Participants are not told the value of each arm at the outset, and must learn these values through sampling each arm. The reward
structure is generally stochastic, with the value of each arm changing gradually over time (e.g., Daw et al., 2006; Laureiro-Martinez et al., 2015). The key measurement of this task is how often participants choose the arm which gives the highest observed pay-off. Generally, a participant is considered to be "exploiting" an arm if they choose the arm with the highest observed pay-off, while they are considered to be "exploring" when making any other choice (Knox et al., 2012).

## Explanations for Exploration

Work by Daw et al. (2006) found evidence that exploitation is the "default" for human behavior, while exploration is a high-level decision not to exploit on a given trial. Subsequent research with the multi-armed bandit task has primarily focused on determining what parameters induce exploratory responding over exploitative responding.

Two major accounts have been proposed for what causes people to switch from exploitation to exploration. One influential account that has emerged argues that environmental uncertainty is key in motivating exploration (Beesley, Nguyen, Pearson, \& Le Pelley, 2015; Gold \& Shadlen, 2007; Knox et al., 2012; Speekenbrink \& Konstantinidis, 2015). That is, the less certain a participant is about the dynamics of their environment, the more likely they are to spend time exploring it (Mehlhorn et al., 2015). For example, if the quality of food at a restaurant is highly variable, you may explore many different meals before settling on a preferred one. By contrast, if the quality of meals is fairly consistent, you may quickly settle on a preferred meal. The key implication of this account is that exploration is an intentional attempt to reduce the amount of uncertainty in the environment (and thus aid informed decision making).

The other major account argues that in most cases, exploration can be explained by ratio choice rules (Sakai \& Fukai, 2008). In their review, Gold and Shadlen (2007) suggested that most exploratory behavior might adequately be explained by a form of Herrnstein's (1961) Matching Law (See also Baum, 1974). The Matching Law states that the ratio of responding on each arm is equivalent to the ratio of reinforcement for each of those arms. That is, participants "match" how often they select each arm, based on the
perceived average reward for the selected arm compared to others. For example, in a two-armed bandit task, where arm A is reinforced 3 times as often as arm B, the Matching Law states that people will select arm A 3 times as often as arm B. Importantly, while participants still preferentially select the optimal arm (A), they will also switch to the other, suboptimal arm (B) on $25 \%$ of trials. In this case, switching away from the optimal arm does not represent an intentional attempt to lessen uncertainty in the environment, but instead represents participants employing a (somewhat crude) ratio choice rule.

Baum (1974) provided an extension to Herrnstein's (1961) "Simple" Matching Law to account for a wider array of choice behavior. This "Generalized Matching Law" included two additional parameters: bias and sensitivity, where bias reflects a tendency for selecting a given option over other available options (irrespective of the reinforcement rate for each option), and sensitivity determines how strictly a participant conforms to the choice ratio for their selections. The Generalized Matching Law has been shown to account for a wider variety of choice behavior than the Simple Matching Law (Baum, 1974; Schneider \& Lickliter, 2010), and is the version applied in this paper.

It is important to note that, even when employing a ratio choice rule like the Matching Law, participants can still update their knowledge of the environment by picking suboptimal responses (as determined by the choice rule). However, the crucial distinction is that exploratory choices occur on the basis of a ratio determined by the choice rule, and are not an intentional attempt to lessen uncertainty. For the purpose of the current paper, this type of behavior may be considered synonymous with the phenomenon known as probability matching (Sakai \& Fukai, 2008; Shanks, Tunney, \& McCarthy, 2002 - though strictly these two phenomena are slightly different, see Shanks et al., 2002).

The main difference between the uncertainty account and the ratio choice rule account of exploration is that, in the former, uncertainty is a catalyst for participants to explore (and thus lessen the total uncertainty in the task); while in the latter, exploration occurs as a product of some choice function. One key issue in attempting to differentiate these two accounts is that it is difficult to increase uncertainty without changing the reward value of the different arms. For example, in the commonly used "walking bandit task" (Daw et al. 2006; Laureiro-Martinez et al., 2015; Speekenbrink \& Konstantinidis, 2015), uncertainty is implemented by stochastically walking (slowly changing) the mean reward for each arm every trial. While this does serve to make the value of each arm uncertain, it also necessarily causes changes to the probability of picking each arm as given by ratio choice rules. Therefore, it is hard to determine whether an attempt to reduce the uncertainty in the task, or a predetermined ratio choice rule, is responsible for motivating exploration under these circumstances.

While it is possible to use cognitive modeling techniques to examine whether uncertainty motivates exploration (e.g.,

Daw et al., 2006; Knox et al., 2012; Speekenbrink \& Konstantinidis, 2015; Stevyers, Lee, \& Wagenmakers, 2009), the conclusions of these methods have been mixed. A recent study by Beesley et al. (2015) argued that attention may be another viable metric for assessing the EE Trade-Off. Beesley et al. conducted a study in which participants were presented with two cues and were asked to make a choice between two responses. One cue was informative about what the optimal response was on that trial, while the other cue was task-irrelevant. Beesley et al. measured participants' attention by tracking eye-gaze on the two cues. They showed that when cues were perfect predictors of the optimal response, participants attended to the informative cue over the task-irrelevant cue. However, when cues were imperfect predictors of the optimal response (i.e., predicted the optimal response on only two-thirds of trials), participants increased their attention to both the informative and task-irrelevant cue, indicating greater exploration of the cues. Beesley et al. argued that these findings were synonymous with the EE trade-off. The implication, therefore, is that exploration can be exhibited in behavioral domains outside of participant choice, and that exploration cannot be solely explained by ratio choice rules (the predictions of which are restricted to the response domain alone).

One limitation to the Beesley et al. task was that the experimenters provided participants with feedback about which response was optimal on each trial. Therefore, while participants appeared to explore the information in the task by altering their attentional processing, they had no incentive to explore different responses to find the one that was most optimal (as they were told which response was optimal regardless of their choice). Thus, the current paper aims to assess whether uncertainty can induce exploratory behavior in both participants' attention and responses, and hence provide wider support for the idea that uncertainty drives exploration. This would imply that exploration itself is perhaps a more complex, intentional process, which would place it outside the scope of ratio choice rules alone. Furthermore, it would suggest that exploration can manifest itself across more than one aspect of behavior (choice and attention). We used a four-armed bandit task where we manipulated uncertainty both within and between subjects. We measured responding and gaze-time to the different arms during the task. In line with our hypotheses, we found that participants made fewer optimal responses and spent longer fixating on task elements when the task had an element of uncertainty, suggesting that uncertainty acts as a catalyst for exploration.

## Method

This experiment aimed to examine the effect of uncertainty on attention and responding in a four-armed bandit task. Participants completed a variant of the bandit task, where on every trial they were presented four arms and asked to pick two. Two of the arms conferred 30 points (the High Value [HV] arms) and the remaining two arms conferred 15 points
(the Low Value [LV] arms). After making their choice, participants were rewarded with the cumulative score associated with each arm. For example, if the participant selected an HV arm worth 30 points and an LV arm worth 15 points, they received a reward of 45 points for that trial.

The experiment was conducted in two stages. Stage 1 of the task had a deterministic reward structure, in which each arm always yielded the same amount. Stage 1 was designed to be simple, such that participants could quickly learn the structure of the task and engage in what might be considered an exploitative pattern of behavior. In Stage 2 of the task, rewards were drawn stochastically from a uniform distribution for each combination of arms, with the mean reward value set at the same value as in the first stage. Stages 1 and 2 were coined the Certain and Uncertain stages respectively. In terms of participants' responding, we hypothesized that when rewards became uncertain, participants would make more exploratory, non-optimal responses, and this exploration would be greater for participants who experienced greater uncertainty (consistent with Knox et al., 2012; Speekenbrink \& Konstantinidis, 2015). In terms of gaze-time, we hypothesized that when rewards became uncertain in Stage 2, participants would increase their gaze-time to all arms in the task. Furthermore, we hypothesized that the greater the level of uncertainty in those rewards in Stage 2, the more gaze time that would be allocated to the arms in the task, with more gaze time to HV arms over LV arms (consistent with Beesley et al., 2015).

## Design

The design of the experiment is shown in Table 1. The key manipulation of the amount of uncertainty present in Stage 2 was manipulated between-subjects. Uncertainty was operationalized as the range of possible scores around the mean reward value that could be received following a trial (in Stage 2). For example, a reward distribution of $\pm 3$ (Low Uncertainty) meant that after the participant made their selections, they received the cumulative score of those arms (e.g., 45 points if they picked one HV arm and one LV arm), $\pm 3$ points (uniformly distributed across trials). Therefore, in this case, the participant could receive a score from 42 to 48. The Low Uncertainty condition had a reward distribution of $\pm 3$, and the High Uncertainty condition had a reward distribution of $\pm 18$. The crucial difference between these two conditions was that in the High Uncertainty condition, the two score distributions overlapped, such that participants could sometimes earn more points after a choice of an HV arm and an LV arm than after a choice of two HV arms. By comparison, for participants in the Low Uncertainty condition, the optimal response was always picking two HV arms. The dependent variables were proportion of HV arms picked, and gaze-time on HV and LV arms as a proportion of trial time.

Table 1: Design of Experiment 1

| Uncertainty <br> condition | HV <br> Reward | LV <br> Reward | Reward uncertainty <br> (Stage 2) |
| :---: | :---: | :---: | :---: |
| Low | 30 | 15 | $\pm 3$ |
| High | 30 | 15 | $\pm 18$ |

## Participants

Sixty-five UNSW Sydney undergraduate students were recruited in exchange for course credit. The two highest scoring participants received a $\$ 20$ prize.

## Apparatus and Materials

Participants were tested individually in a quiet room. During the task, participants' eye-gaze was tracked using a 58.4 cm widescreen Tobii eye-tracking monitor (TX-300). Participants were seated approximately 60 cm from the monitor, and had their heads steadied by a chin rest. The eye-tracker was calibrated at the start of the task. The experiment was run in MATLAB using the Psychophysics Toolbox extension (Brainard, 1997; Kleiner, Brainard, \& Pelli, 2007; Pelli, 1997). Participants mades all responses via a standard keyboard and mouse.

The four arms in the experiment were represented as four colored squares of 200 by 200 pixels (visual angle of approximately $5^{\circ}$ ). The four colors were always red, green, blue, and yellow (Figure 1). Color assignment to design elements (i.e., HV and LV arms) was counterbalanced between participants ( 24 permutations).

## Procedure

At the start of the experiment, participants were instructed that they would be playing a simple guessing game, where the objective of the game was to maximize the number of points they received. On each trial, the four colored arms were presented in the four quadrants of the screen. The location of each arm was counterbalanced between trials, with a full counterbalance of positions taking 24 trials. Participants used the mouse to select two arms. Participants were allowed to deselect arms they had selected by clicking on the arm a second time. Once the participant had selected two arms, a small "Submit" ( 120 by 60 pixels) button appeared in the center of the screen. If the participant selected more than two arms, or deselected an arm, the button disappeared.


Figure 1: A sample screen from Experiment 1.

Once the submit button was clicked, the four arms and the cursor disappeared, and the participant was told how many points they had earned on that trial, as well as the total points accumulated so far. Points were calculated by aggregating the value of the two arms the participant had selected, with the addition of the reward uncertainty in Stage 2 (see Design). Participants then pressed the spacebar to start the next trial. The location of the cursor was reset to the center of the screen on each trial.

Stage 1 consisted of 96 trials and Stage 2144 trials. The start of Stage 2 was not signaled to participants in any way. The only difference between Stage 1 and 2 was the addition of the variability (stochastic noise) for rewards (see Design and Table 1).

## Results

Data were collapsed into blocks of 24 trials for analysis. If a participant had less than $50 \%$ of trials with valid eyetracking data recorded, they were excluded from analysis ( $n$ $=10$ ). In addition, participants who selected the HV arms less than $70 \%$ of the time in the final block of Stage 1 were inferred to have not learnt the associations adequately, and were also excluded ( $n=7$ ). For each exclusion, we ensured a complete counterbalancing of design elements by recruiting a new participant with the same counterbalancing conditions. Trials in which the participant took two standard deviations longer than their mean trial time were excluded from all analyses.

Response data are shown in Figure 2 and were analyzed in three parts, Stage 1 (blocks 1 to 4 ), the between-stage transition period (blocks 4 and 5), and Stage 2 (blocks 5 to 10), using a repeated measures ANOVA with a withinsubjects factor of block and a between-subjects factor of condition. Effect sizes are reported as generalized etasquared, $\eta_{\bar{G}}^{\prime}$ (see Bakeman, 2005). In Stage 1, a significant effect of block was observed, $F(3,138)=98.84, p<.001$, $\eta_{G}=.447$, with participants in both conditions increasing selections of HV arms as they progressed through Stage 1.

During the transition from Stage 1 to Stage 2, a significant effect of block was observed, $F(1,46)=80.4, p<.001, \eta_{\bar{G}}^{2}$ $=.465$, with participants decreasing their selections of HV arms from Stage 1 to Stage 2. A significant effect of condition was also observed, $F(1,46)=18.06, p<.001, \eta_{\bar{G}}$ $=.165$, with participants less likely to choose the HV arms in the High Uncertainty group. Finally, there was a significant interaction between block and condition, $F(1,46)$ $=22.56, p<.001, \eta_{G}^{2}=.197$, with the proportion of HV arm choices showing a greater decrease in the high uncertainty conditions than the low uncertainty condition during the transition period.

In Stage 2, a significant effect of block was observed, $F(5$, 230) $=22.68, p<.001, \eta_{F}^{2}=.165$, with participants increasing their selections of HV arms over the course of Stage 2. A significant effect of condition, $F(1,46)=24.71$, $p<.001, \eta_{\bar{F}}^{2}=.243$, and a significant interaction between condition and block, $F(5,230)=5.02, p<.001, \eta_{G}^{2}=.042$, were observed, with participants picking HV arms less
frequently in the High Uncertainty condition, but also showing a greater increase in their selection of HV arms over the course of Stage 2, compared to participants in the Low Uncertainty condition.

Gaze-time data are shown in Figure 3. Gaze-time was calculated as the summed time of all fixations on the different arms in the task. A fixation was determined to have occurred if a participant's gaze did not deviate more than 75 pixels vertically or horizontally for at least 150 ms . Total fixation time was calculated by extending this time until the participant's gaze exited the 75 pixel limit (in accordance with Beesley et al., 2015; Salvucci \& Goldberg, 2000). Proportion of gaze-time was calculated as the total fixation on each arm divided by the total trial time. Again, these data were analysed using a repeated measures ANOVA, with a within-subjects factor of block, a within-subjects factor of arm value (high and low), and a between-subjects factor of condition. A significant effect of block was observed in Stage $1, F(3,138)=18.85, p<.001, \eta_{-}^{2}=.062$, with participants decreasing their total gaze-time to arms throughout Stage 1. A significant effect of arm value was also observed, $F(1,138)=276.77, p<.001, \eta_{F}^{2}=0.675$, along with a significant interaction between block and arm value, $F(3,138)=17.86, p<.001, \eta_{F}^{2}=.066$, with participants gazing more at HV arms than LV arms, and this difference increasing over the course of Stage 1.


Figure 2: Proportion of HV arms selected in each block. Stage 1 occurred in Blocks 1 to 4, while Stage 2 occurred in Blocks 5 to 10. Error bars represent $\pm 1$ SEM.


Figure 3: Proportion of trial time gazing at HV and LV arms in each block for each condition. Stage 1 occurred in Blocks 1 to 4, while Stage 2 occurred in Blocks 5 to 10. Error bars represent $\pm 1$ SEM.

In the transition from Stage 1 to Stage 2, there was a significant effect of block, $F(1,46)=15.80, p<.001, \eta_{\bar{G}}^{2}$ $=.039$, with participants increasing their total gaze-time at the onset of uncertainty. The significant effect of arm value was maintained, $F(1,46)=282.29, p<.001, \eta_{F}^{2}=.713$. There was no effect of condition observed on gaze-time, $F(1$, 46) $=2.21, p=.144$, and there was no interaction between block and condition, $F<1$. In Stage 2, the significant effect of arm value was maintained, $F(1,46)=362.42, p<.001$, $\eta_{G}^{2}=.721$, and no effect of condition was observed, $F<1$. No effect interaction between block and condition was observed in Stage 2, $F(5,230)=1.19, p=.3161$.

## Discussion

In a reinforcement learning task, participants earned points for combinations of responses. In Stage 1, one combination of responses was optimal and participants readily learnt this relationship. In Stage 2, we introduced variation in the number of points received, while keeping the mean number of points per response constant. When moving from the certainty of Stage 1 to the uncertainty of Stage 2, participants in both conditions reduced their rate of optimal responding, and this reduction was greater for participants who experienced greater uncertainty. Following this change in behavior at the outset of Stage 2, participants in both conditions increased their rate of optimal responding over the course of Stage 2.

In the High Uncertainty condition, the choice behavior of participants is well predicted by the Matching Law. However, crucially in the Low Uncertainty condition the Matching Law fails to predict the drop in optimal responding at the onset of. If a participant in the Low Uncertainty condition were following the Matching Law, they should not show a decrease in optimal responding at the onset of uncertainty. This finding suggests that exploratory choice cannot be solely explained by the Matching Law, and provides support to the idea that uncertainty can drive exploration. The reason why participants in the Low Uncertainty condition chose to switch away from the optimal response at the onset of uncertainty is not immediately clear. However, one possible explanation is that when participants perceived that the nature of the task had changed (i.e., rewards were no longer confined to three set values), they felt compelled to explore the other previously discounted responses to ensure they had not changed in any significant way.

In terms of the attentional data, we found support for two of our three hypotheses. Unsurprisingly, participants began to pay more attention to HV arms over LV arms over the course of the experiment. This is compatible with a host of research from the associative learning literature (See Le

[^539]Pelley et al., 2016, for a review), showing that participants are likely to direct their attention to the most valuable predictors in a task (also see Le Pelley et al., 2015). Furthermore, there is evidence that participants will attend more to arms they are intending to select prior to making their response (e.g., Manohar \& Husain, 2013). As participants selected more HV arms, this likely contributed to participants preferentially attending to them over LV arms.

Crucially, we have shown evidence that an onset of uncertainty is associated with an increase in attention. Once rewards became uncertain at the onset of Stage 2, participants in all conditions increased their gaze-time to all arms in the task. Our data are in line with the findings of Beesley et al. (2015), and provide support to the idea that uncertainty can instigate exploratory behavior in both the choice responses and attentional bias. We argue that these data are beyond the scope of ratio choice rules, which do not provide a natural account of attentional changes under conditions of uncertainty and would not predict changes in response rate across the course of Stage 2. While the notion that uncertainty increases attentional processing of stimuli is not novel (Pearce \& Hall, 1980), very little is known about attentional processing in multi-armed bandit tasks like the one used in the current experiment. The current findings suggest that pursuing this line of research may be important to gaining a more complete understanding of human decision-making.

However, we did not find evidence for gaze-time interacting with the level of uncertainty. If the uncertainty account of exploration is correct, we should have observed greater exploration under greater uncertainty. Instead, the amount of gaze-time participants paid to the arms was comparable under both levels of uncertainty. One possible reason for this is that moving from a completely certain environment to an environment with any level of uncertainty may cause attention to increase. Yu and Dayan (2005) showed that participants behave differently when uncertainty is expected (i.e., present for the entire task) compared to when uncertainty is unexpected (i.e., a period of uncertainty occurs suddenly, following a period of certainty). It may be the case that when unexpected uncertainty occurred, attention increased by a set amount in response (regardless of the degree of that uncertainty). Also, while gaze-time does appear to be affected by uncertainty, the effect-size in our study was much smaller in comparison to the effect of uncertainty on responding $\left(\eta_{G}^{2}=.039\right.$ compared to $\eta_{G}^{2}=.465$ ). This may suggest that while changes in response rate and changes in overt attention are signals of exploration under uncertain conditions, that uncertainty affects these behavioral markers by distinct mechanisms. Alternatively, these data might suggest that gaze-time was less sensitive to uncertainty than was
participants' responding, which made it harder to detect any effect of the different levels of uncertainty and gaze-time.

In summary, we have shown that the introduction of uncertainty into a four-armed bandit task caused a general increase in attending, and a decrease in optimal responding. This provides support for the idea that environmental uncertainty causes an increase in exploratory behavior, and challenges the idea that exploration can be explained purely by ratio choice rules.

## Acknowledgements

This research was supported by an Australian Government Research Training Program (RTP) Scholarship, and by Australian Research Council grant DP140103268, awarded to Tom Beesley, Mike Le Pelley, and Chris Mitchell. The authors thank Ben Newell and Fred Westbrook for their help.

## References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. Behavior Research Methods, 37, 379-384.
Beesley, T., Nguyen, K. P., Pearson, D., \& Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. Quarterly Journal of Experimental Psychology (Hove), 68, 2175-2199.
Brainard, D. H. (1997). The psychophysics toolbox. Spatial Vision, 10, 433-436.
Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching, Journal of the Experimental Analysis of Behaviour, 22, 231-242.
Cohen, J. D., McClure, S. M., \& Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362, 933-942.
Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., \& Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441, 876-879.
Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B (Methodological), 148-177.
Gold, J. I., \& Shadlen, M. N. (2007). The neural basis of decision making. Annual Review of Neuroscience, 30, 535-574.
Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. Journal of the Experimental Analysis of Behavior, 4, 267272.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., \& Broussard, C. (2007). What's new in Psychtoolbox-3. Perception, 36.
Knox, W. B., Otto, A. R., Stone, P., \& Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. Frontiers in Psychology, 2, 398.

Laureiro-Martínez, D., Brusoni, S., Canessa, N., \& Zollo, M. (2015). Understanding the exploration-exploitation dilemma: An fMRI study of attention control and decision-making performance. Strategic Management Journal, 36, 319-338.
Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., \& Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. Psychological Bulletin, 142, 1111-1140.
Le Pelley, M. E., Pearson, D., Griffiths, O., \& Beesley, T. (2015). When goals conflict with values: Counterproductive attentional and oculomotor capture by reward-related stimuli. Journal of Experimental Psychology: General, 144, 158-171.
Manohar, S. G., \& Husain, M. (2013). Attention as foraging for information and value, Frontiers in Human Neuroscience, 62-77.
Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., \& Gonzalez, C. (2015). Unpacking the explorationexploitation tradeoff: A synthesis of human and animal literatures. Decision, 2, 191-215.
Pearce, J. M., \& Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review, 87, 532-552.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.
Sakai, Y., \& Fukai, T. (2008). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. Neural Computation, 20, 227-251.
Salvucci, D. D., \& Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. Proceedings of the Symposium on Eye Tracking Research \& Applications - ETRA '00, 71-78.
Schneider, S, M., \& Lickliter, R. (2010), Choice in quail neonates: The origins of the generalized matching law Journal of the Experimental Analysis of Behaviour, 94, 315-326.
Shanks, D. R., Tunney, R. J., \& McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. Journal of Behavioral Decision Making, 15, 233250.

Speekenbrink, M., \& Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. Topics in Cognitive Science, 7, 351-367.
Steyvers, M., Lee, M. D., \& Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology, 53, 168179.

Sutton, R. S., \& Barto, A. G. (1998). Reinforcement Learning: An Introduction (Vol. 1): MIT press Cambridge. Yu, A. J., \& Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron, 46, 681-692.

# Actively Detecting Patterns in an Artificial Language to Learn Non-Adjacent Dependencies 

Hao Wang<br>University of Pennsylvania

Jason Zevin<br>University of Southern California<br>Calvin Leather<br>University of Southern California<br>Toby Mintz<br>University of Southern California


#### Abstract

Many grammatical dependencies in natural language involve elements that are not adjacent, such as between the subject and verb in "the dog always barks". We recently showed that non-adjacent dependencies are easily learnable without pauses in the signal when speech is presented rapidly. In this study, we used an online measure to look at the relationship between online parsing and the learning performance from the offline assessment of non-adjacent dependency learning. We found that participants who showed current parsing of the language online also learned the dependencies better. However, this pattern disappeared when they are explicitly told where the boundaries are before parsing. Theories of non-adjacent dependency learning are discussed.


# A Plausible Micro Neural Circuit for Decision-Making 

Wei Hui, Dai Dawei, BuYijie (weihui@fudan.edu.cn, dwdai14@fudan.edu.cn)<br>Laboratory of Cognitive Model and Algorithms, Department of Computer Science, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China


#### Abstract

An intermediate level between neural circuits and behaviors is neural computations, various behaviors that animals exhibit following some basic control laws can be implemented by some canonical neural computations [Carandini, 2012]. To explore how the microscopic activity of neurons leads to macroscopic behavioral control strategy, we consider basic logic-like operations as some canonical computations in the brain. In this paper, firstly we designed the functional circuits for basic logic-like operations based on the known neurophysiological properties. Secondly, using basic functional circuits constructed a possible neural network for decision logic of animal's behavior. This study provides a general approach for constructing the neural circuits to implement the behavioral control rules. Furthermore, this study will help us to establish a transitional bridge between the microscopic activity of the nervous system and the macroscopic animal behavior.


Keywords: Neural circuits; Logic; Neural computations;

## Introduction

Brain as a complex system, three distinct levels should be understood, i.e., behavior level, algorithm level and implementational level, which is famously known as Marr's tri-level hypothesis [Marr 1982]. The benefit of this clear distinction is that researchers can focus on a certain level and do researches purposefully. In [Carandini, 2012], the brain was analogized to a computer, as we know, all applications in the computer can be reduced to the most primitive operations (Logic instructions) of CPU, so is the brain. Researches indicate that the brain deal with different problems by combining and repeating a core set of canonical neural computations [Carandini \& Heeger, 2011]. We understand the every detail of instructions, which were implemented in CPU; however, we know little about the details of circuit's constitution in brain. However, without a clear link to behavior and computational mechanism, it is hard to understand what is computed. Therefore, "We need a foundational mechanistic, computational framework to understand how the elements of the brain work together to form functional units and ultimately generate the complex cognitive behaviors" [Brown, 2014].

Obviously, understanding the canonical computations in the brain is helpful to reveal the computational framework from circuit to behavior. In this paper, we consider the basic logical operations as some kind of canonical computations in nervous system. Why can the logical operations be considered a kind of canonical computations in nervous system? We make the rational reasoning from the computational perspective, logic reflects the most basic requirement that any computation can be successfully implemented. Thus, the rules through which animals control their behavior can be described by logic language. In order for a biological nervous system to achieve a specific computation, its structure must be sufficiently complex to achieve the basic logic operations. Therefore, there must be many types of neural circuits to achieve various logical rules in the nervous system. Since, any type of behavioral logic can be formally described by propositional logical. With this reliable and complete formal language, we can describe the basic control rules accurately, with which behaviors comply. Furthermore, with different firing mode of neurons and the synergistic connections between pyramidal neurons and intermediate neurons, how does the nervous system assemble a circuit to achieve a set of specific logical rules?

The aim of our work is not to construct the neural network to achieve the logic operations. In this paper, we attempt to explore computational framework how the microscopic neural activities can systematically explain the macroscopic behavior from the logic view.

## Related works

Research indicates that the brain relies on a core set of computations to apply different functions for different problem [Carandini \& Heeger, 2011]. Neural computations, which occur in populations of neurons, are a transitional level from circuit to behavior. Although, some computations have been discovered in nervous system, there are no details of such circuits' constitution. In order to reveal the true mechanism of nervous system the research works involve in different field. Table1 lists the related works.

Table1. List of related works

| Category |  | Sub-category | Attributes |
| :---: | :---: | :---: | :---: |
|  | Numerical modeling | MP model, BP model, CNN, RBM; <br> [McCulloch,1943; Rumelhart et al,1986; Fischer \& Ige, 2012; ] | Limited function approximation; <br> . Violating basic biological facts *; |
|  | Spike modeling | HH model[Hodgkin \& Huxley, 1952] HR [Hindmarsh\& Rose, 1984] | . Good biological plausibility; <br> . Low efficiency;[Izhikevich, 2004] |
|  |  | A simple Spike model [Izhikevich, 2003, 2004] | . Good biological plausibility; <br> . High efficiency; [Izhikevich, 2003, 2004] |
|  |  | A cortical simulator [Aanthanarayanan \& Modha, 2007] | . Coarse clique-level simulation; <br> . No certain behavior interrelated to; |


|  |  | Model of thalamocortical systems [Izhikevich \& Edelman, 2008] |  | . Good biological plausibility; <br> . No certain behavior interrelated to; |
| :---: | :---: | :---: | :---: | :---: |
| For Physiological Purpose | Sensor-motor circuit | Circuits for C.elegans behaviors: | Using ANN to construct [Fer'ee et al, 1996, 1999]; | .Circuit in ANN-mode is of poor biological plausibility; |
|  |  |  | Using DNN to construct [Jian-Xin \& Xin, 2013] | .Moderate biological plausibility; .No biological neuron was used; |
|  | Reusable and combinable primitive circuit | Canonical neural computations | Linear filtering; <br> Divisive; Normalization; <br> Thresholding; <br> [Carandini \& Heeger, 2011; Wang, 2002; <br> Carandini, 2005, 2012;] | .Hypothesis on functionalism-level, not on implementation level .No constitution details of circuit; |
|  |  | Modulators of [Kenji Doya, 2000, | Decision-making 008] | . Good biological plausibility; <br> . No detail constitution of circuit; |
|  | Decisionmaking circuit | Model of two[Ratcliff \& Rouder, | hoice decisions 998] | . Less biological details; <br> . Numerical approximation only; |
|  |  | Probabilistic m [Wang, 2002; Wei | odel for decision making Dai \& Bu, 2017; Wei \& Bu \& Dai, 2017] | . Good biological plausibility; <br> . Matching behaviorism data; <br> . Statistical abstraction on group-level neural activities. |

* Violating basic biological facts includes:(1) the activation mode of the MP model is two-valued, but that of biological neuron is impulse-firing; (2)the type of ANN's neuron is unitary, however, in the biological neural system, not only multiple types of neurons exist but also their proportion matters; (3)the numerical settings of threshold and connection weights of ANN being able to adjust at will are too idealistic; (4)numerical neurons in the same layer working with perfect synchronization are too idealistic, however, time differences of signal transmitting are more general.


## Biological neuron

## Neuron Model

Izhikevich proposed a simple spiking neuron model that reduces the HH model to a 2-D system [Izhikevich, 2003]. Ordinary differential equations are of the form:

$$
\begin{align*}
& \frac{d v}{d t}=0.04 \mathrm{v}^{2}+5 \mathrm{v}+140-\mathrm{u}+\mathrm{I}  \tag{1}\\
& \frac{\mathrm{du}}{\mathrm{dt}}=\mathrm{a}(\mathrm{bv}-\mathrm{u}) \\
& \text { If } \mathrm{v} \geq 30, \text { Then }\left\{\begin{array}{c}
\mathrm{v} \leftarrow \mathrm{c} \\
\mathrm{u} \leftarrow \mathrm{u}+\mathrm{d}
\end{array}\right.
\end{align*}
$$

Interpretation of parameters refers to [Izhikevich, 2003]. In the paper, typical values of parameters for excitatory neuron were: $a=0.02, b=0.25, c=-65, d=8$. Average firing rate (AFR) of pyramidal neuron was between 0 and 21 Hz . Typical values of parameters for inhibitory neuron were: $a$ $=0.1, b=0.2, c=-55 \sim-48, d=2$. AFR of intermediate neuron was between 0 and 200 Hz .

## Time delays in AP transmission

Delay means the time of AP propagating from pre-synaptic neurons to post-synaptic neurons [Tolnai et al, 2009]. A wide range of time delays (up to 20 ms ) could occur. Since most previous studies did not relate to specific behavioral control logic, which was easy to ignore. In fact, the duration from when the AP is generated to its arrival at the postsynaptic neuron is time-critical or time-sensitive. In this paper, the different delays of AP transmission may be similar to "time multiplexing" in signal processing, which plays an important role in behavioral decision logic.

In this paper, we simulated the propagation delays of AP using different queue lengths. For example, using four different queue lengths, as shown in Fig. 1 (b, Queue 1~4), simulated the different delays of AP propagating from the cell body to positions $1 \sim 4$ in Fig. 1(a). If the length of a
queue is $\boldsymbol{n}$, then the AP is delayed $\boldsymbol{n}$ milliseconds. Four queues with sequential increases in length indicated that as the location of the synapse on the axon moved away from the cell body, the delays increased. If an AP was generated in the pre-synaptic neuron, we added 1 to the head of the queue; otherwise, we added 0 . When the end of queue element was 1 , it indicated that the postsynaptic neuron received an AP. Delays of single neuron were limited; if large delays are required in the nervous system, Fig. 1(c) presents a possible way.


Figure 1. Simulation of the delays in AP transmission along an axon using queues.

## The firing rate of pyramidal neurons is adjustable

A study has indicated that intermediate neurons participate in regulating the firing rates of neural networks [Sanders, et al, 2013]. Fig. 2 shows a possible way of implementation that could achieve this regulation of AFR in the nervous system. This cooperative activity in which excitatory neurons and inhibitory neurons regulate the AFR of downstream neurons is a basic mechanism through which nervous systems function.


Figure 2. AFR of downstream neurons can be regulated by different combinations of upstream excitatory and inhibitory synapses.

In Fig. 2, if pyramidal neuron E received AP with a stable AFR from upstream excitatory neurons (Eneus), then the AFR of $E$ could be regulated successfully by increasing or decreasing the firing rate of upstream inhibitory neurons (Ineus), as shown in Fig. 3. Table 2 shows changes in the range of neuron E's AFR with changes in the AFR of upstream Eneus and Ineus. This basic law revealed that nervous systems could regulate output firing rate through a precise configuration of types of neurons and connections.


Figure 3. (a) The AFR of neuron E increased with a decrease in the AFR of upstream Ineus. (b) The AFR of E decreased with an increase in the AFR of upstream Ineus.

Table2. Regulating the AFR of neurons within a certain range.

| AFR of Eneus | $17 \sim 19 \mathrm{~Hz}$ |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :---: |
| AFR of Ineus | $0 \sim 50 \mathrm{~Hz}$ | $50 \sim 100 \mathrm{~Hz}$ | $100 \sim 150 \mathrm{~Hz}$ | $150 \sim 200 \mathrm{~Hz}$ |  |
| AFR of E | $16 \sim 18 \mathrm{~Hz}$ | $13 \sim 16 \mathrm{~Hz}$ | $9 \sim 13 \mathrm{~Hz}$ | $3 \sim 9 \mathrm{~Hz}$ |  |
| AFR of Ineus | $75 \sim 80 \mathrm{~Hz}$ |  |  |  |  |
| AFR of Eneus | $17 \sim 19 \mathrm{~Hz}$ | $12 \sim 16 \mathrm{~Hz}$ | $8 \sim 12 \mathrm{~Hz}$ | $5 \sim 8 \mathrm{~Hz}$ |  |
| AFR of E | $14 \sim 16 \mathrm{~Hz}$ | $12 \sim 16 \mathrm{~Hz}$ | $8 \sim 12 \mathrm{~Hz}$ | $5 \sim 8 \mathrm{~Hz}$ |  |

## Neural circuit designs for logic-like operations

AP from pre-synaptic neurons can produce excitatory post-synaptic potentials (EPSP) or inhibitory postsynaptic potentials (IPSP). Since, single AP generates too small EPSP or IPSP to activate or inhibit the postsynaptic neurons; we assume that a train of at least 40 AP could activate the postsynaptic neuron. We employ a group of neurons (neuron cluster) that included 50~100 neurons as a functional unit, which is used to construct circuits to achieve the basic logic-like operations. Since, any of the complex logic can be expressed as a logical expression by four basic logical operations: And, Or, Negation, and Conditional. We implement four circuits that are equivalent to the function of these basic logic operations. The circuits contain excitatory neurons and inhibitory neurons.

In the paper, when a constant stimulus 7.5 adding background noise is presented to a neuron cluster, AFR of cluster is higher than 10 Hz ; while a constant stimulus 3.8 adding background noises is presented, AFR of cluster is lower than 5 Hz . If the AFR of a neuron cluster is higher than 10 Hz , then the proposition expressed by the neuron
cluster is True; if the AFR of a neuron cluster is lower than 7 Hz , then the propositions is False.

## And-like operation circuit

As we know that the concept of the neocortex is as an assemblage of the basic functional units [Jean-Vincent Le Bé, 2007]. Neurons in the fourth layer accept the external signal input from the afferent fibers (area-b in Fig. 4). Small pyramidal cells and intermediate neurons in the second and third layers are responsible for processing the signal (area-a in Fig. 4). In the fifth layer, large pyramidal cells are responsible for propagating the "results" out of the cerebral cortex (area-c in Fig. 4). Axons are shown in black and dendrites are shown in blue.


Figure 4. Morphological principles of connectivity between neocortical neurons (Corresponding to [Jean-Vincent Le Bé, 2007]).

And-like operation is equivalent to that upstream neuron clusters A and B both fire AP at a high rate, followed by neuron cluster C firing at a high rate; otherwise, C fires at a low rate. As shown in Fig. 5-Left, neurons in A and B full connect to neurons in C . A and B represent two propositions, and C achieves the function of operation "A And-like B ". As shown in Fig.5-Right, A and B (corresponding to clusters A and B in Fig.5-Left) that represent the incoming information should be distributed in the fourth layer of the neocortex. C (corresponding to C in Fig.5-Left) that achieves the computation of the And-like operation for A and B should be distributed in the second and third layer. At last, the processing results are propagated out of neocortex by the large pyramidal cells in the fifth layers. We re-layout the Fig.5-Left and obtain the Fig.5-Right. The new circuit satisfies the anatomical discoveries and achieves the logic function. It is a feasible implementation in neurobiology.


Figure 5. Circuit of And-like operation
AFR of A and B are stable due to the stable input. Neuron cluster C receives AP from A , and the time span is so long (about 20 ms ) that the number of AP is small at any given
moment. The distribution property of AP from neurons in cluster B is similar to that of A. If neuron clusters A and B both fire at a high rate, AP trains from A and B at least partially overlap. The purpose of this design is that when only one of the two neuron clusters fires at a high rate, the strength of the subsequent EPSP is too weak to activate C fire at a high rate. However, when A and B both fire at a high rate, due to the overlap of EPSP, the strength of the EPSP is sufficiently strong that C fires at a high rate as well.

However, we found that C would not fire with a high rate every time during the experiment. The EPSP from A and B does not necessarily overlap because the overlap is timecritical or time-sensitive. Thus, it is possible that C fires at a low rate even if both A and B fire at high rate. To avoid such a situation, one feasible way that we used neuron clusters as functional units, and the properties of neurons in a cluster are different, including the model parameters and AP delays. Therefore, initiation of neuronal firing is asynchronous. As a result, EPSP always can be overlapped in C. When A and B both fire at high rate, and C fires with a high rate. Typical values for the delays of neuronal AP are 1 $\mathrm{ms}, 2 \mathrm{~ms}, \ldots 20 \mathrm{~ms}$ in A and B; each delay has the same number of neurons, and the model parameters of each neuron is little different. As shown in Fig. 6, only when neuron clusters A and B both fire at a high rate, does C fire at a high rate [Fig. 6(d)]; otherwise, C fires at a low rate [Fig. 6(a), (b), and (c)]. This circuit performs the function of Andlike operation.


Figure 6. AFR of the And-like operation circuit

## Or-like operation circuit

Or-like operation is equivalent that if at least one of upstream neuron clusters A and B fires AP at a high rate, then C fires at a high rate; otherwise, C fires at a low rate. The structure in Fig5 can also achieve the function of Orlike operation through modifying the parameters to make sure that APs from neuron clusters A and B are synchronous and concentrated, and when one of the two clusters fires at a high rate, at least 40 AP have reached C at one given moment. Typical values for the delays of the neuronal AP are 1 ms for A and 5 ms for B . The purpose of this design is that when at least one of the two neuron clusters (A, B) fires at a high rate, the strength of subsequent EPSP is
sufficiently strong to make C fire at a high rate [as shown in Fig. 7(b), (c), and (d)]. Only when neuron clusters A and B both fire at a low rate, C fires at a low rate [as shown in Fig. 7(a)]. This circuit performs the function of "A Or-like B".


Figure 7. AFR of the Or-like operation circuit

## Conditional-like operation, Negation-like operation circuit

Conditional-like operation is equivalent to a simple projection relationship from upstream to downstream neurons, such that if upstream neurons fires with a high rate, then downstream neurons fires with a high rate; otherwise, downstream neurons fire at a low rate. Negation-like operation is equivalent that if A fires AP at a high rate, then C fires with a low rate; otherwise, C fires with a high rate. As shown in Fig.8-left, neurons in A fully connect to neurons in $I_{1} \ldots I_{m}, E_{1} \ldots E_{k}$ and neurons in $I_{1} \ldots I_{m}, E_{1} \ldots E_{k}$ full connect to neurons in C . In addition, for local circuits: $\mathrm{E}_{1}$ Conditional-like C, and $\mathrm{E}_{\mathrm{k}}$ Conditional-like C. Neurons in $\mathrm{I}_{1} \ldots \mathrm{I}_{\mathrm{m}}$ are all inhibitory, and the others are excitatory. A represents a proposition, and C performs the function of operation "Negation-like A". Its possible form in neocortex is shown in Fig. 8-right.


Figure 8. Circuit of Negation-like operation.
The circuit contains excitatory neurons and inhibitory neurons. When neuron cluster A fires at a high rate, which activate intermediate neuron firing at $60 \sim 110 \mathrm{~Hz}$, and the AP are asynchronous, which led to neuron cluster C receiving nearly continuous IPSP. Thus, neurons in C are inhibited, and C could not fire at a high rate. When A fires at a low rate, the intermediate neurons fire less than 60 Hz , which could not inhibit the activity of downstream neurons. Here, the excitatory signal that from A activated C with "time
division multiplexing" by neuron clusters $E_{1}, E_{2}$, and $E_{3}$ (m $=3, \mathrm{k}=3$ ). Thus, the AFR of C is about 3 times greater than that of A. Typical values of delays from A to $I_{1}, I_{2}$, and $I_{3}$ were $1 \mathrm{~ms}, 10 \mathrm{~ms}$, and 20 ms , and AP delays from A to neuron clusters $E_{1}, E_{2}$, and $E_{3}$ are $20 \mathrm{~ms}, 40 \mathrm{~ms}$, and 60 ms , respectively. AP delays from $\mathrm{I}_{1}, \mathrm{I}_{2}$, and $\mathrm{I}_{3}$ to neuron cluster $C$ are 1 ms , and AP delays from $E_{1}, E_{2}$, and $E_{3}$ to $C$ are 30 ms . Each delay had about the same number of neurons. The above settings are not absolute. The circuit performs the transfer of a low firing rate to a high firing rate [Fig. 9(a)], and a high firing rate to a low firing rate [Fig. 9(b)].


Figure 9. AFR of the Negation-like operation circuit

## Constructing a neural network for a specific behavior based on the logic-like operations

Any of the propositional logical expressions can be transferred into an equivalent conjunctive or disjunctive paradigm. Thus, the nervous system could possibly perform a logic function by realizing the corresponding paradigm.

We demonstrated the implementation of a neural circuit for decision-making logic using a rat's behavioral decision. In the behavioral experiment [Yang et al., 2014], the rat was trained to go to alternate arms of a Y-maze for drinking, and after the training, the rat never made a mistake to the same side two times as shown in Fig. 10. This experiment verified that the rat formed a set of accurate rules for decision making (turning left or right), which depended on the information that the side from which the rat obtained the last drink, and whether the rat reached the neck of the Y-maze.


Figure 10. Behavior decision experiment of rat (Corresponding to [Yang et al, 2014]).

We outlined the behavior in a set of logical expressions: Drink_L, the rat last drank on the left side of the Y-maze; Drink_R, the rat last drank on the right side; Thirsty, the rat
was in a thirsty state; At_Neck, the rat reached the neck of the Y-maze; Turn_L, the rat made a decision with turning left; Turn_R, the rat made a decision with turning right. The decision logic could be described such that the rat was thirsty, and working memory retained the left (or right) side of the Y-maze from which the rat last drank. Then, when the rat reached the neck of the Y-maze again, it executed the command of turning right (or turning left). The process of decision logic could be expressed by a proposition logical expression: (Thirsty $\wedge$ Drink_L $\wedge$ At_Neck $\rightarrow$ Turn_R) $\vee$ (Thirsty $\wedge$ Drink_ $R \wedge A t_{-}$Neck $\rightarrow$ Turn_L). A plausible neural network could achieve this decision logic, as shown in Fig. 11.


Figure11. Neural circuit for rat's decision-making.
Taking the rat executing the turning-right command as an example, the details of the circuit were such that: First, four neuron clusters represented the four propositions, Thirsty, Drink_L, At_Neck, and Turn_R. If the AFR of a neuron cluster was higher than 10 Hz , then the corresponding proposition was true; otherwise it was false. Second, we constructed the neural circuit for the logical expression: Thirsty $\wedge$ Drink_L based on the And-like circuit. Third, we constructed the neural circuit for the logical expression: Thirsty $\wedge$ Drink_L $\wedge$ At_Neck based on the And-like circuit. Finally, we constructed the circuit (Thirsty $\wedge$ Drink_ $L \wedge$ At_Neck) $\rightarrow$ Turn_R based on the Conditional-like circuit. In addition, we designed two groups of intermediate neurons (I3 and I4) between Turn_R and Turn_L to avoid misuse; Drink_ $L$ and Drink_ $R$ were also mutually exclusive, if the two propositions were both true, the decision making would be disordered. When the rat was in a given status, the neuron cluster that expressed the opposite status was inhibited. The complete circuit for decision-making is shown in Fig. 11.

We simulated the process of decision-making for a rat in a Y-maze. As shown in Fig. 12, (L-Choice) If the rat last drank at the right side of the Y-maze (Drink_R=True), then when the rat reached the neck of the Y-maze (At_Neck= True) it executed the command turning-left (Turn_L=True); otherwise, the rat executed the command turning-right (Turn_L=False). (R-Choice) If the rat last drank at the left side of the Y-maze, then when the rat reached the neck of Y-maze, rat executed the command turning-left; otherwise, the rat executed the command turning-right.


Figure 12. Decision-making in Y-maze

## Conclusion

Finally, we summarize our work through Marr's threelevel hierarchy. (a) What is computed? Our answer is that some logical rules are computed. Modeling from this perspective can help us to understand the functional base line of it. (b) Why is it computed? For sake of accurate behavior-controlling, these logical rules must be computed, which is the fundamental demand to a specific behavior. (c) How is it computed? In this paper, we design some types of local neural circuits to achieve four basic logic-like operations as canonical computations and assemble them to simulate a rat's decision making behaviors in Y-maze. Firstly, our circuit design is highly faithful to neurobiological facts like neuron firing mode, two major types of neuron, the proportion constrain of their numbers, and pulse-based mode of communication. Secondly, in the scope of cortical column our logicalequivalent local circuits are biologically plausible to be implemented. Thirdly, these basic functional modular are configurable, reusable and combinable.

We lack a bridge theory from circuit to behavior [Carandini, 2012]. For example, how do microscopic activities of neurons and logical relationships in circuits support the achievement of cognitive ability? Our aim is to construct a biological neural network for behavioral control rules from a logic perspective. This study may be useful for gradually transitioning from microscopic neural activity to macroscopic behavioral control. In our future works, we will explore neural computational mechanism about how a proper circuit is formed.

## Acknowledgments

This work was supported by the NSFC project (Project No.61375122), and in part by Shanghai Science and Technology Development Funds (13dz2260200, 13511504300).

## References

Carandini M. From circuits to behavior: a bridge too far? Nature neuroscience.2012; 15(4):507-509.
Carandini M, Heeger D J. Normalization as a canonical neural computation[J].Nature Reviews Neuroscience,2012,13(1): 51-62.
Brown JW. The tale of the neuroscientists and the computer: why mechanistic theory matters. Frontiers in neuroscience. 2014; 8: 349.

Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". Nature 323 (6088): 533-536.
Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of physiology. 1952; 117(4): 500.
Hindmarsh J, Rose R. A model of neuronal bursting using three coupled first order differential equations. Proceedings of the Royal Society of London B: Biological Sciences. 1984; 221(1222):87-102.
Izhikevich EM, et al. Simple model of spiking neurons. IEEE Transactions on neural networks. 2003; 14(6):1569-1572.
Izhikevich EM. Which model to use for cortical spiking neurons? IEEE transactions on neural networks. 2004; 15(5): 1063-1070.
Ananthanarayanan R, Modha D S. Anatomy of a cortical simulator[C]. Proceedings of the 2007 ACM/IEEE conference on Supercomputing. ACM, 2007: 3.
Izhikevich EM, Edelman GM. Large-scale model of mammalian thalamocortical systems. Proceedings of the national academy of sciences.2008; 105(9):3593-3598.
Ferr'ee, T.C., \& Lockery, S.R. (1999). Computational rules for chemotaxis in the nematode C.elegans. Journal of Computational Neuroscience, 6(3), 263-277.
Ferr'ee, T.C., Marcotte, B.A., Lockery, S.R. (1996). Neural network models of chemotaxis in the nematode Caenorhabditis elegans. Advances in Neural Information Processing Systems, 9, 55-61.
Xu J X, Deng X. Biological modeling of complex chemotaxis behaviors for C. elegans under speed regulation-a dynamic neural networks approach[J]. Journal of computational neuroscience, 2013, 35(1): 19-37.
Wang, X.J. Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36, 955-968 (2002).
Wei, H., Dai, D. \& Bu, Y. Cogn Neurodyn (2017). doi: 10.1007/s11571-017-9426-4

Wei, H., Bu, Y. \& Dai, D. Cogn Neurodyn (2017). doi: 10.1007/s11571-017-9436-2

Carandini, M. et al. Do we know what the early visual system does? J. Neurosci. 25, 10577-10597 (2005).

Le Bé J V. Structure and dynamics of the neocortical microcircuit connectivity[J]. 2007.
Tolnai S, Englitz B, Scholbach J, et al. Spike transmission delay at the calyx of Held in vivo: rate dependence, phenomenological modeling, and relevance for sound localization[J]. Journal of neurophysiology, 2009, 102(2): 1206-1217.
Sanders, Z. Josh Huang \& Adam Kepecs. (2013). Cortical interneurons that specialize in disinhibitory control. Nature, vol 503, 521-524
Yang, S. T., Shi, Y., Wang, Q., Peng, J. Y., \& Li, B. M. (2014). Neuronal representation of working memory in the medial prefrontal cortex of rats. Molecular brain, 7(1), 1.

# The Cognitive Reflection Test: familiarity and predictive power in professionals 

Matthew Brian Welsh ${ }^{1,2}$ \& Steve H. Begg ${ }^{1}$<br>(\{matthew.welsh\}\{steve.begg\} @adelaide.edu.au)<br>1. Australian School of Petroleum, 2. School of Psychology<br>University of Adelaide, North Tce, Adelaide, S.A., Australia


#### Abstract

The CRT is an increasingly well-known and used test of bias susceptibility. While alternatives are being developed, the original remains in widespread use and this has led to its becoming increasingly familiar to psychology students (Stieger \& Reips, 2016), resulting in inflated scores. Extending this work, we measure the effect of prior exposure to the CRT in a sample of oil industry professionals. These engineers and geoscientists completed the CRT, seven bias tasks and rated their familiarity with all of these. Key results were that: familiarity increased CRT scores but tended not to reduce bias susceptibility; and industry personnel, even without prior CRT exposure, scored very highly on the CRT greatly reducing its predictive power. Conclusions are that the standard CRT is not a useful tool for assessing bias susceptibility in highly numerate professionals - and doubly so when they have previously been exposed.


Keywords: cognitive reflection test; familiarity; predictive power; bias; industry professionals.

## Introduction

The Cognitive Reflection Test (CRT), due to its impressive predictive power for biases (Frederick, 2005; Toplak, West, \& Stanovich, 2011), is widely used in bias research. Perhaps its most recognisable item is the following:

A bat and a ball together cost $\$ 1.10$. The bat costs $\$ 1$ more than the ball. How much does the ball cost?
This question and its two companions have strongly intuitive but incorrect answers - 10c in the bat-and-ball question's case - such that answering the questions correctly implies greater reflection on one's answer. Thus, the CRT yields a score from $0-3$ with higher values reflecting greater 'cognitive reflection', which has been linked to lessened bias susceptibility.
Despite the CRT's success, concerns have been raised about it. Firstly, it conflates numerical ability with measurement of decision style (see, e.g., Primi et al., 2015; Weller et al., 2013; Welsh, Burns, \& Delfabbro, 2013); and, secondly, consists of only three, quite memorable items.

While these problems have been previously noted and attempts made to improve the CRT by inclusion of additional items and attempts to reduce the mathematical emphasis (Primi et al., 2015; Thomson \& Oppenheimer, 2016; Toplak, West, \& Stanovich, 2014), the original CRT, due to its speed and ease-of-use, remains in widespread use.

This is problematic in that, once a person has been exposed to the CRT, its usefulness may be compromised. Recent work by Stieger and Reips (2016), for example, has shown that familiarity with CRT questions inflated CRT scores amongst psychology students.

Key questions remain, however. First, whether CRT familiarity extends beyond psychology students to people in industries interested in bias reduction strategies. For example, the oil and gas industry has a 4 decade long history of following the judgement and decision making literature - beginning with Capen's (1976) work on overconfidence. With the success of popular science books like Kahneman's (2011) Thinking Fast and Slow, which are often taken up by managers, it seems likely that industry knowledge of the CRT will also expand. This could render it decreasingly useful as a means of distinguishing between individuals because people who have undertaken decision making training may be increasingly likely to have encountered the CRT or similar questions before.

The second question relates to the degree of familiarity required to undermine the CRT's validity. For example, the above bat-and-ball problem is memorable. Its format, however, is nearly as memorable. That is, assume someone has seen the bat-and-ball question; when, then, asked:
A jug and a cup together cost $\$ 2.20$. The jug costs $\$ 2$ more than the cup. How much does the cup cost?

It seems unlikely that anyone would fail to make the connection between the two. That is, despite not having seen the specific question before, recollection of the question format could be sufficient to prime them for reflection on their answer. This would result in them scoring higher on the CRT - not due to superior cognitive reflection but simple familiarity. Given the low score 'ceiling' of the 3 -item CRT, this could reduce the CRT's ability to predict susceptibility to biases by truncating its range of scores.

## Hypotheses

1. Decision making training courses will increase familiarity with bias questions and CRT.
2. Familiarity will inflate CRT scores.
3. This will reduce the CRT's predictive power.
4. Familiarity will increase bias resistance.

## Method

## Participants

Participants were 116 personnel employed at Australian oil companies. Of these, 93 completed all of the (below) tasks in the allotted time. These included 70 males and 23 females, with a mean age of 41.3 years $(S D=10.8)$ and an average of 16.4 years of industry experience $(S D=10.0)$.

## Procedure

Participants were recruited during several visits to oil
companies and tested in groups of 25-30. They were given the pencil and paper battery of questions described below and allowed 45 minutes to complete it.

## Materials

The questionnaire asked demographic questions, Frederick's CRT (spread throughout the questionnaire) and 10 bias measurement tasks commonly seen in managerial decision making books/courses (see, e.g., Bazerman, 2002). Three of these (base rate neglect, optimism and unpacking) were included for separate analyses and are not discussed here. The remaining biases were: anchoring, overconfidence, framing, conjunctive/disjunctive events bias, sample size invariance, the Wason selection task and illusory correlations. Each, except for overconfidence, was tested using a single item and all were scored in line with the CRT; that is, higher scores indicated less bias susceptibility. The specific tasks are described below.

## Demographics.

Participants provided their age, gender, technical specialty and years of industry experience. They also indicated whether they had undertaken training courses in decision making and when, where and with whom this was done.

## Anchoring.

Participants were asked whether world proved oil reserves in 2009 were greater or less than an anchoring value prior to being asked to make an estimate. The assumption here is that oil industry personnel, while unlikely to have a figure for this already in mind, would be capable of constructing a reasonable estimate from their industry knowledge but that, in line with the anchoring and adjustment heuristic (Tversky \& Kahneman, 1974), people's estimates would tend towards the anchor they had seen. Participants saw one of two anchors $-150 \%$ or $50 \%$ of the known true value, although participants were unaware of this - and were assessed as showing the bias if their estimate was closer to the anchor they saw (scored 0 ) than the unseen alternative (scored 1 ).

## Overconfidence.

This task included 10 questions asking the participant to generate an $80 \%$ confidence interval around an unknown quantity related to the oil industry - a task commonly undertaken the oil industry but at which people are known to perform poorly (see, e.g., Lichtenstein, Fischhoff, \& Phillips, 1982; Welsh \& Begg, 2016).

Performance was calculated as the proportion of generated ranges that contained the true value. This was ten converted to a 0 to 1 scale for easier comparison with the other bias scores with 0 indicating the worst performance and 1 the best as follows: Score $=1$-|Hits/10-.8|*1.25

## Framing.

This question, adopted from Pieters (2004), asked participants to select between options for dealing with a hypothetical oil spill - one certain to reduce it by a set amount $(1 / 3)$ and one giving a $1 / 3$ chance of containing it entirely but a $2 / 3$ chance of it spreading to its maximum
extent. That is, both options had an expected value of a $1 / 3$ reduction in the slick. Half of participants had these explained to them in terms of how much the oil would spread (negative frame) while the rest were told how much oil would be contained (positive frame) by each option.

In line with Prospect Theory (Kahneman \& Tversky, 1979), the expectation is that having the problem framed positively will tend to produce risk aversion - causing participants to select the certain option - while a negative frame tends to result in selecting the riskier option. A participant's response was, thus, scored as to whether they conformed to the Prospect Theory prediction (0) or not (1).

## Conjunctive/Disjunctive Events Bias

This question asked participants to select which of three possible responses to a probability question was correct. Specifically, which event was more likely of: a single $50 \%$ prospect finding oil; all of seven $90 \%$ prospects finding oil ( $\sim 48 \%$ ); or at least one of seven $10 \%$ prospects finding oil ( $\sim 52 \%$ ). As noted by Bar-Hillel (1973), people tend to overestimate the likelihood of conjunctive events and underestimate the likelihood of disjunctive events. Given this, participants were scored 1 if they correctly identified the third option and 0 otherwise.

## Sample Size Invariance.

This task asked whether a statistically unlikely result was more likely to occur in a larger or smaller sample - or be similarly likely. Specifically, participants were asked whether, on a given day, it was more likely that $60 \%$ of oil wells would produce above their average rate in a larger (45 well) or smaller ( 15 well) field. As noted by Tversky and Kahneman (1974), people can pay too little attention to the size of the sample and fail to realise that deviant results are more likely in a smaller sample. Given this, selecting the smaller option was scored correct (1) while any other response was scored incorrect (0).

## Selection Task.

Based on Wason's (1968) selection task, participants were asked which of four oil prospects needed to be retested with an alternative tool in order to test a consultant's claim that Tool 2 would always produce a positive result when Tool 1 did. A correct response (scored 1) was to retest prospects where the Tool 1 had given a positive result and those where Tool 2 had given a negative result. Any other combination of choices was deemed incorrect (scored 0 ).

## Illusory Correlations.

The illusory correlations task (Chapman \& Chapman, 1967), asked participants to examine a $2 \times 2$ contingency table and determine whether the data supported a relationship between two events: AVO anomalies (from seismic data) and hydrocarbon presence. In fact, the data offered no support for this despite a preponderance of observations in the AVO present/HC present cell. Participants were scored as correct (1) only if they correctly identified there was no relationship in the data and that all four cells needed to be examined to establish this fact.

Claiming that the data supported a relationship or that only some cells needed to be examined resulted in a score of 0 .

## Cognitive Reflection Test.

The three questions from Frederick's (2005) CRT were spread amongst the other tasks. A person's score on this task is simply the number of questions answered correctly.

## Familiarity.

At the end of the survey, participants were asked to look back and, for each question, indicate whether they had:

1) Never seen it prior to testing (score 0 ).
2) Seen a similar question previously (score 0.5 ).
3) Seen that exact question previously (score 1).

Tasks involving more than one question (CRT and overconfidence) had the familiarity scores for all composite questions averaged to produce a single, familiarity score.

## Results

## Demographic Data

In addition to the data described in the Method section, several demographic questions were asked of participants. Key observations from this are presented in Table 1.

Table 1: Summary of demographic measures.

| Measure |  |
| :--- | :--- |
| Technical Area | 32 Engineers, 52 Geoscientists, 8 Other |
| Training* | 38 trained, 55 untrained |
| Yrs since training | Mean $=4.8$ years, $S D=4.6$ |

* Training courses in decision making, heuristics and biases.


## Descriptive Statistics

Table 2 summarises participant performance on the various measures and their stated familiarity with the questions. Looking, first, at the scores in Table 2 a number of things are immediately clear. The first is that a majority of participants display bias on each of the bias measures. On the six which reflect a simple proportion correct, the highest mean is 0.32 for the Conjuntive/Disjunctive events bias which reflects chance performance on a three-option choice. On the other, single-item tasks, performance ranges from $12 \%$ up to $27 \%$ correct - indicating a significant majority displaying the expected biases. Overconfidence requires more explanation as it indicates the proportion of generated ranges containing the true value compared to the expected number. Thus, the 0.49 average in Table 1 reflects a person achieving around half of their expected calibration - that is $\sim 40 \%$ of their " $80 \%$ " ranges containing the true value, which is a typically strong level of overconfidence.
Finally, the CRT scores are very high. Frederick's (2005) paper listed 11 samples with average CRT scores ranging from 0.57 to 2.18 (and an overall mean of 1.24). A $95 \%$ CI around the industry sample's mean CRT extends from 2.24 to 2.60 - excluding not just the overall average from Frederick's results but that of the highest group as well.

Table 2's familiarity data also shows interesting results. Specifically, while no familiarity scores are particularly high - recalling that a score of 1 would indicate definitely recalling an entire task - participants' highest familiarity rating is observed for the CRT. The average (0.25) score here lies between what would be observed from participants having recalled seeing one of the CRT's actual questions before ( 0.33 ) and having seen one similar one (0.17).

Table 2: Performance on bias and CRT measures.

|  | Score |  | Familiarity |  |
| :--- | :---: | :---: | :---: | :---: |
| Measure | Mean | $S D$ | Mean | $S D$ |
| Anchoring | 0.27 | 0.45 | 0.23 | 0.27 |
| Framing | 0.27 | 0.45 | 0.12 | 0.23 |
| Con/Disjunctive | 0.32 | 0.47 | 0.14 | 0.24 |
| Sample Size | 0.22 | 0.41 | 0.09 | 0.22 |
| Selection Task | 0.12 | 0.32 | 0.17 | 0.27 |
| Illusory Correlation | 0.17 | 0.38 | 0.16 | 0.30 |
| Overconfidence | 0.49 | 0.31 | 0.20 | 0.25 |
| CRT | $2.42 / 3$ | 0.88 | 0.25 | 0.27 |

Note: $\mathrm{N}=93$. The unshaded parts reflect tasks where the Mean value equals the proportion of correct responses.

## Training and Familiarity with Bias and CRT

To test Hypothesis 1 - that industry courses in decision making would increase familiarity with bias and CRT questions - familiarity ratings of participants with and without such training were compared. Looking at Table 3, Hypothesis 1 is clearly supported by the data. In all cases, participants who had undertaken training courses reported significantly higher familiarity with the bias and CRT questions. An interesting observation, however, is that the CRT is an outlier amongst untrained personnel - its mean familiarity more than double that of any other question. This may go some way to explaining the distribution of CRT scores across the trained and untrained groups, shown in Figure 1, where three-quarters of the trained group score $3 / 3$, but so do half of the untrained group.

Table 3: Familiarity with bias and CRT measures by training group.

|  | training group. |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | Trained | Untrained | $t(91)$ | $p$ |
| Anchoring | .45 | .07 | 8.93 | $<.001$ |
| Overconfidence | .37 | .08 | 6.69 | $<.001$ |
| Framing | .21 | .05 | 3.46 | $<.001$ |
| Con/Disjuntive | .24 | .07 | 3.47 | $<.001$ |
| Sample Size | .18 | .03 | 3.59 | $<.001$ |
| Selection | .33 | .05 | 5.56 | $<.001$ |
| Illusory Corr. | .17 | .05 | 5.12 | $<.001$ |
| CRT | .35 | .18 | 3.21 | .002 |

Note: p-values are two-tailed. Independent samples $t$-tests.

## CRT Familiarity and Score

Hypothesis 2 held that familiarity with CRT questions would inflate CRT scores. The correlation between CRT scores and familiarity with CRT questions supported the hypothesis, showing a weak, significant effect, $r(91)=0.29$,
$p=.004$（see Table 5）．That is，participants who had seen CRT（or similar）questions before scored higher．


Figure 1：Distribution of CRT scores by training group．
To better understand the magnitude of the effect，the CRT scores of participants unfamiliar with all of the CRT questions（i．e．，CRT Familiarity $=0$ ）were compared to those who recalled at least one similar question．Looking at Table 4，one sees that the familiar group scored more than half a mark higher on the CRT，which an independent samples t－test confirmed as a significant difference．

Table 4：Mean CRT scores by familiarity group

## CRT Familiarity

| $0(\mathrm{n}=41)$ | $>0(\mathrm{n}=52)$ | $t(91)$ | $p(2$－tailed $)$ |
| :---: | :---: | :---: | :---: |
| $2.12(S D=1.08)$ | $2.65(S D=0.59)$ | 3.0 | .003 |

## Predictive Power of CRT

Hypothesis 3 held that the inflation of CRT scores as a result of familiarity would reduce the its predictive power－ measured herein by correlations calculated between all bias measures，CRT and CRT familiarity，and shown in Table 5.
Looking at Table 5，one sees that the CRT has relatively little predictive power for the seven biases．It very weakly predicts better performance on the Selection task and on Overconfidence questions．This analysis，however，includes participants familiar and unfamiliar with the CRT．To assess the impact of familiarity on CRT＇s predictive power， correlations were calculated separately for participants familiar and unfamiliar with the CRT as seen in Table 6.

Here，one sees that，the CRT does not significantly predict bias for familiar or unfamiliar participants．The pattern of results，however，is for the correlation to be higher in the group familiar with the CRT（ 5 of 7 biases）．While the smaller samples resulting from dividing the group renders these non－significant，the correlations are higher than the significant ones in the full dataset，suggesting prior CRT familiarity may predict better performance on these biases．
Thus，while the overall result does not，technically， support Hypothesis 3，it identifies the lack of predictive power for the CRT in the industry sample that is unfamiliar with the CRT and suggests that what predictive power is observed in the group familiar with the CRT may result from either prior CRT experience somehow priming people to be more aware of biases－or，more likely，that participants with prior exposure to the CRT may also have experience with bias questions and thus perform better．

Table 5：Correlations between CRT，CRT familiarity and bias measures．

|  | $\begin{gathered} \stackrel{\leftarrow}{\sim} \\ \underset{\sim}{\sim} \\ \hline \end{gathered}$ |  | $\begin{aligned} & \dot{\prime} \\ & \frac{1}{C} \\ & \frac{C}{4} \\ & \text { लं } \end{aligned}$ | 4 <br> 0 <br> 0 <br> 0 <br> 0 <br> 0 <br> 0 |  | $\begin{aligned} & \dot{\infty} \\ & \stackrel{-}{\square} \\ & 0 \\ & 0 \\ & 0 \end{aligned}$ | $\begin{aligned} & \stackrel{0}{N} \\ & \underset{N}{N} \\ & \dot{C} \\ & \sim \\ & N \\ & N \end{aligned}$ | $\begin{aligned} & \text { C } \\ & \text { 응 } \\ & \text { O } \\ & \text { W } \\ & \infty \end{aligned}$ | $$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\leftharpoondown$ | ， | \％ | ছ | o | مٌ | ָ | $\begin{aligned} & 6 \\ & \hline 0 \end{aligned}$ | － | ¢ |
| $\sim$ | ¢ | ， | $8$ | $\underset{\sim}{\infty}$ | م̂ | $\stackrel{0}{0}$ | $\stackrel{N}{\mathrm{~N}}$ | $\begin{aligned} & \text { N} \\ & \text { Ṇ } \end{aligned}$ | － |
| $\cdots$ | $\bigcirc$ | $\underset{\sim}{\infty}$ | ＇ | $\stackrel{+}{\infty}$ | $\stackrel{\infty}{\mathrm{P}}$ | ！ | 앙 | ホ | N |
| $\pm$ | $\stackrel{\infty}{\sim}$ | $\cdots$ | $\bigcirc$ | ＇ | $\begin{aligned} & 0 \\ & 0 \\ & \hline \end{aligned}$ | 8 | $\stackrel{\infty}{\square}$ | $\stackrel{J}{\bigcirc}$ | $\ddagger$ |
| 10 | $\stackrel{\bigcirc}{\square}$ | $\bigcirc$ | $\bigcirc$ | $\bigcirc$ | 1 | Ǹ | 巨্ফ | $\bigcirc$ | $\stackrel{\bigcirc}{\stackrel{\circ}{*}}$ |
| $\bullet$ | $\bigcirc$ | N! | 오 | 「 | $\stackrel{\sim}{*}$ | ＇ | $\frac{0}{\infty}$ | $\stackrel{8}{\uparrow}$ | ¢ |
| $N$ | $\bigcirc$ | ? | T | $\stackrel{\square}{\square}$ | $\stackrel{\bigcirc}{\sim}$ | O | ＇ | No | － |
| $\infty$ | 「 | $\stackrel{\sim}{\sim}$ | $\hat{O}_{i}$ | ¢ | $\stackrel{0}{1}$ | $\bigcirc$ | $0$ | 1 | $\bigcirc$ |
| の | $\stackrel{F}{\square}$ | $F$ | $\underset{~}{~}$ | $\infty$ | $\infty$ | $\stackrel{\sim}{*}$ | $\stackrel{\infty}{\sim}$ | $\underset{N}{N}$ | 1 |

Note： $\mathrm{N}=93$ ．Values in the lower triangle are correlation coefficients．Upper triangle data are two－tailed p－values． Bold results are significant．Italic results are significant as directional hypotheses．NB－for binary bias measures，the correlations are equivalent to t－tests and used in preference for consistency and ease of display．

Table 6：Correlations between CRT and biases in participants familiarity and unfamiliar with CRT．

Correlation with CRT
Unfamiliar $n=41)$

|  | Correlation |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | Unfamiliar $(n=41)$ | Familiar（ $n=52$ ） |  |  |
| Bias Task | $r$ | $p(2$－tailed $)$ | $r$ | $p(2$－tailed $)$ |
| Anchoring | .11 | .512 | .04 | .787 |
| Overconf． | .04 | .828 | .26 | .066 |
| Framing | -.07 | .647 | -.11 | .421 |
| Con／Dis．Bias | . .12 | .469 | .09 | .549 |
| Sample Size | .05 | .756 | .07 | .647 |
| Selection | .13 | .416 | .21 | .144 |
| Ill．Corr． | .02 | .914 | .21 | .144 |

## Familiarity and Bias Resistance

As noted above，the results suggest support for Hypothesis 4 －that familiarity with bias questions would improve performance．Data in Table 5 also show CRT Familiarity has stronger relationships with bias performance than a participant＇s CRT score．Given the likely co－occurrence of bias and CRT questions in decision training courses，the effect of familiarity on bias was thus also examined．

To test this, $\chi^{2}$ tests were conducted for the six, binaryscored biases. Given low numbers of participants recalling seeing exact bias questions before, familiarity with a bias was also treated as binary by combining groups who had seen the exact or a similar question together. Table 7 shows the proportion of correct responses for each of these groups for each bias and the results of the corresponding $\chi^{2}$ tests.

Table 7: Proportion correct by bias question and familiarity.
Familiarity

| Bias Task | 0 | $>0$ | $\chi^{2}(1)$ | p |
| :--- | :---: | :---: | :---: | :---: |
| Anchoring | $0.23(\mathrm{n}=53)$ | $0.33(\mathrm{n}=40)$ | 1.13 | .288 |
| Framing | $0.28(\mathrm{n}=72)$ | $0.24(\mathrm{n}=21)$ | 0.13 | .718 |
| Con/Dis | $0.29(\mathrm{n}=68)$ | $0.40(\mathrm{n}=25)$ | 0.94 | .333 |
| Sample Size | $0.22(\mathrm{n}=78)$ | $0.20(\mathrm{n}=15)$ | 0.02 | .877 |
| Selection | $0.08(\mathrm{n}=65)$ | $0.27(\mathrm{n}=22)$ | 3.54 | .060 |
| Illusory Corr. | $0.16(\mathrm{n}=69)$ | $0.21(\mathrm{n}=24)$ | 0.30 | .584 |
| Overconfidence | $r(93)=0.25$ |  |  | .016 |

Note: $p$ values are two-tailed. Overconfidence and its corresponding familiarity are both non-binary, therefore a correlation is used rather than $\chi^{2}$.

Looking at Table 7, one sees that, participants familiar with bias questions do better in 5 of the 7 tasks but significantly only on the Overconfidence and Selection tasks (given a directional hypothesis). That is, Hypothesis 4 is supported for Overconfidence $(r(93)=0.25, p=.016)$ and the Selection Task $\left(\chi^{2}(1)=3.54, p=.060\right)$ - the two biases showing the strongest relationships with CRT amongst participants familiar with the CRT in Table 6.

## Discussion

The results offer support for two hypotheses: that taking part in decision making training courses increases the likelihood of having seen the CRT or bias questions previously; and that having seen CRT-style questions previously results in a significant increase in CRT score - of more than half a mark on the 0-3 scale. The fact that results (largely) failed to support the other hypotheses has, along with observations on the limited predictive power of the CRT herein, implications for the use of the CRT, as expanded on below.

## Predictive Power of the CRT

As noted above, our third hypothesis was that the CRT's predictive power would be eroded by participant's familiarity with CRT questions. The reasoning being that, given a limited set of memorable questions, prior exposure would push results towards ceiling, weakening the relationship between CRT and the biases. Our results, however, showed CRT having little predictive power to start with. This lack of initial, predictive power in our sample may have made it impossible to convincingly demonstrate the impact of familiarity on CRT's predictive power.
The reason for this lack of predictive power, however, seems to be the same as that prompting our Hypothesis 3 that CRT scores are too close to ceiling. As noted above, even the CRT scores of participants with no familiarity with

CRT questions were, at 2.12 , similar to the highest of the 11 groups tested by Frederick (2005) and much higher than his average of 1.24 .

Part of this, we argue, must stem from the nature of our sample. Rather than undergraduate students, we tested oil industry professionals - primarily engineers and scientists. As such, our sample is likely to have much higher than typical numeracy scores and, consequently, higher CRT scores (for discussions of the links between CRT and numeracy, see, e.g., Weller et al., 2013; Welsh et al., 2013).

While this has made certain of our planned comparisons more difficult - effectively rendering our 'control' group of people unfamiliar with the CRT too similar to those who had prior experience, the implications of this for the use of the CRT in expert samples are more troubling. It suggests that, even prior to their first exposure to the CRT, the skewed scores seen in a sample of technical experts will limit the test's ability to differentiate between individuals and predict performance. Combined with the observation that the CRT's highest predictive power was observed amongst people with prior experience on exactly those biases where prior experience aided the most - this argues against the CRT's usefulness.

While these concerns may be lessened when dealing with experts from less numerically-focussed fields, expert decision making and forecasting tends towards exactly these groups, meaning that the CRT may have limited utility.

## Bias vs CRT Familiarity

Analyses of familiarity with both biases and the CRT used to examine Hypothesis 4 found limited evidence of prior experience with biases improving performance. Only for the Overconfidence and the Selection Task did prior exposure lead to better performance - perhaps due to greater memorability or that understanding these biases suggests a solution. For example, overconfidence implies too narrow ranges, which immediately suggests widening ranges. Such awareness generally reduces but does not remove Overconfidence (Welsh, Begg, \& Bratvold, 2006). Amongst the other biases, little evidence was seen of prior bias question experience enabling one to avoid those biases even in an educated, highly numerate sample.

This is doubly important in light of familiarity's effect on CRT. If CRT is inflated by prior exposure more than bias performance, then knowing who has been exposed to the CRT-style questions becomes essential when interpretting results. Adding to this is the fact that the CRT was more familiar than the biases to people who had not completed training, suggesting that these questions occur through other channels or that CRT questions are particularly memorable.

This seems likely to remain true even when 'similar' tasks are used. The structures of CRT questions, once recognized as 'trick' questions, may trigger greater scrutiny of intuitive answers. Certainly, while few participants indicated having seen the exact CRT questions before, reporting having seen similar ones (for now, ignoring questions about the accuracy of their recall) also resulted in higher CRT scores.

## Future Research

Given the problems observed with CRT, an obvious next step is to attempt a replication using one of the newer variants developed to have less reliance on numeracy and a larger number of items (e.g., Primi et al., 2015; Thomson \& Oppenheimer, 2016; Toplak et al., 2014). Whether such a substitution will work depends on whether familiarity is highly specific for particular question types or simply primes generic "I know this is a trick question" responses.

Another necessary step is to look at biases discussed here in greater detail. While a (mostly) single item per bias approach is useful for an exploratory approach - allowing multiple biases to be examined without overloading the goodwill of participants - binary scoring is, of course, a crude measure of susceptibility to any bias. Research using a set of bias questions for each bias (and focusing on fewer biases so as to keep the total number of questions down) would allow finer-grained measurement of susceptibility and shed further light on the findings discussed herein (and allow more detailed discussion of the biases, their modes of action and some of the controversies in the literature regarding their nature - or even existence).

Finally, the very high CRT scores we observed in our oil industry sample suggest that additional work should be conducted to determine how CRT scores vary in other fields amongst both naïve and CRT-familiar personnel.

## Conclusions

Our results have important implications for the use of the CRT as a bias susceptibility measure for decision making research in professional settings. Our technical experts, while susceptible to biases, have inflated CRT scores resulting from greater numerical ability as well as any prior exposure to CRT-style questions. These effects result in the original CRT retaining little to no predictive power.

Given this, future work is required to see whether alternate versions of the CRT, developed to include more items and be less numerically-based, avoid such problems and can provide useful results in professional populations.

## Acknowledgments

The authors thank Santos and Woodside for their support of the CIBP group within the Australian School of Petroleum.

## References

Bar-Hillel, M. (1973). On the subjective probability of compound events. Organizational Behavior and Human Performance, 9(3), 396-406.
Bazerman, M. H. (2002). Judgment in managerial decision making (5th ed.). New York: John Wiley and Sons.
Capen, E. C. (1976). The difficulty of assessing uncertainty. Journal of Petroleum Technology(August), 843-850.
Chapman, L. J., \& Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. Journal of abnormal psychology, 72(3), 193.

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 25-42.
Kahneman, D. (2011). Thinking, Fast and Slow. New York, NY: Farrar, Straus, Giroux.
Kahneman, D., \& Tversky, A. (1979). Prospect Theory: an analysis of decision under risk. Econometrica, 47(2), 263291.

Lichtenstein, S., Fischhoff, B., \& Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic \& A. Tversky (Eds.), Judgment under Uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.
Pieters, D. A. (2004). The influence of framing on oil and has decision making. Marietta, Georgia: Lionheart Publishing Inc.
Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., \& Hamilton, J. (2015). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). Journal of Behavioral Decision Making.
Stieger, S., \& Reips, U.-D. (2016). A limitation of the Cognitive Reflection Test: familiarity. PeerJ, 4, e2395.
Thomson, K. S., \& Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. Judgment and Decision Making, 11(1), 99.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory \& Cognition, 39(7), 1275-1289.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking \& Reasoning, 20(2), 147-168.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 11241131.

Wason, P. C. (1968). Reasoning about a rule. The Quarterly journal of experimental psychology, 20(3), 273-281.
Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., \& Peters, E. (2013). Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach. Journal of Behavioral Decision Making, 26, 198-212. doi: doi: 10.1002/bdm. 1751
Welsh, M. B., \& Begg, S. H. (2016). What have we learnt? Insights from a decade of bias research. APPEA Journal, 56, 435-450.
Welsh, M. B., Begg, S. H., \& Bratvold, R. B. (2006). SPE 102188: Correcting common errors in probabilistic evaluations: efficacy of debiasing. Paper presented at the Society of Petroleum Engineers 82nd Annual Technical Conference and Exhibition., Dallas, Texas, USA.
Welsh, M. B., Burns, N. R., \& Delfabbro, P. H. (2013). The Cognitive Reflection Test: how much more than Numerical Ability? In M. Knauff, M. Pauen, N. Sebanz \& I. Wachsmuth (Eds.), Proceedings of the 35th Meeting of the Cognitive Science Society (pp. 1587-1592). Austin, TX: Cognitive Science Society.

# Desires influence 4- to 6-year-old children's probabilistic judgments 

Adrienne Wente<br>University of California, Berkeley<br>Mariel Goddu<br>University of California, Berkeley<br>Elyanah Posner<br>University of California, Berkeley<br>Teresa Garcia<br>University of California, Berkeley<br>María Fernández Flecha<br>Pontificia Universidad Católica del Perú

Alison Gopnik<br>University of California, Berkeley


#### Abstract

Research on wishful thinking suggests that desires bias adult's probability judgments. Previous research has yet to explore if this extends to young children. In Experiment 1, 2604 - and 6-year-olds in the U.S. and Peru played a card game, where selecting a desirable card was unlikely. In Experiment 2, 2004 - to 6-year-old children were shown a bag of plastic eggs; a few contained desirable prizes. Children were asked to make predictions about what card / egg would be randomly selected. Answers were compared to control conditions in which probability was comparable, but children had no reason to desire a specific outcome. In control conditions, children tended to state that the majority card/ egg would be selected. In the experimental conditions, children were more likely to state that the desirable (and improbable) card/ egg would be selected. Results suggest that a desire bias extends to children as young as 4 .


# The Interaction of Bayesian Pragmatics and Lexical Semantics in Linguistic Interpretation: Using Event-related Potentials to Investigate Hearers' Probabilistic Predictions 

Markus Werning and Erica Cosentino<br>(\{markus.werning, erica.cosentino\}@rub.de)

Department of Philosophy, Ruhr University Bochum


#### Abstract

We contrast two views of how contextual influence on sentence meaning composition can be explained. The Semantic Similarity View maintains that discourse context affects sentence meaning mainly because of the semantic similarity between the words in the discourse context and the words in the sentence (as measured by Latent Semantic Analysis). The Free Pragmatic View, in contrast, defends the claim that also pragmatic aspects of the discourse context can affect sentence meaning composition. This effect can be quantitatively modelled by Bayesian Pragmatics. We introduce a Predictive Completion Task in which the hearer at every moment in a communicative situation has to generate a probabilistic prediction about how a discourse being uttered by the speaker is continued. We test the predictions of the two views in EEG using the well-established observation that the conditional probability of a word given a context is negatively correlated with the amplitude of its N400 component.


Keywords: Latent Semantic Analysis, Bayesian Pragmatics, N400, Generative Lexicon, Telicity, Affordances, Context, Predictive Coding

## Introduction

It has been widely acknowledged that a preceding discourse can influence the way sentence meaning is composed from lexical meaning. In this paper we want to adjudicate between two competing views of how discourse context affects sentence meaning. A prominent view is that the contextual influence is mainly due to the semantic similarity between parts of the discourse context and the words in the target sentence (e.g., as in semantic priming; Otten \& Van Berkum, 2008). It is however highly controversial whether also pragmatic aspects of the discourse context other than the mere resolution of indexicals and anaphors can immediately affect sentence meaning composition. Nieuwland and Van Berkum (2006) have argued that discourse contexts can overturn violations of animacy. For a noun denoting an object that would normally be regarded as inanimate (e.g. peanut) the feature of animacy can be introduced if the preceding discourse context specifies a suitable situation (e.g., a romantic story whose protagonist is a peanut). Here predicates that would normally conflict with the noun because they require animacy (the peanut was in love) were actually more easily predictable than canonical predicates (the peanut was salted), as revealed by an enhanced N 400 for canonical predicates as compared to animacy-violating predicates.

To investigate the contrast between a Semantic Similarity account of contextual influence and a Free Pragmatic account that allows for free pragmatic enrichment in sentence meaning composition, we will look into the way subjects make probabilistic predictions on the completion of
a sentence given a preceding discourse. Quantitatively, the semantic similarity can be determined by Latent Semantic Analysis (Landauer \& Dumais, 1997), whereas we will use the framework of Bayesian Pragmatics (Frank \& Goodman, 2012) to calculate the pragmatic influence - in particular, concerning the rationality of the speaker's intentions in a narrative. As a model of lexical structure we apply the Generative Lexicon approach (Pustejovsky, 1995). In our experimental design we will use the established observation that the conditional probability of a word given a preceding context is negatively correlated with the amplitude of its N400 component measured in EEG.

## Background

The general idea of Bayesian Pragmatics is to account for the rational cooperation between speaker and hearer in an act of communication by modelling the hearer's probabilistic expectations about the speaker's communicative intentions by Bayes's Theorem. Bayesian pragmatics has been successfully used e.g. to explain results in a number of behavioral experiments (e.g., Frank \& Goodman, 2012). It has so far not been validated in EEG studies.

Bayesian pragmatics offers itself as a model also in what one might call the Predictive Completion Task (PCT) of communication. Predictive coding is widely acknowledged in cognitive science as a general mechanism by which the subject at every point in time generates the most probable prediction of the next event on the basis of ongoing perceptual input and learned statistical regularities (Hohwy, 2013). In a PCT the hearer at every moment in a communicative situation has to generate a probabilistic prediction about how a sentence/discourse being uttered by the speaker will be continued. To get a quantitative grasp of this task, we define $P_{T}(a \mid c)$ as the conditional probability of a sentence/discourse being continued with the word $a$ given that the word is preceded by a context $c$ under the assumption that the complete sentence/discourse is true. We will call $P_{T}$ the truth-guided predictive probability function of the hearer. The problem of the hearer is to estimate the truth-guided predictive probability.

The Semantic Similarity and the Free Pragmatic views provide competing theories of how the hearer accomplishes the Predictive Completion Task. To be able to discern between the two theories and given that the overall frequency of the word $a$ in language use and the syntactic congruency of the word $a$ relative to the preceding context
is known to have a major influence on $P_{T}(a \mid c)$, we will in the course of this paper (and in the experiment) presuppose that the frequency of $a$ is held invariant and that the syntactic congruency between $a$ and its context is granted.

According to the first view, $P_{T}(a \mid c)$ should be estimated solely on the basis of the semantic similarities between the lexical meaning of the word $a$ and the semantic properties of the preceding context $c$. The semantic similarity can be quantified by Latent Semantic Analysis, LSA (Landauer \& Dumais, 1997).

The Free Pragmatic View in contrast maintains that pragmatic aspects of the discourse directly interact with meaning components retrieved from the lexicon as well as with any further node in the sentence meaning composition tree. It thus challenges a rigorous notion of compositionality, according to which the meaning of a complex expression is determined by the meanings of its syntactic parts and the way the parts are combined (Werning, 2004, 2005a, 2005b, 2012). Pragmatic enrichment is supposed to be "free" because not only lateral modulations of a word or phrase are allowed, e.g. when the meaning of a word is modulated by the meaning of its argument - cut the cake (vertical cutting) vs. cut the grass (horizontal cutting) - but any, however remote information can in principle modulate the meaning of a linguistic expression at any stage in semantic composition. For example, cut the grass, given a lawn-seller situation, might be interpreted as vertical cutting. Accordingly, a situation introduced in a discourse preceding the sentence may result in the modulation of the meanings of words or phrases in the sentence and of the sentence itself (Cosentino, Adornetti, \& Ferretti, 2013). These modulations will then influence the intuitive truth-conditions of the sentence. This view, as developed for example by Recanati (2012) amounts to a weakening of the rigorous notion of compositionality by introducing context-dependent semantic flexibility by means of modulation. In the Predictive Completion task Bayesian Pragmatics can be used to quantitatively model the Free Pragmatic account (see Predictions).

## Design

The contrasting quantitative predictions of the Semantic Similarity and the Free Pragmatic views can be applied to a previous EEG experiment of ours (Cosentino, Baggio, Kontinen, Garwels, \& Werning, 2014; Cosentino, Baggio, Kontinen, \& Werning, 2017). To design the experiment we combined a particular idea of Pustejovsky's (1995) Generative Lexicon approach with Gibson's (1979) notion of affordances. According to the Generative Lexicon approach, the lexical entry of concrete nouns (e.g. banana) contain a "Qualia Structure" which, among others, specifies a so-called Telic component (e.g. eat) that is retrieved in sentence meaning composition. This retrieval is typically triggered by verbs like use and enjoy that take the respective noun as argument. This explains why sentences such as (a) and (c) are typically understood as having the meaning of (b) and, respectively, (d):
(a) The child enjoyed the banana.
(b) The child enjoyed eating the banana.
(c) The man used his jackknife for the cake.
(d) The man used his jackknife for cutting the cake.

In turn, Gibson proposed that many objects come with subject- and situation-dependent affordances. These are dispositional properties (e.g., sit-ability) that relate to actions to be potentially performed on that object (Werning, 2010). We distinguish between ad-hoc affordances and generic affordances. Generic affordances are affordances of a class of objects that are represented as part of the mental concept of the class (e.g., chair - sit). Ad-hoc affordances are affordances that a particular object has for a particular agent in a particular situation (e.g., this chair - hide under, for a child in a peekaboo game). In line with Pustejovsky, generic affordances are often stored as telic components in the lexicon of nouns and thus in semantic long-term memory.

|  | +TLex | -TLex |
| :--- | :--- | :--- |
| TStd- |  |  |
| Ctx | Clare got herself a funnel to <br> perform a little chemistry <br> experiment at home and to this <br> end she put some dye in water. <br> Once she has done so, she uses <br> the funnel to pour water into a <br> container. | Clare got herself a funnel to <br> perform a little chemistry <br> experiment at home and to this <br> end she put some dye in water. <br> Being an unconventional <br> person, she uses the funnel to <br> hang her coat. |
|  | Clare has an extra funnel and, <br> after having decided what to do <br> with it, she glues it to the wall <br> leaving the narrow end facing <br> outward. <br> Once she has done so, she uses <br> the funnel to pour water into a <br> container. | Clare has an extra funnel and, <br> after having decided what to do <br> with it, she glues it to the wall <br> leaving the narrow end facing <br> outward. <br> Being an unconventional <br> person, she uses the funnel to <br> hang her coat. |
| Ctx |  |  |

Table 1. Sample stimuli for EEG experiment on context effects on Telic lexical component. The table illustrates a $2 \times 2$ design, in which two categories of noun-verb combinations, +TLex and -TLex, are combined with two categories of discourse contexts, Telic Standard Context (TStdCtx) and Telic New Context (TNewCtx). The cue verb is underlined while the corresponding noun preceding the cue verb is in bold. The original stimuli were Italian and here are translated to English.

In our $2 \times 2$ experimental design (see Tab. 1) the first variable - TelicLexicalMatch - refers back to Pustejovsky's notion of a Telic component. The Telic component of the lexical entry specifies the function or the purpose of an object. With regard to the variable TelicLexicalMatch the two conditions, +TLex vs. -TLex varied in whether the cue verb (e.g., pour or, respectively, hang) expresses the telic component in the lexical entry of a given noun $n$ (e.g., funnel). With regard to the second variable - TContext - we varied the discourse context such that in the first condition TStdCtx a standard context preceded the target sentence, whereas in the second condition TNewCtx the preceding discourse context introduced a new telic role as an ad-hoc affordance for the object denoted by the noun, facilitating an action expressible by the -TLex verb (hang).

## Predictions

Semantic Similarity View: With word frequency held constant and syntactic congruency granted, the Semantic Similarity view now entails that the only predictor for a verb given its preceding context is the semantic similarity of the former to the latter. Given that we leave the corresponding noun $n$ (funnel), which precedes the cue verb $v_{i}$ (pour/hang), the same in all four conditions, the semantic similarity is a cumulative (i.e. in both arguments strictly monotonously increasing) function $f^{+}$of the semantic similarity $S\left(v_{i}, n\right)$ between the verb and the noun and the semantic similarity $S\left(v_{i}, c_{j}\right)$ between the verb and the preceding context $c_{j}$ excluding the noun. According to this view the truth-guided predictive probability $P_{T, n}\left(v_{i} \mid c_{j}\right)$ of the verb $v_{i}$ following the noun $n$ given the context $c_{j}$ should hence be estimated by the hearer as follows (see Eqn. (4) in Table 2).

Semantic Similarity view:

$$
P_{T, n}\left(v_{i} \mid c_{j}\right)=f^{+}\left(S\left(v_{i}, n\right), S\left(v_{i}, c_{j}\right)\right)
$$

The experimental settings were chosen such that the semantic similarity - determined by LSA -between the cue verb and the context excluding the noun was invariant across all four conditions (see Eqn. (6) in Table 2). Therefore, $P_{T, n}\left(v_{i} \mid c_{j}\right)$ should depend solely on whether the verb expresses the telic lexical component of the noun, i.e. belongs to +TLex, and correspondingly has a high semantic similarity to the noun - determined again by LSA - as opposed to a verb belonging to -TLex with a low semantic similarity to the noun (see (5)). We can now immediately mathematically derive the four comparative predictions regarding $\quad P_{T, n}\left(v_{i} \mid c_{j}\right)$ for $i \in\{+$ TLex, - TLex $\} \quad$ and $j \in\{$ TStdCtx, TNewCtx $\}$. These predictions are captured by the formulae (11), (15), (16) and (17) shown in Table 2 together with their derivations.

Bayesian Pragmatics: If the Free Pragmatic as opposed to the Similarity View is true the hearer will use a different strategy to estimate $P_{T}(a \mid c)$. From a pragmatic point of view, narratives are goal-directed discourses. In our examples the speaker has the goal to attribute an action to the narrative's protagonist which $\mathrm{s} /$ he performs on a given object: In the above examples, performing the action of pouring or, respectively, hanging on the funnel. In the narrative, the speaker embeds this action in a situation which he may introduce by a discourse context that precedes the description of the action. The speaker in other words has to choose a preceding context to let this action appear rational. To describe this choice situation quantitatively we can define the rationality-guided conditional probability $P_{R, n}\left(c_{j} \mid v_{i}\right)$ as the probability of the speaker to choose under the assumption of narrative rationality - a context $c_{j}$ given that he aims at attributing to the protagonist the action denoted by $v_{i}$ to be performed on the object denoted by $n$. Using Bayes's Theorem $P_{R, n}\left(c_{j} \mid v_{i}\right)$ can be transformed to allow the hearer, in the PCT, to estimate the truth-guided
predictive probability by equating it to the rationality-guided probability of the speaker (see Eqns. (1) and (3) in Table 2):

Free Pragmatic view:
$P_{T, n}\left(v_{i} \mid c_{j}\right)=P_{R, n}\left(v_{i} \mid c_{j}\right)$,
where, according to Bayes's Theorem,
$P_{R, n}\left(v_{i} \mid c_{j}\right)=K\left(c_{j}\right) \cdot P_{R, n}\left(c_{j} \mid v_{i}\right) P_{R, n}\left(v_{i}\right)$.
Here $\quad K\left(c_{j}\right)=\left(\sum_{v \in V} P_{R, n}\left(c_{j} \mid v\right) P_{R, n}(v)\right)^{-1} \quad$ is $\quad$ a normalizing factor and $V$ is the set all (syntactically congruent) verbs. We may assume that the rationalityguided prior probability $P_{R, n}\left(v_{i}\right)$ of an action expressed by the verb $v_{i}$ being performed on the object denoted by the noun $n$ is fully determined by the semantic similarity between the verb and the noun as computed by LSA (see Eqn. (2)):

$$
P_{R, n}\left(v_{i}\right)=S\left(v_{i}, n\right)
$$

For, the lexical entry of a concrete noun can be assumed to reflect the semantic memory of the learned statistical regularities between objects denoted by the noun and actions rationally performed on them. This assumption is part and parcel also of the idea that concrete nouns have telic lexical components in the sense of Pustejovsky's (Pustejovsky, 1995) Generative Lexicon. If the verb corresponds to the telic lexical entry of the nouns, the semantic similarity between them should hence be high. In the experimental settings we can implement a comparative variation of the rationality-guided probability $P_{R, n}\left(c_{j} \mid v_{i}\right)$ of the speaker to choose a context $c_{j}$ given that he aims at attributing to the protagonist the action denoted by $v_{i}$ to be performed on the object denoted by $n$ as captured by the inequalities (7)-(10) in Table 2. This immediately allows us to generate comparative predictions about the conditional predictive probability of the hearer as a consequence of the Bayesian interpretation of the Free Pragmatic view. Making the idealizing assumption that the normalizing factor, which is unknown not only to us, but also to the hearer, is the same for all contexts and in particular, $K\left(c_{1}\right)=K\left(c_{2}\right)=K$, we can generate predictions not only for comparisons within the same context, but also across contexts. The so attained predictions of the Free Pragmatic view are captured in formulae (11), (12), (13) and (14) in Table 2, shown together with their mathematical derivation history.

Correlation with N400 amplitude: To test the predictions, we exploited an empirically already well established relationship between the probability of a word given a preceding context and the amplitude of the N400 component of the event-related potential measured on the onset of the word in EEG. Granted that the cue word is syntactically congruent with its context (i.e., no syntactic violation) and that the frequency as well as length of the word are invariant, the truth-guided predictive probability $P_{T}(a \mid c)$ of the word $a$ given the preceding context $c$ is negatively correlated with the amplitude of the word's N400. Support for the negative correlation between the truth-guided predictive probability of a word given a
preceding context and the word's N 400 comes from multiple sources of evidence. Most importantly, Cloze probability is a strong predictor of the amplitude of the N400 component. Cloze probability values are obtained by asking participants in a Cloze task to complete an incomplete sentence with the word they consider to be the most likely completion. Kutas \& Hillyard (1984) found that the amplitude of the N400 component measured on the target word has a nearly inverse linear relationship with its Cloze probability, that is, relative to more expected words, the N400 amplitude increases as the expectancy of a word in context decreases. DeLong, Urbach, \& Kutas, 2005 confirmed that the preceding words in a sentence are used by readers to estimate relative likelihoods for upcoming words and the differences in the likelihood of the target word are reflected in differences regarding the N400 component. This effect has been observed not only in single sentences (Van Petten, Coulson, Rubin, Plante, \& Parks, 1999), but also for short texts (Otten \& Van Berkum, 2007; Van Berkum, Brown, Zwitserlood, Kooijman, \& Hagoort, 2005). Even more subtle differences in the semantic relatedness between the words in a sentence are found to influence the conditional probability of an upcoming word and affect the amplitude of the N400 measured on that word. For instance, in the sentence The girl was writing letters when her friend spilled coffee on the tablecloth/paper" the semantically unrelated word tablecloth elicits a larger N400 than the semantically related word paper (Baggio, van Lambalgen, \& Hagoort, 2008). The integration of world knowledge during the interpretation of sentences such as The Dutch trains are yellow/white/sour and very crowded (the target words are underlined) also modulates the amplitude of the N400 component. This reflects the role of the (Dutch) subjects' knowledge that Dutch trains are typically yellow for establishing the conditional probability of the target word (Hagoort, Hald, Bastiaansen, \& Petersson, 2004). This stresses the point that the predictive conditional probability, in fact, is guided by expectations regarding the truth of the continuing sentence. Additional evidence of a negative correlation between the predictive conditional probability of a word and the amplitude of its N 400 component is provided by a study previously conducted in our laboratory. In a sentence-picture verification study on scalar implicatures, logical and pragmatic responders provide different truth-value judgements to under-informative sentences (e.g., Some As are B, when it is known that all As are B). Whereas logical responders evaluate these sentences as true, pragmatic responders reject them as false. These divergent responses correlate with significant differences regarding the N400 and can be explained on the basis of expected probabilities of words relative to truth presumed by the subject (Spychalska, Kontinen, \& Werning, 2016). Similar findings have been reported in a study about bare numerals (Spychalska, Kontinen, Noveck, Roesch, \& Werning, 2015).

## Experiment

Method: Twenty-two right-handed native speakers of Italian ( 13 males; mean age $=29.2$ years) were presented with a total of 160 stories in a $2 \times 2$ design (see Table 1). The ERPs recording was time-locked to the onset of the cue words, which were always verbs occurring in the midst of the sentence and matched on word length, number of syllables and mean word frequency. The preceding contexts were pair-wise matched for number of words. The experimental stimuli were translated into English and underwent Latent Semantic Analysis to check for the semantic similarity values between the cue verbs and the preceding nouns or, respectively, between the cue verbs and the preceding contexts. Whereas the difference between $S(+$ TLex, $n)$ and $S(-$ TLex, $n)$ was significant $(\mathrm{t}(39)=5,449, \mathrm{p}<.001)$, there was no significant difference between the cue verbs and the preceding contexts across all experimental conditions. Using average amplitude per condition across all EEG electrodes, a 2(Context: TStdCtx vs. TNewCtx) $\times 2$ (TelicLexicalMatch: +TLex vs. - TLex) repeated measures analysis of variance (ANOVA) was performed in the time window between 400 and 500 ms after critical word onset. A follow-up ANOVA was performed which involved specifically a predetermined region over centro-parietal sites at which the N400 is maximal. In this case, a 2(Context: TStdCtx vs. TNewCtx) $\times 2$ (TelicLexicalMatch: +TLex vs. - TLex $) \times 7$ (Electrodes: $\mathrm{CP} 1, \mathrm{CP} 2, \mathrm{CPz}, \mathrm{Pz}, \mathrm{P} 1, \mathrm{P} 2, \mathrm{POz}$ ) ANOVA was conducted. Bonferroni-adjusted planned comparisons were performed to decompose the effect of trial type in this region.


Figure 1. Crossing over regarding the N400 component. Bars show the average amplitude of the N 400 for the four conditions. The numbers in brackets correspond to the inequalities as predicted by the Free Pragmatic View (see Table 2, (11), (12) and (13)). Note that the fact that the difference between the +TLex and the -TLex verb in TNewCtx is not significant is also consistent with Free Pragmatic View (14).

Results: Given the standard context TStdCtx, the N400 for -TLex $(\mathrm{M}(\mathrm{TStdCtx},-\mathrm{TLex})=-1.67 \mu \mathrm{~V})$ is significantly enhanced compared to +TLex $(\mathrm{M}(\mathrm{TStdCtx},+$ TLex $)=-.64$ $\mu \mathrm{V}, \mathrm{t}(20)=3.069, \mathrm{p}=.006$, CI $1.03 \pm .70)$. Relative to the standard context TStdCtx, the TNewCtx significantly enhances the N400 component for +TLex (M(TStdCtx, + TLex $=-1.15 \mu \mathrm{~V}, \mathrm{t}(20)=2.276, \mathrm{p}=.034$, CI $.51 \pm .47)$, whereas it significantly reduces the N 400 component for TLex $(\mathrm{M}(\mathrm{TStdCtx},-\mathrm{TLex})=-.88 \mu \mathrm{~V}, \mathrm{t}(20)=-2.745$,
$\mathrm{p}=.012$, CI $-.79 \pm .60$. Finally, given a preceding context TNewCtx, the mean amplitude of the N400 component measured on +TLex was not significantly different compared to that measured on $-\mathrm{TLex}(\mathrm{t}(20)=.964, \mathrm{p}=.34)$ See Fig. 1. The follow-up ANOVA of the predetermined N400 region showed a significant Contextx TelicLexicalMatch interaction, $\mathrm{F}(1,20)=11.267, \mathrm{p}<.005$. There was no interaction with electrodes in this region.

## Discussion

In order to test the predictions made by the Free Pragmatic View and the Semantic Similarity View, we rely on the empirically well-founded observation that the truthguided conditional probability $P_{T}(a \mid c)$ of the word $a$ given the preceding context $c$ (granted that no syntactic violation is involved and that features such as the frequency and length of the word are constant) is negatively correlated with the amplitude of the N400 component measured on that word succeeding the context.

In our experimental settings, we determined the semantic similarity $S_{n}\left(v_{i}\right)$ between the meaning of the verb $v_{i}$ and the preceding noun $n$ using LSA. This value gives us the prior probability of $P_{R, n}(+$ TLex $)$ and, respectively, $P_{R, n}(-T L e x)$ with the former (pour corresponds to the telic lexical component of funnel) being higher than the latter (hang does not correspond to the lexical component of funnel), as reported in equation (5). We also determined the semantic similarity values between the verb $v_{i}$ and the discourse context $c_{j}$ excluding the noun and kept these values constant across all experimental conditions (see (6)). With regard to what is relevant for the Semantic Similarity View, the experimental conditions differ only in the semantic similarity values between the verb $v_{i}$ and the preceding noun $n$. The Semantic Similarity view entails that these values are the only predictor of differences in the truth-guided conditional probability of the verb given the preceding context and the noun (see (11), (15), (16) and (17)) and, hence, they are the only predictor of differences regarding the amplitude of the N 400 component.

The Free Pragmatic view focuses instead on the differences in the rationality-guided probability of the speaker choosing a certain context given that he attributes to the narrative subject the aim of performing a certain action with an object. In the Free Pragmatic framework, the rationality-guided probability that the speaker chooses a standard context TStdCtx (e.g., funnel in a chemistry experiment) given that he attributes to the narrative subject the aim of performing the action denoted by the +TLex verb (pour) with the object denoted by the noun $n$ (funnel) is higher than that of choosing this context given the attributed aim of performing with that object the action denoted by the -TLex verb (hang) (see inequality (7)). Furthermore, it is more rational for the speaker to choose a context TNewCtx, which introduces a new ad-hoc affordance for the object (funnel glued to the wall), compared to choose the standard context TStdCtx, given that he attributes to the narrative subject the aim of performing the action denoted by the -

TLex verb (hang) with the object denoted by $n$ (see (8)). As captured by inequality (9), the rationality-guided probability of the speaker choosing the standard context TStdCtx compared to TNewCtx is higher given that he attributes to the narrative subject the aim of performing the action denoted by the +TLex verb with the object denoted by $n$. Finally, as expressed by (10), the rationality-guided probability that the speaker chooses TNewCtx given that he attributes to the narrative subject the aim of performing the action denoted by the -TLex verb with the object denoted by $n$ is higher than the rationality-guided probability of choosing this context given the attribution to the narrative subject of the aim of performing the action denoted by the + TLex verb with the object denoted by $n$.

Given that the Free Pragmatic view estimates the truthguided predictive probability of a word by equating it with its rationality-guided probability (see (1)), the Free Pragmatic view not only predicts (11), in line with the Semantic Similarity View, but, in contrast to the Semantic Similarity View, predicts a crossing-over regarding the N400 component, as expressed by the inequalities (12), (13). With regard to the comparison expressed in (14) the Free Pragmatic View does not make an unambiguous prediction. For, a greater/smaller comparison of the values of the product $P_{R, n}\left(v_{i} \mid c_{j}\right)=K\left(c_{j}\right) \cdot P_{R, n}\left(c_{j} \mid v_{i}\right) P_{R, n}\left(v_{i}\right)$ depends not only on the numerical value of the prior probability $P_{R, n}\left(v_{i}\right)$, which is given through the equation $P_{R, n}\left(v_{i}\right)=S\left(v_{i}, n\right)$, but also on the unknown numerical value of and not just the inequalities between the likelihoods $P_{R, n}\left(c_{j} \mid v_{i}\right)$.

Given the negative correlation between the truth-guided conditional probability of a word given a preceding context and the amplitude of its N400 component, the results of our EEG study can be used to directly evaluate the different predictions of the two views. In our experiment, we found that, if preceded by a standard discourse context TStdCtx, a -TLex verb incongruent with the noun's telic component (funnel-hang) elicited an enhanced N 400 compared to a + TLex verb congruent with the telic component (funnelpour) (confirming (11)). However, given a discourse context TNewCtx, in which a new function for the object is introduced as an ad-hoc affordance, we observed a crossingover regarding the direction of the N400 effect: Comparing TNewCtx with TStdCtx, first, the N400 for the -TLex verb was significantly smaller in TNewCtx than in TStdCtx (disconfirming (15) and confirming (12)). Second, the N400 for the +TLex verb was significantly greater in TNewCtx than in TStdCtx (disconfirming (16) and confirming (13)). Finally, given a preceding context TNewCtx , the N 400 measured on the +TLex verb was not significantly different compared to that measured on the -TLex verb (see Fig. 1). This result is not decisive between the two views (neither confirming nor disconfirming (17) and being consistent with (14)).

The reported differences regarding the N400 component are best explained by the assumption that hearer accomplishes the Predictive Completion Task as envisaged
by the Free Pragmatic View, namely by estimating $P_{T, n}\left(v_{i} \mid c_{j}\right)$ through equating it with $P_{R, n}\left(v_{i} \mid c_{j}\right)$ and applying Bayes's Theorem to it. Indeed, the crossing-over regarding the N400 cannot be explained solely in terms of

|  | Free Pragmatic View | Semantic Similarity View |
| :---: | :---: | :---: |
| $\stackrel{\dot{y y}}{\hat{z}}$ | (1) $P_{T, n}\left(v_{i} \mid c_{j}\right)=P_{R, n}\left(v_{i} \mid c_{j}\right)$ <br> (2) $P_{R, n}\left(v_{i}\right)=S\left(v_{i}, n\right)$ <br> (3) $P_{R, n}\left(v_{i} \mid c_{j}\right)=K \cdot P_{R, n}\left(c_{j} \mid v_{i}\right) P_{R, n}\left(v_{i}\right)$ | (4) $P_{T, n}\left(v_{i} \mid c_{j}\right)=f^{+}\left(S\left(v_{i}, n\right), S\left(v_{i}, c_{j}\right)\right)$ |
|  | (5) $\quad S(+T L e x, n)>S(-T L e x, n)$ <br> (6) $S(+T L e x, T S t d C t x)=S(-T L e x, T S t d C t x)=S(+T L$ <br> (7) $\quad P_{R, n}($ TStdCtx $\mid+T L e x)>P_{R, n}(T S t d C t x \mid-T L e x)$ <br> (8) $\quad P_{R, n}$ (TNewCtx $\left.\mid-T L e x\right)>P_{R, n}($ TStdCtx $\mid-T L e x)$ <br> (9) $\quad P_{R, n}($ TStdCtx $\mid+$ TLex $)>P_{R, n}($ TNewCtx $\mid+$ TLex $)$ <br> (10) $P_{R, n}($ TNewCtx $\mid-T L e x)>P_{R, n}($ TNewCtx $\mid+$ TLex $)$ | $x, T N e w C t x)=S(-T L e x, T N e w C t x)$ |
|  | (11) $P_{T, n}(+T$ Lex $\mid$ TStdCtx $)>P_{T, n}(-$ TLex $\mid$ TStdCtx $)$ (from(1),(2), (3), (5), (7)) <br> (12) $P_{T, n}(-$ TLex $\mid$ TNewCtx $)>P_{T, n}(-T L e x \mid T S t d C t x)$ (from(1),(2), (3), (8), (10)) <br> (13) $\quad P_{T, n}(+$ TLex $\mid$ TStdCtx $)>P_{T, n}(+$ TLex $\mid$ TNewCtx $)$ (from(1),(2), (3), (9)) <br> (14) $\quad P_{T, n}(-T L e x \mid T N e w C t x) \gtreqless P_{T, n}(+T L e x \mid T N e w C t x)$ (from(1),(2), (3), (5), (10)) | (11) $\quad P_{T, n}(+T L e x \mid T S t d C t x)>P_{T, n}(-T L e x \mid T S t d C t x)$ (from (4), (5), (6)) <br> (15) $\quad P_{T, n}(-$ TLex $\mid$ TStdCtx $)=P_{T, n}(-$ TLex $\mid$ TNewCtx $)$ (from (4), (6)) <br> (16) $\quad P_{T, n}(+$ TLex $\mid$ TStdCtx $)=P_{T, n}(+$ TLex $\mid$ TNewCtx $)$ (from (4), (6)) <br> (17) $P_{T, n}(-T L e x \mid T N e w C t x)<P_{T, n}(+T L e x \mid$ TNewCtx) (from (4), (5), (6)) |

Table 2. Overview of the different theoretical assumptions and predictions of the Free Pragmatic View and the Semantic Similarity View given our experimental settings.

## References

Baggio, G., van Lambalgen, M., \& Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. Journal of Memory and Language, 59, 36-53.
Cosentino, E., Adornetti, I., \& Ferretti, F. (2013). Processing Narrative Coherence: Towards a Top-Down Model of Discourse. Proceedings of the 4th Workshop on Computational Models of Narrative, 32, 61-75.
Cosentino, E., Baggio, G., Kontinen, J., Garwels, T., \& Werning, M. (2014). Lexicon in action: N400 contextual effect on affordances and telicity. In P. Bello, M. Guarini, M. McShane, \& B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 2079-2084). Austin, TX: Cognitive Science Society.
Cosentino, E., Baggio, G., Kontinen, J., \& Werning, M. (2017). The timecourse of sentence meaning composition. N400 effects of the interaction between context-induced and lexically stored affordances. Frontiers in Psychology. doi:10.3389/fpsyg.2017.00813
DeLong, K. A., Urbach, T. P., \& Kutas, M. (2005). Probabilistic word preactivation during language comprehension inferred from electrical brain activity. Nature Neuroscience, 8, 1117-1121.
Frank, M. C., \& Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. Science, 336, 998-998.
Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston, MA: Houghton Mifflin.
Hagoort, P., Hald, L., Bastiaansen, M., \& Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. Science (New York, N.Y.), 304, 438-41.
Hohwy, J. (2013). The Predictive Mind. Oxford: Oxford University Press.
Kutas, M., \& Hillyard, S. (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature, 307, 161-163.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
Nieuwland, M. S., \& Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. Journal of Cognitive Neuroscience, 18, 1098-111.
Otten, M., \& Van Berkum, J. J. a. (2008). Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? Discourse Processes, 45, 464-496.
Otten, M., \& Van Berkum, J. J. A. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario
the differences in the semantic similarity values between the target verb and the preceding noun, as assumed by the Semantic Similarity account.

$$
\begin{array}{|ll}
\hline(11) & P_{T, n}(+ \text { TLex } \mid \text { TStdCtx })>P_{T, n}(- \text { TLex } \mid \text { TStdCtx }) \\
& \text { (from (4), (5), (6)) } \\
(15) & P_{T, n}(- \text { TLex } \mid \text { TStdCtx })=P_{T, n}(- \text { TLex } \mid \text { TNewCtx }) \\
& \text { (from (4), (6)) } \\
(16) & P_{T, n}(+ \text { TLex } \mid \text { TStdCtx })=P_{T, n}(+ \text { TLex } \mid \text { TNewCtx }) \\
& \text { (from (4), (6)) } \\
(17) & P_{T, n}(- \text { TLex } \mid \text { TNewCtx })<P_{T, n}(+ \text { TLex } \mid \text { TNewCtx }) \\
& \text { (from (4), (5), (6)) }
\end{array}
$$

fit on the discourse-dependent N400 effect. Brain Research, 1153, 166-177.
Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, MA: MIT Press.
Recanati, F. (2012). Compositionality, Semantic Flexibility, and ContextDependence. In M. Werning, W. Hinzen, \& E. Machery (Eds.), Oxford Handbook of Compositionality (pp. 175-191). Oxford: Oxford University Press.
Spychalska, M., Kontinen, J., Noveck, I., Roesch, L., \& Werning, M. (2015). Exploring the processing costs of the exactly and at least readings of bare numerals with event-related brain potentials. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, \& P. P. Maglio (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 2260-2265). Austin, TX: Cognitive Science Society.
Spychalska, M., Kontinen, J., \& Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: evidence from eventrelated brain potentials. Language, Cognition and Neuroscience, 31, 817-840.
Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., \& Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 443-467.
Van Petten, C., Coulson, S., Rubin, S., Plante, E., \& Parks, M. (1999). Time course of word identification and semantic integration in spoken language. Journal of Experimental Psychology. Learning, Memory, and Cognition, 25, 394-417.
Werning, M. (2004). Compositionality, Context, Categories and the Indeterminacy of Translation. Erkenntnis, 60, 145-178.
Werning, M. (2005a). Right and wrong reasons for compositionality. In M. Werning, E. Machery, \& G. Schurz (Eds.), The Compositionality of Meaning and Content (Vol. I, pp. 285-309). Frankfurt: Ontos Verlag.
Werning, M. (2005b). The Temporal Dimension of Thought: Cortical Foundations of Predicative Representation. Synthese, 146.
Werning, M. (2010). Complex First? On the Evolutionary and Developmental Priority of Semantically Thick Words. Philosophy of Science.
Werning, M. (2012). Non-symbolic compositional representation and its neuronal foundation: Towards an emulative semantics. In M. Werning, W. Hinzen, \& E. Machery (Eds.), The Oxford handbook of compositionality (pp. 633-654). Oxford University Press.

# The Neural Mechanisms of Relational Reasoning: Dissociating Representational Types 

Julia Wertheim (wertheim@tf.uni-freiburg.de)<br>Cognitive Computation Lab, University of Freiburg, Georges-Köhler-Allee 74, 79110 Freiburg, Germany<br>Marco Ragni (ragni@informatik.uni-freiburg.de)<br>Cognitive Computation Lab, University of Freiburg, Georges-Köhler-Allee 5279110 Freiburg, Germany


#### Abstract

The ability to reason about information is an essential human capability. It is less understood from the perspective of neurocognitive processes which can serve to constrain cognitive theories by implications from neuroscientific data. Despite some progress in the last decades, some disagreement about the experimental results and the cognitive processes of reasoning with abstract relations versus visuospatial relations persist. We conducted a cross-study meta-analysis of neuroimaging studies to determine the neural correlates of visuospatial and abstract relational reasoning. We analyzed 884 stereotactic data points from 38 studies and 692 subjects. We found that relational reasoning is mediated by the frontoparietal network, especially the right precuneus and the left pars triangularis. Problems with abstract relations are processed by enhanced activation in the inferior parietal lobe, whereas visuospatial reasoning is promoted by prefrontal domains. Our results disentangle the neurocognitive mechanisms of different representational types of relational reasoning across study designs.


Keywords: meta-analysis; fMRI; reasoning; relational reasoning; analogical reasoning; mental representation, precuneus, pars triangularis

## Introduction

Relational reasoning expedites human everyday life to a great extent. Using relational expressions can decrease the number of statements we would need otherwise to describe the situation. If we want to 'decompress' this information, we draw inferences from the given information to extract what is implicitly present. For example, suppose you are visiting London for the first time and you want to visit London Eye. You take the underground, arrive at Waterloo Station and ask a fellow passenger for the directions. She tells you that the London Eye is behind Jubilee Gardens which are, from your perspective, in front of Waterloo Station. Due to this description, you can easily find your goal destination. But how would you find your way from the station to London Eye without having any explicit information about the spatial orientation? You do need to reason, that means, you need to extract the information from the two statements by inferring that the London Eye is located behind the train station.

Reasoning has been subject to neurocognitive research for the past thirty years. With the advent of neuroimaging methods such as positron emission tomography (PET) and
functional magnetic resonance imaging (fMRI) in the late 1990 s and early 2000 s, seminal work has been accomplished by investigating the neural correlates of reasoning (Prado, Chadha \& Booth, 2011). The potential for research in relational reasoning is manifold. For instance, Knauff and Johnson-Laird (2002) proposed to differentiate deductive relational reasoning tasks based on the kind of mental representation necessary to represent relations. They established the categories of visual, spatial and visuospatial mental representation of the relations which differs from the mental representation of objects they relate. For our study, we decided to reevaluate the differentiations and employed new definitions for the types of mental representation. A relation helps to structure human experiences about one or many objects in everyday life and they can be of different types. A spatial relation reflects implicitly or explicitly an order of objects (e.g., the apple is to the left of the pear). Such a relation can be perceived by the visual system and additionally by other sensory systems that allow to perceive order such as 'touch' for instance. A pure visual relation in turn is only perceivable by the visual system and by no other perception system (e.g., the grass is greener than the tree). Visuospatial relations are perceivable by both, the visual and other systems, such as 'laying on top of something' - this is an ordered relation and at the same time the relation 'top' presupposes a surface perceived visually. Abstract relations are inconceivable by any sensory systems, such as mental attributions (e.g., smarter than) or abstract mathematical operations (e.g., ' $=$ ').

To the best of our knowledge, there is currently no study investigating such aspects of relational reasoning that directly relate to different degrees of imaginability of relations. To collect a sufficient number of data by often diverse study designs, we decided to use a broader, more inclusive relaxation on the data aggregation. Such a relaxation on the rather strict definitions is necessary since methods such as activation likelihood estimation (ALE) are more accurate for larger datasets (Eickhoff et al., 2016).

On the basis of this experimental background, we decided to investigate relational reasoning in general as well as two of its variants, visuospatial and abstract relational reasoning. Our hypotheses are: (1) For the abstract condition, we expect enhanced activation in the posterior parietal cortex (PPC) (Hobeika, Diard-Detoeuf, Garcin, Levy \& Volle,

2016; Maier, Ragni, Wenczel \& Franzmeier, 2014; Wendelken 2015) and the rostrolateral prefrontal cortex (RLPFC) which are regions known to be involved in the processing of abstract information (Christoff, Ream, Geddes \& Gabrieli, 2003). (2) We expect activation in the right superior parietal lobule (SPL) for the visuospatial condition. This is because it is hypothesized that visuospatial mental representations in reasoning are constructed and maintained by the help of these regions (Maier et al., 2014; Ragni, Franzmeier, Maier \& Knauff, 2016).

To study the neural correlates of relational reasoning we considered 38 experiments with a total of 692 participants and 884 foci to find the neural sites which are most likely to be active during relational reasoning in general and in visuospatial and abstract relational reasoning. The results are interpreted in the light of prior neurocognitive research.

Table 1: Overview of the experiments included in the metaanalysis with details about the experimental setup.

| Publication | Foci | Subj. | Rep. | Stim. |
| :---: | :---: | :---: | :---: | :---: |
| Goel et al., 1998 | 6 | 12 | vs | v , sen |
| Goel \& Dolan, 2001 | 36 | 14 | vs, ns | $v$, sen |
| Christoff et al., 2001 | 7 | 10 | vs | v, sha |
| Prabhakaran et al., 2001 | 202 | 7 | ab | v , sen |
| Knauff et al., 2002 | 16 | 12 | vs | aud, sen |
| Acuna et al., 2002 | 17 | 15 | vs | v , sha |
| Knauff \& JohnsonLaird, 2002 | 2 | 12 | $\mathrm{v} / \mathrm{v}, \mathrm{s}$ | aud, sen |
| Knauff et al., 2003 | 28 | 12 | $\begin{aligned} & \text { vs, v, } \\ & \text { s, ab } \end{aligned}$ | aud, sen |
| Ruff et al., 2003 | 20 | 12 | vs | aud, sen |
| Goel et al., 2004 | 19 | 14 | S | v , sen |
| Fangmeier et al., 2006 | 36 | 12 | vs | $v$, let |
| Green et al., 2006 | 2 | 14 | ab | v , wor |
| Lee et al., 2006 | 8 | 36 | vs | v, sha |
| Melrose et al., 2007 | 14 | 19 | v, s | v, sha |
| Wendelken et al., 2008 | 24 | 20 | ab | v , wor |
| Eslinger et al., 2009 | 17 | 16 | vs | v, sha |
| Fangmeier \& Knauff, 2009 | 21 | 12 | vs | aud, let |
| Goel et al., 2009 | 10 | 17 | v | v , sen |
| Prado, Noveck et al., 2010 | 3 | 15 | vs | $v$, sen |
| Wendelken \& Bunge, 2010 | 17 | 16 | vs | v, sha |
| Prado, van der <br> Henst et al., 2010 | 7 | 13 | vs | v , sen |
| Hinton et al., 2010 | 2 | 24 | v | v , let |


| Cho et al., 2010 | 9 | 17 | v | v |
| :---: | :---: | :---: | :---: | :---: |
| Hampshire et al., 2010 | 15 | 16 | ab | v , sha |
| Preusse et al., 2010 | 8 | 17 | vs | v , sha |
| Volle et al., 2010 | 68 | 16 | vs | v , let |
| Preusse et al., 2011 | 6 | 40 | vs | v , sha |
| Jia et al., 2011 | 39 | 20 | ab | v |
| Brzeziczka et al., 2011 | 40 | 17 | ab | v , let |
| Prado et al., 2012 | 3 | 30 | v | v , sen |
| Shokri-Kojori et al., 2012 | 20 | 20 | vS | v , sha |
| Watson \& Chatterjee, 2012 | 3 | 23 | vs | v , sha |
| Kalbfleisch, 2013 | 21 | 34 | vs | v , sha |
| Bazargani et al., 2014 | 12 | 37 | v | v , sha |
| Liang et al., 2014 | 24 | 23 | ab | v , let |
| Parkin et al., 2015 | 53 | 20 | vs | v , sha |
| Jia et al., 2015 | 24 | 15 | ab | v , num |
| Jia \& Liang, 2015 | 21 | 13 | ab | v , num |

Abbreviations: Subj.: Number of subjects, Rep.: Representation type, Stim.: Stimulus, vs: visuospatial, v: visual, ns: nonspatial, ab: abstract, spa: spatial, sen: sentence, aud: auditory, sha: shapes, let: letters.

## Methods

We apply the ALE method (Eickhoff, Bzdok, Laird, Kurth \& Fox, 2012) which has become a standard to conduct meta-analyses to investigate neural correlates (e.g., Hobeika et al., 2016). We include as neuroimaging methods functional magnet resonance imaging (fMRI) and positron emission tomography (PET) data.

## Paper Acquisition and Selection

For acquiring neuroimaging data, we conducted several online search queries via the online platforms PubMed, ScienceDirect and Google Scholar to find peer-reviewed fMRI and PET studies (see Table 1) between 1998 and 2017. We used the search terms 'fMRI OR PET OR Neuroimaging', 'relational OR transitive reasoning' and 'visual reasoning OR spatial reasoning OR visuospatial reasoning' in all queries and additionally for the query in Science direct 'healthy', and for Google Scholar 'MNI OR Talairach'. Additional articles were found via the reference lists of similar papers, the meta-analysis conducted by Prado, Chadha and Booth (2011) and the reviews by Knauff (2006) and Maier et al. (2014). Due to a review of the metaanalysis conducted by Prado, Chadha and Booth (2011), 'reasoning vs. baseline' conditions (such as fixation cross or maintenance tasks, see e.g., Wendelken, Nakhabenko, Donohue, Carter \& Bunge, 2008 and Ruff, Knauff, Fangmeier \& Spreer, 2003, respectively) as well as 'highvs. low-level reasoning' conditions were included since they
represent an aspect of reasoning. Additionally, experimental data were only included when they were reported in MNI or Talairach space and yielded from whole brain analyses.

## Paper Categorization

We decided to categorize the data along different axes: by the type of mental representation of the relation (abstract, spatial, visual, visuospatial, none), by the stimulus (letters, sentences, shapes, words, numbers) and the type of stimulus presentation (visual or auditory) (see Tables 1 and 2). We acknowledge that these differences may reflect differences in neural activation as well but chose to only consider the types of mental representation and subtraction for the sake of including more studies in each group and yielding more robust results.

When reviewing the articles, we realized that our definitions for the mental representations of relations did not fit to what can be found in actual studies. Because of that, we decided to lower our criteria so that we merged the groups 'visual', 'spatial' and 'visuospatial' to the group 'visuospatial' (see Table 2). Also, we redefined the criteria so that visuospatial relations are relations that are easy to mentally envision in a spatial and/or visual manner in the aforementioned sense. For abstract tasks, we included all tasks that are impossible to potentially perceive by senses, such as mathematical tasks and operators (e.g., ' $=$ ', '<', ' $>$ ').

## Activation Likelihood Estimation

ALE is an established method for conducting meta-analyses of neuroimaging data (Eickhoff et al., 2012). It is implemented in the statistical tool GingerALE ${ }^{1}$ (we used version 2.3.6) to determine the likelihood of individual brain regions activating for a specific task. GingerALE features the conduction of either single dataset analyses or conjunction and contrast analyses between datasets.

The ALE algorithm maps the stereotactic data of each experiment on a template brain to generate Modeled Activation (MA) maps. Since the reported data are as points, it reconstructs the scanning data by assigning each data point the center of a Gaussian distribution. The points are blurred by the full width at half maximum (FWHM) which is determined by the subject size of the respective dataset (Eickhoff et al., 2012). The MA maps are merged to render the final ALE file. For each voxel, the likelihood confidence of finding each value is calculated by neglecting spatial information from the dataset and analyzing the probabilities of values being part of an MA map. The information from the two files is combined and a threshold is applied to constitute the final ALE map (Eickhoff et al., 2012). In our analysis, we chose a standard setting of 1000 permutations, the cluster-level family-wise error (FWE) method with a $p$-value of 0.01 and an uncorrected $p$ of 0.001 (Eickhoff et al., 2016). The cluster-level FWE method generates a random dataset tantamount to the set at hand (regarding subject size, number of foci and number of

[^540]studies) which is compared to the actual data set. Foci originally represented in MNI (Montreal Neurological Institute) space were converted to Talairach space. In conjunction and contrast analyses, the ALE maps of two sets are examined in activation likelihood for their overlap and distinctness respectively.

Table 2: Details of the paper categorizations with regard to the quantitative parameters of the groups.

| Representation | Studies | Subjects | Foci |
| :--- | :--- | :--- | :--- |
| all | 38 | 692 | 884 |
| abstract | 10 | 179 | 394 |
| spatial | 2 | 16 | 23 |
| visual | 8 | 161 | 47 |
| visuospatial | 28 | 521 | 445 |
| none | 2 | 26 | 23 |

## Results

## Relational Reasoning

For the relational reasoning condition, activation was most likely found in the right precuneus (BA 7) and the left middle occipital gyrus (BA 31). Concerning the frontal lobe, the left inferior frontal gyrus (pars triangularis), left posterior-medial frontal, left and right middle frontal gyrus (BA 6) and the left middle orbital gyrus (BA 46) were found. Additionally, activation was found in the right basal ganglia (Tables 3 and 4).

Table 3: Overview of brain activation

|  | Frontal |  |  |  | Parietal |  | S | O |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 46 | 6 | 44 | 45 | 7 | 40 | 13 | 31 |
| Relational | - | O |  | - | , |  | , | 1 |
| Visuospatial |  | 0 | 1 | - | 0 |  |  |  |
| Abstract |  |  |  |  | - | 1 |  |  |

Note. Semicircles indicate significant clusters in the respective hemisphere. Filled halves indicate that the respective side's cluster was larger in this half. Abbreviations: S: Sub-Lobar, O: Occipital.

## Reasoning with Visuospatial Relations

In the visuospatial condition, activation was most likely in the left inferior frontal gyrus (triangularis, BA 44), posterior-medial frontal (BA 6), right and left middle frontal gyrus (BA 6) and the left inferior parietal lobule (hIP3, BA 7) and right superior parietal lobule (BA 7A).

## Reasoning with Abstract Relations

Activation in reasoning about abstract relations was found in the right angular gyrus and left superior parietal lobule (hIP $1 / 3$ respectively, both BA 7) and the left and right inferior frontal gyrus (triangularis, BA 45) and precentral gyrus (BA 45).

Table 4: ALE Results. Only significant clusters and a differentiated anatomical localization is reported.

| Macroanatomical Location | BA | Coordinates (Tal) |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  |  | X | Y | Z |
| Relational Reasoning Inference |  |  |  |  |
| R Precuneus | 7 | 0 | -61 | 43 |
| $\mathrm{L} \quad \mathrm{IFG}$ (p. Triangularis) | 45 | -45 | 13 | 32 |
| L Posterior-Medial Frontal | 6 | -1 | 11 | 48 |
| R IFG (p. Triangularis) | 9 | 45 | 18 | 32 |
| L Middle Frontal Gyrus | 6 | -27 | 0 | 53 |
| R Middle Frontal Gyrus | 6 | 28 | -1 | 52 |
| R Basal Ganglia | 13 | 27 | 23 | 3 |
| L Middle Orbital Gyrus | 46 | -41 | 43 | 6 |
| L Middle Occipital Gyrus | 31 | -27 | -77 | 25 |
| Relational Reasoning abstract |  |  |  |  |
| R Angular Gyrus | 7 | 23 | -65 | 42 |
| L Superior Parietal Lobule | 7 | -26 | -65 | 44 |
| $\mathrm{L} \quad \mathrm{IFG}$ (p. Triangularis) | 45 | -44 | 23 | 26 |
| L Precentral Gyrus | 44 | -43 | 3 | 33 |
| Relational Reasoning visuospatial |  |  |  |  |
| L IFG (p. Triangularis) | 44 | -46 | 12 | 33 |
| L Posterior-Medial Frontal | 6 | -1 | 12 | 47 |
| L Inferior Parietal Lobule | 7 | -35 | -55 | 44 |
| R Middle Frontal Gyrus | 6 | 28 | -1 | 53 |
| L Middle Frontal Gyrus | 6 | -26 | -1 | 56 |
| $\mathrm{R} \quad \mathrm{IFG}$ (p. Triangularis) | 45 | 46 | 25 | 33 |
| R Superior Parietal Lobule | 7 | 23 | -65 | 44 |

Abbreviations: BA: Brodmann Area, IFG: Inferior Frontal Gyrus.

## Discussion

## Relational Reasoning involves the fronto-parietal network and occipital lobe

For reasoning about relations, we found activation likelihood in the fronto-parietal network (right pars triangularis, posterior-medial frontal lobe, middle frontal gyrus, precuneus, left dorsolateral prefrontal cortex (DLPFC) and pars triangularis). These results are in accordance with the activation detected in the studies by Hobeika et al. (2016) and Prado, Chadha and Booth (2011). The largest cluster was found in the right precuneus (16800 $\mathrm{mm}^{3}$ ). The weighted center of this cluster is located in the frontal precuneus which is assumed to be involved in mental imagery (Cavanna \& Trimble, 2006). The second largest cluster ( $6256 \mathrm{~mm}^{3}$ ) was found in the left pars triangularis, also known as DLPFC. It is known to be involved in working memory and relational integration (Waltz et al., 1999), as well as speech and language production (Foundas, Eure, Luevano \& Weinberger, 1998)

Furthermore, activation was found in the right basal ganglia (BA 13), which are involved in reasoning and rule application (Melrose, Poulin \& Stern, 2007), a demand inductive reasoning tasks pose. Additionally, activation was found in the occipital lobe (left middle occipital gyrus). Prado, Chadha and Booth (2011) did not find such activation for relational reasoning, though they only considered deductive reasoning tasks. This reliable activation pattern might be due to the portion of tasks with visual contents which are not considered classical deduction tasks. In contrast to Prado, Chadha and Booth (2011), Hobeika et al. (2016) and Wendelken et al. (2008), we did not find any activation of the RLPFC. A rather surprising result when considering the consistent reports thereof in the literature.

## Visuospatial relational processing is executed by prefrontal activation

In visuospatial relational reasoning, the fronto-parietal network exhibited activation as well. Activation was mainly found in the left pars triangularis (BA 44) and posterior medial frontal (BA 9) and the inferior parietal lobule (hIP3) and right middle frontal gyrus (BA 45). This suggests that visuospatial relational reasoning is rather mediated by prefrontal activation, considering the multitude of clusters across the PFC. Also, parietal activation in the inferior parietal lobule (IPL) was found, suggesting that visuospatial processing does more heavily rely on context related processes than on mental imagery.

## Abstract relational reasoning relies on the intraparietal sulcus

In the abstract reasoning condition, we found activation in the right intraparietal sulcus (IPS) and left IPL. Since the analysis consisted of 10 studies only, the results are sparse. Nonetheless, they indicate that the IPL is essential for
abstract relational reasoning. Since the IPL is known to be involved in abstraction (Wurm \& Lingnau, 2015), this might imply the IPL's crucial role in the abstraction of contents from relational information.

## Conclusion

The meta-analysis unraveled some crucial details about the neural mechanisms of relational reasoning. Our results suggest that relational reasoning heavily relies on mental imagery and representation as well as a multitude of regions in the prefrontal cortex such as the DLPFC for relational integration and pars triangularis for language processing. No significant activation in the RLPFC was found, opposed to predictions by previous studies. We found striking differences between the type of representation of relations, so that visuospatial relations seem to rather rely on context, opposed to abstract relations which rely on abstraction of relation and mental imagery. Since the inclusion criteria concerning stimulus presentation and task requirements were relaxed, we assume that these areas mediate the general process of relational reasoning.

## Acknowledgments

Funding for this work was provided by the BrainLiksBrainTools Cluster of Excellence funded by the German Research Foundation (DFG, grant \#EXC1086). This work is additionally supported within DFG-Projects RA 1934/2-1, RA 1934/3-1, and RA 1934/4-1.

## References

Acuna, B. D., Eliassen, J. C., Donoghue, J. P., \& Sanes, J. N. (2002). Frontal and parietal lobe activation during transitive inference in humans. Cerebral Cortex, 12(12), 1312-1321.
Bazargani, N., Hillebrandt, H., Christoff, K., \& Dumontheil, I. (2014). Developmental changes in effective connectivity associated with relational reasoning. Human brain mapping, 35(7), 3262-3276.
Brzezicka, A., Sędek, G., Marchewka, A., Gola, M., Jednoróg, K., Królocki, L., \& Wróbel, A. (2010). A role for the right prefrontal and bilateral parietal cortex in four-term transitive reasoning: An fMRI study with abstract linear syllogism tasks. Acta neurobiologiae experimentalis, 71(4), 479-495.
Cavanna, A. E., \& Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. Brain, 129(3), 564-583.
Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., \& Gabrieli, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. Neuroimage, 14(5), 11361149.

Christoff, K., Ream, J. M., Geddes, L., \& Gabrieli, J. D. (2003). Evaluating self-generated information: anterior prefrontal contributions to human cognition. Behavioral neuroscience, 117(6), 1161-1168.

Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., ... \& Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. Cerebral cortex, 20(3), 524-533.
Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., \& Fox, P. T. (2012). Activation likelihood estimation metaanalysis revisited. Neuroimage, 59(3), 2349-2361.
Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox, P. T., ... \& Eickhoff, C. R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. Neuroimage, 137, 70-85.
Eslinger, P. J., Blair, C., Wang, J., Lipovsky, B., Realmuto, J., Baker, D., ... \& Yang, Q. X. (2009). Developmental shifts in fMRI activations during visuospatial relational reasoning. Brain and cognition, 69(1), 1-10.
Fangmeier, T., Knauff, M., Ruff, C. C., \& Sloutsky, V. (2006). fMRI evidence for a three-stage model of deductive reasoning. Journal of Cognitive Neuroscience, 18(3), 320-334.
Fangmeier, T., \& Knauff, M. (2009). Neural correlates of acoustic reasoning. Brain Research, 1249, 181-190.
Foundas, A. L., Eure, K. F., Luevano, L. F., \& Weinberger, D. R. (1998). MRI asymmetries of Broca's area: the pars triangularis and pars opercularis. Brain and language, 64(3), 282-296.
GingerALE 2.3.6 [Computer software]. (2016). Retrieved from http://www.brainmap.org/ale/
Goel, V., \& Dolan, R. J. (2001). Functional neuroanatomy of three-term relational reasoning. Neuropsychologia, 39(9), 901-909.
Goel, V., Gold, B., Kapur, S., \& Houle, S. (1998). Neuroanatomical correlates of human reasoning. Journal of Cognitive Neuroscience, 10(3), 293-302.
Goel, V., Makale, M., \& Grafman, J. (2004). The hippocampal system mediates logical reasoning about familiar spatial environments. Journal of Cognitive Neuroscience, 16(4), 654-664.
Goel, V., Stollstorff, M., Nakic, M., Knutson, K., \& Grafman, J. (2009). A role for right ventrolateral prefrontal cortex in reasoning about indeterminate relations. Neuropsychologia, 47(13), 2790-2797.
Green, A. E., Fugelsang, J. A., \& Dunbar, K. N. (2006). Automatic activation of categorical and abstract analogical relations in analogical reasoning. Memory \& cognition, 34(7), 1414-1421.
Hampshire, A., Thompson, R., Duncan, J., \& Owen, A. M. (2011). Lateral Prefrontal Cortex Subregions Make Dissociable Contributions during Fluid Reasoning. Cerebral Cortex, 21(1), 1-10.
Hinton, E. C., Dymond, S., Hecker, U. von, \& Evans, C. J. (2010). Neural correlates of relational reasoning and the symbolic distance effect: Involvement of parietal cortex. Neuroscience, 168(1), 138-148.
Hobeika, L., Diard-Detoeuf, C., Garcin, B., Levy, R., \& Volle, E. (2016). General and specialized brain correlates
for analogical reasoning: A meta-analysis of functional imaging studies. Human brain mapping, 37(5), 1953-69.
Jia, X., Liang, P., Lu, J., Yang, Y., Zhong, N., \& Li, K. (2011). Common and dissociable neural correlates associated with component processes of inductive reasoning. Neuroimage, 56(4), 2292-2299.
Jia, X., Liang, P., Shi, L., Wang, D., \& Li, K. (2015). Prefrontal and parietal activity is modulated by the rule complexity of inductive reasoning and can be predicted by a cognitive model. Neuropsychologia, 66, 67-74.
Jia, X., \& Liang, P. (2015). The Relationship of Four Brain Regions to an Information-Processing Model of Numerical Inductive Reasoning Process: An fMRI Study. Journal of Advanced Neuroscience Research, 2, 722.

Kalbfleisch, M. L., Debettencourt, M. T., Kopperman, R., Banasiak, M., Roberts, J. M., \& Halavi, M. (2012). Environmental influences on neural systems of relational complexity. Frontiers in psychology, 4, 631-631.
Knauff, M., Fangmeier, T., Ruff, C. C., \& Johnson-Laird, P. N. (2003). Reasoning, models, and images: behavioral measures and cortical activity. Journal of Cognitive Neuroscience, 15(4), 559-573.
Knauff, M., \& Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. Memory \& Cognition, 30(3), 363371.

Knauff, M., Mulack, T., Kassubek, J., Salih, H. R., \& Greenlee, M. W. (2002). Spatial imagery in deductive reasoning: a functional MRI study. Cognitive Brain Research, 13(2), 203-212.
Lee, K. H., Choi, Y. Y., Gray, J. R., Cho, S. H., Chae, J. H., Lee, S., \& Kim, K. (2006). Neural correlates of superior intelligence: stronger recruitment of posterior parietal cortex. Neuroimage, 29(2), 578-586.
Liang, P., Jia, X., Taatgen, N. A., Zhong, N., \& Li, K. (2014). Different strategies in solving series completion inductive reasoning problems: An fMRI and computational study. International Journal of Psychophysiology, 93(2), 253-260.
Maier, S. J., Ragni, M., Wenczel, F., \& Franzmeier, I. (2014). The role of the posterior parietal cortex in relational reasoning. Cognitive Processing, 15(1).
Melrose, R. J., Poulin, R. M., \& Stern, C. E. (2007). An fMRI investigation of the role of the basal ganglia in reasoning. Brain research, 1142, 146-158.
Parkin, B. L., Hellyer, P. J., Leech, R., \& Hampshire, A. (2015). Dynamic network mechanisms of relational integration. Journal of Neuroscience, 35(20), 7660-7673.
Prabhakaran, V., Rypma, B., \& Gabrieli, J. D. E. (2001). Neural substrates of mathematical reasoning: A functional magnetic resonance imaging study of neocortical activation during performance of the necessary arithmetic operations test. Neuropsychology, 15(1), 115-127.
Prado, J., Chadha, A., \& Booth, J. R. (2011). The brain network for deductive reasoning: A quantitative metaanalysis of 28 neuroimaging studies. Journal of Cognitive Neuroscience, 23(11), 3483-3497.

Prado, J., Mutreja, R., \& Booth, J. R. (2013). Fractionating the neural substrates of transitive reasoning: Taskdependent contributions of spatial and verbal representations. Cerebral Cortex, 23(3), 499-507.
Prado, J., Noveck, I. A., \& van der Henst, J. B. (2010). Overlapping and distinct neural representations of numbers and verbal transitive series. Cerebral Cortex, 20(3), 720-729.
Prado, J., van Der Henst, J.-B., \& Noveck, I. A. (2010). Recomposing a fragmented literature: How conditional and relational arguments engage different neural systems for deductive reasoning. Neuroimage, 51(3), 1213-1221.
Preusse, F., Van Der Meer, E., Deshpande, G., Krueger, F., \& Wartenburger, I. (2011). Fluid intelligence allows flexible recruitment of the parieto-frontal network in analogical reasoning. Frontiers in Human Neuroscience, 5(22), 1-14.
Preusse, F., van der Meer, E., Ullwer, D., Brucks, M., Krueger, F., \& Wartenburger, I. (2010). Long-term characteristics of analogical processing in high-school students with high fluid intelligence: an fMRI study. ZDM, 42(6), 635-647.
Ruff, C. C., Knauff, M., Fangmeier, T., \& Spreer, J. (2003). Reasoning and working memory: Common and distinct neuronal processes. Neuropsychologia, 41(9), 1241-1253.
Shokri-Kojori, E., Motes, M. A., Rypma, B., \& Krawczyk, D. C. (2012). The network architecture of cortical processing in visuo-spatial reasoning. Scientific Reports, 2, 1-7.
Volle, E., Gilbert, S. J., Benoit, R. G., \& Burgess, P. W. (2010). Specialization of the rostral prefrontal cortex for distinct analogy processes. Cerebral Cortex, 20(11), 2647-2659.
Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., ... \& Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. Psychological Science, 10(2), 119-125.
Watson, C. E., \& Chatterjee, A. (2012). A bilateral frontoparietal network underlies visuospatial analogical reasoning. Neuroimage, 59(3), 2831-2838.
Wendelken, C. (2015). Meta-analysis: how does posterior parietal cortex contribute to reasoning?. Frontiers in human neuroscience, 8, 1-11.
Wendelken, C., \& Bunge, S. A. (2010). Transitive inference: distinct contributions of rostrolateral prefrontal cortex and the hippocampus. Journal of Cognitive Neuroscience, 22(5), 837-847.
Wendelken, C., Nakhabenko, D., Donohue, S. E., Carter, C. S., \& Bunge, S. A. (2008). "Brain is to thought as stomach is to??": Investigating the role of rostrolateral prefrontal cortex in relational reasoning. Journal of Cognitive Neuroscience, 20(4), 682-693.
Wurm, M. F., \& Lingnau, A. (2015). Decoding actions at different levels of abstraction. Journal of Neuroscience, 35(20), 7727-7735.

# How the truth can make a great lie: An empirical investigation of the folk concept of lying by falsely implicating 

Alex Wiegmann (Alex.Wiegmann@psych.uni-goettingen.de) ${ }^{1 *}$<br>Pascale Willemsen (Pascale.Willemsen@rub.de) ${ }^{2 *}$<br>${ }^{1}$ University of Göttingen, Department of Cognitive Science, Gosslerstr. 14, 37170 Göttingen<br>${ }^{2}$ Institute for Philosophy II, Ruhr-University Bochum, Universitätsstraße 150, 44801 Bochum<br>* both authors contributed equally to this research


#### Abstract

Is it possible to lie despite not saying anyhing false? While the spontaneous answer seems to be 'no', there is some evidence from ordinary language that a lie does not require what is said to be believed-false. In this paper, we will argue for a pragmatic extension of the standard definition of lying. More specifically, we will present three experiments which show that people's concept of lying is not about what is said, but about what is implied by saying it that way. We test three Gricean conversational maxims. For each one of them we demonstrate that if a speaker implies something misleading, even by saying something semantically true, it is still considered lying.


Keywords: lying; concept of lying, deceiving; Grice; conversational implicature

## Introduction

According to the standard philosophical definition of lying, an agent lies if she makes "a believed-false statement to another person with the intention that that other person believes that statement to be true" (Mahon, 2008). Such a definition of lying entails four necessary conditions, namely the Statement Condition, the Untruhfulness Condition, the Addressee Condition, and the Intention-to-Deceive Condition. According to Intention-to-Deceive Condition, a lying agent aim to deceive. In order to deceive, however, the lie needs to be directed at someone capable of forming false beliefs (Addressee Condition; for a critical perspective see Rutschmann \& Wiegmann, 2017). The means to deceive the addressee is said to be a linguistic statement. The Statement Condition is not limited to verbal or written statements but further includes other linguistic symbols. Finally, this statement need to be uttered untruthfully. Untruthfulness does not require the uttered statement to be objectively false but that the speaker believes his statement to be false. As a consequence, an agent might be lying even if what she believes to be false turns out true. In an empirical study, Wiegmann, Samland, and Waldmann (2016) demonstrated that lay people's intuitions about lying are in line with the Untruthfulness Condition

However, there seem to be cases in which lies do not even require that the agent believes what is said to be false. Benjamin Franklin famously said that "Half a truth is often a great lie". So-called lies of omission seem quite frequent in ordinary conversations. The American elections 2016 provide many interesting examples in which both presidential candidates were criticized for lying, even though what they said was not, strictly speaking, false. Those cases include oversimplifications, using outdated or misleading statistics
(for instance on the murder rate, African American unemployment, or tax deficits), and suspiciously loose speech.

In this paper, we will empirically test the possibility that lying is not tied to semantic falsity but rather about false implicatures. We hypothesise that at the core of people's concept of lying is the discrepancy between what the speaker believes to be true and the belief she believes to create in the addressee by what she says or by what she implies by saying it.

## Conversational Implicatures

Many philosophers and linguists believe that what is relevant for a conversation is not only what we say but also what we imply by saying it. According to Grice (1975), every conversational context rests on the assumption that both speakers share a conversational goal, for instance making a decision on where to travel, or discussing various political opinions. Conversational goals can be introduced in various ways, most easily by direct questions, such as "Where shall we go for the summer holidays next year?". In light of this shared goal, both speakers can expect the other to be cooperative, that is to make their "conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which [they] are engaged." (Grice, 1975, p. 46) In order to make pursuing this joint conversation goal as efficient as possible, four Maxims should be obeyed:

1. Quantity

- Make your contribution as informative as required (for the current purpose of the exchange)
- Do not make your contribution more informative than is required

2. Quality

- Do not say what you believe to be false
- Do not say that for which you lack adequate evidence

3. Relation (Be relevant!)
4. Manner

- Avoid obscurity of expression
- Avoid ambiguity
- Be brief (avoid unnecessary prolixity)
- Be orderly

Violations of these maxims at the level of what is said are very typical in ordinary language-so typical that we are seldom surprised when people do it. Non-literal speech such as irony ("If you want to have a heavenly summer experience,

I recommend London! Warm and sunny all day!"), metaphors ("You're a peach!"), or hyperbole ("This is the best cake I have ever eaten in my entire life.") provide the most obvious examples in which people immediately infer what is meant beyond what is said (cf. Viebahn, in press). Such an inference is possible because we expect our conversational partner to be cooperative, and therefore interpret even violations of the conversational maxims as furthering the joint communicative goal (cf. Dinges, 2015).

## Lying and Falsely Implicating

If there is such a thing as lying by falsely implicating, it might work very similarly. A speaker utters something that violates a conversational maxim. The addressee immediately tries to infer what this utterance might contribute to the joint conversational goal. However, most crucially, in addition to violating a maxim, the speaker also violates the Cooperative Principle. For the addressee, there is yet no indication that the Cooperative Principle has been violated.
We suggest that lying is not only a matter of what is said but also of what is implied (cf. Meibauer, 2005). If a person intentionally violates a maxim in order to create a belief in the addressee which the speaker believes to be false, this violation is considered a case of lying, irrespective of what is said is true at a semantic level. If this hypothesis is correct, then the Untruhfulness Condition under its semantic reading is too limited to adequately capture the folk concept of lying. Rather, central to the folk concept of lying is a discrepancy between what the speaker believes to be true and what he believes to make the listener believe. Whether this belief is generated by what is said or what is implied by so saying seems to be secondary.

## Experiment 1: Violating the Maxim of Quantity

According to Grice, the Cooperative Principle requires the agent to make their contribution as informative as required, and to not omit relevant information. In our first experiment, we test whether violations of the Maxim of Quantity are considered lies. Our hypothesis presupposes that deceiving by violating conversational maxims is only possible of the addressee believes the speaker to be cooperative, that is to share with her a conversational goal. Consequently, we created vignettes in which to agents conversationally interact and agent A asks a question. Agent B, however, intentionally omits information which he believes to be relevant to agent $A$ but which might also get $B$ into trouble.

In the literature on lies of omission, two variants are discussed. On the one hand, a speaker might completely refrain from making a declarative statement. In our experiment, the speaker suggests changing the topic and asks a question in return. On the other hand, lies of omissions might also be told by using half-truths (Vincent and Castelfranchi, 1981). In the experimental condition Half-Truth, the speaker provides an answer to the question, and, thus, makes a statement. However, he omits the relevant, yet potentially troublesome piece of information. Again, the speaker is not untruthful as he believes his statement to be true. We predict
that participants will consider both scenarios instances of lying.

## Methods

The experiment was run online in the U.K (same in all following experiments) using Prolific Academics. 474 participants started the experiment, 451 were included in the analysis (we excluded participants who did not finish the survey, went through it in less than 40 seconds, or failed an attention check - we used the same criteria in all following experiments). $50 \%$ identified as male, $50 \%$ as female. Mean age was 32 years and participants earned $£ 0.20$ for their participation (same amount in all experiments).

Participants first read general instructions (same procedure in all following experiments). They were then randomly assigned to one of four conditions in a 2 (vignette: Couple vs. Car) * 2 (deception: ChangeTopic vs. Half-Truth) betweensubjects design. They were presented with one of two vignettes, namely either Couple or Car. The Couple vignette reads as follows.

Couple: Peter and Jane have been a couple for a year now. They are very happy and just moved in together. Peter trusts Jane, but he knows about her ex- fiancé Steven who still tries to win Jane back. Thus, Peter is very jealous and does not like Jane meeting Steven. Jane is sometimes thinking about getting back together with Steven. As they work in the same company, they have coffee from time to time to talk about their joint projects. Today, Jane and Steven have coffee after lunch to finalize a cost calculation they are supposed to send to their client the next morning. After a few minutes, Steven asks Jane if they could talk about each other and getting back together. Jane tells Steven that they don't have much time and need to focus on the project.

In the evening, Peter and Jane have dinner. Peter asks Jane "What did you do during your lunch break?"
In the ChangeTopic Condition, the story continued as follows:

To avoid mentioning that she saw Steven during lunch, Jane quickly changes the subject and says:
"Let's talk about something other than work. How was the football training?"

Just as Jane intended, Peter never asks about her lunch break and does not believe Jane met with Steven."

In the Half-Truth Condition, the ending was changed to:
"To avoid mentioning that she saw Steven during lunch, Jane tells Peter only half the truth, omitting her meeting with Steven:
"I had lunch at the cafeteria. Then I had a coffee and went back to the office. It was such a busy day."

Just as Jane intended, Peter never asks about her lunch break and does not believe Jane met with Steven.

After reading the vignettes, participants were asked: "Do you rather agree or disagree with the following statement:

Jane lied to Peter"; and they could choose between "I (rather) agree" and "I (rather) disagree" (same test question and response options in all following experiments except of Experiment 3c).

## Results and Discussion

The results of the experiment are summarized in Figure 1. In the Couple condition, only $30 \%$ ( 34 out of 113) consider Jane's behaviour as a case of lying in ChangeTopic. However, $65 \%$ (74 out of 113) did so in Half-Truth. The difference between ChangeTopic and Half-Truth in the Couple vignette is statistically significant, $\chi^{2}(\mathrm{df}=1, \mathrm{~N}=226)=28.37, \mathrm{p}<$ 0.001 . Moreover, agreement in ChangeTopic is significantly below chance ( $50 \%$ ) level (binomial test, 34 out of 113, test value $=.5, \mathrm{p}<.0 .0001$ ), whereas agreement in Half-Truth is significantly above chance level, (binomial test, 74 out of 113, test value $=.5, \mathrm{p}<.0 .01$ ).
In the Car condition, $56 \%$ ( 60 out of 107) agreed that Nick lied to Kathy when Nick changed the topic. In contrast, $69 \%$ ( 81 out of 118) did so in Half-Truth. The difference between ChangeTopic and Half-Truth in Car is marginally significant $\chi^{2}(\mathrm{df}=1, \mathrm{~N}=225)=3.79, \mathrm{p}=0.0516$. ChangeTopic is not significantly different from chance (binomial test, 60 out of 107, test value $=.5, p=.2459$ ), but Half-Truth is above chance level (binomial test, 81 out of 118, test value $=.5, \mathrm{p}<$ .0.0001). Comparing Car and Couple showed a significant difference for ChangeTopic, $\chi^{2}(\mathrm{df}=1, \mathrm{~N}=220)=15.17$, $\mathrm{p}<0.001$, but no difference for Half-Truth, $\chi^{2} \quad(\mathrm{df}=1$, $\mathrm{N}=231)=0.26, \mathrm{p}=0.61$.
The results of our first experiments allow for a more nuanced view on whether we can lie by omission. If an agent deceives by answering to a question but omits facts which are relevant to the question, the agent is judged to have lied. However, changing the topic and omitting an answer altogether is not considered a case of lying.


Figure 1: Proportions of lie judgments as a function of vignette and kind of deception in Experiment 1

## Experiment 2: Violating the Maxim of Relation

The Maxim of Relation requires a speaker to only provide relevant information. In this experiment, we altered the vignettes used in experiment 1 such that the speaker provides information which is completely irrelevant to the actual question, but which seems relevant. All information that is given is true, and, thus, the Untruthfulness Condition of the standard definition of lying is not met.

## Methods

220 participants started the experiment, 208 were included in the analysis ( $44 \%$ identified as male, $56 \%$ as female). Mean age was 34 years. Participants were randomly assigned to one
of two conditions (vignette: Couple vs. Car) in a betweensubjects design. Here is the vignette for Couple.

Couple: Peter and Jane have been a couple for a year now. They are very happy and just moved in together. Peter trusts Jane, but he knows about her ex-fiancé Steven who still tries to win Jane back. Thus, Peter is very jealous and does not like Jane meeting Steven. Jane is sometimes thinking about getting back together with Steven. As they work in the same company, they have coffee from time to time to talk about their joint projects.
Today, Jane and Steven have coffee after lunch to finalize a cost calculation they are supposed to send to their client the next morning. After a few minutes, Steven asks Jane if they could talk about each other and getting back together. Jane tells Steven that they don't have much time and need to focus on the project. Steven has been sick the whole week, but he has nevertheless been at work.
In the evening, Peter and Jane have dinner. Peter asks Jane:
"You told me about this project with your ex-fiancé. Did you see him today?"
To avoid confirming that she saw Steven during lunch, Jane says:
"Steven has been sick the whole week."
Just as Jane intended, Peter does not believe Jane met with Steven.

## Results and Discussion

The results of the experiment are summarized in Figure 2. Let us start with the Couple vignette. The clear majority of participants, $81 \%$ ( 84 out of 104 ) considered Jane's behaviour a case of lying. This proportion is significantly different from chance $(50 \%)$ level (binomial test, 84 out of 104, test value $=.5, \mathrm{p}<.0 .0001$ ). For the Car vignette, we obtained similar results. $76 \%$ (79 out of 104) considered Nick's behaviour a case of lying. This proportion is significantly different from chance ( $50 \%$ ) level (binomial test, 79 out of 104, test value $=.5, \mathrm{p}<.0 .0001$ ). Comparing Car and Couple showed no significant difference, $\chi^{2}(\mathrm{df}=1, \mathrm{~N}=208)=0.71$, $\mathrm{p}=0.40$. Again, even though the agent was being truthful under a said-based definition of lying, people consider the agents' responses as lies.


Figure 2: Proportions of lie judgments as a function of vignette in Experiment 2

## Experiment 3: Violating the Maxim of Manner

The Maxim of Manner is not so much concerned with what is said, but how it is said. In this experiment, we empirically test whether the violation of its sub-maxims lead to an answer being considered a lie. Experiment 3 is therefore divided into sub-chapters, with section a) investigating Ambiguity and

Context-Sensitivity, section b) Obscurity, and section c) the maxim of Order. In all conditions, what the speakers says is true under a semantic reading, and the speaker is being truthful. However, the speaker deceives the addressee by using ambiguous terms, or terms whose reference is highly context-dependent, or by presenting information in an order that does not reflect the temporal order of events.

In Experiment 3 c , we also address a potential objection to our experiments. It might be argued that the results we have gotten so far are a mere artefact of our experimental design. The argument might go like this: Lying is typically considered a case of deceiving or misleading. In all experiments presented so far, we only asked people whether they agreed or disagreed that the agent lied, and we did not give any alternative options. However, it is well-known that deceiving others for one's own benefit is morally condemned. Thus, participants might have agreed to lie statement only to express their disapproval, while not in fact believing that the agent lied. If we had also asked participants whether the agent deceived, agreement with the lie statement might decline. We believe that this point is well taken. In Experiment 3, we therefore added a new independent variable, namely the number of questions asked after the scenarios.

## Exp. 3a) Ambiguity

Methods 204 participants started the experiment, 197 were included in the analysis. $42 \%$ identified as male, $58 \%$ as female. Mean age was 34 years. Participants were randomly assigned to one of two conditions (vignette: Couple vs. Car) in a between-subjects design.
Car: Nick is a car salesman. He is currently trying to sell a small car. Nick's wife is the boss of the company. They have been a couple since high school. She drives the same model as the one Nick is trying to sell. She is satisfied with her car and the thinks that the boot is spacious enough.

Kathy is interested in buying the car. However, she needs the car for grocery shopping, and to take her two kids to their rugby matches. Thus, she needs a spacious boot. Kathy wonders if the car's boot is big enough for her daily needs. She asks:
"I always need to move a lot of stuff in the boot. This one looks rather small to me. Do you believe that the boot is big enough?"
Nick realizes that Kathy might not buy the car if he can't convince her that the boot is big enough for her daily needs. To avoid mentioning that he only knows one person who drives the car and who is very happy with the boot's size, Nick says:
"My wife has the same car and shared your worries, but then she was surprised how spacious the boot is. Also my boss has never had any problems with the boot. And even my first love in high school with her three kids says that the boot is big enough.".

Just as Nick intended, Kathy believes that Nick knows three different people who are satisfied with the size of the boot. She does not believe that the three people are in fact one and the same person.

## Results and Discussion

The results of the experiment are summarized in Figure 3. For the Couple vignette, the clear majority of participants, $83 \%$ (80 out of 96 ) considered Jane's behavior a case of lying. This proportion is significantly different from chance ( $50 \%$ ) level, (binomial test, 80 out of 96 , test value $=.5, \mathrm{p}<.0 .0001$ ). For the Car vignette, we obtained similar results. $80 \%$ ( 81 out of 101) considered Nick's behaviour a lie. This proportion is significantly different from chance ( $50 \%$ ) level (binomial test, 81 out of 101, test value $=.5, \mathrm{p}<.0 .0001$ ). Comparing Car and Couple revealed no significant difference, $\chi^{2} \quad(\mathrm{df}=1$, $\mathrm{N}=197$ ) $=0.32, \mathrm{p}=0.57$.

In both conditions, participants believed that the speaker told a lie, even though what the speaker said was true and believed to be true. The results put additional pressure on advocates of a semantically grounded understanding of the Untruthfulness condition and the Statement Condition.


Figure 3: Proportions of lie judgments as a function of vignette in Experiment 3a

## Exp. 3b) Obscurity

Methods 210 participants started the experiment, 206 were included in the analysis. $45 \%$ identified as male, $55 \%$ as female. Mean age was 34 years. Participants were randomly assigned to one of two conditions (vignette: Couple vs. Car) in a between-subjects design. They were presented with one of the following stories.

Car: Nick is a car salesman. He is currently trying to sell his brother's car which has 55,000 miles on it. Nick's brother needs to get as much money for the car as possible. For this reason, he set the milometer back to 25,000 miles before asking Nick to sell it. As a consequence of the manipulation, the milometer shows 30,000 miles less than actually driven. Nick has decided to ask for 5000 GBP. Nick knows that if a client learned about the manipulated milometer, Nick won't be able to sell the car for 5000 GBP.

Kathy is interested in buying the car. Kathy is surprised that a 10-year old car has only 25,000 miles on it. She says to Nick:
"Is this milometer accurate?"
To avoid mentioning that the milometer does not show the correct distance driven with this car (55,000 miles), Nick mentions that the milometer measures distances perfectly accurate, in the sense that if you drive one mile the tolerance is less than $0.1 \%$ (i.e., the milometer measures something between 1,758 and 1,762 yards):
"The milometer is super accurate and only allows for a $0.1 \%$ tolerance".

Just as Nick intended, Kathy believes that the milometer shows the correct distance driven with the car, and she does not believe that the milometer was set back.

Results and Discussion The results of the experiment are summarized in Figure 4. The clear majority of participants, $90 \%$ ( 93 out of 103) considered Jane's behavior a case of lying. This proportion is significantly different from chance ( $50 \%$ ) level (binomial test, 93 out of 103 , test value $=.5$, $\mathrm{p}<.0 .0001$ ).
For the Car vignette, the results are similar. 78\% (80 out of 103) agreed that Nick lied to Kathy. This proportion is significantly different from chance ( $50 \%$ ) level, (binomial test, 80 out of 103 , test value $=.5, \mathrm{p}<.0 .0001$ ).

Comparing Car and Couple showed a significant difference, $\chi^{2}(\mathrm{df}=1, \mathrm{~N}=206)=6.10, \mathrm{p}<0.05$, with higher lying rates for the Couple vignette.


Figure 4: Proportions of lie judgments as a function of vignette in Experiment 3b

## Exp. 3c) Order

Methods 407 participants started the experiment, 386 were included in the analysis. $46 \%$ identified as male, $54 \%$ as female. Mean age was 35 years.
Participants were randomly assigned to one of four conditions in a 2 (vignette: Couple vs. Car) x 2 (number of questions: OneQuestion vs. FiveQuestion) in a betweensubjects design.
Couple: Peter and Jane have been a couple for a year now. They are very happy and just moved in together. Peter trusts Jane, but he knows about her ex-fiancé Steven who still tries to win Jane back and works in the same company as Jane. Thus, Peter is very jealous and does not like Jane meeting Steven. Jane is sometimes thinking about getting back together with Steven.

Today, the traffic on Jane's way to work was very busy. At work, Jane always do the things that feels most important for her first and the things she does not consider important last. Peter knows about this habit. He also knows that today Jane is supposed to meet to talk about a joint project but also that the project is not very important. Today, Jane first wants to see Steven. This is not because she thinks that the project is the most important thing to do today but rather because she aches for Steven. So she visits him for a few minutes and discusses a few questions about a joint project. She then works on a big project of a client. Afterwards, she writes a couple of emails to clients and before she drives home she started writing an application for a higher position in her company.
In the evening, Peter and Jane have dinner. Peter wants to know about Jane's day:
"When did you meet Steven?"
To avoid mentioning that seeing Steven was the first thing she did, she does mention this event last:
"The traffic on my way to the company was really busy. I worked on this big project I told you about recently. I wrote a couple of emails to clients and started writing my application for the higher position in our company. I shortly visited Steven, discussed a few question about our joint project and drove home."

Just as Jane intended, Peter thinks that meeting Steven was the last thing she did at work.

In the OneQuestion Condition, participants were asked: "Do you rather agree or disagree with the following statement: Jane lied to Peter". Participants could choose between "I (rather) agree" and "I (rather) disagree". In FiveQuestions, participants were asked: "Do you rather agree or disagree with the following statement:

## Jane deceived Peter.

Jane's behaviour was morally bad.
Jane did not want to hurt Peter's feelings.
Jane's behaviour is blameworthy.
Jane lied to Peter.
And could choose for each statement between "I (rather) agree" and "I (rather) disagree".

Results and Discussion The results of the experiment are summarized in Figure 5. Let us start with the standard (One Question) Couple vignette. The clear majority of participants, ( $74 \%, 68$ out of 92 ) agreed that Jane lied to Peter. This proportion is significantly different from chance ( $50 \%$ ) level (binomial test, 68 out of 92 , test value $=.5, \mathrm{p}<.0 .0001$ ). In the 5Question variant of the Couple vignette, we found a similar pattern. The clear majority of participants ( $75 \%, 74$ out of 99) considered Jane's behaviour a case of lying. This proportion is significantly different from chance (50\%) level, (binomial test, 74 out of 99 , test value $=.5, \mathrm{p}<.0 .0001$ ). There was no significant difference between the OneQuestion and the FiveQuestion variant of the Couple vignette, $\chi^{2}(\mathrm{df}=1$, $\mathrm{N}=191$ ) $=0.02, \mathrm{p}=0.90$.

In the standard (OneQuestion) Car condition, the clear majority of participants ( $86 \%, 83$ out of 96) considered Nicks's behaviour a case of lying. This proportion is significantly different from chance ( $50 \%$ ) level (binomial test, 83 out of 96 , test value $=.5, \mathrm{p}<.0 .0001)$.

This pattern was similar for the 5Question variant of the Car vignette. The clear majority of participants $(79 \%, 78$ out of 99) considered Nick's behaviour a case of lying. This proportion is significantly different from chance (50\%) level (binomial test, 78 out of 99 , test value $=.5, p<.0 .0001$ ). There was no significant difference between the OneQuestion and the FiveQuestion variant of the Couple vignette, $\chi^{2}(\mathrm{df}=1$, $\mathrm{N}=195)=1.992, \mathrm{p}=0.16$.

Deceiving by mentioning the relevant facts in reversed order was considered lying. Furthermore, providing participants with the opportunity to express their moral evaluation of the agent and allowing them to indicate that the agent did only deceive another person (in contrast to: lied to) did not affect lie judgments.


Figure 5: Proportions of lie judgments as a function of vignette and number of questions in Experiment 3c

## General Discussion

Do people consider utterances that are not semantically wrong but pragmatically misleading lies? In this paper, we showed that for something to be a lie, subjective falsity at the semantic level is not necessary. A speaker might say something which is both true and which he believes to be true. What seems to be at the heart of people's concept of lying is that the speaker believes to create a belief in the addressee which he himself believes to be false. Whether this false belief is the result of a wrong statement or a false implicature seems to be secondary. In three experiments, we tested whether the violation of three Conversational Maxims would lead to something semantically true being considered a lie. For all the Maxims (Quantity, Relation, and Manner) this effect showed.

When the speaker violates the Maxim of Quantity by omitting relevant information (Exp. 1), participants considered such a statement to be a lie. These results support those philosophers and linguists who have argued for lies of omissions to be actual lies. However, those results put pressure on a semantically grounded understanding of the standard definition of lying and on authors who have denied that lies of omissions can be actual lies (Mahon 2003; Dynel 2011). Furthermore, Experiments 2 and 3 also indicate that said-based definitions cannot account for people's concept of lying. Our results rather indicate that lying occurs at the level of pragmatics, by deceiving others through falsely implicating.
There are two argumentative lines one might want to argue for in line of our results. The most radical way to deal with our results is to reject the standard definition of lying and to search for a radically new definition that focuses on pragmatics alone. We believe such a dismissal of the standard definition to be too rash. The standard definition seems to adequately capture the folk's intuitions in most cases of lying. However, things get messier around the edges. Alternatively, we suggest a reinterpretation that allows us to adequately map folk intuitions by making as few changes as possible. First, the Untruthfulness Condition seems more appropriately understood at the level of pragmatics. In line with previous research, we suggest maintaining a subjective understanding of untruthfulness, but to decouple it from what is said in the semantic sense. Accordingly, an agent is being untruthful if there is a discrepancy between what the agent believes to be
true and what he believes to communicate by saying something. Second, such an adaption allows for untruthfulness to be demonstrated without making a statement. If our account is appropriate, we believe that an agent may lie by falsely implicating by answering with a question in return ("Did you see your ex-fiancé today?""Are going to ask me this question every day now? What is wrong with you?), by requesting ("Don't you ever ask me this again!"), etc. Additional research is required on the extent to which we should re-interpret the Statement Condition in a more pragmatic fashion.

## Acknowledgments

This project was supported by the Leibniz Association through funding for the Leibniz ScienceCampus Primate Cognition at Georg-August University Goettingen., the Konrad-Adenauer Foundation, and the Research School at Ruhr-University Bochum.

We would like to thank Emanuel Viebahn for challenging discussions and his comments on a previous version of this paper.

## References

Dinges, A. (2015). Innocent implicatures. Journal of Pragmatics, 87, 54-63.
Dynel, M. (2011). A web of deceit: A Neo-Gricean view on types of verbal deception, International Review of Pragmatics, 3: 139-167.
Franklin, B. (1758). Poor Richard improved. Accessed Dec. 16, 2013 via USinfo.org maintained by the U.S. Department of State Bureau of International Information Programs, U.S. Department of State.
Grice, P. (1975). Logic and conversation. In P Cole \& J. L. Morgan (Eds.), Syntax and Semantics, Volume 3: Speech Acts (pp. 41-58). New York: Academic Press.
Mahon, J. E. (2003). 'Kant on lies, candour and reticence,' Kantian Review, 7: 101-133.
Mahon, J. E. (2016). The definition of lying and deception. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2016).
Meibauer, J. (2005). Lying and falsely implicating. Journal of Pragmatics, 37, 1373-1399.
Rutschmann, R., \& Wiegmann, A. (2017): No need for an intention to deceive? Challenging the traditional definition of lying. Philosophical Psychology. Advanced online publication.
Viebahn, E. (in press). Non-literal lies. Erkenntnis.
Vincent, J. M., \& Castelfranchi, C. (1981). On the art of deception: How to lie while saying the truth. Possibilities and limitations of pragmatics, 8-14.
Wiegmann, A., Samland, J., \& Waldmann, M. R. (2016). Lying despite telling the truth. Cognition, 150, 37-42.

# Eliciting Middle School Students’ Ideas About Graphs Supports Their Learning from a Computer Model 

Eliane Stampfer Wiese (eliane.wiese@berkeley.edu)<br>Graduate School of Education, 4407 Tolman Hall, Berkeley, CA 94720 USA<br>Anna N. Rafferty (arafferty@carleton.edu)<br>Computer Science Department, One North College Street, Northfield, MN 55057 USA<br>Marcia C. Linn (mclinn@berkeley.edu)<br>Graduate School of Education, 4611 Tolman Hall, Berkeley, CA 94720 USA


#### Abstract

When middle school students learn science content with graphs, the graphing and science knowledge may be mutually reinforcing: understanding the science content may help students interpret a related graph, and information from a graph may illustrate a scientific concept. We examine this relationship between graphing and science by studying how students learn from interactive computer models with accompanying data graphs. The computer models provide an animated simulation that illustrates an unobservable phenomenon, while the data graph tracks one or more quantities over time. This ordering study, on middle school students learning about photosynthesis, indicates that engaging with novel graph concepts helped students interpret their data as they experimented with the computer model. The study also provided some support for the opposite direction: experimenting with the model first helped students make sense of the graphs.


Keywords: Graphing; Photosynthesis; Knowledge Integration

## Learning From Graphs in Science

Graphs are important in science and improving students' integration of graph and science concepts is often neglected, especially at the middle school level (Lai et al. 2016). Graphs can help students test hypotheses to make sense of new science concepts (e.g., Vitale, Madhok, \& Linn, 2016). When linked to an animated simulation, time-series graphs can record and summarize information that points to key relationships. Identifying the main points of a computer simulation is not trivial for students: animations can be deceptively clear, giving a false sense of understanding (Chiu, Chen, \& Linn, 2013), and students may focus on discrete events instead of looking for an underlying pattern (Vitale et al., 2016). While animations provide visualizations of mechanisms and processes, graphs provide visualizations of relationships between variables. Thus, graphs may help students focus on the underlying quantitative relationships in the animations.

While graphs have the potential to be extremely useful for learning science content, middle school students often have difficulty interpreting them (Lai et al., 2016). Graphs are generally taught in math classes, and students may not spontaneously transfer their knowledge to science contexts (Grant, 2013). Prior work notes common student problems in graph interpretation, many stemming from misunderstanding what the graph axes represent (e.g., Clement, 1985). This type of confusion may result from shallow reasoning based on superficial similarity (e.g., Janvier, 1981) or the ease of extracting
irrelevant data from the graph (e.g., Clement, 1985, McDermott, Rosenquist, Popp, \& van Zee, 1983). Therefore, while graphs can be powerful tools, novices need support to work with them. Misinterpretations of graphs could lead students to misunderstand the science being conveyed. On the other hand, if students have a firm grasp of the science content, that knowledge may help them interpret the graph by constraining potential interpretations to ones that are consistent with their science knowledge (Ainsworth, 2006). It is therefore plausible that graph and science knowledge could be mutually reinforcing. To examine if this reinforcement is uni- or bidirectional, we assigned students to see graph-focused steps either before or after an animated computer model of plant growth. Graph First students learned more science from the simulation, with some advantage for Model First students in interpreting a specific graph feature.

## Photosynthesis in the WISE Platform

The ordering experiment on graphing was implemented via a seventh-grade unit on photosynthesis, which was designed with the Knowledge Integration (KI) framework (Linn, Lee, Tinker, Husic, \& Chiu, 2006), using the Web-Based Inquiry Science Environment (WISE: http://www.wise.berkeley.edu). The unit focuses on processes of energy transformation and aligns with the Next Generation Science Standards (NGSS Lead States, 2013). In photosynthesis, glucose is created and stores energy. The energy is released during cellular respiration. A key difference between photosynthesis (making glucose) and cellular respiration (using glucose) is that plants can only make glucose when there is light, but they must use glucose all the time to perform basic cellular functions. This concept is targeted in the plant growth activities (Figure 1.

The plant growth activities begin with predicting when the plant will make and use glucose (multiple-choice text selection). Next, students interact with the animation from the plant growth model, which shows the plant's glucose stores growing when the light is on and shrinking when the light is off. Students then draw a graph to show their predictions for the cumulative glucose that will be made, used, and stored with the light on for a few weeks and then off, and also select which graph shapes match their predictions (multiple choice: all lines increase with light. In the dark, one is horizontal, one decreases, one increases at the same rate, and one in-


Figure 1: Sequence of activities in Studies 1 and 2.
creases more sharply). Finally, students see the experimental steps: interacting with the full plant growth model (with the graph) and graph interpretation. The two experimental steps, described below, are ordered randomly. Test questions assessed how students connect ideas. As shown in Figure 3a, one prompt presented a graph of glucose stored and asked students to explain how turning the light off affected the shape of the graph (Glucose Stored). Another prompt gave an incorrect prediction: that glucose stored would not change when the light was off (Marcia's Prediction). Students explained if the prediction was correct and matched it to a graph (Figure 3b). These questions were intended to measure how well students can connect a graph's shape to its meaning.

## Exploring the Plant Growth Model

The plant growth model allows students to turn a light on or off, while the accompanying time-series graph shows the cumulative amounts of glucose made, used, and stored (Figure 2F). In this model, glucose is made at a constant rate when the light is on, and glucose is used at a constant rate all the time. Students were again asked when the plant makes and uses glucose. The cumulative graph shows these relationships directly: the running total for glucose made only increases while the light is on, but the running total for glucose used increases until the plant dies.

Students were also asked to run a specific trial with the model: leave the light on for 4 weeks in a row, and then turn it off. This trial shows that plants draw on stored glucose to survive in the dark. To structure students' thinking about this trial, students were asked when the plant died, and what best explained why the plant died (Figure 2k). Correctly answering both questions requires reading the graph.

## Ordering Study: Graph First vs. Model First

The graph that accompanies the photosynthesis model is complex: it is cumulative rather than instantaneous and it presents three quantities changing over time. Science knowledge may help students avoid the error of reading the graph as instantaneous (e.g., the line for glucose made is not zero in the dark). While science knowledge could help students interpret the graph, the graph may also help students learn science by illustrating a relationship between quantities of glucose made, used, and stored that explains how plants survive periods of darkness. The KI framework proposes that eliciting students' ideas is crucial for learning because ideas
cannot be leveraged or inspected if they are not first brought to the student's attention. We designed a graph interpretation step to elicit students' ideas about graphs and to prompt them to connect their ideas about a graph's shape to its meaning. Questions asked if an interpretation of a given graph was correct (e.g., the plant is making glucose when the glucose made line is flat). The graph interpretation step was more extensive in Study 1, and is described in more detail for each study.

To examine the effects of graph and science knowledge on each other, we sequenced the graph interpretation step either directly before or after the computer model. Two seventh grade teachers at Bay Area public schools used this WISE unit to teach photosynthesis, and their students were randomly assigned to either the Graph First or Model First sequence. Students did pre- and post-tests individually, and did the unit individually or in teacher-assigned pairs ${ }^{1}$

## Study 1: Extensive Elicitation of Graph Ideas

We ran the first iteration of the study in six 7th grade classes, all taught by Mrs. R ${ }^{2}$ Students completed a pretest, but it is not included in the analyses because it was not matched to the post-test. Since the pre-test and the unit took longer than Mrs. R. expected, she requested a revised and shortened post-test.

Materials. The post-tes ${ }^{3}$ included the questions Glucose Stored and Marcia's Prediction (Figure 3). Students could access the plant growth model on both the pre- and post-test. The experimental step to elicit students' graph ideas asked students to interpret four cumulative graphs, all showing glucose made, used, and stored when the light was shining (three also showed darkness). Two graphs were consistent with the model and two were incorrect. For each graph, students were given one or two interpretations and asked if those interpretations were correct. In total, the elicitation step included five true/false items (one shown in Figure 2 a ) and three openended prompts for students to explain their reasoning.

[^541]

Figure 2: Screenshots. (a) A true/false graph interpretation item in Study 1. (b) The graph interpretation item in Study 2. (c) The plant growth model, with an animation on the left and graphed the output on the right. Correct answers in bold.

Method and Participants. The pretest was administered over 1-2 class days before students began the unit. Students then worked on the unit for seven class days. Most students worked in teacher-assigned pairs, and they were randomly assigned to the Graph First or Model First conditions (withinclass). Most students completed the model-related steps over two class periods. The posttest was administered two school days after the unit, and took one class period.

177 students across Mrs. R's six classes worked on the WISE unit, with 24 to 35 students per class. Five students worked individually, with the rest in pairs. The 89 groups (176 students; 43 Graph First and 46 Model First) who did the experimental activities are included in the analyses of embedded items. Of those, 173 did the posttest and are included in the post-test analyses.

Results: Embedded Assessments. 96\% of groups correctly selected that the plant would make glucose only when the light was on (Graph First: 98\%, Model First: 93\%). 71\% of groups correctly selected that the plant would use glucose all the time (Graph First: 72\%, Model First: 70\%). Students selected which of four shapes (or none) best matched their ideas for glucose made, used, and stored (all lines increase when the light is on. When the light is off, one increases at the same rate, one increases more sharply, one decreases, and one is horizontal.). Most students selected the correct shapes for glucose made and stored, but only $43 \%$ did so for glucose used. The most common incorrect shape was flat after the light was turned off ( $22 \%$ ). Of students who predicted, in text, that plants would use glucose all the time, only $55 \%$ selected the correct graph. Thus either students hold multiple conflicting ideas about when plants use glucose or are unable to connect their science ideas to the graph shape.

Students were engaged with the model; $93 \%$ of groups ran at least one trial where the plant died. However, students' choice to run the suggested trial (light on for four weeks, then off) differed by condition: $88.4 \%$ of Graph First groups vs. $71.7 \%$ of Model First groups. 70\% of Graph First groups and $45 \%$ of Model First groups chose the correct explana-
tion for why the plant died. A logistic regression on answer correctness (with factors for conditions in this study and the unrelated study) confirmed that there was a main effect of condition on whether they responded correctly $(t(86)=2.27$, $p=.023$ ). A logistic regression on whether students ran the suggested trial found a marginally significant effect of condition $(t(86)=1.92, p=.055)$. To determine if running the trial mediated the effect of condition, we re-ran the previous logistic regression on students' answers with running the trial as an additional factor. Running the trial was the only significant factor $(t(85)=3.20, p=.0014 ; 67.6 \%$ correct if trial was run vs. $16.7 \%$ correct without this trial), indicating that running the trial mediated the effect of condition. Thus, structuring exploration with a specific trial is more effective for learning than non-structured exploration, even when both uncover the key information (i.e. the plant dying).

To determine if lack of engagement in general led students to ignore the directions, we coded all five open-response items within the activity (but before the experimental steps) as on-task (answering the question, correctly or not) or offtask (blank, I don't know, or nonsense answers). 20 responses were coded by two authors with $100 \%$ agreement; remaining responses were coded by one of the authors. Both conditions showed high engagement, with means of $98 \%$ on-task responses (means were 4.91/5 for Graph First, 4.89/5 for Model First, $t(87)=.16, p=.87$ ). We repeated the logistic regression on students' answers with all covariates above and engagement as an additional covariate $(t(84)=0.68, p=.50)$ and found that only running the suggested trial was significant $(t(84)=3.39, p=.0007)$. Repeating the logistic regression on whether students ran the suggested trial with factors for conditions and engagement also showed the same effect: ordering condition was marginally significant $(t(85)=1.90$, $p=.058$ ), and engagement was not $(t(85)=0.54, p=.58)$.

Performance on the graph items did not show differences across conditions, suggesting that interacting with the model first did improve graph interpretation in general. Two questions focused on interpretation of a flat line in a cumulative

2: Vague, incorrect, or not focused on stored glucose. (1) When the light was turned off less glucose was used. (2) no because, if you leave the light off it will die
3: Partially correct or simple. (1) The total glucose stored was steadily going up then when the light turned off it went down. (2) Marcia's prediction is not correct. Even though the plant will not make glucose or use anymore, the amount of glucose stored would go down and so will the glucose made.
4S: Plants survive in the dark by using stored glucose. (1) when the light turned off the plant did not make glucose but only used the glucose that the plant stored. (2) No, a plant to use glucose needs glucose. And plant don't get the needed sunlight when the light is off so has to eat into the glucose stored.
4G: Explanation of graph shapes. (1) Turning the light off in addition to causing the glucose stored to go down, it would also cause the glucose made to stop increasing and just stay the same. Turning the light off would not, however, effect the total glucose used, which would keep going up. (2) Marcia's prediction is incorrect because the glucose stored line in the graph goes down and doesn't stay level.

5: Relating the graph's shape to the plant's use of stored glucose. (1) It affected the shape of the graph because the plant cannot make glucose, so the plant has to use the stored glucose. This makes the stored glucose go down in quantity. (2) No, because if the plant doesn't make any at night and still uses it where is it coming from. It must be the storage and that must make it go down.

Table 1: Rubric for scoring (1) Glucose Stored and (2) Marcia's Prediction (0 is blank, 1 is off-topic or "don't know").
graph. Groups in the Graph First condition had lower average scores on these items than groups in the Model First condition (average score $55 \%$ vs. $80 \%$; ANOVA with factors for conditions in both studies found a main effect of ordering condition: $F(1,88)=13.23, p<.001)$. Both conditions scored near chance for the other items, $46 \%-63 \%$.

Results: Post-Test. The test items Glucose Stored and Marcia's Prediction each have one question with an objectively correct answer (when is the light turned off and which graph matches Marcia's prediction) and one open-response that encourages connections between graph and science ideas. The objective questions were scored as correct or not. For the open-ended questions, we developed KI rubrics that required linking the graph shape to processes of plant growth to get the highest score (a 5). Full explanations of either the graph or the science concepts were scored a 4, and were coded for graph or science ideas (see Table 1 ). Two of the authors coded an initial set of responses for each question, reconciling differences and revising the rubric. The same two authors coded a second test set for each question ( 39 answers for Glucose Stored, with $85 \%$ agreement and Cohen's kappa of $.79 ; 31$ answers for Marcia's Prediction, with $90 \%$ agreement and Cohen's kappa of .85). For both questions, the second test sets included responses from both studies.

For Glucose Stored, most students correctly identified that the light was turned off in week 6 (Graph First: 88\%, Model First: $90 \%$ ). Mean scores for their explanations of how turning the light off affected the shape of the graph were 2.9 overall ( 2.94 for Graph First and 2.85 for Model First). All students scored a 2 or higher, indicating engagement with the task ( 0 s are blank and 1 s are off task). $24 \%$ of the Graph First and $21 \%$ of the Model First condition scored a 4 or 5 .

For Marcia's Prediction, students were below ceiling but above chance for selecting the graph that matched her incorrect prediction (Graph First: 62\% correct; Model First: 57\% correct). Mean scores for their explanations of why the prediction was correct or not were 3.4 overall ( 3.5 for Graph First and 3.2 for Model First). $98 \%$ of students gave on-task answers. $60 \%$ of the Graph First and $50 \%$ of the Model First
condition scored a 4 or 5. A MANOVA on the two explanation items with conditions in both studies as factors showed no significant difference by condition for either study (ordering condition: $F(2,168)=1.2, p=.3)$.

## Study 2: Targeted Elicitation of Graph Ideas

Materials. Since Mrs. R's students took longer than we anticipated to complete the experiment, we shortened the materials before repeating the study with Mr. W's classes. The first questions on the pre- and post-test were the same as on the post-test in Study 1, with additional questions afterward. In the unit, on steps that all students saw in the same order, we removed open-response explanation prompts or replaced them with multiple-choice questions. Since Mrs. R's students were confused about the flat segment of the glucose made line, we added a cumulative graph activity. Students were given a table of glucose made each week and filled in a column with the running total. This running total was then shown as a graph, and students were asked to interpret the flat segment. In Study 1, students drew prediction graphs and then characterized the shapes afterward. In Study 2, we reversed this sequence to help students plan their predictions. We also shortened the steps in the ordering experiment. Instead of eliciting ideas about several graphs, we asked students to interpret the flat segment of the glucose made graph with a multiple choice question and an open-response explanation (see Figure 2b). In the modeling activity we removed all open-ended prompts, and replaced them with one prompt that asked students if their initial predictions on when the plant would make and use glucose had been correct. We retained all multiple choice questions.

Method and Participants. As in Study 1, the pre- and posttest were done individually, each on a single class day, with the pretest two days before students started the unit and the posttest right after. Students did the unit in groups of 1-2 over five class days, with a one-week gap between the first three days and the final two days. As in Study 1, most students took 1-2 class days for the experimental steps.

79 students across Mr. W's 7th grade classes used the

(b) Marcia made this prediction: (1) When the light is off: the amount of glucose stored will not change, the plant will not make more glucose, and the plant will still use glucose. (2) When the light is on, the plant will make, use, and store glucose.


Figure 3: (a) Glucose Stored requires graph interpretation and science knowledge of when a plant makes and uses glucose. (b) Marcia's Prediction requires finding a graph shape to show an idea, and critiquing that idea with science knowledge.

WISE unit, with 23-29 students per class. As in Study 1, groups were assigned randomly to conditions (within-class). 24 groups did not reach the experimental steps and are not included in the analysis. The 10 Graph First groups and 18 Model First groups are included in the analysis below (43 students total. 11 did the unit individually, the rest in pairs). Of those students, 39 did both the pre- and post-test and are included in those analyses (13 Graph First and 26 Model First).

Results: Embedded Assessments. 79\% of groups correctly predicted when the plant would make glucose (Graph First: $90 \%$, Model First: $72 \%$ ) and $29 \%$ correctly predicted when the plant would use glucose (Graph First: 30\%, Model First: $27 \%$ ). Students were not above chance ( $11-25 \%$ ) in selecting which graph shapes matched glucose made, used, and stored. In the most common shape for total glucose made ( $35 \%$ ), the line decreased in the dark, a nonsensical answer for cumulative glucose made. Students with the correct text predictions were somewhat more likely to select the correct graphs, but still performed poorly ( $14 \%$ correct for glucose made and $25 \%$ for stored). These predictions suggest that students struggled both with the science content and the graphs.

Only $75 \%$ of groups interacted with the model. To measured engagement, we coded for on-task responses on earlier step. On the two earlier steps, overall $75 \%$ of responses were on-task (means were 1.6/2 for Graph First, 1.4/2 for model first, $t(26)=.56, p=.57$ ). Despite lower engagement, we ran the same analyses as in Study 1. Graph First groups ran the suggested trial more frequently than Model First (50\% vs. $27.8 \%$ ), but were not more correct in explaining why the plant died ( $10 \%$ vs. $11 \%$ ). These trends were not significant for condition (all analyses here are logistic regressions with factors for conditions in both studies; ordering condition $t(25)=1.17, p=.24$ for suggested trial and $t(25)=0.0$, $p=1.0$ for answering correctly). Of the 10 groups who ran the trial, 3 answered correctly, compared with none of the 18 groups who did not run the trial, but this trend was not significant $(t(24)=.0003, p=.99)$. We repeated the logistic regression on running the trial with engagement as an additional covariate, which was not significant $(t(23)=0.96$, $p=.33)$. The replication of trends but not reliable effects may be due to small sample sizes, lack of engagement, or lack of prior knowledge, leading to very low performance on many items. Performance on the graph interpretation item
was above chance but below ceiling; 54\% of groups correctly said that the plant was not making glucose when the line was flat. There was no significant difference by condition.

Results: Pre- and Post-tests. On Glucose Stored, identification of when the light was turned off rose from pre- to posttest (from 4/13 to 9/13 for Graph First, and from 12/26 to 16/26 for Model First; McNemar test from pre- to posttest, $p=.02$ ). A logistic regression on posttest scores with pretest score and study conditions as covariates was not significant for ordering condition. Mean explanation scores also rose from preto post-test (Graph First: 1.2 to 2.3, Model First: 1.4 to 1.9). For selecting which graph matched Marcia's Prediction, students were below chance at pretest ( $2 / 13$ correct for Graph First, $5 / 26$ for Model First) and just above chance at post-test (Graph First: 4/13 correct, Model First: 7/26). Explanation scores rose from pre- to post-test for the Graph First group (. 6 to 1.6) but not for Model First (. 8 to .7). $82 \%$ of responses at pre-test were blank or off task, with the remaining students answering with incorrect or overly simple ideas. At post-test, 1 Model First and 3 Graph First students had a complex idea about the graph or the science.

A full factorial repeated measures ANOVA with test time, question, and conditions in both studies indicates improvement from pre- to posttest on the explanation items, with significant factors for test time $(F(1,36)=10.3, p=.003)$, question $(F(1,36)=36.7, p<.001)$, and a time by ordering condition interaction $(F(1,36)=4.8, p=.035)$. While estimated marginal means for Graph First were lower than Model First at pretest (.9 vs. 1.1) and higher at posttest (1.9 vs. 1.3), ordering condition was not significant in a MANCOVA on the two post-test explanation scores (with pretest scores as covariates and condition in the unrelated study as a factor).

## Discussion

This study provided support for graph understanding helping students to interpret a science model and some support for science understanding helping students interpret specific graph features. While students had difficulty interpreting cumulative graphs, eliciting students' ideas about those graphs helped them make sense of glucose production, use, and storage in the photosynthesis model.

Graph First students were more likely to figure out that the plant dies when it runs out of stored glucose. Interestingly,
this effect was strongly mediated by whether students ran the suggested trial leading to plant death (significant in Study 1, with the same pattern in Study 2). The finding that elicitation of graph ideas, without feedback, improved learning from the model is consistent with other research on knowledge integration (Linn et al., 2006). The cumulative graph in the model was more familiar to Graph First students since they had just been examining cumulative graphs. To run the suggested trial, students needed to read the graph to turn the light off at the right week, and Graph First students may have extracted this information better than Model First students. Furthermore, students in both conditions were equally engaged on prior steps and engagement was not predictive of following the directions to run the suggested trial. We note that condition differences were significant in Mrs. R's classes, and the pattern was similar in Mr. W's classes.

In both studies, performance was not mediated by running any trial where the plant died, only the suggested trial, which was followed by questions to structure students' thinking. This result is consistent with prior work showing that without structure, students may uncover the necessary information but miss key patterns (Reiser, 2004, Vitale et al., 2016).

Study 1 provides some evidence that exploring the model helped students interpret flat lines in cumulative graphs, as Model First groups performed better than Graph First groups on those questions. Science knowledge may have helped constrain students' interpretations of those graphs. However, there were no condition differences when considering all of the graph interpretation questions together; exploring the model may help only with particular graph features.

This research only varied the order of activities. All students were prompted to think about cumulative graphs before the experimental steps (Figure 1). The treatment elicited students' ideas about relevant graphs by suggesting interpretations that connected graph shapes to their meanings. This treatment may have prepared students to distinguish among alternative interpretations while exploring the model.

Further, this study captures difficulties that students have with cumulative graphs in science. In the graph prediction step, many students chose graph shapes that did not match their text predictions, and in the graph interpretation step, many students agreed with interpretations that did not follow from the given graph. Of the $71 \%$ of Mrs. R's groups who said that plants use glucose all the time, only $55 \%$ of them selected a graph shape that matched this prediction. The most common incorrect answer suggested that students were interpreting the graph as instantaneous instead of cumulative. Mr. W's students also seemed to demonstrate this error: the most common graph selected for glucose made decreased when the light was off. This would have been a more reasonable choice if the graphs were instantaneous instead of cumulative, as it would have been the only shape consistent with the plant making less glucose in the dark. While these examples show difficulties going from meaning to graph, students also had difficulty with the reverse in the graph interpretation
steps. Students in Study 1 were near chance for those steps, and almost half of the groups in Study 2 did not recognize that no glucose is made when that line is flat. Graph misconceptions documented in prior work have focused on instantaneous graphs. Our results show that the type of misinterpretations noted in prior work also affect cumulative graphs (specifically, interpreting the graph as if the axes were different, i.e., instantaneous vs. cumulative). Though students exhibited difficulties, Study 2 shows pre- to post-test improvement on graph interpretation and on explanations.

Conclusion. When learning science from computer models with complex graph output, students may benefit from interpreting graphs before interacting with the model. This study demonstrates the power of eliciting student ideas, a component of the knowledge integration framework that is intended to prepare students to distinguish among alternatives in the experiment (Linn et al., 2006). While graphs are currently under-utilized in middle school science, this study shows how graphs can be leveraged to help students connect graph and science ideas.

Acknolwedgements. This material is based upon work supported by the National Science Foundation under Grant No. DRL-1418423 and INT-1451604.

## References

Ainsworth, S. (2006). Deft: A conceptual framework for considering learning with multiple representations. Learning and instruction, 16(3), 183-198.
Chiu, J. L., Chen, J. K., \& Linn, M. C. (2013). Overcoming deceptive clarity by encouraging metacognition in the webbased inquiry science environment. In International handbook of metacognition and learning technologies (pp. 517-531). Springer.
Clement, J. (1985). Misconceptions in graphing. In Proceedings of the Ninth International Conference for the Psychology of Mathematics Education (Vol. 1, pp. 369-375).
Grant, S. (2013). Graph Interpretation: How a Teacher Learns Pedagogical Content Knowledge. Presented at the 2013 Annual Meeting of the American Educational Research Association. San Francisco, CA.
Janvier, C. (1981). Use of situations in mathematics education. Educational Studies in Mathematics, 12(1), 113-122.
Lai, K., Cabrera, J., Vitale, J. M., Madhok, J., Tinker, R., \& Linn, M. C. (2016). Measuring graph comprehension, critique, and construction in science. Journal of Science Education and Technology, 25(4), 665-681.
Linn, M. C., Lee, H.-S., Tinker, R., Husic, F., \& Chiu, J. L. (2006). Teaching and assessing knowledge integration in science. Science, 313(5790), 1049-1050.
McDermott, L., Rosenquist, M., Popp, B., \& van Zee, E. (1983). Student difficulties in connecting graphs, concepts and physical phenomena. Presented at the 1983 Annual Meeting of the American Educational Research Association. Montreal, Quebec.
NGSS Lead States. (2013). Next Generation Science Standards: For States, by States (Matter and Energy in Organisms and Ecosystems). http://www.nextgenscience.org/topic-arrangement/msmatter-and-energy-organisms-and-ecosystems.
Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. The Journal of the Learning Sciences, 13(3), 273-304.
Vitale, J. M., Madhok, J., \& Linn, M. C. (2016). Designing a data-centered approach to inquiry practices with virtual models of density. In Proceedings of the International Conference of the Learning Sciences (pp. 591-598).

# Counterfactual Conditionals and Normative Rules 

Meredith R Wilkinson (mwilkinson@dmu.ac.uk)<br>School of Allied Health Sciences, De Montfort University, The Gateway, Leicester, UK, LE1 9BH<br>Shira Elqayam (selqayam@dmu.ac.uk)<br>Division of Psychology, De Montfort University, The Gateway, Leicester, UK, LE1 9BH<br>Valerie Thompson (valerie.thompson@usask.ca)<br>Department of Psychology, University of Saskatchewan, 9 Campus Drive, Saskatoon, Canada, S7N 5A5<br>David E Over (david.over@durham.ac.uk)<br>Department of Psychology, Durham University Science Site, South Road, Durham, UK, DH1 3LE


#### Abstract

Counterfactual thinking is the consideration of how things could have turned out differently, usually taking the form of counterfactual conditionals. This experiment examined the psychological mechanisms that transform counterfactuals into deontic guidance rules for the future. We examined how counterfactual thinking translates into deontic guidance rules by asking participants to infer these deontic conclusions from the counterfactual premises. Participants were presented with a vignette and a counterfactual conditional, and assigned to either a control condition or a suppression condition in which they were additionally presented with conflicting normative rules. The presence of conflicting norms reduced the likelihood of positive deontic conclusions being endorsed and increased the likelihood of negative deontic conclusions being endorsed. Future intentionality and regret intensity ratings were reduced in the suppression condition. The same conditions that affect normative inference also affect regret and future planning, suggesting similar cognitive mechanisms underlie these processes.


Keywords: conflicting norms; counterfactual thinking; deontic introduction; new paradigm; regret

## Introduction

All of us have to make many deontic judgments about what we 'should' and 'ought' to do. As Elqayam, Thompson, Wilkinson, Evans and Over (2015) note, when we learn about poverty in Somalia, we naturally infer that we ought to donate to famine relief. Such inferences are made on a daily basis, often about such everyday matters as the type of coffee we 'must' or 'should' buy, or the types of food we 'must not' or 'should not' eat. People readily infer these deontic statements from premises that contain no deontic content. However, very little is known about the psychological processes underlying the inferences.

Recently reasoning research has undergone a shift towards the 'new paradigm' of reasoning (Elqayam \& Over, 2013 and see other contributions in this special issue; Evans, 2012). The 'new paradigm' of reasoning rejects binary logic, regards reasoning as strongly related to judgement and decision making, focuses on probabilities and emphasizes pragmatic factors. Such an approach
demonstrates how reasoning can be applied to our everyday judgement and decision making. If this is the case, then it is persuasive that reasoning, judgement and decision making may adopt similar psychological processes.

This paper expands upon our previous work within this area (Elqayam et al., 2015), where we examined how we make such deontic judgements from non-deontic content, a process termed deontic introduction. The next section presents an overview of our previous research on this area with the use of conflicting norms. Our aim for this paper is to see if we can extend Elqayam et al.'s (2015) findings with counterfactual conditionals and this is why the section after examines previous work on counterfactual thinking to set the scene for our own research. Within this section we refer to literature on the functional basis of counterfactual thinking (e.g., Epstude \& Roese, 2008) which provides a rationale for us extending our work on deontic introduction into this domain. The final section before the method section provides an overview of the hypotheses for our current experiment.

## Previous Work on Deontic Introduction

Elqayam et al. (2015) examined the process of deontic introduction which entails making deontic inferences from content that contains no deontic material and which is pragmatic or informal type of inference (e.g., Hahn \& Oaksford, 2007). Deontic introduction is an inference which is socially contextualised, based on previous beliefs and desires and is probabilistic and defeasible (e.g., Oaksford \& Chater, 2007). By defeasible we mean an inference can be withdrawn or suppressed in light of additional information (Elio \& Pelletier, 1997).

Elqayam et al. (2015) propose that deontic introduction depends on a chain of inferences which are largely implicit. We begin with a conditional that bears utility (utility conditional) (e.g., If you pull the dog's tail, then it will bite you). When presented with this statement we make a goal valence inference in that a valenced outcome (positive negative or neutral) is implicitly inferred from the description of the outcome (e.g., being bitten is bad). We
then make a causal inference which infers a causal link between the action and outcome (e.g., pulling the dog's tail makes the dog bite). Given we now have the information that being bitten is bad and that pulling to dog's tail makes the dog bite psychological value passes via the causal link from the inferred goal to the action (e.g., pulling the dog's tail is costly) this is known as valence transference. This then feeds into deontic bridging: the valence can bridge into a novel norm expressed by a deontic operator (e.g., you should not pull the dog's tail).

Elqayam et al. (2015) conducted a series of experiments to test various features of the model. They adopted the same core design throughout their studies: they presented a vignette describing a protagonist and a situation with a utility conditional presented underneath. Then participants had to rate the degree to which it follows that the protagonist must, should and may (positive deontic operator) and must not, should not and need not (negative deontic operator) and going to and not going to (inductive conclusion) take the action to bring around the outcome. This experiment examines Elqayam et al.'s defeasibility hypothesis, and so the inference suppression paradigm was adopted (e.g., Stevenson \& Over, 1995) in which an additional premise is presented to weaken the strength of the initial argument. Our particular focus is on conflicting norms. Say we are informed that going to a particular holiday region will lead us to having a good holiday. Since this is something we wish to do it is likely that we would choose to go to that location. If we are later informed that the particular holiday region has hunts for endangered species, and we are against such hunting, this causes a conflict between the norm generated by deontic introduction and the pre-existing norm presented separately. This blocks deontic bridging by priming a normative conclusion that is opposite to the one generated by deontic bridging. The result is that participants are less likely to infer positive deontic operators when a conflicting norm is presented compared to when it is not, and more likely to endorse negative deontic operators when a conflicting norm is present compared to when it is not, as found by Elqayam et al. (2015). We wish to examine whether this effect occurs for counterfactual conditionals.

## Previous Work on Counterfactual Thinking

Counterfactual thinking can be defined in more than one way, but in the context of this paper, we take it to be considering how things could have turned out differently, for better or worse (see Byrne, 2016 for a review). A student who does not study hard for an exam and then subsequently fails that exam may imagine a world in which they did study and they received a good mark. This process may in turn then lead them to the decision that in future they will study for exams. Whilst such a process of counterfactual thinking elicits emotions such as regret (e.g., Wilkinson, Ball, \& Alford, 2015; Zeelenberg, 1999) it is said to have functional properties in that it can help people
prepare for similar situations in the future (e.g., Epstude \& Roese, 2008).

Research on counterfactual thinking has often employed vignette-based paradigms, where participants are presented with a scenario about a protagonist and are required to make a judgement. This research has yielded some consistent results. One is the temporal order effect, in which actions are regretted more in the short term, and inactions more in the long term (e.g., Gilovich \& Medvec, 1994). Much related research has been done on counterfactual conditionals. These conditionals, of the form if $p$ had been $q$ would have been, presuppose that $p$ and $q$ are false. In the example of us going on holiday to a particular region, it could be asserted, "If I had gone on holiday to that particular region, then I would have enjoyed my holiday better".

When we are faced with a conflicting norm and the outcome turns out negatively we could argue that the outcome was out of our control since taking the action would mean that we behave in a manner which conflicts with our norms in life. In this instance participants may report that the protagonist may feel less regret when a conflicting norm is present relative to when it is not. Previous work has demonstrated that people are more likely to mutate controllable rather than uncontrollable features in the wake of a negative event (e.g., Girotto, Legrenzi \& Rizzo, 1991). If people are more likely to mutate controllable aspects of events, then it may follow that when an outcome is in a protagonist's control participants predict they will feel greater regret in the wake of a negative outcome relative to when it is not as much in their control.

Counterfactual thinking arises from comparing what actually happened to what might have happened. Such comparison, made 'upwards' (it could have been better), or 'downwards' (it could have been worse), help us plan for the future (e.g., Roese, 1994; Epstude \& Roese, 2008). We propose that these future plans are mediated by normative rule generation. For example, faced with a disappointing exam result, a student might think she could have done better had she not been hung over. This in turn leads to the creation of a normative rule, 'I should not drink on an exam's eve'. In the psychology of reasoning, this process is called 'deontic introduction' (Elqayam et al., 2015). Our goal is to study the psychological mechanisms governing this transition from counterfactual thinking to deontic rule, and its effect on future planning

## Hypotheses

This experiment aimed to extend that of Elqayam et al. (2015) by adopting counterfactual conditionals, rather than indicative conditionals, to examine the effect of conflicting norms. Indicative conditionals are of the form 'If $p$ than $q$ ', linking an antecedent $p$ to a consequent $q$. Please see below an example of a set of stimuli for one of the scenarios that participants had to reason about. The predictions were presented with reference to the stimuli.

Martin has a new girlfriend, Gabrielle. He is keen to impress her by cooking a meal and is at the supermarket looking at different oils since he is making an Italian dish. Martin can buy a special olive oil produced in Fontignani. He opts for the cheaper oil and goes home. After he has cooked the meal and serves up he finds that the pasta is a bit greasy. He says to his girlfriend:

If I had opted for the Fontignani olive oil, then our pasta dish would have tasted better. (control condition)

However, the Fontignani olive oil is produced using intensive farming practices. If Martin uses the Fontignani oil, then he will be contributing to environmental degradation of the area (additional information for suppression condition)

There are a number of predictions we have: (1) that conflicting norms will suppress deontic introduction for counterfactual conditionals in the same manner that they do for indicative conditionals. With a conflicting norm present, conclusions with positive deontic operators will be less likely to be rated and conclusions with negative deontic operators more likely to be rated relative to no conflicting norm being present. In the example above, (2) Martin will be viewed as less inclined to use Fontignani oil if there is a conflicting norm present, e.g. a wish to avoid environmental degradation. When participants are asked how likely the protagonist is to take the action in the future, they will rate it as less likely than when a conflicting norm is absent, and (3) the participants will predict the protagonist will feel less regret intensity in the conflicting norm condition relative to the control condition. The reason for this is that they would have needed to take an action that conflicted with a norm (e.g., contribute to environmental degradation) to bring around the desirable outcome.

Both the first and second hypotheses derive out of the work of Elqayam et al. who found that conflicting norms defeated deontic introduction and the third hypothesis relating to regret emerges from the previous literature on counterfactual thinking which indicates that participants are more likely to mutate controllable relative to uncontrollable aspects that led to an unfortunate outcome (e.g., Girotto et al., 1991). It therefore follows that if participants are more likely to mutate controllable aspects then these controllable aspects may in turn lead to the inference that the protagonist will feel greater regret when the situation was in their control and no conflicting norm was present then when the conflicting norm may have taken the outcome somewhat out of their control.

## Method

## Participants

Seventy-eight participants completed the experiment and were recruited via Crowdflower a crowd sourcing platform enabling members of the public to participate in research for a small financial reimbursement. There were 40 females
and 37 males with 1 participant not disclosing gender. Participants age range was 21-75 years. Twenty participants stated Canada was their country of residence, 27 UK, 29 USA and 1 Australia. One participant did not disclose a country of residence. If participants reported a diagnosis of dyslexia, if English was not their first language or they failed to answer the attention checking question correctly their data were excluded from analysis. This left us with 62 participants.

## Experimental Design, Materials and Procedure

A mixed design was adopted. Participants were either randomly assigned to the control condition, in which they were just presented with the vignette and conditional statement or the suppression condition, in which participants were additionally presented with a conflicting norm. Participants had to then complete three tasks which are explained below: (1) a deontic rating task, (2) an intentionality question and (3) a regret intensity question. The independent variables were whether participants were in the condition to which participants were assigned (control or suppression) and the deontic operators (must, should, may and must not, should not and need not). The dependent variables were conclusion rating of the deontic operators from $1=$ definitely does not follow to $7=$ definitely follows, future intentionality rating from $1=$ not at all likely to $7=$ highly likely and regret intensity rating from $1=$ low regret to $7=$ high regret. Each task was presented on a separate page. There was a practice item at the top of each page to get participants used to each task. Participants reasoned about five vignettes. Materials were modified from Elqayam et al. (2015) Experiment 3.

The deontic rating task asked participants to rate the degree to which it followed that the protagonist must, must not, should, may and need not take the action in the vignette. Participants were required to state for each deontic operator whether they thought it definitely does not follow, follows very weakly, follows weakly, follows to some degree, follows strongly or follows very strongly. Participants completed a regret rating task. They had to rate the degree of regret they thought the protagonist would feel on a 7 -point scale from $1=$ low regret to $7=$ a high regret. The future intentionality task asked participants to state the degree to which they thought that the protagonist would be likely to take the action in the future. Again, participants rated this on a 7 -point scale from $1=$ not at all likely to $7=$ highly likely.

## Results

## Deontic Introduction

As can be seen in Table 1, all positive deontic operators receive higher ratings in the control condition relative to the suppression condition and all negative operators receive higher ratings in the suppression condition relative to the control condition. A 2 (condition: suppression or control) x 6 (operator: must, must-not, should, should-not, may and need-not) ANOVA was conducted and found an operator $x$
condition interaction $F(5,255)=12.63, M S E=2.65, p<$ .01, $\eta_{\mathrm{p}}{ }^{2}=.20$. A significant main effect of operator was observed using a Greenhouse-Geisser correction $F(2.73$, $139.25)=45.50, M S E=2.65, p<.05, \eta_{\mathrm{p}}^{2}=.47$ but no main effect of condition $F(1,51)=1.04, M S E=2.56 p=.31$, $\eta_{\mathrm{p}}^{2}=.02$.

Table 1: Mean (and standard deviation) ratings for each deontic operator as a function of condition

| Operator | Control | Suppression |
| :--- | :--- | :--- |
| Must | $4.31(1.62)$ | $2.96(1.15)$ |
| Should | $5.55(0.92)$ | $3.80(0.93)$ |
| May | $5.37(1.21)$ | $4.93(1.19)$ |
| Must not | $1.96(1.28)$ | $2.76(1.38)$ |
| Should not | $2.06(1.35)$ | $3.09(1.24)$ |
| Need not | $2.93(1.24)$ | $3.54(1.62)$ |

In order to unpack the operator x condition interaction, we conducted six independent samples t-tests for each operator separately. This was found to be significant for the operators must, should and should not with a $p<.008$ with the adoption of a Bonferroni correction but not for the must not $p=.03$, may $p=.19$ and need-not $p=.13$. These fit with our first hypothesis that conflicting norms will lead to lower ratings for positive deontic operators compared to the control condition with the reverse the case for negative deontic operators.

## Future Intentionality and Regret Ratings

We then examined the future intentionality ratings comparing the control condition to the suppression condition. We compared mean likelihood ratings across the scenarios for the control and suppression conditions and found future intention ratings were higher in the control condition $(M=5.95, \mathrm{SD}=0.49)$ compared to the suppression condition $(M=4.48, S D=0.96)$ a finding which reached significance when conducting an independent samples t-test $t(51)=6.80 p<.01$. This supports our hypothesis that future intentionality will be weakened in the suppression condition.

Our final analysis considered the reported regret intensity that participants thought the protagonist would feel. It was found that participants thought the protagonists in the control condition would experience greater regret intensity $(\mathrm{M}=5.61, \mathrm{SD}=0.66)$ comparative to the suppression condition $(\mathrm{M}=4.44, \mathrm{SD}=1.09)$ a finding which was significant when undertaking an independent samples t-test $t(51)=4.58, p<.01$. This supports our hypothesis that the level of regret intensity the participant thinks the protagonist will feel is less in the suppression condition compared to the control condition.

## General Discussion

The aim of this experiment was to examine the process of deontic introduction for counterfactual conditionals rather than indicative conditionals. We examined the defeasibility
hypothesis adopting conflicting norms to block the deontic bridging stage of deontic introduction. We proposed three hypotheses at the start of our paper (1) that conflicting norms will suppress deontic introduction in the context of counterfactual conditionals, as they do with indicative conditionals, (2) when a conflicting norm is present participants will rate the protagonist's intention to take the action in the future as lower than when no conflicting norm is present and (3) when a conflicting norm is present regret intensity for the outcome will be rated as lower compared to when no conflicting norm is present.

Support was found for the first hypothesis with positive deontic operators rated as lower in the conflicting norms condition relative to the control condition with the reverse pattern occurring for negative deontic operators. This finding supports the defeasibility hypothesis of Elqayam et al. (2015) and extends to findings of Elqayam et al. to counterfactual conditionals. We propose the same explanation for these findings that Elqayam et al. offer in their paper. Deontic bridging is not able to occur due to a conflict between the pre-existing norm and the invited normative conclusion (generated by deontic introduction).

When it came to our item analysis for each of the operators we observed significant effects for must, should and should not. Taking into consideration the marginal significant effect of must not we note that the significant differences lie in those operators that express obligations and forbidding but not permissions. This finding suggests that perhaps the role of counterfactual thinking is to direct future action, making it functional (e.g., Epstude \& Roese, 2008). In this respect, obligations and forbidding are more powerful than permissions, and this could provide an explanation for the pattern of results we observed.

Our second hypothesis was that, for future intentionality ratings, participants would predict the protagonist would be less likely to take the action when a conflicting norm was present compared to when it was not. This is what we found. We propose that this result occurs because a conflicting norm prevents deontic bridging. Such a finding supports our prediction that deontic introduction can be used to direct future actions as a result of the presence of counterfactuals. This is in line with the notion of counterfactual thinking is functional (e.g., Epstude \& Roese, 2008).

Our third hypothesis was participants would report the protagonist feeling less regret when a conflicting norm was present compared to when it was not. This hypothesis was supported. We propose that this finding may occur for one of two reasons. The first is that the conflicting norm serves to distance the protagonist from the regretted incident by providing a justification or rationale for them not taking the action. In the case of Martin and the Fontignani olive oil, that justification would be the conflict between having the better meal and the fact that he does not want to contribute to environmental degradation. If this process is occurring, the decision could be seen as self-enhancing allowing one to distance oneself from the regretted outcome (e.g., Feeney,

Gardiner, Johnston, Jones \& McEvoy, 2005). A second reason this effect may occur could be linked to the controllability of the outcome. If we take away the conflicting norm in the case of Martin and the Fontignani olive oil the outcome is entirely within Martin's hands: he did not select the correct olive oil for the dish resulting in the dish not being as nice. However, when we add a conflicting norm, that outcome becomes less controllable, since he does not want to behave in a manner that conflicts with his normative framework. We propose that perhaps less regret is predicted in the suppression condition because participants view the outcome as less in control of the protagonist. Girotto et al. (1991) found that participants prefer to mutate controllable relative to uncontrollable events that lead to a negative outcome. We propose that, when a conflicting norm is presented, this makes taking the action to bring about the desired outcome less "controllable" in a normative sense: it becomes less permissible or even forbidden. Since people are more likely to mutate controllable than uncontrollable events that led to a negative outcome, it seems intuitive that greater regret intensity will be predicted for the control condition, where the outcome is within the protagonist's control, than in the suppression condition, in which the conflicting norm serves to block the action, making it uncontrollable in the normative sense. Both hypotheses are possible but the controllability one may be stronger since when it comes to distancing oneself from the outcome controllability may act as a moderator. A future avenue of research could use a controllability manipulation (controllable versus uncontrollable outcome) to examine what effect this direct manipulation has on deontic introduction.

A final suggestion for the result lies in the fact that participants have to make a comparison when presented with a conflicting norm. For Martin it is the choice between using the other oil and the meal not being as tasty to using the Fontignani oil and contributing to environmental degradation. It is possible that in these cases preference construction occurs on the spot.

The fact that conflicting norms demonstrate such consistent results when also accounting for Elqayam et al.'s (2015) findings strongly indicates that people are unwilling to go against their normative framework. Although we did not test it directly in our study one proposal is that whilst people will generally not go against their normative framework for small instances (e.g., having a nice meal) they may do so when the outcome generates sufficient benefit. For example, we may be told as children that it is wrong to lie, and we must tell the truth, and we may hold that norm. However, if we are placed in a situation in which lying could garner a benefit and especially a moral benefit (e.g., saving a life) then it may be the case that we act against our normative framework in this instance.

These findings have extended those of Elqayam et al. (2015) through demonstrating that their proposed model for deontic introduction can be applied to counterfactual conditionals. This is an important theoretical development
since it indicates that similar cognitive processes are at work when reasoning about counterfactual to indicative conditionals. We propose that an avenue for future research could be to test different components of the model of Elqayam et al. to see whether they are applicable for counterfactual counterfactuals in the same manner as they are for indicative conditionals. Elqayam et al.'s work has demonstrated that factors such as utility and probability, which are deeply rooted in the new paradigm, have an impact on deontic introduction for indicative conditionals. We propose that such effects may also occur when counterfactual conditionals are used.

Research on deontic introduction has begun by adopting a vignette-based paradigm like many areas of reasoning research. One challenge of that paradigm though is seeing the degree to which the model can apply to everyday life. We propose an interesting extension would entail asking participants to recall an instance of real life regret, to consider a counterfactual conditional, and then to complete the deontic rating task. Through adopting this approach, we hope to learn how deontic introduction can be applied to real life regrets, and whether the same experimental manipulations, such as conflicting norms, demonstrate the same suppression effects as they do in a vignette-based paradigm, where the participant is reasoning about an unknown protagonist.

From a methodological perspective we believe it would be interesting to examine the cognitive processes participants adopt directly via the adoption of think-aloud protocols. This is a process-tracing technique that requires participants to think aloud whilst working through a problem in order for the researcher to gain insight into participants’ thought processes (see Ericsson \& Simon, 1993). Wilkinson, Ball and Cooper (2010) have utilised think aloud protocols using counterfactual vignettes about mental states to good effect. Stenning and van Lambalgen (2008) show that experimental data can be enriched by the use of think aloud protocols, revealing how participants understand the task, and the trajectory of their reasoning processes. By adopting think aloud protocols whilst asking participants to complete the deontic rating task, future intentionality and regret questions could provide insights into their cognitive processes and potentially add further weight to the model of Elqayam et al. (2015). It would enable participants to state how they deal with the presence of a conflicting norm within their reasoning. This would enable the test of some of the predictions for the findings made within this section.

This experiment has extended one of the findings of Elqayam et al. (2015), demonstrating that deontic introduction in the context of conflicting norms is not only affected by indicative conditionals but also by counterfactual conditionals. This is an important finding for the new paradigm of reasoning (e.g., Evans, 2012; Manktelow, Over, \& Elqayam, 2011; Over, 2009) with subjective degrees of belief at its heart and social pragmatics and subjective psychological value having a significant role
to play (Elqayam \& Over, 2013). We believe that our research adds to this field by showing how counterfactual conditionals can give rise to new deontic norms. It supports the conclusion that, whilst our counterfactual thinking may cause us pain, it is truly functional (e.g., Epstude \& Roese, 2008). As Elqayam et al. (2015) noted, humans are quite ready to infer an 'ought' from an 'is' (see also Hume, 2000/1739-1740). Our findings indicate that humans are also often keen to infer an 'ought' from a 'would have been'.

## Acknowledgments

We acknowledge the Leverhulme trust for funding to the second, third and fourth authors for a grant (RPG-2012573).

## References

Byrne, R. M. J. (2016). Counterfactual thought. Annual Review of Psychology, 67, 135-157.
Elio, R., \& Pelletier, F. J. (1997). Belief change as propositional update. Cognitive Science, 21(4), 419-460.
Elqayam, S., Thompson, V. A., Wilkinson, M. R., Evans, J. St. B. T. \& Over, D. E., (2015). Deontic introduction: A theory of inference from is to ought. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5), 1516-1532.
Elqayam, S., \& Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon and Over. Thinking \& Reasoning, 19(3-4), 249-265.
Epstude, K., \& Roese, N. J. (2008). The functional theory of counterfactual thinking. Personality and Social Psychology Review, 12(2), 168-192.
Ericsson, K. A. \& Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Revised Edition. Cambridge, MA: MIT Press.
Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. Thinking \& Reasoning, 18(1), 5-31.
Feeney, A., Gardiner, D. R., Johnston, K., Jones, E., \& McEvoy, R. J. (2005). Is regret for inaction relatively self-enhancing? Applied Cognitive Psychology, 19(6), 761-777.
Gilovich, T. \& Medvec, V. H. (1994). The temporal pattern to the experience of regret. Journal of Personality and Social Psychology, 67(3), 357-365.
Girotto, V., Legrenzi, P., \& Rizzo, A. (1991). Event controllability in counterfactual thinking. Acta Psychologica, 78(1-3), 111-133.
Hahn, U. \& Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. Psychological Review, 114(3) 704732.

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

Manktelow, K. I., Over, D. E., \& Elqayam, S. (2011). Paradigm shift: Jonathan Evans and the science of reason. In K. I. Manktelow, D. E. Over and S. Elqayam (Eds.), The science of reason: A festschrift in honour of Jonathan St. B. T. Evans (pp. 1-16). Hove, UK: Psychology Press.
Oaksford, M. \& Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. Oxford, UK: Oxford University Press.
Over, D. E. (2009). New paradigm psychology of reasoning. Thinking \& Reasoning, 15(4), 431-438.
Roese, N. J. (1994). The functional basis of counterfactual thinking. Journal of Personality and Social Psychology, 66(5), 805-818.
Stenning, K. \& van Lambalgen, M. (2008). Interpretation, representation, and deductive reasoning. In J. E. Adler and L. J. Rips (Eds.) Reasoning Studies of Human Inference and Its Foundations (pp. 223-248). New York, US: Cambridge University Press.
Stevenson, R. J. \& Over, D. E. (1995). Deduction from uncertain premises. The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 48(3), 613-643.
Wilkinson, M. R., Ball, L. J., \& Alford, D. (2015). Counterfactual reasoning for regretted situations involving controllable versus uncontrollable events: The modulating role of contingent self-esteem. Advances in Cognitive Psychology, 11(1), 22-30.
Wilkinson, M. R., Ball, L. J., \& Cooper, R. (2011). Arbitrating between theory-theory and simulation theory: Evidence from a think-aloud study of counterfactual reasoning. In S. Ohlsson and R. Catrambone Proceedings of the 32nd Annual Meeting of the Cognitive Science Society. (pp. 1008-1013). Austin, TX: Cognitive Science Society.
Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. Journal of Behavioral Decision Making, 12(2), 93-106.

# Modality Differences in Timing: Testing the Pacemaker Speed Explanation 

Emily A. Williams (Emily.A.Williams@Manchester.ac.uk)<br>Andrew J. Stewart (Andrew.J.Stewart @Manchester.ac.uk)<br>Luke A. Jones (Luke.Jones@Manchester.ac.uk)<br>Division of Neuroscience and Experimental Psychology, The University of Manchester, UK


#### Abstract

A classic effect in the timing field is that "sounds are judged longer than lights" (Goldstone, Boardman \& Lhamon, 1959). Recently, judgements for tactile durations have been found to fall between the two (Jones, Poliakoff \& Wells, 2009). These modality differences are commonly interpreted within scalar timing theory as the work of a central pacemaker which runs faster for sounds, then vibrations, and slowest for lights (Wearden, Edwards, Fakhri \& Percival, 1998). We investigated whether verbal estimates and temporal difference thresholds are correlated within each modality, but found this not to be the case. This suggests that differences in pacemaker speed may not be the main driver for modality differences in thresholds. In addition, we investigated sensory bias as an alternative to the pacemaker explanation, but this was found not to correlate with modality differences in timing.


Keywords: Time perception; interval timing; sensory modalities; pacemaker-accumulator; sensory bias.

## Introduction

The timing of stimulus duration by humans has historically been under-researched compared to other perceptual domains. One reason is that although humans possess a very sensitive discrimination for duration (with difference thresholds as low as 10 ms ), there is no sensory organ for time. This forces explanations to draw on hidden processes more heavily than for other sensory systems, such as vision and hearing. To date the most successful models of human timing have centred on the idea that humans possess an internal clock of a pacemaker-accumulator type, such as in scalar timing theory (SET: Gibbon, 1977; Gibbon, Church \& Meck, 1984). The pacemaker generates internal events ('pulses' or 'ticks') which are connected to an accumulator via a switch. The accumulator contents increase linearly with the duration being estimated and forms the basis for timing judgments (further memory and decision modules are also typically added to this clock model - see Gibbon et al., 1984).

Support for the idea of a pacemaker-accumulator internal clock comes from several sources. People (and animals) can stop and start timing like a stopwatch, even managing to 'pause' timing and continue after a short gap (Buhusi \& Meck, 2009). In addition, people can perform ordinality judgements, express one duration as a proportion of another, and average durations together, again suggesting a linear relationship between perceived time and real time (Wearden \& Jones, 2007).

Furthermore, it appears the speed of the internal clock can be altered. A key signature of a change in pacemaker speed
is the 'slope effect'. In verbal estimation ${ }^{1}$ tasks, when the stimulus durations are plotted against estimates of those durations, the difference between the experimental condition and control manifests itself as a difference in slope. This is consistent with a multiplicative increase in pacemaker speed, rather than a simple bias (which manifests as a difference in intercept). Such slope effects have been found for certain drugs (Meck, 1984), body temperature changes (for review see Wearden \& Penton-Voak, 1995), repetitive stimulation (Penton-Voak, Edwards, Percival \& Wearden, 1996), and filled versus unfilled durations (Wearden, Norton, Martin \& Montford-Bebb, 2007).

It is known that durations of sounds are judged longer than lights (Goldstone et al., 1959). This effect manifests as a difference in slope (Wearden et al., 1998), where the auditory slope is steeper than the visual slope. It has been argued that the pacemaker runs at a faster speed for auditory than visual stimuli (Wearden et al., 1998; Penney, Gibbon \& Meck, 2000). This has sparked a debate about whether there is a central pacemaker that runs at different speeds for different modalities, or separate pacemakers for each modality (See Grondin, 2010). These auditory-visual differences have been found to occur on a range of timing tasks, from temporal generalization (Wearden et al., 1998), to temporal bisection (Penney et al., 2000), to temporal difference thresholds (Jones et al., 2009), suggesting the effect is not task-dependent.

Recently, the temporal judgement of tactile stimuli has been investigated. Jones et al. (2009) found that verbal estimation slopes and temporal difference thresholds for tactile stimuli fall between those for auditory and visual stimuli. Additionally, the two tasks share an inverted pattern, where estimation slopes are highest (most accurate) and thresholds are lowest (most sensitive) for auditory stimuli, for example. It has been suggested that a faster pacemaker is a more accurate pacemaker (Troche \& Rammsayer, 2011), which appears to be the case, but this assertion has yet to be empirically investigated. Therefore, the present series of studies will begin with a replication of Jones et al. (2009), but analysis will also examine whether estimation slopes and difference thresholds correlate with each other for auditory, visual, and tactile stimuli.

In contrast to the pacemaker speed explanation, it has been suggested that modality differences could be due to

[^542]intrinsic differences between the different sensory systems (Yuasa \& Yotsumoto, 2015), e.g. some combination of differences in transduction rates (Zampini, Shore \& Spence, 2003) and attentional biases (Spence, Shore \& Klein, 2001). Therefore, we will investigate whether these aspects of sensory bias or salience correlate with differences between auditory, visual and tactile stimuli. The present study will operationalise sensory bias as the 'point of subjective simultaneity' on a temporal order judgement task, i.e. the duration that one modality has to precede another by, in order for the two modalities to be judged as simultaneous.

In summary, the aim of the current work is to investigate whether verbal estimates and temporal difference thresholds correlate within each modality, and whether modality differences can be alternatively explained by a measure of sensory bias.

## Experiment 1a: Verbal Estimation

## Method

Participants 52 right-handed participants (staff and students of the University of Manchester and some members of the general population) completed all three tasks in a counterbalanced order and received $£ 10$ for their time.

Apparatus and Materials Participants were seated at a table in a dark room, with their chin on a chin rest. A PC presented the experiments, written in E-Prime (Psychology Software Tools, Pittsburgh, PA). A 17" Samsung Syncmaster monitor stood at a distance of 60 cm . Participants' eyes were level with the top of the monitor and the fixation cue and questions were displayed $20^{\circ}$ below eye level. A black foam grip ( $5.5 \times 9.5 \times 4.5 \mathrm{~cm}$ ) was secured to the table 30 cm in front of participants in the centreline. Behind the grip was a Sony speaker, which presented the auditory stimuli ( 500 Hz sine wave tones), and to the left of the grip was a numerical keypad ( $8.5 \times 12 \mathrm{~cm}$ ) for use with the left hand.

The grip housed an Oticon-A (100 Ohm) bone conductor with vibrating surface $1.6 \mathrm{~cm} \times 2.4 \mathrm{~cm}$. The bone conductor was inset into the foam in the index finger position when gripped with the right hand, and driven by a 500 Hz sine wave signal through a TactAmp 4.2 amplifier (Dancer Design). Visual stimuli were presented via a 6 mm green LED light ( $87 \mathrm{~cd} / \mathrm{m}^{2}$ ), embedded in a black plastic casing ( 4 x $4 \times 1.75 \mathrm{~cm}$ ) and attached on top of the foam block. The LED was $16^{\circ}$ below the fixation cue ( $36^{\circ}$ below eye level) and 32 cm in front of participants.

Participants wore 3 M Peltor ear protectors with inset earphones, which played white noise ( 56 dB ) during the tasks to mask the sound of the vibrations.

Procedure On each trial participants estimated the duration of a stimulus. The task contained 150 trials, where ten stimulus durations (77, 203, 348, 461, 582, 767, 834, 958, 1065 , and 1183 ms ) were presented in each modality
(auditory, visual, and tactile) five times. Trials were grouped into three counterbalanced blocks by modality.

Each trial began with the presentation of a fixation cross for $500-1000 \mathrm{~ms}$, which was followed by the stimulus. Participants were prompted on-screen to type in their estimate in milliseconds and were reminded that 1 second $=$ 1000 ms . The task lasted approximately 17 minutes.

## Results

Outliers were defined as estimation functions that were invariant to stimulus duration (identified as linear regressions not significantly different to 0 ), which suggested an inability to perform the task. This led to the exclusion of one individual, leaving a sample of 51 participants. See Figure 1 for the mean verbal estimates for each modality.

The hypothesis that verbal estimates would be highest for auditory stimuli and lowest for visual stimuli was examined using a factorial ANOVA with two repeated measures factors: modality (auditory, visual and tactile) and stimulus duration.


Figure 1: Mean verbal estimates for each modality against stimulus durations.

The ANOVA found a main effect of stimulus duration, $F_{(2.60,130.17)}=750.70, p<.001, \eta_{\mathrm{p}}^{2}=.938$. Post hoc analyses revealed that each of the 10 stimulus durations were estimated as significantly differently from each other ( $p$ < . 001 for all comparisons).
There was also a main effect of modality, $F_{(2,100)}=7.50, p$ $=.001, \mathrm{\eta}_{\mathrm{p}}^{2}=.131$. Post hoc analyses revealed that participants estimated auditory stimuli to be significantly longer than visual $(\mathrm{p}=.006)$ and tactile $(\mathrm{p}=.012)$ stimuli. However, estimates for visual and tactile stimuli did not significantly differ ( $\mathrm{p}=.909$ ) .

The interaction between stimulus duration and modality was also significant, $F_{(8.39,419.32)}=4.914, p<.001, \eta_{\mathrm{p}}{ }^{2}$ $=.089$. In order to investigate this interaction, linear regressions were conducted to extract the slope and intercept values of each participant's verbal estimation function for each modality. See Figure 2 for mean slope values.


Figure 2: Mean slope values for auditory, visual and tactile stimuli. Error bars denote standard error.

A repeated measures one-way ANOVA comparing the slopes across modalities found a significant difference between them, $F_{(2,100)}=12.76, p<.001, \mathrm{n}_{\mathrm{p}}^{2}=.203$. Post hoc analyses confirmed that auditory slopes were significantly higher than visual slopes ( $p<.001$ ), but not significantly different to tactile slopes ( $p=1.00$ ). In addition, the tactile slopes were significantly higher than visual slopes ( $p=$ .001).

## Discussion

Verbal estimation slopes for auditory stimuli were significantly and multiplicatively higher than those for visual stimuli, with tactile slopes falling between the two. As perfect estimates would have a slope of 1 , this suggests that people are more accurate when estimating durations of sounds and vibrations than lights, but tend to underestimate all three modalities. This is the same pattern of results found as in Jones et al. (2009).

Auditory and tactile estimates differed significantly in the first ANOVA, but further analysis on slopes (as pacemaker speed differences are said to manifest as slope effects) found the slopes not to differ. The significant difference between the two in the first ANOVA was perhaps due to a difference in intercept.

## Experiment 1b: Temporal Difference Thresholds

## Method

Participants The same participants completed this experiment as in Experiment 1a.

Apparatus and Materials The same apparatus and materials were used as in Experiment 1a.

Procedure Participants completed a 50-trial threshold task in each of the three modalities in a counterbalanced order. The test stimuli were the same as in Experiment 1a.

Each trial began with the presentation of a fixation cross for $500-1000 \mathrm{~ms}$, which was followed by the stimuli. The first stimulus (the standard) was always 700 ms , while the second stimulus (the comparator) began at 1000 ms in duration. A 500-1000 ms delay occurred between the two stimuli, and a $125-250 \mathrm{~ms}$ delay followed the second stimulus. The order of the standard and the comparator was counterbalanced between trials. Participants pressed ' 1 ' or ' 2 ' on the keypad depending on whether they thought the first or the second stimulus was longer.

This task used a weighted 3-up 1-down staircase method (Kaernbach, 1991), which allowed for the calculation of the difference in stimulus durations that participants can discriminate $75 \%$ of the time. The step size was 15 ms for the first 30 trials, then 10 ms for the last 20 trials. Thresholds were calculated as the mean difference between the standard and comparator durations across the last 20 trials. The task took approximately 18 minutes to complete.

## Results

Outliers were defined as thresholds greater than 600 ms (twice the starting difference) which suggested an inability to perform the task. However, no participant had thresholds above this value, giving a full sample of 52 participants.

Figure 3 shows the mean difference between the standard and comparator durations across the 50 trials for the three modalities. The resulting temporal difference thresholds can be seen in Figure 4.


Figure 3: Mean difference between the standard and comparator across the 50 trials for each modality. The vertical dashed line separates the last 20 trials over which the temporal difference thresholds were calculated.

The hypothesis that thresholds would differ according to the modality of the stimuli was examined using a one-way repeated measures ANOVA. This test found a significant difference between thresholds for the different modalities, $F_{(2,102)}=30.89, p<.001, \mathrm{\eta}_{\mathrm{p}}^{2}=.377$. Post hoc analyses
confirmed that thresholds for auditory stimuli and tactile stimuli were significantly lower than thresholds for visual stimuli ( $p<.001$ for each comparison). However, thresholds for auditory and tactile stimuli did not significantly differ ( $p$ $=.079$ ).


Figure 4: Mean temporal difference thresholds for each modality. Error bars denote standard error.

## Discussion

Thresholds for visual stimuli were significantly higher than both auditory and tactile stimuli, while auditory and tactile thresholds did not significantly differ. This suggests that people have greater sensitivity to the durations of sounds and vibrations than lights. This pattern of thresholds was reported previously by Jones et al. (2009).

## Research Question 1: Do Estimates and Thresholds Correlate within Each Modality?

## Results

The same outlier criteria were applied as in the previous sections, with the addition of values 2.5 SDs from the mean. Following removal, this left 51, 49 and 51 participants for auditory, visual and tactile correlations respectively.

Three Pearson's product-moment correlation coefficients found no correlations between estimation slopes and thresholds within each modality (See Table 1).

Table 1: Correlations between verbal estimation slopes and temporal difference thresholds within each modality.

| Modality | $d f$ | $r$ | $p$ |
| :--- | :---: | :---: | :---: |
| Auditory | 49 | -.099 | .490 |
| Visual | 47 | -.123 | .398 |
| Tactile | 49 | -.130 | .362 |

## Discussion

It had been argued that the differences between modalities in these two tasks were due to the pacemaker running at a faster rate for auditory stimuli and a slower rate for visual
stimuli (Wearden et al., 1998; Jones et al., 2009) and that a faster pacemaker leads to greater accuracy and sensitivity (Troche \& Rammsayer, 2011). However, accuracy in estimates (slopes) and sensitivity to duration (thresholds) did not correlate for within any modality. This poses a problem for applying the pacemaker explanation to both of these tasks. It could be argued that estimates (magnitude judgements) and thresholds (discrimination judgements) rely on different mechanisms and are of different levels of abstraction, but we expected small correlations despite the transformative nature of estimations.

## Experiment 2: Sensory Bias, measured by PSS

This experiment will calculate sensory bias, as measured by the point of subjective simultaneity (PSS) on a temporal order judgement task. The PSS measures the duration that one modality has to precede another by, in order for the two modalities to be judged as simultaneous. This can be seen as a measure of relative salience between the different senses and is affected by the intrinsic properties of each sensory system, e.g. transduction rates (Zampini et al., 2003) and attentional biases (Spence et al., 2001).

Previous research has found sensory biases (measured by PSS) in favour of auditory stimuli when compared with visual (Zampini et al., 2003) and tactile stimuli (Zampini, Shore \& Spence, 2005) and in favour of tactile stimuli when compared with visual stimuli (Spence et al., 2001). Therefore, PSSs appear to follow the same modality pattern as estimates and thresholds.

This measure of sensory bias will be investigated as an alternative explanation for the differences between auditory, visual and tactile performance on estimation and threshold tasks in the next section.

## Methods

Participants The same participants completed this experiment as in Experiment 1a and 1b.

Apparatus and Materials The same apparatus and materials were used as in Experiment 1a and 1b.

Procedure Participants were presented with two crossmodal stimuli ( 15 ms each) in quick succession and were asked which occurred first. The task contained 300 trials, where participants were presented with three modality pairs (Aud-Vis, Aud-Tac, Vis-Tac), at 10 different stimulus onset asynchronies (SOAs, -400, -200, -90, -55, -20, +20, +55, $+90,+200$, and +400 ms ), each repeated 10 times. Negative SOAs mean that the first-named stimulus in the pairing came first (e.g. auditory in the Aud-Vis stimulus pair), whereas positive SOAs indicate that the second-named modality came first. Trials were separated into three counterbalanced blocks by modality pair.

On each trial a fixation cross appeared on the screen after a 500 ms delay, where it remained for the rest of the trial. The first stimulus was presented following a random
duration between 500-1000 ms. After the randomly selected SOA, the second stimulus was presented. For example, if the Aud-Vis modality pair was presented with the -400 ms SOA, participants heard a 15 ms tone, followed by a delay of 385 ms , and then saw the green LED illuminate for 15 ms . After a $125-250 \mathrm{~ms}$ delay, participants were then prompted to answer "Which stimulus came first?" and participants pressed ' 1 ' or ' 2 ' on the keypad. The task took approximately 15 minutes to complete.

## Results

Cumulative Gaussian psychometric functions were fitted to participants' individual data, coded according to 'proportion auditory-first', for example. The PSS and just noticeable difference (JND) were extracted for each individual for each modality pair.

Participants' PSSs were inspected for outliers, identified as those with related JNDs greater than 400 ms (Zampini et al., 2003), which suggested an inability to complete the task. This resulted in the exclusion of eight individuals, leaving a sample of 44 participants. See Figure 5 for PSSs for each cross-modal comparison.


Figure 5: Mean Point of Subjective Simultaneity for auditory-visual, visual-tactile and tactile-auditory comparisons. Error bars denote standard error.

No significant sensory bias was found between auditory and visual stimuli, indicated by the PSS not departing from zero $\left(t_{(43)}=.94, p=.354\right)$. However, the PSS for visualtactile comparisons was significantly above zero, $t_{(43)}=$ $3.24, p=.002$. Participants were biased in favour of tactile stimuli in this comparison, and required visual stimuli to be presented 37 ms before tactile stimuli, for the pair to be judged as simultaneous. In addition, the PSS for auditorytactile comparisons was also significantly above zero $\left(t_{(43)}=\right.$ 3.21, $p=.003$ ). Participants were biased in favour of auditory stimuli in this comparison, and required tactile stimuli to be presented 22 ms before auditory stimuli for subjective simultaneity.

## Discussion

Significant sensory biases were found in favour of auditory stimuli when compared with tactile stimuli, and in favour of tactile stimuli when compared with visual stimuli, which concurs with previous research (Spence et al., 2001; Zampini et al., 2005). However, no significant sensory biases were found between auditory and visual stimuli. This was unexpected and is contrary to both previous research (Zampini et al., 2003), and our hypothesis that the large differences between auditory and visual estimates and thresholds may be due to large sensory biases.

Nevertheless, the next section will investigate whether these sensory biases correlate with the differences between modalities in estimates and thresholds.

## Research Question 2: Do cross-modal PSSs correlate with the differences between modalities in estimates and thresholds?

## Results

The same outlier criteria were applied as in the previous sections, leaving a sample 44 participants.

Six Pearson's product-moment correlation coefficients found no correlations between PSSs and estimation slopes or thresholds for any cross-modal pair (See Table 2).

Table 2: Correlations between cross-modal PSSs and slope and threshold differences for each cross-modal pair.

| Variable 1 | Variable 2 | $r$ | $p$ |
| :---: | :--- | :---: | :---: |
| Aud-Vis PSS | Aud - Vis Slope | .004 | 1.00 |
|  | Aud - Vis Threshold | .061 | 1.00 |
| Aud-Tac PSS | Aud - Tac Slope | -.217 | .314 |
|  | Aud - Tac Threshold | .288 | .116 |
| Tac-Vis PSS | Tac - Vis Slope | .323 | .066 |
|  | Tac - Vis Threshold | -.333 | .054 |

## Discussion

Cross-modal sensory biases, as measured by PSSs, were found not to correlate with the differences in estimates and thresholds between each modality pair. This suggests that the differences in auditory, visual and tactile estimates and thresholds cannot be explained by the intrinsic sensory biases of the three different systems.

## General Discussion

We aimed to investigate the pacemaker explanation for differences between auditory, visual and tactile estimates and thresholds, and discovered three main findings. Firstly, the pattern of differences between the modalities appears to be robust as we replicated these in Experiments 1a and 1b (with minor differences in magnitude). Secondly, estimates and thresholds do not correlate within each modality. This poses a problem for the idea that both the slopes in verbal
estimation and the order of thresholds are mostly determined by pacemaker rate. Finally, the modality differences in estimates and thresholds did not correlate with sensory biases, potentially ruling out this alternative explanation. These findings generate several possible conclusions:

1. Pacemaker rate does not determine estimation slopes or threshold values
2. Pacemaker rate determines estimation slopes but not threshold values (or vice-versa)
3. Pacemaker rate contributes to both slopes and thresholds, but this contribution is washed out by other cognitive processes
Despite these theoretical uncertainties, the present research can state that the assertion that faster pacemakers give rise to smaller thresholds (Troche \& Rammsayer, 2011) is flawed, if one assumes that the pacemaker underlies both estimation and threshold tasks.

At present, there is no published model of how the scalar timing theory system operates in threshold tasks, unlike for temporal generalization (Droit-Volet, Clément \& Wearden, 2001), temporal bisection (Wearden, 1991) and verbal estimation (Wearden, 2015). Additionally, the mathematical consequences of increasing pacemaker speed on timing performance (and the assumptions this is based on) in tasks are not explored or predicted in any great detail in the literature. Therefore, our future work will examine the role of the pacemaker in a model of threshold behavior and model mathematical implications of altering pacemaker rate.

Overall, the simple pacemaker speed explanation appears to fail and a more nuanced explanation is required.

## Acknowledgments

Emily Williams was supported by a studentship from the Economic and Social Research Council.

## References

Buhusi, C. V., \& Meck, W. H. (2009). Relativity theory and time perception: single or multiple clocks? PloS one, 4(7), e6268.
Droit-Volet, S., Clément, A., \& Wearden, J. (2001). Temporal generalization in 3- to 8-year-old children. Journal of Experimental Child Psychology, 80, 271-288.
Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. Psychological review, 84(3), 279325.

Gibbon, J., Church, R. M., \& Meck, W. H. (1984). Scalar timing in memory. Annals of the New York Academy of sciences, 423(1), 52-77.
Goldstone, S., Boardman, W. K., \& Lhamon, W. T. (1959). Intersensory comparisons of temporal judgments. Journal of Experimental Psychology, 57(4), 243-248.
Grondin, S. (2010). Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions. Attention, Perception, \& Psychophysics, 72(3), 561-582.

Jones, L. A., Poliakoff, E., \& Wells, J. (2009). Good vibrations: Human interval timing in the vibrotactile modality. The Quarterly Journal of Experimental Psychology, 62(11), 2171-2186.
Meck, W. H. (1986). Affinity for the dopamine D 2 receptor predicts neuroleptic potency in decreasing the speed of an internal clock. Pharmacology Biochemistry and Behavior, 25(6), 1185-1189.
Penney, T. B., Gibbon, J., \& Meck, W. H. (2000). Differential effects of auditory and visual signals on clock speed and temporal memory. Journal of Experimental Psychology: Human Perception and Performance, 26(6), 1770-1787.
Penton-Voak, I. S., Edwards, H., Percival, A., \& Wearden, J. H. (1996). Speeding up an internal clock in humans? Effects of click trains on subjective duration. Journal of Experimental Psychology: Animal Behavior Processes, 22(3), 307-320.
Spence, C., Shore, D. I., \& Klein, R. M. (2001). Multisensory prior entry. Journal of Experimental Psychology: General, 130(4), 799-832.
Troche, S. J., \& Rammsayer, T. H. (2011). Temporal information processing and mental ability: a new perspective. In Multidisciplinary Aspects of Time and Time Perception. Springer: Berlin Heidelberg.
Wearden, J. H. (1991). Human performance on an analogue of an interval bisection task. The Quarterly Journal of Experimental Psychology, 43, 59-81.
Wearden, J. H. (2015). Mission: Impossible? Modelling the verbal estimation of duration. Timing \& Time Perception, 3(3-4), 223-245.
Wearden, J. H., \& Jones, L. A. (2007). Is the growth of subjective time in humans a linear or nonlinear function of real time? The Quarterly Journal of Experimental Psychology, 60(9), 1289-1302.
Wearden, J. H., \& Penton-Voak, I. S. (1995). Feeling the heat: Body temperature and the rate of subjective time, revisited. The Quarterly Journal of Experimental Psychology, 48(2), 129-141.
Wearden, J. H., Edwards, H., Fakhri, M., \& Percival, A. (1998). Why "sounds are judged longer than lights": Application of a model of the internal clock in humans. The Quarterly Journal of Experimental Psychology: Section B, 51(2), 97-120.
Wearden, J. H., Norton, R., Martin, S., \& Montford-Bebb, O. (2007). Internal clock processes and the filled-duration illusion. Journal of Experimental Psychology: Human Perception and Performance, 33(3), 716.
Yuasa, K., \& Yotsumoto, Y. (2015). Opposite distortions in interval timing perception for visual and auditory stimuli with temporal modulations. PloS one, 10(8), e0135646.
Zampini, M., Shore, D. I., \& Spence, C. (2003). Audiovisual temporal order judgments. Experimental brain research, 152(2), 198-210.
Zampini, M., Shore, D. I., \& Spence, C. (2005). Audiovisual prior entry. Neuroscience letters, 381(3), 217-222.

# Modeling cognitive load effects in an interrupted learning task: An ACT-R approach 

Maria Wirzberger (maria.wirzberger@phil.tu-chemnitz.de)<br>E-Learning and New Media, Chemnitz University of Technology, Straße der Nationen 12, 09111 Chemnitz, Germany

Günter Daniel Rey (guenter-daniel.rey@phil.tu-chemnitz.de)
E-Learning and New Media, Chemnitz University of Technology, Straße der Nationen 12, 09111 Chemnitz, Germany

Josef F. Krems (josef.krems@psychologie.tu-chemnitz.de)<br>Cognitive and Engineering Psychology, Chemnitz University of Technology, Wilhelm-Raabe-Str. 43, 09120 Chemnitz, Germany

Without the knowledge of human cognitive processes, instructional design is blind. (Sweller, Ayres, \& Kalyuga, 2011)


#### Abstract

Based on the established framework of Cognitive Load Theory, the presented research focuses on the inspection of cognitive load factors in an interrupted learning task. The task itself is inspired from basic cognitive research and demands participants to learn abstract symbol combinations of varying complexity. In addition, they have to deal with interruptions while performing the task. Experimental results indicate the influence of task complexity on how interruptions effect learning performance. However, questions on underlying learner cognition persist, rising the need for a more in-depth way of examination. For this purpose, a cognitive model within the cognitive architecture ACT-R is developed to clarify cognitive processes and mechanisms within different conditions of the task. Preliminary results from a first model for the easy task condition already indicate some fit between human and model data. Modeling work continues with adjusting the current model and implementing a model for the difficult task condition.


Keywords: Cognitive Load; Interruptions; Learning; ACT-R

## Introduction

Learning constitutes an omnipresent requirement throughout the entire life, whether practicing to bring out the first words as a toddler, preparing for an exam within a course of study or gaining knowledge in a foreign language in mature age. When approaching learning from a psychological perspective, a variety of cognitive processes related to information capture, storage and retrieval come to the fore. They share the commonality to pose load on learners' limited mental resources, raising the need of welldesigned instructional material. Such should support learners' efforts in acquiring the desired knowledge, skills and abilities without overloading their mental capacities.

## Theoretical background

A prominent and often quoted theory in the field of instructional design is the Cognitive Load Theory (Sweller, 1988; Sweller, Ayres, \& Kalyuga, 2011). It deals with the question how certain aspects of a learning scenario demand
learners' cognitive resources. The theory postulates a practically unlimited storage capacity of long-term memory, the mental representation and organization of knowledge via schemata, and a limitation of working memory in terms of duration and capacity. In addition, mental resource demands in learning situations arise from different sources: Schema acquisition and automation build the core focus of each learning process and characterize the facet of germane load. Task complexity in relation to learners' previous knowledge constitutes intrinsic load and is traditionally defined in terms of related information that has to be processed simultaneously, referred to as element interactivity (Sweller, 2010). Extraneous load is increased by inappropriate instructional presentation and situational constraints. The latter comprise, for instance, aspects like performing the learning task in a distracting context with competing goals being present. The activation of such task-irrelevant information detracts cognitive resources needed for the learning task (Gerjets, Scheiter, \& Schorr, 2003). In consequence, learners are prone to switch to simpler tasksolving strategies that are less demanding, but at the same time less effective.

Cognitive Load Theory assumes that learning performance would be impaired if the sum of load imposed by the outlined factors exceeds the provided capacity of human working memory. However, the assumption of pure additivity has been questioned in more recent research (Park, 2010; Kalyuga, 2011; Wirzberger, Beege, Schneider, Nebel, \& Rey, 2016), supporting the need for a theoretical reformulation. A possible time-related extension assumes that intrinsic and extraneous aspects affect performance on a structural and short-term level, while the germane aspect has to be considered on processual and long-term accounts (Wirzberger et al., 2016). In consequence, load induced due to schema acquisition should change over time, while structural load facets should pose a constant level of load. A further essential pre-assumption within the postulated framework comprises the fact that spare cognitive capacity is primarily devoted to foster schema acquisition.

## Research focus

The overall project goal comprises to addresses cognitive processes behind the outlined facets of cognitive load. Within the subsection of research introduced in this paper, the particular influence of structural load components over various stages of the task is queried. In more detail, demands posed by increased task complexity and embedded interruptions are assumed to impair performance to different extents, depending on the achieved progress in the process of schema acquisition.

## Experimental setting

A basic learning task was used to approach the research focus, facilitating the concise definition and control of experimental factors. Since it required no previous knowledge, potential confounding effects of this relevant predictor could be ruled out.

## Methods

The experimental setting comprised 116 student participants ( $M_{\text {age }}=23.25$ years, $S D_{\text {age }}=4.34$, range: 18-44 years, $80 \%$ female) from different courses of study. They were required to figure out and memorize four combinations of arbitrary geometric symbols within 64 trials while being interrupted five times over the task. Interruptions occurred at the same predefined points in time (i.e., after trials 8,24 , 32, 40 and 56) for reasons of comparability across participants. Symbol combinations were either easy (two symbols) or difficult (three symbols) and split up in input (one or two symbols) and response (always one symbol). Participants were randomly assigned to one of the two combination conditions, resulting in a between-subjects manipulation of task complexity.


Figure 1: Schematic structure of a learning trial followed by an interruption in the easy task condition.

As depicted in Figure 1, in the learning part, symbols were presented one after another at the outset of each trial and participants had to indicate which symbol completed the combination. Responses were provided by selecting the correct symbol from an offered choice on the screen via mouse click. For instance, a square being displayed should result in choosing a star. After indicating their response, participants received feedback, as well as the correct solution in the case of an incorrect response. The target combinations represented the knowledge schemata that should be obtained over the task.

Within the interrupting secondary task, participants had to search, count and indicate two out of four types of geometric symbols from a visual search picture. Inspired by evidence from the subitizing task (Jensen, Reese, \& Reese, 1950), seven to nine instances per symbol were displayed, to ensure that equal cognitive mechanisms were used across participants. Performance was recorded continuously during both subtasks via correctness and duration of responses.

Regarding the experimental design, performance efficiency computed as quotient from correct responses and reaction times in seconds (Hoffman \& Schraw, 2010), represented the dependent variable. It reflected the amount of mental resources invested to acquire the task-related schema, characterizing the germane load component. Both structural load components were considered as independent variables: The number of symbols that defined a combination determined the intrinsic load component. Such a priori estimation of task complexity by the number of interacting elements followed Beckmann (2010) and Wirzberger et al. (2016). The interrupting secondary task represented the extraneous load component that was addressed in terms of inappropriate situational constraints.

## Results

The influence of interruptions on task performance in both conditions was inspected by analyses of variance (ANOVAs) based on linear mixed models with Type III sums of squares and Satterthwaite approximation for degrees of freedom of fixed effects.

Results showed significant main effects of pre- vs. postinterruption performance, $F(1,118.12)=16.71, p<.001$, and time of interruption occurrence over the task, $F(4,152.12)=11.72, p<.001$. Moreover, significant interactions between condition and pre- vs. post-interruption performance, $F(1,118.12)=16.86, p<.001$, and the condition and interruption occurrence, $F(4,152.12)=11.75$, $p<.001$, were observed. Post-hoc pairwise comparisons with Tukey's HSD supported the pattern depicted in Figure 2. They indicated a loss in performance efficiency after facing an interruption, but only in the easy task condition. The entire model achieved a conditional pseudo- $\mathrm{R}^{2}$ of .44 , indicating about $44 \%$ of explained variance.


Figure 2: Changes in efficiency due to interruptions. Error bars indicate $95 \%$ confidence intervals.

In terms of interruption performance, a significant main effect showed up for interruption occurrence over the task, $F(4,464.77)=12.53, p<.001$, while no significant difference between conditions was observable. Such pattern also receives visual support from Figure 3. The entire model obtained a conditional pseudo- $\mathrm{R}^{2}$ of .36 , indicating about $36 \%$ of explained variance.


Figure 3: Changes in interruption performance over the task. Error bars indicate $95 \%$ confidence intervals.

By contrast, when comparing the amount of totally recalled and correctly recalled symbol combinations in both conditions, participant achieved nearly equal scores that did not differ significantly.

## Discussion

Taken together, experimental results support influences of both structural load features on the observed task performance. However, the demand to inspect differences between conditions in more detail on a cognitive level arises. Although experimentally manipulated performance measurement provides a controlled way of assessment, it merely operates on indirect means and therefore lacks accessibility. On that point, the method of cognitive
modeling becomes of value, since it offers the opportunity to clarify cognitive processes and mechanisms that underlie observable performance.

## Cognitive modeling approach

Implementing a cognitive model structure raises the need to clearly think about each step within a given task and to ensure compatibility with founded psychological theories on human information processing. The cognitive architecture ACT-R (Anderson \& Lebiere, 1998; Anderson, 2007) provides an elaborated cognitive modeling approach to establish a relationship between underlying biological structures and emerging patterns of behavior. It operates on a set of modules mapping the structure of the brain, illustrated in Figure 4. While the peripheral modules are responsible for handling visual and auditory inputs and motor and vocal outputs, the central modules focus on goal planning, declarative memory, intermediate problem states and action coordination (Anderson, 2007). The predicted BOLD responses in the corresponding brain regions, for instance the basal ganglia in terms of the procedural module, have already been validated by fMRI data (Borst \& Anderson, 2015). Although processes in different modules can be executed in parallel, a limitation in capacity to one element at the same time exists, representing known bottlenecks in information processing resources (Borst, Taatgen, \& van Rijn, 2010; Nijboer, Borst, van Rijn, \& Taatgen, 2016).


Figure 4: Overview of ACT-R core modules. Adapted from Borst \& Anderson (2015) and Anderson (2007).

In contrast to other cognitive modeling approaches, ACTR is characterized by applying both symbolic and subsymbolic features (Anderson, 2007). Amongst the symbolic aspects, information is stored and processed by chunks. Interaction between modules happens by selection of production rules in the procedural module that scans the content of the buffers and, based on the resulting pattern,
chooses a suitable production rule that triggers the related action. If more than one production rule fits, the subsymbolic cost-benefit mechanism of utility decides, which production rule is selected. The level of activation, another important subsymbolic feature, reflects the availability of information in declarative memory and is determined by the context and history of use.

## Model concept

A draft of the steps to be performed during the interrupting task and the learning trials are sketched in Figure 5 and Figure 6. If an intended action cannot be finished within the given timeframe, the model can switch to the next logical step instead.


Figure 5: Outline of steps to perform in each learning trial of the task.

The concept of the cognitive model for the actual task setting is inspired by several sources of research. At first, Whelan (2007) framed a potential fMRI based measurement approach of the outlined cognitive load facets. In line with existing evidence from neuroimaging literature, he states that extraneous load triggers activity in particular in brain regions corresponding with sensory processing. Such aligns well to the extraneous load induction by a visual search task and is incorporated in the model due to the broad occupation of visual resources. The intrinsic load component is proposed to be associated with activity in brain regions responsible for maintaining and manipulating the attentional focus. In more complex tasks, entailing more interrelated elements, higher demands are posed on the corresponding goal and problem state resources. In addition, this provides a toe-hold for subsymbolic mechanisms like spreading activation, directly mapping the concept of activation distribution between related nodes of information.

Regarding the germane load facet, Whelan (2007) postulates a correspondence in particular with brain activation patterns representing motivation. This is plausible, since learners need to be motivated to dedicate available cognitive resources exclusively to schema acquisition. Based on that, in the difficult condition, a strategy shift towards a more heuristic encoding approach with increasing task progress is assumed. In detail, participants more and more tend towards retrieving the potential solution right with encoding the first symbol, which compensates for interruption costs and enables faster responses. Due to the resulting reduction in reaction time, they can achieve a better performance efficiency. The model incorporates such behavior by applying the subsymbolic mechanism of utility learning, which rewards each successful strategy adjustment.


Figure 6: Outline of steps to perform in each occurrence of the interrupting task.

Beyond that, the model bases upon existing modeling work regarding interruption and resumption during task
processing (Trafton, Altmann, Brock, \& Minz, 2003; Wirzberger \& Russwinkel, 2015). In brief, this tradition of research explains the loss in task performance after facing an interruption due to a decay in activation of the task representation. The resulting failure in accessibility of information can be adjusted within the model via subsymbolic chunk-related parameters like retrieval threshold, base-level decay or retrieval latency. On the perceptual level, the cognitive switch between both tasks is triggered bottom-up, at which the change in instruction color represents the salient screen change (Wirzberger \& Russwinkel, 2015). On the processing level, due to this salience, the interrupting task receives immediate attention, represented by a high utility of the task switch. In addition, during both stages of the task, more specific actions are regarded as more useful, for instance attending and encoding available stimuli instead of just searching around. Thus, the related productions receive slightly higher utility and can be performed as soon as they match.

Related to the concept of memory activation is the important question, which components constitute working memory in ACT-R models. The current model follows a recently introduced approach by Nijboer et al. (2016), who discuss a multi-component working memory system that can explain memory interference in dual tasking. It involves the problem state as limited short-term resource to hold and manipulate information, the activated content of the declarative memory as well as the mechanism of subvocalized rehearsal as additional support to prevent activation decay. In particular processes of rehearsal are occupied to a greater extent in the difficult condition, potentially explaining the diverging patterns between conditions.

## Preliminary results ${ }^{1}$

The currently available preliminary model is able to complete the easy task condition, highly demands visual perception, and already employs some subsymbolic parameters. Besides of an enabled base-level learning parameter, defaulting to the well-established value of 0.5 , it operates on increased visual-number finsts, aligning to the available button selection on the screen. Moreover, it induces some instantaneous noise in retrieval-related activation to better account for human variability in memory performance.

Approaching the comparison between human and model data, aside from a graphical inspection, Schunn and Wallach (2005) recommend a combination of numerical goodness-of-fit measures on relative trend magnitude and those assessing deviation from the exact location. In particular, they approve $\mathrm{R}^{2}$ as a measure of relative magnitude, for it relates directly to the accounted proportion of variance and better evaluates models with strong correlations to human data. In order to assess deviation from the exact location, the RMSSD (root mean squared scaled deviation) constitutes

[^543]the measure of choice. It scales the deviation between each mean of human and model data by the corresponding standard error of the human data mean. In this vein, the RMSSD provides a scale invariant opportunity to assess model fit in units of the standard error.


Figure 7: Comparison of performance in trials before and after an interruption.

At first glance, Figure 7 indicates a reasonable fit in terms of impaired task performance due the induced interruptions. This impression receives support by the quite well RMSSD of 3.73 and an explained proportion of variance of $32 \%$ ( $\mathrm{R}=.32$ ) for the selected pre-post interruption trials. When examining task performance in more detail, although the model can relatively map the given amount of correct responses during the learning trials, it shows a decreased match in terms of reaction time. The model constantly reacts much faster than human participants, which degrades the overall fit in performance efficiency. In addition, the model needs to better map interruption performance, indicated by Figure 8 as well as the rather high RMSSD of 7.84 and smaller proportion of explained variance of $29 \%\left(\mathrm{R}^{2}=.29\right)$.


Figure 8: Comparison of performance in interruption task
As a potential solution, the model has to speed up counting during the visual search task, since mostly it is not
able to successfully finish the second counting run or even end counting earlier.

## Further steps

Pending steps within the ongoing modeling project involve the adjustment of outlined weaknesses in model performance. Moreover, due to the theoretical match with the concept of element interactivity, spreading activation has to be included as well. A second stream of work concerns the implementation of the difficult task condition. This involves the inclusion of productions that represent the aforementioned alternating task processing strategies.

## Conclusion

Overall, this project constitutes an elaborated contribution to understanding cognitive processes that underlie knowledge acquisition from given instructional content. In doing so, it provides relevant insights into a so far rather vague defined theoretical framework, and additionally contributes to interconnect methodological approaches from different fields of research.

## Acknowledgments

The presented research is conducted within the Research Training Group "CrossWorlds - connecting virtual and real social worlds" (GRK 1780/1). The authors gratefully acknowledge funding by the German Research Foundation (DFG).

## References

Anderson, J. R., \& Lebiere, C. J. (1998). The atomic components of thought. New York: Psychology Press.
Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York, NY, USA: Oxford University Press.
Beckmann, J. F. (2010). Taming a beast of burden - On some issues with the conceptualisation and operationalisation of cognitive load. Learning and Instruction, 20, 250-264.
Borst, J. P., \& Anderson, J. R. (2015). Using the ACT-R Cognitive Architecture in combination with fMRI data. In B. U. Forstmann, \& E.-J. Wagenmakers (Eds.), An Introduction to Model-Based Cognitive Neuroscience (pp. 339-352). New York, NY, USA: Springer Science + Business Media.
Borst, J. P., Taatgen, N. A., \& van Rijn, H. (2010). The Problem state: A cognitive bottleneck in multitasking. Journal of Experimental Psychology: Learning, Memory and Cognition, 36, 363-382.
Gerjets, P., Scheiter, K., \& Schorr, T. (2003). Modeling processes of volitional action control in multiple-task performance: How to explain effects of goal competition and task difficulty on processing strategies and performance within ACT-R. Cognitive Science Quarterly, 3, 355-400.

Hoffman, B., \& Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. Educational Psychologist, 45, 1-14.
Jensen, E. M., Reese, E. P., \& Reese, T. W. (1950). The subitizing and counting of visually presented fields of dots. Journal of Psychology, 30, 363-392.
Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? Educational Psychology Review, 23, 1-19.
Nijboer, M., Borst, J., van Rijn, H., \& Taatgen, N. (2016). Contrasting single and multi-component working-memory systems in dual tasking. Cognitive Psychology, 86, 1-26.
Park, B. (2010). Testing the Additivity Hypothesis of Cognitive Load Theory (Doctoral Dissertation), Saarbruecken, Saarland, Germany: University of Saarland.
Schunn, C. D., \& Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), Psychologie der Kognition: Reden und Vortraege anlaesslich der Emeritierung von Werner Tack (pp. 115154). Saarbruecken: University of Saarland Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12, 257-285.
Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. Educational Psychology Review, 22, 123-138.
Sweller, J., Ayres, P., \& Kalyuga, S. (2011). Cognitive load theory. New York, NY, USA: Springer Science + Business Media.
Trafton, J. G., Altmann, E. M., Brock, D. P., \& Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. International Journal of Human-Computer Studies, 58, 583-603.
Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. Educational Research Review, 2, 1-12.
Wirzberger, M., Beege, M., Schneider, S., Nebel, S., \& Rey, G. D. (2016). One for all?! Simultaneous examination of load-inducing factors for advancing media-related instructional research. Computers \& Education, 100, 1831.

Wirzberger, M., \& Russwinkel, N. (2015). Modeling interruption and resumption in a smartphone task: An ACT-R approach. i-com, 14, 147-154.

# Does Associative Memory Play a Role in Solving Physics Problems? 

K. Wood, R.P. McLaren \& I.P.L. McLaren (i.p.1.mclaren@exeter.ac.uk)

School of Psychology, University of Exeter, UK


#### Abstract

Previous research has found that people frequently provide incorrect predictions about the path of moving objects when given an idealised physics problem to solve. The aim of this research was to explore whether these incorrect predictions are due to the application of an incorrect naïve physics theory, whether incorrect perceptions generated from past experiences lead to misconceptions of how moving objects behave, or whether it is a combination of both. Thirty-one participants volunteered to take part in the experiment which followed a two (experience congruent/incongruent with naïve physics theory) by two (carried versus free-moving object) within-subject design. The dependent variable was participant response (straight down or curved forwards). Results of the study revealed that participants provided answers both consistent and inconsistent with the naïve physics theory. This suggests that responses were primarily elicited through the retrieval of associatively-mediated memories of similar scenarios - some of which contain perceptual illusions. Possible methodological limitations and alternative theoretical explanations are discussed, along with practical and theoretical implications for education and learning.


## Introduction

Objects are constantly moving all around us. Therefore, it seems realistic that an individual should be able to correctly predict the path of a moving object (Zago \& Lacquaniti, 2005). However, research has found that people are frequently incorrect in their predictions when given an idealised problem to solve (McCloskey, Caramazza \& Green, 1980; McCloskey, Washburn \& Felch, 1983; McLaren, Wood \& McLaren, 2013). The aim of this study is to investigate whether conscious reasoning leads people to apply an incorrect theory to basic physics problems (a propositional account, cf. Mitchell, De Houwer and Lovibond, 2009) or whether automatic associative memories (McLaren, 2011) generate misperceptions of how moving objects behave (the associative account), or even if a combination of both is needed to explain these predictions (a dual-process account (McLaren et al., 2014)). We first introduce the on-going debate as to whether human mental life is best explained in propositional terms, associative terms or both, before going on to consider other related examples of problem solving in humans (and other animals) which may cast light on this research question.

Both the propositional (see also Shanks, 2007) and dualprocess (McLaren et al., 2014) approaches agree that human learning incorporates a conscious, calculating component, and that past experiences play a crucial role. However, disagreement exists on whether human learning can be explained by one unitary set of processes or requires multiple processes for its full characterisation. The dualprocess approach refers to two modes of processing which are fundamentally distinct: associative processes which create links between representations without conscious knowledge and regardless of the individual's intentions, and cognitive processes which consciously employ rules and reasoning (McLaren, Green \& Mackintosh, 1994). It is
argued that whilst both human and non-human species are capable of associative learning, humans alone possess rulebased processes which allow logical reasoning (Povinelli, 2004). However, Mitchell, De Houwer and Lovibond (2009) argue that evidence for these two distinct learning processes is ambiguous. They state that rule-based learning and associative learning are part of a combined system where associative learning relies on conscious, effortful and calculated processes, rather than implicit, automatic processes.

The present study explores all three of the theoretical approaches discussed above, in relation to how people solve physics problems. However, the working hypothesis is that associative processes, in the form of associative memory, primarily drive responses to physics problems in most people. Mental rotation tasks share similarities to the basic physics problems used in this current study, in requiring a simulation of the physical world (Neiworth \& Rilling, 1987). Therefore, they provide a solid starting point for exploring whether associative memory plays a primary role in responding to these types of problems. During a discrimination task involving shape rotations, Shepard and Metzler (1971) found an increased reaction time for larger rotations. This linear relationship led Shepard and Metzler to conclude that people have a general purpose rotation ability. They suggested that participants created mental representations of the first image and then, when shown the second image, mentally rotated it to match the first. However, something resembling mental rotation has also been demonstrated in pigeons (Neiworth \& Rilling, 1987). If humans alone possess rule-based processes which allow logical reasoning (Povinelli, 2004), then mental rotation in pigeons might suggest a role for associative processes in these tasks. Pigeons were trained to respond to images of clocks with rotating hands and were then tested in a discrimination task. Neiworth and Rilling found the pigeons not only performed above chance when discriminating rotations they had been trained on, but also in trials involving rotations they had not been trained on. The pigeons appeared to know where the clock hand should reappear given the time that had gone by, and were able to indicate whether it was right or wrong when the clock hand reappeared. These findings provide support for the primary role of associative memory because the pigeons had considerable experience of successive events during training, and, as a result, were later able to extrapolate novel rotations similar to the rotations they had previously experienced. It can, therefore, be argued that it is also extrapolation based on experience that enables humans to solve similar problems. Rather than utilising a generalpurpose rotation ability, people's ability to apparently mentally rotate objects might instead come from their vast experience of objects in different orientations in the environment (Edelman \& Bülthoff, 1992). A corollary of this position is that if people's responses are driven by associative memory and experience, then changing
contextual variables (i.e. the surface features of the task) should potentially change people's answers. We make use of this logic in the study reported here.

Previous research on naïve physics problems has helped inform the current research. According to McCloskey, Caramazza and Green (1980), when presented with problems involving a ball bearing entering a horizontally positioned curved tube, people tend to utilise the belief that the ball will exit the tube with a "curvy impetus". The majority of participants verbally indicated in a later interview that they believed the ball would acquire a force or momentum in the tube that would cause it to continue to travel in a curved motion upon its exit, gradually losing momentum until the trajectory became straight. This led McCloskey et al. (1980) to conclude that participants were using a naïve theory of motion. McCloskey, Washburn and Felch (1983) later provided evidence for a naïve physics theory in problems involving falling objects. Participants appeared to apply the incorrect non-Newtonian theory that, whilst free-moving objects will fall in a curved motion, carried objects will fall straight down. McCloskey et al. (1983) suggest that a perceptual illusion occurs when people observe a carried object falling. For instance, when an individual drops something whilst cycling, the object ends up behind them due to the cyclist maintaining a constant speed before noticing the object has fallen. Whilst the cyclist continues at a constant speed, the object gradually loses speed as soon as it starts to fall. Although the object falls forward in a curved motion, the observer is likely to believe that the object fell straight down (or backwards) because it is behind the cyclist in their frame of reference.

Whilst it appears that participants are consistently getting basic physics problems wrong due to explicitly applying an incorrect non-Newtonian theory, there is the alternative possibility that these responses are primarily elicited through the retrieval of associatively-mediated memories of similar scenarios - some of which contain perceptual illusions (Zago \& Lacquaniti, 2005). Participants may later offer an explanation congruent with a naïve physics theory to justify their reasoning, in an attempt to rationalise their responses in a scientific way. In order to provide evidence for the use of associatively-mediated memories in these problems, it needs to be demonstrated that individuals can produce systematically different answers to essentially the same problem (Sloman, 1996) when the contextual cues accompanying it are varied. If participants cannot explain their answers using a consistent rule, then this indicates automatic associative processing followed by conscious reasoning to justify their initial automatic responses.

A study by McLaren, Wood and McLaren (2013) explored whether experience is primary in explaining why people incorrectly predict the directions of moving objects. Participants were required to complete a questionnaire containing eight physics problems, each concerning falling objects, and were asked to indicate which path they thought each falling object would take. They found that participants gave responses that were both consistent and inconsistent with the naïve physics theory, depending on the context of the problem. In a series of structurally identical but contextually varied scenarios, around half of the carried objects were predicted (on the basis of extrapolation from experience) to fall in a curved forwards motion whilst the
others were predicted to fall straight down. Equally, around half of the free-moving objects were predicted to fall straight down whilst the others were predicted to fall in a curved forwards motion. A naïve physics view would predict that all the carried objects fall straight down, and all the free-moving objects follow a curved forwards path. McLaren et al. (2013) argued that if participants' responses were produced by applying a naïve physics theory, the theory would have to be consistently applied across all scenarios. In fact, the variation in contextual cues led to responses that were consistent with their predictions rather than the naïve physics view. Therefore, a more feasible explanation for the responses is that participants were primarily responding to these problems using their own experiences of events similar in structure or context to predict the paths of the falling objects.

The present research was predicated on the existence of both propositional and associative processes in humans. Whilst it might appear that people apply an incorrect naïve physics theory when solving basic physics problems (McCloskey et al., 1980; McCloskey et al., 1983) the presence of the inconsistent responses seen in McLaren et al.'s (2013) study suggests that associative memory plays a significant role in generating these incorrect predictions. Thus, it is possible that participants' responses are based on associative memory rather than propositional inference or reasoning and that naïve physics accounts are later presented as a reason for these responses (i.e. an epiphenomenon). This study aims to extend McLaren et al.'s (2013) work in order to determine whether results in support of the associative account can be replicated under more controlled conditions. In order to do this, a number of possible weaknesses and limitations that were identified in McLaren et al.'s (2013) study were eliminated. Firstly, the number of possible paths of falling objects that participants could pick from was reduced from five to two in order to simplify the analysis of participants' responses. Secondly, the number of problems was increased from eight to twelve to improve reliability. Thirdly, each image was carefully refined to avoid any confounding characteristics such as motion lines and object position - these are now equated across problems. Fourthly, each participant was tested on a computer individually and then interviewed afterwards rather than given a questionnaire to complete in class. This was to strengthen validity of the experiment by avoiding confounding variables such as responses of classmates at the same table and noisy distractions. The interview was done to ensure enough information was provided to generate reliable qualitative data. If these findings show responses that are both consistent and inconsistent with the naïve physics theory, depending on the context in which they are presented, this will support the theory that people are making use of associative memories when responding to these problems.

## Experiment

## Method

## Participants

Thirty-one participants volunteered to take part in the experiment. Participants consisted of 17 females and 14 males ( $\mathrm{M}=26.45$ years, $\mathrm{SD}=8.37$ ) living in Exeter and the surrounding areas ( 15 undergraduate students, 5 college
students, 10 professionals and 1 postgraduate student). All participants had a GCSE in Physics and one had an A Level in Physics. Final year psychology students with prior knowledge of the theories explored in this study were excluded from participation.

## Materials and Design

The experiment followed a 2 (experience congruent/ incongruent with naïve physics theory) by 2 (carried/freemoving object) within-subject design. The dependent variable was participant response (straight down/curved forwards). Each participant was presented with the same 12 basic physics problems, constructed using Microsoft Word and Microsoft PowerPoint and programmed using SuperLab software on a Macintosh computer. The experiment contained what was essentially the same physics problem presented in 12 different ways, using different contextual features. All of the problems were framed around falling objects travelling at the same speed. Six problems involved carried objects and the other six problems involved freemoving objects. These two types of problems were divided into subcategories, based on whether the predicted answer was congruent or incongruent with the naïve physics theory. Therefore, the four categories of problems were: Carried Congruent (e.g. Figure 1), Carried - Incongruent (e.g. Figure 2), Free-moving - Congruent (e.g. Figure 3) and Free-Moving - Incongruent (e.g. Figure 4).

There were three problems in each of the four categories. Carried - Congruent: a seagull in flight dropping an icecream, a plane dropping a crate and a running student dropping a book; Carried - Incongruent: a swinging monkey dropping a banana, a plane dropping a bouncing bomb and a running cricketer dropping a ball; Free-Moving Congruent: a cannonball fired off a cliff, a skier approaching a crevasse and a toy car falling off a table; Free-Moving Incongruent: a river flowing off a cliff, a skater dropping into a half-pipe and a toy train falling off a broken track. The congruent problems were based on problems used in previous research that resulted in responses congruent with the naïve physics theory (i.e. straight down for carried objects and curved forwards for free-moving objects). The incongruent problems, which were designed to elicit responses incongruent with the naive physics theory (i.e. curved forwards for carried objects and straight down for free-moving objects), were selected based on the research team's own visualisations of how the objects would appear to behave from associated experiences (e.g. a skater can be associated with dropping straight down into a half-pipe).

## Procedure

Firstly, participants were given a consent form to sign, which briefed them on the procedure of the experiment and their right to withdraw at any time. They then read the onscreen instructions as follows: "You are about to view a series of scenarios in which objects are seen falling to the ground. Firstly, look at the scenario and decide which path you think the object will take on its journey, and then, from the two choices offered, select which path you think most resembles the one you thought of. Indicate your choice by pressing the corresponding number on the keyboard. For example, if you think Path 1 is most similar to the path you thought of, press the key ' 1 '. When making your choice,
please ignore the effects of air/wind resistance and friction. All objects are travelling at the same speed (Speed V). When you have made your choice, let the experimenter know. You will then be asked a couple of brief questions before moving on to the next scenario. Please give as full an answer as possible. If you are happy to continue, press the space bar to begin the experiment."


Figure 4. Free-Moving - Incongruent Problem

The experiment began with one of the 12 scenarios appearing on the screen. Participants were provided with
information about the scenario (see Figures 1-4) and were then asked to press the space bar to see the possible paths of the falling object (see Figure 5). These were shown side by side, with the object falling straight down or curved forwards. Participants were asked to select the path they thought was most likely. They were then asked to let the experimenter know that they had made their decision. The experimenter responded by asking a set list of questions about the scenario: "What was happening in the scenario?", "Which answer did you select?" and "Why did you select that answer?" The 12 scenarios were presented in random order to prevent bias.


Figure 5. Example of choices provided.
At the end of the experiment, participants were asked if they followed any specific rules to help them complete the experiment and if they had any general comments about the experiment. Participants 13 to 31 were also asked if they felt their approach changed at all during the experiment. This was because it became apparent during the experiment that some participants felt they were trying to apply a rule at first and then they later noticed their responses were inconsistent with the rule they had said they were applying. The experimenter then verbally debriefed participants about the nature of the study. The duration of the experiment was approximately 15 minutes for each participant.

## Results

As a first step, the data was analysed using a 4 (condition) x 2 (response) contingency chi-square test to see if the condition (carried-congruent, carried-incongruent, freecongruent or free-incongruent) had a significant effect on the responses given (straight down or curved forwards). A contingency chi-square test was chosen because it investigates whether there is a significant relationship between two variables. A significance level of $p=.05$ was used for all statistical tests. The results of the analysis, $\chi^{2}(3)=66.94, p<.001$ suggest that there was a highly significant effect. In other words, the responses given were not independent of the condition. As the condition did appear to have an effect on the responses given (see Table 1), a number of $2 \times 2$ contingency chi-square analyses were carried out to determine the nature of this effect. Collapsing over carried/free-moving, $\chi^{2}(1)=2.43, p=.119$ (ns) indicates that congruency had no main effect. A similar analysis for carried/free-moving after collapsing over congruency gave a $\chi^{2}(1)=3.89, p=.049$, which suggests there was a marginally significant relationship between this factor and response. Straight down was the most common response for carried problems and curved forwards was the
most popular response for free-moving problems. This finding is consistent with a naïve physics effect.

Table 1
Participants' Responses Categorised by Problem's Condition

| Condition | Response |  |
| :--- | :---: | :---: |
|  | Straight Down | Curved Forwards |
| Carried - Congruent | 64 | 29 |
| Carried - Incongruent | 34 | 59 |
| Free-Moving - Congruent | 17 | 76 |
| Free-Moving - Incongruent | 62 | 31 |

However, a naïve physics theory predicts a preference for straight down responses for carried objects and curved forwards responses for free-moving objects, regardless of congruency classification. These results were further broken down to investigate whether carried versus free-moving had an effect on response when looking at congruent and incongruent problems independently. The $\chi^{2}(1)=48.31, p$ $<.001$ for congruent problems and the $\chi^{2}(1)=16.89, p<$ .001 for incongruent problems suggest that carried versus free-moving had a highly significant but quite opposing effect on responses for both congruent and incongruent problems (see Figure 6). It can clearly be seen from Figure 6 that, whilst it is possible to get a pattern of results consistent with a naïve physics theory (congruent problem data shown in blue), it is also possible to construct similar problems that elicit the reverse pattern of results. In the congruent problems, carried objects tended to produce a straight down response and free-moving objects tended to produce a curved forwards response. Conversely, the incongruent problems, revealed the opposite pattern of responses, directly contradicting the naïve physics theory.


Figure 6. This graph represents the response difference score (No. curved forwards - No. straight down) for each of the four conditions, where a positive score denotes a bias for the curved forwards response over the straight down response.

## Discussion

The aim of this study was to determine whether experiences are primary in predicting the answer of basic physics problems. In order to do this, an extension of McLaren et al.'s (2013) study was developed. Specifically, the interest was whether participants responded both consistently and inconsistently with the naïve physics theory
depending on the context in which the problems were presented. The hypothesis was that associative processes in the form of associative memory primarily drive responses to physics problems in relatively naïve participants. Therefore, it was predicted that participants would provide answers both consistent and inconsistent with the naïve physics theory based on congruency categorisation. The results of the study revealed that they did.

## Evidence for Associative and Cognitive Processes?

Whilst we have strong evidence that contextual variables can influence the responses to essentially the same physics problem, a result which strengthens the case for associative memory influencing these responses, we also have a main effect of carried/free-moving that is consistent with a naïve physics view. Admittedly this effect could be due to the fact that memories of carried objects have a general tendency to elicit a straight down response (and more than a curved forwards response) because of an incorrect perception generated from past experience (McLaren et al., 2013). If the carrier of the object is still moving at a constant speed when the object falls then it will appear that the object has fallen straight down. Nevertheless, this is a post-hoc explanation of this finding, which would have undoubtedly been predicted by a naïve physics account. Thus, we must acknowledge the possibility of a more cognitive component to the responses made to our problems.

When exploring the qualitative data, it becomes clear that many participants do offer something reminiscent of a naïve physics theory when asked to explain their responses. However, the theory provided is inconsistently applied, as congruent questions are predominately explained using a naïve physics theory, whereas incongruent problems are predominately explained by an appeal to experience. If participants were consciously applying different rules in different situations, then they should be able to explain these diverse rules afterwards. However, this was not the case. For instance, in the seagull problem, the majority of participants' explanations were of the form: "I thought it was most likely to fall straight down or backwards because the bird is in motion but the ice cream is not". This sounds like a naïve physics theory, but on the other hand, in the skateboarder problem, the majority of participants' explanations did not refer to theory but to experience: "If you see skaters on a ramp they usually go straight down". Other incongruent problems are explained using an alternative theory. For instance, some participants failed to apply the straight down belief to carried objects but instead attempted to apply the rules of Newtonian physics by simply mentioning gravity: "gravity will pull it straight down". These inconsistent explanations for responses suggest that participants are automatically responding to problems and later coming up with rules when asked to explain why they selected their responses. At the end of the experiment, many participants indicated they had later become aware that the rule they provided as an explanation directly contradicted some of their responses. $74.4 \%$ of participants said that they felt their approach to the task changed as the experiment progressed. When asked if they had any further comments about the experiment, $71.0 \%$ said "no". However, the remaining $29.0 \%$ asked if all the problems had the same answer. This shows that whilst some of the participants suspected that the answers may all be the same due to the
similar structure, perceptions based on their own experiences were so powerful that they overrode these suspicions. The fact that the problems used in this experiment were essentially the same basic problem with different surface features shows how vulnerable associatively-mediated retrieval is to a change in the surface features of a problem.

## Implications

The findings in this study suggest that when participants are presented with problems, such as the ones in this study, memories are elicited first and rules are inferred later. These findings may have significant implications for education. If experiences are primary to predicting the answer of basic physics problems, this needs to be factored into the teaching methods applied to physics. For instance, children who are studying GCSE Physics are likely to have already formed many memories about the behaviour of moving objects, and because they are likely to believe things that they see with their own eyes rather more than what they are told (Wallach, 1987) this will need to be taken into account. Just telling them the correct theory may not be enough; some explicit acknowledgement of their own experience and why it might be misleading in terms of the physics of the situation may be required.

## Strengths and Limitations

This study has successfully replicated McLaren et al.'s (2013) findings using an improved methodology, whereby participants underwent the study in a controlled environment and were able to provide both quantitative data and qualitative data. The results of this study provide strong evidence for the case that associative memory plays an important role in problem solving and learning. However, one limitation of this study is the method of obtaining incongruent problems. The problems which were designed to elicit responses incongruent with the naive physics theory (i.e. curved forwards responses for carried objects and straight down responses for free-moving objects) were selected based solely on the research team's own experience of how the objects appear to behave. It would be useful for future research to find a more independent method of selecting incongruent problems. Another limitation of the study could be that it lacks ecological validity due to the use of images of two-dimensional objects on a computer monitor rather than real-life objects.

It is possible that some propositional theorists may argue that this study does not provide evidence for the primary role of associative memory in problem solving. Mitchell et al. (2009) allow that associative memory may have a role in learning. However, they believe it is propositional reasoning that elicits memories based on previous experience and extracts a rule from them. Therefore, it could be argued that, although the responses appear to be based on the participants' past experiences, it is conscious, propositional reasoning that enables the participants to apply these past experiences to the problems presented in the study. Although this alternative explanation covers most of the facts, it cannot easily explain the pattern of qualitative results found in this study. Participants' inconsistent explanations for their responses suggest that they responded to each problem automatically and later came up with a post-hoc justification to fit the response they had given,
even if their justification directly contradicted an explanation they had given for a previous problem. At the end of the experiment, many participants indicated that they had later become aware that their explanations directly contradicted some of their responses. This shows that perceptions based on their own experiences were so powerful that they overrode any logic.

## Directions for Future Research

As pointed out earlier, the experiment involves people predicting the path of objects drawn in two-dimensional images on a computer monitor. Although the results of this experiment are highly significant, it is important to find out whether the findings can be generalised to other more practical contexts. A possible extension of this study involves participants predicting the path of falling objects in real-life situations. For instance, a student running with a book along a corridor marked with a tape measure. The student could initially run without dropping the book to give participants a real-life illustration. The experimenter would then explain to the participant that the student will drop the book at a specified point the next time he runs along the corridor and ask them to indicate where they expect the book will fall. This may result in more accurate responses due to greater ecological validity. Alternatively, the scenarios could be presented using a video recording.
This study could be extended by running the experiment with a number of samples of participants from different age groups, academic disciplines, professions, or demographics, in order to determine whether the results can be generalised to different groups of participants. It would also be interesting to run the experiment with children in different developmental stages. Although younger children will have less experience with falling objects, they are likely to rely more on automatic associative memory because, according to Piaget (1972), they will not yet have developed the ability to apply a theory when solving problems. However, interestingly, some research has found evidence for children as young as 5 years old applying what appears to be a naïve physics theory to basic physics problems (Blown \& Bryce, 2013; Kaiser, Proffitt \& McCloskey, 1985; Vosniadou, 2002). It would be interesting to explore this further.

Another possible direction for future research is to study the transition from associatively-based to rule-based performance when solving basic physics problems. This could be achieved by firstly running the initial experiment with a group of participants to get a set of responses based on associative memory, and then training the same group of participants on Newtonian mechanics before running the experiment (using different problem variants) with them again. In the second trial, the participants would presumably respond correctly to the problems because they would be able to apply a rule they had recently learnt. This may sound straightforward, but note that most of the participants in this study had some knowledge of physics, and still made systematic errors.

## Conclusion

In conclusion, this study has provided strong evidence for the primary role of associative memory in human learning and problem solving. The highly reliable results show that people are frequently incorrect in their predictions of the
paths of moving objects, highlighting the importance of studying associative processes in humans and their interaction with more cognitive (propositional) processing. We suspect that our results will have practical implications for education, especially instruction in physics and applied mathematics.

## References

Blown, E. J., \& Bryce, T. G. K. (2013). Thought-experiments about gravity in the history of science and in research into children's thinking. Science \& Education, 22(3), 419-481.
Edelman, S., \& Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Research, 32(12), 2385-2400.
Kaiser, M. K., Proffitt, D. R., \& McCloskey, M. (1985). The development of beliefs about falling objects. Perception \& Psychophysics, 38(6), 533-539.
McCloskey, M., Caramazza, A., \& Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. Science, 210, 1139-41.
McCloskey, M., Washburn, A., \& Felch, L. (1983). Intuitive physics: The straight down belief and its origin. Journal of Experimental Psychology: Learning, Memory and Cognition, 9, 636-649.
McLaren, I. P. L. (2011). APECS: An adaptively parameterised model of associative learning and memory, in Alonso, E., \& Mondragón, E. (Eds.). Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications. Hershey, PA: IGI Global.
McLaren, I. P. L., Forrest, C. L. D., McLaren, R. P., Jones, F. W., Aitken, M. R. F., \& Mackintosh, N. J. (2014). Associations and propositions: The case for a dual-process account of learning in humans. Neurobiology of Learning and Memory, 108, 185-195.
McLaren, I. P. L., Green, R. E. A. \& Mackintosh, N. J. (1994) Animal learning and the implicit/explicit distinction. In: Ellis, N. (Eds.). Implicit and Explicit Learning of Languages, London, Academic Press.
McLaren, I. P. L., Wood, K., \& McLaren, R. P. (2013) Naïve Physics: the wrong theory? In Knauff, M., Pauen, M., Sebanz, N., \& Wachsmuth, I. (Eds.). Proceedings of the 35th Annual Conference of the Cognitive Science Society, 1008-1013. Austin, TX: Cognitive Science Society.
Mitchell, C. J., De Houwer, J., \& Lovibond, P. F. (2009). The propositional nature of human associative learning. Behavioural and Brain Sciences, 32(2), 183-198.
Neiworth, J. J., \& Rilling, M. E. (1987). A method for studying imagery in animals. Journal of Experimental Psychology: Animal Behaviour Processes, 13(3), 203.
Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. Human Development, 15, 1-12.
Povinelli, D. J. (2004). Behind the ape's appearance: Escaping anthropocentrism in the study of other minds. Daedalus, 133, 2941.

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. The Quarterly Journal of Experimental Psychology, 60(3), 291-309.
Shepard, R. N. \& Metzler, J. (1971). Mental rotation of threedimensional objects. Science, 191, 701-703.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119, 3-22.
Vosniadou, S. (2002). On the nature of naïve physics. In Reconsidering Conceptual Change: Issues in Theory and Practice (pp. 61-76). Springer Netherlands.
Wallach, H. (1987). Perceiving a stable environment when one moves. Annual Review of Psychology, 38, 1-29.
Zago, M., \& Lacquaniti, F. (2005). Cognitive, perceptual and action-oriented representations of falling objects. Neuropsychologia, 43(2), 178-188.

# Sensorimotor Learning Modulates Automatic Imitation in Visual Speech 

Yuchunzi Wu<br>Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK<br>Bronwen Evans<br>Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2<br>Wakefield Street, London WC1N 1PF, UK<br>Patti Adank<br>Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2<br>Wakefield Street, London WC1N 1PF, UK


#### Abstract

People automatically imitate observed actions, including speech. Automatic Imitation (AI) is linked to observationexecution associations in the mirror neuron system (MNS). AI is measured using interference tasks, in which prompts (say "ba" or "da") are paired with congruent or incongruent distracters (video of someone saying "ba" or "da"). Faster responses for congruent than for incongruent prompt-distracter pairings signal AI. Observation-execution associations for speech actions are thought to be inflexible, unlike associations for manual actions, which have been shown to be flexible. We trained participants to reinforce or abolish their AI response by providing them with compatible (say "ba" for a video of someone saying "ba") or incompatible training (say "ba" for a video of "da"). After training, the AI response was reduced for participants who received incompatible training, thus showing that the MNS for speech actions is also flexible and subject to experience, like the MNS for manual actions.


# Whoa! Aww ... Ohh ... Hee! and Mmm: Infants' nuanced distinctions about the probable causes of emotional expressions 

Yang Wu

MIT
Paul Muentener
Tufts University

## Laura Schulz

MIT


#### Abstract

Can infants map diverse positive emotional expressions to their probable causes? Across two studies (including one pre-registered experiment), we used a preferential-looking task to find that infants as young as 12-17 months (mean: 14.8 months) successfully matched non-verbal vocalizations elicited by funny, exciting, adorable, sympathetic, and delicious images to their probable causes (Experiments 1 and 2). Do infants also posit unobserved causes of emotional expressions? In both exploratory and pre-registered experiments, an adult peeked into a box and made one of two distinct positive emotional vocalizations (Experiment 3: "Aww!" or "Mmm!"; Experiment 4: "Aww!" or "Whoa!"). Infants reaching into the box retrieved either a probable or improbable cause of the reaction. Infants were more likely to search again on incongruent trials. These results suggest that infants make nuanced distinctions among emotions, and infer probable causes of emotional reactions.


# Visual and Audio Aware Bi-Modal Video Emotion Recognition 

Siqi Xiang (sqxiang@buaa.edu.cn)<br>Wenge Rong (w.rong@buaa.edu.cn)<br>Zhang Xiong (xiongz@buaa.edu.cn)<br>School of Computer Science and Engineering, Beihang University, Beijing 100191, China<br>Min Gao (gaomin@cqu.edu.cn)<br>Qingyu Xiong (xiong03@cqu.edu.cn)<br>School of Software Engineering, Chonqing University, Chongqing 401331, China


#### Abstract

With rapid increase in the size of videos online, analysis and prediction of affective impact that video content will have on viewers has attracted much attention in the community. To solve this challenge several different kinds of information about video clips are exploited. Traditional methods normally focused on single modality, either audio or visual. Later on some researchers tried to establish multi-modal schemes and spend a lot of time choosing and extracting features by different fusion strategy. In this research, we proposed an end-toend model which can automatically extract features and target an emotional classification task by integrating audio and visual features together and also adding the temporal characteristics of the video. The experimental study on commonly used MediaEval 2015 Affective Impact of Movies has shown this method's potential and it is expected that this work could provide some insight for future video emotion recognition from feature fusion perspective.


Keywords: videos; multi-modal scheme; modal fusion; end-to-end; temporal characteristics

## Introduction

To better understand and analyse people's emotion response during watching videos, it is essential to study the cognitive determinants beneath the video presentation. Currently, content based approaches are the main trend for video emotion analysis, and a lot of models have been proposed to help identify the emotions evoked by videos (Hanjalic, 2006), among which affective analysis based on video visual contents have been studied for several years. Several approaches which employed different machine learning models such as Bayesian network (Soleymani, Kierkels, Chanel, \& Pun, 2009), Hidden Markov Models (Kang, 2003) have been proposed and proven applicable to tackle with this challenge.

Though visual content based video emotion analysis has proven applicable in real applications, there still exists challenges since even the same scene could cause different emotions (Choe, Chun, Noh, Lee, \& Zhang, 2013). Recently audio related features have also proven its effectiveness in emotion analysis (Cui, Jin, Zhang, Luo, \& Tian, 2010). For example, Xu et al. tried to use audio emotional events (AEE) such as laughing, horror sounds and other features to detect horror and comedy movies (Xu, Chia, \& Jin, 2005).

While previous studies focused on video or audio features alone in detecting video emotion have proven their ease in implementation, to further improve the classification performance, some researchers indicate the possibility by combining visual features with audio features to form a hybrid fea-
ture that can carry information from two different modalities (domains) at the same time. Such methods can be roughly divided into two categories in terms of the way the features are combined, i.e., later fusion of classifiers (Yi, Wang, Zhang, \& Yu, 2015), and early fusion scheme, in which features are concatenated into a final classifier (Dai et al., 2015; P., Hayrapetyan, Tapaswi, \& Stiefelhagen, 2015).

In this research, we employed the idea of modal fusion and then proposed an end-to-end framework to integrate the visual and audio features for video emotion analysis. Recently with the development of deep learning techniques, a lot of advanced methods have been proposed for feature extraction. In this research, we used convolutional neural network (CNN) to extract video emotion related features as CNN has proven its success in learning intermediate representations from lowlevel features (Acar, Hopfgartner, \& Albayrak, 2014). Afterwards, taking into account the temporal characteristics of video, we further use Long Short Term Memory (LSTM) model (Hochreiter \& Schmidhuber, 1997) to integrate the extracted temporal features since it performs well on tasks that require integration of state information over time. Finally a multi-layer perceptron (MLP) is employed to classify the final video emotions.

To confirm the validity of the proposed method, we implement it in the Affective Impact of Movies Task 3 in the MediaEval challenge 2015 (Sjöberg et al., 2015). The task has now become a state-of-the-art benchmark which attracted a large number of research teams to test their models on this data set. The experimental study result against different bench experiments on this dataset shows the proposed method's potential in detecting video's emotion.

## Related Work

In the content-based video research, many researchers have used a lot of models to identify the emotions triggered by the video. Hanjalic argued the possibility to classify films according to their emotions and proposed the concept of "expectation of emotion", which is defined as one or a group of emotions that a filmmaker wishes to use to communicate with a certain culture or a particular audience through the film (Hanjalic, 2006). Through this concept, he proposed the video content information and its underlying characteristics to predict emotion. Later on, Soleymani et al. proposed a Bayesian framework to detect scene affect and the arousal


Figure 1: Visual and Audio Aware Video Emotion Analysis Framework
and valence values with content features were used to classify video emotions into 3 classes, i.e., calm, excited positive and excited negative (Soleymani et al., 2009). Similarly Arifin and Cheung established a framework based on the hierarchical coupling of dynamic Bayesian networks to establish the dependencies of the pleasure-activity-dominance emotion model (Arifin \& Cheung, 2008).

There are also various studies on video affective characterization using audio features, e.g., rhythm, tempo, melfrequency cepstral coefficients (MFCC), pitch, zero crossing rate. For example, three feature sets, i.e., intensity, timbre and rhythm were extracted from audio to classify video emotion using Gaussian mixture models (Lu, Liu, \& Zhang, 2006). Similarly, Xu et al. tried to use audio emotional events (AEE), including laughing, horror sounds and other features to detect horror and comedy movies (Xu et al., 2005).

In fact, there is a complex interaction between the audio and visual contents to determine the perceived mood. As such the video emotion analysis has begun to use feature fusion method to classify emotion into different classes (Yi et al., 2015). Similarly, Trigeorgis et al. selected the low level descriptors with the traditional adaboost as a classifier (Trigeorgis et al., 2015). Wang and Cheong derived the characteristics of multimodality by probabilistic inference based on two SVM models (Wang \& Cheong, 2006), where one SVM model is designed to process audio data and extracts the corresponding advanced audio information, while another SVM model is used to classify the captured video segments.

However, since these framework extracts basic features, they lack the ability to use raw inputs to automatically learn mid-level representations. With the development of deep learning techniques, some deep learning based approaches are also proposed in the literature. For example, Kahou et al. used a deep convolution neural network to analyse facial expressions within a frame and used a deep belief net to capture audio feature (Kahou et al., 2016). Levi and Hassner also
used convolution neural network to capture visual features to classify video into seven emotions (Levi \& Hassner, 2015).

## Proposed Approach

The overall pipeline of the proposed visual and audio aware emotion analysis framework is depicted as Fig. 1, where the whole process is divided into three steps: 1) video segment and low level feature extraction; 2) bi-modal visual and audio feature fusion; and 3) temporal feature integration and emotion classification.

## Video segment and low level feature extraction

To analyze video emotion, it is necessary to firstly divide a video into short videos with a length of $t$ seconds. In this study we set $t=1$ so that a video of length $T$ will have $T$ slices. This segmentation has two benefits. First, since the length of each video is different, this segmentation gives us better access to the visual and audio features. Second, Because of the temporal characteristics of the video, cutting the video into the same segments can be used for subsequent recurrent neural networks.

For each segment, we need to extract its visual and audio features separately. As to the visual features, we extract the $k$ key frames for each segment. Due to the strong correlation among frames within a second, we select $k=1$. The key frame is defined as the frame with the closest RGB histogram to the mean RGB histogram of the whole video clip using the Manhattan distance (Zhu, Jiang, Peng, \& Zhong, 2016). Assume that a video clip $V$ contains $n$ frames, the RGB histogram of $i$-th frame is defined as $h(i)$. The Manhattan distance $D$ between two frames $i$ and $j$ is calculated as follows:

$$
\begin{equation*}
D(i, j)=|h(i)-h(j)| \tag{1}
\end{equation*}
$$

and the key frame will be:

$$
\begin{equation*}
\underset{i}{\arg \min } D\left(i, \frac{1}{n} \sum_{j=1}^{n} h(j)\right) \tag{2}
\end{equation*}
$$

After getting the key frame, it will be resized to $256 * 256$ pixel, as suggested in (Krizhevsky, Sutskever, \& Hinton, 2012) as input for fine-tuning. The concept of fine-tuning is to use a model pre-trained on a large dataset, replacing its last layers, and fine-tune the weights on new task using backpropagation. In this study, AlexNet (Krizhevsky et al., 2012) is employed. AlexNet consists of five convolution layers and three fully connected layers. Here we select the fc7 layer of AlexNet which has 4096 neurons as our visual features.

As to the audio features, the traditional methods for audio emotion analysis need to select proper audio features, e.g., MFCC, energies, flatness, and etc. But they often have to conduct a lot of repeat tests to choose the best features. In order to take full advantage of the depth convolution neural network model in extracting data features, the original features of the data should be kept as much as possible in order to avoid losing information. In this research, we process the audio to spectrogram (Barker \& Virtanen, 2016), which is a visual representation of the spectrum of frequencies in a sound. We set the window function to 40 ms and the hop size to 20 ms to generate a spectrogram every second using short-time Fourier transform with a Hamming window (Allen, 1977). The resulting image is resized to $256 * 256$ pixels, here we also use the method of AlexNet finetune to extract the fc7 layer as a feature of the spectrogram.

## Bi-modal visual and audio feature fusion

From the last step we obtained visual and audio features for video emotion analysis. However, the length of features of both visual and audio is long and there maybe many redundancy in the features. It will be helpful if we can combine the two types of features and then reduce the overall dimension.

Let $x_{a} \in R^{D}$ denotes audio features and $x_{v} \in R^{D}$ denotes visual features, where $D \in R$ is the dimension of audio and visual features, the joint representation of features by fusion modal can be written as:

$$
\begin{equation*}
x_{f}=\alpha_{a} g\left(x_{a} ; w_{a}\right)+\alpha_{v} g\left(x_{v} ; w_{v}\right) \tag{3}
\end{equation*}
$$

where $g($.$) denotes the hidden layer of both audio and visual$ channel. $\alpha_{a}$ defines the weights of audio features and $\alpha_{v}$ defines the weights of visual features at the same time. The hidden layer of audio features is:

$$
\begin{equation*}
g\left(x_{a} ; w_{a}\right)=\theta\left(\left(w_{a}, x_{a}\right)+b_{a}\right) \tag{4}
\end{equation*}
$$

where $\theta$ denotes the activation function (rectified linear units (Zeiler et al., 2013), sigmoid etc.) of the audio hidden layer. Similarly the hidden layer of visual feature is:

$$
\begin{equation*}
g\left(x_{v} ; w_{v}\right)=\theta\left(\left(w_{v}, x_{v}\right)+b_{v}\right) \tag{5}
\end{equation*}
$$

## Temporal feature integration and emotion analysis

Though previous steps we have obtained fused features from visual and audio perspective, there is still a challenge about how to predict corresponding emotion status. Furthermore, in previous step the features are about a single frame, taking into account the temporal characteristics of video, it is necessary to study how these features can be used over time. In this research we will use the LSTM model to fuse sequence features together.

Recurrent Neural Networks (RNNs) are powerful networks and it can model input sequences of different lengths, because the parameters of the network can be shared over different parts (Mikolov, Karafiát, Burget, Cernocký, \& Khudanpur, 2010). RNNs are often trained by Back-Propagation Through Time (BPTT) algorithm, but the main problem with the BPTT is that the gradients tend to vanish or explode which was resulted by propagating the gradients down through layers. Therefore it is difficult to learn efficient long-term dependencies. To overcome this limitation, the Long-Short-Term-Memory (LSTM) (Hochreiter \& Schmidhuber, 1997) units have been created to capture long-term dependencies. LSTMs have the ability to remove or add information to the cell state through a well-designed structure called a "gate". It is believed that the LSTMs can model the temporal aspect of induced emotions in our task. Various units have been proposed in the community to constitute a LSTM. In this research, we employed the LSTM units described in (Zaremba \& Sutskever, 2014). The LSTM unit of time step $t$ consists of three sigmoidal gates, i.e., input gate $i_{t}$, output gate $o_{t}$, forgetting gate $f_{t}$. The most important part of the LSTM unit is a linear self-loop state cell $c_{t}$. The memory cell unit $c_{t}$ is a sum of two terms: the previous memory cell unit $c_{t-1}$ which is modulated by $f_{t}$, and $g_{t}$, a function of the current input and previous hidden state, modulated by the input gate it. $h_{t}$ denotes the hidden layer's output at step $t$. We can update our hidden layer for time step $t$ as follows:

$$
\begin{gather*}
i_{t}=\sigma\left(W_{x i} x_{t}+W_{h i} h_{t-1}+b_{i}\right)  \tag{6}\\
f_{t}=\sigma\left(W_{x f} x_{t}+W_{h f} h_{t-1}+b_{f}\right)  \tag{7}\\
o_{t}=\sigma\left(W_{x o} x_{t}+W_{h o} h_{t-1}+b_{o}\right)  \tag{8}\\
g_{t}=\tanh \left(W_{x c} x_{t}+W_{h c} h_{t-1}+b_{c}\right)  \tag{9}\\
c_{t}=f_{t} \odot c_{t-1}+i_{t} \odot g_{t}  \tag{10}\\
h_{t}=o_{t} \odot \tanh \left(c_{t}\right) \tag{11}
\end{gather*}
$$

where $x_{t}$ is the current fusion feature, $h_{t-1}$ is the previous hidden layer vector. $x \odot y$ denotes the element-wise product of vectors $x$ and $y$. In addition, $W_{x i}, W_{x f}, W_{x o}, W_{x c}, W_{h i}, W_{h f}$, $W_{h o}, W_{h c}$ are weights for the gates, and $b_{i}, b_{f}, b_{o}, b_{c}$ are biases for the gates. $\sigma$ is the nonlinear methods (e.g., sigmoid or tanh).

The output of the last time step of LSTM unit will be the input of the fully connected neural network, also known as multi-layer perception (MLP).The hidden layers and parameters of MLP will discuss in experiment. The prediction layer
will have 3 units $y_{l}(l=0,1,2)$ and the class probability is calculated by taking the softmax as below:

$$
\begin{equation*}
y_{l}: p\left(y_{l}=c\right)=\frac{\exp \left(y_{l}, c\right)}{\sum_{c^{\prime} \in C} \exp \left(y_{l}, c^{\prime}\right)} \tag{12}
\end{equation*}
$$

where $C$ denotes the three emotion states. Finally the label with the max probability will be the expected label.

## Experimental Study

## Dataset

In order to fairly verify the performance of our proposed method, we implement it on the dataset provided by MediaEval 2015 Afective Impact of Movies task (Sjöberg et al., 2015), which consists of 10,900 short video clips extracted from 199 Creative Commons-licensed movies of various genres. It is an extension of the LIRIS-ACCEDE dataset (Baveye, Dellandréa, Chamaret, \& Chen, 2015), which originally contains 9,800 excerpts extracted from 160 movies. The MediaEval 2015 task added 1,100 video clips additionally from 39 movies. The dataset is divided into training set and test set. The training set consists of 6,144 videos extracted from 100 movies while the test set includes 4,756 videos extracted from the remaining 99 movies. These videos last from 8 to 12 seconds and start and end with a cut or fade. The ground truth for each of 10,900 video clips consists of discrete labels for arousal (calm-neutral-active) and valence (negative-neutral-positive).

## Evaluation Metrics \& Baseline

In order to evaluate the affective detection task, the official and complete method is global precision (Sjöberg et al., 2015), which is the proportion of the number of correctly assigned videos in the total video samples and is defined as:

$$
\begin{equation*}
\text { Precision }=N_{c} / N_{t} \tag{13}
\end{equation*}
$$

where $N_{c}$ is the number of videos which are assigned to the correct class, and $N_{t}$ is the total number of test videos. In this research, we only compare the results obtained for the arousal classification. This is because compared to arousal, valence is not sensitive in the dataset. As such comparing the results of the arousal classification is a commonly adopted choice (Sjöberg et al., 2015).

To evaluate applicability of the model fusion approach, in this research we compared it against the proposed approach in predicting arousal values using only the image features or audio features. Furthermore, we also compared the proposed approach against early fusion and later fusion methods, respectively. In the early fusion model we simply concatenate the audio and video features together, while in later fusion schema, we firstly trained two MLP classifiers to represent the two modalities separately. Their predictions are denoted as $p_{a}$ and $p_{t}$ and the overall output emotion class can be assigned by

$$
\begin{equation*}
p=\alpha p_{a}+(1-\alpha) p_{t} \tag{14}
\end{equation*}
$$

where $\alpha$ indicates the relevant importance between audio and visual features. In this research we set $\alpha=0.56$, as indicated in (Goyal, Kumar, Guha, \& Narayanan, 2016).

Afterwards we also compare our results against state-ofart systems in the MediaEval 2015 challenge. These systems include: later fusion models with manually selected features (Yi et al., 2015; Chakraborty et al., 2015), early fusion models with manually selected features (P. et al., 2015; Trigeorgis et al., 2015), later fusion models with automatically selected features (Tiwari et al., 2016), early fusion models with automatically selected features (Dai et al., 2015; Seddati et al., 2015).

## Experiment Settings

We tested the different feature dimensions and found that the final result did not change much in the range of 250 to 1000. We decided to use feature size of 512 for both visual pathway and audio pathway. Therefore the fusion feature as the final LSTM model input has 512 dimensions. LSTM model can handle different video length, the longest video is 18 seconds that is 18 time steps. The system is trained end-to-end to predict the videos emotion class at each time step. It is found that the most significant parameter is the number of LSTM hidden layers. We compared LSTM networks with 64, 128, 256, and 512 hidden units, separately. Finally, we found that 256 hidden units can be selected to achieve the best results, as shown in Fig. 2. Afterwards we selected MLP as our classifier in which rectified linear units were used as nonlinear functions and stochastic gradient descent with minibatches was used for parameter updates (Zeiler et al., 2013). Also we used categorical cross-entropy loss function to get the best results. The hidden layer uses dropout to prevent overfitting, and the factor is set 0.5 . The number of hidden layer of MLP and the units' number can also affect the model results, and ultimately we chose one hidden layer with 64 hidden units.


Figure 2: Arousal Accuracy with Different Number of Hidden Layer Units

## Result and Analysis

Table 1 presents the proposed method's performance using different feature space and fusion strategy. It is observed that the performance of all feature fusion strategies are better than using only single feature. It is because that video images are the main cause of people's emotions, but audio
can complement the lack of information in video images. It is further found that our proposed model fusion method is better than both the simple early fusion and late fusion, affirming the effectiveness of multi-modal emotion classification. This maybe because early fusion leads to the sparsity of input vectors and late fusion has little consideration for visual and audio's correlation (Williams et al., 2009).

Table 1: Comparison of accuracy by different fusion models

| Approaches | Arousal Accuracy(\%) |
| :---: | :---: |
| Visual features only | 55.51 |
| Audio featues only | 55.14 |
| Early fusion | 55.71 |
| Later fusion | 55.70 |
| Modal fusion | $\mathbf{5 5 . 8 9}$ |

Table 2 is the experimental result of the propose method against most recently revealed results. The result demonstrates the feasibility and superiority of end-to-end training for video emotion classification. It is found that one system's result (Yi et al., 2015) is slightly higher (less than $0.1 \%$ ) than the proposed one. However, its features are selected manually, which is time-consuming, not universal and not portable. What's worse, their feature dimension is also long. End-toend training has better transfer learning properties and the training process is convenient. Using a well-trained model for another similar problem only needs a simple refinement. It is also observed from the table that the method proposed in (Tiwari et al., 2016) has the similar feature size to ours, while the proposed model outperforms their final arousal accuracy. This may because their feature fusion approach is rough and does not consider the temporal characteristics. This demonstrates that temporal features could play a role in video emotion analysis to a certain extent. As for the other methods, our result can outweigh them which shows that modal fusion has a great advantage compared with simple early fusion and later fusion. Fusing visual and audio feature in a mid-level is a potential strategy since visual and audio information in video have a certain interaction. It can also inferred that CNN has good performance in visual and audio feature extraction.

## Conclusion and Future Work

Video emotion recognition is an important challenge as detecting affective attitudes is an important research field in cognitive science. It is argued that visual and audio information are both important in detecting video emotion. Therefore in this paper we used a deep learning architecture to fuse visual and audio modalities for video affective classification. This end-to-end framework has the advantages of simple training and convenient transplantation and demonstrates that modal fusion with small size of features can compare against most state-of-art results obtained by participants of the MediaEval 2015 Affective Impact of Movies task. Furthermore, it would be interesting to study if it is feasible to
include information from other domains/modalities, e.g., abstract words (Siakaluk, Knol, \& Pexman, 2014), which deserve future study in the future work.

## Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (No. 61332018), the National Department Public Benefit Research Foundation of China (No. 201510209), and the Basic and Advanced Research Projects in Chongqing (No. cstc2015jcyjA40049).

## References

Acar, E., Hopfgartner, F., \& Albayrak, S. (2014). Understanding affective content of music videos through learned representations. In Proceedings of the 20th Anniversary International Conference on MultiMedia Modeling, Part I (pp. 303-314).
Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete fourier transform. IEEE Transactions on Acoustics Speech and Signal Processing, 25(3), 235-238.
Arifin, S., \& Cheung, P. Y. K. (2008). Affective level video segmentation by utilizing the pleasure-arousal-dominance information. IEEE Transactions on Multimedia, 10(7), 1325-1341.
Barker, T., \& Virtanen, T. (2016). Blind separation of audio mixtures through nonnegative tensor factorization of modulation spectrograms. IEEE/ACM Transactions on Audio, Speech and Language Processing, 24(12), 2377-2389.
Baveye, Y., Dellandréa, E., Chamaret, C., \& Chen, L. (2015). LIRIS-ACCEDE: A video database for affective content analysis. IEEE Transactions on Affective Computing, 6(1), 43-55.
Chakraborty, R., Maurya, A. K., Pandharipande, M., Hassan, E., Ghosh, H., \& Kopparapu, S. K. (2015). TCS-ILAB - mediaeval 2015: Affective impact of movies and violent scene detection. In Working Notes Proceedings of the MediaEval 2015 Workshop.
Choe, W., Chun, H., Noh, J., Lee, S., \& Zhang, B. (2013). Estimating multiple evoked emotions from videos. In Proceedings of 35th Annual Meeting of the Cognitive Science Society (pp. 2046-2051).
Cui, Y., Jin, J. S., Zhang, S., Luo, S., \& Tian, Q. (2010). Music video affective understanding using feature importance analysis. In Proceedings of the 9th ACM International Conference on Image and Video Retrieval (pp. 213-219).
Dai, Q., Zhao, R., Wu, Z., Wang, X., Gu, Z., Wu, W., \& Jiang, Y. (2015). Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning. In Working Notes Proceedings of the MediaEval 2015 Workshop.
Goyal, A., Kumar, N., Guha, T., \& Narayanan, S. S. (2016). A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 2822-2826).

Table 2: Comparison with state-of-the-art approaches of MediaEval 2015 Affective Impact of Movies task

| Approaches | Fusion method | Feature selection type | Feature length | Arousal Accuracy(\%) |
| :---: | :---: | :---: | :---: | :---: |
| (Yi et al., 2015) | Later fusion | Manual | $>4,000$ | 55.93 |
| (Trigeorgis et al., 2015) | Early fusion | Manual | 1000 | 55.72 |
| (P. et al., 2015) | Early fusion | Manual | $>100,000$ | 51.90 |
| (Chakraborty et al., 2015) | Later fusion | Manual | 1,000 | 48.95 |
| (Tiwari et al., 2016) | Early fusion | Automatic | 500 | 55.85 |
| (Seddati et al., 2015) | Early fusion | Automatic | 20,000 | 52.44 |
| (Dai et al., 2015) | Early fusion | Automatic | 10,000 | 48.70 |
| Proposed approach | Modal fusion | Automatic | 512 | 55.89 |

Hanjalic, A. (2006). Extracting moods from pictures and sounds: towards truly personalized TV. IEEE Signal Processing Magazine, 23(2), 90-100.
Hochreiter, S., \& Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
Kahou, S. E., Bouthillier, X., Lamblin, P., Gülçehre, Ç., Michalski, V., Konda, K., . . . Bengio, Y. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal of Multimodal User Interfaces, 10(2), 99-111.
Kang, H. (2003). Affective content detection using HMMs. In Proceedings of the 11th ACM International Conference on Multimedia (pp. 259-262).
Krizhevsky, A., Sutskever, I., \& Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Proceedings of 26th Annual Conference on Neural Information Processing Systems (pp. 1106-1114).
Levi, G., \& Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In Proceedings of 2015 ACM on International Conference on Multimodal Interaction (pp. 503-510).
Lu, L., Liu, D., \& Zhang, H. (2006). Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech and Language Processing, 14(1), 5-18.
Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., \& Khudanpur, S. (2010). Recurrent neural network based language model. In Proceedings of 11th Annual Conference of the International Speech Communication Association (pp. 1045-1048).
P., M. V., Hayrapetyan, S., Tapaswi, M., \& Stiefelhagen, R. (2015). KIT at mediaeval 2015 - evaluating visual cues for affective impact of movies task. In Working Notes Proceedings of the MediaEval 2015 Workshop.
Seddati, O., Kulah, E., Pironkov, G., Dupont, S., Mahmoudi, S., \& Dutoit, T. (2015). Umons at mediaeval 2015 affective impact of movies task including violent scenes detection. In Working Notes Proceedings of the MediaEval 2015 Workshop.
Siakaluk, P. D., Knol, N., \& Pexman, P. M. (2014). Effects of emotional experience for abstract words in the stroop task. Cognitive Science, 38(8), 1698-1717.

Sjöberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., ... Chen, L. (2015). The mediaeval 2015 affective impact of movies task. In Working Notes Proceedings of the MediaEval 2015 Workshop.
Soleymani, M., Kierkels, J. J. M., Chanel, G., \& Pun, T. (2009). A bayesian framework for video affective representation. In Proceedings of 3rd International Conference on Affective Computing and Intelligent Interaction.
Tiwari, S. N., Duong, N. Q., Lefebvre, F., Demarty, C.H., Huet, B., \& Chevallier, L. (2016). Deep features for multimodal emotion classification. Retrieved from https://hal.inria.fr/hal-01289191
Trigeorgis, G., Coutinho, E., Ringeval, F., Marchi, E., Zafeiriou, S., \& Schuller, B. W. (2015). The ICL-TUMPASSAU approach for the mediaeval 2015 "affective impact of movies" task. In Working Notes Proceedings of the MediaEval 2015 Workshop.
Wang, H. L., \& Cheong, L. F. (2006). Affective understanding in film. IEEE Transactions on Circuits and Systems for Video Technology, 16(6), 689-704.
Williams, S., Oliker, L., Vuduc, R. W., Shalf, J., Yelick, K. A., \& Demmel, J. (2009). Optimization of sparse matrix-vector multiplication on emerging multicore platforms. Parallel Computing, 35(3), 178-194.
Xu, M., Chia, L., \& Jin, J. S. (2005). Affective content analysis in comedy and horror videos by audio emotional event detection. In Proceedings of 2005 IEEE International Conference on Multimedia and Expo (pp. 622-625).
Yi, Y., Wang, H., Zhang, B., \& Yu, J. (2015). MIC-TJU in mediaeval 2015 affective impact of movies task. In Working Notes Proceedings of the MediaEval 2015 Workshop.
Zaremba, W., \& Sutskever, I. (2014). Learning to execute. CoRR, abs/1410.4615.
Zeiler, M. D., Ranzato, M., Monga, R., Mao, M. Z., Yang, K., Le, Q. V., ... Hinton, G. E. (2013). On rectified linear units for speech processing. In Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 3517-3521).
Zhu, Y., Jiang, Z., Peng, J., \& Zhong, S. (2016). Video affective content analysis based on protagonist via convolutional neural network. In Proceedings of 17th Pacific-Rim Conference on Multimedia, Part I (pp. 170-180).

# Teaching Versus Active Learning: A Computational Analysis of Conditions that Affect Learning 

Scott Cheng-Hsin Yang (scott.cheng.hsin.yang@gmail.com) \& Patrick Shafto (patrick.shafto@gmail.com)<br>Department of Mathematics \& Computer Science, Rutgers University-Newark


#### Abstract

Researchers have debated whether instructional-based teaching or exploration-based active learning is better for decades with unsatisfying results. A main obstacle is the difficulty in precisely controlling and characterizing the pedagogical methods used and the learning conditions in empirical studies. To address this, we leveraged existing computational models of teaching and active learning to formalize the methods and the learning process. We compared the two pedagogical methods in a concept-learning framework and investigated their effectiveness under various scenarios. Our results show that when the learner and teacher are conceptually aligned, teaching is at least as effective as, and often much more effective than active learning, but when the alignment is broken, active learning can yield moderate improvement over teaching. We conclude by discussing our results' implications for the debate and the prospects of bringing computational models to bear on complex real-world problems that are resistant to simple experimental investigation.


Keywords: pedagogical methods; direct instruction; self exploration; Bayesian teaching; active learning

For centuries, the predominant pedagogical method has been instructional-based teaching such as lecturing. Over the past several decades, constructivism has received increased attention, and educational practitioners have been incorporating constructivist pedagogical methods that emphasize active learning, exploration, and discovery by the learners themselves (Bruner, 1961; Vygotsky, 1978). These methods are typically described as being opposite: the instruction-based method is teacher-centred and passive, while the explorationbased method is learner-centred and active.

Researchers have taken up the debate between instructionbased teaching and active learning, with unsatisfying results. Early cognitive science extolled the virtues of active learning (Bruner, 1961). However, more recently researchers have alternately found evidence for both teaching (Mayer, 2004) and active learning (Sweller, 1988; Gureckis \& Markant, 2012). Researchers have also found instances of equivalence (Klahr \& Nigam, 2004) and nuanced interplay between teaching and active learning when in sequence (Bonawitz et al., 2011; DeCaro \& Rittle-Johnson, 2012). This has led researchers to propose new constructs, such as Guided Play (Weisberg, Hirsh-Pasek, \& Golinkoff, 2013), which have moved forward the debate. However, when and why teaching or active learning may yield better outcomes remain largely unresolved.

One of the greatest barriers to resolving this debate is the difficulty in fully characterizing the pedagogical methods used and precisely controlling the conditions under which they are used (Prince, 2004). We argue that by abstracting away from the particulars of specific learning material (e.g., biology or mathematics) and idealizing the learning process (to be rational and normative), we may use computational
models to formalize the methods and conditions and thus clarify when and why teaching may outperform active learning, and vice versa. Computational models of both teaching and active learning exist in the cognitive science and machine learning literature (e.g., Shafto \& Goodman, 2008; MacKay, 1992). These models have been shown to describe human behavior well in a variety of simple learning settings and higherlevel perceptions (Shafto, Goodman, \& Griffiths, 2014; Castro et al., 2009; Gureckis \& Markant, 2009; Yang, Lengyel, \& Wolpert, 2016). However, the two types of models have never been compared in the same framework.

We provide the first computational comparison between teaching and active learning. Following previous computational work, we use a concept learning task to assess the effectiveness of the two pedagogical methods. To approximate some of the complexity of real-world learning problems, we design the concept space to be hierarchical and introduce partial ambiguity between concepts by introducing overlap between the concepts. As an illustration, the categories "bird" and "sea animals" are partially ambiguous due to the existence of examples that are in both categories. Such ambiguity makes certain examples uninformative for grasping distinctions between the concepts, thus limiting the performance of active learning. In these cases the guidance of a knowledgeable and helpful teacher may be of great import, by avoiding such ambiguous examples. Using varying degrees of ambiguity, we will investigate under what scenario is the effectiveness of teaching more pronounced.

Given perfect knowledge and rational inference, we know that teaching is at least as effective as active learning. This assumes that the teacher is knowledgeable and helpful and that the learner and teacher use the same inference scheme. There is little reason to believe that these assumptions are met in everyday educational settings (Chi, Siler, \& Jeong, 2004). The effect of an unhelpful teacher is easy to imagine (random guidance), but the effects of a teacher with imperfect knowledge of a learner or a teacher operating with incorrect beliefs about the world is less apparent. To investigate this, we introduce a conceptually misaligned teaching model in which we vary the types and degrees of misalignment between the teacher's and learner's concept spaces. Using this, we explore scenarios under which exploration outperforms teaching.

## Framework

The most common and simplest concept learning tasks use concept spaces with a single-layer of Boolean concepts where the features can be discrete or continuous (e.g., Shafto \& Goodman, 2008; Castro et al., 2009; Gureckis \& Markant,
2009). While this is a sensible framework for certain questions, to accentuate differences between the models and ensure that we can model the kinds of challenging concept learning problems that appear in more realistic scenarios, we adopt a more complex concept space with two-layered Boolean concepts, where the features of the higher-level concepts are themselves concepts. In particular, a two-layered concept space allows us to introduce partial ambiguity between the concepts that only teaching, but not exploration, can resolve (explained later in more details with an example). Similarly, this allows us more interesting variations in the ways that teachers may be incorrect about learner's beliefs, or the true state of the world. Below we begin the description of the framework by describing the concept space and setting notations before presenting a detailed example, and the simulations results.

## Concept space

A concept space contains two concepts. A concept contains at least 1 and up to 6 distinct patterns. There are in total 6 types of patterns (Fig. 1 left), and they are all the Boolean concepts that have 4 features with balanced binary labels, that is, two features labeled 0 and two features labeled 1. Figure 1 provides two example concept spaces. Formally, we denote the concept space by $H$, the concepts by $h=\left\{h_{1}, h_{2}\right\}$, the $j^{\text {th }}$ pattern in a concept by $f_{j}$, the features by $x$, and the feature's binary label by $y=\{0,1\}$.

The prior probability on the concept space is hierarchically uniform. This means that $P\left(h_{1}\right)=P\left(h_{2}\right)=\frac{1}{2}$ and that $P\left(f_{j} \mid h_{k}\right)=\frac{1}{N_{k}}$ for all $j$, where $N_{k}$ is the number of patterns in $h_{k}$. We say that the concepts are ambiguous when two concepts have shared patterns, and the degree of ambiguity, $a$, is defined by the number of shared patterns (see Fig. 1 for example). Recall that ambiguous concepts allow a stronger distinction to be made between teaching and active learning.

For the discussion of concept misalignment-where the teacher may be incorrect about the learner's beliefs or the true state of the world-we denote the learner's concept space by $H_{L}$, the teacher's concept space by $H_{T}$, and the true world space by $H_{W}$. We quantify the degree of misalignment, $m$, between two concept spaces by the the minimum number of pattern "moves" within a concept space to make the two concept spaces equivalent. The move operation includes moving a pattern between two concepts, removing a pattern, or adding a pattern. Misalignment with $H_{W}$ is sometimes referred to as misconception. Later we will investigate how the effectiveness of teaching degrades when the learner has misconceptions ( $H_{W}=H_{T} \neq H_{L}$ ) and when the teacher has misconceptions $\left(H_{W}=H_{L} \neq H_{T}\right)$.

## An example trial

First, a concept in $H_{W}$ is chosen as the correct answer; then, a pattern within that concept is chosen as the underlying pattern. On the first trial, a pedagogical method of choice (optimal exploration or teaching) computes scores (according to Eq. 3 or Eq. 6, respectively) for all four potential features.


Figure 1: The six types of patterns, denoted $f$, are used to construct concepts. The positions-top left (TL), top right (TR), bottom left (BL), and bottom right (BR)-of the small squares with blue outline represent the four features. The colors-white or black-represents the binary feature labels which correspond to individual observations. For Concept space A, the prior for each pattern in Concepts 1 and 2 are $P\left(f \mid h_{1}\right) P\left(h_{1}\right)=\frac{1}{5} \times \frac{1}{2}$ and $P\left(f \mid h_{2}\right) P\left(h_{2}\right)=\frac{1}{3} \times \frac{1}{2}$, respectively. The degree of ambiguity, $a$, is 2 because $f_{3}$ and $f_{5}$ in $h_{1}$ are also in $h_{2}$. For Concept space B, $P\left(f \mid h_{1}\right) P\left(h_{1}\right)=\frac{1}{6} \times \frac{1}{2}$, $P\left(f \mid h_{2}\right) P\left(h_{2}\right)=1 \times \frac{1}{2}$, and $a=1$. The degree of misalignment between Concept spaces A and B is 2 : one has to make two "moves" in Concept space A (delete the last pattern in Concept 2 and move the second pattern from Concept 2 to Concept 1) in order to make it equivalent to Concept space B.

The learner queries the world the feature with the highest score, and the world labels the query according to the predetermined underlying pattern. With this observation, the chosen pedagogical method computes the scores for the remaining three features. Then, the learner queries; the world labels; and the process repeats until every feature is queried. Before the first query and after each query, the learner's posterior belief (via Eq. 2) about the correct concept is recorded.

We now give a concrete example that compares optimal exploration with optimal teaching. We first name the features as Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR; see Fig. 1 caption). Figure 2A shows the concept space under consideration, the target concept, and the target underlying pattern. In this case, the teacher's concept space, the learner's concept space, and the world are all aligned ( $H_{W}=H_{T}=H_{L}$ ). Figure 2B shows the scores that exploration and teaching assign to each query. The reasoning behind the scoring is based on predicted outcomes. For example, if TL is queried and labeled black, then one can rule out all patterns in $h_{2}$ and be certain that $h_{1}$ is the answer. This is good. But if TR is queried and labeled black, one can only rule out $f_{2}$ in $h_{1}$ and $f_{1}$ in $h_{2}$, leaving the two concepts equally likely, which is not good. Following this reasoning, the active learner or teacher considers all the possible outcomes (if TL is white...; if TB is black...; if TB is white...; and so on) and chooses the one that best resolves the answer. In this case, before observing anything, both optimal exploration and optimal teaching scores TL or BR the highest. In this trial the learner chooses TL and observes white.

Given this data, $\mathcal{D}=\left\{x_{1}=T L, y_{1}=1\right\}$, the learner rules out $f_{1}$ and $f_{2}$ in $h_{1}$; thus, at this point, the learner believes that $P\left(h_{1} \mid \mathcal{D}\right)=\frac{1}{4}$ and that $P\left(h_{2} \mid \mathcal{D}\right)=\frac{3}{4}$. Optimal exploration chooses a query that reduces the expected uncertainty about arriving at an answer. Intuitively, uncertainty is highest when $h_{1}$ and $h_{2}$ are equally likely and lowest when one is definitely correct. Following the above reasoning, an ac-


Figure 2: An example trial. A. The concept space under consideration. The predetermined underlying pattern is $f_{3}$ in $h_{2}$ (red box), which happen to be equivalent to $f_{3}$ in $h_{1}$. B. Query scorings for optimal exploration according to Eq. 3 (left) and for optimal teaching according to Eqs. 6 (right). The "x" indicates that an observation has been made on that feature, so that feature is excluded from the set of potential queries. The chosen query in each step is highlighted yellow and outlined with a thick border. C-D. Performances of optimal exploration and teaching following the observation sequences given in B. Blue squares indicate that the feature value has not yet been observed.
tive learner computes the expected uncertainty in each predicted case. After summing over all the expected uncertainty weighted by the chance those case would occur, the optimal explorer can then assign a score to each query to indicate how much certainty was gained (or uncertainty is reduced). Here, the sum shows that BR leads to the highest information gain, mainly because it helps reach certainty with a $50 \%$ chance.

Optimal teaching chooses a query to maximize the learner's inference for a desired concept. As such, the query depends on the answers that the teacher has in mind. The reasoning behind optimal teaching is again based on predicted cases and goes as follows: If the real answer is $h_{1}, \mathrm{BR}$ will be white, and TR and BL will be black. Hence, when BR is revealed, the learner will infer that both concepts are equally likely, and when TR or BL is revealed, the learner will consider $h_{2}$ to be more likely. Thus, to help the learner infer the hypothetical answer of $h_{1}$, the teacher will guide the learner to query BR, even though it is, in some sense, ambiguous. Following the same kind of reasoning, the teacher concludes that, for $h_{2}$, TR or BL is the better choice. Now, because the learner knows that the teacher is helpful, the learner can actually infer the answer with certainty just by the teacher's guidance because the guidance is answer-dependent.

This line of reasoning shows that optimal teaching can be better than optimal exploration for two reasons. First, the teacher helps reduce irrelevant search by tailoring guidance based on the answer. This is consistent with theories that support instructional-based teaching (Kirschner, Sweller, \& Clark, 2006). Second, because the guidance depends on teacher's knowledge of the answer, the learner can leverage pedagogical reasoning-the fact that the teacher is knowledgeable and helpful-to make stronger inferences.

Figure 2C-D show the performance with optimal exploration and teaching, respectively. The performance is defined as the learner's posterior belief about the target concept after observing some data (Equation 2 is used for exploration, and Eq. 4 is used for teaching). The performance with optimal exploration eventually reaches chance level because the
underlying pattern, $f_{3}$ in $h_{2}$, is ambiguous in this case. The performance with teaching reaches 1 even for the ambiguous pattern because in optimal teaching the learner can use pedagogical reasoning to break this ambiguity. This type of performance difference via pedagogical reasoning is not possible with single-layered concept space.

## Inference

The learner's inference follows Bayes' rule. Given some data $\mathcal{D}=\left\{x_{i}, y_{i}\right\}_{i=1}^{N}$, the learner's joint posterior is

$$
\begin{align*}
P(h, f \mid \mathcal{D}) & =\frac{P(\mathcal{D} \mid h, f) P(h, f)}{\sum_{k, j} P\left(\mathcal{D} \mid h_{k}, f_{j}\right) P\left(h_{k}, f_{j}\right)}=\frac{1}{Z} P(\mathcal{D} \mid f) P(f \mid h) P(h) \\
& =\frac{1}{Z} \prod_{i} P\left(y_{i} \mid x_{i}, f\right) P(f \mid h) P(h) \tag{1}
\end{align*}
$$

In our framework, labelling is deterministic, so the likelihood $P\left(y_{i} \mid x_{i}, f_{j}\right)$ is either 0 or 1 . The normalizing constant, $Z$, can be computed exactly by enumeration in our simple setting.

The joint posterior of Eq. 1 can be used to obtain the concept posterior,

$$
\begin{equation*}
P(h \mid \mathcal{D})=\sum_{j} P\left(h, f_{j} \mid \mathcal{D}\right), \tag{2}
\end{equation*}
$$

by marginalizing out $f$. It can also be used to obtain the pattern posterior, $P(f \mid \mathcal{D})=\sum_{k} P\left(h_{k}, f \mid \mathcal{D}\right)$, by marginalizing out $h$. This is used for computing the predictive distribution.

## Optimal exploration

We model optimal exploration following a Bayesian active learning model that chooses query $x$ to myopically maximize the expected information gain (MacKay, 1992). The probability of choosing an $x$ is

$$
\begin{equation*}
P(x \mid \mathcal{D})=\lim _{\alpha \rightarrow \infty} \frac{1}{Z}\left[\langle\mathrm{H}[h \mid \mathcal{D}]-\mathrm{H}[h \mid \mathcal{D}, x, y]\rangle_{P(y \mid x, \mathcal{D})}\right]^{\alpha} \tag{3}
\end{equation*}
$$

Here, $Z=\sum_{x^{\prime}}\left[\left\langle\mathrm{H}[h \mid \mathcal{D}]-\mathrm{H}\left[h \mid \mathcal{D}, x^{\prime}, y\right]\right\rangle_{P\left(y \mid x^{\prime}, \mathcal{D}\right)}\right]^{\alpha}$ is the normalizer to produce a probability distribution over $x$, and
$\mathrm{H}[h \mid \cdot]=-\sum_{h} P(h \mid \cdot) \log P(h \mid \cdot)$ is the Shannon entropy. Because $\mathrm{H}[h \mid \cdot]$ is a measure of the uncertainty of the posterior, the difference in entropy before and after receiving a new observation pair $\{x, y\}$ in Eq. 3 quantifies the expected reduction in $h$ uncertainty, which is information gain. The expectation operator, $\langle\cdots\rangle_{P(y \mid x, \mathcal{D})}$, indicates that the learner does not know exactly whether the label for $x^{*}$ will be 0 or 1 but maintains a predictive distribution. The predictive distribution is given by $P(y \mid x, \mathcal{D})=\sum_{j} P\left(y \mid x, f_{j}\right) P\left(f_{j} \mid \mathcal{D}\right)$. The limit $\alpha \rightarrow \infty$ assigns probability uniformly over the $x$ 's that produce the highest argument value. It returns the set of values that have the highest probability when there is more than one; it is equivalent to $\arg \max$ when there is a single highest argument value.

## Optimal teaching

We define optimal teaching to satisfy three assumptions (Shafto \& Goodman, 2008; Shafto et al., 2014). First, the teacher knows the correct answer. Second, $H_{W}=H_{T}=H_{L}$, and the teacher and learner use exactly the same inference scheme. Third, the teacher and the learner are cooperative. From the learner's perspective, this means that the learner reasons about how the teacher, knowing the answer, chooses the most helpful guidance. From the teacher's perspective, this means that the teacher provides guidance while being aware of the learner's inference.

We begin the formulation with the learner's inference. Given the guidance, $x$, from the teacher and the corresponding new observation, $y$, the learner's inference follows

$$
\begin{equation*}
P_{L}(h, f \mid x, y, \mathcal{D})=\frac{P(y \mid x, f) P_{T}(x \mid h, \mathcal{D}) P_{L}(f, h \mid \mathcal{D})}{\sum_{j, k} P\left(y \mid x, f_{j}\right) P_{T}\left(x \mid h_{k}, \mathcal{D}\right) P_{L}\left(f_{j}, h_{k} \mid \mathcal{D}\right)} \tag{4}
\end{equation*}
$$

Note that the teacher's guidance carries information about the answer via the likelihood, $P_{T}(x \mid h, \mathcal{D})$. The cooperative inference mentioned above can be modeled by combining Eq. 4 with

$$
\begin{align*}
& P_{L}(h \mid x, \mathcal{D})=\frac{1}{Z}\left\langle\sum_{j} P_{L}\left(h, f_{j} \mid x, y, \mathcal{D}\right)\right\rangle_{P_{L}(y \mid x, h, \mathcal{D})}  \tag{5}\\
& P_{T}(x \mid h, \mathcal{D})=\lim _{\alpha \rightarrow \infty} \frac{\left[P_{L}(h \mid x, \mathcal{D}) P(x)\right]^{\alpha}}{\sum_{x^{\prime}}\left[P_{L}\left(h \mid x^{\prime}, \mathcal{D}\right) P\left(x^{\prime}\right)\right]^{\alpha}} \tag{6}
\end{align*}
$$

where $P_{L}(y \mid x, h, \mathcal{D})=\sum_{j} P\left(y \mid x, f_{j}\right) P_{L}\left(f_{j}, h \mid \mathcal{D}\right), Z$ is a normalizer, and $P(x)$ is a uniform distribution over $x$. This system of equations (4-6) is first iterated until convergence, then an $x$ is sampled from Eq. 6 conditioned on the teacher's knowledge of the true concept. A sensible initial condition is a uniform $P_{T}(x \mid h, \mathcal{D})$ in Eq. 4. Compared to Eq. 3, the extra $h$ in the expectation operator, $\langle\cdots\rangle_{P_{L}(y \mid x, h, \mathcal{D})}$, shows that teacher and learner reasons about one $h$ at a time. The subscript $L$ in $P_{L}(\cdot)$ emphasizes that the concept-pattern joint prior and posterior are based on the learner's reasoning; the subscript $T$ in $P_{T}(x \mid h, \mathcal{D})$ emphasizes that the $x$ is based on the teacher's reasoning; and the unsubscripted $P(y \mid x, f)$ indicates that the label likelihood is provided by the true world. ${ }^{1}$

[^544]
## Conceptually misaligned teaching

Optimal teaching is always at least as good as optimal exploration because the teacher's guidance offers extra information about the correct answer. But what happens when the assumptions of optimal teaching are violated? We consider two types of violation that breaks the second assumption of the learner and teacher sharing the same inference by introducing misconception in the learner (type 1: $H_{W}=H_{T} \neq H_{L}$ ) and misconception in the teacher (type 2: $H_{W}=H_{L} \neq H_{T}$ ). Note that the first and third assumptions of optimal teaching are still kept. The first assumption poses that regardless of whether the teacher has misconception, the teacher knows the correct concept label. The third assumptions poses that the teacher and learner still reason cooperatively. This leads us to introduce conceptually misaligned teaching, where the two agents, the learner and teacher, reason about each other while wrongly assuming the other agent's concept space. Computationally, the teacher provides $x$ by going through Eqs. 4-6 with $H_{T}$, thinking that the learner also operates with $H_{T}$. Having received $x$, the learner also goes through Eqs. 4-6, while assuming that the teacher also has $H_{L}$ in mind. The first type of violation (misconception in learner) is a common issue in education (Chi et al., 2004), and the second type (misconception in teacher but not in learner) is a natural counterpart simulation to do.

## Simulations: systematic comparison

In the Framework section, we gave a detailed example of how exploration compares to teaching given a particular underlying pattern and concept space. In this section, we compare the performance of the two pedagogical methods in a systematic manner over different classes of concept spaces. This will lead us to address more broadly the condition under which teaching is better than active learning and vice versa. To this end, we consider three scenarios.

For all three scenarios, a concept space with degree of ambiguity, $a$, contains $6+a$ patterns. All 6 patterns in Fig. 1 are used at least once, but no pattern is used more than twice. Figure 1 shows two example concept spaces that satisfy the above criteria. The first scenario assumes that $H_{W}=H_{L}=H_{T}$ and entertains concept spaces of varying degree of ambiguity from $a=0$ to 5 . For a given $a$, the simulation includes all combinations of assignments of $6+a$ patterns to two concepts, with isomorphic concept spaces counted only once. The second scenario assumes $H_{W}=H_{T} \neq H_{L}$ and a fixed $a=1$. We consider all pairings of $H_{T}$ and $H_{L}$ with $a=1$ up to concept-space pair isomorphism, and label each pair with their degree of misalignment, $m$, which can vary from 0 to 6 . The third scenario assumes $H_{W}=H_{L} \neq H_{T}$ and $a=1$. The

[^545]

Figure 3: A. Double-performance plot for the trial in Fig. 2. B. Double-performance plot averaged over concept spaces with differing degree of ambiguity. The numbers indicate the $a$ of the curve, ranging from 0 to 5 in increments of 1 from the rightmost curve to the leftmost curve. C. Double-performance plot for a learner with differing degree of misconception, $m$, as described in the Concept Space section under Framework. The numbers indicate the $m$ of the curves. All concept spaces have $a=1$, so the curve with $m=0$ matches the curve with $a=1$ in A. D. Double-performance plot for a teacher with differing degree of misconception. The numbers again indicate $m$ and increases from the top to bottom in order of the curves' end points.
pairing and labeling are done as in the second scenario.
To visualize the relative performances of exploration and teaching, we plot the two pedagogical methods against each other (Fig. 3A). On this double-performance plot, curves above the diagonal indicates that teaching is better than optimal exploration, and curves below indicates that teaching is worse. To reveal higher-order trends, we average single-trial performances firstly over the patterns in individual concept space, weighted by the patterns' prior probabilities, and secondly over concept spaces that have the same label; that is, the same $a$ or $m$, with each concept space contributing equal weight. Figure 3B, C, and D show the relative performances under the first, second, and third scenario, respectively.

Figure 3B shows that optimal teaching is better than optimal exploration when the concepts are partially ambiguous but no better or worse when the concept spaces are fully unambiguous. When the concept space is fully ambiguous, performances remain at chance level for both teaching and exploration. The advantage of of teaching in absolute performance, as judged by the end points on the double-performance plot, is most pronounced with a $30 \%$ improvement over exploration at a medium degree of ambiguity of $a=3$. Overall, teaching is better than exploration under partial ambiguity because of the reduction in irrelevant search and pedagogical reasoning described in the example task.

Figure 3C shows that on average, teaching and exploration perform similarly when the learner's concept space is wrong. On the one hand, this shows that teaching is robust against learner's misconceptions (in terms of hurting learning) even when the misconception is strong; on the other hand, it shows that the benefits of teaching (in terms of boosting learning) diminishes in the face of a little misconception. The trial-by-trial performances reveal extreme cases when teaching is much worse than exploration ( $0 \%$ vs. $>80 \%$ ). This happens when the learner's concept space looks like Concept space B in Fig. 1, where $h_{1}$ contains all the patterns and $h_{2}$ contains only one pattern which is also in $h_{1}$. This suggests that the learner's strong prior bias for an ambiguous pattern ( $f_{3}$ from $h_{1}$ or $f_{1}$ from $h_{2}$ ) being associated with a particular hypothesis
(i.e., with $h_{2}$ ) can distort the effect of pedagogical reasoning when there is conceptual misalignment.

Figure 3D shows that teaching with the wrong concept space leads to evidently poorer performance (roughly 10-20\% worse) during learning, but only somewhat poorer performance (up to about $10 \%$ worse) at the end of learning (after observing all feature values). The exceptions are when $0<m<3$; then teaching ends up very slightly ( $<5 \%$ ) better. Thus, while the advantage of teaching quickly diminishes with misalignment, the final performance of teaching is rather robust. Interestingly, the first and third scenarios combine to show that exploration before teaching is better than the other way around, as the advantage of teaching comes in after the first observation (Fig. 3C), but its disadvantage due to potential misalignment comes in at the first query (Fig. 3E). This is consistent with previous findings on the interplay between exploration and teaching in learning mathematical concepts (Schwartz \& Martin, 2004; DeCaro \& Rittle-Johnson, 2012).

In summary, our analysis shows that if one knows little about the structures of $H_{W}, H_{T}$, and $H_{L}$ and their alignments, teaching is the preferred pedagogical method because it can potentially be much better and will unlikely be much worse than exploration. If one knows that there is moderate amount of misalignment, exploration is the preferred method. If one knows the detailed structures of $H_{W}, H_{T}$, and $H_{L}$, the alignments among them, and a particular target concept, detailed analysis should be done to choose the better method.

Lastly, it is worth considering our results in the context of popular explanations in favor of active learning. These reasons, including attentional control, enhanced memory, stress relief, and etc. (Springer, Stanne, \& Donovan, 1999; Prince, 2004; Markant, Ruggeri, Gureckis, \& Xu, 2016), all suggest that a learner who explores actively maintains an $H_{L}$ more consistent with $H_{W}$ than a leaner who receives passive teaching guidance. A model that includes pedagogical-methoddependent effect on the concept space is an important direction for future work.

## Discussion

Researchers have debated whether teaching or active learning is better for decades without reaching a consensus. A main barrier is the difficulty in precisely controlling the pedagogical method and learning conditions for meaningful comparison of results that support generalization. We argue that by formalizing the learning conditions and the pedagogical methods, we may clarify when and why which pedagogical method is more effective. We adopt existing models for optimal exploration and optimal teaching and introduce a model for conceptually misaligned teaching. Our computational analysis showed that optimal teaching is much better than exploration when the concepts in play are partially ambiguous, but this effect diminishes very quickly with conceptual misalignment between the learner and teacher.

We expect this trend in our results to generalize to larger concept spaces with richer structures. Optimal teaching should be increasingly more effective than optimal exploration because a larger space allows for greater reduction in irrelevant search and a richer structure allows for finer pedagogical reasoning. However, when there is conceptual misalignment, we expect this advantage of teaching to diminish quickly because the ways of misinterpreting guidance and observation also increases with the complexity of the concept space. The exact scaling between the rate of diminishing benefit and the size and complexity of concept space and is an interesting question for future work.

We have focused on the pedagogical method that best leads the learner to a specific concept. This approach is common to concept learning experiments. However, in many realworld scenarios, we may also be interested in generalization: what does performance on one task predict about future performance on related tasks. To capture this, we would need to consider concept spaces with much richer structure that would support incremental building of compositional concepts and/or transfer learning. Although beyond the scope of the current research, even defining what such conceptual structures should look like in order to capture some of the richness of real-world concept learning problems is another interesting question for future work.

Debates about the relative efficacy of different pedagogical methods have plagued the literature. Because of the complexity of concepts and the variability in the application of the pedagogical methods themselves, empirical tests have been largely inconclusive. Our approach has been to abstract away from some of the details and ask the question in an idealized setting: under what circumstances would we expect teaching or active learning to perform better. Our results yielded the surprising conclusion that, even when the assumptions of teaching are not perfectly met, it is quite robust. While there is more work to be done to capture the richness of psychological theories of active learning, our approach provides a way forward where empirical research has not been as successful as initially hoped. Considerable work remains, but systematic computational analysis of theories themselves provides a po-
tentially promising complement to more traditional empirical methods for uncovering more optimal methods of delivering educational content.

## Acknowledgments

This research was supported in part by NSF CAREER award DRL1149116 to P.S.

## References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., \& Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. Cognition, 120(3), 322-330.
Bruner, J. S. (1961). The act of discovery. Harvard educational review.
Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., \& Zhu, X. (2009). Human active learning. In Advances in neural information processing systems (pp. 241-248).
Chi, M. T., Siler, S. A., \& Jeong, H. (2004). Can tutors monitor students' understanding accurately? Cognition and instruction, 22(3), 363-387.
DeCaro, M. S., \& Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. Journal of experimental child psychology, 113(4), 552-568.
Gureckis, T. M., \& Markant, D. (2009). Active learning strategies in a spatial concept learning game. In Proceedings of the 31st annual conference of the cognitive science society (pp. 3145-3150).
Gureckis, T. M., \& Markant, D. B. (2012). Self-directed learning a cognitive and computational perspective. Perspectives on Psychological Science, 7(5), 464-481.
Kirschner, P. A., Sweller, J., \& Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational psychologist, 41(2), 7586.

Klahr, D., \& Nigam, M. (2004). The equivalence of learning paths in early science instruction effects of direct instruction and discovery learning. Psychological science, 15(10), 661-667.
MacKay, D. J. (1992). Information-based objective functions for active data selection. Neural computation, 4(4), 590-604.
Markant, D. B., Ruggeri, A., Gureckis, T. M., \& Xu, F. (2016). Enhanced memory as a common effect of active learning. Mind, Brain, and Education, 10(3), 142-152.
Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? American psychologist, 59(1), 14.
Prince, M. (2004). Does active learning work? a review of the research. Journal of engineering education, 93(3), 223-231.
Schwartz, D. L., \& Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. Cognition and Instruction, 22(2), 129-184.
Shafto, P., \& Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In Proceedings of the 30th annual conference of the cognitive science society (pp. 1632-1637).
Shafto, P., Goodman, N. D., \& Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive psychology, 71, 55-89.
Springer, L., Stanne, M. E., \& Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. Review of educational research, 69(1), 21-51.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257-285.
Vygotsky, L. (1978). Interaction between learning and development. Readings on the development of children, 23(3), 34-41.
Weisberg, D. S., Hirsh-Pasek, K., \& Golinkoff, R. M. (2013). Guided play: Where curricular goals meet a playful pedagogy. Mind, Brain, and Education, 7(2), 104-112.
Yang, S. C.-H., Lengyel, M., \& Wolpert, D. M. (2016). Active sensing in the categorization of visual patterns. Elife, 5, e12215.

# Training Graph Literacy: Developing the RiskLiteracy.org Outreach Platform 

Vincent T. Ybarra (Vincent.Ybarra@ou.edu)<br>National Institute of Risk and Resilience, 201 Stephenson Parkway, Suite 2300, 5 Partners Place Norman, OK 73019 USA<br>Edward T. Cokely (cokely@ou.edu)<br>National Institute of Risk and Resilience, 201 Stephenson Parkway, Suite 2300, 5 Partners Place Norman, OK 73019 USA<br>Cody Adams (Codyadams@ou.edu)<br>National Institute of Risk and Resilience, 201 Stephenson Parkway, Suite 2300, 5 Partners Place Norman, OK 73019 USA<br>Margo Woller-Carter (Mwoller@mtu.edu)<br>Michigan Technological University, 1400 Townsend Dr. Houghton, MI 49931 USA<br>Jinan N. Allan (jnallan@ou.edu)<br>National Institute of Risk and Resilience, 201 Stephenson Parkway, Suite 2300, 5 Partners Place Norman, OK 73019 USA<br>Adam Feltz (adfeltz@mtu.edu)<br>Michigan Technological University, 1400 Townsend Dr. Houghton, MI 49931 USA<br>Rocio Garcia-Retamero (rretamer@ugr.es)<br>University of Granada, Av. del Hospicio, Granada 18010 Spain


#### Abstract

Visual aids have been found to provide an unusually efficient means of risk communication for diverse and vulnerable individuals facing high-stakes choices (e.g., health, finance, natural hazards). Research indicates the benefits of visual aids follow from scaffolding of cognitive and metacognitive processes that enable independent evaluation and understanding of risk-i.e., risk literacy (see Skilled Decision Theory; Cokely et al.,. 2012; in press). Here, we present a brief review and progress report on the development of an online adaptive graph literacy tutor developed as part of the RiskLiteracy.org decision education platform. We begin with a brief review of theoretical foundations of the current tutor based on graph comprehension theory. Next, we discuss key steps in developing and validating our pseudo-intelligent adaptive tutor with emphasis on cognitive and psychometric item analyses and transfer assessments (i.e., decision-making biases). Finally, we present recent changes in technical implementation of the RiskLiteracy.org platform (i.e., Python based with a NoSQL database) that are designed to facilitate interactive, yet brief ( 5 minute to 3 hour) and easier-to-develop training and risk communication tutors. Discussion focuses on emerging opportunities including cognitive oriented usability analyses that should help promote an effective, enjoyable, and inclusive user experience.


Keywords: Graph literacy, decision making, risk literacy, intelligent tutors, risk communication, brain training, numeracy

## Introduction

Well-informed, skilled decision making is associated with a wide-range of socially and economically valuable decision making outcomes (e.g., health, wealth, happiness; for a review see Skilled Decision Theory, Cokely et al., in press). In part, the benefits of general decision making skill, as measured by tests like the Berlin Numeracy Test, result because skilled decision makers tend to be better prepared to independently evaluate and understand risk as presented in common risk communications (e.g., information about health, finance, natural hazards; RiskLiteracy.org) (Cokely et al., 2012). Unfortunately, individuals with lower skill levels, including many at-risk individuals, are routinely biased by standard and well-intentioned risk communication practices, which can result in dangerous decision errors (e.g., ignoring a heart attack; Petrova et al., 2016).

To help address limitations of current risk communication practices, recent scientific efforts have endeavored to develop more inclusive decision education technologies and outreach platforms (e.g., adaptive decision support and training). For
example, simple, transparent visual aids have been found to dramatically enhance risk literacy and independent decision making, conferring major benefits to diverse decision makers who vary widely in ages, backgrounds, abilities, cultures, and values (Garcia-Retamero \& Cokely, 2013). Consider a recent systematic review by Garcia-Retamero and Cokely (2017) spanning dozens of experiments involving more than 25,000 participants from 60 countries. This work specifically mapped informed, skilled decision-making and how it interacts with graph literacy and visual aids, presenting insights on (a) visual aid effectiveness, (b) heuristics for construction and evaluation of user-friendly visual aids, and (c) the relatively large and robust benefits of visual aids for diverse individuals. While the review documented remarkably large benefits of visual aids for "real world" decision making in general, the review also identified some significant problems, namely: 1) Despite the successes of well-designed visual aids, some at-risk users lack basic graph evaluation and interpretation skills and are not graph literate enough to benefit from effective risk communications (Galesic \& Garcia-Retamero, 2011) and 2) Given conflicts of interests and other factors, it can be hard to get risk communicators to adhere to best design practices (e.g., distorted visual aids can shape attitudes and perceptions without violating truth in advertising regulations, etc.).

In what follows, we present an overview of details, successes, and obstacles in our ongoing efforts to develop a brief and adaptive computerized training programs using the RiskLiteracy.org platform. Focus will be on the development of our Graph Literacy Tutor (Cokely et al., in press; WollerCarter, 2015). A growing body of evidence has documented that substantial, decision-relevant benefits tend to emerge in a relatively short amount of time. Recent advances in the platform also enable more rapid and robust development of pseudo-intelligent (adaptive, but not fully intelligent) interfaces that reduce the costs and time required for development of brief interactive training and risk communication programs (Koedinger \& Corbett, 2006). Accordingly, we begin by reviewing our formal cognitive science based on graph comprehension research and Skilled Decision Theory. We then discuss the development and testing of specific graph literacy modules and assessments and we present results from a recent control trial study documenting near and (relatively) far transfer (i.e., graph literacy training improved graph literacy skills, but also improved text based decision skills including resistance to framing and reference class neglect). Next, we consider advances in platform design and implementation, including efforts to integrate psychometric approaches to circumvent the need for more extensive intelligent tutor engines. We close with brief discussion of future directions, limitations, and ongoing projects focusing on user experience optimization.

## Cognitive Processes in Graph Comprehension

Theoretically, the design of an efficient graph literacy training program will depend on the accuracy of our
understanding of the underlying essential (causal) cognitive processes. As such, we drew from the well-established body of empirical literature on graph comprehension to provide a foundation for our tutor development. Graph comprehension models generally indicate that when an individual views a graph they engage in three processes: 1) encoding of the visual pattern, 2) translation of the identified visual features into conceptual relations, and 3) the selection of referents for the identified concepts (Bertin, 1983; Carpenter \& Shah, 1998; Cleveland \& McGill, 1986; Kosslyn, 1989; Lohse, 1993; Okan, Galesic, \& Garcia-Retamero, 2016; Pinker, 1990; Simkin \& Hastie, 1987; Shah \& Carpenter, 1995). Together these processes allow for individuals to make a piece-wise interpretation of graphs before fully integrating the underlying mental model required for inductive and deductive inferences (e.g., reasoning that goes beyond givens). Theoretically, each step of this evaluation process involves essential processes and judgments that an individual must accurately make to correctly interpret the visualized data.

Broadly, it is often assumed that graph comprehension focuses on encoding the visual pattern, which requires the identification of key features of the graph (e.g., attending to many bars of varying height in a bar chart). Once key features are identified, a relative visual judgement is made to determine relative shape of the graph (e.g., positions of the graph elements within the axis, size, and length of the elements within the graph, the slope or angle of graph items).

Translating the identified visual features into conceptual relations then assigns relative quantitative meaning to the features of the graph. The comparison of size and spatial relations between graph features (e.g., a line graph with one positive and one negatively slopped line). For example, tall bars on a traditional bar graph would be interpreted as "more" compared to short bars. There is reason to believe that the spatial-to-conceptual mappings (e.g., "higher equals more," "steeper equals faster") found with graphs are analogous to ecological heuristics that persist within both adults and children with zero graphing experience (Gattis, 2002; Gattis \& Holyoak, 1996).

Theoretically, the final step in graph comprehension is determining the referents of the concepts identified. Here, one must accurately identify the associations of variables within the graph with numerical values. This is where the conventional features of the graph (title, axes labels, legends, and numerical values) are added into the mental representation of the whole graph. For instance, one must identify the context that the graph represents or the scale at which the y-axis is set before an inference can be made. This process seems to be closer to a skill that is not analogous to real-world conventions (Okan, Garcia-Retamero, Galesic, \& Cokely, 2012). This assumes that the skills needed to create proper schema for the conventional elements are trainable.

## Skilled Graph Comprehension

The idea that reading a graph is trainable is embodied in many theories. One holds that graph schema will be formed in
long-term memory (Maichle, 1994; Peebles \& Cheng, 2001; Pinker, 1990; Ratwani \& Trafton, 2008). Training graph literacy should then aim to increase the available schemas and enhance the already present ones, aiding in the identification of the conventional graph features and improving inferences made. Specifically, the training of skills would be aimed at the increasing of knowledge content, and thus can be relatively independent of limited working memory or visuospatial abilities (Hegarty \& Waller, 2005; Shah et al., 2005). Previous research has found that expertise in a specific domain can increase associations between visual patterns and concepts (Tabachneck-Schijf, Leonardo, \& Simon, 1997) and that inferences become easier to make (Roth \& Bown, 2003). One strategy to train is the use of online adaptive tutors (Anderson, Corbett, \& Koedinger, 1995; Koedinger \& Corbett, 2006; Lovett, Meyer, \& Thille, 2008).

The benefits of tutors are often attributed to factors such as: reduction of cognitive load during learning via worked examples, faster (ideally immediate) performance feedback, easier to understand instructions, frequent and more precise diagnostic tests of knowledge, consistent and direct modes of delivering material, and greater opportunities for detection and self-correction of errors during learning (Corbett \& Anderson, 1991; Koedinger \& Aleven, 2007; Mathan \& Koedinger, 2005; Roediger \& Karnicke, 2006; Sweller, Van Merrieoboer, \& Pass, 1998).

Validated adaptive tutors are currently available for many topics in math, statistics, reading, and physics. However, despite the ubiquitous nature of visual aids in risk communications, there are few validated computerized graph tutors available. However, the available graph tutors are generally designed for specialized, narrow audiences (e.g., geared toward younger high school students). Among the few graph literacy training programs that have been specifically designed for diverse adults, none have been subject to evidence-based validity studies providing estimates of: 1) the efficacy of graph literacy training for various users, 2) the magnitude of associated benefits for naturalistic decision making (e.g., interpreting real high-stakes risk communications about health and natural hazards), and 3) essential usability and user experience outcomes, strengths, and weaknesses.

## RiskLiteracy.org Graph Tutor Methods

Woller-Carter (2015) created an online graph tutor for RiskLiteracy.org that trained participants on the foundations of graph literacy and the application of graphs to everyday risky situations. The graph tutor was the prototype that which the new online graph tutor was created upon. The goal of the graph tutor is to briefly and efficiently train adult learners in essential selection, design, and display of graphs that are common in risk communications and related decision education programs. Broadly, the graph tutor contains two major components. The first consists of graph selection tasks where participants choose the correct graph that (by current standards) is best-suited for depicting specific types of data. The second major training component is the graph design
task, which requires participants to identify the necessary information from data to create a graph that accords with best practices by selecting from four candidate graphs.

Note that all graph selection tasks were chosen from an initial study where 217 participants completed multiple graph selection problems, the Berlin Numeracy Test, and the Graph Literacy Assessment (Cokely, Galesic, Schulz, \& GarciaRetaneri, 2012; Galesic \& Garcia-Retamero, 2011). An item analysis based on Classical Testing Theory was conducted to parse out item difficulty and discriminability, in accordance with Formal Item Response Theory approaches. The same procedure was followed for the graph design task. In total, 862 participants completed a random sample of 10-11 graph design problems (e.g., Approximately 100 participants/problems). Analyses provided a detailed account of the relevant psychometric properties of all task items, facilitating a theoretical optimization of problem type across the underlying skill dimensions (e.g., precisely selected items that were most representative and unbiased problems spanning the difficulty range; see Woller-Carter, 2015).

## Control Trial Results

Woller-Carter (2015) found large pre-test, post-test differences in graph literacy that remained significant even after controlling for initial levels of graph literacy $(t(89)=$ 5.23, $p \leq .001, d=1.10$ ) after participants completed the graph tutor. Interestingly, beyond general competency in graph literacy, compared to a control group that completed a STEM Foundations study skills training, graph literacy training also significantly improved some general decision making skills for decision tasks that did not otherwise include any visual aid or graphical content $(\mathrm{F}(3,87)=10.08, p \leq .001$, $\mathrm{R}^{2}=.033, d=1.30$ ). Findings are consistent with Skilled Decision Theory and theoretical accounts of risk literacy (Cokey et al., in press). Partial mediation between condition and decision task performance indicated that improvements in graph literacy directly mediated observed improvements in general decision making skills (e.g., learning how to represent data in a graph also helps people represent decisionrelevant data in useful ways).

## Additional Decision-Making Items

To further explore how risky decision-making interacts with graph literacy the creation of sensitive measurement tools is needed. The results of the 2015 tutor indicated what type of tools may be necessary. Training graph literacy aided in general decision-making skills that focused on "visualizable" risk situations (e.g., sunk cost). For instance, if someone is confronted with a risky decision and the aid of a graph would increase their ability to decide (e.g., icon arrays and safe sex practices), then graph literacy training will help (Cokely \& Garcia-Retamero, 2015). Our lab took the firsts steps by conducting a battery of previous, validated bias questions. Then we used Formal Item Response Theory to analyze the problems for difficulty and discriminability. The battery focused on three different biases: Sunk cost, reference class
neglect, and framing. Fifty-three University of Oklahoma students completed 90 ( 30 each bias) decision tasks taken from various sources.

The finished product are six psychometrically sensitive questions for each of the three types of bias. Results of the analysis are seen below in Figure 1.

## Sunk Cost



## Reference Class Neglect



## Framing



Figure 1. Visual representations of the Formal Item Response Theory analysis for the three biases in our pilot measurement study: Sunk cost (top), reference class neglect (middle), and framing (bottom).

## Construction of Python Graph Literacy Tutor

To better create a tutor development platform that meets the needs of brief risk interventions the decision was made to transfer the initial graph tutor from a Flash based platform built in Carnagie Mellon University's Cognitive Tutor Authoring Tools (CTAT) to an independent Python built, Flask tutor. For an example of the original graph tutor see figure 2. A large and growing body of research has made CTAT a quintessential and evolving intelligent tool for large scale tutors. Despite many advantages, there are many potentially valuable applications for (pseudo)intelligent and adaptive tutors which may be narrow. For example, many general-use decision support systems or decision aids made for risk communication may only require between 10 and 120
minutes to complete (e.g., mortgage or surgical risk disclosure; medical treatment risk information). There is currently no well-established solution, like CTAT, for the creation of small scale, brief, scientifically validated interventions. To fill this gap, following a survey of the available literature, we developed a "proof of concept" application in Python, following best practices based on CTAT and related efforts (Aleven, Mclaren, Sewall, \& Koedinger, 2009; Anderson, Corbett, Koedinger, \& Pelletier, 1995; Walker, Koedinger, Mclaren, \& Rummel, 2006). Specifically, we implemented the Risk Literacy Graph Tutor platform in Python, to assess viability and trade-offs, as compared to standard approaches implemented in Adobe Flash.


Figure 2. The original graph literacy tutor programmed in CTAT. The tutor was hosted on moodle.com, an open source Learning Management System.

The new graph tutor was built from the ground up with Flask, a micro web framework written in Python. Some notable advantages of Flask include:

- Flask is not tied to specific libraries or tools allowing flexible design of the graph tutor to better suit immediate needs (e.g., database connectivity, form validation, etc.).
- Flask is lightweight (no object-relational mapping, simple routing, and easy set-up) reducing the system requirements and development time.
- Flask is documented and community adopted, reducing the learning curve of implementing new solutions.

Beyond several notable benefits of re-development of the graph tutor, there are also some notable costs. First, the initial tutor programmed in Flash proved problematic with the number of online interfaces pivoting from the platform and potentially requiring extra authentication. These issues were persistent enough that, ultimately, an entirely new web template (e.g., User interface) had to be created. Second, the Learning Management System (LMS) platform (e.g., Moodle or Blackboard) had to be entirely abandoned to better accommodate easier implementation for experimentation, which could prove problematic whenever researchers want to track large numbers of specific users over extended periods
of time (e.g., months). Third, creating online tools that use the Intelligent Tutoring System (ITS) model.

Creating an authorizing adaptive tutor in Python also required development of additional infrastructure components. First, we developed APIs ("Application Programming Interfaces") for user profiles and tutor validation, database connectivity, and a user interface. More specifically, the previous graph tutor used a LMS to track students, which needed a student to enroll in the created class and be approved by the class admin. Now, users can create their own profile that is encrypted and inserted into a database. This design allows for the sharing of the tutor to participants for experimental purposes or for a casual user to independently take the tutor.

Backend and platform development also employed MongoDB, a NoSQL database, to power our tutor application. As a NoSQL database, MongoDB records are structured like that of Python dictionary objects. This feature was emphasized in our selection of a database management system (DBMS) as it relied on existing knowledge of Python, reducing the prerequisites to contribute new features for the graph tutor in the future. There are many technical differences between MongoDB and other DBMS; however, the nuances of SQL versus NoSQL, or variations of DBMS within the NoSQL categories generally seem practically irrelevant for (most) projects of similar size and scope. Finally, authorizing via Java Script allows for the immediate feedback essential to worked-example tutors, which proved essential given the theoretical and practical importance of immediate user feedback during training.

## Conclusions

Graphs are ubiquitous across modern media and risk communications. For many people, graphs simplify and clarify important information about risk, which is essential for informed decision making. In this paper we presented a brief overview of progress and ongoing efforts aimed at developing inclusive decision education programs designed to efficiently improve fundamental adult graph literacy and decision making skills. These efforts represent a significant extension to the RiskLiteracy.org platform, which has been accessed by more than 50,000 people from 166 countries since 2012. The mission of this multinational collaborative effort is to advance the science for informed decision making, with support of a network of scientists who provide validated educational resources such as research instruments (e.g., Berlin Numeracy Test) and inclusive decision education programs (e.g., the Graph Literacy Tutor). Beyond increasing the availability of skilled decision making resources, the current review also provides an overview of the first proof-of-concept for the Python-based (simplified) extension of the RiskLiteracy.org platform. Although this new approach may streamline development of related dynamic risk communications and training programs, several pressing issues remain. For example, we currently have a need for greater integration of iterative (life-cycle) approaches to user-experience and usability optimization.

There is also a need to further investigate the robustness of and longitudinal stability of training effects across diverse participants and naturalistic decision tasks.

In closing, it is useful to note that most consumers should not expect to gain any general cognitive benefit from commercially available products designed to train general cognitive capacities. While this may seem problematic for us given our stated goals, our approach is actually quite different. Our goal is not to train basic abilities or capacities. Instead, we are focused on complex types of cognitive skills that must be acquired through deliberate practice and training (Cokely et al., in press), with an emphasis on acquired skills that are known to be valuable for everyday and high-stakes naturalistic decision making (e.g., numeracy, risk literacy, graph literacy). Accordingly, it should not be surprising that our basic skill tutor results indicate near and far(ish) transfer to applications beyond the specific training context (e.g., learning how bar graphs can be used to deceive in general may help people navigate complicated graphs in political, financial, or health contexts). Just as skilled reading comprehension is a valuable component of many everyday activities, the ability to evaluate and understand risk is also widely-applicable. To the extent our control trial results generalize, we should expect that there are likely many currently under-appreciated opportunities to develop and apply pseudo-intelligent tutoring programs to great effect.

## References

Allan, J.N., Ripberger, J., Ybarra, V.T., Cokely, E.T. (in press) Predicting Decision Resilience: A Model of Tornado Myth Vulnerability
Aleven, V., Mclaren, B. M., Sewall, J., \& Koedinger, K. R. (2006). The Cognitive Tutor Authoring Tools (CTAT): Preliminary Evaluation of Efficiency Gains. Intelligent Tutoring Systems Lecture Notes in Computer Science, 61-70.
Aleven, V., Mclaren, B., Sewall, J., \& Koedinger, K. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. International Journal of Artificial Intelligence in Education, 19(2), 105-154.
Anderson, J., Boyle, C. F., \& Reiser, B. J. (1985). Intelligent Tutoring Systems. Science, 228, 456-461.
Anderson, J. R., Corbett, A. T., Koedinger, K. R., \& Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. Journal of the Learning Sciences,4(2), 167-207.
Bertin, J. (2011). Semiology of graphics: diagrams, networks, maps. Redlands, CA: ESRI Press.
Carpenter, P. A., \& Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. Journal of Experimental Psychology: Applied,4(2), 75-100.
Cleveland, W. S., \& McGill, R. (1986). An experiment in graphical perception. International Journal of ManMachine Studies, 25, 491-500.
Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., \& GarciaRetamero, R. (2012). Measuring risk literacy: The Berlin

Numeracy Test. Judgment and Decision Making. 7(1). 25-47
Cokely, E.T., Feltz, A., Ghazal, S., Allan, J.N., Petrova, D., \& Garcia-Retamero, R., (in press). Decision making skill: From intelligence to numeracy and expertise. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, \& A. M. Williams ( $2^{\text {nd }}$ Eds.), Cambridge Handbook of Expertise and Expert Performance. New York, NY: Cambridge University Press.
Corbett, A., \& Anderson, J. R. (1991). Feedback control and learning to program with the CMU LISP tutor. AERA Annual Meeting.
Galesic, M., \& Garcia-Retamero, R. (2011). Graph Literacy: A Cross-Cultural Comparison. Medical Decision Making,31(3), 444-457.
Garcia-Retamero, R. \& Cokely, E. T. (2013). Communicating Health Risks with Visual Aids. PsycEXTRA Dataset.
Garcia-Retamero, R. \& Cokely, E. T. (2017). Designing visual aids that promote risl literacy: A systematic review of health research and evidence-based design heuristics. Human Factors.
Gattis, M. (2002). Structure mapping in spatial reasoning. Cognitive Development, 17, 1157-1183.
Gattis, M., \& Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 231-239.
Hegarty, M., \& Waller, D. (2005). Individual differences in spatial abilities. In P. Shah \& A. Miyake (Eds.), The Cambridge handbook of visuospatial thinking (pp. 121169). New York, NY: Cambridge University Press.

Koedinger, K. R., \& Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. Educational Psychology Review, 19(3), 239-264.
Koedinger, K. R., \& Corbett, A. (2006). Cognitive Tutor . The Cambridge handbook of the learning sciences, 135.
Kosslyn, S. M. (1989). Understanding charts and graphs. Applied Cognitive Psychology, 3(3), 185-225.
Lohse, G. L. (1993). A Cognitive Model for Understanding Graphical Perception. Human-Computer Interaction,8(4), 353-388.
Lovett, M., Meyer, O., \& Thille, C. (2008). JIME - The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning. Journal of Interactive Media in Education,2008(1), 13.
Maichle, U. (1994). Cognitive processes in understanding line graphs. In W. Schnotz \& R. W. Kulhavy (Eds.), Comprehension of graphics (pp. 207-227). Amsterdam, The Netherlands: North-Holland Elsevier Science.
Mathan, S. A., \& Koedinger, K. R. (2005). Fostering the Intelligent Novice: Learning From Errors With Metacognitive Tutoring. Educational Psychologist,40(4), 257-265.
Nelson, D. E., Hesse, B. W., \& Croyle, R. T. (2009). Making data talk: communicating public health data to the
public, policy makers, and the press. Oxford: Oxford University Press.
Okan, Y., Woller-Carter, M., Simon, S., Russell, K., Ghazal, S., Parpart, P., . . . Cokely, E. T. (2013). Overcoming distortions in political and health communication: mechanisms of graph literacy. PsycEXTRA Dataset.
Peebles, D., \& Cheng, P. C. H. (2001). Graph-based reasoning: From task analysis to cognitive explanation. In J. D. Moore \& K. Stenning (Eds.), Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society (pp. 762-767). Mahwah, NJ: Erlbaum.
Petrova, D., Garcia-Retamero, R., Catena, A., Cokely, E. T., Heredia Carrasco, A., Arrebola Moreno, A., \& Ramírez Hernández, J. A. (2016). Numeracy predicts risk of PreHospital decision delay: A retrospective study of acute coronary syndrome survival. Annals of Behavioral Medicine.
Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), Artificial intelligence and the future of testing (pp. 73-126). Hillsdale, NJ: Erlbaum.
Ratwani, R. M., \& Trafton, J. G. (2008). Shedding light on the graph schema: Perceptual features versus invariant structure. Psychonomic Bulletin \& Review, 15, 757762.

Roediger, H. L., \& Karpicke, J. D. (2006). The Power of Testing Memory . Perspectives on Psychological Science, 1(3), 181-210. d
Shah, P., Freedman, E. G., \& Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah \& A. Miyake (Eds.), The Cambridge handbook of visuospatial thinking (pp. 426-476). New York, NY: Cambridge University Press.
Simkin, D., \& Hastie, R. (1987). An information-processing analysis of graph perception. Journal of the American Statistical Association, 82, 454-465.
Sweller, J., Van Merrienboer, J. J., \& Pass, F. G. (1998). Cognitive Architecture and Instructional Design. Educational Psychology Review,10(3).
Walker, E., Koedinger, K., Mclaren, B., \& Rummel, N. (2006). Cognitive Tutors as Research Platforms: Extending an Established Tutoring System for Collaborative and Metacognitive Experimentation. Intelligent Tutoring Systems Lecture Notes in Computer Science, 207-216.
Woller-Carter, M. M., Okan, Y., Cokely, E. T., \& GarciaRetamero, R. (2012). Communicating and Distorting Risks with Graphs: An Eye-Tracking Study. PsycEXTRA Dataset
Woller-Carter, M. (2015). Development of the intelligent graphs for everyday risky decisions tutor. Open Access Dissertation, Michigan Technological University. Retrieved from http://digitalcommons.mtu.edu/etdr/59/

# Empirical constraints on computational level models of interference effects in human probabilistic judgements. 

James M. Yearsley (james.m.yearsley@ vanderbilt.edu)<br>Department of Psychology, Vanderbilt University, Nashville TN, USA<br>Emmanuel M. Pothos (emmanuel.pothos.1@city.ac.uk)

Albert Barqué-Duran (albert.barque-duran@city.ac.uk)<br>Department of Psychology, City University London, UK


#### Abstract

Decades of research in decision making have established that there are some situations where human judgments cannot be modelled according to classical probability theory. Yet if we abandon classical (Bayesian) probability theory as an overarching framework for constructing cognitive models, what do we replace it with? In this contribution we outline a way to divide the space of possible computational level models of probabilistic judgment into a hierarchy of levels of increasing complexity, with classical Bayesian probability models occupying the lowest level. Each level has a unique experimental signature, and we examine which level is best able to describe human behavior in a particular probabilistic reasoning task.


Keywords: Probabilistic reasoning, disjunction fallacy, quantum theory.

## Introduction

2017 marks 35 years since the publication of the volume "Judgment under Uncertainty: Heuristics and Biases" edited by Kahneman, Slovic and Tversky (1982), which has since become a seminal text for those interested in violations of normative reasoning. The years since have seen a great deal of research aimed at better understanding the conditions under which human decision makers do or do not make 'normative' decisions, that is, decisions that can be thought of as being 'correct' by some measure. Note that in this contribution, in common with many other studies, we take the definition of 'normative' to be essentially equivalent to compatibility with classical/Bayesian probability theory. That is, behaving in a normative way is less about providing judgments that are 'accurate' in the sense of reflecting real likelihoods, and rather about self consistency (Oaksford \& Chater, 2009).

Despite the large amount of research that has been done to understand when and why these deviations from normative prescription occur, there are some notable gaps in our understanding of the structure of non-normative judgments. In part this has been caused by a tendency amongst researchers to define non-normative judgments by the properties they do not possess, ie a simple relation to an underlying probability distribution representing a belief state. This has led to a peculiar splintering of the study of decision making, most visible in the heuristics programme, where different decision making tasks are assumed to be executed with the aid of different heuristics (eg Gigerenzer et al, 2015). Few have considered the relationships between heuristics, or whether they reflect some deeper structure.

A different approach, advocated by some adherents of the Bayesian cognition program (see Jones \& Love, 2011), has been to deal with violations of normative (Bayesian) prescription by including extra variables and relations that, while they may be framed in a Bayesian way, conflict somewhat with the spirit of the Bayesian approach. An example is provided by Bayesian efforts to deal with order effects in probabilistic inference, where, for example, decision makers may judge $p(E \mid A, B) \neq p(E \mid B, A)$. One possible solution is to posit that decision makers are sensitive also to the order in which evidence is presented, and that the probabilities for judgments therefore should also be conditioned on the order of presentation. Formally this saves face for the Bayesian approach, but it is hard to see how this does anything other than redescribe the problem. A genuine solution to the problem of order effects would also need to explain why decision makers possess this sensitivity, and make testable predictions.

What we want to do in this contribution is approach the problem from a different angle. Suppose we assume that there is some sort of computational structure underlying nonnormative behaviour, presumably that can be understood as a generalisation of the Bayesian approach, and we ask whether we can constrain the structure of this theory in some way. Specifically, can we derive predictions that are sensitive only to some general fact about the structure of the theory and not to the precise details? If we can do this we can perhaps sketch out a space of acceptable computational level theories which can then be investigated in more detail by other means.

This approach is inspired in part by recent work in the field of quantum cognition (Pothos \& Buseyemer, 2013), which is the attempt to use the formal probability theory derived from quantum physics to describe human decision making. The advantage of such an approach, so proponents claim, is that quantum probability theory is a formal, all encompassing framework that is nevertheless able to account for deviations from classical, Bayesian, behaviour. However quantum probability theory has a very particular structure (see Busemeyer \& Bruza 2012), and it is not obvious that this is either necessary or sufficient to explain human behaviour. The study reported here can therefore be thought of as an attempt to test the sufficiency of the quantum approach for explaining certain types of non-normative behaviours. However it is also much more general, since we will see that quantum theory is just one example of a specific class of models sharing the
same structural properties. In order to avoid getting bogged down in arguments about the suitability of quantum theory to describe human decision making, we will avoid making reference to specific computational models beyond classical probability theory.

The rest of this contribution is structured as follows: In Section 2 we explain how we can situate different computational level models of decision making in a hierarchy, and we show that classical and quantum probability theories belong to two different levels, each with a unique experimental signature. In Section 3 we describe an experiment designed to test which level in the hierarchy is necessary and sufficient to explain human behaviour in a particular probabilistic reasoning task. In Section 4 we conclude and outline some future directions.

## A Hierarchy of Formal Theories

Given two disjoint events, $A$ and $B$, a key property of classical probability theory is that the probability of the disjunction, $A \cup B$ is equal to the sum of the probabilities of the individual events. This can be generalised for an arbitrary number of disjoint events as,

$$
\begin{equation*}
p\left(\cup_{i} A_{i}\right)=\sum_{i} p\left(A_{i}\right) \tag{1}
\end{equation*}
$$

This result expresses the property of classical probability known as linearity (more formally $\sigma$-additivity, Kolmogorov, 1933/1950). Classical probability models are therefore examples of linear models mapping events to probabilities.

In order to allow for the possibility of response errors it is useful to generalise this notion somewhat. Suppose we consider a more general class of models where, $p(A)=f(A)+\varepsilon$, where $\varepsilon$ is a constant. We will call these models linear if,

$$
\begin{equation*}
f\left(\cup_{i} A_{i}\right)=\sum_{i} f\left(A_{i}\right) \tag{2}
\end{equation*}
$$

Classical probability theory is obviously a special case of these linear models. These models can be used to capture the idea that judgments are noisy, so for example the probability assigned to the null event is not 0 . They can also capture simple types or response biases, for example an aversion to using the endpoints on a response scale.

It is important to note that general linear models can violate some properties of classical probability theory. For example for two disjoint events $A, B$ we have,
$p(A \cup B)=f(A \cup B)+\varepsilon=f(A)+f(B)+\varepsilon=p(A)+p(B)-\varepsilon$
Which is a violation of the classical law for disjoint events. However it is easy to see that any law of classical probability that is violated by a general linear model will be violated by some multiple of $\varepsilon$, and so these sorts of effects are easy to spot. In particular, consider the quantity,

$$
\begin{equation*}
\mathscr{I}_{2}(A, B)=p(A \cup B)-p(A)-p(B)=-\varepsilon \tag{4}
\end{equation*}
$$

We call expressions of this form 'interference' terms, because of the analogy with quantum models. In a simple classical probability theory model this quantity is 0 . In a general linear model this term is no longer zero, but it is instead equal to some constant which should be the same for all events $A, B$. The way to test the general class of linear models is therefore to examine different sets of events and test whether $\mathscr{I}_{2}(A, B)$ depends on the alternatives under consideration. This generalises the notion of the failure of a simple classical model.

The occurrence of widespread violations of classical probability rules in decision making experiments, including conjunction and disjunction fallacies (Tversky \& Kahneman, 1983), has lead many to reject linear models as accounts of the way human decision makers assign probabilities to events. However few have explored proposals for concrete models beyond linear ones.

A natural generalisation of linear models is to consider those which contain a bilinear term. The appropriate generalisation is to have,

$$
\begin{equation*}
p(A)=g(A, A)+f(A)+\varepsilon \tag{5}
\end{equation*}
$$

where $f(\cdot)$ is linear as above, and $g(\cdot, \cdot)$ is linear in both its arguments,

$$
\begin{equation*}
g\left(\cup_{i} A_{i}, B\right)=\sum_{i} g\left(A_{i}, B\right), \quad g\left(A, \cup_{i} B_{i}\right)=\sum_{i} g\left(A, B_{i}\right) \tag{6}
\end{equation*}
$$

If we now consider the probability assigned to the disjunction $A \cup B$ we see,

$$
\begin{align*}
p(A \cup B)= & g(A \cup B, A \cup B)+f(A \cup B)+\varepsilon \\
= & g(A, A)+g(B, B)+g(A, B)+g(B, A)  \tag{7}\\
& +f(A)+f(B)+\varepsilon \\
= & p(A)+p(B)+g(A, B)+g(B, A)-\varepsilon
\end{align*}
$$

Considering again the quantity $\mathscr{I}_{2}(A, B)$ we see,

$$
\begin{align*}
\mathscr{I}_{2}(A, B) & =p(A \cup B)-p(A)-p(B) \\
& =g(A, B)+g(B, A)-\varepsilon \tag{8}
\end{align*}
$$

which is generally non-zero, but also crucially will depend on $A$ and $B$, allowing us to distinguish bilinear models from linear ones.

Although we will not prove this here, the class of bilinear models includes quantum theory as a special case (Sorkin, 1994). However clearly this framework is more general than any specific model.

Once we have made the choice to step beyond linear models, a whole world of possibilities is opened up. Why stop at bilinear models? Why not consider a model containing a trilinear a function $h(A, B, C)$ ? The answer is that such models are possible, and in the same way that the quantity Eq.(8) provides us with a way to test between linear and bilinear models, a generalisation of this lets us distinguish between bilinear models and more complicated theories.

This rule comes from exploring what happens when we have three possible disjoint events $A, B, C$. Because $g(A, B)$ is bilinear it is straightforward, if tedious, to show,

$$
\begin{align*}
p(A \cup B \cup C)= & p(A \cup B)+p(B \cup C)+p(A \cup C)  \tag{9}\\
& -p(A)-p(B)-p(C)+\varepsilon
\end{align*}
$$

So if we define an analogue of $\mathscr{I}_{2}(A, B)$ for three events, $\mathscr{I}_{3}(A, B, C)$, then we have,

$$
\begin{align*}
\mathscr{I}_{3}(A, B, C)= & p(A \cup B \cup C)-p(A \cup B)-p(B \cup C)-p(A \cup C) \\
& +p(A)+p(B)+p(C) \\
& =\varepsilon \tag{10}
\end{align*}
$$

so this three way interference term is constant for all events $A, B, C$ if the underlying model is bilinear. This does not hold in higher order theories, such as ones based on a trilinear function (Sorkin, 1994). Therefore this provides a test of bilinear models versus more complex theories. Note also that this relation holds trivially for a linear model.

Although we will not show it here, it is straightforward to prove by induction that at every level in this hierarchy of possible model types there is a corresponding interference term $\mathscr{I}_{n}(A, B, C \ldots)$ which is constant for any model at that level or below (Sorkin, 1994). In other words we can make a very general statement; The experimental signature of a level $n$ model (ie a model based on a ' $n$-linear' function) is that a) The quantity $\mathscr{I}_{n}(A, B, C \ldots)$ depends on the events $A, B, \ldots$, but b) the quantity $\mathscr{I}_{n+1}(A, B, C \ldots)$ is constant.

We know from previous work that any theory that can capture human behavior, in particular disjunction fallacies, must be at least level two. The question is whether a level two theory is also sufficient, that is, whether there are particular effects in human decision making that arise when considering three alternatives. We will provide experimental evidence below that a level two theory is sufficient to explain behavior in a particular decision making task, however we can also motivate sufficiency on general grounds. Looking at the structure of $\mathrm{Eq}(8)$ for a bilinear theory, we see that the important terms involve the general function $g(\cdot, \cdot)$ evaluated on two different events, $A, B$. This framework is a computational level account, but presumably such a term must arise from a process level account wherein the two events $A, B$ are processed in parallel. In contrast a linear theory only ever involves functions with a single argument, and thus would not require a process level account with simultaneous consideration of multiple events. This strongly suggests that there is a specific sense in which a bilinear model requires a more complex underlying process to instantiate it. By analogy, the analagous term in a level $n$ model will involve a function of $n$ arguments, and, presumably, would require a process model in which $n$ events are considered in parallel.

If a higher order computational theory requires a more complex underlying process to produce it, then we can argue on general grounds that it is unlikely that human decision making is described by a theory of very high order. Of
course, this argument does not tell us whether to expect a bilinear, trilinear etc model, only that a lower level is likely to be preferred over a very high one. The question of exactly what level is needed is an empirical one, which we shall now examine.

## Experiment

We want to test the hypothesis that a bilinear model is necessary and sufficient to capture non-normative effects in human probabilistic reasoning. To do this we need to find sets of at least three disjoint alternatives such that are robust two way disjunction fallacies, in the sense that the interference term $\mathscr{I}_{2}(A, B)$ depends on the events $A, B$. Our approach will be to set up three scenarios, each with three disjoint events, and show that the term $\mathscr{I}_{2}(A, B)$ can be manipulated by introducing joint causes for some of the events. This will prove that a cognitive model capturing these judgments must be at least bilinear. We will then examine the higher order interference term $\mathscr{I}_{3}(A, B, C)$ in each scenario to check that a bilinear model is sufficient. The joint causes we will introduce will either cause the three events to be grouped into two natural sets, with one element shared between sets, or into one set with one singleton event. There are therefore two different ways of presenting each scenario, so we run each as a between participants condition.

## Methods

We recruited 300 participants, equally split into two between participants counterbalancing conditions. Recruitment was through Amazon Mechanical Turk, restricting geographical location to North America. Participants required approximately 20 minutes to complete the task and they were compensated $\$ 1$ for their time.

Both conditions consisted of three scenarios and each scenario described a hospital ward in a fictional town, specializing in a particular type of ailment. For example participants were told of a cancer ward, treating only patients of three types, those with lung cancer, stomach cancer, or throat cancer. For each scenario, participants were given some information creating a common cause between ailment pairs. For example, in one case participants were told that throat and lung cancers are caused by smoking, but throat and stomach cancers by alcoholism. Some rationale was provided to justify each association between ailments. All relations were constructed to look semi-plausible (the authors independently assessed this), but we did not aim for medical accuracy. The between participants condition implemented a counterbalancing manipulation, that presented the same scenarios but varied the common causes.

Participants went through each scenario in a blocked format presentation, so that, for example, no information about a subsequent scenario would be presented prior to finishing all questions relevant to the current scenario (scenario order was randomized). The block for each scenario had analogous format. Participants were first presented with the information about the hospital ward, the ailments treated there, and
the causal relations. Subsequently participants went through four or five multiple choice questions testing knowledge of the causal relations. The questions were meant to be straightforward and answerable on the basis of simple understanding of the presented information. Participants received corrective feedback, specifically if they responded incorrectly they were told so and asked to try again until they answered correctly (there were more than two alternatives for each question).

Once the training part was over, participants were told that they would be asked to make judgments about the proportion of various categories of patients at the fictional hospital. With each question, the text describing the hospital ward and the causal dependencies was included so that participants did not have to memorize anything, just understand the information provided. Each of the questions was prompted with the statement that each patient was brought to the hospital ward for only a single type of ailment (e.g., a single cancer type or a single fracture, depending on the scenario). Then, participants were asked to indicate on a 0 (None of them), to 100 (All of them) slider the proportion of patients likely to be admitted for ailment A in some questions, A or B in other questions, and A or B or C in another question; note, each combination of possibilities was shown only once. The triple disjunction was implemented as a catch question, since the total number of patients was fixed at 100. An additional three catch questions were included, where participants were just told to select a particular response, as a check that they were paying attention.

After participants finished responding to the questions for the three scenarios they were given a version of the the Cognitive Reflection Test (Frederick, 2005). The CRT is designed to discriminate between participants adopting either a more intuitive or a more deliberative thinking style (Toplak et al., 2011). The CRT has previously been shown to correlate well with measures of non-normativity in probabilistic judgments such as conjunction fallacies (Yearsley et al, 2016).
Aside on the physical analogue of this experiment It may be useful for those familiar with interference in physics to outline the analogy between this phenomena and 'interference' effects in human probabilistic decision making. This might help to motivate some of the experimental design, but this subsection can be safely skipped by any reader who wishes.

In the classic two slit interference experiment a particle can arrive at a given point via one of two paths. In quantum theory because of the wave-like nature of particles two things happen: 1) The particle in some sense (which we don't intend to make precise here) takes 'both' paths, and 2) The phase of the particle's wave-function depends on the details of the paths taken. By choosing the paths in a particular way we can cause the two paths to interfere in either a constructive way (so that the total probability of arriving at a point is greater than the sum of the probabilities to follow either path) or a destructive way (so that the total probability of arriving at a point is less than the sum of the probabilities to follow either
path.)
The analogy in decision making is that a disjunction $A \cup B$ of two disjoint events can happen in one of two ways. By manipulating the information we give about the events, for example by introducing a possible common cause, we can, empirically, cause the disjunction to be judged more likely than the sum of the probabilities for the individual events. This is the analogue of constructive interference in the physical set up. This helps us understand why the key experimental manipulation is essentially the stories we tell about the relationship between the different ailments. A pictorial representation is given in Fig 1.

Of course, this analogy is not meant to constitute a formal theory of quantum cognition, but such a theory can be formulated (Busemeyer \& Bruza, 2011; Yearsley \& Busemeyer, 2016).


Figure 1: Sketch of the analogy between the physical experiment and the decision making one. a) In a physical interference experiment, a quantum state can take one of three possible paths to a detector, and the different alternatives interfere. b) In our experiment, a patient ends up in a hospital ward due to one of three ailments. c) A pictorial representation of the different ailments for Condition 1. In Scenario 1 the three ailments were Lung, Throat and Stomach cancers. Lung and Throat cancers were linked (smoking), and Throat and Stomach cancers were linked (alcoholism). In Scenario 2 the three ailments were auto accidents, alcohol poisoning, and falls. Auto accidents and alcohol poisoning were linked (young people) and falls (old people) was a singleton. For Scenario 3 the ailments were fractures to wrists, ankles, or lower legs. Wrist and ankle fractures were given a common cause (everyday falls) and angle and lower leg fractures likewise (playing sports).

## Results

There are two critical tests to perform; firstly we must check that the two way interference terms $\mathscr{I}_{2}(A, B)$ vary depending on the events in each scenario and condition. Second, we must examine three way interference $\mathscr{I}_{3}(A, B, C)$ for each scenario and condition. In this contribution, we only report Bayesian statistical tests that were performed using JASP (JASP team, 2016). In particular we report Bayes factors for the alternative versus the null hypothesis, so that values $>1$
indicate evidence for the alternative hypothesis.
Recall that the two way interference term $\mathscr{I}_{2}(A, B)=$ $p(A \cup B)-p(A)-p(B)$. To check the behaviour of these terms we perform a Bayesian RM ANOVA, with scenario and event pair as the repeated measures, and the counterbalancing condition as a between subjects factor. We omit the CRT from this analysis to save space, but there are no interesting effects of CRT. The analysis of effects is shown in Table 1.

Table 1: Analysis of effects for Bayesian Repeated Measures ANOVA of two way interference terms

| Effect | $p($ incl $)$ | $p($ incl $\mid$ data $)$ | $\mathrm{BF}_{\text {Inclusion }}$ |
| :--- | :--- | :--- | :--- |
| Scenario | 0.737 | 1.000 | $2.44 \times 10^{4}$ |
| Condition | 0.737 | 0.999 | $2.49 \times 10^{2}$ |
| Pair | 0.737 | 1.000 | " $\infty$ |
| Scenario*Condition | 0.316 | 0.998 | $1.37 \times 10^{3}$ |
| Scenario*Pair | 0.316 | 1.000 | $2.11 \times 10^{4}$ |
| Condition*Pair | 0.316 | 0.998 | $1.35 \times 10^{3}$ |
| Scenario*Condition*Pair | 0.053 | 0.998 | $1.12 \times 10^{4}$ |

Recall that if the best description of this situation is via a linear model, ie if non-normative effects are either absent, or due only to response error, then we expect to see no effect of scenario, condition or pair. In contrast JASP actually returns a BF of inclusion for the pair variable of $\infty$, indicating an extremely large effect of pair. The other variables, and all the interaction terms, all have large Bayes factors. The large Bayes factors for the interaction terms are unsurprising given the experimental design - the difference between the same scenario in a different condition and a different scenario in the same condition is really a matter of convention.

This analysis shows that a model beyond a linear one is needed to explain these data. Now we perform the analogous test for $\mathscr{I}_{3}(A, B, C)$ to check if a bilinear model is sufficient.

Recall the three way interference term for each scenario is computed as $\mathscr{I}_{3}(A, B, C)=p(A \cup B \cup C)-p(A \cup B)-p(B \cup$ $C)-p(A \cup C)+p(A)+p(B)+p(C)$. To check the behaviour of these terms we perform a Bayesian RM ANOVA, with scenario as the repeated measure, and the counterbalancing condition and CRT as between subjects factors. The analysis of effects is shown in Table 2.

Table 2: Analysis of effects for Bayesian Repeated Measures ANOVA of three way interference terms

| Effect | $p($ incl $)$ | $p($ incl $\mid$ data $)$ | $\mathrm{BF}_{\text {Inclusion }}$ |
| :--- | :--- | :--- | :--- |
| Scenario | 0.737 | 0.313 | 0.163 |
| Condition | 0.737 | 0.407 | 0.245 |
| CRT | 0.737 | 0.141 | 0.059 |
| Scenario*Condition | 0.316 | 0.299 | 0.923 |
| Scenario*CRT | 0.316 | $2.92 \times 10^{-4}$ | $6.33 \times 10^{-4}$ |
| Condition*CRT | 0.316 | 0.003 | 0.006 |
| Scenario*Condition*CRT | 0.053 | $2.2 \times 10^{-7}$ | $3.97 \times 10^{-6}$ |

The results are striking; none of the Bayes factors for inclusion are greater than 1 , indicating that no model containing any combination of these effects is preferred over a null model. The conclusion then is that the terms $\mathscr{I}_{3}(A, B, C)$
are constant - they do not vary when we manipulate common causes implied for the events in the way that the terms $\mathscr{I}_{2}(A, B)$ do. This implies that a bilinear model is sufficient to explain these effects.

The lack of a significant effect of the CRT is actually reassuring; recall that the terms $\mathscr{I}_{3}(A, B, C)$ should be constant regardless of whether a decision maker is using a linear (classical) or bilinear (quantum) model. The CRT has previously been associated with the strength of various measures of nonnormativity (Yearsley et al, 2015) and the fact that it is not predictive here suggests that these effects behave very differently from other measures such as the size of conjunction fallacies.

## Conclusions and Future Directions

The empirical finding of so-called probabilistic fallacies in decision making has led to an intense debate over how much (if any) of human cognition should be understood in terms of the principles of classical, Bayesian, probability theory (Tversky \& Kahneman, 1983). Those who believe these findings cast doubt on the applicability of classical probability theory have tended to respond by abandoning all together the idea of a formal probabilistic framework for decision making. Recent advances in applying quantum probability theory to modelling human decision making (Pothos \& Busemeyer, 2013) raise the possibility that all (or most) of human cognition can be understood in formal probabilistic terms, but the appropriate approach is not classical probability theory but quantum probability.

However at least one objection to using quantum probability theory (there are many) is that it is unclear how exactly this expands the space of possible models. Most accounts of the relationship between quantum and classical models tend to focus on the issue of incompatibility, but this is notoriously hard to make precise. In addition, it is far from clear that quantum probability theory is the only way to generalise classical probability to include incompatible events.

What we have tried to do in this paper is to show how we may take a particular approach to divide up the space of possible computational level accounts of interference effects in decision making. The space of models is split into different levels, of increasing complexity in the sense of higher level interference effects. Classical probability theory, and its generalisations in the form of linear models, occupy the lowest level of this hierarchy, whereas bilinear models such as quantum theory sit at the next level up in complexity. Above these bilinear models are an infinite number of different levels, although we can argue on general grounds that we expect human behavior to be characterised by a relatively low level model.

Each level in the hierarchy has a unique experimental signature, and we used this to show that behavior in a particular probabilistic reasoning task is consistent with a bilinear theory. Of course, whether all current examples of nonnormative probabilistic reasoning are likewise consistent with a bilinear model is an open question. This level consists of
theories where interference between alternatives is computed pairwise. Quantum theory is situated at this level, however our approach is not able to distinguish between different models in a given level. Further work, for example looking at constraints obeyed by quantum theory but not other non-classical probability theories, could address this.

We finally want to outline some future directions for research. One important question is how well our findings generalise when we consider different kinds of relationships between events. In our scenarios different ailments were related, if at all, by common causes, for example smoking can cause both lung and throat cancer. This means that the associated interference terms $\mathscr{I}_{2}(A, B)=p(A \cup B)-p(A)-p(B)$, tend to be positive. It would be useful to show that we can generate negative interference terms by implying the appropriate relationships between conditions, and check that our conclusions still hold. Another future direction would be to extend this approach to other areas of cognition where quantum models have been proposed, for example perceptual decision making.

Another future direction is to try and extend this analysis to models which fall just outside our framework. One such example is the classical probability plus noise model due to Costello and Watts (2014). They propose a general linear model but with an error term which, rather than be a constant, depends on the type of event being considered, eg a single event, a conjunction or disjunction etc. The idea is that more complex events are associated with larger error terms, and they showed this can lead to conjunction fallacies in participants responses, even though the underlying belief states obey classical probability theory. This theory has a slightly different experimental signature from linear or bilinear models, but it can still be tested against them in a similar way. The results of this analysis will be presented elsewhere.

## Acknowledgments

JMY was supported by NSF grant SES-1556415. EMP was supported by Leverhulme Trust grant RPG-2015-311 and IB by H2020-MSCA-IF-2015 grant 696331.

## References

Busemeyer, JR \& Bruza, P (2011). Quantum models of cognition and decision. CUP: Cambridge, UK.
Costello, F. \& Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. Psychological Review, 121(3):463?480.
Frederick, S (2005). Cognitive reflection and decision making. Journal of Economic Perspectives. 19, 25-42.
Gigerenzer, G, Hertwig, R \& Pachur, T (eds.) (2015). Heuristics: The foundations of adaptive behavior. (OUP).
JASP Team. (2016). Jasp. Retrieved from https://jaspstats.org
Jones, M. \& Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences, 34, 169, 231.

Kahneman, D, Slovic, P \& Tversky, A. (eds) (1982). Judgments under uncertainty: Heuristics and biases. (CUP, UK)
Kolmogorov, AN (1933/1950). Foundations of the theory of probability. N.Y. Chelsea Publishing Co.
Oaksford, M. \& Chater, N. (2009). Prcis of Bayesian rationality: the probabilistic approach to human reasoning. Behavioral and Brain Sciences, 32, 69-120.
Pearl, J (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann.
Pothos, E. M. \& Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modelling? Behavioral \& Brain Sciences, 36, 255-327.
Sorkin, RD (1994). Quantum mechanics as quantum measure theory. Mod. Phys. Lett. A, 09, 3119.
Toplak, ME, West, RF, Stanovich, KE (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory \& Cognition. 39, 1275-1289.
Tversky, A., \& Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjuctive fallacy in probability judgment. Psychological Review, 90, 293-315.
Yearsley, JM \& Busemeyer, JR. (2016). Quantum cognition and decision theories: a tutorial. Journal of Mathematical Psychology. 74, 99-116.
Yearsley, JM, Trueblood, JS \& Pothos, EM (2016). When are representations of causal events quantum versus classical? In Papafragou, A., Grodner, D., Mirman, D., \& Trueswell, J.C. (Eds.) (2016). Proceedings of the 38th Annual Conference of the Cognitive Science Society. (pp. 2447-2452). Austin, TX: Cognitive Science Society.

# Evidence for overt visual attention to hand gestures as a function of redundancy and speech disfluency 

Amelia Yeo (ayeo@wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street<br>Madison, WI 53706 USA<br>Martha W. Alibali (martha.alibali@wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street<br>Madison, WI 53706 USA


#### Abstract

We investigated the effect of gesture redundancy and speech disfluency on listeners' fixations to gestures. Participants watched a speaker producing a redundant or non-redundant gesture, while producing fluent or disfluent speech. Eye movements were recorded. Participants spent little time on a speaker's gestures regardless of condition. Gesture redundancy and speech disfluency did not affect listeners’ percentage dwell time to a speaker's gestures. However, listeners were more likely to fixate to a speaker's gestures when they expected the gesture to be non-redundant. Listeners were also more likely to fixate to a speaker's gestures when the speaker was disfluent. Thus, listeners allocate overt visual attention based on the expected usefulness of a speaker's gestures, although evidence does not suggest that they spend more time fixating on these gestures. Furthermore, listeners are sensitive to disfluency in a speaker's utterance and change how they attend to gestures based on qualities of the speech.


Keywords: gesture; eye tracking; communication; multimodal information processing; spatial features

## Introduction

Speakers usually move their hands when conveying a message. It seems intuitive to suggest that speakers gesture to communicate information to their audience. Indeed, at times speakers appear to produce gesture specifically for the purpose of communicating with the listener (Alibali, Heath \& Myers, 2011).

During the process of comprehension, listeners integrate speech and gesture (Willems, Özyürek, \& Hagoort, 2007). Since co-speech gestures can influence listeners’ comprehension of messages, how then do listeners allocate visual attention resources to speakers' gestures? Some researchers have argued that the content of gestures could be perceived peripherally (Gullberg \& Holmqvist, 1999). If true, this would negate the need for listeners to fixate to gestures during comprehension. However, gestures have also been shown to convey additional semantic content not found in speech (e.g., McNeill, 1992; Alibali, Evans, Hostetter, Ryan \& Mainela-Arnold, 2009; Hostetter, 2011). Fixating to these gestures could help comprehension. Communicating in everyday life is often a multimodal process that involves auditory input from speech and visual input from the speaker's face and body (i.e., MacDonald \&

McGurk, 1978; Ekman, 2004). Hence, understanding how listeners allocate visual attention during the process of face-to-face comprehension is important for understanding the mechanisms involved in the online process of interpersonal communication.

There is some evidence that listeners extend overt visual attention to a speaker's gestures (Nobe, Hayamizu, Hasegawa, \& Takahashi, 1997; 2000). In these studies, the authors presented participants with animations of a speaker uttering short phrases while making hand gestures, and recorded eye movements of the participants during the animations. Participants in the study were found to fixate to gestures consistently on most of the videos presented, preferring to fixate to gestures that occurred more slowly. In the follow-up study (Nobe et al., 2000), participants were found to be able to complete gesture reproduction and comprehension tasks without necessarily fixating to the specific gesture, suggesting that listeners indeed can encode aspects of speakers' gesture without gaze fixations. However, this raises the question of why listeners would consistently fixate to speakers' gestures if comprehension can occur without fixation.

In contrast, other studies of visual attention to gesture have found that listeners rarely fixated to a speaker's gestures, even when those gestures were essential for comprehension (i.e., listeners seldom fixated to gestures that offered information absent from, and thus, non-redundant, with speech). Listeners fixated overwhelmingly on the speaker's face (Gullberg \& Holmqvist, 1999, 2006; Gullberg \& Kita, 2009; Beattie, Webster, Ross, 2010), contrary to the findings by Nobe and colleagues (1997; 2000).

A possible explanation for the differences found in overt visual attention to gestures in these studies is that the speech content of the speakers in the previous experiments was vastly different. Some studies used stimuli that contained a narrative element, while other studies used shorter utterances without a story element, such as "let's count fingers". The difference in speech content could have caused listeners to attend more to the face of the speaker due to the expectation or existence of emotion cues in the speaker's face. Therefore, listeners might spend more time fixating to non-redundant gestures for speakers if the speech content does not contain a narrative element.

In cases where listeners were found to fixate to a speaker's gestures, numerous factors have been cited as potentially driving the fixations. These factors include whether the speaker fixated upon the gesture, the duration of the post-stroke hold (i.e., an aspect of the "form" of the gesture) and the location of the gesture in the speaker's gesture space (e.g., Gullberg \& Holmqvist, 2006; Gullberg \& Kita, 2009). The focus in the literature has thus been on particular physical features of gestures, with little research into the role of listener expectation on overt attention to gestures. Expectations, or predictions, that listeners hold about the usefulness of a speaker's gestures could influence how they attend to the speaker's gestures. In this study, we examine a higher-level feature of gesture, expected redundancy. Keeping all other physical features constant, we test whether the expected redundancy of a speaker's gestures will affect how listeners attend to those gestures. If listeners do allocate attention differently to gestures depending on whether they expect the gesture to be useful for comprehension, then we should see listeners spend more time fixating to gestures and also be more likely to fixate to a speaker's gesture when the gesture offers disambiguating information absent in speech.

As mentioned above, previous studies that examined visual attention to gestures have focused on how physical qualities of a gesture influenced listeners' fixations. In multimodal communication, however, elements of speech can also influence how listeners attend to a speaker's gestures based on existing expectations. To date, no study to our knowledge has examined the role of speech disfluencies on listeners' fixations to a speaker's gestures. Disfluencies such as filled pauses cause a break in speech content and can occur at several points in speech (Ferreira \& Bailey, 2004). A filled pause (i.e., um) that occurs in the middle of a clause has been linked to the need for the speaker to select an option for production from among several competing choices (Clark \& Fox Tree, 2002). Listeners were more likely to remember a word when it was preceded by a filled pause (Corley, MacGregor \& Donaldson, 2007), suggesting that a filled pause could give rise to expectation in listeners that what is to follow is important, signaling listeners to allocate more attentional resources to encode what follows from it. When listeners hear an "um" from a speaker, they might also be more likely to fixate to the speaker's gesture space when the disambiguating information might be produced in gesture as compared to a situation where there is no need for disambiguation.

In this study, we examine the effect of gesture redundancy (i.e., whether a gesture is useful for disambiguating between two options) and speech disfluency on listeners' visual attention to gesture. To do this, we conducted a 2 by 2 fully within-subjects experiment, manipulating gesture redundancy and speech disfluency. We recorded the gaze fixation data (i.e., how long each participant fixated and how many fixations) of each participant as they watched a video of a speaker on each trial. The speaker produced either redundant or non-redundant gestures for a following
task and spoke with either disfluency or without disfluency. We hypothesize that listeners will be more likely to fixate to gestures that are non-redundant with speech. Participants are predicted to fixate at least once to gestures more often for trials with non-redundant gestures than for trials with redundant gestures. Participants are also predicted to spend more time fixating on these gestures. In addition, we also hypothesize that listeners will be more likely to fixate to gestures that accompany disfluent speech than to gestures that accompany fluent speech. This experiment will also allow us to examine whether spatial speech free from narrative content provides a context in which listeners attend less to the speaker's face. However, if the narrative nature of the stimuli used in previous studies was not the reason for the little time listeners spent gazing at gestures, then we expect participants in this study will display similarly low durations of fixations to gesture.

## Method

## Participants

Participants were 30 undergraduate students, all of whom reported being native English speakers. They were recruited from an Introductory Psychology course in exchange for extra credit.

## Materials

There were two sets of stimuli: shape arrays and speaker videos. We created four pairs of shape arrangements using Microsoft PowerPoint, giving eight arrays in total. Each of these eight arrays was repeated twice in the experiment, once paired with a speech-fluent video and once paired with a speech-disfluent video. Thus, there were sixteen target trials in total.

Each pair of shape arrays was identical in every aspect except for a single shape. In the arrays used for the gesture redundant condition, only one triangle was present. In the arrays used for the gesture non-redundant condition, two triangles were present. Thus, to create the arrays for the gesture non-redundant condition, one of the non-triangle shapes in the arrays for the gesture redundant condition was replaced with a triangle (Fig. 1).


Figure 1. An example of a shape array in the gesture redundant condition (left) and in the gesture non-redundant condition (right).

Next, we created eight videos, four featuring fluent speech and four featuring disfluent speech. Each video lasted approximately six seconds and showed a speaker describing a triangle according to a script, while facing the camera (Fig. 2). In the videos with fluent speech, the speaker produced an utterance, such as "the triangle changed color and turned green". In videos with disfluent speech, the speaker produced an utterance with the "um"
disfluency, for example, "the, um, triangle changed color and turned green". In the other of these eight videos, the actor produced exactly the same utterance except with a different color (e.g., orange/red/yellow instead of green). We created the videos such that there were fluent and disfluent pairs containing the same utterance that differed only in the inclusion, or exclusion, of the disfluency "um". In addition, the speaker produced four types of gestures that were paired with corresponding shape arrays. These gestures referred to the triangle that was undergoing the color change. Thus, in the gesture non-redundant condition, the gesture functioned to disambiguate the target triangle from the other triangle in the array. In each video, the speaker's gesture depicted either the pointed tip of the triangle (pointing up or down), or the relative placement of the triangle in the shape array (located high in the array or located above a line). Each gesture was scripted such that the actor began forming the gesture just before the word "triangle" in the utterance and held the gesture for approximately 2 seconds before dropping her hands. In each video, the actor produced only one gesture and gazed at the camera for the duration of the video.


Figure 2. Screen capture of the speaker producing a gesture of an upward-pointing triangle.

We also created shape arrays and speaker videos for filler trials. The purpose of the filler trials was to present the participant with variation in the speaker videos so as to reduce the chances of the participant inferring the purpose of the study. These filler trials contained an assortment of videos where the speaker did not gesture, or gestured while producing a slightly different utterance, such as "the orange triangle changed color and turned green". There were ten filler trials in total. The eight target trials and the filler trials all contained the same actor wearing the same clothing.

## Procedure

Participants were tested individually. Each participant was seated in front of a computer screen and a desk-mounted Eyelink 1000 eye tracker camera. The eye tracker recorded real-time fixations of each participant throughout the entire experiment and was calibrated for each participant before the trials began.

Before the experiment, participants were told that the speaker would always describe a color change of a triangle in the array. Thus, participants began the experiment knowing that it would always be a triangle that changed color. They were not told that the speaker would gesture; participants were not informed in any way that the study
was about gesture or speech disfluency.
During the experiment, each participant viewed 26 trials presented in random order using Experiment Builder from SR Research (Canada). Each trial contained a shape array that was presented onscreen for 5 seconds, followed by a video of the speaker describing the color change occurring to a triangle in the array. The video was programmed to start automatically. After the video, participants were presented with four options of shape arrays and were instructed to say aloud the option that fit the description of the speaker in the video. For example, if a trial presented the array in the gesture non-redundant, speech fluent condition (e.g., the array on the right in Fig. 1) followed by a video of the speaker producing an upward-pointing gesture (Fig. 2) while saying, "the triangle changed color and turned green", the correct option (in Fig. 3) to select would be option C.

The trials in the gesture non-redundant, speech disfluent condition were identical except that the speaker produced a filled pause, for instance, "the um, triangle changed color and turned green". Thus, gesture redundancy was manipulated by having either one or two triangles in the shape array.


Figure 3. Example of four response options in the gesture non-redundant conditions.

An example of a trial in the gesture redundant, speech fluent condition would be the left array in Figure 1 followed by a video of the speaker producing an upward-pointing gesture (Fig. 2) while saying, "the triangle changed color and turned green". The four response options would then contain the same shapes as in the original array but with the single triangle colored in four different colors. The trials in the gesture redundant, speech disfluent condition were identical except that the speaker produced a filled pause, for instance, "the um, triangle changed color and turned green". Thus, gesture redundancy and speech disfluency were perfectly orthogonal.

Each participant's verbal response for each trial was recorded with a microphone that was clipped on to his or her clothing. The verbal responses were recorded to audio files in the computer. At the end of the experiment, participants were debriefed and asked if they could guess the purpose of the study. None of the participants correctly stated the hypothesis about gesture redundancy or speech disfluency on listeners' fixation to a speakers' gestures. Throughout the whole procedure, an experimenter sat in a corner in the room unobtrusively and had no interaction with the participant. The whole experiment lasted for about 20 minutes.

## Coding

Each video was divided into interest areas for eye tracking analysis. The speaker's face was a separate interest area from her gesture space. The fixations of interest for this study were those that occurred to the speaker's gestures from the start to the end of her utterance, since her gestures always occurred as she was speaking. Fixation data that included each dwell time on each area and number of fixations was then exported from Data Viewer (SR Research) for analysis. For each trial, we thus obtained data regarding how long a participant fixated to the speaker's face, how long a participant fixated to the speaker's gesture space, how many fixations a participant made to the speaker's face and how many fixations a participant made to the speaker's gesture space.

## Results

Averaging across all conditions, participants spent the majority of the time fixated on the speaker's face, spending only $9.3 \%$ of the time fixating on the speaker's gestures.

Table 1 displays the average percentage dwell time spent by participants on the listeners' gestures across conditions. We conducted two-way within-subjects analysis of variance on the average percentage dwell time spent fixating on the speaker's gestures as a function of gesture redundancy and speech disfluency. There was no significant main effect of gesture redundancy, $F(1,112)=1.34, p=0.25$, nor was there a significant main effect of speech disfluency, $F<1$, $p=.66$.

There was also no significant interaction between gesture redundancy and speech disfluency on participants' dwell time to speaker's gestures, $F(1,112)=2.30, p=.13$. Even though participants on average spent a higher percentage of dwell time on non-redundant gestures, this difference was not significant.

Table 1. Average dwell time $\%$ to the speaker's gestures as a function of gesture redundancy and speech disfluency.

|  | Gesture |  |
| :---: | :---: | :---: |
| Speech | Redundant | Non-redundant |
| Disfluent | 7.33 | 10.1 |
| Fluent | 9.02 | 10.8 |

Since participants overwhelmingly fixated to the speaker's face in this experiment, we wanted to examine whether gesture redundancy and speech disfluency affected the likelihood of participants fixating at least once to the speaker's gestures. To test whether participants were more likely to fixate to a speaker's gestures as a function of gesture redundancy or speech disfluency, we classified whether each participant fixated on the video speaker's gesture space at least once while the speaker was talking. Thus, the outcome variable for this analysis was dichotomous, i.e., whether or not the participant fixated at least once to the speaker's gesture in each trial.

We analyzed these data using a binomial multilevel model with gesture redundancy and speech disfluency as fixed effects and participant as a random effect. The dependent variable was whether the participant had fixated to the speaker's gesture space (yes/no). The mean proportion of trials on which participants fixated at least once to the speaker's gesture space is displayed as a function of gesture redundancy (Fig. 4) and speech disfluency (Fig. 5).


Figure 4. Proportion of trials on which participants had at least one fixation to the speaker's gesture space as a function of gesture redundancy. Error bars are $\pm$ SE.


Figure 5. Proportion of trials on which participants had at least one fixation to the speaker's gesture space as a function of speech disfluency. Error bars are $\pm$ SE.

Listeners were significantly more likely to fixate to the speaker's gesture in the gesture non-redundant condition than in the gesture redundant condition, Wald's $z=3.06, p<$ .01 , odds ratio $=2.41$. Additionally, listeners were significantly more likely to fixate to the speaker's gesture in the disfluent speech condition than in the fluent speech condition, Wald's $z=2.21, p=.027$, odds ratio $=1.88$. There was no significant interaction between gesture redundancy and speech disfluency on the likelihood of participants fixating to a speaker's gesture, Wald's $z=1.35$, $p=.18$. In sum, participants were more likely to fixate at least once to non-redundant gestures, and they were also more likely to fixate at least once to the speaker's gestures when the speaker was disfluent.

## Discussion

The finding that participants spend little time fixating to a speaker's gestures reflects the results from some past studies. For example, Gullberg and Kita (2009) reported that listeners fixated on gestures only $8 \%$ of the time, even
though these gestures were first fixated by the speaker, showing that gesture fixation duration was low even when there was social impetus (i.e., directed gaze) to fixate at a gesture. Our findings align with this value. Listeners fixated to gestures on average about only $10 \%$ of the time, even for gestures that contained information not present in the speaker's utterance. These findings do not support the hypothesis that the previously reported low fixation durations on gestures were due to the narrative element in speech. Instead, listeners fixate overwhelmingly on the speaker's face even when the narrative element in speech is absent or greatly reduced.

However, we do not yet know if listeners direct so little overt visual attention to gesture because of the communicative context. Past studies, including this one, have featured speakers passively describing objects or actions. Although strengths of this paradigm are its simplicity and ease of experimental control, a limitation is that it tells us little about how people attend to each other's gestures when they are engaging in dialogue. During dialogue, speakers gesture differently depending on the feedback they receive from the listener (Holler \& Wilkin, 2011). This finding reflects observations of research involving instructional gestures. In the classroom, teachers have been found to gesture more when students lack understanding of the lesson (Alibali et al., 2013). Further research could explore how listeners attend to gestures in an instructional setting or in dialogue, using a wearable eye tracker.

As predicted, participants were more likely to fixate to non-redundant gestures than to redundant gestures. This finding implies that listeners preferentially direct overt visual attention to gestures that they expect to be useful for comprehension. Listeners direct overt attention to a speaker's gestures more often when the gesture conveys relevant information not present in speech, implying that listeners generate expectations about the perceived importance of the speaker's gestures and direct attention accordingly. However, we did not find support for the hypothesis that listeners would spend more time fixating to a speaker's gestures. While listeners were more likely to gaze at least once to the speaker's non-redundant gestures, they did not spend more time dwelling on those gestures, implying that the additional fixations to non-redundant gestures occurred very quickly. A potential explanation for this behavior is that visual information from fixated gestures is gleaned very quickly, making it unsurprising that fixation durations across conditions did not differ significantly.

On the surface, it might be unsurprising that listeners are less likely to fixate to gestures that are redundant. This study demonstrates that listeners are less likely to fixate to gestures that are redundant even when those gestures are holds (i.e., the form of the gesture is held in a pause) and occur in the center of the speaker's body, qualities that were reported to best attract listeners' fixations (e.g., Gullberg \& Holmqvist, 1999; 2006) Since we controlled for these features across the gesture redundant and gesture non-
redundant conditions, our findings imply that top-down factors such a redundancy can influence listeners' visual attention to gestures beyond the physical characteristics of those gestures. Few studies to date have explored the role of higher-level cognitive factors, such as expectations, on how listeners process gestures. For example, individuals could hold expectations about the usefulness of gesture based on an individual's communicative fluency, or individual's communicative style. A further direction would be to examine how these factors influence how listeners attend to gestures.

In this study, we also found support for the hypothesis that speech disfluency causes listeners to be more likely to attend to gestures during communication. These findings support the idea that disfluencies in speech can function as a signal to listeners on how to direct their cognitive resources during comprehension. However, we did not find support for the hypothesis that listeners spent more time fixating to gestures that co-occurred with disfluent speech as compared to gestures that occurred with fluent speech. Once again, it is possible that listeners quickly obtained information from gestures. If filled pauses in speech do indeed work as a signal for cross-modal attention shifts, future work could examine if how other forms of speech disfluencies (e.g., false starts) influence visual attention to a speaker's gestures.

As with any investigation, there are some limitations to this experiment. Due to convenience sampling, our sample was comprised of college undergraduates. Undergraduates could offer little variation in terms of cognitive skills as compared to the population at large. While little published research to date exists examining the role of individual differences in cognitive skills on attention to gestures, there is evidence suggesting that people produce gestures differently due to individual differences in spatial abilities (Hostetter \& Alibali, 2007; 2011). It may be the case that listeners with vastly different spatial skills could process a speaker's gestures differently. One way to address this would be to administer measures of verbal and spatial skills to undergraduate participants in future studies. Another way to address this limitation would be to recruit participants outside of the undergraduate pool.

Our participants were English speakers in the Midwestern USA, thus the results might not generalize to speakers of a different language or culture. Past studies on visual attention to gestures have sampled from English-speaking students in the United Kingdom (Beattie, Webster \& Ross, 2010), Dutch-speaking students (Gullberg \& Kita, 2009) and native Swedish speakers (Gullberg \& Holmqvist, 2006). Consistently low fixation durations to gesture across these samples appears to suggest that the effect is generalizable. However, Nobe and colleagues (1997; 2000) sampled from Japanese speakers, raising the question of whether the difference in attention to gestures of a speaker is partly due to cultural norms.

For instance, Graham and Argyle (1975) found that Italian speakers were better able to decode shapes being
described by the speaker when gesture was produced, in contrast to English speakers. If speakers' gestures possess different utility value to listeners depending on the language, we might expect listeners to attend to gestures differently too. Further research should test the assumption that listeners' processing of speakers' gestures is universal. There are undoubtedly common processes involved in multimodal communication across humans, but cultural norms in communication or in the use of hand gestures could also influence how listeners process these gestures.

Another limitation of this study involves the nature of scripted disfluencies. When disfluencies are produced naturally, they could be accompanied by changes in speech rate, tone of voice, or changes in facial expression. Having an actor utter a statement with a scripted disfluency across multiple trials is unnatural. While this choice was made to reduce stimuli variability, further research could use videos of speakers conversing naturally and examine the gaze of listeners when disfluency occurs naturally.

In conclusion, these findings provide another perspective on the question of how listeners process gestures. We show that listeners are more likely to fixate to a speaker's gestures when those gestures are non-redundant, after controlling for physical properties of gesture that have been reported to capture the attention of listeners. We also demonstrate that speech disfluencies can act as signals for listeners to shift attention multimodally. These findings highlight the causal role of expectations in how listeners attend to speakers' gesture.

## Acknowledgments

We thank Maia Ledesma and Kayla Diffee for assistance with production of stimuli, Youn Ku Choi for assistance with data collection, and Mitchell Nathan and Virginia Clinton for assistance with the eye tracker. We also thank Marianne Gullberg for helpful comments during presentation of a version of this work at the conference of the International Society for Gesture Studies in Paris.

## References

Alibali, M. W., Heath, D. C., \& Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. Journal of Memory and Language, 44(2), 169-188.

Alibali, M. W., Evans, J. L., Hostetter, A. B., Ryan, K., \& Mainela-Arnold, E. (2009). Gesture-speech integration in narrative: Are children less redundant than adults?. Gesture, 9(3), 290-311.

Alibali, M. W., Nathan, M. J., Church, R. B., Wolfgram, M. S., Kim, S., \& Knuth, E. J. (2013). Teachers' gestures and speech in mathematics lessons: Forging common ground by resolving trouble spots. $Z D M, 45(3), 425-440$.

Beattie, G., Webster, K., \& Ross, J. (2010). The fixation and processing of the iconic gestures that accompany talk. Journal of Language and Social Psychology, 29(2), 194213.

Clark, H. H. \& Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. Cognition, 84, 73-111.
Corley, M., MacGregor, L. J. \& Donaldson, D. I. (2007). It's the way you, er, say it: Hesitations in speech affect language comprehension. Cognition, 105, 658-668.
Ekman, P. (2004). Emotional and conversational nonverbal signals. In Language, knowledge, and representation (pp. 39-50). Springer Netherlands.
Ferreira, F. \& Bailey, K. G. D. (2004). Disfluencies and human language comprehension. Trends in cognitive sciences, 8, 231-237.
Graham, J. A., \& Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. International Journal of Psychology, 10(1), 5767.

Gullberg, M., \& Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. Pragmatics \& Cognition, 7(1), 35-63.

Gullberg, M., \& Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. Pragmatics \& Cognition, 14(1), 53-82.
Gullberg, M., \& Kita, S. (2009). Attention to speechaccompanying gestures: Eye movements and information uptake. Journal of nonverbal behavior, 33(4), 251-277.
Holler, J., \& Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. Journal of Pragmatics, 43(14), 3522-3536.
Hostetter, A. B., \& Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. Gesture, 7(1), 73-95.
Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. Psychological bulletin, 137(2), 297.

Hostetter, A. B., \& Alibali, M. W. (2011). Cognitive skills and gesture-speech redundancy: Formulation difficulty or communicative strategy?. Gesture, 11(1), 4060.

MacDonald, J., \& McGurk, H. (1978). Visual influences on speech perception processes. Attention, Perception, \& Psychophysics, 24(3), 253-257.
McNeill, D. (1992). Hand and mind: What gestures reveal about thought. University of Chicago Press.
Nobe, S., Hayamizu, S., Hasegawa, O., \& Takahashi, H. (1997, September). Are listeners paying attention to the hand gestures of an anthropomorphic agent? An evaluation using a gaze tracking method. In International Gesture Workshop (pp. 49-59). Springer Berlin Heidelberg.
Nobe, S., Hayamizu, S., Hasegawa, O., \& Takahashi, H. (2000). Hand gestures of an anthropomorphic agent: Listeners' eye fixation and comprehension. Cognitive Studies, 7(1), 86-92.
Willems, R. M., Özyürek, A., \& Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. Cerebral Cortex, 17(10), 2322-2333.

# Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints 

Ilker Yildirim*1 ${ }^{\text {(ilkery @ mit.edu), Tobias Gerstenberg*1 }}{ }^{\text {(tger@ mit.edu), Basil Saeed }}{ }^{1}$ (bsaeed@mit.edu), Marc Toussaint ${ }^{2}$ (marc.toussaint@informatik.uni-stuttgart.de), Joshua B. Tenenbaum ${ }^{1}$ (jbt@mit.edu)<br>${ }^{1}$ Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA<br>${ }^{2}$ Machine Learning and Robotics Lab, University of Stuttgart, Germany


#### Abstract

In this paper, we present a new task that investigates how people interact with and make judgments about towers of blocks. In Experiment 1, participants in the lab solved a series of problems in which they had to re-configure three blocks from an initial to a final configuration. We recorded whether they used one hand or two hands to do so. In Experiment 2, we asked participants online to judge whether they think the person in the lab used one or two hands. The results revealed a close correspondence between participants' actions in the lab, and the mental simulations of participants online. To explain participants' actions and mental simulations, we develop a model that plans over a symbolic representation of the situation, executes the plan using a geometric solver, and checks the plan's feasibility by taking into account the physical constraints of the scene. Our model explains participants' actions and judgments to a high degree of quantitative accuracy.


Keywords: planning; problem solving; logic-geometric programming; intuitive physics; scene understanding

## Introduction

Physical problem solving - converting knowledge into behavior to achieve a goal that involves physical object manipulation - is a core component of human intelligence and ubiquitous in everyday cognition. From young children playing with stacking cups to an adult moving furniture to redesign a room or to load a truck, our intuitive understanding of how to manipulate the physical world in order to meet our goals is remarkable. For instance, when rearranging the furniture in a room, one needs to form and execute a plan which takes into account both spatial and physical constraints, such as how big are the objects, and which objects might be stacked on top of others.

Two independently developed lines of research provide insights and starting points into exploring these computations: reasoning based on mental models, and motor control based on forward models. Firstly, the theoretical and behavioral work on reasoning and problem solving in symbolic domains (e.g., logical reasoning, or visuo-spatial reasoning) emphasizes the importance of common-sense knowledge. For instance, early Artificial Intelligence (AI) systems that were built to reason like humans do, focused on building models that capture aspects of common-sense knowledge about the physical world in the form of knowledge representations and methods to efficiently manipulate them (e.g., Newell, Shaw, \& Simon, 1958). Similarly, in cognitive psychology, the idea that problem solving begins with the construction of a mental model of the situation was explored in more detail by mental model theory (Johnson-Laird, 2005). While still operating over logical representations, mental model theory makes additional assumptions about what aspects of a situation people naturally represent, and how these representa-
tions support reasoning (Johnson-Laird, Khemlani, \& Goodwin, 2015). However, the theoretical and behavioral work on human reasoning and problem solving has tended to focus on symbolic domains (e.g., logical, spatial, and visuo-spatial reasoning Newman, Carpenter, Varma, \& Just, 2003; Byrne \& Johnson-Laird, 1989), and has not yet looked into situations that require reasoning about physical objects, and forming plans about how to interact with them.

Secondly, research on computational motor control and object manipulation emphasizes the knowledge and transformations necessary for skillful manipulation of objects. For instance, work on sensorimotor control and object manipulation extensively studied internal models of the forward dynamics of the arm and the objects, as well as how to choose actions to efficiently achieve one's goals based on internal models (Nagengast, Braun, \& Wolpert, 2009; Franklin \& Wolpert, 2011). However, this line of work has tended to focus on relatively simple actions, instead of settings that involve planning longer sequences of moves.

In this paper, we aim to bring these two different research traditions together. To better understand physical problem solving, we introduce an intuitive, yet complex task in which participants are asked to manipulate a stack of blocks to generate a target configuration. Consider Problem 1 shown in Figure 1. The task is to manipulate the blocks so that the scene on the left is turned to the scene on the right. While participants have no trouble doing this task, and even young children naturally perform such tasks, modeling people's actions is far from trivial and robotic systems rarely implement this kind of flexible manipulation. The task requires representing the initial state, the final state, and making a plan for how to get from A to B . Finding good action sequences in this task not only requires a symbolic high-level plan (e.g., which sequence of actions to take) and visuo-spatial reasoning, it also requires intuitive physical reasoning about how objects support each other (i.e., their dynamics) and actual motor control required to execute the high-level abstract plan. Such combination of rich behavior is common in everyday cognition, but has rarely been studied in the lab. We used two different versions of the task. In one version, participants in the lab were asked to generate the different configurations. In another version of the task, we had online participants judge whether they think the person in the lab used one or two hands to get from A to B (cf. Figure 1E).

We develop a novel computational model of physical problem solving that goes all the way from formulating an abstract symbolic plan to executing the low-level motor commands that are required to realize the plan. The model is com-


Figure 1: Experimental setup. A: Example for an initial and final configuration of the three blocks. B: Illustration for what moves were legal (green border) or illegal (red border). C and D: Some example problems. E: Screenshot of the experimental interface for participants in Experiment 2.
posed of three components: (1) a symbolic representation of the scene, (2) a geometric solver for motion synthesis, and (3) a physics engine for physical reasoning. Planning in the model operates over the symbolic representation of the scene. Each plan is composed of subgoals and finds a sequence of moves that turn the initial into the final configuration (see, e.g., Figure 3C, left side). An optimization-based kinematics solver takes the symbolic plan as its input and generates a full motion plan which we implement in a simulated twoarmed robot (Figure 3C, right side). We use a physics engine to check whether the plan that the kinematic solver came up with is feasible. More specifically, we test at each point when a subgoal is reached, whether the configuration is physically stable. If the plan includes an unstable configuration, it is discarded (Figure 3D for a plan that includes an unstable state). The model's task is to get from the initial stack shown in A to the target stack. However, just taking the red block and moving it to the right so that it's correctly positioned relative to the yellow block, causes the blocks to fall over.

For each pair of initial and target stack, the model is able to generate plans using either only one arm, or both arms. We score each plan based on its efficiency which is a function of the number of the moves it takes to get from the initial to the target stack, as well as the effort that the plan takes. We evaluate the contributions of the three different components of our model through lesion studies (i.e. we remove parts of the model and see how well it does, in order to gauge what components are necessary to capture people's behavior).

The remainder of this paper is organized as follows: first, we describe a novel, physical problem-solving task and show how participants solve the task in the lab and online. Next, we describe our computational model and analyze how well it does in accounting for participants' behavior. We conclude by highlighting the key contributions of the paper, and by suggesting several lines of future research.

## Stack re-configuration problems

Most classical paradigms used to study problem solving, such as the Tower of Hanoi and its variants require visuospatial reasoning and planning for successful solutions. Here we present a novel problem which requires the problemsolver to also take into account physical constraints, such as considering whether a particular configuration of blocks will be stable.

The problems involve an initial stack of three physical blocks on a table paired with an image showing the desired target stack of the same three blocks (Figure 1A). The three wooden blocks had the same size and mass, and were colored in red, yellow, and blue. Given the pair of initial and target stacks, the problem is to re-configure the initial stack such that it will match the target stack in the image. While interacting with the blocks, participants aren't allowed to touch more than one block at a time. Example legal and illegal moves are shown in Fig 1B. To solve each stack re-configuration problem, participants have to plan and execute a set of moves (using one or both hands) that will generate the target stack from the initial stack.

## Experiment 1: Physical task

The goal of Experiment 1 was to assess how participants interact with the scene to get from the initial to the final configuration for each problem. In particular, we were interested in seeing whether they used one hand or two hands to get from A to B.

## Methods

Participants 10 participants $\quad\left(M_{\text {age }}=35, S D_{\text {age }}=\right.$ $16.4, N_{\text {female }}=6$ ) were recruited from MIT's subject pool. The study took about 15 minutes to complete, and all participants were compensated for their participation.
Stimuli The three physical blocks used in the experiment were of size $10 \mathrm{~cm}-5 \mathrm{~cm}-5 \mathrm{~cm}$ (height-width-depth) and


Figure 2: The probability that participants used one hand in the lab (Physical) together with the mean judgments provided by participants online (Mental) for 34 different problems. Note: Error bars indicate $95 \%$ bootstrapped confidence intervals.
weighed about 50 grams. We manually arranged these 3 blocks into 38 different configurations and took a picture of each configuration. The configurations were constrained such that all blocks remained within a spatial boundary on a table, and the block or blocks touching the table were centered at one of three designated spots. Figure 2 shows some examples of initial and final configurations. ${ }^{1}$

Procedure After providing written consent, participants were introduced to the task, including what moves were legal and which ones were illegal. Starting from the initial stack configuration of Problem 1, participants were asked to reconfigure the blocks to the target stack of Problem 1, which was presented on a computer screen in front of them. They clicked on the "Continue" button on the screen to indicate that they were done and the experiment moved on to the next problem.

The initial configuration of the next problem, Problem 2 (Figure 2C), was the target configuration from the previous problem, and so on. This sequence of problems continued for a total number of 37 problems. ${ }^{2}$ The presentation order was the same for all participants. All participant responses were video-recorded. For each problem, we coded whether participants used one or two hands to solve it.

## Results

Figure 2 shows the proportion of participants who used one hand for each trial. In some trials, most participants used only one hand (e.g., Problem 21, Figure 1D), and in others most participants used both hands (e.g., Problem 34, Fig 1D). Across all trials, participants used one or two hands about equally. Participants often solved the problem with one hand if it was possible to do so. Some participants only used their non-dominant hand if it was impossible to achieve the target configuration with one hand only.

[^546]
## Discussion

Overall, we found that participants had no trouble doing the task. There was considerable variance in how participants solved the different problems with some participants almost exclusively using one hand (if possible) and others being more likely to use two hands to get to the target configuration.

Experiment 1 serves as a baseline to see how participants actually interact with the physical scene. In Experiment 2, we were interested to see how people mentally simulate the way in which they would interact with the scene to get from the initial to the final stack. If participants are able to mentally do this task, we would expect a close correspondence between the judgments participants make based on their mental simulation, and the actual behavior of participants in the lab.

## Experiment 2: Mental task

The goal of this experiment was to test whether participants can simulate how another person would interact with a physical scene to get from A to B.

## Methods

Participants 40 participants $\quad\left(M_{\text {age }}=35, S D_{\text {age }}=\right.$ $14, N_{\text {female }}=22$ ) were recruited via Amazon's crowdsourcing service Mechanical Turk. The experiment took 8.7 minutes ( $S D=4.4$ ) to complete and participants were compensated at an hourly rate of $6.0 \$$.
Stimuli The same pairs of initial and target stacks as Experiment 1 were used, with the exception that both stacks were presented on the screen side by side.
Procedure Participants saw two images side by side with the left image showing the initial stack and the right image showing the target stack (example pairs in Figure 1 except panel B). They were instructed that "The image on the left shows you the initial configuration of the blocks. The image on the right shows you the configuration after the person interacted with the blocks." Their task was to judge whether the person had used one hand or two hands to re-configure the stack. They entered their response by adjusting a slider bar at the bottom of the screen (see Figure 1). Then they clicked on the "Continue" button to proceed to the next problem. The different problems were presented in randomized order.

## Results

Figure 2 shows participants' mean judgments for the different problems. To assess how well participants' mental simulations correspond with the actions that participants took in the experiment, we compared the mean responses in Experiment 2 with the proportion of participants who used one hand in Experiment 1.

Overall, we found that participants' judgments about how many hands the person used correlated well with participants' actual behavior in the lab, $r=.73, p<.05$. Whereas there were many trials for which the correspondence between judgments and actions was very high (e.g. Problems 1-6, or 2132), there were also situations in which actions and judgments came apart. For example, in Problem 34 almost all participants in the lab used two hands, whereas online participants believed that it was likely that a person would only use one hand to re-configure the scene.

## Model

The model consists of three components: (1) a set of abstract motion primitives that can be composed to symbolic plans for re-configuring an input stack to a target stack, (2) a hierarchical kinematics-based optimization algorithm to find manipulation trajectories conditioned on the symbolic plan, (3) and a physics engine to evaluate the stability of the intermediate stages produced by the execution of the manipulation trajectories. The first two components of our model are based on the logic-geometric programming framework (Toussaint, 2015).

## Logic-geometric programming framework

The logic-geometric programming framework presents a solution to problems of combined task and motion planning. Such tasks involve sequential manipulation of a scene based on a geometrically defined goal function. It utilizes symbolic task descriptions as (in-)equality constraints within a hierarchical geometric solver to find full manipulation and object trajectories starting from a coarse-level solution to eventually fine-grained full-paths. Below, we present our representations and an algorithm for symbolic planning as well as a general outline of the geometric solver.
Symbolic plans Symbolic plans are sequences of a set of abstract move types defined using actuators, movable objects and fixed objects in a simulated world. The moves change the state of the actuators and the movable objects. The world is described as a linked list of fixed and movable objects with relative world coordinates: the position and rotation of a child object is defined relative to its parent.

In order to model our stack re-configuration tasks, we populated the world with three movable objects (red block R, green block G , and blue block B), and a fixed object (table T ). The world also includes a robotic body with arms and pincer hands (actuators: handL and handR) overall consisting of 12 degrees of freedom (two at each shoulder, two at each wrist, and two at each hand).

There are three types of moves: Grasp (Obj, Act) speci-
fies a grasp action with an actuator on a movable object. For example, $\operatorname{Grasp}(R$, handR) specifies a right hand grasp of the red block. This move changes the position of the object to inside in the actuator while clearing its previous location for moving other objects. The symbolic planning stage doesn't take into account rotation of the objects or the actuators.

Place (Obj, Supp_Obj, Act) specifies any place action that is not final of a movable object on another object using an actuator. For example, Place (B, T, handL) specifies placing the blue block on the table using the left hand. This move changes the position of the object (e.g., the red block) to be on top of the support object (e.g., an empty location on top of the table) while clearing its previous location. The rotation again is not handled at the symbolic planning stage.

Fix(Obj, Supp_Obj, Act) specifies any place action that is final of a movable object on another object using an actuator. For example, Fix (G, R, handR) specifies final fixation of the green block on top of the red block using the right hand. This move changes the position of the object (e.g., the green block) to be on top of the support object (e.g., red block) while clearing its previous location. Fix action is always final - the object isn't moved after.

Given a pair of stack configurations as input, we wish to find sequences of moves (symbolic plans) that transform the initial stack to the target stack. We used Monte Carlo tree search (MCTS) to find satisfying sequences by branching the search tree using the three move types, the three objects, the four support objects, and the two actuators. Our pruning algorithm was efficient to a certain extent - for example, if an object is already grasped, we did not branch the grasp move on it again. We also imposed a condition to produce a specialized set of solutions which we labeled as the efficient set, leaving the label inefficient for the universal set of solutions. To produce the efficient set, we would only branch the search tree to a Place (Obj, ., .) if the Fix (Obj, ., .) was not currently available for the block. We increased the maximum length of move sequences until no new unique solutions could be found.

After a sequence was deemed satisfactory, we assigned integral timestamps to each of the abstract moves that it is composed of. These timestamps indicated the discrete-time values that an abstract move should be executed at. The assignment was done in a way to allow the execution of as many concurrent moves as possible. Of course, when a solution is one-handed, only one move can be executed at a time, thereby each abstract move must be assigned a separate timestamp. However, with two-handed solutions, different blocks can be concurrently actuated by different hands. Example symbolic plans for a pair of initial and target stack configurations are shown in Fig 3.

We assigned a complexity score to every symbolic solution generated, denoted $s_{i, j}$ where $i$ indexes problems and $j$ indexes its solutions. The score for a sequence is equal to the discrete-time that this sequence takes to terminate.


Figure 3: Illustration of how the model works. A: The model successfully went from the inital to the final configuration. B: The symbolic plan for going from Step 1 to Step 2 using two hands. C: A more involved plan that requires 8 moves. D: Example of a scene where a plan fails because it created an unstable configuration (as determined by the physics engine).

Geometric solver The geometric solver can be thought of as compiling a symbolic plan to manipulation trajectories of actuators and movable objects. It is based on a hierarchical optimization procedure for combined task and motion planning where the tasks come from the symbolic plan. Conditioned on the symbolic plan, the geometric solver generates a number of equality and inequality constraints that need to be met by the optimization procedure. These constraints are solved using an optimization package ( k -order motion optimization framework, KOMO Toussaint, 2014) that can handle long-distance dependencies such as the dependencies between actuator and object trajectories across time steps. Due to space limits, we cannot provide any further the details of KOMO and the logic-geometric programming framework (but see Toussaint, 2015, 2014). Snapshots of example manipulation trajectories generated by this optimization procedure for a pair of initial and target stack configurations are shown in Fig 3.

## Physical stability inference

Because the geometric solver only considers kinematics and not the physical dynamics of the scene, it can find solutions that have physically unstable intermediate steps. In-
spired by (Battaglia, Hamrick, \& Tenenbaum, 2013), we infer whether a given intermediate configuration is stable by physically instantiating it in a physics engine (PhysX) and measuring the total kinetic energy over a total simulation duration of 1 sec with a burn-in period of 100 msecs . We reject a solution if the total kinetic energy exceeds an empirically determined threshold of 0.1 joules.

Similar to the complexity score for the symbolic solutions, we assigned an approximately metabolic cost score to every full model solution found (that is, solutions after the physical stability inference step), denoted $f_{i, j}$ where $i$ indexes problems and $j$ indexes its solutions. This score captures the extent to which a particular plan requires effort to execute. The score starts with the symbolic complexity score, $s_{i, j}$, but adds two more quantities: (1) an extra cost of 0.5 for moves involving multiple blocks (e.g., actuating-i.e., grasping, placing or fixing- the red block while the blue block rests on top of it), and (2) an extra cost of 0.5 for moves that result in an intermediate physically unstable configuration from which the solver can recover to reach the correct stable configuration (e.g., moving the yellow block while the red block is leaning on it, and subsequently moving the red block).


Figure 4: Scatter plots showing the relationship between different versions of the model (columns) and participants' actions in the lab (top), or mental simulations online (bottom). Note: $1=$ definitely one hand, $0=$ definitely two hands.

## Simulations and results

In addition to our full model, we also considered a lesioned model which leaves out the physical inference component. We assume that people aim to reach their goal efficiently. Hence, we assume that sequences with higher complexity scores or metabolic costs are less likely to be chosen (in the lab) or simulated (online) than those with lower complexity scores or costs. For a given problem $i$, we obtain the probability of choosing one-hand based on the symbolic complexity scores in the following way $\frac{\sum_{j \in \text { one-hand solutions }} e^{-s_{i, j}}}{\sum_{j \in \text { all solutions }} e^{-s_{i, j}}}$. This means that the model is more likely to choose a one-hand solution the lower the cost of one-hand solutions are relative to all possible solutions.For the full model, the probability of choosing one-hand, $\operatorname{Pr}($ One-hand $)$, is calculated identically but using the full model scores, $f_{i, j}$.

Overall, we found that the model accounted well for the data (see Fig. 4). In particular, we found that both physical stability inferences and efficiency were necessary to account for participants' judgments in Experiment 2 ( $r=.74$, comparisons to symbolic-efficient, symbolic-inefficient and full-model-inefficient $p<.05$ using direct hypothesis testing with the bootstrap samples).

Similarly, in Experiment 1, we found that physical stability inferences were necessary to best explain participants' behavior (with $r=.68$ of the full model compared to $r=0.63$ of a model that doesn't take into account efficiency). But we did not find a statistical difference between using only the efficient solutions versus all solutions ( $p=.06$ ).

## General Discussion

We presented a novel paradigm - the stack re-configuration problems - and studied people's solving these problems in the laboratory (Experiment 1) and mentally simulating what they think a person would do (Experiment 2). We found that participants' judgments about whether they think a person used one or two hands to get from the initial to the target configuration correlated well with participants' actual behavior in the lab.

In order to explain participants' behavior, we developed a computational model that flexibly combines a symbolic, geometric, and physical representation of the scene. It effi-
ciently plans over this representation by first forming a symbolic plan, trying to execute the plan using a geometric solver, and then checking whether the plan was feasible by consulting a physics simulation engine to make sure that each move resulted in a physically stable configuration.

The full model accounts well for participants' actions as well as mental simulations. A model that does not take into account the efficiency of different plans fares worse (particularly when trying to explain mental simulations). Moreover, it is crucial to consider how much effort different plans would take into account well for participants' actions and judgments. Participants chose to use two hands only when a one-hand solution would have required considerably more effort.

A striking aspect of problem solving is that it demands flexible systems that can operate with very little training opportunity, leading many researchers to emphasize the role of common-sense reasoning and model-building as the building blocks of human problem solving (Johnson-Laird, 2005). We find such flexibility and data efficiency in stark contrast with some of the main approaches to artificial intelligence today, in particular to deep learning (Silver et al., 2016). These approaches require huge amounts of data, yet their generalization capacity is limited in contrast to human's flexibility. Turning these data-hungry approaches to flexible problem solvers is a substantial challenge. This paper makes a few (block) moves in this direction.
Acknowledgments This work was supported by the Center for Brains, Minds \& Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333.

## References

Battaglia, P. W., Hamrick, J. B., \& Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45), 1832718332.

Byrne, R. M., \& Johnson-Laird, P. N. (1989). Spatial reasoning. Journal of memory and language, 28(5), 564-575.
Franklin, D. W., \& Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. Neuron, 72(3), 425-442.
Johnson-Laird, P., Khemlani, S. S., \& Goodwin, G. P. (2015). Logic, probability, and human reasoning. Trends in Cognitive Sciences, 19(4), 201-214.
Johnson-Laird, P. N. (2005). Mental models and thought. The Cambridge handbook of thinking and reasoning, 185-208.
Nagengast, A. J., Braun, D. A., \& Wolpert, D. M. (2009). Optimal control predicts human performance on objects with internal degrees of freedom. PLoS Comput Biol, 5(6), e 1000419.
Newell, A., Shaw, J. C., \& Simon, H. A. (1958). Elements of a theory of human problem solving. Psychological Review, 65(3), 151-166.
Newman, S. D., Carpenter, P. A., Varma, S., \& Just, M. A. (2003). Frontal and parietal participation in problem solving in the Tower of London: fMRI and computational modeling of planning and high-level perception. Neuropsychologia, 4l(12), 1668-1682.
Silver, D., van Hasselt, H., Hessel, M., Schaul, T., Guez, A., Harley, T., ... others (2016). The predictron: End-to-end learning and planning. arXiv preprint arXiv:1612.08810.
Toussaint, M. (2014). Newton methods for k-order markov constrained motion problems. arXiv preprint arXiv:1407.0414.
Toussaint, M. (2015). Logic-geometric programming: An optimization-based approach to combined task and motion planning. In IJCAI (pp. 1930-1936).

# Three-Way Bindings in Associative Recognition 

Hyungwook Yim (hyungwook.yim@newcastle.edu.au)<br>School of Psychology, The University of Newcastle, Australia

Adam F. Osth (adam.osth@unimelb.edu.au)<br>School of Psychological Sciences, The University of Melbourne, Australia

Vladimir M. Sloutsky (sloutsky. 1 @ osu.edu)
Department of Psychology, The Ohio State University, USA
Simon J. Dennis (simon.dennis@gmail.com)
School of Psychology, The University of Newcastle, Australia


#### Abstract

To avoid interference among similar memory traces it is required to form complex memory structures that include multiple components of the event, which helps one to distinguish one event from another. In a laboratory setting, these complex binding structures have been studied through a paradigm where one has to form a memory structure that includes two items and the context together (i.e., three-way binding). However, despite the long history of the theoretical concept, its importance, and the existence of the laboratory paradigm, three-way binding structures have only been examined in recall paradigms. Moreover, not all memory models consider the ability to form three-way binding structures as a default. Therefore, the current study examined the use and formation of three-way binding structures in an associative recognition paradigm. Results provide evidence that three-way binding structures are used during recognition, which implies that it is critical for memory models to properly represent them.


Keywords: episodic memory; recognition; three-way binding

## Introduction

LeSean McCoy is a running back in the National Football League (NFL) who started his career with the Philadelphia Eagles. After receiving several awards, and leading the team to the conference finals two times, McCoy was traded with Kiko Alonso, who was Buffalo Bills' linebacker with an "NFL Defensive Rookie of the Year" title. Knowing this fact, how would one later recall which team McCoy was playing before the trade? Even after restricting that there was a trade between McCoy and Alonso, simply recalling which team McCoy played for does not solve the answer since McCoy played for both the Eagles, and the Bills. Additionally retrieving the team that existed in the pre-trade period also does not help since both team existed before and after the trade. The only way to correctly retrieve this information is forming a coherent memory structure of [McCoy]-[Eagles]-[pre-trade] together, and later using the two cues together at retrieval (i.e., [McCoy] and [pre-trade]) as a compound cue.

Memory researchers call this kind of memory structure a three-way binding structure (Humphreys, Bain, \& Pike, 1989), and controlled laboratory experiments using recall paradigms provide evidence that adults robustly form these memory structures (e.g., Postman, 1964). Binding structures in memory research have been mainly studied using a paired associate learning paradigm, where participants are
given (usually two) lists of paired words to study, and are later tested. Especially the three-way binding structure could be examined when the word pairs in one list are repaired in another list (i.e., ABABr condition; Porter \& Duncan, 1953). As in the notation, in an ABABr condition the words in one list (i.e., first two letters 'AB') are identical in the other list (i.e., second two letters 'AB') but paired differently (i.e., last letter ' $r$ ' representing that the words are re-paired). Therefore the structure creates a strong interference between the two lists when trying to retrieve a piece of memory as in the NFL trading example. To correctly retrieve which words were paired in which list, one needs to use a three-way binding structure that includes the context of a specific list, and the two words together (e.g., [list_1]-[word_1]-[word_2]).

Correctly retrieving information from an ABABr condition could be thought as an exclusive or (XOR) problem since the arrangement of the ABABr condition is similar to the XOR operation (Wiles \& Humphreys, 1993). In the XOR operation, which is expressed by using the symbol $\underline{\vee}$, when zero is operated with zero the answer is one (i.e., $0 \underline{\vee} 0=1$ ), and zero operated with one is zero (i.e., $0 \underline{\vee} 1=0$ ). When one is operated with zero the answer is zero (i.e., $1 \underline{\vee} 0=0$ ), and one operated with one is one (i.e., $1 \bigvee 1=1$ ). Considering the first two terms as the cues at test (i.e., first term as the context, and second term as the item cue), and considering the answer of the XOR operation as the to-be-retrieved target, the process of retrieving an answer from the ABABr condition becomes identical to the XOR problem. The solution of the XOR has been well known to be impossible within a two dimensional plane that has independent inputs (e.g., Minsky \& Papert, 1969), and could be solved by increasing the dimension of the inputs such as using multiplicative (configural) coding of the inputs (Sloman \& Rumelhart, 1992). Similarly, the ABABr condition could not be solved with a two-way binding structure, in fact not even with multiple two-way binding structures (e.g., Humphreys, Bain, \& Pike, 1989), and requires a higher dimensional representation such as three-way binding structure.

Empirical evidence for the ability to form and use threeway binding structures implies that our memory system should be able to store three-way binding structures, and use
compound cues when retrieving these structures. However, not all theoretical accounts of episodic memory, and their computational models consider three-way binding structures as a default.

For example, the Search of Associative Memory (SAM; Gillund \& Shiffrin, 1984; Raaijmakers \& Shiffrin, 1981) theory does not consider three-way binding structures as a memory representation. The model assumes that memories are stored as association strengths between two components such as item and context, or item and item. Because these associations contain only two components, it is not possible to represent three-way bindings. On the retrieval side, SAM is unable to use compound cues when multiple cues are provided during retrieval. Rather it treats each cue independently and then combines the retrieved information of each cue by seeking overlapping information that are retrieved by each cue.

The Temporal Context Model (TCM) and its variants also do not employ three-way binding structures (Howard \& Kahana, 2002; Lohnas, Polyn, \& Kahana, 2015; Polyn, Norman, \& Kahana, 2009; Sederberg, Howard, \& Kahana, 2008). TCM employs a two-way binding between an item and the current context representation, which are stored in a matrix of item-context bindings. In TCM, the context is defined as a weighted sum of all past items where more recent items have a stronger weight in representing the current context. These assumptions even hold when word pairs are studied in a paired associate learning paradigm (e.g., Howard et al., 2009). TCM resembles SAM in that the representational structure is fundamentally restricted to two-way associations, and therefore cannot represent three-way bindings without extensive modification.

On the other hand, models that are capable of explaining the use and formation of three-way binding structures have slightly different assumptions. MINERVA 2 (Hintzman, 1984) encodes events as separate traces, in which context and item representations are concatenated into a single vector. Thus, items $\mathrm{A}\left(I_{A}\right)$ and $\mathrm{B}\left(I_{B}\right)$ in context $1\left(C_{1}\right)$ could be represented as $C_{1} \oplus I_{A} \oplus I_{B}$. Recognition decisions in the model are made by matching the cue vector to each memory trace and summing the similarities to produce a single index of memory strength that is then compared to a decision criterion.

MINERVA 2 is sensitive to three-way associations by virtue of a non-linearity at retrieval, where the similarity between a cue and a stored trace vector is raised to the third power. This enables the model to be more sensitive to conjunctions among studied elements, rather than the individual elements themselves. Consider if pairs A, B and C, D were studied in context 1 and pairs $\mathrm{A}, \mathrm{C}$, and $\mathrm{B}, \mathrm{D}$ were studied in context 2. A foil such as $C_{2} \oplus I_{A} \oplus I_{B}$ contains studied elements, but these were not all studied together. After the cubing process (increasing the similarity to the third power), the target $C_{1} \oplus I_{A} \oplus I_{B}$ receives a stronger match than the sum of the partial matches to $C_{2} \oplus I_{A} \oplus I_{B}$. Models in the Retrieving Effectively from Memory (REM; Criss \& Shiffrin, 2005; Shiffrin \& Steyvers, 1997) framework employ a sim-
ilar idea, where instead of a cubing process at retrieval the likelihood of feature match is multiplied across each element in the vector to calculate a likelihood ratio that the trace was a studied item, producing stronger matches to conjunctions of studied elements than studied elements distributed across different memory traces.

The MATRIX model (Humphreys, Bain, \& Pike, 1989; Osth \& Dennis, 2015; Pike, 1984) has a slightly different assumption about its storage representation by using tensor representations. Items and contexts are represented as vectors and the vectors are bound together using outer products to form a third-order tensor (i.e., $C_{1} \otimes I_{A} \otimes I_{B}$ ). Rather than having individual traces for each event, the MATRIX model sums all representations for each event into a single composite tensor. The tensor representation naturally allows the use of compound cues during retrieval by forming a second-order tensor (i.e., an outer product of an item and a context) to represent the compound cue. At retrieval, the cues are combined into a tensor and matched against the memory tensor using the dot product operation which produces a measure of memory strength that reflects how similar the combined cues match the contents of memory.

Interestingly, evidence for the use of three-way binding structures have only been examined with cued recall tasks (e.g., Porter \& Duncan, 1953; Shimamura, Jurica, Mangels, Gershberg, \& Knight, 1995; Yim, Dennis, \& Sloutsky, 2013) where a context and an item is given at test as a cue to recall the paired item during study (e.g., what was the word paired with 'apple' in 'the first list'?). To our knowledge, no recognition paradigm examined the use of three-way binding structures. A number of studies have examined the role of inter-item bindings in recognition memory by using an associative recognition task, where participants study pairs of items such as A-B, C-D, and E-F in a single list. At test, participants have to make discriminations between intact pairs, which were studied on the list (e.g., A-B), and rearranged pairs, which are studied items but in a novel arrangement (e.g., C-F).

Moreover, associative recognition tasks with the ABABr condition have been used in other studies with different purposes (e.g., Aue, Criss, \& Fischetti, 2012; Criss \& Shiffrin, 2005; Postman \& Stark, 1969). However, the design and goal of the studies do not focus on the use of three-way binding structures, mostly by not testing the word pairs from both lists together. This is especially relevant because the role of context has been generally neglected in models of associative recognition. For instance, the TODAM2 model of Murdock (1997) did not bind context to inter-item bindings due to an assumption that context was not employed in associative recognition as it was not relevant to the goal of the task.

Therefore, the current study examined whether three-way binding structures are used during recognition using a associative recognition task with the ABABr condition. We constructed an $A B A B r$ condition by presenting pairs in different contexts, such as A-B and C-D in context 1 and A-D and C-B
in context 2. At test, participants are given a pair and a context cue and asked if the pair occurred in the particular context. Successful discrimination in ABABr conditions would suggest a memory structure that is capable of representing three-way bindings. Demonstration of such bindings in both recognition and recall tasks would provide further evidence that memory models need to consider such representations.

## Experiment

The experiment tested whether people use three-way binding structures in an associative recognition paradigm. In addition to the ABABr condition, which requires a three-way binding structure for a correct recognition, we used an ABAC, and $A B C D$ condition. In the $A B A C$ condition, as in the notation, one item from each pair in the first context (i.e., A in ' $A B$ ') overlap with an item from each pair in the second context (i.e., A in 'AC') which results in a moderate overlap between the two contexts compared to the ABABr condition. At the minimum, it is required to form two two-way binding structures (i.e., item-to-item, and context-to-item) for a correct retrieval (Humphreys, Bain, \& Pike, 1989). Using the same scheme, in the ABCD condition there are no overlapping items between the contexts which results in two contexts with unique items. Since there is no overlap between the two contexts, a correct retrieval only requires a single item-to-item binding at the minimum (i.e., item-item, or cue-target binding). Therefore the level of overlap increases from the ABCD condition to the ABABr condition. Moreover, a more complex binding structure is required for a correct retrieval at test as the level of interference increases. Previous studies using a recall paradigm showed a negative correlation with the level of interference and performance (e.g., Yim, Dennis, \& Sloutsky, 2013). Therefore, the additional two conditions will serve as a reference point for the performance on the ABABr condition.

As part of the design, we defined context as the identity of the speaker that presents the word pair instead of using the temporal order of the 'list' (i.e., first list, and second list) as in previous studies. Embedding the context in the trial enables us to intermix different context in the study phase, and prevents the participants from using the temporal cue. A weakness of previous ABABr designs which use two successive study lists as their contexts is that the first list is naturally expected to have weaker memory strength than the second list. This enables participants to infer that an item is from the first list due to its weaker memory strength even with a two-way binding structure (e.g., Lohnas, Polyn, \& Kahana, 2015). By eliminating the memory strength confound, our design ensures that participants require a three way binding structure to achieve above chance performance in the task.

## Methods

Participants Forty-three undergraduate students at The University of Newcastle participated for course credit (36 females, $M=25.12$ years, $S D=9.87$ ).

Materials The stimuli were video clips of a speaker saying a word. There were nine female and nine male speakers, and each speaker had its own unique background scene (see Figure 1 A ). All words were high frequency words consisting of 54 adjectives, and 63 nouns. Most of the words were selected from the MacArthur-Bates Communicative Development Inventory through the Wordbank repository (Frank, Braginsky, Yurovsky, \& Marchman, 2016).
Procedure Participants were tested individually in the laboratory. There were nine blocks where each block had a study phase followed by a retention interval and a test phase. In the study phase, participants were told that they will be seeing two speakers each presenting different word pairs one at a time. They were also told to exactly remember who said which words together since they will be tested later. Each trial started with a fixation cross for 500 msec followed by a blank screen of 500 msec and a video clip presenting a word pair, which lasted for approximately 3400 msec (see Figure 1B). In all blocks, one of the speakers was always a female, and the other a male. Also, the video clips were presented on one side of the screen throughout the experiment depending on the speaker's sex (e.g., female on the left side, male on the right side), but was randomized across participants. There were eight trials in each study phase consisting of the $A B C D$, ABAC , and ABABr structures (see Table 1 for an example). The first word was always an adjective and the second word was always a noun. The presentation order of the eight trials corresponding for each structure were randomized every block.

Table 1: An example of the eight trials in each study phase. Each triplet in the Trials column represents the speaker's sex (M: male, F: female), first word (adjective), and second word (noun) in order. There are four trials for the ABABr structure, two trials each for the $A B A C$ and $A B C D$ structures.

| Condition | Trials |  |
| :---: | :--- | :--- |
| ABABr | $[\mathrm{M}]-$ green - hand | $[\mathrm{F}]-$ green - toy |
|  | $[\mathrm{M}]-$ hot - toy | $[\mathrm{F}]-$ hot - hand |
|  | $[\mathrm{M}]-$ empty - cat | $[\mathrm{F}]-$ empty - shoe |
| ABCD | $[\mathrm{M}]-$ tall - rain | $[\mathrm{F}]-$ quiet - ball |

During the 60 second retention interval participants were presented with two groups of dots on each side of the screen, and were told to choose the side that had more dots. After a 500 msec fixation (+++) the two groups of dots were presented for 250 msec followed by a random color dot mask, which was presented until a response was made. The number of dots varied between 10 and 40 where the ratio of the two numbers were randomly chosen among the following range:


Figure 1: Design and stimuli used in the experiment. (A) an example of the videos used in the experiment, (B) an example of the study phase, (C) an example of the test phase.
1.0991-1.1915, 1.1915-1.2917, 1.3302-1.4421, and 2.29062.4833 (adapted from Halberda \& Feigenson, 2008).

In the test phase, participant were presented with a video clip as in the study phase and were asked whether it was an old video that they saw during the study phase (i.e., same speaker saying the exact same word pair), or a new one (see Figure 1C). Responses were collected using a computer mouse by clicking the corresponding image on the screen. There were 18 test trials consisting of eight old videos, eight new videos that had the speaker swapped (re-arranged pairs), and two new videos that had a new word pair spoken by the female speaker and the male speaker (lure pairs). The words in the lure pair did not appear in the study phase, and the trials were randomized every block.

Presentation of all stimuli was controlled using MATLAB with Psychtoolbox-3 (Kleiner, Brainard, \& Pelli, 2007) equipped with a 22 inch monitor, and an individual headphone. The combination of the word pairs, and speakers were randomized across participants.

## Results

We analyzed and compared each condition regarding hit rate (HR), false alarm rate (FAR), $d^{\prime}$, and correct reaction time (RT). As shown in Figure 2A, HR was the highest for the ABAC condition ( $M=.76, S D=.14$ ) followed by the ABCD condition ( $M=.71, S D=.15$ ) and ABABr condition $(M=.67$, $S D=.11$ ). A linear mixed-effects model ${ }^{1}$ with subject as a random factor (random intercept model) supported the effect of Condition ( $\chi^{2}(2)=16.70, p<.001$ ), while a Tukey post-

[^547]hoc test only provided evidence for the difference between the ABAC condition and the other two condition (ABAC ABABr: $p<.001$; ABAC - ABCD: $p=.057$; ABCD ABABr: $p<.001$ ). Similarly, the FAR measured by the re-arranged pairs was the highest for the ABABr condition $(M=.40, S D=.15)$ followed by the $\mathrm{ABCD}(M=.25, S D$ $=.16)$, $\operatorname{ABAC}(M=.23, S D=.15)$, and the lures $(M=.17$, $S D=.16$; see Figure 2B). A linear mixed-effects model with subject as a random factor supported the effect of Condition $\left(\chi^{2}(2)=94.62, p<.001\right)$, while a Tukey post-hoc test provided evidence for the difference between the ABABr condition and the other conditions ( $p<.001$ ), and between the lures and the $\operatorname{ABAC}(p=.046)$ and ABCD conditions $(p=$ .001). There was no evidence for a difference between the $A B C D$ and $A B A C$ conditions. Unlike previous studies using a recall paradigm, where the $A B C D$ condition shows a better performance than the ABAC condition, the current results show a higher HR and lower FAR for the ABAC condition than the ABCD condition.

We measured discrimination using $d^{\prime}$ by comparing old pairs against rearranged pairs $\left(d_{\text {rearranged }}^{\prime}\right)$. As shown in Figure 2 C the ABABr condition showed the lowest ( $M=.74$, $S D=.64$ ), followed by the ABCD condition ( $M=1.30, S D$ $=.71)$, and ABAC condition $(M=1.53, S D=.80)$. Also, all conditions showed an above chance performance ( $t \mathrm{~s}(42)$ $>7.60$, Bonferroni adjusted $p \mathrm{~s}<.001$, Cohen's $d>1.16$ ), as evidenced by $d^{\prime}$ scores above zero in each condition. A linear mixed-effects model with subject as a random factor supported the effect of Condition $\left(\chi^{2}(2)=59.12, p<.001\right)$, while a Tukey post-hoc test provided evidence for the difference between the ABABr condition and the other two conditions (ABAC: $p<.001$; ABCD: $p<.001$ ), and between the ABAC and the ABCD condition $(p=.028)$.


Figure 2: Results of the Experiment. (A) hit rate (HR), (B) false alarm rate (FAR), (C) $d^{\prime}$ of distinguishing the re-arranged pairs from studied (old) pairs ( $d_{\text {rearranged }}^{\prime}$ ), (D) $d^{\prime}$ of distinguishing the lures from non-lures ( $d_{\text {lure }}^{\prime}$ ), and (E) correct reaction time (RT). Error bars represent $\pm 1$ SEM.

Discriminability of lures (new pairs) from intact pairs ( $d_{\text {lure }}^{\prime}$ ) was also calculated (see Figure 2D). The ABABr condition showed the lowest $d_{\text {lure }}^{\prime}(M=1.49, S D=.73)$, followed by the ABAC condition ( $M=1.77, S D=.88$ ), and ABCD condition ( $M=1.60, S D=.79$ ), while all conditions showed an above chance performance $(t \mathrm{~s}(42)>13.18$, Bonferroni adjusted $p \mathrm{~s}<.001$, Cohen's $d>2.01$ ). A linear mixed-effects model with subject as a random factor showed supported the effect of Condition ( $\chi^{2}(2)=17.87, p<.001$ ), where evidence for difference was found between the ABAC condition and the other two conditions (Tukey post-hoc test, ABABr: $p$ $<.001$; ABCD: $p=.022$ ), but not between the ABABr and ABCD conditions ( $p=.198$ ).

RT was first pre-processed by taking the median value of each condition for each participant. RT was the slowest for the ABABr condition ( $M=988 \mathrm{msec}, S D=299 \mathrm{msec}$ ) followed by the ABAC condition ( $M=974 \mathrm{msec}, S D=344$ msec ), and ABCD condition ( $M=866 \mathrm{msec}, S D=267 \mathrm{msec}$; see Figure 2E). A linear mixed-effects model with subject as a random effect showed a statistically significant effect of Condition $\left(\chi^{2}(2)=9.11, p=.011\right)$, where a Tukey post-hoc test provided evidence for the difference between the ABCD condition and the other two conditions (ABAC: $p=.035$; ABABr: $p=.015$ ), but not between the ABAC and ABABr conditions ( $p>.25$ ).

## General Discussion

In the current study we examined whether three-way binding structures are formed and used in a recognition task. Even though three-way binding structures are crucial in everyday life since items could be easily re-paired in different contexts, previous studies have only examined the structure with recall paradigms. The results most importantly showed that participants reliably use three-way binding structures during an associative recognition task. Based on both $d_{\text {rearranged }}^{\prime}$, and $d_{\text {lure }}^{\prime}$ measures, participants showed robust above chance level performance in the ABABr condition, which indicates that three-way bindings were formed and used. The overall pattern was similar to previous findings using recall tasks (e.g., Yim, Dennis, \& Sloutsky, 2013) where the ABABr condition
showed above chance accuracy while less accurate than the other two conditions, and required more time to respond due to greater interference. Our results also extend on previous studies by using two contexts that are inter-mixed within a single list. Past studies which have employed two temporally separated study lists allow for the possibility that participants could infer list membership on the basis of a difference in memory strength between the two lists.

One interesting difference from previous results was the performance in the ABAC condition. Previous studies show a better performance in the ABCD structure compared to the ABAC condition since there are less interference in the $A B C D$ condition. However, the ABAC condition showed the best performance in the current results. One possible explanation could be a speed-accuracy trade off since it took longer to respond in the ABAC condition than in the ABCD condition while the accuracy was higher. However, it will be hard to disentangle the cause with only relying on the current data set.

The evidence of using three-way binding structures in both recall and recognition tasks implicate that models that do not represent three-way binding structures should be re-examined (e.g., Gillund \& Shiffrin, 1984; Howard \& Kahana, 2002; Lohnas, Polyn, \& Kahana, 2015). Our results also cast doubt on the proposal that the associative recognition task does not employ a context representation (Murdock, 1997). However, future work may be needed to discriminate between the existing classes of models that are capable of predicting above chance ABABr performance. For instance, our results do not discriminate between multiple trace models such as MINERVA 2 (Hintzman, 1984) and REM (Shiffrin \& Steyvers, 1997) which can predict above chance $A B A B r$ performance by virtue of their non-linear similarity metrics at retrieval, and the class of tensor models (Humphreys, Bain, \& Pike, 1989; Osth \& Dennis, 2015) which employ explicit three-way bindings as third-order tensors. Another interesting possibility for future work concerns the time course of when three way bindings are accessible. Although there are studies showing that associative information arrives after information (e.g., Gronlund \& Ratcliff, 1989), further research should examine these
possibilities with three-way bindings.

## Acknowledgments

This research was supported under Australian Research Council's Discovery Projects funding scheme to SJD (project number DP150100272). We thank Emma K. Lindsay and Emma Kronja for their help on collecting data.

## References

Aue, W. R., Criss, A. H., \& Fischetti, N. W. (2012). Associative information in memory: Evidence from cued recall. Journal of Memory and Language, 66(1), 109-122.
Bates, D., Mächler, M., Bolker, B., \& Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
Criss, A. H., \& Shiffrin, R. M. (2005). List Discrimination in Associative Recognition and Implications for Representation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(6), 1199-212.
Frank, M. C., Braginsky, M., Yurovsky, D., \& Marchman, V. A. (2016). Wordbank: an open repository for developmental vocabulary data. Journal of Child Language, (pp. 1-18).
Gillund, G., \& Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. Psychological Review, 91(1), 1-67.
Gronlund, S. D., \& Ratcliff, R. (1989). Time course of item and associative information: implications for global memory models. Journal of Experimental Psychology. Learning, Memory, and Cognition, 15(5), 846-858.
Halberda, J., \& Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. Developmental Psychology, 44(5), 1457-65.
Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. Behavior Research Methods, Instruments, \& Computers, 16(2), 96-101.
Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., \& Datey, A. V. (2009). Bridging the gap: transitive associations between items presented in similar temporal contexts. Journal of Experimental Psychology: Learning, Memory and Cognition, 35(2), 391-407.
Howard, M. W., \& Kahana, M. J. (2002). A distributed representation of temporal context. Journal of Mathematical Psychology, 46, 269-299.
Humphreys, M. S., Bain, J. D., \& Pike, R. (1989). Different Ways to Cue a Coherent Memory System: A Theory for Episodic, Semantic, and Procedural Tasks. Psychological Review, 96(2), 208-233.
Kleiner, M., Brainard, D., \& Pelli, D. (2007). Whats new in Psychtoolbox-3? Perception, 36, 14.
Lohnas, L. J., Polyn, S. M., \& Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. Psychological Review, 122(2), 337-363.

Minsky, M. L., \& Papert, S. A. (1969). Perceptrons. Cambridge, MA: MIT Press.
Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). Psychological Review, 104(4), 839-862.
Osth, A. F., \& Dennis, S. (2015). Sources of interference in item and associative recognition memory. Psychological Review, 122(2), 260-311.
Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. Psychological Review, 91(3), 281-294.
Polyn, S. M., Norman, K. a., \& Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. Psychological Review, 116(1), 129-156.
Porter, L. W., \& Duncan, C. P. (1953). Negative Transfer in Verbal Learning. Journal of Experimental Psychology, 46(1), 61-64.
Postman, L. (1964). Studies of Learning to Learn II. Changes in Transfer as a Function of Practice. Journal of Verbal Learning and Verbal Behavior, 3, 437-447.
Postman, L., \& Stark, K. (1969). Role of Response Availability in Transfer and Interference. Journal Of Experimental Psychology, 79(1P1), 168-\&.
Raaijmakers, J. G., \& Shiffrin, R. M. (1981). Search of associative memory. Psychological Review, 88(2), 93-134.
Sederberg, P. B., Howard, M. W., \& Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. Psychological Review, 115(4), 893-912.
Shiffrin, R. M., \& Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. Psychonomic Bulletin \& Review, 4(2), 145-166.
Shimamura, A. P., Jurica, P. J., Mangels, J. A., Gershberg, F. B., \& Knight, R. T. (1995). Susceptibility to Memory Interference Effects following Frontal Lobe Damage: Findings from Tests of Paired-Associate Learning. Journal of Cognitive Neuroscience, 7(2), 144-152.
Sloman, S., \& Rumelhart, D. (1992). Reducing Interference in Distributed Memory Through Episodic Gating. In A. F. Healy, S. M. Kosslyn, \& R. M. Shiffrin (Eds.) From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes, (pp. 227-248). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Wiles, J., \& Humphreys, M. S. (1993). Using Artificial Neural Nets to Model Implicit and Explicit Memory Test Performance. In P. Graf, \& M. E. Masson (Eds.) Implicit Memory: New Directions in Cognition, Development, and Neuropsychology, (pp. 141-166). New York, New York, USA: Lawrence Erlbaum Associates, Inc.
Yim, H., Dennis, S. J., \& Sloutsky, V. M. (2013). The Development of Episodic Memory: Items, Contexts, and Relations. Psychological Science, 24(11), 2163-2172.

# Emotional and Cognitive Interest: How Creating Situational Interest Affects Learning with Multimedia 

Angela Yoo (ahyoo79@gatech.edu)<br>Department of Psychology, 654 Cherry St<br>Atlanta, GA 30332 USA

Richard Catrambone (richard.catrambone@psych.gatech.edu)<br>Department of Psychology, 654 Cherry St<br>Atlanta, GA 30332 USA


#### Abstract

Situational interest is the positive affect and sustained attention triggered by particular contexts (Hidi \& Renninger, 2006). Some studies show interesting information enhances learning while others find it hinders learning, producing the seductive detail effect. Limited evidence suggests the seductive detail effect is weakened if emotionally interesting information is relevant to main ideas. The present research shows the seductive detail effect occurs only under certain conditions. Harp and Mayer (1997) proposed that generating cognitive, rather than emotional, interest is more effective for improving learning by cueing relationships among concepts for easier processing. Hidi and Renninger (2006) argue distinguishing between the emotional and cognitive might be artificial. Present research found benefits from cognitive interest but no support as to whether cognitive interest is necessarily a distinct type of interest from emotional interest. There were some challenges with operationalizing cognitive interest, as well as validating strategies utilized to manipulate cognitive interest levels.


Keywords: learning, instruction, situational interest, cognitive interest, seductive detail effect

## Introduction

The idea of creating an interesting experience to enhance learning is at the forefront of educational issues, particularly with science, technology, engineering, and math (STEM) education and massive open online courses (MOOCs) (National Governor's Association, 2011; Norman, 2013). However, ensuring that interesting features are not detracting from the primary instructional purpose is difficult. This challenge of mediating a desire for interest and engagement with the need to produce effective learning outcomes is mirrored in educational and cognitive psychology, with research finding contradictory or mixed results for increasing interest levels in students (Rey, 2012).

## Interest Learning Theory

Empirical support exists for interest learning theory - the idea that the more interesting learning material is the more likely a student is to learn and remember the information (Ainley, Hidi. \& Berndorff, 2002; Schiefele, 1991; Schraw, Bruning, \& Svoboda, 1995). This is especially true when measuring elaborative processing and comprehension, rather than simple recall or recognition (Schiefele, 1991). Statistically significant correlations have been found between interest and
text order choice, interest and positive affect regarding the text, affect and time spent reading, and ultimately between persistence and test scores (Ainley et al., 2002). This implies that students' interests and enjoyment lead to more time and effort spent on a text, enabling them to learn more effectively.

Interest is typically divided into three categories: individual, topic, and situational (Ainley, Hidi, \& Berndorff, 2002). Individual interest and topic interest, which both involve attributes of a person, are not considered here because the focus of the present research is to examine manipulations of learning material. Situational interest refers to environmental stimuli and the general structural features of a situation, such as organization of information, unexpectedness, text cohesion, use of concrete ideas, and intensity of triggered emotions (Ainley et al., 2002; Schraw et al., 1995). Situational interest is briefer in duration than the other forms of interest but is still characterized by positive affect and sustained attention toward the material.

## Seductive Details

Not all research supports interest-based learning. Garner, Alexander, Gillingham, Kulikowich, and Brown (1991) found that efforts to artificially induce interest, particularly with extraneous details, divert the learner's attention and reduce the ability to recall relevant information. They termed these interesting but distracting details "seductive." These seductive details were highly memorable to participants on tests of learning compared to details that were of high importance but lower interest and also could lead to poorer recall and transfer scores (Harp \& Mayer, 1991).

In a meta-analysis of findings on seductive details, nearly two-thirds of the studies included in the analysis supported fully or partially the detrimental effects of seductive details (Rey, 2012). The data in aggregate appear to demonstrate up to a small to medium effect size $(\mathrm{d}=0.3)$ for the reduction in recall and a medium effect size $(d=0.48)$ for transfer of knowledge tasks.

Still one-third of the experiments seductive details did not hurt and sometimes even improved learning (Rey, 2012). These studies show that particular types of interesting detail, the learning domain, time limits, and amount of cognitive load can temper the distracting effects of seductive details. One example of how mitigating factors result in mixed data comes from the study by Garner et al. (1991). Researchers
found that moderately relevant and moderately interesting details were recalled more frequently by participants. Although the most important details, which were rated uninteresting, were not remembered as well, the finding that some of the germane details could be recalled if considered interesting gives interest learning theory some merit.

Studies considering illustrations that accompany text indicate that relevance is the key difference in determining whether details enhance or reduce learning. In the case of reading text, purely decorative illustrations do not benefit the understanding of the content of the text, but illustrations that depict the information, help to organize or interpret information, or provide memory devices for learning can moderately benefit the retention of that information (Carney \& Levin, 2002). Any lack of enhanced learning with purely decorative images could possibly be moderated by other variables, such as the learner's prior knowledge (Magner, Schwonke, Aleven, Popescu, \& Renkl, 2014). While decorative illustrations that incite situational interest can distract learners with little prior knowledge, illustrations can enhance learning for students with more prior knowledge.

## Cognitive Interest

Instead of adding emotional interest with seductive details, Harp and Mayer (1997) suggest using cognitive interest as an alternative way to enhance learning. Harp and Mayer argue that emotionally interesting seductive details do little to help cognitively. To generate cognitive interest, instruction on a topic should signal the underlying structure of relationships of relevant concepts. Such strategies would include identifying main ideas, relating information to prior knowledge, and linking related topics. The idea is that, if topics are presented in a way that learners find easier to understand, the topics will seem more interesting. When comparing performance on both recall and problem-solving transfer, Harp and Mayer (1997) found that students scored higher marks if using text and illustrations that helped to organize and explain a topic (i.e. cognitively interesting details) than if the text and illustrations included extraneous, irrelevant content (i.e. seductive details).

A criticism of cognitive interest comes from a neuroscientific perspective (Hidi \& Renninger, 2006). Separating affective from cognitive processes and, therefore, emotional from cognitive interest is arguably an artificial distinction because of the function of the lateral hypothalamus. The lateral hypothalamus plays a major role in seeking behavior and is responsible for inducing feelings of interest and curiosity (Panksepp, 1998). Hidi and Renninger (2006) claim that, regardless of the stimulus that triggers interest or regardless of whether the person is cognitively processing or affectively responding to the stimulus, the lateral hypothalamus is activated in the brain.

## Current Investigation

The present study focused on multimedia, specifically educational videos. Video was selected as the educational medium because it would be more directly applicable and
relevant to the growing use of technology in the classroom with MOOCs and other online formats. Participants in all conditions watched a video on human digestion that varied by condition.

Relevance of Interesting Details The purpose of the main study was twofold: to explore the mitigation of the seductive detail effect through increased relevance of interesting details to main ideas and to examine the distinction between cognitive and emotional interest.

Many of the previous studies have measured learning by using free recall after learners had been exposed to the learning material (Garner et al, 1991; Harp \& Mayer, 1997; Harp \& Mayer, 1998; Schiefele, 1991; Schraw et al., 1995). Free recall might not account for prior knowledge sufficiently or for variables, such as writing abilities or motivation to write large amounts of text, that could affec performance (Schiefele, 1991). Due to this possibility that free recall alone could be an insufficient measure, a post-test score also measured learning. A pretest score was used to explore the possibility of prior knowledge as a covariate for the experimental groups. Both the pre- and post-test consisted of multiple choice, fill-in-the-blank, and short answer questions, but the questions were different for the two forms. Some of the questions asked students to identify a concept through recall or recognition while other questions measured required students to explain causal relationships and make inferences. The variety of questions was used to overcome any potential issues with using only free response or essay questions.

By creating situational interest with statements that were emotionally engaging but less similar to the main idea of the video, the study was expected to replicate the findings of seductive detail effect researchers (Garner et al., 1991; Harp \& Mayer, 1997; Mayer et al., 2008; Rey, 2012). However, performance was predicted to improve with emotionally interesting details if relevance to the content of the videos was controlled. When the interesting details did not contain less relevant additional information, but instead contained similar information as the main idea, the distracting effect seen with seductive details was not expected to be found.

Distinguishing Emotional and Cognitive Interests The second aim of this research was to examine whether emotional interest and cognitive interest are separate constructs. The distinction between cognitive and emotional aspects contradicts the definition of interest, which necessitates both affective and cognitive changes. Therefore, the phrase "cognitive interest" is used in the remainder of this document to refer to interest incited by strategies used by Harp and Mayer (1997) emphasizing the cognitive processes of interest. "Emotional interest" will refer to strategies emphasizing affective processes, as exemplified by most interest learning researchers, such as Schiefele (1991). The use of "cognitive interest" and "emotional interest" does not indicate necessarily a true differentiation between the cognitive and emotional processes of interest.

For the possibility to remain that emotional interest and cognitive interest are indeed the same phenomenon, as proposed by Hidi and Renninger (2006), there should not be any interactive effects with emotional interest conditions and cognitive interest conditions. Lack of interaction, however, would not necessarily indicate that cognitive and emotional interests are undeniably the same, but the hypothesis that they are the same would remain tenable. On the other hand, the presence of significant interactions possibly would suggest they are indeed different phenomena. Due to the evidence from Harp and Mayer (1997), learning outcomes were expected to demonstrate an interactive effect. Learning was predicted to improve with more cognitive interest compared to conditions with low interest material or material with seductive details by increasing the efficiency of processing the material. However, when there was emotionally interesting material that was also relevant to the learning material, adding cognitive interest should not produce any additional benefits relative to what was already provided by details that were emotionally interesting and relevant.

## Methods

Participants were 93 undergraduates in introductory psychology courses. The experiment was a $3 \times 2$ factorial design with six videos total - one for each condition. Students were randomly assigned to each of the six conditions (Table 1). One hour-long session was required per participant.

Table 1: Participants in each experimental condition.

|  | Emotional Interest |  |  |
| :--- | :---: | :---: | :---: |
|  | Low | Relevant | Irrelevant <br> (Seductive Details) |
| Low Cognitive <br> Interest | 16 | 16 | 16 |
| High Cognitive <br> Interest | 15 | 15 | 15 |

During the experimental session, students completed a preliminary questionnaire and an exit questionnaire. The first questionnaire given before the video viewing contained basic demographic questions including age, year in school, major, GPA, and SAT scores; Likert-type items regarding interest levels on science, biology, and anatomy; and a prior knowledge assessment. The exit questionnaire completed after viewing that video was divided into two sections. The first section repeated the same Likert-type items from the preliminary questionnaire and included a free response prompt. The final section contained items about the content of the learning material. One such item was: "True or false: Chemical digestion begins in the mouth." Thirteen of the post-test questions required students to recall or recognize information presented directly in the videos. Twelve of the items required students to make inferences based on the information presented in the videos. The tests were piloted to determine validity and appropriate levels of difficulty.

Videos ranging from 11.5 to 12 minutes in length and involving screen capture and narration were used in each of the six conditions. The types of illustrations that were used in the screen capture for all conditions were representational pictures. Representational pictures are those that simply depict the concepts being described in the audio but do not provide any type of organizational support for the concepts (Carney \& Levin, 2002). Three of the videos did not include any features that would add cognitive interest and manipulated only the amount and relevance of emotional interest. The first of these three videos for the low cognitive interest conditions contained only basic facts about human digestion and served as the control (Table 2). In the highly relevant emotional interest condition, the video narrative supplemented the information in the control video with facts that had been rated as more interesting and more relevant to the main ideas in a previous pilot study. The video for the irrelevant emotional interest condition - or the seductive detail condition - included anecdotes and facts that had been rated interesting but less relevant to the video's main idea.

Table 2: Sample texts from video scripts.

| Control | Relevant <br> Emotional <br> Interest | Irrelevant Emotional <br> Interest <br> (Seductive Details) |
| :--- | :--- | :--- |
| In addition to | Because its bile- <br> producing bile, the <br> liver functions to <br> looding and <br> job-filtering <br> filter your <br> bloodstream, store <br> some vitamins and <br> important, the <br> liver is the <br> minerals, and help <br> to breakdown <br> some of the excess <br> hormones in the <br> blood. | largest human <br> organ by weight <br> and regenerates <br> all its cells <br> within 30 days. |
| is a shortage of there <br> donors and a great <br> need for donated <br> organs, researchers <br> are experimenting <br> with 3D-printed <br> livers for use in <br> transplant patients. |  |  |

A second set of videos used the same scripts as the previous set emotional interest videos but also included explanative summaries to provide cognitive interest. Explanative summaries were used in the study by Harp and Mayer (1997) to create cognitive interest. These explanative summary paragraphs of 3-6 sentences highlighted major components of the human digestive system, important steps involving these components, and some of the causal processes that occur. Because all the information was presented aurally, the explanative summaries were also presented through narration at 6 different points within each video.

Learning was measured with a free recall exercise and a post-test. For the free recall assessment, participants were instructed to write everything they could remember from the video about the digestive process. Raters awarded a point for each complete and correct statement and 0.5 points for a partially correct or partially complete statement. The free recall assessments were scored by the researcher and a second rater. The interrater reliability for the scores was calculated to be $\alpha=0.96$. For the pre- and post-tests, raters deducted a point for incorrect answers. No partial points were deducted.

The raw score was then converted to a percentage score for the post-test. The free recall score was not converted to a percentage because it is scored on a basis of accumulating points, unlike post-tests scored on a point-deduction basis. Even if an attempt was made to convert recall scores to percentages, comparisons would have been difficult with mean recall scores around $12 \%$ and mean post-test scores of $57 \%$. Comparing patterns of results across the 2 dependent measures, rather than considering comparably scaled means, was more important to underscore the reasons for conflicting results in previous studies.

## Results and Discussion

For the free recall exercise, SAT math scores ( $F=5.92, p=$ 0.02 ), pre-test scores ( $F=8.54, p<0.01$ ), and biology interest levels ( $F=7.13, p=0.01$ ) were found to covary significantly with free recall scores. Pre-test scores ( $F=10.66, p<0.01$ ) and interest in biology ( $F=7.87, p=0.01$ ) also significantly covaried with the post-test scores. Once the indicated covariates were considered, MANCOVAs show main effects for both cognitive ( $F=9.32, p<0.01$ ) and emotional $(F=$ 4.65, $p=0.02$ ) interest for free recall. A significant main effect for only cognitive interest ( $F=4.44, p=0.04$ ) resulted
for the post-test scores. Interactions were not significant for either the free recall scores ( $F=0.65, p=0.53$ ) or post-test scores ( $F=0.30, p=0.74$ ).

The results replicate the findings on cognitive interest of Harp and Mayer's 1997 study (Figure1). Compared to the conditions without cognitive interest, participants learning from the high cognitive interest materials had higher free recall ( $M=8.52, S D=4.82$ ) and post-test scores ( $M=59.02$, $S D=16.95$ ) when compared to free recall scores $(M=6.84$, $S D=3.40)$ and post-test scores ( $M=55.83, S D=17.66$ ) of those in the control condition (Figures 2 and 3).

Such results support the idea that material that is easier to process for learning also provides some level of situational interest - cognitive interest being a form of situational interest - and contributes to improved learning. One possible problem with this interpretation, however, is that adding explanative summaries to create cognitive interest consequently added a second opportunity to hear information that was being presented in the videos. Repeated exposure to learning material can improve performance on immediate recall (Tulving, 1967). Further studies investigating whether the improvement in learning can be attributed to the frequency of exposure or to the cognitive interest that arises from the clarity of organization and concepts is necessary.



Figure 1: Effects of adding explanative summaries for cognitive interest on free recall scores and post-test scores.


Figure 2: Effects of relevance of interesting material on free recall scores and post-test scores.

The results for emotional interest and the relevance of emotionally interesting details were mixed. Pairwise comparisons with Sidak-adjusted p values were statistically significant when comparing free recall results between the control groups and the seductive detail group ( $t=2.80, p=$ 0.02 ) (Figure 2). The control group's scores ( $M=8.10, S D=$ 5.38) were greater on average than the scores for the seductive detail condition ( $M=7.00, S D=3.77$ ). This result lends additional support to the seductive detail findings of Garner et al. (1991), Harp and Mayer (2006), and Mayer et al. (2008). Details in learning material that are not relevant to the main learning object appear to be harmful for learning when learning is measured by the ability to recall information. However, there were no other significant differences $(t=2.00, p=0.15)$ between the control and the relevant interest groups ( $M=7.85, S D=3.24$ ) or between the seductive detail and relevant interest groups ( $t=2.00, p=$ 0.74 ). These results could suggest that, while relevant emotional interest can compensate for any distracting aspects of seductive details, the amount of interest generated in the relevant emotional interest condition is not enough to be advantageous compared to low-interest learning material. However, due to the null results for relevant interest, making any conclusive statements is difficult.

Perhaps any effects of including such emotionally interesting information would have impacted affective states more so than cognitive processes. Further work using a variety of strategies to measure affect and cognition is necessary to determine what effects these details have.

As predicted, the format of assessment appears to affect the measure of learning outcomes (Figure 2). Recall assessments consisting of writing paragraphs tend to be more difficult to write, require more information to be encoded, and produce worse scores compared to assessments that rely on recognition (Tversky, 1973). Because the post-test questions relied on a combination of both recall and recognition items requiring only short answers or marking answer selections, the detrimental effects of seductive details were no longer observed. No discernible effects ( $F=1.03, p=0.37$ ) were found for overall post-test scores across the control ( $M=$ 56.77, $S D=19.33$ ), relevant emotional interest ( $M=58.45$, $S D=14.78$ ), and seductive detail ( $M=56.90, S D=18.00$ ) conditions. The difference between the free recall results and the post-test questionnaire could imply that seductive details are more harmful when deeper encoding is required. In contrast, when less encoding is needed for recognition tasks, seductive details seem to have less of an impact.


Figure 3: Effects of relevance of interesting material on free recall and post-test scores with and without cognitive interest.

Because interactions between cognitive and emotional interest were not statistically significant (Figure 3), the current study was unable to provide any further support to the idea that the two phenomena are distinct constructs. Firstly, the F values $(F=0.65, p=0.53$ for free recall; $F=0.30, p=$ 0.74 for post-test scores) were less than one for the interactions, suggesting other variables at play that could cause the relationships to appear nonlinear and lead to F values smaller than 1. Secondly, Hidi and Renninger's (2006) proposal that emotional and cognitive interests are part of the same construct remains tenable, as does Harp's and Mayer's (1997) idea that a dichotomy exists. However, reassessing the premise of the Harp and Mayer study (1997) provides some indications as to why finding a distinction would be difficult
using their methods. While the positive results found by Harp and Mayer (1997) seemed promising for learning based on cognitive interest, the separation between cognitive and emotional interest is problematic. Interest is defined as having both affective and cognitive dimensions, according to Hidi and Renninger (2006). To distinguish the two components would suggest that either an entirely different construct is being studied or that an essential component was neglected when interpreting the results of the study.

The latter possibility could be the case for the Harp and Mayer study (1997). When participants were asked to rate how interesting the learning material was, the average rating for the passage containing cognitively interesting details was not significantly different from the passage with emotionally
interesting details, showing that both passages were enjoyed equally. The interest ratings in all the conditions were greater than 7 out of a possible 10 points. These results demonstrate that positive affect was experienced in the cognitive interest conditions equal to that in the emotional interest conditions. Even though the researchers conducted a subsequent experiment to have participants distinguish interest based on "entertainment" as an approximation of emotion and interest based on how much the text supported the learner's understanding, participants initially interpreted "interest" as encompassing both these dimensions. Due to the questionable premise of manipulating only cognitive components of interest, in addition to the previously discussed problems of repeated exposure with explanative summaries, finding nonsignificant results for the interaction between emotional interest and cognitive interest is unsurprising.

## Conclusion

Although positive results for relevant interesting details and negative results for seductive details were expected, statistically significant differences were found only for seductive details. The seductive detail effect, however, did not appear with the post-test and could indicate that irrelevant details are problematic only when recall tasks require more encoding.

The current study was unable to demonstrate a distinction between the constructs of emotional and cognitive interest. Therefore, the lack of significant results for any interactive effects should not be interpreted as indicating that the two constructs are the same or different.

The difficulty in developing appropriate manipulations and measures serve to emphasize the importance of careful planning in the design of instructional material. Generating interest and possibly the right type of interest to increase learning outcomes is a challenge. The results of this experiment and the both corroborating and conflicting results in the literature illustrate the need for intentionality in the development of learning content. Failure to make the appropriate considerations can lead to unintended results or no effects for the attempts made to improve instruction.

There remains a need for a more substantive basis for beliefs that interest is a necessary motivating factor for learning. Additional studies with improved materials are needed to further explore whether the relevance of highinterest materials can mitigate the detrimental effects of seductive details and support interest learning theory. Finding more empirical evidence would support popular recommendations for stimulating interest in improving educational outcomes, especially for STEM subjects (National Governors Association, 2011). There would even be value to adding to a possible foundation for creating guidelines on how to select interesting information that is appropriate for a learning purpose, particularly with multimedia. If learning improvements cannot be consistently found, then perhaps this can deter misguided efforts in encouraging instruction that is interesting but ineffective.

## Acknowledgments

Much gratitude is given to Frank Durso and Paul Verhaeghen for their guidance, patience, and support.

## References

Ainley, M., Hidi, S., \& Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. Journal of Educational Psychology, 94(3) 545-561.
Carney, R. N., \& Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. Educational Psychology Review, 14(1), 5-26.
Garner, R., Alexander, P.A., Gillingham, M.G., Kulikowich, J.M., \& Brown, R. (1991). Interest and learning from text. American Educational Research Journal, 28(3), 643-659.
Harp, S. F., \& Mayer, R.E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. Journal of Educational Psychology, 89 (1), 92102.

Hidi, S., \& Renninger, K.A. (2006). The four-phase model of interest development. Educational Psychologist, 41(2), 111-127.
Magner, U. I., Schwonke, R., Aleven, V., Popescu, O., \& Renkl, A. (2014). Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments. Learning and Instruction, 29, 141-152.
Mayer, R. E., Griffith, E., Jurkowitz, I.T.N., \& Rothman, D. (2008). Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. Journal of Educational Psychology: Applied, 13(3), 329-339.
National Governors Association. (2011). Building a Science, Technology, Engineering, and Math Education Agenda: An Update of State Actions. Washington, DC: NGA Center for Best Practices.
Norman, D. (2013, October). MOOCs and Online Education. Seminar presented at the Georgia Institute of Technology GVU Brown Bag, Atlanta, GA.
Panksepp, J. (1998). Affective neuroscience: The foundations of human and animal emotions. New York: Oxford University Press.
Rey, G.D. (2012). A review of research and a meta-analysis of the seductive detail effect. Educational Research Review, 7(3), 216-237.
Schiefele, U. (1991). Interest, learning, and motivation. Educational Psychologist, 26(3-4), 299-323.
Schraw, G., Bruning, R., \& Svoboda, C. (1995). Sources of situation interest. Journal of Reading Behavior, 27(1), 117.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. Journal of Verbal Learning and Verbal Behavior, 6, 175-184.
Tversky, B. (1973). Encoding processes in recognition and recall. Cognitive Psychology, 5(3), 275-287.

# "I won't lie, it wasn't amazing": Modeling polite indirect speech 

Erica J. Yoon, Michael Henry Tessler, Noah D. Goodman and Michael C. Frank<br>\{ejyoon, mtessler, ngoodman, mcfrank\} @stanford.edu<br>Department of Psychology, Stanford University


#### Abstract

Why are we polite when we talk to one another? One hypothesis is that people expect others to choose what to say based on their goals both to transfer information efficiently (an epistemic goal) and to make the listener feel good (a social goal). In our previous work, we found that when these two goals conflict, they sometimes produce white lies. In the current work, we expand on this theory to consider another prominent case of polite speech: indirect remarks using negation (e.g., "It wasn't amazing"). With minimal extensions from our previous framework, our formal model suggests that a pragmatic speaker will produce more indirect remarks when the speaker wants to be informative and seem considerate at the same time. These predictions were borne out in a language production experiment. These findings suggest that the conflict between social and epistemic goals can account for a broad range of politeness phenomena.


Keywords: Politeness; computational modeling; communicative goals; pragmatics

## Introduction

Language users hear and produce polite speech on a daily basis. Adults and even young children spontaneously produce requests in polite forms (Axia \& Baroni, 1985; Clark \& Schunk, 1980), and speakers use politeness strategies even while arguing, preventing unnecessary offense to their interactants (Holtgraves, 1997). But being polite conflicts with one important goal of cooperative communication: exchanging information efficiently and accurately (Grice, 1975). People tell white lies ("Your new dress is gorgeous!") and produce indirect speech that is longer and more nuanced than the simplest form of their intended message ("I don't think that dress looks phenomenal on you" as opposed to "It looks terrible") to make others feel good. Speakers risk potential loss of their intended message (indirect speech), intentionally convey wrong information (lies), and suffer inefficiencies - all in the service of being polite. If information transfer were the only currency in communication, politeness would be both infelicitous and undesirable.

A cooperative speaker, however, can be imagined as one with both an epistemic goal to improve the listener's knowledge as well as a social goal to minimize potential damage to the hearer's (and the speaker's own) self-image, called face (Brown \& Levinson, 1987). If the speaker's intended meaning contains no threat to the speaker or listener's face, then the speaker will choose to convey the meaning in an explicit and efficient manner (putting it "on the record"). As the degree of face-threat becomes more severe, however, a speaker will choose to be polite by producing more indirect utterances.

Inspired by this set of ideas, we have argued that listeners think about polite speech as reflecting a tradeoff between two goals: information transfer (which we called epistemic utility) and face-saving (social utility; Yoon, Tessler, Goodman,
\& Frank, 2016). A speaker with a high weight on social utility will try to save her listener's face: She hides or risks losing information in her intended message by making her utterance false to some degree. On the other hand, a speaker with a high weight on epistemic utility prioritizes truthfulness and informativity, and she may risk a loss of the listener's (or the speaker's own) face. These ideas were formalized in a model of pragmatic language understanding, building on the Rational Speech Act (RSA) theory (for a review, see Goodman \& Frank, 2016). We tested the polite RSA model (pRSA) by examining white lies. The model captured human participants' inferences about a speaker's goals given her utterance (e.g., saying a good talk was "amazing" implies that she is being nice) and about the true state of the world given a speaker's goal (e.g., saying "good" may mean the talk was only okay if the speaker wanted to be nice).

In the current work, we extend our framework to another polite speech act: indirect speech. White lies are produced when a speaker tries to save the listener's face by stretching the truth. But instead of lying, people sometimes try to be polite by being more indirect. Through indirect speech, a speaker can express meaning that is different from the literal meaning of the utterance (Searle, 1975). In this work, we focus on negation ("not"), which has the potential to be indirect. For instance, "Mark isn't the cleanest person I know" may suggest that the speaker thinks Mark is unclean (inferred meaning) rather than not being the person who has the greatest degree of cleanliness (literal meaning). Negation can be used as a hedging or mitigating device to address an undesirable state that is face-threatening to the addressee (Brown \& Levinson, 1987; Grice, 1975).

What may lead a speaker to produce indirect remarks? An indirect remark may be motivated by the speaker's goal to convey some face-threatening information, while being seen as a polite person who avoids threatening others' face. In our previous work, we described a pragmatic listener that jointly inferred the true state and the goals of the speaker. Building on this model, we describe here a speaker whose goal is to lead this pragmatic listener to infer the true state and attribute to the speaker certain goals (e.g., face-saving). For instance, "It wasn't amazing" does not preclude the possibility that the presentation was bad, and may in fact be pragmatically strengthened to mean that it was actually bad. Yet because the speaker does not choose the more direct "It was bad", the listener will infer a face-saving goal. Thus saying "It wasn't amazing" can accomplish the goal of conveying that the presentation was bad while the speaker is seen as not wanting to make the listener feel bad. On the other hand, if the speaker does not care about being seen as face-saving,
she will produce less indirect speech. Further, if the presentation was actually good, or even decent, the speaker will prefer to produce a directly positive remark ("It was good") in either case. Thus we predict more indirect speech when the true state is bad, and an interaction with the speaker's desire to both be informative and be seen as wanting to save face. In what follows, we derive our hypotheses using our formal model and present an empirical test of the hypotheses.

## Computational Model

In the current work, we introduce a minimal extension to our previous RSA model (pRSA; Yoon et al., 2016) to allow for speaker production of indirect remarks using negation.

## Polite RSA

RSA models assume speakers choose utterances approximately optimally given a utility function (Goodman \& Stuhlmuller, 2013). pRSA posited that the speaker's utility function can be decomposed into two components. First, epistemic utility $\left(U_{\text {epi }}\right)$ refers to the standard, informative utility in RSA: the amount of information a literal listener ( $L_{0}$ ) would still not know about world state $s$ after hearing a speaker's utterance $w$. Second, social utility $\left(U_{s o c}\right)$ is the expected subjective utility of the state inferred given the utterance $w$. The expected subjective utility is related to the intrinsic value of the state, and we use a value function $(V)$ to map states to subjective utility values. This captures the affective consequences for the listener of being in state $s$. Finally, some utterances might be costlier than others. The utility of an utterance subtracts the cost $c(w)$ from the weighted combination of the social and epistemic utilities.
$U(w ; s ; \hat{\beta})=\beta_{e p i} \cdot \ln \left(P_{L_{0}}(s \mid w)\right)+\beta_{s o c} \cdot \mathbb{E}_{P_{L_{0}}(s \mid w)}[V(s)]-C(w)$
The speaker $\left(S_{1}\right)$ in pRSA chooses utterances $w$ softmaxoptimally given the state $s$ and his goal weights $\hat{\beta}$. The pragmatic listener $\left(L_{1}\right)$ jointly infers the state $s$ and the utility weights of the speaker, $\beta_{e p i}$ and $\beta_{s o c}$ (Goodman \& Lassiter, 2015; Kao, Wu, Bergen, \& Goodman, 2014).

$$
\begin{align*}
P_{L_{1}}(s, \hat{\beta} \mid w) & \propto P_{S_{1}}(w \mid s, \hat{\beta}) \cdot P(s) \cdot P(\hat{\boldsymbol{\beta}})  \tag{1}\\
P_{S_{1}}(w \mid s, \hat{\beta}) & \propto \exp \left(\lambda_{1} \cdot \mathbb{E}[U(w ; s ; \hat{\beta})]\right)  \tag{2}\\
P_{L_{0}}(s \mid w) & \propto[[w](s) \cdot P(s) \tag{3}
\end{align*}
$$

Within our experimental domain, we assumed there were five possible states of the world corresponding to the value placed on a particular referent (e.g., rating deserved by the presentation the speaker is commenting on, akin to a Yelp rating): $S=\left\{s_{1}, \ldots, s_{5}\right\}$. We assume a uniform prior distribution over possible states of the world. The states have subjective numerical values $V\left(s_{i}\right)=\alpha \cdot i$, where $\alpha$ is a free parameter. $[[w]](s)$ corresponds to the lexical meaning of the utterance $w$ (e.g., "good") when applied to state $s$. We gather independent ratings for these literal meanings.

## Extensions to pRSA

We build on pRSA by adding negative utterances and modeling a more sophisticated speaker. First, we extend the utterance alternatives to include negation. Previously we considered five possible utterances: \{It was terrible, bad, okay, good, and amazing\}, all direct assertions of specific states (e.g., "It was amazing" would be true for the state of 5 but untrue for the states of 1 or 2 ). Now the speaker may say, \{It wasn't terrible, bad, okay, good, and amazing\}. These utterances indirectly address the referent by negating certain state. We assume that it is more costly to say utterances with negation, which makes the utterance morphemically longer and is harder to process (Clark \& Chase, 1972). In our full data analysis, we put a prior on this negation cost parameters and infer its likely values from the data.

Most importantly, we extended the recursive reasoning in the model. For our experiment, we consider the pragmatic speaker ( $S_{2}$ ) who chooses an utterance based on the pragmatic listener model (Eq. 1), thinking about the state as well as goal weights that the pragmatic listener will infer.

$$
P_{S_{2}}(w \mid s, \hat{\beta}) \propto \exp \left(\lambda_{2} \cdot \ln \left(P_{L_{1}}(s, \hat{\beta} \mid w)\right)-C(w)\right)
$$

This crucially captures the idea that the speaker both wants to convey the state $s$, and to be seen as someone with goals $\hat{\beta}$. We simplify from the Yoon et al. (2016) model by including only a single mixture parameter $\phi$ governing the extent to which the speaker is being informative vs. face saving: $\beta_{e p i}=\phi, \beta_{s o c}=1-\phi$.

We implemented this model using the probabilistic programming language WebPPL (Goodman \& Stuhlmuller, $2014 \sqrt{1}$ In the next section, we explore the model's predictions for speaker productions of indirect speech with negation vs. direct speech with no negation.

## Model predictions

Before describing our experimental data, we derive predictions from the pRSA model. In these initial simulations, we use fixed goal weights and parameters - in later fits, we will derive these parameters from the data using Bayesian data analysis. Since the model requires measurements of literal semantics (e.g., what "not good" means on a given dimension), we first describe these measurements and then give model predictions using them.

## Semantic measurement

We probed judgments of literal meanings of the target words assumed by our model and used in all our experiments.
Materials, methods, and results 25 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. We used 13 different context items that were previously used in Yoon et al. (2016), in which someone

[^548]

Figure 1: Semantic measurement results. Proportion of acceptances of utterance types (colors) combined with target words (facets) given the true state represented on a scale of hearts. Error bars represent 95\% confidence intervals.


Figure 2: Schematic model predictions (left), experimental results (center) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).
evaluated a performance of some kind. For example, in one of the contexts, Ann saw a presentation, and Ann's feelings toward the presentation (true state) were shown on a scale out of five hearts (e.g., two out of five hearts filled in red color). The question of interest was "Do you think Ann thought the presentation was / wasn't X?" and participants responded by choosing either "no" or "yes." The target could be one of five possible words: terrible, bad, okay, good, and amazing, giving rise to ten different possible utterances (with negation or no negation). Each participant read 50 scenarios, depicting every possible combination of states and utterances. The order of context items was randomized, and there were a maximum of four repeats of each context item per participant. For this and subsequent experiments, we analyzed the data by collapsing across context items.

For each utterance-state pair, we computed the posterior distribution over the semantic weight (i.e., how consistent X utterance is with Y state) assuming a uniform prior over the weight. Meanings of the words as judged by participants were as one would expect (see Figure 1). We used the fraction of participants that endorsed utterance $w$ for state $s$ to set informative priors to infer posterior credible values of the literal meanings from data in the speaker production experiment.

## Model parameters and predictions

The $S_{2}$ speaker in our model has the goal to convey the state and to be seen as having a particular set of goals. We explore predictions for 3 hypothetical speakers, corresponding to 3 different $\phi$ mixture parameter weights: (a) an informative speaker who wants to convey high epistemic utility (prioritizing information transfer; $\phi=0.9$ ) (b) a social speaker who wants to convey high social utility (making the listener feel good; $\phi=0.1$ ) (c) a both-goal speaker who wants to convey a balance between the two utilities $(\phi=0.5){ }^{2}$

Figure 2 (left) shows the speaker's production probabilities associated with producing an indirect speech act (i.e., an utterance with negation) for the three different speakers as the true state of the world is varied. We see, consistent with our intuition, that indirect speech was relatively more preferred in bad states than in good states. As well, we see higher probability of negation production for the speaker who wants to convey both goals (epistemic and social) relative to each goal

[^549]independently. Indirect speech does not convey that much information and so the informative speaker (a) would disprefer it. The social speaker (b) who wants to convey a face-saving goal would tend to signal a better-that-actual state through direct positive remarks. The both-goal speaker produces indirect remarks to avoid direct remarks that are either true but face-threatening, or face-saving but false.

## Speaker production experiment

To compare against our model predictions, we measured participants' predictions for the most likely utterance ( $w$ ) produced by the speaker, given a description of the true state. For example, given that Ann wanted to make Bob feel good but felt that his poem deserved 2 out of 5 hearts, what would she say? We hypothesized that when there was no tradeoff between informativity and face-threat avoidance (i.e., when the addressee's performance was great), speakers would use truthful and face-saving direct remarks ("[Your poem] was amazing") regardless of their described goals. However, when there was a conflict between the epistemic and social goals (i.e., when the addressee's performance was poor), a speaker who tried to convey both goals would use vague indirect remarks ("[Your poem] wasn't terrible") more often than direct face-threatening remarks ("[Your poem] was bad"; preferred by a speaker who only considered the epistemic goal) or direct face-saving remarks ("[Your poem] was good"; preferred by a speaker who wanted to convey only a social goal).

## Method

Participants 202 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.
Stimuli and Procedure As in the semantics measurements above, we used scenarios in which a person (e.g., Bob) gave some performance and asked for another person (e.g., Ann)'s opinion on the performance. Additionally, we provided information on the speaker Ann's goal - to make Bob feel good, or to give as accurate and informative feedback as possible, or both - and the true state - how Ann actually felt about Bob's performance (e.g., 2 out of 5 hearts). Each participant read 15 scenarios, depicting every possible combination of goals and states. The order of context items was randomized, and there were a maximum of two repeats of each context item per participant.

Each scenario was followed by a question that read, "If Ann wanted to make Bob feel good but not necessarily give informative feedback (or to give accurate and informative feedback but not necessarily make Bob feel good, or BOTH make Bob feel good AND give accurate and informative feedback), what would Ann be most likely to say?" Participants indicated their answer by choosing one of the options on the two dropdown menus, side-by-side, one for choosing between was vs. wasn't and the other for choosing among terrible, bad, okay, good, and amazing (see Figure 3).
Imagine that Justine wrote a review for a book, but Justine didn't know how good it was. Justine approached Kelly, who knows a lot about writing reviews, and asked "How was my review?"
Here's how Kelly actually felt about Justine's review:

If Kelly wanted to make Justine feel good, but not necessarily give informative feedback,
What would Kelly be most likely to say?
What would Kelly be most likely to say?


Next
Next

Figure 3: Example of a trial in Experiment 1.

## Behavioral results

Our hypotheses for utterance production by speakers with different goals were borne out (see full results in Figure 4).
For good states (4 and 5 hearts), positive direct remarks were judged to be the most likely utterances across all three goal conditions. For less-than-perfect, but still decent states, there was a greater degree of expectation of white lies (e.g., "It was amazing" for 4 hearts) given a social goal. For bad states ( 1 and 2 hearts), as predicted, there were more instances of expected indirect remarks overall across all goal conditions given bad states. Critically, speakers with both informative and social goals produced more indirect remarks than were observed in the other two goal conditions (Figure 2, center).

## Model results

Model fitting In this experiment, participants were told what speakers' intentions were (e.g., wanted to make Bob feel good). We assume that the intention descriptions conveyed the weight mixture $\phi$ that the speaker was using. We put uninformative priors on this mixture $(\phi \sim \operatorname{Uniform}(0,1))$ and inferred their credible values separately for each goal condition ("wanted to X") using Bayesian data analytic techniques (Lee \& Wagenmakers, 2014). We also used the fraction of participants that endorsed utterance $w$ for state $s$ to set informative priors to infer posterior credible values of the literal meanings from data.

There were four additional parameters of the model, on which we put uninformative priors: the speaker optimality parameter $\left(\lambda_{S_{1}} \sim \operatorname{Unif}(0,20)\right)$; the pragmatic speaker optimality parameter $\left(\lambda_{S_{2}} \sim \operatorname{Unif}(0,5)\right)$; the value scale parameter ( $\alpha \sim \operatorname{Unif}(0,5)$ ) in the utility function; and the cost parameter $(C(u) \sim \operatorname{Unif}(1,10))$. We inferred their posterior credible values from the data. We ran 4 MCMC chains for 80,000 iterations, discarding the first 40,000 for burnin. The Maximum A-Posteriori (MAP) estimate and $95 \%$ Highest Probability Density Interval (HDI) were: $\lambda_{S_{1}}: 2.16$ [2.02, 3.61]; $\lambda_{S_{2}}: 0.91$ [0.83, 1.75]; $\alpha: 2.71[0.98,4.59] ; C(w): 2.04$ [1.95, 2.25]. To generate utterance predictions, given our model and the inferred parameters, we evaluated the posterior predictive distribution, marginalizing out all parameters.


Figure 4: Experimental results (solid lines) and fitted model predictions (dashed lines) for speaker production. Proportion of utterances chosen (utterance type - direct vs. indirect - in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows). Error bars represent $95 \%$ confidence intervals for the data and $95 \%$ highest density intervals for the model.

Results The inferred weights for each goal condition were largely as expected: For the "wanted to give informative feedback" (informative) condition, the model put a relatively high weight on epistemic utility ( 0.81 ). For the "wanted to make [listener] feel good" (social) condition, the model inferred that the speaker was using a moderate weight on epistemic utility ( 0.51 ). For the "wanted BOTH to make [the listener] feel good and give informative feedback" (both) condition, the model assigned a weight on epistemic utility between the weights for the other two goal conditions (0.57). Overall, the weights tended to be more biased towards prioritizing the epistemic utility.

The predictions for the speaker's utterance were overall highly consistent with the experimental findings (Figure 4). The posterior predictive of the model explained almost all of the variance in the production data $r^{2}(150)=0.962$ (Figure 5). The model successfully predicted distinct patterns for each goal condition. The informative speaker produced direct remarks whose literal meanings mapped onto the true states (e.g., "It was terrible" given 1 heart). The social speaker produced remarks that were positively biased compared to the true states (e.g., "It was okay" given 2 hearts).

While the model in the both condition did produce indirect utterances (e.g., "It wasn't terrible" given 1 heart) it did so slightly less than the empirical data. For this reason, the
model did not yield the expected difference for negation production between both-goal and social conditions (Figure 2, right); though the trend was numerically correct, the effect was much smaller in the fit model than the schematic one. There are several possible explanations for this small deviation. In our experimental data, the social speaker placed a higher weight on epistemic utility than in our schematic predictions. Thus, the particular goal descriptions we used in the experiment may have suggested that the social speaker still wanted to be seen as informative, and have led to little differentiation between the social vs both-goal speaker. Another possible cause is that participants preferred a different kind of indirect speech than the model - in particular, the bothgoal speaker preferred to produce "It wasn't amazing" in the schematic model predictions, whereas participants in our experiment chose "It wasn't terrible." This discrepancy between the two remarks is interesting, because their implied meaning is similar. In a pilot experiment where participants were asked to infer the true state (number of hearts) from an utterance, "It wasn't amazing" and "It wasn't terrible" were very similar ( 2 hearts).

## Discussion

Why are we polite? Here we explored a formal instantiation of the hypothesis that two conflicting speaker goals - epis-


Figure 5: Full distribution of human responses vs. model predictions. Error bars represent $95 \%$ confidence intervals for the data (vertical) and $95 \%$ highest density intervals for the model (horizontal).
temic and social - can be used to explain a range of polite behavior, including white lies and indirect speech acts using negation. Our model predicted that speakers should produce more indirect remarks in cases of greater face threat (given the addressee's poorer performance) and in cases where speakers wanted to be both informative and nice. Our experimental data confirmed these predictions. The model's overall fit to the data was very strong, although it did not show the predicted dominance of indirect speech for the both-goal speaker at low states. Whether this discrepancy between the initial and data-fitted predictions was due to variation in goal weight based on experimental scenarios or a discrepancy in preferences for particular utterances is a question for future work.

An important contribution of this work is in showing the generalizability of our formal model (pRSA) to the case of indirect speech acts. The current work took a step in addressing speakers' self-presentation: Not only do speakers want to save the listener's face, but they also want to save their own face, by appearing informative and considerate to the listener. In future work we hope to explore this aspect more and test how our model's utilities can be extended to capture the speaker's desire to appear polite, genuine, and even modest. Using the model to explore other kinds of polite speech such as indirect requests ("Would you mind closing the window?"; Clark \& Schunk, 1980) and manifestations of polite speech in different cultures (e.g., Holtgraves \& Joong-nam, 1990) are also important future directions.

In sum, our formal model and experimental work represent an advance in polite speech understanding. With a minimal extension to our existing model, we were able to capture subtle patterns in people's inferences about indirect speech production. Our empirical findings suggest that neither epistemic nor social motives alone motivate indirect speech; instead,
the need for indirect speech results from the conflict between these two. These findings provide strong support for a utilitytheoretic framing of politeness, and suggest new directions in understanding of pragmatic language use in social contexts.

## Acknowledgments

This work was supported by NSF grant BCS 1456077 to MCF, ONR grant N00014-13-1-0788 to NDG, NSF Graduate Research Fellowship DGE-114747 to MHT, and NSERC PGS Doctoral scholarship PGSD3-454094-2014 to EJY.

## References

Axia, G., \& Baroni, M. R. (1985). Linguistic politeness at different age levels. Child Development, 918-927.
Brown, P., \& Levinson, S. C. (1987). Politeness: Some universals in language usage (Vol. 4). Cambridge Univ. Press.
Clark, H. H., \& Chase, W. G. (1972). On the process of comparing sentences against pictures. Cognitive Psychology, 3(3), 472-517.
Clark, H. H., \& Schunk, D. H. (1980). Polite responses to polite requests. Cognition, 8(2), 111-143.
Goodman, N. D., \& Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences, 20(11), 818-829.
Goodman, N. D., \& Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin \& C. Fox (Eds.), The handbook of contemporary semantic theory, 2nd edition. Wiley-Blackwell.
Goodman, N. D., \& Stuhlmuller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. Topics in Cognitive Science, 5.
Goodman, N. D., \& Stuhlmuller, A. (2014). The Design and Implementation of Probabilistic Programming Languages. http://dippl.org.
Grice, H. P. (1975). Logic and conversation. In Readings in language and mind. Blackwell.
Holtgraves, T. (1997). YES, but. positive politeness in conversation arguments. Journal of Language and Social Psychology, 16(2), 222-239.
Holtgraves, T., \& Joong-nam, Y. (1990). Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. Journal of Personality and Social Psychology, 59(4), 719.
Kao, J. T., Wu, J. Y., Bergen, L., \& Goodman, N. D. (2014). Nonliteral understanding of number words. Proceedings of the National Academy of Sciences, 111(33), 12002-12007.
Lee, M. D., \& Wagenmakers, E. J. (2014). Bayesian cognitive modeling: A practical course. Cambridge Univ. Press.
Searle, J. R. (1975). Indirect speech acts. In P. Cole \& J. Morgan (Eds.), Syntax and semantics (vol. 3): Speech acts. New York: Academic Press.
Yoon, E. J., Tessler, M. H., Goodman, N. D., \& Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In Proceedings of the thirty-eighth annual conference of the Cognitive Science Society.

# The Structure of Young Children's Numerical and Spatial Abilities 

Christopher Young<br>University of Chicago, Chicago, Illinois, United States<br>Stephen Raudenbush<br>University of Chicago<br>Brittney Fraumeni<br>University of Chicago<br>Susan Levine<br>University of Chicago


#### Abstract

We conducted a study of 400 preschool children to determine whether spatial and numerical skills rely on common processes. Children completed a battery of mathematical tasks as part of an ongoing preschool formative assessment development project. We created theoretically meaningful skills from these tasks and carried out item response theoretic analyses on each skill. We extracted Rasch scores for each of the skills and carried out multiple factor analyses to determine whether one or more factors best characterized spatial and numerical skills. Finally, we regressed factor scores on demographic variables, including age, gender, socioeconomic status, and verbal ability. We discuss how our results add to our understanding of the connection between spatial and numerical processes and their implications for closing the achievement gap in early education.


# Alternation blindness in the perception of binary sequences 

Ru Qi Yu (ruqiyu@psych.ubc.ca)<br>Department of Psychology, University of British Columbia<br>Daniel Osherson (osherson@princeton.edu)<br>Department of Psychology, Princeton University<br>Jiaying Zhao (jiayingz@psych.ubc.ca)<br>Department of Psychology; Institute for Resources, Environment and Sustainability, University of British Columbia


#### Abstract

Binary information is prevalent in the environment. In this study, we examined how people process repetition and alternation in binary sequences. Across four paradigms involving estimation, working memory, change detection, and visual search, we found that the number of alternations is under-estimated compared to repetitions (Experiment 1). Moreover, recall for binary sequences deteriorates as the sequence alternates more (Experiment 2). Changes in bits are also harder to detect as the sequence alternates more (Experiment 3). Finally, visual targets superimposed on bits of a binary sequence take longer to process as alternation increases (Experiment 4). Overall, our results indicate that compared to repetition, alternation in a binary sequence is less salient in the sense of requiring more attention for successful encoding. The current study thus reveals the cognitive constraints in the representation of alternation and provides a new explanation for the over-alternation bias in randomness perception.


Keywords: alternation bias, randomness perception, working memory, attention, numerosity perception

## Introduction

Perceptually, many events in the world can be interpreted as binary, from the outcomes of coin flips to the daily alternations between the sun and the moon. Past research that examines the perception of binary information has focused on the perception of randomness (Bar-Hillel \& Wagenaar, 1991; Nickerson, 2002) and regularities (Julesz, 1962; Lopes \& Oden, 1987). Although it is difficult to define randomness (Ayton et al., 1989; Beltrami, 1999; Chater \& Vitányi, 2003; Fitelson \& Osherson, 2012; Oskarsson et al., 2009), there are systematic biases in people's conception of randomness, such as the gambler's fallacy (Kahneman \& Tversky, 1972) and the hot hand fallacy (Gilovich et al., 1985). One particular bias that has received much attention in the past is the over-alternation bias: a binary sequence that alternates more than expected on the basis of random generation tends to be judged as random (Bar-Hillel \& Wagenaar, 1991; Falk \& Konold, 1997; Lopes \& Oden, 1987; Nickerson, 2002), and people tend to produce random sequences that contain too many alternations (Kahneman \& Tversky, 1972; Wagenaar, 1972). This bias is robust across different stimulus types, sensory modalities, or presentation modes (Yu et al., in press).

A number of accounts have been proposed to explain the over-alternation bias. One explanation focuses on the limits of working memory (Baddeley, 1966; Kareev, 1992). Since people can only hold a limited number of items in working memory at any given time, the amount of bits being processed is constrained, leading to a biased sample of randomness (Hahn \& Warren, 2009; Miller \& Sanjuro, 2015; Yu et al., in press). Another prominent account of the over-alternation bias is the idea of local representativeness, which suggests that people assume equal frequency of outcomes within a local sequence (Tversky \& Kahneman, 1971). A recent account is offered by Falk and Konold (1997) who proposed an encoding hypothesis that states that the probability that a bit sequence is labelled random varies directly with the time needed to correctly memorize or copy it. However, this account has been challenged by recent work showing a discrepancy between encoding difficulty of the binary sequence and labeling the sequence as random (Zhao et al., 2014). While these explanations have offered valuable insights, there remains a possibility that people have an accurate view of randomness, but the cognitive limitations contribute to a biased conception of randomness (Rapoport \& Budescu, 1992).

## The current study

We explore a new explanation focusing on a perceptual limitation in the ability to represent alternations vs. repetitions. If alternations are under-represented compared to repetitions, there needs to be more alternations in the sequence in order for people to perceive a $50 \%$ alternation rate that is typically assumed in a random sequence.
In order to generate a binary sequence that contains different levels of alternations while maintaining the equal probability of outcomes, we used an algorithm that deviates from stochastic independence by allowing previous bit to influence the next one. Specifically, for each $p$ in the unit interval (from 0 to 1 ), let $D(p)$ generate a sequence of bits consisting of zeros and ones as follows:

A fair coin toss determines the $1^{\text {st }}$ bit. Suppose that the $n^{\text {th }}$ bit (for $n \geqslant 1$ ) has been constructed. Then with probability $p$ the $n+1^{\text {st }}$ bit is set equal to the opposite of the $n^{\text {th }}$ bit; with probability $1-p$ the $n+1^{\text {st }}$ bit is set equal to the $n^{\text {th }}$ bit. Repeat this process to generate a sequence of any length.

This procedure was first introduced by Zhao, Hahn, and Osherson (2014). $D(.5)$ is a genuinely random device. For $p<.5, \mathrm{D}(p)$ tends to repeat itself, resulting in long streaks, whereas for $p>.5, D(p)$ tends to alternate. The expected proportion of each bit is $50 \%$ for all $p \in[0,1]$, although empirically, the output might deviate from $50 \%$; however such deviations should be small and random (Yu et al., in press). For any sequence produced by $D(p)$, the expected proportion of alternation, called the "switch rate" of the generating process, is $p$. The expected proportion of repetitions, called the generating "repeat rate", is $1-p$.

We conducted four experiments to examine how people represent alternations vs. repetitions. In Experiment 1, participants viewed a binary sequence and estimated the number of switches or repeats in the sequence. In Experiment 2, participants viewed a binary sequence and recalled the sequence. In Experiment 3, participants viewed two sequences and judged whether the sequences were the same or different. In Experiment 4, participants searched for a target embedded in a binary sequence.

## Experiment 1

This experiment examined if there are differences in the estimation of alternation vs. repetition in binary sequences.

## Participants

Forty-five undergraduate students ( 32 female, mean age $=19.9$ years, $\mathrm{SD}=2.3$ ) from the University of British Columbia (UBC) participated for course credit. Participants in all experiments provided informed consent. All experiments reported here have been approved by the UBC Behavioral Research Ethics Board. We conducted a power analysis using G*Power (Faul, Erdfelder, Lang, \& Buchner, 2007), which showed that given an effect size of 0.53 (based on our prior work, Zhao \& Yu, 2016), a minimum of 38 participants would be required to have $95 \%$ power to detect the effect in our design.

## Stimuli

In each trial, participants viewed a 30-bit sequence. Each sequence contained circles of two colors: green (RGB: 0 2550 ) and blue (RGB: 00255 ). Each circle subtended $0.9^{\circ}$ in diameter (Figure 1a). There were five levels of switch rates in $D(p)$ in generating the sequences, where $p=0.1$, $0.3,0.5,0.7$, and 0.9 . Correspondingly, there were five levels of repeat rates $(1-p)=0.9,0.7,0.5,0.3$, and 0.1 .

Temporal sequences. For half of the trials, participants viewed a temporal sequence where the 30 circles were presented one after another, making simple visual grouping impossible. Each circle was presented at the center of the screen for 100 ms , and the inter-stimulus interval (ISI) was 100 ms with a blank screen (Figure 1a).

Spatial sequences. For the other half of the trials, participants viewed a spatial sequence, where the 30 circles were presented on the screen simultaneously. The circles were presented left to right. The space between two adjacent
circles in the sequence subtended $0.1^{\circ}$. Each sequence was presented on the screen for 1000 ms (Figure 1a).

## Procedure

There were 200 trials in total for each participant. In each trial, participants viewed a sequence with one of the five generating switch rates $(0.1,0.3,0.5,0.7$, or 0.9$)$. Each level of switch rate contained 40 trials, among which 20 trials were temporal sequences and 20 trials were spatial sequences. After viewing the 30 -bit sequence, participants were asked to estimate either the number of the color switches ( 10 trials), or the number of color repeats (10 trials). Specifically, the instruction for estimating color switches was "How many times did a dot have a DIFFERENT color from the previous dot in the sequence?" and the instruction for estimating color repeats was "How many times did a dot have the SAME color as the previous dot in the sequence?". Participants were also told that the range of their estimate was from 0 to 29 (29 was the maximum possible number of switches or repeats in the sequence). Participants typed in their estimate after seeing each sequence. In sum, there were three within-subjects factors: the generating switch rate of the sequence (from 0.1 to 0.9 ), the presentation of the sequence (temporal vs. spatial), and the estimation type (switches vs. repeats). The order of the trials was randomized for each participant. There was no mention of randomness in all experiments.

## Results and Discussion

Estimated switch rate was the derived by dividing the estimated number of switches from the participants by 29 (the maximum possible switches in the sequence). Likewise, estimated repeat rate was calculated by dividing the estimated number of repeats from the participants by 29 (the maximum possible repeats in the sequence). Observed switch rate was the objective switch rate in the sequence presented to the participants in each trial. Likewise, observed repeat rate was the objective repeat rate in the sequence presented in each trial. The generating switch rate was the $p$ in $D(p)$ in the algorithm that generated the sequence. The generating repeat rate was $1-p$. To verify that the presented sequence actually exhibited the generating switch rate or repeat rate, we plotted the observed switch rate or repeat rate for each sequence (Figure 1 b to e), which mapped closely to the generating switch rate or repeat rate.

We computed the signed error (estimated - observed switch rate or repeat rate) at each of the five generating levels. For temporal trials (Figure 1 b and d), a 5 (generating rate: $0.1,0.3,0.5,0.7$, and 0.9 ) $\times 2$ (estimation type: switches vs. repeats) repeated-measures ANOVA revealed a main effect of generating rate $[F(4,176)=162.3, p<.001$, $\left.\eta_{p}{ }^{2}=0.79\right]$ and of estimation type $[F(1,44)=49.34, p<.001$, $\left.\eta_{p}{ }^{2}=0.53\right]$, and a reliable interaction $[F(4,176)=10.75$, $\left.p<.001, \eta_{p}{ }^{2}=0.20\right]$. Pair-wise comparisons at each generating rate showed that participants underestimated the number of switches more than repeats at each of the five generating rates $[p \prime s<.01]$. For spatial trials (Figure 1 c and e), the same

ANOVA revealed a main effect of generating rate $\left[F(4,176)=107.2, p<.001, \eta_{p}^{2}=0.71\right]$ and of estimation type $\left[F(1,44)=114.2, \quad p<.001, \quad \eta_{p}^{2}=0.72\right]$, but no interaction [ $\left.F(4,176)=0.07, \quad p=.99, \quad \eta_{p}{ }^{2}<0.01\right]$. Again, pair-wise comparisons at each generating rate showed that participants underestimated the number of switches more than repeats at each of the five generating rates [ $p$ 's<.001].


Figure 1. Experiment 1. (a) Participants ( $\mathrm{N}=45$ ) were presented with temporal or spatial sequences, and estimated either the number of circles that had a different color from the previous circle (switch) or the number of circles that had the same color as the previous one (repeat). (b) The estimated switch rate and the observed switch rate were plotted for temporal trials. (c) The estimated switch rate and the observed switch rate were plotted for spatial trials. (d) The estimated repeat rate and the observed repeat rate were plotted for temporal trials. (e) The estimated repeat rate and the observed repeat rate were plotted for spatial trials. (Error bars reflect $\pm 1$ SEM; $\left.{ }^{*} p<.05, * * p<.01, * * * p<.001\right)$

We further compared the estimated switch or repeat rate with the observed switch or repeat rate. For temporal trials (Figure 1b), participants over-estimated the switch rate at 0.1 and 0.3 , but under-estimated the switch rate at $0.5,0.7$, and 0.9 . They also over-estimated the repeat rate at 0.1 and 0.3 , but under-estimated the repeat rate at 0.7 and 0.9 (Figure 1d). For spatial trials (Figure 1c), participants overestimated the switch rate only at 0.1 , and under-estimated the switch rate at $0.3,0.5,0.7$, and 0.9 . They over-estimated the repeat rate at $0.1,0.3$, and 0.5 , but under-estimated the repeat rate at 0.7 and 0.9 (Figure 1e).

Interestingly, when estimating the number of repeats, participants were the most accurate around 0.5 where the sequences were truly random. For the same random sequence, participants were significantly under-estimating the number of switches. In fact, for people to perceive a 0.5 switch rate, the sequence must contain more than $50 \%$ switches, with a switch rate of around 0.7 (Figure 1 b and c ).

This perceptual insensitivity to switches may underlie the conceptual over-alternation bias of randomness. Taken together, these results suggest that alternations in a binary sequence were under-represented compared to repetitions.

## Experiment 2

One explanation for the under-estimation of switches could involve working memory. Specifically, people may have trouble representing switches accurately in memory, mistaking them for repeating bits, thus leading to underestimation. To examine this possibility, here participants were asked to recall each sequence.

## Participants

Forty-five students ( 30 female, mean age $=19.6$ years, $\mathrm{SD}=1.2$ ) from UBC participated for course credit.

## Stimuli and Procedure

The stimuli were the same as those in Experiment 1, except for these differences: there were 10 circles per sequence to circumvent a floor effect in the recall task; each circle was slightly larger, subtending $1.4^{\circ}$ in diameter, and the distance between each circle in spatial sequences was $0.2^{\circ}$; and each spatial sequence was presented for 500 ms (Figure 2a).

The procedure was identical to Experiment 1, except for one difference: after seeing each sequence, participants were asked to recall the sequence as accurately as they could, by pressing two keys to produce the green circle (the "G" key) or the blue circle (the "B" key). Participants were instructed to recall the dots in the same order as they appeared. After each key press, the corresponding circle was presented on the screen for 100 ms , and then disappeared. To recall the spatial sequence, participants pressed one key and the corresponding circle appeared from left to right on the screen, and remained on the screen.

## Results and Discussion

Since the observed switch rate of the sequences mapped closely onto the generating switch rates (Experiment 1), for all following experiments task performance was plotted against the five generating switch rates.

To assess the accuracy of participants' recalled sequences, we divided the exact matches between the presented sequence and the recalled sequence by 10 . The accuracy was plotted over the five levels of switch rates. For temporal trials (Figure 2b), a one way repeated-measures ANOVA revealed a significant difference in accuracy across the five switch rates $\left[F(4,176)=75.61, p<.001, \eta_{p}{ }^{2}=0.63\right]$. Post-hoc Tukey HSD analysis showed all pair-wise comparisons were significant except between 0.7 and 0.9 , and 0.5 and 0.9 . For spatial trials (Figure 2c), accuracy was different across the switch rates $\left[F(4,176)=111.5, p<.001, \eta_{p}{ }^{2}=0.72\right]$, and posthoc Tukey HSD analysis showed that all pair-wise comparisons were significant except between 0.7 and 0.9 . These results demonstrate that as the switch rate of the sequence increased, recall accuracy decreased.

To obtain a more fine-grained analysis, from the second bit on, we calculated the recall accuracy of each bit depending on whether the bit repeated or switched from the previous bit. We compared the recall accuracy of switching versus repeating bits. For temporal trials (Figure 2d), a 5 (generating rate: $0.1,0.3,0.5,0.7$, and 0.9 ) $\times 2$ (bit type: repeating vs. switching) repeated-measures ANOVA showed a main effect of generating rate $[F(4,176)=75.61$, $\left.p<.001, \eta_{p}{ }^{2}=0.63\right]$ and of bit type $[F(1,44)=206.7, p<.001$, $\left.\eta_{p}{ }^{2}=0.82\right]$, and a reliable interaction $[F(4,176)=37.4, p<.001$, $\left.\eta_{p}{ }^{2}=0.46\right]$. Pair-wise comparisons at each generating rate showed that the recall accuracy of repeating bits was consistently higher than that of switching bits [ $p$ 's $<.01$ ]. For spatial trials (Figure 2e), the same ANOVA showed a main effect of generating rate $[F(4,176)=111.5, p<.001$, $\left.\eta_{p}{ }^{2}=0.46\right]$ and of bit type $\left[F(1,44)=28.84, p<.001, \eta_{p}{ }^{2}=0.40\right]$, and a reliable interaction $\left[F(4,176)=7.18, p<.001, \eta_{p}{ }^{2}=0.14\right]$. Pair-wise comparisons at each generating rate showed that the recall accuracy of repeating bits was higher than that of switching bits [ $p$ 's $<.001$ ] at switch rates $0.1,0.3$, and 0.5 .


Figure 2. Experiment 2. (a) Participants $(\mathrm{N}=45)$ were presented with 10 -bit temporal or spatial sequences, and recalled the sequences. Accuracy was calculated as the proportion of exact matches between the presented sequence and the recalled sequence for temporal trials (b) and spatial trials (c). From the second bit on in each sequence, recall accuracy of each bit was calculated depending on whether the bit repeated the previous bit, or switched from the previous bit, for temporal sequences (d) and spatial sequences (e). We also calculated the switch rate of the recalled sequences, plotted with observed switch rate of the presented
sequences for temporal trials (f) and spatial trials (g). (Error bars reflect $\left.\pm 1 \mathrm{SEM} ;{ }^{*} p<.05,{ }^{* *} p<.01,{ }^{* * *} p<.001\right)$

One problem with the accuracy measure based on exact matches was that it penalizes cases where participants reversed or misplaced bits but were nonetheless accurate. To circumvent this problem, we conducted another analysis where we calculated the switch rate of the recalled sequence, and compared that to the observed switch rate of the presented sequence (Figure 2 f and g ).

We computed signed error (switch rate of the recalled sequences - observed switch rate) separately for temporal and spatial trials. For temporal trials (Figure 2f), a one way repeated-measures ANOVA revealed a significant difference in signed error across the five generating switch rates $\left[F(4,176)=140.7, p<.001, \eta_{p}{ }^{2}=0.76\right]$. Post-hoc Tukey HSD analysis showed all pair-wise comparisons were significant except between 0.1 and 0.3 , and 0.1 and 0.5 , suggesting that errors were greater at higher switch rates. For spatial trials (Figure 2g), the same ANOVA revealed a significant difference in signed error across the five switch rates $\left[F(4,176)=92.54, p<.001, \eta_{p}^{2}=0.68\right]$. Post-hoc Tukey HSD analysis showed all pair-wise comparisons were significant except between 0.1 and 0.3 , and 0.1 and 0.5 , suggesting errors were greater at higher switch rates.

These results showed that as the sequence alternated more, recall accuracy diminished. The greater recall error in switching bits compared to repeating bits suggests that people are more likely to encode switches as repeats, than to encode repeats as switches.

## Experiment 3

What explains the encoding difficulty of switching bits? One explanation is that switching bits may be less salient than repeating bits. To examine salience, Experiment 3 used a change detection task where participants detected changes in two sequences that were presented one after another.

## Participants

Forty-five students ( 24 female, mean age $=20.6$ years, $\mathrm{SD}=1.8$ ) from UBC participated for course credit.

## Stimuli and Procedure

There were 200 trials in total. In each trial, participants were presented with two back-to-back sequences of 15 green and blue circles (Figure 3a). The color and size of the circles were identical to those used in Experiment 2. The sequences were generated with one of the five switch rates $(0.1$ to 0.9$)$ as before. There were 40 trials per switch rate, 20 of which contained a change where the color of one randomly selected circle was different between the two sequences, and 20 of which contained no change where the two sequences were the same. In each trial, all circles in the first sequence were presented simultaneously at the center of the screen for 500 ms , with an ISI of 500 ms , followed by the second sequence also presented for 500 ms . Participants had to judge whether the two sequences were the same or different
by pressing the " Y " key or the " N " key, respectively. The trials were presented in a random order.

## Results and Discussion

To examine the performance of the change detection task, we calculated A' by dividing the average of correct rejection rate and correct hit rate by two, then adding 0.5 to the resultant number (Pollack \& Noman, 1964). A' was plotted across the five generating switch rates (Figure 3b). There was a reliable difference in A' across the five rates [ $\left.F(4,176)=24.64, p<.001, \eta_{p}{ }^{2}=0.38\right]$. Post-hoc Tukey HSD analysis showed all pair-wise comparisons were significant except for between $0.5,0.7$, or 0.9 .


Figure 3. Experiment 3. (a) Participants ( $\mathrm{N}=45$ ) viewed two back-to-back sequences, and judged if the two sequences were the same or different. (b) Performance was assessed using A'. (c) Trials with changes were categorized into three change groups: 1. repeats to switches, 2. switches to repeats, and 3. switches to switches. (Error bars reflect $\pm 1 \mathrm{SEM} ; * * * p<.001$ )

We also examined change detection accuracy depending on the local environment where the change occurred. For all change trials, we categorized them into three groups: repeats to switches (e.g., 000 to 001,010 , or 100), switches to repeats (e.g., 010,001 , or 100 to 000), and switches to switches (e.g., 001 to 011 or 101,010 to 110 or 011,100 to 101 or 110). Since we only considered trials where a change occurred, there was no false alarm. Therefore, we used accuracy as the measure here (Figure 3c). Among the three types changes, there was a reliable difference in accuracy $\left[F(2,88)=55.95, p<.001, \eta_{p}{ }^{2}=0.56\right]$. Post-hoc Tukey HSD analysis showed that accuracy in the repeats to switches group was reliably higher than that in the switches to repeats and switches to switches groups [ $p$ ' $\ll .001$ ].

These results showed that as the sequence became more alternating, a change in the sequence was harder to detect. This suggests that repetitions were more salient than alternations. Moreover, a change was more salient when a streak was interrupted, than when an alternating pattern became streaky or remained alternating. This differential performance suggests that people may have paid more attention to the streak presented in the first sequence, than to the switches presented in the first sequence.

## Experiment 4

To provide further support for the salience account, Experiment 4 used a visual search task to measure attention to switching vs. repeating sequences.

## Participants

Forty-five students ( 33 female, mean age=19.6 years, $\mathrm{SD}=2.1$ ) from UBC participated for course credit.

## Stimuli and Procedure

As in Experiment 3, there were 200 trials, and in each trial, a sequence containing 15 colored circles were presented simultaneously on the screen. As before, the sequences were generated with one of the five switch rates, and there were 40 trials per switch rate. For each trial, participants had to search for a target (a red arrow pointing left " $<$ " or right " $>$ ") in one of the randomly selected circles in the sequence. They were asked to identify the direction at which the arrow was pointing as fast and as accurately as they could (Figure 4a). Half of the trials contained an arrow pointing left, and the other half contained an arrow pointing right. Each sequence was presented for 1500 ms . The trials were presented in a random order.


Figure 4. Experiment 4. (a) Participants viewed 15 -bit spatial sequences. The target was a small red arrow, pointing to the left or right, in one of the circles. Participants reported the direction of the arrow as fast and as accurately as they could. (b) Response time of correct trials was plotted. (Error bars reflect $\pm 1$ SEM)

## Results and Discussion

The accuracy of the target search task was high (mean $=97.5 \%, \mathrm{SD}=2 \%$ ). Thus, we only examined the response times of correct trials as our measure of attention (Figure 4b). There was a reliable difference in response time across the five switch rates $[F(4,176)=2.55, p<.05$, $\left.\eta_{p}{ }^{2}=0.05\right]$. Post-hoc Tukey HSD analysis showed a reliable difference in response times only between switch rates 0.1 and 0.5 . This result showed that participants were faster to find the target in sequences with more repetitions than with more switches. One explanation is that repeating sequences may draw attention more strongly than switching sequences.

## General Discussion

The goal of the current study was to examine how people represent alternations vs. repetitions in a binary sequence. Across four experiments using estimation, working memory, change detection, and visual search tasks, we found that the number of alternations was under-estimated more strongly than the number of repetitions (Experiment 1). This under-estimation of switches could be explained by the fact that recall accuracy diminished as the sequence became more alternating (Experiment 2). The greater encoding difficulty of alternations could be explained by the finding that changes were harder to detect as the sequence became more alternating (Experiment 3). Finally, visual targets were slower to be found as the sequence became more alternating, suggesting that alternating sequences draw attention less strongly than repeating sequences (Experiment 4). Overall, these results converge to support the same finding that people are more blind or insensitive to alternations than to repetitions, which suggests that alternations are under-represented compared to repetitions.

The current findings support a new account on the overalternation bias. Specifically, there is a perceptual limitation in the ability to accurately represent alternations as opposed to repetitions in a binary sequence. This means that for people to perceive a 0.5 switch rate, the sequence must contain more than $50 \%$ alternations (in fact around $70 \%$ ).

Why are alternations under-represented compared to repetitions? We offer two explanations. First, two alternating bits (e.g., 10) may be perceptually more complex than two repeating bits (e.g., 11), and this higher complexity in an alternation could be more difficult to encode. Second, people may implicitly chunk an alternation into a unit (e.g., perceiving 101010 as three chunks of 10 , Zhao \& Yu, 2016), but rely on numerosity perception for repetitions (e.g., perceiving 111111 as 1 repeating five times).

The current study reveals a perceptual limitation in the representation of alternations. The study is important in several ways: first, it provides a new explanation of the over-alternation bias in randomness perception; second, it reveals new insights on the limits in the perception of binary information; and finally, the same finding was replicated in four different paradigms using different measures. The current findings shed light on how people process binary information, which is fundamental to understanding the limits of the cognitive system.

## Acknowledgments

This work was supported by NSERC Discovery Grant (RGPIN-2014-05617 to JZ), Canada Research Chairs program (to JZ), Leaders Opportunity Fund from the Canadian Foundation for Innovation (F14-05370 to JZ), and Canada Graduate Scholarship Master's program and Elizabeth Young Lacey Fellowship (to RY).

## References

Ayton, P., Hunt, A. J., \& Wright, G. (1989). Psychological conceptions of randomness. Journal of Behavioral Decision Making, 2, 221-238.

Baddeley, A. D. (1966). The capacity for generating information by randomization. Quarterly Journal of Experimental Psychology, 18, 119-129.
Bar-Hillel, M., \& Wagenaar, W. A. (1991). The perception of randomness. Advances in Applied Mathematics, 12, 428-454.
Beltrami, E. (1999). What Is Random? Chance and Order in Mathematics and Life. New York: Springer-Verlag.
Chater, N., \& Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science?. Trends in cognitive sciences, 7, 19-22.
Falk, R., \& Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. Psychological Review, 104, 301-318.
Faul, F., Erdfelder, E., Lang, A. G., \& Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods, 39, 175-191.
Fitelson, B., \& Osherson, D. (2012). Remarks on random sequences. Retrieved from http://arxiv.org/abs/1205.5865
Gilovich, T., Vallone, R., \& Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17, 295-314.
Hahn, U., \& Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. Psychological Review, 116, 454-461.
Julesz, B. (1962). Visual pattern discrimination. IRE Transactions on Information Theory, 8, 84-92.
Kahneman, D., \& Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430454.

Kareev, Y. (1992). Not that bad after all: Generation of random sequences. Journal of Experimental Psychology: Human Perception and Performance, 18, 1189-1194.
Lopes, L. L., \& Oden, G. C. (1987). Distinguishing between random and nonrandom events. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 392-400.
Miller, J. B., \& Sanjurjo, A. (2015). Is It a Fallacy to Believe in the Hot Hand in the NBA Three-Point Contest?
Nickerson, R. S. (2002). The production and perception of randomness. Psychological Review, 109, 330-357.
Olivola, C. Y., \& Oppenheimer, D. M. (2008). Randomness in retrospect: Exploring the interactions between memory and randomness cognition. Psychonomic Bulletin \& Review, 15, 991-996.
Oskarsson, A. T., van Boven, L., McClelland, G. H., \& Hastie, R. (2009). What's next? Judging sequences of binary events. Psychological Bulletin, 135, 262-285.
Pollack, I., \& Norman, D. A. (1964). A non-parametric analysis of recognition experiments. Psychonomic science, 1, 125-126.
Rapoport, A., \& Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. Journal of Experimental Psychology: General, 121, 352-363.
Tversky, A., \& Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.
Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of the literature. Psychological Bulletin, 77, 65-72.
Yu, R., Gunn, J., Osherson, D., \& Zhao, J. (in press). The consistency of the subjective concept of randomness. Quarterly Journal of Experimental Psychology.
Zhao, J., Hahn, U., \& Osherson, D. (2014). Perception and identification of random events. Journal of Experimental Psychology: Human Perception and Performance, 40, 1358.
Zhao, J., \& Yu, R. (2016). Statistical regularities reduce perceived numerosity. Cognition, 146, 217-222.

# Back to ABCs: Clustering Alphabetically, Rather than Semantically, Enhances Vocabulary Learning 

Jingqi Yu (jy45@umail.iu.edu)<br>Department of Psychological and Brain Sciences, 1101 East $10^{\text {th }}$ Street, Bloomington, IN 47405 USA<br>Veronica X. Yan (veronicy@usc.edu)<br>Department of Psychology, University of Southern California, 3620 South McClintock Ave. Los Angeles, CA 90089 USA

Elizabeth Ligon Bjork (elbjork@psych.ucla.edu)<br>Department of Psychology, University of California, Los Angeles, 502 Portola Plaza, Los Angeles, CA 90095 USA

Robert A. Bjork (rabjork@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles, 502 Portola Plaza, Los Angeles, CA 90095 USA


#### Abstract

Optimizing the study of vocabulary words for highstakes tests such as the SAT or GRE prep can be problematic, given that many words are semantically, orthographically, or phonologically confusable. Companies marketing test preparation programs make multiple recommendations, such as clustering words on some basis, but little research has been carried out to examine what that basis should be. Across two experiments, we compare the efficacy of different types of clustering-categorical, alphabetical, and confusable--for the learning of semantically related words (Experiment 1) and confusable words (Experiment 2). We demonstrate that, in contrast to most learners' intuitions, an alphabetical sequence yields superior learning.


Keywords: memory, vocabulary learning, optimal sequencing, semantic clustering, alphabetical clustering

## Introduction

Vocabulary learning is a crucial component of learning languages, not only because knowing some minimum number of words is needed for a basic level of communication in a new language, but also because increasing one's vocabulary in one's own language can be critical in the context of high-stakes testing of various types. Applicants to undergraduate or graduate programs, for example, often spend months preparing for standardized examinations, such as the ACT, SAT, or GRE, and memorization plays an essential role in such preparation. Such vocabulary learning can be especially daunting for international applicants from non-English speaking backgrounds. To meet this demand from anxious testtakers, a huge test-preparation industry has sprung up, with each different business promising a different set of "secrets" to crack the SAT and GRE codes for a price.

Many of these organizations make recommendations for how very large sets of new words may be learned, including (a) using mnemonics (e.g., word imagery), (b) grouping words by their category membership, and (c)
by Greek/Latin roots. Test preparation programs-and hence, their students-often strongly promote one over the other, yet there has been only minimal research testing the relative efficacy of such methods.
The best way(s) to learn new vocabulary, therefore, is still an open question, and what method yields the best learning outcomes may in part depend on the characteristics of the to-be-learned words. When it comes to learning new vocabulary, for example, in preparation for the GRE or SAT exams, there are two large sources of difficulty. First, these tests require individuals to learn and distinguish between many semantically related words (e.g., personality traits: mendacious-callow). Second, there is a need to distinguish between confusable but semantically distinct words, such as words that are similar-looking and/or similarsounding (e.g., decry-descry).
While little to no research has been conducted on optimal sequencing for the learning of confusable pairs, there has been some research on semantic clustering, but evidence in support of semantic clustering, however, has been mixed. In the present studies, we specifically examine two popular methods-clustering by semantic category or by confusability, and alphabetically-clustered-and compare them against a random sequence. We examine the alphabetically-clustered sequence for a practical reason: lists of words are often organized alphabetically, and for this reason, there are many who may attempt to learn words in that order, out of convenience. Indeed-at least anecdotally-studying words alphabetically appears to be common among Chinese students preparing for the SAT and GRE exams. This alphabetical organization is not necessarily a conscious, explicit strategy, but simply a byproduct of how reading typically proceeds (i.e., start on page 1 and work your way through to the end). Studying words alphabetically does offer some structure,
but unlike semantic clustering or clustering of confusable words, clustering by initial letter appears somewhat arbitrary. However, given that it is a strategy that is widely used, and little is known as to whether this strategy is truly beneficial for vocabulary learning, we examine it in the present studies.

## Semantic Clustering

"Semantic clustering" refers to the practice of grouping vocabulary words into different categories based on their meanings (Tinkham, 1993). Such clustering is believed to be effective for several reasons, including that presenting words in semantic clusters allows for intra-category and inter-relational reinforcement (e.g., Seal, 1991), makes the meanings of words clearer by enabling learners to notice fine-grained distinctions between words (e.g., Gairns \& Redman, 1986), better reflects semantic networks in the "mental lexicon" (e.g., Aitchison, 2002), and draws attention to the semantics which may lead to deeper levels of mental processing (Erten \& Tekin, 2008). On the other hand, Schneider, Healy, and Bourne (1998) found that while clustering words semantically aided initial learning, it appeared to hinder relearning a week later.
What constitutes a "semantic cluster" has, however, differed greatly across researchers. Possible constructs include, but are not limited to, near synonyms (e.g., man, fellow, and guy; (Hippner-Page, 2000), topic-related items (e.g., crime: smuggling, jury, and court; Papasanasiou, 2009), and exemplars that fall under a super-ordinate category (e.g., fruit: apple, pear, and peach) (Waring, 1997). There is some, albeit limited, evidence supporting the facilitating effects of semantic clustering (e.g., Finkbeiner \& Nicol, 2003), but these conclusions have largely been drawn from examination of acquisition during training, rather than on the basis of long-term memory tests. Considerable research, however, has demonstrated that the manipulations that boost performance during training do not always boost learning (see, e.g., Soderstrom \& Bjork, 2015). Furthermore, conclusions are hard to draw because the random and semantically clustered conditions compare learning of different sets of words, making it unclear whether it is the sequence that enhances learning or whether a semantically related set of words is simply easier to learn.

## The Present Studies

We conducted two studies to examine the optimal sequence of vocabulary learning. Although previous studies have used separate lists for different conditions (e.g., Tinkham, 1993, 1997), we created one single set of GRE words for each experiment, ensuring that only sequencing would differ across conditions. Moreover, to test long-term memory, rather than just performance during acquisition,
we employed final criterion tests that were delayed by at least 24 hours.
In Experiment 1, we examined the optimal sequencing for learning semantically related words by comparing the efficacy of alphabetical, categorical (i.e., semantic clusters), and random sequences.

In Experiment 2, we examined the optimal sequencing for learning confusable words by comparing the efficacy of alphabetical, paired (confusable clusters), and random sequences. Whether confusable clusters or random sequencing creates more difficulties is unclear. Because confusable pairs are similar-looking and/or similarsounding, similarities in orthography and/or phonology could well interfere with performance during training, but then might also enhance retention performance. Alternatively, an alphabetical sequence resembles a hybrid of random and paired schedules to some degree, and thus might incorporate the best (or worst) of both worlds.

## Experiment 1

The purpose of Experiment 1 is to address the first challenge of vocabulary learning-the need to distinguish between semantically related words.

## Participants and Design

Participants were 152 undergraduates from the University of California, Los Angeles (UCLA) who participated in the experiment in exchange for course credit. Eight-two of them were native English-speakers, and 70 spoke English as a Second Language (ESL; $M_{\text {length of speaking English }}=5.34$ years). ESL students were expected to be at a decent English level due to the University's requirements. Participants were randomly assigned to one of three sequencing conditions: categorical, alphabetical, and random.

## Materials

A pool of 36 GRE word-synonym pairs was selected from GREedge, a theme-based word list. This word list contained six words from each of six different categories: Communication, Crime \& Law, Nature, Personality, Thoughts \& Ideas, and Time. College-level participants were expected to know the meaning of each paired synonym. Table 1 shows an example of the "Communication" related words. Within each category, there was one word that began with one of six initial letters-a, $c, i, m, p$, and $s-$ which allowed us to construct the six sets of words for the alphabetical condition. Table 2 shows an example of the words that begin with the letter $a$.

Table 1
Example of one of the six-word categorical sets: words relating to "Communication"

| Category | Initial | GRE Word | Synonym |
| :--- | ---: | :--- | :--- |
| Communication | $a$ | acrimonious | bitter |
|  | $c$ | circumlocution | rambling |
|  | $i$ | importune | beg |
|  | $m$ | missive | letter |
|  | $p$ | prattle | babble |
|  | $s$ | sententious | pithy |

Table 2
Example of one of the six-word alphabetical sets: words beginning with the letter $a$

| Initial | Category | GRE Word | Synonym |
| :--- | :--- | :--- | :--- |
| a | Communication | acrimonious | bitter |
|  | Crime \& Law | abjure | withdraw |
|  | Nature | arroyo | environment |
|  | Personality | abstemious | restrained |
|  | Thoughts \& Ideas | apotheosis | exaltation |
|  | Time | antediluvian | ancient |

## Procedure

Participants were told that their task was to learn 36 GRE words. Participants were presented one GRE wordsynonym pair at a time, and, for each pair, they were first asked to generate the synonym ( 8 sec ) before they were shown the correct answer ( 3 sec ). Therefore, the study phase consisted of tests-with-feedback trials (for a discussion of the benefits of testing and feedback on long-term retention, see Roediger \& Karpicke, 2006). The six pairs of a given set (e.g., a random set, a category set, or an alphabetical set) were always presented consecutively. Each set were presented three times, with the order of pairs randomized each time (hence, the alphabetical sequence was not strictly alphabetical, but grouped words starting with the same initial letter). Thus, any given pair was presented on an average spacing interval of 5-5. Between individuals, the order of the six sets was randomized.
After they completed studying all 36 pairs ( 3 times each), participants were asked to predict how many of the 36 pairs they would be able to answer correctly on the test the next day. They were then informed of the three different study sequences that had been used for different participants and asked to judge which one they believed would be most effective for learning vocabulary.
Twenty-four hours later, participants were emailed a link to take the final test. The test phase consisted of two portions: First, they were asked to complete a cued-recall task in which they were presented with the GRE word and asked to generate the synonym. Second, they were given a multiple-choice in which they were given the synonym and four options. The options, illustrated in Figure 1, were constructed such that one was the correct answer, one was a wrong answer from the same category, one was a wrong
answer with the same initial letter, and one was a random lure selected from the other GRE words they had studied.
Finally, we collected information about participants' GRE preparation status and language fluency and number of languages spoken.


Figure 1: Example of a multiple-choice trial in Experiment 1. Participants were only shown contents inside the square. The rest is only for illustration purpose.

## Results and Discussion

Sixteen participants were removed from subsequent analysis for the following reasons: five looked up answers, seven reported serious technical issues with the experiment, and four people indicated both. Among the remaining 136 participants, 74 were native English-speakers and 62 were ESL students ( $M_{\text {length of speaking English }}=5.42$ years).

Acquisition The increase in acquisition from the to the third time a given pair was presented during the study phase is shown in the left panel of Figure 2. By the end of the final trial, there was a trend for a difference between the three conditions, $F(2,133)=.2 .24, M S E=.07, p=.11$. Post-hoc analyses revealed that the random condition ( $M=$ $.73, S D=.18$ ) was marginally worse than the alphabetical ( $M=.79, S D=.18, p=.06$ ) and categorical conditions ( $M=.79, S D=.18, p=.08$ ). No significant difference between the alphabetical and categorical conditions was obtained, $p=.96$.

Final test Performance on the final cued recall test, illustrated in the right panel of Figure 2, suggested some differences (although limited) between conditions, $F(2,133)$ $=3.105, M S E=.04, p=.048$, but the pattern was somewhat different from that of the acquisition pattern. Post-hoc pairwise comparisons revealed that participants in the alphabetical condition $(M=.50, S D=.03)$ performed significantly better than those in the random condition ( $M=.40, S E=.03$ ), $p=.02$ ), but neither was significantly different from the categorical condition ( $M=$ $.44, S D=.21$ ), $p \mathrm{~s}>.10$. No significant differences in performance on the multiple-choice test among conditions were observed, $F(2,133)=1.764, p=.18$.


Figure 2: Study phase acquisition curves and final cued recall test in Experiment 1. Error bars represent one standard error of the mean.

Metacognitive responses With respect to the question "What sequence do you think is best for learning?" asked of participants at the end of the acquisition phase, 94 (70\%) of the 134 participants responding reported believing that categorical clustering would lead to the best learning, 35 ( $26 \%$ ) reported believing that a random sequence would be best, and only 5 (4\%) reported believing that an alphabetical sequence would be best. These metacognitive responses did not differ by assigned condition, $\chi^{2}(4)=7.34, p=.12$.
Overall, while Experiment 1 neither show evidence in favor of nor against categorical sequencing, it surprisingly suggested some benefits associated with an alphabetical sequencing strategy. Participants' metacognitive beliefs, however, showed limited faith in this "new" strategy.

## Experiment 2

Experiment 2 aimed to address the second challenge in GRE word learning: students' ability to distinguish between similar-looking and/or similar-sounding words.

## Participants and Design

Participants were 112 undergraduates from UCLA who participated in the experiment in exchange for course credit. Sixty-five of those participants were native Englishspeakers, and 47 ESL students ( $M_{\text {length of speaking English }}=5.60$ years). Despite being non-native English-speakers, these ELS students were expected to be at a decent English level due to the University's requirements. Participants were randomly assigned to one of three sequencing conditions: paired, alphabetical, and random.

## Materials

As in Experiment 1, we selected 36 GRE word- synonym pairs from several sites (e.g., Magoosh). However, each GRE word was also matched with another GRE word that was highly confusable. That is, 36 GRE words were broken
down into 18 confusable pairs. The words in nine of these pairs shared the same initial letters (e.g., augur-bode and auger- drill) and the remaining nine of these pairs did not (e.g., astringent-bitter and stringent-strict). Across all the GRE words, any initial letter was also made to appear at least three times (e.g., three words that began with ' $a$ '). College-level participants were expected to know the meaning of each paired synonym.
In the random condition, one randomized sequence of the 36 words was created for each participant. In the paired condition, the confusable paired-GRE words were always presented consecutively (e.g., augur-bode followed by auger- drill, or vice versa), and the order of pairs was randomized for each participant. In the alphabetical condition, the GRE words were presented in alphabetical order. Table 3 shows an example of two confusable pairs in alphabetical order: One same-initial pair (veracious and voracious) and one different-initial pair (pabulum and vinculum). As demonstrated below, Same-initial pairs by definition would still appear close to each other in alphabetical order (not necessarily consecutively), but different-initial pairs would for sure be shuffled.

## Table 3

Example of two confusable pairs in alphabetical order: One same-initial pair and one different-initial pair.

| Initial | GRE Word | Synonym |
| :---: | :--- | :--- |
| p | pabulum | sustenance |
| $\ldots$ | $\ldots$ | $\ldots$ |
| v | veracious | honest |
|  | vinculum | bond |
|  | voracious | greedy |

## Procedure

The procedure of Experiment 2 was similar to that of Experiment 1 with three exceptions: (a) Instead of repeating in sets of six, all 36 pairs were presented before they were repeated, yielding an average spacing interval of 35-35; (b) the alphabetical order here was strictly alphabetical, so the order of words in each of the three cycles of 36 trials was the same; and (c) the multiplechoice test used the GRE word as the cue and presented four studied synonyms as the options (the correct answer, the synonym of the confusable GRE word, a synonym of a word sharing the same initial letter, and another random synonym).

## Results and Discussion

Sixteen participants were removed from subsequent analysis: five looked up answers, and 11 reported serious technical issues with the experiment. Among the remaining 96 participants, 57 were native English-speakers and 39 were ESL students ( $M_{\text {length of speaking English }}=5.76$ years).

Acquisition The increase in acquisition from the first to the third time a given pair was presented during the study phase is shown in the left panel of Figure 3. Participants in different conditions exhibited a similar amount of accuracy boost by the end of the study phase, and the proportions of synonyms correctly recalled were not different, although performance in the alphabetical condition ( $M=.43, S D=.20$ ) was numerically better than that in the random $(M=.39, S D=.18)$ and paired $(M=$ $.37, S D=.18$ ) conditions. Unlike in Experiment 1, the curves were nearly linear with no indication of deceleration in learning rates at the end of the $3^{\text {rd }}$ presentation. One interpretation of this finding was that intrinsic confusion in confusable words made them naturally more difficult to learn than regular ones (indeed, accuracy rates were lower). Thus, three learning trials still left space for learning to improve at an accelerating rate before starting to slow down.


Figure 3: Study phase acquisition curves and final cued recall test in Experiment 2. Error bars represent one standard error of the mean.

Final test Performance on the final cued recall test is illustrated in the right panel of Figure 3. A 3 (sequencing condition) x 2 (pair type) mixed effects ANOVA performed on the cued-recall test performance, revealed a main effect of condition, $F(2,93)=3.45, M S E=.10, p=$ .036. Additionally, post-hoc pairwise comparisons revealed that the alphabetical condition $(M=.53, S D=25)$ yielded significantly better learning than the paired condition ( $M=$ $.39, S D=.19, p=.01)$. The random condition $(M=.44$, $S D=.25$ ) was not significantly different from either the paired condition ( $p=.33$ ) or the alphabetical condition ( $p$ $=.12$ ).

A main effect of pair type was also observed, $F(1,93)=$ $50.55, M S E=.01, p<.001$, with the GRE words from the pairs having different initial letters $(M=.50, S D=.23)$ being learned significantly better than the words from pairs sharing the same initial letter $(M=.39, S D=.24)$. There
was no interaction between condition and pair type, $F(2,93)=1.12, M S E=.01, p=.33$.

Figure 4 shows multiple-choice test performance by condition and pair type. A 3 (sequencing condition) x 2 (pair type) mixed effects ANOVA performed on the multiple-choice test performance revealed similar patterns to that obtained for the cued-recall test performance. Again, there was a trend-level effect of condition, $F(2,93)=2.22$, $M S E=.10, p=.12$. Pairwise comparisons revealed a similar pattern as found with the cued-recall test: the alphabetical condition ( $M=.77, S D=.03$ ) was marginally better than the paired $(M=.70, S D=.03, p=.06)$ and the random $(M=.70, S D=.03, p=.08)$ conditions.

Again, a significant effect of pair type, $F(1,93)=53.03$, $M S E=.01, p<.001$, was observed, with the differentinitial letter pairs $(M=.77, S D=.18)$ learned better than the same-initial letter pairs $(M=.67, S D=.18)$. Finally, there was no condition $x$ pair type interaction, $F(2,93)=1.57, M S E=.01, p=.21$


Figure 4: Multiple-choice test performance on Experiment 2 by condition and pair type. Error bars represent one standard error of the mean.

Metacognitive Responses When asked what sequence they believed was best for learning, 42 (44\%) participants reported that they believed that pairing confusable words would lead to the best learning, 39 ( $41 \%$ ) a random sequence would be best, and 15 ( $16 \%$ ) believed that an alphabetical sequence would be best. These metacognitive responses did not differ by experienced condition, $\chi^{2}(4)=6.30, p=.18$.

Overall, the findings of Experiment 2 demonstrate that when trying to learn confusable words, contrary to many people's belief (a majority of $84 \%$ thought that either random or paired order would be the best), an alphabetical order led to the greatest learning.

## General Discussion

The current study is one of the very few instances where the alphabetical order has been studied. Experiments 1 and 2 revealed some preliminary evidence on the merits of an
alphabetical word learning sequence. There was some suggestion that this strategy can be equally, if not more effective than strategies that are traditionally considered good (e.g., categorical or paired clustering).
It is worth pointing that while both being referred to as an "alphabetical sequence," the structures of the alphabetical order in Experiments 1 and 2 were not identical. In Experiment 1, the structure was not strictly alphabetical, but rather grouped words beginning with the same initial letter together. Participants went through an alphabetical cluster three times, so they were only exposed to one initial letter at any given time. Thus, it was more of an "alphabet-informed grouping." We speculate that one possibility is that this grouping offers an optimal level of support and difficulty: The support from the small degree of structure (i.e., shared initial letters) may ease extraneous cognitive processing load from the difficult word learning task; the otherwise-random nature of the words maintains a sense of difficulty to promote deeper processing. An alternative explanation could simply be that clustering alphabets gives learners another, redundant, cue to aid learning. The initial letter is another cue to the context in which words were learned. For example, the initials may have already narrowed down the list from 36 to six words, which may well require less cognitive effort to identify the correct answer, given that words in the same initial cluster are not too similar orthographically/phonologically. Therefore, while shared categories may be a contextual cue in categorical clustering, list position could be a contextual cue in an alphabetical sequencing. Each clustering method has its own advantage and both lead to respectable results.
In Experiment 2, the alphabetical sequence was truly alphabetical, with half same-initial and half different-initial confusable pairs. The advantage of this strategy was primarily reported in learning different-initial than sameinitial pairs. Thus, the benefit of alphabetizing a confusable word-list was observed at a global (the entire list with multiple initials) rather than a local (same-initial pairs grouped to the same alphabetical cluster) level. Consequently, an alphabetical condition represents a hybrid of randomization of the easier pairs (i.e., different-initial pairs) and confusability- clustering of the more difficult pairs (i.e., same initial pairs), which may have incidentally created a degree of "desirable difficulty" (Bjork, 1994). Alternatively, because learners in Experiment 2 learned all 36 words before repeating, those in the alphabetical condition may have simply used list position as a contextual cue to aid learning. For example, they may have linked to-be-learned words to some known knowledge (i.e., an alphabetical list) to help memorize.
We have already demonstrated some caveats when alphabetizing to-be-learned words. As suggested above, there might be differences in the role of "alphabet-informed grouping vs. alphabetical order. It is therefore unclear how far the benefits of an alphabetical sequence would extend. In the present studies, we presented participants relatively difficult words in a language that they were familiar with

Hence, knowledge about word etymology (e.g., Latin roots) or even passing familiarity with the to-be-learned words themselves (note the large jump in performance between the first and second trial of the study phase) may have supported learning. It is unclear how the optimal sequence might change for foreign language learning, where learners do not have this type of background knowledge.

As part of a critical factor in high-stakes tests, GRE word learning is a major concern of many students. The present studies extend the literature by suggesting the powerful potential of learning words in alphabetical order, a widely used, yet under-investigated alternative to clustering or random sequencing. Whether, however, the benefits of an alphabetical sequencing might generalize to vocabulary words that are less difficult and abstract than GRE words remains to be seen. In the meantime, it appears that generations of Chinese students who have been learning English vocabulary words grouped alphabetically may not have been engaging in what may seem, by some arguments, to be a misguided practice.

## References

Aitchison, J. (2002). Words in the Mind: An Introduction to the Mental Lexicon (3rd Ed.). Blackwell Publishers: Great Britain.
Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe \& A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185-205). Cambridge, MA: MIT Press.
Erten, I.H., \& Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semanticallyunrelated sets. System, 36 (3), 407-422.
Finkbeiner, M., \& Nicol, J. (2003). Categorical category effects in second language word learning. Applied Psycholinguistics, 24(3), 369-383.
Gairns, R., \& Redman, S. (1986). Working with words: A guide to teaching and learning vocabulary. New York: Cambridge University Press.
Hippner-Page, T. (2000). Semantic clustering versus thematic clustering of English vocabulary words for second language instruction: Which method is more effective? Retrieved from ERIC database. (ED445550)
Seal, B. D. (1991). Vocabulary learning and teaching. In M. CelceMurcia (Ed.), Teaching English as a second or foreign language (2nd ed., pp. 296-311). Boston: Heinle \& Heinle.
Schneider, V. I., Healy, A. F., \& Bourne, L. E., Jr (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy \& L. E. Bourne, Jr. (Eds.), Foreign Language Learning: Psycholinguistic Studies on Training and Retention (pp. 77-90). Mahwah, NJ: Erlbaum.
Soderstrom, N. C., \& Bjork, R. A. (2015). Learning versus performance: An integrative review. Perspectives on Psychological Science, 10, 176-199.
Tinkham, T. (1993). The effects of semantic clustering on the learning of second language vocabulary. System, 21, 371-380.
Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. Second Language Research, 13, 138-163.
Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. System, 25, 261-274.

# Domain-General Learning of Neural Network Models to Solve Analogy Tasks - A Large-Scale Simulation 

Arianna Yuan (xfyuan@stanford.edu)<br>Department of Psychology, Stanford University, Stanford, CA 94305 USA


#### Abstract

Several computational models have been proposed to explain the mental processes underlying analogical reasoning. However, previous models either lack a learning component or use limited, artificial data for simulations. To address these issues, we build a domain-general neural network model that learns to solve analogy tasks in different modalities, e.g., texts and images. Importantly, it uses word representations and image representations computed from large-scale naturalistic corpus. The model reproduces several key findings in the analogical reasoning literature, including relational shift and familiarity effect, and demonstrates domain-general learning capacity. Our model also makes interesting predictions on cross-modality transfer of analogical reasoning that could be empirically tested. Our model makes the first step towards a computational framework that is able to learn analogy tasks using naturalistic data and transfer to other modalities.


Keywords: analogical reasoning; learning; cross-modality transfer; neural network models.

## Introduction

Analogy is arguably one of the most important mechanisms through which people acquire new knowledge (Gentner, Holyoak, \& Kokinov, 2001). Verbal analogy task such as "PIG:BOAR::DOG:? a. WOLF b. CAT" and visual analogy task such as Raven's Progressive Matrices has been widely used in standardized test to assess students' intellectual ability (Buck et al., 1998). For a long time, there has been a heated debate over whether analogy-mapping is a domain-specific or domain-general process (Forbus, Gentner, Markman, \& Ferguson, 1998). Nowadays, it has been increasingly clear that there are both a domain-general component and a domain-specific component in learning analogical reasoning.

One phenomenon related to the domain-general component is the relational shift in child development, i.e., early in development children tend to choose the item that is more associated to the third item in the analogy question A:B::C:[D1|D2] (Sternberg \& Nigro, 1980), and only older kids are able to use relational matching rather than associations to perform the task. Some researchers have argued that the lack of inhibitory control, which requires a full-blown pre-frontal cortex is partially responsible for the associative response (Richland \& Burchinal, 2013). The ability to inhibit associative responding and maintain the relational constraints imposed by $\mathrm{A}: \mathrm{B}$ is domain-general, i.e., it is a universal prerequisite for analogy-making no matter which sensory modality or semantic domain the analogy task is built on. Hence, if one person is trained to perform analogy task in a particularly domain or modality, it is likely that the training will help them do better in other domain or modality,
due to enhanced ability to suppress associative responses and to maintain contexts.

The evidence for a domain-specific component comes from the finding that children's ability to perform analogy task depends on their familiarity with the test material. Goswami and Brown (1990a, 1990b) found that when familiar concepts were used in analogy tasks, performances were much better. These findings underscore the contribution of domain-specific knowledge in analogy-making, yet change in this aspect is unlikely to boost task performance in other domain or modality.

Several computational models have been proposed to account for how people solve analogy tasks. A useful classification scheme is to group them into three types of models (French, 2002; Gentner \& Forbus, 2010): symbolic models (Kuehne, Forbus, Gentner, \& Quinn, 2000; Falkenhainer, Forbus, \& Gentner, 1989), connectionist models (Holyoak \& Thagard, 1989; Hummel \& Holyoak, 1997; Kollias \& McClelland, 2013) and hybrid models (Mitchell, 1993; Kokinov \& Petrov, 2000). Most of symbolic models represent analogy questions using predicates and logical forms. When asked to solve an analogy task such as $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ?, they use symbolic manipulations and search algorithms to find the correct answer. One of the most influential symbolic approach to analogy-mapping is the Structure Mapping Engine (SME) (Falkenhainer et al., 1989; Gentner, 1983). It represents the base and source using predicate-calculus and compares the two representations to see if there is any structural similarities between them. Once optimal matching structures are identified, the system then transfer structure in the source to the target. Later version of SME have relaxed matching criteria to allow similar, but not identical predicate to match (Brian, 1990), but whether two predicates are similar need to be explicitly computed.

Contrary to these symbolic approaches, connectionist models often use distributed representation for the items in an analogy task and encode their semantic and structural similarity in a more implicit and continuous way, e.g., Kollias and McClelland (2013). The connectionist models learn to make a correct response by adjusting connection weights so that the spreading activation within the neural networks could reveal the distributed representation of the target, thus leading to the correct answer.

Previous computational models of analogy-making have significantly deepened our understanding of the mental processes underling analogical reasoning. Many of such models claim that they provide a domain-general explanation of how people perform analogy tasks. For example, the

Structure Mapping Engine, which was originally introduced to solve analogy task in discrete semantic space, was later used to answer analogy questions in continuous visual domain (Lovett, Forbus, \& Usher, 2007). Most of the connectionist models are theoretically domain-general as well, since we can easily feed the distributed representation of stimuli from different domain/modality to the input placeholders of those models. It is important for these models to be domain-general, since humans are able to make within-modality generalization, such as recognizing examples that they have never encountered before (Lake, Salakhutdinov, \& Tenenbaum, 2015), or make cross-modality transfer (Hupp \& Sloutsky, 2011). Specifically, it has been shown that relational knowledge is critical to the development of analogical reasoning (Goswami, 1991). If one person has received sufficient training to solve verbal analogy tasks, they would gain some experience in relational knowledge. When later asked to solve a visual analogy-making task, they do not need to learn it from scratch (Figure 1).


Figure 1: Stimuli used in the current study. Left: Verbal analogy task. Right: Visual analogy task.

Despite the generalization ability that previous computational models may have claimed, they are often missing some critical components. Symbolic approaches, for example, rarely address how people learn to make an analogy. Since the knowledge representations and the search algorithms in those models are preprogrammed, it is not clear how experience of analogy-making in one domain could facilitate analogy-making in another unfamiliar domain.

As for the connectionist models, despite their theoretically domain-general nature, none of the previous studies has directly tested cross-modality transfer. Particularly, they only demonstrate within-modality generalization, i.e., the models are trained on some examples and tested on a different set of examples in the same modality. Also, the distributed representation of stimuli are either manually defined according to the semantic features of the items, or randomly assigned to some localist codes, which are not very naturalistic.

Finally, all the previous modeling works have used small datasets, containing hundreds of examples at most. We are wondering if we could build a model that scales up to handle a very large and naturalistic dataset. Particularly, we want to understand if we use a dataset that reflects the statistical distribution of stimuli in real life, can the model still learn analogical reasoning and even make cross-modality transfer? This idea is motivated by the statistical learning account of language acquisition (Frost, Armstrong, Siegelman,
\& Christiansen, 2015), which proposes that language acquisition relies partially on a domain-general mechanism, which is learning and processing sensory stimuli unfolding across time and space (Saffran, Aslin, \& Newport, 1996). Early since 1990s, researchers have found that if you train a recurrent neural network to predict the next word in a sentence, the word representation it learns reveals the syntactic and semantic role of the word (Elman, 1990). Inspired by these previous studies, we use distributed representations of words that reflect the statistics of word co-occurrence in everyday life, which are more naturalistic.

As for the representations of visual stimuli, we process the images (geometric figures) using a deep convolutional neural network that has been trained to perform object recognition task (Krizhevsky, Sutskever, \& Hinton, 2012), and use the activation of the 7th hidden layer as the representations of the visual stimuli. Previous studies have shown that deep convolutional networks share a lot of similarities with human visual system (Yamins, Hong, Cadieu, \& DiCarlo, 2013). After we obtain the representations (embeddings) of the words and the images, we build a simple, light-weighted neural network to learn the analogy tasks.

## Experiment

Data. We first describe the representation we use for the word. The distributed representation of words are computed using the continuous Skip-Gram model (Mikolov, Sutskever, Chen, Corrado, \& Dean, 2013). It takes the current word to predict the surrounding window of context words. Hence, the estimated word embeddings capture the semantic and syntactic role of the words (Mikolov, Yih, \& Zweig, 2013). We download the pre-trained word vectors from Google Word2 Vec ${ }^{1}$, which have 300 dimensions. For computational simplicity and efficiency, we reduce the dimensionality to 30 using principle component analysis (PCA) so that each word has a $30-\mathrm{d}$ vector representation.

We use the same verbal analogy dataset from Mikolov, Yih, and Zweig (2013) ${ }^{2}$, which contains 19529 examples in total with 907 unique words. We divide the dataset into three sets, a training set (percentage: $80 \%, 15634$ examples) ${ }^{3}$, a validation set ( $10 \%, 1955$ examples) and a test set $(10 \%$, 1955 examples). We use the accuracy on validation data to tune the hyper-parameters of the model and report accuracies on the test data. We run two types of tasks. The first one is $\mathrm{A}: \mathrm{B}:: \mathrm{C}:[\mathrm{D} 1|\mathrm{D} 2| \ldots \mid \mathrm{Dn}]$, in which the model is given some choices and has to select the correct one (Task 1). We simulate Task 1 with different number of choices ranging from 2 to 5 . The second type of task takes the form of $\mathrm{A}: \mathrm{B}:: \mathrm{C}: ?$, i.e., the model is required to find the correct D from all words in its vocabulary (Task 2).

To simulate associative responses children usually give

[^550]when first learning analogy tasks, we construct three datasets from the original analogy dataset. All of them are binary-choice questions, but the incorrect alternatives have different levels of associations with the third item C in the question. In the High Association Dataset, each questions has an incorrect response alternative (foil) that is strongly associated with the third word in the analogy question, whereas examples in the Low Association Dataset contain alternatives that is weakly associated with the third item. Finally, in the Random Association Dataset, the incorrect response alternatives are randomly selected from the vocabulary so that it does not necessarily has a strong or weak association with the third item C. We determine word associations by calculating the cosine distance between the word vectors of the two words. The smaller the distance, the stronger the association.

As for the visual stimuli, we use the Shape dataset from Reed, Zhang, Zhang, and Lee (2015). It is a dataset of 2-D colored shapes, with 8 colors, 4 shapes, 4 scales, 5 row and column positions, and 24 rotation angles. We only use one value for the rotation variable to avoid potential confusion (e.g., a square rotated $180^{\circ}$ would be the same figure as the original figure, but it has a different label in the dataset), and vary the other 5 variables to create a dataset. An example question is showed in Figure 1, right. We generate 19080 examples in total and randomly split them into a training set ( $80 \%$ ), a validation set ( $10 \%$ ) and a test set ( $10 \%$ ). Next, we use the AlexNet, a deep neural network trained to recognize objects (Krizhevsky et al., 2012), to process these images. We use the pre-trained connection weights from Caffe (Jia et al., 2014) to process each image in our dataset and use the hidden activation of the 7th layer as its embedding. We also reduce the dimensionality of the image embeddings to 30 using PCA.

Model. The model architecture is fairly simple (Figure 2). There are three layers, the input layer, the hidden layer and the output layer. The input layer contains three pools that encode the first three items (A, B and C in the analogy question). Each pool has 30 nodes, which corresponds to the dimensionality of the word/visual embeddings. The connection weights from the input pool encoding $A$ to the hidden layer $H$ and the ones from the pool encoding $C$ to $H$ are the same, denoted by $W_{1}$. The connection weights from the pool encoding $B$ to the hidden layer $H$ are denoted by $W_{2}$. The connection weights from the hidden layer $H$ to the output layer $O$ is the embedding matrix of either the choices in the current example (Task 1) or the whole vocabulary (Task 2). Mathematically, the model can be described by the following equations:

$$
\begin{align*}
H & =W_{1} v_{A}+W_{2} v_{B}+W_{1} v_{C}+b  \tag{1}\\
O & =\phi\left(W_{0} H\right)
\end{align*}
$$

where $v_{A}, v_{B}, v_{C} \in \mathbb{R}^{30}$ are the word/image embeddings for the stimuli, $W_{1}, W_{2} \in \mathbb{R}^{30 \times 30}$ are the connection weights from the input pools to the hidden layer, $b \in \mathbb{R}^{30}$ is the bias in the hidden layer, and $W_{0}=\left[V_{D_{1}} ; V_{D_{2}} ; \ldots ; V_{D_{n}}\right]^{T} \in \mathbb{R}^{30 \times 30}$ is a matrix composed of embeddings of all the choices. We use
the softmax function $\phi(\boldsymbol{x})_{i}=\frac{e^{x_{i}}}{\sum_{j=1}^{n} e^{x_{j}}}$ to normalize the input $\boldsymbol{x}$ to the final layer $O$, which amounts to $W_{0} H$. The $i$ th value of the output, $\phi(\boldsymbol{x})_{i}$, indicates the probability of the $i$ th choice being correct.


Figure 2: Model Architecture with an exemplar question "Boy:Girl::Brother:[Sister|Mom]"

Training. The model is trained by back-propagation using the TensorFlow framework (Abadi et al., 2015). We only update the weights $W_{1}, W_{2}$ and $b$. We train the neural networks with a batch size of 50 for each task. We find that the model cannot learn well in the Task 2 setting, where it needs to pick up the correct $D$ from all the possible words. Therefore, in the following section we only report our results for Task 1. We run 2 simulations. In the first simulation, we examine whether we could reproduce the relational shift phenomenon. To this end, we train the model on a verbal analogy dataset with random association. As the training proceeds, we test it periodically on verbal High Association test set and verbal Low Association test set (within-modality), as well as the visual High Association test set and visual Low Association test set (cross-modality). In the second simulation, we look at the influences of number of choices on within-modality generalization and cross-modality transfer. Particularly, we first train the model to perform verbal analogy tasks with different numbers of choices, then test it on visual analogy tasks. We also conduct another version of the experiment in which the visual analogy tasks are learned first. We run each of the simulations 10 times with different random initialization of parameters, and in each run the model is trained for 41 epochs.

## Results

Simulation 1. We first test if our model reproduces the relational shift observed during child development. Figure 3 shows the accuracy curves of four test sets: verbal High Association Dataset, verbal Low Association Dataset, visual High Association Dataset and visual Low Association Dataset.

First of all, we notice that our model clearly demonstrates the tendency of associative responding in the early stage of learning, since the accuracy on the High Association Dataset grows slowly (solid curves), compared with the accuracies
on the Low Association Dataset. As the training proceeds, the model gradually learns to inhibit associative responses for High Association Dataset (black solid curve).

Second, the test accuracies on verbal sets are consistently higher than those on the visual test sets, which is not surprising given that the model is only trained on verbal stimuli. However, we still find a decent amount of cross-modality transfer. For one thing, the tendency to give associative responses earlier in the training is carried over to the visual modality, even though no visual stimuli has been used to train the network. In addition, as the accuracy on verbal Datasets gradually increases, the model also becomes better at answering questions in visual Low Association Dataset (gray dash-dot curve). However, this improvement is not reliably transferred to the visual High Association Dataset, as the accuracy of this dataset remains near chance-level after prolonged training (gray solid curve).

Third, we find that when the accuracy on verbal Low Association test stops to grow after roughly 3 epochs, the accuracy on the corresponding visual dataset continues to improve until after 11 epochs, which then slowly decreases (gray dash-dot curve). This can be explained by the domain-general and the domain-specific component of analogical reasoning. The model first learns the domain-general component of analogical reasoning from the training in the verbal domain, but later the training becomes detrimental to cross-modality transfer since it continuously shapes the model to be specific to the verbal domain, thus reducing the accuracy in the corresponding visual domain.


Figure 3: The effect of association on accuracy and cross-modality transfer. The dotted line indicates the chance-level.

Simulation 2. To understand the effect of number of choices on accuracy for different data type (train/test) and modality (verbal/visual), we run two linear models. In the first linear model, we look at two sets of data, which are obtained from the experiment "learning verbal analogy first" and the one "learning visual analogy first". In the "learning verbal analogy first" experiment (Figure 4, left), we find that both training and same-modality test accuracy are almost perfect, whereas the cross-modality accuracy is not. The linear model shows that for all of these three conditions, the accuracy decreases as the number of choices increases (train: $\beta=-0.005, t(114)=-2.57, p=0.011$,
same-modality test: $\beta=-0.004, t(114)=-2.12, p=0.036$, different-modality test: $\beta=-0.1, t(114)=-54.63, p<$ .001). However, we also find an interaction between test conditions and choice numbers. Particularly, the influence of choice numbers on different-modality test is much larger than the one on same-modality test, $\beta=0.096, t(114)=37.13$, $p<.001$. In the "learning visual analogy first" experiment (Figure 4, right), we find a similar effect of choice numbers on the different-modality test condition, as well as a similar interaction between test conditions and choice numbers, $\beta=$ $0.071, t(114)=14.07, p<.001$.

Although both modalities demonstrate near perfect performance of within-modality generalization after sufficient training ( $\sim 40$ epochs), the performance of "learning verbal analogy first" is consistently better than the one of "learning visual analogy first" throughout the training. For instance, half way through the training, the same-modality test accuracy of "learning verbal analogy first" is higher than the one of "learning visual analogy first" (mean difference is $4.77 \%, t(234)=3.279, p=0.001$ ). This implies that the semantic space of word embeddings may have a stronger structural regularity, which makes it easier to discover relations between words than images.


Figure 4: The influence of number of choices on accuracy. Solid dots indicate chance-levels.

## Visualization of connection weights

To get a deeper understanding of what the model has learned, we visualize the weight matrix $W_{1}$ and $W_{2}$. We find that $W_{1}$ is very much like a identity matrix (Figure 5, left), whereas $W_{2}$ does not have a easily describable pattern (Figure 5, right).

We compare our model with the vector offset method, which was used by Mikolov, Yih, and Zweig (2013) to solve analogy tasks. Given the problem $\mathrm{A}: \mathrm{B}:: \mathrm{C}: ?$, they found the word $D$ such that its embedding vector had the greatest cosine similarity to $x_{B}-x_{A}+x_{C}$. Their method amounts to assigning an identity matrix $I$ to $W_{1},-I$ to $W_{2}$, and a zero vector to the bias $b$ in our model. The weights of our neural network show that our model is not doing exactly the same thing as the vector offset method does, since $B$ does not approximate the negative identity matrix. Hence, its weights are tuned to solve the current analogy task, and the same weights are also capable of solving analogy task in another modality.


Figure 5: (a): Connections from A to the hidden layer. (b): Connections from B to the hidden layer. Lighter areas represent larger weights.

## Discussion

In this article, we build the first neural network model that can learn to solve analogy tasks and make cross-modality transfer. It uses word representations and image representations estimated from large-scale naturalistic corpora. The model demonstrates both the domain-general and the domain-specific component of analogical reasoning.

Specifically, we see that the accuracy on the same-modality test set is consistently higher than the accuracy on the cross-modality test set. This is aligned with the empirical finding that domain-specific knowledge boosts performance in analogical reasoning (Goswami \& Brown, 1990b). On the other hand, the model demonstrates the domain-general property of analogy-making by showing cross-modality transfer. This is relevant to a broader topic in cognitive science, the zero-shot learning. Zero-shot learning refers to the ability to solve a task despite not having received any training examples of that task. As human beings, we do zero-shot learning all the time. Only recently did researchers begin to simulate zero-shot learning using neural network models. For instance, in Socher, Ganjoo, Manning, and Ng (2013), they showed that learning the distributions of words in texts as a semantic space helps the model understand the visual appearances of objects, and enables the model to recognize objects even if no training data is available for that category. Our model contribute to the zero-shot learning literature by showing that zero-learning is possible for analogy-making task as well. It also makes some interesting predictions that can be empirically tested. The success of our model suggests the possibility that there might be some similar structural regularities in the word embeddings extracted from naturalistic corpus and in the image embeddings extracted from object recognition models. This similarity explains why our model makes cross-modal transfer.

Our results are also relevant to another line of research, the one-shot learning. One-shot learning refers to the problem of learning from one or very few examples. Classic deep learning neural networks could not perform one-shot learning, which is a common criticism of neural networks being plausible cognitive models of human learning (Lake, Ullman, Tenenbaum, \& Gershman, 2016). However, recently Vinyals, Blundell, Lillicrap, Wierstra, et al. (2016) showed
that if you match the training task with the test task, neural networks are able to learn from few examples. In their paper, they trained a network to map a query example to one of the four candidates example so that both of them belong to the same category. The model learned the task very well. Our results lend further support to their approach. We find that our model only learns efficiently under the Task 1 setting, where it chooses among a few choices rather than the whole vocabulary. Our work extends Vinyals and colleagues' results by showing that our network model can make an inference not only on unseen stimuli, but also on unseen stimuli from a completely different modality.

There are some limitations of the current work. First, the model is a simplification of the actual mental processes underlying analogy-making. A lot of previous computational model have given very insightful explanations of those mental processes (Gentner, 1983; Morrison et al., 2004; Kollias \& McClelland, 2013; Gergel' \& Farkaš, 2015), and our goal is not to argue against those models or to provide a better model. Instead, our goal is to demonstrate the possibility that domain-general neural network models can learn from large-scale, realistic datasets to solve analogy tasks. Second, we have not directly compared our model performance with human performance. It would be interesting to see how human would respond to the analogy questions in the current study and whether our model predictions align with human data in the future.

## Acknowledgments

A.Y. is grateful to Jay McClelland for helpful discussions about this work, and the whole Parallel Distributed Processing Lab at Stanford for their useful comments.

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. (Software available from tensorflow.org)
Brian, F. (1990). Analogical interpretation in context. In Proceedings of the 12th Annual Meeting of the Cognitive Science Society.
Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., \& Phelps, M. (1998). Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I Verbal: Analogy section. ETS Research Report Series, 1998(1).
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.
Falkenhainer, B., Forbus, K. D., \& Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial Intelligence, 41(1), 1-63.
Forbus, K. D., Gentner, D., Markman, A. B., \& Ferguson, R. W. (1998). Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. Journal of Experimental \& Theoretical Artificial Intelligence, 10(2), 231-257.

French, R. M. (2002). The computational modeling of analogy-making. Trends in Cognitive Sciences, 6(5), 200-205.
Frost, R., Armstrong, B. C., Siegelman, N., \& Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. Trends in Cognitive Sciences, 19(3), 117-125.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155-170.
Gentner, D., \& Forbus, K. D. (2010). Computational models of analogy. Wiley Interdisciplinary Reviews: Cognitive Science, 2(3), 266-276.
Gentner, D., Holyoak, K. J., \& Kokinov, B. N. (2001). The analogical mind: Perspectives from cognitive science. MIT press.
Gergel', P., \& Farkaš, I. (2015). Connectionist modeling of part-whole analogy learning. In Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science.
Goswami, U. (1991). Analogical reasoning: What develops? a review of research and theory. Child Development, 62(1), 1-22.
Goswami, U., \& Brown, A. L. (1990a). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. Cognition, 36(3), 207-226.
Goswami, U., \& Brown, A. L. (1990b). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. Cognition, 35(1), 69-95.
Holyoak, K. J., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive Science, 13(3), 295-355.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104(3), 427-466.
Hupp, J. M., \& Sloutsky, V. M. (2011). Learning to learn: From within-modality to cross-modality transfer during infancy. Journal of Experimental Child Psychology, 110(3), 408-421.
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093.
Kokinov, B., \& Petrov, A. (2000). Dynamic extension of episode representation in analogy-making in AMBR. In Proceedings of the 22nd Annual Meeting of the Cognitive Science Society.
Kollias, P., \& McClelland, J. (2013). Context, cortex, and associations: a connectionist developmental approach to verbal analogies. Frontiers in Psychology, 4, 857.
Krizhevsky, A., Sutskever, I., \& Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1097-1105).
Kuehne, S., Forbus, K., Gentner, D., \& Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. In Proceedings of the 22nd Annual Meeting of the Cognitive Science Society.

Lake, B. M., Salakhutdinov, R., \& Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science, 350(6266), 1332-1338.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., \& Gershman, S. J. (2016). Building machines that learn and think like people. arXiv preprint arXiv:1604.00289.
Lovett, A., Forbus, K., \& Usher, J. (2007). Analogy with qualitative spatial representations can simulate solving Raven's Progressive Matrices. In Proceedings of the 29th Annual Meeting of the Cognitive Science Society.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., \& Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (pp. 3111-3119).
Mikolov, T., Yih, S. W.-t., \& Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013). Association for Computational Linguistics.
Mitchell, M. (1993). Analogy-making as perception: A computer model. MIT Press.
Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., \& Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. Journal of Cognitive Neuroscience, 16(2), 260-271.
Reed, S. E., Zhang, Y., Zhang, Y., \& Lee, H. (2015). Deep visual analogy-making. In Advances in Neural Information Processing Systems (pp. 1252-1260).
Richland, L. E., \& Burchinal, M. R. (2013). Early executive function predicts reasoning development. Psychological Science, 24(1), 87-92.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8 -month-old infants. Science, 274(5294), 1926-1928.
Socher, R., Ganjoo, M., Manning, C. D., \& Ng, A. (2013). Zero-shot learning through cross-modal transfer. In Advances in Neural Information Processing Systems (pp. 935-943).
Sternberg, R. J., \& Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. Child Development, 27-38.
Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In Advances in Neural Information Processing Systems (pp. 3630-3638).
Yamins, D. L., Hong, H., Cadieu, C., \& DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In Advances in Neural Information Processing Systems (pp. 3093-3101).

# Convincing Conversations: Using a Computer-Based Dialogue System to Promote a Plant-Based Diet 

Emma L. Zaal (E.L.Zaal@rug.nl)<br>Center for Language and Cognition Groningen, Oude Kijk In't Jatstraat 26<br>Groningen, 9712 EK, The Netherlands<br>Gregory J. Mills (G.J.Mills@rug.nl)<br>Center for Language and Cognition Groningen, Oude Kijk In't Jatstraat 26<br>Groningen, 9712 EK, The Netherlands<br>Afke Hagen (A.Hagen.1@student.rug.nl)<br>Center for Language and Cognition Groningen, Oude Kijk In't Jatstraat 26<br>Groningen, 9712 EK, The Netherlands<br>Carlijn A. Huisman (C.A.Huisman@student.rug.nl)<br>Center for Language and Cognition Groningen, Oude Kijk In't Jatstraat 26<br>Groningen, 9712 EK, The Netherlands<br>John C.J. Hoeks (J.C.J.Hoeks@rug.nl)<br>Center for Language and Cognition Groningen, Oude Kijk In't Jatstraat 26<br>Groningen, 9712 EK, The Netherlands


#### Abstract

In this study, we tested the effectiveness of a computer-based persuasive dialogue system designed to promote a plant-based diet. The production and consumption of meat and dairy has been shown to be a major cause of climate change and a threat to public health, bio-diversity, animal rights and human rights. A system promoting plant-based diets was developed, comprising conversational, motivational and argumentational elements. 280 participants were randomly assigned to one of four conditions, each representing a particular combination of motivational and argumentational modules. Male participants showed higher intention scores in the motivational conditions compared to the argumentation-only or control condition. Female participants scored higher overall, unaffected by condition. These results suggest that men and women are differentially sensitive to persuasive strategies regarding the adoption of a plant-based diet. It seems to be particularly worthwhile to use motivational - as opposed to merely argumentational - elements in a persuasive conversation.


Keywords: human-machine interaction; dialogue system; persuasive communication; cognitive dissonance; motivational interviewing.

## Introduction

When individuals experience a mismatch between their beliefs and their actual behavior, the phenomenon of cognitive dissonance (Festinger, 1957) kicks in: "an aversive state which motivates cognitive or behavioural actions to lower itself" (Dijkstra, 2009, p.792). For instance, most people do have a desire to behave ethically, but this desire is often not reflected in their actual behavior. People are often 'wilfully ignorant' when it comes to their ethical beliefs, actively ignoring or discarding relevant information
about those beliefs (Zane, Irwin \& Reczek, 2013).There are two ways in which you can deal with and/or solve cognitive dissonance. The first one is to change your cognitions in line with your current behavior. For example, you could change your belief: "behaving ethically is not so important for me (anymore)." The second way is to change your behavior in line with their current beliefs: actually acting in an ethical manner (Dijkstra, 2009; Hewstone, Jonas \& Stroebe; 2012).

The latter way of solving cognitive dissonance is often an ultimate goal of - for instance - health and sustainability communication professionals: they want people to change their behaviour, like: exercise more often (Riet et al., 2010); quit smoking (Ballast \& Dijkstra, 2011); reduce fat-intake (Wright, Velicer \& Prochaska, 2008); behave environmentally friendly (Bolderdijk et al., 2012) etc. However, people often change their cognitions instead of their actual behavior. Therefore, persuaders often start by trying to change people's cognitions. In order to achieve this, one has to address the most focal cognitions that people hold about a particular behaviour (Fishbein \& Yzer, 2003).

One problem area that is often associated with cognitive dissonance, is the consumption of meat and other animal products. Meat eaters are assumed to experience cognitive dissonance resulting for instance from the perceived aversive consequences of their diet ("I eat meat but eating meat hurts animals") (Cooper \& Fazio, 1984); or because it threatens their self-integrity ("ethical people don't eat meat, but I do") (Aronson, 1968). Most people believe it is wrong to hurt animals, while at the same time around $95 \%$ of the consumers in the U.S. eat meat. Researchers have referred to this phenomenon as the 'meat-paradox', and claim that
cognitive dissonance lies at the heart of this phenomenon (Bastian, Loughnan, Haslam \& Radke 2012; Rothgerber, 2014).

Besides solving the unpleasant feeling of cognitive dissonance that consumers may experience due to eating meat, there are several other important reasons to address meat consumption. First, the livestock industry is one of the largest polluters worldwide. In order to meet stringent climate change targets, the consumption of animal products should be reduced at least by half (Donham et al., 2007; Hedenus, Wirsenius \& Johansson, 2014; Hertwich et. al., 2010; Steinfeld et al., 2006). Second, the production of animal products is threatening rights and wellbeing of both humans and animals due to poor working conditions for the first, and poor living conditions for the latter (Pew Commission, 2008). Third, individual and public health is negatively affected by meat and dairy consumption. For instance, the growing consumption animal products is associated with an increase of diseases like obesity, type II diabetes (Cooney, 2014; Montonen et al., 2013), heart and vascular diseases (Yokoyama et al., (2014) and a variety of cancers, like colorectal, lung and bladder cancer (Lippi, Mattiuzzi \& Cervellin, 2016). Thus, a successful promotion of plant-based diets - diets without meat, dairy and eggs could have a great many positive outcomes for animals, humans, and the environment.

It has been claimed that one of the most effective strategies to reach persuasion is through conversation and interaction (Helme et al., 2011; Noar, Carlyle \& Cole, 2006). However, even if that were the case, face-to-face conversations are obviously very time consuming and costly (Southwell \& Yzer, 2007). A fruitful solution for this could be the deployment of human-machine interaction systems. There is a huge benefit in using online dialogue systems as larger target groups can be reached at lower costs. In addition, automated dialogues can also easily be tailored, which lowers resistance to persuasion and makes messages individually more relevant (Dijkstra, 2008). An added benefit for experimental research is that manipulations can be held more constant, making it easier to measure the effect(s) of the strategy/strategies employed within the dialogue system separately - and/or combined.

## The Dialogue System

We devised and tested a persuasive online dialogue system that promotes plant-based diets. Our system incorporates persuasive strategies aimed at reducing the consumption of animal products. First, an 'argumentation' module was designed to target cognitive dissonance by addressing focal beliefs about meat consumption. This module provides individually tailored arguments that address the individual's so-called 'disengagement' beliefs: beliefs that may be true in themselves, but that are not valid arguments in the discussion at hand (e.g., "Our ancestors ate meat" is true, but it is not a valid reason why a present-day individual should still be eating meat).

In addition, a second module was designed to reduce the likelihood of experiencing negative affect and resistance to counter arguments and persuasion. In building this module, we borrowed heavily from the theoretical framework of Motivational Interviewing (MI). MI is "a collaborative conversation style for strengthening a person's own motivation and commitment to change" (Miller \& Rollnick, 2002). In this method, the receiver formulates its own goals, capacities and reasons regarding his/her behavior change towards the targeted behavior. Dialogue in MI is framed in such a way to emphasize one's autonomy and to avoid any direct confrontation with the target individual. Important elements of MI are: 1) an explicit consent question, asking whether an individual agrees with talking about a specific topic; 2) a 'motivation ruler', which consists of asking one to number or 'grade' their motivation to change and subsequent questioning and giving feedback on the number they choose (e.g., "You chose " 2 ", what would it take for you to get a higher level of motivation?"); and 3) a 'confidence ruler', which asks about a person's confidence in his/her own capacity to change towards the target behavior. Digital applications of this conversational method have proved to be effective in achieving positive intentional and behavioral outcomes (Shingleton \& Palfai, 2015).

To increase the 'feel' of an actual conversation, we include a picture (of a young female) to visualize the 'person' talking to the participant. We use personal pronouns like " I " and "you" in conversation; and talk is individually tailored by the system throughout the conversation based on responses of the - non-digital conversational partner. In the future, we would like to develop a more sophisticated dialogue system, capable of reacting to natural language input. Previous research has shown that technological social agents - like robots - are able to induce behavior change by providing interactive feedback with regard to for instance sustainability-related behavior like energy conservation (Ham \& Midden, 2014). However - to our knowledge - no research has yet looked into computerized agents that are active in the field of plantbased eating. We used survey-builder Qualtrics to design a straightforward, tree-based conversational system with feedback based on answers on - for the most part - multiple choice questions.

Table 1: Example of tree-based conversation with Eliza

| "When are you planning to make changes to your diet?" |  |  |
| :---: | :--- | :--- |
| Participant's answer | Eliza's answer |  |
| $\bullet$ | Within a week | "That's pretty soon, good to |
| • | Within a month | hear! Can I ask you more?" |
| $\bullet$ | Within 3 months | "You're taking your time, |
| • | Within a year | but that's OK! Are there any <br> ways to potentially speed up <br> this process for you?" |

## Method

## Participants

Three-hundred-and-seventy-one Dutch participants took part in the research. Participants were recruited from social media, mailing lists and the researchers' personal networks. Data from participants who had a vegetarian or vegan diet or failed to complete the questionnaire - were excluded from analysis. Two-hundred-and-eighty participants remained. Of this sample, $76 \%$ was female $(\mathrm{N}=212)$ and $24 \%$ male ( $\mathrm{N}=68$ ). Participants' mean age was 26 years ( $\mathrm{SD}=9.9$ ), ranging from 17 to 65 . Male and female participants were dispersed evenly across four conditions (15-20 men and 5560 women in each condition).

## Measures

Disengagement Belief Strength The strength of the various disengagement beliefs was measured by asking participants about the extent they (1) totally disapprove - (7) totally approve of fourteen disengagement beliefs about eating meat (i.e. "Without meat you cannot be healthy"; "Lions eat meat too"). These fourteen disengagement beliefs were formulated based on a belief elicitation study in which twenty-three participants stated the most important reasons for them to keep eating meat. In the present study, participants were asked which three of these fourteen beliefs were the most important reasons for them to keep eating meat. Subsequently, they received tailored feedback based on the answers they provided. This feedback was framed as a short text in a what-if question format (i.e. "You state that meat is too tasty. That is hard to deny! But what if you find out that some meat substitutes are quite good and sometimes even as good as real meat. Would you then be open to change your current diet towards a more plant-based diet?"). When a participant answered "no", the dialogue system would give similar 'what-if' feedback for the second and/or third reason, until the participant either answered "yes" or all three reasons were addressed. Note that all responses of the dialogue system consisted of valid arguments, based on scientific findings.

MI module In the full MI version, participants were explicitly asked for their approval to talk about their dietary habits ("In this program, I will talk about the advantages of having a plant-based diet [...] Are you open to talk about your own eating habits and possible changes to it?"). When participants did not consent, there was a short feedback page after which the experiment ended. As we described above, motivation to change towards a plant-based diet was measured by asking one's motivation on scales ranging from (0) No motivation at all - (10) very highly motivated. (e.g. "It looks like reducing your meat consumption is not that high of a priority for you! Could you tell me why?") A similar procedure was used for measuring how much participants trusted their own capacity to perform the requested behavior (self-efficacy).

Attitude and Behavioral Intention Attitude was measured by evaluating the following statement on a semantic differential scale: "If I would change my current diet to a vegan diet, that would be...": i.e. good-bad; foolish-wise; unnecessary-necessary. In addition, participants evaluated several statements on a seven-point scale ranging from (1) strongly disapprove - (7) strongly approve (i.e. "A diet without animal product is more environmentally friendly."). Behavioral intention towards three kinds of behavior was measured: going completely or partially vegan; going completely or partially vegetarian; going completely or partially organic where meat was concerned. These intention measures were weighted and summed into one intention-score that indicated their willingness to change their diet towards a more - or less - plant-based diet (the higher, the more willing).

Evaluation questions At the end of the session, we asked 1) whether participants felt they were addressed in a nice manner, and 2) whether they liked to communicate with Eliza by means of evaluating statements on a Likert-type scale: (1) completely disagree - (7) completely agree.

## Procedure and Design

When starting the dialogue system, participants saw a picture of a young girl who was introduced as "Eliza". She asked if they cared to join her in a conversation about their eating habits. Subsequently, participants were asked about their gender, age, education and actual eating habits. Next, participants received information about several benefits of a plant-based diet. In three of the four conditions the disengagement belief handling module ('DBH') was included to address the participant's most focal beliefs about eating meat. Two of the four versions also included a motivational interviewing module ('MI'), either a full (with an explicit consent question) or partial (without an explicit consent question) module. A fourth version did not include DBH and MI modules and served as - baseline - control condition. Table 2 shows how the modules were combined in the different versions of the system that were used in this study.

Table 2: Experimental conditions

|  | Full MI | Partial <br> MI | DBH | Control |
| :--- | :---: | :---: | :---: | :---: |
| Consent | + | - | - | - |
| MI | + | + | - | - |
| DBH | + | + | + | - |
| INFO | + | + | + | + |

Note. MI = Motivational Interviewing; DBH = Disengagement Belief Handling; Consent $=$ consent question; $\mathrm{INFO}=$ information on plant-based diet.

## Analysis

We used between-groups univariate Analysis of Variance to investigate the patterns of results for Attitude, Intention and Evaluation. Condition was one factor, with four levels (Full MI, Partial MI, DBH and Control), and Sex of Participant the other factor (male versus female). This latter factor has consistently been shown to influence behavioral outcomes concerning reduction of meat consumption (e.g. Cooney, 2014). In the analysis on Evaluation there were only three levels of Condition, as the Control condition did not feature a conversation with Eliza.

## Results

Table 3 shows the means for attitude towards adopting a plant-based diet, and the composite scores for intention to reduce meat consumption per condition.

Table 3: Means and Composite Scores (plus standard errors) of intention and attitude

|  | Intention |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| Attitude |  |  |  |  |
| Condition | Men | Women | Men | Women |
| Full MI | 95.45 | 99.67 | 5.56 | 5.62 |
|  | $(7.95)$ | $(5.24)$ | $(.21)$ | $(.14)$ |
| Partial MI | 95.67 | 106.80 | 5.21 | 5.56 |
|  | $(10.72)$ | $(5.24)$ | $(.28)$ | $(.14)$ |
| DBH | 80.65 | 101.72 | 4.74 | 5.45 |
|  | $(8.89)$ | $(4.80)$ | $(.23)$ | $(.12)$ |
| Control | 65.95 | 102.23 | 5.11 | 5.77 |
|  | $(7.76)$ | $(4.41)$ | $(.20)$ | $(.11)$ |

Note. MI = Motivational Interviewing; DBH = Disengagement Belief Handling

## Effects of Condition and Sex on Attitude

A Univariate Analysis of Variance showed a significant main effect of Condition on Attitude towards adopting a more plant-based diet $(\mathrm{F}(3,269)=2.67, \mathrm{p}=.048)$. Attitude scores of participants in the Control ( $\mathrm{M}=5.44, \mathrm{SE}=.12$ ) and Full MI ( $\mathrm{M}=5.59, \mathrm{SE}=.12$ ) condition were significantly higher than Attitude scores of participants in the DBH condition ( $\mathrm{M}=5.10$, $\mathrm{SE}=.13$ ). In addition, there was a main effect of Sex on Attitude $(\mathrm{F}(1.269)=11.54, \mathrm{p}=.001)$. Women ( $\mathrm{M}=5.60$, $\mathrm{SE}=.06$ ) scored significantly higher on attitude measures than men ( $\mathrm{M}=5.16, \mathrm{SE}=.06$ ) regardless of exposure to experimental condition. No interaction effect of Condition x Sex was found (p-value>.23).

## Effects of Condition and Sex on Intention

A Univariate Analysis of Variance showed a marginally significant main effect of Condition on Intention to adopt a more plant-based diet $\mathrm{F}(3,272)=2.29, \mathrm{p}=.078$. In addition, there was a main effect of Sex on Intention. Men ( $\mathrm{M}=84.44$, $\mathrm{SE}=4.56$ ) scored significantly lower on intention to adopt a more plant-based diet than women ( $\mathrm{M}=102.61, \mathrm{SE}=2.47$ ), $\mathrm{F}(1,272)=12.74, \mathrm{p}=.000$. The interaction of Condition x

Sex was found to be marginally significant $(\mathrm{F}(3,272)=2.2$, $\mathrm{p}=.09$ ).

Further exploration of these effects showed that intention scores were only significant between conditions for male participants $(\mathrm{F}(3,64)=2.85, \mathrm{p}=.044)$ but not for female participants $(\mathrm{F}(3,208)=.332, \mathrm{p}=.80)$. A post-hoc test showed that intention scores of men who were exposed to the Control condition ( $\mathrm{M}=66.0$, $\mathrm{SE}=7.8$ ), were significantly lower than mean intention scores of men who were exposed to either the full MI condition ( $\mathrm{M}=95.5, \mathrm{SE}=8.0$ ), $\mathrm{p}=.011$; or the partial MI condition ( $\mathrm{M}=95.67, \mathrm{SE}=10.8$ ), $\mathrm{p}=.030$. Intention scores between men in the DBH condition and the control condition did not differ.

## Evaluation questions

Table 4 shows the mean scores about the participants' appreciation for the way they were addressed by Eliza and the extent they liked talking to her.

Table 4: Means (plus SE) of appreciation and liking

|  | Appreciation |  | Liking |  |
| :--- | :--- | :--- | :--- | :--- |
| Condition | Men | Women | Men | Women |
| Full MI | 5.26 | 5.33 | 4.42 | 4.04 |
|  | $(.32)$ | $(.21)$ | $(.32)$ | $(.20)$ |
| Partial MI | 4.91 | 5.23 | 4.46 | 4.18 |
|  | $(.43)$ | $(.22)$ | $(.42)$ | $(.21)$ |
| DBH | 5.53 | 5.02 | 4.73 | 4.06 |
|  | $(.36)$ | $(.19)$ | $(.36)$ | $(.19)$ |

Note. MI = Motivational Interviewing; DBH = Disengagement Belief Handling

In general, participants felt they were addressed in a nice manner: mean scores ranged from 4.9 to 5.5 (maximum 7). A $3 \times 2$ ANOVA was carried out - no evaluation questions about Eliza were asked in the control condition - which showed no significant main or interaction effects (p-values > .37). Participants also seemed to like to communicate with Eliza, though the marginally significant main effect of Sex of Participant $(\mathrm{F}(1,182)=3.39, \mathrm{p}=.067)$ suggested that male participants enjoyed talking to Eliza a little bit more ( $\mathrm{M}=4.54$, $\mathrm{SE}=.21$ ) than female participants $\quad(\mathrm{M}=4.09$, $\mathrm{SE}=.12$ ). No other effects were significant (p-values > .77).

## Discussion

The aim of this research was to find out 1) whether it would be effective to design a dialogue system to promote plantbased diets, and 2) which elements or modules contribute to the persuasive power of the system. To this end, four versions of a computer-based dialogue system were developed, which 'talked' about the benefits of plant-based diets and provided - individually tailored - conversations based on different persuasive strategies.

Most importantly, we saw positive effects of incorporating persuasive strategies in a dialogue system. This implicates the usefulness of using a dialogue system to
promote sustainable behavior concerning the promotion of adopting a (more) plant-based diet.

We discovered effects of the different versions on behavioral intention towards adopting a plant-based diet. These effects were only present for men. Women were in general more willing to change their diet, regardless of the condition they were assigned to. So contrary to expectations, men seemed to be most sensitive to persuasion in the area of moving towards a plant-based diet. Men who were exposed to the Full MI and the Partial MI conditions which both contained a motivational interviewing module showed higher intentions to adopt a more plant-based diet than men in the control condition. The DBH condition that only contained disengagement belief handling did not significantly differ from the control condition.

Apparently, if we restrict our persuasive attempt to mere argumentation, we will not get far, at least not with our male conversational partners. This conclusion is underscored by the findings regarding participants' attitudes. Incorporating an argumentational component even seems to negatively affect those: attitude scores were significantly lower in the argumentation-only ( DBH ) condition than in the Control condition for both men and women. Because a DBH module is included in all experimental conditions, it is still possible that DBH is effective, but only in combination with a motivational module. Future research including an 'only MI' condition could perhaps answer that question.

A second important finding concerns the effects of Sex of the Participant. Our results support the notion that men and women think differently about adopting a more plant-based diet and that they are persuaded by different means (e.g. Cooney, 2014). However, as we said, the outcome was not quite as we anticipated, as we expected that women would be more susceptible to persuasion in this area. What our results do suggest is that women are generally more likely to have or adopt a plant-based diet. While it is true that in this research only men were positively affected by two versions of the intervention, women showed more positive attitudes and higher intentions to adopt a more plant-based diet than men overall.

Of specific interest is that there was also a (trend towards a) main effect of Sex on the scores of one of the evaluation questions, when participants were asked whether they had enjoyed talking to Eliza. Men had enjoyed the conversation with Eliza more than women. Perhaps if we would use a picture of a young male, this 'Elisus' would work better with a majority of female participants. We will tackle this issue in a future version of the present experiment.

Future studies in our lab will focus on a number of issues. First of all, we will perform a replication of this study in a different participant group to gauge the extent to which we can generalize the specific findings of this study

Second, we measured intention immediately after participants were exposed to the intervention. It is very likely the case that interventions need much more time before they have a detectable effect. Especially with difficult behaviors like changing dietary habits, the
occurrence of what is sometimes called 'sleeper effects' seems very plausible (Kumkale \& Albarracín, 2004).

In addition, as in most studies, the present research measured behavioral intention and not actual behavior. While intention is thought to be a potent predictor of actual behavior (Fishbein \& Ajzen, 2010), being able to measure and predict actual behavioral outcomes is the ultimate goal of persuasion research. We would like to use some form of longitudinal design, where we measure - self-reported eating behavior over a longer period.

Finally, the present research measured the effects of a single intervention. In the real world, people are often exposed more than once to the same, or related persuasive information - they read advertisements in newspapers and magazines, see posters, look at commercials etc. Perhaps we need multiple exposures to create more persuasive results.

In conclusion, then, our research suggests that a dialogue system can induce behavior change in the field of a plantbased life-style. However, strategy-wise, only giving people arguments in trying to persuade them is not going to work. People, especially men, may want to feel their own autonomous motivation, which can be fueled by a supporting dialogue based on - for instance - elements from motivational interviewing. Future research should test designs that incorporate different combinations of persuasive strategies (and pictures/avatars); multiple exposures to/conversations with the system; over a longer time span; and should measure actual consumption behavior instead of intention as outcome.

## References

Aronson, E. (1968) Dissonance theory: Progress and problems. Ableson, R.P. (1968) Theories of cognitive consistency: A sourcebook. Rand Mcnally, Chicago, 5-27
Ballast, K., \& Dijkstra, A. (2011) Personalization and perceived personal relevance in computer-tailored persuasion in smoking cessation. British Journal of Health Psychology, 17 (1), 60-73
Bastian, B., Loughnan, S., Haslam, N., \& Radke, H. (2012) Don't mind meat? The denial of mind to animals used for animal consumption. Personality and Social psychology Bulletin, 38, 247-256.
Bolderdijk, J.W., Steg, L., Geller, E.S., Lehmand, P.K., Postmes, T. (2013) Comparing the effectiveness of monetary versus moral motives in environmental campaigning. Nature Climate Change, 3, 413-416
Cooney, N. (2014) Veganomics. The Surprising Science on What Motivates Vegetarians from the Breakfast Table to the Bedroom. Brooklyn: Lantern Books
Cooper, J., Fazio, R.H. (1984) A new look at dissonance theory. Advances in Experimental Social Psychology, 17, 229-266
Dijkstra (2009) Disengagement beliefs in smokers: Do they influence the effects of a tailored persuasive message advocating smoking cessation? Psychology \& Health, 24 (7), 791-804

Dijkstra, A. (2008). The psychology of tailoring-ingredients of computer-tailored persuasion. Personality and Social Psychology Compass, 2, 765-784.
Donham, K. J., Wing, S., Osterberg, D., Flora, J. L., Hodne, C., Thu, K. M., \& Thorne, P. S. (2007). Community Health and Socioeconomic Issues Surrounding Concentrated Animal Feeding Operations. Environmental Health Perspectives, 115(2), 317-320.
Festinger, L. (1957) A theory of cognitive dissonance, Vol 2. Stanford University press, Palo Alto, CA
Fishbein, M. \& Ajzen, I. (2010) Predicting and changing behavior: the reasoned action approach. New York: Psychology Press
Fishbein, M. \& Yzer, M. (2003) Using Theory to Design Effective Health Behavior Interventions. Communication Theory 13 (2), 164-183
Ham, J., \& Midden, C. H. (2014). A persuasive robot to stimulate energy conservation: The influence of positive and negative social feedback and task similarity on energy-consumption behavior. International Journal Of Social Robotics, 6(2), 163-171. doi:10.1007/s12369-013-0205-z
Hedenus, F., Wirsenius, S., \& Johansson, D. (2014). The importance of reduced meat and dairy consumption for meeting stringent climate change targets. Climatic Change, 124(1/2), 79-91. doi:10.1007/s10584-014-11045.

Helme, D.W. \& Noar, S.M. \& Allard, S. \& Zimmerman, R.S. \& Palmgreen, P. \& McClanahan, K.J. (2011) InDepth Investigation of Interpersonal Discussions in Response to a Safer-Sex Mass Media Campaign. Health Communication, 26, 366-378
Hertwich, E., van der Voet, E., Suh, S., Tukker, A., Huijbregts M., Kazmierczyk, P., Lenzen, M., McNeely, J. \& Moriguchi, Y. (2010). Assessing the Environmental Impacts of Consumption and Production: Priority Products and Materials, A Report of the Working Group on the Environmental Impacts of Products and Materials to the International Panel for Sustainable Resource Management. UNEP
Hewstone, M., Stroebe, W. \& Jonas, K. (2012) An Introduction to social psychology. Glasgow: The British Psychological Society and John Wiley \& Sons Ltd.
Kumkale, G. T., \& Albarracín, D. (2004). The Sleeper Effect in Persuasion: A Meta-Analytic Review. Psychological Bulletin, 130(1), 143-172.
Lippi, G., Mattiuzzi, C., \& Cervellin, G. (2016). Meat consumption and cancer risk: a critical review of published meta-analyses. Critical Reviews In Oncology/Hematology,971-14.
Miller, W. R. \& Rollnick, S. (2002) Motivational interviewing, preparing people for change (second edition). New York: The Guilford Press
Montonen, J., Boeing, H., Fritsche, A., Schleicher, E., Joost, H., Schulze, M., Steffen, A. \& Pischon, T. (2013). Consumption of red meat and whole-grain bread in
relation to biomarkers of obesity, inflammation, glucose metabolism and oxidative stress.
Noar, S.M. \& Carlyle, K. \& Cole, C. (2006) Why communication is crucial: meta-analysis of the relationship between safer sexual communication and condom use. Journal of health communication, 11(4), 365-90
Pew Commission on Industrial Farm Animal Production. (2008). Putting meat on the table: Industrial farm animal production in America. Washington, DC.
Riet, van 't, J.P., Ruiter, R.A.C., Werrij, M.Q., Vries, de, H. (2010) Investigating message-framing effects in the context of a tailored intervention promoting physical activity. Health education research, 25(2), 343-354
Rothgerber, H. (2014) Efforts to overcome vegetarianinduced dissonance among meat eaters. Appetite, 79, 3241
Shingleton, R.M. \& Palfai, T.P. (2015) Technologydelivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research, Patient Education and Counseling (99), pp. 17-35
Southwell, B. G., \& Yzer, M. C. (2007). The Roles of Interpersonal Communication in Mass Media Campaigns. Communication Yearbook, 31420-462.
Steinfeld, H., Gerber, P., Wassenaar, T., Castel, V., Rosales, M. \& de Haan, C. (2006). Livestock's long shadow: environmental issues and options. Food and Agriculture Organization of the United Nations
Wright, J.A., Velicer, W.F., Prochaska, J.O. (2008) Testing the Predictive Power of the Transtheoretical Model of Behavior Change Applied to Dietary Fat Intake. Health Education Research, 24 (2), 224-23.
Yokoyama, Y., Nishimura, K., Barnard, N. D., Takegami, M., Watanabe, M., Sekikawa, A., \& ... Miyamoto, Y. (2014). Vegetarian Diets and Blood Pressure: A Metaanalysis. JAMA Internal Medicine, 174(4),577-587. doi:10.1001/jamainternmed.2013.14547.
Zane, D., Irwin, J. \& Reczek, R. (2015) Do less ethical consumers denigrate more ethical consumers? The effect of wilful ignorance on judgments of others. Journal of Consumer Psychology, 26, 337-349.

# Pseudoneglect and development: Age-related spatial bias in bisection and drawing 

Yordanka Zafirova, Asenia Giagtzidou, Dara Vassileva, Elena Andonova<br>(jovasileva@abv.bg, nia.giagtz@gmail.com, daravassileva@gmail.com, eandonova@nbu.bg)<br>Department of Cognitive Science and Psychology, 21 Montevideo Street<br>Sofia, 1618, Bulgaria


#### Abstract

The numerous studies on pseudoneglect have generated inconsistent results and disagreement concerning the underlying mechanisms. Most research supports the hypothesis that hemispheric lateralization is the main reason for the persistent leftward bias in spatial tasks. Findings on the influence of reading direction, handedness and participant age are largely contradictory. As a result of brain maturation adults usually perform with significant leftward bias. However, both hemispheric activation and scanning habits exert an influence on space representation, which varies across age groups. Preschoolers, middle school children and adults were tested on the line and word bisection tasks and on house-person-tree drawing tasks. The analysis of their performance produced results consistent with an explanatory account that the direction of the spatial bias shifts leftwards in the course of development.


Keywords: pseudoneglect, line bisection, age differences, house-person-tree drawing task

## Introduction

Cognitive processes have their limitations that may lead to distortion in perception or judgment. Different biases arise when cognitive resources are challenged, but are also partially rooted in the cultural context and can be learned implicitly, including our ability to navigate through space and construct adequate spatial representations of the close environment. The temporal and spatial structure of the viewing behavior is independent from the goal of the task and is attributed to the horizontal asymmetries of the visual and attentional systems.

The hemispatial neglect syndrome is a neuropsychological disorder where patients have difficulties processing stimuli from the contralesional hemispace. The condition is generally due to impairment of the ability to direct attention and movement, although it might as well be attributed to the inability to form spatial representations (Bisiach, 1996). Patients with left visuospatial neglect bisect horizontally presented lines to the right of their objective center. In contrast, neurologically normal people usually perform the task with significant leftward bias. This exploratory inclination to look slightly to the left of the center of the presented stimuli is known in the literature as pseudoneglect (Bowers \& Heilman, 1980).

Pseudoneglect is typically interpreted in terms of the hemispheric lateralization of the brain and has two complementary components - visual and motor. The visual component incorporates the scanning habits and the attentional and perceptual effects, while the motor
component refers to the overdriven movements, like directional hypometria, perceptual motor activation and cueing (Macdonald-Nethercott, Kinnear, \& Venneri, 2000). In one of the first studies on pseudoneglect leftward errors were made only in the paper-and-pencil version of the line bisection task and not in the computer version of the task (Luh, 1995). This result supports the contribution of the above mentioned motor factor to the leftward spatial bias. Both hemispheres are engaged in directing attention to the contralateral space, but as the right hemisphere is generally more active during spatial tasks this could lead to subsequent enhanced attendance to the left visual hemispace. This was demonstrated in a cancellation task designed to deliberately activate the left or the right hemisphere (Vingiano, 1991). A more recent fMRI study also showed increased activation in the right intra-parietal sulcus and lateral peristriate cortex during judgment and performance of line bisection tasks (Cicek, Deouell, \& Knight, 2009).

In a series of eye-tracking experiments on viewing behavior in exploring complex scenes participants demonstrated a marked initial leftward bias that was independent of the category of the presented images (Ossandon, Onat, \& Konig, 2014). It is generally accepted that people tend to scan the visual field in the direction that they read. Opposite reading habits give rise to opposite spatial bias when performing the line bisection task - left-to-right readers bisect the lines to the left of their veridical center, while right-to-left readers deviate to the right (Chokron \& Imbert, 1993). In a similar study the line bisection performance of adults, 8-year-olds and preschoolers coming from cultures with opposite reading habits was compared and differences were found in all groups (Chokron \& De Agostini, 1995). Usually the developmental shift in the observed bias in line bisection is from right to left, as demonstrated in an experiment with 4-5 and 10-12 years old children (Dellatolas, Coutin, \& De Agostini, 1996).

There is evidence that with the maturation of the corpus callosum during puberty, spatial processing shifts from the contralateral to the right hemisphere. Right-handed prepuberty children showed rightward bias in line bisection tasks when using their right hand and a leftward bias with their left hand. The adult and puberty groups bisected the lines to the left with both hands (Hausmann, Waldie, \& Corballis, 2003). Another study also showed that learned directionality and movement preferences have little influence on placement of single object on a page. Both

French and Moroccan dextral children manifested similar leftward bias in a draw-a-tree task, supporting the advantage of brain lateralization over learned cultural habits, including reading direction (Picard \& Zarhbouch, 2014).

Both neglect and pseudoneglect, despite the difference in the underlying neurological mechanisms, have similar behavioral manifestation and are assessed with the same tasks, such as line and word bisection, drawing or bisecting a simple figure and exploration of complex scenes. Basically, both lines and words are bisected to the left of their objective center by adults, in accordance with the hemispheric activation hypothesis and the reading direction hypothesis. However, the effect is more pronounced in word bisection showing that an additional mechanism might be involved. In a series of experiments with lines and orthographic strings (words, pseudowords, symbols), lines were bisected always to the left, while words yielded a different result depending on their length - short ones (3-4 letters) were bisected with a general right-side bias (Arduino, Previtali, \& Girelli, 2010).

As the beginning of the word is more informative than its end, oculomotor behavior shows that attention is usually directed to the left of the word center, while trying to access the mental lexicon in establishing a matching cohort. According to the Attentional scaling hypothesis, proposed by Fischer (1996), orthographic strings are processed differently than other symbolic or pictorial material. Word bisection is biased toward the beginning of the horizontally presented words, depends on their length and the ease of lexical access - Hebrew-American bilinguals showed a greater leftward bias in their second language (English) than native English readers (Fischer, 1996). Both English and Hebrew readers showed greater leftward bias for words and pseudowords than for lines (greatest for low-frequency words), and people with developmental dyslexia deviated more to the left than controls (Gabay, Gabay, Henik, Schiff, \& Behrmann, 2015). Interestingly, dyslexic children have been reported to show inversed pseudoneglect in line bisection tasks, shifting their subjective center to the right of the veridical one (Michel, Bidot, Bonnetblanc, \& Quercia, 2011).

There are numerous studies on pseudoneglect with inconsistent results as well as disagreement concerning the underlying mechanisms. The goal of the current study was to combine explicit and implicit measures to dissociate the attentional component in spatial processing. We tested three age groups - preschoolers, sixth-graders and adults on line and word bisection tasks and draw-a-house-person-tree tasks. In accordance with previous studies, we used paper-and-pencil tasks and only long and low-frequency words.

As different spatial tasks supposedly tap into different aspects of spatial awareness, we compared performance on active (implicit) and passive (explicit) spatial tasks across three developmental stages with different reading expertise (see Barrett, Kim, Crucian, \& Heilman, 2002 for a discussion on the difference of implicit and explicit tasks). Both the hemispheric activation and the directionality
hypotheses predict that right-handed preschool children would exhibit a right spatial bias in all tasks, because of the prevalence of the motor component in spatial judgments and their inability to read. Sixth-graders should have mixed results due to already established reading habits but incomplete brain lateralization. Adults were expected to have a strong leftward bias, especially in the word bisection tasks, in accordance with the Attentional scaling hypothesis.

## Method

In order to address the research questions above, we carried out an experimental study protocol consisting of a series of five tasks used in previous research on pseudoneglect in three population age groups - preschool children, middle school children, and young adults. The experimental tasks were line bisection, word bisection, and drawing a house, a person, and a tree on a blank sheet of paper in landscape orientation. The dependent variable was the degree of lateral (left or right) spatial bias in participants' performance.

Participants were asked to perform two passive (line and word bisection) and three active spatial tasks (drawing a simple figure). Line bisection is the most widely used and rigid measure for neglect in clinical settings, and for pseudoneglect in neurologically normal people, and is more connected to attention. The word bisection task is thought to activate semantic processing together with the spatial representations. Drawing reflects higher order cognitive functions - instead of passive judgment, it involves the ability to plan and execute a simple task in peripersonal space. We hypothesized that with age the direction of the bias should shift from right to left, and this would be reflected in the performance of the middle school group.

## Participants

60 Bulgarian speaking participants ( 22 men) took part in the study. The preschool group consisted of 19 children (8 boys), with a mean age of 4 years and 4 months ( $M=53.37$, $S D=3.47$, range 49-59, calculated in months). The middle school group consisted of 20 children ( 4 boys), with a mean age of 148 months ( 12 years and 3 months), $S D=3.65$ and range 139-154 months. The adults were 21 ( 10 men), with a mean age of 28 years, $S D=11.02$ and range $18-51$ years. All adults gave their written informed consent before the study. Informed written consent for the children was given by their parents.

All participants underwent assessment for handedness, given that some studies report a significant influence of handedness on bisection and drawing tasks (e.g. Jewell \& McCourt, 2000; Picard \& Zarhbouch, 2014). Only righthanded participants' data were considered further. From the analyses were excluded data from two preschoolers who used predominantly their left hand in the drawing tasks (in accordance with Kastner-Koller, Deimann, and Bruckner, 2007), two participants from the 12 -year-old group who self-reported as left-handed; three adult participants who self-reported as ambidextrous but reported predominant use of their left hand on the handedness questionnaire.

Thus, statistical analyses were performed on the data from 17 preschool children ( 7 boys), 18 middle-schoolers ( 4 boys) and 18 adults ( 9 men). The mean age of the adult and middle school groups remained the same, and a t-test showed no significant difference between the age of the recruited and the analyzed groups of preschoolers ( $p=$ .890). None of the preschool children had reading or writing habits.

## Stimuli

Two types of stimuli were used in the study - five straight horizontal lines and ten words, presented on two separate sheets of paper A4 format, portrait orientation. The lines and the words were printed in advance, while the drawing tasks were executed on blank sheets of paper format A4 with landscape orientation. The lines had a mean length of 10.9 $\mathrm{cm}(S D=2.2)$, and were positioned to the left or to the right side of the sheet, in a way that no two lines was exactly below each other. Only very long and low frequency words were chosen for the task. The mean length of the words in characters was $13.4(S D=2.1)$, or $3.9 \mathrm{~cm}(S D=0.7)$ and their mean objective frequency was 0.19 ( $S D=0.23$, range $0.00-0.59)$. Objective frequency data for the Bulgarian words (Simov, Osenova, Kolkovska, Balabanova, \& Doikoff, 2004) was converted into frequency score per million and 10 -base logarithm of the score was taken with one added to the score per million to avoid the undefined $\mathrm{Lg}(0)$. The words had an odd number of characters in order to avoid an overlap between their orthographic and physical center. For the same reason a handwriting script and not block letters, was used. The words were written in Segoe Script, bold, font size 13, again on different positions to the left or to the right of the A4 sheet.

Table 1: Means and SDs (in parentheses) of the line and word length in centimeters; word length in number of characters and their objective frequency (Simov et al., 2004)

|  | Length in cm | Length in characters | Frequency |  |
| :---: | :---: | :---: | :---: | :---: |
| Words | $3.9(0.7)$ | $13.4(2.1)$ | $0.19(0.23)$ |  |
| Lines | $10.9(2.2)$ |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

## Procedure

All participants had to perform line and word bisection, and also draw a house, a person and a tree on three separate sheets of paper. The order of the line and word bisection tasks and the drawing tasks was counterbalanced for the adults and the middle-schoolers. Preschoolers performed the tasks in random order. The sheets of paper were placed one by one in front of the participants and taken away after the execution of each task. Participants were asked to cross the middle of the lines and words as fast and as accurately as possible, and to draw a house, a person and a tree, on their own terms. For the 4 -year-olds the instructions were more detailed, as the experimenter had to be sure that they understood the task well.

## Results

The performance on the active and passive tasks was measured with two different types of indices. For the analyses of the line and word bisection data, a Percent deviation score was calculated as the difference between the left bisected part and the true half, divided by the true half and multiplied by 100 ((left bisected-half)/half*100) (see Fujii, Fukatsu, Yamadori, \& Kimura, 1995; Failla, Sheppard, \& Bradshaw, 2003).
Also, a novel drawing bias index was developed that took into account both the size of the drawing and its deviation from the center. First, the distance between the two outmost points on the lateral axis of the drawing was divided by two. The distance from this central point $\mathrm{C}_{1}$ to the outer points was taken as the drawing`s radius R. Each Bias index was calculated as the proportion of the shortest distance ( $\perp$ ) from $\mathrm{C}_{1}$ to the sheet's midline $\mathrm{C}_{2}\left( \pm \perp \mathrm{C}_{1} \mathrm{C}_{2}\right.$, negative values coded left), and the absolute sum of $\pm \perp \mathrm{C}_{1} \mathrm{C}_{2}$ and R .

$$
\text { Drawing Bias Index }=\frac{ \pm \perp C_{1} C_{2}}{\left| \pm C_{1} C_{2}+R\right|}
$$

This calculation yielded values between -1 and +1 , with zero for the centrally positioned objects. For both measures negative values indicated left bias and positive values indicated right bias.

## Preschool children

The data of 17 four-year old children from the preschool age group were subjected to analyses of Bias for each of the five tasks. Table 2 shows the means, standard deviations, and range of values for Line Bisection and Word Bisection Bias.

Table 2: Means, standard deviations, and range for Deviation Percent Score (Line/Word Bisection) in 4 year olds.

| Experimental <br> Task (Bias) | M | SD | Min | Max |
| :--- | :---: | :---: | :---: | :---: |
| Line Bisection | 6.88 | 11.84 | -12.83 | 30.20 |
| Word Bisection | 6.05 | 9.47 | -16.20 | 22.18 |

Note: Negative values correspond to a Left side preference.
In the Line Bisection Task, children's choices were significantly biased towards the right-hand side of the lines, as seen in the percent deviation score, $M=6.88, S D=$ $11.84, t_{(16)}=2.40, p=.03$. Children's performance in the Word Bisection Task was similarly biased rightwards, $M=$ $6.05, S D=9.47, t_{(16)}=2.63, p=.02$.

The drawing placement choices of the preschool children were analyzed in terms of the House bias index, Person bias index, and Tree bias index, respectively.

One-sample t-tests evaluated whether children's drawings were positioned with significant lateral bias. Table 3 shows the means, standard deviations, and range for each drawing task.

Table 3: Means, standard deviations, and range for the drawing tasks bias indices in preschool children.

| Experimental <br> Task: Bias Indices | M | SD | Min | Max |
| :--- | :---: | :---: | :---: | :---: |
| Draw a House | .24 | .54 | -.83 | .90 |
| Draw a Person | -.01 | .56 | -.82 | .93 |
| Draw a Tree | .11 | .49 | -.73 | .71 |

Note: Negative values correspond to a Left side preference.
There was no lateral bias in the Draw-a-house task, $M=.24$, $S D=.54, t_{(16)}=1.79, p=.09$, the Draw-a-person Task, $M=$ -. $01, S D=.56, t_{(16)}=.11, p>.1$, or the Draw-a-tree Task, $M$ $=.11, S D=.49, t_{(16)}=.96, p>.1$. In neither of the three object drawing tasks did children's object placements deviate reliably from the sheet's midline.

## Middle school children

The data of 18 twelve-year olds from the middle school age group were analyzed for Bias for each of the five tasks. Table 4 shows the means, standard deviations, and range for the Line Bisection and Word Bisection Bias measures.

Table 4: Means, standard deviations, and range for Deviation Percent Score (Line/Word Bisection) in the middle school group.

| Experimental <br> Task (Bias) | M | SD | Min | Max |
| :--- | :---: | :---: | :---: | :---: |
| Line Bisection | .22 | 4.94 | -9.69 | 9.25 |
| Word Bisection | -.97 | 9.71 | -26.09 | 12.31 |

Note: Negative values correspond to a Left side preference.
In both bisection tasks, twelve-year olds' choices were not reliably biased to one of the sides, all $p$ 's $>.1$.

The drawing choices of the 12-year olds were analyzed in a similar way to the data of the preschool children. Onesample $t$-tests yielded a reliable lateral bias for the Person and Tree Bias index measures. Table 5 shows the means, standard deviations, and range for each drawing task.

Table 5: Means, standard deviations, and range for the drawing tasks bias indices in 12-year old children.

| Experimental <br> Task: Bias Indices | M | SD | Min | Max |
| :--- | :---: | :---: | :---: | :---: |
| Draw a House | -.06 | .17 | -.39 | .21 |
| Draw a Person | -.40 | .14 | -.60 | -.13 |
| Draw a Tree | -.13 | .15 | -.41 | .15 |

Note: Negative values correspond to a Left side preference.
The analyses revealed no lateral bias in the Draw-a-house task, $p>.1$, but we found a reliable Left bias in the Draw-aperson and Draw-a-tree tasks, $t_{(17)}=11.87, p<.001$, and $t_{(17)}$ $=3.59, p=.002$, respectively.

## Adult group

The data of 18 adult participants were analyzed for Bias. Table 6 shows the means, standard deviations, and range for the Line Bisection and Word Bisection Bias measures.

Table 6: Means, standard deviations, and range for Deviation Percent Score (Line/Word Bisection) in adults.

| Experimental <br> Task (Bias) | M | SD | Min | Max |
| :--- | :--- | :---: | :---: | :---: |
| Line Bisection | -1.21 | 5.49 | -13.10 | 7.25 |
| Word Bisection | -.83 | 3.21 | -8.45 | 4.78 |

Note: Negative values correspond to a Left side preference.
In neither task, did adults' performance show any lateral bias, p's > .1.

The drawing choices of the adults were analyzed in terms of the House, Person, and Tree bias indices. Table 7 shows the means, standard deviations, and range for each object drawing task.

Table 7: Means, standard deviations, and range for the Draw-a-House, Draw-a-Person, and Draw-a-Tree indices, in the adult group.

| Experimental <br> Task: Bias Indices | M | SD | Min | Max |
| :--- | :---: | :---: | :---: | :---: |
| Draw a House | -.27 | .26 | -.72 | .25 |
| Draw a Person | -.45 | .22 | -.77 | .03 |
| Draw a Tree | -.25 | .21 | -.58 | .13 |

Note: Negative values correspond to a Left side preference.
In all three object drawing tasks, adults exhibited a reliable Left bias: in the Draw-a-house task, $M=-.27, S D=.26, t_{(17)}$ $=4.23, p=.001$, in the Draw-a-person Task, $M=-.45, S D=$ $.22, t_{(17)}=8.79, p<.001$, and in the Draw-a-tree Task, $M=-$ $.25, S D=.21, t_{(17)}=5.07, p<.001$.
In summary, we found a reliable Right bias in preschool children's performance on the two bisection tasks, and a reliable Left bias in adults' performance on the drawing tasks. Twelve-year olds had no lateral bias on the bisection tasks, and a reliable Left bias on two of the three drawing tasks. In addition, in order to evaluate whether their performance differed from each of the other age groups we analyzed their performance measures in combination with the other groups in separate multiple analyses of variance on the bisection tasks and on the drawing tasks. A MANOVA revealed a significant main effect of age group in line and word bisection $\left(F(4,98)=3.75, p=.007 ; \eta_{p}^{2}=.133\right)$. A post-hoc analysis (Scheffé) revealed that the difference was between the adults and the preschoolers ( $p=.016$ for the line bisection task; and $p=.05$ for the word bisection task). No difference was found between the performance of the middle school children and either of the other two groups (Figure 1).


Figure 1. Percent deviation line and word scores as a function of age group. Positive values indicate deviation to the right and negative values indicate deviation to the left.

Error bars denote .95 confidence intervals. $* p<.05$.
A MANOVA showed a significant main effect of age group in the drawing tasks $\left(F(6,96)=4.73, p<.001 ; \eta_{p}^{2}=\right.$ .228). Post-hoc analysis (Scheffé) revealed that the adults and the preschoolers performed differently in all drawing tasks - in the draw-a-house task $(p=.001)$, in the draw-atree task $(p=.006)$ and in the draw a person task $(p=.003)$. Middle school children differed from preschoolers only in the draw-a-person task $(p=.01)$. Again, no difference was found between the performance of the middle school children and the adults (Figure 2).


Figure 2. Drawing bias index as a function of age group. Positive values indicate deviation to the right and negative values indicate deviation to the left. Error bars denote .95 confidence intervals. ${ }^{* * *} p<.001 ; * * p=<.01 ; * p<.05$.

## Discussion

The present study showed spatial bias in different cognitive spatial tasks across three age groups. Interestingly, the direction of the demonstrated bias depended significantly on the nature of the tasks. In the passive tasks (line and word bisection), results showed bias only for the
preschool group, and in the active (drawing) tasks - only for the middle school children and the adults. This suggests that implicit and explicit tasks might reflect different types of spatial processing. Furthermore, with development the direction of the bias was indeed shifted from right to left, but also across the tasks. Generally, all participants deviated more to the left in the implicit tasks. One explanation would be that the explicit instruction in the bisection task made participants more attentive, resulting in overdriven movements for the preschoolers because of the pervasive motor component. Another speculation would be that a certain familiarity effect could enhance the ability to form object representations in the peripersonal space, resulting in difference in performance.

Preschoolers exhibited significant rightward bias in the line and word bisection tasks, consistent with reports of previous studies (Dellatolas et al., 1996; Dobler, Manly, Atkinson, Wilson, Ioannou, \& Robertson, 2001). As reported by Dellatolas et al. (1996), when four-year-old children use their right hand, a low degree of hemispheric interaction due to callosal immaturity may be the reason which leads to enhanced left hemisphere involvement, shift of attention to the right side and rightwards overestimation.

The 12-year-old middle school children were not reliably biased in the bisection tasks. As reported by Hausmann et al. (2003), there is a robust developmental step to the adult pattern of pseudoneglect between the ages of 10-12 and 1315. In their study, a group of $10-12$ year old children demonstrated a symmetrical neglect for the line bisection task, whereas their eldest group (13-15 year olds) showed a leftwards bias. As proposed by Hausmann et al. (2003), this developmental step can be attributed both to corpus callosum maturation and hormonal change during puberty.

Furthermore, for the adult group, the line and word bisection task showed no significant bias although the overall mean scores indicated a leftward directionality.

In order to assess the leftward bias we used implicit measures of visual-spatial computation as in the person-treehouse drawing task (Barrett et al., 2002). As expected, the adult group demonstrated a leftward bias for all three types of object drawings, similar to the results in the study of Barret et al. (2002). The middle-schoolers` drawings also showed a leftward bias though to a smaller degree. This leftward bias is consistent with previous studies using the draw-a-person task (Heller, 1991) and the draw-a-tree task (Picard \& Zarhbouch, 2014) and indicates that at this age the right hemisphere asymmetric activation is evident and reading habits affect spatial attention.

Notably, Barrett et al. (2002) reported that reading habits (left to right or right to left) could not reverse the leftward bias in the house-tree-person drawings of adults. In our study we found that preschoolers exhibited no lateral bias and tended to place their drawings in the middle of the page. Thus, when reading habits are not established and there is no imbalanced left hemispheric activation, children's perceptual right space is not attenuated.

## Conclusion

The present study used both implicit and explicit tasks in three age groups with different reading skills. Age had significant effect in all tasks, but results depended on the nature of the tasks. Adults differed from preschoolers in all tasks. Middle school children performed like the adults, and differed from the preschoolers in only one of the implicit tasks. These results are consistent with previous studies stressing the importance of corpus callosum maturation in the the asymmetric activation of the right hemisphere, as well as the importance of the reading habits.

Unlike previous research, we did not find significant spatial bias for the adults and the middle school children in the explicit tasks. One explanation might be the enhanced executive control over motor performance. This would mean that when attention is engaged, pseudoneglect might be attenuated with maturation. However, due to the small sample sizes, not definite conclusion can be made.

Notably, in the present study was used a novel index to assess the spatial bias in the drawings` placement, taking into account not only the deviation of the drawing but its size. That is why it is difficult to compare our results with earlier results. Subsequent studies are planned with 7-8 year old children and illiterate adults in order to examine further the influence of brain maturation and scanning habits on the spatial representation of the immediate surroundings.

## References

Arduino, L. S., Previtali, P., \& Girelli, L. (2010). The centre is not in the middle: Evidence from line and word bisection. Neuropsychologia, 48(7), 2140-2146.
Barrett, A. M., Kim, M., Crucian, G. P., \& Heilman, K. M. (2002). Spatial bias: Effects of reading direction on Korean subjects. Neuropsychologia, 40, 1003-1012.
Bradshaw, J. L., Spataro, J. A., Harris, M., Nettleton, N. C., \& Bradsaw, J. (1988). Crossing the midline by four to eight year old children. Neuropsychologia, 26, 221-235.
Bowers, D., \& Heilman, K. M. (1980). Pseudoneglect: effects of hemispace on a tactile line bisection task. Neuropsychologia, 18(4), 491-498.
Chokron, S., \& De Agostini, M. (1995). Reading habits and line bisection: A developmental approach. Cognitive Brain Research, 3(1), 51-58.
Chokron, S., \& Imbert, M. (1993). Influence of reading habits on line bisection. Cognitive Brain Research, 1(4), 219-222.
Cicek, M., Deouell, L. Y., \& Knight, R. T. (2009). Brain activity during landmark and line bisection tasks. Frontiers in Human Neuroscience, 3, 1-8.
Dellatolas, G., Coutin, T., \& De Agostini, M. (1996). Bisection and perception of horizontal lines in normal children. Cortex, 32(4), 705-715.
Dobler, V., Manly, T., Atkinson, J., Wilson, B. A., Ioannou, K., \& Robertson, I. H. (2001). Interaction of hand use and spatial selective attention in children. Neuropsychologia, 39, 1055-1064.

Failla, C. V., Sheppard, D. M., \& Bradshaw, J. L. (2003). Age and responding-hand related changes in performance of neurologically normal subjects on the line-bisection and chimeric-faces tasks. Brain and Cognition, 52(3), 353-363.
Fischer, M. H. (1996). Bisection performance indicates spatial word representation. Cognitive Brain Research, 4(3), 163-170.
Fujii, T., Fukatsu, R., Yamadori, A., \& Kimura, I. (1995). Effect of age on the line bisection test. Journal of Clinical and Experimental Neuropsychology, 17(6), 941-944.
Gabay, Y., Gabay, S., Henik, A., Schiff, R., \& Behrmann, M. (2015). Word and line bisection in typical and impaired readers and a cross-language comparison. Brain and Language, 150, 143-152.
Hausmann, M., Waldie, K. E., \& Corballis, M. C. (2003). Developmental changes in line bisection: A result of callosal maturation?. Neuropsychology, 17(1), 155-160.
Heller, W. (1991). Hemispatial biases in children on the Draw-A-Person Test. Developmental Neuropsychology, 7, 151-160.
Jewell, G., \& McCourt, M. E. (2000). Pseudoneglect: a review and meta-analysis of performance factors in line bisection tasks. Neuropsychologia, 38(1), 93-110.
Kastner-Koller, U., Deimann, P., \& Bruckner, J. (2007). Assessing handedness in pre-schoolers: Construction and initial validation of a hand preference test for 4-6-yearolds. Psychology Science, 49(3), 239-254.
Luh, K. E. (1995). Line bisection and perceptual asymmetries in normal individuals: What you see is not what you get. Neuropsychology, 9(4), 435-448.
Macdonald-Nethercott, E. M., Kinnear, P. R., \& Venneri, A. (2000). Bisection of shapes and lines: analysis of the visual and motor aspects of pseudoneglect. Perceptual and Motor Skills, 91(1), 217-226.
Michel, C., Bidot, S., Bonnetblanc, F., \& Quercia, P. (2011). Left minineglect or inverse pseudoneglect in children with dyslexia?. Neuroreport, 22(2), 93-96.
Ossandon, J. P., Onat, S., \& Konig, P. (2014). Spatial biases in viewing behavior. Journal of Vision, 14(2), 20-20. doi:10.1167/14.2.20
Picard, D., \& Zarhbouch, B. (2014). Leftward spatial bias in children's drawing placement: Hemispheric activation versus directional hypotheses. Laterality: Asymmetries of Body. Brain and Cognition, 19(1), 96-112.
Simov, K., Osenova, P., Kolkovska, S., Balabanova, E., Doikoff, D. (2004). A language resources infrastructure for Bulgarian. Proceedings of Language Resources and Evaluation Conference (pp.1685-1688). Lisbon, Portugal (http://www.bultreebank.org/papers/BulgarianL RI316.pdf).
Vingiano, W. (1991). Pseudoneglect on a cancellation task. International Journal of Neuroscience, 58(1-2), 63-67.

# Is Red Fire Warmer than Blue Fire? Colored Thermal Words in a Stroop Task 

Yordanka Zafirova<br>Department of Cognitive Science and Psychology, New Bulgarian University, Bulgaria<br>Armina Janyan<br>Research Center for Cognitive Science, New Bulgarian University, Bulgaria


#### Abstract

In many languages there are concepts for warm and cold colors. Research on color-temperature correspondence and their interaction is quite scarce, and based mostly on subjective measures. It is still unknown whether and to what extent colors bear the thermal information. The current study explored the relationship between warm and cold colors (red and blue) and thermal aspects of the word semantics (sun, snow), using the Stroop paradigm in a color categorization task. It was hypothesized that if colors activate the thermal meaning then Stroop effect should occur. The results suggested a colortemperature compatibility effect - faster responses when associated color and thermal meaning corresponded (e.g. sun presented in red). This provides important information on the automaticity of thermal activation during word processing, and on the strength of conceptual associations in color perception. It was suggested that words induced mental simulation of the thermal concepts, together with the associated color.


# The Redundancy Effect in Human Causal Learning: Evidence Against a Comparator Theory Explanation 

Tara Zaksaite (gintare.zaksaite@plymouth.ac.uk)<br>School of Psychology, Plymouth University, Plymouth, UK<br>Peter M. Jones (peter.m.jones@plymouth.ac.uk)<br>School of Psychology, Plymouth University, Plymouth, UK


#### Abstract

The blocking effect, canonical in the study of associative learning, is often explained as a failure of the blocked cue to become associated with the outcome. However, this perspective fails to explain recent findings that suggest learning about a blocked cue is superior to a different type of redundant cue. We report an experiment designed to test the proposal that blocking is not a failure of association, but a performance effect arising from a comparator process (Denniston, Savastano, \& Miller, 2001). Participants received A $+\mathrm{AX}+\mathrm{BY}+\mathrm{CY}-$ training containing a blocked cue X and another redundant cue Y , before rating outcome expectancies for individual cues. These ratings were inconsistent with the association-failure view. After subsequent A- Y+ training, participants rated cues again. Ratings in the second test were inconsistent with the comparator theory. Our data suggest that neither perspective is likely to provide a complete account of causal learning.


Keywords: associative learning; comparator theory; redundancy effect; blocking; cue competition

## Introduction

In a typical causal learning task, participants are required to learn which cues cause an outcome. Many such tasks involve presentation of more than one cue on each trial, and this typically results in cue competition. That is, learning about a cue is dependent to some extent on accompanying cues. Probably the best-known example of cue competition is blocking (e.g. Dickinson, Shanks, \& Evenden, 1984). In a blocking task, participants receive trials on which cue A is paired with an outcome (denoted A+) and trials on which A is presented alongside a second cue, X , and paired with the outcome (AX+). Blocking is said to have occurred if learning about $X$ is restricted by the presence of $A$, relative to a control condition in which A+ trials are omitted. Learning about X is therefore influenced by the presence and associative history of A . This finding is analogous to classic demonstrations of blocking in nonhuman animals (e.g. Kamin, 1969).

Following the discovery of cue competition effects, Rescorla and Wagner (1972) outlined an elegant and muchcited model according to which an outcome will only support learning if it is surprising. Surprise is equivalent to prediction error, or the discrepancy between the outcome that is expected and the outcome that occurs. When an unexpected outcome occurs, the resulting prediction error enables the formation of an association between any cues
present and the outcome. Critically however, outcome expectancy is based on all the cues that are present rather than individual cues. To illustrate this, consider the blocking effect. On AX+ trials, expectancy of the outcome is based on the extent to which it is predicted by both A and X . Because A is established as a predictor of the outcome on A+ trials, the outcome is expected on AX+ trials and little learning can take place. Learning about X is therefore 'blocked' by the presence of A. If A were not separately paired with the outcome, blocking would not occur. Informally, we can say that X is blocked because it is informationally redundant; it indicates no change in the outcome that is predicted by A. According to the RescorlaWagner model, this is operationalized as a failure by X to become associated with the outcome.

This description of blocking as a failure of association formation has been called into question by a recent result comparing learning about a blocked cue with another kind of redundant cue. Uengoer, Lotz, and Pearce (2013) compared learning about the blocked cue X with cue Y from a BY+CY- discrimination. Here, the outcome was predicted by B and its absence was predicted by C . We refer to the common cue, Y , as an uncorrelated cue because it is paired with both the presence and the absence of the outcome. Uengoer et al. gave participants $\mathrm{A}+\mathrm{AX}+\mathrm{BY}+\mathrm{CY}-$ training, followed by test trials on which they were asked to rate outcome expectancy for each cue. The Rescorla-Wagner (1972) model predicts that learning about X should be blocked by A, as described above. The prediction for Y is perhaps more complex, but the model predicts that the strength of the association between Y and the outcome will increase overall during training. This results from the use of a combined prediction error in determining learning, as follows: On BY+ trials, the associations between B and Y and the outcome should strengthen. On CY- trials, the association between Y and the outcome should lead to expectation of the outcome, and its non-occurrence will in turn lead to decreases in the extent to which both C and Y predict the outcome. As a consequence, C should be established as an inhibitor of the outcome. This will enable Y to maintain its association with the outcome to some extent. Informally, we can say that Y could be a cause of the outcome if its absence on CY- trials is explained by the preventative status of C . The Rescorla-Wagner model, then, predicts that Y will become better associated with the outcome than will X. Contrary to this prediction, Uengoer et
al.'s participants rated $X$ as a more likely cause of the outcome than Y. This finding is known as the redundancy effect (for corresponding results using rats and pigeons, see Jones \& Pearce, 2015; Pearce, Dopson, Haselgrove, \& Esber, 2012).

## Comparator theory

While the redundancy effect is not predicted by the Rescorla-Wagner (1972) model, it perhaps makes intuitive sense because X is consistently paired with the outcome. Y, on the other hand, is paired with the outcome only intermittently. Ignoring any effect of cue competition, we might expect X to become better associated with the outcome than Y. Accordingly, Uengoer et al. (2013) considered whether their results might be better accounted for by supposing that blocking occurs not because X fails to become associated with the outcome, but because of an additional process that acts during the test. According to comparator theory (Denniston, Savastano, \& Miller, 2001), association formation is non-competitive and driven by an individual prediction error for each cue. Cue competition is then accounted for by a comparator process that operates at test to influence performance. This process compares the associative status of the target cue with that of any cues that have previously been presented alongside the target. This results in a decrease if companion cues have a strong association with the outcome, and an increase if the association is weak. In the case of X, outcome expectancy will be reduced because A is strongly associated with the outcome, and blocking will occur. This model also predicts the redundancy effect, because association formation is governed by each cue's relationship with the outcome. X should be better associated with the outcome than Y, because Y is only followed by the outcome on $50 \%$ of trials. The comparator theory therefore seems like a promising candidate for explaining both blocking and the redundancy effect.

However, two attempts have been made to test this account and both have cast doubt on its validity. Jones and Pearce (2015) conducted an experiment in which rats were given $\mathrm{A}+\mathrm{AX}+\mathrm{BY}+\mathrm{CY}$ - training, where each cue was an auditory or visual stimulus and the outcome was the delivery of a sucrose solution. Rats were subsequently tested in extinction with $\mathrm{B}, \mathrm{X}$, and Y. A larger response was elicited by X than Y , demonstrating the redundancy effect. Responding was also higher for B than for X . Jones and Pearce suggested that this was important, because it allowed a further test of the comparator theory. According to this theory, because B and X were both consistently paired with the outcome, they should have become associated with the outcome to the same extent. The larger response for B than for X at test must therefore have been the result of the comparator process. Because B had been presented alongside Y, which was only weakly associated with the outcome, the response to B was left largely intact. For X, however, the response was moderated because X had been trained alongside A, which was strongly associated with the
outcome. To test this account, rats were given A- Y+ training. Following this, they were again tested with B and X . The comparator theory now predicts greater responding for X than for B , but the results closely resembled those from the first test. B elicited more responding than X despite revaluation of the comparator cues A and Y , apparently in contradiction of the theory. An objection may be raised, however, because of the nature of the outcome used in this experiment. Miller and Matute (1996) suggested that, once a target cue becomes associated with an outcome of motivational significance, the target cue itself acquires motivational significance. As a result, attempts to deflate responding to the target cue by further conditioning of an associate cue may be unsuccessful. In the experiment reported by Jones and Pearce, the appetitive outcome is likely to have had substantial motivational significance. It is therefore possible that responding to $B$ was unaffected by Y+ training, not because the comparator theory is incorrect but because the manner in which it was tested was inadequate. Urushihara and Miller (2010) noted that such revaluation effects are difficult to observe in nonhuman animals because of the use of motivationally significant outcomes, but occur frequently in human causal learning.

There also exists a test of whether the comparator theory can account for the redundancy effect in humans, reported by Uengoer et al. (2013). Since blocking is dependent on a comparison between X and A , it follows that revaluation of A should increase outcome expectancy for X . In one experiment, following initial $\mathrm{A}+\mathrm{AX}+\mathrm{BY}+\mathrm{CY}-$ training and subsequent individual cue testing, participants were given A- training and a further test. They found that outcome expectancy for X was equivalent for the two tests, contrary to the predictions of the comparator theory. This conclusion should be treated with caution, however. The crucial comparison is between outcome expectancy for X during the first and second tests. This means that the results are likely to have been contaminated to some extent by order effects. In the present paper, we report an experiment intended to provide a fairer test of the comparator theory. The experiment is conceptually similar to the Jones and Pearce experiment, except that it used human participants and a causal learning task. It therefore combines the better aspects of the existing evaluations of the comparator theory described above, while eliminating the shortcomings. The use of human participants should provide ideal conditions for observing revaluation effects and, because the adequacy of the comparator theory can be assessed by comparing B and $X$ in the same test, the confounding effect of order present in the Uengoer et al. experiment is avoided.

## A test of the comparator theory

The design of this experiment is summarized in Table 1. Stage 1 of the experiment was designed to establish the causal status of B, X, and Y. Each participant received four types of trial: A+, AX+, BY+, and CY-. Following Uengoer et al. (2013), training was embedded in a variant of the classic allergist task (Aitken, Larkin, \& Dickinson, 2000).

On each trial, participants were shown one or two food pictures and asked to predict whether they would lead to stomach ache in a fictional patient, Mr. X. After participants made their predictions, they received feedback on whether stomachache did (+) or did not (-) occur. After the completion of Stage 1, a test stage was administered in which participants were shown the five individual food cues and asked to rate the likelihood of stomach ache for each food using a rating scale. These ratings served as the measure of outcome expectancy for each cue. We expected these ratings to resemble those obtained by Uengoer et al. That is, we expected ratings to be higher for X than for Y (the redundancy effect) and to be higher for B than for X . We also expected ratings to be high for A and low for Y . After this test, participants received further training in Stage 2. This training was designed to revalue A and Y , and consisted of A- and Y+ trials. Following this training, outcome expectancies were again measured in the same way as in the earlier test. If the comparator theory (Denniston, Savastano, \& Miller, 2001) is correct, ratings for X should be higher than ratings for B in this test. Alternatively, if the outcome expectancy for B was higher than for X at Test 1 because of a difference in the strength of associations formed between these cues and the outcome during Stage 1, then ratings should still be higher for B than for X at Test 2.

Table 1: The design of the experiment.

| Stage 1 | Test 1 | Stage 2 | Test 2 |
| :--- | :--- | :--- | :--- |
| A+ | A | A- | A |
| AX + | B | $\mathrm{Y}+$ | B |
| BY+ | C |  | C |
| CY- | X |  | X |
|  | Y |  | Y |
| 8 blocks | 2 blocks | 8 blocks | 2 blocks |

## Method

Participants The participants were 50 Plymouth University undergraduate students studying Psychology. They received course credit for their participation in this experiment. They were aged $18-53$ years $(M=21.86, S D=7.1)$ and five were male.

Materials The experiment was run using computers attached to 22 -inch monitors with a $1920 \times 1080$ resolution. The experiment was designed, cues presented and responses recorded, using E-prime 2.0 software (Psychology Software Tools, PA, US).

The cues were five images of foods on a white background, each measuring $300 \times 300$ pixels. The foods were: apple, cherry, grape, lemon and strawberry. Foods were randomly assigned to serve as each cue ( $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{X}$, Y ) for each participant. Outcomes were stomach ache, signified by text and a sad face on a red background, and no stomachache, indicated by text and a happy face on a green
background. Cues and outcomes were presented on a black background with white text. Participants responded using the mouse.

Procedure Each participant was initially asked to read onscreen instructions that were identical to those used by Uengoer et al. (2013). In the first stage of the experiment participants were presented with eight blocks of trials. Each of the four trial types ( $\mathrm{A}+, \mathrm{AX}+, \mathrm{BY}+, \mathrm{CY}-)$ were presented once per block, and were randomized within each block. Each trial started with the presentation of either one or two images of foods, below the phrase "The patient ate the following food(s):" The sentence "Which reaction do you expect?" was presented below the images. Participants responded by clicking one of two response buttons placed at the bottom of the screen. The left-hand button was labelled "No stomach ache", and the right-hand button was labelled "Stomach ache". As soon as the participant responded, the response buttons and the sentence above them were replaced by a statement and image showing the outcome of the trial. When the outcome was stomach ache, the statement was "The patient has stomach ache" and the picture of a sad face was shown. When the outcome was no stomach ache, the statement was "The patient has no stomach ache", and the picture of a happy face was shown. This feedback display remained on the screen for 3 s and was then followed by the next trial.

After the completion of Stage 1, Test 1 began. Here the participants were instructed to judge the probability with which specific foods will cause stomach ache in the absence of feedback. On each trial a single food was presented on the screen below the sentence "What is the probability that the food causes stomach ache?" Participants responded by clicking on an 11-point rating scale ranging from 0 (Certainly not) to 10 (Very certain). After participants chose a rating for each food, a blank screen was shown for 1 s . Each food that appeared in Stage 1 was presented twice, with the order of trials randomly determined for each participant. For each participant, the average of the two outcome expectancy ratings was calculated and used in subsequent analyses.

Participants then received further training in Stage 2. Training consisted of eight blocks of two trial types (A-, $\mathrm{Y}+$ ) appearing once per block in a random order. The procedure for this stage was otherwise identical to Stage 1. Test 2 then measured final outcome expectancies, using the same procedure as Test 1 .

## Results

We applied an inclusion criterion of $60 \%$ correct predictions in Stage 1, commonly used in similar work (e. g. Le Pelley \& McLaren, 2003). Four participants failed to meet this, and are excluded from all subsequent analyses. The remaining 46 participants learned readily, and made $98 \%$ correct responses during the final block of trials of Stage 1. These participants also made correct predictions on $98 \%$ of trials
during the final block of Stage 2. Our analyses here focus on the critical test data.

Mean ratings from Tests 1 and 2 are shown in Figure 1. The pattern of results in Test 1 closely resembles those obtained by Uengoer et al. (2013), with higher ratings for X than for Y , and higher ratings for B than for X . For the comparator theory, the crucial comparison is between B and X at Test 2. As for Test 1, ratings for B were higher than for X. A two-way ANOVA with test and cue variables was conducted. This revealed a significant effect of test, $F(1,45)$ $=6.59, p=.014, \eta_{\mathrm{p}}^{2}=.128$, a significant effect of cue, $F(4$, 180) $=80.76, p^{<}<.001, \eta_{\mathrm{p}}{ }^{2}=.642$, and a significant interaction, $F(4,180)=207.85, p<.001, \eta_{\mathrm{p}}{ }^{2}=.822$. To explore this interaction, simple effects analyses were used to compare ratings from Tests 1 and 2 for each cue. Ratings differed between tests for $\mathrm{A}, F(1,45)=431.15, p<.001, \eta_{\mathrm{p}}{ }^{2}$ $=.905$, for $\mathrm{B}, F(1,45)=29.30, p<.001, \eta_{\mathrm{p}}^{2}=.394$, for C , $F(1,45)=23.28, p<.001, \eta_{\mathrm{p}}^{2}=.341$, and for $\mathrm{Y}, F(1,45)=$ 359.38, $p<.001, \eta_{p}^{2}=.889$. Ratings for X did not differ between tests, $F<1$.

Separate analyses were conducted to test the most informative comparisons. Firstly, in order to check that the redundancy effect was obtained, we used a within-subjects $t$-test to compare ratings for X and Y at Test 1. Ratings for X were significantly higher than for $\mathrm{Y}, t(45)=7.58, p<$ .001. Secondly, to confirm that the revaluation of A and Y was successful, we conducted a two-way ANOVA comparing ratings for A and Y in the two tests. We found an effect of cue, $F(1,45)=6.37, p=.0151, \eta_{\mathrm{p}}{ }^{2}=.124$, an effect of test, $F(1,45)=5.12, p=.029, \eta_{p}^{2}=.102$, and importantly, a significant interaction, $F(1,45)=559.83, p<$ $.001, \eta_{\mathrm{p}}^{2}=.926$. Exploring this interaction, we found that ratings for A were higher than for Y at Test $1, F(1,45)=$ $559.62, p<.001, \eta_{\mathrm{p}}^{2}=.926$, but that ratings were higher for Y than for A at Test $2, F(1,45)=112.41, p<.001, \eta_{\mathrm{p}}^{2}=$ .714. The revaluation of A and Y was therefore successful. Thirdly, to test the predictions of the comparator theory, we conducted a similar two-way ANOVA to compare ratings for $B$ and $X$ at Test 1 and Test 2 . We found an effect of cue, $F(1,45)=46.82, p<.001, \eta_{\mathrm{p}}^{2}=.510$, an effect of test, $F(1$, $45)=21.40, p<.001, \eta_{\mathrm{p}}^{2}=.322$, and a significant interaction, $F(1,45)=18.81, p<.001, \eta_{\mathrm{p}}^{2}=.295$. Exploring the interaction, we found that ratings for B were higher than for X at both Test $1, F(1,45)=62.86, p<.001, \eta_{\mathrm{p}}^{2}=.583$, and Test 2, $F(1,45)=16.52, p<.001, \eta_{\mathrm{p}}^{2}=.268$. This disconfirms the predictions of the comparator theory. If outcome expectancies for B and X were determined by a combination of direct associations with the outcome and comparison with Y and A respectively, then ratings for X should have been higher than for B at Test 2. One notable feature of the data that might suggest some role for a comparator process is the change in ratings for B between the two tests. Participants rated B as a less likely cause of the outcome after Y+ training than they did before, which is consistent with the comparator theory. However, an opposite effect was observed for C . Ratings for C were higher at Test 2 than at Test 1, which is the opposite change
to that predicted by the comparator theory. It therefore seems likely that these changes are not the result of a comparator process, but rather a general decrease in certainty at Test 2 . Since A and Y had been revalued in Stage 2, some participants may have assumed that associations learned during Stage 1 were no longer reliable.


Figure 1: Mean ratings for Test 1 and Test 2, for each cue. Error bars show the standard error of the mean.

## Discussion

The experiment reported here was designed to test an explanation of the redundancy effect based on the comparator theory (Denniston, Savastano, \& Miller, 2001). Following A+AX+BY+ CY- training, participants were asked to rate the probability of the outcome occurring for each individual cue in Test 1. Ratings were higher for X than for Y; we therefore replicated the redundancy effect (Uengoer et al., 2013). This finding is consistent with the comparator theory, which states that the strength of the association formed between a cue and an outcome is determined by an individual (i.e. non-competitive) prediction error. Since $X$ was consistently paired with the outcome and Y was not, it follows that X should have become better associated with the outcome than Y . Participants also gave higher ratings for B than for X during Test 1. Again, this is consistent with the comparator theory. Although the theory predicts that each of these cues will have become associated with the outcome to the same extent, it also states that outcome expectancies should have been moderated by the comparator process at test. Specifically, outcome expectancy for X should have been reduced because it had been trained alongside A, which was strongly associated with the outcome. Any reduction in outcome expectancy for $B$ should have been smaller, because it had been trained alongside Y, which was only weakly associated with the outcome. However, the comparator theory is not consistent with the results of Test
2. Following Stage-2 A- Y+ training, participants again rated the probability of the outcome occurring for each cue. Ratings for B were again higher than for X . The comparator theory, however, predicts the opposite pattern of results. This is because, although the associations with the outcome should have remained unchanged for both B and X , the associative status of their comparator cues had changed. We therefore conclude that the comparator theory cannot account for our results. Of course, this conclusion relies on the assumption that Stage-2 training was successful in revaluation of A and Y . This is apparent in the higher ratings given for Y than for A at Test 2.

Our results are also difficult to reconcile with the model of learning proposed by Rescorla and Wagner (1972). Because it describes learning as being the result of a combined prediction error, X should have failed to become associated with the outcome and should have been rated as a less likely cause of the outcome than Y at Test 1. In other words, the Rescorla-Wagner model fails to account for the redundancy effect because it incorrectly predicts that learning about the blocked cue will be prevented. However, Vogel and Wagner (2017) have suggested a way in which the prediction of blocking can be modified to accommodate the redundancy effect. Their modification is based on the assumption that each cue shares some common features, denoted K. The training given in Stage 1 here could therefore be re-described as AK+ AXK+ BYK + CYK-. K should become associated with the outcome, with two consequences that are relevant for interpreting the redundancy effect. Firstly, because K is present on CYKtrials, overexpectation of the outcome is increased and the weakening of the association between Y and the outcome on these trials is more substantial than when K is omitted. Secondly, When XK is presented at test, outcome expectancy is boosted by K ; the model can therefore predict greater outcome expectancy for XK as a result of including the common features. Combination of these two changes allows the model to predict the redundancy effect. This version of the model also makes an interesting prediction regarding the effect of adding further trial types to Stage-1 training. Because the extent to which K becomes associated with the outcome is critical, adding extra trials on which the outcome does not occur (e.g. DK-) should reduce the influence of common features and eliminate the redundancy effect. This prediction remains untested. If it is correct, it would lend support to an account that provides a way to reconcile the Rescorla-Wagner model with the redundancy effect.

Another possibility is that learning is governed by quite different rules. Not all models of learning make such strong predictions about the restriction of learning about blocked cues. Pearce's $(1987,1994)$ configural model, for instance, predicts substantial outcome expectancy for blocked cues. According to this model, participants learn about configural representations that include all cues present on a given trial, rather than each cue entering into its own association with the outcome. In the case of blocking, participants would
come to associate A with the outcome on A+ trials, and to associate AX with the outcome on AX+ trials. Outcome expectancy for X alone would then be determined by generalization from AX, based on their similarity. Outcome expectancy for X would therefore be weaker than for AX , but considerably stronger than it would have been without any training. However, Pearce at al. (2012) note that the theory is unable to predict the redundancy effect because it predicts that outcome expectancy for Y will be higher still. As with the Rescorla-Wagner (1972) model, it is possible that some modification of the configural model would alter this prediction, but it is not clear at present what that modification might be.

Whether cues are learned about individually or as configurations, the redundancy effect might be accommodated if we suppose that the amount of attention paid to blocked and uncorrelated cues changes during training. For instance, it is commonly assumed (Le Pelley, 2004; Mackintosh, 1975) that cues are processed to the extent that they have predictive value. Since blocked and uncorrelated cues are both redundant, we might expect the amount of attention they are paid to be reduced. In order to explain the redundancy effect, however, we need to propose that this reduction in attention differs in magnitude for blocked and uncorrelated cues. If we suppose that participants learn quickly that Y is irrelevant during $\mathrm{BY}+$ CY- training, then we might expect substantial decreases in the amount of attention paid to Y and a weak association between Y and the outcome as a result. Attention to X, on the other hand, might be maintained for longer, allowing a stronger association to form between X and the outcome. In an attempt to evaluate this claim empirically, Jones and Zaksaite (2017) monitored participants' eye gaze during A+ AX + BY + CY- training. The duration of eye gaze for each cue has been used extensively as a measure of overt attention in learning tasks (e.g. Beesley \& Le Pelley, 2011). Jones and Zaksaite found that participants spent more time looking at Y than at X , but that this was likely to have been a consequence of differing trial durations. When X and Y were presented on the screen together in a subsequent stage of training, gaze was equivalent for each. This experiment therefore failed to provide any evidence that the amount of attention paid to blocked and uncorrelated cues differs.

In addition to associative accounts of blocking, others (e.g. Lovibond, Been, Mitchell, Bouton, \& Frohardt, 2003) have argued that blocking is the result of inferential reasoning. According to this view, blocking occurs because participants do not have independent evidence that the blocked cue causes the outcome (i.e. training trials on which X is presented without A ). However, since participants also lack evidence that the blocked cue does not cause the outcome, they should be uncertain about the causal status of $X$ and blocking should be relatively weak. This uncertainty might be enhanced because the magnitude of the outcome is fixed, meaning that compound presentation of two causal cues would lead to the same outcome as either cue alone. Lovibond et al. provided support for this position by
showing that blocking is enhanced when the magnitude of the outcome varies in accordance with the number of causes present, allowing participants to infer that the blocked cue is not a cause of the outcome. In light of this account, we should consider whether the intermediate ratings for X in the present experiment were the result of an intermediate level of learning, or of uncertainty about its causal status. An unpublished experiment from our laboratory suggests that this might be a promising approach. In addition to rating the probability of the outcome for each cue, participants rated their confidence in these judgments. Confidence ratings were lower for blocked than for uncorrelated cues, suggesting that the redundancy effect might be due at least in part to uncertainty about X .

## Conclusion

We have considered theories that account for learning by using individual and combined prediction errors. While combined prediction error models (e.g. Rescorla \& Wagner, 1972) are difficult to reconcile with the redundancy effect, individual prediction error does not result in cue competition effects such as blocking, unless an additional process is invoked. We tested a theory that includes such a process (Denniston, Savastano, \& Miller, 2001), but found that it was not consistent with the results of Test 2 . We suggest two lines of future enquiry. Firstly, data should be collected that evaluate the predictions arising from Vogel and Wagner's (2017) addition of common features to simulations of the Rescorla-Wagner model. While the cues used in the present experiment are likely to have shared some common features, we cannot currently evaluate the claim that learning about these features enables the redundancy effect to occur. Secondly, since neither combined nor individual prediction errors seem capable of producing our results, attempts should be made to evaluate some combination of the two. In particular, any models containing both kinds of prediction error should be tested against the idea that cue competition occurs because of inferential reasoning processes.

## References

Aitken, M. R. F., Larkin, M. J. W., \& Dickinson, A. (2000). Super-learning of causal judgements. Quarterly Journal of Experimental Psychology, 53B, 59-81.
Beesley, T., \& Le Pelley, M. E. (2011). The influence of blocking on overt attention and associability in human learning. Journal of Experimental Psychology: Animal Behavior Processes, 37, 114-120.
Denniston, J. C., Savastano, H, I., \& Miller, R. R. (2001). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In R. R. Mowrer \& S. B. Klein (Eds.), Handbook of contemporary learning theories. Mahwah, NJ: Erlbaum.
Dickinson, A., Shanks, D., \& Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. Quarterly Journal of Experimental Psychology, 36A, 29-50.

Jones, P. M., \& Pearce, J. M. (2015). The fate of redundant cues: Further analysis of the redundancy effect. Learning and Behavior, 43, 72-82.
Jones, P. M., \& Zaksaite, T. (2017). The redundancy effect in human causal learning: no evidence for changes in selective attention. Manuscript submitted for publication.
Kamin, L. J. (1969). Selective attention and conditioning. In N. J. Mackintosh \& W. K. Honig (Eds.), Fundamental issues in associative learning. Halifax, Nova Scotia: Dalhousie University Press.
Le Pelley, M. E. (2004). The role of associative history in models of associative learning: a selective review and a hybrid model. Quarterly Journal of Experimental Psychology, 57B, 193-243.
Le Pelley, M. E., \& McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. Quarterly Journal of Experimental Psychology, 56B, 68-79.
Lovibond, P. E., Been, S. L., Mitchell, C. J., Bouton, M. E., \& Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. Memory and Cognition, 31, 133-142.
Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review, 82, 276-298.
Miller, R. R., \& Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgement. Journal of Experimental Psychology: General, 125, 370-386.
Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. Psychological Review, 94, 61-73.
Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. Psychological Review, 101, 587-607.
Pearce, J. M., Dopson, J. C., Haselgrove, M., \& Esber, G. R. (2012). The fate of redundant cues during blocking and a simple discrimination. Journal of Experimental Psychology: Animal Behavior Processes, 38, 167-179.
Rescorla, R. A., \& Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black \& W. F. Prokasy (Eds.), Classical conditioning II: Current theory and research. New York, NY: Appleton-CenturyCrofts.
Uengoer, M., Lotz, A., \& Pearce, J. M. (2013). The fate of redundant cues in human predictive learning. Journal of Experimental Psychology: Animal Behavior Processes, 39, 323-333.
Urishihara, K., \& Miller, R. R. (2010). Backward blocking in first-order conditioning. Journal of Experimental Psychology: Animal Learning and Cognition, 36, 281295.

Vogel, E. H., \& Wagner, A. R. (2017). A theoretical note on the interpretation of the "redundancy effect" in associative learning. Journal of Experimental Psychology: Animal Learning and Cognition, 43, 119-125.

# Modeling Semantic Fluency Data as Search on a Semantic Network 

Jeffrey C. Zemla (zemla@wisc.edu)<br>Joseph L. Austerweil (austerweil@wisc.edu)<br>Department of Psychology, University of Wisconsin-Madison, 1202 W. Johnson Street<br>Madison, WI 53706 USA


#### Abstract

Psychologists have used the semantic fluency task for decades to gain insight into the processes and representations underlying memory retrieval. Recent work has suggested that a censored random walk on a semantic network resembles semantic fluency data because it produces optimal foraging. However, fluency data have rich structure beyond being consistent with optimal foraging. Under the assumption that memory can be represented as a semantic network, we test a variety of memory search processes and examine how well these processes capture the richness of fluency data. The search processes we explore vary in the extent they explore the network globally or exploit local clusters, and whether they are strategic. We found that a censored random walk with a priming component best captures the frequency and clustering effects seen in human fluency data.


Keywords: memory; search; semantic networks; fluency

## Introduction

One important task for the human mind is to retrieve knowledge from memory when it is needed. To investigate how the mind solves this task, psychologists have used the semantic fluency task (Bousfield, \& Sedgewick, 1944), in which participants generate as many unique items as they can from a category (e.g., "Name as many animals as you can") in a fixed amount of time (e.g., one minute). Semantic fluency data are richly structured. For example, researchers have found a frequency effect: Items that occur more often in the world are also produced more often in fluency lists. Under the assumption that knowledge is represented as a semantic network, we evaluate a number of possible process models of memory retrieval by determining how well they can reproduce the rich structure of human fluency data.

A debate has emerged as to whether a censored random walk over a semantic network (Abbott, Austerweil, \& Griffiths, 2015) or a strategic search in a high-dimensional space (Hills, Jones, \& Todd, 2012) better describes how the mind retrieves knowledge from memory. Central to the debate has been one property of fluency data: People tend to retrieve items in clusters in a manner consistent with optimal foraging (Hills et al., 2012). This is the tendency to search in memory within a cluster (sub-category) until the gains from further search in that cluster are outweighed by the benefits of switching to a new cluster. Both models can retrieve items in a manner consistent with optimal foraging. One issue in resolving this debate is that the semantic network account has only made computational-level claims and qualitative comparisons to human data. To advance the debate, we explore different possible search processes and compare their retrieval behavior to human retrieval behavior
by examining two traditional effects seen in semantic fluency data: frequency and clustering effects. We test processes that vary in the extent that they search strategically and explore the network.

The semantic fluency task is often scored by a simple count of the number of items named. While healthy controls typically have no trouble generating many items, patients with memory deficits such as Alzheimer's disease or semantic dementia recall fewer items (Troyer, Moscovitch, Winocur, Leach, \& Freedman, 1998). In addition, items that are typical of a category are reported more frequently than items that are atypical (Henley, 1969). For instance, cat is more likely to be named by a participant than lynx. This is particularly pronounced in patients with memory deficits, who often only recall items that are frequent in natural statistics (Sailor, Antoine, Diaz, Kuslansky, \& Kluger, 2004). Another well-studied property of semantic fluency data is clustering (Troyer, Moscovitch, \& Winocur, 1997). Healthy control participants tend to cluster items together in recall. For example, a participant may list cat, dog, and hamster in sequence because all three items belong to a common sub-category: pets.

In this paper, we evaluate how well different search processes on a semantic network reproduce frequency and clustering effects found in human fluency data. To do so, we first describe several search processes. Next, we measure frequency and clustering performance in humans using previously collected semantic fluency data (Zemla, Kenett, Jun, \& Austerweil, 2016). Then, we implement several network search procedures on a standardized semantic network and calculate those same measures for comparison.

## Memory Retrieval as Search

A model of memory retrieval is a search process over some representation. A search process begins with a cue (e.g., a category label) and uses that cue to locate relevant information (e.g., category members). A crucial component to any search process is the procedure used to navigate the representation (i.e., how the next item in search is determined). Search processes vary in whether they are local or global, or have aspects of both. A local search process will move from the current location to one nearby (in the representation) using information associated with the current location. A global search process may move from the current location to one far away using information encoded across the representation. Search is typically performed in conjunction with some executive process, to determine the relevance of the information encountered. For example, when searching through memory for "animals", it
is necessary to recognize that some items in the search path are not animals, and that some animals have already been found. This executive process does not need to be overt; in the search processes described below, it is assumed that search can traverse over items without conscious awareness and without being reported. In addition, search processes vary in whether they are strategic or not. Strategic search processes involve greater use of working memory and executive functioning to determine where to direct search next. Note that these dimensions are meant to help organize search processes and are idealized (e.g., most search processes are not purely local or global).

From behavioral evidence alone, it is difficult to infer properties of human search. For example, clustering in semantic fluency data has been taken as evidence of a strategic search process that leverages local cues (Troyer et al., 1997). To list animals we may start search at a random animal, say elephant. At this point, it may be more efficient to limit search to only African animals, as these items are more accessible. Once search exhausts its store of African animals, we switch back to a global search of any animal. This strategic cluster-and-switch process produces clustered fluency data as seen in participants. However, the switch between global and local cues does not need to be strategic if the underlying memory representation is organized in clusters (Abbott et al., 2015). Under this view, even a simple search process can produce clustered data simply by listing items in the order they are encountered in memorythe burden of efficiently retrieving items is built into the representation, rather than the process.

In this paper, we assume semantic memory is best represented as a semantic network and evaluate different search processes on it. Without specifying both a representation and a process, it is not possible to make claims from behavioral results (Abbott et al., 2015). We do so as a first step towards resolving the semantic network vs. space debate.

## Search Processes Over a Semantic Network

In this section, we define a semantic network and outline different possible search processes on it. Some processes, such as node degree search (NDS), rely on global cues, selecting the next item independent of the current item. Others, such as cluster-based depth first search (CbDFS) or the censored random walk (CRW), exploit local clusters by constraining search to nearby nodes. Most processes use a mixture of these cues, including three variations of the CRW which implement random jumping ( $\mathrm{CRW}+\mathrm{RJ}$ ), strategic jumping (CRW + SJ), and priming (CRW + PV) We also implement a basic spreading activation model (SA), which has been conceptually influential in the history of semantic networks (e.g., Collins \& Loftus, 1975). Figure 1 illustrates how the search processes are approximately distributed over the local-global dimension, and Figure 2 depicts a hypothetical semantic network and a possible search outcome for each search process.


Figure 1. Search processes ranked in terms of how far they tend to move on each step. Note that the ordering is given for the parameter values used in the paper and the precise ordering depends on the parameter values (e.g., CRW+RJ is more global-like when the jump probability is close to 1 ).


Figure 2. (top) A hypothetical network. (bottom) Example observed search paths for each method.

Semantic networks A common way to represent knowledge in memory is using a semantic network. A semantic network consists of nodes and edges. A node encodes a specific item in memory, such as dog or cat. Two nodes are connected to each other via an edge whenever those two items are semantically similar. For instance, $\operatorname{dog}$ and cat may be connected by an edge because they are both pets, but $d o g$ and elephant are unlikely to be connected.

In this work, we examine a semantic network that is both undirected and unweighted. Undirected means that if there is an edge between camel and horse, search could go from camel to horse or horse to camel (even though people might be more likely to say horse after camel than vice versa). Unweighted means that all edges imply the same amount of relatedness between two nodes. For example, if the network has one edge connecting horse and pony and another edge connecting horse and camel, the network encodes that a horse is as related to a pony as it is to a camel. Nodes sharing an edge are called neighbors. Although these assumptions may seem unrealistic, previous work found that a random walk over an undirected and unweighted semantic network captures optimal foraging behavior (Abbott et al., 2015). These distinctions may influence the clustering and frequency properties of semantic fluency data, but using directed and weighted edges increases the complexity of model substantially. So, we use an undirected network without weights and determine whether fluency data can be approximated by different search processes over it.

Node Degree Search (NDS) Node degree search selects nodes with probability proportional to a node's degree (the number of edges connected to a node). This corresponds to the relative frequency each node is visited by an "infinite" length random walk, which is a predictor of phonetic fluency data (Griffiths, Steyvers, \& Firl, 2007). Nodes with a large degree have many neighbors, and are typically encountered more frequently than nodes with a small degree. This search process chooses items based on their approximate frequency within the network regardless of the current location. Thus, it is a global and non-strategic process.

Cluster-based Depth First Search (CbDFS) Cluster-based depth-first search is equivalent to traditional depth-first search, except that the primary unit is a node and its neighbors (cluster) rather than a single node. Search begins at a starting node and outputs all of the neighbors of that node (in random order), skipping any node that has already been output. The process then moves to the most recently output node that has new neighbors and outputs those neighbors. Search is local, always emitting the current node's nearest neighbors and traversing one edge at a time.

Censored Random Walk (CRW) A censored random walk (Abbott et al., 2015) begins at a starting node and proceeds to follow a random walk, outputting each node the first time it is traversed (subsequent traversals over the same node are not output, i.e., they are "censored"). It is a local search process because it only depends on the neighborhood of the current node and can only move to a neighbor of the current node. However, because it only outputs the nodes it observes for the first time, items adjacent in its output may be far apart on the network. Nonetheless, it is much more likely to output a sequence of nodes that are close (in network space).

Censored Random Walk with Random Jumps ( $\mathbf{C R W}+\mathbf{R J}$ ) A censored random walk with random jumps is equivalent to a CRW with one key exception: At each step, the walk may jump to another node in the network (possibly one unconnected to the current node) with probability $\theta_{\mathrm{RJ}}$. (Goñi et al., 2010). The target node is chosen proportional to the node's degree (number of edges). As such, nodes that have more edges are more probable jump targets. As with the CRW, this search process is partially local, but due to random jumps, it has a global component. The decision to switch between local and global cues is random, and not a strategic decision.

Censored Random Walk with Strategic Jumps ( $\mathbf{C R W}+\mathbf{S J}$ ) A censored random walk with strategic jumps is similar to one with random jumps, except jump points are not chosen at random. Rather, the jumps occur after encountering $\theta_{\text {SJ }}$ censored nodes. The number of censored nodes is a proxy for time spent without outputting a new item, and is as a metacognitive cue that the current cluster is
exhausted. As with the CRW+RJ, this model will switch between local and global search. However unlike the random jump model, this switch is strategic: the switch is performed when there is evidence that the local cluster has been exploited sufficiently.

Censored Random Walk with Priming Vector (CRW+PV) One artifact of collecting multiple fluency lists from the same individual is that they are not independent. This is particularly pronounced when multiple lists are collected during a single session, as in our data set (Zemla et al., 2016). This results in search being biased by transitions made in a previous search (priming effects).

The censored random walk with priming vector attempts to capture this by biasing transitions toward those transitions produced in the previous list. Search is still a random walk, but whenever it reaches a node present in the previous list, with probability $\theta_{\mathrm{PV}}$ it transitions to the next observed node in the previous list (if such a transition exists) and with probability $1-\theta_{\mathrm{PV}}$ it moves to a random adjacent node. This search is primarily local, and does not have a strategic component.

Spreading Activation (SA) Classic models of semantic networks (e.g. Collins \& Loftus, 1975) explain priming effects using spreading activation. Each node has an activation value attached to it. An initial activation value of 1.0 is given to the starting node, with all other nodes given an activation of 0.0 . Activation spreads between all nodes through edges, decaying as it propagates through the network with proportion $\theta_{\mathrm{SA}}$ at each step. At each step, after performing a batch update of all node activation values, the search process chooses a node with probability proportional to its activation value. This node is then assigned an activation of 1.0. Note that we bound all activation values to be between 0.0 and 1.0. As activation begins to spread throughout the network, search quickly resembles global search as every node's activation eventually reaches and stays at 1.0. Once this happens, the search returns unobserved nodes with uniform probability.

## Experiment and Simulation Details

In this section, we describe the previous data used to evaluate the search processes, as well as the simulation and parameter fitting procedures.

## Human Data

We use human data from a previously reported experiment (Zemla et al., 2016). In their study, twenty participants were recruited from Amazon's Mechanical Turk. Each participant performed the semantic fluency task three times for three categories (animals, vegetables, and fruit). Participants entered items as they came to mind and hit "Enter" after each item, which notified the participant that the item was recorded and cleared it from the screen. Participants were instructed not to repeat an item within a list, but could repeat items across lists. Categories were pseudo-
randomized so that no participant received the same category twice in a row and each triad of lists contained each category once. For each list, participants were asked to generate as many items as they could from the category in three minutes (with a visible timer). We only analyze the results for the animal category. The data were cleaned after collection, correcting any spelling mistakes, removing pluralizations, and standardizing synonymous animals.

## Simulations

Following previous work (Abbott et al., 2015), we used the University of South Florida (USF) free association data to construct a semantic network (Nelson, McEvoy, \& Schreiber, 2004). This network was constructed by pooling the free association data of 149 participants. Given a set of cue words, participants were asked to generate the first word that came to mind. A semantic network was constructed by drawing edges between each cue-response pair. For our simulations, we used only the largest connected component of the animal subset of the USF network. This network contains 160 nodes, 786 edges, and has an average node degree of 4.91 .

Simulated fluency data was generated for each participant using every search process. The simulated data were yoked to real participant data in two ways: First, the simulated fluency lists were matched in length to real participant lists. For instance, if a participant generated lists containing 25, 30 , and 35 items, a corresponding set of simulated fluency lists would also contain 25, 30, and 35 items. Second, the yoked list always started with the first item of the participant's real list. In some cases, participants generated items that were not in the USF network. For these cases ( $15 \%$ of items), the simulated lists were instead seeded with a close semantic neighbor (as judged by the first author). For example if a participant list started with beagle (not in the USF network), the yoked list would start with dog.

This seeding process ensures that the lists explore different parts of the USF network when applicable. Moreover, it mimics the strong primacy effects seen in the experimental data: thirteen of twenty participants started at least two lists with the same animal, six of whom started all three lists with the same animal. Removing this constraint is likely to overestimate the extent to which participants are able to generate novel animals from list to list. One hundred yoked data sets (sets of three lists, matched for length) were generated for each participant. Clustering and frequency metrics were calculated as the average across all 100 data sets for each participant.

## Parameter fitting

Four of the seven search processes contained one free parameter. The best-fit parameter was found using a grid search which minimized the maximum z-score compared to the human data across all clustering and frequency measures (described below). CRW + RJ, CRW + PV, and SA models searched parameters 0.0 through 1.0 in intervals of 0.05 . For CRW + RJ, the best fitting parameter was 0.0 (no jumping)
and so we chose the second best fit for comparison. The best fit parameters were $\theta_{\mathrm{RJ}}=.6, \theta_{\mathrm{SJ}}=1, \theta_{\mathrm{PV}}=.75$, and $\theta_{\mathrm{SA}}=.25$. CRW, NDS, and CbDFS have no free parameters.

## Quantifying Cluster and Frequency Effects

Although there are many possible statistics based on clusters and frequency, we opt for simple, transparent measures to evaluate and compare the above search processes.

Clustering The clustering of fluency data is evaluated with three measures: cluster size, number of cluster switches, and number of cluster types. Cluster size is the average number of items output from a given cluster before switching to a new cluster. The number of cluster switches is the average number of times a participant switches clusters within a list.

Clusters are determined by assigning each animal to different categories as coded by Troyer et al. (1997) and extended by Hills et al. (2012). We have further extended this coding scheme by including any animals in the data that were not in the coding scheme ( $14 \%$ of animals). Because each animal may belong to multiple categories, determining cluster switch points can be done in multiple ways. We used a fluid switch measure, which counts a cluster switch as any two adjacent items that do not share any categories.

Our third cluster-based statistic is the number of unique cluster types. This is calculated by counting the total number of categories within a list (counting all categories to which an item belongs). Intuitively, the number of unique clusters appears to measure the same thing as cluster switches-but note that a cluster switch does not imply switching to a novel cluster. That is, a participant may switch back and forth between the same two clusters. Nonetheless, participant cluster switches and number of cluster types are highly correlated ( $\mathrm{r}=.74, p<.001$ ).

Frequency We evaluate frequency effects in three ways: the number of unique animals named (unigrams), the number of unique ordered pairs of animals named (bigrams), and the distribution of unigrams. The number of unigrams and bigrams was counted across all three lists. Finally, we calculated the distribution of unigrams in the data: How many items appeared only once, twice, or three times?

## Results

Because simulations were performed on an idealized (USF) semantic network, our discussion of the results focus on the relative patterns of fit. Figure 3 shows error bars (standard error of the mean) to help gauge quantitative fit.

Cluster Size Figure 3a depicts cluster sizes for participants and search processes. On average, participants generated clusters with 2.1 items (SD .44). Processes that behaved like global search (node-degree search and spreading activation) strongly underestimated cluster sizes. Cluster-based depth first search also underestimated cluster size, despite using a local search procedure. This is surprising because clusters are close in a semantic network.

Figure 3. Each clustering and frequency measure is shown for the human data (left) and each search process. (a) Average cluster size per list, (b) Average number of cluster switches per list, (c) Average number of cluster types per list, (d) Number of unique unigrams across three lists, (e) Distribution of unigrams across three lists, (f) Number of bigrams across three lists


The censored random walk produced clusters close in size to the actual human data. Variations of the CRW that included a priming vector ( $\mathrm{CRW}+\mathrm{PV}$ ) or strategic jumps (CRW+SJ) showed very little difference in cluster size compared to the CRW. However the censored random walk with random jumps (CRW+RJ) produced smaller clusters, as compared to CRW and compared to human data. Abbott et al. (2015) had previously explored censored random walks with and without random jumps and found no discernible difference with respect to optimal foraging. Our results suggest that when cluster size is taken into account, the random jump model fits worse.

Cluster Switches Figure 3b shows the average number of cluster switches for participants and search processes. Participants switched clusters an average of 16.85 times per list (SD 5.71). The pattern of cluster switches mimicked the inverse of cluster size-models that underestimated the cluster size overestimated the number of cluster switches. This is not surprising, as participant cluster size and cluster switches are negatively correlated in the data ( $\mathrm{r}=-.45, p=$ .045). Again, processes that relied strictly on global (NDS, SA ) or local (CbDFS) cues overestimated the number of cluster switches. In contrast, the CRW, CRW+SJ, and CRW+PV all closely resembled human performance. However the CRW+RJ suffered from the inclusion of random jumps, overestimating the number of cluster switches.
Number of Cluster Types Figure 3c shows the number of cluster types for participants and search processes. Participants produced an average of 16.7 cluster types per list (SD 2.7). In contrast to the cluster size and switch data,
counting the average number of cluster types per list produced a different pattern of results. Search processes that behaved more like global search (NDS, SA, CRW+RJ) performed closest to the actual data, while other search processes tended to underestimate the number of cluster types encountered.

Of course, processes that rely on global cues will tend to have more breadth than processes that rely on local cues. It is interesting that our participants were able to generate a breadth of cluster types resembling global search, but switch clusters less often (as expected by local search). This suggests that participants try to exploit local clusters, but that when they do switch, they tend to avoid old clusters.

Number of Unique Items (Unigrams) Figure 3d plots the average number of unique items listed by participants and the search processes. Across three lists, participants generated an average of 54.4 unique items (SD 19.8), though an average of 99.4 token items (SD 29.9). CRW+PV and CbDFS both produced lists containing a similar number of unique items as the human participants. Both of these models do so by limiting exploration in different ways. CbDFS will tend to generate similar fluency lists on successive trials as there is no mechanism to make longdistance transitions within the network. CRW + PV will tend to produce similar lists because it will make the same transitions as it has in previous lists with high probability.

Distribution of Unigrams Figure 3e depicts the distribution of unigrams for participants and the search processes. On average, participants listed 23.2 items once, 17.3 twice, and 13.9 three times (SDs 16.3, 7.1, 6.1). The large number of
items listed twice and three times is indicative of priming effects from earlier lists. CRW + PV was the only search process that produced a similar distribution of unigrams (in particular, the large number of items produced three times). The other search processes strongly overestimate the number of items that appear only once, and strongly underestimate the number of items that appear three times.

Number of Bigrams Figure 3f plots the number of bigrams produced by participants and the search processes. Participants listed 87.1 unique bigrams (SD 28.4). Nearly all models produced a similar number of bigrams as people, except for CRW +PV , which produced too few bigrams. Because CRW+PV often follows the same transitions as in previous lists, fewer unique bigrams are generated.

## General Discussion

We explored whether several search processes on a semantic network adequately captured the frequency and clustering effects seen in semantic fluency data. We found that local search processes captured the average cluster sizes and number of clusters, but failed to capture the number of cluster types produced by people. A priming component to the search process is needed to capture the unigram frequency statistics, but this priming component interferes with producing the appropriate number of bigrams. Although none of the search processes captured human behavior on every measure, the censored random walk with priming performed well across many.

More broadly, the censored random walk model (CRW), and its variations that include strategic jumping (CRW+SJ) or recent memory (CRW+PV), captured much of the clustering behavior seen in human data. Search processes that heavily favored only global or local search tended to produce too many cluster switches and underestimate cluster size. Thus, people balance between local and global search. While previous research was not able to discriminate between the CRW model with or without jumps (Abbott et al., 2015), our results suggest that a random jumping model does not capture human performance well.

The search processes were less successful at modeling frequency effects in the data, though CbDFS and CRW +PV produced the best fits. Overall, this suggests that for our experimental procedure-collecting multiple fluency lists from a single participant in a single setting-the censored random walk with priming vector produces the closest fits to human data with respect to clustering and frequency effects. Its performance may be improved by including a strategic component, where it jumps to unvisited nodes or clusters after censoring multiple items in a row. Future research should investigate this and whether the different processes also replicate optimal foraging.

The current work is limited in several ways. It relies on the validity of the USF semantic network, and the assumption of unweighted and undirected edges (De Deyne, Navarro, \& Storms, 2013). This network is constructed from an aggregate of participants, and does not reflect the
variability across participants. While these may not be unreasonable assumptions, we have established a baseline to compare additional search processes in the future. One possibility is to use crossvalidation: with enough lists from each participant, some lists could be used to estimate an individual's semantic network using U-INVITE (Zemla et al., 2016), while the remaining lists are used to evaluate the search processes.

## Acknowledgements

Support for this research was provided by NIH R21AG0534676 and the Office of the VCGRE at UW-Madison with funding from the WARF.

## References

Abbott, J. T., Austerweil, J. L., \& Griffiths, T. L. (2015).
Random walks on semantic networks can resemble
optimal foraging. Psychological Review, 122(3), 558569. doi:10.1037/a0038693

Bousfield, W. A., \& Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. The Journal of General Psychology, 30(2), 149-165.
Collins, A. M., \& Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological Review, 82(6), 407-428.
De Deyne, S., Navarro, D. J., \& Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. Behavior Research Methods, 45(2), 480-498.
Goñi, J., Martincorena, I., Corominas-Murtra, B., Arrondo, G., Ardanza-Trevijano, S., \& Villoslada, P. (2010). Switcher-random-walks: A cognitive-inspired mechanism for network exploration. International Journal of Bifurcation and Chaos, 20(3), 913-922.
Griffiths, T. L., Steyvers, M., \& Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. Psychological Science, 18(12), 1069-1076.
Henley, N. M. (1969). A psychological study of the semantics of animal terms. Journal of Verbal Learning and Verbal Behavior, 8(2), 176-184.
Hills, T. T., Jones, M. N., \& Todd, P. M. (2012). Optimal foraging in semantic memory. Psychological Review, 119(2), 431-440. doi:10.1037/a0027373
Nelson, D. L., McEvoy, C. L., \& Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, \& Computers, 36(3), 402-407.
Sailor, K., Antoine, M., Diaz, M., Kuslansky, G., \& Kluger, A. (2004). The effects of Alzheimer's disease on item output in verbal fluency tasks. Neuropsychology, 18(2), 306-314.
Troyer, A. K., Moscovitch, M., \& Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. Neuropsychology, 11(1), 138-146.
Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., \& Freedman, M. (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. Journal of the International Neuropsychological Society, 4(02), 137-143.
Zemla, J. C., Kenett, Y. N., Jun, K-S., \& Austerweil, J. L. (2016). U-INVITE: Estimating individual semantic networks from fluency data. In Proceedings of the 38th Annual Meeting of the Cognitive Science Society (pp. 1907-1912). Austin, TX: Cognitive Science Society.

# The Impact of Decision Agency \& Granularity on Aptitude Treatment Interaction in Tutoring 

Guojing Zhou and Min Chi \{gzhou3,mchi\}@ncsu.edu<br>Department of Computer Science, North Carolina State University


#### Abstract

In this study, we explored the impact of the decision agency (Student vs. Tutor) and granularity (Problem vs. Step) across students with different levels of incoming competence (High vs. Low). Students were randomly assigned to four conditions and split into High and Low groups based on their pre-test scores. All students used the same Intelligent Tutoring System (ITS) called Pyrenees, followed the same general procedure, studied the same training materials, and worked through the same training problems. The only substantive differences among the four conditions were decision agency and granularity. That is: who decided to present an example or to solve a problem: the student or the ITS tutor; and was the decision made problem-by-problem or step-by-step? Our overall results showed that there were significant different impacts of the decision agency and granularity between High and Low students on learning performance. More specifically, for High students granularity was the more dominant factor in that step level decisions can be more effective than problem level decisions regardless of the decision agency; for Low students there was a significant interaction effect in that: Low students benefit significantly more when they were making problem-level decisions than making step-level decisions, but no significant difference was not found when the decisions were made by the tutor. Much to our surprise, both High and Low groups showed strong decision-making preference for problem solving over worked example at both problem and step levels.


Keywords: Aptitude Treatment Interaction, Pedagogical decisions, granularity, student-centered learning,

## Introduction

Certain learners are less sensitive to learning environments and can always learn; while others are more sensitive to variations in learning environments and may fail to learn. In order to fully honor their promises, effective learning environments should exhibit an aptitude-treatment interaction (ATI), that is, its instruction should match to the aptitude of the learner (Cronbach \& Snow, 1977). Intelligent Tutoring Systems (ITSs) are powerful educational technologies that support learning by providing step-by-step support and contextualized feedback adapted to individual learners and ITSs have demonstrated great success in many complex domains (Koedinger \& et al., 1997; Vanlehn, 2006). In our work, we explored the possibility of improving the effectiveness of ITSs from two perspectives: decision agency and granularity. Here we split students into High and Low groups based on their incoming competence and investigated the impact of these two perspectives on ATI: how the decision agency and granularity would impact students' learning across the High and Low groups.
Decision Agency: ITSs are generally designed to support users' learning by providing instructions, scaffolded problem-solving practice and on demand help. Most of existing ITSs are tutor-centered. The tutor is responsible for
selecting the next action to take at any given time. Each of these decisions affects student's successive actions and performance. In learning literature, the skills used to make such decisions are generally referred to as pedagogical skills. More formally, Chi et al. defined pedagogical skills are those "involve skillful execution of tactics, such as giving explanations and feedback or selecting the appropriate problems or questions to ask the students" (Chi, Siler, \& Jeong, 2004). Most ITSs generally employ fixed pedagogical policies that do not adapt to users' needs. For example on most ITSs students are asked to solve a series of training problems while research showed that studying worked examples can be more effective than solving problems and the former generally takes much less time (McLaren \& Isotani, 2011).

On the other hand, previous research showed that it is desirable for students to experience a sense of control over their own learning. For example, Cordova and Lepper (Cordova \& Lepper, 1996) found that offering student choices over their learning could lead to significantly better learning outcome than those who were not offered. Letting students make decisions during the tutorial process should make them feel that they are actively directing their own learning process and not just passively following it. Furthermore, prior research suggested that offering student learning choices often exhibits an ATI effect: students with different levels of competence should be offered with different choices. For example, Young split learners into High vs. Low based on survey results and found that the performance difference between the High and Low learners was significantly greater under learner's control than under system control (Young, 1996). In this paper, we provided the students with different yet both reasonable choices and let them decide how they would like to study the materials and our goal is to investigate how these choices would impact their learning differently across High vs. Low students.

Granularity: Tutoring in domains such as math and science can be viewed as a two-loop procedure (Vanlehn, 2006). In the outer loop, the tutor makes tasks or problem-level decisions such as deciding what problem to solve next, while the inner loop controls step level decisions such as whether or not to give a hint. In educational literature, 'steps' often refer to the application of a major domain principle such as Newton's Third Law of Thermodynamics. Solving a complete problem generally involves applying many individual principles in a logical order. In theory, problem-level decisions are at a larger grain size and thus once students make one 'big' decision, they can focus on comprehending an example or solving a problem. However, such "big" decision
might not be very sensitive to students' specific moment-bymoment needs. For example, offering a complete worked example to students facing difficulty with a single principle may rob them of the chance to exercise other skills. When making step-level decisions, by contrast, students may be better able to tailor their decisions to their immediate needs and current knowledge level. However, making fine-grained decisions often requires more sophisticated decision-making skills. Prior research has shown interleaving worked examples with problem solving in both problem level and step level could result in improved learning performance comparing to doing problem solving only (Van Gog et al., 2011; Salden et al., 2010). However, it remained unclear how worked example and problem solving tasks should be provided to maximize the tutoring effectiveness. Therefore, in this paper, we are going to examine the impact of different decision granularity across learners with different incoming competence.

In this study, we strictly controlled the content to be equivalent for all participants by: 1) using an ITS which provides equal support for all learners; and 2) investigating on tutorial decisions that cover the same domain content at both problem and step levels, in this case Worked Examples (WE) versus Problem-Solving (PS). In WE, students were given a detailed example showing the expert solution for the problem or step. In PS, the students were tasked with solving the same problem or step using the ITS.

Previously we investigated the impact of granularity on the effectiveness of students' pedagogical decisions by comparing students' decisions against tutor's random yet reasonable decisions. Overall, our results showed that there was a significant interaction effect between decision agency (Tutor vs. Student) and granularity (Problem vs. Step) on learning. We found that step level decisions can be more effective than problem level decisions but the students were more likely to make effective pedagogical decisions at problem level than at step level (Zhou et al., 2016). In this paper, we further investigate the impact of decision agency and granularity across students with different levels of incoming competence. Following prior research, we divided students into High and Low groups based on their pretest scores and our primary research question is: would the impact of decision agency and granularity on learning differ between the High and Low students?

## Background

WE/PS, vs. FWE: A number of researchers have examined the impacts of problem-level PS, problem-level WE, vs. Faded Worked Example (FWEs) (Renkl et al., 2002; Schwonke et al., 2009; Najar et al., 2014; Salden et al., 2010). FWEs interleave problem-solving steps with worked example steps within a single problem. Renkl et al. compared WE-PS pairs with FWE using a fixed fading policy (Renkl et al., 2002). In that study the number of example steps and problem-solving actions were strictly equal between the conditions. They found that FWEs with fixed fading policy significantly outperformed the WE-PS pairs, but no significant time-on-task differences were found. Schwonke et
al. compared FWE with a fixed fading policy to tutored PS (Schwonke et al., 2009). Over the course of two studies, they found no significant differences between the two conditions in terms of their learning outcomes. However the FWE group spent significantly less time on task than the tutored PS group. Najar and colleagues compared FWE with an adaptive fading policy to WE-PS pairs. They found that the FWE condition significantly outperformed the WE-PS condition in their learning outcomes and spent significantly less time on task (Najar et al., 2014). Finally, Salden et al. compared three conditions: FWE with a fixed fading policy, FWE with an adaptive fading policy, and PS-only (Salden et al., 2010). They found that the adaptive FWE group outperformed the fixed FWE who, in turn, outperformed PS-only and there is no significant time-on-task differences among three groups.

Thus prior researchers have shown that FWEs with effective pedagogical polices can outperform fixed WE-PS pairs. It has also been shown that the former may need significantly less time on task than the latter. However all of these studies relied on hand-coded tutor pedagogical polices whereas in this study, we investigated how students with different levels of incoming competence would differ on pedagogical decision-making at both problem and step level.
Students Pedagogical Decision on ITS: Prior research on student problem-level decision-making has primarily focused on decisions of choosing instructional content, e.g. problem selection, but not how, e.g. WE vs. PS. Mitrovic et al. showed that learners, even college students, often make poor problem selections (Mitrovic \& Martin, 2003). Long et al. compared the impact of joint student/system control again full system control over problem selection (Long \& Aleven, 2014). In joint control, the system adaptively selects the problem type while the students select the individual problems. They found no significant difference on learning between the joint control groups and the full control group. In another study, Long et al. augmented a ITS with features that help students develop effective problem selection strategies with shared student/system control and compared its effectiveness with full system control ITS (Long \& Aleven, 2016). They found that students in the shared control group learned significantly better than those in the full system control group. The results for student step-level pedagogical decision-making are inconclusive. Aleven \& Koedinger studied students' help-seeking behaviors in the Cognitive Tutor (Aleven \& Koedinger, 2000). They found that students cannot use hints effectively in that they tended to wait too long before asking for hints. Roll et al. by contrast examined the relationship between students' help-seeking patterns and their learning (Roll et al., 2014). They found that asking for help on challenging steps was generally productive while help abusing behaviors were correlated with poor learning. Finally, Wood et al. found that learners with high prior knowledge can exhibit more effective help-seeking behaviors than those with low prior knowledge learners(Wood \& Wood, 1999).

Therefore prior research on students' decision suggests that
students can benefit substantially from effective pedagogical decision-making. Yet they often lack the necessary metacognitive skills to do so. On the other hand, help in ITSs is generally provided on demand, and some students might never need to request. In this study, we controlled for this possible conflict by focusing on WE/PS decisions, and by examining both problem and step-level decision-making. By letting both High vs. Low students make pedagogical decisions, we can fully investigate the impact of decision agency and granularity on learning across students of different levels of incoming competence.

## Our Approach

We will investigate the impact of students' pedagogical decisions on learning by comparing students' decisions to tutors' random decisions at either problem or step level in order to avoid the impact of possibly misguided pedagogical policies. This study is $2\{$ Student, Tutor $\} \times 2$ \{Problem, Step $\}$ design with four conditions: 1) Stud ${ }_{P r o b}$ : problem-level student decisions; 2) Stud $d_{\text {Step }}$ : step-level student decisions; 3) Tut Prob : problem-level random tutor decisions and 4) Tut Step: steplevel random tutor decisions.

## Methods

Participants: This study was conducted in the undergraduate Discrete Mathematics course at the Department of Computer Science at NC State University in the Fall of 2015. 279 students participated in this study, which was given as their final homework assignment.
Conditions: The students were assigned to the four conditions via balanced random assignment based upon their course section and performance on the class mid-term exam. Since the two tutor-random decision groups were already compared in our prior study (Zhou et al., 2015) and the primary goal of this work is to examine the nature and effectiveness of students' pedagogical decision-making and ATI effect, we assigned twice more students to the two studentdecision groups, Stud ${ }_{\text {Prob }} \&$ Stud $_{\text {Step }}$, than the two tutorrandom groups, Tut $_{\text {Prob }} \&$ Tut $_{\text {Step }}$. The final group sizes are as follows: $N=92$ for Stud ${ }_{\text {Prob }}, N=93$ for Stud $_{\text {Step }}, N=47$ for Tut $_{\text {Prob }}$, and $N=47$ for Tut Step.

Due to the holiday break, preparations for final exams, and length of the experiment, 212 students completed the experiment. 11 students were excluded from our subsequent analysis because they performed perfectly on the pretest. The remaining 201 students were distributed as follows: $N=70$ for Stud $_{\text {Prob }} ; N=59$ for Stud $_{\text {Step }} ; N=38$ for Tut Prob $; N=34$ for Tutstep. A $\chi^{2}$ test examining the relation between condition and completion rate showd no significant difference: $\chi^{2}(3)=1.159, p=0.763$.
Probability Tutor -Pyrenees Pyrenees is a web-based ITS for probability. It covers 10 major principles of probability, such as the Complement Theorem and Bayes' Rule. Pyrenees provides step-by-step instruction, immediate feedback and on-demand hints prompting students with what they should
do next. As with other systems, help in Pyrenees is provided via a sequence of increasingly specific hints. The last hint in the sequence, the bottom-out hint, tells the student exactly what to do. For the purposes of this study we incorporated four distinct pedagogical decision modes into Pyrenees to match the four conditions.

Procedure In this experiment, students were required to complete 4 phases: 1) pre-training, 2) pre-test, 3) training on Pyrenees, and 4) post-test. During the pre-training phase, all students studied the domain principles through a probability textbook, reviewed some examples, and solved certain training problems. The students then took a pre-test which contained 10 problems. The textbook was not available at this phase and students were not given feedback on their answers, nor were they allowed to go back to earlier questions. This was also true of the post-test.

During phase 3 , students in all four conditions received the same 12 problems in the same order on Pyrenees. Each primary domain principle was applied at least twice. The minimum number of steps needed to solve each training problem ranged from 20 to 50 . The steps included variable definitions, principle applications and equation solving. The number of domain principles required to solve each problem ranged from 3 to 11 . For the FWE problems, the $S t u d_{\text {Step }}$ students were asked to make decision only on two types of steps: principle selection and principle application. To apply a principle, students need to first choose the principle they will use (principle selection) and then write the appropriate equation to apply it (principle application). We evaluated the students' decisions on both types of steps in our analysis below. The only procedural differences among the four conditions were the decision agency: Student vs. Tutor and the granularity of the decision: Problem vs. Step. Apart from this, the system was identical.

Finally, all of the students took a post-test with 16 problems. Ten of the problems were isomorphic to the pre-test problems given in phase 2. Note that the rest of six questions are non-isomorphic complicated problems.

Grading Criteria: The test problems required students to derive an answer by writing and solving one or more equations. We used three scoring rubrics: binary, partial credit, and one-point-per-principle. Under the binary rubric, a solution was worth 1 point if it was completely correct or 0 if not. Under the partial credit rubric, each problem score was defined by the proportion of correct principle applications evident in the solution. A student who correctly applied 4 of 5 possible principles would get a score of 0.8 . The One-point-per-principle rubric in turn gave a point for each correct principle application. All of the tests were graded in a doubleblind manner by a single experienced grader. The results presented below were based upon the partial-credit rubric but the same results hold for the other two. For comparison purposes, all test scores were normalized to the range of $[0,1]$.

Table 1: Learning Performance

| High Group Students |  |  |  |
| :--- | :--- | :--- | :--- |
| Cond | Pre | Iso Post | Overall Post |
| Stud Prob $(31)$ | $.851(.059)$ | $.909(.111)$ | $.843(.143)$ |
| Stud $_{\text {Step }}(28)$ | $.846(.074)$ | $.936(.062)$ | $.882(.104)$ |
| Tut $_{\text {Prob }}(20)$ | $.857(.074)$ | $.889(.088)$ | $.785(.141)$ |
| Tut $_{\text {Step }}(20)$ | $.868(.058)$ | $.931(.061)$ | $.877(.113)$ |
| Low Group Students |  |  |  |
| Cond | Pre | Iso Post | Overall Post |
| Stud $_{\text {Prob }}(39)$ | $.551(.144)$ | $.863(.107)$ | $.731(.126)$ |
| Stud $_{\text {Step }}(31)$ | $.512(.164)$ | $.772(.182)$ | $.658(.195)$ |
| Tut $_{\text {Prob }}(18)$ | $.603(.188)$ | $.764(.272)$ | $.693(.282)$ |
| Tut $_{\text {Step }}(14)$ | $.591(.132)$ | $.856(.158)$ | $.773(.167)$ |

## Results

We split students into High and Low groups based on their pre-test scores. Using a median split of 0.75 and students were divided into: High $(n=99)$ and Low $(n=102)$ groups. As expected, the High group scored significant higher than the Low group: $t(199)=17.462, p<0.0001, d=2.464$. The numbers in the parentheses in the first column of Table 1 shows the numbers of High vs. Low students across the four conditions. No significant difference was found among the four conditions on the distribution of High vs. Low students: $\chi^{2}(3)=1.1879, p=0.7559$.

Fortunately, random assignment balanced the four conditions and this balance persisted even with the groups were subdivided into High and Low. The second column in Table 1 shows the pretest scores of High and Low groups. A one-way ANOVA test on students' pre-test score shows that there is no significant difference among the four conditions: $F(3,197)=1.969, p=0.12$, or the four conditions in High group: $F(3,95)=0.581, p=0.629$ or the four conditions in the Low group: $F(3,95)=0.449, p=0.719$.

## Learning Performance

Table 1 shows a comparison of the pre-test, isomorphic posttest ( 10 isomorphic questions), and overall post-test scores among the four conditions, showing the mean (and SD) for each score.

To investigate the impact of decision agency and granularity on learning performance across High and Low students, two three-way ANOVAs were conducted using decision agency (tutor vs. student), granularity (problem vs. step), and incoming competence (High vs. Low) on the isomorphic post-test scores and the overall post-test scores respectively. For the isomorphic post-test scores, there is a significant three-way interaction effect: $F(1,193)=4.079$, $p=0.045$, a significant two-way interaction effect on decision agency and granularity: $F(1,193)=5.324, p=0.022$, a significant main effect on incoming competence: $F(1,193)=$ 26.23, $p<0.0001$ and a marginal interaction effect on granularity and incoming competence: $F(1,193)=2.854, p=$
0.093. For the overall post-test scores, there is a significant two-way interaction effect on decision agency and granularity: $F(1,193)=4.415, p=0.037$, a significant main effect on incoming competence: $F(1,193)=38.96, p<0.0001$ and a marginal significant interaction effect on granularity and incoming competence: $F(1,193)=3.521, p=0.062$. Overall, our results showed that the impact of decision agency and granularity on learning performance differs significantly between the High and the Low groups. Next we will examine the learning performance of High and Low groups separately.
High Groups A repeated measure analysis using test type (pre-test vs. isomorphic post-test) as factors and test score as the dependent measure showed a main effect for test type $F(1,95)=34.74, p<0.0001$. Thus, overall the High students learned significantly by training on Pyrenees. However, further comparisons on the condition by condition basis revealed that: no significant improvement was found from pre-test to isomorphic post-test for the High Tut Prob group: $F(1,19)=1.817, p=0.194$, but the remaining three High groups showed significant improvement: $F(1,30)=$ $6.385, p=0.017$ for $\operatorname{Stud}_{\text {Prob }} ; F(1,27)=22.58, p<0.0001$ for Stud $_{\text {Step }}$ and $F(1,19)=16.37, p=0.0007$ for Tut Step . This suggests that random problem level pedagogical decisions may not be very effective for High students.

To fully compare the learning performance among the four High groups, a two-way ANOVA analysis using decision agency and granularity as factors was conducted on the overall post-test scores. Our results showed while there is no significant interaction effect, there is a significant main effect on granularity: $F(1,95)=5.504, p=0.021$, that is, the step level decisions are significantly more effective than problem level decisions across the decision agencies. More specifically, the two step level decision groups, Stud Step and Tut $_{\text {Step }}$ scored significantly higher than the Tut Prob group: $t(38)=-2.263, p=0.029, d=0.716$ for the Tut Step group and $t(46)=-2.749, p=0.009, d=0.805$ for the Stud ${ }_{\text {Step }}$ group respectively. For isomorphic post-test scores. Twoway ANOVA analysis showed a marginal main effect on granularity: $F(1,95)=3.563, p=0.062$. Pairwise t-tests showed the StudStep group outperformed the Tut Prob group significantly: $t(46)=-2.178, p=0.035, d=0.638$ and the Tut Step group tended to outperform the Tut Prob group: $t(38)=-1.757, p=0.087, d=0.556$. Therefore, our results showed that step-level decisions are more effective for High group students than problem-level ones.
Low Groups A repeated measure analysis using test type as the repeated factor shows that Low group students learned significantly after training on Pyrenees. $F(1,98)=200.01$, $p<0.0001$. In fact, all four groups made significant improvement from pre-test to isomorphic test: $F(1,38)=117.99$, $p<0.0001$ for Stud $_{\text {Prob }} ; ~ F(1,30)=63.89, p<0.0001$ for Stud $_{\text {Step }} ; F(1,17)=8.537, p=0.010$ for Tut $_{\text {Prob }}$; and $F(1,13)=39.98, p<0.0001$ for Tut Step . This suggests that for Low students, the basic practice and problems, domain exposure, and interactivity of Pyrenees might help students

Table 2: Student Decisions

| Problem Level Decisions |  |  |  |
| :--- | :--- | :--- | :--- |
| Competence | WE | PS | Total |
| High | $1.55(1.31)$ | $8.45(1.31)$ | 10 |
| Low | $1.56(1.48)$ | $8.44(1.48)$ | 10 |
| Step Level Decisions |  |  |  |
| Competence | WE | PS | Total |
| High | $21.14(24.34)$ | $115.07(27.03)$ | 136 |
| Low | $18.84(17.28)$ | $114.26(16.84)$ | 133 |

to learn even when the problem- and step-level decisions are made randomly.

A two-way ANOVA analysis using decision agency and granularity on isomorphic post-test showed a significant interaction effect across the four Low groups: $F(1,98)=5.819$, $p=0.018$. Post hoc Pairwise t-test reveals that the Stud $d_{\text {Prob }}$ Low group scored significantly higher than the Stud Step Low group: $t(68)=2.591, p=0.012, d=0.624$. For overall posttest, our two-way ANOVA showed a marginal interaction effect: $F(1,98)=3.591, p=0.061$. Pairwise t -tests showed a trend that the $S_{t u d_{\text {Prob }}}$ group outperformed the $S_{t u d_{\text {Step }}}$ group: $t(68)=1.903, p=0.061, d=0.458$. Therefore, our results showed that Low group students benefited more from making problem level decisions than step level ones and no significantly difference was found between the two tutor decision groups: Tut Prob and Tut Step.

To summarize, our results showed that: 1) for the High group, step level decision was more effective 2) for the Low group, letting students make problem level decisions can be more beneficial than letting them make step level decisions.

## Student Pedagogical Decisions and Training Time

Student Decisions Much to our surprise, our analysis on students' decision-making preference revealed that both High and Low Groups are far more likely to choose problem solving than worked examples. For the tutor decision groups, our random policies generated a balanced 50-50 selection of WE and PS. Table 2 shows the number of pedagogical decisions made by students at both problem level and step level. Columns 2 and 3 show the average number of worked examples and problem-solving decisions made by each group. We required all students to solve two problems in order to familiarize them with Pyrenees. Therefore each student in problem-level condition made or received 10 problem-level decisions. Within each of the 10 problems, there are 6 to 24 step-level decisions. Therefore each student in step-level condition made or received about 135 step level decisions. In the following, we will compare the decision making preference across High and Low groups.

We compared the percentage of WEs students selected among different groups. For problem level decisions, both High and Low groups selected around $15 \%-16 \%$ of WEs on average. That is, both groups chose significantly less WEs than the two corresponding tutor groups: $t(49)=8.717, p<$
$0.0001, d=2.500$ for the High groups and $t(55)=-10.668$, $p<0.0001, d=3.040$ for the Low groups. The results for step level decisions are similar. High group students chose an average of $15.52 \%$ WE steps; while Low group students chose $14.16 \%$. Again, both groups chose significantly less WEs than the two corresponding tutor decision groups: $t(46)=8.920, p<0.0001, d=2.612$ for the High groups and $t(43)=10.27, p<0.0001, d=3.308$ for the Low groups. The results suggested that students were significantly more likely to choose PS than WE at both levels.

Training Time Given that our results showed that the type of student decisions was not impacted by granularity and our preliminary results showed that similar patterns were found across the two different granularities on the training time. In the following, we will combine the step and problem level decision groups and mainly focus on the impact of the decision agency on time on task for High vs. Low students.

Despite the fact that students selected more PS, surprisingly, not all of them spent more time on learning comparing to those received equal number of PS and WE from the tutor. Table 3 shows the average total training time on Pyrenees (in seconds). A two-way ANOVA analysis examining the effect of incoming competence and decision agency shows a marginal significant interaction effect: $F(1,196)=3.345$, $p=0.069$. More specifically, while no significant difference was found between the two High groups, there is a significant difference between the two Low groups in that the student decision group spent significantly more time on training than the tutor decision group: $t(99)=-2.272, p=0.025, d=0.490$.

Since student decision groups chose more PS than WE and PS is generally more time consuming than WE, we further investigated the impact of decision agency on training time by comparing the average time on each WE step and PS step. The third and fourth columns in Table 3 shows the average amount of time students spent on each WE and PS steps respectively. For the average WE step time, no significant difference was found among the four groups. For the average time on PS steps, a two-way ANOVA on decision agency and incoming competence showed a significant main effect of decision agency: $F(1,196)=14.53, p=0.0002$. That is the student decision groups spent significantly less time on each PS step than the tutor decision groups. Pairwise t-test showed that this difference is significant for both High and Low groups: $t(97)=6.118, p<0.0001, d=1.253$ for the

Table 3: Time Results

| High Group Students |  |  |  |
| :--- | :--- | :--- | :--- |
| Cond | Total | WE | PS |
| High $_{\text {Stud }}$ | $7977(1811)$ | $9(10)$ | $35(8)$ |
| High $_{\text {Tut }}$ | $8041(2503)$ | $8(5)$ | $51(18)$ |
| Low Group Students |  |  |  |
| Low $_{\text {Stud }}$ | $8612(2428)$ | $9(9)$ | $39(10)$ |
| Low $_{\text {Tut }}$ | $7457(2179)$ | $9(7)$ | $50(16)$ |

High group and $t(99)=3.888, p=0.0002, d=0.839$ for the Low group. Therefore, students worked faster on PS steps when they made decisions than when tutor decided.

## Discussion

In this study, we investigated the impact of decision agency (student vs. tutor) and granularity (problem vs. step) on learning across students with different levels of incoming competence (High vs. Low). Students were randomly assigned to four experiment conditions and split into High and Low groups based on their pre-test scores. Our results showed that all four Low groups and three out of four High groups (except the High Tut Prob group) learned significantly after training on Pyrenees. In general, the Low students learned more than their High peers. This suggests that the training of Pyrenees is generally effective especially for low students.

We found that there were significantly different impacts of decision agency and granularity across High and Low students. For the High ones, granularity is the more dominant factor in that the two step-level groups significantly outperformed the two problem-level decision groups on the overall post-test. For the Low groups, there is a significant decision agency and granularity interaction effect: while no significant difference was found between the two Low tutor decision groups, the Low Student Problem-level group learned significantly more than the Low Student Step-level group. The results suggest that for High students step level decisions can be more effective than problem level decisions, but for Low students making problem level decisions are more beneficial than making step level ones.

Surprisingly, both High and Low students selected more problem solving than worked example at both problem and step level. However, students worked faster on PS steps when they selected them than received them. A potential explanation is that the control of their own learning process produced increases in motivation and depth of engagement. Currently, we are applying Reinforcement Learning (RL) to induce effective pedagogical policies based on which we will derive a methodology for teaching effective pedagogical decisionmaking strategy. After that, we will augment our ITS with decision-making development features to help students learn those strategies and examine its effectiveness.

## Acknowledgements

This research was supported by the NSF Grant \#1432156 and \#1651909.

## References

Aleven, V., \& Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In Intelligent tutoring systems (pp. 292-303).
Chi, M. T. H., Siler, S., \& Jeong, H. (2004). Can tutors monitor students' understanding accurately? Cognition and Instruction, 22(3), 363-387.
Cordova, D. I., \& Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contex-
tualization, personalization, and choice. Journal of educational psychology, 88(4), 715.
Cronbach, L. J., \& Snow, R. E. (1977). Aptitudes and instructional methods: A handbook for research on interactions. Irvington.
Koedinger, K. R., \& et al. (1997). Intelligent tutoring goes to school in the big city. IJAIED, 8(1), 30-43.
Long, Y., \& Aleven, V. (2014). Gamification of joint student/system control over problem selection in a linear equation tutor. In Its (pp. 378-387).
Long, Y., \& Aleven, V. (2016). Mastery-oriented shared student/system control over problem selection in a linear equation tutor. In Its (pp. 90-100).
McLaren, B. M., \& Isotani, S. (2011). When is it best to learn with all worked examples? In Aied (pp. 222-229).
Mitrovic, A., \& Martin, B. (2003). Scaffolding and fading problem selection in sql-tutor. In Aied (pp. 479-481).
Najar, A. S., Mitrovic, A., \& McLaren, B. M. (2014). Adaptive support versus alternating worked examples and tutored problems: Which leads to better learning? In Umap (pp. 171-182).
Renkl, \& et al. (2002). From example study to problem solving: Smooth transitions help learning. The Journal of Experimental Education, 70(4), 293-315.
Roll, I., Baker, R. S. d., Aleven, V., \& Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. Journal of the Learning Sciences, 23(4), 537-560.
Salden, R. J., Aleven, V., Schwonke, R., \& Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. Instructional Science, 38(3), 289307.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., \& Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. Computers in Human Behavior, 25(2), 258-266.
Van Gog, T., Kester, L., \& Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices learning. Contemporary Educational Psychology, 36(3), 212-218.
Vanlehn, K. (2006). The behavior of tutoring systems. IJAIED, 16(3), 227-265.
Wood, H., \& Wood, D. (1999). Help seeking, learning and contingent tutoring. Computers \& Education, 33(2), 153169.

Young, J. D. (1996). The effect of self-regulated learning strategies on performance in learner controlled computerbased instruction. Educational Technology Research and Development, 44(2), 17-27.
Zhou, \& et al. (2015). The impact of granularity on worked examples and problem solving. In Cogsci (pp. 28172822).

Zhou, \& et al. (2016). The impact of granularity on the effectiveness of students' pedagogical decisions. In Cogsci (pp. 2801-2806).

# Information Seeking as Chasing Anticipated Prediction Errors 

Jian-Qiao Zhu (j.zhu@warwick.ac.uk)<br>Wendi Xiang (wendi.xiang11@gmail.com)<br>Elliot A. Ludvig (e.ludvig@warwick.ac.uk)<br>Department of Psychology, University of Warwick<br>Coventry, UK CV4 7AL


#### Abstract

When faced with delayed, uncertain rewards, humans and other animals usually prefer to know the eventual outcomes in advance. This preference for cues providing advance information can lead to seemingly suboptimal choices, where less reward is preferred over more reward. Here, we introduce a reinforcement-learning model of this behavior, the anticipated prediction error (APE) model, based on the idea that prediction errors themselves can be rewarding. As a result, animals will sometimes pick options that yield large prediction errors, even when the expected rewards are smaller. We compare the APE model against an alternative information-bonus model, where information itself is viewed as rewarding. These models are evaluated against a newly collected dataset with human participants. The APE model fits the data as well or better than the other models, with fewer free parameters, thus providing a more robust and parsimonious account of the suboptimal choices. These results suggest that anticipated prediction errors can be an important signal underpinning decision making.


Keywords: information seeking; early resolution of uncertainty; anticipated prediction errors; forward sampling.

## Introduction

Humans and other animals have a strong preference for informative options. They are inherently curious and will explore unknown options, even sacrificing rewards to resolve an uncertain outcome early. Sometimes the search for predictive information can be independent of profit and have no effect on the delivery of primary rewards, as if consuming information itself was rewarding (Wyckoff, 1952; Prokasy, 1956; Bromberg-Martin \& Hikosaka, 2009; Iigaya, Story, KurthNelson, Dolan, \& Dayan, 2016). On occasion, this information seeking can lead to seemingly suboptimal behaviors with animals preferring options with lower expected values (Spetch, Belke, Barnet, Dunn, \& Pierce, 1990; Roper \& Zentall, 1999). In this paper, we develop a new computational model of this information-seeking behaviour based on the idea that animals' choices reflect both the expected rewards and the anticipated prediction errors from any upcoming cues.

This preference for advanced information has been widely observed across species, including rats (Prokasy, 1956; Chow, Smith, Wilson, Zentall, \& Beckmann, 2016), pigeons (Spetch et al., 1990), starlings (Vasconcelos, Monteiro, \& Kacelnik, 2015), monkeys (Bromberg-Martin \& Hikosaka, 2009, 2011; Blanchard, Hayden, \& Bromberg-Martin, 2015), and humans (Iigaya et al., 2016). In some cases, animals even give up food or water for advance information about impending rewards, even though these advanced signals do not change the eventual reward. For example, pigeons reliably choose an alternative that provides delayed access to food $50 \%$ of the time over one that always provides the same
amount of food with the same delay, but only when an immediate cue is provided, which signals to the pigeons whether or not food will eventually be available on that trial (Spetch et al., 1990; Gipson, Alessandri, Miller, \& Zentall, 2009). The choice of the $50 \%$ option is clearly suboptimal in terms of reward-intake maximization. Similarly, when choosing between delayed, probabilistic rewards, monkeys and humans will prefer an option that informs them about the eventual outcome of that trial over one that leaves the resolution of uncertainty to the time of reward delivery (Bromberg-Martin \& Hikosaka, 2009, 2011; Iigaya et al., 2016).

In addition to the presence of advance information, a few variables have proven critical to the emergence of this suboptimal choice (see McDevitt, Dunn, Spetch, and Ludvig (2016) for review). First, the contingencies between the predictive cues and the outcomes is important because it influences the amount of uncertainty resolved by the cues: The more information conveyed by the predictive cues, the more preferred the associated choice target (Bromberg-Martin \& Hikosaka, 2009). Second, humans and other animals also exhibit a preference for earlier advanced notice of the eventual outcome. Increased information seeking has been found in the case of longer delays (Spetch et al., 1990; Iigaya et al., 2016). Third, the subjective value of advance information scales with the reward magnitude of the potential outcomes (Blanchard et al., 2015). Finally, aversive outcomes can sometimes produce outright information avoidance in human subjects, as in the Ostrich effect (Karlsson, Loewenstein, \& Seppi, 2009).

Given this rich set of empirical findings, we endeavored to build a computational model that can capture as many of these empirical results as possible, but first we briefly review the existing computational models for this information-seeking behaviour.

## Existing Computational Models

The apparent departures from optimality observed when advance information is available poses a significant computational challenge to standard models of reinforcement learning (RL) (Niv \& Chan, 2011). Previous research has explored several possible extensions and refinements to the usual RL framework, including the information bonus model (Bromberg-Martin \& Hikosaka, 2011), the disengagement model (Beierholm \& Dayan, 2010), and the anticipatory utility model (Iigaya et al., 2016). The information bonus model encapsulates the idea that receiving advance information acts as if it were a primary reward. This information bonus has alternatively been operationalized as either a free parameter
(Bromberg-Martin \& Hikosaka, 2011) or as the Shannon entropy of the reward probability (Bennett, Bode, Brydevall, Warren, \& Murawski, 2016). These ideas successfully explain the observed preference for more informative options (Bromberg-Martin \& Hikosaka, 2009). Formalizing the information bonus as the Shannon entropy, however, fails to deal with, for instance, the fact that animals prefer to observe more even when the number of bits they receive by doing so is less (Roper \& Zentall, 1999). On the other hand, without using Shannon entropy, the information bonus cannot capture the relationship between information seeking and probability, which resembles an inverse U-shaped function (Green \& Rachlin, 1977).

The anticipatory utility model is a recently proposed alternative model for these data, which formalizes the economic idea of savouring (Loewenstein, 1987; Iigaya et al., 2016). According to this model, animals are hypothesized to enjoy or savour the anticipation of guaranteed rewards to come. Anticipatory utility alone, however, cannot explain why the delay to reward would influence how much the informative option is preferred (Spetch et al., 1990; Iigaya et al., 2016). This limitation emerges because delay renders anticipatory utility less rewarding at the same speed as the primary reward. To rectify this, an additional boosting mechanism was introduced to enhance anticipatory utility, and thereby slow down effect of discounting future rewards (Iigaya et al., 2016). The full model, including this boosting mechanism, explains a wide range of information-seeking behaviours, including many of the properties of sub-optimal choice. One challenge for the anticipatory utility model is how such a mechanism could be learned locally, as the computations require full knowledge of the eventual time to reward in advance (Niv \& Chan, 2011).

## The Anticipated Prediction Error Model

Given these limitations on prior models, here, we develop an alternative formalism centered around the idea of anticipated prediction errors (APE). According to the APE model, animals draw one-step samples of their anticipated futures from a simple model of the world and calculate the prediction error that would be associated with that sample. These anticipated prediction errors are then treated as though they were rewarding in and of themselves, reminiscent of how momentary subjective well-being correlates with prediction errors (Rutledge, Skandali, Dayan, \& Dolan, 2014). These samples are biased such that futures which contain positive prediction errors are more likely to be sampled. This forward sampling (i.e. anticipation) from the current state using imagined experiences and learned environmental dynamics, such as developed in the Dyna-2 architecture (Silver, Sutton, \& Müller, 2008), can provide useful anticipatory signals that guide decision making. The critical difference between the APE model and the standard RL model is that the APE model maintains two separate valuation systems: one estimated from actual experience (model-free), and the other estimated through this forwardsampling process (model-based). The prediction errors gen-
erated via the forward traces are called anticipated prediction errors (APEs). Together with the conventional value functions, these APEs drive the preference to seek or avoid certain future states. The bias in the sampling process toward positive prediction errors can even induce suboptimal choices.

## Model Specification

We extend the standard Temporal-difference (TD) model (Sutton \& Barto, 1998) where agents are assumed to estimate an action-value function for each experimental stimulus:

$$
\begin{equation*}
Q\left(s_{t}, a_{t}\right)=\mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k} r_{t+k-1}\right] \tag{1}
\end{equation*}
$$

where $t$ indexes time, $s_{t}$ specifies the state visited at time $t$, $r_{t}$ indicates the immediate reward delivered at time $t$, and $\gamma \in$ $[0,1)$ is a discount factor, which devalues delayed rewards. This action-value function represents the expected discounted future reward. In TD learning, this action-value function is estimated through a simple incremental update mechanism:

$$
\begin{equation*}
Q\left(s_{t}, a_{t}\right) \leftarrow Q\left(s_{t}, a_{t}\right)+\alpha \delta_{t+1} \tag{2}
\end{equation*}
$$

where $\alpha$ is the learning rate and $\delta$ is the reward prediction error (RPE), calculated as follows:

$$
\begin{equation*}
\delta_{t+1}=r_{t+1}+\gamma \max _{a} Q\left(s_{t+1}, a\right)-Q\left(s_{t}, a_{t}\right) \tag{3}
\end{equation*}
$$

This RPE signal represents the difference between the value of the current state-action pair and the value of the best next-state-action plus the reward achieved in the transition. Thus, RPEs are triggered by each state transition; the mechanics of the APE model hinge on the transition from choice to the predictive cues which reveals what the eventual outcomes will be on that trial. In particular, a good cue, which resolves reward uncertainty appealingly, will generate positive RPEs, whereas a bad cue will generate negative RPEs. The RPEs will be zero in response to non-predictive cues, once the values are well learned.

Here, we define anticipated prediction errors (APE) as the perceived discrepancy between the current state (what it is like at present) and an anticipated future state (how it would be in the future) (see Figure 1). Formally, if there is no immediate primary reward delivered during the trajectory from state $s$ to $s^{\prime}$ (e.g., the transition from choice state to cue states in the information choice task), then the value of APE in state $s$ when anticipating future state $s^{\prime}$ is defined as the product of prediction errors between the two states and the transition probability:

$$
\begin{equation*}
A P E\left(s^{\prime} \mid s, a\right)=T\left(s^{\prime} \mid s, a\right) \times\left[\gamma_{s s^{\prime}}^{D_{a^{\prime}}} \max Q\left(s^{\prime}, a^{\prime}\right)-Q(s, a)\right] \tag{4}
\end{equation*}
$$

where $D_{s s^{\prime}}$ is the time taken to travel from $s$ to $s^{\prime}$, and $T\left(s^{\prime} \mid s, a\right)$ is the transition probability from $s$ to $s^{\prime}$ by taking action $a$. In the simulations here, this travel time is always taken to be 1 , but the formulation is more general.

$\longmapsto$ choice $\longmapsto$ cue period $\longrightarrow$, outcomes $\longrightarrow$

Figure 1: Formal representation of the information-choice task as a Markov Decision Process (MDP). Two offers (red and blue circles) are presented and the animal must choose one of them. A cue then appears after this initial choice, which is either informative (green $S^{+}$indicates a rewarding outcome, and yellow $S^{0}$ indicates a neutral outcome) or uninformative (black $S^{*}$ leaving the animal in a state of uncertainty). Following a delay ( $T_{\text {delay }}$ ), the animal obtains the outcome (reward or no reward). The anticipatory signals proposed by the APE model are illustrated as purple dashed lines.

Note that this computation relies on the samples generated based on the learned environment dynamics. The primary assumption of the APE model is that humans and other animals treat APEs as though they were rewarding, whereby positive APEs are reinforcing and negative APEs punishing. The APEs are positive when the anticipated value of the future sampled state is better than value of the present state and negative in the opposite case. These quantities can also be understood as measurements of the pleasure (displeasure) one derives from anticipating the good cue (bad cue). Furthermore, attention weights are assigned to each individual APE, specifying the relative likelihood that a particular future state will be sampled. Accordingly, the decision value $\bar{Q}$ of taking action $a$ is defined as the weighted sum of APEs for anticipated future outcomes plus the value function for the corresponding state:

$$
\begin{equation*}
\bar{Q}(s, a)=\sum_{s_{k} \in \mathcal{S}} w_{k} A P E\left(s_{k} \mid s, a\right)+Q(s, a) \tag{5}
\end{equation*}
$$

where $\mathcal{S}$ denotes the set of all possible future states after taking action $a$ at the state $s$ that a subject can attend to.

Given the decision values of both the cued and the uncued options, the softmax function is then used to compute the probability of choosing the cued option:

$$
\begin{equation*}
P(a)=\frac{e^{\beta \bar{Q}(s, a)}}{\sum_{a^{\prime} \in \mathcal{A}} e^{\overline{\bar{Q}}\left(s, a^{\prime}\right)}} \tag{6}
\end{equation*}
$$

where $\mathcal{A}$ is the set of all possible actions at state $s$, and $\beta$ is an inverse temperature parameter, which controls the degree of exploration.

## Experiment

We conducted an empirical experiment to evaluate the quality of the APE model in comparison with the informationbonus model discussed above. In the experiment, people were repeatedly given a choice between an informative or non-informative option, where the outcomes were delayed 20s. Outcomes were either positive (erotic images), neutral (images of objects) or negative (aversive images). Good trials involved positive or neutral images, Mixed trials involved positive or negative images, and bad trials involved negative or neutral images. These outcomes were always delivered with 50/50 odds on each trial. Qualitatively, the APE model predicts that people will seek information in the positive and mixed cases, but not the negative cases. This prediction emerges from bias toward sampling future states with positive outcomes. The information bonus model would expect equivalent information seeking in all cases, as the amount of information present is equal in all three types of choices.

## Methods

Participants Eighty human participants were recruited from the Warwick University SONA system. All participants gave informed consent and were paid a flat rate of 5 pounds for their participation.

Task Participants performed the experiment on Windows PCs running PsychoPy (Peirce, 2007). The task was a simple two-alternative forced choice between an uncued target (Keep It Secret), which was followed by a non-predictive cue, and a cued target (Find Out Now), which was immediately followed by predictive cues that signalled the eventual outcome. Choosing either the cued or the uncued option did not alter the odds nor the eventual outcomes. The only difference between the two options was the presence or absence of advance information about those eventual outcomes. After choice, the cue was present for 20 seconds in all trials. The outcome image was presented immediately at the end of this cue. To ensure participants viewed the image, they had to press a randomly selected key (indicated on the image proper) to advance to the next trial.

The experiment consisted of three different conditions in terms of the valence of eventual outcomes. In the Good condition, the gamble included $50 \%$ erotic images and $50 \%$ neutral images (as illustrated in Figure 2). In the Bad condition, the gamble included $50 \%$ aversive images and $50 \%$ neutral images. In the Mixed condition, the gamble included $50 \%$ erotic images and $50 \%$ aversive images. The images used in the experiment were previously validated in the International Affective Picture System (Lang, Bradley, Cuthbert, et al., 1999). Sixteen images from the "EroticCouple" category were selected as positive images for heterosexual subjects and another 16 images from "Mutilation" category were selected as the aversive images. Images were chosen as the rewards
so that they could be consumed immediately, as opposed to monetary rewards (Crockett et al., 2013). All participants completed 16 interleaved trials for each condition, making 48 trials in total. Participants were informed about the nature of the potential outcomes before the experiment started.


Figure 2: Human information choice task. The diagram illustrates the Good condition, which contains a gamble of $50 \%$ erotic and $50 \%$ neutral images. The experiment also tested a Bad condition ( $50 \%$ aversive and $50 \%$ neutral images) and a Mixed condition (50\% erotic and 50\% aversive images).

## Results

A total of 69 heterosexual participants ( 48 female and 21 male) completed the task. Eleven participants were excluded (6 non-heterosexual, 4 did not disclose their sexual orientation, and 1 did not complete the task). Only the data from the last nine trials per condition are reported here.

As shown in Figure 3, participants chose the cued option on average $42.8 \% \pm 8.2 \%, 60.2 \% \pm 7.6 \%$, and $71.8 \% \pm$ $6.5 \%$ in the Bad, Mixed, and Good conditions respectively. Choices in the Good condition and the Mixed condition were significantly higher than chance responding (Good: $t(68)=6.72, p<0.001, d=1.143$; Mixed: $t(68)=2.68, p=$ $0.009, d=0.456$ ). In the Bad condition, people chose the informative slightly below chance, but not significantly so $(t(68)=-1.75, p=0.085, d=-0.297)$. There were, however, considerable individual differences in the preferences for advance information (dashed grey lines in Figure 3). This pattern of responses qualitatively agree with the predictions of the APE model, but not the information-bonus model.

## Model Comparisons

Next, we attempted a quantitative model comparison, fitting both the APE model and the information-bonus model to the individual choice proportions in the current dataset.

To fit the APE model to the data, first note that the expected rewards for both options are held constant in the ex-


Figure 3: Mean percentage of choosing cued option (Find Out Now) in the Bad, Mixed, and Good conditions. Error bars indicate $\pm$ SEM in mean choice proportions. The dashed black line indicates the $50 \%$ choice probability. The dashed grey lines are individual choice probabilities.
perimental task, $Q($ cued $)=Q$ (uncued). In addition, receiving non-predictive cues leaves participants equally uncertain about the eventual outcomes, and thus sampling from those states does not generate any anticipated prediction errors, $A P E\left(S^{*} \mid\right.$ uncued $)=0$. This analysis suggests that only the APEs related to the predictive cues determine choices in the current task. Following this logic, from equation (5), we can calculate differences in the action values of the cued and uncued options in the Good, Bad, and Mixed conditions, respectively as follows:

$$
\begin{align*}
\Delta \bar{Q}_{\text {Good }} & =w^{+} A P E\left(S^{+} \mid \text {cued }\right)+w^{0} A P E\left(S^{0} \mid \text { cued }\right)  \tag{7}\\
& =\left(w^{+}-w^{0}\right) \frac{\gamma^{T} R}{4}  \tag{8}\\
\Delta \bar{Q}_{\text {Bad }} & =w^{-} A P E\left(S^{-} \mid \text {cued }\right)+w^{0} A P E\left(S^{0} \mid \text { cued }\right)  \tag{9}\\
& =\left(w^{0}-w^{-}\right) \frac{\gamma^{T} R}{4}  \tag{10}\\
\Delta \bar{Q}_{\text {Mixed }} & =w^{+} A P E\left(S^{+} \mid \text {cued }\right)+w^{-} A P E\left(S^{-} \mid \text {cued }\right)  \tag{11}\\
& =\left(w^{+}-w^{-}\right) \frac{\gamma^{T} R}{2} \tag{12}
\end{align*}
$$

where $T$ is the length of the delay, $R$ is the absolute magnitude of rewards or punishment, and $S^{+}, S^{-}, S^{0}$ are the cues indicating positive, negative, or neutral outcomes. The weight factors are associated with their corresponding future states. Note that only the differences in weights matter for model behavior.

Any individual differences are reflected by the weight pa-
rameters in the APE model. The APE model predicts no preference for advance information when $w^{+}=w^{-}=w^{0}$. We hypothesize that the differences in weights for various future outcomes give rise to information seeking or avoidance behaviors. The preferences for advance information would arise, for instance, in the Mixed condition if the model weights $S^{+}$more heavily than $S^{-}: w^{+}>w^{-}$.

As described above, the information-bonus model (Bromberg-Martin \& Hikosaka, 2011) assumes that information has an intrinsic value, $r_{\text {info }}$, which in our setting was delivered upon each state transition to an informative cue state. This model would predict the differences in decision values as follows:

$$
\begin{equation*}
\Delta \bar{Q}=r_{\text {info }} \tag{13}
\end{equation*}
$$

for all situations where information is sometimes available. We also considered a potential extension to the informationbonus model which would assign different values to different types of information: $r_{\text {info }}^{+}, r_{\text {info }}^{-}, r_{\text {info }}^{0}$ for viewing reward predicting cues, punishment predicting cues, and neutrality predicting cues respectively.

## Model Fitting

The models were fit to the the data using hierarchical Bayesian modeling (Huys et al., 2011), in which the maximum a posteriori estimate of each parameter $h_{i}$ for each participant $i$ is calculated. These parameters are treated as a random sample from a Gaussian population distribution with means and variance $\theta=\left\{\mu_{\theta}, \Sigma_{\theta}\right\}$. Model comparison was based on the integrated Bayesian Information Criteria (iBIC) scores with an uninformative prior. As such, we analyzed the log likelihood $p(D \mid M)$ of each model directly:

$$
\begin{align*}
p(D \mid M) & =\int p(D \mid \theta) p(\theta \mid M) d \theta  \tag{14}\\
& \approx-\frac{1}{2} \mathrm{iBIC}  \tag{15}\\
& =\log p\left(D \mid \theta^{M L}\right)-\frac{1}{2}|M| \log |D| \tag{16}
\end{align*}
$$

where $|D|$ is the number of choices made by all participants, and $|M|$ is the number of parameters fitted. We compute the $\log p\left(D \mid \theta^{M L}\right)$ as the sum of integrals over individual parameters:

$$
\begin{align*}
p\left(D \mid \theta^{M L}\right) & =\sum_{i} \log \int p\left(D_{i} \mid h\right) p\left(h \mid \theta^{M L}\right) d h  \tag{17}\\
& \approx \sum_{i} \log \frac{1}{K} \sum_{k=1}^{K} p\left(D_{i} \mid h_{k}\right) \tag{18}
\end{align*}
$$

where the integrals are replaced by a sum over samples from the empirical prior. This step ensures that we evaluate how well the model fits the data only using information about the group parameters.

As a result, the iBIC penalizes for model complexity, and the model with the lowest iBIC is taken as the best-fitting
model. As shown in Table 1, the best-fitting APE model vastly outperformed the different information-bonus models. We use the iBIC score of the best fitted model as a baseline and derive the differences in iBIC as $\Delta \mathrm{iBIC}$.

Table 1: Quality of model fit to behavioral data

| Model | Free parameters | Model iBIC | $\Delta \mathrm{iBIC}$ |
| :--- | :--- | :--- | :--- |
| APE | $w^{+}-w^{-}, w^{-}-w^{0}$ | 2065.1 | 40.4 |
| APE | $w^{+}-w^{-}, w^{-}-w^{0}, \gamma$ | 2126.3 | 101.6 |
| APE | $w^{+}-w^{-}, w^{-}-w^{0}, \beta$ | $\mathbf{2 0 2 4 . 7}$ | $\mathbf{0 . 0}$ |
| APE | $w^{+}-w^{-}, w^{-}-w^{0}, \beta, \gamma$ | 2137.2 | 112.5 |
| Info Bonus | $r_{\text {info }}$ | 2361.5 | 336.8 |
| Info Bonus | $r_{\text {info }}, \beta$ | 2430.8 | 406.1 |
| Info Bonus | $r_{\text {info }}^{+}, r_{\text {info }}^{-}, r_{\text {info }}^{0}$ | 2098.8 | 74.1 |
| Info Bonus | $r_{\text {info }}^{+}, r_{\text {info }}^{-}, r_{\text {info }}^{0}, \beta$ | 2081.0 | 56.3 |

## Discussion

We have introduced a novel model of information-seeking in choice, which assumes that preferences are driven by anticipated prediction errors (APEs) accumulated through simulated forward trajectories. These APEs are treated like rewards, which combined with a bias toward sampling trajectories with positive outcomes, leads to information seeking in situations with potential positive outcomes. The model was compared against an information-bonus model through a novel empirical experiment, whereby people were given the opportunity to get early information about rewarding or aversive outcomes. As the APE model predicted, and contrary to the information-bonus model, people only sought early information for positive outcomes. Quantitative model selection supported these conclusions.

In addition to better fitting the novel dataset, the APE model provides potential insights into other types of information-induced sub-optimal choices (McDevitt et al., 2016). For example, the positive APE scales with the probability of reward (larger with lower probabilities), providing a mechanism through which a lower probability reward could be preferred to a higher probability one, as sometimes observed in animals (Spetch et al., 1990; Roper \& Zentall, 1999; Gipson et al., 2009). In addition, unlike an information bonus, the APE is sensitive to the magnitude of reward and would grow with larger rewards leading to greater preference for informative options, as observed with information-seeking in monkeys (Blanchard et al., 2015). Future work will require direct simulation of these other findings, as well as further comparison to existing models, including the different information-bonus models (BrombergMartin \& Hikosaka, 2011; Bennett et al., 2016) and the anticipatory utility model (Iigaya et al., 2016).

The current experimental protocol only involved a shallow decision tree, and the corresponding APE model presented here only used one-step anticipation. For decision
trees with high branching factors and/or larger depths, however, it would be computationally intractable to sample from all possible forward trajectories. For example, one recent study used a four-stage information-seeking game, and observed systematic deviations from the optimal strategy (Hunt, Rutledge, Malalasekera, Kennerley, \& Dolan, 2016). This type of task poses yet another computational challenge for all the models discussed here. The APE model, which already involves look-ahead experiences, is readily adaptable to incorporate other, more sophisticated, planning algorithms such as Monte-Carlo tree search (Coulom, 2006). This potential extension of the model to more complex tree search remains a question for further research.

## References

Beierholm, U. R., \& Dayan, P. (2010). Pavlovianinstrumental interaction in observing behavior. PLoS Comput Biol, 6(9), e1000903.
Bennett, D., Bode, S., Brydevall, M., Warren, H., \& Murawski, C. (2016). Intrinsic valuation of information in decision making under uncertainty. PLoS Comput Biol, 12(7), e1005020.
Blanchard, T. C., Hayden, B. Y., \& Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. Neuron, 85(3), 602-614.
Bromberg-Martin, E. S., \& Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron, 63(1), 119-126.
Bromberg-Martin, E. S., \& Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. Nature neuroscience, 14(9), 1209-1216.
Chow, J. J., Smith, A. P., Wilson, A. G., Zentall, T. R., \& Beckmann, J. S. (2016). Suboptimal choice in rats: incentive salience attribution promotes maladaptive decisionmaking. Behavioural Brain Research.
Coulom, R. (2006). Efficient selectivity and backup operators in monte-carlo tree search. In International conference on computers and games (pp. 72-83).
Crockett, M. J., Braams, B. R., Clark, L., Tobler, P. N., Robbins, T. W., \& Kalenscher, T. (2013). Restricting temptations: neural mechanisms of precommitment. Neuron, 79(2), 391-401.
Gipson, C. D., Alessandri, J. J., Miller, H. C., \& Zentall, T. R. (2009). Preference for $50 \%$ reinforcement over $75 \%$ reinforcement by pigeons. Learning \& Behavior, 37(4), 289298.

Green, L., \& Rachlin, H. (1977). Pigeons' preferences for stimulus information: Effects of amount of information1. Journal of the experimental analysis of behavior, 27(2), 255-263.
Hunt, L. T., Rutledge, R. B., Malalasekera, N., Kennerley, S. W., \& Dolan, R. J. (2016). Approach-induced biases in human information sampling. bioRxiv, 047787.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., \& Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput Biol, 7(4), e1002028.
Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J., \& Dayan, P. (2016). The modulation of savouring by prediction error and its effects on choice. eLife, 5, e13747.
Karlsson, N., Loewenstein, G., \& Seppi, D. (2009). The ostrich effect: Selective attention to information. Journal of Risk and uncertainty, 38(2), 95-115.
Lang, P. J., Bradley, M. M., Cuthbert, B. N., et al. (1999). International affective picture system (iaps): Instruction manual and affective ratings. The center for research in psychophysiology, University of Florida.
Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. The Economic Journal, 97(387), 666684.

McDevitt, M. A., Dunn, R. M., Spetch, M. L., \& Ludvig, E. A. (2016). When good news leads to bad choices. Journal of the experimental analysis of behavior, 105(1), 2340.

Niv, Y., \& Chan, S. (2011). On the value of information and other rewards. Nature neuroscience, 14(9), 1095.
Peirce, J. W. (2007). Psychopypsychophysics software in python. Journal of neuroscience methods, 162(1), 8-13.
Prokasy, W. F. (1956). The acquisition of observing responses in the absence of differential external reinforcement. Journal of Comparative and Physiological Psychology, 49(2), 131.
Roper, K. L., \& Zentall, T. R. (1999). Observing behavior in pigeons: The effect of reinforcement probability and response cost using a symmetrical choice procedure. Learning and Motivation, 30(3), 201-220.
Rutledge, R. B., Skandali, N., Dayan, P., \& Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. Proceedings of the National Academy of Sciences, 111(33), 12252-12257.
Silver, D., Sutton, R. S., \& Müller, M. (2008). Sample-based learning and search with permanent and transient memories. In Proceedings of the 25th international conference on machine learning (pp. 968-975).
Spetch, M. L., Belke, T. W., Barnet, R. C., Dunn, R., \& Pierce, W. D. (1990). Suboptimal choice in a percentagereinforcement procedure: Effects of signal condition and terminal-link length. Journal of the experimental analysis of behavior, 53(2), 219-234.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction (Vol. 1) (No. 1). MIT press Cambridge.
Vasconcelos, M., Monteiro, T., \& Kacelnik, A. (2015). Irrational choice and the value of information. Scientific reports, 5.
Wyckoff, L. B. (1952). The role of observing responses in discrimination learning. part i. Psychological review, 59(6), 431.

# Interruptions Reduce Confidence Judgments: Predictions of Three Sequential Sampling Models 

Kevin Zish (kzish@gmu.edu)<br>George Mason University<br>4400 University Dr, Fairfax, VA 22030

Malcolm McCurry (malcolm.mccurry.ctr@nrl.navy.mil)<br>Harris<br>4555 Overlook Ave SW, Washington, DC 20375

Nathan Aguiar (naguiar@masonlive.gmu.edu)
George Mason University
4400 University Dr, Fairfax, VA 22030
J. Gregory Trafton (greg.trafton@nrl.navy.mil)
U.S. Naval Research Laboratory

4555 Overlook Ave SW, Washington, DC 20375


#### Abstract

The relationship between confidence and accuracy has been modeled many times. This paper compares and contrasts three decision-making mathematical models (2DSD, Poisson, RTCON2) of confidence and investigates how each model predicts the effects of interruptions on accuracy, decision response time, confidence, and confidence response time.


Keywords: 2DSD; Poisson model; RTCON2; confidence; accuracy; interruptions; response time; decision-making

## Introduction

In 2016, the U.S. Justice Department released guidelines for law enforcement on how to collect confidence judgments for witness identification ("Justice Department Issues New Guidance On Securing Eyewitness IDs," 2016). These evidence-based guidelines consider the role memory plays in confidence judgments.

Memory researchers have found that the more an item is rehearsed in memory, the more confident a person will be in the accuracy of that retrieval (Busey, Tunnicliff, Loftus, \& Loftus, 2000). For witnesses, repeating a testimony prior to trial is not uncommon. As a result, by the time a trial occurs, the confidence a witness has about their testimony has increased beyond the confidence of their first testimony (Wixted, Mickes, Clark, Gronlund, \& Roediger III, 2015).

Inflated confidence is a concern for the justice system because of the often replicated finding that the relationship between accuracy in memory and confidence is positive (DeSoto \& Roediger, 2014; Dunlosky \& Metcalfe, 2008; Roediger \& DeSoto, 2014; Roediger III \& Desoto, 2012; Roediger III \& DeSoto, 2014; Wixted et al., 2015). Because this positive relationship is often noticed by laypersons, jury's mistake inflated witness confidence for accuracy. The Justice Department encourages law enforcement to guard against inflated confidence by recording confidence during
the first testimony so as to better reflect the accuracy that the testimony happened as described.
.Although this policy change at the U.S. Justice Department is likely to result in higher quality evidence in courtrooms, formally modeling the relationship between confidence and accuracy has been very difficult because it has been hard to determine when confidence judgments begin. There have been many attempts to model the relationship between confidence and accuracy (see Dunlosky \& Metcalfe, 2008; Pleskac \& Busemeyer, 2010; Ratcliff \& Starns, 2013 for a review). It seems intuitive that the process of forming a confidence judgment should begin after some choice has been made. However, Petrusic \& Baranski (2003) showed that when a confidence judgment was required, participant's response times for the primary choice were longer than when the confidence judgment was not required. Petrusic \& Baranski (2003) interpreted this finding to mean that at least some of the processing for a confidence judgment occurs during the primary judgment. As a result, many researchers have attempted to extend previous models of primary choice to account for confidence judgments.

Three of the most popular models to attempt to explain confidence judgments are 2DSD (Pleskac \& Busemeyer, 2010), the Poisson model (Merkle \& Van Zandt, 2006; Van Zandt \& Maldonado-Molina, 2004), and RTCON2 (Ratcliff \& Starns, 2009, 2013). Each model relies on sequential sampling to determine the selection of a choice. Sequential sampling models assume that information is collected from memory or sensory input and summatively translated to evidence towards a particular choice.

Evidence collection through sequential sampling is a common theme across all three models. According to each model, choice is based on the collection of evidence. Evidence is collected until a threshold is reached for one of
the choice alternatives. Crossing a threshold and the subsequent response (e.g key-press) is the decision time for a primary choice. The primary choice is the evidence accumulation and response to questions like "Did you see item ' $A$ ' or item ' $B$ ' before?' Confidence time is the time it takes to make a confidence judgment about the probability that the primary choice was correct. The confidence judgment is a secondary choice in response to questions like "How would you rate your confidence on a Likert scale of 1 to 6 ?" Thresholds are variable and so can be determined in a trial-by-trial and person-to-person basis (Audley, 1960; D. Vickers, 1970; Douglas Vickers, 2014).

Below is a description of each of the three models. The goal is to briefly summarize each model and highlight how each model suggests confidence judgments are calculated from evidence accumulation and how primary choice may relate to confidence judgments (when specified).

## 2DSD

The 2-stage-dynamic-signal-detection theory (2DSD) was first introduced by Pleskac \& Busemeyer in 2010 and suggests that confidence judgments involve post-decision processing of the primary choice. 2DSD is specifically adapted to a 2-alternative forced-choice task (2AFC).

In the 2DSD model, participants make a primary choice by collecting evidence that pushes a single counter towards one choice alternative or the other. When the counter reaches a criterion for one alternative, a primary choice is made. If a decision is prompted before a threshold is passed, the alternative choice that is closest to the counter is selected.

After a primary choice is made, evidence continues to be collected for a confidence judgment. Evidence continues to accrue for the single counter until evidence passes a threshold for a particular confidence judgment and a secondary choice is made. Each possible confidence response has a separate threshold. Similar to primary choice, if a choice is prompted before a threshold is passed; the confidence judgment that is closest to the counter is selected.

According to 2DSD decision time for primary and secondary choice is the product of drift rate. Drift rate is determined by the quality of the evidence collected. Because drift rate towards one alternative or the other determines how fast a choice is made, response time is a function of the quality of the evidence.

2DSD extends the idea that high drift rates lead to fast choices to confidence judgments. The often replicated finding that there is a negative relationship between confidence judgments and response times (Baranski \& Petrusic, 1998; Petrusic \& Baranski, 2003) suggests that lower drift rates will produce lower confidence responses.

## Poisson Model

The Poisson model was introduced by Pike (1971;1973) and modified for confidence by Merkle \& Van Zandt, (2006). The Poisson model assumes that in a 2AFC task,
there is one counter of evidence for each of the possible primary choice alternatives. Counters accrue evidence for each alternative. Whichever counter reaches its respective choice threshold first is the primary choice. If the decision is prompted, whichever counter is closest to the criterion threshold is selected.

After the primary choice is made, evidence collection stops and a secondary choice for confidence is ready to be made. In the Poisson model, confidence is a function of the difference in evidence between counters. If the difference between the collected evidence is large, confidence is high. If the difference between the collected evidence is small, confidence is low.

The idea of confidence being generated by separate counters was made most popular by the balance of evidence hypothesis (D. Vickers, 1970; Douglas Vickers, 2001, 2014). The balance of evidence hypothesis suggests that the difference in evidence between counters is scaled to produce a confidence response. The Poisson model is unclear about whether or not confidence is scaled immediately after a choice is made or a computation is required first.

According to the Poisson model, decision time for the primary choice is the sum of the time to retrieve each piece of evidence and increment the winning counter. The model does not specify confidence as a second choice that is calculated as a post-decision. However, Van Zandt \& Maldonado-Molina (2004) have suggested that additional evidence collection and post-processing could occur after a primary choice but may not always be necessary.

## RTCON2

RTCON2 is a model of making confidence judgments only and not of primary choice (see Ratcliff, 1978 for their diffusion model of primary choice). As a result, it is unclear from RTCON2 how primary choice influences confidence judgments.

RTCON2 suggests that there is a separate counter for each of the possible confidence responses. A confidence judgment is selected when a counter reaches a predefined threshold. According to RTCON2, participants do not have access to the amount of evidence each counter has accrued and therefore can only make a choice when a counter has reached a threshold. In addition, because people do not have access to the evidence, there is no comparison between the amount of evidence for different confidence judgments.

Importantly RTCON2 differs from the original RTCON (Ratcliff \& Starns, 2009) in that each counter for a confidence judgment is affected by the behavior of other counters so as to maintain no net difference. As a result, if evidence facilitates an increment in one counter, the other counters decrease so as to have a net zero effect.

Similar to 2DSD, RTCON2 also uses higher drift rates to explain faster response times and higher confidence responses.

## Memory

One major component of the 2DSD, Poisson model, and RTCON2 models is their reliance on memory for their evidence counters. A sufficient model should be able to predict what happens when memory quality changes.

Many researchers have investigated the relationship between accuracy and confidence by manipulating memory (see Dunlosky \& Metcalfe, 2008 for a review). One wellresearched way of manipulating memory is using interruptions. A long history of research has shown a decrease in task performance (e.g increased response time, decreased accuracy, increased time to return to the task) following an interruption (Altmann \& Trafton, 2007; Altmann, Trafton, \& Hambrick, 2014; Cades, BoehmDavis, Trafton, \& Monk, 2011; Gillie \& Broadbent, 1989; Trafton, Altmann, \& Ratwani, 2011; Trafton, Jacobs, \& Harrison, 2012).

More recently, confidence in the accuracy of a memory has been shown to be lower after an interruption (Aguiar, Zish, McCurry, \& Trafton, 2016; Zish, Hassanzadeh, McCurry, \& Trafton, 2015). In an experiment we replicate these findings and discuss the predictions of each of the three models.

## Model Predictions

All of the models suggest that decision-making is the result of some form of sequential evidence collection. There are two primary differences between each of the models. First, decisions are either the result of the use of one counter for evidence (2DSD) or multiple competing counters (Poisson and RTCON2). Second, confidence judgments are the result of the winning counter choosing the confidence response (2DSD and RTCON2) or the winning counter choosing the time when the delta between multiple counters is used to calculate confidence (Poisson).

The number of counters (one or multiple) and the role of the counter (choosing a response or calculating the delta between counters) results in testable predictions for how interruptions will affect performance.

For all three models accuracy should be lower after an interruption because drift rates will experience more fluctuations when the quality of the evidence collected decreases. Fluctuations in drift rates can result in an error if noise allows for evidence to increment towards the incorrect choice threshold.

Drift rates also drive response time for all three models. Any decrease in the drift rate should increase response time. Because interruptions increase the amount of time to retrieve an item from memory, interruptions should increase decision response times and confidence response times.

As for confidence, 2DSD and RTCON2 predict that confidence decreases whenever drift rate decreases. Although 2DSD relies on one counter and RTCON2 uses multiple counters, both models suggest that confidence is chosen when one counter crosses a threshold for a confidence judgment. Alternatively the Poisson model uses two counters where confidence is the delta between the two
counters. An interruption is likely to slow the increment of both counters equally. Therefore, the Poisson model predicts that confidence should be no different after an interruption trial than a non-interruption trial.

Given the predictions of these three models, the number of counters clearly does not matter when it comes to predicting accuracy or response time. The primary difference between the predictions of the three models is whether or not confidence will be the same after an interruption trial as compared to a non-interruption trial.

## Methods

## Participants

Fifty-five George Mason University undergraduates participated for course credit.

## Tasks

Primary Task The primary task consisted of a simulated stock exchange where participants filled out Buy and Sell orders. Each order had 12 widgets that needed different information about the state of the stock market and the Buy or Sell request (e.g Stock Symbol, Exchange, Transaction Type).

To begin, participants were presented with an autoselected Buy or Sell request at the bottom of the screen (colored gray) and a red arrow designating which of the 12 widgets required information first. The red arrow's location was randomized so that participants learned to start from multiple widgets.

Participants located and selected a "Start" button on the side of the widget designated by the red arrow. Selecting "Start" would teleport the widget to the bottom middle of the screen so that it became the main focus of the task.

Participants would use information from the gray-colored request and the stock market information along the middle of the screen to fill in the widget with the correct information. When the correct information was selected from the widget's dropdown menu, the widget would return to its original place on the screen (Figure 1). Participants repeated the process by finding information for the next widget. Widgets were completed left-to-right and top-down.

A trial ended when the active order was replaced by another auto-selected Buy or Sell request.


Figure 1: Primary task with auto-selected order and widget.
Interruption Task For half of the trials, participants were given a secondary task that served as an interruption. The interruption lasted for 20 -seconds after completing an order. The interruption consisted of a series of addition problems. Addition problems completely occluded the screen until the secondary task was complete. Participants were instructed to complete the addition problems as quickly and as accurately as possible.

Signal Position Question After a trial ended or after a trial and interruption ended, participants were presented with a facsimile of the stock order screen. A blue arrow pointed to one of the 12 widgets with the question: "Is the arrow pointing to the next correct step?" Participants would respond by clicking the word "Yes" in the top left corner or the word "No" in the top right corner (Figure 2). Once the participant made a selection, they were presented with the next order to complete with a new Buy or Sell request.

The placement of the blue arrow was evenly split between the next correct or incorrect step.


Figure 2: Signal Position Question with a blue arrow pointing at a possible next correct step.

Confidence Question Once the signal detection question was complete, the screen was replaced with a question that asked: "How confident are you that the [widget name] was the next correct step?" The participant selected a button on the bottom of the screen that represented their confidence on a scale of 1 through 6 with 1 being "Not at all Confident" and 6 being "Entirely Confident."

## Design

The study was a 2 factor (interruption/non-interruption) repeated measures design.

Each participant had 32 interruptions across 64 trials. The order of screens participants saw was the primary task for 25 completed widgets, a 20 -second secondary task after half of the trials, a signal position question, and a confidence question.

The 64 trials were equally divided between $2,3,4$, and 5 completed widgets in length. The length of the trial was varied to reduce the likelihood that participants could prepare for an interruption and/or signal position question.

Each participant had half of the signal position arrows pointing to the next correct step.

## Procedure

Participants filled out an approved IRB consent form as well as biographical information. Participants were seated approximately 47 cm from the computer monitor. The task was first described using screenshots of the primary and secondary tasks as well as the signal position and question.

Three practice trials were completed that were each 12 widgets long. This was to give the participant the opportunity to experience the order of the widgets before being given partial orders to fill. The experimenter provided the opportunity for participants to ask clarifying questions about the behavior of the task. Participants could begin once the experimenter left the room and were debriefed and dismissed once finished.

## Measures

Behavioral data based on mouse clicks was collected for all participants in addition to screen recordings. Accuracy for identifying the next correct step in the task, response time (RT) for identifying the next correct step in the task, confidence in identifying the next step of the task, and response time for the confidence judgment were calculated.

## Results

Fifty-five participants made 3176 correct responses and 3520 confidence responses.

## Behavioral Results

A within-subjects ANOVA between interruption and noninterruption trials show that interruptions hurt performance metrics for accuracy $(\mathrm{F}(1,54)=132.4, \mathrm{MSE}=0.59, \mathrm{p}<.05$, $\left.\eta^{2}=.56\right)$, confidence $(\mathrm{F}(1,54)=135.7, \mathrm{MSE}=22.813, \mathrm{p}<$ $\left..05, \eta^{2}=.73\right)$, decision response time $(\mathrm{F}(1,54)=134.8$, MSE
$\left.=78,277,425, \mathrm{p}<.05, \eta^{2}=.75\right)$, and confidence response time $\left(\mathrm{F}(1,54)=30.47, \mathrm{MSE}=844,342, \mathrm{p}<.05, \eta^{2}=.11\right)$. A summary of the means of each performance metric across interruption condition can be found in Table 1.

Table 1: Means of performance measures.

| Performance Metric | Interrupted <br> Trials | Non- <br> Interruption <br> Trials |
| :--- | :--- | :--- |
| Accuracy | $82.9 \%$ | $97.5 \%$ |
| Confidence | 4.95 | 5.86 |
| Decision RT (ms) | 4493.96 | 2806.82 |
| Confidence RT (ms) | 1492.30 | 1317.08 |

## Empirical Data and Model Predictions

All three models predicted the decrease in accuracy, the increase in decision response time, and the increase in confidence response time that appeared after an interruption trial compared to a non-interruption trial. Most importantly, only 2DSD and RTCON2 predicted a decrease in confidence after an interruption.

## Discussion

In this paper we describe three sequential-sampling models of decision making and apply their predictions to the results of an experiment.

2DSD used one counter to make a primary choice and then subsequent post-decision processing to provide a confidence judgment once the single counter passed a threshold. The Poisson model used two counters to make a primary choice where the first counter to cross a threshold determines the time when a delta between the two counters is calculated for confidence. Finally, RTCON2 is a multiple counter confidence-only model where the winning counter crosses a threshold that informs the confidence judgment.

We compared the predictions of these three models to the results of an experiment where participants completed a task in a simulated stock market. Participants were interrupted $50 \%$ of the time and then asked to choose if a widget was the next correct step of the task followed by a confidence judgment.

This study replicated the results of many other interruption-based studies in that performance suffered after an interruption. An addition to previous work is that confidence response times are slower following an interruption.

In terms of model predictions, all three models can account for a decrease in accuracy after an interruption. All three models suggest that accuracy is a function of the fluctuations in the drift rate. Counters can reach the threshold for the incorrect choice first when there is more noise in the drift rate, particularly when the thresholds are lowered. To fully explain a decrease in accuracy after an interruption, each model would need to explain why there is
more noise in drift rate and/or lower choice thresholds after an interruption trial than a non-interruption trial.

A strength of each model is a fairly comprehensive explanation for decision response time and confidence response time after an interruption. All three models suggest that response times are a function of the speed of evidence collection. As a result, anything that reduces the speed of evidence collection should reduce response time. While no link has been produced that shows that quality of evidence decreases after an interruption, there is certainly a large amount of literature showing interruptions increase retrieval time (Altmann \& Trafton, 2007; Altmann et al., 2014; Cades et al., 2011; Trafton et al., 2011, 2012). An increase in retrieval time would slow down evidence collection and lead to longer decision and confidence response times.

As for confidence only 2DSD and RTCON2 were able to predict lower confidence judgments after an interruption. In both models confidence decreases when the drift rate is lower. Therefore, interruptions are likely to lower the drift rate.

The Poisson model is unclear about explaining a change in confidence after an interruption. Confidence in the Poisson model is driven by the balance of evidence hypothesis and calculated by scaling the delta between the counters. According to the Poisson model, confidence after an interruption should stay the same as before an interruption because both counters in a 2 AFC would be affected by the interruption equally. Thus, the Poisson model does not predict the difference in confidence found in this study.

This paper compared and contrasted three mathematical models of decision making. The 2DSD, Poisson model, and RTCON2 can explain changes in accuracy with drift rate but do not provide a clear mechanism to explain the effect of interruptions. However, each model does have a fairly robust explanation of decision response times and confidence response times via speed of evidence collection. As for confidence judgments, it is easier for models to explain changes in confidence judgments when there is a single counter that chooses a response than models that use multiple counters to calculate the delta between alternatives.

## Acknowledgments

This work was supported by the Office of Naval Research to JGT. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies of the U. S. Navy.

## References

Aguiar, N., Zish, K., McCurry, J. M., \& Trafton, J. G. (2016). Interruptions Reduce Performance across All Levels of Signal Detection When Estimations of Confidence are Highest. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 60, pp. 254258). SAGE Publications. Retrieved from http://pro.sagepub.com/content/60/1/254.short

Altmann, E. M., \& Trafton, J. G. (2007). Timecourse of recovery from task interruption: Data and a model. Psychonomic Bulletin \& Review, 14(6), 1079-1084.
Altmann, E. M., Trafton, J. G., \& Hambrick, D. Z. (2014). Momentary interruptions can derail the train of thought. Journal of Experimental Psychology: General, 143(1), 215.

Audley, R. J. (1960). A stochastic model for individual choice behavior. Psychological Review, 67(1), 1 .

Baranski, J. V., \& Petrusic, W. M. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. Journal of Experimental Psychology: Human Perception and Performance, 24(3), 929.

Busey, T. A., Tunnicliff, J., Loftus, G. R., \& Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. Psychonomic Bulletin \& Review, 7(1), 26-48.

Cades, D. M., Boehm-Davis, D. A., Trafton, J. G., \& Monk, C. A. (2011). Mitigating disruptive effects of interruptions through training: what needs to be practiced? Journal of Experimental Psychology: Applied, 17(2), 97.
DeSoto, K. A., \& Roediger, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. Psychological Science, 0956797613516149.
Dunlosky, J., \& Metcalfe, J. (2008). Metacognition. Sage Publications. Retrieved from https://books.google.com/books?hl=en\&lr=\&id=eVUXBA $A A Q B A J \& o i=f n d \& p g=P R 7 \& d q=$ dunlosky + and + metcalfe \& ots=pQQ0zHb8sp\&sig=T44q3ZPgczKMKHKEIGsf20Ez2Y

Gillie, T., \& Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. Psychological Research, 50(4), 243-250.
Justice Department Announces Department-Wide Procedures for Eyewitness Identification. (n.d.). Retrieved January 23, 2017, from https://www.justice.gov/opa/pr/justice-department-announces-department-wide-procedures-eyewitnessidentification

Merkle, E. C., \& Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. Journal of Experimental Psychology: General, 135(3), 391.

Petrusic, W. M., \& Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. Psychonomic Bulletin \& Review, 10(1), 177183.

Pike, A. R. (1971). The latencies of correct and incorrect responses in discrimination and detection tasks: Their interpretation in terms of a model based on simple counting. Attention, Perception, \& Psychophysics, 9(6), 455-460.

Pike, R. (1973). Response latency models for signal detection. Psychological Review, $80(1), 53$.
Pleskac, T. J., \& Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychological Review, 117(3), 864.
Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59.

Ratcliff, R., \& Starns, J. J. (2009). Modeling confidence and response time in recognition memory. Psychological Review, 116(1), 59.

Ratcliff, R., \& Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. Psychological Review, 120(3), 697.
Roediger, H. L., \& DeSoto, K. A. (2014). Understanding the relation between confidence and accuracy in reports from memory. Remembering: Attributions, Processes, and Control in Human Memory: Essays in Honor of Larry Jacoby, 347-367.

Roediger III, H. L., \& Desoto, K. A. (2012). The Curious Complexity between Confidence and Accuracy in Reports from Memory. Memory and Law, 84.
Roediger III, H. L., \& DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. Memory, 22(1), 76-91.
Trafton, J. G., Altmann, E. M., \& Ratwani, R. M. (2011). A memory for goals model of sequence errors. Cognitive Systems Research, 12(2), 134-143.

Trafton, J. G., Jacobs, A., \& Harrison, A. M. (2012). Building and verifying a predictive model of interruption resumption. Proceedings of the IEEE, 100(3), 648-659.
Van Zandt, T., \& Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(6), 1147.
Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. Ergonomics, 13(1), 37-58.

Vickers, Douglas. (2001). Where Does the Balance of Evidence Lie with Respect to Confidence? na. Retrieved from
https://pdfs.semanticscholar.org/d16d/ba06d352494962e03c 1c00bed1c 1944bd7f4.pdf

Vickers, Douglas. (2014). Decision processes in visual perception. Academic Press. Retrieved from https://books.google.com/books?hl=en\&lr=\&id=LXA-
AwAAQBAJ\&oi=fnd\&pg=PP1\&dq=Vickers,+D.+(1979).+ Decision+processes+in+visual+perception.+New+York:+A cademic+Press.\&ots=KJ1kVOIGDJ\&sig=vx20Bf_g8X_4y W10c2DuWXse510

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., \& Roediger III, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. American Psychologist, 70(6), 515.
Zish, K., Hassanzadeh, S., McCurry, J. M., \& Trafton, J. G. (2015). Interruptions can Change the Perceived Relationship between Accuracy and Confidence. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 59, pp. 230-234). SAGE Publications. Retrieved from http://pro.sagepub.com/content/59/1/230.short

# Simulation and heuristics in flexible tool use 

Kelsey Allen<br>Massachusetts Institute of Technology, Cambridge, MA, US

Kevin Smith
Massachusetts Institute of Technology
Josh Tenenbaum
Massachusetts Institute of Technology


#### Abstract

Humans are remarkably flexible tool users. We not only recognize a wide range of existing tools, but also produce new tools by seeing objects in new ways, or by making or repurposing objects to solve a problem confronting us. Here we study the cognitive processes supporting flexible tool use, including deciding what makes a good tool, and how it should be used. Participants played a video game which requires selecting an object from a set of options and placing it in a virtual physical scene in order to accomplish goals such as tipping another object over or launching it into a container. People appear to use a combination of simulation-based planning and experience-based heuristics: fast heuristics drive the initial selection and placement of a candidate tool, and that solution can then be refined by several rounds of mental simulation interspersed with trial-and-error experimentation to rapidly converge on goal-satisfying solutions.


# Predicting Preschool-Aged Children's Behavior Regulation from Attention Tasks in the Lab 

Chelsea Andrews<br>University of Wisconsin Madison<br>Emily Coates<br>Ripon College<br>Kristine Kovack-Lesh<br>Ripon College<br>Vanessa Simmering<br>University of Wisconsin - Madison


#### Abstract

One challenge in studying cognition over the lifespan is designing tasks that measure the same construct in different age groups and relate reliably to real-world outcomes. The current study confronts this challenge by testing a new paradigm to assess attention in preschool-aged children for comparison with other measures. Children completed the new "Pop-the-Bubbles" paradigm plus Flanker and Visual Search tasks, for comparison with parental reports of behavioral regulation. Correlations between behavioral regulation and measures from both Flanker and Pop-the-Bubbles suggest that children's ability to ignore irrelevant stimuli in these lab tasks relates to their ability to behave appropriately in everyday situations. Further development of Pop-the-Bubbles for eye-tracking and a color version of Flanker are underway to test these relationships more extensively in young children.


# Why do we punish negligent behaviors? 

Sarin Arunima

Harvard University

Cushman Fiery<br>Harvard University


#### Abstract

Prior research suggests that negligent harms are punished because of the resulting negative outcomes. Under this account, negligent but completely harmless acts should not be punished. An alternative possibility is that negligence is punished as a way of modifying future thought and behavior. Across three studies we find support for this second proposal. Study 1 demonstrates that punishment is assigned to negligent agents, irrespective of whether or not a harm actually occurs. Study 2 demonstrates that non-negligent agents who cause harm are punished less than negligent agents who do not cause harm. Study 3 shows that the punishment of harmful negligent actions is only judged to be successful when it results in the agent ceasing to act negligently, and not when it results in the harm ceasing to occur. Together, these results suggest that a primary function of punishment in cases of negligence is modify future thought.


# When I say 'Black Lives Matter,' why do some people hear 'Others Lives Matter Less'? 

E. Malemma Azumah<br>Agnes Scott College

Jason Shepard<br>Agnes Scott College


#### Abstract

The statement "Black lives matter" is commonly construed as implying other lives matter less, even though the statement does not explicitly reference other lives. Bias is a common explanation for this construal. However, other factors may contribute. We hypothesized that the linguistic structure of "Black lives matter" plays an important role. "Black lives matter" takes the form of a generic, or statements in which a property is attributed to members of a set (e.g., "lions have manes"). Generics are often interpreted as implicit comparisons (e.g., "lions are more likely to have manes than other animals"). We report two experiments in which we find evidence that the statement "Black lives matter" is often construed as an implicit comparative claim, similarly to other generics. This research contributes to our understanding of generics, while providing a novel explanation for why when I say "Black lives matter," some people hear "Other lives matter less."


# The Use of Ambiguous Messages as a Strategy to Appeal to Multiple Decision Makers 

Devika Banerji<br>Purdue University, West Lafayette, IN, United States<br>Torsten Reimer<br>Purdue University, West Lafayette, IN, United States<br>Chris Roland<br>Purdue University, West Lafayette, IN, United States


#### Abstract

Messages that are tailored to specific audiences (matched messages) are typically more persuasive compared to messages that are crafted for a general audience (Hirsh, Kang, \& Bodenhauser, 2012). However, tailoring messages can have the effect that messages are less persuasive for audiences for which they were not tailored (mismatched messages; Sillince, Jarzabkowski, \& Shaw, 2012). Eisenberg (1984) introduced the concept of strategic ambiguity to appeal to multiple audiences simultaneously. We systematically compared effects of matched/mismatched tailored messages with the effects of ambiguous messages on multiple-criteria choice behavior. We found evidence that ambiguous messages can be used under certain conditions to simultaneously appeal to multiple audiences within the context of credit card choices. Using the financial control typology developed by Shefrin and Nicols (2014) to define different audiences, the study ( 154 participants) provided some support for the use of ambiguity as a tool for tailoring messages to diverse credit-card holders.


# Social wayfinding in complex environments 

Iva Barisic<br>ETH Zurich, Zurich, Switzerland

Tyler Thrash
ETH Zurich, Zurich, Switzerland
Victor R. Schinazi
ETH Zurich, Zurich, Switzerland

## Christoph Hoelscher

ETH Zurich, Zurich, Switzerland


#### Abstract

Wayfinders in a group can be influenced by various factors, including other group members and environmental structure, but social wayfinding is an underexplored topic. This experiment investigated differences in wayfinding decisions between individuals and groups and their dependence on environmental structure. Participants navigated through a train station with or without market stalls, either as individuals or as groups. There was a significant main effect of environmental structure on task efficiency, and an inconclusive interaction between environmental structure and group membership on task efficiency $(\mathrm{p}=0.05)$. Because of heterogeneity of variance, we conducted targeted t -tests. T-tests revealed that groups were slower than individuals with market stalls $(\mathrm{p}=0.02)$ but not without $(\mathrm{p}=0.91)$. There was significant main effect of the environmental structure on number of turns. The main effect of group membership on number of turns and the interaction were not significant. We will analyze walked and Levenstein distance as wayfinding efficiency indicators.


# Projecting space into the future: peripersonal space remaps in anticipation of an object manipulation 

Anna Belardinelli<br>University of Tübingen<br>Johannes Lohmann<br>University of Tübingen

Martin Butz<br>University of Tübingen


#### Abstract

Manipulation planning relies on anticipatory processes, aimed at achieving the desired goal state, such as a grasp. This implies that peripersonal space is remapped to the anticipated grasp posture on the targeted object. Vibrotactile-visual interactions were probed at different times during a grasp-and-place task. Thumb or index finger were stimulated concurrently with a visual distractor on the to-be-grasped object. Object orientation (upright/upside down) afforded a thumb-up or thumbdown grasp, inverting the congruency between haptic and visual stimulation. Response times about which finger was stimulated show the expected crossmodal congruency effect already before motion onset, with shorter times when the visual distractor and the future position of the stimulated finger overlapped. Moreover, eye-tracking data show that the tactile stimulation influences the gaze in anticipation of the upcoming grasp. Thus peripersonal hand space is mapped into the future, predictively mediating between tactile and visual perceptions as a function of the final state.


# Characterizing Human-Machine Teams with Process Algebras 

Leslie Blaha<br>Pacific Northwest National Laboratory, Richland, Washington, USA<br>Robert Jasper<br>Pacific Northwest National Laboratory, Richland, Washington, USA


#### Abstract

We conceptualize human-machine (computer, robot) teams as concurrent processes. Such a conceptualization means: (1) the human and machine agents have a common goal or mission; (2) each agent may have different subtasks within the goal space; (3) they do not have a shared memory, but (4) they do have a means of communicating with each other. Process algebras, such as communicating sequential processes (Hoare, 1977), are formal languages for describing the ways in which two concurrent processes interact through message passing across information channels. In this research, we enumerate the ways in which human-machine interactions can be structured, such as strictly serial, parallel, and cascade-like architectures. We use process algebras to characterize the interactions in candidate architectures. We discuss design implications for active and interactive machine learning systems.


# Bottom-up attentional cueing in category learning in children 

Nathaniel Blanco<br>Ohio State University<br>Vladimir Sloutsky<br>Ohio State University


#### Abstract

Young children tend to differ from adults in how they learn new categories. In comparison to adults (who rely on selective attention and tend to form explicit rules), children distribute attention widely, forming similarity-based category representations. But, when attention is explicitly directed toward the rule with top-down feedback, children exhibit rulebased classification-though memory performance still indicates distributed attention. Little is known, however, how bottom-up attentional cueing affects the category representations that children form. In our experiment 4 -year-olds learned to classify alien creatures composed of binary features. A single "deterministic" feature perfectly predicted category membership, while other features were probabilistically predictive. We manipulated the saliency of the deterministic feature, making it grow and shrink. This manipulation was remarkably effective at facilitating category learning and rule-based classification, but recognition memory still showed evidence of distributed attention. These results help elucidate the important role of attentional processes in the development of categorization.


# Infusing Cognitive Science Content in Teacher Preparation 

Julie Booth<br>Temple University


#### Abstract

Surprisingly, foundational knowledge about cognitive science (CS) is not included in all teacher preparation programs or required for certification in all states. Here, I examine the impact of infusing CS content into teachercandidates' coursework by providing half of the pedagogy instructors with professional development on big ideas in CS (memory/attention/transfer/problem solving/practice/expertise) and encouraging them to use the materials to deliver mini-lectures on these topics and discuss their relevance to instructional practice. Control instructors did not receive PD or CS materials. In both experimental and control classes, CS knowledge was measured at the beginning and end of the semester; we also collected lesson plans where teacher-candidates explained their reasoning for each instructional decision. We saw no CS knowledge improvement, but teacher-candidates exposed to CS reduced their use of folk reasons (e.g., buzz words such as learning styles, concrete thinking, active learning, etc.) when planning lessons compared with peers in control classes.


# Using Analogical Processing to Categorize Musical Patterns 

Janet Bourne<br>Bates College, Lewiston, Maine, USA<br>Elliot Chun<br>Bates College, Lewiston, Maine, USA


#### Abstract

Participants often categorize musical melodies ("themes") based on perceptual features (e.g. loudness, fastness) instead of structural or relational features (e.g. pitch, rhythm) (Lamont \& Dibben, 2001; Ziv \& Eitan, 2007). In the present study, we investigate whether within-category analogical comparison (Markman \& Gentner, 1993) influences participants to categorize musical themes based on relational features, a prediction from structure-mapping theory (Gentner, 1983). Participants completed a forced-choice triad task where they had to choose whether one theme (relational match) or another (perceptual match) best fit the target theme. In a "no-compare" condition (between-subjects), participants heard one target theme. In a "compare" condition, participants heard and compared two target themes. Initial results indicate that participants who compared two themes chose more relational matches. We found this result for Western Classical themes and popular music chord progressions. These results and their implications are discussed with respects to analogical processing and musical categorization.


# How infants map nonce phrases to scenes with objects and predicates. 

Angelica Buerkin-Pontrelli<br>University of Pennsylvania<br>Daniel Swingley<br>University of Pennsylvania


#### Abstract

When infants hear sentences containing unfamiliar words, are some language-world links (such as noun-object) more readily formed than others (verb-predicate)? What if the context renders verb-predicate and noun-object interpretations equally plausible? We examined 14-15-month-olds' capacity for linking semantic elements of scenes with simple bisyllabic nonce utterances. Each syllable either referred to the object, or the object's motion. Infants heard the syllables in either a VSor SV-consistent order. Learning was tested using 2AFC language-guided looking. Infants learned the nouns and verbs equally well, showing no bias favoring nouns. In all conditions, infants learned the meaning of the utterance-final syllable, but not the initial one, suggesting that noun or verb biases played a smaller role than utterance position when noun- and verb-learning were equally supported by context.


# No Tranfer of Training in Simple Addition 

Jamie Campbell

University of Saskatchewan, Saskatoon, SK, Canada

Yalin Chen
University of Saskatchewan, Saskatoon, SK, Canada


#### Abstract

Several researchers have proposed that skilled adults may solve single-digit addition problems (e.g. $3+1=4,4+3=7$ ) using a fast counting procedure. Practicing a procedure often leads to transfer of learning and faster performance of unpracticed items. Such transfer has been demonstrated using a counting-based alphabet arithmetic task (e.g., B+4 = C D E F) that indicated robust RT gains when untrained transfer problems at test had been implicitly practiced (e.g., practice $\mathrm{B}+3$, test $\mathrm{B}+2$ or $\mathrm{B}+1$ ). Here we constructed analogous simple addition problems (practice $4+3$, test $4+2$ or $4+1$ ). In three experiments ( $n=108$ ) there was no evidence of generalization for these items, but there was robust speed up when the items were repeated. The results are consistent with direct retrieval of addition facts from long-term memory rather than a counting procedure.


# Vowel Harmony as a Distributional Learning Problem 

## Spencer Caplan

University of Pennsylvania, Philadelphia, Pennsylvania, United States
Jordan Kodner
University of Pennsylvania, Philadelphia, Pennsylvania, United States


#### Abstract

Vowel harmony is a class of phonotactic restrictions in which vowels in a language are divided into two or more subclasses, and words must contain only vowels from only one such subclass regardless of intervening consonants. Languages worldwide (Turkish, Finnish, Mongolian, Warlpiri, but not English) exhibit vowel harmony. The opacity of such potentially long distance alternations poses a challenge for the learner. Nevertheless, infants are sensitive to vowel harmony alternations at as young as seven months. We present a computational model for vowel harmony acquisition. By normalizing transitional probabilities over the vowel tier, and making minimal assumptions about the phonology, we successfully determine which test languages have harmony processes and correctly categorize their vowels into harmonizing classes. Using universal typological patterns to inform the search space, we find that phenomena which appear opaque can be captured by simple distributional learning.


# A comparative assessment of embodied and computational topic extraction 

James Carney<br>Department of Psychology, Lancaster University, Lancaster, Lancashire, United Kingdom


#### Abstract

Word embedding algorithms like word2vec (Mikolov et al., 2013) have enabled advances in topic modelling by training shallow neural networks on the co-occurrence of words in corpuses of sentences. However, it is not clear how this process reflects human cognition. This poster will compare the results of document classification using the word2vec skipgram model and the 20k sensorimotor word norms collected by the presenter and colleagues (Lynott \& Connell 2013; Carney et al., in prep.) (These latter norms establish how concepts are processed by way of perceptual and motor schemes, and thus offer a useful proxy for human conceptual classification.) The results of the comparison will generate insights into the different ways in which higher-order concepts are inferred, and allow systematic biases in concept formation to be identified. It will also allow for machine learning processes to be finessed so as to more accurately reflect human-level modes of cognition.


# Mapping hand to world; Development of iconic representation in gesture and homesign 

Erica Cartmill<br>University of California Los Angeles

Lilia Rissman
University of Chicago
Miriam Novack
Northwestern University
Susan Goldin-Meadow
University of Chicago


#### Abstract

In both gesture and sign, objects and events can be represented by reproducing some of their features iconically. Iconic gestures do not typically appear until well into children's second year of life, suggesting that the cognitive and/or communicative resources required are not trivial. Here we investigate how manual iconicity develops in two different communicative systems. Using longitudinal video corpora, we compare the emergence of manual iconicity in 52 hearing children learning a spoken language (co-speech gesture) to a deaf child creating a manual communication system (homesign). We focus on the shape of the hand, asking how handshape use changes between age 1 and 5 , and how handshape choice relates to semantic content. We find broadly similar patterns of handshape development in co-speech gesture and homesign. This suggests that the cognitive building blocks underlying children's ability to iconically map forms to meanings are shared across vastly different communicative systems.


# One-shot word learning under high and low sentential constraints in adult L2 learners of Chinese 

Jenn-Yeu Chen<br>National Taiwan Normal University<br>Yu-Shu Huang<br>National Taiwan Normal University


#### Abstract

New words were embedded in high- and low-constraint sentences and presented three times in a random order to adult learners of Chinese as a second language. The learners explained the meaning of each word in their native languages and their answers were scored by other native speakers. The overall accuracy was .47 with no effect of constraint or frequency. When the data were limited to those words the learners reported having no prior encounters with and those sentences they reported comprehending, the accuracy was .45 . The results demonstrated fast mapping of word learning in adult L2 learners but indicated that extended mapping was necessary to achieve ultimate attainment. The results are also consistent with Krashen's (1982) "comprehensible input" and " $i+1$ " hypothesis.


# Comparing Human Use of Fast \& Frugal Tree with Machine-Learning Tree 

Yee Siang Chng<br>UNIVERSITY COLLEGE LONDON, LONDON, LONDON, United Kingdom


#### Abstract

Previous studies have shown that the predictive accuracy of fast and frugal decision trees (FFTs) is comparable to decision trees generated by machine-learning (Martignon et al., 2008). FFTs are thought to be useful decision tools that are cognitively plausible to internalise, as opposed to complex machine-learning algorithms. Nonetheless, there seems to be a lack of behavioural studies in the literature to support such a claim. In this between-group experiment, we examined the human use of an FFT versus a C 4.5 algorithm tree when completing a car evaluation task. Participants had to learn the rules of their given tree before making evaluations based on their memory. Preliminary results show that FFTs may indeed be easier to use, even when the number of cues for both trees are the same. Interestingly, participants who were successful in using the C 4.5 tree exhibited tree pruning strategies, resulting in a heuristic similar to an FFT.


# Incorrect responses salience affects strategy use in a figural analogy task 

Iwona Ciechanowska<br>Jagiellonian University, Krakow<br>Bartlomiej Kroczek<br>Jagiellonian University, Krakow<br>Adam Chuderski<br>Jagiellonian University, Krakow


#### Abstract

Previous studies of multiple-choice analogy problems suggested that some people use a more efficient but also harder constructive strategy (they build the complete representation of analogy), whereas others tend to use a less effective but simpler response elimination. We tested whether salience of incorrect options (five per figural analogy problem) affected strategy use. Salient options in 18 problems missed many features from the (sixth) correct option; options in 18 non-salient problems missed only few features. When controlling for working memory capacity, eye tracking yielded strongly correlating patterns of data that suggested, in line with previous reports, large individual variance in strategy use. However, participants overall spent $50 \%$ less time analyzing salient than non-salient options, suggesting that salience promoted the constructive strategy. This conclusion was supported by pupil size significantly predicting accuracy on problems with salient options, but not on those with non-salient options (which additionally yielded lower accuracy).


# Learning Object Names from Visual Pervasiveness: the Visual Statistics Predict 

Elizabeth Clerkin<br>Indiana University, Bloomington, IN, United States

Chen Yu<br>Indiana University

Linda Smith<br>Indiana University


#### Abstract

Recent analysis of a corpus of infant-perspective head-camera images found an extremely right-skewed frequency distribution of objects present in 8 - to 10 -month-old infants' visual environments (Clerkin, et al., 2017). Furthermore, the objects most pervasively present in these scenes have names normatively acquired first by learners of English. New analyses show that the names for these objects occur only sparsely in infants' environments, and object name frequency is not correlated with object visual frequency. Therefore, we designed a simple associative model simulating word-object co-occurrence in order to investigate how visual pervasiveness without high-frequency naming could lead to learning of word-object correspondences. With random sampling from distributions reflecting the actual frequency of words and objects in infants' environments, we find that the most frequent objects have a distinct advantage over less frequent objects in their conditional probability. This suggests visual experience with objects may be the principal predictor of early word-referent learning.


# Relations Between Intuitive Biological Thought and Scientific Misconceptions 

John Coley<br>Northeastern University

Nicole Betz
Northeastern University
Jessica Leffers
Northeastern University

Yian Xu<br>Northeastern University

Michal Fux<br>Northeastern University

Kristin de Nesnera
San Francisco State University
Kimberly Tanner
San Francisco State University


#### Abstract

Students enter educational settings with complex and well-established intuitive conceptual understandings of the world, which have important educational consequences. In biology, intuitive thinking can be characterized in terms of cognitive construals (anthropocentric, teleological, and essentialist thinking, Coley \& Tanner, 2015). We examined relations between intuitive thinking and biological misconceptions, and how formal biology education might influence such relations. 137 biology and non-science majors completed measures of anthropocentric, teleological, and essentialist thinking, and indicated agreement/disagreement with common misconceptions and explained their responses. Teleological thinking (but not anthropocentric or essentialist thinking) predicted teleological misconceptions. Anthropocentric and teleological thinking (but not essentialist thinking) predicted anthropocentric misconceptions. Agreement with essentialist misconceptions was unrelated to intuitive thinking. Similar patterns for non-majors and majors suggests formal biology education may have little influence on relations between intuitive reasoning and misconceptions. These findings demonstrate a clear impact of intuitive thinking on learning biology at the university level.


# A picture falls under many categories: How ancient mathematical marks became extinct 

Peter Coppin<br>Perceptual Artifacts Lab, Faculty of Design, OCAD University and the Department of Mechanical and Industrial Engineering, University of Toronto<br>Daemon Retren<br>Perceptual Artifacts Lab, Faculty of Design, OCAD University

Ambrose Li<br>Perceptual Artifacts Lab, OCAD University


#### Abstract

The development of mathematical marking conventions from prehistory to the present is characterized by a trend from conventions with more iconic relationships to concrete structures of the physical world (such as more pictorial ancient land surveying marks) to marking systems with less-iconic relationships to physical structures (that represent numbers, operations, infinity, and other more abstract concepts). We propose how certain constraints of perception-cognition induced conventions that made more-iconic (pictorial) marks controversial. These became too conceptually ambiguous to convey more abstract conceptual categories during the formalization of mathematics: Iconic properties of ancient proto-mathematical conventions recruited lower level perceptual capabilities developed to perceive-act in a concrete world of occluded surfaces-edges and were suitable for conveying concrete structures (such as landforms during surveying). However, these were too conceptually ambiguous to convey more abstract conceptual categories that emerged when mathematics was formalized because a (pictured) concrete structure can fall under many possible conceptual categories


# ANCHORING is amodal: evidence from a signed language. 

Qatherine Dana<br>Dep. of Psychology, Northeastern University, Boston, MA, United States<br>Diane Brentari<br>Dep. of Linguistics, University of Chicago<br>Outi Bat-EI<br>Dep. of Linguistics, Tel-Aviv University<br>Iris Berent<br>Dep. of Psychology, Northeastern University


#### Abstract

Modern linguistic theory posits the existence of universal constraints. But whether these constraints concern language structure, generally, or speech, specifically, is unknown. To address this question, here we ask whether the constraints identified in spoken languages transfer to sign languages. ANCHORING (McCarthy \& Prince, 1993) is a putatively universal constraint on reduplication. ANCHORING requires that the final element of a suffixed reduplicant match the final element of the base (e.g., pana 'chase' —>panana, 'run' not panapa). Here, we examine whether ANCHORING is likewise operative in a signed language. In our experiments, native ASL signers rated novel reduplicated forms: either ones consistent or inconsistent with ANCHORING (i.e., ABB vs. ABA , where A and B are syllables). Results showed that signers reliably favored ABB forms over ABA. These findings show for the first time that ANCHORING constrains a sign language. This conclusion is consistent with the existence of amodal linguistic principles.


# Don't forget to bind: Memory binding and interference in development 

Kevin Darby<br>The Ohio State University, Columbus, OH, USA<br>Vladimir Sloutsky<br>The Ohio State University, Columbus, OH, USA


#### Abstract

This work investigates the development and causes of memory interference effects. Specifically, we measured proactive and retroactive interference effects in children and adults when learning multiple sets of contingencies, as well as individuals' memory binding for the same contingencies. We measured proactive interference by examining memory for a second set of contingencies after learning a first set, and retroactive interference by examining memory for the first set of contingencies after learning the second set. We measured memory binding by presenting participants with partial information about each contingency and measuring their accuracy and pattern of errors when asked to identify the completed contingency. Results indicate that both children and adults experienced substantial interference effects, but children were more prone to interference and substantially worse at memory binding. Additionally, individuals' memory binding abilities were predictive of the magnitude of interference effects, suggesting that memory binding is an important mechanism modulating memory interference.


# Developmental Changes in Visual Scene Statistics 

Christina Deserio<br>Indiana University, Bloomington, IN, United States<br>Jason Gold<br>Indiana University, Bloomington, IN, United States<br>Swapnaa Jayaraman<br>Indiana University, Bloomington, IN, United States<br>Rowan Candy<br>Indiana University, Bloomington, IN, United States<br>Linda Smith<br>Indiana University, Bloomington, IN, United States


#### Abstract

Mature visual experience is tuned by inputs to the developing visual system. However, little is known about the low-level statistics of available visual input as infants interact with the world in rapidly changing ways. Recent studies of the contents of infant-perspective scenes (sampled from a corpus of over 5 million head camera images) indicate that these contents change dramatically over the first year of life. Faces, ceilings, wall edges, and high-contrast patterns characterize younger babies (below 3 months), while more crowded images characterize older babies. These differences suggest possible developmental changes in lower-level visual statistics. After analyzing a sample of infant-perspective scenes from 4- to 10 -week-old infants, and from 28- to 34-week-old infants, we found that mean Feature Congestion and Subband Entropy-measures of visual clutter in natural scenes-increase with age. The full analyses include spatial frequency, orientation, contrast, and clutter measures across $1,821,021$ frames.


# Speed and accuracy trade-off of semantic composition involving highlighting and adjustment 

Diana Dimitrova<br>University of Cologne, Cologne, Germany<br>Brian McElree<br>New York University, New York City, USA<br>Petra Schumacher<br>University of Cologne, Cologne, Germany


#### Abstract

In a Speed-Accuracy-Tradeoff (SAT) paradigm we investigated how adjective type and polarity modulate the online semantic composition of noun phrases (NPs). 22 German speakers read sentences like "The tradesman - buys - a real diamond". Enriched adjectives ("real/fake") highlighted or adjusted the noun's meaning, whereas non-enriched adjectives ("white/flawed") simply specified a property. Adjectives had positive ("white/real") or negative polarity ("flawed/fake"). Upon the display of critical NPs, participants indicated by a series of key presses if the sentence was correct. For the SAT response function we computed the (i) asymptote (response accuracy as d'), (ii) rate (response speed) and (iii) intercept (point when accuracy departs from chance). Accuracy was significantly lower for semantically enriched vs. non-enriched NPs, suggesting that highlighting and adjusting certain properties during composition is costly. Polarity affected temporal dynamics with negative NPs showing a slower rate than positive NPs, indicating that negative information is processed in more depth.


# A Neurodynamical Model of How Prior Knowledge Influences Visual Perception 

Dražen Domijan<br>University of Rijeka

Mateja Marić<br>University of Rijeka


#### Abstract

Recent behavioral studies showed that prior knowledge can directly influence visual perception. In the current work, we offer an explanation of the observed findings based on the adaptive resonance theory (ART). The ART neural network was designed to solve the problem of catastrophic forgetting during learning in non-stationary environment. In the ART, stability of learning is achieved by matching bottom-up sensory signals with top-down expectations. Resonant state that corresponds with conscious perception develops in the network when the bottom-up and top-down signals are closely aligned. On the other hand, mismatch produces global reset signal that clears the traces of erroneous top-down expectations. Therefore, prior knowledge can influence conscious perception only when it already closely matches with sensory signals. We performed computer simulations with real-time implementation of the ART circuit that confirm our analysis. Simulations also showed how observed behavioral findings arise from response bias.


# Couples Emotion Dynamics During Conversations Involving Stress and Enjoyment 

Nicholas Duran<br>Arizona State University

Ashley Randall<br>Arizona State University


#### Abstract

While conversing face-to-face, romantic partners are thought to form a coupled and co-regulatory system, unintentionally shaping each other's emotional states on a moment-by-moment basis. What has been less explored, however, are the ways in which this coupling is modulated by high-level interpersonal factors, such as discussing topics that are stressful for one or both partners. We provide an initial exploration by examining the emotion ratings of 42 romantic, heterosexual couples during conversations involving stress or enjoyment. Ratings were generated via continuous dials (sampled every second) as participants watched video playback of their interactions. The resulting time series were assessed for time-lagged patterns of emotional coupling using cross recurrence quantification analysis. Initial results show that for topics that involved a mutual sense of stress or enjoyment, overall coupling was high, but this coupling was largely disrupted once the stress was more asymmetrically experienced.


# Information Signatures in Children's Language Environment 

Steven L. Elmlinger<br>Cornell University

Drew H. Abney<br>Indiana University

David W. Vinson
University of California, Merced
Linda B. Smith
Indiana University

Chen Yu<br>Indiana University


#### Abstract

In auditory statistical learning, children are sensitive to the transitional structure of their language environment. Variability and stability of utterances in the language environment are important properties of statistical learning but are currently understudied across laboratory and naturalistic research contexts. In this study, we quantify variability and stability in the language environment of children as measured by amount of information within the temporal structure of caregivers' utterances. In this work we present a new method for understanding information signatures in the temporal structure of parent-child free play contexts and document information signatures of caregiver utterances at multiple timescales. Our results suggest information signatures of parental utterances increase across development (9-24 months), but decrease within individual play sessions (5-6 minutes). We speculate that the dynamics of information signatures varies across multiple timescales. Possible implications of the observed information signatures inherent in caregivers' naming of objects to their young children are explored.


# The acquisition of verb morphology in Polish and Finnish: Model and experiment 

Felix Engelmann<br>University of Manchester, Manchester, United Kingdom<br>Joanna Kolak<br>University of Manchester, Manchester, United Kingdom<br>Sonia Granlund<br>University of Liverpool, Liverpool, United Kingdom<br>Ben Ambridge<br>University of Liverpool, Liverpool, United Kingdom<br>Julian Pine<br>University of Liverpool, Liverpool, United Kingdom<br>Anna Theakston<br>University of Manchester, Manchester, United Kingdom<br>Elena Lieven<br>University of Manchester, Manchester, United Kingdom


#### Abstract

Usage-based approaches suggest that language acquisition is a function of the statistical properties of the input. We compare predictions from neural network models with results of two elicited-production experiments on verb inflection with children in the morphologically complex languages Polish and Finnish. Three-layer neural networks were trained to produce person/number-inflected present-tense verb forms in Polish and Finnish from phoneme representations of verb stems using frequency information from child-directed speech corpora. Simulated acquisition in both languages was affected by token frequency and phonological neighbourhood density (PND) as well as an interaction such that low-frequency forms benefited more from PND than high-frequency forms. Suffix errors showed overgeneralisation and substitutions of low-frequency forms with higher-frequency forms. The model predictions are consistent with our empirical findings, except for the frequency X PND interaction. We discuss the experimental and simulated data and their implications.


# What makes a joke funny: Analysing joke humor through single-word ratings. 

Tomas Engelthaler

University of Warwick

## Thomas Hills

University of Warwick


#### Abstract

The appreciation of humor is a universal phenomenon and a key aspect of cognition. It has been studied in the context of jokes, where the incongruity in expected and observed context results in the perception of humor. The present study examines how the humor appreciation of single words relates to the humor of the whole joke - is a joke simply a sum of its parts? Using a novel dataset of single-word humor ratings, collections of jokes from the JESTER database were analyzed. A multiple regression analysis showed joke length and individual word arousal were the best predictors of joke funniness. Longer jokes with fewer individually arousing words were found funnier. Individual word humor did not contribute to the humor of the overall joke. These findings suggest the cognitive aspects of humor are likely driven by broader semantic context, whereas appreciating humor on a per-word basis links to separate factors.


# What's on your wandering mind? The content of mind wandering during textand film comprehension 

Myrthe Faber<br>University of Notre Dame, Notre Dame, Indiana, United States<br>Sidney Dmello<br>University of Notre Dame


#### Abstract

What do we think about when our mind wanders? We asked 88 students to read an instructional text and watch a film (each 20 minutes) and report whenever they found themselves zoning out. Each time they did, we asked them to report their thoughts and what, if anything, triggered them. We then categorized these thoughts ( 1208 in total) based on their content, and found that in contrast with previous studies, only $17 \%$ involved prospection whereas $33 \%$ consisted of autobiographical and semantic memory retrieval. This discrepancy might be driven by the rich content of stimuli: $71 \%$ of autobiographical and semantic retrieval was explicitly triggered by the text or film, compared to $28 \%$ of prospection. Latent semantic analysis revealed that memories were more similar to their triggers than prospective thoughts, suggesting that a substantial proportion of mind wandering is driven by the content of our environment.


# Interaction with a robot changes human motor behavior 

Laura Fademrecht<br>Max Planck Institute for Biological Cybernetics, Tübingen, Germany<br>Tobias Meilinger<br>Max Planck Institute for Biological Cybernetics, Tübingen, Germany<br>Stephan Streuber<br>École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland<br>Aurelie Saulton<br>Max Planck Institute for Biological Cybernetics, Tübingen, Germany<br>Heinrich Bülthoff<br>Max Planck Institute for Biological Cybernetics, Tübingen, Germany<br>Rouwen Cañal-Bruland<br>Friedrich-Schiller-University Jena, Jena, Germany<br>Stephan de La Rosa<br>Max Planck Institute for Biological Cybernetics, Tübingen, Germany


#### Abstract

Social judgments about other people are often made based on visual appearance. In this study, we investigated whether visual appearance of an interaction partner influences action coordination in social interactions. In a novel interactive augmented reality setup participants interacted (i.e. carried out a high-five) with a life-sized 3D avatar that was either humanlooking or robot-looking. Importantly, the kinematics of the avatars were identical for both appearances. We examined whether motion trajectories of a high-five action and other motion trajectory parameters such as velocity, radial error, synchrony, and variability were modulated by the visual appearance of the avatar. Results showed that participants carried out the high-five faster and applied different motion trajectories for the human-looking than for the robot-looking avatar. These findings suggest that visual appearance does not only influence social judgments but also the immediate behavior towards the interaction partner.


# The differential effects of transmission and interaction on linguistic variation 

Olga Feher<br>University of Edinburgh<br>Kenny Smith<br>University of Edinburgh


#### Abstract

Variation in natural language is constrained: languages tend to lose competing variants over time, and where variation persists, its use tends to be conditioned on grammatical or sociolinguistic context. We had adult participants learn and communicate with artificial languages exhibiting unpredictable variation in plural marking. Using an iterated learning procedure, the languages produced by participants were used as training languages for other participants. We passed on either the language produced during a post-training recall test (Recall condition) or the language used while communicating with another participant (Interaction condition). We found that alignment during interaction leads to the elimination of variability: in Interaction chains, one plural marker typically came to dominate. However, in Recall chains, variation became conditioned on linguistic context, rather than being eliminated. This suggests that the pattern of restricted, conditioned variation in natural language reflects the combined influences of biases in learning, recall and interaction.


# Production of morphologically complex words as revealed by a typing task: Morphological influences on keystroke dynamics 

Laurie Feldman<br>SUNY Albany \& Haskins Labs<br>Rick Dale<br>UC Merced<br>Jacolien van Rij<br>U Groningen, NL<br>\section*{David Vinson}<br>UC Merced


#### Abstract

In a production by typing task, with extraneous factors (e.g., length) controlled, measures such as latency to initial keystroke as well as mean inter keystroke interval typically vary systematically according to the word's lexical properties. Conventionally, lexical effects in production tasks get interpreted as evidence of cascaded processing between central and peripheral levels. We compare mean and distribution of keystroke latencies within the same stem as it undergoes affixation in sets such as DEPRESS, DEPRESSION, DEPRESSIVE. Novel is the comparison of stems that differ with respect to number of affixes like SUPER, SUPERIOR, SUPERIORITY. Results provide new insights into the ways in which morphological structure can influence purportedly peripheral motor processing.


# Probability matching as a cognitive basis of cultural drift 

Vanessa Ferdinand<br>Santa Fe Institute, Santa Fe, New Mexico, USA


#### Abstract

In the field of cultural evolution, cognitive agents are either seen as perfect imitators who reproduce cultural variants veridically (e.g. Boyd \& Richerson 1985) or as imperfect imitators who transform the variants as they replicate them (e.g. Sperber 1996). In this poster, I explain how the transformative view of cognition applies to not only to the generation of variants, but also to the way we learn frequency distributions. Probability matching is a widely-observed human behavior where learners reproduce a frequency distribution over variants with a small amount of error and is equivalent to WrightFisher drift when the variance in error is binomial/multinomial. However, humans and learning algorithms can produce error distributions that are non-binomial/non-multinomial, which constitute a broader class of drift processes than those that exist in genetic evolution or in perfect-imitator models of cultural evolution.


# Is gender-fair langauge needed? How grammatical gender influences representations of discourse referents 

Evelyn Ferstl<br>Centre for Cognitive Science, University of Freiburg<br>Lena Dietsche<br>Centre for Cognitive Science, University of Freiburg


#### Abstract

The use of gender-fair language is an important measure to boost gender equality. However, there is wide-spread scepticism as to the usefulness of avoiding male bias in language, even in gendered languages. For instance, in German all nouns carry grammatical gender, and role names are considered generic, even when their gender is masculine. We used a sentence-picture matching task to test whether male references in language induce gendered representations. After presenting a sentence with a role name, a picture of a person was shown. In 48 trials, the factors gender of the role name (masculine vs. feminine) and sex of the person in the picture (woman vs. man) were crossed. The results of 40 participants showed that women after masculine referents were more readily accepted than men after feminine referents, but reaction times increased. Thus, readers interpret some masculine forms as generic, but only with considerable cognitive effort.


# Introducing a New JavaScript Framework for Professional Online Studies. 

Holger Finger<br>University of Osnabrueck<br>\section*{Dorena Diekamp}<br>University of Osnabrueck<br>Caspar Goeke<br>University of Osnabrueck


#### Abstract

New possibilities such as online crowdsourcing (Amazon Mechanical Turk), open data repositories (Open Science Framework), and online analysis (Ipython notebook) offer rich possibilities to improve, validate, and speed up research. However, until today there is no cross-platform integration of these subsystems. Furthermore, implementation of online studies still suffers from the complex implementation (server infrastructure, database programming, security considerations etc.). Here we present LabVanced, a JavaScript framework that constitutes methodological innovation by combining three essential aspects for online research. With our framework studies can be implemented in an intuitive graphical user interface without programming. Second, the framework takes care about participant recruitment and third, it outlines options for data visualizations and statistical analysis. Additionally, the framework can be used for sharing not only the recorded data, but also the study design and the analysis. In summary, we introduce a new powerful JavaScript framework for improving and accelerating online research.


# How to communicate uncertainty in severe weather forecasts? 

Nadine Fleischhut<br>Max Planck Institute for Human Development<br>Stefan Herzog<br>Max Planck Institute for Human Development<br>Ralph Hertwig<br>Max Planck Institute for Human Development


#### Abstract

Communicating uncertainty to lay audiences is as challenging as indispensible if people are to understand medical test results, gains from financial investments, or weather warnings.

Compared to risk communication in the medical domain, there is so far only limited evidence on how to best communicate uncertainty for continuous quantities, such as financial returns or wind speeds (Spiegelhalter et al., 2011).

The poster presents results from a longitudinal study investigating this question within a real-life setting. We implemented different representations communicating probabilistic weather forecasts within an online information system operated by the German National Weather Service. The system is used by fire brigade coordination centers throughout Germany to prepare for severe weather conditions.

By analyzing web usage and search behavior, we investigate which representations users rely upon under real operational constraints. We link the analysis to tests which representations are best understood and could thus aid emergency managers in their decisions.


# Developing Visual Closure in Infancy 

Samuel Forbes<br>University of Oxford

Kim Plunkett

University of Oxford


#### Abstract

Visual closure is the ability to complete a picture from partial information. In children it is a requisite in many skills such as fluent reading, and is also used in many tests of colour vision. Here we present a test of visual closure in infants across two age groups, $1 ; 4$ and $1 ; 7$, both to test their abilities in visual closure, but also as a prototype for a colour vision test for younger infants. The results of the study show evidence of development in visual closure abilities across those two ages, suggesting that visual closure is a perceptual ability that continues to develop in the second year of life. The results of this study are discussed in terms of perceptual development in infants and toddlers, and have consequences for both medical and scientific understanding of visual closure in children.


# Executive Functions and Academic Achievement in a High-Poverty Sample 

Gill A. Francis (1)<br>University of Cambridge<br>Zewelanji Serpell (2)<br>Virginia Commonwealth University<br>Teresa Parr (3)<br>Ashley Parr LLC<br>Michelle R. Ellefson (1)<br>University of Cambridge


#### Abstract

Research exploring cognitive theories of executive function (EF) report positive associations with academic outcomes, but whether such general cognitive theories generalise to when children are exposed to social or economic poverty contexts require more in-depth investigating. This study explores associations between EF and academic achievement for an ethnic minority sample aged 8-10 years, from high poverty, urban backgrounds. EF skills were measured using stop-signal (inhibition), continuous performance (sustained attention), task-switching (cognitive flexibility), spatial span (working memory) and Tower of Hanoi (planning). In addition, we included a popular standardized test of academic ability commonly used by schools to measure literacy, numeracy and science skills and the Raven's Progressive Matrices task to measure general cognitive ability. EF is differentiated in this sample and is linked to academic achievement. The role of important mediators like cognitive ability are considered in the context of children with high-poverty urban backgrounds.


# Neural Phase Synchrony on Understanding Meanings of Symbols 

Masayuki Fujiwara<br>Japan Advanced Institute of Science and Technology<br>Takashi Hashimoto<br>Japan Advanced Institute of Science and Technology<br>\section*{Guanhong Li}<br>Japan Advanced Institute of Science and Technology<br>Jiro Okuda<br>Kyoto Sangyo University<br>Takeshi Konno<br>Kanazawa Institute of Technology<br>Kazuyuki Samejima<br>Tamagawa University<br>Junya Morita<br>Shizuoka University


#### Abstract

The establishment of symbolic communication system, i.e., making a shared meaning system from meaningless signals, is studied in experimental semiotics (Galantucci, 2005). Local neural activities within a brain region during a symbolic communication task (Konno et al., 2013), where two participants try to establish a symbolic communication system from scratch, has been studied (Li et al, 2015). It is, however, not certain how information bindings between different brain regions is involved in a cognitive process associated with the establishment process. We analyzed EEG phase synchronization, as a measure of functional connectivity, of participants engaged in the symbolic communication task. We found the recruitment of fronto-occipital synchronization at 40 Hz frequency (gamma band), when a symbolic message was displayed, became fast when establishing a symbolic communication system. This finding suggests that frontal-occipital information binding by phase synchronization becomes efficiently used in the course of mutual understanding of symbolic messages.


# Jumping in Japanese: Converting linguistic instructions into physical performances 

Chie Fukada<br>Kyoto Institute of Technology<br>Noriyuki Kida<br>Kyoto Institute of Technology<br>Hiromichi Hagihara<br>Graduate School of Human and Environmental Studies, Kyoto University<br>Takatsugu Kojima<br>Shiga University of Medical Science


#### Abstract

This study explores the difficulties in physically realizing linguistic instructions concerning the action of jumping. We carried out a questionnaire on the understandability and physical feasibility of various jumping actions in Japanese, and then conducted an experiment in which participants were asked to jump according to these instructions. After the physical performances the participants were asked to rate the easiness of the actions in a second questionnaire, and the results of the two questionnaires were compared. The results show that the understandability of the instructions and the participants' beliefs about the physical feasibility of the instructions were closely correlated. However, the results of the two questionnaires did not correlate. The results suggest that although participants believe they can convert jump instructions into physical performances if the instructions are easy to understand, there are some gaps between the understandability of the linguistic instructions and the physical realization of them.


# The Price of Fear: Developing a behavioural assessment of fear-related avoidance incorporating dynamic response measures. 

Santiago Garcia-Guerrero

National University of Ireland, Galway<br>Denis O’Hora<br>National University of Ireland, Galway

Arkady Zgonnikov
National University of Ireland, Galway


#### Abstract

In economics, "willingness to pay" reflects subjective value which has been employed to price goods, and more recently, negative outcomes. The current project proposes a protocol for the behavioural assessment of fear-related avoidance based on how much an individual is willing to pay to avoid their fears.

The proposed protocol consists of a "card game" interface in which participants make choices in several stages. During baseline, participants chose between two decks that provide differential point rewards. Across a series of experimental blocks, feared stimuli (e.g. a spider image) were presented in addition to rewards when the richer deck was chosen. Rewards were then manipulated, in a staircase fashion, to establish the value of the feared stimulus. Mouse and eye movements were tracked in an attempt to track cognitive processes during decision-making and avoidance. Preliminary results indicate sensitivity of the protocol, and strengths and weaknesses will be discussed.


# On the road to ... somewhere? Change-blindness in event description tasks is informative about the interrelation between visual perception and language planning 

Johannes Gerwien<br>Heidelberg University, Heidelberg, Germany<br>Ines Marberg<br>Heidelberg University, Heidelberg, Germany


#### Abstract

The visual processing of complex event stimuli and the planning of utterances to describe them happen rapidly and partly overlap in time, posing a challenge to researchers on vision and language: How exactly do the processes interact? As a test case we investigate how sudden content-changes in visual scenes affect speakers of different languages. In a novel approach, we elicit event descriptions from naturalistic video stimuli of motion events consisting of two segments (240ms each), each followed by a mask ( 80 ms ). A potential change-blindness situation regarding the presence/absence of the goal of motion is created. We exploit typological differences between French and German regarding the verbal encoding of goal-orientation. Analyses of the linguistic data (content and timing) reveals a language-specific effect regarding how subjects accommodate to seemingly unnoticed changes (e.g., distribution of hesitations, temporal onsets of words). Furthermore, we find differences in overt change detection frequency depending on conditions.


# Mental computations underlying morphosyntax acquisition 

Heidi Getz<br>Georgetown University<br>Elissa Newport<br>Georgetown University


#### Abstract

Research in theoretical linguistics has shown that human languages require abstract and highly detailed grammatical representations. However, we understand surprisingly little about the mechanisms through which these representations are acquired. What kinds of statistical relationships would learners need to compute to construct representations like those posited by linguistic theory? We created miniature languages containing patterns found in natural languages and also patterns not found in natural languages. We showed that complex word-order contingencies are acquired only when they correlate with morphological patterns like those in natural languages. We then asked how learning changes when the statistical evidence for these patterns is manipulated. These experiments illuminate the nature of learners' computations and the units over which they are performed.


# Computational Modelling of Embodied Semantic Cognition: A Deep Learning Approach 

Ajitesh Ghose<br>Centre for Cognition, Computation and Modelling Department of Psychological Sciences, Birkbeck, University of London Malet Street, London, WC1E 7HX, United Kingdom<br>Rick Cooper<br>Centre for Cognition, Computation and Modelling Department of Psychological Sciences, Birkbeck, University of London Malet Street, London, WC1E 7HX, United Kingdom


#### Abstract

Barsalou's $(1999,2003)$ perceptual symbol systems hypothesis describes how semantic knowledge is grounded in sensorimotor experience. According to the theory, knowledge is acquired through sensorimotor simulations. This challenges the classical view supported by the disembodied cognition hypothesis, which generally favours an abstract and symbolic system. We propose a unified perspective, in which, the embodied cognition hypothesis, with a particular focus on the semantic domain, is provided with a mechanistically tractable computational framework based on the parallel distributed processing (PDP) paradigm. A critical difference between the current approach and previous mechanistic accounts of embodied cognition is that the current approach avoids using hand-coded representations and instead, relies on an agent-based simulation with environmental interaction for the creation of situated inputs and outputs, supplemented with supervised and unsupervised deep learning mechanisms, from which semantic cognition emerges.


# The Perceived Duration of Vast Spaces 

Devin Gill<br>University of Utah, Salt Lake City, UT, USA<br>Jeanine Stefanucci<br>University of Utah, Salt Lake City, UT, USA


#### Abstract

Experiencing awe may make us believe we have more time (Rudd, Vohs, \& Aaker, 2012). Awe can be evoked by encountering a vast experience (Keltner \& Haidt, 2003), for example an endless ocean or large mountains (Klatzky, Thompson, Stefanucci, Gill, \& McGee, 2017). Vast environments may lead to distortions in perceived time that are reported after awe experiences. Participants reproduced the perceived duration of images of natural environments that varied in vastness and estimated the degree awe they would experience in those spaces. Results show that as actual duration increased, perceived duration of the image decreased, whereas estimated awe increased. The perceived duration of highly vast images was underestimated less than other images. Participants reported they would experience more awe in highly vast images compared to low and medium vast images. These findings suggest that distortions of time associated with awe may be related to the vastness of the environment.


# The role of spatial skills in mathematics cognition: Evidence from children aged 5-10 years 

Katie Gilligan<br>UCL Institute of Education, London, United Kingdom<br>Alex Hodgkiss<br>UCL Institute of Education, London, United Kingdom<br>Michael Thomas<br>Birkbeck College, University of London<br>Emily Farran<br>UCL Institute of Education, London, United Kingdom


#### Abstract

While there is evidence of associations between spatial skills and mathematics, relatively few studies explore these associations in children aged 5-10 years. I will present findings from longitudinal and cross-sectional studies to highlight the importance of spatial skills as both longitudinal and concurrent predictors of mathematics. First, secondary data analysis of the Millennium Cohort Study indicates that spatial performance at both 5 and 7 years is a significant predictor of mathematics at age $7(\mathrm{~N}=12099)$. Second, cross-sectional findings from children aged $5-10$ years $(\mathrm{N}=156)$, suggest that spatial skills explain 10$12 \%$ of the variation in standardised maths performance and approximate number sense, even after accounting for vocabulary skills. That is, spatial scaling was a significant predictor of mathematics for all age groups, while the role of mental rotation and mental folding varied with age. These findings have implications for the design of mathematics interventions customised for specific developmental stages.


# Investigating the predictions of a memory-based account of statistical learning 

Sandrine Girard<br>Carnegie Mellon University<br>Erik Thiessen<br>Carnegie Mellon University


#### Abstract

Statistical learning (SL) refers to the ability to extract statistical regularities from the environment. Many researchers believe that SL arises as a consequence of the way that information is stored and accessed in memory (Thiessen, Kronstein, \& Hufnagle, 2013). Accordingly, manipulations that influence memory should have similar effects in SL experiments. In the current study, participants were presented with artificial languages that varied along two dimensions known to impact memory: number of distractors in the input and timing of presentation (e.g., spaced vs. massed). Participants' performance was clearly influenced by these manipulations; for example, the ability to segment a word (e.g., "dupona") differed as a function of whether there was one frequent competitor (e.g., "dugalo") or several less frequent competitors (e.g., "dugalo," "dufalu," "dumiso"). Experimental results were compared to two memory-based computational models of SL (PARSER and TRACX). Implications of the experimental results and model comparisons will be discussed.


# Motor Fluency Effects on Causal Judgment: The Role of Grip-Strength Asymmetries and Spatial-Numeric Associations 

Kelly Goedert<br>Seton Hall University<br>Daniel Czarnowski<br>Seton Hall University


#### Abstract

Human understanding of causation may be grounded in our experience of physical forces in the world. We investigated whether right-handers, who exert greater force with their right than left hands, judge candidate causes on the right side as more causal. In two experiments, subjects simultaneously learned about a moderately effective and an ineffective cause on a trial-by-trial basis. Subjects rated the moderately effective cause as more causal when it appeared on the right side of space. This effect was also present in subjects' trial-by-trial predictions, but the effect reversed with a left-right reversal in the spatial-numeric mapping of the causal judgment scale. The results are not consistent with the notion that our understanding of causation is grounded in our ability to exert force. However, they are consistent with influences of motor fluency and polarity correspondence on judgment.


# Comparison of directed gaze during vocalizations in bonobo and human infants 

Ulrike Griebel<br>The University of Memphis, Memphis, TN, USA<br>Josep Call<br>Max Planck Institute for Evolutionary Anthropology. Leipzig, Germany<br>Eugene Buder<br>The University of Memphis, Memphis, TN, USA<br>D. Kimbrough Oller<br>The University of Memphis, Memphis, TN, United States, USA


#### Abstract

A crucial step in language evolution was likely joint attention with alternating gaze between vocalizing individuals and an object. This triadic interaction likely formed a foundation for labeling of objects. We have argued that vocalizations used for "social glue" - flexible low intensity and low arousal vocalizations given during e.g. grooming, keeping in contact with the group, etc. - are a probable source of raw material for first labels. It is critical that these vocalizations be socially directed, by gaze contact. We longitudinally investigated directed gaze during vocalizations in low arousal interactions during the first year in three bonobo mother-infant pairs and compared them with 9 human mother-infant pairs. We found that bonobo infants directed their gaze to a conspecific during vocalizations only $8 \%$ of the time while human infants directed it $44 \%$ of the time.


# How does of initial inaccuracy benefit cross-situational word learning? 

Chris Grimmick<br>New York University<br>George Kachergis<br>Radboud University<br>\section*{Todd Gureckis}<br>New York University


#### Abstract

Both children and adults are able to extract several intended word-referent mappings from a series of scenes containing multiple words and objects. Known as cross-situational learning, this ability is thought to be an important means of acquiring language. Proposed models of this ability range from hypothesis-testing accounts to associative accounts, but most formal models assume learners store one or more feasible word-referent mappings per experience, and that the correct mappings emerge through consistent co-occurrence. These theories would all predict that presenting unambiguous evidence for a correct pair would benefit learning, but recent evidence indicates the reverse is true: giving unambiguous evidence for incorrect pairs improves subsequent cross-situational learning (Fitneva and Christiansen, 2015). With some nuances, we replicate these results, and show why future models may need to include an error-driven learning mechanism to explain word learning.


# On the Detection of "Alternative Facts" in Environmental Messages: The Effects of a Sequential versus a Simultaneous Presentation Format 

Mona Guath<br>Uppsala University, Uppsala, Sweden<br>Peter Juslin<br>Uppsala University, Uppsala, Sweden


#### Abstract

Reasonable rational information processing is important in people's in everyday decision-making. A number of features affect how environmental messages are processed, including the presentation format and the reliability of the information source. One way to measure the importance assigned to the source reliability is to frame the question in terms of Bayes' theorem (Hahn \& Harris, 2009). In two online experiments, we investigated how people process environmental messages in a Bayesian integration task where the participants rate the probability of an energy crisis. The information about the prior, likelihood ratio, and source reliability were presented either sequentially or simultaneously. The results showed that, as prescribed by Bayes' theorem, participants integrated the sentences multiplicatively. However, with sequential presentation they assigned more weight to source reliability, and this effect remained when the source reliability was presented next to last, suggesting that participants assigned more weight to the source regardless of its position.


# Exploring inductive bias of visual scenes 

Jessica Hamrick<br>University of California, Berkeley, Berkeley, CA, USA<br>David Bourgin<br>University of California, Berkeley, Berkeley, CA, USA<br>Thomas Langlois<br>University of California, Berkeley, Berkeley, CA, USA<br>Tom Griffiths<br>University of California, Berkeley, Berkeley, CA, USA


#### Abstract

When people encode a representation of a scene, they do not necessarily represent the exact locations and orientations of the constituent elements. Instead, people rely on preexisting inductive biases to simplify their encoding of new scene configurations. We investigated people's inductive biases in their memory for configurations of simple 2D shapes (such as circles, triangles, etc.) using a serial reproduction paradigm (Bartlett, 1932). This paradigm establishes an iterative process in which information is transmitted through a chain of people (like the "telephone" game). In our experiment, we asked people to memorize configurations of simple shapes (which were either generated at random or by other participants) and then asked them to reproduce those configurations. In analyzing the final generation of reproductions, we found that people have strong preferences for the scale of individual shapes, as well as the alignment, distance, overlap, and relative rotation between pairs of shapes.


# Choosing while Losing: The Effects of Valence and Relative Magnitude on Decision Dynamics. 

Avril Hand<br>National University of Ireland, Galway, Ireland<br>Denis O'Hora<br>National University of Ireland, Galway, Ireland<br>Rick Dale<br>University of California, Merced, USA<br>Petri Piiroinen<br>National University of Ireland, Galway, Ireland


#### Abstract

Framing decision options as gains or as losses affects how we evaluate those options. The current study assessed the effects of gain- and loss-framing on the acquisition of outcome values across decisions and on the dynamics of computer mouse responses to those decisions. In a series of 36 decisions per block, four arbitrary symbols were presented, two of which were assigned high points (e.g., 20) and two of which were assigned low points (e.g., 5). Participants ( $\mathrm{N}=86$ ) learned to choose high values and avoid low values when values were positive and to choose low values and avoid high values when they were negative. Loss-framed outcomes (i.e., negative valence) were learned faster and more reliably. Response trajectories following acquisition were slower, more curved and exhibited greater vacillation when choosing between two poor outcomes. These effects were stronger when poor outcomes were negatively valenced.


# The Relationship between Anxiety, Mind Wandering and Task-switching: A Diffusion Model Analysis 

Andree Hartanto<br>Singapore Management University<br>Hwajin Yang<br>Singapore Management University


#### Abstract

The current study aims to examine the mechanisms underlying the negative impact of anxiety on task-switching. To do so, we employed a stochastic diffusion model analysis along with a thought-probe technique in task-switching paradigm. Across 152 participants, we found state anxiety was associated to higher switch costs in nondecision time but not drift rate parameter of diffusion model, implying that the locus of task-switching impairment in anxious individuals is pertinent to the efficiency of task-set reconfiguration but not proactive interference processes. Furthermore, we found boundary separation parameter - which quantifies conservative decisional styles - heightened as a function of anxiety, supporting the existence of compensatory strategy in anxious individuals. Lastly, we found that impaired performance by anxiety was not attributed to the frequency of worrisome thoughts during task-switching. These findings elucidate several theoretical assumptions on the relationship between anxiety and task-switching.


# Do people behave dishonestly easily? 

Hajimu Hayashi

Kobe University


#### Abstract

This study examines whether dishonest behaviors occur easily. In 60 trials, 100 undergraduate students viewed 20 dots on a square divided into right and left sides and had to decide which side contained more dots within one second (developed by Gino et al., 2010). In with-reward condition, participants received 0.1 point for each left decision and 1 point for each right decision, and they received more sweets depending on points. Therefore, this asymmetrical payment structure triggered motivation to dishonestly report more right-side dots, even when there are actually more left-side dots. The results demonstrate that participants decided at greater frequencies that more dots were on the right side in with-reward condition than in without-reward condition, indicating dishonest behaviors occurred. Furthermore, participants with greater right-side frequencies in with-reward condition showed lower points on a morality scale. These results suggest dishonest behaviors occur easily and are related with a decline in morality.


# Hierarchical Processing of Response Production and Categorisation 

Liusha He<br>Centre for Cognition, Computation and Modelling Department of Psychological Sciences, Birkbeck, University of London, London, London, United Kingdom

Rick Cooper
Birkbeck, University of London, London, England, United Kingdom


#### Abstract

Early research on categorisation suggested that verbalizable and nonverbalizable category-learning are qualitatively different. Toward this end, the implementational-level model (COVIS-COmpetition between Verbal and Implicit Systems) of categorisation assumes that category-learning involves separate but parallel sub-systems. Specifically, for verbalizable tasks abstract category-labels are learned by a hypothesis-testing sub-system, while for nonvertbalizable tasks response position is learned by a procedural-learning sub-system. However, recent research has revealed that: 1) regardless of category structure, reversal learning is easier than learning novel categories; 2) qualitative difference between verbalizable and nonverbalizable tasks disappears when automaticity has developed; and 3) control of automatic categorisation is different from both proposed sub-systems. These challenges suggest a fundamental revision of the mechanisms of categorisation. Contrary to the separate, parallel-processing sub-systems theory, we argue that categorisation involves hierarchical-processing sub-systems of responseproduction and category-label association. This framework, when combined with Supervisory Attentional System theory, may facilitate the unification of human categorisation.


# Where are you? The Effect of Uncertainty and its Visual Representation on Location Judgments in GPS-like Displays 

Mary Hegarty<br>University of California, Santa Barbara

Alinda Friedman<br>University of Alberta

Alexander P. Boone<br>University of California, Santa Barbara, Santa Barbara, CA, USA

Trevor J. Barrett
University of California, Santa Barbara


#### Abstract

Two experiments revealed how non-experts interpret visualizations of positional uncertainty on GPS-like displays and how the visual representation of uncertainty affects their judgments. Participants were shown maps with representations of their current location; locational uncertainty was visualized as either a circle (confidence interval) or a faded glyph (indicating the probability density function directly). When shown a single circle or faded glyph, participants assumed they were located at the center of the uncertain region. In a task that required combining two uncertain estimates of their location, the most common strategy - integration - was to take both estimates into account, with more weight given to the more certain estimate. Participants' strategies were not affected by how uncertainty was visualized, but visualization affected the consistency of their responses. The results indicate that non-experts have an intuitive understanding of uncertainty and that the best visualization method is task dependent.


# Semantic Bootstrapping in Frames: A Computational Model of Syntactic Category Acquisition 

John Hewitt<br>University of Pennsylvania<br>\section*{Charles Yang}<br>University of Pennsylvania


#### Abstract

Semantic Bootstrapping in Frames: A Computational Model of Syntactic Category Acquisition According to the semantic bootstrapping hypothesis, children map certain (prototypical) semantic concepts to syntactic categories (e.g., objects as nouns, actions as verbs), which are then used to learn other elements of language. We report a computational model of syntactic category acquisition that combines semantic bootstrapping with the distributional learning of language. The model has access to a small set of "seed" words, with labeled categories. It then iteratively constructs syntactic frames from the seeds; sufficiently frequent frames are used to categorize non-seeded words which then contribute to the construction of additional frames, including frames that incorporate category information. The model is online and effective. Simulation on child-directed English corpus shows that with only 100 seed words, classification precision exceeds $70 \%$.


# Tangible rhythm: Sensorimotor representations of metrical structure and learning musical rhythm with gesture 

Courtney Hilton<br>The University of Sydney, Sydney, NSW, Australia<br>Micah Goldwater<br>The University of Sydney, Sydney, NSW, Australia<br>Michael Jacobson<br>The University of Sydney, Sydney, NSW, Australia


#### Abstract

When we listen to music, we can mentally control how we perceive the beat. This ability is thought to be subserved by sensorimotor imagery, having top-down effects on attentional-allocation and perception. Here, we examine whether imagined "up and down" gestures can support an internal generation of metrical accent in rhythmic sequences. We also examine how this type of motor imagery interacts with either metrically congruent or incongruent auditory imagery. This is explored using EEG with a frequency-tagging approach, quantifying the strength of metrical accent with the amplitude of beat-related SSEPs. Gesture supports our ability to think and learn by fostering an alignment between sensorimotor representations and more abstract conceptual structure. Therefore, the imagined gestures may act as a bridge between perceptual and action-oriented understandings of metrical structure and the more abstract conceptual ones that musicians struggle with in their training. These imagery strategies may then be beneficial to music education.


# Eye movement-based probabilistic models for physical scene understanding 

Eghbal Hosseini<br>Massachusetts Institute of Technology<br>Eli Pollock<br>Massachusetts Institute of Technology<br>Tobias Gerstenberg<br>Massachusetts Institute of Technology


#### Abstract

Humans make prediction about physical environments and future events through inference. Previous research has proposed that a common sense engine implementing probabilistic programming is used to build an internal model of the environment, and simulations of that internal model are used for inferences. Battaglia et al.(2013) have demonstrated an application of this formulation in physical scene understanding and stability judgment in the case of block tower. Here we augment this formulation by including the subjects' eye movements as a process of sampling the environment, and propose that the underlying common sense model guides gaze toward sampling the features of the space with relevant information for the judgments about stability. We compare a base probabilistic model with one that takes the statistics of the saccades into account, and argue that the additional information improves the model predictions about subjects' judgment.


# Data Driven Eye Gaze Path Segmentation 

Joseph Houpt<br>Wright State University, Dayton, Ohio, USA<br>Mary Frame<br>Miami University, Oxford, Ohio, USA<br>Leslie Blaha<br>Pacific Northwest National Laboratory, Richland, Washington, USA


#### Abstract

The first stage of analyzing eye-tracking data is commonly to code the data into sequences of fixations and saccades. This is usually automated using simple, predetermined rules for classifying ranges of the time series into events, such as "if the dispersion of gaze samples is lower than the threshold, then code as a fixation." More recent approaches incorporate additional eye-movement categories in automated parsing algorithms, particularly glissades, by using time-varying, data-driven thresholds. We describe an alternative approach using the beta-process auto-regressive hidden Markov model (BP-AR HMM). The BP-AR-HMM offers two main advantages over existing frameworks. First, it provides a statistical model for eye movement classification rather than a single estimate. Second, the BP-AR-HMM uses a latent process to model the number and nature of the types of eye-movements and hence is not constrained to predetermined categories. We present comparisons between BP-AR-HMM parsing and standard analyses on multiple datasets.


# Case Markers Facilitate Abstraction of Syntax among Mandarin-speaking preschoolers 

Dong-Bo Hsu<br>National Taiwan Normal University, Taipei, Taiwan, Taiwan, R.O.C.


#### Abstract

In light of the studies that investigate when and how English-speaking young children exhibit adult-like abstraction of syntax, this study explores these issues by examining Mandarin-speaking preschoolers ranging from 2- to 5-year-olds' comprehension of the Mandarin SVO-, ba-, long-passive, and short-passive constructions using the forced-choice pointing paradigm. The results indicated that at the age of 2, Mandarin preschoolers exhibited abstraction of syntax in these four constructions. These results went against the predictions of accounts derived from the structure mapping account and from the competition model. Instead, Mandarin ba- (used in the ba-construction) and bei-markers (used in the long- and short-passive constructions) play an important role in Mandarin-speaking young children's demonstrations of abstraction of syntax.


# An Exploratory Study on Remote Associates Problem Solving: Evidence of Eye Movement Indicators 

Po-Sheng Huang<br>Hsuan Chuang University, Hsinchu, Taiwan, R.O.C.<br>Shu-Ling Peng<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC<br>Jon-Fan Hu<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC


#### Abstract

Remote associates problems (RAP) have been widely used to measure creative processes. However, studies have rarely explored the RAP processes. The main purpose of this study was to record eye movements while solving RAP. The results show that: (1) The mean fixation duration increases throughout the problem-solving process, which indicates that more problem solvers encounter impasses. This result supports the "impasse encounter" phase of insight. (2) During the initial period of problem solving, individuals display more regression counts in the fixation region than in the key region, which supports that the impasses are caused by inappropriate initial representation. (3) During the middle period of the process, the time individuals spend gazing at the key region increases, while the time spend at the fixation region decreases. This supports the "impasse resolution and insight" phase of insight.


# Processing of Filler-gap Dependency in Island Constraints and its Relation to Working Memory for Non-native Speakers of English 

Kuan-Jung Huang<br>National Cheng Kung University, Tainan, Taiwan<br>Jon-Fan Hu<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC<br>Anastazija Rajkovic<br>National Cheng Kung University, Tainan, Taiwan


#### Abstract

Whether non-native speakers' on-line processing can be native-like remains a hot issue. Recently many have shown that qualitatively native-like processing is attainable, especially for learners with high proficiency. However, most of the studies recruited learners who had been immersed to the English-speaking country. The current study investigated processing of fillergap dependency and island constraints for Chinese learners of English as a second but foreign language. We also attempted to look into individual differences by taking different variables into account. The results showed that native-like active gap-filling strategy positively correlated with L2 proficiency, native-like island effect negatively correlated with age of acquisition, but neither one correlated with working memory capacity. These findings lent more support to the grammar-based account for island effects, though future studies adopting more precise measure of working memory would be needed. The study also called for further investigation into L1 background on processing islands in an L2.


# Linguistic processes in translation: Eye-tracking reveals differential effects of phrase order and lexical choice 

Xin Huang<br>University of Macau<br>Wei (sophia) Deng University of Macau<br>University of Macau

## Defeng Li

University of Macau


#### Abstract

What are the processes underlying the judgments of translation? And what is the role of language proficiency? This study addresses these questions by examining how Chinese-English bilinguals evaluate poetry translations. Participants were shown haikus in Chinese and the corresponding English translations and were asked to rate the translation quality. The English translations ranged from literal to free style and differed in two source text factors - phrase order and lexical choice. Results indicated an interaction between translation style and language proficiency, with the high proficiency bilinguals giving free translations higher ratings. Furthermore, the analyses of eye movements revealed that, (a) in contrast to low proficiency bilinguals, high proficiency bilinguals tended to integrate discourse information regardless of intra-text re-ordering, and (b) among the good quality translations, the phrase order effect was more prominent than the lexical effect. These findings suggest the interplaying roles of language proficiency and linguistic factors in translation.


# How reactivation strength affects memory updating 

Scully Iiona<br>Lehigh University<br>Hupbach Almut<br>Lehigh University


#### Abstract

Memory reactivation induces plasticity, rendering reactivated memories susceptible to interference. The current study examined whether the method and strength of reactivation modulates retroactive interference effects. Two days after learning AB word pairs, memory for these pairs was either not reactivated, moderately reactivated (presentation of A cues in an unrelated task), or strongly reactivated (restudy of AB pairs or cued recall of B targets). Immediately afterwards, participants either learned AC word pairs, DE word pairs, or performed an unrelated distractor task. Cued recall of target words was tested two days later. Strong reactivation before learning new material protected memory from retroactive interference and intrusions, whereas moderate reactivation resulted in both. This finding suggests that strong reactivation enhances event-based distinctiveness, counteracting memory modification. Results are discussed in reference to the testing effect literature and the reconsolidation account, and implications for educational practice are outlined.


# An fNIRS Hyperscanning Study on Brain-Brain Interactions of a Dyad during a Joint Sentence Reading Task 

Erdinc Isbilir<br>Middle East Technical University, Ankara, Ankara, Turkey<br>Murat Perit Cakir<br>Middle East Technical University, Ankara, Turkey<br>Fred Cummins<br>University College Dublin<br>Hasan Ayaz<br>Drexel University


#### Abstract

Existing studies in cognitive neuroscience predominantly focus on a single participant's behavioral and brain responses. Lack of an interactive context for joint action particularly limited social neuroscience studies to simulated social contexts. Advances in portable brain imaging technologies have made it practical to simultaneously monitor the brain activity of two or more people in an interactive context to investigate neural correlates of social interaction. In this study, the relationship between behavioral synchrony and inter-brain coherence is investigated during simultaneous reading of matching and mismatching sentences in different auditory conditions. A dual-fNIRS hyperscanning setup was used to obtain simultaneous recordings of hemodynamic activity from the prefrontal cortices of the participants while they jointly read-aloud the sentences displayed on their screens. The results suggest that the level of inter-brain coherence in the right superior cortex tends to increase depending on the level of behavioral synchrony among the participants.


# Pre-term infants exhibit impaired prediction and learning in Audio-Visual association paradigm 

Sagi Jaffe-Dax<br>Princeton University

Alex M Boldin<br>Princeton University<br>Nathaniel D Daw<br>Princeton University<br>Lauren L Emberson<br>Princeton University


#### Abstract

Forming reliable predictions about upcoming events are both essential to and the product of successful learning. Using fNIRS recording of cortical hemodynamics, we measured infants' prediction of upcoming visual events that were preceded by auditory cues in infants who are at-risk for poor development due to premature birth and their full-term peers. We compared prediction and learning across groups by fitting their occipital cortex response (which we assumed to reflect the magnitude of the prediction error) to a reinforcement learning model with a dynamic learning rate. We found that preemies had a lower learning rate than full-terms. These findings shed light on the origins of the developmental difficulties associated with prematurity.


# Endpoints and Midpoints in Event Perception 

Yue Ji

University of Delaware

Anna Papafragou<br>University of Delaware


#### Abstract

Events unfold over time, i.e., they have a beginning and endpoint. Previous studies have illustrated the importance of endpoints for the perception and memory of various events (Lakusta \& Landau, 2005, 2012; Papafragou, 2010; Regier \& Zheng, 2009; Strickland \& Keil, 2011; Zacks \& Swallow, 2007). However, this work has not compared endpoints to other potentially salient points in the internal temporal profile of events (e.g., midpoints). Building on the "picky puppet task" (Waxman \& Gelman, 1986), we presented 4-to-5-year-old children and adults with a puppet that liked clips of events containing brief screen blanks that disrupted either the midpoint or the endpoint of the event. Both children and adults learned the puppet's preferences better (as evidenced by their extension to novel events) when the puppet liked midpoint compared to endpoint interruptions. These findings suggest a bias for event endpoints that is present from an early age.


# Dynamic and multiplexed networks for working memory 

Elizabeth Johnson<br>University of California, Berkeley<br>Robert Knight<br>University of California, Berkeley


#### Abstract

Working memory (WM) provides the neurobiological infrastructure for human cognition. Dominant models posit that prefrontal cortex (PFC) supports WM by coordinating control over distributed memory representations. In two studies, multimodal electrophysiology data reveal that PFC control over WM is rhythmic, fundamentally dynamic, and not altogether necessary. Direct brain recordings ( $\mathrm{n}=10$ ) demonstrate that PFC and medial temporal lobe (MTL) theta-band rhythms direct a complex system of higher-frequency neural activity across regions, uncovering initial support for bidirectional PFC-MTL interactions related to WM demands. Then, data from patients with unilateral PFC damage ( $\mathrm{n}=14$ ) challenge dominant models on the central role of PFC (note $8 \%$ accuracy decrease in patients). In healthy controls ( $\mathrm{n}=20$ ), delta-theta-band rhythms precess from PFC toward parieto-occipital sites, concurrent with alpha-beta-band rhythms precessing in the opposite direction. All PFC effects are diminished with unilateral damage, revealing an independent posterior WM mechanism. These results reveal that rapid, parallel processing governs WM.


# Improving Perceptual Reasoning in School Children through Chess Training 

Ebenezer Joseph<br>Madras University, Chennai, Tanilnadu, India<br>Sumathi Chandrasekaran<br>Emmanuel Chess Centre - DST Project, Chennai, Tamilnadu, India<br>Uma T.r.<br>Emmanuel Chess Centre,Chennai, Tamilnadu, India<br>Veena Easvaradoss<br>Women's Christian College, Chennai, Tamilnadu, India


#### Abstract

Perceptual reasoning is the ability that incorporates fluid reasoning, spatial processing, and visual motor integration. Several theories of cognitive functioning emphasize the importance of fluid reasoning. Tasks that require fluid reasoning involve the process of manipulating abstractions, rules, generalizations, and logical relationships. A pretest-posttest with control group design was used, with 43 ( 28 boys, 15 girls) children in the experimental group and 42 ( 26 boys, 16 girls) children in the control group. The sample was selected from children studying in two private schools from South India, which included both the genders. The experimental group underwent weekly one-hour chess training for one year. Perceptual reasoning was measured by three subtests of WISC-IV INDIA. Pre-equivalence of means was established. Statistical analyses revealed that the experimental group shows statistically significant improvement in perceptual reasoning compared to the control group. The present study establishes a correlation between chess learning and perceptual reasoning.


# Measuring Demand Avoidance with the Demand Selection Task: Challenges and Opportunities 

Ion Juvina<br>Wright State University<br>Jeffrey Nador<br>Wright State University<br>Othalia Larue<br>Wright State University<br>Randall Green<br>Wright State University<br>Brandon Minnery<br>Wright State University

Assaf Harel<br>Wright State University


#### Abstract

When given the chance to choose between two tasks, one will more likely choose the easier, less demanding task. This effect has been shown in various domains and referred to as the law of minimum effort or demand avoidance. Kool and colleagues $(2010,2013)$ designed the demand selection task (DST) and showed that most of their participants exhibited clear demand avoidance. We attempted to replicate and extend their results in a series of three studies. Here we argue that DST confounds demand detection and demand selection, which weakens its ability to reliably measure demand avoidance in different populations. In our first study, most participants did not show reliable demand avoidance and those who showed it had higher working memory capacity. The following two studies aimed to de-confound the two processes. We define a new measure of demand avoidance that affords a more robust estimation of demand avoidance in different populations.


# Optimizing Mathematic Learning: Effects of Continuous and Nominal Practice Format on Transfer of Arithmetic Skills 

Charles Kalish<br>University of Wisconsin-Madison, Madison, Wisconsin, USA<br>Rui Meng<br>University of Wisconsin-Madison, Madison, Wisconsin, USA<br>Ayon Seng<br>University of Wisconsin-Madison, Madison, Wisconsin, USA<br>Percival Matthews<br>University of Wisconsin-Madison, Madison, Wisconsin, USA


#### Abstract

Should we give learners a lot of practice with a few problems, or a little practice with a variety of problems? The best practice set depends on the way people are learning. We describe two models people employ when learning arithmetic problems. We show that features of the task environment influence model use. When problems are presented in a purely symbolic format, people learn an item-specific model. When the task format linked problems to representations of magnitudes, people learn a continuous model. We also test the effects of different practice sets on learning. In both formats people learned the practice sets well with a few repeated examples. With a continuous magnitude format people showed better transfer with a wide variety of practice problems. Variety led to poor learning in the symbolic format. In ongoing research we are attempting to identify the optimal practice set for each type of learning model.


# Priors, informative cues and ambiguity aversion 

Lauren Kennedy<br>University of Adelaide

Amy Perfors<br>University of Adelaide

## Daniel Navarro

University of New South Wales


#### Abstract

Ambiguity aversion, or the preference for options with known rather than unknown probabilities, is a robust finding within the decision making literature (see Camerer \& Weber, 1992, for a review). There are some suggestions this averseness is due differences in the inferred prior distribution (Güney \& Newell, 2015). In this study we investigated the relationship between prior distributions and information cues on decision making and participants' judgments of underlying distribution. We used three different prior cues; a positive underlying distributional cue, a negative underlying distributional cue, and a neutral cue. We also used five different information cues which varied both the bias of the information and the degree of ambiguity. Whilst we found that both prior and information manipulations had the expected impact for participants' judgments of underlying distributions, they only impacted the decisions participants made in some cases. There were also interesting interactions between the prior and information manipulations.


# Moral Judgments in Trolley Like Dilemmas: An Eye-Tracking Study 

Vjeran Keric<br>New Bulgarian University

## Evgeniya Hristova

New Bulgarian University


#### Abstract

Previous research suggests that participants may be susceptible to confirmation bias after making decisions in moral dilemmas. We manipulated the type of moral dilemmas (personal or impersonal) and the framing of the question prompting participants to respond (emphasizing saving five people or sacrificing one person). The actors in the dilemmas were represented by a series of silhouettes. Eye tracking data revealed that both manipulations had an effect on participants' gaze. Further analysis of utilitarian choices has shown that there were no framing effects of the prompting question when the dilemmas were impersonal. The data suggests that participants' subsequent gaze patterns are sensitive to both how the situation is described and the framing of their hypothetical actions. Taken together, our results provide some support to the claim that confirmation bias may arise after making moral decisions.


# Word retrieval decline in midlife: a voxel-based morphometry study 

Vanja Kljajevic

University of the Basque Country, Vitoria \& IKERBASQUE, Basque Foundation for Science, Bilbao, Spain


#### Abstract

There is currently little understanding on whether significant word retrieval difficulties appear during midlife and if so, whether they relate to decrease in grey matter (GM) density that accompanies aging. To answer this question, we studied retrieval of proper names in 125 cognitively healthy middle-aged persons ( $49.7, \pm 3.2$ ) comparing their performance on a tip-of-the-tongue (TOT) task with that of 86 young persons $(25.4, \pm 3.5)$ from the Cam-Can data repository (http://www.mrccbu.cam.ac.uk/datasets/camcan/). The middle age (MA) group was worse in word retrieval ( $\mathrm{U}=23950.5, \mathrm{p}=0.003$ ) and had less GM volume in a range of left fronto-temporal areas relative to the young group, but there were no statistically significant correlations between volumes of the regions known to be implicated in word retrieval and MA's TOT scores. Thus, midlife word retrieval decline is not associated with GM volume reduction; more likely it is due to changes in connectivity.


# The relationship between verbal route descriptions and personal characteristics of empathy 

Takatsugu Kojima<br>Shiga University of Medical Science


#### Abstract

Empathy is important for good verbal navigation instructions. The present study examined the relationship between verbal route descriptions and personal characteristics of empathy using a verbal navigation task and the Japanese Interpersonal Reactivity Index (IRI-J). In the verbal navigation task, participants were presented a maze map and were instructed to provide verbal route instructions to reach a goal, to a person who is lost in the maze. Then, the participants answered a questionnaire about their own navigation abilities and responded to the IRI-J. The descriptive data were objectively evaluated with reference to the following three points: types of spatial description (survey or route), consideration for the other person's point of view, and unambiguity of instruction. We analyzed the relationships between the descriptive traits and scores on the two questionnaires, and found that the three points were good predictors for some empathy-related factors measured by the IRI-J.


# Explanatory Completeness: Evidence from Causal Chains 

Joanna Korman<br>Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington DC, USA<br>\section*{Sangeet Khemlani}<br>Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, USA


#### Abstract

Explanations have no bound in principle, but in practice, people prefer explanations that are complete (Zemla et al., 2017), and the explanations that they generate are bounded (Miyake, 1986). We tested reasoners' ability to assess whether some explanations are incomplete. Participants in three experiments received explanations, i.e., chains of causal events, e.g., A causes B causes C. Their task was to choose questions relevant to links in the chain. Some explanations contained "breaks" in the chain, whereas others did not. Participants in three studies were able to detect the breaks, and preliminary data suggest that they assess explanations with breaks as less complete than those without breaks. Many participants also chose to ask questions about the initial event in a causal chain (e.g., A in the chain above), suggesting that such initial events are themselves seen as incomplete. The studies reveal a novel pattern in reasoners' ability to formulate explanations.


# Different alternative explanations can render different information relevant to explaining an event 

Barbara Koslowski<br>Cornell University<br>Francoise Vermeylen<br>Cornell University


#### Abstract

Scientific reasoning includes deciding whether information is relevant to explaining an event. In some cases, seeing information as relevant requires having a background theory or explanation that can make sense of the information. College students were shown a possible explanation for an event, along with two pieces of possibly relevant information (Infol and Info2), and one of two possible alternative explanations (Alt1 or Alt2). Info1 was seen as more relevant when Alt1 rather than Alt2 was presented; Info2 was seen as more relevant when Alt2 rather than Alt1 was presented. In addition, relevance ratings of the information increased as did initial ratings of the Alternative. People from different backgrounds might bring with them different alternative theories that can hinder the understanding of why some information is relevant and other information not. In addition, finding the initial alternative compelling might enable people to better assess the relevance of additional information.


# One-shot Learning and Classification in Children 

Eliza kosoy<br>MIT, cambridge<br>\section*{Brenden Lake}<br>New York University, New York City, NY, USA<br>Josh Tenenbaum<br>MIT, Cambridge


#### Abstract

People can often generalize concepts from just a single example, while machine learning algorithms typically require hundreds. Lake, Salakhutdinov and Tenenbaum (2016) studied this ability in the domain of handwritten characters, and proposed a model for one-shot learning of new concepts based on inferring compositionally structured generative models, and transfer (or learning to learn) from familiar concepts. Lake et al showed that their model fit well with the classifications and drawings of adults, but provided no direct evidence for the role of learning to learn which presumably occurs mostly in children learning to draw. Here we study the drawing and classification abilities of children ages 3-5, asking whether their ability to classify novel objects and handwritten characters is related to their ability to infer an appropriate motor program for drawing or tracing characters. Preliminary results suggest at least a weak relationship between these abilities, independent of age.


# Mentioning atypical properties of objects is communicatively efficient 

Elisa Kreiss<br>University of Osnabrueck<br>Robert X.d. Hawkins<br>Stanford University<br>Judith Degen<br>Stanford University<br>Noah Goodman<br>Stanford University


#### Abstract

What governs how much information speakers include in referring expressions? Atypical properties of objects are more likely to be included in referring expressions than typical ones. E.g., speakers are more likely to call a blue banana a "blue banana" and a yellow banana a "banana". A unified account of this phenomenon is lacking. When should a rational speaker mention an object's color? Reference production is modeled within the Rational Speech Act framework. Utterances ("banana", "blue", and "blue banana") are taken to have a graded semantics: rather than assuming all bananas are equally good instances of "banana", we empirically elicited object-utterance typicality values for all possible utterances. Pragmatic speakers select utterances proportionally to the probability that a literal listener using a graded semantics will select the intended referent. We evaluate the proposed model on a dataset of freely produced referring expressions collected in an interactive reference game experiment via the web.


# Why Would 'Same' Go With 'Same'? Exploring New Factors Required For Relational Reasoning 

Ivan Kroupin<br>Harvard University<br>Susan Carey<br>Harvard University


#### Abstract

Relational Match to Sample (RMTS) is a common test of relational reasoning involving matching cards based on the relations "same" and "different". Children below the age of five fail RMTS, even with corrective feedback. Given that success on RMTS depends on the ability to represent and compare "same" and "different", such failure has been interpreted as indicative of the absence of these abilities (Penn, Holyoak \& Povinelli, 2008; Hochmann, Mody \& Carey, 2016).

In the current studies three, four and five-year-old children were provided explicit instructions on RMTS. Results show success in all groups, including three-year-olds - two years earlier than previous work. This suggests the ability to represent and compare "same" and "different" emerges significantly earlier than spontaneous success on RMTS, undermining previous interpretations. More generally, this work begins to explore the nature of the development which allows existing relational reasoning capacities to be spontaneously deployed in RMTS.


# When reading is harder than a mother kucker: Top-down effects of the taboo-ness on novel word pronunciation 

Sarah Kucker<br>University of Wisconsin Oshkosh<br>Lynn Perry<br>University of Miami, Coral Gables, Florida, United States


#### Abstract

When pronouncing novel/unknown words, readers often use prior experience with similar, neighbor words. Comparison to neighbors can be helpful for unknown or novel words (wug is like pug), but it can also lead to errors (pint is not like mint). We investigate whether pronunciation can be affected by top-down influences, specifically the perceived taboo-ness of a known neighbor. While orthographic similarity typically biases novel-word pronunciation to be similar to a known word, taboo-ness might bias pronunciation away from a likely one. Adults read aloud words from three lists- novel words that were neighbors to taboo words, novel words that were neighbors to benign words, and known control words. All known neighbors and controls were frequency matched. Results show differences in the correspondence between pronunciation of novel words and known neighbors depending on the relative taboo-ness of the known neighbor. Findings suggest that perceived taboo-ness has top-down influences on reading.


# Effects of Variable Response-Stimulus Interval (RSI) On Sequence Learning 

Sneha Kummetha

Cognitive Science Lab, International Institute of Information Technology, Hyderabad, India
Anuj Shukla
Cognitive Science Lab, International Institute of Information Technology, Hyderabad, India

## Raju Bapi

Cognitive Science Lab, International Institute of Information Technology, Hyderabad, India and School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India


#### Abstract

The objective of this study was to investigate the effects of varying Response-to-Stimulus interval (RSI) on sequence learning by systematically varying it across three different groups (Group1: 0-300ms, Group2: 400-700ms and Group3: 8001100 ms ) and to assess the implicitness and explicitness of the knowledge acquired through such learning. Serial Reaction time task followed by generation task and recognition task were used for this purpose. Results of the SRT task showed learning in all the three groups and the results of the free generation task and recognition task revealed that the sequence learning was implicit in Groups 1 and 2 while it was explicit in Group3. These results were discussed in the context of a recent theoretical framework that proposes conditions in which a switch from implicit to explicit knowledge acquisition is facilitated.


# Object Understanding: Exploring the Path from Percept to Meaning 

Kenneth Kurtz<br>Binghamton University<br>Daniel Silliman<br>Binghamton University


#### Abstract

This research addresses the generation of meaningful interpretations of real-world perceptual stimuli. According to a widespread framework we will call the features-first view, a stimulus is initially encoded via semantically-laden, symbollike properties that are compared to stored category representations to find the best match. Alternative theoretical perspectives challenge the features-first view, but there has been no direct empirical test. In our experiment, participants were shown photographic images of everyday objects and asked to judge as quickly as possible whether a provided verbal descriptive matched the picture. We tested different levels of delay between image and descriptive and found evidence that basic-level category labels were verified faster than clearly manifested descriptions of physical or functional properties. Accordingly, people know the category of the stimulus before knowing its semantic properties. The present evidence suggests that the category is used to achieve a property-level description of the meaning of the stimulus, not vice-versa.


# Effect of Touch-produced Sounds on Surface Texture Perception 

Jinhwan Kwon<br>Department of Informatics, The University of Electro-Communications<br>Suguru Hata<br>Department of Informatics, The University of Electro-Communications<br>Natsumi Komoto<br>Department of Informatics, The University of Electro-Communications<br>Maki Sakamoto<br>Department of Informatics, The University of Electro-Communications


#### Abstract

Texture is an important source of information for distinguishing surface properties. We are able to perceive various textural properties of surfaces from tactile or visual inputs. However, it is unclear how touch-produced sounds influence the various surface texture perceptions. In this study, we examined whether the touch sounds produced by different surface textures influence the various surface perceptions. Consequently, the surface textures with high height and wide interval resulted in rough, bumpy, soft and cool perceptions and the surface textures with the low height and narrow interval resulted in smooth, flat, hard and warm perception. Also, there were statistically significant differences in these measures between two surface texture groups. Furthermore, significantly positive correlations were found in "rough - smooth", "bumpy - flat", "sticky slippery", "wet - dry" and "unpleasant - pleasant" measures between touch-produced sounds and actual touch. This indicate that the touch-produced sounds influence various surface perceptions.


# Identifying Causal Direction in the Two-Variable Case 

Pablo León-Villagrá<br>University of Edinburgh, Edinburgh, United Kingdom<br>Christopher Lucas<br>University of Edinburgh, Edinburgh, United Kingdom


#### Abstract

One of the key characteristics of human cognition is the ability to learn causal structure from data. An influential thread of research into causal learning relies on causal graphical models as a theoretical foundation, and emphasizes the role of prior knowledge, interventions, and statistical independence as tools with which people learn causal structure. What if these sources of information are all absent, as in the problem of identifying causal direction from observations of just two variables? Most work has either ignored this problem or asserted that it is inherently unsolvable. However, recent machine learning algorithms can sometimes infer causal directionality in this setting, by exploiting simple assumptions about the relationship between causes and the noise observed in their effects (Mooij, et al 2016). We investigate whether humans are able to exploit these assumptions or others in order to infer the causal connection between two statistically dependent variables.


# Intolerance to uncertainty is associated with diminished exploration 

Jennifer Lenow<br>Department of Psychology, New York University, New York, NY, USA<br>Nathaniel Daw<br>Princeton Neuroscience Institute, Princeton University, Princeton, NJ<br>Elizabeth Phelps<br>Department of Psychology, New York University, New York, NY, USA


#### Abstract

Across diverse cognitive and behavioral domains, humans confront a fundamental tension between exploiting current knowledge about the environment and exploring the environment in order to acquire new knowledge. Individuals differ idiosyncratically in how they balance this explore/exploit tradeoff, although the sources of these individual differences have not been systematically studied. In the current study, we sought to do so, in terms of trait-level affective phenotypes. Specifically, we investigated whether intolerance to uncertainty (IU), characterized by a negative disposition toward uncertainty, predicted both random and directed exploration in a two-armed bandit task which manipulated decision horizon. We found that greater IU was associated with diminished exploration, both random ( $\mathrm{p}<0.001$ ) and directed ( $\mathrm{p}<0.05$ ). These results suggest the importance of explicitly considering affective states and dispositions in human decision-making and also have psychiatric implications, to the extent that IU is a transdiagnostic dimension central to a range of anxiety-related disorders.


# Interplay between semantic and emotional information in visual scene processing 

Anaïs Leroy<br>Laboratoire d'Anthropologie et de Psychologie Cliniques Cognitives et Sociales, Université Côte d'Azur, France<br>Sylvane Faure<br>Laboratoire d'Anthropologie et de Psychologie Cliniques Cognitives et Sociales, Université Côte d'Azur, France<br>Sara Spotorno<br>Institute of Neuroscience and Psychology, University of Glasgow, UK


#### Abstract

We examined whether and how image's semantics and emotion content interact during visual processing. In each trial, we briefly presented two emotional or neutral images (a scene context and an object), manipulating the semantic consistency and the emotional consistency of the pair. Participants categorised one image semantically or emotionally. Semantic categorisation was overall better than emotional categorisation, but reduced in emotional compared to neutral images, and especially in negative images. Emotional categorisation was better for positive than neutral or negative images; moreover, it was facilitated by emotional consistency and, for accuracy in context images, by semantic consistency. Our results show easiness of semantic compared to emotional categorisation. They suggest that semantic and emotion processes are interdependent, although emotional influence on semantic processing seems stronger than the counterpart, with in particular an interfering effect of aversive images. Conversely, image's attractiveness seems beneficial when evaluating the quality of the emotional content.


# The effect of binaural beats on inhibition 

Kimery Levering<br>Marist College, Poughkeepsie, NY<br>Molly Poinan<br>Marist College, Poughkeepsie, NY<br>\section*{Kristin Jay}<br>Marist College, Poughkeepsie, NY


#### Abstract

A binaural beat is the perceptual experience that occurs when two tones of slightly different frequencies are presented dichotically, creating the experience of a third tone corresponding to the difference in frequencies. Many temporary cognitive effects have been linked to the presentation of a binaural beat, including increased working memory capacity. In the present study, a version of the flanker test was used to investigate the effect of short-term alpha wave binaural beat stimulation on inhibition processes specifically. Participants were presented with 10 minutes of either mid-alpha range binaural beats combined with a recording of waves or only the sound of waves. After this, participants completed a flanker test. The difference between reaction times of congruent and incongruent trials on the flanker task was significantly lower in the binaural beats condition than the wave condition, suggesting that even brief exposure to binaural beats aids in the inhibition of irrelevant stimuli.


# Perception of others: Representation of immigrant groups in newspaper articles 

Ying Li

University of Warwick

Thomas Hills<br>University of Warwick


#### Abstract

Immigrants always have difficulties in integrating into a local society, and sometimes these difficulties come from the fact that they are subconsciously regarded as someone from the outside. How they are represented in the main stream media could reflect people's perceptions towards the 'otherness' of different social members. We analyzed the representation of 29 US immigrant groups in newspaper articles in 2 related studies. The favorability of an immigrant groups is highly associated with its perceived social distance (reflected through usage of concrete language) in our research. To further understand what caused the positive or negative image of immigrants, we applied Latent Dirichlet Allocation to identify topics associated with immigrant groups. We also investigated into how these news topics differ in terms of lingual social distance and favorability. The results provide both qualitative and quantitative insights in how the image of immigrants are reproduced in social media.


# Modelling the dynamics of integrating context into perception: in good and in poor readers 

Itay Lieder<br>ELSC, Hebrew University<br>Vincent Adam<br>Gatsby Computational Neuroscience Unit, UCL<br>Maneesh Sahani<br>Gatsby Computational Neuroscience Unit, UCL<br>Merav Ahissar<br>ELSC, Hebrew University


#### Abstract

Individuals implicitly learn the statistics of environmental stimuli. We used "contraction bias", the tendency to perceive stimuli closer to the estimated mean of similar previous stimuli, to characterize the dynamics of these implicit inference processes. Using a simple auditory discrimination task we found that listeners build a rich representation of the distribution of past stimuli, and yet over represent very recent events. This combined pattern allows both learning of the stable environment, and flexibility to fast changes.

We further characterized populations who have difficulties in acquiring specific expertise, i.e. specific developmental disorders, focusing on reading (dyslexia) and non-verbal communication (high functioning ASD, autism spectrum, individuals) disability, respectively. We found that the pattern of their perceptual inference differs from controls'. Both underweight previous events. However, dyslexics' implicit memory decays fast and they underweight earlier events, whereas ASD individuals underweight recent events. This pattern parallels, and perhaps underlies, their strengths and weaknesses.


# The Effects of Familiarity and Typicality on Naming Objects and Faces 

Yu-Ching Lin<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC<br>Jon-Fan Hu<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC<br>Shu-Ling Peng<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC<br>Shulan Hsieh<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC


#### Abstract

It is found that when we name an object or a face, we often use basic level name (e.g., dog) rather than a name at superordinate level (e.g., animal) or subordinate level (e.g., Labrador). In addition, although abundant evidence generally suggested that both familiarity and typicality influence object recognition, how each of the two factors involves categorization in terms of naming is not fully investigated yet. The present studies were performed to examine the familiarity and typicality effects on naming either an object or a face. Names for basic, superordinate, and subordinate levels were prepared for testing the speed and correctness of object/face identification. As a result, familiarity, not typicality, induced a down-shift pattern for naming. In contrast, typicality led to overall faster responses. The findings of the study indicated that familiarity and typicality have dissimilar effects on categorization by naming.


# The Effect of CSAL AutoTutor on Deep Comprehension of Text in Low-Literacy Adult Readers 

Anne Lippert<br>University of Memphis<br>Breya Walker<br>University of Memphis<br>Raven Davis<br>University of Memphis<br>Qinyu Cheng<br>University of Memphis

## Zhiqiang Cai

University of Memphis

## Danielle Clewley

University of Memphis

## Genghu Shi

University of Memphis

## Arthur Graesser

University of Memphis


#### Abstract

It is well documented that reading strategies of low-literacy readers are suboptimal when text requires deeper levels of comprehension. Deep comprehension demands causal or goal-oriented reasoning and functional conceptual knowledge. Alternatively, shallow comprehension entails recall of definitions or text features without necessitating a coherent understanding of the text. The Center for Adult Literacy (CSAL) AutoTutor is an interactive intelligent tutoring system designed to foster deep and shallow comprehension in low-literacy readers. The present work represents the first empirical study of the effect of CSAL AutoTutor on comprehension type in low- and high-literacy readers. Community members and students interacted with CSAL Autotutor and then were assessed on recall ability for the structure (shallow) and meaning (deep) of sentences from lesson text. Preliminary analysis suggests CSAL AutoTutor promotes comparable deep level comprehension in low- and high-literacy readers. Implications for CSAL AutoTutor as a literacy intervention and future goals are discussed.


# Adaptive response priors in context-dependent decision-making 

Olga Lositsky<br>Princeton University, Princeton, NJ, USA<br>Michael Shvartsman<br>Princeton University<br>Robert C. Wilson<br>University of Arizona<br>Jonathan D. Cohen<br>Princeton University


#### Abstract

Context (such as our location or current goal) informs everyday decisions, both by predicting stimuli and determining relevant responses. How do we develop priors that are general enough to apply in various contexts yet specific enough to maximize reward in a given context? We investigated this using the AX-CPT, a task in which a cue determines which button to press for a probe that appears seconds later. We manipulated the frequency of the probe given the cue across participants and built a diffusion model to estimate how the cue informs participants' priors for the decision. We found that participants' context-dependent priors were closer to each other and less extreme than those predicted by a model that maximizes reward rate given the true stimulus frequencies. However, participants' priors were optimal given their subjective frequency estimates, which showed that they averaged response probabilities across cues when the cues made sufficiently similar predictions.


# Comprehension of Chinese Classifiers in Preschool Children 

Yu-Han Luo<br>National Cheng Kung University, Tainan City, Taiwan, R.O.C<br>Jon-Fan Hu<br>National Cheng Kung Univeristy, Tainan City, Taiwan, R.O.C


#### Abstract

The present research aimed to investigate children's comprehension of Chinese classifiers. Sixty-five Chinesespeaking children between the ages of 4 and 6 recruited in Taiwan participated in the experiment. The results indicate that children can make generalization based on their understanding of classifiers instead of solely relying on classifier-noun associations. The results also show that the participants performed equally on both shape-based and feature-shared classifiers, which suggests that children not only use shape salience to learn Chinese classifiers, but are also sensitive to other relations between objects classified by the same Chinese classifier. Besides, the complex patterns in the results imply that in spite of the exposure to classifiers, the semantic transparency between classifiers and objects varies considerably in both semantic types of classifier, which might be the primary reason that some classifiers are more difficult for children to acquire.


# Cumulative response probabilities: Estimating time course of lexical activation from single-point response times 

Sahil Luthra<br>University of Connecticut, Storrs, Connecticut, USA<br>James Magnuson<br>University of Connecticut, Storrs, Connecticut, USA


#### Abstract

An aim of research on spoken word recognition is to characterize the influence of various lexical characteristics (e.g., word frequency, neighborhood size) on lexical access. Dynamics can be coarsely estimated from single-point measures like naming or more directly assessed using time course measures like fixation proportions over time in the visual world paradigm (e.g., Tanenhaus et al., 1995). We propose that cumulative response probabilities (CRPs) over time may allow a new characterization of the activation dynamics of lexical access from single-point measures. We assume that the timing of responses in a naming task reflects probabilistic sampling of underlying continuous activation dynamics that can be recovered by CRPs. We applied CRP analyses to visual word recognition data collected for 40,481 words from 472 participants (the English Lexicon Project; Balota et al., 2007) and report initial efforts to validate this new approach.


# Guardian and Daily Mail Readers' Implicit Attitudes to Immigration 

Dermot Lynott<br>Lancaster University, UK<br>Michael Walsh<br>University of Stuttgart, Germany<br>Tony McEnery<br>Lancaster University, UK<br>Louise Connell<br>Lancaster University, UK<br>Liam Cross<br>University of Buckingham, UK


#### Abstract

The implicit association test (IAT) measures bias towards often controversial topics (race/religion), while newspapers typically take strong positive/negative stances on such issues. In a pre-registered study, we developed and administered an immigration IAT to readers of the Daily Mail (typically anti-immigration) and Guardian (typically pro-immigration) newspapers. IAT Materials were constructed based on co-occurrence frequencies from each newspapers' website for immigration-related terms (migrant) and positive/negative attributes (skilled/unskilled). Target stimuli showed stronger negative associations with immigration concepts in the Daily Mail corpus compared to the Guardian corpus, and stronger positive associations in the Guardian corpus compared to the Daily Mail. Consistent with these linguistic distributional differences, Daily Mail readers exhibited a larger IAT bias, revealing stronger negative associations to immigration concepts compared to Guardian readers. This difference in overall bias was not explained by other variables, and raises the possibility that exposure to biased language contributes to biased implicit attitudes.


# How and when does the syllable become a reading unit? Developmental evidence in French children 

Norbert Maïonchi-Pino<br>Université Clermont Auvergne - LAPSCO - CNRS UMR 6024, Clermont-Ferrand, France<br>Virginie Loiseau<br>Université Clermont Auvergne - LAPSCO - CNRS UMR 6024, Clermont-Ferrand, France


#### Abstract

French beginning readers might rely on syllables during reading acquisition. However, no in-depth developmental study has been carried out to determine how and when the syllable becomes this prelexical and segmental unit used in the time course of reading acquisition. We recruited 800 French-speaking children distributed in grade 1-5. We used a lexical decision task in a visual masked priming paradigm and a visual identification task. We manipulated the initial syllable frequency, the initial bi/trigram frequency, and the initial syllable structure (CV; CVC). Our main results describe a clear developmental course. The syllable-based effects are early (G1) and sustainably observed (G5), and primarily depend on the syllable frequency. From G2, we found the systematic, automatic use of the syllable as prelexical and segmental unit but the syllable frequency has facilitatory syllable-based effects in the task with lexical access, while it has inhibitory effects in the task without lexical access.


# Epistemically Suspect Beliefs can be partly explained by individual's propensity towards contradiction 

Yoshimasa Majima<br>Hokusei Gakuen University, Sapporo, Hokkaido, JAPAN


#### Abstract

Studies on epistemically suspect beliefs (ESB) have suggested that individual's analytic cognition suppresses unwarranted beliefs, however, our previous studies also showed that an inhibitory effect of analytic cognition was higher among Westerners than Easterners. Rather, intuitive cognition was a common predictor of beliefs between two cultures. Among several cultural differences in cognitive style, we suspect that tendency towards dialectic thinking, i.e., tolerance for contradiction may contribute cultural differences on ESB. The present study aimed to explore this possibility and investigated the association between beliefs and other cognitive measures including individual's cognitive abilities, thinking dispositions, personality traits and propensity towards dialectic thinking. The results showed that the ESB resulted from our intuitive cognition for the most part, and that the effect of culture diminished whilst controlling individual's tendency towards dialectic thinking and style of causal cognition. The cultural difference in a relationship between beliefs and cognitive style was discussed.


# Practicing an auditory working memory task recruits lower-level auditory areas in a task-specific manner 

Tamar Malinovitch<br>Hebrew University of Jerusalem<br>Philippe Albouy<br>Montreal Neurological Institute, McGill University, Montreal<br>Merav Ahissar<br>Hebrew University of Jerusalem<br>Robert Zatorre<br>Montreal Neurological Institute, McGill University, Montreal


#### Abstract

We studied the impact of the trained auditory task on the pattern of behavioural improvement, and its relation to the underlying neural mechanisms. Specifically, we asked whether training with tone retention and manipulation (working memory, WM) transferred to pitch discrimination and vice versa, and whether training modified the brain areas that underlie task performance. Training substantially improved performance, but did not transfer across tasks, even when using the same stimuli. Pre and post training fMRI scans revealed that WM training enhanced activity in bilateral auditory cortices, but not in frontal areas that are initially associated with higher cognitive functions. These results suggest that training-induced improvement is associated with back-tracking along the reverse hierarchy in a task specific manner, as predicted by the Reverse Hierarchy Theory of perceptual learning (Ahissar \& Hochstein, 2004). Thus, low-level areas are recruited, but there is no general upgrade in WM or in auditory skills.


# Do Relationality and Aptness Influence Conventionalization? 

Francisco Maravilla<br>Northwestern University, Chicago, IL, United States

Dedre Gentner

Northwestern University


#### Abstract

The conventionalization of figurative comparisons is one source of lexical evolution. For example, anchor once only meant a device for mooring a ship, but may now be used to describe any source of stability or confidence. Our goal is to understand this process. Following the Career of Metaphor framework, figurative mappings are interpreted through a structure-mapping process, rendering common structure salient. As figurative terms become conventionalized, (1) the figurative sense becomes associated with the base term; and (2) there is shift from simile form to metaphor form. In two studies we investigated psycholinguistic properties that may influence this process: relationality and aptness. We use relative preference for the metaphor form as an estimate of degree of conventionalization; by determining the preferred form for a set of figuratives, we find evidence that both aptness and relationality influence this process. We speculate that figurative comparisons may give rise to new relational terms.


# Mutual Exclusivity Revisited - When Pragmatics overrides Novelty 

Hanna Marno<br>Central European University, Budapest, Hungary<br>Dan Sperber<br>Central European University, Budapest, Hungary; Institut Nicod, Paris, France


#### Abstract

Children typically apply a novel label to a novel object, rather than to a familiar object; a phenomenon called Mutual Exclusivity (Markman et al., 2003). A recent explanation is that children tend to associate novel stimuli together (Horst et al., 2011). We show that pragmatic factors may override novelty. In our study two-year-old children first played with a novel object together with E1. Then E1 left the room and E2 brought another three novel objects for the child to manipulate on his/her own. Finally, E1 came back and requested the child to give her the 'Bitye'. Most children chose the first object, with which they had a common history with E1, even though it was the least novel. This suggests that children understand a novel word by considering to which object the speaker is most likely to have intended to refer.


# Search Your Feelings, Luke: Emotional Fluency Predicts Well-being and Emotional Intelligence 

Masitah Masitah<br>University of Warwick

Ying Li
University of Warwick

Thomas Hills<br>University of Warwick


#### Abstract

How we feel reflects a combination of recalled and recognized emotions. All existing self-report measures are based solely on recognized emotions. To understand the influence of recalled emotions, we developed a new method to recover human emotional states based on emotional free association, in a task we call the emotional fluency. The present work investigated the differences between recall and recognition in human emotional states. We compared the emotional fluency task with self-report measures, including PANAS, WEMWBS, and the Emotional Intelligence Scale. Using language statistics computed from the emotional fluency task, we developed multiple models for predicting self-report measures. We find that while recalled emotions can predict recognized emotions, they highlight important problems with existing recognition measures, including emotional coverage and the difference between availability and accessibility. We also investigate the search process in emotional memory, supporting the role of unbiased memory sampling and higher emotional intelligence and mental well-being.


# Children's reasoning about data sets 

Amy Masnick<br>Hofstra University, Hempstead, NY, US<br>Bradley Morris<br>Kent State University, Kent, Ohio, USA

## Christopher Was

Kent State University


#### Abstract

When reasoning about several numbers, past work has shown that adults mentally summarize data sets and reason based on set characteristics such as the mean and variance (Morris \& Masnick, 2015). In the current study, we asked 10- and 12 -year-old children to look at two columns of numbers (framed as the distances two golfers drove a golf ball, when doing so repeatedly), and to choose which golfer hit the ball farther. We examined reaction time, accuracy, and eye movements, in addition to self-reported strategy use. We found children reasoned using some of the same summary characteristics as adults, though less consistently, and had more varied strategy uses. For example, some children focused only on one number in each set, a pattern not seen in adults. These findings suggest that instruction building on these intuitions may help develop children's numerical cognition skills.


# Attention Modulation Effects on Visual Feature-selectivity of Neurons in Brain-inspired Categorization Models 

Saeed Masoudnia<br>Cognitive Robotics Lab, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran.

Abdol-Hossein Vahabie<br>School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.<br>Majid Nili Ahmadabadi<br>Cognitive Robotics Lab, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran. School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.<br>Babak Nadjar Araabi<br>Cognitive Robotics Lab, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran. School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.


#### Abstract

Most Brain-inspired Visual Object Recognition Models(BVORMs) do not consider local and global reciprocal connections in visual pathway. We addressed this weakness and implemented an attention modulation mechanism based on feedback connections in BVORMs, where feature-selectivity is shaped and modulated by categorization of objects based on their visual features. This modification is inspired by the top-down neuromodulatory signals that make changes in post-synaptic activities of the feature-selective neurons. We also incorporated an implicit memory unit in BVORMs to accumulate recent Hebbian synaptic plasticity's of the neurons in each task. This mechanism guides the top-down feature-based attention modulation to retrieve the interrelated feature-selectivity pattern for each task.HMax and CNN models were used as two BVORMs and tested on a visual categorization problem: natural versus artificial objects in CALTECH-256. Based on experimental results, our proposed modifications not only increased their biological-plausibility but also significantly improved their categorization accuracies compared to the original models.


# Simple and Complex Working Memory Tasks Allow Similar Benefits of Information Compression 

fabien mathy<br>Université Côte d'Azur<br>mustapha chekaf<br>Université Côte d'Azur<br>Nelson Cowan<br>University of Missouri, Columbia


#### Abstract

Because complex span tasks were designed to create a demanding concurrent task, the average span is usually lower ( $4 \pm 1$ items) than in simple span tasks ( $7 \pm 2$ items). One possible reason for the higher span of simple span tasks is that participants can take profit of the spare time to chunk a few stimuli into $4 \pm 1$ groups. It follows that the respective spans of these two types of tasks could be equal (at around $4 \pm 1$ ) when regularities are absent. We therefore predicted an interaction between task and chunkability, supporting a single higher span for a simple span task using chunkable items. However, observation of the spans in the non-chunkable vs. chunkable series refuted the idea that chunking is important solely in simple spans. Indeed, information compression processes contributed to performance levels to a similar extent in simple and complex span tasks.


# Causal asymmetry and the intuitive physics of collision events 

Ralf Mayrhofer<br>University of Göttingen<br>Michael Waldmann<br>University of Göttingen


#### Abstract

In the Michotte (1963) launching scenario, an object (X) moves toward a resting object (Y), eventually colliding with $i t$. In the moment of contact, X stops und Y starts moving - creating the strong impression that X caused Y 's motion and that X exerted a force on Y (but not vice versa). These asymmetries contradict the (symmetrical) laws of Newtonian mechanics, which are at the heart of the popular "noisy Newton" theories of intuitive physics. As an alternative, we propose that inferences in physical scenarios operate over pre-Newtonian representations that are based on the asymmetrical concept of impetus, a motive force that keeps objects moving and that is transferred and reflected in object collisions. We present a formal model of impetus and show that, unlike noisy Newton theories, it provides an explanation of asymmetrical judgments. Other related findings can also be modeled (e.g., biases in mass judgments).


# The development of interpersonal regret and its relation to prosocial choice 

## Teresa McCormack

Queen's University Belfast

## Brian Uprichard

Queen's University Belfast


#### Abstract

We examined whether children feel regret when their failure to make a prosocial choice negatively affects a peer. Five-to-six-year-olds and 7-to-9-year-olds played a game in which they completed a sticker sheet to win a prize. Children then decided whether to donate a spare sticker to another child; most children did not donate. Children discovered that the next child did not have enough stickers to win a prize, and rated their emotions. At this point, children did not know whether the next child could have been able to win the prize if they had donated the sticker in question. This counterfactual information was then provided, and children rated whether they felt happier, sadder, or the same as before. Only the 7-to-9-year-olds' responses suggested that they experienced interpersonal regret. We also showed that experiencing interpersonal regret in the sticker task resulted in children making more prosocial choices in a separate task.


# Interactive and embodied repair: Displaying, recognizing, and negotiating misalignment in an emerging language context 

Ashley Micklos<br>Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands


#### Abstract

As problems of understanding arise in conversational interaction, we must find a means to indicate to our interlocutor the reason for our misunderstanding. However, we are simultaneously constrained by social interactive practices that limit face threat and adhere to epistemic rights. Thus, the challenge is to communicate our own misunderstanding - as specifically as possible - while avoiding explicitness. This challenge may be increased in contexts of language emergence in which alignment is necessary to promote communicative efficiency and conventionalization. Participants in novel communication tasks relied on certain gesture-driven other-initiated repair strategies to gain interactive alignment. The embodied display of cognitive and interactive misalignment cues the interlocutor to repair in a way that reflects their own understanding of the repair initiation and trouble source. The breakdown of intersubjectivity - and its subsequent re-building - is observed in the negotiation of evolving signal-meaning matches through interactive repair sequences.


# Misalignment increases abstraction of referring expressions 

Gregory Mills<br>University of Groningen, Groningen, Netherlands<br>Gisela Redeker<br>University of Groningen


#### Abstract

A central finding in dialogue research is that interlocutors rapidly converge on referring expressions which become progressively contracted and abstract. However there is currently no consensus on which mechanisms underpin convergence: The interactive alignment model (Pickering and Garrod, 2009) favours alignment processes, the grounding model (Clark, 1996) prioritizes positive feedback, while Healey (2002) demonstrates the importance of miscommunication in identifying differences of interpretation.

To investigate convergence we report a variant of the maze-task in which both participants are given misaligned instructions: One participant is primed with instructions that conceptualize the maze as consisting of horizontal vectors (e.g ." 4 th row, 2nd square"); the other is primed with instructions that conceptualize the maze as consisting of vertical vectors (e.g ."3rd column, 2nd square"). Compared with a baseline, misaligned dyads converged on more abstract referring expressions. We argue this pattern is due to participants interactively combining their perspective with that of their partner.


# The spontaneous creation of systems of conventions 

Jennifer Misyak<br>University of Warwick, Coventry, UK<br>Nick Chater<br>University of Warwick, Coventry, UK


#### Abstract

We used a non-linguistic experimental paradigm to explore the instantaneous creation and adaption of novel communicative systems of conventions. Groups of participants played a computer game, in which they sent and interpreted minimal signals to obtain shared rewards within a virtual scene. Within groups, trials manipulated the space of possible signals that could be sent, and the set of meanings to be expressed (the range of cases for the locations and quantities of rewards). Between groups, initial conditions were manipulated through early exposure to different sets of communicative cases.

We observed participants spontaneously develop systems of conventions that were adapted to the full range of signal-meaning mappings encountered. Groups favoured systems optimised to their particular initial learning environment. These systems become entrenched and transferred to new signal-mapping environments to which they were not adapted.


# Refining the cognitive semantic web: The tensor method to represent the topographic emplacement of different word categories 

Eduardo Mizraji<br>Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay<br>Andrés Pomi<br>Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay


#### Abstract

A central problem concerning the organization of the cognitive semantic web is to understand how different categories of words are stored in the brain with a well-defined topographical organization. This topography is a natural construction that plausibly is strongly related with the syntactic and semantic organization of natural languages. An eloquent experimental evidence of the existence of a continuous semantic representation of object and action categories in the human brain has been published by Huth et al (Neuron $76: 1210,2012$ ). One of the ways to explain the emergence of a topographical organization in the brain cortex using neurocomputational models, is by means of Kohonen's self-organizing maps. Here we show that these topographies can be operationally represented with associative memories spatially organized by tensor contexts. We illustrate formally and numerically this fact. In addition, we show that, consistently with evidence from pathology, different semantic categories can be specifically damaged.


# Picture book reading in the lives of 18-30 month old children: A diary study 

Jessica Montag<br>University of California, Riverside

Linda Smith

Indiana University


#### Abstract

Picture book reading is a common activity in the lives of many children. This work describes the frequency and character of picture book reading in American homes. Seventy-seven monolingual English-speaking families with children between the ages of 18-30 months took part in a 5 -day diary study in which caregivers recorded details about picture book reading activities. This sample is characteristic of some laboratory samples but less nationally representative; $92.2 \%$ of caregivers held a college degree. Relative to previously reported averages, caregivers reported reading to children more often ( $6.8 \mathrm{x} / \mathrm{day}$ ), reported beginning reading at a younger age ( 2.2 months) and reported more books in the home ( 111.1 books). Caregivers reported both reading the book text and discussing the pictures with their children. These numbers suggest an extremely high upper-limit to the amount of language input some children receive from picture book reading. Consequences for language environments and language development will be discussed.


# Resemblance among similarity measures in semantic representation 

Maria Montefinese<br>Department of Experimental Psychology, University College London, London, UK<br>David Vinson<br>Department of Experimental Psychology, University College London, London, UK


#### Abstract

Similarity measures, extent to which two concepts have similar meanings, are key to understanding how concepts are represented, with different theoretical perspectives relying on very different sources of data from which similarity can be calculated. Experiential/embodied theories use verbal features or property ratings; distributional/relational ones use cooccurrence. Similarity may also be estimated from tasks like word-association, priming, and other rating studies. Often the different theoretical perspectives are placed in opposition; here we test the extent to which similarity representations based on different measures converge. We used Representational Similarity Analysis and multidimensional scaling on 31 similarity measures. Similarity in age-of-acquisition and word-length were related to similarity in naming and priming; and affective similarity and co-occurrence were also related. More importantly, representational resemblance was shown among embodied, distributional and association-based representations, demonstrating that different data sources are employed in a similar way in building meaningful conceptual representation.


# Effects of Question Format on Test-Taker Cognition 

Jung Aa Moon<br>Educational Testing Service<br>Irvin Katz<br>Educational Testing Service<br>\section*{Madeleine Keehner}<br>Educational Testing Service


#### Abstract

Technology-based, interactive test questions are common in large-scale assessments, yet how alternative question formats influence test-taker cognition is not well understood. In a series of studies, we investigated test-taker performance on isomorphic questions using alternative presentation layouts and modes of responding. Adult participants solved math problems in three formats, each of which regularly appear in many large-scale assessments: 1) forced-choice (explicit True-False options) presented in a table format, 2) check-all-that-apply (implicit True-False options) presented in a table format, and 3) check-all-that apply presented as separate questions. Participants' solution time and affirmative selection rate suggested different cognitive processes for the question formats, particularly when they were uncertain of their answers. We propose a cognitive model to account for the results and predict the impact of alternative question formats on test-takers. We discuss how principles of cognitive science and human-computer interaction provide direct implications for designing assessment questions and understanding test-taker cognition.


# Magnitude of metaphor and its effect on reasoning about immigration 

Karie Moorman<br>University of California, Merced<br>Teenie Matlock<br>University of California, Merced


#### Abstract

Metaphor is replete in discourse about immigration. Recent work shows that metaphoric framing can influence attitudes toward immigration (e.g., Landau et al., 2009). However, we know little about how and when specific information in the source domain drives this effect. Our study takes a novel approach, examining how varying intensity of information in the source domain frame influences attitudes toward immigrants and immigration in the U.S. We analyze various metaphors but we focus especially on intensity effects in the conceptual metaphor IMMIGRATION IS FLUID TRANSFER. For the FLUID TRANSFER source domain, we investigate how varying intensity of flow (e.g., rate) influences attitudes about immigration, including whether immigrants should have access to social services and what type of wall should be built, if at all. Our results make a valuable contribution to metaphor research by revealing what information within the source domain has the most (or least) robust effects on reasoning.


# Investigating the Impact of Sleep on Eyewitness Memory 

David Morgan<br>Royal Holloway, University of London<br>Jakke Tamminen<br>Royal Holloway, University of London<br>Laura Mickes<br>Royal Holloway, University of London


#### Abstract

How sleep impacts the accuracy of identifications that eyewitnesses make from lineups is unknown. For a comprehensive understanding of eyewitness performance, two types of eyewitness ID accuracy are considered: discriminability (the ability to distinguish innocent from guilty suspects) and reliability (the probability that the identified suspect was the offender). The well-known role sleep plays in memory consolidation should apply to an eyewitness's ability to discriminate, but not necessarily their reliability. That is what we investigated in a large-scale forensically-relevant experiment. We compared discriminability and reliability from sleep (sleep occurs between witnessing a crime and lineup test) and wake (remains awake between crime and lineup) conditions. Furthermore, theorists have long been using signal-detection models to understand recognition memory, but its use is new to the field of eyewitness ID research. Thus, we compared signal-detection models with different decision rules. Our findings shed light on the impact sleep has on eyewitness IDs.


# Impact of testimony and prior knowledge on children's beliefs about category homogeneity 

Kelsey Moty

Lehigh University

Amanda Brandone<br>Lehigh University


#### Abstract

Previous work has shown that preschoolers-in comparison to older children and adults-tend to view categories as homogeneous, generalizing properties of individuals broadly to all category members (e.g., this dax has wings, so all daxes do). Here, we explore whether the testimony used to describe category individuals as well as children's prior knowledge of categories attenuates their homogeneity expectations. Using a novel induction task, 4 to 7 -year-olds were asked to predict the distribution of properties among members of familiar/unfamiliar animal categories based on a single exemplar. Exemplars were introduced as "special" to half of participants. Preliminary findings $(\mathrm{N}=71)$ suggest that prior knowledge may contribute to beliefs about category homogeneity: responses for familiar animals varied appropriately given the real-world prevalence of each property whereas children overestimated the property's prevalence for unfamiliar animals. The complete dataset will speak to how language choice in testimony shifts children's beliefs about homogeneity.


# Learning to Consider Alternative Causes: Can Practice Make Us More Aware of Our Imperfection? 

Edward Munnich<br>University of San Francisco

Ana Maria Hoffmann<br>University of San Francisco

Nancy Ortega
University of San Francisco
Emma Weinberger
University of San Francisco

Dana-Lis Bittner

University of San Francisco
Jacqueline Tiongco
University of San Francisco
Jasmine West
University of San Francisco


#### Abstract

In hindsight bias, upon learning an outcome, one is overly confident that one would have "known it all along." Several researchers have been able to neutralize hindsight bias by prompting participants to consider alternative outcomes, but can we learn to avoid bias for novel outcomes, without prompting? Foresight participants read brief summaries of five psychology studies, and learned the mean performance of one group in each study. They estimated the other group's perfor-mance-reflecting their sense of the effect size—stated possible causes, and then learned the other group's mean performance. Hindsight participants learned both groups' mean performance at the outset, then indicated what they would have estimated. We asked whether (1) participants would show superior estimation and/or consideration of alternative causes for novel stimuli one week later, and (2) whether Foresight participants would benefit more given the feedback they received on the accuracy of their estimates.


# Bidirectional effect of emotional contagion for pain during face-to-face interaction 

Aiko Murata<br>Waseda University<br>Tatsuya Kameda<br>The University of Tokyo

Katsumi Watanabe<br>Waseda University


#### Abstract

The automatic contagion of emotion is considered crucial in interpersonal communication. In face-to-face interactions, people could be both the receiver and sender of emotional content. Thus, contagion may have bidirectional influences on the emotional states of individuals. However, many studies have mainly dealt with unidirectional contagion, such that the expression of pain in a target entails a reaction of pain in the observer. In this study, we demonstrated bidirectional emotional contagion in the experience of thermal pain during interaction. Firstly, we showed that the physiological responses of dyad members were correlated with each other when they could interact compared to when they were impaired to see each other. Further, we demonstrated that individuals showed higher or lower physiological responses when their partners experienced stronger or weaker stimuli respectively. Thus, people can develop similar physiological responses through interactions, and this effect seems to induce a change in the responsivity to stimuli.


# Agent's symmetry elicits egocentric transformations for spatial perspective-taking 

Hiroyuki Muto<br>Graduate School of Human Sciences, Osaka University Japan Society for the Promotion of Science<br>Soyogu Matsushita<br>Graduate School of Human Sciences, Osaka University<br>Kazunori Morikawa<br>Graduate School of Human Sciences, Osaka University


#### Abstract

Spatial perspective-taking is an ability to understand in which direction an object is located relative to an agent (e.g., another person or a chair). Previous studies showed that left/right judgments prompted an egocentric transformation strategy (i.e., mental rotation of the self) whereas front/behind judgments prompted other strategies (e.g., tracing a line of sight). To examine whether the symmetrical shape of an agent could affect the choice of strategies, we used as an agent a cuboid which has a prong on one of its sides. We labeled the prong side as the front (Experiment 1) or right (Experiment 2) of the agent, about which participants made left/right and front/behind judgments. The results revealed that egocentric transformations were more favored for judgments about directions along symmetrical than asymmetrical axes of the agent, regardless of whether the judgment was about left/right or front/behind. This suggests similar processing underlies left/right and front/behind judgments.


# Robot as Moral Agent: A Philosophical and Empirical Approach 

Shoji Nagataki<br>Chukyo University<br>Masayoshi Shibata<br>Kanazawa University

# Tatsuya Kashiwabata <br> Keio University 

Takashi Hashimoto
Japan Advanced Institute of Science and Technology

Takeshi Konno<br>Kanazawa Institute of Technology<br>Hideki Ohira<br>Nagoya University<br>Toshihiko Miura<br>University of Tokyo<br>Shinichi Kubota<br>Kanazawa University


#### Abstract

What is necessary for robots to coexist with human beings? In order to do so, we suppose, robots must be moral agents. To be a moral agent is to bear its own responsibility which others cannot take for it. We will argue that such an irreplaceability consists in its having an inner world - one which others cannot directly experience, just as pleasure and pain. And personality of a moral agent, which is to be irreducible to a mere difference of traits or features of individuals, is firmly rooted in such an inner world.

We will support our theses by referring to our experiment in which humans and robots interact with each other doing a coordination task. This experiment will provide an empirical analysis of the human-robot relationship with regard to learning mechanism, moral judgement, and the ascription of the inner world.


# Iconicity vs. Systematicity in Artificial Language Learning 

Alan Ks Nielsen<br>Max Planck Institute for Psycholinguistics<br>Julia Simner<br>University of Sussex<br>Simon Kirby<br>University of Edinburgh<br>\section*{Kenny Smith}<br>University of Edinburgh


#### Abstract

A foundational assumption in linguistics has been that words and their meanings are arbitrarily related; however, this position has been challenged recently. Experiments have shown that both systematic (where similar objects have similar labels) and iconic (words 'resemble' the objects they label) associations between words and objects facilitate learning. However, these two literatures remain confounded: the degree to which increased learnability is driven by iconicity rather than systematicity has not been disentangled. Here we present the results of two studies testing the differences in learnability between artificial lexica that are either conventionally systematic, or both systematic and cross-modally iconic. In the first study we find that both conventional and iconic systematic lexicons are equally learnable, but iconic mappings provide an early learnability advantage. In the second study we find that the presence of sound-symbolic associations for one dimension can interfere with the learning of conventional associations on another dimension.


# The space and time of contamination: Complete, continual, spreading effects 

Laura Niemi<br>Harvard University, Cambridge, MA, USA<br>Liane Young<br>Boston College, Chestnut Hill, MA


#### Abstract

People sometimes report feeling "totally" different (complete affectedness) and that "they'll never be the same again" (continual affectedness) after negative events. It's been proposed that complete, and continual negative effects characterize contamination or impurity. Meanwhile, whether impurity is a legitimate moral domain apart from harm has been debated in moral psychology. We address these matters using novel approaches from cognitive linguistics. First, according to a prominent theory of verb semantics, verbs that convey impurity (contaminate, taint) belong to a class that implies complete affectedness (the "fill" class), such that contaminated entities are seen as completely contaminated. Second, people rated perpetrators equally, and highly, "contaminated", "contaminating", and "injuring", whereas victims were rated straightforwardly "injured" (Turk; $\mathrm{n}=126$, replicated twice). For "contaminated" perpetrators, the taint carried on - they were continually "contaminating". In sum, impurity is distinct from harm: the process underlying impurity, contamination, involves inferences of complete, continual negative effects that spread.


# Construction of design activity index based on the value of artifact 

Kunio Nikata<br>Kanazawa College of Art<br>\section*{Kentaro Inomata}<br>Kwansei Gakuin University<br>Toru Sato<br>Kwansei Gakuin University<br>Keigo Kawasaki<br>Kanazawa College of Art<br>Noriko Nagata<br>Kwansei Gakuin University


#### Abstract

Digital personal fabrication refers to the creation of products using ICT-tools by individuals. In order to support the users of such tools who are untrained in design, it is necessary to develop a support system that makes it possible to use expert knowledge on designing products. For this study, we selected 26 items from the values that are considered important for design (e.g., unique, modern; Inomata et al., 2016), and investigated the design activities for realizing these values. Eighty professionals in design participated in the survey. Many design activities concerning shapes and colors were observed as ways to realize the values. In addition, various activities such as improvements on materials and motifs or advices to satisfy the practical design activity were observed. We created an index of the frequently used activities to realize each value and discussed its potential as an actual design support tool.


# Indirection Explains Flexible Tuning of Neurons in Prefrontal Cortex 

David Noelle<br>University of California, Merced


#### Abstract

The prefrontal cortex (PFC) is broadly seen as supporting cognitive flexibility - quickly adapting behavior in response to changing circumstances. Some PFC neurons appear to actively maintain rule-like information associated with the current task, with firing changing with task context. Some PFC neurons, however, have been found to exhibit activity related to specific stimulus features or action options, but the tuning of these neurons appears to dynamically change with task shifts (Duncan, 2001). Short-term synaptic plasticity has been proposed as the primary mechanism for rapidly adapting the response profiles of these cells. Using a computational cognitive neuroscience model of hierarchical structure in PFC (Kriete, Noelle, Cohen, \& O'Reilly, 2013), an alternative account is offered in which flexible neural tuning arises not from fast synaptic change but from a frontal representational scheme involving neurons that encode references to other PFC areas rather than directly encoding task relevant sensory/motor information.


# Five-Year-Old Children Transfer a Metacognitive Strategy to a Novel Task 

Allison O'Leary<br>The Ohio State University<br>Vladimir Sloutsky<br>The Ohio State University


#### Abstract

Previous work has demonstrated that interventions like 1) giving in-the-moment performance feedback and 2) providing a strategy rule can improve children's metacognitive learning. However, there is little evidence to suggest that this learning transfers to a novel task. We trained 5-year-olds' metacognitive control in a task requiring participants to select the easier of two games to acquire the highest amount of points. Compared to a control group who received no training, children who were trained to control behavior (by selecting an easier dot discrimination task) showed greater evidence of transfer to a novel task (by selecting an easier line length discrimination task). This suggests that the learned strategy rule (i.e., to select an easier task) was not stimulus-specific, and was abstract enough to apply to a novel task with new stimuli. In sum, 5 -year-olds were able to learn a strategy rule and spontaneously apply the strategy to a novel task.


# Do forgiving God primes strengthen support for state sanctioned punishment? 

Katherine O'Lone<br>Royal Holloway, University of London, LONDON<br>Ryan McKay<br>Royal Holloway, University of London


#### Abstract

Do forgiving God primes strengthen support for state-sanctioned punishment? Laurin et al (2012) found that beliefs in powerful, intervening Gods (both in general and when made salient) reduce people's endorsement of state-sanctioned punishment. In light of this, we investigated whether the manner in which God intervenes (via forgiveness or punishment) influences people's endorsement of state-sanctioned punishment.

Across four studies we explored a) whether priming participants with a forgiving God and b) whether salient, forgiving God beliefs increase endorsements of state-sanctioned punishment. The rationale being that a forgiving God will lead people to view punishment as a responsibility that lies with them rather than one outsourced to God. Our results revealed no evidence for effects of forgiving God primes or salient forgiving god beliefs on endorsements of state-sanctioned punishment. We discuss the implications of these findings for extant theories of religious prosociality and proportionality-based accounts of morality.


# Effects of motives of search and prior experiences on online browsing performance: Considerations from searchers cognitive load 

Kayoko Ohtsu<br>Waseda University<br>Takako Sakawaki<br>Waseda University


#### Abstract

The present study aimed to develop effective education methods of online search for unskilled college students. In the preparatory stage of the study, an experiment using simple browsing tasks was conducted to examine the effects of important factors of searching focusing on cognitive load. Under two conditions (Casual and Formal) promoting different motivations, search result lists were displayed to fifty-nine college students to look for two types of information: seeking statistical data (task A) and seeking views and opinions to answer open questions (task B). Analyses of each task using two factors (the conditions and their presentation orders) revealed that in task A, only when the Casual condition was first, the participants performed better in the Formal condition. In task B, only when the Formal condition was first, browsing time in the Casual condition was shorter. We assume that these effects are associated with the workload of browsing.


# Anticipatory Active Inference from Learned Recurrent Neural Forward Models 

Sebastian Otte<br>University of Tübingen, Tübingen, Baden-Württemberg, Germany<br>Theresa Schmitt<br>University of Tübingen, Tübingen, Baden-Württemberg, Germany<br>Martin V. Butz<br>University of Tübingen, Tübingen, Baden-Württemberg, Germany


#### Abstract

We demonstrate that inference-based goal-directed behavior can be done by utilizing the temporal gradients in recurrent neural network (RNN). The RNN learns a dynamic sensorimotor forward model. Once the RNN is trained, it can be used to execute active-inference-based, goal-directed policy optimization. The internal neural activities of the trained RNN essentially model the predictive state of the controlled entity. The implemented optimization process projects the neural activities into the future via the RNN recurrences following a tentative sequence of motor commands (encoded in neurons akin to recurrent parametric biases). This sequence is adapted by back-projecting the error between the forward-projected hypothetical states and desired (goal-like) system states onto the motor commands. Few cycles of forward projection and goal-based error backpropagation yield the sequences of motor commands that control the dynamical systems. As an example, we show that a trained RNN model can be used to effectively control a quadrocopter-like system.


# Application of fuzzy logic in dyslexia user modelling to design customizing assistive technology 

Tereza Pařilová

Masaryk University

## Eva Hladká

Masaryk University


#### Abstract

Cognitive psychology studies phenomena that cannot be directly observed. Scientific knowledge about the brain is extensive, but there is still a lot to be understood about its functions. Cognitive functions are weakened in dyslexic children; this is reflected in highly individual problems regarding the reading skills. Reading is a process which consists of decrypting graphic characters (perceptual level) and understanding the meaning of words (cognitive level). These levels cannot be separated. An approach - fuzzy logic - is used in order to address this issue and create a model of the dyslexic user, based on which technologies can be individually tailored to a particular dyslexic. We discusse the possibilities of the use of the mathematical apparatus for the categorisation of users with regard to their "black box". Further, we focuse on the development of new assistive technologies targeted at specific attention disorders, reading disorders, as well as information processing disorders.


# An Exploratory Study of the Influence of Pretend Play on Children's Self-Regulation and Language Skills 

Tanya Paes<br>University of Cambridge, Cambridge, Cambridgeshire, United Kingdom<br>Michelle Ellefson<br>University of Cambridge


#### Abstract

Recently, there has been increased interest regarding how pretend play contributes to children's cognitive development. This study examines the efficacy of a pretend play intervention on self-regulation and language skills of 4-to 5-year-olds and explores parents' perceptions about children's engagement in pretend play. The small-scale intervention includes eight 30minute sessions over 6 weeks, in groups of five children. Each session included: (1) shared storybook reading; (2) role-playing; and (3) review. During shared story-book reading the children were read two books with explicit phonological awareness and vocabulary instruction for 18 words in each book. Role-playing included providing the children with props, which allow for engagement in pretend play activities. Several measures were used pre- and post-intervention to evaluate children's self- regulation and language skills. The improvements that occurred in the intervention are considered alongside other cognitive and educational factors to better understand the role of pretend play in educational settings.


# Interpreting nonsignificant findings in psychological research 

Bence Palfi<br>University of Sussex, Falmer, United King

Aba Szollosi<br>University of New South Wales<br>Barnabas Szaszi<br>Eotvos Lorand University<br>Marton Kovacs<br>Eotvos Lorand University<br>Mark Zrubka<br>Eotvos Lorand University<br>Peter Szecsi<br>Eotvos Lorand University<br>Balazs Aczel<br>Eotvos Lorand University<br>\section*{Eric-Jan Wagenmakers}<br>University of Amsterdam


#### Abstract

In this study, we examined the current practice and alternative methods for interpreting nonsignificant findings in psychological research. The traditional null-hypothesis testing presents a challenge for researchers to interpret nonsignificant findings. We reviewed the abstracts of all empirical articles published in three high-esteem psychological journals in 2015 and selected those which referred to a nonsignificant result $(\mathrm{N}=134)$. We found that the majority of the statements interpreted the results only within the sample, yet in $23 \%$ the authors inferred from the results to the absence of an effect. Bayes factor analyses on these statistics indicated that the support of these results for the null-hypothesis is strong only in $4 \%$, moderate in $70 \%$ and anecdotal in $26 \%$. The results revealed that Bayes factor analysis can help researchers in interpreting nonsignificant results and also highlight that psychological studies with traditional sample sizes are unlikely to present strong evidence for the null-hypothesis.


# Perceived control in bounded-rational decision-making 

Saee Paliwal<br>Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland<br>\section*{Frederike Petzschner}<br>Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland<br>\section*{Ekaterina Lomakina}<br>Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland<br>Klaas Enno Stephan<br>Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland


#### Abstract

The amount of control perceived by an agent governs their ability to learn. Bounded rationality, or the idea that we are limited by the amount cognitive work we can perform, provides an appealing framework within which perceived control could be formulated. When modeling the world, the bounded-rational agent balances the trade-off between the utility and complexity of this constructed model in order to choose an optimal policy. Here, we present a novel formulation of behavioral control, bounded inference, which explicitly models control as the perceived constraint experienced by an agent during the inference process, employing a version of the free energy functional with an additional boundedness parameter as the variational principle of this constrained optimization. The utility of bounded inference is demonstrated in simulations that capture various characteristics of dysfunctional behavioral patterns as observed in a range of psychiatric disorders for which control beliefs play a central role.


# Applications of Cognitive Science to Enhancing Scholarly Communication 

Purav Patel<br>University of Minnesota - Twin Cities, Minneapolis, Minnesota, USA


#### Abstract

Learning from and building on the accomplishments of scholarly publications is often difficult. To address this challenge, this work leverages well-replicated cognitive science phenomena to promote people's understanding of research found in journal articles. It forms the conceptual groundwork for a digital platform through which users can author and learn from interactive multimedia documents that communicate research more effectively. One of the many recommendations is to reduce the split-attention effect by integrating text and graphics in figures. Doing so may help readers understand complex visuospatial representations. Encouraging active processing via comprehension questions and responsive simulations of experimental procedures embedded in articles may boost learning even more. To promote the creative extension of research, evidence-based brainstorming prompts that trigger analogical reasoning and episodic specificity induction should be adopted. If scholarly communication is centered on scientific principles like these, then the dissemination and dynamics of science may both advance.


# Novel metacognitive problem solving task 

Jwalin Patel<br>Faculty of Education, University of Cambridge<br>\section*{Amanda Aldercotte}<br>Faculty of Education, University of Cambridge

Teresa Parr
Ashley-Parr, LLC
Zewelanji Serpell
Department of Psychology, Virginia Commonwealth University
Michelle Ellefson
Faculty of Education, University of Cambridge


#### Abstract

Metacognition is important for decision making, problem solving and learning. Despite the widespread interest in metacognitive skills and their development, it is challenging to measure metacognitive skills in children. Some excellent qualitative and observational measures exist, but use metrics that are different from traditional metacognition tasks for adults. Some meta-cognition tasks of memory have been developed for children, but these only offer a narrow range of the skills involved in metacognition. Here, we compared performance on a meta-memory task for children with a new task of metacognition for problem solving. Our sample includes about 800 children aged $8-10$ years who were part of a larger study exploring the development of thinking skills. The results indicated similarities and differences between the memory and problem solving tasks, suggesting that the new task could be a bridge between existing qualitative and quantitative measures of metacognition in children.


# Comparing comparison indices: Assessing the validity of different magnitude comparison measures across presentation formats and age groups 

Marcie Penner-Wilger<br>King's University College at Western University<br>Aaron Cecala<br>Elizabethtown College<br>Melissa Elfers<br>King's University College at Western University


#### Abstract

Magnitude comparison tasks are used to assess the precision of numerical representations. Recent research, however, questions the validity of different measures of magnitude comparison. We investigated the validity of five performance measures: overall RT, overall accuracy, numerical ratio effect (RT), numerical ratio effect (accuracy), and Weber fraction. Kindergarten and university students completed symbolic and non-symbolic magnitude comparison tasks and a math skill measure. For children and adults, we calculated Chronbach's $\alpha$ separately for each presentation format. All values were in the unacceptable range, indicating that the different indices were not measuring the same construct. For children, a multiple regression predicting KeyMath scores from symbolic and non-symbolic indices showed that only non-symbolic overall accuracy and symbolic overall RT were predictors. For adults, a multiple regression predicting French Kit scores showed that only the symbolic numerical ratio effect (RT) was a predictor. No index demonstrated predictive validity across formats or age groups.


# Language input and development during a year in an early intervention classroom 

Lynn Perry<br>University of Miami, Coral Gables, Florida, United States<br>Emily Prince<br>University of Miami, Coral Gables, Florida, United States<br>Adriana Valtierra<br>University of Miami, Coral Gables, Florida, United States<br>Camila Rivero-Fernández<br>University of Miami, Coral Gables, Florida, United States<br>Mary Anne Ullery<br>University of Miami, Coral Gables, Florida, United States<br>Lynne Katz<br>University of Miami, Coral Gables, Florida, United States<br>\section*{Daniel Messinger}<br>University of Miami, Coral Gables, Florida, United States


#### Abstract

By the time they are three-years-old, children raised in poverty hear 30 million fewer words than their socioeconomically advantaged peers. This word gap predicts later school readiness outcomes and sets the stage for achievement gaps that can follow the child through life. Although parent speech has become a subject of increasing study and intervention, less is known about speech in childcare settings. We conducted a longitudinal study in an early-intervention classroom for $2-3$-year-old children from low-income, at-risk backgrounds. We examine the relationship between language input from teachers and peers and children's language skills over one year. Results show that vocabulary knowledge influences children's talkativeness in the classroom, and talkativeness and the amount of language they hear positively relates to increases in their language abilities. Our application of automated measurement provides new insight into the dynamics of the classroom language environment and consequences for language development in at-risk children.


# Early Visual Evoked Potentials (VEPs) in Infant Siblings of Children with ASD, ADHD and Age-Matched Controls 

Elena Serena Piccardi<br>Centre for Brain and Cognitive Development, Birkbeck University of London<br>Mark H. Johnson<br>Centre for Brain and Cognitive Development, Birkbeck University of London<br>Teodora Gliga<br>Centre for Brain and Cognitive Development, Birkbeck University of London


#### Abstract

Atypicalities in sensory perception are observed in individuals diagnosed with ASD and ADHD but have rarely been contrasted in experimental studies. In the visual domain, superior performance on visual search tasks and hypersensitivity to flickering lights have been cited as evidence of unusual sensory profiles.

To measure a reliable visual response, black-and-white checkerboards were presented under free-viewing conditions to three groups of 10 -month-olds: siblings of children with ASD ( $\mathrm{N}=47$ ), ADHD ( $\mathrm{N}=21$ ) and controls $(\mathrm{N}=18)$. Continuous EEG was recorded and VEPs time-locked to checkerboards presentation computed.

Analysis of VEPs amplitude and latency revealed statistically significant group differences in the first 200ms post-stimulus onset. Early components were enhanced in amplitude (P100) and delayed in latency (P100-N100) in at-risk infants compared to controls ( $\mathrm{p}<.05$ ).

Atypical VEPs to low-level information might index a domain-general aberration in at-risk populations. The nature of this atypicality will be further investigated by analyzing its association with background EEG.


# How does social touch modulate arousal states? An investigation in early development. 

Laura Pirazzoli<br>Birkbeck College, London, United Kingdom<br>Emily Jones<br>Birkbeck College, London, United Kingdom<br>Teodora Gliga<br>Birkbeck College, London, United Kingdom<br>Mark H. Johnson<br>Birkbeck College, London, United Kingdom


#### Abstract

Caregiver-infant interaction through touch was shown to have long-term effects on child's cognitive development, but the mechanisms are poorly understood. Our aim is to investigate how affective touch (slow gentle caressing) affects arousal states in young infants. Previous work showed that slow-touch decreases heart rate in 9-month-old infants.

We tested two groups of 6-months-old ( $n=26$ ) and 9 -months-old infants ( $n=23$ ). We measured heart rate and saccadic reaction time while infants performed a visual orienting task, where speed of re-orienting from a central fixation to a peripheral target was measured. During the experiment, infants received either slow or fast-touch on their back in blocked trials. We found no effects of touch on heart rate in either age-group, and only marginal effects of slow-touch on reaction times in 9-month-old infants. We are currently testing 2 months-old infants to investigate if these effects are observed earlier in life; these new results will be discussed.


# Hand, spoon or toothbrush? Towards the understanding of the neural underpinnings of affective touch in 5 months-old infants. 

Laura Pirazzoli<br>Birkbeck College, London, United Kingdom<br>Sarah Lloyd-Fox<br>Birkbeck College, London, United Kingdom<br>Mark H. Johnson<br>Birkbeck College, London, United Kingdom<br>Teodora Gliga<br>Birkbeck College, London, United Kingdom


#### Abstract

It is known that affective touch leads to broad cortical activations including posterior STS, key region of the social brain. Our goal is to discover if a similar pattern of activation can be observed in 5-months-old infants, or whether the development of this cortical specialization results from extensive postnatal experience.

Over two studies we used functional-Near-InfraRed-Spectroscopy (fNIRS) to compare social touch (a human caress) to non-social touch (a caress performed with a spoon in study $1-n=22$ - or with an electric toothbrush in study2-n=17-).

In study1 we found similar patterns of activation to the social and non-social stimulus. In study2 we report broad responses to the non-social stimulus, but, to our surprise, we found no activations to the human caress.

In light of these results we conclude that it is possible that at this age discrimination between social and non-social touch in the posterior temporal lobe is still undergoing specialization.


# Do we see things better when we know grammar? 

Fenna Poletiek<br>Leiden University<br>Maartje Van de Velde<br>Leiden University


#### Abstract

Language affects perception. But how? Recent findings (Boutonnet \& Lupyan, 2015; Bocanegra, Poletiek \& Zwaan, submitted) suggest a dissociation between perception that is mediated as compared to not mediated by language. One explanation is that language -that is combinatorial in nature- stresses the separate features of objects. We investigated the effect of combinatorial (two words) and non-combinatorial (one word) labels on the perceptual separation of features in visual recognition. Participants were trained to categorize meaningless objects with two dimensions: shape and height. Each category had either a one word name; or a two words name reflecting its features. Participants then were tested on new objects . Combinatorial labels enhanced categorization performance as compared to single labels. This suggests that language, by decomposing objects into parts, might drive dimension separation in vision as well.


# Patterns of Cortical Activation Correlate With Speech Understanding After Cochlear Implantation 

Luca Pollonini<br>University of Houston<br>Heather Bortfeld<br>University of California, Merced<br>Michael Beauchamp<br>Baylor College of Medicine<br>John Oghalai<br>Stanford University


#### Abstract

Cochlear implantation is a standard intervention for deafness, yet the ability of implanted patients to understand speech varies widely. To better understand this variability, we used functional near-infrared spectroscopy to image auditory cortex activation in response to different classes of sound and compared that to behavioral measures of speech perception. Both control and implanted participants with good speech perception exhibited greater cortical activity to natural speech than to unintelligible speech. In contrast, implanted participants with poor speech perception produced pronounced cortical activation across stimulus classes. Moreover, the ratio of cortical activation in response to normal speech relative to that of scrambled speech directly correlated with their comprehension scores, though not with auditory threshold, age, side of implantation, or time after implantation. Because implanted adults with low speech perception scores produced indistinguishable cortical activation across stimulus classes without preferential response to speech, we interpret this as demonstration of compensatory processing effort.


# Encouraging Fruit and Vegetable Consumption Through Intuitive Theory Building 

Derek Powell<br>Stanford University

Ellen Markman<br>Stanford University


#### Abstract

Although routinely informed of the benefits of fruits and vegetables, Americans eat far short of the recommended amounts. Instead of just telling people that fruits and vegetables are healthy, providing a compelling causal and teleological explanatory framework could increase both people's conviction about their health benefits and commitment to increasing the amount and variety of fruits and vegetables they consume. Our brief intervention: (1) emphasizes that fruits and vegetables have thousands of health-promoting phytochemicals, well beyond just vitamins, (2) describes clear causal mechanisms by which these foods ensure cellular health, (3) draws an analogy between the benefits of plant-based foods and the power of plantderived medicines, and (4) explains that plants contain abundant nutrients because they must manufacture these chemicals for their own survival. This novel intervention improved understanding and increased participants' intentions to eat more fruits and vegetables, illustrating how intuitive theories can shape motivation for behavioral change.


# Scientific Reasoning Ability in Middle Schoolers related to MasterMind Discovery Strategy 

Jean-Baptiste Quillien<br>University of Minnesota, Minneapolis, MN, United States<br>Keisha Varma<br>University of Minnesota, Minneapolis, MN, United States<br>Purav Patel<br>University of Minnesota, Minneapolis, MN, United States<br>Martin Van Boekel<br>University of Illinois at Urbana Champaign


#### Abstract

A study, investigating the relationship between scientific reasoning and the capacity to discover the strategy to play an hypothetico-deductive game (MasterMind), posits that students being able to discover Complex Strategies (vs. General Strategy, Feedback Related, No Strategy) were also, on average, performing higher on our measure of scientific reasoning, itself composed of evaluative, experimental and scientific knowledge measures. In addition to bridge the discovery of complex strategies with higher SR ability, the finding also suggests the necessity to integrate rule discovery exercises in curriculum to 1practice while 2- recognize valid reasoning procedures. Finally, inquiring about the middle schooler's capacity to recognize the most effective strategy, will help to assess the class level as a whole. Such assessment will help the teacher identify some needs and target effective lessons to explain and facilitate the transfer of CoV strategies to novel situations, as suggested by "real life" problem demands.


# Neural responses decrease while performance increases with practice: A neural network model 

Milena Rabovsky<br>Freie Universitaet, Berlin

Steven S. Hansen

Stanford University

James L. McClelland

Stanford University


#### Abstract

Why do neural responses decrease with practice? We used a predictive neural network model of sentence processing (St. John \& McClelland, 1990) to simulate neural responses during language understanding, and examined the model's correlate of neural responses (specifically, the N400 component), measured as stimulus-induced change in hidden layer activation, across training. N400 magnitude first increased and then gradually decreased over training while comprehension performance at the output steadily rose with practice. These results fit the developmental trajectory of N400 amplitudes. Importantly, they also address the reduction of neural activation with practice. In the model, the reduction is due to continuous adaptation of connection weights over training. As connection weights between hidden and output layer grow stronger, less hidden layer activation is necessary to efficiently modulate the output. This shift of labor from activation to connection weights might be an important mechanism contributing to the reduction of neural activation with practice.


# Age-related top-down and bottom-up guidance on eye movements when searching in real-world scenes 

Hanane Ramzaoui<br>Laboratoire d'Anthropologie et de Psychologie Cliniques, Cognitives et Sociales, Université Côte d'Azur, France<br>Sylvane Faure<br>Laboratoire d'Anthropologie et de Psychologie Cliniques, Cognitives et Sociales, Université Côte d'Azur, France<br>\section*{Sara Spotorno}<br>Institute of Neuroscience and Psychology, University of Glasgow, UK


#### Abstract

Efficient selection of targets is crucial in everyday activities across the lifespan. Studies reporting age-related decline have, however, typically utilised arrays of simple, unrealistic objects. Using real-world scenes, we investigated how reliability of scene semantics (consistent vs. inconsistent targets), target template specificity (name vs. precise picture) and target perceptual salience influence oculomotor search behaviour in older vs. young viewers. Aging resulted in slower search considering initial saccade latency, time and number of fixations to locate the target, and verification of object-template matching. No group differences emerged in accuracy and in search facilitation due to a pictorial template or a semantically consistent target. Target high salience enhanced efficiency in both groups, with stronger effects in older viewers. Aging seems therefore to lead to an overall search speed reduction not due to specific deficits in utilisation of scene semantic guidance or in target recognition, and possibly reduced by enhancing target perceptual guidance.


# The Long and Short of It: The Role of Verb Stem Vowel Duration in Sentence Processing 

Gwendolyn Rehrig
Rutgers University
Sten Knutsen
Rutgers University
Nicolaus Schrum
Rutgers University
Paul de Lacy
Rutgers University

Karin Stromswold<br>Rutgers University


#### Abstract

When native English speakers say active and passive sentences, verb stems are longer in passive sentences than in their active counterparts (Stromswold et al., 2002; Rehrig et al., 2015) because phrase-final lengthening and polysyllabic shortening cause the verb stem vowel to be longer in passives (Aveni et al., 2016; Mayro et al., 2016). Eye-tracking and gating studies of unaltered sentences revealed that listeners are able to predict whether a sentence is active or passive prior to hearing the inflection on the verb (Stromswold et al., 2002; 2016). To examine whether listeners use vowel duration in online sentence comprehension, we lengthened the vowel in half of the active verb stems and shortened it in half of the passive verb stems. Reaction times were longer for sentences with altered verb stem vowels ( $\mathrm{p}<.001$ ), consistent with listeners using verb stem vowel duration as a predictive cue in online comprehension.


# Recycling or Trash Bin? Modeling Consumers' Recycling Behavior in a Field Study 

Torsten Reimer<br>Purdue University, West Lafayette, IN, United States<br>Juan Pablo Ramirez<br>National University of Colombia, Bogota, Colombia<br>Chris Roland<br>Purdue University, West Lafayette, IN, United States<br>\section*{Devika Banerji}<br>Purdue University, West Lafayette, IN, United States


#### Abstract

What affects people's behavior when they dispose items? The distance hypothesis predicts that the number of misplaced items is a function of the distance of an appropriate bin. We categorized and mapped bins at 140 locations on the campus of a major research university in the Midwest and calculated the distances between adjacent bins. The distance hypothesis predicts that users dispose more recyclables in single, isolated trash bins than in trash bins that are paired with recycling bins. Likewise, it is expected that more trash items can be found in isolated compared to paired recycling bins. We conducted a field study that involved systematic comparisons of matched locations and focused on behavioral data that were obtained through systematic audits of trash and recycling bins. The study provided partial support for the distance hypothesis, which was supported for certain items. The role of item difficulty and weather conditions will be discussed.


# Is Neurocomputational Self-Organization a Core Mechanism of AGI Systems? 

Spyridon Revithis<br>University of New South Wales (UNSW), Sydney, NSW, Australia


#### Abstract

Artificial General Intelligence (AGI) is a term that describes a variant of a Strong AI revival in the mind sciences. Irrespective of its definition limits, and leaving aside the non-scientific metaphysical or philosophical aspirations, AGI studies the feasibility and implementation aspects of artificial systems that would have the capacity of domain non-specific (domaingeneral) human-level intelligence.

The importance of self-organization in natural neural systems as well as in neuromimetic computational systems, especially the class of Self-Organizing Map (SOM) neural networks, has been extensively demonstrated and supported in the literature. Neurocomputational self-organization exhibits unique characteristics, including non-deterministic epigenetic (post-genetic) behavior, which enable direct functional and structural comparisons with the neocortex more than most existing relevant computational mechanisms. If the problem of artificial general intelligence is approached from a biologically relevant computational standpoint then SOM mechanisms are currently a very strong candidate as a core component of a computational AGI system.


# Does a present bias influence exploratory choice? 

Alexander S Rich<br>New York University<br>Todd M Gureckis<br>New York University


#### Abstract

Balancing exploration and exploitation is difficult, and across a wide variety of situations under-exploration of uncertain alternatives appears prevalent. We propose that one possible cause of under-exploration is present bias, whereby immediate rewards (like those gained from exploitation) loom larger than future rewards (like those gained from exploration). This possible cause of under-exploration is not addressed by past studies, in which choices generally yield token rewards that are converted to money at the end of the experiment, removing the inter-temporal aspect of the decision-making process. To address this issue, we developed an exploratory choice task with immediately-consumed rewards. We then tested whether whether imposing a temporal delay before the consumption of rewards increased exploration by decreasing present bias, and report on our results.


# Everyday object affordance enhances automatic inhibitory control: an ERP study 

Stefania Righi<br>University of Florence<br>Giorgio Gronchi<br>University of Florence

## Lapo Pierguidi

University of Florence

## Maria Pia Viggiano

University of Florence


#### Abstract

High affordance stimuli are associated with an enhancement in the activation of the corresponding motor programs. Such over-activation of motor programs may imply a decrease in performances based on inhibitory control. However, recent data suggest that high affordance stimuli are associated with a widespread privileged neural activation that goes beyond motor representations. In this case, we can expect that high affordance objects will be associated to a higher level of flexibility in an oddball task with Go-NoGo procedure. By measuring ERPs, we observed that, in the case of high affordance objects, the amplitude of the N200 is decreased when the inhibition of the motor response is more difficult. Data suggests that high affordance objects facilitate inhibitory control, probably due to a higher activation of automatic attentional resources.


# Metacognitive Monitoring of Internal and External Storage and Retrieval 

Evan Risko<br>University of Waterloo, waterloo<br>Connor Gaspar<br>University of Waterloo, waterloo<br>Dave McLean<br>University of Waterloo, waterloo<br>Tim Dunn<br>University of Waterloo, waterloo<br>Derek Koehler<br>University of Waterloo, waterloo


#### Abstract

The ability to monitor our cognitive performance (i.e., metacognitive monitoring) is central to efficient functioning. Research investigating this ability has focused largely on tasks that rely exclusively on internal processes. However, our day-today cognitive activities often consist of mixes of internal and external processes. For example, we can offload memory demands onto external media (e.g., computers, paper). In the present investigation, we expand research on the metacognitive monitoring of performance to this domain. Specifically, we examine participant's ability to accurately monitor their performance in tasks that require them to rely on only their internal processes (e.g., short term memory to remember a series of letters) and tasks that require them to rely on both (e.g., paper and pencil to remember a series of letters). Results suggest that the former results in superior monitoring relative to the latter. Implications for understanding metacognition in more distributed cognitive domains will be discussed.


# From Concrete Examples to Abstract Relations: A model-based neuroscience approach to how people learn new categories 

Gareth Roberts<br>School of Psychology, University of Sydney<br>Micah Goldwater<br>School of Psychology, University of Sydney<br>Evan Livesey<br>School of Psychology, University of Sydney<br>Josue Giron<br>School of Psychology, University of Sydney<br>Tyler Davis<br>Department of Psychological Sciences, Texas Tech University


#### Abstract

The ability to form relational categories for objects that share few features in common is a hallmark of human cognition. However until recently, neuroimaging research largely focused solely on how people acquire categories defined by features. In the current electroencephalography (EEG) study, we examine how relational and feature-based category learning compare in well-matched learning tasks. Building on a previous functional magnetic resonance imaging study by our laboratory, we capitalise on the rich temporal information offered by EEG. Focusing on the neural dynamics of how people learn category memberships over individual trials in an experimental task, we investigate how these single trial dynamics modulate computational estimates from decision-making modelling frameworks. Specifically, by sorting participants’ individual trials by their position in the experimental sequence we observe striking relationships between EEG dynamics (e.g., frontal theta oscillations and P300 component) and feature-based and relational categorisation behaviour.


# Progress in building a machine that can ask interesting and informative questions 

Anselm Rothe<br>New York University<br>Brenden Lake<br>New York University<br>Todd Gureckis<br>New York University


#### Abstract

Asking creative questions is a hallmark of human cognition. In comparison, machine learning systems that attempt to mimic this ability are still extremely limited (e.g., current chatbots ask questions based on preprogrammed routines). In the present work, we developed a computational model of question generation. Based on a corpus of questions collected from online participants playing an information-seeking game, we designed a "grammar of questions." The grammar is powerful enough to represent all human questions we collected and thus defines the "question space." Given a particular context (game scenario), people are more likely to ask (generate) some questions that others. Our computational model predicts these likelihoods, that is, a probability distribution over the question space. In addition, the model can generalize to novel contexts. Key model ingredients are informativity, compositionality, and length of a question.


# Shifting backward to say what's front? Spatial referencing of dorsal object arrangements 

Annelie Rothe-Wulf<br>University of Freiburg, Freiburg, Germany<br>Andrea Bender<br>University of Bergen, Bergen, Norway<br>Julia Fischer<br>University of Freiburg, Freiburg, Germany<br>Nadine Lobmüller<br>University of Freiburg, Freiburg, Germany<br>Sieghard Beller<br>University of Bergen, Bergen, Norway


#### Abstract

When referring to spatial arrangements of two objects in the visual field, German native speakers prefer reflection as a subtype of the relative frame of reference. Whether this preference transfers to objects in one's back and whether a mental turn has to precede such dorsal references (turn hypothesis), has recently been explored in studies implementing questionnaires. However, the results hardly supported the turn hypothesis and rather suggested backward projection as an alternative strategy for dorsal references. To test the two assumptions more rigorously, a series of experiments implemented dorsal object arrangements in interview situations and induced dorsal perspectives via turning, shifting or reflecting the actual view of participants. Across experiments and conditions, backward projection consistently emerged as the preferred referencing strategy and only a small proportion of dorsal references accorded with the turn hypothesis. Participants' retrospective descriptions supported this pattern and suggested backward projection to be involved in dorsal referencing.


# The time course of Intentional Binding 

Miriam Ruess<br>Albert-Ludwigs-Universität Freiburg, Freiburg i. Br., Germany<br>Roland Thomaschke<br>Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

## Andrea Kiesel

Albert-Ludwigs-Universität Freiburg


#### Abstract

Environmental stimuli caused by actions (i.e., effects) are perceived earlier than stimuli not caused by actions. This phenomenon is termed intentional binding (IB), and serves as implicit measure of sense of agency. We investigated the influence of effect delay and temporal predictability on IB, measured with the classic clock procedure as the bias to perceive the effect as temporally shifted towards the action. For short delays, IB increased with delay (Experiment 1: $200 \mathrm{~ms}, 250 \mathrm{~ms}, 300 \mathrm{~ms}$ ) and this initial increase declined for longer delays (Experiment 2: $100 \mathrm{~ms}, 250 \mathrm{~ms}, 400 \mathrm{~ms}$ ). These results extend previous findings showing IB to decrease with increasing delays for delay ranges of 250 ms to 650 ms . Further, the indication that IB, that is, sense of agency, might be maximal for different delays depending on the specific characteristics and context of action and effect, has important implications for human-machine interfaces.


# A positive attitude increases subjective life expectancy 

Susana Ruiz Fernández<br>Leibniz-Institut für Wissensmedien, Tübingen, Germany<br>Martin Lachmair<br>Leibniz-Institut für Wissensmedien, Tübingen, Germany<br>Juan José Rahona<br>Leibniz-Institut für Wissensmedien, Tübingen, Germany<br>Lotte Sophia Roessler<br>Leibniz-Institut für Wissensmedien, Tübingen, Germany<br>Peter Gerjets<br>Leibniz-Institut für Wissensmedien, Tübingen, Germany


#### Abstract

Subjective life expectancy (SLE) has been related to psychological variables, such as optimism. Based on previous studies where positive attitude was related with longer lifetime, the present study examined whether modifying participants' attitude would influence their SLE. Therefore, 50 participants were randomly assigned either to a positive or to a neutral attitude group. During one week, participants of the positive (neutral) group, had to choose the three most accurate positive (neutral) sentences (among 22) to describe their day. After this week, they had to estimate the probability of being $60,70,80$, or 90 years old (traditional measure) and to situate themselves on a line representing their lifetime (spatial based measure). Results show that 1) a more positive attitude increased SLE more than a neutral one, 2) the spatial based measure was sensitive to the intervention and 3) both measures correlated positively with participants' optimism.


# Probability matching in choice behavior influenced by virtual rewards 

Yuichi Saito<br>Graduate School of Engineering, Kogakuin University

Miyuki G. Kamachi<br>Faculty of informatics, Kogakuin University


#### Abstract

We recognize the amount of "reward" according to our choices. In repeated binary choice tasks, human behave according to the theoretical basis of "probability matching" (Shanks et al., 2002), which has been advocated in several studies. However, the quality of reward may influence their choice-behavior. It is acknowledged that the sensitivity of values for gains or losses differs among individuals because of risk aversion (Kahneman \& Tversky, 1979). We conducted a series of experiments to investigate how participants' choices change when the ratio of hit-items was set up. Virtual rewards, $-3 / 0,-3 /+3$, and $0 /+3$ point for each choice, were given to participants. The results showed that the choice-ratio of the weighted correct side was higher in conditions involving losses, suggesting that participants' choices indicate risk aversion even though rewards were virtual. Our results suggest that probability matching can be found only when people implicitly recognize their choices have no loss.


# ASR Systems as Models of Phonetic Category Perception in Adults 

Thomas Schatz<br>LSCP (ENS/EHESS/CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, France<br>Francis Bach<br>SIERRA Project-Team (ENS/INRIA/CNRS), Département d'Informatique, Ecole Normale Supérieure, PSL Research University, France<br>Emmanuel Dupoux<br>LSCP (ENS/EHESS/CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, France


#### Abstract

Adult speech perception is tuned to efficiently process native phonetic categories, causing difficulties with certain non-native categories. For example, Japanese has no equivalent of the distinction between American English $/ \mathrm{r} /$ and $/ \mathrm{l} /$ and native speakers of Japanese have a hard time discriminating between these two sounds. Here, we ask whether standard Automatic Speech Recognition (ASR) systems trained on large corpora of continuous speech can make correct quantitative predictions regarding such non-native phonetic category perception effects. By training an ASR system on language L1 and evaluating it on language L2, we obtain predictions for a native L1 speaker tested on L2 phonetic contrasts. Using a variety of L1 and L2, we show that ASR models correctly predict several well-documented effects. Beyond the immediate results, our evaluation methodology, based on a machine version of ABX discrimination tasks, opens the possibility of a more systematic investigation of computational models of phonetic category perception.


# Which test to perform? Modeling utility of medical tests: information gain, patient risk and financial costs 

Clara Schirren<br>Humboldt-Universität zu Berlin

Stefanie C. Hautz

AO Educational Institute

Wolf E. Hautz

Inselspital - Universitätsspital Bern
Juliane E. Kämmer
Charité Universitätsmedizin Berlin
Björn Meder
Max Planck Institute for Human Development, Berlin
Jonathan D. Nelson
Max Planck Institute for Human Development, Berlin


#### Abstract

In medical diagnosis, as in many cognitive domains, asking the right questions is crucial. Medical tests differ not only in the type of information they provide, but in their financial costs and physical risks to a patient. We develop a model that combines informational and cost constraints, describing specific medical scenarios of a patient with realistic symptoms. We then define a finite number of existing medical tests that are available in this situation. The tests differ in their sensitivity and specificity concerning different possible underlying diseases as well as in their financial costs and the physical risks they pose to a patient. Combining these, we compare the utilities of the different tests if performed alone as well as if performed in combination. We show how purely informational considerations are not adequate for the analysis of such a scenario; test costs and patient outcomes must also be taken into account.


# Intuitive system control: Challenging the standard model of dynamic decision making 

Wolfgang Schoppek<br>University of Bayreuth


#### Abstract

Dynamic decision making (DDM) is usually operationalized in a way that subjects explore and control computersimulated dynamic systems consisting of interconnected variables. Most authors in the field agree on a "standard model of DDM" that assumes that decision makers have to learn the causal structure of the system through appropriate explorative behavior before they can bring the system to given goal states. This strategy draws heavily on cognitive ressources, such as working memory. The standard model predicts that performance in DDM, as well as structural knowledge should be severely impaired when a second cognitive task has to be executed while exploring the system. An experiment with a dual task as the main factor revealed no differences in knowledge and performance between the conditions. Participants in both conditions appeared to rely on rudimentary structural knowledge and adopted intuitive strategies. We interprete the findings within a dual processing framework.


# Learning Temporal Generative Neural Codes for Biological Motion Perception and Inference 

Fabian Schrodt<br>University of Tübingen, Tübingen, Baden-Württemberg, Germany<br>Martin V. Butz<br>University of Tübingen, Tübingen, Baden-Württemberg, Germany


#### Abstract

We introduce a modular recurrent neural architecture, which learns distributed, generative temporal models of biological motion. It encodes modal visual and proprioceptive (angular) biological motions separately by means of autoencoders, structuring respective postures, motion directions, and motion magnitudes separately. The submodal encoders are interdependent by predicting each other's next autoencoder states temporally. As a result, distributed attractor states can develop from self-generated motions. We show that the architecture is able to synchronize its activities across modalities towards overall consistent action-encoding attractors. Moreover, the developing spatial and temporal structures can complete partially observable actions, e.g., when only providing visual information. Furthermore, we show that the network is capable of simulating whole-body actions without any sensory stimulation, thus imagining unfolding actions. Finally, we show that the network is able to infer the visual perspective on a biological motion. Thus, the neural architecture enables embodied perspective taking and action inference.


# The Sufficiency Principle: Predicting when children will regularize inconsistent language variation 

Kathryn D. Schuler<br>Georgetown University<br>Jaclyn E. Horowitz<br>Georgetown University

Charles Yang
University of Pennsylvania
Elissa L. Newport
Georgetown University


#### Abstract

Children exposed to inconsistent language variation regularize this variation in their productions (Hudson-Kam \& Newport, 2005). Existing demonstrations of regularization observe this behavior when the signal-to-noise ratio is greater-than-or-equal-to $40 \%$, but whether regularization occurs when the dominant form is less widespread has not been investigated. A recent computational model, the Sufficiency Principle, quantifies when a pattern is widespread enough to generalize (Yang, 2016): Let $R$ be a generalization over $N$ items, of which $M$ are attested to follow $R$. $R$ extends to all $N$ items iff: $N-M<N / \ln (N)$. To test this model, we exposed children to artificial languages in which the dominant form occurred either above or below this threshold for generalization. We found that, as predicted, children regularized only under circumstance in which the Sufficiency Principle threshold for generalization is met. Thus, regularization may be governed by a basic principle of generalization that is well captured by this model.


# Impaired phonological processing of lexical tones in Cantonese speakers with congenital amusia 

Jing Shao<br>The Hong Kong Polytechnic University<br>Xunan Huang<br>The Hong Kong Polytechnic University<br>Caicai Zhang<br>The Hong Kong Polytechnic University


#### Abstract

Congenital amusia is a lifelong musical disorder. It has been found that tonal-language speakers with amusia are impaired in lexical tone perception. But it has also been found that tonal-language experience compensates the deficit in certain scenario, reducing prevalence rate of amusia in speakers of a highly complex tonal-language - Cantonese. Thus it remains unclear whether lexical tone perception, especially its phonological processing, is impaired in Cantonese-speaking amusics. This study investigated the categorical perception of a continuum of lexical tone stimuli and pure tone analogues in Cantonesespeaking amusics and controls. The amusics showed reduced discrimination peak across the categorical boundary compared to controls in lexical tone condition, suggesting impaired categorical perception; in pure tone condition, the amusics showed inferior performance on both between- and within-category discriminations, suggesting a deficit in auditory pitch processing. These findings indicate that phonological processing of tone is impaired in Cantonese-speaking amusics, despite possible compensation effect.


# Discovering kinds of future-oriented thought using automated machine-learning techniques 

Jason Shepard<br>Agnes Scott College

Aiming Nie<br>Stanford University

Bridget Copley
CNRS/Universite Paris 8
Phillip Wolff
Emory University


#### Abstract

Humans have a remarkable ability to think about the future. Our abilities to think about the future are essential for the level of goal construction, planning, and execution of plans that is only observable in humans. Thinking about the future has also been found to be important for the development of sense of self and for health and well-being. In spite of the importance of future-oriented thought, very little empirical work has been conducted on the nature of future-oriented thought. In this research, we demonstrate how automated methodologies can be used to identify references to the future from natural text (Study 1) and how machine-learning techniques can be used to identify categories of future-oriented thought (Study 2). We also demonstrate how the categories that emerge from these analyses can help us better understand the relation between future-oriented thought and many of the positive outcomes associated with future-oriented thought (Study 3).


# The influence of word-order harmony on structural priming in artificial languages 

Kenny Smith<br>University of Edinburgh<br>Olga Feher<br>University of Edinburgh<br>Jennifer Culbertson<br>University of Edinburgh


#### Abstract

Structural priming occurs when interlocutors copy the syntactic structure of their partners' utterances, and is diagnostic of their underlying representations. We trained adult participants on an artificial 'alien' language in which nouns appeared with adjectives or numerals in two-word phrases; participants then used that language to communicate with an alien interlocutor. Input languages had variable word-order with the two modifier types tending to appear on the same side of the noun (harmonic) or on different sides of the noun (non-harmonic). Participants in all conditions acquired the dominant order of their input; however, structural priming only occurred within modifier types (e.g. encountering Numeral-Noun primed Numeral-Noun order only, not Adjective-Noun), even for participants exposed to harmonic input where both modifier types patterned the same way. This suggests that the abstract representations tapped by structural priming in rapidly-learnt artificial languages encode distinctions that are not based purely on distributional properties of the input.


# Recently rewarded task-irrelevant stimuli do not distract 2-year-olds during visual search 

Hayley K Smith<br>University of Massachusetts, Boston<br>Erik Blaser<br>University of Massachusetts, Boston<br>Zsuzsa Kaldy<br>University of Massachusetts, Boston


#### Abstract

In adults, stimuli associated with reward capture attention, even when task-irrelevant, resulting in distraction (Awh et al., 2012). Here we examine whether rewarded stimuli capture attention in 2 -year-old children. Toddlers ( $\mathrm{N}=46$, mean age: 28;10, range: 19;16-36;18) performed a visual search task where the target switched between blocks. Search arrays consisted of the current target, a previous target, and six feature conjunction distractors. On each trial, the current target was cued, and following a fixed search period, rotated as a reward. We used a Tobii T120 eye-tracker to record toddlers' eye-movements. Following a target switch, toddlers fixated the current target before the previous target, despite the previous target's recent reward history $\mathrm{F}(1,44)=31.183, \mathrm{p}<0.001$ ). Our study is one of the first to investigate the early development of reward-based attentional selection.


# Disambiguating Disfluencies: What Do Speech Disfluencies Tell Us About Speech Production? 

Mikhail Sokolov<br>Carleton University

John Logan<br>Carleton University


#### Abstract

Speech disfluencies occur frequently in spontaneous speech but their source is unclear. Disfluencies can take several forms, most commonly as verbalized disfluencies such as "um", "uh", and "so", as well as silent pauses. In the present exploratory study we examined the relationship between disfluencies as distinct entities, individual differences in working memory capacity, and linguistic markers of complexity. We found that disfluencies diverge in their relationship with these variables. The "um" disfluency was most closely related to working memory capacity and linguistic complexity. The "uh" disfluency was associated with infrequent word production. The "so" disfluency predicted of the number of words produced. Silent disfluencies were not related to working memory capacity. However, micro-pauses were related to word production, and macro-pauses were negatively correlated with the "so" disfluency. Results are discussed in terms of potential relationships between disfluencies and speech production processes.


# Phonological Competition during Spoken-Word Recognition in Infants and Adults 

Marlene Spangenberg<br>University of Oxford, Oxford, United Kingdom<br>Kim Plunkett<br>University of Oxford, Oxford, United Kingdom


#### Abstract

An ongoing debate concerns whether spoken word recognition happens in an incremental or continuous manner (Marslen-Wilson \& Zwitserlood, 1989; McClelland \& Elman, 1986). In the current study, participants ( 31 adults and 49 infants aged $24-30$ months) were presented with four images while they heard a sentence like "Look at the cat". Among the images was one object that rhymed with the spoken word, one object that shared its onset and two phonologically unrelated objects. Growth curve analysis of eye-tracking data revealed that adults preferentially fixated onset competitors over unrelated objects soon after word onset but did not preferentially fixate rhyme competitors. Fixations of the onset competitors were modulated by the degree to which the onsets of the three remaining competitors were phonologically similar to the spoken word. Infants showed no preference for either type of phonologically related competitor. The absence of a rhyme effect contradicts continuous theories of spoken word recognition.


# Word order rules in business name binomials 

Samuel Spevack
University of California, Merced, Merced, California, United States

## Patricia Lichtenstein

University of California, Merced, Merced, CA, US

Stephanie Shih

University of California, Merced


#### Abstract

Naming practices offer a window onto linguistic processes of productivity that rely on input from interacting streams of information. Previous studies have looked at proper personal names and binomial combinations of proper personal names to show that phonological features such as rhythm, semantic features such as gender, and corpus features such as word frequency play an important role in naming and ordering of names. In comparison to personal names, business names tend to be more diverse in terms of constituent structure, often incorporating binomial constructions that may or may not consist of proper names themselves. In this study, we investigate whether the ordering of binomials in business names reflects the features identified in previous work, with a focus on the following: syllable count, metrical stress, animacy, concreteness, word frequency, and binomial frequency. We report here on an initial analysis of data from the Yelp Dataset Challenge.


# The time course of colour guidance in realistic scene search 

Sara Spotorno<br>University of Glasgow, Glasgow, Scotland, United Kingdom<br>Ben Tatler<br>University of Aberdeen, Aberdeen, Scotland, United Kingdom


#### Abstract

Colour is a source of attentional guidance and object segmentation when viewing a scene. In an eye-tracking study, we examined its role during search of targets placed in consistent or inconsistent locations within realistic scene contexts. Both the target template and the whole scene were presented in full colour or grayscale. Colour presence did not influence early search, considering latency, direction or gain of the first saccade, but affected later phases, with longer scene scanning and more fixations required to locate the target in the grayscale condition, which also lengthened verification of template-object matching. These effects were enhanced in inconsistent scenes. Our results suggest that observers may not utilise colour cues when initiating scene inspection during search but also that colour information modulates efficiency of the search process in terms of attentional selection and object recognition, in particular when the context of the scene does not provide reliable high-level guidance.


# Perceptual decision making from correlated samples 

Oana Stanciu<br>Central European University, Budapest, Hungary<br>Mate Lengyel<br>University of Cambridge<br>Daniel Wolpert<br>University of Cambridge<br>Jozsef Fiser<br>Central European University, Budapest, Hungary


#### Abstract

The optimal perceptual decision making strategy for weighting serially presented information depends on the degree of sample dependence. Uniform weighting produces optimal estimates from independent samples, but increases in autocorrelation should be matched by increasing and symmetric overweighting of early and late samples in order to maintain optimal performance.

In the current experiment, participants $(\mathrm{N}=30)$ observed briefly presented sequences of eight dots and were asked to estimate their center of mass by dragging the cursor. The autocorrelation of the series was manipulated in two distinct blocks (either 0 or .7). Preliminary results show that the weight assignment to uncorrelated inputs did not differ significantly from the optimal uniform allocation. In contrast, in the high-dependence block participants used different weighting profiles - overweighting the first and/or last samples of the sequence. This suggests that humans flexibly adapt to changes in statistical structure in the predicted direction of optimality.


# Optimality of visual search under ambiguous stimuli 

Elina Stengård<br>Department of Psychology, Uppsala, Sweden<br>Ronald van den Berg<br>Department of psychology, Uppsala, Sweden


#### Abstract

A hallmark of optimal decision making is that cues are weighted by their reliability. Previous studies have reported evidence for reliability weighting in several human perceptual decision-making tasks in which sensory noise was the only possible source of errors. Here we use a target detection task to test whether optimality generalizes to situations where stimulus ambiguity is an additional source of uncertainty. Target and distractor orientations were drawn from distributions with different means and the level of ambiguity was varied through the amount of overlap between the two distributions. In line with previous studies, we found clear evidence for sensory reliability-weighting, regardless of the level of ambiguity. However, using a richer set of models than before, we also found that the estimated weights deviated from the optimal ones. Finally, we found no effect of ambiguity level on task efficiency, which suggests that subjects optimally accounted for this source of uncertainty.


# It is new, but will it be good? Context-driven exploration of novel options 

Hrvoje Stojic<br>University College London, London, London, United Kingdom<br>Eric Schulz<br>University College London, London, London, United Kingdom<br>Maarten Speekenbrink<br>University College London, London, London, United Kingdom


#### Abstract

How do people decide whether to try out novel options? We argue that they utilize contextual information to efficiently generalize from learned functional relations in order to decide between known or novel options. In a contextual multi-armed bandit task, in which rewards are a noisy function of observable features, we assess participants' preferences for newly introduced options. We show that participants preferably choose a novel option if its features indicate high rewards, but shun the option if its features indicate low rewards, a behavior that can only be explained by functional generalization. Moreover, we assess people's preferences for novel options that have medium rewards to test whether they prefer options less similar to experienced options, consistent with choices guided by uncertainty. Given that novel options normally come with observable features, we argue that contextual learning is a parsimonious yet powerful explanation of behavior in the face of novelty.


# Mathematical Symbol Recognition in Children 

Sandra Street<br>University of Wisconsin-Whitewater, Whitewater, WI, US


#### Abstract

In early mathematics development the development of symbolic skills is critical to math learning. Our mathematical system is based on Arabic number symbols which do not provide any semantic meaning relevant to the number words and symbols. In order to succeed in math one must be able to recognize and understand the meaning of numeric and other mathematical symbols. Little is known about the development of these symbolic skills. The current study examines $4-7$-year-old children's understanding and recognition of number and arithmetic symbols. The youngest children made significant errors in identifying numbers as well as confusing letter symbols with number symbols. Results reveal a developmental progression of numeric symbol recognition.


# Speaking in English, sorting in Chinese: interaction in L2 can reinforce existing categories in L1 

Ellise Suffill<br>University of Edinburgh, Edinburgh, Midlothian, United Kingdom<br>Holly Branigan<br>University of Edinburgh, Edinburgh, UK<br>Martin Pickering<br>University of Edinburgh, Edinburgh, UK


#### Abstract

How does interaction affect categorization, and how might this vary between native and non-native speakers? When people use shared labels to categorize objects, they categorize more similarly to each other. We investigated whether interaction leads non-native speakers to categorize in the same way as native speakers. In six rounds, L1-English and L2-English/L1Mandarin speakers individually categorized dishware using labels (BOWL, PLATE), then discussed their categories or an unrelated topic after each round. L2 speakers' categories shifted following category-relevant interaction with L1 speakers, but their categories did not become more L1-like. Unexpectedly, category-relevant interaction reduced alignment within pairs and within language groups; however, this effect was weaker in the L2 than L1 group. Hence, L2 speakers showed a stronger tendency than L1 speakers to use categories that were similar to other speakers from their language group. This suggests that interaction in an L2 can reinforce L2 speakers' categories in their L1.


# The effect of overt language use on category induction 

Justin Sulik<br>Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands<br>Haily Merritt<br>Indiana University, Bloomington, IN, USA<br>Gary Lupyan<br>University of Wisconsin, Madison, WI, USA


#### Abstract

Successfully solving a problem should help people solve similar problems, but such generalization is often surprisingly limited. We investigated generalization performance when people explicitly verbalized solutions to open-ended categoryinduction "Bongard problems", compared to tacitly indicating that they had found a solution. In a Bongard problem people are presented with an array of items falling into two classes, and have to induce the basis for the classification by working out what (sometimes quite abstract) feature of the items is relevant, from a vast set of possibilities. We measured objective performance by testing people with new items, and observed how explicitly vs. tacitly expressed solutions affected generalization across concretely similar or abstractly similar problems. For the concretely similar problems, explicitness boosted transfer of correct solutions. For the abstractly similar problems, there was no evidence of transfer, though there was a general positive effect of explictness.


# Respecting UP and Despising DOWN: Emotional and Body-based Image in Japanese Verbs 

Tomohiro Taira<br>Osaka City University


#### Abstract

The aim of this study was to examine the image-schematic representations that arise from sentences referring to concrete/abstract action in Japanese verbs. We used a free positioning task that required the participants to draw the position of an object in a sentence referring to an agent's concrete/abstract action and a simple rating task that investigated the agent's need for body movement and emotional evaluation for the object. The results showed that the drawn object's position in not only a concrete but also an abstract action sentence is changed before and after the action. Further, the results indicated that the height and distance from the agent to the object in the sentence is related to the emotional evaluation of the agent for the object in the sentence.


# A time-series eye-fixation analysis of the similarity-compromise effect in multi-alternative choice 

Tsuzuki Takashi<br>Rikkyo University

Chiba Itsuki<br>National Institute of Advanced Industrial Science and Technology


#### Abstract

In decision-making tasks with two attributes and three alternatives, the similarity effect implies that, if the total expected utility is equal between two opposite alternatives (i.e., the target and competitor), the probability of the target being chosen decreases with the addition of the decoy similar to the target. This study demonstrated the similarity-compromise effect, wherein the probability of the target being chosen increased with the addition of the decoy, under the same conditions as the similarity effect, by setting all attribute values of three alternatives to broken numbers rather than rounded numbers. To determine the mechanism underlying this effect, we examined information acquisition patterns using eye-movement data collected from 37 undergraduates who made 10 hypothetical purchase tasks with two attributes and three alternatives. Timeseries analysis of fixation time for the three alternatives revealed dynamic temporal features distinct from those of attraction and compromise effects observed in our previous research.


# Relationship between four measures reflecting representations of fraction magnitude in adults: number line estimation, comparison, calculation of fractions, and immediate serial recall of fractions 

Yuki Tanida<br>Osaka University, Suita-city, Osaka, Japan<br>Yu Koshima<br>Osaka Prefecture University, Sakai, Osaka, Japan<br>Masahiko Okamoto<br>Osaka Prefecture University, Sakai, Osaka, Japan


#### Abstract

Our previous studies (presented at the London meeting of EPS in 2017 and submitted for ICPS 2017) suggested that immediate serial recall tasks access magnitude representation of fraction. A subsequent research task is to explore the intercorrelations among four tasks stimulating representations of fraction magnitude: an immediate serial recall task with fraction stimuli and three typical tasks, number line estimation, comparison, and calculation of fractions. The purpose of this study is to examine whether our new measure, the size of the magnitude similarity effect on immediate serial recall of fractions, relates to other typical measures for adults. The results from 36 university students showed a significant correlation between the size of the magnitude similarity effect and the RT of fraction calculation tasks but no correlations among any other tasks. This result suggests that it is necessary to reexamine what tasks could access the magnitudinal representation of fraction in adults.


# Slow Change: The Visual Context for Real World Learning 

Charlene Tay<br>Indiana University Bloomington, Bloomington, Indiana, United States<br>Linda Smith<br>Indiana University Bloomington, Bloomington, Indiana, United States<br>Chen Yu<br>Indiana University Bloomington, Bloomington, Indiana, United States


#### Abstract

The visual world can be a noisy and dynamic place. This poses problems for novice word learners who must map heard names to objects in scenes with their many potential referents. In this study, we consider how the visual stability and selectivity of scenes from the first-person perspective may simplify the learning problem. 12- and 30-month-old children wore head cameras and played with a large set of toys. Through analyzing head-camera video frame by frame, we measured the rate of change of scene information in the natural world of children in this context, and found that the visual world from the child's perspective changes continuously. However, this change is slow and incremental - tiny steps - even across increasingly larger timescales. We discuss the importance of understanding the dynamics of real world environments for understanding visual processing, sustained attention, and early object name learning.


# Effects of attention to emergent phenomena on rule discovery 

Hitoshi Terai<br>Kindai University<br>Kazuhisa Miwa<br>Nagoya University<br>Sho Yokoyama<br>Kindai University<br>Souta Fujimura<br>Kindai University<br>Gotaro Nakayama<br>Kindai University


#### Abstract

In this study, we focused on effects of finding of emergent phenomena in rule discovery. In the experiment, we used Conway's Game of Life, which generates high-order phenomena from fundamental rules. Our research question is to realize the effects of attention to emergent phenomena on finding the fundamental rules. The two experimental conditions (chaotic and static) differed only in initial states. In the chaotic condition, the initial state consisted some Methuselahs, which take long period until they become stable. On the other hand, in the static condition, the initial condition consisted many emergent patterns: still lifes and oscillators, which repeat same pattern in short period. We classified the hypotheses reported by the participants to either mentioning about emergent phenomena or not. This result revealed that people might see emergent phenomena not only in the static condition but also in the chaotic condition, which do not include the emergent patters.


# Scheduling system delays for optimal user performance: Don't predict time; let time predict! 

Roland Thomaschke<br>Albert-Ludwigs-Universität Freiburg, Freiburg, Germany<br>Lennart Koch<br>Albert-Ludwigs-Universität Freiburg, Freiburg, Germany<br>Miriam Ruess<br>Albert-Ludwigs-Universität Freiburg, Freiburg, Germany<br>Andrea Kiesel<br>Albert-Ludwigs-Universität Freiburg, Freiburg, Germany


#### Abstract

System delays affect user performance and experience when interacting with computers. We investigated the effects of different prediction relations between delay duration and response requirements on user performance. In one experiment, delay duration predicted, to different degrees ( $50 \%$ vs. $75 \%$ vs. $100 \%$ ), the following system response. Predictability substantially increased users' response speed, while adaptation was highly flexible, between different prediction regimes. In a second experiment, users' responses predicted system delay duration. Compared to the first experiment, users' response speed was moderately increased, while the adaptation was rather inflexible across different prediction regimes. In a third experiment, we directly compared both types of predictability. The results confirmed a stronger and more flexible adaptation effect when time predicted the system response, compared to when users' responses predicted time. These findings have important implications for scheduling data transmission rates across different users in internet-based parallel computing.


# Computational Foundations of Cultural Evolution: Modeling the Emergence of Systems from Higher-order Probabilistic Inference 

Bill Thompson<br>Max Planck Institute for Psycholinguistics


#### Abstract

Cumulative cultural evolution in humans is the process through which behaviours gain structure and complexity as they are transmitted from one generation of learners to the next. A central challenge in the cultural evolution literature is to understand how the unique computational principles of human cognition scaffold the emergence of complex behavioural systems. I explore how the human ability to make inferences at higher order levels of abstraction can lead to cultural complexity, in two ways: by allowing initially independent behaviours to gradually acquire group-like structure as new learners repeatedly impose an expectation for statistical dependence; and by allowing inferences in one domain to be rapidly transferred to new domains which share features at higher-order levels of abstraction. I model these processes in populations using a probabilistic cognitive model for the acquisition of vowel systems in human language.


# Counterfactual thoughts and judgments about morally good actions 

Shane Timmons<br>TRINITY COLLEGE DUBLIN, UNIVERSITY OF DUBLIN, IRELAND<br>Ruth Byrne<br>TRINITY COLLEGE DUBLIN, UNIVERSITY OF DUBLIN, IRELAND


#### Abstract

Evaluating the morality of an action is affected by thoughts about whether the outcome might have turned out differently. We report experimental results that show a moral action effect occurs for judgments about morally good actions. Participants read stories about a morally elevating situation, e.g., an agent is found to be a match as a bone-marrow donor for someone else. The agent decided to act or not to act, and the outcome turned out well or it did not turn out well. Participants created counterfactual thoughts and they also made judgments about whether the agent should have acted, and whether the agent was morally responsible for the outcome. The results show a moral action effect: participants judged that the action should have been taken, and that the agent was morally responsible for the outcome, when the agent acted compared to when they did not act, regardless of the outcome.


# Gradually ascending sound with accelerating automatic driving vehicle might change passengers' tension or anxiety: analysis of biometrical index. 

Akitoshi Tomita<br>University of Tsukuba<br>Etsuko Harada<br>University of Tsukuba, Tsukuba, Ibaraki, Japan<br>Kozue Miyashiro<br>University of Tsukuba, Tsukuba, Ibaraki, Japan<br>Satoshi Ando<br>University of Tsukuba<br>Maito Ohmori<br>University of Tsukuba<br>\section*{Hiroaki Yano}<br>University of Tsukuba


#### Abstract

When people ride an autonomous car, they might feel anxiety because they cannot know how it may move. Adding artificially augmented signals, which represent coming changes of the vehicle, it may be useful to reduce anxiety by change expectation. Thus we executed an experiment examining whether ascending sound could decrease passenger's anxiety, while riding on virtual autonomous car. In the experiment, participants saw 360-degree computer-graphics world through a head-mounted-display. The stimuli were views from a moving car with 2 speed ( 19 and $320 \mathrm{~km} / \mathrm{h}$ ), half of which was added ascending and descending sound at first / last 6 secs. Results of the heart-wave analysis as biometric index, i.e., index of sympathetic nervous (LF/HL), showed a marginal interaction between existence of sounds and the vehicle speed; while sounds reduced participants' anxiety with high-speed condition, they showed higher tension with sound at slow-speed conditions.


# Does sonority influence the syllable segmentation in visual identification? Evidence in French skilled readers. 

Méghane Tossonian<br>Clermont Auvergne University, LAPSCO-CNRS, UMR 6024<br>Norbert Maïonchi-Pino<br>Clermont Auvergne University, LAPSCO-CNRS UMR 6024


#### Abstract

Many studies focused on the importance of statistical and distributional properties to account for the prelexical and segmental role of syllable-sized units in silent reading in French. We explored how skilled readers segmented printed (pseudo)words when no reliable statistical cues were available around and within the syllable boundary. We were interested in how sonority, a universal phonological element, might be a reliable source for syllable segmentation. We tested 160 native French-speaking adults with pseudowords in which orthographic and phonological statistical properties were (quasi)null for the first three letters including the syllable boundary in a revisited version of the paradigm used by Treiman and Chafetz (1987). Five sonority profiles within the syllable boundaries along a continuum from legal to illegal clusters were designed. Our results showed that segmentation does not strictly depend on statistical cues; participants were also sensitive to the legality of the sonority profile to locate the syllable boundary.


# Poverty of materials makes recursive combination operation evolvable 

Genta Toya<br>Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan<br>Takashi Hashimoto<br>Japan Advanced Institute of Science and Technology


#### Abstract

Humans can use recursive combination operation in various behaviors; other primates, however, rarely perform this operation. In our previous research, using an evolutionary simulation of combination behavior, we showed that recursive combination was more adaptive than repetitive combination in cases where the robustness of production or the diversity of products was required. In this research, we examined the evolvability of recursive combination in combinatorial space parametrized by kinds of elemental materials and the number of elements per product. Recursive combination evolved in the region of low kinds of elemental materials and large number of elements per product. This region may be compared with the situation of the middle stone age when invented diversified tools with limited kinds of materials such as stone, bone, and woods. The recursive combinatorial operation could scaffold the evolution of general recursive combination abilities including language, technology, music, and mathematics.


# Fitting a Stochastic Model to Eye Movement Time Series in a Categorization Task 

Paul Tupper<br>Simon Fraser University<br>Thuan Pham Nguyen<br>Simon Fraser University<br>Yunlong Nie<br>Simon Fraser University<br>Jiguo Cao<br>Simon Fraser University<br>Liangliang Wang<br>Simon Fraser University


#### Abstract

Our goal is to develop an efficient framework for fitting stochastic continuous-time models to experimental data in cognitive psychology. As a simple test problem, we consider data from an eye-tracking study of attention in learning. For each subject, the data for each trial consists of the sequence of stimulus features that the subject fixates on, together with the duration of each fixation. We fit a stochastic differential equation model to this data, using the Approximate Bayesian Computation framework. For each subject we infer posterior distributions for the unknown parameters in the model.


# Gaze during utterances and silence in L1 and L2 Conversations 

Ichiro Umata<br>KDDI Research, Inc.<br>Koki Ijuin<br>Doshisha University<br>Tsuneo Kato<br>Doshisha University<br>Seiichi Yamamoto<br>Doshisha University


#### Abstract

Gazing activities during utterances and silence were analyzed in a face-to-face three party conversation setting in a native language (L1) and in a second language (L2). The function of each utterance was categorized according to the Grounding Acts defined by Traum (Traum, 1994) so that gazes during utterances could be analyzed from the viewpoint of grounding in communication (Clark, 1996). Factor analyses of gaze activities showed similar factor structures in L1 and L2 conversations: the first factor was characterized by speakers' gazes and gazes during silence, and another factor was characterized by listeners' gazes in each condition. Analyses of the participants based on the factor scores, however, showed different tendencies between the two conditions, suggesting that language proficiency affects gaze activities during utterances.


# Statistical Learning Contributions to Semantic Knowledge Development 

Layla Unger<br>Carnegie Mellon University, Pittsburgh, Pennsylvania, USA<br>Anna Fisher<br>Carnegie Mellon University, Pittsburgh, Pennsylvania, USA<br>Robert Powers<br>Carnegie Mellon University, Pittsburgh, PA, United States


#### Abstract

The organization of semantic knowledge according to relations between concepts influences many facets of higher cognition. Therefore, understanding the origins of relations knowledge is a key focus of cognitive development research. This study investigated the contributions of environmental statistical regularities to relations knowledge in preschool-age children. Using CHILDES to estimate co-occurrence between familiar items, we constructed triads consisting of a target, related distractor, and unrelated distractor in which targets and related distractors consistently co-occurred (e.g., sock-foot), belonged to the same taxonomic category (e.g., sock-coat), or both (e.g., sock-shoe). Using a Visual World paradigm, we then measured relations knowledge as the degree to which children looked at related versus unrelated distractors when asked to look for targets. The results revealed that co-occurrence, regardless of taxonomic relatedness, influenced whether participants looked significantly more at related versus unrelated distractors. These findings demonstrate that co-occurrence regularities between entities in the environment shape knowledge organization.


# Improving Number Foundations in Preschoolers: ANS versus Symbolic Knowledge 

Jo Van Herwegen<br>Kingston University London<br>Hiwet Costa<br>Kingston University London

Bethany Nicholson

Kingston University London
Chris Donlan
University of London


#### Abstract

The current study examined whether preschoolers who are low achievers (LA) on mathematical tasks benefit more from a training programme that focuses on magnitude comparisons or ANS abilities (PLUS games) compared to games that target symbolic knowledge (DIGIT games).

Twenty-four preschoolers played PLUS games and 21 children played DIGIT games 3 times per week for 5 weeks. Performance scores were compared to 25 typical control children who did not play any games. All children were assessed pre and post-intervention on Test of Early Mathematics (TEMA), a computerized ANS task, the Give a Number task (Wynn, 1990) to assess cardinality and a counting and Digit Recognition task.

The results showed that, although the DIGIT and PLUS groups performed lower than the Control group, both PLUS and DIGIT games improved mathematical abilities in LA children. These results suggest that there is a complex interaction between ANS, symbolic, and formal mathematical abilities.


# Biological and Artificial Perspectives on Metacognition 

Katherine Wagner<br>DePaul University, Chicago, Illinois, USA


#### Abstract

Metacognition may be broadly understood as awareness, monitoring, and regulation of an intelligent agent's own internal processing, a "thinking about thinking". The cognitive complexity and self-maintenance value of this introspective skillset has considerable current interest in the study of both biological and artificial intelligence, with intriguing parallels. Study of metacognition in some nonhuman species and Biologically-Inspired Cognitive Architecture (BICA) systems reflect evidence of, at best, an attenuated form of the elaborated human manifestation, with ongoing difficulties in operationalizing metacognitive components and traits. A linked exploration of these "inhuman" forms of metacognition may better clarify the locus of divergence from the human form and illuminate the role of the skill in supporting potentially emergent cognitive traits, from self-recognition to Theory of Mind understanding. The current review will take a comparative approach in assessing metacognitive systems in nonhuman biological and artificial agents in pursuit of clarity for future methodological and conceptual directions.


# Predicting Future Performance in an ITS system via Gradient Boosting Classification 

Breya Walker<br>University of Memphis, Memphis, TN, United States<br>Anne lippert<br>University of Memphis, Memphis, TN<br>Raven Davis<br>University of Memphis, Memphis, TN<br>Zhiqiang Cai<br>The University of Memphis, Memphis, TN<br>\section*{Cheng Qinyu}<br>The University of Memphis, Memphis, TN<br>\section*{Genghu Shi}<br>The University of Memphis, Memphis, TN<br>Arthur Graesser<br>The University of Memphis


#### Abstract

Gradient Boosting Classification (GBC) models are well known to machine learning and artificial intelligence. Having the ability to predict user performance is imperative to the outcomes and purpose of an intelligent tutoring system. The Center for the Study of Adult Literacy (CSAL) intelligent tutoring system aims to improve reading comprehension in low-literacy adult learners. A GBC was applied to preliminary data gathered from high-literacy adult readers ( $\mathrm{N}=1800$ observations). Our model was shown high accuracy in predicting users' correct/incorrect responses to our multiple choice items. Specifically, users' reaction times and order of question presentation are important features of the model to consider. Less important features are difficulty of the item and the users reading ability. Our next steps are to apply GBC to high-literacy college students, followed by low-literacy readers, as a test set. Our eventual goal is to predict correctness prior to scoring.


# Interactivity, Stereotype threat, and Working memory 

Anna-Stiina Wallinheimo<br>University of Surrey, Guildford, United Kingdom<br>Harriet Tenenbaum<br>University of Surrey, Guildford, United Kingdom<br>Adrian Banks<br>University of Surrey, Guildford, United Kingdom


#### Abstract

The purpose of the current study was to investigate the role of interactivity (the use of pen and paper) in defusing the impact of stereotype threat on difficult mental arithmetic tasks, covering all four operations of mathematics. Eighty-four 16-year-old girls from secondary schools in South East England (UK) participated in this study. Participants carried out (in an educational setting) difficult, multi-digit mental arithmetic tasks in a stereotype threat or control condition, crossed with interactivity or no interactivity. The primary dependent variables were the overall performance of the participants in accuracy, latency to solution, working memory, and mathematics anxiety. Increased interactivity enhanced mental arithmetic performance. Girls in the stereotype condition performed worse in the working memory test than the participants in the control condition. However, there was no causal role of working memory in reduced mathematics performance under stereotype threat. Reasons for this finding and recommendations for future studies are discussed.


# Judging Magnitude: Is there a Common Cognitive System for Different Types of Magnitude Judgments? 

Rylan J. Waring<br>University of Guelph

Lana M. Trick

University of Guelph


#### Abstract

It has been suggested that a common cognitive system is employed in magnitude judgments across multiple modalities (Walsh, 2003). To test this theory, we examined whether performance on magnitude judgments of number, surface area, duration, and loudness correlated with each other in both magnitude comparison (e.g., determine which is more), and magnitude estimation (e.g., if magnitude 1 value $=100$, estimate the value of magnitude 2) tasks. For magnitude comparison, significant correlations were observed between number, surface area, and loudness (but not duration) tasks (percent correct measured). Similar results were observed for magnitude estimation (mean absolute percent deviation of value estimates from correct measured). These results are indicative of a common cognitive system for at least some magnitude judgment modalities, and suggest that such a system may play a role not only in more-than/less-than magnitude judgments, but also in the process of assigning numerical values to magnitudes.


# Effects of Auditory-Feedback Delays and Musical Roles on Coordinated Timing Asymmetries in Piano Duet Performance 

Auriel Washburn<br>Stanford University<br>Matthew Wright<br>Stanford University

Takako Fujioka
Stanford University


#### Abstract

Recent research in human behavioral dynamics has demonstrated that co-actors often successfully achieve joint goals by adopting functionally asymmetric patterns of behavior. To better understand the evolution of such patterns in a naturalistic musical context, the current study examined how auditory-feedback delays and individual musical roles affect collective temporal stability and relative adaptability during duet performance. The delays between pianists were short (1040 ms ), bidirectional, and remained constant during a single trial, simulating those typical in internet-mediated performance. Preliminary results show increasingly reduced collective stability for longer delays along with a distinct pattern of asynchronies across the points where temporal synchrony would be expected, in which individuals exhibited consistent alternation between playing before or after their co-performer. Furthermore, asynchronies became greater when the two musical parts were less similar. Thus, emerging coordinative dynamics appear to be shaped both by asymmetries in co-performers' assigned roles and external constraints on shared information.


# What influences the impact of warning labels in decisions from description-plus-experience 

Leonardo Weiss-Cohen<br>University College London, London, London, United Kingdom<br>Emmanouil Konstantinidis<br>University of New South Wales<br>Maarten Speekenbrink<br>University College London<br>Nigel Harvey<br>University College London, London, London, United Kingdom


#### Abstract

Warning labels can be considered as descriptions added to repeated decisions-from-experience. Limited research so far has looked at the theoretical integration of decisions from descriptions and decisions from experience when the two are available concurrently. We explore how the presence and timing of such warning labels influence behaviour. We expected the provision of warning labels to subsequently reduce risk taking, and that more prior experience before the appearance of such labels would lead to stronger habit formation and reduce their behavioural impact. Instead, we show how the appearance of descriptions warning against risks can have a perverse effect of increasing risk taking. And counter-intuitively, we also observe that the amount of previous experience prior to the appearance of descriptions does not impact behaviour. Briefly presented warning labels also have the same effect as constantly present ones. All of these findings have strong implications on the design of effective warning labels.


# Later lexical development in bilinguals 

Anne White<br>KU Leuven - University of Leuven<br>Barbara Malt<br>Lehigh University<br>Steven Verheyen<br>Institut Jean-Nicod, PSL Research University, École Normale Supérieure<br>Gert Storms<br>KU Leuven - University of Leuven


#### Abstract

We investigated object naming in Dutch-French bilingual children to determine the developmental trajectory of the cross-language convergence in naming patterns shown by bilingual adults. We collected name choices for nearly 200 common household containers from French-Dutch simultaneous bilinguals of 6 different age groups, along with monolingual control groups. Multidimensional scaling analysis on a group level suggests that convergence is present in bilinguals at all ages. On the individual level, pairwise between-subject correlations show that monolingual naming patterns in different languages show a remarkable correspondence at younger ages. Between age 5 and adulthood, the naming patterns of monolingual children demonstrate increasing divergence as they learn the language specificities of their L1. Bilingual children, however, maintain a fully converged naming pattern up till age 10 . They start learning some language-specific idiosyncrasies from age 12 onwards, but never to the extent of monolinguals. We propose a gradual divergence perspective for bilingual lexical development.


# An immersive binaural horizon for sonic data analytics 

Richard Windeyer<br>The Centre for Drama, Theatre and Performance Studies, University of Toronto and the Perceptual Artifacts Lab, OCAD University<br>Dan MacDonald<br>Biomedical Simulation Lab, Department of Mechanical and Industrial Engineering, University of Toronto<br>\section*{David Steinman}<br>Biomedical Simulation Lab, Department of Mechanical and Industrial Engineering, University of Toronto<br>Ambrose Li<br>Perceptual Artifacts Lab, OCAD University<br>Peter Coppin<br>Perceptual Artifacts Lab, Faculty of Design, OCAD University and The Department of Mechancial and Industrial Engineering, University of Toronto


#### Abstract

Accessible data analytics-that can be experienced through vision, hearing, and touch-poses a challenge to interaction design. It is also a human rights requirement because many societies mandate that all individuals have the right to experience products and services, yet not every individual accesses media visually. As more data is presented through visualization, accessibility for populations who do not access data through vision decreases.

Guidelines that claim to make visual media accessible through text fail to translate the iconic properties of visual shapes, thus subtracting affordances for pattern recognition. Non-linguistic sonication can be a means for non-visual pattern recognition.

Hearing is optimized for detecting locations on a horizontal plane, and our approach for presenting data analytics recruits this optimization by using an immersive binaural horizontal plane. We will demonstrate our approach via two case studies: A sonic translation of a map and a sonic translation of a computational fluid dynamics simulation.


# A model of cultural co-evolution of language and mindreading 

Marieke Woensdregt<br>University of Edinburgh, Edinburgh, United Kingdom<br>Simon Kirby<br>University of Edinburgh<br>\section*{Chris Cummins}<br>University of Edinburgh<br>\section*{Kenny Smith}<br>University of Edinburgh


#### Abstract

Language requires mindreading for entertaining communicative intentions, and mindreading in turn profits from language as a means for sharing mental states. Hence it has been hypothesised that the two skills have co-evolved.

We present a Bayesian agent-based model to formalise this hypothesis. This model combines referential signalling with mental states, such that a speaker's topic choice is probabilistically dependent on their perspective on the world. In order to learn the language, a learner has to simultaneously infer the speaker's lexicon and perspective. Learners can solve this task by bootstrapping one with the other, but only if the speaker uses an informative language.

We will present results of an iterated learning version of this model, showing that selection on communication results in the emergence of a fully informative lexicon from scratch. However, selection on perspective-taking alone also results in the emergence of partially-informative lexicons, which is sufficient for inferring others' perspectives.


# Language as a process: An exploration among pre-adolescent Chinese EFLs 

Hing Yi Wong<br>The Education University of Hong Kong, Hong Kong, Canton, China<br>Susanna Yeung<br>The Education University of Hong Kong<br>Wen Chung Wang<br>The Education University of Hong Kong<br>Chi Wing Yip<br>The Education University of Hong Kong


#### Abstract

This paper reports a reading intervention programme, the LMVS (Linguistically Mediated Visual Search) among pre-adolescent Chinese EFLs. It sets out to test whether managing the process of silent reading might modify text complexity as perceived. The paper is a combination of two studies. The first study was the development and assessment of a reading comprehension test. The second study piloted an intervention for pre-adolescents. Item-by-item analysis of students' performance in the post-test show changes in the perception of item difficulty after the intervention. Chinese EFL struggling readers were found to be weaker in lexical analysis. They also faced difficulties in decoding main ideas in compound/complex sentences. In response to the analysis, strategies were developed for automatic syntactic processing. The paper proposes seeing language as a process, rather than a product so that learner management skills might be prepared for reading intervention.


# Influences of the Matching Effects of Cognitive and Emotional Factors on Attitude Change 

Tung-Cheng Wu<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC<br>Jon-Fan Hu<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC<br>Shu-Ling Peng<br>National Cheng Kung Univeristy, Tainan, Taiwan, ROC


#### Abstract

This research is aimed to study whether we will have the same attitude change when we have the same intensity of cognitive or emotional level for attitude. People who have higher involvement will be persuaded by central route and low involvement will be persuaded by peripheral route: the matching effect in attitude change. The present study controlled the intensity of both cognitive and emotional factor and instructed participants to express their initial attitude as well as attitude change under four experimental manipulations. Results showed that only matching effect of emotional factor was found but not cognitive factor. A connectionist model was therefore built to simulate the processes and found that there would have different thresholds for cognitive and emotional routes and the threshold of cognitive route should be higher than emotional route. Implications are proposed based on the behavioral and simulation investigations.


# Is infants' mutual exclusivity response based on preference to novelty or non-name of an object? 

Qinmei Xu<br>Center for Learning and Cognitive Science, College of Education, Zhejiang University, Hangzhou, China<br>Ye Tao<br>Department of Special Education, Zhejiang Normal University, Hangzhou, China<br>Yuyan Wang<br>Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China<br>\section*{Jon-Fan Hu}<br>National Cheng Kung University, Tainan, Taiwan, ROC


#### Abstract

Although "mutual exclusivity (ME)" is the term to refer to the behavior that infants map a novel label onto a novel object rather than a familiar object, two studies, using preferential looking paradigm, aimed to investigate whether infants' ME is based on preference to novelty or non-name of an object. In Study 1, 18-month-olds were tested on 2 conditions: familiar-object/novel-object trials with known label and familiar-object/novel-object trials with unknown label. The infants preferred to novel objects before naming but no naming effect found for both conditions. In Study 2, 18-month-olds in the same two conditions as Study 1 were pre-familiarized to both of novel and familiar objects. The results showed that the naming effects were found for both conditions, indicating that ME occurred. The findings of the present studies suggest that pre-familiarization could be used to validate if 18 -month-olds' ME response is based on non-name preference of an object.


# Children's Attention to Semantic Content versus Emotional Tone: Differences between Two Cultural Groups 

Yang Yang<br>Cornell University, Ithaca, NY, United States

Li Wang
Peking University, Beijing, China
Qi Wang
Cornell University, Ithaca, NY, United States


#### Abstract

People from varied cultural backgrounds differ in their attention to particular aspects of emotional cues. Whereas semantic content explicitly expresses feelings, vocal tone conveys implicit information regarding emotions. This study examined the attention to different emotional cues in European-American and Chinese children. Participants were 121 EuropeanAmerican and 120 Chinese children (4-9 years old). They played two games in which they listened to spoken words and judged the pleasantness of the word meaning while ignoring the vocal tone (meaning game) or judged the pleasantness of the vocal tone while ignoring the word meaning (tone game). Preliminary results showed that European-American children paid more attention to word meaning than did Chinese children. Additionally, older (8-9 years old) Chinese children attended more to vocal tone than did their European-American counterparts. The results suggest that children acquire culturally specific attention bias by 8-9 years old.


# Knowledge partitioning in forecasting 

Lee-Xieng Yang<br>National Chengchi University<br>\section*{Tzu-Hsi Lee}<br>National Chengchi University


#### Abstract

In this study, we would like to examine whether the learned forecasting function can be separated for use by context. The participants were asked to learn to forecast the position of a target, defined as a sine function of trial number. A context cue was paired with the moving of the target systematically and randomly in two conditions. The learning performance was quite good in both conditions. In the transfer phases, in the systematic-context condition, some participants learned to rely on context to direct their prediction (i.e., knowledge partitioning), whereas some others and those in the randomized-context condition learned to rely on the concept about the function for forecasting. However, contrary to the precedent knowledge partitioning studies, the variety of using context or not was found within participants across transfer phases. The modeling results favored the associative account over the rule account on accommodating the training and transfer response patterns.


# Working Memory and lexical ambiguity resolution in Chinese 

Michael Yip<br>The Education University of Hong Kong


#### Abstract

Two cross-modal priming experiments were conducted to examine the underlying mechanism of lexical disambiguation process was in activational nature or in inhibitory approach. In experiment one, forty native Cantonese listeners were recruited to participate in two tasks (1) a Chinese version reading span task (Daneman \& Carpenter, 1980) to measure their WM capacity and (2) a cross-modal priming task (Yip, 2015). In experiment two, another group of native Mandarin listeners were recruited to participate in the same two tasks in Mandarin. The results revealed that sentence context had an early effect on the disambiguation processes for both high- and low-WM span groups and the underlying mechanism of the disambiguation process for the high-WM span group seemed to be in an inhibitory nature.


# Impact of Polarity, Rationality, and Math Ability on Numerical Magnitude Knowledge 

Laura Young<br>Temple University, Philadelphia, PA , United States<br>Julie Booth<br>Temple University, Philadelphia, PA , United States<br>Kelly McGinn<br>Temple University, Philadelphia, PA , United States


#### Abstract

Previous research has shown that numerical magnitude knowledge is related to current mathematic abilities and predictive of future mathematics performance. Much of this early research examined magnitude knowledge of positive whole numbers, more recently this has been extended to positive rational numbers. However, research about negative number magnitude knowledge is less abundant. The present study aims to understand how different types of magnitude knowledge relate to one another and whether performance differs according to the type of number line scale. Thirteen number line scales were used to assess 7th grade students' $(\mathrm{N}=180)$ magnitude knowledge of positive and negative, whole and rational numbers. Correlational analyses illustrate that performance on most scales are significantly related. Further analyses reveal that students' performance differed depending on the scale's polarity and the number type of the scale. Moreover, performance differences were found to vary according to students' mathematics classroom ability level. Educational implications are discussed.


# Global consequences of local complexity: evidence from recall of visually presented nonwords 

Mackenzie Young<br>Johns Hopkins University

Colin Wilson<br>Johns Hopkins University


#### Abstract

There is extensive evidence that structural regularities affect the processing of visually-presented words. However, it is not known whether the processing consequences of an orthographic violation are limited to the offending subpart (e.g., an unattested onset cluster) or apply more globally (e.g., to the entire word). We provide evidence of global disruption from the recall of briefly-presented nonwords that were manipulated for degree of orthographic markedness and length. Error rates were higher for both the initial and final portions of nonwords beginning with more marked onsets; symmetrically, report of marked onsets was degraded in words with longer endings. These effects suggest that, as in other visual tasks, the fidelity with which one element can be represented depends on the overall stimulus complexity. We present a modified version of rational models of visual word perception in which global effects result from the distribution of a limited processing resource over letter positions.


# Children's Reasoning about Geometric Footprints 

Christopher Young<br>University of Chicago, Chicago, Illinois, United States

Alana Foley<br>University of Chicago

## Susan Levine

University of Chicago


#### Abstract

We explored preschool's children's understanding of the correspondence of 3-D objects and 2-D faces in a novel task. In the "footprints" task children were shown a geometric solid, such as a pyramid or a prism, and asked to select which shape the solid would make if it were dipped in ink and stamped on a piece of paper. Through a latent class analysis of children's errors we found children differed significantly in their misconceptions about object structure. Three distinct classes of children emerged: children who could only match visible faces, children who believed solids have an 'essential' face irrespective of rotation, and children who differentiated faces based on a solid's rotation. We examined the characteristics of children in each of these classes using a battery of spatial tasks and numeric tasks. Our results suggest errors found in older children's and adults' reasoning about geometric concepts develop prior to formal schooling.


# Relevance Theory, Pragmatic Inference and Cognitive Architecture 

Wen Yuan<br>Beihang University, Beijing, China<br>Richard Cooper<br>Birkbeck, University of London


#### Abstract

Relevance Theory (RT: Sperber\&Wilson, 1986) argues that human interpretative processes maximize relevance and postulates that there is a relevance-based procedure that a listener follows when trying to understand utterances. However, Mazzone (2013) points out that RT fails to explain how speaker-related information, such as the speaker's abilities or preferences, is incorporated into pragmatic processes. He proposes that situational or goal schemata, with the speaker represented as a component, are sufficient to activate the hearer's speaker-related knowledge and further asserts that human communication is driven by goal management and action rather than relevance maximization. Yet Mazzone cannot fully explain how linguistic meaning and speaker-related knowledge are integrated within a modular system. Based on RT's cognitive requirements and contemporary cognitive theory, we argue that this integration is realized within working memory via production-like conversational rules with which the constructed utterance interpretation should be consistent, and present a simple model of this process.


# Do you forgive past mistakes of virtual assistant? A study on changing impressions of virtual character when using its assistance multiple times 

Masahide Yuasa<br>Shonan Institute of Technology


#### Abstract

We investigated the gain-loss effect of virtual/personal assistant character, which provides intelligent assistance to humans, and focused on not only the first impression of using the assistance but also changing impression about the character when used multiple times for assistance. The experiment used a fictive retrieval system (searching onomatopoeia), and the virtual assistant character looked up for suitable words for the user (success), or failed to find the words (mistake). There were three sessions, differing by the task of character's mistake; two tasks were successes and one was a failure in each session. The results showed that the group of people who had low expectation from its first appearance, formed negative impressions after the final mistake, significantly. Consequently, final mistake influenced the formation of negative impression more than other mistakes, thus showing that the final mistake in multiple times of assistance was associated with loss effect.


# Knowledge acquiring on event chronology in Russian-language texts 

Vera Zabotkina<br>Russian State University for the Humanities<br>Mikhail Epifanov<br>Russian State University for the Humanities<br>Elena Pozdnyakova<br>Moscow State Institute of Foreign Relations (University),


#### Abstract

Chaining the sequence of events through awareness of their temporal relations is an important aspect of text understanding. As a rule, text provides only partial knowledge of event unfolding however various types of additional sources (documents, personal diaries, etc.) provide an added knowledge to make chronology more precise. The paper argues the novel approach to automated retrieval of information on temporal relations between events marked in the text. The data retrieved will provide additions to computer ontology which formally represents the actual events chronology.

A system of linguistic algorithms for analyzing the contexts with specific verbal (or linguistic situations) inputs is suggested within the present approach. We use syntactic graphs of the sentences and some grammatical characteristics of the words produced by the system for the automatic syntactic analysis of Russian texts.


# The Impact of Presentation Order on Category Learning Strategies: Behavioral Data and Self-Reports 

Christina Zeller<br>University of Bamberg, Germany

## Ute Schmid

University of Bamberg, Germany


#### Abstract

The presentation order in supervised categorization learning can influence the category representation. For example, the order can bias a rule-based approach focusing the identification of relevant features or an exemplar-based approach focusing the similarity of category members. In a blocked design stimuli can either be presented in a way that relevant features over stimuli become obvious or that two succeeding stimuli share as many common features as possible (cf. Mathy \& Feldman, 2016). In an empirical study with 21 participants we investigated both orders for the 5-4 category structure (Medin \& Schaffer, 1978) and assessed categorization behavior and self-reports in the first trials. Results suggest that the answer behavior and selfreports during the first trials can be influenced by the presentation order. However, in both groups feature-based and similaritybased explanations were reported. Additionally, the similarity-based group named more feature-based rules including irrelevant features.


# Walking dynamics of intertemporal choice 

Arkady Zgonnikov<br>School of Psychology, National University of Ireland, Galway<br>Iñaki Rañó<br>Intelligent Systems Research Centre, University of Ulster, Derry/Londonderry, UK<br>Denis O’Hora<br>School of Psychology, National University of Ireland, Galway<br>\section*{Kongfatt Wong-Lin}<br>Intelligent Systems Research Centre, University of Ulster, Derry/Londonderry, UK


#### Abstract

The notion that cognitive processes 'leak" into motor output of decisions inspired much recent process-tracing research. In mouse-tracking, an increasingly popular decision-making paradigm, difficult choices lead to increased curvature of the mouse trajectories towards the unchosen option. Here we explore whether traces of a decision process can be found in its motor output in a more naturalistic setting. Our subjects performed a series of choices between a smaller reward now and a larger reward at some delay. Using Kinect camera, we recorded subjects' walking trajectories when they moved towards their preferred option displayed in one of the corners across the room. We found that deviation of subjects' trajectories from the ideal trajectory increased with delay when they preferred the "later" option, and decreased with delay in trials where the "now" option was chosen. Our results suggest that walking trajectory of a person can provide information about their ongoing thought processes.


# A Spatial-Temporal Analysis of a Visual Working Memory Task with EEG and ECoG 

Qiong Zhang<br>Carnegie Mellon University, Pittsburgh, Pennsylvania, United States<br>Marieke van Vugt<br>University of Groningen, Groningen, Groningen, Netherlands<br>Jelmer Borst<br>University of Groningen, Groningen, Groningen, Netherlands<br>John Anderson<br>Carnegie Mellon University, Pittsburgh, Pennsylvania, United States


#### Abstract

In this study, we investigated the neural correlates of a visual working memory task. Two experiments were carried out using scalp electroencephalography (EEG) and Electrocorticography (ECoG), respectively. In each trial, participants judged whether a test face had been among a small set of recently studied faces. We used a combination of hidden semi-Markov models (HSMMs) and multi-variate pattern analysis (MVPA) to decompose the neural signal into a sequence of latent stages. Analyzed separately, EEG and ECoG data yielded converging results on the durations of recovered stages. Combining these stages with the high spatial resolution of ECoG suggested that activity in the temporal cortex reflected item familiarity in the retrieval stage; and that once retrieval is complete, there is active maintenance of the studied face set in the medial temporal lobe (MTL). During this same period, the frontal lobe guides the decision by means of theta coupling with the MTL.


# To organize or not to organize? Examining biases in search strategies using Lego building blocks 

Mona Zhu<br>University of Waterloo, Waterloo, Ontario, Canada<br>Kimberley Yuen<br>University of Waterloo, Waterloo, Ontario, Canada<br>Evan Risko<br>University of Waterloo, Waterloo, Ontario, Canada


#### Abstract

A widely-accepted notion is that organization can improve task performance and generally allow us to better function within a given task environment (Kirsh, 1995; 1996). However, it remains unclear the extent to which individuals believe that organization will help to improve task performance when they are asked to carry out mundane tasks in the real world. To examine this, individuals were asked to search through a pile of Lego building blocks for specific pieces. Prior to the search task, they were asked their preferred strategy for this task (e.g., organizing vs. not organizing the Lego pile prior to search) and to estimate how much time and effort each strategy would take for task completion. While both strategies were comparable in terms of objective task completion time and subjective time and effort estimations, participants were overwhelmingly biased against choosing the organization strategy. Implications for the current study will be discussed.


# Emotion in Deceptive Language 

Iraide Zipitria

University of the Basque Country (UPV/EHU)

## Basilio Sierra

University of the Basque Country (UPV/EHU)

## Imanol Sopena-Garaikoetxea

University of the Basque Country (UPV/EHU)


#### Abstract

Deception involves emotions of fear and guilt. These negative emotions are expressed in language in terms of psychological distance from the deception object. The psychological distance and emotional experience reflect an attempt to control the negative mental representation. More especifically emotional distance is represented in deceptive language by means of cues of reference, verb tense and detail avoidance. Then, hints of emotions of fear and guilt should be displayed in language.The present work analyses emotional language cues for deception detection by means of Machine Learning(ML) techniques and Linguistic Inquiry and Word Count (LIWC). Results show that Support Vector Machines (SVM) best represents the discrimination between true and false information (up to $74.15 \%$ of accuracy rates) based only on the effect of emotion in deceptive speech.


## Author Index

Abashidze, Dato ..... 1483
Abney, Drew H. ..... 1369, 3698
Abney, Drew ..... 39, 1489
Abramova, Ekaterina ..... 45
Acartürk, Cengiz ..... 1495
Acharya, Aditya ..... 51
Aczel, Balazs ..... 3806
Adam, Vincent ..... 3764
Adams, Cody ..... 3566
Adams, Katherine A. ..... 57
Adams, Katherine ..... 1501
Adank, Patti ..... 3552
Aguiar, Nathan ..... 3664
Aguirre-Celis, Nora E ..... 1507
Ahissar, Merav 2267, 2419, 3764, 3773
Ahluwalia, Daljit ..... 1513
Ahmed, Nesreen K. ..... 895
Ainooson, James ..... 1519
Albert, Saul ..... 7, 1525
Albouy, Philippe ..... 3773
Aldercotte, Amanda ..... 3809
Aleni, Andrea ..... 1424
Alhama, Raquel G. ..... 1531
Alibali, Martha ..... 3578
Alikhani, Malihe ..... 2876
Alireza, Hala ..... 63
Allaham, Mowafak ..... 1035
Allan, Jinan ..... 3566
Allen, Kelsey ..... 3670
Allen, Laura ..... 1884, 2573
Almut, Hupbach ..... 3738
Alonso-Diaz, Santiago ..... 3411
Althaus, Nadja ..... 1537, 3267
Altman, Sara ..... 162
Amatuni, Andrei ..... 1538
Ambridge, Ben ..... 3699
Aminoff, Elissa ..... 136
Andersen-Green, Cleo ..... 1937
Anderson, Erin ..... 1544
Anderson, John ..... 1175, 3890
Anderson, Maxwell ..... 2114
Anderson, Michael ..... 1150
Ando, Satoshi ..... 3860
Andonova, Elena ..... 3633
Andrews, Chelsea ..... 3671
Andrews, Janet ..... 1550
Andrews, Mark ..... 69, 3398
Arai, Noriko ..... 1556
Arai, Teiko ..... 1556
Aravena, Carla ..... 1562
Armstrong, Blair ..... 2693
Arroyo, Jeffrey ..... 1925
Arshamian, Artin ..... 258
Arunima, Sarin ..... 3672
Asaba, Mika ..... 458
Aslanidou, Asimina ..... 2925
Asmuth, Jennifer ..... 277
Astuti, Rita ..... 21
Atalla, Chad ..... 1096
Au, Terry Kit-fong ..... 3386
Aulet, Lauren ..... 75
Austerweil, Joseph L. ..... 3646
Austerweil, Joseph ... 526, 1010, 1035, ..... 1812
Ayaz, Hasan ..... 3739
Azar, Zeynep ..... 81
Azevedo, Roger ..... 1666, 2766
Azumah, E. Malemma ..... 3673
B. Tenenbaum, Joshua ..... 1284
Bach, Francis ..... 3833
Backus, Ad ..... 81
Baggs, Ed ..... 1568, 1568
Bagley, Aaron ..... 2482
Bago, Bence ..... 87
Bailey, Richard ..... 2657
Baker, Lewis ..... 1574
Balasubramaniam, Ramesh ..... 2120
Balboa Bandeira, Yolanda ..... 2432
Balcı, Fuat ..... 1477
Ballarini, Cristina ..... 1580
Banerji, Devika ..... 3674, 3822

## Author Index

Bangerter, Adrian ..... 1641
Bernabeu, Pablo ..... 1629
Banks, Adrian ..... 3869
Banner, Stephanie ..... 488
Bapi, Raju ..... 3052, 3756
Bardolph, Megan D. ..... 1586
Barisic, Iva ..... 3675
Barner, David ..... 1193
Barnes, Jordan ..... 588
Barque-Duran, Albert ..... 93, 1592, 3572
Barrett, Trevor J. ..... 3729
Bartels, Daniel ..... 2748
Bass, Ilona ..... 99, 105
Bat-El, Outi ..... 3692
Bateman, Joel Edward ..... 1598
Batinic, Jelena ..... 1604
Battaglia, Peter ..... 1, 2956
Battleday, Ruairidh ..... 111
Bauer, Drake ..... 1610
Beauchamp, Michael ..... 3816
Beekhuizen, Barend ..... 112, 853
Beesley, Tom ..... 3484
Begg, Steve ..... 3497
Begum, Habiba ..... 1836
Beierholm, Ulrik ..... 1381
Belardinelli, Anna ..... 3676
Belledonne, Mario ..... 1393
Beller, Sieghard ..... 21, 118, 1617, 3829
Bellingham, Erika ..... 343Bello, Paul1611, 1672, 2640
Bender, Andrea ..... 21, 118, 1617, 3829
Benjamin, Aaron ..... 1623
Bennett-Pierre, Grace ..... 458
Bennett, Stephen ..... 1623
Bentzen, Martin ..... 2494
Berent, Iris ..... 3692
Bergelson, Elika ..... 711, 1538
Bergen, Benjamin ..... 865, 2193
Bergenholtz, Carsten ..... 11
Bergmann, Christina ..... 124
Berkowitsch, Nicolas ..... 2457
Berner, Catherine ..... 3241
Betts, Hannah ..... 1463
Betz, Nicole ..... 33, 3690
Bhatia, Sudeep ..... 421, 1436, 1635
Bicknell, Klinton ..... 301
Bietti, Lucas ..... 1641
Birney, Damian Patrick ..... 1598
Birney, Damian ..... 1943
Bittner, Dana-Lis ..... 3792
Bixter, Michael ..... 1770
Bjork, Elizabeth ..... 3615
Bjork, Robert ..... 3615
Black, Alexis ..... 124
Black, John B. ..... 1163
Blaha, Leslie ..... 3677, 3733
Blair, Mark ..... 588
Blake, Peter ..... 1764
Blakeman, Sam ..... 130
Blancke, Stefaan ..... 33
Blanco, Nathaniel ..... 3678
Blaser, Erik ..... 3841
Blauch, Nicholas ..... 136
Blokpoel, Mark ..... 1418
Blonder, Aviva ..... 780
Blott, Lena M. ..... 1647
Blouw, Peter ..... 142
Blundell, Charles ..... 1319
Bnaya, Zahy ..... 1254
Bogaerts, Louisa ..... 2592
Boghrati, Reihane ..... 1648, 2320
Bohnemeyer, Juergen ..... 3332
Bohnemeyer, Jürgen ..... 343
Boldin, Alex M ..... 3740
Bolognese, Martina ..... 3039
Bonawitz, Elizabeth ..... 99, 105, 1406, 1574
Bonnefon, Jean-Francois ..... 148
Boone, Alexander P. ..... 3729
Booth, Julie ..... 2687, 3679, 3882
Bopp, Julia A. ..... 11

## Author Index

Bornstein, Aaron ..... 637
Boroditsky, Lera ..... 1060
Borst, Jelmer ..... 3890
Bortfeld, Heather ..... 3816
Bos, Nathan ..... 2992
Bosch, David ..... 149
Bott, Lewis ..... 2969
Bott, Oliver ..... 1654
Botvinick, Matt ..... 1319
Botvinick, Matthew ..... 1, 1327
Bourgin, David ..... 1660, 3724
Bourne, Janet ..... 3680
Bowers, Jeffrey ..... 1442
Bradbury, Amanda ..... 1666
Bramley, Neil ..... 150, 1047, 1788
Brand, James ..... 817
Brandone, Amanda ..... 3791
Branigan, Holly ..... 1836, 3850
Braquet, Guillaume ..... 156
Braus, Niels ..... 3460, 3466
Brentari, Diane ..... 3692
Bridewell, Will ..... 1672, 2640
Bridgers, Sophie ..... 162
Briggs, Gordon ..... 1611, 1672, 2085
Brochhagen, Thomas ..... 1678
Broersma, Mirjam ..... 3392
Brondi, Marco ..... 2138
Broschard, Matthew ..... 1453
Brouillet, Denis ..... 2729
Brown, Scott ..... 847
Brucker, Birgit ..... 168, 174
Brunel, Lionel ..... 1684, 1854
Bu, Yijie ..... 3491
Bucciarelli, Monica ..... 1690
Bucher, Leandra ..... 1806
Buchsbaum, Daphna ..... 501, 2840, 2894
Buckle, Leone ..... 1691
Buder, Eugene ..... 3721
Buechel, Sven ..... 180
Buehner, Marc ..... 3137
Buerkin-Pontrelli, Angelica ..... 3681
Bullinaria, John ..... 2549
Bülthoff, Heinrich H. ..... 525, 1156
Bülthoff, Heinrich H ..... 650
Bülthoff, Heinrich ..... 3702
Bunji, Kyosuke ..... 1556
Burchill, Zachary ..... 2616
Burns, Bruce ..... 1697, 1908
Buschkuehl, Martin ..... 2747
Bushong, Wednesday ..... 186
Butz, Martin V. ..... 2634, 3803, 3836
Butz, Martin ..... 25, 452, 525, 3676
Byrd Hornburg, Caroline ..... 600
Byrne, Ruth ..... 2839, 3859
Cai, Zhenguang ..... 192
Cai, Zhiqiang ..... 3766, 3868
Cakir, Murat Perit ..... 3739
Call, Josep ..... 3721
Callaway, Frederick ..... 1703
Calvillo, Jesus ..... 1709
Campbell, Jamie ..... 3682
Campbell, Susan ..... 2827
Campero, Andres ..... 1715
Can, Burcu ..... 1495
Cañal-Bruland, Rouwen ..... 3702
Candy, Rowan ..... 3694
Cao, Jiguo ..... 3863
Caplan, Spencer ..... 3683
Carey, Susan ..... 3754
Carlson, Craig ..... 3155
Carney, James ..... 1444, 3684
Carstensen, Alexandra ..... 1721
Cartmill, Erica ..... 3685
Carvalho, Maira B. ..... 1727
Carvalho, Paulo ..... 198, 204
Casarotto, Blair ..... 2986
Caso, Andrea ..... 210
Cassani, Giovanni ..... 216, 433
Castro, Leyre ..... 1453, 1469
Cathcart, Chundra ..... 1733

## Author Index

Catrambone, Richard ..... 3596
Cipollini, Ben ..... 2931
Cecala, Aaron ..... 3810
Cerezo Falco, Christopher ..... 2438
Cevolani, Gustavo ..... 1459
Chai, Jun Ho ..... 1739
Chakrabarti, Bhismadev ..... 2815
Chambers, Craig ..... 730
Chan, Antoni B. ..... 1430
Chan, Hei Yan Veronica ..... 2561
Chandran. E, David ..... 2333
Chandrasekaran, Sumathi ..... 3743
Chang, Ya-Ning ..... 811, 1740
Chang, Yin-Juei ..... 1544
Chater, Nick ..... 3784
Chatterjee, Anjan ..... 1471
chekaf, mustapha ..... 3779
Chemero, Anthony ..... 871, 2784
Chen, Danlei ..... 1752
Chen, Dawn ..... 1746
Chen, Hangyu ..... 3386
Chen, Jenn-Yeu ..... 3686
Chen, Xiuli ..... 51
Chen, Yalin ..... 3682
Cheng, Patricia ..... 1758, 3033
Cheng, Qinyu ..... 3766
Chernyak, Nadia ..... 1764
Cheyette, Samuel ..... 222
Chi, Min ..... 3652
Chng, Yee SiangChong, Amy Wing-Lam1770
Chopin, Nicolas ..... 1278
Chou, Yung-Yi Juliet ..... 1776
Christakou, Anastasia ..... 2815
Christiansen, Morten H. ..... 564, 2681
Christiansen, Morten ..... 2055
Chuderski, Adam ..... 2290, 3688
Chuey, Aaron ..... 3368
Chun, Elliot ..... 3680
Cichon, Kamil ..... 651
Ciechanowska, Iwona ..... 3688
Civile, Ciro ..... 1782
Clarke, Alasdair ..... 2279
Clément, Evelyne ..... 2957
Clerkin, Elizabeth ..... 3689
Clewley, Danielle ..... 3766
Clothier, Karen ..... 228
Coates, Emily ..... 3671
Coenen, Anna ..... 1788
Cohen, Jonathan D. ..... 895, 3767
Cohen, Jonathan ..... 637, 829
Cohen, Samantha ..... 234
Cohn, Neil ..... 23, 1794
Cokely, Edward ..... 3566
Coley, John ..... 33, 3690
Collins, Michael ..... 1800
Collins, Peter ..... 2476
Colombo, Matteo ..... 1806
Colon-Acosta, Nirmaliz ..... 1163
Coman, Alin ..... 3453
Conaway, Nolan ..... 1812
Concannon, Shauna ..... 1818
Connell, Louise ..... 1444, 2604, 3770
Connor Desai, Saoirse ..... 1824
Cooper, Richard P. ..... 210
Cooper, Richard ..... 29, 3885
Cooper, Rick ..... 3716, 3728
Cooperrider, Kensy ..... 240
Copley, Bridget ..... 3839
Coppin, Peter ..... 3691, 3874
Corbin, Jonathan ..... 2518
Corral, Daniel ..... 1830
Corredor, Javier Alejandro ..... 2073
Cortesa, Cathryn S. ..... 246
Cosentino, Erica ..... 3504
Costa, Hiwet ..... 3866
Costello, Fintan ..... 252, 2224
Cottrell, Garrison ..... 2931
Cottrell, Gary ..... 1096
Coulson, Seana ..... 1586

## Author Index

Coventry, Kenny ..... 451, 1445
Davis, Isaac ..... 1902
Cowan, Benjamin R. ..... 1836
Cowan, Nelson ..... 3779
Cowan, William ..... 2475
Crawford, L. Elizabeth ..... 2518
Creem, Daniel ..... 2803
Cress, Ulrike ..... 650
Crocker, Matthew .. 2061, 2261, 3168, 3356
Davis, Raven ..... 3766, 3868
Davis, Tehran J. ..... 706
Davis, Tyler ..... 2102, 3827
Davis, Zachary ..... 1896
Daw, Nathaniel D ..... 3740
Daw, Nathaniel ..... 637, 3760
Day, Samuel ..... 277
Croijmans, Ilja ..... 258, 3227
Dayan, Peter ..... 2956
Cross, Liam ..... 3770
Crossley, Scott ..... 1884
Crupi, Vincenzo ..... 1459
Cruz, Nicole ..... 259
Culbertson, Jennifer .. 156, 643, 1023, 2760, 3095, 3840
Cummins, Chris ..... 3875
Cummins, Fred ..... 3739
Cusimano, Corey ..... 1842, 1848
Cutler, Anne ..... 1447
Czarnowski, Daniel ..... 3720
Czernecka, Karolina ..... 2037
Da Silva, Fabrice ..... 1854
Daelemans, Walter ..... 216, 433
Dai, Dawei ..... 3491
Dai, Junyi ..... 265
Daikhin, Luba ..... 2419
Dalca, Ioana ..... 1060
Dale, Rick ..... $39,3704,3725$
Dam, Gregory ..... 1860
Damen, Debby ..... 1866
Dana, Qatherine ..... 3692
Dancy, Christopher ..... 1872
Darby, Kevin ..... 3693
Das, Arindam ..... 1878
Das, Olivia ..... 1878
Dasari, Aneesha ..... 3374
Dascalu, Mihai ..... 1884
Dasgupta, Ishita ..... 1890
Davenport, Jodi ..... 600
Davies, Jim ..... 271, 2296
Davis-Stober, Clintin ..... 1436
de Koning, Björn ..... 168
de La Rosa, Stephan ..... 650, 3702
de Lacy, Paul ..... 3821
de Leeuw, Joshua ..... 1550
de Marneffe, Marie-Catherine ..... 1144
de Nesnera, Kristin ..... 3690
De Neys, Wim ..... 87, 148, 391
De Ruiter, J.p. ..... 1525
Deane, George ..... 2395
Degen, Judith ..... 3753
Dehghani, Morteza 1648, 2223, 2320, 2407, 2413
Dekel, Shir ..... 1908
Delaney-Busch, Nathaniel ..... 283
Delfabbro, Paul ..... 2211
Delogu, Francesca ..... 3356
demayo, Benjamin ..... 495
Demberg, Vera ..... 2156
Demestre, Josep ..... 2628
Deng, Wei (sophia) ..... 3737
Dennis, Simon ..... 295, 3590
Dennis, Tammy ..... 289
Depascale, Mary ..... 786
Deserio, Christina ..... 3694
Dessalles, Jean-Louis ..... 1914
Devlin, Sam ..... 3249
Dewhurst, Joe ..... 2395
Dey, Biswadip ..... 829, 895
Dias Martins, Mauricio de Jesus ..... 1920
Diederich, Adele ..... 1351
Diekamp, Dorena ..... 3707
Dietsche, Lena ..... 3706

## Author Index

Dietz Saldanha, Emmanuelle-Anna ..... 883
Dils, Alexia Toskos ..... 1925
Dimitrova, Diana ..... 3695
Dimov, Cvetomir ..... 1931
Dingemanse, Mark ..... 2055
Dmello, Sidney ..... 3701
Doebel, Sabine ..... 1937
Dolscheid, Sarah ..... 3227
Domijan, Dražen ..... 3696
Donelson, Katharine ..... 343
Donkin, Chris ..... 847
Donkin, Christopher ..... 736
Donlan, Chris ..... 3866
Donoghue, Thomas ..... 2067
Double, Kit ..... 1943
Douglas, Hannah M. ..... 1949
Doumas, Alex ..... 1955
Drach, Rae ..... 3453
Drachen, Anders ..... 3249
Drenhaus, Heiner ..... 2261, 3168
Du, Jing ..... 343
Duan, Yunyan ..... 301
Dubey, Rachit ..... 307
Duffy, Sam ..... 1961
Dujmović, Marin ..... 3429
Dunn, John ..... 625, 1138
Dunn, Tim ..... 3826
Dupoux, Emmanuel ..... 3833
Duran, Nicholas ..... 3697
Durgin, Frank ..... 1967
Dweck, Carol ..... 1333
Dye, Melody ..... 313, 319, 325, 1973
Dziobek, Isabel ..... 2815
Easvaradoss, Veena ..... 2333, 3743
Eddy, Caleb ..... 2383
Edmiston, Pierce ..... 331
Edmunds, Charlotte ..... 1979Ehlis, Ann-Christine168
Eidels, Ami ..... 3338
Ejova, Anastasia ..... 397
Elfers, Melissa ..... 3810
Eliasmith, Chris ..... 142, 631, 2125, 3131
Ellefson (1), Michelle R. ..... 3710
Ellefson, Michelle ..... 3805, 3809
Elman, Jeffrey ..... 337
Elmlinger, Steven L. ..... 3698
Elqayam, Shira ..... 3528
Elsbeth, Stern ..... 2809
Emberson, Lauren L ..... 3740
Eng, Cassondra M. ..... 415
Engelmann, Felix ..... 3699
Engelthaler, Tomas ..... 3700
Epifanov, Mikhail ..... 3887
Eryılmaz, Kerem ..... 760
Espino, Orlando ..... 2839
Esposito, Roberto ..... 1997
Evans, Bronwen ..... 3552
Evans, Nathan ..... 295
Evers, Stephanie ..... 343
Faber, Myrthe ..... 3701
Fabry, Regina E. ..... 349
Fademrecht, Laura ..... 3702
Fan, Lifeng ..... 1066
Fang, Siyuan ..... 1985
Farran, Emily ..... 538, 3718
Farrell, Simon ..... 3410
Faure, Sylvane ..... 3761, 3820
Favela, Luis H. ..... 2796
Fay, Nicolas ..... 494, 1991
Fedor, Anna ..... 63
Feeney, Aidan ..... 1205
Feher, Olga ..... 3703, 3840
Fein, Deborah ..... 913
Feldman, Laurie ..... 2537, 3704
Feltz, Adam ..... 3566
Fenoglio, Giorgia ..... 1997
Ferdinand, Vanessa ..... 3705
Fernald, Anne ..... 780
Fernández Flecha, María ..... 3503
Ferrigno, Stephen ..... 3411

## Author Index

Ferstl, Evelyn ..... 3706
Field, Andy ..... 2006
Fiery, Cushman ..... 3672
Finger, Holger ..... 3707
Fink, Clay ..... 2992
Fischer, Julia ..... 3829
Fiser, Jozsef ..... 3846
Fisher, Anna V. ..... 415
Fisher, Anna ..... $670,1242,3417,3865$
Fisher, Christopher ..... 355
Fisher, Cynthia ..... 754
Fitneva, Stanka ..... 2003
Flack, Zoe ..... 2006, 2007
Fleischhut, Nadine ..... 3708
Fliervoet, Yorick ..... 2439
Flusberg, Stephen ... 361, 1925, 3320, 3326
3884
Foley, Alana
367, 3709
Forbes, Samuel ..... 2949
Forbus, Kenneth ..... 1449
Forestier, Sébastien ..... 2013
Foulsham, Tom ..... 23, 2174
Fourtassi, Abdellah ..... 373
379, 2019 Foushee, Ruthe
2729
2729
Fraisse, Philippe
Fraisse, Philippe ..... 3733
Francis (1), Gill A. ..... 3710
Francl, Andrew ..... 1715
Frank, Michael .... 373, 495, 576, 780, 3602
Frank, Mike ..... 482
Frank, Stefan ..... 385, 1451, 3392
Frank, Till D. ..... 706
Franke, Michael ..... 507, 1678, 3070
Franke, Nicole ..... 2025
Franz, Allison ..... 2180
Fraumeni, Brittney ..... 3608
Freeman, John ..... 1453
Freiwald, Winrich ..... 1393
French, Robert M. ..... 2031
Frey, Darren ..... 391
Friedman, Alinda ..... 3729
Frost, Rebecca ..... 817
Fryt, Joanna ..... 2037
Fujimura, Souta ..... 3856
Fujioka, Takako ..... 3871
Fujisaki, Itsuki ..... 2043
Fujiwara, Masayuki ..... 3711
Fukada, Chie ..... 3712
Fukuda, Eren ..... 1193
Fukuoka, Misa ..... 2049
Furbach, Ulrich ..... 9
Furst-Holloway, Stacie ..... 1949
Fusaroli, Riccardo ..... 494, 913, 2055
Futrell, Richard ..... 319, 1022
Fux, Michal ..... 3690
G. Kamachi, Miyuki ..... 3832
Galbiati, Gianni ..... 1254
Gale, Maggie ..... 3180
Gambi, Chiara ..... 2061, 2585
Ganea, Patricia ..... 730
Gao, Min ..... 3554
Gao, Richard ..... 2067
Gao, Tao ..... 1562
Garcia-Albea, Jose E. ..... 2628
Garcia-Guerrero, Santiago ..... 3713
Garcia-Retamero, Rocio ..... 3566
Garcia, Teresa ..... 3503
Garly, Katrine ..... 2055
Garret, Paul ..... 3338
Garrett, Paul ..... 295
Gaspar, Connor ..... 3826
Gates, Monica ..... 403
Gauer, Gustavo ..... 2144
Gauvrit, Nicolas ..... 651
Gaviria, Christian ..... 2073
Gaylord, Nick ..... 2102
Gazzo Castaneda, Lupita Estefania ..... 2079
Gearhart, Jeffery ..... 355
Gelastopoulos, Alexandros ..... 949
Gelman, Andrew ..... 1272

## Author Index

Gelpi, Rebekah ..... 1967Gentner, Dedre $240,835,1072,1544,3332$,3774
Gerjets, Peter 168, 174, 3831
Gershman, Samuel J. ..... 1890Gerstenberg, Tobias .. 150, 409, 1132, 3584,3732
Gervits, Felix ..... 2085
Gerwien, Johannes ..... 3714
Getz, Heidi ..... 3715
Ghose, Ajitesh ..... 3716
Giagtzidou, Asenia ..... 3633
Gibson, Edward ..... 1022
Gilbert, Jared ..... 2413
Gilbert, Rebecca ..... 1463
Gill, Devin ..... 3717
Gilligan, Katie ..... 538, 3718
Gillis, Steven ..... 216, 433
Gimbel, Sarah ..... 2407
Girard, Sandrine ..... 3719
Giron, Josue ..... 3827
Glaescher, Jan ..... 3228
Glebkin, Vladimir ..... 2091, 2096
Gleitman, Lila ..... 1320
Gliga, Teodora ..... 3812, 3813, 3814
Gliozzi, Valentina ..... 1997
Gluck, Kevin ..... 1800
Gobet, Fernand ..... $25,1465,3039$
Goddu, Mariel ..... 3503
Godwin, Karrie E. ..... 415
Goedert, Kelly ..... 3720
Goeke, Caspar ..... 3707
Gökaydin, Dinis ..... 397
Göksun, Tilbe ..... 1471, 1477
Gold, Jason ..... 3694
Goldin-Meadow, Susan ..... 240, 1072, 3685
Goldsmith, Timothy ..... 2338
Goldstone, Rob ..... 198
Goldstone, Robert L. ..... 1206
Goldstone, Robert ..... 2675
Goldwater, Micah .. 1908, 2102, 3308, 3731,3827
Golke, Stefanie 2108, 2937 Gunzelmann, Glenn ..... 355
Golman, Russell ..... 421
Golob, Edward ..... 2114
Gómez-Veiga, Isabel ..... 2839
Gonzalez, Richard ..... 2144
Goodman, Noah D. ..... 1890
Goodman, Noah ..... 482, 1181, 3602, 3753
Goodwin, Geoffrey ..... 1842
Gopnik, Alison ..... 99, 1266, 1302, 3503
Gordon, Chelsea ..... 2120
Gosmann, Jan ..... 631, 2125
Goudbeek, Martijn ..... 2846
Graben, Katharina ..... 2131
Grabitz, Clara ..... 3273
Graesser, Arthur ..... 3766, 3868
Granlund, Sonia ..... 3699
Grant, Erin ..... 427
Gratch, Jonathan ..... 2413
Gray, Wayne ..... 25, 3167
Green, Randall ..... 3744
Greenberg, Clayton ..... 3168
Griebel, Ulrike ..... 3721
Griffiths, Thomas L. ..... 427, 859
Griffiths, Thomas ..... 403, 712, 919
Griffiths, Tom ... 111, 307, 742, 1660, 1703,1746, 2284, 2469, 2699, 3724
Grigoroglou, Myrto ..... 2132
Grimm, Robert ..... 216, 433
Grimm, Scott ..... 2500
Grimmick, Chris ..... 3722
Grinberg, Maurice ..... 2236
Gronchi, Giorgio ..... 2138, 3825
Gros, Hippolyte ..... 439
Gu, Yan ..... 445
Guath, Mona ..... 3723
Gudde, Harmen ..... 451, 1445
Guedes de Nonohay, Roberto ..... 2144
Guest, Duncan ..... 3398
Guest, Olivia ..... 1457
Gulz, Agneta ..... 1169
Gumbsch, Christian ..... 452

## Author Index

Gureckis, Todd M ..... 3824
Gureckis, Todd ........ 464, 1788, 3722, 3828
Gütschow, Jakob ..... 2634
Gvozdic, Katarina ..... 2150
Gweon, Hyowon ..... $105,162,458,570$
H. Siegel, Max ..... 1284
Haake, Magnus ..... 1169
Hachigian, Amy ..... 1163
Hack, Charles ..... 355
Hagen, Afke ..... 3627
Hager, Gregory D. ..... 246
Hagihara, Hiromichi ..... 3712
Hahn, Udo ..... 180
Hahn, Ulrike . 1084, 2476, 2663, 3454, 3460
Hai Weiss, Atalia ..... 2419
Haikka, Pinja ..... 11
Halpern, David ..... 464
Hamari, Juho ..... 11
Hamburger, Kai ..... 2131
Hamer, Aaron ..... 1955
Hamilton, Antonia ..... 1455
Hampton, James ..... 470, 1592Hamrick, JessicaHamrick, JessicaHand, Avril3725
Hansen, Steven S. ..... 3819
Harada, Etsuko ..... 3860
Harbison, Isaiah ..... 2992
Hardcastle, Valerie ..... 476
Harel, Assaf ..... 3744
Harmon, Zara ..... 2366
Harris, Adam ..... 582
Harrison, Stephen ..... 2506
Hartanto, Andree ..... 3726
Hartmann, Stephan ..... 3454
Harvey, Nigel ..... 2162, 3872
Hashimoto, Takashi ..... 3711, 3795, 3862
Hassabis, Demis ..... 2956
Hassan, Reem ..... 355
Hata, Suguru ..... 3758
Häuser, Katja ..... 2156

## Author Index

Hladká, Eva ..... 3804
Ho, Mark ..... 526
Hoareau, Violette ..... 532
Hodas, Nathan ..... 937
Hodgkiss, Alex 538, 3718
Hodossy, Lilla ..... 544
Hoeks, John ..... 3627
Hoelscher, Christoph ..... 3675
Hofer, Matthias ..... 550
Hoffmann, Ana Maria ..... 3792
Hofmann, Markus J. ..... 2025
Hohenberger, Annette ..... 3423
Hölldobler, Steffen ..... 9, 883
Hollins, Timothy ..... 2524
Holmes, Kevin ..... 556, 3320
Holzmann, Julia ..... 3088
557, 2043, 3143
Honda, Hidehito
2223
2223
Hoover, Joe
Hoover, Joe
148
148
Hopfensitz, Astrid
Hopfensitz, Astrid
2413
2413
Horowitz, Charlotte
3837
Horowitz, Jaclyn E.
2006, 2007
Horst, Jessica
2223
2223
Horton, Katie
Horton, Katie ..... 3732
Hough, Alexander ..... 718
Hough, Julian ..... 2230
Houpt, Joseph ..... 3733
Howard, Marc ..... 1187
Howard, Zachary ..... 3338
Howe, Rita ..... 2224
Howes, Andrew ..... 51
Howes, Christine ..... 2230
Hoyos, Christian ..... 2949Hristova, Evgeniya2236, 3747
Hristova, Penka ..... 967, 3191
Hsiao, Janet H. ..... 1430, 2561
Hsiao, Janet ..... 3386
Hsiao, Yaling ..... 563
Hsieh, Shulan ..... 3765
Hsu, Chun-Ting ..... 2815
Hsu, Dong-Bo ..... 3734
Hsu, Shaw ..... 2512
Hu, Jane C ..... 2894
Hu, Jon-Fan . 3735, 3736, 3765, 3768, 3877,3878
Huang, Kuan-Jung ..... 3736
Huang, Po-Sheng ..... 3735
Huang, Shanshan ..... 853
Huang, Xin ..... 3737
Huang, Xunan ..... 3838
Huang, Yu-Shu ..... 3686
Huang, Yujing ..... 2242
Huettig, Falk ..... 730
Huisman, Carlijn ..... 3627
Hunt, Amelia ..... 2279
Husney, Sarah ..... 556
Hutton, Elizabeth A. ..... 2243
Ibarra, Alyssa ..... 2249
Icard, Thomas ..... 2255
Igami, Tsuyoshi ..... 2669
Iiona, Scully ..... 3738
Ijuin, Koki ..... 3864
Inomata, Kentaro ..... 3798
Inuzuka, Miwa ..... 1556
Ireland, Molly ..... 2102
Isbilen, Erin S. ..... 564, 2681
Isbilir, Erdinc ..... 3739
Itsuki, Chiba ..... 3853
Izquierdo, Eduardo J. ..... 3119
Jachmann, Torsten ..... 2261
Jacobson, Michael ..... 3731
Jacoby, Hilla ..... 2419
Jacoby, Nori ..... 712
Jaeger, T. Florian ..... 186, 2616
Jaeggi, Susanne ..... 2747
Jaffe-Dax, Sagi ..... 2267, 3740
Jain, Saransh ..... 2273
James, Warren ..... 2279
Janner, Michael ..... 1393
Jansen, Rachel ..... 2019, 2284
Janyan, Armina ..... 3639

## Author Index

Jara-Ettinger, Julian ..... 570, 1284
Jarecki, Jana ..... 11
Jarecki, Julian ..... 11
Jasbi, Masoud ..... 576
Jaspaert, Koen ..... 3435
Jasper, Robert ..... 3677
Jastrzebski, Jan ..... 2290
Jay, Kristin ..... 3762
Jayaraman, Swapnaa ..... 3694
Jenkins, Gavin ..... 588
Jenkins, Sarah ..... 582
Jennings, Jay ..... 2296
Jew, Carol ..... 1752
Ji Ma, Wei ..... 1254
Ji, Yue ..... 594, 3741
Jiang, Chenfanfu ..... 700
Jimenez, Eva ..... 2302
Jódar Sánchez, José Antonio ..... 343
Johannes, Kristen ..... 600, 2308
Johannsen, Niels Nørkjær ..... 494
Johanson, Megan ..... 2132
Johnson-Laird, Phil ..... 980
Johnson-Laird, Philip ..... 663, 1690
Johnson, Anthony ..... 2992
Johnson, Elizabeth ..... 3742
Johnson, Kate M. ..... 2320
Johnson, Kate ..... 1648
Johnson, Mark H. ..... 3812, 3813, 3814
Johnson, Samuel ..... 606, 612, 618, 2314
Johnson, Scott ..... 3089
Johnston, Angie ..... 612
Johnston, Benjamin ..... 3447
Jokinen, Jussi ..... 624
Jones, Emily ..... 3813
Jones, Fergal ..... 2327
Jones, Gary ..... 3398
Jones, Jonathan D. ..... 246
Jones, Luke ..... 3534
Jones, Matt ..... 1830
Jones, Matthew ..... 2321
,
Jones, Michael ..... 313, 325, 1973
Jones, Peter M. ..... 3640
Jordan, Fiona ..... 21
Joseph, Ebenezer ..... 2333, 3743
Juanchihch, Marie ..... 2610
Jung, Kyung Hun ..... 2338
Juslin, Peter ..... 3723
Juvina, Ion ..... 718, 3744
Kachergis, George 57, 519, 1501, 2205, 3722
Kaesler, Matthew ..... 625
Kahl, Sebastian ..... 2350
Kajic, Ivana ..... 631
Kakouros, Sofoklis ..... 998
Kaldy, Zsuzsa ..... 3841
Kalish, Charles ..... 31, 3745
Kalkstein, David ..... 149
Kallen, Rachel W. ..... 1949, 2784
Kallen, Rachel ..... 871, 1321, 2506
Kameda, Tatsuya ..... 3793
Kamermans, Kevin ..... 2925
Kaminski, Jennifer ..... 2356, 3197
Kämmer, Juliane E. ..... 3834
Kampa, Alyssa ..... 2362
Kane, Gary ..... 637
Kanwal, Jasmeen ..... 643
Kao, Yvonne ..... 600
Kapatsinski, Vsevolod ..... 2366
Kaplan, Jonas ..... 2407, 2413
Kapur, Manu ..... 649
Karadöller, Dilay Z ..... 2372
Karlinsky, April ..... 2377
Kary, Arthur ..... 847
Kashiwabata, Tatsuya ..... 3795
Kato, Tsuneo ..... 3864
Katsos, Napoleon ..... 724
Kätsyri, Jari ..... 2344
Katz, Benjamin ..... 2747
Katz, Irvin ..... 3788
Katz, Lynne ..... 3811
Kawasaki, Keigo ..... 3798

## Author Index

Keehner, Madeleine ..... 3788
Knauff, Markus ..... 2079, 2131
Keil, Frank 606, 612, 618, 3368
Keilmann, Franziska ..... 650
Kelemen, Deborah ..... 33
Kellen, David ..... 1084
Kello, Chris ..... 39
Kemp, Charles ..... 2383
Kempe, Vera ..... 651
Kenett, Yoed ..... 657, 1035
Kennedy, Lauren ..... 3746
Kent, Chris ..... 3410
Keric, Vjeran ..... 3747
Kersten, Luke ..... 2389, 2395
Keshava, Ashima ..... 1296
Khemlani, Sangeet ... 663, 1611, 1690, 3750Khemlani, Sunny3362
Khudanpur, Sanjeev ..... 246
Kida, Noriyuki ..... 3712
Kidd, Celeste ..... 3411
Kidd, Evan ..... 564
Kiesel, Andrea ..... 3830, 3857
Kießner, Ann-Kathrin ..... 2401
Kim, Cora ..... 1060
Kim, Dan ..... 669, 2943
Kim, Eunkyung ..... 2407, 2413
Kim, Jaeah ..... 670
Kim, Jangjin ..... 1453
Kimel, Eva ..... 2419
Kingstone, Alan ..... 1296
Kirby, Simon .. 643, 1023, 2760, 3095, 3796,3875
Kirfel, Lara ..... 2420
Kirschhock, Maximilian E. ..... 525
Kitamura, Miho ..... 2426
Kılıç, Özkan ..... 1495
Kırkıcı, Bilal ..... 1495
Klauer, Karl Christoph ..... 1084
Kleiman-Weiner, Max ..... 13, 676, 2536
Kleiman, Tali ..... 149
Klenske, Edgar ..... 1047
Kljajevic, Vanja 2432, 3748
Knight, Robert ..... 3742
Knoeferle, Pia ..... 1483, 2463
Knutsen, Sten ..... 3821
Kobayashi, Harumi ..... 3291
Koch, Lennart ..... 3857
Kochenderfer, Mykel J. ..... 2488
Kodner, Jordan ..... 2438, 3683
Koedinger, Ken ..... 204
Koehler, Derek ..... 3826
Kojima, Takatsugu ..... 3712, 3749
Kokkonen, Tommi ..... 2445
Kola, Ilir ..... 980
Kolak, Joanna ..... 3699
Komer, Brent ..... 631
Kominsky, Jonathan ..... 931
Komoto, Natsumi ..... 3758
Kon, Elizabeth ..... 682
König, Peter ..... 1296
Konno, Takeshi ..... 3711, 3795
Konovalova, Elizaveta ..... 688
Konstantinidis, Emmanouil . 694, 736, 3872
Koolen, Ruud ..... 2439
Koponen, Ismo ..... 2445
Kopp, Stefan ..... 2350
Korman, Joanna ..... 2451, 3750
Kornkasem, Sorachai ..... 1163
Koshima, Yu ..... 3854
Koslowski, Barbara ..... 3751
kosoy, Eliza ..... 3752
Koster, Ernst H. W. ..... 3022
Kovack-Lesh, Kristine ..... 3671
Kovacs, Marton ..... 3806
Koven, Marissa ..... 612
Kovic, Vanja ..... 1604
Kovtunenko, Alexander ..... 2091
Krahmer, Emiel ..... 1866, 2846
Kranjec, Alexander ..... 1290, 1477
Kray, Jutta ..... 2156
Krefeld-Schwalb, Antonia ..... 2457

## Author Index

Kreiss, Elisa ..... 3753
Langsford, Steven ..... 2530
Krems, Josef F. ..... 3540
Kroczek, Bartlomiej ..... 3688
Kroeger, Julia Marina ..... 2463
Kroupin, Ivan ..... 3754
Krueger, Paul ..... 742, 2469
Krusmark, Michael ..... 1800
Krysova, Ekaterina ..... 2091
Kryven, Marta ..... 2475
Krzyzanowska, Karolina ..... 2476
Kubota, Shinichi ..... 3795
Kubricht, James ..... 700, 2579
Kucker, Sarah ..... 2482, 3755
Kuefler, Alex ..... 2488
Kuhnert, Barbara ..... 2494, 3478
Kumaran, Dharshan ..... 1319
Kummetha, Sneha ..... 3756
Kunda, Maithilee ..... 1519, 2754
Kuntay, Aylin ..... 3273
Kuperberg, Gina ..... 283, 283
Kurth-Nelson, Zeb ..... 1319
Kurtz, Kenneth ..... 907, 3215,3757
Kurumada, Chigusa ..... 2500
Kutas, Marta ..... 1212
Kwisthout, Johan ..... 2907
Kwon, Jinhwan ..... 3758
La Torre, Amelia ..... 2037
Lachmair, Martin ..... 3831
Lagnado, DavidLahiri, Aditi1537
Laing, Catherine ..... 711
Lake, Brenden ..... 3752, 3828
Lam, Melanie ..... 2377
Lamb, Maurice ..... 1321, 2506, 2784
Lampinen, Andrew ..... 2512
Landau, Barbara ..... 246, 2308, 3235
Landy, David ..... 907, 2518, 2675, 3016
Lange, Nicholas ..... 2524
Langley, Pat ..... 29
Langlois, Thomas 712,3724
Lanning, Guilherme ..... 2144
Larigaldie, Nathanael ..... 1381
Lark, Murray ..... 582
Larson, Calais ..... 1550
Larue, Othalia ..... 718, 3744
Lau, Ellen ..... 283
Lau, Esther E.y. ..... 1430
Lavelle, Mary ..... 2230
Lavelle, Suilin ..... 2821
Lazaridou-Chatzigoga, Dimitra ..... 724
Le Guen, Olivier ..... 21
Le Mens, Gael ..... 688
Le Pelley, Mike ..... 3484
Leake, David ..... 2717
Leather, Calvin ..... 3490
Lee, Ruth ..... 730
Lee, Tzu-Hsi ..... 3880
Lee, Yuna ..... 2536
Leffers, Jessica ..... 3690
Lei, Yuqing ..... 3228
Leibo, Joel ..... 1319
Lemaire, Benoit ..... 532
Lengyel, Mate ..... 3846
Lenow, Jennifer ..... 3760
León-Villagrá, Pablo ..... 3759
Leonard, Julia ..... 2536
Leroy, Anaïs ..... 3761
Leroy, Caroline ..... 525
Leslie, David ..... 767
Lester, Nicholas ..... 2537
Leuker, Christina ..... 2543
Levering, Kimery ..... 3762
Levine, Susan ..... 3608, 3884
Levy, Joseph ..... 2549
Levy, Roger ..... 550
Lew-Williams, Casey ..... 3101
Lewandowsky, Stephan ..... 532, 767
Lewis, Richard ..... 51
Li, Ambrose ..... 3691, 3874

## Author Index

Li, Defeng ..... 3737
Livins, Katherine ..... 15
Li, Fuyin ..... 343
Li, Guanhong ..... 3711
Li, Luhe ..... 2561
Li, Qi ..... 2790
Li, Tze Kwan ..... 2561
Li, Xiaoqian ..... 1400, 2555
Li, Ying ..... 3763, 3776
Li, Yunqi ..... 1254
Liang, Garston ..... 736
Lichtenstein, Patricia ..... 2567, 3844
Lieder, Falk$742,1660,2469$
2419, 3764Lieder, Itay ............................ 2419, 3764
Lieven, Elena ..... 1691, 3699
Likens, Aaron ..... 2573Liljeholm, Mimi1758, 3261
Lin, Deborah ..... 748
Lin, Jenny ..... 2579
Lin, Yi ..... 754
Lin, Yu-Ching ..... 3765
Lindner, Felix ..... 2494
Lindsay, Laura ..... 2585
Linjie, Li ..... 1096
Linn, Marcia ..... 3522
Lins, Jonas ..... 2586
Linzen, Tal ..... 2592
lippert, Anne ..... 3766, 3868
Liquin, Emily ..... 2598
Little, DanielLittle, Hannah760
Littman, Michael ..... 526
Litvinova, Aleksandra ..... 2407
Liu, Dawn ..... 2610
Liu, Hsi-wen ..... 2622
Liu, Huihui ..... 192
Liu, Linda ..... 2616
Liu, Pei Q ..... 2604
Liu, Shari ..... 766
Liu, Xun ..... 2790
Livesey, Evan ..... 3827
Lloyd-Fox, Sarah ..... 3814
Lloyd, Kevin ..... 767
Lobina, David J. ..... 2628
Lobinger, Babett ..... 823
Lobmüller, Nadine ..... 3829
Lobue, Vanessa ..... 1574
Lockhart, Kristi ..... 3368
Logan, John ..... 3842
Loh, Vanessa ..... 1598
Lohmann, Johannes ..... 2634, 3676
Lohse, Keith ..... 2377
Loiseau, Virginie ..... 3771
Loiselle, Tayler ..... 1610
Lomakina, Ekaterina ..... 3807
Lombard, Marlize ..... 494
Lombrozo, Tania ..... 682, 1266, 2598
Lopez-Brau, Michael ..... 1181
López-Felip, Maurici A. ..... 706
López, Cristina ..... 2432
Lorenz, Tamara ..... 2506
Lositsky, Olga ..... 3767
Lou-Magnuson, Matt ..... 773
Lourenco, Stella ..... 75
Louwerse, Max M. ..... 1727
Louwerse, Max ..... 39, 1629
Love, Bradley ..... 1457
Lovett, Andrew ..... 2640
Low, Hui Min ..... 1739
Lowrie, Tom ..... 27
Lu, Hongjing ..... 700, 1066, 2579
Luby, Erin ..... 556
Lucas, Christopher ..... 501, 3759
Ludvig, Elliot ..... 3658
Luo, Yu-Han ..... 3768
Luo, Yu ..... 774
Lupyan, Gary ..... 331, 3851
Luthra, Sahil ..... 3769
Lyngs, Ulrik ..... 2646
Lynott, Dermot 1444, 2604, 3770

## Author Index

Ma, Lizhi ..... 3404
MacDonald, Dan ..... 3874
MacDonald, Kyle ..... 780
Mackiewicz, Robert ..... 1690
Madruga, Juan Antonio Garcia ..... 2839
Madsen, Jens Koed . 2651, 2657, 2663, 2902Maehigashi, Akihiro .............. 2049, 2669
Magid, Rachel ..... 786, 792
Magnuson, James ..... 3769
Mahal, Kuldeep K. ..... 1078
Maïonchi-Pino, Norbert ..... 3771, 3861
Majid, Asifa ..... 258, 3227
Majima, Yoshimasa ..... 3772
Mak, Joanna Chor Yan ..... 3125
Malinovitch, Tamar ..... 3773
Malle, Bertram ..... 1035, 1848, 2451
Mallot, Hanspeter ..... 3088
Malt, Barbara ..... 986, 3873
Malykhina, Katsiaryna ..... 1471
Manoharan, Sundar ..... 2333
Maravilla, Francisco ..... 3774
Marberg, Ines ..... 3714
Marchman, Virginia ..... 780
Marcus, Gary ..... 3089
Mareschal, Denis ..... 130, 2031
Marewski, Julian ..... 877, 1931
Marghetis, Tyler ..... 865, 2193, 2675
Margni, Adrián ..... 2735
Marić, Mateja ..... 3696
Markant, Doug ..... 1501
Markman, Ellen ..... 1333, 3817
Marno, Hanna ..... 3775
Marshall, Peter ..... 2711
Martin, Andrea ..... 1955
Martin, Jay ..... 15
Masitah, Masitah ..... 3776
Masnick, Amy ..... 3777
Masoudnia, Saeed ..... 3778
mathy, fabien ..... 3779
Matlock, Teenie ..... 361, 3789
Matsui, Tatsunori ..... 1985
Matsuka, Toshihiko ..... 557, 3143
Matsushita, Soyogu ..... 3794
Matsuzaki, Takuya ..... 1556
Matthews, Percival ..... 2705, 3745
Mayor, Eric ..... 1641
Mayor, Julien ..... 1739, 3161
Mayrhofer, Ralf ..... 150, 3780
McCabe, Rose ..... 2230
McCauley, Stewart M. ..... 564, 2681
McClelland, James L. ..... 2488, 2512, 3819
McCormack, Teresa ..... 35, 3781
McCormick, Samantha ..... 2549
McCurry, Malcolm ..... 3362, 3664
McElree, Brian ..... 3695
McEnery, Tony ..... 3770
McFate, Clifton ..... 2949
McGinn, Kelly ..... 2687, 3882
McKay, Ryan ..... 2646, 3801
McKenna, Lucy ..... 1836
McLaren, Ipl ..... 1782, 2327, 3546
McLaren, Rossy ..... 2327, 3546
McLaughlin, Elizabeth ..... 204
McLean, Dave ..... 3826
McMahan, Brian ..... 2876
McNamara, Danielle ..... 1884, 2573
McNeil, Nicole ..... 600
McRae, Ken ..... 337
Medeiros, Joyse ..... 2693
Meder, Björn ..... 1357, 1459, 3834
Meilinger, Tobias 525, 650, 1156, 3702
Meng, Rui ..... 2705, 3745
Meng, Yuan ..... 2699
Mercier, Hugo ..... 2710
Meredith Weiss, Staci ..... 2711
Merritt, Haily ..... 3851
Messinger, Daniel ..... 3811
Metcalf, Katherine ..... 2717
Meylan, Stephan C ..... 859
Michalland, Arthur-Henri ..... 2729, 2729

## Author Index

Mickes, Laura ..... 3790
Micklos, Ashley ..... 3782
Miikkulainen, Risto ..... 1507
Milin, Petar ..... 319
Miłkowski, Marcin ..... 2723
Miller, Brent ..... 793
Mills, Gregory ..... 3627, 3783
Milton, Fraser ..... 1979
Minai, Ali ..... 2506
Minervino, Ricardo A. ..... 2735, 3296
Minnery, Brandon ..... 3744
Mintz, Toby ..... 3490
Mistry, Percy ..... 2741
Misyak, Jennifer ..... 3784
Miton, Helena ..... 2710
Miura, Toshihiko ..... 3795
Miwa, Kazuhisa ..... 2049, 2669, 3856
Miyashiro, Kozue ..... 3860
Mizraji, Eduardo ..... 3785
Mohammed, Shafee ..... 2747
Mohler, Betty J ..... 650
Mollica, Francis ..... 799, 805
Molouki, Sarah ..... 2748
Monaghan, Padraic ... 811, 817, 1307, 1740
35
Monier, Florie
3786
Montag, Jessica
3787
Montefinese, Maria
2754
Monteiro Eliott, Fernanda
3788
Moon, Jung Aa
3789
Moorman, Karie ..... 3790
Morgan, David
283
Morgan, Emily .....
2669 .....
2669
Mori, Kensaku
3794
Morikawa, Kazunori
3711
Morita, Junya
3344, 3777
Morris, Bradley
3, 2537
3, 2537
Moscoso del Prado
Motamedi, Yasamin ..... 2760
Moty, Kelsey ..... 3791
Mudrick, Nicholas V. ..... 2766
Muenster, Katja ..... 2463
Muentener, Paul ..... 3553
Muhlstein, Larry ..... 2772
Munakata, Yuko ..... 1937
Munnich, Edward ..... 3792
Munos, Remi ..... 1319
Murata, Aiko ..... 3793
Musculus, Lisa ..... 823
Musslick, Sebastian ..... 829, 895
Muto, Hiroyuki ..... 3794
Myers, Christopher ..... 51, 355
Myers, Matthew ..... 835
Nadjar Araabi, Babak ..... 3778
Nador, Jeffrey ..... 3744
Nagata, Noriko ..... 3798
Nagataki, Shoji ..... 3795
Naigles, Letitia ..... 913
Nakamura, Yoshihiko ..... 2669
Nakano, Yoshiki ..... 2778
Nakayama, Gotaro ..... 3856
Nalepka, Patrick ..... 2784
Nambu, Satoshi ..... 228
Nan, Weizhi ..... 2790
Nasraoui, Olfa ..... 1375
Nataraja, N.p. ..... 2273
Nation, Kate ..... 563
Navarrete, Jairo A. ..... 841
Navarro, Daniel ..... 3746
Navarro, Victor ..... 1457
Neemeh, Zachariah A. ..... 2796
Nelson, Jonathan D. ..... 1357, 1459, 3834
Nematzadeh, Aida ..... 427, 853, 859
Neufeld, Chris ..... 2802
Newell, Ben R. ..... 694, 736
Newport, Elissa L ..... 3837
Newport, Elissa ..... 3715
Nicenboim, Bruno ..... 1278
Nicholson, Bethany ..... 3866
Nie, Aiming ..... 3839
Nie, Yunlong ..... 3863

## Author Index

Niedopytalski, Scott ..... 1925
Nielsen, Alan Ks ..... 3796
Niemi, Laura ..... 3797
Nikata, Kunio ..... 3798
Nili Ahmadabadi, Majid ..... 3778
Nishimi, Andrea ..... 865
Niv, Yael ..... 2217
Noelle, David ..... 3799
Norbury, Courtenay ..... 3441
Nordbeck, Patric ..... 871
Norris, Catherine ..... 2803
Novack, Miriam ..... 3685
Noviski, Sarah ..... 2986
Nozari, Nazbanou ..... 1218
Nulty, Paul ..... 2963
Nunez, Rafael ..... 1004
Nussbaumer, Daniela ..... 2809
O'Connell, Garret ..... 2815
O'Grady, Cathleen ..... 2821
O'Hora, Denis ..... 1424, 3713, 3725, 3889
O'Leary, Allison ..... 3800
O'Lone, Katherine ..... 3801
O'Rourke, Polly ..... 2827
Oaksford, Mike ..... 259
Oberauer, Klaus ..... 532
Obhi, Sukhvinder ..... 1782
Oda, Masahiro ..... 2669
Oei, Adam C. ..... 1400
Oghalai, John ..... 3816
Ohira, Hideki ..... 3795
Ohmori, Maito ..... 3860
Ohtsu, Kayoko ..... 3802Ojha, Suman2833
Okada, Ryo ..... 2168
Okamoto, Masahiko ..... 3854
Okuda, Jiro ..... 3711
Olds, Justin ..... 877
Olguín, Valeria ..... 3296
Oliveira da Costa, Ana ..... 883
Oller, D. Kimbrough ..... 3721
Omaki, Akira ..... 228
Onnis, Luca ..... 773
Ono, Hajime ..... 228
Opfer, John ..... 669, 2943
Orenes, Isabel ..... 2839
Orr, Ryan W. ..... 631
Ortega, Gerardo ..... 889
Ortega, Nancy ..... 3792
Osherson, Daniel ..... 3609
Ostarek, Markus ..... 961
Osth, Adam ..... 3590
Otsubo, Kay ..... 2840
Otte, Sebastian ..... 452, 3803
Oudeyer, Pierre-Yves ..... 2013
Out, Charlotte ..... 2846
Over, David ..... 259, 3528
Overmann, Karenleigh ..... 21
Ozaki, Koken ..... 1556
Ozcimder, Kayhan ..... 829, 895
Ozzer, Demet ..... 1471, 1477
Özer, Ege Ekin ..... 1471
Özkan, Ayșegül ..... 1495
Ozyurek, Asli ..... 889, 2372
Ozyürek, Aslı ..... 81
Paas, Fred ..... 2925
Pachur, Thorsten ..... 2543
Paes, Tanya ..... 3805
Palfi, Bence ..... 3806
Paliwal, Saee ..... 3807
Palokangas, Lauri ..... 2852
Pandza, Nick ..... 2827
Pankka, Hanna ..... 925
Papafragou, Anna 594, 2132, 2362, 3 ..... 3741
Pařilová, Tereza ..... 3804
Parker, Jessica ..... 2858
Parlett, Chelsea ..... 2747
Parr (3), Teresa ..... 3710
Parr, Teresa ..... 3809
Pastorino Campos, Carla ..... 2864
Pate, John ..... 901, 1157

## Author Index

Patel, Jwalin ..... 3809
Pipergias Analytis, Pantelis ..... 949Patel, Purav .................. 1610, 3808, 3818
Patterson, John ..... 907
Paul, Celeste ..... 2992
Pavlova, Margarita ..... 2870
Pearl, Lisa ..... 3064
Pederson, Kaitlin ..... 1794
Pei, Kevin ..... 2876
Peng, Shu-Ling ..... $3735,3765,3877$
Peng, Yujia ..... 1066
Penner-Wilger, Marcie ..... 1150, 3810
Peres, Filipe ..... 3209
Perez Baez, Gabriela ..... 3332
Perez, Ray ..... 25
Perfors, Amy ..... 847, 992, 2530, 3746
Perlman, Marcus ..... 331, 760, 913
Perone, Sammy ..... 3174
Perry, Lynn ..... 1461, 3755, 3811
Persaud, Kimele ..... 1513,2876
Peterson, Joshua ..... $111,919,1746$
Petri, Giovanni ..... 895
Petzschner, Frederike ..... 3807
Pfeifer, Niki ..... 925, 2882, 2888
Pham Nguyen, Thuan ..... 3863
Pham, Theresa ..... 2894
Phelps, Elizabeth ..... 3760
Phillips, Jonathan ..... 931
Phillips, Lawrence ..... 937
Phillips, StevenPhillips, Webb1060
Piantadosi, Steve ..... 222, 799, 805
Piccardi, Elena Serena ..... 3812
Pickering, Martin ..... 3850
Pierguidi, Lapo ..... 3825
Pierrehumbert, Janet ..... 1059
Piiroinen, Petri ..... 1424, 3725
Pilditch, Toby D. ..... 943, 2651, 2657
Pilucik, Dariusz Lukasz ..... 2902
Pine, Julian ..... 3699
Pink-Hashkes, Sarit ..... 2907
Pirazzoli, Laura ..... 3813, 3814
Pirrone, Angelo ..... 955
Plancher, Gaën ..... 532
Pleskac, Timothy J. ..... 2543
Pleskac, Timothy ..... 265
Plunkett, Kim ..... 367, 3267, 3709, 3843
Poinan, Molly ..... 3762
Poletiek, Fenna ..... 3815
Pollock, Eli ..... 3732
Pollonini, Luca ..... 3816
Pomi, Andrés ..... 3785
Ponari, Marta ..... 3441
Popov, Vencislav ..... 961, 967, 968, 2913
Portrat, Sophie ..... 532
Posid, Tasha ..... 2919
Posner, Elyanah ..... 3503
Pothos, Emmanuel ..... 93, 1592, 3572
Pouw, Wim ..... 2925
Powell, Derek ..... 3817
Powers, Robert ..... 3865
Pozdnyakova, Elena ..... 3887
Prasad, Vishaal ..... 2931
Prince, Emily ..... 3811
Prinz, Anja ..... 2937
Psaromiligkos, Ioannis N. ..... 3046
Puebla-Ramirez, Guillermo ..... 1955
Purver, Matthew ..... 1818
Qin, Jike ..... 2943
Qinyu, Cheng ..... 3868
Quillien, Jean-Baptiste ..... 1610, 3818
Raab, Markus ..... 823
Rabagliati, Hugh ..... 2585, 3241
Rabkina, Irina ..... 2949
Rabovsky, Milena ..... 3819
Rácz, Peter ..... 21
Radach, Ralph R. ..... 2025
Radulescu, Angela ..... 2217
Rafferty, Anna ..... 2284, 3522
Raftopoulos, Athanassios ..... 974

## Author Index

Ragni, Marco 9, 883, 980, 2401, 2494, 3380, Reynolds, Katherine ..... 29863478, 3510
Rahona, Juan José ..... 3831
Raizada, Rajeev ..... 1752
Raja, Vicente ..... 2955
Rajkovic, Anastazija ..... 3736
Rakow, Tim ..... 2174
Ralph, Schumacher ..... 2809
Ramarajan, Dhaya ..... 99
Ramirez, Juan Pablo ..... 3822
Ramiro, Christian ..... 986
Ramscar, Michael ..... 313, 319, 1973
Ramzaoui, Hanane ..... 3820
Randall, Ashley ..... 3697
Rañó, Iñaki ..... 3889
Ransom, Keith ..... 992
Raposo, David ..... 2956
Rapp, Brenda ..... 1339
Rasanen, Okko ..... 998
Rastle, Kathleen ..... 1740
Raudenbush, Stephen ..... 3608
Raynal, Lucas ..... 2957
Rebuschat, Patrick ..... 1307
Recchia, Gabriel ..... 2963
Redeker, Gisela ..... 3783
Reder, Lynne ..... 968, 2913
Redshaw, Jonathan ..... 35
Rees, Alice ..... 2969
Reeves, Stuart ..... 25
Rehder, Bob ..... 1896
Rehrig, Gwendolyn ..... 3821
Reichman, Daniel ..... 1660
Reijula, Samuli ..... 2975
Reimer, Torsten ..... 3674, 3822
Reimers, Stian ..... 1824, 2981
Relaford-Doyle, Josephine ..... 1004
Ren, Jie ..... 1010
Retren, Daemon ..... 3691
Revithis, Spyridon ..... 3823
Rey, Günter Daniel ..... 3540
Reyna, Valerie ..... 2180
Rhodes, Rebecca ..... 2992
Rich, Alexander S ..... 3824
Richardson, Daniel ..... 544
Richardson, Michael J. ..... 1949, 2784
Richardson, Michael ..... 871, 1321, 2506
Richland, Lindsey ..... 1072
Riede, Felix ..... 494
Rieskamp, Jörg ..... 2457
Righi, Stefania ..... 3825
Risko, Evan ..... 3826, 3891
Rissman, Lilia ..... 3685
Ritter, Frank ..... 1872
Rivero-Fernández, Camila ..... 3811
Robbins, Talia ..... 2998
Roberts, Gareth ..... 3004, 3827
Roberts, Sean ..... 1090
Robinson, Chris ..... 2858
Roche, Jennifer ..... 2986
Rodd, Jennifer M. ..... 1647
Rodd, Jennifer ..... 1463
Roessler, Lotte Sophia ..... 3831
Roettger, Timo B. ..... 3010
Rogers, Brad ..... 3016
Rogers, Timothy ..... 31
Rogers, Wendy ..... 1770
Rojo, Sergio Gonzalez de La Higuera ..... 494
Roland, Chris ..... 3674, 3822
Rolison, Jonathan J. ..... 1205
Rölke, André ..... 2025
Romigh, Griffin D. ..... 1016
Rong, Wenge ..... 3554
Rosenbaum, David ..... 168
Rothe-Wulf, Annelie ..... 1617, 3829
Rothe, Anselm ..... 3828
Rothwell, Clayton D. ..... 1016
Ruess, Miriam ..... 3830, 3857
Ruggeri, Azzurra ..... 823, 1788
Ruiz Fernández, Susana ..... 3831
Runner, Jeffery ..... 2249

## Author Index

Rusch, Tessa .............................. 3228
Ruthruff, Eric .............................. 2338
Rutten, Isa ................................. 3022
Ryder, Robin .............................. 1278
Ryskin, Rachel ............................ 1022
Saeed, Basil ................................ 3584
Safronov, Nikita ............................ 2096
Sagi, Eyal .................................. 3028
Sahani, Maneesh .......................... 3764
Saito, Motoyuki ............................ 3033
Saito, Yuichi ............................... 3832
Sakamoto, Maki ........................... 3758
Sakawaki, Takako .......................... 3802
Sala, Giovanni ..................... 1465, 3039
Salay, Nancy ............................. 3045
Saldana, Carmen .......................... 1023
Salehi Nobandegani, Ardavan ... 1029, 3046
Saltzman, Elliot .......................... 2784
Samejima, Kazuyuki ...................... 3711
Sanborn, Adam ............. 767, 1102, 3285
Sander, Emmanuel ......... 439, 2150, 2957
Sandhofer, Catherine .................... 1758
Sangster, Matthew ......................... 25
Sankar, Remya ............................. 3052
Sarathy, Vasanth ......................... 1035
Satkunarajah, Praveena ................ 2199
Sato, Toru .................................. 3798
Saulton, Aurelie ........................... 3702
Savic, Andrej ............................... 1604
Savic, Olivera .............................. 3058
Savinelli, K.j. ............................. 3064
Saxe, Andrew ................................. 829
Scatturin, Lara ............................. 174
Schatz, Thomas .......................... 3833
Scheibehenne, Benjamin ................. 2457
Scheibling-Sève, Calliste ................. 3076
Schepens, Job ............................. 3082
Scherbaum, Stefan ........................ 1041
Scheuerman, Jaelle ......................... . 2114
Scheutz, Matthias ............... 1035, 2085

Schick, Wiebke ............................ 3088
Schiefner, Annika ........................... 889
Schinazi, Victor R. ......................... 3675
Schirren, Clara ............................ 3834

Schmitt, Theresa ........................... 3803
Schneegans, Sebastian ........................ 5
Schoemann, Martin ...................... 1041
Schoener, Gregor .......................... 2586
Schoetensack, Christine .................. 1307
Schöller, Anthea ............................. 3070
Schon, Claudia ................................... 9
Schonberg, Christina ...................... 3089
Schooler, Lael ............................... 1931
Schoppek, Wolfgang .................... 3835
Schouwstra, Marieke ............ 2760, 3095
Schrodt, Fabian .......................... 3836
Schrum, Nicolaus ....................... 3821
Schuler, Kathryn D. ...................... 3837
Schuler, William .......................... 1260
Schultheis, Holger ........................ 1445
Schulz, Eric ......... 1047, 1357, 1890, 3848
Schulz, Laura ... 786, 792, 1363, 2536, 3553
Schulze, Christin ......................... 1053
Schumacher, Petra ....................... 3695
Schumacher, R. Alexander ............... 1059
Schunn, Christian ......................... 1467
Schwab, Jessica F. ........................ 3101
Schwan, Stephan ............................ 650
Scontras, Gregory ......................... 3064
Scott-Phillips, Thom ...................... 2821
Scott, Rose ..................................... 3203
Sederberg, Per ............................... 295
Sekicki, Mirjana ........................... 3107
Sekine, Kazuki ............................... 3113
Semenuks, Arturs ......................... 1060
Semmler, Carolyn .......................... 625
Seng, Ayon ..................................... 3745
Serpell (2), Zewelanji .................... 3710
Serpell, Zewelanji ......................... 3809

## Author Index

Servajean, Philippe ..... 1684
Setzler, Matthew ..... 3119
Shafto, Patrick .. 99, 1375, 1406, 1574, 3560
Shalin, Valerie L. ..... 1016
Shankar, Karthik ..... 1187
Shao, Jing ..... 3125,3838
Sharma, Sugandha ..... 3131
Shaver, John ..... 17
Shaw, Alex ..... 676
Sheffield, David ..... 3180
Shelton, Amy L. ..... 246
Shenhav, Amitai ..... 637
Shepard, Jason ..... 3673, 3839
Sherson, Jacob ..... 11
Shi, Genghu 3766, 3868
Shibata, Masayoshi ..... 3795
Shiffrin, Richard ..... 325
Shih, Stephanie ..... 3844
Shiloh, Daniel ..... 3137
Shirasuna, Masaru ..... 3143
Shiratori, Takaaki ..... 1562
Shockley, Kevin ..... 1321
Shtulman, Andrew ..... 33, 3149
Shu, Tianmin ..... 1066
Shukla, Anuj ..... 3052, 3756
Shultz, Thomas R. ..... 1029
Shutova, Ekaterina ..... 2567
Shvartsman, Michael ..... 3767
Si, Mei ..... 3155
Sia, Ming Yean ..... 3161
Sibert, Catherine ..... 3167
Siegel, Max ..... 792
Siegelman, Noam ..... 2592
Sierra, Basilio ..... 3892
Sifa, Rafet ..... 3249
Sikos, Les ..... 3168
Silliman, Daniel ..... 3757
Silvey, Catriona ..... 1072
Sim, Zi L. ..... 1078
Simmering, Vanessa ..... $5,3174,3671$
Simner, Julia ..... 3796
Simonovic, Boban ..... 3180
Sims, Chris ..... 2217
Sinnett, Scott ..... 1296
Sjåfjell, Kristin ..... 1617
Skach, Sophie ..... 3186
Skovgaard-Olsen, Niels ..... 1084
Slavov, Slavi ..... 3191
Sloman, Steven ..... 1580
Slomka, Piotr ..... 1457
Slonimska, Anita ..... 1090
Slors, Marc ..... 45
Sloutsky, Vladimir . 1453, 1469, 2919, 3058,3590, 3678, 3693, 3800
Smalley, Evans ..... 3197
Smith, Hayley K ..... 3841
Smith, Kenny . 643, 1023, 2760, 2821, 3703,3796, 3840, 3875
Smith, Kevin ..... 409, 3209, 3670
Smith, Linda B. ..... 3698
Smith, Linda ... 37, 1236, 1412, 1489, 3689, 3694, 3786, 3855
Smith, Megan ..... 3203
Smith, Tim ..... 23
Smolen, Tomasz ..... 2037
Snedeker, Jesse ..... 2242
Snoddy, Sean ..... 3215
Soderstrom, Melanie ..... 998
Sokolov, Mikhail ..... 3842
Solberg, Mads ..... 21
Somers, Sterling ..... 3221
Song, Amanda ..... 1096
Sonina, Varvara ..... 2096
Sopena-Garaikoetxea, Imanol ..... 3892
Soter, Laura ..... 871
Soyer, Hubert ..... 1319
Spangenberg, Marlene ..... 3843
Speed, Laura ..... 258, 3227
Speekenbrink, Maarten ..... 192, 1047, 1357,
2162, 3848, 3872
Speer, Shari ..... 1144
Spelke, Elizabeth ..... 766
Spencer, John ..... 5
Sperber, Dan ..... 3775

## Author Index

Spevack, Samuel ..... 3844
Spezio, Michael ..... 3228
Spicer, Jake ..... 1102
Spike, Matthew ..... 1108
Spivey, Michael ..... 3229
Spotorno, Sara ..... 3761, 3820, 3845
Sprenger, Jan ..... 1806
Sreekumar, Vishnu ..... 295
Srinivas, Sadhwi ..... 3235
Srinivasan, Mahesh 379, 986, 1120, 2019,3241
Srivastava, Nisheeth ..... 1114, 3247, 3248
Stafford, Tom ..... 25, 955, 3249
Stanciu, Oana ..... 3846
Starr, Ariel ..... 1120
Stas, Aurélie ..... 3435
Stassun, Keivan ..... 2754
Staudte, Maria ..... 2261, 3107
Steensig, Jakob ..... 2055
Stefanucci, Jeanine ..... 3717
Steffensen, Sune Vork ..... 1345
Steinemann, Sharon ..... 11
Steinman, David ..... 3874
Stemmler, Antje ..... 2079
Stengård, Elina ..... 3847
Stephan, Klaas Enno ..... 3807
Stephan, Simon ..... 1126, 1132
Stephens, Rachel ..... 1138
Stepp, Nigel ..... 1321
Stevens, Christopher ..... 355
Stevens, Jon ..... 1144, 3004, 3255
Stevenson, Suzanne ..... 112, 853
Stewart, Andrew ..... 3534
Stewart, Neil ..... 1635
Stewart, Rebecca ..... 3186
Stewart, Terrence ..... 631, 1150
Steyvers, Mark ..... 793, 1623
Stieff, Mike ..... 27
Stockall, Linnaea ..... 724
Stoeber, Mathias ..... 3010
Stojic, Hrvoje ..... 3848

## Author Index

Tanida, Yuki ..... 3854
Tanner, Kimberly ..... 3690
Tansan, Merve ..... 1471
Tao, Ye ..... 3878
Tärning, Betty ..... 1169
Tarr, Michael ..... 136
Tatler, Ben ..... 3845
Taub, Michelle ..... 1666, 2766
Tavernini, Lucía Micaela ..... 1206, 3296
Taves, Ann ..... 17
Tay, Charlene ..... 3855
Taylor, Gemma ..... 817
Taylor, J. S. H. ..... 1740
Taylor, Ryan ..... 1794
Tenebaum, Joshua ..... 3209
Tenenbaum, Harriet ..... 3869
Tenenbaum, Josh $13,19,37,676,766,792$,
$1393,2475,2536,3584,3670,3752$
Tenenbaum, Joshua B. ..... 1715
Tenenbaum, Joshua ..... 409
Tenison, Caitlin ..... 961, 1175
Tentori, Katya ..... 1459
Terai, Hitoshi ..... 3856
Terzopoulos, Demetri ..... 700
Tessler, Michael Henry ..... 1181, 3602
Thapa Magar, Stuti ..... 1848
Theakston, Anna ..... 1691, 3699
Theodorou, Lida ..... 3302
Thibaut, Jean-Pierre ..... 439, 3308, 3314Thibodeau, Paul H. .......................... 2187
Thibodeau, Paul ..... 361, 3320, 3326
Thiessen, Erik ..... 670, 3719
Thomas, Michael ..... $63,538,3718$
Thomaschke, Roland ..... 3830, 3857
Thompson-Schill, SharonThompson, Bill3858
Thompson, Valerie ..... 3528
Thrash, Tyler ..... 3675
Thurn, Christian ..... 2809
Tiganj, Zoran ..... 1187
Tilbe, Timothy ..... 3332
Tillman, Gabriel ..... 3338
Tillman, Katharine ..... 35, 1193
Timmons, Shane ..... 3859
Tiongco, Jacqueline ..... 3792
Tirumala, Dhruva ..... 1319
Todaro, Rachael ..... 3344
Todd, Peter ..... 234
Todo, Naoya ..... 1556
Tolmie, Andrew ..... 538
Tomita, Akitoshi ..... 3860
Tomm, Brandon ..... 3350
Tonhauser, Judith ..... 1144
Torii, Takuma ..... 519, 2205
Tossonian, Méghane ..... 3861
Tourtouri, Elli ..... 3356
Toussaint, Marc ..... 3584
Tovar, Angel ..... 1199
Toya, Genta ..... 3862
Trafton, Greg ..... 3362
Trafton, J. Gregory ..... 3664
Trausan-Matu, Stefan ..... 1884
Travers, Eoin ..... 1205
Trench, Máximo ..... 1206, 2735, 3296
Trick, Lana M. ..... 3870
Trouche, Emmanuel ..... 3368
Troyer, Melissa ..... 1212
Trude, Alison ..... 1218
Trueblood, Jennifer . 1387, 1394, 2741, 3374
Trueswell, John ..... 37, 1320
Truschzinski, Martina ..... 1224
Tse, Alice Ping Ping ..... 3380
Tsetsos, Konstantinos ..... 1230
Tso, Ricky Van-yip ..... 3386
Tsoukala, Chara ..... 3392
Tulkki, Leena ..... 2888
Tupper, Paul ..... 588, 3863
Turk, Russell ..... 3398
Tversky, Barbara ..... 1776
Twomey, Katherine ..... 3404
Twyman, Matt ..... 2162

## Author Index

Tylén, Kristian ..... 494, 2055
Udale, Rob ..... 3410
Ueda, Kazuhiro 557, 2043, 3143
Ugarte Gómez, Estibaliz ..... 2432
Ullery, Mary Anne ..... 3811
Ullman, Tomer ..... 766, 2475, 3411
Umata, Ichiro ..... 3864
Unger, Layla ..... 1242, 3417, 3865
Uprichard, Brian ..... 3781
Urminsky, Oleg ..... 2748
Usal, Kerem Alp ..... 3423
Uttal, David ..... 27
Vahabie, Abdol-Hossein ..... 3778
Valenti, J. J. ..... 606
Valenzuela-Rendon, Manuel ..... 1507
Valerjev, Pavle ..... 3429
Vales, Catarina ..... 1236, 1242
Vallee-Tourangeau, Frederic .... 1248, 2181Vallee-Tourangeau, Gaelle ............... 2181
Valtierra, Adriana ..... 3811
Van Allan, Jason ..... 2102
van Amelsvoort, Marije ..... 1866
Van Boekel, Martin ..... 1610, 3818
Van de Velde, Maartje ..... 3815
van den Berg, Ronald ..... 3847
van den Bos, Wouter ..... 3082
van der Wijst, Per ..... 1866
Van Herwegen, Jo ..... 3866
van Opheusden, Bas ..... 1254
van Ravenzwaaij, Don ..... 694
van Rij, Jacolien ..... 3704
van Rooij, Iris ..... 45, 2907
van Schijndel, Marten ..... 1260
van Vugt, Marieke ..... 3890
Vande Velde, Anna ..... 670
Vanpaemel, Wolf ..... 3022
Varma, Keisha ..... 1610, 3818
Varma, Sashank ..... 1610
Vasilyeva, Nadya ..... 1266
Vasishth, Shravan 1272, 1278

## Author Index

Walker, Neil ..... 1307
Wallace, Sarah ..... 1290
Wallinheimo, Anna-Stiina ..... 3869
Wallraven, Christian ..... 1393
Walmsley, Lachlan ..... 1313
Walsh, Matthew ..... 1800
Walsh, Michael ..... 3770
Wang, Felix Hao ..... 2243
Wang, Hao ..... 1320, 3490
Wang, Hongbin ..... 2790, 3279
Wang, Jane ..... 1319
Wang, Li ..... 3879
Wang, Liangliang ..... 3863
Wang, Qi ..... 3879
Wang, Ruiming ..... 192
Wang, Wen Chung ..... 3876
Wang, Yuyan ..... 3878
Waring, Rylan J. ..... 3870
Waring, Rylan ..... 1150
Warren, Jane E. ..... 1647
Was, Christopher ..... 3777
Washburn, Auriel ..... 1321, 3871
Wasserman, Edward ..... 1453, 1457, 1469
Wasylyshyn, Christina ..... 1611
Watanabe, Katsumi ..... 2426, 3793
Watson, Julia ..... 112
Watts, Paul ..... 252
Wei, Hui ..... 3491
Weinberger, Emma ..... 3792
Weinstein, Ari ..... 1327
Weisman, Kara ..... 1333
Weiss-Cohen, Leonardo ..... 3872
Welbourne, Stephen ..... 811
Wellman, Henry ..... 99
Welsh, Matthew ..... 2211, 3497
Wente, Adrienne ..... 3503
Werning, Markus ..... 3504
Wertheim, Julia ..... 3510
West, Jasmine ..... 3792
Westermann, Gert ..... 1199, 3404

## Author Index

Wu, Yang ..... 1363, 3553
Wu, Yuchunzi ..... 3552
Xiang, Siqi ..... 3554
Xiang, Wendi ..... 3658
Xiong, Qingyu ..... 3554
Xiong, Zhang ..... 3554
Xu, Fei ..... 1078, 2699
Xu, Qinmei ..... 3878
Xu, Tian Linger ..... 1369, 1412
Xu , Xiaoqing ..... 1550
Xu, Yang ..... 986
Xu, Yian ..... 3690
Xygalatas, Dimitris ..... 17
Yama, Hiroshi ..... 2882
Yamamoto, Seiichi ..... 3864
Yan, Veronica ..... 3615
Yang, Charles ..... 37, 3730, 3837
Yang, Hwajin ..... 3726
Yang, Jin-Biao ..... 1451
Yang, Lee-Xieng ..... 3880
Yang, Scott Cheng-Hsin ..... 1375, 3560
Yang, Yang ..... 3879
Yano, Hiroaki ..... 3860
Yarlett, Daniel ..... 313
Yasuda, Tetsuya ..... 3291
Yates, Tim ..... 1381
Ybarra, Vincent ..... 3566
Yearsley, James ..... 1387, 1592, 3572
Yengin, Ilker ..... 2199
Yeo, Amelia ..... 3578
Yeung, Susanna ..... 3876
Yeung, Yui Andrew ..... 3386
Yildirim, Ilker ..... 19, 1393, 3584
Yim, Hyungwook .................. 295, 3590
Yin, Siyuan ..... 1394
Yip, Chi Wing ..... 3876
Yip, Michael ..... 3881
Yokoyama, Sho ..... 3856
Yoo, Angela ..... 3596
Yoon, Erica ..... 3602

## Author Index

Zhao, Joyce Wenjia ....................... 1436 Zhu, Song-Chun ..... 700, 1066, 2579
Zhao, Mintao ..... 525
Zhao, Yibiao ..... 13
Zheng, Yeqiu ..... 445
Zhou, Guojing ..... 3652
Zhou, Liang ..... 409
Zhu, Jian-Qiao ..... 3658
Zhu, Mona ..... 3891
Zhu, Yixin ..... 700, 2579
Zinszer, Benjamin ..... 1752
Zipitria, Iraide ..... 3892
Zish, Kevin ..... 3362, 3664
Zrubka, Mark ..... 3806
Zuidema, Willem ..... 1531
Zuluaga-Rendón, Zadkiel ..... 2073

## Reviewers

| Linda Abarbanell | Nada Attar | Francesca Benuzzi |
| :--- | :--- | :--- |
| Ahmed M. H. Abdel-Fattah | Luc Augier | Elika Bergelson |
| Akinori Abe | Joseph Austerweil | Leon Bergen |
| Drew Abney | Martijn Baart | Christina Bergmann |
| Steven Abney | Stephanie Badde | Kirsten Bergmann |
| Dor Abrahamson | Gi-Yeul Bae | Luc Berthouze |
| Ekaterina Abramova | Rebekah Baglini | Tommaso Bertolotti |
| Cengiz Acarturk | Jonathan Bakdash | Tarek R. Besold |
| Vincent Adam | Chris Baker | Brad Best |
| Staub Adrian | Lewis Baker | Sudeep Bhatia |
| Javier Aguado-Orea | Ryan Baker | Francesco Bianchini |
| Sheeraz Ahmad | M Bakker | Irving Biederman |
| Woo-kyoung Ahn | Pragathi Priyadharsini Balasub- | Perrin Bignoli |
|  | ramani |  |
| Ozge Alacam | Adriana Alexandra Baltaretu | Nik Nailah Binti Abdullah |
| Vincent Aleven | Valentina Bambini | Tamas Biro |
| Martha Alibali | Bonny Bannerjee | Jennifer Bittner |
| Dimitrios Alikaniotis | David Barack | Michael Bixter |
| Afra Alishahi | Libby Barak | Leslie Blaha |
| David Allbritton | Michael Baranski | Peter Blake |
| Laura Allen | Jean Baratgin | Julia Blau |
| Cecilia Alm | Roxana-Maria Barbu | Stephen Blessing |
| Richard Alterman | Mike Barley | Elma Blom |
| Nadja Althaus | Daniel Bartels | Megan Blossom |
| Nicholas Altieri | Andrew Bauer | Peter Blouw |
| Ercan Altinsoy | Malcolm Bauer | Richard Blythe |
| Erik Altmann | Merja Bauters | Aysecan Boduroglu |
| Eric Amazeen | Paul Baxter | Reihane Boghrati |
| Ben Ambridge | Louise Beattie | Sander Bohte |
| Elena Andonova | Fabian Beck | Luca Bonatti |
| Mark Andrews | Nicole Beckage | Rebecca Boncoddo |
| Caterina Ansuini | Barend Beekhuizen | Jean-Francois Bonnefon |
| Ian Apperly | Kreshnik Begolli | Alexander Boone |
| Scott Appling | Marlene Behrmann | Anna Borghi |
| Engin Arik | Ulrik Beierholm | Arielle Borovsky |
| Blair Armstrong | Peter beim Graben | Jelmer Borst |
| Tracy Arner | Anna Belardinelli | Heather Bortfeld |
| Jennifer Arnold | Roman Belavkin | Solveig Bosse |
| Maria Arredondo | Daniel Belenky | Mona Roxana Botezatu |
| Burcu Arslan | Sieghard Beller | Lewis Bott |
| Sudha Arunachalam | Paul Bello | Oliver Bott |
| Mika Asaba | Andrea Bender | Roberto Bottini |
| Richard Ashley | Viridiana L. Benitez | Howard Bowman |
| Natsuki Atagi | Tessa Bent | Huseyin Boyaci |
|  |  |  |
|  |  |  |
|  |  |  |

## Reviewers

Ty Boyer
David Braithwaite
Neil Bramley
Amanda Brandone
Martin Brandt
Brooke Breaux
Cameron Brick
Will Bridewell
Sophie Bridgers
Gordon Briggs
Henry Brighton
Renaud Brochard
Joseph Brooks
Thomas Brooks
Harm Brouwer
Gordon Brown
Meredith Brown
Riccardo Brunetti
Angela Brunstein
Joerg Brunstein
Tad Brunye
David Bryant
Joanna Bryson
David Buchanan
Daphna Buchsbaum
Cameron Buckner
Marc Buehner
John Bullinaria
John Bunce
Ann Bunger
Michele Burigo
Joseph Burling
Bruce Burns
Edwin Burns
Hendrik Buschmeier
Wednesday Bushong
Lucas Butler
stephen Butterfill
Martin Butz
Esteban Buz
Krista Byers-heinlein
Patrick Byers
Matt Cain

Murat Perit Cakir
Josep Call
Charles Callaway
Andres Campero
Angelo Cangelosi
Reality Canty
valentina cardella
Nancy Carlisle
Richard Carlson
Floriana Carmeci
Alexandra Carstensen
Diana Carter
Paulo Carvalho
Leslie Carver
Marisa Casillas
Gianluca Caterina
Richard Catrambone
Gabrielle Cayton-Hodges
Gustavo Cevolani
Chayan Chakrabarti
Jenny Yun-Chen Chan
Margaret Chan
wen-hsuan chan
Sanjay Chandrasekharan
Franklin Chang
Nick Chater
Emmanuel Chemla
Dar-Wei Chen
Yi-Chia Chen
Peter Cheng
Erik Cheries
Nadia Chernyak
Dana Chesney
Fabienne Chetail
Pierina Cheung
Samuel Cheyette
Jane Childers
Seth Chin-Parker
Jessie Chin
Enrico Chiovetto
Eleanor Chodroff
Elizabeth Chrastil
Peer Christensen

Wayne Christensen
Stella Christie
Evangelia Chrysikou
Mingyuan Chu
Emily Cibelli
Philipp Cimiano
Benjamin Cipollini
John Clapper
Christophe Claramunt
Timothy Clausner
Meghan Clayards
Jennifer Clegg
Charles Clifton
Adam Cobb
Moreno I. Coco
Yann Coello
Anna Coenen
Uriel Cohen Priva
Andrew Cohen
Neil Cohn
Jeremy Cole
John Coley
Michael Collins
Matteo Colombo
Arzu Coltekin
Nolan Conaway
Kensy Cooperrider
Wright Cory
Fintan Costello
Gary Cottrell
Helen Couclelis
Caitlin Coughlin
Christopher Cox
Gregory Cox
Scotty Craig
Christopher Crick
Ilja Croijmans
Vincenzo Crupi
Jennifer Culbertson
Chris Cummins
Fred Cummins
Christine Cuskley
Ilaria Cutica

## Reviewers

Sidney D'Mello
Rick Dale
Gregory Dam
Debby Damen
Christopher Dancy
David Danks
Judith Danovitch
Geetanjali Date
Isabelle Dautriche
Eddy Davelaar
Jim Davies
Nick Davis
Bart de Boer
Felipe De Brigard
Simon De Deyne
Iwan de Kok
Josh de Leeuw
Rita de Oliveira
J.p. de Ruiter

Virginia de Sa
Gedeon Deák
Christelle Declercq
Judith Degen
Morteza Dehghani
Jasmine Dejesus
Yvonne Delevoye-turrell
Sara Dellantonio
Claire Delle Luche
Wei (sophia) Deng
Stephanie Denison
Jean-Louis Dessalles
Emilie Destruel-Johnson
Massimiliano Di Luca
Julien Diard
Zoltan Dienes
Berkeley Dietvorst
Brian Dillon
Alexia Toskos Dils
Mark Dingemanse
Tuan Do
Kristina Doing-Harris
Tim Donovan
Alex Doumas

Gabe Doyle
Colleen Dragovich
Trafton Drew
Felix Dreyer
Adam Dubé
John Duff
Susan Duncan
Kristen Dunfield
Frank Durgin
Melody Dye
Sarah Dziura
Catharine Echols
Peter Edelsbrunner
Charlotte Edmunds
Brian Edwards
Alyson Eggleston
Christian Eichhorn
Michelle Ellefson
Mark Elliott
Sarah Elliott
Susan Epstein
Edgar Erdfelder
Goker Erdogan
Zachary Estes
Sara Irina Fabrikant
Zoe Falomir
Mirko Farina
Igor Farkas
George Farmer
Thomas Farmer
Simon Farrell
Caitlin Fausey
Luis H. Favela
Mark Fedyk
Maryia Fedzechkina
Roman Feiman
Michele Feist
Naomi Feldman
Maggie Feng
Vanessa Ferdinand
Fernanda Ferreira
Alissa Ferry
Mary Anne Fields

Benjamin Files
Mark Finlayson
Sara Finley
Helen Fischer
Jozsef Fiser
Anna Fisher
Christopher Fisher
Julia Fisher
Paula Fitzpatrick
Nick Flor
Stephen Flusberg
Jonathan Folstein
Anouschka Foltz
Ken Forbus
Carolyn Forsyth
Emily Foster-Hanson
Chris Foster
Meadhbh Foster
Alice Foucart
Abdellah Fourtassi
Amy Fox
Elizabeth Fox
Mary Frame
Elaine Francis
Mike Frank
Stefan Frank
Stella Frank
Till Frank
Michael Franke
Diego Frassinelli
Alexandra Frazer
Christian Freksa
Bob French
Jim Friedrich
Steven Frisson
Chris Frith
Caren Frosch
Rebecca Frost
Shimin Fu
Wai-Tat Fu
Wai Fu
Kim Kristina Fuellenbach
Richard Futrell

## Reviewers

| Rami Gabriel | Adele Goldberg | John Hale |
| :--- | :--- | :--- |
| Alexia Galati | Rob Goldstone | Jessica Hall |
| Chiara Gambi | Micah Goldwater | Matthew Hall |
| Joanna Ganczarek | Roberta Golinkoff | Tim Halverson |
| Song Gao | David Gomez | Kai Hamburger |
| Raquel Garcia Jurado | Rebecca Gomez | Aaron Hamer |
| Ricardo Garcia | Laura Gonnerman | Jessica Hamrick |
| Kelly Garner | Monica Gonzalez-Marquez | Agness Hara |
| Bärbel Garsoffky | Cleotilde Gonzalez | J. Isaiah Harbison |
| Justin Garten | Noah Goodman | Jesse Harris |
| Remi Gau | Luise Gootjes-Dreesbach | Kevin Harris |
| Nirit Gavish | Martijn Goudbeek | Anthony Harrison |
| Ross Gayler | Art Graesser | Joshua Hartshorne |
| Lupita Estefania Gazzo | Cas- | Erin Graham |
| taneda | Charlotte Hartwright |  |
| Hector Geffner | Klaus Gramann |  |
| Charles Geier | Pär-Anders Granhag | Olaf Hauk |
| Susan Gelman | Erin Grant | Robert X.d. Hawkins |
| Dedre Gentner | Suzanne Grassmann | Kara Hawthorne |
| György Gergely | C. Shawn Green | Brett Hayes |
| Samuel Gershman | Nathan Greenauer | Mary Hayhoe |
| Tobias Gerstenberg | Stefan Th. Gries | Angela Xiaoxue He |
| Bryce Gessell | Zenzi Griffin | Patrick Healey |
| Moojan Ghafurian | Sascha Griffiths | Tobias Heed |
| Saeideh Ghahghaei | Tom Griffiths | Mary Hegarty |
| Shadi Ghajar-khosravi | Lisa Grimm | Dietmar Heinke |
| Rikhiya Ghosh | Maurice Grinberg | Evan Heit |
| Valeria Giardino | Giorgio Gronchi | Steve Hekkanen |
| Todd Gibson | Ricardo Gudwin | Daphna Heller |
| Nina Gierasimczuk | Amal Guha | Emily Hembacher |
| Anne Giersch | Curry Guinn | Marlone Henderson |
| Sandrine Gil | Christine Gunlogson | Andrew Hendrickson |
| Daniel Gildea | Glenn Gunzelmann | Li Heng |
| Nicholas Giudice | Todd Gureckis | José Hernández-Orallo |
| Joseph Glavan | Rusan Hespos |  |
| Vladimir Glebkin |  | Lisa Guthrie |

## Reviewers

Jean-Remy Hochmann
Marieke Hoetjes
Janina Anna Hoffmann
Annette Hohenberger
Andreas Højlund
John Holden
Diessel Holger
Judith Holler
Kevin Holmes
garrett honke
Joe Hoover
Ryan Hope
Michael Horn
Zachary Horne
R Horry
William Horton
Joseph Houpt
Michael Hout
Andrew Howes
Christine Howes
Evgenia Hristova
Penka Hristova
Janet Hsiao
Nina Hsu
Edward Hubbard
Timothy Hubbard
Roland Hubscher
Eva Hudlicka
Carla Hudson Kam
Julie Hupp
E. Matthew Husband

Erika Hussey
Thomas Icard
Mutsumi Imai
Alistair Isaac
Toru Ishikawa
Eduardo Izquierdo
Tanner Jackson
Cassandra Jacobs
Robert Jacobs
T. Florian Jaeger

Saransh Jain
Yoonhee Jang

Stefan Janke
Petra Jansen
Julian Jara-Ettinger
Andreas Jarvstad
Brett Jefferson
Alan Jern
Patrick Jeuniaux
William Jimenez-Leal
Mark Johansen
Elizabeth Johnson
Kate Johnson
Mina Johnson
Samuel Johnson
Angie Johnston
Gary Jones
William Jones
Pamela Jordan
David Joyner
Barbara Juhasz
Ion Juvina
George Kachergis
Sebastian Kahl
rachelle kaldani
Charles Kalish
Michael Kalish
Irene Kan
Subu Kandaswamy
Vsevolod Kapatsinski
Effie Kapnoula
White Katherine
Irvin Katz
John Kearns
Madeleine Keehner
Gitit Kehat
Christopher Kello
Damian Kelty-Stephen
Yoed Kenett
William Kennedy
Chris Kent
Christopher Kent
Trina Kershaw
Alan Kersten
Bryan Kerster

Shaiyan Keshvari
Azam Khan
Sangeet Khemlani
Peter Khooshabeh
Celeste Kidd
Douwe Kiela
Dahee Kim
Shinwoo Kim
Simon Kirby
Szabolcs Kiss
Özkan Kılıç
Max Kleiman-Weiner
Dave Kleinschmidt
Christian Kliesch
Melissa Kline
Alexander Klippel
Heidi Kloos
Joshua Knobe
Pia Knoeferle
Dominique Knutsen
Adam Koblinger
Ken Koedinger
Judith Koehne
Sungryoung Koh
Takanori Komatsu
Elizabeth Kon
Piotr Konderak
Agnieszka Konopka
Emmanouil Konstantinidis
Ruud Koolen
Stefan Kopp
Mark Koranda
Hatice Kose
Janet Koster van Groos
Aditi Kothiyal
Peter Krafft
Emiel Krahmer
Attila Krajcsi
Alexander Kranjec
Ekaterina Kravtchenko
Josef Krems
Nikhil Krishnaswamy
Julia Marina Kröger

## Reviewers

| Edmundo Kronmüller | Yuan Chang Leong | Dermot Lynott |
| :---: | :---: | :---: |
| Jakub Krukar | Rachel Lerch | Chris MacDonald |
| Ulf Krumnack | Nicholas Lester | Kyle MacDonald |
| Marta Kryven | Beth Levin | Maryellen MacDonald |
| Karolina Krzyżanowska | William Levine | Ross Macdonald |
| Stefanie Kuchinsky | Roger Levy | Korey MacDougall |
| Sarah Kucker | Ashley Lewis | Luís Macedo |
| Carolina Kuepper-Tetzel | Molly Lewis | Michael Mack |
| Tim Kühl | Jixing Li | John Mackay |
| Kai-Uwe Kühnberger | Rui Li | Colin MacLeod |
| Anuenue Kukona | Peipeng Liang | Neil Macrae |
| Philipp Kulms | Patricia Lichtenstein | Alex Madva |
| Bora Kumova | Karen Lichtman | Akihiro Maehigashi |
| Maithilee Kunda | Amy Lieberman | Alfons Maes |
| Victor Kuperman | Falk Lieder | Hagit Magen |
| Tsuyoshi Kuroda | Constantine Lignos | Brian Magerko |
| Chigusa Kurumada | Shane Lindsay | Rachel Magid |
| Peter Kvam | Marcus Lindskog | Paul Maglio |
| Johan Kwisthout | Daniel Little | Lorenzo Magnani |
| Joy Kwon | Hannah Little | Phil Maguire |
| Tei Laine | Pei Q Liu | Steve Majerus |
| Brenden Lake | Jeffrey Loewenstein | Asifa Majid |
| Michelle Lamar | Johannes Lohmann | Vicente Malave |
| Karen Lander | Tania Lombrozo | Hanspeter Mallot |
| David Landy | Lyle Long | Kenneth Malmberg |
| Gabriella Lapesa | Deryle Lonsdale | Laurence Maloney |
| Othalia Larue | Tamara Lorenz | Kingson Man |
| Daniel Lassiter | Sebastian Loth | Eric Mandelbaum |
| Suilin Lavelle | Jeff Loucks | Olivier Mangin |
| Chris Lawson | Brad Love | Liliann Manning |
| Olivier Le Guen | Andrew Lovett | Francisco Maravilla |
| Brooke Lea | Jason Low | Denis Mareschal |
| David Leake | Hongjing Lu | Tyler Marghetis |
| Tania Leal | Christopher Lucas | Doug Markant |
| Gianluca Lebani | Elliot Ludvig | Julie Markant |
| Christian Lebiere | Casimir Ludwig | Art Markman |
| Jessica Lee | George Luger | Hanna Marno |
| Blair Lehman | Christian Luhmann | Jessecae Marsh |
| Mallorie Leinenger | Steven Luke | Andrea Martin |
| Jaroslaw Lelonkiewicz | Jian Luo | Amy Masnick |
| Alessandro Lenci | Stephen Lupker | Robert Mason |
| Julia Leonard | Gary Lupyan | J. Douglas Mastin |
| Giuseppe Leonardi | Michael Lynch | Emily Mather |

## Reviewers

| Bryan Matlen | Laura Morett | Eva Nunnemann |
| :---: | :---: | :---: |
| Miki Matsumuro | Brent Morgan | Padraig O'Seaghdha |
| Percival Matthews | Emily Morgan | Tim Oates |
| Jon Matthis | Matteo Morganti | Unaizah Obaidellah |
| Laura Matzen | Tiffany Morisseau | Natalie Obrecht |
| Ralf Mayrhofer | Bradley Morris | Marta Olivetti |
| Stewart McCauley | Fermin Moscoso del Prado Martin | Maria Olkkonen |
| Matt McCrudden | Jarrod Moss | Andrew Olney |
| David McDonald | Larry Moss | Ana-Maria Olteteanu |
| Danielle McNamara | Benjamin Motz | Sasha Ondobaka |
| Marjorie McShane | Linda Moxey | Desmond Ong |
| Ben Meadows | Johannes Mueller-Trede | Luca Onnis |
| Bjoern Meder | Tim Mullett | John Opfer |
| Jared Medina | Edward Munnich | Isabel Orenes |
| Tim Meese | Paul Munro | Daniel Osherson |
| Tobias Meilinger | Dafne Muntanyola | Jake Oury |
| Catherine Mello | Stefan Münzer | Iris Chuoying Ouyang |
| Melisa Menceloglu | Gregory Murphy | David Over |
| David Mendonca | Wayne Murray | Karenleigh Overmann |
| Mythili Menon | Christopher Myers | Tolga Esat Ozkurt |
| Hugo Mercier | Myrto Mylopoulos | Michael Pacer |
| Marie Meteer | Nashid Nabian | Thorsten Pachur |
| Meredith Meyer | Jonas Nagel | Mike Page |
| Stephan Meylan | Letitia Naigles | Bozena Pajak |
| Christos Michaelides | Kuninori Nakamura | Kiran pala |
| Laura Mickes | Laura Namy | Zsolt Palatinus |
| Gareth Miles | Bhuvana Narasimhan | Prajakt Pande |
| Brent Miller | Mitchell Nathan | John Pani |
| Craig Miller | David Natvig | Peter Pantelis |
| Hilary Miller | Jonathan Nelson | Anna Papafragou |
| Daniel Mirman | Aida Nematzadeh | Carita Paradis |
| Percy Mistry | Dezso Nemeth | Natasha Parikh |
| Krishna Miyapuram | Nora Newcombe | Dan Parker |
| Takashi Mizuguchi | Alan Nielsen | Jeff Parker |
| Reza Moazzezi | Laura Niemi | Philip Pärnamets |
| Jeff Moher | David Noelle | Andrea Patalano |
| Francis Mollica | Takao Noguchi | Purav Patel |
| Monika Molnar | Nicholaus Noles | Sturt Patrick |
| Padraic Monaghan | Michael Noll-Hussong | John Patterson |
| Jessica Montag | Jonas Nölle | David Pautler |
| Daniel Montello | Uta Noppeney | Alexandra Paxton |
| Jung Aa Moon | Ira Noveck | Stephen Payne |
| Claire Moore-Cantwell | Jared Novick | Diane Pecher |

## Reviewers

Sarah Bro Pedersen
David Peebles
Jonathan Peelle
Marcie Penner-Wilger
Antonino Pennisi
Gordon Pennycook
Brandon Perelman
Amy Perfors
Marcus Perlman
Lynn Perry
Kimele Persaud
Mathias Pessiglione
Megan Peters
Ryan Peters
Joshua Peterson
Matthew Peterson
Georgi Petkov
Niki Pfeifer
Eugene Philalithis
Jonathan Phillips
Steven Phillips
Steve Piantadosi
Maria M. Pinango
Yair Pinto
Pantelis Pipergias Analytis
Alex Pitti
Myrthe Plaisier
Alessio Plebe
Amanda Pogue
Rüdiger F. Pohl
Ellen Poliakoff
Stephen Politzer-Ahles
Guy Politzer
Kimberly Pollard
Vencislav Popov
Jan Pöppel
Vincent Porretta
Sophie Portrat
Marie Postma-Nilsenová
Nausicaa Pouscoulous
Georgina Powell
Ranjani Prabhakaran
Sandeep Prasada

Vitor Profeta
Athanassios Protopapas
James Pustejovsky
Cynthia Putnam
Aryn Pyke
David Pynadath
Ting Qian
Zhiying Qian
Jennifer Queen
Hugh Rabagliati
Eva Rafetseder
Anna Rafferty
Marco Ragni
jason ralph
David Rapp
Okko Johannes Räsänen
Martina Rau
Limor Raviv
Devin Ray
Florencia Reali
Gisela Redeker
Stephen Reed
Patricia Reeder
Alice Rees
Bob Rehder
Nicholas Reinholtz
David Reitter
Michael Rescorla
Kate Revill
Russell Revlin
Constantin Rezlescu
Darren Rhodes
Theo Rhodes
Alexander Rich
Mike Richardson
Elizabeth Richey
Lindsey Richland
Kai-Florian Richter
Hannes Rieser
Lucia Riggio
Michael Riley
Lance Rips
Lilia Rissman

Frank Ritter
Talia Robbins
Jennifer Roche
Stuart Rodgers
Florian Roehrbein
Tanja Roembke
Neal Roese
Hannah Rohde
Katharina Rohlfing
Alexa Romberg
Samuel Ronfard
Paul Rosenbloom
Karl Rosengren
Constantin Rothkopf
Timo Röttger
Benjamin Rottman
Brandon Roy
Paula Rubio-Fernandez
Azzurra Ruggeri
Simon Rushton
Rachel Ryskin
Selma Sabanovic
Matthew Sachs
Eyal Sagi
Antoine Saillenfest
Joanna Salapska-Gelleri
Adam Sanborn
Guilherme Sanches de Oliveira
Yuliya Sandamirskaya
Lelyn Saner
Sudeep Sarkar
Barbara Sarnecka
Cristina Savin
Andrew Saxe
Brian Scassellati
Walter Schaeken
Graham Schafer
Job Schepens
Fabian Schlotterbeck
Ute Schmid
Stefan Schneider
David Schnyer
H.steven Scholte

## Reviewers

Marc Scholten
Christoph Schommer
Liz Schotter
Fabian Schrodt
Carl Schultz
Eric Schulz
Christian Schunn
Nora Schwartz
Stephanie Schwartz
Michael Schwartze
John Schwenkler
Angela Schwering
Travis Seale-Carlisle
Natalie Sebanz
Julie Sedivy
Colleen Seifert
Andrea J. Sell
Florian Sense
Maria Sera
Michael J. Serra
Thomas Serre
Scott Seyfarth
Natela Shanidze
Kailen Shantz
Dinkar Sharma
Timothy Shea
Wenlei Shi
Zhuanghua Shi
Thomas Shipley
Kevin Shockley
Andrew Shtulman
Keith Shubeck
Mo Shukla
Anna Shusterman
Michael Shvartsman
Michael Siebers
Max Siegel
Noam Siegelman
Catriona Silvey
Patrick Simen
Luca Simione
Vanessa Simmering
Nina Simms

Chris Sims
Scott Sinnett
Jeffrey Mark Siskind
Paul Skarratt
Dimitrios Skordos
John Skowronski
William Skylark
Heleen Slagter
Lauren Slone
Vladimir Sloutsky
Paul Smaldino
Andy Smith
David Smith
Kenny Smith
Kevin Smith
Erica Snow
Gregg Solomon
Katya Solovyeva
Eisenbeiss Sonja
Katharina Spalek
George Spanoudis
Jesse Sparks
Shannon Spaulding
Laura Speed
Maarten Speekenbrink
Jennifer Spenader
Charles Spence
John Spencer
Michael Spezio
Stephen Spiller
Joseph Spino
Michael Spivey
Sara Spotorno
Michael Spranger
Maria Spychalska
Mahesh Srinivasan
Nisheeth Srivastava
Nitish Srivastava
Kevin Stadler
Tom Stafford
Joe Stafura
Sarah Stager
Matthew Stanley

Jeffrey Starns
Maria Staudte
Mark Steedman
Fabio Stefanini
Nancy L Stein
Shane Steinert-Threlkeld
John Stell
Keith Stenning
Simon Stephan
Christopher Stevens
Andrew Stewart
Terrence Stewart
Andrea Stocco
Kurt Stocker
Stephanie Stokes
Dietrich Stout
Sumarga Suanda
Jordan Suchow
Jeremiah Sullins
Jess Sullivan
Chen Sun
Naveen Sundar G
Abigail Sussman
John Sutton
Richard Sutton
Michael Swart
Marc Swerts
Kristen Syrett
Jakub Szymanik
Niels Taatgen
Whitney Tabor
Satohiro Tajima
Yasutake Takahashi
Durk Talsma
James Tanaka
Margaret Tarampi
Caglar Tas
Loukia Taxitari
Julia Taylor
Erno Teglas
farnaz tehranchi
Eva Telzer
David Temperley

## Reviewers

Thora Tenbrink
Josh Tenenbaum
Caitlin Tenison
Michael Tessler
Susan Teubner-Rhodes
Khanh-Phuong Thai
Jean-Pierre Thibaut
Paul Thibodeau
Erik Thiessen
Serge Thill
Rick Thomas
Bill Thompson
Clarissa Thompson
Jim Thompson
Robert Thomson
Charles Tijus
Alex Tillas
Sabine Timpf
Maurizio Tirassa
Rachael Todaro
Ida Toivonen
Martin Tomko
John Tomlinson
Judith Tonhauser
Peter Tornille
John Towler
Greg Trafton
Eoin Travers
Annie Tremblay
Bastien Trémolière
Susan Trickett
Melissa Troyer
Jennifer Trueblood
John Trueswell
Dido Tsigaridi
Ricky Van Yip Tso
Shannon Tubridy
Luca Tummolini
Paul Tupper
Barbara Tversky
Katherine Twomey
Kristian Tylen
Tomer Ullman

Monika Undorf
Christoph Ungemach
Layla Unger
Oleg Urminsky
David Uttal
Prakaiwan Vajrabhaya
Pavle Valerjev
Catarina Vales
Juan Valle-Lisboa
Frederic Vallee-Tourangeau
Loes Van dam
Tim Van de Cruys
Ronald van den Berg
Natalie van der Wal
Robrecht van der Wel
Jo Van Herwegen
Kurt Vanlehn
Alexandra Varga
Sashank Varma
Nadya Vasilyeva
Argiro Vatakis
Richard Veale
Bella Veksler
Natalia Vélez
Francesco Vespignani
Ricard Viñas de Puig
David Vinson
Peter Vishton
Jonathan Vitale
Haley Vlach
Carl Vogel
Jorrig Vogels
Paul Vogt
John Voiklis
Regina Vollmeyer
Titus von der Malsburg
Bettina von Helversen
Momme von Sydow
Wai Keen Vong
Wouter Voorspoels
Soroush Vosoughi
Ed Vul
Loan Vuong

Matt Wagers
Abby Walker
Caren Walker
Erin Walker
Alan Wallington
Clare Walsh
Matthew Walsh
Felix Wang
Frances Wang
Jane Wang
Panqu Wang
Todd Wareham
Walter Warick
Felix Warneken
Paul Warren
Christopher Was
Kirsten Weber
Andrew Wedel
Christoph Weidemann
Kristin Weingartner
Deena Weisberg
Matthew Welsh
Katharina Wendler
Janet Werker
Chris Westbury
Gert Westermann
Aaron Steven White
Jeffrey White
Bruno Wicker
Alex Wiegmann
Katja Wiemer-Hastings
Eliane Wiese
Eva Wiese
Andy Wills
Maria Wirzberger
John Wixted
Ralph Wojtowicz
Phillip Wolff
Sharon Wood
Charley Wu
Jiajun Wu
Xin Xie
Yang Xu

## Reviewers

Shaorong Yan
Brenda Yang
Lee-Xieng Yang
Scott Cheng-Hsin Yang
Mark Yates
Yaara Yeshurun-Dishon
Saiwing Yeung
Ilker Yildirim
Hyungwook Yim
Michael Yip
Erica Yoon
Si On Yoon

Galit Yovel
Yue Yu
Lei Yuan
Wen Yuan
Jiwon Yun
Dan Yurovsky
Ron Zacharski
Alessandra Zarcone
Jeffrey Zemla
Cristina Zepeda
Arkady Zgonnikov
Jiaxiang Zhang

Qiong Zhang
Jiaying Zhao
Xuan Zhao
Sheng-hua Zhong
Jayden Ziegler
Brian Zikmund-Fisher
Iraide Zipitria
Kevin Zish
Laura Ziznewski
Filio Zourou


[^0]:    ${ }^{1}$ See http://bit.ly/2kPpZtF, checked 1/30/2017

[^1]:    ${ }^{1}$ http://ratiolog.uni-koblenz.de/bridging.html

[^2]:    Organized by Wayne D. Gray. Address all correspondence to Wayne Gray [wayne.gray.cogsci@gmail.com](mailto:wayne.gray.cogsci@gmail.com).

[^3]:    ${ }^{1}$ For an accessible introduction to enactivism see (McGee, 2005).

[^4]:    ${ }^{2}$ This interpretation is suggested by a version of the task by (Froese, Iizuka, \& Ikegami, 2014) in which participants' experience of the other's presence is one of the dependent variables.
    ${ }^{3}$ The reader should be aware that this is a necessary oversimplification of the variety of positions held within philosophy of science overall and that even within cognitive science there are heated debates on the precise understanding of deductivenomological, probabilistic, rational, mechanistic and other explanations. We merely focus on the distinction that has been most widely used in discussions between traditional cognitive science and REC.

[^5]:    ${ }^{4}$ This view composed of several statements can of course be translated into a continuum of positions. Arguing against the explanatory primacy of components might mean rejecting explanations that (a) ignore interactions, (b) assume only linear interactions, (c) assume only static interactions ignoring dynamics, (d) ignore the effect of parameters external to the system. We thank anonymous reviewer for pointing this out.

[^6]:    ${ }^{5} \mathrm{We}$ should note here that it might be that the mechanistic approach and enactivism rely on different notions of causality that make them incompatible. Addressing this possibility would require not merely examining the respective commitments but also the plausibility of particular models of causality assumed, e.g. whether circular causality typically adopted by enactivists is a helpful notion.

[^7]:    ${ }^{1}$ The integration of information across fixation is a local update for each cell.

[^8]:    ${ }^{1}$ In this paper, we use the term text to refer generally to any coherent or self-contained piece of spoken or written language.

[^9]:    ${ }^{2}$ According to a bag-of-words model, a language corpus is a set of texts, where each text is an unordered set, or bag, of words.

[^10]:    ${ }^{3}$ Full details about how the corpus was created, including all the code used to create it, is available at https://github.com/lawsofthought/tantalum
    ${ }^{4}$ Full details about the Gibbs sampler for the HDPMM, including the code implementing it, can be found at https://lawsofthought.github.io/gustavproject.
    ${ }^{5}$ Full details about how we sampled the texts, including the code implementing the sampling and the sampled texts themselves, can be found at https://github.com/lawsofthought/berkelium.
    ${ }^{6}$ Full details about how we sampled from the posterior predictive distribution, including the code implementing the sampling, can be found at https://github.com/lawsofthought/gallium.

[^11]:    ${ }^{7}$ Here, as above, we use the term text to denote any coherent and

[^12]:    self-contained piece of language.

[^13]:    ${ }^{8}$ Full details about how these two associative models were created, including the code implementing them, can be found at https://github.com/lawsofthought/gallium.

[^14]:    ${ }^{9}$ Full details about the recognition test word lists were created, including the code implementing this, can be found at https://github.com/lawsofthought/berkelium.

[^15]:    ${ }^{10}$ This is open-source software and is available at https://github.com/lawsofthought/wilhelmproject
    ${ }^{11}$ All raw data, and code for all analyses, can be found at https://github.com/lawsofthought/gallium.

[^16]:    ${ }^{1}$ Across all three experiments, we find no evidence of spatialnumerical associations in the vertical dimension.
    ${ }^{2}$ Consistent with previous research on pseudoneglect (Lourenco \& Longo, 2007; for review, see Jewel \& McCourt, 2000), there is an overall leftward bias in all three experiments.

[^17]:    ${ }^{3}$ Mean accuracy was not significantly different across the three experiments as determined by a one-way analysis of variance (ANOVA).

[^18]:    ${ }^{1}$ Note that we will be using the label "normative", "correct", or "logical" response as a handy shortcut to refer to "the response that has traditionally been considered as correct or normative according to standard logic or probability theory". The appropriateness of these traditional norms has sometimes been questioned in the reasoning field (e.g.,see Stanovich \& West, 2000, for a review). Under this interpretation, the heuristic response should not be labeled as "incorrect" or "biased". For the sake of

[^19]:    ${ }^{1}$ To help children understand the hypotheses under consideration, the confederate announced the full set of hypotheses before the child began teaching (e.g., "I see, Dax could mean fork, or Dax could mean white, or it could mean white fork.").

[^20]:    ${ }^{1}$ In an ongoing replication with adults, we are finding the same pattern of results as we did with children: Adults' ratings are influenced by the informant's knowledge state and the degree of information omission, but not by the value of the functions taught.

[^21]:    ${ }^{1}$ The set of languages is (per language family, in ISO 639-2): (Semitic) ar, he; (Indo-European) bg, bs, cs, da, de, el, en, es, fr, hr, it, nl, no, pl, pt, ro, ru, sl, sr, sv; (Finno-Ugric) et, fi, hu; (Austronesian) id; (isolate) ja; (Turkic) tr; (Vietic) vi; (Sino-Tibetan) zh.
    ${ }^{2}$ We used $k$-clique percolation (Palla et al., 2005) with $k=9$.
    ${ }^{3}$ We used a manually compiled stemming dictionary to lemmatize the words and correct spelling and alphabetic variation.

[^22]:    ${ }^{4}$ Annotation was done for English only. It is possible that Haspelmath's functions are not always translated: a conditional may be translated as an declarative. Being relatively infrequent, we consider these cases noise.
    ${ }^{5}$ Evaluating the method on a parallel Bible corpus against a gold standard of Strong number annotations gives a cluster purity of .89 and a cluster recall of . 90 .

[^23]:    ${ }^{6}$ We use sets of translated terms, because an indefinite pronoun in English may be translated to multiple terms in other languages.

[^24]:    ${ }^{7}$ These clusters do not completely coincide with the further analysis of Haspelmath (1997, par. 5.6), but we leave that comparison for future research.

[^25]:    ${ }^{8}$ Here and elsewhere, we observe evidence of a finer-grained semantic space for PEOPLE than for THINGS, in line with typological observations such as Silverstein (1976): The distribution of functions given each ontological category is more spread out for PEOPLE than for THINGS (Tab. 2), and decreasing the number of clusters to 3 or 2 deteriorates the Rand score less for THINGS than for PEOPLE (Tab. 3). We note that usage data reveals distinctions that remain obscured when glossing over ontological categories.

[^26]:    ${ }^{1}$ Inspecting the oriented items with inanimate objects versus living beings indicated no differences in how often the intrinsic FoR was applied. The two types of items were therefore pooled.

[^27]:    ${ }^{1}$ A conceptual replication was defined as a study that did not introduce an additional, non TP-based cue, and did not differ from the original study protocol in a significant way. For example, studies that included a priming phase pre-familiarization, or a test phase involving carrier phrases were not included. See Github repository for a full list of included papers and the subset of conceptual replications.
    ${ }^{2}$ https://github.com/christinabergmann/StatLearnDB
    ${ }^{3}$ Given that infant looking-time studies generally accept either familiarity or novelty preferences, one might argue that we should instead use the absolute value of looking-time difference as dependent measure. Indeed, in the studies reported here that pit statistical learning against other cues, a switch in looking-time

[^28]:    preference is explicitly predicted. We address these cues and their impact in the Complete Literature section of the paper. We would also like to address the general idea of absolute values in metaanalysis, and point to why this method may not be appropriate: 1) Theories of infant cognition and language acquisition have long sought to motivate the direction of looking-time preference; metaanalysis offers the potential power to test those theories and generate new possibilities when the theory is found to be inadequate. 2) Two opposing outcomes should reflect two underlying effects. Using raw effect sizes and testing the value of proposed moderators is a much more powerful use of metaanalytic techniques. Furthermore, it is important to recognize that allowing for two opposing outcomes, without the ability to predict those outcomes, increases the risk for false positives and might violate basic assumptions of sampling and null hypothesis significance testing.
    ${ }^{4}$ To assess the impact of this imputation, we re-ran our analysis with imputations based on varying means and verified that our conclusions about key findings do not change.

[^29]:    ${ }^{1}$ We use the SpaCy python library, available at https://spacy.io

[^30]:    ${ }^{2}$ We use the PyEnchant python library, available at http://pythonhosted.org/pyenchant/.

[^31]:    ${ }^{3}$ Note that all of the sentence pairs in SNLI were generated by providing subjects with a caption for an unseen image and asking them to produce a further caption that is either true, false, or maybe true of the image. So all of the sentences in SNLI can be described as image captions. The point of using this caption-based strategy in the construction of the dataset is to eliminate co-reference ambiguities that make it difficult to determine the appropriate inferential relationship between two sentences. See Bowman et al. (2015) for more details.
    ${ }^{4}$ Two of the main captions had no associated contradictions in SNLI, so subjects in the contradiction condition only rated 18 captions.

[^32]:    ${ }^{1}$ In this specific case, we assume the parameters (i.e., causal strength $w_{S}$, expected length of delays $\mu$, and delay variability $\alpha$ ) to be known which is consistent with the setup of the experiment.

[^33]:    ${ }^{2}$ Where necessary, we ruled out paths that implied an implausibly high number of failed connections, or extreme cause-effect delays, until the number of possible paths fell below 100,000 .

[^34]:    ${ }^{3}$ Try the task https://www.ucl.ac.uk/lagnado-lab/el/it or watch a trial https://www.ucl.ac.uk/lagnado-lab/el/itv.

[^35]:    ${ }^{1}$ For a more complete discussion of the flexibility of adjective ordering in French, including features which predict whether an adjective can be pre-nominal, see Fox and Thuilier (2012).

[^36]:    ${ }^{2}$ Note that the standard error on estimates for coefficients encoding contrasts between the French-like condition and the postnominal harmonic condition is very high. This is likely due to individual variation in the French-like condition, where some participants strongly regularized the input order (similar to behavior in the post-nominal harmonic condition), while others shifted away from the input dramatically. This is illustrated in Figure 4.

[^37]:    ${ }^{3}$ The numeral 'one' in French, un(e), corresponds to the indefinite article ' $a$ ', and therefore may be acquired differently or much earlier than other numerals.

[^38]:    ${ }^{1}$ Another common name for the Valence dimension is Pleasure (PAD). Our choice of terminology (VAD) follows the more recent stimulus sets we use here (Warriner, Kuperman, \& Brysbært, 2013; Ferré, Guasch, Martínez-García, Fraga, \& Hinojosa, 2016).

[^39]:    ${ }^{2}$ http://topepo.github.io/caret/index.html

[^40]:    ${ }^{3}$ https://github.com/JULIELab/EmoBank

[^41]:    ${ }^{1}$ We note that analyses in VOT space do not tell us about the context effect on the basis of individual participants' subjective perceptual ambiguity, however.

[^42]:    ${ }^{2}$ This perceptual ambiguity measure can also be computed on a by-subject basis and does not change the results.

[^43]:    ${ }^{1}$ Source code for the simulation, including a complete list of parameters and their values, is available from the first author on request.

[^44]:    ${ }^{2}$ That is, those cortical units that represent cognitive schemas.

[^45]:    ${ }^{3}$ Note also that the dopamine system is vulnerable to change in normal ageing (Cham et al., 2008).

[^46]:    ${ }^{1}$ Results from Mintz (2003) show that merging pronouns with nouns, and auxiliaries, copulas, and non-finite forms with verbs does not bias categorization results.
    ${ }^{2} X_{-} d o g \sim n$ and $X_{\_} d o g s \sim n$ are different contexts, just as light $\sim n-X$, light $\sim v_{-}$, and light $\sim a d j-X$

[^47]:    ${ }^{3}$ The bias towards nouns and verbs in categorization does not result from an imbalance in the set of target words, consisting of 40 adjectives, 47 adverbs, 76 function words, 145 nouns, and 148 verbs.

[^48]:    ${ }^{1}\left(\left(A^{2} B\right)^{N}\right)$ and $\left(\left(A^{3} B\right)^{N}\right)$ are, in full, N repetitions of AAB and AAAB , respectively. In general $X^{N}$ means the symbol or sequence X repeated N times.

[^49]:    ${ }^{1}$ Since both local and non-local antecedents are acceptable for the LD reflexive, there were no mismatch trials for this anaphor in the embedded clause sentence types discussed here.

[^50]:    ${ }^{2}$ However, the main model shows that overall participants in the Local condition were more likely to accept the LD reflexive with the local antecedent ( $\beta=0.517, \mathrm{z}=3.444, \mathrm{p}=.001$ ).

[^51]:    ${ }^{1}$ For a given model with $n$ blocks, we define the set of possible correct states as the set of unique states traversed along any path which begins at the null state, ends at the goal state, and contains $n$ states. These states can be enumerated computationally.

[^52]:    ${ }^{1}$ The model performance results (reported in a later section) were virtually the same or worse when the standard deviation instead of variance of the relative distribution was set to be proportional to the mean and/or the tradeoff parameter $\kappa$ in the static model was incorporated into $w(t)$ as a multiplicative constant to put subjective value and time weight on the same scale.
    ${ }^{2}$ Mathematical proof on the necessity of this assumption for accommodating the relevant phenomena is available upon request.

[^53]:    ${ }^{3}$ Overall BICs across participants showed the same pattern.

[^54]:    ${ }^{1}$ The amplitude of the sine waves was $.5, .5$, and a third amplitude sampled from a normal distribution with a mean of zero and a standard deviation .1(low), 2 (medium), or 3 (high).

[^55]:    ${ }^{2}$ Brightness and saturation were adjusted by a similar adjustment with a standard deviation of 15 （low）， 25 （medium），or 35 （high）．

[^56]:    ${ }^{1}$ In general, the most efficient place to move the eyes next in a rational model depends not just on visual information already obtained but also contextual information. For the present paper, we ignore contextual information for simplicity.

[^57]:    ${ }^{2}$ Note that motor error in a rational model has only random error (variance), but not systematic error (bias).

[^58]:    ${ }^{1}$ Our results may seem to invite comparison with those reported in associative priming, where the facilitation provided by forwards and backwards priming is frequently indistinguishable (Koriat, 1981; Thompson-Schill et al., 1998). However, the association norms employed in such studies are distinct from the type of association built through temporal co-occurrence patterns (Jones et al., 2006; Lund, Burgess, \& Audet, 1996), and are thus not directly comparable to our findings.

[^59]:    ${ }^{1}$ To better understand adjective-noun relations in English, we retrieved the top 50k most frequency nouns from the ukWaC, along with information about the adjectives preceding them, including: their number, average frequency, and entropy (i.e., the uncertainty over the noun's prior distribution). These results were log-transformed to approximate normality. For convenience, a base-two logarithmic transform was used for adjective-number, yielding adjective maximum entropy in bits.

[^60]:    Arnon, I. \& Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order of acquisition affects
    what gets learned. Cognition, 122(3), 292-305.
    Atkinson, Q. D., \& Gray, R. D. (2005). Curious parallels and curious connections--phylogenetic thinking in biology and
    Baayen, R. H., \& Ramscar, M. (2015). Abstraction, storage and naive discriminative learning. In Dabrowska, E., and Divjak, D. (Eds.) Handbook of Cognitive Linguistics., 99-120. Berlin: De Gruyter Mouto
    Baayen, R. H., Milin, P., \& Ramscar, M. (2015). Frequency in lexical processing. Aphasiology, 30(11), 1174-1220
    Baroni, M., Bernardini, S., Ferraresi, A., \& Zanchetta. E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3), 209-226.
    Baroni, M., \& Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. Proceedings of the 2010 Conference on Empirical Methods in Natural Language
    Processing, 1183-1193. Processing, 1183-1193.
    Beckner, C. et al. (2009). Language is a complex adaptive system: Position paper. Language Learning, 59(1), 1-26.
    Boyd, R., \& Richerson) P. J. (2005). The origin and evolution of cultures New York: Oxford University Press
    Bhristiansen, M. H., \& Chater, N. (2008). Language as shaped by the brain. Behavioral and Brain Sciences, $31(05), 489$
    509 .
    Clark, H. H., \& Wasow, T. (1998). Repeating words in spontaneous speech. Cognitive Psychology, 37, 201-242.
    Curzan, A. (2003). Gender shifts in the history of English. Cambridge: Cambridge University Press.
    Dahan, D., Swingley, D., Tanenhaus, M. K., \& Magnuson, J. S. ( 2000). Linguistic Gender and Spoken-Word Recognition
    in French. Journal of Memory and Language, 42(4), 465-480.
    Dahl, Ö. (2004). Studies in Language Companion Series : Growth and Maintenance of Linguistic Complexity. John
     Danks, J. H., \& Glucksberg, S. (1971). Psychological scaling of adjective orders. Journal of Verbal Learning and Vernal
    Behavior, 10(1), 63-67. Dawson, H. C. (2003): Defining the Outcome of Language Contact: Old English and Old Norse. OSUWPL, 57, 40-57.
    Dye, M., Milin, P., Futrell, R., \& Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J. P. Blevin
    Dye, M., Milin, P., Futrell, R., \& Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J.P. Blevins,
    \& H. Bartos (Eds.) Perspectives on Morphological Organization: Data and Analyses. Brill: Leiden.
    Engelhardt, P., Bailey, K., \& Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity?
    Journal of Memory and Language, 54(4), 554-573.
    Evans, N., \& Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences, 32(05), 429-448.
    Faaß, G., \& Eckart, K. (2013). SdeWaC - A Corpus of Parsable Sentences from the Web. Gurevych, Iryna, Chris Biemann \& Torsten Zesch (eds.): GSCL 2013, LNCS 8105. (Heidelberg: Springer).
    Fedzechkina, M., Jaeger, T.F, \& Newport, EL, (2011). Functional biases in languag order and case-marking interaction. Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX.
    Austin, IX.
    Fernald, A., Thorpe, K., \& Marchman, V.A. (2010). Blue Car, Red Car: Developing Efficiency in Online Interpretation of
    Adjective-N. Adjective-Noun Phrases. Cognitive Psychology, 60(3), 190-217.
    Ferraresi, A., Zanchetta, E., Baroni, M. \& Bernardini, S. (2008). Introducing and evaluating ukWaC
    derived corpus of English. In Proceedings of the th Web as Corpus Workshop (WAC-4). 47-54.
    Fodor, J.A. \& Lepore, E. (2002). The Compositionality Papers Oxf
    Fodor, J.A. \& Lepore, E. (2002). The Compositionality Papers. Oxford University Press.
    Genzel, D., \& Charriai, E. (2002). Entropy rate constancy in text (pp. 199-206). In Proceedings of the 40th Annual
    Meting on Association for Computational Linguistics. Association for Computanal Lingistics: Morristown Graff, D. et al. (2007). English Gigaword Third Edition LDC2007T07. Web Download. Philadelphia: Linguistic Data
    Consortium.
    Hopper, P., \& Traugott, E. (1993). Grammaticalization. Cambridge: Cambridge University Press.
    Hudson Kam, C. L., \& Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages
    Conitive Psycholog, 59(1), 30.66 . Cognitive Psychology, 59(1), $30-66$.
    Jaeger, T. F. (2010). Redundancy and red 61(1), 23-62.
    Kamp, H. \& Partee, B. (1995). Prototype theory and compositionality. Cognition, 57, 129-191
    Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory. PNAS, 105(31), 10681-
    Kemmerer, D., Weber-Fox, C., Price, K., Zdanczyk, C., \& Way, H. (2007). Big brown dog or brown big dog? An electrophysiological study of semantic constraints on prenominal adjective order. Brain and Language, 100(3), 238-
    256. Labov, S. (1972). Sociolinguistic Patterns. University of Pennsylvania Press.
    Lahav, R. (1989). Against compositionality: The case of adjectives. Philosophical Studies, 57(3), 261-279
    ambert, W. E., \& Paivio, A. (1956). The influence of noun-adjective order on learning. Canadian Journal of
    Lupyan, G., \& Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. PLoS ONE, 5(1), e8559
    Malouf, R., (2000). The order of prenominal adjectives in natural language generation. Proceedings of the 38 th Annual Meeting of the Association for Computational Linguistics, 85-92
    Maratsos, M. P. (1n D. Aaronson \& R. Rieber (eds.), Perspectives in psycholinguistics. Hillsdale, N.J.: Erbaum
    Monaghan, P., Christiansen, M. H., \& Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. Journal of Experimental Psychology: General, 140(3), 325-347.
    Pellegrino, F., Coupé, C., \& Marsico, E. (2011). A cross-language perspective on speech information rate. Language,
    $87(3), 539-558$.
    Percy, E. J., Sherman, S. J., Garcia-Marques, L., Mata, A., \& Garcia-Marques, T. (2009). Cognition and native-language grammar: The organizational role of adjective-noun word order in information representation. Psychonomic Bulletin scar, M. \& Baayen, H. (2013).
    Frontiers in Language Sciences, 4, 233.
    Ramscar, M. \& Gitcho, N. (2007) Developmental change and the nature of learning in childhood. Trends In Cognitive Science, 11(7), 274-279.
    anscar, M., Yarlett, D., Dye, M., Denny, K., \& Thorpe, K. (2010). The effects of feature-label-order and their
     name grammars and the impact of social engineering on the evolution of natural information systems. Proceedings of the 35th Meeting of the Cognitive Science Society, Berlin, Germany.
    
    Schmid, H. \& Laws, F. (2000). Estimation of Conditional Probabilities with Decision Trees and an Application to FineGrained POS Tagging. COLING 2008: Manchester, Great Britain.
    annon, C.E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423, 623-
    656.
    Simon, H. A. (1989). Cognitive Architectures and Rational Analysis: Comment. Technical Report AIP, 58, 1-25.
    Skut, W., B. Krenn, T. Brants., \& H. Uszkoreit. (1997). An annotation scheme for free word order languages. In
    Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP). Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP).
    Slobin, D. (2006). Cross-linguistic comparative approaches to language acquisition. Encyclopedia of Language \& Slobin, D. (2006). Cross-linguistic comparative approaches to language acquisition. Encyclopedia of Language \&
    Linguistics, 299-301-
    Tanenhaus M. K., Spivey-Knowlton M. J., Eberhard K. M., Sedivy J. C. (1995). Integration of visual and linguistic Tanenhaus M. K., Spivey-Knowlton M. J., Eberhard K. M., Sedivy J. C. (1
    information in mpoken language comprehension. Science 268 1632-1634.
    information in spoken anguage comprehension. Science 268 1632-1634.
    Thompson-Schill, S., Ramscar, M., \& Chrysikou, E. (2009) Cognition without control: When a little frontal lobe goes a
    long way. Current Directions in Psychological Science long way. Current Directions in Psychological Science, 8(5), 259-263.
    Tily, H., Gahi, S., Arnon, I., Snider, N., Kothari, A., \& Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. Language and Cognition, 1, (2), $147-165$.
    Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, MA: Harvard University Press.
    Trudgill, P. (2002). Sociolinguistic Variation and Change. Edinburgh University Press.
    Trudgill, P. (2010). Investigations in sociohistorical linguistics: Stories of colonisation and contact. Cambridge
    University Press.
    Vigliocco, G., Antonini, T., \& Garrett, M. F. (1997). Grammatical Gender Is on the Tip of Italian Tongues. Psychological
    Science, 8(4), 314-317.
    Vigliocco, G., Vinson, D.P., Paganelli, F., \& Dworzynski, K. (2005). Grammatical Gender Effects on Cognition: Implications for Language Learning and Language Use. Journal of Experimental Psychology: General, 134,
    Ziff, P. (1960). Semantic Analysis. Cornell University Press: Ithaca, NY.

[^61]:    ${ }^{1}$ This conclusion fits well with the finding that word frequency effects in recognition are closely related to list composition. Systematically varying the frequency of targets and foils in pure list conditions neatly illustrates this point. While a list of LF targets is similarly well-discriminated when paired with a set of HF or LF foils, foil-frequency dramatically affects discrimination of HF targets. When paired with LF foils, the error rate is close to zero; when paired with other HF foils, the error rate far exceeds that of LF targets (Underwood \& Freund, 1970). In line with this, raising the proportion of HF items on a list increases the magnitude of the WFE (Dorfman \& Glanzer, 1988; Malmberg \& Murnane, 2002).

[^62]:    ${ }^{1}$ Items that are represented in terms of the same description and configuration of variables in Table 1 differed from one another in terms of 1) the use of an instrument by the Causee, 2) for unintentional Causees, the medium of interaction between the Causer and the Causee (physical (e.g., pushing) vs non-physical (e.g., yelling loudly to startle) manipulation). The impact of these further variables has not yet been analyzed.

[^63]:    1 Intriguingly, recent data from fMRI experiments indicate that the left vOT area is also significantly associated with writing and its acquisition (Ludersdorfer, Kronbichler, \& Wimmer, 2015; Purcell, Jiang, \& Eden, 2017; Purcell, Napoliello, \& Eden, 2011; Rapp \&

[^64]:    ${ }^{1} \mathrm{We}$ interpreted the disappearance of the stimulus as a response deadline in each task. In the discussion, we note that this assumption

[^65]:    does not change core findings.

[^66]:    ${ }^{1}$ We created the composite outcome measures using two separate PCAs because we wanted the outcome measures to clearly reflect participants' responses to the original questions. For example, as shown in Table 2, pcl primarily reflects variability in judgments of urgency, risk perception, and willingness to change one's behavior, but it also loads onto the other three judgments; $p c 2$ loads positively on the first two questions (goal and solvable) and negatively on the remaining four (inevitable, urgent, risk, behavioral intentions). These patterns of weighting the original questions present some difficulty in interpreting what each factor actually reflects (e.g., $p c 2$ mostly reflects an optimistic outlook regarding our capacity to address climate change, but also, to a lesser degree, reflects the inverse of judgments of urgency, risk perception, and willingness to change one's behavior). In contrast, conducting separate principal components analyses-one on the first three questions, one on the last three questions-yields two outcome variables that clearly correspond to the original questions: tractable is tightly correlated with the three questions used to create it (goal, $r=.73$; solvable, $r=.80$; inevitable, $r=-.64$ ) and only slightly correlated with the other three questions (urgency, $r=$ .10 ; risk perception, $r=.09$; behavior, $r=.10$ ); serious is tightly correlated with the three questions used to create it (urgency, $r=$ .83; risk perception, $r=.82$; behavior, $r=.85$ ), and only slightly correlated with the other three questions (goal, $r=.18$; solvable, $r$ $=-.02$; inevitable, $r=-.11$ ).

[^67]:    ${ }^{1}$ In the general case, language can of course be visual as well as auditory, and object identification can be done through many modalities. For simplicity, we focus on audio-visual matching here.

[^68]:    ${ }^{2}$ Note that we are considering environmental noise, which is different from the noise inherent to perception.

[^69]:    ${ }^{1}$ We included plain exemplars of the critical object kinds after most children during piloting for Exp. 3 cited the pimwit's spots to explain why the speaker who had denied it was pretty was wrong.

[^70]:    ${ }^{1}$ If both the current and previous word's semantic distance had been included as factors in a single regression model, this would have greatly reduced the amount of usable data because both adjacent words would have to be content words.
    ${ }^{2}$ The displayed coefficients for current (previous) surprisal come from the regression model that includes current (previous) semantic distance. Consequently, exactly the same set of words was involved in estimating the coefficients for the surprisal and semantic distance measures, even though surprisal (unlike semantic distance) is also defined for function words.

[^71]:    ${ }^{1}$ Given these fairly fundamental disagreements, it should come as no surprise that what counts as a normative response in any number of contexts is hotly debated. For the sake of simplicity, terms like "normative," "correct," and "logical" will be used to indicate conclusions that are considered correct in classical logic and probability.

[^72]:    ${ }^{2}$ This analysis was meant to assess whether there was a rather diffuse and non-specific conflict detection signal that predicted an individual's detection on a given task by a composite of their relative effects on other tasks. Were the data not binary, a factor analysis would perhaps be appropriate here. Essentially, this was the most liberal test we could devise to check for generality of conflict detection effects. However, it is worth noting that a more conventional test, using multiple regressions with all tasks as predictors except the one being predicted, generated the exact same pattern of results, with all models being uninformative except where BR \& CON items were concerned.

[^73]:    ${ }^{1}$ Note that in RT this is the opposite: a positive sign means going with the pattern, and negative against it.

[^74]:    ${ }^{1}$ Data, materials, figures, and code are available here: https: / / github.com/tobiasgerstenberg/tower_counterfactual
    An interface to view the stimuli and play around with the different noise models may be accessed here: http://web.mit.edu/tger/ www/demos/towers/physics_interface.html

[^75]:    ${ }^{1}$ A running record was included in the test battery to obtain an online measure of decoding accuracy and was included after the experiment commenced. Thus, 13/24 participants have running record data available.

[^76]:    ${ }^{1}$ We vary the dimensionality of the memory and word embedding, the number of computational hops (accesses to the memory cache to answer a single question), the number of training and testing examples ( 1000 vs. 10000), and the size of the world from which the dataset of stories is generated ( 5 vs. 10 vs. 30 entities per entity type, which correspond to the objects, container, etc. in the story).

[^77]:    ${ }^{1}$ The code for running our experiments is available online: https://github.com/RobGrimm/CogSci2017-MultiWordUnits

[^78]:    ${ }^{2}$ The same argument can be made for adults, who are faster to recognize and produce frequent four-word MWUs in similar experiments (Arnon \& Snider, 2010). Such results also support theories of adult linguistic competence which include MWU-like constituents (O’Donnell, 2015).

[^79]:    ${ }^{3}$ Note that we do not claim that the acquisition of MWUs always precedes the acquisition of single words, but merely that this happens often enough to have a measurable impact on word learning.

[^80]:    ${ }^{4}$ Corpora names (see http://childes.talkbank.org/access/ for references): Bates, Bernstein, Bliss, Bloom70, Bloom73, Bohannon, Braunwald, Brent, Brown, Carterette, Clark, Cornell, Demetras1, Demetras2, ErvinTripp, Evans, Feldman, Garvey, Gathercole, Gleason, HSLLD, Hall, Higginson, Kuczaj,MacWhinney, McCune, McMillan, Morisset, Nelson, NewEngland, Peters, Post,Providence, Rollins, Sachs, Snow, Soderstrom, Sprott, Suppes, Tardif, Valian, VanHouten, VanKleeck, Warren, Weist
    ${ }^{5}$ Belfast, Fletcher, Manchester, Thomas, Tommerdahl, Wells, Forrester, Lara

[^81]:    ${ }^{6}$ available online: http://www.cs.toronto.edu/~jbrooke

[^82]:    ${ }^{7}$ http://www.speech.cs.cmu.edu/cgi-bin/cmudict
    ${ }^{8} \mathrm{http}: / / \mathrm{crr} . \mathrm{ug}$ nt.be/archives/1330

[^83]:    ${ }^{1}$ When a calculation error resulted in a difference of +1 or -1 compared to the correct value, problems were still considered correctly solved.

[^84]:    ${ }^{1}$ Participants did Study 2 first，followed by the clock question （Study 1）that was inserted in the middle of a questionnaire．As participants were not given any feedback，no significant influence was expected from the first task on the second one．

[^85]:    ${ }^{1}$ "Moreover, it must be confessed that perception and that which depends upon it are inexplicable on mechanical grounds, that is to say, by means of figures and motions. And supposing there were a machine, so constructed as to think, feel, and have perception, it might be conceived as

[^86]:    increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, (we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. Thus it is in a simple substance, and not in a compound or in a machine, that perception must be sought for. Further, nothing but this (namely, perceptions and their changes) can be found in a simple substance. It is also in this alone that all the internal activities of simple substances can consist" (Leibniz 1714, section 17).

[^87]:    ${ }^{1}$ Data are available at https://cocolab.stanford.edu/ datasets/tangrams.html

[^88]:    ${ }^{2}$ Specifically, we used the Universal Dependencies tags csubj, ccomp, xcomp, and advcl for subordinate clauses and acl for adjectival clauses (Schuster \& Manning, 2016)

[^89]:    ${ }^{3}$ There is a broader debate over the timescales at which lexicons and lexicon learning mechanisms operate; here, we assume a discourse-level structure to the lexicon, where there is uncertainty over how words are used in the given conversation. See Frank, Goodman, \& Tenenbaum (2009) for a related approach at the scale of cross-situational word learning.
    ${ }^{4}$ All results can be reproduced running our code in the browser at http://forestdb.org/models/conventions.html
    ${ }^{5}$ In our implementation, we enumerate over coarse-grained bins; preliminary experiments using variational inference on the full continuous distribution give similar results.

[^90]:    ${ }^{6}$ In our simulations, we used $\alpha=10$ and found the basic reduction effect over a range of different biases

[^91]:    ${ }^{1}$ Planned sample size, exclusion criteria, and analysis plan preregistered at https://osf.io/y7mvt

[^92]:    ${ }^{2}$ https://github.com/emilyfae/socref_uncert

[^93]:    ${ }^{3}$ Planned sample size, exclusion criteria, and analysis plan preregistered at https://osf.io/y7mvt/.

[^94]:    ${ }^{1}$ A recent exception is (Moss, 2015).

[^95]:    ${ }^{2}$ For convenience, the prior distribution over states is assumed to be a symmetric betabinomial distribution between 0 and 10 with shape parameters $\alpha=\beta$ free in the model.

[^96]:    ${ }^{3} \mathrm{We}$ assume $\theta_{\text {probably }}=\theta_{\text {likely }}$ and $\theta_{\text {might }}=\theta_{\text {possible }}$. The threshold of the remaining modifier $\theta_{\text {certainly }}$ is free in the model.
    ${ }^{4}$ The delta function $\delta_{s \in \llbracket m \rrbracket}$ gives 1 as output if the state $s$ belongs to the meaning of $s, 0$ otherwise. The expression rat.bel $(. \mid o, a)$ refers to the belief distribution over states induced in a rational agent by the observation of $o$ red balls out of $a$.

[^97]:    ${ }^{5}$ Goodman and Stuhlmüller (2013) use Kullback-Leibler divergence as a measure of discrepancy between speaker and listener beliefs. We found Hellinger distance a more adequate measure in the present setting because utilities in terms of KL-divergence lead to speakers who will never use messages that are semantically false, whereas HD allows messages to be send if they are "true enough." The Hellinger distance between two discrete distributions $P$ and $Q$ is defined as $\operatorname{HD}(P, Q)=\frac{1}{\sqrt{2}} \sqrt{\sum_{i}\left(\sqrt{P_{i}}-\sqrt{Q_{i}}\right)^{2}}$.
    ${ }^{6}$ As $\lambda \rightarrow \infty$, the choice approaches perfect rationality.

[^98]:    ${ }^{7}$ The possible choices included not only likely but also probable in embedded position. However, having not found interesting differences in the behavior of these two expressions, the results reported in this paper, the visualization in Figure 5 and the model evaluation are all based on data in which the counts of participants' choices of messages containing probable and likely have been aggregated.

[^99]:    ${ }^{8}$ The other parameters of interest are $\alpha$ : mean $=6.373$, HDI: $5.546-7.178$; and $\lambda:$ mean $=5.429$, HDI: 5.192-5.659.

[^100]:    ${ }^{1}$ While other variations of this category structure were used to counterbalance features and category locations, they preserved this same main category structure and so we discuss the experiment in terms of this one. We do this for the other experiments, as well.

[^101]:    ${ }^{1}$ The encoding duration of the distractor is an arbitrary value, as we did not know the exact duration of the attentional capture of the reading task of non-words, but we were interested on the relative effect of an additional free time and not on the absolute effect of the free time in this study.

[^102]:    ${ }^{2}$ All the results discussed below are represented by the solid lines in all panels B, which correspond to the simulation of the model with the default value $(r=1.5)$ of the removal strength parameter. The dashed lines will be discussed latter.

[^103]:    ${ }^{1}$ The constraint that words always have four CVCV segments is a simplification for convenience. In principle, our model should be applicable to any set of phonemes and syllable structures. See Pajak et al. (2016) for the complete list of phonemes and stimuli used.

[^104]:    ${ }^{2}$ In particular, the distance between dissimilar phonemes in feature space is large because their means differ in multiple units across multiple dimensions. The distance between highly-similar phonemes on the other hand is small since they are only one unit apart along a single dimension.

[^105]:    ${ }^{3}$ Depending on how model fits are quantified it may also be possible to fit the discrimination results very well but overestimate accuracy on the word learning task. Critically, however, discrimination accuracy will always be lower than word learning accuracy in the $M_{-} A_{-}$model.

[^106]:    1 Hereafter, we call verbal probability phrases verbal probabilities.

[^107]:    *(the treatment will be helpful in that case.)

[^108]:    ${ }^{2}$ In the following analyses, the ratings were mapped onto 0-1 scale.

[^109]:    ${ }^{3}$ In this analysis, we used density values for 99 probabilities ( $1 \%, 2 \%, 3 \%, \ldots, 97 \%, 98 \%$, and $99 \%$ ).

[^110]:    1/https://github.com/jasbi/cogsci2017

[^111]:    ${ }^{2}$ For the data, full results including non-critical trials, and statistical analyses visit the paper's online repository.

[^112]:    ${ }^{3}$ We used a tight prior in this case to decrease posterior correlations between cutpoints and intercept.

[^113]:    ${ }^{1}$ Although many students in this age range have poor spelling, coders did not have trouble determining what a given child had written, even when words were misspelled (e.g., "the toltal", "write the anser next"). Inter-rater agreement between coders on whether a given definition was relational ranged from 95-100\%.

[^114]:    ${ }^{1}$ The model and data analysis source code are available at https://github.com/ctn-archive/kajic-cogsci2017.

[^115]:    ${ }^{2}$ We use italics to refer to the name of a population of neurons or the vector that is represented by that population, which is to be inferred from the context. The bold font is used to refer to labels assigned to vectors representing a word. For example, cue $\cdot$ animal refers to the dot product between the vector represented by the population of neurons labeled "cue" and the vector corresponding to the word "animal".

[^116]:    ${ }^{3}$ See Troyer et al. (1997) for more detailed description of the categorization procedure.

[^117]:    ${ }^{1}$ Which object (the blue fruit or the red stalk) appeared more frequently with which frame, as well as which object was paired with which long name, were both counterbalanced between participants, giving a total of 4 possible object-frame-name pairings which a participant might be trained on. This ensured that potential factors such as sound symbolism, or higher saliency or learnability of any specific object-word pairing, could not systematically bias our results.

[^118]:    ${ }^{1}$ Children also participated in a homogeneous condition, in which all distractors are identical (but distinct from the target). This condition is designed to require less SSA because target tracking is supported exogenously by increased salience of the target. We did not analyze this condition as distinguishing exogenous vs. endogenous effects was not of interest for this study.

[^119]:    ${ }^{2}$ We omitted difficulty level 3 , as only one age group (7 year olds) completed it.

[^120]:    Alex and Josh are both employees at a large company. Their coworker Max has been asked to decide how to assign bonuses to Alex and Josh. Due to company policy, Max can either: give $\$ 1,000$ to one employee and $\$ 100$ to the other or give [ $\$ 0$ / \$100 / \$500 / \$1000 / \$1,100] to both. Alex and Josh currently make the same amount each year, do the same job, [and have received identical work evaluations / but Alex has received a better work evaluation].
    Participant group 1: What would you do? (Give Alex the \$1,000 bonus and Josh the $\$ 100$ bonus / Give Josh the \$1,000

[^121]:    ${ }^{1}$ Without exclusion criteria, $\left(\chi^{2}=1.87, p=0.17\right)$

[^122]:    ${ }^{1}$ This is because the variance estimator we used is statistically unbiased, see eq. 2 .

[^123]:    ${ }^{2}$ Similar results hold if the mean is unknown, but the formulas are much more complicated.
    ${ }^{3}$ The prior hyperparameters were chosen to match the setup of the example discussed above $(N(0,1)$ payoff distribution and $t=$

[^124]:    15). The interpretation of the parameters is that the prior is based on a sample of size $2 \alpha=15$ with variance $\beta / \alpha=1$.

[^125]:    ${ }^{1}$ This analysis is equivalent to a chi-squared test of independence, but it accounts for individual heterogeneity in the data (see Willemsen \& Johnson, 2011).

[^126]:    ${ }^{2}$ All pairwise contrasts were significant, $p<.05$, apart from the contrast between probability and delay regarding the last item, and the contrast between amount and probability regarding intermediate items.

[^127]:    ${ }^{3}$ This is the case in our experiment. In fact, it is twice as likely for an alternative-wise transition to occur (6 transitions) compared to a dimension-wise ( 3 transitions).

[^128]:    1A third condition with misleading descriptions called the conflicting descriptions experience condition has not been included here as it was not relevant to our current investigation.

[^129]:    2 In cases where the sphericity assumption has been violated, the Greenhouse-Geisser correction to $d f$ has been used.

[^130]:    ${ }^{1}$ A programming error meant that all participants in Experiment 1 completed the brightness-relevant task.

[^131]:    ${ }^{2}$ For the correlated condition, this provides an index of the bypass strategy (Dyson \& Quinlan, 2010). The bypass strategy describes a strategy whereby participants monitor only the trial-by-trial sequences making the same response as on the previous trial when the stimulus is the same as the previous trial and switching responses when the stimulus changes.

[^132]:    ${ }^{1}$ We report the mean and the $95 \%$ highest density interval (HDI) of the posterior distributions for each parameter. The HDI represents the range of credible values given the model specification and the data. We chose not to interpret the DDM fits for the Bullseye/Face tasks since there was no suggestion of any non-guessing signal.

[^133]:    ${ }^{1}$ It might seem surprising that any children failed to remember whether the Other child was older or younger. However, recall that the photos of the Other child used in the study reflected a range of ethnicities. Anecdotally (given the small number of excluded children) children who missed this question tended to miscategorize the age of a photo of a child of another race than themselves, possibly reflecting an own-race bias in processing faces (Anastasi \& Rhodes, 2005).

[^134]:    ${ }^{1}$ All family trees, feature matrices (and code) can be found at https://github.com/MollicaF/LogicalWordLearning

[^135]:    ${ }^{2}$ This may be due to data sparsity for UnCLE in the trees. As UNCLE is the most complex concept learned here, it may be that UNCLE requires more unique data points to be learned. Under our Zipfian data sampling, the model receives data for less than half of the unique uncles in the trees. When you relax the sampling assumption to uniform, the model does learn UNCLE and having the correct hypothesis in the space alters the time scale of the shift (to around 30 data points).

[^136]:    ${ }^{3}$ This pattern holds if the data distribution is uniform or becomes more peaked-i.e., a Zipfian exponent of 0,1 or 2.

[^137]:    ${ }^{1}$ We assumed each word conveys on average 10 bits, which is

[^138]:    true of our stimuli, and an average sentence length of 10 words.

[^139]:    ${ }^{2}$ All code and data are available at:

[^140]:    ${ }^{3}$ We set our presentation rate at $147 \mathrm{~ms} /$ word even though previous tasks have gone as fast as $84 \mathrm{~ms} /$ word because we were concerned about enervating participants after one story.

[^141]:    ${ }^{4}$ Analyses excluding their data do not significantly alter the results we present here.
    ${ }^{5}$ We did not have an exclusionary cut-off for participant accuracy because we designed our questions to be difficult and inspire engagement with the text.

[^142]:    ${ }^{6}$ Although, it is clear that there are several alternative proposals to be explored in the future including parallel processing and retrieval accounts (e.g., Lewis \& Vasishth, 2005).

[^143]:    ${ }^{7}$ Average reading time $516 \mathrm{~ms}(\mathrm{SEM}=2.25)$ and average bits per five word chunk 50 bits ( $\mathrm{SEM}=0.5$ ) give an information processing rate around $100 \mathrm{bits} / \mathrm{sec}$.

[^144]:    ${ }^{8}$ The optimal preparation model (Smith \& Levy, 2008) also suggests this decoupling, although the predictions are less clear.

[^145]:    ${ }^{1}$ Multitasking can, in some situations, be achieved by rapid sequential processing (e.g., switching between asynchronous serial processes, as is common in computers), rather than through true synchronous processing. Here, our focus is on forms of multitasking that reflect truly concurrent processing, sometimes referred to as perfect timesharing or pure parallelism.

[^146]:    ${ }^{2}$ Note that it is ensured that, for the uniform distribution $U[0,1]$ of stimulus unit activations in the task-relevant set of input units, every relevant output unit is equally likely to be required for execution.

[^147]:    ${ }^{3}$ All reported results were obtained using gradient decent to minimize the MSE of each training pattern. However, we observed the same qualitative effects using the cross-entropy loss function.

[^148]:    ${ }^{1}$ Holyoak and colleagues (Holyoak et al., 2010; Lee \& Holyoak, 2008) at points extend this claim even further, suggesting that SMT establishes analogical inferences "solely on the logical form of representations and not on their meaning" (Lee \& Holyoak, 2008 p 1120). This however would suggest that SMT is insensitive to the content of the higher-order relations that bind predicates. This is incorrect. SMT distinguishes higher-order constraining relations that confer systematicity (such as cause and prevent) from nonconstraining relations (such as and), which do not. (Falkenhainer, Forbus, \& Gentner, 1989; Gentner, 1983). Here we focus on the aspects of their argument that would prove challenging for the theory.

[^149]:    ${ }^{2}$ One alternative way participants could implement a global strategy would be to generate a posterior probability for the system of base relations, and either increase or decrease their estimate contingent upon which cause is absent in the target. For example, if participants are given $G_{1} P_{1} P_{2}$ as a base and $G_{1} P_{1}$ as a target, dropping a preventative relation (e.g., $P_{2}$ ) may lead them to boost their probability estimate. However, while this is certainly possible, we

[^150]:    believe that such an explanation is far less parsimonious than an account based on higher-order relational mapping (i.e., $G_{1}>P_{1} P_{2}$ ).

[^151]:    ${ }^{1}$ CRT is correlated with cognitive ability (Frederick, 2005; Toplak, West, \& Stanovich, 2011), but can still predict rational thinking and performance on heuristics and biases tasks after controlling for the variance associated with assessments of intelligence, thinking dispositions, executive functions and cognitive skills. Thus, people who score highly on the CRT can be categorized as being more likely to engage in rational, analytic thinking (Shah, Michal, Ibrahim, Rhodes, \& Rodriguez, 2017).

[^152]:    ${ }^{2}$ See https://www.youtube.com/watch?v=n2gRewFVssY
    ${ }^{3}$ See https://www.youtube.com/watch?v=d56jnzPwBOU
    ${ }^{4}$ See https://www.youtube.com/watch?v=ALris6B4yj4

[^153]:    ${ }^{5}$ A possible drawback of normalized scores (such as the NSPM) is that they are prone to overestimations. An extreme example can illustrate this situation: If a participant assigned a preference of 2 to the mirror-alternative and 1 to the shift-alternative, her NSPM score would be $66.6 \%$, thus having a strong additive effect in computing averages, even though the mirror-alternative was ranked as negligible. This can be controlled by using geometric means: when a score is a ratio such as $X_{i} / Y_{i}$, the geometric mean is the only mean with the property $\mathrm{GM}\left(\mathrm{X}_{\mathrm{i}} / \mathrm{Y}_{\mathrm{i}}\right)=\mathrm{GM}\left(\mathrm{X}_{\mathrm{i}}\right) / \mathrm{GM}\left(\mathrm{Y}_{\mathrm{i}}\right)$ i.e. it "normalizes" the ranges being averaged in such a way that no range dominates the weighting (Fleming \& Wallace, 1986).

[^154]:    ${ }^{1}$ We use the term referent to denote anything referred to by a word - an object or event, or set of semantic properties (e.g., $\{$ INDEFINITE, SINGULAR $\}$ for $a n$ ).

[^155]:    ${ }^{1}$ The pre-trained Word2Vec model is available at https://code.google.com/archive/p/word2vec/; The pre-trained GloVe model is available at http:// nlp.stanford.edu/projects/glove/

[^156]:    ${ }^{1}$ The authors are mentioned in alphabetical order.
    ${ }^{2} \mathrm{We}$ are using the classical abbreviations.

[^157]:    ${ }^{3}$ http://mentalmodels.princeton.edu/models/mreasoner/

[^158]:    ${ }^{4} A \leftarrow \perp$ is called an assumption because it can be overwritten under the Weak Completion Semantics, as we will discuss later.

[^159]:    ${ }^{5}$ If $\mathcal{P}=\{A \leftarrow \perp, A \leftarrow \top\}$ then $w c \mathcal{P}=\{A \leftrightarrow \perp \vee \top\}$. This is semantically equivalent to $w c \mathcal{P}=\{A \leftrightarrow \top\} . A \leftarrow \perp$ is overwritten.

[^160]:    ${ }^{6}$ The wheels of a lawnmower are wet if the gras is wet; the wheels are wet if the sprinkler is on; the gras is wet if it is raining.

[^161]:    ${ }^{7}$ We remove the fact from the program that generated the observation, because otherwise the explanation would be empty.

[^162]:    ${ }^{1}$ In Dutch, verb/noun distinctions are differentiated through affixes to the root. For example, telefoneren (to phone) is a verb and telefoon (telephone) is a noun. The English translations do not reflect that participants responded with a single word.

[^163]:    ${ }^{1}$ A task/process/input-output mapping is defined as a unique mapping from all possible vectors in the input subspace to corresponding vectors in the output subspace, that is independent of the mappings for all other combinations of input and output components in the network.

[^164]:    ${ }^{2}$ Although we are restricting ourselves to logistic functions with unit steepness, in a more general setting one can use the steepness as another design parameter.

[^165]:    ${ }^{3}$ Simulation details are omitted due to space constraints.

[^166]:    ${ }^{1}$ All the results follow straightforwardly for structured messages as long as there is a deterministic linearization, such as reverse Polish notation for tree-structured messages.

[^167]:    ${ }^{2}$ I here consider only discrete signals and messages. The continuous case requires either a limit on the power of the signal or for source and channel coding to be considered simultaneously.

[^168]:    ${ }^{3}$ Composition operations that involve copying, such as Suffixaufnahme in Old Georgian (Michaelis \& Kracht, 1996), present an interesting wrinkle. If they can be handled by introducing an integer coefficient for each $l_{m_{i}}$, the ultimate independence result of this section still holds. In any case, they make the signal longer, so they should not present a more efficient bound than Equation 7 .

[^169]:    ${ }^{4}$ I obtained similar results when using $b=6,780$, the number of distinct characters.

[^170]:    ${ }^{5}$ As the CEDict pronouncing dictionary has an average length of about 3.1 non-tone pinyin letters, or 2.8 phonemes, per character type, the optimization rate of Mandarin is similar to the others.

[^171]:    ${ }^{1}$ We thank Kevin T. Kelly for pointing us to the Ellsberg paradox.

[^172]:    ${ }^{1}$ In 2016 a PewResearch poll found the majority ( $62 \%$ ) of US adults get their news through social media. Source: http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

[^173]:    ${ }^{2}$ Further simulations in which total network size has been varied, but density has been kept constant at $1 \%$ (i.e. 10 agent-links for 1000 agents, 50 links for 5000 agents) have shown clustering effects remain constant (i.e. depend on relational, not absolute links / network size).
    ${ }^{3}$ Mathematically, P(Declaration) starts to have an impact when it effectively reduces the average number of "functional" links to a point below the absolute threshold for a singular cohesive network (i.e. if it reduces the average number of active links below 4 in the present model; left-hand panel of Fig. 3).
    ${ }^{4}$ The present model demonstrates this with fixed, neutral (0.5) priors for all agents. If variance in priors is included, such that $S D$

[^174]:    ${ }^{1}$ this problem is similar, but not identical, to the problem of reverse inference (Poldrack, 2006)
    ${ }^{2}$ Whereas a minor degree of systematicity does seem to exist

[^175]:    ${ }^{3}$ What makes representational codes different is a surprisingly difficult question to answer. Due to space

[^176]:    ${ }^{4}$ in fact, because of that linear one-to-one relationship, replicating our simulation with these two examples leads to

[^177]:    ${ }^{1}$ We thank Xiaonan Liu for pointing us to these representations

[^178]:    ${ }^{2}$ i.e., trials with the same target regardless of distractors

[^179]:    ${ }^{1}$ Because each of the models we examined is parameter-free, metrics that take into account model complexity such as the Bayesian Information Criterion would give identical results to those only taking into account likelihood.

[^180]:    ${ }^{1}$ Admittedly, this could well be considered a bald-faced lie.

[^181]:    ${ }^{2}$ These exclusions were of participants who failed to select the Helpful message in a Control condition (where the goal was to help) on at least $40 \%$ of trials. We also excluded those who selected the Helpful message in the Low Suspicion condition on $40 \%$ or more trials (where the goal was to hinder, and double bluffing was unreasonable). There was no difference in the significance of our findings if these people were included..

[^182]:    ${ }^{1}$ Unless speech is assumed to be a normally distributed IID process without temporal contiguity, a larger F0 range does not guarantee lower temporal predictability (c.f., e.g., a sine wave).

[^183]:    ${ }^{1}$ The model may also account for other asymmetries in speech production. We plan to pursue this in future work.

[^184]:    ${ }^{1}$ Other experimental manipulations in the original corpus were not the focus of our analysis.
    ${ }^{2}$ To address the within-subjects nature of the data, we removed variance due to teams in a secondary analysis (summarized below)

[^185]:    and found a similar pattern of results. We did not test for order effects or learning, but if learning has occurred it would increase the performance variability in the dataset.
    ${ }^{3}$ Fusaroli \& Tylén's lexical choices

[^186]:    ${ }^{1}$ In counterfactual terms: Had a human seen a green-striped elephant, s/he would have yet recognized it as an elephant. Geoffrey Hinton once told a similar story about a pink elephant!

[^187]:    ${ }^{2}$ Fahlman and Lebiere (1989) also suggest linear, Gaussian, and asymmetric sigmoidal (with range 0 to +1 ) activation functions as alternatives. Our proposed framework can be straightforwardly adapted to handle all such activation functions.

[^188]:    ${ }^{3}$ More precisely, it has been shown how the continuous-time version of MAL, Langevin dynamics, can be implemented in a neurally-plausible manner. But note that MAL amounts to sampling from the underlying Langevin dynamics.
    ${ }^{4}$ Formally, $f\left(\cdot ; W^{*}\right): \prod_{i=1}^{n} D_{i} \rightarrow \prod_{j=1}^{m} R_{j}$ where $D_{i}$ and $R_{j}$ denote the set of values that input unit $i$ and output unit $j$ can take on, respectively.
    ${ }^{5}$ In counterfactual terms, this is equivalent to saying: Had input instance $X$ been presented to the network, it would have classified $X$ in class $j$.

[^189]:    ${ }^{6}$ More specifically, $Z$ renders the computation of the gradient of the log-likelihood for those models intractable.
    ${ }^{7}$ The MAL algorithm inherits this property from the MetropolisHasting algorithm, which it uses as a subroutine.

[^190]:    ${ }^{8}$ Due to the inherent randomness in CCNN construction, training could lead to networks with different structures. However, since in this work we are solely concerned with generating examples using

[^191]:    CCNNs rather than how well CCNNs could learn a given discriminitive task, we arbitrarily pick a learned network. Note that our proposed framework can handle CCNNs with arbitrary structures; in that light, the choice of network is without loss of generality.
    ${ }^{9}$ Yet, too large a $\beta$ is not good either, leading to a sparse and coarse-grained exploration of the input space. Some measures have been proposed in computational statistics for properly choosing $\tau$; cf. (Roberts \& Rosenthal, 1998).

[^192]:    ${ }^{1}$ The use of deontic logic for normative reasoning is the subject of active debate. Although further discussion of this debate is outside the scope of this paper, we note that our proposed approach does not require using deontic operators. We can still reason about norms and learn them using the schema described in Definitions 2 and 3. We would simply need to replace the deontic operators and modify the predicates slightly. Norm $\mathcal{N}_{5}$ in example 1 would become: $\mathcal{N}_{5}:=[0.3,0.6]::$ in (library, $\left.X\right) \Longrightarrow$ forbidden $(X$, talking $)$

[^193]:    ${ }^{2}$ We ensured that the imported actions were physically plausible in the given scene/context.
    ${ }^{3}$ We focus here on permissions. Prescriptions and prohibitions show very similar patterns overall, but prohibitions differ from the other two norm types in interesting ways (e.g., less consensus, slower activation) that will be treated in a separate investigation.

[^194]:    ${ }^{1}$ For instance, a subject's indifference point at interval 1 is 0.8 (see Figure 4). Given a manipulation point of -0.3 , the respective manipulated trial must yield a subjective value ratio (SS/LL) of 0.5 at an interval of 1 . The same logic applies over all manipulation points and intervals.

[^195]:    2 The fitting of the logistic regression model was performed using the StixBox mathematical toolbox by Anders Holtsberg (http://www.maths.lth.se/matstat/stixbox/). The fit was based on the model $\log \left[\frac{p}{1-p}\right]=X b$, where $p$ is the probability that the choice is 1 (SS) and not $0(\mathrm{LL}), X$ represents value differences, and $b$ represents the point estimates for the logistic function.

[^196]:    ${ }^{3}$ The fit of the hyperbolic function was based on minimizing the summed squared errors (SSE). $R^{2}$ is defined as the ratio of the sum of squares of the regression ( $S S R$ ) and the total sum of squares (SST). Since $S S T$ is defined as $S S R+S S E, R^{2}$ is defines by $1-$ SSE/SST.

[^197]:    ${ }^{1}$ We scored the top 30 male and top 30 female first names between 1911 and 2010 in Germany (Bielefeld, 2016) on a scale from 30 (for the most popular male/female name in a year) to zero (for names not listed during a year). These scores were then weighted, for each gender separately, by the proportion of people in the German population who belong to the cohort (Statistisches Bundesamt, 2014). We selected the most popular male and female name in each decade based on the summed raw scores each name received across these ten-year periods. In addition to these 20 most popular names from each decade, we selected the most frequent male and female name in the population (that was not already in the list) based on the total sum of the weighted scores across all years. Finally, the 22 selected names were ranked based on the sum of their weighted scores across all years.

[^198]:    ${ }^{2}$ For few participants, this resulted in the mixture collapsing on the SCM. For these participants, we used a prior that assigned low initial weight to the SCM (e.g., .001) and equal weight to the other two strategies. To ensure unbiased estimation of latent groupmembership, these unequal priors were taken into account in the calculation of membership probabilities.
    ${ }^{3}$ Yet children's inferences were well calibrated to their cohort's instances. Evaluating inferences based on a ranking derived from children's reported number of instances, flips the accuracy pattern such that children significantly outperform adults, $t(78)=-2.40$, $p=.019, d=-.536, B F_{10}=2.70$.

[^199]:    ${ }^{1}$ English does have pronominal gender, but not a full-fledged grammatical gender system.

[^200]:    *These two authors contributed equally.

[^201]:    ${ }^{1}$ Since no comparisons were produced before session 4 (26 months), graphs \& analyses focus on sessions 5-12 (26-58 months).

[^202]:    ${ }^{2}$ Utterances using the word 'favorite' were not coded, since it was not clear that children understood its meaning as comparative.

[^203]:    1 To make the saturated model identifiable, we constrained $\sigma^{2}$ to be the same for both latent classes.

[^204]:    ${ }^{\dagger}$ These authors contributed equally.

[^205]:    a. Masha didn't run $[\mathrm{QUICKLY}]_{\mathrm{F}}$.
    b. $[\mathrm{MASHA}]_{\mathrm{F}}$ didn't run quickly.

[^206]:    ${ }^{1}$ Model-generated probabilities are mapped onto a 7-point scale with the following formula: RATING $=1+6 * P(\phi \mid M)$.
    ${ }^{2}$ The Python code used to implement the model can be downloaded at https://github.com/jonscottstevens/Prosody-Projection.

[^207]:    ${ }^{1}$ Theoretical work in semantics has instead focused on how in-

[^208]:    formation from a comparison class is used and what representations might be preferred (Bale, 2011; Solt, 2009).

[^209]:    ${ }^{2}$ Corpus accessed via https://corpora.linguistik. uni-erlangen.de/cgi-bin/demos/Web1T5/Web1T5_freq. perl. Due to potential polysemy and idiosyncracies of our expermental materials (Table 1), we made the following substitutions when querying the database for emprical frequency: produce $\rightarrow$ "fruits and vegetables"; things you watch online $\rightarrow$ "online videos"; days in $\{$ season $\} \rightarrow$ " $\{$ season $\}$ days"; dishwashers $\rightarrow$ "dishwashing machines"; videos of cute animals $\rightarrow$ "animal videos".

[^210]:    ${ }^{3}$ This was the maximal mixed-effects structure that converged.

[^211]:    ${ }^{1}$ The addition of random intercepts and slopes involving Items (Egg, Rose, etc) did not improve the fit of these models.

[^212]:    ${ }^{2}$ While we did not assess children's emergent literacy skills here, ongoing studies are employing parent surveys to do so.

[^213]:    ${ }^{1}$ Due to limited space, we do not report main effects of or interactions with Channel. Main effects of and interactions with

[^214]:    ${ }^{2}$ More fine-grained analyses of HP sentences, however, do suggest a relationship between knowledge and late positivities that is mediated by offline, knowledge-based Cloze measures. For lower-

[^215]:    Cloze (and, by inference, less-frequently-accessed) compared to higher-Cloze items, post-N400 activity was more positive-going, but only for high-knowledge individuals.

[^216]:    ${ }^{1} \rho^{*}=$ Spearman-Brown adjusted correlation, a measure of splithalf reliability

[^217]:    ${ }^{1}$ It is suggested that the simultaneous processing of two competing samples emulates deliberation over two multiattribute options. This analogy works under the assumtpion that, in real-life multiattribute choices and on each moment, one attribute is considered and the corresponding attribute values serve as inputs into a preference formation process (Tsetsos et al., 2012)

[^218]:    ${ }^{1}$ The corpus in this work is about 25 times larger.

[^219]:    ${ }^{2}$ In fact, the number of possible syntactic derivations is constrained by a very large beam.

[^220]:    ${ }^{3} 5$-gram surprisal predicts conditional frequency effects based on $n$-gram co-occurrence counts. Previous work has shown that 5 -gram frequency controls are sufficiently able to control for frequency effects that syntactic frequency controls are sometimes unable to predict reading times over them (van Schijndel \& Schuler, 2016), so 5 -grams create a strong baseline with which to test other frequency influences.

[^221]:    ${ }^{4} \mathrm{An}$ alternative to the approach taken in this paper would be to maintain a constant vocabulary size but to train the conditional probabilities of that vocabulary over a much larger training set. Such an approach would only help if the weakness of lexical entropy is due to poor probability estimates rather than to unknown words.

[^222]:    ${ }^{1}$ Translating research on norms into predictions about structural reasoning is not straightforward. First, moral norms carry deontic content, which distinguishes them from other kinds of structural constraints (such as a wage gap) that do not. Second, categoryspecific norms can be interpreted in either essentialist or structural terms (e.g., if girls are not allowed to go out after 9 pm , this could stem from inherent characteristics of girls, or structural forces). Existing studies about norms have not made these distinctions, complicating their interpretation with regard to structural reasoning.

[^223]:    ${ }^{1}$ In order to compute power, we need to have an estimate of the true effect, the sample size, and an estimate of the standard deviation.

[^224]:    ${ }^{2}$ The presentation below generalizes to the two-sided test.

[^225]:    ${ }^{3}$ This can be confirmed by running the following command using R ( R Core Team, 2014): power.t.test(delta=0.1,sd=1, $n=36$, alternative $=$ "one.sided", type="one.sample").
    ${ }^{4}$ Treating lower p-values as furnishing more evidence against the null hypothesis reflects a misunderstanding about the meaning of the p-value; given a continuous dependent measure, when the null hypothesis that $\mu=0$ is true, under repeated sampling the p -value has a uniform distribution (see proof in the Appendix). This has the consequence that, when the null is true, a p-value near 0 is no more surprising than a p-value near 0.05 .

[^226]:    ${ }^{5}$ Note that this range is an error; power cannot be less than $5 \%$ if Type I error is set at $5 \%$.

[^227]:    ${ }^{1}$ The dependency could be equally well be between the relative clause verb and the head noun; nothing hinges on assuming a gaphead noun dependency.

[^228]:    ${ }^{2}$ In the published paper, Gibson and Wu (2013) did not delete any data, leading to the results shown in Table 1.

[^229]:    ${ }^{3}$ We also used a simpler method than k -fold cross-validation to compare the models; this method is described in Vehtari et al. (2016). The results are the same regardless of the model comparison method used.

[^230]:    ${ }^{4}$ A reviewer suggests that the direct-access model may simply be an elaboration of the distance model. This is by definition not the case: direct access (i.e., distance-independent access) is incompatible with the distance account.

[^231]:    ${ }^{1} \mathrm{CoV}$ is the standard deviation of a data set divided by its mean. Studies of magnitude estimation in animals and humans have found that response variability correlates with target magnitude, producing a CoV of 0.15 (see Whalen, Gallistel, \& Gelman, 1999).

[^232]:    ${ }^{1}$ Radical enactivism is just one position in the greater antirepresentationalist movement in the philosophy of cognitive science. Less ambitious positions within this movement will not be discussed in this paper.
    ${ }^{2}$ This understanding does not presuppose that cognition must be explained by appealing to representation.

[^233]:    ${ }^{3}$ There is another class of alternatives to the representational explanation as I characterize it in $\S 2$, which I call cognitiveneuroscience explanations. REC does not offer these explanations so they won't be considered here. Relating the representational explanation to these alternatives will be left for future work.
    ${ }^{4}$ CCTC can be made more rigorous with any number of formal theories of computation. Here, I follow Gallistel and King's and assume a functioning homomorphism view of computation and the Turing machine mathematical formalism because these apply to the desert ant example in section 4.1 (see Gallistel \& King 2009: 196-206).

[^234]:    ${ }^{5}$ This is a minimal conception of CCTC. Many classical computationalists, such as Jerry Fodor (1975), endorse a language of thought, but CCTC is compatible with the absence of a language of thought. Furthermore, CCTC need not (and I think should not) posit anything like that found in our folk psychological theories (Stich 1983).
    ${ }^{6}$ This is not the only reason RECers have for rejecting representations. For example, another reason, which I deal with below, concerns the causal efficacy of semantic properties. For brevity's sake I can present no more reasons here.

[^235]:    ${ }^{7}$ This is only a sketch of Wittgenstein's view, but a sketch will do here.

[^236]:    8 As Kaplan (2015) argues, dynamical models can be explanatory when construed as representing the dynamics of the mechanism responsible for the phenomena to be explained, but not when they are merely phenomenological. In these cases, the mechanism and not just its dynamic behaviour constitute the explanation.

[^237]:    ${ }^{9}$ Some might want to resist this claim and argue that detection systems do involve information-bearing structures. For good reasons to not to resist see (Ramsey 2007).

[^238]:    ${ }^{1}$ Motion tracking ( $\sim 5.32 \mathrm{~ms}$ ) and data transfer ( $\sim 5$ to 8 ms ) time, plus screen refresh rate $(\sim 13.33 \mathrm{~ms})$ resulted in a minimal delay between a participant's movement and rendering of 26.67 ms .
    ${ }^{2}$ Individuals creating similar chaotic movement sequences produced behavior with the average frequency for a given trial between .14 and .57 Hz , and an overall average frequency of .32 Hz (Washburn et al., 2015).
    ${ }^{3}$ Washburn et al. (2015) used these sequences to train individuals to act as master systems during interpersonal anticipatory synchronization and demonstrated that the training consistently led to individuals producing chaotic movement behavior.
    ${ }^{4}$ Harmonic spring systems are flexible with relatively few intrinsic dynamics. For slave systems with inherently chaotic dynamics it will be harder to evaluate whether anticipatory behavior of another chaotic system is primarily a product of coordination.

[^239]:    ${ }^{1}$ Belief states (Kaelbling et al., 1998) or predictive state representations (Littman et al., 2001) are sufficient for planning, but can be computed internally in each module and do not need to be communicated.

[^240]:    ${ }^{1}$ Compared to adults, children credited robots with slightly greater physiological capabilities. This is particularly obvious in examining modal responses for items like feel safe (adults: no, $n=82$; children: yes, $n=40$ ), feel tired (adults: no, $n=88$; children: no, $n=38$, yes, $n=36$ ), and feel scared (both age groups: no, $n=93$ adults, 50 children), which each have emotional and cognitive connotations in addition to their relevance for biological life.

[^241]:    ${ }^{2}$ Note, however, that have goals loaded more strongly on the social-emotional factor, and two potentially "moral" items (have self-control; communicate with somebody else), loaded equally on the social-emotional and perceptual-cognitive factors.

[^242]:    ${ }^{1}$ An alternative definition of complexity, perimeter squared over ink area, has been used successfully by Pelli, Burns, Farell, \& Moore-page (2006) to account for human efficiency in letter identification. However, it was found that this measure of complexity was a significantly weaker predictor of RT in the same/different judgment, as originally reported in Wiley, Wilson, \& Rapp (2016).

[^243]:    ${ }^{2}$ Only the "same" pairs are used here because they are used to measure the effects of visual complexity. The "different" pairs are discussed in detail in Wiley, Wilson, \& Rapp (2016), where they were used to determine the relative importance of various visual features (e.g. lines, curves) for letter perception and how that relative importance differed between naïve and expert observers.

[^244]:    ${ }^{3}$ The monoscriptal participants had varying degrees of knowledge of languages written in the Roman alphabet other than English, primarily Spanish or French.

[^245]:    ${ }^{1}$ We also considered Probability of Improvement and Probability of Maximum Utility (Speekenbrink \& Konstantinidis, 2015) as alternate decision strategies, but have omitted them because they failed to reach performance comparable to UCB.

[^246]:    ${ }^{2}$ Note that the peak in average reward for the GP-UCB is due to the use of human parameter estimates, whereas a GP-UCB model with optimized hyper-parameters and a dynamic $\beta$ is known to achieve sublinear regret bounds (i.e., monotonically increasing average reward; Srinivas et al., 2010)

[^247]:    ${ }^{3}$ Because horizon length varied within subjects, we compare the aggregate mean of the cross-validated parameter estimates for $\beta$.

[^248]:    ${ }^{1}$ These two questions, although indirect assays of the agents' desires, were selected as being more natural to the context. To

[^249]:    answer these two questions, however, children have to infer the two agents' desires.

[^250]:    ${ }^{1}$ This and subsequent numbers represent statistically significant results, as described above.

[^251]:    ${ }^{1}$ Multiple imputation is the recommended tool to predict missing data when missingness depended on other observed variables, but not the missing variable itself (Sinharay, Stern, \& Russell, 2001). In our case, because missingness was a result of the parent's decision, it was associated with patterns observed in parent-child interactions (as shown in the logistic regression), but was not directly associated with children's behavior in the test. Therefore, multiple imputation is suitable to simulate behavior for the notconsented children.
    ${ }^{2}$ The two test measurements regarding activation of target function were imputed as binary variables, whereas all other test measurements were imputed as continuous variables. For each test measurement, a logistic regression model (for binary variables) or a linear regression model (for continuous variables) was first fitted on the data from the consented group. From the fitted model, posterior distributions were computed for the 8 parameters in the logistic regression model or the 9 parameters in the linear regression model (intercept, coefficients for the 7 observational measurements, and the residual variance for linear regression only). Then the values for the not-consented group were imputed for $m=$ 100 runs. For each run, a new set of parameters were randomly drawn from their respective posterior distributions, and were used to compute the expected values plus random errors for each child in the not-consented group. The means and standard deviations of the not-consented group and of the whole population were then calculated for this test measurement. The procedure is then repeated for the remaining runs of simulations, and applied to the other test measurements.

[^252]:    ${ }^{1}$ In a recent book chapter (Zawidzki, forthcoming), Zawidzki takes a stronger stance on mindshaping. Although Zawidzki seemingly proposes mindshaping as a separate alternative to inferential mindreading, in doing so one may throw out the baby with the

[^253]:    ${ }^{2}$ Blokpoel et al. proved that Bayesian inverse planning can encode and 'solve' computational problems that are amongst some of the hardest problems known in computer science. For details (and a full tutorial) see Blokpoel et al. (2013).

[^254]:    ${ }^{3}$ No results are known for 1-p by itself. It is, however, prudent to assume that restricting $1-p$ by itself also does not lead to tractability of Bayesian inverse planning.

[^255]:    ${ }^{1}$ Due to space limit, these results were not reported here.

[^256]:    ${ }^{2}$ The log-likelihood of participants' eye movement patterns being classified as the eye-mouth pattern was also correlated with the percentages of "fearful" responses, but with a smaller Pearson's r, $r=.394, p=.007$. This effect may be due to the similarities between the two representative patterns. Note that the two patterns were significantly different: eye movements classified as the eye-mouth pattern had higher likelihoods of being generated by the eye-mouth model, and vice versa (Chuk et al., 2014).

[^257]:    ${ }_{1}$ Publication-based author.

[^258]:    ${ }^{1} \mathrm{http}: / /$ www.freeciv.org/

[^259]:    ${ }^{1}$ The coefficients $(\beta)$ are interpretable as log-odds, but can also be transformed to an odds ratio $\left(O R=e^{\beta}\right)$.

[^260]:    ${ }^{1}$ Due to the difference, the Google graphic was not included in the analysis since it was designed differently compared to the other questions.

[^261]:    ${ }^{2}$ The 6 Google logo related questions tested participants on correctly recalling the colors of each of the letters in 'google.'

[^262]:    ${ }^{1}$ An earlier version of R\&R (Alhama, Scha, and Zuidema (2016), Alhama and Zuidema (2016)) features a different probability for retention, with a binary switch over an attenuation parameter. This design was inspired by experimental studies in which the stimuli eventually contained 25 ms pauses, a duration that is supposed to be perceived by humans only subliminally. The stimuli we plan to use for our simulations, based on Frank et al. (2010), differ significantly in the use of pauses, which have a duration of 500 ms (and therefore should be clearly perceived). The retention probability we present here is more general, since the effect of pause length could be accounted for with different values of $\mu$.

[^263]:    ${ }^{2}$ The only parameter that we keep fixed in our search is $\mu_{n p}=$ 1.0 , since the interpretation of the relative importance of pauses is clearer if only one of the $\mu$ parameter is varied.
    ${ }^{3}$ We optimize our parameters on the same data we evaluate the model on, as seems to have been the case for the models we compare with. This brings the risk of overfitting, so in the discussion section we briefly discuss better ways of evaluating models.

[^264]:    ${ }^{4}$ Even though the values that parameter $D$ can take range from 0.0 to 1.0 , the number of types stored by $\mathrm{R} \& \mathrm{R}$ grow very rapidly in our model due to the memorization segments of any length. For this reason, small values are impracticable, since the probability of recognizing a segment quickly drops close to zero.

[^265]:    ${ }^{1}$ Graphs with high clustering coefficients and low average path lengths, as in small-world networks, are efficient to search and relay information through, while scale invariance allows a single algorithm to operate across seemingly disparate representational frames.

[^266]:    ${ }^{2}$ SEEDLingS networks contain many more nouns than the WordBank network (resulting in different connectivity patterns), but AoA data is only available for the 369 CDI nouns for all networks
    ${ }^{3}$ http://nlp.stanford.edu/projects/glove/

[^267]:    ${ }^{4}$ We omit these due to space, but thank an anonymous reviewer for this suggestion; they will be presented at CogSci.
    ${ }^{5}$ https://github.com/andreiamatuni/wordgraph
    ${ }^{6}$ https://github.com/BergelsonLab/semspace

[^268]:    ${ }^{7}$ Scale free networks are inherently ultrasmall (Cohen \& Havlin, 2003)
    ${ }^{8}$ Phonological neighborhood effects are saved for future work

[^269]:    ${ }^{9}$ Node degrees are from the SEEDLings-All network at peak similarity threshold of $\varepsilon=0.13$

[^270]:    ${ }^{1}$ The p-curve app conducts three different statistical tests to detect whether there is right skew. All of these tests were statistically significant ( $\mathrm{p}<0.001$ for all), indicating that the curve is very unlikely to be not right skewed.

[^271]:    ${ }^{2}$ This study was not included in the preliminary meta-analysis reported here because it did not provide the relevant standard statistical information.
    ${ }^{3}$ Another difficulty with sorting out the different potential effects of learned CP is that nearly all the published work is not designed to address this question. This makes it impossible in most cases to distinguish between, as an example, acquired distinctiveness of a dimension plus compression versus just expansion. These two cases have the same behavioral outcome, and thus must be distinguished through experimental controls, such as different kinds of training.

[^272]:    ${ }^{1}$ One research tradition which arguably falls outside these debates is the distributed cognition research programme (e.g., Hutchins, 1995a). While distributed cognition is explicitly cognitivist, it succeeds in avoiding the group actor assumption by analysing individual actions as part of a larger system encompassing physical artefacts as well as actors within a space (Hutchins, 1995b).

[^273]:    ${ }^{2}$ It might be argued that this is not really a joint action phenomenon at all. Just as Tomasello argued that the individual chimpanzees were all merely hunting in parallel, all in their own self-

[^274]:    interest, so one might argue that the pedestrian's activity and the driver's activity are merely two individual action phenomena that happen to overlap in space. But this sort of argument again commits us to dividing the world up a priori into individual and joint actions. The argument is only valid, in other words, if we already accept the group actor assumption.

[^275]:    ${ }^{1}$ The facial recognition algorithm was adapted from the Open Source Computer Vision Library (OpenCV v2.4.13; Bradski, 2000) and programmed in Python 2.7. OpenCV is an open source library that provides a common infrastructure to machine vision applications in academia and industry.

[^276]:    ${ }^{1}$ In the Smartphone condition participants could do the experiment with the following devices: iPhone, Android, Windows Mobile Phone and BlackBerry. In the PC condition participants could use a desktop or a laptop computer. No tablets were allowed.

[^277]:    ${ }^{2}$ In the Smartphone condition, $39 \%$ of participants used an iPhone during the experiment, $58.5 \%$ an Android, $2.2 \%$ a Windows Mobile Phone and $0.2 \%$ a BlackBerry.

[^278]:    ${ }^{1}$ Each set of items also included some non-temporal items, which are not discussed here.

[^279]:    ${ }^{2}$ That references using complementary prepositions (such as "in front of" vs. "behind") need not result in perfectly complementary response patterns was also observed for the spatial domain (e.g., Grabowski \& Weiß, 1996).

[^280]:    ${ }^{1}$ Olfactory and gustatory words were not used because there were too few in our candidate stimuli, and were not required.

[^281]:    ${ }^{2}$ Segment selection (partly based on the Brian Vision Analyzer tutorials at http://www.erpinfo.org/the-erp-bootcamp.html). The critical period spanned from 300 ms before target onset to 800 ms after target onset (the period before onset is 100 ms longer than the general baseline of the ERPs because that improved the selection of segments). Gradient: $75 \mu \mathrm{~V} / \mathrm{ms}$. Threshold for difference between maximum and minimum voltage in segment: $\pm 150 \mu \mathrm{~V}$ (this was increased or decreased by up to $40 \mu \mathrm{~V}$ in a minority of cases where the automatic selection yielded too noisy waveforms), interval length 200 ms . Amplitude: $-100 \mu \mathrm{~V}$, $+100 \mu \mathrm{~V}$. Low activity: $0.5 \mu \mathrm{~V}$, interval length 50 ms .

[^282]:    ${ }^{3}$ Quick group: Auditory-to-visual: $62 \%$ (SD=48 pp.). Haptic-tovisual: $61 \%$ ( $S D=49 \mathrm{pp}$.). Visual-to-visual: $63 \%$ ( $S D=48 \mathrm{pp}$.). Slow group: Auditory-to-visual: $64 \%$ ( $S D=48$ pp.). Haptic-tovisual: $64 \%$ ( $S D=48 \mathrm{pp}$.). Visual-to-visual: $64 \%$ ( $S D=48 \mathrm{pp}$.).

[^283]:    ${ }^{1}$ In all analyses in the present paper including the three-level factor DISCOURSE CONDITION the degrees of freedom were corrected by applying the Greenhouse-Geisser correction. In the text the uncorrected degrees of freedom are reported.

[^284]:    ${ }^{1}$ One exception is the work of Carruthers, Masson, and Stege (2012) which found that the planarity of graphs has no effect on human performance in the Vertex Cover problem.

[^285]:    ${ }^{2}$ Here we will only consider undirected graphs.
    ${ }^{3}$ We included the spectral gap as it admits an efficient polynomial algorithm as opposed to the other three predictors which are NPhard. As such, it may prove as a practical predictor of performance.

[^286]:    ${ }^{1}$ Specifically, the previous model possessed six components used to achieve the primary cross-dimension comparison task: the object height reporter and object length reporter; the length comparator, which compared the results from the object height/length reporters; comparison recorder; and the center highlighter and orientation inhibitor, which generated a fixation on the center proto-object and inhibition on peripheral proto-objects, respectively.

[^287]:    ${ }^{2}$ When there are three dots, only one dot goes uncounted, which is approximated with high accuracy.

[^288]:    ${ }^{3}$ Our usage of the terms serial and parallel in this paper align with this strong and weak sense of attentional dependency, respectively.

[^289]:    ${ }^{1}$ To do so, agents would have to infer or come equipped with knowledge about $P_{N}\left(\cdot \mid s_{a}\right)$, which could itself be subject to updates. We stick to the simpler case of noise-free inference here, but as long as the actual state is not always recoverable our general results also hold for agents that reason about noise.

[^290]:    ${ }^{2}$ Knowing $k$ allows learners to compute the likelihood of a type not reporting $k-\left|d_{l}\right|$ state observations. A better but more complex alternative is to specify a prior over $k$ with learners performing a joint inference on $k$ and the teacher's type. For simplicity, we opt for the former, albeit admittedly artificial, assumption.

[^291]:    ${ }^{3}$ Concretely, results are obtained for probabilistic speaker behavior following the definitions of Brochhagen et al. (2016). Nothing essential to our main argument and simulation results hinges on these details, so we background them here for ease of exposition.

[^292]:    ${ }^{1}$ We collapsed congruent RTs (Yellow prime/Yellow target and Green prime/Green target) and we did the same for the incongruent RTs (Yellow prime/Green target and Green prime/Yellow target.

[^293]:    ${ }^{1}$ This vector is computed by calculating the dot product between the situation-state matrix and the original $25 k$-dimensional situation vector, and then normalizing each dimension of the resulting vector by the sum over the dimensions of the original $25 k$-dimensional situation vector.

[^294]:    ${ }^{1}$ For the case of $\alpha_{k}$, the conditional posterior cannot be computed analytically and the parameter is sampled with the MetropolisHastings rule (Yildirim, 2012).

[^295]:    ${ }^{2}$ Taken from Salakhutdinov et al. (2012)

[^296]:    ${ }^{3}$ Averages across 10 random repetitions and all categories are reported.
    ${ }^{4}$ Each category contains a varying number of examples

[^297]:    ${ }^{1}$ One-tailed binomial tests reflect the directional nature of our hypothesis, but the outcome is comparable with two-tailed tests (same: $p=.052$; different: $p=.020$ ).

[^298]:    ${ }^{2}$ We compared performance of Mandarin-speaking preschoolers in the current study with English-speaking preschoolers in Walker et al. (2016). Considering each condition separately, we find a significant difference between Mandarin- and English-speaking preschoolers in the different condition (one-tailed $p=.022$, Fisher's exact) and a marginal difference in the same condition (one-tailed $p=.068$, Fisher's exact). Combining across different and same conditions, we find that Mandarin-speaking preschoolers significantly outperform English speakers (one-tailed $p=.004$, Fisher's exact).

[^299]:    ${ }^{1}$ Precision and recall scores consider the "high" class as positive label.

[^300]:    ${ }^{1}$ I also experimented with a procedure that excluded any TENS/DIGITS pairs from the proposal distribution for a given form that were assigned to any previous forms within a window of arbitrarily chosen size. However, this exacerbated the label-switching problem (a trivial issue); less trivially, it was difficult to motivate a window size which plausibly paralleled working memory.

[^301]:    ${ }^{2}$ It is well known that South Asia is home to rigid societal hierarchies, and historically, exact number systems may have been the preserve of elite groups, while marginal groups relied on an alternative system (prior to language standardization). While this theory is blatant speculation, it is worth briefly noting that Sinhala, an Indic language whose speakers chiefly practice Buddhism (which preaches a doctrine of egalitarianism), developed a transparent number system.

[^302]:    ${ }^{1}$ In order to avoid biasing the results with children who were responding with nonsensical numbers, we excluded responses on which PAE scores were over $500 \%(n=2)$ for this analysis.

[^303]:    ${ }^{2}$ Because Overall PAE and Work Expended PAE were highly collinear and conceptually and empirically confounded, we ran separate models using each PAE type (see Table 2).

[^304]:    Such a result would advance our understanding of both the mechanisms controlling perceptual learning and the face inversion effect in a number of ways. We would have found an experimental procedure (anodal tDCS at Fp 3 brain site) able to selectively affect perceptual learning and its expression, and this would help in discriminating between competing theories. Furthermore, we would have additional evidence that perceptual learning is a contributor (at least in part) to the face inversion effect. Finally, this would be the first demonstration in the literature of how relatively brief

[^305]:    ${ }^{1}$ The precise details of the strategy classification had to be omitted from this paper for space reasons, but will appear in a longer version of this manuscript that is currently under review.

[^306]:    ${ }^{1}$ Note that Jern and Kemp (2013)'s model is slightly different, as they used a non-conjugate model. Their model acts very similar to our version of it and receives comparable fits.

[^307]:    ${ }^{1}$ The experiment took place in an open plan office and there were other colleagues working quietly in the room. Participants were made aware of this before the experiment started, so that they were not distracted by this.
    ${ }^{2}$ The full list of words and phrases used for each category, are as follows:- Certainty adverbials: absolutely, actually, certainly, clearly, plainly, definitely, evidently, indeed, obviously, really, surely, undoubtedly, unquestionably, for certain, for sure, of course; Uncertainty adverbials: allegedly, apparently, arguably, conceivably, inexplicably, likely, maybe, perhaps, possibly, potentially, presumably, probably, reportedly, seemingly, supposedly; Uncertainty modals: may, might, can, could and Hedges: quite, sort of, kind of , might, a bit, a little bit, just, at least, approximately, about, around, something like, almost, pretty, sometimes.

[^308]:    ${ }^{1}$ In this context the term 'false information' refers to incorrect or inaccurate information that is initially presented as true.

[^309]:    ${ }^{2}$ Planned contrasts are reported for predicted differences. Tukey's posthoc tests are reported when no difference between conditions was predicted.

[^310]:    ${ }^{1}$ It is important to note that there are numerous domains within cognitive science that formalize representation in various ways. In the present paper, we work within the framework of structuremapping theory.

[^311]:    ${ }^{1}$ Software available at gitlab.bucknell.edu/AI-CogSci-Group/IGT-Open/

[^312]:    ${ }^{1}$ Although more sophisticated amortization schemes have been developed in the machine learning literature (e.g., Stuhlmüller et al., 2013; Rezende et al., 2014; Paige \& Wood, 2016), they are difficult to test experimentally in humans.

[^313]:    ${ }^{1}$ Fits were carried out per subject and identified parameters that minimized squared error. The parameters were $w_{Y}$ (the marginal probabilities of $Y_{A}$ and $Y_{B}$ ), $w_{Y X}$ (the strength, or causal power, of the links between the $Y \mathrm{~s}$ and $X$ ), and $w_{X}$ (the strength of alternative causes of $X$ ). Predicted conditional probabilities [0-1] were multiplied by a scaling parameter $s$ to bring them into the range of subjects' ratings [0-100]. The best fitting parameters averaged over subjects were $w_{Y}=.401, w_{Y X}=.483, w_{X}=.178$, and $s=158.6$.
    ${ }^{2}$ Rather than explicit sampling, the causal samplers predictions for a given chain length has an analytic solution involving repeated multiplication of the matrix of transition probabilities between graph states defined by the Metropolis-Hastings rule. Fractional values of chain length involve computing the weighted average of the joint probability distributions that obtain when chain length is rounded up and down. The best fitting parameters averaged over subjects

[^314]:    were $w_{Y}=.440, w_{Y X}=.469, w_{X}=.233, s=137.7$, and chain length $=17.9$.

[^315]:    ${ }^{3}$ Our lab has subsequently extended these finding to a more traditional hypothesis testing task in which subjects rate the posterior probability of the graph in Figure 1A relative to alternative hypotheses (those formed by removing one or both of the causal links).

[^316]:    ${ }^{4}$ The best fitting parameters $\left(w_{Y}=.519, w_{Y X}=0.440, w_{X}=.243\right.$ averaged over subjects), were those that maximized the likelihood of the distribution of pennies.
    ${ }^{5} w_{Y}=.534, w_{Y X}=0.410, w_{X}=.328$, chain length $=10.1$.

[^317]:    ${ }^{1}$ Technically, exact back-propagation cannot be applied to a stochastic network, though there are several versions of "approximate" back-propagation for stochastic networks, e.g. (Gu, Levine, Sutskever, \& Mnih 2015)

[^318]:    ${ }^{1}$ The word 'argumentation' in English sometimes conveys an idea of dispute or may refer to situations in which some individual tries to convince another. We do not consider such restrictions.

[^319]:    ${ }^{2}$ Note that speaker and hearer may adopt different attitudes, e.g. if the latter proposes a solution (go and grab the child) to what was a 'look at this!' utterance.

[^320]:    ${ }^{3}$ One such signalled event is a news about the near future. It is listed as 'narrative', though narratives are typically about past events.

[^321]:    ${ }^{4}$ Note that the 'resource-bounded' restriction makes the notions of complexity and of unexpectedness computable (Saillenfest \& Dessalles, 2015).

[^322]:    ${ }^{1}$ Due to an error in setting up the study, one trial was lost in the unattractive last condition in one third of participants. Therefore, only 24 trials instead of 25 went into data analysis for those participants in that condition.

[^323]:    ${ }^{1}$ Children have been successfully trained to use rehearsal to support delayed serial recall, but they stopped using the strategy when told they did not have to use it (Keeney, Canizzo, \& Flavell, 1967). Whether children stopped because they could not sustain the strategy or because the experimenter's instructions influenced motivation to use it is not clear.

[^324]:    ${ }^{1}$ For the time being we leave the question of realism of the acoustic part to future work.

[^325]:    ${ }^{2}$ https://github.com/tensorflow/models/tree/master/ inception

[^326]:    ${ }^{3}$ http://scikit-learn.org/stable/modules/clustering

[^327]:    ${ }^{4} \mathrm{An}$ area in the map is associated to a word if the activation of the neurons within it are at their peak when they respond to a stimulus of that particular word.

[^328]:    ${ }^{1}$ Source code and notebooks available as a Github repository at https://github.com/sebastien-forestier/CogSci2017

[^329]:    ${ }^{1}$ Because participants were on average enthusiastic to learn new topics in math ( $M=4.29, S D=0.94$ ), more so even than other topics, we did not further analyze the results of this measure.

[^330]:    ${ }^{1}$ This article is an abridged, modified version of Mareschal, D. \& French, R. M. (2017) TRACX2: a connectionist autoencoder using graded chunks to model infant visual statistical learning. Phil. Trans. R. Soc. B 2017372 20160057; DOI: 10.1098/rstb.2016.0057.

[^331]:    ${ }^{1}$ In this paper, a $95 \%$ confidence interval was calculated by bootstrapping, based on 1000 sampling, with replacement.

[^332]:    ${ }^{2}$ Two participants in self-others, two in dialectical, and one in selftwice conditions were excluded from analysis, as confidence data was not collected.

[^333]:    ${ }^{1}$ In reality, compound words (especially very frequent ones) might be planned as a single phonological sequence (Jacobs \& Dell, 2014). This does not affect the interpretation of our results, as we did not ask our participants to produce compounds, but rather sequences of two unrelated words.
    ${ }^{2}$ While Griffin (2003) found a reversed word-length effect on speech latencies (i.e., longer latencies when $\operatorname{word}_{l}$ was shorter) as well as on gaze-speech lags, Meyer et al. (2007) only found this effect on gaze-speech lags.

[^334]:    ${ }^{3}$ Although Griffin (2003) varied word2 frequency and length, between-items differences were very small.

[^335]:    ${ }^{4}$ Variation in the Long condition was not sufficient to allow treating this variable as a continuous predictor in the analyses. Instead, Length was treated as a categorical predictor (Short vs. Long) throughout.

[^336]:    ${ }^{5}$ Perhaps parafoveal information was sufficient for speakers to identify intact right pictures more often than degraded ones.

[^337]:    ${ }^{6}$ Alternatively, retrieval difficulties with word $_{2}$ could have interfered with preparation of the articulatory code for the first syllable of word ${ }_{1}$, and slowed it down. However, note that articulatory retrieval does not appear to impose huge demands on central attention (Roelofs \& Piai, 2011). Also, this interference account would have also predicted a smaller effect of degradation on the gaze-speech lag when word $_{l}$ was long, because with long words the temporal overlap between articulatory encoding for the first syllable of word $_{1}$ and word $_{2}$ retrieval should have been shorter.

[^338]:    ${ }^{1}$ We have the original raw data file from the norming study, but, alas, have lost the handwritten surveys on paper. We therefore conducted a second norming study with the same materials. The results were similar and so we used the data from the original norming study. Interestingly, the few items were we found slight differences were the ones related to immigration - a topic that recently became highly controversial in Germany and many other countries.

[^339]:    ${ }^{1}$ e.g., JIBO: http://www.jibo.com

[^340]:    ${ }^{2}$ Ideally, an operationalization of brevity should obtain some metric from the surface realization of a potential utterance (e.g., phoneme count, simulated speech output time, etc.). This architectural integration is still a work in progress.
    ${ }^{3}$ While DIARC has the capacity to handle unconventional indirect requests (e.g., "My batteries are running low..."), for sake of clarity we focused on more conventional cases in our demonstration.

[^341]:    ${ }^{4}$ Scenario 1 involved an elder care setting in which a robot asked the nurse for a sick patient's medication ("Hand me the red pills."). Scenario 2 involved a household robot running low on battery that asked to be plugged in before important data was lost ("Plug me in."). Scenario 3 involved a service robot that requested to take a child's coat at a fancy reception ("Hand me your coat."). Finally, Scenario 4 involved a mine-sweeping robot that asked its superior officer to step aside as it searched a room in a training exercise ('Move out of the way.").

[^342]:    ${ }^{5}$ For example, what does it mean for utterance A to be equally less polite (e.g., 0.4) than utterance $B$ as utterance $B$ is less informative than utterance A?

[^343]:    1 There are two basic interpretations of the concept counterfactual in cognitive science: 'contrary to reality' (Pearl 2000; Hiddleston 2005; Rips \& Edwards 2013), and 'possible, but not implemented in some situation' (Roese \& Olson 1995; Roese 1997). For this paper only the first interpretation is actual.

[^344]:    ${ }^{1}$ Only 37 from 50 recordings were valid for the pause measuring.

[^345]:    ${ }^{1}$ Furthermore, in the princeps study that introduced the SSF framework, the small $b$ values, and the ones close to $a$ were equally present in the Si - and MA-versions of the problems.

[^346]:    ${ }^{1}$ This formula has been adapted across a range of studies, such as in political psychology research (Hardisty, Johnson \& Weber, 2010).

[^347]:    ${ }^{2}$ In each analysis 175 responses were analyzed (4 responses for each participant, one for each scenario, apart from one participant whose responses were not recorded for the final scenario due to an eye-tracking malfunction).
    ${ }^{3}$ GEE model was chosen as it represents a flexible approach to handling correlated data structures. A full discussion of this method can be found in Honish Edwards, Elden \& Leonard (2010).

[^348]:    ${ }^{4}$ Analysis conducted via a one-sample t-test with test value of 50.

[^349]:    ${ }^{5}$ Degrees of freedom for each study scenario:
    Study 1 df's: between (1, 101-104) Study 2 df's: $(1,39-40)$
    ${ }^{6}$ Indeed, in our study, as the one risk introduced four different side effects, we used information about three of the four of these as our three separate risks.

[^350]:    1 Three of the 62 students who initially participated were excluded from the analyses due to a technical error (one leading to a posttest log not being created, and the other to one of the topics between a pair not being discussed). There was a procedure error for two additional participants (i.e. they had kept their learningmaterial window open and used it for the posttest). Lastly, one student withdrew participation assent during the posttest. In addition, of the remaining 56 , due to an ID entry error one participant lacked a pretest log, and therefore was excluded from the multiple-choice question analysis, and for two participants we could not link their short-answer log IDs to their dialog screen IDs; they were excluded from the dialog metric analyses.

[^351]:    ${ }^{1}$ Gives the probability of $k$ successes in $n$ draws, from a population of size $N$ with exactly $K$ successes. Thus, similar to the binomial distribution, but drawing without replacement.

[^352]:    ${ }^{1}$ https://bitbucket.org/julianhough/stir

[^353]:    ${ }^{1}$ It is even possible for an a.s.-terminating program to have infinite expected run time. Consider a program defining a geometric distribution that repeatedly flips a fair coin until first flipping a heads after $n$ steps, then continutes for $2^{n}$ more steps thereafter.

[^354]:    ${ }^{1}$ For a concrete description of the removal see Section "Data Analysis" on the following page.

[^355]:    ${ }^{2}$ As every trial contained two gaze actions, one aligned to the first noun and one aligned to the second noun, the total number of gaze actions throughout the course of the experiment was 288 per list/participant.

[^356]:    "This diagram is intended to show a curved tube. In the diagram you are looking down on the tube from above. A metal ball is put into the end of the tube indicated by the arrow. The ball is then shot out of the other end. Your task is to draw the path the ball will follow after it comes out of the tube. Please flip the page and proceed."

[^357]:    ${ }^{1}$ As just one example, Bennett (1975, p. 71) defines in and on using the notions of location at the interior of an object (for in), and location at the surface of an object (for on).
    ${ }^{2}$ We examined the 7 items for which the rate of expression use

[^358]:    ${ }^{2}$ We examined the 7 items for which the rate of expression use exceeded the truth-value acceptance rate and found that, for all but one item, the absolute difference between use and acceptance was less than 0.3.

[^359]:    1 For the document pairs used in Lee et al. (2005), the highest reported correlation between a model and humans was .77 (Yeh, Ramage, Manning, Agirre, \& Soroa, 2009).

[^360]:    3 We considered the approach of transforming scores into a 0-1 scale, as in Lee et al. (2005). However, this approach is overly sensitive to the minimum and maximum values. On the other hand,

[^361]:    ${ }^{1}$ Data of native signers were collected as part of a bigger project conducted between 2010 and 2015 (Sümer, 2015).
    ${ }^{2}$ Pictures used in the study were originally developed by Dr Jennie Pyers and were adapted further for the purposes of this present study.

[^362]:    ${ }^{1}$ In total, the data of 2079 and 583 participants went towards the joint Simon ( $M=21.88 /$ group, $S D=9.02$ ) and $\operatorname{IGNG}(M=$ 16.66/group, $S D=5.79$ ) random-effects meta-analyses, respectively.
    ${ }^{2}$ When the number of participants per group was not specified, the total number of participants reported was assumed to be distributed evenly amongst groups. Standard errors (SE) were converted into standard deviations (SD) for future computations.
    ${ }^{3}$ Control condition criteria included a physically present, human co-actor, actively responding to an alternative stimulus.

[^363]:    ${ }^{4}$ In cases where authors provided alternative hypotheses regarding whether the JSE would manifest itself, or not, we could not definitively code the condition as a 'wipeout' or 'non-wipeout,' and the dataset was excluded from the analysis ( $n=10$ ).
    ${ }^{5}$ Large samples were defined as $n>24$, reflecting the $75^{\text {th }}$ percentile of sample size.

[^364]:    ${ }^{6}$ The summary effect size of wipeout conditions was also significantly smaller than that of control conditions, but we have omitted this additional analysis for brevity, given that the reported difference between wipeout and non-wipeout conditions is stronger evidence of the former's impact on the size of the JSE.

[^365]:    ${ }^{7}$ A supplementary analysis of the control versus non-control task conditions, with the wipeout data removed from the latter to diminish potential cancellation effects, also yielded no significant difference between the summary effect sizes. This supports the notion that there is some flexibility in the task conditions that can be applied and still elicit the JSE.

[^366]:    ${ }^{1}$ Although it would be more appropriate to speak about entities including humans, in the following we refer to spatial objects.

[^367]:    ${ }^{2}$ This was due to a better readability. For the purpose of calculating Odds Ratio and the further meta-analysis, absolute frequency of accuracy based on the number of participants and the percentage of the correctness of the given responses are used.

[^368]:    Together with two former classmates from university, Agent 1 and Agent 2, you have founded a small startup company of which you are the CEO. You are in charge of the management and finances, while Agent 1 and Agent 2 manage the creative direction part. Your team works from Mondays to Thursdays. The printer in your office works fine, only when it receives two printing orders at the same time it crashes.

[^369]:    ${ }^{1}$ Each of these two single-mental state conditions involves more specific tests of integration capacity. For example, in explicit desire (alone) condition, the participant's ability to produce a false belief indicates that she can recognize the relevance of the provided desire to an action that produces an unforeseen outcome.

[^370]:    ${ }^{1}$ The compensation-consistency parameter is scaled so that the the maximum compensation is twice the number of steps in the accumulation process, because due to the scaling of $\rho$ to $[-1,1]$ the maximum evidence is $S$ and the minimum evidence is $-S$. Thus, the maximum compensation can not be bigger than twice the number of accumulation steps

[^371]:    ${ }^{2}$ The Bayes factors were calculated with ttestBF function of the R-package "BayesFactor"

[^372]:    *This equation assumes a constant branching factor and an upper bound on the complexity of planning. People's planning time likely increases less than exponentially fast with the planning horizon but our approximation may be sufficient for small problems.

[^373]:    ${ }^{1}$ Throughout this paper, we will use the term "conditionals" to refer specifically to indicatives.

[^374]:    ${ }^{2}$ cf. Strong Ramsey test, proposed by Rott (1986), according to which "If $p$, then $q$ " is acceptable if and only if $q$ is acceptable under the supposition of $p$ but not acceptable under the supposition of $\neg p$.

[^375]:    ${ }^{3}$ In the "sensibleness" data, this effect was also significant, $t(61.65)=2.86, p=.006, r=.35$.

[^376]:    ${ }^{4}$ Note that, in the context of our experiment, the differences in people's assertability judgments could have only resulted from the semantics of the evaluated sentences, since the conversational contexts remain constant.

[^377]:    ${ }^{1}$ There was a marginal effect of order; Forced Choice (FC) Retention before Free Recall (FR) led to higher Retention (FC: $\mathrm{t}(46)=2.38, p=.021$, and $\mathrm{FR}: \mathrm{t}(44)=1.88, p=.067$ ). However, main effects and interactions in the overall ANOVA remain the same with and without order included.

[^378]:    ${ }^{1}$ Our implementation adapts source code from https://github.com/devsisters/DQN-tensorflow

[^379]:    ${ }^{1}$ http://www.hera-project.com

[^380]:    ${ }^{1}$ For each side of the table, subtask trajectories examined include: rest-to-pickup, pickup-to-target, pickup-to-pass, rest-to-receive, receive-to-target, pass-to-rest, and target-to-pickup.

[^381]:    ${ }^{1}$ Where there are duplicated singular values, the SVD is not unique, so more precisely we mean there exists a basis which makes the SVD is block diagonal.

[^382]:    ${ }^{2}$ In the general case representations will be distributed across the hidden units, and so there will not be a unit which responds to the analogy and nothing else, but this is simply a rotation of the representation space, and because of the linearity of derivatives the same general pattern will emerge.

[^383]:    ${ }^{1}$ Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set.

[^384]:    ${ }^{1}$ The linear nature of this relation was confirmed using a Generalized Additive Model with penalized spline-based smoothers.

[^385]:    ${ }^{1}$ The ACL Wiki lists the performance of some key approaches: https://www.aclweb.org/aclwiki/index.php?title=TOEFL_Synony m_Questions_(State_of_the_art)

[^386]:    ${ }^{1}$ Three children in the inaccurate condition did not provide answers to the "good or not good" question and one child in the accurate condition did not provide answers to the "ask for help" question.

[^387]:    _—_ Musicians - English non-words preceded by musical segments
    Musicians - English non-words preceded by Tibetan letter strings
    Non-musicians - English non-words preceded by musical segments
    _ Non-musicians - English non-words preceded by Tibetan letter strings

[^388]:    ${ }^{1}$ http://www.crowdflower.com

[^389]:    ${ }^{2}$ The mean proportion of dimension-congruent judgments is the mean of congruent responses per participant over 24 responses.
    $3^{3}$ http://cran.r-project.org/web/packages/lme4/

[^390]:    4http://www.cran.r-project.org/package=ordinal/

[^391]:    ${ }^{1}$ We also estimated models that included polynomial terms. However, none of the polynomial models improved model fit and were not reported here.

[^392]:    ${ }^{1}$ Participants were also prompted to report their confidence after both observe and bet trials; in the interest of space we do not report analyses of confidence here.

[^393]:    ${ }^{1}$ We have provided a full set of our materials on the Open Science Framework, which can be accessed via the following link: https://osf.io/49nf7/

[^394]:    ${ }^{1}$ http://www.wsi.uni-tuebingen.de/lehrstuehle/cognitive-modeling/staff/staff/johannes-lohmann.html
    ${ }^{2}$ Based on the inertial data from the IGS suit, it is possible to calculate a kinematic chain with the hips as root. Hence, the position of the hip joint in the virtual scene is the reference point for all body

[^395]:    movements.
    ${ }^{3}$ While most participants remained unaware to the manipulation and attributed the variance in their grip aperture to inaccuracies of the tracking equipment, two participants reported to be aware of the manipulation after the experiment. Seeing that conscious awareness was not critical in this experiment, we did not perform a behavioral manipulation check in terms of a signal detection task to determine whether participants were able to consciously detect the manipulation of the visual grip aperture.

[^396]:    ${ }^{4}$ Please note that the data pattern remains nearly unaffected if the data is not filtered. Removing the size estimates from error trials only reduces the effect size of the three-way interaction.

[^397]:    ${ }^{5}$ Available at http://www.igroup.org/pq/ipq/index.php
    ${ }^{6}$ The considerable overestimation might be partially due to the initial slider position in the visual reproduction, starting at $90 \%$ of the sliding range for larger cubes.

[^398]:    ${ }^{1}$ Object locator constructs a priority map in three steps. First, following Bouma's law for visual crowding (Whitney \& Levi, 2011), object locator generates Marr wavelets centered at each tracked object's location, scaled so that the sizes of the suppressive fields increase as those objects enter the periphery. Second, since untracked stimuli produce visual crowding at a weaker rate than tracked stimuli (Whitney \& Levi, 2011), for each protoobject, we include the negative component of a wavelet whose amplitude is set to $20 \%$ of that for tracked objects. Third, Holcombe, Chen, and Howe (2014) report a general cost for having more tracked objects. We account for this effect with long-range suppression in the visual field, implemented as a constant value ( 0.04 in the model) subtracted from the entire visual field outside of each tracked object's enhanced region.

[^399]:    ${ }^{1}$ Thornton et al. had observed the largest effect size in the digit cancellation task, so we chose to include only this task to make room for additional measures testing the opportunity cost framework.
    ${ }^{2}$ Thornton et al. found no effects of gender in the original study, so we did not attempt to balance the gender representation in our sample of participants.

[^400]:    ${ }^{3}$ The distributions of cancellation scores were not normal, so we applied Wilcoxon's rank-sum test. ( $t$-test gave similar results)

[^401]:    ${ }^{4}$ The distribution of effort ratings was not normal for the simple cancellation task, but conformed to normality for the additive task.

[^402]:    ${ }^{1} \mathrm{P}(\mathrm{E})=$ perceived expertise $(0-1) ; \mathrm{P}(\mathrm{T})=$ perceived trust $(0-1)$; Rep is the represented statement (e.g., Hypothesis =1)

[^403]:    ${ }^{2} \mathrm{p}($ rep $\mid \mathrm{h})=\mathrm{p}($ rep $\mid \mathrm{h}, \exp$, trust $) * \mathrm{p}(\exp ) * \mathrm{p}($ trust $)+\mathrm{p}($ rep $\mid \mathrm{h}$, $\neg \exp$, trust $) * p(\neg \exp ) * p($ trust $)+p($ rep $\mid \mathrm{h}, \neg \exp , \neg$ trust $) * \mathrm{p}(\neg \exp )$ * $\mathrm{p}(\neg$ trust $)+\mathrm{p}($ rep $\mid \mathrm{h}, \exp , \neg$ trust $) * \mathrm{p}(\exp ) * \mathrm{p}(\neg$ trust $)$, mutatis mutandis for $\mathrm{p}($ rep $\mid \neg \mathrm{h})$

[^404]:    ${ }^{3}$ While it is not possible in the current model, this enables negative campaigning, as candidates could provide a negative representation (e.g., $\mathrm{p}($ candidate $\neg$ rep)) and attacks designed to undermine the trustworthiness or expertise of the opposing candidate. Intentionally, the model is built compartmentally to allow for increasingly complex persuasion campaigns.

[^405]:    ${ }^{1}$ For the sake of clarity, we provide the background story as an appendix after the bibliography.

[^406]:    ${ }^{2}$ Alongside the background story, full witness descriptions can be found in the appendix.

[^407]:    ${ }^{1} \mathrm{http}: / /$ semarch.linguistics.fas.nyu.edu/barker/Syllables/

[^408]:    ${ }^{2}$ We found that raw NWR performance scores resulted in weaker linear models and did not reach significance as a predictor. Therefore, we focus on the Phonological Chunk Sensitivity metric in the analyses (see Part 2).

[^409]:    ${ }^{1}$ Residual familiarity was derived by regressing out NoM, NoS, and NoM vs. NoS from raw familiarity.

[^410]:    ${ }^{1}$ Feldman and Tremoulet (2008) only briefly mentioned this hypothesis at the end of their paper, the bulk of which focused on detecting agency from observed motion patterns. The idea of using computational models (e.g., finite state automaton) to formalize agent detection is shared by our paper, but we dived into a very different domain-static visual sequences, where there is a substantial literature on the formalism of subjective randomness to draw on.

[^411]:    ${ }^{2}$ Griffiths and Tenenbaum (2003) only showed each participant half (128) of the stimuli-that is, one either saw a sequence (e.g., HTHHTHTT) or its complement (e.g., THTTHTHH).

[^412]:    ${ }^{3}$ This was suggested by two anonymous reviewers.

[^413]:    ${ }^{1}$ The structure-mapping theory and the multiconstraint theory postulate that the new hypothetical entities should be capable of filling the role played by their corresponding base objects, but they do not give further specifications about how to identify these entities in the target domain (see, e.g., Falkenhainter et al., 1989; Holyoak et al., 1994). With respect to the transferred relations, the structure-mapping theory (see, e.g., Falkenhainer et al., 1989) maintains that a generated relation in the target will be assumed to be identical to the corresponding source predicate. However, to the extent that this theory has always treated identicality as a "tiered" condition (Forbus, Ferguson, Lovett, \& Gentner, in press), we suppose the theory would admit inferences that include nonidentical relations as long as they can be regarded as identical at a higher level of abstraction. The multiconstraint theory adheres to the "copy of relations" postulate, but it treats it as a default rule that is adequate for initial explorations in the target (see, e.g., Holyoak et al., 1994). Under these considerations, we assume that CWSG involves either copying the base relations or replacing them by similar ones.

[^414]:    Note. ${ }^{*}$ Significant at $\alpha=.05 ;{ }^{* *}$ Significant at $\alpha=.01$

[^415]:    ${ }^{1}$ Two week test-retest reliabilities (from separate sample, $N=$ 215): $r=0.795$ ( 5 -item measure) and $r=0.923$ ( 10 -item measure).
    ${ }^{2}$ Conformity, for which ratings of both increases and decreases did not significantly differ from the scale midpoint of 50 , was excluded, though overall results do not change if it is included.

[^416]:    ${ }^{3}$ Only the 5 -item personality measure in Study 2, where participants on average experienced greater positive net change over the study period, substantially deviated from this pattern.

[^417]:    ${ }^{1}$ Although eye-tracking data were collected, they were not analyzed for this study.

[^418]:    ${ }^{1}$ In a roughly Bayesian sense
    ${ }^{2}$ Though the construction of a privileged and precise version of this concept is of considerable value, it is regrettably beyond the scope of the present article.

[^419]:    ${ }^{3}$ Such as in the earlier example involving misaligned concepts of justice.
    ${ }^{4}$ At least in this author's experience

[^420]:    ${ }^{5}$ Context is an incredibly complex topic, but for the purposes of this article, we will leave its nuances aside.
    ${ }^{6}$ Except possibly some component of the context, but the variance in context between participants in an interaction is beyond the scope of this article.

[^421]:    1 Wolf packs are only distributed cognitive systems in the context of the hunt. Outside of this context, they retain only a loose association and they do not otherwise share in joint tasks.

[^422]:    ${ }^{2}$ O'Donnell and colleagues (2015) describe wasp swarms as distributed cognitive systems, a usage inconsistent with the established definitions. This is an example of how the lack of an array of different concepts of nonreductive social cognition can leave some researchers forcing square pegs into round holes.

[^423]:    ${ }^{3}$ This may provide an alternate explanation for why individual vigilance against predators in Quelea quelea decreases with flock size (cf. Lazarus, 1979). Lazarus explains this effect in terms of an economy of energy, but it may simply arise from a situation in which only a local region of birds within the flock actively respond to predators.

[^424]:    ${ }^{4}$ The properties of the collective need not resemble the properties of its members (Hutchins, 1995a).

[^425]:    ${ }^{1}$ Slope estimates represent log transformed RT data.

[^426]:    ${ }^{1}$ Note that there are other appraisal variables proposed by the theory. Describing all the appraisal variables and their meanings is out of the scope of this paper.

[^427]:    ${ }^{2}$ Our computational model is inspired by the work of S. Marsella, Gratch, and Petta (2010) but implemented with completely different computation mechanism.

[^428]:    ${ }^{3}$ Currently, our model can generate and express eight different emotions described by (Ortony et al., 1990) namely Joy, Distress, Appreciation, Reproach, Gratitude, Anger, Liking and Disliking.

[^429]:    ${ }^{1}$ The materials were kindly provided by Stella Christie.

[^430]:    ${ }^{1}$ In fact, models of informativeness are often operationalized in terms of ruling out competing referential interpretations (Frank \& Goodman, 2012). The only way to apply such models to our experiment is with the trivial assumption that all true descriptions are equally informative.

[^431]:    ${ }^{2}$ Saturation and value were held constant to reduce the dimensionality of color space and to understand how people use color words to partition the range of hue values. This procedure is common practice when assessing expectations for basic color categories (Persaud \& Hemmer, 2016; Sims, Ma, Allred, Lerch, \& Flombaum, 2016)

[^432]:    It is not the case, that: If the toy block is not a cube, then the toy block is a cube.

[^433]:    If the figure on the upward facing side of the die is a circle, then the figure is black.

[^434]:    ${ }^{1}$ Analysis using a mixed logistic regression model-structuring outcomes as binary responses-did not change our findings

[^435]:    ${ }^{1}$ The authors wish to express their gratitude to both sets of authors for supplying the stimuli that allows for a direct test of the hypotheses.

[^436]:    ${ }^{2}$ Paolacci, Chandler \& Ipeirotis (2010) for validation of using Mechanical Turk. Note that Asian origin may also include Bangladeshi, Indian, and Pakistani participants. For the current experiment, however, participants were limited to people with Far Eastern origin.
    ${ }^{3}$ Further analysis using $2 \times 3$ two-way ANOVA with all three groups (Black, White, and Asian) revealed no significant interaction: $F(2,92)=1.76, p=.18$. This confirms that all three groups had the same pattern of response.

[^437]:    ${ }^{1}$ https://deepdreamgenerator.com/

[^438]:    ${ }^{2}$ This is known as the N170 event-related potential (ERP).

[^439]:    * Parameter was fit individually for each participant (Mean $\pm$ SD)

[^440]:    ${ }^{1}$ Enactive accounts may be disentangled from each other on the basis of their differing commitment to the necessary involvement of non-neural bodily states in mental imagery (see Foglia \& O'Regan, 2016; cf. Thomas, 1999). For the purposes of the present paper the commonalities rather than the differences between these approaches are highlighted.

[^441]:    ${ }^{2}$ We failed to obtain age for two participants.

[^442]:    ${ }^{3}$ Studies generally show low ambiguity detection rates across a range of more and less complex figures (detection rates are always less than $50 \%$; Brandimonte \& Gerbino, 1993; Hyman \& Neisser, 1991; Peterson, Kihlstrom, Rose, \& Glisky, 1992; Mast \& Kosslyn, 2002)

[^443]:    ${ }^{1}$ https://github.com/guruucsd/DifferentialEncoding/releases/tag/ slotnick

[^444]:    ${ }^{1}$ Sequential Analogical Generalization Engine, Working Memory

[^445]:    ${ }^{1} 4$ participants chose one of the two distractors or did not report any retrieval and were not included in the analyses

[^446]:    ${ }^{2} 2$ participants chose one of the two distractors or did not report any retrieval and were not included in the analyses

[^447]:    ${ }^{1}$ These are distributional semantic models that "frame the vector estimation problem directly as a supervised task, where the weights in a word vector are set to maximize the probability of the contexts in which the word is observed in the corpus" (Baroni et al., 2014, p. 238.)

[^448]:    ${ }^{2}$ Co-occurrences are generally viewed as symmetric relations, and we will keep with that tradition here: tiger co-occurs with bright twice in this sentence, and vice versa.
    ${ }^{3}$ This is the approach of Church \& Hanks (1989) and Islam \& Inkpen (2008), except that their contexts are defined as windows of text (i.e., strings containing $n$ words); the size of the window is an additional parameter for the model.

[^449]:    ${ }^{4}$ Pantel \& Lin (2002)
    ${ }^{5}$ Washtell \& Markert (2009)
    ${ }^{6}$ Evert (2005)
    ${ }^{7}$ Levy, Goldberg, \& Dagan (2015)

[^450]:    ${ }^{8}$ If computing a metric resulted in an undefined value $(\log 0)$, the value of the metric was replaced with zero.
    ${ }^{9}$ That is, a window size of 10 for SGNS and 5 for CBOW (as recommended at https://code.google.com/archive/p/word2vec/), and all other parameters left on their default settings.

[^451]:    ${ }^{10}$ Specifically, the frequency of the lowest-frequency word in each word pair.

[^452]:    ${ }^{1}$ We refer to implicit utterances made by the speaker as implications and inferences made by the listener as inferences.

[^453]:    ${ }^{1}$ To be clear, probabilistic models are also used for purposes other than explanation (e.g., prediction, hypothesis generation). This paper, however, only examines their explanatory import.

[^454]:    ${ }^{2}$ In Oaksford \& Chater 2007, P(q|p) was set to 0.9. See ibid. for the underlying account of conditional inference and for the mathematical details.

[^455]:    ${ }^{3}$ Uncertainty $I\left(M_{i}\right)$ given n mutually exclusive and exhaustive hypotheses, is $-\sum_{i=1}^{n} P\left(M_{i}\right) \log _{2} P\left(M_{i}\right)$.

[^456]:    ${ }_{5}^{4}$ See Goodman et al. (2015) for the modelers' response.
    ${ }^{5}$ See Griffiths \& Tenenbaum (2006) for an empirical study that attempts to directly test the optimality assumption.

[^457]:    ${ }^{6}$ As an anonymous referee pointed out, also the dynamical models used in ecological psychology are often understood as formalizations of affordances. This calls for a systematic comparison between the two modeling paradigms.

[^458]:    ${ }^{1}$ All procedures in the pilot study and experiments were approved by the Johns Hopkins University Institutional Review Board.

[^459]:    * COPD stands for Chronic Obstructive Pulmonary Disease

[^460]:    ${ }^{1}$ Given that the orientation of the two figures was never varied and that the two line figures were never identical, it is important to note that there was always at least one cell that would, on its own, allow the two line figures to be distinguished.

[^461]:    ${ }^{1}$ Our description generalizes across typical candidate inferences approaches (Falkenhainer, 1990; Gentner \& Colhoun, 2010; Gentner \& Markman, 1997) and other related approaches such as copy with substitution \& generation (Holyoak \& Hummel, 2000; Lee \& Holyoak, 2008). For our purposes, the distinctions matter less than the commonalities.

[^462]:    ${ }^{2}$ The current stimulus was designed with a near analogy rather than a distant analogy. We expected the homological nature of mammalian brains to make the analogy prima facie plausible. The rat brain to human brain analogy is often used in experimental study, although here we reverse source and target. Other stimuli (discussed later) have produced consistent, but less pronounced effects for analogies across more distant domains.

[^463]:    ${ }^{3}$ We intended the explanations in each domain to be interpreted by participants as mutually exclusive and exhaustive, but we did not assume this in their response structure. We allowed the model to account for the possibility that both explanations within a domain are correct or that both are incorrect. This compelled the use of the multivariate Dirichlet distribution rather than the Beta distribution.

[^464]:    ${ }^{1}$ This corpus was used because it was pre-analyzed and the ratings were successfully used in previous studies.

[^465]:    ${ }^{2}$ Since the hypothesized effect is due to mere exposure to the text, participants' agreement with the primes was not predicted to affect the results and it is therefore omitted from the analysis. Importantly,

[^466]:    ${ }^{1}$ It must be noticed that transfer of training to multivariate measures of a particular skill (e.g., spatial ability) does not necessarily mean that that skill has been successfully enhanced (Shipstead, Redick, \& Engle, 2012). In fact, the improvement in a

[^467]:    ${ }^{2}$ The item "striking a match" was considered inappropriate for primary school children and thus replaced with the item "dealing cards" (Groen, Whitehouse, Badcock, \& Bishop, 2013).

[^468]:    ${ }^{1}$ In an informative example on Hidden Markov Models (HMMs), Icard and Goodman (2015) present a setting wherein the relevant submodel is infinitely large-an example which highlights what is wrong with the sequential approach stated earlier.
    ${ }^{2}$ The terms "consult" and "retrieve" will be used interchangeably. We elaborate on the rationale behind that in Sec. 5, where we connect our work to Long Term Memory and Working Memory.

[^469]:    ${ }^{3}$ We elaborate more on this in the Discussion section.
    ${ }^{4}$ The term "nested" implies that the thus-far retrieved submodel is subsumed by every later submodel (provided that the reasoner proceeds with the retrieval process).

[^470]:    ${ }^{5}$ More precisely, PL induces a topological order on the nodes of a CBN, with temporal interpretations suggested in Def. 1.

[^471]:    ${ }^{6}$ There are cases, however, that, despite the precedence of cause to effect, quantifying the amount of time between their occurrences may bear no meaning, e.g., when dealing with hypothetical constructs. In such cases, PL should be simply construed as a topological ordering. From a purely computational perspective, PL is a generalization of topological sorting in computer science.

[^472]:    ${ }^{7}$ Taking one step backwards from variable $\boldsymbol{q}$ amounts to retrieving all the parents of $\boldsymbol{q}$.
    ${ }^{8} \mathrm{We}$ do not consider interventions in this work. However, with some modifications, the presented analysis/results can be extended to handle a generic causal query of the form $\mathbb{P}(\mathbf{O}=O \mid \mathbf{E}=$ $E, \operatorname{do}(\mathbf{Z}=Z)$ ) where $\mathbf{Z}$ denotes the set of intervened variables.

[^473]:    ${ }^{9}$ For a discussion on the special case (b), the reader is referred to: https://arxiv.org/pdf/1701.08100
    ${ }^{10}$ For a formal proof of Lemma 1, and the rationale behind Remark 1, the reader is referred to: https://arxiv.org/pdf/1701.08100

[^474]:    ${ }^{11}$ Note that the trend of the upper- and lower-bound curves as well as the size of the intervals shown in Fig. 3(right) are insensitive with regard to the choice of PLs for variables $\left\{\mathbf{x}_{t-i}\right\}_{i=-1}^{+\infty}$.

[^475]:    ${ }^{12}$ The very property that the submodel gets constructed incrementally in a nested fashion guarantees that the obtained lower and upper bounds get tighter as the reasoner adopts smaller ITs; see Fig. 3(left).

[^476]:    ${ }^{1}$ There is in fact converging evidence from formal semantics, syntax, and functional linguistics to support this analysis (Partee, 1987; Adger, 2003; Rijkhoff, 2004).

[^477]:    ${ }^{2}$ Note that we do not wish to make claims about the nature of the gestural elements produced by participants, and use the linguistic terms $\mathrm{N}, \mathrm{Adj}$, Num, and Dem for convenience.

[^478]:    ${ }^{3}$ Different t-tests have different degrees of freedom. This is due to the fact that not all data could be included for each test: some participants produced only incomplete gesture strings.
    ${ }^{4}$ Taken from: http://www.flaticon.com/packs/essential-set-2.

[^479]:    ${ }^{5}$ Adjectives can be incorporated in the noun in sign languages (Sutton-Spence \& Woll, 1999), and it is interesting that participants in Experiment 1 sometimes did this as well, but for the purposes of our experiment we wanted to discourage it.

[^480]:    ${ }^{6}$ Because in this experiment, some participants produced English orders, we ran a separate T-test on the orders that did not follow English structure, and found that subset of the data to be significantly more isomorphic than chance $(\mathrm{t}=3.3347, \mathrm{df}=10, \mathrm{p}<0.01)$.

[^481]:    ${ }^{7}$ Participants were as likely to produce structures like $\mathrm{N}-\mathrm{Num}-\mathrm{Adj}$ and Num-Adj-N as the isomorphic variants N -Adj-Num and Num-Adj-N.

[^482]:    ${ }^{1}$ The scale ranged from "very plausible" (1), to "not plausible,

[^483]:    difficult to imagine" (7).
    ${ }^{2}$ Images used (for both experiments) were taken from an open source database (www.openclipart.org) and pretested for naming.

[^484]:    ${ }^{3}$ This is the required setup for the subsequent extraction of the ICA events from EyeWorks Workload Module software (version 3.12).
    ${ }^{4}$ The middle of the referent noun was calculated by taking the audio duration of the whole word and using its half as the starting

[^485]:    point (for each word individually).
    ${ }^{5}$ NewInspections $\sim$ Constraint $+(1+$ Constraint | Subject) $+(1$ + Constraint | Item), family = "binomial"
    ${ }^{6}$ TargetInspections $\sim$ Gaze $+(1+$ Gaze | Subject $)+(1+$ Gaze | Item), family $=$ "binomial"

[^486]:    ${ }^{7}$ ICA $\sim$ Constraint*Plausibility + Half*Gaze $+(1+$ Constraint*Plausibility || Subject) + (1 + Constraint*Plausibility || Item), family $=$ poisson (link $=$ " $\log ^{\prime \prime}$ )

[^487]:    ${ }^{8}$ None of the students were familiar with Experiment 1.
    ${ }^{9}$ In order to counterbalance the referent nouns the experiment was run in two versions. Version a) included a verb fitting to one noun in the item (spill water vs. sausage), while the verb in version b) fit the other noun (grill sausage vs. water).

[^488]:    ${ }^{10}$ TargetInspections $\sim$ Gaze $+(1+$ Gaze | Subject $)+(1+$ Gaze | Item), family $=$ "binomial"
    ${ }^{1}$ IICA $\sim$ Gaze*Fit + Half*Gaze + ( $1+$ Gaze*Fit + Half*Gaze \| Subject $)+(1+$ Gaze*Fit || Item $)$, family $=$ poisson (link $=$ "log" $)$
    ${ }^{12}$ ICA $\sim$ Gaze*Fit $+(1+$ Gaze*Fit || Subject $)+(1+$ Fit | Item $)$, family $=$ poisson (link $=$ "log")

[^489]:    ${ }^{1}$ The two databases were "Kikuzo II Visual" (online database of Asahi Shimbun; date range: January 1, 1984 to May 23, 2016) and "Yomidasu Rekishikan" (online database of Yomiuri Shimbun; date range: January 1, 1986 to May 23, 2016).

[^490]:    ${ }^{1}$ Choices in linguistic encoding are known to be influenced by aspects of information structure, including contrastive focus,

[^491]:    ${ }^{2}$ Lowercased, replaced punctuation with space, replaced digits with NUM, removed empty lines, replaced tabs with spaces, removed multiple spaces, removed multiple NUMs, replaced umlauts by their conventional character bigrams, and added sentence begin and end markers.
    ${ }^{3}$ A threshold of 15 was selected since this was the highest possible while maintaining a less than $1 \%$ out of vocabulary rate on a different corpus (EUROPARL).

[^492]:    ${ }^{4}$ The sharp difference in perplexity scores between the test corpus and stimuli suggests that the German Wikipedia corpus is not an ideal match for our stimuli. We return to this point in the discussion.
    ${ }^{5}$ Foils were created in a two-stage process. First, a custom Python script randomly selected a foil candidate for each word in each experimental and filler item. Foil candidates were constrained such that they did not appear in bigrams with the correct word at the previous position in the sentence within a large German corpus. Second, each foil was then hand checked by at least two trained native-German linguists to ensure that it was not a grammatical continuation of the sentence. The same foil was used for identical words (or derivationally related words) across conditions.

[^493]:    ${ }^{6}$ Qualitatively identical results are obtained when raw RTs are used.

[^494]:    ${ }^{7}$ Note, however, that the cloze results validate both the stimuli and the RT findings.

[^495]:    ${ }^{1}$ Although there is some disagreement about the aspect of analytic thinking that the CRT measures (e.g. Stupple, Gale, \& Richmond, 2013; Toplak, West, \& Stanovich, 2011), it is a useful tool in measuring analytic ability and reflective processing.
    ${ }^{2}$ Six participants were excluded from the analysis due to incomplete pupil dilation data ( $\mathrm{N}=4$ ) or extreme outlier pupil dilation data ( $\mathrm{N}=2$ )
    ${ }^{3}$ We extended the original IGT to 140 trials to assess the learning effect (e.g. Bagneux, Font, \& Bollon, 2013).

[^496]:    ${ }^{1}$ see https://www.tekscan.com/

[^497]:    ${ }^{1}$ The Radiation problem is about a doctor who wishes to destroy a tumor in his patient's stomach using a ray. However, if he emits the rays at high intensity, the tumor will be destroyed, but so will be the healthy tissues of the patient. If a lower intensity is applied, the tissues will not be affected, but neither will be the tumor. In fact, the solution to this problem is analogous to the base AttackDispersion problem and requires the doctor to emit the rays at lower intensity from different directions simultaneously in order for the concentrated forces of the rays to destroy the tumor.

    2 The Attack-Desperation Problem was considered to be the superficially dissimilar analog of the Radiation Problem in Gick and Holyoak's study (1980). In that story a general wants to capture a fortress located in the center of the country. The problem arises when the general realizes he cannot send his troops all at once due to the mined roads, but if he divides his troops to small

[^498]:    units and attacks from many directions, they will not be affected and the combined forces will capture the fortress.

[^499]:    3 Algorithmic mind can be associated to fluid intelligence capacities. It is a Type 2 processing, which is typically linked to situations that require an optimal performance and a correct answer should be obtained (Stanovich, 2012).

[^500]:    ${ }^{4}$ Standard deviation in seconds.

[^501]:    ${ }^{5} 64$ out of the $67 ; 3$ participants had studied technical specialties

[^502]:    ${ }^{1}$ We also measured response time starting from when the stimuli came on the screen; results were qualitatively similar regardless of the choice of starting time.

[^503]:    ${ }^{2}$ We modeled log-RT using a linear mixed effects model with random effects for participant, trial, and a trial-by-motion direction interaction.

[^504]:    ${ }^{3}$ Because it was unclear how far the ball should travel in the "no motion" condition, we did not include those trials in this analysis.
    ${ }^{4}$ If we include the non-topological trials in this analysis, we still find a relationship between ball travel time and reaction time ( $r=$ $0.29, t(238)=4.6, p=7.2 * 10^{-6}$ ) but do not find statistical evidence that the slopes differ between non-topological, topological towards, and topological away conditions $(F(2,236)=1.01, p=0.37)$. Furthermore, if we remove topological trials that were not fully contained (Porous level 3 and Stopper levels $2 \& 3$ ), this relationship remains $(r=0.34, t(106)=3.7, p=0.00037)$ and we still do not find changes with motion condition $(F(1,105)=0.78, p=0.38)$.

[^505]:    ${ }^{5}$ This analysis considers all topological conditions, including the porousness and stopper trials in which simulation in theory could reach the incorrect goal. However, limiting this analysis to just the size and complexity trials (which were all fully contained) produces the same qualitative results.

[^506]:    ${ }^{1}$ This model did not detect a significant effect of age, which was treated as a continuous variable ( $\beta=0.03, S E=0.02, z=1.43, p$ $=.15$ )

[^507]:    ${ }^{1}$ The effect of $\lambda$ on model performance is gradual enough, and the differences between the different model variants large enough, that not much fine-tuning is required to make our point.

[^508]:    ${ }^{2}$ Franke and Degen (2016) also consider the combination $\mathcal{S}_{1} /$ $\mathcal{H}_{1}$, i.e., a non-iterative model. This would not do any better here, as we see in Table 3 that $\mathcal{H}_{1}$ never makes better predictions than $\mathcal{H}_{2}$.

[^509]:    ${ }^{3}$ Change in deviance between the two models, $\Delta D=59.26$, where deviance is -2 times log likelihood, follows a chi-square distribution with degrees of freedom equal to the number of parameters added to the more complex model. Six prior values must be specified to the labeling model-two each for color, shape and size itemsyielding $\chi^{2}=59.26, d f=6, p<0.001$.

[^510]:    ${ }^{1}$ Note that even the possibility of simultaneously performing multiple responses falls outside the scope of Niv et al.'s (2007) model, according to which all action-latency pairs are serially implemented (i.e., no alternative actions may be executed while the time indicated by the chosen latency passes).
    ${ }^{2}$ Indeed, Niv et al.'s (2007) model hard-codes into the definition of each state several variables that are needed to discover an optimal policy in the environments addressed by the model (e.g., the time elapsed since the last response when modeling interval schedules and the number of presses since the last reward when modeling ratio schedules).

[^511]:    ${ }^{1}$ Only participants who contributed to all trials were included in this analysis

[^512]:    ${ }^{1}$ The method of generating functions by Markov chains also applies to independent Bernoulli trials parameterized the probabilities of single elements (e.g., the probability of heads or tails), and it also generates higher-moment statistics such as variance (e.g., Sun \& Wang, 2015). Here we only present the main results and the exact generating functions are omitted.

[^513]:    ${ }^{1}$ We removed words such as 'excellent' from the task to reduce the likelihood that word valence in this task would influence future participant responses. This quiz block was followed by three more blocks.
    ${ }^{2}$ Due to an off-by-one error in our code, the first statement from the block after the word search task was erroneously displayed before the word search task. We excluded this error trial from the analysis.
    ${ }^{3}$ It should be noted that in this ANOVA, we are treating the block number as a factor variable rather than a numeric variable due to the non-linear relationship between block number and participant response; including the block number as a numeric variable yields qualitatively the same results.

[^514]:    ${ }^{1}$ Traditionally, general object similarity was manipulated as the membership of two objects to a same taxonomic category (e.g., Gentner \& Kurtz, 2006). However, general object similarities could also refer to other commonalities between objects taken in isolation, like, for example, intrinsic and functional properties.

[^515]:    ${ }^{1}$ In pilot testing we found no effect of subtle wording differences for these metaphorical phrases (i.e. "Working on a puzzle" versus "Solving a puzzle"; "Climbing a mountain" versus "Scaling a mountain"). In every case, about $90 \%$ of participants from the general public chose puzzle.
    ${ }^{2}$ Of note, the order of the metaphor preference judgment and the entailment ranking tasks was counterbalanced for the sample from the general public. Since there were no differences in how people responded on the two orderings, we did not counterbalance these tasks for the academics.

[^516]:    ${ }^{1}$ We do not view the boundary between the "literal" and the "metaphorical" as so sharp, "metaphoricity" is best thought of a continuous rather than categorical variable (cf., Rumelhart, 1979). That said, the distinction is still useful and informative in the context of understanding the nature of abstract thought (Lakoff \& Johnson, 1980).

[^517]:    1 MacLaury (1989) characterizes the application of Zapotec body part terms to parts of inanimate objects as analogical or metaphorical mappings from human body parts. An alternate view, taken by Levinson (1994) in his work on Tseltal, is that the terms are general abstractions and not metaphorical.

[^518]:    2 Since the instruction for the practice trial contained the meronyms 'head' and 'leg', a subvocal rehearsal effect on the test trials cannot be ruled out.

[^519]:    ${ }^{3}$ An anonymous reviewer points out that, since the instructions indicated there was a correct answer for each triad, the English speakers' responses may have been motivated by trying to find counterintuitive answers, as in an IQ test.
    ${ }^{4}$ A total of nine Novel Objects were originally designed by the third author and produced for the project Spatial language and cognition in Mesoamerica ('MesoSpace'; NSF Award \#BCS0723694) directed by the second author.

[^520]:    ${ }^{5}$ However, this result is also consistent with subvocal rehearsal being already at ceiling even in the non-primed condition.

[^521]:    ${ }^{1}$ Since in NR more entities bear the mentioned feature, and therefore attention is distributed across more objects, we do not take this difference to reflect any preference for a gricean/contrastive reading of the adjective.

[^522]:    ${ }^{1}$ The participant size was unbalanced because the original authors were concerned about unpredictability. For our purposes it should not impact our results.

[^523]:    ${ }^{2}$ Note that social influence also could have occurred at other places in the model (e.g., search). However, preliminary testing showed that the model actually fit worse when social influence occurred in a stage other than consideration.

[^524]:    ${ }^{1}$ Syllogisms are deductive reasoning problems in which one or more conclusions are derived from two premises. The two premises are categorical propositions which are assumed to be true. For example, the conclusion "All As are Cs" can be drawn from the two premises "All As are Bs" and "All Bs are Cs". The abstract terms, A, B and C can be substituted by concrete categorical terms like "apple" and "fruit".

[^525]:    ${ }^{2}$ The syllogism mentioned here is the MM4 syllogism in the study: Premise 1 as "Most As are Bs"; Premises 2 as "Most Bs are Cs "; and conclusion as "Most Cs are As".

[^526]:    ${ }^{3}$ Quantifiers with higher word frequencies are supposed be processed faster due to the availability heuristic or ease of retrieval, having a faster response time in a spontaneous timed task. A significant decrease of the recognition time for words with higher word frequency (e.g., O'Malley \& Besner, 2008) is generally found. The word frequency measures were taken from the British National Corpus: http://www.natcorp.ox.ac.uk.

[^527]:    ${ }^{4}$ The list can be retrieved from www.cc.uni-freiburg.de/data.

[^528]:    ${ }^{5}$ We performed the Levene's test for equality of variances and the results were not significant, i.e. equal variance can be assumed.

[^529]:    ${ }^{1}$ Left-side bias in face perception refers to the phenomenon that people often judge chimeric faces formed by two left halves of the original face to be more similar to the original face than those formed with two right halves (Brady, Campbell, \& Flaherty, 2005). This effect was also observed in Chinese literates viewing mirrorsymmetric Chinese characters (Hsiao \& Cottrell, 2009).

[^530]:    ${ }^{2}$ The part-whole paradigm (Tanaka \& Farah, 1993) can also demonstrate HP (Maurer et al., 2002). However, it involves memory performance heavily (Piepers \& Robbins, 2012). As the focus of this study is to examine perceptual effects, the complete composite paradigm was used to minimize memory demands and response biases (Richler, Cheung, \& Gauthier, 2011).

[^531]:    ${ }^{3}$ Hsiao and Cottrell (2009) showed that when character halves were misaligned, the HP effect of Chinese characters disappeared, suggesting that the effect reflected the inability to selectively attend to aligned character halves rather than inhibition control.

[^532]:    ${ }^{1}$ The percentages are calculated by Antón-Méndez (p. 129, Table 6 ) and they represent the frequency of the gender pronoun mistake (68 and 10 , respectively) with respect to the total number of pronouns produced where this particular mistake could have occurred.

[^533]:    ${ }^{1}$ Weight, mass, and density as physical notions have certainly been studied in development, across infancy (e.g. Baillargeon 2004), childhood (e.g. Carey 1999 2009), and adulthood (e.g. Hamrick et al. 2016. But there the concern is with questions such as 'When do infants realize big things move small things' and 'When do children understand weight and density are separate', and 'Can adults tell which block is heavy', not 'Do children think being weighed down relates to being sad'.

[^534]:    ${ }^{2}$ That is, with the intuition of the Western adult authors of this paper.
    ${ }^{3}$ Although, in contrast, consider the value and importance currently associated with slim technology products, or human figures. By the same logic, slim would imply less "stuff". Thus, this explanation is insufficient on its own.

[^535]:    ${ }^{4}$ At this age children possess a sufficiently large vocabulary and can correctly point to a heavier object when prompted, but have not received much formal education.
    ${ }^{5}$ One participant designated the heavy box 'The Hulk' and the light box 'Captain America'. Captain America was funny, and The Hulk was not smart.

[^536]:    1 "The text was a two-page, 1385 -word passage entitled Space Travel (see Appendix A) that described effects of space travel on the human body at an introductory level. It was adapted from several sources including two Web-pages from an educational Web

[^537]:    ${ }^{1}$ In this paper we will use the notation $\tilde{M}$ to denote a matrix having the same column vectors of another matrix $M$, but sorted in an inverse order.

[^538]:    ${ }^{1}$ Actually an analogous model of subjective belief update of the cell probabilities (using Dirichlet distributions) which is then compared to ideal patterns yields the same results (not elaborated here).

[^539]:    1 All data and analyses can be accessed on the Open Science Framework at osf.io/y6hqp.

[^540]:    ${ }^{1}$ http://www.brainmap.org/ale/

[^541]:    ${ }^{1}$ A separate, unrelated study comparing forms of scaffolding for essays was run concurrently in the same unit. The target essays did not involve the computer model or graphs, and condition assignment was independent for the two studies. Our results do not show condition effects of the unrelated study.
    ${ }^{2}$ The full unit is available at http://wise.berkeley.edu/ previewproject.html?projectId=18309
    ${ }^{3}$ The post-test is available at http://wise.berkeley.edu/ previewproject.html?projectId=18463

[^542]:    ${ }^{1}$ 'Verbal estimation' is a misnomer stemming from experiments where participants verbalized their estimates of duration, before the introduction of computers.

[^543]:    ${ }^{1}$ Based on $\mathrm{n}=55$ model runs and $\mathrm{n}=55$ human participants.

[^544]:    ${ }^{1}$ The formulation here appears different from (Shafto et al.,

[^545]:    2014) because the setup is different in two ways. First, the query and the label are kept distinct to match the setup of exploration. This gives rise the two likelihoods in Eq. 4, one for $y$ and one for $x$. It also gives rise to the expectation in Eq. 5 with respective to the predictive distribution over $y$. Second, in this setup, the teacher knows the concept from which the pattern is drawn from but not the pattern itself; thus, there is the marginalization over patterns in Eq. 5.
[^546]:    ${ }^{1}$ For the full set of problems as well as example videos for how the model described below solves the different trials please see: https://github.com/iyildirim/ stack-reconfiguration-problems
    ${ }^{2}$ Because several participants had trouble to successfully generate the trials 35-37, we will focus on the first 34 trials.

[^547]:    ${ }^{1}$ All mixed-effects models here and hereafter were implemented using the lmer4 package in R (Bates, Mächler, Bolker, \& Walker, 2015), and all effects were calculated by a likelihood-ratio test against the null-model that only had the random effect of subject (random intercept model).

[^548]:    ${ }^{1}$ A complete implementation of the model, raw data and analyses, and links to the experiments and pre-registration of hypotheses and method can be found at https://github.com/ejyoon/ cogsci2017

[^549]:    ${ }^{2}$ In addition, the model has a few parameters not of theoretical interest. For the purposes of generating model predictions a priori, we assign values to these parameters consistent with the previous literature with this class of models: the speaker optimality parameter ( $\lambda_{1}$ assigned to 2 ); the pragmatic speaker optimality parameter $\left(\lambda_{2}\right.$ to 2 ); the value scale parameter ( $\alpha$ to 1 ) in the utility function; and the parameter governing the cost of producing a negation $(C(u)$ to 2).

[^550]:    ${ }^{1} \mathrm{https}: / /$ code.google.com/archive/p/word2vec/
    ${ }^{2} \mathrm{http}: / /$ download.tensorflow.org/data/questions-words.txt
    ${ }^{3}$ Training with fewer data (e.g., 50\%) does not lead to qualitatively different results.

